Journal of Medical Internet Research

Journal Impact Factorâ ¢ (JIF) (2023): 5.8
Volume 26 (2024) ISSN 1438-8871 Editor in Chief: Gunther Eysenbach, MD, MPH

Contents

Editorials

Redefining Virtual Assistants in Health Care: The Future With Large Language Models (e53225)
Emre Sezgin. 80

Chatbots and COVID-19: Taking Stock of the Lessons Learned (e54840)
Virginia Arnold, Tina Purnat, Robert Marten, Andrew Pattison, Hebe Gouda. 84

Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS) (e52508)
Khaled El Emam, Tiffany Leung, Bradley Malin, William Klement, Gunther Eysenbach. 88

Converge or Collide? Making Sense of a Plethora of Open Data Standards in Health Care (e55779)
Guy Tsafnat, Rachel Dunscombe, Davera Gabriel, Grahame Grieve, Christian Reich. 94

Original Papers

Developing Methods for Assessing Mental Activity Using Human-Smartphone Interactions: Comparative Analysis of Activity Levels and Phase Patterns in General Mental Activities, Working Mental Activities, and Physical Activities (e56144)
Hung-Hsun Chen, Chen Lin, Hsiang-Chih Chang, Jen-Ho Chang, Hai-Hua Chuang, Yu-Hsuan Lin. 99

Rest-Activity Rhythm Differences in Acute Rehabilitation Between Poststroke Patients and Non–Brain Disease Controls: Comparative Study (e49530)
Huey-Wen Liang, Chueh-Hung Wu, Chen Lin, Hsiang-Chih Chang, Yu-Hsuan Lin, Shao-Yu Chen, Wei-Chen Hsu. 111

High-Quality eHealth Websites for Information on Endometriosis: Systematic Search (e48243)
Diksha Sirohi, Cecilia Ng, Niranjan Bidargaddi, Helen Slater, Melissa Parker, Mary Hull, Rebecca O'Hara. 138

Smartphone Apps for Food Purchase Choices: Scoping Review of Designs, Opportunities, and Challenges (e45904)
Remco Benthem de Grave, Christopher Bull, Diogo Monjardino de Souza Monteiro, Eleni Margariti, Gareth McMurchy, Joseph Hutchinson, Jan Smeddinck. 562

Clinical Decision Support System Used in Spinal Disorders: Scoping Review (e53951)
Zheng Toh, Bjørnar Berg, Qin Han, Hwee Hey, Minna Pikkarainen, Margreth Grotle, Hong-Gu He. 673

XSL•FO
RenderX

The SoCAP (Social Communication, Affiliation, and Presence) Taxonomy of Social Features: Scoping Review of Commercially Available eHealth Apps (e49714)
Ian Kwok, Melanie Freedman, Lisa Kamsickas, Emily Lattie, Dershung Yang, Judith Moskowitz. 1972

Application of Chatbots to Help Patients Self-Manage Diabetes: Systematic Review and Meta-Analysis (e60380)
Yibo Wu, Jinzi Zhang, Pu Ge, Tingyu Duan, Junyu Zhou, Yiwei Wu, Yuening Zhang, Siyu Liu, Xinyi Liu, Erya Wan, Xinying Sun. 2949

Perceptions of HIV-Related Comorbidities and Usability of a Virtual Environment for Cardiovascular Disease Prevention Education in Sexual Minority Men With HIV: Formative Phases of a Pilot Randomized Controlled Trial (e57351)
S Ramos, Harmony Reynolds, Constance Johnson, Gail Melkus, Trace Kershaw, Julian Thayer, Allison Vorderstrasse. 3375

Usability of Health Care Price Transparency Data in the United States: Mixed Methods Study (e50629)
Negar Maleki, Balaji Padmanabhan, Kaushik Dutta. 3414

Cost Evaluation of the Ontario Virtual Urgent Care Pilot Program: Population-Based, Matched Cohort Study (e50483)
Jean-Eric Tarride, Justin Hall, Shawn Mondoux, Katie Dainty, Joy McCarron, J Paterson, Lesley Plumptre, Emily Borgundvaag, Howard Ovens, Shelley McLeod. 3426

Policy Guidance for Direct-to-Consumer Genetic Testing Services: Framework Development Study (e47389)
Suzanne Onstwedder, Marleen Jansen, Martina Cornel, Tessel Rigter. 3438

Development and Evaluation of a Smartphone-Based Chatbot Coach to Facilitate a Balanced Lifestyle in Individuals With Headaches (BalanceUP App): Randomized Controlled Trial (e50132)
Sandra Ulrich, Andreas Gantenbein, Viktor Zuber, Agnes Von Wyl, Tobias Kowatsch, Hansjörg Künzli. 4136

Digital Group–Based Intervention for Physical Activity Promotion Among Thai Adults During the COVID-19 Lockdown: Randomized Controlled Trial (e43366)
Nanthawan Pomkai, Piyawat Katewongsa, Aphichat Chamratrithirong, Kanokwan Tharawan, Teeranong Sakulsri, Bhubate Samutachak, Dyah Widyastari, Niramon Rasri, Boonyanuch Wijarn, Yodchanan Wongsawat. 4165

The Evaluation of the GET.ON Nationwide Web-Only Treatment Service for Depression- and Stress-Related Symptoms: Naturalistic Trial (e42976)
Anne Etzelmueller, Elena Heber, Hanne Horvath, Anna Radkovsky, Dirk Lehr, David Ebert. 4178

Effect of a Mobile Game–Based Intervention to Enhance Child Safety: Randomized Controlled Trial (e51908)
Rosa Wong, Keith Tung, Frederick Ho, Wilfred Wong, Chun Chow, Ko Chan, King Fu, Patrick Ip. 4199

Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial (e49839)
Oh Kwon, Mi Lee, Hye Lee, Hyerang Kim, Jae Lee, Yeonsoo Jang. 4212

Comparing the Effectiveness of the Blended Delivery Mode With the Face-to-Face Delivery Mode of Smoking Cessation Treatment: Noninferiority Randomized Controlled Trial (e47040)
Lutz Siemer, Marcel Pieterse, Somaya Ben Allouch, Marloes Postel, Marjolein Brusse-Keizer. 4230

A Social Media–Based Mindfulness Psycho-Behavioral Intervention (MCARE) for Patients With Acute Coronary Syndrome: Randomized Controlled Trial (e48557)
Huijing Zou, Sek Chair, Bilong Feng, Qian Liu, Yu Liu, Yu Cheng, Dan Luo, Xiao Wang, Wei Chen, Leiqing Huang, Yunyan Xianyu, Bing Yang.
4 2 4 8

XSL•FO
RenderX

Internet-Based Cognitive Behavioral Therapy for Children and Adolescents With Dental or Injection Phobia: Randomized Controlled Trial (e42322)
Robert Schibbye, Erik Hedman-Lagerlöf, Viktor Kaldo, Göran Dahllöf, Shervin Shahnavaz. 4264

Efficacy of Internet-Based Cognitive Behavioral Therapy for Subthreshold Depression Among Older Adults in Institutional Long-Term Care Settings: Pragmatic Randomized Controlled Trial (e40187)
Fanqian Kong, Libo Yu, Yanbin Hou, Lijie Zhu, Jing Zhou, Lingjie Huang, Yueer Lv, Li Wang, Li Zhang, Yiling Yang, Yuchen Ying. 4279

Layperson-Supported, Web-Delivered Cognitive Behavioral Therapy for Depression in Older Adults: Randomized Controlled Trial (e53001)
Xiaoling Xiang, Jay Kayser, Skyla Turner, Samson Ash, Joseph Himle. 4294

Effects of a Social Media Intervention on Vaping Intentions: Randomized Dose-Response Experiment (e50741)
William Evans, Jeffrey Bingenheimer, Jennifer Cantrell, Jennifer Kreslake, Shreya Tulsiani, Megumi Ichimiya, Alexander D'Esterre, Raquel Gerard, Madeline Martin, Elizabeth Hair. 4311

Long-Term Effectiveness of a Multi-Strategy Choice Architecture Intervention in Increasing Healthy Food Choices of High-School Students From Online Canteens (Click & Crunch High Schools): Cluster Randomized Controlled Trial (e51108)
Tessa Delaney, Jacklyn Jackson, Christophe Lecathelinais, Tara Clinton-McHarg, Hannah Lamont, Sze Yoong, Luke Wolfenden, Rachel Sutherland, Rebecca Wyse. 4323

Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial (e50330)
Hailey Miller, John Gallis, Miriam Berger, Sandy Askew, Joseph Egger, Melissa Kay, Eric Finkelstein, Mia de Leon, Abigail DeVries, Ashley Brewer, Marni Holder, Gary Bennett. 4336

Effectiveness of a Web-Based Individual Coping and Alcohol Intervention Program for Children of Parents With Alcohol Use Problems: Randomized Controlled Trial (e52118)
Håkan Wall, Helena Hansson, Ulla Zetterlind, Pia Kvillemo, Tobias Elgán. 4349

Digital Therapeutic (Mika) Targeting Distress in Patients With Cancer: Results From a Nationwide Waitlist Randomized Controlled Trial (e51949)
Franziska Springer, Ayline Maier, Michael Friedrich, Jan Raue, Gandolf Finke, Florian Lordick, Guy Montgomery, Peter Esser, Hannah Brock, Anja Mehnert-Theuerkauf. 4364

A Digital Cognitive-Physical Intervention for Attention-Deficit/Hyperactivity Disorder: Randomized Controlled Trial (e55569)
Licong Zhao, Heather Agazzi, Yasong Du, Hongdao Meng, Renya Maku, Ke Li, Peter Aspinall, Cynthia Garvan, Shuanfeng Fang. 4381

Effect of Adding Personalized Instant Messaging Apps to a Brief Smoking Cessation Model in Community Smokers in Hong Kong: Pragmatic Randomized Clinical Trial (e44973)
Yongda Wu, Yee Cheung, Jay Lee, Carlos Wong, Sai Ho, William Li, Ying Yao, Tai Lam, Man Wang. 4395

Comparison of the Working Alliance in Blended Cognitive Behavioral Therapy and Treatment as Usual for Depression in Europe: Secondary Data Analysis of the E-COMPARED Randomized Controlled Trial (e47515)
Asmae Doukani, Matteo Quartagno, Francesco Sera, Caroline Free, Ritsuko Kakuma, Heleen Riper, Annet Kleiboer, Arlinda Cerga-Pashoja, Anneke van Schaik, Cristina Botella, Thomas Berger, Karine Chevreul, Maria Matynia, Tobias Krieger, Jean-Baptiste Hazo, Stasja Draisma, Ingrid Titzler, Naira Topooco, Kim Mathiasen, Kristofer Vernmark, Antoine Urech, Anna Maj, Gerhard Andersson, Matthias Berking, Rosa Baños, Ricardo Araya. 4427

Implementation of a Web-Based Outpatient Asynchronous Consultation Service: Mixed Methods Study (e48092)
Magdalena Rzewuska Díaz, Louise Locock, Andrew Keen, Mike Melvin, Anthony Myhill, Craig Ramsay. 4445

XSL•FO
RenderX

Six-Month Outcomes of a Theory- and Technology-Enhanced Physical Activity Intervention for Latina Women (Pasos Hacia La Salud II): Randomized Controlled Trial (e51708)
Lauren Connell Bohlen, Shira Dunsiger, Tayla von Ash, Britta Larsen, Dori Pekmezi, Becky Marquez, Tanya Benitez, Andrea Mendoza-Vasconez, Sheri Hartman, David Williams, Bess Marcus. 4461

App Engagement as a Predictor of Weight Loss in Blended-Care Interventions: Retrospective Observational Study Using Large-Scale Real-World Data (e45469)
Marco Lehmann, Lucy Jones, Felix Schirmann. 4474

Development of an Artificial Intelligence–Based Tailored Mobile Intervention for Nurse Burnout: Single-Arm Trial (e54029)
Aram Cho, Chiyoung Cha, Gumhee Baek. 4487

A Symptom-Checker for Adult Patients Visiting an Interdisciplinary Emergency Care Center and the Safety of Patient Self-Triage: Real-Life Prospective Evaluation (e58157)
Andreas Meer, Philipp Rahm, Markus Schwendinger, Michael Vock, Bettina Grunder, Jacopo Demurtas, Jonas Rutishauser. 4504

Digital Interventions to Modify Skin Cancer Risk Behaviors in a National Sample of Young Adults: Randomized Controlled Trial (e55831)
Carolyn Heckman, Anna Mitarotondo, Yong Lin, Olga Khavjou, Mary Riley, Sharon Manne, Amy Yaroch, Zhaomeng Niu, Karen Glanz. 4517

The Effects of a Complex Interactive Multimodal Intervention on Personalized Stress Management Among Health Care Workers in China: Nonrandomized Controlled Study (e45422)
Wenhua Liu, Quan Wang, Danli Zheng, Junhua Mei, Jiajia Lu, Guohua Chen, Wei Wang, Fengfei Ding. 4534

Evaluating the Acceptability of the Drink Less App and the National Health Service Alcohol Advice Web Page: Qualitative Interview Process Evaluation (e42319)
Melissa Oldham, Larisa-Maria Dina, Gemma Loebenberg, Olga Perski, Jamie Brown, Colin Angus, Emma Beard, Robyn Burton, Matt Field, Felix Greaves, Matthew Hickman, Eileen Kaner, Susan Michie, Marcus Munafò, Elena Pizzo, Claire Garnett. 4547

A Web-Based Resilience-Enhancing Program to Improve Resilience, Physical Activity, and Well-being in Geriatric Population: Randomized Controlled Trial (e53450)
Yi-Chen Wu, Shu-Fen Shen, Liang-Kung Chen, Heng-Hsin Tung. 4561

Efficacy of Mobile App–Based Cognitive Behavioral Therapy for Insomnia: Multicenter, Single-Blind Randomized Clinical Trial (e50555)
Jiyoon Shin, Sujin Kim, Jooyoung Lee, Hyerin Gu, Jihye Ahn, Chowon Park, Mincheol Seo, Jeong Jeon, Ha Lee, Ji Yeom, Sojeong Kim, Yeaseul Yoon, Heon-Jeong Lee, Seog Kim, Yu Lee. 4584

Remote Exercise Training Intervention During the COVID-19 Pandemic: Randomized Controlled Trial (e53145)
Keito Philippi, Peter Zeier, Alexandra Brahmer, Elmo Neuberger, Magdalena Sandner, Matthias Hagenah, Thilo Porten, Regina Lenz, David Ochmann, Florian Wedekink, Jörg Wischhusen, Beat Lutz, Klaus Lieb, Michèle Wessa, Perikles Simon. 4599

Sex Differences in Pain Sensitivity in a Dutch Cohort: Cross-Sectional and Web-Based Multidimensional Study (e53926)
Rachel Smits, Selina van der Wal, Regina van Boekel, Hans Timmerman, Ewald Bronkhorst, Diana Abrar, Kris Vissers, Esmeralda Blaney Davidson, Monique Steegers. 4612

Effect of a Mobile Health–Based Remote Interaction Management Intervention on the Quality of Life and Self-Management Behavior of Patients With Low Anterior Resection Syndrome: Randomized Controlled Trial (e53909)
Peng Zhou, Hui Li, Xueying Pang, Ting Wang, Yan Wang, Hongye He, Dongmei Zhuang, Furong Zhu, Rui Zhu, Shaohua Hu. 4626

XSL•FO
RenderX

Association of Wearable Device–Measured Step Volume and Variability With Blood Pressure in Older Chinese Adults: Mobile-Based Longitudinal Observational Study (e50075)
Han Xiao, Zechen Zhou, Yujia Ma, Xiaoyi Li, Kexin Ding, Xiaotong Dai, Dafang Chen. 4639

Expectations and Preferences for Digital Cessation Treatment: Multimethods Study Among Older Adults Who Smoke Cigarettes (e52919)
Margaret Fahey, Mathew Carpenter, Riley O'Neal, Kinsey Pebley, Melissa Schick, Emily Ware, Benjamin Toll, Jennifer Dahne. 4655

Sociodemographic, Health, and Lifestyle-Related Characteristics Associated With the Commencement and Completion of a Web-Based Lifestyle Educational Program for People With Multiple Sclerosis: Randomized Controlled Trial (e58253)
Jeanette Reece, Maggie Yu, William Bevens, Steve Simpson-Yap, Rebekah Davenport, George Jelinek, Sandra Neate. 4668

Web-Based Coping Skills Training and Coach Support for Women Living With a Partner With an Alcohol Use Disorder: Randomized Controlled Trial (e56119)
Robert Rychtarik, Brian Danaher, Neil McGillicuddy, Milagra Tyler, Christopher Barrick, Florence Leong, Derek Kosty. 4695

Investigating Patient Perspectives on Using eHealth Technologies for the Self-Management of Inflammatory Bowel Disease: Mixed Methods Study (e53512)
Sander Hermsen, Danielle Tump, Eva Wentink, Marjolijn Duijvestein. 4713

Comparing the Efficacy of 2 WeChat Mini Programs in Reducing Nonmarital Heterosexual Contact by Male Factory Workers: Randomized Controlled Trial (e49362)
Kechun Zhang, Bolin Cao, Yuan Fang, Xue Liang, Danhua Ye, Ya Chen, Ruilan Zhong, He Cao, Tian Hu, Ting Li, Yong Cai, Huachun Zou, Zixin Wang. 4735

A Web-Based Antiretroviral Therapy Adherence Intervention (Thrive With Me) in a Community-Recruited Sample of Sexual Minority Men Living With HIV: Results of a Randomized Controlled Study (e53819)
Keith Horvath, Sara Lammert, Darin Erickson, K Amico, Ali Talan, Ore Shalhav, Christina Sun, H Rendina. 4753

Efficacy of WeChat-Based Digital Intervention Versus Metformin in Women With Polycystic Ovary Syndrome: Randomized Controlled Trial (e55883)
Diliqingna Dilimulati, Xiaowen Shao, Lihua Wang, Meili Cai, Yuqin Zhang, Jiayi Lu, Yao Wang, Hongying Liu, Ming Kuang, Haibing Chen, Manna Zhang, Shen Qu. 4768

Enhancing Patient Safety Through an Integrated Internet of Things Patient Care System: Large Quasi-Experimental Study on Fall Prevention (e58380)
Ming-Huan Wen, Po-Yin Chen, Shirling Lin, Ching-Wen Lien, Sheng-Hsiang Tu, Ching-Yi Chueh, Ying-Fang Wu, Kelvin Tan Cheng Kian, Yeh-Liang Hsu, Dorothy Bai. 4784

The Effectiveness of a Digital Mental Fitness Program (Positive Intelligence) on Perceived Stress, Self-Compassion, and Ruminative Thinking of Occupational Therapy Undergraduate Students: Longitudinal Study (e49505)
Shermain Puah, Ching Pua, Jing Shi, Sok Lim. 4795

Effectiveness of a Digital Health Intervention Leveraging Reinforcement Learning: Results From the Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) Randomized Clinical Trial (e60834)
Adrian Aguilera, Marvyn Arévalo Avalos, Jing Xu, Bibhas Chakraborty, Caroline Figueroa, Faviola Garcia, Karina Rosales, Rosa Hernandez-Ramos, Chris Karr, Joseph Williams, Lisa Ochoa-Frongia, Urmimala Sarkar, Elad Yom-Tov, Courtney Lyles. 4809

Comparing a Digital Health Check With Traditional Nurse-Led Health Examinations Among Long-Term Unemployed Individuals: Comparison Study (e49802)
Venla Raussi, Sari Kujala, Iiris Hörhammer, Kaisa Savolainen, Reija Autio, Tuomas Koskela. 4834

XSL•FO
RenderX

Outcomes of a Remotely Delivered Complementary and Integrative Health Partnered Intervention to Improve Chronic Pain and Posttraumatic Stress Disorder Symptoms: Randomized Controlled Trial (e57322)
Jolie Haun, Christopher Fowler, Hari Venkatachalam, Amy Alman, Lisa Ballistrea, Tali Schneider, Rachel Benzinger, Christine Melillo, Neil Alexander, S Klanchar, William Lapcevic, Matthew Bair, Stephanie Taylor, Jennifer Murphy, Dustin French. 4845

Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study (e55075)
Angela Karellis, Duncan Webster, Jean Boulanger, Kate Harland, Paige Feltmate, Stefanie Materniak, Gabriel Daunais-Laurin, Christine Mesa, Olivia Vaikla, John Kim, Nitika Pant Pai. 4869

Effects of Monitoring Frailty Through a Mobile/Web-Based Application and a Sensor Kit to Prevent Functional Decline in Frail and Prefrail Older Adults: FACET (Frailty Care and Well Function) Pilot Randomized Controlled Trial (e58312)
Myriam Valdés-Aragonés, Rodrigo Pérez-Rodríguez, José Carnicero, Pedro Moreno-Sánchez, Myriam Oviedo-Briones, Elena Villalba-Mora, Pedro Abizanda-Soler, Leocadio Rodríguez-Mañas. 4889

A Universal Digital Stress Management Intervention for Employees: Randomized Controlled Trial with Health-Economic Evaluation (e48481)
Johanna Freund, Filip Smit, Dirk Lehr, Anna-Carlotta Zarski, Matthias Berking, Heleen Riper, Burkhardt Funk, David Ebert, Claudia Buntrock.
4 9 0 5

Effects of HIV Self-Testing on Testing Promotion and Risk Behavior Reduction Among Transgender Women in China: Randomized Controlled Trial (e58591)
Yan-Yan Zhu, Ze-Hao Ye, Zhen-Xing Chu, Yingjie Liu, Jie Wei, Le Jia, Yong-Jun Jiang, Hong Shang, Qing-Hai Hu. 4939

Effectiveness of King's Theory of Goal Attainment in Blood Glucose Management for Newly Diagnosed Patients With Type 2 Diabetes: Randomized Controlled Trial (e59142)
Man Yan, Yingchun Yu, Shuping Li, Peiling Zhang, Jiaxiang Yu. 4952

Effectiveness of the Offer of the Smoke Free Smartphone App Compared With No Intervention for Smoking Cessation: Pragmatic Randomized Controlled Trial (e50963)
Sarah Jackson, Dimitra Kale, Emma Beard, Olga Perski, Robert West, Jamie Brown. 4978

The Impact of Comment Slant and Comment Tone on Digital Health Communication Among Polarized Publics: A Web-Based Survey Experiment (e57967)
Fangcao Lu, Caixie Tu. 4990

Digital Exclusion and Cognitive Function in Elderly Populations in Developing Countries: Insights Derived From 2 Longitudinal Cohort Studies (e56636)
Sainan Duan, Dongxu Chen, Jinping Wang, Mohammed Paramboor, Zhen Xia, Wanting Xu, Kun Han, Tao Zhu, Xiaoqin Jiang. 5007

A Personalized, Transdiagnostic Smartphone App (Mello) Targeting Repetitive Negative Thinking for Depression and Anxiety: Qualitative Analysis of Young People's Experience (e63732)
Lee Valentine, Chelsea Arnold, Jennifer Nicholas, Emily Castagnini, Jessi Malouf, Mario Alvarez-Jimenez, Imogen Bell. 5064

The Digital Divide and Cognitive Disparities Among Older Adults: Community-Based Cohort Study in China (e59684)
Yumeng Li, Chen Liu, Jiaqing Sun, Junying Zhang, Xin Li, Zhanjun Zhang. 5079

Predicting Early Dropout in a Digital Tobacco Cessation Intervention: Replication and Extension Study (e54248)
Linda Yu, Michael Amato, George Papandonatos, Sarah Cha, Amanda Graham. 5091

XSL•FO
RenderX

Internet-Delivered Psychoeducation (SCOPE) for Transition-Aged Autistic Youth: Pragmatic Randomized Controlled Trial (e49305)
Anna Backman, Lise Roll-Pettersson, Are Mellblom, Elisabet Norman-Claesson, Emma Sundqvist, Eric Zander, Sarah Vigerland, Tatja Hirvikoski. 5100

Messenger App–Based Information Provision for Promoting Social Participation to Enhance Well-Being Among Community-Dwelling Adults: Randomized Controlled Trial (e57205)
Atsushi Nakagomi, Noriyuki Abe, Yu-Ru Chen, Kazushige Ide, Shuhei Kobayashi, Masamichi Hanazato, Katsunori Kondo. 5115

A Smart Glass Telemedicine Application for Prehospital Communication: User-Centered Design Study (e53157)
Zhan Zhang, Enze Bai, Yincao Xu, Aram Stepanian, Jared Kutzin, Kathleen Adelgais, Mustafa Ozkaynak. 5136

Patients' and Clinicians' Experiences Using a Real-Time Remote Monitoring System for Chemotherapy Symptom Management (ASyMS): Qualitative Study (e53834)
Lisa McCann, Liane Lewis, Olubukola Oduntan, Jenny Harris, Andrew Darley, Geir Berg, Simone Lubowitzki, Katy Cheevers, Morven Miller, Jo Armes, Emma Ream, Patricia Fox, Eileen Furlong, Alexander Gaiger, Grigorios Kotronoulas, Elisabeth Patiraki, Stylianos Katsaragakis, Paul McCrone, Christine Miaskowski, Antonella Cardone, Dawn Orr, Adrian Flowerday, Simon Skene, Margaret Moore, Nicosha De Souza, Peter Donnan, Roma Maguire. 5154

Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers (e60353)
Anna Jolliff, Richard Holden, Rupa Valdez, Ryan Coller, Himalaya Patel, Matthew Zuraw, Anna Linden, Aaron Ganci, Christian Elliott, Nicole Werner. 5168

Effects of 10 Weeks of Walking With Mobile Step-Tracking Apps on Body Composition, Fitness, and Psychological State in Adolescents Who Are Overweight and Obese: Randomized Controlled Trial (e55243)
Adrián Mateo-Orcajada, Cristina Ponce-Ramírez, Lucía Abenza-Cano, Raquel Vaquero-Cristóbal. 5220

Personalized Mobile App–Based Program for Preparation and Recovery After Radical Prostatectomy: Initial Evidence for Improved Outcomes From a Prospective Nonrandomized Study (e55429)
Alberto Martini, Claudia Kesch, Alae Touzani, Giorgio Calleris, Bogdan Buhas, Rawad Abou-Zahr, Razvan-George Rahota, Benjamin Pradère, Christophe Tollon, Jean-Baptiste Beauval, Guillaume Ploussard. 5244

Effects of a Smartphone-Based Breastfeeding Coparenting Intervention Program on Breastfeeding-Related Outcomes in Couples During First Pregnancy: Randomized Controlled Trial (e51566)
Yi-Yan Huang, Rong Wang, Wei-Peng Huang, Tian Wu, Shi-Yun Wang, Sharon R. Redding, Yan-Qiong Ouyang. 5252

Association Between the Composite Cardiovascular Risk and mHealth Use Among Adults in the 2017-2020 Health Information National Trends Survey: Cross-Sectional Study (e46277)
Yuling Chen, Ruth-Alma Turkson-Ocran, Binu Koirala, Patricia Davidson, Yvonne Commodore-Mensah, Cheryl Himmelfarb. 5267

Preliminary Efficacy of a Digital Intervention for Adolescent Depression: Randomized Controlled Trial (e48467)
Emily Peake, Ian Miller, Jessica Flannery, Lang Chen, Jessica Lake, Aarthi Padmanabhan. 5280

Assessing the Influence of Patient Empowerment Gained Through Mental Health Apps on Patient Trust in the Health Care Provider and Patient Compliance With the Recommended Treatment: Cross-sectional Study (e48182)
Julien François, Anne-Françoise Audrain-Pontevia, Sana Boudhraâ, Stéphane Vial. 5302

Media Use Behavior Mediates the Association Between Family Health and Intention to Use Mobile Health Devices Among Older Adults: Cross-Sectional Study (e50012)
Jinghui Chang, Yanshan Mai, Dayi Zhang, Xixi Yang, Anqi Li, Wende Yan, Yibo Wu, Jiangyun Chen. 5315

XSL•FO
RenderX

Living Lab Data of Patient Needs and Expectations for eHealth-Based Cardiac Rehabilitation in Germany and Spain From the TIMELY Study: Cross-Sectional Analysis (e53991)
Boris Schmitz, Svenja Wirtz, Manuela Sestayo-Fernández, Hendrik Schäfer, Emma Douma, Marta Alonso Vazquez, Violeta González-Salvado, Mirela Habibovic, Dimitris Gatsios, Willem Kop, Carlos Peña-Gil, Frank Mooren. 5331

Assessing the Effectiveness of STAPP@Work, a Self-Management Mobile App, in Reducing Work Stress and Preventing Burnout: Single-Case Experimental Design Study (e48883)
Sevda Demirel, Yvette Roke, Adriaan Hoogendoorn, Jamie Hoefakker, Kirsten Hoeberichts, Peter van Harten. 5345

Mobile-Delivered Mindfulness Intervention on Anxiety Level Among College Athletes: Randomized Controlled Trial (e40406)
Yu Gao, Lu Shi, Ning Fu, Nan Yang, Tracy Weeks-Gariepy, Yuping Mao. 5360

Understanding Heterogeneity in Individual Responses to Digital Lifestyle Intervention Through Self-Monitoring Adherence Trajectories in Adults With Overweight or Obesity: Secondary Analysis of a 6-Month Randomized Controlled Trial (e53294)
Shiyu Li, Yan Du, Hongyu Miao, Kumar Sharma, Chengdong Li, Zenong Yin, Bradley Brimhall, Jing Wang. 5372

Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation (e46098)
Hyo-Jeong Ahn, Eue-Keun Choi, So-Ryoung Lee, Soonil Kwon, Hee-Seok Song, Young-Shin Lee, Seil Oh. 5386

Assessment of the Barriers and Enablers of the Use of mHealth Systems in Sub-Saharan Africa According to the Perceptions of Patients, Physicians, and Health Care Executives in Ethiopia: Qualitative Study (e50337)
Genet Aboye, Gizeaddis Simegn, Jean-Marie Aerts. 5398

Effectiveness of the Minder Mobile Mental Health and Substance Use Intervention for University Students: Randomized Controlled Trial (e54287)
Melissa Vereschagin, Angel Wang, Chris Richardson, Hui Xie, Richard Munthali, Kristen Hudec, Calista Leung, Katharine Wojcik, Lonna Munro, Priyanka Halli, Ronald Kessler, Daniel Vigo. 5420

The Impact of Digital Self-Monitoring of Weight on Improving Diabetes Clinical Outcomes: Quasi-Randomized Study (e54940)
Yifat Fundoiano-Hershcovitz, Marilyn Ritholz, David Horwitz, Ephraim Behar, Omar Manejwala, Pavel Goldstein. 5438

The Effectiveness of a Digital App for Reduction of Clinical Symptoms in Individuals With Panic Disorder: Randomized Controlled Trial (e51428)
KunJung Kim, Hyunchan Hwang, Sujin Bae, Sun Kim, Doug Han. 5478

Problems and Barriers Related to the Use of mHealth Apps From the Perspective of Patients: Focus Group and Interview Study (e49982)
Godwin Giebel, Carina Abels, Felix Plescher, Christian Speckemeier, Nils Schrader, Kirstin Börchers, Jürgen Wasem, Silke Neusser, Nikola Blase. 5490

Smartphone-Based Speech Therapy for Poststroke Dysarthria: Pilot Randomized Controlled Trial Evaluating Efficacy and Feasibility (e56417)
Yuyoung Kim, Minjung Kim, Jinwoo Kim, Tae-Jin Song. 5502

Effectiveness of a Smartphone App–Based Intervention With Bluetooth-Connected Monitoring Devices and a Feedback System in Heart Failure (SMART-HF Trial): Randomized Controlled Trial (e52075)
Minjae Yoon, Seonhwa Lee, Jah Choi, Mi-Hyang Jung, Jong-Chan Youn, Chi Shim, Jin-Oh Choi, Eung Kim, Hyungseop Kim, Byung-Su Yoo, Yeon Son, Dong-Ju Choi. 5518

XSL•FO
RenderX

Digital Self-Management Platform for Adult Asthma: Randomized Attention-Placebo Controlled Trial (e50855)
Aaron Kandola, Kyra Edwards, Joris Straatman, Bettina Dührkoop, Bettina Hein, Joseph Hayes. 5532

Evaluating the World Health Organization's SkinNTDs App as a Training Tool for Skin Neglected Tropical Diseases in Ghana and Kenya: Cross-Sectional Study (e51628)
Mireia Cano, José Ruiz-Postigo, Paul Macharia, Yaw Ampem Amoako, Richard Odame Phillips, Esther Kinyeru, Carme Carrion. 5544

Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation (e46036)
Han Chew, Nicholas Chew, Shaun Loong, Su Lim, Wai Tam, Yip Chin, Ariana Chao, Georgios Dimitriadis, Yujia Gao, Jimmy So, Asim Shabbir, Kee Ngiam. 5564

Characteristic Changes of the Stance-Phase Plantar Pressure Curve When Walking Uphill and Downhill: Cross-Sectional Study (e44948)
Christian Wolff, Patrick Steinheimer, Elke Warmerdam, Tim Dahmen, Philipp Slusallek, Christian Schlinkmann, Fei Chen, Marcel Orth, Tim Pohlemann, Bergita Ganse. 5578

Just-in-Time Adaptive Intervention for Stabilizing Sleep Hours of Japanese Workers: Microrandomized Trial (e49669)
Hiroki Takeuchi, Tetsuro Ishizawa, Akifumi Kishi, Toru Nakamura, Kazuhiro Yoshiuchi, Yoshiharu Yamamoto. 5609

Association of Smartwatch-Based Heart Rate and Physical Activity With Cardiorespiratory Fitness Measures in the Community: Cohort Study (e56676)
Yuankai Zhang, Xuzhi Wang, Chathurangi Pathiravasan, Nicole Spartano, Honghuang Lin, Belinda Borrelli, Emelia Benjamin, David McManus, Martin Larson, Ramachandran Vasan, Ravi Shah, Gregory Lewis, Chunyu Liu, Joanne Murabito, Matthew Nayor. 5627

Longitudinal Assessment of Seasonal Impacts and Depression Associations on Circadian Rhythm Using Multimodal Wearable Sensing: Retrospective Analysis (e55302)
Yuezhou Zhang, Amos Folarin, Shaoxiong Sun, Nicholas Cummins, Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Pauline Conde, Heet Sankesara, Petroula Laiou, Faith Matcham, Katie White, Carolin Oetzmann, Femke Lamers, Sara Siddi, Sara Simblett, Srinivasan Vairavan, Inez Myin-Germeys, David Mohr, Til Wykes, Josep Haro, Peter Annas, Brenda Penninx, Vaibhav Narayan, Matthew Hotopf, Richard Dobson, RADAR-CNS consortium. 5640

Web-Based Warfarin Management (Alfalfa App) Versus Traditional Warfarin Management: Multicenter Prospective Cohort Study (e46319)
Wenfei Chen, Jiana Chen, Shaojun Jiang, Chunhua Wang, Jinhua Zhang. 5656

An Effective Deep Learning Framework for Fall Detection: Model Development and Study Design (e56750)
Jinxi Zhang, Zhen Li, Yu Liu, Jian Li, Hualong Qiu, Mohan Li, Guohui Hou, Zhixiong Zhou. 5668

Investigating Best Practices for Ecological Momentary Assessment: Nationwide Factorial Experiment (e50275)
Michael Businelle, Emily Hébert, Dingjing Shi, Lizbeth Benson, Krista Kezbers, Sarah Tonkin, Megan Piper, Tianchen Qian. 5682

Multicomponent Support Program for Secondary Prevention of Stroke Using Digital Health Technology: Co-Design Study With People Living With Stroke or Transient Ischemic Attack (e54604)
David Silvera-Tawil, Jan Cameron, Jane Li, Marlien Varnfield, Liam Allan, Mitch Harris, Natasha Lannin, Christian Redd, Dominique Cadilhac. 5 6 9 7

Establishing an Evaluation Indicator System for User Satisfaction With Hypertension Management Apps: Combining User-Generated Content and Analytic Hierarchy Process (e60773)
Yunfan He, Han Chen, Peng Xiang, Min Zhao, Yingjun Li, Yongcheng Liu, Tong Wang, Jun Liang, Jianbo Lei. 5719

Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study (e54617)
Daun Shin, Hyoseung Kim, Seunghwan Lee, Younhee Cho, Whanbo Jung. 5736

XSL•FO
RenderX

mHealth Physical Activity and Patient-Reported Outcomes in Patients With Inflammatory Bowel Diseases: Cluster Analysis (e48020)
Ashley Griffin, Lucas Mentch, Feng-Chang Lin, Arlene Chung. 5748

A Mobile Gaming App to Train Teenage Mothers on Appropriate Child Feeding Practices: Development and Validation Study (e53560)
Mercy Sosanya, Folake Samuel, Sadia Bashir, Victoria Omoera, Jeanne Freeland-Graves. 5761

Mobile Phones and HIV Testing: Multicountry Instrumental Variable Analysis From Sub-Saharan Africa (e48794)
Francesco Iacoella, Nyasha Tirivayi. 5780

Effects of a Mobile App to Promote Social Participation on Older Adults: Randomized Controlled Trial (e64196)
Kenjiro Kawaguchi, Atsushi Nakagomi, Kazushige Ide, Katsunori Kondo. 5797

Health Maintenance Organization–mHealth Versus Face-to-Face Interaction for Health Care in Israel: Cross-Sectional Web-Based Survey Study (e55350)
Avi Zigdon, Moti Zwilling, Ofek Zigdon, Orna Reges. 5811

Challenges and Alternatives to Evaluation Methods and Regulation Approaches for Medical Apps as Mobile Medical Devices: International and Multidisciplinary Focus Group Discussion (e54814)
Laura Maaß, Robert Hrynyschyn, Martin Lange, Alexandra Löwe, Kathrin Burdenski, Kaley Butten, Sebastian Vorberg, Mariam Hachem, Aldo Gorga, Vittorio Grieco, Vincenzo Restivo, Giuseppe Vella, Marlien Varnfield, Felix Holl. 5823

Brief Peer-Supported Web-Based Skills Training in Affective and Interpersonal Regulation (BPS webSTAIR) for Trauma-Exposed Veterans in the Community: Randomized Controlled Trial (e52130)
Laura Ong, Sarah Speicher, Diana Villasenor, Jamie Kim, Adam Jacobs, Kathryn Macia, Marylene Cloitre. 5841

A Behavioral Change–Based Mobile Intervention for Promoting Regular Physical Activity in Medical Rehabilitation Maintenance of Patients With Coronary Artery Disease: Controlled Trial (e56480)
Melina Waranski, René Garbsch, Mona Kotewitsch, Marc Teschler, Boris Schmitz, Frank Mooren. 5855

Menstrual Cycle Management and Period Tracker App Use in Millennial and Generation Z Individuals: Mixed Methods Study (e53146)
Minji Hong, Vasuki Rajaguru, KyungYi Kim, Suk-Yong Jang, Sang Lee. 5874

Associations of Wearable Activity Tracker Use With Physical Activity and Health Outcomes in Patients With Cancer: Findings from a Population-Based Survey Study (e51291)
Weijiao Zhou, Shaomei Shang, Youmin Cho. 5891

Digital Phenotypes of Mobile Keyboard Backspace Rates and Their Associations With Symptoms of Mood Disorder: Algorithm Development and Validation (e51269)
Qimin Liu, Emma Ning, Mindy Ross, Andrea Cladek, Sarah Kabir, Amruta Barve, Ellyn Kennelly, Faraz Hussain, Jennifer Duffecy, Scott Langenecker, Theresa Nguyen, Theja Tulabandhula, John Zulueta, Alexander Demos, Alex Leow, Olusola Ajilore. 5903

Discriminant Power of Smartphone-Derived Keystroke Dynamics for Mild Cognitive Impairment Compared to a Neuropsychological Screening Test: Cross-Sectional Study (e59247)
Jin-Hyuck Park. 5917

Experiences With mHealth Use Among Patient-Caregiver Dyads With Chronic Heart Failure: Qualitative Study (e57115)
Xiaorong Jin, Yimei Zhang, Min Zhou, Xiong Zhang, Qian Mei, Yangjuan Bai, Wei Wei, Fang Ma. 5942

XSL•FO
RenderX

Efficacy of a Wearable Activity Tracker With Step-by-Step Goal-Setting on Older Adults' Physical Activity and Sarcopenia Indicators: Clustered Trial (e60183)
Mu-Hsing Ho, Chi-Yuan Peng, Yung Liao, Hsin-Yen Yen. 5955

Implementation of a Mobile Health Approach to a Long-Lasting Insecticidal Net Uptake Intervention for Malaria Prevention Among Pregnant Women in Tanzania: Process Evaluation of the Hati Salama (HASA) Randomized Controlled Trial Study (e51527)
Trinity Vey, Eleonora Kinnicutt, Nicola West, Jessica Sleeth, Kenneth Nchimbi, Karen Yeates. 5966

The Effectiveness of Remote Exercise Rehabilitation Based on the "SCeiP" Model in Homebound Patients With Coronary Heart Disease: Randomized Controlled Trial (e56552)
Dandan Xu, Dongmei Xu, Lan Wei, Zhipeng Bao, Shengen Liao, Xinyue Zhang. 5977

Service Quality and Patient Satisfaction of Internet Hospitals in China: Cross-Sectional Evaluation With the Service Quality Questionnaire (e55140)
Tao Han, Qinpeng Wei, Ruike Wang, Yijin Cai, Hongyi Zhu, Jiani Chen, Zhiruo Zhang, Sisi Li. 5992

Design Guidelines for Improving Mobile Sensing Data Collection: Prospective Mixed Methods Study (e55694)
Christopher Slade, Roberto Benzo, Peter Washington. 6004

Smartphone App for Improving Self-Awareness of Adherence to Edoxaban Treatment in Patients With Atrial Fibrillation (ADHERE-App Trial): Randomized Controlled Trial (e65010)
Minjae Yoon, Ji Lee, In-Cheol Kim, Ju-Hee Lee, Mi-Na Kim, Hack-Lyoung Kim, Sunki Lee, In Kim, Seonghoon Choi, Sung-Ji Park, Taeho Hur, Musarrat Hussain, Sungyoung Lee, Dong-Ju Choi. 6019

Long-Term Efficacy of an AI-Based Health Coaching Mobile App in Slowing the Progression of Nondialysis-Dependent Chronic Kidney Disease: Retrospective Cohort Study (e54206)
Jianwei Ma, Jiangyuan Wang, Jiapei Ying, Shasha Xie, Qin Su, Tianmeng Zhou, Fuman Han, Jiayan Xu, Siyi Zhu, Chenyi Yuan, Ziyuan Huang, Jingfang Xu, Xuyong Chen, Xueyan Bian. 6030

Characterizing Performance on a Suite of English-Language NeuroUX Mobile Cognitive Tests in a US Adult Sample: Ecological Momentary Cognitive Testing Study (e51978)
Emily Paolillo, Jessica Bomyea, Colin Depp, Ashley Henneghan, Anunay Raj, Raeanne Moore. 6044

Recruitment for Voluntary Video and Mobile HIV Testing on Social Media Platforms During the COVID-19 Pandemic: Cross-Sectional Study (e54420)
Piao-Yi Chiou, Wei-Wen Tsao, Chia-Lin Li, Jheng-Min Yu, Wen-Han Su, Zhi-Hua Liu, Cheng-Ru He, Yu-Chun Chang, Yi-Hsuan Tsai. 6063

Challenges and Opportunities in Digital Screening for Hypertension and Diabetes Among Community Groups of Older Adults in Vietnam: Mixed Methods Study (e54127)
Trang Nong, Giang Nguyen, Alexander Lepe, Thuy Tran, Lan Nguyen, Jaap Koot. 6077

Predicting and Monitoring Symptoms in Patients Diagnosed With Depression Using Smartphone Data: Observational Study (e56874)
Arsi Ikäheimonen, Nguyen Luong, Ilya Baryshnikov, Richard Darst, Roope Heikkilä, Joel Holmen, Annasofia Martikkala, Kirsi Riihimäki, Outi Saleva, Erkki Isometsä, Talayeh Aledavood. 6090

Dynamic Bidirectional Associations Between Global Positioning System Mobility and Ecological Momentary Assessment of Mood Symptoms in Mood Disorders: Prospective Cohort Study (e55635)
Ting-Yi Lee, Ching-Hsuan Chen, I-Ming Chen, Hsi-Chung Chen, Chih-Min Liu, Shu-I Wu, Chuhsing Hsiao, Po-Hsiu Kuo. 6104

Analysis of Physical Activity Using Wearable Health Technology in US Adults Enrolled in the All of Us Research Program: Multiyear Observational Study (e65095)
Rujul Singh, Macy Tetrick, James Fisher, Peter Washington, Jane Yu, Electra Paskett, Frank Penedo, Steven Clinton, Roberto Benzo. 6118

XSL•FO
RenderX

User Experience and Extended Technology Acceptance Model in Commercial Health Care App Usage Among Patients With Cancer: Mixed Methods Study (e55176)
Ye-Eun Park, Yae Tak, Inhye Kim, Hui Lee, Jung Lee, Jong Lee, Yura Lee. 6136

Methodological Challenges in Randomized Controlled Trials of mHealth Interventions: Cross-Sectional Survey Study and Consensus-Based Recommendations (e53187)
Jesus Lopez-Alcalde, L Wieland, Yuqian Yan, Jürgen Barth, Mohammad Khami, Siddharudha Shivalli, Cynthia Lokker, Harleen Rai, Paul Macharia, Sergi Yun, Elvira Lang, Agnes Bwanika Naggirinya, Concepción Campos-Asensio, Leila Ahmadian, Claudia Witt. 6150

mHealth Engagement for Antiretroviral Medication Adherence Among People With HIV and Substance Use Disorders: Observational Study (e57774)
Ranran Mi, Ellie Yang, Alexander Tahk, Adati Tarfa, Lynne Cotter, Linqi Lu, Sijia Yang, David Gustafson Sr, Ryan Westergaard, Dhavan Shah.
6 1 7 3

Issues Related to the Use of Visual Social Networks and Perceived Usefulness of Social Media Literacy During the Recovery Phase: Qualitative Research Among Girls With Eating Disorders (e53334)
Elena Faccio, Margherita Reggiani, Michele Rocelli, Sabrina Cipolletta. 6188

Gamified Crowdsourcing as a Novel Approach to Lung Ultrasound Data Set Labeling: Prospective Analysis (e51397)
Nicole Duggan, Mike Jin, Maria Duran Mendicuti, Stephen Hallisey, Denie Bernier, Lauren Selame, Ameneh Asgari-Targhi, Chanel Fischetti, Ruben Lucassen, Anthony Samir, Erik Duhaime, Tina Kapur, Andrew Goldsmith. 6201

Using Twitter (X) to Mobilize Knowledge for First Contact Physiotherapists: Qualitative Study (e55680)
Laura Campbell, Jonathan Quicke, Kay Stevenson, Zoe Paskins, Krysia Dziedzic, Laura Swaithes. 6215

Twitter Discussions on #digitaldementia: Content and Sentiment Analysis (e59546)
Hyeongchan Cho, Kyu-Min Kim, Jee-Young Kim, Bo-Young Youn. 6230

Identifying Medicine Shortages With the Twitter Social Network: Retrospective Observational Study (e51317)
Doerine Postma, Magali Heijkoop, Peter De Smet, Kim Notenboom, Hubert Leufkens, Aukje Mantel-Teeuwisse. 6247

Peer Support for Chronic Pain in Online Health Communities: Quantitative Study on the Dynamics of Social Interactions in a Chronic Pain Forum (e45858)
Aaron Necaise, Mary Amon. 6260

Quality Measurement of Consumer Health Questions: Content and Language Perspectives (e48257)
Ashwag Alasmari, Lina Zhou. 6279

Alternative Health and Conventional Medicine Discourse About Cancer on TikTok: Computer Vision Analysis of TikTok Videos (e60283)
Roxana Muenster, Kai Gangi, Drew Margolin. 6292

Spillover Effects of Paid Functions on Physicians' Unpaid Knowledge Activities: Quasi-Experimental Approach (e58688)
Xuan Liu, Xiaotong Chi, Ming Chen, Wen Sun, Jia Li. 6308

Implementation of a Social Media Strategy for Public Health Promotion in Black, American Indian or Alaska Native, and Hispanic or Latino Communities During the COVID-19 Pandemic: Cross-Sectional Study (e58581)
Maria Mora Pinzon, Ornella Hills, George Levy, Taryn James, Ashley Benitez, Sacheen Lawrence, Tiffany Ellis, Venus Washington, Lizbeth Solorzano, Patricia Tellez-Giron, Fernando Cano Ospina, Melissa Metoxen, Carey Gleason. 6330

XSL•FO
RenderX

Characteristics and Popularity of Videos of Abusive Head Trauma Prevention: Systematic Appraisal (e60530)
Luc Goethals, Victoria Prokofieva Nelson, Fabien Fenouillet, Karine Chevreul, Manon Bergerat, Christine Lebreton, Yacine Refes, Flora Blangis, Martin Chalumeau, Enora Le Roux. 6345

Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences (e54321)
Julia Walsh, Jonathan Cave, Frances Griffiths. 6360

Exploring the Relationship Between Public Social Media Accounts, Adolescent Mental Health, and Parental Guidance in England: Large Cross-Sectional School Survey Study (e57154)
Wakithi Mabaso, Sascha Hein, Gabriela Pavarini, The OxWell Study Team, Mina Fazel. 6382

Influencers in Policy Fields on Social Media: Global Longitudinal Study of Dietary Sodium Reduction Posts, 2006-2022 (e54506)
Alana Montoya, Lingchao Mao, Adam Drewnowski, Joshua Chen, Ella Shi, Aileen Liang, Bryan Weiner, Yanfang Su. 6400

Information and Communication Technology for Managing Social Isolation and Loneliness Among People Living With Parkinson Disease: Qualitative Study of Barriers and Facilitators (e48175)
Gomathi Thangavel, Mevludin Memedi, Karin Hedström. 6415

Addressing Cyberscams and Acquired Brain Injury (“I Desperately Need to Know What to Do”): Qualitative Exploration of Clinicians’ and Service Providers’ Perspectives (e51245)
Kimberly Chew, Jennie Ponsford, Kate Gould. 6430

The EMPOWER Occupational e–Mental Health Intervention Implementation Checklist to Foster e–Mental Health Interventions in the Workplace: Development Study (e48504)
Alberto Raggi, Renaldo Bernard, Claudia Toppo, Carla Sabariego, Luis Salvador Carulla, Sue Lukersmith, Leona Hakkaart-van Roijen, Dorota Merecz-Kot, Beatriz Olaya, Rodrigo Antunes Lima, Desirée Gutiérrez-Marín, Ellen Vorstenbosch, Chiara Curatoli, Martina Cacciatore. 6446

The Impact of COVID-19 and Exposure to Violent Media Content on Cyber Violence Victimization Among Adolescents in South Korea: National Population-Based Study (e45563)
Eugene Lee, Peter Schulz, Hye Lee. 6459

Moderating Effect of Coping Strategies on the Association Between the Infodemic-Driven Overuse of Health Care Services and Cyberchondria and Anxiety: Partial Least Squares Structural Equation Modeling Study (e53417)
Richard Xu, Caiyun Chen. 6471

Empowering School Staff to Support Pupil Mental Health Through a Brief, Interactive Web-Based Training Program: Mixed Methods Study (e46764)
Emma Soneson, Emma Howarth, Alison Weir, Peter Jones, Mina Fazel. 6489

The Impact of Video-Based Microinterventions on Attitudes Toward Mental Health and Help Seeking in Youth: Web-Based Randomized Controlled Trial (e54478)
Diana Lemmer, Markus Moessner, Nicolas Arnaud, Harald Baumeister, Agnes Mutter, Sarah-Lena Klemm, Elisa König, Paul Plener, Christine Rummel-Kluge, Rainer Thomasius, Michael Kaess, Stephanie Bauer. 6511

Rapport Building in Written Crisis Services: Qualitative Content Analysis (e42049)
Laura Schwab-Reese, Caitlyn Short, Larel Jacobs, Michelle Fingerman. 6532

User Dynamics and Thematic Exploration in r/Depression During the COVID-19 Pandemic: Insights From Overlapping r/SuicideWatch Users (e53968)
Jianfeng Zhu, Ruoming Jin, Deric Kenne, NhatHai Phan, Wei-Shinn Ku. 6545

XSL•FO
RenderX

Long-Term Emotional Impact of the COVID-19 Pandemic and Barriers and Facilitators to Digital Mental Health Tools in Long-Term Care Workers: Qualitative Study (e47546)
Leticia González-Spinoglio, Anna Monistrol-Mula, Cecilia Vindrola-Padros, Salvatore Aguilar-Ortiz, Bernat Carreras, Josep Haro, Mireia Felez-Nobrega. 6560

Tracking and Profiling Repeated Users Over Time in Text-Based Counseling: Longitudinal Observational Study With Hierarchical Clustering (e50976)
Yucan Xu, Christian Chan, Evangeline Chan, Junyou Chen, Florence Cheung, Zhongzhi Xu, Joyce Liu, Paul Yip. 6571

A Transdiagnostic Video-Based Internet Intervention (Uni Virtual Clinic-Lite) to Improve the Mental Health of University Students: Randomized Controlled Trial (e53598)
Louise Farrer, Hayley Jackson, Amelia Gulliver, Alison Calear, Liana Leach, Penelope Hasking, Natasha Katruss, Philip Batterham. 6590

Impact of an Online Discussion Forum on Self-Guided Internet-Delivered Cognitive Behavioral Therapy for Public Safety Personnel: Randomized Trial (e59699)
Hugh McCall, Heather Hadjistavropoulos. 6609

Effectiveness of a Smartphone-Based Stress Management Program for Depression in Hospital Nurses During COVID-19 in Vietnam and Thailand: 2-Arm Parallel-Group Randomized Controlled Trial (e50071)
Kazuhiro Watanabe, Thuy Tran, Narisara Sripo, Asuka Sakuraya, Kotaro Imamura, Plernpit Boonyamalik, Natsu Sasaki, Thanate Tienthong, Hiroki Asaoka, Mako Iida, Quynh Nguyen, Nga Nguyen, Son Vu, Thuy Ngo, Tham Luyen, Long Nguyen, Nga Nguyen, Binh Nguyen, Yutaka Matsuyama, Yukie Takemura, Daisuke Nishi, Akizumi Tsutsumi, Huong Nguyen, Orawan Kaewboonchoo, Norito Kawakami. 6632

Clinical Effects of Asynchronous Provider-Guided Practice Sessions During Blended Care Therapy for Anxiety and Depression: Pragmatic Retrospective Cohort Study (e60502)
Hallie Espel-Huynh, Lu Wang, Emily Lattie, Robert Wickham, Alethea Varra, Connie Chen, Anita Lungu, Jennifer Lee. 6646

Assessing the Usage and Usability of a Mental Health Advice Telephone Service in Uganda: Mixed Methods Study (e65692)
Johnblack Kabukye, Juliet Nakku, Jackline Niwemuhwezi, James Nsereko, Rosemary Namagembe, Iris Groen, Ritah Neumbe, Denis Mubiru, Caroline Kisakye, Roseline Nanyonga, Marie Sjölinder, Susanne Nilsson, Caroline Wamala-Larsson. 6663

Patient Perspectives on Blended Internet-Based and Face-to-Face Cognitive Behavioral Therapy for Alcohol Use Disorder: Qualitative Study (e47083)
Kristine Tarp, Regina Christiansen, Randi Bilberg, Simone Borkner, Caroline Dalsgaard, Marie Paldam Folker, Anette Søgaard Nielsen. 6681

The Impact of Different Types of Social Media Use on the Mental Health of UK Adults: Longitudinal Observational Study (e56950)
Yue Yu, Jennifer Dykxhoorn, Ruth Plackett. 6696

Predictors of Engagement in Multiple Modalities of Digital Mental Health Treatments: Longitudinal Study (e48696)
Molly Nowels, Meghan McDarby, Lilla Brody, Evan Kleiman, Sara Sagui Henson, Cynthia Castro Sweet, Elissa Kozlov. 6708

Effectiveness of an Internet-Based Self-Help Acceptance and Commitment Therapy Program on Medical Students' Mental Well-Being: Follow-Up Randomized Controlled Trial (e50664)
Difan Wang, Bingyan Lin, Shuangxi Zhang, Wei Xu, Xinying Liu. 6723

Machine Learning–Based Suicide Risk Prediction Model for Suicidal Trajectory on Social Media Following Suicidal Mentions: Independent Algorithm Validation (e49927)
Zachary Kaminsky, Robyn McQuaid, Kim Hellemans, Zachary Patterson, Mysa Saad, Robert Gabrys, Tetyana Kendzerska, Alfonso Abizaid, Rebecca Robillard. 6740

XSL•FO
RenderX

An Automated Conversational Agent Self-Help Program: Randomized Controlled Trial (e53829)
Heather Foran, Christian Kubb, Janina Mueller, Spencer Poff, Megan Ung, Margaret Li, Eric Smith, Akinniyi Akinyemi, Melanie Kambadur, Franziska Waller, Mario Graf, Y-Lan Boureau. 6751

Effectiveness of an Internet-Based Acceptance and Commitment Therapy Intervention for Reducing Psychological Distress in Health Care Professionals: Randomized Controlled Trial (e59093)
Lin Zhang, Shuang Huang, Sha Liu, Yuanxiu Huang, Shan Chen, Jinsong Hu, Mingzhong Xu. 6765

A Digital Approach for Addressing Suicidal Ideation and Behaviors in Youth Mental Health Services: Observational Study (e60879)
Min Chong, Ian Hickie, Antonia Ottavio, David Rogers, Gina Dimitropoulos, Haley LaMonica, Luke Borgnolo, Sarah McKenna, Elizabeth Scott, Frank Iorfino. 6785

Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media (e47508)
Feipeng Guo, Zixiang Liu, Qibei Lu, Shaobo Ji, Chen Zhang. 6799

Perceptions of the Agency and Responsibility of the NHS COVID-19 App on Twitter: Critical Discourse Analysis (e50388)
Dan Heaton, Elena Nichele, Jérémie Clos, Joel Fischer. 6816

Analyzing Reddit Forums Specific to Abortion That Yield Diverse Dialogues Pertaining to Medical Information Seeking and Personal Worldviews: Data Mining and Natural Language Processing Comparative Study (e47408)
Danny Valdez, Lucrecia Mena-Meléndez, Brandon Crawford, Kristen Jozkowski. 6838

Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study (e49139)
Michael Deiner, Natalie Deiner, Vagelis Hristidis, Stephen McLeod, Thuy Doan, Thomas Lietman, Travis Porco. 6854

Impact of Traditional and New Media on Smoking Intentions and Behaviors: Secondary Analysis of Tasmania's Tobacco Control Mass Media Campaign Program, 2019-2021 (e47128)
James Kite, Anne Grunseit, Glenn Mitchell, Pip Cooper, Lilian Chan, Bo-Huei Huang, Margaret Thomas, Blythe O'Hara, Abby Smith. 6870

Agendas on Nursing in South Korea Media: Natural Language Processing and Network Analysis of News From 2005 to 2022 (e50518)
Daemin Park, Dasom Kim, Ah-hyun Park. 6881

Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study (e47826)
Annika Molenaar, Dickson Lukose, Linda Brennan, Eva Jenkins, Tracy McCaffrey. 6893

Bridging and Bonding Social Capital by Analyzing the Demographics, User Activities, and Social Network Dynamics of Sexual Assault Centers on Twitter: Mixed Methods Study (e50552)
Jia Xue, Qiaoru Zhang, Yun Zhang, Hong Shi, Chengda Zheng, Jingchuan Fan, Linxiao Zhang, Chen Chen, Luye Li, Micheal Shier. 6914

Medical Misinformation in Polish on the World Wide Web During the COVID-19 Pandemic Period: Infodemiology Study (e48130)
Małgorzata Chlabicz, Aleksandra Nabo ny, Jolanta Koszelew, Wojciech Łaguna, Anna Szpakowicz, Paweł Sowa, Wojciech Budny, Katarzyna Guziejko, Magdalena Róg-Makal, Sławomir Pancewicz, Maciej Kondrusik, Piotr Czupryna, Beata Cudowska, Dariusz Lebensztejn, Anna Moniuszko-Malinowska, Adam Wierzbicki, Karol Kami ski. 6937

The Influence of Joe Wicks on Physical Activity During the COVID-19 Pandemic: Thematic, Location, and Social Network Analysis of X Data (e49921)
Wasim Ahmed, Opeoluwa Aiyenitaju, Simon Chadwick, Mariann Hardey, Alex Fenton. 6950

XSL•FO
RenderX

Public Discourse, User Reactions, and Conspiracy Theories on the X Platform About HIV Vaccines: Data Mining and Content Analysis (e53375)
Jueman Zhang, Yi Wang, Magali Mouton, Jixuan Zhang, Molu Shi. 6965

Emerging Trends in Information-Seeking Behavior for Alpha-Gal Syndrome: Infodemiology Study Using Time Series and Content Analysis (e49928)
Jamie Romeiser, Nicole Jusko, Augusta Williams. 6985

News Coverage of the COVID-19 Pandemic on Social Media and the Public's Negative Emotions: Computational Study (e48491)
Hanjing Wang, Yupeng Li, Xuan Ning. 6996

Monitoring Adverse Drug Events in Web Forums: Evaluation of a Pipeline and Use Case Study (e46176)
Pierre Karapetiantz, Bissan Audeh, Akram Redjdal, Théophile Tiffet, Cédric Bousquet, Marie-Christine Jaulent. 7017

Nutrients and Foods Recommended for Blood Pressure Control on Twitter in Japan: Content Analysis (e49077)
Marina Terada, Tsuyoshi Okuhara, Rie Yokota, Takahiro Kiuchi, Kentaro Murakami. 7036

Assessing Message Deployment During Public Health Emergencies Through Social Media: Empirical Test of Optimizing Content for Effective Dissemination (e50871)
Paola Pascual-Ferrá, Neil Alperstein, Julia Burleson, Amelia Jamison, Ananya Bhaktaram, Sidharth Rath, Rohini Ganjoo, Satyanarayan Mohanty, Daniel Barnett, Rajiv Rimal. 7047

Cancer Prevention and Treatment on Chinese Social Media: Machine Learning–Based Content Analysis Study (e55937)
Keyang Zhao, Xiaojing Li, Jingyang Li. 7058

Investigation of Public Acceptance of Misinformation Correction in Social Media Based on Sentiment Attributions: Infodemiology Study Using Aspect-Based Sentiment Analysis (e50353)
Ning Ma, Guang Yu, Xin Jin. 7073

Digital Epidemiology of Prescription Drug References on X (Formerly Twitter): Neural Network Topic Modeling and Sentiment Analysis (e57885)
Varun Rao, Danny Valdez, Rasika Muralidharan, Jon Agley, Kate Eddens, Aravind Dendukuri, Vandana Panth, Maria Parker. 7096

Assessment of the Quality and Readability of Web-Based Arabic Health Information on Halitosis: Infodemiological Study (e54072)
Khalid Aboalshamat. 7115

The Normalization of Vaping on TikTok Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis: Mixed Methods Study (e55591)
Sungwon Jung, Dhiraj Murthy, Bara Bateineh, Alexandra Loukas, Anna Wilkinson. 7127

Uncovering the Complexity of Perinatal Polysubstance Use Disclosure Patterns on X: Mixed Methods Study (e53171)
Dezhi Wu, Hannah Shead, Yang Ren, Phyllis Raynor, Youyou Tao, Harvey Villanueva, Peiyin Hung, Xiaoming Li, Robert Brookshire, Kacey Eichelberger, Constance Guille, Alain Litwin, Bankole Olatosi. 7147

Racial Biases Associated With Pulse Oximetry: Longitudinal Social Network Analysis of Social Media Advocacy Impact (e56034)
Wasim Ahmed, Mariann Hardey, Bradford Winters, Aarti Sarwal. 7162

XSL•FO
RenderX

Selling Misleading “Cancer Cure” Books on Amazon: Systematic Search on Amazon.com and Thematic Analysis (e56354)
Marco Zenone, May van Schalkwyk, Greg Hartwell, Timothy Caulfield, Nason Maani. 7174

Exploring Public Emotions on Obesity During the COVID-19 Pandemic Using Sentiment Analysis and Topic Modeling: Cross-Sectional Study (e52142)
Jorge Correia, Sarmad Ahmad, Ahmed Waqas, Hafsa Meraj, Zoltan Pataky. 7184

Impact of Concurrent Media Exposure on Professional Identity: Cross-Sectional Study of 1087 Medical Students During Long COVID (e50057)
Manli Wu, Jun Yan, Chongming Qiao, Chu Yan. 7196

Association Between School-Related Google Trends Search Volume and Suicides Among Children and Adolescents in Japan During 2016-2020: Retrospective Observational Study With a Time-Series Analysis (e51710)
Takahiro Arai, Hiroe Tsubaki, Ayako Wakano, Yasuyuki Shimizu. 7213

Tracking the Spread of Pollen on Social Media Using Pollen-Related Messages From Twitter: Retrospective Analysis (e58309)
Martín Pérez-Pérez, María Fernandez Gonzalez, Francisco Rodriguez-Rajo, Florentino Fdez-Riverola. 7227

Discussion of Heated Tobacco Products on Twitter Following IQOS’s Modified-Risk Tobacco Product Authorization and US Import Ban: Content Analysis (e53938)
Minji Kim, Julia Vassey, Dongmei Li, Artur Galimov, Eileen Han, Matthew Kirkpatrick, Cassandra Stanton, Jenny Ozga, Sarah Lee, Jennifer Unger. 7258

Examining Quitting Experiences on Quit Vaping Subreddits From 2015 to 2021: Content Analysis (e52129)
Elexis Kierstead, Nathan Silver, Michael Amato. 7273

Online Interest in Elf Bar in the United States: Google Health Trends Analysis (e50343)
Akshaya Bhagavathula, Page Dobbs. 7285

Social Media Marketing Strategies for Electronic Cigarettes: Content Analysis of Chinese Weibo Accounts (e51594)
Xinyi Zhou, Xinyu Hao, Yuhang Chen, Hui Deng, Ling Fang, Lingyun Zhang, Xiaotao Yan, Pinpin Zheng, Fan Wang. 7294

Tobacco and Alcohol Content in Top Vietnamese YouTube Music Videos: Content Analysis (e55555)
Thi Tran, Thu Vu, Yachao Li, Lucy Popova. 7307

Predicting Prefecture-Level Well-Being Indicators in Japan Using Search Volumes in Internet Search Engines: Infodemiology Study (e64555)
Myung Yang, Kazuya Taira. 7321

Concurrent Mentions of Vaping and Alcohol on Twitter: Latent Dirichlet Analysis (e51870)
Lynsie Ranker, David Tofu, Manyuan Lu, Jiaxi Wu, Aruni Bhatnagar, Rose Robertson, Derry Wijaya, Traci Hong, Jessica Fetterman, Ziming Xuan. 7335

Wellness Influencer Responses to COVID-19 Vaccines on Social Media: A Longitudinal Observational Study (e56651)
Gabrielle O'Brien, Ronith Ganjigunta, Paramveer Dhillon. 7347

Long COVID Discourse in Canada, the United States, and Europe: Topic Modeling and Sentiment Analysis of Twitter Data (e59425)
Ahmed AbuRaed, Emil Prikryl, Giuseppe Carenini, Naveed Janjua. 7362

XSL•FO
RenderX

Toxicity on Social Media During the 2022 Mpox Public Health Emergency: Quantitative Study of Topical and Network Dynamics (e52997)
Lizhou Fan, Lingyao Li, Libby Hemphill. 7380

Discovering Time-Varying Public Interest for COVID-19 Case Prediction in South Korea Using Search Engine Queries: Infodemiology Study (e63476)
Seong-Ho Ahn, Kwangil Yim, Hyun-Sik Won, Kang-Min Kim, Dong-Hwa Jeong. 7404

Introducing and Evaluating the Effectiveness of Online Cognitive Behavior Therapy for Gambling Disorder in Routine Addiction Care: Comparative Cohort Study (e54754)
Olof Molander, Anne Berman, Miriam Jakobson, Mikael Gajecki, Hanna Hällström, Jonas Ramnerö, Johan Bjureberg, Per Carlbring, Philip Lindner. 7421

Optimizing Telehealth for Diabetes Management in the Deep South of the United States: Qualitative Study of Barriers and Facilitators on the Patient and Clinician Journey (e43583)
Alessandra Bazzano, Tejal Patel, Elizabeth Nauman, Dana Cernigliaro, Lizheng Shi. 7436

Key Factors Influencing the Operationalization and Effectiveness of Telemedicine Services in Henan Province, China: Cross-Sectional Analysis (e45020)
Xianying He, Fangfang Cui, Minzhao Lyu, Dongxu Sun, Xu Zhang, Jinming Shi, Yinglan Zhang, Shuai Jiang, Jie Zhao. 7451

Costs and Cardiovascular Benefits of a Fourth-Generation Synchronous Telehealth Program on Mortality and Cardiovascular Outcomes for Patients With Atrial Fibrillation: Retrospective Cohort Study (e48748)
Hao-Yun Chang, Hui-Wen Wu, Chi-Sheng Hung, Ying-Hsien Chen, Ching-Chang Huang, Li-Tan Yang, Shin-Tsyr Hwang, Jiun-Yu Yu, Jen-Kuang Lee, Yi-Lwun Ho. 7462

Reliability of Telepsychiatry Assessments Using the Attention-Deficit/Hyperactivity Disorder Rating Scale-IV for Children With Neurodevelopmental Disorders and Their Caregivers: Randomized Feasibility Study (e51749)
Shunya Kurokawa, Kensuke Nomura, Nana Hosogane, Takashi Nagasawa, Yuko Kawade, Yu Matsumoto, Shuichi Morinaga, Yuriko Kaise, Ayana Higuchi, Akiko Goto, Naoko Inada, Masaki Kodaira, Taishiro Kishimoto. 7475

Wait Times for Scheduling Appointments for Prevention of Macrovascular and Microvascular Complications of Diabetes: Cross-Sectional Descriptive Study (e55351)
Corey Basch, Grace Hillyer, Charles Basch. 7484

Evaluation of Telehealth Services that are Clinically Appropriate for Reimbursement in the US Medicaid Population: Mixed Methods Study (e46412)
Sanjeev Saravanakumar, Andrey Ostrovsky. 7491

Telehealth Care Through Internet Hospitals in China: Qualitative Interview Study of Physicians' Views on Access, Expectations, and Communication (e47523)
Yuqiong Zhong, Jessica Hahne, Xiaomin Wang, Xuxi Wang, Ying Wu, Xin Zhang, Xing Liu. 7500

Impacts of an Acute Care Telenursing Program on Discharge, Patient Experience, and Nursing Experience: Retrospective Cohort Comparison Study (e54330)
Courtenay Bruce, Steve Klahn, Lindsay Randle, Xin Li, Kelkar Sayali, Barbara Johnson, Melissa Gomez, Meagan Howard, Roberta Schwartz, Farzan Sasangohar. 7514

Longitudinal Monitoring of Clinician-Patient Video Visits During the Peak of the COVID-19 Pandemic: Adoption and Sustained Challenges in an Integrated Health Care Delivery System (e54008)
Jessica Palakshappa, Erica Hale, Joshua Brown, Carol Kittel, Emily Dressler, Gary Rosenthal, Sarah Cutrona, Kristie Foley, Emily Haines, Thomas Houston II. 7524

XSL•FO
RenderX

Patients' Representations of Perceived Distance and Proximity to Telehealth in France: Qualitative Study (e45702)
Amélie Loriot, Fabrice Larceneux, Valérie Guillard, Jean-Philippe Bertocchio. 7547

IT-Related Barriers and Facilitators to the Implementation of a New European eHealth Solution, the Digital Survivorship Passport (SurPass Version 2.0): Semistructured Digital Survey (e49910)
Ismay de Beijer, Selina van den Oever, Eliana Charalambous, Giorgio Cangioli, Julia Balaguer, Edit Bardi, Marie Alfes, Adela Cañete Nieto, Marisa Correcher, Tiago Pinto da Costa, Alexander Degelsegger-Márquez, Vanessa Düster, Anna-Liesa Filbert, Desiree Grabow, Gerald Gredinger, Hannah Gsell, Riccardo Haupt, Maria van Helvoirt, Ruth Ladenstein, Thorsten Langer, Anja Laschkolnig, Monica Muraca, Saskia Pluijm, Jelena Rascon, Günter Schreier, Zuzana Tomášikova, Florian Trauner, Justas Trink nas, Kathrin Trunner, Anne Uyttebroeck, Leontien Kremer, Helena van der Pal, Catherine Chronaki, PanCareSurPass Consortium. 7558

Prevalence and Disparities in Telehealth Use Among US Adults Following the COVID-19 Pandemic: National Cross-Sectional Survey (e52124)
Erin Spaulding, Michael Fang, Yvonne Commodore-Mensah, Cheryl Himmelfarb, Seth Martin, Josef Coresh. 7576

Travel Distance Between Participants in US Telemedicine Sessions With Estimates of Emissions Savings: Observational Study (e53437)
Mollie Cummins, Sukrut Shishupal, Bob Wong, Neng Wan, Jiuying Han, Jace Johnny, Amy Mhatre-Owens, Ramkiran Gouripeddi, Julia Ivanova, Triton Ong, Hiral Soni, Janelle Barrera, Hattie Wilczewski, Brandon Welch, Brian Bunnell. 7586

Public and Research Interest in Telemedicine From 2017 to 2022: Infodemiology Study of Google Trends Data and Bibliometric Analysis of Scientific Literature (e50088)
Andrea Maugeri, Martina Barchitta, Guido Basile, Antonella Agodi. 7598

Findings From a National Survey of Older US Adults on Patient Willingness to Use Telehealth Services: Cross-Sectional Survey (e50205)
Olufeyisayo Odebunmi, Tamera Hughes, Austin Waters, Benjamin Urick, Caroline Herron, Mary Wangen, Catherine Rohweder, Renée Ferrari, Macary Marciniak, Stephanie Wheeler, Alison Brenner, Parth Shah. 7611

Effect of a Structured Multilevel Telehealth Service on Hospital Admissions and Mortality During COVID-19 in a Resource-Limited Region in Brazil: Retrospective Cohort Study (e48464)
Clara Oliveira, Magda Pires, Karina Meira, Jordana de Jesus, Isabela Borges, Maria Paixão, Mayara Mendes, Leonardo Ribeiro, Milena Marcolino, Maria Alkmim, Antonio Ribeiro. 7623

Care Integration for Hepatitis C Virus Treatment Through Facilitated Telemedicine Within Opioid Treatment Programs: Qualitative Study (e53049)
Suzanne Dickerson, Saliyah George, Ana Ventuneac, Arpan Dharia, Andrew Talal. 7636

Antibiotic Prescribing by Digital Health Care Providers as Compared to Traditional Primary Health Care Providers: Cohort Study Using Register Data (e55228)
Andy Wallman, Kurt Svärdsudd, Kent Bobits, Thorne Wallman. 7650

Perception and Acceptance of Telemedicine Use in Health Care Among the General Public in China: Web-Based Cross-Sectional Survey (e53497)
Yulan Lin, Xiaonan Xu, Yiyang Liu, Haridah Alias, Zhijian Hu, Li Wong. 7681

Smartwatch-Detected Arrhythmias in Patients After Transcatheter Aortic Valve Replacement (TAVR): Analysis of the SMART TAVR Trial (e41843)
Jiaqi Fan, Hanyi Dai, Yuchao Guo, Jianguo Xu, Lihan Wang, Jubo Jiang, Xinping Lin, Cheng Li, Dao Zhou, Huajun Li, Xianbao Liu, Jian'an Wang. 7693

The Use of Online Consultation Systems and Patient Experience of Primary Care: Cross-Sectional Analysis Using the General Practice Patient Survey (e51272)
Xiaochen Ge, Paul Chappell, Jean Ledger, Minal Bakhai, Geraldine Clarke. 7703

XSL•FO
RenderX

Potential, Pitfalls, and Future Directions for Remote Monitoring of Chronic Respiratory Diseases: Multicenter Mixed Methods Study in Routine Cystic Fibrosis Care (e54942)
Martinus Oppelaar, Yvette Emond, Michiel Bannier, Monique Reijers, Hester van der Vaart, Renske van der Meer, Josje Altenburg, Lennart Conemans, Bart Rottier, Marianne Nuijsink, Lara van den Wijngaart, Peter Merkus, Maud Heinen, Jolt Roukema. 7725

Patient and Provider Experiences With Compassionate Care in Virtual Physiatry: Qualitative Study (e51878)
Marina Wasilewski, Abirami Vijayakumar, Zara Szigeti, Amanda Mayo, Laura Desveaux, James Shaw, Sander Hitzig, Robert Simpson. 7742

Identification of Motivational Determinants for Telemedicine Use Among Patients With Rheumatoid Arthritis in Germany: Secondary Analysis of Data From a Nationwide Cross-Sectional Survey Study (e47733)
Felix Muehlensiepen, Pascal Petit, Johannes Knitza, Martin Welcker, Nicolas Vuillerme. 7753

Health Care Usage During the COVID-19 Pandemic and the Adoption of Telemedicine: Retrospective Study of Chronic Disease Cohorts (e54991)
Margrét Bjarnadóttir, David Anderson, Kelley Anderson, Omar Aljwfi, Alina Peluso, Adam Ghannoum, Gayle Balba, Nawar Shara. 7768

Statistical Analysis of Telehealth Use and Pre- and Postpandemic Insurance Coverage in Selected Health Care Specialties in a Large Health Care System in Arkansas: Comparative Cross-Sectional Study (e49190)
Aysenur Cengil, Sandra Eksioglu, Burak Eksioglu, Hari Eswaran, Corey Hayes, Cari Bogulski. 7780

Patient-, Provider-, and Facility-Level Contributors to the Use of Cardiology Telehealth Care in the Veterans Health Administration: Retrospective Cohort Study (e53298)
Rebecca Tisdale, Jacqueline Ferguson, James Van Campen, Liberty Greene, Charlie Wray, Donna Zulman. 7803

Exploring the Importance of Race and Gender Concordance Between Patients and Physical Therapists in Digital Rehabilitation for Musculoskeletal Conditions: Observational, Longitudinal Study (e65354)
Anabela Areias, Dora Janela, Maria Molinos, Virgílio Bento, Carolina Moreira, Vijay Yanamadala, Steven Cohen, Fernando Correia, Fabíola Costa. 7812

Characterization of Telecare Conversations on Lifestyle Management and Their Relation to Health Care Utilization for Patients with Heart Failure: Mixed Methods Study (e46983)
Mojisola Erdt, Sakinah Yusof, Liquan Chai, Siti Md Salleh, Zhengyuan Liu, Halimah Sarim, Geok Lim, Hazel Lim, Nur Suhaimi, Lin Yulong, Yang Guo, Angela Ng, Sharon Ong, Bryan Choo, Sheldon Lee, Huang Weiliang, Hong Oh, Maria Wolters, Nancy Chen, Pavitra Krishnaswamy. . . 7 8 2 5

Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey (e64159)
Yuka Sugawara, Yosuke Hirakawa, Masao Iwagami, Haruo Kuroki, Shuhei Mitani, Ataru Inagaki, Hiroki Ohashi, Mitsuru Kubota, Soichi Koike, Rie Wakimizu, Masaomi Nangaku. 7841

Impact of Internet Hospital Consultations on Outpatient Visits and Expenses: Quasi-Experimental Study (e57609)
Yayuan Liu, Haofeng Jin, Zhuoyuan Yu, Yu Tong. 7857

Barriers and Facilitators Associated With Remote Concussion Physical Assessments From the Perspectives of Clinicians and People Living With Workplace Concussions: Focus Group Study (e56158)
Keely Barnes, Heidi Sveistrup, Motahareh Karimijashni, Mark Bayley, Mary Egan, Martin Bilodeau, Michel Rathbone, Monica Taljaard, Shawn Marshall. 7870

Comparing the Quality of Direct-to-Consumer Telemedicine Dominated and Delivered by Public and Private Sector Platforms in China: Standardized Patient Study (e55400)
Faying Song, Xue Gong, Yuting Yang, Rui Guo. 7894

XSL•FO
RenderX

Telemonitoring of Active Inflammatory Bowel Disease Using the App TECCU: Short-Term Results of a Multicenter Trial of GETECCU (e60966)
Mariam Aguas, Javier Del Hoyo, Raquel Vicente, Manuel Barreiro-de Acosta, Luigi Melcarne, Alejandro Hernandez-Camba, Lucía Madero, María Arroyo, Beatriz Sicilia, María Chaparro, María Martin-Arranz, Ramón Pajares, Francisco Mesonero, Miriam Mañosa, Pilar Martinez, Silvia Chacón, Joan Tosca, Sandra Marín, Luciano Sanroman, Marta Calvo, David Monfort, Empar Saiz, Yamile Zabana, Ivan Guerra, Pilar Varela, Virginia Baydal, Raquel Faubel, Pilar Corsino, Sol Porto-Silva, Eduard Brunet, Melodi González, Ana Gutiérrez, Pilar Nos. 7911

Telemedicine Integrated Care Versus In-Person Care Mode for Patients With Short Stature: Comprehensive Comparison of a Retrospective Cohort Study (e57814)
Yipei Wang, Pei Zhang, Yan Xing, Huifeng Shi, Yunpu Cui, Yuan Wei, Ke Zhang, Xinxia Wu, Hong Ji, Xuedong Xu, Yanhui Dong, Changxiao Jin. 7931

Quantitative Impact of Traditional Open Surgery and Minimally Invasive Surgery on Patients' First-Night Sleep Status in the Intensive Care Unit: Prospective Cohort Study (e56777)
Chen Shang, Ya Yang, Chengcheng He, Junqi Feng, Yan Li, Meimei Tian, Zhanqi Zhao, Yuan Gao, Zhe Li. 7947

Impact of Telehealth on Health Disparities Associated With Travel Time to Hospital for Patients With Recurrent Admissions: 4-Year Panel Data Analysis (e63661)
Youyou Tao, Ace Vo, Dezhi Wu, Junyuan Lin, Kala Seal, Abhay Mishra, Arindam Brahma. 7956

The Benefits of Using Active Remote Patient Management for Enhanced Heart Failure Outcomes in Rural Cardiology Practice: Single-Site Retrospective Cohort Study (e49710)
William Craig, Suzanne Ohlmann. 7971

Preferences of Individuals With Obesity for Online Medical Consultation in Different Demand Scenarios: Discrete Choice Experiments (e53140)
Yaolin Hu, Jian Wang, Jianbo Zhou, Yuanyuan Gu, Stephen Nicholas, Elizabeth Maitland. 7978

Opportunities to Address Specialty Care Deserts and the Digital Divide through the Veterans Health Administration's Telehealth Hub-and-Spoke Cardiology Clinic: Retrospective Cohort Study (e53932)
Rebecca Tisdale, Colin Purmal, Neil Kalwani, Alexander Sandhu, Paul Heidenreich, Donna Zulman, Tanvir Hussain. 8018

Perspectives of Clients and Health Care Professionals on the Opportunities for Digital Health Interventions in Cerebrovascular Disease Care: Qualitative Descriptive Study (e52715)
Henna Härkönen, Kirsi Myllykangas, Mikko Kärppä, Kirsi Rasmus, Julius Gomes, Milla Immonen, Piia Hyvämäki, Miia Jansson. 8030

Accuracy, Reproducibility, and Responsiveness to Treatment of Home Spirometry in Cystic Fibrosis: Multicenter, Retrospective, Observational Study (e60892)
Martinus Oppelaar, Hanneke van Helvoort, Michiel Bannier, Monique Reijers, Hester van der Vaart, Renske van der Meer, Josje Altenburg, Lennart Conemans, Bart Rottier, Marianne Nuijsink, Lara van den Wijngaart, Peter Merkus, Jolt Roukema. 8047

Postimplementation Evaluation in Assisted Living Facilities of an eHealth Medical Device Developed to Predict and Avoid Unplanned Hospitalizations: Pragmatic Trial (e55460)
Jacques-Henri Veyron, François Deparis, Marie Al Zayat, Joël Belmin, Charlotte Havreng-Théry. 8059

Comparing Contact Tracing Through Bluetooth and GPS Surveillance Data: Simulation-Driven Approach (e38170)
Weicheng Qian, Aranock Cooke, Kevin Stanley, Nathaniel Osgood. 8069

Investigating Health and Well-Being Challenges Faced by an Aging Workforce in the Construction and Nursing Industries: Computational Linguistic Analysis of Twitter Data (e49450)
Weicong Li, Liyaning Tang, Jed Montayre, Celia Harris, Sancia West, Mark Antoniou. 8088

Impact Analysis of COVID-19 Pandemic on Hospital Reviews on Dianping Website in Shanghai, China: Empirical Study (e52992)
Weixue Huo, Mengwei He, Zhaoxiang Zeng, Xianhao Bao, Ye Lu, Wen Tian, Jiaxuan Feng, Rui Feng. 8104

XSL•FO
RenderX

Determining an Appropriate Sample Size for Qualitative Interviews to Achieve True and Near Code Saturation: Secondary Analysis of Data (e52998)
Claudia Squire, Kristen Giombi, Douglas Rupert, Jacqueline Amoozegar, Peyton Williams. 8123

Contextual Acceptance of COVID-19 Mitigation Mobile Apps in the United States: Mixed Methods Survey Study on Postpandemic Data Privacy (e57309)
Yuanyuan Feng, Brad Stenger, Shikun Zhang. 8141

Assessing the Response Results of an mHealth-Based Patient Experience Survey Among People Receiving HIV Care in Lusaka, Zambia: Cohort Study (e54304)
Jacob Mutale, Kombatende Sikombe, Boroma Mwale, Mwansa Lumpa, Sandra Simbeza, Chama Bukankala, Njekwa Mukamba, Aaloke Mody, Laura Beres, Charles Holmes, Carolyn Bolton Moore, Elvin Geng, Izukanji Sikazwe, Jake Pry. 8160

Development and Validation of a Computed Tomography–Based Model for Noninvasive Prediction of the T Stage in Gastric Cancer: Multicenter Retrospective Study (e56851)
Jin Tao, Dan Liu, Fu-Bi Hu, Xiao Zhang, Hongkun Yin, Huiling Zhang, Kai Zhang, Zixing Huang, Kun Yang. 8171

Lessons From 3 Longitudinal Sensor-Based Human Behavior Assessment Field Studies and an Approach to Support Stakeholder Management: Content Analysis (e50461)
Johanna Kallio, Atte Kinnula, Satu-Marja Mäkelä, Sari Järvinen, Pauli Räsänen, Simo Hosio, Miguel Bordallo López. 8187

Comparing Health Survey Data Cost and Quality Between Amazon's Mechanical Turk and Ipsos' KnowledgePanel: Observational Study (e63032)
Patricia Herman, Mary Slaughter, Nabeel Qureshi, Tarek Azzam, David Cella, Ian Coulter, Graham DiGuiseppi, Maria Edelen, Arie Kapteyn, Anthony Rodriguez, Max Rubinstein, Ron Hays. 8206

Resting Heart Rate and Associations With Clinical Measures From the Project Baseline Health Study: Observational Study (e60493)
Kent Feng, Sarah Short, Sohrab Saeb, Megan Carroll, Christoph Olivier, Edgar Simard, Susan Swope, Donna Williams, Julie Eckstrand, Neha Pagidipati, Svati Shah, Adrian Hernandez, Kenneth Mahaffey. 8226

Willingness to Use Digital Health Screening and Tracking Tools for Public Health in Sexual Minority Populations in a National Probability Sample: Quantitative Intersectional Analysis (e47448)
Wilson Vincent. 8248

Online Health Information Seeking and Preventative Health Actions: Cross-Generational Online Survey Study (e48977)
Jayati Sinha, Nuket Serin. 8264

Evaluating the Digital Health Experience for Patients in Primary Care: Mixed Methods Study (e50410)
Melinda Choy, Kathleen O'Brien, Katelyn Barnes, Elizabeth Sturgiss, Elizabeth Rieger, Kirsty Douglas. 8285

Investigating Patient Use and Experience of Online Appointment Booking in Primary Care: Mixed Methods Study (e51931)
Helen Atherton, Abi Eccles, Leon Poltawski, Jeremy Dale, John Campbell, Gary Abel. 8304

Sociodemographic Factors Associated With Using eHealth for Information Seeking in the United States: Cross-Sectional Population-Based Study With 3 Time Points Using Health Information National Trends Survey Data (e54745)
Christian Vazquez, Rebecca Mauldin, Denise Mitchell, Faheem Ohri. 8323

XSL•FO
RenderX

Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study (e54186)
Dianna Vidas, Romina Carrasco, Ryan Kelly, Jenny Waycott, Jeanette Tamplin, Kate McMahon, Libby Flynn, Phoebe Stretton-Smith, Tanara Sousa, Felicity Baker. 8335

Mediating Effects of Information Access on Internet Use and Multidimensional Health Among Middle-Aged and Older Adults: Nationwide Cross-Sectional Study (e49688)
Liping Fu, Caiping Liu, Yongqing Dong, Xiaodong Ma, Quanling Cai, Dongli Li, Kaisheng Di. 8352

Perceptions and Satisfaction With the Use of Digital Medical Services in Urban Older Adults of China: Mixed Methods Study (e48654)
Ning Wang, Siyu Zhou, Zhuo Liu, Ying Han. 8367

Factors Associated With the Use of Digital Technology Among Youth in Zimbabwe: Findings From a Cross-Sectional Population-Based Survey (e52670)
Kevin Martin, Rachel Peh, Mandikudza Tembo, Constancia Mavodza, Aoife Doyle, Chido Dziva Chikwari, Ethel Dauya, Tsitsi Bandason, Steven Azizi, Victoria Simms, Rashida Ferrand. 8390

From Digital Inclusion to Digital Transformation in the Prevention of Drug-Related Deaths in Scotland: Qualitative Study (e52345)
Hadi Daneshvar, Hannah Carver, Graeme Strachan, Jessica Greenhalgh, Catriona Matheson. 8406

The Digitalization of Social Care in England and Implications for Older, Unpaid Carers: Constructionist Thematic Analysis (e60056)
Anastasia Rousaki, Efpraxia Zamani, Laura Sbaffi, Kate Hamblin, Rachael Black. 8423

Racial and Demographic Disparities in Susceptibility to Health Misinformation on Social Media: National Survey-Based Analysis (e55086)
Ranganathan Chandrasekaran, Muhammed Sadiq T, Evangelos Moustakas. 8435

Technology Acceptance Among Low-Income Asian American Older Adults: Cross-Sectional Survey Analysis (e52498)
Pauline DeLange Martinez, Daniel Tancredi, Misha Pavel, Lorena Garcia, Heather Young. 8448

Relationship Between Internet Use and Cognitive Function Among Middle-Aged and Older Chinese Adults: 5-Year Longitudinal Study (e57301)
Bowen Chen, Chun Yang, Shanshan Ren, Penggao Li, Jin Zhao. 8467

The Effects of the COVID-19 Pandemic on Age-Based Disparities in Digital Health Technology Use: Secondary Analysis of the 2017-2022 Health Information National Trends Survey (e65541)
Yuanbo Qiu, Huang Huang, Junjie Gai, Gianluca De Leo. 8483

Dietary Intake Assessment Using a Novel, Generic Meal–Based Recall and a 24-Hour Recall: Comparison Study (e48817)
Cathal O'Hara, Eileen Gibney. 8497

Validation of 3 Computer-Aided Facial Phenotyping Tools (DeepGestalt, GestaltMatcher, and D-Score): Comparative Diagnostic Accuracy Study (e42904)
Alisa Reiter, Jean Pantel, Magdalena Danyel, Denise Horn, Claus-Eric Ott, Martin Mensah. 8514

Mining Real-World Big Data to Characterize Adverse Drug Reaction Quantitatively: Mixed Methods Study (e48572)
Qi-Xuan Yue, Ruo-Fan Ding, Wei-Hao Chen, Lv-Ying Wu, Ke Liu, Zhi-Liang Ji. 8527

XSL•FO
RenderX

Comparing Human-Smartphone Interactions and Actigraphy Measurements for Circadian Rhythm Stability and Adiposity: Algorithm Development and Validation Study (e50149)
Hai-Hua Chuang, Chen Lin, Li-Ang Lee, Hsiang-Chih Chang, Guan-Jie She, Yu-Hsuan Lin. 8541

Multicentric Assessment of a Multimorbidity-Adjusted Disability Score to Stratify Depression-Related Risks Using Temporal Disease Maps: Instrument Validation Study (e53162)
Rubèn González-Colom, Kangkana Mitra, Emili Vela, Andras Gezsi, Teemu Paajanen, Zsófia Gál, Gabor Hullam, Hannu Mäkinen, Tamas Nagy, Mikko Kuokkanen, Jordi Piera-Jiménez, Josep Roca, Peter Antal, Gabriella Juhasz, Isaac Cano. 8557

The Potential of Evidence-Based Clinical Intake Tools to Discover or Ground Prevalence of Symptoms Using Real-Life Digital Health Encounters: Retrospective Cohort Study (e49570)
Eden Avnat, Michael Samin, Daniel Ben Joya, Eyal Schneider, Elia Yanko, Dafna Eshel, Shahar Ovadia, Yossi Lev, Daniel Souroujon. 8574

ChatGPT for Automated Qualitative Research: Content Analysis (e59050)
Rimke Bijker, Stephanie Merkouris, Nicki Dowling, Simone Rodda. 8584

Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews (e52758)
Kentaro Matsui, Tomohiro Utsumi, Yumi Aoki, Taku Maruki, Masahiro Takeshima, Yoshikazu Takaesu. 8602

The Dual Task Ball Balancing Test and Its Association With Cognitive Function: Algorithm Development and Validation (e49794)
Barry Greene, Sean Tobyne, Ali Jannati, Killian McManus, Joyce Gomes Osman, Russell Banks, Ranjit Kher, John Showalter, David Bates, Alvaro Pascual-Leone. 8617

The Validity of Impressions as a Media Dose Metric in a Tobacco Public Education Campaign Evaluation: Observational Study (e55311)
Kevin Davis, Laurel Curry, Brian Bradfield, David Stupplebeen, Rebecca Williams, Sandra Soria, Julie Lautsch. 8630

Identifying and Estimating Frailty Phenotypes by Vocal Biomarkers: Cross-Sectional Study (e58466)
Yu-Chun Lin, Huang-Ting Yan, Chih-Hsueh Lin, Hen-Hong Chang. 8639

Impact of Image Content on Medical Crowdfunding Success: A Machine Learning Approach (e58617)
Renwu Wang, Huimin Xu, Xupin Zhang. 8651

Evaluation of an App-Based Mobile Triage System for Mass Casualty Incidents: Within-Subjects Experimental Study (e65728)
Martin Schmollinger, Jessica Gerstner, Eric Stricker, Alexander Muench, Benjamin Breckwoldt, Manuel Sigle, Peter Rosenberger, Robert Wunderlich. 8666

Reliable Web-Based Auditory Cognitive Testing: Observational Study (e58444)
Meher Lad, John-Paul Taylor, Timothy Griffiths. 8683

Research Trends on Metabolic Syndrome in Digital Health Care Using Topic Modeling: Systematic Search of Abstracts (e53873)
Kiseong Lee, Yoongi Chung, Ji-Su Kim. 8695

Understanding Symptom Self-Monitoring Needs Among Postpartum Black Patients: Qualitative Interview Study (e47484)
Natalie Benda, Sydney Woode, Stephanie Niño de Rivera, Robin Kalish, Laura Riley, Alison Hermann, Ruth Masterson Creber, Eric Costa Pimentel, Jessica Ancker. 8726

XSL•FO
RenderX

Short-Term Effects of an eHealth Care Experiential Learning Program Among Patients With Type 2 Diabetes: Randomized Controlled Trial (e53509)
Yu-Shan Cheng, Cheng-Pei Lin, Lu-Yen Chen, Wei-Ren Hwang, Yi-Chun Lin, Yu-Chi Chen. 8743

Impacts of Clinical Decision Support Systems on the Relationship, Communication, and Shared Decision-Making Between Health Care Professionals and Patients: Multistakeholder Interview Study (e55717)
Florian Funer, Diana Schneider, Nils Heyen, Heike Aichinger, Andrea Klausen, Sara Tinnemeyer, Wenke Liedtke, Sabine Salloch, Tanja Bratan. 8758

"Doctor ChatGPT, Can You Help Me?" The Patient's Perspective: Cross-Sectional Study (e58831)
Jonas Armbruster, Florian Bussmann, Catharina Rothhaas, Nadine Titze, Paul Grützner, Holger Freischmidt. 8772

Exploring Client Preferences for Psychological Counselors in a Chinese Online Health Community: Longitudinal Study (e58428)
Xiandong Feng, Yinhuan Hu, Holger Pfaff, Sha Liu, Jinzhu Xie, Zemiao Zhang. 8785

The Quality and Cultural Safety of Online Osteoarthritis Information for Affected Persons and Health Care Professionals: Content Analysis (e57698)
Smita Dhakal, Shermeen Merani, Vandana Ahluwalia, Marisa Battistella, Cornelia Borkhoff, Glen Hazlewood, Aisha Lofters, Deborah Marshall, Crystal MacKay, Anna Gagliardi. 8800

Assessing the Role of the Generative Pretrained Transformer (GPT) in Alzheimer's Disease Management: Comparative Study of Neurologist- and Artificial Intelligence–Generated Responses (e51095)
Jiaqi Zeng, Xiaoyi Zou, Shirong Li, Yao Tang, Sisi Teng, Huanhuan Li, Changyu Wang, Yuxuan Wu, Luyao Zhang, Yunheng Zhong, Jialin Liu, Siru Liu. 8813

Heuristics Identified in Health Data–Sharing Preferences of Patients With Cancer: Qualitative Focus Group Study (e63155)
Anna Hermansen, Samantha Pollard, Kimberlyn McGrail, Nick Bansback, Dean Regier. 8828

The Environmental Impacts of Electronic Medical Records Versus Paper Records at a Large Eye Hospital in India: Life Cycle Assessment Study (e42140)
Cordelia Kwon, Lernik Essayei, Michael Spencer, Tom Etheridge, Rengaraj Venkatesh, Natrajan Vengadesan, Cassandra Thiel. 8840

Designing and Implementation of a Digitalized Intersectoral Discharge Management System and Its Effect on Readmissions: Mixed Methods Approach (e47133)
Christoph Strumann, Lisa Pfau, Laila Wahle, Raphael Schreiber, Jost Steinhäuser. 8851

Sociotechnical Cross-Country Analysis of Contextual Factors That Impact Patients' Access to Electronic Health Records in 4 European Countries: Framework Evaluation Study (e55752)
Jonas Moll, Isabella Scandurra, Annika Bärkås, Charlotte Blease, Maria Hägglund, Iiris Hörhammer, Bridget Kane, Eli Kristiansen, Peeter Ross, Rose-Mharie Åhlfeldt, Gunnar Klein. 8868

Building and Sustaining Public Trust in Health Data Sharing for Musculoskeletal Research: Semistructured Interview and Focus Group Study (e53024)
Zainab Yusuf, William Dixon, Charlotte Sharp, Louise Cook, Søren Holm, Caroline Sanders. 8884

Patient Health Record Protection Beyond the Health Insurance Portability and Accountability Act: Mixed Methods Study (e59674)
Hemang Subramanian, Arijit Sengupta, Yilin Xu. 8899

Patients' Use of Electronic Health Records Facilitates Patient-Centered Communication: Findings From the 2017 Health Information National Trends Survey (e50476)
Suhwoo Ahn, Chul-joo Lee, Inhwan Bae. 8937

XSL•FO
RenderX

Association Between Patient Portal Engagement and Weight Loss Outcomes in Patients After Bariatric Surgery: Longitudinal Observational Study Using Electronic Health Records (e56573)
Xinmeng Zhang, Kaidi Kang, Chao Yan, Yubo Feng, Simon Vandekar, Danxia Yu, S Rosenbloom, Jason Samuels, Gitanjali Srivastava, Brandon Williams, Vance Albaugh, Wayne English, Charles Flynn, You Chen. 8956

Insights and Trends in Open Note Access: Retrospective Observational Study (e55982)
Randeep Badwal, Paul Cavo, Mandip Panesar. 8969

Psychometric Evaluation of a Tablet-Based Tool to Detect Mild Cognitive Impairment in Older Adults: Mixed Methods Study (e56883)
Josephine McMurray, AnneMarie Levy, Wei Pang, Paul Holyoke. 8978

Digital Food Frequency Questionnaire Assessing Adherence to the Norwegian Food–Based Dietary Guidelines and Other National Lifestyle Recommendations: Instrument Validation Study (e53442)
Hege Henriksen, Markus Knudsen, Anette Hjartåker, Rune Blomhoff, Monica Carlsen. 8995

Cerebral Microbleeds in Different Brain Regions and Their Associations With the Digital Clock-Drawing Test: Secondary Analysis of the Framingham Heart Study (e45780)
Samia Akhter-Khan, Qiushan Tao, Ting Ang, Cody Karjadi, Indira Itchapurapu, David Libon, Michael Alosco, Jesse Mez, Wei Qiu, Rhoda Au. 9 0 3 6

A Multidimensional Approach for Evaluating Reality in Social Media: Mixed Methods Study (e52058)
HyunYi Cho, Wenbo Li, Rachel Lopez. 9046

True Mitotic Count Prediction in Gastrointestinal Stromal Tumors: Bayesian Network Model and PROMETheus (Preoperative Mitosis Estimator Tool) Application Development (e50023)
Salvatore Renne, Manuela Cammelli, Ilaria Santori, Marta Tassan-Mangina, Laura Samà, Laura Ruspi, Federico Sicoli, Piergiuseppe Colombo, Luigi Terracciano, Vittorio Quagliuolo, Ferdinando Cananzi. 9063

Web-Based Health Information Seeking by People Living With Multiple Sclerosis: Qualitative Investigation of the Multiple Sclerosis Online Course (e53372)
William Bevens, Rebekah Davenport, Sandra Neate, Maggie Yu, Pia Jelinek, George Jelinek, Jeanette Reece. 9079

Dental Students' Satisfaction With Web-Based Learning During the Initial Phase of the COVID-19 Pandemic: Mixed Methods Study (e50278)
Minjung Lee, So An, Jungjoon Ihm. 9092

Utility of Large Language Models for Health Care Professionals and Patients in Navigating Hematopoietic Stem Cell Transplantation: Comparison of the Performance of ChatGPT-3.5, ChatGPT-4, and Bard (e54758)
Elisabetta Xue, Dara Bracken-Clarke, Giovanni Iannantuono, Hyoyoung Choo-Wosoba, James Gulley, Charalampos Floudas. 9106

Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial (e57037)
Wenyi Gan, Jianfeng Ouyang, Hua Li, Zhaowen Xue, Yiming Zhang, Qiu Dong, Jiadong Huang, Xiaofei Zheng, Yiyi Zhang. 9112

A Blended Learning Course on the Diagnostics of Mental Disorders: Multicenter Cluster Randomized Noninferiority Trial (e54176)
Gabriel Bonnin, Svea Kröber, Silvia Schneider, Jürgen Margraf, Verena Pflug, Alexander Gerlach, Timo Slotta, Hanna Christiansen, Björn Albrecht, Mira-Lynn Chavanon, Gerrit Hirschfeld, Tina In-Albon, Meinald Thielsch, Ruth von Brachel. 9125

Effectiveness of Virtual Simulations Versus Mannequins and Real Persons in Medical and Nursing Education: Meta-Analysis and Trial Sequential Analysis of Randomized Controlled Trials (e56195)
Nan Jiang, Yuelun Zhang, Siyu Liang, Xiaohong Lyu, Shi Chen, Xiaoming Huang, Hui Pan. 9140

XSL•FO
RenderX

An e-Learning Intervention for Professionals to Promote Family-Centered Cancer Care When a Significant Caregiver for Children Is at End of Life: Mixed Methods Evaluation Study (e65619)
Cherith Semple, Carla O'Neill, Sarah Sheehan, Tanya McCance, Amanda Drury, Jeffrey Hanna. 9153

Effects of a Rice-Farming Simulation Video Game on Nature Relatedness, Nutritional Status, and Psychological State in Urban-Dwelling Adults During the COVID-19 Pandemic: Randomized Waitlist Controlled Trial (e51596)
Seulki Lee, Chisung Yuh, Yu-Bin Shin, Heon-Jeong Lee, Young-Mee Lee, Jungsil Lee, Chul-Hyun Cho. 9169

Efficacy of Digital Dance on Brain Imagery, Cognition, and Health: Randomized Controlled Trial (e57694)
Heng-Hsin Tung, Chen-Yuan Kuo, Pei-Lin Lee, Chih-Wen Chang, Kun-Hsien Chou, Ching-Po Lin, Liang-Kung Chen. 9185

The Impact of a Gamified Intervention on Daily Steps in Real-Life Conditions: Retrospective Analysis of 4800 Individuals (e47116)
Alexandre Mazéas, Cyril Forestier, Guillaume Harel, Martine Duclos, Aïna Chalabaev. 9202

COVID-19–Related Social Isolation, Self-Control, and Internet Gaming Disorder Among Chinese University Students: Cross-Sectional Survey (e52978)
Yufang Guo, Fangyan Yue, Xiangyu Lu, Fengye Sun, Meixing Pan, Yannan Jia. 9217

Effects of Virtual Reality Motor-Cognitive Training for Older People With Cognitive Frailty: Multicentered Randomized Controlled Trial (e57809)
Rick Kwan, Justina Liu, Olive Sin, Kenneth Fong, Jing Qin, Joe Wong, Claudia Lai. 9227

Game-Based Assessment of Peripheral Neuropathy Combining Sensor-Equipped Insoles, Video Games, and AI: Proof-of-Concept Study (e52323)
Antao Ming, Vera Clemens, Elisabeth Lorek, Janina Wall, Ahmad Alhajjar, Imke Galazky, Anne-Katrin Baum, Yang Li, Meng Li, Sebastian Stober, Nils Mertens, Peter Mertens. 9245

Interaction Effects Between Low Self-Control and Meaning in Life on Internet Gaming Disorder Symptoms and Functioning in Chinese Adolescents: Cross-Sectional Latent Moderated Structural Equation Modeling Study (e59490)
Ted Fong, Kunjie Cui, Paul Yip. 9263

Facilitating Thought Progression to Reduce Depressive Symptoms: Randomized Controlled Trial (e56201)
Shai-Lee Yatziv, Paola Pedrelli, Shira Baror, Sydney DeCaro, Noam Shachar, Bar Sofer, Sunday Hull, Joshua Curtiss, Moshe Bar. 9276

Open-World Games' Affordance of Cognitive Escapism, Relaxation, and Mental Well-Being Among Postgraduate Students: Mixed Methods Study (e63760)
Ailin Anto, Arunima Basu, Rania Selim, Thomas Foscht, Andreas Eisingerich. 9295

Factors That Influence Young Adults' Preferences for Virtual Reality Exergames in a Weight Control Setting: Qualitative Study (e58422)
Yanya Chen, Lina Guan, Weixin Wu, Liang Ye, Yan He, Xiaofen Zheng, Sicun Li, Bingsheng Guan, Wai-kit Ming. 9314

Identifying the Risk Factors of Allergic Rhinitis Based on Zhihu Comment Data Using a Topic-Enhanced Word-Embedding Model: Mixed Method Study and Cluster Analysis (e48324)
Dongxiao Gu, Qin Wang, Yidong Chai, Xuejie Yang, Wang Zhao, Min Li, Oleg Zolotarev, Zhengfei Xu, Gongrang Zhang. 9324

What Adolescents Say in Text Messages to Motivate Peer Networks to Access Health Care and Sexually Transmitted Infection Testing: Qualitative Thematic Analysis (e44861)
Marguerita Lightfoot, Chadwick Campbell, Allysha Maragh-Bass, Joi Jackson-Morgan, Kelly Taylor. 9340

Exploring the Types of Social Support Exchanged by Survivors of Pediatric Stroke and Their Families in an Online Peer Support Community: Qualitative Thematic Analysis (e49440)
William Wright, Charlotte Howdle, Neil Coulson, Anna De Simoni. 9347

XSL•FO
RenderX

Factors Influencing Recovery From Pediatric Stroke Based on Discussions From a UK-Based Online Stroke Community: Qualitative Thematic Study (e49409)
Charlotte Howdle, William Wright, Jonathan Mant, Anna De Simoni. 9359

Effect of Prosocial Behaviors on e-Consultations in a Web-Based Health Care Community: Panel Data Analysis (e52646)
Xiaoxiao Liu, Huijing Guo, Le Wang, Mingye Hu, Yichan Wei, Fei Liu, Xifu Wang. 9371

Diverse Mentoring Connections Across Institutional Boundaries in the Biomedical Sciences: Innovative Graph Database Analysis (e47560)
Toufeeq Syed, Erika Thompson, Zainab Latif, Jay Johnson, Damaris Javier, Katie Stinson, Gabrielle Saleh, Jamboor Vishwanatha. 9388

Association Between Prosuicide Website Searches Through Google and Suicide Death in the United States From 2010 to 2021: Lagged Time-Series Analysis (e53404)
Nora Kelsall, Catherine Gimbrone, Mark Olfson, Madelyn Gould, Jeffrey Shaman, Katherine Keyes. 9409

Exploring the Relationship Between Instagram Use and Self-Criticism, Self-Compassion, and Body Dissatisfaction in the Spanish Population: Observational Study (e51957)
Andrea Varaona, Miguel Alvarez-Mon, Irene Serrano-Garcia, Marina Díaz-Marsá, Jeffrey Looi, Rosa Molina-Ruiz. 9420

Using Topic Modeling to Understand Patients' and Caregivers' Perspectives About Left Ventricular Assist Device: Thematic Analysis (e50009)
Semyon Melnikov, Stav Klein, Moni Shahar, David Guy. 9434

How Does an Online Mental Health Community on Twitter Empower Diverse Population Levels and Groups? A Qualitative Analysis of #BipolarClub (e55965)
Horeya AbouWarda, Mateusz Dolata, Gerhard Schwabe. 9444

Examining the Popularity, Content, and Intersections With the Substance Abuse and Mental Health Services Administration's Definition of Recovery in a Nonclinical Online Cannabis Cessation Community: Infodemiology Study of Reddit Posts (e47357)
Elyse Thulin, Maureen Walton, Erin Bonar, Anne Fernandez. 9464

Understanding e-Cigarette Addictiveness: Triangulation of Focus Group and Netnographic Data (e57970)
Marike Andreas, Nadja Grundinger, Nadine Wolber, Daria Szafran, Tatiana Görig, Ute Mons, Valerie Lohner, Sabine Vollstädt-Klein, Sven Schneider. 9479

Online Depression Communities as a Complementary Approach to Improving the Attitudes of Patients With Depression Toward Medication Adherence: Cross-Sectional Survey Study (e56166)
Runnan Chen, Xiaorong Fu, Mochi Liu, Ke Liao, Lifei Bai. 9490

Communal Load Sharing of Miscarriage Experiences: Thematic Analysis of Social Media Community Support (e56680)
Julia Dubbelman, Jonelle Ooms, Laura Havgry, Lianne Simonse. 9513

Development and Validation of a Robust and Interpretable Early Triaging Support System for Patients Hospitalized With COVID-19: Predictive Algorithm Modeling and Interpretation Study (e52134)
Sangwon Baek, Yeon Jeong, Yun-Hyeon Kim, Jin Kim, Jin Kim, Eun Kim, Jae-Kwang Lim, Jungok Kim, Zero Kim, Kyunga Kim, Myung Chung.
9 5 2 6

Development and Validation of an Interpretable Conformal Predictor to Predict Sepsis Mortality Risk: Retrospective Cohort Study (e50369)
Meicheng Yang, Hui Chen, Wenhan Hu, Massimo Mischi, Caifeng Shan, Jianqing Li, Xi Long, Chengyu Liu. 9541

XSL•FO
RenderX

Making Science Computable Using Evidence-Based Medicine on Fast Healthcare Interoperability Resources: Standards Development Project (e54265)
Andrey Soares, Lisa Schilling, Joshua Richardson, Bhagvan Kommadi, Vignesh Subbian, Joanne Dehnbostel, Khalid Shahin, Karen Robinson, Muhammad Afzal, Harold Lehmann, Ilkka Kunnamo, Brian Alper. 9557

Collaborative Human–Computer Vision Operative Video Analysis Algorithm for Analyzing Surgical Fluency and Surgical Interruptions in Endonasal Endoscopic Pituitary Surgery: Cohort Study (e56127)
Chia-En Wong, Pei-Wen Chen, Heng-Jui Hsu, Shao-Yang Cheng, Chen-Che Fan, Yen-Chang Chen, Yi-Pei Chiu, Jung-Shun Lee, Sheng-Fu Liang. 9569

Electronic Health Record–Oriented Knowledge Graph System for Collaborative Clinical Decision Support Using Multicenter Fragmented Medical Data: Design and Application Study (e54263)
Yong Shang, Yu Tian, Kewei Lyu, Tianshu Zhou, Ping Zhang, Jianghua Chen, Jingsong Li. 9580

A Computed Tomography–Based Fracture Prediction Model With Images of Vertebral Bones and Muscles by Employing Deep Learning: Development and Validation Study (e48535)
Sung Kong, Wonwoo Cho, Sung Park, Jaegul Choo, Jung Kim, Sang Kim, Chan Shin. 9601

Diagnostic Accuracy of a Mobile AI-Based Symptom Checker and a Web-Based Self-Referral Tool in Rheumatology: Multicenter Randomized Controlled Trial (e55542)
Johannes Knitza, Koray Tascilar, Franziska Fuchs, Jacob Mohn, Sebastian Kuhn, Daniela Bohr, Felix Muehlensiepen, Christina Bergmann, Hannah Labinsky, Harriet Morf, Elizabeth Araujo, Matthias Englbrecht, Wolfgang Vorbrüggen, Cay-Benedict von der Decken, Stefan Kleinert, Andreas Ramming, Jörg Distler, Peter Bartz-Bazzanella, Nicolas Vuillerme, Georg Schett, Martin Welcker, Axel Hueber. 9614

Building Dual AI Models and Nomograms Using Noninvasive Parameters for Aiding Male Bladder Outlet Obstruction Diagnosis and Minimizing the Need for Invasive Video-Urodynamic Studies: Development and Validation Study (e58599)
Chung-You Tsai, Jing-Hui Tian, Chien-Cheng Lee, Hann-Chorng Kuo. 9625

Automated Behavioral Coding to Enhance the Effectiveness of Motivational Interviewing in a Chat-Based Suicide Prevention Helpline: Secondary Analysis of a Clinical Trial (e53562)
Mathijs Pellemans, Salim Salmi, Saskia Mérelle, Wilco Janssen, Rob van der Mei. 9643

Factors Influencing the Implementation of Digital Advance Care Planning: Qualitative Interview Study (e50217)
Andy Bradshaw, Jacqueline Birtwistle, Catherine Evans, Katherine Sleeman, Suzanne Richards, Robbie Foy, Pablo Millares Martin, Paul Carder, Matthew Allsop, Maureen Twiddy. 9657

AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation (e54616)
Xuan Zou, Weijie He, Yu Huang, Yi Ouyang, Zhen Zhang, Yu Wu, Yongsheng Wu, Lili Feng, Sheng Wu, Mengqi Yang, Xuyan Chen, Yefeng Zheng, Rui Jiang, Ting Chen. 9672

Creating a Modified Version of the Cambridge Multimorbidity Score to Predict Mortality in People Older Than 16 Years: Model Development and Validation (e56042)
Debasish Kar, Kathryn Taylor, Mark Joy, Sudhir Venkatesan, Wilhelmine Meeraus, Sylvia Taylor, Sneha Anand, Filipa Ferreira, Gavin Jamie, Xuejuan Fan, Simon de Lusignan. 9690

Combining Clinical-Radiomics Features With Machine Learning Methods for Building Models to Predict Postoperative Recurrence in Patients With Chronic Subdural Hematoma: Retrospective Cohort Study (e54944)
Cheng Fang, Xiao Ji, Yifeng Pan, Guanchao Xie, Hongsheng Zhang, Sai Li, Jinghai Wan. 9704

Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study (e62890)
Yun Kim, Won-Doo Seo, Sun Lee, Ja Koo, Gyung Kim, Hee Song, Minji Lee. 9722

XSL•FO
RenderX

Evaluating a Natural Language Processing–Driven, AI-Assisted International Classification of Diseases, 10th Revision, Clinical Modification, Coding System for Diagnosis Related Groups in a Real Hospital Environment: Algorithm Development and Validation Study (e58278)
Hong-Jie Dai, Chen-Kai Wang, Chien-Chang Chen, Chong-Sin Liou, An-Tai Lu, Chia-Hsin Lai, Bo-Tsz Shain, Cheng-Rong Ke, William Wang, Tatheer Mir, Mutiara Simanjuntak, Hao-Yun Kao, Ming-Ju Tsai, Vincent Tseng. 9744

Implementation of a Clinical Decision Support System for Antimicrobial Prescribing in Sub-Saharan Africa: Multisectoral Qualitative Study (e45122)
Nathan Peiffer-Smadja, Sophie Descousse, Elsa Courrèges, Audrey Nganbou, Pauline Jeanmougin, Gabriel Birgand, Séverin Lénaud, Anne-Lise Beaumont, Claire Durand, Tristan Delory, Josselin Le Bel, Elisabeth Bouvet, Sylvie Lariven, Eric D'Ortenzio, Issa Konaté, Marielle Bouyou-Akotet, Abdoul-Salam Ouedraogo, Gisèle Kouakou, Armel Poda, Corinne Akpovo, François-Xavier Lescure, Aristophane Tanon. 9765

Reinforcement Learning to Optimize Ventilator Settings for Patients on Invasive Mechanical Ventilation: Retrospective Study (e44494)
Siqi Liu, Qianyi Xu, Zhuoyang Xu, Zhuo Liu, Xingzhi Sun, Guotong Xie, Mengling Feng, Kay See. 9780

Effects of Implementing a Barcode Information Management System on Operating Room Staff: Comparative Study (e56192)
Chia-Yen Li, Mei-Hui Huang, Yu-Shiue Lin, Chi-Ming Chu, Hsueh-Hsing Pan. 9794

Establishing Medical Intelligence—Leveraging Fast Healthcare Interoperability Resources to Improve Clinical Management: Retrospective Cohort and Clinical Implementation Study (e55148)
Alexander Brehmer, Christopher Sauer, Jayson Salazar Rodríguez, Kelsey Herrmann, Moon Kim, Julius Keyl, Fin Bahnsen, Benedikt Frank, Martin Köhrmann, Tienush Rassaf, Amir-Abbas Mahabadi, Boris Hadaschik, Christopher Darr, Ken Herrmann, Susanne Tan, Jan Buer, Thorsten Brenner, Hans Reinhardt, Felix Nensa, Michael Gertz, Jan Egger, Jens Kleesiek. 9807

Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study (e65994)
Minseok Hong, Ri-Ra Kang, Jeong Yang, Sang Rhee, Hyunju Lee, Yong-gyom Kim, KangYoon Lee, HongGi Kim, Yu Lee, Tak Youn, Se Kim, Yong Ahn. 9819

Development and Validation of a Machine Learning–Based Early Warning Model for Lichenoid Vulvar Disease: Prediction Model Development Study (e55734)
Jian Meng, Xiaoyu Niu, Can Luo, Yueyue Chen, Qiao Li, Dongmei Wei. 9834

Identification of a Susceptible and High-Risk Population for Postoperative Systemic Inflammatory Response Syndrome in Older Adults: Machine Learning–Based Predictive Model (e57486)
Haiyan Mai, Yaxin Lu, Yu Fu, Tongsen Luo, Xiaoyue Li, Yihan Zhang, Zifeng Liu, Yuenong Zhang, Shaoli Zhou, Chaojin Chen. 9846

A Prediction Model to Identify Clinically Relevant Medication Discrepancies at the Emergency Department (MED-REC Predictor): Development and Validation Study (e55185)
Greet Van De Sijpe, Matthias Gijsen, Lorenz Van der Linden, Stephanie Strouven, Eline Simons, Emily Martens, Nele Persan, Veerle Grootaert, Veerle Foulon, Minne Casteels, Sandra Verelst, Peter Vanbrabant, Sabrina De Winter, Isabel Spriet. 9865

Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study (e60231)
Scott Silvey, Jinze Liu. 9883

Effect of a Feedback Visit and a Clinical Decision Support System Based on Antibiotic Prescription Audit in Primary Care: Multiarm Cluster-Randomized Controlled Trial (e60535)
Pauline Jeanmougin, Stéphanie Larramendy, Jean-Pascal Fournier, Aurélie Gaultier, Cédric Rat. 9898

XSL•FO
RenderX

A Machine Learning–Based Prediction Model for Acute Kidney Injury in Patients With Community-Acquired Pneumonia: Multicenter Validation Study (e51255)
Mengqing Ma, Caimei Chen, Dawei Chen, Hao Zhang, Xia Du, Qing Sun, Li Fan, Huiping Kong, Xueting Chen, Changchun Cao, Xin Wan. . . 9 9 0 9

Efficient Screening in Obstructive Sleep Apnea Using Sequential Machine Learning Models, Questionnaires, and Pulse Oximetry Signals: Mixed Methods Study (e51615)
Nai-Yu Kuo, Hsin-Jung Tsai, Shih-Jen Tsai, Albert Yang. 9926

Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey (e60944)
Xiaoli Tang, Xiaochen Yang, Jiajun Yuan, Jie Yang, Qian Jin, Hanting Zhang, Liebin Zhao, Weiwei Guo. 9939

Development and Validation of a Nomogram-Based Model to Predict Primary Hypertension Within the Next Year in Children and Adolescents: Retrospective Cohort Study (e58686)
Chenlong Qin, Li Peng, Yun Liu, Xiaoliang Zhang, Shumei Miao, Zhiyuan Wei, Wei Feng, Hongjian Zhang, Cheng Wan, Yun Yu, Shan Lu, Ruochen Huang, Xin Zhang. 9956

Leveraging Machine Learning to Identify Subgroups of Misclassified Patients in the Emergency Department: Multicenter Proof-of-Concept Study (e56382)
Sage Wyatt, Dagfinn Lunde Markussen, Mounir Haizoune, Anders Vestbø, Yeneabeba Sima, Maria Sandboe, Marcus Landschulze, Hauke Bartsch, Christopher Sauer. 9970

Multiple Perspectives on the Adoption of SMART Technologies for Improving Care of Older People: Mixed Methods Study (e45492)
Steriani Elavsky, Lenka Knapova, Kamil Janiš, Richard Cimler, Jitka Kuhnova, Tomas Cernicky. 9981

Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study (e51058)
Amy Thomas, Andrea Asnes, Kyle Libby, Allen Hsiao, Gunjan Tiyyagura. 9995

Differences in Physicians' Ratings of Work Stressors and Resources Associated With Digital Transformation: Cross-Sectional Study (e49581)
Magdalena Wekenborg, Katharina Förster, Florian Schweden, Robin Weidemann, Felix Bechtolsheim, Clemens Kirschbaum, Jürgen Weitz, Beate Ditzen. 10010

Patients' and Clinicians' Perceptions of the Clinical Utility of Predictive Risk Models for Chemotherapy-Related Symptom Management: Qualitative Exploration Using Focus Groups and Interviews (e49309)
Morven Miller, Lisa McCann, Liane Lewis, Christine Miaskowski, Emma Ream, Andrew Darley, Jenny Harris, Grigorios Kotronoulas, Geir V Berg, Simone Lubowitzki, Jo Armes, Elizabeth Patiraki, Eileen Furlong, Patricia Fox, Alexander Gaiger, Antonella Cardone, Dawn Orr, Adrian Flowerday, Stylianos Katsaragakis, Simon Skene, Margaret Moore, Paul McCrone, Nicosha De Souza, Peter Donnan, Roma Maguire. 10027

Facilitators of and Barriers to Teachers' Engagement With Consumer Technologies for Stress Management: Qualitative Study (e50457)
Julia Manning, Ann Blandford, Julian Edbrooke-Childs. 10044

Comparison of the Response to an Electronic Versus a Traditional Informed Consent Procedure in Terms of Clinical Patient Characteristics: Observational Study (e54867)
Anna Zondag, Marieke Hollestelle, Rieke van der Graaf, Hendrik Nathoe, Wouter van Solinge, Michiel Bots, Robin Vernooij, Saskia Haitjema, UCC-CVRM study group. 10064

Using Short-Form Videos to Get Clinical Trial Newcomers to Sign Up: Message-Testing Experiment (e49600)
Sisi Hu, Ciera Kirkpatrick, Namyeon Lee, Yoorim Hong, Sungkyoung Lee, Amanda Hinnant. 10075

XSL•FO
RenderX

Acceptance of Social Media Recruitment for Clinical Studies Among Patients With Hepatitis B: Mixed Methods Study (e54034)
Theresa Willem, Bettina Zimmermann, Nina Matthes, Michael Rost, Alena Buyx. 10087

Strategies for Identifying and Recruiting Women at High Risk for Breast Cancer for Research Outside of Clinical Settings: Observational Study (e54450)
Claire Conley, Jennifer Rodriguez, McKenzie McIntyre, Bethany Niell, Suzanne O'Neill, Susan Vadaparampil. 10100

Challenges and Approaches to Recruitment for and Retention in a Dyad-Focused eHealth Intervention During COVID-19: Randomized Controlled Trial (e51877)
Chunxuan Ma, Rachel Adler, Daria Neidre, Ronald Chen, Laurel Northouse, Christine Rini, Xianming Tan, Lixin Song. 10112

Improving Participant Recruitment in Clinical Trials: Comparative Analysis of Innovative Digital Platforms (e60504)
Alexia Bikou, Elena Deligianni, Foteini Dermiki-Gkana, Nikolaos Liappas, José Teriús-Padrón, Maria Beltrán Jaunsarás, Maria Cabrera-Umpiérrez, Christos Kontogiorgis. 10126

Examining the Type, Quality, and Content of Web-Based Information for People With Chronic Pain Interested in Spinal Cord Stimulation: Social Listening Study (e48599)
Maarten Moens, Leen Van Doorslaer, Maxime Billot, Edgard Eeckman, Manuel Roulaud, Philippe Rigoard, Maaike Fobelets, Lisa Goudman.
1 0 1 4 7

Quality of Patient-Centered eHealth Information on Erosive Tooth Wear: Systematic Search and Evaluation of Websites and YouTube Videos (e49514)
Lena Holland, Amelie Kanzow, Annette Wiegand, Philipp Kanzow. 10165

Retrospecting Digital Media Use, Negative Emotions, and Trust Gaps During the COVID-19 Pandemic in China: Cross-Sectional Web-Based Survey (e49422)
Lu Wei, Qing Huang. 10178

Quality Assessment of TikTok as a Source of Information About Mitral Valve Regurgitation in China: Cross-Sectional Study (e55403)
Nannan Cui, Yuting Lu, Yelin Cao, Xiaofan Chen, Shuiqiao Fu, Qun Su. 10191

Videos in Short-Video Sharing Platforms as Sources of Information on Colorectal Polyps: Cross-Sectional Content Analysis Study (e51655)
Jia-Lun Guan, Su-Hong Xia, Kai Zhao, Li-Na Feng, Ying-Ying Han, Ji-Yan Li, Jia-Zhi Liao, Pei-Yuan Li. 10206

Perceived Experiences and Needs of Digital Resources Among Postpartum Women in the United Arab Emirates: Qualitative Focus Group Study (e53720)
Nivine Hanach, Roba Saqan, Hadia Radwan, Wegdan Baniissa, Nanne de Vries. 10221

The Costs of Anonymization: Case Study Using Clinical Data (e49445)
Lisa Pilgram, Thierry Meurers, Bradley Malin, Elke Schaeffner, Kai-Uwe Eckardt, Fabian Prasser, GCKD Investigators. 10236

Finding Medical Photographs of Patients Online: Randomized, Cross-Sectional Study (e55352)
Zack Marshall, Maushumi Bhattacharjee, Meng Wang, Abdul Cadri, Hannah James, Shabnam Asghari, Rene Peltekian, Veronica Benz, Vanessa Finley-Roy, Brynna Childs, Lauren Asaad, Michelle Swab, Vivian Welch, Fern Brunger, Chris Kaposy. 10250

Ethical, Legal, and Practical Concerns Surrounding the Implemention of New Forms of Consent for Health Data Research: Qualitative Interview Study (e52180)
Svenja Wiertz, Joachim Boldt. 10260

XSL•FO
RenderX

Pediatric and Young Adult Household Transmission of the Initial Waves of SARS-CoV-2 in the United States: Administrative Claims Study (e44249)
Ming Chung, Brian Hart, Mauricio Santillana, Chirag Patel. 10273

Consumption of Health-Related Videos and Human Papillomavirus Awareness: Cross-Sectional Analyses of a US National Survey and YouTube From the Urban-Rural Context (e49749)
Ashvita Garg, Alan Nyitray, James Roberts, Nicholas Shungu, Kenneth Ruggiero, Jessica Chandler, Haluk Damgacioglu, Yenan Zhu, Naomi Brownstein, Katherine Sterba, Ashish Deshmukh, Kalyani Sonawane. 10285

Depression and Anxiety in Adolescents During the COVID-19 Pandemic in Relation to the Use of Digital Technologies: Longitudinal Cohort Study (e45114)
Chen Shen, Rachel Smith, Joel Heller, Alexander Spiers, Rhiannon Thompson, Helen Ward, Jonathan Roiser, Dasha Nicholls, Mireille Toledano. 10296

Machine Learning–Based Prediction of Suicidality in Adolescents With Allergic Rhinitis: Derivation and Validation in 2 Independent Nationwide Cohorts (e51473)
Hojae Lee, Joong Cho, Jaeyu Park, Hyeri Lee, Guillaume Fond, Laurent Boyer, Hyeon Kim, Seoyoung Park, Wonyoung Cho, Hayeon Lee, Jinseok Lee, Dong Yon. 10314

Status and Influencing Factors of Social Media Addiction in Chinese Workers: Cross-Sectional Survey Study (e48026)
Weitao Kong, Yuanyuan Li, Aijing Luo, Wenzhao Xie. 10328

Outcomes and Costs of the Transition From a Paper-Based Immunization System to a Digital Immunization System in Vietnam: Mixed Methods Study (e45070)
Thi Dang, Emily Carnahan, Linh Nguyen, Mercy Mvundura, Sang Dao, Thi Duong, Trung Nguyen, Doan Nguyen, Tu Nguyen, Laurie Werner, Tove Ryman, Nga Nguyen. 10343

Moving Biosurveillance Beyond Coded Data Using AI for Symptom Detection From Physician Notes: Retrospective Cohort Study (e53367)
Andrew McMurry, Amy Zipursky, Alon Geva, Karen Olson, James Jones, Vladimir Ignatov, Timothy Miller, Kenneth Mandl. 10360

Twitter Analysis of Health Care Workers' Sentiment and Discourse Regarding Post–COVID-19 Condition in Children and Young People: Mixed Methods Study (e50139)
Macarena Chepo, Sam Martin, Noémie Déom, Ahmad Khalid, Cecilia Vindrola-Padros. 10373

Examining the Role of Information Behavior in Linking Cancer Risk Perception and Cancer Worry to Cancer Fatalism in China: Cross-Sectional Survey Study (e49383)
Lianshan Zhang, Shaohai Jiang. 10388

COVID-19 Vaccine Effectiveness and Digital Pandemic Surveillance in Germany (eCOV Study): Web Application–Based Prospective Observational Cohort Study (e47070)
Anna-Lena Lang, Nils Hohmuth, Vukašin Viškovi , Stefan Konigorski, Stefan Scholz, Felix Balzer, Cornelius Remschmidt, Rasmus Leistner. .
1 0 4 0 4

Creation of Standardized Common Data Elements for Diagnostic Tests in Infectious Disease Studies: Semantic and Syntactic Mapping (e50049)
Caroline Stellmach, Sina Hopff, Thomas Jaenisch, Susana Nunes de Miranda, Eugenia Rinaldi, The NAPKON, LEOSS, ORCHESTRA, and ReCoDID Working Groups. 10422

Physiological Response to the COVID-19 Vaccine: Insights From a Prospective, Randomized, Single-Blinded, Crossover Trial (e51120)
Andjela Markovic, Vladimir Kovacevic, Timo Brakenhoff, Duco Veen, Paul Klaver, Marianna Mitratza, George Downward, Diederick Grobbee, Maureen Cronin, Brianna Goodale, COVID-19 Remote Early Detection (COVID-RED) consortium. 10435

XSL•FO
RenderX

Linking Opinions Shared on Social Media About COVID-19 Public Health Measures to Adherence: Repeated Cross-Sectional Surveys of Twitter Use in Canada (e51325)
José Denis-Robichaud, Erin Rees, Patrick Daley, Christina Zarowsky, Assane Diouf, Bouchra Nasri, Simon de Montigny, Hélène Carabin. . . 10404

From Tweets to Streets: Observational Study on the Association Between Twitter Sentiment and Anti-Asian Hate Crimes in New York City from 2019 to 2022 (e53050)
Hanxue Wei, Yulin Hswen, Junaid Merchant, Laura Drew, Quynh Nguyen, Xiaohe Yue, Heran Mane, Thu Nguyen. 10465

Quality, Usability, and Trust Challenges to Effective Data Use in the Deployment and Use of the Bangladesh Nutrition Information System Dashboard: Qualitative Study (e48294)
Berhaun Fesshaye, Shivani Pandya, Lena Kan, Anna Kalbarczyk, Kelsey Alland, SM Rahman, Md. Bulbul, Piyali Mustaphi, Muhammad Siddique, Md. Tanim, Mridul Chowdhury, Tajkia Rumman, Alain Labrique. 10475

Public Response to Federal Electronic Cigarette Regulations Analyzed Using Social Media Data Through Natural Language Processing: Topic Modeling Study (e58919)
Shuo-Yu Lin, Sahithi Tulabandu, J Koch, Rashelle Hayes, Andrew Barnes, Hemant Purohit, Songqing Chen, Bo Han, Hong Xue. 10487

Evaluation of Machine Learning to Detect Influenza Using Wearable Sensor Data and Patient-Reported Symptoms: Cohort Study (e47879)
Kamran Farooq, Melody Lim, Lawrence Dennison-Hall, Finn Janson, Aspen Olszewska, Muhammad Ahmad Zabidi, Anna Haratym-Rojek, Karol Narowski, Barry Clinch, Marco Prunotto, Devika Chawla, Victoria Hunter, Vincent Ukachukwu. 10499

Mobile Phone Syndromic Surveillance for Respiratory Conditions in an Emergency (COVID-19) Context in Colombia: Representative Survey Design (e50184)
Andres Vecino-Ortiz, Deivis Guzman-Tordecilla, Vidhi Maniar, Sandra Agudelo-Londoño, Oscar Franco-Suarez, Nathaly Aya Pastrana, Mariana Rodríguez-Patarroyo, Marino Mejía-Rocha, Jaime Cardona, Mariangela Chavez Chamorro, Dustin Gibson. 10513

Association Between Daily Internet Use and Intrinsic Capacity Among Middle-Aged and Older Adults in China: Large Prospective Cohort Study (e54200)
Xing-Ling Chen, Jin Li, Shu-Ning Sun, Qiang-Qiang Zhao, Sheng-Rong Lin, Ling-Jun Wang, Zhong-Qi Yang, Shi-Hao Ni, Lu Lu. 10529

Differences in Use of a Patient Portal Across Sociodemographic Groups: Observational Study of the NHS App in England (e56320)
Sukriti KC, Chrysanthi Papoutsi, Claire Reidy, Bernard Gudgin, John Powell, Azeem Majeed, Felix Greaves, Anthony Laverty. 10545

Performance of a Full-Coverage Cervical Cancer Screening Program Using on an Artificial Intelligence– and Cloud-Based Diagnostic System: Observational Study of an Ultralarge Population (e51477)
Lu Ji, Yifan Yao, Dandan Yu, Wen Chen, Shanshan Yin, Yun Fu, Shangfeng Tang, Lan Yao. 10557

Themes in Abortion Forum Discussions in a Restrictive Access Context: Qualitative and Quantitative Analyses (e59544)
Krzysztof Klimiuk, Amelia Kot, Ewa Majcherek, Katarzyna Kubiak, Łukasz Balwicki. 10569

Ecological Momentary Assessment of Mental Health Problems Among University Students: Data Quality Evaluation Study (e55712)
Ana Portillo-Van Diest, Philippe Mortier, Laura Ballester, Franco Amigo, Paula Carrasco, Raquel Falcó, Margalida Gili, Glenn Kiekens, Francisco H Machancoses, Jose Piqueras, Marisa Rebagliato, Miquel Roca, Tíscar Rodríguez-Jiménez, Jordi Alonso, Gemma Vilagut. 10584

An Extensible Evaluation Framework Applied to Clinical Text Deidentification Natural Language Processing Tools: Multisystem and Multicorpus Study (e55676)
Paul Heider, Stéphane Meystre. 10605

XSL•FO
RenderX

Augmenting K-Means Clustering With Qualitative Data to Discover the Engagement Patterns of Older Adults With Multimorbidity When Using Digital Health Technologies: Proof-of-Concept Trial (e46287)
Yiyang Sheng, Raymond Bond, Rajesh Jaiswal, John Dinsmore, Julie Doyle. 10618

Service Quality and Residents' Preferences for Facilitated Self-Service Fundus Disease Screening: Cross-Sectional Study (e45545)
Senlin Lin, Yingyan Ma, Yanwei Jiang, Wenwen Li, Yajun Peng, Tao Yu, Yi Xu, Jianfeng Zhu, Lina Lu, Haidong Zou. 10641

Exploring the Experiences of Community-Dwelling Older Adults on Using Wearable Monitoring Devices With Regular Support From Community Health Workers, Nurses, and Social Workers: Qualitative Descriptive Study (e49403)
Arkers Wong, Jonathan Bayuo, Jing Su, Karen Chow, Siu Wong, Bonnie Wong, Athena Lee, Frances Wong. 10676

Value of Engagement in Digital Health Technology Research: Evidence Across 6 Unique Cohort Studies (e57827)
Sarah Goodday, Emma Karlin, Alexa Brooks, Carol Chapman, Christiana Harry, Nelly Lugo, Shannon Peabody, Shazia Rangwala, Ella Swanson, Jonell Tempero, Robin Yang, Daniel Karlin, Ron Rabinowicz, David Malkin, Simon Travis, Alissa Walsh, Robert Hirten, Bruce Sands, Chetan Bettegowda, Matthias Holdhoff, Jessica Wollett, Kelly Szajna, Kallan Dirmeyer, Anna Dodd, Shawn Hutchinson, Stephanie Ramotar, Robert Grant, Adrien Boch, Mackenzie Wildman, Stephen Friend. 10688

Predicting Long-Term Engagement in mHealth Apps: Comparative Study of Engagement Indices (e59444)
Yae Tak, Jong Lee, Junetae Kim, Yura Lee. 10702

Evaluating the Impact of a Game (Inner Dragon) on User Engagement Within a Leading Smartphone App for Smoking Cessation: Randomized Controlled Trial (e57839)
Justin White, Séverine Toussaert, Bethany Raiff, Marie Salem, Amy Chiang, David Crane, Edward Warrender, Courtney Lyles, Lorien Abroms, J Westmaas, Johannes Thrul. 10712

Effectiveness of the Support From Community Health Workers and Health Care Professionals on the Sustained Use of Wearable Monitoring Devices Among Community-Dwelling Older Adults: Feasibility Randomized Controlled Trial (e52435)
Arkers Wong, Jonathan Bayuo, Jing Su, Frances Wong, Karen Chow, Bonnie Wong, Siu Wong, Vivian Hui. 10732

Exploring Discussions About Virtual Reality on Twitter to Inform Brain Injury Rehabilitation: Content and Network Analysis (e45168)
Sophie Brassel, Melissa Brunner, Andrew Campbell, Emma Power, Leanne Togher. 10755

Evaluating the Impact of Virtual Reality on the Behavioral and Psychological Symptoms of Dementia and Quality of Life of Inpatients With Dementia in Acute Care: Randomized Controlled Trial (VRCT) (e51758)
Lora Appel, Eva Appel, Erika Kisonas, Samantha Lewis-Fung, Susanna Pardini, Jarred Rosenberg, Julian Appel, Christopher Smith. 10774

Assessing the Clinical Efficacy of a Virtual Reality Tool for the Treatment of Obesity: Randomized Controlled Trial (e51558)
Dimitra Anastasiadou, Pol Herrero, Paula Garcia-Royo, Julia Vázquez-De Sebastián, Mel Slater, Bernhard Spanlang, Elena Álvarez de la Campa, Andreea Ciudin, Marta Comas, Josep Ramos-Quiroga, Pilar Lusilla-Palacios. 10798

Integrating Biomarkers From Virtual Reality and Magnetic Resonance Imaging for the Early Detection of Mild Cognitive Impairment Using a Multimodal Learning Approach: Validation Study (e54538)
Bogyeom Park, Yuwon Kim, Jinseok Park, Hojin Choi, Seong-Eun Kim, Hokyoung Ryu, Kyoungwon Seo. 10816

Exploring Relations Between Unique Patient Characteristics and Virtual Reality Immersion Level on Anxiety and Pain in Patients Undergoing Venipuncture: Secondary Analysis of a Randomized Control Trial (e53196)
Jeffrey Gold, Krystal Akbar, Sandra Avila, Nhat Ngo, Margaret Klein. 10834

The Influence of Billboard-Based Tobacco Prevention Posters on Memorization, Attitudes, and Craving: Immersive Virtual Reality Study (e49344)
Solenne Bonneterre, Oulmann Zerhouni, Marilisa Boffo. 10845

XSL•FO
RenderX

Factors Associated With Continuous Use of a Cancer Education Metaverse Platform: Mixed Methods Study (e57762)
Sunghak Kim, Timothy Jung, Dae Sohn, Mina Suh, Yoon Chang. 10859

Assessing the Clinical Effectiveness of an Exergame-Based Exercise Training Program Using Ring Fit Adventure to Prevent and Postpone Frailty and Sarcopenia Among Older Adults in Rural Long-Term Care Facilities: Randomized Controlled Trial (e59468)
Sheng-Hui Tuan, Lin-Hui Chang, Shu-Fen Sun, Chien-Hui Li, Guan-Bo Chen, Yi-Ju Tsai. 10875

Nurses' Knowledge and Skills After Use of an Augmented Reality App for Advanced Cardiac Life Support Training: Randomized Controlled Trial (e57327)
Wan-Na Sun, Min-Chai Hsieh, Wei-Fang Wang. 10932

Examining the Use of Virtual Reality to Support Mindfulness Skills Practice in Mood and Anxiety Disorders: Mixed Methods Study (e45640)
Rebecca Blackmore, Claudia Giles, Hailey Tremain, Ryan Kelly, Fiona Foley, Kathryn Fletcher, Maja Nedeljkovic, Greg Wadley, Elizabeth Seabrook, Neil Thomas. 10944

Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis (e47562)
Nuša Fari , Henry Potts, James Heilman. 10959

Added Value of Medical Subject Headings Terms in Search Strategies of Systematic Reviews: Comparative Study (e53781)
Victor Leblanc, Aghiles Hamroun, Raphaël Bentegeac, Bastien Le Guellec, Rémi Lenain, Emmanuel Chazard. 10973

Generation Z's Health Information Avoidance Behavior: Insights From Focus Group Discussions (e54107)
Chenjin Jia, Pengcheng Li. 10988

Patients' Experiences With Digitalization in the Health Care System: Qualitative Interview Study (e47278)
Christian Gybel Jensen, Frederik Gybel Jensen, Mia Loft. 11000

Serbian Version of the eHealth Literacy Questionnaire (eHLQ): Translation, Cultural Adaptation, and Validation Study Among Primary Health Care Users (e57963)
Branko Vujkovic, Voin Brkovic, Ana Paji i , Vedrana Pavlovic, Dejana Stanisavljevic, Dušanka Krajnovi , Aleksandra Jovic Vranes. 11012

Spanish and Catalan Versions of the eHealth Literacy Questionnaire: Translation, Cross-Cultural Adaptation, and Validation Study (e49227)
Eulàlia Hernández Encuentra, Noemí Robles, Ariadna Angulo-Brunet, David Cullen, Ignacio del Arco. 11026

Electronic Health Literacy Scale-Web3.0 for Older Adults with Noncommunicable Diseases: Validation Study (e52457)
Wenfei Cai, Wei Liang, Huaxuan Liu, Rundong Zhou, Jie Zhang, Lin Zhou, Ning Su, Hanxiao Zhu, Yide Yang. 11046

eHealth Literacy and Web-Based Health Information–Seeking Behaviors on COVID-19 in Japan: Internet-Based Mixed Methods Study (e57842)
Seigo Mitsutake, Koichiro Oka, Orkan Okan, Kevin Dadaczynski, Tatsuro Ishizaki, Takeo Nakayama, Yoshimitsu Takahashi. 11059

Age Differences in Electronic Mental Health Literacy: Qualitative Study (e59131)
Richard Xu, Lidan Tian, Liling Zhu, Yuan Cao, Sherry Chan, Dong Dong, Wai-ling Cheung, Eliza Wong. 11086

XSL•FO
RenderX

Associations Between Sociodemographic Characteristics, eHealth Literacy, and Health-Promoting Lifestyle Among University Students in Taipei: Cross-Sectional Validation Study of the Chinese Version of the eHealth Literacy Scale (e52314)
Dan-Ping Chao. 11097

Developing an Educational Resource Aimed at Improving Adolescent Digital Health Literacy: Using Co-Design as Research Methodology (e49453)
Callum Lewis, Melody Taba, Tiffany Allen, Patrina Caldwell, S Skinner, Melissa Kang, Hamish Henderson, Liam Bray, Madeleine Borthwick, Philippa Collin, Kirsten McCaffery, Karen Scott. 11114

Development and Validation of a Mobile-Centered Digital Health Readiness Scale (mDiHERS): Health Literacy and Equity Scale (e58497)
Hana Kim, Rebecca Schnall, Nagyeom Yoon, Seong-Joon Koh, Jisan Lee, Jae Cheon. 11129

Evaluating Expert-Layperson Agreement in Identifying Jargon Terms in Electronic Health Record Notes: Observational Study (e49704)
John Lalor, David Levy, Harmon Jordan, Wen Hu, Jenni Smirnova, Hong Yu. 11142

Exploring Physicians' Perceptions of Digital Health's Impact on the Patient-Physician Relationship in the Primary Health Care Setting: Qualitative Descriptive Study (e53705)
Kai Sze, Qi Fong, Jacqueline De Roza, Eng Lee, Shu Tan. 11155

Online Health Information Seeking, eHealth Literacy, and Health Behaviors Among Chinese Internet Users: Cross-Sectional Survey Study (e54135)
Diyi Liu, Shuhang Yang, Calvin Cheng, Lin Cai, Jing Su. 11170

How Do Scholars Conceptualize and Conduct Health and Digital Health Literacy Research? Survey of Federally Funded Scholars (e57040)
Mayank Sakhuja, Brooks Yelton, Simone Kavarana, Lauren Schaurer, Jancham Rumthao, Samuel Noblet, Michelle Arent, Mark Macauda, Lorie Donelle, Daniela Friedman. 11188

Relationships Among eHealth Literacy, Physical Literacy, and Physical Activity in Chinese University Students: Cross-Sectional Study (e56386)
Shan Jiang, Johan Ng, Siu Choi, Amy Ha. 11201

Factors Associated With Digital Capacity for Health Promotion Among Primary Care Workers: Cross-Sectional Survey Study (e63054)
Yining Wang, Hui Ren, Shaotan Xiao, Tian Meng, Shuyue Sun, Siyu Yu, Qing Liu, Fan Wang. 11217

Integration of Patient-Reported Outcome Data Collected Via Web Applications and Mobile Apps Into a Nation-Wide COVID-19 Research Platform Using Fast Healthcare Interoperability Resources: Development Study (e47846)
Johannes Oehm, Sarah Riepenhausen, Michael Storck, Martin Dugas, Rüdiger Pryss, Julian Varghese. 11231

Development and External Validation of Machine Learning Models for Diabetic Microvascular Complications: Cross-Sectional Study With Metabolites (e41065)
Feng He, Clarissa Ng Yin Ling, Simon Nusinovici, Ching-Yu Cheng, Tien Wong, Jialiang Li, Charumathi Sabanayagam. 11247

The Effects of Displaying the Time Targets of the Manchester Triage System to Emergency Department Personnel: Prospective Crossover Study (e45593)
Jonas Bienzeisler, Guido Becker, Bernadett Erdmann, Alexander Kombeiz, Raphael Majeed, Rainer Röhrig, Felix Greiner, Ronny Otto, Fabian Otto-Sobotka, AKTIN Research Group. 11261

Experiences of Electronic Health Records' and Client Information Systems' Use on a Mobile Device and Factors Associated With Work Time Savings Among Practical Nurses: Cross-Sectional Study (e46954)
Satu Paatela, Maiju Kyytsönen, Kaija Saranto, Ulla-Mari Kinnunen, Tuulikki Vehko. 11275

XSL•FO
RenderX

The Paradoxes of Digital Tools in Hospitals: Qualitative Interview Study (e56095)
Marie Wosny, Livia Strasser, Janna Hastings. 11288

Five-Feature Models to Predict Preeclampsia Onset Time From Electronic Health Record Data: Development and Validation Study (e48997)
Hailey Ballard, Xiaotong Yang, Aditya Mahadevan, Dominick Lemas, Lana Garmire. 11307

Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study (e46936)
Isabel Straw, Geraint Rees, Parashkev Nachev. 11319

Evaluation of a Natural Language Processing Approach to Identify Diagnostic Errors and Analysis of Safety Learning System Case Review Data: Retrospective Cohort Study (e50935)
Azade Tabaie, Alberta Tran, Tony Calabria, Sonita Bennett, Arianna Milicia, William Weintraub, William Gallagher, John Yosaitis, Laura Schubel, Mary Hill, Kelly Smith, Kristen Miller. 11337

Caregiver and Youth Characteristics That Influence Trust in Digital Health Platforms in Pediatric Care: Mixed Methods Study (e53657)
Eric Chow, Alice Virani, Susan Pinkney, Fatema Abdulhussein, Tibor van Rooij, Matthias Görges, Wyeth Wasserman, Jeffrey Bone, Holly Longstaff, Shazhan Amed. 11350

Influence of Blood Sampling Service Process Reengineering on Medical Services Supply: Quasi-Experimental Study (e51412)
Wenmin Liao, Rong He, Zhonglian He, Nan Shi, Dan Li, Aihua Zhuang, Feng Gan, Ying Sun, Chaofeng Li. 11387

Hospital Length of Stay Prediction for Planned Admissions Using Observational Medical Outcomes Partnership Common Data Model: Retrospective Study (e59260)
Haeun Lee, Seok Kim, Hui-Woun Moon, Ho-Young Lee, Kwangsoo Kim, Se Jung, Sooyoung Yoo. 11397

Antihypertensive Drug Recommendations for Reducing Arterial Stiffness in Patients With Hypertension: Machine Learning–Based Multicohort (RIGIPREV) Study (e54357)
Iván Cavero-Redondo, Arturo Martinez-Rodrigo, Alicia Saz-Lara, Nerea Moreno-Herraiz, Veronica Casado-Vicente, Leticia Gomez-Sanchez, Luis Garcia-Ortiz, Manuel Gomez-Marcos, RIGIPREV Investigators Group. 11412

A Computable Phenotype Algorithm for Postvaccination Myocarditis/Pericarditis Detection Using Real-World Data: Validation Study (e54597)
Matthew Deady, Raymond Duncan, Matthew Sonesen, Renier Estiandan, Kelly Stimpert, Sylvia Cho, Jeffrey Beers, Brian Goodness, Lance Jones, Richard Forshee, Steven Anderson, Hussein Ezzeldin. 11433

Longitudinal Model Shifts of Machine Learning–Based Clinical Risk Prediction Models: Evaluation Study of Multiple Use Cases Across Different Hospitals (e51409)
Patricia Cabanillas Silva, Hong Sun, Mohamed Rezk, Diana Roccaro-Waldmeyer, Janis Fliegenschmidt, Nikolai Hulde, Vera von Dossow, Laurent Meesseman, Kristof Depraetere, Joerg Stieg, Ralph Szymanowsky, Fried-Michael Dahlweid. 11445

Effect of Negative Online Reviews and Physician Responses on Health Consumers' Choice: Experimental Study (e46713)
Xi Han, Yongxi Lin, Wenting Han, Ke Liao, Kefu Mei. 11459

Youth is Prized in Medicine, Old Age is Valued in Law: Analysis of Media Narratives Over 200 Years (e45855)
Reuben Ng, Nicole Indran. 11477

XSL•FO
RenderX

Examining the Role of Physician Characteristics in Web-Based Verified Primary Care Physician Reviews: Observational Study (e51672)
Neil Sehgal, Benjamin Rader, John Brownstein. 11488

Classification of Patients' Judgments of Their Physicians in Web-Based Written Reviews Using Natural Language Processing: Algorithm Development and Validation (e50236)
Farrah Madanay, Karissa Tu, Ada Campagna, J Davis, Steven Doerstling, Felicia Chen, Peter Ubel. 11500

An Electronic Health Record–Based Automated Self-Rescheduling Tool to Improve Patient Access: Retrospective Cohort Study (e52071)
Smitha Ganeshan, Andrew Liu, Anne Kroeger, Prerna Anand, Richard Seefeldt, Alexis Regner, Diana Vaughn, Anobel Odisho, Michelle Mourad. 11512

The Impact of Broadband Infrastructure Construction on Medical Resource Mismatch: Quasi-Natural Experiment From the Broadband China Policy (e53921)
Yulin Chai, Xiaoping Yuan, Lin Guo, Zhongming Chen. 11523

Information and Communications Technology–Based Monitoring Service for Tailored Chronic Disease Management in Primary Care: Cost-Effectiveness Analysis Based on ICT-CM Trial Results (e51239)
Sung-Hee Oh, Jae-Heon Kang, Jin-Won Kwon. 11560

Evaluating the Economic Impact of the PedAMINES App in Reducing Medication Errors in Pediatric Emergency Care: Cost-Effectiveness Analysis (e52077)
Loïc Brunner, Johan Siebert, Frédéric Ehrler, Sergio Manzano, Joachim Marti. 11578

A Typology of Social Media Use by Human Service Nonprofits: Mixed Methods Study (e51698)
Jia Xue, Micheal Shier, Junxiang Chen, Yirun Wang, Chengda Zheng, Chen Chen. 11591

User-Centered Framework for Implementation of Technology (UFIT): Development of an Integrated Framework for Designing Clinical Decision Support Tools Packaged With Tailored Implementation Strategies (e51952)
Jessica Ray, Emily Finn, Hollyce Tyrrell, Carlin Aloe, Eliana Perrin, Charles Wood, Dean Miner, Randall Grout, Jeremy Michel, Laura Damschroder, Mona Sharifi. 11608

Developing and Evaluating Digital Public Health Interventions Using the Digital Public Health Framework DigiPHrame: A Framework Development Study (e54269)
Tina Jahnel, Chen-Chia Pan, Núria Pedros Barnils, Saskia Muellmann, Merle Freye, Hans-Henrik Dassow, Oliver Lange, Anke Reinschluessel, Wolf Rogowski, Ansgar Gerhardus. 11620

Toward Guidelines for Designing Holistic Integrated Information Visualizations for Time-Critical Contexts: Systematic Review (e58088)
Ahmed Patel, Weston Baxter, Talya Porat. 11632

Impact of Reverse Empowerment and Proactive Motivations on Physicians' Online Knowledge Sharing in Digital Platforms: Survey Study (e59904)
Jingyuan Su, Kathy Shen, Xitong Guo. 11656

Application of Patient-Reported Outcome Measurements in Adult Tumor Clinical Trials in China: Cross-Sectional Study (e45719)
Yan Jia, Qi Li, Xiaowen Zhang, Yi Yan, Shiyan Yan, Shunping Li, Wei Li, Xiaowen Wu, Hongguo Rong, Jianping Liu. 11674

Implementation of Patient-Reported Outcomes in a Medical Oncology Setting (the iPROMOS Study): Type II Hybrid Implementation Study (e55841)
Natasha Roberts, Anita Pelecanos, Kimberly Alexander, David Wyld, Monika Janda. 11695

XSL•FO
RenderX

A Collection of Components to Design Clinical Dashboards Incorporating Patient-Reported Outcome Measures: Qualitative Study (e55267)
Anja Bischof, David Kuklinski, Irene Salvi, Carla Walker, Justus Vogel, Alexander Geissler. 11704

Enhancing Clinicians' Use of Electronic Patient-Reported Outcome Measures in Outpatient Care: Mixed Methods Study (e60306)
Veerle van Engen, Igna Bonfrer, Kees Ahaus, Monique Den Hollander-Ardon, Ingrid Peters, Martina Buljac-Samardzic. 11715

The Association Between Linguistic Characteristics of Physicians' Communication and Their Economic Returns: Mixed Method Study (e42850)
Shuang Geng, Yuqin He, Liezhen Duan, Chen Yang, Xusheng Wu, Gemin Liang, Ben Niu. 11737

Individual Characteristics in the Comprehension of Pandemic Video Communication: Randomized Controlled Between-Subjects Design (e48882)
Daniel Lungu, Jo Røislien, Ionica Smeets, Siri Wiig, Kolbjørn Brønnick. 11753

Discussions of Cannabis Over Patient Portal Secure Messaging: Content Analysis (e63311)
Vishal Shetty, Christina Gregor, Lorraine Tusing, Apoorva Pradhan, Katrina Romagnoli, Brian Piper, Eric Wright. 11765

Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study (e52113)
Anne Herrmann-Werner, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, Moritz Mahling. 11778

Identifying Functional Status Impairment in People Living With Dementia Through Natural Language Processing of Clinical Documents: Cross-Sectional Study (e47739)
John Laurentiev, Dae Kim, Mufaddal Mahesri, Kuan-Yuan Wang, Lily Bessette, Cassandra York, Heidi Zakoul, Su Lee, Li Zhou, Kueiyu Lin. . . 11791

Adverse Event Signal Detection Using Patients' Concerns in Pharmaceutical Care Records: Evaluation of Deep Learning Models (e55794)
Satoshi Nishioka, Satoshi Watabe, Yuki Yanagisawa, Kyoko Sayama, Hayato Kizaki, Shungo Imai, Mitsuhiro Someya, Ryoo Taniguchi, Shuntaro Yada, Eiji Aramaki, Satoko Hori. 11803

Machine Learning for Predicting Risk and Prognosis of Acute Kidney Disease in Critically Ill Elderly Patients During Hospitalization: Internet-Based and Interpretable Model Study (e51354)
Mingxia Li, Shuzhe Han, Fang Liang, Chenghuan Hu, Buyao Zhang, Qinlan Hou, Shuangping Zhao. 11820

Improving the Prognostic Evaluation Precision of Hospital Outcomes for Heart Failure Using Admission Notes and Clinical Tabular Data: Multimodal Deep Learning Model (e54363)
Zhenyue Gao, Xiaoli Liu, Yu Kang, Pan Hu, Xiu Zhang, Wei Yan, Muyang Yan, Pengming Yu, Qing Zhang, Wendong Xiao, Zhengbo Zhang. . . 11840

Development and Validation of an Explainable Deep Learning Model to Predict In-Hospital Mortality for Patients With Acute Myocardial Infarction: Algorithm Development and Validation Study (e49848)
Puguang Xie, Hao Wang, Jun Xiao, Fan Xu, Jingyang Liu, Zihang Chen, Weijie Zhao, Siyu Hou, Dongdong Wu, Yu Ma, Jingjing Xiao. 11851

Potential of Large Language Models in Health Care: Delphi Study (e52399)
Kerstin Denecke, Richard May, LLMHealthGroup, Octavio Rivera Romero. 11865

Machine Learning–Based Prediction of Suicidal Thinking in Adolescents by Derivation and Validation in 3 Independent Worldwide Cohorts: Algorithm Development and Validation Study (e55913)
Hyejun Kim, Yejun Son, Hojae Lee, Jiseung Kang, Ahmed Hammoodi, Yujin Choi, Hyeon Kim, Hayeon Lee, Guillaume Fond, Laurent Boyer, Rosie Kwon, Selin Woo, Dong Yon. 11886

XSL•FO
RenderX

Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis (e53164)
Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, Caroline Ruetsch-Chelli. 11899

The Effect of Artificial Intelligence on Patient-Physician Trust: Cross-Sectional Vignette Study (e50853)
Anna Zondag, Raoul Rozestraten, Stephan Grimmelikhuijsen, Karin Jongsma, Wouter van Solinge, Michiel Bots, Robin Vernooij, Saskia Haitjema. 11941

Evaluation of the Clinical Efficacy and Trust in AI-Assisted Embryo Ranking: Survey-Based Prospective Study (e52637)
Hyung Kim, Hyoeun Kang, Chaeyoon Lee, Jong Park, Mi Chung, Miran Kim, Na Kim, Hye Lee. 11953

Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4 (e54571)
Adi Lahat, Kassem Sharif, Narmin Zoabi, Yonatan Shneor Patt, Yousra Sharif, Lior Fisher, Uria Shani, Mohamad Arow, Roni Levin, Eyal Klang. 1 1 9 6 4

Artificial Intelligence–Based Electrocardiographic Biomarker for Outcome Prediction in Patients With Acute Heart Failure: Prospective Cohort Study (e52139)
Youngjin Cho, Minjae Yoon, Joonghee Kim, Ji Lee, Il-Young Oh, Chan Lee, Seok-Min Kang, Dong-Ju Choi. 11979

Evaluating and Enhancing Large Language Models' Performance in Domain-Specific Medicine: Development and Usability Study With DocOA (e58158)
Xi Chen, Li Wang, MingKe You, WeiZhi Liu, Yu Fu, Jie Xu, Shaoting Zhang, Gang Chen, Kang Li, Jian Li. 12008

Human-AI Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation (e50130)
Nadine Bienefeld, Emanuela Keller, Gudela Grote. 12021

Early Detection of Pulmonary Embolism in a General Patient Population Immediately Upon Hospital Admission Using Machine Learning to Identify New, Unidentified Risk Factors: Model Development Study (e48595)
Ori Ben Yehuda, Edward Itelman, Adva Vaisman, Gad Segal, Boaz Lerner. 12035

A Machine Learning Model for Predicting In-Hospital Mortality in Chinese Patients With ST-Segment Elevation Myocardial Infarction: Findings From the China Myocardial Infarction Registry (e50067)
Jingang Yang, Yingxue Li, Xiang Li, Shuiying Tao, Yuan Zhang, Tiange Chen, Guotong Xie, Haiyan Xu, Xiaojin Gao, Yuejin Yang. 12053

A 3D and Explainable Artificial Intelligence Model for Evaluation of Chronic Otitis Media Based on Temporal Bone Computed Tomography: Model Development, Validation, and Clinical Application (e51706)
Binjun Chen, Yike Li, Yu Sun, Haojie Sun, Yanmei Wang, Jihan Lyu, Jiajie Guo, Shunxing Bao, Yushu Cheng, Xun Niu, Lian Yang, Jianghong Xu, Juanmei Yang, Yibo Huang, Fanglu Chi, Bo Liang, Dongdong Ren. 12065

Differences in Fear and Negativity Levels Between Formal and Informal Health-Related Websites: Analysis of Sentiments and Emotions (e55151)
Abigail Paradise Vit, Avi Magid. 12086

Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study (e52506)
Yueye Wang, Xiaotong Han, Cong Li, Lixia Luo, Qiuxia Yin, Jian Zhang, Guankai Peng, Danli Shi, Mingguang He. 12100

Using Domain Adaptation and Inductive Transfer Learning to Improve Patient Outcome Prediction in the Intensive Care Unit: Retrospective Observational Study (e52730)
Maruthi Mutnuri, Henry Stelfox, Nils Forkert, Joon Lee. 12112

XSL•FO
RenderX

Predictors of Medical and Dental Clinic Closure by Machine Learning Methods: Cross-Sectional Study Using Empirical Data (e46608)
Young-Taek Park, Donghan Kim, Ji Jeon, Kwang Kim. 12131

Enhancing Patient Selection in Sepsis Clinical Trials Design Through an AI Enrichment Strategy: Algorithm Development and Validation (e54621)
Meicheng Yang, Jinqiang Zhuang, Wenhan Hu, Jianqing Li, Yu Wang, Zhongheng Zhang, Chengyu Liu, Hui Chen. 12143

An Advanced Machine Learning Model for a Web-Based Artificial Intelligence–Based Clinical Decision Support System Application: Model Development and Validation Study (e56022)
Tai-Han Lin, Hsing-Yi Chung, Ming-Jr Jian, Chih-Kai Chang, Cherng-Lih Perng, Guo-Shiou Liao, Jyh-Cherng Yu, Ming-Shen Dai, Cheng-Ping Yu, Hung-Sheng Shang. 12161

Using AI to Differentiate Mpox From Common Skin Lesions in a Sexual Health Clinic: Algorithm Development and Validation Study (e52490)
Nyi Soe, Zhen Yu, Phyu Latt, David Lee, Ranjit Samra, Zongyuan Ge, Rashidur Rahman, Jiajun Sun, Jason Ong, Christopher Fairley, Lei Zhang. 12173

Trial Factors Associated With Completion of Clinical Trials Evaluating AI: Retrospective Case-Control Study (e58578)
David Chen, Christian Cao, Robert Kloosterman, Rod Parsa, Srinivas Raman. 12185

Using Natural Language Processing (GPT-4) for Computed Tomography Image Analysis of Cerebral Hemorrhages in Radiology: Retrospective Analysis (e58741)
Daiwen Zhang, Zixuan Ma, Ru Gong, Liangliang Lian, Yanzhuo Li, Zhenghui He, Yuhan Han, Jiyuan Hui, Jialin Huang, Jiyao Jiang, Weiji Weng, Junfeng Feng. 12191

Enhancing Performance of the National Field Triage Guidelines Using Machine Learning: Development of a Prehospital Triage Model to Predict Severe Trauma (e58740)
Qi Chen, Yuchen Qin, Zhichao Jin, Xinxin Zhao, Jia He, Cheng Wu, Bihan Tang. 12203

Machine Learning–Based Prediction of Neurodegenerative Disease in Patients With Type 2 Diabetes by Derivation and Validation in 2 Independent Korean Cohorts: Model Development and Validation Study (e56922)
Hyunji Sang, Hojae Lee, Jaeyu Park, Sunyoung Kim, Ho Woo, Ai Koyanagi, Lee Smith, Sihoon Lee, You-Cheol Hwang, Tae Park, Hyunjung Lim, Dong Yon, Sang Rhee. 12218

Evolution of the “Internet Plus Health Care” Mode Enabled by Artificial Intelligence: Development and Application of an Outpatient Triage System (e51711)
Lingrui Yang, Jiali Pang, Song Zuo, Jian Xu, Wei Jin, Feng Zuo, Kui Xue, Zhongzhou Xiao, Xinwei Peng, Jie Xu, Xiaofan Zhang, Ruiyao Chen, Shuqing Luo, Shaoting Zhang, Xin Sun. 12233

Automated Speech Analysis for Risk Detection of Depression, Anxiety, Insomnia, and Fatigue: Algorithm Development and Validation Study (e58572)
Rachid Riad, Martin Denais, Marc de Gennes, Adrien Lesage, Vincent Oustric, Xuan Cao, Stéphane Mouchabac, Alexis Bourla. 12244

Use of Artificial Intelligence in Cobb Angle Measurement for Scoliosis: Retrospective Reliability and Accuracy Study of a Mobile App (e50631)
Haodong Li, Chuang Qian, Weili Yan, Dong Fu, Yiming Zheng, Zhiqiang Zhang, Junrong Meng, Dahui Wang. 12262

Machine Learning–Based Prediction for Incident Hypertension Based on Regular Health Checkup Data: Derivation and Validation in 2 Independent Nationwide Cohorts in South Korea and Japan (e52794)
Seung Hwang, Hayeon Lee, Jun Lee, Myeongcheol Lee, Ai Koyanagi, Lee Smith, Sang Rhee, Dong Yon, Jinseok Lee. 12276

XSL•FO
RenderX

Development and Validation of Deep Learning–Based Infectivity Prediction in Pulmonary Tuberculosis Through Chest Radiography: Retrospective Study (e58413)
Wou Chung, Jinsik Yoon, Dukyong Yoon, Songsoo Kim, Yujeong Kim, Ji Park, Young Kang. 12290

Pioneering Klebsiella Pneumoniae Antibiotic Resistance Prediction With Artificial Intelligence-Clinical Decision Support System–Enhanced Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: Retrospective Study (e58039)
Ming-Jr Jian, Tai-Han Lin, Hsing-Yi Chung, Chih-Kai Chang, Cherng-Lih Perng, Feng-Yee Chang, Hung-Sheng Shang. 12304

EDAI Framework for Integrating Equity, Diversity, and Inclusion Throughout the Lifecycle of AI to Improve Health and Oral Health Care: Qualitative Study (e63356)
Samira Abbasgholizadeh Rahimi, Richa Shrivastava, Anita Brown-Johnson, Pascale Caidor, Claire Davies, Amal Idrissi Janati, Pascaline Kengne Talla, Sreenath Madathil, Bettina Willie, Elham Emami. 12316

A Taxonomy and Archetypes of AI-Based Health Care Services: Qualitative Study (e53986)
Marlene Blaß, Henner Gimpel, Philip Karnebogen. 12369

Development and Validation of a Prediction Model Using Sella Magnetic Resonance Imaging–Based Radiomics and Clinical Parameters for the Diagnosis of Growth Hormone Deficiency and Idiopathic Short Stature: Cross-Sectional, Multicenter Study (e54641)
Kyungchul Song, Taehoon Ko, Hyun Chae, Jun Oh, Ho-Seong Kim, Hyun Shin, Jeong-Ho Kim, Ji-Hoon Na, Chae Park, Beomseok Sohn. . .
1 2 3 8 6

Early Detection of Dementia in Populations With Type 2 Diabetes: Predictive Analytics Using Machine Learning Approach (e52107)
Phan Thanh Phuc, Phung-Anh Nguyen, Nam Nguyen, Min-Huei Hsu, Nguyen Le, Quoc-Viet Tran, Chih-Wei Huang, Hsuan-Chia Yang, Cheng-Yu Chen, Thi Le, Minh Le, Hoang Nguyen, Christine Lu, Jason Hsu. 12405

EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model (e60063)
Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Yue Wu, Mingpu Xu, Le Gao, Yinwen Li, Xianwen Shang, Danli Shi, Mingguang He. . .
1 2 4 2 1

Leveraging Large Language Models for Improved Understanding of Communications With Patients With Cancer in a Call Center Setting: Proof-of-Concept Study (e63892)
Seungbeom Cho, Mangyeong Lee, Jaewook Yu, Junghee Yoon, Jae-Boong Choi, Kyu-Hwan Jung, Juhee Cho. 12455

Development and Validation of a Literature Screening Tool: Few-Shot Learning Approach in Systematic Reviews (e56863)
Phongphat Wiwatthanasetthakarn, Wanchana Ponthongmak, Panu Looareesuwan, Amarit Tansawet, Pawin Numthavaj, Gareth McKay, John Attia, Ammarin Thakkinstian. 12465

Investigating Older Adults' Perceptions of AI Tools for Medication Decisions: Vignette-Based Experimental Survey (e60794)
Sarah Vordenberg, Julianna Nichols, Vincent Marshall, Kristie Weir, Michael Dorsch. 12477

Artificial Intelligence–Aided Diagnosis System for the Detection and Classification of Private-Part Skin Diseases: Decision Analytical Modeling Study (e52914)
Wei Wang, Xiang Chen, Licong Xu, Kai Huang, Shuang Zhao, Yong Wang. 12493

What's in a Name? Experimental Evidence of Gender Bias in Recommendation Letters Generated by ChatGPT (e51837)
Deanna Kaplan, Roman Palitsky, Santiago Arconada Alvarez, Nicole Pozzo, Morgan Greenleaf, Ciara Atkinson, Wilbur Lam. 12506

Parents' Perceptions of Their Parenting Journeys and a Mobile App Intervention (Parentbot—A Digital Healthcare Assistant): Qualitative Process Evaluation (e56894)
Joelle Chua, Mahesh Choolani, Cornelia Chee, Huso Yi, Yiong Chan, Joan Lalor, Yap Chong, Shefaly Shorey. 12520

XSL•FO
RenderX

Comparison of the Quality of Discharge Letters Written by Large Language Models and Junior Clinicians: Single-Blinded Study (e57721)
Joshua Tung, Sunil Gill, Gerald Sng, Daniel Lim, Yuhe Ke, Ting Tan, Liyuan Jin, Kabilan Elangovan, Jasmine Ong, Hairil Abdullah, Daniel Ting, Tsung Chong. 12537

Effectiveness and User Experience of a Smoking Cessation Chatbot: Mixed Methods Study Comparing Motivational Interviewing and Confrontational Counseling (e53134)
Linwei He, Erkan Basar, Emiel Krahmer, Reinout Wiers, Marjolijn Antheunis. 12544

Health Professionals' Views on the Use of Conversational Agents for Health Care: Qualitative Descriptive Study (e49387)
A MacNeill, Lillian MacNeill, Alison Luke, Shelley Doucet. 12562

Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis (e55164)
Michelle Nguyen, João Sedoc, Casey Taylor. 12573

Large Language Models May Help Patients Understand Peer-Reviewed Scientific Articles About Ophthalmology: Development and Usability Study (e59843)
Reza Kianian, Deyu Sun, William Rojas-Carabali, Rupesh Agrawal, Edmund Tsui. 12593

Advanced Technology in a Real-World Rehabilitation Setting: Longitudinal Observational Study on Clinician Adoption and Implementation (e60374)
Louise Pearce, Julie Pryor, Jason Redhead, Catherine Sherrington, Leanne Hassett. 12618

Designing A Blockchain-Empowered Telehealth Artifact for Decentralized Identity Management and Trustworthy Communication: Interdisciplinary Approach (e46556)
Xueping Liang, Nabid Alam, Tahmina Sultana, Eranga Bandara, Sachin Shetty. 12635

Automatic Recommender System of Development Platforms for Smart Contract–Based Health Care Insurance Fraud Detection Solutions: Taxonomy and Performance Evaluation (e50730)
Rima Kaafarani, Leila Ismail, Oussama Zahwe. 12659

Redefining Health Care Data Interoperability: Empirical Exploration of Large Language Models in Information Exchange (e56614)
Dukyong Yoon, Changho Han, Dong Kim, Songsoo Kim, SungA Bae, Jee Ryu, Yujin Choi. 12687

Digital Maturity as a Predictor of Quality and Safety Outcomes in US Hospitals: Cross-Sectional Observational Study (e56316)
Anne Snowdon, Abdulkadir Hussein, Melissa Danforth, Alexandra Wright, Reid Oakes. 12716

Assessing Opportunities and Barriers to Improving the Secondary Use of Health Care Data at the National Level: Multicase Study in the Kingdom of Saudi Arabia and Estonia (e53369)
Janek Metsallik, Dirk Draheim, Zlatan Sabic, Thomas Novak, Peeter Ross. 12730

Acceptance of Electronic Labeling for Medicinal Product Information Among Malaysian Hospital Patients: Cross-Sectional Study (e56591)
Xin Loh, Ai Woo, Azwa Haris, Cheryl Pereira, Bee Tan. 12745

Construction and Application of a Private 5G Standalone Medical Network in a Smart Health Environment: Exploratory Practice From China (e52404)
Baozhan Chen, Xiaobing Shi, Tianyi Feng, Shuai Jiang, Yunkai Zhai, Mingxing Ren, Dongqing Liu, Chengzeng Wang, Jinghong Gao. 12761

XSL•FO
RenderX

In-Depth Examination of the Functionality and Performance of the Internet Hospital Information Platform: Development and Usability Study (e54018)
Guang-Wei Zhang, Bin Li, Zheng-Min Gu, Wei-Feng Yang, Yi-Ran Wang, Hui-Jun Li, Han-Bing Zheng, Ying-Xu Yue, Kui-Zhong Wang, Mengchun Gong, Da-Xin Gong. 12779

Investigating the Norwegian eHealth Governance Model: Document Study (e59717)
Line Linstad, Hilde Bjørnå, Anne Moen, Truls Tunby Kristiansen, Anne Hansen. 12793

Multifactor Quality and Safety Analysis of Semaglutide Products Sold by Online Sellers Without a Prescription: Market Surveillance, Content Analysis, and Product Purchase Evaluation Study (e65440)
Amir Ashraf, Tim Mackey, Róbert Vida, Gy z Kulcsár, János Schmidt, Orsolya Balázs, Bálint Domián, Jiawei Li, Ibolya Csákó, András Fittler. .
1 2 8 0 9

Big 5 Personality Traits and Individual- and Practice-Related Characteristics as Influencing Factors of Digital Maturity in General Practices: Quantitative Web-Based Survey Study (e52085)
Lisa Weik, Leonard Fehring, Achim Mortsiefer, Sven Meister. 12828

Attitudes Toward and Use of eHealth Technologies Among German Dermatologists: Repeated Cross-Sectional Survey in 2019 and 2021 (e45817)
Matthias Augustin, Patrick Reinders, Toni Janke, Klaus Strömer, Ralph von Kiedrowski, Natalia Kirsten, Alexander Zink, Marina Otten. 12851

The Acceptance and Use of Digital Technologies for Self-Reporting Medication Safety Events After Care Transitions to Home in Patients With Cancer: Survey Study (e47685)
Yun Jiang, Misun Hwang, Youmin Cho, Christopher Friese, Sarah Hawley, Milisa Manojlovich, John Krauss, Yang Gong. 12863

Exploring the Impact of In Basket Metrics on the Adoption of a New Electronic Health Record System Among Specialists in a Tertiary Hospital in Alberta: Descriptive Study (e53122)
Melita Avdagovska, Craig Kuziemsky, Helia Koosha, Maliheh Hadizadeh, Robert Pauly, Timothy Graham, Tania Stafinski, David Bigam, Narmin Kassam, Devidas Menon. 12878

Influence of Physical Attractiveness and Gender on Patient Preferences in Digital Doctor Consultations: Experimental Study (e46551)
Xia Wei, Shubin Yu, Changxu Li. 12890

How Informal Carers Support Video Consulting in Physiotherapy, Heart Failure, and Cancer: Qualitative Study Using Linguistic Ethnography (e51695)
Lucas Seuren, Sara Shaw. 12903

eHealth Literacy and the Use of NHS 111 Online Urgent Care Service in England: Cross-Sectional Survey (e50376)
Joanne Turnbull, Jane Prichard, Jennifer MacLellan, Catherine Pope. 12920

Effect of Digital Early Warning Scores on Hospital Vital Sign Observation Protocol Adherence: Stepped-Wedge Evaluation (e46691)
David Wong, Timothy Bonnici, Stephen Gerry, Jacqueline Birks, Peter Watkinson. 12933

Insights From a Clinically Orientated Workshop on Health Care Cybersecurity and Medical Technology: Observational Study and Thematic Analysis (e50505)
Isabel Straw, Irina Brass, Andrew Mkwashi, Inika Charles, Amelie Soares, Caroline Steer. 12947

A Patient-Driven Mobile Health Innovation in Cystic Fibrosis Care: Comparative Cross-Case Study (e50527)
Pamela Mazzocato, Jamie Luckhaus, Moa Malmqvist Castillo, Johan Burnett, Andreas Hager, Gabriela Oates, Carolina Wannheden, Carl Savage. 12958

XSL•FO
RenderX

Exploring How Patients Are Supported to Use Online Services in Primary Care in England Through “Digital Facilitation”: Survey Study (e56528)
Rachel Winder, John Campbell, Nurunnahar Akter, Abodunrin Aminu, Jeffrey Lambert, Emma Cockcroft, Chloe Thomas, Christopher Clark, Carol Bryce, Jon Sussex, Helen Atherton, Christine Marriott, Gary Abel. 12972

Digital Media Usage Behavior and Its Impact on the Physician-Patient Relationship: Cross-Sectional Study Among Individuals Affected by Psoriasis in Germany (e57823)
Mert Erbas, Stefanie Ziehfreund, Hannah Wecker, Tilo Biedermann, Alexander Zink. 12992

Long-Term Glycemic Control Improvement After the Home and Self-Care Program for Patients With Type 1 Diabetes: Real-World–Based Cohort Study (e60023)
Dae-Jeong Koo, Sun-Joon Moon, Suhyeon Moon, Se Park, Eun-Jung Rhee, Won-Young Lee, Cheol-Young Park. 13005

Evaluating the National Rollout of a Type 2 Diabetes Self-Management Intervention: Qualitative Interview Study With Local National Health Service Leads Responsible for Implementation (e55546)
Lisa Brunton, Sarah Cotterill, Paul Wilson. 13020

Evaluating the “Zindagi Mehfooz” Electronic Immunization Registry and Suite of Digital Health Interventions to Improve the Coverage and Timeliness of Immunization Services in Sindh, Pakistan: Mixed Methods Study (e52792)
Patricia Mechael, Sara Gilani, Ahsan Ahmad, Amnesty LeFevre, Diwakar Mohan, Asra Memon, Mubarak Shah, Danya Siddiqi, Subhash Chandir, Riswana Soundardjee. 13034

Accessibility, Cost, and Quality of an Online Regular Follow-Up Visit Service at an Internet Hospital in China: Mixed Methods Study (e54902)
Kun Wang, Wenxin Zou, Yingsi Lai, Chun Hao, Ning Liu, Xiang Ling, Xiaohan Liu, Ting Liu, Xin Yang, Chenxi Zu, Shaolong Wu. 13048

Evaluating the Implementation and Clinical Effectiveness of an Innovative Digital First Care Model for Behavioral Health Using the RE-AIM Framework: Quantitative Evaluation (e54528)
Samuel Nordberg, Brittany Jaso-Yim, Pratha Sah, Keke Schuler, Mara Eyllon, Mariesa Pennine, Georgia Hoyler, J Barnes, Lily Murillo, Heather O'Dea, Laura Orth, Elizabeth Rogers, George Welch, Gabrielle Peloquin, Soo Youn. 13064

Clarifying the Concepts of Personalization and Tailoring of eHealth Technologies: Multimethod Qualitative Study (e50497)
Iris ten Klooster, Hanneke Kip, Sina Beyer, Lisette van Gemert-Pijnen, Saskia Kelders. 13076

Digital Health Platform for Improving the Effect of the Active Health Management of Chronic Diseases in the Community: Mixed Methods Exploratory Study (e50959)
Zhiheng Zhou, Danian Jin, Jinghua He, Shengqing Zhou, Jiang Wu, Shuangxi Wang, Yang Zhang, Tianyuan Feng. 13096

Effects of Virtual Care on Patient and Provider Experience of the Clinical Encounter: Qualitative Hermeneutic Study (e52552)
Graham McCaffrey, Erin Wilson, Lela Zimmer, Anurag Singh, Steinunn Jonatansdottir, Peter Zimmer, David Snadden, Ian Graham, Martha MacLeod. 13114

The Views and Experiences of Integrated Care System Commissioners About the Adoption and Implementation of Virtual Wards in England: Qualitative Exploration Study (e56494)
Laura McGowan, Fiona Graham, Jan Lecouturier, Louis Goffe, Carlos Echevarria, Michael Kelly, Falko Sniehotta. 13124

Dimensions and Subcategories of Digital Maturity in General Practice: Qualitative Study (e57786)
Timo Neunaber, Achim Mortsiefer, Sven Meister. 13140

XSL•FO
RenderX

Needs Expressed in Peer-to-Peer Web-Based Interactions Among People With Depression and Anxiety Disorders Hospitalized in a Mental Health Facility: Mixed Methods Study (e51506)
Dawid Storman, Paweł Jemioło, Zuzanna Sawiec, Mateusz Swierz, Ewa Antonowicz, Malgorzata Bala, Anna Prokop-Dorner. 13167

Evaluation and Comparison of Academic Impact and Disruptive Innovation Level of Medical Journals: Bibliometric Analysis and Disruptive Evaluation (e55121)
Yuyan Jiang, Xue-li Liu, Zixuan Zhang, Xinru Yang. 13206

Mapping Knowledge Landscapes and Emerging Trends in AI for Dementia Biomarkers: Bibliometric and Visualization Analysis (e57830)
Wenhao Qi, Xiaohong Zhu, Danni He, Bin Wang, Shihua Cao, Chaoqun Dong, Yunhua Li, Yanfei Chen, Bingsheng Wang, Yankai Shi, Guowei Jiang, Fang Liu, Lizzy Boots, Jiaqi Li, Xiajing Lou, Jiani Yao, Xiaodong Lu, Junling Kang. 13224

Evaluating Algorithmic Bias in 30-Day Hospital Readmission Models: Retrospective Analysis (e47125)
H Wang, Jonathan Weiner, Suchi Saria, Hadi Kharrazi. 13270

The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research (e46777)
Joseph Romano, Van Truong, Rachit Kumar, Mythreye Venkatesan, Britney Graham, Yun Hao, Nick Matsumoto, Xi Li, Zhiping Wang, Marylyn Ritchie, Li Shen, Jason Moore. 13288

Searching COVID-19 Clinical Research Using Graph Queries: Algorithm Development and Validation (e52655)
Francesco Invernici, Anna Bernasconi, Stefano Ceri. 13304

Comparing Federal Communications Commission and Microsoft Estimates of Broadband Access for Mental Health Video Telemedicine Among Veterans: Retrospective Cohort Study (e47100)
Amy O'Shea, Kailey Mulligan, Knute Carter, Bjarni Haraldsson, Charlie Wray, Ariana Shahnazi, Peter Kaboli. 13322

Public Maternal Health Dashboards in the United States: Descriptive Assessment (e56804)
Jennifer Callaghan-Koru, Paige Newman Chargois, Tanvangi Tiwari, Clare Brown, William Greenfield, Güne Koru. 13335

Data Visualization Preferences in Remote Measurement Technology for Individuals Living With Depression, Epilepsy, and Multiple Sclerosis: Qualitative Study (e43954)
Sara Simblett, Erin Dawe-Lane, Gina Gilpin, Daniel Morris, Katie White, Sinan Erturk, Julie Devonshire, Simon Lees, Spyridon Zormpas, Ashley Polhemus, Gergely Temesi, Nicholas Cummins, Matthew Hotopf, Til Wykes, RADAR-CNS Consortium. 13351

Internet-Based Social Connections of Black American College Students in Pre–COVID-19 and Peri–COVID-19 Pandemic Periods: Network Analysis (e55531)
Eun Lee, Heejun Kim, Yildiz Esener, Terika McCall. 13361

An Electronic Medical Record–Based Prognostic Model for Inpatient Falls: Development and Internal-External Cross-Validation (e59634)
Rex Parsons, Robin Blythe, Susanna Cramb, Ahmad Abdel-Hafez, Steven McPhail. 13375

Factor Analysis of Patients Who Find Tablets or Capsules Difficult to Swallow Due to Their Large Size: Using the Personal Health Record Infrastructure of Electronic Medication Notebooks (e54645)
Masaki Asano, Shungo Imai, Yuri Shimizu, Hayato Kizaki, Yukiko Ito, Makoto Tsuchiya, Ryoko Kuriyama, Nao Yoshida, Masanori Shimada, Takanori Sando, Tomo Ishijima, Satoko Hori. 13388

Tanzania's and Germany's Digital Health Strategies and Their Consistency With the World Health Organization's Global Strategy on Digital Health 2020-2025: Comparative Policy Analysis (e52150)
Felix Holl, Jennifer Kircher, Attila Hertelendy, Felix Sukums, Walter Swoboda. 13400

Comparing Open-Access Database and Traditional Intensive Care Studies Using Machine Learning: Bibliometric Analysis Study (e48330)
Yuhe Ke, Rui Yang, Nan Liu. 13432

XSL•FO
RenderX

Effects of a Digital Mental Health Intervention on Perceived Stress and Rumination in Adolescents Aged 13 to 17 Years: Randomized Controlled Trial (e54282)
Eliane Boucher, Haley Ward, Cynthia Miles, Robert Henry, Sarah Stoeckl. 13445

Perceptions of Illness Control, Coherence, and Self-Efficacy Following a Web-Based Lifestyle Program for Multiple Sclerosis: A Qualitative Analysis of Semistructured Interviews (e60240)
Rebekah Davenport, William Bevens, Sandra Neate, Pia Jelinek, Maggie Yu, George Jelinek, Jeanette Reece. 13462

Pharmacogenetics Clinical Decision Support Systems for Primary Care in England: Co-Design Study (e49230)
Videha Sharma, John McDermott, Jessica Keen, Simon Foster, Pauline Whelan, William Newman. 13476

Association of Remote Patient-Reported Outcomes and Step Counts With Hospitalization or Death Among Patients With Advanced Cancer Undergoing Chemotherapy: Secondary Analysis of the PROStep Randomized Trial (e51059)
Christopher Manz, Emily Schriver, William Ferrell, Joelle Williamson, Jonathan Wakim, Neda Khan, Michael Kopinsky, Mohan Balachandran, Jinbo Chen, Mitesh Patel, Samuel Takvorian, Lawrence Shulman, Justin Bekelman, Ian Barnett, Ravi Parikh. 13505

Efficacy of Save Medical Corporation (SMC)–01, a Smartphone App Designed to Support Type 2 Diabetes Self-Management Based on Established Guidelines: Randomized Controlled Trial (e53740)
Nicholas Leung, Kayo Waki, Satoshi Nozoe, Shunpei Enomoto, Ryo Saito, Sakurako Hamagami, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe, Yukiko Onishi. 13515

Outpatient Video Visits During the COVID-19 Pandemic: Cross-Sectional Survey Study of Patients' Experiences and Characteristics (e49058)
Stefanie van den Bosch, Demi van Dalen, Marjan Meinders, Harry van Goor, Stefaan Bergé, Martijn Stommel, Sandra van Dulmen. 13625

#ProtectOurElders: Analysis of Tweets About Older Asian Americans and Anti-Asian Sentiments During the COVID-19 Pandemic (e45864)
Reuben Ng, Nicole Indran. 13642

Identification of Predictors for Clinical Deterioration in Patients With COVID-19 via Electronic Nursing Records: Retrospective Observational Study (e53343)
Sumi Sung, Youlim Kim, Su Kim, Hyesil Jung. 13653

Understanding the Impact of Communicating Uncertainty About COVID-19 in the News: Randomized Between-Subjects Factorial Experiment (e51910)
Rui Zhao, Xuerong Lu, Jiayi Yang, Biao Li. 13663

Difference in Rumor Dissemination and Debunking Before and After the Relaxation of COVID-19 Prevention and Control Measures in China: Infodemiology Study (e48564)
Xiaoqi Liu, Qingyuan Hu, Jie Wang, Xusheng Wu, Dehua Hu. 13674

Implementation of Inpatient Electronic Consultations During the COVID-19 Crisis and Its Sustainability Beyond the Pandemic: Quality Improvement Study (e55623)
Anna Aledia, Amish Dangodara, Aanya Amin, Alpesh Amin. 13688

Prevalence and Factors Associated With Willingness to Sustain Pandemic-Induced Digital Work in the General Population and Moderating Effects of Screen Hours: Cross-Sectional Study (e53321)
Jiaying Li, Daniel Fong, Mandy Ho, Edmond Choi, Kris Lok, Jung Lee, WenJie Duan, Janet Wong, Chia-Chin Lin. 13696

Identifying Reddit Users at a High Risk of Suicide and Their Linguistic Features During the COVID-19 Pandemic: Growth-Based Trajectory Model (e48907)
Yifei Yan, Jun Li, Xingyun Liu, Qing Li, Nancy Yu. 13709

XSL•FO
RenderX

Messenger Use and Video Calls as Correlates of Depressive and Anxiety Symptoms: Results From the Corona Health App Study of German Adults During the COVID-19 Pandemic (e45530)
Johanna-Sophie Edler, Yannik Terhorst, Rüdiger Pryss, Harald Baumeister, Caroline Cohrdes. 13725

Crisis and Emergency Risk Communication and Emotional Appeals in COVID-19 Public Health Messaging: Quantitative Content Analysis (e56854)
Shirley Ho, Agnes Chuah, Vanessa Ho, Sonny Rosenthal, Hye Kim, Shannon Soh. 13741

Online Media Use and COVID-19 Vaccination in Real-World Personal Networks: Quantitative Study (e58257)
Iulian Oan , Marian-Gabriel Hâncean, Matjaž Perc, Jürgen Lerner, Bianca-Elena Mih il , Marius Geant , José Molina, Isabela Tinc , Carolina Espina. 13755

Chinese Public Attitudes and Opinions on Health Policies During Public Health Emergencies: Sentiment and Topic Analysis (e58518)
Min Liu, Shuo Yuan, Bingyan Li, Yuxi Zhang, Jia Liu, Cuixia Guan, Qingqing Chen, Jiayi Ruan, Lunfang Xie. 13773

Ever Use of Telehealth in Nebraska by March 2021: Cross-Sectional Analysis (e53320)
Lisa Smith, George Johnson, Snehal Jadhav, Josiane Kabayundo, Muskan Ahuja, Hongmei Wang, Kendra Ratnapradipa. 13786

Studies of Artificial Intelligence/Machine Learning Registered on ClinicalTrials.gov: Cross-Sectional Study With Temporal Trends, 2010-2023 (e57750)
Shoko Maru, Michael Matthias, Ryohei Kuwatsuru, Ross Simpson Jr. 13814

Adolescent and Parent Perspectives on Digital Phenotyping in Youths With Chronic Pain: Cross-Sectional Mixed Methods Survey Study (e47781)
Bridget Nestor, Justin Chimoff, Camila Koike, Elissa Weitzman, Bobbie Riley, Kristen Uhl, Joe Kossowsky. 13833

Natural Language Processing of Referral Letters for Machine Learning–Based Triaging of Patients With Low Back Pain to the Most Appropriate Intervention: Retrospective Study (e46857)
Sebastian Fudickar, Carsten Bantel, Jannik Spieker, Heinrich Töpfer, Patrick Stegeman, Henrica Schiphorst Preuper, Michiel Reneman, André Wolff, Remko Soer. 13848

Comparison of Two Symptom Checkers (Ada and Symptoma) in the Emergency Department: Randomized, Crossover, Head-to-Head, Double-Blinded Study (e56514)
Johannes Knitza, Ragip Hasanaj, Jonathan Beyer, Franziska Ganzer, Anna Slagman, Myrto Bolanaki, Hendrik Napierala, Malte Schmieding, Nizam Al-Zaher, Till Orlemann, Felix Muehlensiepen, Julia Greenfield, Nicolas Vuillerme, Sebastian Kuhn, Georg Schett, Stephan Achenbach, Katharina Dechant. 13862

Unveiling Usage Patterns and Explaining Usage of Symptom Checker Apps: Explorative Longitudinal Mixed Methods Study (e55161)
Anna-Jasmin Wetzel, Christine Preiser, Regina Müller, Stefanie Joos, Roland Koch, Tanja Henking, Hannah Haumann. 13873

User-Centered Development of a Patient Decision Aid for Choice of Early Abortion Method: Multi-Cycle Mixed Methods Study (e48793)
Kate Wahl, Melissa Brooks, Logan Trenaman, Kirsten Desjardins-Lorimer, Carolyn Bell, Nazgul Chokmorova, Romy Segall, Janelle Syring, Aleyah Williams, Linda Li, Wendy Norman, Sarah Munro. 13984

Effective Privacy Protection Strategies for Pregnancy and Gestation Information From Electronic Medical Records: Retrospective Study in a National Health Care Data Network in China (e46455)
Chao Liu, Yuanshi Jiao, Licong Su, Wenna Liu, Haiping Zhang, Sheng Nie, Mengchun Gong. 14000

Social and Ethical Implications of Digital Crisis Technologies: Case Study of Pandemic Simulation Models During the COVID-19 Pandemic (e45723)
Gabriel Bartl. 14013

XSL•FO
RenderX

Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study (e54800)
Kelly Quinn, Sarah Leiser Ransom, Carrie O'Connell, Naoko Muramatsu, David Marquez, Jessie Chin. 14024

Ecological Momentary Intervention to Replace Sedentary Time With Physical Activity to Improve Executive Function in Midlife and Older Latino Adults: Pilot Randomized Controlled Trial (e55079)
Ulf Bronas, David Marquez, Cynthia Fritschi, Katherine Petrarca, Spyros Kitsiou, Olu Ajilore, Nathan Tintle. 14040

Gaming-Based Tele-Exercise Program to Improve Physical Function in Frail Older Adults: Feasibility Randomized Controlled Trial (e56810)
Lakshmi Kannan, Upasana Sahu, Savitha Subramaniam, Neha Mehta, Tanjeev Kaur, Susan Hughes, Tanvi Bhatt. 14065

Targeting Daily Positive Events to Improve Emotional and Functional Well-Being in Adults With Fibromyalgia: Insights From the LARKSPUR Randomized Controlled Trial (e54678)
Anthony Ong, Kenneth Wilcox, M Reid, Elaine Wethington, Dakota Cintron, Elizabeth Addington, Selin Goktas, Judith Moskowitz. 14077

Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study (e48996)
Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Michael Paget, Christopher Naugler. 14092

Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study (e51926)
Xiaocong Liu, Jiageng Wu, An Shao, Wenyue Shen, Panpan Ye, Yao Wang, Juan Ye, Kai Jin, Jie Yang. 14103

Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study (e48443)
You-Qian Lee, Ching-Tai Chen, Chien-Chang Chen, Chung-Hong Lee, Peitsz Chen, Chi-Shin Wu, Hong-Jie Dai. 14115

Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study (e51069)
Ziru Fu, Yu Hsu, Christian Chan, Chaak Lau, Joyce Liu, Paul Yip. 14137

Quality and Dependability of ChatGPT and DingXiangYuan Forums for Remote Orthopedic Consultations: Comparative Analysis (e50882)
Zhaowen Xue, Yiming Zhang, Wenyi Gan, Huajun Wang, Guorong She, Xiaofei Zheng. 14151

An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study (e54580)
Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, Yuxiang Li. 14162

Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study (e52935)
Joseph Mugaanyi, Liuying Cai, Sumei Cheng, Caide Lu, Jing Huang. 14176

Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study (e56655)
Zhe He, Balu Bhasuran, Qiao Jin, Shubo Tian, Karim Hanna, Cindy Shavor, Lisbeth Arguello, Patrick Murray, Zhiyong Lu. 14183

ChatGPT's Performance in Cardiac Arrest and Bradycardia Simulations Using the American Heart Association's Advanced Cardiovascular Life Support Guidelines: Exploratory Study (e55037)
Cecilia Pham, Romi Govender, Salik Tehami, Summer Chavez, Omolola Adepoju, Winston Liaw. 14200

XSL•FO
RenderX

Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study (e54419)
Annessa Kernberg, Jeffrey Gold, Vishnu Mohan. 14210

Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content (e55847)
Xiaolei Lv, Xiaomeng Zhang, Yuan Li, Xinxin Ding, Hongchang Lai, Junyu Shi. 14229

Physician Versus Large Language Model Chatbot Responses to Web-Based Questions From Autistic Patients in Chinese: Cross-Sectional Comparative Analysis (e54706)
Wenjie He, Wenyan Zhang, Ya Jin, Qiang Zhou, Huadan Zhang, Qing Xia. 14238

Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study (e53297)
Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven Meuth, Lennert Böhm, Marc Pawlitzki. 14248

Assessing the Reproducibility of the Structured Abstracts Generated by ChatGPT and Bard Compared to Human-Written Abstracts in the Field of Spine Surgery: Comparative Analysis (e52001)
Hong Kim, Jae Yang, Dong-Gune Chang, Lawrence Lenke, Javier Pizones, René Castelein, Kota Watanabe, Per Trobisch, Gregory Mundis Jr, Seung Suh, Se-Il Suk. 14258

ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis (e56110)
John Hoppe, Matthias Auer, Anna Strüven, Steffen Massberg, Christopher Stremmel. 14268

Patient-Representing Population's Perceptions of GPT-Generated Versus Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment (e60336)
Thomas Huang, Conrad Safranek, Vimig Socrates, David Chartash, Donald Wright, Monisha Dilip, Rohit Sangal, Richard Taylor. 14294

Performance of Large Language Models in Patient Complaint Resolution: Web-Based Cross-Sectional Survey (e56413)
Lorraine Yong, Joshua Tung, Zi Lee, Win Kuan, Mui Chua. 14306

Characterizing the Adoption and Experiences of Users of Artificial Intelligence–Generated Health Information in the United States: Cross-Sectional Questionnaire Study (e55138)
Oluwatobiloba Ayo-Ajibola, Ryan Davis, Matthew Lin, Jeffrey Riddell, Richard Kravitz. 14319

Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment (e55939)
Damien Gibson, Stuart Jackson, Ramesh Shanmugasundaram, Ishith Seth, Adrian Siu, Nariman Ahmadi, Jonathan Kam, Nicholas Mehan, Ruban Thanigasalam, Nicola Jeffery, Manish Patel, Scott Leslie. 14331

ChatGPT and Google Assistant as a Source of Patient Education for Patients With Amblyopia: Content Analysis (e52401)
Gloria Wu, David Lee, Weichen Zhao, Adrial Wong, Rohan Jhangiani, Sri Kurniawan. 14346

Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study (e56500)
Kevin Li, Adrian Fernandez, Rachel Schwartz, Natalie Rios, Marvin Carlisle, Gregory Amend, Hiren Patel, Benjamin Breyer. 14353

The Diagnostic Ability of GPT-3.5 and GPT-4.0 in Surgery: Comparative Analysis (e54985)
Jiayu Liu, Xiuting Liang, Dandong Fang, Jiqi Zheng, Chengliang Yin, Hui Xie, Yanteng Li, Xiaochun Sun, Yue Tong, Hebin Che, Ping Hu, Fan Yang, Bingxian Wang, Yuanyuan Chen, Gang Cheng, Jianning Zhang. 14366

XSL•FO
RenderX

Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage of Simulated Patients Using the Simple Triage and Rapid Treatment (START) Protocol: Gage Repeatability and Reproducibility Study (e55648)
Jeffrey Franc, Attila Hertelendy, Lenard Cheng, Ryan Hata, Manuela Verde. 14377

Ascle—A Python Natural Language Processing Toolkit for Medical Text Generation: Development and Evaluation Study (e60601)
Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha Dave, Tiarnan Keenan, Yuhe Ke, Chuan Hong, Nan Liu, Emily Chew, Dragomir Radev, Zhiyong Lu, Hua Xu, Qingyu Chen, Irene Li. 14386

Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study (e60291)
Jonathan Yau, Soheil Saadat, Edmund Hsu, Linda Murphy, Jennifer Roh, Jeffrey Suchard, Antonio Tapia, Warren Wiechmann, Mark Langdorf. 1 4 4 0 0

Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study (e58041)
Dingqiao Wang, Jiangbo Liang, Jinguo Ye, Jingni Li, Jingpeng Li, Qikai Zhang, Qiuling Hu, Caineng Pan, Dongliang Wang, Zhong Liu, Wen Shi, Danli Shi, Fei Li, Bo Qu, Yingfeng Zheng. 14439

Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study (e60226)
Shuai Ming, Xi Yao, Xiaohong Guo, Qingge Guo, Kunpeng Xie, Dandan Chen, Bo Lei. 14451

Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study (e59439)
Yuhe Ke, Rui Yang, Sui Lie, Taylor Lim, Yilin Ning, Irene Li, Hairil Abdullah, Daniel Ting, Nan Liu. 14465

Using Large Language Models to Abstract Complex Social Determinants of Health From Original and Deidentified Medical Notes: Development and Validation Study (e63445)
Alexandra Ralevski, Nadaa Taiyab, Michael Nossal, Lindsay Mico, Samantha Piekos, Jennifer Hadlock. 14476

Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study (e58329)
Junhyuk Seo, Dasol Choi, Taerim Kim, Won Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Lee, Edward Choi. 14491

AI in Dental Radiology—Improving the Efficiency of Reporting With ChatGPT: Comparative Study (e60684)
Daniel Stephan, Annika Bertsch, Matthias Burwinkel, Shankeeth Vinayahalingam, Bilal Al-Nawas, Peer Kämmerer, Daniel Thiem. 14504

Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis (e66114)
Hui Zong, Rongrong Wu, Jiaxue Cha, Jiao Wang, Erman Wu, Jiakun Li, Yi Zhou, Chi Zhang, Weizhe Feng, Bairong Shen. 14519

Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration (e54047)
Ziwei Zhao, Weiyi Zhang, Xiaolan Chen, Fan Song, James Gunasegaram, Wenyong Huang, Danli Shi, Mingguang He, Na Liu. 14535

Twenty-Five Years of Progress—Lessons Learned From JMIR Publications to Address Gender Parity in Digital Health Authorships: Bibliometric Analysis (e58950)
Annika Meyer, Thomas Streichert. 14600

Digital Health Applications (DiGAs) on a Fast Track: Insights From a Data-Driven Analysis of Prescribable Digital Therapeutics in Germany From 2020 to Mid-2024 (e59013)
Moritz Goeldner, Sara Gehder. 14670

XSL•FO
RenderX

Evaluation of the Veterans Health Administration's Digital Divide Consult for Tablet Distribution and Telehealth Adoption: Cohort Study (e59089)
Jacqueline Ferguson, James Van Campen, Cindie Slightam, Liberty Greene, Leonie Heyworth, Donna Zulman. 14686

Development Trends and Prospects of Technology-Based Solutions for Health Challenges in Aging Over the Past 25 Years: Bibliometric Analysis (e63367)
Lu Liu, Xiu-Ling Wang, Nuo Cheng, Fu-Min Yu, Hui-Jun Li, Yang Mu, Yonghui Yuan, Jia-Xin Dong, Yu-Dan Wu, Da-Xin Gong, Shuang Wang, Guang-Wei Zhang. 14804

Mapping the Evolution of Digital Health Research: Bibliometric Overview of Research Hotspots, Trends, and Collaboration of Publications in JMIR (1999-2024) (e58987)
Jing Hu, Chong Li, Yanlei Ge, Jingyi Yang, Siyi Zhu, Chengqi He. 14867

The Evolution of Health Information Technology for Enhanced Patient-Centric Care in the United States: Data-Driven Descriptive Study (e59791)
Wesley Barker, Wei Chang, Jordan Everson, Meghan Gabriel, Vaishali Patel, Chelsea Richwine, Catherine Strawley. 14913

Decoding the Digital Pulse: Bibliometric Analysis of 25 Years in Digital Health Research Through the Journal of Medical Internet Research (e60057)
Robert Kaczmarczyk, Theresa Wilhelm, Jonas Roos, Ron Martin. 14952

Telemonitoring for Chronic Heart Failure: Narrative Review of the 20-Year Journey From Concept to Standard Care in Germany (e63391)
Sebastian Spethmann, Gerhard Hindricks, Kerstin Koehler, Stefan Stoerk, Christiane Angermann, Michael Böhm, Birgit Assmus, Sebastian Winkler, Martin Möckel, Mirja Mittermaier, Monika Lelgemann, Daniel Reuter, Ralph Bosch, Alexander Albrecht, Stephan von Haehling, Thomas Helms, Stefan Sack, Tarek Bekfani, Jan Gröschel, Magdalena Koehler, Christoph Melzer, Jan Wintrich, Bettina Zippel-Schultz, Georg Ertl, Claus Vogelmeier, Nikolaos Dagres, Jasmin Zernikow, Friedrich Koehler. 15042

Reviews

Embodied Conversational Agents for Chronic Diseases: Scoping Review (e47134)
Zhili Jiang, Xiting Huang, Zhiqian Wang, Yang Liu, Lihua Huang, Xiaolin Luo. 123

Machine Learning Approaches for the Image-Based Identification of Surgical Wound Infections: Scoping Review (e52880)
Juan Tabja Bortesi, Jonathan Ranisau, Shuang Di, Michael McGillion, Laura Rosella, Alistair Johnson, PJ Devereaux, Jeremy Petch. 149

Engaging Parents in Technology-Assisted Interventions for Childhood Adversity: Systematic Review (e43994)
Grace Aldridge, Alessandra Tomaselli, Clare Nowell, Andrea Reupert, Anthony Jorm, Marie Yap. 164

The Reporting Quality of Machine Learning Studies on Pediatric Diabetes Mellitus: Systematic Review (e47430)
Zsombor Zrubka, Gábor Kertész, László Gulácsi, János Czere, Áron Hölgyesi, Hossein Nezhad, Amir Mosavi, Levente Kovács, Atul Butte, Márta Péntek. 193

Predictive Value of Machine Learning for Platinum Chemotherapy Responses in Ovarian Cancer: Systematic Review and Meta-Analysis (e48527)
Qingyi Wang, Zhuo Chang, Xiaofang Liu, Yunrui Wang, Chuwen Feng, Yunlu Ping, Xiaoling Feng. 215

Understanding the Integrated Health Management System Policy in China From Multiple Perspectives: Systematic Review and Content Analysis (e47197)
Yang Yu, Sufen Wang, Lijue You. 233

XSL•FO
RenderX

Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review (e45209)
Fouzia Amar, Alain April, Alain Abran. 254

The Performance of Wearable AI in Detecting Stress Among Students: Systematic Review and Meta-Analysis (e52622)
Alaa Abd-alrazaq, Mohannad Alajlani, Reham Ahmad, Rawan AlSaad, Sarah Aziz, Arfan Ahmed, Mohammed Alsahli, Rafat Damseh, Javaid Sheikh. 277

Digitally Delivered Interventions to Improve Nutrition Behaviors Among Resource-Poor and Ethnic Minority Groups With Type 2 Diabetes: Systematic Review (e42595)
Nazgol Karimi, Rachelle Opie, David Crawford, Stella O'Connell, Kylie Ball. 297

Race, Ethnicity, and Other Cultural Background Factors in Trials of Internet-Based Cognitive Behavioral Therapy for Depression: Systematic Review (e50780)
Robinson De Jesús-Romero, Amani Holder-Dixon, John Buss, Lorenzo Lorenzo-Luaces. 324

Effectiveness of Telemedicine for Musculoskeletal Disorders: Umbrella Review (e50090)
Silvia Bargeri, Greta Castellini, Jacopo Vitale, Stefania Guida, Giuseppe Banfi, Silvia Gianola, Federico Pennestrì. 338

Timing, Indicators, and Approaches to Digital Patient Experience Evaluation: Umbrella Systematic Review (e46308)
Tingting Wang, Guido Giunti, Richard Goossens, Marijke Melles. 353

Mapping Theories, Models, and Frameworks to Evaluate Digital Health Interventions: Scoping Review (e51098)
Geneviève Rouleau, Kelly Wu, Karishini Ramamoorthi, Cherish Boxall, Rebecca Liu, Shelagh Maloney, Jennifer Zelmer, Ted Scott, Darren Larsen, Harindra Wijeysundera, Daniela Ziegler, Sacha Bhatia, Vanessa Kishimoto, Carolyn Steele Gray, Laura Desveaux. 378

The Implementation of Federated Digital Identifiers in Health Care: Rapid Review (e45751)
Karishini Ramamoorthi, Vess Stamenova, Rebecca Liu, Onil Bhattacharyya. 393

Effectiveness and Feasibility of Telehealth-Based Dietary Interventions Targeting Cardiovascular Disease Risk Factors: Systematic Review and Meta-Analysis (e49178)
Rupal Trivedi, Shaimaa Elshafie, Randall Tackett, Henry Young, Elisabeth Sattler. 408

Effectiveness of eHealth Interventions Promoting Physical Activity in Children and Adolescents: Systematic Review and Meta-Analysis (e41649)
Irene Sequí-Domínguez, Iván Cavero-Redondo, Celia Álvarez-Bueno, Jose López-Gil, Vicente Martínez-Vizcaíno, Carlos Pascual-Morena. . .
4 2 3

Effects of eHealth Interventions on 24-Hour Movement Behaviors Among Preschoolers: Systematic Review and Meta-Analysis (e52905)
Shan Jiang, Johan Ng, Kar Chong, Bo Peng, Amy Ha. 442

Virtual Reality–Based Exercise Rehabilitation in Cancer-Related Dysfunctions: Scoping Review (e49312)
Zhenzhen Su, Liyan Zhang, Xuemin Lian, Miaomiao Guan. 466

Virtual Reality–Based Training in Chronic Low Back Pain: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e45406)
Ran Li, Yinghao Li, Youli Kong, Hanbin Li, Danrong Hu, Chenying Fu, Quan Wei. 487

Chatbots That Deliver Contraceptive Support: Systematic Review (e46758)
Rhiana Mills, Emily Mangone, Neal Lesh, Gayatri Jayal, Diwakar Mohan, Paula Baraitser. 505

XSL•FO
RenderX

Attrition in Conversational Agent–Delivered Mental Health Interventions: Systematic Review and Meta-Analysis (e48168)
Ahmad Jabir, Xiaowen Lin, Laura Martinengo, Gemma Sharp, Yin-Leng Theng, Lorainne Tudor Car. 519

The Effect of Digital Mental Health Literacy Interventions on Mental Health: Systematic Review and Meta-Analysis (e51268)
GeckHong Yeo, Stephanie Reich, Nicole Liaw, Elizabeth Chia. 543

The Impact of Digital Hospitals on Patient and Clinician Experience: Systematic Review and Qualitative Evidence Synthesis (e47715)
Oliver Canfell, Leanna Woods, Yasaman Meshkat, Jenna Krivit, Brinda Gunashanhar, Christine Slade, Andrew Burton-Jones, Clair Sullivan. . 5 8 4

Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review (e47923)
Karen O'Connor, Su Golder, Davy Weissenbacher, Ari Klein, Arjun Magge, Graciela Gonzalez-Hernandez. 608

Reducing Loneliness and Social Isolation of Older Adults Through Voice Assistants: Literature Review and Bibliometric Analysis (e50534)
Rachele Marziali, Claudia Franceschetti, Adrian Dinculescu, Alexandru Nistorescu, Dominic Kristály, Adrian Mo oi, Ronny Broekx, Mihaela Marin, Cristian Vizitiu, Sorin-Aurel Moraru, Lorena Rossi, Mirko Di Rosa. 632

The Use of Mobile Technologies to Promote Physical Activity and Reduce Sedentary Behaviors in the Middle East and North Africa Region: Systematic Review and Meta-Analysis (e53651)
Huong Tong, Aroub Alnasser, Najim Alshahrani, Rowaedh Bawaked, Reem AlAhmed, Reem Alsukait, Severin Rakic, Volkan Cetinkaya, Hazzaa Al-Hazzaa, Saleh Alqahtani. 652

Methodological Frameworks and Dimensions to Be Considered in Digital Health Technology Assessment: Scoping Review and Thematic Analysis (e48694)
Joan Segur-Ferrer, Carolina Moltó-Puigmartí, Roland Pastells-Peiró, Rosa Vivanco-Hidalgo. 700

Digital Interventions for Recreational Cannabis Use Among Young Adults: Systematic Review, Meta-Analysis, and Behavior Change Technique Analysis of Randomized Controlled Studies (e55031)
José Côté, Gabrielle Chicoine, Billy Vinette, Patricia Auger, Geneviève Rouleau, Guillaume Fontaine, Didier Jutras-Aswad. 720

Mobile Apps to Support Mental Health Response in Natural Disasters: Scoping Review (e49929)
Nwamaka Ezeonu, Attila Hertelendy, Medard Adu, Janice Kung, Ijeoma Itanyi, Raquel Dias, Belinda Agyapong, Petra Hertelendy, Francis Ohanyido, Vincent Agyapong, Ejemai Eboreime. 742

Investigating the Cost-Effectiveness of Telemonitoring Patients With Cardiac Implantable Electronic Devices: Systematic Review (e47616)
Sarah Raes, Andrea Prezzi, Rik Willems, Hein Heidbuchel, Lieven Annemans. 756

Patient and Staff Experience of Remote Patient Monitoring—What to Measure and How: Systematic Review (e48463)
Valeria Pannunzio, Hosana Morales Ornelas, Pema Gurung, Robert van Kooten, Dirk Snelders, Hendrikus van Os, Michel Wouters, Rob Tollenaar, Douwe Atsma, Maaike Kleinsmann. 778

Electronic Media Use and Sleep Quality: Updated Systematic Review and Meta-Analysis (e48356)
Xiaoning Han, Enze Zhou, Dong Liu. 794

Behavior Change Approaches in Digital Technology–Based Physical Rehabilitation Interventions Following Stroke: Scoping Review (e48725)
Helen Gooch, Kathryn Jarvis, Rachel Stockley. 808

The Applications of Artificial Intelligence for Assessing Fall Risk: Systematic Review (e54934)
Ana González-Castro, Raquel Leirós-Rodríguez, Camino Prada-García, José Benítez-Andrades. 827

XSL•FO
RenderX

Suitability of the Current Health Technology Assessment of Innovative Artificial Intelligence-Based Medical Devices: Scoping Literature Review (e51514)
Line Farah, Isabelle Borget, Nicolas Martelli, Alexandre Vallee. 843

Person-Generated Health Data in Women's Health: Scoping Review (e53327)
Jalisa Karim, Rachel Wan, Rhea Tabet, Derek Chiu, Aline Talhouk. 861

Reporting of Ethical Considerations in Qualitative Research Utilizing Social Media Data on Public Health Care: Scoping Review (e51496)
Yujie Zhang, Jiaqi Fu, Jie Lai, Shisi Deng, Zihan Guo, Chuhan Zhong, Jianyao Tang, Wenqiong Cao, Yanni Wu. 884

Mobile Health Apps, Family Caregivers, and Care Planning: Scoping Review (e46108)
Marjorie Kelley, Tia Powell, Djibril Camara, Neha Shah, Jenna Norton, Chelsea Deitelzweig, Nivedha Vaidy, Chun-Ju Hsiao, Jing Wang, Arlene Bierman. 902

Digital Behavior Change Intervention Designs for Habit Formation: Systematic Review (e54375)
Yujie Zhu, Yonghao Long, Hailiang Wang, Kun Lee, Lie Zhang, Stephen Wang. 916

Advances in the Application of AI Robots in Critical Care: Scoping Review (e54095)
Yun Li, Min Wang, Lu Wang, Yuan Cao, Yuyan Liu, Yan Zhao, Rui Yuan, Mengmeng Yang, Siqian Lu, Zhichao Sun, Feihu Zhou, Zhirong Qian, Hongjun Kang. 936

Effectiveness of mHealth Apps for Maternal Health Care Delivery: Systematic Review of Systematic Reviews (e49510)
Edward Ameyaw, Padmore Amoah, Obidimma Ezezika. 960

Methods, Indicators, and End-User Involvement in the Evaluation of Digital Health Interventions for the Public: Scoping Review (e55714)
Vera Weirauch, Clarissa Soehnchen, Anja Burmann, Sven Meister. 979

Vulnerability to Cyberattacks and Sociotechnical Solutions for Health Care Systems: Systematic Review (e46904)
Pius Ewoh, Tero Vartiainen. 995

Use of Patient-Generated Health Data From Consumer-Grade Devices by Health Care Professionals in the Clinic: Systematic Review (e49320)
Sharon Guardado, Maria Karampela, Minna Isomursu, Casandra Grundstrom. 1025

Examining the Effectiveness of Social Media for the Dissemination of Research Evidence for Health and Social Care Practitioners: Systematic Review and Meta-Analysis (e51418)
Sarah Roberts-Lewis, Helen Baxter, Gill Mein, Sophia Quirke-McFarlane, Fiona Leggat, Hannah Garner, Martha Powell, Sarah White, Lindsay Bearne. 1041

Physician Perspectives on Internet-Informed Patients: Systematic Review (e47620)
Qianfeng Lu, Peter Schulz. 1061

Integrated Real-World Data Warehouses Across 7 Evolving Asian Health Care Systems: Scoping Review (e56686)
Wen-Yi Shau, Handoko Santoso, Vincent Jip, Sajita Setia. 1077

The Effects of mHealth Interventions on Quality of Life, Anxiety, and Depression in Patients With Coronary Heart Disease: Meta-Analysis of Randomized Controlled Trials (e52341)
Qiao Hou, Le Liu, Ying Wu. 1096

XSL•FO
RenderX

Intensive Longitudinal Methods Among Adults With Breast or Lung Cancer: Scoping Review (e50224)
Joran Geeraerts, Kim de Nooijer, Lara Pivodic, Mark De Ridder, Lieve Van den Block. 1116

Characteristics and Effects of Home-Based Digital Health Interventions on Functional Outcomes in Older Patients With Hip Fractures After Surgery: Systematic Review and Meta-Analysis (e49482)
Suphawita Pliannuom, Kanokporn Pinyopornpanish, Nida Buawangpong, Nutchar Wiwatkunupakarn, Poppy Mallinson, Wichuda Jirapornchaoren, Chaisiri Angkurawaranon. 1150

Effects of Peer- or Professional-Led Support in Enhancing Adherence to Wearable Monitoring Devices Among Community-Dwelling Older Adults: Systematic Review of Randomized Controlled Trials (e53607)
Colette Chan, Mandy Kan, Arkers Wong. 1168

Wearable Technologies for Detecting Burnout and Well-Being in Health Care Professionals: Scoping Review (e50253)
Milica Barac, Samantha Scaletty, Leslie Hassett, Ashley Stillwell, Paul Croarkin, Mohit Chauhan, Sherry Chesak, William Bobo, Arjun Athreya, Liselotte Dyrbye. 1178

Effectiveness of Web-Based Mindfulness-Based Interventions for Patients With Cancer: Systematic Review and Meta-Analyses (e47704)
Ting Wang, Chulei Tang, Xiaoman Jiang, Yinning Guo, Shuqin Zhu, Qin Xu. 1191

Design of Digital Mental Health Platforms for Family Member Cocompletion: Scoping Review (e49431)
Ellen Welsh, Jennifer McIntosh, An Vuong, Zoe Cloud, Eliza Hartley, James Boyd. 1212

Cultural and Contextual Adaptation of Digital Health Interventions: Narrative Review (e55130)
Aila Naderbagi, Victoria Loblay, Iqthyer Zahed, Mahalakshmi Ekambareshwar, Adam Poulsen, Yun Song, Laura Ospina-Pinillos, Michael Krausz, Mostafa Mamdouh Kamel, Ian Hickie, Haley LaMonica. 1265

Telemedicine in Improving Glycemic Control Among Children and Adolescents With Type 1 Diabetes Mellitus: Systematic Review and Meta-Analysis (e51538)
Kun Zhang, Qiyuan Huang, Qiaosong Wang, Chengyang Li, Qirong Zheng, Zhuoyue Li, Dan Xu, Cuiling Xie, Mingqi Zhang, Rongjin Lin. . . .
1 2 7 9

The Use of Telepsychiatry Services in Emergency Settings: Scoping Review (e51814)
Ligat Shalev, Renana Eitan, Adam Rose. 1301

Effectiveness of eHealth Interventions in Improving Medication Adherence Among Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis (e58013)
Yiqun Miao, Yuan Luo, Yuhan Zhao, Mingxuan Liu, Huiying Wang, Ying Wu. 1313

Evaluating the Potential and Pitfalls of AI-Powered Conversational Agents as Humanlike Virtual Health Carers in the Remote Management of Noncommunicable Diseases: Scoping Review (e56114)
Sadia Anisha, Arkendu Sen, Chris Bain. 1328

Digital Health Interventions for Chronic Wound Management: A Systematic Review and Meta-Analysis (e47904)
Xinrui Bai, Hongyan Zhang, Yanxia Jiao, Chenlu Yuan, Yuxia Ma, Lin Han. 1375

Mapping Digital Public Health Interventions Among Existing Digital Technologies and Internet-Based Interventions to Maintain and Improve Population Health in Practice: Scoping Review (e53927)
Laura Maaß, Konstantinos Angoumis, Merle Freye, Chen-Chia Pan. 1393

Digital Interventions for Self-Management of Type 2 Diabetes Mellitus: Systematic Literature Review and Meta-Analysis (e55757)
David Kerr, David Ahn, Kayo Waki, Jing Wang, Boris Breznen, David Klonoff. 1424

XSL•FO
RenderX

Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review (e56930)
Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, Bertrand Lebouché. 1440

Supportive eHealth Technologies and Their Effects on Physical Functioning and Quality of Life for People With Lung Cancer: Systematic Review (e53015)
Suriya Kirkpatrick, Zoe Davey, Peter Wright, Catherine Henshall. 1468

Learning Outcomes of e-Learning in Psychotherapy Training and Comparison With Conventional Training Methods: Systematic Review (e54473)
Kasperi Mikkonen, Eeva-Eerika Helminen, Samuli Saarni, Suoma Saarni. 1486

eHealth-Integrated Psychosocial and Physical Interventions for Chronic Pain in Older Adults: Scoping Review (e55366)
Annalisa De Lucia, Cinzia Perlini, Alessandro Chiarotto, Sara Pachera, Ilenia Pasini, Lidia Del Piccolo, Valeria Donisi. 1506

Online Medical Consultation Service–Oriented Recommendations: Systematic Review (e46073)
Hongxun Jiang, Ziyue Mi, Wei Xu. 1526

Examining the Effectiveness of Electronic Patient-Reported Outcomes in People With Cancer: Systematic Review and Meta-Analysis (e49089)
Melissa Perry, Sally Taylor, Binish Khatoon, Amy Vercell, Corinne Faivre-Finn, Galina Velikova, Antonia Marsden, Calvin Heal, Janelle Yorke. . 1 5 5 2

Training Service Users in the Use of Telehealth: Scoping Review (e57586)
Emer Galvin, Shane Desselle, Blánaid Gavin, Fiona McNicholas, Shane Cullinan, John Hayden. 1574

Assessing the Effectiveness of eHealth Interventions to Manage Multiple Lifestyle Risk Behaviors Among Older Adults: Systematic Review and Meta-Analysis (e58174)
Beibei Shi, Guangkai Li, Shuang Wu, Hongli Ge, Xianliang Zhang, Si Chen, Yang Pan, Qiang He. 1585

Telehealth-Supported Exercise or Physical Activity Programs for Knee Osteoarthritis: Systematic Review and Meta-Analysis (e54876)
Xiao-Na Xiang, Ze-Zhang Wang, Jing Hu, Jiang-Yin Zhang, Ke Li, Qi-Xu Chen, Fa-Shu Xu, Yue-Wen Zhang, Hong-Chen He, Cheng-Qi He, Si-Yi Zhu. 1609

Implementing AI in Hospitals to Achieve a Learning Health System: Systematic Review of Current Enablers and Barriers (e49655)
Amir Kamel Rahimi, Oliver Pienaar, Moji Ghadimi, Oliver Canfell, Jason Pole, Sally Shrapnel, Anton van der Vegt, Clair Sullivan. 1629

Predictors of Health Care Practitioners' Intention to Use AI-Enabled Clinical Decision Support Systems: Meta-Analysis Based on the Unified Theory of Acceptance and Use of Technology (e57224)
Julius Dingel, Anne-Kathrin Kleine, Julia Cecil, Anna Sigl, Eva Lermer, Susanne Gaube. 1658

Longitudinal Coadaptation of Older Adults With Wearables and Voice-Activated Virtual Assistants: Scoping Review (e57258)
Kristina Kokorelias, Alisa Grigorovich, Maurita Harris, Umair Rehman, Louise Ritchie, AnneMarie Levy, Kerstin Denecke, Josephine McMurray. 1 6 8 3

Enhancing Patient Understanding of Laboratory Test Results: Systematic Review of Presentation Formats and Their Impact on Perception, Decision, Action, and Memory (e53993)
Frederieke van der Mee, Fleur Schaper, Jesse Jansen, Judith Bons, Steven Meex, Jochen Cals. 1701

Leadership for AI Transformation in Health Care Organization: Scoping Review (e54556)
Abi Sriharan, Nigar Sekercioglu, Cheryl Mitchell, Senthujan Senkaiahliyan, Attila Hertelendy, Tracy Porter, Jane Banaszak-Holl. 1720

XSL•FO
RenderX

Digital Serious Games to Promote Behavior Change in Children With Chronic Diseases: Scoping Review and Development of a Self-Management Learning Framework (e49692)
Made Sarasmita, Ya-Han Lee, Fan-Ying Chan, Hsiang-Yin Chen. 1734

Prevalence of Health Misinformation on Social Media—Challenges and Mitigation Before, During, and Beyond the COVID-19 Pandemic: Scoping Literature Review (e38786)
Dhouha Kbaier, Annemarie Kane, Mark McJury, Ian Kenny. 1753

Immersive and Nonimmersive Virtual Reality–Assisted Active Training in Chronic Musculoskeletal Pain: Systematic Review and Meta-Analysis (e48787)
Hermione Lo, Mengting Zhu, Zihui Zou, Cho Wong, Suzanne Lo, Vincent Chung, Samuel Wong, Regina Sit. 1769

The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review (e48320)
Laura Swinckels, Frank Bennis, Kirsten Ziesemer, Janneke Scheerman, Harmen Bijwaard, Ander de Keijzer, Josef Bruers. 1789

A Comprehensive Analysis of COVID-19 Misinformation, Public Health Impacts, and Communication Strategies: Scoping Review (e56931)
Sezer Kisa, Adnan Kisa. 1806

Data Quality–Driven Improvement in Health Care: Systematic Literature Review (e57615)
Anthony Lighterness, Michael Adcock, Lauren Scanlon, Gareth Price. 1825

Automated Interpretation of Lung Sounds by Deep Learning in Children With Asthma: Scoping Review and Strengths, Weaknesses, Opportunities, and Threats Analysis (e53662)
Isabelle Ruchonnet-Métrailler, Johan Siebert, Mary-Anne Hartley, Laurence Lacroix. 1846

Provenance Information for Biomedical Data and Workflows: Scoping Review (e51297)
Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, Atinkut Zeleke. 1 8 6 6

Barriers and Facilitators to the Implementation of Digital Health Services for People With Musculoskeletal Conditions in the Primary Health Care Setting: Systematic Review (e49868)
Mark van Tilburg, Ivar Spin, Martijn Pisters, J Staal, Raymond Ostelo, Miriam van der Velde, Cindy Veenhof, Corelien Kloek. 1887

Framework Development for Reducing Attrition in Digital Dietary Interventions: Systematic Review and Thematic Synthesis (e58735)
Jian Wang, Jinli Mahe, Yujia Huo, Weiyuan Huang, Xinru Liu, Yang Zhao, Junjie Huang, Feng Shi, Zhihui Li, Dou Jiang, Yilong Li, Garon Perceval, Lindu Zhao, Lin Zhang. 1914

Approaches to Evaluating Digital Health Technologies: Scoping Review (e50251)
Anneloek Rauwerdink, Pier Spinazze, Harm Gijsbers, Juul Molendijk, Sandra Zwolsman, Marlies Schijven, Niels Chavannes, Marise Kasteleyn. 1 9 3 4

Characteristics, Influence, Prevention, and Control Measures of the Mpox Infodemic: Scoping Review of Infodemiology Studies (e54874)
XiangYu Yan, Zhuo Li, Chunxia Cao, Longxin Huang, Yongjie Li, Xiangbin Meng, Bo Zhang, Maohe Yu, Tian Huang, Jiancheng Chen, Wei Li, Linhui Hao, Dongsheng Huang, Bin Yi, Ming Zhang, Shun Zha, Haijun Yang, Jian Yao, Pengjiang Qian, Chun Leung, Haojun Fan, Pei Jiang, Tiejun Shui. 1954

Remote Patient Monitoring and Machine Learning in Acute Exacerbations of Chronic Obstructive Pulmonary Disease: Dual Systematic Literature Review and Narrative Synthesis (e52143)
Henry Glyde, Caitlin Morgan, Tom Wilkinson, Ian Nabney, James Dodd. 1987

Detection of Sleep Apnea Using Wearable AI: Systematic Review and Meta-Analysis (e58187)
Alaa Abd-alrazaq, Hania Aslam, Rawan AlSaad, Mohammed Alsahli, Arfan Ahmed, Rafat Damseh, Sarah Aziz, Javaid Sheikh. 2004

XSL•FO
RenderX

Prompt Engineering Paradigms for Medical Applications: Scoping Review (e60501)
Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, Christian Lovis. 2037

Usability Evaluation Methods Used in Electronic Discharge Summaries: Literature Review (e55247)
Wubshet Tesfaye, Margaret Jordan, Timothy Chen, Ronald Castelino, Kamal Sud, Racha Dabliz, Parisa Aslani. 2059

Artificial Intelligence–Augmented Clinical Decision Support Systems for Pregnancy Care: Systematic Review (e54737)
Xinnian Lin, Chen Liang, Jihong Liu, Tianchu Lyu, Nadia Ghumman, Berry Campbell. 2071

Smartphone-Based Hand Function Assessment: Systematic Review (e51564)
Yan Fu, Yuxin Zhang, Bing Ye, Jessica Babineau, Yan Zhao, Zhengke Gao, Alex Mihailidis. 2089

Exploring Adaptive Virtual Reality Systems Used in Interventions for Children With Autism Spectrum Disorder: Systematic Review (e57093)
Luna Maddalon, Maria Minissi, Thomas Parsons, Amaia Hervas, Mariano Alcaniz. 2118

Investigating the Use of Telemedicine by Health Care Providers to Diagnose and Manage Patients With Musculoskeletal Disorders: Systematic Review and Meta-Analysis (e52964)
Raphaël Vincent, Maxime Charron, Simon Lafrance, Audrey-Anne Cormier, Dahlia Kairy, François Desmeules. 2134

User Engagement With mHealth Interventions to Promote Treatment Adherence and Self-Management in People With Chronic Health Conditions: Systematic Review (e50508)
Cyd Eaton, Natalie Vallejo, Xiomara McDonald, Jasmine Wu, Rosa Rodríguez, Nishanth Muthusamy, Nestoras Mathioudakis, Kristin Riekert.
2 1 5 2

Processing of Short-Form Content in Clinical Narratives: Systematic Scoping Review (e57852)
Amila Kugic, Ingrid Martin, Luise Modersohn, Peter Pallaoro, Markus Kreuzthaler, Stefan Schulz, Martin Boeker. 2184

Just-in-Time Adaptive Interventions for Behavior Change in Physiological Health Outcomes and the Use Case for Knee Osteoarthritis: Systematic Review (e54119)
Janis Fiedler, Matteo Bergmann, Stefan Sell, Alexander Woll, Bernd Stetter. 2203

Effectiveness of Internet-Based Telehealth Programs in Patients With Hip or Knee Osteoarthritis: Systematic Review and Meta-Analysis (e55576)
Hao-Nan Wang, Pei Luo, Shuyue Liu, Yunyi Liu, Xiao Zhang, Jian Li. 2225

Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review (e55315)
Hans Eguia, Carlos Sánchez-Bocanegra, Franco Vinciarelli, Fernando Alvarez-Lopez, Francesc Saigí-Rubió. 2244

Patient Safety Incident Reporting and Learning Guidelines Implemented by Health Care Professionals in Specialized Care Units: Scoping Review (e48580)
Tusiwe Gqaleni, Sipho Mkhize, Geldine Chironda. 2258

Considering User Experience and Behavioral Approaches in the Design of mHealth Interventions for Atrial Fibrillation: Systematic Review (e54405)
Sagar Suresh Kumar, Patricia Connolly, Anja Maier. 2266

Digital Interventions for Managing Medication and Health Care Service Delivery in West Africa: Systematic Review (e44294)
Emmanuel Oluokun, Festus Adedoyin, Huseyin Dogan, Nan Jiang. 2281

XSL•FO
RenderX

Exploring the Need for Medical Futures Studies: Insights From a Scoping Review of Health Care Foresight (e57148)
Bertalan Meskó, Tamás Kristóf, Pranavsingh Dhunnoo, Nóra Árvai, Gellért Katonai. 2295

Impact of Skin Pigmentation on Pulse Oximetry Blood Oxygenation and Wearable Pulse Rate Accuracy: Systematic Review and Meta-Analysis (e62769)
Sanidhya Singh, Miles Bennett, Chen Chen, Sooyoon Shin, Hamid Ghanbari, Benjamin Nelson. 2314

Spaced Digital Education for Health Professionals: Systematic Review and Meta-Analysis (e57760)
Laura Martinengo, Matthew Ng, Tony Ng, Yi-Ian Ang, Ahmad Jabir, Bhone Kyaw, Lorainne Tudor Car. 2335

Technology-Based Interventions in Tobacco Use Treatment Among People Who Identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native: Scoping Review (e50748)
Emily Hichborn, Avery Turner, Sarah Moore, Phoebe Gauthier, Kathleen Bell, LaTrice Montgomery, Jesse Boggis, Chantal Lambert-Harris, Elizabeth Saunders, Jesse Dallery, Bethany McLeman, Lisa Marsch. 2353

Quality Assessment of Digital Health Apps: Umbrella Review (e58616)
Maciej Zych, Raymond Bond, Maurice Mulvenna, Jorge Martinez Carracedo, Lu Bai, Simon Leigh. 2370

Co-Designing Remote Patient Monitoring Technologies for Inpatients: Systematic Review (e58144)
Jennifer Sumner, Si Tan, Yuchen Wang, Camille Keck, Eunice Xin Lee, Emily Chew, Alexander Yip. 2382

Exploring the Role of Mobile Apps for Insomnia in Depression: Systematic Review (e51110)
Yi-Hang Chiu, Yen-Fen Lee, Huang-Li Lin, Li-Chen Cheng. 2396

Technology-Supported Physical Activity and Its Potential as a Tool to Promote Young Women's Physical Activity and Physical Literacy: Systematic Review (e52302)
Kimberley Watson-Mackie, Lauren Arundell, Natalie Lander, Fiona McKay, Alethea Jerebine, Fotini Venetsanou, Lisa Barnett. 2412

Designing and Evaluating IT Applications for Informal Caregivers: Scoping Review (e57393)
Shweta Premanandan, Awais Ahmad, Åsa Cajander, Pär Ågerfalk, Lisette van Gemert-Pijnen. 2433

Acceptability of Digital Mental Health Interventions for Depression and Anxiety: Systematic Review (e52609)
Carrie Lau, Anthony Saad, Bettina Camara, Dia Rahman, Blanca Bolea-Alamanac. 2460

Implications of Big Data Analytics, AI, Machine Learning, and Deep Learning in the Health Care System of Bangladesh: Scoping Review (e54710)
Md Alam, Md Sajib, Fariya Rahman, Saraban Ether, Molly Hanson, Abu Sayeed, Ema Akter, Nowrin Nusrat, Tanjeena Islam, Sahar Raza, K Tanvir, Mohammod Chisti, Qazi Rahman, Akm Hossain, MA Layek, Asaduz Zaman, Juwel Rana, Syed Rahman, Shams Arifeen, Ahmed Rahman, Anisuddin Ahmed. 2477

Experiences of Patients With Cancer Using Electronic Symptom Management Systems: Qualitative Systematic Review and Meta-Synthesis (e59061)
Siying Zhu, Yan Dong, Yumei Li, Hong Wang, Xue Jiang, Mingen Guo, Tiantian Fan, Yalan Song, Ying Zhou, Yuan Han. 2504

Digital Mental Health Interventions for Adolescents in Low- and Middle-Income Countries: Scoping Review (e51376)
Carolina Wani, Lisa McCann, Marilyn Lennon, Caterina Radu. 2517

Provision of Digital Primary Health Care Services: Overview of Reviews (e53594)
Virgínia Fava, Luís Lapão. 2537

Health Care Professionals' Experience of Using AI: Systematic Review With Narrative Synthesis (e55766)
Abimbola Ayorinde, Daniel Mensah, Julia Walsh, Iman Ghosh, Siti Ibrahim, Jeffry Hogg, Niels Peek, Frances Griffiths. 2555

XSL•FO
RenderX

The Relation Between Passively Collected GPS Mobility Metrics and Depressive Symptoms: Systematic Review and Meta-Analysis (e51875)
Yannik Terhorst, Johannes Knauer, Paula Philippi, Harald Baumeister. 2574

The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis (e56532)
William Waldock, Joe Zhang, Ahmad Guni, Ahmad Nabeel, Ara Darzi, Hutan Ashrafian. 2588

Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review (e52738)
Jie Gong, Dandan Gu, Suyun Dong, Wangqin Shen, Haiou Yan, Juan Xie. 2603

Exploration of Features of Mobile Applications for Medication Adherence in Asia: Narrative Review (e60787)
Tzu Wang, Yen-Ming Huang, Hsun-Yu Chan. 2617

Participation in eHealth Communication Interventions Among Patients Undergoing Hemodialysis: Scoping Review (e51900)
Anne Deinboll, Cathrine Moe, Mette Ludvigsen. 2632

The Effects of Virtual Reality–Based Reminiscence Therapies for Older Adults With Cognitive Impairment: Systematic Review (e53348)
Qian Mao, Zhen Zhao, Lisha Yu, Yang Zhao, Hailiang Wang. 2652

Telehealth Use and Legal Considerations in Drug Health Services During Pandemics: Systematic Scoping Review (e46394)
Meryem Jefferies, Robert Graham, Marguerite Tracy, Scott Read, Mohammed Eslam, Mark Douglas, Jacob George. 2671

Machine Learning Methods to Personalize Persuasive Strategies in mHealth Interventions That Promote Physical Activity: Scoping Review and Categorization Overview (e47774)
Annette Brons, Shihan Wang, Bart Visser, Ben Kröse, Sander Bakkes, Remco Veltkamp. 2682

Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review (e51432)
Phawinpon Chotwanvirat, Aree Prachansuwan, Pimnapanut Sridonpai, Wantanee Kriengsinyos. 2702

Examining the Role of Large Language Models in Orthopedics: Systematic Review (e59607)
Cheng Zhang, Shanshan Liu, Xingyu Zhou, Siyu Zhou, Yinglun Tian, Shenglin Wang, Nanfang Xu, Weishi Li. 2725

Human Factors, Human-Centered Design, and Usability of Sensor-Based Digital Health Technologies: Scoping Review (e57628)
Animesh Tandon, Bryan Cobb, Jacob Centra, Elena Izmailova, Nikolay Manyakov, Samantha McClenahan, Smit Patel, Emre Sezgin, Srinivasan Vairavan, Bernard Vrijens, Jessie Bakker, Digital Health Measurement Collaborative Community (DATAcc) hosted by DiMe. 2745

Use of Digital Tools in Arbovirus Surveillance: Scoping Review (e57476)
Carolina Melo, Larissa Mageste, Lusiele Guaraldo, Daniela Paula, Mayumi Wakimoto. 2767

AI-Based Noninvasive Blood Glucose Monitoring: Scoping Review (e58892)
Pin Chan, Eric Jin, Miia Jansson, Han Chew. 2781

Effectiveness of Digital Health Interventions in Promoting Physical Activity Among College Students: Systematic Review and Meta-Analysis (e51714)
Siyuan Bi, Junfeng Yuan, Yanling Wang, Wenxin Zhang, Luqin Zhang, Yongjuan Zhang, Rui Zhu, Lin Luo. 2798

XSL•FO
RenderX

Cybersecurity Interventions in Health Care Organizations in Low- and Middle-Income Countries: Scoping Review (e47311)
Kaede Hasegawa, Niki O'Brien, Mabel Prendergast, Chris Ajah, Ana Neves, Saira Ghafur. 2817

The Uses and Experiences of Synchronous Communication Technology for Home-Dwelling Older Adults in a Home Care Services Context: Qualitative Systematic Review (e59285)
Martin Bavngaard, Anne Lund, Björg Thordardottir, Erik Rasmussen. 2832

Elements Influencing User Engagement in Social Media Posts on Lifestyle Risk Factors: Systematic Review (e59742)
Yan Yip, Mohd Makmor-Bakry, Wei Chong. 2851

Artificial Intelligence Applications to Measure Food and Nutrient Intakes: Scoping Review (e54557)
Jiakun Zheng, Junjie Wang, Jing Shen, Ruopeng An. 2868

Digital Interventions for Symptoms of Borderline Personality Disorder: Systematic Review and Meta-Analysis (e54941)
Julia Lindsay, Niall McGowan, Thomas Henning, Eli Harriss, Kate Saunders. 2882

Virtual Primary Care for People With Opioid Use Disorder: Scoping Review of Current Strategies, Benefits, and Challenges (e54015)
Shawna Narayan, Ellie Gooderham, Sarah Spencer, Rita McCracken, Lindsay Hedden. 2902

Effectiveness and Acceptability of Asynchronous Digital Health in Asthma Care: Mixed Methods Systematic Review (e57708)
Nazim Uzzaman, Victoria Hammersley, Kirstie McClatchey, Jessica Sheringham, Diksha Singh, GM Habib, Hilary Pinnock. 2932

Application of Chatbots to Help Patients Self-Manage Diabetes: Systematic Review and Meta-Analysis (e60380)
Yibo Wu, Jinzi Zhang, Pu Ge, Tingyu Duan, Junyu Zhou, Yiwei Wu, Yuening Zhang, Siyu Liu, Xinyi Liu, Erya Wan, Xinying Sun. 2949

Reviewing the Evidence Base for Topical Steroid Withdrawal Syndrome in the Research Literature and Social Media Platforms: An Evidence Gap Map (e57687)
Noreen Orr, Morwenna Rogers, Abigail Stein, Jo Thompson Coon, Kenneth Stein. 2963

Health Insurance Payment for Telehealth Services: Scoping Review and Narrative Synthesis (e56699)
Qingqing Li, Feng Cheng, Huatang Zeng, Junfang Xu. 2978

Psychological Factors That Contribute to the Use of Video Consultations in Health Care: Systematic Review (e54636)
Helen Haydon, James Fowler, Monica Taylor, Anthony Smith, Liam Caffery. 2991

Factors Associated With Digital Intervention Engagement and Adherence in Patients With Cancer: Systematic Review (e52542)
Lucile Montalescot, Louise Baussard, Elodie Charbonnier. 3009

Large Language Models and Empathy: Systematic Review (e52597)
Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, Eyal Klang. 3029

Current Implementation of Digital Health in Chronic Disease Management: Scoping Review (e53576)
Candelyn Pong, Rachel Tseng, Yih Tham, Elaine Lum. 3039

Effects of Prehabilitation With Advanced Technologies in Patients With Musculoskeletal Diseases Waiting for Surgery: Systematic Review and Meta-Analysis (e52943)
Stefania Guida, Jacopo Vitale, Eva Swinnen, David Beckwée, Silvia Bargeri, Federico Pennestrì, Greta Castellini, Silvia Gianola. 3072

XSL•FO
RenderX

Prevention and Health Promotion Interventions for Young People in the Context of Digital Well-Being: Rapid Systematic Review (e59968)
Michelle Colder Carras, Dahlia Aljuboori, Jing Shi, Mayank Date, Fatima Karkoub, Karla García Ortiz, Fasika Abreha, Johannes Thrul. 3085

When and Why Adults Abandon Lifestyle Behavior and Mental Health Mobile Apps: Scoping Review (e56897)
Patrick Kidman, Rachel Curtis, Amanda Watson, Carol Maher. 3109

Investigating eHealth Lifestyle Interventions for Vulnerable Pregnant Women: Scoping Review of Facilitators and Barriers (e54366)
Ashley Smit, Isra Al-Dhahir, Lieke Schiphof-Godart, Linda Breeman, Andrea Evers, Koen Joosten. 3121

Large Language Models in Gastroenterology: Systematic Review (e66648)
Eun Gong, Chang Bang, Jae Lee, Jonghyung Park, Eunsil Kim, Subeen Kim, Minjae Kimm, Seoung-Ho Choi. 3146

Tracking Upper Limb Motion via Wearable Solutions: Systematic Review of Research From 2011 to 2023 (e51994)
Eirini Karoulla, Maria Matsangidou, Fotos Frangoudes, Panayiotis Paspalides, Kleanthis Neokleous, Constantinos Pattichis. 3159

Characteristics and Outcomes of mHealth Interventions in Psychosis: Systematic Mapping Review (e55924)
Pei Loh, Laura Martinengo, Creighton Heaukulani, Xin Tan, Moses Hng, Yong Cheah, Robert Morris, Lorainne Tudor Car, Jimmy Lee. 3177

Machine Learning and Deep Learning for Diagnosis of Lumbar Spinal Stenosis: Systematic Review and Meta-Analysis (e54676)
Tianyi Wang, Ruiyuan Chen, Ning Fan, Lei Zang, Shuo Yuan, Peng Du, Qichao Wu, Aobo Wang, Jian Li, Xiaochuan Kong, Wenyi Zhu. 3193

Exploring the Applications of Explainability in Wearable Data Analytics: Systematic Literature Review (e53863)
Yasmin Abdelaal, Michaël Aupetit, Abdelkader Baggag, Dena Al-Thani. 3211

Consideration of Cybersecurity Risks in the Benefit-Risk Analysis of Medical Devices: Scoping Review (e65528)
Oscar Freyer, Fatemeh Jahed, Max Ostermann, Christian Rosenzweig, Pascal Werner, Stephen Gilbert. 3242

Platform-Based Patient-Clinician Digital Health Interventions for Care Transitions: Scoping Review (e55753)
Chantal Backman, Rosie Papp, Aurelie Tonjock Kolle, Steve Papp, Sarah Visintini, Ana Schaefer Ferreira de Mello, Gabriela de Melo Lanzoni, Anne Harley. 3266

Applying AI to Structured Real-World Data for Pharmacovigilance Purposes: Scoping Review (e57824)
Stella Dimitsaki, Pantelis Natsiavas, Marie-Christine Jaulent. 3299

Use of Extrinsic Motivators to Improve the BMI of Obese or Overweight Adolescents: Systematic Review (e57458)
Ana Gonçalves, Pedro Simões, Bernardo Sousa-Pinto, Tiago Taveira-Gomes. 3324

The Effectiveness of Video Animations as a Tool to Improve Health Information Recall for Patients: Systematic Review (e58306)
Steffen Hansen, Tue Jensen, Anne Schmidt, Janni Strøm, Peter Vistisen, Mette Høybye. 3343

Assessing the Content and Effect of Web-Based Decision Aids for Postmastectomy Breast Reconstruction: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e53872)
Lin Yu, Jianmei Gong, Xiaoting Sun, Min Zang, Lei Liu, Shengmiao Yu. 4411

XSL•FO
RenderX

The Influence of Physical Activity and Diet Mobile Apps on Cardiovascular Disease Risk Factors: Meta-Review (e51321)
Erica Bushey, Yin Wu, Alexander Wright, Linda Pescatello. 4820

Behavior Change Techniques Used in Self-Management Interventions Based on mHealth Apps for Adults With Hypertension: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e54978)
You Zhou, Si-Jia Li, Ren-Qian Huang, Hao-Ming Ma, Ao-Qi Wang, Xing-Yi Tang, Run-Yuan Pei, Mei-Hua Piao. 4918

Effectiveness and Feasibility of Digital Pulmonary Rehabilitation in Patients Undergoing Lung Cancer Surgery: Systematic Review and Meta-Analysis (e56795)
Taiping Lu, Ting Deng, Yangyang Long, Jin Li, Anmei Hu, Yufan Hu, Li Ouyang, Huiping Wang, Junliang Ma, Shaolin Chen, Jiale Hu. 4963

Investigating the Effectiveness of Technology-Based Distal Interventions for Postpartum Depression and Anxiety: Systematic Review and Meta-Analysis (e53236)
Sarah Brocklehurst, Alyssa Morse, Tegan Cruwys, Philip Batterham, Liana Leach, Alysia Robertson, Aseel Sahib, Colette Burke, Jessica Nguyen, Alison Calear. 5022

Usability and Effectiveness of eHealth and mHealth Interventions That Support Self-Management and Health Care Transition in Adolescents and Young Adults With Chronic Disease: Systematic Review (e56556)
ZhiRu Li, FangYan Lu, JingYun Wu, RuiJie Bao, YuXin Rao, Yun Yang, Huafen Wang. 5045

Biomarkers of Response to Internet-Based Psychological Interventions: Systematic Review (e55736)
Giulia Gotti, Chiara Gabelli, Sophia Russotto, Fabio Madeddu, Philippe Courtet, Jorge Lopez-Castroman, Patrizia Zeppegno, Carla Gramaglia, Raffaella Calati. 5124

Technology-Based Interventions for Substance Use Treatment Among People Who Identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native: Scoping Review (e53685)
Sarah Moore, Jesse Boggis, Phoebe Gauthier, Chantal Lambert-Harris, Emily Hichborn, Kathleen Bell, Elizabeth Saunders, LaTrice Montgomery, Eilis Murphy, Avery Turner, Nico Agosti, Bethany McLeman, Lisa Marsch. 5185

Regulatory Standards and Guidance for the Use of Health Apps for Self-Management in Sub-Saharan Africa: Scoping Review (e49163)
Benard Bene, Sunny Ibeneme, Kayode Fadahunsi, Bala Harri, Nkiruka Ukor, Nikolaos Mastellos, Azeem Majeed, Josip Car. 5453

Patients' Perspectives on the Data Confidentiality, Privacy, and Security of mHealth Apps: Systematic Review (e50715)
Nasser Alhammad, Mohannad Alajlani, Alaa Abd-alrazaq, Gregory Epiphaniou, Theodoros Arvanitis. 5589

Electronic Health Interventions and Cervical Cancer Screening: Systematic Review and Meta-Analysis (e58066)
Xiaoxia Liu, Lianzhen Ning, Wenqi Fan, Chanyi Jia, Lina Ge. 5928

Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis (e51250)
Jian Huo, Yan Yu, Wei Lin, Anmin Hu, Chaoran Wu. 7532

Evaluation of Telemedicine Consultations Using Health Outcomes and User Attitudes and Experiences: Scoping Review (e53266)
Pranavsingh Dhunnoo, Bridie Kemp, Karen McGuigan, Bertalan Meskó, Vicky O'Rourke, Michael McCann. 7660

Teledentistry Applied to Health and Education Outcomes: Evidence Gap Map (e60590)
Júlia Nascimento da Silva Mulder, Marcelo Ramos Pinto, Isabelle Aníbal, Ana Dornellas, Deise Garrido, Camila Huanca, Ana Haddad, Carmen Mendes Abdala. 7993

XSL•FO
RenderX

Patient and Public Willingness to Share Personal Health Data for Third-Party or Secondary Uses: Systematic Review (e50421)
Rebecca Baines, Sebastian Stevens, Daniela Austin, Krithika Anil, Hannah Bradwell, Leonie Cooper, Inocencio Maramba, Arunangsu Chatterjee, Simon Leigh. 8707

The Impact of Patient Access to Electronic Health Records on Health Care Engagement: Systematic Review (e56473)
Dalia Alomar, Maryam Almashmoum, Iliada Eleftheriou, Pauline Whelan, John Ainsworth. 8916

Instruments for Measuring Psychological Dimensions in Human-Robot Interaction: Systematic Review of Psychometric Properties (e55597)
Roberto Vagnetti, Nicola Camp, Matthew Story, Khaoula Ait-Belaid, Suvobrata Mitra, Massimiliano Zecca, Alessandro Di Nuovo, Daniele Magistro. 9013

The Acceptability, Engagement, and Feasibility of Mental Health Apps for Marginalized and Underserved Young People: Systematic Review and Qualitative Study (e48964)
Holly Bear, Lara Ayala Nunes, Giovanni Ramos, Tanya Manchanda, Blossom Fernandes, Sophia Chabursky, Sabine Walper, Edward Watkins, Mina Fazel. 10654

Effects of Immersive Technology–Based Education for Undergraduate Nursing Students: Systematic Review and Meta-Analysis Using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) Approach (e57566)
Subin Park, Hui Shin, Hyoeun Kwak, Hyun Lee. 10891

The Effectiveness of Therapeutic Exercise Interventions With Virtual Reality on Balance and Walking Among Persons With Chronic Stroke: Systematic Review, Meta-Analysis, and Meta-Regression of Randomized Controlled Trials (e59136)
Maria Krohn, Aki Rintala, Jaakko Immonen, Tuulikki Sjögren. 10908

Question Answering for Electronic Health Records: Scoping Review of Datasets and Models (e53636)
Jayetri Bardhan, Kirk Roberts, Daisy Wang. 11365

Economic Evaluations of Digital Health Interventions for Patients With Heart Failure: Systematic Review (e53500)
Neily Zakiyah, Dita Marulin, Mohammed Alfaqeeh, Irma Puspitasari, Keri Lestari, Ka Lim, Julia Fox-Rushby. 11540

AI Quality Standards in Health Care: Rapid Umbrella Review (e54705)
Craig Kuziemsky, Dillon Chrimes, Simon Minshall, Michael Mannerow, Francis Lau. 11910

Harnessing Artificial Intelligence to Predict Ovarian Stimulation Outcomes in In Vitro Fertilization: Scoping Review (e53396)
Rawan AlSaad, Alaa Abd-alrazaq, Fadi Choucair, Arfan Ahmed, Sarah Aziz, Javaid Sheikh. 11990

Accuracy of Machine Learning in Discriminating Kawasaki Disease and Other Febrile Illnesses: Systematic Review and Meta-Analysis (e57641)
Jinpu Zhu, Fushuang Yang, Yang Wang, Zhongtian Wang, Yao Xiao, Lie Wang, Liping Sun. 12330

Implementation of Machine Learning Applications in Health Care Organizations: Systematic Review of Empirical Studies (e55897)
Luigi Preti, Vittoria Ardito, Amelia Compagni, Francesco Petracca, Giulia Cappellaro. 12349

Accuracy of Machine Learning in Detecting Pediatric Epileptic Seizures: Systematic Review and Meta-Analysis (e55986)
Zhuan Zou, Bin Chen, Dongqiong Xiao, Fajuan Tang, Xihong Li. 12436

XSL•FO
RenderX

Identification of Ethical Issues and Practice Recommendations Regarding the Use of Robotic Coaching Solutions for Older Adults: Narrative Review (e48126)
Cécilia Palmier, Anne-Sophie Rigaud, Toshimi Ogawa, Rainer Wieching, Sébastien Dacunha, Federico Barbarossa, Vera Stara, Roberta Bevilacqua, Maribel Pino. 12600

Contextual Barriers to Implementing Open-Source Electronic Health Record Systems for Low- and Lower-Middle-Income Countries: Scoping Review (e45242)
Sarah Bostan, Owen Johnson, Lena Jaspersen, Rebecca Randell. 12700

Success Factors of Growth-Stage Digital Health Companies: Systematic Literature Review (e60473)
Estelle Pfitzer, Laura Bitomsky, Marcia Nißen, Christoph Kausch, Tobias Kowatsch. 13186

Self-Management Using eHealth Technologies for Liver Transplant Recipients: Scoping Review (e56664)
Soo Kim, Kyoung-A Kim, Sang Chu, Hyunji Kim, Dong Joo, Jae Lee, JiYeon Choi. 13411

Definitions and Characteristics of Patient Digital Twins Being Developed for Clinical Use: Scoping Review (e58504)
David Drummond, Apolline Gonsard. 13489

Patient Organizations' Digital Responses to the COVID-19 Pandemic: Scoping Review (e58566)
Simon Wallraf, Marie-Luise Dierks, Cosima John, Jonas Lander. 13798

General Characteristics and Design Taxonomy of Chatbots for COVID-19: Systematic Review (e43112)
Wendell Lim, Razel Custodio, Monica Sunga, Abegail Amoranto, Raymond Sarmiento. 13884

Innovative Approaches to Menstruation and Fertility Tracking Using Wearable Reproductive Health Technology: Systematic Review (e45139)
Lynnette Lyzwinski, Mohamed Elgendi, Carlo Menon. 13972

Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis (e60807)
Mingxin Liu, Tsuyoshi Okuhara, XinYi Chang, Ritsuko Shirabe, Yuriko Nishiie, Hiroko Okada, Takahiro Kiuchi. 14275

Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review (e22769)
Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, Zhijun Yin. 14415

Impact of 25 Years of Mobile Health Tools for Pain Management in Patients With Chronic Musculoskeletal Pain: Systematic Review (e59358)
Jenny Shi, Regina Sit. 14613

The Unintended Consequences of Telehealth in Australia: Critical Interpretive Synthesis (e57848)
Sagda Osman, Kate Churruca, Louise Ellis, Dan Luo, Jeffrey Braithwaite. 14628

A 25-Year Retrospective of the Use of AI for Diagnosing Acute Stroke: Systematic Review (e59711)
Zhaoxin Wang, Wenwen Yang, Zhengyu Li, Ze Rong, Xing Wang, Jincong Han, Lei Ma. 14715

Data Analytics in Physical Activity Studies With Accelerometers: Scoping Review (e59497)
Ya-Ting Liang, Charlotte Wang, Chuhsing Hsiao. 14741

The Journey From Nonimmersive to Immersive Multiuser Applications in Mental Health Care: Systematic Review (e60441)
Iveta Fajnerova, Lukáš Hejtmánek, Michal Sedlák, Markéta Jablonská, Anna Francová, Pavla Stopková. 14933

XSL•FO
RenderX

AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges (e59225)
David Owen, Amy Lynham, Sophie Smart, Antonio Pardiñas, Jose Camacho Collados. 14968

Telemonitoring for Chronic Heart Failure: Narrative Review of the 20-Year Journey From Concept to Standard Care in Germany (e63391)
Sebastian Spethmann, Gerhard Hindricks, Kerstin Koehler, Stefan Stoerk, Christiane Angermann, Michael Böhm, Birgit Assmus, Sebastian Winkler, Martin Möckel, Mirja Mittermaier, Monika Lelgemann, Daniel Reuter, Ralph Bosch, Alexander Albrecht, Stephan von Haehling, Thomas Helms, Stefan Sack, Tarek Bekfani, Jan Gröschel, Magdalena Koehler, Christoph Melzer, Jan Wintrich, Bettina Zippel-Schultz, Georg Ertl, Claus Vogelmeier, Nikolaos Dagres, Jasmin Zernikow, Friedrich Koehler. 15042

Looking Back on Digital Medical Education Over the Last 25 Years and Looking to the Future: Narrative Review (e60312)
Oluwadamilola Ogundiya, Thahmina Rahman, Ioan Valnarov-Boulter, Tim Young. 15066

Short Papers

Limitations of the Cough Sound-Based COVID-19 Diagnosis Artificial Intelligence Model and its Future Direction: Longitudinal Observation Study (e51640)
Jina Kim, Yong Choi, Young Lee, Seung Yeo, Kyung Kim, Min Kim, Masoud Rahmati, Dong Yon, Jinseok Lee. 3353

Patterns of Ownership and Usage of Wearable Devices in the United States, 2020-2022: Survey Study (e56504)
Ashwini Nagappan, Adriana Krasniansky, Madelyn Knowles. 3359

Investigating How the Use of Technology Can Reduce Missed Appointments: Quantitative Case Study at a General Practitioner Surgery (e43894)
Teresa Sides, Dhouha Kbaier. 3368

Retention in HIV Primary Care Using a Web-Based Patient Engagement Platform: Multistate Case-Control Study (e64092)
Adam Sukhija-Cohen, Henna Patani, Michael Blasingame, Kathy Vu, Ramin Bastani. 3390

Accuracy, Quality, and Misinformation of YouTube Abortion Procedural Videos: Cross-Sectional Study (e50099)
Nicole Acero, Emma Herrero, Juanita Foncham, Jamie McIlvaine, Emre Kayaalp, Melissa Figueora, Antonia Oladipo. 3397

Use of ChatGPT to Explore Gender and Geographic Disparities in Scientific Peer Review (e57667)
Paul Sebo. 3403

Virtual Clinic Telehealth Abortion Services in the United States One Year After Dobbs: Landscape Review (e50749)
Leah Koenig, Jennifer Ko, Ushma Upadhyay. 7718

Equivalence of Alcohol Use Disorder Symptom Assessments in Routine Clinical Care When Completed Remotely via Online Patient Portals Versus In Clinic via Paper Questionnaires: Psychometric Evaluation (e52101)
Theresa Matson, Amy Lee, Malia Oliver, Katharine Bradley, Kevin Hallgren. 8132

XSL•FO
RenderX

Commentary

Commentary on “Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line” (e42144)
Timothy Reierson. 3409

Viewpoints

Public Health Dashboards in Overdose Prevention: The Rhode Island Approach to Public Health Data Literacy, Partnerships, and Action (e51671)
Jesse Yedinak, Maxwell Krieger, Raynald Joseph, Stacey Levin, Sarah Edwards, Dennis Bailer, Jonathan Goyer, Colleen Daley Ndoye, Cathy Schultz, Jennifer Koziol, Rachael Elmaleh, Benjamin Hallowell, Todd Hampson, Ellen Duong, Abdullah Shihipar, William Goedel, Brandon Marshall. 3454

Extended Reality—New Opportunity for People With Disability? Practical and Ethical Considerations (e41670)
Karen Stendal, Rosemarie Bernabe. 3468

Building Pandemic-Resilient Primary Care Systems: Lessons Learned From COVID-19 (e47667)
Yejin Jeong, Trevor Crowell, Anna Devon-Sand, Theadora Sakata, Amelia Sattler, Shreya Shah, Timothy Tsai, Steven Lin. 3473

Improvement of Cancer Prevention and Control: Reflection on the Role of Emerging Information Technologies (e50000)
Pei Dong, Ayan Mao, Wuqi Qiu, Guanglin Li. 3478

Promises, Pitfalls, and Clinical Applications of Artificial Intelligence in Pediatrics (e49022)
Hansa Bhargava, Carmela Salomon, Srinivasan Suresh, Anthony Chang, Rachel Kilian, Diana Stijn, Albert Oriol, Daniel Low, Ashley Knebel, Sharief Taraman. 3488

The Digital Therapeutics Real-World Evidence Framework: An Approach for Guiding Evidence-Based Digital Therapeutics Design, Development, Testing, and Monitoring (e49208)
Meelim Kim, Kevin Patrick, Camille Nebeker, Job Godino, Spencer Stein, Predrag Klasnja, Olga Perski, Clare Viglione, Aaron Coleman, Eric Hekler. 3496

Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges (e53008)
Yan Chen, Pouyan Esmaeilzadeh. 3513

Mitigating Health-Related Uncertainties During Pregnancy: The Role of Smart Health Monitoring Technologies (e48493)
Max Tretter. 3532

The Political Economy of Digital Health Equity: Structural Analysis (e46971)
James Shaw, Wiljeana Glover. 3538

Digital Alcohol Interventions Could Be Part of the Societal Response to Harmful Consumption, but We Know Little About Their Long-Term Costs and Health Outcomes (e44574)
Katarina Ulfsdotter Gunnarsson, Martin Henriksson, Marcus Bendtsen. 3550

The Challenges in Using eHealth Decision Resources for Surrogate Decision-Making in the Intensive Care Unit (e47017)
Wan-Na Sun, Chi-Yin Kao. 3556

A Perspective on Crowdsourcing and Human-in-the-Loop Workflows in Precision Health (e51138)
Peter Washington. 3565

XSL•FO
RenderX

The Effect of Everyday-Life Social Contact on Pain (e53830)
Martin Weiß, Marthe Gründahl, Annalena Jachnik, Emilia Lampe, Ishitaa Malik, Heike Rittner, Claudia Sommer, Grit Hein. 3573

Envisioning the Future of Personalized Medicine: Role and Realities of Digital Twins (e50204)
Alexandre Vallée. 3585

The Power of Rapid Reviews for Bridging the Knowledge-to-Action Gap in Evidence-Based Virtual Health Care (e54821)
Megan MacPherson, Sarah Rourke. 3598

Efficient Use of Biological Data in the Web 3.0 Era by Applying Nonfungible Token Technology (e46160)
Guanyi Wang, Chen Chen, Ziyu Jiang, Gang Li, Can Wu, Sheng Li. 3607

An Infrastructure Framework for Remote Patient Monitoring Interventions and Research (e51234)
Jennifer Claggett, Stacie Petter, Amol Joshi, Todd Ponzio, Eric Kirkendall. 3616

Harnessing ChatGPT for Thematic Analysis: Are We Ready? (e54974)
V Lee, Stephanie van der Lubbe, Lay Goh, Jose Valderas. 3630

A Taxonomy for Health Information Systems (e47682)
Anna Janssen, Candice Donnelly, Tim Shaw. 3641

Does an App a Day Keep the Doctor Away? AI Symptom Checker Applications, Entrenched Bias, and Professional Responsibility (e50344)
Ma'n Zawati, Michael Lang. 3660

Trust but Verify: Lessons Learned for the Application of AI to Case-Based Clinical Decision-Making From Postmarketing Drug Safety Assessment at the US Food and Drug Administration (e50274)
Robert Ball, Andrew Talal, Oanh Dang, Monica Muñoz, Marianthi Markatou. 3669

Key Considerations for Designing Clinical Studies to Evaluate Digital Health Solutions (e54518)
Elaina Bolinger, Benoit Tyl. 3683

Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses (e56780)
Xufei Luo, Fengxian Chen, Di Zhu, Ling Wang, Zijun Wang, Hui Liu, Meng Lyu, Ye Wang, Qi Wang, Yaolong Chen. 3691

A Nordic Perspective on Patient Online Record Access and the European Health Data Space (e49084)
Maria Hägglund, Anna Kharko, Annika Bärkås, Charlotte Blease, Åsa Cajander, Catherine DesRoches, Asbjørn Fagerlund, Josefin Hagström, Isto Huvila, Iiris Hörhammer, Bridget Kane, Gunnar Klein, Eli Kristiansen, Jonas Moll, Irene Muli, Hanife Rexhepi, Sara Riggare, Peeter Ross, Isabella Scandurra, Saija Simola, Hedvig Soone, Bo Wang, Maedeh Ghorbanian Zolbin, Rose-Mharie Åhlfeldt, Sari Kujala, Monika Johansen.
3 7 0 1

Patient Portals Fail to Collect Structured Information About Who Else is Involved in a Person's Care (e49394)
Liz Salmi, Danielle Peereboom, David Dorr, Leilani Graham, Jennifer Wolff, Catherine DesRoches. 3712

Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support (e50295)
Carlos Sáez, Pablo Ferri, Juan García-Gómez. 3718

Enhancing Agency in Posttraumatic Stress Disorder Therapies Through Sensorimotor Technologies (e58390)
Vladimir Adrien, Nicolas Bosc, Claire Peccia Galletto, Thomas Diot, Damien Claverie, Nicco Reggente, Marion Trousselard, Eric Bui, Thierry Baubet, Félix Schoeller. 3732

XSL•FO
RenderX

Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions (e47645)
Yu-Qing Cai, Da-Xin Gong, Li-Ying Tang, Yue Cai, Hui-Jun Li, Tian-Ci Jing, Mengchun Gong, Wei Hu, Zhen-Wei Zhang, Xingang Zhang, Guang-Wei Zhang. 3753

Expanding a Behavioral View on Digital Health Access: Drivers and Strategies to Promote Equity (e51355)
Maura Kepper, Lauren Fowler, Isabelle Kusters, Jean Davis, Manal Baqer, Sara Sagui-Henson, Yunyu Xiao, Adati Tarfa, Jean Yi, Bryan Gibson, Kristin Heron, Nicole Alberts, Marissa Burgermaster, Veronica Njie-Carr, Lisa Klesges. 3774

5G Key Technologies for Helicopter Aviation Medical Rescue (e50355)
Wei Han Sr, Yuanting Li 2nd, Changgen Chen 3rd, Danni Huang, Junchao Wang, Xiang Li, Zhongliang Ji, Qin Li, Zhuang Li. 3790

Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint (e60083)
Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, Ren Wei. 3803

Evaluating Artificial Intelligence in Clinical Settings—Let Us Not Reinvent the Wheel (e46407)
Kathrin Cresswell, Nicolette de Keizer, Farah Magrabi, Robin Williams, Michael Rigby, Mirela Prgomet, Polina Kukhareva, Zoie Wong, Philip Scott, Catherine Craven, Andrew Georgiou, Stephanie Medlock, Jytte Brender McNair, Elske Ammenwerth. 3816

Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies (e57896)
Tianhui Xu, Huiting Weng, Fang Liu, Li Yang, Yuanyuan Luo, Ziwei Ding, Qin Wang. 3825

The Digital Platform and Its Emerging Role in Decentralized Clinical Trials (e47882)
Rachel Copland, Sten Hanke, Amy Rogers, Lampros Mpaltadoros, Ioulietta Lazarou, Alexandra Zeltsi, Spiros Nikolopoulos, Thomas MacDonald, Isla Mackenzie. 3838

Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook (e59505)
Rawan AlSaad, Alaa Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, Javaid Sheikh. 3851

Engine of Innovation in Hospital Pharmacy: Applications and Reflections of ChatGPT (e51635)
Xingang Li, Heng Guo, Dandan Li, Yingming Zheng. 3875

Applying Implementation Science to Advance Electronic Health Record–Driven Learning Health Systems: Case Studies, Challenges, and Recommendations (e55472)
Katy Trinkley, Anna Maw, Cristina Torres, Amy Huebschmann, Russell Glasgow. 3885

Using Social Media to Recruit Participants in Health Care Research: Case Study (e51751)
Amy Wright, Ysabella Willett, Era Ferron, Vithusa Kumarasamy, Sarah Lem, Ossaid Ahmed. 3902

Enhancing Research Involvement of Young People With Lived Expertise: Reflecting on Experiences in Digital Mental Health Research (e55441)
Josephine Brogden, Zsofi de Haan, Carla Gorban, Samuel Hockey, Alexis Hutcheon, Frank Iorfino, Yun Song, Elizabeth Scott, Ian Hickie, Sarah McKenna. 3914

Video Abstracts in Research (e64221)
Sophie Nachman, Esteban Ortiz-Prado, Joseph Tucker. 3923

Economics and Equity of Large Language Models: Health Care Perspective (e64226)
Radha Nagarajan, Midori Kondo, Franz Salas, Emre Sezgin, Yuan Yao, Vanessa Klotzman, Sandip Godambe, Naqi Khan, Alfonso Limon, Graham Stephenson, Sharief Taraman, Nephi Walton, Louis Ehwerhemuepha, Jay Pandit, Deepti Pandita, Michael Weiss, Charles Golden, Adam Gold, John Henderson, Angela Shippy, Leo Celi, William Hogan, Eric Oermann, Terence Sanger, Steven Martel. 3927

XSL•FO
RenderX

The Challenges and Lessons Learned Building a New UK Infrastructure for Finding and Accessing Population-Wide COVID-19 Data for Research and Public Health Analysis: The CO-CONNECT Project (e50235)
Emily Jefferson, Gordon Milligan, Jenny Johnston, Shahzad Mumtaz, Christian Cole, Joseph Best, Thomas Giles, Samuel Cox, Erum Masood, Scott Horban, Esmond Urwin, Jillian Beggs, Antony Chuter, Gerry Reilly, Andrew Morris, David Seymour, Susan Hopkins, Aziz Sheikh, Philip Quinlan. 3953

Exploring Participants' Experiences of Digital Health Interventions With Qualitative Methods: Guidance for Researchers (e62761)
Kristin Harrison Ginsberg, Katie Babbott, Anna Serlachius. 3964

A Cross-Disciplinary Analysis of the Complexities of Scaling Up eHealth Innovation (e58007)
Sanne Allers, Chiara Carboni, Frank Eijkenaar, Rik Wehrens. 3971

Digital Information Ecosystems in Modern Care Coordination and Patient Care Pathways and the Challenges and Opportunities for AI Solutions (e60258)
You Chen, Christoph Lehmann, Bradley Malin. 3984

Reflections From the Pandemic: Is Connectivism the Panacea for Clinicians? (e53344)
Jennifer Benjamin, Tyson Pillow, Heather MacNeill, Ken Masters, Anoop Agrawal, Neil Mehta. 3999

Unlocking the Potential for Implementation of Equitable, Digitally Enabled Citizen Science: Multidisciplinary Digital Health Perspective (e50491)
Lucio Naccarella, Jonathan Rawstorn, Jaimon Kelly, Eleanor Quested, Stuart Jenkinson, Dominika Kwasnicka. 4008

From Doubt to Confidence—Overcoming Fraudulent Submissions by Bots and Other Takers of a Web-Based Survey (e60184)
Jeffrey Hardesty, Elizabeth Crespi, Joshua Sinamo, Qinghua Nian, Alison Breland, Thomas Eissenberg, Ryan Kennedy, Joanna Cohen. . . .
4 0 1 6

Remote Passive Sensing of Older Adults' Activities and Function: User-Centered Design Considerations for Behavioral Interventions Conducted in the Home Setting (e54709)
Lyndsey Miller, Jeffrey Kaye, Allison Lindauer, Wan-Tai Au-Yeung, Nathaniel Rodrigues, Sara Czaja. 14055

Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals (e56764)
Avishek Choudhury, Zaira Chaudhry. 14221

The Role of Assistive Technology in Enabling Older Adults to Achieve Independent Living: Past and Future (e58846)
Anna Sweeting, Katie Warncken, Martyn Patel. 14546

The McMaster Health Information Research Unit: Over a Quarter-Century of Health Informatics Supporting Evidence-Based Medicine (e58764)
Cynthia Lokker, K McKibbon, Muhammad Afzal, Tamara Navarro, Lori-Ann Linkins, R Haynes, Alfonso Iorio. 14555

Beyond Step Count: Are We Ready to Use Digital Phenotyping to Make Actionable Individual Predictions in Psychiatry? (e59826)
Abigail Ortiz, Benoit Mulsant. 14563

A Quarter-Century of Online Informatics Education: Learners Served and Lessons Learned (e59066)
William Hersh. 14569

XSL•FO
RenderX

The Second Life Metaverse and Its Usefulness in Medical Education After a Quarter of a Century (e59005)
Francisco Sendra-Portero, Rocío Lorenzo-Álvarez, Teodoro Rudolphi-Solero, Miguel Ruiz-Gómez. 14579

Deceptively Simple yet Profoundly Impactful: Text Messaging Interventions to Support Health (e58726)
Brian Suffoletto. 14653

Research Into Digital Health Intervention for Mental Health: 25-Year Retrospective on the Ethical and Legal Challenges (e58939)
Charlotte Hall, Aislinn Gómez Bergin, Stefan Rennick-Egglestone. 14702

From Fax to Secure File Transfer Protocol: The 25-Year Evolution of Real-Time Syndromic Surveillance in England (e58704)
Alex Elliot, Helen Hughes, Sally Harcourt, Sue Smith, Paul Loveridge, Roger Morbey, Amardeep Bains, Obaghe Edeghere, Natalia Jones, Daniel Todkill, Gillian Smith. 14782

We Have Spent Time, Money, and Effort Making Self-Help Digital Mental Health Interventions: Is Anyone Going to Come to the Party? (e58198)
Skye Fitzpatrick, Alexander Crenshaw, Victoria Donkin, Alexis Collins, Angela Xiang, Elizabeth Earle, Kamya Goenka, Sonya Varma, Julianne Bushe, Tara McFadden, Andrea Librado, Candice Monson. 14792

Equity in Digital Mental Health Interventions in the United States: Where to Next? (e59939)
Athena Robinson, Megan Flom, Valerie Forman-Hoffman, Trina Histon, Monique Levy, Alison Darcy, Toluwalase Ajayi, David Mohr, Paul Wicks, Carolyn Greene, Robert Montgomery. 14816

Evolutionary Trends in the Adoption, Adaptation, and Abandonment of Mobile Health Technologies: Viewpoint Based on 25 Years of Research (e62790)
Jennifer Portz, Susan Moore, Sheana Bull. 14830

Hospital Is Not the Home: Lessons From Implementing Remote Technology to Support Acute Inpatient and Transitional Care in the Home in the United States and United Kingdom (e58888)
Matt Wilkes, Annabel Kramer, Juliana Pugmire, Christopher Pilkington, Benjamin Zaniello, Nicole Zahradka. 14844

Primary Care Informatics: Vitalizing the Bedrock of Health Care (e60081)
Jacqueline You, Tiffany Leung, Deepti Pandita, Matthew Sakumoto. 14855

Transitioning Perspectives in Digital Health Through Phenomenology Integration (e62691)
Maddalena Fiordelli. 14887

The Value of Smartwatches in the Health Care Sector for Monitoring, Nudging, and Predicting: Viewpoint on 25 Years of Research (e58936)
Charlotte Köhler, Alexander Bartschke, Daniel Fürstenau, Thorsten Schaaf, Eduardo Salgado-Baez. 14895

A 25-Year Retrospective of Health IT Infrastructure Building: The Example of the Catalonia Region (e58933)
Jordi Piera-Jiménez, Gerard Carot-Sans, Marina Ramiro-Pareta, Maria Nogueras, Júlia Folguera-Profitós, Pepi Ródenas, Alba Jiménez-Rueda, Thais de Pando Navarro, Josep Mira Palacios, Joan Fajardo, Joan Ustrell Campillo, Emili Vela, David Monterde, Damià Valero-Bover, Tara Bonet, Guillermo Tarrasó-Urios, Roser Cantenys-Sabà, Pau Fabregat-Fabregat, Beatriz Gómez Oliveros, Jesús Berdún, Xabier Michelena, Isaac Cano, Rubèn González-Colom, Josep Roca, Oscar Solans, Caridad Pontes, Pol Pérez-Sust. 14994

Perspectives of Digital Health Innovations in Low- and Middle-Income Health Care Systems From South and Southeast Asia (e57612)
Siyan Yi, Esabelle Yam, Kochukoshy Cheruvettolil, Eleni Linos, Anshika Gupta, Latha Palaniappan, Nitya Rajeshuni, Kiran Vaska, Kevin Schulman, Karen Eggleston. 15016

Ten Myths About the Effect of Social Media Use on Well-Being (e59585)
Jeffrey Hall. 15027

XSL•FO
RenderX

Don't Forget the Humble Text Message: 25 Years of Text Messaging in Health (e59888)
Rosie Dobson, Robyn Whittaker, Lorien Abroms, Dale Bramley, Caroline Free, Hayden McRobbie, Melanie Stowell, Anthony Rodgers. 15058

Decades in the Making: The Evolution of Digital Health Research Infrastructure Through Synthetic Data, Common Data Models, and Federated Learning (e58637)
Jodie Austin, Elton Lobo, Mahnaz Samadbeik, Teyl Engstrom, Reji Philip, Jason Pole, Clair Sullivan. 15087

Tutorials

Selecting an Ecological Momentary Assessment Platform: Tutorial for Researchers (e51125)
Lauren Henry, Eleanor Hansen, Justin Chimoff, Kimberly Pokstis, Miryam Kiderman, Reut Naim, Joe Kossowsky, Meghan Byrne, Silvia Lopez-Guzman, Katharina Kircanski, Daniel Pine, Melissa Brotman. 4028

Developing a Chatbot to Support Individuals With Neurodevelopmental Disorders: Tutorial (e50182)
Ashwani Singla, Ritvik Khanna, Manpreet Kaur, Karen Kelm, Osmar Zaiane, Cory Rosenfelt, Truong Bui, Navid Rezaei, David Nicholas, Marek Reformat, Annette Majnemer, Tatiana Ogourtsova, Francois Bolduc. 4045

Machine Learning and Health Science Research: Tutorial (e50890)
Hunyong Cho, Jane She, Daniel De Marchi, Helal El-Zaatari, Edward Barnes, Anna Kahkoska, Michael Kosorok, Arti Virkud. 4064

Mental Wellness Self-Care in Singapore With mindline.sg: A Tutorial on the Development of a Digital Mental Health Platform for Behavior Change (e44443)
Janice Weng, Yanyan Hu, Creighton Heaukulani, Clarence Tan, Julian Chang, Ye Phang, Priyanka Rajendram, Weng Tan, Wai Loke, Robert Morris. 4079

Transforming Digital Phenotyping Raw Data Into Actionable Biomarkers, Quality Metrics, and Data Visualizations Using Cortex Software Package: Tutorial (e58502)
James Burns, Kelly Chen, Matthew Flathers, Danielle Currey, Natalia Macrynikola, Aditya Vaidyam, Carsten Langholm, Ian Barnett, Andrew Byun, Erlend Lane, John Torous. 4094

Adapting a Telehealth Physical Activity and Diet Intervention to a Co-Designed Website for Self-Management After Stroke: Tutorial (e58419)
Dina Pogrebnoy, Lee Ashton, Brian Beh, Meredith Burke, Richard Cullen, Jude Czerenkowski, Julie Davey, Amy Dennett, Kevin English, Erin Godecke, Nicole Harper, Elizabeth Lynch, Lesley MacDonald-Wicks, Amanda Patterson, Emily Ramage, Ben Schelfhaut, Dawn Simpson, Karly Zacharia, Coralie English. 4121

Reporting Guidelines for the Early-Phase Clinical Evaluation of Applications Using Extended Reality: RATE-XR Qualitative Study Guideline (e56790)
Johan Vlake, Denzel Drop, Jasper Van Bommel, Giuseppe Riva, Brenda Wiederhold, Pietro Cipresso, Albert Rizzo, Barbara Rothbaum, Cristina Botella, Lotty Hooft, Oscar Bienvenu, Christian Jung, Bart Geerts, Evert-Jan Wils, Diederik Gommers, Michel van Genderen, RATE-XR Expert Group. 13154

Research Letters

The Triage and Diagnostic Accuracy of Frontier Large Language Models: Updated Comparison to Physician Performance (e67409)
Michael Sorich, Arduino Mangoni, Stephen Bacchi, Bradley Menz, Ashley Hopkins. 12402

Where Do Oncology Patients Seek and Share Health Information? Survey Study (e36441)
Eric Freeman, Darshilmukesh Patel, Folasade Odeniyi, Mary Pasquinelli, Shikha Jain. 13902

XSL•FO
RenderX

Using Longitudinal Twitter Data for Digital Epidemiology of Childhood Health Outcomes: An Annotated Data Set and Deep Neural Network Classifiers (e50652)
Ari Klein, José Gutiérrez Gómez, Lisa Levine, Graciela Gonzalez-Hernandez. 13907

Response of Unvaccinated US Adults to Official Information About the Pause in Use of the Johnson & Johnson–Janssen COVID-19 Vaccine: Cross-Sectional Survey Study (e41559)
Vishala Mishra, Joseph Dexter. 13911

Evaluation of Prompts to Simplify Cardiovascular Disease Information Generated Using a Large Language Model: Cross-Sectional Study (e55388)
Vishala Mishra, Ashish Sarraju, Neil Kalwani, Joseph Dexter. 13918

Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties (e54948)
Felix Busch, Tianyu Han, Marcus Makowski, Daniel Truhn, Keno Bressem, Lisa Adams. 13924

Using Large Language Models to Support Content Analysis: A Case Study of ChatGPT for Adverse Event Detection (e52499)
Eric Leas, John Ayers, Nimit Desai, Mark Dredze, Michael Hogarth, Davey Smith. 13928

Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases (e53724)
Wan Chiu, Wei Ko, William Cho, Sin Hui, Wing Chan, Michael Kuo. 13933

Multimodal ChatGPT-4V for Electrocardiogram Interpretation: Promise and Limitations (e54607)
Lingxuan Zhu, Weiming Mou, Keren Wu, Yancheng Lai, Anqi Lin, Tao Yang, Jian Zhang, Peng Luo. 13937

Benchmarking State-of-the-Art Large Language Models for Migraine Patient Education: Performance Comparison of Responses to Common Queries (e55927)
Linger Li, Pengfei Li, Kun Wang, Liang Zhang, Hongwei Ji, Hongqin Zhao. 13941

Quality and Accountability of ChatGPT in Health Care in Low- and Middle-Income Countries: Simulated Patient Study (e56121)
Yafei Si, Yuyi Yang, Xi Wang, Jiaqi Zu, Xi Chen, Xiaojing Fan, Ruopeng An, Sen Gong. 13946

Internet Search Activity for Intentional Self-Harm Forums After a High-Profile News Publication: Interrupted Time Series Analysis (e52354)
Nora Kelsall, Catherine Gimbrone, Mark Olfson, Madelyn Gould, Jeffrey Shaman, Katherine Keyes. 13951

Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials (e60695)
Tony Hung, Gilad Kuperman, Eric Sherman, Alan Ho, Chunhua Weng, David Pfister, Jun Mao. 13955

Gender Bias in AI's Perception of Cardiovascular Risk (e54242)
Margaux Achtari, Adil Salihu, Olivier Muller, Emmanuel Abbé, Carole Clair, Joëlle Schwarz, Stephane Fournier. 13961

Preliminary Screening for Hereditary Breast and Ovarian Cancer Using an AI Chatbot as a Genetic Counselor: Clinical Study (e48914)
Ann Sato, Eri Haneda, Yukihiko Hiroshima, Hiroto Narimatsu. 13964

Advancing the United Nations Sustainable Development Goals Through Digital Health Research: 25 Years of Contributions From the Journal of Medical Internet Research (e60025)
Raghu Raman, Monica Singhania, Prema Nedungadi. 14928

XSL•FO
RenderX

Letters to the Editor

Concerns About the Generalizability Associated With a South African Randomized Controlled Trial on Prenatal Mothers (e53861)
Yongjian Lin. 13532

Authors' Reply: Concerns About the Generalizability Associated With a South African Randomized Controlled Trial on Prenatal Mothers (e55930)
Maya Adam, V n Nguy n. 13534

Ambiguity in Statistical Analysis Methods and Nonconformity With Prespecified Commitment to Data Sharing in a Cluster Randomized Controlled Trial (e54090)
Yasaman Jamshidi-Naeini, Lilian Golzarri-Arroyo, Deependra Thapa, Andrew Brown, Daniel Kpormegbey, David Allison. 13536

Authors' Reply: Ambiguity in Statistical Analysis Methods and Nonconformity With Prespecified Commitment to Data Sharing in a Cluster Randomized Controlled Trial (e57422)
Diego Arguello. 13539

Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications (e56997)
Kuan-Ju Huang. 13541

Authors' Reply: "Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications" (e57778)
Anne Herrmann-Werner, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, Moritz Mahling. 13543

How to Enhance Adherence in Patients with Nonalcoholic Fatty Liver Disease: A Thought-Provoking Issue (e59208)
Chunlan Liu, Run Zhou, Junping Shi. 13545

Mental Health Problems Among Children and Adolescents From a Sports Sociology Perspective (e60513)
Yuan Li, Qun Zhai, Weihang Peng. 13547

Author's Reply: Mental Health Problems Among Children and Adolescents From a Sports Sociology Perspective (e62775)
Grace Aldridge. 13549

The Power of Collaboration in Facilitating Mobile Technology Adoption in Health Care (e60350)
Weijing Fan, Guobin Liu. 13551

Authors' Reply: The Power of Collaboration in Facilitating Mobile Technology Adoption in Health Care (e62891)
Huong Tong, Severin Rakic, Hazzaa Al-Hazzaa, Saleh Alqahtani. 13553

Considerations and Challenges in the Application of Large Language Models for Patient Complaint Resolution (e65527)
Bin Wei, Xin Hu, XiaoRong Wu. 13555

Expanding the Scope: Reflections on Digital Smoking Cessation Strategies for Diverse Age Groups (e65929)
Bin Wei, Xin Hu, XiaoRong Wu. 13557

Author's Reply: Expanding the Scope: Reflections on Digital Smoking Cessation Strategies for Diverse Age Groups (e67749)
Margaret Fahey. 13559

XSL•FO
RenderX

The Vast Potential of ChatGPT in Pediatric Surgery (e66453)
Ran Tang, Shi-qin Qi. 13561

Reassessing AI in Medicine: Exploring the Capabilities of AI in Academic Abstract Synthesis (e55920)
Zijian Wang, Chunyang Zhou. 13563

Authors' Reply: Reassessing AI in Medicine: Exploring the Capabilities of AI in Academic Abstract Synthesis (e65123)
Tien-Wei Hsu, Chih-Sung Liang. 13565

Corrigenda and Addendas

Correction: Social Media Use and Its Concurrent and Subsequent Relation to a Biological Marker of Inflammation: Short-Term Longitudinal Study (e56808)
David Lee, Tao Jiang, Jennifer Crocker, Baldwin Way. 13567

Correction: Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media (e57233)
Feipeng Guo, Zixiang Liu, Qibei Lu, Shaobo Ji, Chen Zhang. 13569

Correction: Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis (e59628)
Jian Huo, Yan Yu, Wei Lin, Anmin Hu, Chaoran Wu. 13571

Correction: Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial (e60137)
Hailey Miller, John Gallis, Miriam Berger, Sandy Askew, Joseph Egger, Melissa Kay, Eric Finkelstein, Mia de Leon, Abigail DeVries, Ashley Brewer, Marni Holder, Gary Bennett. 13573

Correction: Habit Strength, Medication Adherence, and Habit-Based Mobile Health Interventions Across Chronic Medical Conditions: Systematic Review (e63980)
Sherif Badawy, Richa Shah, Usman Beg, Mallorie Heneghan. 13575

Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties (e64411)
Felix Busch, Tianyu Han, Marcus Makowski, Daniel Truhn, Keno Bressem, Lisa Adams. 13577

Correction: Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis (e66590)
Nuša Fari , Henry Potts, James Heilman. 13579

Correction: Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study (e67994)
Angela Karellis, Duncan Webster, Jean Boulanger, Kate Harland, Paige Feltmate, Stefanie Materniak, Gabriel Daunais-Laurin, Christine Mesa, Olivia Vaikla, John Kim, Nitika Pant Pai. 13581

Correction: Quality Assessment of Digital Health Apps: Umbrella Review (e67849)
Maciej Zych, Raymond Bond, Maurice Mulvenna, Jorge Martinez Carracedo, Lu Bai, Simon Leigh. 13584

Correction: Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review (e68783)
Jie Gong, Dandan Gu, Suyun Dong, Wangqin Shen, Haiou Yan, Juan Xie. 13585

XSL•FO
RenderX

Correction: Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study (e69042)
Minseok Hong, Ri-Ra Kang, Jeong Yang, Sang Rhee, Hyunju Lee, Yong-gyom Kim, KangYoon Lee, HongGi Kim, Yu Lee, Tak Youn, Se Kim, Yong Ahn. 13586

Correction: Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line (e67880)
JMIR Editorial Office. 13588

Correction: Electronic Health Record–Nested Reminders for Serum Lithium Level Monitoring in Patients With Mood Disorder: Randomized Controlled Trial (e55805)
Tomotsugu Seki, Morio Aki, Toshi Furukawa, Hirotsugu Kawashima, Tomotaka Miki, Yujin Sawaki, Takaaki Ando, Kentaro Katsuragi, Takahiko Kawashima, Senkei Ueno, Takashi Miyagi, Shun'ichi Noma, Shiro Tanaka, Koji Kawakami. 13591

Correction: Creatively Adapting Touch-Based Practices to the Web Format During the COVID-19 Pandemic: Systematic Review (e55713)
Greta Gauhe, Rosemary Kostic Cisneros, Jade Ward, David Hohenschurz-Schmidt. 13593

Correction: Correlation Between Remote Symptom Reporting by Caregivers and Adverse Clinical Outcomes: Mixed Methods Study (e56368)
Ingrid Oakley-Girvan, Reem Yunis, Stephanie Fonda, Michelle Longmire, Tess Veuthey, Jennifer Shieh, Sara Aghaee, Ai Kubo, Sharon Davis, Raymond Liu, Elad Neeman. 13595

Table Correction: Effectiveness of eHealth Smoking Cessation Interventions: Systematic Review and Meta-Analysis (e56438)
Yichen Fang, Zhixian Zhang, Ray Wang, Bolu Yang, Chen Chen, Claudia Nisa, Xin Tong, Lijing Yan. 13597

Correction: Exploring the Effect of the Dynamics of Behavioral Phenotypes on Health Outcomes in an mHealth Intervention for Childhood Obesity: Longitudinal Observational Study (e57013)
Sarah Woo, Sunho Jung, Hyunjung Lim, YoonMyung Kim, Kyung Park. 13599

Correction: Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial (e57499)
Oh Kwon, Mi Lee, Hye Lee, Hyerang Kim, Jae Lee, Yeonsoo Jang. 13602

Correction: Internet Tool to Support Self-Assessment and Self-Swabbing of Sore Throat: Development and Feasibility Study (e59420)
Mark Lown, Kirsten Smith, Ingrid Muller, Catherine Woods, Emma Maund, Kirsty Rogers, Taeko Becque, Gail Hayward, Michael Moore, Paul Little, Margaret Glogowska, Alastair Hay, Beth Stuart, Efi Mantzourani, Christopher Wilcox, Natalie Thompson, Nick Francis. 13604

Correction: Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation (e59984)
Hyo-Jeong Ahn, Eue-Keun Choi, So-Ryoung Lee, Soonil Kwon, Hee-Seok Song, Young-Shin Lee, Seil Oh. 13606

Correction: Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study (e60444)
Amy Thomas, Andrea Asnes, Kyle Libby, Allen Hsiao, Gunjan Tiyyagura. 13608

Correction: Experiences of Using Digital Mindfulness-Based Interventions: Rapid Scoping Review and Thematic Synthesis (e60967)
Emma Osborne, Ben Ainsworth, Nic Hooper, Melissa Atkinson. 13610

XSL•FO
RenderX

Correction: Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation (e62767)
Han Chew, Nicholas Chew, Shaun Loong, Su Lim, Wai Tam, Yip Chin, Ariana Chao, Georgios Dimitriadis, Yujia Gao, Jimmy So, Asim Shabbir, Kee Ngiam. 13612

Author Correction: Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study (e65977)
Dianna Vidas, Romina Carrasco, Ryan Kelly, Jenny Waycott, Jeanette Tamplin, Kate McMahon, Libby Flynn, Phoebe Stretton-Smith, Tanara Vieira Sousa, Felicity Baker. 13614

Correction: Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study (e67135)
Yun Kim, Won-Doo Seo, Sun Lee, Ja Koo, Gyung Kim, Hee Song, Minji Lee. 13616

Correction: Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study (e66813)
Kelly Quinn, Sarah Leiser Ransom, Carrie O'Connell, Naoko Muramatsu, David Marquez, Jessie Chin. 13619

Correction: Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey (e68439)
Yuka Sugawara, Yosuke Hirakawa, Masao Iwagami, Haruo Kuroki, Shuhei Mitani, Ataru Inagaki, Hiroki Ohashi, Mitsuru Kubota, Soichi Koike, Rie Wakimizu, Masaomi Nangaku. 13621

Correction: Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers (e69714)
Anna Jolliff, Richard Holden, Rupa Valdez, Ryan Coller, Himalaya Patel, Matthew Zuraw, Anna Linden, Aaron Ganci, Christian Elliott, Nicole Werner. 13623

XSL•FO
RenderX

Editorial

Redefining Virtual Assistants in Health Care: The Future With Large Language Models

Emre Sezgin[1,2], PhD

[1]The Abigail Wexner Reseach Institute at Nationwide Children's Hospital, Columbus, OH, United States
[2]The Ohio State University College of Medicine, Columbus, OH, United States

Corresponding Author:
Emre Sezgin, PhD
The Abigail Wexner Reseach Institute at Nationwide Children's Hospital
700 Children's Drive
Columbus, OH, 43205
United States
Phone: 1 6147223179
Email: emre.sezgin@nationwidechildrens.org

Abstract

This editorial explores the evolving and transformative role of large language models (LLMs) in enhancing the capabilities of virtual assistants (VAs) in the health care domain, highlighting recent research on the performance of VAs and LLMs in health care information sharing. Focusing on recent research, this editorial unveils the marked improvement in the accuracy and clinical relevance of responses from LLMs, such as GPT-4, compared to current VAs, especially in addressing complex health care inquiries, like those related to postpartum depression. The improved accuracy and clinical relevance with LLMs mark a paradigm shift in digital health tools and VAs. Furthermore, such LLM applications have the potential to dynamically adapt and be integrated into existing VA platforms, offering cost-effective, scalable, and inclusive solutions. These suggest a significant increase in the applicable range of VA applications, as well as the increased value, risk, and impact in health care, moving toward more personalized digital health ecosystems. However, alongside these advancements, it is necessary to develop and adhere to ethical guidelines, regulatory frameworks, governance principles, and privacy and safety measures. We need a robust interdisciplinary collaboration to navigate the complexities of safely and effectively integrating LLMs into health care applications, ensuring that these emerging technologies align with the diverse needs and ethical considerations of the health care domain.

(J Med Internet Res 2024;26:e53225) doi:10.2196/53225

KEYWORDS

large language models; voice assistants; virtual assistants; chatbots; conversational agents; health care

Virtual assistants (VAs)—mostly voice assistants, chatbots, and dialogue-based interactive applications—have been leading conversational technologies, being used for health care communications and remote monitoring [1-4]. However, their accuracy and reliability in understanding and responding to medical questions have been limitations [5], which have been slowly improving over the years [6]. Large language models (LLMs) offer scalable and customizable solutions for these limitations of VAs. The body of literature demonstrating the capabilities of LLMs in medicine and health care has been growing, and a number of studies benchmarking LLMs' performance against each other or humans' medical knowledge, decision-making processes, and empathic responses have been published [7-11]. Furthermore, LLM-based services improve equitable access to information and reduce language barriers via contextual and culturally aware systems [12] and privacy-preserving, local solutions for low-resource settings [13-16]. This research and evidence give a glimpse at the future of personalized VAs.

In the context of mental health and health information–seeking behavior, we investigated the performance of VAs in responding to postpartum depression–related frequently asked questions. The evidence from our two studies (conducted in 2021 with VAs [17] and 2023 with LLMs [18]) provides comparable findings on the 2-year difference in technology; illuminates the evolving roles of artificial intelligence, natural language processing, and LLMs in health care; and shows the promise of a more accurate and reliable digital health landscape.

In our first study in 2021 [17], we investigated the clinical accuracy of Google Assistant, Amazon Alexa, Microsoft Cortana, and Apple Siri voice assistant responses to postpartum depression–related questions. In our second study in 2023 [18], we replicated our research by using LLMs, and the new study

XSL•FO
RenderX

showed significant improvements in the accuracy and clinical relevance of responses. Specifically, GPT-4's responses to all postpartum depression–related questions were more accurate and clinically relevant in contrast to the VAs, of which the proportion of clinically relevant responses did not exceed the 29% reported in our earlier study. In addition, the interrater reliability score for LLMs (GPT-4: $\kappa=1$; $P<.05$) was higher than that for VAs ($\kappa=0.87$; $P<.001$), underscoring LLMs' consistency in clinically relevant responses, which is vital to achieve for health care applications. LLMs also recommended consultations with health care providers—a feature that adds an extra layer for safety—which was not observed in our earlier study with VAs. This dramatic improvement suggests a paradigm shift in the capabilities of digital health tools for health information–seeking activities. The high clinical accuracy and reliability of LLMs point toward a promising future for their integration into existing VA platforms. LLMs can offer dynamic adaptability for VAs via custom applications and decentralized LLM architectures [19,20]. Given their capabilities for open-source developments and collaborations, LLMs could serve as cost-effective and inclusive frameworks for collaborative developments (ie, among technology providers, patients, and clinical experts) in fine-tuning and training VAs for specific medical purposes.

The empirical data from our studies, as well as the literature [21], indicate a compelling trajectory toward LLMs being used to potentially improve the clinical and instructional capabilities of conversational technologies. This suggests a shift in our earlier spectrum model for VAs in health care (Figure 1) [22], in which we proposed 4 service levels for a spectrum of VA use that were associated with the risk, value, and impact of VAs. These levels were the "information" (eg, asking Amazon Alexa to start self-care guidance), "assistance" (eg, setting up reminders for medication or self-therapy), "assessment" (eg, identification, detection, prediction with digital biomarkers, and management), and "support" (prescribing, substituting, or supplementing medication and therapy tools) levels. In 2020, the evidence on the utilization of VAs in health care indicated that VAs were at the "information" and "assistance" levels [22]. However, LLMs are opening up opportunities for VAs, potentially toward the "assessment" and "support" levels. As Figure 1 shows, the level of a service and the associated risk, value, and impact of the service can change based on the targeted problems and solutions. A digital health ecosystem represents the ecosystem where we may envision a future of support from VAs enhanced by LLMs with speech interaction and audio-based sensing capabilities [23]. Such enhancements may include quantifying human behavior–related factors beyond VA engagement, such as social engagement, emotions, neurodevelopmental and behavioral health, sleep health (snoring, heart rate, and movements), respiratory symptoms (sneezing and coughing), and motion (gait, exercise, and sedentary behavior).

Figure 1. Spectrum of virtual assistants (outlines the risk, value, and impact in health care services) and applications in digital health ecosystems. These can change based on the targeted problems and solutions. This figure was created with BioRender.com (BioRender).

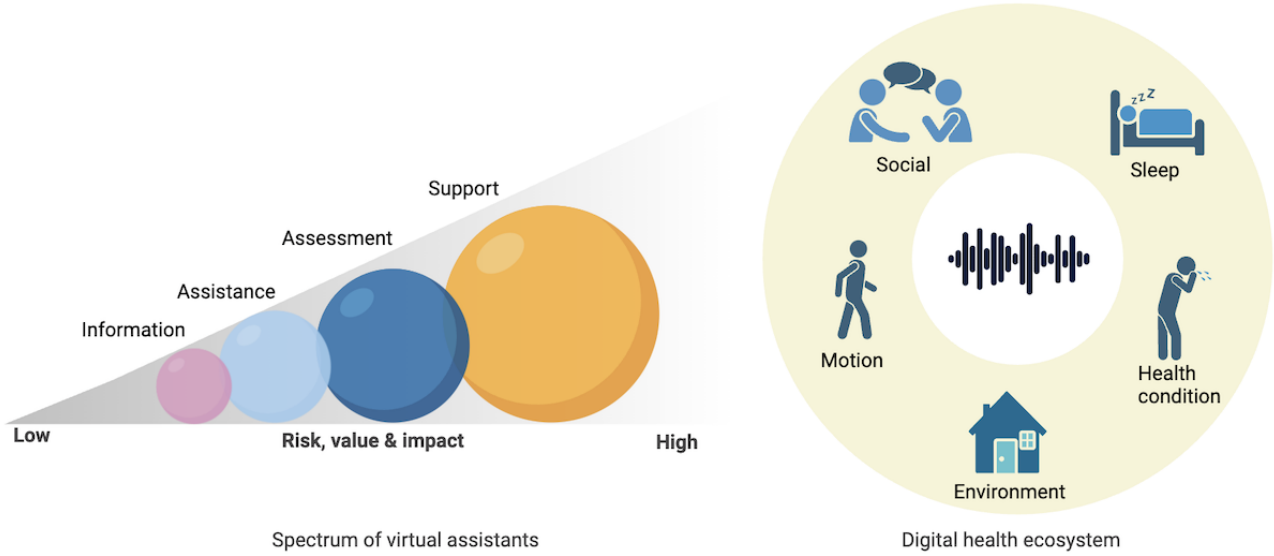

Despite this promising horizon, we need to approach cautiously. As LLMs are becoming highly appealing tools that can be used as VAs in health care, it is imperative to establish a platform that facilitates democratized access and interdisciplinary collaboration during the development of such applications [19,24]. This platform should be designed to bring together a diverse range of stakeholders, including technologists, ethicists, researchers, health care professionals, and patients. This would ensure that the development and integration of VAs are guided by a balanced perspective that considers ethical guidelines and regulatory oversight [25,26], governance principles [27,28], privacy and safety measures [29], feasibility, efficacy, and patient-centric approaches and assessment methods [30,31]. By prioritizing such collaborative and inclusive dialogues, we can better navigate the complex challenges and harness the full potential of these advanced technologies in health care, ensuring that they are developed responsibly, ethically, and in alignment with the diverse needs of all users.

XSL•FO
RenderX

Acknowledgments

The author thanks Yasemin Sezgin for her constructive feedback. Figure 1 was created with BioRender.com (BioRender).

Conflicts of Interest

ES serves on the editorial board of JMIR Publications.

References

1. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. NPJ Digit Med 2020 Sep 16;3:122 [FREE Full text] [doi: 10.1038/s41746-020-00332-0] [Medline: 33015374]
2. Corbett CF, Combs EM, Chandarana PS, Stringfellow I, Worthy K, Nguyen T, et al. Medication adherence reminder system for virtual home assistants: mixed methods evaluation study. JMIR Form Res 2021 Jul 13;5(7):e27327 [FREE Full text] [doi: 10.2196/27327] [Medline: 34255669]
3. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021 Nov 29;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
4. Sawad AB, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, et al. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. Sensors (Basel) 2022 Mar 29;22(7):2625 [FREE Full text] [doi: 10.3390/s22072625] [Medline: 35408238]
5. Palanica A, Thommandram A, Lee A, Li M, Fossat Y. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. NPJ Digit Med 2019 Jun 20;2:55 [FREE Full text] [doi: 10.1038/s41746-019-0133-x] [Medline: 31304401]
6. Palanica A, Fossat Y. Medication name comprehension of intelligent virtual assistants: a comparison of Amazon Alexa, Google Assistant, and Apple Siri between 2019 and 2021. Front Digit Health 2021 May 19;3:669971 [FREE Full text] [doi: 10.3389/fdgth.2021.669971] [Medline: 34713143]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the Applied Knowledge Test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
9. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659 [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
10. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023 Jan 23;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
11. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
12. Yao B, Jiang M, Yang D, Hu J. Empowering LLM-based machine translation with cultural awareness. arXiv Preprint posted online on May 23, 2023. [FREE Full text]
13. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. From text to tables: a local privacy preserving large language model for structured information retrieval from medical documents. medRxiv Preprint posted online on Dec 8, 2023. [FREE Full text] [doi: 10.1101/2023.12.07.23299648]
14. Cai W. Feasibility and prospect of privacy-preserving large language models in radiology. Radiology 2023 Oct;309(1):e232335. [doi: 10.1148/radiol.232335] [Medline: 37815443]
15. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Science 2023 Jul 24;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61]
16. Ogueji K, Zhu Y, Lin J. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In: Ataman D, Birch A, Conneau A, Firat O, Ruder S, Sahin GG, editors. Proceedings of the 1st Workshop on Multilingual Representation Learning. Kerrville, TX: Association for Computational Linguistics; 2021:116-126.
17. Yang S, Lee J, Sezgin E, Bridge J, Lin S. Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. JMIR Mhealth Uhealth 2021 Jan 11;9(1):e24045 [FREE Full text] [doi: 10.2196/24045] [Medline: 33427680]
18. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023 Sep 11;25:e49240 [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]
19. Xu B, Liu X, Shen H, Han Z, Li Y, Yue M, et al. Gentopia: a collaborative platform for tool-augmented LLMs. arXiv Preprint posted online on Aug 8, 2023. [FREE Full text] [doi: 10.18653/v1/2023.emnlp-demo.20]

XSL•FO
RenderX

20. Chen C, Feng X, Zhou J, Yin J, Zheng X. Federated large language model: a position paper. arXiv Preprint posted online on Jul 18, 2023. [FREE Full text]
21. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
22. Sezgin E, Militello LK, Huang Y, Lin S. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. Transl Behav Med 2020 Aug 7;10(3):606-628. [doi: 10.1093/tbm/ibz141] [Medline: 32766865]
23. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health 2022 Dec 22;1(12):e0000168 [FREE Full text] [doi: 10.1371/journal.pdig.0000168] [Medline: 36812634]
24. AMA issues new principles for AI development, deployment and use. American Medical Association. 2023 Nov 28. URL: https://www.ama-assn.org/press-center/press-releases/ama-issues-new-principles-ai-development-deployment-use [accessed 2023-12-25]
25. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
26. Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery; Jun 12, 2023:1112-1123.
27. Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. AI Ethics 2023 May 30 [FREE Full text] [doi: 10.1007/s43681-023-00289-2]
28. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of clinical AI applications to facilitate safe and equitable deployment in a large health system: key elements and early successes. Front Digit Health 2022 Aug 24;4:931439 [FREE Full text] [doi: 10.3389/fdgth.2022.931439] [Medline: 36093386]
29. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023 Apr 25;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
30. Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. J Am Med Inform Assoc 2023 Dec 09 Epub ahead of print. [doi: 10.1093/jamia/ocad222] [Medline: 38070173]
31. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. arXiv Preprint posted online on Dec 29, 2023. [FREE Full text]

Abbreviations

LLM: large language model
VA: virtual assistant

Edited by T Leung, T de Azevedo Cardoso; submitted 29.09.23; this is a non–peer-reviewed article;accepted 02.01.24; published 19.01.24.

Please cite as:
Sezgin E
Redefining Virtual Assistants in Health Care: The Future With Large Language Models
J Med Internet Res 2024;26:e53225
URL: https://www.jmir.org/2024/1/e53225
doi:10.2196/53225
PMID:38241074

©Emre Sezgin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Editorial

Chatbots and COVID-19: Taking Stock of the Lessons Learned

Virginia Arnold[1], MA, MBA; Tina D Purnat[2,3], BA, MSc; Robert Marten[4], MPH, MPP, PhD; Andrew Pattison[5], MSc; Hebe Gouda[6], MPH, PhD

[1]Department of Social Determinants of Health, Division of UHC Healthier Populations, World Health Organization, Geneva, Switzerland
[2]Department of Epidemic and Pandemic Preparedness and Prevention, Emergencies Programme, World Health Organization, Geneva, Switzerland
[3]University of Memphis School of Public Health, Memphis, TN, United States
[4]Alliance for Health Policy and Systems Research, Science Division, World Health Organization, Geneva, Switzerland
[5]Department of Digital Health and Innovation, Science Division, World Health Organization, Geneva, Switzerland
[6]Department of Health Promotion, Division of UHC Healthier Populations, World Health Organization, Geneva, Switzerland

Corresponding Author:
Hebe Gouda, MPH, PhD
Department of Health Promotion
Division of UHC Healthier Populations
World Health Organization
Avenue Appia 20
Geneva, 1211
Switzerland
Phone: 41 793865070
Email: goudah@who.int

Abstract

While digital innovation in health was already rapidly evolving, the COVID-19 pandemic has accelerated the generation of digital technology tools, such as chatbots, to help increase access to crucial health information and services to those who were cut off or had limited contact with health services. This theme issue titled “Chatbots and COVID-19” presents articles from researchers and practitioners across the globe, describing the development, implementation, and evaluation of chatbots designed to address a wide range of health concerns and services. In this editorial, we present some of the key challenges and lessons learned arising from the content of this theme issue. Most notably, we note that a stronger evidence base is needed to ensure that chatbots and other digital tools are developed to best serve the needs of population health.

(J Med Internet Res 2024;26:e54840) doi:10.2196/54840

KEYWORDS

chatbots; COVID-19; health; public health; pandemic; health care

Introduction

Like many areas of technology, digital innovation in health is evolving rapidly [1]. It is frequently used to expand access to health services for populations whose options are otherwise restricted due to geographical, economic, or sociocultural barriers. Digital technology has been increasingly used in health care and public health to help disseminate important health-related information to the public as well as health care workers, and to support health services by scheduling medical appointments, for example, or even helping practitioners with the diagnosis of health problems and monitoring health outcomes. During 2020 to 2023, the COVID-19 pandemic led to increased use of digital technology as countries limited in-person access to health care facilities, paused or slowed immunization and other health campaigns, and reallocated health resources to the COVID-19 response. Suddenly, access to health care and essential health services was drastically reduced at a time when health care needs and concerns were increasing.

As a result, there was a tremendous surge of interest in digital technology, including chatbots—computer programs that present a conversation-like interface through which people can access information and services [2]. During the early stages of the COVID-19 pandemic, national and subnational governments, as well as international health organizations such as the World Health Organization (WHO) [3,4] and the United States Centers for Disease Control and Prevention (CDC) [5], private companies, nongovernmental organizations, and researchers explored these new channels to reach people who were otherwise cut off from health services.

Chatbots have been designed to provide a range of health-related services, from information sharing [6] to offering support for

XSL•FO
RenderX

mental well-being [7,8], and diagnostic tools to assist health care workers [9]. They have been designed for a range of subgroups such as specific age-groups (eg, youth who are often considered the most likely group to accept and adopt mobile technology), people with specific diseases (eg, young people with diabetes [10]), or those in health care roles (eg, frontline health workers [11]).

While this theme issue in the *Journal of Medical Internet Research* titled "Chatbots and COVID-19" [12] focuses on chatbots related to COVID-19, the span of topics covered by these chatbots is broad. They range from vaccine hesitancy and equity in access to relevant information [13-15] to pandemic-related mental health [16-19]. A number of topics focus on health promotion interventions, such as smoking cessation [20], as well as encouraging physical activity [21] and tooth-brushing [22]. Although the topics are diverse, some common themes have emerged from these explorations, namely, the importance of considering user experience in chatbot design, the need to make chatbots inclusive, and the importance of user trust [23].

The perception of chatbots in these studies was generally positive: user experiences demonstrate that the chatbots are seen as valuable tools for managing personal health [24-26] and are considered useful in health care and clinical settings [27]. However, there was a clear paucity of robust evidence on the impact these chatbots have on health outcomes. Chatbots are a medium for delivering content, but the quality, relevance, and impact of the content itself is as essential. Not all chatbots are equal, and potential efficiency gains should not be assumed to deliver automatically. Some chatbots may potentially do more harm than good. There is an urgent need for more implementation research on best-practice monitoring and evaluation of chatbots in terms of their contribution to health, as well as their strengths and weaknesses compared to other mediums when used in different contexts or with different populations.

It is also important to not lose sight of the audience. In some cases, focusing on a chatbot at the expense of alternative approaches may lead to the marginalization of certain population groups. There are inherent access barriers created by chatbots. Most notably, of course, is the access to the mobile technology itself, which is not equally distributed globally. Although mobile phones have become ubiquitous globally, mobile data can be unreliable and expensive in many contexts and the exclusion of certain groups due to age or language restrictions, for example, can be a major challenge [28]. Nonetheless, chatbots can also be a valuable tool for increasing access by overcoming literacy barriers for some populations and some globally focused chatbots have already been developed in multiple languages. Pahola (a chatbot designed to help people reduce their alcohol consumption [3]) and Florence (a chatbot designed to help people with multitude of health concerns, including smoking cessation [4]), for example, "speak" 4 and 7 different languages, respectively.

There is reason to be optimistic about chatbots and new digital tools for health. They hold the potential to remove access barriers and engage marginalized populations. Over the past 2 decades, there has been some *hype* around these innovations, however, and excessive confidence about the potential benefits of chatbots despite a lack of evidence on their quantifiable impact on health outcomes. Furthermore, as recent debates on the role of artificial intelligence (AI) showcase, technology can trigger unintended consequences to both its users and wider society. Although chatbots are not all the same as AI—indeed, many in this series are often closer to traditional SMS text messaging than to a machine learning model—a similar awareness of context and opportunity costs is essential for their healthy development. To fully understand the benefits of health chatbots, providers must work closely with a broad spectrum of perspectives, including local communities, social scientists, anthropologists, and ethicists, to ensure that their development addresses growing concerns about digital harms [2,29]. Without adequate ethical guidance, transparency, and accountability of developers and products, there is a risk of harming users through intentional or unintentional discrimination, marginalization, or lack of privacy [29].

In May 2023, WHO officially declared the end of the COVID-19 international health emergency. Nevertheless, its consequences for health systems and digital health development will continue to reverberate for some time. A wider interest in health chatbots may have started as a stopgap, but their use and value have not ended with the pandemic. However, chatbots must adhere to the same evidence-based standards as any new tool in a public health or clinical setting. Without this robust foundation, chatbots will remain tools that are at best a nice-to-have, and at worst only considered when other options are unavailable.

Authors' Contributions

VA conceived the project. VA and HG drafted the editorial. All authors reviewed and edited the manuscript.

Conflicts of Interest

All authors acted as editors of the special theme issue "Chatbots and COVID-19" in the *Journal of Medical Internet Research* at the time of this publication.

References

1. Luengo-Oroz M, Hoffmann Pham K, Bullock J, Kirkpatrick R, Luccioni A, Rubel S, et al. Artificial intelligence cooperation to support the global response to COVID-19. Nat Mach Intell 2020 May 22;2(6):295-297. [doi: 10.1038/s42256-020-0184-3]
2. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR Hum Factors 2022 Oct 05;9(4):e35882 [FREE Full text] [doi: 10.2196/35882] [Medline: 36197708]

XSL•FO
RenderX

3. Monteiro MG, Pantani D, Pinsky I, Hernandes Rocha TA. Using the Pan American Health Organization digital conversational agent to educate the public on alcohol use and health: preliminary analysis. JMIR Form Res 2023 Apr 06;7:e43165 [FREE Full text] [doi: 10.2196/43165] [Medline: 36961920]
4. Loveys K, Lloyd E, Sagar M, Broadbent E. Development of a virtual human for supporting tobacco cessation during the COVID-19 pandemic. J Med Internet Res 2023 Dec 05;25:e42310 [FREE Full text] [doi: 10.2196/42310] [Medline: 38051571]
5. Shah A, Oyegun E, Hampton W, Neri A, Maddox N, Raso D, et al. Engagement with the Centers for Disease Control and Prevention coronavirus self-checker and guidance provided to users in the United States from March 23, 2020, to April 19, 2021: thematic and trend analysis. J Med Internet Res 2023 Mar 10;25:e39054 [FREE Full text] [doi: 10.2196/39054] [Medline: 36745776]
6. Nehme M, Schneider F, Perrin A, Sum Yu W, Schmitt S, Violot G, et al. The development of a chatbot technology to disseminate post-COVID-19 information: descriptive implementation study. J Med Internet Res 2023 Jun 05;25:e43113 [FREE Full text] [doi: 10.2196/43113] [Medline: 37195688]
7. Islam A, Chaudhry BM. Design validation of a relational agent by COVID-19 patients: mixed methods study. JMIR Hum Factors 2023 Jun 08;10:e42740 [FREE Full text] [doi: 10.2196/42740] [Medline: 36350760]
8. Wang H, Lin T, Yao Y, Hsu C, Yang C, Chen TH, et al. Community-based digital contact tracing of emerging infectious diseases: design and implementation study with empirical COVID-19 cases. J Med Internet Res 2023 Nov 08;25:e47219 [FREE Full text] [doi: 10.2196/47219] [Medline: 37938887]
9. Han J, Montagna M, Grammenos A, Xia T, Bondareva E, Siegele-Brown C, et al. Evaluating listening performance for COVID-19 detection by clinicians and machine learning: comparative study. J Med Internet Res 2023 May 9;25:e44804. [doi: 10.2196/44804]
10. Boggiss A, Consedine N, Hopkins S, Silvester C, Jefferies C, Hofman P, et al. Improving the well-being of adolescents with Type 1 Diabetes during the COVID-19 pandemic: qualitative study exploring acceptability and clinical usability of a self-compassion chatbot. JMIR Diabetes 2023 May 05;8:e40641 [FREE Full text] [doi: 10.2196/40641] [Medline: 36939680]
11. Jackson-Triche M, Vetal D, Turner E, Dahiya P, Mangurian C. Meeting the behavioral health needs of health care workers during COVID-19 by leveraging chatbot technology: development and usability study. J Med Internet Res 2023 Jun 08;25:e40635 [FREE Full text] [doi: 10.2196/40635] [Medline: 37146178]
12. Theme Issue 2022: Chatbots and COVID-19. Journal of Medical Internet Research.: JMIR Publications URL: https://www.jmir.org/themes/1164-theme-issue-2022-chatbots-and-covid19 [accessed 2023-12-28]
13. Perez-Ramos JG, Leon-Thomas M, Smith SL, Silverman L, Perez-Torres C, Hall WC, et al. COVID-19 vaccine equity and access: case study for health care chatbots. JMIR Form Res 2023 Jan 25;7:e39045 [FREE Full text] [doi: 10.2196/39045] [Medline: 36630649]
14. Trzebiński W, Claessens T, Buhmann J, De Waele A, Hendrickx G, Van Damme P, et al. The effects of expressing empathy/autonomy support using a COVID-19 vaccination chatbot: experimental study in a sample of Belgian adults. JMIR Form Res 2023 May 08;7:e41148 [FREE Full text] [doi: 10.2196/41148] [Medline: 37074978]
15. Luk TT, Lui JHT, Wang MP. Efficacy, usability, and acceptability of a chatbot for promoting COVID-19 vaccination in unvaccinated or booster-hesitant young adults: pre-post pilot study. J Med Internet Res 2022 Oct 04;24(10):e39063 [FREE Full text] [doi: 10.2196/39063] [Medline: 36179132]
16. Ludin N, Holt-Quick C, Hopkins S, Stasiak K, Hetrick S, Warren J, et al. A chatbot to support young people during the COVID-19 pandemic in New Zealand: evaluation of the real-world rollout of an open trial. J Med Internet Res 2022 Nov 04;24(11):e38743 [FREE Full text] [doi: 10.2196/38743] [Medline: 36219754]
17. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-Blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719 [FREE Full text] [doi: 10.2196/40719] [Medline: 36355633]
18. Goonesekera Y, Donkin L. A cognitive behavioral therapy chatbot (Otis) for health anxiety management: mixed methods pilot study. JMIR Form Res 2022 Oct 20;6(10):e37877 [FREE Full text] [doi: 10.2196/37877] [Medline: 36150049]
19. Sinha C, Meheli S, Kadaba M. Understanding digital mental health needs and usage with an artificial intelligence-led mental health app (Wysa) during the COVID-19 pandemic: retrospective analysis. JMIR Form Res 2023 Jan 26;7:e41913 [FREE Full text] [doi: 10.2196/41913] [Medline: 36540052]
20. Whittaker R, Dobson R, Garner K. Chatbots for smoking cessation: scoping review. J Med Internet Res 2022 Sep 26;24(9):e35556 [FREE Full text] [doi: 10.2196/35556] [Medline: 36095295]
21. Wlasak W, Zwanenburg SP, Paton C. Supporting autonomous motivation for physical activity with chatbots during the COVID-19 pandemic: factorial experiment. JMIR Form Res 2023 Jan 25;7:e38500 [FREE Full text] [doi: 10.2196/38500] [Medline: 36512402]
22. Pithpornchaiyakul S, Naorungroj S, Pupong K, Hunsrisakhun J. Using a chatbot as an alternative approach for in-person toothbrushing training during the COVID-19 pandemic: comparative study. J Med Internet Res 2022 Oct 21;24(10):e39218 [FREE Full text] [doi: 10.2196/39218] [Medline: 36179147]

XSL•FO
RenderX

23. Weeks R, Sangha P, Cooper L, Sedoc J, White S, Gretz S, et al. Usability and credibility of a COVID-19 vaccine chatbot for Yyoung adults and health workers in the United States: formative mixed methods study. JMIR Hum Factors 2023 Jan 30;10:e40533 [FREE Full text] [doi: 10.2196/40533] [Medline: 36409300]
24. Chagas BA, Pagano AS, Prates RO, Praes EC, Ferreguetti K, Vaz H, et al. Evaluating user experience with a chatbot designed as a public health response to the COVID-19 pandemic in Brazil: mixed methods study. JMIR Hum Factors 2023 Apr 03;10:e43135 [FREE Full text] [doi: 10.2196/43135] [Medline: 36634267]
25. White BK, Martin A, White JA. User experience of COVID-19 chatbots: scoping review. J Med Internet Res 2022 Dec 27;24(12):e35903 [FREE Full text] [doi: 10.2196/35903] [Medline: 36520624]
26. Shan Y, Ji M, Xie W, Zhang X, Qian X, Li R, et al. Use of health care chatbots among young people in China during the Omicron wave of COVID-19: evaluation of the user experience of and satisfaction with the technology. JMIR Hum Factors 2022 Jun 09;9(2):e36831 [FREE Full text] [doi: 10.2196/36831] [Medline: 35576058]
27. Daniel T, de Chevigny A, Champrigaud A, Valette J, Sitbon M, Jardin M, et al. Answering hospital caregivers' questions at any time: proof-of-concept study of an artificial intelligence-based chatbot in a French hospital. JMIR Hum Factors 2022 Oct 11;9(4):e39102 [FREE Full text] [doi: 10.2196/39102] [Medline: 35930555]
28. Phiri M, Munoriyarwa A. Health chatbots in Africa: scoping review. J Med Internet Res 2023 Jun 14;25:e35573 [FREE Full text] [doi: 10.2196/35573] [Medline: 35584083]
29. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. J Med Internet Res 2023 Jul 26;25:e43068 [FREE Full text] [doi: 10.2196/43068] [Medline: 37224277]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 23.11.23; this is a non–peer-reviewed article;accepted 20.12.23; published 21.03.24.

Please cite as:
Arnold V, Purnat TD, Marten R, Pattison A, Gouda H
Chatbots and COVID-19: Taking Stock of the Lessons Learned
J Med Internet Res 2024;26:e54840
URL: https://www.jmir.org/2024/1/e54840
doi:10.2196/54840
PMID:38512309

©Virginia Arnold, Tina D Purnat, Robert Marten, Andrew Pattison, Hebe Gouda. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Editorial

Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS)

Khaled El Emam[1,2], BEng, PhD; Tiffany I Leung[3,4], MD, MPH; Bradley Malin[5], BA, MSc, PhD; William Klement[2], PhD; Gunther Eysenbach[3,6], MD, MPH

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[2]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
[3]JMIR Publications, Inc, Toronto, ON, Canada
[4]Department of Internal Medicine (adjunct), Southern Illinois University School of Medicine, Springfield, IL, United States
[5]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, United States
[6]School of Health Information Science, University of Victoria, Victoria, BC, Canada

Corresponding Author:
Khaled El Emam, BEng, PhD
School of Epidemiology and Public Health
University of Ottawa
401 Smyth Road
Ottawa, ON, K1H 8L1
Canada
Phone: 1 6137377600
Email: kelemam@ehealthinformation.ca

Abstract

The number of papers presenting machine learning (ML) models that are being submitted to and published in the *Journal of Medical Internet Research* and other JMIR Publications journals has steadily increased. Editors and peer reviewers involved in the review process for such manuscripts often go through multiple review cycles to enhance the quality and completeness of reporting. The use of reporting guidelines or checklists can help ensure consistency in the quality of submitted (and published) scientific manuscripts and, for example, avoid instances of missing information. In this Editorial, the editors of JMIR Publications journals discuss the general JMIR Publications policy regarding authors' application of reporting guidelines and specifically focus on the reporting of ML studies in JMIR Publications journals, using the Consolidated Reporting of Machine Learning Studies (CREMLS) guidelines, with an example of how authors and other journals could use the CREMLS checklist to ensure transparency and rigor in reporting.

(J Med Internet Res 2024;26:e52508) doi:10.2196/52508

KEYWORDS

reporting guidelines; machine learning; predictive models; diagnostic models; prognostic models; artificial intelligence; editorial policy

Introduction

The number of papers presenting machine learning (ML) models that are being submitted to and published in the *Journal of Medical Internet Research* and other JMIR Publications journals has steadily increased over time. The cross-journal JMIR Publications e-collection "Machine Learning" includes nearly 1300 articles as of April 1, 2024 [1], and there are additional sections in other journals, which collate articles related to the field (eg, "Machine Learning from Dermatological Images" [2] in *JMIR Dermatology*). From 2015 to 2022, the number of published articles with "artificial intelligence" (AI) or "machine learning" in the title and abstract in JMIR Publications journals increased from 22 to 298 (13.5-fold growth), and there are already 312 articles in 2023 (14-fold growth). For *JMIR Medical Informatics*, the number of articles increased from 10 to 160 (16-fold growth) until 2022. This is consistent with the growth in the research and application of medical AI in general where a similar PubMed search (with the keyword "medicine") revealed a 22-fold growth (from 640 to 14,147 articles) between 2015 and 2022, and there are already 11,272 matching articles in 2023.

Many papers reporting the use of ML models in medicine have used a large clinical data set to make diagnostic or prognostic predictions [3-6]. However, the use of data from electronic

XSL•FO
RenderX

health records and other resources is often not without pitfalls as these data are typically collected and optimized for other purposes (eg, medical billing) [7].

Editors and peer reviewers involved in the review process for such manuscripts often go through multiple review cycles to enhance the quality and completeness of reporting [8]. The use of reporting guidelines or checklists can help ensure consistency in the quality of submitted (and published) scientific manuscripts and, for instance, avoid instances of missing information. For example, in the experiences of the editors-in-chief of *JMIR AI*, missing information is especially notable because for manuscripts reporting on ML models, which are submitted to *JMIR AI*, this can delay the overall review interval by adding more revision cycles.

According to the EQUATOR (Enhancing the Quality and Transparency of Health Research) network, a reporting guideline is "a simple, structured tool for health researchers to use while writing manuscripts. A reporting guideline provides a minimum list of information needed to ensure a manuscript can be, for example: understood by a reader, replicated by a researcher, used by a doctor to make a clinical decision, and included in a systematic review" [9]. These can be presented in the form of a checklist, flow diagram, or structured text.

In this Editorial, we discuss the general JMIR Publications policy regarding authors' application of reporting guidelines. We then focus specifically on the reporting of ML studies in JMIR Publications journals.

JMIR Publications Policy on the Use of Reporting Guidelines

Accumulating evidence suggests that when authors apply reporting guidelines and reporting checklists in health research, they can be beneficial for authors, readers, and the discipline overall by enabling the replication or reproduction of studies. Recent evidence suggests that asking reviewers to use reporting checklists, instead of authors, offers no added benefits regarding reporting quality [10]. However, Botos [11] reported a positive association between reviewer ratings of adherence to reporting guidelines and favorable editorial decisions, while Stevanovic et al [12] reported a significant positive correlation between adherence to reporting guidelines and citations and between adherence to reporting guidelines and publication in higher-impact-factor journals.

JMIR Publications' editorial policy recommends that authors adhere to applicable study design and reporting guidelines when preparing manuscripts for submission [13]. Authors should note that most reporting guidelines are strongly recommended, particularly because they can improve the quality, completeness, and organization of the presented work. At this time, JMIR Publications *requires* reporting checklists to be completed and supplied as multimedia appendices for randomized controlled trials without [14-16] or those with eHealth or mobile health components [17], systematic and scoping literature reviews across the portfolio, and Implementation Reports in *JMIR Medical Informatics* [18]. Although some medical journals have mandated the use of certain reporting guidelines and checklists, JMIR Publications recognizes that authors may have concerns about the additional burden that the formalized use of checklists may bring to the submission process. As such, JMIR Publications has chosen to begin recommending the use of ML reporting guidelines and will evaluate their benefits and gather feedback on implementation costs before considering more stringent requirements.

Reporting on ML Models

Regarding the reporting of prognostic and diagnostic ML studies, multiple directly relevant checklists have been developed. Klement and El Emam [19] have consolidated these guidelines and checklists into a single set that we refer to as the Consolidated Reporting of Machine Learning Studies (CREMLS) checklist. CREMLS serves as a reporting checklist for journals publishing research describing the development, evaluation, and application of ML models, including all JMIR Publications journals, which have officially adopted these guidelines. CREMLS was developed by identifying existing relevant reporting guidelines and checklists. The initial item list was identified through a structured literature review and expert curation, and then the quality of the methods used for their development was assessed to narrow them down to a high-quality subset. This high-quality item subset was further filtered to reveal those that meet specific inclusion and exclusion criteria. The resultant items were converted to guidelines and a checklist that was reviewed by the editorial board of *JMIR AI*, followed by a preliminary application to assess articles published in *JMIR AI*. The final checklist offers present-day best practices for high-quality reporting of studies using ML models.

Examples of the application of the CREMLS checklist are presented in Table 1. In doing so, we identified 7 articles published in JMIR Publications journals, which exemplify each checklist item. Note that not all of the items are relevant to each article, and some articles are particularly good examples of how to operationalize a checklist item.

XSL•FO
RenderX

Table 1. Illustration of how various articles published in JMIR Publications journals implement each of the CREMLS (Consolidated Reporting of Machine Learning Studies) checklist items.

Item number	Item	Example illustrating the item
Study details		
1.1	The medical or clinical task of interest	Examines chronic disease management—a clinical problem with 4 example solutions using ML[a] models [20]
1.2	The research question	Proposes a framework to transfer old knowledge to a new environment to manage drifts [21]
1.3	Current medical or clinical practice	Provides a review of current practice and issues associated with chronic disease management [20]
1.4	The known predictors and confounders of what is being predicted or diagnosed	Describes variables defined as part of a well-established health test available to the public [20]
1.5	The overall study design	Presents experimental design with data flow and data partitions used at various steps of the experiment (Figure 1 [22])
1.6	The medical institutional settings	Describes the institution as an academic (teaching) community hospital where the data were collected [23]
1.7	The target patient population	Clear partitioning of target patient populations and the comparator group [20]
1.8	The intended use of the ML model	Describes how the prediction model fits in the clinical practice of scheduling operating theater procedures [5]
1.9	Existing model performance benchmarks for this task	Reviews existing research and presents achieved performance (eg, AUC[b]) [20]
1.10	Ethical and other regulatory approvals obtained	Ethics approvals [5]
The data		
2.1	Inclusion or exclusion criteria for the patient cohort	Defined in Figure 1 in the paper by Kendale et al [5]
2.2	Methods of data collection	Describes sources and methods of data collection, what type of data were used, and potential implied bias in interpretation [23]
2.3	Bias introduced due to the method of data collection used	Discusses potential bias in data collection and outcome definition [23]
2.4	Data characteristics	Uses descriptive statistics to show data characteristics for different types of data (demographics and clinical measurements) [23]
2.5	Methods of data transformation and preprocessing applied	Imputation is discussed [5]
2.6	Known quality issues with the data	Missingness and outlier detection were discussed [5]
2.7	Sample size calculation	Brief section dedicated to power analysis [5]
2.8	Data availability	Explains how to obtain a copy of the data [24]
Methodology		
3.1	Strategies for handling missing data	Describes how missing values were replaced [20]
3.2	Strategies for addressing class imbalance	Describes the approach of using SMOTE[c] to adjust class ratios to address imbalance [23]
3.3	Strategies for reducing dimensionality of data	Describes the vectorization of a dimension of 100 into a 2D space using an established algorithm [22]
3.4	Strategies for handling outliers	The authors stated the threshold values used to detect outliers [5]
3.5	Strategies for data augmentation	Showed how variable similarity is achieved between synthetic and real data in the context of augmentation [24]
3.6	Strategies for model pretraining	Describes and illustrates (Figure 1) how models from other data sets were trained and used in the new model [23]
3.7	The rationale for selecting the ML algorithm	Discusses properties of the selected algorithm relevant to the problem at hand as motivation [20]
3.8	The method of evaluating model performance during training	Presents a separate discussion of evolution in cross-validation settings and external evaluation while also describing hyperparameter tuning [23]

XSL•FO
RenderX

Item number	Item	Example illustrating the item
3.9	The method used for hyperparameter tuning	Comprehensive description of tuning within nested cross-validation (this is a tutorial but illustrates how to describe the process) [25]
3.10	Model's output adjustments	Describes the final model, how it was calibrated and discusses the impact of embedding on patient data for interpretation [22]
Evaluation		
4.1	Performance metrics used to evaluate the model	Comprehensive and detailed discussion of evaluation and quality metrics [24]
4.2	The cost or consequence of errors	Comprehensive error analysis [25]
4.3	The results of internal validation	Detailed validation discussion (internally and externally) [25]
4.4	The final model hyperparameters	Presents details of the final model and the winning parameters [5]
4.5	Model evaluation on an external data set	Detailed and comprehensive external validation that is separate from model testing [5]
4.6	Characteristics relevant for detecting data shift and drift	Implements performance monitoring, addresses data shifts over time, and illustrates them in detail [21]
Explainability and transparency		
5.1	The most important features and how they relate to the outcomes	Presents variable importance (SHAP[d] values) in the context of interpretation and compares it to existing literature [5]
5.2	Plausibility of model outputs	Shows sample output (Figure 4 in the paper by Kendale et al [5])
5.3	Interpretation of a model's results by an end user	Good discussion about interpretability and use of the final model [5]

[a]ML: machine learning.

[b]AUC: area under the curve.

[c]SMOTE: synthetic minority oversampling technique.

[d]SHAP: Shapely additive explanations.

We strongly advise authors who seek to submit their manuscripts describing the development, evaluation, and application of ML models to the *Journal of Medical Internet Research*, *JMIR AI*, *JMIR Medical Informatics*, or other JMIR Publications journals to adhere to the CREMLS guidelines and checklist to ensure that they have considered and addressed all relevant details for their work before initiating their submission and review process. More complete and high-quality reporting benefits the authors by accelerating the review cycle and reducing the burden on reviewers. Hence, the need exists for reporting guidelines and checklists for papers describing prognostic and diagnostic ML studies. This is expected to assist, for example, in reducing missing documentation on hyperparameters for an ML model and to clarify how data leakage was avoided. We have observed that peer reviewers have, in practice, been asking authors to improve reporting on the same topics covered in the CREMLS checklist. This is not a surprise given that peer reviewers are experts in the field and would note important information that is missing. Nevertheless, we would encourage reviewers to use the checklist regularly to ensure completeness and consistency.

The CREMLS checklist's scope is limited to ML models using structured data that are trained and evaluated in silico and in shadow mode. This provides a significant opportunity to expand on the CREMLS to different data modalities and additional phases of model deployment. Should such extended reporting guidelines and checklists be developed, they may be considered for recommendation for submissions to JMIR Publications journals, incorporating lessons learned from the initial checklist for studies reporting the use of ML models.

Conclusion

There is evidence that the completeness of reporting of research studies is beneficial to the authors and the broader scientific community. For prognostic and diagnostic ML studies, many reporting guidelines have been developed, and these have been consolidated into CREMLS, capturing the combined value of the source guidelines and checklists in one place. In this Editorial, we extend journal policy and recommend that authors follow these guidelines when submitting articles to journals in the JMIR Publications portfolio. This will improve the reproducibility of research studies using ML methods, accelerate review cycles, and improve the quality of published papers overall. Given the rapid growth of studies developing, evaluating, and applying ML models, it is important to establish reporting standards early.

Authors' Contributions

KEE and BM conceptualized this study and drafted, reviewed, and edited the manuscript. TIL and GE reviewed and edited the manuscript. WK prepared the literature summary and reviewed the manuscript.

XSL•FO
RenderX

Conflicts of Interest

KEE and BM are co–editors-in-chief of *JMIR AI*. KEE is the cofounder of Replica Analytics, an Aetion company, and has financial interests in the company. TIL is the scientific editorial director at JMIR Publications, Inc. GE is the executive editor and publisher at JMIR Publications, Inc, receives a salary, and owns equity.

References

1. Machine Learning. JMIR Medical Informatics. URL: https://medinform.jmir.org/themes/500-machine-learning [accessed 2024-04-01]
2. Machine Learning from Digital Images in Dermatology. JMIR Dermatology. URL: https://derma.jmir.org/themes/922-machine-learning-from-digital-images-in-dermatology [accessed 2023-09-22]
3. Lee S, Kang WS, Kim DW, Seo SH, Kim J, Jeong ST, et al. An artificial intelligence model for predicting trauma mortality among emergency department patients in South Korea: retrospective cohort study. J Med Internet Res 2023 Aug 29;25:e49283 [FREE Full text] [doi: 10.2196/49283] [Medline: 37642984]
4. Deng Y, Ma Y, Fu J, Wang X, Yu C, Lv J, et al. Combinatorial use of machine learning and logistic regression for predicting carotid plaque risk among 5.4 million adults with fatty liver disease receiving health check-ups: population-based cross-sectional study. JMIR Public Health Surveill 2023 Sep 07;9:e47095 [FREE Full text] [doi: 10.2196/47095] [Medline: 37676713]
5. Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: algorithm development and validation study. JMIR AI 2023 Sep 8;2:e44909. [doi: 10.2196/44909]
6. Williams DD, Ferro D, Mullaney C, Skrabonja L, Barnes MS, Patton SR, et al. An "All-Data-on-Hand" deep learning model to predict hospitalization for diabetic ketoacidosis in youth with type 1 diabetes: development and validation study. JMIR Diabetes 2023 Jul 18;8:e47592 [FREE Full text] [doi: 10.2196/47592] [Medline: 37224506]
7. Maletzky A, Böck C, Tschoellitsch T, Roland T, Ludwig H, Thumfart S, et al. Lifting hospital electronic health record data treasures: challenges and opportunities. JMIR Med Inform 2022 Oct 21;10(10):e38557 [FREE Full text] [doi: 10.2196/38557] [Medline: 36269654]
8. Emam KE, Klement W, Malin B. Reporting and methodological observations on prognostic and diagnostic machine learning studies. JMIR AI 2023:e47995 [FREE Full text]
9. What is a reporting guideline? Enhancing the QUAlity and Transparency Of health Research. URL: https://www.equator-network.org/about-us/what-is-a-reporting-guideline/ [accessed 2023-09-22]
10. Speich B, Mann E, Schönenberger CM, Mellor K, Griessbach AN, Dhiman P, et al. Reminding peer reviewers of reporting guideline items to improve completeness in published articles: primary results of 2 randomized trials. JAMA Netw Open 2023 Jun 01;6(6):e2317651 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.17651] [Medline: 37294569]
11. Botos J. Reported use of reporting guidelines among authors, editorial outcomes, and reviewer ratings related to adherence to guidelines and clarity of presentation. Res Integr Peer Rev 2018 Sep 27;3(1):7 [FREE Full text] [doi: 10.1186/s41073-018-0052-4] [Medline: 30275983]
12. Stevanovic A, Schmitz S, Rossaint R, Schürholz T, Coburn M. CONSORT item reporting quality in the top ten ranked journals of critical care medicine in 2011: a retrospective analysis. PLoS One 2015 May 28;10(5):e0128061 [FREE Full text] [doi: 10.1371/journal.pone.0128061] [Medline: 26020246]
13. What reporting guidelines should I follow for my article? JMIR Publications Knowledge Base and Help Center. URL: https://support.jmir.org/hc/en-us/articles/115001575267-What-reporting-guidelines-should-I-follow-for-my-article [accessed 2024-01-30]
14. Schulz KF, Altman D, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Trials 2010 Mar 24;11:32 [FREE Full text] [doi: 10.1186/1745-6215-11-32] [Medline: 20334632]
15. Schulz KF, Altman D, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010 Jun 01;152(11):726-732 [FREE Full text] [doi: 10.7326/0003-4819-152-11-201006010-00232] [Medline: 20335313]
16. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340(mar23 1):c869-c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
17. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
18. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]

XSL•FO
RenderX

19. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. J Med Internet Res 2023 Aug 31;25:e48763 [FREE Full text] [doi: 10.2196/48763] [Medline: 37651179]
20. Lee C, Jo B, Woo H, Im Y, Park RW, Park C. Chronic disease prediction using the common data model: development study. JMIR AI 2022 Dec 22;1(1):e41030 [FREE Full text] [doi: 10.2196/41030]
21. Zhang X, Xue Y, Su X, Chen S, Liu K, Chen W, et al. A transfer learning approach to correct the temporal performance drift of clinical prediction models: retrospective cohort study. JMIR Med Inform 2022 Nov 09;10(11):e38053 [FREE Full text] [doi: 10.2196/38053] [Medline: 36350705]
22. Steiger E, Kroll LE. Patient embeddings from diagnosis codes for health care prediction tasks: Pat2Vec machine learning framework. JMIR AI 2023 Apr 21;2:e40755 [FREE Full text] [doi: 10.2196/40755]
23. Sang S, Sun R, Coquet J, Carmichael H, Seto T, Hernandez-Boussard T. Learning from past respiratory infections to predict COVID-19 outcomes: retrospective study. J Med Internet Res 2021 Feb 22;23(2):e23026 [FREE Full text] [doi: 10.2196/23026] [Medline: 33534724]
24. Kang HYJ, Batbaatar E, Choi D, Choi KS, Ko M, Ryu KS. Synthetic tabular data based on generative adversarial networks in health care: generation and validation using the divide-and-conquer strategy. JMIR Med Inform 2023 Nov 24;11:e47859 [FREE Full text] [doi: 10.2196/47859] [Medline: 37999942]
25. Wilimitis D, Walsh CG. Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial. JMIR AI 2023 Dec 18;2:e49023 [FREE Full text] [doi: 10.2196/49023]

Abbreviations

AI: artificial intelligence
CREMLS: Consolidated Reporting of Machine Learning Studies
EQUATOR: Enhancing the Quality and Transparency of Health Research
ML: machine learning

Edited by T Leung; submitted 04.04.24; this is a non–peer-reviewed article;accepted 04.04.24; published 02.05.24.

Please cite as:
El Emam K, Leung TI, Malin B, Klement W, Eysenbach G
Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS)
J Med Internet Res 2024;26:e52508
URL: https://www.jmir.org/2024/1/e52508
doi:10.2196/52508
PMID:38696776

©Khaled El Emam, Tiffany I Leung, Bradley Malin, William Klement, Gunther Eysenbach. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Editorial

Converge or Collide? Making Sense of a Plethora of Open Data Standards in Health Care

Guy Tsafnat[1,2,3], PhD; Rachel Dunscombe[4,5*], MICT; Davera Gabriel[1,3,6*], RN; Grahame Grieve[7,8*], PhD; Christian Reich[3,9*], BSc, MD

[1]Evidentli Pty Ltd, Surry Hills, Australia
[2]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie Univeristy, Macquarie Park, Australia
[3]OHDSI OMOP + FHIR Working Group
[4]openEHR International, St. Helens, United Kingdom
[5]Imperial College London, London, United Kingdom
[6]School of Medicine, Johns Hopkins University, Baltimore, MD, United States
[7]Health Level 7 International, Ann Arbor, MI, United States
[8]Health Intersections Pty Ltd, Melbourne, Australia
[9]Odysseus Data Services, Cambridge, MA, United States
[*]these authors contributed equally

Corresponding Author:
Guy Tsafnat, PhD
Evidentli Pty Ltd
50 Holt St
Suite 516
Surry Hills, 2010
Australia
Phone: 61 415481043
Email: guyt@evidentli.com

Abstract

Practitioners of digital health are familiar with disjointed data environments that often inhibit effective communication among different elements of the ecosystem. This fragmentation leads in turn to issues such as inconsistencies in services versus payments, wastage, and notably, care delivered being less than best-practice. Despite the long-standing recognition of interoperable data as a potential solution, efforts in achieving interoperability have been disjointed and inconsistent, resulting in numerous incompatible standards, despite the widespread agreement that fewer standards would enhance interoperability. This paper introduces a framework for understanding health care data needs, discussing the challenges and opportunities of open data standards in the field. It emphasizes the necessity of acknowledging diverse data standards, each catering to specific viewpoints and needs, while proposing a categorization of health care data into three domains, each with its distinct characteristics and challenges, along with outlining overarching design requirements applicable to all domains and specific requirements unique to each domain.

(J Med Internet Res 2024;26:e55779) doi:10.2196/55779

KEYWORDS

interoperability; clinical data; open data standards; health care; digital health; health care data

Introduction

The digital health sector presents unique and diverse data challenges that manifest as a fractured ecosystem and often fails to deliver effective collaboration among interconnected entities [1]. The consequences include inconsistencies between services rendered and payments, waste, and—most importantly—suboptimal care. Interoperable data has been hailed as a near-future solution to many of these challenges for decades. Paradoxically, interoperability efforts have been fractured and inconsistent, resulting in many incompatible interoperability standards, despite widespread acknowledgment that fewer standards would provide better interoperability [2,3]. This paper presents a typology of healthcare data requirements and describes the challenges and opportunities of open data standards in health care.

Recognizing that different data standards represent different points of view and respond to different needs, and that no single

XSL•FO
RenderX

standard would necessarily be able to meet all the requirements of all health care systems, we distinguish three domains of health care data with their unique characteristics and challenges, and outline high-level design requirements. We distinguish between requirements that are common across all domains and those that are specific to each domain.

Three Domains of Health Care Data

Clinical Care and Administration

Health care data are generated as documentation of clinical service delivery for various purposes, including continuity of care and billing. A repository of clinical records containing the complete patient history of every patient is required to deliver continuous and consistent care efficiently and effectively. As clinical care sometimes lacks a strong evidence base from which best practices can be gleaned, care documentation should be able to address a broad range of data types, including "one off" events reflecting unforeseen patient circumstances [4].

The main focus of standards for electronic medical record interoperability is to protect users from vendor lock-in, in which the cost of changing to a better system outweighs the benefits expected from the switch. An electronic health care record system that uses an open standard can replace another that uses the same standard far more easily and with less disruption to clinical practice than two incompatible systems where data transformation from one system to the next represents the lion's share of the cost.

Data Exchange

Records of care provided are used in communication between different actors. For example, continuity of care depends on multiple health care providers sharing information about the patient, and effective billing requires sharing information between providers and payers. It is often critical that information is passed between systems immediately to inform urgent decisions. Standards for the transmission of health care data should be able to be arranged in small quantities that include all and only the relevant information.

The main focus of data exchange standards is fidelity to the operational health care processes. For example, one focus is on capturing and faithfully representing the terminology used in the local processes, whatever that is, as opposed to restricting use to a common terminology.

Longitudinal Analysis

The main focus of longitudinal analysis are patterns, trends, and their predictors. A practical consideration is that such patterns tend to be subtle, and hence, data from multiple systems and organizations are required to be aggregated or compared to elucidate reliable phenomena. As an empirical discipline, medicine has incrementally improved using inferences made from big data. Moreover, medical evidence should produce predictable outcomes across multiple settings and for all patients.

Analytical methods use generalizations such as grouping patients with similar characteristics like age. "One off" events and subtle differences in clinical practice are inconsequential or even harmful to population statistics. For example, as age is often measured in years and stratified into a larger age bracket, the exact birth date of each patient can be substituted by an approximate date without changing the results of the analysis. Another common reason to lose fidelity is patient privacy and prevention of unauthorized reidentification of individuals from analytic data sets.

Practicality, Feasibility, and Community

Technical Trade-Offs

The technical suitability of a standard to the problem it addresses is important, but digital health and data interoperability are applied disciplines. Practical and socio-technical considerations, such as available skill sets, are as important as, if not more important than, technical superiority. While no standard is perfect, there are major differences in what standards can do "out of the box" versus how much customization is required before they can be put to use. To provide data interoperability in practice, standards have to find the right balance between customizability and rigidity. The United States Core Data for Interoperability is a prime example of such balance [5]. These core data element definitions are regulated for use in the United States and can be thought of as a midpoint between customizations needed to be made at every site and rigidity with regard to specific data elements.

Community

Data standards are only as useful as their adoption by collaborating organizations, so standards that are in widespread use in the community are more likely to be beneficial to organizations seeking to join collaboration networks that use these standards. Engagement from communities that not only use a standard but also contribute and improve it over time, differentiates legacy standards from more sustainable ones. An actively engaged community will, over time, evolve even an infantile standard into a superior one. An engaged and vibrant community is therefore a major advantage for the longevity of the data standards it uses [6]. Open standards have a clear advantage over proprietary ones, as they are available for contribution from a larger audience and are tested in more diverse environments. Conversely, it is usually easier to govern the development of proprietary standards, which can lead to shorter improvement cycles [7].

Terminologies

A common aspect of all data standards, regardless of application, is the use of common terminologies. Many standards rely on terminologies that are developed independently. Terminologies can be flat or ontological, specific to one or a handful of domains or more general, proprietary or open. Depending on the intended use of the standard and available resources, the inclusion policy for terminologies in a standard may differ. Standards designed for continuity of care may allow the use of external terminologies if they are available to all parties providing care, whereas standards used for data mining may restrict their use to only vetted terminologies. As is the case of data standards, terminologies benefit from active communities to keep them up to date. As evident in differing design criteria, a truly

XSL•FO
RenderX

interoperable, agile health care system that provides the best possible clinical care implements multiple standards. For example, standard records from a repository of complete patient histories may be transformed to a standard for transmission and transformed again when entered into a data warehouse for analysis. Systematic harmonization between standards would reduce or eliminate information loss in the transformation processes [8].

Emerging Standards

Paradox of Interoperability

A paradox of health care interoperability is the existence of a large number of standards with significant overlap among them. Each may be more or less equally suited for its intended use, so selecting a standard for implementation can be difficult. Harmonization from all standards to all others is impractical. The interoperability paradox is that more standards lead to less interoperability. We therefore propose that the health informatics community converge on just three standards. We have chosen standards that are open, are backed by active communities, and have been tried and tested in their respective areas.

Clinical Care and Administration

openEHR [9] provides a common data structure for health IT systems, including but not limited to, electronic medical records and patient administration systems. The standard uses a centralized database with a tree structure for efficient retrieval of a complete patient record. The schema of the database is patient-centric, clinical, and extensible with reusable archetypes and templates. openEHR is increasingly preferred over proprietary formats, as multiple vendors that offer commercial systems based on openEHR means that a very expensive process for switching vendors can be greatly simplified.

Data Exchange

Health Level 7's (HL7) Fast Healthcare Interoperability Resources (FHIR) [10] is an open standard particularly suited for data transfer. At the center of the FHIR specification is a representational state transfer (RESTful) application programming interface (API) that provides real-world health care services such as "record patient details," "prescribe a medication," "suggest the correct treatment for a patient," and "make an appointment with your doctor." Its resource structure only contains information relevant to a particular API call so it can be transmitted efficiently. Specific minimum data sets of FHIR (called cores) are required in the United States, with other countries currently developing additional core requirements. Using FHIR allows different entities in that country to streamline continuity of care by exchanging information between systems and organizations.

Longitudinal Analysis

Observational Health Data Sciences and Informatics' (OHDSI) Observational Medical Outcomes Partnership Common Data Model (OMOP) [11] is an open standard designed for longitudinal interrogation of clinical records. Its use of a patient-centric relational model makes it suited for complex queries across many patients' records to discover patterns and trends using statistical and machine learning techniques.

As the requirements for each data standard mean that one standard cannot replace the others, harmonization projects are led by community volunteers to provide standard transformations from one standard to another. For example, the collaboration between OHDSI and HL7 manifests as a common working group with weekly meetings to produce an implementation guide for algorithms that convert data from FHIR to OMOP and from OMOP to FHIR. Similar projects for harmonizing openEHR with OMOP and FHIR with openEHR provide a standard architecture for effective data interoperability and reuse that will create a virtuous cycle of growing community participation, improving the standards and simplifying harmonization. Such volunteer groups are highly beneficial, not only to the standards they are dedicated to but also to health care systems globally.

Which Standard to Use?

Most users will easily choose whether their needs' focus is operational care delivery or analytics and identify which systems are involved. Figure 1 provides a high-level key, making it easy to choose between openEHR, OMOP, or FHIR and invest in extract, transform, load processes to change data representations from one to another. An interoperable health system would use openEHR to collect data, FHIR to transmit data between systems and organizations, and OMOP to find insights in the data.

It is possible for transformations from one standard to another to be "lossy." For example:

- Fidelity may be lost when converting from a specific term (eg, "fracture of the fourth metatarsal on the left foot") to a more general one (eg, "fracture of the fourth metatarsal").
- Information can be lost when information captured in one standard (eg, the time drug administration was witnessed in FHIR) has no place in another standard (eg, OMOP).
- Relationships may be lost when information has to be split (eg, a diagnosis and subsequent treatment).
- Accuracy may be lost when converting to a standard that requires information not found in the other (eg, the time a drug was stopped is required for every exposure in OMOP and may need to be estimated whenever it is not present in the data set).

Consideration should be given to minimizing such losses. These approaches can use tacit or implicit knowledge about the data set, external knowledge, or metadata [12].

XSL•FO
RenderX

Figure 1. Alignment between domain (outer circle) and open standard (inner circle). FHIR: Fast Healthcare Interoperability Resources; OMOP: Outcomes Partnership Common Data Model.

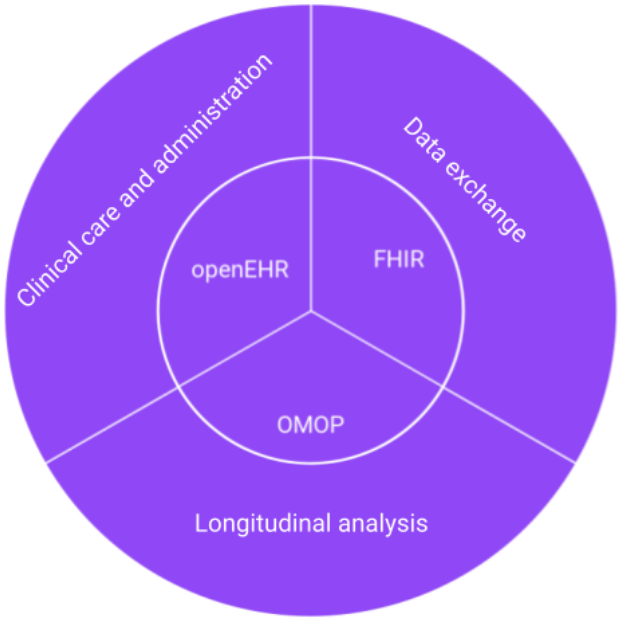

Conclusion

Open standards, backed by engaged communities, hold an advantage over proprietary ones. They enable broader contributions and testing, fostering continual improvement. However, achieving true interoperability requires harmonization among standards and a pathway for migration from other standards and proprietary models. Three open standards are particularly suited to their respective health care domains: openEHR for clinical care and administration, FHIR for data exchange, and OMOP for longitudinal analysis. The most important aspect of every standard is its community, and all three have active and growing communities that continue to use and improve these standards.

Conflicts of Interest

GT is the founder and chairman of Evidentli Pty Ltd and an associate editor of the *Journal of Medical Internet Research*. RD is the chief executive officer of openEHR International. GG is the founder of Health Intersections Pty Ltd. CR is the chief executive officer of Odysseus Data Services. DG is an employee of Evidentli and serves on the Evidentli Scientific Advisory Board.

References

1. Coiera E. Guide to Health Informatics, 3rd Edition. London: CRC Press; 2015.
2. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79. [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
3. Ahmadi N, Zoch M, Kelbert P, Noll R, Schaaf J, Wolfien M, et al. Methods used in the development of common data models for health data: scoping review. JMIR Med Inform 2023 Aug 03;11:e45116 [FREE Full text] [doi: 10.2196/45116] [Medline: 37535410]
4. Handler T, Holtmeier R, Metzger J, Overhage M, Taylor S, Underwood C. HIMSS electronic health record definitional model version 1.0. ResearchGate. 2003. URL: https://www.researchgate.net/publication/247151484_HIMSS_electronic_health_record_definitional_model_version_10 [accessed 2024-04-04]
5. United States Core Data for Interoperability (USCDI). Office of the National Coordinator for Health Information Technology. 2020 Jul. URL: https://www.healthit.gov/isa/sites/isa/files/2020-10/USCDI-Version-1-July-2020-Errata-Final_0.pdf [accessed 2024-04-04]
6. Coiera E. The standard problem. J Am Med Inform Assoc 2023 Nov 17;30(12):2086-2097. [doi: 10.1093/jamia/ocad176] [Medline: 37654094]
7. Reynolds CJ, Wyatt JC. Open source, open standards, and health care information systems. J Med Internet Res 2011 Feb 17;13(1):e24 [FREE Full text] [doi: 10.2196/jmir.1521] [Medline: 21447469]
8. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
9. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]

XSL•FO
RenderX

10. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR (Health Information Technology Standards). Cham: Springer; 2016.
11. Hripcsak G, Duke J, Shah N, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
12. Bönisch C, Kesztyüs D, Kesztyüs T. Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. Sci Data 2022 Oct 28;9(1):659. [doi: 10.1038/s41597-022-01792-7] [Medline: 36307424]

Abbreviations

API: application programming interface
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Outcomes Partnership Common Data Model
RESTful: representational state transfer

Edited by A Mavragani; submitted 23.12.23; this is a non–peer-reviewed article;accepted 13.03.24; published 09.04.24.

Please cite as:
Tsafnat G, Dunscombe R, Gabriel D, Grieve G, Reich C
Converge or Collide? Making Sense of a Plethora of Open Data Standards in Health Care
J Med Internet Res 2024;26:e55779
URL: https://www.jmir.org/2024/1/e55779
doi:10.2196/55779
PMID:38593431

©Guy Tsafnat, Rachel Dunscombe, Davera Gabriel, Grahame Grieve, Christian Reich. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing Methods for Assessing Mental Activity Using Human-Smartphone Interactions: Comparative Analysis of Activity Levels and Phase Patterns in General Mental Activities, Working Mental Activities, and Physical Activities

Hung-Hsun Chen[1,2], PhD; Chen Lin[3], PhD; Hsiang-Chih Chang[3,4], PhD; Jen-Ho Chang[5,6], PhD; Hai-Hua Chuang[7,8,9,10], MD, PhD; Yu-Hsuan Lin[4,11,12], MD, PhD

[1]Department of Mathematics, Fu Jen Catholic University, Taipei, Taiwan
[2]Program of Artificial intelligence and Information Security, Fu Jen Catholic University, Taipei, Taiwan
[3]Department of Biomedical Sciences and Engineering, National Central University, Taoyuan City, Taiwan
[4]Institute of Population Health Sciences, National Health Research Institutes, Miaoli County, Taiwan
[5]Institute of Ethnology, Academia Sinica, Taipei, Taiwan
[6]Department of Psychology, National Taiwan University, Taipei, Taiwan
[7]College of Medicine, Chang Gung University, Taoyuan, Taiwan
[8]Department of Family Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan
[9]Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, Taiwan
[10]School of Medicine, National Tsing Hua University, Hsinchu, Taiwan
[11]Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan
[12]Department of Psychiatry, College of Medicine, National Taiwan University, Taipei, Taiwan

Corresponding Author:
Yu-Hsuan Lin, MD, PhD
Institute of Population Health Sciences
National Health Research Institutes
35 Keyan Road
Zhunan
Miaoli County
Taiwan
Phone: 886 37 206 166 ext 36383
Email: yuhsuanmed@gmail.com

Abstract

Background: Human biological rhythms are commonly assessed through physical activity (PA) measurement, but mental activity may offer a more substantial reflection of human biological rhythms.

Objective: This study proposes a novel approach based on human-smartphone interaction to compute mental activity, encompassing general mental activity (GMA) and working mental activity (WMA).

Methods: A total of 24 health care professionals participated, wearing wrist actigraphy devices and using the “Staff Hours” app for more than 457 person-days, including 332 workdays and 125 nonworkdays. PA was measured using actigraphy, while GMA and WMA were assessed based on patterns of smartphone interactions. To model WMA, machine learning techniques such as extreme gradient boosting and convolutional neural networks were applied, using human-smartphone interaction patterns and GPS-defined work hours. The data were organized by date and divided into person-days, with an 80:20 split for training and testing data sets to minimize overfitting and maximize model robustness. The study also adopted the M10 metric to quantify daily activity levels by calculating the average acceleration during the 10-hour period of highest activity each day, which facilitated the assessment of the interrelations between PA, GMA, and WMA and sleep indicators. Phase differences, such as those between PA and GMA, were defined using a second-order Butterworth filter and Hilbert transform to extract and calculate circadian rhythms and instantaneous phases. This calculation involved subtracting the phase of the reference signal from that of the target signal and averaging these differences to provide a stable and clear measure of the phase relationship between the signals.

XSL•FO
RenderX

Additionally, multilevel modeling explored associations between sleep indicators (total sleep time, midpoint of sleep) and next-day activity levels, accounting for the data's nested structure.

Results: Significant differences in activity levels were noted between workdays and nonworkdays, with WMA occurring approximately 1.08 hours earlier than PA during workdays ($P<.001$). Conversely, GMA was observed to commence about 1.22 hours later than PA ($P<.001$). Furthermore, a significant negative correlation was identified between the activity level of WMA and the previous night's midpoint of sleep ($\beta=-0.263$, $P<.001$), indicating that later bedtimes and wake times were linked to reduced activity levels in WMA the following day. However, there was no significant correlation between WMA's activity levels and total sleep time. Similarly, no significant correlations were found between the activity levels of PA and GMA and sleep indicators from the previous night.

Conclusions: This study significantly advances the understanding of human biological rhythms by developing and highlighting GMA and WMA as key indicators, derived from human-smartphone interactions. These findings offer novel insights into how mental activities, alongside PA, are intricately linked to sleep patterns, emphasizing the potential of GMA and WMA in behavioral and health studies.

(J Med Internet Res 2024;26:e56144) doi:10.2196/56144

KEYWORDS

digital phenotyping; human-smartphone interaction; labor or leisure; machine learning; mental activity; physical activity

Introduction

Human biological rhythms are commonly assessed through the measurement of physical activity (PA), such as wrist-worn actigraphy, which estimates sleep and wake times based on patterns of PA [1]. Actigraphy uses wrist-worn accelerometers to collect data on individuals' physical activities, providing a practical means to approximate sleep and circadian rhythm measurements. The rest-activity rhythm throughout the day serves as a valuable metric for quantifying circadian rhythms [2]. However, human beings distinguish themselves from other animals by engaging in highly active mental activity, particularly those individuals who endure long working hours in contemporary society. Given its greater quantity and relevance, mental activity may offer a more substantial reflection of human biological rhythms than PA alone. This is especially pertinent in the digital age, where mental engagement often supersedes PA in daily life. Further differentiation within mental activity can be observed through various theories related to labor or leisure and cognitive control [3,4]. Nevertheless, the current landscape lacks continuous methods for measuring mental activity and working-related mental activity throughout the day. Although some studies have endeavored to explore these aspects, their ecological validity often faces challenges stemming from prolonged monitoring, reliance on self-reports, and limited accuracy [5].

In response to these challenges, our study addressed the nuanced relationship between mental activities and daily interactions with technology, especially with smartphones. We advanced a novel methodology centered on human-smartphone interaction to quantify both general mental activity (GMA) and working mental activity (WMA). This method notably distinguished between work and leisure periods based on interaction patterns. By doing so, our study bridges a significant gap in understanding how modern technology influences daily mental rhythms, a topic underexplored in current literature.

Previous research in cognitive psychology [3,4,6-9], which focused on the dynamics of mental effort between labor and leisure, primarily used experience-sampling studies. However, these approaches often fell short of capturing the complex interplay of smartphone use across different stages of mental engagement. Additionally, the dual role of smartphones as tools for work and leisure blurred the distinction between professional and personal use, a phenomenon that prior studies had noted [10]. For example, smartphones were frequently used for work-related purposes at home and for private purposes at work, reflecting their role as leisure surrogates. Nevertheless, the true value of a smartphone in these contexts was less about the device itself and more about the behaviors it facilitated, such as responding to notifications [8]. Traditional experience-sampling methods struggled to categorize and analyze these screen events in terms of their relevance to work-related or GMAs.

To overcome these challenges, our study used a more nuanced approach, building on our previous work where we developed an application named Staff Hours [6]. This app harnessed GPS background data to automatically calculate users' work hours while also continuously recording screen events, including time stamps of notifications, screen on/off instances, and app use. This rich data collection led us to embrace the concept of "digital phenotyping"—the moment-by-moment quantification of the individual-level human phenotype in situ, using data from personal digital devices [5,9,11,12]. Using this innovative approach, our previous studies successfully identified various behaviors linked to digital interactions, such as sleep-wake cycles [13], circadian rhythms [14], and patterns of problematic smartphone use [15]. This methodology allowed for a more nuanced analysis of smartphone interactions, transcending the limitations of conventional experience-sampling studies. By continuously recording screen events, we were able to more accurately assess the complex relationship between smartphone interactions and the user's mental state, whether related to work or leisure.

In this study, we used standard wrist-worn actigraphy to assess sleep-wake cycles while simultaneously recording the time stamps of human-smartphone interaction patterns through our developed app. This app automatically captures the screen events associated with human-smartphone interaction. We devised a

XSL•FO
RenderX

method that uses an algorithm similar to actigraphy to calculate the biological rhythms derived from human-smartphone interaction, which we interpret as "general mental activity." Additionally, we developed a machine-learning algorithm that combines human-smartphone interaction patterns with GPS-defined work hours to generate probabilities indicating engagement in work-related tasks, which we interpret as "working mental activity." Our study aims to compare these 2 novel methods, GMA and WMA, developed by our team, with standard actigraphy serving as the reference for measuring PA. Both the GMA and WMA methods passively and continuously capture data, exhibiting similar characteristics to PA measurements. We would investigate the hypotheses that (1) the phase of the 3 activities would differ, (2) activity levels might vary between workdays and nonworkdays, and (3) we would identify mutual influences between activity levels, such as the impact of sleep on WMA and the influence of sleep duration on PA levels.

Methods

Participants

A total of 24 health care professionals (21 females; mean age 39.7, SD 8.3 years) were recruited from a workplace health promotion program for medical staff at 7 hospitals in northern Taiwan, for the study period between September 2021 and February 2023. Of the 24 participants in this study, 14 participants were health care professionals involved in primary care, including physicians, nurses, and dietitians. The remaining 10 participants were administrative staff not directly involved in patient care. All participants worked in hospital environments, with none engaged in remote work. Their job nature was classified as office work rather than physical work, and none of the participants were engaged in shift work during the study period. Prior to participation, all individuals provided informed consent. Eligible participants were required to have an Android operating smartphone that would be exclusively used by them throughout the study period.

Participants were instructed to install the Staff Hours app on their smartphones and wear a wrist actigraphy device continuously for a minimum of 1 week (Figure 1). Throughout the study, which involved 24 participants, a collective duration of 481 person-days was recorded. It is worth noting that there were 24 days during this period where no smartphone use data were recorded at all. Therefore, the analysis was conducted on a data set of 457 person-days, comprising 323 workdays and 115 nonworkdays.

Figure 1. The measurement methods in the study. (A) Physical activity was recorded using a wrist-worn actigraphy device. (B) Records of human-smartphone interaction were captured using the Staff Hours app. The recorded events included timestamps indicating the type of app used, screen on/off, and notifications. (C) The Staff Hours app also recorded the GPS location data, enabling the identification of work hours based on the user's location at their workplace and nonworkplace locations.

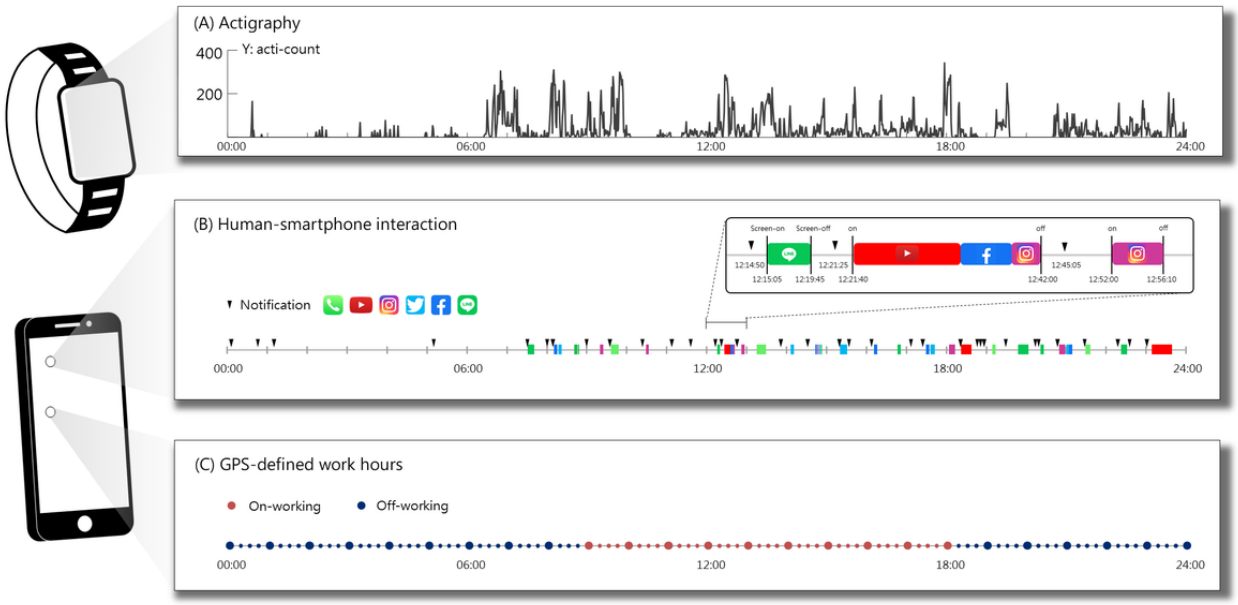

The study used only Android users due to technical limitations. The use of only Android smartphones was due to technical limitations on iOS devices, which at the time only allowed the collection of GPS data. Unlike Android, the iOS version did not support the retrieval of detailed human-smartphone interaction patterns such as time stamps of screen on/off, notifications, and app use types. These data points were critical for calculating GMA and WMA in our study. It was ensured that participants did not share their smartphones with other individuals during the study period. While wearing the actigraphy device was mandatory, no specific instructions were given regarding the placement of smartphones.

Ethical Considerations

The study adhered to the principles set forth in the Declaration of Helsinki. The institutional review boards of Chang-Gung Memorial Hospital (202100434B0A3 and 202002452A3) approved the research protocol. All participants provided informed consent prior to their inclusion in the study. They signed consent forms that specifically addressed the use and publication of their data. Regarding privacy and confidentiality,

XSL•FO
RenderX

all data were deidentified to ensure the protection of participant information. No additional compensation was provided to the participants for their involvement in this study. This decision was made clear to the participants during the informed consent process, ensuring transparency and fairness in the study's conduct.

Study Design

Physical Activity

Participants were instructed to wear a research-grade wrist actigraphy device (MiCorTM A100, MiTAC Inc, Taiwan) on their nondominant wrist. This device collected data on acceleration across 3 axes, with an internal sampling rate of 30 Hz. The actigraphy watches recorded accelerations along 3 axes, which were combined using the Euclidean distance of the accelerations' deviation from zero (Z). The data were then bandpass filtered within the range of 0.5-3 Hz. The Z values exceeding a predefined threshold were integrated over 2 seconds, and 1-second epochs of activity counts (acti-counts) were derived by averaging the integrated segments within 1 minute [1,16].

To estimate sleep and wake times, we used the standard Cole-Kripke algorithm on the acti-count data with slight modifications [16]. A crucial component of this algorithm, the minute-by-minute categorization of data into rest-active states, based on the weighted sum of the current minute with that of the contiguous minutes, was extracted to investigate the PA state during smartphone use. MATLAB (MathWorks) was used for the implementation of the algorithm using preexisting codes.

General Mental Activity

The Staff Hours app automatically recorded various smartphone events: GPS location data, time stamps of screen on/off events, notifications, and the labels of apps used in the background, all while maintaining a power-saving design. Using geofencing technology, it automatically logged the work hours spent in the workplace. The accuracy of the app-recorded work hours, with a sensitivity of 94.6% and specificity of 93.9%, was validated in our previous study [6]. Users were able to view their work hours on the app and, for enhanced accuracy and flexibility, adjust their actual work hours within a 7-day window.

These recordings formed time series of screen events, including notifications, screen on/off events, and app use. These data effectively charted human-smartphone interactions with significant temporal stability, as highlighted in [17]. Moreover, these time stamps reflected stimuli, individual responses, and content types typically examined in laboratory settings [5]. By harnessing the advanced data recording capabilities of smartphones, we successfully measured these interactions continuously and precisely in ecologically valid, real-world environments.

The use counts were defined as the number of apps used per minute (Figure 2). To account for the potentially patchy nature of app use behavior compared with activity data, we used longer durations to represent app use behaviors, referred to as "app-counts." App-counts were obtained by summing up the minute-by-minute use counts in 5-minute nonoverlapping epochs (288 epochs per day) to avoid excessive segments with zero counts. These data were then used to mimic the activity data obtained from the wrist actigraphy device.

Figure 2. Methods for calculating general mental activity (GMA) and working mental activity (WMA). Human-smartphone interaction records were analyzed to determine the frequency of app switches, used as a proxy for GMA. These data were then used to simulate activity levels comparable to those obtained from the wrist actigraphy device. WMA was assessed by calculating the probability of an individual being in work mode using human-smartphone interaction patterns and corresponding GPS location data.

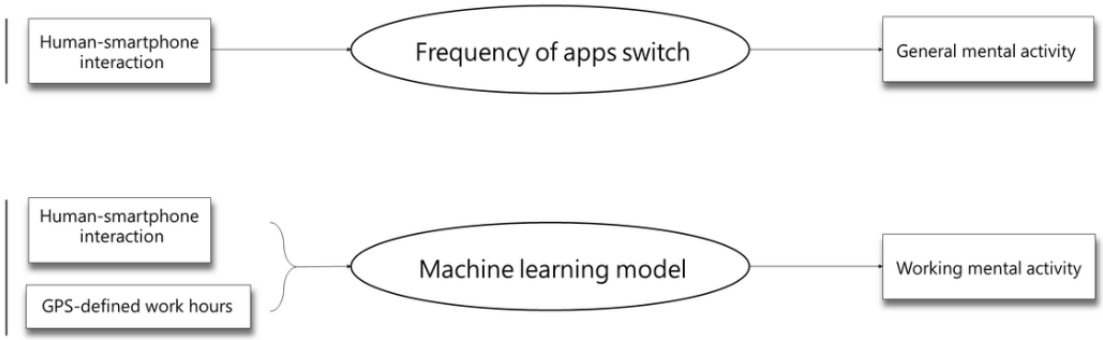

Working Mental Activity

To evaluate WMA, we developed a model that calculated the probability of an individual being in work mode, using human-smartphone interaction patterns and GPS location data.

The process was divided into 2 stages. In the first stage, extreme gradient–boosted trees were applied to transform interaction patterns into probabilities [18]. Subsequently, given a threshold, we calculated the length of consecutive time during which the probability values differed by less than the threshold before and after each time point. Ultimately, the transformed values were used as the grayscale values for that time. Data were then converted into 30×30–pixel images using 30-minute intervals as the unit of time. In the second stage, a 2D convolutional neural network model used these 30×30 grayscale images to train a binary classification model (distinguishing between "at work" and "off work"). The neural network architecture was based on ResNet-34 for feature extraction and was fine-tuned accordingly. Ultimately, the model could generate a sequence of continuous probabilities based on the previous probabilities in the sequence, as illustrated in Figure 3 and Multimedia Appendix 1.

XSL•FO
RenderX

Figure 3. The algorithm to generate working mental activity. Smartphone use logs, that is, human-smartphone interaction patterns, correlated with GPS data, establish the context of working and nonworking periods. An advanced machine learning model, integrating XGBoost and a 2D convolutional neural network (2D-CNN) processes these inputs. (Further details are also referred to in Multimedia Appendix 1).

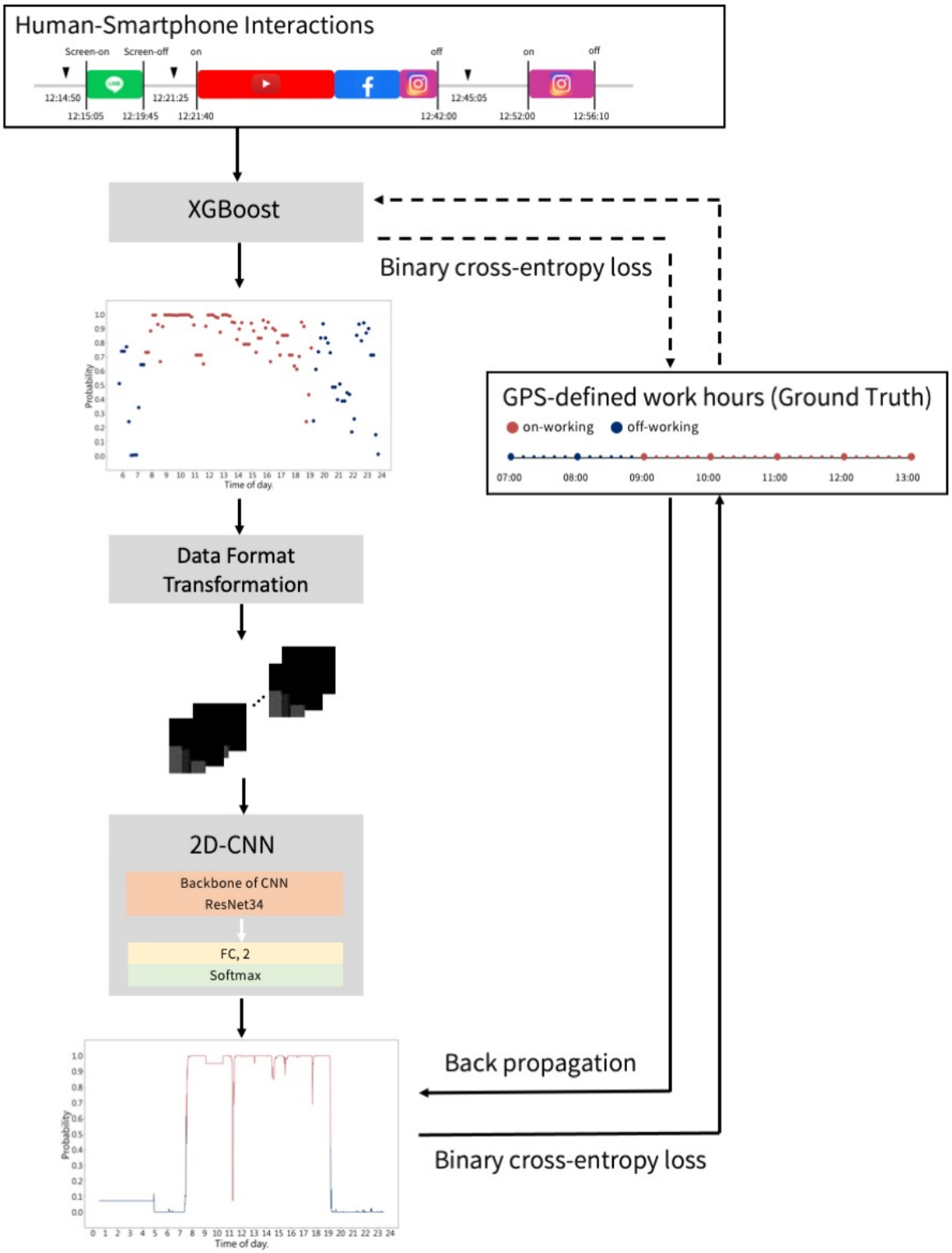

During the training phase for WMA, models were individually tailored for each participant. We excluded 5% (n=24) of person-days from our analysis due to the complete absence of smartphone use data on these days. For the training of individual user models, data were integrated by the minute, and a 10-minute time window was set to generate the training data set. Given the sequential nature of user data, we sorted it by date and divided it into person-days in an 80:20 ratio (80% for the training data set and 20% for the testing data set). As a result, the average size of the testing data set per user was 6.4 days, with the smallest data set comprising 2 days. The accuracy of the testing data set was minutely assessed, ensuring at least 2760 data points per user to calculate accuracy. We applied the same algorithm across data sets for 24 participants.

XSL•FO
RenderX

In the testing data set, the model demonstrated the following performance metrics: the average accuracy was 0.89 (SD 0.08), the average precision was 0.84, the average recall was 0.79 (SD 0.21), and the average F_1-score was 0.79 (SD 0.18). The average area under the receiver operating characteristic curve (AUC) per participant was 0.8, ranging from 0.62 to 0.97. AUC calculations were deliberately performed using the testing data set, excluding the training set, to prevent overfitting and demonstrate the model's robustness.

Machine learning algorithms were implemented using the Python scikit-learn library (Python Software Foundation), using 5-fold cross-validation. Missing values were imputed by sampling with replacement from nonmissing values. All data processing and analysis were performed using Python (version 3.8.5; Python Software Foundation), NumPy (version 1.19.2), scikit-learn (version 0.23.2), and PyTorch (version 2.0.1).

Activity Levels

In this study, we expanded the use of the M10 value as a representative indicator of the participant's activity levels, encompassing PA, GMA, and WMA, on a daily basis (Figure 4). The M10 value was derived by calculating the average acceleration during the 10-hour period with the highest activity level, providing insight into the participant's peak activity during the day. Through the application of the M10 value, we were able to consistently assess and compare the levels of PA, GMA, and WMA. This metric, widely acknowledged as an informative indicator of an individual's PA level in previous studies [19,20], was successfully used in both the assessment of GMA and WMA within this study. The statistical distributions of the main variables in this study ranged from –0.53 to 0.96 for skewness, and –1.13 to 1.06 for kurtosis, all falling within the normal and acceptable range (absolute values<1.5) for further statistical analysis.

Figure 4. Comparison of physical activity (PA), general mental activity (GMA), and working mental activity (WMA). The figure presents the data from a single participant for 2 consecutive days, one being a nonworkday (represented in blue) and the other being a workday (represented in red). The color-coded segments indicate the time periods during which the participant was engaged in work or nonwork activities. To assess activity levels, we used the M10 value as a representative indicator, which encompasses (A) PA, (B) GMA, and (C) WMA, on a daily basis. The M10 value was derived by calculating the average acceleration during the 10-hour period (highlighted in green) with the highest activity level, providing insights into the participant's peak activity during the day. Notably, there were significant differences observed in the M10 values of WMA between nonworkdays and workdays (0.030 vs 0.994). During workdays, the M10 period of WMA (7:39-17:39) occurred earliest, followed by PA (11:10-21:10), while GMA (13:32-23:32) exhibited the latest M10 period. Although the phase of PA is later than the phase of GMA, the calculated sleep periods (in orange) for both activities (23:43-05:28 vs 23:53-05:23) are remarkably similar.

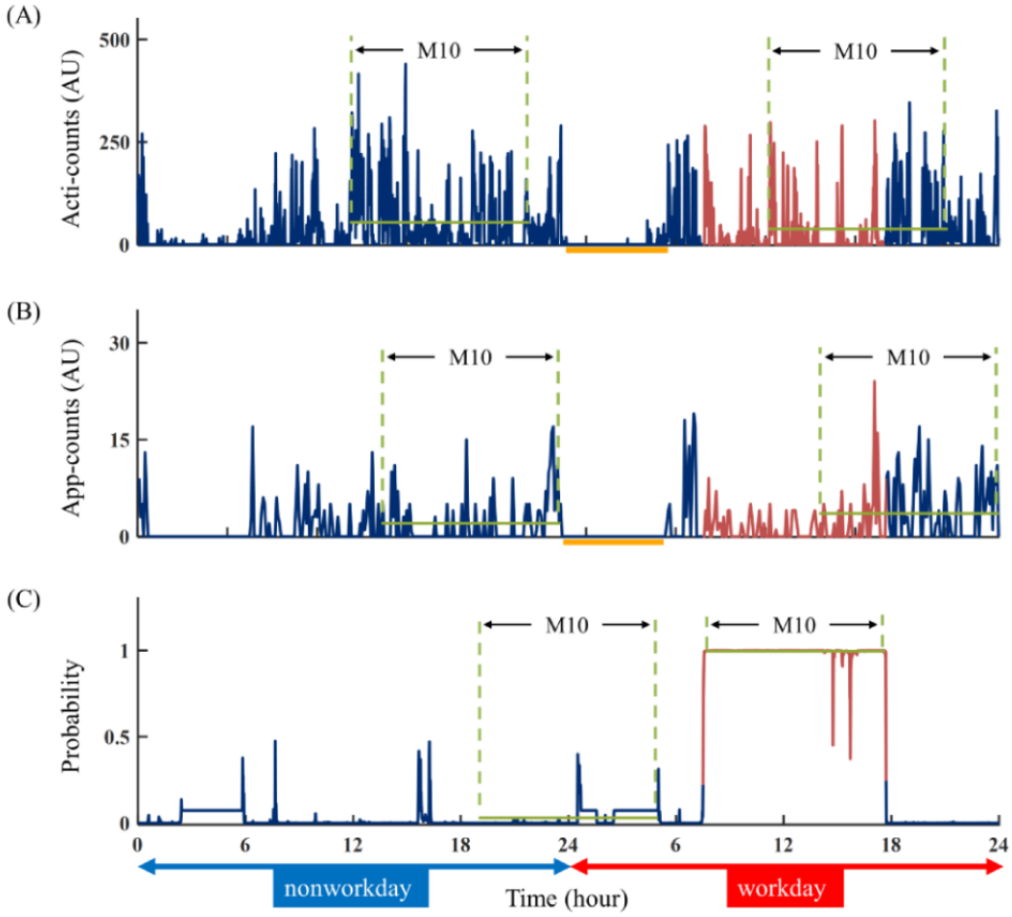

XSL•FO
RenderX

Statistical Analysis

To examine the impact of employment status on activity levels, several statistical analyses were conducted. An independent *t* test compared PA, GMA, and WMA levels (M10 value) between workdays and nonworkdays, aiming to identify significant differences based on employment status. Additionally, a paired *t* test assessed agreement between actigraphy-measured sleep indicators and app-measured counterparts, detecting variations and evaluating consistency. Another independent *t* test compared sleep indicators specifically between workdays and nonworkdays, exploring differences associated with employment status.

Phase difference quantifies the relative shift in phase between interacting signals r [21,22]. To calculate the phase difference (eg, PA and GMA), signals were filtered using a second-order Butterworth filter within a frequency range of 16-28 hours to extract circadian rhythms (Figure 5).

Figure 5. Phase difference calculation process. The illustrative example of calculating the circadian phase differences between physical activity and human-smartphone interaction patterns. (A) Raw data of physical activity (in blue) and general mental activity (in orange) generated by human-smartphone interaction patterns for a 7-day period of the same individual. (B) Filtered data for the cycles between 16 and 28 hours, retaining approximately 24-hour cycles of physical activity (in blue) and general mental activity (in orange). (C) Phase calculation using Hilbert transformation, showing the phase range of each cycle from –π to π. The graph shows that general mental activity (in orange) exhibits a later phase compared with physical activity (in blue). (D) Subtracting the phase of physical activity from the phase of general mental activity and converting the phase differences between physical activity and general mental activity to a 24-h time format (where 24 h=2π), the averaged phase difference reveals that general mental activity lags behind average physical activity by 268 min.

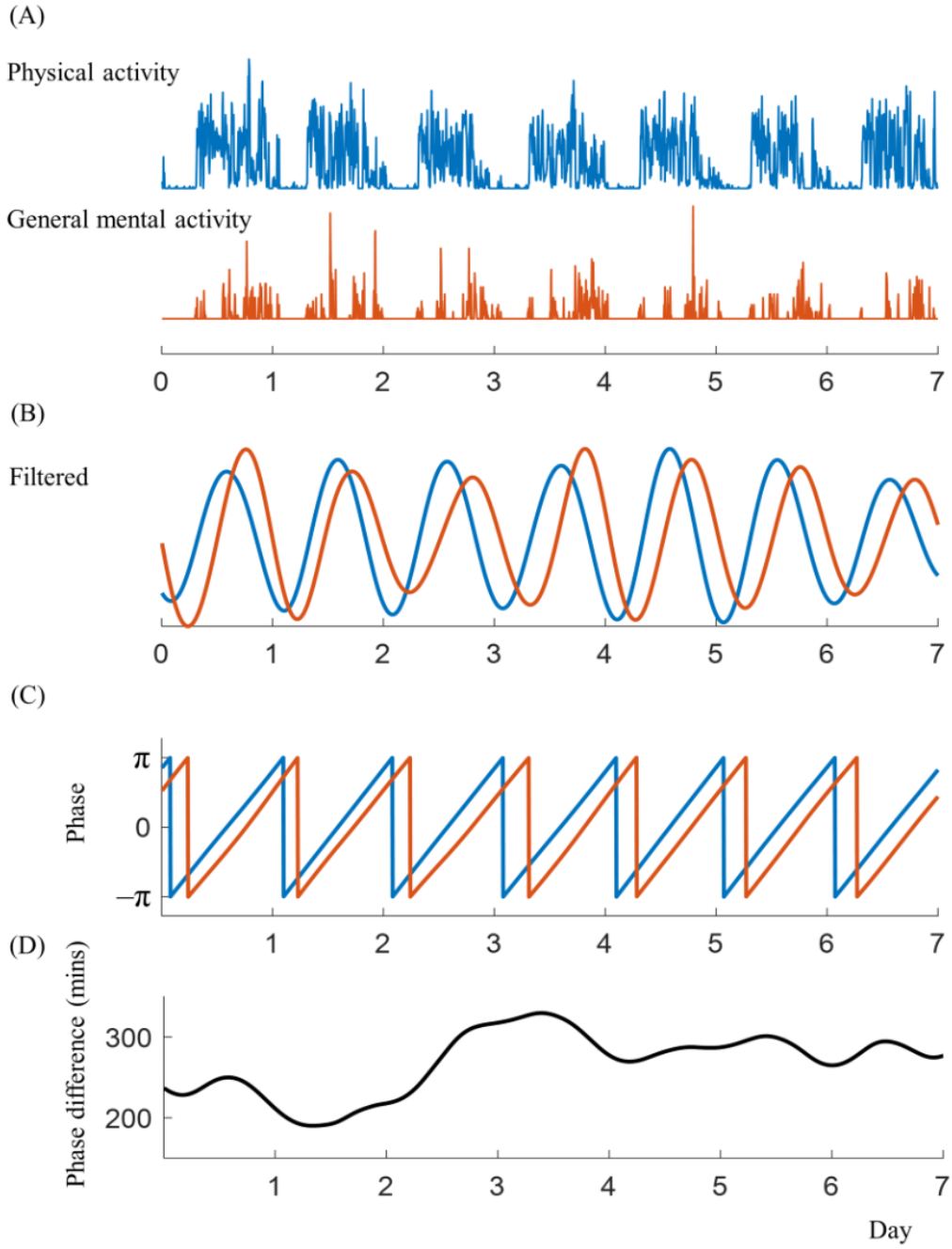

XSL•FO
RenderX

The filtered signal then undergoes a Hilbert transform, which can be represented as:

where * denotes the convolution operation. The Hilbert transform extracts the instantaneous phase of the filtered signal by getting the complex numbers of the resultant signals. Next, the phase of the reference signal is subtracted from the phase of the target signal to obtain the instantaneous phase difference:

For example, when the reference is the GMA signal and the target signal is PA signal, we can evaluate whether the circadian of PA is advanced or delayed in associated with that of GMA. Finally, to obtain a representative measure, averaging is performed on the time-varying phase difference signal over a specific time interval:

where *N* is the number of samples. This averaging step provides a more stable and meaningful measure of the overall phase relationship between the 2 signals, reducing the influence of short-term fluctuations or noise.

Considering the impact of sleep quality on daily performances, this study examined the effects of sleep indicators on mental activities. Using multilevel modeling via SPSS (version 25; IBM Corp) mixed model due to the nested data structure (daily information within each participant), we predicted the next day's PA, GMA, and WMA levels (M10) based on the previous night's sleep patterns (total sleep time and midpoint of sleep), controlling for participants' demographic backgrounds (gender and age). In the multilevel modeling, the previous night's sleep time, sleep midpoint, PA, GMA, and WMA levels are considered Level 1 data (within-person), while gender and age are considered Level 2 data (between-person).

Results

There were significant differences in activity levels between workdays and nonworkdays for WMA, as indicated by the M10 value (0.774, SD 0.214 vs 0.153, SD 0.166; *P*<.001). However, no significant differences were observed for PA (73.5, SD 22.5 vs 72.3, SD 23.3; *P*=.60) and GMA (2.9, SD 1.9 vs 2.9, SD 2.0; *P*=.81) between workdays and nonworkdays.

The phase differences among the 3 activity levels were found to be significant. WMA exhibited an earlier phase than PA by 1.08 (SD 1.22) hours (*P*<.001), while PA showed an earlier phase than GMA by 1.22 (SD 1.22) hours (*P*<.001). However, when comparing sleep onset (23:41, SD 1:11 vs 23:35, SD 1:17; *P*=.14), wake time (6:45, SD 1:17 vs 6:43, SD 1:14; *P*=.73), and total sleep time (416.1, SD 77.8 min vs 416.8, SD 81.6 min; *P*=.89) using PA and GMA, there were no significant differences observed. Therefore, the sleep time estimated by PA was chosen for further analysis since actigraphy is the standard measurement for calculating long-term sleep-wake cycles [1].

The comparison between PA and GMA on workdays and nonworkdays revealed no significant phase differences. However, the sleep-wake cycle on nonworkdays tended to be delayed compared with workdays, with a later sleep onset by 23.0 (SD 7.7) minutes (*P*=.003) and a later wake time by 37.5 (SD 8.5) minutes (*P*<.001). Nevertheless, the total sleep time did not significantly differ between nonworkdays and workdays (412.7, SD 72.5 vs 425.5, SD 90.7; *P*=.13).

Table 1 displays the associations among different mental activities and sleep indicators from the previous night (total sleep time and midpoint of sleep) while controlling the participants' demographics.

The results indicate a significant negative correlation between WMA's activity level and the previous night's midpoint of sleep. This suggests that a later bedtime and wake time were associated with lower activity levels in WMA on the following day. However, no significant correlation was found between WMA's activity level and total sleep time. Additionally, no significant correlations were observed between the activity level (M10 value) of PA and GMA and the sleep indicators from the previous night.

Table 1. Multilevel model coefficients (β) of activity level (M10) and sleep indicators in the previous day.

Characteristics	Physical activity		General mental activity		Working mental activity	
	β	*P* value	β	*P* value	β	*P* value
Midpoint of Sleep	–0.079	.06	–0.026	.42	–0.263	<.001
Total sleep time	0.056	.19	–0.009	.78	–0.089	.07
Gender[a]	0.670	.11	–0.684	.15	–0.187	.42
Age	–0.017	.30	–0.034	.07	–0.007	.48

[a]Gender was dummy-coded as 1=male and 2=female.

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we introduced a novel method based on human-smartphone interaction to assess GMA and WMA. Our approach offers a unique perspective on mental effort allocation, which has been a subject of scientific inquiry in cognitive psychology. Previous research has explored the labor or leisure tradeoff in cognitive control [3], often focusing on smartphone use at work as a proxy for leisure activity [8,10,23]. However, we argue that the definition of WMA is more explicit compared with the definition of leisure mental activity. Leisure tasks are typically described as unproductive and mentally undemanding relative to the cognitive demands of labor tasks [3]. Our study used human-smartphone interaction patterns, combined with workplace identification, to generate WMA measures. This approach provided a clear distinction between work-related and nonwork-related mental activity. By considering the specific context of WMA and incorporating smartphone interaction patterns within a defined workplace, our approach provides a more nuanced understanding of the allocation of mental effort.

Our study revealed a notable finding regarding the phase differences among WMA, PA, and GMA. Specifically, we observed that WMA occurred earlier in the day compared with PA, and PA occurred earlier than GMA. This finding correlates with the results reported by Dora et al [23], where participants were more likely to interact with their smartphones when fatigued or bored but did not necessarily use their smartphones for longer durations when experiencing higher levels of fatigue or boredom. The earlier occurrence of WMA compared with GMA can be explained by the possibility that participants, feeling fatigued from work, engaged in WMA activities before transitioning to GMA. This suggests that participants may have started using their smartphones after experiencing fatigue, corresponding to the active phase of WMA. However, once participants became actively engaged in GMA, characterized by smartphone use, there seemed to be a decrease in the activity levels of both PA and WMA. This finding is consistent with the results where participants reported increased fatigue and boredom after using their smartphones [23]. Consequently, following the active phase of GMA, the subsequent phase appears to be associated with sleep. These findings indicate a sequential pattern of mental and physical activities throughout the day.

In the method to describe WMA, we acknowledge the pervasive use of smartphones in modern-day work [7,10], particularly among health care professionals who heavily rely on these devices for their daily tasks. Hence, we used a machine learning model for WMA by associating human-smartphone interaction patterns with the context of the workplace. The introduction of smartphones has significantly transformed people's lives, blurring the boundaries between work and private life [7,10]. While Dora et al [10] examined the smartphone use characteristics of employees, distinguishing work-related use at home from private use at work, our machine learning approach offers a more personalized distinction based on human-smartphone interaction patterns. By considering the context-incongruent purpose of smartphone use for each individual, our method provides a more nuanced understanding of the interplay between work and private life. Furthermore, our approach complements the findings in previous research [10] that the current value of the smartphone did not significantly contribute to predicting switches from labor to leisure when considering the current task value. However, participants reacted strongly to naturally incoming notifications, which strongly influenced labor-to-leisure switches. In our study, human-smartphone interaction patterns, including screen on/off, app type, and notifications, allowed us to observe participants' smartphone use patterns and assess whether they reacted strongly to naturally incoming notifications, aligning with the findings in the previous study [10]. By using human-smartphone interaction patterns to generate WMA and considering the specific workplace context, our research methodology offers a more comprehensive understanding of the dynamics between work and private life in relation to smartphone use. These insights contribute to the ongoing discussion on the impact of smartphones on work-life boundaries and provide a basis for future investigations in this area.

Our study used an innovative method of assessing mental activity using human-smartphone interaction patterns. We directly computed GMA and WMA using this approach. Interestingly, we observed distinct patterns between GMA and WMA, with WMA exhibiting a significantly earlier phase compared with GMA, approximately 2 hours earlier. Furthermore, while GMA did not show differences between workdays and nonworkdays, WMA demonstrated significantly higher activity levels on workdays compared with nonworkdays. Although WMA was derived using a machine learning model, often referred to as a "black box," which may be challenging to validate directly, the significant difference in activity levels between workdays and nonworkdays provides supporting evidence that WMA indicator in this study represents WMA. Additionally, WMA was not solely influenced by work-related time periods. Our findings indicate that WMA activity levels were influenced by delayed bedtime and wake time. Specifically, if individuals have a later bedtime and wake time, their average WMA activity levels during the 10 most active hours of the day will be lower. This finding aligns with the previous research suggesting a link between late sleep schedules, and decreased work efficiency the following day [24-27], and supports the notion that WMA captures the impact of sleep patterns on working-related mental activity.

Limitations

Our study has several limitations that should be considered when interpreting the results. First, the small sample size and the use of convenience sampling focused exclusively on health care professionals may restrict the generalizability of our findings to other occupational groups. Additionally, the specific setting of the participants, who interacted with smartphones in a manner facilitated by GPS data for assessing WMA, might not accurately represent interaction patterns in a more diverse workforce where job roles and locations vary widely. Furthermore, the preliminary nature of these findings limits the scope and applicability of the research, suggesting that the conclusions drawn should be viewed as initial insights rather

XSL•FO
RenderX

than definitive results. This emphasizes the need for further studies with broader and more varied samples to validate and expand upon these initial findings. Additionally, the WMA assessment's reliance on the contrast between workplace and nonworkplace human-smartphone interactions might not directly translate to occupations without clear geographical boundaries or those that have experienced significant shifts to remote work arrangements. As we collected data during a period marked by the pandemic, from September 2021 to February 2023, our study inadvertently captured the implications of an extraordinary period where the health care professionals' work patterns were more stable compared with other professions affected by the pandemic's disruptions. It is, therefore, crucial to consider the extent to which the pandemic's impact on work patterns across different industries may influence the applicability of our findings. As industries gradually return to prepandemic operational modes, the generalizability of our results may require further validation. These insights have been integrated into the discussion section of our paper to elucidate the specific occupational context of our participants during the pandemic and to underscore the need for the cautious application of our results to the general working population. Second, 5% (n=24) of the person-days could not be included in the analysis of GMA and WMA. This could be attributed to participants turning off their smartphones throughout the day, using different smartphones at work and off work, or relying on personal computers or tablets as alternatives to smartphones. Consequently, the probability in work mode failed to accurately differentiate between on-working and off-working situations in these cases. Third, the use of the Staff Hours app limited the study's applicability to the Android operating system, and future research should consider compatibility with other operating systems. Fourth, we encountered challenges with potential interruptions in data collection from both the wrist-worn actigraphy device and the smartphone app, mainly due to low battery or nonuse by participants. To improve data robustness in future research, we plan to integrate the app with actigraphy. This approach aims to reduce data gaps and provide a more holistic perspective in our future studies. In addition, for activities where work locations are not fixed, our Staff Hours app allows for manual annotation, albeit this method is less consistent compared with the automatic 10-minute GPS logging due to potential human error or omissions. Fifth, the validation of the relationships between WMA, PA, and GMA was not fully achieved in this study. Despite observing no statistically significant differences in sleep onset and wake times measured by PA and GMA, GMA, when assessed against the widely accepted standard of actigraphy validated for PA, was confirmed to be reliable in assessing sleep dimensions. This result was also consistent with our previous research findings [28]. Furthermore, the performance of WMA in this study on the testing data set validated the use of human-smartphone interaction patterns to distinguish between work and nonwork locations, as the average AUC per participant was 0.80, ranging from 0.62 to 0.97. However, further research is needed to validate the relationships between WMA, PA, and GMA through interviews or additional qualitative methods with the participants. Finally, all participants were blinded to the locally stored records to minimize "biofeedback" effects, which could lead to self-recognition and behavioral modification. This blinding was crucial to ensure the authenticity of the data collected, as awareness of being monitored might have influenced the participants' natural behavior patterns. However, this approach also limited our ability to gather immediate, subjective feedback from participants about their experiences and the results obtained. While this might be seen as a limitation, it was a necessary tradeoff to preserve the integrity and ecological validity of the data. Future studies could explore the feasibility of incorporating participant feedback mechanisms that do not compromise the naturalistic observation essential to our research objectives. Despite these limitations, our study serves as a stepping stone toward comprehensive investigations into the neurobehavioral aspects of human-smartphone interaction.

Conclusions

In conclusion, our study introduced a novel approach to assessing mental activity using human-smartphone interaction patterns. The observed phase differences among WA, PA, and GMA shed light on the temporal organization of activities throughout the day.

Acknowledgments

We express our gratitude for Ms Hsiao-Han Lin's excellent technical assistance. This work was supported by the National Health Research Institutes of Taiwan (grant numbers 12A1-PHPP14-014), the Chang Gung Medical Foundation, Taiwan (grant numbers CORPG1L0011-12, CORPG1I0021), and National Science and Technology Council, Taiwan (grant number MOST 110-2314-B-182-062 [NMRPD1L1141]).

Data Availability

The data sets generated or analyzed during this study are not publicly available due to restrictions imposed by informed consent from participants in this study, which has been approved by the institutional review boards of the Chang Gung Memorial Hospital. However, they are available from the institutional review board of the Chang Gung Memorial Hospital for researchers who meet the criteria for access to confidential data.

XSL•FO
RenderX

Authors' Contributions

HHC and YHL contributed to the study's conception and design. HHC, CL, HCC, and JHC performed data generation and statistical analysis. HHC, CL, HHC, and YHL drafted the manuscript, conceived and designed the study, and acquired data. All authors contributed to data analysis and interpretation and were involved in reviewing, editing, and approval of the final revision. YHL is responsible for the integrity of the work as a whole.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Development and evaluation of a convolutional neural network model for predicting work mode based on human-smartphone interaction data.
[DOCX File , 16 KB - jmir_v26i1e56144_app1.docx]

References

1. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
2. Witting W, Kwa IH, Eikelenboom P, Mirmiran M, Swaab DF. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. Biol Psychiatry 1990;27(6):563-572. [doi: 10.1016/0006-3223(90)90523-5] [Medline: 2322616]
3. Kool W, Botvinick M. A labor/leisure tradeoff in cognitive control. J Exp Psychol Gen 2014;143(1):131-141 [FREE Full text] [doi: 10.1037/a0031048] [Medline: 23230991]
4. Kool W, Botvinick M. Mental labour. Nat Hum Behav 2018;2(12):899-908. [doi: 10.1038/s41562-018-0401-9] [Medline: 30988433]
5. Chen IM, Chen YY, Liao SC, Lin YH. Development of digital biomarkers of mental illness via mobile apps for personalized treatment and diagnosis. J Pers Med 2022;12(6):936 [FREE Full text] [doi: 10.3390/jpm12060936] [Medline: 35743722]
6. Chiang TW, Chen SY, Pan YC, Lin YH. Automatic work-hours recorder for medical staff (Staff Hours): mobile app development. JMIR Mhealth Uhealth 2020;8(2):e16063 [FREE Full text] [doi: 10.2196/16063] [Medline: 32130165]
7. Derks D, van Duin D, Tims M, Bakker AB. Smartphone use and work–home interference: the moderating role of social norms and employee work engagement. J Occup Organ Psychol 2014;88(1):155-177. [doi: 10.1111/joop.12083]
8. Dora J, van Hooff M, Geurts S, Kompier M, Bijleveld E. Labor/leisure decisions in their natural context: the case of the smartphone. Psychon Bull Rev 2021;28(2):676-685 [FREE Full text] [doi: 10.3758/s13423-020-01844-2] [Medline: 33219457]
9. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry 2018;17(3):276-277 [FREE Full text] [doi: 10.1002/wps.20550] [Medline: 30192103]
10. Dora J, van Hooff MLM, Geurts SAE, Hooftman WE, Kompier MAJ. Characterizing work-related smartphone use at home and private smartphone use at work using latent class analysis. Occup Health Sci 2019;3(2):187-203 [FREE Full text] [doi: 10.1007/s41542-019-00040-6]
11. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
12. Lin YH, Lin YC, Lee YH, Lin PH, Lin SH, Chang LR, et al. Time distortion associated with smartphone addiction: identifying smartphone addiction via a mobile application (App). J Psychiatr Res 2015;65:139-145. [doi: 10.1016/j.jpsychires.2015.04.003] [Medline: 25935253]
13. Lin YH, Wong BY, Lin SH, Chiu YC, Pan YC, Lee YH. Development of a mobile application (App) to delineate "digital chronotype" and the effects of delayed chronotype by bedtime smartphone use. J Psychiatr Res 2019;110:9-15 [FREE Full text] [doi: 10.1016/j.jpsychires.2018.12.012] [Medline: 30611008]
14. Lin YH, Wong BY, Pan YC, Chiu YC, Lee YH. Validation of the mobile app-recorded circadian rhythm by a digital footprint. JMIR Mhealth Uhealth 2019;7(5):e13421 [FREE Full text] [doi: 10.2196/13421] [Medline: 31099340]
15. Lin YH, Lin YC, Lin SH, Lee YH, Lin PH, Chiang CL, et al. To use or not to use? compulsive behavior and its role in smartphone addiction. Transl Psychiatry 2017;7(2):e1030 [FREE Full text] [doi: 10.1038/tp.2017.1] [Medline: 28195570]
16. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992;15(5):461-469 [FREE Full text] [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
17. Pan YC, Lin HH, Chiu YC, Lin SH, Lin YH. Temporal stability of smartphone use data: determining fundamental time unit and independent cycle. JMIR Mhealth Uhealth 2019;7(3):e12171 [FREE Full text] [doi: 10.2196/12171] [Medline: 30912751]
18. Chen HH, Lu HHS, Weng WH, Lin YH. Developing a machine learning algorithm to predict the probability of medical staff work mode using human-smartphone interaction patterns: algorithm development and validation study. J Med Internet Res 2023;25:e48834 [FREE Full text] [doi: 10.2196/48834] [Medline: 38157232]

XSL•FO
RenderX

19. Jones SE, van Hees VT, Mazzotti DR, Marques-Vidal P, Sabia S, van der Spek A, et al. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. Nat Commun 2019;10(1):1585 [FREE Full text] [doi: 10.1038/s41467-019-09576-1] [Medline: 30952852]
20. Gonçalves BSB, Adamowicz T, Louzada FM, Moreno CR, Araujo JF. A fresh look at the use of nonparametric analysis in actimetry. Sleep Med Rev 2015;20:84-91. [doi: 10.1016/j.smrv.2014.06.002] [Medline: 25065908]
21. Yamazaki S, Kerbeshian MC, Hocker CG, Block GD, Menaker M. Rhythmic properties of the hamster suprachiasmatic nucleus in vivo. J Neurosci 1998;18(24):10709-10723 [FREE Full text] [doi: 10.1523/JNEUROSCI.18-24-10709.1998] [Medline: 9852606]
22. Lin C, Lin PF, Wang CH, Juan CH, Tran TT, Pham VT, et al. Probing age-related changes in cardio-respiratory dynamics by multimodal coupling assessment. Chaos 2020;30(3):033118 [FREE Full text] [doi: 10.1063/1.5134868] [Medline: 32237792]
23. Dora J, van Hooff M, Geurts S, Kompier M, Bijleveld E. Fatigue, boredom and objectively measured smartphone use at work. R Soc Open Sci 2021;8(7):201915 [FREE Full text] [doi: 10.1098/rsos.201915] [Medline: 34295513]
24. Eliasson AH, Lettieri CJ, Eliasson AH. Early to bed, early to rise! sleep habits and academic performance in college students. Sleep Breath 2010;14(1):71-75. [doi: 10.1007/s11325-009-0282-2] [Medline: 19603214]
25. Sonnentag S, Cheng BH, Parker SL. Recovery from work: advancing the field toward the future. Annu Rev Organ Psychol Organ Behav 2022;9(1):33-60 [FREE Full text] [doi: 10.1146/annurev-orgpsych-012420-091355]
26. Medeiros ALD, Mendes DBF, Lima PF, Araujo JF. The relationships between sleep-wake cycle and academic performance in medical students. Biological Rhythm Research 2010;32(2):263-270. [doi: 10.1076/brhm.32.2.263.1359]
27. Hennig T, Krkovic K, Lincoln TM. What predicts inattention in adolescents? an experience-sampling study comparing chronotype, subjective, and objective sleep parameters. Sleep Med 2017;38:58-63 [FREE Full text] [doi: 10.1016/j.sleep.2017.07.009] [Medline: 29031757]
28. Lin C, Chen IM, Chuang HH, Wang ZW, Lin HH, Lin YH. Examining human-smartphone interaction as a proxy for circadian rhythm in patients with insomnia: cross-sectional study. J Med Internet Res 2023;25:e48044 [FREE Full text] [doi: 10.2196/48044] [Medline: 38100195]

Abbreviations

acti-count: activity count
AUC: area under the curve
GMA: general mental activity
PA: physical activity
WMA: working mental activity

Edited by S Ma, T Leung; submitted 08.01.24; peer-reviewed by M Stach, LA Lee; comments to author 09.04.24; revised version received 17.04.24; accepted 18.04.24; published 17.06.24.

Please cite as:
Chen HH, Lin C, Chang HC, Chang JH, Chuang HH, Lin YH
Developing Methods for Assessing Mental Activity Using Human-Smartphone Interactions: Comparative Analysis of Activity Levels and Phase Patterns in General Mental Activities, Working Mental Activities, and Physical Activities
J Med Internet Res 2024;26:e56144
URL: https://www.jmir.org/2024/1/e56144
doi:10.2196/56144
PMID:38885499

©Hung-Hsun Chen, Chen Lin, Hsiang-Chih Chang, Jen-Ho Chang, Hai-Hua Chuang, Yu-Hsuan Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Rest-Activity Rhythm Differences in Acute Rehabilitation Between Poststroke Patients and Non–Brain Disease Controls: Comparative Study

Huey-Wen Liang[1,2,3], MD, PhD; Chueh-Hung Wu[1,2], MD, PhD; Chen Lin[4], PhD; Hsiang-Chih Chang[5], PhD; Yu-Hsuan Lin[5,6,7], MD, PhD; Shao-Yu Chen[3], MD; Wei-Chen Hsu[1], MD

[1]Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu, Taiwan
[2]Department of Physical Medicine and Rehabilitation, College of Medicine, National Taiwan University, Taipei, Taiwan
[3]Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, Taipei, Taiwan
[4]Department of Biomedical Sciences and Engineering, National Central University, Taoyuan City, Taiwan
[5]Institute of Population Health Sciences, National Health Research Institutes, Miaoli County, Taiwan
[6]Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan
[7]Department of Psychiatry, College of Medicine, National Taiwan University, Taipei, Taiwan

Corresponding Author:
Yu-Hsuan Lin, MD, PhD
Institute of Population Health Sciences
National Health Research Institutes
No 35 Keyan Road
Miaoli County, 35053
Taiwan
Phone: 886 37 206166 ext 36383
Email: yuhsuanlin@nhri.edu.tw

Abstract

Background: Circadian rhythm disruptions are a common concern for poststroke patients undergoing rehabilitation and might negatively impact their functional outcomes.

Objective: Our research aimed to uncover unique patterns and disruptions specific to poststroke rehabilitation patients and identify potential differences in specific rest-activity rhythm indicators when compared to inpatient controls with non–brain-related lesions, such as patients with spinal cord injuries.

Methods: We obtained a 7-day recording with a wearable actigraphy device from 25 poststroke patients (n=9, 36% women; median age 56, IQR 46-71) and 25 age- and gender-matched inpatient control participants (n=15, 60% women; median age 57, IQR 46.5-68.5). To assess circadian rhythm, we used a nonparametric method to calculate key rest-activity rhythm indicators—relative amplitude, interdaily stability, and intradaily variability. Relative amplitude, quantifying rest-activity rhythm amplitude while considering daily variations and unbalanced amplitudes, was calculated as the ratio of the difference between the most active 10 continuous hours and the least active 5 continuous hours to the sum of these 10 and 5 continuous hours. We also examined the clinical correlations between rest-activity rhythm indicators and delirium screening tools, such as the 4 A's Test and the Barthel Index, which assess delirium and activities of daily living.

Results: Patients who had a stroke had higher least active 5-hour values compared to the control group (median 4.29, IQR 2.88-6.49 vs median 1.84, IQR 0.67-4.34; P=.008). The most active 10-hour values showed no significant differences between the groups (stroke group: median 38.92, IQR 14.60-40.87; control group: median 31.18, IQR 18.02-46.84; P=.93). The stroke group presented a lower relative amplitude compared to the control group (median 0.74, IQR 0.57-0.85 vs median 0.88, IQR 0.71-0.96; P=.009). Further analysis revealed no significant differences in other rest-activity rhythm metrics between the two groups. Among the patients who had a stroke, a negative correlation was observed between the 4 A's Test scores and relative amplitude (ρ=–0.41; P=.045). Across all participants, positive correlations emerged between the Barthel Index scores and both interdaily stability (ρ=0.34; P=.02) and the most active 10-hour value (ρ=0.42; P=.002).

Conclusions: This study highlights the relevance of circadian rhythm disruptions in poststroke rehabilitation and provides insights into potential diagnostic and prognostic implications for rest-activity rhythm indicators as digital biomarkers.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e49530) doi:10.2196/49530

KEYWORDS

circadian rhythms; stroke rehabilitation, rest-activity rhythms, relative amplitude, delirium screening, interdaily stability

Introduction

Sleep-wake cycle disruptions are prevalent in patients who had a stroke and are marked by disturbances in sleep architecture and circadian rhythms [1,2]. These disruptions often lead to fragmented sleep patterns and increase nighttime wakefulness. Poor sleep quality negatively impacts neuroplasticity and memory consolidation post stroke [3,4], potentially affecting recovery outcomes. Additionally, it correlates with reduced daily motor function [5], diminishes physical activity levels, and increases fatigue during inpatient rehabilitation [6]. Higher levels of sleep disruption are tied to a slower recovery of functional independence and motor skills during inpatient rehabilitation [7,8].

During hospital stays, sleep-wake cycle disturbances are frequent, often because of clinical care interventions and a noisy environment [9]. Such impacts are particularly detrimental to stroke survivors, hampering neurological recovery and draining the energy necessary for optimal rehabilitation. Despite the high prevalence of sleep-wake cycle in patients who had a stroke, many questions about the predictors of these disorders, their link to stroke characteristics and severity, and their progression over time remain [10]. Sleep-wake disturbances encompass various factors crucial to the rehabilitation process, such as activity levels, sleep-wake cycles, rest-activity rhythms (RARs), and delirium. These factors often lead to severe circadian rhythm disruptions [11,12], which could directly and negatively influence rehabilitation outcomes by diminishing engagement and impairing learning and consolidation.

Neurocognitive disorders, mainly delirium and dementia [13], represent significant challenges within the realm of poststroke rehabilitation, primarily due to their shared core manifestation of cognitive dysfunction often coupled with disturbances in the sleep-wake cycle. Although dementia typically progresses slowly and over an extended period, delirium presents with an acute onset, making their management particularly complex in the context of stroke recovery. The significance of addressing dementia and delirium in poststroke rehabilitation becomes evident when considering their profound impact on the recovery process. Delirium, with an occurrence rate ranging from 21% to 32% in patients with acute stroke [14], poses a considerable hurdle to rehabilitation efforts. Defined by the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [13] as an acute and fluctuating disturbance of attention, cognition, or consciousness, delirium not only complicates the clinical picture but also predicts a poor functional outcome for patients who have a stroke [15]. The diagnostic challenge lies in distinguishing delirium from preexisting neurocognitive disorders, underscoring the importance of early identification. The coexistence of sleep-wake disturbance with delirium further exacerbates the situation, emphasizing the critical need for vigilant monitoring and management strategies. Using tools such as actigraphy, which effectively documents altered patterns of motor activity or immobility [16-18], presents an opportunity for early detection and intervention [19]. Despite its proven utility in various clinical settings, its application in patients who had a stroke and experienced delirium remains underexplored [20], highlighting a gap in current research. Moreover, the pioneering work of Witting et al in 1990 laid the foundation for analyzing RARs in neurodegenerative disorders, including dementia [21]. By quantifying parameters such as interdaily stability (IS) and intradaily variability (IV), their framework provided valuable insights into the relationship between RARs and aging-related cognitive decline. This expanded interest in RARs underscores the importance of considering sleep-wake disturbances in the context of dementia and delirium within poststroke rehabilitation.

However, evaluations of sleep-wake disturbances have not been comprehensive within the clinical framework. Most studies primarily focused on sleep and traditionally depended on subjective assessments, such as self-reported questionnaires, daily sleep records, or observations by health care professionals. Objective assessments like polysomnography were also used. Yet these methods did not characterize the activity levels throughout the entire day, which would reflect the full impact of sleep-wake cycle disturbances. Recent advancements in wearable technology have made continuous and noninvasive monitoring of physiological outputs, including motor activity, possible. Ambulatory motor activity recordings, such as RARs based on actigraphy, have proven instrumental in understanding the changes in daily rhythms in patients who had a stroke [11,22,23] and in assessing stroke risk and poststroke outcomes [24]. Considering the well-documented long-term adverse outcomes in poststroke survivors, including a heightened risk of disability and mortality, grasping the relationship between altered RARs and these outcomes is essential for long-term management of post-stroke recovery and rehabilitation [25,26].

While past research has explored the sleep-wake cycle for patients undergoing rehabilitation post stroke, most studies have compared them to age- and gender-matched control groups of healthy community-dwelling adults [11,12]. This design did not account for potential confounding factors like significantly reduced physical activity levels and the influences of hospitalization and acute illness. Therefore, a gap exists in the literature concerning research that included inpatient control participants, especially those with reduced mobility.

Both poststroke patients and those with non–brain-related conditions who were admitted for rehabilitation likely experienced circadian rhythm disruptions. Given that patients who had a stroke faced added challenges from brain injuries, specific RAR indicators were expected when they were compared to inpatient controls with non–brain-related lesions, such as those with spinal cord or orthopedic injuries. This study mainly focused on poststroke rehabilitation patients, compared with the age-matched control group consisting of individuals without brain injuries undergoing rehabilitation. Both groups

XSL•FO
RenderX

likely experienced similar reductions in physical activity levels during rehabilitation, providing a controlled setting to investigate RAR indicators specific to poststroke rehabilitation. The primary goal was to analyze RAR indicators between these groups, emphasizing patterns unique to poststroke patients. Additionally, the aim was to correlate RAR indicators with activities of daily living (ADL) and delirium screening tools, ultimately validating the RAR indicators as potential digital biomarkers with diagnostic and prognostic implications.

Methods

Study Design and Participants

This study was part of a prospective investigation into the changes and impacts of circadian rhythm and depression on the functional outcomes for patients who had a stroke. We recruited participants from rehabilitation wards in 2 affiliated hospitals of a university in northern Taiwan, with admissions occurring between February and August 2023. They were invited within the first week following their admission to the rehabilitation wards. To qualify, participants of the stroke group needed to be aged ≥20 years, with a confirmed diagnosis of stroke by brain computerized tomography or magnetic resonance imaging, and had been admitted for intensive rehabilitation. The criteria for inclusion in the inpatient control group were being aged ≥20 years and having been admitted for intensive rehabilitation for diagnoses other than brain conditions. Both groups were required to be medically and neurologically stable and to have been healthy and independent before the current admission. Common exclusion criteria for both groups included the following: having an acute illness or unstable vital signs, having an underlying terminal disease, having previous brain conditions (eg, Parkinson disease, stroke, or dementia), being pregnant, and being unable or refusing to use a wearable device for 7 days. Inpatient control participants were age-matched (±3 years) with the stroke group participants.

A power analysis conducted using G*Power (version 3.1; Heinrich-Heine University) was used to determine the required sample size to compare the two groups. The Type I error rate, denoted as alpha (α), was set to a value of .05, while the power was set to 0.95 for a 2-tailed comparison. A total sample size of 50 (25 per group) was required according to the mean (SD) obtained from a previous study using an actigraphy device to evaluate circadian rhythm [27].

Ethical Considerations

The study adhered to the Declaration of Helsinki, and the protocol received approval from the ethical committee of the National Taiwan University Hospital on January 11, 2023 (approval number 202212047RINA). All participants had provided written informed consent before participation and were given the opportunity to opt out at any time. They signed consent forms that specifically addressed the use and publication of their data. Regarding privacy and confidentiality, all data were deidentified to ensure the protection of participant information. No additional compensation was provided to the participants for their involvement in this study. This decision was made clear to the participants during the informed consent process, ensuring transparency and fairness in the study's conduct.

Data Collection

Medical Records

We thoroughly reviewed the medical records of all participants. This review covered their demographic details, previous medical and personal histories, specifics about their current condition leading to rehabilitation, medication history—especially concerning sedatives and hypnotics, and their ADL as measured by the Barthel Index (BI) at the time of admission. The data were retrieved from the nurses' records as a routine evaluation in the rehabilitation wards. The BI is a commonly used and reliable tool to measure functional disability in neurorehabilitation [28,29]. There are 10 items regarding the level of assistance required for basic ADL and the sphincter control. The total scores range from 0 to 100, with a higher score indicating a higher independence.

For those diagnosed with a stroke, we delved deeper, collecting data about the stroke type (infarction or hemorrhage) and category as proposed by the Oxfordshire Community Stroke Project [30]. We also noted any interventions they underwent during the acute stage, such as the administration of recombinant tissue plasminogen activator or mechanical thrombectomy. A particular focus of our review was on the medication history, emphasizing benzodiazepines and antipsychotics, given their potential sleep disturbances or delirium during the study.

Shortly after participants began wearing the actigraphy device, within the first week, nurses screened patients with stroke for delirium using the 4 A's Test (4AT). Recognized as a rapid delirium assessment instrument, the 4AT consists of 4 items regarding alertness, cognition, attention, and any sudden changes or fluctuations in these areas [31]. The total score ranges from 0 to 12, and a score of 4 or more suggests delirium. It is not a diagnostic tool, but excellent reliability and validity is established through multiple studies, including among patients who had a stroke [32]. Before the onset of the project, all involved nurses underwent a 30-minute training session to ensure they were familiar with the screening's purpose and procedures. To ensure the process's accuracy and adherence to standards, a senior nurse, well versed in the evaluation, observed the initial screenings.

Actigraphy and RAR measures

Participants were instructed to wear the research-grade wrist actigraphy device (MiCor A100, MiTAC Inc) for a minimum of 7 days. For patients who had a stroke, the actigraphy device was placed on the unaffected side to minimize the impact of limb weakness. For inpatient controls, it was attached to the nondominant side, which most participants preferred. This placement choice was supported by research suggesting that the specific wrist used for actigraphy does not significantly affect the measurement of key sleep variables [33]. Additionally, it was noted that none of the participants experienced bilateral paralysis. Accelerations along 3 axes were collected by the actigraphy watches, combined using Euclidean distance of the accelerations' deviation from 0, and bandpass filtered from 0.5-3 Hz. The 0 values exceeding a predefined threshold were integrated within 2 seconds, and the activities of 1-second

XSL•FO
RenderX

epochs (acti-counts) were derived from averaging the integrated segments within 1 minute [21,34].

We used nonparametric approaches to quantify RARs. The nonparametric method was used to calculate 3 of the circadian rhythm indicators [35]: relative amplitude (RA), IS, and IV.

RA, the ratio of the differences between the most active 10 continuous hours (M10) and the least active 5 continuous hours (L5) over the summation of M10 and L5, was calculated to measure the amplitude of RARs while considering the daily variations and unbalanced amplitudes of peak and trough in daily activity rhythms, as follows:

IS quantified the stability of the rhythms between days, that is, the coupling strength of the rhythms with supposedly stable environmental factors. It could vary between 0 and 1, with higher values indicating more stable daily rhythms, as follows:

where N represents the number of hourly averaged acti-counts, indicating the total data points across all days considered; p is the number of data points in the daily circadian template, such as 24 points for an hourly template; X_h denotes the template values at each hour; and X_i refers to the observed values at each hour.

IV indicated the fragmentation of the rhythms, that is, the frequency and extent of transitions between rest and activity. It could vary roughly between 0 and 2, with higher values indicating higher fragmentations, as follows:

where N denotes the total number of hourly averaged acti-counts, and X_i indicates the values observed at each hour.

RA, IS, and IV were calculated for a minimum period of 1 week [35].

Physical Activity

The physical activity level of each participant was quantified by using an accelerometer to measure their daily movements. Key features were calculated for each day, including the M10 and L5 values. The M10 and L5 represent the 10 hours of the day when the participant was most active and the 5 hours when they were least active, respectively. These are widely recognized indicators of a person's circadian activity patterns [36,37]. To determine the M10, a 10-hour moving average was used to estimate the period of the day with the highest average acceleration, which was considered to be the participant's overall physical activity level.

Statistical Analysis

The normality of distribution for all circadian parameters derived from the actigraphy was assessed using the Shapiro-Wilk test, although it is noteworthy that in health sciences, most variables do not adhere to a normal distribution due to the nature of the field. Depending on the data distribution, descriptive statistics were either reported as means (SDs) or medians (IQR). To investigate relationships among circadian parameters, BI scores, and 4AT scores, the Spearman ρ correlation coefficient (for nonnormally distributed data) were used. It is important to note that while the Pearson correlation coefficient r measures linear relationships in variables following a normal distribution, the use of Spearman ρ is more appropriate for our data set due to its nonnormal distribution. Comparative analyses between the stroke and inpatient control groups, including demographic features, BI scores, 4AT scores, and circadian parameters, were conducted. For nonnormally distributed data, the Mann-Whitney U test was used, while the independent 2-tailed t test was used for normally distributed data. Furthermore, categorical variables were compared using the chi-square test, with frequencies and percentages reported accordingly.

All statistical computations were executed using SPSS for Windows (version 15.0; IBM Inc). A threshold of $P<.05$ was determined for statistical significance.

Results

As depicted in Figure 1, a total of 25 participants from each group were analyzed. Since the variance of the data was mostly unequal between the two groups, nonparametric methods were used for statistical analysis. The primary diagnoses for the inpatient control group consisted of traumatic spinal cord injuries (n=7, 28%), nontraumatic myelopathy (n=12, 48%), and other neurological diseases (n=6, 24%). Figure 2 provides sample recordings from 2 participants of each group. Age, sex, and other demographic parameters were comparable between the groups according to the Mann-Whitney U test, chi-square test, or Fisher exact test (Table 1). The stroke group demonstrated a significantly higher prevalence of hypertension and hyperlipidemia. There was no statistically significant difference in the proportion of previous medication history or current medication with sleep medications or antipsychotics (predominantly quetiapine) between the two groups. The average BI scores for both were below 40, indicating pronounced dependence.

XSL•FO
RenderX

Figure 1. The CONSORT (Consolidated Standards of Reporting Trials) charts of the study participants.

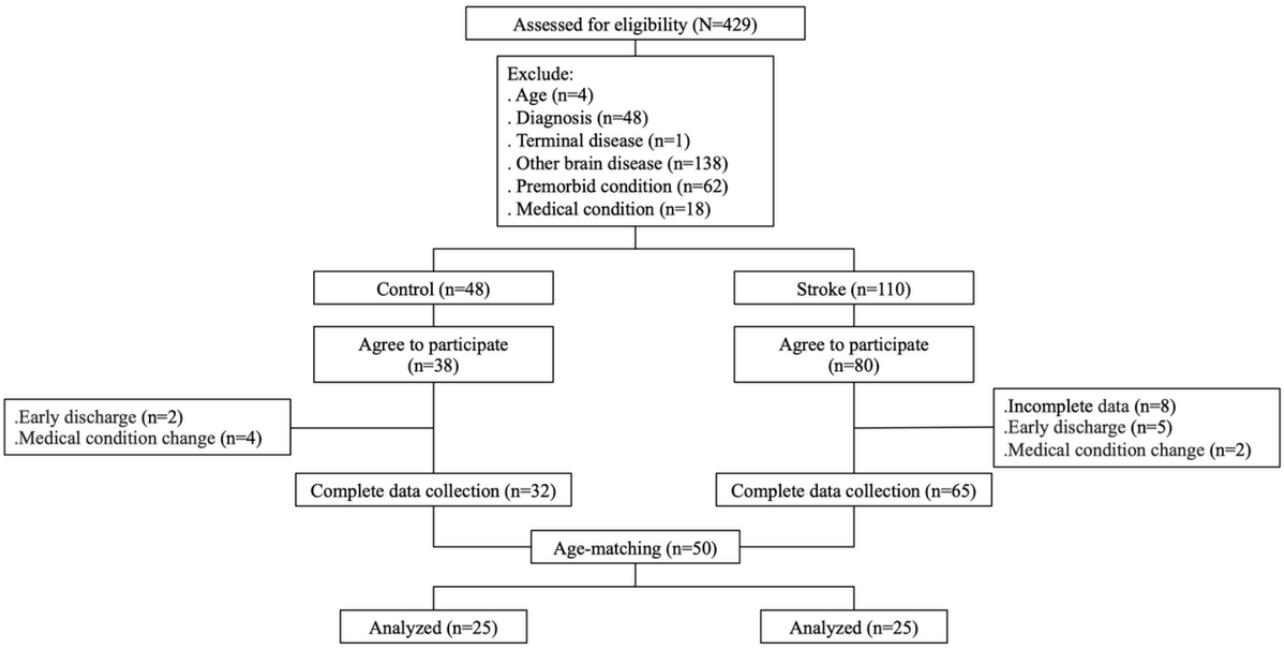

Figure 2. Seven-day rest-activity rhythm analysis of a patient who had a stroke versus an inpatient control. On the left, daily data over the span of a week is presented. Red lines represent the most active 10 continuous hours (M10), while blue lines denote the least active 5 continuous hours (L5), both of which are marked for every 24 hours. On the right, average 7-day data display the composite values of M10, L5, and the relative amplitude (RA), calculated as RA = (M10 – L5) / (M10 + L5). For the stroke group patient, the data are as follows: intradaily variability (IV)=1.55; interdaily stability (IS)=0.57; L5=9.67; M10=38.91; and RA=0.60. In comparison, the inpatient control group presents with the following data: IV=1.05; IS=0.66; L5=1.84; M10=32.48; and RA=0.89.

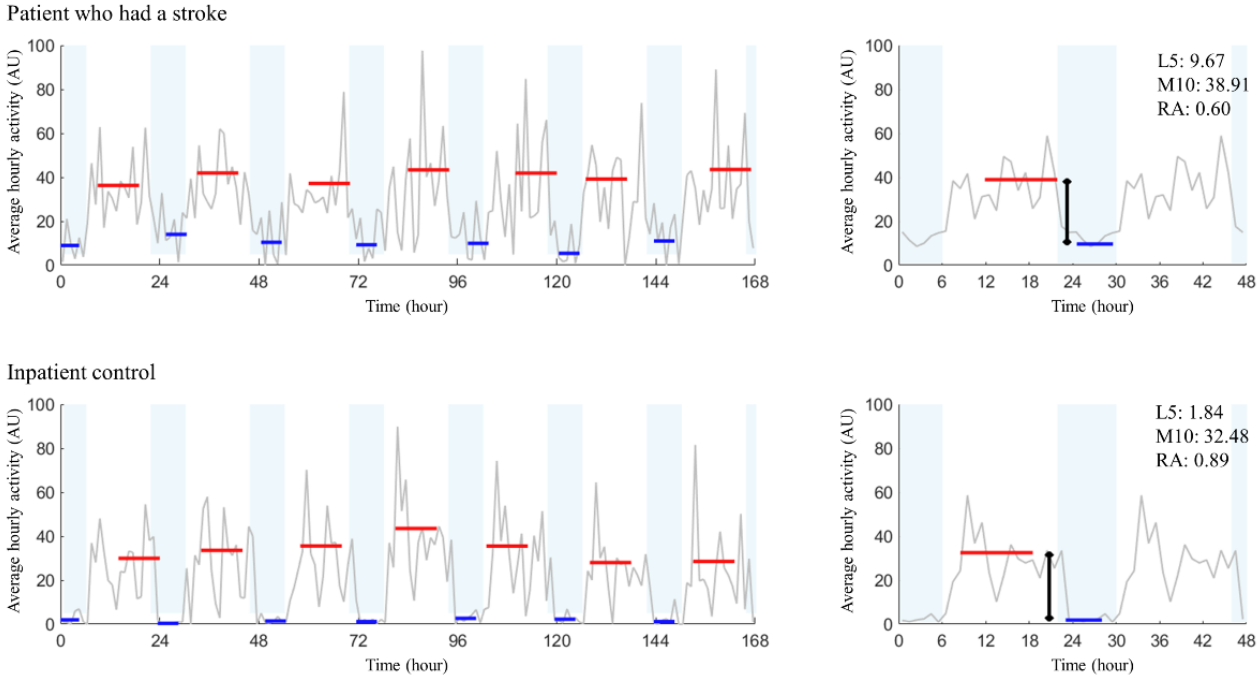

XSL•FO
RenderX

Table 1. Demographic and medical characteristics of the participants. Numerical data comparisons were conducted using the Mann-Whitney *U* test, and categorical data were analyzed using the chi-square test or Fisher exact test.

Variables	Stroke group (n=25)	Inpatient control group (n=25)	*P* value[a]
Women, n (%)	9 (36)	15 (60)	.09
Age (years), median (range)	56 (46-71)	57 (46.5-68.5)	.83
BMI (kg/m^2), median (range)	25.2 (23.3-27.5)	23.2 (19.4-26.8)	.11
Days after admission to hospitals, median (range)	13 (6.5-22)	18 (5-24)	.40
Days after admission to rehabilitation wards, median (range)	7 (7-7)	7 (7-7)	.42
Marital status, n (%)			.59
Married	18 (72)	21 (84)	
Divorced or widowed	2 (8)	1 (4)	
Unmarried or others	5 (20)	3 (12)	
Smoking, n (%)			.89
Nonsmoker	18 (72)	20 (80)	
Past smoker	3 (12)	3 (12)	
Smoker	4 (16)	2 (8)	
Occupation, n (%)			>.99
Retired or no occupation	9 (36)	8 (32)	
Part-time	2 (8)	2 (8)	
Full-time	14 (56)	15 (60)	
Past history, n (%)			
Hypertension	17 (68)	7 (28)	<.001
Diabetes mellites	4 (16)	3 (12)	.70
Hyperlipidemia	11 (44)	4 (16)	.03
Sleep medication	4 (16)	5 (20)	>.99
Barthel Index on admission, median (range)	40 (2.5-55)	35 (20-60)	.55
Using sleep medication, n (%)	8 (32)	11 (44)	.38
Type of stroke, n (%)			—[b]
Large-artery atherosclerosis	11 (44)	—	
Cardioembolism	11 (44)	—	
Small-vessel occlusion	6 (24)	—	
Cardioembolism	3 (12)	—	
Hemorrhage	9 (36)	—	
Other	3 (12)	—	
Side of stroke, n (%)			—
Right	10 (40)	—	
Left	11 (44)	—	
Bilateral	4 (16)	—	
Treated with recombinant tissue plasminogen activator, n (%)	5 (20)	—	—
Treated with mechanical thrombectomy, n (%)	3 (12)	—	—
4A's Test score			—
Total	2 (0-3)	—	
0	10 (40)	—	
1-3	12 (48)	—	

XSL•FO
RenderX

Variables	Stroke group (n=25)	Inpatient control group (n=25)	*P* value[a]
≥4	3 (12)	—	

[a]*P* values were derived from statistical comparisons between poststroke patient and inpatient control groups.

[b]Not available.

An in-depth comparison of RAR indicators between stroke and the control group is detailed in Table 2. The data highlight unique patterns and provide insights into poststroke patients' circadian rhythm disruptions. Patients who had a stroke exhibited significantly higher L5 values (representing the least active 5 continuous hours) than the control group, with median and IQR values of 4.29 (2.88-6.49) and 1.84 (0.67-4.34), respectively (*P*=.008 as per the Mann-Whitney *U* test). Conversely, M10 values, which encapsulate the most active 10 continuous hours, showed no significant differences between the groups (stroke group: median 38.92, IQR 14.60-40.87; control group: median 31.18, IQR 18.02-46.84; *P*=.93). This infers that peak activity durations were largely uniform across both cohorts. However, despite the congruence in M10 values, the stroke group (median 0.74, IQR 0.57-0.85) displayed a notably lower RA when juxtaposed with the control group (median 0.88, IQR 0.71-0.96; *P*=.009). Further analysis revealed no significant differences in other nonparametric or parametric RAR metrics between the groups.

Table 2. Comparative analysis of rest-activity rhythm indicators and physical activity levels between the stroke and inpatient control groups.

Variables	Stroke group (n=25)	Inpatient control group (n=25)	*P* value[a]
Interdaily stability	0.48 (0.38-0.62)	0.54 (0.35-0.69)	.18
Intradaily variability	1.10 (0.85-1.48)	1.19 (0.89-1.52)	.98
Relative amplitude	0.74 (0.57-0.85)	0.88 (0.71-0.96)	.009
Most active 10 continuous hours	38.92 (14.60-40.87)	31.18 (18.02-46.84)	.93
Least active 5 continuous hours	4.29 (2.88-6.49)	1.84 (0.67-4.34)	.008

[a]*P* values were calculated based on statistical comparisons between the stroke patient and inpatient control groups, using the Mann-Whitney *U* test.

Diving deeper into the stroke group patients, a significantly negative correlation was observed between 4AT scores (a delirium metric) and RA, as presented in Table 3 (ρ=–0.41; *P*=.045). For all participants, positive correlations emerged between total BI scores, denoting basic ADL independence, and both IS (ρ=0.34; *P*=.02) and M10 values (ρ=0.42; *P*=.002).

Table 3. Spearman correlation analysis between rest-activity rhythm indicators, 4A's Test (4AT) scores, and Barthel Index scores.

Instrument	Interdaily stability	Intradaily variability	Relative amplitude	M10[a]	L5[b]
Barthel Index					
ρ	0.34	–0.14	0.19	0.42	–0.01
P value	.02	.32	.18	.002	.96
4AT score[c]					
ρ	–0.32	0.26	–0.41	–0.36	–0.00
P value	.12	.21	.045	.07	.98

[a]M10: the most active 10 continuous hours.

[b]L5: the least active 5 continuous hours.

[c]Correlations for the 4AT scores pertain specifically to patients who had a stroke.

Discussion

Principal Findings

Our study unveiled a multitude of key findings emphasizing the clinical significance of specific RAR indicators in the context of poststroke rehabilitation. These findings provided valuable insights into the circadian rhythm disruptions experienced by patients who had a stroke during their recovery journey and underscored the need to consider circadian rhythms as a vital aspect of poststroke care. Our research highlighted the clinical relevance of RA in assessing circadian rhythms in patients who had a stroke. The lower RA observed in these patients, in comparison to the inpatient control group, suggested that patients who experienced a stroke exhibited a less pronounced diurnal activity variation, which could negatively predict their overall recovery and functional outcomes. It became apparent that circadian rhythm disruptions, as indicated by RA, had far-reaching implications for the management and assessment of poststroke patients. Additionally, our study explored the clinical implications of RA components, noting that disruptions in circadian patterns during nighttime and daytime significantly impacted patients' cognitive function and ADL. In particular, disturbances in nighttime activity patterns were linked to

XSL•FO
RenderX

potential sleep disturbances, adversely affecting the overall well-being of these patients. This highlighted the importance of managing these disruptions in comprehensive stroke care. Similarly, robust daytime activity was associated with independence in basic ADL, illustrating the importance of maintaining such patterns to improve functional independence during poststroke rehabilitation.

Comparison to Prior Research

Compared to prior research, our study represented a significant advancement in our understanding of the clinical significance of specific RAR indicators in poststroke rehabilitation. Our research contributed several key distinctions and valuable insights that enriched our comprehension of circadian rhythm disruptions in patients who had a stroke. Our study's observations concerning RAR measures, particularly the interplay between RA, L5, and M10, aligned with the findings of earlier large-scale investigations, such as those conducted within the UK Biobank [24]. These preceding studies consistently underscored the pivotal role of circadian disturbances in influencing poststroke outcomes. Significantly, our study reaffirmed these established trends. Moreover, our research extended beyond the scope of a previous study using accelerometer data from the UK Biobank, which encompassed a substantial sample size of 91,105 participants and meticulously controlled for numerous confounding factors [38]. This previous investigation primarily highlighted the association of lower RA with mood disorders, subjective well-being, and cognitive function. However, it did not delve into the specific components of lower RA, particularly how daytime and nighttime activities contributed to the observed associations with well-being and the risk of mood disorders [38]. Our study, on the other hand, delved deeper into the complexities of circadian disruptions and their specific manifestations in patients who had a stroke. We provided a comprehensive comparative analysis between stroke patients and inpatient control groups, focusing on distinct RAR indicators.

In our study, we observed that patients who had a stroke exhibited significantly higher L5 values compared to the control group, which indicated potentially poor sleep quality. This aligns with existing research suggesting that elevated L5 values are associated with sleep fragmentation, adversely affecting motor and cognitive performance [39-42]. Typically, the least active 5 hours in individuals sleeping more than 5 hours should coincide with the main sleep period, and disruptions in this period are reflected in increased L5 values [43]. For instance, in someone who sleeps 7 hours, the 1-hour periods just after falling asleep and before waking might involve some minor movements, but the middle 5 hours, which are usually without movement, are considered the L5 timing. This period represents the L5 value. Thus, an increase in L5 value often corresponds with shorter or more fragmented sleep, particularly during deep sleep phases where increased movement can be detected.

We also considered potential confounding factors, particularly the use of benzodiazepines for sedation and sleep. Although benzodiazepines are typically prescribed to alleviate sleep disturbances and could theoretically serve as indicators of sleep quality [44], we chose not to use them as such in our study due to their complex interplay with delirium and sleep apnea in poststroke patients. Indeed, while benzodiazepines have sedative and muscle-relaxing effects that might theoretically reduce the L5 values and improve sleep quality, their use was common among our study population (poststroke patients: n=8, 42.1%; control group: n=11, 57.9%), yet this did not translate into better sleep quality. This is because in patients who had a stroke, benzodiazepines might increase the risk of delirium or disturb the circadian rhythm, potentially leading to higher L5 values by increasing restlessness. Additionally, the high prevalence of sleep apnea in this group [45], causing fragmented sleep patterns and increased nocturnal activity, could also contribute significantly to elevated L5 values. Clinical guidelines recommend against using benzodiazepines for nighttime agitation in delirium, as they can worsen the condition [46]. Furthermore, the muscle-relaxing effects of benzodiazepines can exacerbate sleep apnea [47] in patients who had a stroke, further degrading sleep quality. These considerations underscore the importance of a holistic approach to managing sleep disturbances in stroke rehabilitation, incorporating both pharmacological and nonpharmacological strategies.

In addition to these revelations, our research reinforced the paramount importance of daily activity patterns in the context of poststroke rehabilitation. We identified positive correlations between total scores on the BI, IS, and the M10 value, firmly establishing the connection between circadian rhythm disruptions and functional independence during rehabilitation.

Furthermore, our cross-sectional observational study has provided insights into potential interventions that could be derived from enhancing digital biomarkers associated with circadian rhythms. A key finding is the significant role of RA in indicating circadian disruptions, where patients who had a stroke exhibited lower RA compared to the control group patients, and higher RA correlated with increased delirium scores. Given that RA is calculated as (M10–L5)/(M10+L5), interventions could focus on increasing M10 to enhance daytime activity or decreasing L5 to stabilize nighttime sleep, thus improving day-night distinctions and potentially enhancing daily living functions, as indicated by the positive correlation with the BI.

Moreover, stabilizing daily routines—not only by increasing daytime activities but also by ensuring regularity in day-to-day schedules—could enhance functional outcomes, as suggested by the positive correlation with IS. Lastly, improving sleep stability, which is particularly disrupted in patients who had a stroke (higher L5), could reduce delirium symptoms, further supporting the recovery process. These tailored interventions highlight the importance of precise and targeted approaches to managing circadian rhythm disruptions to improve overall patient outcomes in poststroke recovery.

Methodological Strengths

Our research design incorporated several methodological strengths that contributed to the significance of our study. First, we focused on circadian disruption, delirium, and other indicators specific to poststroke patients undergoing rehabilitation. While sleep disturbances were well recognized after stroke, our study delved into less explored aspects of sleep

XSL•FO
RenderX

quality in this population, shedding light on how these factors may impact recovery and rehabilitation outcomes. Second, unlike many previous studies that used healthy community-dwelling individuals as control groups [11,12], we used a more robust control group design. By comparing poststroke rehabilitation patients with inpatient control participants with non–brain-related lesions undergoing rehabilitation, we accounted for the unique circumstances and challenges faced by both groups during their inpatient stay. This approach enabled us to better isolate the effects of circadian rhythm disruptions from other factors, such as activity levels, environmental effects, and daily routines, which could significantly impact circadian rhythms. Moreover, our study used wearable sensors for monitoring circadian rhythms, aligning with recent advancements in health care technology. The use of wearable sensors provided a low-cost and practical method for monitoring sleep quality in hospital settings. This methodological approach addressed the need for more accessible and cost-effective ways to assess sleep quality in clinical environments, potentially making it easier for health care professionals to incorporate sleep monitoring into routine care. Lastly, our study aligned with recent research trends in the field. We acknowledged the study [20], which explored delirium detection using wearable sensors and machine learning in patients with intracerebral hemorrhage. This study emphasized the growing importance of wearable sensor technology in health care and highlighted the relevance of our approach in the broader context of neurological and rehabilitation research.

Limitations

Several methodological limitations should be taken into account when interpreting the findings of our study. First, the evaluation of delirium using the 4AT tool was exclusively administered to patients who had a stroke, resulting in a lack of corresponding delirium assessment for the control group. The participants in the control group had no history of brain-related lesions and were not routinely assessed for delirium in the wards. This discrepancy limits our ability to draw direct comparisons between the two groups in terms of delirium incidence. Second, both the stroke patient group and the control group exhibited limited levels of physical activity, making it impractical to apply standard actigraphy algorithms to distinguish sleep and wake states accurately. However, despite these constraints, our study provided valuable insights into sleep quality by analyzing nocturnal activity patterns, with a particular emphasis on L5 values. These observations allowed us to make informed inferences regarding potential sleep fragmentation and suboptimal sleep quality among patients who had a stroke. Finally, our assessment of ADL was based on cross-sectional observations, which may not fully capture the longitudinal dynamics of their relationship with rehabilitation outcomes. A more extended, longitudinal study would be necessary to explore the long-term effects of circadian rhythm disruptions on poststroke recovery comprehensively.

In conclusion, our study illuminated the critical role of specific RAR indicators in poststroke rehabilitation, shedding light on the clinical significance of circadian rhythm disruptions in this context. The lower RA observed in poststroke patients suggested a potential avenue for intervention, indicating that circadian disturbances might be modifiable risk factors for stroke. By addressing these disturbances and promoting robust daytime and nighttime activity patterns aligned with the circadian cycle, health care professionals could potentially enhance stroke rehabilitation and improve the overall quality of poststroke care. Furthermore, our research underscored the emergence of “digital biomarkers” in the field of neurological and rehabilitation research. Wearable sensor technology, as demonstrated in our study, offered a practical and cost-effective means of monitoring circadian rhythms and sleep quality in clinical settings. The integration of digital biomarkers into routine health care assessments holds great promise for enhancing the precision and personalization of patient care, enabling early interventions, and ultimately improving stroke outcomes.

Acknowledgments

This work was supported by the National Taiwan University Hospital Hsin-Chu Branch and National Health Research Institutes (grants 112-HR0008 and 12A1-PHSP05-014). The funding source had no involvement in the study design; the collection, analysis, and interpretation of data; the writing of the report; and the decision to submit the manuscript for publication. The authors would like to thank the joint funding from the National Taiwan University Hospital Hsin-Chu Branch and the National Health Research Institutes. We used Generative AI only for grammar correction and did not use it for content drafting in any portion of the manuscript.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YHL, HWL, and CHW contributed to the study conception and design. HWL, SYC, and WCH conducted data collection. YHL, HWL, CL, HCC, and CHW were in charge of the analysis and interpretation of the results. YSL and HWL drafted the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest

YHL is an associate editor of JMIR at the time of this publication

XSL•FO
RenderX

References

1. Sterr A, Kuhn M, Nissen C, Ettine D, Funk S, Feige B, et al. Post-stroke insomnia in community-dwelling patients with chronic motor stroke: physiological evidence and implications for stroke care. Sci Rep 2018 May 30;8(1):8409 [FREE Full text] [doi: 10.1038/s41598-018-26630-y] [Medline: 29849087]
2. Terzoudi A, Vorvolakos T, Heliopoulos I, Livaditis M, Vadikolias K, Piperidou H. Sleep architecture in stroke and relation to outcome. Eur Neurol 2009 Oct 24;61(1):16-22. [doi: 10.1159/000165344] [Medline: 18948695]
3. Duss SB, Seiler A, Schmidt MH, Pace M, Adamantidis A, Müri RM, et al. The role of sleep in recovery following ischemic stroke: A review of human and animal data. Neurobiol Sleep Circadian Rhythms 2017 Jan;2:94-105 [FREE Full text] [doi: 10.1016/j.nbscr.2016.11.003] [Medline: 31236498]
4. Siengsukon C, Al-Dughmi M, Al-Sharman A, Stevens S. Sleep parameters, functional status, and time post-stroke are associated with offline motor skill learning in people with chronic stroke. Front Neurol 2015 Oct 30;6:225 [FREE Full text] [doi: 10.3389/fneur.2015.00225] [Medline: 26579068]
5. Kim J, Kim Y, Yang KI, Kim D, Kim SA. The relationship between sleep disturbance and functional status in mild stroke patients. Ann Rehabil Med 2015 Aug;39(4):545-552 [FREE Full text] [doi: 10.5535/arm.2015.39.4.545] [Medline: 26361590]
6. Hei Chow C, Fraysse F, Hillier S. The relationship between sleep and physical activity in an in-patient rehabilitation stroke setting: a cross-sectional study. Top Stroke Rehabil 2023 Jan 29;30(1):43-52. [doi: 10.1080/10749357.2021.2006982] [Medline: 34842058]
7. Fleming MK, Smejka T, Henderson Slater D, van Gils V, Garratt E, Yilmaz Kara E, et al. Sleep disruption after brain injury is associated with worse motor outcomes and slower functional recovery. Neurorehabil Neural Repair 2020 Jul 07;34(7):661-671 [FREE Full text] [doi: 10.1177/1545968320929669] [Medline: 32508249]
8. Delaney LJ, Currie MJ, Huang HC, Lopez V, Van Haren F. "They can rest at home": an observational study of patients' quality of sleep in an Australian hospital. BMC Health Serv Res 2018 Jul 05;18(1):524 [FREE Full text] [doi: 10.1186/s12913-018-3201-z] [Medline: 29976191]
9. Christensen J, Yamakawa GR, Shultz SR, Mychasiuk R. Is the glymphatic system the missing link between sleep impairments and neurological disorders? Examining the implications and uncertainties. Prog Neurobiol 2021 Mar;198:101917. [doi: 10.1016/j.pneurobio.2020.101917] [Medline: 32991958]
10. Bassetti CL, Randerath W, Vignatelli L, Ferini-Strambi L, Brill A, Bonsignore MR, et al. EAN/ERS/ESO/ESRS statement on the impact of sleep disorders on risk and outcome of stroke. Eur Respir J 2020 Apr 21;55(4):1901104 [FREE Full text] [doi: 10.1183/13993003.01104-2019] [Medline: 32317355]
11. Cavalcanti P, Campos T, Araujo J. Actigraphic analysis of the sleep-wake cycle and physical activity level in patients with stroke: implications for clinical practice. Chronobiol Int 2012 Nov 24;29(9):1267-1272. [doi: 10.3109/07420528.2012.719960] [Medline: 23003002]
12. English C, Healy G, Coates A, Lewis L, Olds T, Bernhardt J. Sitting and activity time in people with stroke. Phys Ther 2016 Feb;96(2):193-201. [doi: 10.2522/ptj.20140522] [Medline: 26112254]
13. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Association Publishing; 2013.
14. Shaw RC, Walker G, Elliott E, Quinn TJ. Occurrence rate of delirium in acute stroke settings: systematic review and meta-analysis. Stroke 2019 Nov;50(11):3028-3036. [doi: 10.1161/STROKEAHA.119.025015] [Medline: 31554501]
15. Nerdal V, Gjestad E, Saltvedt I, Munthe-Kaas R, Ihle-Hansen H, Ryum T, et al. The relationship of acute delirium with cognitive and psychiatric symptoms after stroke: a longitudinal study. BMC Neurol 2022 Jun 27;22(1):234 [FREE Full text] [doi: 10.1186/s12883-022-02756-5] [Medline: 35761180]
16. Jaiswal SJ, Bagsic SRS, Takata E, Kamdar BB, Ancoli-Israel S, Owens RL. Actigraphy-based sleep and activity measurements in intensive care unit patients randomized to ramelteon or placebo for delirium prevention. Sci Rep 2023 Jan 26;13(1):1450 [FREE Full text] [doi: 10.1038/s41598-023-28095-0] [Medline: 36702822]
17. Godfrey A, Conway R, Leonard M, Meagher D, Olaighin G. A classification system for delirium subtyping with the use of a commercial mobility monitor. Gait Posture 2009 Aug;30(2):245-252. [doi: 10.1016/j.gaitpost.2009.05.014] [Medline: 19539473]
18. Maybrier HR, King CR, Crawford AE, Mickle AM, Emmert DA, Wildes TS, ENGAGES Study Investigators. Early postoperative actigraphy poorly predicts hypoactive delirium. J Clin Sleep Med 2019 Jan 15;15(1):79-87 [FREE Full text] [doi: 10.5664/jcsm.7576] [Medline: 30621829]
19. Osse RJ, Tulen JHM, Hengeveld MW, Bogers AJJC. Screening methods for delirium: early diagnosis by means of objective quantification of motor activity patterns using wrist-actigraphy. Interact Cardiovasc Thorac Surg 2009 Mar;8(3):344-8; discussion 348. [doi: 10.1510/icvts.2008.192278] [Medline: 19103609]
20. Ahmed A, Garcia-Agundez A, Petrovic I, Radaei F, Fife J, Zhou J, et al. Delirium detection using wearable sensors and machine learning in patients with intracerebral hemorrhage. Front Neurol 2023;14:1135472 [FREE Full text] [doi: 10.3389/fneur.2023.1135472] [Medline: 37360342]
21. Witting W, Kwa I, Eikelenboom P, Mirmiran M, Swaab D. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. Biol Psychiatry 1990 Mar 15;27(6):563-572. [doi: 10.1016/0006-3223(90)90523-5] [Medline: 2322616]

XSL•FO
RenderX

22. Cavalcanti PRA, Campos TF, Araújo JF. Circadian and homeostatic changes of sleep-wake and quality of life in stroke: implications for neurorehabilitation. NRE 2013 Apr 11;32(2):337-343. [doi: 10.3233/nre-130853]
23. de Oliveira DC, Ferreira PRC, Silveira Fernandes ABG, Pacheco TBF, Avelino MML, Cavalcanti FADC, et al. Circadian activity rhythm and fragmentation are associated with sleep-wake patterns and sleep quality in patients with stroke. NRE 2019 Jun 20;44(3):353-360. [doi: 10.3233/nre-182665]
24. Gao L, Zheng X, Baker S, Li P, Scheer F, Nogueira R, et al. Associations of rest-activity rhythm disturbances with stroke risk and post-stroke adverse outcomes. medRxiv 2023 May 16:2023 [FREE Full text] [doi: 10.1101/2023.05.14.23289966] [Medline: 37292791]
25. Robinson RG, Jorge RE. Post-stroke depression: a review. Am J Psychiatry 2016 Mar 01;173(3):221-231. [doi: 10.1176/appi.ajp.2015.15030363] [Medline: 26684921]
26. Crichton SL, Bray BD, McKevitt C, Rudd AG, Wolfe CDA. Patient outcomes up to 15 years after stroke: survival, disability, quality of life, cognition and mental health. J Neurol Neurosurg Psychiatry 2016 Oct 22;87(10):1091-1098. [doi: 10.1136/jnnp-2016-313361] [Medline: 27451353]
27. Lin C, Chen I, Chuang H, Wang Z, Lin H, Lin Y. Examining Human-Smartphone Interaction as a Proxy for Circadian Rhythm in Patients With Insomnia: Cross-Sectional Study. J Med Internet Res 2023 Dec 15;25:e48044 [FREE Full text] [doi: 10.2196/48044] [Medline: 38100195]
28. Hsieh Y, Wang C, Wu S, Chen P, Sheu C, Hsieh C. Establishing the minimal clinically important difference of the Barthel Index in stroke patients. Neurorehabil Neural Repair 2007;21(3):233-238. [doi: 10.1177/1545968306294729] [Medline: 17351082]
29. Mahony FI, Barthel DW. Functional evaluation: the Barthel Index. Md State Med J 1965 Feb;14:61-65. [Medline: 14258950]
30. Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. Classification and natural history of clinically identifiable subtypes of cerebral infarction. Lancet 1991 Jun 22;337(8756):1521-1526. [doi: 10.1016/0140-6736(91)93206-o] [Medline: 1675378]
31. Bellelli G, Morandi A, Davis DHJ, Mazzola P, Turco R, Gentile S, et al. Validation of the 4AT, a new instrument for rapid delirium screening: a study in 234 hospitalised older people. Age Ageing 2014 Jul 02;43(4):496-502 [FREE Full text] [doi: 10.1093/ageing/afu021] [Medline: 24590568]
32. Lees R, Corbet S, Johnston C, Moffitt E, Shaw G, Quinn TJ. Test accuracy of short screening tests for diagnosis of delirium or cognitive impairment in an acute stroke unit setting. Stroke 2013 Nov;44(11):3078-3083. [doi: 10.1161/STROKEAHA.113.001724] [Medline: 23988641]
33. Driller MW, O'Donnell S, Tavares F. What wrist should you wear your actigraphy device on? Analysis of dominant vs non-dominant wrist actigraphy for measuring sleep in healthy adults. Sleep Sci 2017;10(3):132-135 [FREE Full text] [doi: 10.5935/1984-0063.20170023] [Medline: 29410743]
34. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003 May 01;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
35. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992 Oct;15(5):461-469. [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
36. Gonçalves BSB, Adamowicz T, Louzada FM, Moreno CR, Araujo JF. A fresh look at the use of nonparametric analysis in actimetry. Sleep Med Rev 2015 Apr;20:84-91. [doi: 10.1016/j.smrv.2014.06.002] [Medline: 25065908]
37. Jones SE, van Hees VT, Mazzotti DR, Marques-Vidal P, Sabia S, van der Spek A, et al. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. Nat Commun 2019 Apr 05;10(1):1585 [FREE Full text] [doi: 10.1038/s41467-019-09576-1] [Medline: 30952852]
38. Lyall LM, Wyse CA, Graham N, Ferguson A, Lyall DM, Cullen B, et al. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. Lancet Psychiatry 2018 Jun;5(6):507-514. [doi: 10.1016/s2215-0366(18)30139-1]
39. Wright K, Hull J, Hughes R, Ronda J, Czeisler C. Sleep and wakefulness out of phase with internal biological time impairs learning in humans. J Cogn Neurosci 2006 Apr;18(4):508-521. [doi: 10.1162/jocn.2006.18.4.508] [Medline: 16768357]
40. Curcio G, Ferrara M, De Gennaro L. Sleep loss, learning capacity and academic performance. Sleep Med Rev 2006 Oct;10(5):323-337. [doi: 10.1016/j.smrv.2005.11.001] [Medline: 16564189]
41. Goldman S, Stone K, Ancoli-Israel S, Blackwell T, Ewing S, Boudreau R, et al. Poor sleep is associated with poorer physical performance and greater functional limitations in older women. Sleep 2007 Oct;30(10):1317-1324 [FREE Full text] [doi: 10.1093/sleep/30.10.1317] [Medline: 17969465]
42. Oosterman JM, van Someren EJW, Vogels RLC, Van Harten B, Scherder EJA. Fragmentation of the rest-activity rhythm correlates with age-related cognitive deficits. J Sleep Res 2009 Mar 24;18(1):129-135 [FREE Full text] [doi: 10.1111/j.1365-2869.2008.00704.x] [Medline: 19250179]
43. Meyer N, Harvey AG, Lockley SW, Dijk D. Circadian rhythms and disorders of the timing of sleep. Lancet 2022 Sep;400(10357):1061-1078. [doi: 10.1016/s0140-6736(22)00877-7]
44. De Crescenzo F, D'Alò GL, Ostinelli EG, Ciabattini M, Di Franco V, Watanabe N, et al. Comparative effects of pharmacological interventions for the acute and long-term management of insomnia disorder in adults: a systematic review and network meta-analysis. Lancet 2022 Jul 16;400(10347):170-184 [FREE Full text] [doi: 10.1016/S0140-6736(22)00878-9] [Medline: 35843245]

XSL•FO
RenderX

45. McKee Z, Auckley DH. A sleeping beast: obstructive sleep apnea and stroke. Cleve Clin J Med 2019 Jun;86(6):407-415 [FREE Full text] [doi: 10.3949/ccjm.86a.18033] [Medline: 31204980]
46. Marcantonio ER. Delirium in hospitalized older adults. N Engl J Med 2017 Oct 12;377(15):1456-1466 [FREE Full text] [doi: 10.1056/NEJMcp1605501] [Medline: 29020579]
47. Gottlieb DJ, Punjabi NM. Diagnosis and management of obstructive sleep apnea: a review. JAMA 2020 Apr 14;323(14):1389-1400. [doi: 10.1001/jama.2020.3514] [Medline: 32286648]

Abbreviations

ADL: activities of daily living
BI: Barthel Index
IS: interdaily stability
IV: intradaily variability
L5: the least active 5 continuous hours
M10: the most active 10 continuous hours
RA: relative amplitude
RAR: rest-activity rhythm
4AT: 4 A's Test

Edited by T de Azevedo Cardoso; submitted 08.01.24; peer-reviewed by LA Lee, J Fernandez-Saez; comments to author 10.04.24; revised version received 08.05.24; accepted 30.05.24; published 04.07.24.

Please cite as:
Liang HW, Wu CH, Lin C, Chang HC, Lin YH, Chen SY, Hsu WC
Rest-Activity Rhythm Differences in Acute Rehabilitation Between Poststroke Patients and Non–Brain Disease Controls: Comparative Study
J Med Internet Res 2024;26:e49530
URL: https://www.jmir.org/2024/1/e49530
doi:10.2196/49530
PMID:

©Huey-Wen Liang, Chueh-Hung Wu, Chen Lin, Hsiang-Chih Chang, Yu-Hsuan Lin, Shao-Yu Chen, Wei-Chen Hsu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Embodied Conversational Agents for Chronic Diseases: Scoping Review

Zhili Jiang[1], MSN; Xiting Huang[1], MM; Zhiqian Wang[1], MSN; Yang Liu[1], MM; Lihua Huang[1], MBBS; Xiaolin Luo[2]

[1]Department of Nursing, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Quality Evaluation, Zhejiang Evaluation Center for Medical Service and Administration, Hangzhou, China

Corresponding Author:
Lihua Huang, MBBS
Department of Nursing
The First Affiliated Hospital
Zhejiang University School of Medicine
Building 17, 3rd Floor
79 Qingchun Road
Hangzhou, 310003
China
Phone: 86 13867129329
Email: lihuahuang818@zju.edu.cn

Abstract

Background: Embodied conversational agents (ECAs) are computer-generated animated humanlike characters that interact with users through verbal and nonverbal behavioral cues. They are increasingly used in a range of fields, including health care.

Objective: This scoping review aims to identify the current practice in the development and evaluation of ECAs for chronic diseases.

Methods: We applied a methodological framework in this review. A total of 6 databases (ie, PubMed, Embase, CINAHL, ACM Digital Library, IEEE Xplore Digital Library, and Web of Science) were searched using a combination of terms related to ECAs and health in October 2023. Two independent reviewers selected the studies and extracted the data. This review followed the PRISMA-ScR (Preferred Reporting Items of Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) statement.

Results: The literature search found 6332 papers, of which 36 (0.57%) met the inclusion criteria. Among the 36 studies, 27 (75%) originated from the United States, and 28 (78%) were published from 2020 onward. The reported ECAs covered a wide range of chronic diseases, with a focus on cancers, atrial fibrillation, and type 2 diabetes, primarily to promote screening and self-management. Most ECAs were depicted as middle-aged women based on screenshots and communicated with users through voice and nonverbal behavior. The most frequently reported evaluation outcomes were acceptability and effectiveness.

Conclusions: This scoping review provides valuable insights for technology developers and health care professionals regarding the development and implementation of ECAs. It emphasizes the importance of technological advances in the embodiment, personalized strategy, and communication modality and requires in-depth knowledge of user preferences regarding appearance, animation, and intervention content. Future studies should incorporate measures of cost, efficiency, and productivity to provide a comprehensive evaluation of the benefits of using ECAs in health care.

(J Med Internet Res 2024;26:e47134) doi:10.2196/47134

KEYWORDS

embodied conversational agent; ECA; chronic diseases; eHealth; health care; mobile phone

Introduction

Background

With the rapid aging of the population and lifestyle changes, chronic diseases have become a significant global public health problem, arousing great concern in people from all walks of life. In 2018, at least 1 chronic disease was experienced by 51.8% of American adults, and 27.2% dealt with multiple chronic diseases [1]. In China, chronic diseases accounted for 86.6% of total deaths and approximately 70% of the total burden of diseases [2]. Given the prolonged duration and severe health damage associated with chronic diseases, patients often require

XSL•FO
RenderX

assistance in long-term care [3]. To relieve this growing burden, particularly in health care services and related costs, advancements in network communication technology have shown promise in improving the availability and quality of support services. eHealth applications allow remote patient monitoring and provide patient-tailored support in their home settings. However, many eHealth applications have faced the problem of their actual use decreasing after several weeks [4]. This decline may be attributed to the fact that most existing eHealth applications provide such support in the form of plain text or via a text-based question-answer module, whereas person-to-person interaction remains one of the best ways to communicate health information [5]. Face-to-face consultations, with the use of verbal and nonverbal behaviors such as empathy and immediacy, can foster trust and satisfaction among patients, leading to better health communication and understanding [6].

An embodied conversational agent (ECA) is a computer-based dialogue system with an embodiment that simulates a face-to-face conversation with humanlike physical properties, including verbal and nonverbal behavioral cues (eg, speech, facial expressions, and gestures) [7]. Compared with a static character image or a text-only display, the interactive, conversational modes of communication used by ECAs may potentially improve engagement by providing additional motivational and emotional support [8,9]. In health care, ECAs have been designed to assist with various tasks such as providing diabetes self-management education [10], promoting cancer screening [11], and delivering cognitive behavioral therapy for depression [12]. Despite the exciting potential for using ECAs for health purposes, their use could be ineffective or even have unintended negative consequences if the design, including visual appearance and intervention content, does not meet the user's expectations [13]. Research has shown that design decisions related to the look and feel of ECAs significantly influence users' psychological and emotional responses and their engagement with the applications [14]. However, how ECAs should be designed and used to maximize their effectiveness in the context of chronic diseases is still unknown. Therefore, it is crucial to systematically review the development and evaluation of ECAs in a specific context to optimize their design to provide a positive user experience and promote engagement.

Currently, there is a lack of comprehensive reviews on the development and evaluation of ECAs, particularly in the context of chronic diseases. Although there have been literature reviews on conversational agents in eHealth, they often focus on the impact of ECAs rather than the design processes involved. For example, Kramer et al [15] conducted a scoping review of ECAs in a healthy lifestyle and pointed out that the design of an ECA could have a major effect on both impact and uptake. However, reports on design activities and their results were generally incomplete or missing. Similarly, 2 other reviews by ter Stal et al [16] and Loveys et al [17] identified the most common design features of ECAs and their impact on user perception but ignored the design activities of ECAs. Another scoping review by Provoost et al [18] aimed to provide an overview of the technological and clinical possibilities of ECAs but only for patients with mental disorders. Therefore, a comprehensive literature review focusing on the design processes of ECAs in the context of chronic diseases is required.

Existing studies have primarily focused on design features rather than design processes and have examined a broader context beyond health or a specific subarea of health such as clinical psychology. In this study, we aimed to review relevant studies to understand how ECAs have been designed and evaluated specifically in the context of chronic diseases. After a preliminary exploration of the relevant literature on ECAs to determine the review method, it was found that the traditional systematic review or meta-analysis method seemed unsuitable because of the variability in populations, study designs, and measured outcomes. Compared with traditional systematic reviews, scoping reviews cover a broader range of topics in which many different study designs may be applicable, and the quality evaluation of the included research is not emphasized [19]. Therefore, we adopted the scoping review method, which provides a clear and systematic means to outline this large and diverse body of literature, using rigorous methods to minimize bias [20].

Objectives

In this study, we undertook a scoping review focused on the development and evaluation of ECAs in the context of chronic diseases. In particular, the aims of our scoping review are as follows: (1) provide an overview of all studies on the developmental practices of ECAs for chronic diseases, (2) summarize the design and design processes of ECAs in chronic diseases, and (3) identify evaluation and outcomes reported in studies. Conducting this scoping review will benefit both technology developers and health care professionals. For technology developers, this review will provide a comprehensive understanding of the different approaches and techniques that have been used in the development of ECAs for chronic diseases, enabling them to develop a more intelligent ECA that provides a natural experience for users. For health care professionals, this review will offer actionable advice that helps them better manage and provide medical services using ECAs, ultimately improving patient outcomes.

Methods

Study Design

The framework for scoping reviews by Arksey and O'Malley [19] was adopted. The main five stages were as follows: (1) identifying the research questions; (2) identifying relevant studies; (3) selecting studies; (4) charting the data; and (5) collating, summarizing, and reporting the results. We followed the process outlined in the published protocol and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines (Multimedia Appendix 1) [21].

Identifying the Research Questions

This study mainly addressed the following questions: (1) What are the basic characteristics of the included studies? (2) How to design ECAs to guide self-care of chronic diseases? and (3) How to evaluate the impact of ECA interventions for chronic diseases?

XSL•FO
RenderX

Identifying Relevant Studies

The databases used to locate the relevant studies were PubMed, Embase, CINAHL, Web of Science, IEEE Xplore Digital Library, and ACM Digital Library. These databases were chosen because they cover relevant aspects in the fields of health and information technology and have been used in other systematic literature reviews covering similar topics [22,23]. The search terms were identified based on a preliminary literature scan and the opinions of a research librarian to ensure a comprehensive search for relevant studies. The final search terms included an extensive list of items describing the constructs "embodied conversational agents" and "health." A complete overview of the search terms for each construct and the inclusion criteria implemented by selecting different options and limits during the search can be found in Multimedia Appendix 2. An exemplary search strategy for PubMed is presented in Textbox 1. In addition, we used the snowball method. The search was limited to English papers published before the search date of October 1, 2023.

Textbox 1. The search strategy used in PubMed.

Health

"Health*"[Title/Abstract] OR "mHealth"[Title/Abstract] OR "m-Health"[Title/Abstract] OR "Telehealth"[Title/Abstract] OR "Tele-health"[Title/Abstract] OR "eHealth"[Title/Abstract] OR "e-Health" [Title/Abstract] OR "Telemedicine"[Title/Abstract] OR "Tele-medicine"[Title/Abstract] OR "well-being"[Title/Abstract] OR "wellbeing"[Title/Abstract] OR "medic*"[Title/Abstract] OR "illness"[Title/Abstract] OR "patient*"[Title/Abstract] OR "disorder*"[Title/Abstract] OR "disease*"[Title/Abstract]

Embodied conversational agents

"Conversational agent*"[Title/Abstract] OR "Conversational assistant*"[Title/Abstract] OR "Embodied agent*"[Title/Abstract] OR "Animated character*"[Title/Abstract] OR "Animated agent*" [Title/Abstract] OR "Virtual agent*"[Title/Abstract] OR "Virtual assistant*"[Title/Abstract] OR "Virtual health assistant*"[Title/Abstract] OR "Virtual coach*"[Title/Abstract] OR "Virtual character*"[Title/Abstract] OR "Virtual human"[Title/Abstract] OR "Virtual therapist*"[Title/Abstract] OR "Virtual nurse*"[Title/ Abstract] OR "Virtual companion*"[Title/Abstract] OR "Virtual counselor*"[Title/Abstract] OR "Virtual health counselor*"[Title/Abstract] OR "Virtual clinician*"[Title/Abstract] OR "Interactive agent*"[Title/Abstract] OR "Relational agent*"[Title/Abstract]

Combined

1 AND 2

Study Selection

The results of the search query were uploaded to the NoteExpress (version 3.3; Beijing Aegean Software Company) reference manager and independently assessed by 2 reviewers (ZJ and XH) to decide on their inclusion based on the title, abstract, and full text. Following an initial screening of titles and abstracts, the full texts were obtained and screened by 2 reviewers. If the eligibility of the full text was unclear, any discrepancies were reviewed by an additional author (YL) and resolved in a consensus meeting. For this review, we included chronic diseases identified by the Public Health Agency of Canada, including cancer, heart disease, hypertension, stroke, chronic respiratory diseases (eg, asthma, chronic obstructive pulmonary disease, and sleep apnea), diabetes, inflammatory bowel diseases, neurological conditions (eg, Alzheimer disease and other dementia, Parkinson disease, traumatic brain injury, and traumatic spinal cord injury), arthritis, and osteoporosis [24]. Mental illness was excluded from the list, given that support interventions of ECAs for this group may have unique features that are not generalizable to other chronic diseases. In addition, we decided to include other diseases that require self-care outside the list of the Public Health Agency of Canada, such as obesity and chronic pain. We included full papers that met the following criteria: (1) participants were adults aged ≥18 years, (2) papers were published in English, and (3) ECAs were made available to the general public (eg, general population or patient). We applied the following exclusion criteria: (1) reviews, editorials, opinions, theses, and conference abstracts; (2) papers for which full texts were unavailable; (3) ECAs used for training or educating medical professionals or not used in the context of chronic diseases; and (4) papers that did not involve ECAs (computer-generated virtual individuals with an animated appearance to enable face-to-face interaction between the user and the system) [25].

Charting the Data and Collating and Summarizing the Results

Data extraction was conducted independently by 2 reviewers (ZJ and ZW) using an Excel (Microsoft Corp) spreadsheet. Any discrepancies in the extracted data were discussed between the authors and resolved by discussion and consensus. Extracted data included (1) paper information, (2) study information, (3) details about the ECAs (eg, identity, communication modality, and personalized content), and (4) evaluation outcomes. The content of concepts can be predefined based on the study by Kramer et al [15] (Multimedia Appendix 3). In cases where a paper included multiple studies, data extraction was performed only for the studies that met the eligibility criteria. Data were extracted separately if multiple eligible studies were included in 1 paper. Once all the study data were collected, we conducted a thematic analysis and categorized them into 3 main topics. The first topic described the identities (including ECA's names, roles, and appearances), communication modalities, and personalization in intervention content and delivery. The second topic focused on the technologies and theories or principles used. The third topic described the evaluation measures and outcomes.

XSL•FO
RenderX

Results

Papers Retrieved

The initial search identified 6332 references in October 2023; after the removal of duplicates, 4341 references (68.56%) remained. The titles and abstracts of these references were screened by both reviewers, resulting in the exclusion of 4066 references. After further evaluation, 245 papers were excluded, and the remaining 30 papers were considered eligible for a comprehensive review. In addition, 4 more studies were identified through snowballing [26-29]. This resulted in a total of 36 studies as 2 papers [30,31] included 2 studies each. Figure 1 describes the search process and outcomes.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection. ECA: embodied conversational agent.

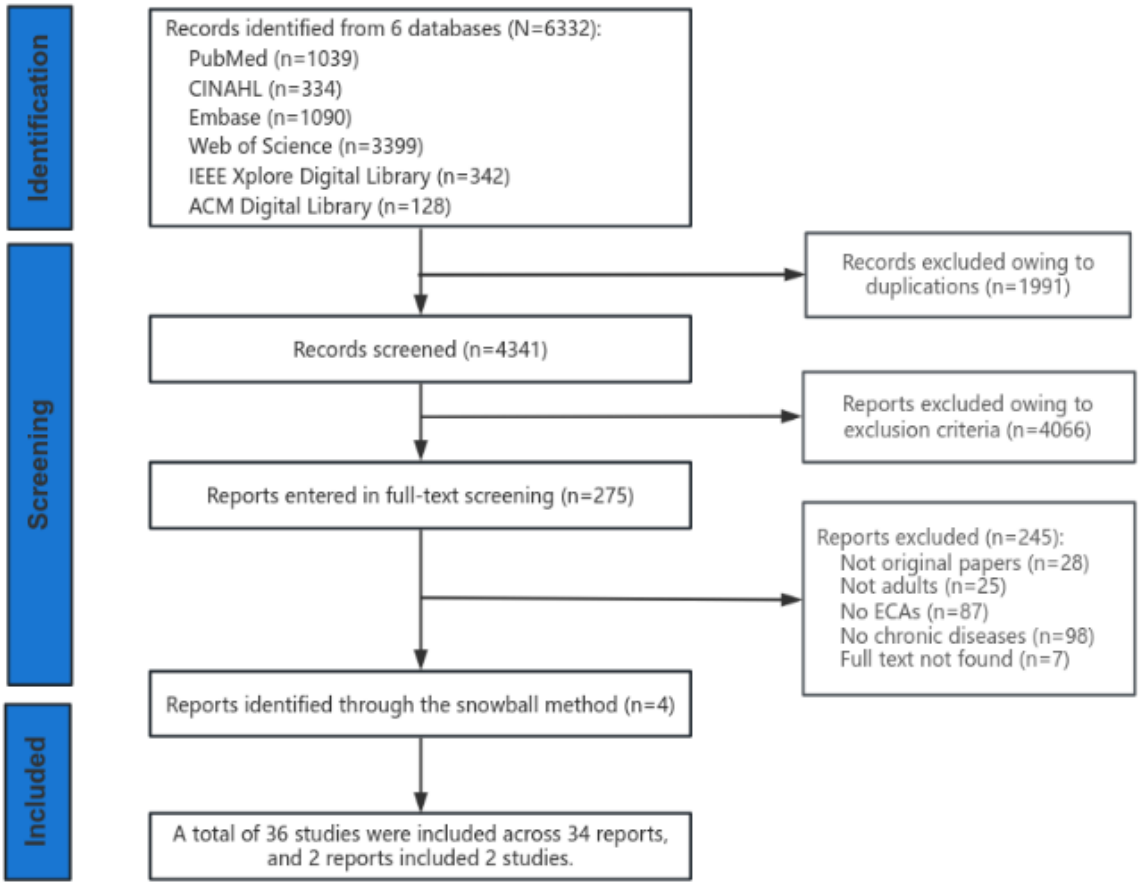

Description of Included Studies

Multimedia Appendix 4 [10,11,26-57] presents a summary of the studies included in this review. The characteristics of the included studies are presented in detail in Table 1.

The first study [32] was published in 2012, and 22% (8/36) of the studies were published before 2020 [27-29,32-36]. Most of the included studies were conducted in the United States (27/36, 75%) [11,27-34,36,43-57] and Australia (3/36, 8%) [10,26,37]. Other study locations included France (1/36, 3%) [38], Portugal (1/36, 3%) [39], the Netherlands (1/36, 3%) [40], and Italy (1/36, 3%) [41]. One study was conducted in 3 countries (Italy, Luxembourg, and the Netherlands) [42], and another study did not provide any information [35]. Among the studies that reported a study design, mixed method studies (8/36, 22%) [10,28,39-44] and pilot studies (8/36, 22%) [31,34,38,45-48] were the most commonly used study designs, followed by qualitative studies (6/36, 17%) [11,26,35,49-51], randomized controlled trials (5/36, 14%) [32,37,52-54], ongoing trials (4/36, 11%) [33,55-57], quasiexperimental designs (2/36, 6%) [30], longitudinal study (1/36, 3%) [29], single-group nonrandomized trial (1/36, 3%) [36], and pre- and posttest design (1/36, 3%) [27]. Of the 36 studies, 30 (83%) were conducted on patients (15/36, 42%) [10,26,32,34-37,40,42,47,52,54-57] or healthy individuals (15/36, 42%) [11,27,29-31,44-46,48-51,53] and 6 (17%) [28,33,38,39,41,43] on stakeholders such as homecare providers or experts. The mean age of the target population across studies was >40 years, except for 19% (7/36) of the studies [30,33,35,47,53,57].

XSL•FO
RenderX

Table 1. Characteristics of the included studies (n=36).

Study characteristics	Studies, n (%)
Publication year	
Before 2020	8 (22)
2020 or after	28 (78)
Study country	
United States	27 (75)
Australia	3 (8)
Portugal	1 (3)
Netherlands	1 (3)
Italy	1 (3)
France	1 (3)
Italy, Luxembourg, and the Netherlands	1 (3)
Not available	1 (3)
Study design	
Mixed methods study	8 (22)
Pilot study	8 (22)
Qualitative study	6 (17)
Randomized controlled trial	5 (14)
Ongoing trial	4 (11)
Quasiexperimental design	2 (6)
Longitudinal study	1 (3)
Pre- and posttest design	1 (3)
Single-group nonrandomized design	1 (3)
Sample population	
Patients	15 (42)
Healthy adults	15 (42)
Multiple stakeholders	6 (17)
Targeted chronic disease	
Colorectal cancer	10 (28)
Atrial fibrillation	4 (11)
Breast cancer	4 (11)
Type 2 diabetes	4 (11)
Prostate cancer	3 (8)
Chronic pain	3 (8)
Dementia	2 (6)
Chronic obstructive pulmonary disease and chronic heart failure	1 (3)
Obesity	1 (3)
Parkinson disease	1 (3)
Sleep apnea syndrome	1 (3)
Spinal cord injury	1 (3)
Heart failure	1 (3)
Main purpose of embodied conversational agent	
Promote screening	13 (36)

XSL•FO
RenderX

Study characteristics	Studies, n (%)
Promote self-management	11 (31)
Genetic counseling	3 (8)
Therapy	3 (8)
Assisted living	2 (6)
Promote exercise	2 (6)
Diagnosis	1 (3)
Education	1 (3)
Delivery channel	
Web	15 (42)
Smartphone app	10 (28)
Tablet	7 (19)
Computer	4 (11)

The most common chronic condition reported in the studies was cancer, accounting for 47% (17/36) of the studies. Among these 36 studies, 10 (28%) [11,31,44-46,48-51] focused on colorectal cancer, 4 (11%) [30,43,53] focused on breast cancer, and 3 (8%) [27-29] focused on prostate cancer. Of the 36 studies, 8 (22%) addressed atrial fibrillation [34,52,55,56] and type 2 diabetes [10,26,37,39]. Other conditions included dementia (2/36, 6%) [41,42], obesity (1/36, 3%) [32], Parkinson disease (1/36, 3%) [36], sleep apnea syndrome (1/36, 3%) [38], spinal cord injury (1/36, 3%) [35], and chronic pain (3/36, 8%) [47,54,57]. One study focused on 2 diseases [40]. Of the 36 studies, 13 (36%) [11,27-29,31,44-46,48-51] used ECAs as auxiliary aids to deliver health information to promote screening. The goals of the ECA-led interventions varied, with 31% (11/36) of the studies aiming to promote self-management [10,26,33-35,37,39,40,52,55,56], including activities such as daily walking, medication taking, healthy eating, monitoring symptoms, and interaction management. ECAs were also used for mental therapy (3/36, 8%) [47,54,57], genetic counseling (3/36, 8%) [30,53], assisted living (2/36, 6%) [41,42], diagnosis (1/36, 3%) [38], and education (1/36, 3%) [43]. Of the 36 studies, 2 (6%) [32,36] used ECAs to promote exercise. ECAs were delivered through a variety of means in the included studies, with the majority being deployed on web-based platforms (15/36, 42%) [31,32,43-51,53,54,57], followed by smartphone apps (10/36, 28%) [10,11,26,33,34,37,38,52,55,56], tablet-based systems (7/36, 19%) [27,29,36,39-42], and computer-based systems (4/36, 11%) [28,30,35,46].

Design and Design Processes of ECAs

ECA Identities and Communication Modalities

Across 67% (24/36) of the studies, we found that ECAs were given 9 different names, including ALEX [31,45,46,48,51], Tanya [30,34,35,52,53], Danya [43], Anne [41,42,47,57], Sylvia [40], Laura [10,26,37], Vitoria [39], iHeartHelper [33], and Ellie [54]. Of 36 studies, 14 (39%) reported that ECAs assumed the role of virtual health coaches or counselors [10,26,30,32,34-37,43,47,52,53,57], whereas others described them as virtual health assistants (n=9, 25%) [11,31,33,39,44,48-50], virtual clinicians (n=3, 8%) [45,46,51], health providers (n=2, 6%) [27,29], a virtual human interviewer (n=1, 3%) [54], and a semiexpert (n=1, 3%) [40]. Of the 36 studies, only 1 (3%) [40] provided basic information about the ECA, such as age, place of residence, and education level. In terms of embodiment, ECAs were typically presented as middle-aged women, as shown in the screenshots provided in 64% (23/36) of the studies [10,26,30,32-43,47,52-57]. The visual appearance of ECAs varied across studies. Of the 36 studies, 6 (17%) offered 4 different ECAs, whose appearances could be matched to the participants' race and gender [11,31,45,46,48], and 3 (8%) studies [27,28,51] reported that ECAs had the appearance of an African American man. Of the 36 studies, 5 (14%) [44-46,50,51] reported that the visual appearances of ECAs were based on formative focus groups and think aloud interviews with the target population, improving the details of the character's appearance (eg, adding white coats, removing white hair, and changing hairstyles). Of the 36 studies, 3 (8%) used a 3D model of a digital examination room as the virtual environment behind the ECA, resembling local clinical rooms [31,45,46].

Regarding the communication modalities of ECAs, 14 agents were able to communicate with users through verbal and nonverbal behaviors [10,27,30,32,33,35,36,38,39,41,45,54,55,57]. Nonverbal behavior included facial displays of emotions, gaze shifts, eyebrow raises, head nods, body posture shifts, and hand gestures [30,37,53]. The specific presentation of these conversational modalities was detailed in 33% (12/36) of the studies. For example, 3% (1/36) of the studies reported that an agent with 2D animation would blink her eyes every 10 seconds and move her mouth for a fixed period after a new sentence appeared on the screen [40]. Another study mentioned that the virtual human had an idle breathing animation in a sitting posture and featured high-fidelity voices recorded by professional voice talents of the same race and gender [31]. To address communication shortcomings, such as lower eyesight accuracy or hearing impairment, the development of the interface took into account the needs of patients. This included customized speech speed [33] and text captions of the audio with the recommended font size [31]. To address safety concerns, user contributions to the dialogue were fully

XSL•FO
RenderX

constrained and made by selecting an utterance from a menu [30,43,53]. Users had the option to respond to the agent by speaking, inputting an SMS text message, or touching an option on the screen in 8% (3/36) of the studies [10,26,37]. Only 1 ECA was designed to portray a listener responsive to the respondent's nonverbal speaking behavior [54].

Personalization in Intervention Content and Delivery

Personalization was a common feature in the ECAs used in the studies to customize the content and delivery to suit individual users. This involved addressing users by their names and appropriate time contexts [34,52]. In addition, reminders, warnings, or alerts were provided based on individually reported data [55], and feedback on current progress toward the set goals was given [36]. Out of the 36 studies, in 1 (3%) study, users could respond to ECA's queries using forced-choice text options that would trigger different responses, allowing the system to interact responsively in a personalized manner with the users [57]. The personalization in other studies was achieved based on various channels of information. Out of the 36 studies, 3 (8%) personalized the script logic based on the clinical targets provided by users' health care professionals and users' responses during the interactions [10,26,37]. To tailor the therapy for each patient, the ECA utterances, the patient's responses, custom goals, and overall objective metrics such as time in the simulation were stored in an SQL database [47]. Only 2 (6%) out of the 36 studies provided dynamic adaptive genetic counseling for breast cancer based on the user model's current state and the discourse context [30,53].

ECA Technology and Theories or Principles

From a technological perspective, the physical appearances of ECAs were primarily created using 3D character modeling and animation software, such as the Unity 3D (Unity Technologies) game engine [30,35,38,39,47,53], iClone (Reallusion) [28], and Fuse (Adobe Systems) [31]. Only 1 study used 2D animation implemented with scalable vector graphics and HTML animations [40]. Regarding communication modalities, speech recognition technology was used in 17% (6/36) of the studies to allow users to answer the agent's questions orally [10,26,37,38,41,42], whereas speech synthesis technology was used to generate the agent's spoken responses [30,34,35,39,41,42,52,53,56]. For example, a text-to-speech software, such as Speech2Go (Harpo), was used to convert written dialogues into audio files [39]. Nonverbal behaviors of the ECAs were generated using the Behavior Expression Animation Toolkit text-to-embodied speech engine [30,35,47], and the LipSync Generator was used to synchronize the agent's lip movements with the spoken words [39]. Motion capture technology was used in 6% (2/36) of the studies to record the voice and gestures of a real person, adding a level of realism to the ECA's behavior [28,38]. Various approaches were used in dialogue management. These included using a rules engine to determine the agent's responses based on contextual information [39], a hierarchical transition network–based dialogue engine [35,47,53], and a scenario manager based on decision trees [38]. In addition, 8% (3/36) of the studies used voice recognition with prescripted conversational elements and a sophisticated script logic [10,26,37].

Most ECAs in the studies applied theories or therapy-derived principles to guide their content, visual design, and communication strategies. Table 2 shows 13 different theories and principles mentioned across 14 studies. Of these 14 studies, 4 (29%) incorporated behavior change theories to guide intervention content and delivery [32,33,37,39] and 10 (71%) based their interventions on theories derived from communication [30,53], technology [27,28,44,50,51], or psychological domains [27,28,32,37,49,53]. A total of 11 studies have reported the role of theory in design processes [27,28,30,37,39,44,49-51,53]. For example, the design of the software prototype intervention and dialogue creation was guided by the Behavior Change Wheel [39].

Table 2. Theories or principles informing the embodied conversational agent (ECA)–based interventions (n=14).

Theories or principles guiding ECA designs	Value, n (%)
Information, Motivation, and Behavioral skills model	1 (7)
Behavioral Theory	1 (7)
Social cognitive theory	1 (7)
Technology Acceptance Model	1 (7)
Transtheoretical Model	1 (7)
Behavior Change Wheel	1 (7)
Behavior change techniques	1 (7)
Unified Theory of Acceptance and Use of Technology	2 (14)
Cognitive Theory of Multimedia Learning	3 (21)
Gamification elements	2 (14)
Heuristic-Systematic Model	1 (7)
Risk Communication theories	3 (21)
Modality, Agency, Interactivity, and Navigability Model	2 (14)

XSL•FO
RenderX

Evaluation Measures and Outcomes

Overview

The studies included in this review reported on both acceptability and effectiveness. Acceptability refers to the emotional attitude toward new digital health interventions, use intentions, actual use, and satisfaction [58]. Effectiveness refers to the effect of ECA-based interventions on health-related outcomes. A total of 4 studies only described protocols [33,55-57], which were not considered in this section.

Acceptability

A total of 31 studies reported the acceptability of ECAs. Of the 31 studies, 8 (26%) measured acceptability using validated questionnaires or adapted versions of previously published questionnaires [27,32,38,39,41,42,44,47]. The System Usability Scale was used in 13% (4/31) of the studies to measure the acceptability of the entire system [39,41,42,47]. Other instruments used included the Acceptability E-Scale [38], the Almere Model [41,42], the closeness scale [41,42], the Technology Acceptance Model [47], the Computer Self-Efficacy Scale [27], the Working Alliance Inventory [32], and the ECA Trust Questionnaire [38]. For example, 1 study assessed participants' perceived ease of use of the system using a 24-item scale adapted from the Unified Theory of Acceptance and Use of Technology [27]. Another study used the Portuguese version of the System Usability Scale and calculated an aggregate average score of 73.75, corresponding to a borderline rating of excellent [39]. A total of 2 studies on dementia illustrated that ECA Anne received a mean score of 66.2 [42] and 67.1 [41]. In addition to questionnaires, customized items were used in 14 studies to assess users' satisfaction with ECAs and the overall systems [10,27,29-31,34,36,38,40,43,45-48,53,54], whereas interviews and focus groups were conducted in 18 studies to explore more topics [10,11,26,28-31,35,39-44,49-52]. User satisfaction concerns items related to constructs such as liking, trust, ease of use, and desire to continue using the ECA; for example, "How much did you like Tanya?" [30]. Objective measures of user engagement with ECAs were reported in 8 studies [32,34,36,37,40-42,52]. These measures included the number of log-ins to the agent application, the time and number of interactions with the ECAs, and the time of program use. For example, 1 study showed that the time for participants to use relationship agents ranged from 3 to 30 days, and the number of log-ins to the relational agent ranged from 4 to 43 days [34]. Telemetry data were used in 2 studies to detect problems and evaluate the status quo [41,42]. In total, 2 studies measured use over time, all showing a decrease [37,40]; for example, "the program use, including the number of chats and number of blood glucose uploads, reduced over time of the program access."

Effectiveness

The effectiveness of ECA intervention was evaluated in 16 studies [27,30-32,34,36,37,43,46-48,52-54]. Behavioral outcomes were examined in 4 studies [32,34,36,52]. Participants in the intervention arm showed higher daily activity and adherence to therapy than those in the control arm [52]. Gait speed and the 6-minute walk test significantly improved after the intervention (all P=.02) [36]. A total of 5 studies assessed participants' knowledge using questionnaires, and it was found that their cancer knowledge significantly improved after the intervention [27,30,43,53]. Participants in the adaptive condition had significantly greater knowledge gain than participants in the nonadaptive condition and control condition [53]. A total of 3 studies assessed changes in quality of life using questionnaires [34,37,52], and 2 studies [37,52] found a positive difference in quality of life levels between participants who engaged with the ECA and those who did not. However, there was no significant difference in glycated hemoglobin change between participants in the intervention and control groups [37]. For motivational outcomes, 3 studies assessed changes in users' motivation [27,32,34]. Self-efficacy [27,32] and patient activation [34] were assessed using questionnaires. Symptom improvements were assessed in 3 studies using questionnaires [37,47,54]. The results showed that the use of ECAs led to a greater reduction in pain interference and a marginally greater reduction in pain intensity compared with standard interviews [54]. Safety was assessed in 1 study, where participants were asked standardized questions about adverse events such as falls, diseases, injuries, and the use of any medical services [36].

Discussion

Principal Findings

This scoping review specifically targeted ECAs applied for chronic diseases in health care, as we aimed to inform technology developers and health care professionals of the technological possibilities and the evidence base. Our scoping review identified 32 studies and 4 ongoing clinical trials, with most papers published from 2020 onward. The most commonly reported chronic diseases were cancer, atrial fibrillation, and type 2 diabetes. The review found that ECAs were predominantly defined as female coaches or counselors interacting with users through voice and nonverbal behavior. In addition, multiple technologies and theories were applied to the design activities of ECA-delivered interventions. A combination of effectiveness and acceptability was typically assessed. Results from the studies reveal that ECAs have the potential to enhance engagement in self-care for chronic diseases, although the evidence on their effectiveness remains inconclusive.

The identified studies were not geographically diverse, with 75% (27/36) conducted in the United States and none conducted in Africa or Asia. This lack of diversity in research locations limits the generalizability of the findings, as they are embedded in Western cultures. Given the global prevalence of chronic conditions and the need for health care system–specific solutions, future research should strive to include diverse geographies to ensure the relevance of interventions in different health care systems. Among the 36 included studies, 6 (17%) explored stakeholders' opinions. It has been shown that there were very positive relationships between homecare providers' and patients' perceptions of virtual agents [38], mirroring the findings by Heerink et al [59], which reaffirmed the finding that social influence plays an important role in the user acceptance of a social agent. In addition, people who were retired, highly educated, and engaged with the app were overrepresented in

XSL•FO
RenderX

some studies in the interviewed sample of participants [10,26,39]. Health literacy is relevant to the development, accessibility, and successful implementation of eHealth. eHealth interventions focused on health literacy have the potential to reduce disparities in susceptible populations, where limited health literacy is more prevalent [60]. To incorporate interventions into clinical practice effectively and ensure widespread adoption, it is necessary to identify the experiences and needs of stakeholders and users who are less autonomous and less experienced with technology.

Owing to the diversity of the design activities reported in the studies, it can be challenging to draw general findings. However, some commonalities can be observed regarding the identities of the ECAs. A common design feature is giving an ECA a name, which may enhance its social presence. Through the lens of the Technology Acceptance Model, social presence, or the general sense of being with another person, is relevant to patients accepting agents because perceptions of social presence can lead to a desire for future interaction [44]. Another design feature is portraying an ECA as a coach, indicating that a relaxed and nonjudgmental role may be more successful in building a supportive relationship than an authoritative role. It is important to note that preferences may vary among different patient populations, as a recent systematic review found that racial and ethnic minority groups most often prefer a paternalistic model of health decision-making [61]. The research found that the agent's role, such as being called a virtual physician or health care assistant, influenced the user's expectations for the agent's appearance [31]. Some studies reported that users tested the prototype and commented on character details that informed refinement. For example, in 1 study, an ECA was regarded as a medical authority, and changes were made to enhance its appearance, such as adding a name badge, updating clothing to include a white medical coat, and adjusting perceptions of the agent's age [44]. Most ECAs are depicted as middle-aged women, which aligns with previous reports [15]. This may be related to the gender stereotypes associated with health guidance tasks [62]. For chronic diseases involving sensitive information, such as cancer, agents of the same race and gender are often preferred. In conclusion, developers of ECAs will have to learn to determine which identities to prioritize for their ECA but can begin analyzing this by determining their target population and their specific contexts.

Regarding communication modalities, ECAs reported in the literature mimic human conversation using interactive voice recognition, allowing users to interact with the system through voice rather than only navigating with the touch screen. When applied in the context of health communication, unconstrained speech input from conversational assistants has been found to pose patient safety risks [63]. In addition, the quality of automatic speech recognition and synthesis remains a technical problem, and there is room for improvement [41]. Nonverbal behavior has a deep impact on the process and outcome of communication, with approximately 65% of social meaning derived from nonverbal behavior [64]. In 1 study, an ECA lacked inflection in voice and exhibited a limited number of random body movements that did not align with the context of the conversation. This mismatch between a character's speech and expected facial expressions and body movements can create an unnatural dissonance and affect acceptability, known as the "uncanny valley" [65]. Perfecting natural communication via congruence between verbal and nonverbal cues is critical [66]. This requires understanding natural behaviors and the biological processes underlying them, so as to develop efficient algorithms to implement a convincing simulation via ECAs.

Our scoping review showed that various forms of personalization are used for building content and delivery. The most common approach to personalization is providing more specific feedback based on user responses and health-related data. In this review, examples of conversational adaptation included using an individual's current knowledge state, preferred information processing method, and other user traits such as health literacy and breast cancer risk level according to the user models [30]. From a technical point of view, future personalization may involve tailoring interventions to each user individually or even to their current status at the moment of interaction [67]. For example, Sripian et al [68] developed methods to measure and estimate the user's emotions based on biological information. Using these methods, the ECA could react to negative user's emotions and provide assistance. Johal et al [69] used a multisensory environment to detect when a user would like to interact with a device. Promoting the health and well-being of people through human-centered technologies requires the partnership of research networks, medical scientists, technology developers, patients, and their formal and informal caregivers.

A total of 13 theories and principles were applied to the included studies to create ECA-based interventions. However, it is difficult to determine which theory or principle is most suitable in the context of chronic diseases, which is consistent with the findings of Kramer et al [15]. Different theories and principles may need to be applied in different situations. Behavior change theory is usually combined with various behavior change technologies to guide conversational scripts and algorithms. The Modality, Agency, Interactivity, and Navigability Model, used in 2 studies, is an organizational framework for designing a multimedia learning environment. It helps to understand how interface features affect the user's psychology through 4 affordances [50,51]. For example, navigability refers to the user's ability to access information and complete tasks, which can aid developers in designing the navigation structure and interface layout of ECAs (eg, clear menus, navigation options, and visual cues), making it easier for users to find the desired information and functionalities. The cognitive theory of multimedia learning, which is a more recent iteration of the cognitive load theory, is also mentioned. The cognitive theory of multimedia learning recognizes that the capacity to process and store new information is limited, and presenting multimodal information redundantly may overload the cognitive capacity [70]. ECAs can be considered as a form of multimodal learning, using multiple sensory channels to enhance learning and information processing. In addition to information processing theories, future research may adopt other interpersonal communication theories to maximize the persuasiveness of ECA, as tailoring to other constructs beyond information

XSL•FO
RenderX

processing modes may improve the ECA's effectiveness in adherence motivation [53].

The evaluation measures used to assess ECAs and their effects on chronic conditions are broad and not unified. This is similar to other reviews on this topic focusing on conversational agent–delivered mental health interventions [22,71]. Several researchers have developed surveys to measure specific issues. The perceptions of the agents are most frequently reported in evaluating ECAs; for example, "How well do the words below describe Laura?" [10]. The respondents rated a range of positive and negative traits (eg, friendly, expert, reliable, annoying, and boring) using a Likert-type scale. However, there is variation in opinions and limited evidence regarding which agent characteristics are especially important. Future ECA design studies should explore both the perception of the characteristics of the designed agent and the perceived importance of these characteristics for an agent in a specific context [40]. Research suggests that incorporating humanlike characteristics in ECA design did increase user engagement [26]. However, the program use was reduced over time of the program access, and there was a dose-response relationship between the number of chats and the change in the quality of life score [37]. The dose-response relationship between the level of app use and its effectiveness suggests that more effort is required to improve the maintenance of program use over time. Regarding effectiveness-related outcomes, improvements in behavior, knowledge, symptoms, and quality of life were observed in the reviewed studies, whereas physiological data did not show significant changes. Future studies should aim to measure the efficacy of the relational agent using objective measures rather than relying on self-reporting, which is subject to multiple biases [72]. In addition to the evaluation measures mentioned earlier, the studies reviewed in this paper did not measure the cost, efficiency, or productivity improvements associated with using ECAs in health care settings. This is a significant limitation as it prevents us from determining whether ECAs are cost-effective compared with alternative approaches or if they can enhance the work of health care professionals.

Strengths and Limitations

This scoping review has several strengths and some limitations. One of the strengths of this study is that it offers some important insights into the condition of ECAs in health care, with a focus on design activities. The study was reported according to the PRISMA-ScR guidelines, which enhanced the quality of the review. In addition, the study selection process and data extraction process were conducted by 2 reviewers independently to reduce selection bias. Agreement between the reviewers was very good for the study of both the selection process and the data extraction process.

The comprehensive literature search included 6 databases; however, the lack of standardized terminology in this field may have resulted in the omission of some related papers. In addition, limiting the search to papers published in English may result in the exclusion of ECAs developed for chronic diseases in other languages. Finally, as this is a scoping review without synthesis of the evidence, we present the outcomes as reported by the authors of primary studies; therefore, no determinations or recommendations are made regarding the appropriateness or utility of the outcomes.

Recommendations for Future Research and Practice

One important direction for future work is to improve the interaction between ECAs and users to minimize the potential impact of uncontrolled variables (eg, user preferences) and allow for a better study of the impact of agent adaptation. First, researchers can focus on leveraging the unique characteristics of the conversational medium to personalize interactions. For example, capturing prosodic features in users' speech can be used to automatically detect changes in mood or speech pathologies. This information will allow ECAs to provide adaptive information and services tailored to the user's needs. Second, the behavior data of users can be technically collected and analyzed. Compared with the data collected by the questionnaires, this data-driven quantitative approach is mediated through data objectively describing their behavior, and the domain behind the data is qualitatively understood through interviews [42,73]. Finally, larger longitudinal studies need to be conducted to measure the effects of ECAs on specified subgroups over time, which will help to identify the components of the multifaceted intervention contributing to increased acceptability and positive outcomes.

Conclusions

This scoping review followed a strict methodology to summarize and discuss the development and evaluation of ECAs in chronic diseases. The findings showed that the ECAs have been increasingly used in recent years and may have significant potential to deliver effective health interventions to promote self-care. However, there is a requirement for technological advancements in embodiment; communication methods; personalized strategies; and a deeper comprehension of user preferences regarding appearance, animation, and personalized content. In addition, there is a lack of reliable and comparable evidence in user-centered evaluation approaches. Future studies should incorporate cost, efficiency, and productivity measures to provide a comprehensive evaluation of the benefits of implementing ECAs in health care.

Acknowledgments

This study was funded by General Public Welfare Projects of Science and Technology Department of Zhejiang Province of China (LGF20G030009).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items of Systematic Reviews and Meta-Analyses) checklist.
[DOC File , 57 KB - jmir_v26i1e47134_app1.doc]

Multimedia Appendix 2
Search terms and database search.
[DOCX File , 15 KB - jmir_v26i1e47134_app2.docx]

Multimedia Appendix 3
Data extraction form.
[DOCX File , 16 KB - jmir_v26i1e47134_app3.docx]

Multimedia Appendix 4
Description of the included studies.
[XLSX File (Microsoft Excel File), 34 KB - jmir_v26i1e47134_app4.xlsx]

References

1. Boersma P, Black LI, Ward BW. Prevalence of multiple chronic conditions among US adults, 2018. Prev Chronic Dis 2020 Sep 17;17:E106 [FREE Full text] [doi: 10.5888/pcd17.200130] [Medline: 32945769]
2. Shilian H, Jing W, Cui C, Xinchun W. Analysis of epidemiological trends in chronic diseases of Chinese residents. Aging Med (Milton) 2020 Dec;3(4):226-233 [FREE Full text] [doi: 10.1002/agm2.12134] [Medline: 33392427]
3. Alonso J, Ferrer M, Gandek B, Ware JEJ, Aaronson NK, Mosconi P, et al. Health-related quality of life associated with chronic conditions in eight countries: results from the International Quality of Life Assessment (IQOLA) Project. Qual Life Res 2004 Mar;13(2):283-298. [doi: 10.1023/b:qure.0000018472.46236.05] [Medline: 15085901]
4. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
5. Bickmore TW, Pfeifer LM, Jack BW. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009 Presented at: CHI '09; April 4-9, 2009; Boston, MA. [doi: 10.1145/1518701.1518891]
6. Robertson S, Solomon R, Riedl M, Gillespie TW, Chociemski T, Master V. The visual design and implementation of an embodied conversational agent in a shared decision-making context (eCoach). In: Proceedings of the International Conference on Learning and Collaboration Technologies. 2015 Presented at: LCT 2015; August 2-7, 2015; Los Angeles, CA. [doi: 10.1007/978-3-319-20609-7_40]
7. Martínez-Miranda J. Embodied conversational agents for the detection and prevention of suicidal behaviour: current applications and open challenges. J Med Syst 2017 Sep;41(9):135. [doi: 10.1007/s10916-017-0784-6] [Medline: 28755270]
8. Grekin ER, Beatty JR, Ondersma SJ. Mobile health interventions: exploring the use of common relationship factors. JMIR Mhealth Uhealth 2019 Apr 15;7(4):e11245. [doi: 10.2196/11245] [Medline: 30985281]
9. Baptista S, Trawley S, Pouwer F, Oldenburg B, Wadley G, Speight J. What do adults with type 2 diabetes want from the "Perfect" app? Results from the second diabetes MILES: Australia (MILES-2) study. Diabetes Technol Ther 2019 Jul;21(7):393-399. [doi: 10.1089/dia.2019.0086] [Medline: 31166804]
10. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. Acceptability of an embodied conversational agent for type 2 diabetes self-management education and support via a smartphone app: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e17038 [FREE Full text] [doi: 10.2196/17038] [Medline: 32706734]
11. Vilaro MJ, Wilson-Howard DS, Neil JM, Tavassoli F, Zalake MS, Lok BC, et al. A subjective culture approach to cancer prevention: rural Black and White adults' perceptions of using virtual health assistants to promote colorectal cancer screening. Health Commun 2022 Aug;37(9):1123-1134. [doi: 10.1080/10410236.2021.1910166] [Medline: 33876658]
12. Burton C, Szentagotai Tatar A, McKinstry B, Matheson C, Matu S, Moldovan R, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare 2016 Sep;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]
13. Curtis RG, Bartel B, Ferguson T, Blake HT, Northcott C, Virgara R, et al. Improving user experience of virtual health assistants: scoping review. J Med Internet Res 2021 Dec 21;23(12):e31737 [FREE Full text] [doi: 10.2196/31737] [Medline: 34931997]

XSL•FO
RenderX

14. Bhandari U, Chang K, Neben T. Understanding the impact of perceived visual aesthetics on user evaluations: an emotional perspective. Inf Manage 2019 Jan;56(1):85-93. [doi: 10.1016/j.im.2018.07.003]
15. Kramer LL, Ter Stal S, Mulder BC, de Vet E, van Velsen L. Developing embodied conversational agents for coaching people in a healthy lifestyle: scoping review. J Med Internet Res 2020 Feb 06;22(2):e14058 [FREE Full text] [doi: 10.2196/14058] [Medline: 32022693]
16. ter Stal S, Kramer LL, Tabak M, op den Akker H, Hermens H. Design features of embodied conversational agents in eHealth: a literature review. Int J Hum Comput Stud 2020 Jun;138:102409. [doi: 10.1016/j.ijhcs.2020.102409]
17. Loveys K, Sebaratnam G, Sagar M, Broadbent E. The effect of design features on relationship quality with embodied conversational agents: a systematic review. Int J of Soc Robotics 2020 Sep 24;12(6):1293-1312. [doi: 10.1007/s12369-020-00680-7]
18. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 09;19(5):e151 [FREE Full text] [doi: 10.2196/jmir.6553] [Medline: 28487267]
19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
20. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. J Med Internet Res 2020 Jun 05;22(6):e18301 [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]
23. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
24. Chronic diseases. Public Health Agency of Canada. URL: https://www.canada.ca/en/public-health/services/chronic-diseases.html [accessed 2023-07-07]
25. Ruttkay Z, Dormann C, Noot H. Embodied conversational agents on a common ground. In: Ruttkay Z, Pelachaud C, editors. From Brows to Trust. Dordrecht, The Netherlands: Springer; 2004:27-66.
26. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. User experiences with a type 2 diabetes coaching app: qualitative study. JMIR Diabetes 2020 Jul 17;5(3):e16692 [FREE Full text] [doi: 10.2196/16692] [Medline: 32706649]
27. Owens OL, Felder T, Tavakoli AS, Revels AA, Friedman DB, Hughes-Halbert C, et al. Evaluation of a computer-based decision aid for promoting informed prostate cancer screening decisions among African American men: iDecide. Am J Health Promot 2019 Feb;33(2):267-278. [doi: 10.1177/0890117118786866] [Medline: 29996666]
28. Owens OL, Friedman DB, Brandt HM, Bernhardt JM, Hébert JR. An iterative process for developing and evaluating a computer-based prostate cancer decision aid for African American men. Health Promot Pract 2015 Sep;16(5):642-655. [doi: 10.1177/1524839915585737] [Medline: 25983084]
29. Owens OL, Kim S, Tavakoli AS. Are decision aids leading to shared prostate cancer screening decisions among African-American men?: iDecide. Cancer Causes Control 2019 Jul 15;30(7):713-719. [doi: 10.1007/s10552-019-01179-x] [Medline: 31093861]
30. Zhou S, Bickmore T. Automating cancer genetic counseling with an adaptive pedagogical agent. In: Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI EA '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411763.3451675]
31. Zalake M, Tavasolli F, Griffin L, Krieger J, Lok B. Internet-based tailored virtual human health intervention to promote colorectal cancer screening: design guidelines from two user studies. Intell Virtual Agents 2021 Jun;15(2):147-162 [FREE Full text] [Medline: 34027518]
32. Watson A, Bickmore T, Cange A, Kulshreshtha A, Kvedar J. An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial. J Med Internet Res 2012 Jan 26;14(1):e1 [FREE Full text] [doi: 10.2196/jmir.1629] [Medline: 22281837]
33. Zhang L, Babu SV, Jindal M, Williams JE, Gimbel RW. A patient-centered mobile phone app (iHeartU) with a virtual human assistant for self-management of heart failure: protocol for a usability assessment study. JMIR Res Protoc 2019 May 23;8(5):e13502 [FREE Full text] [doi: 10.2196/13502] [Medline: 31124472]
34. Magnani JW, Schlusser CL, Kimani E, Rollman BL, Paasche-Orlow MK, Bickmore TW. The atrial fibrillation health literacy information technology system: pilot assessment. JMIR Cardio 2017;1(2):e7 [FREE Full text] [doi: 10.2196/cardio.8543] [Medline: 29473644]

XSL•FO
RenderX

35. Shamekhi A, Trinh H, Bickmore TW, DeAngelis TR, Ellis T, Houlihan BV, et al. A virtual self-care coach for individuals with spinal cord injury. In: Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility. 2016 Presented at: ASSETS '16; October 23-26, 2016; Reno, Nevada. [doi: 10.1145/2982142.2982199]
36. Ellis T, Latham NK, DeAngelis TR, Thomas CA, Saint-Hilaire M, Bickmore TW. Feasibility of a virtual exercise coach to promote walking in community-dwelling persons with Parkinson disease. Am J Phys Med Rehabil 2013 Jun;92(6):472-81; quiz 482 [FREE Full text] [doi: 10.1097/PHM.0b013e31828cd466] [Medline: 23552335]
37. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My Diabetes Coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 05;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]
38. Dupuy L, Micoulaud-Franchi JA, Philip P. Acceptance of virtual agents in a homecare context: evaluation of excessive daytime sleepiness in apneic patients during interventions by continuous positive airway pressure (CPAP) providers. J Sleep Res 2021 Apr;30(2):e13094. [doi: 10.1111/jsr.13094] [Medline: 32557996]
39. Balsa J, Félix I, Cláudio AP, Carmo MB, Silva IC, Guerreiro A, et al. Usability of an intelligent virtual assistant for promoting behavior change and self-care in older people with type 2 diabetes. J Med Syst 2020 Jun 13;44(7):130. [doi: 10.1007/s10916-020-01583-w] [Medline: 32533367]
40. Ter Stal S, Sloots J, Ramlal A, Op den Akker H, Lenferink A, Tabak M. An embodied conversational agent in an eHealth self-management intervention for chronic obstructive pulmonary disease and chronic heart failure: exploratory study in a real-life setting. JMIR Hum Factors 2021 Nov 04;8(4):e24110 [FREE Full text] [doi: 10.2196/24110] [Medline: 34734824]
41. Stara V, Vera B, Bolliger D, Rossi L, Felici E, Di Rosa M, et al. Usability and acceptance of the embodied conversational agent Anne by people with dementia and their caregivers: exploratory study in home environment settings. JMIR Mhealth Uhealth 2021 Jun 25;9(6):e25891 [FREE Full text] [doi: 10.2196/25891] [Medline: 34170256]
42. Stara V, Vera B, Bolliger D, Paolini S, de Jong M, Felici E, et al. Toward the integration of technology-based interventions in the care pathway for people with dementia: a cross-national study. Int J Environ Res Public Health 2021 Oct 02;18(19):10405 [FREE Full text] [doi: 10.3390/ijerph181910405] [Medline: 34639704]
43. Gunn C, Maschke A, Bickmore T, Kennedy M, Hopkins MF, Fishman MD, et al. Acceptability of an interactive computer-animated agent to promote patient-provider communication about breast density: a mixed method pilot study. J Gen Intern Med 2020 Apr;35(4):1069-1077 [FREE Full text] [doi: 10.1007/s11606-019-05622-2] [Medline: 31919723]
44. Vilaro MJ, Wilson-Howard DS, Zalake MS, Tavassoli F, Lok BC, Modave FP, et al. Key changes to improve social presence of a virtual health assistant promoting colorectal cancer screening informed by a technology acceptance model. BMC Med Inform Decis Mak 2021 Jun 22;21(1):196 [FREE Full text] [doi: 10.1186/s12911-021-01549-z] [Medline: 34158046]
45. Cooks EJ, Duke KA, Flood-Grady E, Vilaro MJ, Ghosh R, Parker N, et al. Can virtual human clinicians help close the gap in colorectal cancer screening for rural adults in the United States? The influence of rural identity on perceptions of virtual human clinicians. Prev Med Rep 2022 Dec;30:102034 [FREE Full text] [doi: 10.1016/j.pmedr.2022.102034] [Medline: 36531088]
46. Cooks EJ, Duke KA, Neil JM, Vilaro MJ, Wilson-Howard D, Modave F, et al. Telehealth and racial disparities in colorectal cancer screening: a pilot study of how virtual clinician characteristics influence screening intentions. J Clin Transl Sci 2022 Apr 08;6(1):e48 [FREE Full text] [doi: 10.1017/cts.2022.386] [Medline: 35619640]
47. Reilly ED, Volonte M, Bickmore T. An online therapeutic intervention for veterans patients suffering with chronic pain. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2023 Presented at: IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW); March 25-29, 2023; Shanghai, China. [doi: 10.1109/vrw58643.2023.00039]
48. Krieger JL, Neil JM, Duke KA, Zalake MS, Tavassoli F, Vilaro MJ, et al. A pilot study examining the efficacy of delivering colorectal cancer screening messages via virtual health assistants. Am J Prev Med 2021 Aug;61(2):251-255 [FREE Full text] [doi: 10.1016/j.amepre.2021.01.014] [Medline: 33888362]
49. Vilaro MJ, McAlindon K, Mertens G, Ashley T, Zalake M, Cooks EJ, et al. Information architects: using community-engaged and qualitative methods to design a technology-based nutrition and cancer risk intervention for rural adults. Cancer Control 2023;30:10732748221130162 [FREE Full text] [doi: 10.1177/10732748221130162] [Medline: 36919704]
50. Vilaro MJ, Wilson-Howard DS, Griffin LN, Tavassoli F, Zalake MS, Lok BC, et al. Tailoring virtual human-delivered interventions: a digital intervention promoting colorectal cancer screening for Black women. Psychooncology 2020 Dec;29(12):2048-2056 [FREE Full text] [doi: 10.1002/pon.5538] [Medline: 32893399]
51. Wilson-Howard D, Vilaro MJ, Neil JM, Cooks EJ, Griffin LN, Ashley TT, et al. Development of a credible virtual clinician promoting colorectal cancer screening via telehealth apps for and by Black men: qualitative study. JMIR Form Res 2021 Dec 24;5(12):e28709 [FREE Full text] [doi: 10.2196/28709] [Medline: 34780346]
52. Guhl E, Althouse AD, Pusateri AM, Kimani E, Paasche-Orlow MK, Bickmore TW, et al. The atrial fibrillation health literacy information technology trial: pilot trial of a mobile health app for atrial fibrillation. JMIR Cardio 2020 Sep 04;4(1):e17162 [FREE Full text] [doi: 10.2196/17162] [Medline: 32886070]
53. Zhou S, Bickmore T. A virtual counselor for breast cancer genetic counseling: adaptive pedagogy leads to greater knowledge gain. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3517553]

XSL•FO
RenderX

54. Junghaenel DU, Schneider S, Lucas G, Boberg J, Weinstein FM, Richeimer SH, et al. Virtual human-delivered interviews for patients with chronic pain: feasibility, acceptability, and a pilot randomized trial of standard medical, psychosocial, and educational interviews. Psychosom Med 2023 Sep 01;85(7):627-638. [doi: 10.1097/PSY.0000000000001228] [Medline: 37363989]
55. Althouse AD, Abebe KZ, Paasche-Orlow MK, Lalama CM, Ferry D, Lancet M, et al. Design, rationale, and baseline characteristics of a randomized controlled trial evaluating a mobile relational agent to enhance atrial fibrillation self-care. Contemp Clin Trials 2023 Jan;124:107015. [doi: 10.1016/j.cct.2022.107015] [Medline: 36402276]
56. Magnani JW, Ferry D, Swabe G, Martin D, Chen X, Brooks MM, et al. Design and rationale of the mobile health intervention for rural atrial fibrillation. Am Heart J 2022 Oct;252:16-25 [FREE Full text] [doi: 10.1016/j.ahj.2022.05.023] [Medline: 35691371]
57. Reilly ED, Kathawalla UK, Robins HE, Heapy AA, Hogan TP, Waring ME, et al. An online acceptance and mindfulness intervention for chronic pain in veterans: development and protocol for a pilot feasibility randomized controlled trial. JMIR Res Protoc 2023 Mar 07;12:e45887 [FREE Full text] [doi: 10.2196/45887] [Medline: 36881446]
58. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 06;22(7):e17256 [FREE Full text] [doi: 10.2196/17256] [Medline: 32628122]
59. Heerink M, Kröse B, Evers V, Wielinga B. Assessing acceptance of assistive social agent technology by older adults: the Almere model. Int J of Soc Robotics 2010 Sep 4;2(4):361-375. [doi: 10.1007/s12369-010-0068-5]
60. Bickmore TW, Paasche-Orlow MK. The role of information technology in health literacy research. J Health Commun 2012;17 Suppl 3:23-29. [doi: 10.1080/10810730.2012.712626] [Medline: 23030559]
61. Perez Jolles MP, Richmond J, Thomas KC. Minority patient preferences, barriers, and facilitators for shared decision-making with health care providers in the USA: a systematic review. Patient Educ Couns 2019 Jul;102(7):1251-1262. [doi: 10.1016/j.pec.2019.02.003] [Medline: 30777613]
62. Forlizzi J, Zimmerman J, Mancuso V, Kwak S. How interface agents affect interaction between humans and computers. In: Proceedings of the 2007 conference on Designing pleasurable products and interfaces. 2007 Presented at: DPPI '07; August 22-25, 2007; Helsinki, Finland. [doi: 10.1145/1314161.1314180]
63. Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J Med Internet Res 2018 Sep 04;20(9):e11510 [FREE Full text] [doi: 10.2196/11510] [Medline: 30181110]
64. Rosenthal-von der Pütten AM, Straßmann C, Yaghoubzadeh R, Kopp S, Krämer NC. Dominant and submissive nonverbal behavior of virtual agents and its effects on evaluation and negotiation outcome in different age groups. Comput Hum Behav 2019 Jan;90:397-409. [doi: 10.1016/j.chb.2018.08.047]
65. Seyama J, Nagayama RS. The uncanny valley: effect of realism on the impression of artificial human faces. Presence Teleoperators Virtual Environ 2007 Aug 01;16(4):337-351. [doi: 10.1162/pres.16.4.337]
66. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
67. Rai HK, Cavalcanti Barroso A, Yates L, Schneider J, Orrell M. Involvement of people with dementia in the development of technology-based interventions: narrative synthesis review and best practice guidelines. J Med Internet Res 2020 Dec 03;22(12):e17531 [FREE Full text] [doi: 10.2196/17531] [Medline: 33270034]
68. Sripian P, Mohd Anuardi MN, Kajihara Y, Sugaya M. Empathetic robot evaluation through emotion estimation analysis and facial expression synchronization from biological information. Artif Life Robotics 2021 Sep 16;26(4):379-389. [doi: 10.1007/s10015-021-00696-w]
69. Johal W, Vaufreydaz D. Multi-sensors engagement detection with a robot companion in a home environment. In: Proceedings of the Workshop on Assistance and Service robotics in a human environment at IEEE International Conference on Intelligent Robots and Systems (IROS2012). 2012 Presented at: Workshop on Assistance and Service robotics in a human environment at IEEE International Conference on Intelligent Robots and Systems (IROS2012); October 7-12, 2012; Algarve, Portugal.
70. van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010 Jan;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
71. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548 [FREE Full text] [doi: 10.2196/44548] [Medline: 37074762]
72. Stirratt MJ, Dunbar-Jacob J, Crane HM, Simoni JM, Czajkowski S, Hilliard ME, et al. Self-report measures of medication adherence behavior: recommendations on optimal use. Transl Behav Med 2015 Dec;5(4):470-482 [FREE Full text] [doi: 10.1007/s13142-015-0315-2] [Medline: 26622919]
73. Bickmore TW, Kimani E, Trinh H, Pusateri A, Paasche-Orlow MK, Magnani JW. Managing chronic conditions with a smartphone-based conversational virtual agent. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. 2018 Presented at: IVA '18; November 5-8, 2018; Sydney, Australia. [doi: 10.1145/3267851.3267908]

XSL•FO
RenderX

Abbreviations

ECA: embodied conversational agent
PRISMA-ScR: Preferred Reporting Items of Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by A Mavragani; submitted 13.03.23; peer-reviewed by A Berry, M Zalake, C You; comments to author 19.07.23; revised version received 19.10.23; accepted 29.11.23; published 09.01.24.

Please cite as:
Jiang Z, Huang X, Wang Z, Liu Y, Huang L, Luo X
Embodied Conversational Agents for Chronic Diseases: Scoping Review
J Med Internet Res 2024;26:e47134
URL: https://www.jmir.org/2024/1/e47134
doi:10.2196/47134
PMID:38194260

©Zhili Jiang, Xiting Huang, Zhiqian Wang, Yang Liu, Lihua Huang, Xiaolin Luo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

High-Quality eHealth Websites for Information on Endometriosis: Systematic Search

Diksha Sirohi[1], MBA, MPH; Cecilia Hoi Man Ng[2,3], PhD; Niranjan Bidargaddi[4], PhD; Helen Slater[5], PhD, FACP; Melissa A Parker[6], MNurs (Research); Mary Louise Hull[1], PhD, FRANZCOG; Rebecca O'Hara[1], PhD

[1]Robinson Research Institute, Adelaide Medical School, University of Adelaide, North Adelaide, Australia

[2]School of Clinical Medicine, Division of Obstetrics and Gynaecology, University of New South Wales, Sydney, Australia

[3]Jean Hailes for Women's Health, Melbourne, Australia

[4]Digital Health, College of Medicine and Public Health, Flinders University, Adelaide, Australia

[5]Curtin School of Allied Health, Curtin University, Perth, Australia

[6]Canberra Endometriosis Centre, Centenary Hospital for Women and Children, ACT Health, Canberra, Australia

Corresponding Author:
Diksha Sirohi, MBA, MPH
Robinson Research Institute
Adelaide Medical School
University of Adelaide
55 Ground Floor
King William Road
North Adelaide, 5006
Australia
Phone: 61 0883131527
Email: d.sirohi@uqconnect.edu.au

Abstract

Background: eHealth websites are increasingly being used by community members to obtain information about endometriosis. Additionally, clinicians can use these websites to enhance their understanding of the condition and refer patients to these websites. However, poor-quality information can adversely impact users. Therefore, a critical evaluation is needed to assess and recommend high-quality endometriosis websites.

Objective: This study aimed to evaluate the quality and provide recommendations for high-quality endometriosis eHealth websites for the community and clinicians.

Methods: PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines informed 2 Google searches of international and Australian eHealth websites. The first search string used the terms “endometriosis,” “adenomyosis,” or “pelvic pain,” whereas “Australia” was added to the second search string. Only free eHealth websites in English were included. ENLIGHT, a validated tool, was used to assess the quality across 7 domains such as usability, visual design, user engagement, content, therapeutic persuasiveness, therapeutic alliance, and general subjective evaluation. Websites with a total score of 3.5 or more were classified as “good” according to the ENLIGHT scoring system and are recommended as high-quality eHealth websites for information on endometriosis.

Results: In total, 117 eHealth websites were screened, and 80 were included in the quality assessment. Four high-quality eHealth websites (ie, those that scored 3.5 or more) were identified (Endometriosis Australia Facebook Page, Endometriosis UK, National Action Plan for Endometriosis on EndoActive, and Adenomyosis by the Medical Republic). These websites provided easily understood, engaging, and accurate information. Adenomyosis by the Medical Republic can be used as a resource in clinical practice. Most eHealth websites scored well, 3.5 or more in the domains of usability (n=76, 95%), visual design (n=64, 80%), and content (n=63, 79%). However, of the 63 websites, only 25 provided references and 26 provided authorship details. Few eHealth websites scored well on user engagement (n=18, 23%), therapeutic persuasiveness (n=2, 3%), and therapeutic alliance (n=22, 28%). In total, 30 (38%) eHealth websites scored well on general subjective evaluation.

Conclusions: Although geographical location can influence the search results, we identified 4 high-quality endometriosis eHealth websites that can be recommended to the endometriosis community and clinicians. To improve quality, eHealth websites must provide evidence-based information with appropriate referencing and authorship. Factors that enhance usability, visual design,

XSL•FO
RenderX

user engagement, therapeutic persuasiveness, and therapeutic alliance can lead to the successful and long-term uptake of eHealth websites. User engagement, therapeutic persuasiveness, and therapeutic alliance can be strengthened by sharing lived experiences and personal stories and by cocreating meaningful content for both the community and clinicians. Reach and discoverability can be improved by leveraging search engine optimization tools.

Trial Registration: PROSPERO CRD42020185475; https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=185475&VersionID=2124365

(J Med Internet Res 2024;26:e48243) doi:10.2196/48243

KEYWORDS

digital health; endometriosis; eHealth websites; eHealth; pelvic pain; adenomyosis

Introduction

Endometriosis is a chronic condition causing pain and fertility problems in 5%-10% of natal females globally [1]. It is associated with an average of 6-8 years delay in diagnosis [2,3], which is compounded by uncertainty for health care providers over optimal management [4]. Endometriosis requires long-term therapeutic strategies and appropriate access to medical services to reduce the negative impacts on quality of life such as anxiety, depression, pain during sex, difficulty in doing household tasks, or caring for children due to chronic pelvic pain [5].

People with endometriosis commonly use eHealth [6] websites to seek information about endometriosis when their symptoms persist and are not effectively addressed in traditional health care settings [7,8]. In this context, “What is endometriosis?” was the third highest trending health-related question on Google in 2018 [8,9]. More than 400,000 Google searches on endometriosis are carried out per month in the United States alone [10]. An Australian study showed that a Google search for “endometriosis” increased by 26.4% after the announcement of the 2018 National Action Plan for Endometriosis in Australia [11]. Digital information seeking, including the use of eHealth websites, can contribute to improved health literacy [12].

Clinicians use evidence-based eHealth websites for medical education and when providing more information to patients [13,14]. Of 108 surveyed clinicians, 59% (n=64) had recommended a website to a patient [15]. However, the difficulty in determining the evidence base for eHealth websites and concerns over the quality of the content were identified as barriers to their use in clinical practice [8,13].

Incorrect and inaccurate information on eHealth websites can adversely affect people’s health [16,17]. In a systematic review by Hirsch et al [10] that included 54 eHealth websites providing information on endometriosis, over one-third did not cite authorship and almost half did not report references or sources of information. In a study that screened 25 eHealth websites providing information on dysmenorrhea (painful periods), a symptom commonly associated with endometriosis, only 28% included the name and credentials of the author [18]. The aim of this systematic review using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [19] was to evaluate the quality of endometriosis-related eHealth websites.

Methods

Search Strategy

Two Google searches were performed following PRISMA guidelines on July 27, 2020, to assess both international (.com) and Australian (.com.au) websites (Multimedia Appendix 1). The first search was conducted on Google.com using the search terms “endometriosis” OR “adenomyosis,” OR “pelvic pain.” The second search was conducted on Google.com.au and “Australia” was added to the search string (Figure 1). Google accounts for 92.26% of the global market share compared to Bing (Microsoft Corp; 2.83%), and Yahoo Search provides results generated by Bing [20,21]. Hence, we reported results based on Google search only. To minimize the impact of any previous search history, the search was conducted in “incognito” mode. The first 30 eHealth websites listed were screened as most people do not investigate beyond this number [22]. The search was later updated on August 24, 2023, to include more recent websites. Duplicate results were removed, and the eHealth websites were screened for eligibility.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flowchart.

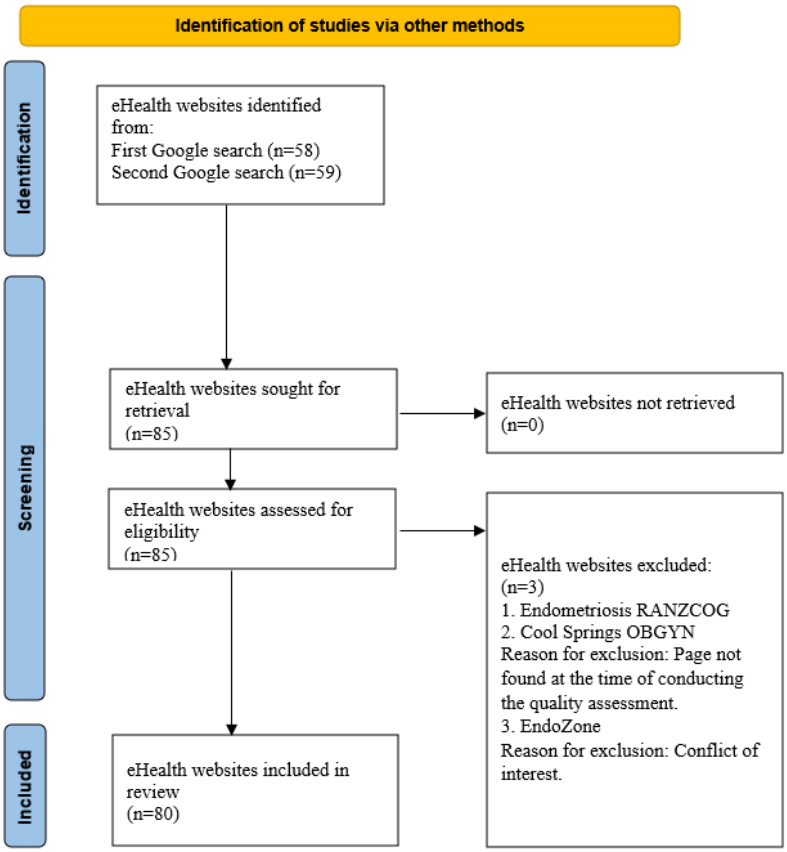

Inclusion and Exclusion Criteria

eHealth websites were included if they related to endometriosis or adenomyosis or pelvic pain, were written in English, and were free. eHealth websites that did not meet the inclusion criteria were excluded (Textbox 1).

Textbox 1. Inclusion and exclusion criteria for including eHealth websites on endometriosis.

Inclusion criteria

- eHealth websites that relate to endometriosis, adenomyosis, and pelvic pain in women
- Free eHealth websites (no associated cost)
- eHealth websites written in the English language

Exclusion criteria

- eHealth websites that did not relate to endometriosis, adenomyosis, and pelvic pain in women
- eHealth websites that require a payment or subscription to access them
- eHealth websites written in a language other than English

Data Extraction

Descriptive data were manually extracted by 1 researcher (DS) after reading the initial description and purpose of each eHealth website. These data were collated in an Excel spreadsheet (Microsoft Corp), under the following categories: (1) eHealth website name, (2) hyperlink, (3) developer, (4) funder, (5) intended purpose, (6) target audience, (7) category, (8) country of origin, and (9) last updated (Multimedia Appendix 2).

Quality Assessment

The ENLIGHT quality assessment tool [23] was used to evaluate all included eHealth websites. The ENLIGHT tool assesses seven criteria: (1) usability, (2) visual design, (3) user engagement, (4) content, (5) therapeutic persuasiveness, (6) therapeutic alliance, and (7) general subjective evaluation (Table 1). Each ENLIGHT quality assessment criterion is scored using a rating scale of 1-5 (very poor to very good and not applicable) [23].

XSL•FO
RenderX

Table 1. Description of the ENLIGHT quality assessment criteria, objectives, and factors assessed [23].

Quality assessment criteria	Objective	Factors assessed
Usability	Assesses the ease of learning how to use an eHealth website and the ease of using it appropriately	• Navigation • Learnability • Ease of use
Visual design	Assesses the look and feel of the eHealth website and the visual quality of the graphical user interface	• Aesthetics • Layout • Size
User engagement	Assesses the extent to which the eHealth website's design attracts users to use it.	• Content presentation • Interactive • Not irritating • Targeted or tailored or personalized reports • Captivating
Content	Assesses the content provided or learned while using the eHealth website	• Evidence-based content • Quality of information provided • Complete and concise • Clarity about the program's purpose
Therapeutic persuasiveness	Assesses the extent to which the eHealth website is designed to encourage users to make positive behavior changes or to maintain positive aspects of their life	• Call to action • Load reduction of activities • Therapeutic rationale and pathway • Rewards • Real data-driven or adaptive content • Ongoing feedback • Expectations and relevance
Therapeutic alliance	Assesses the ability of the eHealth website to create an alliance with the user in order to effect a beneficial change.	• Basic acceptance and support • Positive therapeutic expectations • Relatability
General subjective evaluation of the program's potential	Examines the eHealth website's general potential to benefit its target audience based on the rater's subjective evaluation	• Appropriate features to meet the clinical aim • Right mix of ability and motivation • I like the program

The eHealth websites were reviewed in 2 stages. Initially, 1 researcher (DS) reviewed all included eHealth websites. Then, the eHealth websites were divided and independently reviewed by another member of the team (RO, MLH, NB, HS, MAP, and CHMN). The ENLIGHT scores of each eHealth website were collated in an Excel spreadsheet.

Discrepancies in ratings (any deviation greater than 1 rating unit) were resolved by discussion between pairs of reviewers. If evaluation differences were not resolved, a third independent assessor was consulted. After a detailed assessment, the average of the 2 reviewers' (DS and RO or MLH or NB or HS or MAP or CHMN) ratings was used to calculate a score for each of the 7 domains (Multimedia Appendix 3).

High-Quality eHealth Websites

A total score for each eHealth website was calculated according to the ENLIGHT formula [24] (Multimedia Appendix 3). eHealth websites with a total score of ≥3.5 are classified as "good" according to the ENLIGHT scoring system [24] and are recommended as high-quality eHealth websites for information on endometriosis. Interrater reliability was described using an intraclass correlation coefficient, which is estimated from a 2-way mixed effects model using an absolute definition of agreement [25].

Ethical Considerations

An ethics approval was not required for this study because this was a systematic review of eHealth websites and did not involve the recruitment of participants.

Results

Overview

A total of 117 eHealth websites were returned in the search, 58 from the first Google search (International) and 59 from the second Google search (Australian). Thirty-two eHealth websites were duplicates (duplicate websites identified in the international and Australian search), leaving 85 that were screened for this systematic review (Figure 1). Two eHealth websites were excluded, as they were not related to the topic. There was a conflict of interest in assessing 1 website (EndoZone) [26] since the authors were responsible for its development. Two eHealth websites—Royal Australian and New Zealand College of Obstetricians and Gynaecologists [27] and Cool Springs OBGYN [28]—could not be included in the final review, as the

XSL•FO
RenderX

link was no longer available. A total of 80 eHealth websites were included in the final assessment and analysis (Multimedia Appendix 2).

There were discrepancies between the first and second reviewers (DS and RO or MLH or NB, HS, MAP or CHMN) in 7% of the ratings (316 variances across 4480 total ratings). All discrepancies were resolved without needing a third reviewer. The intraclass correlation coefficient was 0.61 (95% CI 0.45-0.73), indicating that interrater reliability was moderate [25].

Characteristics of eHealth Websites

Of the 80 eHealth websites, 44 (55%) belonged to Australian organizations, while 36 (45%) belonged to international organizations. The majority of the eHealth websites (n=49, 61%) provided education to the community on natal female pain (eg, period pain or conditions that cause pelvic pain in women), 25 (31%) were business pages of private organizations, 5 (6%) were related to endometriosis research, and 1 (1%) was a media release of a study at a university. Of the 49 eHealth websites that provided community education, 12 also provided support features to the endometriosis community. These included resources (eg, booklets, webinars), links to support group networks, and social media platforms for digital engagement with the endometriosis community (Multimedia Appendix 2). Of the 80 eHealth websites included in this study, none required payment or a subscription fee to access content.

Target Users

The majority of the eHealth websites (n=70, 88%) were designed for use by the endometriosis community. Five (6%) eHealth websites provided information or education to health care providers and 5 (6%) provided information for researchers.

Quality of eHealth Websites

Overview

The eHealth websites were evaluated using the ENLIGHT quality assessment criteria [23]. Table 2 presents the top 4 eHealth websites (ie, those with a total score of ≥3.5) according to the ENLIGHT scoring formula [24] and are recommended as high-quality eHealth websites for information on endometriosis.

Table 2. Top 4 eHealth websites according to the ENLIGHT scoring system.

eHealth website	Usability score	Visual design score	User engagement score	Content score	Therapeutic persuasiveness score	Therapeutic alliance score	General subjective evaluation score	Total score
Endometriosis Australia Facebook Page [29]	4.50	4.67	4.40	4.00	4.21	4.50	4.00	4.24
Understanding endometriosis—Endometriosis UK [30]	5.00	4.00	4.30	4.38	3.57	4.33	4.33	3.92
Adenomyosis: The poor cousin of endometriosis—The Medical Republic[a] [31]	4.67	4.50	4.00	5.00	3.07	3.50	3.83	3.63
National Action Plan for Endometriosis—EndoActive [32]	4.50	4.17	2.00	5.00	3.36	4.67	4.50	3.53

[a]High-quality eHealth website for clinicians.

Usability

The majority of eHealth websites (n=76, 95%) scored well (≥3.5) for usability (Multimedia Appendix 3). These eHealth websites were characterized by smooth, nearly frictionless navigation. They had an intuitive interface that was easy to learn and straightforward to use. Examples of websites that scored well for usability included the Endometriosis Australia Facebook Page [29], Endometriosis UK [30], and the Endometriosis page on the Jean Hailes for Women's Health website [33].

Of the 76 eHealth websites that scored well for usability, 42 (55%) provided education on natal female pain to the endometriosis community, 23 (30%) were business pages of private organizations that encouraged users to book appointments, 5 (7%) eHealth websites provided education to health care providers, 5 (7%) were eHealth websites related to endometriosis research, and 1 (1%) was a media release article.

Visual Design

In total, 64 (80%) eHealth websites scored ≥3.5 on visual design (Multimedia Appendix 3). These eHealth websites were assessed as having an attractive visual design, an appealing color scheme, were well structured with a consistent layout, and the content was easy to read. They also displayed appropriately sized fonts, buttons, and menus. Some examples include Pelvic Pain-Pain Australia [34], Endometriosis-Healthline [35], and Endometriosis Practice Essentials-Medscape [36].

Of the 64 eHealth websites that scored well on visual design, 35 (54%) provided education on natal female pain to the endometriosis community, 22 (34%) were business pages of organizations, 4 (6%) provided education to health care providers on the management of conditions that cause natal female pain, and 2 (3%) were related to endometriosis research.

XSL•FO
RenderX

User Engagement

A low number of eHealth websites (n=18, 23%) scored highly (≥3.5) on user engagement (Multimedia Appendix 3). Websites that scored highly were characterized by a good mix of text, images, and videos. The content was presented interactively and engagingly. User engagement was further enhanced by avoiding features like pop-up ads, notifications, alerts, and sounds. Some examples include Adenomyosis—Sydney Morning Herald [37], Adenomyosis—The Centre for Innovative Gyn Care [38], and Adenomyosis—The Medical Republic [31]. Of the 18 eHealth websites that scored well on user engagement, 11 (61%) provided education on natal female pain to the community, 5 (28%) were business pages of private organizations, and 2 (11%) provided education to health care providers.

Content

A total of 63 (79%) eHealth websites scored well (≥3.5) on the content domain (Multimedia Appendix 3). These eHealth websites contained appropriate, complete, and concise information with clarity about the eHealth website's purpose. However, of 63 websites, only 40% (n=25) of the eHealth websites provided references or mentioned the sources of information, and only 38% (n=26) provided the name of the author. Examples of websites that scored highly for content included Endometriosis—Jean Hailes for Women's Health [33], NewsGP—RACGP [39], and Endometriosis—Better Health Channel [40].

Thirty-five (56%) of these eHealth websites provided education on natal female pain to the endometriosis community, 17 (27%) were business pages of private organizations, 5 (8%) provided education to health care providers, 5 (8%) were related to endometriosis research, and 1 (2%) was a media release article on endometriosis by an Australian University.

Therapeutic Persuasiveness

Only 2 (2%) eHealth websites (ie, Endometriosis Australia Facebook Page [29] and Endometriosis UK [30]) scored highly (≥3.5) on the therapeutic persuasiveness domain (Multimedia Appendix 3). One page [29] is designed for the endometriosis community on Facebook, while the other page [30] is the official website of a United Kingdom–based endometriosis charitable organization. Both websites provide resources to raise awareness and educate people about endometriosis. These pages provide opportunities for a call to action, which can be described as activities that prompt the user to take action (eg, goal setting). Both websites facilitate interactions between the digital endometriosis community, release information about upcoming events, and enable engagement with page content such as watching informative videos [29] or engaging in a web chat [30].

Therapeutic Alliance

Only 28% (n=22) of the eHealth websites scored highly (≥3.5) on the therapeutic alliance domain (Multimedia Appendix 3). Seventeen (77%) eHealth websites provide education on natal female pain to the endometriosis community, 3 (14%) were business pages of private organizations, 1 (5%) provided education to health care providers, and 1 (5%) was related to endometriosis research. These eHealth websites incorporated features that sought to foster a therapeutic alliance with the user. Examples of support include personal stories of people affected by endometriosis, which creates a sense of a shared digital endometriosis community, support group information, and helpline numbers. Examples include Endometriosis UK [30], Endometriosis—Jean Hailes for Women's Health [33], Endometriosis Australia [41] and Pelvic Pain Foundation of Australia [42].

General Subjective Evaluation of eHealth Websites

This criterion evaluates the eHealth website's potential to benefit users based on reviewers' subjective scores. Only 38% (n=30) of eHealth websites scored highly (≥3.5) under this criterion (Multimedia Appendix 3). Nineteen (63%) eHealth websites provided education on natal female pain to the endometriosis community, 4 (13%) provided education to health care providers, 6 (20%) were business pages of private organizations, and 1 (3%) was a journal article.

Discussion

Summary

We conducted a comprehensive, multidimensional quality assessment of endometriosis eHealth websites using the ENLIGHT tool that captures quality constructs like persuasive design and therapeutic alliance, which are considered central to the successful uptake of eHealth websites among end users [22]. Our systematic review identified 4 high-quality endometriosis eHealth websites that can be used as educational resources for the community and health care providers. This is the first systematic review to use the ENLIGHT tool and comprehensively assess endometriosis eHealth websites.

Principal Results

Quality of Endometriosis eHealth Websites

Overview

The proliferation of incorrect digital health information is a major concern [43]. A quality assessment of eHealth websites in the United States revealed that only 58% (n=58) met the criteria for accuracy and credibility of content [44]. In a systematic review by Hirsch et al [10] that included 54 eHealth websites providing information on endometriosis, over a third did not cite authorship and almost a half did not report references or sources of information.

Recommendation 1

The need for the development of accurate and evidence-based endometriosis eHealth websites for the community. We found similar results in our assessment indicating the need for eHealth websites to integrate information such as referencing and authorship to provide credibility. Furthermore, we found that eHealth websites that provide education to health care providers scored better than those that provided education to the community. Therefore, to reduce the proliferation of incorrect information, improved referencing and authorship on community-targeted eHealth websites will improve credibility.

XSL•FO
RenderX

Furthermore, some evidence in the literature states that the quality of endometriosis-related information on the internet centers around content that can be inaccurate and misleading [10]; however, there is little evidence to describe what the quality of information means to the endometriosis community. To some, quality centers around improving self-awareness about endometriosis to help make informed decisions [45]. While credible evidence-based content is a significant part of quality, factors that facilitate user engagement, therapeutic persuasiveness, and therapeutic alliance are also worthy of quality assessment for information obtained over the internet since these factors ensure successful and long-term uptake of eHealth websites [46].

Need to Enhance User Engagement, Therapeutic Persuasiveness, and Therapeutic Alliance

Overview

Most of the eHealth websites scored well (≥3.5) on usability (n=76, 95%) and visual design (n=64, 80%). Usability and visual design influence the user's first impression and subsequent uptake and use [47,48]. However, we found a low percentage of eHealth websites scored well on user engagement (n=18, 23%). Current endometriosis eHealth websites are primarily informative and lack user interaction. A Cochrane review found that health platforms with interactive features have positive effects on users (improved knowledge, self-efficacy, behavior, and clinical outcomes) as compared to nonusers [49].

Recommendation 2

User engagement can be improved by providing interactive features that enable users to input and receive a reaction by providing personalized feedback. For example, the web chat feature on the Endometriosis UK website [30] allows the user to make an enquiry and receive feedback. User engagement is also strengthened by sharing lived experiences and stories and cocreating content that is meaningful to users, such as clinicians sharing clinical insights.

Therapeutic persuasiveness and therapeutic alliance could further enhance user engagement [22,46]. However, we found few eHealth websites that scored well on therapeutic persuasiveness (n=2, 3%) and therapeutic alliance (n=22, 28%). Therapeutic persuasiveness is positively correlated with real-world usage of eHealth websites, while therapeutic alliance enhances positive user engagement by fostering relatability [46].

Recommendation 3

Since most endometriosis eHealth websites are informative only, therapeutic persuasiveness can be increased by adding a "call to action." This means the eHealth website could suggest when to see a general practitioner, what to discuss at the medical appointment, or provide evidence-based self-management strategies to cope with endometriosis. Additionally, eHealth websites can incorporate conversational agents or chatbots such as Alexa (Google), Siri (Apple), S Voice or Bixby (Samsung), and Cortana (Microsoft Corp) [46,50], which mimic human conversations [46,50] to foster relatability. Therapeutic persuasiveness and therapeutic alliance can be further strengthened by sharing lived experiences through digital community engagement and cocreating meaningful content.

Evidence-Based Insights on Optimization of eHealth Websites

Overview

Our Google search for "endometriosis" identified a mix of eHealth websites with eclectic primary purposes including education, marketing for business, and dissemination of academic research. During internet searches, people are most likely to click on the first 5 websites that come up on the Google search result pages [51]. To attract traffic, an eHealth website should appear in the first 5 rankings [51].

Recommendation 4

Organizations could benefit from investing in search engine optimization (SEO) tools. SEO improves the ranking of a website in Google search results. There are various tools to achieve this including the creation of fresh, unique, and qualitative content [52] and improving keyword density, which is the amount of time the keywords that the users are searching for appear on the eHealth website. Including keywords in the headers, main titles, and content of the eHealth website also improves search ranking [51,52]. The use of meta descriptions, a short description (160 characters) of the website's content that appears below the page title on the search result page and is managed by the website owner, can also improve ranking. The meta descriptions should include words that the target audience are likely to search for [51]. The use of permalinks (permanent and specific URL links) improves SEO. Proper link architecture helps the search engine discover the eHealth website [51,52]. Adding backlinks (other websites linking back to the main eHealth website) increases discoverability and is an important factor in improving rankings [51,52]. Social media promotions on popular platforms such as Facebook, Instagram, and Twitter help to popularize the content of the eHealth website, thereby improving a website's ranking [51,52].

Other tools that can improve an eHealth website's ranking include submitting the sitemap (ie, a list of pages, videos, and other content on the website and the relationship between them) to Google. This helps Google to download and index information. Google analyses this information to produce search results [53].

Endometriosis eHealth Websites as Resources for Health Care Providers

Overview

Our systematic review found only 1 high-quality eHealth website for health care providers Adenomyosis—The Medical Republic [31], compared to 3 for the endometriosis community (Table 2), indicating a lack of high-quality eHealth websites for health care providers. Evidence suggests that health care providers benefit from using eHealth websites in daily practice [54]. However, barriers such as difficulty accessing full-text documents, subscription fees, concerns about quality, and limited relevance of the information in day-to-day clinical practice may limit use [13]. Coupled with an identified need to improve endometriosis education among health care providers

XSL•FO
RenderX

[4], it is necessary to develop digital learning tools to address this gap.

Recommendation 5

There is a need to develop easily accessible, evidence-based endometriosis eHealth websites for health care providers that provide a valuable and easy reference in daily practice and enhance professional development for endometriosis management.

Strengths

Our systematic review has several strengths. Novel findings are presented on high-quality eHealth websites assessed using the ENLIGHT quality assessment tool and guide the community and health care providers toward quality, credible, and supportive health information. Health care providers can use these eHealth websites as educational resources and recommend high-quality eHealth websites to their patients. Finally, this review presents recommendations (ie, good design features and use of SEO tools) when designing or updating endometriosis eHealth websites as evidence suggests that good design features can help improve a website's ranking, reach, and discoverability [52].

Limitations

Our study has the following limitations. The ENLIGHT quality assessment criteria [22] were challenging to apply to websites. Most eHealth websites did not provide a health or behavior-related intervention for natal female pain. Hence, it was difficult to evaluate therapeutic persuasiveness and therapeutic alliance criteria in their entirety. We did not evaluate the entire eHealth website. We only evaluated the landing page or article the Google search engine result produced to mimic real-world circumstances. However, in some cases, this led to an uploaded document (the National Action Plan for Endometriosis on the EndoActive website) rather than the actual website, so it was not truly assessing an eHealth website. Google accounts for 92.26% of the global market share as compared to Microsoft Bing (2.83%) [20]. Additionally, Yahoo Search provides results generated by Microsoft Bing [21]. Hence, we reported results based on Google search only. Although, we searched in "incognito mode," the Google algorithm may have automatically incorporated our location when searching, which may have influenced the results. Furthermore, the digital world is changing rapidly. The time and geographical location of the search may influence results conducted today versus the results presented above. Due to the lack of translation services, we did not include eHealth websites in languages other than English. We did not assess if the eHealth websites included in this study were developed using a genuine cocreation process hence, we do not know whether they are representative of all ethnicities or races. Finally, while we recommend the top 4 high-quality eHealth websites based on the ENLIGHT scoring system as good sources of information on endometriosis, we believe that the word "good" may well be influenced in context, that is, the generalizability of interpretation of good may vary for users in low income versus high-income countries. Nevertheless, this study has shown how eHealth websites can be assessed using the ENLIGHT checklist, which is a validated tool with its 7 quality assessment criteria.

Conclusions

Our systematic review presents novel findings on the quality assessment of eHealth websites using the ENLIGHT checklist to obtain endometriosis-related information. The findings of our study are (1) suggestive of high-quality eHealth websites for community use and (2) can be used by health care providers for educational purposes and recommendations. We recommend the development of (1) accurate and evidence-based endometriosis eHealth websites for the community; (2) accessible endometriosis eHealth websites for health care providers supporting daily practice and professional development; (3) interactive eHealth websites that promote user engagement, therapeutic persuasiveness, and therapeutic alliance; and (4) leveraging SEO tools to improve Google search ranking.

Acknowledgments

We thank Ms Vikki Langton (Librarian—Adelaide Medical School, University of Adelaide) for rendering assistance in formulating the search strategy. Diksha Sirohi is a PhD (Medicine) candidate and is supported by the Australian Government, Research Training Program Stipend.

Data Availability

All data generated or analyzed during this study are included in this published article and Multimedia Appendices 1-3.

Authors' Contributions

DS searched eHealth websites, data extraction, quality assessment, data collation, data analysis, and drafting of the manuscript. RO searched eHealth websites, contributed to quality assessment, and drafted the manuscript. CHMN contributed to the quality assessment and drafting of the manuscript. NB, HS, MAP, and MLH contributed to the quality assessment and review of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

Although this systematic review has not received separate funding, it is part of the EndoZone, a digital platform for endometriosis project funded by the Australian Government Department of Health and Aged Care and Jean Hailes for Women's Health. CHMN

XSL•FO
RenderX

was employed by Jean Hailes for Women's Health, which precluded her from conducting the quality assessment of eHealth websites from Jean Hailes for Women's Health. CHMN received grant funding from the Australian Government Department of Health and Aged Care, under the Medical Research Future Fund and was a previous employee of CSL Vifor (formerly Vifor Pharma Pty Ltd). NB has shares in goAct Pty Ltd and received research funding support from Digital Health CRC, Medical Research Future Fund and Barossa Fleurieu Local Health Network.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist and PRISMA 2020 abstracts checklist.
[DOCX File , 23 KB - jmir_v26i1e48243_app1.docx]

Multimedia Appendix 2
Details of eHealth websites included in this study (listed alphabetically).
[DOCX File , 49 KB - jmir_v26i1e48243_app2.docx]

Multimedia Appendix 3
ENLIGHT quality assessment total score.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e48243_app3.xlsx]

References

1. Taylor HS, Kotlyar AM, Flores VA. Endometriosis is a chronic systemic disease: clinical challenges and novel innovations. Lancet 2021;397(10276):839-852 [FREE Full text] [doi: 10.1016/S0140-6736(21)00389-5] [Medline: 33640070]
2. Arruda MS, Petta CA, Abrão MS, Benetti-Pinto CL. Time elapsed from onset of symptoms to diagnosis of endometriosis in a cohort study of Brazilian women. Hum Reprod 2003 Apr;18(4):756-759 [FREE Full text] [doi: 10.1093/humrep/deg136] [Medline: 12660267]
3. Nnoaham KE, Hummelshoj L, Webster P, d'Hooghe T, de Cicco Nardone F, de Cicco Nardone C, World Endometriosis Research Foundation Global Study of Women's Health consortium. Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries. Fertil Steril 2011 Aug;96(2):366-373.e8 [FREE Full text] [doi: 10.1016/j.fertnstert.2011.05.090] [Medline: 21718982]
4. van der Zanden M, Teunissen DAM, van der Woord IW, Braat DDM, Nelen WLDM, Nap AW. Barriers and facilitators to the timely diagnosis of endometriosis in primary care in the Netherlands. Fam Pract 2020;37(1):131-136 [FREE Full text] [doi: 10.1093/fampra/cmz041] [Medline: 31414120]
5. Culley L, Law C, Hudson N, Denny E, Mitchell H, Baumgarten M, et al. The social and psychological impact of endometriosis on women's lives: a critical narrative review. Hum Reprod Update 2013;19(6):625-639 [FREE Full text] [doi: 10.1093/humupd/dmt027] [Medline: 23884896]
6. Hertling S, Hertling D, Schleußner E, Loos F, Graul I. E-health–the importance of the internet as an informative digital health application for gynecological patients in times of SARs-CoV2: a national cross-sectional survey. Inform Med Unlocked 2022;30:100942 [FREE Full text] [doi: 10.1016/j.imu.2022.100942]
7. Kwan G, Shaw JA, Murnane L. Internet usage within healthcare: how college students use the internet to obtain health information. J Consumer Health Internet 2019 Dec 02;23(4):366-377 [FREE Full text] [doi: 10.1080/15398285.2019.1681247]
8. Sbaffi L, King K. Living with endometriosis: the role of the internet in supporting the diagnosis and treatment process. J Consumer Health Internet 2020 Dec 07;24(4):370-390 [FREE Full text] [doi: 10.1080/15398285.2020.1816803]
9. Welch A. Google's top trending health questions of 2018. CBS News. 2018 Dec 29. URL: https://www.cbsnews.com/news/top-10-google-trending-health-questions-of-2018/ [accessed 2022-06-10]
10. Hirsch M, Aggarwal S, Barker C, Davis CJ, Duffy JMN. Googling endometriosis: a systematic review of information available on the Internet. Am J Obstet Gynecol 2017 May;216(5):451-458.e1 [FREE Full text] [doi: 10.1016/j.ajog.2016.11.1007] [Medline: 27840143]
11. Hu H, Mcintyre A. Using Google Trends to assess the Australian public's interest in topical gynaecology issues. Aust N Z J Obstet Gynaecol 2021 Dec;61(6):978-981 [FREE Full text] [doi: 10.1111/ajo.13435] [Medline: 34570377]
12. Armour M, Parry K, Curry C, Ferfolja T, Parker MA, Farooqi T, et al. Evaluation of a web-based resource to improve menstrual health literacy and self-management in young women. J Psychosom Res 2022 Nov;162:111038 [FREE Full text] [doi: 10.1016/j.jpsychores.2022.111038] [Medline: 36179421]
13. Bernard E, Arnould M, Saint-Lary O, Duhot D, Hebbrecht G. Internet use for information seeking in clinical practice: a cross-sectional survey among French general practitioners. Int J Med Inform 2012 Jul;81(7):493-499 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.02.001] [Medline: 22425281]
14. Romano M, Gesualdo F, Pandolfi E, Tozzi AE, Ugazio AG. Use of the internet by Italian pediatricians: habits, impact on clinical practice and expectations. BMC Med Inform Decis Mak 2012 Mar 28;12(1):23 [FREE Full text] [doi: 10.1186/1472-6947-12-23] [Medline: 22455671]

XSL•FO
RenderX

15. Usher WT. Gold Coast general practitioners' recommendations of health websites to their patients. Med J Aust 2007 Jul 16;187(2):82-83 [FREE Full text] [doi: 10.5694/j.1326-5377.2007.tb01145.x] [Medline: 17635088]
16. Armstrong-Heimsoth A, Johnson ML, McCulley A, Basinger M, Maki K, Davison D. Good Googling: a consumer health literacy program empowering parents to find quality health information online. J Consumer Health Internet 2017 Jun 12;21(2):111-124 [FREE Full text] [doi: 10.1080/15398285.2017.1308191]
17. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107 [FREE Full text] [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
18. Lovett J, Gordon C, Patton S, Chen CX. Online information on dysmenorrhoea: an evaluation of readability, credibility, quality and usability. J Clin Nurs 2019 Oct;28(19-20):3590-3598 [FREE Full text] [doi: 10.1111/jocn.14954] [Medline: 31162870]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906 [FREE Full text] [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
20. Leung C, Chan W. A study on key elements for successful and effective search engine optimization. Int J Technol Knowl Society 2021;17(2):23-39. [doi: 10.18848/1832-3669/cgp/v17i02/23-39]
21. Chris A. Top 10 search engines in the world (2023 Update). Reliable Soft Digital Marketing Agency. 2021. URL: https://www.reliablesoft.net/top-10-search-engines-in-the-world [accessed 2023-08-24]
22. Baumel A, Faber K, Mathur N, Kane JM, Muench F. ENLIGHT: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based eHealth interventions. J Med Internet Res 2017 Mar 21;19(3):e82 [FREE Full text] [doi: 10.2196/jmir.7270] [Medline: 28325712]
23. Baumel A, Kane JM, Muench F. Enlight summary report of [eHealth intervention product name]. Mindtools.io. 2016. URL: https://mindtools.io/wp-content/uploads/2018/06/Enlight-Complete.pdf [accessed 2020-07-27]
24. Scientific approach. Mindtools.io. 2023. URL: https://mindtools.io/scientific-approach/ [accessed 2023-11-07]
25. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
26. A digital health platform for endometriosis. EndoZone. Adelaide: University of Adelaide; 2022 Mar 31. URL: https://www.endozone.com.au [accessed 2022-03-31]
27. Endometriosis. Royal Australian and New Zealand College of Obstetricians and Gynaecologists. URL: https://ranzcog.edu.au/womens-health/patient-information-resources/endometriosis [accessed 2020-07-27]
28. Adenomyosis. Cool Springs Obstetrics & Gynecology. URL: http://tinyurl.com/mkeahnp3 [accessed 2023-08-24]
29. Endometriosis Australia. URL: https://www.facebook.com/EndometriosisAustralia/ [accessed 2020-07-27]
30. What is endometriosis. Endometriosis UK. URL: https://www.endometriosis-uk.org/what-endometriosis [accessed 2020-07-27]
31. Liang E, Brown B. Adenomyosis: the poor cousin of endometriosis. The Medical Republic. 2020 Feb 17. URL: http://tinyurl.com/39wsnmk4 [accessed 2020-07-27]
32. National action plan for endometriosis. EndoActive. 2018 Jul. URL: http://tinyurl.com/29c589pd [accessed 2020-07-27]
33. Endometriosis. Jean Hailes for Women's Health. URL: https://www.jeanhailes.org.au/health-a-z/endometriosis [accessed 2020-07-27]
34. Pelvic Pain. Pain Australia. URL: https://www.painaustralia.org.au/about-pain/forms-of-pain-2021/pelvic-pain-2021 [accessed 2020-07-27]
35. Amstrong M. Endometriosis. Healthline. 2019 Mar 28. URL: https://www.healthline.com/health/endometriosis [accessed 2020-07-27]
36. Davila G. Endometriosis Practice Essentials. Medscape. 2023. URL: https://emedicine.medscape.com/article/271899-overview [accessed 2020-07-27]
37. Power J. A new solution to women's severe pain that isn't a hysterectomy. Sydney Morning Herald. 2018 Oct 02. URL: http://tinyurl.com/mwje747e [accessed 2020-07-27]
38. Adenomyosis. The Centre for Innovative Gyn Care. URL: https://innovativegyn.com/conditions/adenomyosis [accessed 2023-07-27]
39. Liotta M. One in nine Australian women aged 40–44 has endometriosis. NewsGP by The Royal Australian College of General Practitioners (RACGP). 2019 Aug 29. URL: http://tinyurl.com/4atv4m92 [accessed 2020-07-27]
40. Endometriosis. Better Health Channel. URL: http://tinyurl.com/mpt9udkj [accessed 2020-07-27]
41. Some facts about endo. Endometriosis Australia. URL: https://endometriosisaustralia.org/ [accessed 2023-08-24]
42. How we help. The Pelvic Pain Foundation. URL: https://www.pelvicpain.org.au/ [accessed 2020-07-27]
43. Portillo IA, Johnson CV, Johnson SY. Quality evaluation of consumer health information websites found on Google using DISCERN, CRAAP, and HONcode. Med Ref Serv Q 2021;40(4):396-407 [FREE Full text] [doi: 10.1080/02763869.2021.1987799] [Medline: 34752199]
44. Devine T, Broderick J, Harris LM, Wu H, Hilfiker SW. Making quality health websites a national public health priority: toward quality standards. J Med Internet Res 2016;18(8):e211 [FREE Full text] [doi: 10.2196/jmir.5999] [Medline: 27485512]

XSL•FO
RenderX

45. Whelan E. 'No one agrees except for those of us who have it': endometriosis patients as an epistemological community. Sociol Health Illn 2007 Nov;29(7):957-982 [FREE Full text] [doi: 10.1111/j.1467-9566.2007.01024.x] [Medline: 18092978]
46. Baumel A, Kane JM. Examining predictors of real-world user engagement with self-guided eHealth interventions: analysis of mobile apps and websites using a novel dataset. J Med Internet Res 2018 Dec 14;20(12):e11491 [FREE Full text] [doi: 10.2196/11491] [Medline: 30552077]
47. Lazard AJ, King AJ. Objective design to subjective evaluations: connecting visual complexity to aesthetic and usability assessments of eHealth. Int J Human–Comput Interact 2019 Apr 24;36(1):95-104 [FREE Full text] [doi: 10.1080/10447318.2019.1606976]
48. Lee K, Hoti K, Hughes JD, Emmerton LM. Consumer use of "Dr Google": a survey on health information-seeking behaviors and navigational needs. J Med Internet Res 2015 Dec 29;17(12):e288 [FREE Full text] [doi: 10.2196/jmir.4345] [Medline: 26715363]
49. Murray E, Burns J, See TS, Lai R, Nazareth I. Interactive health communication applications for people with chronic disease. Cochrane Database Syst Rev 2005 Oct 19(4):CD004274 [FREE Full text] [doi: 10.1002/14651858.CD004274.pub4] [Medline: 16235356]
50. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
51. Stephan G. Using website analytics in search engine optimization for the domain of LIS links, in India. Library Philosophy and Practice (e-journal). 2020. URL: http://tinyurl.com/2kvz96ff [accessed 2022-06-10]
52. Mavridis T, Symeonidis AL. Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms. Eng Appl Artifi Intelli 2015 May;41:75-91 [FREE Full text] [doi: 10.1016/j.engappai.2015.02.002]
53. In-depth guide to how Google search works. Google Search Central. URL: http://tinyurl.com/598h8fyn [accessed 2022-06-10]
54. Blumenthal D, Glaser JP. Information technology comes to medicine. N Engl J Med 2007 Jun 14;356(24):2527-2534. [doi: 10.1056/NEJMhpr066212] [Medline: 17568035]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RACGP: Royal Australian College of General Physicians
SEO: search engine optimization

Edited by A Mavragani; submitted 20.04.23; peer-reviewed by D Carvalho, K Malhotra, G Delanerolle; comments to author 20.08.23; revised version received 20.11.23; accepted 27.11.23; published 10.01.24.

Please cite as:
Sirohi D, Ng CHM, Bidargaddi N, Slater H, Parker MA, Hull ML, O'Hara R
High-Quality eHealth Websites for Information on Endometriosis: Systematic Search
J Med Internet Res 2024;26:e48243
URL: https://www.jmir.org/2024/1/e48243
doi:10.2196/48243
PMID:38198205

©Diksha Sirohi, Cecilia Hoi Man Ng, Niranjan Bidargaddi, Helen Slater, Melissa A Parker, Mary Louise Hull, Rebecca O'Hara. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning Approaches for the Image-Based Identification of Surgical Wound Infections: Scoping Review

Juan Pablo Tabja Bortesi[1], HBSc; Jonathan Ranisau[1], MASc; Shuang Di[1,2], MSc; Michael McGillion[3], PhD; Laura Rosella[2], PhD; Alistair Johnson[4], DPhil; PJ Devereaux[3], PhD; Jeremy Petch[1,3,5,6], BA, MA, PhD

[1]Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[3]Population Health Research Institute, Hamilton, ON, Canada
[4]SickKids Research Institute, Toronto, ON, Canada
[5]Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
[6]Division of Cardiology, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Jeremy Petch, BA, MA, PhD
Centre for Data Science and Digital Health
Hamilton Health Sciences
175 Longwood Road South
Suite 207
Hamilton, ON, L8P 0A1
Canada
Phone: 1 9055212100
Email: petchj@hhsc.ca

Abstract

Background: Surgical site infections (SSIs) occur frequently and impact patients and health care systems. Remote surveillance of surgical wounds is currently limited by the need for manual assessment by clinicians. Machine learning (ML)–based methods have recently been used to address various aspects of the postoperative wound healing process and may be used to improve the scalability and cost-effectiveness of remote surgical wound assessment.

Objective: The objective of this review was to provide an overview of the ML methods that have been used to identify surgical wound infections from images.

Methods: We conducted a scoping review of ML approaches for visual detection of SSIs following the JBI (Joanna Briggs Institute) methodology. Reports of participants in any postoperative context focusing on identification of surgical wound infections were included. Studies that did not address SSI identification, surgical wounds, or did not use image or video data were excluded. We searched MEDLINE, Embase, CINAHL, CENTRAL, Web of Science Core Collection, IEEE Xplore, Compendex, and arXiv for relevant studies in November 2022. The records retrieved were double screened for eligibility. A data extraction tool was used to chart the relevant data, which was described narratively and presented using tables. Employment of TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines was evaluated and PROBAST (Prediction Model Risk of Bias Assessment Tool) was used to assess risk of bias (RoB).

Results: In total, 10 of the 715 unique records screened met the eligibility criteria. In these studies, the clinical contexts and surgical procedures were diverse. All papers developed diagnostic models, though none performed external validation. Both traditional ML and deep learning methods were used to identify SSIs from mostly color images, and the volume of images used ranged from under 50 to thousands. Further, 10 TRIPOD items were reported in at least 4 studies, though 15 items were reported in fewer than 4 studies. PROBAST assessment led to 9 studies being identified as having an overall high RoB, with 1 study having overall unclear RoB.

Conclusions: Research on the image-based identification of surgical wound infections using ML remains novel, and there is a need for standardized reporting. Limitations related to variability in image capture, model building, and data sources should be addressed in the future.

(J Med Internet Res 2024;26:e52880) doi:10.2196/52880

XSL•FO
RenderX

KEYWORDS

surgical site infection; machine learning; postoperative surveillance; wound imaging; mobile phone

Introduction

Postoperative complications are associated with significant morbidity and mortality [1,2]. Wound-related issues following surgery remain common and represent a considerable cost to patients and health care systems [3,4]. The global incidence of surgical site infections (SSIs)—which include superficial or deep infections occurring at the incision site as well as organ-space infections related to the surgery [5]—has been estimated to be 11% [6]. Many of these events occur after hospital discharge, highlighting the need for remote posthospital discharge monitoring. Early research suggests that remote postoperative wound follow-up is associated with high patient satisfaction and reduced costs [7,8].

Artificial intelligence tools have been applied to various aspects of health care and are contributing to the shift toward precision medicine [9-11]. Specifically, machine learning (ML) techniques can leverage health data and develop predictive models to assist in clinical decision-making [12], and can be used in conjunction with computer vision. An important medical task is the classification and detection of various objects, ranging from skin lesions to cell nuclei [13]. Recently, ML-enabled computer vision methods have been used to contribute to the automation of wound segmentation [14,15], evaluation of postoperative outcomes [16,17], and improvement of wound assessment practices [18,19], often outperforming existing approaches.

Wound care involves cleaning and dressing, monitoring healing, addressing possible infection, and other wound type-specific measures [20]. Current image-based wound management practices, often involving manual wound photography and assessment carried out by nurses, are time- and labor-intensive [21]. In contrast, models of care augmented with ML-enabled methods can be automated [22,23]. The portability of these methods might also be employed to conduct such assessments remotely [24], reducing patient travel burden and improving access to wound care in rural areas [25,26]. A recent clinical trial (Post-Discharge After Surgery Virtual Care With Remote Automated Monitoring-1) found that virtual care with remote monitoring that included wound evaluation shows promise in improving outcomes important to patients and to optimal health system function [27]. These results highlight the utility of digital approaches to care, which can be integrated with automated ML systems to increase scalability.

The research landscape of ML-based methods for wound surveillance is evolving rapidly. Several reviews have addressed the use of ML for various aspects of wound care from different perspectives. One scoping review focused on mapping the use cases for ML in the management of various types of chronic wounds (eg, visual assessment and predicting evolution) [28]. Another review addressed image-based chronic wound assessment from a technical standpoint, characterizing existing rule-based and ML methods for wound feature extraction and classification, as well as systems for wound imaging [29]. However, chronic and acute wounds differ in terms of the clinical signs associated with infection as those in chronic wound infections are often less discernible [30], and there is a need to establish the state of the science with respect to how ML-based tools are being used for postoperative wounds. One systematic review specifically characterized the effectiveness of ML algorithms that use textual or structured data for the detection and prediction of SSIs [31], though a survey of image-based methods has not been undertaken. Likewise, other systematic reviews have found that reporting in ML-based prediction model studies is generally poor and that most are at high risk of bias (RoB) [32,33]. Considering these results, assessments of RoB and the employment of reporting guidelines—which have not been included in previous reviews of image-based ML for wound care—can further provide insights into the current state of research in this field.

The scope and purpose of this review was to provide an in-depth overview of ML approaches that use visual data for the identification of SSIs. Specifically, this review describes the nature of the methods used in this context, the ways in which they have been validated, the extent to which the reporting of these studies follows guideline recommendations, and their RoB.

Methods

Review Methodology

This scoping review was conducted in accordance with the appropriate JBI (Joanna Briggs Institute) methodology [34]. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist was used to guide the writing of this review [35]. We opted for a scoping review approach as we sought to analyze the methods employed in conducting research in this field, an indication for scoping reviews [36], rather than synthesize model performance.

Search Strategy and Study Selection

Following our protocol [37], participants of any age (or other demographic variable) who underwent any type of surgery were considered. The main concept being addressed is the use of ML-based computer vision in the image-based identification of surgical wound infections. Only wounds that were directly the product of surgery were included. Other types of wounds, such as pressure ulcers, were excluded. We included studies that described detection of infection of such wounds (as defined by study authors). Studies solely focusing on tasks other than identification (eg, segmentation) and using sources other than images or videos for prediction were not considered. Studies conducted in any postoperative context, including postdischarge settings, were included.

Studies that developed or validated one or more prediction models were included in this review, including those that gathered data from experimental, quasi-experimental, and observational studies (eg, randomized controlled trials, and prospective and retrospective studies). Only primary sources

XSL•FO
RenderX

were considered. Select grey literature sources, such as conference proceedings and preprints, were also considered. Animal studies were excluded.

An initial limited search of MEDLINE (Ovid) and CINAHL (EBSCO) was undertaken to identify relevant papers. Text words used in the titles and abstracts of retrieved records, as well as index terms used to describe them, were used to develop the full search strategy (Multimedia Appendix 1), which was adapted for each database. The databases we searched were MEDLINE (Ovid), CENTRAL (Ovid), Embase (Ovid), CINAHL (EBSCO), Web of Science Core Collection, IEEE Xplore, and Compendex. We also searched arXiv for relevant preprints. All databases were searched from inception to November 24, 2022. Reference lists of all included records were likewise searched for other records. Only English-language records were considered.

After the search was completed, duplicate citations were removed and all identified citations were uploaded into Rayyan [38] for title and abstract and full-text screening by 2 independent reviewers. An abstract screening tool was used to aid in the screening process (Multimedia Appendix 2). The texts of potentially relevant records were retrieved in full and assessed in the same manner. Disagreements were resolved through discussion or by consultation with an additional reviewer.

Assessment of the Employment of Reporting Guidelines and RoB

A data extraction tool (Multimedia Appendix 3)—that had been piloted with 20% (2/10) of the included reports by 2 independent reviewers—was used to abstract the relevant data. After piloting the tool, a single reviewer extracted data from the remaining sources with validation by an additional reviewer. The data were summarized using tables and presented narratively.

We determined the extent to which the included reports employed TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines using the TRIPOD adherence assessment form [39], and used the PROBAST (Prediction Model Risk of Bias Assessment Tool) to conduct critical appraisal [40]. Further, 2 reviewers assessed both employment of reporting guidelines and RoB for 20% (2/10) of the included reports; the remaining assessments were carried out by 1 reviewer (with an additional reviewer available for validation). In studies that developed multiple models, we only evaluated reporting and RoB for those that were image-based. To facilitate comparison between the reporting level of TRIPOD items, we chose arbitrary thresholds to denote high (≥70%), moderate (40%-69%), and low (1%-39%) adherence.

The TRIPOD adherence form and PROBAST were modified as needed for the purposes of this review. As has been noted in other reviews [33,41-43], it is difficult to assess RoB in the predictors of deep learning (DL) models that use images for prediction, as the image features are automatically selected by the algorithm. Still, we deemed image capture considerations important (eg, whether images were systematically captured) and altered the relevant TRIPOD and PROBAST items accordingly. The full list of modifications can be found in Multimedia Appendix 4.

Results

Study Inclusion

The search retrieved 796 records, or 699 records after duplicates were removed (Figure 1). We excluded 684 records during initial screening and full-text screened 15 reports. We identified 10 reports that met the eligibility criteria. The reference lists of these reports had an additional 16 potentially relevant records, though none met the eligibility criteria.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram showing study selection process. ML: machine learning; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; SSI: surgical site infection.

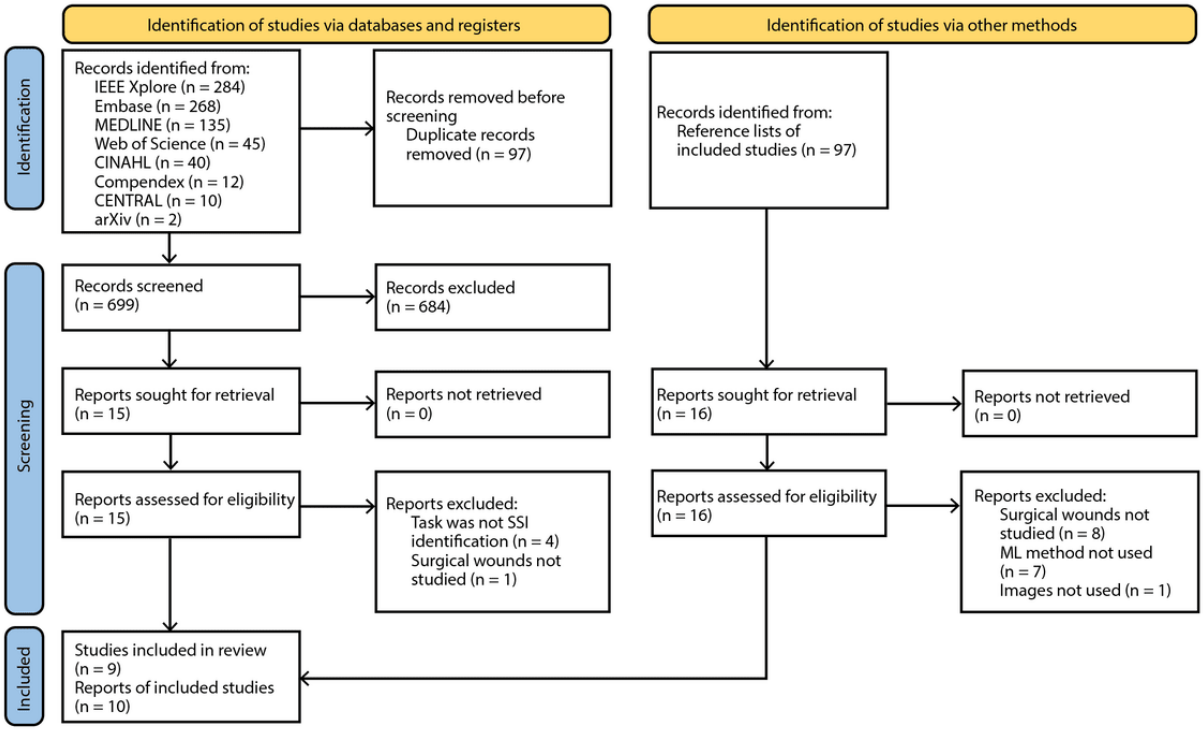

Review Findings

The included studies took place in a variety of settings, across a wide range of cohort sizes (Table 1). Important study characteristics were sometimes unclear or not reported. The full data extraction sheet can be found in Multimedia Appendix 5.

XSL•FO
RenderX

Table 1. Study characteristics.

Author	Purpose	Setting	Cohort	Events
Fletcher et al [44]	To develop a model for predicting SSI[a] in C-section wounds from thermal images taken with smartphones	Women who underwent C-section at a particular hospital in Kigali, Rwanda, between September 2019 and February 2020 prospectively enrolled on postoperative day 1	In total, 530 participants	In total, 30 participants with infected wounds
Fletcher et al [45]	To develop a model for predicting SSI in C-section wounds from color images taken with mobile devices	Women aged >18 years who underwent C-section at a particular hospital in Kigali, Rwanda, between March and October 2017 enrolled prior to discharge	In total, 572 participants (out of 729) that returned for follow-up	In total, 62 participants with infected wounds
Wu et al [46]	To develop an automatic monitoring tool for surgical wounds based on smartphone images	Prospectively collected wound database of patients who had undergone laparotomy, minimal invasive surgery, or hernia repair at an Asian academic center	In total, 480 wound images from 100 patients	In total, 136 images of infected wounds
Fletcher et al [47]	To develop models for predicting SSI in C-section wounds from questionnaire and image data	Women aged 18+ years who underwent C-section at a particular hospital in Kigali, Rwanda, between March and October 2017 enrolled prior to discharge	In total, 572 participants (out of 729) that returned for follow-up; images available for 568 patients	In total, 62 participants with infected wounds
Hsu et al [48]	To develop an automatic wound interpretation app for automated wound monitoring	Images of chest, abdomen, back, hand, and podiatry wounds collected from the Department of Surgery and Department of Internal Medicine of National Taiwan University Hospital	In total, 293 wound images	In training set, 27 infection images; total number unclear
Lüneburg et al [49]	To explore ML[b] approaches for remote LVAD[c] patient monitoring using images	Images of LVAD driveline exit sites obtained from Schüchtermann-Schiller'sche Kliniken and Hannover Medical School	In total, 745 images from 61 patients, though only 732 are labeled	In total, 212 images of mild infection and 37 images of severe infection
Shenoy et al [50]	To develop a model that can identify the onset of wound ailments from smartphone images	Images collected primarily from patients and surgeons at the Palo Alto Veterans Affairs Hospital and the Washington University Medical Center	In total, 1335 images	In total, 355 images of infection
Hsu et al [51]	To develop a model for recognizing SSI	Images collected from the Department of Surgery of National Taiwan University Hospital	In total, 42 images	In total, 30 images of infection
Zeng et al [52]	To develop a system for automatic wound detection and subsequent infection detection	Not reported	Total unclear; 6 images for testing	Unclear
Wang et al [53]	To develop an integrated system for automatic wound segmentation and analysis of wound conditions from wound images	Images collected from New York University Wound Database	In total, 3400 images	In total, 155 images of infection

[a]SSI: surgical site infection.

[b]ML: machine learning.

[c]LVAD: left ventricular assist device.

The earliest included paper was published in 2015 [53], 6 papers were published between 2017 and 2019 [47-52], and 3 papers were published between 2020 and November 2022 [44-46].

The objective of the included studies was generally to develop models for identifying surgical wound infection from images. In some cases, the purpose was broader; 2 studies sought to identify the presence of various wound attributes (eg, granulation) [46,50] and 4 studies developed models for automatic wound segmentation [48,49,51,53]. Other objectives included healing progress prediction [53], surface area estimation [53], and wound detection [52].

Patients, Procedures, and Image Capture

The types of patients and surgical procedures studied varied. In total, 3 papers focused on C-section patients in rural Rwanda [44,45,47], while another study examined patients implanted with a left ventricular assist device in Germany [49]. Further, 2 studies conducted in Asia described the surgical procedures

XSL•FO
RenderX

more broadly; for instance, 1 paper included patients that had undergone laparotomy, minimal invasive surgery, or hernia repair [46], while another included surgical wounds of the chest, abdomen, back, hands, and feet [48]. In 4 papers, this information was not specified [50-53].

The context of image capture likewise varied (Table 2). Most studies simply stated that images were obtained from one or more sites or data sets [48-51,53], without further details on how the images were selected; though 1 study additionally indicated that the data were "prospectively collected" [46] and the studies conducted in Rwanda described their cohorts in the greatest detail [44,45,47].

Table 2. Study data collection and ML[a] methodology.

Author	Time of image capture	Imaging modality	Outcome determination	Modeling methods	Performance metric for best-performing model
Fletcher et al [44]	Approximately 10 days after surgery	Thermal images taken by community health workers with a thermal camera module connected to smartphone that produces a JPG thermal image and a separate 2D temperature array	Physical examination performed by general practitioner	CNN[b]	Median AUC[c]: 0.90
Fletcher et al [45]	Approximately 10 days after surgery	Color images taken by community health workers with Android tablets	Physical examination performed by general practitioner	CNN	Median AUC: 0.655
Wu et al [46]	Just after surgery, during hospitalization, and in outpatient clinic follow-up	Color images taken by surgeons with smartphones	Annotation of abnormal wound features on images performed by surgeons	CNN, SVM[d], RF[e], GB[f]	Median AUC: 0.833
Fletcher et al [47]	Approximately 10 days after surgery	Color images taken by community health workers with Android tablets	Physical examination performed by general practitioner	Unclear; potentially both SVM and logistic regression	Median AUC: 1.0
Hsu et al [48]	Not reported	Color images taken with smartphones	Unclear, but likely annotation of images by 3 physicians	SVM	Overall accuracy: 0.8358
Lüneburg et al [49]	Not reported	Color images; device not reported	Unclear, but likely based on physical examination performed by physicians	CNN	Overall accuracy: 0.670
Shenoy et al [50]	Not reported	Color images taken by patients and surgeons with smartphones	Not reported	CNN	AUC: 0.82
Hsu et al [51]	Not reported	Color images; device not reported	Not reported	SVM	Overall accuracy: 0.9523
Zeng et al [52]	Not reported	Color images; device not reported	Not reported	SVM	AUCs varied depending on the infection-related wound attribute, but ranged from 0.7682 to 0.9145.
Wang et al [53]	Not reported	Color images; device not reported	Not reported	SVM using CNN features	AUC: 0.847

[a]ML: machine learning.

[b]CNN: convolutional neural network.

[c]AUC: area under curve.

[d]SVM: support vector machine.

[e]RF: random forest.

[f]GB: gradient boosting.

In the studies conducted with C-section patients, the wounds were photographed approximately 10 days after surgery, with infection assessment taking place on the same day [44,45,47]. Another study collected images at multiple time points: immediately after surgery, during hospitalization, and at a later follow-up, though the number of days post surgery was not indicated [46]. However, the time at which the images were taken relative to surgery and the time at which infection was assessed relative to image capture were not reported in 6 records [48-53].

XSL•FO
RenderX

In terms of the images themselves, 9 studies used color images [45-53], and 1 used thermal images [44]. Further, 6 studies used a mobile device (either smartphone or tablet) to capture the image [44-48,50], while others did not report the device used [49,51-53]. Across studies that reported the persons responsible for capturing the images, community health workers were typically responsible [44,45,47]; 1 study used images taken by surgeons [46]; and another used images collected by both patients and surgeons [50].

Assessment of surgical wound infection establishes the model ground truth and occurred mainly through face-to-face physical examination [44,45,47,49], through manual annotation of the wound images [46,48,51], or was not reported [50,52,53].

ML Approaches

All the included records were model development studies (ie, no external validation). In total, 4 papers used convolutional neural networks (CNNs) [44,45,49,50], 3 used support vector machines (SVMs) developed using handcrafted features [48,51,52], 1 trained an SVM classifier using CNN-derived features [53], 1 used a CNN, an SVM, a random forest model, and a gradient boosting classifier [46], and 1 paper's methods were not entirely clear but may have involved both logistic regression and SVMs [47]. Additional technical details are available in Multimedia Appendix 6.

The number of images used for developing an infection detection model ranged from just 42 [51] to 3400 [53]. Likewise, the proportion of images of infected wounds ranged from 4.6% (155/3400) [53] to 71.4% (30/42) [51]. In some cases, there was 1 image per patient [44,45,47], while in others, there were multiple per patient [46,49] or the number of patients was not reported [48,50-53].

In 5 papers, the classification task was binary [44-47,53], while in most others, the task was multiclass. In 1 paper, multiclass classification entailed distinguishing between mild, severe, and no infection [49], while in 3 others, the model differentiated between various infection-related wound attributes, such as granulation and swelling [48,51,52]. In contrast, 1 paper addressed a multilabel task in which the model identified the presence of a wound, infection, granulation, and drainage per image [50].

All studies reported model performance. In total, 7 studies reported area under the receiver operating characteristic curve values, which ranged from 0.655 [45] to 1.0 [47] for the best-performing models. The remaining studies reported overall accuracies, ranging from 0.670 [49] to 0.952 [51] for the best-performing models, as well as other performance metrics (eg, F_1-scores).

Employment of Reporting Guidelines

There were a few TRIPOD items that were highly employed (ie, employed by at least 7 out of the 10 included studies). For instance, all papers reported their objectives, and most reported background information, overall interpretations of study results, and descriptions of whether actions were taken to standardize image capture or otherwise systematically identify wounds from the images. In addition, 6 TRIPOD items had moderate employment (employed by between 4 and 6 studies); namely, the reporting of data sources and study setting, descriptions of model-building procedures, the number of participants or images (and the number showing infection), study limitations, as well as the potential clinical use of the models and future research directions.

Employment of 8 TRIPOD items was low (employed by between 1 and 3 studies), including items related to the reporting of participant selection methods, descriptions of how and when images were taken, rationales for sample sizes, the flow of participants within the paper, explanations of how to use the models, and funding details. Most studies did not completely use these guidelines in terms of outcome assessment, as there was often no indication of the criteria used to diagnose surgical wound infection or the time interval between surgery and assessment was unclear.

An additional 7 TRIPOD items were not reported in any of the included studies. Titles and abstracts did not employ reporting guidelines, and participant demographics were not reported. Similarly, model calibration was not discussed, and in studies that did not exclusively use DL methods for infection detection [46-48,51-53], reporting of feature modeling details did not meet TRIPOD guidelines.

About RoB

The RoB assessment led to 9 studies being identified as having an overall high RoB, while the remaining study was determined to have overall unclear RoB (Table 3). The participants domain was determined to be unclear in terms of RoB because little information about the source of data and recruitment methods was reported [46,48-53]. The 3 papers on C-section patients in Rwanda were at low RoB for this domain, as the nature of these works was cross-sectional and the cohorts were well defined [44,45,47]. In terms of predictors, we identified 5 papers as being at high RoB since there was variability in image capture conditions without later accounting for this variability [45-47,50,53]. In contrast, other papers were judged to be at low RoB for this domain because they segmented the wound prior to infection detection [48,49,51] or placed a frame around the wound prior to image capture [44], improving the uniformity of images processed for model training. Likewise, most studies were rated as having unclear RoB in the outcome domain, largely because the specific criteria used to gauge the presence of surgical wound infection were not reported. In other cases, the outcome domain was determined to be at high RoB because the presence of infection was determined solely from images, as opposed to by face-to-face review. In 8 studies, the analysis domain was assessed as being at high RoB for many reasons [45,47-53], including omission of participants in model development, an absence of both discrimination and calibration measures, and failure to appropriately account for overfitting.

XSL•FO
RenderX

Table 3. PROBAST[a] RoB[b] assessment of the 10 included reports.

Study	RoB				
	Participants	Predictors	Outcome	Analysis	Overall
Fletcher et al [44]	+[c]	+	?[d]	?	?
Fletcher et al [45]	+	–[e]	?	–	–
Wu et al [46]	?	–	–	?	–
Fletcher et al [47]	+	–	?	–	–
Hsu et al [48]	?	+	–	–	–
Lüneburg et al [49]	?	+	?	–	–
Shenoy et al [50]	?	–	?	–	–
Hsu et al [51]	?	+	–	–	–
Zeng et al [52]	?	?	?	–	–
Wang et al [53]	?	–	?	–	–

[a]PROBAST: Prediction Model Risk of Bias Assessment Tool.

[b]RoB: risk of bias.

[c]+ indicates low RoB.

[d]? indicates unclear RoB.

[e]– indicates high RoB.

Discussion

Principal Findings

This scoping review aimed to characterize the available research on ML approaches for the image-based identification of surgical wound infections. Such research is important as it can be integrated with remote patient monitoring, which enables improved health care decision-making and management, with additional benefits such as reduced travel burden. Initial work has suggested that remote image-based monitoring of wounds is feasible and associated with higher patient satisfaction [54-56], and is at least comparable to routine in-person care in terms of time to infection diagnosis [57]. Other aspects of wound assessment targeted by image-based remote patient monitoring include identification of dehiscence and surface area and temperature measurements [58,59], though much has not been automated or ML-based.

Despite the extensive body of ML-based work using medical images in other specialties [60,61], there is scarce ML research on the identification of surgical wound infections from digital images. We identified only 10 such papers, 7 of which were conference papers, which limits the space for reporting and likely contributed to the low reporting of TRIPOD items. In contrast, a recent review of ML for SSI detection identified 32 papers that used structured electronic health record, free-text, or administrative data for prediction [31], suggesting that ML-based SSI detection research has mostly used these more readily available data sources. While models based on such in-hospital data perform well in the context of inpatient SSI detection, they may be limited in their practical application during clinical care, as visual inspection is the essential mode by which infection is identified. In terms of incorporating innovative imaging techniques, thermal imaging has recently emerged as a potentially valuable tool in the management of surgical wounds [62-64]. Thermal imaging can be used with mobile devices [44,65], which facilitates its application for postdischarge monitoring, and may better generalize to different skin colors. On the other hand, the utility of electronic health record– or text-based models for postdischarge surveillance is perhaps less clear. Current postdischarge surgical wound surveillance largely depends on evaluation at follow-up visits, which may be infrequent and not timely [66], or on patient self-assessment, which is not reliable [67,68]. ML for the image-based identification of surgical wound infections presents the opportunity to automate this practice.

Reporting Data Collection Details

ML hinges on effective data collection, which can be challenging in outpatient or remote monitoring settings; hence, this type of research is still in early development. Although virtual care as a model of health care is relatively new, progress has been made in terms of data collection technology [64,69], similar telemedicine research without ML [70-72], and monitoring of other wound types [73,74]. As almost three-fourths of individuals worldwide own a mobile phone [75], leveraging this technology for remote monitoring holds potential. Still, it is worth noting that mobile phone ownership and mobile network coverage is lower in certain geographical areas and in low-income groups. In these contexts, alternative approaches, such as in-hospital follow-up with pictures taken by a community health worker [44], may be more appropriate. In terms of the data used in the included studies, it has mainly been collected in non-Western settings, and there are no publicly available data sets of surgical wound infection images, which presents a challenge to reproducibility and further development in the field. Likewise, the lack of reporting on image metadata (eg, gender and age distributions, procedures received, and occurrence of surgical complications) and eligibility criteria

XSL•FO
RenderX

limits the understanding of the populations that this research can be generalized to and contributes to RoB in terms of participant selection. Reporting of such details needs improvement for the progression of different prototypes for different subpopulations in this domain.

Transparency and Standardization in Model Development

The nature of the models developed in the included studies was diagnostic rather than prognostic. Similarly, none of the included papers performed out-of-sample external validation, highlighting the newness of this field of work and opportunity for further maturity. Interestingly, 4 papers published between 2017 and 2019 did not use DL methods, perhaps because the expertise required for development of such models was not yet widely available. Model performance is likewise not well-standardized in its reporting, as no papers reported on calibration and some did not include discrimination measures, which gives rise to RoB in analysis methods. Many papers did not report on measures to address overfitting, which calls the developed models' generalizability into question. Despite the partial reporting, the performance of the models in the included papers suggest that image-based ML for identification of surgical wound infection holds promise. In order to better understand their generalizability and reliability, future studies should externally validate and calibrate the developed models and report areas under the curve (as opposed to solely reporting other measures such as accuracy), and provide transparent documentation (eg, open-source code) to promote reproducibility and collaboration. Considering that interpretability and explainability support clinician trust [76], researchers may likewise wish to explore these concepts in future work.

Employment of Reporting Guidelines and RoB

Standardization of Image Capture

Use of TRIPOD guidelines was mostly low and RoB was generally unclear or high. This was in part due to participant- and analysis-related considerations discussed above; however, there were also concerns with the images themselves. In most studies, the way in which the images were taken, environmental conditions, the persons responsible for taking them, and the time of image capture relative to surgery, were not reported in detail. Still, there was often variability in the conditions of image capture, which might be attributed to unique challenges associated with collection and standardization in this particular modality. As opposed to other modalities, surgical wound infection images are largely taken by hand, without explicit training or guidance, which makes for considerable differences among images and introduces RoB in terms of model predictors. Efforts to standardize image capture help reduce RoB by minimizing systematic differences between images of infected versus noninfected wounds. Recent approaches such as instructions for patient-generated surgical wound images [77] or automated color calibration, scaling, and rotation correction [78] suggest that these considerations are receiving attention. Some studies created segmentation algorithms to capture the wound more reliably from the nonuniform images, which may have hindered the development of infection detection models. Segmentation and classification represent distinct areas of research, though many studies developed their own segmentation models rather than using or building on existing segmentation algorithms. In future work, specific directions detailing the time (relative to surgery), method, and conditions of image capture should be provided in order to reduce unwanted variability, and image processing steps can be undertaken for further standardization.

Transparency in Outcome Assessment

Outcome assessment was also not well reported in most papers. While there is no universally accepted and objective gold standard for SSI detection [79], clinical examination (involving direct observation of the wound) is frequently used as a reference standard [68,70,72,80,81]. Although some studies did perform in-person clinical examination, none reported the specific criteria used to gauge the presence of infection. Considering that there are differences in the rates of reported SSIs depending on the criteria used [82], specifying these criteria is important to more accurately assess the RoB arising from outcome assessment. It is worth noting, however, that there are challenges associated with in-person postoperative wound assessment. Surgical wound infections progress variably, with some only apparent after the 30-day postoperative surveillance benchmark [83,84]. However, extended in-person follow-up timeframes may require additional administrative resources. In practice, the criteria employed for SSI assessment typically consider both feasibility and validity [79]. This may necessitate striking a balance between resources, time constraints, and quality of assessment, which can pose challenges to the comprehensive evaluation of surgical wound infections. On a smaller scale, interrater reliability of in-person SSI assessment using established criteria can be modest [85,86], and in rural areas, there may be limited access to high-quality in-person wound care. Where feasible, determination of ground truth should use established criteria for infection and employ multiple independent assessors to minimize RoB.

Limitations

There are some limitations to this review. For instance, additional searching (eg, forward citation searching) may have led to more relevant reports being identified, as may have searching grey literature sources, which would reduce selection bias. We may have missed other relevant non–English-language papers, potentially excluding valuable studies. The included studies are from diverse locations (eg, Rwanda, Germany, and Taiwan), though this does not fully compensate for the potential language bias. Similarly, data extraction and the TRIPOD and PROBAST assessments were mainly completed by 1 reviewer, which introduces a potential source of bias in our findings. The modifications made to the TRIPOD and PROBAST tools may limit the ability to compare the results of our assessments to those of other reviews. Artificial intelligence–oriented extensions of both tools are in development [87] and will facilitate their use in appraising ML-based studies.

Conclusions

The use of ML for the image-based identification of surgical wound infections remains in the early stages, with only 10 studies available and a need for reporting standardization. Future

XSL•FO
RenderX

development and validation of such models should carefully consider image variability, overfitting concerns, and criteria for determination of infection. These considerations are important to advance the state of image-based ML for wound management, which has the potential to automate traditionally labor-intensive practices.

Acknowledgments

This study was funded through an award from the Hamilton Health Sciences Research Institute. The funders had no role in the collection, analysis, and interpretation of data; in the writing of this paper; and in the decision to submit this paper for publication.

Conflicts of Interest

JP declares research funding support from Roche Canada. PJD has received grants from Abbott Diagnostics, Siemens Canada, and Roche Diagnostics; received consulting fees from Abbott Laboratories, Renibus Therapeutics, Roche Diagnostics, and Trimedic Canada; received monitoring devices from CloudDX and Philips Healthcare; participated in advisory board meetings for Bayer AG; and is a member of the Data Safety Monitoring Board for the PEPPER (Comparative Effectiveness of Pulmonary Embolism Prevention After Hip and Knee Replacement) Study, New Hampshire. The other authors have no conflicts of interest to disclose.

Multimedia Appendix 1
Search strategy.
[DOCX File , 22 KB - jmir_v26i1e52880_app1.docx]

Multimedia Appendix 2
Abstract screening tool.
[DOCX File , 14 KB - jmir_v26i1e52880_app2.docx]

Multimedia Appendix 3
Data extraction tool.
[DOCX File , 14 KB - jmir_v26i1e52880_app3.docx]

Multimedia Appendix 4
TRIPOD and PROBAST altered or excluded items.
[DOCX File , 17 KB - jmir_v26i1e52880_app4.docx]

Multimedia Appendix 5
Complete results of data extraction, assessment of TRIPOD employment, and PROBAST assessment.
[DOCX File , 70 KB - jmir_v26i1e52880_app5.docx]

Multimedia Appendix 6
Additional technical details.
[DOCX File , 28 KB - jmir_v26i1e52880_app6.docx]

Multimedia Appendix 7
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 101 KB - jmir_v26i1e52880_app7.pdf]

References

1. Tevis SE, Kennedy GD. Postoperative complications and implications on patient-centered outcomes. J Surg Res 2013;181(1):106-113 [FREE Full text] [doi: 10.1016/j.jss.2013.01.032] [Medline: 23465392]
2. Endo I, Kumamoto T, Matsuyama R. Postoperative complications and mortality: are they unavoidable? Ann Gastroenterol Surg 2017;1(3):160-163 [FREE Full text] [doi: 10.1002/ags3.12045] [Medline: 29863116]
3. Monahan M, Jowett S, Pinkney T, Brocklehurst P, Morton DG, Abdali Z, et al. Surgical site infection and costs in low- and middle-income countries: a systematic review of the economic burden. PLoS One 2020;15(6):e0232960 [FREE Full text] [doi: 10.1371/journal.pone.0232960] [Medline: 32497086]
4. Guest JF, Fuller GW, Vowden P. Costs and outcomes in evaluating management of unhealed surgical wounds in the community in clinical practice in the UK: a cohort study. BMJ Open 2018;8(12):e022591 [FREE Full text] [doi: 10.1136/bmjopen-2018-022591] [Medline: 30552253]

XSL•FO
RenderX

5. Onyekwelu I, Yakkanti R, Protzer L, Pinkston CM, Tucker C, Seligson D. Surgical wound classification and surgical site infections in the orthopaedic patient. J Am Acad Orthop Surg Glob Res Rev 2017;1(3):e022 [FREE Full text] [doi: 10.5435/JAAOSGlobal-D-17-00022] [Medline: 30211353]
6. Gillespie BM, Harbeck E, Rattray M, Liang R, Walker R, Latimer S, et al. Worldwide incidence of surgical site infections in general surgical patients: a systematic review and meta-analysis of 488,594 patients. Int J Surg 2021;95:106136 [FREE Full text] [doi: 10.1016/j.ijsu.2021.106136] [Medline: 34655800]
7. Wang SC, Au Y, Ramirez-GarciaLuna JL, Lee L, Berry GK. The promise of smartphone applications in the remote monitoring of postsurgical wounds: a literature review. Adv Skin Wound Care 2020;33(9):489-496. [doi: 10.1097/01.ASW.0000694136.29135.02] [Medline: 32810062]
8. Gunter RL, Chouinard S, Fernandes-Taylor S, Wiseman JT, Clarkson S, Bennett K, et al. Current use of telemedicine for post-discharge surgical care: a systematic review. J Am Coll Surg 2016;222(5):915-927 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2016.01.062] [Medline: 27016900]
9. Mesko B. The role of artificial intelligence in precision medicine. Expert Rev Precis Med Drug Dev 2017;2(5):239-241 [FREE Full text] [doi: 10.1080/23808993.2017.1380516]
10. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
11. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019;8(7):2328-2331 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
12. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc ;2020:191-200 [FREE Full text] [Medline: 32477638]
13. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. npj Digit Med 2021;4(1):5 [FREE Full text] [doi: 10.1038/s41746-020-00376-2] [Medline: 33420381]
14. Carrión H, Jafari M, Bagood MD, Yang HY, Isseroff RR, Gomez M. Automatic wound detection and size estimation using deep learning algorithms. PLoS Comput Biol 2022;18(3):e1009852 [FREE Full text] [doi: 10.1371/journal.pcbi.1009852] [Medline: 35275923]
15. Wang C, Anisuzzaman DM, Williamson V, Dhar MK, Rostami B, Niezgoda J, et al. Fully automatic wound segmentation with deep convolutional neural networks. Sci Rep 2020;10(1):21897 [FREE Full text] [doi: 10.1038/s41598-020-78799-w] [Medline: 33318503]
16. Sikka K, Ahmed AA, Diaz D, Goodwin MS, Craig KD, Bartlett MS, et al. Automated assessment of children's postoperative pain using computer vision. Pediatrics 2015;136(1):e124-e131 [FREE Full text] [doi: 10.1542/peds.2015-0029] [Medline: 26034245]
17. Şimşek İ, Şirolu C. Analysis of surgical outcome after upper eyelid surgery by computer vision algorithm using face and facial landmark detection. Graefes Arch Clin Exp Ophthalmol 2021;259(10):3119-3125. [doi: 10.1007/s00417-021-05219-8] [Medline: 33963919]
18. Barakat-Johnson M, Jones A, Burger M, Leong T, Frotjold A, Randall S, et al. Reshaping wound care: evaluation of an artificial intelligence app to improve wound assessment and management amid the COVID-19 pandemic. Int Wound J 2022;19(6):1561-1577 [FREE Full text] [doi: 10.1111/iwj.13755] [Medline: 35212459]
19. Mukherjee R, Manohar DD, Das DK, Achar A, Mitra A, Chakraborty C. Automated tissue classification framework for reproducible chronic wound assessment. Biomed Res Int ;2014:851582 [FREE Full text] [doi: 10.1155/2014/851582] [Medline: 25114925]
20. Dreifke MB, Jayasuriya AA, Jayasuriya AC. Current wound healing procedures and potential care. Mater Sci Eng C 2015;48:651-662 [FREE Full text] [doi: 10.1016/j.msec.2014.12.068] [Medline: 25579968]
21. Aldaz G, Shluzas LA, Pickham D, Eris O, Sadler J, Joshi S, et al. Hands-free image capture, data tagging and transfer using Google Glass: a pilot study for improved wound care management. PLoS One 2015;10(4):e0121179 [FREE Full text] [doi: 10.1371/journal.pone.0121179] [Medline: 25902061]
22. Zahia S, Garcia-Zapirain B, Elmaghraby A. Integrating 3D model representation for an accurate non-invasive assessment of pressure injuries with deep learning. Sensors (Basel) 2020;20(10):2933 [FREE Full text] [doi: 10.3390/s20102933] [Medline: 32455753]
23. Lau CH, Yu KHO, Yip TF, Luk LYF, Wai AKC, Sit TY, et al. An artificial intelligence-enabled smartphone app for real-time pressure injury assessment. Front Med Technol 2022;4:905074 [FREE Full text] [doi: 10.3389/fmedt.2022.905074] [Medline: 36212608]
24. Zoppo G, Marrone F, Pittarello M, Farina M, Uberti A, Demarchi D, et al. AI technology for remote clinical assessment and monitoring. J Wound Care 2020;29(12):692-706 [FREE Full text] [doi: 10.12968/jowc.2020.29.12.692] [Medline: 33320742]
25. Sood A, Granick MS, Trial C, Lano J, Palmier S, Ribal E, et al. The role of telemedicine in wound care: a review and analysis of a database of 5,795 patients from a mobile wound-healing center in Languedoc-Roussillon, France. Plast Reconstr Surg 2016;138(Suppl 3):248S-256S. [doi: 10.1097/PRS.0000000000002702] [Medline: 27556769]

XSL•FO
RenderX

26. Zhang J, Mihai C, Tüshaus L, Scebba G, Distler O, Karlen W. Wound image quality from a mobile health tool for home-based chronic wound management with real-time quality feedback: randomized feasibility study. JMIR mHealth uHealth 2021;9(7):e26149 [FREE Full text] [doi: 10.2196/26149] [Medline: 34328440]
27. McGillion MH, Parlow J, Borges FK, Marcucci M, Jacka M, Adili A, et al. Post-discharge after surgery Virtual Care with Remote Automated Monitoring-1 (PVC-RAM-1) technology versus standard care: randomised controlled trial. BMJ 2021;374:n2209 [FREE Full text] [doi: 10.1136/bmj.n2209] [Medline: 34593374]
28. Dabas M, Schwartz D, Beeckman D, Gefen A. Application of artificial intelligence methodologies to chronic wound care and management: a scoping review. Adv Wound Care (New Rochelle) 2023;12(4):205-240. [doi: 10.1089/wound.2021.0144] [Medline: 35438547]
29. Anisuzzaman DM, Wang C, Rostami B, Gopalakrishnan S, Niezgoda J, Yu Z. Image-based artificial intelligence in wound assessment: a systematic review. Adv Wound Care (New Rochelle) 2022;11(12):687-709. [doi: 10.1089/wound.2021.0091] [Medline: 34544270]
30. Hurlow J, Bowler PG. Acute and chronic wound infections: microbiological, immunological, clinical and therapeutic distinctions. J Wound Care 2022;31(5):436-445 [FREE Full text] [doi: 10.12968/jowc.2022.31.5.436] [Medline: 35579319]
31. Wu G, Khair S, Yang F, Cheligeer C, Southern D, Zhang Z, et al. Performance of machine learning algorithms for surgical site infection case detection and prediction: a systematic review and meta-analysis. Ann Med Surg (Lond) 2022;84:104956 [FREE Full text] [doi: 10.1016/j.amsu.2022.104956] [Medline: 36582918]
32. Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021;375:n2281 [FREE Full text] [doi: 10.1136/bmj.n2281] [Medline: 34670780]
33. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
34. Peters MDJ, Godfrey C, McInerney P, Munn Z, Trico AC, Khalil H. Chapter 11: scoping reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2020.
35. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
36. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
37. Machine learning approaches for computer vision tasks related to the identification of surgical site infections: a scoping review protocol. OSF Registries. URL: https://osf.io/3k9xq [accessed 2023-03-23]
38. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
39. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open 2019;9(4):e025611 [FREE Full text] [doi: 10.1136/bmjopen-2018-025611] [Medline: 31023756]
40. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
41. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199-217 [FREE Full text] [doi: 10.1038/s42256-021-00307-0]
42. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ 2020;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
43. Frizzell TO, Glashutter M, Liu CC, Zeng A, Pan D, Hajra SG, et al. Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: a systematic review. Ageing Res Rev 2022;77:101614. [doi: 10.1016/j.arr.2022.101614] [Medline: 35358720]
44. Fletcher RR, Schneider G, Bikorimana L, Rukundo G, Niyigena A, Miranda E, et al. The use of mobile thermal imaging and deep learning for prediction of surgical site infection. In: Annu Int Conf IEEE Eng Med Biol Soc. 2021 Presented at: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Mexico; November 1-5, 2021 p. 5059-5062 URL: https://ieeexplore.ieee.org/document/9630094 [doi: 10.1109/EMBC46164.2021.9630094]
45. Fletcher RR, Schneider G, Hedt-Gauthier B, Nkurunziza T, Alayande B, Riviello R, et al. Use of convolutional neural nets and transfer learning for prediction of surgical site infection from color images. In: Annu Int Conf IEEE Eng Med Biol Soc. 2021 Presented at: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); November 1-5, 2021; Mexico p. 5047-5050. [doi: 10.1109/EMBC46164.2021.9630430]

XSL•FO
RenderX

46. Wu JM, Tsai CJ, Ho TW, Lai F, Tai HC, Lin MT. A unified framework for automatic detection of wound infection with artificial intelligence. Appl Sci 2020;10(15):5353 [FREE Full text] [doi: 10.3390/app10155353]
47. Fletcher RR, Olubeko O, Sonthalia H, Kateera F, Nkurunziza T, Ashby JL, et al. Application of machine learning to prediction of surgical site infection. In: Annu Int Conf IEEE Eng Med Biol Soc. 2019 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany p. 2234-2237. [doi: 10.1109/EMBC.2019.8857942]
48. Hsu JT, Chen YW, Ho TW, Tai HC, Wu JM, Sun HY, et al. Chronic wound assessment and infection detection method. BMC Med Inform Decis Mak 2019;19(1):99 [FREE Full text] [doi: 10.1186/s12911-019-0813-0] [Medline: 31126274]
49. Lüneburg N, Reiss N, Feldmann C, van der Meulen P, van de Steeg M, Schmidt T, et al. Photographic LVAD driveline wound infection recognition using deep learning. Stud Health Technol Inform 2019;260:192-199. [doi: 10.3233/978-1-61499-971-3-192] [Medline: 31118337]
50. Shenoy VN, Foster E, Aalami L, Majeed B, Aalami O. Deepwound: automated postoperative wound assessment and surgical site surveillance through convolutional neural networks. In: IEEE Int Conf Bioinform Biomed. 2018 Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3-6, 2018; Madrid, Spain p. 1017-1021. [doi: 10.1109/bibm.2018.8621130]
51. Hsu JT, Ho TW, Shih HF, Chang CC, Lai F, Wu JM. Automatic wound infection interpretation for postoperative wound image. 2017 Presented at: Proceedings SPIE 10225, Eighth International Conference on Graphic and Image Processing (ICGIP 2016); February 8, 2017; Tokyo, Japan p. 1022526. [doi: 10.1117/12.2266110]
52. Zeng YC, Liao KH, Wang CH, Lin Y, Chang WT. Implementation of post-operative wound analytics. In: IEEE Int Conf Consum Electron-Taiwan. 2017 Presented at: 2017 IEEE International Conference on Consumer Electronics—Taiwan (ICCE-TW); June 12-14, 2017; Taipei, Taiwan p. 93-94. [doi: 10.1109/icce-china.2017.7991011]
53. Wang C, Yan X, Smith M, Kochhar K, Rubin M, Warren SM, et al. A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks. In: Annu Int Conf IEEE Eng Med Biol Soc. 2015 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); August 25-29, 2015; Milan, Italy p. 2415-2418. [doi: 10.1109/EMBC.2015.7318881]
54. Gunter RL, Fernandes-Taylor S, Rahman S, Awoyinka L, Bennett KM, Weber SM, et al. Feasibility of an image-based mobile health protocol for postoperative wound monitoring. J Am Coll Surg 2018;226(3):277-286 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2017.12.013] [Medline: 29366555]
55. Gunter R, Fernandes-Taylor S, Mahnke A, Awoyinka L, Schroeder C, Wiseman J, et al. Evaluating patient usability of an image-based mobile health platform for postoperative wound monitoring. JMIR mHealth uHealth 2016;4(3):e113 [FREE Full text] [doi: 10.2196/mhealth.6023] [Medline: 27683059]
56. Mousa AY, Broce M, Monnett S, Davis E, McKee B, Lucas BD. Results of telehealth electronic monitoring for post discharge complications and surgical site infections following arterial revascularization with groin incision. Ann Vasc Surg 2019;57:160-169. [doi: 10.1016/j.avsg.2018.09.023] [Medline: 30500646]
57. McLean KA, Mountain KE, Shaw CA, Drake TM, Pius R, Knight SR, et al. Remote diagnosis of surgical-site infection using a mobile digital intervention: a randomised controlled trial in emergency surgery patients. npj Digit Med 2021;4(1):160 [FREE Full text] [doi: 10.1038/s41746-021-00526-0] [Medline: 34795398]
58. Wang SC, Anderson JAE, Evans R, Woo K, Beland B, Sasseville D, et al. Point-of-care wound visioning technology: reproducibility and accuracy of a wound measurement app. PLoS One 2017;12(8):e0183139 [FREE Full text] [doi: 10.1371/journal.pone.0183139] [Medline: 28817649]
59. Wiseman JT, Fernandes-Taylor S, Gunter R, Barnes ML, Saunders RS, Rathouz PJ, et al. Inter-rater agreement and checklist validation for postoperative wound assessment using smartphone images in vascular surgery. J Vasc Surg Venous Lymphat Disord 2016;4(3):320-328.e2 [FREE Full text] [doi: 10.1016/j.jvsv.2016.02.001] [Medline: 27318052]
60. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. npj Digit Med 2021;4(1):65 [FREE Full text] [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
61. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
62. Childs C, Wright N, Willmott J, Davies M, Kilner K, Ousey K, et al. The surgical wound in infrared: thermographic profiles and early stage test-accuracy to predict surgical site infection in obese women during the first 30 days after caesarean section. Antimicrob Resist Infect Control 2019;8:7 [FREE Full text] [doi: 10.1186/s13756-018-0461-7] [Medline: 30637101]
63. Siah CJR, Childs C, Chia CK, Cheng KFK. An observational study of temperature and thermal images of surgical wounds for detecting delayed wound healing within four days after surgery. J Clin Nurs 2019;28(11-12):2285-2295. [doi: 10.1111/jocn.14832] [Medline: 30791157]
64. Li F, Wang M, Wang T, Wang X, Ma X, He H, et al. Smartphone-based infrared thermography to assess progress in thoracic surgical incision healing: a preliminary study. Int Wound J 2023;20(6):2000-2009 [FREE Full text] [doi: 10.1111/iwj.14063] [Medline: 36542868]

XSL•FO
RenderX

65. Kanazawa T, Nakagami G, Goto T, Noguchi H, Oe M, Miyagaki T, et al. Use of smartphone attached mobile thermography assessing subclinical inflammation: a pilot study. J Wound Care 2016;25(4):177-182. [doi: 10.12968/jowc.2016.25.4.177] [Medline: 27064366]
66. Saunders RS, Fernandes-Taylor S, Rathouz PJ, Saha S, Wiseman JT, Havlena J, et al. Outpatient follow-up versus 30-day readmission among general and vascular surgery patients: a case for redesigning transitional care. Surgery 2014;156(4):949-958 [FREE Full text] [doi: 10.1016/j.surg.2014.06.041] [Medline: 25239351]
67. Whitby M, McLaws ML, Doidge S, Collopy B. Post-discharge surgical site surveillance: does patient education improve reliability of diagnosis? J Hosp Infect 2007;66(3):237-242. [doi: 10.1016/j.jhin.2007.04.014] [Medline: 17582652]
68. Richter V, Cohen MJ, Benenson S, Almogy G, Brezis M. Patient self-assessment of surgical site infection is inaccurate. World J Surg 2017;41(8):1935-1942. [doi: 10.1007/s00268-017-3974-y] [Medline: 28271262]
69. Maddah E, Beigzadeh B. Use of a smartphone thermometer to monitor thermal conductivity changes in diabetic foot ulcers: a pilot study. J Wound Care 2020;29(1):61-66. [doi: 10.12968/jowc.2020.29.1.61] [Medline: 31930943]
70. Nkurunziza T, Williams W, Kateera F, Riviello R, Niyigena A, Miranda E, et al. mHealth-community health worker telemedicine intervention for surgical site infection diagnosis: a prospective study among women delivering via caesarean section in rural Rwanda. BMJ Glob Health 2022;7(7):e009365 [FREE Full text] [doi: 10.1136/bmjgh-2022-009365] [Medline: 35902205]
71. Hedt-Gauthier B, Miranda E, Nkurunziza T, Hughes O, Boatin AA, Gaju E, et al. Telemedicine for surgical site infection diagnosis in rural Rwanda: concordance and accuracy of image reviews. World J Surg 2022;46(9):2094-2101. [doi: 10.1007/s00268-022-06597-8] [Medline: 35665833]
72. Totty JP, Harwood AE, Wallace T, Smith GE, Chetter IC. Use of photograph-based telemedicine in postoperative wound assessment to diagnose or exclude surgical site infection. J Wound Care 2018;27(3):128-135. [doi: 10.12968/jowc.2018.27.3.128] [Medline: 29509108]
73. Wang L, Pedersen PC, Strong DM, Tulu B, Agu E, Ignotz R. Smartphone-based wound assessment system for patients with diabetes. IEEE Trans Biomed Eng 2015;62(2):477-488 [FREE Full text] [doi: 10.1109/TBME.2014.2358632] [Medline: 25248175]
74. Howell RS, Liu HH, Khan AA, Woods JS, Lin LJ, Saxena M, et al. Development of a method for clinical evaluation of artificial intelligence-based digital wound assessment tools. JAMA Netw Open 2021;4(5):e217234 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.7234] [Medline: 34009348]
75. Facts and figures 2022—mobile phone ownership. International Telecommunication Union. 2022. URL: https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-mobile-phone-ownership/ [accessed 2023-10-03]
76. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In: Proceedings of the 4th Machine Learning for Healthcare Conference. 2019 Presented at: Machine Learning for Healthcare; August 2019; Ann Arbor, MI p. 359-380 URL: https://proceedings.mlr.press/v106/tonekaboni19a.html
77. Macefield RC, Blazeby JM, Reeves BC, King A, Rees J, Pullyblank A, et al. Remote assessment of surgical site infection (SSI) using patient-taken wound images: development and evaluation of a method for research and routine practice. J Tissue Viability 2023;32(1):94-101 [FREE Full text] [doi: 10.1016/j.jtv.2023.01.001] [Medline: 36681617]
78. Alayande BT, Prasad S, Abimpaye M, Bakorimana L, Niyigena A, Nkurunziza J, et al. Image-based surgical site infection algorithms to support home-based post-cesarean monitoring: lessons from Rwanda. PLOS Glob Public Health 2023;3(2):e0001584 [FREE Full text] [doi: 10.1371/journal.pgph.0001584] [Medline: 36963016]
79. Bruce J, Russell EM, Mollison J, Krukowski ZH. The quality of measurement of surgical wound infection as the basis for monitoring: a systematic review. J Hosp Infect 2001;49(2):99-108. [doi: 10.1053/jhin.2001.1045] [Medline: 11567554]
80. Lathan R, Sidapra M, Yiasemidou M, Long J, Totty J, Smith G, et al. Diagnostic accuracy of telemedicine for detection of surgical site infection: a systematic review and meta-analysis. npj Digit Med 2022;5(1):108 [FREE Full text] [doi: 10.1038/s41746-022-00655-0] [Medline: 35922663]
81. Guerra J, Guichon C, Isnard M, So S, Chan S, Couraud S, et al. Active prospective surveillance study with post-discharge surveillance of surgical site infections in Cambodia. J Infect Public Health 2015;8(3):298-301 [FREE Full text] [doi: 10.1016/j.jiph.2014.09.007] [Medline: 25466596]
82. Ashby E, Haddad FS, O'Donnell E, Wilson APR. How will surgical site infection be measured to ensure "high quality care for all"? J Bone Joint Surg Br 2010;92(9):1294-1299 [FREE Full text] [doi: 10.1302/0301-620X.92B9.22401] [Medline: 20798451]
83. Hopkins B, Eustache J, Ganescu O, Ciopolla J, Kaneva P, Fiore JF, et al. At least ninety days of follow-up are required to adequately detect wound outcomes after open incisional hernia repair. Surg Endosc 2022;36(11):8463-8471. [doi: 10.1007/s00464-022-09143-8] [Medline: 35257211]
84. Holihan JL, Flores-Gonzalez JR, Mo J, Ko TC, Kao LS, Liang MK. How long is long enough to identify a surgical site infection? Surg Infect (Larchmt) 2017;18(4):419-423. [doi: 10.1089/sur.2016.132] [Medline: 28004989]
85. Carpini GD, Giannella L, Di Giuseppe J, Fioretti M, Franconi I, Gatti L, et al. Inter-rater agreement of CDC criteria and ASEPSIS score in assessing surgical site infections after cesarean section: a prospective observational study. Front Surg 2023;10:1123193 [FREE Full text] [doi: 10.3389/fsurg.2023.1123193] [Medline: 37675248]

XSL•FO
RenderX

86. Allami MK, Jamil W, Fourie B, Ashton V, Gregg PJ. Superficial incisional infection in arthroplasty of the lower limb. Interobserver reliability of the current diagnostic criteria. J Bone Joint Surg Br 2005;87-B(9):1267-1271 [FREE Full text] [doi: 10.1302/0301-620X.87B9.16672] [Medline: 16129756]
87. Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7):e048008 [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]

Abbreviations

CNN: convolutional neural network
DL: deep learning
JBI: Joanna Briggs Institute
ML: machine learning
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PROBAST: Prediction Model Risk of Bias Assessment Tool
RoB: risk of bias
SSI: surgical site infection
SVM: support vector machine
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 18.09.23; peer-reviewed by D Hu, M Zhou, P Okoro, K McLean; comments to author 21.09.23; revised version received 09.11.23; accepted 12.12.23; published 18.01.24.

Please cite as:
Tabja Bortesi JP, Ranisau J, Di S, McGillion M, Rosella L, Johnson A, Devereaux PJ, Petch J
Machine Learning Approaches for the Image-Based Identification of Surgical Wound Infections: Scoping Review
J Med Internet Res 2024;26:e52880
URL: https://www.jmir.org/2024/1/e52880
doi:10.2196/52880
PMID:38236623

©Juan Pablo Tabja Bortesi, Jonathan Ranisau, Shuang Di, Michael McGillion, Laura Rosella, Alistair Johnson, PJ Devereaux, Jeremy Petch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Engaging Parents in Technology-Assisted Interventions for Childhood Adversity: Systematic Review

Grace Aldridge[1], BPsych(Hons); Alessandra Tomaselli[1], BSc; Clare Nowell[1], BPsych(Hons); Andrea Reupert[2], DipEd, GradDip(CounsPsychol), BA, PhD; Anthony Jorm[3], BA, MPsychol, PhD, DSc; Marie Bee Hui Yap[1,3], BA, BSc(Hons), MPsych, PhD

[1]Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton, Australia

[2]School of Educational Psychology and Counselling, Monash University, Clayton, Australia

[3]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia

Corresponding Author:
Marie Bee Hui Yap, BA, BSc(Hons), MPsych, PhD
Turner Institute for Brain and Mental Health, School of Psychological Sciences
Monash University
Level 5, 18 Innovation Walk
Clayton Campus
Clayton
Australia
Phone: 61 3 9905 0723
Email: marie.yap@monash.edu

Abstract

Background: Youth mental health problems are a major public health concern and are strongly associated with adverse childhood experiences (ACEs). Technology-assisted parenting programs can intervene with ACEs that are within a parent's capacity to modify. However, engagement with such programs is suboptimal.

Objective: This review aims to describe and appraise the efficacy of strategies used to engage parents in technology-assisted parenting programs targeting ACEs on the behavioral and subjective outcomes of engagement.

Methods: Using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines, we conducted a systematic review of peer-reviewed papers that described the use of at least 1 engagement strategy in a technology-assisted parenting program targeting ACEs that are within a parent's capacity to modify. A total of 8 interdisciplinary bibliographic databases (CENTRAL, CINAHL, Embase, OVID MEDLINE, OVID PsycINFO, Scopus, ACM, and IEEE Xplore) and gray literature were searched. The use of engagement strategies and measures was narratively synthesized. Associations between specific engagement strategies and engagement outcomes were quantitatively synthesized using the Stouffer method of combining *P* values.

Results: We identified 13,973 articles for screening. Of these, 156 (1.12%) articles were eligible for inclusion, and 29 (18.2%) of the 156 were associated with another article; thus, 127 studies were analyzed. Preliminary evidence for a reliable association between 5 engagement strategies (involving parents in a program's design, delivering a program on the web compared to face-to-face, use of personalization or tailoring features, user control features, and provision of practical support) and greater engagement was found. Three engagement strategies (professional support features, use of videos, and behavior change techniques) were not found to have a reliable association with engagement outcomes.

Conclusions: This review provides a comprehensive assessment and description of the use of engagement strategies and engagement measures in technology-assisted parenting programs targeting parenting-related ACEs and extends the current evidence with preliminary quantitative findings. Heterogeneous definition and measurement of engagement and insufficient engagement outcome data were caveats to this synthesis. Future research could use integrated definitions and measures of engagement to support robust systematic evaluations of engagement in this context.

Trial Registration: PROSPERO CRD42020209819; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=209819

(J Med Internet Res 2024;26:e43994) doi:10.2196/43994

XSL•FO
RenderX

KEYWORDS

systematic review; intervention; digital technology; parenting; children; technology; parenting program; engagement; support; adverse childhood experiences

Introduction

Parenting Programs to Prevent or Reduce Depression and Anxiety Disorders in Young People

Depression and anxiety disorders are major sources of the global disease burden in children and young people [1,2]. Parenting factors are known to influence the risk of developing these disorders in children and adolescents (ie, offspring aged 0-18 years, henceforth referred to as "young people") [3,4]. Unlike other known systemic or biological factors (eg, poverty or family history of psychopathology), parenting factors are within a parent's capacity to intervene. Parenting programs capitalize on the central role that parents and caregivers play in a young person's development by improving the parenting skills involved in supporting their young person's outcomes. There is good evidence to support the efficacy of parenting programs in improving young people's mental health outcomes [5] in between-group comparisons over time when compared with a control condition (no treatment) [6] and when compared with a range of control conditions such as usual care or attention controls [7]. However, patterns of poor engagement (such as low rates of enrollment and program completion) are common in studies of face-to-face parenting programs, which limit the potential benefits of these programs at the family and population level [8,9]. Common barriers include time constraints, conflicting schedules, perceived stigma, and stress from involvement in parenting programs [10,11].

Parenting Programs for Parents of Young People With Adverse Childhood Experiences

Barriers to engaging in parenting programs are especially prevalent in families who experience adversity, marginalization, and stress owing to socioeconomic pressure [12,13]. Parental stress is associated with maladaptive parenting behaviors (defined as parenting behaviors characterized by high hostility and low warmth [14], henceforth referred to as "maladaptive parenting") [15]. Recent evidence suggests that maladaptive parenting is as predictive of mental disorders and suicidality in young people as more commonly known family-level adverse childhood experiences (ACEs; such as child maltreatment and interparental conflict) [16]. Therefore, there is a clear potential for parenting programs to intervene with these family-level ACEs and to reduce or prevent the risk of mental disorders in young people. Further, benefits from engaging with such programs may potentially buffer against the stress experienced because of other systemic ACEs. However, a better understanding of strategies to enhance engagement with programs targeting family-level ACEs is needed, given that target families are likely to experience greater barriers to engagement.

Technology-Assisted Parenting Programs

Technology has the potential to minimize or overcome common barriers associated with engaging in face-to-face parenting programs. For example, technology can offer a user flexibility and choice regarding how and when they access a parenting program, as well as increased privacy. Functions such as automated reminders and content tailoring may also enhance the relevance and relationship between the program and its user [17]. The delivery and reach of existing services for parents can also be enhanced with technology, as it can carry out progress monitoring and content updates in a time-efficient manner and with fewer human resources. The potential benefits of technology-assisted parenting programs have therefore been widely explored over the past 2 decades [18], especially in recent years owing to the effects of the COVID-19 pandemic on families' ability to access face-to-face services [19]. There is a growing body of evidence supporting the efficacy of technology-assisted parenting programs in improving parenting outcomes (including maladaptive parenting), reducing their young person's internalizing problems [20,21] and externalizing problems [22-24], and promoting their physical and mental health [25]. Importantly, the efficacy of these programs has also been found for parenting outcomes and child problem behaviors for families experiencing socioeconomic disadvantage [26]. However, program effect sizes are often found to be small, warranting further exploration of how these effects might be enhanced.

Parental Engagement in Technology-Assisted Parenting Programs

Engagement in a program is a key mechanism for improving target behavioral outcomes [27]. In the context of parenting programs, engagement has specifically been conceptualized as 3 discrete behavioral components: *initial* engagement, measured by both intended and actual enrollment in a program; *ongoing* engagement, as indicated by measures of attendance or program completion; and *quality* of engagement, as shown through measures of active participation (such as completing specific program components either within or beyond the program itself) [28]. It has been suggested that the *quality* of engagement is most closely related to program outcomes and, hence, is suggested to be a key mechanism for positive parenting change [29].

Several systematic reviews evaluating the effects of technology-assisted parenting programs on target parents and young people's outcomes have also explored the use of specific engagement strategies or program features. For instance, Florean et al [24] found that parenting programs to reduce elevated or diagnosed behavior problems in young people that are delivered via videoconferencing yielded comparable effects on young people's outcomes with the programs delivered face-to-face. They also found that the effects on both young people and parenting outcomes were comparable between conditions that provided specialized support (ie, understanding and applying program content) and conditions that provided technical support (ie, using and navigating the program) [24]. Similarly, Spencer et al [21] found that web-based parenting programs with

XSL•FO
RenderX

additional clinical support (ie, access to a specialist or therapist in addition to the program) did not significantly influence the strength of program effects on a range of parenting and young people's outcomes compared with programs without clinical support. Corralejo and Domenech Rodrigues [23] noted that although more than half of 31 included studies on behavioral parent training programs included coaching components, no study compared program effects between conditions with and without coaching. Thongseiratch et al [30] used qualitative comparative analysis to identify specific program components associated with stronger program effects on child behavioral problems. Their analysis revealed that sending reminders to parents was the only effective feature, whereas additional phone calls were associated with weaker program effects. Notably, these reviews did not consider or explore the effect of these features on program engagement outcomes despite endorsing the utility of technology in improving accessibility.

Hansen et al [20] attempted to explore the relationship between program engagement strategies and engagement outcomes in a systematic review of randomized controlled trials (RCTs) of technology-assisted parenting programs. They found that recruitment and tailoring strategies were linked to higher postintervention *study* retention rates, but limited and inconsistent reporting of *program* adherence outcomes across studies precluded drawing conclusions about the effects of such strategies on parents' engagement with the program itself [20]. In studies on face-to-face parenting programs, reporting of data on program adherence outcomes has also been limited [8,27] or absent [28]. As program adherence measures *quality* of parental engagement with a parenting program, and *quality* of engagement is a key mechanism for changes in parenting outcomes [29], greater focus on measuring and evaluating *quality* of engagement is warranted.

Conceptualizing Engagement in Technology-Assisted Parenting Programs

Studies of engagement with technology-assisted parenting programs are largely undertaken within the behavioral science discipline, hence engagement is typically conceptualized and measured in behavioral terms (eg, use of the program as a whole or per components, known as "dose" or "adherence," respectively) [31]. However, the design and delivery of technology-assisted programs is informed by multiple disciplines, meaning that there are often differences in theory that result in highly varied conceptualizations of engagement in the literature [32]. For example, in the computer science and human-computer interaction disciplines, engagement is conceptualized in both behavioral *and* subjective terms, with subjective terms referring to experiences that emerge in the momentary interaction with the program [31]. Thus, conclusions drawn by behavioral science evaluations of engagement with technology-assisted programs may be both deepened and advanced by using behavioral *and* subjective measures of engagement.

To reduce the fragmentation of research objectives and findings between disciplines, Perski et al [31] proposed an integrated definition and conceptual framework of engagement, which maps both evidence-based and hypothesized influences of engagement based on available interdisciplinary literature. This comprehensive conceptualization of engagement may advance the behavioral science understanding of enhancing engagement. For instance, evidence on the influences of engagement may assist program developers to consider and design program features that enhance these influences and broaden the range of a program's engagement strategies. In addition to program features, this framework includes *context* in its conceptualization of engagement [31]. Prior research has suggested that a program's level of prevention (ie, universal, selective, or indicated) is an important contextual factor to consider in enhancing engagement, as the intensity of program involvement at each level (and consequently, the effort typically required of program users at each level) may require different strategies to engage users [33].

This Systematic Review

Intervening with family-level ACEs represents an important focus in efforts to reduce or prevent the risk of mental disorders among young people. Engagement represents a key mechanism by which programs yield desired improvements in target parenting outcomes [29], and technology can potentially improve engagement by overcoming common barriers. However, to date, no study has systematically synthesized engagement strategies and outcomes in technology-assisted parenting programs. Further, the behavioral science understanding of engagement strategies and measures that can be used in the design and delivery of technology-assisted programs has likely not been sufficiently conceptualized to account for the full experience of engagement [31]. The primary aim of this review is to address this knowledge gap by (1) describing the range of engagement strategies reported in the design and delivery phases of technology-assisted parenting programs targeting an ACE, (2) exploring any patterns in the use of engagement strategies based on the program's level of prevention, and (3) describing the range of behavioral and subjective engagement measures used in studies of technology-assisted parenting programs targeting a family-level ACE. The secondary aim of this review is to synthesize, where possible, the effects of specific engagement strategies on engagement outcomes and the associated target ACE outcomes.

Methods

The search strategy, inclusion criteria, primary and secondary outcomes, and proposed data synthesis methods were prespecified, registered, and published on the PROSPERO database (CRD42020209819).

Information Sources

Following consultation with academic librarians within the School of Psychological Sciences and Faculty of Information Technology at Monash University, the following 8 electronic bibliographic databases were searched: CENTRAL, CINAHL, Embase, OVID MEDLINE, OVID PsycINFO, Scopus, ACM, and IEEE Xplore.

An effective combination of search terms was designed by the first author (GA) according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement

XSL•FO
RenderX

[34] and in consultation with an academic librarian. The terms were identified based on relevant prior research and keyword search. Syntax was specific to each database. Keywords and terms used in IT databases were less specific than those used in health databases to maximize the yield of potentially eligible studies. Search terms for all databases included multiple terms for target ACE concepts (maladaptive parenting, child maltreatment, and interparental conflict), parents, interventions and programs, and technology (Multimedia Appendix 1 provides a full list).

Search Strategy

The first search was conducted on November 26, 2020. No language or date filter was applied to ensure that a diverse range of studies was retrieved. Abstracts from studies published in languages other than English were entered into Google Translate during screening to ascertain whether they were eligible for a full-text review. No such studies were eligible, hence further translational resources were not required. Although no date filter was applied during the search, a recency criterion was later applied following consultation with other authors and academics, leading to the exclusion of studies published before 2010. This decision was made given that technology-assisted interventions are more susceptible to changes over time owing to rapid advances in technology, hence research regarding these interventions' engagement capacity may be quickly outdated. This approach is consistent with the decisions made in previous reviews [26,35]. Reference lists of relevant systematic reviews identified in the search were manually searched to identify additional studies that were either overlooked or missed in the initial electronic database search. A gray literature search was also conducted to fully exploit available data, defined as targeted website browsing of relevant authorities and organizations and search engine searching (Multimedia Appendix 1 provides documented results). The flow of studies identified, screened, and excluded based on recency criterion can be found in the PRISMA diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the studies selection process.

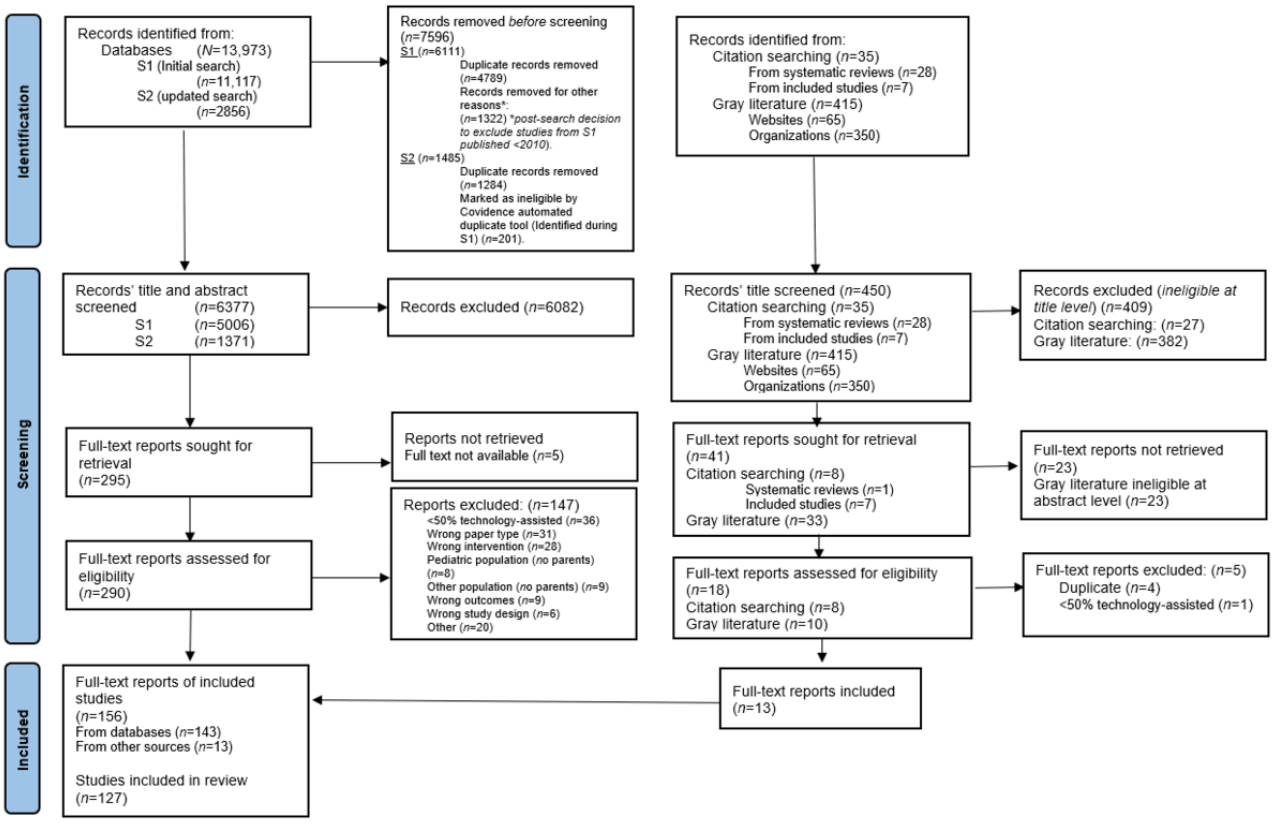

Following the initial search, it was decided among the review team to exclude 2 target ACEs (parental physical and mental illness and bullying) on the basis that they are not specifically parenting behaviors and hence less within a parent's capacity to modify. Thus, the studies returned from these 2 searches did not undergo further screening following retrieval. The PROSPERO record was updated to reflect this decision, along with the decision to apply a recency criterion and an elaboration on the proposed data synthesis methods. To ensure that the latest data were included in the review, an updated database search was conducted by the first author (GA) on November 30, 2021, to include studies published between December 2020 and November 2021. The reference lists of relevant systematic reviews identified in this search were manually searched to identify any additional studies that had been missed in the electronic database search.

Eligibility Criteria

Studies were included in the review if the study met the following criteria: (1) published in a peer-reviewed journal or publication (except for gray literature); (2) reported on a delivered intervention targeting ≥1 of the 3 predefined, modifiable ACEs (maladaptive parenting, child maltreatment, or interparental conflict); (3) >50% of the intervention was

XSL•FO
RenderX

directed at parents or caregivers of young people aged 0-18 years; (4) >50% of the intervention was delivered through technology-assisted methods or platforms; (5) described at least 1 strategy used to engage parents in the design or the delivery of the intervention; and (6) published during or after 2010. It is understandably typical for interventions targeting ACEs to self-describe from a strengths-based perspective, hence the study's background and intervention outcome measures were checked to verify whether the intervention was intended to target ACEs if otherwise unclear. Included studies that met the following additional criteria were included for answering this review's secondary aim: (1) a comparison group whose engagement was measured and compared with the experimental group (eg, treatment as usual, active control, attention control) and (2) between-group statistical analyses were conducted, with statistical significance reported. The peer-review criterion was also required for any gray literature that was otherwise eligible for inclusion in answering the secondary aim. Multimedia Appendix 2 provides further detail with regard to the inclusion and exclusion criteria.

Selection Process

Following the removal of duplicate references (using both human and automated tools via Covidence v2815 systematic review software), 2 authors (GA and CN) independently double screened the titles and abstracts of studies that were identified through the first search. One author (GA) screened the titles and abstracts identified through the second database search.

Data Collection Process

Full-text articles of potentially eligible studies identified through both searches were assessed independently by 3 authors (GA, AT, and CN), with each study assessed by 2 authors. Each author worked blinded and independently until all studies were assessed, both at the level of title and abstract and full-text screening. Discrepancies in eligibility assessment were resolved through discussion with MY. All reasons for exclusion are documented in the PRISMA diagram (Figure 1).

Extraction of key study characteristics and outcomes was completed using a standardized form on Excel (Microsoft Corporation), which was prepiloted on 10 studies (ranging in study design) by the first author (GA). All extractions were independently completed by 2 of 3 researchers (GA, AT, and CN), with the first author (GA) extracting from all studies.

Data Items: Study Characteristics and Coding of Predictors and Outcomes

The extracted study characteristics included the country of sample population; sample size; demographics (age, sex, and socioeconomic position) of the target parent or caregiver and young person; mode and function of technology-assisted intervention components; level of intervention (eg, universal, selective, or indicated prevention); intervention design frameworks; proportion and intended duration of technology-assisted component; and type of ACE outcome targeted (ie, maladaptive parenting, child maltreatment, or interparental conflict).

The extracted predictors included reported engagement strategies or program features designed to engage parents, strategies, or features not specifically reported but identified according to definitions from prior research [17,31]. Engagement strategies (predictors) were primarily identified as attributes of a technology-assisted parenting program that prior research has identified as having an evidence-based or hypothesized influence on engagement with the program [17,31,36]. An overview of these attributes and definitions is provided in Table 1.

XSL•FO
RenderX

Table 1. Overview of engagement strategy codes.

Strategies and attributes	Definition
Program design strategies	
Consultation	Processes or activities that seek to obtain verbal or conceptual input about the design or delivery of an intervention
End user consultation [a]	Consultation with users of the program, defined as parents or caregiver program participants
Stakeholder consultation	Consultation with stakeholders associated with the program, defined as service or program providers or clinicians
Academic expert consultation	Consultation with academics with expertise in the program's content
Testing	Processes or activities which allow users to engage with a prototype, and provide input about the experience of using the intervention
End user testing	Testing with users of the program, defined as parents or caregiver program participants
Stakeholder testing	Testing with stakeholders associated with the program, defined as service or program providers or clinicians
Academic expert testing	Testing with academics with expertise in the program's content
Enrollment strategies	
Targeted recruitment strategy	Efforts to increase the likelihood that the intended population or sample will be recruited[b]
Partnerships	Efforts to collaborate with and involve relevant communities or services in the recruitment of intervention users[b]
Practical support	Provision of services or materials that facilitate users to use and engage with the program (eg, loaning of technological equipment, technical assistance, provision of childcare).
Program-specific strategies	
Content- behavior change techniques	Tools, features, or strategies used in the intervention to promote behavior change
Goal-setting	Encourages the user to list and set goals relevant to the intervention's content[b]
Action plans	Supports the user to list specific behaviors or actionable strategies that they intend to complete[b]
Feedback	Allows the user to receive feedback (may be automated or via an interventionist)[b]
Self-monitoring tools	Tracks a user's performance or status to support achieving goals[b]
Content: rewards	Features that offer reward upon user's performance of a target behavior
Praise	Offers praise to the end user on any occasion[b]
Content: summaries	Features which provide the user with a summary of intervention content or activity[b]
Content: social support features	Features that facilitate the receipt of social support
Web-based discussion forums	Provides the opportunity for users to see other users use the intervention or performing the target behavior[b]
Content: supplementary resources	Features which centralize a range of resources relevant to the user or intervention[b]
Content: reminders	Features that serve to remind users about using the intervention or performing target behaviors
Delivery: mode	Intentional use of a given technology mode, thought to enhance engagement, to deliver intervention content[b]
Delivery professional support features	Features that enable remote contact with a professional
Clinical support	Provides clinical, coaching, or therapeutic support to facilitate behavioral change[b]
Nonclinical support	Provides nonclinical support to facilitate use, eg, technical, encouragement, progress monitoring[b]
Delivery: control features	Features that make users feel in control of and free to make choices about how to interact and use the intervention
All-at-once	Users receive the whole intervention at once and are not required to complete predetermined steps to progress through the intervention
Reviewability	Users have the ability to review old content[b]

XSL•FO
RenderX

Strategies and attributes	Definition
Delivery: tunneling	Features that lead users through predetermined steps to progress through intervention
Delivery: novelty	Features that provide the user with regular content updates
Delivery: ease of use	Features that support the use of the intervention to feel natural
Delivery: personalization	Features that deliver content in a way that is adapted to the user on an individual level
Delivery tailoring	Features that deliver content in a way that is adapted to factors relevant to the user's potential needs, interest, use context or other factors relevant to the user's social group
Delivery: message tone	The terminology and wording used to communicate the content's message
Credibility features	Features whose presence is designed to inculcate a feeling of trust or familiarity within the user
Narrative	The presence of a storyline to deliver intervention content or concepts
Delivery: esthetics and design	Visual features that are designed to be attractive to users
Delivery: guidance	Features that provide the user with tutorials or how-to-use guides to inculcate feelings of comfort and ease of use
Delivery: interactivity	Features that promote a two-way flow of information between the intervention and its user
Rehearsal	Features that invite the user to rehearse a behavior or content of the intervention
Challenge	Features that invite the user to test or apply their knowledge of program concepts[b]
Gamification	Features that stimulate users to compete with themselves or the program in achieving a target behavior
Reflection	Features that invite the user to reflect on the program content or their behavior[b]
Research involvement strategies	
Reminders	Features that serve to remind users about using or completing the program's research component (eg, questionnaires, or measures)
Rewards and incentives	Features that offer reward upon user's use or completion of the program's research component (eg, questionnaires, or measures)

[a]Strategy subtypes are italicized.

[b]Defined by authors.

Engagement outcomes were identified as measures of engagement. Measures were categorized by the component of engagement and defined with reference to prior research [28,31,37]. Components were also defined with reference to prior research both from the parenting intervention literature that describes behavioral components (*initial, ongoing, and quality*) [28] and the technology-assisted intervention literature that describes experiential components (*qualitative*) [31,37]. An overview of the engagement outcome measure categories and definitions, organized by components of engagement, is provided in Table 2.

Additional outcomes for studies included in the secondary analysis included between-group statistical analyses and associated *P* values for engagement outcomes and target ACE outcomes.

XSL•FO
RenderX

Table 2. Engagement outcome measures categorized by phases of program engagement or as qualitative measures.

Component of engagement and measure category	Definition
Initial	Parents' intent or actual enrollment in a program
Recruitment rates	Rates of parents recruited per recruitment site per month
Enrollment rates	Rates of parents enrolling in the program
Expressions of interest	Rates of parents expressing interest in enrolling in the program
Ongoing	Parents' behavioral engagement with the program or session or module as a whole
Ongoing frequency	Measures that provide information on how often a user visits the program
Attendance [a]	Visits indicated by attendance to a session
Log-ins	Visits indicated by log-ins to a module
Ongoing intensity	Measures that provide information on the depth of users' engagement within the program
Session or module interaction	Measures that provide information on users' interaction with interactive features within sessions or modules.
Ongoing completion rates	Measures that indicate when a user has completed the program or a session or module
Program	Measures that indicate when a user has completed the entire program
Session or module	Measures that indicate when a user has completed a session or a module in the program
Ongoing time or duration	Measures of the duration of engagement during a visit to the program or the session or module
Program	Measures of duration of engagement during the entire program
Session or module	Measures of duration of engagement during a session or a module
Retention, attrition, or dropout rates	Measures indicating users' status of engagement with the treatment, program or study
Treatment or program	Measures indicating users' status of engagement with the treatment or program itself
Study	Measures indicating users' status of engagement with the study in which the treatment or program is delivered
Quality	Parents' behavioral engagement with the programs or session or module's specific components
Intensity of specific component use	Measures that provide information on the depth of users' engagement with specific components within the program or session or module
Completion rates of specific components	Measures that indicate when a user has completed specific components within the program or a session or module
Time or duration spent in specific components	Measures of the duration of engagement during a visit to a specific component within the program
Adherence	Measures the extent to which the program is engaged with as intended or agreed by the parent or program developer
Qualitative	Parents' subjective feedback regarding their experience of engaging with the program
Satisfaction measures	Captures extent to which parent perceives the program met their needs and expectations
Feedback measures	Captures a range of parents' reactions or opinions about the program
Semistructured interviews	Captures parents' experiences and feelings in relation to using the program
Think-aloud	Captures parents' experiences of using the program in real time
Focus groups	Captures social and contextual factors in specific population subgroups that influence engagement with the program
Other	Other measures of subjective feedback

[a]Measure category subtypes are italicized.

Methodological Quality Assessment and Appraisal

The included studies were found to use qualitative, quantitative, and mixed methods for reporting on and evaluating interventions. Therefore, the Mixed Methods Appraisal Tool (MMAT; version 2018; [38]) was used to analyze the methodological quality of the included studies. The MMAT includes 2 screening items to check whether a study reports on empirical data, and 5 subsequent items (which differ depending on the category of empirical study design selected) to assess the methodological quality of empirical studies. The MMAT's 2-item screener and 5-item quality criteria were included in the prepiloted extraction template. Each included study that met

XSL•FO
RenderX

the 2-item screener was then rated according to the 5-item MMAT scoring criteria (Multimedia Appendix 2 provides the methodological quality assessment decision rules). As recommended by the MMAT, the spread of ratings was interpreted per criterion to better summarize the quality of the included studies (rather than an overall score per criterion). If an included study did not meet the 2-item screener criteria, its quality could not be assessed. However, it was still included in the narrative synthesis for answering this review's primary aim but was excluded from the quantitative synthesis for answering this review's secondary aim. Risk of bias was independently assessed by 2 researchers, with the first author (GA) assessing all studies, and 2 researchers (AT and CN) assessing alternating studies. Discrepancies in ratings were resolved through discussion between researchers.

Data Synthesis Methods

Primary Outcome

A narrative synthesis [39] was used to address the primary aim of describing the range of engagement strategies and measures used in technology-assisted parenting interventions that target ACEs and exploring any patterns to the use of engagement strategies based on the program's level of prevention. The total number of observations of a given strategy and measure was reported along with the proportion of included studies that used a given strategy or measure.

Secondary Outcome

Heterogeneity in intervention features, settings, populations, and statistical tests did not allow a meta-analysis of effect sizes to be conducted. When studies examine a common variable, but results are represented by a variety of effect magnitude measures, combined significance tests are indicated [40]. The Stouffer method of combining *P* values [41] was used to synthesize results from studies eligible for the secondary outcome analysis. Where there were ≥2 independent observations of the association between the same engagement strategy and engagement outcome category in the included studies, the Stouffer method was applied to test the combined significance of this association. (all subcategories for each engagement strategy were included in pairs, as all subcategories fall under the same definition). Stouffer *z* was also applied to test the combined significance of associations between engagement outcomes and target ACE outcomes, where ≥2 independent observations of this association were identified. Stouffer *z* was calculated by dividing the sum of the $z(p_i)$ values by the square root of *k*, where *k* is the number of associations per pairing of the engagement strategy and outcome. If the resulting *P* value corresponded to a probability level <.01, the null hypothesis of no effect was rejected.

Before analysis, the engagement outcomes were coded into discrete types and measures for consistency (Table 2). Raw *P* values were converted to 1-tailed values before analysis to test the directional hypothesis that engagement strategies significantly increase engagement. Direction of effect was assumed toward the experimental group, given the rationale provided in all studies for testing engagement strategies, based on theory or evidence that they might have an effect.

Results

Study Selection

Figure 1 provides the flow of the systematic search process. A total of 13,973 records were identified by searching electronic databases, which were reduced to 6377 records after removal of duplicates and the postsearch decision to exclude studies published before 2010. Of the 6377 records whose titles and abstracts were screened, 290 (4.55%) full texts were assessed for eligibility and 147 (2.3%) were excluded (Figure 1 provides the reasons for exclusion). Manual searching of the reference lists of both relevant systematic reviews identified in the database search and included studies revealed 8 additional records that met the study criteria. A gray literature search was conducted (Multimedia Appendix 1 provides the search strategy and results); 415 records were identified, 33 records were retrieved and screened, and 10 full-text records were assessed for eligibility. Five reports were excluded (Figure 1), leaving 5 additional reports that met the study criteria. A total of 156 records were included in the review, comprising 127 separate studies (Multimedia Appendix 3 indicates which records were merged). The included studies are summarized in the *Characteristics of Included Studies* section using narrative synthesis methods for the primary outcome and Stouffer *P* analysis for the secondary outcome. Multimedia Appendix 4 details each study's characteristics, engagement strategies, and measures used.

Characteristics of the Included Studies

Study Designs

Most of the included studies were RCTs (77/127, 60.6%), followed by nonrandomized studies that estimated the effectiveness of the program (31/127, 24.4%). Other included studies had a descriptive design (5/127, 3.9%), mixed methods design (6/127, 4.7%), and qualitative design (3/127, 2.4%).

Participants

Studies were conducted across 16 different countries, although most were conducted in the United States (69/127, 54.3%), followed by Australia (33/127, 26%). Programs most often catered to parents of young people with a mean age between 5 and 12 years (31/127, 24.4%), although many studies (51/127, 40.2%) did not report the age of parent or caregiver's young person. Mothers or female caregivers represented most (ie, >80%) of the sample in just more than half (69/127, 50.4%) of the included studies, whereas fathers or male caregivers comprised most of the sample in far fewer studies (8/127, 5.5%). A very small percentage of the studies reported an even spread of male and female caregivers (6/127, 4.7%). Just more than one-third (42/127, 33.1%) of the samples in the identified studies were reported as taking place in the context of socioeconomic difficulty or vulnerability, and approximately half (61/127, 48%) did not specifically report or state participants' socioeconomic position (where possible, participant characteristics were extracted only for study participants with access to the technology-assisted intervention).

XSL•FO
RenderX

Programs

Most programs (97/127, 76.4%) did not report a design framework by which the program was designed or developed, although a small number reported user-centered design approaches (9/127, 7.1%). The most common ACE targeted by the programs was maladaptive parenting behaviors (98/127, 77.2%), followed by interparental conflict (19/127, 15%), and child maltreatment (10/127, 7.9%). A small percentage of the studies (6/127, 4.7%) reported targeting more than one of the target ACEs. Programs were almost equal either at the selective (48/127, 37.8%) or indicated (47/127, 37%) level of prevention, with universal programs being less common (32/127, 35.2%). The programs were primarily delivered in the participants' home (84/127, 66.1%), and approximately a quarter of the programs were delivered at home through a health (20/127, 15.8%) or community service (11/127, 8.7%). One-fifth (26/127, 20.5%) of the programs involved the young person of the participating parent or caregiver. The programs' technology most commonly functioned to facilitate self-directed (ie, asynchronous) learning (67/127, 52.8%), with one-third (39/127, 30.7%) of the programs combining remote clinician contact with self-directed learning (ie, synchronous). A few programs included technologies that functioned to enhance existing services (5/127, 3.9%). Most programs comprised one (54/127, 42.5%) or two (42/127, 33.1%) modes of technology in the delivery of the program, with web-based modules being the most common mode (76/127, 59.8%). Videos, videoconferencing, emails, telephone calls, and text or application messaging were also commonly used. Table 3 provides a detailed breakdown of the participants and program characteristics of the included studies.

XSL•FO
RenderX

Table 3. Characteristics of included studies (N=127).

Characteristics	Studies, n (%)
Participant	
Country	
Australia	33 (26)
Canada	2 (1.6)
Finland	2 (1.6)
New Zealand	4 (3.1)
The Netherlands	3 (2.4)
Sweden	2 (1.6)
United Kingdom	3 (2.4)
United States	69 (54.3)
Other	9 (7.1)
Mean age of young person at recruitment	
Infant age (0-12 mo)	4 (3.1)
Toddler (1-3 y)	4 (3.1)
Preschool age (>3-5 y)	22 (17.3)
Primary School age (>5-12 y)	31 (24.4)
Adolescence (>12-18 y)	15 (11.8)
Not reported	51 (40.1)
Age range reported	24 (19.3)
Not applicable due to study design	5 (3.9)
Parent sex	
>80% female	69 (54.3)
<20% female (100% male)	8 (6.3)
50% male and female	6 (4.7)
20%-79% female	18 (14.2)
N/A[a] or not reported	26 (20.5)
Socioeconomic position	
Low socioeconomic position or vulnerability	42 (33.1)
Not reported	61 (48)
Program	
Level of prevention	
Universal	32 (25.2)
Selective	48 (37.8)
Indicated	47 (37)
Primary target ACEs[b]	
Maladaptive parenting	98 (77.2)
Maladaptive parenting+secondary ACE	6 (4.7)
Child maltreatment	10 (7.9)
Interparental conflict	19 (15)
Target recipient	
Parents	96 (75.6)

XSL•FO
RenderX

Characteristics	Studies, n (%)
Parents (adoptive)	5 (3.9)
Parents+young person	26 (20.5)
Setting	
Home	84 (66.1)
Home, via health or hospital	20 (15.7)
Home, via community or social service	11 (8.7)
Health service or hospital	3 (2.4)
Community setting	3 (2.4)
Participants' choice	5 (3.9)
Other	1 (0.8)
Function of technology	
Self-directed learning	67 (52.7)
Self-directed learning+remote clinician contact	39 (30.7)
Service enhancement	5 (3.9)
Peer support	0 (0)
Remote clinician contact	7 (5.5)
Remote clinician contact+peer support	1 (0.8)
Self-directed learning+peer support	3 (2.4)
Self-directed learning+remote clinician contact+peer support	1 (0.8)
Other	4 (3.1)
Types of modes	
Web-based modules	76 (59.8)
Website	8 (6.3)
Computer, tablet, or phone apps	3 (2.4)
Videos	25 (19.7)
Videoconferencing	24 (18.9)
emails	25 (19.7)
Telephone calls	30 (23.6)
Text or application messaging	24 (18.9)
Podcasts	4 (3.1)
Digital feedback	4 (3.1)
Social media	13 (10.2)
Multimodal; participant chooses mode	3 (2.4)
Total number of technology modes	
1	54 (42.5)
2	42 (33.1)
3	21 (16.5)
4	9 (7.1)
5	1 (0.8)
Design framework	
Community-based involvement	2 (1.6)
User-centered design	9 (7.1)
User-involved design	2 (1.6)

XSL•FO
RenderX

Characteristics	Studies, n (%)
Iterative approach	7 (5.5)
Other	9 (7.1)
Not reported	97 (76.4)

[a]N/A: not applicable.

[b]ACE: adverse childhood experience.

Quality Assessment of Included Studies

All studies were assessed for quality according to the MMAT, except for 4 studies [42-45] that did not meet the MMAT screening criteria. As results from the MMAT are best understood via ratings of each study design's criterion rather than calculating an overall score per study [38], a brief summary of the key results is provided (Multimedia Appendix 5 provides a full summary of results and results from each included study). All the criteria were met in 10% (8/77) of the included RCTs. The most commonly unmet criteria for RCT studies were insufficient outcome data (28/77, 36% of RCT studies) and insufficient adherence to assigned interventions (24/77, 31% of RCT studies). All criteria were met in 19% (6/31 of the included quantitative nonrandomized studies, with complete outcome data again being the most commonly unmet criterion, although interventions were more commonly administered as planned compared with quantitative RCT studies. Quantitative descriptive studies mostly included appropriate sampling strategies and statistical analyses for the research question, although criteria regarding sample representativeness and the risk of nonresponse bias were commonly unmet. Mixed methods studies overall adhered to the quality criteria for each method involved, but criteria regarding integration between each method's results and divergences or inconsistencies were less commonly met. All qualitative studies met all the criteria.

Primary Outcome

Engagement Strategies Used in the Design and Delivery of Programs in the Included Studies

Strategies used to influence engagement during the program's design phase of the intervention cycle were reported in 43 studies [42-44,46-85]. The total number of strategies used per study ranged from 1 to 6 (mean 2.0). End user consultation was the most commonly used strategy (26/43, 60%), followed by user testing, stakeholder consultation, and expert consultation. Multimedia Appendix 6 provides a detailed overview of the types of strategies identified in the design phase, number of observations of each strategy, and number of studies reporting the use of each strategy (definitions for "Type of strategy" are provided in Table 1).

Strategies and program features designed to influence engagement during the delivery phase of the intervention cycle were reported in 123 studies [42,45-76,78-167]. The total number of strategies used per study ranged from 1 to 16 (mean 6.2). Interactive program features were the most commonly used strategy (85/123, 73%), followed by videos or animations providing guidance, user control features, professional support features, and behavior change techniques. Multimedia Appendix 6 provides a detailed overview of the types of strategies identified in the delivery phase, the total observations of each strategy, and the total studies reporting the use of each strategy (definitions for "Type of strategy" are provided in Table 1).

Overview of Measures Used in the Included Studies

Engagement was measured in 111 of the included studies [42,46,48-55,57-60,62-76,78-98,100-115,117,118,121-125,127-146,148-159,162-167]. The most common component of behavioral engagement measured was initial engagement (83/111, 74.8%), followed by ongoing engagement (78/111, 70.3%) and quality of engagement (42/111, 37.8%). Moreover, 70.3% (78/111) of the studies used qualitative measures of engagement, which was comparable with the number of studies that used behavioral measures of ongoing engagement. Enrollment and recruitment rates were the most commonly reported measures of initial engagement, with a small number reporting expressions of interest. The most commonly used measure of ongoing engagement was session or module completion rates, followed by study retention, attrition, or dropout rates. The most used measure of quality of engagement was use of specific program components, followed by completion and time spent on specific program components. Satisfaction ratings or questionnaires were the most used measures of qualitative engagement, followed by feedback measures. Multimedia Appendix 6 provides a detailed overview of measures used in the included studies (definitions for "Component of engagement, Measure" are provided in Table 2).

Secondary Outcome

Engagement outcomes between groups were statistically compared in 22 studies, in which 1 group received unique engagement strategies. Most studies in this subgroup were RCTs (15/22, 68%) that evaluated a program targeting maladaptive parenting (19/22, 86%) at the indicated level (10/22, 45%), involving parents only (15/22, 68%) in a home setting (14/22, 63%). The mean number of engagement strategies used was highest at the selective level of prevention (mean 10), followed by the indicated level (mean 7), and universal level (mean 7). No study evaluated measures related to initial or quality components of engagement, hence the following synthesis explores the effects of engagement strategies on the ongoing and qualitative components of engagement. Multimedia Appendix 4 details each study's between-group engagement outcomes. Of these studies, 18 (82%) reported program engagement outcomes with *P* value data and were therefore appropriate for the analysis [49,73,75,86,87,94,97,98,100-102,122,124,125,142,149,152].

XSL•FO
RenderX

Stouffer *P* analyses indicated that ongoing and qualitative engagement outcomes were positively and reliably associated with both user and stakeholder involvement (consultation and testing) in the program's design, web-based formats (compared with face-to-face equivalents), provision of practical support to use the technology, personalization or tailoring program features, control features, and use of engagement strategies during the program's design phase. Interactive program features were also reliably associated with ongoing outcomes of engagement, but not with qualitative outcomes of engagement. Clinical professional support features, videos, behavior change techniques, and reminders were not reliably associated with ongoing and qualitative outcomes of engagement. Table 4 provides a complete overview of the results.

Only 2 studies included in this subgroup statistically analyzed the relationships between engagement outcomes in the experimental group and target ACE outcomes [87,149]. Stouffer *P* analysis indicated that session completion (ongoing engagement) was positively and reliably associated with improvements in maladaptive parenting behaviors ($P<.01$). One study analyzed the relationship between parent-rated therapeutic alliance (qualitative engagement) and changes in child outcomes, but no significant association was found [142].

Among the 22 studies eligible for inclusion in the secondary analysis, 3 (14%) studies [49,87,102] reported significant positive effects of the experimental group (where unique engagement strategies were used) on changes in maladaptive parenting compared with control groups where such engagement strategies were not used. Moreover, 3 (14%) studies [73,94,149] reported nonsignificant group effects on maladaptive parenting and 5 (23%) studies reported no difference between groups on changes in maladaptive parenting [72,84,105,124] and interparental conflict [159]. One study found a significant negative effect of the experimental group (where unique engagement strategies were used) on changes in interparental conflict compared with a control group where strategies were not used [152]. Another study identified a negative effect of both experimental and control groups on interparental conflict, although this reduction was significantly lower for the experimental group than in the control group [124]. Five studies assessed engagement outcomes only and did not include a measure to assess changes in target ACE outcomes in the parenting domain [66,75,100,101,122] or included a measure that was not aligned with this study's definition of ACEs [97]. Multimedia Appendix 4 details each study's program effects on ACE outcomes.

XSL•FO
RenderX

Table 4. Findings from Stouffer P calculations between engagement strategies and measures of engagement.

Engagement strategy	Measure of engagement	
	Ongoing	Qualitative
Design: user consultation and testing		
n of studies	3	0
n of associations in Stouffer *P* value	3	0
Stouffer *P* value	<.001[a]	N/A[b]
Design: stakeholder consultation and testing		
n of studies	0	2
n of associations in Stouffer *P* value	0	2
Stouffer *P* value	N/A	<.001[a]
Delivery mode: web based		
n of studies	6	9
n of associations in Stouffer *P* value	7	10
Stouffer *P* value	.001[a]	<.001[a]
Practical support		
n of studies	5	6
n of associations in Stouffer *P* value	5	6
Stouffer *P* value	<.001[a]	<.001[a]
Delivery: interactivity		
n of studies	4	5
n of associations in Stouffer *P* value	4	6
Stouffer *P* value	<.001[a]	.015
Delivery: professional support features (clinical)		
n of studies	4	0
n of associations in Stouffer *P* value	4	0
Stouffer *P* value	.02	N/A
Delivery: guidance (videos)		
n of studies	3	3
n of associations in Stouffer *P* value	3	4
Stouffer *P* value	.04	.04
Content: behavior change techniques		
n of studies	3	4
n of associations in Stouffer *P* value	3	4
Stouffer *P* value	.08	.02
Delivery: personalization or tailoring		
n of studies	3	2
n of associations in Stouffer *P* value	3	2
Stouffer *P* value	<.001[a]	<.001[a]
Content: reminders		
n of studies	3	3
n of associations in Stouffer *P* value	3	3
Stouffer *P* value	.08	.03

XSL•FO
RenderX

Engagement strategy	Measure of engagement	
	Ongoing	Qualitative
Delivery: control features		
n of studies	2	4
n of associations in Stouffer *P* value	2	4
Stouffer *P* value	.01[a]	<.001[a]

[a]Indicates significance (*P*<.01).

[b]N/A: not applicable.

Discussion

Principal Findings

This review aimed to systematically describe the range of engagement strategies and measures used in the design and delivery of technology-assisted parenting programs targeting ACEs related to modifiable parental behavior. A secondary aim of this review was to synthesize the findings from studies that examined the effects of using specific engagement strategies on engagement outcomes and explore any patterns in the associations between engagement outcomes and ACE outcomes. The following discussion synthesizes the primary and secondary outcome findings for the most commonly identified engagement strategies and other strategies included in the secondary outcome analysis. The available data did not permit the synthesis of engagement strategies and the initial and quality of engagement outcomes in the secondary outcome analysis, hence this discussion explores patterns of associations between engagement strategies and ongoing and qualitative engagement outcomes.

Summary of Evidence

Use of Engagement Strategies in Technology-Assisted Programs Targeting Parent-Behavior ACEs

Design Strategies

Just more than one-third of the studies reported using a strategy in the program's design phase to enhance engagement during the program's delivery. This finding is comparable with the results from the systematic review of technology-assisted parenting programs by Hansen et al [20], which found that just more than one-third of the RCTs reported strategies to enhance engagement during the program's design phase. In addition, this review found that approximately 1-2 strategies were used on average per study, with program user consultation and testing being among the most commonly reported strategies. Programs that involved users and stakeholders in their design were more likely to better engage users in their ongoing use and subjective experience of the program, respectively, compared with programs developed without these strategies. This finding supports the central claim of user-centered design approaches in that incorporating users' unique experiences and knowledge into the design of a program is likely to increase acceptability and relevance to other parents during the program's delivery [168].

Delivery Strategies

This review found that the use of program-specific engagement strategies or engaging program features reported in the delivery of a program was significantly more common than the use of design strategies, with studies using approximately 7 strategies on average. Although previous reviews exploring the use of engagement strategies in technology-assisted programs for parenting [20] and mental health [169] did not report the number of strategies used per study, it appears that the average number of strategies identified in this review was higher than the number of strategies identified in previous reviews. The review by Hansen et al [20] defined engagement strategies a posteriori, whereas the review by Saleem et al [169] defined engagement strategies a priori using the same conceptual framework as this study. This review used a conceptual framework to identify engagement strategies a priori as well as to identify and categorize additional strategies a posteriori. This approach allowed identification of strategies reported by the study's authors and strategies that were not reported but were consistent with the conceptual framework, which may explain the higher number of strategies identified in the studies included in this review.

Interactivity was the most commonly used strategy, often in the form of challenges posed by the program to the user such as multiple-choice quizzes, check-in questions, or problem-solving exercises. Programs with interactive features were more likely to better engage users in their ongoing use (indicated by greater attendance, lower treatment dropout, greater module log-ins, and completion) compared with programs without these features. However, users' subjective experience of programs with these features was not significantly different from the experience of users in a program without these features. Perski et al [31] identified interactivity as primarily having a hypothesized influence on engagement. Findings from this review provide preliminary support for the reliability of the influence that interactivity has on a user's ongoing engagement in a technology-assisted parenting program.

Guidance provided by videos in the form of roleplays or vignettes was often used for demonstrating or modeling skills to users. Videos were not, however, more likely to better engage users in their ongoing use or subjective experience of a program compared with programs without videos. Interestingly, a reliable positive association between ongoing program engagement and target ACE outcomes (maladaptive parenting) was found in 2 programs [87,150] that included guidance provided by videos. According to social cognitive theories, demonstration and

XSL•FO
RenderX

modeling are the key mechanisms by which learning occurs [170]. Therefore, it is likely that video guidance might be more relevant to measures of the *quality* of engagement (ie, what parents specifically invest in and receive from the program), which in turn may form a key mechanism for change [28,29]. No study with video guidance measured parents' *quality* of engagement, hence this hypothesis could not be explored.

Control features frequently referred to features that allowed the parent to choose how they engage with the program, such as self-selection of modules or topics, and the ability to self-pace (which often took the form of delivering the content "all-at-once" or in a way that permitted end users to review content). Control features such as self-pacing and ability to review content were found overall to better engage users in their ongoing use of a program compared with programs without these features. However, this finding should be interpreted with caution as there were only 2 studies included in this analysis, one of which reported that users of a program with control features were *less* likely to engage compared with users of a program without such features [124]. Control features were also more likely to improve users' subjective experience of a program. This finding is consistent with recent research with parents from a low socioeconomic background, which suggests that convenience and flexibility were key preferences for engaging with technology-assisted programs [171,172]. These results suggest that greater freedom in choosing how to engage with the program may enhance parents' perceived benefit and satisfaction.

Professional support features were most frequently clinical in nature, such as coaching or therapist support, or professional facilitation of other clinical interventions (eg, groups or forums). The presence of professional support features was not found to better engage users with a program compared with programs without the presence of a professional. This finding does not support findings from similar reviews on digital health or mental health programs for adults [31,169], both of which found that professional support features or coaching had positive influences on ongoing measures of engagement (eg, number of log-ins and time spent on the internet). One key difference between these reviews and the current review is that programs in the current review targeted parents, however parents from low socioeconomic backgrounds have previously identified professional support features as a potentially engaging feature for technology-assisted programs [171]. Interestingly, one study in this analysis that reported a positive link between professional support and engagement also involved user consultation as a design strategy. It is possible that consulting parents about the type of professional support they would like to receive may positively influence ongoing engagement. Further, data from the included studies in this review did not permit the assessment of *professional support features'* effects on the *quality* of parents' engagement (such as adherence) or their *qualitative* engagement (such as therapeutic alliance), hence it remains to be seen whether *professional support features* may influence parents' engagement in other ways. Closer inspection of the studies included in this review revealed that parents who used a program with professional support features were not more likely to improve target ACE outcomes compared with parents who did not receive professional support features (Multimedia Appendix 4). This finding is consistent with previous reviews of technology-assisted parenting programs, which indicated that additional support did not make programs more effective in changing young people and parent outcomes than those without such support [21,24,30]. Spencer et al [21] argued that this suggests programs both with and without professional support may be beneficial, whereas Florean et al [24] argued that given prior research has highlighted the importance of professional support, and the number of studies in their meta-analysis was small, the lack of effect they identified should be interpreted with caution. Overall, further research is required to clarify whether professional support features can enhance both engagement with technology-assisted parenting programs and parenting outcomes. This may be achieved by designing professional support features that suit parents' contexts based on parent user consultation and measuring the effects of *professional support features* on the *quality* of parents' engagement or their *qualitative* engagement. These measures are more closely related both to what parents specifically invest in and receive from a program and to their experience of receiving professional support, respectively.

Behavior change techniques most frequently took the form of feedback (eg, as a tailored report or in response to interactive challenges) and goal setting. Including behavior change techniques in programs was not found to better engage users in their ongoing use and subjective experience of a program. However, the effects of behavior change techniques may be more relevant to target ACE outcomes than engagement outcomes, given that they are specifically designed to promote change in target parenting behavior. Inspection of the studies included in this analysis (n=3) found that all programs with behavior change techniques were more likely to improve target ACE outcomes compared with programs without behavior change techniques [47,122,147].

Delivery mode being web based, compared with a face-to-face mode of delivery of the same program, was more likely to engage parents in their ongoing use and subjective experience of a program. Up to 10 studies were included in this analysis, which enhanced the robustness of this finding. This suggests that technology-assisted programs' ability to overcome a range of barriers for attending and completing sessions (ongoing engagement) is likely to enhance parents' perceived acceptability and satisfaction (qualitative engagement) with the program. Although half of the studies included in this analysis did not measure ACE outcomes, the studies that did generally find no differences between web-based programs compared with programs delivered face-to-face. To our knowledge, this is the first review to directly compare both engagement and ACE outcomes between web-based programs versus programs delivered face-to-face. Our findings suggest that web-based programs were as effective as face-to-face programs in improving ACE outcomes related to modifiable parenting behavior; however, engaging in a web-based program may have been a more positive experience than attending face-to-face.

Personalization or tailoring strategies often included content recommendations based on user data, or artifacts that permitted the user to personalize their experience (eg, scrapbooks or

XSL•FO
RenderX

journals). The review found that these strategies were more likely to better engage parents, consistent with prior research [31,169], providing further support for personalizing or tailoring parenting programs. Although *practical support* for using the technology-assisted components appears to have generally been included for research purposes, this review found that programs involving *practical support* were more likely to engage parents, indicating that this may be an important strategy to enhance continued engagement with a parenting program. For example, one low-cost strategy to provide *practical support* was to provide instructions, user manuals, or orientation sessions on using the technology.

Finally, this review found that the use of reminders was not more likely to engage with parents' ongoing use or subjective experience of a program. Previous reviews have found mixed support for the use of reminders in engaging users [173,174]. This review found that *control features* and *personalization or tailoring* were more likely to improve engagement, hence offering parents control over how they receive reminders and personalizing or tailoring reminders may increase their relevance and enhance engagement.

Use of Engagement Strategies Across Levels of Prevention

Programs at the level of selective prevention were found to use a higher number of engagement strategies, whereas the use of engagement strategies was roughly equal between programs at the level of indicated and universal prevention. Prevention strategies are most effective when they account for variation in families' needs, as well as families and services' available resources and capacities [33]. Given programs at the indicated prevention level are more intensive, they may require more intense effort and greater use of strategies to promote parental engagement. All programs at the indicated level included professional support features, whereas professional support features were much less frequently included for programs at the selective level and not at all for programs at the universal level. This association may be because fewer strategies were needed to deliver programs at the indicated level owing to the presence of a clinician facilitating and supporting engagement, whereas programs delivered at the selective level made use of other features to facilitate and support parents whose young people are identified as "at risk" in self-directed learning and engagement. On the basis of the findings from this review, parents' ongoing engagement with self-directed learning programs aimed at selective or universal prevention may be enhanced by using personalization or tailoring strategies, control features, and interactivity.

Measures

This review found that most studies in this review reported measuring initial engagement in a program, which was likely because most studies were RCTs where reporting on measures of initial engagement or recruitment rates is required [175]. Technology-specific engagement measures, such as frequency of use, time or duration spent in the technology-assisted component, and interaction with components (eg, modules viewed and links clicked), were relatively low. However, studies including such measures tended to include >1 measure, which may in fact provide greater insight into engagement compared with focusing on one measure or domain [37]. Qualitative measures of engagement were most often in the form of satisfaction and feedback measures, which are efficient methods for understanding the perceived acceptability and usefulness of programs, evaluating the program's acceptability and usefulness from a user perspective, and informing continued program development and refinement [37]. A very small number of studies (n=16) used measures to explore parents' subjective experience, such as semistructured interviews, focus groups or "think aloud" techniques. Although such measures may require greater time resources, they are likely to support users to reflect on how the program affected changes in target domains. Such data may assist researchers and program developers to better understand the features or components that are specifically related to behavior change. The number of studies that measured the quality of engagement was approximately half the number of studies that measured other domains of engagement. Given that the quality of parents' engagement is suggested to be a key mechanism for positive parenting change [76], more frequent evaluation of engagement quality is required to extend our understanding of engagement and associated change in target outcomes in technology-assisted programs.

Strengths and Limitations of This Review

To our knowledge, this is the first systematic review of technology-assisted parenting programs to comprehensively assess and describe the use of engagement strategies and measurement of engagement outcomes. Unlike previous reviews, which have identified specific strategies to assess a priori, this review did not place any restrictions on the type of program engagement strategy or outcome, and furthermore, used a highly inclusive approach to study selection. This permitted identification of a broad range of technology-assisted parenting programs, and subsequently a broad range of strategies and measures were identified for synthesis. Findings from this review were also reported using an existing conceptual framework, as use of a shared terminology can support both advancements in understanding and integration of knowledge across disciplines. Although a meta-analysis was not possible in this review, the range of strategies and outcomes identified through an inclusive approach meant that it was possible to perform a quantitative synthesis of associations between engagement strategies and engagement outcomes. This extends the current evidence, which to date has been narratively synthesized [20]. Overall, the methodology in this review permitted a comprehensive description and preliminary quantitative assessment of engagement strategies used in technology-assisted parenting programs. This in turn is a step toward responding to calls in the wider digital health literature for assessment of all available engagement strategies to generate more robust evidence, as well as contributing to a reduction in conceptual and empirical fragmentation in digital health research [31,32].

Several limitations of the current review should be noted. First, although the use of existing conceptual frameworks [28,31,37] supported consistent definitions of identified engagement strategies and measures, some strategies and measures required new definitions based on available information from included

XSL•FO
RenderX

studies. Such information was not consistently clear or adequate, hence these definitions and results associated with these strategies and measures should be interpreted with this caveat in mind. Second, available data for associations between engagement strategies and engagement outcomes were heterogeneous both in program delivery and statistical analysis. This precluded meta-analysis to estimate effect sizes, so the Stouffer *P* method was used to estimate the reliability of these associations. This method is argued to be limited by its inability to weigh studies according to their sample size, although some authors argue that *P* values are already weighted, as the *P* value itself depends on the sample size for which it is calculated [176]. However, available data for associations between engagement strategies and engagement outcomes were limited. The resulting small sample size of the studies included in the Stouffer *P* analyses should be considered when interpreting the reliability of these associations. Another critique of *P* values is that they are less clinically relevant than other measures of statistical inference such as effect sizes and CIs. However, given that engagement is not specifically a clinical outcome, we considered use of *P* values to demarcate significant change as appropriate in this context. Third, although the quality assessment of the included studies was primarily to provide a summary of the quality of the available literature, the overall quality of the studies included in the quantitative analysis was low. Most notably, just more than half of the RCTs (42/77, 55%) reported sufficient complete outcome data, so reported *P* values for engagement outcomes may have been biased by completer versus noncompleter characteristics. Fourth, fewer than half (32/77, 42%) of the study participants sufficiently adhered to the assigned program. The cutoff value applied for acceptable adherence to the program in this review was based on findings from a previous review that examined adherence to technology-assisted parenting programs [20], however this cutoff may be above the norm for technology-assisted parenting programs (see the *Recommendations for Future Research* section for further elaboration). Finally, samples of parents in the included studies were overwhelmingly mothers or female participants, underscoring the well-recognized need for better representing fathers or male caregivers in technology-assisted parenting program development and evaluation [177].

Recommendations for Future Research

This review identified a broad range of engagement strategies that future research can draw on in the design and delivery of technology-assisted parenting interventions. The findings from this review also suggest that consulting and testing program components with parents in the design phase of a program may lead to better engagement outcomes. It is recommended that future research meaningfully involves parents in the program's design to more effectively identify strategies that will be perceived as acceptable and useful in the delivery phase of a program, and hence enhance engagement. Furthermore, the process of engaging, incorporating, and triangulating multiple stakeholder perspectives may uncover assumptions about engagement within the literature and potentially advance the understanding of why technology-assisted parenting programs are frequently undermined by poor engagement.

Heterogeneity in measuring and evaluating engagement has been cited as precluding meta-analysis of engagement outcomes in technology-assisted parenting programs [20]. This review integrated measurement concepts from the digital intervention and parenting intervention literature to synthesize heterogeneous measures in a conceptually meaningful way, which allowed a preliminary quantitative synthesis. To continue reducing heterogeneity in measuring and evaluating engagement, future studies of technology-assisted parenting programs may consider these integrated measurement concepts in selecting a range of measures, at various stages of engagement, appropriate to a given research question. Future studies should also consider selecting nonbehavioral measures of engagement to complement behavioral measures to sufficiently capture the full experience of engaging in a technology-assisted parenting program.

Of the 87 studies whose research design involved a treatment or experimental group and a comparison group, only 22 (25%) performed statistical comparisons of parents' engagement outcomes. Low proportions have been reported in other reviews of engagement in digital interventions in the mental health field [169]. To respond to calls for a better understanding of *how* technology-assisted parenting programs work, future research should consider using experimental study designs and statistical between-groups comparisons of both intervention *and* engagement outcomes and report standardized effect sizes to facilitate meta-analysis of engagement outcomes. Such data can extend our knowledge of the effects of specific engagement strategies on engagement outcomes, as well as better understand the associations between engagement and intervention outcomes. Yardley et al [32] argued that promoting "more engagement" in technology-assisted programs may not always be associated with positive intervention outcomes, as greater demands on a user to engage with an intervention may lead to user burden and fatigue [31]. Current standards for assessing sufficient engagement on outcomes such as attrition or adherence may therefore be too high. Therefore, this knowledge may also contribute to better understanding of what an optimal dose or "effective engagement" [32] looks like in technology-assisted parenting programs. Better understanding of sufficient engagement with a program to accomplish desired effects may have implications for how these concepts are evaluated in future studies.

Summary and Conclusions

This review describes and appraises the range of engagement strategies and measures used in the design and delivery of technology-assisted parenting programs targeting ACEs that are within parents' capacity to modify. Preliminary evidence was found for including involvement of users and stakeholders in the program's design, personalization or tailoring features, control features, and provision of practical support for enhancing ongoing and qualitative outcomes of engagement. Preliminary evidence was also found for the notion that web-based parenting programs are effective in promoting ongoing engagement, which in turn may enhance overall satisfaction. Engagement strategies that were not found to enhance ongoing or qualitative engagement outcomes (ie, professional support features, videos, and behavior change techniques) may be related to the quality of parents' engagement in a program. Using a broad range of

XSL•FO
RenderX

engagement measures to sufficiently capture parents' experience of engagement in a technology-assisted parenting program and statistically comparing engagement outcomes between groups receiving different programs to facilitate meta-analysis can advance current knowledge on the potential effects of specific strategies on engagement outcomes. However, such knowledge should serve to complement knowledge about user context when designing technology-assisted parenting programs [32,168]. There is much yet to learn about the relationship between engagement in technology-assisted parenting programs and change in target ACE outcomes, but efforts to better understand this potential mechanism for change hold significant implications for preventing or reducing the impact of ACEs at the family level and associated mental health outcomes of young people.

Acknowledgments

This project is affiliated with The Centre of Research Excellence in Childhood Adversity and Mental Health, a 5-year research program (2019-2023) cofunded by the National Health and Medical Research Council (NHMRC) and Beyond Blue. GA was funded by the Research Training Program Scholarship (Monash University). GA has also been awarded a top-up scholarship from the NHMRC Centres of Research Excellence Grant APP1153419 titled "Centre of Research Excellence in Childhood Adversity and Associated Depression and Anxiety." AJ was supported by NHMRC Investigator Grant 1172889.

The first author would like to acknowledge the contribution of specialist librarians Anne Young and Mario Sos from Monash University for their input in developing the search strategy and Dr Sarah Khor for her input in supporting the secondary data analysis of this review.

Authors' Contributions

GA was involved in conceptualization, data curation, formal analysis, investigation, preparation of methodology, project administration, visualization, and writing the original draft. AT and CN were involved in the investigation and writing, reviewing, and editing the manuscript. AR and AJ were involved in writing, reviewing, and editing the manuscript. MBHY was involved in conceptualization, methodology, supervision, and writing, reviewing, and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File , 64 KB - jmir_v26i1e43994_app1.docx]

Multimedia Appendix 2
Decision rules.
[DOCX File , 23 KB - jmir_v26i1e43994_app2.docx]

Multimedia Appendix 3
Additional references merged with main references.
[DOCX File , 22 KB - jmir_v26i1e43994_app3.docx]

Multimedia Appendix 4
Results of individual studies.
[DOCX File , 298 KB - jmir_v26i1e43994_app4.docx]

Multimedia Appendix 5
Results of quality assessment.
[DOCX File , 49 KB - jmir_v26i1e43994_app5.docx]

Multimedia Appendix 6
Primary outcome tables.
[DOCX File , 22 KB - jmir_v26i1e43994_app6.docx]

Multimedia Appendix 7
PRISMA checklist.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 70 KB - jmir_v26i1e43994_app7.pdf]

References

1. Erskine HE, Moffitt TE, Copeland WE, Costello EJ, Ferrari AJ, Patton G, et al. A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. Psychol Med 2015 May;45(7):1551-1563 [FREE Full text] [doi: 10.1017/S0033291714002888] [Medline: 25534496]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. Yap MB, Jorm AF. Parental factors associated with childhood anxiety, depression, and internalizing problems: a systematic review and meta-analysis. J Affect Disord 2015 Apr 01;175:424-440. [doi: 10.1016/j.jad.2015.01.050] [Medline: 25679197]
4. Yap MB, Pilkington PD, Ryan SM, Jorm AF. Parental factors associated with depression and anxiety in young people: a systematic review and meta-analysis. J Affect Disord 2014 Mar;156:8-23. [doi: 10.1016/j.jad.2013.11.007] [Medline: 24308895]
5. Sandler IN, Schoenfelder EN, Wolchik SA, MacKinnon DP. Long-term impact of prevention programs to promote effective parenting: lasting effects but uncertain processes. Annu Rev Psychol 2011;62:299-329 [FREE Full text] [doi: 10.1146/annurev.psych.121208.131619] [Medline: 20822438]
6. Dretzke J, Davenport C, Frew E, Barlow J, Stewart-Brown S, Bayliss S, et al. The clinical effectiveness of different parenting programmes for children with conduct problems: a systematic review of randomised controlled trials. Child Adolesc Psychiatry Ment Health 2009 Mar 04;3(1):7 [FREE Full text] [doi: 10.1186/1753-2000-3-7] [Medline: 19261188]
7. Yap MB, Morgan AJ, Cairns K, Jorm AF, Hetrick SE, Merry S. Parents in prevention: a meta-analysis of randomized controlled trials of parenting interventions to prevent internalizing problems in children from birth to age 18. Clin Psychol Rev 2016 Dec;50:138-158. [doi: 10.1016/j.cpr.2016.10.003] [Medline: 27969003]
8. Chacko A, Jensen SA, Lowry LS, Cornwell M, Chimklis A, Chan E, et al. Engagement in behavioral parent training: review of the literature and implications for practice. Clin Child Fam Psychol Rev 2016 Sep;19(3):204-215. [doi: 10.1007/s10567-016-0205-2] [Medline: 27311693]
9. McCurdy K, Daro D. Parent involvement in family support programs: an integrated theory. Fam Relat 2004 Feb 19;50(2):113-121 [FREE Full text] [doi: 10.1111/j.1741-3729.2001.00113.x]
10. Finan SJ, Warren N, Priest N, Mak JS, Yap MB. Parent non-engagement in preventive parenting programs for adolescent mental health: stakeholder views. J Child Fam Stud 2019 Oct 19;29(3):711-724 [FREE Full text] [doi: 10.1007/s10826-019-01627-x]
11. Mytton J, Ingram J, Manns S, Thomas J. Facilitators and barriers to engagement in parenting programs: a qualitative systematic review. Health Educ Behav 2014 Apr 02;41(2):127-137. [doi: 10.1177/1090198113485755] [Medline: 23640123]
12. Finan SJ, Yap MB. Engaging parents in preventive programs for adolescent mental health: a socio‐ecological framework. J Fam Theory Rev 2021 Nov 02;13(4):515-527 [FREE Full text] [doi: 10.1111/jftr.12440]
13. Berry V, Melendez-Torres GJ, Axford N, Axberg U, de Castro BO, Gardner F, et al. Does social and economic disadvantage predict lower engagement with parenting interventions? An integrative analysis using individual participant data. Prev Sci 2023 Nov;24(8):1447-1458 [FREE Full text] [doi: 10.1007/s11121-022-01404-1] [Medline: 35870094]
14. Farrant B. Mal adaptive parenting and child emotional symptoms in the early school years: findings from the longitudinal study of Australian children. Australas J Early Child 2014 Jun 01;39(2):118-125 [FREE Full text] [doi: 10.1177/183693911403900215]
15. Gard AM, McLoyd VC, Mitchell C, Hyde LW. Evaluation of a longitudinal family stress model in a population-based cohort. Soc Dev 2020 Nov 24;29(4):1155-1175 [FREE Full text] [doi: 10.1111/sode.12446] [Medline: 33953492]
16. Sahle BW, Reavley NJ, Li W, Morgan AJ, Yap MB, Reupert A, et al. The association between adverse childhood experiences and common mental disorders and suicidality: an umbrella review of systematic reviews and meta-analyses. Eur Child Adolesc Psychiatry 2022 Oct;31(10):1489-1499. [doi: 10.1007/s00787-021-01745-2] [Medline: 33638709]
17. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
18. Hall CM, Bierman KL. Technology-assisted interventions for parents of young children: emerging practices, current research, and future directions. Early Child Res Q 2015;33:21-32 [FREE Full text] [doi: 10.1016/j.ecresq.2015.05.003] [Medline: 27773964]
19. Taylor CB, Fitzsimmons-Craft EE, Graham AK. Digital technology can revolutionize mental health services delivery: the COVID-19 crisis as a catalyst for change. Int J Eat Disord 2020 Jul 25;53(7):1155-1157 [FREE Full text] [doi: 10.1002/eat.23300] [Medline: 32449523]
20. Hansen A, Broomfield G, Yap MB. A systematic review of technology-assisted parenting programs for mental health problems in youth aged 0–18 years: applicability to underserved Australian communities. Aust J Psychol 2020 Nov 20;71(4):433-462. [doi: 10.1111/ajpy.12250]

XSL•FO
RenderX

21. Spencer CM, Topham GL, King EL. Do online parenting programs create change?: a meta-analysis. J Fam Psychol 2020 Apr;34(3):364-374. [doi: 10.1037/fam0000605] [Medline: 31697102]
22. Baumel A, Pawar A, Kane JM, Correll CU. Digital parent training for children with disruptive behaviors: systematic review and meta-analysis of randomized trials. J Child Adolesc Psychopharmacol 2016 Oct;26(8):740-749. [doi: 10.1089/cap.2016.0048] [Medline: 27286325]
23. Corralejo SM, Domenech Rodríguez MM. Technology in parenting programs: a systematic review of existing interventions. J Child Fam Stud 2018 Jun 5;27(9):2717-2731 [FREE Full text] [doi: 10.1007/s10826-018-1117-1]
24. Florean IS, Dobrean A, Păsărelu CR, Georgescu RD, Milea I. The efficacy of internet-based parenting programs for children and adolescents with behavior problems: a meta-analysis of randomized clinical trials. Clin Child Fam Psychol Rev 2020 Dec;23(4):510-528. [doi: 10.1007/s10567-020-00326-0] [Medline: 32897527]
25. Flujas-Contreras JM, García-Palacios A, Gómez I. Technology-based parenting interventions for children's physical and psychological health: a systematic review and meta-analysis. Psychol Med 2019 Aug;49(11):1787-1798. [doi: 10.1017/S0033291719000692] [Medline: 30977462]
26. Harris M, Andrews K, Gonzalez A, Prime H, Atkinson L. Technology-assisted parenting interventions for families experiencing social disadvantage: a meta-analysis. Prev Sci 2020 Jul 15;21(5):714-727. [doi: 10.1007/s11121-020-01128-0] [Medline: 32415543]
27. Ingoldsby EM. Review of interventions to improve family engagement and retention in parent and child mental health programs. J Child Fam Stud 2010 Oct 01;19(5):629-645 [FREE Full text] [doi: 10.1007/s10826-009-9350-2] [Medline: 20823946]
28. Finan SJ, Swierzbiolek B, Priest N, Warren N, Yap M. Parental engagement in preventive parenting programs for child mental health: a systematic review of predictors and strategies to increase engagement. PeerJ 2018 Apr 27;6:e4676 [FREE Full text] [doi: 10.7717/peerj.4676] [Medline: 29719737]
29. Piotrowska PJ, Tully LA, Lenroot R, Kimonis E, Hawes D, Moul C, et al. Mothers, fathers, and parental systems: a conceptual model of parental engagement in programmes for child mental health-connect, attend, participate, enact (CAPE). Clin Child Fam Psychol Rev 2017 Jun 2;20(2):146-161 [FREE Full text] [doi: 10.1007/s10567-016-0219-9] [Medline: 27914017]
30. Thongseiratch T, Leijten P, Melendez-Torres GJ. Online parent programs for children's behavioral problems: a meta-analytic review. Eur Child Adolesc Psychiatry 2020 Nov 10;29(11):1555-1568. [doi: 10.1007/s00787-020-01472-0] [Medline: 31925545]
31. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
32. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
33. Randolph KA, Fincham F, Radey M. A framework for engaging parents in prevention. J Fam Soc Work 2009 Feb 20;12(1):56-72 [FREE Full text] [doi: 10.1080/10522150802654278]
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
35. Breitenstein SM, Gross D, Christophersen R. Digital delivery methods of parenting training interventions: a systematic review. Worldviews Evid Based Nurs 2014 Jun;11(3):168-176. [doi: 10.1111/wvn.12040] [Medline: 24842341]
36. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
37. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
38. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291 [FREE Full text] [doi: 10.3233/efi-180221]
39. Popay J, Roberts H, Sowden A, Petticrew M. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme. Evaluation 2009 Jan 01;15(1):49-73. [doi: 10.13140/2.1.1018.4643]
40. Borenstein M. Meta-analysis methods based on direction and p-values. In: Borenstein M, Hedges LV, Higgins JP, Rothstein HR, editors. Introduction to Meta‐Analysis. Hoboken, NJ: John Wiley & Sons; 2009.
41. Stouffer S, Suchman E, DeVinney L, Star S, Williams RS. The American Soldier, Vol 1: Adjustment During Army Life. Princeton, NJ: Princeton University Press; 1949.
42. Fox J, Archard PJ. Developing a digital toolkit for adoptive parents. J Soc Work Pract 2017 Mar 30;32(1):87-98. [doi: 10.1080/02650533.2017.1300876]

XSL•FO
RenderX

43. Lawrence KA, Yap MB, Rapee RM, Cardamone-Breen MC, Green J, Jorm AF. Partners in parenting: a multi-level web-based approach to support parents in prevention and early intervention for adolescent depression and anxiety. JMIR Ment Health 2017 Dec 19;4(4):e59 [FREE Full text] [doi: 10.2196/mental.8492] [Medline: 29258974]
44. May CD, Fletcher R. The development and application of a protocol for the writing, assessing, and validating of a corpus of relationship-focused text messages for new and expecting fathers. Health Informatics J 2019 Jun;25(2):240-246 [FREE Full text] [doi: 10.1177/1460458217704249] [Medline: 28464725]
45. Tiwari A, Yuk H, Pang P, Fong DY, Yuen F, Humphreys J, et al. Telephone intervention to improve the mental health of community-dwelling women abused by their intimate partners: a randomised controlled trial. Hong Kong Med J 2012 Dec;18 Suppl 6:14-17 [FREE Full text] [Medline: 23249846]
46. Baggett KM, Davis B, Sheeber LB, Ammerman RT, Mosley EA, Miller K, et al. Minding the gatekeepers: referral and recruitment of postpartum mothers with depression into a randomized controlled trial of a mobile internet parenting intervention to improve mood and optimize infant social communication outcomes. Int J Environ Res Public Health 2020 Dec 02;17(23):8978 [FREE Full text] [doi: 10.3390/ijerph17238978] [Medline: 33276610]
47. Becher EH, Cronin S, McCann E, Olson KA, Powell S, Marczak MS. Parents forever: evaluation of an online divorce education program. J Divorce Remarriage 2015 Apr 27;56(4):261-276 [FREE Full text] [doi: 10.1080/10502556.2015.1025900]
48. Bloom TL, Glass NE, Case J, Wright C, Nolte K, Parsons L. Feasibility of an online safety planning intervention for rural and urban pregnant abused women. Nurs Res 2014;63(4):243-251. [doi: 10.1097/NNR.0000000000000036] [Medline: 24977721]
49. Breitenstein SM, Fogg L, Ocampo EV, Acosta DI, Gross D. Parent use and efficacy of a self-administered, tablet-based parent training intervention: a randomized controlled trial. JMIR Mhealth Uhealth 2016 Apr 20;4(2):e36 [FREE Full text] [doi: 10.2196/mhealth.5202] [Medline: 27098111]
50. Breitenstein SM, Shane J, Julion W, Gross D. Developing the eCPP: adapting an evidence-based parent training program for digital delivery in primary care settings. Worldviews Evid Based Nurs 2015 Feb;12(1):31-40. [doi: 10.1111/wvn.12074] [Medline: 25588505]
51. Cardamone-Breen MC, Jorm AF, Lawrence KA, Rapee RM, Mackinnon AJ, Yap MB. A single-session, web-based parenting intervention to prevent adolescent depression and anxiety disorders: randomized controlled trial. J Med Internet Res 2018 Apr 26;20(4):e148 [FREE Full text] [doi: 10.2196/jmir.9499] [Medline: 29699964]
52. Choi H, Kim S, Ko H, Kim Y, Park CG. Development and preliminary evaluation of culturally specific web-based intervention for parents of adolescents. J Psychiatr Ment Health Nurs 2016 Oct 08;23(8):489-501. [doi: 10.1111/jpm.12327] [Medline: 27500792]
53. Chu JT, Wadham A, Jiang Y, Whittaker R, Stasiak K, Shepherd M, et al. Development of MyTeen text messaging program to support parents of adolescents: qualitative study. JMIR Mhealth Uhealth 2019 Nov 20;7(11):e15664 [FREE Full text] [doi: 10.2196/15664] [Medline: 31746767]
54. DeGarmo DS, Jones JA. Fathering Through Change (FTC) intervention for single fathers: preventing coercive parenting and child problem behaviors. Dev Psychopathol 2019 Sep 06;31(5):1801-1811 [FREE Full text] [doi: 10.1017/s0954579419001019]
55. Fletcher R, Knight T, Macdonald JA, StGeorge J. Process evaluation of text-based support for fathers during the transition to fatherhood (SMS4dads): mechanisms of impact. BMC Psychol 2019 Sep 13;7(1):63 [FREE Full text] [doi: 10.1186/s40359-019-0338-4] [Medline: 31519207]
56. Fletcher R, Campbell L, Sved Williams A, Rawlinson C, Dye J, Baldwin A, et al. SMS4 perinatal parents: designing parenting support via text messages for mothers with severe mental illness (SMI) and their partners. Adv Ment Health 2018 Nov 29;17(1):85-95 [FREE Full text] [doi: 10.1080/18387357.2018.1550367]
57. Fletcher R, Hammond C, Faulkner D, Turner N, Shipley L, Read D, et al. Stayin' on Track: the feasibility of developing Internet and mobile phone-based resources to support young aboriginal fathers. Aust J Prim Health 2017 Sep;23(4):329-334. [doi: 10.1071/PY16151] [Medline: 28449728]
58. Fletcher R, May C, Wroe J, Hall P, Cooke D, Rawlinson C, et al. Development of a set of mobile phone text messages designed for new fathers. J Reprod Infant Psychol 2016 Aug 09;34(5):525-534. [doi: 10.1080/02646838.2016.1214250]
59. Fulgoni CM, Melvin GA, Jorm AF, Lawrence KA, Yap MB. The Therapist-assisted Online Parenting Strategies (TOPS) program for parents of adolescents with clinical anxiety or depression: development and feasibility pilot. Internet Interv 2019 Dec;18:100285 [FREE Full text] [doi: 10.1016/j.invent.2019.100285] [Medline: 31890632]
60. Gelatt VA, Adler-Baeder F, Seeley JR. An interactive web-based program for stepfamilies: development and evaluation of efficacy. Fam Relat 2010 Dec 02;59(5):572-586 [FREE Full text] [doi: 10.1111/j.1741-3729.2010.00624.x] [Medline: 33071416]
61. Mohammadinasab H, Mazaheri M, Reazaeizade M, Heydari M. The effectiveness of ABFT-based parenting training via cell phone on parent-adolescent conflict: a single case study. In: Proceedings of the 2020 International Serious Games Symposium. 2020 Presented at: ISGS '20; December 23-25, 2020; Tehran, Iran p. 107-114 URL: https://ieeexplore.ieee.org/document/9375430 [doi: 10.1109/isgs51981.2020.9375430]

XSL•FO
RenderX

62. Hamil J, Gier E, Garfield CF, Tandon D. The development and pilot of a technology-based intervention in the United States for father's mental health in the perinatal period. Am J Mens Health 2021;15(5):15579883211044306 [FREE Full text] [doi: 10.1177/15579883211044306] [Medline: 34587839]
63. Hinton S, Sheffield J, Sanders MR, Sofronoff K. A randomized controlled trial of a telehealth parenting intervention: a mixed-disability trial. Res Dev Disabil 2017 Jun;65:74-85. [doi: 10.1016/j.ridd.2017.04.005] [Medline: 28458049]
64. Huebner DM, Rullo JE, Thoma BC, McGarrity LA, Mackenzie J. Piloting lead with love: a film-based intervention to improve parents' responses to their lesbian, gay, and bisexual children. J Prim Prev 2013 Oct 13;34(5):359-369 [FREE Full text] [doi: 10.1007/s10935-013-0319-y] [Medline: 23943135]
65. James Riegler L, Raj SP, Moscato EL, Narad ME, Kincaid A, Wade SL. Pilot trial of a telepsychotherapy parenting skills intervention for veteran families: implications for managing parenting stress during COVID-19. J Psychother Integr 2020 Jun;30(2):290-303 [FREE Full text] [doi: 10.1037/int0000220]
66. Jones DJ, Forehand R, Cuellar J, Parent J, Honeycutt A, Khavjou O, et al. Technology-enhanced program for child disruptive behavior disorders: development and pilot randomized control trial. J Clin Child Adolesc Psychol 2014 Aug 07;43(1):88-101 [FREE Full text] [doi: 10.1080/15374416.2013.822308] [Medline: 23924046]
67. Kaplan K, Solomon P, Salzer MS, Brusilovskiy E. Assessing an Internet-based parenting intervention for mothers with a serious mental illness: a randomized controlled trial. Psychiatr Rehabil J 2014 Sep;37(3):222-231. [doi: 10.1037/prj0000080] [Medline: 24978623]
68. Khor SP, Fulgoni CM, Lewis D, Melvin GA, Jorm AF, Lawrence K, et al. Short-term outcomes of the therapist-assisted online parenting strategies intervention for parents of adolescents treated for anxiety and/or depression: a single-arm double-baseline trial. Aust N Z J Psychiatry 2022 Jun 07;56(6):695-708 [FREE Full text] [doi: 10.1177/00048674211025695] [Medline: 34231423]
69. Love SM, Sanders MR, Metzler CW, Prinz RJ, Kast EZ. Enhancing accessibility and engagement in evidence-based parenting programs to reduce maltreatment: conversations with vulnerable parents. J Public Child Welf 2013 Jan;7(1):20-38 [FREE Full text] [doi: 10.1080/15548732.2012.701837] [Medline: 23710156]
70. May CD, St George JM, Lane S. From presence to participation: engagement with an SMS program for fathers of children on the autism spectrum. J Child Fam Stud 2020 Nov 07;30(1):29-37 [FREE Full text] [doi: 10.1007/s10826-020-01845-8]
71. Metcalfe RE, Matulis JM, Cheng Y, Stormshak EA. Therapeutic alliance as a predictor of behavioral outcomes in a relationally focused, family-centered telehealth intervention. J Marital Fam Ther 2021 Apr 15;47(2):473-484 [FREE Full text] [doi: 10.1111/jmft.12517] [Medline: 33855726]
72. Metzler CW, Sanders MR, Rusby JC, Crowley RN. Using consumer preference information to increase the reach and impact of media-based parenting interventions in a public health approach to parenting support. Behav Ther 2012 Jun;43(2):257-270 [FREE Full text] [doi: 10.1016/j.beth.2011.05.004] [Medline: 22440064]
73. Morgan AJ, Rapee RM, Bayer JK. Prevention and early intervention of anxiety problems in young children: a pilot evaluation of cool little kids online. Internet Interv 2016 May;4:105-112 [FREE Full text] [doi: 10.1016/j.invent.2016.05.001] [Medline: 30135796]
74. Morgan AJ, Rapee RM, Salim A, Goharpey N, Tamir E, McLellan LF, et al. Internet-delivered parenting program for prevention and early intervention of anxiety problems in young children: randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2017 May;56(5):417-25.e1. [doi: 10.1016/j.jaac.2017.02.010] [Medline: 28433091]
75. Murry VM, Berkel C, Liu N. The closing digital divide: delivery modality and family attendance in the pathways for African American Success (PAAS) program. Prev Sci 2018 Jul 22;19(5):642-651 [FREE Full text] [doi: 10.1007/s11121-018-0863-z] [Medline: 29356928]
76. Piotrowska PJ, Tully LA, Collins DA, Sawrikar V, Hawes D, Kimonis ER, et al. ParentWorks: evaluation of an online, father-inclusive, universal parenting intervention to reduce child conduct problems. Child Psychiatry Hum Dev 2020 Aug 24;51(4):503-513 [FREE Full text] [doi: 10.1007/s10578-019-00934-0] [Medline: 31650461]
77. Ragavan MI, Ferre V, Bair-Merritt M. Thrive: a novel health education mobile application for mothers who have experienced intimate partner violence. Health Promot Pract 2020 Mar 24;21(2):160-164. [doi: 10.1177/1524839919890870] [Medline: 31874566]
78. Rizzo CJ, Houck C, Barker D, Collibee C, Hood E, Bala K. Project STRONG: an online, parent-son intervention for the prevention of dating violence among early adolescent boys. Prev Sci 2021 Feb 17;22(2):193-204 [FREE Full text] [doi: 10.1007/s11121-020-01168-6] [Medline: 32940857]
79. Sawyer A, Kaim A, Le HN, McDonald D, Mittinty M, Lynch J, et al. The effectiveness of an app-based nurse-moderated program for new mothers with depression and parenting problems (eMums plus): pragmatic randomized controlled trial. J Med Internet Res 2019 Jun 04;21(6):e13689 [FREE Full text] [doi: 10.2196/13689] [Medline: 31165715]
80. Self-Brown SR, C Osborne M, Rostad W, Feil E. A technology-mediated approach to the implementation of an evidence-based child maltreatment prevention program. Child Maltreat 2017 Nov;22(4):344-353. [doi: 10.1177/1077559516678482] [Medline: 27837009]
81. Sheeber LB, Seeley JR, Feil EG, Davis B, Sorensen E, Kosty DB, et al. Development and pilot evaluation of an internet-facilitated cognitive-behavioral intervention for maternal depression. J Consult Clin Psychol 2012 Oct;80(5):739-749 [FREE Full text] [doi: 10.1037/a0028820] [Medline: 22663903]

XSL•FO
RenderX

82. Sung JY, Mumper E, Schleider JL. Empowering anxious parents to manage child avoidance behaviors: randomized control trial of a single-session intervention for parental accommodation. JMIR Ment Health 2021 Jul 06;8(7):e29538 [FREE Full text] [doi: 10.2196/29538] [Medline: 34255718]
83. Tully LA, Piotrowska PJ, Collins DA, Frick PJ, Anderson V, Moul C, et al. Evaluation of 'the father effect' media campaign to increase awareness of, and participation in, an online father-inclusive parenting program. Health Commun 2019 Nov 09;34(12):1423-1432. [doi: 10.1080/10410236.2018.1495160] [Medline: 29985642]
84. Wade SL, Cassedy AE, Shultz EL, Zang H, Zhang N, Kirkwood MW, et al. Randomized clinical trial of online parent training for behavior problems after early brain injury. J Am Acad Child Adolesc Psychiatry 2017 Nov;56(11):930-9.e2. [doi: 10.1016/j.jaac.2017.09.413] [Medline: 29096775]
85. Yap MB, Mahtani S, Rapee RM, Nicolas C, Lawrence KA, Mackinnon A, et al. A tailored web-based intervention to improve parenting risk and protective factors for adolescent depression and anxiety problems: postintervention findings from a randomized controlled trial. J Med Internet Res 2018 Jan 19;20(1):e17 [FREE Full text] [doi: 10.2196/jmir.9139] [Medline: 29351895]
86. Agazzi H, Hayford H, Thomas N, Ortiz C, Salinas-Miranda A. A nonrandomized trial of a behavioral parent training intervention for parents with children with challenging behaviors: in-person versus internet-HOT DOCS. Clin Child Psychol Psychiatry 2021 Oct 22;26(4):1076-1088. [doi: 10.1177/13591045211027559] [Medline: 34156883]
87. Antonini TN, Raj SP, Oberjohn KS, Cassedy A, Makoroff KL, Fouladi M, et al. A pilot randomized trial of an online parenting skills program for pediatric traumatic brain injury: improvements in parenting and child behavior. Behav Ther 2014 Jul;45(4):455-468. [doi: 10.1016/j.beth.2014.02.003] [Medline: 24912459]
88. Baggett K, Davis B, Feil E, Sheeber L, Landry S, Leve C, et al. A randomized controlled trial examination of a remote parenting intervention: engagement and effects on parenting behavior and child abuse potential. Child Maltreat 2017 Nov 07;22(4):315-323 [FREE Full text] [doi: 10.1177/1077559517712000] [Medline: 28587520]
89. Baggett KM, Davis B, Feil EG, Sheeber LL, Landry SH, Carta JJ, et al. Technologies for expanding the reach of evidence-based interventions: preliminary results for promoting social-emotional development in early childhood. Topics Early Child Spec Educ 2010 Feb 01;29(4):226-238 [FREE Full text] [doi: 10.1177/0271121409354782] [Medline: 20454545]
90. Baker M, Biringen Z, Meyer-Parsons B, Schneider A. Emotional attachment and emotional availability tele-intervention for adoptive families. Infant Ment Health J 2015 Feb 20;36(2):179-192. [doi: 10.1002/imhj.21498] [Medline: 25704237]
91. Baker S, Sanders MR, Turner KM, Morawska A. A randomized controlled trial evaluating a low-intensity interactive online parenting intervention, triple p online brief, with parents of children with early onset conduct problems. Behav Res Ther 2017 Apr;91:78-90. [doi: 10.1016/j.brat.2017.01.016] [Medline: 28167330]
92. Bodenmann G, Hilpert P, Nussbeck FW, Bradbury TN. Enhancement of couples' communication and dyadic coping by a self-directed approach: a randomized controlled trial. J Consult Clin Psychol 2014 Aug;82(4):580-591. [doi: 10.1037/a0036356] [Medline: 24660673]
93. Boekhorst MG, Hulsbosch LP, Nyklíček I, Spek V, Kastelein A, Bögels S, et al. An online mindful parenting training for mothers raising toddlers: assessment of acceptability, effectiveness, and personal goals. Mindfulness 2020 Dec 09;12(2):519-531. [doi: 10.1007/s12671-020-01542-z]
94. Breaux R, Shroff DM, Cash AR, Swanson CS, Carlton C, Bertollo JR, et al. Telehealth delivery of the RELAX intervention for families of adolescents diagnosed with ADHD: preliminary treatment outcomes and evidence of acceptability and feasibility. Evid Based Pract Child Adolesc Ment Health 2021 Nov 02;8(1):24-38 [FREE Full text] [doi: 10.1080/23794925.2021.1970053]
95. Breitenstein SM, Fehrenbacher C, Holod AF, Schoeny ME. A randomized trial of digitally delivered, self-administered parent training in primary care: effects on parenting and child behavior. J Pediatr 2021 Apr;231:207-14.e4 [FREE Full text] [doi: 10.1016/j.jpeds.2020.12.016] [Medline: 33338496]
96. Brophy-Herb HE, Moyses K, Shrier C, Rymanowicz K, Pilkenton A, Dalimonte-Merckling D, et al. A pilot evaluation of the Building Early Emotional Skills (BEES) curriculum in face-to-face and online formats. J Community Psychol 2021 Jul 09;49(5):1505-1521. [doi: 10.1002/jcop.22478] [Medline: 33561319]
97. Cefai J, Smith D, Pushak RE. Parenting wisely: parent training via CD-ROM with an Australian sample. Child Fam Behav Ther 2010 Mar 03;32(1):17-33. [doi: 10.1080/07317100903539709]
98. Comer JS, Furr JM, Miguel EM, Cooper-Vince CE, Carpenter AL, Elkins RM, et al. Remotely delivering real-time parent training to the home: an initial randomized trial of Internet-delivered parent-child interaction therapy (I-PCIT). J Consult Clin Psychol 2017 Sep;85(9):909-917. [doi: 10.1037/ccp0000230] [Medline: 28650194]
99. Cotter KL, Bacallao M, Smokowski PR, Robertson CI. Parenting interventions implementation science: how delivery format impacts the parenting wisely program. Res Soc Work Pract 2013 May 31;23(6):639-650. [doi: 10.1177/1049731513490811]
100. Czymoniewicz-Klippel M, Chesnut R, DiNallo J, Perkins D. Patterns of participation in the Grow parenting program. J Child Serv 2019 Mar 18;14(1):27-41 [FREE Full text] [doi: 10.1108/jcs-06-2018-0014]
101. Dadds MR, Thai C, Mendoza Diaz A, Broderick J, Moul C, Tully LA, et al. Therapist-assisted online treatment for child conduct problems in rural and urban families: Two randomized controlled trials. J Consult Clin Psychol 2019 Aug;87(8):706-719. [doi: 10.1037/ccp0000419] [Medline: 31204839]

XSL•FO
RenderX

102. Day JJ, Sanders MR. Do parents benefit from help when completing a self-guided parenting program online? A randomized controlled trial comparing triple P online with and without telephone support. Behav Ther 2018 Nov;49(6):1020-1038. [doi: 10.1016/j.beth.2018.03.002] [Medline: 30316482]
103. Donovan CL, March S. Online CBT for preschool anxiety disorders: a randomised control trial. Behav Res Ther 2014 Jul;58:24-35. [doi: 10.1016/j.brat.2014.05.001] [Medline: 24927471]
104. Doss BD, Roddy MK, Llabre MM, Georgia Salivar E, Jensen-Doss A. Improvements in coparenting conflict and child adjustment following an online program for relationship distress. J Fam Psychol 2020 Feb;34(1):68-78 [FREE Full text] [doi: 10.1037/fam0000582] [Medline: 31380688]
105. Ehrensaft MK, Knous-Westfall HM, Alonso TL. Web-based prevention of parenting difficulties in young, urban mothers enrolled in post-secondary education. J Prim Prev 2016 Dec 13;37(6):527-542. [doi: 10.1007/s10935-016-0448-1] [Medline: 27624608]
106. Enebrink P, Högström J, Forster M, Ghaderi A. Internet-based parent management training: a randomized controlled study. Behav Res Ther 2012 Apr;50(4):240-249. [doi: 10.1016/j.brat.2012.01.006] [Medline: 22398153]
107. Epstein M, Oesterle S, Haggerty KP. Effectiveness of Facebook groups to boost participation in a parenting intervention. Prev Sci 2019 Aug 23;20(6):894-903 [FREE Full text] [doi: 10.1007/s11121-019-01018-0] [Medline: 31124023]
108. Farris JR, Bert SS, Nicholson JS, Glass K, Borkowski JG. Effective intervention programming: improving maternal adjustment through parent education. Adm Policy Ment Health 2013 May 14;40(3):211-223. [doi: 10.1007/s10488-011-0397-1] [Medline: 22246615]
109. Feil EG, Baggett K, Davis B, Landry S, Sheeber L, Leve C, et al. Randomized control trial of an internet-based parenting intervention for mothers of infants. Early Child Res Q 2020;50(Pt 1):36-44 [FREE Full text] [doi: 10.1016/j.ecresq.2018.11.003] [Medline: 32863565]
110. Ferraro AJ, Oehme K, Bruker M, Arpan L, Opel A. The impact of training videos on attitudes about parenting after divorce. J Divorce Remarriage 2018 May 04;59(7):590-600 [FREE Full text] [doi: 10.1080/10502556.2018.1466253]
111. Fleming GE, Kohlhoff J, Morgan S, Turnell A, Maiuolo M, Kimonis ER. An effectiveness open trial of internet-delivered parent training for young children with conduct problems living in regional and rural Australia. Behav Ther 2021 Jan;52(1):110-123. [doi: 10.1016/j.beth.2020.03.001] [Medline: 33483109]
112. Flujas-Contreras JM, García-Palacios A, Gómez I. Effectiveness of a web-based intervention on parental psychological flexibility and emotion regulation: a pilot open trial. Int J Environ Res Public Health 2021 Mar 13;18(6):2958 [FREE Full text] [doi: 10.3390/ijerph18062958] [Medline: 33805835]
113. Ford-Gilboe M, Varcoe C, Scott-Storey K, Perrin N, Wuest J, Wathen CN, et al. Longitudinal impacts of an online safety and health intervention for women experiencing intimate partner violence: randomized controlled trial. BMC Public Health 2020 Feb 26;20(1):260 [FREE Full text] [doi: 10.1186/s12889-020-8152-8] [Medline: 32098633]
114. Franke N, Keown LJ, Sanders MR. An RCT of an online parenting program for parents of preschool-aged children with ADHD symptoms. J Atten Disord 2020 Oct 08;24(12):1716-1726. [doi: 10.1177/1087054716667598] [Medline: 27609783]
115. Ghaderi A, Kadesjö C, Björnsdotter A, Enebrink P. Randomized effectiveness trial of the family check-up versus internet-delivered parent training (iComet) for families of children with conduct problems. Sci Rep 2018 Jul 31;8(1):11486 [FREE Full text] [doi: 10.1038/s41598-018-29550-z] [Medline: 30065246]
116. Gülırmak K, Orak OS. Effectiveness of web-based distance education for parents in the prevention of emotional neglect and abuse: a randomized controlled study. Perspect Psychiatr Care 2021 Apr 15;57(2):573-582 [FREE Full text] [doi: 10.1111/ppc.12580] [Medline: 32668039]
117. Hegarty K, Tarzia L, Valpied J, Murray E, Humphreys C, Taft A, et al. An online healthy relationship tool and safety decision aid for women experiencing intimate partner violence (I-DECIDE): a randomised controlled trial. Lancet Public Health 2019 Jun;4(6):e301-e310. [doi: 10.1016/s2468-2667(19)30079-9]
118. Hemdi A, Daley D. The effectiveness of a psychoeducation intervention delivered via WhatsApp for mothers of children with autism spectrum disorder (ASD) in the kingdom of Saudi Arabia: a randomized controlled trial. Child Care Health Dev 2017 Nov 20;43(6):933-941. [doi: 10.1111/cch.12520] [Medline: 28929582]
119. Holden GW, Brown AS, Baldwin AS, Croft Caderao K. Research findings can change attitudes about corporal punishment. Child Abuse Negl 2014 May;38(5):902-908. [doi: 10.1016/j.chiabu.2013.10.013] [Medline: 24246718]
120. Hudson DB, Campbell-Grossman C, Hertzog M. Effects of an internet intervention on mothers' psychological, parenting, and health care utilization outcomes. Issues Compr Pediatr Nurs 2012;35(3-4):176-193. [doi: 10.3109/01460862.2012.734211] [Medline: 23140413]
121. Irvine AB, Gelatt VA, Hammond M, Seeley JR. A randomized study of internet parent training accessed from community technology centers. Prev Sci 2015 May 29;16(4):597-608. [doi: 10.1007/s11121-014-0521-z] [Medline: 25351866]
122. Jones DJ, Loiselle R, Zachary C, Georgeson AR, Highlander A, Turner P, et al. Optimizing engagement in behavioral parent training: progress toward a technology-enhanced treatment model. Behav Ther 2021 Mar;52(2):508-521 [FREE Full text] [doi: 10.1016/j.beth.2020.07.001] [Medline: 33622517]
123. Jones S, Calam R, Sanders M, Diggle PJ, Dempsey R, Sadhnani V. A pilot web based positive parenting intervention to help bipolar parents to improve perceived parenting skills and child outcomes. Behav Cogn Psychother 2014 May;42(3):283-296. [doi: 10.1017/S135246581300009X] [Medline: 23544953]

XSL•FO
RenderX

124. Kavanagh D, Connolly J, Fisher J, Halford WK, Hamilton K, Hides L, et al. The baby steps web program for the well-being of new parents: randomized controlled trial. J Med Internet Res 2021 Nov 26;23(11):e23659 [FREE Full text] [doi: 10.2196/23659] [Medline: 34842534]
125. Kirkman JJ, Hawes DJ, Dadds MR. An open trial for an e-health treatment for child behavior disorders I: social acceptability, engagement, and therapeutic process. Evid Based Pract Child Adolesc Ment Health 2016 Sep 20;1(4):196-212. [doi: 10.1080/23794925.2016.1230481]
126. Kopystynska O, Turner JJ, Schramm DG, Higginbotham B. Evaluation of topics in Utah's ONE - HOUR divorce education program. Fam Court Rev 2020 Jul 27;58(3):804-815. [doi: 10.1111/fcre.12469]
127. Koziol-McLain J, Vandal AC, Wilson D, Nada-Raja S, Dobbs T, McLean C, et al. Efficacy of a web-based safety decision aid for women experiencing intimate partner violence: randomized controlled trial. J Med Internet Res 2018 Jan 10;19(12):e426 [FREE Full text] [doi: 10.2196/jmir.8617] [Medline: 29321125]
128. Loew B, Rhoades G, Markman H, Stanley S, Pacifici C, White L, et al. Internet delivery of PREP-based relationship education for at-risk couples. J Couple Relatsh Ther 2012 Oct;11(4):291-309. [doi: 10.1080/15332691.2012.718968]
129. Love SM, Sanders MR, Turner KM, Maurange M, Knott T, Prinz R, et al. Social media and gamification: engaging vulnerable parents in an online evidence-based parenting program. Child Abuse Negl 2016 Mar;53:95-107. [doi: 10.1016/j.chiabu.2015.10.031] [Medline: 26880281]
130. Mast JE, Antonini TN, Raj SP, Oberjohn KS, Cassedy A, Makoroff KL, et al. Web-based parenting skills to reduce behavior problems following abusive head trauma: a pilot study. Child Abuse Negl 2014 Sep;38(9):1487-1495. [doi: 10.1016/j.chiabu.2014.04.012] [Medline: 24844734]
131. Mello MJ, Bromberg JR, Baird J, Wills H, Gaines BA, Lapidus G, et al. Feasibility and acceptability of an electronic parenting skills intervention for parents of alcohol-using adolescent trauma patients. Telemed J E Health 2019 Sep 01;25(9):833-839 [FREE Full text] [doi: 10.1089/tmj.2018.0201] [Medline: 30484743]
132. Morawska A, Tometzki H, Sanders MR. An evaluation of the efficacy of a triple P-positive parenting program podcast series. J Dev Behav Pediatr 2014;35(2):128-137. [doi: 10.1097/DBP.0000000000000020] [Medline: 24343191]
133. Narad M, Minich N, Taylor HG, Kirkwood MW, Brown TM, Stancin T, et al. Effects of a web-based intervention on family functioning following pediatric traumatic brain injury. J Dev Behav Pediatr 2015;36(9):700-707 [FREE Full text] [doi: 10.1097/DBP.0000000000000208] [Medline: 26461100]
134. Nickerson AB, Livingston JA, Kamper-DeMarco K. Evaluation of second step child protection videos: a randomized controlled trial. Child Abuse Negl 2018 Feb;76:10-22. [doi: 10.1016/j.chiabu.2017.10.001] [Medline: 28992513]
135. O'Shea A, Kaplan K, Solomon P, Salzer MS. Randomized controlled trial of an internet-based educational intervention for mothers with mental illnesses: an 18-month follow-up. Psychiatr Serv 2019 Aug 01;70(8):732-735. [doi: 10.1176/appi.ps.201800391] [Medline: 31023190]
136. Olthuis JV, McGrath PJ, Cunningham CE, Boyle MH, Lingley-Pottie P, Reid GJ, et al. Distance-delivered parent training for childhood disruptive behavior (Strongest Families™): a randomized controlled trial and economic analysis. J Abnorm Child Psychol 2018 Nov 8;46(8):1613-1629. [doi: 10.1007/s10802-018-0413-y] [Medline: 29516341]
137. Owen DA, Hutchings J. An evaluation of the online universal programme : a feasibility study. J Public Health Res 2017 Apr 13;6(1):819 [FREE Full text] [doi: 10.4081/jphr.2017.819] [Medline: 28660177]
138. Perrino T, Estrada Y, Huang S, St George S, Pantin H, Cano M, et al. Predictors of participation in an eHealth, family-based preventive intervention for Hispanic youth. Prev Sci 2018 Jul 4;19(5):630-641. [doi: 10.1007/s11121-016-0711-y] [Medline: 27704326]
139. Porzig-Drummond R, Stevenson RJ, Stevenson C. Preliminary evaluation of a self-directed video-based 1-2-3 Magic parenting program: a randomized controlled trial. Behav Res Ther 2015 Mar;66:32-42. [doi: 10.1016/j.brat.2015.01.003] [Medline: 25687560]
140. Potharst ES, Boekhorst MG, Cuijlits I, van Broekhoven KE, Jacobs A, Spek V, et al. A randomized control trial evaluating an online mindful parenting training for mothers with elevated parental stress. Front Psychol 2019 Jul 17;10:1550 [FREE Full text] [doi: 10.3389/fpsyg.2019.01550] [Medline: 31379646]
141. Prinz RJ, Metzler CW, Sanders MR, Rusby JC, Cai C. Online-delivered parenting intervention for young children with disruptive behavior problems: a noninferiority trial focused on child and parent outcomes. J Child Psychol Psychiatry 2022 Feb 08;63(2):199-209 [FREE Full text] [doi: 10.1111/jcpp.13426] [Medline: 33829499]
142. Rabbitt SM, Carrubba E, Lecza B, McWhinney E, Pope J, Kazdin AE. Reducing therapist contact in parenting programs: evaluation of internet-based treatments for child conduct problems. J Child Fam Stud 2016 Jun 23;25(6):2001-2020 [FREE Full text] [doi: 10.1007/s10826-016-0363-3] [Medline: 27453678]
143. Raj S, Shultz E, Zang H, Zhang N, Kirkwood MW, Taylor HG, et al. Effects of web-based parent training on caregiver functioning following pediatric traumatic brain injury: a randomized control trial. J Head Trauma Rehabil 2018;33(6):E19-E29. [doi: 10.1097/HTR.0000000000000388] [Medline: 29601344]
144. Razuri EB, Howard AR, Parris SR, Call CD, DeLuna JH, Hall JS, et al. Decrease in behavioral problems and trauma symptoms among at-risk adopted children following web-based trauma-informed parent training intervention. J Evid Inf Soc Work 2016 Jun 14;13(2):165-178 [FREE Full text] [doi: 10.1080/23761407.2015.1014123] [Medline: 26072917]

XSL•FO
RenderX

145. Richardson HL, Damashek A. Examining the use of a brief online intervention in primary care for changing low-income caregivers' attitudes toward spanking. J Interpers Violence 2022 Nov 22;37(21-22):NP20409-NP20427. [doi: 10.1177/08862605211054101] [Medline: 34802302]
146. Ristkari T, Kurki M, Suominen A, Gilbert S, Sinokki A, Kinnunen M, et al. Web-based parent training intervention with telephone coaching for disruptive behavior in 4-year-old children in real-world practice: implementation study. J Med Internet Res 2019 Apr 11;21(4):e11446 [FREE Full text] [doi: 10.2196/11446] [Medline: 30973337]
147. Rudd BN, Holtzworth-Munroe A, Reyome JG, Applegate AG, D'Onofrio BM. Randomized control trial: online parent program and waiting period for unmarried parents in title IV-D court. J Fam Psychol 2015 Oct;29(5):679-686. [doi: 10.1037/fam0000106] [Medline: 26075738]
148. Russell BS, Maksut JL, Lincoln CR, Leland AJ. Computer-mediated parenting education: digital family service provision. Child Youth Serv Rev 2016 Mar;62:1-8. [doi: 10.1016/j.childyouth.2016.01.012]
149. Sanders MR, Dittman CK, Farruggia SP, Keown LJ. A comparison of online versus workbook delivery of a self-help positive parenting program. J Prim Prev 2014 Jun 6;35(3):125-133. [doi: 10.1007/s10935-014-0339-2] [Medline: 24500106]
150. Sanders MR, Baker S, Turner KM. A randomized controlled trial evaluating the efficacy of Triple P online with parents of children with early-onset conduct problems. Behav Res Ther 2012 Nov;50(11):675-684. [doi: 10.1016/j.brat.2012.07.004] [Medline: 22982082]
151. Scholer SJ, Hudnut-Beumler J, Dietrich MS. A brief primary care intervention helps parents develop plans to discipline. Pediatrics 2010 Feb;125(2):e242-e249. [doi: 10.1542/peds.2009-0874] [Medline: 20083523]
152. Schramm DG, McCaulley G. Divorce education for parents: a comparison of online and in-person delivery methods. J Divorce Remarriage 2012 Nov;53(8):602-617. [doi: 10.1080/10502556.2012.721301]
153. Self-Brown S, Cowart-Osborne M, Baker E, Thomas A, Boyd C, Chege E, et al. Dad2K: an adaptation of SafeCare to enhance positive parenting skills with at-risk fathers. Child Fam Behav Ther 2015 Jun 11;37(2):138-155. [doi: 10.1080/07317107.2015.1035992]
154. Self-Brown SR, Reuben K, Perry EW, Bullinger LR, Osborne MC, Bielecki J, et al. The impact of COVID-19 on the delivery of an evidence-based child maltreatment prevention program: understanding the perspectives of SafeCare® providers. J Fam Violence 2022;37(5):825-835 [FREE Full text] [doi: 10.1007/s10896-020-00217-6] [Medline: 33173254]
155. Sim WH, Fernando LM, Jorm AF, Rapee RM, Lawrence KA, Mackinnon AJ, et al. A tailored online intervention to improve parenting risk and protective factors for child anxiety and depression: medium-term findings from a randomized controlled trial. J Affect Disord 2020 Dec 01;277:814-824. [doi: 10.1016/j.jad.2020.09.019] [Medline: 33065822]
156. Sourander A, McGrath PJ, Ristkari T, Cunningham C, Huttunen J, Lingley-Pottie P, et al. Internet-assisted parent training intervention for disruptive behavior in 4-year-old children: a randomized clinical trial. JAMA Psychiatry 2016 Apr 01;73(4):378-387. [doi: 10.1001/jamapsychiatry.2015.3411] [Medline: 26913614]
157. Spence SH, Prosser SJ, March S, Donovan CL. Internet-delivered cognitive behavior therapy with minimal therapist support for anxious children and adolescents: predictors of response. J Child Psychol Psychiatry 2020 Aug 19;61(8):914-927. [doi: 10.1111/jcpp.13257] [Medline: 32424896]
158. Stalker KC, Rose RA, Bacallao M, Smokowski PR. Parenting wisely six months later: how implementation delivery impacts program effects at follow-up. J Prim Prev 2018 Apr 16;39(2):129-153. [doi: 10.1007/s10935-017-0495-2] [Medline: 29453502]
159. Stevens J, Scribano PV, Marshall J, Nadkarni R, Hayes J, Kelleher KJ. A trial of telephone support services to prevent further intimate partner violence. Violence Against Women 2015 Dec 28;21(12):1528-1547. [doi: 10.1177/1077801215596849] [Medline: 26223696]
160. Taylor LC, Leary KA, Boyle AE, Bigelow KE, Henry T, DeRosier M. Parent training and adolescent social functioning: a brief report. J Child Fam Stud 2015 Jan 14;24(10):3030-3037. [doi: 10.1007/s10826-014-0106-2]
161. Thomson M. Alternative modes of delivery for family dispute resolution: the telephone dispute resolution service and the online FDR project1. J Fam Stud 2011;17(3):253-257 [FREE Full text] [doi: 10.5172/jfs.2011.17.3.253]
162. Traube DE, Hsiao HY, Rau A, Hunt-O'Brien D, Lu L, Islam N. Advancing home based parenting programs through the use of telehealth technology. J Child Fam Stud 2019 May 25;29(1):44-53. [doi: 10.1007/s10826-019-01458-w]
163. Turner JJ, Kopystynska O, Bradford K, Schramm DG, Higginbotham BJ. Evaluating perceived impact of Utah's online divorce education program. J Divorce Remarriage 2021 Feb 08;62(3):179-198. [doi: 10.1080/10502556.2021.1871829]
164. van der Zanden RA, Speetjens PA, Arntz KS, Onrust SA. Online group course for parents with mental illness: development and pilot study. J Med Internet Res 2010 Dec 19;12(5):e50 [FREE Full text] [doi: 10.2196/jmir.1394] [Medline: 21169178]
165. White L, Delaney R, Pacifici C, Nelson C, Dickinson SL, Golzarri-Arroyo L. Understanding and parenting children's noncompliant behavior: the efficacy of an online training workshop for resource parents. Child Youth Serv Rev 2019 Apr;99:246-256 [FREE Full text] [doi: 10.1016/j.childyouth.2019.01.045] [Medline: 31130765]
166. Yap MB, Martin PD, Jorm AF. Online parenting guidelines to prevent adolescent depression and anxiety: evaluating user characteristics and usefulness. Early Interv Psychiatry 2018 Oct;12(5):951-958. [doi: 10.1111/eip.12503] [Medline: 29076246]

XSL•FO
RenderX

167. Zlotnick C, Tzilos Wernette G, Raker CA. A randomized controlled trial of a computer-based brief intervention for victimized perinatal women seeking mental health treatment. Arch Womens Ment Health 2019 Jun 7;22(3):315-325 [FREE Full text] [doi: 10.1007/s00737-018-0895-1] [Medline: 30088145]
168. Shen S, Doyle-Thomas KA, Beesley L, Karmali A, Williams L, Tanel N, et al. How and why should we engage parents as co-researchers in health research? A scoping review of current practices. Health Expect 2017 Aug;20(4):543-554 [FREE Full text] [doi: 10.1111/hex.12490] [Medline: 27516003]
169. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
170. Bandura A. Social cognitive theory of mass communication. Media Psychol 2001 Aug;3(3):265-299. [doi: 10.1207/S1532785XMEP0303_03]
171. Broomfield G, Wade C, Yap MB. Engaging parents of lower-socioeconomic positions in internet- and mobile-based interventions for youth mental health: a qualitative investigation. Int J Environ Res Public Health 2021 Aug 28;18(17):9087 [FREE Full text] [doi: 10.3390/ijerph18179087] [Medline: 34501675]
172. Broomfield G, Brown SD, Yap MB. Socioeconomic factors and parents' preferences for internet- and mobile-based parenting interventions to prevent youth mental health problems: a discrete choice experiment. Internet Interv 2022 Apr;28:100522 [FREE Full text] [doi: 10.1016/j.invent.2022.100522] [Medline: 35309756]
173. Geraghty AW, Torres LD, Leykin Y, Pérez-Stable EJ, Muñoz RF. Understanding attrition from international internet health interventions: a step towards global eHealth. Health Promot Int 2013 Sep 10;28(3):442-452 [FREE Full text] [doi: 10.1093/heapro/das029] [Medline: 22786673]
174. Lin H, Wu X. Intervention strategies for improving patient adherence to follow-up in the era of mobile information technology: a systematic review and meta-analysis. PLoS One 2014 Aug 6;9(8):e104266 [FREE Full text] [doi: 10.1371/journal.pone.0104266] [Medline: 25100267]
175. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med 2010 Mar 24;7(3):e1000251 [FREE Full text] [doi: 10.1371/journal.pmed.1000251] [Medline: 20352064]
176. Becker BJ. Combining significance levels. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York, NY: Russell Sage Foundation; 1994:216-229.
177. Panter-Brick C, Burgess A, Eggerman M, McAllister F, Pruett K, Leckman JF. Practitioner review: engaging fathers--recommendations for a game change in parenting interventions based on a systematic review of the global evidence. J Child Psychol Psychiatry 2014 Nov;55(11):1187-1212 [FREE Full text] [doi: 10.1111/jcpp.12280] [Medline: 24980187]

Abbreviations

ACE: adverse childhood experience
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by T Leung, G Eysenbach; submitted 01.11.22; peer-reviewed by H Holmen, P Mcgrath; comments to author 05.03.23; revised version received 28.04.23; accepted 30.11.23; published 19.01.24.

Please cite as:
Aldridge G, Tomaselli A, Nowell C, Reupert A, Jorm A, Yap MBH
Engaging Parents in Technology-Assisted Interventions for Childhood Adversity: Systematic Review
J Med Internet Res 2024;26:e43994
URL: https://www.jmir.org/2024/1/e43994
doi:10.2196/43994
PMID:38241066

©Grace Aldridge, Alessandra Tomaselli, Clare Nowell, Andrea Reupert, Anthony Jorm, Marie Bee Hui Yap. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Reporting Quality of Machine Learning Studies on Pediatric Diabetes Mellitus: Systematic Review

Zsombor Zrubka[1], MBA, MD, PhD; Gábor Kertész[2], PhD; László Gulácsi[1], MD, DSC; János Czere[3], MSc; Áron Hölgyesi[1,4], DPharm; Hossein Motahari Nezhad[1,5], PhD; Amir Mosavi[2], PhD; Levente Kovács[6], PhD; Atul J Butte[7], MD, PhD; Márta Péntek[1], MD, DSC

[1]HECON Health Economics Research Center, University Research and Innovation Center, Óbuda University, Budapest, Hungary
[2]John von Neumann Faculty of Informatics, Óbuda University, Budapest, Hungary
[3]Doctoral School of Innovation Management, Óbuda University, Budapest, Hungary
[4]Doctoral School of Molecular Medicine, Semmelweis University, Budapest, Hungary
[5]Doctoral School of Business and Management, Corvinus University of Budapest, Budapest, Hungary
[6]Physiological Controls Research Center, University Research and Innovation Center, Óbuda University, Budapest, Hungary
[7]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, United States

Corresponding Author:
Zsombor Zrubka, MBA, MD, PhD
HECON Health Economics Research Center
University Research and Innovation Center
Óbuda University
Bécsi út 96/b
Budapest, 1034
Hungary
Phone: 36 302029415
Email: zrubka.zsombor@uni-obuda.hu

Abstract

Background: Diabetes mellitus (DM) is a major health concern among children with the widespread adoption of advanced technologies. However, concerns are growing about the transparency, replicability, biasedness, and overall validity of artificial intelligence studies in medicine.

Objective: We aimed to systematically review the reporting quality of machine learning (ML) studies of pediatric DM using the Minimum Information About Clinical Artificial Intelligence Modelling (MI-CLAIM) checklist, a general reporting guideline for medical artificial intelligence studies.

Methods: We searched the PubMed and Web of Science databases from 2016 to 2020. Studies were included if the use of ML was reported in children with DM aged 2 to 18 years, including studies on complications, screening studies, and in silico samples. In studies following the ML workflow of training, validation, and testing of results, reporting quality was assessed via MI-CLAIM by consensus judgments of independent reviewer pairs. Positive answers to the 17 binary items regarding sufficient reporting were qualitatively summarized and counted as a proxy measure of reporting quality. The synthesis of results included testing the association of reporting quality with publication and data type, participants (human or in silico), research goals, level of code sharing, and the scientific field of publication (medical or engineering), as well as with expert judgments of clinical impact and reproducibility.

Results: After screening 1043 records, 28 studies were included. The sample size of the training cohort ranged from 5 to 561. Six studies featured only in silico patients. The reporting quality was low, with great variation among the 21 studies assessed using MI-CLAIM. The number of items with sufficient reporting ranged from 4 to 12 (mean 7.43, SD 2.62). The items on research questions and data characterization were reported adequately most often, whereas items on patient characteristics and model examination were reported adequately least often. The representativeness of the training and test cohorts to real-world settings and the adequacy of model performance evaluation were the most difficult to judge. Reporting quality improved over time (r=0.50; P=.02); it was higher than average in prognostic biomarker and risk factor studies (P=.04) and lower in noninvasive hypoglycemia detection studies (P=.006), higher in studies published in medical versus engineering journals (P=.004), and higher in studies

XSL•FO
RenderX

sharing any code of the ML pipeline versus not sharing (*P*=.003). The association between expert judgments and MI-CLAIM ratings was not significant.

Conclusions: The reporting quality of ML studies in the pediatric population with DM was generally low. Important details for clinicians, such as patient characteristics; comparison with the state-of-the-art solution; and model examination for valid, unbiased, and robust results, were often the weak points of reporting. To assess their clinical utility, the reporting standards of ML studies must evolve, and algorithms for this challenging population must become more transparent and replicable.

(J Med Internet Res 2024;26:e47430) doi:10.2196/47430

KEYWORDS

diabetes mellitus; children; adolescent; pediatric; machine learning; Minimum Information About Clinical Artificial Intelligence Modelling; MI-CLAIM; reporting quality

Introduction

Background

In recent years, the regulatory authorization of medical devices and digital health technologies based on big data and machine learning (ML) has accelerated [1]. ML solutions have the potential to transform clinical practice by automating diagnosis, enhancing clinical decision-making, improving patient monitoring, and personalizing treatment [2]. Diabetes care has been among the first clinical areas to adapt ML technologies [1,2].

Pediatric Diabetes Mellitus

Worldwide, pediatric diabetes mellitus (DM) is one of the most common chronic conditions among children, with growing incidence and increasingly complex presentation [3-6]. Type 1 DM (T1DM) is characterized by the lack of insulin secretion mainly due to autoimmune etiology, whereas type 2 DM (T2DM) is characterized by insulin resistance and metabolic syndrome associated with obesity [5]. Approximately 20% of children have both autoimmunity and insulin resistance [5]. A less common form, maturity-onset DM of the young, is attributed to a monogenic hereditary background [7,8]. Pediatric DM represents a difficult-to-treat population. Glucose targets frequently remain unmet in children and adolescents [9], and chronic complications, such as kidney disease, retinopathy, neuropathy, or hypertension, affect a significant proportion of patients by reaching young adulthood [10,11]. The life expectancy and quality of life of patients with pediatric DM may be reduced to a varying degree [5].

ML in Pediatric DM

Within DM, pediatric DM has been leading the way to adopt digital technologies and intelligent devices [1,12]. Continuous glucose monitoring and automated insulin delivery systems are becoming an essential part of the management of children and adolescents with DM, with superior outcomes compared with alternative treatments [9,13]. Regulated smartphone apps for insulin dosing have become available for pediatric patients [14]. However, none of the currently available algorithms are optimal, and despite the use of advanced technology, many pediatric patients live under suboptimal glycemic control and are at risk of potentially serious long-term consequences [9,15]. In the quest for better disease characterization, prevention, and treatment of pediatric DM, ML has been increasingly applied from glucose sensors and artificial pancreas systems [16,17] to disease management apps (eg, mobile apps for food image–based carbohydrate counting or supporting self-management) [18,19] or risk prediction algorithms [20]. Although the comparison of novel algorithms has been challenging owing to methodological heterogeneity [9], artificial intelligence (AI) or ML algorithms have not been covered by DM technology guidelines [13,21,22]. To meet the needs of this challenging population, it is of utmost importance that algorithms are transparent and that their clinical value can be assessed.

Reporting Quality of ML Studies in Medicine

Although expectations about the potential of technology to improve disease outcomes of pediatric DM have been rising [23,24], there has been growing concern about the transparency, replicability, biasedness, and overall validity of research in the field of AI and ML [25-30]. Indeed, examples of flawed or unfair predictions by algorithms and consequent legislative changes have sparked debate about the explainability, interpretability, and understandability of “black box” systems in ethical [31], philosophical [32,33], legal [34], social [35,36], computer [37], or medical sciences [38,39]. Although the concepts themselves remain vaguely defined or conflated [32], users in general and health care professionals in particular have sought explainable and interpretable ML models instead of predictions made by “black box” systems [40,41]. Although some ML models are “transparent” and others are “opaque” by nature, several post hoc techniques have been developed to make results interpretable, that is, to help medical professionals understand how and why machine decisions were made [40,41].

The outputs of ML models are probably more dependent on the input data than on the algorithm [42]. Therefore, the assessment of the fairness and accuracy of clinical ML studies should involve a thorough understanding of the processes of data production throughout the entire life cycle, from collection to annotation and processing [43]. Although technical aspects of data provenance may surpass the needs of clinicians, detailed reporting of the sources and production of data are indispensable for transparent and reproducible ML research and development [43].

We argue that high reporting quality standards are a prerequisite for the assessment and ultimately the achievement of methodological excellence in biomedical research. Although the association between incomplete reporting and biased treatment effects has been shown in clinical trials [44], evidence

XSL•FO
RenderX

supports the positive effect of using checklists on reporting quality [45,46]. Despite the growing number and widespread adoption of reporting guidelines by leading journals over the past decades, deficient reporting of medical research studies remains a major concern, producing considerable waste [47].

Recognizing the limited usefulness of poorly reported studies in clinical practice, a plethora of reporting guidelines have been proposed for ML studies in medicine. Checklists have been developed for different study types (eg, observational studies, randomized trials, and health economic evaluations) and clinical areas aimed at standardizing the mandatory elements to be included in the study reports. These checklists are increasingly being used by scientific journal editors in medicine as mandatory elements for the submission of a manuscript [48]. Although targeting different apps and audiences, most reporting guidelines aim to ensure that results are reproducible, transparent, and, where appropriate, provide sufficient detail for inclusion in future evidence syntheses [26,30,48,49]. Hence, guidelines may contribute to the adoption of technologies with potential to benefit patients in real-world clinical settings.

Research Aims

Various apps and methods of ML in diabetes care have been systematically reviewed [50,51], including specific use cases such as the prediction of hypoglycemia [52] or complications [53,54], diagnosis [55,56], use in disease management [24,57], and smart devices [58]. However, to the best of our knowledge, the reporting quality of ML studies in pediatric DM has not been systematically reviewed.

By acknowledging the potential of ML methods in addressing the specific treatment challenges of pediatric DM, we aimed to systematically review the reporting quality of ML studies on pediatric DM using a structured reporting checklist. Specifically, we aimed to highlight areas with adequate or poor reporting quality and identify the indicators of reporting quality. Furthermore, we explored the association of reporting quality with expert judgments about the overall clinical usefulness of the reported results.

Methods

Database Search

We considered the updated PRISMA (Preferred Reporting Item for Systematic Reviews and Meta-Analyses) 2020 statement when reporting the results of our study [59] (Multimedia Appendix 1). We searched the PubMed and Web of Science databases for the 5-year period from January 1, 2016, to December 31, 2020, using search syntaxes that combined the terms ML, children, and DM. For ML, we constructed a comprehensive search filter using the Medical Subject Headings (MeSH) terms of ML and AI [60,61]. We extended the search phrase with a list of terms from the caret package [62]. Given the rapidly expanding list and specialized use of methods, terms were added based on expert judgment. For studies on children, we adapted the Cochrane child search filter [63] by removing terms related to infants who were outside the scope of our study. In addition to the MeSH terms, the DM filter also included hyperglycemia, hypoglycemia, ketoacidosis, and insulin resistance. The detailed syntaxes and dates of the search in the PubMed and Web of Science databases are provided in Multimedia Appendix 2 and Multimedia Appendix 3, respectively.

Screening and Selection of Eligible Studies

Original research reports published from 2016 to 2020 were eligible if ML methods were applied to analyze patient data on a population of children aged 2 to 18 years with DM of any subtype. As the primary research goal concerned the reporting quality of the applied ML methods, outcomes and interventions were not specified among the eligibility criteria. We restricted our review to a 5-year window to keep track of recent advances and maintain a feasible range. We included studies if DM or its complications (eg, retinopathy) were the primary diagnosis or DM was a study subpopulation (eg, population screening studies). If the relevant age group was covered, patients aged up to 25 years were accepted. We also included studies involving broader age groups if the results were reported separately for children. Both in vivo and in silico pediatric patients were allowed. No language restrictions were applied.

All records were independently screened by 12 pairs of reviewers formed by 6 authors. An extensive list of ML methods was provided to aid in record screening. Differences were resolved by consensus. Records were excluded if ineligibility could be clearly stated and retained for full-text screening in case of uncertainty or insufficient information.

The full-text reports were independently screened by 12 pairs of reviewers. All eligibility criteria were recorded and had to be reconciled in case of disagreement. Eligibility for the ML criterion was as follows: (1) either a typical ML method was specified (eg, bagging, boosting, bootstrap aggregated models, decision tree, deep belief network, denoising autoencoder, ensemble methods, genetic programming, learning, long short-term memory, model tree, neural network, neuro-fuzzy, random forest, random tree, and support vector) or (2) the data analysis algorithm involved the ML workflow of training, validation, and testing of results using any algorithm including traditional regression methods. In case of uncertainty, a third reviewer (AM) with technical expertise in ML methods made the decision. The third reviewer was omitted only if reviewers mutually agreed on the presence of criterion 1 or the absence of both criterion 1 and 2. Interrater agreement of reviewers was monitored during the screening of records and selection full-text reports via absolute agreement and Cohen κ.

Assessment of Reporting Quality

Minimum Information About Clinical AI Modelling Checklist

Overview

Given the potentially diverse application of ML in pediatric DM, we applied the Minimum Information About Clinical Artificial Intelligence Modelling (MI-CLAIM) checklist, a general-purpose reporting guideline for medical AI studies available from the EQUATOR (Enhancing the Quality and Transparency of Health Research) network [64]. MI-CLAIM has been developed to enable the assessment of clinical impact

XSL•FO
RenderX

(including fairness and bias) and the replication of the technical design process of clinical ML studies. It comprises 17 binary "yes" or "no" items organized into 6 domains: study design (part 1, consisting of 5 items), data and model optimization (parts 2 and 3, consisting of 5 items), model performance (part 4, consisting of 3 items), model examination (part 5, consisting of 5 items), and reproducibility (part 6, consisting of 1 item). Two categorical items ask about the type of data and reproducibility of the entire model pipeline. The data types can be categorized as structured (ie, that can be defined and understood by researchers) or unstructured (ie, the lack of explicitly definable raw features, such as images, natural language, or time series). The reproducibility of the entire model-building pipeline is described by 4 levels: tier 1 (ie, complete sharing of the code), tier 2 (ie, allowing a third party to evaluate the code for accuracy and fairness and share the results of this evaluation), tier 3 (ie, release of a virtual machine for running the code on new data without sharing its details), and tier 4 (no sharing). In this paper, we will refer to MI-CLAIM items by denoting the domain and item number within the domain (eg, item 6.1 denotes reproducibility). To aid the reporting and review process, the MI-CLAIM checklist requires the recording of the page numbers of a paper where relevant information was found concerning the checklist items. To address missing information or inadequate reporting, notes must be taken for each item [64].

Assessment of Reporting Quality Using MI-CLAIM

In this study, MI-CLAIM was applied as follows. Six reviewers were organized into pairs involving a medical expert (ÁH, LG, MP, and ZZ) and an expert in computer science (GK, HMN, and JC), who independently evaluated eligible studies along all items of MI-CLAIM. MI-CLAIM was elaborated in group training sessions before commencing the assessment. For the 17 binary items, response options were "yes" or "no" and "unsure" for cases when information was provided by the authors, but the reviewers could not come to a firm conclusion about whether the reporting was sufficiently clear or detailed. The final ratings relied on the expert judgment and consensus of the involved reviewer pairs. For unanimous "yes" answers, supporting information from the papers was extracted, summarized, and provided in supplementary tables, but no comments were made about missing items or inadequate information. Only those studies were eligible for the evaluation via MI-CLAIM, which followed the typical ML workflow: data-driven model training and validation followed by model testing on a designated data set [64]. Studies in which ML algorithms were used in alternative workflows were not evaluated using MI-CLAIM.

We added two additional items to reflect the overall purpose of MI-CLAIM: (1) "Did the paper enable the direct assessment of clinical impact, including fairness and bias?" (clinical impact) and (2) "Can the technical design process of the paper be rapidly replicated?" (replicability). For all studies, item 1 was rated by a senior medical expert (GL) and item 2 was rated by 2 experts who were well versed in computer science and medical data analysis (GK and ZZ). Responses were captured on a 5-level Likert scale (strongly agree; agree; neither agree nor disagree; disagree; strongly disagree).

Extraction of Additional Data Items

In addition to the assessment of reporting quality, the first author, publication year, Scimago subject category of the publication (engineering and medicine) [65], title, main goal, applied ML method, input data, characteristics of the pediatric training and testing samples, and key findings of the included papers were extracted by a single reviewer (ZZ). Quantitative characteristics were extracted (eg, sample size), and inductive coding was applied to group studies into categories according to their goals, features of the training and test samples, and input data.

Evidence Synthesis

We provided qualitative summaries of the answers to the MI-CLAIM by study and item. For each study, we denoted the count of "yes," "no," and "unsure" answers as well as the categories for data type (item 2.5) and reproducibility (item 6.1) as the MI-CLAIM profile of a study. For quantitative analysis, reproducibility (item 6.1) was dichotomized as "any sharing" (tiers 1-3) and "no sharing" (tier 4). In addition, the count of "yes" ratings for each study was referred to as reporting quality.

We assessed the association between reporting quality and continuous, dichotomous, and polytomous study characteristics using Pearson correlation, 2-sample *t* test, and one-way ANOVA, respectively. The normality of the distribution of reporting quality was tested using the Shapiro-Wilk test. The association between study characteristics and reporting quality (ie, "yes" ratings) was assessed via cross-tabulation and the Fisher exact test for each MI-CLAIM item.

Furthermore, we evaluated the correlation between reporting quality and the overall expert assessment of clinical impact. The association between reproducibility (item 6.1) and overall expert assessment of replicability was tested using a 2-sample *t* test.

Results

Screening and Selection of Eligible Reports

The searches in PubMed and Web of Science databases yielded 717 and 336 records, respectively. After removing 9 duplicates, 1043 publication records were screened, 298 full-text reports were checked for eligibility, and 28 reports were eligible for our review. Research activity increased rapidly over time: 64% (18/28) of papers were published during the last 2 years, and 43% (12/28) were published in 2020 alone. (For web-based papers, the year of publication was subsequently updated in some cases). We assessed 21 reports using MI-CLAIM. Due to initial differences between reviewers, 20.4% (213/1043) of the records were reconciled. The absolute agreement and Cohen κ between reviewers' initial judgments were 79.6% (range 63%-94%) and 0.47 (range 0.15 to 0.83), respectively. In the screening of full-text reports, 12 could not be retrieved due to lack of institutional access (Multimedia Appendix 4), 286 were assessed, 24.1% (69/286) were reconciled between reviewers, and in 13.6% (39/286) of the cases, a third reviewer was invited after the reviewers' discussion to decide on the ML criterion. Altogether, the absolute agreement and κ of reviewers' initial judgments regarding full-text selection were 67% (range

XSL•FO
RenderX

20%-85%) and 0.30 (range 0.06-0.30), respectively (Multimedia Appendix 5). Details of the search, screening, and inclusion are provided in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of included studies. MI-CLAIM: Minimum Information About Clinical Artificial Intelligence Modelling.

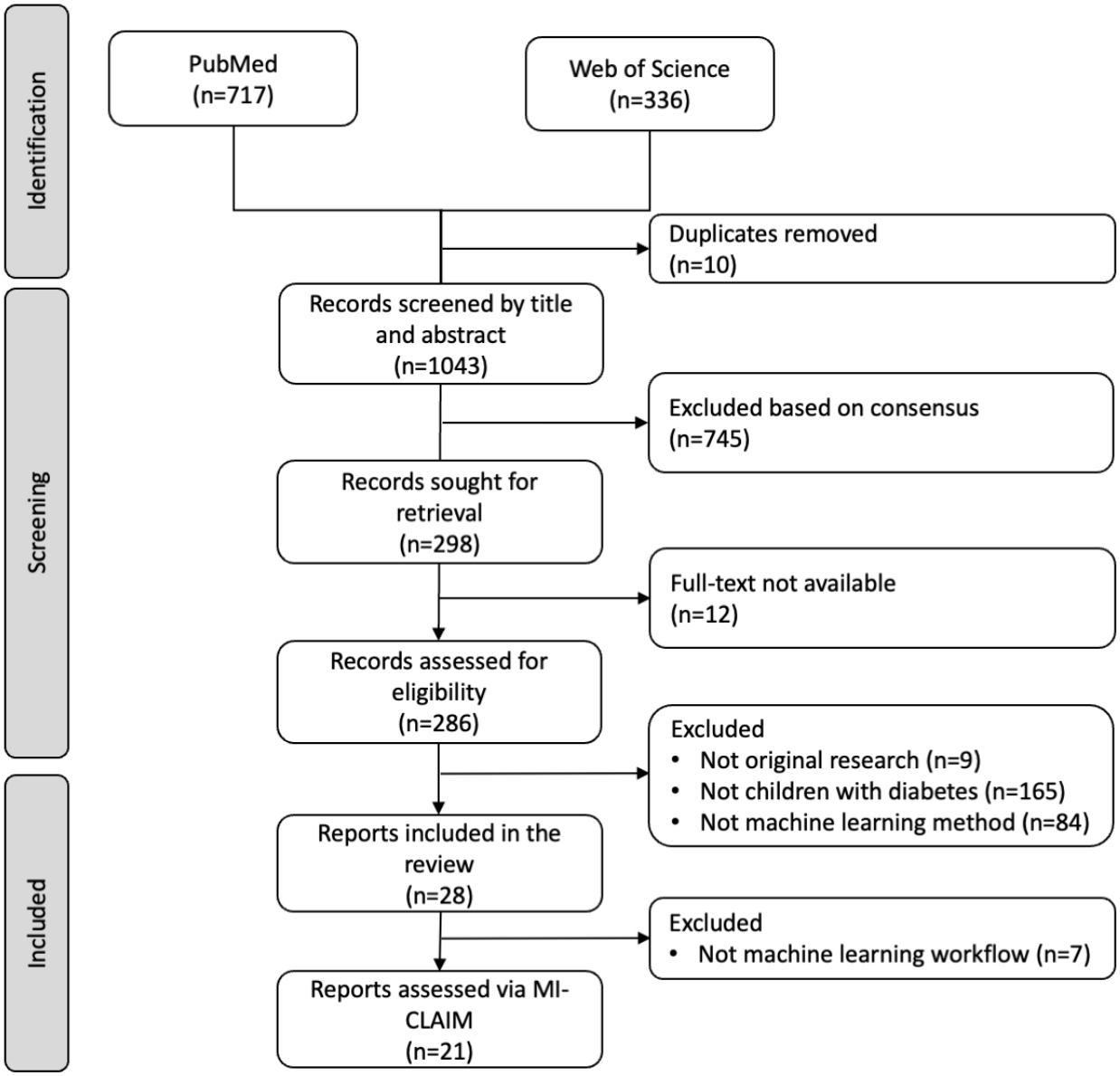

Characteristics of Included Studies

The characteristics of the 28 included studies are summarized in Table 1.

XSL•FO
RenderX

Table 1. Summary of included studies.

Study	Goal	ML[a] method	Training sample	Test sample	Key finding
ML methods applied with an ML workflow (reporting quality was assessed via MI-CLAIM[b])					
Daskalaki et al [66], 2016	Develop an enhanced reinforcement learning model for personalized insulin delivery and glucose control: artificial pancreas system for pediatric T1DM[c]	Model-free actor-critic learning algorithm	In silico adolescents and children from a cohort of 28 patients from the UVA-Padova simulator; 2 outlier children were excluded: approximately 10 adolescents assumed and approximately 8 children. Four-day open-loop period followed by 5-d training.	5-d test period after training on the same in silico patient cohort.	In various parameter settings, time in range 77.8%-86.4% for adolescents and 74.8%-80.5% for children.
Ling et al [67], 2016	Detection of HG[d] in patients with T1DM from the ECG[e] signal	ELM-NN[f] vs PSO-NN[g], MR-FIS[h], FIS[i], and MR[j]	8 children (mean age 14.6, SD 1.5 y) with T1DM from a single center monitored overnight for nocturnal HG using CGM[k] for 360-480 min	8 randomly selected patients with T1DM from the same center	On the basis of a linear combination of sensitivity and specificity (γ), ELM-NN was a superior classifier of HG vs other algorithms with a sensitivity of 0.78 and a specificity of 0.60
Miller et al [68], 2016	Determining subgroups of childhood-onset T1DM based on 25-y CVD[l] risk	TSSA[m]	561 participants from the Pittsburgh Epidemiology of Diabetes Complication prospective cohort study; T1DM onset: age <17 y, between 1950 and 1980	Testing via k-fold cross-validation on the same patient cohort (k not reported)	Distinct subgroups for CVD risk exist within the childhood-onset T1DM cohort
Phyo et al [69], 2016	Detection of HG from ECG signal in pediatric T1DM	DBN[n] vs BBNN[o], WNN[p], FFNN[q], and MR	10 children with T1DM monitored 10h overnight for nocturnal HG using CGM from a single center (training and validation set)	5 randomly selected patients with T1DM from the same center	On the basis of a linear combination of sensitivity and specificity (γ), DBN was a superior classifier of HG vs other algorithms with a sensitivity of 0.80 and specificity of 0.50
Ling et al [70], 2017	Detect HG from ECG signal in patients with T1DM	Combinational MR-NLN[r] with HPSOWM[s] vs MR-NLN, NLN[t], WNN, FFNN, and MR	From 15 T1DM children from a single center (12 with HG, mean age 14.6, SD 1.5 y), monitored 360-480 min overnight for nocturnal HG using CGM and ECG, 5/5 randomly selected for training and validation	5 randomly selected patients with T1DM from the same center	HG in children with T1DM can be detected from an ECG signal with a sensitivity of 0.79 and a specificity of 0.54
Siegel et al [21], 2017	Identify HG biomarkers from breath of children with T1DM.	LDA1[u] with brute force feature selection by testing all possible combinations of predictors.	128 breath samples from 56 patients	Leave-one-out cross-validation using the training sample	HG can be predicted from the identified 7 volatile organic compounds with a sensitivity of 0.91 and a specificity of 0.84
Stawiski et al [71], 2018	Predict insulin resistance from sex, BMI, glucose, and lipid parameters in pediatric T1DM	NN[v] (1000 random models) and MARSplines compared vs reference model (linear regression)	Patients with T1DM from a single center (N=252, mean age 14.95, SD 3.2 y); reference data: euglycemic hyperglycemic clamp	Patients with T1DM from the same center (N=63, mean age 15.15, SD 2.9 y)	NN and MARS[w]: better fit (R^2) and accuracy (median error of prediction) vs reference model
Bois et al [72], 2019a	Compare the performance of ML algorithms in the prediction of HG on a 30-min horizon in pediatric T1DM	FFNN, LSTM[x], ELM[y], SVR[z], GP-RBF[aa], GP-DP[ab]	29-d observation of 10 in silico children using the UVA-Padova simulator, randomly split to 50% training and 25% validation data sets	25% randomly selected data sets from the same 10 in silico patients	GP[ac] with dot-product kernel provided lowest RMSE[ad] and best clinical accuracy on CG-EGA[ae] with approximately 99% acceptable predictions for euglycemia

XSL•FO
RenderX

Study	Goal	ML[a] method	Training sample	Test sample	Key finding
Bois et al [73], 2019b	Develop a model to improve the accuracy of long-term (120 min) glucose predictions in pediatric T1DM	A DCP[af] model with parameter estimates using FFNN, GP with dot-product kernel, and ELM vs 2 alternative predictors: ACP[ag] and AWA[ah]	28-d data from 10 in silico T1DM children from UVA-Padova simulator. 75% of data used for training in 4-fold cross-validation	25% of the data set from the same 10 in silico patients used for testing in 4-fold cross-validation	DCP with novel loss function had promising performance in long-term glucose predictions, with improved clinical acceptability (85.5% accurate predictions CG-EGA)
Khusial et al [74], 2019	Develop a plasma screening panel for NAFLD[ai] in children using metabolomic data	Correlation-based feature selection, then SVM feature selection (information gain), then classification by LR[aj], NB[ak], and RF[al]	Sample size approximately 373; 2 of 3 of the total 559 patients aged 2-25 y, NAFLD: 222 (T2DM: 220), control 337 (T2DM: 328) from 3 studies: Emory University Pediatric Liver Biopsy Data repository, SweetBev Trial, and Yale Pediatric NAFLD Cohort	Sample size approximately 186 (1 of 3 out of the same 559 patients), selected randomly	RF predicted best NAFLD from metabolomic and clinical features with AUROC 0.94 (sensitivity of 0.73, specificity of 0.97)
Langner et al [75], 2019	Quantification of subcutaneous and visceral adipose tissue from MRI scans	2 CNN[am] architectures: U-Net and V-Net	Patients with T2DM from the Tellus study (N=45, age 18-80 y)	Patients selected from Beta-JUDO study (N=10, age 10-18 y)	U-Net provided as accurate results in the test sample as literature reports of human operators and outperformed V-Net
Ngo et al [76], 2019	Detect nocturnal HG from EEG signals	BNN[an]	50% of data from 5 adolescents with T1DM (age 12-18 y) from a single-center overnight HG study (139 total episodes, 45 hypoglycemic)	50% of the same 5 adolescents (139 total episodes, 45 hypoglycemic)	Nocturnal HG could be detected from EEG signal with a sensitivity of 0.82 and specificity of 0.52
Stanfill et al [77], 2019	Develop a data preprocessing method (conditional classifier) to make classification algorithms applicable for omics data in matched case-control studies	Conditional LR, NB, SVM[ao] with radial basis function kernel, SVM with linear kernel, RF, LDA2[ap], and RPCLR[aq] were compared	418 case-control pairs from The Environmental Determinants of Diabetes in the Young (TEDDY) study: genetic, lipidomic, metabolomic biomarkers of IA[ar]	5-fold cross-validation repeated 200 times on the same data set	Conditional SVM and NB outperformed LR in the classification of TEDDY data, with the potential to discover new biomarkers for IA
Amar et al [78], 2020	Compare a novel glucose prediction model vs alternative algorithms on real-life and in silico patients with T1DM	GCN1-3[as] vs ARM[at], RF, GBM[au], FC[av]	In silico: 10 adolescents, 10 children (30-d training, 7-d validation) from UVA-Padova simulator; clinical: retrospective CGM data from 141 patients with T1DM from a single center (mean age 13.5, SD 5.2 y), mean (SD) CGM time: 64.4 (46.6) d	In silico: 7 d test on the same cohort; clinical: 4 d test on the same patients	Clinical accuracy 99.3% for 30-min and 95.8% for 60-min glucose predictions using CG-EGA, improved performance vs the standard ARM
Dave et al [79], 2020	Predict HG from glucose sensor data in pediatric T1DM	LR with LASSO[aw] feature selection, and RF classifiers used on 26 extracted features from CGM data	Sample size approximately 78 (70% of data from 112 children) with T1DM (mean age 12.7, SD 4.8 y), 90 d follow-up using Dexcom G6 CGM device	Sample size approximately 34 (30% of data randomly selected from 112 children) in 10 replications	RF identified 30-60 min HG with a sensitivity of >0.91 and specificity of >0.90
Frohnert et al [80], 2020	Prediction of the development of IA and T1DM from genetic, immunologic, metabolomic, and proteomic biomarkers	ROFI-P3 integrative ML: combining optimal classifiers (LR, RF, KNN, LDA[2], SVM, and NB) with iterative feature set selection for best predictive performance	67 children from the Diabetes Autoimmunity Study in the Young (DAISY) cohort, 22 with T1DM, 20 with persistent IA and 25 from control	The model was tested via 5-fold cross-validation on the training sample	Predictors for IA (ROC AUC[ax] 0.91) and T1DM (ROC AUC 0.92) were identified and should be further validated

XSL•FO
RenderX

Study	Goal	ML[a] method	Training sample	Test sample	Key finding
Garavelli et al [81], 2020	Identify circulating plasma microRNA with prognostic value for the progression of pediatric T1DM (stratification based on C-peptide loss)	FFNN for classification	150 children with T1DM: recent onset (n=88), 12 mo postdiagnosis (n=32), 24 mo postdiagnosis (n=30) from a single center	Two external validation cohorts: 18 children with recent onset T1DM and 26 children with recent onset T1DM from 2 separate clinical centers	Plasma microRNA can be used for the prognostic stratification of children with T1DM based on C-peptide loss
Li et al [82], 2020	GluNet: personalized 30- to 60-min glucose forecast in T1DM from CGM data, meal, insulin dose (GluNet)	DCNN[ay] vs NN, LVR[az] with exogenous input, ARM with exogenous input, SVR	10 in silico adolescents (UVA-Padova simulator); 90-d observation	The same 10 in silico adolescents, 90-d observation	DCNN improved RMSE, MARD[ba], and time lag of glucose forecast vs alternative methods
Zhu et al [83], 2020	Develop and insulin bolus calculator planned for a smartphone app connected to CGM: artificial pancreas system for pediatric T1DM	DRL[bb] vs standard insulin bolus calculator	10 in silico adolescents (UVA-Padova simulator), long-term general training, 180 d of personalized training	Same as the training cohort: after personalized training, 90 d testing with identical parameter settings for the test and control cohorts	DRL improved time in range (61.6%) vs the standard insulin bolus calculator (54.9%)
Zhu et al [84], 2020	Single-hormone (insulin) and dual-hormone (insulin+glucagon) dosing algorithm for closed-loop glucose control of T1DM	Dual-hormone DRL vs single hormone DRL vs state-of the art LGS[bc] control strategy	10 in silico adolescents (UVA-Padova simulator), 1500 d generalized training followed by 30 d personalized training	Testing personalized models in the same cohort for 90 d (10 in silico adolescents)	Best time in range for dual-hormone DRL (78.8%) followed by single hormone DRL (65.9%) followed by LGS (55.5%)
Webb-Robertson et al [85], 2021	Predicting the development of IA and T1DM from metabolomic markers	Integrative ML: product of posterior probability and ROFI[bd]	236 children (118 matched pairs with or without islet autoimmunity or T1DM) from the TEDDY study	78 randomly selected children from the TEDDY study	42 predictor markers identified for T1DM, associated with 3 biological pathways
ML methods applied without an ML workflow (reporting quality was not assessed via MI-CLAIM)					
Alfonsi et al [18], 2020	Test the usability and impact on carbohydrate counting accuracy of an ML-based carbohydrate counting app (iSpy) in pediatric T1DM	CNN to identify food images	NR[be]	NR	Good acceptability, improved carbohydrate accuracy, and fewer counting errors and better HbA1c[bf] vs controls
Adabimohazab et al [86], 2016	Explore the role of inflammation in the pathogenesis of insulin resistance among adolescents with obesity	RF used for classification	345 adolescents (mean age 17.5, SD 1.8 y), lean and overweight or obese	NR	No connection was found between low-grade inflammation and the development of insulin resistance
Brugha et al [87], 2018	Predict cystic fibrosis–related DM from CGM data	Hierarchical and k-means cluster analysis	142 children (age <17 y) with cystic fibrosis from a single center using CGM	NR	Greater median and IQR values for glucose fluctuation predict DM or prediabetes in children with cystic fibrosis
Cabrera et al [88], 2018	On the basis of immunoregulatory profiles, identify pediatric T1DM subtypes at clinical onset to predict postonset insulin secretion and responsiveness to abatacept therapy.	RF, hierarchical clustering	116 children with T1DM, within 100 d from diagnosis (62 control, 54 treated with abatacept)	NR	Innate inflammatory bias levels are associated with T1DM progression rate and responsiveness to abatacept.

XSL•FO
RenderX

Study	Goal	ML[a] method	Training sample	Test sample	Key finding
Biassoni et al [89], 2020	Define gut microbial composition of new-onset pediatric T1DM	Supervised classification: RF, Elastic Net (L_1L_2); unsupervised analysis: WGCNA[bg]	New-onset T1DM pts (N=31, mean age 10.3, SD 4.1 y) vs sex-matched healthy controls (N=25, mean age 10.3, SD 4.1 y)	NR	New-onset T1DM has characteristic fecal microbial flora; not known if cause or consequence of autoimmunity
Nimri et al [90], 2020	Compare the performance of an AI-DSS[bh] vs clinicians for insulin dosing in pediatric T1DM	DreaMed Advisor Pro, continuous glucose monitoring device with insulin pump, details of the ML method not provided	NR	NR	Time in range of AI-DSS was statistically noninferior when compared with physicians

XSL•FO
RenderX

Study	Goal	ML[a] method	Training sample	Test sample	Key finding
Wolf et al [91], 2020	Economic evaluation of AI-based diabetic retinopathy screening vs standard care in pediatric DM from the patient perspective	Artificial intelligence–based diagnosis of diabetic retinopathy from digital fundus image. Methods not described in the paper	NR	NR	With >23% adherence to screening recommendations, artificial intelligence–based diabetic retinopathy screening is cost-saving for patients

[a]ML: machine learning.
[b]MI-CLAIM: Minimum Information About Clinical Artificial Intelligence Modelling.
[c]T1DM: type 1 diabetes mellitus.
[d]HG: hypoglycemia.
[e]ECG: electrocardiogram.
[f]ELM-NN: extreme learning-based feed forward neural network.
[g]PSO-NN: particle swarm optimization based neural network.
[h]MR-FIS: multiple regression–based fuzzy inference system.
[i]FIS: fuzzy inference system.
[j]MR: multiple regression.
[k]CGM: continuous glucose monitor.
[l]CVD: cardiovascular disease.
[m]TSSA: tree-structured survival analysis.
[n]DBN: deep belief neural network.
[o]BBNN: block-based neural network.
[p]WNN: wavelet neural network.
[q]FFNN: feed forward neural network.
[r]MR-NLN: multiple regression–based neural logic network.
[s]HPSOWM: hybrid particle swarm optimization with wavelet mutation.
[t]NLN: neural logic network.
[u]LDA1: linear discriminant analysis.
[v]NN: neural network.
[w]MARS: multivariate adaptive regression splines.
[x]LSTM: long short-term memory.
[y]ELM: extreme learning machine.
[z]SVR: support vector regression.
[aa]RBF: radial basis function.
[ab]GP-DP: Gaussian process regression with dot-product kernel.
[ac]GP: Gaussian process regression.
[ad]RMSE: root mean square error.
[ae]CG-EGA: continuous glucose error grid analysis (Clarke error grid).
[af]DCP: derivatives combination predictor.
[ag]ACP: artificial neural network combination predictor.
[ah]AWA: adaptive weighted average fusion algorithm.
[ai]NAFLD: nonalcoholic fatty liver disease.
[aj]LR: logistic regression.
[ak]NB: naive Bayes.
[al]RF: random forest.
[am]CNN: convolutional neural network.
[an]BNN: Bayesian neural network.
[ao]SVM: support vector machine.
[ap]LDA2: latent Dirichlet allocation.
[aq]RPCLR: random penalized conditional logistic regression.
[ar]IA: islet autoimmunity.
[as]GCN: gradually connected neural network.
[at]ARM: autoregressive model.

XSL•FO
RenderX

[au]GBM: gradient boosting machine.
[av]FC: fully connected neural network.
[aw]LASSO: least absolute shrinkage and selection operator.
[ax]ROC AUC: receiver operating characteristic area under curve.
[ay]DCNN: dilated convolutional neural network.
[az]LVR: latent variable regression.
[ba]MARD: mean absolute relative difference.
[bb]DRL: deep reinforcement learning.
[bc]LGS: low glucose insulin suspension.
[bd]ROFI: repeated optimization for feature interpretation.
[be]NR: not reported.
[bf]HbA1c: hemoglobin A1c.
[bg]WGCNA: weighted correlation network analysis.
[bh]AI-DSS: AI-based decision support system.

Most studies (n=6) focused on the discovery of etiologic or prognostic biomarkers of T1DM [77,80,81,85,88,89], followed by studies aiming to predict hypoglycemia using noninvasive methods such as ECG or EEG signals, breath volatile organic compounds (n=5) [21,67,69,70,76], insulin bolus calculators for closed-loop glucose control (n=5) [66,82-84,90], accurate prediction of glucose levels or hypoglycemia from continuous glucose monitor (CGM) data (n=4) [72,73,78,79], etiologic or risk factors for insulin resistance or T2DM (n=4) [71,74,75,86], and other goals such as long-term cardiovascular risk stratification [68], accurate carbohydrate counting via a smartphone app [18], prediction of cystic fibrosis–related DM from CGM signal [87], and the economic evaluation of diabetic retinopathy screening via AI versus standard care [91].

Reporting the Training and Test Samples

The characteristics of the training sample (including the validation data set) were not reported in 3 studies [18,90,91]. Of the 25 studies reporting details, in 18 (72%), the training sample involved human patients, in 6 (24%) only in silico patients [66,72,73,82-84], and there were both human and in silico patients in one study [78]. Of the 9 studies focusing on bolus calculation or glucose prediction from CGM data, only 2 (22%) had human patients in the training sample [78,79], 6 (67%) had only in silico patients [66,72,73,82-84], and the training sample was not characterized in one study [90]. The size of the training sample ranged from 561 [68] to 5 [76]. Of the 25 studies reporting details, only 10 (48%) had training samples with more than 100 patients [68,71,74,77,78,81,85-88], whereas 9 (36%) involved 10 or fewer patients [67,69,70,72,73,76,82-84]. Etiologic or prognostic studies for T1DM [77,80,81,85,88,89] and T2DM [71,74,75,86] featured the largest training samples ranging between 56 and 418 and 45 and 373, respectively. However, of the 14 studies focusing on insulin bolus calculators, or glucose or hypoglycemia prediction, only 4 (29%) [21,66,78,79] involved more than 10 patients in the training sample, with 141 being the largest sample size [78].

The test sample was characterized in 21 studies that involved a full ML workflow. Testing was performed on the same patients as the training in 12 (57%) studies, involving time-split in 8 studies using CGM data [66,72,73,76,78,82-84] and cross-validation in 4 studies [21,68,77,80]. In 5 (24%) studies, the test sample involved randomly selected patients from the same center as the training sample [67,69-71,79]; in 2 (10%) studies, patients from one or more external centers [75,81]; and in 2 (10%) studies randomly selected patients from multicenter studies [74,85]. The test sample included 10 or fewer patients in 10 (48%) studies [67,69,70,72,73,75,76,82-84], and only 4 (19%) studies featured test samples involving more than 100 patients [68,74,77,78]. The sample sizes of the external test samples ranged between 10 and 186 [74,75,81,85].

ML Methods Used

In the 28 included studies, we identified a plethora of ML methods that sometimes overlapped with the techniques used for feature engineering, dimension reduction, or other steps of the analysis. Altogether, from the 87 proposed or comparator methods used, we identified 61 different techniques, with random forest (RF) mentioned in 8 studies [74,77-80,86,88,89], followed by feed forward neural network in 5 studies [69,70,72,73,81], logistic regression in 4 [74,77,79,80], and multiple regression (MR) [67,69,70] and naive Bayes in to 3-3 studies [77,80]. There were 48 methods mentioned in only one paper, and 2 papers did not specify the ML algorithm [90,91].

Assessment of Reporting Quality Using the MI-CLAIM Profiles

The reporting quality via MI-CLAIM was assessed in 21 studies that followed the ML workflow. The characteristics of the assessed studies are summarized in Multimedia Appendix 6 [21,66-85]. The MI-CLAIM profiles for each study are shown in Figure 2 [21,66-85]. The assessment details of the included studies are provided in Multimedia Appendix 7 [21,66-85]. To support our “yes” ratings, the summaries of reported items for each study by the domains of MI-CLAIM are provided in Multimedia Appendices 8-11 [21,66-85].

XSL•FO
RenderX

Figure 2. Minimum Information About Clinical Artificial Intelligence Modelling (MI-CLAIM) profiles by studies. MI-CLAIM profile: the count of "yes," "no," and "unsure" ratings for items 1.1 to 1.5, 2.1 to 2.4, 4.1 to 4.3, and 5.1 to 5.5; the classification of data type (item 2.5) as "structured" or "unstructured" and the classification of reproducibility (item 6.1) into tier 1 to 4 or "unsure.".

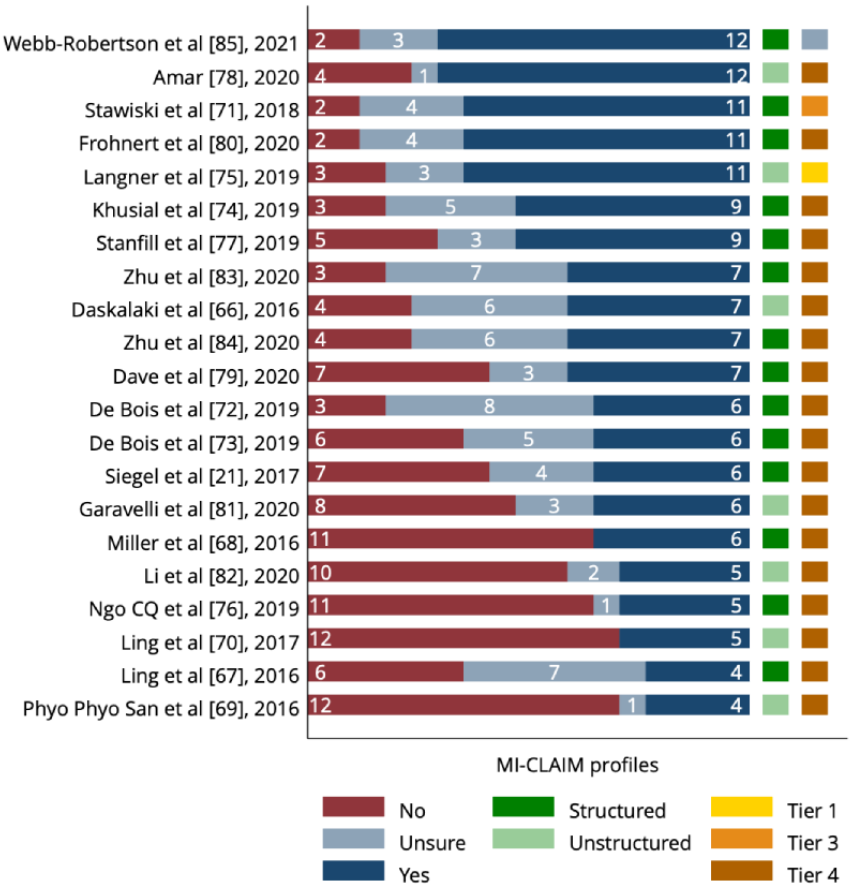

For the 17 binary items, the number of "yes" ratings ranged between 4 and 12 (mean: 7.43), the "unsure" ratings ranged between 0 and 7 (mean 3.62), and the "no" ratings ranged between 2 and 12 (mean 6.95). One study provided a link to the applied unique ML model framework [85], without disclosing the specific code applied in the analysis. We rated reproducibility (item 6.1) for this study as "unsure" and dichotomized it as "any sharing." The distribution of reporting quality was normal (Shapiro-Wilk test; P=.11). Reporting quality correlated positively with the year of publication, showing an improvement over time (r=0.50; P=.02). The reporting quality did not differ between studies using structured [21,67,68,70-74,76,77,79,80,83-85] or unstructured data [66,69,70,75,78,81,82] (t_{19}=0.35; P=.73), or whether the input data were time series [66,69,70,72,73,78,79,82-84], omics [18,21,74,77,80,81,85], or other [68,71,75,76] (ANOVA $F_{2,18}$=2.21; P=.14), in silico [66,72,73,82-84] or human subjects were involved [21,67-71,74-81,85] (t_{19}=1.23; P=.24). However, it differed between studies with different research goals (ANOVA, $F_{5,15}$=4.59; P=.01). Compared with the mean, prognostic biomarker studies in T1DM and risk factor studies of T2DM had significantly higher reporting quality by 2.02 (P=.04) and 2.85 (P=.01) "yes" ratings, respectively, whereas noninvasive hypoglycemia detection studies had fewer "yes" ratings by 2.69 (P=.006). Furthermore, higher reporting quality was observed in studies sharing any code of the ML pipeline [71,75,85] (t_{19}=3.48; P=.003) and in studies published in medical journals [21,66,68,71,74,75,78-81,84,85] (t_{19}=3.24; P=.004). Reporting quality correlated moderately with the overall expert assessment of clinical impact without significant association (r=0.40; P=.07).

Assessment of Reporting Quality by the Items of MI-CLAIM

Figure 3 shows an assessment of reporting quality by the MI-CLAIM items [21,66-85].

XSL•FO
RenderX

Figure 3. Assessment of reporting quality by Minimum Information About Clinical Artificial Intelligence Modelling (MI-CLAIM) items. The wording of MI-CLAIM items are as follows. Study design (part 1)—1.1: the clinical problem in which the model will be used is clearly detailed in the paper; 1.2: the research question is clearly stated; 1.3: the characteristics of the cohorts (training and test sets) are detailed in the text; 1.4: the cohorts (training and test sets) are shown to be representative of real-world clinical settings; 1.5: the state-of-the-art solution used as a baseline for comparison has been identified and detailed. Data and optimization (parts 2 and 3)—2.1: the origin of the data is described and the original format is detailed in the paper; 2.2: transformations of the data before they are applied to the proposed model are described; 2.3: the independence between training and test sets has been proven in the paper; 2.4: details on the models that were evaluated and the code developed to select the best model are provided; 2.5: is the input data type structured or unstructured? Model performance (part 4)—4.1: the primary metric selected to evaluate algorithm performance (eg, area under the curve and F score), including the justification for selection, has been clearly stated; 4.2: the primary metric selected to evaluate the clinical utility of the model (eg, positive predictive value and number needed to treat), including the justification for selection, has been clearly stated; 4.3: the performance comparison between the baseline and proposed model is presented with the appropriate statistical significance. Model examination (part 5)—5.1: examination technique 1a; 5.2: examination technique 2a; 5.3: a discussion of the relevance of the examination results with respect to model or algorithm performance is presented; 5.4: a discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented; 5.5: a discussion of the reliability and robustness of the model as the underlying data distribution shifts is included. Reproducibility (part 6)—choose the appropriate tier of transparency.

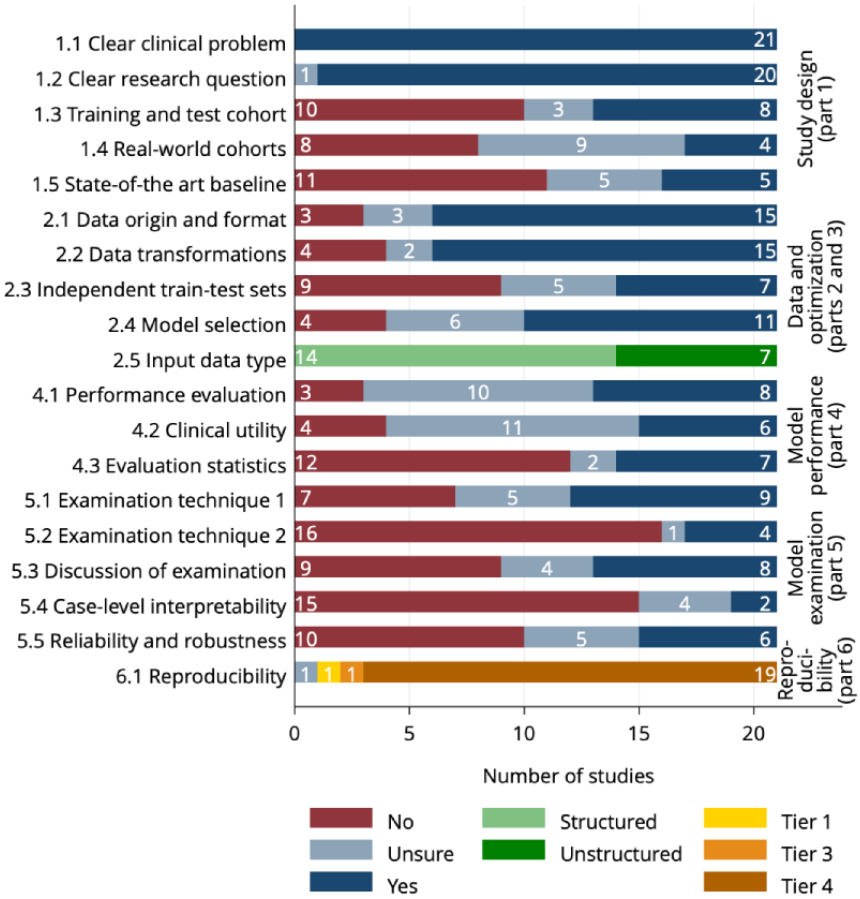

Study Design (Part 1)

The clinical problem (item 1.1) was clearly defined in all studies, and the research question was clearly stated (item 1.2) in nearly all cases. However, in 1 (5%) study [79], our rating was uncertain about the clarity of the main research question. The characteristics of the training and test cohorts were clearly described (item 1.3) in 8 (38%) studies [68,71-74,78,81,85], our rating was "unsure" in 3 (14%) studies [75,83,84]. The cohorts were poorly characterized in all insulin bolus calculator and noninvasive hypoglycemia detection studies, receiving only "no" ratings (Fisher exact test; *P*=.03). Cohort characteristics were reported similarly in silico and in human studies. Our ratings were rather uncertain about whether the representativeness of the cohorts in real-world clinical settings was demonstrated (item 1.4). Only 4 (19%) studies received a "yes" rating [68,71,78,81], involving human sample sizes ranging between 140 and 561. Nine studies were rated as "unsure" (43%) [66,67,72,74,80,82-85], and in 8 (38%) studies, the representativeness of the sample in real-world situations was not demonstrated [21,69,70,73,75-77,79]. The state-of-the-art solution was not included as baseline (item 1.5) in 11 (52%) studies [21,66,68,71,72,75,76,79,81,82,85], and we were uncertain in 5 (24%) studies [67,69,73,74,78].

Data and Optimization (Parts 2 and 3)

Data origin and format (item 2.1) and data transformations (item 2.2) were described in detail in 15 (71%) studies [21,68,71-77,79,80,82-85] and 15 (71%) studies [66-68,71-75,77-79,81,82,84,85] with "unsure" rating in 3 (14%) [66,67,81,89] and 2 (10%) studies [80,83], respectively. The independence of the training and test samples was proven (item 2.3) in 7 (33%) studies [67,70,71,74,75,81,85], and we were "unsure" in 5 (24%) studies [72,73,82-84]. Reporting on this

XSL•FO
RenderX

item differed according to the study aim (Fisher exact test; *P*=.04). The independence of the test and training samples was not demonstrated in studies focusing on insulin bolus calculation [66,82-84] or hypoglycemia detection from CGM data [72,73,78,79]. Item 2.4 requires that the details of the evaluated models and the code to select the best model are detailed. For this item, 11 (52%) studies received "yes" [21,71-73,76,78-80,83-85], and 6 studies received "unsure" (29%) rating [66,67,74,75,77,81]. The data were structured in 14 (67%) studies and unstructured in 7 (33%) studies. Our reviewers rated the 8 studies using glucose time-series data as structured in 5 (63%) and unstructured in 3 (38%) cases. This suggests that our reviewers' notions varied regarding the direct interpretability of glucose time-series data (ie, as structured data were defined by the authors of MI-CLAIM) [64].

Model Performance (Part 4)

The primary metric for the evaluation of model performance was clearly stated and justified (item 4.1) in 8 (38%) studies [21,69-71,77,78,80,85], whereas our rating was frequently "unsure" (10/21, 48%) [66,67,72,74-76,79,81,83,84]. The reporting in this item differed between human and in silico studies (Fisher exact test; *P*=.046), with no in silico studies rated as "yes." The selection and justification of the primary metric of clinical utility (item 4.2) received 6 (29%) "yes" [69,75,76,78,80,85], and 11 (52%) "unsure" ratings [21,66,67,71-74,77,79,83,84]. The performance comparisons were presented with appropriate statistical significance in only 7 (33%) studies [21,77,78,80,82-84], whereas the rating was "no" in over half of the studies (12/21, 57%) [66,68-70,72-76,79,81,85].

Model Examination (Part 5)

Model examination techniques help to validate that model accuracy is related to relevant inputs and explain how complex models work (eg, quantify the importance of predictor variables or characterize subjects with the best or poorest model performance) [64]. While one examination technique (item 5.1) was applied in 9 (43%) studies [66,71,74,75,77-80,85], a second examination technique (item 5.2) was applied in only 4 (19%) [75,77,80,85]. The use of at least one examination technique was more frequent among publications in medical journals (Fisher exact test; *P*=.02). Although model examination was not reported in any of the noninvasive hypoglycemia studies (Fisher exact test; *P*=.04), 2 examination techniques were reported only in prognostic biomarker studies in T1DM and risk factor studies of T2DM (Fisher exact test; *P*=.02) and mainly among studies using omics data (Fisher exact test; *P*=.03). MI-CLAIM suggests that examination results are more relevant for better-performing models and should be discussed in the context of model performance (item 5.3) [64], which was carried out in 8 (38%) studies [66,71,74,75,78-80,85], mainly published in medical publications (Fisher exact test; *P*=.005). Furthermore, if other examination techniques are infeasible, the results should be interpreted at the case level (item 5.4), which we found in only 2 (10%) studies [66,75]. Model reliability and robustness to shifts in data distribution was examined in 6 (29%) studies [66,74,75,78,79,83], and our rating was "unsure" in 5 (24%) cases [21,71,72,80,85]. Discussions on reliability and robustness were reported more often in medical publications (Fisher exact test; *P*=.02).

Reproducibility (Part 6)

All but 3 studies received "tier 4" rating, as the code of the ML workflow was not shared. One study shared the full code ("tier 1"), one provided a link to a downloadable calculator ("tier 3") and one study received "unsure" rating. Our overall expert assessment of replicability did not differ between studies with "any sharing" and "no sharing" of the code of the model pipeline (t_{19}=0.945; *P*=.36), suggesting that beyond the proposed tiers of MI-CLAIM, the reported technical details have influenced the replicability judgments.

Discussion

Principal Findings

This systematic review provides insights into reporting quality and, hence, the potential clinical impact of studies applying ML methods in pediatric DM populations. We applied the MI-CLAIM checklist to assess the reporting quality of 21 studies that followed the ML workflow of model training, validation, and testing. In these studies, reporting quality was generally low, with an improving trend over time. The MI-CLAIM items on research questions and data characterization were reported adequately most often, whereas the items on patient characteristics and model examination were reported adequately least often. The representativeness of the training and test cohorts to real-world settings and the adequacy of model performance evaluation were the most difficult to judge. On average, we found adequate reporting for less than half of the MI-CLAIM items, with considerable differences between studies with different research foci. Medical papers had higher reporting quality compared with articles published in engineering journals, mainly because of more elaborate reporting in the model examination domain. The number of MI-CLAIM items with a "yes" rating showed a moderate correlation with the overall assessment of the clinical impact by independent medical experts. We found no association between the reproducibility ratings on MI-CLAIM and independent experts' assessments of the technical replicability of studies.

Comparison With Prior Work

When writing this paper, MI-CLAIM was used in a single review, focusing on ML in dental and orofacial pain management, in which nearly all included papers were rated with "yes" in 13 or more out of the 15 assessed items [92]. Our study showed a less favorable picture with all studies having 12 or less "yes" ratings and two-thirds of studies having 7 or less "yes" ratings out of 17 items. Our findings corroborate the results of previous studies, raising concerns regarding the reporting quality of ML studies [25,27,29].

Elaboration of Findings

Considering the globally increasing burden and serious consequences of pediatric DM [3] and the rapid growth of ML literature over the past years [48], we found few eligible studies, usually involving small patient populations. Of the 28 eligible studies, the training sample involved more than 100 patients in

XSL•FO
RenderX

only 10 cases. In terms of research aims, applied methods, and data types, the studies were diverse.

According to our experience, compared with highly standardized medical papers, such as randomized clinical studies [93], systematic reviews [94], or economic evaluations [95], the reading and interpretation of the involved ML papers was challenging and time-consuming. The focus of MI-CLAIM on the clinical utility of ML modeling may explain the higher reporting quality of medical papers than those published in engineering journals. Still, the general reporting pattern reflected a "data-driven" mindset: after stating the clinical problem and research question, data-related items were reported most thoroughly. However, important details for clinicians, such as the detailed description of patient cohorts, the state-of-the-art clinical solution, and the clinical utility of the proposed models and model examination for valid, unbiased, and robust results were often the weak points of reporting.

We found an association between reporting quality and research goals. Studies with a strong clinical focus, such as those seeking prognostic biomarkers in T1DM and risk factors in T2DM, had higher reporting quality than the more technically oriented studies aimed at detecting hypoglycemia from CGM data or noninvasive methods or developing insulin bolus calculator algorithms. However, some of the reported differences were clearly attributable to the individual styles of the research teams. While reporting quality of the same teams, such as Webb-Robertson et al [77,80,85], Zhu et al [82-84], Bois et al [72,73], or Ling et al [67,69,70] fell in the same range, the number of "yes" ratings differed up to 6 items between different author teams in similar studies focusing on hypoglycemia prediction or the discovery of prognostic markers for T1DM.

Although MI-CLAIM was developed as a general reporting checklist, not as a measurement tool, our attempt to quantify reporting quality provided several learnings. We observed that the applicable ratings in some items depended on the underlying methods. For example, the independence of the training and test data could not be demonstrated in studies aimed at developing individualized glucose control algorithms. In contrast, the use of transparent models and a thorough examination of feature importance are hallmarks of prognostic marker studies, yielding naturally high scores in the model examination domain. Specific research questions, methods, or data types gave rise to specialized ML reporting guidelines for certain clinical fields or research designs [30,48]. Although MI-CLAIM provides a strong strategic framework for the evaluation of a broad range of ML applications, in our opinion, the evaluation of the validity and potential biases of ML studies in specific clinical use cases of pediatric DM, such as insulin bolus calculation or hypoglycemia detection from various physiological signals, would require more specific technical guidance, preferably from the medical profession, who provides the data and ultimately uses the results for clinical decision-making.

Some MI-CLAIM items received a high proportion of "unsure" ratings from our team. For example, item 1.4 suggests that the representativeness of training and test cohorts in real-world clinical settings should be demonstrated. Despite many studies reporting the parameter variability of T1DM simulators or the recruitment methods and characteristics of patient samples from diabetes clinics, we were unsure about what the established criteria were for representative cohorts of pediatric T1DM or T2DM. In addition, items 4.1 and 4.2 frequently received "unsure" ratings, which require that the primary evaluation metrics for model performance and clinical utility are clearly stated and justified. Although some authors "cherry picked" among multiple performance metrics to declare the superiority of the proposed model, in many cases, the assessment of adequate reporting was more challenging. For example, some studies applied the Clarke Error Grid Analysis [96] to evaluate the clinical accuracy of glucose predictions, reporting results in all 6 regions, but did not specify the primary region of interest (eg, potentially hazardous prediction errors). Other studies reported both sensitivity and specificity for hypoglycemia predictions without specifying a single measure. Furthermore, many studies have reported meaningful measures for the evaluation of model performance and clinical utility, without specifying their purpose. Examples include specificity and sensitivity, or the imaging study of Langner et al [75], which reported 3 metrics: the Dice score [97] and absolute and relative estimation error of adipose tissue volume. Although the metrics were adequate for both technical and clinical evaluation of results, owing to the lack of a single primary evaluation metric, the MI-CLAIM items could not be rated with unanimous "yes." Furthermore, we found an overlap within the model examination domain, where the same piece of information satisfied multiple items, especially when the model examination involved performance testing in special patient populations.

Finally, we must note the diversity of methods and nomenclature used from data processing to feature selection, prediction, or evaluation of results. We found nearly 2 unique ML methods per paper, which makes the systematic search and evidence synthesis of ML challenging. While the term "machine learning" yielded 42,840 hits, our extended search term using specific methods provided 235,042 hits over our study's search period in PubMed (Multimedia Appendices 2 and 11). While keeping track of the novel methods in ML is nearly impossible, the omission of specific terms carries the risk of incomplete search results for evidence synthesis. Furthermore, due to the specific methodological and reporting concerns about ML in medicine, we propose that the "machine learning study" label should be consistently used in the titles of medical studies, whose primary results arise from the typical data-driven ML workflow. This term has been used in the title or abstract of only 101 papers in PubMed in the same search period and could serve as unique identifier of clinical research using ML methods (Multimedia Appendix 12), benefiting future information retrieval and evidence synthesis similarly to the "randomized controlled trial" label.

Limitations

A limitation of our study was that the literature search was closed in early 2021; therefore, recent publications potentially eligible for our review were not covered. In addition, given the plethora of ML methods, although the detailed list was included in the search syntax, some studies may have been missed during the search phase of our study. However, we believe that the key

XSL•FO
RenderX

observations and conclusions of our review remain unaffected. In addition, we did not assess studies using MI-CLAIM, which reported ML results in pediatric DM without applying the ML workflow of model training, validation, and testing. Specific ML reporting guidelines exist [30,48] for clinical pilot studies [18], randomized controlled studies [90], and economic evaluations [91]. Despite the use of MI-CLAIM for a breadth of study types, we considered that MI-CLAIM was not applicable for a comparable assessment of those studies that omitted the steps of model validation and testing [86-89] with those that followed the full ML workflow.

Furthermore, despite the involvement of both medical experts and computer scientists in our research team, we observed high rates of disagreement between reviewers, and our consolidated ratings were "unsure" in over 20% of the assessed items. Although MI-CLAIM items were elaborated in group trainings, due to subjective judgments, some inconsistencies may have remained in our ratings. In particular, our judgments were unsure whether the representativeness of the training and test samples in real-world clinical settings was adequately demonstrated or if the selection and justification of primary model performance evaluation metrics were adequately justified. The specification of adequate sample characteristics and model performance evaluation criteria for clinical decision-making in pediatric DM and other disease areas remains an important area for future research.

Conclusions

The reporting quality of ML studies in the pediatric population with DM was generally low. Important details for clinicians, such as the detailed description of patient cohorts, the state-of-the-art clinical solution, the clinical utility of the proposed models, and model examination for valid, unbiased, and robust results, were often the weak points of reporting. To allow the assessment of their clinical utility, it is of utmost importance that the reporting standards of ML studies evolve and algorithms for this challenging population become more transparent and replicable. MI-CLAIM provided a strong strategic framework for good reporting practices, which could be further supported by disease-specific technical guidance regarding what constitutes an adequate level of detail to inform clinical decision-making. Higher reporting quality standards may indirectly advance science and facilitate the uptake of technologies that have the potential to benefit children with DM.

Acknowledgments

Project number TKP2021-NKTA-36 was implemented with support provided by the National Research, Development, and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme.

Authors' Contributions

ZZ contributed to the conceptualization, methodology development, formal analysis, investigation, project administration, and writing of the original draft. GK contributed to the methodology development and investigation. LG contributed to the conceptualization, methodology development, investigation, funding acquisition, and writing in the review and editing phase. JC, ÁH, and HMN contributed during the investigation. Methodology development and investigation were performed by AM. LK contributed during the conceptualization and provided supervision. AJB contributed to the writing during the review and editing phase. MP contributed to the conceptualization, methodology development, investigation, and funding acquisition.

Conflicts of Interest

AJB is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. AJB receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 123 KB - jmir_v26i1e47430_app1.docx]

Multimedia Appendix 2
Search syntax in PubMed.
[DOCX File , 143 KB - jmir_v26i1e47430_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Search syntax in Web of Science.
[DOCX File , 148 KB - jmir_v26i1e47430_app3.docx]

Multimedia Appendix 4
Excluded full-text papers.
[DOCX File , 191 KB - jmir_v26i1e47430_app4.docx]

Multimedia Appendix 5
Agreement between reviewers.
[DOCX File , 48 KB - jmir_v26i1e47430_app5.docx]

Multimedia Appendix 6
Characteristics of studies assessed via Minimum Information About Clinical Artificial Intelligence Modelling.
[DOCX File , 107 KB - jmir_v26i1e47430_app6.docx]

Multimedia Appendix 7
Assessment details of studies with Minimum Information About Clinical Artificial Intelligence Modelling.
[DOCX File , 66 KB - jmir_v26i1e47430_app7.docx]

Multimedia Appendix 8
Reported items in Minimum Information About Clinical Artificial Intelligence Modelling (part 1).
[DOCX File , 178 KB - jmir_v26i1e47430_app8.docx]

Multimedia Appendix 9
Reported items in Minimum Information About Clinical Artificial Intelligence Modelling (parts 2 and 3).
[DOCX File , 166 KB - jmir_v26i1e47430_app9.docx]

Multimedia Appendix 10
Reported items in Minimum Information About Clinical Artificial Intelligence Modelling (part 4).
[DOCX File , 98 KB - jmir_v26i1e47430_app10.docx]

Multimedia Appendix 11
Reported items in Minimum Information About Clinical Artificial Intelligence Modelling (part 5).
[DOCX File , 143 KB - jmir_v26i1e47430_app11.docx]

Multimedia Appendix 12
Additional searches in PubMed.
[DOCX File , 44 KB - jmir_v26i1e47430_app12.docx]

References

1. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195-e203 [FREE Full text] [doi: 10.1016/S2589-7500(20)30292-2] [Medline: 33478929]
2. Angehrn Z, Haldna L, Zandvliet AS, Gil Berglund E, Zeeuw J, Amzal B, et al. Artificial intelligence and machine learning applied at the point of care. Front Pharmacol 2020 Jun 18;11:759 [FREE Full text] [doi: 10.3389/fphar.2020.00759] [Medline: 32625083]
3. Stanescu DE, Lord K, Lipman TH. The epidemiology of type 1 diabetes in children. Endocrinol Metab Clin North Am 2012 Dec;41(4):679-694. [doi: 10.1016/j.ecl.2012.08.001] [Medline: 23099264]
4. Mayer-Davis EJ, Lawrence JM, Dabelea D, Divers J, Isom S, Dolan L, et al. Incidence trends of type 1 and type 2 diabetes among youths, 2002-2012. N Engl J Med 2017 Apr 13;376(15):1419-1429 [FREE Full text] [doi: 10.1056/NEJMoa1610187] [Medline: 28402773]
5. Shah AS, Nadeau KJ. The changing face of paediatric diabetes. Diabetologia 2020 Apr;63(4):683-691 [FREE Full text] [doi: 10.1007/s00125-019-05075-6] [Medline: 31897525]

XSL•FO
RenderX

6. Guariguata L, Jeyaseelan S. Children and non-communicable disease: global burden report 2019. NCD Child. 2019 Jan 28. URL: https://www.ncdchild.org/wp-content/uploads/2021/03/ncdchild_global_burden-report-2019.pdf [accessed 2022-11-10]
7. Pihoker C, Gilliam LK, Ellard S, Dabelea D, Davis C, Dolan LM, et al. Prevalence, characteristics and clinical diagnosis of maturity onset diabetes of the young due to mutations in HNF1A, HNF4A, and glucokinase: results from the SEARCH for diabetes in youth. J Clin Endocrinol Metab 2013 Oct;98(10):4055-4062 [FREE Full text] [doi: 10.1210/jc.2013-1279] [Medline: 23771925]
8. Dabelea D, Pihoker C, Talton JW, D'Agostino RBJ, Fujimoto W, Klingensmith GJ, et al. Etiological approach to characterization of diabetes type: the SEARCH for diabetes in youth study. Diabetes Care 2011 Jul;34(7):1628-1633 [FREE Full text] [doi: 10.2337/dc10-2324] [Medline: 21636800]
9. Sherr JL, Schoelwer M, Dos Santos TJ, Reddy L, Biester T, Galderisi A, et al. ISPAD clinical practice consensus guidelines 2022: diabetes technologies: insulin delivery. Pediatr Diabetes 2022 Dec;23(8):1406-1431. [doi: 10.1111/pedi.13421] [Medline: 36468192]
10. Dabelea D, Stafford JM, Mayer-Davis EJ, D'Agostino RJ, Dolan L, Imperatore G, et al. Association of type 1 diabetes vs type 2 diabetes diagnosed during childhood and adolescence with complications during teenage years and young adulthood. JAMA 2017 Feb 28;317(8):825-835 [FREE Full text] [doi: 10.1001/jama.2017.0686] [Medline: 28245334]
11. Hamman RF, Bell RA, Dabelea D, D'Agostino RBJ, Dolan L, Imperatore G, et al. The SEARCH for diabetes in youth study: rationale, findings, and future directions. Diabetes Care 2014 Dec 20;37(12):3336-3344 [FREE Full text] [doi: 10.2337/dc14-0574] [Medline: 25414389]
12. Danne T, Limbert C. COVID-19, type 1 diabetes, and technology: why paediatric patients are leading the way. Lancet Diabetes Endocrinol 2020 Jun;8(6):465-467. [doi: 10.1016/s2213-8587(20)30155-8]
13. Tauschmann M, Forlenza G, Hood K, Cardona-Hernandez R, Giani E, Hendrieckx C, et al. ISPAD clinical practice consensus guidelines 2022: diabetes technologies: glucose monitoring. Pediatr Diabetes 2022 Dec;23(8):1390-1405 [FREE Full text] [doi: 10.1111/pedi.13451] [Medline: 36537528]
14. January 2023 510(K) clearances. US Food and Drug Administration. 2023. URL: https://www.fda.gov/medical-devices/510k-clearances/january-2023-510k-clearances [accessed 2023-02-10]
15. Miller KM, Beck RW, Foster NC, Maahs DM. HbA1c levels in type 1 diabetes from early childhood to older adults: a deeper dive into the influence of technology and socioeconomic status on HbA1c in the T1D exchange clinic registry findings. Diabetes Technol Ther 2020 Sep;22(9):645-650 [FREE Full text] [doi: 10.1089/dia.2019.0393] [Medline: 31905008]
16. De Ridder F, den Brinker M, De Block C. The road from intermittently scanned glucose monitoring to hybrid closed-loop systems: part A. Keys to success: subject profiles, choice of systems, education. Ther Adv Endocrinol Metab 2019 Jul 25;10:2042018819865399 [FREE Full text] [doi: 10.1177/2042018819865399] [Medline: 31384420]
17. Leelarathna L, Choudhary P, Wilmot EG, Lumb A, Street T, Kar P, et al. Hybrid closed-loop therapy: where are we in 2021? Diabetes Obes Metab 2021 Mar;23(3):655-660. [doi: 10.1111/dom.14273] [Medline: 33269551]
18. Alfonsi JE, Choi EE, Arshad T, Sammott SA, Pais V, Nguyen C, et al. Carbohydrate counting app using image recognition for youth with type 1 diabetes: pilot randomized control trial. JMIR Mhealth Uhealth 2020 Oct 28;8(10):e22074 [FREE Full text] [doi: 10.2196/22074] [Medline: 33112249]
19. Castensøe-Seidenfaden P, Reventlov Husted G, Teilmann G, Hommel E, Olsen BS, Kensing F. Designing a self-management app for young people with type 1 diabetes: methodological challenges, experiences, and recommendations. JMIR Mhealth Uhealth 2017 Oct 23;5(10):e124 [FREE Full text] [doi: 10.2196/mhealth.8137] [Medline: 29061552]
20. Vangeepuram N, Liu B, Chiu PH, Wang L, Pandey G. Predicting youth diabetes risk using NHANES data and machine learning. Sci Rep 2021 May 27;11(1):11212 [FREE Full text] [doi: 10.1038/s41598-021-90406-0] [Medline: 34045491]
21. Siegel AP, Daneshkhah A, Hardin DS, Shrestha S, Varahramyan K, Agarwal M. Analyzing breath samples of hypoglycemic events in type 1 diabetes patients: towards developing an alternative to diabetes alert dogs. J Breath Res 2017 Jun 01;11(2):026007. [doi: 10.1088/1752-7163/aa6ac6] [Medline: 28569238]
22. Fleming GA, Petrie JR, Bergenstal RM, Holl RW, Peters AL, Heinemann L. Diabetes digital app technology: benefits, challenges, and recommendations. A consensus report by the European association for the study of diabetes (EASD) and the American diabetes association (ADA) diabetes technology working group. Diabetologia 2020 Feb;63(2):229-241. [doi: 10.1007/s00125-019-05034-1] [Medline: 31802144]
23. Prahalad P, Tanenbaum M, Hood K, Maahs DM. Diabetes technology: improving care, improving patient-reported outcomes and preventing complications in young people with type 1 diabetes. Diabet Med 2018 Apr;35(4):419-429. [doi: 10.1111/dme.13588] [Medline: 29356074]
24. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018 May 30;20(5):e10775 [FREE Full text] [doi: 10.2196/10775] [Medline: 29848472]
25. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022 Jan 13;22(1):12 [FREE Full text] [doi: 10.1186/s12874-021-01469-6] [Medline: 35026997]

XSL•FO
RenderX

26. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform 2021 Sep;153:104510 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104510] [Medline: 34108105]
27. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc 2018 Aug 01;25(8):963-968 [FREE Full text] [doi: 10.1093/jamia/ocy028] [Medline: 29669066]
28. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
29. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
30. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform 2021 Aug;28(1):e100385 [FREE Full text] [doi: 10.1136/bmjhci-2021-100385] [Medline: 34426417]
31. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc 2016 Dec 01;3(2). [doi: 10.1177/2053951716679679]
32. Erasmus A, Brunet TD, Fisher E. What is interpretability? Philos Technol 2021;34(4):833-862 [FREE Full text] [doi: 10.1007/s13347-020-00435-2] [Medline: 34966640]
33. Zednik C. Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 2019 Dec 20;34:265-288. [doi: 10.1007/s13347-019-00382-7]
34. Casey B, Farhangi A, Vogl R. Rethinking explainable machines: the GDPR's right to explanation debate and the rise of algorithmic audits in enterprise. Berkeley Technol Law J 2019;34(1):143-188. [doi: 10.15779/Z38M32N986]
35. Broniatowski DA. Psychological foundations of explainability and interpretability in artificial intelligence. National Institute of Standards and Technology. 2021. URL: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf [accessed 2023-02-10]
36. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019 Feb;267:1-38. [doi: 10.1016/j.artint.2018.07.007]
37. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. Preprint posted online February 28, 2017 2023:1-13 [FREE Full text] [doi: 10.1201/9780367816377-16]
38. Hatherley J, Sparrow R, Howard M. The virtues of interpretable medical AI. Camb Q Healthc Ethics 2023 Jan 10:1-10. [doi: 10.1017/S0963180122000664] [Medline: 36624634]
39. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? Ann Intern Med 2020 Jan 07;172(1):59-60. [doi: 10.7326/M19-2548] [Medline: 31842204]
40. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312 [FREE Full text] [doi: 10.1002/widm.1312] [Medline: 32089788]
41. Belle V, Papantonis I. Principles and practice of explainable machine learning. Front Big Data 2021 Jul 1;4:688969 [FREE Full text] [doi: 10.3389/fdata.2021.688969] [Medline: 34278297]
42. Kristiansen TB, Kristensen K, Uffelmann J, Brandslund I. Erroneous data: the Achilles' heel of AI and personalized medicine. Front Digit Health 2022 Jul 22;4:862095 [FREE Full text] [doi: 10.3389/fdgth.2022.862095] [Medline: 35937419]
43. Gierend K, Krüger F, Waltemath D, Fünfgeld M, Ganslandt T, Zeleke AA. Approaches and criteria for provenance in biomedical data sets and workflows: protocol for a scoping review. JMIR Res Protoc 2021 Nov 22;10(11):e31750 [FREE Full text] [doi: 10.2196/31750] [Medline: 34813494]
44. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995 Feb 01;273(5):408-412. [doi: 10.1001/jama.273.5.408] [Medline: 7823387]
45. Han S, Olonisakin TF, Pribis JP, Zupetic J, Yoon JH, Holleran KM, et al. A checklist is associated with increased quality of reporting preclinical biomedical research: a systematic review. PLoS One 2017 Sep 13;12(9):e0183591 [FREE Full text] [doi: 10.1371/journal.pone.0183591] [Medline: 28902887]
46. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust 2006 Sep 04;185(5):263-267. [doi: 10.5694/j.1326-5377.2006.tb00557.x] [Medline: 16948622]
47. Simera I, Moher D, Hoey J, Schulz KF, Altman DG. A catalogue of reporting guidelines for health research. Eur J Clin Invest 2010 Jan;40(1):35-53. [doi: 10.1111/j.1365-2362.2009.02234.x] [Medline: 20055895]
48. Zrubka Z, Gulácsi L, Péntek M. Time to start using checklists for reporting artificial intelligence in health care and biomedical research: a rapid review of available tools. In: Proceedings of the IEEE 26th International Conference on Intelligent Engineering Systems (INES). 2022 Presented at: IEEE 26th International Conference on Intelligent Engineering Systems (INES); August 12-15, 2022; Georgioupoli, Greece URL: https://ieeexplore.ieee.org/document/9922639 [doi: 10.1109/ines56734.2022.9922639]

XSL•FO
RenderX

49. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020 Sep 09;370:m3164 [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]
50. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017 Jan 8;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
51. Zhu T, Li K, Herrero P, Georgiou P. Deep learning for diabetes: a systematic review. IEEE J Biomed Health Inform 2021 Jul;25(7):2744-2757. [doi: 10.1109/JBHI.2020.3040225] [Medline: 33232247]
52. Tsichlaki S, Koumakis L, Tsiknakis M. Type 1 diabetes hypoglycemia prediction algorithms: systematic review. JMIR Diabetes 2022 Jul 21;7(3):e34699 [FREE Full text] [doi: 10.2196/34699] [Medline: 35862181]
53. Saputro SA, Pattanaprateep O, Pattanateepapon A, Karmacharya S, Thakkinstian A. Prognostic models of diabetic microvascular complications: a systematic review and meta-analysis. Syst Rev 2021 Nov 01;10(1):288 [FREE Full text] [doi: 10.1186/s13643-021-01841-z] [Medline: 34724973]
54. Tan KR, Seng JJB, Kwan YH, Chen YJ, Zainudin SB, Loh DH, et al. Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review. J Diabetes Sci Technol 2023 Mar;17(2):474-489 [FREE Full text] [doi: 10.1177/19322968211056917] [Medline: 34727783]
55. Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr 2021 Dec 20;13(1):148 [FREE Full text] [doi: 10.1186/s13098-021-00767-9] [Medline: 34930452]
56. Silva KD, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis. Int J Med Inform 2020 Nov;143:104268. [doi: 10.1016/j.ijmedinf.2020.104268] [Medline: 32950874]
57. Kamel Rahimi A, Canfell OJ, Chan W, Sly B, Pole JD, Sullivan C, et al. Machine learning models for diabetes management in acute care using electronic medical records: a systematic review. Int J Med Inform 2022 Apr 02;162:104758. [doi: 10.1016/j.ijmedinf.2022.104758] [Medline: 35398812]
58. Makroum MA, Adda M, Bouzouane A, Ibrahim H. Machine learning and smart devices for diabetes management: systematic review. Sensors (Basel) 2022 Feb 25;22(5):1843 [FREE Full text] [doi: 10.3390/s22051843] [Medline: 35270989]
59. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
60. Machine learning: MeSH descriptor data 2024. National Institutes of Health National Library of Medicine. 2016. URL: https://meshb.nlm.nih.gov/record/ui?ui=D000069550 [accessed 2023-02-10]
61. Artificial intelligence: MeSH descriptor data 2024. National Institutes of Health National Library of Medicine. URL: https://meshb.nlm.nih.gov/record/ui?ui=D001185 [accessed 2023-02-10]
62. Kuhn M. caret: classification and regression training. The Comprehensive R Archive Network. 2017 Dec 10. URL: http://cran.nexr.com/web/packages/caret/index.html [accessed 2023-02-10]
63. Producing a child-relevant cochrane review. The Cochrane Collaboration. URL: https://childhealth.cochrane.org/producing-child-relevant-cochrane-review [accessed 2023-02-10]
64. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020 Sep;26(9):1320-1324 [FREE Full text] [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]
65. Scimago journal and country rank homepage. Scimago Journal and Country Rank. URL: https://www.scimagojr.com/journalrank.php [accessed 2022-10-01]
66. Daskalaki E, Diem P, Mougiakakou SG. Model-free machine learning in biomedicine: feasibility study in type 1 diabetes. PLoS One 2016 Jul 21;11(7):e0158722 [FREE Full text] [doi: 10.1371/journal.pone.0158722] [Medline: 27441367]
67. Ling SH, San PP, Nguyen HT. Non-invasive hypoglycemia monitoring system using extreme learning machine for type 1 diabetes. ISA Trans 2016 Sep;64:440-446. [doi: 10.1016/j.isatra.2016.05.008] [Medline: 27311357]
68. Miller RG, Anderson SJ, Costacou T, Sekikawa A, Orchard TJ. Risk stratification for 25-year cardiovascular disease incidence in type 1 diabetes: tree-structured survival analysis of the Pittsburgh epidemiology of diabetes complications study. Diab Vasc Dis Res 2016 Jul;13(4):250-259 [FREE Full text] [doi: 10.1177/1479164116629353] [Medline: 27190081]
69. San PP, Ling SH, Nguyen HT. Deep learning framework for detection of hypoglycemic episodes in children with type 1 diabetes. Annu Int Conf IEEE Eng Med Biol Soc 2016 Aug;2016:3503-3506. [doi: 10.1109/EMBC.2016.7591483] [Medline: 28269053]
70. Ling SH, San PP, Lam HK, Nguyen HT. Hypoglycemia detection: multiple regression-based combinational neural logic approach. Soft Comput 2015 Jul 25;21(2):543-553. [doi: 10.1007/s00500-015-1809-z]
71. Stawiski K, Pietrzak I, Młynarski W, Fendler W, Szadkowska A. NIRCa: an artificial neural network-based insulin resistance calculator. Pediatr Diabetes 2018 Mar;19(2):231-235. [doi: 10.1111/pedi.12551] [Medline: 28626972]
72. De Bois M, El Yacoubi MA, Ammi M. Study of short-term personalized glucose predictive models on type-1 diabetic children. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). 2019 Presented at: 2019

XSL•FO
RenderX

International Joint Conference on Neural Networks (IJCNN); July 14-19, 2019; Budapest, Hungary URL: https://ieeexplore.ieee.org/document/8852399 [doi: 10.1109/ijcnn.2019.8852399]
73. De Bois M, Ammi M, El Yacoubi MA. Model fusion to enhance the clinical acceptability of long-term glucose predictions. In: Proceedings of the IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). 2019 Presented at: IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE); October 28-30, 2019; Athens, Greece URL: https://ieeexplore.ieee.org/document/8941649 [doi: 10.1109/bibe.2019.00053]
74. Khusial RD, Cioffi CE, Caltharp SA, Krasinskas AM, Alazraki A, Knight-Scott J, et al. Development of a plasma screening panel for pediatric nonalcoholic fatty liver disease using metabolomics. Hepatol Commun 2019 Oct;3(10):1311-1321 [FREE Full text] [doi: 10.1002/hep4.1417] [Medline: 31592078]
75. Langner T, Hedström A, Mörwald K, Weghuber D, Forslund A, Bergsten P, et al. Fully convolutional networks for automated segmentation of abdominal adipose tissue depots in multicenter water-fat MRI. Magn Reson Med 2019 Apr;81(4):2736-2745. [doi: 10.1002/mrm.27550] [Medline: 30311704]
76. Ngo CQ, Chai R, Nguyen TV, Jones TW, Nguyen HT. Nocturnal hypoglycemia detection using EEG spectral moments under natural occurrence conditions. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:7177-7180. [doi: 10.1109/EMBC.2019.8856695] [Medline: 31947490]
77. Stanfill B, Reehl S, Bramer L, Nakayasu ES, Rich SS, Metz TO, et al. Extending classification algorithms to case-control studies. Biomed Eng Comput Biol 2019 Jul 15;10:1179597219858954 [FREE Full text] [doi: 10.1177/1179597219858954] [Medline: 31320812]
78. Amar Y, Shilo S, Oron T, Amar E, Phillip M, Segal E. Clinically accurate prediction of glucose levels in patients with type 1 diabetes. Diabetes Technol Ther 2020 Aug 01;22(8):562-569. [doi: 10.1089/dia.2019.0435] [Medline: 31928415]
79. Dave D, DeSalvo DJ, Haridas B, McKay S, Shenoy A, Koh CJ, et al. Feature-based machine learning model for real-time hypoglycemia prediction. J Diabetes Sci Technol 2021 Jul;15(4):842-855 [FREE Full text] [doi: 10.1177/1932296820922622] [Medline: 32476492]
80. Frohnert BI, Webb-Robertson BJ, Bramer LM, Reehl SM, Waugh K, Steck AK, et al. Predictive modeling of type 1 diabetes stages using disparate data sources. Diabetes 2020 Feb;69(2):238-248 [FREE Full text] [doi: 10.2337/db18-1263] [Medline: 31740441]
81. Garavelli S, Bruzzaniti S, Tagliabue E, Di Silvestre D, Prattichizzo F, Mozzillo E, et al. Plasma circulating miR-23~27~24 clusters correlate with the immunometabolic derangement and predict C-peptide loss in children with type 1 diabetes. Diabetologia 2020 Dec;63(12):2699-2712. [doi: 10.1007/s00125-020-05237-x] [Medline: 32728892]
82. Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: a deep learning framework for accurate glucose forecasting. IEEE J Biomed Health Inform 2020 Feb;24(2):414-423. [doi: 10.1109/jbhi.2019.2931842]
83. Zhu T, Li K, Kuang L, Herrero P, Georgiou P. An insulin bolus advisor for type 1 diabetes using deep reinforcement learning. Sensors (Basel) 2020 Sep 06;20(18):5058 [FREE Full text] [doi: 10.3390/s20185058] [Medline: 32899979]
84. Zhu T, Li K, Herrero P, Georgiou P. Basal glucose control in type 1 diabetes using deep reinforcement learning: an in silico validation. IEEE J Biomed Health Inform 2021 Apr;25(4):1223-1232. [doi: 10.1109/JBHI.2020.3014556] [Medline: 32755873]
85. Webb-Robertson BJ, Bramer LM, Stanfill BA, Reehl SM, Nakayasu ES, Metz TO, et al. Prediction of the development of islet autoantibodies through integration of environmental, genetic, and metabolic markers. J Diabetes 2021 Feb;13(2):143-153 [FREE Full text] [doi: 10.1111/1753-0407.13093] [Medline: 33124145]
86. Adabimohazab R, Garfinkel A, Milam EC, Frosch O, Mangone A, Convit A. Does inflammation mediate the association between obesity and insulin resistance? Inflammation 2016 Jun;39(3):994-1003 [FREE Full text] [doi: 10.1007/s10753-016-0329-z] [Medline: 26956471]
87. Brugha R, Wright M, Nolan S, Bridges N, Carr SB. Quantifying fluctuation in glucose levels to identify early changes in glucose homeostasis in cystic fibrosis. J Cyst Fibros 2018 Nov;17(6):791-797 [FREE Full text] [doi: 10.1016/j.jcf.2017.12.004] [Medline: 29329721]
88. Cabrera SM, Engle S, Kaldunski M, Jia S, Geoffrey R, Simpson P, Type 1 Diabetes TrialNet CTLA4-Ig (Abatacept) Study Group, et al. Innate immune activity as a predictor of persistent insulin secretion and association with responsiveness to CTLA4-Ig treatment in recent-onset type 1 diabetes. Diabetologia 2018 Nov;61(11):2356-2370 [FREE Full text] [doi: 10.1007/s00125-018-4708-x] [Medline: 30167736]
89. Biassoni R, Di Marco E, Squillario M, Barla A, Piccolo G, Ugolotti E, et al. Gut microbiota in T1DM-onset pediatric patients: machine-learning algorithms to classify microorganisms as disease linked. J Clin Endocrinol Metab 2020 Sep 01;105(9):dgaa407. [doi: 10.1210/clinem/dgaa407] [Medline: 32692360]
90. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med 2020 Sep;26(9):1380-1384. [doi: 10.1038/s41591-020-1045-7] [Medline: 32908282]
91. Wolf RM, Channa R, Abramoff MD, Lehmann HP. Cost-effectiveness of autonomous point-of-care diabetic retinopathy screening for pediatric patients with diabetes. JAMA Ophthalmol 2020 Oct 01;138(10):1063-1069 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.3190] [Medline: 32880616]

XSL•FO
RenderX

92. Farook TH, Jamayet NB, Abdullah JY, Alam MK. Machine learning and intelligent diagnostics in dental and orofacial pain management: a systematic review. Pain Res Manag 2021 Apr 24;2021:6659133 [FREE Full text] [doi: 10.1155/2021/6659133] [Medline: 33986900]
93. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010 Mar 24;8:18 [FREE Full text] [doi: 10.1186/1741-7015-8-18] [Medline: 20334633]
94. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021 Mar;18(3):e1003583 [FREE Full text] [doi: 10.1371/journal.pmed.1003583] [Medline: 33780438]
95. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. Value Health 2022 Jan;25(1):3-9 [FREE Full text] [doi: 10.1016/j.jval.2021.11.1351] [Medline: 35031096]
96. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes Care 1987;10(5):622-628. [doi: 10.2337/diacare.10.5.622] [Medline: 3677983]
97. Moore C, Bell D. Dice similarity coefficient. Radiopaedia. URL: https://radiopaedia.org/articles/dice-similarity-coefficient?lang=gb [accessed 2023-02-10]

Abbreviations

AI: artificial intelligence
CGM: continuous glucose monitor
DM: diabetes mellitus
MeSH: Medical Subject Headings
MI-CLAIM: Minimum Information About Clinical Artificial Intelligence Modelling
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus

Edited by T de Azevedo Cardoso; submitted 20.03.23; peer-reviewed by N Jiwani, N Mungoli; comments to author 19.04.23; revised version received 29.04.23; accepted 17.11.23; published 19.01.24.

Please cite as:
Zrubka Z, Kertész G, Gulácsi L, Czere J, Hölgyesi Á, Nezhad HM, Mosavi A, Kovács L, Butte AJ, Péntek M
The Reporting Quality of Machine Learning Studies on Pediatric Diabetes Mellitus: Systematic Review
J Med Internet Res 2024;26:e47430
URL: https://www.jmir.org/2024/1/e47430
doi:10.2196/47430
PMID:38241075

©Zsombor Zrubka, Gábor Kertész, László Gulácsi, János Czere, Áron Hölgyesi, Hossein Motahari Nezhad, Amir Mosavi, Levente Kovács, Atul J Butte, Márta Péntek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Predictive Value of Machine Learning for Platinum Chemotherapy Responses in Ovarian Cancer: Systematic Review and Meta-Analysis

Qingyi Wang[1], MM; Zhuo Chang[2], PhD; Xiaofang Liu[1], PhD; Yunrui Wang[1], PhD; Chuwen Feng[1], PhD; Yunlu Ping[1], PhD; Xiaoling Feng[3], PhD

[1]Department of First Clinical Medical College, Heilongjiang University of Chinese Medicine, Harbin, China
[2]Basic Medical College, Heilongjiang University of Chinese Medicine, Harbin, China
[3]Department of Gynecology, First Affiliated Hospital of Heilongjiang University of Chinese Medicine, Harbin, China

Corresponding Author:
Xiaoling Feng, PhD
Department of Gynecology
First Affiliated Hospital of Heilongjiang University of Chinese Medicine
26 Heping Road
Xiangfang District
Harbin, 150000
China
Phone: 86 13604800585
Email: doctorfxl@163.com

Abstract

Background: Machine learning is a potentially effective method for predicting the response to platinum-based treatment for ovarian cancer. However, the predictive performance of various machine learning methods and variables is still a matter of controversy and debate.

Objective: This study aims to systematically review relevant literature on the predictive value of machine learning for platinum-based chemotherapy responses in patients with ovarian cancer.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we systematically searched the PubMed, Embase, Web of Science, and Cochrane databases for relevant studies on predictive models for platinum-based therapies for the treatment of ovarian cancer published before April 26, 2023. The Prediction Model Risk of Bias Assessment tool was used to evaluate the risk of bias in the included articles. Concordance index (C-index), sensitivity, and specificity were used to evaluate the performance of the prediction models to investigate the predictive value of machine learning for platinum chemotherapy responses in patients with ovarian cancer.

Results: A total of 1749 articles were examined, and 19 of them involving 39 models were eligible for this study. The most commonly used modeling methods were logistic regression (16/39, 41%), Extreme Gradient Boosting (4/39, 10%), and support vector machine (4/39, 10%). The training cohort reported C-index in 39 predictive models, with a pooled value of 0.806; the validation cohort reported C-index in 12 predictive models, with a pooled value of 0.831. Support vector machine performed well in both the training and validation cohorts, with a C-index of 0.942 and 0.879, respectively. The pooled sensitivity was 0.890, and the pooled specificity was 0.790 in the training cohort.

Conclusions: Machine learning can effectively predict how patients with ovarian cancer respond to platinum-based chemotherapy and may provide a reference for the development or updating of subsequent scoring systems.

(J Med Internet Res 2024;26:e48527) doi:10.2196/48527

KEYWORDS

ovarian cancer; platinum chemotherapy response; machine learning; platinum-based therapy; predictive potential

XSL•FO
RenderX

Introduction

Background

Ovarian cancer is one of the most common cancers worldwide [1], and this gynecological cancer is characterized by poor prognosis and high mortality [2]. It is estimated that epithelial ovarian cancer (EOC) represents 90% of the ovarian cancer cases [3], with serous carcinoma being the most common pathological type [4]. Because of the absence of cancer-specific symptoms and effective screening techniques, EOC is frequently diagnosed at a late stage [5,6]. Despite undergoing relevant treatments, patients with ovarian cancer still face high rates of recurrence and mortality, with a 5-year survival rate of <30% [7]. According to GLOBOCAN 2020, the number of new ovarian cancer cases in low and high Human Development Index countries will increase by approximately 96% and 19%, respectively, by 2040 [8].

Currently, the National Comprehensive Cancer Network Guidelines (2023 Edition) recommend the use of paclitaxel or carboplatin for 3 weeks as the first-line treatment for stage 2 to 4 EOC [9]. Although platinum-based chemotherapy is effective in most patients with ovarian cancer, resistance may occur in some patients [10]. In addition, their response to platinum treatment is unknown until chemotherapy is completed. The platinum-free interval is a reliable indicator for predicting treatment efficacy and patient prognosis because it can evaluate whether patients with ovarian cancer respond to platinum drugs and their recurrence [11,12]. The Gynecologic Cancer Group divides responses to platinum chemotherapy into 4 categories based on the duration of platinum-free interval (platinum refractory: <1 mo, platinum-resistant: 1-6 mo, partial platinum response: 6-12 mo, and platinum response: >12 mo) [13]. The chemoresistance of ovarian cancer may be related to genome expression [14,15], tumor heterogeneity, intestinal microbiota, DNA repair [16], DNA methylation [17,18], and mitochondrial function [19,20] related to immunoediting. Within 2 years, approximately 70% of these patients relapse [21]. Therefore, predicting the response to platinum-based chemotherapy in patients with ovarian cancer is critical. Despite the emergence of multiple approaches, including mutational signatures, transcriptomic signatures, tumor mutation burden, and functional biomarkers, there is no conclusive evidence to guide the precise treatment of patients with ovarian cancer [22].

Objectives

In recent years, with the increasing availability of medical data and the continuous improvement in computer analysis capabilities, machine learning has been increasingly used in the medical field [23,24]. Machine learning is a technological application that uses algorithms and data to enable computers to automatically learn and enhance. It excels in handling large amounts of complex and multidimensional information, thereby improving the accuracy and efficiency of decision-making [25-27]. In various domains of oncology, machine learning has been used to identify cancer imaging features [28], predict the risk of cancer recurrence [29], screen cancer drug targets [30], and optimize cancer treatment options [31]. Some researchers have explored machine learning–based methods to construct prediction models for platinum reactions in ovarian cancer. However, in the field of machine learning research, there is a diverse range of methods and variables. The predictive performance of these methods for outcome events remains controversial. Currently, in evidence-based medicine, a comprehensive summary of the predictive performance of machine learning is lacking. Therefore, we conducted this study to explore early risk stratification in response to platinum-based chemotherapy in patients with ovarian cancer. Our aim was to enhance chemotherapy management in patients with ovarian cancer.

Methods

This study was carried out following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Table S1 in Multimedia Appendix 1) and registered on PROSPERO (CRD42022340928).

Data Sources and Searches

Relevant studies published before March 15, 2022, were thoroughly searched in the PubMed, Web of Science, Embase, and Cochrane databases. Search terms included subject headings (Medical Subject Headings in PubMed and Emtree in Embase) and free words, such as “Ovarian Neoplasms,” “machine learning,” “prediction model,” and “Platinum-Based Chemotherapy.” The specific search strategy is presented in Table S1 in Multimedia Appendix 2. To prevent the risk of missing new literature, we conducted a supplementary search of each database until April 26, 2023.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows:

- Patients diagnosed with ovarian cancer
- The research types are case-control, cohort, nested case-control, and case cohort studies.\
- A completely constructed predictive model for platinum chemotherapy (platinum-sensitive or platinum-resistant) response in patients with ovarian cancer
- Studies without external validation
- Different machine learning studies published on the same data set
- The literature written in English

Meanwhile, the following studies were excluded:

- The research type was meta-analysis, review, guideline, expert opinion, etc
- Only a risk factor analysis was carried out, but no complete machine learning prediction model was developed
- The following outcome measures were used: receiver operating characteristic curve, concordance index (C-index), sensitivity, specificity, accuracy, recovery rate, precision rate, confusion matrix, diagnostic 4-grid table, F_1-score, and calibration curve. The original study should include at least one of the above indicators. Missing studies need to be excluded.
- Studies with small sample sizes (<50 cases)
- Research on the accuracy of single-factor prediction

XSL•FO
RenderX

Literature Screening and Data Extraction

The retrieved studies were imported into EndNote (Clarivate Plc) to remove duplicate publications automatically and manually. Subsequently, we reviewed the titles or abstracts of the remaining studies to exclude original studies that did not align with the topic. We proceeded to read the full texts of the remaining studies to screen the original studies that ultimately met the criteria.

The following information was collected from each eligible study: first author, year of publication, location, research duration, population characteristics, background, number of hospitals, study design (prospective or retrospective), number of patients, age, histological classification of enrolled patients, presence of tumor deposit after treatment, treatment methods, prediction objects, chemotherapy methods, number of positive samples, number of training set samples, total number of samples, follow-up time, variable selection method, characteristics of the machine learning approach (specific algorithm or type), validation method (cross-validation, retention method, external validation, or none), number of model variables, included model variables, efficacy evaluation indicators, sample size of the validation set, and prediction results.

The literature screening and data extraction were independently conducted by 2 researchers (QW and ZC). Following completion, crosschecks were performed. In the event of any disputes, a third researcher (XL) was consulted to assist in resolution.

Quality Assessment

The Prediction Model Risk of Bias Assessment (PROBAST), a technique for predicting the model risk of bias, was used to assess the risk of bias in the predictive models reported in eligible studies [32,33]. This tool consists of 4 major domains, participants, predictors, outcomes, and statistical analysis, and it reveals the overall risk of bias and applicability. The 4 domains have 2, 3, 6, and 9 distinct questions, respectively, with 3 possible answers: yes or probably yes, no or probably no, and no information. A domain is deemed high risk if it receives at least 1 no or probably no question, whereas a domain that receives yes or probably yes for all of its questions is considered low risk. When all domains are classified as low risk, the overall risk of bias is graded as low. Meanwhile, when at least one domain is deemed high risk, the overall risk of bias is regarded as high. Two investigators independently assessed the risk of bias and crosschecked their findings using PROBAST. Any disagreements were resolved by discussion with a third researcher. The assessment of the risk of bias was independently conducted by 2 researchers (YW and CF). Upon completion, a crosscheck was performed. In case of any dispute, a third researcher (YP) was consulted to assist in the decision-making process.

Data Analysis

If the C-index lacked a 95% CI and SE, we referred to the study by Debray et al [34], which estimated its SEs. Because machine learning encompasses a wide range of mathematical models and predictive factors, there is high heterogeneity among various studies. Hence, a random effects model was used for the meta-analysis. In addition, we used a bivariate mixed effects model, which is a random effects model, to perform the meta-analysis of sensitivity and specificity. At the same time, we used the heterogeneity index (I^2) to measure the heterogeneity. $P<.05$ indicated a statistically significant difference. Moreover, subgroup analysis was conducted to increase the robustness of the results and reduce heterogeneity between studies, according to the different types of prediction models and the possible influencing factors, for instance, whether it is high-grade serous ovarian cancer and whether there is a tumor deposit.

Ethical Considerations

All analyses were based on previously published studies; therefore, ethics approval and patient consent were not required.

Results

Search Strategy

A total of 1749 articles were obtained from the PubMed, Web of Science, Embase, and Cochrane databases. After removing 752 duplicates, we screened the titles and abstracts and identified 261 potentially eligible articles. On the basis of a full-text review, 242 studies were excluded, with 234 (96.7%) studies deleted for inappropriate outcomes, 6 (2.5%) studies deleted for inadequate data, and 2 (0.8%) studies deleted for no access to the full text. Finally, this study included 19 articles. Figure 1 shows the study search strategy.

XSL•FO
RenderX

Figure 1. Flow diagram of the study selection process.

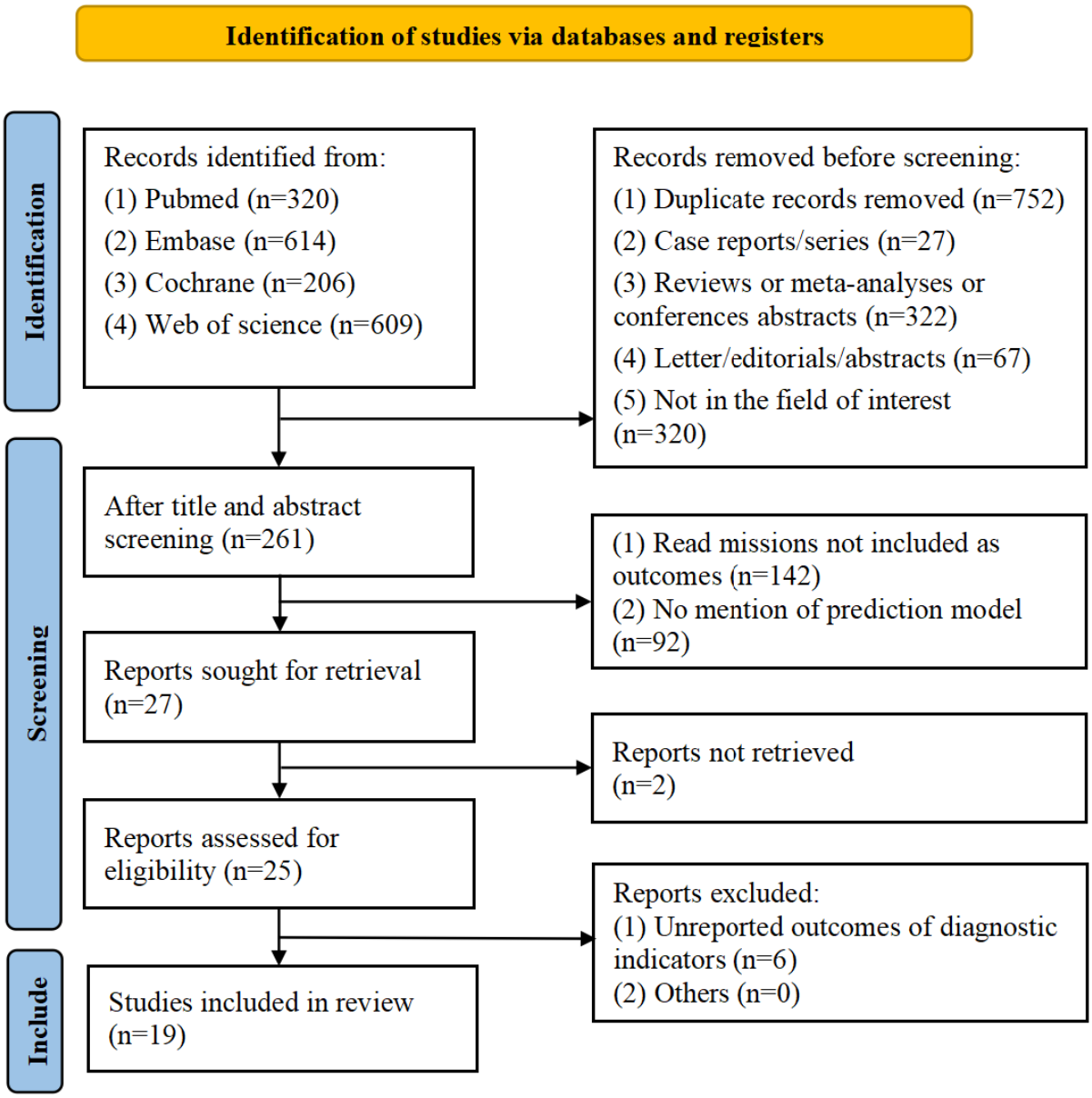

Characteristics of Included Studies

The basic characteristics of the 19 eligible articles [35-53] are presented in Table 1. Of the 19 studies, only one consisted of patients who had recurrent ovarian cancer, whereas the remaining 18 studies involved patients who had primary ovarian cancer. There were 3 multicenter studies, 5 single-center studies, and 11 database-based studies. In total, 15 studies were externally validated. The 19 eligible studies involved 7137 patients, and the number of patients included ranged from 58 to 1002. These eligible studies contained 39 predictive models, of which 22 reported sensitivity and specificity. The most widely used modeling methods in the training cohort were logistic regression (LR; 16/39, 41%), Extreme Gradient Boosting (XGBoost; 4/39, 10%), and support vector machines (SVMs; 4/39, 10%), whereas the common modeling method in the validation cohort was SVM (4/12, 33%).

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study, year	Country	Sample source	Chemotherapy	Positive samples, n/N (%)	Training set samples, n/N (%)	Overall sample size, n	Variable selection method	Type of model	External validation	External validation sample size
Shannon et al [35], 2021	Singapore	GDSC[a], TCGA[b], and GEO[c]	Carboplatin	39/50 (78)	50/60 (83)	60	—[d]	SVM[e], LR[f], KNN[g], DT[h], AdaBoost[i], GBM[j], and XGBoost[k]	1	10
Hwangbo et al [37], 2021	Korea	The Seoul National University Hospital, Asan Medical Center, and Severance Hospital	Platinum-based combination chemotherapy	779/1002 (77.7)	1002/1002 (100)	1002	Univariate and multivariate analysis	LR, RF[l], SVM, and DNN[m]	0	—
Zhao et al [38], 2019	China	GEO and TCGA	Platinum-based combination chemotherapy	42/129 (32)	129/707 (18.2)	707	Univariate and multivariate analysis	LR, COX[n], SVM, and ANN[o]	1	454
Paik et al [39], 2017	Korea	Samsung Medical Center	Platinum-based combination chemotherapy	616/757 (81.4)	757/757 (100)	757	Univariate and multivariate analysis	LR	0	—
Han et al [40], 2012	China	TCGA and GEO	Platinum or paclitaxel-based treatment	177/200 (88.5)	200/322 (62.1)	322	Principal components method	SPC[p]	1	122
Lan et al [36], 2019	China	Sun Yat-Sen University Cancer Center	Platinum or paclitaxel-based treatment	22/71 (31)	71/91 (78)	91	Univariate and multivariate analysis	LR	0	—
Zheng et al [41], 2021	China	Beijing Cancer Hospital, Peking Union Medical College and TCGA	Taxol plus platinum-based chemotherapy	44/60 (73)	60/448 (13)	448	Univariate and multivariate analysis	COX	1	388
Yi et al [42], 2021	China	Hunan Cancer Hospital	Platinum-based combination chemotherapy	26/71 (36)	71/102 (69)	102	LASSO[q]	RF and SVM	1	31
Yu et al [43], 2020	America	TCGA and CPTAC[r]	Platinum-based combination chemotherapy	—	587/587 (100)	—	—	AlexNet[s], GoogLeNet[t], VGGNet[u], SVM, modern deep convolutional neural networks, and multilayer neural networks	1	—
Yu et al [44], 2016	America	TCGA and CPTAC	Platinum-based combination chemotherapy	35/130 (2)	130/130 (100)	130	LASSO	RF, SVM, NB[v], and COX	1	—

XSL•FO
RenderX

Study, year	Country	Sample source	Chemotherapy	Positive samples, n/N (%)	Training set samples, n/N (%)	Overall sample size, n	Variable selection method	Type of model	External validation	External validation sample size
Sun et al [45], 2016	China	Tongji Hospital and Hubei Cancer Hospital	Platinum or taxane-based chemotherapy	43/100 (43)	100/251 (39.8)	251	Univariate analysis	SVM	2	151
Chen et al [46], 2022	China	TCGA or GEO	Platinum-based combination chemotherapy	161/230 (70)	230/305 (75.4)	305	Univariate analysis	RF and COX	1	75
Li et al [47], 2022	China	TCGA or GEO	Platinum-based combination chemotherapy	287/489 (58.7)	489/797 (61.4)	797	LASSO	LR	1	308
Zhao et al [48], 2021	China	TCGA or GEO	Platinum-based combination chemotherapy	—	146/483 (30.2)	483	—	LR	1	337
Buttarelli et al [49], 2022	Italy	TCGA or GEO	Platinum-based combination chemotherapy	7/14 (50)	14/58 (24)	58	—	RF	1	44
Gonzalez Bosquet et al [50], 2016	America	NCI[w] and NHGRI[x]	Platinum-based combination chemotherapy	292/450 (64.9)	450/450 (100)	—	Multivariate analysis	RF, Elastic Net[y], PAM[z], Diagonal Discriminant Analysis, partial least squares–LR, penalized LR, partial least squares, and partial least squares–RF	1	—
Gong et al [51], 2021	China	Shengjing Hospital of China Medical University	Platinum or paclitaxel-based treatment	77/174 (44)	174/174 (100)	—	—	NB, generalized linear model, LR, Fast Large Margin, deep learning, DT, RF, Gradient Boosting Tree, and SVM	1	—
Sun and Yang [52], 2020	China	TCGA	Platinum-based combination chemotherapy	95/320 (29)	320/320 (100)	—	Univariate and multivariate analysis	LR	—	—

XSL•FO
RenderX

Study, year	Country	Sample source	Chemotherapy	Positive samples, n/N (%)	Training set samples, n/N (%)	Overall sample size, n	Variable selection method	Type of model	External validation	External validation sample size
Lei et al [53], 2022	China	The Sun Yat-sen Memorial Hospital	Platinum-based combination chemotherapy	50/62 (80)	62/93 (66)	93	—	Convolutional neural network, principal component analysis, and SVM	1	31

[a]GDSC: Genomics of Drug Sensitivity in Cancer.
[b]TCGA: The Cancer Genome Atlas.
[c]GEO: Gene Expression Omnibus.
[d]Missing data or not applicable.
[e]SVM: support vector machine.
[f]LR: logistic regression.
[g]KNN: k-nearest neighbor.
[h]DT: decision tree.
[i]AdaBoost: Adaptive Boosting.
[j]GBM: Gradient Boosting Machine.
[k]XGBoost: Extreme Gradient Boosting.
[l]RF: random forest.
[m]DNN: deep neural network.
[n]COX: Cox Proportional Hazards Regression and Log-Rank Tests.
[o]ANN: artificial neural network.
[p]SPC: supervised principal component.
[q]LASSO: Least Absolute Shrinkage and Selection Operator.
[r]CPTAC: Clinical Proteomic Tumor Analysis Consortium.
[s]AlexNet: Alexandria Network.
[t]GoogLeNet: Google's Network.
[u]VGGNet: Visual Geometry Group Network.
[v]NB: naive Bayes.
[w]NCI: National Cancer Institute.
[x]NHGRI: National Human Genome Research Institute.
[y]Elastic Net: Elastic Net Regularization.
[z]PAM: Prediction Analysis of Microarrays.

Quality Assessment of Included Studies Using PROBAST

PROBAST was used to evaluate the risk of bias in eligible articles that constituted or externally validated predictive models. Figure 2 summarizes the risk of bias in the 39 predictive models. Overall, 2 models had a low risk of bias in research participants, 2 models had a low risk of bias in predictors, 4 models had a low risk of bias in outcomes, and none had a low risk of bias in statistical analysis (Multimedia Appendix 3). Some models were at a high risk of bias, suggesting that their real predictive performance may be worse than that previously reported. Therefore, we are reasonably concerned that these models may be unreliable when used by others.

XSL•FO
RenderX

Figure 2. Risk of bias assessment for all eligible studies.

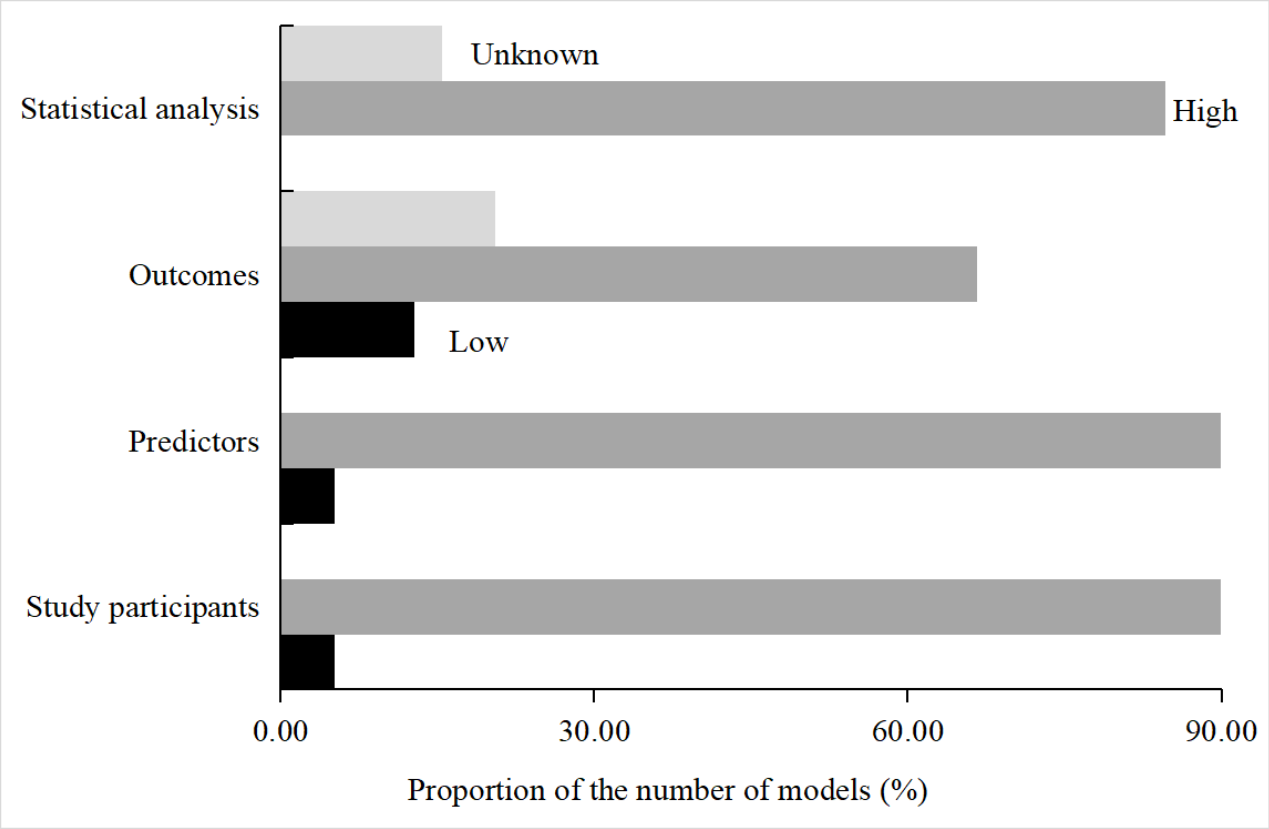

Model Performance Evaluation

Discrimination and calibration are the most commonly used indicators for assessing the prediction model performance. Discrimination is usually evaluated by the area under the receiver operating characteristic curve, namely the C-index, which is between 0.5 and 1. A higher C-index indicates better discrimination in the prediction model. In general, a random effects model was used for the meta-analysis of C-index in 39 predictive models (XGBoost, LR, Least Absolute Shrinkage and Selection Operator [LASSO], SVM, random forest, convolutional neural networks, artificial neural networks, Prediction Analysis of Microarrays, Diagonal Discriminant Analysis, Penalized Logistic Regression, partial least squares, and supervised principal components method). The training cohort reported C-index in 39 predictive models, with a pooled value of 0.806 (95% CI 0.767-0.846); the validation cohort reported C-index in 12 predictive models, with a pooled value of 0.831 (95% CI 0.768-0.895). We conducted subgroup analyses according to the machine learning model type, histological type of ovarian cancer, and whether there was residual tumor after surgery. In terms of the subgroup analysis of model types, the C-index for the models in the training set was as follows—XGBoost: 0.861 (95% CI 0.808-0.914), LR: 0.816 (95% CI 0.775-0.858), SVM: 0.942 (95% CI 0.897-0.988), and ANN: 0.705 (95% CI 0.615-0.796); the C-index for the models in the test set were LR: 0.821 (95% CI 0.767-0.876), LASSO: 0.808 (95% CI 0.723-0.893), SVM: 0.879 (95% CI 0.808-0.949), and random forest: 0.909 (95% CI 0.868-0.950). With regard to the subgroup analysis of pathological types, the C-index in the training cohort was serous adenocarcinoma (0.751, 95% CI 0.682-0.820), high-grade serous ovarian cancers (0.837, 95% CI 0.780-0.894), and unclear (0.800, 95% CI 0.749-0.852); the C-index in the test set was high-grade serous ovarian cancers (0.786, 95% CI 0.679-0.893) and unclear (0.916, 95% CI 0.875-0.958). Meanwhile, in the subgroup analysis of residual tumor, the C-index for residual tumor in the training cohort was 0.767 (95% CI 0.732-0.803) and the C-index for nonresidual tumor was 0.811 (95% CI 0.770-0.852). In the test set, the C-index for residual tumor was 0.719 (95% CI 0.638-0.801) and the C-index for nonresidual tumor was 0.889 (95% CI 0.835-0.943). The forest plot for the subgroup analysis is shown in Figure 3. Table 2 presents the meta-analysis results of the C-index. High heterogeneity was identified among these studies (I^2=97.3%; P≤.001), probably because of the varied machine learning methods and variables used in these studies. Furthermore, a meta-analysis of the sensitivity and specificity of the 22 predictive models was performed. The pooled sensitivity was 0.890 (95% CI 0.800-0.950) and the pooled specificity was 0.790 (95% CI 0.720-0.840) in the training set (Figure 4) [35-37,39,42,46,47,50,52,53]. In the test set, the pooled sensitivity was 0.920 (95% CI 0.810-0.970) and the pooled specificity was 0.910 (95% CI 0.760-0.970; Figure 5) [42,45-48,51,53].

XSL•FO
RenderX

Figure 3. Forest plot of subgroup analysis. ANN: artificial neural network; CNN: convolutional neural network; DDA: Diagonal Discriminant Analysis; HGSOC: high-grade serous ovarian cancer; LASSO: Least Absolute Shrinkage and Selection Operator; LR: logistic regression; PAM: Prediction Analysis of Microarrays; PLR: penalized logistic regression; PLS: partial least squares; RF: random forest; SPC: supervised principal component; SVM: support vector machine; XGBoost: Extreme Gradient Boosting.

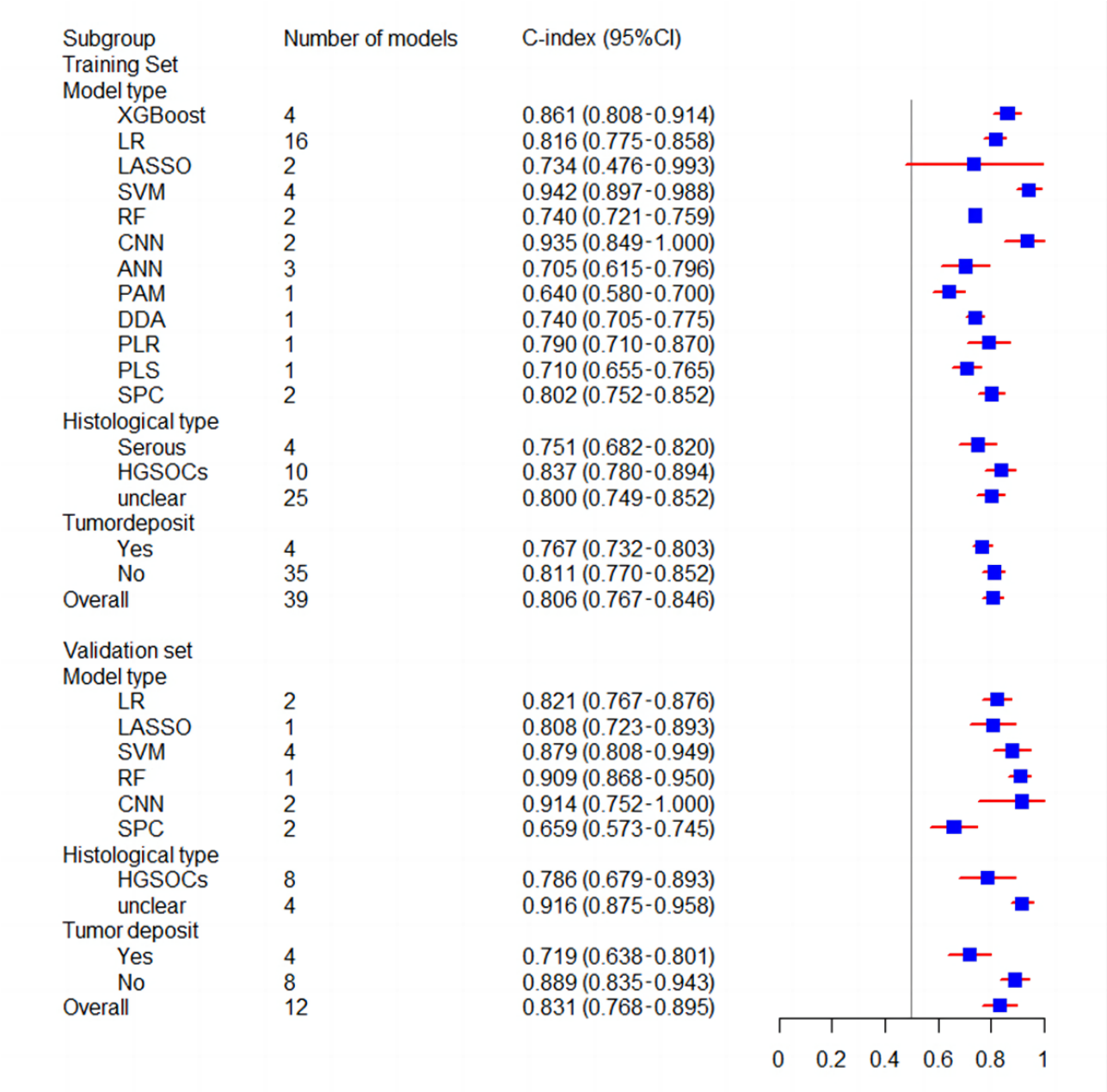

XSL•FO
RenderX

Table 2. Meta-analysis results of predicting platinum chemotherapy responses in ovarian cancer.

	Training set		Test set	
	Number (n=39), n (%)	C-index (95% CI)	Number (n=12), n (%)	C-index (95% CI)
Model				
XGBoost[a]	4 (10)	0.861 (0.808-0.914)	—[b]	—
LR[c]	16 (41)	0.816 (0.775-0.858)	2 (17)	0.821 (0.767-0.876)
LASSO[d]	2 (5)	0.734 (0.476-0.993)	1 (8)	0.808 (0.723-0.893)
SVM[e]	4 (10)	0.942 (0.897-0.988)	4 (33)	0.879 (0.808-0.949)
RF[f]	2 (5)	0.740 (0.721-0.759)	1 (8)	0.909 (0.868-0.950)
CNN[g]	2 (5)	0.935 (0.849-1.000)	2 (17)	0.914 (0.752-1.000)
ANN[h]	3 (8)	0.705 (0.615-0.796)	—	—
PAM[i]	1 (3)	0.640 (0.580-0.700)	—	—
DDA[j]	1 (3)	0.740 (0.705-0.775)	—	—
PLR[k]	1 (3)	0.790 (0.710-0.870)	—	—
PLS[l]	1 (3)	0.710 (0.655-0.765)	—	—
SPC[m]	2 (5)	0.802 (0.752-0.852)	2 (17)	0.659 (0.573-0.745)
Histological type				
Serous	4 (10)	0.751 (0.682-0.820)	—	—
HGSOC[n]	10 (26)	0.837 (0.780-0.894)	8 (67)	0.786 (0.679-0.893)
Unclear	25 (64)	0.800 (0.749-0.852)	4 (33)	0.916 (0.875-0.958)
Residual tumor				
Yes	4 (10)	0.767 (0.732-0.803)	4 (33)	0.719 (0.638-0.801)
No	35 (90)	0.811 (0.770-0.852)	8 (67)	0.889 (0.835-0.943)
Overall	39 (100)	0.806 (0.767-0.846)	12 (100)	0.831 (0.768-0.895)

[a]XGBoost: Extreme Gradient Boosting.
[b]Missing data.
[c]LR: logistic regression.
[d]LASSO: Least Absolute Shrinkage and Selection Operator.
[e]SVM: support vector machine.
[f]RF: random forest.
[g]CNN: convolutional neural network.
[h]ANN: artificial neural network.
[i]PAM: Prediction Analysis of Microarrays.
[j]DDA: Diagonal Discriminant Analysis.
[k]PLR: penalized logistic regression.
[l]PLS: partial least squares.
[m]SPC: supervised principal component.
[n]HGSOC: high-grade serous ovarian cancer.

XSL•FO
RenderX

Figure 4. Meta-analysis of sensitivity and specificity—the training set. The repeated authors in the literature are due to the construction of multiple machine learning models. CNN: convolutional neural network; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBoost: Extreme Gradient Boosting.

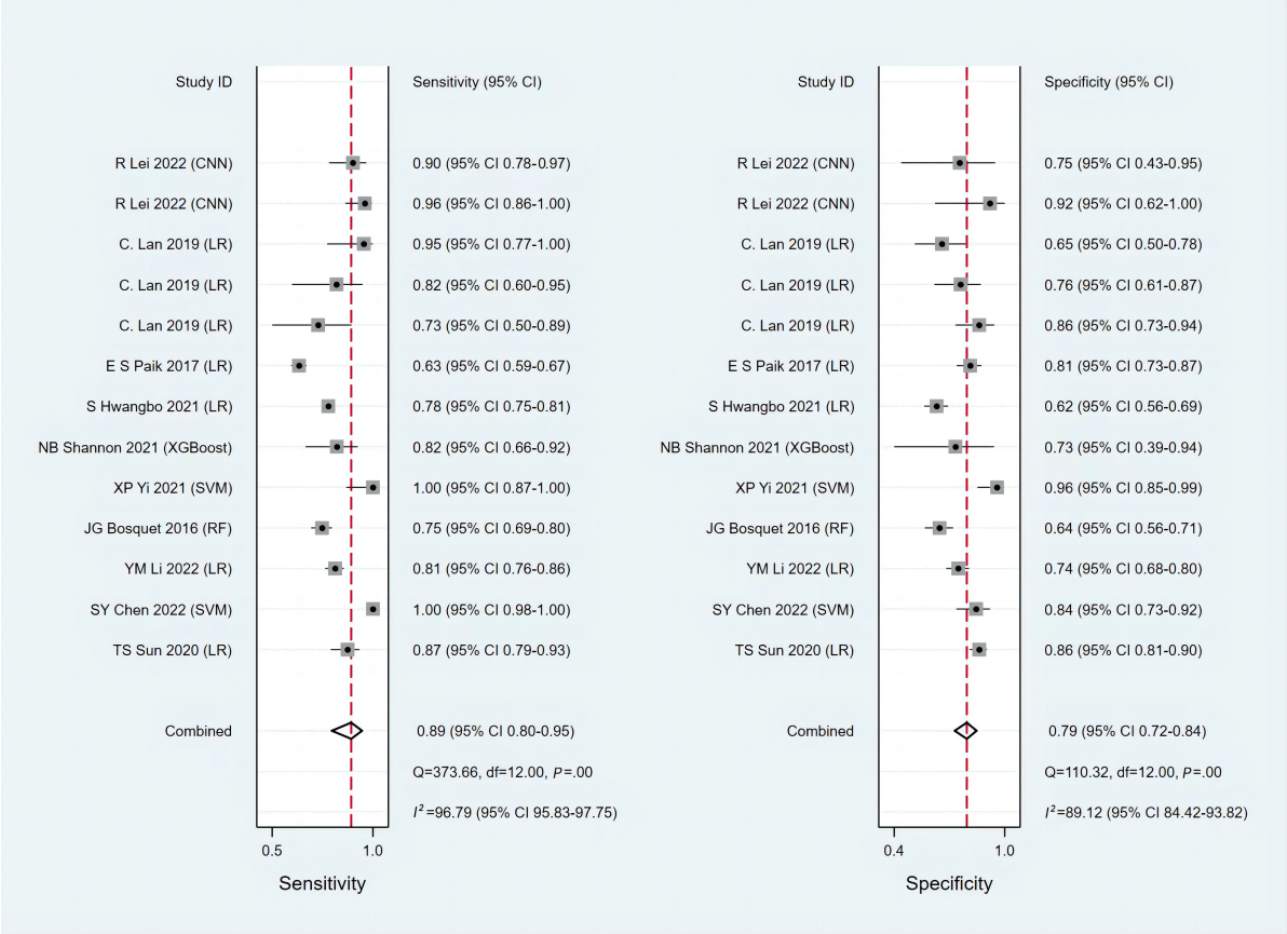

XSL•FO
RenderX

Figure 5. Meta-analysis of sensitivity and specificity—the test set. The repeated authors in the literature are due to the construction of multiple machine learning models. CNN: convolutional neural network; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBoost: Extreme Gradient Boosting.

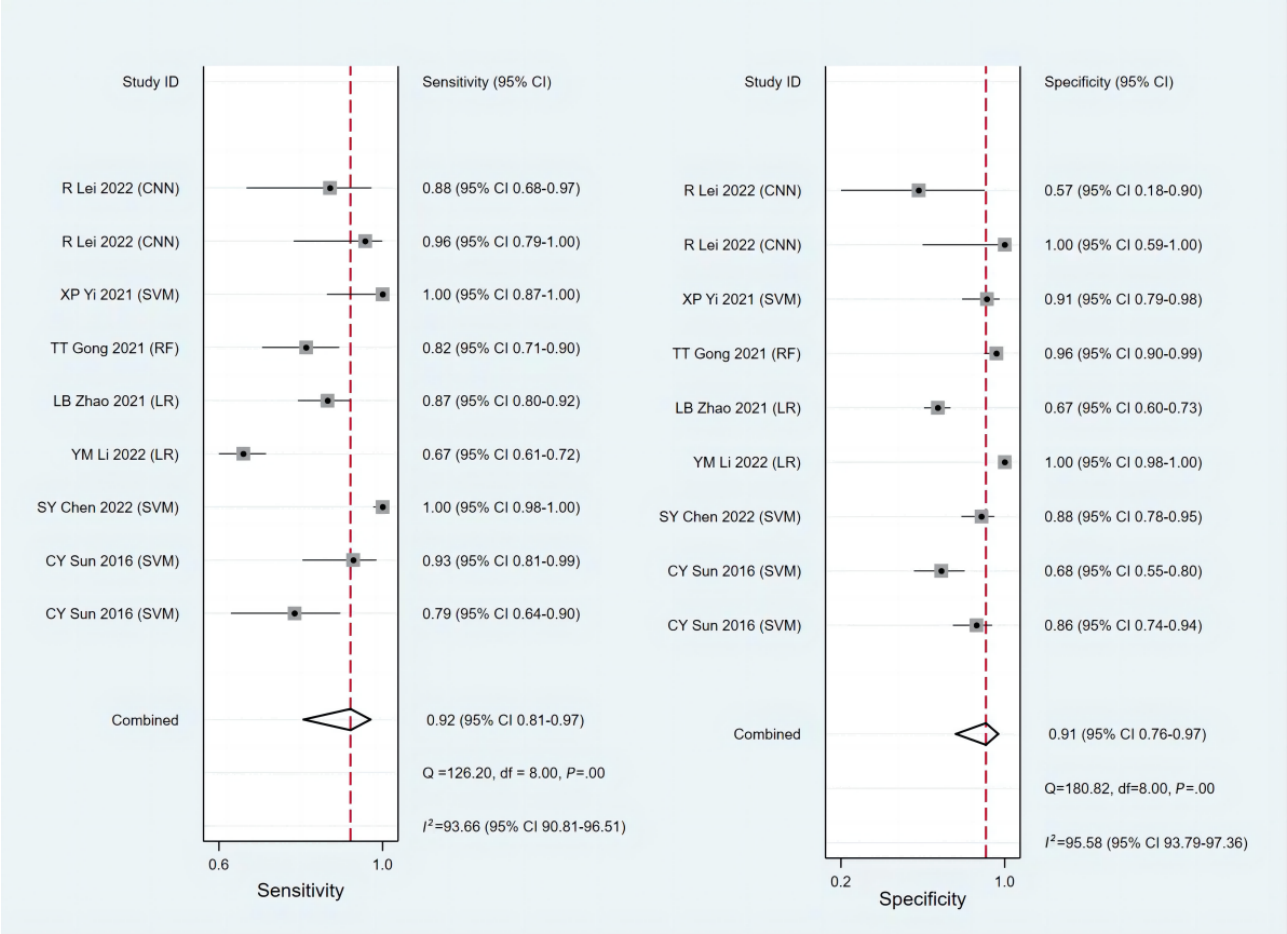

Discussion

Principal Findings

This study conducted a meta-analysis of machine learning models for predicting responses to platinum chemotherapy in patients with ovarian cancer. It delves into the performance, reliability, and influencing factors of models. To our knowledge, this is the first systematic review and meta-analysis on the application of machine learning in predicting responses to platinum-based chemotherapy in patients with ovarian cancer. The search initially yielded 1749 studies, and after applying inclusion criteria, 19 studies (accounting for 1.09% of the total) were ultimately included. This research encompasses 12 machine learning models, such as XGBoost, LR, LASSO, and SVM, built based on various hospital or genomics data sources. The analysis results indicated that these models performed effectively in distinguishing patients' responses to platinum chemotherapy, achieving C-indices of 0.806 and 0.831 in the training and validation sets, respectively. The model demonstrated high overall sensitivity and specificity, underscoring its accuracy and reliability in predicting platinum drug response in ovarian cancer. Subgroup analysis revealed the influence of model type, pathology type, and residual tumor on the prediction performance. SVM stood out on both the training and validation sets because it outperformed other machine learning methods in terms of accuracy and relative error rate measures [54] and exhibited the ability to identify subtle patterns in complex data sets [55]. LR is the most commonly used modeling variable because it can handle not only binary results but also accommodate continuous or categorical predictor variables. This comprehensive approach considers the impact of multiple factors on the results, effectively controls potential confounding factors, and reduces bias [56]. As a result, LR is widely used in machine learning modeling within various fields. The analysis of residual tumor revealed that the model exhibited different performance in predicting patients with or without residual tumor. Compared with nonresidual tumor, the predictive performance of machine learning for residual tumor was more significant, suggesting that residual tumor may be a crucial factor influencing ovarian cancer patients' response to platinum therapy.

Most published meta-analyses on the application of machine learning in ovarian cancer focus on the diagnosis and prediction of ovarian cancer; however, there are some differences in specific research methods, evaluation tools, and presentation of results. Huang et al [57] reviewed the application of computed tomography and magnetic resonance imaging radiomics in ovarian cancer, achieving promising results in differential diagnosis and prognosis prediction. Other studies [58,59] have summarized artificial intelligence methods for gynecological malignant tumors, emphasizing that variable selection, machine learning methods, and end point selection can all influence

XSL•FO
RenderX

model performance. Xu et al [60] systematically reviewed studies that applied artificial intelligence to diagnose ovarian cancer based on medical images and highlighted the good performance of artificial intelligence algorithms in ovarian cancer diagnosis. Koch et al [61] evaluated the accuracy of computer-aided diagnosis, encompassing computer-aided diagnosis for ultrasound, computed tomography, and magnetic resonance imaging, to predict the likelihood of malignancy in ovarian tumors. Given that it is challenging to predict the response of patients with ovarian cancer to platinum therapy before the completion of chemotherapy, accurate prediction of this response is crucial for devising effective treatment plans. This review focuses on the performance of machine learning in predicting responses to platinum-based chemotherapy in patients with ovarian cancer. This not only provides valuable information for clinical prediction but also addresses a long-standing challenge in the development of noninvasive methods for predicting chemotherapy response in patients with ovarian cancer. Feature selection emerges as a critical aspect influencing model performance in this context. Previous studies [41,62] have reported that next-generation sequencing technology can be used to explore correlations between intrinsic genomic features and the response to platinum-based chemotherapy. Radiomics is another approach. A recent study demonstrated that a predictive model based on the combination of radiomics with single nucleotide polymorphisms of Human Sulfatase 1 could predict platinum resistance in ovarian cancer treatment [42]. Previous research [43] has shown that combining whole-slide histopathology scanners and high-throughput omics analysis with cutting-edge machine learning algorithms can help reveal correlations between microscopic tumor cell morphology and molecular pathways. Machine learning models have shown great promise in linking histopathological patterns to patient diagnosis and prognosis. Another study [44] used tumor proteomic features to predict the clinical response to platinum-based chemotherapy in patients with ovarian cancer. The findings revealed a close association between tissue expression levels of 24 proteins and the response to platinum-based chemotherapy. The variables selected in the 19 included articles spanned from molecular-level factors to clinical characteristics, medical imaging, and the microbiome, reflecting the prevailing trend of considering multiple levels and aspects in cancer research. This comprehensive approach facilitates a more in-depth understanding of cancer pathogenesis and predictive factors.

Strengths and Limitations

Strengths

The most noteworthy aspect of our analysis is that it provides a comprehensive map of research on prognostic prediction models for patients with ovarian cancer. We gathered all available predictive models for potential clinical outcomes of platinum chemotherapy responses in patients with ovarian cancer. The characteristics of these models were elucidated in detail. Furthermore, this study critically evaluated the predictive models for platinum chemotherapy response in patients with ovarian cancer using the PROBAST tool. Moreover, a meta-analysis of the C-index using multiple externally verified predictive models was performed. There is currently no meta-analysis that summarizes research on machine learning prediction models for platinum chemotherapy response in ovarian cancer. Hence, this study aimed to explore its performance in prediction. It is critical to systematically review published studies on machine learning and provide guidance for future research. This helps to establish personalized treatment protocols and estimate prognosis by elucidating intrinsic tumor features such as platinum sensitivity in the initial therapy.

Limitations

However, several limitations of the current investigation must be considered. First, the meta-analysis of the C-index had a high degree of heterogeneity, probably because of the various machine learning methods, predictors, and parameters used in model construction, as well as differences in clinical settings, patient characteristics, and research time. At the same time, we should note that the risk of bias assessment of the predictive model is a rigorous tool for the construction of original models; for most of the original studies, the results assessed by this tool have a high risk of bias. In addition, our meta-analysis had several methodological problems in model development, which were reflected in the risk of bias. The PROBAST assessment suggested that some studies had a high or unclear risk of bias in 4 domains: participants, predictors, outcomes, and statistical analysis. Furthermore, the predictive value of machine learning for different diseases may vary. The essence of machine learning is efficient predictors. When the same machine learning model includes more efficient predictors, its predictive value will be significantly improved. This may result in heterogeneity between models. For constructing machine learning, especially for rare events, some studies face challenges in acquiring large data sets, making it difficult to establish an independent validation set. However, the importance of the training process cannot be overlooked as cross-validation may be used during training, although it cannot fully replace an independent validation set. When conducting meta-analysis, it is essential to consider whether the model is overfitting, necessitating attention to the results of the training set. Consequently, our meta-analysis includes studies without independent validation cohorts for a comprehensive evaluation. The most important aspect is the lack of original research with large multicenter samples in the modeling process. Therefore, more high-quality, multicenter, large-scale studies are required. Despite some limitations in this study, we have compiled a comprehensive summary of the current models to provide a reference for the development of more broadly applicable clinical tools in the future. Looking at it this way, it is necessary to conduct a meta-analysis. Although there are frequent disagreements about the predictive value of different studies, this is partially dependent on the selection of the predictive model, which is the most influential factor affecting predictive performance.

Conclusions

Machine learning has excellent predictive performance in predicting response to platinum chemotherapy in patients with ovarian cancer. At the same time, we found that SVM has the best prediction performance among the existing prediction models. Machine learning can be used as a prediction tool for platinum response in ovarian cancer. On the basis of this

XSL•FO
RenderX

research, a large-scale, multicenter, and multiethnic prediction tool can be developed in the future for predicting platinum-based chemotherapy response in patients with ovarian cancer to advance precision chemotherapy for ovarian cancer management.

Acknowledgments

This study was supported by the National Nature Science Foundation of China (grants 81973894 and 82174421).

The authors would like to thank the researchers and study participants for their contributions.

Data Availability

The original contributions presented in this study are included in the paper or multimedia appendices, and further inquiries can be directed to the corresponding author.

Authors' Contributions

QW was involved in study conceptualization and investigation and wrote the original draft. ZC participated in study visualization, supervised the study, and reviewed and edited the draft. XL participated in the investigation and prepared the original draft. YW and CF participated in the investigation and prepared the original draft. YP participated in the investigation and gathered resources. XF was involved in project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 27 KB - jmir_v26i1e48527_app1.docx]

Multimedia Appendix 2
The specific retrieval strategy of this paper.
[DOCX File , 16 KB - jmir_v26i1e48527_app2.docx]

Multimedia Appendix 3
Results of bias risk assessment through Prediction Model Risk of Bias Assessment in 19 included articles.
[DOCX File , 21 KB - jmir_v26i1e48527_app3.docx]

References

1. Zachou G, El-Khouly F, Dilley J. Evaluation of follow-up strategies for women with epithelial ovarian cancer following completion of primary treatment. Cochrane Database Syst Rev 2023 Aug 31;8(8):CD006119. [doi: 10.1002/14651858.CD006119.pub4] [Medline: 37650760]
2. Coburn SB, Bray F, Sherman ME, Trabert B. International patterns and trends in ovarian cancer incidence, overall and by histologic subtype. Int J Cancer 2017 Jun 01;140(11):2451-2460 [FREE Full text] [doi: 10.1002/ijc.30676] [Medline: 28257597]
3. Retamales-Ortega R, Oróstica L, Vera C, Cuevas P, Hernández A, Hurtado I, et al. Role of Nerve Growth Factor (NGF) and miRNAs in Epithelial Ovarian Cancer. Int J Mol Sci 2017 Mar 26;18(3):507 [FREE Full text] [doi: 10.3390/ijms18030507] [Medline: 28245631]
4. Zhu JW, Charkhchi P, Akbari MR. Potential clinical utility of liquid biopsies in ovarian cancer. Mol Cancer 2022 May 11;21(1):114 [FREE Full text] [doi: 10.1186/s12943-022-01588-8] [Medline: 35545786]
5. Torre LA, Trabert B, DeSantis CE, Miller KD, Samimi G, Runowicz CD, et al. Ovarian cancer statistics, 2018. CA Cancer J Clin 2018 Jul;68(4):284-296 [FREE Full text] [doi: 10.3322/caac.21456] [Medline: 29809280]
6. Kania KD, Haręża D, Wilczyński JR, Wilczyński M, Jarych D, Malinowski A, et al. The toll-like receptor 4 polymorphism Asp299Gly is associated with an increased risk of ovarian cancer. Cells 2022 Oct 05;11(19):3137 [FREE Full text] [doi: 10.3390/cells11193137] [Medline: 36231099]
7. Lheureux S, Gourley C, Vergote I, Oza AM. Epithelial ovarian cancer. Lancet 2019 Mar 23;393(10177):1240-1253. [doi: 10.1016/S0140-6736(18)32552-2] [Medline: 30910306]
8. Cabasag CJ, Fagan PJ, Ferlay J, Vignat J, Laversanne M, Liu L, et al. Ovarian cancer today and tomorrow: a global assessment by world region and Human Development Index using GLOBOCAN 2020. Int J Cancer 2022 Nov 01;151(9):1535-1541 [FREE Full text] [doi: 10.1002/ijc.34002] [Medline: 35322413]

XSL•FO
RenderX

9. Xu Y, Liu H, Chen J, Zhou X. NCCN clinical practice guidelines in oncology: ovarian cancer continue including fallopian tube cancer and primary peritoneal cancer(version 2.2023). National Comprehensive Cancer Network. URL: https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1453 [accessed 2023-12-10]
10. Noriega-Rivera R, Rivera-Serrano M, Rabelo-Fernandez RJ, Pérez-Santiago J, Valiyeva F, Vivas-Mejía PE. Upregulation of the long noncoding RNA CASC10 promotes cisplatin resistance in high-grade serous ovarian cancer. Int J Mol Sci 2022 Jul 13;23(14):7737 [FREE Full text] [doi: 10.3390/ijms23147737] [Medline: 35887085]
11. Bogani G, Rossetti D, Ditto A, Martinelli F, Chiappa V, Mosca L, et al. Artificial intelligence weights the importance of factors predicting complete cytoreduction at secondary cytoreductive surgery for recurrent ovarian cancer. J Gynecol Oncol 2018 Sep;29(5):e66 [FREE Full text] [doi: 10.3802/jgo.2018.29.e66] [Medline: 30022630]
12. Kurtz JE, Pujade-Lauraine E, Oaknin A, Belin L, Leitner K, Cibula D, ATALANTE/ENGOT-ov29 Investigators. Atezolizumab combined with bevacizumab and platinum-based therapy for platinum-sensitive ovarian cancer: placebo-controlled randomized phase III ATALANTE/ENGOT-ov29 trial. J Clin Oncol 2023 Oct 20;41(30):4768-4778 [FREE Full text] [doi: 10.1200/JCO.23.00529] [Medline: 37643382]
13. Wilson MK, Pujade-Lauraine E, Aoki D, Mirza MR, Lorusso D, Oza AM, participants of the Fifth Ovarian Cancer Consensus Conference. Fifth Ovarian Cancer Consensus Conference of the Gynecologic Cancer InterGroup: recurrent disease. Ann Oncol 2017 Apr 01;28(4):727-732 [FREE Full text] [doi: 10.1093/annonc/mdw663] [Medline: 27993805]
14. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011 Jun 29;474(7353):609-615 [FREE Full text] [doi: 10.1038/nature10166] [Medline: 21720365]
15. da Costa AA, Baiocchi G. Genomic profiling of platinum-resistant ovarian cancer: the road into druggable targets. Semin Cancer Biol 2021 Dec;77:29-41. [doi: 10.1016/j.semcancer.2020.10.016] [Medline: 33161141]
16. Chatterjee N, Bivona TG. Polytherapy and targeted cancer drug resistance. Trends Cancer 2019 Mar;5(3):170-182 [FREE Full text] [doi: 10.1016/j.trecan.2019.02.003] [Medline: 30898264]
17. Borley J, Brown R. Epigenetic mechanisms and therapeutic targets of chemotherapy resistance in epithelial ovarian cancer. Ann Med 2015;47(5):359-369. [doi: 10.3109/07853890.2015.1043140] [Medline: 26158617]
18. Natanzon Y, Goode EL, Cunningham JM. Epigenetics in ovarian cancer. Semin Cancer Biol 2018 Aug;51:160-169 [FREE Full text] [doi: 10.1016/j.semcancer.2017.08.003] [Medline: 28782606]
19. Kleih M, Böpple K, Dong M, Gaißler A, Heine S, Olayioye MA, et al. Direct impact of cisplatin on mitochondria induces ROS production that dictates cell fate of ovarian cancer cells. Cell Death Dis 2019 Nov 07;10(11):851. [doi: 10.1038/s41419-019-2081-4] [Medline: 31699970]
20. Han Y, Kim B, Cho U, Park IS, Kim SI, Dhanasekaran DN, et al. Mitochondrial fission causes cisplatin resistance under hypoxic conditions via ROS in ovarian cancer cells. Oncogene 2019 Nov;38(45):7089-7105. [doi: 10.1038/s41388-019-0949-5] [Medline: 31409904]
21. Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: evolution of management in the era of precision medicine. CA Cancer J Clin 2019 Jul;69(4):280-304 [FREE Full text] [doi: 10.3322/caac.21559] [Medline: 31099893]
22. Funingana IG, Reinius MA, Petrillo A, Ang JE, Brenton JD. Can integrative biomarker approaches improve prediction of platinum and PARP inhibitor response in ovarian cancer? Semin Cancer Biol 2021 Dec;77:67-82. [doi: 10.1016/j.semcancer.2021.02.008] [Medline: 33607245]
23. Yearley AG, Goedmakers CM, Panahi A, Doucette J, Rana A, Ranganathan K, et al. FDA-approved machine learning algorithms in neuroradiology: a systematic review of the current evidence for approval. Artif Intell Med 2023 Sep;143:102607. [doi: 10.1016/j.artmed.2023.102607] [Medline: 37673576]
24. Kumari J, Kumar E, Kumar D. A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. Arch Comput Methods Eng 2023 Mar 31;30(6):1-29 [FREE Full text] [doi: 10.1007/s11831-023-09915-y] [Medline: 37359744]
25. Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ, Boza G, Racimo F, Gopalakrishnan S, et al. Dealing with dimensionality: the application of machine learning to multi-omics data. Bioinformatics 2023 Feb 03;39(2):btad021 [FREE Full text] [doi: 10.1093/bioinformatics/btad021] [Medline: 36637211]
26. Monaco A, Pantaleo E, Amoroso N, Lacalamita A, Lo Giudice C, Fonzino A, et al. A primer on machine learning techniques for genomic applications. Comput Struct Biotechnol J 2021;19:4345-4359 [FREE Full text] [doi: 10.1016/j.csbj.2021.07.021] [Medline: 34429852]
27. Li Y, Wu X, Yang P, Jiang G, Luo Y. Machine learning for lung cancer diagnosis, treatment, and prognosis. Genomics Proteomics Bioinformatics 2022 Oct;20(5):850-866 [FREE Full text] [doi: 10.1016/j.gpb.2022.11.003] [Medline: 36462630]
28. Li G, Li L, Li Y, Qian Z, Wu F, He Y, et al. An MRI radiomics approach to predict survival and tumour-infiltrating macrophages in gliomas. Brain 2022 Apr 29;145(3):1151-1161 [FREE Full text] [doi: 10.1093/brain/awab340] [Medline: 35136934]
29. Li X, Dowling EK, Yan G, Dereli Z, Bozorgui B, Imanirad P, et al. Precision combination therapies based on recurrent oncogenic coalterations. Cancer Discov 2022 Jun 02;12(6):1542-1559 [FREE Full text] [doi: 10.1158/2159-8290.CD-21-0832] [Medline: 35412613]
30. Issa NT, Stathias V, Schürer S, Dakshanamurthy S. Machine and deep learning approaches for cancer drug repurposing. Semin Cancer Biol 2021 Jan;68:132-142 [FREE Full text] [doi: 10.1016/j.semcancer.2019.12.011] [Medline: 31904426]

XSL•FO
RenderX

31. Kong J, Lee H, Kim D, Han SK, Ha D, Shin K, et al. Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. Nat Commun 2020 Oct 30;11(1):5485 [FREE Full text] [doi: 10.1038/s41467-020-19313-8] [Medline: 33127883]
32. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 01;170(1):W1-33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
33. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
34. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786 [FREE Full text] [doi: 10.1177/0962280218785504] [Medline: 30032705]
35. Shannon NB, Tan LL, Tan QX, Tan JW, Hendrikson J, Ng WH, et al. A machine learning approach to identify predictive molecular markers for cisplatin chemosensitivity following surgical resection in ovarian cancer. Sci Rep 2021 Aug 19;11(1):16829 [FREE Full text] [doi: 10.1038/s41598-021-96072-6] [Medline: 34413360]
36. Lan C, Li J, Huang X, Heindl A, Wang Y, Yan S, et al. Stromal cell ratio based on automated image analysis as a predictor for platinum-resistant recurrent ovarian cancer. BMC Cancer 2019 Feb 18;19(1):159 [FREE Full text] [doi: 10.1186/s12885-019-5343-8] [Medline: 30777045]
37. Hwangbo S, Kim SI, Kim JH, Eoh KJ, Lee C, Kim YT, et al. Development of machine learning models to predict platinum sensitivity of high-grade serous ovarian carcinoma. Cancers (Basel) 2021 Apr 14;13(8):1875 [FREE Full text] [doi: 10.3390/cancers13081875] [Medline: 33919797]
38. Zhao H, Sun Q, Li L, Zhou J, Zhang C, Hu T, et al. High expression levels of AGGF1 and MFAP4 predict primary platinum-based chemoresistance and are associated with adverse prognosis in patients with serous ovarian cancer. J Cancer 2019 Jan 01;10(2):397-407 [FREE Full text] [doi: 10.7150/jca.28127] [Medline: 30719133]
39. Paik ES, Sohn I, Baek S, Shim M, Choi HJ, Kim T, et al. Nomograms predicting platinum sensitivity, progression-free survival, and overall survival using pretreatment complete blood cell counts in epithelial ovarian cancer. Cancer Res Treat 2017 Jul;49(3):635-642 [FREE Full text] [doi: 10.4143/crt.2016.282] [Medline: 27669704]
40. Han Y, Huang H, Xiao Z, Zhang W, Cao Y, Qu L, et al. Integrated analysis of gene expression profiles associated with response of platinum/paclitaxel-based treatment in epithelial ovarian cancer. PLoS One 2012;7(12):e52745 [FREE Full text] [doi: 10.1371/journal.pone.0052745] [Medline: 23300757]
41. Zheng H, Shu T, Zhu S, Zhang C, Gao M, Zhang N, et al. Construction and validation of a platinum sensitivity predictive model with multiple genomic variations for epithelial ovarian cancer. Front Oncol 2021 Sep 16;11:725264 [FREE Full text] [doi: 10.3389/fonc.2021.725264] [Medline: 34604063]
42. Yi X, Liu Y, Zhou B, Xiang W, Deng A, Fu Y, et al. Incorporating SULF1 polymorphisms in a pretreatment CT-based radiomic model for predicting platinum resistance in ovarian cancer treatment. Biomed Pharmacother 2021 Jan;133:111013 [FREE Full text] [doi: 10.1016/j.biopha.2020.111013] [Medline: 33227705]
43. Yu K, Hu V, Wang F, Matulonis UA, Mutter GL, Golden JA, et al. Deciphering serous ovarian carcinoma histopathology and platinum response by convolutional neural networks. BMC Med 2020 Aug 18;18(1):236 [FREE Full text] [doi: 10.1186/s12916-020-01684-w] [Medline: 32807164]
44. Yu KH, Levine DA, Zhang H, Chan DW, Zhang Z, Snyder M. Predicting ovarian cancer patients' clinical response to platinum-based chemotherapy by their tumor proteomic signatures. J Proteome Res 2016 Aug 05;15(8):2455-2465 [FREE Full text] [doi: 10.1021/acs.jproteome.5b01129] [Medline: 27312948]
45. Sun CY, Su TF, Li N, Zhou B, Guo ES, Yang ZY, et al. A chemotherapy response classifier based on support vector machines for high-grade serous ovarian carcinoma. Oncotarget 2016 Jan 19;7(3):3245-3254 [FREE Full text] [doi: 10.18632/oncotarget.6569] [Medline: 26675546]
46. Chen S, Wu Y, Wang S, Wu J, Wu X, Zheng Z. A risk model of gene signatures for predicting platinum response and survival in ovarian cancer. J Ovarian Res 2022 Mar 31;15(1):39 [FREE Full text] [doi: 10.1186/s13048-022-00969-3] [Medline: 35361267]
47. Li Y, Nie Y, Guo H, Guo H, Ha C, Li Y. Establish of an initial platinum-resistance predictor in high-grade serous ovarian cancer patients regardless of homologous recombination deficiency status. Front Oncol 2022 Mar 18;12:847085 [FREE Full text] [doi: 10.3389/fonc.2022.847085] [Medline: 35372049]
48. Zhao L, Ma S, Wang L, Wang Y, Feng X, Liang D, et al. A polygenic methylation prediction model associated with response to chemotherapy in epithelial ovarian cancer. Mol Ther Oncolytics 2021 Feb 20;20:545-555 [FREE Full text] [doi: 10.1016/j.omto.2021.02.012] [Medline: 33738340]
49. Buttarelli M, Ciucci A, Palluzzi F, Raspaglio G, Marchetti C, Perrone E, et al. Identification of a novel gene signature predicting response to first-line chemotherapy in BRCA wild-type high-grade serous ovarian cancer patients. J Exp Clin Cancer Res 2022 Feb 04;41(1):50 [FREE Full text] [doi: 10.1186/s13046-022-02265-w] [Medline: 35120576]

XSL•FO
RenderX

50. Gonzalez Bosquet J, Newtson AM, Chung RK, Thiel KW, Ginader T, Goodheart MJ, et al. Prediction of chemo-response in serous ovarian cancer. Mol Cancer 2016 Oct 19;15(1):66 [FREE Full text] [doi: 10.1186/s12943-016-0548-9] [Medline: 27756408]
51. Gong TT, He XH, Gao S, Wu QJ. Application of machine learning in prediction of chemotherapy resistant of ovarian cancer based on gut microbiota. J Cancer 2021 Mar 15;12(10):2877-2885 [FREE Full text] [doi: 10.7150/jca.46621] [Medline: 33854588]
52. Sun T, Yang Q. Chemoresistance-associated alternative splicing signatures in serous ovarian cancer. Oncol Lett 2020 Jul;20(1):420-430 [FREE Full text] [doi: 10.3892/ol.2020.11562] [Medline: 32565967]
53. Lei R, Yu Y, Li Q, Yao Q, Wang J, Gao M, et al. Deep learning magnetic resonance imaging predicts platinum sensitivity in patients with epithelial ovarian cancer. Front Oncol 2022 Nov 23;12:895177 [FREE Full text] [doi: 10.3389/fonc.2022.895177] [Medline: 36505880]
54. Nedaie A, Najafi AA. Support vector machine with Dirichlet feature mapping. Neural Netw 2018 Feb;98:87-101. [doi: 10.1016/j.neunet.2017.11.006] [Medline: 29223012]
55. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 2018 Jan 2;15(1):41-51 [FREE Full text] [doi: 10.21873/cgp.20063] [Medline: 29275361]
56. LaValley MP. Logistic regression. Circulation 2008 May 06;117(18):2395-2399. [doi: 10.1161/circulationaha.106.682658]
57. Huang ML, Ren J, Jin ZY, Liu XY, He YL, Li Y, et al. A systematic review and meta-analysis of CT and MRI radiomics in ovarian cancer: methodological issues and clinical utility. Insights Imaging 2023 Jul 03;14(1):117 [FREE Full text] [doi: 10.1186/s13244-023-01464-z] [Medline: 37395888]
58. Shrestha P, Poudyal B, Yadollahi S, E Wright D, V Gregory A, D Warner J, et al. A systematic review on the use of artificial intelligence in gynecologic imaging - background, state of the art, and future directions. Gynecol Oncol 2022 Sep;166(3):596-605 [FREE Full text] [doi: 10.1016/j.ygyno.2022.07.024] [Medline: 35914978]
59. Sheehy J, Rutledge H, Acharya UR, Loh HW, Gururajan R, Tao X, et al. Gynecological cancer prognosis using machine learning techniques: a systematic review of the last three decades (1990-2022). Artif Intell Med 2023 May;139:102536. [doi: 10.1016/j.artmed.2023.102536] [Medline: 37100507]
60. Xu HL, Gong TT, Liu FH, Chen HY, Xiao Q, Hou Y, et al. Artificial intelligence performance in image-based ovarian cancer identification: a systematic review and meta-analysis. EClinicalMedicine 2022 Sep 17;53:101662 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101662] [Medline: 36147628]
61. Koch AH, Jeelof LS, Muntinga CL, Gootzen TA, van de Kruis NM, Nederend J, et al. Analysis of computer-aided diagnostics in the preoperative diagnosis of ovarian cancer: a systematic review. Insights Imaging 2023 Feb 15;14(1):34 [FREE Full text] [doi: 10.1186/s13244-022-01345-x] [Medline: 36790570]
62. Kuhlmann JD, Chebouti I, Kimmig R, Buderath P, Reuter M, Puppel SH, et al. Extracellular vesicle-associated miRNAs in ovarian cancer - design of an integrated NGS-based workflow for the identification of blood-based biomarkers for platinum-resistance. Clin Chem Lab Med 2019 Jun 26;57(7):1053-1062. [doi: 10.1515/cclm-2018-1048] [Medline: 30422797]

Abbreviations

C-index: concordance index
EOC: epithelial ovarian cancer
LASSO: Least Absolute Shrinkage and Selection Operator
LR: logistic regression
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment
SVM: support vector machine
XGBoost: Extreme Gradient Boosting

Edited by G Tsafnat; submitted 26.04.23; peer-reviewed by J Zhu, Y Li; comments to author 10.11.23; revised version received 23.11.23; accepted 24.11.23; published 22.01.24.

Please cite as:
Wang Q, Chang Z, Liu X, Wang Y, Feng C, Ping Y, Feng X
Predictive Value of Machine Learning for Platinum Chemotherapy Responses in Ovarian Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e48527
URL: https://www.jmir.org/2024/1/e48527
doi:10.2196/48527
PMID:38252469

XSL•FO
RenderX

©Qingyi Wang, Zhuo Chang, Xiaofang Liu, Yunrui Wang, Chuwen Feng, Yunlu Ping, Xiaoling Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Understanding the Integrated Health Management System Policy in China From Multiple Perspectives: Systematic Review and Content Analysis

Yang Yu[1], MS; Sufen Wang[1], PhD; Lijue You[2], MS

[1]Glorious Sun School of Business and Management, Donghua University, Shanghai, China

[2]Department of Informatics, Huadong Hospital Affiliated to Fudan University, Shanghai, China

Corresponding Author:
Sufen Wang, PhD
Glorious Sun School of Business and Management
Donghua University
1882 West Yan-an Road
Shanghai, 200051
China
Phone: 86 18116350361
Email: sf_wang@dhu.edu.cn

Abstract

Background: The integrated health management system (IHMS), which unites all health care–related institutions under a health-centered organizational framework, is of great significance to China in promoting the hierarchical treatment system and improving the new health care reform. China's IHMS policy consists of multiple policies at different levels and at different times; however, there is a lack of comprehensive interpretation and analysis of these policies, which is not conducive to the further development of the IHMS in China.

Objective: This study aims to comprehensively analyze and understand the characteristics, development, and evolution of China's IHMS policy to inform the design and improvement of the system.

Methods: We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to collect 152 policy documents. With the perspective of policy tools and policy orientation as the core, a comprehensive 6D framework including policy level, policy nature, release time, policy tools, stakeholders, and policy orientation was established by combining the content of policy texts. These dimensions were then analyzed using content analysis.

Results: First, we found that, regarding the coordination of policy tools and stakeholders, China's IHMS policy was more inclined to use environment-based policy tools (1089/1929, 56.45%), which suggests a need for further balance in the internal structure of policy tools. Attention to different actors varied, and the participation of physicians and residents needs further improvement (65/2019, 3.22% and 11/2019, 0.54%, respectively). Second, in terms of level differences, Shanghai's IHMS policy used fewer demand-based policy tools (43/483, 8.9%), whereas the national IHMS policy and those of other provinces and cities used fewer supply-based tools (61/357, 17.1% and 248/357, 69.5%, respectively). The national IHMS strategy placed more emphasis on the construction of *smart health care* (including digital health; 10/275, 3.6%), whereas Shanghai was a leader in the development of *healthy community* and *healthy China* (9/158, 5.7% and 4/158, 2.5%, respectively). Third, in terms of time evolution, the various policy tools showed an increasing and then decreasing trend from 2014 to 2021, with relatively more use of environment-based policy tools and less use of demand-based policy tools in the last 3 years. The growth of China's IHMS policy can be divided into 3 stages: the *disease-centered period* (2014-2017), the *e-health technology development period* (2017-2019), and the *health-centered period* (2018-2021).

Conclusions: Policy makers should make several adjustments, such as coordinating policy tools and the uneven relationships among stakeholders; grasping key policy priorities in the context of local characteristics; and focusing on horizontal, multidimensional integration of health resources starting from the community. This study expands the objects of policy research and improves the framework for policy analysis. The findings provide some possible lessons for future policy formulation and optimization.

(J Med Internet Res 2024;26:e47197) doi:10.2196/47197

XSL•FO
RenderX

KEYWORDS

integrated health management system; medical association; medical consortium; policy tools; content analysis; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Background

The integrated health management system (IHMS) is a form of integrated medical care in China that has gained significant attention in recent years. The World Health Organization defined integrated medical care in its 1996 report as "integrating the management and organization of services, providing residents with health services that are easy to access and willing to accept, so that the expenditure on health services can be controlled, and the final service effect can achieve the expected health benefits" [1]. The IHMS is a health management consortium that places health at the center of its operations. Its core service content focuses on "preventing serious diseases, managing chronic diseases, and promoting health" [2]. Medical and health service institutions are the running subject of IHMS, and it aims to address the problem of unreasonable and uneven distribution of medical resources while meeting the rapidly rising demand for various medical and health services, from disease prevention to postoperative care. Compared with medical associations, the IHMS has a wider range of organizations, emphasizing disease prevention and classified management and expanding the functions and services of medical associations [2]. With the promulgation of *Healthy China 2030* and *Healthy China Action (2019-2030)*, China has started to transition from a *treatment-centered* approach to a *health-centered* one [3].

Existing research on IHMS is mainly carried out from 2 perspectives. The first perspective explores IHMS by analyzing all aspects of integrated medical operation. Early literature limited it to vulnerable groups such as older adults or patients with chronic diseases, cancers, and other diseases [4]. However, with the increasing attention paid to the health rights of individuals throughout their life cycle, the clients of integrated medical care have expanded to encompass the entire population [4]. Integration types can be divided into functional integration [5], professional integration [6], normative integration [7], and system integration [8]. Functional integration refers to the coordination between different health care activities and support functions. Professional integration, in contrast, aims to create synergy and integration between health care providers whether they work within the same organization or in different ones. Normative integration aims to achieve an agreement on the normative aspects among all stakeholders involved in health care integration. It also aims to ensure timely reflection on all relevant changes brought about by integration and collaboration in the provision of health services while reinforcing interdependence and collective responsibility. Systemic integration means looking at the complexity of health care organizations in a holistic context and considering the interrelationships between the stakeholders involved in the integration process and the overall context. The direction of integration is divided into horizontal and vertical integration [9]. Horizontal integration refers to the linking of similar services and organizations at the same level to optimize the supply structure, share technical equipment and services, or reduce costs through collective purchasing, for example, horizontal integration between general practitioners and organizations such as community pharmacies and community hospitals in the United Kingdom. Vertical integration refers to the combination of organizations at different levels under a unified management system to form a scale of health care services to diversify risks and reduce the cost of purchasing health care products while providing seamless services to patients, such as the close-knit county medical communities being implemented across China. In terms of the evaluation of integration effects, scholars have put forward various models. Berwick et al [10] put forward a 3-objective model according to the integration practice in the United States. Sikka et al [11] then improved it and developed a new 4-quadrant model. In the same year, Nurjono et al [12] synthesized dozens of key characteristic indexes, further revised them using the Delphi method, and put forward the Rainbow Model. In 2019, Chaitkin et al [13] proposed a 5-step decision-making theoretical framework. With the development of big data, the internet of things, and other technologies, recent research on integrated medical care is keen to explore the integration of e-medical, telemedicine, and integrated medical care. For example, Tossaint-Schoenmakers et al [14] used the structure-process-outcome framework of the Donabedian model to determine the relevant indicators for integrating eHealth into medical care organizations. Fatoum et al [15] explored the future impact of blockchain technology and its integration with other innovations in the health care ecosystem. de Luca et al [16] designed a comprehensive digital hypertension method to enable citizens to learn how to prevent the development of hypertension and lead a healthy lifestyle. These studies cover the concept, key elements, evaluation methods, and integration of cutting-edge technologies of integrated medical care, but the development of integrated medical care is often related to the development of relevant policy systems.

The second perspective pertains to the policy system related to the IHMS. In China, most of the existing research still centers on medical associations, with few studies on IHMS. There are primarily 3 types of research related to medical association policy. The first is to analyze the current situation by sorting out medical association policy. For instance, Ling et al [17] summarized the development and practice of medical association policy in China, identified issues such as weak 2-way referrals, and provided relevant suggestions. Although they carried out an in-depth qualitative analysis of the policy, they did not statistically analyze its textual content, meaning that, although they made a number of recommendations, these needed to be supported by more comprehensive quantitative data that would have helped provide deeper insights into their recommendations. The second type of research evaluates the implementation effect of medical association policy. For example, Yang et al [18] used a binary logistic regression model to analyze the implementation effect and influencing factors of medical

XSL•FO
RenderX

association policy in Guizhou province and proposed measures such as increasing social health expenditure. Although this research is based on models, the data primarily come from field interviews or questionnaires, neglecting information in policy texts. The third type of research involves multidimensional analysis of policies regarding policy tools. For instance, Chen et al [19] conducted a statistical analysis of medical association policy issued by the central government from the perspective of policy tools and integrated medical care, concluding that the structure requires further adjustment. This type of research tends to be clearer and more operational as the dimensions and categories used are clearly identifiable, allowing the researcher to make judgments and categorize the research based on the definitions of each category, reducing the interference of subjective judgments, and improving the objectivity and reproducibility of the research. However, the number of policy texts they collected and the dimensions of the framework they constructed are relatively limited. In addition, the few studies on the IHMS policy are still in the concept innovation, concept definition, and case analysis phases, such as the studies by Bai [20] and Xiong et al [3].

Objectives

The aim of this study was to comprehensively analyze China's IHMS policy and provide accurate and comprehensive suggestions for policy optimization through quantitative data analysis, such as improving policy level and implementation, strengthening smart medical care and healthy community support, optimizing policy tools, strengthening stakeholder cooperation, and improving policy orientation so as to promote the effective implementation and development of the IHMS in China. On the basis of the contents and objectives of the relevant policy and planning documents issued by the Chinese government, we proposed the concept of China's IHMS policy system, which refers to a set of health-centered policies that integrate various types of health care resources. The policy system covers a variety of policies (Multimedia Appendix 1) on healthy China; graded diagnosis and treatment; medical association; healthy community; and smart health care, which includes digital health.

The collection of these policies expands the number, scope, and subjects of China's IHMS policy, constituting a policy chain–policy cluster–policy network health consortium policy system with common goals, multiple subjects, and numerous means, spanning levels and complex interactions. It covers various aspects of China's IHMS policy, including health objectives, health service patterns, the integration of health resources, the application of IT, and community health services. Multiple policy objectives are unified under a holistic framework, and different policy orientations are organically combined to support and synergize each other to promote the development of health care.

On the basis of the researchers' in-depth investigation and analysis of China's IHMS policy, a 6D framework (Multimedia Appendix 2) was constructed, including policy level, policy nature, release time, policy tools, stakeholders, and policy orientation. Although it fully considers the characteristics of China's IHMS policy, the framework exploits and effectively integrates the meta-information and key elements in the policy text and provides a structured analytical method that allows researchers to systematically conduct quantitative analysis and policy comparison. The framework focuses on policy tools and policy orientations, exploring their characteristics and trends under the dimensions of policy level, policy nature, and release time. Through this comprehensive analytical framework, we expect to provide policy makers and relevant researchers with more operational and practical methods for policy analysis.

This study focused on two main issues:

1. What are the release time, release unit, nature, and specifications of China's IHMS policy? What types of policies, policy tools, and stakeholders are included?
2. With a focus on the 2 core dimensions of policy tools and policy orientation, what are the differences in the distribution of China's IHMS policy in terms of policy level, policy nature, and policy release time?

By answering the aforementioned questions, this study aimed to explore whether the structure of policy tools and policy orientation of China's IHMS is balanced, whether it changes with time, and what the similarities and differences are between local and central governments. The results of this study will aid in developing suggestions for policy optimization.

Methods

Overview

This study used a systematic literature review method [21] and a content analysis approach. First, policy documents related to the IHMS were collected according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) criteria and then sorted by release time and formulation subject, among other criteria. Finally, they were organized, coded, and analyzed using the 6D framework based on policy tool theory.

Research Materials

Overview

The research materials were obtained from various sources, and a systematic review was conducted with the following objectives in mind: (1) being relevant to the background of new medical reform or integrated medical care; (2) involving one or more aspects of health management or disease prevention, treatment, and care; and (3) paying attention to, advocating for, or promoting the development of new technologies in the medical field.

By thoroughly considering these objectives, we compiled a collection of highly relevant and diverse IHMS policy literature. The PRISMA checklist is shown in Multimedia Appendix 3.

Information Sources

We searched for relevant articles in the following databases: Chinese People's Government Network, China National Health Commission Network, China Provincial People's Government Network, China Provincial Health Commission Network, and Bailu think tank. These databases have high authority, comprehensive policy literature, clear classification, and convenient access, making them suitable for this research.

XSL•FO
RenderX

Selection Criteria

To select the literature for analysis, a number of inclusion and exclusion criteria were used, as shown in Textbox 1.

Textbox 1. Inclusion and exclusion criteria for the collection of research materials.

Inclusion criteria

- Article type: public documents, bulletins, and notices
- Article language: Chinese articles
- Article publication time: articles published before the end of 2021
- Article main content: 5 categories of policies as identified by the definition of integrated health management system (IHMS) policies in China

Exclusion criteria

- Article type: other types of articles, such as media reports and policy interpretations
- Article language: articles in other languages
- Article publication time: articles that had not been published by the study launch time
- Article main content: articles not directly related to the definition of China's IHMS policy

Screening of Research Materials

The process of searching for and selecting policy documents was divided into four stages, as outlined in Figure 1:

1. Keyword determination: articles were searched for in electronic databases using the keywords *healthy China*, *graded diagnosis and treatment*, *medical association*, *healthy community*, and *smart health care*.
2. Preliminary review: identified articles were sorted by relevance, and titles and introduction contents were reviewed to select important documents according to the selection criteria.
3. Further review: articles that passed the preliminary review were read.
4. Material determination: articles with a high number of citations were browsed to obtain more research materials.

As a result, a total of 152 policy documents, including plans, outlines, and notices, were obtained, covering a time span from July 2014 to August 2021. Multimedia Appendix 4 provides a complete list.

Figure 1. Search process according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) criteria.

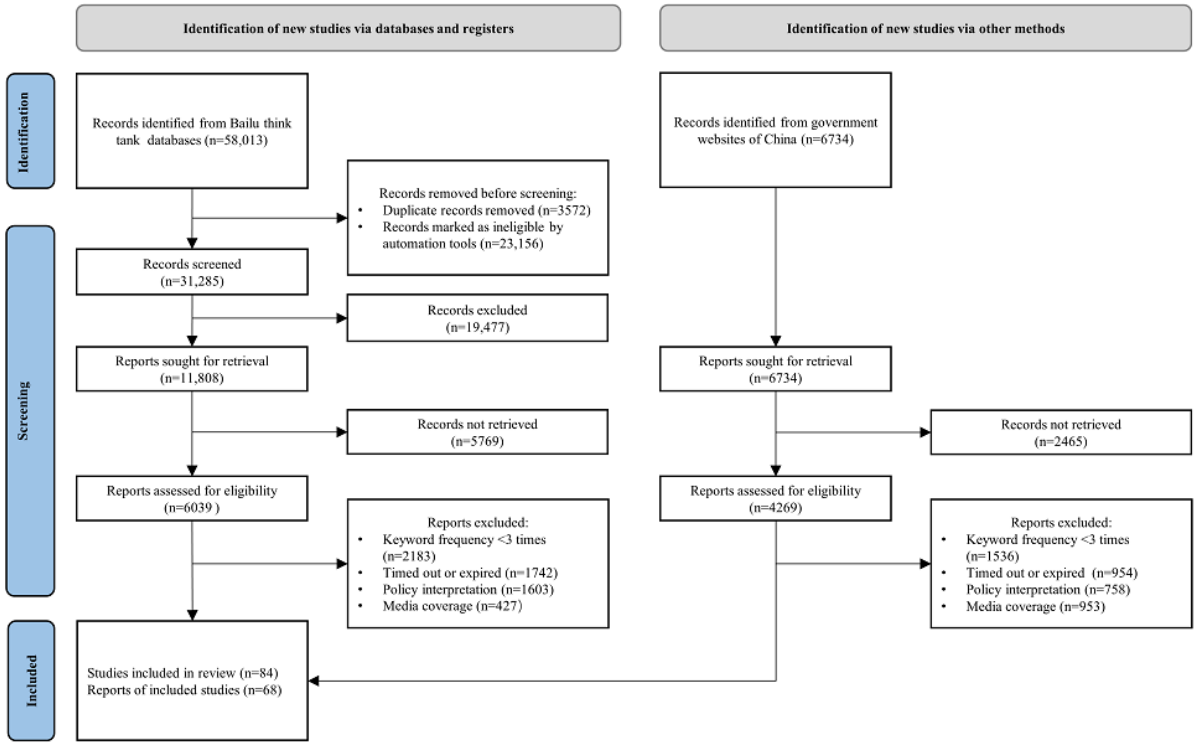

Research Framework

To ensure that the theoretical framework was rigorous and practical, we conducted in-depth analyses of practical application scenarios. From the perspective of practical application, the current policy analysis in the industry focuses on the following three key aspects: (1) policy specifications and nature, with particular emphasis on policies issued by the central

XSL•FO
RenderX

government; (2) policy synergies, that is, whether complementary policies and measures are sufficient; and (3) whether the policies can achieve the expected results, which is also the final point of the policy analysis. In addition, unlike general single-category policies, the IHMS policy is a policy chain–policy cluster–policy network system formed by multicategory policies. The policy orientation, that is, the main elements of the policy objectives, will play a very important role in the implementation effect of the entire system.

However, the 2D or 3D analytical frameworks mainly used in existing studies have limitations in these important aspects. For example, Tang and Si [22] did not make full use of the metainformation of policy documents, such as the release time, issuing agency, and policy type, although they conducted a joint analysis of policy tools and stakeholders, and Shen et al [23], although they conducted an external characterization analysis that included policy year, issuing agency, and policy type and also conducted a joint analysis of policy tools and health systems, did not address the differences in the external characterization dimensions of policy tools. The insufficient exploration of policy synergies limits the scope and diversity of policy analyses.

Therefore, in addition to policy tools and stakeholders, this study added the 4 dimensions of policy level, policy nature, release time, and policy orientation. These dimensions are interrelated and form an overall analytical framework structure, as shown in Figure 2. Policy level, policy nature, release time, and policy orientation are vertical dimensions, indicating the different characteristics of policies. Policy tools and stakeholders are horizontal dimensions, indicating the tools and relevant stakeholders in policy implementation. Meanwhile, policy orientation, as a specific dimension, runs through the entire analytical framework, reflecting the importance of policy objectives.

Figure 2. 6D analysis framework for China's integrated health management system policy.

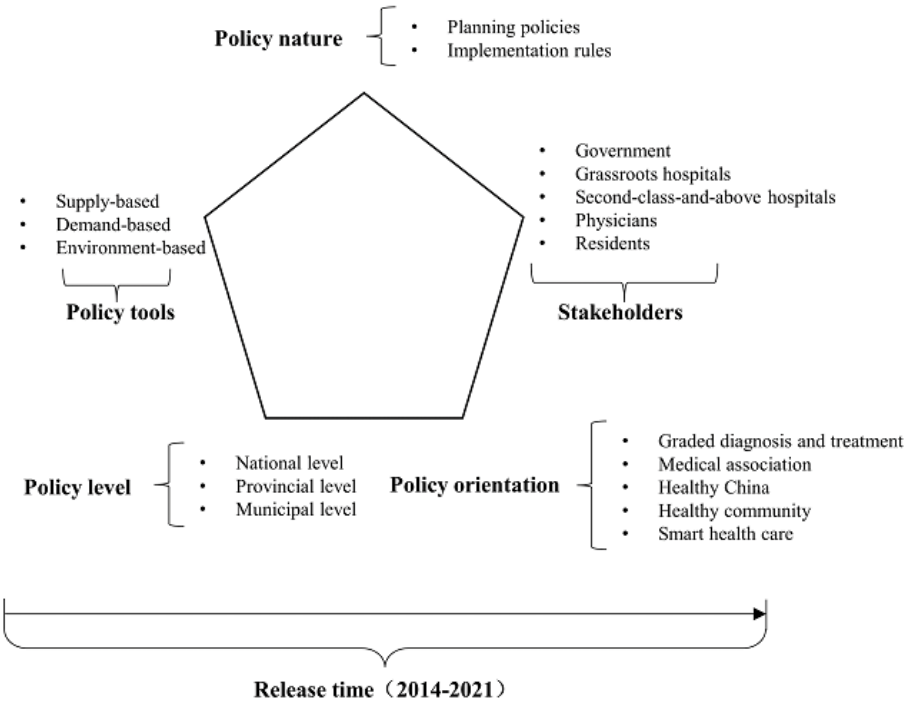

This study adopted the policy tool theory by Rothwell and Zegveld [24] to define 3 types of IHMS policy tools based on the characteristics of the IHMS (Table S1 in Multimedia Appendix 5). Supply-based tools mainly refer to policies that provide direct support for the IHMS, such as funding, talent, technology, and other resources [25]. Demand-based tools mainly refer to policies that promote the IHMS by creating a "pull" effect, such as encouraging the public to use health care services through improved service quality and accessibility. Environment-based tools indirectly promote the IHMS by creating an enabling environment through the establishment of goals, plans, incentives, and other policy measures. By analyzing the use of these policy tools in IHMS policies, we can gain insights into the policy-making strategies and priorities of relevant government departments.

Furthermore, the implementation of the IHMS involves various stakeholders, each with different roles to play. To account for this, we added a stakeholder dimension to create a 2D analysis framework for IHMS policy texts. The stakeholders of the IHMS include the government, grassroots hospitals, second-class-and-above hospitals, physicians, and residents. The government leads and formulates IHMS policies, whereas grassroots hospitals and second-class-and-above hospitals are key players in implementing the IHMS and providing health services to residents through their collaboration. Physicians at all levels serve as the backbone of the IHMS, whereas residents are the direct beneficiaries of the policy [22]. Table S2 in Multimedia Appendix 5 displays the categories and definitions of the stakeholders.

XSL•FO
RenderX

Coding and Statistics

Using the paragraph symbol as a separator, the 152 policy documents obtained from the screening were segmented into a collection of text paragraphs, retaining information such as the title, author, and date of each document but excluding additional tables. We ended up with 1345 policy texts as research material.

We coded the release time, policy level, policy nature, and policy orientation based on the year in which the policy was issued, the administrative level to which the policy-issuing organization belongs, the title and its lineage, and the 5 categories of policies that make up the IHMS. For the policy tools and stakeholders, coding was based on the categorization of policy tools in Table S1 in Multimedia Appendix 5 and the categorization of stakeholders in Table S2 in Multimedia Appendix 5. For example, for the following policy text—"promoting the construction of public health system in medical complexes. Hospitals..."—the policy tools can be coded as *functional localization*, *technical support*, and *talent construction*, whereas stakeholders can be coded as *government*, *grassroots hospitals*, and *second-class and above hospitals*.

Multimedia Appendix 6 shows the specific steps for coding and manual labeling.

We coded the policy texts using Microsoft Excel (Microsoft Corp) and calculated an average κ coefficient value of 74.6%, indicating a high degree of consistency (substantial) for the 2 coders under each dimension of categorization using SPSS (IBM Corp). Following these steps, we ensured that the credibility of the manual annotation results was supported by independent annotation, comparative positioning, and discussion of the changes. This systematic approach to annotation helped reduce subjective interference and improve the accuracy and consistency of annotation, thereby increasing the reliability of the findings and conclusions.

Results

1D Analysis Under the 6D Framework

Policy Release Time

Figure 3 shows the distribution of the IHMS policy release time. The number of IHMS policy texts showed an overall upward trend from 2014 to 2021. This number increased rapidly before 2017, reaching a peak of 531 that year. However, in recent years (not excluding the influence of the statistical period until August 2021), this number declined.

Figure 3. Time-varying chart of marked integrated health management system policy texts.

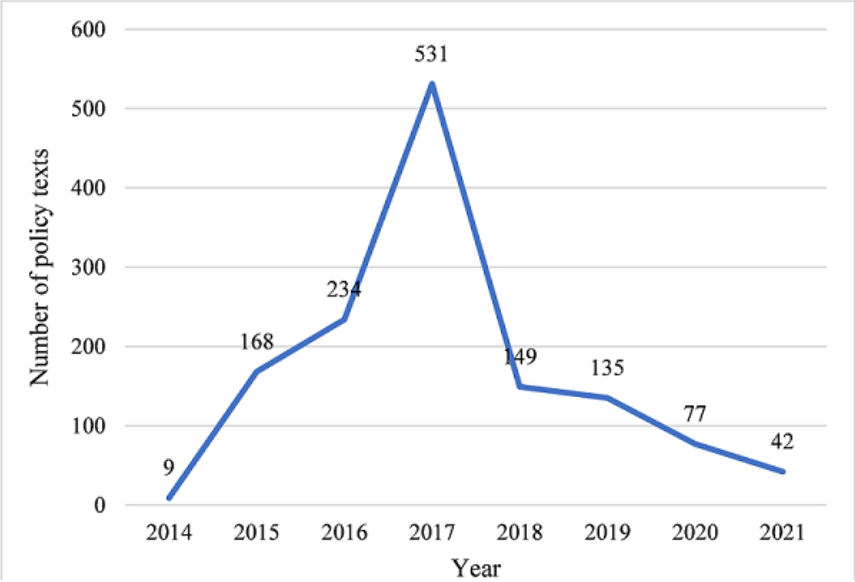

Policy Level

The distribution of the IHMS policy levels is presented in Table 1. The data revealed that there were 659 policy texts at the provincial level followed by 275 at the national level. In addition, there were >150 policy texts at other municipal levels as well as in Shanghai.

Table 1. Distribution of policy level of marked integrated health management system policy texts.

Policy level	Policies (n=1345), n (%)
Provincial level	659 (49)
National level	275 (20.45)
Municipal level	253 (18.81)
Shanghai	158 (11.75)

XSL•FO
RenderX

Nature of Policy

Table 2 shows the distribution of the policy nature of the IHMS. The number of planning and implementation rule policy texts was equivalent, at 674 and 671, respectively.

Table 2. Distribution of policy nature of marked integrated health management system policy texts.

Policy nature	Policies (n=1345), n (%)
Planning policy	674 (50.11)
Detailed regulations	671 (49.89)

Policy Tools

The distribution of the IHMS policy tools is presented in Table 3. Environment-based policy tools were the most frequently used, accounting for 56.45% (1089/1929) of all tools, followed by demand-based policy tools at 25.04% (483/1929), whereas supply-based policy tools were the least commonly used, accounting for only 18.51% (357/1929). This reflects the current situation, in which demand-based policy tools are relatively abundant, whereas supply-based policy tools are relatively scarce in China's IHMS policy system.

Table 3. Distribution of policy tools of the integrated health management system (n=1929).

Policy tool and secondary classification	Policy tools, n (%)
Supply-based (n=357)	
Talent construction	75 (3.89)
Financial guarantee	37 (1.92)
Information support	127 (6.58)
Technical support	18 (0.93)
Rational layout	100 (5.18)
Demand-based (n=483)	
Medical insurance payment	73 (3.78)
Medicine supply	40 (2.07)
Price adjustment	38 (1.97)
Ability building	59 (3.06)
Adhere to public welfare	24 (1.24)
Orderly medical treatment	197 (10.21)
Demonstration project	52 (2.7)
Environment-based (n=1089)	
Objective programming	223 (11.56)
Institutional building	253 (13.12)
Incentive and restraint	187 (9.69)
Policy promotion	49 (2.54)
Functional localization	241 (12.49)
Resource sharing	136 (7.05)

Examining the specific types of policy tools, it was observed that, among environment-based policy tools, *institutional building*, *functional localization*, and *objective programming* were the most frequently used, totaling 37.17% (717/1929), whereas *policy promotion* was used the least, accounting for only 2.54% (49/1929). Among the demand-based policy tools, only *orderly medical treatment* was widely used, and the frequency of use of *adhere to public welfare*, *medicine supply*, and *price adjustment* was <2.5% (24/1929, 1.24%; 40/1929, 2.07%; 38/1929, 1.97%, respectively). The proportion of supply-based policy tools was also relatively small, with the most commonly used tool, *information support*, accounting for only 6.58% (127/1929), whereas *technical support* and *financial guarantee* were the least frequently used, accounting for 2.85% (55/1929) in total.

Stakeholders

Table 4 presents the distribution of stakeholders in IHMS policy. The highest frequency was of governments at 57.16% (1154/2019), followed by *second-class and above hospitals* and *grassroots hospitals* accounting for 39.08% (789/2019) in total. *Residents* had the lowest participation frequency at only 0.54% (11/2019). This indicates that the government plays a leading role in the entire IHMS system, and general and grassroots

XSL•FO
RenderX

hospitals at or above the second level play a significant role as the main service providers. However, the policy pays relatively little attention to family physicians, specialists, and residents.

Table 4. Distribution of stakeholders in the integrated health management system (n=2019).

Stakeholders	Participation, n (%)
Governments	1154 (57.16)
Grassroots hospitals	461 (22.83)
Second-class-and-above hospitals	328 (16.25)
Physicians	65 (3.22)
Residents	11 (0.54)

Policy Orientation

Table 5 displays the distribution of IHMS policy orientation. Medical associations accounted for the highest percentage with 778 policy texts, representing 57.84% (778/1345) of the total. Graded diagnosis and treatment orientation was the second most common with 486 copies, accounting for 36.13% (486/1345). In contrast, smart health care, healthy community, and healthy China had <50 policy texts each. Compared with medical associations and graded diagnosis and treatment, the existing IHMS policy framework covers relatively little content on smart health care, healthy community, and healthy China. However, it is noteworthy that healthy China is a strategic policy orientation, and the related contents are mainly top-level design plans. The specific measures were refined into medical associations, graded diagnosis and treatment, smart health care, healthy community, and other aspects. Therefore, the number of policy texts in this category was the lowest.

Table 5. Distribution of policy orientation of marked integrated health management system policy texts.

Policy orientation	Policies (n=1345), n (%)
Medical association	778 (57.84)
Graded diagnosis and treatment	486 (36.13)
Smart health care	36 (2.68)
Healthy community	24 (1.78)
Healthy China	21 (1.56)

Stakeholder Analysis of Policy Tools

On the basis of the policy tool dimension, a stakeholder dimension was introduced for joint analysis in Table 6. The results show that the 3 types of policy tools involving the government had a relatively even distribution, with environment-based policy tools having the highest frequency of use at 64% (992/1550) and demand-based policy tools having the lowest frequency of use at 46.6% (385/827). For second-class-and-above hospitals, the proportion of the 3 types of policy tools was not significantly different—supply-based policy tools were the most frequently used, accounting for 29.4% (182/620), whereas environment-based policy tools were the least frequently used, accounting for 20.19% (313/1550). For grassroots hospitals, demand-based policy tools had the highest proportion at 20.8% (172/827), and environment-based policy tools had the lowest proportion at 13.94% (216/1550). Examining the internal composition, it was observed that, for the government, the *technical support*,*financial guarantee*, *adhere to public welfare*,*medicine supply*, *price adjustment*, and *policy promotion* tools were relatively lacking compared with other tools of the same type. For second-class-and-above hospitals and grassroots hospitals, although there were few policy items, including *financial guarantee* and *adhere to public welfare*, these 2 items should be led by the government, so a small proportion of items is understandable. However, *technical support*, *price adjustment*, and *policy promotion* were relatively insufficiently used. In addition, the classification of physicians and residents did not contain any *policy promotion* items, highlighting an existing issue.

XSL•FO
RenderX

Table 6. 2D distribution table of policy tools and stakeholders of the integrated health management system (n=2997).

Policy tool and secondary classification	Governments, n (%)	Second-class-and-above hospitals, n (%)	Grassroots hospitals, n (%)	Physicians, n (%)	Residents, n (%)
Supply-based					
Talent construction (n=143)	53 (37.06)	46 (32.17)	36 (25.17)	8 (5.59)	N/A[a]
Financial guarantee (n=42)	37 (88.1)	2 (4.76)	2 (4.76)	1 (2.38)	N/A
Information support (n=225)	121 (53.78)	76 (33.78)	27 (12)	1 (0.44)	N/A
Technical support (n=30)	15 (50)	10 (33.33)	5 (16.77)	N/A	N/A
Rational layout (n=180)	90 (50)	48 (26.67)	40 (22.22)	2 (1.11)	N/A
Total (n=620)	316 (50.97)	182 (29.35)	110 (17.74)	12 (1.94)	0 (0)
Demand-based					
Medical insurance payment (n=94)	73 (77.66)	11 (11.7)	8 (8.51)	2 (2.13)	N/A
Medicine supply (n=63)	35 (55.56)	15 (23.81)	12 (19.05)	1 (1.59)	N/A
Price adjustment (n=52)	38 (73.08)	8 (15.38)	6 (11.54)	N/A	N/A
Ability building (n=117)	43 (36.75)	40 (34.19)	33 (28.21)	1 (0.85)	N/A
Adhere to public welfare (25)	23 (92)	1 (4)	1 (4)	N/A	N/A
Orderly medical treatment (n=396)	121 (30.56)	118 (29.8)	99 (25)	48 (12.12)	10 (2.53)
Demonstration project (n=80)	52 (65)	15 (18.75)	13 (16.25)	N/A	N/A
Total (n=827)	385 (46.55)	208 (25.15)	172 (20.8)	52 (6.29)	10 (1.21)
Environment-based					
Objective programming (n=273)	219 (80.22)	30 (10.99)	20 (7.33)	3 (1.1)	1 (0.37)
Institutional building (n=371)	239 (64.42)	64 (17.25)	51 (13.75)	14 (3.77)	3 (0.81)
Incentive and restraint (n=241)	169 (70.12)	48 (19.92)	20 (8.3)	3 (1.24)	1 (0.41)
Policy promotion (n=56)	48 (85.71)	3 (5.36)	5 (8.93)	N/A	N/A
Functional localization (n=351)	189 (53.85)	94 (26.78)	68 (19.37)	N/A	N/A
Resource sharing (n=258)	128 (49.61)	74 (28.68)	52 (20.16)	4 (1.55)	N/A
Total (n=1550)	992 (64)	313 (20.19)	216 (13.94)	24 (1.55)	5 (0.32)
Grand total (n=2997)	1693 (56.49)	703 (23.46)	498 (16.62)	88 (2.94)	15 (0.5)

[a]N/A: not applicable.

Analysis of Level Differences

Analysis of the Level Differences of Policy Tools

On the basis of the policy tool dimension, the policy level dimension was introduced to analyze differences. Table 7 shows that, when considering all provinces and cities except Shanghai as a whole, the number and proportion of policies was greater than that of the central government and Shanghai, and the distribution proportion was even regarding supply-, demand-, and environment-based tools. Under the second classification of supply-based policy tools, compared with the national IHMS policy, other provinces and cities had a higher proportion of *rational layout*, *talent construction*, and *financial guarantee*, whereas Shanghai had a higher proportion of *technical support* and a lower proportion of *talent construction*. Under the second classification of demand-based policy tools, the proportion of national IHMS policies in *orderly medical treatment* and *demonstration project* was higher, whereas the proportion of *price adjustment* and *adherence to public welfare* was lower. Other provinces and cities had a higher proportion of *orderly medical treatment* and *medical insurance payment*, whereas Shanghai had a higher proportion of *ability building* and the lowest proportion of *medicine supply*. Under the second classification of environment-based policy tools, compared with the national IHMS policy, other provinces and cities had a higher proportion of *institutional building* and *objective programming*, whereas the proportion of *incentive and restraint* was slightly lower. In contrast, the proportion of *objective programming* and *institutional building* was slightly higher in Shanghai.

XSL•FO
RenderX

Table 7. 2D distribution table of policy tools and policy levels of the integrated health management system (n=1929).

Policy tool and secondary classification	National level, n (%)	Provincial and municipal level, n (%)	Shanghai, n (%)
Supply-based			
Talent construction (n=75)	10 (13.33)	59 (78.67)	6 (8)
Financial guarantee (n=37)	4 (10.81)	28 (75.68)	5 (13.51)
Information support (n=127)	33 (25.98)	77 (60.63)	17 (13.39)
Technical support (n=18)	2 (11.11)	8 (44.44)	8 (44.44)
Rational layout (n=100)	12 (12)	76 (76)	12 (12)
Total (n=357)	61 (17.09)	248 (69.47)	48 (13.45)
Demand-based			
Medical insurance payment (n=73)	6 (8.22)	61 (84.56)	6 (8.22)
Medicine supply (n=40)	5 (12.5)	34 (85)	1 (2.5)
Price adjustment (n=38)	2 (5.26)	33 (86.84)	3 (7.89)
Ability building (n=59)	8 (13.56)	39 (66.1)	12 (20.34)
Adhere to public welfare (n=24)	4 (16.67)	17 (70.83)	3 (12.5)
Orderly medical treatment (n=197)	53 (26.9)	131 (66.5)	13 (6.6)
Demonstration project (n=52)	16 (30.77)	31 (59.62)	5 (9.62)
Total (n=483)	94 (19.46)	346 (71.64)	43 (8.9)
Environment-based			
Objective programming (n=223)	21 (9.42)	173 (77.58)	29 (13)
Institutional building (n=253)	38 (15.02)	193 (76.28)	22 (8.7)
Incentive and restraint (n=187)	41 (21.93)	120 (64.17)	26 (13.9)
Policy promotion (n=49)	6 (12.24)	36 (73.47)	7 (14.29)
Functional localization (n=241)	64 (26.56)	150 (62.24)	27 (11.2)
Resource sharing (n=136)	26 (19.12)	92 (67.65)	18 (13.24)
Total (n=1089)	196 (18)	764 (70.16)	129 (11.85)
Grand total (n=1929)	351 (18.2)	1358 (70.4)	220 (11.4)

Analysis of the Level Differences of Policy Orientation

Table 8 displays the distribution of IHMS policy orientations and policy levels. Across all policy levels, IHMS policies contained more content on graded diagnosis and treatment and medical association but less content on the smart health care, healthy community, and healthy China orientations. However, policies issued by Shanghai had a higher proportion of relevant content on healthy community (9/158, 5.7%) and healthy China (4/158, 2.5%) compared with national IHMS policies on these 2 orientations. The proportion of relevant content on smart health care in the national IHMS policies was higher (10/275, 3.6%) than that in policies issued by Shanghai and other provinces and cities. This indicates that Shanghai, as a pioneer city in the implementation of IHMS policies, has played a leading role in the construction of healthy communities and healthy China, but more efforts are needed to strengthen the construction of smart health care.

Table 8. 2D distribution table of policy orientations and policy levels of the integrated health management system.

Policy orientation	National level (n=275), n (%)	Shanghai (n=158), n (%)	Provincial and municipal level (n=912), n (%)
Graded diagnosis and treatment	120 (43.6)	49 (31)	317 (34.8)
Healthy community	9 (3.3)	9 (5.7)	6 (0.7)
Healthy China	6 (2.2)	4 (2.5)	11 (1.2)
Medical association	130 (47.3)	92 (58.2)	556 (61)
Smart health care	10 (3.6)	4 (2.5)	22 (2.4)

XSL•FO
RenderX

Analysis of Nature Differences

Analysis of the Nature Differences of Policy Tools

The dimension of policy nature was introduced to analyze the differences in addition to the dimension of policy tools. As shown in Table 9, demand-based policy tools accounted for 55.9% (270/483) of the planned IHMS policy items, and environment-based policy tools accounted for 53.63% (584/1089) of the implementation details. Under the second classification of supply-based policy tools, the implementation rules had a higher proportion of *technical support* and *rational layout* but a lower proportion of *talent construction* and *financial guarantee* compared with the planned IHMS policy. Under the second classification of demand-based policy tools, the implementation rules had a higher proportion of *demonstration project*, whereas the proportion of *orderly medical treatment* in the planned IHMS policy items was lower than that in the implementation rules. Under the secondary classification of environment-based policy tools, the implementation rules had a higher proportion of *incentive and restraint*, *functional localization*, and *objective programming* but a lower proportion of *policy promotion* and *resource sharing* compared with the planned IHMS policy.

Table 9. 2D distribution table of policy tools and policy nature of the integrated health management system (n=1929).

Policy tool and secondary classification	Planning policy, n (%)	Detailed regulations, n (%)
Supply-based		
Talent construction (n=75)	47 (63)	28 (37)
Financial guarantee (n=37)	21 (57)	16 (43)
Information support (n=127)	58 (45.7)	69 (54.3)
Technical support (n=18)	9 (50)	9 (50)
Rational layout (n=100)	46 (46)	54 (54)
Total (n=357)	181 (50.7)	176 (49.3)
Demand-based		
Medical insurance payment (n=73)	45 (62)	28 (38)
Medicine supply (n=40)	20 (50)	20 (50)
Price adjustment (n=38)	24 (63)	14 (37)
Ability building (n=59)	40 (68)	19 (32)
Adhere to public welfare (n=24)	14 (58)	10 (42)
Orderly medical treatment (n=197)	112 (56.9)	85 (43.1)
Demonstration project (n=52)	15 (29)	37 (71)
Total (n=483)	270 (55.9)	213 (44.1)
Environment-based		
Objective programming (n=223)	101 (45.3)	122 (54.7)
Institutional building (n=253)	123 (48.6)	130 (51.4)
Incentive and restraint (n=187)	67 (35.8)	120 (64.2)
Policy promotion (n=49)	29 (59)	20 (41)
Functional localization (n=241)	108 (44.8)	133 (55.2)
Resource sharing (n=136)	77 (56.6)	59 (43.4)
Total (n=1089)	505 (46.37)	584 (53.63)
Grand total (n=1929)	956 (49.56)	973 (50.44)

Analysis of the Nature Differences of Policy Orientation

Table 10 shows the distribution of IHMS policy orientation and nature. The proportion of planned IHMS policies related to healthy China was relatively high (11/674, 1.6%), whereas the content related to implementing IHMS policies on healthy community and smart health care was relatively high (16/671, 2.4% and 21/671, 3.1%, respectively).

XSL•FO
RenderX

Table 10. 2D distribution table of policy orientation and policy nature of the integrated health management system.

Policy orientation	Planning policy (n=674), n (%)	Detailed regulations (n=671), n (%)
Graded diagnosis and treatment	345 (51.2)	141 (21)
Healthy community	8 (1.2)	16 (2.4)
Healthy China	11 (1.6)	10 (1.5)
Medical association	295 (43.8)	483 (72)
Smart health care	15 (2.2)	21 (3.1)

Analysis of Time Evolution

Analysis of the Time Evolution of Policy Tools

On the basis of the dimension of policy tools, the release time of policies was introduced for the analysis of evolution. As shown in Table 11, the distribution of the number of IHMS policy texts in each year exhibited a trend of initially increasing and then decreasing, reaching its peak in 2017 and declining in the last 2 years.

Regarding supply-based policy tools, the number of policies related to *technical support* was the lowest, but it increased in the last 2 years. The change trend in the number of other secondary tools was more consistent with the overall trend, that is, they first increased and then decreased, with the lack of *financial guarantee* being more prominent in the last 2 years. Demand-based policy tools were used less frequently from 2019 to 2021, with only a small number of policies remaining related to *ability building* and *orderly medical treatment*. The use of environment-based policy tools was relatively high, but the number of secondary tools also dropped significantly since 2017. In the last 2 years, except for *functional localization*, a certain number of policies were maintained, and the frequency of use of other tools was low, especially *policy promotion*.

XSL•FO
RenderX

Table 11. 2D distribution table of policy tools and release time of the integrated health management system (n=1929).

Policy tool and secondary classification	2015 and before, n (%)	2016, n (%)	2017, n (%)	2018, n (%)	2019, n (%)	2020, n (%)	2021, n (%)
Supply-based							
Talent construction (n=75)	13 (17)	12 (16)	29 (39)	8 (11)	8 (11)	1 (1)	4 (5)
Financial guarantee (n=37)	2 (5)	6 (16)	20 (54)	1 (3)	6 (16)	2 (5)	N/A[a]
Information support (n=127)	11 (8.7)	19 (15)	39 (30.7)	22 (17.3)	19 (15)	14 (11)	3 (2.4)
Technical support (n=18)	1 (6)	N/A	9 (50)	2 (11)	1 (6)	2 (11)	3 (17)
Rational layout (n=100)	9 (9)	2 (2)	71 (71)	3 (3)	5 (5)	6 (6)	4 (4)
Total (n=357)	36 (10.1)	39 (10.9)	168 (47.1)	36 (10.1)	39 (10.9)	25 (7)	14 (3.9)
Demand-based							
Medical insurance payment (n=73)	9 (12)	17 (23)	35 (48)	6 (8)	6 (8)	N/A	N/A
Medicine supply (n=40)	6 (15)	5 (13)	21 (53)	6 (15)	1 (3)	1 (3)	N/A
Price adjustment (n=38)	6 (16)	15 (39)	15 (39)	1 (3)	1 (3)	N/A	N/A
Ability building (n=59)	9 (15)	13 (22)	17 (29)	5 (8)	10 (17)	2 (3)	3 (5)
Adhere to public welfare (n=24)	1 (4)	2 (8)	17 (71)	3 (13)	N/A	1 (4)	N/A
Orderly medical treatment (n=197)	43 (21.8)	54 (27.4)	71 (36)	9 (4.6)	15 (7.6)	3 (1.5)	2 (1)
Demonstration project (n=52)	6 (12)	10 (19)	21 (40)	2 (4)	13 (25)	N/A	N/A
Total (n=483)	80 (16.6)	116 (24)	197 (40.8)	32 (6.6)	46 (9.5)	7 (1.4)	5 (1)
Environment-based							
Objective programming (n=223)	24 (10.8)	29 (13)	132 (59.2)	21 (9.4)	9 (4)	5 (2.2)	3 (1.3)
Institutional building (n=253)	25 (9.9)	42 (16.6)	127 (50.2)	33 (13)	10 (4)	12 (4.7)	4 (1.6)
Incentive and restraint (n=187)	15 (8)	23 (12.3)	72 (38.5)	56 (29.9)	9 (4.8)	9 (4.8)	3 (1.6)
Policy promotion (n=49)	9 (18)	14 (29)	19 (39)	2 (4)	3 (6)	1 (2)	1 (2)
Functional localization (n=241)	23 (9.5)	39 (16.2)	103 (42.7)	19 (7.9)	28 (11.6)	17 (7.1)	12 (5)
Resource sharing (n=136)	16 (11.8)	16 (11.8)	72 (52.9)	7 (5.1)	11 (8.1)	10 (7.4)	4 (2.9)
Total (n=1089)	112 (10.28)	163 (15.15)	525 (48.21)	138 (12.67)	70 (6.43)	54 (4.96)	27 (2.48)
Grand total (n=1929)	228 (11.82)	318 (16.49)	890 (46.14)	206 (10.68)	155 (8.04)	86 (4.46)	46 (2.38)

[a]N/A: not applicable.

Analysis of the Time Evolution of Policy Orientation

Table 12 presents the distribution of IHMS policy orientation and release time. Between 2014 and 2021, the proportion of graded diagnosis and treatment, healthy China, medical association, and smart health care increased initially and then decreased, with the highest proportions occurring in 2016, 2019, 2017, and 2018, respectively. However, the proportion of healthy community guidance–related content showed a certain growth trend, particularly in the last 2 years. This suggests that the change order of IHMS policy orientation is roughly as follows: graded diagnosis and treatment, medical association, smart health care, healthy China, and healthy community. This indicates that the government is increasingly prioritizing the all-around health needs of the people with *health as the goal*, promoting the construction of the IHMS, and taking the community as an essential starting point.

Figure 4 presents the changes in the 5 types of policies more intuitively. From 2014 to 2016, policies on graded diagnosis and treatment developed rapidly and then gradually slowed down, with a short stagnation in 2020. Policies on medical association started in 2015 and, after a short decline in 2016, increased in 2017, and the development speed gradually slowed down. Policies on healthy community developed steadily since 2015 and reached a small peak in 2020. Policies on healthy China began in 2016 but experienced a blank period in 2017. They developed rapidly from 2018 to 2019 and then gradually declined. Smart health care policies developed rapidly from 2015 to 2018 and then declined year by year.

XSL•FO
RenderX

Table 12. 2D distribution table of policy orientations and release time of the integrated health management system (n=1345).

Policy orientation	2014 (n=9), n (%)	2015 (n=168), n (%)	2016 (n=234), n (%)	2017 (n=531), n (%)	2018 (n=149), n (%)	2019 (n=135), n (%)	2020 (n=77), n (%)	2021 (n=42), n (%)	Total, n (%)
Graded diagnosis and treatment	9 (100)	104 (61.9)	201 (85.9)	54 (10.2)	55 (36.9)	57 (42.2)	N/A[a]	6 (14.3)	486 (36.1)
Healthy community	N/A	2 (1.2)	1 (0.4)	1 (0.2)	1 (0.7)	4 (3)	10 (13)	5 (11.9)	24 (1.8)
Healthy China	N/A	N/A	2 (0.9)	N/A	1 (0.7)	12 (8.9)	5 (6.5)	1 (2.4)	21 (1.6)
Medical association	N/A	61 (36.3)	28 (12)	475 (89.5)	77 (51.7)	53 (39.3)	55 (71.4)	29 (69)	778 (57.8)
Smart health care	N/A	1 (0.6)	2 (1.2)	1 (0.2)	15 (10.1)	9 (6.7)	7 (9.1)	1 (2.4)	36 (2.7)

[a]N/A: not applicable.

Figure 4. Change chart for the quantity of policies on each policy orientation of the integrated health management system under the time axis.

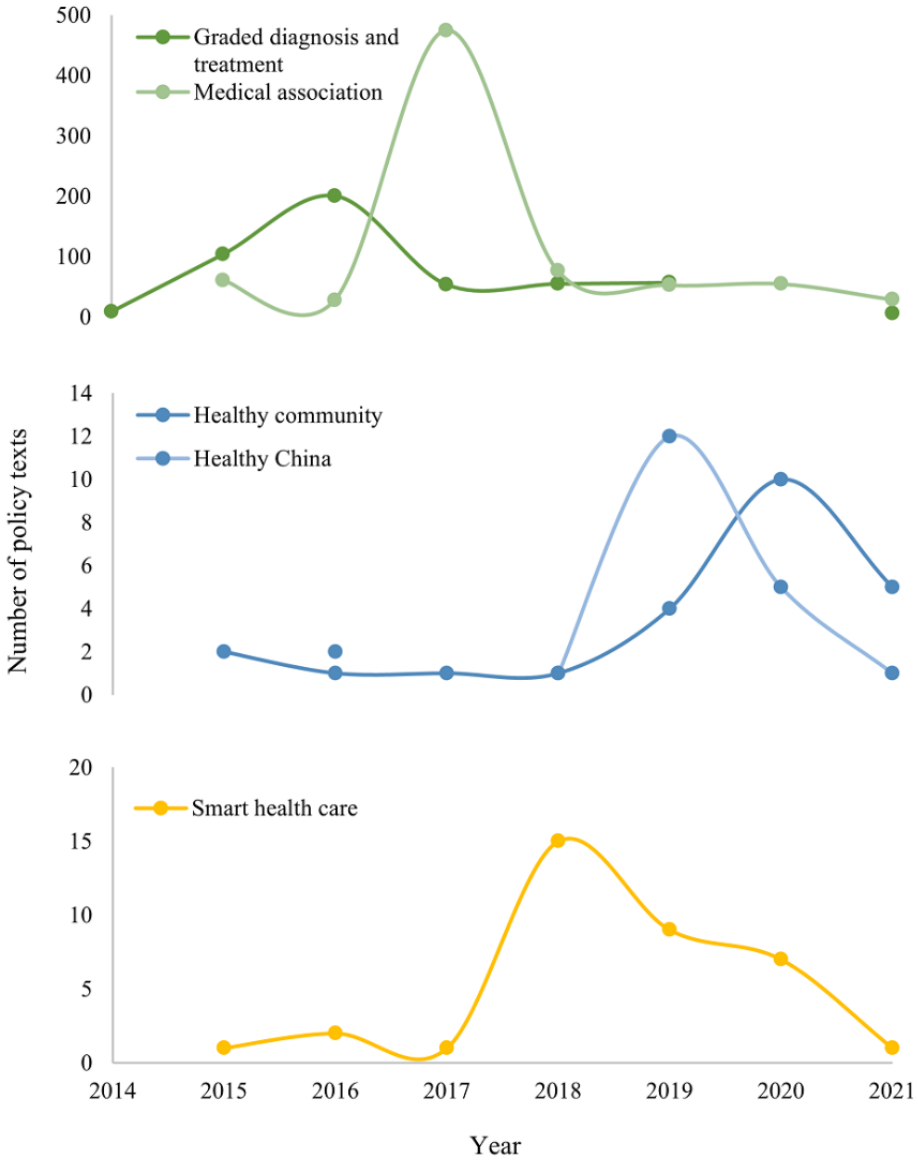

XSL•FO
RenderX

Discussion

Principal Findings

Essential Features

This study found that China's IHMS policy showed more distinctive features under the 3 dimensions of release time, policy type, and policy level. In terms of release time, there was an abrupt increase before 2017, followed by a decline from 2018 to 2019. They remained at a high level, whereas the last 2 years saw a significant downward trend (the effect of the statistical period was not excluded). In terms of policy nature, the number of China's IHMS policy texts on planning and implementation regulations was comparable, but the combination of the time and type of policy documents released shows that there were more general documents such as outlines and plans before 2017, more implementation documents such as opinions and notices from 2018 to 2019, and more prescriptive documents such as programs and methods from 2020 to the present, suggesting that China's IHMS policy construction has gradually improved from top-level design to work requirements. In terms of policy level, IHMS policy texts at the provincial level were the most numerous, followed by the national and other municipal levels and, finally, Shanghai, with the number of policy texts across all of them exceeding 150, indicating that each local government is following the central government's promotion of health consortium construction at a faster pace and with a higher degree of importance.

Coordination of Policy Tools and Stakeholders

The Proportion of Policy Tools Used and the Balance of Internal Structure Still Need to Be Improved

To optimize the proportion of policy tools, we need to focus on increasing the use of supply-based policy tools. This can be achieved by expanding the talent team, increasing scientific and technological information support, providing financial support, introducing advanced diagnosis and treatment technologies, and using public services to improve the coverage of medical services in hospitals at all levels. Demand-based policy tools also need to be further developed to fully use the economic leverage effect of medical insurance funds. This can be done by speeding up the establishment of a unified medical insurance settlement system and medical price system and supporting social medical institutions to join the system construction and expand multilevel and diversified services. It is important to note that, although environment-based policy tools are used more frequently, their limited permeability to the construction of the IHMS requires the government to reduce the proportion of some planning tools. By doing so, we can achieve a more balanced internal structure of policy tools in the construction of the IHMS.

The internal structure of policy tools requires further optimization. Supply-based policy tools such as *technical support* and *financial guarantee* are underused, and demand-based policy tools such as *adherence to public welfare*, *price adjustment*, and *medicine supply* need to be strengthened. Environment-based policy tools such as *functional localization*, *institutional building*, and *objective programming* are more closely related to IHMS, but *policy promotion* is relatively less used. First, the government needs to increase financial input, optimize the investment direction and structure, and pay attention to the proportion of the application of financial support tools. Medical institutions should upgrade their key technologies and treatment methods, especially for chronic and major diseases. Second, the government should provide tax deductions or preferential treatment for different modes of IHMS, establish a dynamic adjustment mechanism for medical service prices, and control drug prices and supply. Finally, the government should strengthen policy propaganda, use new media to popularize and publicize the IHMS, and provide training for medical staff and health education for residents. The government should properly control the use of tools with strong intervention; give relevant units autonomy; and mobilize participants' enthusiasm and sense of responsibility through performance incentives, taxation, and subsidies.

The IHMS Policy Still Needs to Balance the Attention Paid to All Stakeholders

The current IHMS policy tools show a greater focus on the government and pay less attention to hospitals, physicians, and residents at all levels, particularly grassroots hospitals, physicians, and residents. Therefore, adjustments must be made to the use ratio of policy tools, with a greater emphasis on supporting grassroots hospitals to enhance their service capacity and efficiency through talent and technology support. Second-class-and-above hospitals should also collaborate with grassroots hospitals to carry out exchange projects and establish a connection. The government should also urge all hospitals to accelerate the implementation of the multipoint practice system of physicians and the contract system of family physicians, build a multilevel team of physicians, and establish a physician cooperation mechanism to improve the ability of primary medical services through counterpart assistance and business guidance. Furthermore, the government must improve residents' awareness and participation in the policy; guide residents in need to experience the medical treatment process under the IHMS operating system; and encourage them to actively participate in health knowledge forums, IHMS medical treatment exchange meetings, and other activities. Finally, more third-party health institutions such as medical insurance departments and rehabilitation centers should be included so that residents can enjoy health services and protection throughout their life cycle.

Level Differences

Level Differences of Policy Tools

Governments at different levels exhibited similarities and differences in their use of policy tools for IHMS. Environment-based policy instruments were most frequently used at each level of IHMS policy, whereas supply-based policy tools were least commonly used at the national level and in other provincial and local IHMS policies. Demand-based policy tools were least prevalent in Shanghai. Specifically, the proportion of *rational layout*, *talent construction*, *institutional building*, and *objective programming* in the IHMS policies issued by the central government was relatively low. The proportion of *demonstration project*, *resource sharing*, and

XSL•FO
RenderX

incentive and restraint in the IHMS policies issued by other provinces and cities was relatively low. Finally, *talent construction* and *institutional building* were rarely used in the IHMS policies in Shanghai. Therefore, the central government should make use of low-frequency policy tools more frequently, whereas other governments at all levels should respond to the state's initiatives and incorporate local characteristics to enhance the current structure's rationality.

Local authorities can take into account local characteristics and respond to national initiatives by considering the following features to improve the effectiveness of the current structure:

1. Local medical resources: as medical resources are distributed differently across regions, authorities can tailor their policies to local circumstances considering the availability and workforce of local medical facilities. For example, where medical resources are in short supply, the government can introduce incentive schemes to attract medical practitioners and other professionals to work in that region.
2. Demand for medical services: given the differences in the demographic structure, disease spectrum, and health requirements of different regions, local governments can develop targeted policies that cater to specific local conditions. For instance, in areas with a significant aging population, the government can offer more support for older adult care, long-term care, and management of geriatric diseases.
3. Local culture and community characteristics: each region has its own cultural and community features, which local governments can include in smart health care and healthy community policies to improve residents' sense of identification and engagement. For instance, health-related promotional activities can coincide with special folk festivals to inspire local participation in diverse activities focused on health promotion.

Level Differences of Policy Orientation

Different levels of IHMS policies have varying emphases on different aspects. The national IHMS policy places greater emphasis on smart health care (including digital health), whereas Shanghai has made a better effort to create healthy communities and promote a healthy China. To achieve a balanced growth, local governments should accurately understand the central government's primary policy orientation, catch up with the promotion of smart health care policy, and simultaneously increase their efforts to support the construction of healthy communities and healthy China. This will help create a more comprehensive and well-rounded IHMS policy system that can better meet the needs of the people.

Local governments can do the following to strengthen support for smart health care and healthy community:

1. Healthy community construction: local governments can provide financial and resource support to develop healthy communities that offer convenient medical and health services to residents, for instance, set up health consultation centers or community medical stations or organizing regular health checkup activities.
2. Smart medical technology promotion: local governments can organize training and promotion activities to improve medical staff's knowledge of and ability to use smart medical technology and promote the adoption of advanced IT by medical institutions to improve the efficiency of diagnosis and treatment and the quality of service. For example, promoting telemedicine and electronic medical record systems could improve the use of medical resources.
3. Health education and publicity: local governments can increase their efforts in health education and publicity to enhance residents' health awareness and literacy. They can achieve this goal by conducting health information lectures, publishing health promotion materials, and organizing health activities to foster healthy behaviors among residents.

Nature Differences

Differences in the Nature of Policy Tools

Demand-based policy tools constituted a relatively high proportion of the planned IHMS policy, whereas environment-based policy tools constituted a relatively high proportion of the implementation details. However, *technical support*, *adherence to public welfare*, and *policy promotion* were not used very often. Conversely, the planned IHMS policy had a very low use of *demonstration projects* and *incentive and restraint* tools, whereas the comprehensive IHMS policy used fewer mechanisms such as *talent construction*, *financial guarantee*, and *medical insurance payment.* Therefore, the government should prioritize the implementation plan and important measures such as medical insurance and price reform and strengthen the planning and design of pilot systems, assessments, and incentive systems.

Using health insurance reform as an example, the government can take the following steps to strengthen planning and design to ensure smooth implementation and effective outcomes:

1. Creating a detailed implementation program: the government must develop a detailed implementation program that outlines the specific content, timeline, and departments responsible for each policy measure. For example, the mechanism for adjusting drug prices and payment for medical services should be clarified to ensure that the various reform initiatives can be implemented in an orderly manner.
2. Assessing risk factors: during the planning and design process, the government should fully consider potential risks and challenges and create responsive strategies. As an illustration, when adjusting health insurance payments that could provoke protests from some hospitals, the government should anticipate this and create solutions beforehand.
3. Monitoring and evaluation mechanism: establish a robust monitoring and evaluation mechanism to regularly assess the effectiveness of the health care reform implementation so that problems can be identified and policies can be adjusted in a responsive manner. The government can set up a special monitoring team to report regularly on the progress and effectiveness of the reform, thereby increasing transparency and accountability.

XSL•FO
RenderX

Differences in the Nature of Policy Orientation

Both the planned and detailed implementation rules for IHMS policies prioritize medical association and graded diagnosis and treatment while placing less emphasis on smart health care, healthy community, and healthy China. However, in contrast, the planned policies give more importance to healthy China, whereas the implementation rules concentrate more on smart health care and healthy community. To strengthen the implementation of the healthy China policy, the government should supplement it with detailed policies such as implementation plans and methods. For the smart health care and healthy community policies, which have stronger operational attributes, the government should develop comprehensive plans to improve the system of IHMS policies.

For the operational attributes of the smart health care and healthy community policies, the government should develop a comprehensive plan to improve its policy system, including the following:

1. Technical support and standards: the government should establish a unified technical support system for smart health care and healthy communities. Relevant technical standards and specifications should be developed to ensure the consistency and interoperability of implementation across regions.
2. Data management and privacy protection: the government should formulate sound data management policies and privacy protection measures to ensure that the personal health data involved in smart health care and healthy communities are properly managed and protected to prevent data leakage and abuse.
3. Cooperation mechanisms and coordination bodies: the government can develop a cross-sectoral and interdisciplinary cooperation mechanism and set up a special body to facilitate the implementation and coordination of smart health care and healthy community policies. For example, a special committee can be established comprising relevant departments and experts to be responsible for policy formulation, resource allocation, and coordination of promotion activities.

Time Evolution

Time Evolution of Policy Tools

Generally speaking, from 2014 to 2021, there was a trend of rising and falling for various policy tools. Among them, environment-based policy tools were used more frequently, whereas demand-based policy tools were relatively insufficiently used in the past 3 years. In the past 2 years, there has been a shortage of *financial guarantee*, *medical insurance payment*, *medicine supply*, *price adjustment*, *adherence to public welfare*, *demonstration projects*, and *policy promotion*. Although the amount of *technical support* and *capacity building* was small, it has shown an increasing trend in the last 2 years, whereas the number of other policy tools was small and showed a declining trend. To address this issue, the government should first increase the application of demand-based policy tools and then focus on improving the shortcomings of supply-based and environment-based policy tools.

Time Evolution of Policy Orientation

From 2014 to 2021, the IHMS policy orientation shifted from graded diagnosis and treatment to medical association to smart health care and, finally, to healthy China and healthy community. Graded diagnosis and treatment and medical association policies focus on disease diagnosis and treatment, whereas healthy community and healthy China policies focus on health management and smart health care emphasizes the technical means of IHMS operation. We believe that the development of the IHMS policy system can be divided into 3 stages: the *disease-centered period* (2014-2017), dominated by graded diagnosis and treatment and medical association policies; the *electronic health technology development period* (2017-2019), during which smart health care policies were vigorously developed; and the *health-centered period* (2018-2021), which mainly promoted the healthy China and healthy community policies and tended to lean toward healthy community policies. As a result, the IHMS policy system took smart health care policy as a bridge to realize the transformation from *treatment-centered* to *health-centered*. The government should increase multidimensional horizontal integration exploration with the community as the starting point, give primary physicians a pivotal role, and provide people with all-around and full-cycle health services by using electronic health technology.

Contributions and Limitations

Theoretical Contributions

This study makes 3 theoretical contributions.

First, we expanded the object of policy research and developed a comprehensive analysis. The primary contribution of this study to the field of policy analysis is that we have broadened the object of study; conducted in-depth research on multiple types of policies; and constructed an IHMS policy system with common goals, multiple subjects, and complex interactions. This is different from previous studies (eg, Yang et al [26] and Kaló et al [27]) in that we took the concepts and connotations of IHMS as the basis of our study, selected 5 categories of policies (graded diagnosis and treatment, medical association, healthy China, healthy community, and smart health care) to study, and used content analysis to comprehensively understand their characteristics. This is distinctive in that it integrates different policies into a holistic framework and explores their interrelationships. This comprehensive analytical approach integrates quantitative statistics and qualitative analysis and allowed us to examine the characteristics and determinants of IHMS policies from multiple dimensions, making the findings more comprehensive and richer. This comprehensive analytical approach allowed us to gain an in-depth understanding of the interrelationships between the dimensions, thus providing more accurate research conclusions. With the data support of quantitative statistics, we enhanced the credibility of our study and also enriched its content, making it more valuable in both theory and practice.

Second, we enriched the level of policy analysis and provided decision support. The second contribution of this study is that we included different levels of policy factors in our policy

XSL•FO
RenderX

analysis, in particular by considering local policies. This differs from previous studies (eg, Ebrahim et al [28] and Wu et al [29]) that focused on only one level of policy analysis as we showed the diffusion, speed, and impact of policies by comparing the similarities and differences between national and local policies in other dimensions, which helped identify regional characteristics for future research. This is characterized by the fact that we extended the influence of policies to different levels, thus providing a more comprehensive understanding of the multidimensional effects of policies. This contributes to a better understanding of the complexity of policies and provides a more accurate reference and support for policy makers. By analyzing the relationships between the dimensions in depth, we drew conclusions that provide practical guidance for policy formulation and implementation. The value of this practical application not only makes our findings meaningful in the academic field but also provides useful references and decision support for policy makers.

Third, we improved the policy analysis framework to achieve a systematic analysis. The third contribution is that we improved the existing policy analysis framework to make it more applicable to policy cluster or policy network analysis. Compared with previous studies that proposed 2D (Xie and Tian [30]), 3D (Shen et al [31]), or 4D (Yang et al [26]) frameworks, we proposed a 6D policy analysis framework that includes the dimensions of policy tools, stakeholders, policy level, policy nature, policy orientation, and release time and takes policy orientation and policy tool as the core dimensions, focusing on their hierarchical and temporal differences. It is characterized by a more comprehensive analytical framework covering multiple dimensions of consideration. This systematic approach to analysis provided us with a more comprehensive and multidimensional perspective, which helped us gain insights into the characteristics and influences of policies. This improvement enabled us to better grasp the relationships between the dimensions and, thus, gain a more comprehensive understanding of the nature of the policy. This systematic analysis added depth and breadth to our study and helped understand the multilevel and multidimensional characteristics of policy in a more comprehensive way, thus providing a more valuable reference base for future policy formulation and implementation.

Practical Enlightenment

These results have practical value in 2 respects. First, they aid central policy makers in evaluating the macrostructural design of the IHMS to optimize it. The results can also assist local policy makers in better understanding the central authorities' emphasized key contents and in improving common issues, leading to IHMS policy promotion based on local conditions. Second, the results can assist IHMS members in comprehending their respective responsibilities and work priorities and taking measures to enhance their division of labor and cooperation.

Limitations and Future Work

This study has some limitations. First, it does not provide a more in-depth analysis combined with regional division, which limits the regional characteristics obtained. Second, frequency or frequency-based statistics are only applicable to text corpora that do not reach the level of big data and cannot effectively mine the hidden information in policy texts. Third, owing to time and resource constraints, we did not use machine learning models for implicit semantic analysis, and some text-mining algorithms could be incorporated to deal with large-scale public policy texts in future research.

Conclusions

China's IHMS policy has specified the development objectives and core contents of the health consortium and facilitated the promotion of the rational allocation of health resources in China. This paper presents a comprehensive analysis of China's IHMS policy from 6 perspectives, namely, policy level, policy nature, release time, policy tools, stakeholders, and policy orientation, using the PRISMA guidelines and content analysis to reveal the changes and hierarchical differences in the structure of different types of policies and policy tools in the system and provide some possible lessons for future policy formulation and optimization.

Acknowledgments

The authors polished the translation results from the original Chinese manuscript with the assistance of ChatGPT (version 3.5; OpenAI), which helped them improve the English expressions of some parts of the first draft to be more in line with the grammar and reading habits of native English speakers. The ChatGPT conversations used to improve the translation results from the original manuscript are available on the internet [32]. The study was not registered, the protocol was not prepared, and no financial support was received.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YY and SW contributed to study conception and design. YY and LY collected, analyzed, and interpreted the data. YY and SW drafted the manuscript, and SW was responsible for its critical revision.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Specifications of China's integrated health management system policy.
[DOCX File , 19 KB - jmir_v26i1e47197_app1.docx]

Multimedia Appendix 2
Explanation of the 6D framework.
[DOCX File , 18 KB - jmir_v26i1e47197_app2.docx]

Multimedia Appendix 3
The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e47197_app3.docx]

Multimedia Appendix 4
Detailed information on the 152 China integrated health management system policy documents collected.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e47197_app4.xlsx]

Multimedia Appendix 5
Classifications of the policy tools and stakeholders of the integrated health management system.
[DOCX File , 20 KB - jmir_v26i1e47197_app5.docx]

Multimedia Appendix 6
Specific steps for coding and manual labeling.
[DOCX File , 17 KB - jmir_v26i1e47197_app6.docx]

References

1. Integration of health care delivery: report of a WHO study group. World Health Organization. 1996. URL: https://iris.who.int/bitstream/handle/10665/38408/WHO_TRS_861.pdf?sequence=1 [accessed 2022-08-08]
2. Tao S. Health consortium: taking health as the center. Chin Health 2018(09):21-22. [doi: 10.15973/j.cnki.cn11-3708/d.2018.09.005]
3. Xiong M, Wu J, Liu L. Enlightenment of the typical models of integrated delivery system abroad to the development of integrated health management system in China. Chin Gen Pract 2020;23(22):2741-2748.
4. WHO global strategy on people-centred and integrated health services: interim report. World Health Organization. 2015. URL: https://iris.who.int/handle/10665/155002 [accessed 2022-08-22]
5. RAND Europe, Ernst & Young LLP. National evaluation of the DH integrated care pilots. Rand Health Q 2012 Mar 1;2(1):8 [FREE Full text] [Medline: 28083230]
6. Kodner DL, Spreeuwenberg C. Integrated care: meaning, logic, applications, and implications--a discussion paper. Int J Integr Care 2002 Nov 14;2(4):e12 [FREE Full text] [doi: 10.5334/ijic.67] [Medline: 16896389]
7. Fulop N, Mowlem A, Edwards N. Building integrated care: lessons from the UK and elsewhere: No. 37. NHS Confederation. 2005. URL: https://www.abebooks.com/9781859471227/Building-Integrated-Care-Lessons-Elsewhere-1859471226/plp [accessed 2024-01-01]
8. Contandriopoulos AP, Denis JL, Touati N, Rodríguez C. The integration of health care: dimensions and implementation. Université de Montréal. 2003. URL: http://nelhin.on.ca/assets/0/16/2100/3734/3736/6cab135d-87c1-45bd-88cd-2c1d5404ec9b.pdf [accessed 2024-01-01]
9. Goodwin N. Understanding integrated care. Int J Integr Care 2016 Oct 28;16(4):6 [FREE Full text] [doi: 10.5334/ijic.2530] [Medline: 28316546]
10. Berwick DM, Nolan TW, Whittington J. The Triple Aim: care, health, and cost. Health Aff (Millwood) 2008 May 1;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
11. Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf 2015 Oct 2;24(10):608-610. [doi: 10.1136/bmjqs-2015-004160] [Medline: 26038586]
12. Nurjono M, Valentijn PP, Bautista MA, Wei LY, Vrijhoef HJ. A prospective validation study of a rainbow model of integrated care measurement tool in Singapore. Int J Integr Care 2016 Apr 8;16(1):1 [FREE Full text] [doi: 10.5334/ijic.2021] [Medline: 27616946]
13. Chaitkin M, Blanchet N, Su Y, Husband R, Moon P, Rowan A, et al. Integrating vertical programs into primary health care: a decision-making approach for policymakers. Results for Development. 2019. URL: https://r4d.org/resources/

XSL•FO
RenderX

integrating-vertical-programs-into-primary-health-care-a-decision-making-approach-for-policymakers/#:~:text=%E2%80%9CIntegrating%20Vertical%20Programs%20into%20Primary,with%20a%20focus%20on%20PHC [accessed 2024-01-01]
14. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating eHealth into health care: systematic literature review of the Donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180 [FREE Full text] [doi: 10.2196/27180] [Medline: 33970123]
15. Fatoum H, Hanna S, Halamka JD, Sicker DC, Spangenberg P, Hashmi SK. Blockchain integration with digital technology and the future of health care ecosystems: systematic review. J Med Internet Res 2021 Nov 2;23(11):e19846 [FREE Full text] [doi: 10.2196/19846] [Medline: 34726603]
16. de Luca V, Lazic V, Birov S, Piesche K, Beyhan O, Pengo MF, et al. Digitally enabled health service for the integrated management of hypertension: a participatory user-centred design process. Int J Environ Res Public Health 2021 Nov 26;18(23):12442 [FREE Full text] [doi: 10.3390/ijerph182312442] [Medline: 34886167]
17. Ling T, Li J, Ji C, Zhang H, Ge H. The policy review and status analysis of China's medical cluster. Chin Hospitals 2020 Apr 30;24(7):29-32. [doi: 10.19660/j.issn.1671-0592.2020.07.10]
18. Yang L, Wang W, Li S. Analysis on the effect of policy implementation of medical cluster in Guizhou Province based on the comprehensive model of policy implementation. Chin Hospitals 2021 Sep 28;25(11):28-32. [doi: 10.19660/j.issn.1671-0592.2021.11.09]
19. Chen HY, Yan RY, Gao JY, Chen Z, Cao Z, Han C. Quantitative analysis on China's medical consortium construction policy from the perspective of policy tools. Chinese Health Economics 2019 Aug 05;38(11):20-23 [FREE Full text] [doi: 10.7664/CHE20191104]
20. Bai S. Constructing a health management consortium to promote healthy China. Chin J Health Manage 2017;11(6):481-484. [doi: 10.3760/cma.j.issn.1674-0815.2017.06.001]
21. Huang F, You J, Tu L, Zeng W. Systematic review methodology: cases, procedures, and values. e Educ Res 2017;38(11):11-8+25Core [FREE Full text] [doi: 10.13811/j.cnki.eer.2017.11.002]
22. Tang SL, Si Y. Quantitative research on policies related to my country's medical consortium from the perspective of policy tools. Health Soft Sci 2020;34(1):32-36. [doi: 10.3969/j.issn.1003-2800.2020.01.007]
23. Shen H, Zhao J, Ma H, Fu Y, Yang L. Policy text analysis on health care big data in China——from the perspective of policy tools. Health Econ Res 2021;38(8):11-15 [FREE Full text]
24. Rothwell R, Zegveld W. Industrial Innovation and Public Policy: Preparing for the 1980s and the 1990s. London, UK: Frances Pinter; 1981.
25. Liu X, Yan WX, Liu J. Policies for construction of public health emergency management system in China: a policy tool-based quantitative document text analysis. Chin J Public Health 2021 Jan 28;37(5):853-859. [doi: 10.11847/zgggws1134142]
26. Yang F, Shu H, Zhang X. Understanding "internet plus healthcare" in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779 [FREE Full text] [doi: 10.2196/23779] [Medline: 34309581]
27. Kaló Z, Holtorf AP, Alfonso-Cristancho R, Shen J, Ágh T, Inotai A, et al. Need for multicriteria evaluation of generic drug policies. Value Health 2015 Mar;18(2):346-351 [FREE Full text] [doi: 10.1016/j.jval.2014.12.012] [Medline: 25773570]
28. Ebrahim S, Ashworth H, Noah C, Kadambi A, Toumi A, Chhatwal J. Reduction of COVID-19 incidence and nonpharmacologic interventions: analysis using a US county-level policy data set. J Med Internet Res 2020 Dec 21;22(12):e24614 [FREE Full text] [doi: 10.2196/24614] [Medline: 33302253]
29. Wu Z, He Q, Yang K, Zhang J, Xu K. Investigating the dynamics of China's green building policy development from 1986 to 2019. Int J Environ Res Public Health 2020 Dec 29;18(1):196 [FREE Full text] [doi: 10.3390/ijerph18010196] [Medline: 33383834]
30. Xie Q, Tian Z. How innovation policy promotes China's new energy vehicle industry: a text research from policy tools and innovation value chain. Sci Sci Manag S T 2015;36(6):3-14.
31. Shen N, Hu S, Peng H. Can environmental regulation stimulate green innovation? --visual analysis based on point-line-plane three-dimensional frame. China Population, Resources and Environment 2020;30(4):75-84 [FREE Full text] [doi: 10.12062/cpre.20191117]
32. Analyzing China health policy. ChatGPT. 2023 Oct 16. URL: https://chat.openai.com/share/fe38882d-fb0d-4476-b6f2-73041508b914 [accessed 2024-01-04]

Abbreviations

IHMS: integrated health management system
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.03.23; peer-reviewed by Y Du, H Shen, X Zhao; comments to author 05.06.23; revised version received 26.08.23; accepted 25.10.23; published 24.01.24.

Please cite as:
Yu Y, Wang S, You L
Understanding the Integrated Health Management System Policy in China From Multiple Perspectives: Systematic Review and Content Analysis
J Med Internet Res 2024;26:e47197
URL: https://www.jmir.org/2024/1/e47197
doi:10.2196/47197
PMID:38265862

©Yang Yu, Sufen Wang, Lijue You. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review

Fouzia Amar[1], MEng; Alain April[1], BSc, MSc, PhD; Alain Abran[1], BSc, BA, MASc, MSc MGT, PhD

École de technologie supérieure - ETS, Montreal, QC, Canada

Corresponding Author:
Fouzia Amar, MEng
École de technologie supérieure - ETS
1100 Notre Dame Ouest
Montreal, QC, H3C 1K3
Canada
Phone: 1 514 396 8800
Email: famar2022@gmail.com

Abstract

Background: The increasing use of electronic health records and the Internet of Things has led to interoperability issues at different levels (structural and semantic). Standards are important not only for successfully exchanging data but also for appropriately interpreting them (semantic interoperability). Thus, to facilitate the semantic interoperability of data exchanged in health care, considerable resources have been deployed to improve the quality of shared clinical data by structuring and mapping them to the Fast Healthcare Interoperability Resources (FHIR) standard.

Objective: The aims of this study are 2-fold: to inventory the studies on FHIR semantic interoperability resources and terminologies and to identify and classify the approaches and contributions proposed in these studies.

Methods: A systematic mapping review (SMR) was conducted using 10 electronic databases as sources of information for inventory and review studies published during 2012 to 2022 on the development and improvement of semantic interoperability using the FHIR standard.

Results: A total of 70 FHIR studies were selected and analyzed to identify FHIR resource types and terminologies from a semantic perspective. The proposed semantic approaches were classified into 6 categories, namely mapping (31/126, 24.6%), terminology services (18/126, 14.3%), resource description framework or web ontology language–based proposals (24/126, 19%), annotation proposals (18/126, 14.3%), machine learning (ML) and natural language processing (NLP) proposals (20/126, 15.9%), and ontology-based proposals (15/126, 11.9%). From 2012 to 2022, there has been continued research in 6 categories of approaches as well as in new and emerging annotations and ML and NLP proposals. This SMR also classifies the contributions of the selected studies into 5 categories: framework or architecture proposals, model proposals, technique proposals, comparison services, and tool proposals. The most frequent type of contribution is the proposal of a framework or architecture to enable semantic interoperability.

Conclusions: This SMR provides a classification of the different solutions proposed to address semantic interoperability using FHIR at different levels: collecting, extracting and annotating data, modeling electronic health record data from legacy systems, and applying transformation and mapping to FHIR models and terminologies. The use of ML and NLP for unstructured data is promising and has been applied to specific use case scenarios. In addition, terminology services are needed to accelerate their use and adoption; furthermore, techniques and tools to automate annotation and ontology comparison should help reduce human interaction.

(J Med Internet Res 2024;26:e45209) doi:10.2196/45209

KEYWORDS

electronic health record; EHR; Health Level Seven International Fast Healthcare Interoperability Resources; HL7 FHIR; interoperability, web ontology language; OWL; ontology; semantic; terminology; resource description framework; RDF; machine learning; ML; natural language processing; NLP

XSL•FO
RenderX

Introduction

Background

The development and deployment of electronic health records (EHRs) worldwide has led to several interoperability challenges, including extracting valuable information from free-text clinical notes, integrating unstructured EHR patient data, and reusing them without ambiguity for clinical research purposes [1,2].

Interoperability is defined by the Institute of Electrical and Electronics Engineers (IEEE) as the "ability of two or more components to exchange information and to use the information that has been exchanged" [3]. A total of 4 interoperability levels were proposed by the Healthcare Information and Management Systems Society organization: (1) foundational, (2) structural, (3) semantic, and (4) organizational (Figure 1) [4].

Figure 1. Interoperability levels (Healthcare Information and Management Systems Society).

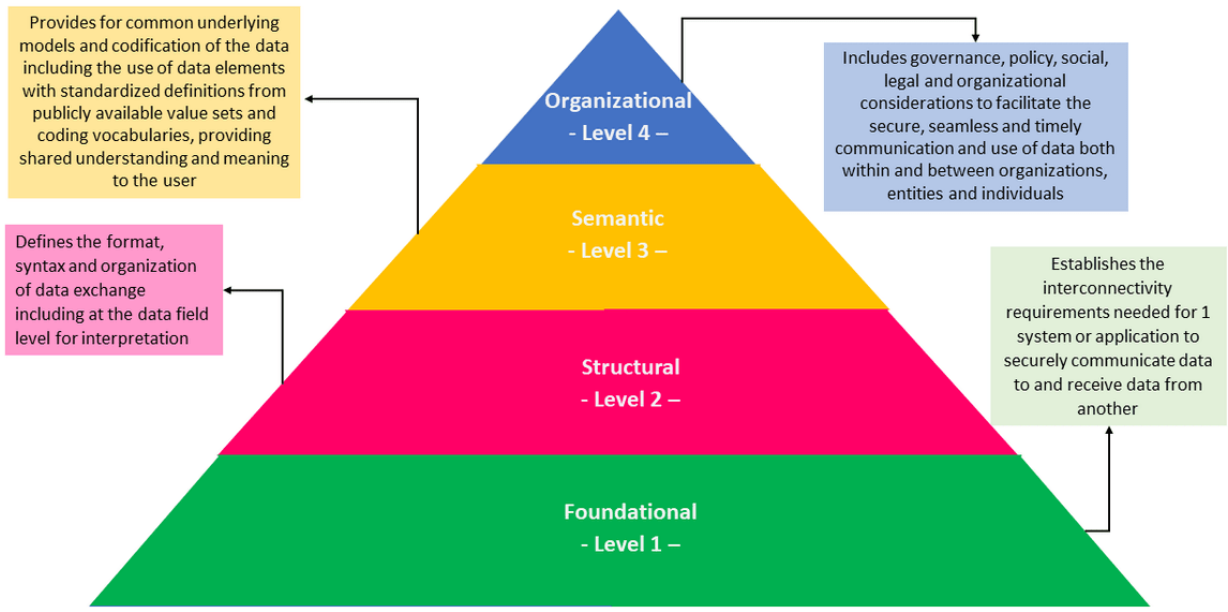

Important research efforts are underway to develop health care standards, including Fast Healthcare Interoperability Resources (FHIR) standards for the electronic exchange of health care information. This standard was developed by the Health Level Seven International (HL7) organization aiming at the evolution of messaging standards to achieve semantic interoperability [5]. FHIR combines the best features of HL7 version 2, version 3, and clinical document architecture, while leveraging the latest web standards.

FHIR interoperability solutions are designed from a set of modular components called "resources" that have been designed as small reusable components defining a set of properties for the capture of the structure of the domain data acquisition for a particular use. The version 4.3.0 release of FHIR R4B proposes 143 different "resources" classified into five categories to handle health data: (1) foundation resources, (2) base resources, (3) clinical resources, (4) financial resources, and (5) specialized resources (Multimedia Appendix 1).

Many elements in FHIR use coded values and allow the use of external terminologies, such as (1) Systematized Nomenclature of Medicine–clinical terms (SNOMED CT)—comprehensive multilingual clinical terminology used in EHR. Its components include concepts (codes), descriptions (terms), and relationships—and (2) Logical Observation Identifiers Names and Codes (LOINC), a set of universal names and ID codes used to identify laboratory and clinical test results.

Common terminology aims to address the problem of ambiguity in exchanged data [6]. However, most legacy systems do not take advantage of a terminology-based approach to data exchange and still use incompatible terminologies or unstructured data in their exchanges with EHR. Thus, to ensure semantic interoperability with these legacy applications, it is mandatory to perform manual data transformations, mapping, or other alternate solutions to ensure that the exchanged data are interpreted correctly by the information systems involved.

The FHIR standard is of great interest to researchers and EHR ecosystem developers [7], and several technology giants, such as Amazon, Google, IBM, Microsoft, Oracle, and Salesforce, offer ready-to-use cloud services to promote the adoption of FHIR [8].

In this study, we report on a systematic mapping Review (SMR) and propose a classification of the FHIR literature based on FHIR resources, terminologies, research approaches, and innovative contributions. The goal of this SMR was to provide state of the art, trends, and analysis of research related to semantic interoperability using FHIR. This activity specifically aimed to identify the FHIR and terminology used, as well as

XSL•FO
RenderX

the different proposed approaches, contribution types, and trends.

Related Work

Of the number of literature reviews on the challenges of health information systems interoperability, only 4 specifically address FHIR standard issues [2,9-11]. This section summarizes the findings of this study.

The study by Ayaz et al [10] reviewed 80 studies published between 2012 and 2019 that identified the opportunities and challenges associated with using FHIR. Another FHIR systematic review by Lehne et al [9] identified the main issues to be considered when exchanging data based on only 2 data sources and presented an overview of 131 selected studies up to February 4, 2019. A third systematic literature review by Setyawan et al [2] on data interoperability issues and possible solutions selected 39 studies published between 2018 and 2021 from 3 data sources (ie, IEEE Xplore, ScienceDirect, and Scopus). Finally, a systematic review of semantic interoperability in the study by de Mello et al [11] selected 28 studies from 2012 to 2020 from 8 different sources.

An overview of these 4 studies is presented in Table 1 in terms of their study goal or objective, period covered, number of studies selected and analyzed, source databases, and their respective mapping and research questions (RQs).

Table 1. Overview of the 4 literature reviews on semantic issues when using Fast Healthcare Interoperability Resources (FHIR).

ID	Study, year	Goal or objective	Period covered	Studies, n	Source databases	Mapping and research questions
R1	Ayaz et al [10], 2021	A systematic review of the literature related to FHIR, including the challenges, implementation, opportunities, and future FHIR applications	January 2012-December 2019	80	• ACM • IEEE • Springer • Google Scholar • PubMed • ScienceDirect	• Q1: Types or models of FHIR implementation? • Q2: Common resources used in FHIR implementation? • Q3: Applications that benefit from the use of FHIR? • Q4: Approaches applied to map or migrate data from previous standards to FHIR? • Q5: Goals of FHIR? • Q6: Challenges and open questions related to the FHIR domain?
R2	Lehne et al [9], 2019	Screen all FHIR publications and identify the main topics	February 4, 2019	131	• Web of science • PubMed	• N/A[a]
R3	Setyawan et al [2], 2021	Identify integration problems of HL7[b] FHIR implementation and provide technical solutions and nontechnical solutions to these problems	2018-2021	39	• ScienceDirect • IEEE • SCOPUS	• Q1: State of the art of data integration and interoperability in HL7 FHIR implementation? • Q2: Problems the researcher has regarding Data Integration and Interoperability in HL7 FHIR Implementation? • Q3: Potential solutions and future research regarding Data Integration and Interoperability in HL7 FHIR Implementation?
R4	de Mello et al [11], 2022	Present a comprehensive systematic literature review of semantic interoperability in electronic health record	2010-2020	28	• ACM • IEEE Xplore • Google scholar • MEDLINE • PubMed • ScienceDirect • SpringerLink • Web of Science	• Q1: State of the art in health standards applied in health records? • Q2: Challenge and open questions to semantic interoperability in health records. • Q3: Health standards adopted in the studies? • Q4: Terminologies or health repositories used? • Q5: Approaches used? • Q6: Main security concerns used? • Q7: Evaluation approaches used?

[a]N/A: not applicable.

[b]HL7: Health Level Seven International.

To the best of our knowledge, no systematic mapping review has been published that specifically addresses the semantic issues raised when using FHIR.

Some of the lessons learned in these studies and their limitations are presented in Table 2. The key findings are summarized as follows:

XSL•FO
RenderX

- Most used FHIR resources from general FHIR implementation up to 2019: Observation, Patient, Practionner, Condition, Medication, AllergyIntolerance, Medication, Device [10].
- The study by Ayaz at al [10] provides a distribution of the different techniques used for the mapping approach (eg, mapping HL7 v2 to FHIR and mapping FHIR to others).
- The formula string using FHIR as the keyword is general and allows maximum results [9]. This study showed that ontology or terminology approaches were subject to a higher number of studies up to 2018.
- One study [2] used a limited number of databases (only 3), and the formula string may limit the obtained results.
- We learned from the study by de Mello et al [11] that the main approaches used are based on ontology and terminology, particularly SNOMED CT, International Classification of Diseases (ICD)-9 or ICD-10.

Table 2. Lessons learned and limitations of related work.

ID	Lessons learned	Limitations
R1	• Interesting mapping framework OR data model into the category of the most selected studies • Common resources used	• More databases are needed • Formula for search may limit the result to semantic FHIR[a] studies • Are the identified resources still of interest for semantic interoperability? • Mapping to terminologies is not covered
R2	• Formula used help to select the maximum number of studies • Topics of interest: ontology, terminology, data models ontology, terminology, data, and information model, mapping • The ontology, terminology, and data models is the subject of a high number of the selected studies	• Limited list of databases • The number of studies continues to increase as of 2018
R3	• N/A[b]	• Limited database sources • The search formula may limit the result as it must have the following: (challenge* OR problem* OR issue* OR trend*)
R4	• Period: 2010-2020 • Different approaches based on ontology and the use of terminology, especially SNOMED CT[c], ICD[d]-9 or ICD-10 • The use of OWL[e] and RDF[f] to solve semantic issues	• HL7[g] FHIR is not in the list of identified standards based on the selected studies • The search based on semantic interoperability may limit the result as it used: “semantic interoperability” AND (“health record” OR “medical record” OR “patient record” OR “hospital record”) AND standard • A small number of articles selected

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]N/A: not applicable.

[c]SNOMED CT: Systematized Nomenclature of Medicine–clinical terms.

[d]ICD: International Classification of Diseases.

[e]OWL: web ontology language.

[f]RDF: resource description framework.

[g]HL7: Health Level Seven International.

These previous studies do not provide an overview of post-2019 developments related to semantic interoperability approaches and trends using FHIR. An extensive and more recent study is needed using many databases and a search string that includes the most possible semantic keywords.

In contrast to related works, this SMR focuses on the semantic level of interoperability using FHIR and reports on research directions over the last decade.

Methods

Overview

SMRs and systematic literature reviews typically use the same methodology, but their goals differ. The systematic literature review is driven by specific questions and the review aims to respond to these specific questions. In contrast, an SMR targets more high-level questions aimed at a specific topic, with the goal of providing a classification of the research results. An SMR can be helpful to researchers because it provides an overview of the literature at a given time for a specific topic of interest [12].

According to the study by Petersen et al [13], the mapping review process follows three steps: (1) planning the mapping, (2) conducting the mapping, and (3) reporting the results of the mapping. Furthermore, specific guidelines for a systematic mapping review in the software engineering domain as mentioned in the study by Brereton et al [14] suggest more detailed activities at each step (Figure 2):

- Planning: identification of the different items used for the research process, including RQs, resources and databases, search terms and formulas, inclusion and exclusion criteria, and questions for quality assessment (QA).

XSL•FO
RenderX

- Conducting: initial extraction of the studies was performed in this step. An initial preselection was performed, and both the inclusion and exclusion criteria were applied. The selected studies were then evaluated and filtered using QA questions.
- Reporting: during this step, mapping was conducted using the final selection of the studies, and a discussion of the results was presented.

Figure 2. Mapping and review process for this Fast Healthcare Interoperability Resources study.

Research Questions

A set of 5 questions was defined (Table 3).

Table 3. Research questions.

RQ[a] ID	RQ
RQ1	What are the main resources types studied or used in FHIR[b] semantic interoperability studies?
RQ2	Which terminologies are used in FHIR semantic interoperability studies?
RQ3	What are the semantic research approaches most frequently applied to achieve semantic interoperability?
RQ4	What are the main types of contributions within the selected studies?
RQ5	What are the publication sources and types?

[a]RQ: research question.

[b]FHIR: Fast Healthcare Interoperability Resources.

Mapping Search Strategy

Search String

The search string was established iteratively: first, by using “FHIR” as a search term in 3 databases and screening the results to look for more keywords dealing with the subject of interest, and second, by identifying additional terms and their synonyms based on the RQs.

The following is the search string composed using Boolean operators based on the identified terms: FHIR AND (Semantic OR terminology OR vocabulary OR Ontology OR mapping OR classification).

The Zotero reference management tool [15] was used to collect, organize, annotate, and review the abstract and full text to include or exclude studies that met the eligibility criteria. Zotero helps manage the large number of studies extracted for research.

Search Databases

The following 10 databases were used to identify and extract studies that satisfied the predefined search string: IEEE Xplore, ScienceDirect, Springer, Scopus, ACM, Web of Science, MEDLINE or PubMed, Compendex, Inspec, and Semantic Scholar.

Eligibility Criteria

Textbox 1 shows the inclusion and exclusion criteria for identified and extracted studies. Articles were excluded based on titles and abstracts as well as full-text reading in some cases.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for identified or extracted studies.

Inclusion criteria

- Published between 2012 and 2022 (start date of 2012 was chosen because it was the date of publication of the first draft of the Fast Healthcare Interoperability Resources [FHIR]).
- Published in English or French

Exclusion criteria

- Studies not accessible in full text
- Books
- Studies reported on 2-3 pages
- Duplicate studies: We retained the most recent data
- Technical reports or summaries of conferences and editorials
- Studies that do not address semantics using FHIR
- Studies addressing enhancements to FHIR terminology and comparisons

Quality Assessment

QA of a research publication is a crucial selection step to ensure that the selected studies have the appropriate impact for mapping reviews and further analysis [12]. For this purpose, we defined quality criteria scores based on the following questions.

- QA1: Are the objectives of the research clearly defined?
- QA2: Does the author present results that meet the stated research objectives?
- QA3: Is the method or technique used to achieve semantic interoperability clearly described?
- QA4: Is a literature review provided?
- QA5: Is the author's contribution clearly stated?
- QA6: Is a result validation strategy presented?

Quality evaluation of the identified and extracted studies was performed by answering each QA question. The selected studies satisfied at least 4 of the 6 QA criteria. Others were rejected.

Data Extraction and Data Synthesis

This SMR was conducted in 2022, with publications published from 2012 to 2022. A summary of the data extraction showing the selection filter effect is presented in Figure 3.

- 2601 results were obtained in the first step
- 2372 remained after applying the exclusion criteria for studies published out of the range of 2012-2022
- 1701 remained after the applying the exclusion criteria on venue type is possible, when available
- 1692 for studies remained after applying the English or French language restriction
- 70 studies remained after removing duplicates, applying the additional inclusion, exclusion criteria and QA.

The 70 studies retained are listed in Multimedia Appendix 2 [16-85]. The extracted data are presented in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 3. Search steps and results from databases. QA: quality assessment.

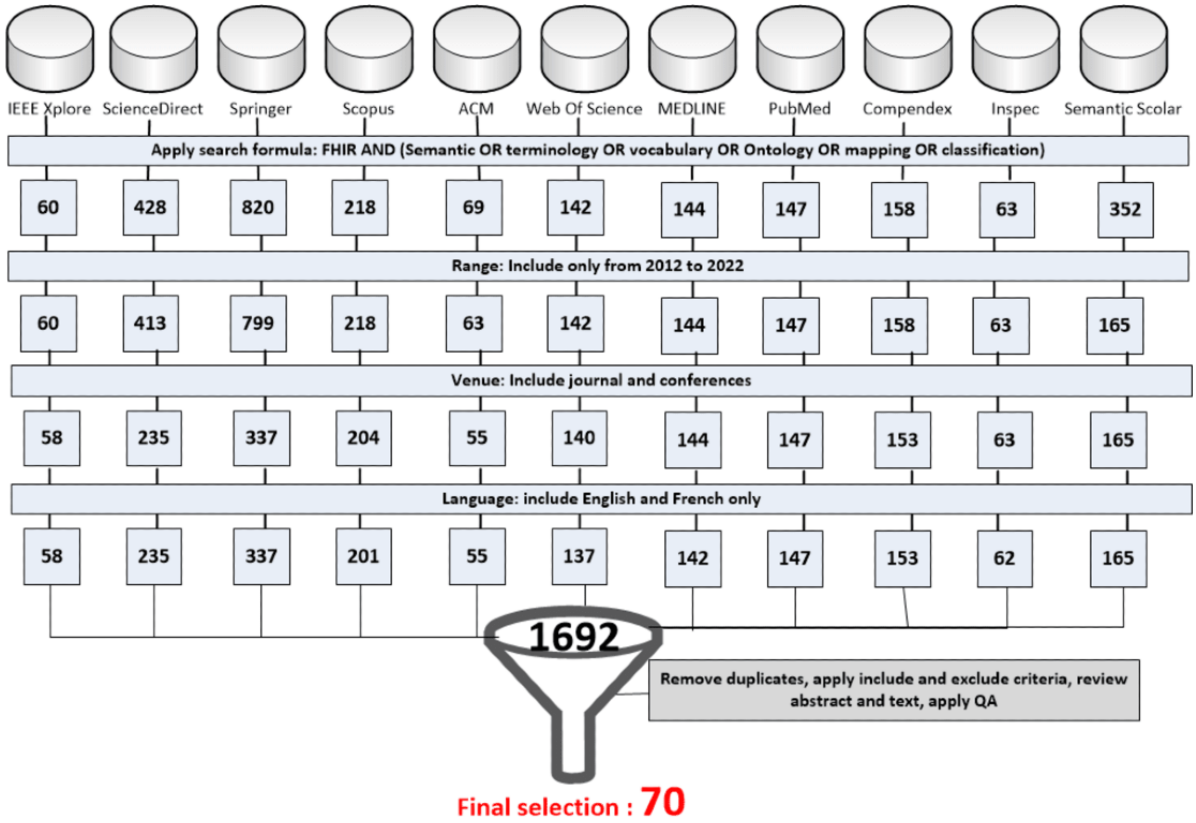

Results

This section presents findings related to the 5 RQs listed in Table 3.

RQ1: Main FHIR Resources Types Studied or Used

We reviewed 70 selected studies and identified the types of FHIR resources that were used. Table 4 lists the 62 types of resources identified in the corresponding studies that have used them. In Figure 4, the resource types are presented, from most frequently used to least used.

The 10 most referred or used resource types in these studies were Observation (31/297, 10.4%), Patient (29/297, 9.8%), Medication (19/297, 6.4%), Condition (16/297, 5.4%), CodeSystem (15/297, 5.1%), Encounter (11/297, 3.7%), Procedure (11/297, 3.7%), Medication Request (10/297, 3.4%), Practitionner (10/297, 3.4%), and ValueSet (9/297, 3%).

The HL7 organization classifies the list of resources into five categories: (1) foundation, (2) base, (3) clinical, (4) financial, and (5) specialized resources. Each of these categories was further divided into 5 subcategories, except for the financial resources category, which was subdivided into 4 subcategories.

The resources inventoried in the 70 selected studies were mostly in 3 clinical subcategories: summary, medications, and diagnosis (Figure 5 and Multimedia Appendix 1). There is one exception to Patient and Encounter resources, which are in the base category and mainly used for identification purposes. We also noted the terminology subcategory (foundation category) among the top 10 most used resources. Terminology resources were used to map concepts to FHIR terminology (eg, SNOMED CT and LOINC).

XSL•FO
RenderX

Table 4. Fast Health Care Interoperability Resources types used and related studies.

Resource type	Study ID	Studies, n
Observation	S2, S42, S43, S68, S47, S60, S48, S6, S7, S8, S9, S11, S12, S40, S28, S13, S14, S15, S54, S55, S56, S38, S22, S61, S23, S62, S5, S34, S69, S70	30
Patient	S1, S42, S43, S68, S65, S46, S47, S60, S48, S50, S39, S9, S11, S13, S14, S15, S16, S54, S55, S30, S56, S38, S22, S23, S5, S34, S52, S69, S70	29
Medication	S1, S68, S36, S47, S63, S4, S60, S48, S6, S39, S9, S14, S54, S55, S31, S22, S61, S5, S69	19
Condition	S25, S68, S60, S6, S39, S8, S9, S11, S12, S40, S14, S15, S54, S22, S61, S62	16
CodeSystem	S37, S9, S29, S10, S40, S41, S17, S18, S20, S61, S33, S14, S15, S21, S28	15
Encounter	S1, S68, S60, S39, S9, S11, S16, S38, S23, S62, S5	11
Procedure	S68, S60, S39, S8, S9, S40, S61, S5, S62, S12, S69	11
Practitionner	S68, S47, S60, S50, S9, S15, S16, S34, S23, S70	10
MedicationRequest	S1, S68, S63, S4, S39, S53, S62, S22, S61, S69	10
ValueSet	S37, S9, S29, S40, S41, S17, S18, S61, S14	9
MedicationStatement	S68, S6, S8, S9, S40, S14, S61	7
FamilyMemberHistory	S68, S60, S39, S8, S15, S54, S61	7
AllergyIntolerance	S68, S60, S9, S67, S28, S15, S62	7
DiagnosisReport	S1, S40, S14, S5, S62, S34, S70	7
CarePlan	S68, S60, S50, S9, S23, S62	6
Device	S2, S43, S68, S60, S7, S39	6
ServiceRequest	S60, S39, S9, S38, S34, S69	6
ConceptMap	S37, S13, S41, S17, S19, S69	6
StructureDefinition	S44, S38, S53, S40, S69	5
Specimen	S60, S9, S40, S34	4
Questionnaire	S28, S15, S30, S23	4
Organization	S60, S16, S38, S23	4
MedicationAdministration	S47, S48, S39, S9	4
Appointment	S60, S9, S28, S69	4
Immunization	S60, S40, S62	3
Composition	S8, S61, S69	3
Communication	S39, S9, S69	3
RiskAssessment	S60, S28, S69	3
HealthcareService	S60, S9	2
Goal	S68, S62	2
EpisodeOfCare	S68, S23	2
DeviceDefinition	S2, S7	2
CareTeam	S68, S60	2
AdverseEvent	S68, S60	2
ActivityDefinition	S23, S43	2
RelatedPerson	S68, S60	2
Coverage	S9, S16	2
ClinicalImpression	S12, S23	2
Location	S68, S69	2
NutritionOrder	S68, S69	2
PlanDefinition	S23, S69	2

XSL•FO
RenderX

Resource type	Study ID	Studies, n
DocumentReference	S60	1
MedicationDispense	S39	1
DeviceUseStatement	S9	1
Consent	S9	1
Person	S28	1
Schedule	S15	1
PractionnerRole	S16	1
QuestionnaireResponse	S23	1
ImagingStudy	S23	1
Substance	S39	1
Account, Library, Claim, DocumentManifest, EnrollmentResponse, CoverageEligibilityRequest, SupplyDelivery, PaymentNotice, OperationDefinition, GuidanceResponse, ImplementationGuide, CapabilityStatement,	S69	1

XSL•FO
RenderX

Figure 4. Distribution of the resources used.

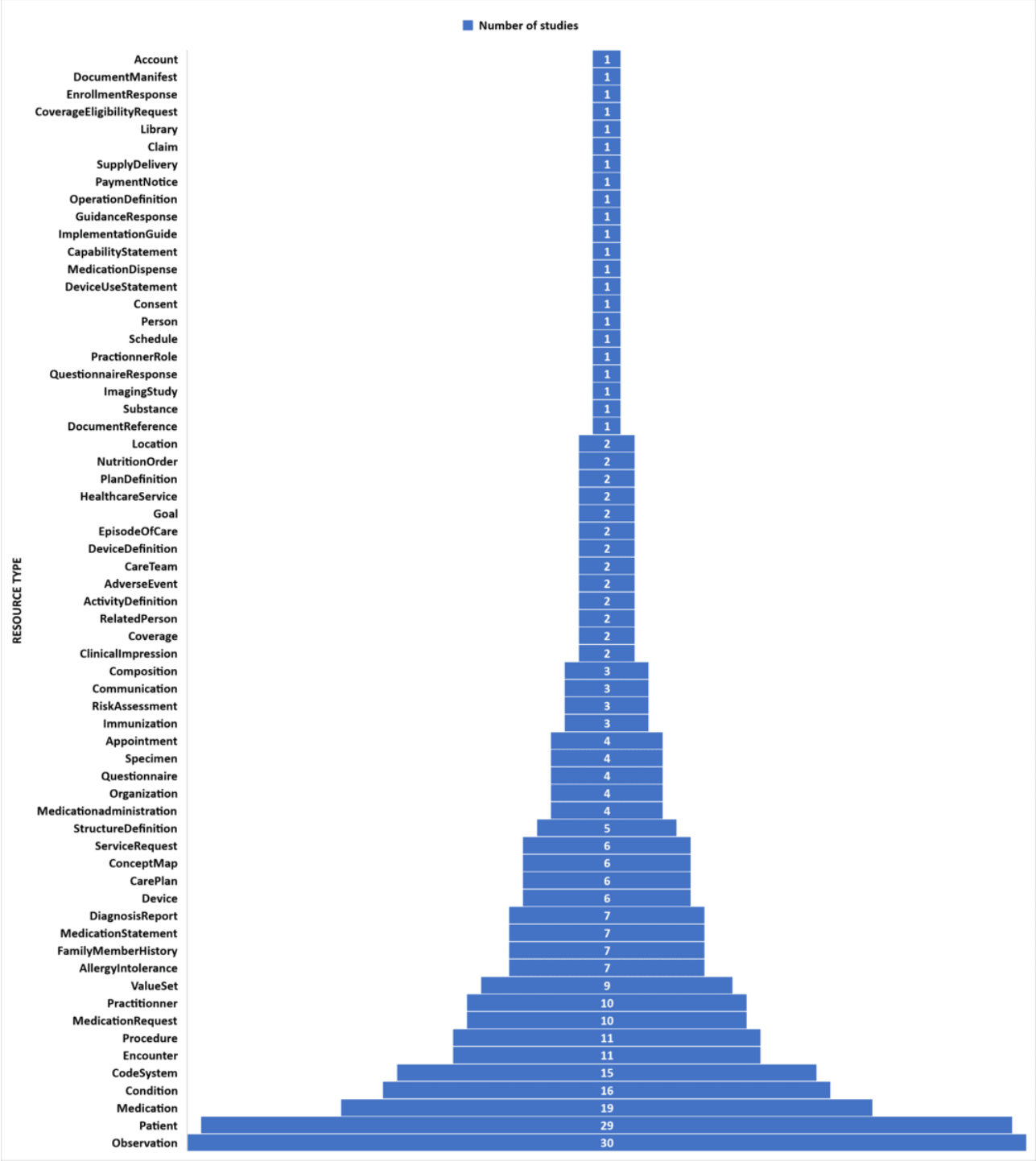

XSL•FO
RenderX

Figure 5. Distribution of the Fast Healthcare Interoperability Resources types used by category.

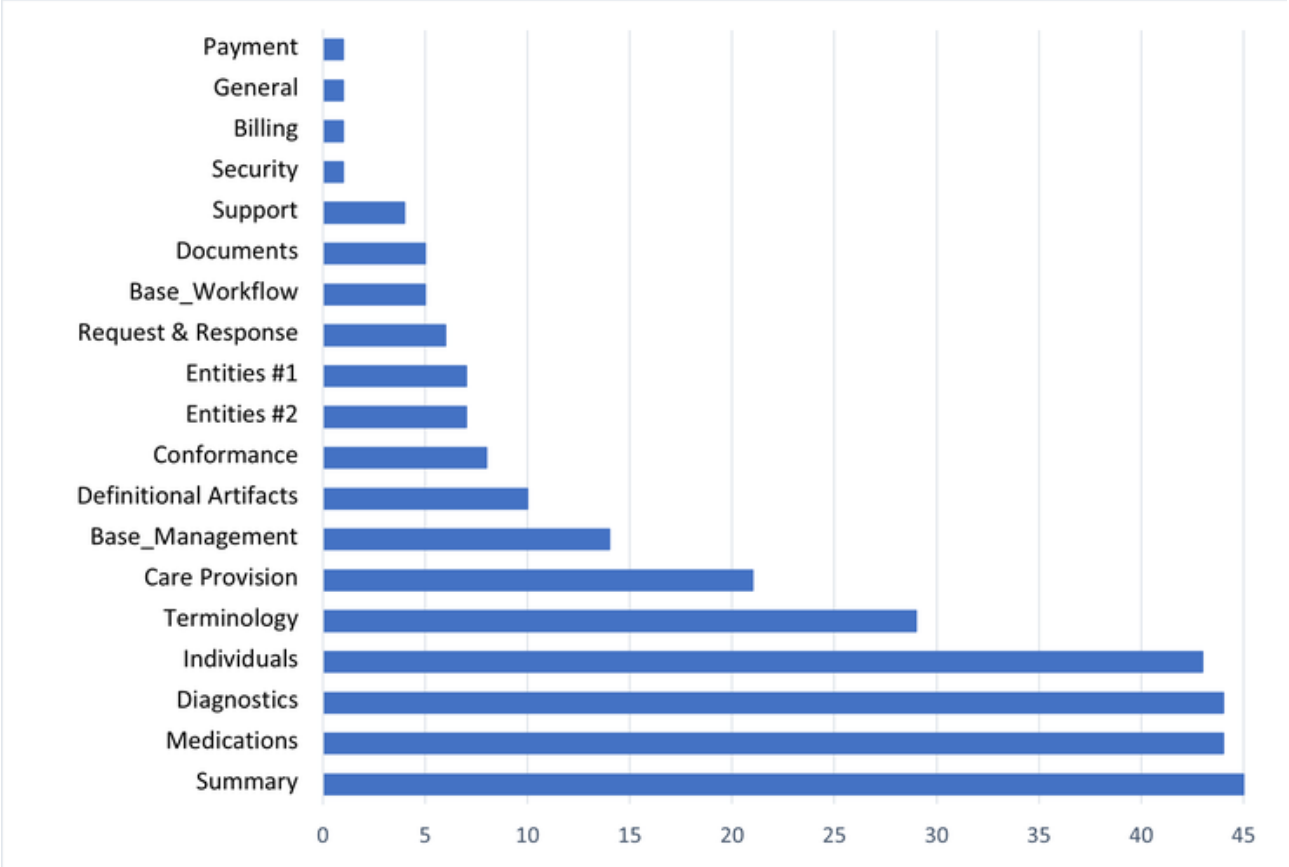

RQ2: Which FHIR Terminologies Were Used

A total of 9 distinct FHIR terminologies were used in the 70 selected studies (Table 5).

Their distribution and percentage are as follows:

- SNOMED CT (31/98, 32%) and LOINC (25/98, 26%) were the most commonly used.
- ICD-9 and ICD-10 (15/98, 15%) and Unified Medical Language System (13/98, 13%)
- RxNORM (9/98, 9%)
- The remaining terminologies (anatomical therapeutic chemical, quality data model, national drug file–reference terminology, and classification of instructional programs code) accounted for <10%.

XSL•FO
RenderX

Table 5. List of terminologies being used.

Terminology	Study ID
SNOMED CT[a]	S2, S34, S3, S35, S4, S48, S5, S49, S7, S8, S9, S29, S10, S11, S12, S40, S28, S53, S14, S16, S41, S31, S38, S18, S66, S22, S61, S23, S13, S25, S64
LOINC[b]	S2, S34, S3, S35, S5, S6, S7, S8, S27, S9, S11, S13, S16, S54, S30, S56, S41, S38, S20, S66, S21, S22, S23, S24, S36
ICD[c]-9 or ICD-10	S34, S1, S35, S49, S6, S9, S54, S30, S41, S17, S18, S66, S22, S33, S26
UMLS[d]	S34, S43, S37, S5, S27, S14, S19, S66, S21, S61, S64, S25, S8
RxNorm[e]	S36, S6, S8, S9, S14, S54, S31, S22, S61
ATC[f]	S1, S8
QDM[g]	S39
NDF-RT[h]	S48
CIP[i] Code	S1

[a]SNOMED CT: Systematized Nomenclature of Medicine–clinical terms.
[b]LOINC: Logical Observation Identifiers Names and Codes.
[c]ICD: International Classification of Diseases.
[d]UMLS: Unified Medical Language System.
[e]RxNORM: medical prescription normalized Medical prescription.
[f]ATC: anatomical therapeutic chemical.
[g]QDM: quality data model.
[h]NDF-RT: national drug file–reference terminology.
[i]CIP: classification of instructional programs.

RQ3: Semantic Research Approaches Most Frequently Applied

A total of six main semantic research approaches were identified from the solutions proposed in the 70 selected studies:

1. Mapping to FHIR terminology (mapping)
2. Terminology services (terminology)
3. Transformation, development, and validation resource description framework (RDF) or web ontology language (OWL)
4. Machine learning (ML) and natural language processing (NLP)
5. Semantic annotation (annotation)
6. Ontology-based approach (ontology)

Table 6 presents the semantic research approaches used to decrease the frequency of use in the selected studies.

Table 6. Semantic research approaches and related studies.

Semantic approach	Studies IDs	Distribution (N=126), n (%)
Mapping to Fast Healthcare Interoperability Resources terminologies (mapping)	S1, S2, S4, S5, S6, S7, S8, S9, S10, S11, S13, S14, S15, S16, S17, S18, S20, S21, S22, S23, S24, S25, S27, S28, S29, S30, S31, S32, S33, S38, S41	31 (24.6)
Terminology services (terminology)	S5, S13, S17, S18, S19, S20, S21, S28, S30, S33, S34, S35, S36, S37, S38, S39, S40, S41	18 (14.3)
Transformation or development or validation resource description framework or web ontology language	S3, S10, S12, S14, S35, S42, S43, S44, S45, S46, S47, S48, S49, S50, S51, S52, S53, S54, S55, S56, S57, S58, S59, S69	24 (19)
Semantic annotation (annotation)	S2, S3, S6, S7, S14, S25, S27, S30, S31, S32, S43, S45, S48, S60, S61, S62, S63, S64	18 (14.3)
Machine learning or natural language processing	S5, S6, S8, S14, S25, S26, S31, S32, S43, S46, S49, S54, S60, S61, S62, S63, S64, S65, S66, S67	20 (15.9)
Ontology-based approach (ontology)	S3, S11, S12, S13, S26, S40, S44, S46, S47, S49, S50, S55, S66, S68, S70	15 (11.9)

The distribution of the 6 semantic approaches is as follow: mapping (31/126, 24.6%), RDF or OWL (24/126, 19%), ML and NLP (20/126, 15.9%), annotation (18/126, 14.3%), terminology (18/126, 14.3%), and ontology (15/126, 11.9%).

XSL•FO
RenderX

The selected studies used at least one of the 6 identified semantic research approaches, and most studies used more than one research approach. For example, 22 of the 31 studies listed in Table 6 used mapping and other approaches (70%), whereas only 29% (9/31) used mapping alone.

To better understand the frequency of combinations of more than one semantic research approach, we identified a number of studies using more than one approach. We performed the same exercise for RDF or OWL followed by annotation. Table 7 shows that most combined semantic approaches are annotation and ML and NLP, mapping and terminology services, mapping and annotation, RDF or OWL, and ontology.

Table 7. Distribution of the combined approaches

Category	Mapping	Terminology	RDF[a] or OWL[b]	Annotation	ML[c] and NLP[d]	Ontology
Mapping	__[e]	11	2	9	7	2
Terminology	—	—	0	1	1	2
RDF and OWL	—	—	—	5	5	8
Annotation	—	—	—	—	11	1
ML and NLP	—	—	—	—	—	4
Ontology	—	—	—	—	—	—

[a]RDF: resource description framework.

[b]OWL: web ontology language.

[c]ML: machine learning.

[d]NLP: natural language processing.

[e]Not applicable.

The next analysis related to the proposed approach examined how these approaches evolved over time. Figure 6 shows how these approaches were used over time. The main finding are (1) ML and NLP-based proposals emerged in 2018 in FHIR semantic interoperability studies, (2) the other approaches span the full period, and (3) no specific research approach dominates the FHIR semantic interoperability research field.

Figure 6. Timeline of semantic research approaches used in the selected studies. ML: machine learning; NLP: natural language processing; RDF: resource description framework.

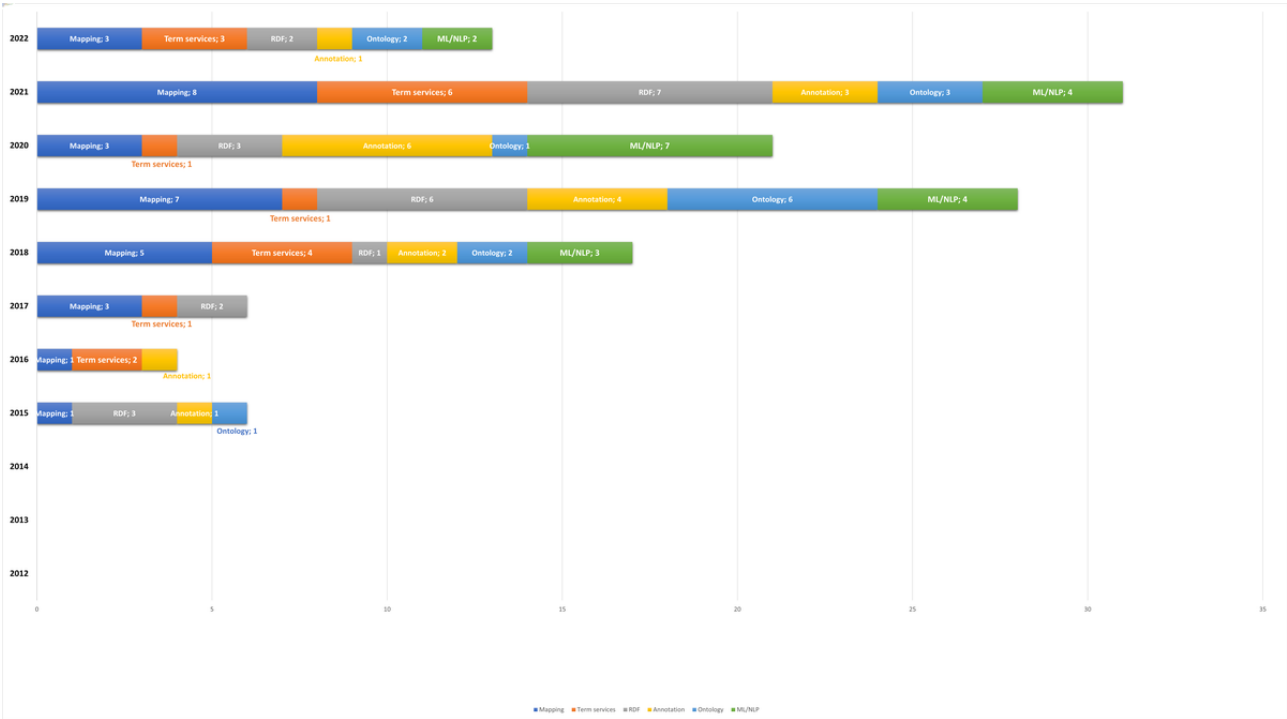

RQ4: Main Types of Contributions

The objective of RQ4 was to identify the contribution types within each of the 70 selected studies. For this purpose, five main contributions were identified: (1) new framework or architecture intended to enable semantic interoperability (framework or architecture proposal), (2) evaluation or comparison of different techniques (eg, comparison of ontologies), (3) new ontology, model or enhancement of an existing one (model proposal), (4) a new technique or

XSL•FO
RenderX

improvement of an existing technique (new technique proposal), and (5) a tool proposal through the development of a new tool or an enhancement to existing ones to facilitate achieving semantic interoperability (tool proposal).

Table 8 presents the distribution of the 70 studies selected according to contribution type.

Figure 7 shows the distribution of studies by contribution type.

The most frequent contribution type was the proposal of a framework or architecture (26/70, 37%), followed by model proposals (13/70, 19%), new or improved techniques (13/70, 19%), comparison services (10/70, 14%), and tool proposals (8/70, 11%).

Table 8. Studies by contribution type.

Contribution type	Studies
Framework or architecture	S2, S5, S7, S8, S14, S16, S20, S22, S23, S25, S26, S29, S31, S34, S35, S38, S41, S43, S51, S53, S54, S61, S62, S65, S66, S68
Comparison	S4, S15, S17, S32, S33, S46, S56, S58, S59, S67
Model	S1, S3, S6, S30, S39, S40, S45, S48, S49, S50, S55, S60, S63
Technique	S9, S10, S12, S18, S19, S21, S24, S27, S42, S47, S64, S69, S70
Tool	S11, S13, S28, S36, S37, S44, S52, S57

Figure 7. Distribution of the studies by contribution type.

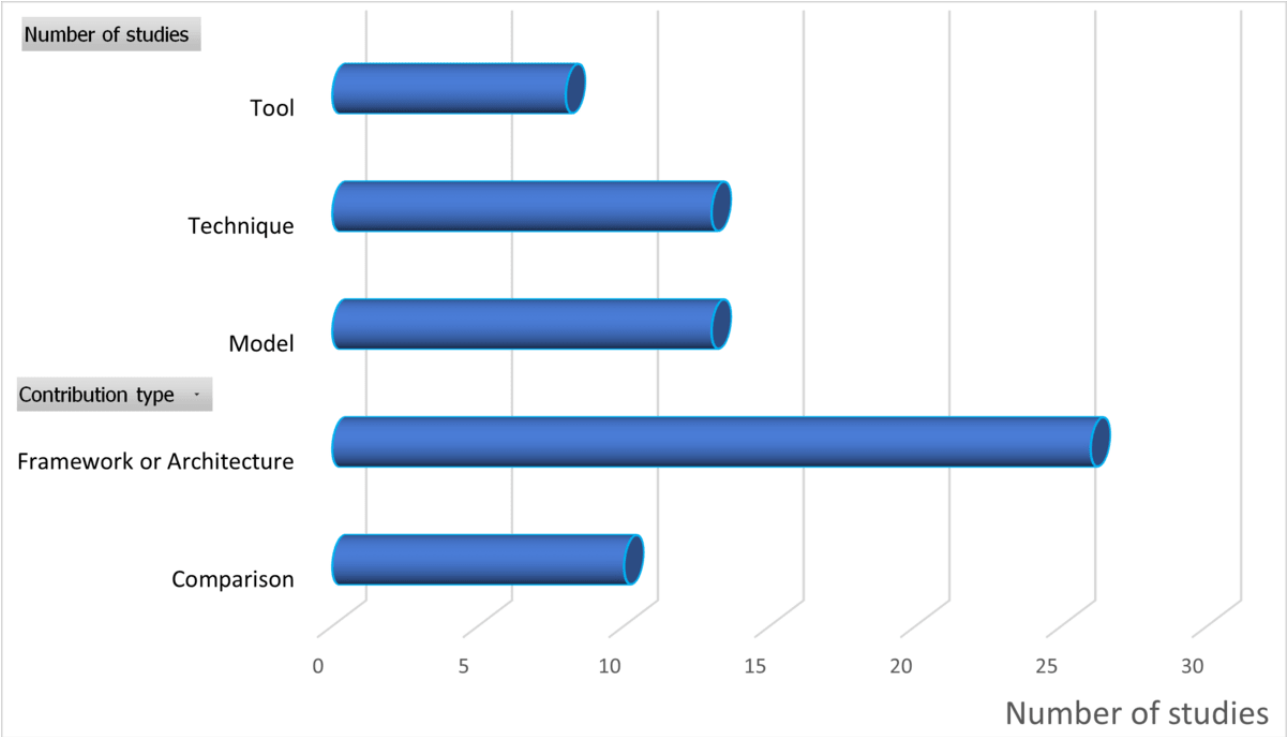

RQ5: Publications Sources and Types

Publication Sources

The distribution of publication sources and the number of studies is summarized as follows (Figure 8):

- The top publication sources were *Studies in Health Technology and Informatics* (8/70, 11%), followed by *CEUR Workshop* and the *Journal of Biomedical Informatics* (6/70, 9%).
- *JMIR Medical Informatics* and *BMC Medical Informatics and Decision Making* follow with (4/70, 6%).
- Approximately (3/70, 4%) of the publications are from 3 sources: *Journal of Biomedical Semantic*, American Medical Informatics Association (AMIA) Joint Summits on Translational Science proceedings, and AMIA Annual Symposium.
- Approximately (2/70, 3%) of the publications are from 4 sources: the *Journal of Medical Systems* , IEEE International Conference on Bioinformatics and Biomedicine, International Conference of the IEEE Engineering in Medicine and Biology Society, and the *Journal of the AMIA*.
- Each of the remaining 25 publications represented 1%, with 1 study per source.

XSL•FO
RenderX

Figure 8. Publication sources and corresponding number of studies.

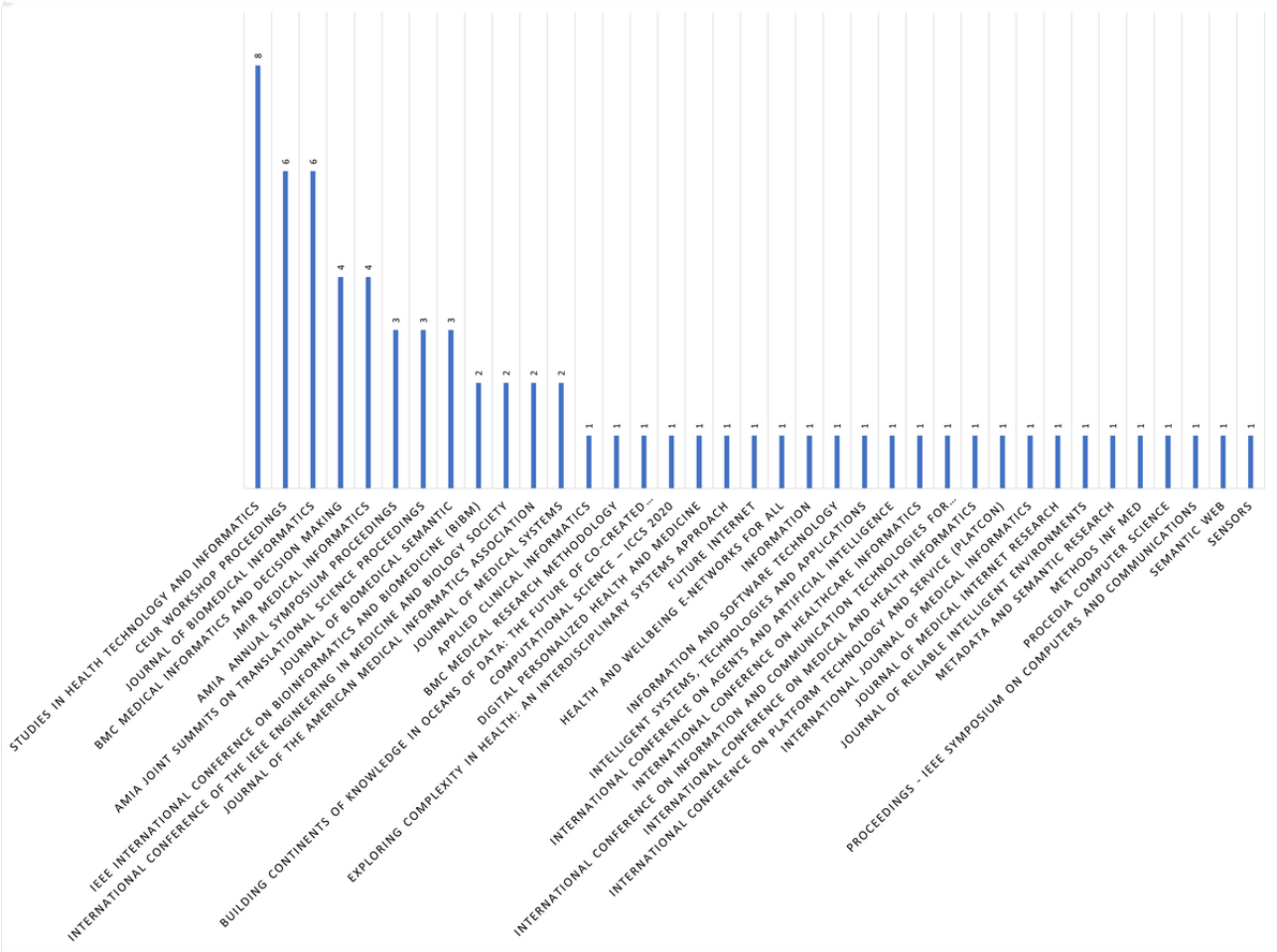

Publication Types

The distribution of the selected studies by venue type (Table 9) showed that most of selected studies were published in a journal (46/70, 66%), followed by conferences (15/70, 21%), workshops (6/70, 9%), and symposiums (3/70, 4%).

Table 9. Number of studies per venue type (N=70).

Venue type	Studies, n (%)
Journal	46 (66)
Conference	15 (21)
Workshop	6 (9)
Symposium	3 (4)

Discussion

Overview

This section presents a discussion of each of the key findings of the RQs.

1. Most studies have concentrated on clinical data resources; therefore, the focus has been on patient summaries.
2. Terminology mapping to SNOMED CT and LOINC was performed in more than half of the selected studies. None of the 5 approaches was dominant in the proposed solutions to semantic interoperability challenges.
3. Most of the selected studies proposed a framework or architecture as a solution to enable semantic interoperability.
4. In recent years, research on semantic interoperability has increased since 2018.

RQ Results

RQ1: What Are the Main Resource Types Studied or Used?

As most studies concentrated on clinical data resources (eg, Observation, Medication, Condition, Procedure, and MedicationRequesL), the focus has been on patient summary including the medical condition, medication, and diagnosis.

The attributes of these resource-use codes (CodeableConcept) and terminology standards, such as SNOMED CT and LOINC, are already included in the FHIR body. For example, the

XSL•FO
RenderX

Observation resource has an attribute code in which the practitioner or user can document the observation type using LOINC code. Similarly, the Medication resource prescribed can be entered either as a code using SNOMED CT or as a textual description.

In contrast, other types of resources (eg, BodyOfStructure) were not addressed in the selected studies.

RQ2: Which Terminologies Were Used?

Terminology mapping to SNOMED CT and LOINC was present in more than half of the selected studies, where SNOMED CT is used mainly for medication, procedure, and condition (problem); SNOMED CT or ICD-10 for diseases and diagnosis; and LOINC for observation and diagnostic report.

Some countries have attempted to develop a consensus on the terminology to be used at the regional or national level. For example, SNOMED CT was adopted as a single terminology in the United Kingdom by the National Health Service (S12). In Switzerland, the federal initiative of the Swiss Personalized Health Network created a semantic framework available for health communities aiming to use the existing standards instead of building a new one, and the semantics were mapped by domain experts on the existing standards, such as SNOMED CT, LOINC, and ICD. RDF is used to store and transport data. In another regional initiative in Italy, the LOINC vocabulary was used to represent concepts already defined in local and standardized terminologies (S20).

RQ3: What Are the Most Frequently Applied Research Approaches to Achieve Semantic Interoperability?

None of the 5 approaches were dominant in the proposed solutions to semantic interoperability challenges. The emerging ML and NLP approach is mostly applied to annotation.

ML and NLP techniques, which have achieved significant penetration in many other information system or information technology domains, have also been used in FHIR semantic interoperability research. On the one hand, the capability of the ML and NLP techniques for automation and prediction is often required for complex use cases. On the other hand, RDF representation has the specific advantage of containing information about semantic links and dependencies between multiple resources. This could explain why transformation to or from RDF is useful for semantic interoperability.

RQ4: What Are the Main Types of Contributions Within the Selected Studies?

Most of the selected studies proposed a framework or architecture as a solution to enable semantic interoperability. Research on achieving semantic interoperability using FHIR has tackled different steps and activities: collecting, extracting, and annotating data; modeling data from legacy systems; and applying transformation and mapping to FHIR models and terminologies. The following research topics were identified.

- One of the main topics in the selected studies was related to unstructured data because valuable information can be extracted from free text for research and health analytics. The standardization of unstructured clinical notes was the focus of the research for specific use cases in the selected studies. The process of concept extraction and mapping text to standardized terminologies or ontologies requires more automation.
- Research on unstructured data using annotation and NLP is a relatively new field for the FHIR community, and challenges related to protecting privacy and security (deidentification and pseudonymization) remain open questions.
- The manual mapping of clinical terminology requires significant effort. The use of a tool can reduce the effort required for mapping and increase the adoption of terminology.
- Another challenge is related to languages other than English. In S24, a tool was developed to map Russian local terminology to LOINC.
- Even though the FHIR standard allows the use of clinical terminologies such as SNOMED CT and LOINC, users still enter information as text or use local terminologies. For example, selecting an appropriate concept from SNOMED CT is not a trivial task. To this end, a proposed solution called Ontoserver (eg, a clinical terminology service) offers a prefix-based search algorithm to help users easily find content and enter coded data. Ontoserver was used in several studies (S5, S10, S11, S17, S18, S40, and S41), and limitations related to its performance were reported.
- Comparison and mapping ontologies: detecting similarities and differences is still of interest and may attract more researchers.

RQ5: What Are the Publication Sources and Types?

Research on FHIR semantic interoperability emerged in 2015 and has increased since 2018. This probably is linked to the first release HL7 FHIR with a “Normative” status in 2018 indicating that its content has been recognized at the international level as stable for use and has been “locked” for several years. The publication trend in Figure 9 that shows the distribution of the selected studies over time confirms that research on this topic is still active.

XSL•FO
RenderX

Figure 9. Number of studies over time.

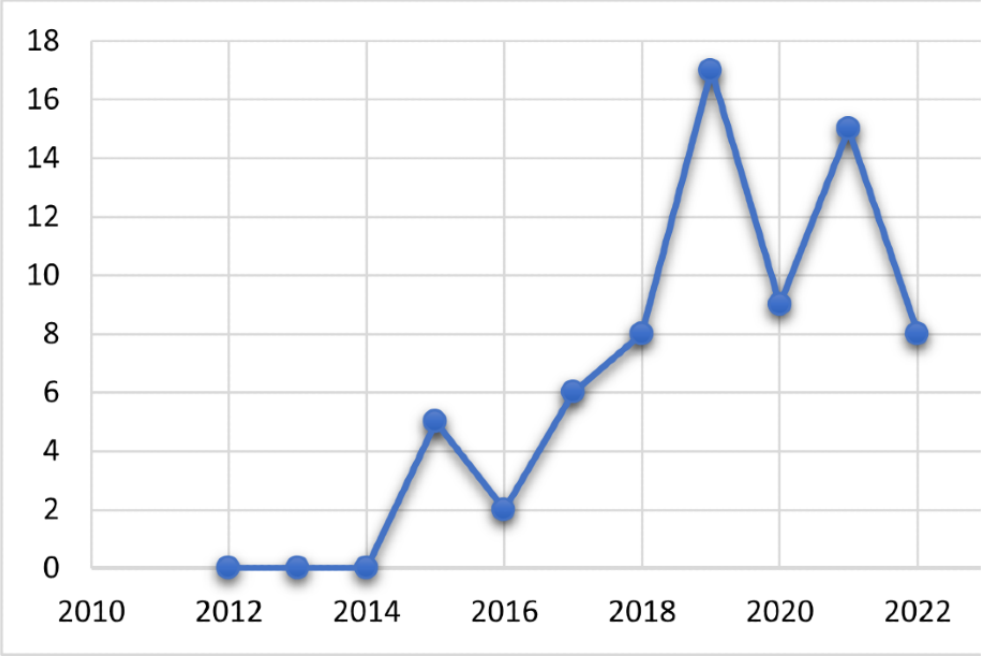

Assumptions and Limitations

This study was limited to semantic studies using only the FHIR standard, and it did not search for other studies related to semantic interoperability in general; therefore, this SMR focused exclusively on studies addressing FHIR implementation.

Furthermore, only studies from the 10 database sources specified were considered, and commercial solutions were excluded.

During the classification of the contribution based on 5 categories, 2 studies (S40 and S49) were considered in 2 different categories (model and technique). Both situations were evaluated and the results were only slightly different.

Conclusions

This SMR inventoried and analyzed the findings on semantic interoperability using FHIR over the last decade (2012-2022) in 70 selected studies from 10 databases. It presents a classification of the selected studies according to the semantic approaches used and their contributions. It also provides the frequency of use of FHIR resources and terminologies.

These findings provide researchers with an inventory of the approaches used to achieve semantic interoperability using the FHIR. The SMR also documented the frequency with which the proposed approaches were used. This may help identify avenues for research on semantic interoperability. It can also be used as a guideline by the FHIR community for the future development of approaches and types of resources that are not well covered in research published to date.

Indeed, understanding and interpreting the exchanged data in the EHR ecosystem needs to be supported by further research using, for instance, FHIR in mapping, RDF or OWL, and comparing ontologies and terminology services.

Semantic interoperability is still an active research field, and there is no broad consensus in the health care community on how to achieve full semantic interoperability between information systems. The main issues reported are related to the difficulty associated with the use of different terminologies and the efforts required to successfully map to FHIR terminology. Even worse, when no terminology is used, unstructured data would still need to be extracted and mapped to the appropriate concept code for FHIR terminology.

This SMR provides a classification of the different solutions proposed to address semantic interoperability using FHIR at different steps: collecting, extracting and annotating data, modeling EHR data from legacy systems, and applying transformation and mapping to FHIR models and terminologies. The use of ML and NLP for unstructured data is promising and has been applied to specific use case scenarios. In addition, terminology services are needed to accelerate their use and adoption; furthermore, techniques and tools to automate annotation and ontology comparison should help reduce human interaction.

Future work will be devoted to the FHIR-based framework for the semantic interoperability of EHR data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Targeted resources from a semantic point of view.

XSL•FO
RenderX

[PNG File , 526 KB - jmir_v26i1e45209_app1.png]

Multimedia Appendix 2
List of the selected studies (N=70).
[PDF File (Adobe PDF File), 152 KB - jmir_v26i1e45209_app2.pdf]

Multimedia Appendix 3
Extracted data and searching strategy results.
[XLSX File (Microsoft Excel File), 134 KB - jmir_v26i1e45209_app3.xlsx]

References

1. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, et al. Use of EHRs data for clinical research: historical progress and current applications. Learn Health Syst 2019 Jan 16;3(1):e10076 [FREE Full text] [doi: 10.1002/lrh2.10076] [Medline: 31245598]
2. Setyawan R, Hidayanto AN, Sensuse DI, Kautsarina, Suryono RR, Abilowo K. Data integration and interoperability problems of HL7 FHIR implementation and potential solutions: a systematic literature review. In: Proceedings of the 5th International Conference on Informatics and Computational Sciences. 2021 Presented at: ICICoS '21; November 24-25, 2021; Semarang, Indonesia p. 293-298 URL: https://ieeexplore.ieee.org/document/9651762/ [doi: 10.1109/icicos53627.2021.9651762]
3. IEEE standard glossary of software engineering terminology. IEEE Std 610.12-1990. 1990 Dec 31. URL: https://ieeexplore.ieee.org/document/159342 [accessed 2024-01-04]
4. Interoperability in healthcare. Healthcare Information and Management Systems Society. URL: https://www.himss.org/resources/interoperability-healthcare [accessed 2024-01-04]
5. Bender D, Sartipi K. HL7 FHIR: an agileRESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013 Presented at: CBMS '13; June 20-22, 2013; Porto, Portugal p. 326-331 URL: https://ieeexplore.ieee.org/document/6627810 [doi: 10.1109/cbms.2013.6627810]
6. Benson T, Grieve G. Principles of Health Interoperability: FHIR, HL7 and SNOMED CT. Cham, Switzerland: Springer; 2021.
7. Dullabh P, Hovey L, Heaney-Huls K, Rajendran N, Wright A, Sittig DF. Application programming interfaces in health care: findings from a current-state sociotechnical assessment. Appl Clin Inform 2020 Jan 09;11(1):59-69 [FREE Full text] [doi: 10.1055/s-0039-1701001] [Medline: 31968383]
8. Cloud providers unite for healthcare interoperability. HL7 International. URL: https://blog.hl7.org/cloud-providers-unite-for-healthcare-interoperability-fhir [accessed 2022-12-11]
9. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. Stud Health Technol Inform 2019 Sep 03;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
10. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
11. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
12. Kitchenham BA, Budgen D, Pearl Brereton O. Using mapping studies as the basis for further research – a participant-observer case study. Inf Softw Technol 2011 Jun;53(6):638-651 [FREE Full text] [doi: 10.1016/j.infsof.2010.12.011]
13. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 2015 Aug;64:1-18 [FREE Full text] [doi: 10.1016/j.infsof.2015.03.007]
14. Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M. Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 2007 Apr;80(4):571-583 [FREE Full text] [doi: 10.1016/j.jss.2006.07.009]
15. Reference management tool. Zotero. URL: https://www.zotero.org/ [accessed 2022-12-12]
16. Boussadi A, Zapletal E. A fast healthcare interoperability resources (FHIR) layer implemented over i2b2. BMC Med Inform Decis Mak 2017 Aug 14;17(1):120 [FREE Full text] [doi: 10.1186/s12911-017-0513-6] [Medline: 28806953]
17. Zampognaro P, Paragliola G, Falanga V. A FHIR based architecture of a multiprotocol IoT home gateway supporting dynamic plug of new devices within instrumented environments. In: Proceedings of the 2021 IEEE Symposium on Computers and Communications. 2021 Presented at: ISCC '21; September 5-8, 2021; Athens, Greece p. 1-7 URL: https://ieeexplore.ieee.org/document/9631446 [doi: 10.1109/iscc53001.2021.9631446]
18. Kilintzis V, Kosvyra A, Beredimas N, Natsiavas P, Maglaveras N, Chouvarda I. A sustainable HL7 FHIR based ontology for PHR data*. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and

XSL•FO
RenderX

Biology Society. 2019 Presented at: EMBC '19; July 23-27, 2019; Berlin, Germany p. 5700-5703 URL: https://ieeexplore.ieee.org/document/8856415/similar [doi: 10.1109/embc.2019.8856415]
19. Goeg KR, Hummeluhr M. An empirical approach to enhancing terminology binding - an HL7 FHIR SNOMED CT example. In: Ugon A, Karlsson D, Klein GO, Moen A, editors. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created Ehealth. Amsterdam, The Netherlands: IOS Press; 2018.
20. Banach A, Ulrich H, Kroll B, Kiel A, Ingenerf J, Kock-Schoppenhauer AK. APERITIF - automatic patient recruiting for clinical trials based on HL7 FHIR. Stud Health Technol Inform 2021 May 27;281:58-62. [doi: 10.3233/SHTI210120] [Medline: 34042705]
21. Daumke P, Heitmann KU, Heckmann S, Martínez-Costa C, Schulz S. Clinical text mining on FHIR. Stud Health Technol Inform 2019 Aug 21;264:83-87. [doi: 10.3233/SHTI190188] [Medline: 31437890]
22. Zampognaro P, Paragliola G, Falanga V. Definition of an FHIR-based multiprotocol IoT home gateway to support the dynamic plug of new devices within instrumented environments. J Reliable Intell Environ 2021 Dec 07;8(4):319-331 [FREE Full text] [doi: 10.1007/s40860-021-00161-2]
23. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen LV, et al. Developing a FHIR-based EHR phenotyping framework: a case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform 2019 Nov;99:103310 [FREE Full text] [doi: 10.1016/j.jbi.2019.103310] [Medline: 31622801]
24. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. J Biomed Inform 2021 Oct;122:103891 [FREE Full text] [doi: 10.1016/j.jbi.2021.103891] [Medline: 34450285]
25. Metke-Jimenez A, Lawley M, Hansen D. FHIR OWL: transforming OWL ontologies into FHIR terminology resources. AMIA Annu Symp Proc 2019;2019:664-672 [FREE Full text] [Medline: 32308861]
26. Metke-Jimenez A, Hansen D. FHIRCap: transforming REDCap forms into FHIR resources. AMIA Jt Summits Transl Sci Proc 2019;2019:54-63 [FREE Full text] [Medline: 31258956]
27. Arguello-Casteleiro M, Martinez-Costa C, Des-Diz J, Maroto N, Fernandez-Prieto MJ, Stevens R. From SNOMED CT expressions to an FHIR RDF representation: exploring the benefits of an ontology-based approach. In: Proceedings of the Joint Ontology Workshops 2019 Episode V: The Styrian Autumn of Ontology. 2019 Presented at: CEUR '19; September 23-25, 2019; Graz, Austria p. 1-13 URL: https://ceur-ws.org/Vol-2518/paper-ODLS1.pdf
28. Paris N, Mendis M, Daniel C, Murphy S, Tannier X, Zweigenbaum P. i2b2 implemented over SMART-on-FHIR. AMIA Jt Summits Transl Sci Proc 2018;2017:369-378 [FREE Full text] [Medline: 29888095]
29. Hong N, Wen A, Shen F, Sohn S, Liu S, Liu H, et al. Integrating structured and unstructured EHR data using an FHIR-based type system: a case study with medication data. AMIA Jt Summits Transl Sci Proc 2018;2017:74-83. [Medline: 29888045]
30. Lenivtseva Y, Kopanitsa G. Investigation of Content Overlap in Proprietary Medical Mappings. Stud Health Technol Inform 2019;258:41-45. [Medline: 30942710]
31. Matney SA, Heale B, Hasley S, Decker E, Frederiksen B, Davis N, et al. Lessons learned in creating interoperable fast healthcare interoperability resources profiles for large-scale public health programs. Appl Clin Inform 2019 Jan;10(1):87-95 [FREE Full text] [doi: 10.1055/s-0038-1677527] [Medline: 30727002]
32. Rajput AM, Drenkhahn C. Perspectives and obstacles for transforming terminologies into FHIR CodeSystems exemplified by Alpha-ID. Stud Health Technol Inform 2021 May 27;281:213-217. [doi: 10.3233/SHTI210151] [Medline: 34042736]
33. Wiedekopf J, Drenkhahn C, Ulrich H, Kock-Schoppenhauer AK, Ingenerf J. Providing ART-DECOR valuesets via FHIR terminology servers - a technical report. Stud Health Technol Inform 2021 Sep 21;283:127-135. [doi: 10.3233/SHTI210550] [Medline: 34545828]
34. Saripalle R. Representing UMLS knowledge using FHIR terminological resources. In: Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine. 2019 Presented at: BIBM '19; November 18-21, 2019; San Diego, CA p. 1109-1112 URL: https://ieeexplore.ieee.org/document/8983305 [doi: 10.1109/bibm47256.2019.8983305]
35. Gazzarata R, Maggi N, Magnoni LD, Monteverde ME, Ruggiero C, Giacomini M. Semantics management for a regional health information system in Italy by CTS2 and FHIR. In: Gazzarata R, Maggi N, Magnoni LD, Monteverde ME, Ruggiero C, Giacomini M, editors. Studies in Health Technology and Informatics. Amsterdam, The Netherlands: IOS Press; 2021.
36. Rajput AM, Gronke A, Johannis W. Semi-automated approach to validate and enrich LOINC codes by FHIR server. In: Proceedings of the 5th International Conference on Medical and Health Informatics. 2021 Presented at: ICMHI '21; May 14-16, 2021; Kyoto, Japan URL: https://dl.acm.org/doi/10.1145/3472813.3473207 [doi: 10.1145/3472813.3473207]
37. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, et al. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc 2017 Mar 01;24(2):398-402 [FREE Full text] [doi: 10.1093/jamia/ocw079] [Medline: 27274012]
38. Leroux H, Metke-Jimenez A, Lawley MJ. Towards achieving semantic interoperability of clinical study data with FHIR. J Biomed Semantics 2017 Sep 19;8(1):41 [FREE Full text] [doi: 10.1186/s13326-017-0148-7] [Medline: 28927443]
39. Kopanitsa G, Taranik M. Using semi-automated approach for mapping local Russian laboratory terms to LOINC. In: Proceedings of the 21st International Conference on Information and Software Technologies. 2015 Presented at: ICIST '15; October 15-16, 2015; Druskininkai, Lithuania p. 37-44 URL: https://link.springer.com/chapter/10.1007/978-3-319-24770-0_4 [doi: 10.1007/978-3-319-24770-0_4]

XSL•FO
RenderX

40. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. J Biomed Inform 2020 Oct;110:103541 [FREE Full text] [doi: 10.1016/j.jbi.2020.103541] [Medline: 32814201]
41. Maga-Nteve C, Kontopoulos E, Tsolakis N, Katakis I, Mathioudis E, Mitzias P, et al. A semantic technologies toolkit for bridging early diagnosis and treatment in brain diseases: report from the ongoing EU-funded research project ALAMEDA. In: Proceedings of the 15th International Conference on Metadata and Semantic Research. 2021 Presented at: MTSR '21; November 29-December 3, 2021; Virtual Event p. 349-354 URL: https://link.springer.com/chapter/10.1007/978-3-030-98876-0_30 [doi: 10.1007/978-3-030-98876-0_30]
42. Rafee A, Riepenhausen S, Neuhaus P, Meidt A, Dugas M, Varghese J. ELaPro, a LOINC-mapped core dataset for top laboratory procedures of eligibility screening for clinical trials. BMC Med Res Methodol 2022 May 14;22(1):141 [FREE Full text] [doi: 10.1186/s12874-022-01611-y] [Medline: 35568796]
43. Chatterjee A, Pahari N, Prinz A. HL7 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: a proof-of-concept study. Sensors (Basel) 2022 May 15;22(10):3756 [FREE Full text] [doi: 10.3390/s22103756] [Medline: 35632165]
44. Odigie E, Lacson R, Raja A, Osterbur D, Ip I, Schneider L, et al. Fast healthcare interoperability resources, clinical quality language, and systematized nomenclature of medicine-clinical terms in representing clinical evidence logic statements for the use of imaging procedures: descriptive study. JMIR Med Inform 2019 May 13;7(2):e13590 [FREE Full text] [doi: 10.2196/13590] [Medline: 31094359]
45. Ulrich H, Kock AK, Duhm-Harbeck P, Habermann JK, Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR. Stud Health Technol Inform 2016;228:162-166. [Medline: 27577363]
46. Wang J, Mathews WC, Pham HA, Xu H, Zhang Y. Opioid2FHIR: a system for extracting FHIR-compatible opioid prescriptions from clinical text. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine. 2020 Presented at: BIBM '20; December 16-19, 2020; Seoul, Korea p. 1748-1751 URL: https://ieeexplore.ieee.org/document/9313258 [doi: 10.1109/bibm49941.2020.9313258]
47. Seveso A, Campagner A, Ciucci D, Cabitza F. Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings. BMC Med Inform Decis Mak 2020 Aug 20;20(Suppl 5):142 [FREE Full text] [doi: 10.1186/s12911-020-01152-8] [Medline: 32819345]
48. Wiedekopf J, Drenkhahn C, Rosenau L, Ulrich H, Kock-Schoppenhauer A, Ingenerf J. TerminoDiff - detecting semantic differences in HL7 FHIR CodeSystems. Stud Health Technol Inform 2022 May 25;294:362-366. [doi: 10.3233/SHTI220475] [Medline: 35612097]
49. Diomaiuta C, Sicuranza M, Ciampi M, De Pietro G. A FHIR-based system for the generation and retrieval of clinical documents. In: Proceedings of the 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health. 2017 Presented at: ICT4AWE '17; April 28-29, 2017; Porto, Portugal p. 135-142 URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0006311301350142 [doi: 10.5220/0006311301350142]
50. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the Swiss personalized health network: methodological study. JMIR Med Inform 2021 Jun 24;9(6):e27591 [FREE Full text] [doi: 10.2196/27591] [Medline: 34185008]
51. Dolin RH, Boxwala A, Shalaby J. A pharmacogenomics clinical decision support service based on FHIR and CDS hooks. Methods Inf Med 2018 Dec;57(S 02):e115-e123 [FREE Full text] [doi: 10.1055/s-0038-1676466] [Medline: 30605914]
52. Saripalle R, Sookhak M, Haghparast M. An interoperable UMLS terminology service using FHIR. Future Internet 2020 Nov 16;12(11):199 [FREE Full text] [doi: 10.3390/fi12110199]
53. Semenov I, Kopanitsa G, Denisov D, Alexandr Y, Osenev R, Andreychuk Y. Patients decision aid system based on FHIR profiles. J Med Syst 2018 Jul 31;42(9):166. [doi: 10.1007/s10916-018-1016-4] [Medline: 30066031]
54. Jiang G, Kiefer RC, Rasmussen LV, Solbrig HR, Mo H, Pacheco JA, et al. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. J Biomed Inform 2016 Aug;62:232-242 [FREE Full text] [doi: 10.1016/j.jbi.2016.07.008] [Medline: 27392645]
55. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, et al. Generation of a fast healthcare interoperability resources (FHIR)-based ontology for federated feasibility queries in the context of COVID-19: feasibility study. JMIR Med Inform 2022 Apr 27;10(4):e35789 [FREE Full text] [doi: 10.2196/35789] [Medline: 35380548]
56. Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semantics 2018 Sep 17;9(1):24 [FREE Full text] [doi: 10.1186/s13326-018-0191-z] [Medline: 30223897]
57. Kober G, Paschke A. A FHIR-to-RDF converter. In: Proceeding of the 12th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. 2019 Presented at: SWAT4HCLS '19; December 11-12, 2019; Edinburgh, UK p. 1-6 URL: https://ceur-ws.org/Vol-2849/paper-03.pdf
58. Dridi A, Sassi S, Chbeir R, Faiz S. A flexible semantic integration framework for fully-integrated EHR based on FHIR standard. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence. 2020 Presented at: ICAART '20; February 22-24, 2020; Valletta, Malta p. 684-691 URL: https://www.scitepress.org/Papers/2020/89815/89815.pdf [doi: 10.5220/0008981506840691]

XSL•FO
RenderX

59. Kilintzis V, Alexandropoulos VC, Beredimas N. A methodology for an auto-generated and auto-maintained HL7 FHIR OWL ontology for health data management. In: Delgado J, Benis A, de Toledo P, Gallos P, Giacomini M, Martínez-García A, et al, editors. Studies in Health Technology and Informatics. Amsterdam, The Netherlands: IOS Press; 2021:99-103.
60. Beredimas N, Kilintzis V, Chouvarda I, Maglaveras N. A reusable ontology for primitive and complex HL7 FHIR data types. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:2547-2550. [doi: 10.1109/EMBC.2015.7318911] [Medline: 26736811]
61. Kiourtis A, Mavrogiorgou A, Kyriazis D. A semantic similarity evaluation for healthcare ontologies matching to HL7 FHIR resources. Stud Health Technol Inform 2020 Jun 16;270:13-17. [doi: 10.3233/SHTI200113] [Medline: 32570337]
62. Kiourtis A, Nifakos S, Mavrogiorgou A, Kyriazis D. Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching. Int J Med Inform 2019 Dec;132:104002. [doi: 10.1016/j.ijmedinf.2019.104002] [Medline: 31629311]
63. Kober G. Annotating FHIR-RDF-graphs with medication knowledge. In: Proceedings of the Workshop on Automation Experience across Domains, AutomationXP 2020, co-located with the ACM Conference on Human Factors in Computing Systems. 2020 Presented at: CEUR '20; April 26, 2020; Honolulu, HI p. 1-8 URL: https://ceur-ws.org/Vol-2759/paper5.pdf
64. Kwon S, Yu J, Park S, Jun JA, Pyo CS. Automatic stroke medical ontology augmentation with standard medical terminology and unstructured textual medical knowledge. In: Proceedings of the 2021 International Conference on Platform Technology and Service. 2021 Presented at: PlatCon '21; August 23-25, 2021; Jeju, Korea p. 1-5 URL: https://ieeexplore.ieee.org/document/9680753 [doi: 10.1109/platcon53246.2021.9680753]
65. Dong X, Samavi R, Topaloglou T. COC: an ontology for capturing semantics of circle of care. Procedia Comput Sci 2015;63:589-594 [FREE Full text] [doi: 10.1016/j.procs.2015.08.389]
66. Jiang G, PrudHommeaux E, Xiao G, Solbrig HR. Developing a semantic web-based framework for executing the clinical quality language using FHIR. In: Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. 2020 Presented at: SWAT4LS '17; December 4-7, 2017; Rome, Italy p. 1-5 URL: https://ceur-ws.org/Vol-2042/paper40.pdf
67. Prud'hommeaux E, Collins J, Booth D, Peterson KJ, Solbrig HR, Jiang G. Development of a FHIR RDF data transformation and validation framework and its evaluation. J Biomed Inform 2021 May;117:103755 [FREE Full text] [doi: 10.1016/j.jbi.2021.103755] [Medline: 33781919]
68. Sharma D, Prud'Hommeaux E, Booth D, Peterson KJ, Stone DJ, Solbrig H, et al. Implementing a new FHIR RDF specification for semantic clinical data using a JSONLD- based approach. In: Proceedings of the 13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. 2022 Presented at: SWAT4HCLS '22; January 10-14, 2022; Leiden, The Netherlands p. 1-5 URL: https://ceur-ws.org/Vol-3127/paper-10.pdf
69. Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, et al. Leveraging genetic reports and electronic health records for the prediction of primary cancers: algorithm development and validation study. JMIR Med Inform 2021 May 25;9(5):e23586 [FREE Full text] [doi: 10.2196/23586] [Medline: 34032581]
70. Kober G, Paschke A. Medical information-graphs, based on ontologies and FHIR. In: Proceedings of the 2021 Conference on Digital Curation Technologies. 2021 Presented at: Qurator '21; February 8-12, 2021; Berlin, Germany p. 1-13 URL: https://ceur-ws.org/Vol-2836/qurator2021_paper_14.pdf
71. Kober G, Robaldo L, Paschke A. Modeling medical guidelines by Prova and SHACL accessing FHIR/RDF. Use case: the medical ABCDE approach. Stud Health Technol Inform 2022 May 16;293:59-66. [doi: 10.3233/SHTI220348] [Medline: 35592961]
72. Luz MP, Nogueira JR, Cavalini LT, Cook TW. Providing full semantic interoperability for the fast healthcare interoperability resources schemas with resource description framework. In: Proceedings of the 2015 International Conference on Healthcare Informatics. 2015 Presented at: ICHI '15; 2015; Dallas, TX p. 463-466 URL: https://ieeexplore.ieee.org/document/7349736 [doi: 10.1109/ichi.2015.74]
73. Szmeja P, Prud'hommeaux E. ShExMap and IPSM-AF: comparison of RDF transformation technologies. In: Proceedings of the 6th International Conference on Intelligent Systems, Technologies and Applications. 2021 Presented at: ISTA '21; June 15-17, 2021; Chennai, India p. 29-45 URL: https://link.springer.com/chapter/10.1007/978-981-16-0730-1_3 [doi: 10.1007/978-981-16-0730-1_3]
74. Thornton K, Solbrig H, Stupp GS, Gayo JE, Mietchen D, Prud'hommeaux E, et al. Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In: Proceedings of the 16th International Conference on the Semantic Web. 2019 Presented at: ESWC '19; June 2-6, 2019; Portorož, Slovenia p. 606-620 URL: https://link.springer.com/chapter/10.1007/978-3-030-21348-0_39 [doi: 10.1007/978-3-030-21348-0_39]
75. De A, Huang M, Feng T, Yue X, Yao L. Analyzing patient secure messages using a fast health care interoperability resources (FIHR)-based data model: development and topic modeling study. J Med Internet Res 2021 Jul 30;23(7):e26770 [FREE Full text] [doi: 10.2196/26770] [Medline: 34328444]
76. Hong N, Wen A, Mojarad MR, Sohn S, Liu H, Jiang G. Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. AMIA Annu Symp Proc 2018;2018:574-583 [FREE Full text] [Medline: 30815098]

XSL•FO
RenderX

77. Soni S, Gudala M, Wang DZ, Roberts K. Using FHIR to construct a corpus of clinical questions annotated with logical forms and answers. AMIA Annu Symp Proc 2019;2019:1207-1215 [FREE Full text] [Medline: 32308918]
78. Aberdeen J, Bayer S, Clark C, Keybl M, Tresner-Kirsch D. An annotation and modeling schema for prescription regimens. J Biomed Semantics 2019 May 31;10(1):10 [FREE Full text] [doi: 10.1186/s13326-019-0201-9] [Medline: 31151407]
79. Hong J, Davoudi A, Yu S, Mowery DL. Annotation and extraction of age and temporally-related events from clinical histories. BMC Med Inform Decis Mak 2020 Dec 30;20(Suppl 11):338 [FREE Full text] [doi: 10.1186/s12911-020-01333-5] [Medline: 33380319]
80. Ciampi M, Sicuranza M, Silvestri S. A privacy-preserving and standard-based architecture for secondary use of clinical data. Information 2022 Feb 13;13(2):87 [FREE Full text] [doi: 10.3390/info13020087]
81. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
82. Lenivtceva I, Slasten E, Kashina M, Kopanitsa G. Applicability of machine learning methods to multi-label medical text classification. In: Proceedings of the 20th International Conference on Computational Science. 2020 Presented at: ICCS '20; June 3-5, 2020; Amsterdam, The Netherland p. 509-522 URL: https://link.springer.com/chapter/10.1007/978-3-030-50423-6_38#:~:text=Medical%20text%20classification%20task%20facilitates,manually%20labeled%20Russian%20medical%20notes [doi: 10.1007/978-3-030-50423-6_38]
83. El-Sappagh S, Ali F, Hendawi A, Jang JH, Kwak KS. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97 [FREE Full text] [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
84. Solbrig HR, Prud'hommeaux E, Grieve G, McKenzie L, Mandel JC, Sharma DK, et al. Modeling and validating HL7 FHIR profiles using semantic web shape expressions (ShEx). J Biomed Inform 2017 Mar;67:90-100 [FREE Full text] [doi: 10.1016/j.jbi.2017.02.009] [Medline: 28213144]
85. Kiourtis A, Mavrogiorgou A, Menychtas A, Maglogiannis I, Kyriazis D. Structurally mapping healthcare data to HL7 FHIR through ontology alignment. J Med Syst 2019 Feb 05;43(3):62. [doi: 10.1007/s10916-019-1183-y] [Medline: 30721349]

Abbreviations

AMIA: American Medical Informatics Association
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
ICD: International Classification of Diseases
IEEE: Institute of Electrical and Electronics Engineers
LOINC: Logical Observation Identifiers Names and Codes
ML: machine learning
NLP: natural language processing
OWL: web ontology language
QA: quality assessment
RDF: resource description framework
RQ: research question
SMR: systematic mapping review
SNOMED CT: Systematized Nomenclature of Medicine–clinical terms

Edited by A Mavragani; submitted 21.12.22; peer-reviewed by A Kiourtis, P Knaup-Gregori, S Thun, D Hansen; comments to author 16.02.23; revised version received 07.03.23; accepted 19.12.23; published 30.01.24.

Please cite as:
Amar F, April A, Abran A
Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review
J Med Internet Res 2024;26:e45209
URL: https://www.jmir.org/2024/1/e45209
doi:10.2196/45209
PMID:38289660

©Fouzia Amar, Alain April, Alain Abran. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Performance of Wearable AI in Detecting Stress Among Students: Systematic Review and Meta-Analysis

Alaa Abd-alrazaq[1], PhD; Mohannad Alajlani[2], PhD; Reham Ahmad[2], MSc; Rawan AlSaad[1], PhD; Sarah Aziz[1], MSc; Arfan Ahmed[1], PhD; Mohammed Alsahli[3], PhD; Rafat Damseh[4], PhD; Javaid Sheikh[1], MD

[1]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Qatar Foundation, Doha, Qatar

[2]Institute of Digital Healthcare, WMG, University of Warwick, Warwick, United Kingdom

[3]Health Informatics Department, College of Health Science, Saudi Electronic University, Riyadh, Saudi Arabia

[4]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

Corresponding Author:
Alaa Abd-alrazaq, PhD
AI Center for Precision Health
Weill Cornell Medicine-Qatar
Qatar Foundation
PO Box 5825, Doha Al Luqta St, Ar-Rayyan
Doha, 2442
Qatar
Phone: 974 5570845212
Email: Aaa4027@qatar-med.cornell.edu

Abstract

Background: Students usually encounter stress throughout their academic path. Ongoing stressors may lead to chronic stress, adversely affecting their physical and mental well-being. Thus, early detection and monitoring of stress among students are crucial. Wearable artificial intelligence (AI) has emerged as a valuable tool for this purpose. It offers an objective, noninvasive, nonobtrusive, automated approach to continuously monitor biomarkers in real time, thereby addressing the limitations of traditional approaches such as self-reported questionnaires.

Objective: This systematic review and meta-analysis aim to assess the performance of wearable AI in detecting and predicting stress among students.

Methods: Search sources in this review included 7 electronic databases (MEDLINE, Embase, PsycINFO, ACM Digital Library, Scopus, IEEE Xplore, and Google Scholar). We also checked the reference lists of the included studies and checked studies that cited the included studies. The search was conducted on June 12, 2023. This review included research articles centered on the creation or application of AI algorithms for the detection or prediction of stress among students using data from wearable devices. In total, 2 independent reviewers performed study selection, data extraction, and risk-of-bias assessment. The Quality Assessment of Diagnostic Accuracy Studies–Revised tool was adapted and used to examine the risk of bias in the included studies. Evidence synthesis was conducted using narrative and statistical techniques.

Results: This review included 5.8% (19/327) of the studies retrieved from the search sources. A meta-analysis of 37 accuracy estimates derived from 32% (6/19) of the studies revealed a pooled mean accuracy of 0.856 (95% CI 0.70-0.93). Subgroup analyses demonstrated that the accuracy of wearable AI was moderated by the number of stress classes (P=.02), type of wearable device (P=.049), location of the wearable device (P=.02), data set size (P=.009), and ground truth (P=.001). The average estimates of sensitivity, specificity, and F_1-score were 0.755 (SD 0.181), 0.744 (SD 0.147), and 0.759 (SD 0.139), respectively.

Conclusions: Wearable AI shows promise in detecting student stress but currently has suboptimal performance. The results of the subgroup analyses should be carefully interpreted given that many of these findings may be due to other confounding factors rather than the underlying grouping characteristics. Thus, wearable AI should be used alongside other assessments (eg, clinical questionnaires) until further evidence is available. Future research should explore the ability of wearable AI to differentiate types of stress, distinguish stress from other mental health issues, predict future occurrences of stress, consider factors such as the placement of the wearable device and the methods used to assess the ground truth, and report detailed results to facilitate the conduct of meta-analyses.

Trial Registration: PROSPERO CRD42023435051; http://tinyurl.com/3fzb5rnp

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52622) doi:10.2196/52622

KEYWORDS

stress; artificial intelligence; wearable devices; machine learning; systematic review; students; mobile phone

Introduction

Background

Students' mental health is now of significant concern in the domain of public health. The academic environment, composed of rigorous curriculum structures, exams, assignments, evaluative metrics, deadlines, and an undercurrent of peer comparisons, not only requires a high cognitive effort but can also be a basis for psychosocial risks [1-3]. Being immersed in such scholarly endeavors at different levels, students become inadvertently vulnerable to mental health issues (eg, anxiety, depression, and stress [3,4]). This can negatively affect a student's life in general and, specifically, their academic performance. Research has revealed that students who are constantly under high levels of stress are more inclined to underperform academically when compared with their peers who are less stressed [5]. Recent studies suggest that failure to manage academic-related stress is detrimental to the well-being of students [6-8].

The psychobiological literature defines stress, including academic stress, as a syndrome of responses combining interactions between the neurocognitive, endocrine, and affective systems [9-11]. Stress manifests when exposed to conditions that present a physical and psychological challenge or threat [11]. Stress is also defined as the result of incompatibility in relation to an individual's ability to handle the burden placed on them by their environment [12]. Stress can be classified based on its duration into 2 forms: chronic and acute stress. Acute stress is a short-term stress response to an immediate situation and usually subsides soon after symptoms first appear, whereas chronic stress is a long-term stress response to ongoing stressors [13]. Typically, acute stress is the most common and does not impose a health burden on young and healthy individuals. In contrast, chronic stress can have a negative impact on both physical and mental health [14,15]. Potential biological risks include inflammation, diabetes, and heart disease [16,17]. Furthermore, studies have shown an association with cognitive dysfunctions because of excessive and chronic stressors [18,19]. Chronic stress can also have negative effects on social behavior, which leads to antisociality [20]; depression; or, in extreme cases, suicide [21,22]. Worldwide, it has been reported that, for every 5 visits by college students to their physician, 3 are for stress-related problems [23]. In addition, most college students who have problems with stress also have sleep-related disorders (76%), are prone to headaches (58%), experience difficult relationships with family and friends (85%), and tend to be short-tempered (70%) [23]. The American Institute of Stress reveals that, in the United States, a combined US $300 billion is spent on treating stress-related ailments and diseases [13].

Given the aforementioned complications of stress, the detection and monitoring of stress in early stages are crucial. Subjective and objective assessments can be used to evaluate stress levels. Subjective assessments of stress can involve 2 different means: standard self-reported questionnaires designed by field experts (eg, the Perceived Stress Scale and Depression, Anxiety, and Stress Scale–21 Items) or clinical interviews with psychologists [15,24]. Beyond this, objective assessments of stress include physical observations and physiological measures. For physical observations, visible changes in the appearance of the human body are noted, such as any facial expressions, the rate of blinking, and any noticeable dilation of the pupils. For physiological measures, blood, urine, or saliva samples can be collected to measure the level of specific hormones (adrenaline and cortisol) that are released as a response to a stressor [10,14,16]. Relying only on subjective assessments is insufficient to accurately detect and monitor stress as they are highly subjective, usually irreproducible, time-consuming, and entangled with other body responses not related to stress [25,26]. Although the aforementioned objective assessments can be deterministic in capturing stress status, they require invasive or obtrusive tests that are commonly deployed in a highly controlled laboratory environment and do not reflect stress in real-life scenarios [13,27]. Therefore, there is a critical need for alternative assessments of stress that are objective, noninvasive, nonobtrusive, and automatic and can continuously monitor biomarkers in real time.

Changes in biomarkers (eg, heart rate [HR], HR variability [HRV], and electrodermal activity [EDA]) can be used to objectively detect stress [15,28]. Real-time and continuous monitoring of these biomarkers must be supported by computational frameworks to automate the process of stress detection, thereby overcoming the aforementioned limitations of the current tools. Wearable artificial intelligence (AI) is a promising solution that has been used to address the issue of assessing stress [29-31]. Wearable AI is an advanced technology that depends on AI techniques to analyze a large amount of data (eg, HR, HRV, EDA, activity level, and skin temperature) collected by sensors in wearable devices to provide personalized feedback. Several types of wearable devices can be used to collect biomarkers: on-body devices, which are fixed directly on the body or skin; near-body devices, which are fixed close to the body but do not have direct contact with the skin; in-body devices, which are implanted in the body; and electronic textiles, which are fabrics that have integrated electronics [30].

Research Problem and Aim

In the past few years, numerous studies have shed light on the performance of wearable AI for the detection of stress. Several reviews of such studies have been conducted, but they have the following limitations. First, none of the previous reviews have focused on students [13,15,24,28,32-35]. Second, none of the previous reviews have used statistical techniques (eg, meta-analysis) to analyze the results of previous studies [13,15,24,28,32-35]. Third, most of the previous reviews have been literature reviews rather than systematic reviews [13,15,24,28,33,34]. Fourth, the search sources in the previous

XSL•FO
RenderX

reviews have not included the main databases in the field, such as MEDLINE [28], ACM Digital Library [28,32,33,35], Scopus [28,32,33,35], Embase [32,35], and IEEE Xplore [32,33,35]. Finally, several previous reviews have focused on wearable and nonwearable devices instead of on wearable devices only [13,15,24,32]. Therefore, this review sought to bridge this gap by assessing the capability of wearable AI to detect and forecast stress among students. Our research question is as follows: what is the effectiveness of wearable AI in diagnosing and predicting stress among students?

Methods

Overview

The authors undertook and reported this review in line with the guidelines set forth by the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy) [36]. The PRISMA-DTA checklist for this review is presented in Multimedia Appendix 1 [36]. The protocol for this review was registered with PROSPERO (ID: CRD42023435051).

Search Strategy

To identify suitable studies, the primary author conducted a comprehensive search across 7 electronic repositories on June 12, 2023: MEDLINE (via Ovid), Embase (via Ovid), PsycINFO (via Ovid), ACM Digital Library, Scopus, IEEE Xplore, and Google Scholar. Using an automated search methodology, biweekly notifications were configured over a span of 3 months, concluding on September 12, 2023. Given the large number of results returned by Google Scholar, the first 100 results (equivalent to 10 pages) were subjected to scrutiny for this review. To find more studies, we checked the lists of sources referenced in the studies that we had already included (backward reference list checking), and we also looked at studies that had cited the ones we had already included (forward reference list checking).

The compilation of search terms for the review was informed by consultations with 2 digital mental health experts and scrutiny of relevant literature reviews. The resultant search query was composed of four distinct categories of terms: (1) terms associated with AI, such as “artificial intelligence,” “machine learning,” and “deep learning”; (2) terms linked to wearable devices, comprising “wearable*,” “smartwatch*,” and “smartband*”; (3) terms related to stress; and (4) terms relevant to students, including “student*,” “postgraduate*,” and “undergraduate*.” The Boolean OR operator was used to combine terms in the same category, whereas the Boolean AND operator was used to combine terms between categories. The precise search formulations are presented in Multimedia Appendix 2 for reference.

Study Eligibility Criteria

This review analyzed research articles centered on the creation or application of AI algorithms for the detection or prediction of stress among students using data from wearable devices. The standards for determining which articles to include and exclude were jointly established through the collaborative expertise of the authors. To be eligible for consideration in this review, studies needed to collect data from students regardless of their educational level, age, gender, and ethnicity. Furthermore, to be included in this review, studies had to assess the effectiveness of AI algorithms in identifying or foreseeing stress. They were also required to present the confusion matrix or performance metrics (such as accuracy, sensitivity, and specificity). Studies that used AI to predict the outcomes of stress interventions or treatments were excluded.

Eligible studies were required to use noninvasive wearable devices (eg, smartwatches, smart glasses, and smart rings) for data collection. This review encompassed studies that used various methods for data collection (such as nonwearable devices, interviews, and questionnaires) in addition to wearable devices. In contrast, studies that solely used the following devices for data collection were excluded: nonwearable devices, handheld devices (eg, mobile phones), near-body wearable devices, in-body wearable devices, wearable devices wired to nonwearable devices, and wearables requiring expert oversight (such as those requiring precise electrode placement).

The inclusion criteria covered peer-reviewed journal articles, conference papers, preprints, and dissertations irrespective of the study setting, reference standard (ie, ground truth), or the country of study. Given the emphasis on contemporary technology and the continuous growth in the field of wearable AI, only articles from 2015 onward were included. Articles not in English or structured as review pieces, editorials, conference abstracts, posters, protocols, and research highlights were all excluded. We included only prospective and retrospective experimental studies; however, studies that used publicly available data sets (eg, Wearable Stress and Affect Detection [WESAD]) or data sets from previous studies were excluded.

Study Selection

The study selection process involved 3 main steps. First, duplicate studies were removed using EndNote X9 (Clarivate Analytics). Then, 2 reviewers independently evaluated the titles and abstracts of the remaining articles. Finally, the reviewers independently assessed the full texts of the remaining articles. Any disagreements were discussed and resolved. The agreement between the reviewers was high, with a κ score of 0.87 for evaluating titles and abstracts and 0.92 for reading full texts.

Data Extraction

The data extraction form was developed and tested with 3 studies first (Multimedia Appendix 3). Using Microsoft Excel (Microsoft Corp), 2 reviewers independently extracted the studies’ metadata, features of the wearable devices, features of the AI algorithms, and the study results. Any disagreements between the reviewers were resolved through discussion. For studies that provided raw data or confusion matrices, we calculated performance measures such as accuracy, specificity, and sensitivity. If this information was not available, we reached out to the first and corresponding authors of those studies to try to obtain it. We did not include results based only on nonwearable device data (eg, data from smartphones or surveys). As many studies conducted multiple experiments with different features, data types, validation methods, and AI algorithms, they reported several results for the same performance measure.

XSL•FO
RenderX

In these cases, we obtained the best result in each performance measure for each algorithm.

Risk-of-Bias and Applicability Appraisal

To assess the quality of the included studies, we tailored the established Quality Assessment of Diagnostic Accuracy Studies–Revised (QUADAS-2) [37] tool to suit the objectives of our review. We made adjustments by replacing certain nonapplicable criteria with more relevant ones from the Prediction Model Risk of Bias Assessment Tool [38]. Our adjusted QUADAS-2 tool includes 4 primary categories: "Participants," "Index Test" (pertaining to AI algorithms), "Reference Standard" (denoting the ground truth), and "Analysis." We formulated 4 specific questions within each category to align with the aims of our review. In addition to examining the potential biases within these 4 categories, we assessed the real-world applicability of the findings for the first 3 categories. To optimize our adjusted tool, we first trialed it on 3 studies to fine-tune it. In total, 2 reviewers independently assessed the included studies using the adapted QUADAS-2 tool (Multimedia Appendix 4). Discrepancies between their evaluations were discussed and resolved through consensus.

Data Synthesis

The data from the studies were combined using narrative and statistical methods. In the narrative synthesis, we used text and tables to provide an overview and describe the key features of the included studies (study metadata, wearable devices, and AI techniques). A statistical approach was used when at least 2 different studies presented sufficient data to conduct meta-analyses. We did not consider the study design when selecting studies for meta-analysis. When estimates (eg, accuracy, sensitivity, and specificity) in the analysis were extracted from different unique studies (ie, independent effect sizes), DerSimonian-Laird random-effects models [39] using the Freeman-Tukey double arcsine transformation [40,41] were conducted to pool the extracted estimates. This approach considers the fluctuations arising from sampling and the heterogeneity in estimates. The analysis was executed using the *meta* toolkit in R (version 4.2.2; R Foundation for Statistical Computing) [42]. However, when there were estimates in the analysis extracted from the same study (ie, dependent effect sizes), we used a multilevel meta-analysis technique [39,43] to account for this dependency in effect sizes, thereby reducing the likelihood of type-I errors. Multilevel meta-analyses were conducted using the *metafor* toolkit in R (version 4.2.2) [40].

Where appropriate, we performed subgroup multilevel meta-analyses to investigate potential relationships between performance estimates and various factors [39,43]. These factors included AI algorithms, number of stress classes, type of wearable device, location of the wearable device, data set size, data sources, data types, stress inducers, ground truth, and validation methods. The difference in estimates between subgroups was considered statistically significant when the P value was <.05.

The Cochrane Q statistic was used to examine between-study heterogeneity, with a P value of <.05 indicating the presence of heterogeneity. The degree of between-study heterogeneity was evaluated using I^2 [40,44], where it was deemed insignificant when I^2 ranged from 0% to 40%, moderate when it ranged from 30% to 60%, substantial when it ranged from 50% to 90%, or considerable when it ranged from 75% to 100% [45].

Results

Search Results

As shown in Figure 1, a total of 327 studies were retrieved when the databases identified previously were searched. Of the 327 retrieved studies, 59 (18%) duplicates were removed using EndNote X9, leaving 268 (82%) studies. Furthermore, 71.3% (191/268) of the studies were removed after we screened the titles and abstracts. After retrieving and reading the full texts of the remaining 77 studies, it was determined that 59 (77%) were ineligible for inclusion. The main reasons for exclusion were that they used a public data set, did not use wearable devices, did not use AI algorithms, did not focus on stress, or were irrelevant publication types. We identified an additional study relevant to this review through backward reference list checking. In total, 19 studies were included in this review [46-64], and 6 (32%) were eligible for the meta-analyses [46,52,53,55,62,63].

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process. AI: artificial intelligence.

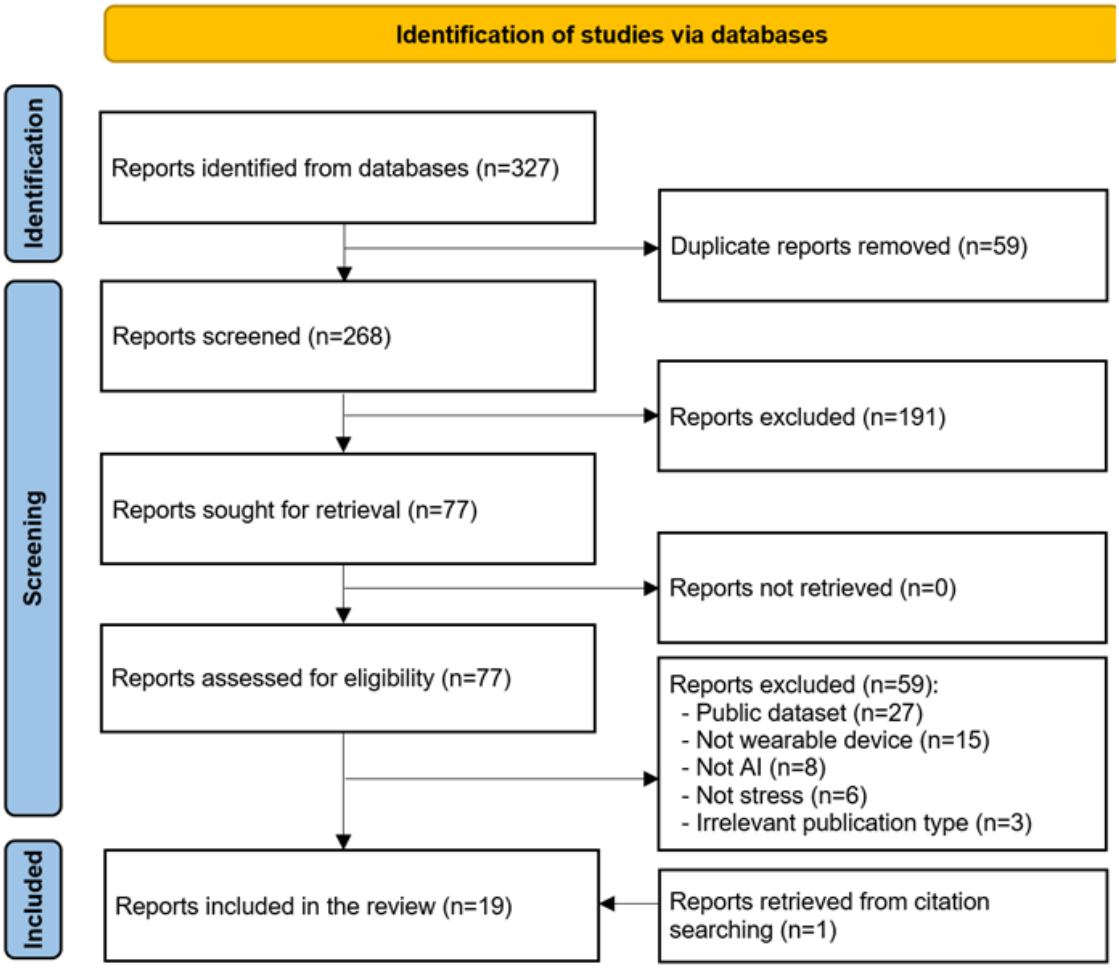

Characteristics of the Included Studies

The key characteristics of the studies included in this review are presented in Table 1. The included studies were published between 2015 and 2023. The year in which the largest number of included studies was published was 2019 (5/19, 26%). The included studies were conducted in 10 distinct countries (Table 1), wherein the United States contributed to over a third of the studies (7/19, 37%). More than half (10/19, 53%) of the studies were conference papers, whereas the remaining studies were journal articles (8/19, 42%) and a preprint (1/19, 5%). The number of participants in the included studies ranged from 5 to 652, with an average of 77 (SD 149.5). The mean age of the participants was reported in 42% (8/19) of the included studies and ranged from 19.8 to 27.5 years, with an average of 22.8 (SD 2.5) years. A total of 58% (11/19) of the studies reported the proportion of female participants, which ranged from 14.3% to 76.8%, with an average of 47.3% (SD 21.5%). Participants in most of the studies (17/19, 89%) were undergraduate students. The characteristics of each included study are presented in Multimedia Appendix 5 [46-64].

XSL•FO
RenderX

Table 1. Characteristics of the included studies (N=19).

Feature	Values	References
Year of publication, n (%)		
2023	2 (11)	[47,54]
2021	3 (16)	[46,53,63]
2020	4 (21)	[49,56,62,64]
2019	5 (26)	[48,52,55,57,58]
2018	3 (16)	[50,60,61]
2017	1 (5)	[51]
2015	1 (5)	[59]
Publication type, n (%)		
Conference paper	10 (53)	[46,51-53,55,57-59,61,64]
Journal article	8 (42)	[48-50,54,56,60,62,63]
Preprint	1 (5)	[47]
Country of publication, n (%)		
United States	7 (37)	[47,51,53-56,60]
India	3 (16)	[57,58,63]
United Kingdom	2 (11)	[49,59]
Others	1 (5) each	[46,48,50,52,61,62,64]
Number of participants, mean (SD; range)	77 (149.5; 5-652)	[46-64]
Age (y)		
Values, mean (SD; range)	22.76 (2.53; 19.8-27.5)	[47-49,54,56,61,62,64]
NR[a], n (%)	11 (58)	[46,50-53,55,57-60,63]
Female participants		
Values (%), mean (SD; range)	47.3 (21.5; 14.3-76.8)	[47-49,52-54,56,60-62,64]
NR, n (%)	8 (42)	[46,50,51,55,57-59,63]
Educational level, n (%)		
High school	1 (5)	[48]
Undergraduate	17 (89)	[47-60,62-64]
Postgraduate	3 (16)	[49,61,63]
NR	1 (5)	[46]

[a]NR: not reported.

Features of the Wearable Devices

As depicted in Table 2, although 89% (17/19) of the studies used commercial wearable devices, 5% (1/19) used noncommercial wearable devices. The included studies used 16 different wearable devices. The wearable devices that were used by the largest number of studies were Empatica E4 (3/19, 16%) and Galaxy (Watch Active2 and Gear S, S2, and S3; 3/19, 16%). The most common type of wearable device used in the included studies was smart bands (12/19, 63%). The wearable devices were placed on 7 different parts of the body; however, the wrist was the most common (16/19, 84%). The periods for which the devices were worn in the included studies ranged from 10 minutes to 365 days. The features of the wearable devices in each included study are presented in Multimedia Appendix 6 [46-64].

XSL•FO
RenderX

Table 2. Features of the wearable devices (N=19).

Feature	Values	References
Status of WD[a], n (%)		
Commercial	17 (89)	[46-54,57-60,62-64]
Noncommercial	1 (5)	[51]
Not reported	2 (11)	[55,56]
Name of WD, n (%)		
Empatica E4	3 (16)	[48,61,64]
Galaxy (Watch Active2 and Gear S, S2, and S3)	3 (16)	[47,48,54]
Fitbit (Charge 2)	2 (11)	[53,58]
Microsoft Band 2	2 (11)	[50,62]
Others	1 (5) each	[46,49,51,52,54,59,60,63]
Not reported	2 (11)	[55,56]
Type of WD, n (%)		
Smart bands	12 (63)	[48-51,55,56,58-62,64]
Smartwatches	6 (32)	[46-48,51-54,63]
Electrodes	2 (11)	[52,63]
Others	1 (5) each	[54,57]
Placement of WD, n (%)		
Wrist	16 (84)	[46-51,53-56,58-62,64]
Chest	3 (16)	[51,52,63]
Finger	3 (16)	[54,59,63]
Palm	2 (11)	[52,63]
Others	1 (5) each	[57,63]
Duration of wearing WD		
Range	10 min to 365 d	N/A[b]
Not reported, n (%)	6 (32)	[46,50-52,57,62]

[a]WD: wearable device.

[b]N/A: not applicable.

Features of the AI

The AI algorithms in most of the included studies (18/19, 95%) were used to solve classification problems (Table 3). Most studies (15/19, 79%) used AI algorithms to categorize the sample into 2 classes (stressed vs not stressed). Among the included studies, 22 different algorithms were used, but the most commonly used algorithms were support vector machine (14/19, 74%) and k-nearest neighbor (13/19, 68%). All the included studies used AI to detect the current stress status, whereas none of them used AI to predict the occurrence of stress in the future. The data set size was reported in 47% (9/19) of the studies and ranged from 30 to 1178, with an average of 517.9 (SD 409). Data were collected using wearable devices (19/19, 100%), self-reported questionnaires (11/19, 58%), and nonwearable devices (eg, smartphones; 6/19, 32%). The included studies used 17 types of data to develop the models. However, the most common data types used to develop the models were HR data (eg, HR, HRV, and interbeat interval; 16/19, 84%), EDA data (10/19, 53%), and physical activity data (eg, step counts, calories, and metabolic rate) (8/19, 42%). The number of features extracted from the data varied between 4 and 498, with an average of 52.2 (SD 115.1). Stress inducers were used in 63% (12/19) of the studies, and the most common method used to induce stress among participants was arithmetic tasks (7/19, 37%). The ground truth was identified using self-reported questionnaires (eg, Perceived Stress Scale; 11/19, 58%), the context of the experiment (eg, stress exposure period and non–stress exposure period; 9/19, 47%), or unsupervised algorithms (1/19, 5%). The most common method used to validate the performance of the models was k-fold cross-validation (11/19, 58%). The most common measure used in the included studies to evaluate the performance of AI algorithms was accuracy (16/19, 84%). The features of the AI in each included study are described in Multimedia Appendix 7 [46-64].

XSL•FO
RenderX

Table 3. Features of the artificial intelligence (AI; N=19).

Feature	Values	References
Problem-solving approaches, n (%)		
Classification	18 (95)	[46-63]
Regression	1 (5)	[50]
Clustering	1 (5)	[64]
Number of stress classes, n (%)		
2	15 (79)	[46,47,49,51,53-62,64]
3	5 (26)	[48,50,52,57,61]
>3	2 (11)	[55,63]
AI algorithms, n (%)		
Support vector machine	14 (74)	[46,48,50-57,59,60,62,63]
K-nearest neighbor	13 (68)	[46,47,50,52-57,61-63]
Random forest	8 (42)	[47,50,51,54,55,61-63]
Logistic regression	6 (32)	[46,48,51,53,62,63]
Decision tree	5 (26)	[50,52,55,58,61]
Naïve Bayes	5 (26)	[46,50,51,58,62]
Artificial neural network	4 (21)	[53,55,62,63]
Multilayer perceptron	4 (21)	[46,48,54,56]
Others	<4	[46-50,52,55,56,60-64]
Aim of AI algorithms, n (%)		
Detection	19 (100)	[46-64]
Data set size, mean (SD; range)	518 (409; 30-1178)	[46,47,52,53,55,61-64]
Data sources, n (%)		
Wearable devices	19 (100)	[46-64]
Self-reported questionnaire	11 (58)	[47,49,51,53-55,58,60,62,63]
Nonwearable devices	6 (32)	[47,53,54,58,60,61]
Data types, n (%)		
Heart rate data	16 (84)	[46-54,56,58,59,61,63,64]
Electrodermal activity data	10 (53)	[48,50-52,54,59-61,63,64]
Activity data	8 (42)	[47,48,53-55,58,60,61]
Sleep data	5 (26)	[53-55,58,60]
Skin temperature	3 (16)	[48,50,60]
Smartphone use data	3 (16)	[47,54,60]
Location	3 (16)	[47,54,60]
Respiratory rate data	3 (16)	[47,54,61]
Others	<3	[47,53,54,57,58,60,61]
Number of features		
Values, mean (SD; range)	52.2 (115.1; 4-498)	[46-64]
Not reported, n (%)	1 (5)	[55]
Stress-induction methods, n (%)		
Arithmetic tasks	7 (37)	[46,50,51,56,59,61,63]
Exams	4 (21)	[49,50,52,62]
Stroop test	4 (21)	[50,51,56,63]

XSL•FO
RenderX

Feature	Values	References
Physical stress	3 (16)	[50,51,56]
Others	<3	[48,50-52,56,59,61,64]
No stress induction	7 (37)	[47,53-55,57,58,60]
Ground truth, n (%)		
Self-reported questionnaire	11 (58)	[47-49,51,53-55,57,58,60,62]
Context	9 (47)	[48,50-52,56,59,61,63,64]
Unsupervised algorithm	1 (5)	[46]
Types of validation, n (%)		
K-fold	11 (58)	[46-48,50-54,56,57,63]
Training-test split	6 (32)	[46,49,55,57,59,62]
Leave-one-out cross-validation	2 (11)	[51,61]
Nested	1 (5)	[60]
Not reported	1 (5)	[58]
Not applicable	1 (5)	[64]
Performance measures, n (%)		
Accuracy	16 (84)	[46,48-50,52-61,63,64]
F_1-score	9 (47)	[46-48,51,55,56,60,61,64]
Precision	7 (37)	[48,51,55,56,58,59,64]
Sensitivity	6 (32)	[48,55,56,58,62,64]
κ	3 (16)	[46,50,51]
Others	<4	[46,50,51,58,62,64]

Results of the Risk-of-Bias Appraisal

All studies (19/19, 100%) reported comprehensive details to determine whether an appropriate consecutive or random sample of eligible patients was used, and none of the studies had inappropriate exclusions. Nearly half (8/19, 42%) of the studies ensured a balanced number of patients across subgroups. Only a third (6/19, 32%) reported a sufficient sample size, leaving ambiguity regarding the adequacy of the sample size in most of the studies (13/19, 68%). As a result, only a little over half (11/19, 58%) of the studies were assessed as having a low risk of bias from the “selection of participants” domain (Figure 2). In terms of matching participants to the predefined requirements in the review question, a low level of concern was identified in most of the studies, accounting for 68% (13/19).

XSL•FO
RenderX

Figure 2. Results of the assessment of risk of bias in the included studies.

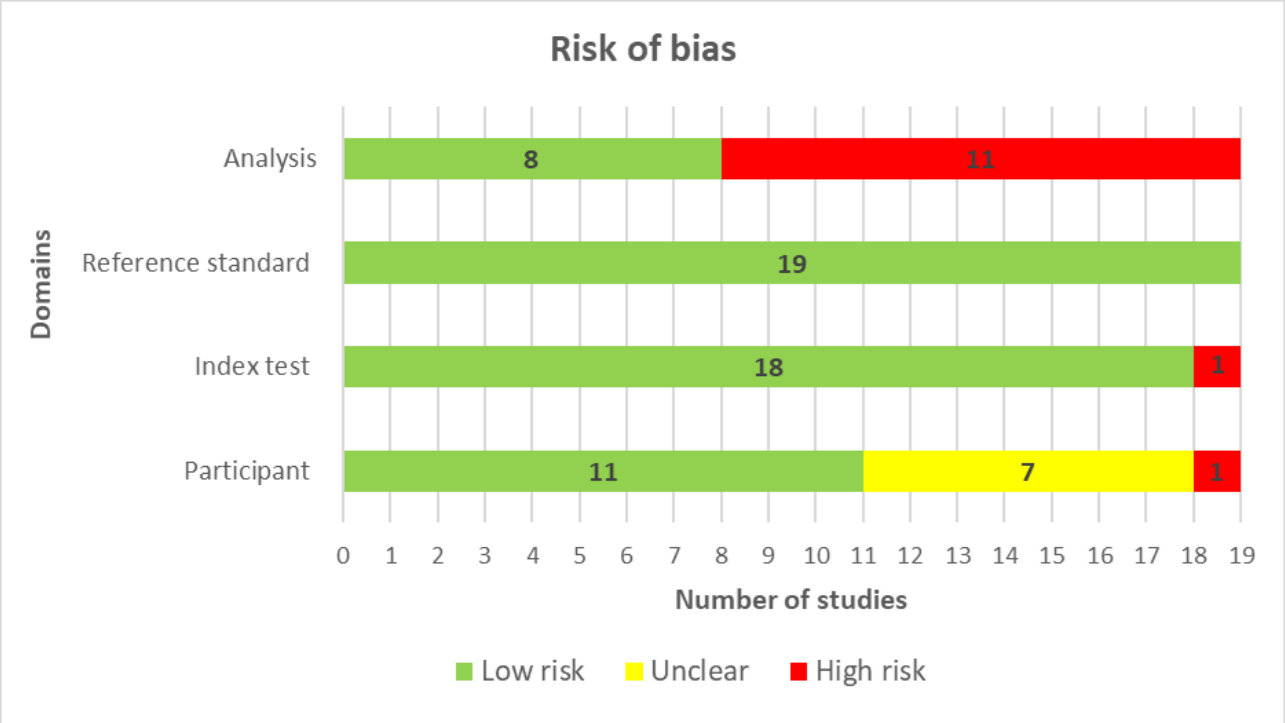

All studies in our review (19/19, 100%) comprehensively detailed the AI models, clearly reported the features (predictors) used, and ensured that these features were sourced without previous knowledge of the outcome data. For nearly every study (18/19, 95%), the features were consistently assessed across participants. Therefore, for the vast majority of the studies (18/19, 95%), the potential for bias in the "index test" was assessed as low (Figure 2). Consistently, all studies (19/19, 100%) were found to have minimal concerns regarding the match between the model's predictors and the review question's criteria (Figure 3).

Figure 3. Results of the assessment of applicability concerns in the included studies.

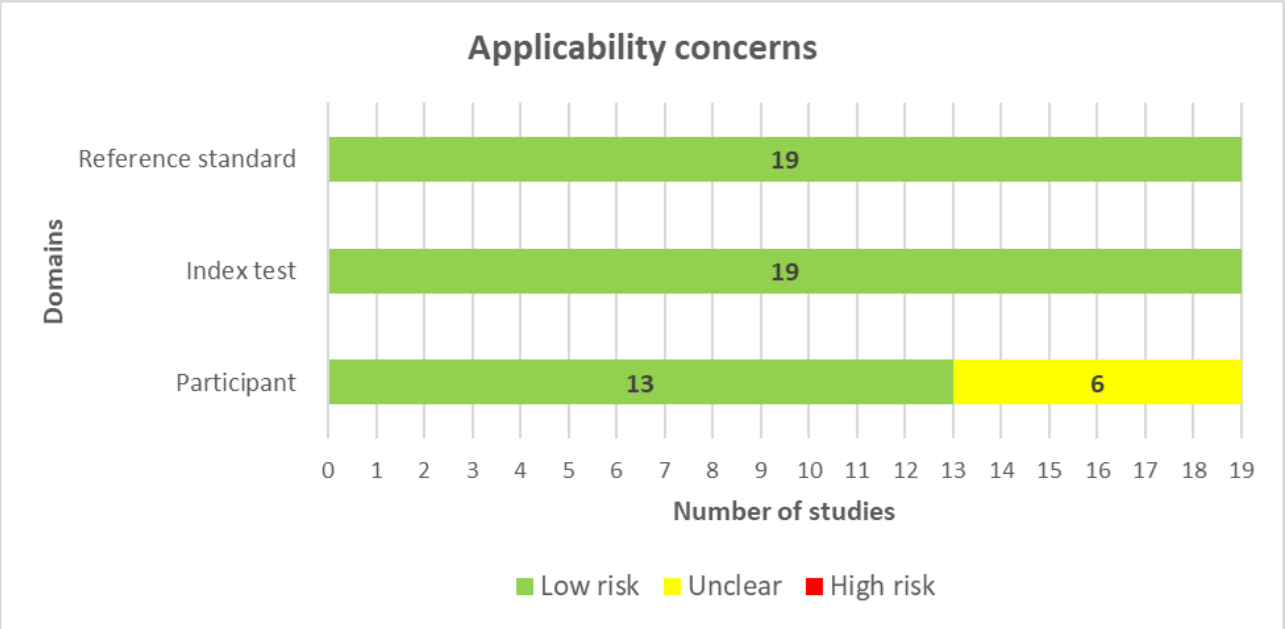

In all the reviewed studies (19/19, 100%), the outcome of interest, specifically the stress level, was consistently assessed using appropriate methodologies. Similarly, outcomes were defined uniformly for participants across studies and determined without access to predictor data, and an appropriate time gap was ensured between the index test and the reference standard. As a result, the potential for bias regarding the "reference standard" domain was deemed low for all studies (19/19, 100%; Figure 2). In addition, all studies (19/19, 100%) showed minimal concerns over discrepancies between the outcome's definition, timing, or determination and the review question's criteria (Figure 3).

XSL•FO
RenderX

Finally, two-thirds of the studies (12/19, 63%) ensured that all the enrolled participants were factored into the data analysis. Over half (10/19, 53%) of these studies executed proper data preprocessing, whereas in 68% (13/19), there was an appropriate split between training, validation, and test sets. The same proportion of studies (13/19, 68%) adopted suitable measures to evaluate the performance of their models. However, considering all these factors, only a minority (8/19, 42%) were deemed to have a minimal "analysis" domain risk of bias (Figure 2). A detailed breakdown of the "risk of bias" and "applicability concerns" for every domain in each study is available in Multimedia Appendix 8 [46-64].

Results of the Studies

Overview

Accuracy, sensitivity, specificity, and F_1-score were reported in 89% (17/19), 32% (6/19), 11% (2/19), and 53% (10/19) of the studies, respectively. Accuracy values were synthesized using both narrative and statistical methods, whereas the values of the remaining measures were synthesized solely through narrative methods. This distinction arises from the fact that 35% (6/17) of the studies that assessed accuracy reported enough data to conduct meta-analyses, whereas none of the studies evaluating the other measures had sufficient data for such analyses. The results related to accuracy and other measures are presented in the following subsections.

Accuracy

Wearable AI accuracy was assessed in 89% (17/19) of the studies. From these studies, we extracted the highest accuracy for each algorithm in each study; therefore, we identified 71 accuracy estimates. The estimates of accuracy ranged from 0.48 to 1.00, with an average of 0.822 (SD 0.135). We conducted meta-analyses of 37 estimates derived from 11,425 participants across 35% (6/17) of the studies (Table 4). The pooled mean accuracy of these estimates was 0.856 (95% CI 0.698-0.934). The meta-analyzed evidence exhibited considerable statistical heterogeneity (P<.001; I^2=98.8%). Furthermore, subgroup meta-analyses were conducted to assess the performance of wearable AI based on different factors. As shown in Table 4, there was a statistically significant difference in pooled accuracy between subgroups in the number of stress classes group (P=.02), type of wearable device group (P=.049), location of wearable device group (P=.02), data set size group (P=.009), and ground truth group (P=.001), whereas no statistically significant difference was found in the pooled accuracy between subgroups in the remaining groups.

XSL•FO
RenderX

Table 4. Pooled mean estimates of accuracy by several factors (N=19).

Group	Number of studies[a]	Sample size, n	Accuracy (%), range	Pooled mean accuracy (%), mean (95% CI)	Heterogeneity measures			Test for subgroup differences (*P* value)
					τ^2	Q (*P* value)	I^2 (%)	
Algorithm								.98
Support vector machine	6	2039	0.65-0.94	0.796 (0.67-0.90)	0.0276	143.0 (<.001)	96.5	
K-nearest neighbor	6	2039	0.58-0.98	0.834 (0.71-0.93)	0.0326	96.2 (<.001)	94.8	
Logistic regression	4	1545	0.66-0.87	0.768 (0.67-0.86)	0.0108	20.7 (<.001)	85.5	
Artificial neural network	4	1686	0.66-0.99	0.811 (0.63-0.94)	0.0436	58.3 (<.001)	94.9	
Random forest	3	1215	0.71-0.98	0.829 (0.61-0.97)	0.0456	43.1 (<.001)	95.4	
Decision tree	2	494	0.64-0.86	0.762 (0.52-0.94)	0.0319	29.6 (<.001)	96.6	
Naïve Bayes	2	990	0.68-0.97	0.841 (0.50-1.00)	0.0672	16.9 (<.001)	94.1	
Number of stress classes								*.02*[b]
2	25	9222	0.58-0.97	0.773 (0.61-0.87)	0.0873	173.4 (<.001)	97.2	
>2	12	2203	0.86-0.99	0.944 (0.77-0.99)	0.2110	302.8 (<.001)	98.2	
Type of WD[c]								*.049*
Smartwatches	11	2094	0.66-0.97	0.836 (0.53-0.95)	0.1466	91.6 (<.001)	96.6	
Smart bands	14	7128	0.58-0.77	0.695 (0.63-0.75)	0.0050	69.2 (<.001)	85.4	
Electrodes	12	2203	0.86-0.99	0.944 (0.77-0.99)	0.2110	302.8 (<.001)	98.2	
Location of WD								*.02*
Wrist	25	9222	0.58-0.97	0.773 (0.61-0.87)	0.0873	173.4 (<.001)	97.2	
Nonwrist	12	2203	0.86-0.99	0.944 (0.77-0.99)	0.2110	302.8 (<.001)	98.2	
Data set size								*.009*
≤100	14	789	0.80-0.99	0.947 (0.81-0.99)	0.1704	166.7 (<.001)	94.1	
>100	23	10,627	0.58-0.91	0.769 (0.62-0.86)	0.0758	485.3 (<.001)	97.4	
Data source								.57
WD-based	12	1825	0.80-0.97	0.895 (0.87-0.91)	0.0000	29.9 (<.001)	60.5	
WD-based data+other data	25	9600	0.58-0.99	0.830 (0.50-0.95)	0.3708	968.9 (<.001)	99.4	
Data type								.21
Heart rate data	13	5970	0.65-0.97	0.832 (0.52-0.95)	0.1594	115.4 (<.001)	98.6	

XSL•FO
RenderX

Group	Number of studies[a]	Sample size, n	Accuracy (%), range	Pooled mean accuracy (%), mean (95% CI)	Heterogeneity measures			Test for subgroup differences (*P* value)
					τ^2	Q (*P* value)	I^2 (%)	
Heart rate data+EDA[d] data	12	2203	0.86-0.99	0.944 (0.77-0.99)	0.2110	302.8 (<.001)	98.2	
Stress inducers								.10
Yes	25	8173	0.65-0.99	0.902 (0.75-0.96)	0.2292	1137.8 (<.001)	98.9	
No	12	3252	0.58-0.80	0.697 (0.62-0.76)	0.0060	51.4 (<.001)	80.6	
Ground truth								*.001*
Objective	19	2413	0.80-0.99	0.933 (0.84-0.97)	0.1266	320.2 (<.001)	96.3	
Subjective	18	9012	0.58-0.80	0.706 (0.66-0.75)	0.0031	95.0 (<.001)	85.1	
Validation method								.16
K-fold	16	4087	0.66-0.99	0.904 (0.64-0.98)	0.3324	784.7 (<.001)	99.0	
Training-test split	14	7128	0.58-0.77	0.695 (0.63-0.75)	0.0050	69.2 (<.001)	85.4	
All studies	37	11,425	0.58-0.99	0.856 (0.698-0.934)	0.2402	1318.2 (<.001)	98.8	N/A[e]

[a]Many studies were included more than once in all meta-analyses except for the meta-analysis related to algorithms given that the studies assessed the performance of more than one algorithm.

[b]Italics indicate statistical significance (*P*<.05).

[c]WD: wearable device.

[d]EDA: electrodermal activity.

[e]N/A: not applicable.

Other Measures

The sensitivity of wearable AI in detecting stress was reported in 32% (6/19) of the studies. From these studies, 27 sensitivity estimates were identified given that we extracted the highest sensitivity for each algorithm in each study. The estimates of sensitivity varied between 0.42 and 0.97, with an average of 0.755 (SD 0.181). From the 11% (2/19) of studies that reported specificity, 8 estimates of specificity were extracted given that more than one algorithm was assessed in these studies. The estimates of specificity ranged between 0.475 and 0.95, with an average of 0.744 (SD 0.147). We identified 46 estimates of F_1-score from 53% (10/19) of the studies. The estimates of the F_1-score ranged from 0.514 to 0.979, with an average of 0.759 (SD 0.139).

Discussion

Principal Findings

This review assessed the performance of wearable AI in detecting stress among students. Our analyses showed that wearable AI has an acceptable performance in detecting stress among students, but there is still room for improvement. Specifically, our meta-analyses revealed that wearable AI was able to correctly classify students with and without stress in 85.6% of cases. Furthermore, our narrative synthesis showed that the performance of wearable AI in detecting students with stress (75.5%) is comparable with its performance in detecting students without stress (74.4%). We considered that the performance of wearable AI in detecting stress among students is suboptimal given that many previous reviews have shown a higher performance of AI in detecting other diseases or disorders, such as cancers [65-68], heart diseases [69,70], ear diseases [71], and ophthalmic disorders [72,73]. Moreover, relying solely on pooled mean accuracy is insufficient for drawing definitive conclusions regarding the performance of wearable AI; thus, we considered sensitivity and specificity when evaluating its performance, which were lower than the pooled accuracy.

The subgroup analyses in this review demonstrated that the performance of wearable AI is moderated by 5 factors. The first factor is the type of wearable device used to collect digital biomarkers. To be more precise, electrodes have higher accuracy (94.4%) in detecting stress than smartwatches (83.6%) and smart bands (69.5%). This can be explained by the practice of typically positioning electrodes on the most suitable areas of the body

XSL•FO
RenderX

for gathering specific biomarkers, thereby collecting more accurate data, as opposed to most commercial and fashionable popular devices being restricted to wrists.

The second moderator identified in this review is the placement of wearable devices on the body. To elaborate, studies that used wrist-worn devices exhibited lower accuracy in detecting stress among students compared with those that used wearable devices placed on other parts of the body (94.4% vs 77.3%). This can be attributed to 4 factors. First, all studies that used non–wrist-worn devices collected data from more than one area of the body (eg, chest, fingers, and palm) and used electrodes, whereas all studies that used wrist-worn devices collected data from only one area (ie, wrist) and used smartwatches or smart bands. Second, the placement of wearable devices on the wrist may lead to less accurate data collection because of the frequent arm movements. Third, all studies that used non–wrist-worn devices collected EDA data, whereas none of the studies that used wrist-worn devices collected EDA data. Finally, the ground truth in all studies that used non–wrist-worn devices was assessed objectively, whereas the ground truth in approximately half of the studies that used wrist-worn devices was identified subjectively.

The third moderator identified in this review is the number of stress classes that wearable AI endeavors to detect. Surprisingly, our analysis revealed that wearable AI exhibits greater accuracy in classifying more than 2 stress levels (94.4%) as opposed to just 2 stress classes (77.3%). This may be attributed to the same 4 aforementioned factors given that the estimates used in the subgroup meta-analysis to assess the performance of wearable AI in classifying more than 2 stress levels were the exact estimates used in the meta-analysis to evaluate the performance of non–wrist-worn devices.

The fourth moderator found in this review is the ground truth. Specifically, the efficacy of wearable AI in stress detection tends to be superior when the ground truth is determined through objective assessments (eg, the experimental context), in contrast to when it relies on subjective assessments (ie, self-reported questionnaires; 93.3% vs 70.6%). This is expected as objective assessments are generally more reliable and accurate than subjective assessments.

The last moderator identified in this review is the data set size. Unexpectedly, our subgroup meta-analyses demonstrated that the accuracy of wearable AI in detecting stress is higher when using a data set size of ≤100 (94.7%) as opposed to data sets with sizes of >100 (76.9%). Although this seems odd at first glance, the reason could be that the ground truth in all studies that used a data set size of ≤100 was examined objectively, whereas the ground truth in approximately half of the studies that used a data set size of >100 was identified subjectively.

As mentioned in the *Introduction* section, none of the previous reviews statistically synthesized the results of previous studies. Thus, we calculated the traditional average of accuracy estimates reported in the studies included in these reviews. We noticed comparable findings despite the differences between our review and these reviews. To elaborate, although the pooled mean in this review was 85.6%, the average accuracy estimates in the previous reviews were 87.6% (50%-100%) [28], 87% (70.8%-99%) [34], 86% (60%-100%) [15], 85.4% (64.5%-98.3%) [35], 84.2% (51.2%-100%) [24], and 82.9% (53%-99%) [13]. Furthermore, this review showed findings comparable with those of previous studies on the performance of wearable AI in detecting depression [74] and anxiety [75]. Specifically, the pooled mean accuracies of wearable AI in detecting depression and anxiety were 89% [74] and 82% [75], respectively.

Research and Practical Implications

This review found that wearable AI holds promise as a valuable tool for detecting stress among students. Nonetheless, we are unable to endorse the immediate integration of wearable AI into clinical and academic practices for the following reasons: (1) its efficiency in detecting stress among students is currently suboptimal, (2) the number of participants was small (≤50) in 74% (14/19) of the studies, (3) the accuracy estimates included in all meta-analyses were extracted from a few studies (≤6), and (4) only 21% (4/19) of the studies were judged to have a low risk of bias in all domains. Hence, it is advisable to use wearable AI in conjunction with other clinical assessments such as self-report questionnaires to detect stress among students.

Most of the studies included in this review (17/19, 89%) targeted undergraduate students, although students at other educational levels are vulnerable to stress. Hence, future research endeavors should recruit students at different educational levels, such as postgraduate and high school students. This review primarily focused on stress among students even though there is a substantial body of research on the effectiveness of wearable AI in identifying stress in other demographic groups. Thus, further systematic reviews are needed to consolidate and synthesize the evidence on the performance of wearable AI in detecting stress across diverse populations.

In this review, all the included studies evaluated the effectiveness of wearable AI in detecting current stress levels rather than predicting the occurrence of stress in the future. Anticipating the onset of stress in the future is equally if not more crucial than identifying the current stress state as this can facilitate the development and implementation of more efficient, timely, and tailored interventions. For this reason, we urge researchers to conduct further studies on the potential of wearable AI for predicting future occurrences of stress.

None of the studies included in this review evaluated the effectiveness of wearable AI in differentiating stress from other mental health conditions such as depression and anxiety, nor did they assess its ability to distinguish between different types of stress (eg, acute stress, chronic stress, emotional stress, physical stress, eustress, distress, and posttraumatic stress disorder). In clinical practice, health care professionals often rely on complex and error-prone diagnostic methods to make distinctions among various patient groups rather than merely distinguishing them from individuals without health issues. Consequently, it is imperative for researchers to investigate how well wearable AI can differentiate between different types of stress and distinguish individuals with stress from those who have other mental disorders exhibiting signs and symptoms similar to those of stress.

XSL•FO
RenderX

Although previous studies have examined the performance of wearable AI based on different factors (eg, algorithms, data types, and number of stress classes), they did not assess its performance based on the placement of the wearable device (eg, wrist, chest, or fingers), the methods used to induce stress (eg, Stroop test, arithmetic tasks, and physical exercises), or the methods used to assess the ground truth (eg, subjective and objective assessments). We strongly encourage researchers to consider these factors in future studies. Furthermore, none of the previous studies have used biological samples as the ground truth, although they are considered as the gold-standard test of stress. For example, elevated cortisol levels in blood, saliva, or urine samples can be indicative of chronic stress. Such gold-standard tests, especially those that can be measured easily and noninvasively, should be considered in future studies.

Even though this review included 19 studies, we were able to include results from only 6 (32%) in our meta-analyses as the remaining studies did not provide sufficient details essential for conducting meta-analyses (eg, confusion matrices, number of cases, and number of controls). In addition, these studies did not present multiple performance measures such as accuracy, sensitivity, specificity, and F_1-score, which are necessary for us to estimate the required details. Hence, researchers should include the aforementioned details in their reports to facilitate the conduct of meta-analyses by others.

A total of 74% (14/19) of the studies had a limited number of participants, with ≤50 individuals involved. This could have posed challenges in identifying potential performance differences in wearable AI when conducting subgroup meta-analyses. In addition, this could have acted as a hindrance for the development of algorithms that rely on a substantial amount of data. To address these issues, it is essential for researchers to recruit a larger number of participants, ensuring robust statistical power and enabling the use of more advanced and efficient algorithms.

Limitations

This review excluded studies that assessed the performance of (1) nonwearable devices, handheld devices, near-body wearable devices, in-body wearable devices, wearable devices wired to nonwearable devices, and wearables requiring expert oversight; (2) wearable AI in detecting other mental disorders (eg, attention-deficit/hyperactivity disorder, bipolar disorder, and schizophrenia); (3) wearable AI in detecting stress among populations other than students; and (4) wearable AI in predicting outcomes of stress treatment. Consequently, the generalizability of our findings to such devices, disorders, populations, and outcomes is limited.

This review could not draw definitive conclusions based on the results of the meta-analyses for several reasons. First, it is probable that we overlooked some studies as we excluded studies that were published in non-English languages and before 2015 and used publicly available data sets (eg, WESAD) or data sets from previous studies. Second, the accuracy estimates included in all meta-analyses were drawn from a limited number of studies (≤6), primarily as the remaining studies lacked the necessary information required for conducting meta-analyses.

Conclusions

Wearable AI holds promise as a valuable tool for detecting stress among students. However, it is not sufficiently ready to be integrated into clinical and academic practices given its suboptimal performance. Until further evidence demonstrates an ideal performance of wearable AI, it should be used in conjunction with other clinical assessments to help detect stress among students. Given the limitations of the current evidence, researchers should investigate how well wearable AI can differentiate between different types of stress and distinguish individuals with stress from those who have other mental disorders exhibiting signs and symptoms similar to those of stress. Furthermore, future research should recruit students at different educational levels, such as postgraduate and high school students. There is also a need to conduct studies on the potential of wearable AI for predicting future occurrences of stress. Future studies should compare the performance of wearable AI based on the placement of the wearable device, the methods used to induce stress, and the methods used to assess the ground truth. Finally, researchers should provide detailed results to facilitate the conduct of meta-analyses by others.

Acknowledgments

No funding was obtained for this study. No custom code or mathematical algorithm was used in this study.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

A Abd-alrazaq, A Ahmed, R AlSaad, and SA developed the protocol with guidance from and under the supervision of JS. A Abd-alrazaq searched the electronic databases and conducted backward and forward reference list checking. The study selection process, data extraction, and risk-of-bias assessment were carried out by MA and R Ahmad under the supervision of R AlSaad and A Abd-alrazaq. Data synthesis was conducted by A Abd-alrazaq. The *Introduction* section was written by MA and RD. The *Methods* section was written by A Ahmed. The *Results* section was written by A Abd-alrazaq and R AlSaad. The *Discussion* and *Conclusions* sections were written by A Abd-alrazaq. The manuscript was critically revised for important intellectual content by all authors. All authors approved the manuscript for publication and agree to be accountable for all aspects of the work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy) checklist.
[DOC File , 68 KB - jmir_v26i1e52622_app1.doc]

Multimedia Appendix 2
Search strategy.
[DOCX File , 42 KB - jmir_v26i1e52622_app2.docx]

Multimedia Appendix 3
Data extraction form.
[DOCX File , 22 KB - jmir_v26i1e52622_app3.docx]

Multimedia Appendix 4
Modified version of the Quality Assessment of Diagnostic Accuracy Studies–Revised.
[DOCX File , 26 KB - jmir_v26i1e52622_app4.docx]

Multimedia Appendix 5
Characteristics of each included study.
[DOCX File , 35 KB - jmir_v26i1e52622_app5.docx]

Multimedia Appendix 6
Features of wearable devices.
[DOCX File , 34 KB - jmir_v26i1e52622_app6.docx]

Multimedia Appendix 7
Features of the artificial intelligence algorithms.
[DOCX File , 38 KB - jmir_v26i1e52622_app7.docx]

Multimedia Appendix 8
Reviewers' judgments about each risk of bias and applicability domain for each included study.
[DOCX File , 42 KB - jmir_v26i1e52622_app8.docx]

References

1. Liu Z, Xie Y, Sun Z, Liu D, Yin H, Shi L. Factors associated with academic burnout and its prevalence among university students: a cross-sectional study. BMC Med Educ 2023 May 06;23(1):317 [FREE Full text] [doi: 10.1186/s12909-023-04316-y] [Medline: 37149602]
2. Yang G, Wang Z, Wu W. Social comparison orientation and mental health: the mediating role of psychological capital. soc behav pers 2021 Jan 06;49(1):1-11 [FREE Full text] [doi: 10.2224/sbp.9767]
3. Kumaraswamy N. Academic stress, anxiety and depression among college students: a brief review. Int Rev Soc Sci humanit 2013;5(1):135-143.
4. Rith-Najarian LR, Boustani MM, Chorpita BF. A systematic review of prevention programs targeting depression, anxiety, and stress in university students. J Affect Disord 2019 Oct 01;257:568-584. [doi: 10.1016/j.jad.2019.06.035] [Medline: 31326690]
5. Mofatteh M. Risk factors associated with stress, anxiety, and depression among university undergraduate students. AIMS Public Health 2021;8(1):36-65 [FREE Full text] [doi: 10.3934/publichealth.2021004] [Medline: 33575406]
6. Bedewy D, Gabriel A. Examining perceptions of academic stress and its sources among university students: the perception of academic stress scale. Health Psychol Open 2015 Jul 30;2(2):2055102915596714 [FREE Full text] [doi: 10.1177/2055102915596714] [Medline: 28070363]
7. Regehr C, Glancy D, Pitts A. Interventions to reduce stress in university students: a review and meta-analysis. J Affect Disord 2013 May 15;148(1):1-11. [doi: 10.1016/j.jad.2012.11.026] [Medline: 23246209]
8. Shiralkar MT, Harris TB, Eddins-Folensbee FF, Coverdale JH. A systematic review of stress-management programs for medical students. Acad Psychiatry 2013 May 01;37(3):158-164. [doi: 10.1176/appi.ap.12010003] [Medline: 23446664]

XSL•FO
RenderX

9. Hobfoll SE. Stress, Culture, and Community: The Psychology and Philosophy of Stress. New York, NY: Springer; 2004.
10. Lazarus RS. From psychological stress to the emotions: a history of changing outlooks. Annu Rev Psychol 1993 Jan;44(1):1-21. [doi: 10.1146/annurev.ps.44.020193.000245] [Medline: 8434890]
11. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. Annu Rev Clin Psychol 2005 Apr 01;1(1):607-628 [FREE Full text] [doi: 10.1146/annurev.clinpsy.1.102803.144141] [Medline: 17716101]
12. Edjah K, Ankomah F, Domey E, Laryea JE. Stress and its impact on academic and social life of undergraduate university students in Ghana: a structural equation modeling approach. Open Educ Stud 2020 Jun;2(1):37-44 [FREE Full text] [doi: 10.1515/edu-2020-0100]
13. Can YS, Arnrich B, Ersoy C. Stress detection in daily life scenarios using smart phones and wearable sensors: a survey. J Biomed Inform 2019 Apr;92:103139 [FREE Full text] [doi: 10.1016/j.jbi.2019.103139] [Medline: 30825538]
14. Aldwin CM. Stress, Coping, and Development: An Integrative Perspective. New York, NY: The Guilford Press; 2009.
15. Elzeiny S, Qaraqe M. Machine learning approaches to automatic stress detection: a review. In: Proceedings of the IEEE/ACS 15th International Conference on Computer Systems and Applications. 2018 Presented at: AICCSA '18; October 28-November 1, 2018; Aqaba, Jordan p. 1-6 URL: https://ieeexplore.ieee.org/document/8612825 [doi: 10.1109/aiccsa.2018.8612825]
16. Stults-Kolehmainen MA, Sinha R. The effects of stress on physical activity and exercise. Sports Med 2014 Jan 13;44(1):81-121 [FREE Full text] [doi: 10.1007/s40279-013-0090-5] [Medline: 24030837]
17. Farrell AK, Simpson JA. Effects of relationship functioning on the biological experience of stress and physical health. Curr Opin Psychol 2017 Feb;13:49-53. [doi: 10.1016/j.copsyc.2016.04.014] [Medline: 28813293]
18. Sandi C. Stress, cognitive impairment and cell adhesion molecules. Nat Rev Neurosci 2004 Dec;5(12):917-930. [doi: 10.1038/nrn1555] [Medline: 15550947]
19. Jonsdottir IH, Nordlund A, Ellbin S, Ljung T, Glise K, Währborg P, et al. Cognitive impairment in patients with stress-related exhaustion. Stress 2013 Mar 17;16(2):181-190. [doi: 10.3109/10253890.2012.708950] [Medline: 22746338]
20. Sandi C, Haller J. Stress and the social brain: behavioural effects and neurobiological mechanisms. Nat Rev Neurosci 2015 May 20;16(5):290-304 [FREE Full text] [doi: 10.1038/nrn3918] [Medline: 25891510]
21. Gradus JL. Prevalence and prognosis of stress disorders: a review of the epidemiologic literature. Clin Epidemiol 2017 May 03;9:251-260 [FREE Full text] [doi: 10.2147/CLEP.S106250] [Medline: 28496365]
22. Vrshek-Schallhorn S, Stroud CB, Mineka S, Hammen C, Zinbarg RE, Wolitzky-Taylor K, et al. Chronic and episodic interpersonal stress as statistically unique predictors of depression in two samples of emerging adults. J Abnorm Psychol 2015 Nov;124(4):918-932 [FREE Full text] [doi: 10.1037/abn0000088] [Medline: 26301973]
23. Kumar S, Bhukar JP. Stress level and coping strategies of college students. J Physic Educ Sport Manag 2013 Jan;4(1):5-11 [FREE Full text]
24. Arsalan A, Anwar S, Majid M. Mental stress detection using data from wearable and non-wearable sensors: a review. arXiv. Preprint posted online on Feb 7, 2022 [FREE Full text]
25. Frisone F, Sicari F, Settineri S, Merlo EM. Clinical psychological assessment of stress: a narrative review of the last 5 years. Clin Neuropsychiatry 2021 Apr;18(2):91-100 [FREE Full text] [doi: 10.36131/cnfioritieditore20210203] [Medline: 34909024]
26. Epel ES, Crosswell AD, Mayer SE, Prather AA, Slavich GM, Puterman E, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169 [FREE Full text] [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
27. Dhama K, Latheef SK, Dadar M, Samad HA, Munjal A, Khandia R, et al. Biomarkers in stress related diseases/disorders: diagnostic, prognostic, and therapeutic values. Front Mol Biosci 2019 Oct 18;6:91 [FREE Full text] [doi: 10.3389/fmolb.2019.00091] [Medline: 31750312]
28. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066 [FREE Full text] [doi: 10.1109/access.2021.3085502]
29. Ahmed A, Aziz S, Abd-Alrazaq A, Farooq F, Sheikh J. Overview of artificial intelligence-driven wearable devices for diabetes: scoping review. J Med Internet Res 2022 Aug 09;24(8):e36010 [FREE Full text] [doi: 10.2196/36010] [Medline: 35943772]
30. Abd-Alrazaq A, AlSaad R, Aziz S, Ahmed A, Denecke K, Househ M, et al. Wearable artificial intelligence for anxiety and depression: scoping review. J Med Internet Res 2023 Jan 19;25:e42672 [FREE Full text] [doi: 10.2196/42672] [Medline: 36656625]
31. Iqbal SM, Mahgoub I, Du E, Leavitt MA, Asghar W. Advances in healthcare wearable devices. Npj Flex Electron 2021 Apr 12;5(1):9 [FREE Full text] [doi: 10.1038/s41528-021-00107-x]
32. Hickey BA, Chalmers T, Newton P, Lin CT, Sibbritt D, McLachlan CS, et al. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: a systematic review. Sensors (Basel) 2021 May 16;21(10):3461 [FREE Full text] [doi: 10.3390/s21103461] [Medline: 34065620]
33. Long N, Lei Y, Peng L, Xu P, Mao P. A scoping review on monitoring mental health using smart wearable devices. Math Biosci Eng 2022 May 27;19(8):7899-7919 [FREE Full text] [doi: 10.3934/mbe.2022369] [Medline: 35801449]

XSL•FO
RenderX

34. Gradl S, Wirth M, Richer R, Rohleder N, Eskofier B. An overview of the feasibility of permanent, real-time, unobtrusive stress measurement with current wearables. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2019 Presented at: PervasiveHealth '19; May 20-23, 2019; Trento, Italy p. 360-365 URL: https://dl.acm.org/doi/10.1145/3329189.3329233 [doi: 10.1145/3329189.3329233]
35. Vos G, Trinh K, Sarnyai Z, Rahimi Azghadi M. Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review. Int J Med Inform 2023 May;173:105026 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105026] [Medline: 36893657]
36. McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
37. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
38. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
39. Moltu C, Stefansen J, Svisdahl M, Veseth M. Doing Meta-Analysis With R: A Hands-On Guide. Boca Raton, FL: CRC Press; 2012.
40. Borenstein M, Hedges LV, Higgins PT, Rothstein HR. Introduction to Meta‐Analysis. Hoboken, NJ: John Wiley & Sons; 2009.
41. Freeman MF, Tukey JW. Transformations related to the angular and the square root. Ann Math Statist 1950 Dec;21(4):607-611 [FREE Full text] [doi: 10.1214/aoms/1177729756]
42. Schwarzer G. Meta: an R package for meta-analysis. R news 2007;7(3):40-45 [FREE Full text]
43. Assink M, Wibbelink CJ. Fitting three-level meta-analytic models in R: a step-by-step tutorial. Quant Method Psychol 2016 Oct 01;12(3):154-174 [FREE Full text] [doi: 10.20982/tqmp.12.3.p154]
44. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
45. Deeks J, Higgins JP, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2019:241-284.
46. Altaf H, Ibrahim SN, Olanrewaju RF. Non invasive stress detection method based on discrete wavelet transform and machine learning algorithms. In: Proceedings of the 11th IEEE Symposium on Computer Applications & Industrial Electronics. 2021 Presented at: ISCAIE '21; April 3-4, 2021; Penang, Malaysia p. 106-111 URL: https://ieeexplore.ieee.org/document/9431805 [doi: 10.1109/iscaie51753.2021.9431805]
47. Aqajari SA, Labbaf S, Tran PH, Nguyen B, Mehrabadi MA, Levorato M, et al. Context-aware stress monitoring using wearable and mobile technologies in everyday settings. medRxiv. Preprint posted online on Apr 24, 2023 [FREE Full text] [doi: 10.1101/2023.04.20.23288181]
48. Can YS, Chalabianloo N, Ekiz D, Ersoy C. Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study. Sensors (Basel) 2019 Apr 18;19(8):1849 [FREE Full text] [doi: 10.3390/s19081849] [Medline: 31003456]
49. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inform 2020 Dec;112:103610 [FREE Full text] [doi: 10.1016/j.jbi.2020.103610] [Medline: 33137470]
50. de Arriba Pérez F, Santos-Gago JM, Caeiro-Rodríguez M, Fernández Iglesias MJ. Evaluation of commercial-off-the-shelf wrist wearables to estimate stress on students. J Vis Exp 2018 Jun 16(136):57590 [FREE Full text] [doi: 10.3791/57590] [Medline: 29985338]
51. Egilmez B, Poyraz E, Zhou W, Memik G, Dinda P, Alshurafa N. UStress: understanding college student subjective stress using wrist-based passive sensing. In: Proceedings of the 2021 International Conference on Pervasive Computing and Communications Workshops. 2021 Presented at: PerCom '21; March 22-26, 2021; Kona, HI p. 673-678 URL: https://ieeexplore.ieee.org/document/7917644 [doi: 10.1109/percomw.2017.7917644]
52. Hasanbasic A, Spahic M, Bosnjic D, Adzic HH, Mesic V, Jahic O. Recognition of stress levels among students with wearable sensors. In: Proceedings of the 18th International Symposium on INFOTEH-JAHORINA. 2019 Presented at: INFOTEH '19; March 20-22, 2019; East Sarajevo, Bosnia and Herzegovina p. 1-4 URL: https://ieeexplore.ieee.org/document/8717754 [doi: 10.1109/infoteh.2019.8717754]
53. Islam TZ, Liang PW, Sweeney F, Pragner C, Thiagarajan JJ, Sharmin M, et al. College life is hard! - shedding light on stress prediction for autistic college students using data-driven analysis. In: Proceedings of the 45th Annual Computers, Software, and Applications Conference. 2021 Presented at: COMPSAC '21; July 12-16, 2021; Madrid, Spain p. 428-437 URL: https://ieeexplore.ieee.org/document/9529720 [doi: 10.1109/compsac51774.2021.00066]

XSL•FO
RenderX

54. Jafarlou S, Lai J, Azimi I, Mousavi Z, Labbaf S, Jain RC, et al. Objective prediction of next-day's affect using multimodal physiological and behavioral data: algorithm development and validation study. JMIR Form Res 2023 Mar 15;7:e39425 [FREE Full text] [doi: 10.2196/39425] [Medline: 36920456]
55. Jiang JY, Chao Z, Bertozzi AL, Wang W, Young SD, Needell D. Learning to predict human stress level with incomplete sensor data from wearable devices. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019 Presented at: CIKM '19; November 3-7, 2019; Beijing, China p. 2773-2781 URL: https://dl.acm.org/doi/10.1145/3357384.3357831 [doi: 10.1145/3357384.3357831]
56. Moridani MK, Mahabadi Z, Javadi N. Heart rate variability features for different stress classification. Bratisl Lek Listy 2020;121(9):619-627. [doi: 10.4149/BLL_2020_107] [Medline: 32990009]
57. Nagar P, Sethia D. Brain mapping based stress identification using portable EEG based device. In: Proceedings of the 11th International Conference on Communication Systems & Networks. 2019 Presented at: COMSNETS '19; April 23-24, 2019; Bengaluru, India p. 601-606 URL: https://ieeexplore.ieee.org/document/8711009 [doi: 10.1109/comsnets.2019.8711009]
58. Padmaja B, Prasad VV, Sunitha KV, Reddy CS, Anil CH. DetectStress: a novel stress detection system based on smartphone and wireless physical activity tracker. In: Proceedings of the 1st International Conference on Artificial Intelligence and Cognitive Computing. 2018 Presented at: AICC '18; November 24-25, 2018; Virtual Event p. 67-80. [doi: 10.1007/978-981-13-1580-0_7]
59. Sandulescu V. Stress detection using wearable physiological sensors. In: Proceedings of the 2015 International Work-Conference on the Interplay Between Natural and Artificial Computation on Artificial Computation in Biology and Medicine. 2015 Presented at: IWINAC '15; June 1-5, 2015; Elche, Spain p. 526-532 URL: http://tinyurl.com/3xjmecjm [doi: 10.1007/978-3-319-18914-7_55]
60. Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, et al. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. J Med Internet Res 2018 Jun 08;20(6):e210 [FREE Full text] [doi: 10.2196/jmir.9410] [Medline: 29884610]
61. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018 Presented at: ICMI '18; October 16-20, 2018; Boulder, CO p. 400-408 URL: https://dl.acm.org/doi/10.1145/3242969.3242985 [doi: 10.1145/3242969.3242985]
62. Silva E, Aguiar J, Reis LP, Sá JO, Gonçalves J, Carvalho V. Stress among Portuguese medical students: the EuStress solution. J Med Syst 2020 Jan 02;44(2):45. [doi: 10.1007/s10916-019-1520-1] [Medline: 31897774]
63. Tiwari S, Agarwal S. A shrewd artificial neural network-based hybrid model for pervasive stress detection of students using galvanic skin response and electrocardiogram signals. Big Data 2021 Dec 01;9(6):427-442. [doi: 10.1089/big.2020.0256]
64. Wu Y, Daoudi M, Amad A, Sparrow L, D'Hondt F. Unsupervised learning method for exploring students' mental stress in medical simulation training. In: Companion Publication of the 2020 International Conference on Multimodal Interaction. 2020 Presented at: ICMI '20; October 25-29, 2020; Virtual Event p. 165-170 URL: https://dl.acm.org/doi/10.1145/3395035.3425191 [doi: 10.1145/3395035.3425191]
65. Zhang SM, Wang YJ, Zhang ST. Accuracy of artificial intelligence-assisted detection of esophageal cancer and neoplasms on endoscopic images: a systematic review and meta-analysis. J Dig Dis 2021 Jun 18;22(6):318-328 [FREE Full text] [doi: 10.1111/1751-2980.12992] [Medline: 33871932]
66. Jiang K, Jiang X, Pan J, Wen Y, Huang Y, Weng S, et al. Current evidence and future perspective of accuracy of artificial intelligence application for early gastric cancer diagnosis with endoscopy: a systematic and meta-analysis. Front Med (Lausanne) 2021 Mar 15;8:629080 [FREE Full text] [doi: 10.3389/fmed.2021.629080] [Medline: 33791323]
67. Dumitrescu EA, Ungureanu BS, Cazacu IM, Florescu LM, Streba L, Croitoru VM, et al. Diagnostic value of artificial intelligence-assisted endoscopic ultrasound for pancreatic cancer: a systematic review and meta-analysis. Diagnostics (Basel) 2022 Jan 25;12(2):309 [FREE Full text] [doi: 10.3390/diagnostics12020309] [Medline: 35204400]
68. Morozov A, Taratkin M, Bazarkin A, Rivas JG, Puliatti S, Checcucci E, Young Academic Urologists (YAU) Working Group in Uro-technology of the European Association of Urology. A systematic review and meta-analysis of artificial intelligence diagnostic accuracy in prostate cancer histology identification and grading. Prostate Cancer Prostatic Dis 2023 Dec 25;26(4):681-692. [doi: 10.1038/s41391-023-00673-3] [Medline: 37185992]
69. Hoodbhoy Z, Jiwani U, Sattar S, Salam R, Hasan B, Das JK. Diagnostic accuracy of machine learning models to identify congenital heart disease: a meta-analysis. Front Artif Intell 2021 Jul 8;4:708365 [FREE Full text] [doi: 10.3389/frai.2021.708365] [Medline: 34308341]
70. Li XM, Gao XY, Tse G, Hong SD, Chen KY, Li GP, et al. Electrocardiogram-based artificial intelligence for the diagnosis of heart failure: a systematic review and meta-analysis. J Geriatr Cardiol 2022 Dec 28;19(12):970-980 [FREE Full text] [doi: 10.11909/j.issn.1671-5411.2022.12.002] [Medline: 36632204]
71. Habib AR, Kajbafzadeh M, Hasan Z, Wong E, Gunasekera H, Perry C, et al. Artificial intelligence to classify ear disease from otoscopy: a systematic review and meta-analysis. Clin Otolaryngol 2022 May 15;47(3):401-413 [FREE Full text] [doi: 10.1111/coa.13925] [Medline: 35253378]
72. Bulloch G, Seth I, Zhu Z, Stapleton FJ, Fung J, Tan Z, et al. Artificial intelligence applications and performance metrics in ophthalmology: a systematic review and meta-analysis. Vestn Oftalmol Vis Sci 2023;6(1):1037 [FREE Full text]

XSL•FO
RenderX

73. Wu JH, Nishida T, Weinreb RN, Lin JW. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis. Am J Ophthalmol 2022 May;237:1-12. [doi: 10.1016/j.ajo.2021.12.008] [Medline: 34942113]
74. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023 May 05;6(1):84 [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
75. Abd-Alrazaq A, AlSaad R, Harfouche M, Aziz S, Ahmed A, Damseh R, et al. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. J Med Internet Res 2023 Nov 08;25:e48754 [FREE Full text] [doi: 10.2196/48754] [Medline: 37938883]

Abbreviations

AI: artificial intelligence
EDA: electrodermal activity
HR: heart rate
HRV: heart rate variability
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies–Revised
WESAD: Wearable Stress and Affect Detection

Edited by T de Azevedo Cardoso; submitted 10.09.23; peer-reviewed by M Zhou, Y Su, E Ikefuama; comments to author 19.10.23; revised version received 24.10.23; accepted 19.12.23; published 31.01.24.

Please cite as:
Abd-alrazaq A, Alajlani M, Ahmad R, AlSaad R, Aziz S, Ahmed A, Alsahli M, Damseh R, Sheikh J
The Performance of Wearable AI in Detecting Stress Among Students: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e52622
URL: https://www.jmir.org/2024/1/e52622
doi:10.2196/52622
PMID:38294846

©Alaa Abd-alrazaq, Mohannad Alajlani, Reham Ahmad, Rawan AlSaad, Sarah Aziz, Arfan Ahmed, Mohammed Alsahli, Rafat Damseh, Javaid Sheikh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digitally Delivered Interventions to Improve Nutrition Behaviors Among Resource-Poor and Ethnic Minority Groups With Type 2 Diabetes: Systematic Review

Nazgol Karimi[1], PhD; Rachelle Opie[1], PhD; David Crawford[1], PhD; Stella O'Connell[1], PhD; Kylie Ball[1], PhD

Institute for Physical Activity and Nutrition, School of Exercise and Nutrition Sciences, Deakin University, Melbourne, Australia

Corresponding Author:
Nazgol Karimi, PhD
Institute for Physical Activity and Nutrition, School of Exercise and Nutrition Sciences
Deakin University
221 Burwood Highway, Burwood
Melbourne, 3125
Australia
Phone: 61 410825300
Email: nazgol.karimi@gmail.com

Abstract

Background: Resource-poor individuals, such as those with a low income, are disproportionately affected by diabetes and unhealthy eating patterns that contribute to poor disease self-management and prognosis. Digitally delivered interventions have the potential to address some of the barriers to healthy eating experienced by this group. However, little is known about their effectiveness in disadvantaged populations.

Objective: This systematic review is conducted to assess the effectiveness of digitally delivered interventions in improving nutritional behaviors and nutrition - related health outcomes among disadvantaged people with type 2 diabetes (T2D).

Methods: MEDLINE complete, Global Health, Embase, CINAHL complete, Informit Health, IEEE Xplore, and Applied Science and Technology Source databases were searched for studies published between 1990 and 2022 on digitally delivered nutrition interventions for disadvantaged people with T2D. Two reviewers independently assessed the studies for eligibility and determined the study quality using the Cochrane Risk-of-Bias Assessment Tool. The Behavioral Change Technique Taxonomy V1 was used to identify behavior change techniques used in the design of interventions.

Results: Of the 2434 identified records, 10 (0.4%), comprising 947 participants, met the eligibility criteria and were included in the review. A total of 2 digital platforms, web and messaging services (eg, SMS text messaging interventions or multimedia messaging service), were used to deliver interventions. Substantial improvements in dietary behaviors were reported in 5 (50%) of the 10 studies, representing improvements in healthier food choices or increases in dietary knowledge and skills or self-efficacy. Of the 10 studies, 7 (70%) examined changes in blood glucose levels, of which 4 (57%) out of 7 achieved significant decreases in hemoglobin A_{1C} levels ranging from 0.3% to 1.8%. The most frequently identified behavior change techniques across all studies were *instruction on how to perform the behavior*, *information about health consequences*, and *social support*.

Conclusions: This review provided some support for the efficacy of digitally delivered interventions in improving healthy eating behaviors in disadvantaged people with T2D, an essential dietary prerequisite for changes in clinical metabolic parameters. Further research is needed into how disadvantaged people with T2D may benefit more from digital approaches and to identify the specific features of effective digital interventions for supporting healthy behaviors among disadvantaged populations.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42020149844; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=149844

(J Med Internet Res 2024;26:e42595) doi:10.2196/42595

KEYWORDS

digitally delivered; eHealth; type 2 diabetes; nutrition; socioeconomically disadvantaged; mobile phone

XSL•FO
RenderX

Introduction

Background

Type 2 diabetes (T2D) is a common and increasingly prevalent noncommunicable disease [1]. This disease, considered an epidemic of the 21st century, is a serious challenge for public health, affecting the health of millions of people worldwide [2]. The burden of T2D is growing, specifically among resource-poor and ethnic minority groups [2,3]. These groups are more likely to experience higher rates of suboptimal glycemic control, diabetes-related complications, hospitalizations, as well as increased mortality and morbidity compared with more advantaged populations [4,5]. T2D and its associated complications pose a huge economic cost to the health system and patients through rising health care costs associated with direct medical costs and loss of work and income [5]. The direct and indirect annual costs of diabetes (with approximately 90% comprising T2D) and its various complications account for >2 million deaths each year and cost >US $827 billion worldwide [6,7].

Diabetes Self-Management

Self-management is considered to be the hallmark of diabetes control [8]. T2D self-management involves a combination of nutritional, lifestyle, and medication therapies to optimize glycemic control and reduce complications associated with diabetes. A total of seven self-care behaviors are important in T2D self-management: (1) healthy eating, (2) being physically active, (3) taking medication, (4) monitoring blood glucose, (5) reducing risk, (6) healthy coping, and (7) problem-solving [9].

Healthy Eating

Healthy eating is a key component of T2D self-management [10]. Evidence shows that improved diet quality in people with T2D can decrease hemoglobin A_{1c} (HbA_{1c}) levels to a similar or even better level than the provision of medication [11]. Healthy eating remains an important element of the overall treatment plan for people with T2D, although medication is necessary. In addition, improved diet quality may prevent obesity, high blood pressure, and abnormal lipid profile, which in turn decrease the risk of cardiovascular diseases in people with T2D [11].

People who are socioeconomically disadvantaged, including those with lower incomes, lower education levels, or lower status occupations, are at an increased risk of unhealthy dietary intake, with diets characterized by fewer fruits and vegetables and higher intakes of less healthy discretionary foods and beverages [12-14]. The less optimal eating behaviors observed among socioeconomically disadvantaged groups have been attributed to a range of factors, including lower levels of nutrition knowledge, lower prioritization of health during food selection, less access to nutrition education, lower social support, limited access to higher quality foods, and cost-related barriers [3,15-19]. Furthermore, these barriers to healthy eating partly explain the poor clinical outcomes and higher incidence of diabetic complications in these groups [20].

Limitations of In-Person Diabetes Self-Management Education

Generally, supporting people in T2D management, including healthy eating, involves one-on-one or group-based, in-person education provided by a health care provider (HCP). These clinical approaches can provide the education and support needed by people with T2D. However, the proportion of patients receiving any type of diabetes education or behavioral support is low globally. For instance, only 11% of patients with T2D in the United Kingdom, 23% to 66% in the United States [21], and 40% in Australia receive diabetes education, and this proportion is even lower for disadvantaged populations [22]. This has been attributed to factors such as the limited availability and accessibility of HCPs, lack of time, travel costs, limited insurance reimbursement for diabetes self-management (DSM) education, or disabilities that make travel to DSM education services challenging [3].

Potential for New Digital Technologies

Digital technologies provide a unique delivery mode for promoting T2D self-management. These include mobile technologies, such as tablets, phones, smartphones, and physical activity tracking devices, to help and improve public health practices [23]. These technologies can provide remote access to information that forms the basis of the current in-person T2D self-management support programs.

Digitally delivered health interventions include any web-based application that uses processing or communication via digital technologies to facilitate ≥1 aspect of self-management. These include technologies that can promote (1) self-tailoring goals without requiring continuous professional input, (2) decision-making, (3) emotional management, (4) resource use, (5) action planning for behavior change, or (6) problem-solving [22]. These interventions provide opportunities to overcome some of the barriers of in-person T2D self-management education such as limited access to HCPs and could be a more effective approach to helping people learn self-management skills [24]. Moreover, digital interventions are easily distributable and not limited to a specific location; therefore, they can be delivered in clinics and community health centers, at home, or on the move at a time of greatest convenience to the users [22,24,25]. In addition, they can potentially offer ongoing support and provide extensive intervention as needed [24]. Furthermore, these interventions can accommodate different literacy levels, learning styles, and user-specific preferences and needs (eg, linguistic or cultural background) [26].

Importance of This Review

Many earlier systematic reviews investigating digitally delivered interventions for T2D self-management reported positive outcomes [22,24-37]. However, in nearly all of these studies, remote technologies were used to aid data collection (eg, records of medication, physical activity, or diet) and as a supplement to other intervention components rather than the core intervention component. Furthermore, to the best of our knowledge, all existing reviews, except the one by Dening et al [31] that specifically focused on the effect of digitally

XSL•FO
RenderX

delivered interventions on nutrition behaviors such as the intake of vegetables and fruits, predominantly combined different target behaviors (eg, medication adherence, diet, and physical activity) [24,25,38,39] or examined single nutrition - related health outcome (eg, weight loss) [40] in general (not specifically disadvantaged) populations.

We were able to identify 2 reviews that focused on evaluating the effects of digitally delivered interventions for disadvantaged people with T2D or a mixed sample of both type 1 and type 2 diabetes [25,26], although the main focus of these reviews was on HbA_{1c} changes, and none of these specifically addressed nutritional outcomes.

Although healthy nutritional behavior is an important part of DSM and nutrition education and support is an urgent need for people with T2D, digitally delivered interventions to date have chiefly focused on overall self-management; although some studies have included a dietary component within the intervention package, the assessment of dietary behavior changes or adherence remains scarce. To date, we are aware of only 1 review by Dening et al [31] that has investigated the effects of digital interventions on changes in dietary behavior in people with T2D, and we could not find any analogous review that targeted the assessment of dietary behavior changes in disadvantaged populations. Conducting systematic reviews with a focus on people of lower socioeconomic position is important because the health intervention features that work well in well-resourced settings may not be available or effective for disadvantaged groups [25,26]. Therefore, a considerable evidence gap remains in investigating the effects of digitally delivered interventions on nutrition behaviors and nutrition-related health outcomes among socioeconomically disadvantaged people with T2D, and this systematic review was conducted to address this gap.

To be effective, T2D self-management interventions need to help people increase their understanding of their condition and alter their adherence patterns to effective self-management behaviors (eg, healthy eating). To achieve this aim, many interventions apply behavioral theories and behavior change techniques (BCTs). BCTs are “observable, replicable, and irreducible” components of interventions [41,42]. Extracting information about the theoretical framework and behavioral content of interventions using an established behavior change taxonomy can provide insight into the active ingredients of digital interventions and can help guide future intervention development. Therefore, as part of the aim of this review, we extended the scope of previous systematic reviews by identifying the behavioral theory and techniques underpinning each intervention.

Methods

Reporting Guidelines

This systematic review was reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [43]. The review protocol was registered at PROSPERO (2019#CRD42020149844). Study searches, selection, and synthesis were guided by the population, intervention, control, and outcomes (PICO) statement described in the *Eligibility Criteria* Section.

Eligibility Criteria

The eligibility criteria for the systematic review and search strategy were defined a priori to minimize bias in the selection of studies. The criteria encompassed various aspects, including population characteristics, intervention types, control parameters, study outcomes, acceptable study types, and intervention duration. Below is a detailed breakdown of each criterion:

- Population: studies were eligible for inclusion if they were original research studies on adult participants aged ≥18 years who were clinically diagnosed with T2D and were from disadvantaged populations (≥50% of the sample). Disadvantaged populations included people of low socioeconomic status (SES; ie, people with low educational level, low income, low occupational status, or people living in low SES residential areas) or members of a racial or ethnic minority group. The latter were included because these groups often experience socioeconomic disadvantage [44], are at increased risk of chronic disease [45], and face significant challenges in accessing health promotion services [46-49]. Studies were deemed ineligible if the participants were children aged <18 years, women with gestational diabetes, or people with prediabetes. No criteria were established for the maximum and minimum number of participants in each study.
- Intervention: all types of mobile- or web-based interventions that delivered nutritional knowledge or skills were considered eligible. We defined interventions as mobile based or web based if participants directly received information and skills from digital devices and interfaced with the World Wide Web. On the basis of this definition, studies that used digital devices only as a data collection tool and did not provide any information or feedback on participants’ inputs through the web or telehealth interventions that only provided access to a health professional without further digital support were not deemed mobile based or web based and were, therefore, ineligible for the review.
  Studies with combined or composite interventions, such as those that provided nutritional information as a component of a multifaceted mobile- or web-based intervention, were considered eligible.
- Control: no limitations were imposed on the control group, and studies with or without a control group were considered eligible.
- Outcomes: intended study outcomes for eligibility were measures focused on nutrition-related behaviors (eg, food intake, nutritional knowledge, and skills) or nutrition-related health outcomes (eg, glycemic control as indicated by HbA_{1c} levels and biomarkers such as weight and BMI). Studies with outcomes related to changes in blood lipid levels, exercise levels, and medication adherence were not included in this review.
- Study type: randomized and nonrandomized controlled trials were considered eligible for inclusion. However, protocol papers, letters, and conference proceedings were

XSL•FO
RenderX

excluded. The reference lists of systematic reviews were searched for relevant papers.

- Intervention duration: no limitation was placed on the duration of the intervention or follow-up period, and studies of any duration were considered eligible for inclusion.

Information Sources

Relevant papers were obtained by systematic search using the following electronic databases: MEDLINE complete, Global Health, Embase, CINAHL complete, Informit Health, IEEE Xplore, and Applied Science and Technology Source. Additional publications were identified from the reference lists of the original papers. In some instances, when we found a relevant conference abstract or protocol paper but not the entire result paper, we contacted the study authors to identify additional studies. Each source was last searched on April 6, 2022.

Search Strategy

The search strategy was developed in cooperation with Deakin University's expert librarian team. The strategy comprised 4 main category blocks with medical subject headings terms related to digital health, healthy eating, SES, and diabetes. The search was limited to peer-reviewed studies that were written in the English language only and published between January 1990 and April 2022. Full search strategies for the MEDLINE complete, Global Health, CINAHL complete, and Applied Science and Technology Source databases can be found in Multimedia Appendix 1.

Selection Process

The reference management software EndNote X8 (Clarivate) [50] was used to merge the search results and remove the duplicate papers. All papers were then exported to a web-based systematic review management tool, Covidence (Veritas Health innovation) [51], for other stages of screening, data extraction, and quality assessment. Using the inclusion and exclusion criteria, all potentially relevant papers were independently screened by 2 reviewers (NK and DC or KB) for eligibility based on titles and abstracts. If deemed potentially eligible, the full-text publication was extracted and reviewed by 2 reviewers (NK and DC or KB) to be included in the systematic review. The other team members were consulted in cases of uncertainty.

Finally, studies that fulfilled the eligibility criteria for the review were examined by extracting data on the key criteria in a tabulated form for comparison, qualitative evaluation, and discussion.

Data Extraction and Synthesis of Results

Data extraction was performed by 2 independent authors (NK and DM or KB). Data were extracted from the final selected studies using the following parameters: basic characteristics (including author, date, country, study design, sample size, gender, age, race and ethnicity, attrition, study location, and delivery modes), intervention characteristics (including the type of digital intervention and intervention components, behavior change theory, BCTs, control group protocol, total study duration, and follow-up), and research outcomes (including dietary intake; nutritional knowledge or skills; diet-related health outcomes, such as glycemic control as indicated by HbA_{1c} levels; and biomarkers, such as weight and BMI). Owing to insufficient data and heterogeneity of the included studies, conducting a formal meta-analysis was not possible. Therefore, the findings of this systematic literature review are presented as a narrative summary.

Reporting Bias Assessment

The methodological quality of each paper was independently assessed by 2 reviewers (NK, KB, or RO). Disagreements were resolved by consensus, consultation with a third reviewer (KB, or DC), or discussion in a group meeting of coauthors where the final decisions were made. The Cochrane Collaboration tool [44] was used to assess the risk of bias in all included studies [52].

Results

Study Selection

The initial search resulted in 2434 studies, of which 674 (27.69%) duplicates were removed and 1760 (72.31%) studies were screened against the title and abstract. Furthermore, of the 1760 studies, 1681 (95.51%) were excluded after preliminary screening for not meeting the inclusion criteria. Of the remaining 79 studies reviewed completely, 10 (13%) met the inclusion criteria and were included in this review. Figure 1 displays the results of study selection and the number of included and excluded studies.

Most studies excluded from the review were telehealth interventions. The reason for their exclusion was that the main aim of these studies was to enable health professional–patient interaction at a distance rather than providing patients with knowledge or skills to improve healthy eating behaviors as part of T2D self-management education. Interventions that used mobile apps only for daily recording of diet were also excluded because they did not meet our inclusion criteria that participants received educational information and behavioral change support through the World Wide Web.

Studies with general populations only (eg, participants with average or higher SES) or with less than half of the participants belonging to low SES groups were the other common group of studies that were excluded. Other reasons for exclusion were studies with no clear dietary intervention components, studies with no diet-related outcomes, or studies focused on children or participants with type 1 diabetes only.

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

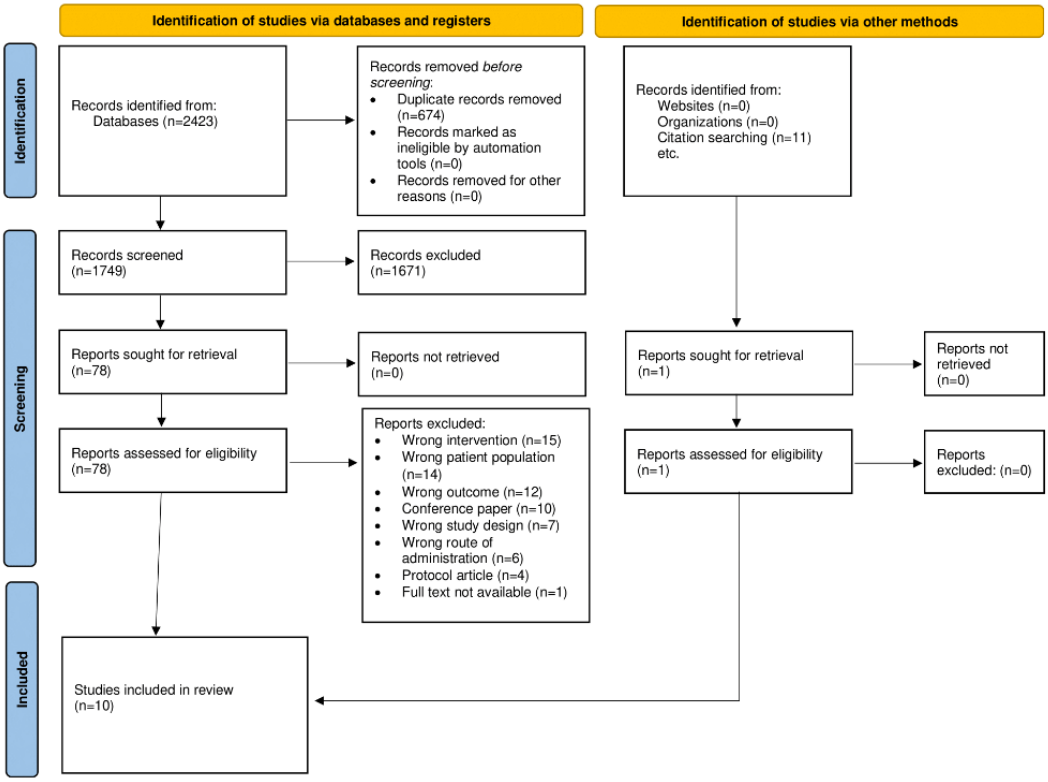

Participants and Study Characteristics

Table 1 presents the characteristics of all the included studies. Out of 10 studies, most (n=7, 70%) were conducted in the United States [53-57], with 1 (10%) study conducted in the Kingdom of Saudi Arabia [58], 1 (10%) in Iran [59], and 1 (10%) in Mexico [60]. The participants of all studies were mainly recruited from the community or primary health care centers, apart from 20% (2/10) of the studies that recruited participants from a hospital emergency department [53,59]. The sample sizes ranged between 13 [55] and 506 participants [61]. The mean age of participants ranged from 45.4 years [53] to 55.8 years [57]. Regarding the length of interventions, 10% (1/10) of the studies had 12 months intervention period and 3-month follow-up period [61], 40% (4/10) of the studies had 6 months intervention period [56,57,60,62], and the remaining studies (5/10, 50%) had intervention periods lasting <3 months or for a single session [53-55,58,59] without any follow-up.

Text messaging, in the form of SMS or multimedia messaging service (MMS), and websites were the 2 digital platforms used to deliver the interventions. In 5 (50%) of the 10 included studies, the intervention was administered via text messaging only [53,54,59,61,62]. In these interventions, educational materials and regular reminders to eat healthier food options and to perform exercises were sent to participants via SMS or MMS. The option for participants to send text messages to providers or educators was also a feature of these studies. In 4 (40%) of the 10 included studies [55-58], website platforms were used to deliver T2D self-management materials. Access to tailored educational materials, videos and links to other resources, and recommended websites were some of the features of these studies. For example, 1 of these studies designed an interactive website [56], in which participants could record the type and amount of food they consumed, and the system provided them with feedback on the calories and grams of proteins, fats, and carbohydrates in their food [56]. Another study incorporated a mix of 7 weekly in-person sessions and 6 months of delivering SMS text messages and MMS messages to reinforce content and promote participants' intention and maintenance of behavior change [60].

Participants' access to computers, mobile phones, and the internet was a barrier addressed in some of the studies. For example, in the study by Robertson et al [56], participants in the intervention group received a computer and 6 months of internet connection. Similarly, in the study by Ruggeiro et al [57], all participants in the intervention group were provided with the same type of laptop to minimize barriers to access. Furthermore, support for internet access was provided where necessary. By contrast, in 4 (40%) of the 10 studies [53,54,59,60], participants were required to own a mobile phone for inclusion in the intervention.

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
Aligholipour et al [59], 2019; Iran	2-arm randomized controlled trial for a period of 3 months	• Having T1D[a] • No severe heart disease • No renal disease • No admission because of ketoacidosis and nonketogenic hyperosmolar syndrome • No severe visual, hearing, and cognitive defect • Resident of Tabriz or near villages • Having a cell phone with the Telegram app	• N=66 (33.3% female), intervention group (n=33) and control group (n=33), and mean age 44.91 (SD 11.8) years • T2D[b]%: ND[c] • Study setting or recruitment site: emergency department or the outpatient clinic of the hospital • Race and ethnicity: 100% Iranian • Attrition rate: 4.54%	Mobile phone: MMS[d]	• Intervention delivery: MMS+nurse diabetes educator (as a support) • Program name: ND • Intervention: at least 2 daily MMS in the form of text, image, video, or GIF sent through a specific Telegram channel, to educate regarding nutrition, exercise, insulin therapy, blood sugar monitoring, foot care, and prevention of diabetes complications	• Tailored design (program designed for adults with low literacy) • Personalized content (tailored feedback and training messages were provided by a nurse diabetes educator)	Usual care (3 in-person educational session by a nurse diabetes educator)
Arora et al [53], 2012; United States	1-arm nonrandomized pilot trial for a period of 3 weeks	• Age ≥18 years • T1D or T2D • Having an SMS text message–capable mobile phone and knowing how to receive SMS text messages • Ability to speak and read English or Spanish	• N=23 (39% female), intervention group (n=23), control group (n=0), and mean age 45.4 (SD 8.3) years • T2D%: 70% • Study setting and recruitment site: emergency department of the hospital • Race and ethnicity: 70% Latinx or Hispanic • Attrition rate: 0%	Mobile phone: SMS text messages	• Intervention delivery: fully automated • Program name: TExT-MED • Intervention: a total of 3 unidirectional daily SMS text messages to deliver educational and motivational material, medication reminders, and healthy living challenges • Trivia questions with answers were sent 1 hour after the SMS text messages and links were provided to access free diabetes management tools.	• Tailored digital design (English or Spanish)	N/A[e]

XSL•FO
RenderX

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
Burner et al [54], 2017; United States	2-arm randomized controlled trial for a period of 3 months	• HbA_{1c}[f] level ≥8 • Having a text-capable mobile phone • Ability to send and receive SMS texts messages	• N=44 (50% female), Intervention group (n=22) and control group (n=22), and mean age ND • T2D%: ND • Study setting or recruitment site: medical Center • Race and ethnicity: 77% Latinx • Attrition rate: 18%	Mobile phone:text messages	• Intervention delivery: fully automated • Program name: TExT-MED+FANS • Intervention: the program consisted of 2 curricula, 1 for patients and another for supporters (eg, family members or friends), and sent 2 unidirectional daily SMS text messages delivering education and motivation (1 per week) medication reminders (3 per week), healthy living challenges (2 per week), and trivia questions (2 per week) • The pair of messages was sent to the patient and the supporter synchronously.	• Tailored digital design (English or Spanish)	A total of 2 SMS text messages daily for patients plus a pamphlet for supporter (no SMS text message for supporters)
Fortmann et al [62], 2017; United States; the study provided phone to participants for the intervention	2-arm randomized controlled trial for a period of 6 months	• Age 18-75 years • T2D • HbA_{1c}≥7.5% • Spanish-speaking and English-speaking Hispanic • Uninsured or underinsured	• N=126 (74.6% female), Intervention group (n=63) and control group (n=63), and mean age 48.4 (SD 9.0) years • T2D%: 100% • Study setting or recruitment site: a network of federally qualified health centers • Race and ethnicity: 100% Hispanic • Attrition rate: 10.3%	Mobile phone: SMS text messages	• Intervention delivery: fully automated • Program name: Dulce digital • Intervention: a total of 2-3 SMS text messages per day (covering topics such as diabetes and its complications; the role of diet, exercise, and medication; and the importance of self-monitoring) were sent at study start, with the frequency tapering over 6 months.	• Tailored content (culturally sensitive content and recipes for Hispanic population)	Usual care including visit with primary care physician, certified diabetes educator, and group DSM[g]
				Website			

XSL•FO
RenderX

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
Moussa et al [58], 2013; Kingdom of Saudi Arabia	2-arm randomized controlled trial for a period of 1 month	• Age 40-65 years • T1D or T2D • Low diabetes literacy	• N=46 (60.9% female), Intervention group (n=23) and control group (n=23), and mean age 52 (SD 7.99) years) • T2D%: 78% • Study setting or recruitment site: free health clinic and community health center • Race and ethnicity: 100% African American • Attrition rate: 0%		• Intervention delivery: trained graduate research assistants+website • Program name: eCare We Care • Intervention: the program provided 4-week web-based diabetes education sessions delivering introduction to diabetes, eye complications, foot care, and meal planning plus providing links to freely accessible and credible government health information websites.	• Tailored digital design (designed for adults with low health and electronic literacy with simple language, option for audio, clearly labeled buttons, and text and arrows for easy navigation)	Paper-based and text-only tutorial
Nelson et al [61], 2021; United States	3-arm randomized controlled trial with a 12-month intervention perio+3 months postintervention follow-up	• Adults • T2D • Use daily diabetes medication		Mobile phone: SMS text messages		• Personalized content (tailored to participants' self-identified barriers to medication adherence and their prescribed diabetes medications)	Nonintervention

XSL•FO
RenderX

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
			• N=506 (54.2% female), intervention group (n=139) and control group (n=135), and mean age of intervention group 55.8 (SD 9.8) years • T2D%: 100% • Study setting or recruitment site: community health center locations and primary care locations • Race and ethnicity: non–Hispanic White 48%, non–Hispanic Black 39.3%, non–Hispanic other races 6.3%, and Hispanic 6.3% • Attrition rate: 4.8%		• Intervention delivery: fully automated • Program name: REACH • Intervention: the program included daily interactive SMS text messages and tailored SMS text messages addressing medication adherence and nontailored SMS text messages supporting other self-care behaviors (ie, diet, exercise, and self-monitoring of blood glucose) for the first 6 months. Then, 3-4 self-care promotion 1-way SMS text messages were sent each week, and 1 weekly interactive SMS text message was sent until the intervention period ended (if a participant chose the low-dose option). • Half of the participants assigned to receive REACH also received FAMS for the first 6 months. FAMS included monthly phone coaching to set diabetes self-care goals and improve family and friend involvement in self-care and the option to invite an adult friend or family member to receive SMS text messages about self-care goals.		
Porter et al [55], 2009; United States	1-arm nonrandomized trial for 1 session	• T2D • First language Spanish		Website			N/A

XSL•FO
RenderX

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
			• Intervention group (N=13) • T2D%: 100% • Race and ethnicity: 100% Hispanic • Attrition rate: 0%		• Intervention delivery: fully automated • Program name: Your Guide to Diet and Diabetes • Intervention: the program delivered information about general overview of diabetes and food, eating for target blood glucose levels, and eating for cardiovascular health plus provided links to other resources and recommended websites.	• Tailored content (culturally sensitive content and recipes for Hispanic population) • Tailored digital design (Spanish)	
Robertson et al [56], 2007; United States	2-arm randomized controlled trial for a period of 6 months	• Age ≥18 years • No renal failure • English literate • Able to provide informed consent	• N=35 (51.4% female), intervention group (n=29) and control group (n=19), and mean age 52.5 (SD 8.6) years • T2D%: 100% • Study setting or recruitment site: Northern Plains Indian reservation • Race and ethnicity: 100% American Indian • Attrition rate: 12.12%	Website	• Intervention delivery: interactive website • Program name: Keya Tracker • Intervention: the program provided materials about physical activity, nutrition, cultural activity, and social activity. Participants recorded their exercise and food consumption daily (for minimum 3 times per week) and recorded their cultural and social data once a week. • Keya Tracker provided feedback on the calories and grams of proteins, fats, and carbohydrates of participants' food choices.	• Tailored content (culturally sensitive content with input from Northern Plains' Indian tribal members about a list of the specific physical, cultural, and social activities)	Nonintervention
Ruggiero et al [57], 2014; United States	1-arm nonrandomized trial for a period of 6 months			Computer or website			N/A

XSL•FO
RenderX

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
		• T2D • African American • Age ≥18 years • Fluent in English • Receiving medication therapy for diabetes (insulin, oral agents, or both) • Not pregnant or planning a pregnancy during the study period • Able to provide informed consent	• Intervention group (N=41, 71% female) and mean age 55.2 (SD 9.6) years • T2D%: 100% • Race and ethnicity:100% African American • Attrition rate: 33.3%		• Intervention delivery: health professionals+website • Program name: Diabetes Island • Intervention: the program provided a virtual environment for DSM education and support and a study website for communication and tracking • Health professionals (as avatars) provided real-time formal and informal educational sessions. The program had a series of 10 formal presentation sessions, including 6 healthy eating topics led by a registered dietitian, 1 on medication adherence led by a pharmacist, 1 on physical activity led by an exercise researcher, and 2 on the "ABCs" of diabetes and proper foot care led by a nurse certified diabetes educator. Several dietitian-facilitated real-time discussions on healthier eating were offered in the grocery store, fast food restaurant, and home.	• Tailored content (culturally sensitive content [57,60] with input from the multidisciplinary research team and feedback from the advisory committees regarding special needs of this population)	
Whittemore et al [60,63], 2020; Mexico	2-arm randomized waitlist-controlled pilot trial with 7 weekly sessions and 6 months of mHealth[g] intervention	• T2D • Age 21-70 years • Medically stable and able to exercise • HbA_{1c} >7.5% • Receiving health care at a Seguro Popular clinic • Access to a mobile phone		Mobile phone: SMS text messages or MMS			Usual care

XSL•FO
RenderX

Study and country	Study design and intervention duration	Inclusion criteria	Sample characteristics, race or ethnicity, and attrition rate	Technology type	Intervention characteristics	Intervention tailoring (eg, tailored digital design, tailored content, or personalized content)	Control group
			• N=47 (68% female), intervention group (n=26) and control group (n=21), and mean age 55.5 (SD 9.2) years • T2D%: 100% • Race and ethnicity: 100% Mexican • Attrition rate: 6.4%		• Intervention delivery: registered nurse and social worker+SMS text messages • Program name: ¡Sí, Yo Puedo Vivir Sano con Diabetes! • Intervention: 7 weekly sessions to progressively increase the knowledge and abilities of participants to improve their metabolic control through healthier eating, and 6 months of daily SMS text or picture messages to reinforce class content and promote participants' intention and maintenance of behavior change	• Tailored content (culturally relevant content that addressed cultural misconception) • Tailored digital design (designed for adults with low health literacy, SMS text messages designed for third to fourth grade reading level, -used images for most the SMS text messages, and colloquial language)	

[a]T1D: type 1 diabetes.

[b]T2D: type 2 diabetes.

[c]ND: none described.

[d]MMS: multimedia messaging service.

[e]N/A: not applicable.

[f]HbA_{1c}: hemoglobin A_{1c}.

[g]mHealth: mobile health.

Most studies (8/10, 80%) focused on samples belonging to an ethnic minority population [53-59,61,62], and in others (2/10, 20%), the samples were disadvantaged in terms of income level, working status, or food insecurity [59,60]. Three different approaches were used to tailor the interventions to the participants: (1) tailoring the digital design using digital elements or language suitable for people with low literacy or electronic literacy or incorporating highly visual messages, (2) adapting content to the culture of target groups, and (3) providing personalized educational content or feedback. Regarding tailoring to language, 2 (20%) of the 10 included studies offered bilingual intervention components to the participants [53,54], and in 1 (10%) study, educational materials were completely translated from English into Spanish, the first language of the participants [55]. In 5 (50%) of the 10 studies, the educational components were informed through input from the participants, multidisciplinary research teams, and feedback from advisory committees regarding patients' special needs and cultural preferences [55-57,60,62]. Of the 10 studies, 2 (20%) provided personalized content or feedback based on participants' special needs [59,61].

Nutritional Content of Interventions

Of the 10 included studies, 5 (50%) [54,55,61,62,64] were fully automated interventions without HCPs' coaching or support, and they mainly provided general healthy eating advice based on standard diabetes dietary guidelines. In these studies, recommendations and educational materials emphasized setting goals to follow a healthy diet with more vegetables and fruits, reading food labels, choosing foods from the 5 food groups, drinking water, avoiding saturated fats, and planning meals. Because most of these studies were conducted in the United States, the dietary guidelines mentioned more often across these studies were those of the American Diabetes Association and the National Diabetes Education Program. In addition to these guidelines, Porter et al [55] referred participants to the Joslin Diabetes Center website, the National Diabetes Information Clearinghouse website, the National Institutes of Diabetes & Digestive & Kidney Diseases website, the Endocrine Society website, and the Centers for Disease Control and Prevention Diabetes Public Health Resource website. Furthermore, the remaining studies [56-60], besides digitally delivering general

XSL•FO
RenderX

health advice on eating well from credible sources, provided some instruction from HCPs (eg, registered nurses, diabetes educators, or nutritionists) for participants via phone or in-person visits.

Nutritional Outcomes and Measures

Studies have explored various outcomes such as nutrition-related cognitive or behavioral outcomes (including dietary consumption, dietary self-efficacy, nutrition knowledge, and skills), nutrition-related health outcomes (including HbA_{1c} levels, fasting blood sugar [FBS], BMI, weight, and blood lipids), medication adherence, and physical activity levels. However, because this systematic review focuses only on nutrition-specific outcomes, we only selected and summarized outcomes related to nutritional behavior and nutrition-related health outcomes.

The scales used for evaluating these 2 groups of outcomes varied across studies. Arora et al [64] used the Diabetes Knowledge Test developed by the Michigan Diabetes Research Training Center and the Diabetes Empowerment Scale-Short Form. Similarly, Burner et al [54] used the Diabetes Empowerment Scale-Short Form to evaluate changes in self-efficacy and the Problem Areas in Diabetes Scale and the Summary of Diabetes Self-Care Activities (SDSCA) Scale to evaluate the diabetes-related quality of life and healthy behaviors, respectively. The SDSCA scale was also used in the study by Whittemore et al [63], and its revised version was used in the study by Ruggeiro et al [57]. The revised version of the SDSCA scale included all the original items, and additional items to assess specific nutritional content areas emphasized in the program (ie, reading food labels, choosing foods from the 5 food groups, and drinking water). Furthermore, Ruggeiro et al [57] used the Fat-Related Diet Habits Questionnaire to assess dietary intake. Moussa et al [58] and Porter et al [58] used surveys developed by their research teams specific to assessing their intervention content. Robertson et al [56] used the United States Department of Agriculture Nutritional Nutrient Database for Standard Reference to enter participants' food records on the website. Participants entered daily food records, specifying the type and amount of food items consumed. The first and last 3 days of each participant's nutrition entries were analyzed using the Food Processor to evaluate changes in calories and amounts of carbohydrates, protein, and fat intake. Nelson et al [61] used the Personal Diabetes Questionnaire subscale to assess the use of dietary information for decision-making, and Aligholipour et al [59] used the Persian version of Toobert's Self-Care Activities Questionnaire.

Intervention Effects on Nutrition-Related Cognitive or Behavioral Outcomes and Nutrition-Related Health Outcomes

Of the 10 included studies, 9 (90%) reported outcomes in the category of nutrition-related cognitive or behavioral outcomes (Table 2) [54-59,61,63,64]. Of these 9 studies, 2 (22%) reported changes in dietary consumption [53,56], with one 1-arm nonrandomized pilot trial by Arora et al [53] showing positive changes in the intervention participants. In their study, Arora et al [53] reported a 26.5% increase in the number of people reporting eating fruits and vegetables after receiving daily educational, motivational, and reminder SMS text messages.

Of 10 studies, 3 (30%), including 2 one-arm trials [53,55] and 1 randomized controlled trial [58], reported changes in nutritional knowledge and skills, and 2 of these by Porter et al [55] and Moussa et al [58] reported positive outcomes. In this group, Moussa et al [58] conducted a controlled trial on a 4-week web-based educational intervention called *eCare We Care* and reported significant (P=.01) improvements in intervention participants' diabetes knowledge scores (including nutritional knowledge questions) in comparison with the control group participants who received paper-based, text-only tutorials. Porter et al [55] also reported that the *Your Guide to Diet and Diabetes* web-based intervention led to a 39% increase in diabetes-related knowledge (including nutritional knowledge questions), a 39% increase in carbohydrate counting skills, and a 31% increase in the ability to plan a meal plate.

A total of 4 studies [54,57,59,60] in the category of nutrition-related cognitive or behavioral outcomes, including 3 randomized controlled trials [54,59,60] and 1 one-arm trial [57], evaluated the effect of digitally delivered interventions on improving dietary self-efficacy, with 3 of these studies [57,59,60] showing positive impacts on at least 1 measure. Ruggeiro et al [57], for example, reported significant improvements in 1 specific subscale of the dietary self-efficacy scale (modifying meat) among participants who were provided with a virtual environment for DSM education and support and a study website for communication and tracking (*Diabetes Island*) in a 1-arm trial. However, this study did not show any significant improvement in other items of the dietary self-efficacy scale including general diet and specific diet. By contrast, Whittemore et al (using the same scale) [63] and Aligholipour et al [59] reported significant improvement in total scores of participants' dietary self-efficacy. Nelson et al [61] evaluated the effect of daily SMS text messages supporting self-care behaviors (ie, diet, exercise, and self-monitoring of blood glucose) on the use of dietary information for decision-making using the Personal Diabetes Questionnaire scale and reported no significant changes in participants' total scores.

The effects of interventions on improving nutrition-related health outcomes including FBS, $HbA_{1c,}$ and anthropometrics measurements were examined in 1 one-arm trial [57] and 6 randomized controlled trials [54,56,59,61-63]. All 7 studies reported improvement in the blood glucose levels (FBS or HbA_{1c} levels) of participants postintervention. However, these changes were not significant in 3 of these studies [54,56,57].

Of the 10 included studies, 3 (30%; one 1-arm trial [57] and 2 randomized controlled trials [60,62]) reported the effects of the interventions on anthropometric measures. However, in the 1-arm study by Ruggeiro et al [57], the mean BMI in the intervention group decreased significantly by 0.6 kg/m^2. This reduction was attributed to improvements in the participants' dietary choices. None of the studies reported improvements in the participants' weight or other clinical indicators.

XSL•FO
RenderX

Table 2. Nutrition-related primary or secondary outcomes of the included studies.

Study	Study design	Outcome category	Results of the intervention group
Aligholipour et al [59], 2019	2-arm randomized controlled trial	• Nutrition-related health outcome • FBS[a] level • Nutrition-related cognitive or behavioral outcomes • Dietary self-efficacy	• Significant reduction in FBS level (−36.9; *P*<.001) • Significant improvement in dietary self-efficacy (*P*<.001)
Arora et al [53], 2012	1-arm nonrandomized pilot trial	• Nutrition-related cognitive or behavioral outcomes • Dietary consumption • Knowledge and skills	• Significant increase in fruits and vegetables consumption by 26.5% of participants (*P*<.05) • Nonsignificant change in diabetes knowledge score
Burner et al [54], 2017	2-arm randomized controlled trial	• Nutrition-related health outcome • HbA_{1c}[b] level • Nutrition-related cognitive or behavioral outcomes • Dietary self-efficacy	• Nonsignificant reduction in HbA_{1c} level by −1.4% (*P*=.30) • Nonsignificant change in dietary self-efficacy score (including general diet and specific diet)
Fortmann et al [62], 2017	2-arm randomized controlled trial	• Nutrition-related health outcome • HbA_{1c} level • BMI • Weight	• Significant reduction in HbA_{1c} levels by −0.8% (*P*=.03) at 3 months and by −0.8% (*P*=.03) at 6 months postbaseline • Nonsignificant change in BMI • Nonsignificant change in weight
Moussa et al [58], 2013	2-arm randomized controlled trial	• Nutrition-related cognitive or behavioral outcomes • Knowledge and skills	• Significant increase in meal planning score (*P*=.01)
Nelson et al [61], 2021	3-arm randomized controlled trial	• Nutrition-related health outcome • HbA_{1c} level • Nutrition-related cognitive or behavioral outcomes • PDQ[c] score	• Significant reduction in HbA_{1c} levels by −0.26% (*P*=.02) at 3 months and by −0.31% (*P*=.04) at 6 months postbaseline • Nonsignificant change in PDQ
Porter et al [55], 2009	1-arm nonrandomized trial	• Nutrition-related cognitive or behavioral outcomes • Knowledge and skills	• Significant improvement in diabetes-related knowledge by 39% (*P*<.05) • Significant improvement in carbohydrate counting skill by 39% (*P*<.05) • Significant improvement in ability to plan a meal plate by 31% (*P*<.05)
Robertson et al [56]; 2007	2-arm randomized controlled trial	• Nutrition-related health outcome • HbA_{1c} level • Nutrition-related cognitive or behavioral outcomes • Dietary consumption	• Nonsignificant reduction in HbA_{1c} level (−0.8; *P*=.06) • Nonsignificant change in total calorie intake • Nonsignificant change in calories obtained from carbohydrates, proteins, or fats

XSL•FO
RenderX

Study	Study design	Outcome category	Results of the intervention group
Ruggiero et al [57]; 2014	1-arm nonrandomized trial	• Nutrition-related health outcome • HbA_{1c} level • BMI • Nutrition-related cognitive or behavioral outcomes • Dietary self-efficacy	• Nonsignificant reduction in HbA_{1c} level (−0.3; *P*=.053) • Significant reduction in BMI level (−0.5, *P*=.02, at 3 months; and −0.7, *P*=.02, at 6 months) • Nonsignificant change for general diet (*P*=.22) • Nonsignificant change for general diet [61] (*P*=.09) • Nonsignificant improvement in specific diet (*P*=.42) • Nonsignificant change for specific diet rev (*P*=.91) • Significant improvement in modifying meat (*P*=.01) • Nonsignificant changes for avoiding fat (*P*=.70) • Nonsignificant changes for diet subscale (*P*=.01)
Whittemore et al [60], 2020	2-arm randomized waitlist-controlled pilot trial	• Nutrition-related health outcome • HbA_{1c} level • BMI • Nutrition-related cognitive or behavioral outcomes • Dietary self-efficacy	• Significant reduction in HbA_{1c} level by −1.76% (*P*<.01) • Nonsignificant change in BMI (*P*=.22) • Significant improvement in dietary self-efficacy (*P*<.01)

[a]FBS: fasting blood sugar.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]PDQ: Personal Diabetes Questionnaire.

Behavior Change Theories and Techniques

Of the 10 included studies, only 2 (20%) [57,60,63] described the use of behavioral change theories in designing interventions. The theories mentioned in these studies were the social cognitive theory, the health action process approach model, and the empowerment theory.

The studies included in this review used substantially heterogeneous terms to describe the intervention content and rarely reported their BCTs in detail, which limited the possibility of identifying their mechanisms of action. We used the BCT taxonomy [42] to develop a list of BCTs that appeared to be used according to the reported intervention content for each study. Table 3 summarizes the frequencies of the BCT clusters implemented across the studies in this review.

The coding of intervention strategies showed that 10 (62%) of the 16 BCT clusters were used across the included studies. *Shaping knowledge* (BCT 4; 10/10, 100% studies) and *natural consequences* (BCT 5; 9/10, 90% studies) were implemented in almost all studies, whereas *repetition and substitution* (BCT 8; 1/10, 10% study) and *regulation* (BCT 11; 1/10, 10% study) were only implemented in 1 study.

In addition to the 16 main behavior change clusters, intervention components can be further classified into 93 single BCTs [42]. A detailed summary of the BCTs included in each study is presented in Table 4.

The number of BCTs implemented across all the studies ranged from 2 [55] to 12 [63]. The most frequently implemented individual BCT was *instruction on how to perform the behavior* (BCT 4.1; 10/10, 100% studies), which was used across all studies. The second most frequently used individual BCT was *information about health consequences* (BCT 5.1; 9/10, 90% studies), which was implemented in 9 studies in this review [53-55,57,58]. The other common individual BCTs in the included studies were *credible source* (BCT 9.1; 8/10, 80% studies) and *social support (unspecified)* (BCT 3.1; 8/10, 80% studies).

XSL•FO
RenderX

Table 3. Studies implementing each BCT[a] cluster (n=10).

Behavior change cluster	Studies, n (%)
4. Shaping knowledge[b]	10 (100)
5. Natural consequences	9 (90)
3. Social support	8 (80)
9. Comparison of outcomes	8 (80)
2. Feedback and monitoring	7 (70)
1. Goals and planning	4 (40)
10. Reward and threat	3 (30)
6. Comparison of behavior	2 (20)
8. Repetition and substitution	1 (10)
11. Regulation	1 (10)

[a]BCT: behavior change technique.

[b]Cluster numbers are based on the actual numbers used in the BCT taxonomy developed by Michie et al [42].

XSL•FO
RenderX

Table 4. Theoretical approaches, BCT[a] clusters, and BCTs implemented across studies.

Study	Theoretical approach	Behavior change clusters[b]	BCTs[b]
Aligholipour et al [59], 2019	• ND[c]	• 2. Feedback and monitoring • 3. Social support • 4. Shaping knowledge • 5. Natural consequences • 9. Comparison of outcomes	• 2.2. Feedback on behavior • 2.7. Feedback on outcome of behavior • 3.1. Social support (unspecified) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 9.1. Credible source
Arora et al [53], 2012	• ND	• 1. Goals and planning • 3. Social support • 4. Shaping knowledge • 5. Natural consequences • 9. Comparison of outcomes • 10. Reward and threat	• 1.1. Goal setting (behavior) • 1.4. Action planning • 3.1. Social support (unspecified) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 9.1. Credible source • 9.2. Pros and cons • 9.3. Comparative imagining of future outcomes • 10.2. Material reward
Burner et al [54], 2017	• ND	• 2. Feedback and monitoring • 3. Social support • 4. Shaping knowledge • 5. Natural consequences • 9. Comparison of outcomes • 10. Reward and threat	• 2.2. Feedback on behavior • 2.7. Feedback on outcome of behavior • 3.1. Social support (unspecified) • 3.3. Social support (emotional) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 9.1. Credible source • 10.4. Social reward
Fortmann et al [62], 2017	• ND	• 2. Feedback and monitoring • 3. Social support • 4. Shaping knowledge • 5. Natural consequences • 9. Comparison of outcomes	• 2.3. Self-monitoring of behavior • 3.1. Social support (unspecified) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 9.1. Credible source
Moussa et al [58], 2013	• ND	• 3. Social support • 4. Shaping knowledge • 5. Natural consequences • 9. Comparison of outcomes	• 3.1. Social support (unspecified) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 9.1. Credible source
Nelson et al [61], 2021	• ND	• 1. Goals and planning • 2. Feedback and monitoring • 3. Social support • 4. Shaping knowledge • 5. Natural consequences 9. Comparison of outcomes	• 1.1. Goal setting (behavior) • 2.2. Feedback on behavior • 2.7. Feedback on outcome of behavior • 3.1. Social support (unspecified) • 3.3. Social support (emotional) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 9.1. Credible source
Porter et al [55], 2009	• ND	• 4. Shaping knowledge • 5. Natural consequences	• 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences

XSL•FO
RenderX

Study	Theoretical approach	Behavior change clusters[b]	BCTs[b]
Robertson et al [56], 2007	• ND	• 2. Feedback and monitoring • 3. Social support • 4. Shaping knowledge	• 2.2. Feedback on behavior • 2.3. Self-monitoring of behavior • 3.1. Social support (unspecified) • 3.3. Social support (emotional) • 4.1. Instruction on how to perform the behavior
Ruggiero et al [57], 2014	• Social cognitive theory	• 1. Goals and planning • 2. Feedback and monitoring • 4. Shaping knowledge • 5. Natural consequences • 6. Comparison of behavior • 8. Repetition and substitution • 9. Comparison of outcomes • 10. Reward and threat	• 1.1. Goal setting (behavior) • 2.2. Feedback on behavior • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 6.1. Demonstration of the behavior • 6.2. Social comparison • 8.1. Behavioral practice and rehearsal • 9.1. Credible source • 9.2. Pros and cons • 10.2. Material reward
Whittemore et al [60], 2020	• HAPA[d] model • Empowerment theory • Social cognitive theory	• 1. Goals and planning • 2. Feedback and monitoring • 3. Social support • 4. Shaping knowledge • 5. Natural consequences • 6. Comparison of behavior • 9. Comparison of outcomes 11. Regulation	• 1.1. Goal setting (behavior) • 1.2. Problem-solving • 1.3 Goal setting (outcome) • 1.4. Action planning • 2.3. Self-monitoring of behavior • 3.1. Social support (unspecified) • 4.1. Instruction on how to perform the behavior • 5.1. Information about health consequences • 6.1. Demonstration of the behavior • 9.1. Credible source • 9.2. Pros and cons 11.2. Reduce negative emotions

[a]BCT: behavior change technique.

[b]The clusters and BCT numbers are based on the actual numbers used in the BCT taxonomy developed by Michie et al [42].

[c]ND: none described.

[d]HAPA: health action process approach.

Risk of Bias Across Studies

Details of the risk of bias of the included studies are summarized in Figures 2 and 3 [53-62], using the Cochrane Risk-of-Bias Assessment Tool (version 2) [65], and in Multimedia Appendix 2 [53-62].

XSL•FO
RenderX

Figure 2. Risk-of-bias summary: review authors' judgments about each risk-of-bias item for each included study.

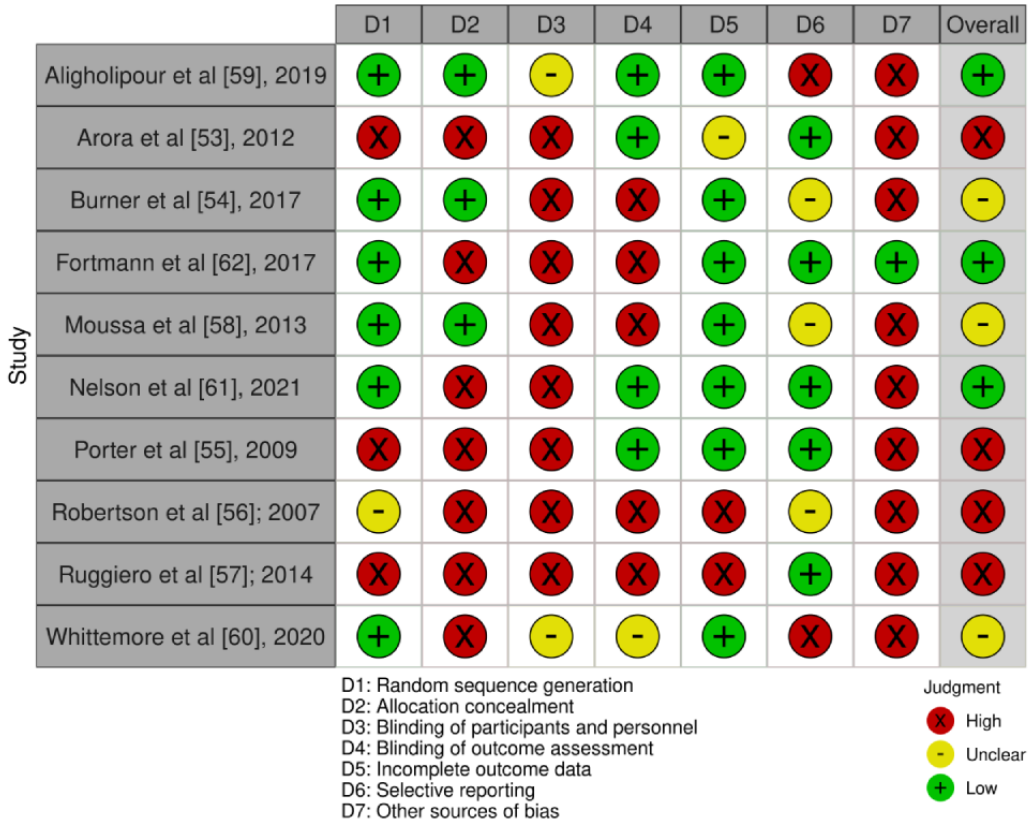

Figure 3. Risk-of-bias judgments about each risk-of-bias item presented as percentages across all included studies.

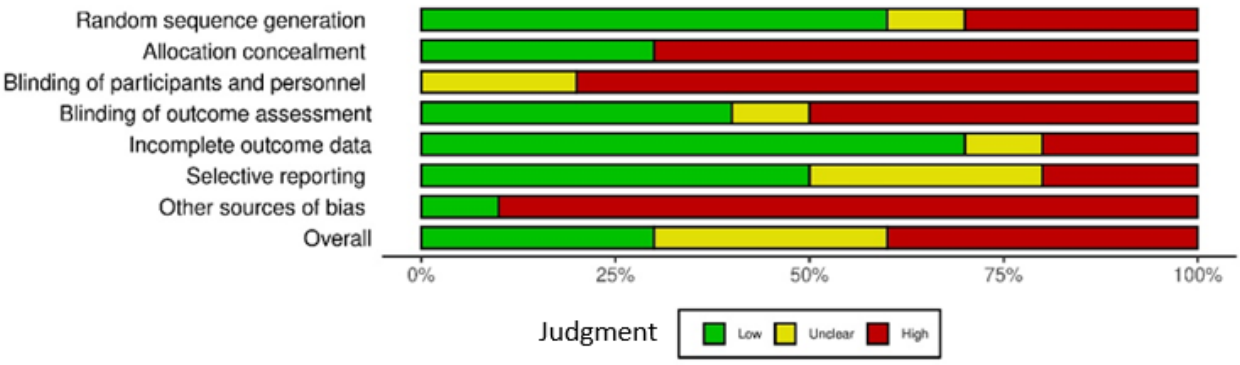

Allocation

Of the 10 studies included in this review, 6 (60%) had acceptable sequence generation [54,58-62], and 3 (30%) [53,55,57] were 1-arm nonrandomized controlled trials, and the sequence generation of participants was not possible with their design. These 3 studies were categorized as having a high risk of bias. In the study by Robertson et al [56], the method of generating random number sequences was not described, and this study was classified as having an unclear risk of bias.

A total of 3 (30%) [54,58,59] of the 10 studies in this review described an appropriate allocation concealment method. In these studies, the methods for allocation concealment were the use of a closed envelope [54,59] or a container, including equal *intervention* and *control* choices to draw [58]. In 4 (40%) of the 10 studies [56,60-62], the method used to conceal the allocation sequence was not provided. Hence, intervention allocations could have been foreseen before or during the enrollment. Another 3 studies [53,55,57] were 1-arm nonrandomized controlled trials, and the allocation concealment of participants was not possible with their design.

Blinding

Blinding of participants is often not feasible in dietary behavior studies, and none of the studies in this review blinded the participants. Furthermore, there were no comments on blinding of program personnel in 4 studies [53,55,57,59] and outcome assessors in 6 studies [53,55,57,59-61]. However, in 3 of these studies [53,55,59], data were collected through surveys; therefore, blinding of assessors was not necessary, and the outcome measurement was not likely to be influenced by lack of blinding. These studies were assigned a low risk of bias.

XSL•FO
RenderX

Incomplete Outcome Data

Of the 10 included studies, 7 (70%) [54,55,58-62] were assigned a low risk of bias. These studies either had 0 or minimal missing data (which was unlikely to have a great influence on outcome data) [55,58,59], or missing data were imputed using appropriate methods [54,60-62].

The study by Arora et al [53] had modest attrition (13%) by the 3-week assessment, but reasons for attrition were not provided, and analyses for handling incomplete data were not performed. Consequently, this study was identified as having an unclear risk of bias [53].

The remaining 2 studies failed to provide complete outcome data and had a high risk of bias [56,57]. In 1 of these studies [57], a series of mixed-effects analyses with time as a single repeated measures factor was performed to address missing data; however, as the attrition rate was high and differences in characteristics between completers and noncompleters were not provided, we felt that the results were at a high risk of attrition bias. The study [56] had incomplete demographic data, and consequently, group differences between completers and noncompleters were not reported accurately.

Selective Reporting

We were able to identify a published protocol for 4 (40%) of the 10 included studies [59-62]. In 2 of these studies [61,62], the reported outcomes matched those reported in the registered protocol; therefore, these studies were considered to have a low risk of bias.

Other Potential Sources of Bias

We did not observe any strong source of other bias in 1 (10%) [62] of the 10 studies included in this review. However, other studies had high [53,54,56-58] or unclear [55] risks of having other biases. Studies were judged to have a high risk of bias because of having ungeneralizable data [53,59], small sample sizes without formal power calculations or convenience sampling [54,55,58], using unvalidated surveys [58], providing monetary incentives [57,60,61], recruiting participants who owned mobile phones only [53,54,59], not providing sufficient demographic data [56], and having a potentially active intervention for the control group [54,58,59].

Discussion

Principal Findings

This review advances the existing evidence based on previous reviews [25,26,31] by assessing studies that focused on nutritional behaviors and nutrition-related health outcomes and were conducted among socioeconomically disadvantaged people with T2D. In this systematic review, although 10 studies (including 947 participants) were identified, smaller subsets assessed particular outcomes, and both methods and findings were somewhat mixed; therefore, evidence remains limited but is suggestive of the potential of digitally delivered nutrition interventions in socioeconomically disadvantaged people with T2D. Studies in this review included controlled and noncontrolled interventions covering mobile- and web-based interventions, ranging from 1 day to 1 year in duration, with or without health coaching, among individuals aged 45.4 to 55.8 years from disadvantaged population groups.

Effect of the Interventions on Nutrition Behavior Outcomes

Studies comparing digitally delivered interventions versus controls and pre - and posttest comparisons yielded broadly positive results. However, considerable diversity was observed in the range of outcomes measured.

A total of 9 studies in this review, including 3 one-arm trials [53,55,57] and 6 randomized controlled trials [54,56,58,60,61], reported nutrition-related cognitive or behavioral outcomes. Of the 3 one-arm trials, all showed improvements in nutritional behavior outcomes, and 1 specifically provided evidence regarding healthier food choices, such as increased consumption of fruits and vegetables, among participants [53]. By contrast, of the 6 controlled trials, only 2 provided significant improvement in nutritional behaviors [58,59]. All 9 studies showed positive effects of interventions on nutritional knowledge, skills, and self-efficacy, which are the key mediators of healthy eating. Given the rapid increase in the disease burden and prevalence of diabetes worldwide, and especially in disadvantaged populations, it is notable that the weight of evidence suggests that these digitally delivered interventions can have a positive effect on dietary choices and engagement in healthy eating behaviors or on their cognitive or behavioral determinants among socioeconomically disadvantaged populations.

Differences across nutritional outcomes might reflect the heterogeneous dietary recommendations provided to participants in the interventions or the type and amount of tailoring of the materials. It appears that interventions that incorporated feedback and suggestions from members of ethnic communities and provided culturally appropriate materials (eg, cooking lessons or tailored nutritional education or counseling) resulted in better outcomes [55-57,60,62]. Another explanation for the mixed results may be intervention complexity or dose, as interventions varied in terms of the number of components, quantity of materials, and support provided to participants. In this review, interventions with fewer elements and simple design characteristics to address the needs of people with low literacy and electronic literacy [55,58] reported more positive outcomes and higher retention rates than those with more complex designs [56,57]. Previous systematic reviews have also indicated that overly complex interventions may lead to a lack of motivation owing to confusion, provision of irrelevant content, and technical difficulties [66,67]. Furthermore, the studies included in this review used substantially different nutrition assessment methods, which were often study specific. This diversity in methods for assessing dietary intake might have also contributed to the observed heterogeneity in the results.

Finally, based on the results observed in 7 randomized controlled trials and 3 one-arm trials, the overall effectiveness of the interventions did not seem to vary according to whether the interventions were fully automated or whether they were combined with additional intervention components such as the presence of an educator. This reinforces the idea that automated programs, with their potential to save labor, could be seen as a

XSL•FO
RenderX

way to deliver diabetes interventions efficiently and effectively. Such programs can also increase the predictability of outcomes. When people follow a standardized care path supported by automation, they are more likely to stay on track toward the accepted outcomes [31]. Moreover, automation can assist in detecting when a patient deviates from the recommended care plan, enabling the health care team to intervene at the right moment [28,29,31].

Effect of the Interventions on Nutrition-Related Health Outcomes

This review also provided some evidence that improvement in blood glucose levels can be achieved by using digitally delivered interventions for people from socioeconomically disadvantaged backgrounds.

A total of 7 studies examined changes in blood glucose levels (FBS or HbA_{1c}), of which 4 [59-62] revealed significant improvement. In these 4 studies, decreases in HbA_{1c} levels ranged between 0.3% [53] and 1.8% [60], and FBS decreased by 36.9 mmol/L [59]. The other 3 studies [54,56,57] reported positive but nonstatistically significant improvements.

This range of changes in HbA_{1c} levels is somewhat similar to that reported in in-person and digitally delivered T2D self-management education interventions in general (not specifically disadvantaged) [68,69] and disadvantaged populations [70]. For instance, web-based programs have been shown to decrease HbA_{1c} levels by 0.47% to 1.49%, mobile apps by 0.4% to 1.9% [71], other mobile-based programs by an average of 0.8% [72], and telehealth interventions by an average of 0.17% [73]. In addition, in socioeconomically disadvantaged people with T2D, a previous systematic review and meta-analysis on the effects of digitally delivered T2D self-management education interventions on glycemic control found modest but clinically significant reductions in HbA_{1c} levels, ranging from 0.2% to 0.8% at 6 months. [26]. Although the observed changes in HbA_{1c} levels are rather small across all types of studies, these changes are clinically meaningful, as previous research has demonstrated that even a 1% decrease in HbA_{1c} levels for 10 years is related to a 37% reduction in microvascular complications and 21% reduction in morbidity from diabetes [74].

BCTs Used by the Interventions

Despite previous evidence that theoretically grounded interventions are more effective than atheoretical approaches, 8 (80%) of the 10 studies did not report a theoretical underpinning. Those that did [57,63] reported significant intervention effects for some but not all outcomes. Further work is needed to examine the most appropriate evidence-based theoretical frameworks for developing intervention content for these target groups.

There were considerable variations in the type and number of BCTs included in the interventions. On average, 5 (31%) of the 16 different BCT clusters were implemented in each study. The BCT clusters that formed the central *building blocks* of interventions in this review were providing information, informing natural consequences, providing social support, delivering feedback and self-monitoring, and setting goals. These BCT clusters are similar to those incorporated in successful in-person individual and group - based interventions [75] or those reported in reviews of digitally delivered interventions [22,23,40,75-78].

The number of single BCTs implemented in the interventions varied from 2 to 12 across studies. Although in studies with a minimum number of BCTs, significant improvements in healthy eating and diet-related health outcomes were reported [55], several studies included >6 BCTs and yet showed no significant outcomes. Therefore, these results do not support the previously suggested associations between greater intervention effectiveness and increasing number of BCTs [75,77-79]. A similar result was reported in a review by Villinger et al [27], which revealed no association between the number of BCTs and the efficacy of app-based mobile interventions on nutritional behaviors. However, it should be noted that most studies did not explicitly report on BCTs; hence, they were inferred by the review authors based on the intervention content described.

The most frequently identified single BCTs were *instruction on how to perform the behavior*, *information about health consequences*, and *social support*. These results suggest that researchers rely heavily on the assumption that providing instructions on how to perform the behavior and information about health consequences are the most effective tools for bringing about changes in people's healthy behaviors. However, evidence shows that health behaviors are motivated by a broader range of perceived benefits than health considerations alone [80]. Moreover, social support provided in the included interventions mainly focused on encouragement and counseling from HCPs or family supporters, and mentions of practical support were not present, although evidence suggests that such sources of social support are also beneficial for improving healthy behaviors and clinical outcomes [81]. There were no similar reviews that have reported on BCTs used in digitally delivered nutritional support interventions. However, the most commonly used BCT categories identified in this review correspond closely with those found in a previous review by Seppälä et al [81] that identified targets, mediators, and change strategies for physical activity and nutrition behavior changes in Finnish policy papers on workplace health promotion.

Although we acknowledge that the taxonomy of BCTs is comprehensive and not all techniques may be needed to change nutrition behavior, among the 93 BCTs proposed in the taxonomy by Michie et al [42], only 19 were used in the interventions. These findings suggest the need for researchers to explore the potential of not commonly used BCTs, but others, such as problem-solving, that may contribute to increasing the effectiveness of interventions to promote nutritional behaviors.

Tailoring of Interventions to Target Population

In this review, tailoring the digital design using digital elements or language suitable for people with low literacy or electronic literacy [53,54] and targeting intervention content based on the cultural background and language of intended participants comprised the main forms of intervention tailoring [55-57,60,62]. These results are in line with the results of previous systematic reviews on the effectiveness of

XSL•FO
RenderX

self-management support interventions in patients with diabetes or other chronic conditions [3,22,26], which reported that the most common ways of targeting the intervention to the population group interventions are by (1) modifying content to culture or language, (2) using personnel familiar with study participants (eg, community health workers, health educators, and nurses), or (3) individually tailoring an intervention to self-reported participants' data (eg, blood glucose level, diet and physical activity data, diabetes barriers, social support, and psychosocial challenges).

Although tailoring the content to language has been shown to be helpful in some studies [3,25,55,58], for maximum effectiveness, researchers should also consider participants' literacy and numeracy levels in the design of digital interventions for people who are disadvantaged. Low literacy is a prevalent barrier to understanding and acting on health information among these populations [26]. People with lower levels of literacy and numeracy often struggle with navigating web-based portals and understanding and trusting web-based health information and prefer to rely on verbal communication about their health-related issues [82,83]. Some studies that factored in electronic literacy considerations suggested that this led to an increased understanding of educational content by end users [83].

Strengths and Limitations

To the best of our knowledge, this is the first review to evaluate the effect of digitally delivered interventions on nutrition-related cognitive or behavioral outcomes and nutrition-related health outcomes among socioeconomically disadvantaged people with T2D. A comprehensive search strategy was developed in consultation with an expert librarian, and the search was performed using 7 electronic databases from medical and nonmedical sources. Furthermore, we used a reliable BCT taxonomy [42] to develop a list of effective BCTs for each study. This helped us to handle heterogeneity across multicomponent and complex interventions and to unravel the effects of intervention components on nutrition behavior outcomes. However, a potential limitation is that, given the limited reporting of these constructs, not all BCTs may have been captured accurately or at all.

Our search was limited to English language studies, and only published peer review studies were included. Owing to the multifaceted and complex nature of digitally delivered interventions, there were a multitude of primary and secondary outcomes for which data were extracted; however, these were specified a priori, and we have only reported on the outcomes listed in the registered protocol. The findings should also be interpreted with some caution, given the limited number of studies in each group of technologies and in general. Comparable to similar reviews, the methodological quality of studies varied; for example, almost none of the studies blinded outcome assessors or participants to the treatment group. However, blinding participants in behavioral interventions is difficult, and this criticism is common in many similar reviews. In addition, most studies were conducted in the United States, limiting the generalizability to low- to middle-income countries. Given that three-quarters of the world's diabetes population lives in these countries, more digitally delivered interventions should be developed and tested for people with diabetes in low- to middle-income countries. Such interventions must be properly contextualized and adapted to the culture, language, literacy and numeracy skills, health systems, and access to the internet and digital devices of the populations in these countries.

Another limitation of this review was that definitions of and thresholds for belonging to disadvantaged populations differed to some extent between studies, revealing that there is no one agreed-on *cut-off* for this. We specified that the terms related to SES, low income, or minority had to be used in the title or body of the paper to refer to participants for studies to be included, and we searched an extensive range of concepts to target low SES, such as residential areas, belonging to a health clinic serving disadvantaged groups, belonging to certain ethnic groups, and indicators of income. Thus, we captured studies considering disadvantage in different ways. However, relevant papers that did not use these terms may have been missed.

One challenge in answering our research question was that although it is well known that a healthy diet plays an important role in managing diabetes, several included interventions did not have a main focus on nutrition but included nutrition education as part of their multicomponent interventions to improve T2D self-management. Participants might have responded more effectively to interventions focused intensively on 1 area of behavior change (such as nutrition) at a time rather than multicomponent interventions [27]. This highlights a further gap in the current literature that needs to be addressed in future studies.

Conclusions

Despite research highlighting the serious need for effective behavior change support for people from susceptible populations to assist in reducing health inequities, relatively few studies of digital interventions have included these populations or particularly tailored materials to fit their cultural, linguistic, literacy, numeracy, or other needs.

Owing to the limited evidence base, few claims can be made regarding the effectiveness of digitally delivered interventions on nutrition behavior outcomes for disadvantaged people with T2D. However, we found some support for the efficacy of these interventions on healthy behaviors among disadvantaged populations, which is an essential prerequisite for changes in the clinical metabolic parameters. Our review showed small but positive changes in some eating behaviors and related health outcomes such as HbA_{1c} levels, although the results were highly heterogeneous. Heterogeneity is likely because of variations in the study design (eg, comparator groups and technologies used) and methodological quality (eg, instruments to measure outcomes assessed). Differences in the intervention content, including the BCTs used, is also a possible contributing factor.

In summary, a small body of evidence suggests that digital interventions can influence eating behaviors among certain disadvantaged groups with T2D. If this effect is generalized on a population, it could potentially have a considerable impact on public health. Specifically, digital technology can be used to meet the specific needs of disadvantaged patients for T2D

XSL•FO
RenderX

self-management with regard to eLiteracy, literacy, numeracy, and cultural considerations. Further evidence is needed regarding the specific features of effective digital interventions for supporting healthy behaviors in the most disadvantaged people and helping address health inequities.

Acknowledgments

The authors would like to thank the Institute for Physical Activity and Nutrition at the Deakin University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies for the MEDLINE complete, CINAHL complete, Global Health, and Applied Science databases.
[DOCX File , 18 KB - jmir_v26i1e42595_app1.docx]

Multimedia Appendix 2
Checklist for the quality assessment of studies.
[DOCX File , 34 KB - jmir_v26i1e42595_app2.docx]

References

1. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet 2017 Jun;389(10085):2239-2251. [doi: 10.1016/s0140-6736(17)30058-2]
2. Liu M, Lv X, Li Y, Li J, He Y. Prevalence and control status of diabetes and related risk factors among 4196 chinese male older elderly aged ≥80 years. Int J Gerontol 2018 Jun;12(2):122-126. [doi: 10.1016/j.ijge.2017.08.005]
3. van Hecke A, Heinen M, Fernández-Ortega P, Graue M, Hendriks JM, Høy B, et al. Systematic literature review on effectiveness of self-management support interventions in patients with chronic conditions and low socio-economic status. J Adv Nurs 2017 Apr;73(4):775-793. [doi: 10.1111/jan.13159] [Medline: 27653960]
4. Australia's health 2018. Australia's health series no. 16. AUS 221. Australian Institute of Health and Welfare. Canberra: Australian Institute of Health and Welfare; 2018. URL: https://www.aihw.gov.au/getmedia/7c42913d-295f-4bc9-9c24-4e44eff4a04a/aihw-aus-221.pdf [accessed 2023-12-25]
5. Global report on diabetes. World Health Organization. 2016 Apr 21. URL: https://www.who.int/publications/i/item/9789241565257 [accessed 2023-12-25]
6. Menon K, Mousa A, de Courten MP, Soldatos G, Egger G, de Courten B. Shared medical appointments may be effective for improving clinical and behavioral outcomes in type 2 diabetes: a narrative review. Front Endocrinol (Lausanne) 2017 Oct 4;8:263 [FREE Full text] [doi: 10.3389/fendo.2017.00263] [Medline: 29046662]
7. Seuring T, Archangelidi O, Suhrcke M. The economic costs of type 2 diabetes: a global systematic review. Pharmacoeconomics 2015 Aug;33(8):811-831 [FREE Full text] [doi: 10.1007/s40273-015-0268-9] [Medline: 25787932]
8. Funnell MM, Brown TL, Childs BP, Haas LB, Hosey GM, Jensen B, et al. National standards for diabetes self-management education. Diabetes Care 2008 Jan;31 Suppl 1(Suppl 1):S97-104 [FREE Full text] [doi: 10.2337/dc08-S097] [Medline: 18165344]
9. American Association of Diabetes Educators. AADE guidelines for the practice of diabetes self-management education and training (DSME/T). Diabetes Educ 2009 Nov 24;35(3_suppl):85S-107S. [doi: 10.1177/0145721709352436]
10. Povey RC, Clark-Carter D. Diabetes and healthy eating: a systematic review of the literature. Diabetes Educ 2007 Dec 07;33(6):931-59; discussion 960. [doi: 10.1177/0145721707308408] [Medline: 18057263]
11. Evert AB, Dennison M, Gardner CD, Garvey WT, Lau KH, MacLeod J, et al. Nutrition therapy for adults with diabetes or prediabetes: a consensus report. Diabetes Care 2019 May;42(5):731-754 [FREE Full text] [doi: 10.2337/dci19-0014] [Medline: 31000505]
12. Ball K, McNaughton SA, Le HN, Abbott G, Stephens LD, Crawford DA. ShopSmart 4 Health: results of a randomized controlled trial of a behavioral intervention promoting fruit and vegetable consumption among socioeconomically disadvantaged women. Am J Clin Nutr 2016 Aug;104(2):436-445 [FREE Full text] [doi: 10.3945/ajcn.116.133173] [Medline: 27413129]
13. Ball K, Crawford D, Mishra G. Socio-economic inequalities in women's fruit and vegetable intakes: a multilevel study of individual, social and environmental mediators. Public Health Nutr 2006 Aug 02;9(5):623-630 [FREE Full text] [doi: 10.1079/phn2005897] [Medline: 16923294]
14. Lee A. Affordability of fruits and vegetables and dietary quality worldwide. Lancet Global Health 2016 Oct;4(10):e664-e665. [doi: 10.1016/s2214-109x(16)30206-6]

XSL•FO
RenderX

15. Elgart JF, Caporale JE, Asteazarán S, De La Fuente JL, Camilluci C, Brown JB, et al. Association between socioeconomic status, type 2 diabetes and its chronic complications in Argentina. Diabetes Res Clin Pract 2014 May;104(2):241-247. [doi: 10.1016/j.diabres.2014.02.010] [Medline: 24629409]
16. Wang J, Cai C, Padhye N, Orlander P, Zare M. A behavioral lifestyle intervention enhanced with multiple-behavior self-monitoring using mobile and connected tools for underserved individuals with type 2 diabetes and comorbid overweight or obesity: pilot comparative effectiveness trial. JMIR Mhealth Uhealth 2018 Apr 10;6(4):e92 [FREE Full text] [doi: 10.2196/mhealth.4478] [Medline: 29636320]
17. Addo J, Agyemang C, de-Graft Aikins A, Beune E, Schulze MB, Danquah I, et al. Association between socioeconomic position and the prevalence of type 2 diabetes in Ghanaians in different geographic locations: the RODAM study. J Epidemiol Community Health 2017 Jul;71(7):633-639 [FREE Full text] [doi: 10.1136/jech-2016-208322] [Medline: 28348205]
18. Ciapponi A, Lewin S, Herrera CA, Opiyo N, Pantoja T, Paulsen E, et al. Delivery arrangements for health systems in low-income countries: an overview of systematic reviews. Cochrane Database Syst Rev 2017 Sep 13;9(9):CD011083 [FREE Full text] [doi: 10.1002/14651858.CD011083.pub2] [Medline: 28901005]
19. Walker RJ, Gebregziabher M, Martin-Harris B, Egede LE. Independent effects of socioeconomic and psychological social determinants of health on self-care and outcomes in type 2 diabetes. Gen Hosp Psychiatry 2014;36(6):662-668 [FREE Full text] [doi: 10.1016/j.genhosppsych.2014.06.011] [Medline: 25103544]
20. Mumu SJ, Saleh F, Ara F, Haque MR, Ali L. Awareness regarding risk factors of type 2 diabetes among individuals attending a tertiary-care hospital in Bangladesh: a cross-sectional study. BMC Res Notes 2014 Sep 03;7(1):599 [FREE Full text] [doi: 10.1186/1756-0500-7-599] [Medline: 25187113]
21. Pereira K, Phillips B, Johnson C, Vorderstrasse A. Internet delivered diabetes self-management education: a review. Diabetes Technol Ther 2015 Jan;17(1):55-63. [doi: 10.1089/dia.2014.0155] [Medline: 25238257]
22. Pal K, Eastwood SV, Michie S, Farmer AJ, Barnard ML, Peacock R, et al. Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. Cochrane Database Syst Rev 2013 Mar 28;2013(3):CD008776 [FREE Full text] [doi: 10.1002/14651858.CD008776.pub2] [Medline: 23543567]
23. Direito A, Carraça E, Rawstorn J, Whittaker R, Maddison R. mHealth technologies to influence physical activity and sedentary behaviors: behavior change techniques, systematic review and meta-analysis of randomized controlled trials. Ann Behav Med 2017 Apr;51(2):226-239. [doi: 10.1007/s12160-016-9846-0] [Medline: 27757789]
24. Rollo ME, Aguiar EJ, Williams RL, Wynne K, Kriss M, Callister R, et al. eHealth technologies to support nutrition and physical activity behaviors in diabetes self-management. Diabetes Metab Syndr Obes 2016 Nov 4;9:381-390 [FREE Full text] [doi: 10.2147/DMSO.S95247] [Medline: 27853384]
25. Heitkemper EM, Mamykina L, Travers J, Smaldone A. Do health information technology self-management interventions improve glycemic control in medically underserved adults with diabetes? a systematic review and meta-analysis. J Am Med Inform Assoc 2017 Sep 01;24(5):1024-1035 [FREE Full text] [doi: 10.1093/jamia/ocx025] [Medline: 28379397]
26. Mayberry LS, Lyles CR, Oldenburg B, Osborn CY, Parks M, Peek ME. mHealth interventions for disadvantaged and vulnerable people with type 2 diabetes. Curr Diab Rep 2019 Nov 25;19(12):148 [FREE Full text] [doi: 10.1007/s11892-019-1280-9] [Medline: 31768662]
27. Villinger K, Wahl DR, Boeing H, Schupp HT, Renner B. The effectiveness of app-based mobile interventions on nutrition behaviours and nutrition-related health outcomes: a systematic review and meta-analysis. Obes Rev 2019 Oct;20(10):1465-1484 [FREE Full text] [doi: 10.1111/obr.12903] [Medline: 31353783]
28. McGill R, Anwar E, Orton L, Bromley H, Lloyd-Williams F, O'Flaherty M, et al. Are interventions to promote healthy eating equally effective for all? systematic review of socioeconomic inequalities in impact. BMC Public Health 2015 May 02;15:457 [FREE Full text] [doi: 10.1186/s12889-015-1781-7] [Medline: 25934496]
29. Yu CH, Bahniwal R, Laupacis A, Leung E, Orr MS, Straus SE. Systematic review and evaluation of web-accessible tools for management of diabetes and related cardiovascular risk factors by patients and healthcare providers. J Am Med Inform Assoc 2012;19(4):514-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000307] [Medline: 22215057]
30. Paramastri R, Pratama SA, Ho DK, Purnamasari SD, Mohammed AZ, Galvin CJ, et al. Use of mobile applications to improve nutrition behaviour: a systematic review. Comput Methods Programs Biomed 2020 Aug;192:105459. [doi: 10.1016/j.cmpb.2020.105459] [Medline: 32234632]
31. Dening J, Islam SM, George E, Maddison R. Web-based interventions for dietary behavior in adults with type 2 diabetes: systematic review of randomized controlled trials. J Med Internet Res 2020 Aug 28;22(8):e16437 [FREE Full text] [doi: 10.2196/16437] [Medline: 32857059]
32. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. J Diabetes Sci Technol 2017 Sep;11(5):1015-1027 [FREE Full text] [doi: 10.1177/1932296817713506] [Medline: 28560898]
33. Dobson R, Whittaker R, Pfaeffli Dale L, Maddison R. The effectiveness of text message-based self-management interventions for poorly-controlled diabetes: a systematic review. Digit Health 2017 Nov 09;3:2055207617740315 [FREE Full text] [doi: 10.1177/2055207617740315] [Medline: 29942620]

XSL•FO
RenderX

34. Sahin C, Courtney KL, Naylor PJ, E Rhodes R. Tailored mobile text messaging interventions targeting type 2 diabetes self-management: a systematic review and a meta-analysis. Digit Health 2019 Apr 22;5:2055207619845279 [FREE Full text] [doi: 10.1177/2055207619845279] [Medline: 31041110]
35. Adu MD, Malabu UH, Callander EJ, Malau-Aduli AE, Malau-Aduli BS. Considerations for the development of mobile phone apps to support diabetes self-management: systematic review. JMIR Mhealth Uhealth 2018 Jun 21;6(6):e10115 [FREE Full text] [doi: 10.2196/10115] [Medline: 29929949]
36. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
37. Holmen H, Wahl AK, Cvancarova Småstuen M, Ribu L. Tailored communication within mobile apps for diabetes self-management: a systematic review. J Med Internet Res 2017 Jun 23;19(6):e227 [FREE Full text] [doi: 10.2196/jmir.7045] [Medline: 28645890]
38. McKay FH, Cheng C, Wright A, Shill J, Stephens H, Uccellini M. Evaluating mobile phone applications for health behaviour change: a systematic review. J Telemed Telecare 2018 Jan;24(1):22-30. [doi: 10.1177/1357633X16673538] [Medline: 27760883]
39. Palmer M, Sutherland J, Barnard S, Wynne A, Rezel E, Doel A, et al. The effectiveness of smoking cessation, physical activity/diet and alcohol reduction interventions delivered by mobile phones for the prevention of non-communicable diseases: a systematic review of randomised controlled trials. PLoS One 2018 Jan 5;13(1):e0189801 [FREE Full text] [doi: 10.1371/journal.pone.0189801] [Medline: 29304148]
40. Schippers M, Adam PC, Smolenski DJ, Wong HT, de Wit JB. A meta-analysis of overall effects of weight loss interventions delivered via mobile phones and effect size differences according to delivery mode, personal contact, and intervention intensity and duration. Obes Rev 2017 Apr;18(4):450-459 [FREE Full text] [doi: 10.1111/obr.12492] [Medline: 28187246]
41. Michie S, Johnston M. Theories and techniques of behaviour change: developing a cumulative science of behaviour change. Health Psychol Rev 2012 Mar;6(1):1-6. [doi: 10.1080/17437199.2012.654964]
42. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
43. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
44. Williams DR, Priest N, Anderson NB. Understanding associations among race, socioeconomic status, and health: patterns and prospects. Health Psychol 2016 Apr;35(4):407-411 [FREE Full text] [doi: 10.1037/hea0000242] [Medline: 27018733]
45. Kington RS, Smith JP. Socioeconomic status and racial and ethnic differences in functional status associated with chronic diseases. Am J Public Health 1997 May;87(5):805-810. [doi: 10.2105/ajph.87.5.805] [Medline: 9184510]
46. Anderson NB, Bulatao RA, Cohen B, National Research Council (U.S.). Panel on Race, Ethnicity, and Health in Later Life. Critical Perspectives on Racial and Ethnic Differences in Health in Late Life. Washington, DC: National Academies Press; 2004.
47. Ashton CM, Haidet P, Paterniti DA, Collins TC, Gordon HS, O'Malley K, et al. Racial and ethnic disparities in the use of health services: bias, preferences, or poor communication? J Gen Intern Med 2003 Feb;18(2):146-152 [FREE Full text] [doi: 10.1046/j.1525-1497.2003.20532.x] [Medline: 12542590]
48. D'Anna LH, Ponce NA, Siegel JM. Racial and ethnic health disparities: evidence of discrimination's effects across the SEP spectrum. Ethn Health 2010 Apr;15(2):121-143. [doi: 10.1080/13557850903490298] [Medline: 20131130]
49. Gardiner T, Abraham S, Clymer O, Rao M, Gnani S. Racial and ethnic health disparities in healthcare settings. BMJ 2021 Mar 08;372:n605 [FREE Full text] [doi: 10.1136/bmj.n605] [Medline: 33685939]
50. EndNote X8. Clarivate Analytics. URL: https://endnote.com/?language=en [accessed 2023-12-25]
51. Covidence - better systematic review management. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2023-12-20]
52. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.4. London, UK: The Cochrane Collaboration; 2023.
53. Arora S, Peters AL, Agy C, Menchine M. A mobile health intervention for inner city patients with poorly controlled diabetes: proof-of-concept of the TExT-MED program. Diabetes Technol Ther 2012 Jun;14(6):492-496. [doi: 10.1089/dia.2011.0252] [Medline: 22524591]
54. Burner E, Lam CN, DeRoss R, Kagawa-Singer M, Menchine M, Arora S. Using mobile health to improve social support for low-income Latino patients with diabetes: a mixed-methods analysis of the feasibility trial of TExT-MED + FANS. Diabetes Technol Ther 2018 Jan;20(1):39-48 [FREE Full text] [doi: 10.1089/dia.2017.0198] [Medline: 29227155]
55. Porter SJ, Chapman-Novakofski KM, Scherer JA. Your guide to diet and diabetes: web-based diabetes education tailored to Hispanics. J Nutr Educ Behav 2009;41(5):374-376. [doi: 10.1016/j.jneb.2009.04.003] [Medline: 19717123]

XSL•FO
RenderX

56. Robertson C, Kattelmann K, Ren C. Control of type 2 diabetes mellitus using interactive internet-based support on a northern plains Indian reservation: a pilot study. Top Clin Nutr 2007 Apr;22(2):2007-2093. [doi: 10.1097/01.TIN.0000270137.00099.91]
57. Ruggiero L, Moadsiri A, Quinn LT, Riley BB, Danielson KK, Monahan C, et al. Diabetes island: preliminary impact of a virtual world self-care educational intervention for African Americans with type 2 diabetes. JMIR Serious Games 2014;2(2):e10 [FREE Full text] [doi: 10.2196/games.3260] [Medline: 25584346]
58. Moussa M, Sherrod D, Choi J. An e-health intervention for increasing diabetes knowledge in African Americans. Int J Nurs Pract 2013 Sep 01;19 Suppl 3:36-43. [doi: 10.1111/ijn.12167] [Medline: 24090296]
59. Aligholipour M, Feizollahzadeh H, Ghaffari M, Jabbarzadeh F. Comparison of in-person and MMS -based education in telegram on self-care and fasting blood sugar of patients with diabetes mellitus: a randomized clinical trials. J Caring Sci 2019 Sep;8(3):157-164 [FREE Full text] [doi: 10.15171/jcs.2019.023] [Medline: 31598508]
60. Whittemore R, Vilar-Compte M, de la Cerda S, Delvy R, Jeon S, Burrola-Méndez S, et al. ¡Sí, Yo Puedo Vivir Sano con Diabetes! a self-management randomized controlled pilot trial for low-income adults with type 2 diabetes in Mexico city. Curr Dev Nutr 2020 May;4(5):nzaa074 [FREE Full text] [doi: 10.1093/cdn/nzaa074] [Medline: 32368713]
61. Nelson LA, Greevy RA, Spieker A, Wallston KA, Elasy TA, Kripalani S, et al. Effects of a tailored text messaging intervention among diverse adults with type 2 diabetes: evidence from the 15-month REACH randomized controlled trial. Diabetes Care 2021 Jan;44(1):26-34 [FREE Full text] [doi: 10.2337/dc20-0961] [Medline: 33154039]
62. Fortmann AL, Gallo LC, Garcia MI, Taleb M, Euyoque JA, Clark T, et al. Dulce digital: an mHealth SMS-based intervention improves glycemic control in Hispanics with type 2 diabetes. Diabetes Care 2017 Oct;40(10):1349-1355 [FREE Full text] [doi: 10.2337/dc17-0230] [Medline: 28600309]
63. Whittemore R, Vilar-Compte M, Burrola-Méndez S, Lozano-Marrufo A, Delvy R, Pardo-Carrillo M, et al. Development of a diabetes self-management + mHealth program: tailoring the intervention for a pilot study in a low-income setting in Mexico. Pilot Feasibility Stud 2020;6:25 [FREE Full text] [doi: 10.1186/s40814-020-0558-7] [Medline: 32082611]
64. Arora S, Peters AL, Burner E, Lam CN, Menchine M. Trial to examine text message-based mHealth in emergency department patients with diabetes (TExT-MED): a randomized controlled trial. Ann Emerg Med 2014 Jun;63(6):745-54.e6. [doi: 10.1016/j.annemergmed.2013.10.012] [Medline: 24225332]
65. Risk of Bias 2 (RoB 2) tool. Cochrane Methods. URL: https://methods.cochrane.org/risk-bias-2 [accessed 2023-12-25]
66. Pal K, Dack C, Ross J, Michie S, May C, Stevenson F, et al. Digital health interventions for adults with type 2 diabetes: qualitative study of patient perspectives on diabetes self-management education and support. J Med Internet Res 2018 Jan 29;20(2):e40 [FREE Full text] [doi: 10.2196/jmir.8439] [Medline: 29463488]
67. Lie SS, Karlsen B, Oord ER, Graue M, Oftedal B. Dropout from an eHealth intervention for adults with type 2 diabetes: a qualitative study. J Med Internet Res 2017 May 30;19(5):e187 [FREE Full text] [doi: 10.2196/jmir.7479] [Medline: 28559223]
68. Steinsbekk A, Rygg L, Lisulo M, Rise MB, Fretheim A. Group based diabetes self-management education compared to routine treatment for people with type 2 diabetes mellitus. a systematic review with meta-analysis. BMC Health Serv Res 2012 Jul 23;12:213 [FREE Full text] [doi: 10.1186/1472-6963-12-213] [Medline: 22824531]
69. Gary TL, Genkinger JM, Guallar E, Peyrot M, Brancati FL. Meta-analysis of randomized educational and behavioral interventions in type 2 diabetes. Diabetes Educ 2003;29(3):488-501. [doi: 10.1177/014572170302900313] [Medline: 12854339]
70. Ferguson S, Swan M, Smaldone A. Does diabetes self-management education in conjunction with primary care improve glycemic control in Hispanic patients? a systematic review and meta-analysis. Diabetes Educ 2015 Aug;41(4):472-484. [doi: 10.1177/0145721715584404] [Medline: 25941192]
71. Ramadas A, Quek KF, Chan CK, Oldenburg B. Web-based interventions for the management of type 2 diabetes mellitus: a systematic review of recent evidence. Int J Med Inform 2011 Jun;80(6):389-405 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.02.002] [Medline: 21481632]
72. Kitsiou S, Paré G, Jaana M, Gerber B. Effectiveness of mHealth interventions for patients with diabetes: an overview of systematic reviews. PLoS One 2017 Mar;12(3):e0173160 [FREE Full text] [doi: 10.1371/journal.pone.0173160] [Medline: 28249025]
73. Nangrani N, Malabu U, Vangaveti V. Outcomes of telehealth in the management of type 2 diabetes—a systematic review and meta-analysis of randomised controlled trials. Diabetes 2018 May;67(Supplement 1):1298-P. [doi: 10.2337/db18-1298-P]
74. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412 [FREE Full text] [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
75. Direito A, Dale LP, Shields E, Dobson R, Whittaker R, Maddison R. Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? BMC Public Health 2014 Jun 25;14:646 [FREE Full text] [doi: 10.1186/1471-2458-14-646] [Medline: 24965805]
76. Matthews J, Win KT, Oinas-Kukkonen H, Freeman M. Persuasive technology in mobile applications promoting physical activity: a systematic review. J Med Syst 2016 Mar;40(3):72. [doi: 10.1007/s10916-015-0425-x] [Medline: 26748792]

XSL•FO
RenderX

77. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014 Jul 25;11:97 [FREE Full text] [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]
78. Samdal GB, Eide GE, Barth T, Williams G, Meland E. Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. Int J Behav Nutr Phys Act 2017 Mar 28;14(1):42 [FREE Full text] [doi: 10.1186/s12966-017-0494-y] [Medline: 28351367]
79. Tang MY, Smith DM, Mc Sharry J, Hann M, French DP. Behavior change techniques associated with changes in postintervention and maintained changes in self-efficacy for physical activity: a systematic review with meta-analysis. Ann Behav Med 2019 Aug 16;53(9):801-815. [doi: 10.1093/abm/kay090] [Medline: 30534971]
80. Hankonen N, Absetz P, Kinnunen M, Haukkala A, Jallinoja P. Toward identifying a broader range of social cognitive determinants of dietary intentions and behaviors. Appl Psychol Health Well Being 2013 Mar;5(1):118-135. [doi: 10.1111/j.1758-0854.2012.01081.x] [Medline: 23457087]
81. Seppälä T, Hankonen N, Korkiakangas E, Ruusuvuori J, Laitinen J. National policies for the promotion of physical activity and healthy nutrition in the workplace context: a behaviour change wheel guided content analysis of policy papers in Finland. BMC Public Health 2017 Aug 02;18(1):87 [FREE Full text] [doi: 10.1186/s12889-017-4574-3] [Medline: 28764754]
82. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 07;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
83. Taha J, Sharit J, Czaja SJ. The impact of numeracy ability and technology skills on older adults' performance of health management tasks using a patient portal. J Appl Gerontol 2014 Jun;33(4):416-436 [FREE Full text] [doi: 10.1177/0733464812447283] [Medline: 24781964]

Abbreviations

BCT: behavior change technique
DSM: diabetes self-management
FBS: fasting blood sugar
HbA_{1c}: hemoglobin A1C
HCP: health care provider
MMS: multimedia messaging service
PICO: population, intervention, control, and outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SDSCA: Summary of Diabetes Self-Care Activities
SES: socioeconomic status
T2D: type 2 diabetes

Edited by T Leung; submitted 10.09.22; peer-reviewed by YL Leung, P Bower; comments to author 24.01.23; revised version received 22.06.23; accepted 30.07.23; published 01.02.24.

Please cite as:
Karimi N, Opie R, Crawford D, O'Connell S, Ball K
Digitally Delivered Interventions to Improve Nutrition Behaviors Among Resource-Poor and Ethnic Minority Groups With Type 2 Diabetes: Systematic Review
J Med Internet Res 2024;26:e42595
URL: https://www.jmir.org/2024/1/e42595
doi:10.2196/42595
PMID:38300694

©Nazgol Karimi, Rachelle Opie, David Crawford, Stella O'Connell, Kylie Ball. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Race, Ethnicity, and Other Cultural Background Factors in Trials of Internet-Based Cognitive Behavioral Therapy for Depression: Systematic Review

Robinson De Jesús-Romero[1], MS; Amani R Holder-Dixon[1,2], BA; John F Buss[1], BS; Lorenzo Lorenzo-Luaces[1], PhD

[1]Department of Psychological and Brain Sciences, Indiana University - Bloomington, Bloomington, IN, United States

[2]Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Robinson De Jesús-Romero, MS
Department of Psychological and Brain Sciences
Indiana University - Bloomington
1101 East 10th Street
Bloomington, IN, 47405
United States
Phone: 1 7872409456
Email: robdejes@iu.edu

Abstract

Background: There is a growing interest in developing scalable interventions, including internet-based cognitive behavioral therapy (iCBT), to meet the increasing demand for mental health services. Given the growth in diversity worldwide, it is essential that the clinical trials of iCBT for depression include diverse samples or, at least, report information on the race, ethnicity, or other background indicators of their samples. Unfortunately, the field lacks data on how well diversity is currently reported and represented in the iCBT literature.

Objective: Thus, the main objective of this systematic review was to examine the overall reporting of racial and ethnic identities in published clinical trials of iCBT for depression. We also aimed to review the representation of specific racial and ethnic minoritized groups and the inclusion of alternative background indicators such as migration status or country of residence.

Methods: Studies were included if they were randomized controlled trials in which iCBT was compared to a waiting list, care-as-usual, active control, or another iCBT. The included papers also had to have a focus on acute treatment (eg, 4 weeks to 6 months) of depression, be delivered via the internet on a website or a smartphone app and use guided or unguided self-help. Studies were initially identified from the METAPSY database (n=59) and then extended to include papers up to 2022, with papers retrieved from Embase, PubMed, PsycINFO, and Cochrane (n=3). Risk of bias assessment suggested that reported studies had at least some risk of bias due to use of self-report outcome measures.

Results: A total of 62 iCBT randomized controlled trials representing 17,210 participants are summarized in this study. Out of those 62 papers, only 17 (27%) of the trials reported race, and only 12 (19%) reported ethnicity. Reporting outside of the United States was very poor, with the United States accounting for 15 (88%) out of 17 of studies that reported race and 9 (75%) out of 12 for ethnicity. Out of 3,623 participants whose race was reported in the systematic review, the racial category reported the most was White (n=2716, 74.9%), followed by Asian (n=209, 5.8%) and Black (n=274, 7.6%). Furthermore, only 25 (54%) out of the 46 papers conducted outside of the United States reported other background demographics.

Conclusions: It is important to note that the underreporting observed in this study does not necessarily indicate an underrepresentation in the actual study population. However, these findings highlight the poor reporting of race and ethnicity in iCBT trials for depression found in the literature. This lack of diversity reporting may have significant implications for the scalability of these interventions.

(J Med Internet Res 2024;26:e50780) doi:10.2196/50780

KEYWORDS

diversity; cognitive behavioral therapy; internet-based; depression; race; racial; ethnicity; culture; depressive; diverse; inclusive; inclusivity; DEI; diversity, equity, and inclusion; internet-based cognitive behavioral therapy; mental health; ethnic; cultures;

XSL•FO
RenderX

culturally; review methods; review methodology; systematic; clinical trial; clinical trials; randomized controlled trial; randomized controlled trials; controlled trial; controlled trials; reporting; immigrant; migrant; migrants; immigrants; psychotherapy; underrepresented; underrepresentation; representation; mobile phone

Introduction

Background

In 2020, approximately 34% of the US population identified themselves as belonging to racial minoritized groups, and 19% identified as members of ethnic minoritized groups [1]. Following the terminology used by other scholars [2], we employ the term "minoritized" rather than "minorities" to emphasize that these individuals' experiences are not intrinsic qualities of statistically small groups; instead, they are the result of dominant groups subordinating and, consequently, "minoritizing" them. The proportion of individuals belonging to a racial or ethnic minoritized group in the United States is expected to continue growing [3]. This increase in the proportion of minoritized groups is also expected in other countries. There are currently an estimated 272 million immigrants worldwide, with 82 million residing in Europe, 59 million in North America, and 49 million in Northern Africa and Western Asia [4]. Therefore, it is essential that mental health research responds to this increase in racial and ethnic diversity.

Mental health is one of the leading causes of disability in the United States and worldwide [5]. Among mental disorders, depression is the leading cause of significant disability [6] with an estimated economic burden of around US $210.5 billion per year in the United States alone [6]. Thus, given that mental health is strongly correlated with racial-ethnic identity and with other factors that are themselves linked to mental health (eg, socioeconomic status), there is a strong imperative for mental health research to represent the racial-ethnic diversity in our populations [7-9].

Current Reporting on Race and Ethnicity

Current research in mental health continues to use primarily non-Hispanic White populations, which fails to reflect the demographic makeup of countries worldwide. For example, in a study by Mak and colleagues [10], they reviewed 379 National Institute of Mental Health–funded clinical trials for various mental health disorders published between 1995 and 2004 to investigate how many trials reported sex, race, and ethnicity. They found that 91.6% of the National Institute of Mental Health–funded published trials reported sex. However, only 47.8% included race or ethnicity in their demographics, and 25.6% had incomplete race or ethnicity information.

Since then, the overall pattern of reporting demographic information has improved slightly. A more recent meta-analysis by Polo and colleagues [11] examines the trends in reporting and representation of racial-ethnic diversity in randomized controlled trials (RCT) of psychotherapy for depression for over 36 years. They found that reporting of racial-ethnic group membership increased from 16% to 55% during this time. This increase was attributed to the introduction of new guidelines aimed at increasing gender, race, and ethnicity reporting on RCTs. These guidelines included the National Institutes of Health Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research in 1994 [12], the CONSORT (Consolidated Standards of Reporting Trials) [13] in 2001, and the American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards (JARS) in 2008. However, the reporting of treatment effects by ethnic groups remained at a low at 2.1% [11]. Additionally, only non-Hispanic Black and Latino individuals were represented significantly more than in previous years. Asian Americans, multiracial individuals, Native Americans or Native Alaskans, and Native Hawaiians or Pacific Islanders were still underrepresented in the literature, with no significant change across time [11] despite evidence that participants are willing to engage in web-based health-related research [14].

Depression and Health Disparities

Existing research on depression and race-ethnicity presents a somewhat complex picture. In the United States, apart from Native Americans, who have the highest rates of depression, minoritized groups tend to have lower rates of depression than non-Hispanic White individuals [15]. However, minoritized individuals may be at a higher risk for more severely debilitating depression when compared to non-Hispanic White individuals [16]. Additionally, the costs associated with treating depression may impact racial groups differently. For example, those who are middle class and Black tend to encounter more obstacles to upward mobility (ie, moving from one social class to another) and are more susceptible to downward mobility, which may impede their access to mental health care [17,18]. In 2020, 53% Black individuals experienced worse access to care than White individuals. Similarly, 29% of Asian individuals had worse access to care, while American Indian or Alaska Native individuals presented the largest disparity with 50% receiving worse access to care than White individuals. Native Hawaiian and Other Pacific Islander individuals, however, experienced the same level of access to care as White individuals [19].

Furthermore, substantial barriers in access to mental health care tend to affect minoritized racial-ethnic groups more than nonminoritized groups. For example, there is evidence that interpersonal factors (eg, stigma and personal shame), sociocultural factors (eg, fear of negative evaluation by family or peers, preference for traditional coping strategies such as withdrawal and "accepting fate"), and systemic factors (eg, lack of culturally centered clinical environments and culturally responsive services and language barriers) can make it difficult for members of racial-ethnic minoritized groups to access mental health care [20]. These barriers may be particularly pronounced in face-to-face mental health treatment settings, where patients need to engage fully with another person for it to be successful; however, innovative internet-based delivery methods (eg, internet-based cognitive behavioral therapy [iCBT] self-help) may reduce barriers to treatment by making it more accessible [21]. Moreover, there is evidence that shows that Latinx and non-Hispanic Black individuals may be willing to engage in

XSL•FO
RenderX

smartphone-based interventions and bibliotherapy, respectively [22,23].

In addition to race and ethnicity, migration status is another crucial sociodemographic factor associated with depressive and anxious symptoms. For example, in a study involving 37,076 individuals from 20 European countries, first-generation migrants exhibited higher levels of depression, with rates significantly elevated for those born outside Europe [24]. Research on refugees resettled in high-income countries indicates that they experience increased rates of anxiety and depression compared to the general population [25]. These findings suggest that aspects such as migration status, country of origin, refugee status, and other related demographics play a significant role in predicting depression and anxiety symptoms, emphasizing the importance of providing mental health services to cater to these populations.

iCBT for Depression

One of the most studied and promising internet-based interventions for depression is iCBT. iCBT involves the provision of self-help materials (eg, websites, apps, and videos) that impart psychoeducation and teach skills that individuals can use to manage their symptoms. iCBT is usually delivered in 1 of 2 formats: guided or unguided. In guided self-help, individuals access self-help material and are assisted by a trained mental health professional or paraprofessional [26]. In unguided self-help, individuals access materials independently, without assistance or support. Both guided and unguided iCBT treatments have been shown to be more efficacious than waiting list controls [27].

Although unguided forms of self-help are more scalable and easier to access, guided self-help has been shown to be more effective than unguided self-help and as effective as face-to-face treatment [27]. Individuals may also adhere to it more closely [28]. Considering the wide support for the efficacy of iCBT, this format has the potential to reduce the public health burden of untreated depression, especially among individuals from a racial-ethnic minoritized group [21,29]. Given the potential for iCBT to reach minoritized communities, it is important to understand the reporting and representation of race-ethnicity in iCBT studies. If individuals from racial-ethnic minoritized groups are being underreported, it would be difficult to determine how well our current interventions meet the needs of these individuals.

For these reasons, we explored the reporting and representation of racial-ethnic diversity in clinical trials of iCBT for depression. Our first aim was to examine the overall reporting of individuals from racial-ethnic minoritized groups in published RCTs of iCBT for depression. Our second aim was to explore the representation of specific racial-ethnic minoritized groups in RCTs of iCBT for depression. To achieve these aims, we conducted a systematic review of RCTs of iCBT for depression and examined the reporting of race-ethnicity along with the representation (ie, the sample composition) [30-90]. We also explored the reporting of other background factors (eg, migration status).

Methods

Search Strategies

To identify RCTs, we employed a 2-pronged search strategy. First, we searched METAPSY, a database of randomized clinical trials for depression created by Cuijpers [91] and colleagues [92]. The public version of the database covers a search of trials between January 1, 1966, and January 2018, including Risk of Bias (ROB) assessments. We obtained an updated database version from one of the METAPSY lead researchers (P Cuijpers), which included studies up to January 1, 2021. The database of 763 studies was created from a search on PubMed, PsycINFO, Embase, and the Cochrane Library. The search was performed using terms involving psychotherapy (eg, “psychotherapy” and “cognitive-behavioral therapy”) and depression (eg, “depressive symptoms” and “major depression”). An example of the search string for PubMed can be found in Multimedia Appendix 1. Studies were included if they were an RCT in which a psychotherapy condition was compared to another. Control conditions included waiting list, treatment as usual, pill placebo, pharmacotherapy, alternate therapy delivery modalities (eg, single vs group therapy), or any other active control condition. Exclusion criteria included (1) no statement of randomization; (2) depression not being an inclusion criterion; (3) not being focused on acute treatment (eg, maintenance or relapse prevention); (4) studies with children or adolescents; (5) dissertation studies; (6) studies in which the depression was not the target of treatment; (7) if effect sizes could not be calculated; and (8) studies in languages other than English, Spanish, German, or Dutch. METAPSY contains a risk-of-bias assessment using the risk-of-bias assessment tool from the Cochrane Collaboration for papers published before 2018 [92]. Two raters (ARH-D and RDJ-R) examined the titles and abstracts of the 763 studies identified in METAPSY to decide whether they were relevant for review.

To obtain more recent papers, we used the same search string used for METAPSY [92] with the addition of a search term for internet-based studies to search Cochrane, PubMed, PsycINFO, and Embase for papers published between January 1, 2021, and July 18, 2022. An example search string for PubMed is available in Multimedia Appendix 2. This second search yielded an additional 2159 studies, of which 590 were duplicates. We also identified an additional paper by cross-referencing a study protocol. Further, 4 raters used a randomized, counterbalanced design to review the titles and abstracts of the remaining 1570 studies (ARH-D, JFB, LL-L, and RDJ-R). Disagreements between the raters were resolved by consensus.

We included randomized controlled trials in which iCBT was compared to a waiting list, care-as-usual, active control, or another iCBT treatment (eg, 2 different iCBTs or unguided vs guided iCBT). The included papers must have also focused on acute treatment (eg, 4 weeks to 6 months) of depression, delivered via the internet on a website or a smartphone app as guided or unguided self-help iCBT. In addition to the exclusion criteria used for METAPSY [91,92], we also excluded studies if the (1) interventions were not delivered via the internet, and (2) interventions focused on specific medical or psychiatric

XSL•FO
RenderX

subpopulations (eg, individuals with cannabis use disorder). We excluded studies that used specific subpopulation samples because we aimed to study the general reporting and representation of race-ethnicity in trials of iCBT for depression [93,94]. Additionally, we concentrated on studies that examined depression to prevent overlapping treatment outcomes (eg, chronic pain, sleep, and biases due to the specificity of the pool of participants) [95]. Of the 2333 studies, 282 were identified as relevant to self-help iCBT for depression. We then examined the full text of these studies and excluded those that did not meet our inclusion criteria. See Figure 1 for the PRISMA (Preferred Reporting Items for Systematic Reviews) flowchart. We assessed the ROB for 20 papers that did not already have an ROB assessment in the public version of the METAPSY database, 17 that were provided in the METAPSY update and 3 we found in our systematic search (Multimedia Appendix 3).

Figure 1. Flow diagram of papers at each stage of the screening process. CBT: cognitive behavioral therapy; RCT: randomized controlled trial; TAU: treatment as usual.

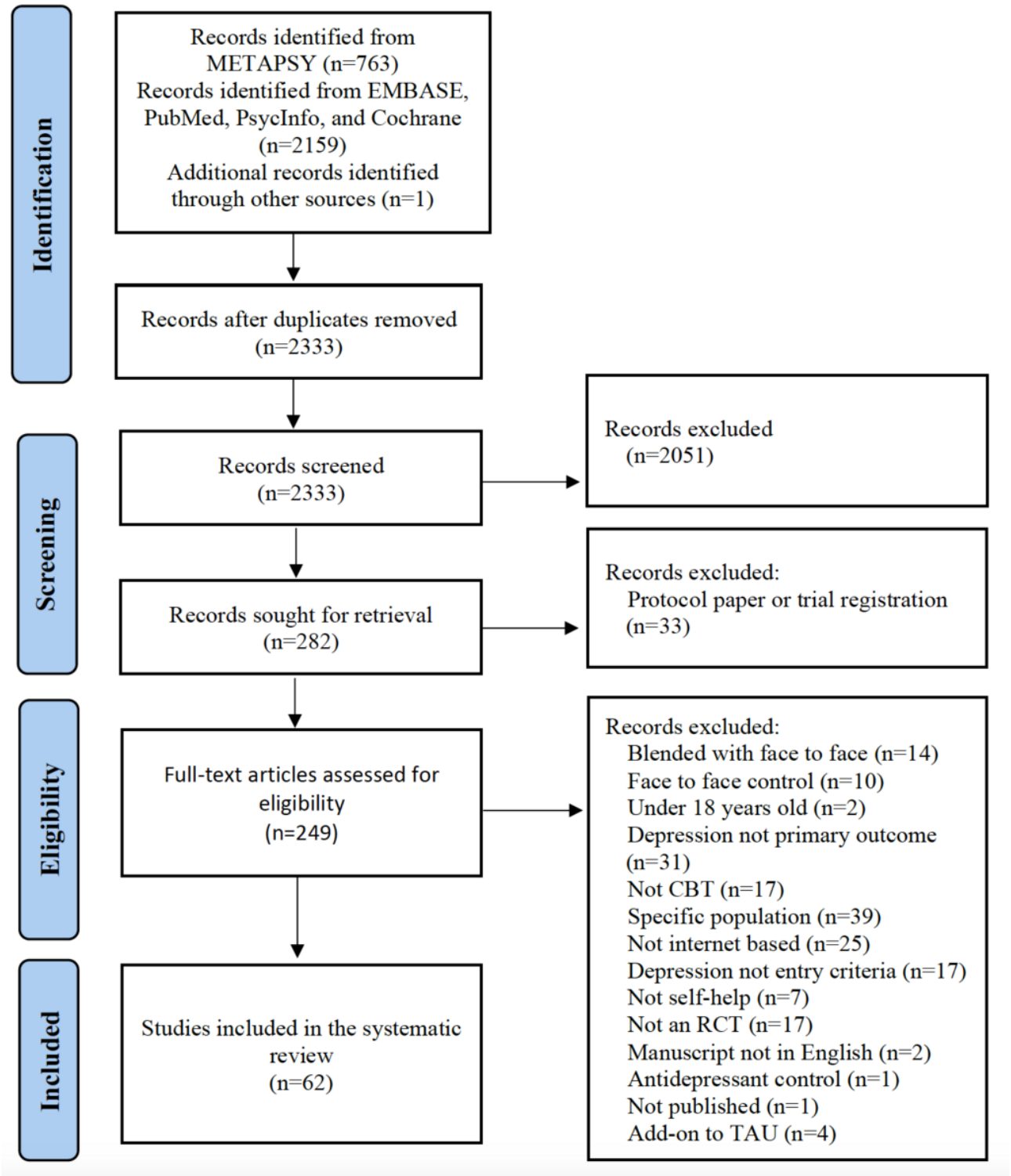

XSL•FO
RenderX

Coding

Coders rated study characteristics, trial design aspects, intervention type, and control group. They also coded various study-level features, including year of publication, whether the studies used guided or unguided iCBT and the country from which the study was sampled. For our primary question of interest, we first coded whether studies included race or ethnicity as "1" if they did or "0" if they did not. Additionally, when studies reported race or ethnicity, we recorded the number of participants in the sample that belonged to the different racial and ethnic groups. We then calculated the proportion of minoritized group members in the sample for the studies that reported overall race or ethnicity. Finally, given that we included studies conducted outside the United States, where there may be different conceptualizations of race and ethnicity, we also recorded other alternative variables indicative of participant background, including migration status, country of birth, country of residence, nationality, or native language. Agreement between raters was based on consensus, with disagreements being resolved by the principal investigator (LL-L).

Results

Study Selection and Characteristics

In total, 62 papers met the criteria of an RCT of iCBT for depression in adults. Some studies used samples from multiple countries or reported multiple arms, therefore, there are more interventions and control conditions than the number of studies (ie, 62). These studies included a total of 64 interventions of which 41 (64%) included guided self-help, and 23 (36%) included unguided self-help. Regarding control groups, 34 (52%) out of the 65 control groups included were compared to waiting lists, 12 (18%) to treatment as usual, 12 (18%) to active controls, and 7 (11%) used an internet-based active treatment comparator (ie, guided or unguided iCBT). The studies drew on samples from 13 countries: United States (n=17), Germany (n=12), Australia (n=10), Sweden (n=10), the Netherlands (n=6), China (n=2), Finland (n=2), United Kingdom (n=2), Canada (n=1), Colombia (n=1), Ireland (n=1), New Zealand (n=1), and Switzerland (n=1). The 62 papers included in this study had a total sample of 17,210 participants. ROB information is available in Multimedia Appendix 3 and suggests that reported studies had at least some ROB. However, most of this bias was due to self-reporting as an outcome measure. In the other ROB domains, 14 (70%) out of 20 of studies had low risk concerning missing outcome data, while 18 (90%) out of 20 of studies showed low risk for deviation from intended interventions.

Reporting of Race and Ethnicity

Of the 62 papers, only 17 (27%) reported race, and 12 (19%) reported ethnicity. Despite this overall low reporting rate, we observed that reporting differed markedly by the country sampled in the study (Fisher exact test $P<.001$). Of the studies conducted in the United States (n=17), almost all included race (n=15, 88%). Outside of the United States, only 2 studies conducted in the United Kingdom (n=1) and Australia (n=1) reported race. RCTs using samples from Germany, the Netherlands, Sweden, Ireland, Switzerland, China, Finland, Canada, Colombia, and New Zealand did not report race. Similarly, ethnicity was only reported in 12 studies, 4 of which grouped ethnicity with race. A fifth trial described reporting ethnicity, but it combined racial and ethnic categories; therefore, it was labeled as reporting both race and ethnicity. Additionally, 2 of them were conducted specifically on a minoritized group. The results are summarized in Table 1. Translating the reporting numbers to individuals, of the 17,210 participants in all these RCTs, 13,587 (78.9%) had no reported race, and 14,810 (86.1%) had no reported ethnicity.

Table 1. Number of published randomized controlled trials of internet-based cognitive behavioral therapy that reported race or ethnicity, by reported source country (N=62).

Country	Papers	Reported race, n (%)	Report ethnicity, n (%)
United States	17	15 (88)	9 (53)
Germany	12	0 (0)	0 (0)
Australia	10	1(10)	1 (10)
Sweden	10	0 (0)	0 (0)
Netherlands	6	0 (0)	1 (17)
China	2	0 (0)	0 (0)
Finland	2	0 (0)	0 (0)
United Kingdom	2	1 (50)	1 (50)
Canada	1	0 (0)	0 (0)
Colombia	1	0 (0)	0 (0)
Ireland	1	0 (0)	0 (0)
New Zealand	1	0 (0)	0 (0)
Switzerland	1	0 (0)	0 (0)

XSL•FO
RenderX

Representation of Individuals From Racial and Ethnic Minoritized Groups

Of the 17 papers that reported race, 4 (24%) categorized groups into "minority" versus "non-minority" or "White" versus "other." The other papers included the following groups on their breakdown: American Indian or Alaskan Native (n=3); Asian (n=8); Black or African American (n=12); Middle Eastern or North African (n=1); multiracial (n=8); Native American (n=1); other (n=8) which included "not specified," "declined to answer," and not reported as White; Pacific Islander (n=3); unknown (n=18); and White (n=16) (see Table 2).

Table 2. Total number of participants by race for trials of internet-based cognitive behavioral therapy for depression that reported race (N=3623).

Race	Participants, n (%)
AIAN[a]	14 (0.4)
Asian	209 (5.8)
Black	274 (7.6)
MENA[b]	1 (0)
Multiple	143 (3.9)
Native American	3 (0.1)
Not specified	75 (2.1)
Other	76 (2.1)
Pacific Islander	6 (0.2)
Unknown	106 (2.9)
White	2716 (74.9)

[a]AIAN: American Indian or Alaskan Native.

[b]MENA: Middle Eastern or North African.

The papers that reported ethnicity (Table 3) reported the following ethnic groups: American (n=1), Chinese (n=1), European (n=1), Hispanic or Latino (n=9), other (n=1), Turkish (n=1), and White British (n=1). Many papers specified ethnicity for minoritized member groups (eg, Hispanic) but not for majority members (eg, non-Hispanic).

Table 3. Total number of participants by ethnicity for trials of internet-based cognitive behavioral therapy for depression that reported ethnicity (N=2496).

Ethnicity	Participants, n (%)
American	2 (0)
Chinese	148 (5.9)
European	11(0)
Hispanic	281 (11.3)
Not specified (eg, reported ethnicity for some but not other individuals)	1886 (75.6)
Other	9 (0)
Turkish	96 (3.8)
White British	63 (2)

Representation of Race in iCBT Trials in the United States

Given that reporting of race was highest in the United States, we explored the representation of individuals from racial-ethnic minoritized groups in the 15 RCTs from that country that reported race (n=3354). A total of 2568 out of 3354 (76.6%, 95% CI 75.1-78.0) of individuals in these samples were identified as White. By way of comparison, 61.6% of adults in the United States were reported as being White in the US census [1]. These rates are significantly different (χ^2_1=316.9, P<.001). Extrapolating the number of individuals that would be expected to be White based on the 12-month prevalence of depression reported in the latest epidemiological study (National Epidemiologic Survey on Alcohol and Related Conditions-III: 10.4%) [96] and racial-ethnic differences reported in National Epidemiologic Survey on Alcohol and Related Conditions-III, one would expect 67.3% of individuals in a 12-month depression sample to be non-Hispanic White. The rate of White individuals represented in our study was statistically significantly different from 67.3% (χ^2_1=130.4, P<.001). When using lifetime prevalence rates, one would expect 72.5% of individuals with lifetime depression to be White. The rate of White individuals

XSL•FO
RenderX

we found was significantly higher (χ^2_1=27.6, P<.001), though it was a relatively small difference (4.1%).

Reporting of Other Group Demographics Outside of the United States

Out of the 46 articles conducted outside of the United States, only 25 (54%) reported on other demographic factors, such as language proficiency (n=13), country of residence (n=11), country of origin (n=5), migration status (n=2), and nationality (n=2). Most (n=11) of the studies that reported language only did so indirectly, as an inclusion criterion (eg, "participants must speak Dutch fluently"). Similarly, most (n=9) of the studies that reported the country of residence did so indirectly, (eg, "participants must reside in Sweden"). Some studies reported on multiple of the aforementioned demographics, therefore, there are more reported demographic variables (ie, 33) than the number of studies reporting these (ie, 25).

Discussion

Principal Findings

Our first aim was to examine the overall reporting of racial and ethnic diversity in RCTs of iCBT for depression. The second aim was to explore the representation of specific minoritized groups within the studies reporting race-ethnicity. Our search revealed substantial gaps in the literature. Out of the 62, only 17 (27%) of the papers included in this paper had data on race, and 12 (19%) included ethnicity. Focusing on the United States, even when race was reported, White individuals were overrepresented in samples relative to US census population estimates (61.6%) [97] and the expected estimate of non-Hispanic White individuals in depression samples (67.3%) [98].

Limitations

Before interpreting our findings, several caveats are worth noting. First, we recognize that failure to report racial-ethnic identity information does not guarantee that racial-ethnic diversity is not well-represented in a trial's sample. However, we believe that it is unlikely that researchers are collecting very racially diverse samples and choosing not to report on that aspect of the sample composition. In some countries reporting race or ethnicity might be more limited by regulations like ethical review board approval. Nonetheless, only 25 (54%) out of 46 of studies conducted outside of the United States reported other characteristics such as nationality, country of origin, or migration status. Additionally, it is worth noting that it is not always clear what the implications of underreporting are. For example, in a large individual-patient data meta-analysis, Karyotaki et al [99] reported that individuals from racial-ethnic minoritized groups experienced poorer outcomes in guided iCBT than "native-born" participants. These findings are worth replicating as they imply that racial-ethnic groups may respond differently to iCBTs and are concerning regarding the underrepresentation of minoritized groups in iCBT research. Nonetheless, even if race-ethnicity was not a predictor of outcomes in treatment, the differential enrollment of participants from different racial-ethnic groups is itself a racial-ethnic disparity. We excluded gray literature such as posters and other unpublished works. While these exclusions represent a limitation of this study (ie, fewer RCTs included), it is not clear that including these would have changed our conclusions that the reporting of representation is poor in RCTs of self-help iCBT. Finally, we chose to focus on iCBT, given its popularity and potential to reduce the public health burden of depression [100,101]. However, self-help iCBTs are not the only low-intensity treatment available. For example, bibliotherapy (ie, printed self-help media) is also a low-intensity treatment that may be preferred by many individuals, with some evidence suggesting that some racial-ethnic minoritized groups may prefer bibliotherapy over digital interventions [23].

Implications of Underrepresentation and Methodological Considerations

Our results suggest that the rate of reporting of ethnicity in iCBT studies in the United States has improved relative to the prior report by Polo and colleagues [11] for RCTs of face-to-face psychotherapy for depression. However, reporting of background characteristics in iCBT RCTs is relatively poor in Europe and elsewhere. Indeed, only 2 RCTs of iCBT for depression conducted outside of the United States reported race or ethnicity, one of which was a cultural adaptation. This finding is surprising given that an estimated 23 million non-European citizens (5% of the population) live in Europe [102]. This underreporting makes it challenging to determine how well our current interventions fit with the experiences of minoritized groups. While it is a possibility that logistic barriers such as internet access may be responsible for the lower representation of individuals from racial-ethnic minoritized groups in iCBT studies, it is unclear if these differences are large enough to account for the disparities in trial representation [103]. Ramos et al [29] noted that diversity, equity, and inclusion were not guiding principles in iCBT research. Similarly, our findings suggest a need to understand how individuals from racial-ethnic minoritized groups engage in iCBT research and how to increase that engagement.

Another explanation for the lack of reporting and representation could be group differences in the acceptability of iCBT (ie, internet-based self-help not being a preferred format). However, there is evidence that compared to non-Hispanic White adults, Asian, Hispanic, non-Hispanic Black, and other racial-ethnic groups report either equal or greater willingness to use and choose to learn about iCBTs. Thus, overall willingness to use iCBTs is unlikely to explain differences in initiating iCBT trials.

Furthermore, the relevance of race and ethnicity as terms in countries outside the United States warrants additional examination. It is essential to investigate alternative methods of describing samples that are appropriate for different countries while allowing for cross-country comparisons. The low number of iCBT studies focusing on minoritized individuals' mental health makes it challenging to advocate for the using of iCBT to reduce health disparities in these populations, especially outside the United States. Hence, it is essential to conduct research that encompasses diverse populations and reports on their characteristics.

XSL•FO
RenderX

Future Directions

This lack of knowledge regarding race and ethnicity has implications for implementing iCBT programs globally. Providers, for example, worry about the lack of diversity represented in the literature since this could lead to engagement challenges [104]. Acknowledging that a primarily White sample does not represent the general population may be an essential step to improving engagement. Numerous efforts to culturally adapt interventions to increase cultural competence of clinicians have been made [105]. However, we still lack information on how these adaptations translate to a self-help format to reduce barriers to access for minoritized groups. Currently, hundreds of online mental health resources are available to the public, but it is sometimes unclear what they offer users [106]. Clinicians and clients have no easy way of knowing which resources are evidence-based and suitable for individuals from racial-ethnic minoritized groups. The label "CBT" (cognitive behavioral therapy) could be a way to identify evidence-based mental health services. However, it is hard to determine whether digital interventions apply CBT principles correctly and whether this delivery format is ideal for members of racial-ethnic minoritized groups.

The extent to which racial-ethnic minoritized groups, other than US non-Hispanic Black adults, use self-help iCBT has received little attention [11,29]. It has been more than a decade since the National Institutes of Health, CONSORT, and JARS guidelines [12,13,107] introduced criteria regarding race and ethnicity, and the outlook has not improved significantly. Although reporting of race has improved, it is not being reported to the extent that it could be helpful. Further, a potential future direction includes developing new guidelines similar to the CONSORT and JARS [13,107] for journals oriented toward internet-based mental health. These guidelines could involve (1) including racial-ethnic groups with a breakdown; (2) describing the type and intensity of support being provided for guided self-help; (3) ensuring that samples are representative of gender, racial-ethnic groups, and other underreported identities; and (4) detailing which elements of CBT are included in the protocol. This standardization could facilitate the implementation of similar protocols in hospitals, mental health foundations, and businesses interested in providing mental health services to a broader population. The lack of reporting could also be improved by increasing outreach efforts and ensuring racial-ethnic diversity is included in RCTs for self-help iCBT.

Conclusions

There is a lack of reporting of racial and ethnic diversity in clinical trials of iCBT, and the representation of individuals from racial-ethnic minoritized groups is quite poor. This gap has significant implications for the generalizability of the findings currently in the literature, as these might not apply to individuals from racial-ethnic minoritized groups. Therefore, improving the representation of racial-ethnic diversity in trials of iCBT should be a key direction for the field going forward.

Acknowledgments

This research was partially funded by the Global Mental Health Fellowship (principal investigator: LL-L), the National Institute of Mental Health (grant T32 MH103213-06) and (grants KL2TR002530 and UL1TR002529) (A Shekhar, principal investigator) from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award (principal investigator: LL-L).

Conflicts of Interest

LL-L has received consulting fees from Syra Health, Inc who were not involved with the current work.

Multimedia Appendix 1
Search string from MetaPsy.
[DOCX File , 18 KB - jmir_v26i1e50780_app1.docx]

Multimedia Appendix 2
Search string for articles published from 2021 to 2022.
[DOCX File , 14 KB - jmir_v26i1e50780_app2.docx]

Multimedia Appendix 3
Risk of bias assessment for articles not included in MetaPsy (n=20).
[DOCX File , 305 KB - jmir_v26i1e50780_app3.docx]

References

1. U.S. Census Buerau. 2020 census illuminates racial and ethnic composition of the country. Census.gov. 2020. URL: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html [accessed 2022-06-06]
2. Wingrove-Haugland E, McLeod J. Not "minority" but "minoritized". Teach Ethics 2021;21(1):1-11. [doi: 10.5840/tej20221799]

XSL•FO
RenderX

3. Colby SL, Ortman JM. Projections of the size and composition of the U.S. population: 2014 to 2060. United States Census Bureau. 2015. URL: https://www.census.gov/library/publications/2015/demo/p25-1143.html [accessed 2023-12-05]
4. United Nations. International Migration 2019. Department of Economic Affaris. 2019. URL: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2019_internationalmigration_highlights.pdf [accessed 2023-12-05]
5. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 2007;370(9590):841-850 [FREE Full text] [doi: 10.1016/S0140-6736(07)61414-7] [Medline: 17826169]
6. Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry 2015;76(2):155-162 [FREE Full text] [doi: 10.4088/JCP.14m09298] [Medline: 25742202]
7. Pascoe EA, Richman LS. Perceived discrimination and health: a meta-analytic review. Psychological Bulletin 2009;135(4):531-554. [doi: 10.1037/a0016059]
8. Carter RT, Lau MY, Johnson V, Kirkinis K. Racial discrimination and health outcomes among racial/ethnic minorities: a meta-analytic review. J Multicult Couns Devel 2017;45(4):232-259. [doi: 10.1002/jmcd.12076]
9. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. Int Rev Psychiatry 2014;26(4):392-407. [doi: 10.3109/09540261.2014.928270] [Medline: 25137105]
10. Mak WWS, Law RW, Alvidrez J, Pérez-Stable EJ. Gender and ethnic diversity in NIMH-funded clinical trials: review of a decade of published research. Adm Policy Ment Health 2007;34(6):497-503. [doi: 10.1007/s10488-007-0133-z] [Medline: 17690976]
11. Polo AJ, Makol BA, Castro AS, Colón-Quintana N, Wagstaff AE, Guo S. Diversity in randomized clinical trials of depression: a 36-year review. Clin Psychol Rev 2019;67:22-35 [FREE Full text] [doi: 10.1016/j.cpr.2018.09.004] [Medline: 30292439]
12. NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research. National Institutes of Health. 2001. URL: http://grants1.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm [accessed 2022-06-06]
13. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Int J Surg 2011;9(8):672-677 [FREE Full text] [doi: 10.1016/j.ijsu.2011.09.004] [Medline: 22019563]
14. King DB, O'Rourke N, DeLongis A. Social media recruitment and online data collection: a beginner's guide and best practices for accessing low-prevalence and hard-to-reach populations. Can Psychol 2014;55(4):240-249. [doi: 10.1037/a0038087]
15. Budhwani H, Hearld KR, Chavez-Yenter D. Depression in racial and ethnic minorities: the impact of nativity and discrimination. J Racial Ethn Health Disparities 2015;2(1):34-42 [FREE Full text] [doi: 10.1007/s40615-014-0045-z] [Medline: 26863239]
16. Woodward AT, Taylor RJ, Abelson JM, Matusko N. Major depressive disorder among older African Americans, Caribbean Blacks, and non-Hispanic Whites: secondary analysis of the National Survey of American Life. Depress Anxiety 2013;30(6):589-597 [FREE Full text] [doi: 10.1002/da.22041] [Medline: 23319438]
17. McBrier DB, Wilson G. Going down? Race and downward occupational mobility for white-collar workers in the 1990s. Work Occup 2004;31(3):283-322. [doi: 10.1177/0730888404266383]
18. Hardaway CR, McLoyd VC. Escaping poverty and securing middle class status: how race and socioeconomic status shape mobility prospects for African Americans during the transition to adulthood. J Youth Adolesc 2009;38(2):242-256 [FREE Full text] [doi: 10.1007/s10964-008-9354-z] [Medline: 19636721]
19. 2021 National Healthcare Quality and Disparities Report. Agency for Healthcare Research and Quality. Rockville, MD; 2021. URL: https://www.ahrq.gov/research/findings/nhqrdr/nhqdr21/index.html [accessed 2023-12-05]
20. Holden KB, McGregor BS, Blanks SH, Mahaffey C. Psychosocial, socio-cultural, and environmental influences on mental health help-seeking among African-American men. J Mens Health 2012;9(2):63-69 [FREE Full text] [doi: 10.1016/j.jomh.2012.03.002] [Medline: 22905076]
21. Ramos G, Chavira DA. Use of technology to provide mental health care for racial and ethnic minorities: evidence, promise, and challenges. Cogn Behav Pract 2022;29(1):15-40. [doi: 10.1016/j.cbpra.2019.10.004]
22. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
23. De Jesús-Romero R, Wasil A, Lorenzo-Luaces L. Willingness to use internet-based versus bibliotherapy interventions in a representative US sample: cross-sectional survey study. JMIR Form Res 2022;6(8):e39508 [FREE Full text] [doi: 10.2196/39508] [Medline: 36001373]
24. Levecque K, Van Rossem R. Depression in Europe: does migrant integration have mental health payoffs? A cross-national comparison of 20 European countries. Ethn Health 2015;20(1):49-65. [doi: 10.1080/13557858.2014.883369] [Medline: 24517205]

XSL•FO
RenderX

25. Henkelmann JR, de Best S, Deckers C, Jensen K, Shahab M, Elzinga B, et al. Anxiety, depression and post-traumatic stress disorder in refugees resettling in high-income countries: systematic review and meta-analysis. BJPsych Open 2020;6(4):e68 [FREE Full text] [doi: 10.1192/bjo.2020.54] [Medline: 32611475]
26. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
27. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019;76(7):700-707 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
28. Mohr DC, Duffecy J, Ho J, Kwasny M, Cai X, Burns MN, et al. A randomized controlled trial evaluating a manualized TeleCoaching protocol for improving adherence to a web-based intervention for the treatment of depression. PLoS One 2013;8(8):e70086 [FREE Full text] [doi: 10.1371/journal.pone.0070086] [Medline: 23990896]
29. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021;147:103990 [FREE Full text] [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
30. Andersson G, Bergström J, Holländare F, Carlbring P, Kaldo V, Ekselius L. Internet-based self-help for depression: randomised controlled trial. Br J Psychiatry 2005;187:456-461. [doi: 10.1192/bjp.187.5.456] [Medline: 16260822]
31. Ly KH, Trüschel A, Jarl L, Magnusson S, Windahl T, Johansson R, et al. Behavioural activation versus mindfulness-based guided self-help treatment administered through a smartphone application: a randomised controlled trial. BMJ Open 2014;4(1):e003440 [FREE Full text] [doi: 10.1136/bmjopen-2013-003440] [Medline: 24413342]
32. Lindner P, Olsson E, Johnsson A, Dahlin M, Andersson G, Carlbring P. The impact of telephone versus e-mail therapist guidance on treatment outcomes, therapeutic alliance and treatment engagement in internet-delivered CBT for depression: a randomised pilot trial. Internet Interventions 2014;1(4):182-187. [doi: 10.1016/j.invent.2014.09.001]
33. Lappalainen P, Langrial S, Oinas-Kukkonen H, Tolvanen A, Lappalainen R. Web-based acceptance and commitment therapy for depressive symptoms with minimal support: a randomized controlled trial. Behav Modif 2015;39(6):805-834. [doi: 10.1177/0145445515598142] [Medline: 26253644]
34. Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther 2011;40(4):251-266. [doi: 10.1080/16506073.2011.616531] [Medline: 22060248]
35. Löbner M, Pabst A, Stein J, Dorow M, Matschinger H, Luppa M, et al. Computerized cognitive behavior therapy for patients with mild to moderately severe depression in primary care: A pragmatic cluster randomized controlled trial (@ktiv). J Affect Disord 2018;238:317-326. [doi: 10.1016/j.jad.2018.06.008] [Medline: 29902736]
36. Meyer B, Bierbrodt J, Schröder J, Berger T, Beevers C, Weiss M, et al. Effects of an Internet intervention (Deprexis) on severe depression symptoms: randomized controlled trial. Internet Interventions 2015;2(1):48-59. [doi: 10.1016/j.invent.2014.12.003]
37. Moritz S, Schilling L, Hauschildt M, Schröder J, Treszl A. A randomized controlled trial of internet-based therapy in depression. Behav Res Ther 2012;50(7-8):513-521. [doi: 10.1016/j.brat.2012.04.006] [Medline: 22677231]
38. Nyström MBT, Stenling A, Sjöström E, Neely G, Lindner P, Hassmén P, et al. Behavioral activation versus physical activity via the internet: a randomized controlled trial. J Affect Disord 2017;215:85-93 [FREE Full text] [doi: 10.1016/j.jad.2017.03.018] [Medline: 28319696]
39. Oehler C, Görges F, Rogalla M, Rummel-Kluge C, Hegerl U. Efficacy of a guided web-based self-management intervention for depression or dysthymia: randomized controlled trial with a 12-month follow-up using an active control condition. J Med Internet Res 2020;22(7):e15361 [FREE Full text] [doi: 10.2196/15361] [Medline: 32673233]
40. Stiles-Shields C, Montague E, Kwasny MJ, Mohr DC. Behavioral and cognitive intervention strategies delivered via coached apps for depression: pilot trial. Psychol Serv 2019;16(2):233-238 [FREE Full text] [doi: 10.1037/ser0000261] [Medline: 30407055]
41. O'moore KA, Newby JM, Andrews G, Hunter DJ, Bennell K, Smith J, et al. Internet cognitive-behavioral therapy for depression in older adults with knee osteoarthritis: a randomized controlled trial. Arthritis Care Res (Hoboken) 2018;70(1):61-70. [doi: 10.1002/acr.23257] [Medline: 28426917]
42. Salamanca-Sanabria A, Richards D, Timulak L, Connell S, Mojica Perilla M, Parra-Villa Y, et al. A culturally adapted cognitive behavioral internet-delivered intervention for depressive symptoms: randomized controlled trial. JMIR Ment Health 2020;7(1):e13392 [FREE Full text] [doi: 10.2196/13392] [Medline: 32003749]
43. Titov N, Dear BF, Staples LG, Terides MD, Karin E, Sheehan J, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: a randomized controlled trial. J Anxiety Disord 2015;35:88-102 [FREE Full text] [doi: 10.1016/j.janxdis.2015.08.002] [Medline: 26422822]
44. Ruwaard J, Schrieken B, Schrijver M, Broeksteeg J, Dekker J, Vermeulen H, et al. Standardized web-based cognitive behavioural therapy of mild to moderate depression: a randomized controlled trial with a long-term follow-up. Cogn Behav Ther 2009;38(4):206-221. [doi: 10.1080/16506070802408086] [Medline: 19221919]

XSL•FO
RenderX

45. Smith J, Newby JM, Burston N, Murphy MJ, Michael S, Mackenzie A, et al. Help from home for depression: a randomised controlled trial comparing internet-delivered cognitive behaviour therapy with bibliotherapy for depression. Internet Interv 2017;9:25-37 [FREE Full text] [doi: 10.1016/j.invent.2017.05.001] [Medline: 30135834]
46. Schröder J, Brückner K, Fischer A, Lindenau M, Köther U, Vettorazzi E, et al. Efficacy of a psychological online intervention for depression in people with epilepsy: a randomized controlled trial. Epilepsia 2014;55(12):2069-2076 [FREE Full text] [doi: 10.1111/epi.12833] [Medline: 25410633]
47. Richards D, Timulak L, O'Brien E, Hayes C, Vigano N, Sharry J, et al. A randomized controlled trial of an internet-delivered treatment: its potential as a low-intensity community intervention for adults with symptoms of depression. Behav Res Ther 2015;75:20-31. [doi: 10.1016/j.brat.2015.10.005] [Medline: 26523885]
48. Vernmark K, Lenndin J, Bjärehed J, Carlsson M, Karlsson J, Oberg J, et al. Internet administered guided self-help versus individualized e-mail therapy: a randomized trial of two versions of CBT for major depression. Behav Res Ther 2010;48(5):368-376. [doi: 10.1016/j.brat.2010.01.005] [Medline: 20152960]
49. Schure MB, Lindow JC, Greist JH, Nakonezny PA, Bailey SJ, Bryan WL, et al. Use of a fully automated internet-based cognitive behavior therapy intervention in a community population of adults with depression symptoms: randomized controlled trial. J Med Internet Res 2019 18;21(11):e14754 [FREE Full text] [doi: 10.2196/14754] [Medline: 31738173]
50. Titov N, Andrews G, Davies M, McIntyre K, Robinson E, Solley K. Internet treatment for depression: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010;5(6):e10939 [FREE Full text] [doi: 10.1371/journal.pone.0010939] [Medline: 20544030]
51. Ünlü Ince B, Cuijpers P, van 't Hof E, van Ballegooijen W, Christensen H, Riper H. Internet-based, culturally sensitive, problem-solving therapy for Turkish migrants with depression: randomized controlled trial. J Med Internet Res 2013;15(10):e227 [FREE Full text] [doi: 10.2196/jmir.2853] [Medline: 24121307]
52. Williams AD, Blackwell SE, Mackenzie A, Holmes EA, Andrews G. Combining imagination and reason in the treatment of depression: a randomized controlled trial of internet-based cognitive-bias modification and internet-CBT for depression. J Consult Clin Psychol 2013;81(5):793-799 [FREE Full text] [doi: 10.1037/a0033247] [Medline: 23750459]
53. Warmerdam L, van Straten A, Twisk J, Riper H, Cuijpers P. Internet-based treatment for adults with depressive symptoms: randomized controlled trial. J Med Internet Res 2008;10(4):e44 [FREE Full text] [doi: 10.2196/jmir.1094] [Medline: 19033149]
54. Zhao C, Wampold BE, Ren Z, Zhang L, Jiang G. The efficacy and optimal matching of an internet-based acceptance and commitment therapy intervention for depressive symptoms among university students: a randomized controlled trial in China. J Clin Psychol 2022;78(7):1354-1375. [doi: 10.1002/jclp.23329] [Medline: 35191525]
55. Pots WTM, Fledderus M, Meulenbeek PAM, ten Klooster PM, Schreurs KMG, Bohlmeijer ET. Acceptance and commitment therapy as a web-based intervention for depressive symptoms: randomised controlled trial. Br J Psychiatry 2016;208(1):69-77. [doi: 10.1192/bjp.bp.114.146068] [Medline: 26250745]
56. Pihlaja S, Lahti J, Lipsanen JO, Ritola V, Gummerus EM, Stenberg JH, et al. Scheduled telephone support for internet cognitive behavioral therapy for depression in patients at risk for dropout: pragmatic randomized controlled trial. J Med Internet Res 2020;22(7):e15732 [FREE Full text] [doi: 10.2196/15732] [Medline: 32706658]
57. Zagorscak P, Heinrich M, Sommer D, Wagner B, Knaevelsrud C. Benefits of individualized feedback in internet-based interventions for depression: a randomized controlled trial. Psychother Psychosom 2018;87(1):32-45. [doi: 10.1159/000481515] [Medline: 29306945]
58. Reins JA, Boß L, Lehr D, Berking M, Ebert DD. The more I got, the less I need? efficacy of internet-based guided self-help compared to online psychoeducation for major depressive disorder. J Affect Disord 2019;246:695-705. [doi: 10.1016/j.jad.2018.12.065] [Medline: 30611913]
59. Bücker L, Schnakenberg P, Karyotaki E, Moritz S, Westermann S. Diminishing effects after recurrent use of self-guided internet-based interventions in depression: randomized controlled trial. J Med Internet Res 2019;21(10):e14240 [FREE Full text] [doi: 10.2196/14240] [Medline: 31579014]
60. Roepke AM, Jaffee SR, Riffle OM, McGonigal J, Broome R, Maxwell B. Randomized controlled trial of SuperBetter, a smartphone-based/internet-based self-help tool to reduce depressive symptoms. Games Health J 2015;4(3):235-246. [doi: 10.1089/g4h.2014.0046] [Medline: 26182069]
61. Carlbring P, Hägglund M, Luthström A, Dahlin M, Kadowaki Å, Vernmark K, et al. Internet-based behavioral activation and acceptance-based treatment for depression: a randomized controlled trial. J Affect Disord 2013;148(2-3):331-337. [doi: 10.1016/j.jad.2012.12.020] [Medline: 23357657]
62. Bur OT, Krieger T, Moritz S, Klein JP, Berger T. Optimizing the context of support of web-based self-help in individuals with mild to moderate depressive symptoms: a randomized full factorial trial. Behav Res Ther 2022;152:104070 [FREE Full text] [doi: 10.1016/j.brat.2022.104070] [Medline: 35306266]
63. Choi I, Zou J, Titov N, Dear BF, Li S, Johnston L, et al. Culturally attuned internet treatment for depression amongst Chinese Australians: a randomised controlled trial. J Affect Disord 2012;136(3):459-468. [doi: 10.1016/j.jad.2011.11.003] [Medline: 22177742]
64. Perini S, Titov N, Andrews G. Clinician-assisted internet-based treatment is effective for depression: randomized controlled trial. Aust N Z J Psychiatry 2009;43(6):571-578. [doi: 10.1080/00048670902873722] [Medline: 19440890]

XSL•FO
RenderX

65. Jelinek L, Arlt S, Moritz S, Schröder J, Westermann S, Cludius B. Brief web-based intervention for depression: randomized controlled trial on behavioral activation. J Med Internet Res 2020;22(3):e15312 [FREE Full text] [doi: 10.2196/15312] [Medline: 32213470]
66. Johansson O, Bjärehed J, Andersson G, Carlbring P, Lundh LG. Effectiveness of guided internet-delivered cognitive behavior therapy for depression in routine psychiatry: a randomized controlled trial. Internet Interv 2019;17:100247 [FREE Full text] [doi: 10.1016/j.invent.2019.100247] [Medline: 31249791]
67. Flygare AL, Engström I, Hasselgren M, Jansson-Fröjmark M, Frejgrim R, Andersson G, et al. Internet-based CBT for patients with depressive disorders in primary and psychiatric care: is it effective and does comorbidity affect outcome? Internet Interv 2020;19:100303 [FREE Full text] [doi: 10.1016/j.invent.2019.100303] [Medline: 32055451]
68. Kenter RMF, Cuijpers P, Beekman A, van Straten A. Effectiveness of a web-based guided self-help intervention for outpatients with a depressive disorder: short-term results from a randomized controlled trial. J Med Internet Res 2016;18(3):e80 [FREE Full text] [doi: 10.2196/jmir.4861] [Medline: 27032449]
69. Krämer R, Köhler S. Evaluation of the online-based self-help programme "Selfapy" in patients with unipolar depression: study protocol for a randomized, blinded parallel group dismantling study. Trials 2021;22(1):264 [FREE Full text] [doi: 10.1186/s13063-021-05218-4] [Medline: 33836810]
70. Kivi M, Eriksson MCM, Hange D, Petersson EL, Vernmark K, Johansson B, et al. Internet-based therapy for mild to moderate depression in Swedish primary care: short term results from the PRIM-NET randomized controlled trial. Cogn Behav Ther 2014;43(4):289-298 [FREE Full text] [doi: 10.1080/16506073.2014.921834] [Medline: 24911260]
71. Fischer A, Schröder J, Vettorazzi E, Wolf OT, Pöttgen J, Lau S, et al. An online programme to reduce depression in patients with multiple sclerosis: a randomised controlled trial. Lancet Psychiatry 2015;2(3):217-223. [doi: 10.1016/S2215-0366(14)00049-2] [Medline: 26359900]
72. Tomasino KN, Lattie EG, Ho J, Palac HL, Kaiser SM, Mohr DC. Harnessing peer support in an online intervention for older adults with depression. Am J Geriatr Psychiatry 2017;25(10):1109-1119 [FREE Full text] [doi: 10.1016/j.jagp.2017.04.015] [Medline: 28571785]
73. Rosso IM, Killgore WDS, Olson EA, Webb CA, Fukunaga R, Auerbach RP, et al. Internet-based cognitive behavior therapy for major depressive disorder: a randomized controlled trial. Depress Anxiety 2017;34(3):236-245 [FREE Full text] [doi: 10.1002/da.22590] [Medline: 28009467]
74. Guo Y, Hong YA, Cai W, Li L, Hao Y, Qiao J, et al. Effect of a WeChat-based intervention (Run4Love) on depressive symptoms among people living with HIV in China: a randomized controlled trial. J Med Internet Res 2020;22(2):e16715 [FREE Full text] [doi: 10.2196/16715] [Medline: 32044751]
75. Hallgren M, Kraepelien M, Öjehagen A, Lindefors N, Zeebari Z, Kaldo V, et al. Physical exercise and internet-based cognitive-behavioural therapy in the treatment of depression: randomised controlled trial. Br J Psychiatry 2015;207(3):227-234. [doi: 10.1192/bjp.bp.114.160101] [Medline: 26089305]
76. Christensen H, Griffiths KM, Mackinnon AJ, Brittliffe K. Online randomized controlled trial of brief and full cognitive behaviour therapy for depression. Psychol Med 2006;36(12):1737-1746. [doi: 10.1017/S0033291706008695] [Medline: 16938144]
77. Christensen H, Griffiths KM, Jorm AF. Delivering interventions for depression by using the internet: randomised controlled trial. BMJ 2004;328(7434):265 [FREE Full text] [doi: 10.1136/bmj.37945.566632.EE] [Medline: 14742346]
78. Donker T, Bennett K, Bennett A, Mackinnon A, van Straten A, Cuijpers P, et al. Internet-delivered interpersonal psychotherapy versus internet-delivered cognitive behavioral therapy for adults with depressive symptoms: randomized controlled noninferiority trial. J Med Internet Res 2013;15(5):e82 [FREE Full text] [doi: 10.2196/jmir.2307] [Medline: 23669884]
79. Kingston J, Becker L, Woeginger J, Ellett L. A randomised trial comparing a brief online delivery of mindfulness-plus-values versus values only for symptoms of depression: does baseline severity matter? J Affect Disord 2020;276:936-944. [doi: 10.1016/j.jad.2020.07.087] [Medline: 32745830]
80. Forand NR, Barnett JG, Strunk DR, Hindiyeh MU, Feinberg JE, Keefe JR. Efficacy of guided iCBT for depression and mediation of change by cognitive skill acquisition. Behav Ther 2018;49(2):295-307 [FREE Full text] [doi: 10.1016/j.beth.2017.04.004] [Medline: 29530267]
81. Hobfoll SE, Blais RK, Stevens NR, Walt L, Gengler R. Vets prevail online intervention reduces PTSD and depression in veterans with mild-to-moderate symptoms. J Consult Clin Psychol 2016;84(1):31-42. [doi: 10.1037/ccp0000041] [Medline: 26322788]
82. Dahne J, Collado A, Lejuez CW, Risco CM, Diaz VA, Coles L, et al. Pilot randomized controlled trial of a Spanish-language behavioral activation mobile app (¡Aptívate!) for the treatment of depressive symptoms among United States Latinx adults with limited English proficiency. J Affect Disord 2019;250:210-217 [FREE Full text] [doi: 10.1016/j.jad.2019.03.009] [Medline: 30870770]
83. Clarke G, Reid E, Eubanks D, O'Connor E, DeBar LL, Kelleher C, et al. Overcoming Depression on the Internet (ODIN): a randomized controlled trial of an internet depression skills intervention program. J Med Internet Res 2002;4(3):E14 [FREE Full text] [doi: 10.2196/jmir.4.3.e14] [Medline: 12554545]

XSL•FO
RenderX

84. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
85. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR mHealth uHealth 2016;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]
86. Clarke G, Eubanks D, Reid E, Kelleher C, O'Connor E, DeBar LL, et al. Overcoming Depression on the Internet (ODIN) (2): a randomized trial of a self-help depression skills program with reminders. J Med Internet Res 2005;7(2):e16 [FREE Full text] [doi: 10.2196/jmir.7.2.e16] [Medline: 15998607]
87. Beevers CG, Pearson R, Hoffman JS, Foulser AA, Shumake J, Meyer B. Effectiveness of an internet intervention (Deprexis) for depression in a United States adult sample: a parallel-group pragmatic randomized controlled trial. J Consult Clin Psychol 2017;85(4):367-380. [doi: 10.1037/ccp0000171] [Medline: 28230390]
88. Dahne J, Lejuez CW, Diaz VA, Player MS, Kustanowitz J, Felton JW, et al. Pilot randomized trial of a self-help behavioral activation mobile app for utilization in primary care. Behav Ther 2019;50(4):817-827 [FREE Full text] [doi: 10.1016/j.beth.2018.12.003] [Medline: 31208690]
89. Clarke G, Kelleher C, Hornbrook M, Debar L, Dickerson J, Gullion C. Randomized effectiveness trial of an internet, pure self-help, cognitive behavioral intervention for depressive symptoms in young adults. Cogn Behav Ther 2009;38(4):222-234 [FREE Full text] [doi: 10.1080/16506070802675353] [Medline: 19440896]
90. de Graaf LE, Gerhards SAH, Arntz A, Riper H, Metsemakers JFM, Evers SMAA, et al. Clinical effectiveness of online computerised cognitive-behavioural therapy without support for depression in primary care: randomised trial. Br J Psychiatry 2009;195(1):73-80 [FREE Full text] [doi: 10.1192/bjp.bp.108.054429] [Medline: 19567900]
91. Cuijpers P. Four decades of outcome research on psychotherapies for adult depression: an overview of a series of meta-analyses. Can Psychol 2017;58(1):7-19. [doi: 10.1037/cap0000096]
92. Cuijpers P, Karyotaki E, Ebert D, Harrer M. METAPSY Database. 2021. URL: https://www.metapsy.org/database/depression-psychotherapy [accessed 2022-06-06]
93. Lorenzo-Luaces L, Zimmerman M, Cuijpers P. Are studies of psychotherapies for depression more or less generalizable than studies of antidepressants? J Affect Disord 2018;234:8-13 [FREE Full text] [doi: 10.1016/j.jad.2018.02.066] [Medline: 29522947]
94. Lorenzo-Luaces L, Johns E, Keefe JR. The generalizability of randomized controlled trials of self-guided internet-based cognitive behavioral therapy for depressive symptoms: systematic review and meta-regression analysis. J Med Internet Res 2018;20(11):e10113 [FREE Full text] [doi: 10.2196/10113] [Medline: 30413400]
95. Lopez-Quintero C, de los Cobos JP, Hasin DS, Okuda M, Wang S, Grant BF, et al. Probability and predictors of transition from first use to dependence on nicotine, alcohol, cannabis, and cocaine: results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Drug Alcohol Depend 2011;115(1-2):120-130 [FREE Full text] [doi: 10.1016/j.drugalcdep.2010.11.004] [Medline: 21145178]
96. Smith SM, Goldstein RB, Grant BF. The association between post-traumatic stress disorder and lifetime DSM-5 psychiatric disorders among veterans: data from the National Epidemiologic Survey on Alcohol and Related Conditions-III (NESARC-III). J Psychiatr Res 2016;82:16-22 [FREE Full text] [doi: 10.1016/j.jpsychires.2016.06.022] [Medline: 27455424]
97. Race and ethnicity in the United States: 2010 census and 2020 census. US Census Bureau. 2021. URL: https://www.census.gov/library/visualizations/interactive/race-and-ethnicity-in-the-united-state-2010-and-2020-census.html [accessed 2023-12-05]
98. Hasin DS, Sarvet AL, Meyers JL, Saha TD, Ruan WJ, Stohl M, et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. JAMA Psychiatry 2018;75(4):336-346 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.4602] [Medline: 29450462]
99. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, et al. Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. Clin Psychol Rev 2018;63:80-92 [FREE Full text] [doi: 10.1016/j.cpr.2018.06.007] [Medline: 29940401]
100. Kazdin AE, Rabbitt SM. Novel models for delivering mental health services and reducing the burdens of mental illness. Clin Psychol Sci 2013;1(2):170-191. [doi: 10.1177/2167702612463566]
101. Kazdin AE, Blase SL. Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspect Psychol Sci 2011;6(1):21-37. [doi: 10.1177/1745691610393527] [Medline: 26162113]
102. Migration and migrant population statistics. Eurostat. 2022. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Migration_and_migrant_population_statistics [accessed 2023-12-05]
103. Internet/broadband fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2023-12-05]
104. Southam-Gerow MA, Rodríguez A, Chorpita BF, Daleiden EL. Dissemination and implementation of evidence based treatments for youth: challenges and recommendations. Prof Psychol Res Pr 2012;43(5):527-534. [doi: 10.1037/a0029101]

XSL•FO
RenderX

105. Huey SJ, Tilley JL, Jones EO, Smith CA. The contribution of cultural competence to evidence-based care for ethnically diverse populations. Annu Rev Clin Psychol 2014;10:305-338 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032813-153729] [Medline: 24437436]
106. Jennings K. Venture funding for mental health startups hits record high as anxiety, depression skyrocket. Forbes Magazine. 2021. URL: https://www.forbes.com/sites/katiejennings/2021/06/07/venture-funding-for-mental-health-startups-hits-record-high-as-anxiety-depression-skyrocket/?sh=76021c8c1116 [accessed 2022-06-06]
107. APA Publications and Communications Board Working Group on Journal Article Reporting Standards. Reporting standards for research in psychology: why do we need them? what might they be? Am Psychol 2008;63(9):839-851 [FREE Full text] [doi: 10.1037/0003-066X.63.9.839] [Medline: 19086746]

Abbreviations

CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
iCBT: internet-based cognitive behavioral therapy
JARS: Journal Article Reporting Standards
PRISMA: Preferred Reporting Items for Systematic Reviews
RCT: randomized controlled trial
ROB: Risk of Bias

Edited by G Eysenbach; submitted 12.07.23; peer-reviewed by N Mathews; comments to author 25.09.23; revised version received 05.11.23; accepted 16.11.23; published 01.02.24.

Please cite as:
De Jesús-Romero R, Holder-Dixon AR, Buss JF, Lorenzo-Luaces L
Race, Ethnicity, and Other Cultural Background Factors in Trials of Internet-Based Cognitive Behavioral Therapy for Depression: Systematic Review
J Med Internet Res 2024;26:e50780
URL: https://www.jmir.org/2024/1/e50780
doi:10.2196/50780
PMID:38300699

©Robinson De Jesús-Romero, Amani R Holder-Dixon, John F Buss, Lorenzo Lorenzo-Luaces. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Telemedicine for Musculoskeletal Disorders: Umbrella Review

Silvia Bargeri[1], MSc; Greta Castellini[1], PhD; Jacopo Antonino Vitale[2], PhD; Stefania Guida[1], MSc; Giuseppe Banfi[3,4], MD; Silvia Gianola[1], PhD; Federico Pennestrì[3], PhD

[1]Unità di Epidemiologia Clinica, IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
[2]Spine Center, Schulthess Klinik, Zürich, Switzerland
[3]IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
[4]Università Vita-Salute San Raffaele, Milan, Italy

Corresponding Author:
Greta Castellini, PhD
Unità di Epidemiologia Clinica
IRCCS Istituto Ortopedico Galeazzi
Via Cristina Belgioioso, 173
Milan, 20161
Italy
Phone: 39 02 8350 2260
Email: greta.castellini@grupposandonato.it

Abstract

Background: Several systematic reviews (SRs) assessing the use of telemedicine for musculoskeletal conditions have been published in recent years. However, the landscape of evidence on multiple clinical outcomes remains unclear.

Objective: We aimed to summarize the available evidence from SRs on telemedicine for musculoskeletal disorders.

Methods: We conducted an umbrella review of SRs with and without meta-analysis by searching PubMed and EMBASE up to July 25, 2022, for SRs of randomized controlled trials assessing telemedicine. We collected any kind of patient-reported outcome measures (PROMs), patient-reported experience measures (PREMs), and objective measures, including direct and indirect costs. We assessed the methodological quality with the AMSTAR 2 tool (A Measurement Tool to Assess systematic Reviews 2). Findings were reported qualitatively.

Results: Overall, 35 SRs published between 2015 and 2022 were included. Most reviews (n=24, 69%) were rated as critically low quality by AMSTAR 2. The majority of reviews assessed “telerehabilitation” (n=29) in patients with osteoarthritis (n=13) using PROMs (n=142 outcomes mapped with n=60 meta-analyses). A substantive body of evidence from meta-analyses found telemedicine to be beneficial or equal in terms of PROMs compared to conventional care (n=57 meta-analyses). Meta-analyses showed no differences between groups in PREMs (n=4), while objectives measures (ie, “physical function”) were mainly in favor of telemedicine or showed no difference (9/13). All SRs showed notably lower costs for telemedicine compared to in-person visits.

Conclusions: Telemedicine can provide more accessible health care with noninferior results for various clinical outcomes in comparison with conventional care. The assessment of telemedicine is largely represented by PROMs, with some gaps for PREMs, objective measures, and costs.

Trial Registration: PROSPERO CRD42022347366; https://osf.io/pxedm/

(J Med Internet Res 2024;26:e50090) doi:10.2196/50090

KEYWORDS

telemedicine; systematic review; patient-reported outcomes measures; PROMs; patient-reported experience measures; PREMs; rehabilitation; musculoskeletal disorder; condition; musculoskeletal; patient-reported; telerehabilitation; orthopedics; meta-analyses; osteoarthritis; accessibility; health care

XSL•FO
RenderX

Introduction

Telemedicine is a broad term encompassing many applications, such as diagnostic asynchronous evaluation, continuous monitoring using biosensors, and synchronous video consultations, including multiple variations on each theme. This definition includes "telerehabilitation," "health technologies," "digital medicine," and similar keywords [1,2]. During the COVID-19 pandemic, the spread of smart devices and accessible internet connections made telemedicine grow exponentially, become increasingly popular, and be tested for many health conditions [3,4]. Musculoskeletal disorders usually require multidisciplinary and multifacility treatment throughout different settings and providers (physiotherapy, rehabilitation, prehabilitation, and orthopedics; inpatient, outpatient, and home); therefore, the adoption of telemedicine can improve clinical and patient-reported outcomes along with organizational arrangements and cost savings [5,6]. Telemedicine in rehabilitation first appeared in a scientific publication in 1998 [7]; rehabilitation is an old branch of medicine, and in the last 20 years, new telemedicine practices have been developed showing an interest in understanding its effectiveness. However, there is not yet a universal definition of telemedicine nor a consensus on its effects.

In addition, in recent years, an increasing number of studies have used patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) to evaluate telemedicine services [8]. With the increasing maturity of telemedicine applications and higher evidence levels, the use of PROMs has increased. PREMs in turn are useful to describe the health care service experience from the perspective of patients [9-11] in order to identify real-world factors (eg, organizational, relational, environmental) that may improve or hamper the access to, quality of, and safety of care [12]: in the case of telemedicine, the user-friendliness of a certain technology, its actual functioning in ordinary settings (eg, considering backlogs and poor internet connections), the clarity of the instructions received, and the degree of interoperability among different providers (health care facilities, professionals, and technology suppliers). Otherwise, excellent technical care may be wasted by lack of compliance, poor health literacy, and insufficient patient engagement.

Different systematic reviews (SRs) [13-15] have been published examining different types of telemedicine solutions in specific musculoskeletal populations. However, it is crucial to investigate evidence on not just a single question but across multiple questions pertaining to a specific topic [16]. This emphasizes the importance of providing the best available evidence on the effectiveness of telemedicine and telerehabilitation in the entire musculoskeletal field, encompassing all telemedicine applications.

The aim of this umbrella review is to explore the effectiveness of telemedicine and rehabilitation in the treatment of musculoskeletal conditions in terms of physical impairment, function, health-related quality of life (HRQoL), adverse events, adherence, and costs, including PROMs and PREMs. This would help professionals improve decision-making and yield better clinical outcomes for patients.

Methods

Study Design

We conducted an umbrella review according to the Cochrane Handbook's chapter on overviews of reviews and the *JBI Manual for Evidence Synthesis*. Reviews of SRs are referred to by several different names in the scientific literature, including "umbrella reviews," "overviews of reviews," "reviews of reviews," "summaries of systematic reviews," and "synthesis of reviews" [17,18]. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [19] guidelines for the flow chart and the Preferred Reporting Items for Overviews of Reviews (PRIOR) [20,21] as a reporting checklist. The review protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO) database (CRD42022347366).

Criteria for Considering Reviews for Inclusion

According to Cochrane's definition, an SR is a review of the literature that "attempts to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a specific research question by using explicit, systematic methods that are selected with a view aimed at minimizing bias, producing more reliable findings to inform decision making" [22]. In this umbrella review, we considered inclusion criteria according to the PICOS (population, intervention, comparison, outcomes, study design) format: patients with any musculoskeletal or orthopedic condition (population); any kind of interventions based on advanced technology systems named as "telemedicine," "telerehabilitation," "health technologies" and "digital medicine," delivered both in synchronous and asynchronous modalities (intervention); in-person treatment or usual care or no treatment (comparison); PROMs, PREMs, or objective measures (outcomes); SRs of randomized controlled trials (RCTs) (study design).

SRs were excluded if they assessed (1) observational studies, (2) mixed populations (eg, if they combined effects for cardiovascular and musculoskeletal patients, or if it was not possible to separate data for a population of interest, such as musculoskeletal patients), or (3) interventions that aimed to configure technical aspects of devices and apps.

Main Outcomes

We considered outcomes related to the following domains: physical impairment, function, health-related quality of life (HRQoL), adverse events, and adherence to therapy or care pathways, expressed as the following categories: PROMs, PREMs, or objective measurements. PROMs are used to assess a patient's health status at a particular point in time (eg, common symptoms, pain, stiffness, HRQoL, and disease-specific interference with domestic activities and leisure time). Furthermore, we considered PREMs as any patient-reported information about the experience of treatment (eg, inclusivity of a technology, adequate communication with health care professionals, availability of professionals, possibility to ask

XSL•FO
RenderX

questions, clarity of information received, spared visits to hospital, access to treatment, and perceived safety). Costs related to treatments were also collected.

Search Strategy

A search of SRs was performed in PubMed and Embase from inception to July 25, 2022. The search was restricted to English-language publications. No restriction on year was applied (Multimedia Appendix 1 [8,23], Table S1).

Data Extraction (Selection and Coding)

Two independent reviewers consulted information screening sources by title and abstract. The full text of relevant studies was downloaded and evaluated for final inclusion according to the inclusion criteria. Any conflict was resolved through discussion. EndNote (version 20; Clarivate) and Rayyan (Qatar Computing Research Institute) were used to manage the study selection phase. The selection process is shown in the PRISMA flow chart [24].

Two independent reviewers extracted general characteristics of the SRs (eg, country, funding, and conflicts of interest), population characteristics, interventions, comparisons, and outcomes. Patient populations were classified according to the International Classification of Diseases, 10th revision (ICD-10) codes for diseases of the musculoskeletal system and connective tissue (M00-M99) [25].

Interventions were classified into three categories, building on a published taxonomy [8]: (1) teleconsultation (providing health care over a distance), telediagnostics (identifying diseases over a distance), and telemonitoring (collecting data over a distance to allow medical decisions); (2) telerehabilitation (collecting data over a distance to help patients cope with the long-term consequences of disease or impairment); and (3) digital self-management (to promote patient health responsibility and encourage health literacy).

Outcomes were also classified into subgroups for 3 categories: PROMs, PREMs, and objective outcomes. Among PROMs, we examined the following outcome domains: pain, HRQoL, physical function, social function, emotional function, cognitive function, health literacy, side effects, and adherence. We categorized PREMs into 2 outcome domains: treatment and technology. In Multimedia Appendix 1 [8,23], Table S2, we report all details about PROM and PREM taxonomy [8]. Objective measures were mapped into the same categories (eg, physical function) and costs.

Quantitative results were extracted from meta-analyses if reviews included more than 1 study in each analysis. For continuous outcomes, we extracted mean difference (MD) or standardized mean difference (SMD) with their 95% CIs. For binary outcomes, we extracted odds ratios (ORs) or relative risks (RRs) with their 95% CIs. In absence of quantitative syntheses, results were reported descriptively.

Risk of Bias (Quality) Assessment

The methodological quality of reviews was assessed using AMSTAR 2 (A Measurement Tool to Assess systematic Reviews 2) [26]. It consists of 16 items rating the quality of each SR as high, moderate, low, or critically low. Two reviewers independently performed the assessments; a third reviewer resolved any disagreement between reviewers.

Strategy for Data Synthesis

We followed the methodology outlined in the *Cochrane Handbook* chapter on overviews of reviews and the *JBI Manual for Evidence Synthesis* [18]. The characteristics of the included SRs were described with narrative synthesis. The main results reported in the reviews at the shortest follow-up were summarized. Qualitative results reported by the reviews were narratively summarized, while meta-analyzed results were visually presented in terms of directions of effects (favor intervention, no difference between groups, favor control) by a bubble plot map in which bubbles were organized into subgroups based on the AMSTAR 2 assessment, proportional to the number of participants for each meta-analysis and colored by outcomes and population [27].

Since we expected heterogeneity across reviews due to the different populations, interventions, controls, and outcomes included, the number of overlapping primary studies included in the SRs was not assessed.

Results

Study Selection

After removing duplicates, 3598 records were identified. Of the 157 full texts assessed, 122 SRs were excluded and 35 were included. Figure 1 shows the PRISMA flowchart. The reasons for excluding certain studies after reading their full texts are presented in Multimedia Appendix 2 [13-15,28-59], Table S1.

XSL•FO
RenderX

Figure 1. Flow diagram.

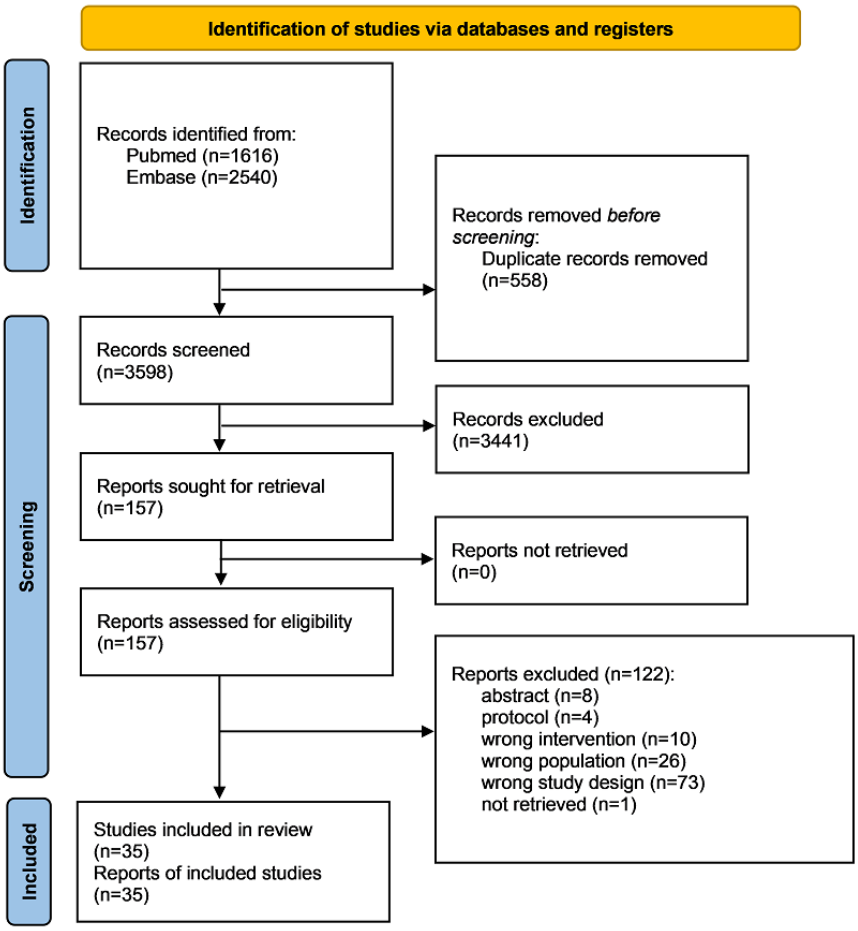

Characteristics of Included SRs

Most SRs were published in Europe (n=15), followed by the Americas (n=10), Asia (n=7), and Oceania (n=3). The median publication year was 2021 (minimum 2015, maximum 2022), with 77% (n=27) of SRs published in the last 3 years (2020-2022). Most SRs assessed telerehabilitation (n=29). A few assessed digital self-management (n=4) and teleconsultation, telediagnostics, and telemonitoring (n=2). The most common musculoskeletal disorder investigated was osteoarthritis (eg, hip and knee replacement; n=13) and mixed conditions (eg, chronic musculoskeletal pain; n=14). Twenty-two SRs (63%) reported a meta-analysis. More details are reported in Table 1.

XSL•FO
RenderX

Table 1. General characteristics of included studies (n=35).

Characteristics	Studies, n (%)
Country	
China	6 (17)
Spain	4 (11)
United Kingdom	4 (11)
United States	4 (11)
Australia	3 (9)
Brazil	3 (9)
Canada	3 (9)
Finland	2 (6)
Germany	2 (6)
Ireland	1 (3)
Italy	1 (3)
Netherlands	1 (3)
Pakistan	1 (3)
Years	
2020-2022	27 (77)
2017-2019	7 (20)
2014-2016	1 (3)
Conflicts of interest	
No	31 (88)
Not reported	2 (6)
Reported	2 (6)
Funding	
Not reported	21 (60)
No profit	11 (31)
Mixed	3 (9)
Population	
Mixed	14 (40)
Osteoarthritis	13 (37)
Other dorsopathies	4 (11)
Other disorders of the musculoskeletal system and connective tissue	3 (9)
Autoinflammatory syndromes	1 (3)
Intervention	
Telerehabilitation	29 (83)
Digital self-management	4 (11)
Teleconsultation, telediagnostics, monitoring	2 (6)
Meta-analysis	
Yes	22 (63)
No	13 (27)

XSL•FO
RenderX

Risk of Bias (Quality) Assessment

Methodological quality was critically low in 69% (n=24) of the SRs, low in 29% (n=10), and moderate in 1. Among critical items, 80% of SRs did not report reasons for exclusion (n=28), 48% did not account for risk of bias in individual studies when interpreting or discussing the results of the review (n=17), and 31% did not carry out an adequate investigation of publication bias (eg, small study bias; n=11). A summary plot is shown in Figure 2 [13-15,28-59].

Figure 2. AMSTAR 2 (A Measurement Tool to Assess systematic Reviews 2) summary plot. Item 1: "Did the research questions and inclusion criteria for the review include the components of PICO?"; item 2: "Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?"; item 3: "Did the review authors explain their selection of the study designs for inclusion in the review?"; item 4: "Did the review authors use a comprehensive literature search strategy?"; item 5: "Did the review authors perform study selection in duplicate?"; item 6: "Did the review authors perform data extraction in duplicate?"; item 7: "Did the review authors provide a list of excluded studies and justify the exclusions?"; item 8: "Did the review authors describe the included studies in adequate detail?"; item 9: "Did the review authors use a satisfactory technique for assessing the risk of bias RoB in individual studies that were included in the review?"; item 10: "Did the review authors report on the sources of funding for the studies included in the review?"; item 11: "If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?"; item 12: "If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?"; item 13: "Did the review authors account for RoB in primary studies when interpreting/discussing the results of the review?"; item 14: "Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?"; item 15: "If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias small study bias and discuss its likely impact on the results of the review?"; item 16: "Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?".

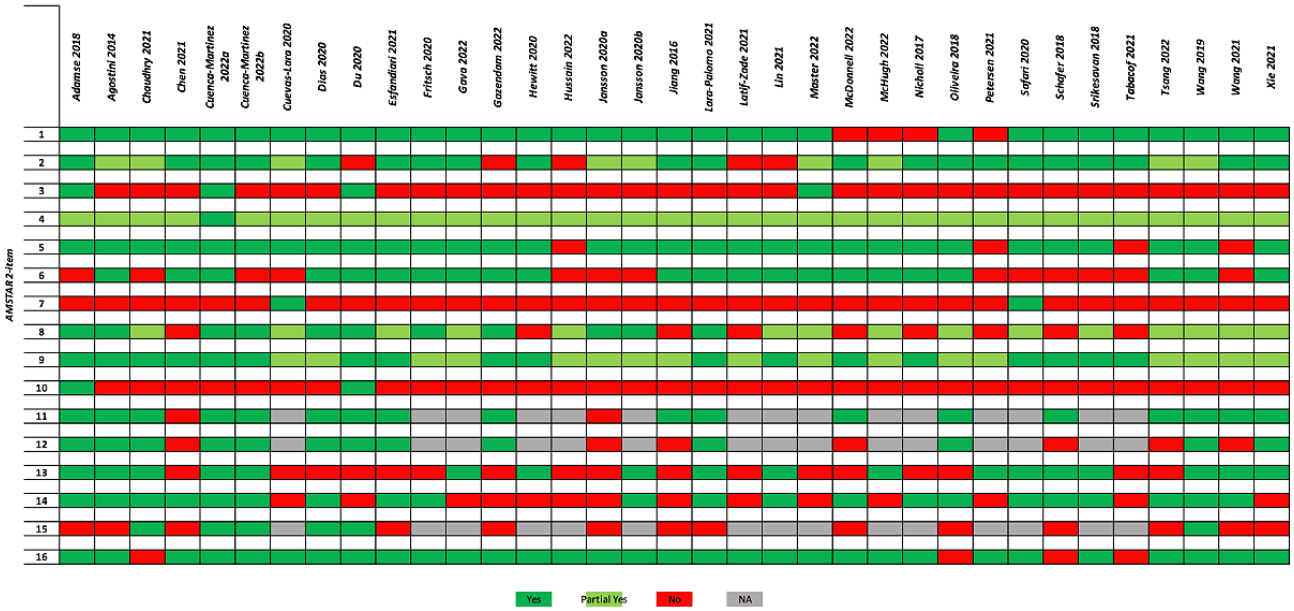

Outcome Characteristics

Overall, 190 outcomes were collected, among which were 142 PROMs and 15 PREMs; 31 objective outcomes were also collected, including 19 related to physical function, 4 to physical activity (eg, steps per day), and 8 to costs; 2 were composite outcomes (ie, PROMs plus objective measures). The most reported PROMs were related to pain assessment (n=40 in 29 reviews), physical function (n=35 in 27 reviews), and HRQoL (n=20 in 20 reviews). Among PREMs, technology was assessed in 4 reviews (11%) and treatment in 8 reviews (23%). Figure 3 illustrates the outcomes assessed by the reviews in subgroup for PROMs, PREMs, and objective measures.

XSL•FO
RenderX

Figure 3. Outcomes reported by reviews. The frequencies of each outcome category reflect the number of reviews addressing them. Inner circles represent the 3 categories defined in the methods: objective outcomes, patient-reported outcome measures (PROMs), and patient-reported experience measures (PREMs); outer circles represent the outcome domains (eg, physical function). HrQoL: health-related quality of life.

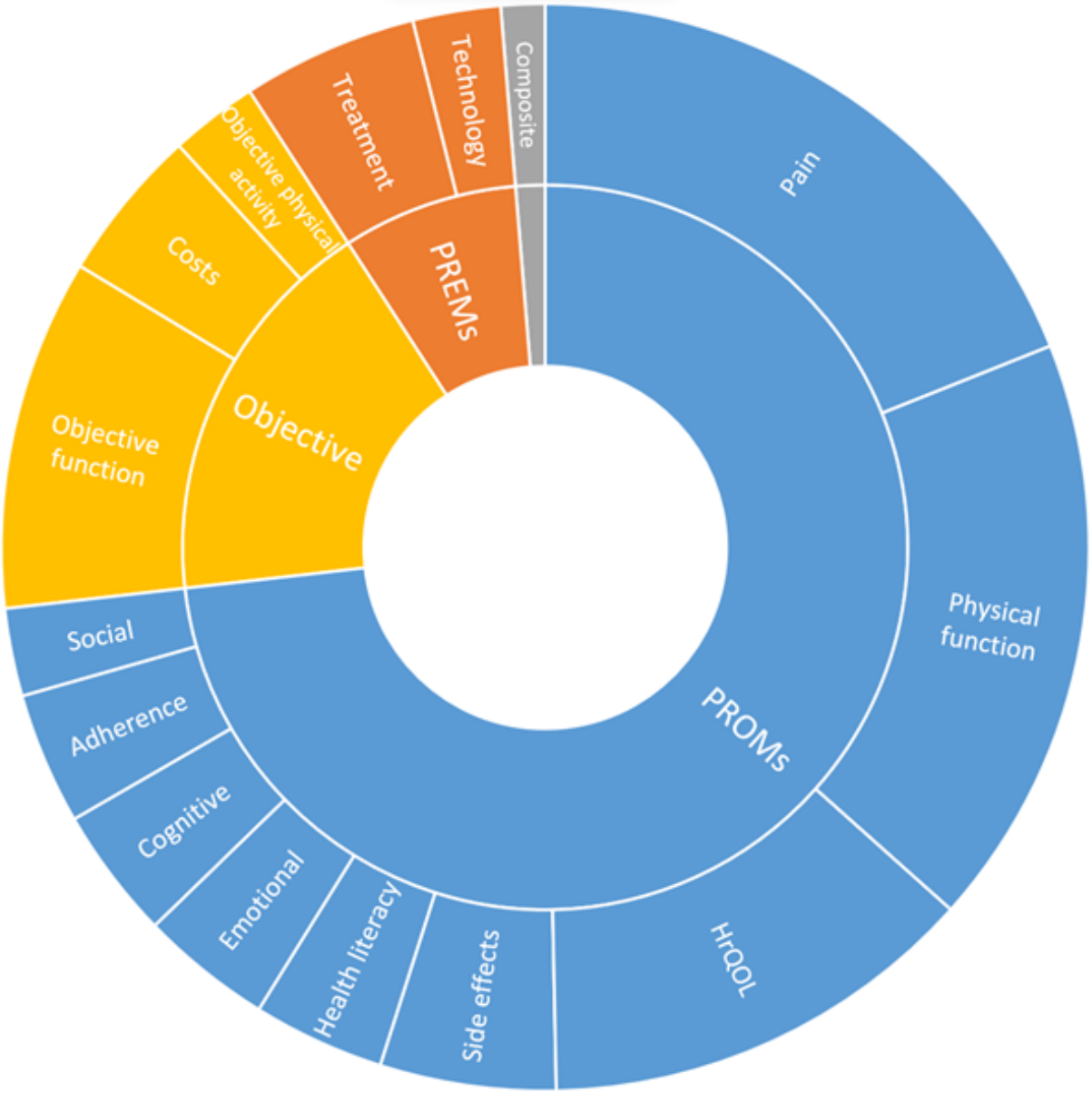

Summary of Quantitative Analyses

A total of 23 reviews (66%) reported a meta-analysis for a total of 79 analyses. Among these, 60 meta-analyses were performed on PROMs, 4 on PREMs, 13 on objective outcomes, and 2 on composite outcomes (ie, PROMs plus objective measurements).

Most PROMs were quantitatively analyzed in the following categories: pain assessment (n=23), physical function (n=21), HRQoL (n=8), cognitive function (n=2), emotional function (n=4), health literacy (n=1), and social function (n=1). The 37% (n=22) of meta-analyses assessing PROMs had results favoring telemedicine, 58% (n=35) found no differences between groups, and a few favored controls (n=3). No quantitative analyses for side effects or adherence were found. Meta-analyses of PREMs analyzed only treatment experiences and none showed any difference between groups. All the 13 objective outcomes meta-analyzed were on physical function (eg, balance test, Time Up and Go test, range of motion, 6-minute walking test), mainly favoring telemedicine (n=2) or showing no differences between groups (n=7), while a few (n=4) favored controls. Composite outcomes including PROMs and objective measures were analyzed in 2 meta-analyses; these showed no differences between groups. Overall, reported heterogeneity ranged from 0 to 97 (median I^2 28.2, IQR 0-67.8)

Figure 4 represents, with bubble plots, the direction of results and methodological quality of reviews in subgroups for outcome categories and populations.

XSL•FO
RenderX

Figure 4. Bubble plot showing directions of effects and AMSTAR 2 (A Measurement Tool to Assess systematic Reviews 2) results by (A) outcomes and (B) type of population. This graphic provides information in three dimensions: (1) in the x-axis, the authors' conclusions are rated as "beneficial for intervention," "no effect," or "beneficial for control" (this is further described in the Data Extraction section); (2) in the y-axis, the quality assessment (AMSTAR 2) is shown; and (3) the bubble size is proportional to the number of participants included in each systematic review. PREM: patient-reported outcome measure; PROM: patient-reported experience measure.

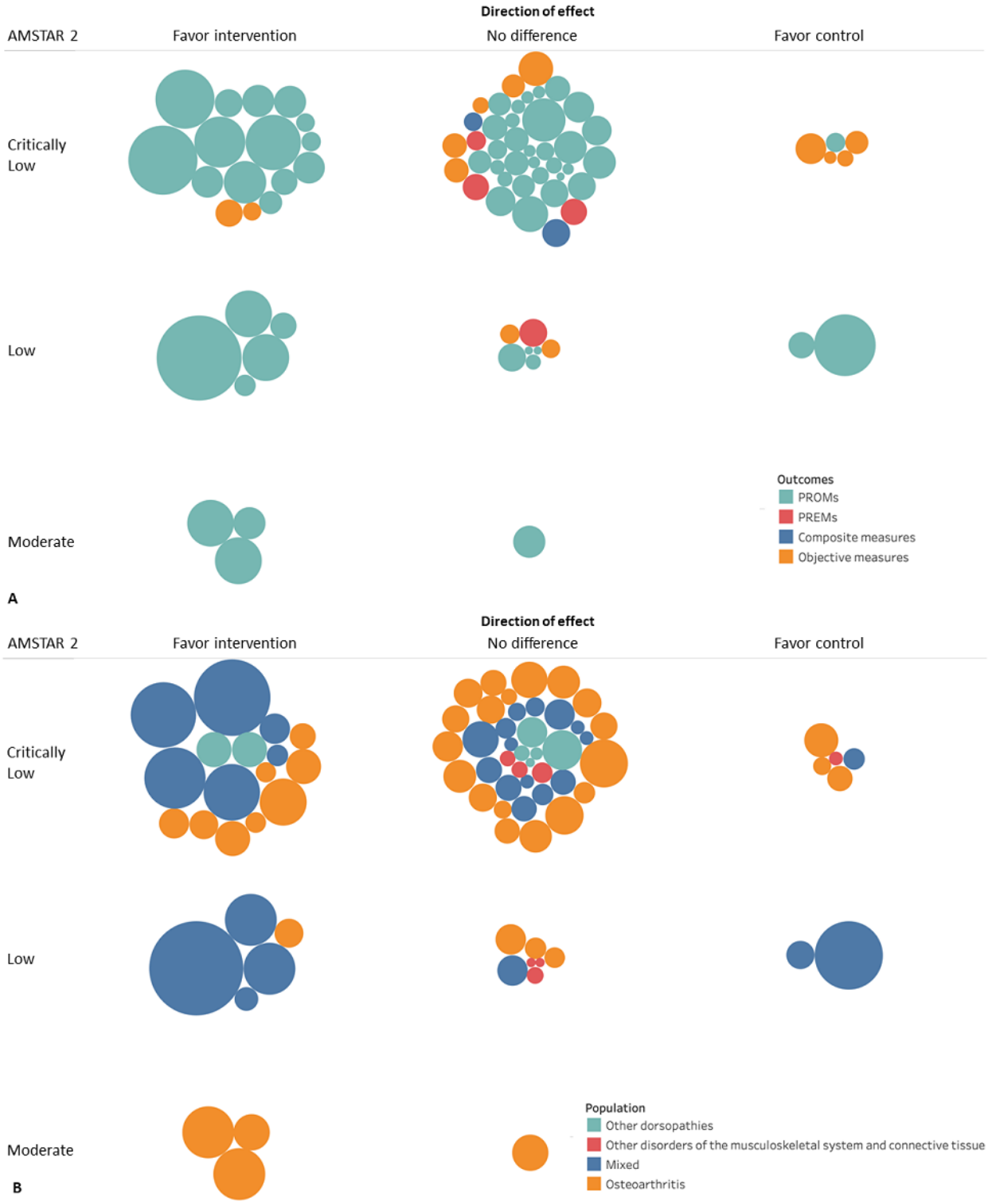

Summary of Qualitative Analyses

Twenty-four SRs reported qualitative results. Overall, 82 PROMs were described in the following areas: pain assessment (n=16), physical function (n=14), HRQoL (n=13), health literacy (n=9), cognitive function (n=8), side effects (n=8), adherence (n=7), emotional function (n=4), and social function (n=3). The SRs reported heterogeneous results in terms of physical function and HRQoL. Most of the SRs reported beneficial results in terms of health literacy and cognitive function in favor of telemedicine intervention, that is, the effects of the experimental intervention were superior or equal to control in all other

XSL•FO
RenderX

outcomes. Eight of the 12 PREMs included by the SRs were related to treatment and 4 to technology; 9 of the 21 objective outcomes were related to physical function, 4 to physical activity, and 8 to costs. Considering costs, all SRs showed that telemedicine cost significantly less than in-person visits or usual care. Multimedia Appendix 3 qualitatively describes these outcomes divided by domain of interest.

Discussion

Main Findings

This umbrella review summarizes the results from 35 SRs on telemedicine for musculoskeletal disorders, mainly published in the last 3 years (2020-2022) and predominantly conducted in Europe and North America. The type of telemedicine most often assessed was telerehabilitation for mixed chronic musculoskeletal pain and osteoarthritis. Overall, we retrieved many PROMs assessing pain, physical function, and HRQoL, whereas PREMs were less investigated.

According to our results, PROMs are more frequently analyzed than objective measures and PREMs, with most meta-analyses showing an improvement or similar effects compared to any other kind of intervention (eg, in-person treatment, usual care, or waiting list). The same effects were found in different subgroups of populations when visualized by direction of effects. However, some PROMs, such as side effects and adherence, were not quantitatively analyzed. It can be hypothesized that primary studies did not offer useful data to be pooled in a synthesis due to different taxonomies or poor outcome reporting, as well as the presence of outcome nonreporting bias [60,61].

Narrative syntheses on PREMs were reported in few SRs. Superior or equal effects were reported by telemedicine interventions in terms of acceptability of technology (usability, enjoyment, patient experience) and beneficial effects in terms of treatment (patient satisfaction and motivation). Costs were poorly investigated, but it seems that direct and indirect costs were reduced when telemedicine intervention is provided in comparison to in-person visits or usual care.

Overall, SR results should be interpreted with caution considering their methodological quality, which was generally critically low.

Comparison With Previous Overviews

Our findings are consistent with a previous overview [62] on the use of telemedicine across the 53 member states of the World Health Organization European Region, which showed clear benefits of telemedicine interventions in the screening, diagnosis, management, treatment, and long-term follow-up of many clinically and epidemiologically relevant diseases. Other overviews found that telemedicine delivered in different forms (eg, telerehabilitation, teleconsultation, and telemonitoring) has the potential to improve clinical outcomes in patients with cardiovascular disease [63] and chronic obstructive respiratory diseases [64]; patients who have survived cancer [65]; patients requiring neurorehabilitation [66] or urology care [67]; and patients with diabetes, dyslipidemia, and hypertension [68]. However, there are some areas of intervention that have not yet been covered by telemedicine and have uncertain effectiveness [68]. Nevertheless, in the musculoskeletal field, our umbrella review agrees with a previous study that supported the use of devices, tools, or software applications to facilitate remote rehabilitation and health care in general [69].

Clinical Implications

Clinicians and stakeholders should consider the adoption of the best available telemedicine technologies to treat patients' acute and chronic conditions, both in ordinary [70] and extraordinary situations [71]; evidence-based exercise and education [72] can be tailored and delivered remotely, for instance, to increase patients' compliance to treatment [28,73], reduce withdrawal rates from follow-up (so-called no-show patients), optimize workforce efficiency [74], and sometimes reduce health care costs [75].

Patient-centered care builds on the consideration of individual preference, easier access to treatment, and digital literacy enhancement. However, several barriers still exist in terms of ethical issues, privacy, accessibility, and data security. Telemedicine may help reach people living and working in rural and remote areas [76] with limited medical facilities and personnel. Indeed, patients living in rural areas can have poorer health outcomes in comparison to their urban counterparts [77], highlighting unequal health coverage [78]. An equal use of telemedicine technologies needs countries to invest in effective information policies and communication infrastructures [79]. One example is an Australian Commonwealth government program that aims to expand the Medical Benefit Schedule (MBS) by including telephone or online health consultations to reduce inequalities in favor of rural or remote patients.

Research Implications

Further efforts should be pursued to standardize collection of PROMs and PREMs in studies evaluating telemedicine. One significant challenge for certain musculoskeletal conditions is the lack of uniformity in outcome measurement across clinical trials; better standardization might help to identify and include PROMs and PREMs in core outcome sets to be measured and reported in all trials of a specific condition [80]. At present, for musculoskeletal disorders, there is still no clear consensus on PROMs as a core outcome set, even though some sets have been developed for Norway and the United Kingdom. Future research is needed to validate these in other countries [81,82].

Considering the lack of evidence on the use of PREMs in the evaluation of telemedicine technologies, they should be used more frequently and consistently before and after interventions. Few SRs reported data regarding costs, which is worth studying in more detail, taking into account the comparative difficulties related to different health care systems [83] and the different means of economic evaluation generally adopted [6,84].

Currently, research is ongoing to identify common core outcome domains from core outcome sets of musculoskeletal conditions [85].

Strengths and Limitations

To our knowledge, this is the first umbrella review encompassing any kind of telemedicine for different

XSL•FO
RenderX

musculoskeletal disorders, including multiple clinical outcomes and costs.

However, some limitations should be mentioned. Only 2 databases were explored and some relevant SRs from other sources (eg, gray literature) may have been missed. Inclusion criteria were focused on SRs of RCTs only. This criterion might have influenced the ratio of PROMs to PREMs, as this may depend on the study type and evidence level [8]. In fact, a previous study [8] investigating the use of PROMs and PREMs in any population with a telemedicine prescription found that the frequency of PREMs decreased with an increasing evidence level (ie, RCTs). Nevertheless, RCTs and reports with the highest quality of evidence should also include information on the usability and acceptance of the technology, expressed as PREMs, in addition to PROMs.

Moreover, we collected and interpreted quantitative effects in the SRs only at the shortest follow-up, limiting generalizability to all time end points. As well, it is possible that the overall positive effects might be biased by the wide range of heterogeneity within the included meta-analyses.

Finally, published taxonomies were used to standardize populations, interventions, and outcomes; however, the categorization of interventions was made difficult by the heterogeneous definition of telemedicine given by the SRs (ie, some used hybrid telemedicine or telemedicine mixed with conventional care, some did not; some used synchronous telemedicine, some did not). Indeed, *telemedicine* is considered an umbrella term for all health care services [1,2].

Conclusion

Telemedicine for musculoskeletal conditions can provide more accessible health care, with noninferior results in multiple clinical outcomes and no increase in side effects in comparison with more conventional care. The assessment of telemedicine is largely represented by PROMs, reflecting the relevance of patient-centered care. From a cost-effectiveness point of view, future studies should put effort into investigating PREMs, objective measures, and costs, filling the gaps in this promising area.

Acknowledgments

We thank our MSc (master of science) students, Gaia Ravot and Paolo Caneparo, for supporting us in the assessment of the methodological quality of the systematic reviews. This study was funded by the Italian Ministry of Health (Ricerca Corrente L3054), which had no role in the study design, data collection, analysis, interpretation, or report writing. The article processing charge was funded by the Italian Ministry of Health (Ricerca Corrente). This research did not receive a specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data Availability

The data sets generated during and/or analyzed during this study are available in the Open Science Framework repository [86].

Authors' Contributions

S Gianola and SB conceptualized and designed the study, drafted the initial manuscript, and coordinated and supervised the data collection. FP conceptualized the study and selected and collected data. GC performed the statistical analysis and contributed to results interpretation. S Guida selected and collected data and reviewed the manuscript. JAV and GB contributed to results interpretation and reviewed the manuscript. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy, taxonomy, and methodological quality assessment.
[DOCX File , 25 KB - jmir_v26i1e50090_app1.docx]

Multimedia Appendix 2
List of included and excluded studies, general characteristics of included systematic reviews, and effect sizes.
[DOCX File , 79 KB - jmir_v26i1e50090_app2.docx]

Multimedia Appendix 3
Narrative synthesis of patient-reported outcomes and experiences.
[DOCX File , 35 KB - jmir_v26i1e50090_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

PRISMA checklist.
[PDF File (Adobe PDF File), 72 KB - jmir_v26i1e50090_app4.pdf]

References

1. Cottrell MA, Russell TG. Telehealth for musculoskeletal physiotherapy. Musculoskelet Sci Pract 2020 Aug;48:102193 [FREE Full text] [doi: 10.1016/j.msksp.2020.102193] [Medline: 32560876]
2. Russell TG. Physical rehabilitation using telemedicine. J Telemed Telecare 2007;13(5):217-220. [doi: 10.1258/135763307781458886] [Medline: 17697506]
3. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
4. Bordas-Martinez J, Matéu Gómez L, Cámara Menoyo D, López-Sánchez M, Santos S, Molina-Molina M, et al. Patient-reported outcomes measures (PROMs) and patient-reported experience measures (PREMs) of COVID-19 telerehabilitation: Prospective pilot program. Medicine (Baltimore) 2022 Aug 05;101(31):e29639 [FREE Full text] [doi: 10.1097/MD.0000000000029639] [Medline: 35945781]
5. Foni NO, Costa LAV, Velloso LMR, Pedrotti CHS. Telemedicine: Is it a tool for orthopedics? Curr Rev Musculoskelet Med 2020 Dec;13(6):797-801 [FREE Full text] [doi: 10.1007/s12178-020-09680-6] [Medline: 33119841]
6. Vanni F, Foglia E, Pennestrì F, Ferrario L, Banfi G. Introducing enhanced recovery after surgery in a high-volume orthopaedic hospital: a health technology assessment. BMC Health Serv Res 2020 Aug 24;20(1):773 [FREE Full text] [doi: 10.1186/s12913-020-05634-3] [Medline: 32829712]
7. Rogante M, Grigioni M, Cordella D, Giacomozzi C. Ten years of telerehabilitation: A literature overview of technologies and clinical applications. NeuroRehabilitation 2010;27(4):287-304. [doi: 10.3233/NRE-2010-0612] [Medline: 21160118]
8. Knapp A, Harst L, Hager S, Schmitt J, Scheibe M. Use of patient-reported outcome measures and patient-reported experience measures within evaluation studies of telemedicine applications: systematic review. J Med Internet Res 2021 Nov 17;23(11):e30042 [FREE Full text] [doi: 10.2196/30042] [Medline: 34523604]
9. Jamieson Gilmore K, Corazza I, Coletta L, Allin S. The uses of patient reported experience measures in health systems: a systematic narrative review. Health Policy 2023 Feb;128:1-10 [FREE Full text] [doi: 10.1016/j.healthpol.2022.07.008] [Medline: 35934546]
10. De Rosis S, Cerasuolo D, Nuti S. Using patient-reported measures to drive change in healthcare: the experience of the digital, continuous and systematic PREMs observatory in Italy. BMC Health Serv Res 2020 Apr 16;20(1):315 [FREE Full text] [doi: 10.1186/s12913-020-05099-4] [Medline: 32299440]
11. Corazza I, Gilmore KJ, Menegazzo F, Abols V. Benchmarking experience to improve paediatric healthcare: listening to the voices of families from two European children's university hospitals. BMC Health Serv Res 2021 Jan 27;21(1):93 [FREE Full text] [doi: 10.1186/s12913-021-06094-z] [Medline: 33504331]
12. Pennestrì F, Banfi G. The experience of patients in chronic care management: applications in health technology assessment (HTA) and value for public health. Int J Environ Res Public Health 2022 Aug 10;19(16):9868 [FREE Full text] [doi: 10.3390/ijerph19169868] [Medline: 36011499]
13. Jiang S, Xiang J, Gao X, Guo K, Liu B. The comparison of telerehabilitation and face-to-face rehabilitation after total knee arthroplasty: a systematic review and meta-analysis. J Telemed Telecare 2018 May;24(4):257-262. [doi: 10.1177/1357633X16686748] [Medline: 28027679]
14. Du S, Liu W, Cai S, Hu Y, Dong J. The efficacy of e-health in the self-management of chronic low back pain: A meta analysis. Int J Nurs Stud 2020 Jun;106:103507. [doi: 10.1016/j.ijnurstu.2019.103507] [Medline: 32320936]
15. McDonnell J, Ahern D, Ross T, Gibbons D, Synnott K, Butler J. The efficacy of remote virtual care in comparison to traditional clinical visits for elective orthopaedic patients: A meta-analysis of prospective randomised controlled trials. Surgeon 2022 Jun;20(3):177-186. [doi: 10.1016/j.surge.2021.02.008] [Medline: 33762159]
16. Belbasis L, Bellou V, Ioannidis JPA. Conducting umbrella reviews. BMJ Med 2022;1(1):e000071 [FREE Full text] [doi: 10.1136/bmjmed-2021-000071] [Medline: 36936579]
17. Aromataris E, Fernandez R, Godfrey C, Holly C, Khalil H, Bhatarasakoon P. JBI Reviewer's Manual, Chapter 10: Umbrella Reviews. Joanna Briggs Institute. URL: https://jbi-global-wiki.refined.site/space/MANUAL/4685848/Previous+versions?attachment=/rest/api/content/4685848/child/attachment/att4698146/download&type=application/pdf&filename=JBI_Reviewers_Manual_2020March [accessed 2023-12-27]
18. Pollock M, Fernandes R, Becker L, Pieper D, Hartling L. Cochrane Handbook for Systematic Reviews of Interventions Handbook, Part 1, Chapter V: Overviews of Reviews. Wiley Online Library. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470712184.ch22 [accessed 2023-12-27]
19. Page M, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

20. Pollock M, Fernandes RM, Pieper D, Tricco AC, Gates M, Gates A, et al. Preferred Reporting Items for Overviews of Reviews (PRIOR): a protocol for development of a reporting guideline for overviews of reviews of healthcare interventions. Syst Rev 2019 Dec 23;8(1):335 [FREE Full text] [doi: 10.1186/s13643-019-1252-9] [Medline: 31870434]
21. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022 Aug 09;378:e070849 [FREE Full text] [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
22. About Cochrane reviews. Cochrane Library. URL: https://www.cochranelibrary.com/about/about-cochrane-reviews [accessed 2022-01-28]
23. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007 Feb 15;7:10 [FREE Full text] [doi: 10.1186/1471-2288-7-10] [Medline: 17302989]
24. Rethlefsen M, Kirtley S, Waffenschmidt S, Ayala A, Moher D, Page M, PRISMA-S Group. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. ICD-10-CM codes M00-M99: diseases of the musculoskeletal system and connective tissue. ICD10data. URL: https://www.icd10data.com/ICD10CM/Codes/M00-M99 [accessed 2023-12-27]
26. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
27. Miake-Lye IM, Mak S, Lee J, Luger T, Taylor SL, Shanman R, et al. Massage for pain: an evidence map. J Altern Complement Med 2019 May;25(5):475-502 [FREE Full text] [doi: 10.1089/acm.2018.0282] [Medline: 30892910]
28. Adamse C, Dekker-Van Weering MG, van Etten-Jamaludin FS, Stuiver MM. The effectiveness of exercise-based telemedicine on pain, physical activity and quality of life in the treatment of chronic pain: A systematic review. J Telemed Telecare 2018 Sep;24(8):511-526. [doi: 10.1177/1357633X17716576] [Medline: 28696152]
29. Agostini M, Moja L, Banzi R, Pistotti V, Tonin P, Venneri A, et al. Telerehabilitation and recovery of motor function: a systematic review and meta-analysis. J Telemed Telecare 2015 Jun;21(4):202-213 [FREE Full text] [doi: 10.1177/1357633X15572201] [Medline: 25712109]
30. Chaudhry H, Nadeem S, Mundi R. How satisfied are patients and surgeons with telemedicine in orthopaedic care during the COVID-19 pandemic? a systematic review and meta-analysis. Clin Orthop Relat Res 2021 Jan 01;479(1):47-56 [FREE Full text] [doi: 10.1097/CORR.0000000000001494] [Medline: 33009231]
31. Chen T, Or CK, Chen J. Effects of technology-supported exercise programs on the knee pain, physical function, and quality of life of individuals with knee osteoarthritis and/or chronic knee pain: A systematic review and meta-analysis of randomized controlled trials. J Am Med Inform Assoc 2021 Feb 15;28(2):414-423 [FREE Full text] [doi: 10.1093/jamia/ocaa282] [Medline: 33236109]
32. Cuenca-Martínez F, López-Bueno L, Suso-Martí L, Varangot-Reille C, Calatayud J, Herranz-Gómez A, et al. Implementation of online behavior modification techniques in the management of chronic musculoskeletal pain: a systematic review and meta-analysis. J Clin Med 2022 Mar 24;11(7):1806 [FREE Full text] [doi: 10.3390/jcm11071806] [Medline: 35407414]
33. Cuenca-Martínez F, Suso-Martí L, Herranz-Gómez A, Varangot-Reille C, Calatayud J, Romero-Palau M, et al. Effectiveness of telematic behavioral techniques to manage anxiety, stress and depressive symptoms in patients with chronic musculoskeletal pain: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Mar 09;19(6):3231 [FREE Full text] [doi: 10.3390/ijerph19063231] [Medline: 35328917]
34. Cuevas-Lara C, Izquierdo M, Sáez de Asteasu ML, Ramírez-Vélez R, Zambom-Ferraresi F, Zambom-Ferraresi F, et al. Impact of game-based interventions on health-related outcomes in hospitalized older patients: a systematic review. J Am Med Dir Assoc 2021 Feb;22(2):364-371.e1. [doi: 10.1016/j.jamda.2020.07.027] [Medline: 32873472]
35. Dias JF, Oliveira VC, Borges PRT, Dutra FCMS, Mancini MC, Kirkwood RN, et al. Effectiveness of exercises by telerehabilitation on pain, physical function and quality of life in people with physical disabilities: a systematic review of randomised controlled trials with GRADE recommendations. Br J Sports Med 2021 Feb;55(3):155-162. [doi: 10.1136/bjsports-2019-101375] [Medline: 33060156]
36. Esfandiari E, Miller WC, Berardi A, King S, Ashe MC. Telehealth interventions for mobility after lower limb loss: a systematic review and meta-analysis of randomized controlled trials. Prosthet Orthot Int 2022 Apr 01;46(2):108-120. [doi: 10.1097/PXR.0000000000000075] [Medline: 35412520]
37. Fritsch CG, Ferreira PH, Prior JL, McLachlan AJ, Ferreira ML. Effects of using text message interventions for the management of musculoskeletal pain: a systematic review. Pain 2020 Nov;161(11):2462-2475. [doi: 10.1097/j.pain.0000000000001958] [Medline: 32569088]
38. Gava V, Ribeiro LP, Barreto RPG, Camargo PR. Effectiveness of physical therapy given by telerehabilitation on pain and disability of individuals with shoulder pain: a systematic review. Clin Rehabil 2022 Jun;36(6):715-725. [doi: 10.1177/02692155221083496] [Medline: 35230167]

XSL•FO
RenderX

39. Gazendam A, Zhu M, Chang Y, Phillips S, Bhandari M. Virtual reality rehabilitation following total knee arthroplasty: a systematic review and meta-analysis of randomized controlled trials. Knee Surg Sports Traumatol Arthrosc 2022 Aug;30(8):2548-2555 [FREE Full text] [doi: 10.1007/s00167-022-06910-x] [Medline: 35182172]
40. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020 Jun 05;22(6):e15617 [FREE Full text] [doi: 10.2196/15617] [Medline: 32501277]
41. Jansson MM, Hyvämäki P, Pikkarainen M. Computer- and telephone-delivered interventions on patient outcomes and resource utilization in patients with orthopaedic conditions: a systematic review and narrative synthesis. Orthop Nurs 2020;39(5):340-352. [doi: 10.1097/NOR.0000000000000699] [Medline: 32956277]
42. Jansson MM, Rantala A, Miettunen J, Puhto AP, Pikkarainen M. The effects and safety of telerehabilitation in patients with lower-limb joint replacement: a systematic review and narrative synthesis. J Telemed Telecare 2022 Feb;28(2):96-114. [doi: 10.1177/1357633X20917868] [Medline: 32316837]
43. Lara-Palomo IC, Gil-Martínez E, Ramírez-García JD, Capel-Alcaraz AM, García-López H, Castro-Sánchez AM, et al. Efficacy of e-health interventions in patients with chronic low-back pain: a systematic review with meta-analysis. Telemed J E Health 2022 Dec;28(12):1734-1752. [doi: 10.1089/tmj.2021.0599] [Medline: 35532971]
44. Latif-Zade T, Tucci B, Verbovetskaya D, Bialkin E, Ng B, Heddon S, et al. Systematic review shows tele-rehabilitation might achieve comparable results to office-based rehabilitation for decreasing pain in patients with knee osteoarthritis. Medicina (Kaunas) 2021 Jul 28;57(8):764 [FREE Full text] [doi: 10.3390/medicina57080764] [Medline: 34440970]
45. Lin H, Han K, Ruan B. Effect of virtual reality on functional ankle instability rehabilitation: a systematic review. J Healthc Eng 2021;2021:7363403 [FREE Full text] [doi: 10.1155/2021/7363403] [Medline: 34880979]
46. Master H, Bley JA, Coronado RA, Robinette PE, White DK, Pennings JS, et al. Effects of physical activity interventions using wearables to improve objectively-measured and patient-reported outcomes in adults following orthopaedic surgical procedures: A systematic review. PLoS One 2022;17(2):e0263562 [FREE Full text] [doi: 10.1371/journal.pone.0263562] [Medline: 35167599]
47. McHugh CG, Kostic AM, Katz JN, Losina E. Effectiveness of remote exercise programs in reducing pain for patients with knee osteoarthritis: A systematic review of randomized trials. Osteoarthr Cartil Open 2022 Sep;4(3):100264 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100264] [Medline: 36474946]
48. Nicholl BI, Sandal LF, Stochkendahl MJ, McCallum M, Suresh N, Vasseljen O, et al. Digital support interventions for the self-management of low back pain: a systematic review. J Med Internet Res 2017 May 21;19(5):e179 [FREE Full text] [doi: 10.2196/jmir.7290] [Medline: 28550009]
49. Oliveira CB, Franco MR, Maher CG, Ferreira PH, Morelhão PK, Damato TM, et al. Physical activity-based interventions using electronic feedback may be ineffective in reducing pain and disability in patients with chronic musculoskeletal pain: a systematic review with meta-analysis. Arch Phys Med Rehabil 2018 Sep;99(9):1900-1912. [doi: 10.1016/j.apmr.2017.10.013] [Medline: 29122581]
50. Petersen W, Karpinski K, Backhaus L, Bierke S, Häner M. A systematic review about telemedicine in orthopedics. Arch Orthop Trauma Surg 2021 Oct;141(10):1731-1739 [FREE Full text] [doi: 10.1007/s00402-021-03788-1] [Medline: 33635399]
51. Safari R, Jackson J, Sheffield D. Digital self-management interventions for people with osteoarthritis: systematic review with meta-analysis. J Med Internet Res 2020 Jul 20;22(7):e15365 [FREE Full text] [doi: 10.2196/15365] [Medline: 32706657]
52. Schäfer AGM, Zalpour C, von Piekartz H, Hall TM, Paelke V. The efficacy of electronic health-supported home exercise interventions for patients with osteoarthritis of the knee: systematic review. J Med Internet Res 2018 Apr 26;20(4):e152 [FREE Full text] [doi: 10.2196/jmir.9465] [Medline: 29699963]
53. Srikesavan C, Bryer C, Ali U, Williamson E. Web-based rehabilitation interventions for people with rheumatoid arthritis: a systematic review. J Telemed Telecare 2019 Jun;25(5):263-275. [doi: 10.1177/1357633X18768400] [Medline: 29669470]
54. Tabacof L, Baker TS, Durbin JR, Desai V, Zeng Q, Sahasrabudhe A, et al. Telehealth treatment for nonspecific low back pain: a review of the current state in mobile health. PM R 2022 Sep;14(9):1086-1098. [doi: 10.1002/pmrj.12738] [Medline: 34786870]
55. Tsang MP, Man GCW, Xin H, Chong YC, Ong MTY, Yung PSH. The effectiveness of telerehabilitation in patients after total knee replacement: a systematic review and meta-analysis of randomized controlled trials. J Telemed Telecare 2022 May 12:1357633X221097469. [doi: 10.1177/1357633X221097469] [Medline: 35549756]
56. Wang Q, Lee RLT, Hunter S, Chan SWC. The effectiveness of internet-based telerehabilitation among patients after total joint arthroplasty: a systematic review and meta-analysis of randomised controlled trials. J Telemed Telecare 2023 May;29(4):247-260. [doi: 10.1177/1357633X20980291] [Medline: 33459120]
57. Wang X, Hunter DJ, Vesentini G, Pozzobon D, Ferreira ML. Technology-assisted rehabilitation following total knee or hip replacement for people with osteoarthritis: a systematic review and meta-analysis. BMC Musculoskelet Disord 2019 Nov 03;20(1):506 [FREE Full text] [doi: 10.1186/s12891-019-2900-x] [Medline: 31679511]
58. Xie SH, Wang Q, Wang LQ, Wang L, Song KP, He CQ. Effect of internet-based rehabilitation programs on improvement of pain and physical function in patients with knee osteoarthritis: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Jan 05;23(1):e21542. [doi: 10.2196/21542] [Medline: 33399542]

XSL•FO
RenderX

59. Hussain A, Haroon H, Ahmed A, Gilani SA. Digital technologies in management of chronic pain - a systematic review. J Pak Med Assoc 2022 Jun;72(6):1158-1165 [FREE Full text] [doi: 10.47391/JPMA.3885] [Medline: 35751328]
60. Page MJ, Higgins JPT. Rethinking the assessment of risk of bias due to selective reporting: a cross-sectional study. Syst Rev 2016 Jul 08;5(1):108 [FREE Full text] [doi: 10.1186/s13643-016-0289-2] [Medline: 27392044]
61. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research - a narrative review. Trials 2010 Apr 13;11(1):37 [FREE Full text] [doi: 10.1186/1745-6215-11-37] [Medline: 20388211]
62. Saigí-Rubió F, Borges do Nascimento IJ, Robles N, Ivanovska K, Katz C, Azzopardi-Muscat N, et al. The current status of telemedicine technology use across the World Health Organization European region: an overview of systematic reviews. J Med Internet Res 2022 Oct 27;24(10):e40877 [FREE Full text] [doi: 10.2196/40877] [Medline: 36301602]
63. Arian M, Valinejadi A, Soleimani M. Quality of life in heart patients receiving telerehabilitation: an overview with meta-analyses. Iran J Public Health 2022 Nov;51(11):2388-2403 [FREE Full text] [doi: 10.18502/ijph.v51i11.11157] [Medline: 36561264]
64. Barbosa MT, Sousa CS, Morais-Almeida M. Telemedicine in the management of chronic obstructive respiratory diseases: an overview. Exon Publications. URL: https://www.ncbi.nlm.nih.gov/books/NBK580636/ [accessed 2023-12-27]
65. Chan R, Crichton M, Crawford-Williams F, Agbejule O, Yu K, Hart N, Multinational Association of Supportive Care in Cancer (MASCC) Survivorship Study Group. The efficacy, challenges, and facilitators of telemedicine in post-treatment cancer survivorship care: an overview of systematic reviews. Ann Oncol 2021 Dec;32(12):1552-1570 [FREE Full text] [doi: 10.1016/j.annonc.2021.09.001] [Medline: 34509615]
66. Matamala-Gomez M, Bottiroli S, Realdon O, Riva G, Galvagni L, Platz T, et al. Telemedicine and virtual reality at time of COVID-19 pandemic: an overview for future perspectives in neurorehabilitation. Front Neurol 2021;12:646902 [FREE Full text] [doi: 10.3389/fneur.2021.646902] [Medline: 33841313]
67. Ellimoottil C. Implementing telemedicine in urology: an overview of the benefits and barriers. J Urol 2019 Jul;202(1):47-48. [doi: 10.1097/JU.0000000000000285] [Medline: 30991010]
68. Timpel P, Oswald S, Schwarz PEH, Harst L. Mapping the evidence on the effectiveness of telemedicine interventions in diabetes, dyslipidemia, and hypertension: an umbrella review of systematic reviews and meta-analyses. J Med Internet Res 2020 Mar 18;22(3):e16791 [FREE Full text] [doi: 10.2196/16791] [Medline: 32186516]
69. Edwards D, Williams J, Carrier J, Davies J. Technologies used to facilitate remote rehabilitation of adults with deconditioning, musculoskeletal conditions, stroke, or traumatic brain injury: an umbrella review. JBI Evid Synth 2022 Aug 01;20(8):1927-1968. [doi: 10.11124/JBIES-21-00241] [Medline: 35971198]
70. Baroni MP, Jacob MFA, Rios WR, Fandim JV, Fernandes LG, Chaves PI, et al. The state of the art in telerehabilitation for musculoskeletal conditions. Arch Physiother 2023 Jan 04;13(1):1 [FREE Full text] [doi: 10.1186/s40945-022-00155-0] [Medline: 36597130]
71. Pennestrì F, Gaudioso A, Jani A, Bottinelli E, Banfi G. Is administered competition suitable for dealing with a public health emergency? Lessons from the local healthcare system at the centre of early COVID-19 outbreak in Italy. Cent Eur J Public Health 2021 Jun;29(2):109-116 [FREE Full text] [doi: 10.21101/cejph.a6533] [Medline: 34245550]
72. de Campos TF. Low back pain and sciatica in over 16s: assessment and management NICE Guideline [NG59]. J Physiother 2017 Apr;63(2):120 [FREE Full text] [doi: 10.1016/j.jphys.2017.02.012] [Medline: 28325480]
73. Morrison KS, Paterson C, Toohey K. The feasibility of exercise interventions delivered via telehealth for people affected by cancer: a rapid review of the literature. Semin Oncol Nurs 2020 Dec;36(6):151092. [doi: 10.1016/j.soncn.2020.151092] [Medline: 33223409]
74. Bhavsar NA, Doerfler SM, Giczewska A, Alhanti B, Lutz A, Thigpen CA, et al. Prevalence and predictors of no-shows to physical therapy for musculoskeletal conditions. PLoS One 2021;16(5):e0251336 [FREE Full text] [doi: 10.1371/journal.pone.0251336] [Medline: 34048440]
75. Snoswell CL, Taylor ML, Comans TA, Smith AC, Gray LC, Caffery LJ. Determining if telehealth can reduce health system costs: scoping review. J Med Internet Res 2020 Oct 19;22(10):e17298 [FREE Full text] [doi: 10.2196/17298] [Medline: 33074157]
76. Balestri R, Cavina E, Aliferis A, Goletti O, Rocci R, Lippolis PV, et al. Telemedicine on a small island. J Telemed Telecare 1999;5 Suppl 1(1_suppl):S50-S52. [doi: 10.1258/1357633991932531] [Medline: 10534840]
77. Wakerman J, Humphreys JS. Sustainable primary health care services in rural and remote areas: innovation and evidence. Aust J Rural Health 2011 Jun;19(3):118-124. [doi: 10.1111/j.1440-1584.2010.01180.x] [Medline: 21605224]
78. Bradford N, Caffery L, Smith A. Telehealth services in rural and remote Australia: a systematic review of models of care and factors influencing success and sustainability. Rural Remote Health 2016;16(4):4268 [FREE Full text] [Medline: 27817199]
79. Meso P, Mbarika V, Sood S. An overview of potential factors for effective telemedicine transfer to sub-Saharan Africa. IEEE Trans Inf Technol Biomed 2009 Sep;13(5):734-739. [doi: 10.1109/TITB.2007.899807] [Medline: 19273036]
80. Macefield RC, Jacobs M, Korfage IJ, Nicklin J, Whistance RN, Brookes ST, et al. Developing core outcomes sets: methods for identifying and including patient-reported outcomes (PROs). Trials 2014 Feb 05;15(1):49 [FREE Full text] [doi: 10.1186/1745-6215-15-49] [Medline: 24495582]

XSL•FO
RenderX

81. Burgess R, Lewis M, McRobert C, Hill J. Developing a core outcome set for community and primary care musculoskeletal services: A consensus approach. Musculoskelet Sci Pract 2021 Oct;55:102415. [doi: 10.1016/j.msksp.2021.102415] [Medline: 34171606]
82. Klokkerud M, Dagfinrud H, Uhlig T, Dager T, Furunes K, Klokkeide, et al. Developing and testing a consensus-based core set of outcome measures for rehabilitation in musculoskeletal diseases. Scand J Rheumatol 2018 May;47(3):225-234. [doi: 10.1080/03009742.2017.1347959] [Medline: 28988517]
83. Lim KK, Chan M, Navarra S, Haq SA, Lau CS. Development and implementation of models of care for musculoskeletal conditions in middle-income and low-income Asian countries. Best Pract Res Clin Rheumatol 2016 Jun;30(3):398-419. [doi: 10.1016/j.berh.2016.08.007] [Medline: 27886939]
84. Pennestrì F, Lega F, Banfi G. From volume to value: Improving peri-operative elective pathways through a roadmap from fast-track orthopedic surgery. Health Serv Manage Res 2023 Nov;36(4):284-290 [FREE Full text] [doi: 10.1177/09514848221127623] [Medline: 36444939]
85. Sabet TS, Anderson DB, Stubbs PW, Buchbinder R, Terwee CB, Chiarotto A, et al. Identifying common core outcome domains from core outcome sets of musculoskeletal conditions: protocol for a systematic review. Syst Rev 2022 Nov 19;11(1):248 [FREE Full text] [doi: 10.1186/s13643-022-02120-1] [Medline: 36403060]
86. Efficacy of telemedicine for musculoskeletal disorders: an umbrella review. OSFHome. URL: https://osf.io/pxedm [accessed 2023-12-27]

Abbreviations

AMSTAR 2: A Measurement Tool to Assess systematic Reviews 2
HRQoL: health-related quality of life
ICD-10: International Classification of Diseases, 10th revision
MD: mean difference SMD: standardized mean difference
OR: odds ratio
PICOS: population, intervention, comparison, outcomes, study design
PREM: patient-reported experience measure
PRIOR: Preferred Reporting Items for Overviews of Reviews
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROM: patient-reported outcome measure
PROSPERO: Prospective Register of Systematic Reviews
RCT: randomized controlled trial
RR: relative risk
SR: systematic review

Edited by T de Azevedo Cardoso; submitted 19.06.23; peer-reviewed by L Harst, N Shah; comments to author 13.10.23; revised version received 02.11.23; accepted 29.11.23; published 02.02.24.

Please cite as:
Bargeri S, Castellini G, Vitale JA, Guida S, Banfi G, Gianola S, Pennestrì F
Effectiveness of Telemedicine for Musculoskeletal Disorders: Umbrella Review
J Med Internet Res 2024;26:e50090
URL: https://www.jmir.org/2024/1/e50090
doi:10.2196/50090
PMID:38306156

©Silvia Bargeri, Greta Castellini, Jacopo Antonino Vitale, Stefania Guida, Giuseppe Banfi, Silvia Gianola, Federico Pennestrì. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Timing, Indicators, and Approaches to Digital Patient Experience Evaluation: Umbrella Systematic Review

Tingting Wang[1], MA; Guido Giunti[2,3], PhD; Richard Goossens[1], PhD; Marijke Melles[1], PhD

[1]Department of Human-Centered Design, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands

[2]Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

[3]Clinical Medicine Neurology, School of Medicine, Trinity College Dublin, Dublin, Ireland

Corresponding Author:
Tingting Wang, MA
Department of Human-Centered Design
Faculty of Industrial Design Engineering
Delft University of Technology
Gebouw 32
Landbergstraat 15
Delft, 2628 CE
Netherlands
Phone: 31 623018218
Email: t.wang-8@tudelft.nl

Abstract

Background: The increasing prevalence of DH applications has outpaced research and practice in digital health (DH) evaluations. Patient experience (PEx) was reported as one of the challenges facing the health system by the World Health Organization. To generate evidence on DH and promote the appropriate integration and use of technologies, a standard evaluation of PEx in DH is required.

Objective: This study aims to systematically identify evaluation timing considerations (ie, *when to measure*), evaluation indicators (ie, *what to measure*), and evaluation approaches (ie, *how to measure*) with regard to digital PEx. The overall aim of this study is to generate an evaluation guide for further improving digital PEx evaluation.

Methods: This is a 2-phase study parallel to our previous study. In phase 1, literature reviews related to PEx in DH were systematically searched from Scopus, PubMed, and Web of Science databases. Two independent raters conducted 2 rounds of paper screening, including title and abstract screening and full-text screening, and assessed the interrater reliability for 20% (round 1: 23/115 and round 2: 12/58) random samples using the Fleiss-Cohen coefficient (round 1: $k1=0.88$ and round 2: $k2=0.80$). When reaching interrater reliability ($k>0.60$), TW conducted the rest of the screening process, leaving any uncertainties for group discussions. Overall, 38% (45/119) of the articles were considered eligible for further thematic analysis. In phase 2, to check if there were any meaningful novel insights that would change our conclusions, we performed an updated literature search in which we collected 294 newly published reviews, of which 102 (34.7%) were identified as eligible articles. We considered them to have no important changes to our original results on the research objectives. Therefore, they were not integrated into the synthesis of this review and were used as supplementary materials.

Results: Our review highlights 5 typical evaluation objectives that serve 5 stakeholder groups separately. We identified a set of key evaluation timing considerations and classified them into 3 categories: intervention maturity stages, timing of the evaluation, and timing of data collection. Information on evaluation indicators of digital PEx was identified and summarized into 3 categories (intervention outputs, patient outcomes, and health care system impact), 9 themes, and 22 subthemes. A set of evaluation theories, common study designs, data collection methods and instruments, and data analysis approaches was captured, which can be used or adapted to evaluate digital PEx.

Conclusions: Our findings enabled us to generate an evaluation guide to help DH intervention researchers, designers, developers, and program evaluators evaluate digital PEx. Finally, we propose 6 directions for encouraging further digital PEx evaluation research and practice to address the challenge of poor PEx.

(J Med Internet Res 2024;26:e46308) doi:10.2196/46308

XSL•FO
RenderX

KEYWORDS

digital health; eHealth; telemedicine; mobile health; mHealth; patient experience; user experience; evaluation timing; evaluation indicators; evaluation approaches; user-centered design; patient-centered care; human-computer interaction; mobile phone

Introduction

Background

Emerging digital technologies promise to shape the future health care industry [1,2]. According to our previous review [3], most researchers had a positive impression of digital health interventions (DHIs). The number of DHIs is proliferating [4-6], which is affecting the way patients receive their health care services compared with face-to-face health care services and ultimately influencing the patient journey and overall patient experience (PEx) [7,8]. Good PEx is a key intent of patient-centered care [9] and a core measure of care quality in digital health (DH) [10,11]. Digital technologies have the potential to enhance or provide comparable PEx compared with some face-to-face health care services [8,12-14]. However, the uptake of digital technologies in health care is not as rapid as it has been in many other industries [15], and their potential in health care remains unfulfilled [16]. According to a report by the World Health Organization (WHO) on the classification of DHIs, the health system is not responding adequately to the need for improved PEx [17].

Despite the growing number of DHIs, evaluations that are timely, cost-effective, and robust have not kept pace with this growth [7,18,19]. PExs in the wide range of DHIs are mixed [20,21]. Few published DHIs have resulted in high download numbers and active users [22]; most are released with minimal or no evaluation and require patients to assess the quality for themselves and take responsibility for any consequences [23]. Low-quality DH may disrupt user experience (UX) [24], resulting in low acceptance, and some may even be harmful [25]. In addition, a DHI may be popular with patients but not valued by clinicians [26]. To generate evidence and promote the appropriate integration and use of digital technologies in health care, an overview of how to evaluate PEx or UX in varied DHIs is needed [3,27].

Evaluating the Digital PEx

In this study, we used the definition of digital PEx from our previous review [3]: “the sum of all interactions affected by a patient’s behavioral determinants, framed by digital technologies, and shaped by organizational culture, that influence patient perceptions across the continuum of care channeling digital health.” This incorporates influencing factors of digital PEx [3] and the existing definitions of DHIs [28,29], PEx [30], and UX [31]. Compared with the general PEx and UX, it highlights patient perceptions that are affected by technical, behavioral, and organizational determinants when interacting with a DHI. DHI has become an umbrella term that often encompasses broad concepts and technologies [32], such as DH applications, ecosystems, and platforms [28]. In this study, we followed the WHO’s definition of DHIs [29], that is, the use of digital, mobile, and wireless technologies to support the achievement of health objectives. It refers to the use of information and communication technologies for health care, encompassing both mobile health and eHealth [29,33]. Compared with evaluating DHIs, PEx, and UX, little is known about evaluating digital PEx. However, combining the definition of digital PEx with the extensively explored measurement of PEx, UX, and DHIs can lead to an improved understanding of and enable the development of evaluation approaches for measuring digital PEx. Therefore, the evaluations of PEx, UX, and DHIs will be used as a starting point in this study to clarify when to measure, what to measure, and how to measure digital PEx.

When to Measure

First, the timing of measuring and evaluating digital PEx is an important consideration and must align with the contextual situation, such as evaluation objectives and stakeholders, to ensure practicality and purposefulness [34,35]. According to the European Union [36] and the Department of Health of The King’s Fund [37], an evaluation can be scheduled during the design phase or during or after the implementation phase. Similarly, the WHO [29] introduced 3 DHI evaluation stages: efficacy, effectiveness, and implementation. The evaluation of efficacy refers to where the intervention is under highly controlled conditions, the evaluation of effectiveness is carried out in a real world context, and the evaluation of implementation occurs after efficacy and effectiveness have been established. Furthermore, an evaluation can be performed before, during, or after the evaluated intervention in both research and nonresearch settings [36]. However, decision-making on *when* to collect PEx data can be more complicated. As argued in earlier studies [35,37], immediate feedback has the benefit of gaining real-time insights, but patients may be too unwell, stressed, or distracted to provide detailed opinions. In contrast, when the feedback is related to medical outcomes or quality of life, it often requires a lengthy period after the intervention to observe any changes. However, responses gathered long after a care episode may be inferior because of recall bias.

What to Measure

Second, there is a need for a decision on *what* is required to measure to assess digital PEx. The frequently mentioned UX evaluation concepts, such as usability, functionality, and reliability, from studies [38-40] investigating UX can be applied to evaluate the intervention outputs to anticipate digital PEx at a service level. Moreover, according to the existing constructs and frameworks of understanding or evaluating PEx [41-45], such as emotional support, relieving fear and anxiety, patients as active participants in care, and continuity of care and relationships, they can be adjusted to evaluate digital PEx by understanding patient outcomes at an individual level. In addition, the National Quality Forum [11] proposed a set of measurable concepts to be used to evaluate PEx in telehealth, for example, patients’ increased confidence in, understanding of, and compliance with their care plan; reduction in diagnostic errors and avoidance of adverse outcomes; and decrease in waiting times and eliminated travel. Some of these concepts

XSL•FO
RenderX

can be used to understand digital PEx at an organizational level by assessing the impact of the health care system.

How to Measure

The third consideration is *how* to choose evaluation approaches appropriate for evaluating the digital PEx [35], starting from widely used theories, study designs, methods, and tools for evaluating DHIs and the related PEx or UX. There is rapidly evolving guidance for guiding DH innovators [18], such as the National Institute for Health and Care Excellence Evidence Standards Framework for Digital Health Technologies [46]. The strength of the evidence in the evaluation of DHIs often depends on the study design [18]. However, the high bar for evidence in health care usually requires a longer time for evidence generation, such as prospective randomized controlled trials (RCTs) and observational studies, which often conflicts with the fast-innovation reality of the technology industry [16,18]. In addition, many traditional approaches, such as qualitative and quantitative methods, can be used to collect experience-related data to evaluate the DHIs [18,29]. Qualitative methods such as focus groups, interviews, and observations are often used to obtain an in-depth understanding of PEx [37] in the early intervention development stages [29]. Surveys using structured questionnaires, such as patient satisfaction ratings [37,47], patient-reported experience measures (PREMs) [35,48], and patient-reported outcome measures (PROMs) [35,37,48], are often used to examine patterns and trends from a large sample. Hodgson [49] believed that strong evidence results from UX data that are valid and reliable, such as formative and summative usability tests, and stated that behavioral data are strong, but opinion data are weak.

Objectives

This study aims to systematically identify (1) evaluation timing considerations (ie, *when to measure*), (2) evaluation indicators (ie, *what to measure*), and (3) evaluation approaches (ie, *how to measure*) with regard to digital PEx. The overall aim of this study is to generate an evaluation guide for further improving digital PEx evaluation research and practice.

Methods

Overview

This study consists of 2 phases. In phase 1, we followed the same study search and selection process as our previous research [3] but focused on a different data extraction and analysis process to achieve our objectives in this study. In the previous study [3], we identified the influencing factors and design considerations of digital PEx, provided a definition, constructed a design and evaluation framework, and generated 9 design guidelines to help DH designers and developers improve digital PEx. To highlight the connections between “design” and “evaluation” works in the development of DH and provide readers with a clear road map, we included some evaluation-related information in the previous paper as well. However, it was limited and described at a very abstract level. In this study, detailed information on the evaluation was provided, including evaluation timing considerations, evaluation indicators, and evaluation approaches, and we aimed to generate an evaluation guide for improving the measurement of digital PEx. Given that this is an evolving area, after we finished phase 1, we conducted an updated literature search as a subsequent investigation to determine whether an update of a review was needed in this study.

Phase 1: The Original Review

Study Search and Selection

Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [50], we conducted an umbrella systematic review [51] on literature reviews related to PEx and UX in DH. The term DH was first introduced in 2000 by Frank [52]. Therefore, Scopus, PubMed, and Web of Science databases were used for searching related articles that were published between January 1, 2000, and December 16, 2020. Furthermore, Google Scholar was used to search for additional studies that were identified during the review process through the snowballing method. The computer search resulted in 173 articles, of which 58 (33.5%) were duplicates. After removing the duplicates, the titles and abstracts of a small random sampling (23/115, 20%) were reviewed by 2 independent raters to assess the interrater reliability by using the Fleiss-Cohen coefficient, which resulted in $k1$=0.88 (SE 0.07; 95% CI 0.74-1.03). This was followed by a group discussion to reach an agreement on the selection criteria. Subsequently, the remaining titles and abstracts (92/115, 80%) were reviewed by TW individually. After screening the titles and abstracts, half of the articles (58/115, 50.4%) remained for the full-text review. Meanwhile, 4 additional articles were identified through snowballing and were included in the full-text screening. Another small random sample (12/62, 19%) was reviewed by the 2 raters to screen the full texts. After achieving interrater reliability, $k2$=0.80 (SE 0.13; 95% CI 0.54-1.05) and reaching a consensus on the inclusion criteria through another group discussion, TW reviewed the full texts of the remaining papers (50/62, 80%). Google Sheets was used for performing the screening process and assessments. Finally, as shown in Figure 1 [3], a total of 45 articles were included for data extraction. A detailed search strategy, selection criteria, and screening process can be found in our previously published study [3]. Multimedia Appendix 1 [53-97] presents the included and excluded articles.

XSL•FO
RenderX

Figure 1. Study flow diagram. ICT: information and communication technology.

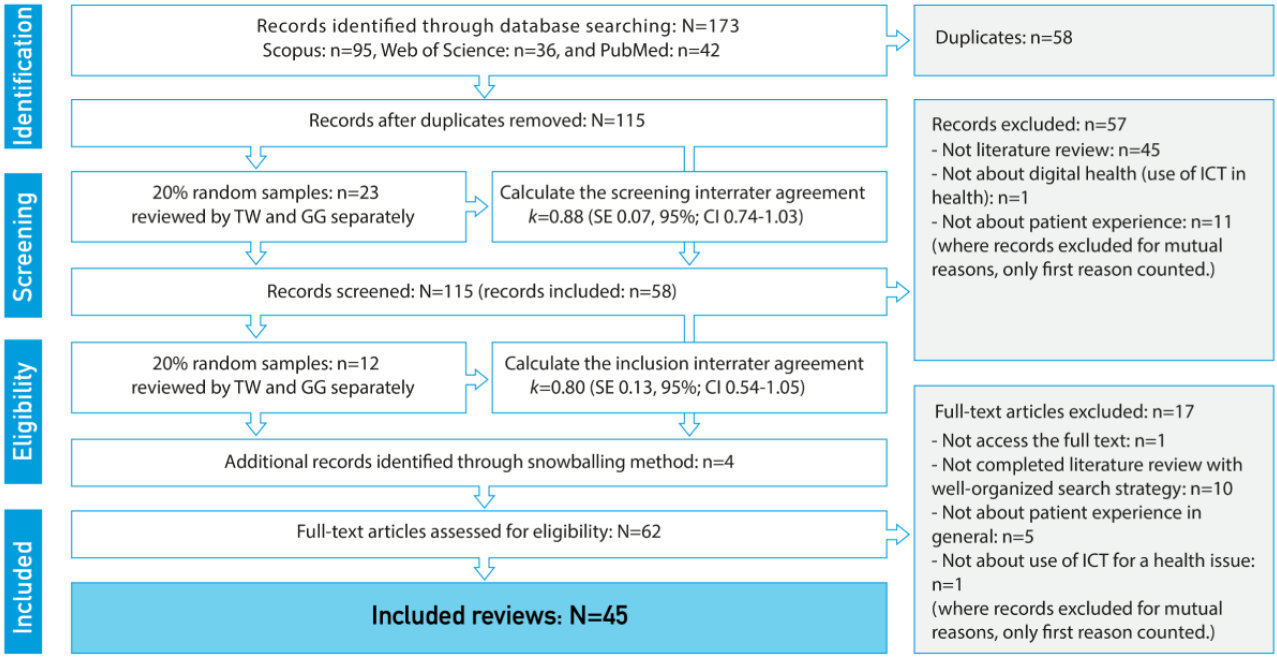

Data Extraction and Thematic Analysis

We used ATLAS.ti (Scientific Software Development GmbH; version 9.0.7) for data extraction. Data were extracted for the three predefined objectives: (1) evaluation timing considerations, (2) evaluation indicators, and (3) evaluation approaches of the digital PEx. In addition, we collected data related to evaluation objectives among the included studies. Data analysis followed the 6-phase thematic analysis method proposed by Braun and Clarke [98,99]: familiarization, coding, generating themes, reviewing themes, defining and naming themes, and writing up. First, we became familiar with the 45 articles included in the study. Second, after a thorough review, TW started iteratively coding the data related to the predefined objectives based on existing frameworks, including the Performance of Routine Information System Management framework [100], monitoring and evaluation guide [29], measures of PEx in hospitals [37], and an overview of research methodology [101]. This resulted in 25 initial codes. After no additional new codes were identified, TW proposed a coding scheme to summarize the recurring points throughout the data. Then, GG, RG, and MM reviewed and discussed the coding scheme until they reached an agreement. Third, TW followed the coding scheme to code the data more precisely and completely and searched for themes among the generated codes. Fourth, TW, GG, RG, and MM reviewed and discussed these codes and themes to address any uncertainties. Fifth, the definitions and names of the generated themes were adjusted through team discussions. Finally, the analytical themes related to the evaluation timing, indicators, and approaches were produced and reported. Both deductive and inductive approaches [99] were used to identify and generate themes. Four researchers were involved in the review process.

We first highlighted the evaluation timing considerations in terms of *intervention maturity stages*, the *timing of evaluation*, and the *timing of datacollection*, which were adopted from the description of the WHO and European Union (Table 1) [36,37].

We then determined the evaluation indicators and classified them into 3 categories (Table 2). *Intervention outputs* are the direct products or deliverables of process activities and refer to the different stages of evaluation that correspond to the various stages of maturity of the DHI. *Patient outcomes* describe the intermediate changes in patients, including patients' emotions, perceptions, capabilities, behaviors, and health conditions as determined by DHIs in terms of influencing factors and interaction processes. *Health care system impact* is the medium- to long-term, large-scale financial (intended and unintended) effects produced by a DHI.

Finally, we concluded evaluation approaches in terms of study designs, data collection methods and instruments, and data analysis approaches (Table 3). According to the WHO [29], study designs are intended to assist in decision-making on evidence generation and clarify the scope of evaluation activities. Data collection and analysis are designed through an iterative process that involves strategies for collecting and analyzing data and a series of specifically designed tools [36].

XSL•FO
RenderX

Table 1. Initial codes of evaluation timing considerations of the digital patient experience.

Categories and initial codes	Description
Intervention maturity stages [29,36,37]	
Efficacy	Assess whether the DHI[a] achieves the intended results in research or controlled setting
Effectiveness	Assess whether the DHI achieves the intended results in nonresearch or uncontrolled setting
Implementation	Assess the uptake, institutionalization, and sustainability of evidence-based DHIs in a given context, including policies and practices
Timing of the evaluation [36]	
Before intervention	A baseline test is performed before individuals adopt or implement the intervention. It assesses individuals' initial status and their anticipated perception of the intervention
During intervention	An evaluation performed during intervention's use aims to monitor individuals' real-time feedback and reactions
After intervention	An evaluation that is performed right after or a long time after the completion of the interventions by individuals. It assesses individuals' changes regarding using the intervention
Timing of data collection [35,37]	
Immediate evaluation	Aims to collect real-time data on patients' experiences during or immediately after their treatment
Delayed evaluation	Aims to obtain more substantial responses after the intervention's completion over a long period
Momentary evaluation	Aims to collect transient information from individuals at a specific moment
Continuous evaluation	Aims to gather feedback from individuals at different points along the care pathway

[a]DHI: digital health intervention.

Table 2. Initial codes of evaluation indicators of the digital patient experience.

Categories and initial codes	Description
Intervention outputs [29,38-40,102]	
Functionality	Assess whether the DHI[a] works as intended. It refers to the ability of the DH[b] system to support the desired intervention.
Usability	Assess whether the DHI is used as intended. It refers to the degree to which the intervention is understandable and easy to use.
Quality of care	Assess whether the DHI delivers effective, safe, people-centered, timely, accessible, equitable, integrated, and efficient care services. It refers to the degree to which health services for individuals and populations increase the likelihood of desired health outcomes.
Patient outcomes [11,41-45]	
Emotional outcomes	Assess whether patients' feelings and well-being change positively or negatively because of the use or anticipated use of DHIs. It refers to what the patients feels.
Perceptual outcomes	Assess whether the informed state of mind that patients achieve as intended before, during, or after using the DHIs. It refers to what the patient thinks and believes.
Capability outcomes	Assess whether patients' health literacy, communication skills, or computer confidence in managing diseases, communicating with health care providers, or operating digital devices increased as expected. It refers what the patient knows and acquires.
Behavior outcomes	Assess whether patients engage in activities to cope with the disease and treatments through DHIs. It refers to what the patient acts and does.
Clinical outcomes	Assess whether patients' health improvements meet the intentions of the DHIs. It refers to what medical condition the patient is in and aims to maintain.
Health care system impact [29]	
Economic outcomes	Assess whether the DHIs are cost-effective, whether the organization and DH users can afford the DHI system, and whether there is a probable return on investment. It refers to the use of health care resources.

[a]DHI: digital health intervention.

[b]DH: digital health.

XSL•FO
RenderX

Table 3. Initial codes of evaluation approaches of the digital patient experience.

Categories and initial codes	Description
Study designs [29]	
Descriptive study	Aims to define the "who, what, when, and where" of the observed phenomena and include qualitative research concerning both individuals and populations.
Analytical study	Aims to quantify the relationship between the intervention and the outcomes of interest, usually with the specific aim of demonstrating a causative link between the 2, including experimental and observational studies.
Data collection methods and instruments [103]	
Qualitative methods	Qualitative research is expressed in words. It is used to understand concepts, thoughts, or experiences. Common qualitative methods include interviews with open-ended questions, observations described in words, and literature reviews that explore concepts and theories.
Quantitative methods	Quantitative research is expressed in numbers and graphs. It is used to test or confirm theories and assumptions. Common quantitative methods include experiments, observations recorded as numbers, and surveys with closed-ended questions.
Qualitative analysis	Qualitative data consist of text, images, or videos instead of numbers. Content analysis, thematic analysis, and discourse analysis are the common approaches used to analyze these types of data.
Quantitative analysis	Quantitative data are based on numbers. Simple math or more advanced statistical analysis is used to discover commonalities or patterns in the data.

Phase 2: The Updated Scoping Search

The decision to undertake an update of a review requires several considerations. Review authors should consider whether an update for a review is necessary and when it will be more appropriate [104]. In light of the "decision framework to assess systematic reviews for updating, with standard terms to report such decisions" [105], we consider that research on PEx in DH remains important and evolves rapidly. In case we missed some newly published articles that would bring significant changes to our initial findings, we conducted a rapid scoping search for articles published after our last search. We reran the search strategy as specified before with the addition of date (from December 16, 2020, to August 18, 2023) limits set to the period following the most recent search. After removing duplicates (73/367, 19.8%), we collected 294 articles in total. Following the same screening process and selection criteria, we finally identified 102 new eligible articles. The excluded articles were either not a literature review with systematic search (74/294, 25.2%), not about DH (87/294, 29.6%), not about PEx (26/294, 8.8%), our own parallel publications (2/294, 0.7%), or not accessible in full text (3/294, 1%). The eligible and ineligible articles in this phase are available in Multimedia Appendix 2. We found that the outcomes in the new studies were almost consistent with the existing data. For example, these articles either aimed to investigate what factors influence the feasibility, efficacy, effectiveness, design, and implementation of DH; examine how patients expect, perceive, and experience the DHIs; or intend to compare the DHIs with conventional face-to-face health care services. The research objectives of these new eligible articles are available in Multimedia Appendix 3. We considered that their findings were unlikely to meaningfully impact our findings on when to measure, what to measure, and how to measure digital PEx. As suggested by Cumpston and Chandler [104], review authors should decide whether and when to update the review based on their expertise and individual assessment of the subject matter. We decided to use these new articles as supplementary materials (Multimedia Appendices 2 and 3) but did not integrate them into the synthesis of this review.

Results

General Findings

This paper is a part of a larger study, and we have presented results related to study characteristics in a previous publication [3]. Multimedia Appendix 4 [53-97] provides detailed information regarding the characteristics of the included reviews, including research questions or aims, review types, analysis methods, number of included studies, target populations, health issues, and DHIs reported in each review. In this study, to achieve our research objectives, we identified reviews that reported different intervention maturity stages, timing of the evaluation, and timing of data collection. In addition, we identified a set of evaluation indicators of digital PEx and classified them into 3 predefined categories (ie, intervention outputs, patient outcomes, and health care system impact), which in turn included 9 themes and 22 subthemes. Furthermore, we highlighted evaluation approaches in terms of evaluation theories, study designs, data collection methods and instruments, and data analysis methods. We found that it was valuable to compare the evaluation objectives of the included studies. Therefore, we captured 5 typical evaluation objectives and the stakeholders involved, which clarified why and for whom DH evaluators carried out the evaluation tasks. The detailed findings are presented in the Evaluation Objectives section.

Evaluation Objectives

Our review findings highlighted 5 typical evaluation objectives.

The first objective was to *broaden the general understanding of the digital PEx and guide evaluation research and practice* (11/45, 24%) [53-63]. For instance, 1 review [61] aimed to identify implications for future evaluation research and practice on mental health smartphone interventions by investigating UX evaluation approaches.

XSL•FO
RenderX

The second was to *improve the design, development, and implementation of the DHI in terms of a better digital PEx* (15/45, 33%) [53-62,64-68]. As demonstrated in an included review [58], the evaluation of DHIs is critical to assess progress, identify problems, and facilitate changes to improve health service delivery and achieve the desired outcomes.

The third was to *achieve evidence-based clinical use and increase DHIs' adoption and uptake* (14/45, 31%) [53,55,56,58-62,64,65,67,69-71].

The fourth was to *drive ongoing investment* (3/45, 7%) [53,70,71]; without compelling economic supporting evidence, the proliferation of DHIs will not occur. Therefore, ensuring the sustained clinical use, successful implementation, and adoption of and continued investment in DHIs require more evaluative information. This helps ensure that resources are not wasted on ineffective interventions [53].

The fifth was to *inform health policy practice* (3/45, 7%) [55,56,61]. As the 2 included articles stated [55,56], ongoing evaluation and monitoring of DHIs is critical to inform health policy and practice. In addition, in terms of the varied evaluation objectives, the evaluation activities serve different stakeholder groups, including program investigators, evaluators, and researchers; designers, developers, and implementers; end users, patients, and health care providers (HCPs); clients and investors; and governments and policymakers.

Evaluation Timing Considerations

Among the included studies, evaluations were carried out at various stages of the intervention to fulfill the 5 evaluation objectives. Our findings showed that most reviews reported feasibility, efficacy, and pilot studies (32/45, 71%) [53,54,56,58-64,67,69-89] and then investigated effectiveness (20/45, 44%) [53,58,60,61,63,65,69,70,72,75,79,80,82-84,86,90-93] and implementation studies (20/45, 44%) [54,56,58,61,62,64,68,70,73,75,78,81-83,85,87,88,90,91,94]. Notably, some reviews included >1 type of study. Our findings show that the timing of evaluation can be directly at pre- or postintervention [53,56,59,60,62-65,69-72,77,78,81,82,84,85,87,88,90,91], at the baseline point or after a short- or long-term follow-up intervention [58,59,62,65,67,70-72,76,77,79,81,82,84,85,88,91,95,96], during intervention use [76,85], continued monitoring [56,75], and even at dropout [53]. One study [84] suggested providing a period of technical training and conducting a baseline test to reduce the evaluation bias caused by individual technology familiarity and novelty. As demonstrated by another study [53], pre- and postintervention assessments using clinical trials can measure intervention effectiveness (eg, patients' blood glucose levels). In terms of the timing of data collection, 1 included study [53] suggested that evaluations directly after the intervention are appropriate so that the users retain fresh memories of the experience. To sustain intervention outcomes over a longer period, longitudinal evaluations and long-term follow-up evaluations were recommended in 2 studies [55,84].

Evaluation Indicators

Overview

Evaluation indicators relate to the goal to which the research project or commercial program intends to contribute. Indicators are defined as "a quantitative or qualitative factor or variable that provides a simple and reliable means to measure achievement, to reflect the changes connected to an intervention, or to help assess the performance of a development actor" [106]. On the basis of our initial codes, we grouped the evaluation indicators into 3 main categories: intervention outputs, patient outcomes, and health care system impact. Each category contains several themes and subthemes (Tables 4-6) and is discussed in detail in the below 3 sections: Intervention Outputs, Patient Outcomes, and Health Care System Impact.

XSL•FO
RenderX

Table 4. Themes, subthemes, and evaluation indicators of the intervention outputs of the digital patient experience.

Themes and subthemes	Studies (n=45), n (%)	Evaluation indicators	References
Functionality (n=36, 80%)			
Intended values	21 (47)	• Ability to either change or maintain the user's health state in a beneficial way: support self-management, shared decision-making, trigger actions, and track and respond to changes • Ability to collect clinical metrics: the number of monitored variables and the frequency, accuracy, concordance, timeliness, and visibility of monitoring	[53,54,56,61-63,65-69,74,79,81-83,88,91,93,96,97]
Content and information	20 (44)	• Quality of the content: evidence based, tailored, relevance, practicality, consistency, and clarity • Amount of the information: comprehensible, completeness, glanceability (understandability), and conciseness • Language of the information: simple nontechnical language; actionable message; and a nonauthoritarian, friendly, and nonjudgmental tone of voice	[54-56,59,61,64,65,67,68,71,76,80,81,83,84,86,89,90,93,95]
Intervention features	20 (44)	• Appropriate features that meet the intended values: activity planning, activity scheduling, activity tracking, diary, alerts, journal, feedback, and reminders • Degree of setup, maintenance, and training: ready to use, initial training, and ongoing education • Channel or mode of delivery: phone calls, social media, mobile apps, web, video, devices, and wearable kit	[53-56,60-65,67-69,76,80,81,83,86,89,94]
Theory-based interventions	11 (24)	• Presence or absence of an underlying theoretical basis: behavior change theory, social presence, and a quality certification	[54,59,68,71,78,83,86,88-90,92]
Usability (n=26, 58%)			
Technology quality attributes	24 (53)	• Technology operability: the ease of use, learnability, memorability, readability, efficiency, system errors, product, or service • Technology standards and specifications: interoperability, integration, scalability, ergonomics, connectivity, adaptability, flexibility, accuracy, and reliability	[53-57,59-62,64,65,67,68,71,73,75,76,82,86,88,91,93,95,96]
Interaction design	17 (38)	• Use of human-centered design methodologies during the development process: co-design, user-centered design, and inclusive design • Design quality of system architecture, layout, and interface: intuitive, interactive, personalized, and esthetic	[53-57,59-62,67,68,71,73,83,86,88,94]
Care quality (n=30, 67%)			
Accessible care	27 (60)	• Accessibility of care services: data, information, and HCPs[a] • Involvement of related stakeholders: family, friends, and peer-to-peer communication • Accessibility to high-quality care: timely, integrated, continuous, improved (more predictable daily life), convenient (fits into daily routines), and personalized care	[53-56,59,62-64,66-69,71,75-81,84,86>,88,91,92,95,97]
Safe and credible care	14 (31)	• Credibility and accountability of care: the owners' credibility and third-party verification • Security of care: the number of medical errors • Privacy of care: the presence of general privacy notifications, the documentation of individual access to user private data, and regulation compliance	[53-56,67,68,71-73,79,80,92-94]

[a]HCP: health care provider.

XSL•FO
RenderX

Table 5. Themes, subthemes, and evaluation indicators of patient outcomes of the digital patient experience.

Themes and subthemes	Studies (n=45), n (%)	Evaluation indicators	References
Emotional outcomes (n=32, 71%)			
Positive emotions	31 (69)	• Patient satisfaction • A sense of reassurance • Well-being • A sense of security • Peace of mind • A sense of belonging	[53,55,57,61-67,69,71,72,74-78,81,82,84,85,88-92,94-97]
Negative emotions	16 (36)	• Concerns • Fears • A sense of uncertainties • Dissatisfaction • A sense of frustration • A sense of insecurity • Worries	[55,61,62,64,66,67,71,76,78,85,86,88,91,92,95,97]
Perceptual outcomes (n=32, 71%)			
Empowerment	23 (51)	• Perceived values • Quality of life • Confidence • Self-efficacy • Comfort	[53,55,56,61,62,65-67,69,71,74-78,84,85,88-92,95]
Acceptability	19 (42)	• Degree to which technology, treatment, and care services are accepted: willingness to use, intention to use, intention to continue using, and likelihood to recommend	[53,56,59-62,65,67,69,72,74-76,78,80,82,86,91,95]
Connectedness	16 (36)	• Relationships between patient and provider: closeness, detachment, trust, or doubts	[53,61,64,66-69,71,72,75-78,88,92,97]
Attitudes	14 (31)	• Initial beliefs, preferences, and expectations • Impression of the excellence of the DHIs[a] • Interpretation of the DHIs • Motivation to change behavior	[53,55,64,65,67,69,72,74-78,80,88]
Burden	12 (27)	• Perceived burden and restriction • Discomfort • Unconfident	[55,66,67,72,76-78,85,89,92,95,97]
Capability outcomes (n=19, 42%)			
Autonomy and knowledge-gaining	19 (42)	• Participants' level of informed state of mind after using the DHIs: clinical awareness • Patients' level of health knowledge: health literacy, skills, and understanding • Patients' ability to make clinical decisions: problem-solving and shared decision-making	[53,56,62,64,65,69,71,72,75-79,84,88,90,92,95,97]
Behavioral outcomes (n=26, 58%)			
Adherence	19 (42)	• Initial, sustained use of certain features • Download and deletion rates • Completion rates • Dropout rates • Speed of task completion	[55,61-63,65,67,69,74-76,78,79,82,84,88-91,95]
Self-management behaviors	17 (38)	• Number of individuals exercising regularly or using dietary behaviors compared with the total number of participants • Engagement of treatment, self-care, and help-seeking behavior	[53,59,61,62,65,67,69,74,75,78,81,84,85,88-90,92]

XSL•FO
RenderX

Themes and subthemes	Studies (n=45), n (%)	Evaluation indicators	References
Patient-provider communication	11 (24)	• Number and frequency of patient-provider contacts • Engagement of patient-provider communication • Quality of patient-provider communication (eg, percentage of patients reporting that HCPs[b] communicated well)	[59,63,64,69,71,75,77,79,81,88,92]
Clinical outcomes (n=23, 51%)			
Health conditions	23 (51)	• Level of pain and symptoms control • Status of physical health • Level of health or treatment-related anxiety, depression, and stress • Mortality rates • Morbidity rates • Adverse effects	[59-65,67,69,71,74-76,78,79,81,82,88-91,95,96]

[a]DHI: digital health intervention.

[b]HCP: health care provider.

Table 6. Themes, subthemes, and evaluation indicators of health care system impact of the digital patient experience.

Themes and subthemes	Studies (n=45), n (%)	Evaluation indicators	References
Economic outcomes (n=16, 36%)			
Cost-effectiveness	14 (31)	• Out-of-pocket expenses for patients: care costs and travel costs • Time efficiency of using the DHIs[a]: waiting time, travel time, and consultation time • Reduction in overuse of services: printed materials	[56,63,64,66,72,76,77,79,81,88,90,91,96,97]
Health care service use	8 (18)	• Duration of consultations • Number of hospitals, primary care, and emergency department visits • Hospital admissions • Hospitalization • Proportion of referrals	[62,64,76,81,88,90-92]

[a]DHI: digital health intervention.

Intervention Outputs

Intervention outputs are partially determined by the intervention inputs and processes (ie, influencing factors and design considerations, such as personalized design) [3]. We identified 3 themes and 8 subthemes within this category (Table 2). The first theme, *functionality*, refers to the assessment of whether the DHIs work as intended. The subthemes included (1) the consistency of intended value (eg, the ability of the DHIs to collect the amount of accurate clinical metrics in real time [56,62,74,88]), (2) the quality of content and information (eg, tailored content [56,64,76,81,83,86,89,90]), (3) the appropriateness of intervention features (eg, the degree of system setup [54,69]), and (4) the use of intervention theories (eg, the presence of an underlying theoretical basis [54,59,68,78,83,86,88,90,92]). The second theme, *usability*, refers to whether the DH system is used as intended [29]. Both technology quality attributes (eg, ease of use [53-56,59,61,62,67,68,71,76,86,95]) and interaction design (eg, intuitive interface design [67,68,94]) can be used for usability evaluations. The third theme, *care quality*, refers to effective, safe, people-centered, timely, accessible, equitable, integrated, and efficient care services [102]. For example, the assessment of convenient care accessibility (eg, care that fits into daily routines [53,59,62,76,77,81,86,88] and the credibility of DHIs' owners [53,54]).

Patient Outcomes

Studies used a variety of quantitative and qualitative factors and variables to measure and describe patient outcomes (Table 3), referring to 5 themes (emotional outcomes, perceptual outcomes, capability outcomes, behavioral outcomes, and clinical outcomes) and 12 subthemes. *Emotional outcomes* relate to patients' positive or negative feelings that result from the use or anticipated use of DHIs. For example, a high level of patient satisfaction [53,55,57,61-67,69,72,74-76,82,84,89-91,94-96] is a typical positive feeling. Increased concerns about data privacy and security [55,64,67,71,76,86,95,97] is a frequently mentioned negative feeling. *Perceptual outcomes* are the informed states of mind or nonemotional feelings the patients achieve before, during, or after using the DHIs [69], including patients' initial attitudes toward the DHIs (eg, internal motivation [53,64,69,75,77,78,88]); patient-to-provider relationships, for example, those that are enhanced by perceived improved accessibility to HCPs [53,67,69,71,75,76,78,88,92] versus those that are interfered with by perceived loss of face-to-face contacts [61,64,66,71,76,77,97]; perceived empowerment (eg, increased confidence in managing their health conditions

XSL•FO
RenderX

[56,69,75,77,78,90]) and burden (eg, increased perception of restriction [55,76-78,85,92,95,97]); and overall acceptance of the DHIs (eg, willingness to use [61,62,67,72]). *Capability outcomes* refer to the improvement in patients' self-management autonomy, health knowledge, and clinical awareness. DHIs may be effective at improving their independency, self-management autonomy, problem-solving, and decision-making skills [53,62,64,65,69,71,75-79,84,88,92,95]; gaining health literacy, knowledge, or understanding of their health conditions or care plans [53,56,72,75,79,88,90,92,97]; and raising their clinical awareness to be more certain of when it was necessary to seek medical attention [69,71,72,78,92]. *Behavioral outcomes* include activities that the patients adopt owing to DHIs [69], including adherence to the intervention (eg, dropout rates [61,65,69,74,76,82,84]), self-management behaviors (eg, physical and diet activities [65,67,74,78,84,88,89]), and patient-to-provider communication (eg, increased interactions between patients and HCPs [59,63,64,69,71,75,77,79,81,88,92]). *Clinical outcomes* are related to individual health conditions and the main intentions of the DHIs. For example, a reduction in anxiety, depression, and stress [59,61-65,69,71,75,76,78,81,82,89,95] and increased symptom control [67,69,71,75,78,88-90,96] can help to measure the individual health conditions.

Health Care System Impact

Health care system impact contains 1 theme and 2 subthemes. *Economic outcomes* refer to the cost-effectiveness and health care services use. In terms of cost-effectiveness, for example, studies report less out-of-pocket expenses for patients because of reduced care and travel costs [56,63,64,79,81,88,90,91,97] and greater time efficiency owing to shorter waiting, travel, and consultation time [66,72,76,77,81,91,96]. Furthermore, indicators related to health care service use, such as the reduced number of hospital [62,64,76,90,91] and emergency department visits [90,91], can be used to assess savings regarding health care services.

Evaluation Approaches

Overview of the Approaches

In addition to evaluation timing considerations and indicators, strategies and specifically designed tools for collecting and analyzing data are required to set up the evaluation plan. Various evaluation approaches were identified based on our initial codes; these are depicted in 3 aspects (Tables 7-9): study designs, data collection methods and instruments, and data analysis approaches. Furthermore, we collected data related to evaluation theories that were used to guide the study designs, data collection, and analysis.

XSL•FO
RenderX

Table 7. Study designs for evaluating the digital patient experience.

Study designs	Studies, n (%)	References
Mode of inquiry (n=36, 80%)		
• Qualitative research	35 (78)	[53,55,56,58,59,61,62,64-72,74-83,85,86,88,90-92,95-97]
• Phenomenology • Ethnography		
• Quantitative research	21 (47)	[53,55,58,61,62,64,66-68,70,71,74,76,83,85,86,88,90,91,95,96]
• Mixed methods research (and multiple methods research)	17 (38)	[53,55,56,61-63,66,68,71,76,78,80,85,86,88,91,95]
Nature of the investigation (n=33, 73%)		
• Experimental research	25 (56)	[53,58-60,62-64,70-72,75,76,78-85,88-91,95,96]
• Randomized controlled trials • Nonrandomized trials		
• Observational research	9 (20)	[60,72,76,80,84-86,88,91]
• Descriptive research	7 (16)	[55,56,68,71,72,74,91]
• Case reports • Case series • Cross-sectional		
• Analytical research	6 (13)	[55,60,71,88,91,94]
• Case control • Cohort		
Number of contacts (n=21, 47%)		
• Cross-sectional	8 (18)	[55,56,68,72,74,91]
• Longitudinal	6 (13)	[55,62,71,93,95,97]
• Before and after	4 (9)	[53,59,60,62-65,71,72,81,82,87,90]
Reference period (n=10, 22%)		
• Prospective	8 (18)	[60,62,71,72,81,89,91,94]
• Retrospective	4 (9)	[56,60,91,95]
Research through design (n=4, 9%)		
• User research	3 (7)	[55,60,87]
• Participatory design or contextual design	1 (2)	[69]
• Design sessions	1 (2)	[55]

XSL•FO
RenderX

Table 8. Data collection methods of evaluating the digital patient experience.

Data collection methods	Studies, n (%)	References
Questionnaires	33 (73)	[53,55,56,58,59,61,62,64-69,71,72,75,76,79-87,89-93,95]
Surveys	32 (71)	[53,55,58,59,61-69,71-74,76,77,79-83,86,87,89-91,93-95]
Interviews	31 (69)	[53,55,56,59-62,65-69,71-73,75-78,80-87,91,92,95,97]
Focus groups	19 (42)	[55,56,58-60,63,66-68,76,78,80,81,83,85-87,95,97]
Observations	17 (38)	[55,60,66,67,69,71,72,76,78,80,84-88,91,97]
Log data	13 (29)	[55,61,69,71-74,81,83,84,90,95,97]
Open-ended questions	10 (22)	[53,56,59,62,65,67,75,77,80,86]
Likert scales	10 (22)	[53,58,65,67,71,82,84,89,91,93]
Usability testing	8 (18)	[53,57,60,64,67,81-83]
Diaries	6 (13)	[53,55,68,80,90,97]
Contextual inquiry	5 (11)	[53,56,69,80,87]
Needs assessment	5 (11)	[53,77,82,83,87]
Performance tests	5 (11)	[60,61,65,72,84]
Field notes	4 (9)	[56,69,85,97]
Workshops	4 (9)	[67,68,82,83]
Forms	3 (7)	[53,72,82]
Think-aloud method	3 (7)	[53,68,69]
Benchmark testing	2 (4)	[61,87]
Human impact assessment methodologies	1 (2)	[95]
Personas	1 (2)	[87]

Table 9. Data analysis approaches of evaluating the digital patient experience.

Data analysis approaches	Studies, n (%)	References
Statistical analysis	15 (33)	[59-61,65,70-72,74-76,82,84,90,91,96]
Thematic analysis	11 (24)	[56,61,69,76,77,80,85,88,92,95,97]
Content analysis	9 (20)	[53,56,63,76,77,80,86,92,97]
Grounded theory	7 (16)	[53,56,61,80,85,92,97]
Framework analysis	5 (11)	[56,80,85,92,97]
Heuristic analysis	4 (9)	[61,67,80,87]
Cost analysis	4 (9)	[63,70,88,91]
Task analysis	3 (7)	[61,83,87]
Text analysis	2 (4)	[66,92]
Document analysis	2 (4)	[71,80]
Failure analysis	2 (4)	[83,87]
Inductive analysis	2 (4)	[56,97]
Deductive analysis	1 (2)	[56]
Formal analysis	1 (2)	[73]
Decision analytic approach	1 (2)	[91]

Evaluation Theories

Our findings showed that in some cases, theories are used to guide the evaluation process. An included review [58] mapped various DHI evaluation frameworks and models into conceptual, results, and logical frameworks as well as theory of change. Among the included reviews, the National Quality Forum [63,79], UX model [93], American Psychiatric Association App Evaluation Model [61], Markov model [88], and Consolidated

XSL•FO
RenderX

Framework for Implementation Research [56] were mentioned as evaluation frameworks or models for setting up, conducting, or analyzing the evaluation activities. In addition, theories from other fields such as frameworks or models related to *health care* (eg, diabetes theory [56,69], triple aims framework [91], and chronic disease management model [58]), *behaviors* (eg, social cognitive theory [59,82,93], behavior change theory [58,59,90]), *design* (eg, human factors principles [87], and inclusive design [57]), and *technology* (eg, the Unified Theory of Acceptance and Use of Technology [57,64] and Health Information Technology Usability Evaluation Model [67]) can be adopted to assess specific outputs, outcomes, or impact. For example, the behavior change theory can be used to guide the evaluation of patient behavioral outcomes [59].

Study Designs

The terminologies used to describe the study designs were mixed in terms of different classification bases. Following the work on research methodology by Kumar [101], we identified 4 standards for classifying study designs in DH: the perspectives of mode of inquiry, nature of the investigation, reference period, and number of contacts with the study population. From the perspectives of "mode of inquiry," we found 3 types of study. The first used a qualitative study design, such as phenomenology or ethnography studies. The second were quantitative studies. The third type used mixed methods research and multiple methods research (ie, >1 qualitative or quantitative method, such as using both focus groups and interviews to collect data). In addition, based on *the nature of the investigation*, the collected primary studies among the included reviews were reported as observational studies versus experimental studies (RCTs and nonrandomized trials) and descriptive studies (case reports, case series, and cross-sectional) versus analytical studies (case-control or cohort studies). On the basis of the *number of contacts* with the study population, cross-sectional, before-and-after, and longitudinal studies were mentioned. Furthermore, in terms of the *reference period* (the time frame in which a study explores a phenomenon, situation, event, or problem), some studies included prospective designs, whereas others reported retrospective study designs. In addition, we note that others reported study designs from a design perspective, such as user studies, participatory design or contextual design, and design sessions.

Data Collection Methods and Instruments

Various data collection methods were used among the included reviews: questionnaires, surveys, interviews, focus groups, observations, log data, open-ended questions, Likert scales, usability testing, diaries, contextual inquiry, needs assessment, performance tests, field notes, workshops, forms, think-aloud method, benchmark testing, human impact assessment methodologies, and personas. Notably, these data collection techniques appeared as a mixed combination in some studies. In addition, we found various standard evaluation tools and performance tests used to collect the digital PEx–related data in 18 of the included papers [53,55,57,59,61,65,67,71,75,76,79,82,84,87,89,91,94,96], including the System Usability Scale [53,61,82], Patient Activation Measure [75,84], Patient Health Questionnaire-9 [75,89], and Beck Depression Inventory [75,89]. However, none of these tools are designed for evaluating the digital PEx; most are designed or modified to evaluate UX, PEx in general, or the usability of specific DHIs.

Data Analysis Approaches

Our findings showed that different types of data were used to evaluate digital PEx, such as self-reported data [74] and observable or monitored data [61]. To analyze the evaluative information, various data analysis methods were reported among the included reviews, including statistical analysis, thematic analysis, content analysis, grounded theory, framework analysis, heuristic analysis, cost analysis, task analysis, text analysis, document analysis, failure analysis, inductive analysis, deductive analysis, formal analysis, and decision analytic approach.

Discussion

Principal Findings

The goals of this umbrella review were to systematically review the evaluation timing considerations, indicators, and approaches of digital PEx. Furthermore, we identified 5 typical evaluation objectives and related audiences. The timing of a digital PEx evaluation should be a critical consideration when conducting an evaluation study; however, we found limited information about when to measure digital PEx. Moreover, the identified evaluation indicators are often heterogeneous and appear to be related to the different aspects of digital PEx. In terms of evaluation approaches, various theories were reported in the included papers. Furthermore, we noted that not only did the evaluation methods differ between the reviews but also the classification bases or perspectives used to describe these methods. Following our findings on when to measure, what to measure, and how to measure digital PEx, we generated a step-by-step evaluation guide and proposed 6 research directions for future studies.

When to Measure

DHIs change throughout the product life cycle, so to provide better-quality results and evidence-based health practice, evaluations need to be incorporated into the intervention maturity stages [55,56,58,80]. Our findings showed that many studies were not performed in a real-world setting for a long period, and most studies were either feasibility or pilot studies; these results are directly in line with previous findings [56,65,73,82,85,89,97]. Pilot or feasibility studies can help improve new intervention development but only provide limited evidence for increasing sustained clinical use and large-scale practice [58,80]. Two studies [55,77] reported a lack of information on the long-term experience. Others have shown that some solutions may be less sustainable outside the trial context [80,85]. In addition, it is possible that participants were more adherent during the study period and decreased their use of the apps over time [74]. Therefore, some authors call for further research on digital PEx when incorporating the DHIs into existing health care services and processes [76]; there is a need to move DHIs from promise into policy and practice [56].

One study [72] reported significantly different evaluation results before and after the treatment. It is likely that patients' initial

XSL•FO
RenderX

emotional state or understanding of DHIs may affect their final PEx evaluation outcomes. Therefore, a baseline test on individual differences would be a valuable step to limit evaluation bias, as noted in a previous study [84]. We found that the data gathered could occur at a specific moment or at different time points along the care pathway to reflect a rapid or delayed digital PEx. Thus, posttreatment evaluations should account for the recall bias caused by the time delay between treatment and recollection of experience, as has been noted in previous studies [53,76]. In line with other studies [29,64], we believe that real-world testing and direct feedback from actual users will help improve the usability of DHIs and directly benefit new users.

What to Measure

In comparison with intervention outputs and health care system impact, we discovered more evaluation indicators related to patient outcomes. We assume that this is owing to the consideration of the strength of the evidence and duration of the study. Patient outcomes enable the identification of patients' actual experiences and reactions in uncontrolled settings, providing evidence for clinical use and further improvements. However, intervention outputs seem more suitable for exploring experts' (eg, designers, health care professionals, and policymakers) or patients' anticipated understandings of DHIs in the early stages of design and for addressing any potential system barriers. The health care system impact can be useful in predicting the sustainability of the DHIs on a large scale through a long-term study.

We used a set of themes and subthemes to describe each category. For instance, patient outcomes include emotional, perceptual, capability, behavioral, and clinical outcomes, as noted in 2 studies [60,75]: one study categorized the variables of patient engagement as behavioral, cognitive, and emotional outcomes, whereas the other study used biomarkers, perceptions, and behaviors to describe patient clinical outcomes with regard to DHIs. Furthermore, we noted that the evaluation outcome of one indicator is often unable to anticipate the outcome of another indicator. For instance, some patients reported high acceptance of a certain DHI, but they rarely used it [95]. Aligned with the arguments among the differences between patient satisfaction, PEx, PREMs, and PROMs [35,37,47,48], our findings indicate that digital PEx evaluations are not equivalent to the measurement of patient satisfaction, PEx, PREMs, or PROMs, but that these measures can be used to assess some of the digital PEx. We showed that the priorities of the evaluation indicators can differ between projects. In terms of what to measure first, as stated in a previous study [107], the goal of evaluations should be to focus on those processes that should be optimized by the digital catalyst. Furthermore, the evaluation indicators need to be continually updated as the DH landscape is rapidly evolving and the technology infrastructure is constantly shifting [54].

How to Measure

As demonstrated in an included review [58], an evidence-based theoretical evaluation framework is helpful in informing the evaluation process. Across the included reviews, we found that not only specifically designed evaluation theories were used to guide the evaluation activities but also theories from other fields were adopted to assess the evaluative data. We identified various traditional approaches across the included reviews. In addition, our results showed that more than half of the included reviews reported RCTs in their studies. RCTs were recommended in 2 reviews [108,109] to evaluate DHIs for stronger evidence. However, a recent systematic review [110] noted that only a handful of clinical decision support systems have been tested in this way. Others argued that there is a tension between the amount of time needed for evidence generation with traditional approaches and the speed of digital product development and iterative upgrading [16,18], which requires more innovative methods for fast evidence generation [18].

We identified a wide range of evaluation methods and instruments, although most were modified based on the evaluations for traditional face-to-face treatment or usability testing in human-computer interactions. This is also in line with the findings from previous studies [53,61,72,79]. Semistructured interviews and questionnaires were the most common evaluation methods for collecting evaluative data among the included reviews, which is in line with previous studies [53,76]. Semistructured interviews are the key methods used to understand the details of UX [59,61,62,65,66], whereas questionnaires are often modified from existing assessments to assess large-scale interventions [53,61]. It is likely that more in-depth, observational data collection methods are necessary to better capture experience data [53,66]. The use of a descriptive approach might be appropriate for a smaller sample size, collecting qualitative data through surveys, focus groups, and interviews [76]. Standard functional questionnaires may be preferred when DHIs are compared with other interventions [53]. However, we found that detailed interview outlines or questionnaires were generally not published, as mentioned in another study [53]. Comprehensive information on user evaluation methods and results is often lacking [65]. The determination of evaluation approaches depends on the specific context. In alignment with 2 studies [4,40], we state that the choice of evaluation approaches heavily depends on evaluation objectives, timing, indicators, and evaluation requirements and resources. An included review [58] recommended using multiple research methods, such as combining qualitative, quantitative, co-design principles, and process measures, for evaluation designs.

Thanks to the use of digital technologies [54,75], patients' illness experience and what they feel when participating in a health care intervention can be monitored. However, we found that these may blur the boundaries between interventions, monitoring, and evaluations. For example, the diary function can be used as an intervention feature (eg, a self-management diary to track symptoms and identify exacerbations [78]), as a monitoring tool (eg, diary entries [97] or adherence [90]), or as an evaluation method (eg, to capture user feedback [53]). Furthermore, a study indicated that with the advancement of technology, the ability of DHIs to collect "passive data" for assessing digital PEx may gain more attention and eventually eclipse the utility of DH-aided self-report [74]. Finally, we believe that involving multiple stakeholders is not only essential in the design process but is also a requirement for the evaluation

XSL•FO
RenderX

process. Both end users and experts can contribute to the evaluation activities [53]. This aligns with a recent study that suggests that digital solution evaluation requires collective efforts from multiple parties, such as health authorities, HCPs, and manufacturers [18].

Design Implications

Our analysis showed that the evaluation of a DHI follows the same evaluative process as that of traditional interventions, which supports a previous study [53]. To make the evaluation findings more comparable, more rigorous studies and standardized evaluations are suggested, including unified terminology [53,65,68], predefined measurable indicators [79,81], standardized methods [61,66], validated instruments [84,96], uniform time intervals [84], and adequate patient selection [81]. Intervention characteristics (eg, aims, expected outcomes, elements, length, frequency, and duration), study designs (eg, sample size, period, regulations, investigator, evaluators, recruitment, ethics, topic guides, or questions asked by the researchers), objectively measured patient health outcomes, and adverse events should be carefully considered when conducting and reporting an evaluation study [53,60,69,84].

Inspired by the challenges for the evaluation of DHIs [18]; shaped by the Performance of Routine Information System Management framework [100], the monitoring and evaluation DHIs guide [29], PEx measures [37], and our previous publications on influencing factors and design considerations of digital PEx [3,111]; and based on the findings of this study, we have developed a step-by-step evaluation guide for DH innovators, such as designers, developers, and evaluators (Figure 2): The first step is to clarify the evaluation objectives and determine the target audiences for the evaluation. We proposed 5 typical evaluation purposes and their related audiences. The selection of evaluation objectives can help determine the stages for evaluating the DHI. For example, we consider effectiveness and implementation studies more appropriate for achieving evidence-based clinical use and increasing adoption and uptake compared with efficacy studies. The second step is to determine the intervention contexts and foci in terms of the intervention maturity stages, including efficacy, effectiveness, and implementation. The determination of the evaluation stage is not only because of the evaluation objective but also because of the current condition of the DHI. The determination of the evaluation objectives and identification of the evaluation stage affect the consideration of influencing factors and evaluation indicators at the next step. For example, the evaluation of patient outcomes in an uncontrolled setting can provide evidence for clinical use and further improvement. The third step includes a set of influencing factors (ie, inputs and processes) and evaluation indicators (ie, outputs, outcomes, and impacts) that can be used for further formulating evaluation constructs. The former is more appropriate for formative evaluations, which often occur during the design and development process, whereas the latter is suitable for summative evaluations, which often occur during and after the implementation process. In the fourth step, we present 2 types of evaluations. On the basis of the frequency of evaluations, we can capture momentary experiences before, during, and following an intervention or monitor continuous feedback throughout the intervention. With regard to the time interval between the intervention and evaluation, assessments can reflect immediate experiences directly after the intervention or recalled experiences over an extended period. In the fifth step, we present various evaluation approaches that can be used to plan and carry out specific evaluation activities, such as study designs, data collection methods and instruments, and data analysis approaches. The consideration of study designs often affects the strength of the evidence and determines the data collection and analysis methods. Behavioral data may provide stronger evidence than opinion data. Qualitative methods, such as interviews, are more appropriate for collecting in-depth experience data for a smaller sample size in the early intervention development stages, and quantitative methods, such as questionnaires, are more suitable for investigating experience data at a large scale or comparing it with other interventions during or after the implementation stages. In the sixth step, we proposed 6 questions for the evaluation investigators to guide them in reporting the evaluation results and 5 questions to inspire them to generate theoretical or practical implications for responding to the related stakeholder groups. The answers to these 11 questions should reflect the evaluation processes and serve the evaluation objectives.

The guide can be used when setting up a digital PEx evaluation plan or guiding evaluation practice. Notably, the interrelationships between these 6 steps are not fixed; the entire evaluation plan is an iterative process; and the decisions made at the previous steps may influence the following steps, and vice versa. In addition, other considerations beyond this guide can also impact the evaluation process, such as human, time, and financial resources. Our guide presents an ideal way to conduct the evaluation of digital PEx; however, in the real world, the order of these steps may be changed or some steps may even be skipped depending on the specific project context. For instance, in certain assessment procedures, selecting an evaluation construct, such as usability, may come first, rather than taking evaluation objectives or target audiences into account. We developed this guide based on our literature analysis. It provides an overview of the most common evaluation timing considerations, indicators, and approaches used to collect digital PEx–related data. However, it may be incomplete and require updating in the future. For example, owing to the methodological limitations, we did not provide concrete recommendations on which evaluation approaches are superior for what types of DHIs. We believe that without providing a specific context and concrete project requirements, it is difficult to draw a conclusion.

XSL•FO
RenderX

Figure 2. Digital patient experience evaluation guide.

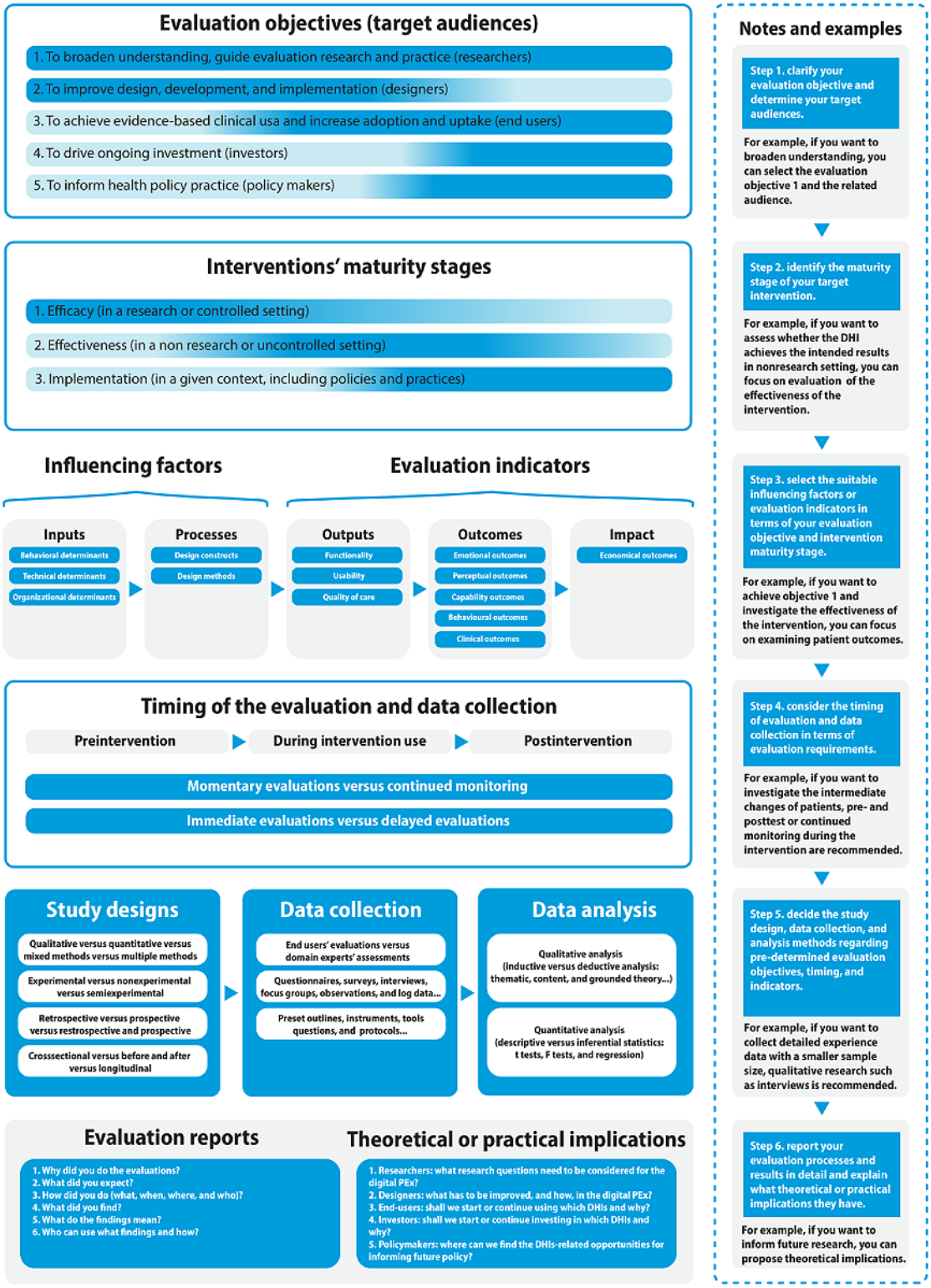

Limitations

This study had several limitations. First, we noted possible resource restrictions and the newness of the field, which may have led to missing articles. To overcome this, we searched 3 databases and used the snowballing method. In addition, we performed an updated literature search to check whether there were any meaningful new insights that would significantly change our conclusions. To our knowledge, although there were some newly published reviews in this area, we confirmed that our results were quite stable, and the newly identified studies were unlikely to significantly impact our results. Second, we could not perform a quality assessment because of the diversity in reviews and methodological limitations. As previous studies on investigating umbrella reviews have indicated, there are currently no official standards for determining the certainty of evidence when performing umbrella reviews [112,113]. In addition, among the included reviews, only 53% (24/45) of the

XSL•FO
RenderX

studies assessed the risk of bias and used diverse quality assessment instruments. After a thorough attempt, we found that none of these instruments were suitable for assessing the various reviews included. These encompass systematic reviews, scoping reviews, comprehensive overviews, and general literature reviews, which incorporate various primary and secondary studies extending beyond RCTs and nonrandomized studies of interventions. This diversity makes it challenging to use a standardized method for assessing the quality of evidence across the extensive range of included reviews. However, we tried to reduce the risk of bias by only including reviews that were published in peer-reviewed journals. Third, reviewing secondary research may have led to the omission of crucial information and reporting bias. To minimize potential bias, we used the most common terms used across the included papers as themes and subthemes. Owing to the cross-disciplinary nature of the topic, there is a lack of consistency or clarity in the terminology used to describe the evaluation indicators and approaches. For instance, in one study [55], a user study was pitched at the same level as interviews or observations, whereas in another study [56], interviews and questionnaires are methods that form part of "user study" research. In addition, information related to the evaluation approaches was reported at different levels among the included studies. For example, one study provided information related to data collection methods, such as focus groups, design sessions, and questionnaires [55], whereas another study reported information related to study designs, including qualitative, quantitative, and mixed methods designs [71]. These inconsistencies complicated the comparison between different studies. To counter this, we analyzed the different classification bases behind these study designs. Finally, we could not draw firm conclusions regarding which evaluation approaches are better suited for which types of DHIs. Owing to the nature of this study being a review of reviews, details such as the characteristics of DHIs are not always adequately covered in each included review. In addition, the included reviews contained a large number of primary studies, which makes referring back to each primary study challenging. The lack of details about the characteristics of these primary studies limits the classification of DHIs in this study. Moreover, the included reviews represent a wide range of studies, making comparison across the included reviews challenging. In addition, we decided that this was out of the scope of this study. When planning this study, we deemed it more appropriate to initially offer an overview of diversities rather than begin with a best practice recommendation. Consequently, we aimed to map possible evaluation considerations and approaches for evaluating digital PEx, instead of discussing which approach is better. However, we encourage future research to address this issue.

Future Research

Considering our research limitations, to further facilitate evaluations of digital PEx, we propose 6 future research directions. First, further research into how one indicator mediates another indicator's impact on digital PEx is required. For example, is there a correlation between clinical outcomes and perceptual outcomes? To explore this, we performed an experimental study to investigate whether patients' initial pain perception and technology acceptance (using virtual reality distraction) affected their experienced pain during wound care treatment. Our findings will be published in a future article. Second, the variables that influence the selection or prioritization of evaluation indicators and approaches should be further investigated. For example, it would be valuable to investigate whether some evaluation indicators and approaches are better suited for evaluating certain types of DHIs according to the strength of the evidence and the length of the evidence generation time. Third, agreement is needed on standardized measures to evaluate digital PEx, particularly innovative approaches for faster and high-quality evidence generation. In a follow-up interview study, we aim to summarize the often-used agile evaluation approaches based on designers' experiences. Furthermore, in cases where an interview or questionnaire is used to collect evaluative information, we recommend reporting the detailed interview outlines or questionnaires together with the evaluation results. Fourth, research is needed on how the intervention maturity stages and timing of the evaluation of the evaluation affect the evaluation results. Fifth, future studies should not only investigate whether DHIs achieve the intended results in a research setting but also assess the long-term digital PEx regarding the uptake, institutionalization, and sustainability of evidence-based DHIs in a given context and a real-world setting, including policies and practices. Finally, research is required on how to analyze and respond to the evaluative data. We recommend that future evaluation research and practice provide theoretical and practical guidance on how to use the evaluative information.

Conclusions

To effectively improve the digital PEx, knowing how to evaluate the digital PEx is as important as knowing what factors influence the digital PEx and how to design the digital PEx. Evaluating digital PEx requires clarifying the evaluation objectives, identifying stakeholder groups, considering reasonable evaluation timings, choosing relevant evaluation indicators, and selecting appropriate evaluation approaches. Following our previous publication on the influencing factors and design considerations of digital PEx [3], we first identified 5 typical evaluation objectives and related stakeholder groups. We then described potential evaluation timing considerations in terms of 4 intervention maturity stages and 3 evaluation timings. We collected knowledge on evaluation indicators of digital PEx and grouped them into 3 categories: intervention outputs, patient outcomes, and health care system impact. These were then classified into 9 themes (intervention functionality, usability, care quality, patient emotional outcomes, perceptual outcomes, capability outcomes, behavioral outcomes, clinical outcomes, and system financial outcomes) and 22 subthemes. Furthermore, we noted a set of common study designs, data collection methods and instruments, as well as data analysis methods, which can be used or adapted to evaluate digital PEx. On the basis of our findings, we developed an evaluation guide to help DHI researchers, designers, and developers further evaluate digital PEx. Finally, we recommend 6 directions for further research on digital PEx evaluation. Multimedia Appendix 5 (the PRISMA checklist) provides more detail on the structure of this review.

XSL•FO
RenderX

Acknowledgments

This work was supported by the China Scholarship Council (grant 201906790084).

Authors' Contributions

TW conceived, designed, and led the overall study conduct; carried out the data collection; led the analysis and interpretation of the data; and drafted the manuscript. TW and GG selected the papers. MM, RG, and GG participated in the overall study conduct; contributed to the analysis and interpretation of study data; and conceptualized, reviewed, and suggested modifications to the presentation of results. TW, GG, RG, and MM have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The first and second rounds of the review.
[XLSX File (Microsoft Excel File), 193 KB - jmir_v26i1e46308_app1.xlsx]

Multimedia Appendix 2
Results of a rapid scoping search.
[XLSX File (Microsoft Excel File), 58 KB - jmir_v26i1e46308_app2.xlsx]

Multimedia Appendix 3
The research aims or questions of the newly eligible reviews.
[XLSX File (Microsoft Excel File), 19 KB - jmir_v26i1e46308_app3.xlsx]

Multimedia Appendix 4
Study characteristics and digital health intervention characteristics of the included reviews.
[DOCX File , 51 KB - jmir_v26i1e46308_app4.docx]

Multimedia Appendix 5
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 32 KB - jmir_v26i1e46308_app5.docx]

References

1. Jandoo T. WHO guidance for digital health: what it means for researchers. Digit Health 2020;6:2055207619898984 [FREE Full text] [doi: 10.1177/2055207619898984] [Medline: 31949918]
2. Kellermann AL, Jones SS. What it will take to achieve the as-yet-unfulfilled promises of health information technology. Health Aff (Millwood) 2013 Jan;32(1):63-68. [doi: 10.1377/hlthaff.2012.0693] [Medline: 23297272]
3. Wang T, Giunti G, Melles M, Goossens R. Digital patient experience: umbrella systematic review. J Med Internet Res 2022 Aug 04;24(8):e37952 [FREE Full text] [doi: 10.2196/37952] [Medline: 35925651]
4. IQVIA Institute for Human Data Science Study: impact of digital health grows as innovation, evidence and adoption of mobile health apps accelerate. IQVIA. 2017 Nov 07. URL: https://www.iqvia.com/newsroom/2017/11/impact-of-digital-health-grows-as-innovation-evidence-and-adoption-of-mobile-health-apps-accelerate/ [accessed 2022-01-14]
5. Asadzadeh A, Kalankesh LR. A scope of mobile health solutions in COVID-19 pandemics. Inform Med Unlocked 2021;23:100558 [FREE Full text] [doi: 10.1016/j.imu.2021.100558] [Medline: 33842688]
6. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020 Feb 03;3(1):14 [FREE Full text] [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
7. Alkire (née Nasr) L, O'Connor GE, Myrden S, Köcher S. Patient experience in the digital age: an investigation into the effect of generational cohorts. J Retail Consum Serv 2020 Nov;57:102221 [FREE Full text] [doi: 10.1016/j.jretconser.2020.102221]
8. Shaw J, Agarwal P, Desveaux L, Palma DC, Stamenova V, Jamieson T, et al. Beyond "implementation": digital health innovation and service design. NPJ Digit Med 2018 Sep 20;1(1):48 [FREE Full text] [doi: 10.1038/s41746-018-0059-8] [Medline: 31304327]
9. Constand MK, MacDermid JC, Dal Bello-Haas V, Law M. Scoping review of patient-centered care approaches in healthcare. BMC Health Serv Res 2014 Jun 19;14(1):271 [FREE Full text] [doi: 10.1186/1472-6963-14-271] [Medline: 24947822]

XSL•FO
RenderX

10. Philpot LM, Khokhar BA, DeZutter MA, Loftus CG, Stehr HI, Ramar P, et al. Creation of a patient-centered journey map to improve the patient experience: a mixed methods approach. Mayo Clin Proc Innov Qual Outcomes 2019 Dec;3(4):466-475 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2019.07.004] [Medline: 31993565]
11. Hollander J, Ward M, Alverson D, Bashshur R, Darkins A, DePhillips H. Creating a framework to support measure development for telehealth. National Quality Forum. 2017. URL: https://www.aristamd.com/wp-content/uploads/2018/10/telehealth_final_report.pdf [accessed 2024-01-12]
12. Whitten P, Love B. Patient and provider satisfaction with the use of telemedicine: overview and rationale for cautious enthusiasm. J Postgrad Med 2005;51(4):294-300 [FREE Full text] [Medline: 16388172]
13. Riley PE, Fischer JL, Nagy RE, Watson NL, McCoul ED, Tolisano AM, et al. Patient and provider satisfaction with telemedicine in otolaryngology. OTO Open 2021 Jan 08;5(1):2473974X20981838 [FREE Full text] [doi: 10.1177/2473974X20981838] [Medline: 33474522]
14. Altinisik Ergur G, Nuhoglu S, Cobanoglu C, Sengul M, Eryildiz N, Ergur A. The patient perspective of telemedicine in the context of COVID-19 pandemic. Bull Sci Technol Soc 2022 Apr 12;42(1-2):39-53 [FREE Full text] [doi: 10.1177/02704676221094735]
15. Keown OP, Parston G, Patel H, Rennie F, Saoud F, Al Kuwari H, et al. Lessons from eight countries on diffusing innovation in health care. Health Aff (Millwood) 2014 Sep;33(9):1516-1522. [doi: 10.1377/hlthaff.2014.0382] [Medline: 25201655]
16. Desveaux L, Shaw J, Wallace R, Bhattacharyya O, Bhatia RS, Jamieson T. Examining tensions that affect the evaluation of technology in health care: considerations for system decision makers from the perspective of industry and evaluators. JMIR Med Inform 2017 Dec 08;5(4):e50 [FREE Full text] [doi: 10.2196/medinform.8207] [Medline: 29222075]
17. Classification of digital health interventions v1. 0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://iris.who.int/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf?sequence=1 [accessed 2024-01-12]
18. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020 Aug 27;3(1):110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
19. Monitoring the implementation of digital health: an overview of selected national and international methodologies. World Health Organization. 2022. URL: https://iris.who.int/bitstream/handle/10665/364227/WHO-EURO-2022-5985-45750-65816-eng.pdf?sequence=1 [accessed 2024-01-12]
20. Kamillah S, Panduragan SL, Poddar S, Abdullah BF. Patients' experiences in using diabetes self-management application: a scoping review. Kesmas: J Kesehatan Masyarakat Natl 2022 Aug 19;17(sp1):67-73 [FREE Full text] [doi: 10.21109/kesmas.v17isp1.5936]
21. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1):bjgpopen20X101020 [FREE Full text] [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
22. mHealth app economics 2017/2018: current status and future trends in mobile health. Research 2 Guidance. 2017 Nov. URL: https://research2guidance.com/?download-link=6912 [accessed 2024-01-12]
23. Koh A, Swanepoel DW, Ling A, Ho BL, Tan SY, Lim J. Digital health promotion: promise and peril. Health Promot Int 2021 Dec 13;36(Supplement_1):i70-i80 [FREE Full text] [doi: 10.1093/heapro/daab134] [Medline: 34897444]
24. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002 May 22;287(20):2691-2700. [doi: 10.1001/jama.287.20.2691] [Medline: 12020305]
25. Bindhim NF, Naicker S, Freeman B, Mcgeechan K, Trevena L. Apps promoting illicit drugs—a need for tighter regulation? J Consum Health Internet 2014 Feb 24;18(1):31-43 [FREE Full text] [doi: 10.1080/15398285.2014.869166]
26. Singh K, Diamantidis CJ, Ramani S, Bhavsar NA, Mara P, Warner J, et al. Patients' and nephrologists' evaluation of patient-facing smartphone apps for CKD. Clin J Am Soc Nephrol 2019 Apr 05;14(4):523-529 [FREE Full text] [doi: 10.2215/CJN.10370818] [Medline: 30898873]
27. Fraser H, Bailey C, Sinha C, Mehl G, Labrique A. Call to action on global eHealth evaluation: consensus statement of the WHO Global eHealth Evaluation Meeting, Bellagio, September 2011. World Health Organization. 2011. URL: https://www.who.int/publications/m/item/call-to-action-on-global-ehealth-evaluation [accessed 2024-01-12]
28. Digital implementation investment guide (DIIG): integrating digital interventions into health programmes. World Health Organization. 2020. URL: https://iris.who.int/bitstream/handle/10665/334306/9789240010567-eng.pdf?sequence=1 [accessed 2024-01-12]
29. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://iris.who.int/bitstream/handle/10665/252183/9789241511766-eng.pdf?sequence=1 [accessed 2024-01-12]
30. Wolf JA, Niederhauser V, Marshburn D, LaVela SL. Defining patient experience. Patient Exp J 2014;1(1):7-19 [FREE Full text] [doi: 10.35680/2372-0247.1004]

XSL•FO
RenderX

31. Jokela T, Iivari N, Matero J, Karukka M. The standard of user-centered design and the standard definition of usability: analyzing ISO 13407 against ISO 9241-11. In: Proceedings of the 2003 Latin American conference on Human-computer interaction. 2003 Presented at: CLIHC '03; August 17-20, 2003; Rio de Janeiro, Brazil p. 53-60. [doi: 10.1145/944519.944525]
32. Tecco H. 2016 Year end funding report: a reality check for digital health. Rock Health. 2016. URL: https://rockhealth.com/insights/2016-year-end-funding-report-a-reality-check-for-digital-health/ [accessed 2022-05-17]
33. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-01-14]
34. Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? Soc Sci Med 2005 Feb;60(4):833-843. [doi: 10.1016/j.socscimed.2004.06.022] [Medline: 15571900]
35. LaVela SL, Gallan AS. Evaluation and measurement of patient experience. Patient Exp J 2014;1(1):28-36 [FREE Full text] [doi: 10.35680/2372-0247.1003]
36. Timing of the evaluation and intervention cycle. European Union. URL: https://europa.eu/capacity4dev/evaluation_guidelines/wiki/timing-evaluation-and-intervention-cycle-0#anchor2 [accessed 2024-01-14]
37. Coulter A, Fitzpatrick R, Cornwell J. The Point of Care Measures of Patients' Experience in Hospital: Purpose, Methods and Uses. London, UK: King's Fund; 2009.
38. Bolton RN, McColl-Kennedy JR, Cheung L, Gallan A, Orsingher C, Witell L, et al. Customer experience challenges: bringing together digital, physical and social realms. J Serv Manag 2018 Sep 07;29(5):776-808 [FREE Full text] [doi: 10.1108/josm-04-2018-0113]
39. Richardson B, Campbell-Yeo M, Smit M. Mobile application user experience checklist: a tool to assess attention to core UX principles. Int J Hum-Comput Int 2021 Feb 01;37(13):1283-1290. [doi: 10.1080/10447318.2021.1876361]
40. Norman D, Nielsen J. The definition of user experience (UX). Nielsen Norman Group. 2016. URL: https://www.nngroup.com/articles/definition-user-experience/#:~:text=Summary%3A%20%22User%20experience%22%20encompasses,Jakob%20Nielsen [accessed 2024-01-14]
41. Reeves R, Coulter A, Jenkinson C, Cartwright J, Bruster S, Richards N. Development and pilot testing of questionnaires for use in the acute NHS trust inpatient survey programme. Picker Institute. 2002. URL: https://nhssurveys.org/Filestore/documents/Re-development_report_of_the_acute_adult_inpatients_survey_2005.pdf [accessed 2024-01-12]
42. Baker A. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press; 2001.
43. Staniszewska S, Boardman F, Gunn L, Roberts J, Clay D, Seers K, et al. The Warwick patient experiences framework: patient-based evidence in clinical guidelines. Int J Qual Health Care 2014 Apr 20;26(2):151-157. [doi: 10.1093/intqhc/mzu003] [Medline: 24556816]
44. NHS patient experience framework. NHS National Quality Board. 2011. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/215159/dh_132788.pdf [accessed 2024-01-14]
45. Shandley LM, Hipp HS, Anderson-Bialis J, Anderson-Bialis D, Boulet SL, McKenzie LJ, et al. Patient-centered care: factors associated with reporting a positive experience at United States fertility clinics. Fertil Steril 2020 Apr;113(4):797-810 [FREE Full text] [doi: 10.1016/j.fertnstert.2019.12.040] [Medline: 32147181]
46. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence. 2022. URL: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [accessed 2024-01-12]
47. Rockville M. What is patient experience? Agency for Healthcare Research and Quality. 2016. URL: https://www.ahrq.gov/cahps/about-cahps/patient-experience/index.html [accessed 2024-01-12]
48. Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. BJA Educ 2017 Apr;17(4):137-144 [FREE Full text] [doi: 10.1093/bjaed/mkw060]
49. Hodgson P. User experience research and strength of evidence. Userfocus. 2017. URL: https://userfocus.co.uk/articles/strength-of-evidence.html [accessed 2023-05-03]
50. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009 Oct;62(10):e1-34 [FREE Full text] [doi: 10.1016/j.jclinepi.2009.06.006] [Medline: 19631507]
51. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009 Jun 27;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
52. Frank SR. Digital health care--the convergence of health care and the internet. J Ambul Care Manage 2000 Apr;23(2):8-17. [doi: 10.1097/00004479-200004000-00003] [Medline: 10848396]
53. Feather JS, Howson M, Ritchie L, Carter PD, Parry DT, Koziol-McLain J. Evaluation methods for assessing users' psychological experiences of web-based psychosocial interventions: a systematic review. J Med Internet Res 2016 Jun 30;18(6):e181 [FREE Full text] [doi: 10.2196/jmir.5455] [Medline: 27363519]
54. Baumel A, Birnbaum ML, Sucala M. A systematic review and taxonomy of published quality criteria related to the evaluation of user-facing eHealth programs. J Med Syst 2017 Aug 22;41(8):128. [doi: 10.1007/s10916-017-0776-6] [Medline: 28735372]
55. Sakaguchi-Tang DK, Bosold AL, Choi YK, Turner AM. Patient portal use and experience among older adults: systematic review. JMIR Med Inform 2017 Oct 16;5(4):e38 [FREE Full text] [doi: 10.2196/medinform.8092] [Medline: 29038093]

XSL•FO
RenderX

56. Slater H, Campbell JM, Stinson JN, Burley MM, Briggs AM. End user and implementer experiences of mHealth technologies for noncommunicable chronic disease management in young adults: systematic review. J Med Internet Res 2017 Dec 12;19(12):e406 [FREE Full text] [doi: 10.2196/jmir.8888] [Medline: 29233804]
57. Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: a literature based framework (MOLD-US). Int J Med Inform 2018 Jun;114:66-75. [doi: 10.1016/j.ijmedinf.2018.03.012] [Medline: 29673606]
58. Bashi N, Fatehi F, Mosadeghi-Nik M, Askari MS, Karunanithi M. Digital health interventions for chronic diseases: a scoping review of evaluation frameworks. BMJ Health Care Inform 2020 Mar 09;27(1):e100066 [FREE Full text] [doi: 10.1136/bmjhci-2019-100066] [Medline: 32156751]
59. Brigden A, Anderson E, Linney C, Morris R, Parslow R, Serafimova T, et al. Digital behavior change interventions for younger children with chronic health conditions: systematic review. J Med Internet Res 2020 Jul 31;22(7):e16924 [FREE Full text] [doi: 10.2196/16924] [Medline: 32735227]
60. Choi W, Wang S, Lee Y, Oh H, Zheng Z. A systematic review of mobile health technologies to support self-management of concurrent diabetes and hypertension. J Am Med Inform Assoc 2020 Jun 01;27(6):939-945 [FREE Full text] [doi: 10.1093/jamia/ocaa029] [Medline: 32357368]
61. Lemon C, Huckvale K, Carswell K, Torous J. A narrative review of methods for applying user experience in the design and assessment of mental health smartphone interventions. Int J Technol Assess Health Care 2020 Jan 24;36(1):64-70. [doi: 10.1017/s0266462319003507]
62. Leonardsen AL, Hardeland C, Helgesen AK, Grøndahl VA. Patient experiences with technology enabled care across healthcare settings- a systematic review. BMC Health Serv Res 2020 Aug 24;20(1):779 [FREE Full text] [doi: 10.1186/s12913-020-05633-4] [Medline: 32838784]
63. O'Keefe M, White K, Jennings JA. Asynchronous telepsychiatry: a systematic review. J Telemed Telecare 2019 Jul 29;27(3):137-145. [doi: 10.1177/1357633x19867189]
64. Palacholla RS, Fischer N, Coleman A, Agboola S, Kirley K, Felsted J, et al. Provider- and patient-related barriers to and facilitators of digital health technology adoption for hypertension management: scoping review. JMIR Cardio 2019 Mar 26;3(1):e11951 [FREE Full text] [doi: 10.2196/11951] [Medline: 31758771]
65. Wesselman LM, Hooghiemstra AM, Schoonmade LJ, de Wit MC, van der Flier WM, Sikkes SA. Web-based multidomain lifestyle programs for brain health: comprehensive overview and meta-analysis. JMIR Ment Health 2019 Apr 09;6(4):e12104 [FREE Full text] [doi: 10.2196/12104] [Medline: 30964438]
66. Ingemann C, Hansen NF, Hansen NL, Jensen K, Larsen CV, Chatwood S. Patient experience studies in the circumpolar region: a scoping review. BMJ Open 2020 Oct 05;10(10):e042973 [FREE Full text] [doi: 10.1136/bmjopen-2020-042973] [Medline: 33020108]
67. Molina-Recio G, Molina-Luque R, Jiménez-García AM, Ventura-Puertos PE, Hernández-Reyes A, Romero-Saldaña M. Proposal for the user-centered design approach for health apps based on successful experiences: integrative review. JMIR Mhealth Uhealth 2020 Apr 22;8(4):e14376 [FREE Full text] [doi: 10.2196/14376] [Medline: 32319965]
68. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020 Dec 09;22(12):e21687 [FREE Full text] [doi: 10.2196/21687] [Medline: 33295292]
69. Jalil S, Myers T, Atkinson I. A meta-synthesis of behavioral outcomes from telemedicine clinical trials for type 2 diabetes and the Clinical User-Experience Evaluation (CUE). J Med Syst 2015 Mar 13;39(3):28. [doi: 10.1007/s10916-015-0191-9] [Medline: 25677954]
70. Eze ND, Mateus C, Cravo Oliveira Hashiguchi T. Telemedicine in the OECD: an umbrella review of clinical and cost-effectiveness, patient experience and implementation. PLoS One 2020 Aug 13;15(8):e0237585 [FREE Full text] [doi: 10.1371/journal.pone.0237585] [Medline: 32790752]
71. Steindal SA, Nes AA, Godskesen TE, Dihle A, Lind S, Winger A, et al. Patients' experiences of telehealth in palliative home care: scoping review. J Med Internet Res 2020 May 05;22(5):e16218 [FREE Full text] [doi: 10.2196/16218] [Medline: 32369037]
72. Swanepoel DW, Hall 3rd JW. A systematic review of telehealth applications in audiology. Telemed J E Health 2010 Mar;16(2):181-200. [doi: 10.1089/tmj.2009.0111] [Medline: 20187743]
73. Memon M, Wagner SR, Pedersen CF, Beevi FH, Hansen FO. Ambient assisted living healthcare frameworks, platforms, standards, and quality attributes. Sensors (Basel) 2014 Mar 04;14(3):4312-4341 [FREE Full text] [doi: 10.3390/s140304312] [Medline: 24599192]
74. Firth J, Torous J. Smartphone apps for schizophrenia: a systematic review. JMIR Mhealth Uhealth 2015 Nov 06;3(4):e102 [FREE Full text] [doi: 10.2196/mhealth.4930] [Medline: 26546039]
75. Barello S, Triberti S, Graffigna G, Libreri C, Serino S, Hibbard J, et al. eHealth for patient engagement: a systematic review. Front Psychol 2015 Jan 08;6:2013 [FREE Full text] [doi: 10.3389/fpsyg.2015.02013] [Medline: 26779108]
76. Jones L, Grech C. The patient experience of remote telemonitoring for heart failure in the rural setting: a literature review. Contemp Nurse 2016 Dec 15;52(2-3):230-243. [doi: 10.1080/10376178.2015.1111154] [Medline: 26525077]

XSL•FO
RenderX

77. Cox A, Lucas G, Marcu A, Piano M, Grosvenor W, Mold F, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 09;19(1):e11 [FREE Full text] [doi: 10.2196/jmir.6575] [Medline: 28069561]
78. Morton K, Dennison L, May C, Murray E, Little P, McManus RJ, et al. Using digital interventions for self-management of chronic physical health conditions: A meta-ethnography review of published studies. Patient Educ Couns 2017 Apr;100(4):616-635 [FREE Full text] [doi: 10.1016/j.pec.2016.10.019] [Medline: 28029572]
79. Rising KL, Ward MM, Goldwater JC, Bhagianadh D, Hollander JE. Framework to advance oncology-related telehealth. JCO Clin Cancer Inform 2018 Dec;2:1-11 [FREE Full text] [doi: 10.1200/CCI.17.00156] [Medline: 30652582]
80. Ames HM, Glenton C, Lewin S, Tamrat T, Akama E, Leon N. Clients' perceptions and experiences of targeted digital communication accessible via mobile devices for reproductive, maternal, newborn, child, and adolescent health: a qualitative evidence synthesis. Cochrane Database Syst Rev 2019 Oct 14;10(10):CD013447 [FREE Full text] [doi: 10.1002/14651858.CD013447] [Medline: 31608981]
81. De La Cruz Monroy MF, Mosahebi A. The use of smartphone applications (apps) for enhancing communication with surgical patients: a systematic review of the literature. Surg Innov 2019 Apr 02;26(2):244-259. [doi: 10.1177/1553350618819517] [Medline: 30602332]
82. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
83. Søgaard Neilsen A, Wilson RL. Combining e-mental health intervention development with human computer interaction (HCI) design to enhance technology-facilitated recovery for people with depression and/or anxiety conditions: an integrative literature review. Int J Ment Health Nurs 2019 Feb 22;28(1):22-39. [doi: 10.1111/inm.12527] [Medline: 30133096]
84. Kuijpers W, Groen WG, Aaronson NK, van Harten WH. A systematic review of web-based interventions for patient empowerment and physical activity in chronic diseases: relevance for cancer survivors. J Med Internet Res 2013 Feb 20;15(2):e37 [FREE Full text] [doi: 10.2196/jmir.2281] [Medline: 23425685]
85. Brunton L, Bower P, Sanders C. The contradictions of telehealth user experience in chronic obstructive pulmonary disease (COPD): a qualitative meta-synthesis. PLoS One 2015 Oct 14;10(10):e0139561 [FREE Full text] [doi: 10.1371/journal.pone.0139561] [Medline: 26465333]
86. Lim S, Tan A, Madden S, Hill B. Health professionals' and postpartum women's perspectives on digital health interventions for lifestyle management in the postpartum period: a systematic review of qualitative studies. Front Endocrinol (Lausanne) 2019 Nov 8;10:767 [FREE Full text] [doi: 10.3389/fendo.2019.00767] [Medline: 31787932]
87. Fouquet SD, Miranda AT. Asking the right questions-human factors considerations for telemedicine design. Curr Allergy Asthma Rep 2020 Aug 29;20(11):66 [FREE Full text] [doi: 10.1007/s11882-020-00965-x] [Medline: 32862299]
88. Greenhalgh T, A'Court C, Shaw S. Understanding heart failure; explaining telehealth - a hermeneutic systematic review. BMC Cardiovasc Disord 2017 Jun 14;17(1):156 [FREE Full text] [doi: 10.1186/s12872-017-0594-2] [Medline: 28615004]
89. Rincon E, Monteiro-Guerra F, Rivera-Romero O, Dorronzoro-Zubiete E, Sanchez-Bocanegra CL, Gabarron E. Mobile phone apps for quality of life and well-being assessment in breast and prostate cancer patients: systematic review. JMIR Mhealth Uhealth 2017 Dec 04;5(12):e187 [FREE Full text] [doi: 10.2196/mhealth.8741] [Medline: 29203459]
90. Morrison D, Wyke S, Agur K, Cameron EJ, Docking RI, Mackenzie AM, et al. Digital asthma self-management interventions: a systematic review. J Med Internet Res 2014 Feb 18;16(2):e51 [FREE Full text] [doi: 10.2196/jmir.2814] [Medline: 24550161]
91. Liddy C, Drosinis P, Keely E. Electronic consultation systems: worldwide prevalence and their impact on patient care-a systematic review. Fam Pract 2016 Jun 13;33(3):274-285. [doi: 10.1093/fampra/cmw024] [Medline: 27075028]
92. Barken TL, Söderhamn U, Thygesen E. A sense of belonging: a meta-ethnography of the experience of patients with chronic obstructive pulmonary disease receiving care through telemedicine. J Adv Nurs 2019 Dec 05;75(12):3219-3230. [doi: 10.1111/jan.14117] [Medline: 31225664]
93. Cheung KL, Durusu D, Sui X, de Vries H. How recommender systems could support and enhance computer-tailored digital health programs: a scoping review. Digit Health 2019 Jan 24;5:2055207618824727 [FREE Full text] [doi: 10.1177/2055207618824727] [Medline: 30800414]
94. Werder M. Health information technology: a key ingredient of the patient experience. Patient Exp J 2015 Apr 29;2(1):143-147. [doi: 10.35680/2372-0247.1071]
95. Stokke R. The personal emergency response system as a technology innovation in primary health care services: an integrative review. J Med Internet Res 2016 Jul 14;18(7):e187 [FREE Full text] [doi: 10.2196/jmir.5727] [Medline: 27417422]
96. Chaudhry H, Nadeem S, Mundi R. How satisfied are patients and surgeons with telemedicine in orthopaedic care during the COVID-19 pandemic? A systematic review and meta-analysis. Clin Orthop Relat Res 2021 Jan 01;479(1):47-56 [FREE Full text] [doi: 10.1097/CORR.0000000000001494] [Medline: 33009231]
97. Walker RC, Tong A, Howard K, Palmer SC. Patient expectations and experiences of remote monitoring for chronic diseases: systematic review and thematic synthesis of qualitative studies. Int J Med Inform 2019 Apr;124:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.01.013] [Medline: 30784430]

XSL•FO
RenderX

98. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
99. Caulfield J. How to do thematic analysis: step-by-step guide & examples. Scribbr. 2019. URL: https://www.scribbr.com/methodology/thematic-analysis/ [accessed 2023-05-04]
100. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan 2009 May 20;24(3):217-228 [FREE Full text] [doi: 10.1093/heapol/czp010] [Medline: 19304786]
101. Kumar R. Research Methodology: A Step-By-Step Guide for Beginners. Thousand Oaks, CA: Sage Publications; 2018.
102. Quality of care. World health organization. URL: https://www.who.int/health-topics/quality-of-care#tab=tab_1 [accessed 2022-02-10]
103. Streefkerk R. Qualitative vs. quantitative research: differences, examples & methods. Scribbr. URL: https://www.scribbr.com/methodology/qualitative-quantitative-research/ [accessed 2022-04-20]
104. Cumpston M, Chandler IJIV. Updating a review. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2020.
105. Allen S. 2019 Global health care outlook: shaping the future. Deloitte. 2019. URL: https://www2.deloitte.com/tr/en/pages/life-sciences-and-healthcare/articles/2019-global-healthcare-outlook.html [accessed 2024-01-14]
106. WHO evaluation practice handbook. World Health Organization. 2013. URL: https://cdn.who.int/media/docs/default-source/evaluation-office/who-evaluation-practice-handbook-2013.pdf?sfvrsn=2513394e_3&download=true [accessed 2024-01-12]
107. Labrique A, Vasudevan L, Weiss W, Wilson K. Establishing standards to evaluate the impact of integrating digital health into health systems. Glob Health Sci Pract 2018 Oct 10;6(Supplement 1):S5-17. [doi: 10.9745/ghsp-d-18-00230] [Medline: 30305335]
108. Chung KC, Swanson JA, Schmitz D, Sullivan D, Rohrich RJ. Introducing evidence-based medicine to plastic and reconstructive surgery. Plast Reconstr Surg 2009 Apr;123(4):1385-1389 [FREE Full text] [doi: 10.1097/PRS.0b013e31819f25ff] [Medline: 19337107]
109. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg 2010 Dec;126(6):2234-2242 [FREE Full text] [doi: 10.1097/PRS.0b013e3181f44abc] [Medline: 20697313]
110. Pawloski PA, Brooks GA, Nielsen ME, Olson-Bullis BA. A systematic review of clinical decision support systems for clinical oncology practice. J Natl Compr Canc Netw 2019 Apr 01;17(4):331-338 [FREE Full text] [doi: 10.6004/jnccn.2018.7104] [Medline: 30959468]
111. Wang T, Giunti G, Melles M, Goossens R. Design-relevant factors affecting the patient experience in digital health: preliminary results of an umbrella systematic review. Stud Health Technol Inform 2022 Jun 06;290:862-866. [doi: 10.3233/SHTI220202] [Medline: 35673141]
112. Sadoyu S, Tanni KA, Punrum N, Paengtrai S, Kategaew W, Promchit N, et al. Methodological approaches for assessing certainty of the evidence in umbrella reviews: A scoping review. PLoS One 2022 Jun 8;17(6):e0269009 [FREE Full text] [doi: 10.1371/journal.pone.0269009] [Medline: 35675337]
113. Choi GJ, Kang H. Introduction to umbrella reviews as a useful evidence-based practice. J Lipid Atheroscler 2023 Jan;12(1):3-11 [FREE Full text] [doi: 10.12997/jla.2023.12.1.3] [Medline: 36761061]

Abbreviations

DH: digital health
DHI: digital health intervention
HCP: health care provider
PEx: patient experience
PREM: patient-reported experience measure
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROM: patient-reported outcome measure
RCT: randomized controlled trial
UX: user experience
WHO: World Health Organization

XSL•FO
RenderX

Edited by G Eysenbach; submitted 06.02.23; peer-reviewed by KL Mauco, T Judson; comments to author 10.04.23; revised version received 05.06.23; accepted 29.11.23; published 05.02.24.

Please cite as:
Wang T, Giunti G, Goossens R, Melles M
Timing, Indicators, and Approaches to Digital Patient Experience Evaluation: Umbrella Systematic Review
J Med Internet Res 2024;26:e46308
URL: https://www.jmir.org/2024/1/e46308
doi:10.2196/46308
PMID:38315545

©Tingting Wang, Guido Giunti, Richard Goossens, Marijke Melles. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Mapping Theories, Models, and Frameworks to Evaluate Digital Health Interventions: Scoping Review

Geneviève Rouleau[1,2,3], RN, PhD; Kelly Wu[2], MSc; Karishini Ramamoorthi[2], MSc; Cherish Boxall[4], MSc; Rebecca H Liu[2], PhD; Shelagh Maloney[5], BSc Hons; Jennifer Zelmer[6], MA, PhD; Ted Scott[7], MAppSC, PhD; Darren Larsen[8,9], MD, CCFP, MPLc; Harindra C Wijeysundera[10], MD, PhD; Daniela Ziegler[11], MSI; Sacha Bhatia[12,13], MD, MBA; Vanessa Kishimoto[2], MPH; Carolyn Steele Gray[14,15], PhD; Laura Desveaux[2,15,16], MPT, PhD

[1]Nursing department, Université du Québec en Outaouais, Saint-Jérôme, QC, Canada
[2]Institute for Health System Solutions and Virtual Care Toronto, Women's College Hospital, Toronto, ON, Canada
[3]Institut du Savoir Montfort, Montfort Hospital, Ottawa, ON, Canada
[4]Southampton Clinical Trials Unit, University of Southampton, Southampton, United Kingdom
[5]Canada Infoway, Toronto, ON, Canada
[6]Healthcare Excellence Canada, Ottawa, ON, Canada
[7]School of Nursing, Hamilton Health Sciences, McMaster University, Hamilton, ON, Canada
[8]Telus Healthcare Delivery, Women's College Hospital, Toronto, ON, Canada
[9]Women's College Hospital Family Health Team, Women's College Hospital, Toronto, ON, Canada
[10]Schulich Heart Program, Sunnybrook Health Sciences Centre, Toronto, ON, Canada
[11]Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada
[12]Ontario Health, Toronto, ON, Canada
[13]Peter Munk Cardiac Centre, University Health Network, Toronto, ON, Canada
[14]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, ON, Canada
[15]Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
[16]Institute for Better Health, Trillium Health Partners, Toronto, ON, Canada

Corresponding Author:
Geneviève Rouleau, RN, PhD
Nursing department
Université du Québec en Outaouais
5, rue Saint-Joseph
Saint-Jérôme, QC, J7Z 0B7
Canada
Phone: 1 819 595 3900
Email: genevieve.rouleau02@uqo.ca

Abstract

Background: Digital health interventions (DHIs) are a central focus of health care transformation efforts, yet their uptake in practice continues to fall short of their potential. In order to achieve their desired outcomes and impact, DHIs need to reach their target population and need to be used. Many factors can rapidly intersect between this dynamic of users and interventions. The application of theories, models, and frameworks (TMFs) can facilitate the systematic understanding and explanation of the complex interactions between users, practices, technology, and health system factors that underpin research questions. There remains a gap in our understanding of how TMFs have been applied to guide the evaluation of DHIs with real-world health system operations.

Objective: This study aims to map TMFs used in studies to guide the evaluation of DHIs. The objectives are to (1) describe the TMFs and the constructs they target, (2) identify how TMFs have been prospectively used (ie, their roles) in primary studies to evaluate DHIs, and (3) to reflect on the relevance and utility of our findings for knowledge users.

Methods: This scoping review was conducted in partnership with knowledge users using an integrated knowledge translation approach. We included papers (eg, reports; empirical quantitative, qualitative, and mixed methods studies; conference proceedings; and dissertations) if primary insights resulting from the application of TMFs were presented. Any type of DHI was eligible. Papers

XSL•FO
RenderX

published from 2000 and onward were mainly identified from the following databases: MEDLINE (Ovid), CINAHL Complete (EBSCOhost), PsycINFO (Ovid), EBM Reviews (Ovid), and Embase (Ovid).

Results: A total of 156 studies published between 2000 and 2022 were included. A total of 68 distinct TMFs were identified across 85 individual studies. In more than half (85/156, 55%) of the included studies, 1 of following 6 prevailing TMFs were reported: Consolidated Framework for Implementation Research (n=39); the Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework (n=17); the Technology of Acceptance Model (n=16); the Unified Theory on Acceptance and Use of Technology (n=12); the Diffusion of Innovation Theory (n=10); and Normalization Process Theory (n=9). The most common intended roles of the 6 TMFs were to inform data collection (n=86), to inform data analysis (n=69), and to identify key constructs that may serve as barriers and facilitators (n=52).

Conclusions: As TMFs are most often reported to be applied to support data collection and analysis, researchers should consider more clearly synthesizing key insights as practical use cases to both increase the relevance and digestibility of their findings. There is also a need to adapt or develop guidelines for better reporting DHIs and the use of TMFs to guide evaluation. Hence, it would contribute to ensuring ongoing technology transformation efforts are evidence and theory informed rather than anecdotally driven.

(J Med Internet Res 2024;26:e51098) doi:10.2196/51098

KEYWORDS

digital health interventions; evaluation; implementation; integrated knowledge translation, theories; models; frameworks; scoping review

Introduction

Background

Digital health interventions (DHIs) are a central focus of health care transformation efforts worldwide [1-4], yet their uptake in practice continues to fall short of their potential [5-8]. DHIs are complex interventions with multiple components that fulfill a range of functionalities such as supporting communication, decision-making, documentation and maintenance of patient records, diagnosis, and access to therapies. They target a range of patients, health care providers (HCPs), and health system users and are deployed in a variety of settings (eg, hospital, community, and home) [9] in hopes of delivering on the Quadruple Aim [10,11]. The Quadruple Aim is intended to improve population health, patients' and caregivers' experiences, and providers' experience and to reduce costs. To achieve the desired outcomes, the ideal first step would involve the "determination and optimisation of reach and uptake by the intended population, in the context in which the DHI will be used" [10]. In reality, and at an increasing rate, DHIs are implemented in practice in the paucity of fulsome evidence of their effect; studies being limited to pilot or feasibility ones [12]. This is partly a product of the timelines of traditional research and the rapid pace of technology progress [13]. As a result, evaluations of DHIs may seek to answer various research questions about their effectiveness, associated implementation outcomes [10,14], or both at different stages of the research cycle. For instance, it is appropriate to evaluate the feasibility of DHIs by focusing on implementation outcomes such as reach, adoption, practicability, and acceptability, as well as to determine the impacts of DHIs components on the expected outcomes [10,15]. We refer to evaluation throughout this paper in this broad sense, consistent with our previous work [16], encompassing the systematic assessment of an intervention's design, implementation, and outcomes that can judge merit, worth, or significance by combining evidence and values [17,18]. The term "evaluation" is then inclusive of various evaluation activities, different types of evaluation (eg, process evaluation, implementation evaluation, as well as impact and outcome evaluation) purposes, and research questions [17,18].

In reality, a variety of other factors influence the successful (or failed) implementation of DHIs [19], including but not limited to funding structure, policy, organizational settings, the complex interactions between users, existing routines and processes, the value proposition, and the technology itself [20]. One way of facilitating the systematic understanding and explanation of the complex interactions between users, practices, technology, and health system factors that underpin research questions [20,21] is to use theories, models, and frameworks (TMFs). There is a wide range of TMFs that have been used in studies of knowledge translation [22,23], and implementation science [24] (examples of more than 40 TMFs are cited). Heinsch et al [24] identified 36 theories for informing and explaining eHealth implementation, and Greenhalgh et al [20] identified 28 technology-specific implementation frameworks. In their systematic review, Bashi et al [12] identified 11 evaluation frameworks applied in the management of chronic diseases. The authors of those works use either the terms "implementation frameworks" or "evaluation frameworks." Despite this body of knowledge on TMFs, there remains a gap in our understanding of how TMFs have been and could be applied to guide the reported prospective evaluation of DHIs with real-world health system operations.

Objectives

The aim of this work was to map the TMFs used in studies to evaluate DHIs. Specifically, our objectives were to (1) describe the TMFs and the constructs they target, (2) identify how TMFs have been used in primary studies (hereafter referred to as the roles of the TMFs), and (3) reflect on the relevance and utility of our findings for knowledge user partners as a post hoc objective. A scoping review was the suited knowledge synthesis approach to map in a comprehensive way the current state of

XSL•FO
RenderX

evidence in this given area and to cover a breadth of the literature [25,26].

Methods

Protocol and Registration

We conducted a scoping review, informed by the Joanna Briggs Institute methodology [27]. The protocol has been published previously [28]. We used the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist to inform reporting [29].

Integrated Knowledge Translation Approach

This work was conducted in partnership with knowledge users using an integrated knowledge translation strategy [30,31]. The aim of the strategy was to inform the objectives and approach, develop a shared understanding of the findings, and work with knowledge users to understand how the resulting knowledge could be synthesized to support its application in practice. The knowledge users were identified by the leadership of the Centre for Digital Health Evaluation and, especially, by its director and scientific lead at the time of the initiation of the study. Only 1 person out of 7 declined the email invitation. The advisory panel included senior leaders (DL, HCW, SM, JZ, TS, and SB), policy makers (JZ and SB), a researcher (CSG), clinicians (DL, HCW, and SB), and a DHI developer (DL) who are involved in health decision making regarding the evaluation of DHIs. The knowledge users advisory panel provided input to the protocol of the scoping review [28], supported the refinement of the eligibility criteria of included papers, identified relevant data abstraction elements to prioritize, and assisted in the interpretation of findings. Furthermore, as a mechanism of reflexivity [32], they shared their vision and experience about using TMFs in their respective context of work. Consequently, it helped to inform the results. Details regarding knowledge users' profiles, area of expertise, and their application of TMFs are available in Multimedia Appendix 1. Two newsletters were sent out to inform knowledge users about progress updates and upcoming activities (eg, titles and abstracts screening and data extraction). Knowledge users participated in 6 meetings over Zoom (Zoom Video Communications) between March 2021 and December 2022, with each meeting lasting 60 minutes. Figure 1 illustrates the timeline of the meetings, the topics covered, and some examples of questions discussed.

Figure 1. Timeline and content of meetings with knowledge users. DHIs: digital health interventions; TMF: theory, model, and framework.

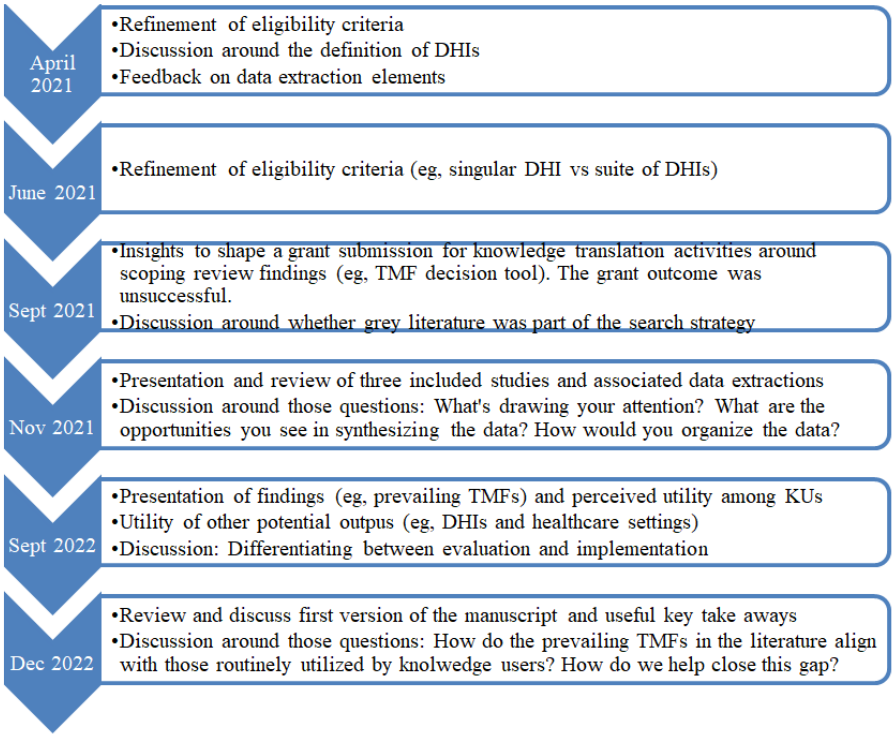

Critical Appraisal of Individual Sources of Evidence

Given that the purpose of our scoping review was to provide a broad overview of the TMFs used in relation to DHIs and not to recommend evidence to guide clinical practice, the methodological quality of included studies was not assessed; this was prespecified in the protocol [28] and was still in line with methodological guidance [27].

Eligibility Criteria

To enhance translation into practice, the eligibility criteria focused on papers in which authors reported having used TMFs prospectively to guide DHI evaluation. Retrospective application of TMFs was excluded. We were interested in understanding how TMFs can be prospectively used when undertaking a theoretical-based evaluation of DHIs.

Type of Interventions

DHIs are defined as the use of various digital technologies (eg, eHealth, telemedicine, patient remote monitoring, smartphone

XSL•FO
RenderX

apps, patient sensors, and artificial intelligence) to improve health care delivery [33]. We included papers that reported on a single DHI and excluded those in which a collection or suite of DHIs was presented (eg, variety of electronic medical records) and those in which DHIs were described broadly (eg, eHealth systems having no identified features or components). This exclusion criterion was mainly decided to narrow down the number of included studies in order to increase the feasibility of performing the knowledge synthesis with the constrained resources we had.

Type of Literature

We included peer-reviewed quantitative, qualitative, and mixed methods empirical studies reporting on the application of TMFs to prospectively guide the evaluation of DHIs in health care. Published gray literature, including conference abstracts or proceedings, dissertations, reports, and white papers, have been included if primary insights resulting from the application of TMFs were presented. Reviews, study protocols, commentaries, and letters to the editor were excluded as they had no primary data.

Type of Participants

No limitations were placed on the user population as long as the evaluation of DHIs in a health care context was described.

Type of TMFs

We included studies that pursued different research questions in which TMFs were applied to guide the evaluation of DHIs. We excluded papers that described the theoretical underpinnings and the overall process of intervention development. We made this decision at the early stage of the scoping review given the abundance of papers focusing on the intervention development; and the limited resources we had in completing the knowledge synthesis.

Information Sources and Search Strategy

The following electronic databases were searched: MEDLINE (Ovid), CINAHL Complete (EBSCOhost), PsycINFO (Ovid), EBM Reviews (Ovid), and Embase (Ovid). A complementary search of Google Scholar was conducted to identify relevant studies. Search strategies were designed by a librarian (DZ) and were peer-reviewed by another senior information specialist prior to execution using the PRESS Checklist [34]. We imposed no language restrictions and the search extended to studies published in 2000 and onward. The search was initially run on March 10, 2021, and then updated on March 16, 2022. The full electronic search strategy is presented in Multimedia Appendix 2. Due to resource constraints and delays introduced by the pandemic, we did not apply supplementary search strategies as planned in the protocol (ie, checking the reference lists of included studies, and conducting a forward citation search).

Eligibility Screening Process

Citations obtained from the literature were stored in Endnote (Version X9; Clarivate) [35] and then uploaded to Covidence (Veritas Health Information) [36], a web-based collaboration software platform that streamlines the production of systematic and other literature reviews. This software allows multiple reviewers to participate in various stages of the review (ie, screening titles, abstracts, and full texts and identifying discrepancies). We applied a 2-step process for identifying relevant citations. At stage 1, titles and abstracts were independently assessed by 5 reviewers (RHL, GR, KR, CB, and VK). Studies with abstracts fulfilling the criteria were passed to level 2 full-text screening. At this stage, each full text was reviewed by 1 person (a total of 7 team members: RHL, GR, KR, VK, SM, CB, and KW). All reviewers flagged full texts they were unsure about, and these were validated by a second reviewer. A pilot test of the screening strategy was completed using a random sample of 10% of citations and full-text papers prior to full implementation, with the expressed purpose of assessing agreement between reviewers at each level (interrater reliability ≥80% was considered adequate). When agreement was not reached, a third reviewer (RHL and GR) mediated any disagreements.

Data Extraction Process and Data Items

Studies fulfilling the eligibility criteria were extracted in Microsoft Excel. The following study characteristics were collected: reference, country of origin of the first author, study design, and DHI user. We also extracted data specific to the TMFs, including name; constructs; variables or mechanisms; and roles of framework in the study, that is, how it has been applied in research. We pilot-tested the data extraction form by extracting data from the same study as a team and iteratively adapted the form. We had biweekly working meetings to discuss the process, highlight challenges, and identify strategies to mitigate those challenges.

Data Synthesis

We undertook a descriptive quantitative and qualitative data analysis. First, we did simple frequency counts in line with Joanna Briggs Institute guidance [27] of interventions, frameworks, and roles of the TMFs. Descriptive qualitative content analysis involved categorizing 2 sets of data: DHI users and roles of frameworks. We used the World Health Organization taxonomy [37] to categorize the DHIs according to their primary targeted users: (1) clients (potential or actual users of health services, including caregivers), (2) HCPs (deliverers of health services), (3) health system managers ("involved in the administration and oversight of public health systems"), and (4) data services (crosscutting functionality supporting various activities focusing on data collection, management, use, and exchange). A single DHI could be categorized in more than 1 domain (eg, it can target both clients and HCPs). We coded TMF roles according to the use classifications outlined by Birken et al [38] (Textbox 1).

The knowledge users advisory panel informed the synthesis of findings, including the level of detail abstracted from included papers and the approach to DHI classification.

XSL•FO
RenderX

Textbox 1. Roles of theories, models, and frameworks according to the classifications by Birken et al .

- To identify key constructs that may serve as barriers and facilitators
- To inform data collection
- To guide implementation planning
- To enhance conceptual clarity
- To specify the process of implementation
- To frame an evaluation
- To inform data analysis
- To guide the selection of implementation strategies
- To specify outcomes
- To clarify terminology
- To convey the larger context of the study
- To specify hypothesized relationships between the constructs

Results

Search Results

A total of 10,567 titles or abstracts were identified from the 5 databases, Google Scholar, and other methods, from which 3192 were removed in EndNote by the librarian (DZ). After removing duplicate references, 7375 titles or abstracts were assessed for eligibility. Of these, 6561 papers were excluded based on title and abstract screening and application of the eligibility criteria previously outlined. A total of 814 full-text papers were sought and screened, and 658 were excluded. The list of excluded studies is presented in Multimedia Appendix 3. Between 2000 and 2022, a total of 156 published papers met the eligibility criteria. The list of these included papers is presented in Multimedia Appendix 4. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) study flow diagram [39] is illustrated in Figure 2 to show the overall process of review selection.

Figure 2. The PRISMA (preferred reporting items for systematic reviews and meta-analyses) study flow diagram. DHI: digital health intervention.

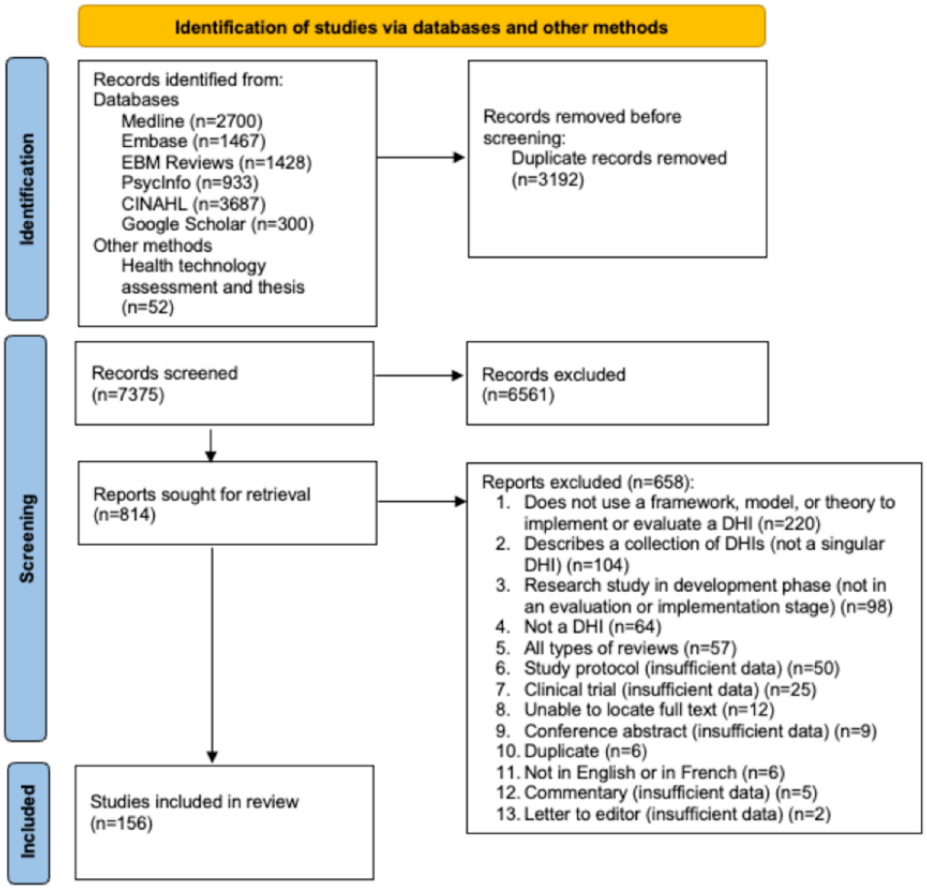

XSL•FO
RenderX

Study Characteristics

The majority of papers were published by researchers in the United States (n=64), Canada (n=19), the United Kingdom (n=14), the Netherlands (n=12), and Australia (n=7). Study designs were largely qualitative (n=63) and mixed methods (n=62), with a smaller number of quantitative studies (n=31). Most DHIs targeted either HCPs (n=67) or clients (n=63), with a few targeting health system managers (n=8), data services (n=3), and a combination of users (n=15).

Identification of Theories and the Most Reported TMFs

In total, 68 distinct TMFs were identified (see Multimedia Appendix 5) across 85 individual studies. More than half (85/156, 55%) of included studies used 1 of 6 TMFs, which included the CFIR (Consolidated Framework for Implementation Research; 39 studies), the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework; 17 studies), the TAM (Technology of Acceptance Model; 16 studies), the UTAUT (Unified Theory on Acceptance and Use of Technology; 12 studies), the DOI (Diffusion of Innovation Theory; 10 studies), and the NPT (Normalization Process Theory; 9 studies). It should be noted that the number of studies across the 6 TMFs is 103; however, because TMFs are used in combination with other theoretical approaches (Figure 3), this number represents 85 individual studies. UTAUT is the theory most frequently used in combination with other TMFs. A descriptive table of those studies that includes references, DHI user, study type, TMF used in combination, and roles of TMFs is presented in Multimedia Appendix 6. Our results focus on synthesizing insights across the 6 prevailing TMFs, being reported in 9 studies or more, allowing us to synthesize their application across different studies and contexts of evaluation. The constructs of the prevailing TMFs are described and summarized in the Multimedia Appendix 7.

The most common intended roles of the 6 TMFs were to inform data collection (n=86), to inform data analysis (n=69), to identify key constructs that may serve as barriers and facilitators (n=52), to organize and report the study findings (n=47), and to frame an evaluation (n=18; see Table 1). TMFs were applied to pursue various roles, that is, they served multiple purposes. The average number of distinct roles per TMFs is as follows: RE-AIM (n=3), CFIR (n=2.9), NPT (n=2.56), DOI (n=1.9), UTAUT (n=1.75), and TAM (n=1.69).

Figure 3. Nature of application of theories, models, and frameworks across included studies.

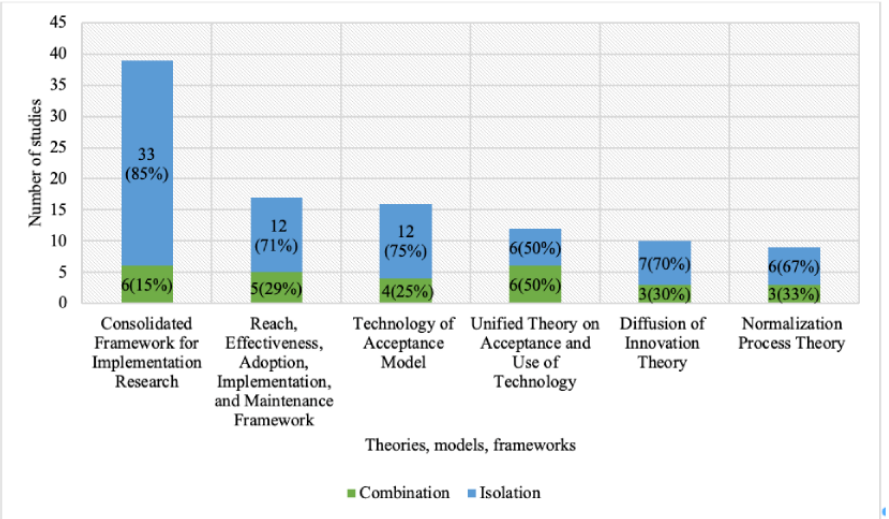

XSL•FO
RenderX

Table 1. Intended roles of TMFs[a] as specified across the studies.

Roles of TMFs	6 prevailing TMFs; number of roles identified across all studies					
	CFIR[b]	RE-AIM[c]	TAM[d]	UTAUT[e]	DOI[f]	NPT[g]
To identify barriers and facilitators	31	8	3	4	4	2
To inform data collection	25	26	14	9	3	8
To inform data analysis	32	18	7	3	3	6
To organize and report study findings	15	16	5	3	4	4
To guide implementation planning	7	2	0	0	0	1
To guide the selection of implementation strategies	3	1	0	1	1	1
To frame an evaluation	3	9	1	1	2	2
To clarify terminology	0	0	0	0	1	0
To specify the process of implementation	1	0	0	0	0	0
To specify hypothesized relationships between constructs	0	0	0	0	1	0

[a]TMF: theory, model, and framework.

[b]CFIR: Consolidated Framework for Implementation Research.

[c]RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework.

[d]TAM: Technology of Acceptance Model.

[e]UTAUT: Unified Theory on Acceptance and Use of Technology.

[f]DOI: Diffusion of Innovation Theory.

[g]NPT: Normalization Process Theory.

DHIs and Intended Users Associated With Top 6 Frameworks

DHIs targeting clients and HCPs are the most frequently reported (see Table 2). RE-AIM and DOI were used for DHIs designed for clients, while CFIR, TAM, UTAUT, and NPT have been used primarily with DHIs involving HCPs. DHIs targeting clients included patient portals [40,41], web-based self-management interventions [42,43], and mobile health diet apps [44]. DHIs targeting HCPs included (but are not limited to) mobile apps targeting patients' smoking cessation [45] and medication adherence counseling [46,47], telemedicine and telehealth [48,49], cancer prevention decision support tools [50], e-consultation between primary care providers and specialty care expertise [51,52], and e-learning for dementia caregiver education [53]. Two studies targeted health system managers and an information system for case-based surveillance [54] and a patient-reported outcome data collection system [55]. Two studies focused on data services including big data analytics [56], the former being reported in combination with clients, and an electronic patient falls reporting system [57].

XSL•FO
RenderX

Table 2. Target users of the DHI[a] across included studies.

Framework	Clients, n	HCPs[b], n	Health system managers, n	Data services, n	Combination, n
CFIR[c]	13	18	2	0	5
RE-AIM[d]	9	3	0	0	5
TAM[e]	3	9	2	1	1
UTAUT[f]	5	7	0	0	0
NPT[g]	3	5	0	0	1
DOI[h]	5	3	0	0	2

[a]DHI: digital health intervention.
[b]HCPs: health care providers.
[c]CFIR: Consolidated Framework for Implementation Research.
[d]RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework.
[e]TAM: Technology of Acceptance Model.
[f]UTAUT: Unified Theory on Acceptance and Use of Technology.
[g]NPT: Normalization Process Theory.
[h]DOI: Diffusion of Innovation Theory.

Current Gaps Between Prevailing TMFs Used in Research and in Practice by Knowledge Users

The knowledge users reflected that most of the prevailing TMFs identified in this scoping review were not familiar to them. They have used different TMFs in practice (see Multimedia Appendix 1) that reflect their interest in capturing the process of implementation and for outcome-driven evaluation approaches that would help them understand whether DHIs work or not. One first example is the NPT, applied to implement and evaluate the effectiveness of the electronic patient-reported outcome (ePRO) mobile app and portal system. ePRO was designed to enable goal-oriented care delivery in interprofessional primary care practices [58]. In this study, many types of outcomes were of interest to produce early evidence of effectiveness (or ineffectiveness) of the ePRO and its mechanisms of action: the context (eg, sociodemographic data and barriers of adopting ePRO), process (eg, usability), and outcome measures (eg, patients' quality of life, provider-level effectiveness in delivering care to patients with chronic illness). The Quadruple Aim [59] was also used by 5 knowledge users to evaluate the impact of DHIs on health system performance with outcomes such as equitable access, cost reduction, patient-provider relationships, providers' burnout, and work-life balance [11]. The Benefits Evaluation Framework [60] was also well-known and used by knowledge users. Similarly to TAM and UTAUT, it aims to describe factors influencing eHealth success (eg, system quality, information quality, and user satisfaction), with the addition of the resulting impacts (or outcomes) of DHIs in terms of care quality (eg, effectiveness and health outcomes), access services, and productivity (eg, efficiency).

Discussion

Principal Findings

While a wide range of TMFs (n=68) have been used to guide the evaluation of DHIs, 6 main TMFs are used consistently by researchers. These TMFs were used in a variety of roles and were broadly applied across types of DHI and target user groups, demonstrating their flexibility in academic practice. These 6 TMFs were not commonly used by nor familiar to many of the knowledge users, highlighting the disconnect between academic and health system practice. Our discussion presents the 3 key insights from these conversations in relation to our results: specifically, how the application of prevailing TMFs in the literature could be used in health system decision-making, how to bridge the persistent gap between academic knowledge and health system practice, and lessons learned about how future work might bridge this gap.

Insights From the Application of Prevailing TMFs

The findings allowed us to identify a higher number of TMFs (n=68) than those reported previously by Heinsch et al [24] (n=36) and Greenhalgh et al [20] (n=28). Our findings corroborate the ones in Heinsh et al [24], in which 5 of our prevailing TMFs (except RE-AIM) have been identified. Furthermore, at least 3 TMFs (TAM, DOI, and NPT) identified in our review were cited as a groundwork for technology implementation frameworks as identified by Greenhalgh et al [20]. While this highlights the variability of TMFs used in the evaluation process of DHIs, they are most often used to inform data collection and data analysis, aligning with the findings of Birken et al [38].

While different types of outcomes were of interest to knowledge users deriving from their use of TMFs (see Multimedia Appendix 1), that is, service (eg, efficiency or cost, effectiveness, and access to care), client (eg, patients' quality

XSL•FO
RenderX

of life), and implementation (eg, adoption and sustainability), as aligned with the literature [15], only RE-AIM included an explicit effectiveness outcome domain. Recently, CFIR has been extended to include implementation outcomes and innovation (ie, intervention) outcomes as part of CFIR 2.0 [61,62]. CFIR 2.0 outcomes are inclusive of both purchase and operating costs (the innovation cost) [61]—an element that is central to decision-making within resource-constrained systems [63]. As an example, the perceived advantage of a mobile app from the perspective of HCPs (an implementation determinant) may impact their uptake and referral rate to their patients (an implementation outcome). This is distinct from patient motivation to use the app (an innovation determinant) which will impact weight loss (an innovation outcome). These distinct categories were included to focus attention "squarely on the way that context shapes intermediate results and conditions, such as user acceptance, which in turn influence classic measures of an intervention's ultimate aims or outcomes" [64]. This highlights the need to consider a chain of short-term proxy outcomes (eg, acceptability of DHIs and adaptation to novel contexts), including the attributes of context [65], if we want to capture the likely benefit of the DHIs [10]. This would help to address the disconnect between (less) attention paid to context in comparison with effectiveness outcomes [10,65].

Bridging the Gap Between Academic and Health System Practice

Our work echoes the opportunity for researchers to better understand the realities of health care practice and operations [66]. This can be achieved by understanding the context in which knowledge users operate, their values and professional experience from the early beginning of the project [67,68], and assessing the usefulness of TMFs in supporting their routine decision-making. Relatedly, there is an opportunity for researchers to support knowledge users in understanding how to leverage insights from the literature to better achieve their desired outcomes. Greenhalgh et al [20] observed a tendency across DHI implementations "to assume the issues to be addressed were simple or complicated (hence knowable, predictable, and controllable) rather than complex (that is, inherently not knowable or predictable but dynamic and emergent)." Explicitly highlighting how TMFs can mitigate the inherent challenges that knowledge users face in evaluating interventions may help to address this gap.

Our knowledge user panel recommended presenting the TMFs as practical use cases to illustrate their real-world application potential—specifically in nonacademic settings, where implementation and evaluation activities are part of routine operations. The knowledge users shared an interest in TMFs that could support scaling up DHIs and understanding their effectiveness and impact on patient outcomes. Through our discussion with the advisory panel, the CFIR and RE-AIM frameworks were identified as aligning with the dual purpose of guiding the implementation effectiveness (CFIR), as well as scalability and sustainability (RE-AIM) [69]. The use cases were constructed to highlight the utility of the TMFs as well as to demonstrate how and to what end they have been used in research (Multimedia Appendices 8 and 9).

Lessons Learned and Limitations

A natural evolution of this work would be to provide knowledge users with an easy-to-use tool to select a TMF that aligns with their operational needs and local context. The Theory, Model, and Framework Comparison and Selection Tool [70] can help scientists and practitioners select the most appropriate TMF to meet their needs and realize the potential that a given DHI may, or may not, bring in its intended context. Users are directed to a web tool and repository of TMFs [71] which includes a tutorial for novice users and guidance on how to address their research or practice questions.

Despite the desire among our team to classify DHIs according to their primary function (eg, to communicate with clients and to transmit information), we were constrained by the variability in how DHI-related information was reported. A standardized reporting structure inclusive of DHI function, setting, target users, and intended outcomes would help to facilitate learning across systems and studies as health care becomes increasingly technology enabled. Guideline for reporting evidence-based practice educational interventions and teaching [72] and the Template for Intervention Description and Replication checklist [73] are both used to better report interventions. However, adapting those guidelines to the specificities of DHIs would be valuable in future work. In addition, Krick et al [74] developed a comprehensive digital nursing technology outcome framework that allows the identification of effective outcomes. This outcome framework can be a good starting from which other types of outcomes (such as implementation) can be added. This would potentially address the desire of knowledge users to understand effectiveness outcomes at a categorical level (eg, the effectiveness of DHIs by functional category or setting), which we were unable to achieve due to the heterogeneity of outcomes and terminology used to describe DHIs.

While this work engaged knowledge users from the study conception, our search strategy did not capture the Benefits Evaluation Framework—the primary framework they used in practice despite its application in more than 50 organizationally-led evaluations [75]. This limited our ability to systematically compare TMFs routinely used in the academic literature with those routinely used in practice, which is likely to provide further insights into how knowledge users collect, synthesize, report, and digest evaluation insights. Future integrated knowledge translation projects would benefit from investing time upfront to better understand how knowledge users and team members approach their work and which resources and tools they rely on to ensure research is better positioned to address persistent gaps between academic knowledge to operational practice. Another limitation in the process is that we used the information as reported by the authors to classify the roles of TMFs. We did not interpret the various roles, such as "to convey the larger context of the study" or "to frame an evaluation." Simply put, if authors did not clearly report their intended purposes for using TMFs, we did not extract the information. Hence, we did not explore to what extent the claimed theory was used. Birken et al [38] highlight that providing guidance for theory selection may encourage implementation scientists to use theories in a meaningful way and discourage superficial use and misuse. Our findings pointed

XSL•FO
RenderX

out that the prevailing TMFs were used in combination with other TMFs: adding to the challenges of aligning and using meaningfully the use of multiple TMFs. Reporting guidelines for the use of TMFs to guide evaluation would be an avenue for future research.

Conclusions

The findings of this scoping review illustrate the range of TMFs applied to support the evaluation of a breadth of DHIs. As TMFs are most often applied to support data collection and analysis, researchers should consider more clearly synthesizing key insights as practical use cases to both increase the relevance and digestibility of their findings. The opportunity to develop a standardized reporting structure inclusive of DHI function, setting, target users, and intended outcomes is quickly becoming a crucial need to ensure ongoing technology transformation efforts are evidence informed rather than anecdotally driven. Finally, guidance on how to effectively report the use of TMFs to guide evaluation would also be needed.

Acknowledgments

This work was supported by the Centre for Digital Health Evaluation and by research funds of LD. GR was granted a Fonds de recherche du Québec Santé postdoctoral scholarship at the time of this review (#302821). GR had protected time to lead and work on that scoping review. The authors would like to acknowledge the work of those people who took part in the activities at different stages of the project: Simon Minich, Abigail Appiahene-Affriyie, Marlena Dang Nguyen, Meagan Lacroix, and Julie Vo.

The publication costs for this article have been covered by Canada Health Infoway Inc, a non-profit corporation funded by the Government of Canada. The funders had no role in study design, data collection, and analysis, or decision to publish the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Knowledge users profiles and their use of theories, models, and frameworks.
[DOCX File , 25 KB - jmir_v26i1e51098_app1.docx]

Multimedia Appendix 2
Full search strategy.
[DOCX File , 37 KB - jmir_v26i1e51098_app2.docx]

Multimedia Appendix 3
List of excluded papers.
[DOCX File , 137 KB - jmir_v26i1e51098_app3.docx]

Multimedia Appendix 4
List of included studies.
[DOCX File , 35 KB - jmir_v26i1e51098_app4.docx]

Multimedia Appendix 5
List of all theories, models and frameworks.
[DOCX File , 28 KB - jmir_v26i1e51098_app5.docx]

Multimedia Appendix 6
Descriptive table most prevailing theories, models and frameworks.
[DOCX File , 51 KB - jmir_v26i1e51098_app6.docx]

Multimedia Appendix 7
Constructs of prevailing theories, models, and frameworks.
[DOCX File , 24 KB - jmir_v26i1e51098_app7.docx]

Multimedia Appendix 8
Use cases of the Consolidated Framework for Implementation Research.
[DOCX File , 22 KB - jmir_v26i1e51098_app8.docx]

XSL•FO
RenderX

Multimedia Appendix 9
Use cases of the Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework.
[DOCX File , 19 KB - jmir_v26i1e51098_app9.docx]

Multimedia Appendix 10
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 113 KB - jmir_v26i1e51098_app10.pdf]

References

1. Notice: Health Canada's approach to digital health technologies. Government of Canada. 2018. URL: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/activities/announcements/notice-digital-health-technologies.html [accessed 2021-10-02]
2. Delivering 21st century IT support for the NHS. Department of Health. United Kingdom: Department of Health; 2002. URL: https://webarchive.nationalarchives.gov.uk/ukgwa/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4067112.pdf [accessed 2024-01-03]
3. Walsh M, Chipperfield A. Australian national telehealth think tank. J Telemed Telecare 2000;6(6):353 [FREE Full text] [doi: 10.1258/1357633001935978] [Medline: 11265107]
4. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2021-10-02]
5. de Bont A, Bal R. Telemedicine in interdisciplinary work practices: on an IT system that met the criteria for success set out by its sponsors, yet failed to become part of every-day clinical routines. BMC Med Inform Decis Mak 2008;8:47 [FREE Full text] [doi: 10.1186/1472-6947-8-47] [Medline: 18954428]
6. Kuipers P, Humphreys JS, Wakerman J, Wells R, Jones J, Entwistle P. Collaborative review of pilot projects to inform policy: a methodological remedy for pilotitis? Aust N Z Health Policy 2008;5:17 [FREE Full text] [doi: 10.1186/1743-8462-5-17] [Medline: 18638412]
7. Wyatt JC, Sullivan F. eHealth and the future: promise or peril? BMJ 2005;331(7529):1391-1393 [FREE Full text] [doi: 10.1136/bmj.331.7529.1391] [Medline: 16339252]
8. Sundin P, Callan J, Mehta K. Why do entrepreneurial mHealth ventures in the developing world fail to scale? J Med Eng Technol 2016;40(7-8):444-457. [doi: 10.1080/03091902.2016.1213901] [Medline: 27686003]
9. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019;19(1):400 [FREE Full text] [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
10. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
11. Bhatti S, Dahrouge S, Muldoon L, Rayner J. Using the quadruple aim to understand the impact of virtual delivery of care within Ontario community health centres: a qualitative study. BJGP Open 2022;6(4):BJGPO.2022.0031 [FREE Full text] [doi: 10.3399/BJGPO.2022.0031] [Medline: 36109022]
12. Bashi N, Fatehi F, Mosadeghi-Nik M, Askari MS, Karunanithi M. Digital health interventions for chronic diseases: a scoping review of evaluation frameworks. BMJ Health Care Inform 2020;27(1):e100066 [FREE Full text] [doi: 10.1136/bmjhci-2019-100066] [Medline: 32156751]
13. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
14. Hrynyschyn R, Prediger C, Stock C, Helmer SM. Evaluation methods applied to digital health interventions: what is being used beyond randomised controlled trials?-a scoping review. Int J Environ Res Public Health 2022;19(9):5221 [FREE Full text] [doi: 10.3390/ijerph19095221] [Medline: 35564616]
15. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
16. Brual J, Rouleau G, Fleury C, Strom M, Koshy M, Rios P, et al. The Pan-Canadian digital health evaluation framework and toolkit: final report (version 1.0). Canadian Network for Digital Health Evaluation. 2022. URL: https://cndhe.womenscollegehospital.ca/network/conceptual-model/ [accessed 2024-01-03]
17. What is evaluation? Better Evaluation. 2022. URL: https://www.betterevaluation.org/getting-started/what-evaluation [accessed 2023-11-03]
18. Treasury HM. The Green Book: Central Government Guidance on Appraisal and Evaluation. UK: OGL Press; 2022.

XSL•FO
RenderX

19. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
20. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
21. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decision Sci 2008;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
22. Colquhoun HL, Letts LJ, Law MC, MacDermid JC, Missiuna CA. A scoping review of the use of theory in studies of knowledge translation. Can J Occup Ther 2010;77(5):270-279. [doi: 10.2182/cjot.2010.77.5.3] [Medline: 21268509]
23. Esmail R, Hanson HM, Holroyd-Leduc J, Brown S, Strifler L, Straus SE, et al. A scoping review of full-spectrum knowledge translation theories, models, and frameworks. Implement Sci 2020;15(1):11 [FREE Full text] [doi: 10.1186/s13012-020-0964-5] [Medline: 32059738]
24. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021;23(5):e18500 [FREE Full text] [doi: 10.2196/18500] [Medline: 34057427]
25. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
26. Munn Z, Pollock D, Khalil H, Alexander L, Mclnerney P, Godfrey CM, et al. What are scoping reviews? providing a formal definition of scoping reviews as a type of evidence synthesis. JBI Evid Synth 2022;20(4):950-952 [FREE Full text] [doi: 10.11124/JBIES-21-00483] [Medline: 35249995]
27. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126 [FREE Full text] [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
28. Soobiah C, Cooper M, Kishimoto V, Bhatia RS, Scott T, Maloney S, et al. Identifying optimal frameworks to implement or evaluate digital health interventions: a scoping review protocol. BMJ Open 2020;10(8):e037643 [FREE Full text] [doi: 10.1136/bmjopen-2020-037643] [Medline: 32792444]
29. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
30. A guide to researcher and knowledge-user collaboration in health research. Canadian Institutes of Health Research. 2012. URL: https://cihr-irsc.gc.ca/e/44954.html [accessed 2022-11-18]
31. Guide to knowledge translation planning at CIHR: integrated and end-of-grant approaches. Canadian Institutes of Health Research. 2012. URL: https://cihr-irsc.gc.ca/e/45321.html#a3 [accessed 2022-11-18]
32. Alley S, Jackson SF, Shakya YB. Reflexivity: a methodological tool in the knowledge translation process? Health Promot Pract 2015;16(3):426-431 [FREE Full text] [doi: 10.1177/1524839914568344] [Medline: 25631521]
33. Snowdon A. Digital health: a framework for healthcare transformation. Healthcare Information and Management Systems Society. 2020. URL: https://www.himss.org/resources/digital-health-framework-healthcare-transformation-white-paper [accessed 2021-08-30]
34. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
35. EndNote Version X9. Philadelphia, PA: Clarivate Version; 2013. URL: https://support.clarivate.com/Endnote/s/article/Citing-the-EndNote-program-as-a-reference?language=en_US [accessed 2024-01-03]
36. Covidence. Veritas Health Innovation. Melbourne, Australia; 2022. URL: https://www.covidence.org/ [accessed 2024-01-03]
37. WHO | Classification of digital health interventions v1.0. World Health Organization.: WHO. World Health Organization URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2021-10-02]
38. Birken SA, Powell BJ, Shea CM, Haines ER, Kirk MA, Leeman J, et al. Criteria for selecting implementation science theories and frameworks: results from an international survey. Implement Sci 2017;12(1):124 [FREE Full text] [doi: 10.1186/s13012-017-0656-y] [Medline: 29084566]
39. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
40. Emani S, Peters E, Desai S, Karson AS, Lipsitz SR, LaRocca R, et al. Who adopts a patient portal?: an application of the diffusion of innovation model. J Innov Health Inform 2018;25(3):149-157 [FREE Full text] [doi: 10.14236/jhi.v25i3.991] [Medline: 30398458]
41. Lalitaphanit K, Theeraroungchaisri A. Factors affecting community pharmacy customers' decision to use personal health records via smartphone. TJPS 2016;40(Supplement Issue):163-167 [FREE Full text]

XSL•FO
RenderX

42. Knoerl R, Dudley WN, Smith G, Bridges C, Kanzawa-Lee G, Smith EML. Pilot testing a web-based system for the assessment and management of chemotherapy-induced peripheral neuropathy. Comput Inform Nurs 2017;35(4):201-211. [doi: 10.1097/CIN.0000000000000320] [Medline: 28002115]
43. Myall M, May CR, Grimmett C, May CM, Calman L, Richardson A, et al. RESTORE: an exploratory trial of a web-based intervention to enhance self-management of cancer-related fatigue: findings from a qualitative process evaluation. BMC Med Inform Decis Mak 2015;15:94 [FREE Full text] [doi: 10.1186/s12911-015-0214-y] [Medline: 26577690]
44. Akdur G, Aydin MN, Akdur G. Adoption of mobile health apps in dietetic practice: case study of diyetkolik. JMIR Mhealth Uhealth 2020;8(10):e16911 [FREE Full text] [doi: 10.2196/16911] [Medline: 33006566]
45. Meijer E, Korst JS, Oosting KG, Heemskerk E, Hermsen S, Willemsen MC, et al. "At least someone thinks i'm doing well": a real-world evaluation of the quit-smoking app StopCoach for lower socio-economic status smokers. Addict Sci Clin Pract 2021;16(1):48 [FREE Full text] [doi: 10.1186/s13722-021-00255-5] [Medline: 34321088]
46. Bardosh KL, Murray M, Khaemba AM, Smillie K, Lester R. Operationalizing mHealth to improve patient care: a qualitative implementation science evaluation of the WelTel texting intervention in Canada and Kenya. Global Health 2017;13(1):87 [FREE Full text] [doi: 10.1186/s12992-017-0311-z] [Medline: 29208026]
47. McCreesh-Toselli S, Torline J, Gouse H, Robbins RN, Mellins CA, Remien RH, et al. Staff perceptions of preimplementation barriers and facilitators to a mobile health antiretroviral therapy adherence counseling intervention in South Africa: qualitative study. JMIR Mhealth Uhealth 2021;9(4):e23280 [FREE Full text] [doi: 10.2196/23280] [Medline: 33821806]
48. Hale-Gallardo JL, Kreider CM, Jia H, Castaneda G, Freytes IM, Ripley DCC, et al. Telerehabilitation for rural veterans: a qualitative assessment of barriers and facilitators to implementation. J Multidiscip Healthc 2020;13:559-570 [FREE Full text] [doi: 10.2147/JMDH.S247267] [Medline: 32669850]
49. Peracca SB, Jackson GL, Lamkin RP, Mohr DC, Zhao M, Lachica O, et al. Implementing teledermatology for rural veterans: an evaluation using the RE-AIM framework. Telemed J E Health 2021;27(2):218-226. [doi: 10.1089/tmj.2020.0013] [Medline: 32343924]
50. Harry ML, Truitt AR, Saman DM, Henzler-Buckingham HA, Allen CI, Walton KM, et al. Barriers and facilitators to implementing cancer prevention clinical decision support in primary care: a qualitative study. BMC Health Serv Res 2019;19(1):534 [FREE Full text] [doi: 10.1186/s12913-019-4326-4] [Medline: 31366355]
51. Haverhals LM, Sayre G, Helfrich CD, Battaglia C, Aron D, Stevenson LD, et al. E-consult implementation: lessons learned using consolidated framework for implementation research. Am J Manag Care 2015;21(12):e640-e647 [FREE Full text] [Medline: 26760426]
52. Liddy C, Bello A, Cook J, Drimer N, Pilon MD, Farrell G, et al. Supporting the spread and scale-up of electronic consultation across Canada: cross-sectional analysis. BMJ Open 2019;9(5):e028888 [FREE Full text] [doi: 10.1136/bmjopen-2018-028888] [Medline: 31152043]
53. Levinson AJ, Ayers S, Butler L, Papaioannou A, Marr S, Sztramko R. Barriers and facilitators to implementing web-based dementia caregiver education from the clinician's perspective: qualitative study. JMIR Aging 2020;3(2):e21264 [FREE Full text] [doi: 10.2196/21264] [Medline: 33006563]
54. Ndlovu K, Mauco KL, Keetile M, Kadimo K, Senyatso RY, Ntebela D, et al. Acceptance of the district health information system version 2 platform for malaria case-based surveillance by health care workers in Botswana: web-based survey. JMIR Form Res 2022;6(3):e32722 [FREE Full text] [doi: 10.2196/32722] [Medline: 35289760]
55. Adeoye-Olatunde OA, Curran GM, Jaynes HA, Hillman LA, Sangasubana N, Chewning BA, et al. Preparing for the spread of Patient-Reported Outcome (PRO) data collection from primary care to community pharmacy: a mixed-methods study. Implement Sci Commun 2022;3(1):29 [FREE Full text] [doi: 10.1186/s43058-022-00277-3] [Medline: 35287764]
56. Wang SL, Lin HI. Integrating TTF and IDT to evaluate user intention of big data analytics in mobile cloud healthcare system. Behav Inf Technol 2019;38(9):974-985. [doi: 10.1080/0144929x.2019.1626486]
57. Mei YY, Marquard J, Jacelon C, DeFeo AL. Designing and evaluating an electronic patient falls reporting system: perspectives for the implementation of health information technology in long-term residential care facilities. Int J Med Inform 2013;82(11):e294-e306. [doi: 10.1016/j.ijmedinf.2011.03.008] [Medline: 21482183]
58. Gray CS, Chau E, Tahsin F, Harvey S, Loganathan M, McKinstry B, et al. Assessing the implementation and effectiveness of the electronic patient-reported outcome tool for older adults with complex care needs: mixed methods study. J Med Internet Res 2021;23(12):e29071 [FREE Full text] [doi: 10.2196/29071] [Medline: 34860675]
59. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12(6):573-576 [FREE Full text] [doi: 10.1370/afm.1713] [Medline: 25384822]
60. Lau F, Hagens S, Muttitt S. A proposed benefits evaluation framework for health information systems in Canada. Healthc Q 2007;10(1):112-116. [Medline: 17326376]
61. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR outcomes addendum. Implement Sci 2022;17(1):7 [FREE Full text] [doi: 10.1186/s13012-021-01181-5] [Medline: 35065675]
62. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]

XSL•FO
RenderX

63. Schünemann HJ, Reinap M, Piggott T, Laidmäe E, Köhler K, Põld M, et al. The ecosystem of health decision making: from fragmentation to synergy. Lancet Public Health 2022;7(4):e378-e390 [FREE Full text] [doi: 10.1016/S2468-2667(22)00057-3] [Medline: 35366410]
64. Hung D, Gray C, Martinez M, Schmittdiel J, Harrison MI. Acceptance of lean redesigns in primary care: a contextual analysis. Health Care Manage Rev 2017;42(3):203-212. [doi: 10.1097/HMR.0000000000000106] [Medline: 26939032]
65. Squires JE, Aloisio LD, Grimshaw JM, Bashir K, Dorrance K, Coughlin M, et al. Attributes of context relevant to healthcare professionals' use of research evidence in clinical practice: a multi-study analysis. Implement Sci 2019;14(1):52 [FREE Full text] [doi: 10.1186/s13012-019-0900-8] [Medline: 31113449]
66. Wensing M, Grol R. Knowledge translation in health: how implementation science could contribute more. BMC Med 2019;17(1):88 [FREE Full text] [doi: 10.1186/s12916-019-1322-9] [Medline: 31064388]
67. Deverka PA, Lavallee DC, Desai PJ, Esmail LC, Ramsey SD, Veenstra DL, et al. Stakeholder participation in comparative effectiveness research: defining a framework for effective engagement. J Comp Eff Res 2012;1(2):181-194 [FREE Full text] [doi: 10.2217/cer.12.7] [Medline: 22707880]
68. Haddaway NR, Kohl C, Rebelo da Silva N, Schiemann J, Spök A, Stewart R, et al. A framework for stakeholder engagement during systematic reviews and maps in environmental management. Environ Evid 2017;6(1):11 [FREE Full text] [doi: 10.1186/s13750-017-0089-8]
69. Rangachari P, Mushiana SS, Herbert K. A scoping review of applications of the Consolidated Framework for Implementation Research (CFIR) to telehealth service implementation initiatives. BMC Health Serv Res 2022;22(1):1450 [FREE Full text] [doi: 10.1186/s12913-022-08871-w] [Medline: 36447279]
70. Theory, model, and framework comparison and selection tool (T-CaST). ImpSciX. Birken SA: The North Carolina Translational and Clinical Sciences (NC TraCS) Institute at The University of North Carolina at Chapel Hill C; 2023. URL: https://impsci.tracs.unc.edu/tcast/ [accessed 2023-02-16]
71. University of Colorado. URL: https://dissemination-implementation.org/tool/ [accessed 2023-02-16]
72. Phillips AC, Lewis LK, McEvoy MP, Galipeau J, Glasziou P, Moher D, et al. Development and validation of the guideline for reporting evidence-based practice educational interventions and teaching (GREET). BMC Med Educ 2016;16(1):237 [FREE Full text] [doi: 10.1186/s12909-016-0759-1] [Medline: 27599967]
73. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
74. Krick T, Huter K, Seibert K, Domhoff D, Wolf-Ostermann K. Measuring the effectiveness of digital nursing technologies: development of a comprehensive digital nursing technology outcome framework based on a scoping review. BMC Health Serv Res 2020;20(1):243 [FREE Full text] [doi: 10.1186/s12913-020-05106-8] [Medline: 32209099]
75. Lau F, Kuziemsky C, editors. Handbook of eHealth Evaluation: An Evidence-based Approach. Victoria (BC): University of Victoria; 2017.

Abbreviations

CFIR: Consolidated Framework for Implementation Research
DHI: digital health intervention
DOI: Diffusion of Innovation Theory
ePRO: electronic patient-reported outcome
HCP: health care provider
NPT: Normalization Process Theory
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework
TAM: Technology of Acceptance Model
TMF: theory, model, and framework
UTAUT: Unified Theory on Acceptance and Use of Technology

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.07.23; peer-reviewed by A Finucane, T Rapley; comments to author 18.09.23; revised version received 10.11.23; accepted 27.12.23; published 05.02.24.

Please cite as:
Rouleau G, Wu K, Ramamoorthi K, Boxall C, Liu RH, Maloney S, Zelmer J, Scott T, Larsen D, Wijeysundera HC, Ziegler D, Bhatia S, Kishimoto V, Steele Gray C, Desveaux L
Mapping Theories, Models, and Frameworks to Evaluate Digital Health Interventions: Scoping Review
J Med Internet Res 2024;26:e51098
URL: https://www.jmir.org/2024/1/e51098
doi:10.2196/51098
PMID:38315515

©Geneviève Rouleau, Kelly Wu, Karishini Ramamoorthi, Cherish Boxall, Rebecca H Liu, Shelagh Maloney, Jennifer Zelmer, Ted Scott, Darren Larsen, Harindra C Wijeysundera, Daniela Ziegler, Sacha Bhatia, Vanessa Kishimoto, Carolyn Steele Gray, Laura Desveaux. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Implementation of Federated Digital Identifiers in Health Care: Rapid Review

Karishini Ramamoorthi[1], MSc; Vess Stamenova[1], PhD; Rebecca H Liu[1], PhD; Onil Bhattacharyya[1,2], MD, PhD

[1]Institute for Health System Solutions and Virtual Care, Women's College Hospital, Toronto, ON, Canada

[2]Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Karishini Ramamoorthi, MSc
Institute for Health System Solutions and Virtual Care
Women's College Hospital
76 Grenville St
Toronto, ON, M5S 1B2
Canada
Phone: 1 4163236400
Email: karishini.ramamoorthi@wchospital.ca

Abstract

Background: Federated digital identifiers (FDIs) have been cited to improve the interoperability of data and information management while enhancing the privacy of individuals verifying their identity on the web. Many countries around the world have implemented FDIs in various sectors, such as banking and government. Similarly, FDIs could improve the experience for those wanting to access their health care information; however, they have only been introduced in a few jurisdictions around the world, and their impact remains unclear.

Objective: The main objective of this environmental scan was to describe how FDIs have been established and implemented to enable patients' access to health care.

Methods: We conducted this study in 2 stages, with the primary stage being a rapid review, which was supplemented by a targeted gray literature search. Specifically, the rapid review was conducted through a database search of MEDLINE and Embase, which generated a list of countries and their services that use FDIs in health care. This list was then used to conduct a targeted gray literature search using the Google search engine.

Results: A total of 93 references from the database and targeted Google searches were included in this rapid review. FDIs were implemented in health care in 11 countries (Australia, Belgium, Canada, Denmark, Estonia, Finland, Iceland, Norway, Singapore, Sweden, and Taiwan) and exclusively used with a patient-accessible electronic health record system through a single sign-on interface. The most common FDIs were implemented nationally or provincially, and establishing them usually required individuals to visit a bank or government office in person. In contrast, some countries, such as Australia, allow individuals to verify their identities entirely on the web. We found that despite the potential of FDIs for use in health care to facilitate the amalgamation of health information from different data sources into one platform, the adoption of most health care services that use FDIs remained below 30%. The exception to this was Australia, which had an adoption rate of 90%, which could be correlated with the fact that it leveraged an opt-out consent model.

Conclusions: This rapid review highlights key features of FDIs across regions and elements associated with higher adoption of the patient-accessible electronic health record systems that use them, like opt-out registration. Although FDIs have been reported to facilitate the collation of data from multiple sources through a single sign-on interface, there is little information on their impact on care or patient experience. If FDIs are used to their fullest potential and implemented across sectors, adoption rates within health care may also improve.

(J Med Internet Res 2024;26:e45751) doi:10.2196/45751

KEYWORDS

digital identity; electronic health record; environmental scan; identity management; identity verification; national electronic health record; online access; PAEHR; patient records; patient-accessible electronic health records

XSL•FO
RenderX

Introduction

Overview

Medical information is increasingly available to users through the digitization of health care [1]. This is reshaping how patients interact with the health care system by facilitating information exchange with providers and institutions. Today, patients who receive care from different organizations access their personal health information through multiple patient portals, and this information is stored separately without one place to organize or process the information [2]. Contributing to this fragmented system is the use of decentralized identities, which require patients to create, verify, and remember multiple usernames and passwords for the services owned and controlled by multiple distinct entities [3]. Although having unique electronic credentials may be perceived by some as being more secure, users are prone to creating the same logins and passwords for multiple applications in a "do-it-yourself" password management strategy, which negates the perceived increase in security [4]. This ultimately impacts the quality of the user experience. To mitigate this issue, federated identity management solutions are being introduced into health care from other sectors to improve the interoperability of health care data among institutions [3]. These solutions use federated digital identifiers (FDIs) as part of the identity proofing process, which involves binding a user with their credentials (such as a driver's license) and using that information for the authentication process [4]. Authentication is performed by a trusted identity provider, who then uses the identity proof to verify that the user is who they claim to be. Once the user has been authenticated, a relying party authorizes the user to access their services [5]. FDIs can reduce repetitive logins and the need to remember multiple passwords for patients, which reduces the number of siloed systems and facilitates a continuum of care that enables improved health care decision-making [4]. In fact, according to the Digital Identity and Authentication Council of Canada, interoperable health management systems that maintain patient privacy and autonomy can be achieved by using FDIs [6]. A good FDI permits authentication of a person's identity and has the following features: (1) It needs to be verified and authenticated with high assurance (ie, the initial registration process is accomplished with high standards), (2) it needs to be unique such that each individual only has 1 identity within a system, (3) it must be established with an individual's consent so that users are aware of what personal data is being shared, and (4) it must protect the user's privacy and allow them to control how they use their personal data [6,7]. Some researchers have even said governments have a formal responsibility to ensure that the digital identity infrastructure will not result in the disempowerment of individual citizens [8]. Considering the boom in the digital identity market, developing trustworthy processes and infrastructure is key in order to avoid security risks that can occur or are created at a rapid speed and on a large scale [9].

Many governments have implemented FDIs in various sectors, including health care. However, only a few countries (like Estonia and Australia) have applied their citizen FDI across multiple sectors (eg, financial, government, and health care) [6,7]. In contrast, other countries have applied it narrowly, and its use within health care tends to be lower [8]. Barriers to widespread adoption across sectors have been said to include user trust, unintended effects that may arise from requirements needed to establish an FDI, and policies and regulations [1]. As a result, FDIs have largely matured in the financial and government sectors, while their use in health care remains in its infancy [6]. Therefore, as the use of patient-facing digital platforms increases and more jurisdictions aspire to develop patient-centric systems, there is a need to understand how FDIs have been used, specifically in health care. Previous reviews largely focus on the implementation of platforms that use FDIs to provide patients with access to health care services, and they place an emphasis on factors such as policies, stakeholder engagement, and infrastructure [7]. Alternatively, the reviews that aim to understand how FDIs are used in health care focus on the login procedures used by patients to access their own data [8-11]. Therefore, there is a lack of studies that aim to understand how the FDIs used in health care are established or how their use impacts the implementation and uptake of the platforms that leverage them [7,12]. Therefore, given that governments globally have recently placed an emphasis on designing and implementing FDIs that transform how citizens access their data, we sought to characterize how FDIs being used in health care are currently established and leveraged to integrate different data sources [1,4,6,12].

Objectives

This rapid review aims to explore how FDIs have been established and implemented globally to provide patients with access to health care resources and services.

Methods

This environmental scan was conducted in 2 stages: a literature review using a rapid review methodology, followed by a targeted gray literature search [13].

Literature Review

Search Strategy

We conducted a literature search in the MEDLINE and Embase databases on February 1, 2021, to identify services that leveraged digital IDs within a health care setting. Our search strategy was built using key terms surrounding the concepts of "national electronic health records," "digital identity," and "single sign-on" (Multimedia Appendix 1 contains the detailed search strategy). Upon identifying services that applied digital identifiers to access health data, we also performed targeted searches on MEDLINE and Embase using the names of these services. All articles were exported into Zotero (Corporation for Digital Scholarship, Roy Rosenzweig Center for History and New Media), a citation management software.

Article Screening

For the literature search, as per commonly used rapid review methodologies [13-15], a 2-step screening consisting of a title-abstract and full text screening was performed by a single reviewer (KR). During the first step of screening, articles were excluded based on the title, and abstracts were read in cases

XSL•FO
RenderX

where the title was not conclusive. During the full-text screening stage, all articles were read in detail. The inclusion of papers was restricted to full-text publications in the English language published between January 1, 2011, and February 1, 2021, since articles examining FDIs in health care were uncommon before 2011. Specifically, since 2011, there has been a general upward trend in the total number of related articles. Articles focusing on the implementation and deployment of FDIs or single sign-on services within health care were included, along with those that focused on stakeholder experiences when interacting with these services. Research articles, commentaries, reviews, and nonresearch articles were included. Articles that focused on the implementation of a national electronic health record without a patient access component or articles that did not describe how the FDI was used in health care were excluded, as our focus was on the patients' user experience and the implementation of the FDI for health care purposes.

A standard data extraction form was used by a single reviewer (KR) to extract the following details from the included studies: author, year, country, name and features of the patient-accessible electronic health record (PAEHR), consent model, infrastructure, name and type of digital identity, process citizens use to access the data, name and type of identity provider, year implemented, and adoption rate. A 10% quality check was completed by a second reviewer (VS), as per common rapid review methodologies [13], wherein 10% of articles selected randomly were screened by VS based on the title, then abstracts, and when needed, full texts were consulted if additional details were required. Any discrepancies in the inclusion and exclusion of articles were discussed. Although this rapid review does not entail a quality assessment, a narrative approach consistent with the data extraction of a rapid review was used [13]. All data extractions were also discussed between KR and VS.

Gray Literature Search

The literature review generated a list of countries and services that have implemented FDIs to access health care resources. We used this list to conduct an additional gray literature search using Google (Multimedia Appendix 1). The Google search results were limited to the first 10 pages, allowing us to focus our data collection on the most relevant identified services from their respective service and government websites. For the supplemental gray literature search, a single reviewer (KR) screened the search results and extracted the data.

Ethical Considerations

This review has not been registered, and a protocol has not been developed.

Results

Literature and Gray Literature Search

The results of the search are summarized in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram in Figure 1. A total of 93 references were included in the review [10,11,16-106].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram detailing the search and selection process applied to the rapid review. FDI: federated digital identifier.

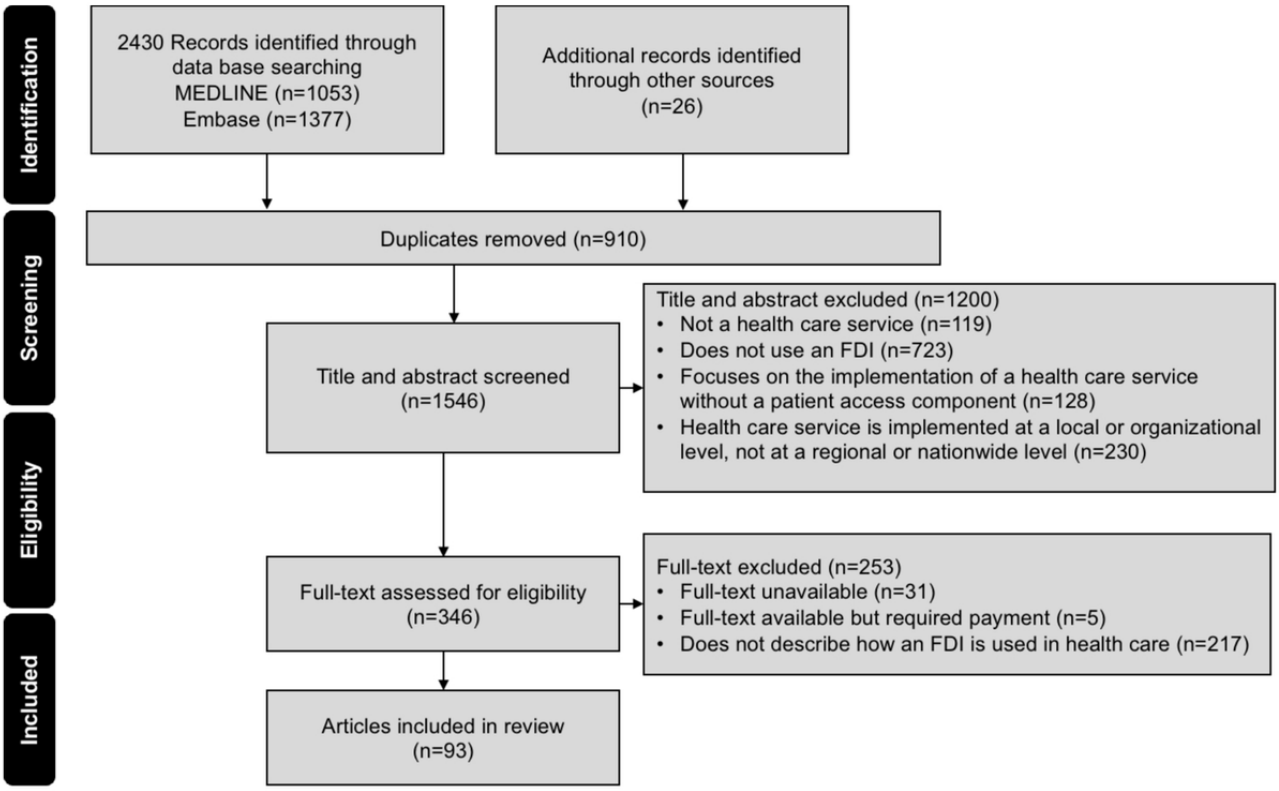

Through our literature search, we identified 9 countries that used digital identifiers for citizens' access to health care information. Our targeted searches identified 2 additional countries, for a total of 11 countries identified using FDIs. These were Estonia, Finland, Australia, Taiwan, Denmark, Norway, Sweden, Iceland, Singapore, Canada, and Belgium.

XSL•FO
RenderX

Application of FDIs to Health Care Platforms

Among the 11 identified countries, the most common application of digital identifiers in health care was in the context of the national PAEHR. The exception to this was Canada, where citizens of Alberta and Quebec access their health care information on a provincial level because health care is provincially managed [17,76]. Most countries that have implemented digital identifiers at a national level are in Europe.

Identifiers can be verifiable (ie, they can be used to authenticate a citizen's identity) and reusable (ie, they can be used across multiple sectors to provide citizens with access to multiple services) [9,16]. Across the jurisdictions we identified, the FDIs were always verifiable, and all except for Taiwan were also reusable [10,17,74]. Countries used 1 of 2 implementation frameworks [9]. Most countries (eg, Australia) use a centralized implementation framework that facilitates the collation of data from existing repositories and data sources into a central storage system [9]. Distributed implementation frameworks were used less frequently (eg, in Denmark, Sweden, and Iceland), and they stored the data presented on the single sign-on interface in disparate repositories and pulled relevant data into the system only when required [9].

Table 1 provides a comparison of how patients receive access to their health care services through a national or provincial PAEHR that uses FDIs. The earliest use of digital IDs for patient access was in Denmark in 2003 [9,11].

XSL•FO
RenderX

Table 1. Overview of countries around the world that have implemented a federated digital identifier to enable patient access, the name of the federated digital identifier, and the consent model used by the country.

Country; name of PAEHR[a]	Year of implementation	Description of platform	Consent model	FDI[b] used for access	References
Estonia; Digilugu.ee	2009	Centralized her system, which contains information about every interaction a patient has with the health care system. This is visible to both patients and to all clinicians who treat them.	Opt-out	Estonian eID	[10,11,16-21]
Finland; My Kanta Pages	2010	A web-based service where citizens can view data entered by private and public health care services concerning their interactions with the health care system as well as data they have entered themselves.	Opt-out except for ePrescription	Banking ID, mobile ID, or chip-based identity card enabling web-based transactions	[10,11,22-29]
Australia; My Health Record	2012	A web-based summary of citizens' key health information in one place. Health care providers can also see the record when they need to (eg, in an emergency).	Opt-out	MyGovID	[10,11,30-72]
Taiwan; My Health Bank	2014	A web-based service that allows people with National Health Insurance to view their own medical history.	Opt-in	National Health Insurance Card (NHI Card)	[17,73,74]
Denmark; Sundhed.dk	2003	A public portal that gives citizens and health care providers access to and information about all Danish health care services.	Opt-out	NemID	[9,11,26,75-78]
Norway; Helsenorge	2011	A web-based portal that provides citizens access to health care services and digital access to documents from patient portals.	Opt-out	BankID, BuyPass ID, or Commfides e-ID	[10,11,75,79,80]
Sweden; Journalen	2012	A web-based national patient portal that allows patients access to all of their health care information and is compatible with the electronic health record used by any of the patient's providers.	Opt-in	Freja eID plus or Bank ID	[10,11,26,77,81,94]
Iceland; Heilsuvera	2014	A portal that provides educational material about health, including prevention, while allowing citizens to manage their health care information.	Not Reported	Electronic ID (eID)	[11,95]
Singapore; HealthHub	2015	A one-stop health service that provides access to a family's health records, health information, and services.	Not Reported	SingPass ID	[96-99]
Canada; Carnet Sante´ Quebec	2018	A web-based service that provides access to citizens' health information.	Opt-in	clicSE´QUR	[26,100,101]
Canada; MyHealth Records Alberta	2019	A single account that provides citizens access to store health information in a secure place.	Opt-in	MyAlberta Digital ID	[11,102-104]
Belgium; Summarised Electronic Health Record	2019	A set of documents that health professionals, in consultation with the patient, decide to share as they are deemed necessary and relevant for care.	Not Reported	Bankcard, or Belgian ID	[105,106]

[a]PAEHR: patient-accessible electronic health record.

[b]FDI: federated digital identifier.

Establishing and Verifying FDIs

IDs were usually obtained at a municipal office, a government service center, a post office, or a bank that served as the identity provider. Among the jurisdictions studied, only Australia and Taiwan allowed individuals to obtain and verify a digital identity entirely on the web [45,74]. In contrast, the other jurisdictions (Belgium, Denmark, Estonia, Iceland, Norway, Singapore, Sweden, and Quebec) required citizens to obtain further documentation in person or had information mailed to their home address as an additional step to web-based verification [19,24,45,76,79,86,101,102,107].

Among the jurisdictions included in the study, Table 2 summarizes the common forms of identification used by identity

XSL•FO
RenderX

providers in establishing a digital ID. The most common documents used to verify one's identity when establishing a digital ID were a passport or banking information. Other common methods included jurisdictional ID cards and a driver's license.

Table 2. Forms of verification used by various identity providers when establishing a digital identifier.

Infrastructure and country; name of PAEHR[a] [references]	FDI[b] used for access	Login ID verification
Centralized infrastructure[c]		
Estonia; Digilugu.ee [10,11,19-21]	Estonian eID	Government ID and mobile ID
Finland; My Kanta Pages [10,11,24]	Banking ID, mobile ID, or chip-based identity card enabling web-based transactions	Bank ID and mobile ID
Australia; My Health Record [10,11,45-47]	MyGovID	Government ID and additional verification using health care or banking information
Taiwan; My Health Bank [73,74]	NHI Card (Not reusable)	Other
Distributed infrastructure[d]		
Denmark; Sundhed.dk [10,11,72]	NemID	Government ID
Norway; Helsenorge [10,11,79,80]	BankID, BuyPass ID, or Commfides e-ID	Government ID, bank ID, mobile ID, and other
Sweden; Journalen [11,17,86]	Freja eID plus or Bank ID	Government ID, bank ID, and mobile ID
Iceland; Heilsuvera [95]	eID	Government ID and mobile ID
Infrastructure not reported		
Singapore; HealthHub [99]	SingPass ID	Government ID
Canada; Carnet Sante´ Quebec [101]	clicSE´QUR	Government ID and additional verification using Health Insurance number to obtain an activation code
Canada; MyHealth Records Alberta [11,102]	MyAlberta Digital ID	Government ID and additional verification using Alberta Health Card number
Belgium; Summarised Electronic Health Record [106]	Bankcard, or Belgian ID	Government ID, bank ID, and additional verification using the identification token and verification code sent to mobile phone

[a]PAEHR: patient-accessible electronic health record.

[b]FDI: federated digital identifier.

[c]Centralized PAEHR uses a central store and is implemented as a separate layer on top of electronic health records that are already in use.

[d]Distributed PAEHR encourages health information exchange between different electronic health records and other data sources without creating a central store.

Different countries also instated either an opt-in or opt-out consent model for their PAEHRs. Opt-in consent models required users to agree to participate in the program, whereas in an opt-out consent model, users were automatically enrolled in the service and had to "opt-out" from the service to be excluded [11,42]. Some countries that used an opt-out consent model used it strategically to improve the adoption of their PAEHR platform [50]. For example, Australia switching from an opt-in to an opt-out consent model correlated with an increase in adoption from 20% to 90% [11,50].

Integrated Health Services in Platforms Using FDIs

Table 3 compares the features found in various PAEHR platforms that use FDIs through a single sign-on interface to provide patients access to their health information. Most platforms provide citizens with the ability to share and restrict access to their records or portions of their records to specific providers and provide a log of all personnel that have accessed their records. In some locations, such as Alberta and Estonia, health care providers can only access a specific user's information if they already have an existing patient-provider relationship [10,104]. In contrast, citizens in Australia cannot restrict access to the summary portion of their health record but can limit what is included in the summary, who can see the additional documents outside the summary, and who can remove documents if they want to [10,51]. In addition, most platforms, such as the Journalen in Sweden, allow citizens to share their health information with other users (eg, family members) [82].

XSL•FO
RenderX

They also often provide the ability for patients to view their record, or in some instances, such as in Australia's My Health Record, to view and edit their record to provide information on their personal health summary or to provide advance care directives [10,11,35,56]. Other features commonly available to citizens on most platforms include the ability to access their laboratory records, medications, and clinical summaries [11].

Table 3. Features of patient accessible electronic health record platforms that use federated digital identifiers (n=12). The most commonly implemented features include access to laboratory records, medications, and clinical summaries.

Features [references]	PAEHRs with feature, n (%)
Access to health information	
Access logs [10,11,17,23,75,96,100,103,105]	10 (83)
Share or restrict access to providers [10,11,17,23,73,105]	9 (75)
Provide access to trusted individual [10,11,23,96,103]	8 (67)
Parental access to child's record [10,11,17,23,96]	7 (58)
Appointment information	
Visits (time, date, or provider) [10,11,17,23,96,103]	9 (75)
Book appointments [10,11,17,103]	8 (67)
Referrals [23,75]	6 (50)
Reports or notes	
Clinical summaries [10,11,17,96,103,105]	10 (83)
Laboratory results [10,11,17,23,74,96,100,105]	10 (83)
Medical reports (radiology, pathology, etc) [17,23,75,96,105]	9 (75)
Diagnoses and conditions [10,11,17,96,103,105]	8 (67)
Child growth and development [17,51]	5 (42)
Medication information	
Medication and dispensing [10,11,17,96,100,103,105]	11 (92)
Allergies and adverse reactions [23,51,96,103,105]	7 (58)

Adoption Rates

The adoption rates of these PAEHR services that used FDIs varied considerably around the world (Figure 2; [10,11,16,18,24,25,39,50,74,77,78,84,90,95,96,98,100,102,104,106]). Australia and Singapore had the highest and lowest adoption rates at 90% and 1.5%, respectively, and most countries had adoption rates below 30% [23,40].

XSL•FO
RenderX

Figure 2. Countries and provinces (adoption rate in parentheses where available) and their respective PAEHR (year implemented in parentheses) that use federated digital identifies to provide patients with access to their health records and the names of their respective patient-accessible electronic health records. The adoption rate for Taiwan (920,000/23,726,185) was calculated as a percentage of the total population [74].

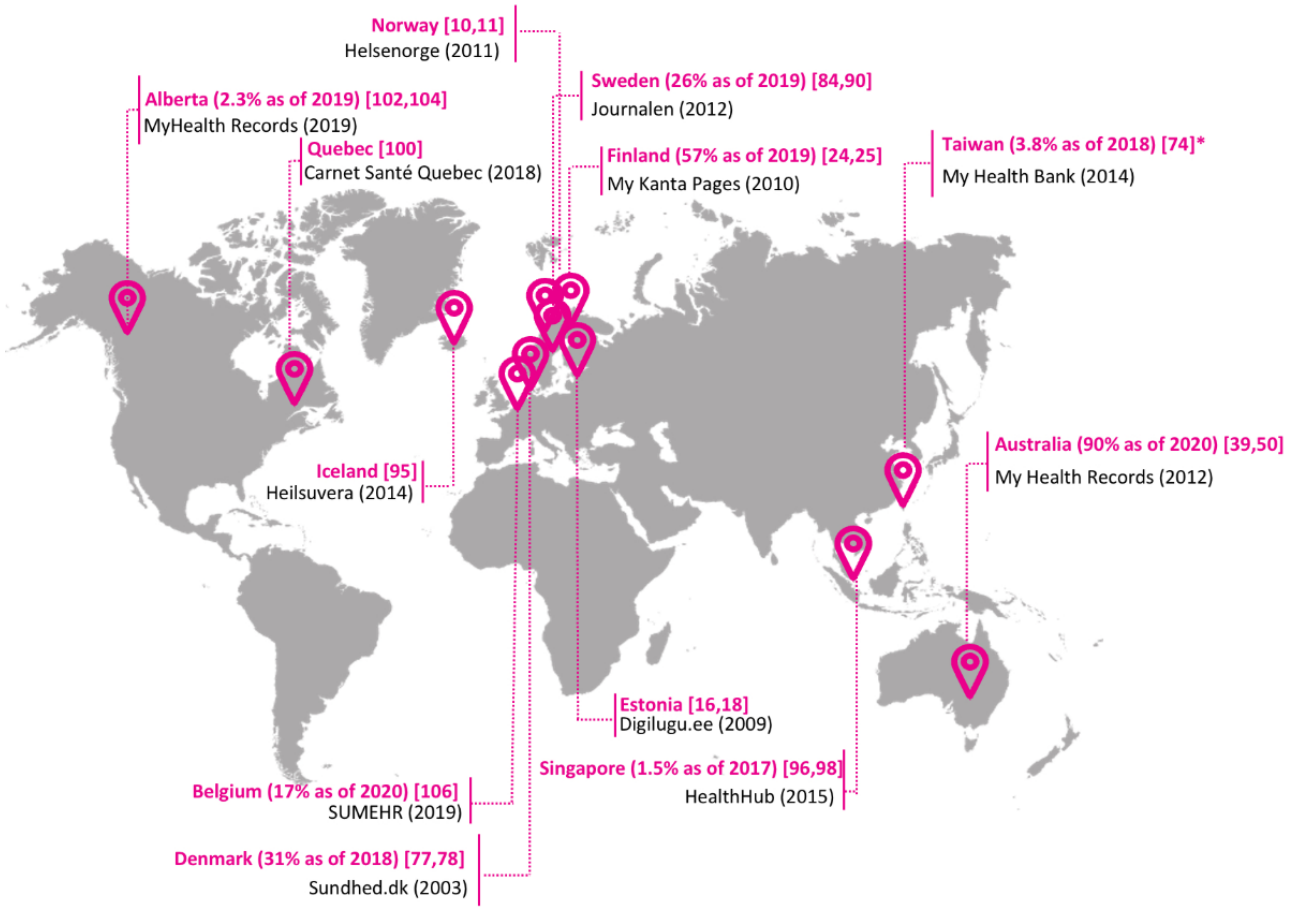

Discussion

Overview

This study found 11 countries that leveraged FDIs with a health care platform to provide patients access to their health information. FDIs were typically established and implemented by the government on a national level, and they were always used in the context of a PAEHR. PAEHRs allow patients to access data from multiple health care services through a single sign-on interface using either a centralized implementation framework (most common) or a distributed framework (less common). Most authentication procedures to verify a person's identity when establishing an FDI (to use with the single sign-on interface) were completed in person, and the identity providers most often included the bank or government. Additionally, with the exception of Taiwan, all countries used a reusable FDI, meaning the digital identity could be used in other sectors as well. Overall, the studies in the literature do not focus on describing patients' experiences with the FDI login and registration process but rather on the experience with the platforms these FDIs provided access to, that is, the PAEHRs.

Comparison to Other Work

To our knowledge, this is the first environmental scan of FDIs that describes how they are established and used in health care to provide patients with access to their health information. Previous reviews have only described user experiences focusing on their interaction with the PAEHR platforms themselves, as opposed to the experience of creating an FDI and using it to access their personal health information [10,11,108]. Although understanding these user experiences can inform factors that impact the uptake of the platform itself, it does not speak to the factors that impact the uptake of the digital ID or the influence that the digital ID creation experience has on the overall experience of the PAEHR user. Therefore, given that FDIs are increasingly being explored as an option to improve patient access to health care information [1,8], it is important for health system leaders to understand what identity verification and authentication processes are commonly used globally today. This will, in turn, inform how leaders design FDI establishment processes in the future.

Factors to Consider When Implementing FDIs in Health Care

One of the main perceived benefits of FDIs is their ability to reduce repetitive logins, eliminate the need to remember multiple passwords, and facilitate care across different institutions [6]. Given that FDIs used in the context of a national PAEHR do not require separate logins to access their health care information to begin with since electronic access to health care is already consolidated into 1 platform (from the users' perspective), the unique benefits of FDIs may not be as apparent. In these situations, the benefit of implementing an FDI may derive from its use across sectors (eg, financial and other governments),

XSL•FO
RenderX

preventing users from having to remember different login information for their different needs (eg, banking vs health care logins). Therefore, in these situations, the true benefits of FDIs only become apparent if one assesses their benefits across sectors [109,110]. In countries where the health care system is more fragmented and where there are multiple institutional, reimbursement, and regional portals, the benefits of FDIs to the health care sector alone may become more obvious [109,110]. Specifically, implementing FDIs across different jurisdictions could drive adoption of digital identity since it increases the value of having a digital identifier. This could, in turn, drive users to use the services that leverage the digital ID to provide access. Our environmental scan found that current literature is limited to describing how FDIs are established (ie, identity verification process using banking information, a passport, etc) or how they are used to provide access to health care services (ie, the login procedures) [11,17] However, there were no studies that aim to understand users' perceptions of the process of establishing an FDI and users' experiences with using an FDI when accessing health care resources. This makes it difficult to understand the barriers and facilitators to the adoption of FDIs in health care settings.

Factors Affecting PAEHR Adoption Irrespective of FDI Use

Through this environmental scan, we hoped to ascertain how patients perceive the benefits of FDIs, but as most FDIs have been used in the context of PAEHRs, the literature focused on how patients perceive the use of their country-specific PAEHR as opposed to the use of the FDI that enabled access. Common PAEHR platform features included providing patients with access to their laboratory and medical reports, booking appointments, and allowing a trusted individual to access their health care information [10,17,24]. The adoption rates of these PAEHR platforms remained below 30% in most countries even years after implementation, and this was true irrespective of the type of identity management being used (federated or not) [110,111]. Similar trends are also observed in countries such as France and the United Kingdom, which have also implemented PAEHRs for the entire nation but provide access without an FDI [111,112]. The adoption rates in these countries have been as low as 0.5% and 1.5%, respectively [111,112]. Our environmental scan found that among the countries that used FDI-managed PAEHRs, platforms with an opt-out consent model, such as Australia and Finland, generally had higher adoption rates [51,113]. This strategy to improve PAEHR adoption has also been reported in non-FDI PAEHR platforms [110]. Similarly, the factors that influenced adoption of PAEHRs were not unique to platforms that use FDIs with a single sign-on interface and were instead common to all digital health technologies. For example, facilitators such as supportive legislation, clear government guidelines, recognized standards, and proper stakeholder engagement worked across PAEHR platforms [114-117]. Negative attitudes and beliefs about health care professionals, a lack of leadership engagement, a lack of comprehensive information, stakeholder disagreements, and the presence of multiple local initiatives were, on the other hand, common barriers across PAEHRs, irrespective of the type of identity management system used [108,117,118].

Our environmental scan found that the information available to patients on the PAEHR platforms using FDIs varied in terms of who had access to the information, what information was presented, and how the information could be changed and modified [23,35,54,82]. For example, an individual in Taiwan was able to share or restrict information access to specific health care providers but could not provide access to a trusted individual like a caregiver, whereas someone from Singapore was unable to restrict access to a specific provider that belongs to their care circle but was able to provide access to an individual that they trust [18,74,97,98]. This variation can also be observed in PAEHRs not managed by FDIs [111,117]. Previous work by Essen et al [10] examining policy documentation of PAEHRs around the world also highlighted a heterogenous PAEHRs landscape with distinct services, regulation approaches, and patients' access across jurisdictions.

Limitations

All instances of FDI use in the literature were in the context of PAEHRs, with no studies focused exclusively on FDIs in the context of health care, so it was difficult to identify what benefits related to the use of FDI versus. those related to the use of PAEHR in general. We have reported the features and adoption rates of PAEHRs in this study, as it may be that the use of FDIs makes most sense in the context of regional PAEHRs. The low rate of adoption of these platforms could be attributed to challenges with the implementation and adoption of PAEHRs, not necessarily with the use of FDIs. Among the 11 countries that we identified as using FDIs for health care access, 46 of 93 included studies describing Australia's My Health Record, which was the most developed platform with the highest adoption rates among those reviewed. In comparison, the PAEHRs in other countries were not well established or studied, which made broader generalizations more limited. Finally, as this was an environmental scan with a rapid review, we only reviewed 2 databases and only included resources available in English, so it is possible that some platforms that leverage FDIs for health care access were missed. To mitigate this issue, we supplemented our academic search with gray literature and included governmental reports and websites in our resources.

Conclusion

Federated digital IDs have been leveraged for health care use around the world as their use facilitates the amalgamation of information through a single sign-on interface, which allows patients to access their information from multiple data sources. FDIs have been used exclusively in the context of PAEHRs, and adoption of PAEHRs remains low in many countries. As a result, it is difficult to disentangle the unique contribution of FDIs to these adoption rates. As FDIs provide patients with the opportunity to have a single point of access to health care services and information from multiple sources, future studies could focus on exploring patients' perceptions about the benefits and drawbacks of FDIs specifically. If FDIs are used to their fullest potential and implemented across sectors, adoption rates within health care may improve.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the Ontario Ministry of Health for funding this study. The funders played no part in the study design, collection, analysis, or interpretation of the data, in the writing of the report, or in the decision to submit the manuscript for publication.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOC File , 88 KB - jmir_v26i1e45751_app1.doc]

References

1. DIACC industry insights: digital ID in health care. Digital ID and Authentication Council of Canada. URL: https://diacc.ca/wp-content/uploads/2019/08/Industry-Insights-Digital-ID-in-Health-Care_August-2019.pdf [accessed 2024-01-11]
2. Suter E, Oelke ND, Adair CE, Armitage GD. Ten key principles for successful health systems integration. Healthc Q 2009;13(Spec No):16-23 [FREE Full text] [doi: 10.12927/hcq.2009.21092] [Medline: 20057244]
3. Satybaldy A, Hasselgren A, Nowostawski M. Decentralized identity management for e-Health applications: state-of-the-art and guidance for future work. Blockchain Healthc Today 2022;5:195 [FREE Full text] [doi: 10.30953/bhty.v5.195] [Medline: 36779018]
4. Federated identity management in health care. Canada Health Infoway. 2014. URL: https://infocentral.infoway-inforoute.ca/en/resources/docs/enterprise-architecture-health-care/310-federated-identity-management-in-health-care-full-report [accessed 2024-01-11]
5. Orion Health chronic care index report 2018. Orion Health. 2018. URL: https://orionhealth.com/global/media-releases/canadians-with-chronic-conditions-encounter-medication-errors-and-unnecessary-repeat-procedures/ [accessed 2024-01-11]
6. Building Canada's digital identity future. Digital ID and Authentication Council of Canada. 2015. URL: https://diacc.ca/2015/05/06/building-canadas-digital-future-2/ [accessed 2024-01-11]
7. Connecting health and care for the nation: a shared nationwide interoperability roadmap. The Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf [accessed 2024-01-11]
8. Giannopoulou A. Digital identity infrastructures: a critical approach of self-sovereign identity. Digit Soc 2023;2(2):18 [FREE Full text] [doi: 10.1007/s44206-023-00049-z] [Medline: 37200582]
9. Nøhr C, Parv L, Kink P, Cummings E, Almond H, Nørgaard JR, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. BMC Health Serv Res 2017;17(1):534 [FREE Full text] [doi: 10.1186/s12913-017-2482-y] [Medline: 28784173]
10. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
11. Nohr C, Wong MC, Turner P, Almond H, Parv L, Gilstad H, et al. Citizens' access to their digital health data in eleven countries—a comparative study. Stud Health Technol Inform 2016;228:685-689. [Medline: 27577472]
12. Beduschi A. Rethinking digital identity for post-COVID-19 societies: data privacy and human rights considerations. Data Policy 2021;3:e15 [FREE Full text] [doi: 10.1017/dap.2021.15]
13. Tricco AC, Langlois EV, Straus SE. Rapid reviews to strengthen health policy and systems: a practical guide. The Alliance for Health Policy and Systems Research. 2017. URL: https://ahpsr.who.int/docs/librariesprovider11/publications/supplementary-material/alliancehpsr_rrguide_trainingslides.pdf [accessed 2024-01-11]
14. Tricco AC, Antony J, Zarin W, Strifler L, Ghassemi M, Ivory J, et al. A scoping review of rapid review methods. BMC Med 2015;13:224 [FREE Full text] [doi: 10.1186/s12916-015-0465-6] [Medline: 26377409]
15. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
16. Habicht T, Reinap M, Kasekamp K, Sikkut R, Aaben L, van Ginneken E. Estonia: health system review. Health Syst Transit 2018;20(1):1-189 [FREE Full text] [Medline: 30277217]
17. Yeh MJ, Saltman RB. Creating online personal medical accounts: recent experience in two developed countries. Health Policy Technol 2019;8(2):171-178. [doi: 10.1016/j.hlpt.2019.05.004]

XSL•FO
RenderX

18. Metsallik J, Ross P, Draheim D, Piho G. Ten years of the e-health system in Estonia. CEUR Workshop Proc 2018:6-15 [FREE Full text]
19. Personal identity documents and applying for them. eesti.ee. URL: https://www.eesti.ee/en/citizenship-and-documents/personal-identity-documents-and-applying-for-them/#applyingforanidcard2 [accessed 2021-03-30]
20. Patient portal. digilugue.ee. URL: https://www.digilugu.ee/login?locale=en [accessed 2021-03-04]
21. de Lusignan S, Ross P, Shifrin M, Hercigonja-Szekeres M, Seroussi B. A comparison of approaches to providing patients access to summary care records across old and new europe: an exploration of facilitators and barriers to implementation. Stud Health Technol Inform 2013;192:397-401. [Medline: 23920584]
22. Lämsä E, Timonen J, Mäntyselkä P, Ahonen R. Pharmacy customers' experiences with the national online service for viewing electronic prescriptions in Finland. Int J Med Inform 2017;97:221-228. [doi: 10.1016/j.ijmedinf.2016.10.014] [Medline: 27919380]
23. Maria S, Emma A, Elina L, Riitta A, Johanna T. Use and non-use of a nationwide patient portal—a survey among pharmacy customers. J Pharm Health Serv Res 2020;11(4):335-342 [FREE Full text] [doi: 10.1111/jphs.12368]
24. Jormanainen V, Lindgren M. PHP189—national my kanta pages patient portal service: implementation, use and users in 2010-2017 in Finland. Value Health 2018;21:S182 [FREE Full text] [doi: 10.1016/j.jval.2018.09.1083]
25. Hyppönen H, Lumme S, Reponen J, Vänskä J, Kaipio J, Heponiemi T, et al. Health information exchange in Finland: usage of different access types and predictors of paper use. Int J Med Inform 2019;122:1-6. [doi: 10.1016/j.ijmedinf.2018.11.005] [Medline: 30623778]
26. Nordic eHealth benchmarking. Nordic Council of Ministers. 2014. URL: https://www.norden.org/en/publication/nordic-ehealth-benchmarking-0 [accessed 2024-01-11]
27. Makeham M. Role of digital technology in delivering 'healthy futures' and 'healthy cities'. Intern Med J 2020;50(11):1408-1409. [doi: 10.1111/imj.15062] [Medline: 33215836]
28. Number of users of My Kanta pages by municipality. Kanta.fi. URL: https://www.kanta.fi/en/blog/-/asset_publisher/1QjC602jKPR6/content/omakannan-kayttajat-kunnittain [accessed 2021-03-10]
29. The Social Insurance Institution of Finalnd. More than half of all Finnish residents use My Kanta pages—user numbers have grown especially during the corona outbreak—press releases. kela.en. URL: https://www.kela.fi/web/en/-/more-than-half-of-all-finnish-residents-use-my-kanta-pages-user-numbers-have-grown-especially-during-the-corona-outbreak [accessed 2021-03-10]
30. Black A, Sahama T, Gajanayake R. eHealth-as-a-Service (eHaaS): a data-driven decision making approach in Australian context. Stud Health Technol Inform 2014;205:915-919. [Medline: 25160321]
31. Bidargaddi N, van Kasteren Y, Musiat P, Kidd M. Developing a third-party analytics application using Australia's national personal health records system: case study. JMIR Med Inform 2018;6(2):e28 [FREE Full text] [doi: 10.2196/medinform.7710] [Medline: 29691211]
32. Lupton D. 'I'd like to think you could trust the government, but I don't really think we can': Australian women's attitudes to and experiences of My Health Record. Digit Health 2019;5:1-12 [FREE Full text] [doi: 10.1177/2055207619847017] [Medline: 31069106]
33. Kerai P, Wood P, Martin M. A pilot study on the views of elderly regional Australians of personally controlled electronic health records. Int J Med Inform 2014;83(3):201-209. [doi: 10.1016/j.ijmedinf.2013.12.001] [Medline: 24382474]
34. Statistics and insights. Digital Health. 2020. URL: https://www.digitalhealth.gov.au/sites/default/files/2021-06/myhealthrecord_stats_sept2020.pdf [accessed 2024-01-11]
35. Walsh L, Hill S, Allan M, Balandin S, Georgiou A, Higgins I, et al. A content analysis of the consumer-facing online information about My Health Record: implications for increasing knowledge and awareness to facilitate uptake and use. Health Inf Manag 2018;47(3):106-115 [FREE Full text] [doi: 10.1177/1833358317712200] [Medline: 28589741]
36. Mooranian A, Emmerton L, Hattingh L. The introduction of the national e-health record into Australian community pharmacy practice: pharmacists' perceptions. Int J Pharm Pract 2013;21(6):405-412 [FREE Full text] [doi: 10.1111/ijpp.12034] [Medline: 23560554]
37. Lehnbom EC, Douglas HE, Makeham MAB. Positive beliefs and privacy concerns shape the future for the personally controlled electronic health record. Intern Med J 2016;46(1):108-111. [doi: 10.1111/imj.12956] [Medline: 26813902]
38. Andrews L, Gajanayake R, Sahama T. The Australian general public's perceptions of having a Personally Controlled Electronic Health Record (PCEHR). Int J Med Inform 2014;83(12):889-900. [doi: 10.1016/j.ijmedinf.2014.08.002] [Medline: 25200198]
39. de Mesquita RC, Edwards I. Systematic literature review of My Health Record system. Asia Pac J Health Manag 2020;15(1):14-25 [FREE Full text] [doi: 10.24083/apjhm.v15i1.311]
40. Kariotis TC, Harris KM. Clinician perceptions of My Health Record in mental health care: medication management and sharing mental health information. Aust J Prim Health 2019;25(1):66-71. [doi: 10.1071/PY17181] [Medline: 30636668]
41. Armani R, Mitchell LE, Allen-Graham J, Heriot NR, Kotsimbos T, Wilson JW. Current patient and healthcare worker attitudes to eHealth and the personally controlled electronic health record in major hospitals. Intern Med J 2016;46(6):717-722. [doi: 10.1111/imj.13086] [Medline: 27040359]

XSL•FO
RenderX

42. Torrens E, Walker SM. Demographic characteristics of Australian health consumers who were early registrants for opt-in personally controlled electronic health records. Health Inf Manag 2017;46(3):127-133. [doi: 10.1177/1833358317699341] [Medline: 28537210]
43. Almond H, Cummings E, Turner P. An approach for enhancing adoption, use and utility of shared digital health records in Rural Australian Communities. Stud Health Technol Inform 2017;235:378-382. [Medline: 28423818]
44. Williams PAH. Does the PCEHR mean a new paradigm for information security? Implications for health information management. Health Inf Manag 2013;42(2):31-36. [doi: 10.12826/18333575.2013.0007.Williams] [Medline: 23695935]
45. How to set up. myGovID. URL: https://www.mygovid.gov.au/set-up [accessed 2021-03-30]
46. Australian Government. Register for a My Health Record. My Health Record. 2018. URL: https://www.myhealthrecord.gov.au/for-you-your-family/howtos/register-for-my-health-record [accessed 2021-03-04]
47. Australian Government. Using myGovID with myGov. myGovID. URL: https://www.mygovid.gov.au/using-mygovid-mygov#benefits [accessed 2021-03-04]
48. Lehnbom EC, Brien JE, McLachlan AJ. Knowledge and attitudes regarding the personally controlled electronic health record: an Australian national survey. Intern Med J 2014;44(4):406-409. [doi: 10.1111/imj.12384] [Medline: 24754689]
49. Khan UR, Zia TA, Pearce C, Perera K. The MyHealthRecord system impacts on patient workflow in general practices. Stud Health Technol Inform 2019;266:162-167 [FREE Full text] [doi: 10.3233/SHTI190789] [Medline: 31397318]
50. Vimalachandran P, Liu H, Lin Y, Ji K, Wang H, Zhang Y. Improving accessibility of the Australian My Health Records while preserving privacy and security of the system. Health Inf Sci Syst 2020;8(1):31 [FREE Full text] [doi: 10.1007/s13755-020-00126-4] [Medline: 33088487]
51. Garrety K, McLoughlin I, Wilson R, Zelle G, Martin M. National electronic health records and the digital disruption of moral orders. Soc Sci Med 2014;101:70-77. [doi: 10.1016/j.socscimed.2013.11.029] [Medline: 24560226]
52. Kariotis T, Prictor M, Chang S, Gray K. Evaluating the contextual integrity of Australia's My Health Record. Stud Health Technol Inform 2019;265:213-218 [FREE Full text] [doi: 10.3233/SHTI190166] [Medline: 31431601]
53. Carroll J, Butler-Henderson K. MyHealthRecord in Australian primary health care: an attitudinal evaluation study. J Med Syst 2017;41(10):158. [doi: 10.1007/s10916-017-0807-3] [Medline: 28866846]
54. Hemsley B, Georgiou A, Carter R, Hill S, Higgins I, van Vliet P, et al. Use of the My Health Record by people with communication disability in Australia: a review to inform the design and direction of future research. Health Inf Manag 2016;45(3):107-115. [doi: 10.1177/1833358316652060] [Medline: 27269277]
55. Duckett S. Australia's new digital health record created ethical dilemmas. Healthc Manage Forum 2019;32(3):167-168. [doi: 10.1177/0840470419827719] [Medline: 30935232]
56. Hanna L, Gill SD, Newstead L, Hawkins M, Osborne RH. Patient perspectives on a personally controlled electronic health record used in regional Australia. Health Inf Manag 2017;46(1):42-48. [doi: 10.1177/1833358316661063] [Medline: 27486184]
57. Jackson S, Peterson G. My Health Record: a community pharmacy perspective. Aust Prescr 2019;42(2):46-47 [FREE Full text] [doi: 10.18773/austprescr.2019.009] [Medline: 31048935]
58. Quinlivan JA, Lyons S, Petersen RW. Attitudes of pregnant women towards personally controlled electronic, hospital-held, and patient-held medical record systems: a survey study. Telemed J E Health 2014;20(9):810-815. [doi: 10.1089/tmj.2013.0342] [Medline: 25046543]
59. Almond H, Cummings E, Turner P. Recommendations for enhancing the implementation and utility of shared digital health records in rural Australian communities. Stud Health Technol Inform 2018;252:15-20. [Medline: 30040676]
60. Bainbridge M, Pearce C, Taggart R. A maturity model to deliver a national shared medications system, options, practice, and pitfalls. Stud Health Technol Inform 2013;192:1070. [Medline: 23920844]
61. Showell CM. Citizens, patients and policy: a challenge for Australia's national electronic health record. Health Inf Manag 2011;40(2):39-43. [doi: 10.1177/183335831104000206] [Medline: 28683627]
62. Walsh L, Hemsley B, Allan M, Adams N, Balandin S, Georgiou A, et al. The e-health literacy demands of Australia's My Health Record: a heuristic evaluation of usability. Perspect Health Inf Manag 2017;14(Fall):1f [FREE Full text] [Medline: 29118683]
63. Royle R, Blackburn S, Walduck A. Review of the personally controlled electronic health record. Ministry of Health. 2013. URL: https://delimiter.com.au/wp-content/uploads/2014/05/FINAL-Review-of-PCEHR-December-2013.pdf [accessed 2024-01-11]
64. Partel K. Toward better implementation: Australia;s My Health Record. Deeble Institute. 2015. URL: https://apo.org.au/sites/default/files/resource-files/2015-10/apo-nid58339.pdf [accessed 2024-01-11]
65. Walsh L, Hemsley B, Allan M, Dahm MR, Balandin S, Georgiou A, et al. Assessing the information quality and usability of within a health literacy framework: what's changed since 2016? Health Inf Manag 2021;50(1-2):13-25 [FREE Full text] [doi: 10.1177/1833358319864734] [Medline: 31370712]
66. Van Kasteren Y, Maeder A, Williams PA, Damarell R. Consumer perspectives on MyHealth Record: a review. Stud Health Technol Inform 2017;239:146-152. [Medline: 28756450]
67. Kosari S, Yee KC, Mulhall S, Thomas J, Jackson SL, Peterson GM, et al. Pharmacists' perspectives on the use of My Health Record. Pharmacy (Basel) 2020;8(4):1-12 [FREE Full text] [doi: 10.3390/pharmacy8040190] [Medline: 33066569]

XSL•FO
RenderX

68. Vimalachandran P, Wang H, Zhang Y, Zhuo G. The Australian PCEHR system: ensuring privacy and security through an improved access control mechanism. EAI Endorsed Trans. Scalable Inf. Syst 2016;3:e4 [FREE Full text] [doi: 10.4108/eai.9-8-2016.151633]
69. My Health Record statistics. My Health Record. 2018. URL: https://www.myhealthrecord.gov.au/statistics [accessed 2021-03-16]
70. Pearce C, Bainbridge M. A personally controlled electronic health record for Australia. J Am Med Inform Assoc 2014;21(4):707-713 [FREE Full text] [doi: 10.1136/amiajnl-2013-002068] [Medline: 24650635]
71. Pearce C, Bartlett J, Mcleod A, Eustace P, Amos R, Shearer M. Effectiveness of local support for the adoption of a national programme--a descriptive study. Inform Prim Care 2014;21(4):171-178 [FREE Full text] [doi: 10.14236/jhi.v21i4.70] [Medline: 25479347]
72. Hemsley B, Georgiou A, Balandin S, Carter R, Hill S, Higgins I, et al. The Personally Controlled Electronic Health Record (PCEHR) for adults with severe communication impairments: findings of pilot research. Stud Health Technol Inform 2015;214:100-106. [Medline: 26210425]
73. Li YCJ, Yen JC, Chiu WT, Jian WS, Syed-Abdul S, Hsu MH. Building a national electronic medical record exchange system—experiences in Taiwan. Comput Methods Programs Biomed 2015;121(1):14-20. [doi: 10.1016/j.cmpb.2015.04.013] [Medline: 26001420]
74. Lee IC. 'My Health Bank' service to be expanded, NHIA says. Taipei Times. 2018. URL: https://www.taipeitimes.com/News/taiwan/archives/2018/11/08/2003703825 [accessed 2021-03-10]
75. Fragidis LL, Chatzoglou PD. Development of Nationwide Electronic Health Record (NEHR): an international survey. Health Policy Technol 2017;6(2):124-133. [doi: 10.1016/j.hlpt.2017.04.004]
76. Valid ID. NemID. URL: https://www.nemid.nu/dk-en/get_started/request_nemid/valid_id/ [accessed 2021-03-30]
77. A brief overview of health IT collaboration and interoperability in five countries in 2018. Ehelse. 2018. URL: https://www.ehelse.no/publikasjoner/a-brief-overview-of-health-it-collaboration-and-interoperability-in-five-countries-in-2018 [accessed 2024-01-11]
78. Government of Denmark. Background. Sundhed.dk. 2022. URL: https://www.sundhed.dk/borger/service/om-sundheddk/om-organisationen/ehealth-in-denmark/background/ [accessed 2024-01-11]
79. bankID. URL: https://www.bankid.no/privat/ [accessed 2021-03-04]
80. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]
81. E-health services available for everyone and everyone at all times in Iceland. Government Offices of Sweden, The Global Goals, Nordic Council of Ministers, Region Vasterbotten. URL: https://www.healthcareatdistance.com/media/1121/e-health-services-available-for-everyone-and-at-all-times-in-iceland.pdf [accessed 2024-01-11]
82. Nurgalieva L, Cajander Å, Moll J, Åhlfeldt RM, Huvila I, Marchese M. 'I do not share it with others. No, it's for me, it's my care': on sharing of patient accessible electronic health records. Health Informatics J 2020;26(4):2554-2567 [FREE Full text] [doi: 10.1177/1460458220912559] [Medline: 32264734]
83. Grünloh C, Myreteg G, Cajander Å, Rexhepi H. "Why do they need to check me?" Patient participation through eHealth and the doctor-patient relationship: qualitative study. J Med Internet Res 2018;20(1):e11 [FREE Full text] [doi: 10.2196/jmir.8444] [Medline: 29335237]
84. Bärkås A, Scandurra I, Hägglund M. Analysis of voluntary user feedback of the Swedish national PAEHR service. Stud Health Technol Inform 2019;264:1126-1130 [FREE Full text] [doi: 10.3233/SHTI190401] [Medline: 31438100]
85. Scandurra I, Lyttkens L, Eklund B. Implications of Swedish national regulatory framework of the patient accessible electronic health record. Stud Health Technol Inform 2016;228:695-699. [Medline: 27577474]
86. Beetzen K. Get Freja eID. Freja eID. URL: https://frejaeid.com/en/get-freja-eid/ [accessed 2021-03-30]
87. Your medical records. Region Skåne. URL: https://www.skane.se/en/Health-care/rules-and-rights-in-health-care/your-medical-records/ [accessed 2021-03-04]
88. Hägglund M, Scandurra I. A socio-technical analysis of patient accessible electronic health records. Stud Health Technol Inform 2017;244:3-7. [Medline: 29039366]
89. Hägglund M, Scandurra I. Patients' online access to electronic health records: current status and experiences from the implementation in Sweden. Stud Health Technol Inform 2017;245:723-727. [Medline: 29295193]
90. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
91. Scandurra I, Pettersson M, Eklund B, Lyttkens L. Analysis of the updated Swedish regulatory framework of the patient accessible electronic health record in relation to usage experience. Stud Health Technol Inform 2017;245:798-802. [Medline: 29295208]
92. Moll J, Cajander Å. On patient accessible electronic health records and the experienced effect on the work environment of nurses. Stud Health Technol Inform 2020;270:1021-1025 [FREE Full text] [doi: 10.3233/SHTI200316] [Medline: 32570536]

XSL•FO
RenderX

93. Hägglund M, Moll J, Åhlfeldt R, Scandurra I. Timing it right—patients' online access to their record notes in Sweden. Stud Health Technol Inform 2018;247:336-340. [Medline: 29677978]
94. Wass S, Vimarlund V. The role of PAEHRs in patient involvement. J Med Syst 2018;42(11):210 [FREE Full text] [doi: 10.1007/s10916-018-1070-y] [Medline: 30255404]
95. Electronic ID by phone. Islandbanki.is. URL: https://www.islandsbanki.is/en/article/electronic-id-and-security [accessed 2021-03-30]
96. See QY. Attitudes and perceptions of general practitioners towards the National Electronic Health Record (NEHR) in Singapore. Emerg Med J 2020;5(1):86-93 [FREE Full text] [doi: 10.33590/emj/19-00184]
97. Register for SingPass. Government of Singapore. URL: https://www.singpass.gov.sg/singpass/register/instructions [accessed 2021-03-04]
98. Portal gives S'poreans access to recent health records. Mtoday. URL: https://www.todayonline.com/singapore/new-portal-and-app-launched-allow-users-access-health-records-check-future-appointments [accessed 2021-03-04]
99. Singapore's journey to build a national electronic health record system. Accenture. 2012. URL: https://www.hhmglobal.com/knowledge-bank/articles/singapores-journey-to-build-a-national-electronic-health-record-system [accessed 2024-01-11]
100. Government of Quebec. The Québec health booklet. Carnet Sante. URL: https://carnetsante.gouv.qc.ca/portail?Langue=en [accessed 2021-03-03]
101. Government of Quebec. What are the steps for creating a clicSÉQUR account when your identity is verified by the Régie de l'assurance maladie du Québec? clicSÉQUR—Citoyens. URL: https://www.info.clicsequr.gouv.qc.ca/en/citoyens/create-clicsequr-account/steps-creation-clicsequr-account-ramq/ [accessed 2021-03-30]
102. My Alberta Digital Id. URL: https://account.alberta.ca/ [accessed 2021-03-30]
103. McEachern A, Cholewa D. Digital health services and digital identity in Alberta. Stud Health Technol Inform 2017;234:222-227. [Medline: 28186045]
104. Review of connect care, Alberta netcare and MyHealth Records. Ernst & Young LLP. 2020. URL: https://open.alberta.ca/publications/review-of-connect-care-alberta-netcare-and-myhealth-records-final-report [accessed 2024-01-11]
105. Getting started with itsme. itsme. URL: https://www.itsme.be/en/get-started [accessed 2021-03-04]
106. Kingdom of Belgium Foreign Affairs, Foreign Trade and Development Cooperation. The identity card for adults and children older than 12. Federal Public Service Foreign Affairs. 2016. URL: https://diplomatie.belgium.be/en/belgians-abroad/identity-card-belgians-living-abroad/identity-card-adults-and-children-older-12-eid [accessed 2021-03-30]
107. Benjamins J, Haveman-Nies A, Gunnink M, Goudkuil A, de Vet E. How the use of a patient-accessible health record contributes to patient-centered care: scoping review. J Med Internet Res 2021;23(1):e17655 [FREE Full text] [doi: 10.2196/17655] [Medline: 33427683]
108. Fujioka JK, Bickford J, Gritke J, Stamenova V, Jamieson T, Bhatia RS, et al. Implementation strategies to improve engagement with a multi-institutional patient portal: multimethod study. J Med Internet Res 2021;23(10):e28924 [FREE Full text] [doi: 10.2196/28924] [Medline: 34709195]
109. Otte-Trojel T, de Bont A, Aspria M, Adams S, Rundall TG, van de Klundert J, et al. Developing patient portals in a fragmented healthcare system. Int J Med Inform 2015;84(10):835-846. [doi: 10.1016/j.ijmedinf.2015.07.001] [Medline: 26190049]
110. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. BMJ 2010;341:c5814 [FREE Full text] [doi: 10.1136/bmj.c5814] [Medline: 21081595]
111. Seroussi B, Bouaud J. Use of a nationwide personally controlled electronic health record by healthcare professionals and patients: a case study with the French DMP. Stud Health Technol Inform 2017;235:333-337. [Medline: 28423809]
112. Vehko T, Ruotsalainen S, Hyppönen H. E-health and e-welfare of Finland: check point 2018. National Institute for Health and Welfare, Helsinki. 2019. URL: http://www.julkari.fi/handle/10024/138244 [accessed 2024-01-11]
113. Ammenwerth E. From eHealth to ePatient: the role of patient portals in fostering patient empowerment. Eur J Biomed Inform 2018;14(2):20-23 [FREE Full text] [doi: 10.24105/ejbi.2018.14.2.4]
114. Kluge EH. Advanced patient records: some ethical and legal considerations touching medical information space. Methods Inf Med 1993;32(2):95-103. [Medline: 8321138]
115. Pyper C, Amery J, Watson M, Crook C, Thomas B. Patients' access to their online electronic health records. J Telemed Telecare 2002;8(Suppl 2):103-105. [doi: 10.1177/1357633X020080S247] [Medline: 12217158]
116. Cijvat CD, Cornet R, Hägglund M. Factors influencing development and implementation of patients' access to electronic health records-a comparative study of Sweden and the Netherlands. Front Public Health 2021;9:621210 [FREE Full text] [doi: 10.3389/fpubh.2021.621210] [Medline: 34169054]
117. Liu J, Gong X, Weal M, Dai W, Hou S, Ma J. Attitudes and associated factors of patients' adoption of patient accessible electronic health records in China—a mixed methods study. Digit Health 2023;9:1-17 [FREE Full text] [doi: 10.1177/20552076231174101] [Medline: 37188077]
118. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]

XSL•FO
RenderX

Abbreviations

FDI: federated digital identifier
PAEHR: patient-accessible electronic health record
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 16.01.23; peer-reviewed by H Naseri, KM Kuo; comments to author 20.03.23; revised version received 04.08.23; accepted 19.12.23; published 08.02.24.

Please cite as:
Ramamoorthi K, Stamenova V, Liu RH, Bhattacharyya O
The Implementation of Federated Digital Identifiers in Health Care: Rapid Review
J Med Internet Res 2024;26:e45751
URL: https://www.jmir.org/2024/1/e45751
doi:10.2196/45751
PMID:38329799

©Karishini Ramamoorthi, Vess Stamenova, Rebecca H Liu, Onil Bhattacharyya. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness and Feasibility of Telehealth-Based Dietary Interventions Targeting Cardiovascular Disease Risk Factors: Systematic Review and Meta-Analysis

Rupal Trivedi[1], RDN, MS, PhD; Shaimaa Elshafie[1], BPharm, MS; Randall Tackett[1], PhD; Henry Young[1], PhD; Elisabeth Lilian Pia Sattler[1,2], RPh, PhD

[1]Department of Clinical and Administrative Pharmacy, College of Pharmacy, University of Georgia, Athens, GA, United States

[2]Department of Nutritional Sciences, College of Family and Consumer Sciences, University of Georgia, Athens, GA, United States

Corresponding Author:
Elisabeth Lilian Pia Sattler, RPh, PhD
Department of Clinical and Administrative Pharmacy
College of Pharmacy
University of Georgia
250 W Green St
Athens, GA
United States
Phone: 1 7065421040
Email: lilian@uga.edu

Abstract

Background: Telehealth-based dietary interventions were recommended for cardiovascular disease (CVD) management during the COVID-19 pandemic; however, data regarding their effectiveness and feasibility are limited.

Objective: We aimed to examine (1) the effectiveness of telehealth-based dietary interventions in improving clinical CVD risk factors and (2) the feasibility of these interventions among individuals with CVD.

Methods: To conduct this systematic review and meta-analysis of randomized controlled trials (RCTs), 2 investigators searched PubMed, Cochrane Library, Web of Science, and ClinicalTrials.gov databases based on predetermined search terms and included English-language RCTs published between January 2000 and July 2022. The Cochrane Risk of Bias tool was used to assess RCT quality. To evaluate intervention effectiveness, weight, BMI, systolic and diastolic blood pressure, and levels of total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, or blood glucose were compared postintervention in telehealth and usual care (UC) groups. Feasibility was determined through the number of participants retained in intervention and UC groups. Pooled data for each CVD outcome were analyzed using a random effects model. Mean difference (MD), standardized MD, or risk ratio were calculated using R software.

Results: A total of 13 RCTs with 3013 participants were included in the analysis to assess the effectiveness and feasibility of telehealth-based dietary interventions among individuals with CVD. Participants had a mean age of 61.0 (SD 3.7) years, and 18.5% (n=559) were women. Approximately one-third of RCTs were conducted in the United States (n=4, 31%). Included studies used telephone, app, text, audio-visual media, or website-based interventions. Of the 13 included studies, 3 were of high quality, 9 were of moderate quality, and only 1 was of low quality. Pooled estimates showed systolic blood pressure (MD –2.74, 95% CI –4.93 to –0.56) and low-density lipoprotein cholesterol (standardized MD –0.11, 95% CI –0.19 to –0.03) to be significantly improved among individuals with CVD as a result of telehealth-based dietary interventions compared to UC. No significant difference in effectiveness was detected for weight, BMI, and levels of diastolic blood pressure, total cholesterol, high-density lipoprotein, and triglycerides between telehealth-based dietary interventions and UC among those with CVD. There was no significant difference between the feasibility of telehealth-based dietary interventions versus UC. Significant I^2 indicated moderate to considerable heterogeneity.

Conclusions: Telehealth-based dietary interventions show promise in addressing CVD risk factors.

(J Med Internet Res 2024;26:e49178) doi:10.2196/49178

XSL•FO
RenderX

KEYWORDS

telehealth; dietary interventions; cardiovascular diseases; risk factors; self-management; systematic review and meta-analysis

Introduction

Cardiovascular diseases (CVDs) impact 126.9 million American adults and are associated with 5.0 million annual hospitalizations, 874,613 annual deaths, and US $378 billion in medical expenditures in the United States [1]. CVD diagnoses are adversely impacted by lifestyle-related CVD risk factors [2]. About 80% of CVD risk in the United States can be attributed to lifestyle factors, including the consumption of an unhealthy diet, smoking, physical inactivity, alcohol use, obesity, uncontrolled levels of blood pressure, total cholesterol, and blood glucose [2]. Out of these lifestyle-related factors, the consumption of an unhealthy diet contributes to the largest proportion of CVD risk among Americans [2].

For primary and secondary CVD prevention, clinical guidelines endorse the maintenance of a healthy body weight, engagement in physical activity, smoking cessation, limited alcohol consumption, and a healthy diet [3-6]. Registered dietitian nutritionist (RDN)-led medical nutrition therapy has been shown to successfully improve dietary consumption among individuals with CVD [3,5,6]; however, limited access to health care services due to long travel distances and lack of adequate transportation to health care facilities presents significant barriers to obtaining effective medical nutrition therapy, particularly in rural communities [7-11].

Telehealth technologies provide a solution for overcoming these barriers by administering and supporting clinical health care over a long distance [12,13]. Through the use of the internet, videoconferencing, streaming media, electronic health records, and terrestrial and wireless communication tools, a variety of medical conditions have been successfully managed, including CVD [14,15]. Furthermore, the American Society for Preventative Cardiology recommended telehealth technologies for the dissemination of RDN-led medical nutrition therapy among those diagnosed with CVD during the COVID-19 pandemic [16]. While RDN-led telehealth-based medical nutrition therapy showed great promise in expanding patient care infrastructure during the pandemic, data regarding the effectiveness and feasibility of telehealth services among individuals with CVD are limited [16]. Previous research used small sample sizes, lacked inclusion of individuals with CVD as the target population, and did not assess telehealth-based dietary interventions [17]. Therefore, the objective of this study was to determine the effectiveness and feasibility of telehealth-based dietary interventions among individuals with CVD.

Methods

Study Design

We conducted a systematic review and meta-analysis, guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [18].

Inclusion Criteria

Randomized controlled trials (RCTs) published in the English language between January 2000 and July 31, 2022, that compared telehealth-based dietary interventions and usual care (UC) among adults aged ≥18 years with CVD were included in this study. Eligible RCTs involved participants with CVDs, including coronary heart disease, heart failure, hypertension, individuals with CVD with a history of cardiac events (myocardial infarction, acute coronary syndromes, angina, and revascularization), and individuals with CVD with a history of cardiac procedures (percutaneous coronary intervention and coronary artery bypass graft surgery) [19,20]. Interventions were considered as telehealth-based dietary interventions if diet-related information was conveyed through synchronous or asynchronous points-of-contacts between a patient and a health care professional via medical education systems or information and communications technologies, such as telephones, cellular phones, web-based systems, and video technologies [21,22]. Participants in UC groups received some form of in-person nutritional care as a part of CVD management. If the involvement of nutritional care in UC was not clearly stated in the RCTs' papers, authors R Trivedi and ELPS contacted the study authors to gain clarity. In the case where authors did not reply with clarification, the corresponding RCT was excluded from this systematic review and meta-analysis. To determine the effectiveness of telehealth-based dietary interventions, clinical CVD risk factors (weight, BMI, systolic and diastolic blood pressures, and levels of total cholesterol, low-density lipoprotein [LDL] cholesterol, high-density lipoprotein [HDL] cholesterol, triglycerides, or blood glucose) were assessed in postintervention in telehealth, and compared to UC groups. To determine patient feasibility toward the telehealth-based dietary interventions, the included studies reported the number of participants retained by the RCT in both intervention and UC groups.

Search Strategy

PubMed, Cochrane Library, Web of Science, and ClinicalTrials.gov databases were searched using the following key terms and Medical Subject Headings terms: ("diet" or "nutrition" or "nutrition status" or "status, nutrition" or "nutritional science" or "science, nutritional" or "nutrition science" or "science, nutrition") and ("telehealth" or "telemedicine" or "mobile health" or "health, mobile" or "mhealth" or "telehealth" or "ehealth") and ("cardiovascular disease" or "cardiovascular diseases" or "disease, cardiovascular" or "diseases, cardiovascular" or "heart disease" or "cardiac disease" or "cardiac disorder" or "heart disorder" or "vascular disease" or "disease, vascular") and ("rand*"). An experienced librarian reviewed and confirmed this search strategy (Multimedia Appendix 1).

Using the predetermined search string, 2 investigators (R Trivedi and SE) first screened titles and abstracts to determine eligibility independently. Further scrutiny was given to the full text of the selected papers to determine whether eligibility was sustained.

XSL•FO
RenderX

The included and excluded records were compared between the 2 investigators, and no unresolved disagreements needed to be reconciled through the input of a third investigator (ELPS).

Data Extraction

Study characteristics manually extracted from each RCT included National Clinical Trial identifier number, first author's last name, year of publication, cardiovascular condition, interventions and their duration, follow-up durations after intervention initiation, country in which the study took place, location type (urban or rural), and total number of randomized participants. Participant characteristics extracted from each RCT included age, gender, race or ethnicity, and education and income levels. Continuous clinical outcome data included postintervention weight (kg or lbs), BMI (kg/m^2), systolic and diastolic blood pressure (mmHg), and levels of total cholesterol (mg/dL or mmol/L), LDL cholesterol (mg/dL or mmol/L), HDL cholesterol (mg/dL or mmol/L), triglycerides (mg/dL or mmol/L), or blood glucose (mg/dL or mmol/L) from intervention and UC groups, respectively. The dichotomous feasibility outcome data included the number of participants retained in intervention and UC groups [23-28].

Risk of Bias

Study quality was determined independently by R Trivedi and SE through the Cochrane risk of bias tool, assessing randomization, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete or selective reporting, and external sources of bias [29]. The decision on the RCTs' overall quality was based on the number of unmet or insufficiently described criteria and the following thresholds: more than 3 (low quality), 2-3 (moderate quality), and less than 2 (high quality). All conflicts were resolved through discussion by the 2 investigators, and no further input was required by ELPS as a third investigator.

Analytic Plan

Pooled data for each outcome was assessed using a random effects model. Data for continuous variables were assessed either as weighted mean difference (MD) or standardized mean difference (SMD) using the inverse-variance model while our dichotomous variable was assessed as a risk ratio using the Mantel-Haenszel (DerSimonian-Laird method) model to account for the variations among studies. An MD was calculated when outcomes were reported in a uniform measurement scale and SMD was calculated when nonuniform measurement scales were used. Since participants from the included studies were assigned to intervention and UC groups at random, baseline characteristics of participants between groups of each study were assumed to be similar. Therefore, no analysis was conducted to detect differences between the intervention and the UC groups at baseline. Visual inspection of the forest plots and the statistically measured I^2 determined the heterogeneity among the analyzed studies. The level of heterogeneity was based on the following I^2 thresholds: up to 29.99% (low), 30%-59.99% (moderate), 60%-74.99% (substantial), and 75%-100% (considerable). A sensitivity analysis was performed if significant heterogeneity was detected through analysis and the outcome involved more than 7 RCTs, ensuring that the I^2 estimate was reliable and not overestimated by a small number of studies analyzed [30]. The sensitivity analysis was performed based on the type of telehealth intervention and the follow-up duration after intervention initiation. Further assessment to determine the source of heterogeneity included influence analysis. Influence analysis, or "leave-one-out" approach, assesses between-study heterogeneity by quantifying the influence of any one RCT on the overall summary estimate that is calculated in a meta-analysis, which is beneficial in identifying potential sources of error or bias introduced by an RCT [31-33]. Visual inspection of the funnel plot and the statistical Egger test were used to assess publication bias for outcomes with more than 10 RCTs to ensure adequate power for the test [29]. All analyses were conducted using the *meta* and *metafor* packages in R (version 4.2.2; R Foundation for Statistical Computing) with a $P \leq .05$ (95% CI) set as a statistically significant level.

Results

Literature Search Results

Our search string initially returned 230 records from PubMed, Cochrane Library, Web of Science, and ClinicalTrials.gov databases. Out of the 230 records, 61 were duplicates and the remaining 169 records' titles and abstracts were initially screened. After applying the inclusion and exclusion criteria, 145 records were excluded from this study. A total of 24 RCTs' reports were further assessed for eligibility, out of which 11 RCTs were excluded in secondary screening. Of these excluded RCTs, 4 included non-CVD participants, 5 did not report any primary outcomes of interest, and 2 did not incorporate an eligible UC. Based on the screening completed, a total of 13 RCTs were included in this study (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA flow diagram of screened and selected studies.

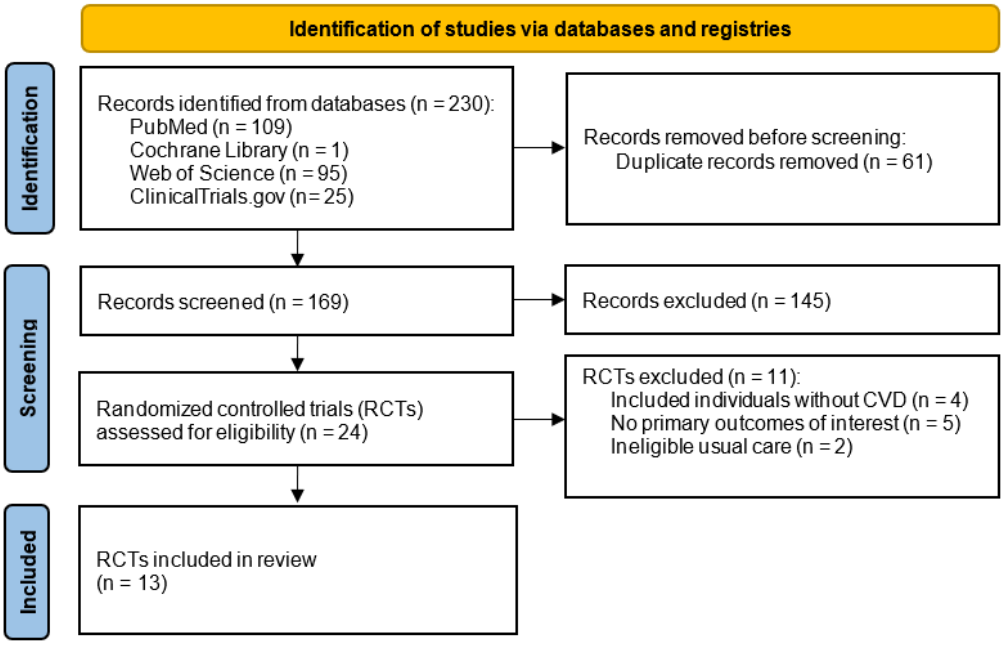

Study Characteristics

A summary of the study and participant characteristics of the included 13 RCTs [34-46] are shown in Multimedia Appendix 2 [34-46]. A total of 3013 participants were analyzed in this study. The mean age was 61.0 (SD 3.7) years and 18.5% (n=559) were women. Cardiovascular conditions in the included studies were a "general CVD diagnosis" (n=3), hypertension (n=3), individuals with CVD who experienced cardiac events (ie, myocardial infarction, acute coronary syndromes, angina, and revascularization, n=3), individuals with CVD who underwent cardiac procedures (ie, percutaneous coronary intervention and coronary artery bypass graft surgery, n=2), heart failure (n=1), and coronary heart disease (n=1). The trial locations included United States (n=4), Europe (n=4: Netherlands, United Kingdom, Spain, and Sweden), New Zealand (n=2), Australia (n=1), Korea (n=1), and Japan (n=1). Except for 1 trial [38], all RCTs took place in an urban setting. Intervention durations ranged from 4 weeks to 48 weeks. Follow-up durations after intervention initiation varied from 4 weeks (n=1), 6 weeks (n=1), 8 weeks (n=1), 12 weeks (n=2), 16 weeks (n=1), 24 weeks (n=4), 40 weeks (n=1), 48 weeks (n=1), to 48-56 weeks (n=1). The types of telehealth-based dietary interventions were based on apps (n=7), text messaging (n=3), telephone calls (n=1), web interaction (n=1), and an audio-visual media device (n=1). Of note, few RCTs included in this study involved RDNs in the dissemination of the telehealth-based dietary interventions (n=4) and UC (n=1).

Risk of Bias

The quality of the included RCTs, assessed based on the Cochrane risk of bias tool [29], is presented in Table 1. Out of the 13 included studies, 3 were assessed to be of high overall quality, while 9 were of moderate quality, and only 1 was of low quality. The majority of RCTs did not mask their participants; however, the absence of masking outcome assessors and unclear description of allocation concealment additionally contributed to the moderate and low assessments of overall quality.

XSL•FO
RenderX

Table 1. Overall quality assessment of included randomized controlled trials based on the Cochrane risk of bias tool.

Study	Random sequence generation	Allocation concealment	Masking of participants	Masking of outcome assessors	Avoided incomplete or selective reporting	Free from other external sources of bias	Overall quality
Friedberg et al [34]	++[a]	++	–[b]	++	++	++	High
Chow et al [35]	++	++	–	++	++	++	High
Dale et al [36]	++	++	–	–	++	++	Moderate
Eyles et al [37]	++	++	–	–	++	++	Moderate
Barnason et al [38]	++	+[c]	–	–	++	++	Moderate
Choi et al [39]	++	+	–	–	++	–	Low
Engelen et al [40]	++	–	–	–	++	++	Moderate
Dorsch et al [41]	++	+	–	–	++	++	Moderate
Riches et al [42]	++	++	–	–	++	++	Moderate
Bae et al [43]	++	++	–	++	++	++	High
Peydró et al [44]	++	++	–	++	++	–	Moderate
Michelsen et al [45]	++	++	–	–	++	++	Moderate
Nagatomi et al [46]	++	+	–	–	++	++	Moderate

[a]Low risk of bias.

[b]High risk of bias.

[c]Moderate risk of bias.

Meta-Analysis

A summary of the pooled estimates for all outcomes is reported in Table 2. Telehealth-based dietary interventions significantly reduced systolic blood pressure (MD –2.74, 95% CI –4.93 to –0.56) and LDL cholesterol (SMD –0.11, 95% CI –0.19 to –0.03) when compared to UC among individuals with CVD. The pooled estimates for weight (SMD –0.15, 95% CI –0.34 to 0.04), BMI (MD –0.01, 95% CI –1.70 to 1.68), diastolic blood pressure (MD –1.29, 95% CI –2.85 to 0.28), and levels for total cholesterol (SMD –0.10, 95% CI –0.28 to 0.08), HDL (SMD –0.10, 95% CI –0.22 to 0.01), and triglycerides (SMD –0.07, 95% CI –0.32 to 0.18) showed no significant difference among individuals with CVD who received telehealth-based dietary interventions versus UC. The forest plots and pooled estimate results for individual clinical outcomes are presented in Figures 2-9. The feasibility of intervention groups of the included RCTs did not significantly differ from UC groups (risk ratio 0.99, 95% CI 0.95 to 1.02), as shown in Figure 10, indicating similar feasibility between the 2 groups.

XSL•FO
RenderX

Table 2. Summary of meta-analysis results.

Outcomes	Number of studies (n=13), n (%)	Participants, n	Statistical methods	Effect estimates (95% CI)	Heterogeneity, I^2 (%)[a] (*P* value)
Weight	4 (31)	425	SMD[b] (IV[c], random, 95% CI)	–0.15 (–0.34 to 0.04)	0 (.49)
BMI	6 (46)	2070	MD[d] (IV, random, 95% CI)	–0.01 (–1.70 to 1.68)	92 (<.01)
Systolic blood pressure	10 (77)	2866	MD (IV, random, 95% CI)	–2.74 (–4.93 to –0.56)	64 (<.01)
Diastolic blood pressure	7 (54)	1388	MD (IV, random, 95% CI)	–1.29 (–2.85 to 0.28)	13 (.33)
Total cholesterol	6 (46)	1321	SMD (IV, random, 95% CI)	–0.10 (–0.28 to 0.08)	47 (.09)
LDL[e] cholesterol	6 (46)	2170	SMD (IV, random, 95% CI)	–0.11 (–0.19 to –0.03)	21 (.27)
HDL[f] cholesterol	5 (38)	1208	SMD (IV, random, 95% CI)	–0.10 (–0.22 to 0.01)	0 (.82)
Triglycerides	3 (23)	1018	SMD (IV, random, 95% CI)	–0.07 (–0.32 to 0.18)	66 (.05)
Feasibility	13 (100)	3013	RR[g] (M-H[h], random, 95% CI)	0.99 (0.95 to 1.02)	54 (<.01)

[a]Values for n not provided for I^2 percentages.
[b]SMD: standardized mean difference.
[c]IV: inverse-variance model.
[d]MD: mean difference.
[e]LDL: low-density lipoprotein.
[f]HDL: high-density lipoprotein.
[g]RR: risk ratio.
[h]M-H: Mantel-Haenszel model.

Figure 2. Forest plot of standardized mean differences in weight for telehealth-based dietary intervention (Int) and usual care (UC) groups [38-40,44].

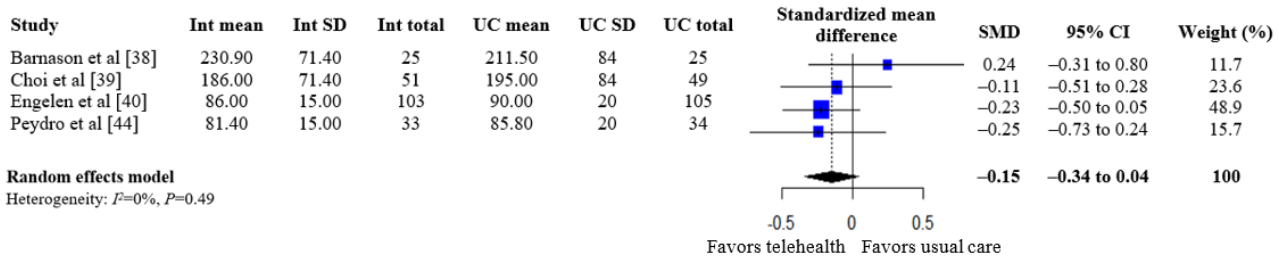

Figure 3. Forest plot of standardized mean differences in body mass index for telehealth-based dietary intervention (Int) and usual care (UC) groups [35,36,38-40,43].

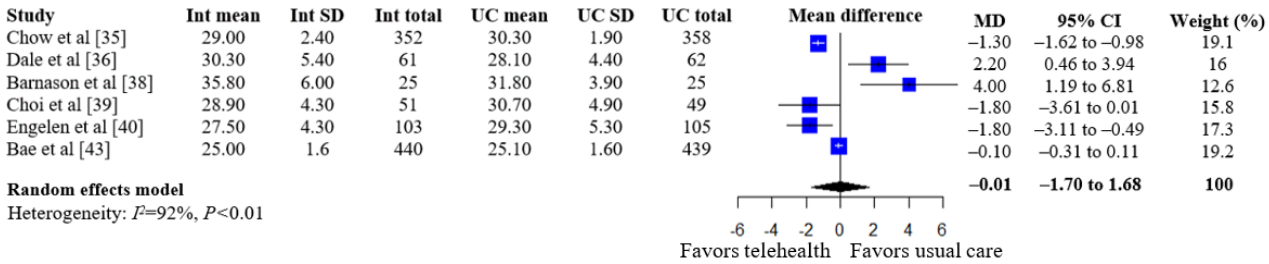

XSL•FO
RenderX

Figure 4. Forest plot of mean differences in systolic blood pressure for telehealth-based dietary intervention (Int) and usual care (UC) groups [34-37,39-43,45].

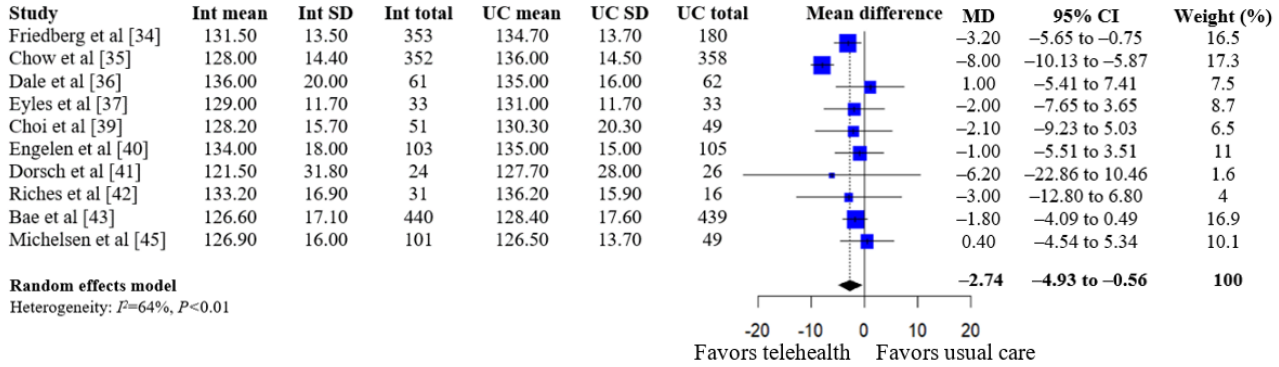

Figure 5. Forest plot of mean differences in diastolic blood pressure levels for telehealth-based dietary intervention (Int) and usual care (UC) groups [35,36,39-42,45].

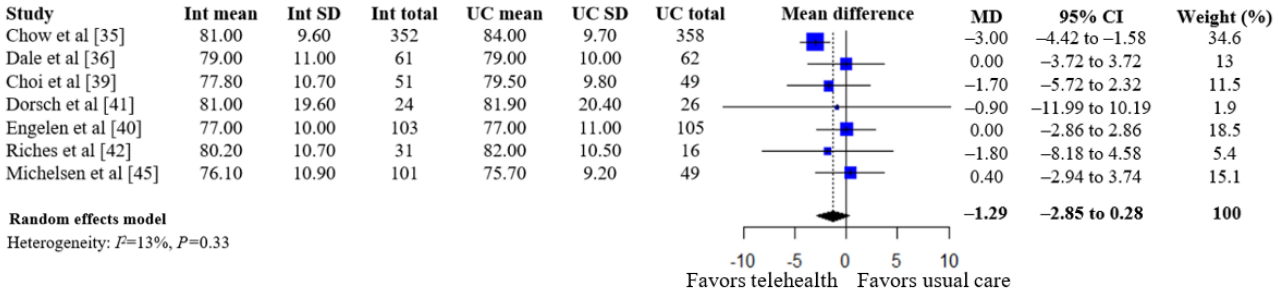

Figure 6. Forest plot of standardized mean differences in total cholesterol levels for telehealth-based dietary intervention (Int) and usual care (UC) groups [35,36,39,40,45,46].

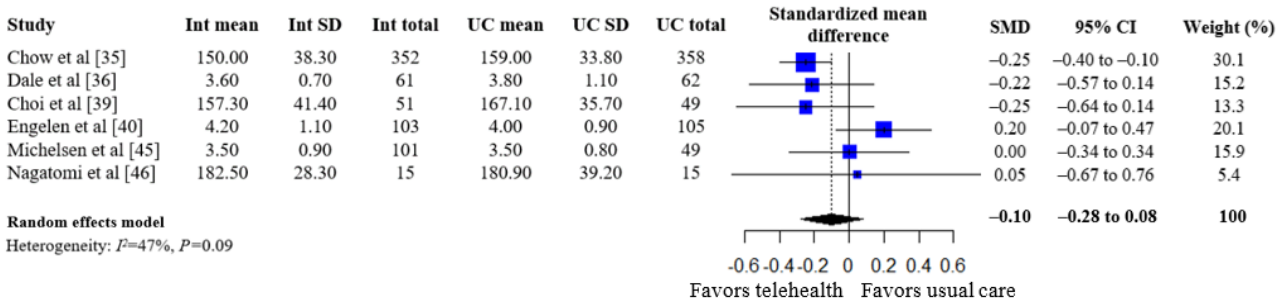

Figure 7. Forest plot of standardized mean differences in low-density lipoprotein cholesterol levels for telehealth-based dietary intervention (Int) and usual care (UC) groups [35,36,39,40,43,45].

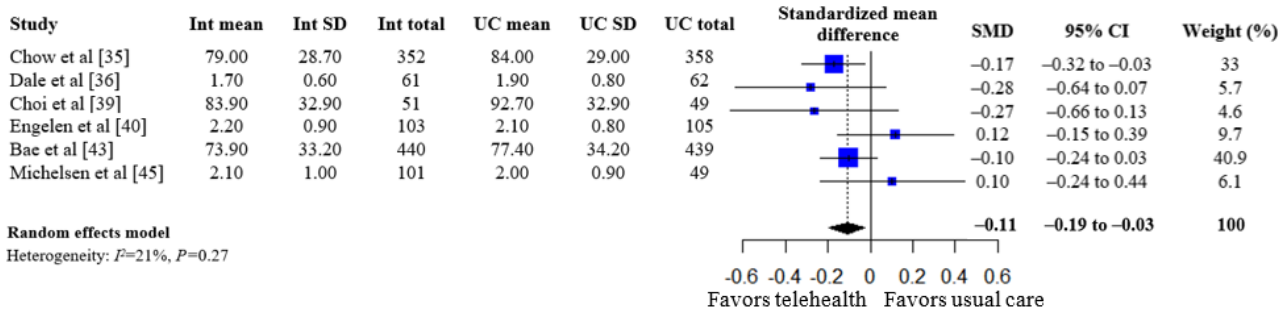

XSL•FO
RenderX

Figure 8. Forest plot of standardized mean differences in high-density lipoprotein cholesterol levels for telehealth-based dietary intervention (Int) and usual care (UC) groups [35,36,39,40,44].

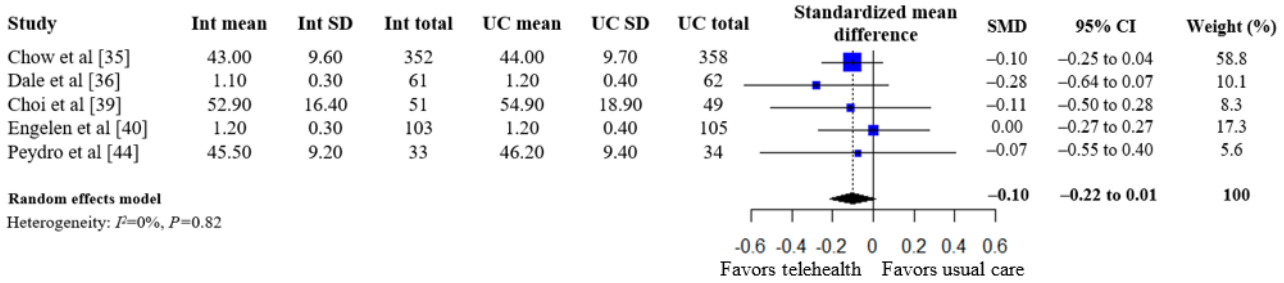

Figure 9. Forest plot of standardized mean differences in triglyceride levels for telehealth-based dietary intervention (Int) and usual care (UC) groups [35,39,40].

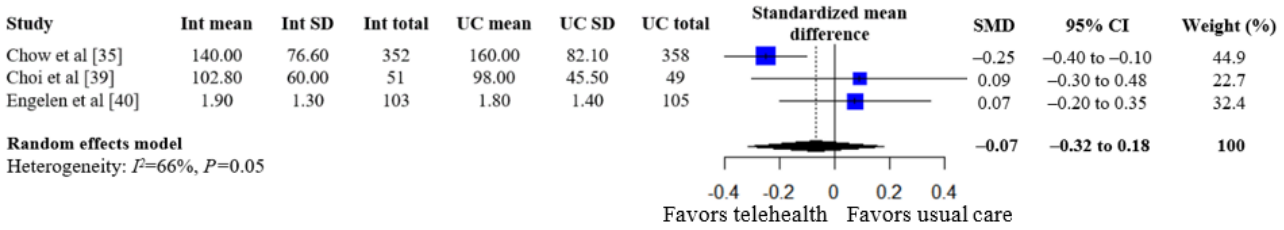

Figure 10. Forest plot of risk ratio (RR) for feasibility [34-46].

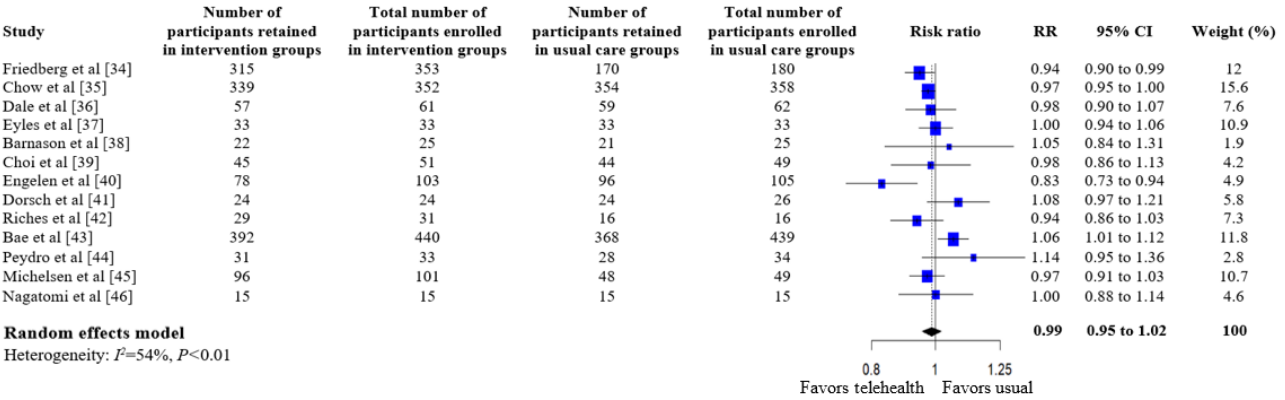

Funnel plot and the Egger statistic were assessed only for the feasibility outcome because its analysis included more than 10 RCTs, which ensured adequate power to test for publication bias [29]. Visual symmetry was noted in the funnel plot for the outcome, indicating no significant evidence for publication bias, as shown in Multimedia Appendix 3. This result was confirmed by the 1-tailed Egger regression test estimate for the feasibility outcome (t_{11}=0.51, P=.62).

Sensitivity Analyses

Sensitivity analyses were conducted for the systolic blood pressure and the feasibility (ie, feasibility) outcomes to investigate the significant heterogeneity detected. These outcomes were deemed appropriate for further assessment because they involved more than 7 RCTs, ensuring that the I^2 estimate was reliable and unbiased by the small number of studies analyzed [30]. The results for the sensitivity analyses on intervention type and follow-up duration after intervention initiation are shown in Multimedia Appendix 4 [34-37,39-43,45], Multimedia Appendix 5 [34-37,39-43,45], Multimedia Appendix 6 [34-46], and Multimedia Appendix 7 [34-46]. App-based (MD –1.33, 95% CI –4.41 to 1.74) and text-based (MD –3.44, 95% CI –8.61 to 1.72) interventions did not show any significant difference between telehealth-based dietary interventions and UC for systolic blood pressure (Multimedia Appendix 4 [34-37,39-43,45]). The RCTs that involved app-based interventions were nonsignificantly homogeneous (I^2=0%, P=.91); however, RCTs with text-based interventions had significant and considerable heterogeneity (I^2=89%, P<.01; Multimedia Appendix 4 [34-37,39-43,45]). The RCTs with a 24-week follow-up duration after intervention initiation showed a significant reduction in systolic blood pressure when compared to UC (MD –3.53, 95% CI –7.05 to –0.01) but they had significant and considerable heterogeneity (I^2=85%, P<.01; Multimedia Appendix 5 [34-37,39-43,45]).

The sensitivity analyses for the feasibility outcome consistently showed no significant difference between telehealth-based dietary interventions and UC groups (Multimedia Appendices 6 [34-46] and 7 [34-46]). The RCTs involving app-based interventions had insignificant and low heterogeneity (I^2=14%, P=.33), while RCTs with text-based interventions demonstrated significant and considerable heterogeneity (I^2=78%, P=.01; Multimedia Appendix 6 [34-46]). Within the feasibility outcome, RCTs with a 12-week follow-up duration after intervention

XSL•FO
RenderX

initiation were insignificantly homogeneous (I^2=0%, P=.86), while RCTs with a 24-week follow-up duration showed significant and considerable heterogeneity (I^2=74%, P<.01; Multimedia Appendix 7 [34-46]).

Influence Analysis

An influence analysis was conducted to determine whether any particular RCT impacted the heterogeneity for the systolic blood pressure and the feasibility outcomes, as shown in Multimedia Appendices 8 [34-37,39-43,45] and 9 [34-46]. The results for the systolic blood pressure outcome confirmed that the RCT by Chow et al [35] was the main source of heterogeneity. After omitting this RCT from the analysis, the MD summary estimate still showed telehealth-based dietary interventions to reduce systolic blood pressure significantly more than UC (MD –1.94, 95% CI –3.31 to –0.58); however, the I^2 percent of this outcome dropped from 64.4% (substantial heterogeneity) to 0% (low heterogeneity; Multimedia Appendix 8 [34-37,39-43,45]).

For the feasibility outcome, no single RCT was determined to be drastically different from the others in that its omission from analysis would reduce the outcome's moderate heterogeneity to the low category (Multimedia Appendix 9 [34-46]).

Discussion

Principal Findings

To the best of our knowledge, this study is the first to assess the effectiveness and feasibility of telehealth-based dietary interventions among individuals with CVD. Our study showed that telehealth-based dietary interventions improved levels of systolic blood pressure and LDL cholesterol when compared to UC. We further found no significant difference between telehealth-based dietary interventions and UC in improving weight, BMI, levels of diastolic blood pressure, total and HDL cholesterol, and triglycerides among individuals with CVD. There was no significant difference in feasibility between telehealth-based dietary interventions and UC.

Our finding of a positive clinical effect of telehealth-based dietary interventions on LDL cholesterol levels was in contrast to a previous meta-analysis conducted by Kelly et al [47] that found no significant differences in LDL cholesterol when comparing the effectiveness of telehealth-based dietary interventions to UC in changing dietary habits among chronic disease patients. However, our results of telehealth interventions improving systolic blood pressure compared to UC were confirmed by a meta-analysis by Kelly et al [47]. Plausible explanations for the clinical effectiveness of telehealth-based dietary interventions observed in our study are rooted in the unique benefits that telehealth-based technologies provide in modifying health behaviors. Telephone-based dietary telehealth interventions likely motivated participants through easily accessible, monthly scheduled, individually tailored counseling sessions, aimed to improve multiple components of their diet (ie, the intake of sodium, fat, fruit, and vegetables) [34]. According to Bandura's social cognitive theory (SCT), a motivated attitude can lead an individual to carry out healthy behaviors [48]. Telehealth-based dietary interventions that involved text messages were significantly more effective than UC in improving systolic blood pressure levels, possibly as a result of increasing and reinforcing participants' knowledge of healthy behaviors through receipt of regular messages, which may impact an individual's self-efficacy, and thereby leading to healthy behaviors per the SCT [36,48]. An RCT by Dale et al [36], which was included in our analysis, aimed to determine the effectiveness of a text message-based cardiac rehabilitation intervention among adults with coronary heart disease. Through the intervention in this trial, participants received recommendations to improve lifestyle behaviors, including diet [36]. Adherence to the recommendations and clinical biomarkers were measured as the primary and secondary outcomes, respectively [36]. The number of text messages sent to participants in the RCT's intervention arm was reduced from the first 12 weeks (7 messages per week) to the last 12 weeks (5 messages per week) of the study period [36]. As the number of text messages from the study team decreased per week, participants may have relapsed into unhealthy dietary behaviors [36]. This suggests a need for sustained implementation of telehealth-based dietary interventions as a continued source of knowledge and support of self-efficacy for participants' healthy behaviors and their clinical benefits. However, future research should confirm whether effectiveness is sustained if intervention durations are prolonged. Furthermore, findings from an RCT included in our study assessed the effectiveness and feasibility of a telehealth-based weight management intervention with cardiac rehabilitation in reducing the weight of overweight and obese individuals with CVD postcardiac procedure when compared to those who only received cardiac rehabilitation [38]. This RCT showed that participants who received nutritional intervention through an audio-visual media device in addition to cardiac rehabilitation had higher scores of perceived self-efficacy, knowledge, and skills in CVD self-management during follow-up than those who only received cardiac rehabilitation [38]. Greater self-efficacy, knowledge, and skill are all known constructs of the SCT, shown to help participants improve their health behaviors [48]. Lastly, a telehealth-based dietary intervention that involved smartphone apps provided notifications on heart-healthy alternatives to participants at the time of eating in a restaurant or purchasing food at a grocery store [41]. Such real-time app notifications provided participants with tailored nutritional information during critical decision-making time, which may be a key to the clinical benefits of telehealth-based dietary interventions observed in our study [41]. Various types of app-based interventions have been implemented through telehealth in the included RCTs that engage participants in different frequencies and timings during the day and target different constructs of behaviors, all aiming to improve cardiovascular health outcomes. For example, some app-based telehealth interventions may provide clinical benefits among individuals with CVD because they require participants to enter their food consumption into the app daily, which may improve motivation toward adhering to a heart-healthy diet [44,48]. According to the SCT, improved motivation can increase skills toward health behaviors by increasing participants' capacity for forethought and goal setting [48].

While all included RCTs provided a nutritional component to participants (UC and telehealth intervention groups), the benefits

XSL•FO
RenderX

of the telehealth-based dietary intervention over UC may imply the need for improved access to and standard of nutritional care for individuals diagnosed with CVD [34-46]. Participants enrolled in usual cardiac rehabilitation postcardiac events and procedures were provided with weekly educational sessions for at least 2 weeks, and nutrition was only 1 out of several other topics covered [36,38,44,45]. Once cardiac rehabilitation was complete, participants were typically not followed up by the rehabilitation unit, which indicates that nutrition was likely a very small component of the overall rehabilitation care given to the participants [44]. Other forms of UC involved regular outpatient clinic visits where participants received disease-specific pharmacological treatment, evaluation of cardiovascular risk factors, and general feedback on improving their lifestyle habits [35,37,40,43,46]. In some of the included RCTs, UC simply implied that participants received nutritional and disease-specific advice at 1 time point only (baseline) [34]. In one of the included RCTs, participants in the UC group received only a postal copy of publicly available booklets from the British Heart Foundation focused on reducing salt intake, titled “Cut Down on Salt” or “Taking Control of Salt” [42]. Apart from 1 RCT, where an RDN provided nutritional counseling to participants in the UC group [39], nutritionally trained health care professionals may not have been consistently involved in the dissemination of nutritional care in UC groups of the included RCTs.

While most included RCTs did not clearly state that they involved RDNs to disseminate the nutritional component to participants in intervention and UC groups, there is strong evidence suggesting that patients referred to RDNs have significant improvements in their clinical CVD risk factor outcomes [49-52]. According to the Academy of Nutrition and Dietetics, RDNs are increasingly using telehealth services in their clinical practice [53]. A qualitative study of 200 RDNs reported that medical nutrition therapy delivered through telehealth services positively impacted patient outcomes [54]. Previous studies suggest that telehealth services enable RDNs to better understand their patients’ nutritional habits since they were able to observe and discuss patients’ home environment more efficiently than in the clinical setting [54,55]. Furthermore, telehealth services may allow RDNs to facilitate group visits with a patient to involve their family members more conveniently during dietary education sessions, which has been shown to improve patient adherence to dietary advice and attendance to medical nutrition therapy appointments with RDNs [54,55].

Despite the benefits of telehealth-based dietary interventions on clinical CVD risk factors, some policy and patient-based barriers may prevent the implementation of telehealth use in clinical practice. First, each state has its own policies and license requirements for telehealth-based dietary practices in the clinical setting [53]. This may create inconsistencies in nutritional care provided to individuals with CVD nationwide. Therefore, more research showing the clinical benefits of telehealth-based dietary interventions is needed to demonstrate their effectiveness to policymakers in an attempt to standardize policies and strengthen dietetics practices across the United States [53]. Additionally, RDNs may not be able to use telehealth services to communicate with specific patients who have a limited understanding of communication technologies [53]. This patient-related barrier must be considered by RDNs on a case-by-case basis before engaging in virtual care.

Limitations

Our study is not without limitations. First, the included RCTs did not consistently provide participant demographic information on race, and levels of education and income, which limits our understanding of participant characteristics and the potential impact they may have on the overall results. Next, a majority of the included RCTs were of moderate overall quality, which may have impacted the overall results of our study. While we did not assess the certainty of evidence through the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) approach, we were able to assess the quality of each RCT using the Cochrane risk of bias tool, analyze publication bias when appropriate, and evaluate heterogeneity in the meta-analysis results to determine any source of bias. The included RCTs did not provide adequate data to analyze the effectiveness of telehealth-based dietary intervention on improving blood glucose levels or hemoglobin A_{1C}. While one of the assessed outcomes involved 13 RCTs, most other clinical CVD risk factor outcomes only involved 3-7 RCTs, which limited our analysis of heterogeneity and publication bias. Additionally, our sensitivity analysis was limited by the few number of RCTs available for each type of telehealth-based dietary intervention and the length of follow-up duration after intervention initiation. Therefore, more RCTs are needed to build a stronger evidence base on the effectiveness of different types of telehealth-based dietary interventions and for various follow-up durations to confirm our results. While we were able to determine that the RCT by Chow et al [35] was the source of heterogeneity for the systolic blood pressure outcome, this was less clear for the feasibility outcome. Nonetheless, our study contributes valuable knowledge on the effectiveness and feasibility of telehealth-based dietary interventions among individuals with CVD using robust meta-analysis techniques.

Conclusions

Telehealth-based dietary interventions significantly improved systolic blood pressure and LDL cholesterol when compared to UC among individuals with CVD. Future updates on this meta-analysis are needed to evaluate data from an increased number of RCTs for each type of telehealth intervention and follow-up duration in individuals with CVD.

Acknowledgments

None declared. No funding was received for this study, and there is no other financial information to disclose.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in the tables and figures of this published paper.

Authors' Contributions

R Trivedi and ELPS conceptualized the research questions, designed this study, interpreted the results, and wrote the original version of this paper. R Trivedi extracted the data and SE repeated extraction to confirm the results. R Trivedi managed and analyzed the data. SE, R Tackett, and HY interpreted the results and contributed to the discussion. All authors reviewed and commented on subsequent drafts of this paper, and approved the final version of this paper for publication. No generative AI was used in any portion of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File (Adobe PDF File), 73 KB - jmir_v26i1e49178_app1.pdf]

Multimedia Appendix 2
Characteristics of the included randomized controlled trials.
[DOCX File , 39 KB - jmir_v26i1e49178_app2.docx]

Multimedia Appendix 3
Funnel plot of risk ratio for feasibility.
[PNG File , 28 KB - jmir_v26i1e49178_app3.png]

Multimedia Appendix 4
Forest plot of mean differences in systolic blood pressure levels for telehealth-based dietary intervention (Int) and usual care (UC) groups by intervention type.
[PNG File , 235 KB - jmir_v26i1e49178_app4.png]

Multimedia Appendix 5
Forest plot of mean differences in systolic blood pressure levels for telehealth-based dietary intervention (Int) and usual care (UC) groups by follow-up duration after intervention initiation.
[PNG File , 164 KB - jmir_v26i1e49178_app5.png]

Multimedia Appendix 6
Forest plot of risk ratio (RR) for feasibility by intervention type.
[PNG File , 201 KB - jmir_v26i1e49178_app6.png]

Multimedia Appendix 7
Forest plot of risk ratio (RR) for feasibility by follow-up duration after intervention initiation.
[PNG File , 211 KB - jmir_v26i1e49178_app7.png]

Multimedia Appendix 8
Influence analysis of mean differences (MD) for systolic blood pressure levels.
[PNG File , 142 KB - jmir_v26i1e49178_app8.png]

Multimedia Appendix 9
Influence analysis of risk ratio (RR) for feasibility.
[PNG File , 220 KB - jmir_v26i1e49178_app9.png]

Multimedia Appendix 10
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 64 KB - jmir_v26i1e49178_app10.pdf]

XSL•FO
RenderX

References

1. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart disease and stroke statistics-2022 update: a report from the American Heart Association. Circulation 2022;145(8):e153-e639 [FREE Full text] [doi: 10.1161/CIR.0000000000001052] [Medline: 35078371]
2. Roth GA, Johnson CO, Abate KH, Abd-Allah F, Ahmed M, Alam K, et al. The burden of cardiovascular diseases among US states, 1990-2016. JAMA Cardiol 2018;3(5):375-389 [FREE Full text] [doi: 10.1001/jamacardio.2018.0385] [Medline: 29641820]
3. Bansilal S, Castellano JM, Fuster V. Global burden of CVD: focus on secondary prevention of cardiovascular disease. Int J Cardiol 2015;201(Suppl 1):S1-S7 [FREE Full text] [doi: 10.1016/S0167-5273(15)31026-3] [Medline: 26747389]
4. Pearson GJ, Thanassoulis G, Anderson TJ, Barry AR, Couture P, Dayan N, et al. 2021 Canadian Cardiovascular Society guidelines for the management of dyslipidemia for the prevention of cardiovascular disease in adults. Can J Cardiol 2021;37(8):1129-1150 [FREE Full text] [doi: 10.1016/j.cjca.2021.03.016] [Medline: 33781847]
5. Riegel B, Moser DK, Buck HG, Dickson VV, Dunbar SB, Lee CS, et al. Self-care for the prevention and management of cardiovascular disease and stroke: a scientific statement for healthcare professionals from the American Heart Association. J Am Heart Assoc 2017;6(9):e006997 [FREE Full text] [doi: 10.1161/JAHA.117.006997] [Medline: 28860232]
6. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines. J Am Coll Cardiol 2013;62(16):e147-e239 [FREE Full text] [doi: 10.1016/j.jacc.2013.05.019] [Medline: 23747642]
7. Vilme H, Duke NN, Muiruri C, Wordlaw L, Skinner AC. Using telehealth to disseminate primary, secondary, and tertiary CVD interventions to rural populations. Curr Hypertens Rep 2019;21(12):92 [FREE Full text] [doi: 10.1007/s11906-019-0998-8] [Medline: 31701257]
8. Healthcare access in rural communities. Rural Health Information Hub. 2021. URL: https://www.ruralhealthinfo.org/topics/healthcare-access [accessed 2022-01-03]
9. MacQueen IT, Maggard-Gibbons M, Capra G, Raaen L, Ulloa JG, Shekelle PG, et al. Recruiting rural healthcare providers today: a systematic review of training program success and determinants of geographic choices. J Gen Intern Med 2018;33(2):191-199 [FREE Full text] [doi: 10.1007/s11606-017-4210-z] [Medline: 29181791]
10. Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. J Community Health 2013;38(5):976-993 [FREE Full text] [doi: 10.1007/s10900-013-9681-1] [Medline: 23543372]
11. Pathman DE, Ricketts TC, Konrad TR. How adults' access to outpatient physician services relates to the local supply of primary care physicians in the rural southeast. Health Serv Res 2006;41(1):79-102 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00454.x] [Medline: 16430602]
12. Telehealth. Health Resources & Services Administration. 2021. URL: https://www.hrsa.gov/rural-health/telehealth [accessed 2022-02-02]
13. Strehle EM, Shabde N. One hundred years of telemedicine: does this new technology have a place in paediatrics? Arch Dis Child 2006;91(12):956-959 [FREE Full text] [doi: 10.1136/adc.2006.099622] [Medline: 17119071]
14. Wijesooriya NR, Mishra V, Brand PL, Rubin BK. COVID-19 and telehealth, education, and research adaptations. Paediatr Respir Rev 2020;35:38-42 [FREE Full text] [doi: 10.1016/j.prrv.2020.06.009] [Medline: 32653468]
15. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
16. Khera A, Baum SJ, Gluckman TJ, Gulati M, Martin SS, Michos ED, et al. Continuity of care and outpatient management for patients with and at high risk for cardiovascular disease during the COVID-19 pandemic: a scientific statement from the American Society for Preventive Cardiology. Am J Prev Cardiol 2020;1:100009 [FREE Full text] [doi: 10.1016/j.ajpc.2020.100009] [Medline: 32835347]
17. Levy CE, Silverman E, Jia H, Geiss M, Omura D. Effects of physical therapy delivery via home video telerehabilitation on functional and health-related quality of life outcomes. J Rehabil Res Dev 2015;52(3):361-370 [FREE Full text] [doi: 10.1682/JRRD.2014.10.0239] [Medline: 26230650]
18. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. What is cardiovascular disease? American Heart Association. 2017. URL: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease [accessed 2022-02-15]
20. Cardiovascular diseases (CVDs). World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2022-02-15]
21. Marx W, Kelly JT, Crichton M, Craven D, Collins J, Mackay H, et al. Is telehealth effective in managing malnutrition in community-dwelling older adults? A systematic review and meta-analysis. Maturitas 2018;111:31-46. [doi: 10.1016/j.maturitas.2018.02.012] [Medline: 29673830]
22. Klaassen B, van Beijnum BJF, Hermens HJ. Usability in telemedicine systems-a literature survey. Int J Med Inform 2016;93:57-69. [doi: 10.1016/j.ijmedinf.2016.06.004] [Medline: 27435948]

XSL•FO
RenderX

23. O'Grady HK, Farley C, Takaoka A, Mayens E, Bosch J, Turkstra L, et al. Retention in RCTs of physical rehabilitation for adults with frailty: a systematic review and meta-analysis. Trials 2022;23(1):235 [FREE Full text] [doi: 10.1186/s13063-022-06172-5] [Medline: 35346320]
24. Baillot A, St-Pierre M, Bernard P, Burkhardt L, Chorfi W, Oppert JM, et al. Exercise and bariatric surgery: a systematic review and meta-analysis of the feasibility and acceptability of exercise and controlled trial methods. Obes Rev 2022;23(9):e13480. [doi: 10.1111/obr.13480] [Medline: 35695385]
25. Mizrahi D, Wakefield CE, Fardell JE, Quinn VF, Lim Q, Clifford BK, et al. Distance-delivered physical activity interventions for childhood cancer survivors: a systematic review and meta-analysis. Crit Rev Oncol Hematol 2017;118:27-41. [doi: 10.1016/j.critrevonc.2017.08.008] [Medline: 28917267]
26. Toohey K, Chapman M, Rushby AM, Urban K, Ingham G, Singh B. The effects of physical exercise in the palliative care phase for people with advanced cancer: a systematic review with meta-analysis. J Cancer Surviv 2023;17(2):399-415. [doi: 10.1007/s11764-021-01153-0] [Medline: 35040076]
27. Miller KJ, Adair BS, Pearce AJ, Said CM, Ozanne E, Morris MM. Effectiveness and feasibility of virtual reality and gaming system use at home by older adults for enabling physical activity to improve health-related domains: a systematic review. Age Ageing 2014;43(2):188-195 [FREE Full text] [doi: 10.1093/ageing/aft194] [Medline: 24351549]
28. Gowing L, Ali R, White JM. Opioid antagonists with minimal sedation for opioid withdrawal. Cochrane Database Syst Rev 2009(4):CD002021. [doi: 10.1002/14651858.CD002021.pub3] [Medline: 19821290]
29. Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: Cochrane; 2019.
30. von Hippel PT. The heterogeneity statistic I(2) can be biased in small meta-analyses. BMC Med Res Methodol 2015;15:35 [FREE Full text] [doi: 10.1186/s12874-015-0024-z] [Medline: 25880989]
31. Patsopoulos NA, Evangelou E, Ioannidis JPA. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. Int J Epidemiol 2008;37(5):1148-1157 [FREE Full text] [doi: 10.1093/ije/dyn065] [Medline: 18424475]
32. Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. Res Synth Methods 2010;1(2):112-125. [doi: 10.1002/jrsm.11] [Medline: 26061377]
33. Willis BH, Riley RD. Measuring the statistical validity of summary meta-analysis and meta-regression results for use in clinical practice. Stat Med 2017;36(21):3283-3301 [FREE Full text] [doi: 10.1002/sim.7372] [Medline: 28620945]
34. Friedberg JP, Rodriguez MA, Watsula ME, Lin I, Wylie-Rosett J, Allegrante JP, et al. Effectiveness of a tailored behavioral intervention to improve hypertension control: primary outcomes of a randomized controlled trial. Hypertension 2015;65(2):440-446 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.114.03483] [Medline: 25403606]
35. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015;314(12):1255-1263 [FREE Full text] [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
36. Dale LP, Whittaker R, Jiang Y, Stewart R, Rolleston A, Maddison R. Text message and internet support for coronary heart disease self-management: results from the Text4Heart randomized controlled trial. J Med Internet Res 2015;17(10):e237 [FREE Full text] [doi: 10.2196/jmir.4944] [Medline: 26490012]
37. Eyles H, McLean R, Neal B, Jiang Y, Doughty RN, McLean R, et al. A salt-reduction smartphone app supports lower-salt food purchases for people with cardiovascular disease: findings from the SaltSwitch randomised controlled trial. Eur J Prev Cardiol 2017;24(13):1435-1444 [FREE Full text] [doi: 10.1177/2047487317715713] [Medline: 28631933]
38. Barnason S, Zimmerman L, Schulz P, Pullen C, Schuelke S. Weight management telehealth intervention for overweight and obese rural cardiac rehabilitation participants: a randomised trial. J Clin Nurs 2019;28(9-10):1808-1818 [FREE Full text] [doi: 10.1111/jocn.14784] [Medline: 30667588]
39. Choi BG, Dhawan T, Metzger K, Marshall L, Akbar A, Jain T, et al. Image-based mobile system for dietary management in an American cardiology population: pilot randomized controlled trial to assess the efficacy of dietary coaching delivered via a smartphone app versus traditional counseling. JMIR mHealth uHealth 2019;7(4):e10755 [FREE Full text] [doi: 10.2196/10755] [Medline: 31012860]
40. Engelen MM, van Dulmen S, Puijk-Hekman S, Vermeulen H, Nijhuis-van der Sanden MW, Bredie SJ, et al. Evaluation of a web-based self-management program for patients with cardiovascular disease: explorative randomized controlled trial. J Med Internet Res 2020;22(7):e17422 [FREE Full text] [doi: 10.2196/17422] [Medline: 32706708]
41. Dorsch MP, Cornellier ML, Poggi AD, Bilgen F, Chen P, Wu C, et al. Effects of a novel contextual just-in-time mobile app intervention (LowSalt4Life) on sodium intake in adults with hypertension: pilot randomized controlled trial. JMIR mHealth uHealth 2020;8(8):e16696 [FREE Full text] [doi: 10.2196/16696] [Medline: 32663139]
42. Riches SP, Piernas C, Aveyard P, Sheppard JP, Rayner M, Albury C, et al. A mobile health salt reduction intervention for people with hypertension: results of a feasibility randomized controlled trial. JMIR mHealth uHealth 2021;9(10):e26233 [FREE Full text] [doi: 10.2196/26233] [Medline: 34673535]
43. Bae JW, Woo SI, Lee J, Park SD, Kwon SW, Choi SH, et al. mHealth interventions for lifestyle and risk factor modification in coronary heart disease: randomized controlled trial. JMIR Mhealth Uhealth 2021;9(9):e29928 [FREE Full text] [doi: 10.2196/29928] [Medline: 34559058]

XSL•FO
RenderX

44. Peydró ED, Sevilla NS, Tuzón Segarra MT, Palau VM, Torrijos JS, Sales JC. A randomized controlled clinical trial of cardiac telerehabilitation with a prolonged mobile care monitoring strategy after an acute coronary syndrome. Clin Cardiol 2022;45(1):31-41 [FREE Full text] [doi: 10.1002/clc.23757] [Medline: 34952989]
45. Michelsen HO, Sjölin I, Bäck M, Garcia MG, Olsson A, Sandberg C, et al. Effect of a lifestyle-focused web-based application on risk factor management in patients who have had a myocardial infarction: randomized controlled trial. J Med Internet Res 2022;24(3):e25224 [FREE Full text] [doi: 10.2196/25224] [Medline: 35357316]
46. Nagatomi Y, Ide T, Higuchi T, Nezu T, Fujino T, Tohyama T, et al. Home-based cardiac rehabilitation using information and communication technology for heart failure patients with frailty. ESC Heart Fail 2022;9(4):2407-2418 [FREE Full text] [doi: 10.1002/ehf2.13934] [Medline: 35534907]
47. Kelly JT, Reidlinger DP, Hoffmann TC, Campbell KL. Telehealth methods to deliver dietary interventions in adults with chronic disease: a systematic review and meta-analysis. Am J Clin Nutr 2016;104(6):1693-1702 [FREE Full text] [doi: 10.3945/ajcn.116.136333] [Medline: 27935523]
48. Glanz K, Rimer BK, Viswanath K. Health Behavior and Health Education: Theory, Research, and Practice, 4th Edition. San Francisco, CA: Jossey-Bass; 2008.
49. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. J Am Coll Cardiol 2019;74(10):1376-1414 [FREE Full text] [doi: 10.1016/j.jacc.2019.03.009] [Medline: 30894319]
50. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. J Am Coll Cardiol 2019;73(24):e285-e350 [FREE Full text] [doi: 10.1016/j.jacc.2018.11.003] [Medline: 30423393]
51. Jacobson TA, Maki KC, Orringer CE, Jones PH, Kris-Etherton P, Sikand G, et al. National Lipid Association recommendations for patient-centered management of dyslipidemia: part 2. J Clin Lipidol 2015;9(6 Suppl):S1-S122.e1 [FREE Full text] [doi: 10.1016/j.jacl.2015.09.002] [Medline: 26699442]
52. Rozga M, Latulippe ME, Steiber A. Advancements in personalized nutrition technologies: guiding principles for registered dietitian nutritionists. J Acad Nutr Diet 2020;120(6):1074-1085. [doi: 10.1016/j.jand.2020.01.020] [Medline: 32299678]
53. Peregrin T. Telehealth is transforming health care: what you need to know to practice telenutrition. J Acad Nutr Diet 2019;119(11):1916-1920. [doi: 10.1016/j.jand.2019.07.020] [Medline: 31655720]
54. Brunton C, Arensberg MB, Drawert S, Badaracco C, Everett W, McCauley SM. Perspectives of registered dietitian nutritionists on adoption of telehealth for nutrition care during the COVID-19 pandemic. Healthcare (Basel) 2021;9(2):235 [FREE Full text] [doi: 10.3390/healthcare9020235] [Medline: 33672179]
55. Mehta P, Stahl MG, Germone MM, Nagle S, Guigli R, Thomas J, et al. Telehealth and nutrition support during the COVID-19 pandemic. J Acad Nutr Diet 2020;120(12):1953-1957 [FREE Full text] [doi: 10.1016/j.jand.2020.07.013] [Medline: 32792328]

Abbreviations

CVD: cardiovascular disease
GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
HDL: high-density lipoprotein
LDL: low-density lipoprotein
MD: mean difference
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RDN: registered dietitian nutritionist
SCT: social cognitive theory
SMD: standardized mean difference
UC: usual care

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 01.06.23; peer-reviewed by M Rozga; comments to author 29.06.23; revised version received 17.08.23; accepted 24.11.23; published 16.02.24.

Please cite as:
Trivedi R, Elshafie S, Tackett R, Young H, Sattler ELP
Effectiveness and Feasibility of Telehealth-Based Dietary Interventions Targeting Cardiovascular Disease Risk Factors: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e49178
URL: https://www.jmir.org/2024/1/e49178
doi:10.2196/49178
PMID:38363635

©Rupal Trivedi, Shaimaa Elshafie, Randall Tackett, Henry Young, Elisabeth Lilian Pia Sattler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of eHealth Interventions Promoting Physical Activity in Children and Adolescents: Systematic Review and Meta-Analysis

Irene Sequí-Domínguez[1,2,3], PhD; Iván Cavero-Redondo[1,4], PhD; Celia Álvarez-Bueno[1,5], PhD; Jose Francisco López-Gil[6], PhD; Vicente Martínez-Vizcaíno[1,4], MD; Carlos Pascual-Morena[1,3], PhD

[1]Health and Social Research Center, Universidad de Castilla-La Mancha, Cuenca, Spain
[2]Research Network on Chronicity, Primary Care and Health Promotion, Cuenca, Spain
[3]Facultad de Enfermería de Albacete, Universidad de Castilla-La Mancha, Albacete, Spain
[4]Facultad de Ciencias de la Salud, Universidad Autónoma de Chile, Talca, Chile
[5]Universidad Politécnica y Artística del Paraguay, Asunción, Paraguay
[6]One Health Research Group, Universidad de las Americas, Quito, Ecuador

Corresponding Author:
Vicente Martínez-Vizcaíno, MD
Health and Social Research Center, Universidad de Castilla-La Mancha
Edificio Melchor Cano, Centro de Estudios Socio-Sanitarios,
Santa Teresa Jornet s/n,
Cuenca, 16071
Spain
Phone: 34 969179100
Email: vicente.martinez@uclm.es

Abstract

Background: eHealth interventions have been postulated as a feasible, acceptable, and possibly effective tool to promote physical activity (PA) among children and adolescents; however, a comprehensive quantitative analysis of the effects of eHealth interventions promoting PA is lacking.

Objective: This study aims to conduct a systematic review and meta-analysis on experimental studies reporting the effects of eHealth interventions aimed at promoting PA on PA parameters and sedentary behavior parameters in children and adolescents.

Methods: The CENTRAL, MEDLINE, Embase, and Web of Science databases were searched from inception to February 2022 for randomized controlled trials that analyzed the effects of eHealth interventions aimed at promoting PA on PA and sedentary parameters in children and adolescents. The Hartung-Knapp-Sidik-Jonkman random effects method was used to determine the mean differences (MDs) with their respective 95% CIs. The risk of bias was assessed using the Risk of Bias 2 (RoB2; Cochrane) tool and its extension for cluster randomized controlled trials. The certainty of evidence was evaluated using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) tool.

Results: A total of 20 trials reporting the effects of different eHealth interventions aimed at promoting PA were included. Results for each parameter were as follows: counts per minute (MD −16.11 counts, 95% CI −122.76 to 90.53; k=3; n=402; I^2=69%; favoring control), steps per day (MD 593.46 steps, 95% CI −2102.27 to 3289.19; k=2; n=152; I^2=0%; favoring intervention [FI]), moderate to vigorous PA (MD −1.99 min/d, 95% CI −8.95 to 4.96; k=14; n=2336; I^2=86%; favoring control), light PA (MD 3.28 min/d, 95% CI −15.48 to 22.04; k=5; n=355; I^2=67%; FI), screen time (MD −31.48 min/d, 95% CI −68.62 to 5.65; k=5; n=904; I^2=0%; FI), and sedentary time (MD −33.12 min/d, 95% CI −57.27 to −8.97; k=8; n=819; I^2=75%; FI). Our results should be interpreted cautiously because of important limitations such as the scarcity of evidence, overall risk of bias, and low to very low certainty of evidence.

Conclusions: We did not find conclusive evidence regarding the impact of PA-targeted eHealth interventions on PA parameters, but the very low certainty of evidence suggests that eHealth interventions may reduce sedentary time in children and adolescents. Our results may have important scientific implications as they highlight that the rapid development of eHealth interventions to promote PA lacks robust supporting evidence.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD42020211020; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=211020

(J Med Internet Res 2024;26:e41649) doi:10.2196/41649

KEYWORDS

eHealth technologies; physical activity; sedentary behaviors; children; mobile phone

Introduction

Physical inactivity is considered one of the most important modifiable risk factors for noncommunicable diseases, as 1.6 million deaths per year can be attributed to insufficient physical activity (PA) [1]. Moreover, physical inactivity is associated with an increased risk of cardiovascular disease, hypertension, type 2 diabetes, multiple types of cancer, dementia, depression, as well as cardiovascular disease–related mortality and all-cause mortality in adults [2,3]. As childhood PA-related behaviors are associated with cardiometabolic risk and tend to endure from childhood through adolescence into adulthood [4,5], the promotion of PA at early ages is not only a cardiometabolic prevention strategy but can also be considered an adult prevention intervention [6-8].

According to the World Health Organization guidelines, children and adolescents (aged 5 to 17 years) should engage in an average of 60 minutes of moderate to vigorous PA (MVPA) per day and limit their sedentary time. In addition, these guidelines recommend that vigorous PA and muscle- and bone-strengthening activities should each be incorporated at least 3 days per week [9,10]. However, approximately 70% of this age group do not meet these recommendations [11]. Consequently, the intervention strategies for PA promotion in children and adolescents remain an important public health topic. Understanding waking activities as a continuum, where sedentary behaviors are waking activities characterized by low energy expenditure and performed in a sitting or reclining posture [12], interventions aimed at increasing energy expenditure during waking time may be able to increase PA while diminishing sedentary time.

Despite evidence supporting the substantial health benefits of engaging in adequate levels of PA and decreasing sedentary behaviors during childhood [13-15], interventions promoting PA targeting children and adolescents have shown serious shortcomings, including a lack of scalability, high costs, or difficulties in sustaining or implementing them [16]. Therefore, finding feasible, scalable, and effective interventions to encourage children and adolescents to adopt active lifestyles is a public health priority [17].

eHealth is defined as a cost-effective and safe use of information and communication technologies in support of health and health-related domains, enabling better communication between practitioners and patient practitioners, better monitoring and data management, and acting as a vehicle to deliver health information and interventions for prevention and care [18]. eHealth has been postulated as a feasible, acceptable, and possibly effective tool for promoting PA among children and adolescents [19]. The potential of eHealth interventions to engage young population not only resides in the fact that they are *digital natives* but also on the characteristics and diverse possibilities of eHealth delivery; these include features such as allowance for tailored feedback, monitoring, direct interaction, report on goals and track of progress, or access to a community [19,20].

Previous systematic reviews and meta-analyses studying the effects of mobile health (mHealth) technologies [21-28], SMS text message interventions [29], and digital interventions [30] in children and adolescents on PA and sedentary behavior parameters are inconclusive. However, when the horizon is broadened to analyze eHealth interventions aimed at promoting PA as a whole, they are postulated as a successful approach to increase PA; however, there is a lack of quantitative synthesis to support such conclusions [19,20]. This makes it difficult to make a pronouncement on the pillars of solid evidence as to whether the eHealth interventions that promote PA can be more effective than traditional PA promotion interventions.

Furthermore, owing to the scarcity and heterogeneity of available data, previous evidence [19-30] supports the necessity for more comparable studies to accumulate meaningful evidence.

The most recent study [22] showed the most promising results by only analyzing smartphone-based interventions, but the effect size pooled estimates show a mix of very different PA-related outcome measures (steps, counts, distance, and intensity of PA) that, in most cases, are not interchangeable with each other. In accordance with these inconsistencies, a comprehensive quantitative analysis of the effects of different eHealth interventions aimed at promoting PA on PA and sedentary behaviors is lacking.

Consequently, this systematic review and meta-analysis aims to comprehensively analyze the available scientific literature on the effect of eHealth interventions promoting PA on PA parameters (count/min, steps/d, light PA [LPA], and MVPA) and sedentary behavior parameters (screen time and sedentary time) in children and adolescents and to clarify the potential reasons for the conflicting results. For this purpose, we will analyze the effect of eHealth interventions promoting PA compared with a control (waitlist, nontreatment, or minimal alternative intervention), hence analyzing the combination of intervention content (PA promotion) and intervention delivery mechanism (eHealth).

Methods

Systematic Review

This systematic review and meta-analysis has been reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [31] and the recommendations of the Cochrane Handbook for Systematic Reviews of Interventions [32] and was registered in PROSPERO

XSL•FO
RenderX

(CRD42020211020), which has been modified to broaden the scope, changing the intervention scope (mHealth to eHealth) and the population (children to children and adolescents).

Search Strategy

A systematic search of the literature was conducted in the CENTRAL, Embase (via Scopus), MEDLINE (via PubMed), and Web of Science databases from inception of the database to February 2022. The search strategy was designed using the PICO (patient, intervention, comparison, and outcomes) strategy, as shown in Textbox 1. The search strategy was designed combining the following relevant terms: (1) "m-health," "ehealth," "ICT," "technology assisted," "mobile technology," "health technology," "internet based," "mobile health" and "mobile phone-based"; (2) "physical activity," "exercise," "fitness," "cardiorespiratory fitness," "aerobic fitness," "physical fitness," "step-count," "daily steps," "daily activity counts," "sedentar*" and "screen-time"; (3) "effect," "effecti*" and "evaluation"; and (4) "child*," "infant," "kids," "young," "adolescents." Moreover, a reverse search was performed by checking the reference lists of previous systematic reviews and meta-analyses for other relevant studies. Moreover, previous systematic reviews [19-30] were reviewed to perform a reverse search. Detailed search strategies for each database are included in Table S1 of Multimedia Appendix 1 [33-57].

Textbox 1. Summary of the eligibility criteria following the PICO (patient, intervention, comparison, and outcomes) strategy.

Population

- Inclusion: children and adolescents with mean age between 5 and 17 years without physical or psychological morbidities that would prevent the realization of the respective interventions
- Exclusion: studies including adult populations where disaggregation of data for children, adolescents, and adults was not possible

Intervention

- Inclusion: active monitoring or active interventions including eHealth technologies aimed to promote physical activity (PA) or reduce sedentary behaviors delivered to children and adolescents; multicomponent interventions including eHealth as the main component
- Exclusion: passive interventions (eg, nonfeedback monitoring), and multicomponent interventions not including eHealth as the main component or interventions aimed at physical exercise coaching

Comparator

- Inclusion: studies with waitlist, nonintervention, or usual care control groups
- Exclusion: studies without a control group or with a technology-delivered comparator

Outcome

- Inclusion: PA parameters (count/min, steps/d, light PA, and moderate to vigorous PA [MVPA]) and sedentary behavior parameters (screen time and sedentary time)
- Exclusion: eHealth intervention studies not involving PA or sedentary behavior parameters as a primary or secondary outcome

Study design

- Inclusion: experimental studies including randomized controlled trial (RCT) and feasibility studies with an RCT design
- Exclusion: nonexperimental study designs and non–peer-reviewed studies (eg, letters, comments, conference proceedings, reviews, and narrative articles)

Selection of Studies

The eligibility criteria according to the PICO strategy are summarized in Textbox 1. For inclusion in this systematic review and meta-analysis, studies had to meet the following criteria: (1) designed as experimental studies (randomized controlled trials [RCTs] or feasibility studies with an RCT design); (2) interventions delivered to children and adolescents with a mean age between 5 and 17 years without physical or psychological morbidity that would prevent the respective interventions; (3) include active monitoring or active interventions with eHealth technologies as a main component aimed at promoting PA or reducing sedentary behaviors; (4) compared with a waitlist, no to minimal intervention, or usual care control group; and (5) report its effects on PA parameters (count/min, steps/d, LPA, and MVPA) and sedentary behavior parameters (screen time and sedentary time). The study population was limited to ages between 5 and 17 years to use homogeneous samples in physical and psychological development consistent with those used in different PA recommendations [9,10,58].

Studies were included when PA and sedentary behavior parameters were reported according to objective or subjective measures using established thresholds or cut points, not necessarily in the same units. When data were reported using both objective and subjective measures, the objective measures were prioritized. When trials reported multiple follow-up points, the longest available follow-up period was prioritized.

Studies were excluded if they were not written in English or Spanish or if they included adult populations where disaggregation of data for children, adolescents, and adults was not possible. Crossover studies were also excluded because of

XSL•FO
RenderX

potential carryover effects resulting from a PA intervention that is not unreasonable to expect to continue for a prolonged period, which limited generalizability. Screening and trial selection were conducted independently by 2 reviewers (ISD and ICR), and disagreements were resolved by consensus.

Data Extraction and Quality Assessment

The following information was extracted from the included studies: (1) study reference, (2) country, (3) study design, (4) population characteristics (sample size, percentage of female participants, mean age, and type of population), (5) outcome variables included, and (6) type of eHealth intervention.

The Cochrane Collaboration tool for assessing the Risk of Bias 2 (RoB2; version 2) [59] was used to assess the potential bias of the included RCTs. Five bias domains were reviewed: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Studies could be rated as "low risk of bias" if all domains are classified as "low risk," "some concerns" if there is at least 1 domain rated as "some concern," and "high risk of bias" if there is at least 1 domain rated as "high risk" or ≥3 domains rated as "some concerns." For cluster RCT, a specific version of the RoB2 tool with additional considerations was used [60], which includes a modification in the first domain to assess the identification or recruitment bias. A risk of bias (RoB) analysis was performed by analyzing the allocation to the intervention in both cases.

Data extraction and quality assessment were independently performed by 2 reviewers (ISD and ICR). Inconsistencies were resolved by consensus or by involving a third researcher (CPM).

Grading the Quality of Evidence

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework was used to evaluate the certainty of evidence provided by this meta-analysis on the different outcomes [61]. Each outcome (counts/min, steps/d, LPA, MVPA, sedentary time, and screen time) could be scored as high, moderate, low, and very low evidence value, depending on the design of the studies, RoB, inconsistency, indirect evidence, imprecision, and publication bias. In this sense, some factors could increase or decrease the quality of the evidence score as follows: (1) RoB (−1 when <75% of the analyzed studies were at low RoB); (2) inconsistency (−1 when I^2>50%); (3) indirect evidence related to indirect population, intervention, control, or outcomes; (4) imprecision related to wide CIs; and (5) publication bias (−1 when it exists).

Statistical Analyses

When extracting data from the same trial, priority was given to the most recent and complete objectively measured data. For the statistical analysis, means before and after for both intervention and control groups (longest time after the intervention reported) with their respective SDs were preferably extracted. Data that did not meet these criteria as means before and after and SEs or mean differences (MDs) were extracted and converted. When data were reported as median (IQR), they were converted to mean (SD) following the most correct method according to the sample size devised by Hozo et al [62]. Furthermore, when primary data were reported in different units, unit conversions were performed.

To minimize *unit-of-analysis errors* when analyzing multiarm trials, similar intervention arms were combined into 1 arm. When interventions were not sufficiently similar, multiple entries were maintained in the splitting control group according to the number of intervention arms. Split samples were also combined.

To check for the normality of each outcome, we evaluated whether the authors tested the normality of the outcome through specific tests, and we calculated the mean/SD ratio for each intervention group (IG) to evaluate skewness.

When cluster RCTs were included, considering that the unit of allocation is the cluster, we checked for proper cluster analysis to avoid *unit-of-analysis error* [63]. When cluster RCTs were not appropriately analyzed, approximately correct analyses were performed using the inflated SEs method [64].

As opposed to final scores, change scores were analyzed, and when change was not available, final scores were converted to change scores using the following formula: *([pre-T mean – post-T mean] – [pre-C mean – post-C mean])*. When studies did not report change scores SDs and the exact correlation coefficients were not available, we assumed a correlation coefficient of 0.5 and calculated change scores SDs using Comprehensive Meta-Analysis software (version 2.2.064; Biostat, Inc.). The Hartung-Knapp-Sidik-Jonkman random effects method was used to compute the pooled MD estimates with their respective 95% CIs in PA parameters (counts/min, steps/d, LPA, and MVPA) and sedentary behavior parameters (sedentary time and screen time) [65]. Only the most recent analyses were included for quantitative synthesis when performed on the same population to avoid sample duplication. Moreover, when studies presented results stratified by sex, age, or weight status, they were analyzed as pooled.

The heterogeneity of results across studies was evaluated using the I^2 statistic [30] considering the following values: might not be important (0%-30%), may represent moderate heterogeneity (30%-50%), substantial heterogeneity (50%-75%), or considerable heterogeneity (75%-100%). The corresponding P values were also taken into account [66]. Raw data and all data conversions are provided in the Multimedia Appendix 1.

Exploratory sensitivity analyses were performed to analyze the results when skewed data, high RoB data, and nonobjectively assessed data were removed. To assess the robustness of the pooled estimates and to detect whether any particular study accounted for a large proportion of heterogeneity, sensitivity analyses using the leave-one-out method were conducted. Furthermore, to analyze whether children with comorbidities could have influenced the results, an exploratory sensitivity analysis was performed.

Exploratory random effects meta-regression models were used to evaluate whether summary estimates were influenced by the percentage of female participants and the mean age of the participants. Exploratory multigroup analyses were performed based on the type of eHealth intervention (mobile phone app,

XSL•FO
RenderX

multicomponent intervention, telephone-delivered intervention, text messages, or web-based intervention).

Finally, to evaluate publication bias, visual examination of funnel plots and the regression asymmetry test proposed by Egger [67] were used, and a *P* value <.10 was considered statistically significant.

Statistical analyses were performed using STATA SE software (version 15; StataCorp) and the *metagen* package from R statistical software (version 4.1.2; R Foundation for Statistical Computing).

Results

Systematic Review

The systematic search, which included an inverse search, retrieved a total of 4087 articles. One study that had not been retrieved through the systematic search was included by contact with experts [33]. After removing duplicates, 146 articles were selected based on the title and abstract screening for a full content review. Finally, 25 reports of 20 studies [33-57] met the inclusion criteria and were included in the systematic review (Figure 1). One study [68] was classified as "awaiting assessment" because more information was needed to clarify whether it was a duplicate report of 1 of the included studies [49]. Excluded studies and their respective reason for exclusion are displayed in Table S2 of Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the systematic literature search strategy. WOS: Web of Science.

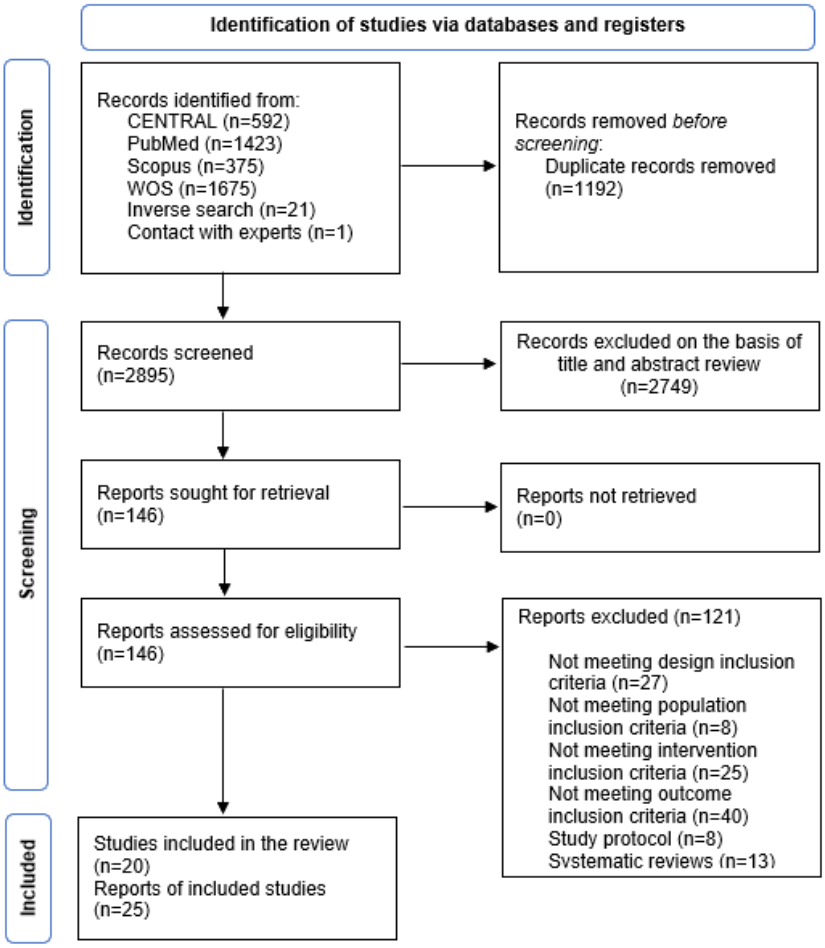

Table 1 summarizes the characteristics of the included studies. All studies meeting the inclusion criteria had an RCT design; 3 studies were pilot randomized trials [43,48,53], and 6 studies [36,37,41,50,52,54] had cluster RCT designs. Studies were published between 2006 and 2020 and were conducted in 7 different countries: Australia (3/20, 15%), Belgium (1/20, 5%), Finland (1/20, 5%), New Zealand (2/20, 10%), the Netherlands (3/20, 15%), the United Kingdom (1/20, 5%), and the United States (7/20, 35%), and 1 study [50] was conducted in 6 different European countries (Austria, Belgium, Crete, Germany, Greece, and Sweden).

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study type and study name	Reference	Design	Country	Population characteristics			Outcome	
				Type of population	Participants, n (female)	Age (y), mean (SD)	PA[a] parameters	Sedentary parameters
App intervention								
AIMFIT[b]	Direito et al [34], 2015	RCT[c]	New Zealand	Healthy adolescents	IG[d]1: 17 (9); IG2: 16 (10); CG[e]: 18 (10)	IG1: 15.78 (1.11); IG2: 15.69 (1.04); CG: 15.55 (1.32)	Counts per minute, PAQ-A[f], PACES[g], light PA, moderate PA, vigorous PA, and MVPA[h] (min/d)	Sedentary time (min/d)
Exergame intervention								
__[i]	Staiano et al [35], 2017	RCT	United States	Adolescents with overweight and obesity	IG: 19 (19); CG: 18 (18)	IG: 15.30 (1.3); CG: 16.10 (1.3)	Light PA, moderate PA, and vigorous PA (min/d)	Sedentary time (min/d)
GameSquad	Staiano et al [57], 2018	RCT	United States	Children with overweight and obesity	46 (21); IG: 23; CG: 23	11.2 (0.8)	MVPA (min/d)	—
Multicomponent intervention								
Switch-off 4 Healthy Minds	Babic et al [36], 2016	Clustered RCT	Australia	Healthy adolescents	IG: 167 (107); CG: 155 (104)	IG: 14.47 (0.6); CG: 14.33 (0.5)	MVPA (min/d)	Screen time (min/d)
Nutrition and Enjoyable Activity for Teen Girls	Dewar et al [37], 2013 Dewar et al [38], 2014; Lubans et al [39], 2012	Clustered RCT	Australia	Healthy adolescents	IG: 178 (178) CG: 179 (179)	IG: 13.2 (0.5) CG: 13.2 (0.4)	Counts per minute and MVPA (min/d, %)	Screen time (min/d) and sedentary time (min/d)
Fit4Life	Huang et al [40], 2014	RCT	United States	Children surviving acute lymphoblastic leukemia	IG: 19 (7); CG: 19 (8)	IG: 13 (1.74); CG: 13 (1.74)	MVPA (min/d)	—
Avoiding Screen-time ATLAS[j]	Lubans et al [41], 2016; Smith et al [42], 2014	Clustered RCT	Australia	Healthy adolescents	IG: 181 (0); CG: 180 (0)	IG: 12.7 (0.5); CG: 12.7 (0.5)	Counts per minute and MVPA (%)	Screen time (min/d)
—	Mendoza et al [43], 2017	Pilot RCT	United States	Childhood cancer survivors	IG: 29 (17); CG: 30 (18)	IG: 16.9 (1.5); CG: 16.3 (1.5)	MVPA (min/d)	Sedentary time (min/d)
Pace-Internet for Diabetes Prevention Intervention (PACEi-DP)	Patrick et al [44], 2013	RCT	United States	Adolescents with overweight and obesity at risk of type 2 DM[k]	IG1 (W[l]): 26 (16); IG2 (WSMS[m]): 14 (12); IG3 (WG[n]): 26 (18); CG: 25 (18)	IG1 (W): 14.1 (1.4); IG2 (WSMS): 14.3 (1.8); IG3 (WG): 14.3 (1.5); CG: 14.5 (1.5)	MVPA (min/wk)	Sedentary time (h/d)

XSL•FO
RenderX

Study type and study name	Reference	Design	Country	Population characteristics			Outcome	
				Type of population	Participants, n (female)	Age (y), mean (SD)	PA[a] parameters	Sedentary parameters
—	Ruotsalainen et al [33], 2015	RCT	Finland	Adolescents with overweight and obesity	IG1 (FB+Act[o]): 15 (10); IG2 (FB): 16 (11); CG: 15 (11)	IG1 (FB+Act): 14.8 (0.8); IG2 (FB): 14.8 (0.8); CG: 15 14.7 (0.8)	Light PA (min/d) and MVPA (min/d)	Sedentary time (min/d)
—	Thompson et al [45], 2016	RCT	United States	Healthy adolescents	IG: 40 (17); CG: 40 (20)	14 to 17	Steps/d and MVPA (min/d)	—
Telephone-delivered interventions								
Healthy Eating and Activity Today study	Wright et al [46], 2013	RCT	United States	Children with obesity	IG: 24 (9); CG: 26 (12)	IG: 10.9 (1.3); CG: 10.5 (1.2)	—	Screen time (h/d)
Text messages intervention								
—	Newton et al [47], 2009	RCT	New Zealand	Adolescents with type 1 DM	IG: 38 (22); CG: 40 (20)	14.4 (2.37)	Steps/d and MVPA (min/wk)	—
—	Shapiro et al [48], 2008	Pilot RCT	United States	Healthy adolescents	IG: 18 (13); CG: 22 (13)	IG: 8.4 (2.3); CG: 8.5 (2.3)	Steps/d	Screen time (min/d)
Web-based intervention								
Rheumates@Work	Armbrust et al [49], 2017	RCT	The Netherlands	Children with juvenile idiopathic arthritis	IG: 28 (21); CG: 21 (12)	IG: 9.9 (0.78) CG: 10 (0.45)	Light PA (min/d) and MVPA (min/d)	—
Activ-O-Meter HELENA study	de Bourdeaudhuij et al [50], 2010; Cook et al [51], 2014	Clustered RCT	Europe	Healthy adolescents	1050 (515); IG: 581 (NR[p]); CG: 469 (NR)	14.5 (1.4)	MVPA (min/wk)	—
FATaint-PHAT	Ezendam et al [52], 2012	Clustered RCT	The Netherlands	Healthy adolescents	IG: 485 (198); CG: 398 (200)	IG: 12.7 (0.7); CG: 12.6 (2.7)	Steps/d	Screen time (min/d)
Families Reporting Every Step to Health	Guagliano et al [53], 2020	Pilot RCT	United Kingdom	Healthy children	IG: 30 (15); CG: 29 (14)	IG: 10.1 (2.8); CG: 8.9 (0.6)	MVPA (min/d)	Sedentary time (min/d)
—	Haerens et al [54], 2006; Haerens et al [55], 2007	Clustered RCT	Belgium	Healthy children	IG1: 1194 (479); IG2: 911 (142); CG: 735 (432)	IG1: 13 (0.8); IG2: 13.2 (0.9); CG: 12.9 (0.7)	Light PA (min/d) and MVPA (min/d)	Sedentary time (min/d)

XSL•FO
RenderX

Study type and study name	Reference	Design	Country	Population characteristics			Outcome	
				Type of population	Participants, n (female)	Age (y), mean (SD)	PA[a] parameters	Sedentary parameters
—	Slootmaker et al [56], 2010	RCT	The Netherlands	Healthy adolescents	IG: 41 (26); CG: 46 (29)	IG: 15.4 (1.1); CG: 14.9 (1.3)	Light PA (min/wk) and MVPA (min/wk)	Sedentary time (min/wk)

[a]PA: physical activity.
[b]AIMFIT: Apps for Improving Fitness.
[c]RCT: randomized controlled trial.
[d]IG: intervention group.
[e]CG: control group.
[f]PAQ-A: Physical Activity Questionnaire for Adolescents.
[g]PACES: perceived enjoyment using the Physical Activity Enjoyment Scale.
[h]MVPA: moderate to vigorous physical activity.
[i]Not available.
[j]ATLAS: Active Teen Leaders Avoiding Screen-time.
[k]DM: diabetes mellitus.
[l]W: website only group.
[m]WSMS: website and SMS group.
[n]WG: website, monthly group sessions, and follow-up calls group.
[o]FB+Act: Facebook-delivered lifestyle counseling+physical activity self-monitoring group
[p]NR: not reported.

The sample size of the included studies ranged from 37 to 2840 participants (with a mean of 2436/4978, 58.7% female participants), whose ages ranged from 8.4 to 17 years; these studies were performed in children 6-12 years (5/20, 25%) and adolescents 13-17 years (15/20, 75%). Most studies included healthy children or adolescents (11/20, 55%), although several studies included children with physical health problems (overweight or obesity, 5/20, 25%; survivors of cancer, 2/20, 10%; and type 1 diabetes mellitus, 1/20, 5%).

Studies were classified according to the type of eHealth delivering the intervention (mobile apps, 1/20, 5%; exergames, 2/20, 10%; multicomponent interventions, 8/20, 40%; telephone-delivered interventions, 1/20, 5%; text messages, and 2/20, 10%; web-based interventions, 6/20, 30%). With regard to intervention content, most interventions were based on self-determination and social cognitive theories (5/20, 25% and 6/20, 30%, respectively), and included, in most cases, several of the following components: coaching (9/20, 45%), counseling (8/20, 40%), parental counseling (2/20, 10%), monitoring (2/20, 10%) and self-monitoring (3/20, 15%), group sessions (5/20, 25%), and goal setting (4/20, 20%). The duration of the interventions ranged from 8 to 96 weeks, and the dosage was left undetermined owing to the characteristics of the intervention. Most interventions required clinicians' assistance (13/20, 65%); 20% (4/20) of the interventions were autonomous and 15% (3/20) of the interventions involved the school setting.

For the comparison, most studies used usual behavior (6/20, 30%), waitlist control (5/20, 25%), or no intervention (4/20, 20%). However, minimal interventions such as printed information, 1 group session, or generic advice were also used as control in 25% (5/20) of the studies.

Outcomes measured with objective measures such as accelerometers or pedometers were steps (4/20, 20%), counts (3/20, 15%), LPA (6/20, 30%), MVPA (13/20, 65%), sedentary time (6/20, 30%), and screen time (2/20, 10%). Several outcomes were also measured with self-reported instruments as questionnaires (screen time, 6/20, 30%; sedentary time, 2/20, 10%; or MVPA, 2/20, 10%).

Further details on intervention characteristics and outcome measurement methods are shown in Table S3 in Multimedia Appendix 1.

RoB2 Tool

According to the Cochrane Collaboration's tool for assessing the RoB, RoB2 for RCTs, 43% (6/14) of the trials showed a high RoB for overall bias, and 43% (6/14) the trials showed some concerns. By domain, no trial showed a high risk for the randomization process, missing outcome data, or selection of the reported result. Only 1 trial showed high risk for assignment to the intervention. The most compromised domain was measurement of the outcome, where 43% (6/14) showed a high RoB, probably owing to the nature of the interventions (Figure S1 in Multimedia Appendix 1).

Cluster RCTs RoB was assessed using a specific version of the RoB2 tool with additional considerations. According to the results, only 6% (1/16) of the trials showed a high RoB owing to a compromise in the measurement of the outcome domain,

XSL•FO
RenderX

whereas the rest of the trials showed some concerns for the overall RoB (Figure S2 in Multimedia Appendix 1).

Quality of Evidence

According to the GRADE summary of findings, the main limitations were the high RoB, the *substantial* heterogeneity, and imprecision among the RCTs included. Therefore, the level of certainty of the findings was very low for counts per minute, LPA, and sedentary time and low for steps per day, MVPA, and screen time (Table S4 in Multimedia Appendix 1).

Meta-Analyses

When analyzing trials with several reports, the most recent and complete objectively measured data were considered for the quantitative synthesis. One study did not report sufficient numerical data to be included in the meta-analysis [57]. Only 1 of the included cluster RCTs [52] did not report an appropriate analysis to avoid "unit-of-analysis error" and did not report sufficient data to perform the inflated SEs method and was therefore excluded from the meta-analysis. Raw data and all conversions are reported in Multimedia Appendix 1.

To minimize the unit-of-analysis errors, studies with similar intervention arms were combined into 1 arm, as in the AIMFIT [34] study (combining IG1 and IG2), the PACEi-DP study [44] (combining IG1, IG2, and IG3), the studies by Haerens et al [54,55] combining IG1 and IG2, and the study by Slootmaker et al [56] combining boys and girls that had been analyzed separately.

None of the included studies reported change score SDs, and the exact correlation coefficients were only available for 1 study [49] for MVPA and LPA outcomes. However, because of the specific characteristics of the population included in this study, such coefficients were not used, and we assumed a correlation coefficient of 0.5 for the calculation of change score SDs.

The pooled MD for counts per minute (Figure 2 [34,37,41,45,47]) was −16.11 counts per minute (95% CI −122.76 to 90.53), with the number of trials (k=3) and number of participants (n=402) favoring control. For steps per day (Figure 2), MD was 593.46 steps (95% CI −2102.27 to 3289.19; k=2; n=152; favoring intervention [FI]). Both had substantial and insignificant heterogeneity values (I^2=69% for counts/min I^2=0%, for steps/d). The trial by Shapiro et al [48] was excluded from the steps per day analysis owing to its lack of a control group for this parameter.

Figure 2. Forest plot showing the effect size for (A) counts per minute and (B) steps per day outcomes. Positive mean difference (MD) values favor intervention. Bold values highlight pooled effects (MD, 95% CI) and their respective prediction intervals. HK: Hartung-Knapp method.

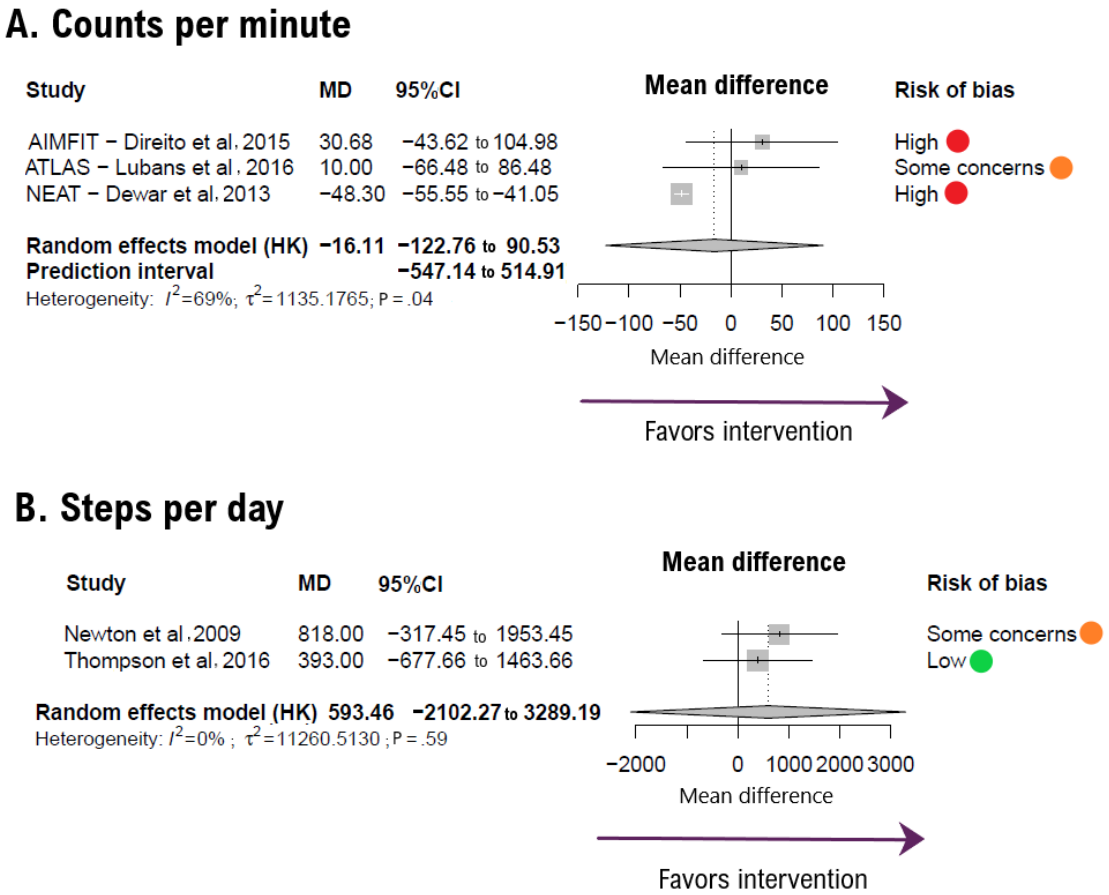

The pooled MD for MVPA (Figure 3 [33,34,36,39,40,43,45,47,49,50,53,54,56,57]) was −1.99 minute per day (95% CI −8.95 to 4.96; k=14; n=2336; favoring control). After extracting data for the analysis, 1 study [43] showed log transformed data for MVPA and was consequently excluded from the analysis. For LPA (Figure 3), MD was 3.28 minute per day (95% CI −15.48 to 22.04; k=5; n=355; FI). The heterogeneity values

XSL•FO
RenderX

were substantial to considerable (I^2=86% for MVPA and I^2=67%, for LPA).

For sedentary behavior parameters, the pooled MD for screen time (Figure 4 [33,34,36-38,41,43,44,46,48,53,54,56]) was −31.48 minute per day (95% CI −68.62 to 5.65; *k*=5; n=904, FI). For sedentary time (Figure 3), the MD was −33.12 minute per day (95% CI −57.27 to −8.97; *k*=8; n=819; FI). There were no substantial heterogeneity estimates (I^2=0% for screen time and I^2=75%, for sedentary time).

Figure 3. Forest plot showing the effect size for (A) moderate to vigorous physical activity (MVPA) and (B) light physical activity (LPA) outcomes. Positive mean difference (MD) values favor intervention. Bold values highlight pooled effects (MD, 95% CI) and their respective prediction intervals.

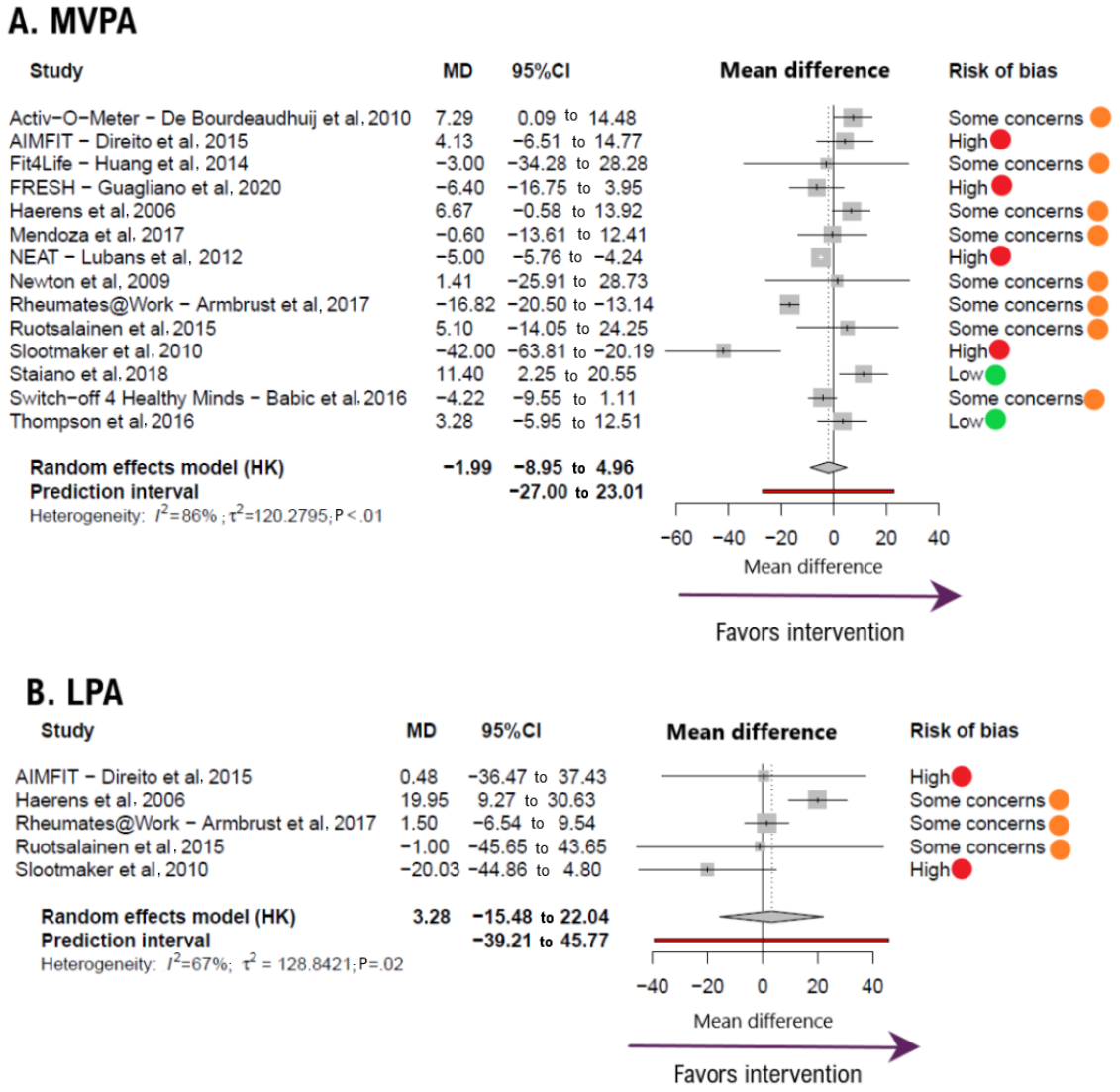

XSL•FO
RenderX

Figure 4. Forest plot showing the effect size for (A) screen time and (B) sedentary time outcomes. Negative mean difference (MD) values favor intervention. Bold values highlight pooled effects (MD, 95% CI) and their respective prediction intervals. HK: Hartung-Knapp method.

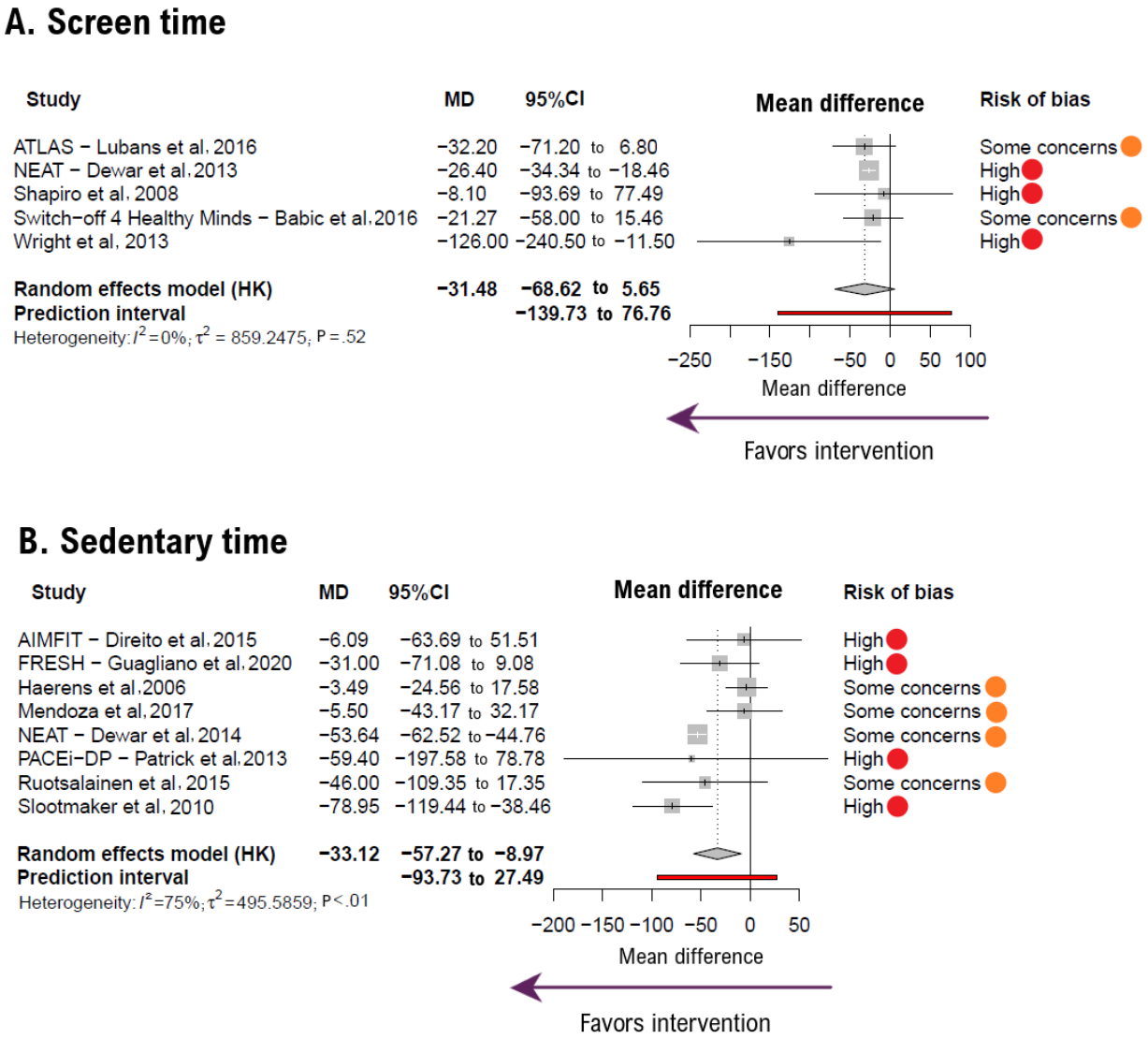

Sensitivity Analysis

After evaluating skewness to check for normality of primary data, evidence of skewness was found for several reports in steps per day, MVPA, screen time, and sedentary time analyses, and taking into account the nature of the data, a sensitivity analysis excluding skewed data was performed. Such analysis showed no significant statistical differences from the main analyses, except for the loss of statistical significance in the sedentary time MD (Table 2).

XSL•FO
RenderX

Table 2. Sensitivity analysis results.

Outcome	Number of trials	Pooled MD[a] estimates (95% CI)	I^2 (%)
Pooled MD estimates excluding skewed data			
Steps per day	*k*=1 [44]	393 (−677.66 to 1463.66)	__[b]
MVPA[c] (min/d)	*k*=3 [38,48,56]	−18.89 (−63.09 to 25.32)	95.9
Screen time (min/d)	*k*=1 [36]	*−48.30 (−55.55 to −41.05)*[d]	—
Sedentary time (min/d)	*k*=4 [33,37,54,56]	−36.69 (−93.97 to 20.58)	86.9
Pooled MD estimates excluding high risk of bias data			
Counts per minute	*k*=1 [40]	10 (−66.48 to 86.48)	—
MVPA (min/d)	*k*=10 [34,35,39,42,44,48,50,54,56,67]	0.54 (−6.07 to 7.14)	88.4
LPA[e] (min/d)	*k*=3 [48,54,67]	9.27 (−19.10 to 37.65)	73.3
Screen time (min/d)	*k*=2 [35,40]	−26.41 (−95.73 to 42.91)	0.0
Sedentary time (min/d)	*k*=4 [37,42,54,67]	−27.08 (−70.79 to 16.63)	86.8
Pooled MD estimates excluding nonobjectively measured data			
MVPA (min/d)	*k*=11 [33-35,38,39,42,44,48,53,54,67]	−1.33 (−7.02 to 4.35)	85.6
LPA (min/d)	*k*=4 [33,48,54,67]	8.57 (−8.37 to 25.51)	60.7
Screen time	*k*=0	—	—
Sedentary time (min/d)	*k*=6 [33,37,42,53,54,67]	*−26.45 (−51.80 to −1.10)*	79.8
Pooled MD estimates excluding converted data			
Counts per minute	*k*=1 [33]	10.00 (−66.48 to 86.48)	—
Steps per day	*k*=0	—	—
MVPA (min/d)	*k*=6 [33,34,39,50,53,67]	3.66 (−2.17 to 9.50)	7.3
LPA (min/d)	*k*=3 [33,54,67]	14.94 (−11.52 to 41.40)	0.0
Screen time (min/d)	*k*=2 [45,47]	−60.32 (−804.46 to 683.81)	61.7
Sedentary time (min/d)	*k*=4 [33,53,54,67]	−15.23 (−44.22 to 13.75)	0.0
Pooled MD estimates including only healthy population			
Steps per day	*k*=1 [44]	393 (−677.66 to 1463.66)	—
MVPA (min/d)	*k*=8 [33,35,38,44,50,53,54,56]	−2.86 (−14.45 to 8.74)	81.4
LPA (min/d)	*k*=3 [33,54,56]	2.66 (−50.21 to 55.54)	77.4
Screen time (min/d)	*k*=4 [35,36,40,47]	*−26.20 (−31.06 to −21.34)*	0.0
Sedentary time (min/d)	*k*=5 [33,37,53,54,56]	−35.81 (−74.77 to 3.14)	82.9

[a]MD: mean difference.

[b]Not available.

[c]MVPA: moderate to vigorous physical activity.

[d]Italicized values indicate statistical significance ($P \leq .05$).

[e]LPA: light physical activity.

In addition, exploratory sensitivity analyses were conducted, excluding trials classified as high RoB and those using nonobjective measurement methods. These results did not significantly differ from the main analyses but differed for sedentary time (Table 2).

Furthermore, owing to the different formats of primary data, an exploratory sensitivity analysis was performed by analyzing only primary data reported as means (SD), which did not significantly differ from the main analysis for the loss of statistical significance in the sedentary time MD (Table 2).

Moreover, studies were removed one at a time from each analysis to examine their individual impact on pooled MD estimates, which were not significantly modified in magnitude or direction when each trial was removed in any of the outcomes.

XSL•FO
RenderX

When performing an exploratory sensitivity analysis including only children and adolescents reported as healthy, only screen time was statistically significantly modified (*k*=4; –26.20 min/d, 95% CI –31.06 to –21.34).

Meta-Regression and Multiple Group Analysis

Exploratory random effects meta-regression models showed that neither the age of the participants nor the percentage of female participants included in the samples could have influenced the pooled standardized MD estimates for any of the outcomes studied (Table S5 in Multimedia Appendix 1).

Table S6 in Multimedia Appendix 1 shows the results of exploratory multigroup analyses, which, although there were not enough studies to have a meaningful impact on our results, were included along with tests for subgroup differences to show how evidence was distributed between the different types of eHealth interventions. Furthermore, it is important to note that these analyses are derived from observations and lack sufficient statistical power.

Publication Bias

After visually examining the funnel plots (Figures S3-S8 in Multimedia Appendix 1) and performing the Egger tests for each parameter (Table S7 in Multimedia Appendix 1), no evidence of significant publication bias was found.

Discussion

Principal Findings

This systematic review and meta-analysis provides a quantitative analysis of the effects of eHealth interventions on the different parameters of PA and sedentary behavior in children and adolescents. Overall, our results showed that eHealth interventions aimed at increasing PA and decreasing time spent in sedentary behaviors in children did not have a significant effect on increasing PA or decreasing screen time. However, they showed a statistically significant reduction in time spent on sedentary behavior that was not robust to any sensitivity analysis, highlighting the need for further research to confirm these findings.

According to our results, eHealth interventions aimed at promoting PA showed a greater effect on reducing some sedentary behavior parameters, such as time spent in sedentary behaviors or screen time, than on increasing PA, especially MVPA. These findings are in line with the results of previous systematic reviews and meta-analyses [21-23], and considering the importance of PA, especially MVPA for a healthy development [10], the lack of effect of increasing MVPA highlights one of the main limitations of this type of intervention. However, further research is required to elucidate whether more specific interventions or modified interventions could improve MVPA.

The statistically significant reduction in sedentary time, which has not been shown previously when analyzing only mHealth interventions [26], is particularly important because sedentary behaviors, especially screen-associated sedentary behaviors such as screen time, have been recognized as significant contributors to adverse health [69]. Moreover, eHealth interventions have been previously questioned owing to their potentially adverse consequences of increased sedentary screen time and decreased focused attention owing to technology use [70]; hence, our results suggest the possibility of benefits outweighing potential harm. Although it is true that our estimates are based on data from only a few studies and are not robust to sensitivity analyses performed, they need to be confirmed by subsequent studies.

Although the results of previous meta-analyses in older people raise expectations about the usefulness of eHealth interventions for increasing PA [71-73], the results in children have not shown strong beneficial effects on daily waking activities [19-30]. Our findings are in line with previous systematic reviews and meta-analyses [21-23,25-27], as the heterogeneity of the designs, small sample sizes, and variety of outcomes and devices analyzed suggest that further research is needed. This is particularly important given the rapid growth of eHealth technologies, which, considering the high replacement rate, makes studies obsolete in a short period and consequently challenges research to keep pace [74].

Several reasons could be hypothesized to explain these contradictory differences in the effect of eHealth technologies on children and older people, including the following: (1) most interventions in children lack a behavioral change model, or the model is inappropriate; (2) because children are "born connected," they are assumed to have a literacy in the use of these technologies that may be lacking or may not be sufficient for appropriate use; (3) the use of these technologies may require a change in children's daily routines, thus requiring parental involvement; and (4) owing to intervention strategies being unable to meet children's needs, it may be difficult to achieve and maintain motivation and benefit perception to preserve adherence.

In view of these results, should we discard any hope that such interventions are effective in promoting PA? It seems logical that to answer this question, studies must first be conducted to overcome the aforementioned shortcomings (small number of studies, small sample sizes, high heterogeneity, and poor quality) of the studies conducted to date.

eHealth interventions promoting PA, despite controversies about their effectiveness, are described as acceptable, usable, and feasible, with benefits such as cost-effectiveness, potential for real-time data collection, feedback capability, minimized participant burden, and increased dissemination capability [70]. Thus, they are postulated as promising alternatives to truly scalable interventions; however, such benefits must be confirmed through studies involving larger sample sizes. Hence, future research should focus on developing eHealth interventions adapted to the pediatric population, which could be tested in large RCTs with a clearer and comparable methodology.

Limitations

Our results should be interpreted cautiously, as our study has some limitations that should be acknowledged. First, the paucity of comparable RCTs, RCTs registered but not yet published, or gray literature that may have been overlooked in our systematic search may have affected the magnitude of our

XSL•FO
RenderX

estimates. Second, the heterogeneity of some of the included interventions owing to differences in their components, length, main outcomes, or measurement instruments may also have affected our findings. Third, the small sample size of some of the included studies decreased the precision of our estimates. Fourth, challenges arise in estimating the independent effect of each component when analyzing multicomponent interventions or interventions whose primary objective is not promoting PA. Fifth, approximately 50% (6/14) of the included trials showed a high RoB, which could compromise the consistency of our findings. Sixth, several exploratory multigroup analyses were performed, even knowing that the Cochrane Handbook for Systematic Reviews of Interventions [32] recommends that these analyses should include at least 10 studies for the same outcome, recognizing the threat it poses to the inferences based on our estimates. Finally, none of the included studies used the mHealth evidence reporting and assessment checklist [75].

Conclusions

Our results provide a comprehensive quantitative analysis of the effects of eHealth interventions on PA and sedentary behavior parameters in children and adolescents according to PRISMA standards (Multimedia Appendix 2). We did not find evidence of an effect of PA-targeted eHealth interventions on PA parameters, but the very low certainty of evidence suggests that eHealth interventions may reduce sedentary time in children and adolescents. Our findings may have an important scientific impact as they highlight that fast-paced advancements in eHealth interventions aimed at promoting PA lack robust supporting evidence.

Acknowledgments

This study was funded by the European Regional Development Fund. CPM was supported by a grant from the University of Castilla-La Mancha (2018-CPUCLM-7939).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material including in-depth step-by-step analyses.
[DOCX File , 230 KB - jmir_v26i1e41649_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 50 KB - jmir_v26i1e41649_app2.pdf]

References

1. GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016 Oct 08;388(10053):1659-1724 [FREE Full text] [doi: 10.1016/S0140-6736(16)31679-8] [Medline: 27733284]
2. Patterson R, McNamara E, Tainio M, de Sá TH, Smith AD, Sharp SJ, et al. Sedentary behaviour and risk of all-cause, cardiovascular and cancer mortality, and incident type 2 diabetes: a systematic review and dose response meta-analysis. Eur J Epidemiol 2018 Sep;33(9):811-829 [FREE Full text] [doi: 10.1007/s10654-018-0380-1] [Medline: 29589226]
3. Katzmarzyk PT, Friedenreich C, Shiroma EJ, Lee IM. Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries. Br J Sports Med 2022 Jan;56(2):101-106 [FREE Full text] [doi: 10.1136/bjsports-2020-103640] [Medline: 33782046]
4. Yavuz HM, van Ijzendoorn MH, Mesman J, van der Veek S. Interventions aimed at reducing obesity in early childhood: a meta-analysis of programs that involve parents. J Child Psychol Psychiatry 2015 Jun;56(6):677-692 [FREE Full text] [doi: 10.1111/jcpp.12330] [Medline: 25292319]
5. Garrido-Miguel M, Cavero-Redondo I, Álvarez-Bueno C, Rodríguez-Artalejo F, Moreno LA, Ruiz JR, et al. Prevalence and trends of overweight and obesity in European children from 1999 to 2016: a systematic review and meta-analysis. JAMA Pediatr 2019 Oct 01;173(10):e192430 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.2430] [Medline: 31381031]
6. Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. JAMA 2013 Jan 02;309(1):71-82 [FREE Full text] [doi: 10.1001/jama.2012.113905] [Medline: 23280227]
7. Juhola J, Magnussen CG, Viikari JS, Kähönen M, Hutri-Kähönen N, Jula A, et al. Tracking of serum lipid levels, blood pressure, and body mass index from childhood to adulthood: the cardiovascular risk in young Finns study. J Pediatr 2011 Oct;159(4):584-590 [FREE Full text] [doi: 10.1016/j.jpeds.2011.03.021] [Medline: 21514597]
8. Juonala M, Lewis S, McLachlan R, Hammarberg K, Kennedy J, Saffery R, et al. American Heart Association ideal cardiovascular health score and subclinical atherosclerosis in 22-35-year-old adults conceived with and without assisted

XSL•FO
RenderX

reproductive technologies. Hum Reprod 2020 Jan 01;35(1):232-239 [FREE Full text] [doi: 10.1093/humrep/dez240] [Medline: 31834929]
9. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec 25;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
10. Chaput JP, Willumsen J, Bull F, Chou R, Ekelund U, Firth J, et al. 2020 WHO guidelines on physical activity and sedentary behaviour for children and adolescents aged 5-17 years: summary of the evidence. Int J Behav Nutr Phys Act 2020 Nov 26;17(1):141 [FREE Full text] [doi: 10.1186/s12966-020-01037-z] [Medline: 33239009]
11. Tremblay MS, Gray C, Akinroye K, Harrington D, Katzmarzyk P, Lambert EV, et al. Physical activity of children: a global matrix of grades comparing 15 countries. J Phys Act Health 2014 May;11 Suppl 1:S113-S125 [FREE Full text] [doi: 10.1123/jpah.2014-0177] [Medline: 25426906]
12. Tremblay MS, Aubert S, Barnes J, Saunders T, Carson V, Latimer-Cheung A, et al. Sedentary behavior research network (SBRN) - terminology consensus project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75 [FREE Full text] [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
13. Janssen I, Leblanc AG. Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. Int J Behav Nutr Phys Act 2010 May 11;7:40 [FREE Full text] [doi: 10.1186/1479-5868-7-40] [Medline: 20459784]
14. Wu XY, Han LH, Zhang JH, Luo S, Hu JW, Sun K. The influence of physical activity, sedentary behavior on health-related quality of life among the general population of children and adolescents: a systematic review. PLoS One 2017 Nov 09;12(11):e0187668 [FREE Full text] [doi: 10.1371/journal.pone.0187668] [Medline: 29121640]
15. Tremblay MS, LeBlanc AG, Kho ME, Saunders TJ, Larouche R, Colley RC, et al. Systematic review of sedentary behaviour and health indicators in school-aged children and youth. Int J Behav Nutr Phys Act 2011 Sep 21;8:98 [FREE Full text] [doi: 10.1186/1479-5868-8-98] [Medline: 21936895]
16. Nooijen CF, Galanti MR, Engström K, Möller J, Forsell Y. Effectiveness of interventions on physical activity in overweight or obese children: a systematic review and meta-analysis including studies with objectively measured outcomes. Obes Rev 2017 Feb;18(2):195-213 [FREE Full text] [doi: 10.1111/obr.12487] [Medline: 28067022]
17. Lau PW, Lau EY, Wong DP, Ransdell L. A systematic review of information and communication technology-based interventions for promoting physical activity behavior change in children and adolescents. J Med Internet Res 2011 Jul 13;13(3):e48 [FREE Full text] [doi: 10.2196/jmir.1533] [Medline: 21749967]
18. Monitoring and evaluating digital health interventions. World Health Organization. 2016 Dec 12. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-02-06]
19. McIntosh JR, Jay S, Hadden N, Whittaker PJ. Do e-health interventions improve physical activity in young people: a systematic review. Public Health 2017 Jul;148:140-148 [FREE Full text] [doi: 10.1016/j.puhe.2017.04.001] [Medline: 28482322]
20. Norman GJ, Zabinski MF, Adams MA, Rosenberg DE, Yaroch AL, Atienza AA. A review of eHealth interventions for physical activity and dietary behavior change. Am J Prev Med 2007 Oct;33(4):336-345 [FREE Full text] [doi: 10.1016/j.amepre.2007.05.007] [Medline: 17888860]
21. Böhm B, Karwiese SD, Böhm H, Oberhoffer R. Effects of mobile health including wearable activity trackers to increase physical activity outcomes among healthy children and adolescents: systematic review. JMIR Mhealth Uhealth 2019 Apr 30;7(4):e8298 [FREE Full text] [doi: 10.2196/mhealth.8298] [Medline: 31038460]
22. He Z, Wu H, Yu F, Fu J, Sun S, Huang T, et al. Effects of smartphone-based interventions on physical activity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Feb 01;9(2):e22601 [FREE Full text] [doi: 10.2196/22601] [Medline: 33522980]
23. Darling KE, Sato AF. Systematic review and meta-analysis examining the effectiveness of mobile health technologies in using self-monitoring for pediatric weight management. Child Obes 2017 Oct;13(5):347-355 [FREE Full text] [doi: 10.1089/chi.2017.0038] [Medline: 28471699]
24. Shin Y, Kim SK, Lee M. Mobile phone interventions to improve adolescents' physical health: a systematic review and meta-analysis. Public Health Nurs 2019 Nov;36(6):787-799 [FREE Full text] [doi: 10.1111/phn.12655] [Medline: 31456259]
25. Turner T, Spruijt-Metz D, Wen CK, Hingle MD. Prevention and treatment of pediatric obesity using mobile and wireless technologies: a systematic review. Pediatr Obes 2015 Dec;10(6):403-409 [FREE Full text] [doi: 10.1111/ijpo.12002] [Medline: 25641770]
26. Baumann H, Fiedler J, Wunsch K, Woll A, Wollesen B. mHealth interventions to reduce physical inactivity and sedentary behavior in children and adolescents: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 May 11;10(5):e35920 [FREE Full text] [doi: 10.2196/35920] [Medline: 35544294]
27. Langarizadeh M, Sadeghi M, As'habi A, Rahmati P, Sheikhtaheri A. Mobile apps for weight management in children and adolescents; an updated systematic review. Patient Educ Couns 2021 Sep;104(9):2181-2188 [FREE Full text] [doi: 10.1016/j.pec.2021.01.035] [Medline: 33573915]

XSL•FO
RenderX

28. Yau KW, Tang TS, Görges M, Pinkney S, Kim AD, Kalia A, et al. Effectiveness of mobile apps in promoting healthy behavior changes and preventing obesity in children: systematic review. JMIR Pediatr Parent 2022 Mar 28;5(1):e34967 [FREE Full text] [doi: 10.2196/34967] [Medline: 35343908]
29. Ludwig K, Arthur R, Sculthorpe N, Fountain H, Buchan DS. Text messaging interventions for improvement in physical activity and sedentary behavior in youth: systematic review. JMIR Mhealth Uhealth 2018 Sep 17;6(9):e10799 [FREE Full text] [doi: 10.2196/10799] [Medline: 30224335]
30. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017 Dec;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
31. Page MJ, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
32. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: Wiley; Sep 20, 2019.
33. Ruotsalainen H, Kyngäs H, Tammelin T, Heikkinen H, Kääriäinen M. Effectiveness of Facebook-delivered lifestyle counselling and physical activity self-monitoring on physical activity and body mass index in overweight and obese adolescents: a randomized controlled trial. Nurs Res Pract 2015;2015:159205 [FREE Full text] [doi: 10.1155/2015/159205] [Medline: 26697218]
34. Direito A, Jiang Y, Whittaker R, Maddison R. Apps for improving fitness and increasing physical activity among young people: the AIMFIT pragmatic randomized controlled trial. J Med Internet Res 2015 Aug 27;17(8):e210 [FREE Full text] [doi: 10.2196/jmir.4568] [Medline: 26316499]
35. Staiano AE, Marker AM, Beyl RA, Hsia DS, Katzmarzyk PT, Newton RL. A randomized controlled trial of dance exergaming for exercise training in overweight and obese adolescent girls. Pediatr Obes 2017 Apr;12(2):120-128 [FREE Full text] [doi: 10.1111/ijpo.12117] [Medline: 26918815]
36. Babic MJ, Smith JJ, Morgan PJ, Lonsdale C, Plotnikoff RC, Eather N, et al. Intervention to reduce recreational screen-time in adolescents: outcomes and mediators from the 'Switch-Off 4 Healthy Minds' (S4HM) cluster randomized controlled trial. Prev Med 2016 Oct;91:50-57 [FREE Full text] [doi: 10.1016/j.ypmed.2016.07.014] [Medline: 27471018]
37. Dewar DL, Morgan PJ, Plotnikoff RC, Okely AD, Collins CE, Batterham M, et al. The nutrition and enjoyable activity for teen girls study: a cluster randomized controlled trial. Am J Prev Med 2013 Sep;45(3):313-317 [FREE Full text] [doi: 10.1016/j.amepre.2013.04.014] [Medline: 23953358]
38. Dewar DL, Morgan PJ, Plotnikoff RC, Okely AD, Batterham M, Lubans DR. Exploring changes in physical activity, sedentary behaviors and hypothesized mediators in the NEAT girls group randomized controlled trial. J Sci Med Sport 2014 Jan;17(1):39-46 [FREE Full text] [doi: 10.1016/j.jsams.2013.02.003] [Medline: 23506657]
39. Lubans DR, Morgan PJ, Okely AD, Dewar D, Collins CE, Batterham M, et al. Preventing obesity among adolescent girls: one-year outcomes of the nutrition and enjoyable activity for teen girls (NEAT Girls) cluster randomized controlled trial. Arch Pediatr Adolesc Med 2012 Sep 01;166(9):821-827 [FREE Full text] [doi: 10.1001/archpediatrics.2012.41] [Medline: 22566517]
40. Huang JS, Dillon L, Terrones L, Schubert L, Roberts W, Finklestein J, et al. Fit4Life: a weight loss intervention for children who have survived childhood leukemia. Pediatr Blood Cancer 2014 May;61(5):894-900 [FREE Full text] [doi: 10.1002/pbc.24937] [Medline: 24436138]
41. Lubans DR, Smith JJ, Plotnikoff RC, Dally KA, Okely AD, Salmon J, et al. Assessing the sustained impact of a school-based obesity prevention program for adolescent boys: the ATLAS cluster randomized controlled trial. Int J Behav Nutr Phys Act 2016 Aug 20;13:92 [FREE Full text] [doi: 10.1186/s12966-016-0420-8] [Medline: 27542825]
42. Smith JJ, Morgan PJ, Plotnikoff RC, Dally KA, Salmon J, Okely AD, et al. Smart-phone obesity prevention trial for adolescent boys in low-income communities: the ATLAS RCT. Pediatrics 2014 Sep;134(3):e723-e731 [FREE Full text] [doi: 10.1542/peds.2014-1012] [Medline: 25157000]
43. Mendoza JA, Baker KS, Moreno MA, Whitlock K, Abbey-Lambertz M, Waite A, et al. A Fitbit and Facebook mHealth intervention for promoting physical activity among adolescent and young adult childhood cancer survivors: a pilot study. Pediatr Blood Cancer 2017 Dec;64(12) [FREE Full text] [doi: 10.1002/pbc.26660] [Medline: 28618158]
44. Patrick K, Norman GJ, Davila EP, Calfas KJ, Raab F, Gottschalk M, et al. Outcomes of a 12-month technology-based intervention to promote weight loss in adolescents at risk for type 2 diabetes. J Diabetes Sci Technol 2013 May 01;7(3):759-770 [FREE Full text] [doi: 10.1177/193229681300700322] [Medline: 23759410]
45. Thompson D, Cantu D, Ramirez B, Cullen KW, Baranowski T, Mendoza J, et al. Texting to increase adolescent physical activity: feasibility assessment. Am J Health Behav 2016 Jul;40(4):472-483 [FREE Full text] [doi: 10.5993/AJHB.40.4.9] [Medline: 27338994]
46. Wright JA, Phillips BD, Watson BL, Newby PK, Norman GJ, Adams WG. Randomized trial of a family-based, automated, conversational obesity treatment program for underserved populations. Obesity (Silver Spring) 2013 Sep;21(9):E369-E378 [FREE Full text] [doi: 10.1002/oby.20388] [Medline: 23512915]

XSL•FO
RenderX

47. Newton KH, Wiltshire EJ, Elley CR. Pedometers and text messaging to increase physical activity: randomized controlled trial of adolescents with type 1 diabetes. Diabetes Care 2009 May;32(5):813-815 [FREE Full text] [doi: 10.2337/dc08-1974] [Medline: 19228863]
48. Shapiro JR, Bauer S, Hamer RM, Kordy H, Ward D, Bulik CM. Use of text messaging for monitoring sugar-sweetened beverages, physical activity, and screen time in children: a pilot study. J Nutr Educ Behav 2008;40(6):385-391 [FREE Full text] [doi: 10.1016/j.jneb.2007.09.014] [Medline: 18984496]
49. Armbrust W, Bos GJ, Wulffraat NM, van Brussel M, Cappon J, Dijkstra PU, et al. Internet program for physical activity and exercise capacity in children with juvenile idiopathic arthritis: a multicenter randomized controlled trial. Arthritis Care Res (Hoboken) 2017 Jul;69(7):1040-1049 [FREE Full text] [doi: 10.1002/acr.23100] [Medline: 27696793]
50. de Bourdeaudhuij I, Maes L, de Henauw S, de Vriendt T, Moreno LA, Kersting M, et al. Evaluation of a computer-tailored physical activity intervention in adolescents in six European countries: the Activ-O-Meter in the HELENA intervention study. J Adolesc Health 2010 May;46(5):458-466 [FREE Full text] [doi: 10.1016/j.jadohealth.2009.10.006] [Medline: 20413082]
51. Cook TL, de Bourdeaudhuij I, Maes L, Haerens L, Grammatikaki E, Widhalm K, et al. Moderators of the effectiveness of a web-based tailored intervention promoting physical activity in adolescents: the HELENA Activ-O-Meter. J Sch Health 2014 Apr;84(4):256-266 [FREE Full text] [doi: 10.1111/josh.12140] [Medline: 24617909]
52. Ezendam NP, Brug J, Oenema A. Evaluation of the web-based computer-tailored FATaintPHAT intervention to promote energy balance among adolescents: results from a school cluster randomized trial. Arch Pediatr Adolesc Med 2012 Mar;166(3):248-255 [FREE Full text] [doi: 10.1001/archpediatrics.2011.204] [Medline: 22064878]
53. Guagliano JM, Armitage SM, Brown HE, Coombes E, Fusco F, Hughes C, et al. A whole family-based physical activity promotion intervention: findings from the families reporting every step to health (FRESH) pilot randomised controlled trial. Int J Behav Nutr Phys Act 2020 Sep 22;17(1):120 [FREE Full text] [doi: 10.1186/s12966-020-01025-3] [Medline: 32962724]
54. Haerens L, Deforche B, Maes L, Cardon G, Stevens V, de Bourdeaudhuij I. Evaluation of a 2-year physical activity and healthy eating intervention in middle school children. Health Educ Res 2006 Dec;21(6):911-921 [FREE Full text] [doi: 10.1093/her/cyl115] [Medline: 17032704]
55. Haerens L, de Bourdeaudhuij I, Maes L, Cardon G, Deforche B. School-based randomized controlled trial of a physical activity intervention among adolescents. J Adolesc Health 2007 Mar;40(3):258-265 [FREE Full text] [doi: 10.1016/j.jadohealth.2006.09.028] [Medline: 17321427]
56. Slootmaker SM, Chinapaw MJ, Seidell JC, van Mechelen W, Schuit AJ. Accelerometers and internet for physical activity promotion in youth? feasibility and effectiveness of a minimal intervention [ISRCTN93896459]. Prev Med 2010 Jul;51(1):31-36 [FREE Full text] [doi: 10.1016/j.ypmed.2010.03.015] [Medline: 20380847]
57. Staiano AE, Beyl RA, Hsia DS, Katzmarzyk PT, Newton RL. Twelve weeks of dance exergaming in overweight and obese adolescent girls: transfer effects on physical activity, screen time, and self-efficacy. J Sport Health Sci 2017 Mar;6(1):4-10 [FREE Full text] [doi: 10.1016/j.jshs.2016.11.005] [Medline: 28491483]
58. Tremblay MS, Carson V, Chaput JP. Introduction to the Canadian 24-hour movement guidelines for children and youth: an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2016 Jun;41(6 Suppl 3):iii-iiv. [doi: 10.1139/apnm-2016-0203] [Medline: 27306430]
59. Sterne JA, Savović J, Page M, Elbers R, Blencowe N, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
60. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol 2020 Oct;126:37-44 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.06.015] [Medline: 32562833]
61. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the journal of clinical epidemiology. J Clin Epidemiol 2011 Apr;64(4):380-382 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.09.011] [Medline: 21185693]
62. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005 Apr 20;5:13 [FREE Full text] [doi: 10.1186/1471-2288-5-13] [Medline: 15840177]
63. Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. Med Care 1984 Dec;22(12):1101-1114. [doi: 10.1097/00005650-198412000-00005] [Medline: 6513619]
64. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MW, et al. Including variants on randomized trials. In: Cochrane Handbook for Systematic Reviews of Interventions, Second Edition. Hoboken, NJ: John Wiley & Sons; 2019.
65. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
66. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558 [FREE Full text] [doi: 10.1002/sim.1186] [Medline: 12111919]

XSL•FO
RenderX

67. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ 2001 Jul 14;323(7304):101-105 [FREE Full text] [doi: 10.1136/bmj.323.7304.101] [Medline: 11451790]
68. Lelieveld OT, Armbrust W, Geertzen JH, de Graaf I, van Leeuwen MA, Sauer PJ, et al. Promoting physical activity in children with juvenile idiopathic arthritis through an internet-based program: results of a pilot randomized controlled trial. Arthritis Care Res (Hoboken) 2010 May;62(5):697-703 [FREE Full text] [doi: 10.1002/acr.20085] [Medline: 20191468]
69. Saunders TJ, Chaput J, Tremblay M. Sedentary behaviour as an emerging risk factor for cardiometabolic diseases in children and youth. Can J Diabetes 2014 Mar;38(1):53-61 [FREE Full text] [doi: 10.1016/j.jcjd.2013.08.266] [Medline: 24485214]
70. Tate EB, Spruijt-Metz D, O'Reilly G, Jordan-Marsh M, Gotsis M, Pentz MA, et al. mHealth approaches to child obesity prevention: successes, unique challenges, and next directions. Transl Behav Med 2013 Dec;3(4):406-415 [FREE Full text] [doi: 10.1007/s13142-013-0222-3] [Medline: 24294329]
71. Kwan RY, Salihu D, Lee PH, Tse M, Cheung DS, Roopsawang I, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020;17:7 [FREE Full text] [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
72. Núñez de Arenas-Arroyo S, Cavero-Redondo I, Alvarez-Bueno C, Sequí-Domínguez I, Reina-Gutiérrez S, Martínez-Vizcaíno V. Effect of eHealth to increase physical activity in healthy adults over 55 years: a systematic review and meta-analysis. Scand J Med Sci Sports 2021 Apr;31(4):776-789 [FREE Full text] [doi: 10.1111/sms.13903] [Medline: 33280182]
73. Martin A, Fitzsimons C, Jepson R, Saunders DH, van der Ploeg HP, Teixeira PJ, et al. Interventions with potential to reduce sedentary time in adults: systematic review and meta-analysis. Br J Sports Med 2015 Aug;49(16):1056-1063 [FREE Full text] [doi: 10.1136/bjsports-2014-094524] [Medline: 25907181]
74. Lee AM, Chavez S, Bian J, Thompson LA, Gurka MJ, Williamson VG, et al. Efficacy and effectiveness of mobile health technologies for facilitating physical activity in adolescents: scoping review. JMIR Mhealth Uhealth 2019 Mar 12;7(2):e11847 [FREE Full text] [doi: 10.2196/11847] [Medline: 30747716]
75. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, et al. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174 [FREE Full text] [doi: 10.1136/bmj.i1174] [Medline: 26988021]

Abbreviations

FI: favoring intervention
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
IG: intervention group
LPA: light physical activity
MD: mean difference
mHealth: mobile health
MVPA: moderate to vigorous physical activity
PA: physical activity
PICO: patient, intervention, comparison, and outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB: risk of bias
RoB2: Risk of Bias 2

Edited by A Mavragani; submitted 03.08.22; peer-reviewed by A Bailey, W Wang; comments to author 07.10.22; revised version received 30.05.23; accepted 19.12.23; published 21.02.24.

Please cite as:
Sequí-Domínguez I, Cavero-Redondo I, Álvarez-Bueno C, López-Gil JF, Martínez-Vizcaíno V, Pascual-Morena C
Effectiveness of eHealth Interventions Promoting Physical Activity in Children and Adolescents: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e41649
URL: https://www.jmir.org/2024/1/e41649
doi:10.2196/41649
PMID:38381490

©Irene Sequí-Domínguez, Iván Cavero-Redondo, Celia Álvarez-Bueno, Jose Francisco López-Gil, Vicente Martínez-Vizcaíno, Carlos Pascual-Morena. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of eHealth Interventions on 24-Hour Movement Behaviors Among Preschoolers: Systematic Review and Meta-Analysis

Shan Jiang[1], MSc; Johan Y Y Ng[1], PhD; Kar Hau Chong[2], PhD; Bo Peng[1], MSc; Amy S Ha[1], PhD

[1]Department of Sports Science and Physical Education, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]School of Health and Society and Early Start, Faculty of the Arts, Social Sciences and Humanities, University of Wollongong, Wollongong, Australia

Corresponding Author:
Amy S Ha, PhD
Department of Sports Science and Physical Education
The Chinese University of Hong Kong
G05 Kwok Sports Building, Shatin, N.T.
Hong Kong
China (Hong Kong)
Phone: 852 3943 6083
Email: sauchingha@cuhk.edu.hk

Abstract

Background: The high prevalence of unhealthy movement behaviors among young children remains a global public health issue. eHealth is considered a cost-effective approach that holds great promise for enhancing health and related behaviors. However, previous research on eHealth interventions aimed at promoting behavior change has primarily focused on adolescents and adults, leaving a limited body of evidence specifically pertaining to preschoolers.

Objective: This review aims to examine the effectiveness of eHealth interventions in promoting 24-hour movement behaviors, specifically focusing on improving physical activity (PA) and sleep duration and reducing sedentary behavior among preschoolers. In addition, we assessed the moderating effects of various study characteristics on intervention effectiveness.

Methods: We searched 6 electronic databases (PubMed, Ovid, SPORTDiscus, Scopus, Web of Science, and Cochrane Central Register of Controlled Trials) for experimental studies with a randomization procedure that examined the effectiveness of eHealth interventions on 24-hour movement behaviors among preschoolers aged 2 to 6 years in February 2023. The study outcomes included PA, sleep duration, and sedentary time. A meta-analysis was conducted to assess the pooled effect using a random-effects model, and subgroup analyses were conducted to explore the potential effects of moderating factors such as intervention duration, intervention type, and risk of bias (ROB). The included studies underwent a rigorous ROB assessment using the Cochrane ROB tool. Moreover, the certainty of evidence was evaluated using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) assessment.

Results: Of the 7191 identified records, 19 (0.26%) were included in the systematic review. The meta-analysis comprised a sample of 2971 preschoolers, which was derived from 13 included studies. Compared with the control group, eHealth interventions significantly increased moderate to vigorous PA (Hedges g=0.16, 95% CI 0.03-0.30; P=.02) and total PA (Hedges g=0.37, 95% CI 0.02-0.72; P=.04). In addition, eHealth interventions significantly reduced sedentary time (Hedges g=−0.15, 95% CI −0.27 to −0.02; P=.02) and increased sleep duration (Hedges g=0.47, 95% CI 0.18-0.75; P=.002) immediately after the intervention. However, no significant moderating effects were observed for any of the variables assessed (P>.05). The quality of evidence was rated as “moderate” for moderate to vigorous intensity PA and sedentary time outcomes and “low” for sleep outcomes.

Conclusions: eHealth interventions may be a promising strategy to increase PA, improve sleep, and reduce sedentary time among preschoolers. To effectively promote healthy behaviors in early childhood, it is imperative for future studies to prioritize the development of rigorous comparative trials with larger sample sizes. In addition, researchers should thoroughly examine the effects of potential moderators. There is also a pressing need to comprehensively explore the long-term effects resulting from these interventions.

Trial Registration: PROSPERO CRD42022365003; http://tinyurl.com/3nnfdwh3

(J Med Internet Res 2024;26:e52905) doi:10.2196/52905

XSL•FO
RenderX

KEYWORDS

preschooler; movement behaviors; eHealth; physical activity; sedentary behavior; sleep; mobile phone; review; systematic review

Introduction

Background

Physical activity (PA), sedentary behavior (SB), and sleep are integrated as "24-hour movement behaviors" owing to the collective effect on daily movement patterns. The 24-hour movement paradigm acknowledges the possibility of categorizing these behaviors according to their intensity levels across a full day. This encompasses a diverse range of activities, including sleep; SB (eg, screen time, reclining, or lying down); and light, moderate, or vigorous PA [1]. Globally, the "24-hour movement behaviors" paradigm has already been recognized and adopted into movement guidelines [2]. In 2020, the World Health Organization (WHO) released guidelines on PA and SB that incorporate all 3 movement behaviors [3]. The health benefits of engaging in PA, getting the recommended sleep, and reducing sedentary time are well documented. Recent reviews have shown a positive association between PA; sleep; and a wide range of child outcomes related to mental health, cognition, and cardiometabolism [4-6]. In addition, it is worth mentioning that different domains of SB can have varying health effects. For instance, non–screen-based sedentary activities such as reading or studying have been associated with favorable cognitive development in children [7]. Conversely, screen-based sedentary time, also referred to as "screen time," has been found to have adverse effects on health-related outcomes [8]. Moreover, prior research has indicated that imbalances in 24-hour movement behaviors—specifically, elevated sedentary screen time coupled with diminished levels of PA and sleep—could potentially increase the risk of depression [9] and result in poor health-related quality of life [10]. Therefore, any change in one of the movement behaviors may lead to a compensatory increase or decrease in one or both behaviors.

However, insufficient healthy levels of 24-hour movement behaviors in early childhood have remained one of the most critical global public health challenges [11,12]. According to the WHO guidelines [3], preschool children are recommended to engage in adequate daily PA, consisting of 180 minutes, with 60 minutes dedicated to moderate to vigorous PA (MVPA). In addition, they should ensure sufficient sleep, ranging from 10 to 13 hours, while limiting sedentary recreational screen time to no more than 60 minutes per day. Unfortunately, a significant proportion of preschoolers do not meet the PA guidelines (<50% across studies) [13]. Furthermore, previous studies have consistently demonstrated that preschoolers exceed the screen time recommendations set by the WHO. A comprehensive meta-analysis of 44 studies revealed that only 35.6% of children aged between 2 and 5 years met the guideline of limiting daily screen time to 1 hour. Moreover, when examining the integration of 24-hour movement behaviors [8], another meta-analysis discovered that only 13% of children worldwide adhere to all 3 behavior guidelines [14].

Preschoolers play a crucial role in laying the foundation for long-term physical health and overall well-being [15,16]. Improving PA levels, minimizing SB, and prioritizing quality sleep in young children have multiple benefits, including positively influencing their physical fitness [17,18], promoting the development of motor and cognitive skills [19,20], and preventing childhood obesity [21] and associated health issues [14,22,23]. Several studies have shown that these healthy behavior patterns can shape lifelong habits that extend from childhood through adolescence and into adulthood [5,24].

Although these statistics are concerning, attempts to address the issue through various interventions have yielded inconsistent findings [25-28]. For instance, a meta-analysis of PA intervention studies involving preschoolers revealed only small to moderate effects in enhancing PA, suggesting room for improvement in achieving the desired outcomes [29]. In a meta-analysis conducted by Fangupo et al [30], no intervention effect was observed on daytime sleep duration for young children. Interestingly, earlier research has also elucidated overflow effects stemming from interventions focusing on a specific behavior, impacting other behaviors that were not the primary target. A systematic review highlighted that interventions aimed at enhancing PA in children aged <5 years led to a reduction in screen time by approximately 32 minutes [31]. It is crucial to understand that as time is finite, the durations dedicated to PA, sedentary time, and sleep are interconnected within 24 hours. Thus, we need effective interventions for preschool children that holistically address all components of 24-hour movement behaviors.

eHealth broadly refers to a diverse array of information and communication technologies used to facilitate the delivery of health care [32,33]. The rapid evolution of digitalization in recent decades has led to the widespread adoption of eHealth in interventions [28,34]. Recent reviews [35-38] suggest that with the global proliferation of eHealth interventions, health promotion via these platforms is evolving to become more accessible and user-friendly, garnering acceptance among adolescents and adults. Previous reviews have underscored the effectiveness of these digital platforms in enhancing various movement behavior outcomes across diverse age groups, including children aged 6 to 12 years [39], adolescents [40], adults [41], and older adults [42]. Specifically, a meta-analysis indicated that eHealth interventions have successfully promoted PA among individuals with noncommunicable diseases [43]. Another review showed that computer, mobile, and wearable technologies have the potential to mitigate sedentary time effectively [41]. Previous studies have targeted different participant groups to investigate the impact of eHealth on sleep outcomes. Deng et al [44] conducted a meta-analysis demonstrating that eHealth interventions for adults with insomnia are effective in improving sleep and can be considered a promising treatment. Nevertheless, a review focusing on healthy adolescents found that there has not been any school-based eHealth interventions focusing on sleep outcomes [45].

XSL•FO
RenderX

Indeed, child-centered strategies such as gamification are used in some digital apps and have been shown to encourage children's PA [46-48]. A considerable body of work has addressed the pivotal role of parental influence and role modeling in cultivating healthy lifestyle habits in children [49,50]. Physical literacy, a multidimensional concept encompassing various aspects of PA such as the affective, physical, cognitive, and behavioral dimensions, plays a vital role in enhancing PA engagement [51]. Ha et al [52] conducted a web-based parent-focused intervention, revealing that enhancing parents' physical literacy can effectively support children's participation in PA. By understanding and promoting physical literacy, parents can provide valuable support to their children, fostering a lifelong commitment to healthy and active lifestyles. Although eHealth interventions offer promise, there are conflicting findings regarding their impact, especially when they are parent supported and targeted at young children. A previous meta-analysis examining eHealth interventions targeted at parents found no significant impact on children's BMI. In addition, no studies have included children aged <5 years [50]. Similarly, a recent systematic review observed that eHealth interventions aimed at parents showed no significant effectiveness in enhancing PA levels in young children [53]. However, the prevalence of digital device use in young children has become widespread. For instance, studies conducted in England (the United Kingdom), Estonia, and the United States have reported that, on average, 83% of children aged 5 years use a digital device at least once a week [54]. Research also revealed that in the United States, approximately three-fourths of children had their own mobile device by the age of 4 years, and nearly all children (96.6%) used mobile devices [55]. Consequently, there is an urgent need to harness the potential of digital platforms and explore whether they can effectively deliver interventions to preschoolers [56].

Objectives

In previous research, there has been a lack of studies examining the effectiveness of eHealth behavior change interventions among preschoolers. Although a systematic review found a significant effect of digital health interventions on the PA of preschoolers [53], this review did not include sedentary time and sleep in its inclusion criteria, and there is a lack of conclusive statements owing to the insufficient number of studies, and no quantitative methods were available for synthesizing the evidence on the effectiveness of eHealth interventions. To our knowledge, no systematic review or meta-analysis has distinctly investigated the effects of eHealth interventions on 24-hour movement behaviors in preschoolers or the factors that may influence their implementation. Therefore, the aims of this study were (1) to assess the effectiveness of eHealth interventions on 24-hour movement behaviors (improving PA and sleep duration and decreasing sedentary time) and (2) to examine the moderating effects of study characteristics (eg, intervention duration, intervention type, and outcome measurement tools) on intervention effectiveness.

Methods

This review was registered with PROSPERO (CRD42022365003) and conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [57].

Eligibility Criteria

This review included trials with a randomization procedure that examined the outcomes of interventions using information and communication technology. These interventions targeted at least 1 movement behavior in preschool children aged 2 to 6 years. Studies were excluded if (1) the control groups received intervention using eHealth technology and (2) published in a non-English language. Full details are provided in Multimedia Appendix 1 [58].

Search and Selection

The following databases were systematically searched from inception to February 08, 2023: PubMed, Ovid, SPORTDiscus, Scopus, Web of Science, and Cochrane Central Register of Controlled Trials. We used the search terms "eHealth," "Physical activity," "Sedentary behavior," "Sleep," "preschooler," and their Medical Subject Headings terms. The complete search strategy is described in Multimedia Appendix 2 [59-61]. A manual search of the reference lists of the included publications was performed to identify additional eligible studies for potential inclusion. Two independent reviewers (SJ and BP) screened the titles and abstracts and subsequent full-text articles for eligibility. Discrepancies that emerged during the selection process were effectively resolved through a discussion involving 3 authors (SJ, BP, and JYYN).

Data Extraction

A comprehensive data extraction form was developed (SJ) and refined (SJ and BP) based on the Cochrane Handbook for Systematic Reviews of Interventions [62]. Extracted information included bibliographic details (authors, title, journal, and year); study details (country, design, retention rate); participants' characteristics (number of children and demographics); intervention type (parent supported, teacher led, or child centered), intervention's theoretical basis, duration, delivery tool, and intensity; comparison (sample size, activity type); outcomes (behavioral variables with baseline and postintervention means with SDs), and measurement tools. Regarding the categorization of intervention types, we have established a clear classification. Specifically, in child-centered interventions, children are the direct beneficiaries, participating autonomously with less guidance from guardians. This can be accomplished using an exergaming system or designed mobile health games. In parent-supported interventions, parents are involved in educational programs and instructions that improve parents' knowledge of preschoolers' healthy movement behaviors. A teacher-led intervention involves supervising preschoolers' PA during school time or participating in structured PA sessions aimed at improving healthy indicators. For data that were either incomplete or absent within the main text, we sought to reach out to the respective authors through email correspondence.

XSL•FO
RenderX

Risk of Bias

The included studies were assessed for risk of bias (ROB) using the revised Cochrane ROB2 tool [63]. The following domains of bias were assessed for each study: selection (random sequence generation and allocation concealment), performance and detection (masking of participants, personnel, and assessors), deviations from intended interventions, missing outcome data, measurement of the outcome, appropriateness of analysis (selection of the reported outcome), and bias arising from period and carryover effects (for crossover studies) [63]. The studies were ranked as low risk, some concerns, or high risk for each domain. The ROB was evaluated independently by 2 authors (SJ and BP). Any discrepancies were resolved through discussions with the author (JYYN).

Outcomes and Data Synthesis

Our outcome targeted any of the following movement behaviors: PA (MVPA and total PA), sedentary time (screen time and sitting time), or sleep duration. Meta-analysis was conducted in R (version 4.2.1; R Group for Statistical Computing) using the *meta*, *metafor*, and *metareg* packages [64]. A random-effects model (Hartung-Knapp method) was used to calculate pooled estimates (Hedges *g*, a type of standardized mean difference) to account for variations in participants and measurement methods of movement behavior outcomes [65]. Multimedia Appendix 3 [63-65] describes the processing of missing data. Hedges *g* and their corresponding variances were calculated using the pre- and postintervention mean scores and SDs. However, if some studies had changes in baseline and postintervention data or if there were significant differences in their baseline data [59-61], we used the within-group difference in means and their SDs for intervention and control groups to calculate the effect size. Values of 0.2, 0.5, and 0.8 represent small, moderate, and large effect sizes, respectively. A positive effect size indicated a beneficial effect on the intervention group compared with the control group. The between-group heterogeneity of the synthesized effect sizes was examined using the Cochran Q test and I^2 statistics. I^2 values of 25%, 50%, and 75% indicated low, moderate, and high levels of heterogeneity, respectively. Subgroup analyses were conducted based on the following factors: (1) intervention duration (0-3 months vs >3 months) and (2) type of intervention (child centered, parent focused, or teacher led). (3) Types of outcome measurement tools (objective vs self-reported) and (4) ROB (low risk, some concerns, or high risks).

Furthermore, we performed meta-regression analyses to examine the impact of potential moderators on the overall effect size. Potential moderators included 5 variables, as specified in the subgroup analyses, and 2 continuous variables (sample size and intervention length). These variables were selected based on existing evidence that highlights their significant moderating effects on eHealth interventions targeting movement behaviors [53,66,67]. Sensitivity analyses were performed using the leave-one-out method. Publication bias was visualized using funnel plot symmetry and quantified using the Eggertest score, for which $P<.05$ indicates a significant publication bias [68].

Quality Assessment of the Overall Evidence

GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) 2 criteria were used to assess the certainty of evidence for the effect of eHealth interventions on the targeted outcomes [69,70]. The GRADE assessment was completed using GRADEpro, and the quality of evidence was classified as high (≥4 points overall), moderate (3 points), low (2 points), or very low (≤1 point) [70].

Results

Study Selection

The database search yielded 7140 records, with an additional 51 records identified from the reference lists of relevant systematic reviews. There were 64 articles screened for full text, and 45 articles were excluded. The reasons for exclusion are listed in Multimedia Appendix 4. A total of 19 studies reporting the effectiveness of interventions on movement behaviors were included in the systematic review [17,59-61,71-85], and 13 studies were included in the meta-analysis [59-61,76-85]. The PRISMA flowchart of the study selection process is shown in Figure 1 and PRISMA checklists are in Multimedia Appendices 5 and 6.

XSL•FO
RenderX

Figure 1. Flowchart of study selection.

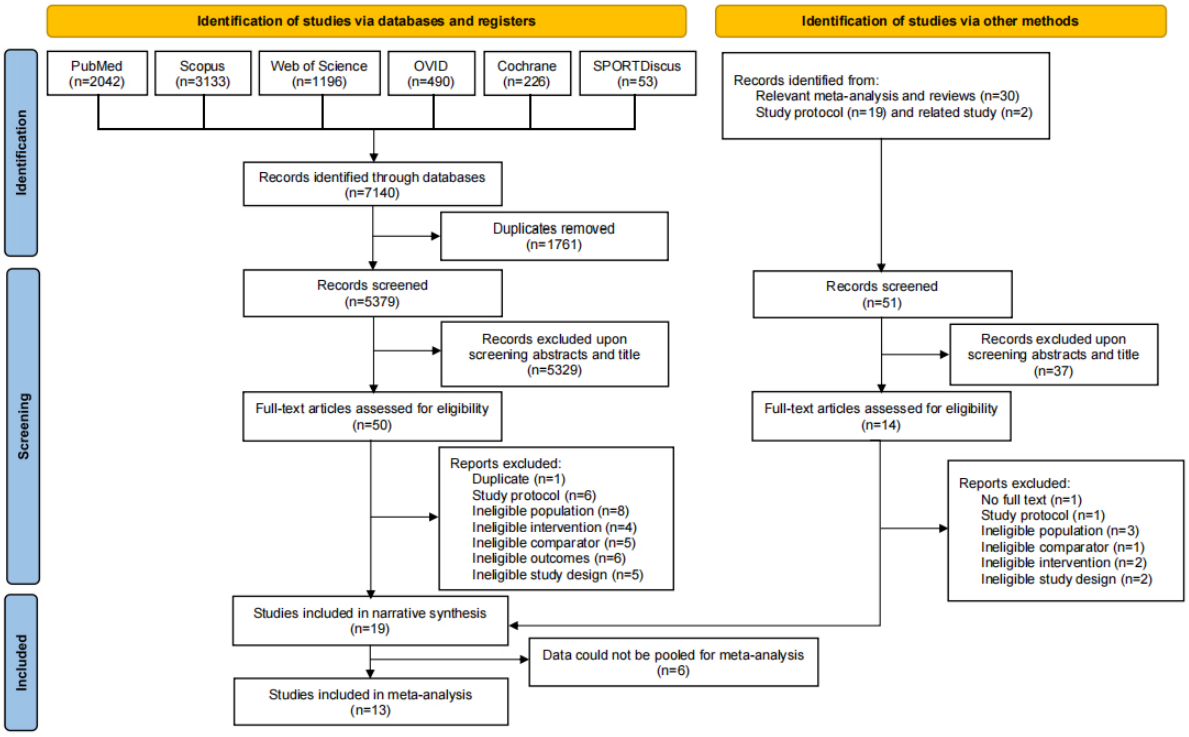

Study Characteristics

The study characteristics are described in Table 1. In the 19 studies, 2971 preschoolers from 6 regions were included. A total of 18 studies were conducted in high-income countries, and only 1 study was conducted in an upper middle–income country, according to the World Bank classification (Multimedia Appendix 7) [86]. Most included studies were conducted during and after 2017. For the study design, 16 studies were 2-arm randomized controlled trials (RCTs), with 11 using a parallel group design [17,59-61,71-74,76,77,84], 2 being cluster RCTs [82,83], 2 pilot RCTs [79,81], and 1 crossover study [85]. The remaining 3 studies consisted of 2-arm experimental studies with a randomization procedure [75,78,80]. The sample size ranged from 34 preschoolers to 617 preschoolers. The study details are presented in Multimedia Appendices 8 and 9 [59-61,76-85].

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
Andersen et al [82]; Norway	Two-arm cluster-randomized controlled trial; 3-4 y; 116 (I: 67 and C: 49); 89.66%	Teacher led; Facebook group	• Premeetings (1 h) and follow-up meetings (1 h) in each ECEC[c] institution, • Start-up seminar (6 h) and 2 follow-up courses (3+3 h) with all the staff members, • Ongoing planning and collective reflections practice in the ECEC institutions (weekly regular), • Facebook group (4 mo) • An equipment package.	4 mo	Maintain usual routines	PA[d] and SB[e]
Delisle Nyström et al [60]; Sweden	Two-arm parallel randomized controlled trial; mean 4.5 y (SD 2 mo); 315 (I: 156 and C: 159); 83.49%	Parent supported; mHealth[f] app	• A smartphone app is available to parents for 6 mo, which contains 12 healthy behavior change–related themes, with a new theme being released every 2 wk (6 mo for parents involved into mHealth intervention with researchers' support and instruction). • There is no support, and parents are free to use the apps for the next 6 mo with follow-up.	6+6-mo follow-up without support	The control group received a basic pamphlet on dietary and PA behaviors.	PA and SB
Nyström et al [61]; Sweden	Two-arm parallel randomized controlled trial; mean 4.5 y (SD 2 mo); 315 (I: 156 and C: 159); 89.2%	Parent supported; mHealth app	• A smartphone app is available to parents for 6 mo, which contains 12 healthy behavior change–related themes, with a new theme being released every 2 wk (6 mo for parents involved into mHealth intervention with researchers' support and instruction). • There is no support, and parents are free to use the apps for the next 6 mo with follow-up.	6 mo	The control group received a basic pamphlet on dietary and PA behaviors.	PA and SB
Haines et al [76]; United States	Two-arm randomized controlled trial; 2-5 y; 121 (I: 62 and C: 59); 94.21%	Parent supported; telephone call and texting	• Motivational coaching by a health educator during 4 home visits and 4 health coaching telephone calls • Mailed educational materials and incentives. • Weekly texting on adoption of household routines	6 mo (monthly coaching calls and text messages twice weekly)	The control group received 4 monthly mailed packages that included educational materials.	SB and sleep

XSL•FO
RenderX

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
Ling et al [80]; United States	Two-group experimental study with randomization procedure; 3-5 y; 73 (I: 41 and C: 32); 94.52%	Parent supported; Facebook group	• The caregiver Facebook-based program aimed to (a) offer health information, engage in family fun activities, and provide behavioral change strategies and tips. (b) it encouraged interactive positive communication among caregivers, fostering peer support. Caregivers were instructed to gradually increase their weekly step count by 500, with the goal of becoming positive role models for their preschoolers. • Three face-to-face caregiver meetings were provided. • The center-based preschooler program has been conducted. Preschoolers received 10-wk healthy eating and PAe participatory learning and fun games codelivered by teachers and interventionists 4 d per wk. (total 40 d) at their attended Head Start center. The session duration was 30 min per d. • The child's weekly letter to the caregiver was privately sent through Facebook Messenger. Caregivers were encouraged to engage in discussions with their preschooler about the letter over the weekend and share their thoughts and plans in response to the letter within the Facebook private group.	10 wk	Maintain usual routines	PA and SB
Downing et al [79]; Australia	Two-arm pilot randomized controlled trial; 2-4 y; 57 (I: 30 and C: 27); 92.98%	Parent supported; text message	• A one-on-one discussion with each participant individually, either in person or over the phone, to set their goals for the program. Two goals were set around reducing their child's SB, one is a screen time goal and another is an overall SB goal. • Three personalized, interactive text messages (ie, the main mode of intervention delivery) per week were sent to parents for 6 wk (19 texts in total). The standard text messages included 2 behavioral messages with practical ideas and suggestions for limiting and displacing their child's screen and sitting time, active play ideas, and monitoring and encouraging achievement of individual goals.	6 wk	Waitlist control	SB
				3 mo	Waitlist control	PA and sleep

XSL•FO
RenderX

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
Yoong et al [81]; Australia	Two-arm pilot randomized controlled trial; 3-6 y; 76 (I: 38 and C: 38); 48.68%	Parent supported; Web-based video, telephone call, and text messages	• Web-based video: education prerecorded videos were sent to parents. The content targeted brief information about sleep benefits and targeted behaviors to implement sleep practices. • Telephone call: approximately 2-4 wk after the video link was emailed, a trained psychologist provides parents support to implement the messages in the video. • Text messages: 2 text messages were sent approximately 1 wk after the phone call. Text messages were tailored to parent's engagement with the intervention and were aimed at encouraging parents to access the video and persist with implementing any new routines that had been put in place.			
Hoffman et al [83]; United States	Two-arm cluster-randomized controlled trial; 3-5 y; 58 (I: 27 and C: 31); 98.28%	Teacher led; study website	• Teachers were trained by receiving an email about a WE PLAY website link, and the training was self-paced over a 4-wk period. Training activities were designed to be completed in five steps: (1) complete the web-based training, (2) view the video library, (3) read the game sheets and select games, (4) lead a structured and unstructured active play session and complete the teacher self-assessments, and (5) complete the supervisor support component (ie, observations, performance feedback, and teacher recognition)	4 wk	No other instructions related to PA	PA
Marsh et al [84]; New Zealand	Two-arm randomized controlled trial; 2-4 y; 54 (I: 27 and C: 27); 98.15%	Parent supported; study website	• Parent-child dyads participated in a half-day, face-to-face workshop at the University of Auckland. The workshop is designed to introduce the program framework and provide a blueprint for parents to develop healthy lifestyle behaviors among children. • After the workshop, all participants were given access to a study website, which provided all the workshop content, for a period of 6 wk.	6 wk	Waitlist control	SB and sleep
Marsh et al [84]; New Zealand		Parent supported; study website		12 wk	Waitlist control	SB and sleep

XSL•FO
RenderX

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
	Two-arm randomized controlled trial; 2-4 y; 54 (I: 27 and C: 27); 98.15%		• Parent-child dyads participated in a half-day, face-to-face workshop at the University of Auckland. The workshop is designed to introduce the program framework and provide a blueprint for parents to develop healthy lifestyle behaviors among children. • After the workshop, all participants were given access to a study website, which provided all the workshop content, for a period of 6 wk.			
Barkin et al [77]; United States	Two-arm randomized controlled trial; 3-5 y; 610 (I: 304 and C: 306); 75.74%	Parent supported; telephone call	• A 12-wk intensive phase with weekly 90-min skills-building sessions via either in-person groups or telephone calls, • A 9-mo maintenance phase with monthly coaching telephone calls. • A 24-mo sustainability phase providing frequent cues to action (eg, texts, personalized letters, and monthly calls) to use parks and recreation programing for healthy family behaviors	12 mo (12-wk with 90 min per wk skills-building session via phone call and 9-mo weekly telephone call)	School-readiness program	PA and SB
Barkin et al [77]; United States	Two-arm randomized controlled trial; 3-5 y; 610 (I: 304 and C: 306); 75.74%	Parent supported; telephone call	• A 12-wk intensive phase with weekly 90-min skills-building sessions via either in-person groups or telephone calls, • A 9-mo maintenance phase with monthly coaching telephone calls. • A 24-mo sustainability phase providing frequent cues to action (eg, texts, personalized letters, and monthly calls) to use parks and recreation programing for healthy family behaviors	24 mo	School-readiness program	PA and SB
Barkin et al [77]; United States	Two-arm randomized controlled trial; 3-5 y; 610 (I: 304 and C: 306); 75.74%	Parent supported; telephone call	• A 12-wk intensive phase with weekly 90-min skills-building sessions via either in-person groups or telephone calls, • A 9-mo maintenance phase with monthly coaching telephone calls. • A 24-mo sustainability phase providing frequent cues to action (eg, texts, personalized letters, and monthly calls) to use parks and recreation programing for healthy family behaviors	36 mo	School-readiness program	PA and SB
				1 wk	Maintain usual routines	PA and SB

XSL•FO
RenderX

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
Byun et al [78]; United States	Two parallel group experimental study with randomization procedure; 2-5 y; 115 (I: 57 and C: 58); 80.87%	Teacher led; wearable device monitor	• One-week real-time PA monitoring system provided teachers with instant feedback on children's PA levels. On the basis of the feedback, teachers were asked to self-regulate their classroom strategies to provide more PA opportunities as needed. Teachers were given the autonomy to design the intervention strategies with expectations for encouraging children to be more physically active.			
Zeng et al [85]; United States	Two-arm randomized with 2-period crossover trial; 3-5 y; 34 (I: 18 and C: 16); 94.12%	Child centered; exergaming	• Families in the exergaming group received a LeapTV educational video game console (LeapFrog Enterprises, Inc), with several age-appropriate exergames, including Sports!, Dance & Learn, Kart Racing Supercharged, Pet Play World, Dora and Friends, Ultimate Spider-Man, Bubble Guppies, Sofia the First, Paw Patrol, and Jake and the Never Land Pirates. • Children used the active gaming system and engaged in exergaming 30 min per d for 5 d per wk.	12 wk, 30 min per session, 5 times per week	Maintain usual routines	PA and SB
Alexandrou et al [59]; Sweden	Two-arm parallel group individually randomized control trial; 2.5-3 y; 552 (I: 277 and C: 275); 91.12%	Parent supported; mHealth app	• Participants in the intervention group were given immediate access to the MINISTOP[g] 2.0 app (a mobile app). • The intervention builds on the MINISTOP 1.0 digital platform and content over 6 mo. • Parents used this app, receiving an extensive health-related information and feedback, to increase parental knowledge, skills, and self-efficacy to support and enable behavior change for improved diet and PA behaviors in children.	6 mo	Maintain usual routines	PA and SB
Garrison and Christakis [71][g]; United States	Two-arm randomized controlled trial; 3-5 y; 617 (I: 303 and C: 314); 91.6%	Parent supported; a telephone call and DVD		6 mo	A nutrition intervention, including monthly mailings, that encouraged families to decrease consumption of sugary drinks and increase fruit and vegetable intake.	Sleep

XSL•FO
RenderX

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
			• Intervention sessions began with the initial home visit and was then followed by mailings and follow-up telephone calls with the case manager for 12 mo. • The first 6 mailings also included DVDs with 5- to 10-min clips of suggested educational and prosocial shows. Case managers used their training to motivate parents to replace violent and age-inappropriate media content with content that was age appropriate and educational or prosocial in nature.			
Sun et al [72][g]; United States	Two-arm randomized controlled trial; 3-5 y; 32 (I: 16 and C: 16); 90.6%	Parent supported; tablet computer, videos, and telephone calls	• The intervention consisted of 8 weekly 30-min, interactive, Cantonese, educational modules accessed on the web via tablet computers. The topics were as follows: (1) introduction to the 5-4-3-2-1-0 program;(2) energy balance—maintain a healthy weight; (3) what to feed my family—energy IN; (4) grocery shopping; (5) find fun in PA—energy OUT; (6) less sit, more fit—decrease screen time; (7) smart parenting; and (8) maintain a healthy weight for life. A total of 6 of the 8 lessons were in the format of a 10- to 15-min animated short video in Cantonese, and 2 lessons were in a talk show format hosted by a bicultural registered dietitian or health educator with Cantonese-speaking mothers of young children.	8 wk 30 min per week	Received weekly mailings of printed health information to preschool-aged children over the 8 wk	PA
Fu et al [73][g]; United States	Two-arm randomized controlled trial; mean 4.9 (SD 0.7); 65 (I: 36 and C: 29); 98.5%	Child centered; exergaming	• Exergaming was incorporated into the regular school routine. During the 12-wk exergaming intervention, children had one 30-min exergaming session per day for 5 d per wk, supervised by 1 trained research assistant and teachers. • During the 30-min exergaming session, several active videogames were offered.	12 wk, 30 min per session for 5 times per week	Children in the control group had five 30-min active free-play sessions per week for 12 consecutive weeks, supervised by schoolteachers.	PA
Gao et al [17][g]; United States		Child centered; exergaming		12 wk, 30 min per session for 5 times per week	Maintain regular PA patterns without any exergaming gameplay	PA

XSL•FO
RenderX

Study; country	Study design; participants' age; sample size (I[a] and C[b]); retention rate	Intervention type; delivery	Intervention components	Intervention duration, frequency	Control group	Targeted 24-h movement behavior outcomes
	Two-arm randomized controlled trial; 4-6 y; 34 (I: 18 and C: 16); 94.12%		• Parents were instructed to have their children perform a home-based educational exergaming using LeapTV gaming console for at least 30 min per session for 5 times per week beyond usual PA (duration: 12 wk) • A phone call to parents 2 d after the LeapTV installation; and researchers also visited each family once or twice within the first 2 wk to encourage using LeapTV.			
Trost et al [74][g]; Australia	Two-arm randomized controlled trial; 3-6 y; 34 (I: 17 and C: 17); 94.12%	Child centered; mHealth app	• Moovosity is a mobile app designed to promote the development of FMS[h] and increase PA in young children. • The game is started by touching the "LETS GO!" button, initiating a 10-min timer for the child and parent to actively play the game without the use of the app.	8 wk	Waitlist control	PA
Yarimkaya et al [75][g]; Turkey	Two-group experimental study with randomization procedure; mean age 5.7; 42 (I: 21 and C: 21); 100%	Parent supported; WhatsApp	• The parents were included in this private WhatsApp group. • A 20- to 30-min PA session was held for 7 d per wk for 6 wk. Each PA session consisted of (1) a warm-up period of approximately 10 min, (2) a PA period of approximately 10 min, and (3) a cooling down period of approximately 10 min. Parents were sent 3 YouTube videos (approximately 10 min each) illustrating the 3 components of a 30-min session.	6 wk, 20- to 30-min PA session for 7 d per wk	Maintain usual routines	PA

[a]I: intervention.

[b]C: control.

[c]ECEC: early childhood education and care.

[d]PA: physical activity.

[e]SB: sedentary behavior.

[f]mHealth: mobile health.

[g]MINISTOP: mobile-based intervention intended to stop obesity in preschoolers.

[h]FMS: fundamental movement skills.

Intervention Details

The included studies used various delivery channels of eHealth technologies for the intervention. Seven studies used smartphone apps [59-61,74] and social media (Facebook and WhatsApp) [75,80,82]; 3 studies used an exergaming program [17,73,85]; 3 studies used the internet, with interventions including informational websites [83,84] and tablet computers [72]; and several studies used technology to dispatch reminders to exercise and send motivational messages encouraging persistence. Specifically, studies sent text messages and made telephone calls [71,76-79,81].

The intervention duration ranged from 1 week [78] to 36 months [77]. Seven studies had interventions that lasted >3 months [59,61,71,76,77,80,82]. Only 3 studies included follow-up

XSL•FO
RenderX

assessment after intervention, with durations of 6 weeks [84], 3 months [72], and 6 months [60]. Regarding intervention types, this study consisted of 12 studies supported by parents [59-61,71,72,75-77,79-81,84], 3 studies led by teachers [78,82,83], and 4 studies involving eHealth interventions directed at children [17,73,74,85].

The comparison groups included a waitlist control group (n=4) [74,79,81,84], education as usual (n=7) [17,59,75,78,80,82,85], and an additional non-eHealth intervention (n=8) [59-61,71-73,76,77]. A total of 14 studies targeted PA [17,59-61,72-75,77,78,80,81,83,85], 12 studies targeted SB [59-61,71,76-80,82,84,85], and 4 studies targeted sleep duration [71,76,81,84]. Notably, no studies examined all 3 movement behaviors.

Meta-Analyses

Overview

Meta-analyses demonstrated that eHealth interventions produced significant improvements in MVPA (Hedges g=0.16, 95% CI 0.03-0.30; P=.02; 7/13, 54%) and total PA (Hedges g=0.37, 95% CI 0.02-0.72; P=.04; 2/13, 15%), as shown in Figure 2A [77,78,80-83,85]. For SB outcomes, another meta-analysis showed a significant decrease (Hedges g=−0.15, 95% CI −0.27 to −0.02; P=.02; 8/13, 62%), as shown in Figure 2B [76-80,82,84,85]. Finally, meta-analysis also showed that there were significant improvements in sleep duration (Hedges g=0.47, 95% CI 0.18-0.75; P<.01; 3/13, 23%), as shown in Figure 2C [76,81,84].

Owing to the heterogeneity among the included studies, the mobile-based intervention intended to stop obesity in preschoolers (MINISTOP) project's 3 studies solely reported the difference in pre-to-post comparison [60,61,76]. Consequently, their inclusion in the pooled analysis with other studies was deemed inappropriate. We analyzed a series of MINISTOP studies separately and presented the findings using a forest plot. The pooled analysis indicated that no significant change in MVPA (Hedges g=−0.03, 95% CI −0.15 to 0.09; P=.66; 3/6, 50%; Multimedia Appendix 10 [59-61,76-85]) was observed between the intervention and control groups. An intervention effect was found in reducing SB (Hedges g=0.02, 95% CI −0.13 to 0.16; P=.83; 3/6, 50%; Multimedia Appendix 10) immediately after the intervention, as indicated in Multimedia Appendix 10. Nonetheless, this effect was not statistically significant. All the results showed negligible heterogeneity (I^2=0).

XSL•FO
RenderX

Figure 2. Forest plots: eHealth intervention versus control group (A) effect on moderate to vigorous physical activity (MVPA), (B) effect on sedentary time (SED), and (C) effect on sleep [76-85].

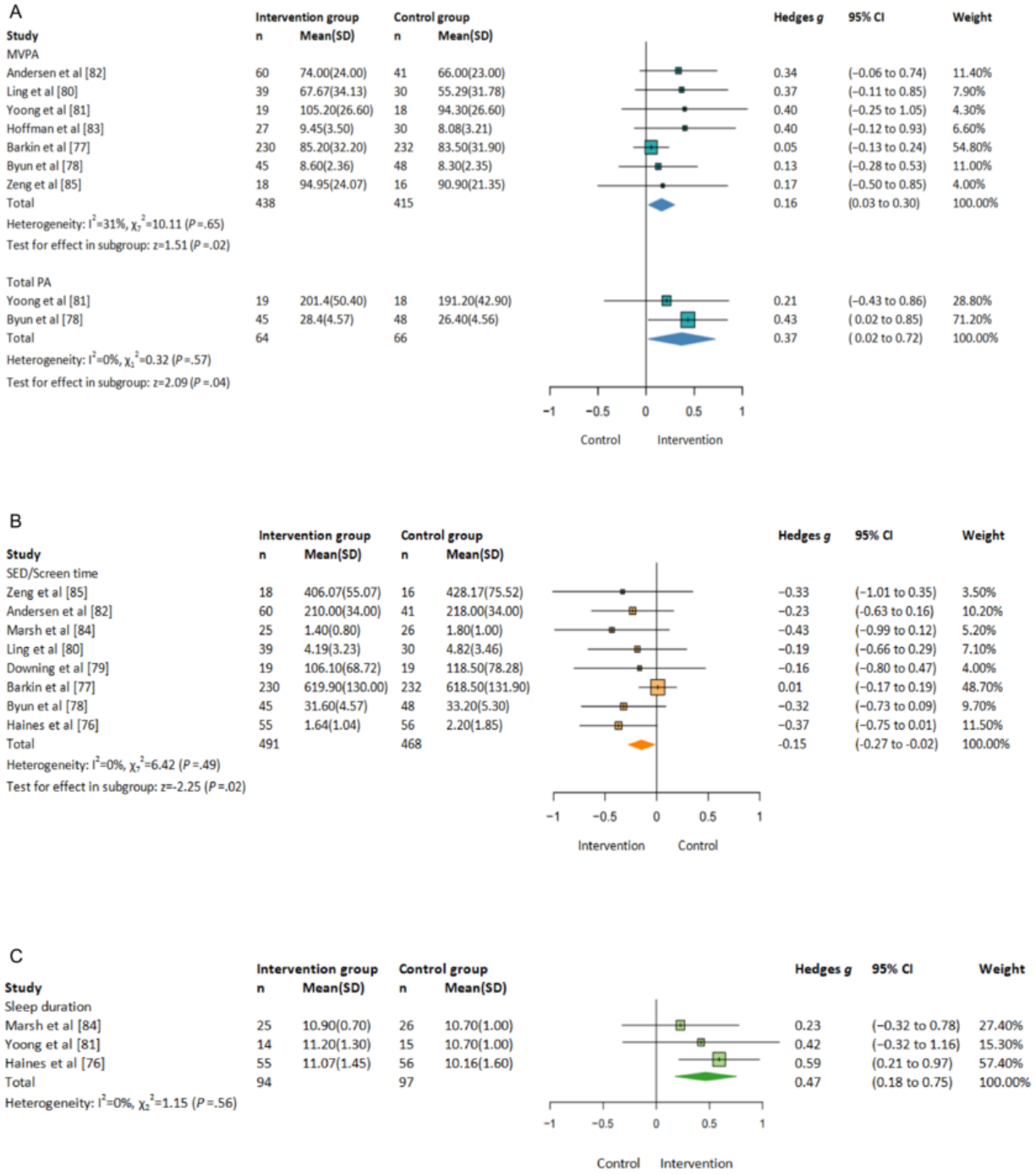

Subgroup Analyses and Meta-Regression

Table 2 shows the subgroup analysis and meta-regression results of MVPA and sedentary time according to study characteristics. No significant moderating effects were observed for any of the variables assessed (P>.05). The complete results of the subgroup analyses are presented in Multimedia Appendix 11 [59-61,76-85].

XSL•FO
RenderX

Table 2. Subgroup analysis and meta-regression results of MVPA[a] and sedentary time.

	MVPA				Sedentary time			
	Studies, n	Hedges *g* (95% CI)	I^2 (%)	*P* value	Studies, n	Hedges *g* (95% CI)	I^2 (%)	*P* value
Overall	7	0.16 (0.03 to 0.30)	0	.02	8	−0.15 (−0.27 to −0.02)	0	.02
Duration				.59				.30
0-3 mo	4	0.25 (−0.02 to 0.52)	25.9		4	−0.32 (−0.59 to −0.05)	22.4	
>3 mo	3	0.17 (−0.04 to 0.38)	74.1		4	−0.13 (−0.32 to 0.05)	77.6	
Risk of bias				.14				.33
Low risk	2	0.06 (−0.11 to 0.24)	58.8		4	−0.13 (−0.34 to 0.08)	67.8	
Some concerns	5	0.3 (0.09 to 0.51)	41.2		4	−0.28 (−0.51 to −0.06)	32.2	
Measurement				N/A[b]				.20
Objective	7	N/A	N/A		6	−0.1 (−0.24 to 0.04)	16.7	
Self-reported	0	N/A	N/A		2	−0.39 (−0.70 to −0.08)	83.3	
Type				N/A				N/A
Teacher focused	3	0.27 (0.02 to 0.52)	29	N/A	2	−0.28 (−0.56 to 0.01)	19.9	N/A
Parent supported	3	0.24 (−0.07 to 0.34)	67	.40[c]	5	−0.15 (−0.33 to 0.04)	76.6	.57[c]
Child centered	1	0.16 (0.03 to 0.30)	4	.80[c]	1	−0.15 (−0.27 to −0.02)	3.5	.90[c]

[a]MVPA: moderate to vigorous physical activity.

[b]N/A: not applicable.

[c]Teacher focused studies as a reference group.

Sensitivity Analyses and Publication Bias

Sensitivity analysis indicated that no individual study had an excessive influence on the results. The omitted meta-analytic estimates were not significantly different from those associated with the combined analysis, and all estimates were within the 95% CI. Forest plots of the sensitivity analysis for MVPA, sedentary time, and sleep are summarized in Multimedia Appendix 12 [59-61,76-85]. The significance of Egger's test results provided evidence for asymmetry of the funnel plots (MVPA: t_5=3.27; *P*=.02; Multimedia Appendix 13; sedentary time: t_6=−3.37; *P*=.02; Multimedia Appendix 14). However, we could not distinguish chance from true asymmetry using the funnel plot asymmetry test because <10 studies were included in our meta-analysis [86].

ROB of Studies

Multimedia Appendix 15 [59-61,76-85] summarizes the overall ROB assessment for all the included papers. Six studies were considered to have a low ROB [59,74,76,77,79,85], and the remaining 13 were considered to have some concerns regarding the ROB [17,60,61,71-73,75,78,80-84]. Furthermore, 7 studies did not disclose randomization methods clearly [17,72,75,78,80,82,83], so they were rated as having some concerns about random sequence generation. All studies were rated as having a low risk for the measurement of outcomes based on the use of objective measurement tools or reliable questionnaires in each study. Four studies were rated as 'some concerns' of reporting bias because neither published study protocols nor registered trial records were presented [72,75,78,80].

Quality of the Evidence

The GRADE scores are shown in Multimedia Appendix 16, and we deemed the overall quality of evidence to be moderate to low. The quality of evidence for MVPA and sedentary time outcomes was rated as "moderate," considering the low ROB, absence of heterogeneity in participants' outcomes, and high precision in results. As eHealth interventions are often combined with other intervention approaches, all evaluations of directness were assessed as "Indirectness." There were high imprecisions with the sample size included in the study for total PA and sleep, which were graded as "Low."

Discussion

Principal Findings

This study systematically reviewed the effectiveness of eHealth interventions targeting 24-hour movement behaviors among preschool-aged children. Most studies assessed interventions aimed at increasing PA and decreasing SB. Few studies targeted sleep, and no studies have addressed a combination of all 24-hour movement behaviors. Overall, these studies showed trends supporting the effectiveness of eHealth interventions in increasing PA and sleep duration and reducing sedentary time immediately after the intervention; however, only short-term effects were found, and all trials were judged to be of low to moderate quality.

XSL•FO
RenderX

This review demonstrates a small positive effect of eHealth interventions targeting increases in preschooler's MVPA (Hedges g=0.16) and total PA (Hedges g=0.37) immediately after the intervention. One possible explanation could be that eHealth interventions, while providing new opportunities for PA, might not be sufficient to result in significant overall activity increases. This might require expanding activity opportunities, extending new activity options, and enhancing broader activity strategies to achieve substantial benefits. Our findings echo the argument made in a previous study of young children that PA interventions had a small effect on MVPA [87]. Another meta-analysis found a positive impact of PA interventions with small to moderate effects on total PA (Hedges g=0.44) and moderate effects on MVPA (Hedges g=0.51) [29]. There is no conclusive explanation as to why MVPA and total PA were seen to have a smaller effect in our study, but this could be attributed to most interventions thus far concentrating on devising PA programs of diverse intensities without distinct objectives, including low-intensity PA, MVPA, and total PA (eg, activities such as outdoor active play and structured gross motor activity sessions in childcare environments). Moreover, our results are consistent with previous review findings that digital platforms can potentially increase PA among preschoolers [53]. Hence, future interventions should aim to optimize their effectiveness in increasing PA among young children. In addition, further research is warranted to investigate the mechanisms of the changes associated with these PA outcomes. This will help enhance the size and sustainability of the effects observed in eHealth interventions.

We found no significant improvement in MVPA for mobile app interventions (MINISTOP project). This is in contrast to a review of studies focusing on mobile apps and technologies, which highlighted the significant potential to enhance PA [88]. It is worth noting that the MINISTOP project aimed to reduce obesity as its primary outcome rather than targeting MVPA. In addition, studies concentrating solely on educating parents without implementing direct interventions for children have not achieved the desired enhancements in MVPA. Thus, we cannot draw conclusions about mobile apps because few intervention studies have used these means of communication for young children and their guardians. Given the small number of studies included in our meta-analysis, the positive, negative, and null findings of the individual studies may have attenuated the results. Thus, considering the popularity and cost-effectiveness of mobile apps in the new generation, future research should investigate the potential of using emerging and novel technologies, such as mobile health, for preschoolers.

Our meta-analysis suggests that eHealth interventions may be an effective strategy for decreasing sedentary time in preschoolers, although the magnitude of the effect was small (Hedges g=−0.15) and short term. Nonetheless, the significance should not be understated, given that many studies indicate that reduced sedentary time during childhood correlates with improved physical and mental health outcomes in subsequent years [16,21,89]. In the subgroup analysis, the effect of eHealth interventions on sedentary time varied depending on whether accelerometer or questionnaire measures were used. The questionnaire measures yielded higher levels of sedentary time, although this difference was not statistically significant. This observation aligns with findings from the existing literature, suggesting that questionnaire-based assessments tend to overestimate the actual sedentary time. For a more accurate evaluation of the impact of eHealth interventions, future research should consider using device-based measurement methods [90].

Interestingly, most eHealth interventions aimed to increase children's PA and reduce sedentary time with parental support. Previous research has shown that parental and family involvement were among the key intervention components that encouraged significant improvement in children's health behaviors and a decrease in sedentary time [91,92]. Likewise, Ha et al [49] found that parents' physical literacy predicts children's values toward PA, and concurrent interventions that target enhancing parents' physical literacy for PA in the family context may be more effective in raising children's PA values. However, our subgroup analysis showed no significant improvements in MVPA or reductions in sedentary time with the parent-supported interventions. This result also aligns with a prior review indicating that parent-directed digital interventions were ineffective in improving PA [53]. In that review, 8 studies, all published before 2020, primarily used digital platforms to convey health information and education to parents. Notably, in the wake of the COVID-19 pandemic, there has been a marked increase in research centered on leveraging technology to improve children's PA, leading to more recent studies in 3 years [93]. Furthermore, the discourse regarding the comparative value of targeting either parents or children exclusively is not a novel debate within intervention research. In contrast to the review, our study featured a larger sample size and included a quantitative analysis of effect sizes in the interventions. These insights indicate that prevailing eHealth interventions, even with parental support, may fail to effectively engage preschoolers. Recognizing the reciprocal dynamics between parents and young children can offer insights for refining digital interventions. Therefore, preliminary research is imperative to comprehensively understand the perceptions, attitudes, and driving factors of parents. Recognizing the reciprocal dynamics between parents and young children is crucial in understanding how they influence their children's PA and SB.

Intervention duration is also an essential component for conducting acceptable and highly effective interventions. Another subgroup analysis found that interventions with a duration of <3 months had a significantly greater effect on PA and sedentary time than those with a duration of >3 months, although the results were not significant. This notion is corroborated by another systematic review, which demonstrated the difficulty in sustaining long-term behavior change, potentially attributed to the diminishing effects of behavior change interventions mediated by digital technology [41].

The meta-analysis, involving 3 studies, revealed an immediate improvement in sleep duration following the intervention. Previous research has extensively examined the influence of sleep duration during the preschool years on physical, cognitive, and psychosocial development. For instance, the systematic review by Chaput et al [6] involving 25 studies revealed a correlation between shorter sleep duration and diminished

XSL•FO
RenderX

emotion regulation in children aged 0 to 4 years. Recent findings also suggest that maintaining an extended sleep duration during the early preschool stages is significant for subsequent behavioral outcomes [24]. However, few studies have focused on effective interventions to improve sleep outcomes [45,94]. Consequently, further research is warranted to explore the impact of eHealth interventions on sleep outcomes among preschoolers.

Increasing awareness of the interconnected nature of 24-hour movement behaviors highlights their intrinsic interdependence [14]. However, none of the studies in our review specifically investigated the intervention effects on all 3 movement behaviors. Generally, conventional analytical methods do not adequately consider these indicators during analysis. Therefore, future research should explore alternative approaches, such as compositional analyses, to attain a more profound comprehension of whether an optimal equilibrium is present among SB, light PA, MVPA, and sleep [90,95,96]. Furthermore, most studies in our review examined the immediate postintervention effect. Consequently, insights into the enduring nature of alterations in 24-hour movement behaviors remain elusive. Further studies should include long-term follow-up assessments. In addition, it would be interesting to obtain more insights into the feasibility of incorporating wearable devices and apps into the design of eHealth interventions. This information could inform the design of wearables and apps that effectively enhance PA, diminish sedentary time, and enhance sleep, thereby maximizing their impact on public health. Moreover, the overall quality of the interventions was suboptimal, lacking thorough descriptions or proper execution in areas such as randomization, blinded outcome assessment, valid measurement of 24-hour movement behaviors, and adjusted differences between groups. In our meta-analysis, we observed that lower-quality studies exhibited a more pronounced positive impact on the targeted outcomes. Thus, it is essential to interpret the results cautiously, recognizing that there could be an overestimation of the effect of eHealth interventions in studies of lower quality owing to potential bias. This mirrors the findings from previous reviews on eHealth childhood PA [53] and behavior change interventions among adolescents [45].

Strengths and Limitations

This systematic review has some strengths. First, this study is the first meta-analysis to quantitatively assess the effects of previously conducted RCTs using eHealth interventions on 24-hour movement behaviors in preschoolers. Second, the review was conducted rigorously, encompassing comprehensive terms and using an extensive systematic search strategy. We focused on robust evidence from RCT studies, assessed the quality using the GRADE approach, and adhered to a preregistered protocol. This meticulous approach reduces the heterogeneity and provides a more precise estimation of the effects.

Nonetheless, several limitations of our study should be noted. First, the quality of the studies included in this review was generally low and lacked rigorous study designs. Second, the small number of studies discerned over the decade spanned by this meta-analysis underscores the nascent state of this research domain, even considering significant technological advancements and their widespread acceptance. Third, although we systematically screened relevant electronic databases to identify studies, the search was restricted to studies published in English. Finally, the lack of evidence regarding sustained effects beyond the immediate postintervention period underscores the need for extended follow-up. Future studies should strive to elucidate strategies for maintaining the intervention effects over the preschooler's trajectory.

Future Research and Implications

This study highlights the significant avenues for future research. First, further research is warranted to develop eHealth interventions that yield larger effect sizes and higher quality, specifically in identifying effective 24-hour movement behaviors. It is worth noting that none of the eligible eHealth interventions addressed the comprehensive integration of 24-hour movement behaviors in preschoolers, despite the increasing recognition of the interdependence between PA, SB), and sleep. Second, many studies were conducted in Western and high-income countries, prompting the need for further exploration of the effectiveness of eHealth behavior change interventions in other country settings. Third, our study's focus was primarily on the quantitative aspects of 24-hour movement behaviors, warranting future studies to also delve into the qualitative facets, such as motor skills and sleep quality. In addition, it is crucial to recognize the pivotal role of objective measurement tools in comprehending movement behaviors among young children. Given the sporadic and unstructured nature of preschoolers' activities, it becomes challenging for parents and teachers to accurately discern shifts in MVPA and SB, even if they have occurred. This highlights the importance of using objective measurement tools for precise insights into these behaviors. Finally, future research in this field should prioritize broadening the focus and incorporate additional dimensions, such as physical, affective, and cognitive indicators. This approach may promote the holistic development of young children and contribute to advancements in the field of health outcomes. By considering these dimensions, researchers can also gain a comprehensive understanding of the various factors that influence children's overall well-being and physical literacy development.

Given the multifaceted nature of intervention moderators, further research is warranted to establish optimal patterns of daily movement behaviors and to gain deeper insights into the mechanisms underlying change when addressing the amalgamation of 24-hour movement behaviors in preschoolers. Indeed, future interventions should also draw from the effective behavior change techniques used in single-behavior eHealth interventions and apply them to interventions targeting multiple healthy movement behaviors. Moreover, collaborative engagement with parents and teachers throughout both the developmental and implementation phases of these interventions will play a pivotal role in their success. In addition, capitalizing on emerging and novel technologies may offer a valuable avenue to enhance the effectiveness and feasibility of these interventions.

XSL•FO
RenderX

Conclusions

The findings suggest that eHealth interventions may hold promise in improving 24-hour movement behaviors, particularly by increasing PA, improving sleep duration, and reducing sedentary time among preschoolers. However, these effects were relatively modest and transient and were observed primarily immediately after the intervention. Furthermore, the overall quality of the evidence was rated as moderate to low. As a result, there is a pressing need for rigorous and high-quality research endeavors to develop eHealth interventions capable of effectively enhancing both the quantity and quality of 24-hour movement behaviors simultaneously. These interventions should strive to maintain their effects over extended periods.

Acknowledgments

The authors of this study would like to express their sincere gratitude to the authors who responded to their emails and generously provided detailed information and data regarding their studies. Their cooperation has been instrumental in advancing this study.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SJ drafted the manuscript. SJ, ASH, and JYYN were responsible for the concept and design of the study. SJ and BP screened all abstracts full texts, extracted all data, performed the risk of bias, and conducted the quality assessment. SJ performed the statistical analyses. SJ, JYYN, KHC, and ASH critically revised the manuscript for important intellectual content. All authors participated in developing the review's methodology, contributed to multiple manuscript drafts, and gave their approval for the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Eligibility criteria for study inclusion.
[DOCX File , 15 KB - jmir_v26i1e52905_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 32 KB - jmir_v26i1e52905_app2.docx]

Multimedia Appendix 3
Missing data processing.
[DOCX File , 22 KB - jmir_v26i1e52905_app3.docx]

Multimedia Appendix 4
Exclusion studies.
[DOCX File , 27 KB - jmir_v26i1e52905_app4.docx]

Multimedia Appendix 5
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 27 KB - jmir_v26i1e52905_app5.docx]

Multimedia Appendix 6
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) abstract checklist.
[DOC File , 21 KB - jmir_v26i1e52905_app6.doc]

Multimedia Appendix 7
Number of studies included per country and income economy.
[DOCX File , 17 KB - jmir_v26i1e52905_app7.docx]

Multimedia Appendix 8
Summary of intervention details in the included studies.

XSL•FO
RenderX

[XLSX File (Microsoft Excel File), 14 KB - jmir_v26i1e52905_app8.xlsx]

Multimedia Appendix 9
Characteristics of the included studies including physical activity, sedentary behavior, and sleep outcomes.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e52905_app9.xlsx]

Multimedia Appendix 10
Forest plot of the mobile-based intervention intended to stop obesity in preschoolers (MINISTOP) results.
[DOCX File , 552 KB - jmir_v26i1e52905_app10.docx]

Multimedia Appendix 11
Forest plots of the subgroup analyses of moderate to vigorous physical activity and sedentary behavior.
[DOCX File , 3185 KB - jmir_v26i1e52905_app11.docx]

Multimedia Appendix 12
Sensitive analysis.
[PDF File (Adobe PDF File), 568 KB - jmir_v26i1e52905_app12.pdf]

Multimedia Appendix 13
Moderate to vigorous physical activity bias funnel.
[PDF File (Adobe PDF File), 11 KB - jmir_v26i1e52905_app13.pdf]

Multimedia Appendix 14
Sedentary behavior bias funnel.
[PDF File (Adobe PDF File), 11 KB - jmir_v26i1e52905_app14.pdf]

Multimedia Appendix 15
Risk of bias.
[PDF File (Adobe PDF File), 1496 KB - jmir_v26i1e52905_app15.pdf]

Multimedia Appendix 16
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) assessment results.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v26i1e52905_app16.xlsx]

References

1. Shirazipour CH, Raines C, Diniz MA, Salvy SJ, Haile RW, Freedland SJ, et al. The 24-hour movement paradigm: an integrated approach to the measurement and promotion of daily activity in cancer clinical trials. Contemp Clin Trials Commun 2023 Apr;32:101081 [FREE Full text] [doi: 10.1016/j.conctc.2023.101081] [Medline: 36875555]
2. WHO guidelines on physical activity and sedentary behaviour. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240015128 [accessed 2024-01-18]
3. Guidelines on physical activity, sedentary behaviour and sleep for children under 5 years of age. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/311664/9789241550536-eng.pdf?sequence=1 [accessed 2024-01-18]
4. Kuzik N, Poitras VJ, Tremblay MS, Lee EY, Hunter S, Carson V. Systematic review of the relationships between combinations of movement behaviours and health indicators in the early years (0-4 years). BMC Public Health 2017 Nov 20;17(Suppl 5):849 [FREE Full text] [doi: 10.1186/s12889-017-4851-1] [Medline: 29219071]
5. Rollo S, Antsygina O, Tremblay MS. The whole day matters: understanding 24-hour movement guideline adherence and relationships with health indicators across the lifespan. J Sport Health Sci 2020 Dec;9(6):493-510 [FREE Full text] [doi: 10.1016/j.jshs.2020.07.004] [Medline: 32711156]
6. Chaput JP, Gray CE, Poitras VJ, Carson V, Gruber R, Birken CS, et al. Systematic review of the relationships between sleep duration and health indicators in the early years (0-4 years). BMC Public Health 2017 Nov 20;17(Suppl 5):855 [FREE Full text] [doi: 10.1186/s12889-017-4850-2] [Medline: 29219078]
7. Hu R, Zheng H, Lu C. The association between sedentary screen time, non-screen-based sedentary time, and overweight in Chinese preschool children: a cross-sectional study. Front Pediatr 2021 Dec;9:767608 [FREE Full text] [doi: 10.3389/fped.2021.767608] [Medline: 35004541]
8. McArthur BA, Volkova V, Tomopoulos S, Madigan S. Global prevalence of meeting screen time guidelines among children 5 years and younger: a systematic review and meta-analysis. JAMA Pediatr 2022 Apr 01;176(4):373-383 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.6386] [Medline: 35157028]

XSL•FO
RenderX

9. da Costa BG, Chaput JP, Lopes MV, Malheiros LE, Silva KS. Movement behaviors and their association with depressive symptoms in Brazilian adolescents: a cross-sectional study. J Sport Health Sci 2022 Mar;11(2):252-259 [FREE Full text] [doi: 10.1016/j.jshs.2020.08.003] [Medline: 32791204]
10. Del Pozo-Cruz B, Perales F, Parker P, Lonsdale C, Noetel M, Hesketh KD, et al. Joint physical-activity/screen-time trajectories during early childhood: socio-demographic predictors and consequences on health-related quality-of-life and socio-emotional outcomes. Int J Behav Nutr Phys Act 2019 Jul 08;16(1):55 [FREE Full text] [doi: 10.1186/s12966-019-0816-3] [Medline: 31286983]
11. Okely AD, Kariippanon KE, Guan H, Taylor EK, Suesse T, Cross PL, et al. Global effect of COVID-19 pandemic on physical activity, sedentary behaviour and sleep among 3- to 5-year-old children: a longitudinal study of 14 countries. BMC Public Health 2021 May 17;21(1):940 [FREE Full text] [doi: 10.1186/s12889-021-10852-3] [Medline: 34001086]
12. Okely AD, Reilly JJ, Tremblay MS, Kariippanon KE, Draper CE, El Hamdouchi A, et al. Cross-sectional examination of 24-hour movement behaviours among 3- and 4-year-old children in urban and rural settings in low-income, middle-income and high-income countries: the SUNRISE study protocol. BMJ Open 2021 Oct 25;11(10):e049267 [FREE Full text] [doi: 10.1136/bmjopen-2021-049267] [Medline: 34697112]
13. Tucker P. The physical activity levels of preschool-aged children: a systematic review. Early Child Res Q 2008 Oct;23(4):547-558 [FREE Full text] [doi: 10.1016/j.ecresq.2008.08.005]
14. Feng J, Zheng C, Sit CH, Reilly JJ, Huang WY. Associations between meeting 24-hour movement guidelines and health in the early years: a systematic review and meta-analysis. J Sports Sci 2021 Nov 28;39(22):2545-2557 [FREE Full text] [doi: 10.1080/02640414.2021.1945183] [Medline: 34176439]
15. Feng J, Huang WY, Reilly JJ, Wong SH. Compliance with the WHO 24-h movement guidelines and associations with body weight status among preschool children in Hong Kong. Appl Physiol Nutr Metab 2021 Oct;46(10):1273-1278 [FREE Full text] [doi: 10.1139/apnm-2020-1035] [Medline: 33945770]
16. Rodriguez-Ayllon M, Cadenas-Sánchez C, Estévez-López F, Muñoz NE, Mora-Gonzalez J, Migueles JH, et al. Role of physical activity and sedentary behavior in the mental health of preschoolers, children and adolescents: a systematic review and meta-analysis. Sports Med 2019 Sep 16;49(9):1383-1410. [doi: 10.1007/s40279-019-01099-5] [Medline: 30993594]
17. Gao Z, Lee JE, Zeng N, Pope ZC, Zhang Y, Li X. Home-based exergaming on preschoolers' energy expenditure, cardiovascular fitness, body mass index and cognitive flexibility: a randomized controlled trial. J Clin Med 2019 Oct 21;8(10):1745 [FREE Full text] [doi: 10.3390/jcm8101745] [Medline: 31640158]
18. Wong RS, Tung KT, Chan BN, Ho FK, Rao N, Chan KL, et al. Early-life activities mediate the association between family socioeconomic status in early childhood and physical fitness in early adolescence. Sci Rep 2022 Jan 07;12(1):81 [FREE Full text] [doi: 10.1038/s41598-021-03883-8] [Medline: 34996964]
19. Zeng N, Ayyub M, Sun H, Wen X, Xiang P, Gao Z. Effects of physical activity on motor skills and cognitive development in early childhood: a systematic review. Biomed Res Int 2017;2017:2760716 [FREE Full text] [doi: 10.1155/2017/2760716] [Medline: 29387718]
20. Carson V, Hunter S, Kuzik N, Wiebe SA, Spence JC, Friedman A, et al. Systematic review of physical activity and cognitive development in early childhood. J Sci Med Sport 2016 Jul;19(7):573-578. [doi: 10.1016/j.jsams.2015.07.011] [Medline: 26197943]
21. Talarico R, Janssen I. Compositional associations of time spent in sleep, sedentary behavior and physical activity with obesity measures in children. Int J Obes (Lond) 2018 Aug;42(8):1508-1514. [doi: 10.1038/s41366-018-0053-x] [Medline: 29568110]
22. Cliff DP, McNeill J, Vella SA, Howard SJ, Santos R, Batterham M, et al. Adherence to 24-hour movement guidelines for the early years and associations with social-cognitive development among Australian preschool children. BMC Public Health 2017 Nov 20;17(Suppl 5):857 [FREE Full text] [doi: 10.1186/s12889-017-4858-7] [Medline: 29219104]
23. Christian H, Murray K, Trost SG, Schipperijn J, Trapp G, Maitland C, et al. Meeting the Australian 24-hour movement guidelines for the early years is associated with better social-emotional development in preschool boys. Prev Med Rep 2022 Jun;27:101770 [FREE Full text] [doi: 10.1016/j.pmedr.2022.101770] [Medline: 35321215]
24. Guerlich K, Avraam D, Cadman T, Calas L, Charles MA, Elhakeem A, et al. Sleep duration in preschool age and later behavioral and cognitive outcomes: an individual participant data meta-analysis in five European cohorts. Eur Child Adolesc Psychiatry 2024 Jan;33(1):167-177. [doi: 10.1007/s00787-023-02149-0] [Medline: 36749392]
25. Johnstone A, Hughes AR, Martin A, Reilly JJ. Utilising active play interventions to promote physical activity and improve fundamental movement skills in children: a systematic review and meta-analysis. BMC Public Health 2018 Jun 26;18(1):789 [FREE Full text] [doi: 10.1186/s12889-018-5687-z] [Medline: 29940923]
26. Lee AM, Chavez S, Bian J, Thompson LA, Gurka MJ, Williamson VG, et al. Efficacy and effectiveness of mobile health technologies for facilitating physical activity in adolescents: scoping review. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e11847 [FREE Full text] [doi: 10.2196/11847] [Medline: 30747716]
27. Morgan EH, Schoonees A, Sriram U, Faure M, Seguin-Fowler RA. Caregiver involvement in interventions for improving children's dietary intake and physical activity behaviors. Cochrane Database Syst Rev 2020 Jan 05;1(1):CD012547 [FREE Full text] [doi: 10.1002/14651858.CD012547.pub2] [Medline: 31902132]

XSL•FO
RenderX

28. Hammersley ML, Wyse RJ, Jones RA, Okely AD, Wolfenden L, Eckermann S, et al. Telephone and web-based delivery of healthy eating and active living interventions for parents of children aged 2 to 6 years: mixed methods process evaluation of the time for healthy habits translation trial. J Med Internet Res 2022 May 26;24(5):e35771 [FREE Full text] [doi: 10.2196/35771] [Medline: 35616994]
29. Gordon ES, Tucker P, Burke SM, Carron AV. Effectiveness of physical activity interventions for preschoolers: a meta-analysis. Res Q Exerc Sport 2013 Sep;84(3):287-294. [doi: 10.1080/02701367.2013.813894] [Medline: 24261007]
30. Fangupo LJ, Haszard JJ, Reynolds AN, Lucas AW, McIntosh DR, Richards R, et al. Do sleep interventions change sleep duration in children aged 0-5 years? A systematic review and meta-analysis of randomised controlled trials. Sleep Med Rev 2021 Oct;59:101498 [FREE Full text] [doi: 10.1016/j.smrv.2021.101498] [Medline: 34029803]
31. Downing KL, Hnatiuk JA, Hinkley T, Salmon J, Hesketh KD. Interventions to reduce sedentary behaviour in 0-5-year-olds: a systematic review and meta-analysis of randomised controlled trials. Br J Sports Med 2018 Mar 06;52(5):314-321 [FREE Full text] [doi: 10.1136/bjsports-2016-096634] [Medline: 29449219]
32. Oh H, Rizo C, Enkin M, Jadad A. What is eHealth (3): a systematic review of published definitions. J Med Internet Res 2005 Feb 24;7(1):e1 [FREE Full text] [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]
33. Boogerd EA, Arts T, Engelen LJ, van de Belt TH. "What is eHealth": time for an update? JMIR Res Protoc 2015 Mar 12;4(1):e29 [FREE Full text] [doi: 10.2196/resprot.4065] [Medline: 25768939]
34. Liu S, Li J, Wan DY, Li R, Qu Z, Hu Y, et al. Effectiveness of eHealth self-management interventions in patients with heart failure: systematic review and meta-analysis. J Med Internet Res 2022 Sep 26;24(9):e38697 [FREE Full text] [doi: 10.2196/38697] [Medline: 36155484]
35. Brown HE, Atkin AJ, Panter J, Wong G, Chinapaw MJ, van Sluijs EM. Family-based interventions to increase physical activity in children: a systematic review, meta-analysis and realist synthesis. Obes Rev 2016 Apr;17(4):345-360 [FREE Full text] [doi: 10.1111/obr.12362] [Medline: 26756281]
36. He Z, Wu H, Yu F, Fu J, Sun S, Huang T, et al. Effects of smartphone-based interventions on physical activity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Feb 01;9(2):e22601 [FREE Full text] [doi: 10.2196/22601] [Medline: 33522980]
37. Bonvicini L, Pingani I, Venturelli F, Patrignani N, Bassi MC, Broccoli S, et al. Effectiveness of mobile health interventions targeting parents to prevent and treat childhood obesity: systematic review. Prev Med Rep 2022 Oct;29:101940 [FREE Full text] [doi: 10.1016/j.pmedr.2022.101940] [Medline: 36161123]
38. Western MJ, Armstrong ME, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Nov 09;18(1):148 [FREE Full text] [doi: 10.1186/s12966-021-01218-4] [Medline: 34753490]
39. Lau PW, Lau EY, Wong DP, Ransdell L. A systematic review of information and communication technology-based interventions for promoting physical activity behavior change in children and adolescents. J Med Internet Res 2011 Jul 13;13(3):e48 [FREE Full text] [doi: 10.2196/jmir.1533] [Medline: 21749967]
40. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017 Dec;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
41. Stephenson A, McDonough SM, Murphy MH, Nugent CD, Mair JL. Using computer, mobile and wearable technology enhanced interventions to reduce sedentary behaviour: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2017 Aug 11;14(1):105 [FREE Full text] [doi: 10.1186/s12966-017-0561-4] [Medline: 28800736]
42. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
43. Duan Y, Shang B, Liang W, Du G, Yang M, Rhodes RE. Effects of eHealth-based multiple health behavior change interventions on physical activity, healthy diet, and weight in people with noncommunicable diseases: systematic review and meta-analysis. J Med Internet Res 2021 Feb 22;23(2):e23786 [FREE Full text] [doi: 10.2196/23786] [Medline: 33616534]
44. Deng W, M J J van der Kleij R, Shen H, Wei J, Brakema EA, Guldemond N, et al. eHealth-based psychosocial interventions for adults with insomnia: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Mar 14;25:e39250 [FREE Full text] [doi: 10.2196/39250] [Medline: 36917145]
45. Champion KE, Parmenter B, McGowan C, Spring B, Wafford QE, Gardner LA, Health4Life team. Effectiveness of school-based eHealth interventions to prevent multiple lifestyle risk behaviours among adolescents: a systematic review and meta-analysis. Lancet Digit Health 2019 Sep;1(5):e206-e221 [FREE Full text] [doi: 10.1016/S2589-7500(19)30088-3] [Medline: 33323269]
46. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
47. Kozak AT, Buscemi J, Hawkins MA, Wang ML, Breland JY, Ross KM, et al. Technology-based interventions for weight management: current randomized controlled trial evidence and future directions. J Behav Med 2017 Feb;40(1):99-111 [FREE Full text] [doi: 10.1007/s10865-016-9805-z] [Medline: 27783259]

XSL•FO
RenderX

48. Benzing V, Schmidt M. Exergaming for children and adolescents: strengths, weaknesses, opportunities and threats. J Clin Med 2018 Nov 08;7(11):422 [FREE Full text] [doi: 10.3390/jcm7110422] [Medline: 30413016]
49. Ha AS, Jia J, Ng FF, Ng JY. Parent's physical literacy enhances children's values towards physical activity: a serial mediation model. Psychol Sport Exerc 2022 Nov;63:102297 [FREE Full text] [doi: 10.1016/j.psychsport.2022.102297]
50. Hammersley ML, Jones RA, Okely AD. Parent-focused childhood and adolescent overweight and obesity eHealth interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jul 21;18(7):e203 [FREE Full text] [doi: 10.2196/jmir.5893] [Medline: 27443862]
51. Cornish K, Fox G, Fyfe T, Koopmans E, Pousette A, Pelletier CA. Understanding physical literacy in the context of health: a rapid scoping review. BMC Public Health 2020 Oct 19;20(1):1569 [FREE Full text] [doi: 10.1186/s12889-020-09583-8] [Medline: 33076887]
52. Ha AS, He Q, Lubans DR, Chan CH, Ng JY. Parent-focused online intervention to promote parents' physical literacy and support children's physical activity: results from a quasi-experimental trial. BMC Public Health 2022 Jul 12;22(1):1330 [FREE Full text] [doi: 10.1186/s12889-022-13739-z] [Medline: 35820873]
53. Swindle T, Poosala AB, Zeng N, Børsheim E, Andres A, Bellows LL. Digital intervention strategies for increasing physical activity among preschoolers: systematic review. J Med Internet Res 2022 Jan 11;24(1):e28230 [FREE Full text] [doi: 10.2196/28230] [Medline: 35014962]
54. International early learning and child well-being study. Organisation for Economic Co-operation and Development. 2020. URL: https://www.oecd.org/education/school/early-learning-and-child-well-being-study/ [accessed 2024-01-18]
55. Kabali HK, Irigoyen MM, Nunez-Davis R, Budacki JG, Mohanty SH, Leister KP, et al. Exposure and use of mobile media devices by young children. Pediatrics 2015 Dec;136(6):1044-1050. [doi: 10.1542/peds.2015-2151] [Medline: 26527548]
56. McCloskey ML, Thompson DA, Chamberlin B, Clark L, Johnson SL, Bellows LL. Mobile device use among rural, low-income families and the feasibility of an app to encourage preschoolers' physical activity: qualitative study. JMIR Pediatr Parent 2018 Dec 06;1(2):e10858 [FREE Full text] [doi: 10.2196/10858] [Medline: 31518295]
57. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
58. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009 Jul 21;339(jul21 1):b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
59. Alexandrou C, Henriksson H, Henström M, Henriksson P, Delisle Nyström C, Bendtsen M, et al. Effectiveness of a smartphone app (MINISTOP 2.0) integrated in primary child health care to promote healthy diet and physical activity behaviors and prevent obesity in preschool-aged children: randomized controlled trial. Int J Behav Nutr Phys Act 2023 Feb 21;20(1):22 [FREE Full text] [doi: 10.1186/s12966-023-01405-5] [Medline: 36810069]
60. Delisle Nyström C, Sandin S, Henriksson P, Henriksson H, Maddison R, Löf M. A 12-month follow-up of a mobile-based (mHealth) obesity prevention intervention in pre-school children: the MINISTOP randomized controlled trial. BMC Public Health 2018 May 24;18(1):658 [FREE Full text] [doi: 10.1186/s12889-018-5569-4] [Medline: 29793467]
61. Nyström CD, Sandin S, Henriksson P, Henriksson H, Trolle-Lagerros Y, Larsson C, et al. Mobile-based intervention intended to stop obesity in preschool-aged children: the MINISTOP randomized controlled trial. Am J Clin Nutr 2017 Jun;105(6):1327-1335 [FREE Full text] [doi: 10.3945/ajcn.116.150995] [Medline: 28446496]
62. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2019.
63. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
64. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019 Nov 28;22(4):153-160 [FREE Full text] [doi: 10.1136/ebmental-2019-300117] [Medline: 31563865]
65. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
66. Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022 Aug;4(8):e615-e626 [FREE Full text] [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]
67. Ho RS, Chan EK, Liu KK, Wong SH. Active video game on children and adolescents' physical activity and weight management: a network meta-analysis. Scand J Med Sci Sports 2022 Aug;32(8):1268-1286. [doi: 10.1111/sms.14176] [Medline: 35485251]
68. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
69. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ, GRADE Working Group. What is "quality of evidence" and why is it important to clinicians? BMJ 2008 May 03;336(7651):995-998 [FREE Full text] [doi: 10.1136/bmj.39490.551019.BE] [Medline: 18456631]

XSL•FO
RenderX

70. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
71. Garrison MM, Christakis DA. The impact of a healthy media use intervention on sleep in preschool children. Pediatrics 2012 Sep;130(3):492-499 [FREE Full text] [doi: 10.1542/peds.2011-3153] [Medline: 22869826]
72. Sun A, Cheng J, Bui Q, Liang Y, Ng T, Chen JL. Home-based and technology-centered childhood obesity prevention for Chinese mothers with preschool-aged children. J Transcult Nurs 2017 Nov 08;28(6):616-624. [doi: 10.1177/1043659617719139] [Medline: 28826348]
73. Fu Y, Burns RD, Constantino N, Zhang P. Differences in step counts, motor competence, and enjoyment between an exergaming group and a non-exergaming group. Games Health J 2018 Oct;7(5):335-340. [doi: 10.1089/g4h.2017.0188] [Medline: 29989430]
74. Trost SG, Brookes DS. Effectiveness of a novel digital application to promote fundamental movement skills in 3- to 6-year-old children: a randomized controlled trial. J Sports Sci 2021 Feb 27;39(4):453-459. [doi: 10.1080/02640414.2020.1826657] [Medline: 32985373]
75. Yarımkaya E, Esentürk OK, İlhan EL, Karasu N. A WhatsApp-delivered intervention to promote physical activity in young children with autism spectrum disorder. Int J Dev Disabil 2022 Feb 18;68(5):732-743 [FREE Full text] [doi: 10.1080/20473869.2021.1887436] [Medline: 36210901]
76. Haines J, McDonald J, O'Brien A, Sherry B, Bottino CJ, Schmidt ME, et al. Healthy habits, happy homes: randomized trial to improve household routines for obesity prevention among preschool-aged children. JAMA Pediatr 2013 Nov 01;167(11):1072-1079. [doi: 10.1001/jamapediatrics.2013.2356] [Medline: 24019074]
77. Barkin SL, Heerman WJ, Sommer EC, Martin NC, Buchowski MS, Schlundt D, et al. Effect of a behavioral intervention for underserved preschool-age children on change in body mass index: a randomized clinical trial. JAMA 2018 Aug 07;320(5):450-460 [FREE Full text] [doi: 10.1001/jama.2018.9128] [Medline: 30088008]
78. Byun W, Lau EY, Brusseau TA. Feasibility and effectiveness of a wearable technology-based physical activity intervention in preschoolers: a pilot study. Int J Environ Res Public Health 2018 Aug 23;15(9):1821 [FREE Full text] [doi: 10.3390/ijerph15091821] [Medline: 30142911]
79. Downing KL, Salmon J, Hinkley T, Hnatiuk JA, Hesketh KD. Feasibility and efficacy of a parent-focused, text message-delivered intervention to reduce sedentary behavior in 2- to 4-year-old children (mini movers): pilot randomized controlled trial. JMIR Mhealth Uhealth 2018 Feb 09;6(2):e39 [FREE Full text] [doi: 10.2196/mhealth.8573] [Medline: 29426816]
80. Ling J, Robbins LB, Zhang N, Kerver JM, Lyons H, Wieber N, et al. Using Facebook in a healthy lifestyle intervention: feasibility and preliminary efficacy. West J Nurs Res 2018 Dec 09;40(12):1818-1842. [doi: 10.1177/0193945918756870] [Medline: 29421989]
81. Yoong SL, Grady A, Stacey F, Polimeni M, Clayton O, Jones J, et al. A pilot randomized controlled trial examining the impact of a sleep intervention targeting home routines on young children's (3-6 years) physical activity. Pediatr Obes 2019 Apr 11;14(4):e12481. [doi: 10.1111/ijpo.12481] [Medline: 30417593]
82. Andersen E, Øvreås S, Jørgensen KA, Borch-Jenssen J, Moser T. Children's physical activity level and sedentary behaviour in Norwegian early childhood education and care: effects of a staff-led cluster-randomised controlled trial. BMC Public Health 2020 Nov 04;20(1):1651 [FREE Full text] [doi: 10.1186/s12889-020-09725-y] [Medline: 33148218]
83. Hoffman JA, Schmidt EM, Arguello DJ, Eyllon MN, Castaneda-Sceppa C, Cloutier G, et al. Online preschool teacher training to promote physical activity in young children: a pilot cluster randomized controlled trial. Sch Psychol 2020 Mar;35(2):118-127. [doi: 10.1037/spq0000349] [Medline: 31916788]
84. Marsh S, Taylor R, Galland B, Gerritsen S, Parag V, Maddison R. Results of the 3 Pillars Study (3PS), a relationship-based programme targeting parent-child interactions, healthy lifestyle behaviours, and the home environment in parents of preschool-aged children: a pilot randomised controlled trial. PLoS One 2020 Sep 17;15(9):e0238977 [FREE Full text] [doi: 10.1371/journal.pone.0238977] [Medline: 32941530]
85. Zeng N, Lee JE, Gao Z. Effects of home-based exergaming on preschool children's cognition, sedentary behavior, and physical activity: a randomized crossover trial. Brain Behav Immun Integr 2023 Jan;1:100002 [FREE Full text] [doi: 10.1016/j.bbii.2023.100002]
86. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
87. Hnatiuk JA, Brown HE, Downing KL, Hinkley T, Salmon J, Hesketh KD. Interventions to increase physical activity in children 0-5 years old: a systematic review, meta-analysis and realist synthesis. Obes Rev 2019 Jan;20(1):75-87 [FREE Full text] [doi: 10.1111/obr.12763] [Medline: 30257277]
88. Wu CC, Huang CW, Wang YC, Islam MM, Kung WM, Weng YC, et al. mHealth research for weight loss, physical activity, and sedentary behavior: bibliometric analysis. J Med Internet Res 2022 Jun 08;24(6):e35747 [FREE Full text] [doi: 10.2196/35747] [Medline: 35675126]

XSL•FO
RenderX

89. Zhang J, Yang SX, Wang L, Han LH, Wu XY. The influence of sedentary behaviour on mental health among children and adolescents: a systematic review and meta-analysis of longitudinal studies. J Affect Disord 2022 Jun 01;306:90-114 [FREE Full text] [doi: 10.1016/j.jad.2022.03.018] [Medline: 35304232]
90. Cliff DP, Hesketh KD, Vella SA, Hinkley T, Tsiros MD, Ridgers ND, et al. Objectively measured sedentary behaviour and health and development in children and adolescents: systematic review and meta-analysis. Obes Rev 2016 Apr;17(4):330-344. [doi: 10.1111/obr.12371] [Medline: 26914664]
91. Kelishadi R, Azizi-Soleiman F. Controlling childhood obesity: a systematic review on strategies and challenges. J Res Med Sci 2014 Oct;19(10):993-1008 [FREE Full text] [Medline: 25538786]
92. Marsh S, Foley LS, Wilks DC, Maddison R. Family-based interventions for reducing sedentary time in youth: a systematic review of randomized controlled trials. Obes Rev 2014 Feb;15(2):117-133. [doi: 10.1111/obr.12105] [Medline: 24102891]
93. Neville RD, Lakes KD, Hopkins WG, Tarantino G, Draper CE, Beck R, et al. Global changes in child and adolescent physical activity during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2022 Sep 01;176(9):886-894 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.2313] [Medline: 35816330]
94. Teesson M, Champion KE, Newton NC, Kay-Lambkin F, Chapman C, Thornton L, Health4Life Team. Study protocol of the Health4Life initiative: a cluster randomised controlled trial of an eHealth school-based program targeting multiple lifestyle risk behaviours among young Australians. BMJ Open 2020 Jul 13;10(7):e035662 [FREE Full text] [doi: 10.1136/bmjopen-2019-035662] [Medline: 32665344]
95. Chastin SF, Palarea-Albaladejo J, Dontje ML, Skelton DA. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. PLoS One 2015;10(10):e0139984 [FREE Full text] [doi: 10.1371/journal.pone.0139984] [Medline: 26461112]
96. Chong KH, Parrish AM, Cliff DP, Dumuid D, Okely AD. Changes in 24-hour movement behaviours during the transition from primary to secondary school among Australian children. Eur J Sport Sci 2022 Aug 02;22(8):1276-1286 [FREE Full text] [doi: 10.1080/17461391.2021.1903562] [Medline: 33719925]

Abbreviations

GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MINISTOP: mobile-based intervention intended to stop obesity in preschoolers
MVPA: moderate to vigorous physical activity
PA: physical activity
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
ROB: risk of bias
SB: sedentary behavior
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 19.09.23; peer-reviewed by W Liang, M Zhou, Y Zhang, EJ Buckler; comments to author 11.10.23; revised version received 04.11.23; accepted 18.01.24; published 21.02.24.

Please cite as:
Jiang S, Ng JYY, Chong KH, Peng B, Ha AS
Effects of eHealth Interventions on 24-Hour Movement Behaviors Among Preschoolers: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e52905
URL: https://www.jmir.org/2024/1/e52905
doi:10.2196/52905
PMID:38381514

©Shan Jiang, Johan Y Y Ng, Kar Hau Chong, Bo Peng, Amy S Ha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Virtual Reality–Based Exercise Rehabilitation in Cancer-Related Dysfunctions: Scoping Review

Zhenzhen Su[1], BSc; Liyan Zhang[2], MSc; Xuemin Lian[1], BSc; Miaomiao Guan[1], BSc

[1]School of Nursing, Peking University, Beijing, China

[2]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Gastrointestinal Oncology, Peking University Cancer Hospital & Institute, Beijing, China

Corresponding Author:
Liyan Zhang, MSc
Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing)
Department of Gastrointestinal Oncology
Peking University Cancer Hospital & Institute
52 Fucheng Road, Haidian District
Beijing, 100142
China
Phone: 86 18210187087
Email: zhangliyand@126.com

Abstract

Background: Virtual reality–based exercise rehabilitation (VRER) is a promising intervention for patients with cancer-related dysfunctions (CRDs). However, studies focusing on VRER for CRDs are lacking, and the results are inconsistent.

Objective: We aimed to review the application of VRER in patients with CRDs.

Methods: This scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist framework. Publications were included from the time of database establishment to October 14, 2023. The databases were PubMed, Embase, Scopus, Cochrane, Web of Science, ProQuest, arXiv, IEEE Xplore, MedRxiv, CNKI, Wanfang Data, VIP, and SinoMed. The population included patients with cancer. A virtual reality (VR) system or device was required to be provided in exercise rehabilitation as an intervention. Eligible studies focused on VRER used for CRDs. Study selection and data extraction were performed by 2 reviewers independently. Extracted data included authors, year, country, study type, groups, sample size, participant age, cancer type, existing or potential CRDs, VR models and devices, intervention programs and durations, effectiveness, compliance, satisfaction, and safety.

Results: We identified 25 articles, and among these, 12 (48%) were randomized clinical trials, 11 (44%) were other experimental studies, and 2 (8%) were observational studies. The total sample size was 1174 (range 6-136). Among the 25 studies, 22 (88%), 2 (8%), and 1 (4%) included nonimmersive VR, immersive VR, and augmented reality, respectively, which are models of VRER. Commercial game programs (17/25, 68%) were the most popular interventions of VRER, and their duration ranged from 3 to 12 weeks. Using these models and devices, VRER was mostly applied in patients with breast cancer (14/25, 56%), leukemia (8/25, 32%), and lung cancer (3/25, 12%). Furthermore, 6 CRDs were intervened by VRER, and among these, postmastectomy syndromes were the most common (10/25, 40%). Overall, 74% (17/23) of studies reported positive results, including significant improvements in limb function, joint range of motion, edema rates, cognition, respiratory disturbance index, apnea, activities of daily living, and quality of life. The compliance rate ranged from 56% to 100%. Overall, 32% (8/25) of studies reported on patient satisfaction, and of these, 88% (7/8) reported satisfaction with VRER. Moreover, 13% (1/8) reported mild sickness as an adverse event.

Conclusions: We found that around half of the studies reported using VRER in patients with breast cancer and postmastectomy dysfunctions through nonimmersive models and commercial game programs having durations of 3-12 weeks. In addition, most studies showed that VRER was effective owing to virtualization and interaction. Therefore, VRER may be an alternate intervention for patients with CRDs. However, as the conclusions were drawn from data with acknowledged inconsistencies and limited satisfaction reports, studies with larger sample sizes and more outcome indictors are required.

(J Med Internet Res 2024;26:e49312) doi:10.2196/49312

XSL•FO
RenderX

KEYWORDS

virtual reality; cancer; virtual reality–based exercise rehabilitation; cancer-related dysfunction; rehabilitation; scoping review

Introduction

Use of Traditional Exercise Rehabilitation in Patients With Cancer-Related Dysfunctions

Cancer is a leading cause of death, and cancer and its treatments cause varying degrees of cancer-related dysfunctions (CRDs) [1-3]. Studies have shown that the probability of CRDs in pediatric patients is close to 20%, whereas the incidence of CRDs in adult patients exceeds 50% [4]. Common CRDs include chemotherapy-induced peripheral neuropathy (CIPN), sexual dysfunction, cancer-related fatigue (CRF), cancer-related sleep disorder, cancer-related cognitive impairment (CRCI), postoperative syndromes in breast cancer, and cardiopulmonary dysfunction, with incidence rates of up to 96% [5], 95% [6,7], 90% [8], 90% [9], 83% [10], 60% [11], and 28% [12], respectively. These CRDs significantly reduce the quality of life (QoL) of patients and increase the economic burden on the health care system. Exercise rehabilitation, including aerobic, resistance, flexibility, and neuromuscular training, has been widely discussed as a treatment method [13]. It has been shown that exercise rehabilitation can significantly improve CRDs [14-18]. Palm et al [18] demonstrated that exercise rehabilitation, such as physical exercises and pelvic floor muscle exercises, could improve sexual function. Zimmer et al [19] reported the same results, showing a significant alleviation of CIPN after an 8-week exercise rehabilitation program that included endurance, resistance, and balance training [19].

However, traditional exercise rehabilitation has limitations. First, most patients with cancer have few opportunities to participate in exercise rehabilitation. A study on the unmet needs of patients with gynecologic cancers found that only one-third of these patients had access to an exercise rehabilitation program [20]. In addition, 86% of pediatric embryonal brain cancer survivors in Norway had unmet rehabilitation needs [21]. Second, patients with cancer have poor compliance with exercise rehabilitation. Because exercise rehabilitation was performed over a long period, was monotonous, and lacked supervision, up to 50% of patients could not continue their recovery, even after they arrived at the rehabilitation center [22]. Third, exercise rehabilitation programs must be individualized to the characteristics of each patient, which poses a challenge to medical staff [23]. Therefore, research exploring new exercise rehabilitation measures to improve CRDs and the QoL of patients with cancer is urgently needed. In addition, such research will reduce the burden on the medical system. These exercise rehabilitation measures must also ensure patient compliance and satisfaction.

Use of Virtual Reality–Based Exercise Rehabilitation in CRDs

Virtual reality (VR) is a technology that integrates visual and auditory stimuli through devices such as head-mounted displays, virtual headsets, and virtual glasses. While wearing a VR device, users can interact with the virtual environment through hand controllers and sensors [24]. Four main models of VR systems are commonly used in the medical field: desktop VR, immersive VR, augmented reality (AR), and distributed VR [25]. Desktop VR, also known as nonimmersive VR, allows users to interact with the virtual environment through devices such as a keyboard, mouse, joystick, or touch screen. Immersive VR systems temporarily isolate users from the real world. Users are immersed in virtual environments using interactive devices, such as head-mounted displays, which affect their visual, auditory, and other senses. AR enhances VR by superimposing virtual objects onto the real world to create a more realistic experience. Finally, distributed VR systems connect virtual environments from different locations across the internet, enabling users in multiple locations to participate in the same virtual space and interact effectively. The advancement of technology has enabled the gradual application of VR in patients.

Virtual reality–based exercise rehabilitation (VRER) is a promising intervention that combines VR and exercise rehabilitation. VRER uses VR to create a 3D environment coupled with body tracking to provide safe and realistic scenes [26]. Studies have suggested that VRER effectively improves dysfunctions in patients with Parkinson disease, stroke, and cardiovascular disease. In patients with Parkinson disease, VR technology can provide a virtual scenario with gait and balance rehabilitation, in which patients can engage in multisensory VRER to improve their dysfunctions [27]. In addition, this research found that VRER interventions improved the limb function and walking ability of patients with stroke, through various devices for motor or balance exercises. The improvement of both outcomes exceeded that of traditional exercise rehabilitation [28]. Similarly, patients with cardiovascular disease also benefited from VRER through a series of exergames. The treatment improved participants' symptoms and cardiorespiratory fitness [29]. Abbas et al [30] proposed that VR could achieve desired rehabilitation outcomes without a therapist, which would significantly reduce rehabilitation costs. Furthermore, VRER is beneficial when traditional rehabilitation services are inadequate or rehabilitation environments are unsafe [30].

Research Gaps and Aims

As mentioned above, traditional exercise rehabilitation is considered to be boring, lacks accessibility, has low compliance, and lacks individualization. There is an urgent need to improve the rehabilitation effectiveness in patients with CRDs. VRER, as a promising intervention, aimed to overcome the shortage of traditional exercise rehabilitation. Original studies have investigated its application and have reported the models, contents, effectiveness, and other outcomes, but the results of the available research have been inconsistent [24,31]. With the increasing number of VRER interventions, it is important to aggregate the current research via a scoping review, so that researchers can review the applications and limitations of VRER in CRDs.

XSL•FO
RenderX

We searched the Cochrane Database and PubMed for previous systematic reviews and scoping reviews, but no comprehensive reviews were found. Therefore, we conducted a scoping review on the application of VRER in CRDs, with the aim to assess the following: (1) models and contents of VRER in CRDs; (2) types of cancers and CRDs in VRER; (3) effectiveness of VRER; and (4) patient compliance, satisfaction, and safety of VRER in CRDs.

Methods

Protocol

A scoping review was conducted to explore the relevant articles on the effectiveness of VRER in CRDs. To conduct and report this scoping review, we followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist framework [32,33], but modified some parts to fit our review. The PRISMA-ScR checklist of our review can be found in Multimedia Appendix 1. The protocol details can be found in Multimedia Appendix 2 [4,14-18,20-24,27,29-31]. A scoping review is an ideal method for reporting because it provides comprehensive information about studies of interest [34]. It lists how the study is conducted, defines certain concepts and characterizations, pinpoints key factors or essential issues, analyzes insufficient information, and examines the future direction of a specific field [34]. We chose the scoping review approach as it was suitable for our aims.

Eligibility Criteria

The PICOS (Population, Intervention, Comparator, Outcome, Study design) framework was used to clarify our search eligibility criteria (Textbox 1). The population involved patients with cancer. The interventions included exercise rehabilitation through VR systems or devices. There were no restrictions on comparators applied. The eligible studies were focused on VRER used for the recovery of CRDs to assess effectiveness or feasibility, and patient satisfaction, compliance, or safety. In addition, the contents of interventions of VRER were required to be displayed. Experimental studies or observational studies with full text in English or Chinese were included, while qualitative research or other such types were excluded because we attempted to explore quantitative outcomes. Publications from the establishment of the database to October 14, 2023, were considered for inclusion.

Textbox 1. PICOS (Population, Intervention, Comparator, Outcome, Study design) framework with the inclusion and exclusion criteria, language, and time limit.

Inclusion criteria

- Population (P): Patients with cancer
- Intervention (I): Exercise rehabilitation through a virtual reality (VR) system or device
- Comparator (C): No comparator was needed. However, when studies were designed with a comparator, such as a control group, it was required to be another intervention or a blank control, and to have baseline data.
- Outcome (O): Feasibility indicators of virtual reality–based exercise rehabilitation (VRER); Effectiveness of VRER; Patient satisfaction, compliance, or safety of VRER
- Study design (S): Experimental design (randomized clinical trial [RCT], quasi-RCT, non-RCT, and before-after study); Observational design
- Language: English or Chinese
- Time limit: From the date of database establishment to October 14, 2023

Exclusion criteria

- Population (P): Dysfunction not caused by cancer or its therapy; Not cancer patients
- Intervention (I): No exercise rehabilitation; Exercise rehabilitation without VR; No description of VRER contents
- Comparator (C): No restrictions on the comparators applied
- Outcome (O): Not reported
- Study design (S): Review meeting comment, letter, and editorial protocol only
- Invention: Case report; Guidelines; Qualitative research
- Language: Not in English or Chinese
- Time limit: Time not mentioned in the inclusion criterion

Information Sources

We searched both English and Chinese databases to identify relevant studies in this review. The databases included PubMed, Embase, Scopus, Cochrane, Web of Science, ProQuest, arXiv, IEEE Xplore, MedRxiv, CNKI, Wanfang Data, VIP, and SinoMed. The databases of arXiv, IEEE Xplore, and MedRxiv were searched because they are professional computer technology databases. Moreover, we searched reference lists to explore further studies of interest.

Search Strategy

We used a combination of Medical Subject Healings (MeSH) terms and free words to build our strategy in order to achieve

XSL•FO
RenderX

perfection and optimization. Two researchers (ZS and XL) independently constructed the search terms based on the PICOS framework and previous research. If their opinions differed, the search terms were decided by another author (LZ). A specific search strategy for PubMed is listed in Textbox 2. The search terms of all databases are presented in Multimedia Appendix 3.

Textbox 2. Specific search strategy for PubMed.

- #1 “Neoplasms”[MeSH] OR “Cancer”[Title/Abstract] OR “Neoplas*”[Title/Abstract] OR “Carcinoma”[Title/Abstract] OR “Tumo*”[Title/Abstract] OR “Adenocarcinoma”[Title/Abstract] OR “Malignan*”[Title/Abstract]
- #2 “Virtual Reality”[MeSH] OR (“Virtual”[Title/Abstract] AND “Reality”[Title/Abstract])
- #3 “Virtual Reality Exposure Therapy”[MeSH] OR “Exergaming”[MeSH] OR “Exercise Therapy”[MeSH] OR “Exercise”[MeSH] OR “Sports”[MeSH] OR (“Reality Therap*”[Title/Abstract] AND “Virtual”[Title/Abstract]) OR “Active-Video Gaming*”[Title/Abstract] OR “Exergam*”[Title/Abstract] OR “Rehabilitation Exercise*”[Title/Abstract] OR “Remedial Exercise*”[Title/Abstract] OR “Exercise*”[Title/Abstract] OR “Athletic*”[Title/Abstract] OR “Training”[Title/Abstract]
- #4 #1 AND #2 AND #3

Selection of Sources of Evidence

Two researchers (ZS and MG) independently screened and cross-checked the literature. If their opinions differed, the search terms were decided by another researcher (LZ). All search results were exported to EndNote X9 (Clarivate), and then, we screened and eliminated duplicate articles manually with the software. The study selection process involved 2 steps. First, we eliminated irrelevant studies by reading the titles and abstracts. Second, we read the full text of the remaining studies to identify studies for inclusion (same 2 researchers).

Data Extraction and Result Synthesis

A data extraction form was developed by ZS and reviewed by LZ. The extracted data were all from full-text studies. The extracted contents included metadata (authors, year, country, study type, groups, sample size, and age), features of VRER application in CRDs (cancer type, existing or potential CRDs, VR models and devices, intervention programs, and duration), effectiveness, satisfaction, compliance, and safety. Two researchers (ZS and XL) independently performed data extraction and cross-checking. If their opinions differed, the search terms were decided by another researcher (LZ).

Results

Selection of Evidence

The search was conducted from the date of database establishment to October 14, 2023. A total of 2697 English and Chinese records were retrieved from 13 databases, including PubMed, Embase, Scopus, Cochrane, Web of Science, ProQuest, arXiv, IEEE Xplore, MedRxiv, CNKI, Wanfang Data, VIP, and SinoMed. All forms were included in EndNote X9, and 857 duplicate articles were manually removed. The titles and abstracts of 1840 reports were read, and then, 1770 irrelevant reports were excluded. The reasons for exclusion were research at cellular and molecular levels, inconsistent research type and patient type, and VR applications in auxiliary medical methods. After screening the titles and abstracts, 70 reports were included in the full-text screening, and 11 reports were excluded because the full text of the articles was unavailable. Moreover, 42 reports were excluded after reading the full text because the reports were reviews (n=4), protocols (n=10), qualitative studies (n=3), or invention reports (n=4); were not in English or Chinese (n=1); did not address VRER (n=14); did not involve cancer patients (n=3); and did not report results or outcomes (n=3). In addition, we simultaneously conducted a citation search of interest, generating a total of 17 records. Of these, 9 reports were excluded because the reports were reviews (n=1), did not address VRER (n=5), did not involve cancer patients (n=2), and were not in English or Chinese (n=1). Through database and citation searches, 17 and 8 articles were included, respectively. A total of 25 articles ultimately met the predefined inclusion and exclusion criteria and were cited for data extraction [35-59]. Details of the screening process and the included and excluded articles at each stage are shown in the PRISMA-ScR flowchart (Figure 1). Multimedia Appendix 4 shows the data extracted from all 25 studies [35-59], which we will also introduce in the following text.

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flowchart of the study selection process.

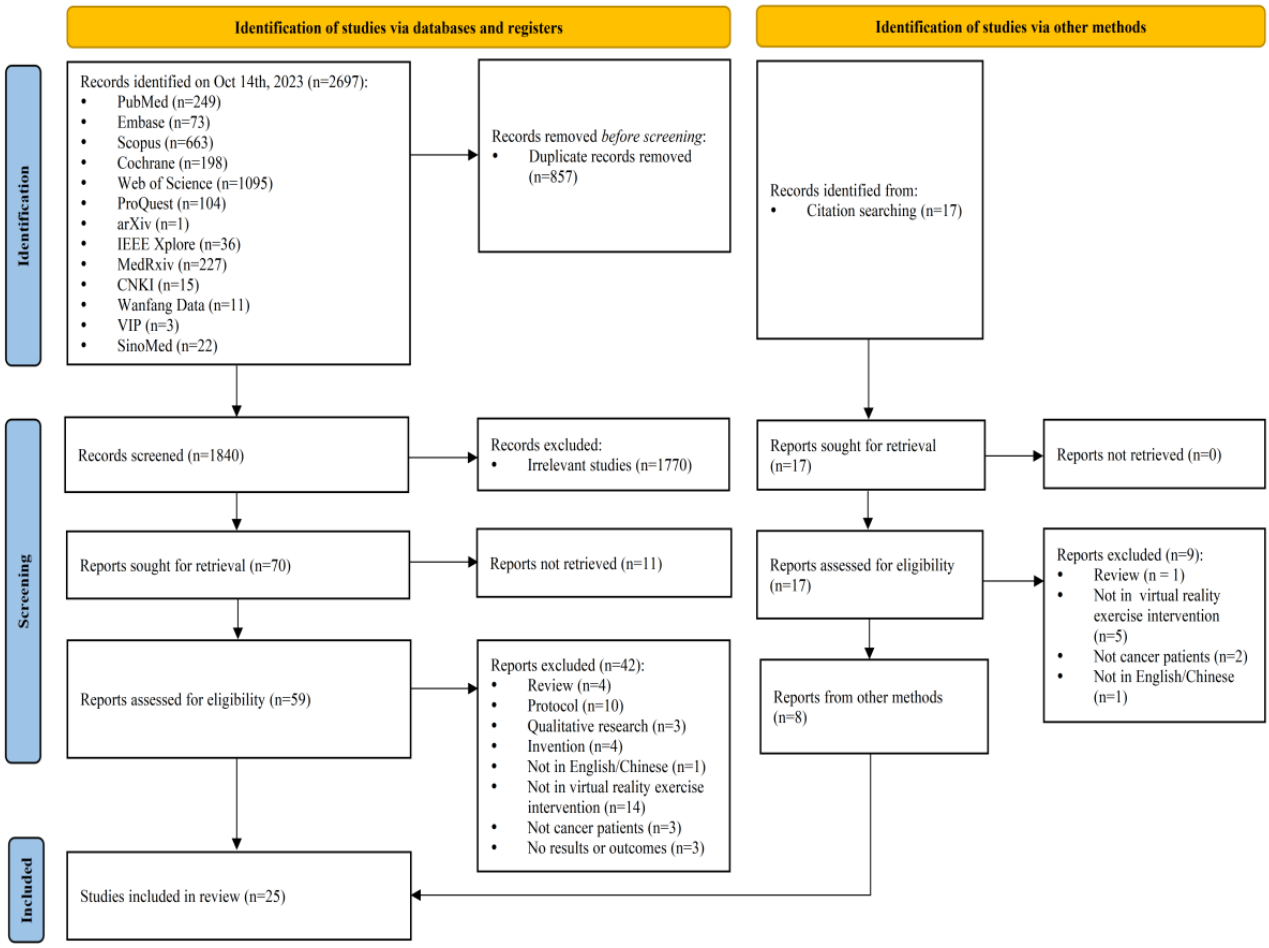

Characteristics of the Sources of Evidence

The general characteristics of the included studies are displayed in Table 1. It can be seen that primary published research on VRER dates back as far as 2013. Most studies (12/25, 48%) were randomized clinical trials (RCTs), followed by before-after studies (8/25, 32%), quasi-RCTs (2/25, 8%), observational studies (2/25, 8%), and non-RCTs (1/25, 4%). The total sample size was 1174, ranging from 6 to 136. The results showed that most (10/25, 40%) studies had fewer than 30 participants. The second largest sample size was between 31 and 50 (7/25, 28%). Less than one-third of studies (8/25, 32%) had over 50 participants.

XSL•FO
RenderX

Table 1. The time frame, study type, and sample size of the included studies.

Variable	Value (N=25), n (%)	References
Study year		
2013-2018	25 (100)	[35-59]
Study type		
RCT[a]	12 (48)	[36,37,41,42,46,48-51,54,55,59]
Quasi-RCT	2 (8)	[35,47]
Non-RCT	1 (4)	[56]
Before-after study	8 (32)	[38-40,43-45,52,53]
Observational study	2 (8)	[57,58]
Sample size		
≤30	10 (40)	[35,43-45,50-53,57,58]
31-50	7 (28)	[39-42,54-56]
51-80	5 (20)	[36-38,47,59]
≥81	3 (12)	[46,48,49]

[a]RCT: randomized clinical trial.

Models and Contents of VRER in CRDs

We explored the models and devices of VR (Table 2). The results showed that nonimmersive VR was the most popular model chosen by researchers (22/25, 88%), while immersive VR and AR were used in 8% (2/25) and 4% (1/25) of included studies, respectively. Regarding VR devices, most studies (16/25, 64%) used commercial devices manufactured by Nintendo, Microsoft, or other technology companies (eg, Nintendo Wii, Xbox Kinect, UINCARE Home+, and IREX system), while 36% (9/25) of studies used self-built VR devices.

Table 2. Models and devices of virtual reality systems in the included studies.

Variable	Value (N=25), n (%)	References
VR[a] models		
Nonimmersive VR	22 (88)	[35-48,50-56,59]
Immersive VR	2 (8)	[57,58]
Augmented reality	1 (4)	[49]
VR device		
Nintendo Wii	7 (28)	[35,42-44,50,52,53]
Xbox Kinect	6 (24)	[36,37,39-41,54]
Self-built VR	9 (36)	[38,45-48,51,57-59]
UINCARE Home+	1 (4)	[49]
IREX system	2 (8)	[55,56]

[a]VR: virtual reality.

Regarding the contents of VRER, we identified commercial game programs and individualized exercise programs. Most (17/25, 68%) of the studies adopted commercial game programs through commercial VR devices like Nintendo Wii and Xbox Kinect, while around one-third (8/25, 32%) of the studies applied individualized exercise programs, with 88% (7/8) using self-built systems (Table 3).

XSL•FO
RenderX

Table 3. Programs and durations of virtual reality–based exercise rehabilitation in cancer-related dysfunctions.

Author, year	Contents	
	Programs	Duration
Atef et al [35], 2020	Tennis, triceps extension, and rhythmic boxing	Two sessions per week for 4 weeks
Basha et al [36], 2022	"Macarena" dance and others like darts, bowling, boxing, table tennis, fruit ninja, and beach volleyball	Once per day, 5 times a week for 8 weeks
Benzing et al [37], 2020	Six different "workouts" such as "Waterfall Jump" (jumps onto wood without falling down) and "Derby Skate" (like dance, remember movements, and imitation)	Three times a week for approximately 45 minutes, for 8 weeks
Chen et al [38], 2019	Perform "Tai Chi," "Cake Cutting," "Rowing," "Virtual Shopping," and other trainings	Once per day, 5 times a week for 8 weeks
da Silva Alves et al [39], 2018	Three games: "Wall Breaker" (hit projected cubes), "Stomp It" (hit lights by moving limbs), and "Run the World" (simulate walking using knee flexion and hip turn movements)	About 45 minutes per session, 3 sessions per week for 8 weeks.
da Silva Alves et al [40], 2017	Three games: "Wall Breaker" (hit projected cubes), "Stomp It" (hit lights by moving limbs), and "Run the World" (simulate walking using knee flexion and hip turn movements)	About 45 minutes per session, 3 sessions per week for 8 weeks
Feyzioğlu et al [41], 2020	Darts, bowling, and boxing for the first 3 weeks, and beach volleyball, table tennis, and "Fruit Ninja" for the last 3 weeks.	45 minutes per session, 2 sessions a week for 6 weeks
Hamari et al [42], 2019	Various Nintendo Wii games	30 minutes a day, 7 days a week for 8 weeks
Hoffman et al [43], 2013	Walking, balance exercises, including downhill skiing, soccer, golf, and video game activities	30 minutes a day, 5 days a week for 6 weeks
Hoffman et al [44], 2014	Walking, balance exercises, including downhill skiing, soccer, golf, and video game activities	30 minutes a day, 5 days a week for 10 weeks
House et al [45], 2016	Nine custom games: "Breakout 3D" (bounce a virtual ball by paddle avatars), "Card Island," "Remember that Card" (match card), "Musical Drums" (strike notes using drum stick avatars), "Xylophone" game (repeat notes using mallet avatars), "Pick & Place" (grasp and move balls), "Arm Slalom" (rotate the shoulder to guide a skier avatar through a downhill slalom), "Avalanche" (control a pick axe and a shovel avatar to break and clear ice walls), and "Treasure Hunt" game (control shovel avatars to find buried treasure)	20 to 50 minutes per session, 2 sessions per week for 8 weeks
Jin et al [46], 2018	Breath and raise the leg, ball throw, hair comb, pendulum, shoulder shrug, arm rotation, hand raise, chest expansion, jump, and rollers	30 minutes per session, 1 session per day, 6 days per week for 3 months
Jin et al [47], 2018	Four phases: Phase 1 (1-7 days after surgery): finger strengthening, ball hold and squeeze, knead, and fingertip rub; Phase 2 (8-14 days after surgery): move like a pendulum shrug, wall climb, upper limb movement, and arm rotation; Phase 3 (15 days to 1 month after surgery): flying, rope swing, ball throw, dumbbell exercise, and elastic band; Phase 4 (1-3 months after surgery): hand shake, arm extension, waist turn, and circle exercise	15 to 30 minutes per session, 2 sessions per day for 3 months
Lin et al [48], 2021	Four phases: Phase 1 (1 week after surgery): finger stretch, ball squeeze, ball hold, and fingertip rub; Phase 2 (1-2 weeks after surgery): upper arm movement, body rotation, shoulder lift, shoulder shrug, wall climb, arm swing, and others; Phase 3 (3-4 weeks after surgery): contraction, lateral push and pull, hand swing, chest expansion, lateral lift, circumference, abdominal and back muscle training, lifts, body turns, and others; Phase 4 (5-8 weeks after surgery): head shake, arm extension, waist turn, and circle exercise	For 8 weeks
Park et al [49], 2023	Two parts: Part 1 included passive and active forward flexion, external rotation and abduction, and trunk rotation with both shoulder forward flexion; Part 2 added exercises with dumbbell and pectoralis stretching in addition to part 1.	30 minutes per session, 5 sessions per day, 7 days per week for 8 weeks
Sabel et al [50], 2016	"Sports," "Sports Resort," "Fit," "Fit plus," "Dance," and "Michael Jackson Dance"	30 minutes per day, 5 days per week for 10 weeks
Schwenk et al [51], 2016	Three tasks: Balance exercises (forward, backward, sideward, and diagonal leaning tasks); Ankle point-to-point reaching task (forward, backward, sideward, and diagonal leaning and partial weight transfer); Virtual obstacle crossing task (cross obstacles)	45 minutes per session, 2 sessions per week for 4 weeks
Tanriverdi et al [52], 2022	Jogging and hula hoop	45 minutes per session, 2 sessions per day, 2 days per week for 12 weeks

XSL•FO
RenderX

Author, year	Contents	
	Programs	Duration
Tsuda et al [53], 2016	Two activities: “Hula Hoop” and “Basic Step” at light to moderate intensity	20 minutes per session, once per day, 5 times a week until hospital discharge
Villumsen et al [54], 2019	Aerobic and strength exercises for 1 hour using “Your Shape Fitness Evolved 2012,” “Sport,” and “Adventure” games	About 1 hour per session, 3 sessions a week for 12 weeks
Yang et al [55], 2014	Five programs: “Conveyor” (move boxes from one side to the other), “Coconut” (catch coconuts in a basket), “Bird and Balls” (catch birds), “Soccer” (stop balls from entering the net), and “Juggler” (hit balls)	30 minutes a day, 3 times a week for 4 weeks
Yoon et al [56], 2015	Six programs: “Birds and Balls,” “Conveyor,” “Drums” (use hands to play a drum along with the music), “Juggler,” “Coconuts,” and “Soccer”	30 minutes per day, 3 days per week for 3 weeks
Zeng et al [57], 2020	Whack-a-mole, golf, fruit cutting, and shooting	NA[a]
Zhou et al [58], 2021	Two models: Model 1 included making fists, rotating the wrists, and bending the elbows; Model 2 included making fists, rotating the wrists, bending the elbows, lifting up, wrapping the shoulders, touching the ears, climbing walls, putting the hands behind the back, holding the head, and abducting	NA
Zhu et al [59], 2019	Four phases: Phase 1 (1 week after surgery): joint functional training like finger stretch, ball squeeze and hold, and fingertip knead; Phase 2 (2 weeks after surgery): shoulder joint abduction, adduction, extension, and forward flexion exercises; Phase 3 (3-4 weeks after surgery): resistance training with dumbbell; Phase 4 (1-3 months after surgery): normal exercise with low intensity	30 minutes per session, 2 sessions per day for 3 months

[a]NA: not available.

The intervention duration of VRER programs was different among studies (Table 3). In the included studies, the programs were conducted for 20 to 60 minutes per session, and 3 to 7 sessions were conducted per week for 3 to 12 weeks. Most were conducted for 8 weeks (9/25, 36%) or 3 months (5/25, 20%). In terms of the time per session, the majority (11/25, 44%) of programs were conducted for 30 minutes per session, and the second most common time was 45 minutes per session.

Types of Cancers and CRDs Studied in VRER

Table 4 presents the overall general characteristics of the included populations, and types of cancers and CRDs studied in VRER. The population’s age ranged from 3 years to over 70 years. Among all studies, 16% (4/25) reported pediatric cancer patients aged <18 years, who were mainly diagnosed with leukemia, lymphoma, or brain cancer [37,42,50,52]. The remaining 84% (21/25) of studies reported adults with various cancers, including breast cancer, leukemia, lymphoma, brain cancer, nervous system cancer, lung cancer, gastrointestinal tract cancer, abdominal and pelvic cancer, oropharyngeal cancer, multiple myeloma, colorectal cancer, melanoma, bladder cancer, prostate cancer, pancreatic cancer, and ovarian cancer. Among all 25 studies included in the analysis, patients with breast cancer constituted the largest group (14/25, 56%). The second largest cancer type was leukemia (8/25, 32%), and the third largest was lung cancer (3/25, 12%). With these cancers, the most reported CRDs were postmastectomy syndromes, including upper limb dyskinesia and lymphedema (10/25, 40%). The second most reported CRDs were CRCI and CRF (both 5/25, 20%). The third most reported CRDs were extensive physical function damage and neuropathy (3/25, 12%; 1 study about CIPN and 2 studies about cancer-related central neuropathy). Cancer-related sleep disorders (1/25, 4%) were also covered in the included population.

XSL•FO
RenderX

Table 4. Population characteristics, types of cancers, and cancer-related dysfunctions studied in virtual reality–based exercise rehabilitation.

Author, year	Age (years), mean (SD) or mean (range)		Cancer type	Types of existing or potential CRDs[a]
	Intervention group	Control group		
Atef et al [35], 2020	54.1 (8.3)	53.1 (7.2)	Breast cancer	Postmastectomy lymphedema
Basha et al [36], 2022	48.8 (7.0)	52.1 (7.5)	Breast cancer	Postmastectomy lymphedema
Benzing et al [37], 2020	11.8 (2.4) and 10.7 (2.5)	11.1 (2.5)	Various types (leukemia, lymphoma, and nervous system cancer)	Central neuropathy
Chen et al [38], 2019	NR[b]	NR	Breast cancer	CRCI[c]
da Silva Alves et al [39], 2018	57.1 (16.7) and 63.3 (7.3)	56.7 (11.9)	Various types (gastrointestinal tract, breast, abdominal, pelvic, and oropharyngeal cancer)	CRF[d]
da Silva Alves et al [40], 2017	57.1 (16.7) and 63.3 (7.3)	56.7 (11.9)	Various types (gastrointestinal tract, breast, abdominal, pelvic, and oropharyngeal cancer)	CRF
Feyzioğlu et al [41], 2020	50.8 (8.5)	51.0 (7.1)	Breast cancer	Postmastectomy upper limb dyskinesia
Hamari et al [42], 2019	7.8 (range, 3-16)	7.9 (range, 3-15)	Various types (Wilms cancer, acute lymphocytic leukemia, and lymphoma)	Physical function damage and CRF
Hoffman et al [43], 2013	64.6 (range, 53-73)	NR	Lung cancer	CRF
Hoffman et al [44], 2014	64.6 (range, 53-73)	NR	Lung cancer	CRF
House et al [45], 2016	57.8 (20.4)	NR	Breast cancer	CRCI and postmastectomy upper limb dyskinesia
Jin et al [46], 2018	NR	NR	Breast cancer	Postmastectomy upper limb dyskinesia
Jin et al [47], 2018	Range 33-36	Range 35-69	Breast cancer	Postmastectomy upper limb dyskinesia
Lin et al [48], 2021	33.0 (2.8)	33.1 (2.8)	Breast cancer	CRCI
Park et al [49], 2023	42.6 (9.1)	47.3 (8.6)	Breast cancer	Postmastectomy syndrome
Sabel et al [50], 2016	11.9 (3.6)	13.2 (1.9)	Brain cancer	Central neuropathy
Schwenk et al [51], 2016	68.7 (8.7)	71.8 (8.9)	Various types (multiple myeloma, chronic lymphoid leukemia, lung, breast, colorectal, melanoma, bladder, prostate, pancreas, and ovarian cancer)	Chemotherapy-induced peripheral neuropathy
Tanriverdi et al [52], 2022	12.9 (5.8)	NR	Acute lymphoblastic leukemia	Cancer-related sleep disorder
Tsuda et al [53], 2016	66.0 (5.4)	NR	Lymphoma and acute leukemia	Physical function damage
Villumsen et al [54], 2019	67.6 (4.6)	69.8 (4.4)	Prostate cancer	Physical function damage
Yang et al [55], 2014	47.9 (14.5)	52.9 (14.0)	Brain cancer	CRCI
Yoon et al [56], 2015	48.6 (11.3)	50.0 (17.5)	Brain cancer	Postmastectomy upper limb dyskinesia
Zeng et al [57], 2020	NR	NR	NR	CRCI
Zhou et al [58], 2021	54.7 (7.8)	NR	Breast cancer	Postmastectomy upper limb dyskinesia
Zhu et al [59], 2019	58.3 (15.4)	58.6 (15.1)	Breast cancer	Postmastectomy upper limb dyskinesia

[a]CRDs: cancer-related dysfunctions.

[b]NR: not reported.

[c]CRCI: cancer-related cognitive impairment.

[d]CRF: cancer-related fatigue.

XSL•FO
RenderX

Effectiveness of VRER for CRDs

The effectiveness of VRER for CRDs is shown in Table 5. Two observational studies were excluded in this table because they only observed the current situation without any intervention. The effectiveness included positive (results with significant improvements) and negative results (results with no significant improvements, results without statistical analysis, or inconsistent results). Overall, 74% (17/23) of studies reported positive results. Among the 23 studies, 70% (16/23) reported the results compared with preintervention groups and 65% (15/23) reported the results compared with control groups. Among the studies that reported the results compared with preintervention groups, 75% (12/16) reported positive results after the intervention, including significant improvements in limb function, excessive limb volume, activities of daily living (ADLs), QoL, muscle strength, fear of movement, CRF, balance, body coordination, respiratory disturbance index, apnea, sleep habits, and memory. Conversely, 25% (4/16) of studies reported negative results. Among the studies that reported the results compared with control groups, 60% (9/15) reported positive results after the intervention, including significant improvements in pain intensity, range of motion (ROM) of the joint, fear of movement, edema rate, cognition, ADLs, and QoL.

XSL•FO
RenderX

Table 5. Effectiveness of exercise rehabilitation with virtual reality devices (N=23).

Author, year	Results (preintervention or control group vs VRER[a] group)	
	Compared with preintervention	Compared with control
Atef et al [35], 2020	There were significant improvements in excessive arm volume and affected upper limb dysfunction (9655.27 vs 6854.23 and 55 vs 38, respectively; both *P*<.05). The percentage improvement in excessive arm volume was 26.47% and that of upper limb dysfunction was 33.66%.	No significant differences (*P*>.05).
Basha et al [36], 2022	There were significant improvements in excessive limb volume by 2.63%; disability of the arm, shoulder, and hand by 14.7; ROM[b] by 25.83°; strength by 1.28 kg; general health by 6.07; mental health by 4.53; vitality by 5.03; physical functioning by 14.7; social functioning by 2.77; and pain by 31.0 (all *P*<.05).	There were significant improvements in pain intensity (53.76 vs 40.33), disability of the arm (27.1 vs 22.9), shoulder ROM (including flexion, abduction, and external rotation, with scores of 126.0 vs 139.0, 102.1 vs 123.3, and 66.1 vs 75.0, respectively), general health (54.9 vs 58.4), and vitality (53.5 vs 55.7) (all *P*<.05).
Benzing et al [37], 2020	NR[c]	No significant differences (*P*>.05).
Chen et al [38], 2019	There were significant improvements in cognitive function (26.22 vs 17.41) and daily life ability (16.05 vs 20.79) (all *P*<.05).	NA[d]
da Silva Alves et al [39], 2018	There were significant improvements in the general functional assessment (88.13 vs 95.37), functional outcomes index (76.00 vs 92.68), and fatigue assessment (121.20 vs 138.99) (all *P*<.05).	NA
da Silva Alves et al [40], 2017	There were significant improvements in maximal voluntary isometric contraction of the right dorsiflexor muscles (6.62 vs 10.9), right plantar flexion (15.01 vs 22.59), and left medial gastrocnemius (158.61 vs 185.23) (all *P*<.05).	NA
Feyzioğlu et al [41], 2020	There were significant improvements in pain (6.53 vs 1.53), ROM (by 55.39°), muscle strength (5.97 vs 6.25), grip strength (21.47 vs 23.71), functionality (44.67 vs 16.49), and fear scores (42.37 vs 29.47) (all *P*<.05).	Fear of movement was significantly improved by VRER, while the traditional exercise rehabilitation displayed more improvement in functionality (5.94 vs 12.9 and −36.54 vs −28.18, respectively) (both *P*<.05).
Hamari et al [42], 2019	No significant differences (*P*>.05).	No significant differences (*P*>.05).
Hoffman et al [43], 2013	There were improvements in CRF[e] severity (4.8 vs 2.5), perceived self-efficacy for fatigue self-management (7 vs 8.8), walking and balance (47.4% vs 93.3% and 72.8% vs 83.7%, respectively), and functional performance (steps taken per day, 4650 vs 6393). No statistical analysis.	NA
Hoffman et al [44], 2014	There were improvements in CRF severity (4.8 vs 1.3), perceived self-efficacy for fatigue self-management (7 vs 9), walking and balance (47.4% vs 99.4% and 72.8% vs 88.9%, respectively), and functional performance (steps taken per day, 4650 vs 7683). No statistical analysis.	NA
House et al [45], 2016	There were significant improvements in ROM by 8°, strength by 8.2 N, daily life ability by 13.8, depression by −5.7, and visuospatial memory by 8 (all *P*<.05).	NA
Jin et al [46], 2018	NR	There was a significant improvement in QoL[f] (75.1 vs 85.4; *P*<.05).
Jin et al [47], 2018	NR	There were significant improvements in shoulder ROM (57.8° vs 81.6°), compliance (84% vs 100%), and edema rate of the affected limbs (42% vs 11%) (all *P*<.05).
Lin et al [48], 2021	NR	There were significant improvements in postoperative complication rates (67% vs 11%), compliance (71% vs 95%), shoulder ROM (57.9° vs 81.6°), daily life ability (20.84 vs 16.07), QoL (15.39 vs 18.69), and cognition (24.60 vs 29.46) (all *P*<.05).
Park et al [49], 2023	There were significant improvements in ROM (109.56 vs 169.12), limb dysfunction (22.05 vs 17.07), and QoL (0.69 vs 0.85) (all *P*<.05).	No significant differences (*P*>.05).
Sabel et al [50], 2016	The body coordination score improved by 15% (*P*<.05).	No significant differences (*P*>.05).

XSL•FO
RenderX

Author, year	Results (preintervention or control group vs VRER[a] group)	
	Compared with preintervention	Compared with control
Schwenk et al [51], 2016	NR	There were significant improvements in the medial-lateral center of mass sway (21.2% vs 55.5%), hip sway (42.3% vs 67.5%), and ankle sway (21.4% vs 68.2%) (all *P*<.05).
Tanriverdi et al [52], 2022	There were significant improvements in the Respiratory Disturbance Index (3.10 vs 1.97), number of apnea episodes (15.55 vs 10.14), and sleep habits (57.45 vs 48.27) (all *P*<.05).	NA
Tsuda et al [53], 2016	No significant differences (*P*>.05).	NA
Villumsen et al [54], 2019	NR	There was a significant improvement in the 6-min walking test by 4.2% (*P*<.05).
Yang et al [55], 2014	There were significant improvements in the results of the continuous performance test (0.65 vs 0.55), digit span test (3.9 vs 5.1), visual span test (3.5 vs 4.6), learning test-recognition (30.5 vs 40.7), Trail Making Test-type A (137.9 vs 70.2), Korean version of the Modified Barthel Index (43.4 vs 73.7), and Korean version of the Mini-Mental Status Examination (19.8 vs 25.0) (all *P*<.05).	There were significant improvements in the results of the continuous performance test (−0.2 vs 0.0), backward digit span test (1.4 vs 0.3), backward visual span test (1.4 vs 0.7), and Trail Making Test-type A (−67.7 vs −24.5) (all *P*<.05).
Yoon et al [56], 2015	There were significant improvements in manual dexterity (30.5 vs 38.0), manual function test results (70.3 vs 82.8), Fugl-Meyer scale scores of upper extremity motor ability (52.0 vs 58.0), and activity of life (52.5 vs 70.5) (all *P*<.05).	There were significant improvements in manual dexterity (8.0 vs 11.0), manual function test results (shoulder/elbow/forearm section, 5.0 vs 7.0), and Fugl-Meyer scale scores (shoulder/elbow/forearm section, 2.0 vs 3.5) (all *P*<.05).
Zhu et al [59], 2019	NR	There were significant improvements in compliance (72.5 vs 90), shoulder ROM (57.84° vs 81.65°), and postoperative complication rates (12.5% vs 32.5%) (all *P*<.05).

[a]VRER: virtual reality–based exercise rehabilitation.

[b]ROM: range of motion.

[c]NR: not reported.

[d]NA: not available.

[e]CRF: cancer-related fatigue.

[f]QoL: quality of life.

Compliance, Satisfaction, and Safety of VRER

Patient-reported compliance, satisfaction, and safety of VRER are presented in Table 6. Some studies did not report some indicators. Compliance and satisfaction rates or scores were either explicitly stated by the authors or calculated from flow charts. If the satisfaction rate was above 85% or the score was over 3.4/4, 4.3/5, or 5.1/6, we defined the patient-reported result as satisfactory. If the satisfaction rate was lower than 60% or the score was less than 2.4/4, 3.0/5, or 3.6/6, we defined the patient-reported result as unsatisfactory.

XSL•FO
RenderX

Table 6. Patient compliance, satisfaction, and safety of exercise rehabilitation with virtual reality devices.

Variable	Value (N=25), n (%)
Compliance (%)	22 (88)
100	6 (27)
85-99	8 (36)
70-84	6 (27)
60-69	1 (5)
<60	1 (5)
Satisfaction	8 (32)
Satisfactory	7 (88)
Neutral	1 (13)
Unsatisfactory	0 (0)
Safety (adverse events)	8 (32)
None	7 (88)
Sickness	1 (13)

As shown in Table 6, 88% (22/25) of studies reported compliance rates, ranging from 56% to 100%. Among them, the majority (13/22, 55%) reported compliance rates between 85% and 99%. Unfortunately, 5% (1/22) reported a compliance rate lower than 60%. With regard to satisfaction, 32% (8/25) of studies provided results. Among them, 88% (7/8) reported satisfaction with VRER and 13% (1/8) reported a neutral result. None of the studies reported unsatisfactory results. With regard to adverse events, 8 studies provided results. Among them, 88% (7/8) reported no adverse events and 12% (1/8) reported mild sickness.

Discussion

Principal Findings

VRER is a promising intervention for rehabilitation in patients with CRDs. However, studies focusing on this intervention are lacking and have inconsistent results. In this scoping review, we aimed to explore the applications of VRER used for CRDs, in order to provide a comprehensive summary of studies and to provide a reference for clinical practice.

Of 2714 studies, 25 records were identified that fit our eligibility criteria through screening and analysis. Most of the studies were RCTs (12/25, 48%) and before-after studies (8/25, 32%). The total sample size was 1174, ranging from 6 to 136. According to our objectives, there were 3 categories of results, and each of them provided classification of studies from a different aspect.

First, there were several models and contents of VRER. The models of VR included nonimmersive VR, immersive VR, and AR. Nonimmersive VR was the most (22/25, 88%) used branch through both commercial (16/25, 74%) and self-built devices (9/25, 36%). Within these 3 models, researchers preferred choosing commercial games (17/25, 68%). The other 32% (8/25) of studies applied individualized exercise programs, which were mainly conducted by self-built systems (7/8, 88%). The duration of VRER ranged from 3 to 12 weeks.

Second, there were various types of cancers and CRDs intervened by different VRER approaches. In our review, the most common cancers were breast cancer (14/25, 56%), leukemia (8/25, 32%), and lung cancer (3/25, 12%). Postmastectomy syndromes, including upper limb dyskinesia and lymphedema, were the main CRDs (10/25, 40%).

Third, the effectiveness, compliance, satisfaction, and safety of VRER were inconsistent in the publications. Positive results included significant improvements in limb function, joint ROM, edema rates, cognition, respiratory disturbance index, apnea, ADLs, and QoL. The compliance rate ranged from 56% to 100%. None of the studies reported unsatisfactory results, while 1 study reported mild sickness.

Figure 2 provides an illustration of the distribution and efficacy of VRER applications across different types of cancers.

XSL•FO
RenderX

Figure 2. Distribution and efficacy of virtual reality–based exercise rehabilitation applications across different types of cancers. Inconsistent indicates that some studies showed positive results while others showed negative results for the same cancer. AR: augmented reality; VR: virtual reality.

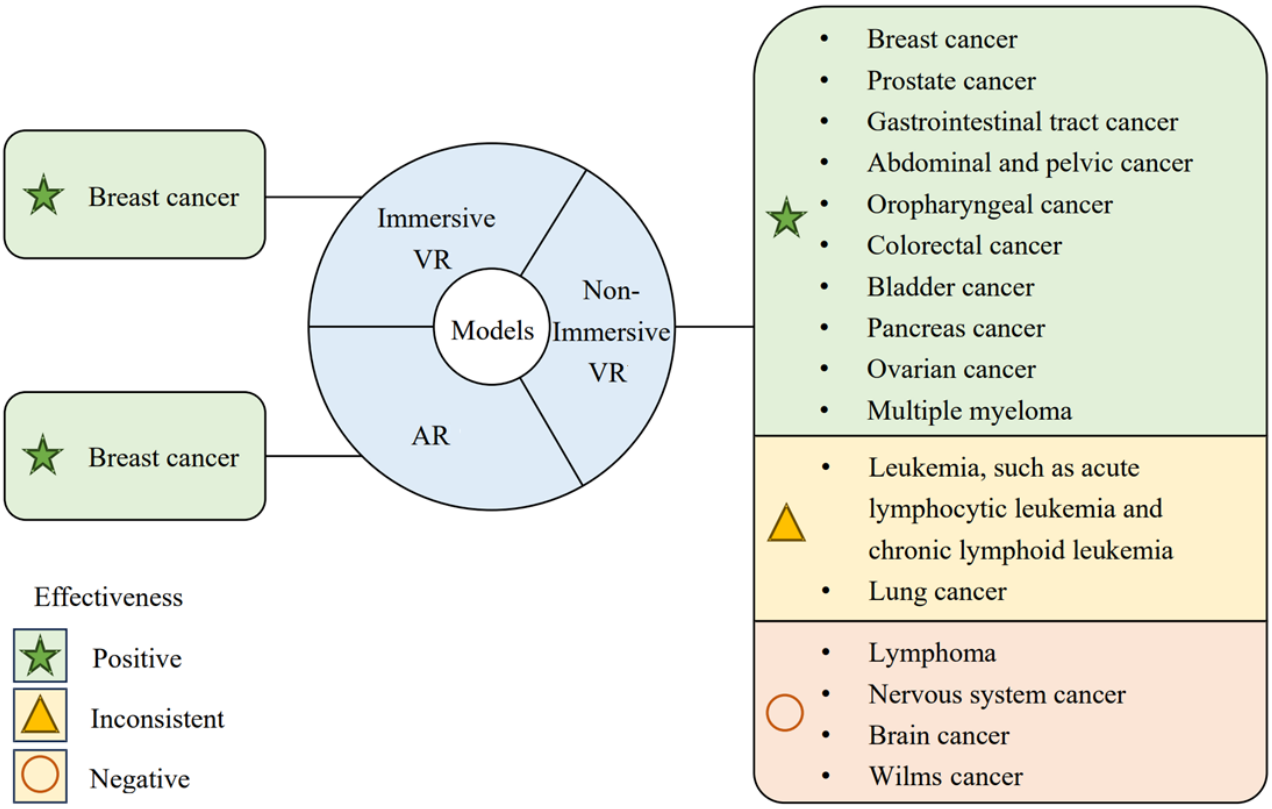

Comparison With Prior Work

Models and Contents of VRER

In our review, 88% (22/25) of studies applied nonimmersive VR systems for rehabilitation, while the remaining 12% (3/25) applied immersive VR and AR. Among them, the contents were mainly commercial game programs, which were reported in 68% (17/25) of studies. Commercial devices like Nintendo Wii and Xbox Kinect were mostly (13/25, 52%) used. These findings are consistent with the results of other systematic reviews [60]. The reasons for the dominance of nonimmersive VR and commercial game programs include low cost, easy use, and high accessibility [61]. In addition, commercial devices have mature game systems that involve various fun rehabilitation approaches [62].

We also discovered different views from other studies. In our review, nonimmersive VR more frequently led to less adverse events than immersive VR, and thus, people preferred nonimmersive VR [58]. This finding is not consistent with the results in the studies by Tuena et al [63] and Buche et al [64], who found that immersive VR appeared to be associated with less motion sickness. The difference in the findings may be related to the intervention duration. In our review, studies reported that the duration ranged from 3 to 12 weeks, whereas over half of the studies in the review by Tuena et al [63] conducted 1 session of less than 45 minutes. Studies with a larger duration and with controls can be implemented in the future with nonimmersive and immersive VR to obtain definite conclusions.

We noted that over one-third (9/25, 36%) of studies developed self-built VRER systems. Most of them considered the differences in the disease durations and exercise preferences of patients. These studies aimed to improve the outcomes through individual design. Unfortunately, the effectiveness of these systems was compared with that of traditional exercise rehabilitation but not with that of commercial VRER. Further research is needed in this field.

Types of Cancers and CRDs in Studies on VRER

We found that around half of the studies on VRER involved patients with breast cancer and postmastectomy dysfunction. This finding is similar to the results in other studies [60,65]. The reasons may be that breast cancer is the most common cancer worldwide and postoperative dysfunction continues from treatment to the survival stage. Thus, there is a need for symptom management, and this area is worthy of attention [66].

Moreover, we identified VRER for several other types of cancers and CRDs. Some studies (13/25, 52%) reported the use of VRER in patients with leukemia, lymphoma, brain cancer, nervous system cancer, lung cancer, gastrointestinal cancer, abdominal and pelvic cancer, oropharyngeal cancer, multiple myeloma, colorectal cancer, melanoma, bladder cancer, prostate cancer, pancreas cancer, and ovarian cancer. Furthermore, research (15/25, 60%) on CRDs, including CRCI, CRF, extensive physical dysfunction, CIPN, cancer-related central neuropathy, and sleep disorders, reported that intervention by VRER was effective. Considering the diversity of cancers and CRDs, we adopted a broader scope to explore the application of VRER in CRDs. However, the number of publications was limited. Future studies should consider a greater number of cancers and CRDs.

XSL•FO
RenderX

Effectiveness of VRER

Another significant finding of our study was that VRER can effectively improve CRDs in patients with cancer. Buche et al [64] reviewed VRER application in CRDs and reported a similar result.

In addition to postmastectomy syndromes, other types of CRDs could also be intervened effectively by VRER. Studies showed that VRER can improve limb function, joint ROM, edema rates, cognition, respiratory disturbance index, apnea, ADLs, and QoL in cases of CRDs like CRCI, CIPN, and cancer-related sleep disorder [38-40,43-45,48,50-52,54,55]. The potential mechanism was speculated to be closely related to the virtualization and interaction of VR. On one hand, the virtual environment created by VRER could alleviate the fear of movement and the pain of exercise among patients, enabling them to execute exercise rehabilitation better through distraction [36,67]. Moreover, the virtual environment provided stimuli and pleasure, encouraging participants to be more active [68-70]. On the other hand, VR can directly affect exercise rehabilitation through interaction. The interaction associated with VR generates visual stimuli through which patients can identify differences between their movements and the correct ones [71]. Thus, patients' objective indicators and stimulus feedback allow VR games to be continuously adjusted to achieve superior results [41]. Therefore, it can be more personalized and accurate [38]. In addition, interaction encourages competition and repetition, improving participants' focus and executive function [72,73]. The potential mechanisms of VRER in CRDs are depicted in Figure 3.

Figure 3. The potential mechanisms of virtual reality–based exercise rehabilitation (VRER) in cancer-related dysfunctions (CRDs).

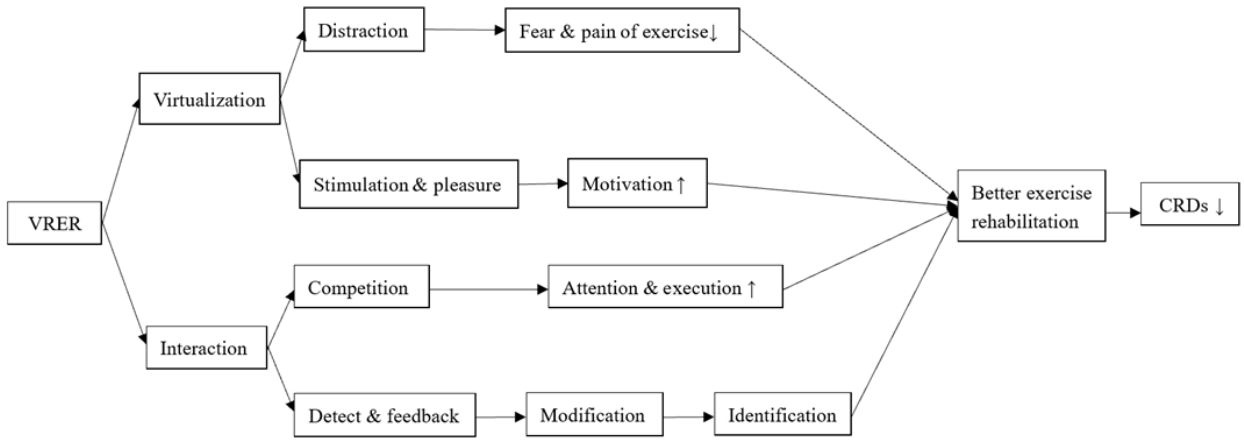

Undeniably, there is no certainty that VRER is better than traditional exercise rehabilitation. Many (10/25, 40%) studies had a small sample size or lacked a randomized and controlled design, which made the results of the studies somewhat different, and the studies lacked sufficient convincing power. We look forward to adding large-sample RCTs in the future.

Compliance, Satisfaction, and Safety of VRER

In addition to clinical effectiveness, we found that it is highly feasible for medical staff and patients to use VRER for CRDs. The median compliance rate was 91% (IQ1-IQ3=82%-99%) in the 25 studies, which is much higher than the rate of 71% or less for traditional exercise rehabilitation [74]. Furthermore, high satisfaction and mild VRER adverse events indicated that VRER was widespread and readily accepted. These results were similar to the findings of a systematic review of VR for symptom management in cancer patients, which reported high retention rates in most VR interventions [75]. However, only 8 studies (8/25, 32%) reported on satisfaction, with 88% (7/8) reporting satisfactory results on VRER (satisfaction rate of over 85% or scores of 3.4/4, 4.3/5, or 5.1/6). When we explored the safety of VRER devices, we found that only 1 study (1/8, 13%) reported mild sickness. Although adverse events have been reported less frequently and the reported adverse event was mildly symptomatic, such events may have some impact on patients. As Zhou et al [58] reported, cybersickness is unavoidable in the use of VR owing to the desynchronization of visual stimuli with neural signals. Therefore, when VRER is used, we suggest to introduce it slowly and carefully according to the feedback received from patients. In addition, when reporting outcomes, future studies on VRER for CRDs need to include compliance, satisfaction, and occurrence of adverse events as indicators to better understand the feasibility and safety of VRER.

Limitations and Future Implications

Our study has some limitations. First, we found that the sample sizes of the studies varied widely, ranging from 6 to 136, which may have introduced variability in our findings. Additionally, more than one-third (10/25, 40%) of the studies had a sample size of less than 30, making it difficult to obtain convincing results. Second, not all studies reported all the indicators we required, which may have reduced the accuracy of our findings, and the validity of the results should be interpreted with caution. In view of the variety of CRDs, VRER application is still relatively limited. Third, there was a trend of using self-built VR systems, but comparisons with other VR systems were lacking. For these reasons, it is difficult to determine the effectiveness of VRER for CRDs, and its adoption should be treated with caution.

Based on these limitations, we have made 3 suggestions for future research. When assessing VRER for CRDs, researchers

XSL•FO
RenderX

should (1) increase the sample size to enable more people to use VRER in order to obtain valuable and credible outcomes; (2) include more types of CRDs and more comprehensive indicators like compliance, satisfaction, and safety; and (3) develop more self-built VRER systems with individual content and perform comparisons with traditional VRER systems.

Conclusion

To the best of our knowledge, this is the first scoping review to provide the most comprehensive data for VRER in patients with CRDs. We summarized the types, models, contents, effectiveness, compliance, satisfaction, and safety of VRER for CRDs, as well as mapped the potential mechanism. Our study found that VRER was mainly used in patients with breast cancer and postmastectomy dysfunctions through nonimmersive models and commercial game programs. Moreover, we found that VRER is an effective intervention accompanied with high compliance and satisfaction for CRDs. However, our findings regarding the effectiveness of VRER are drawn from data with acknowledged inconsistencies and limited satisfaction reports. Thus, it is critical to consider these conclusions with caution. In addition, the sample size, types of CRDs, reported indicators, and VR systems were limited. Nevertheless, we believe that this review can help clinical practices to better understand the applications of VRER for CRDs and to better determine whether to use this approach. We believe that VRER has further unexploited potential in rehabilitation and health care for CRDs, but additional research is needed to solidify these findings. For VRER to be properly accepted in the real word, studies involving larger sample sizes, more CRDs with individual content, and more outcome indictors are required.

Acknowledgments

We thank TopEdit for linguistic assistance during the preparation of this manuscript.

Data Availability

The data sets generated during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ZS and LZ contributed to the conception and design of this paper. ZS conducted the literature searches and drafted the first version of the manuscript. ZS, LZ, and XL contributed to the construction of search terms and the screening of the literature. ZS and XL contributed to the data extraction. LZ, MG, and XL revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 116 KB - jmir_v26i1e49312_app1.pdf]

Multimedia Appendix 2
Protocol.
[PDF File (Adobe PDF File), 296 KB - jmir_v26i1e49312_app2.pdf]

Multimedia Appendix 3
Search terms used to identify studies.
[PDF File (Adobe PDF File), 114 KB - jmir_v26i1e49312_app3.pdf]

Multimedia Appendix 4
Data extracted from all 25 studies.
[PDF File (Adobe PDF File), 252 KB - jmir_v26i1e49312_app4.pdf]

References

1. Kline RM, Arora NK, Bradley CJ, Brauer ER, Graves DL, Lunsford NB, et al. Long-Term Survivorship Care After Cancer Treatment - Summary of a 2017 National Cancer Policy Forum Workshop. J Natl Cancer Inst 2018 Dec 01;110(12):1300-1310 [FREE Full text] [doi: 10.1093/jnci/djy176] [Medline: 30496448]
2. Granström B, Tiblom Ehrsson Y, Holmberg E, Hammerlid E, Beran M, Tano K, Swedish HeadNeck Cancer Register (SweHNCR). Return to work after oropharyngeal cancer treatment-Highlighting a growing working-age population. Head Neck 2020 Aug 24;42(8):1893-1901. [doi: 10.1002/hed.26123] [Medline: 32092226]

XSL•FO
RenderX

3. Lee Y, Goo-Yoshino S, Lew HL, Chi W, Yen C, Liao H, et al. Social participation in head and neck cancer survivors with swallowing disorder: World Health Organization Disability Assessment Schedule 2.0 study. Head Neck 2020 May 30;42(5):905-912. [doi: 10.1002/hed.26062] [Medline: 31886608]
4. Stubblefield MD, Schmitz KH, Ness KK. Physical functioning and rehabilitation for the cancer survivor. Semin Oncol 2013 Dec;40(6):784-795. [doi: 10.1053/j.seminoncol.2013.09.008] [Medline: 24331197]
5. Shah A, Hoffman EM, Mauermann ML, Loprinzi CL, Windebank AJ, Klein CJ, et al. Incidence and disease burden of chemotherapy-induced peripheral neuropathy in a population-based cohort. J Neurol Neurosurg Psychiatry 2018 Jun 08;89(6):636-641 [FREE Full text] [doi: 10.1136/jnnp-2017-317215] [Medline: 29439162]
6. Franzoi MA, Agostinetto E, Perachino M, Del Mastro L, de Azambuja E, Vaz-Luis I, et al. Evidence-based approaches for the management of side-effects of adjuvant endocrine therapy in patients with breast cancer. Lancet Oncol 2021 Jul;22(7):e303-e313. [doi: 10.1016/S1470-2045(20)30666-5] [Medline: 33891888]
7. Neal D, Metcalfe C, Donovan J, Lane J, Davis M, Young G, ProtecT Study Group. Ten-year Mortality, Disease Progression, and Treatment-related Side Effects in Men with Localised Prostate Cancer from the ProtecT Randomised Controlled Trial According to Treatment Received. Eur Urol 2020 Mar;77(3):320-330 [FREE Full text] [doi: 10.1016/j.eururo.2019.10.030] [Medline: 31771797]
8. Fabi A, Bhargava R, Fatigoni S, Guglielmo M, Horneber M, Roila F, ESMO Guidelines Committee. Cancer-related fatigue: ESMO Clinical Practice Guidelines for diagnosis and treatment. Ann Oncol 2020 Jun;31(6):713-723 [FREE Full text] [doi: 10.1016/j.annonc.2020.02.016] [Medline: 32173483]
9. Chen D, Yin Z, Fang B. Measurements and status of sleep quality in patients with cancers. Support Care Cancer 2018 Feb 23;26(2):405-414. [doi: 10.1007/s00520-017-3927-x] [Medline: 29058128]
10. Wu X, Li J, Wang F, Zhao Q, Wu Q. Cognitive function and its influencing factors in patients with first-episode brain tumor. Chin Nurs Res 2019;33(04):564-570. [doi: 10.12102/j.issn.1009-6493.2019.04.005]
11. Rockson SG. Lymphedema after Breast Cancer Treatment. N Engl J Med 2018 Nov 15;379(20):1937-1944. [doi: 10.1056/NEJMcp1803290] [Medline: 30428297]
12. Yu AF, Flynn JR, Moskowitz CS, Scott JM, Oeffinger KC, Dang CT, et al. Long-term Cardiopulmonary Consequences of Treatment-Induced Cardiotoxicity in Survivors of ERBB2-Positive Breast Cancer. JAMA Cardiol 2020 Mar 01;5(3):309-317 [FREE Full text] [doi: 10.1001/jamacardio.2019.5586] [Medline: 31939997]
13. Chinese Association of Rehabilitation Medicine Tumor Rehabilitation Professional Committee, Jiangsu Integrative Medical Research Institute. The consensus of Chinese cancer patients' exercise rehabilitation experts centered on disability. Chin J Rehabil Med 2023;38(01):1-7. [doi: 10.3969/j.issn.1001-1242.2023.01.001]
14. Brassel S, Power E, Campbell A, Brunner M, Togher L. Recommendations for the Design and Implementation of Virtual Reality for Acquired Brain Injury Rehabilitation: Systematic Review. J Med Internet Res 2021 Jul 30;23(7):e26344 [FREE Full text] [doi: 10.2196/26344] [Medline: 34328434]
15. Campbell KL, Winters-Stone KM, Wiskemann J, May AM, Schwartz AL, Courneya KS, et al. Exercise Guidelines for Cancer Survivors: Consensus Statement from International Multidisciplinary Roundtable. Med Sci Sports Exerc 2019 Nov;51(11):2375-2390 [FREE Full text] [doi: 10.1249/MSS.0000000000002116] [Medline: 31626055]
16. Hayes SC, Singh B, Reul-Hirche H, Bloomquist K, Johansson K, Jönsson C, et al. The Effect of Exercise for the Prevention and Treatment of Cancer-Related Lymphedema: A Systematic Review with Meta-analysis. Med Sci Sports Exerc 2022 Aug 01;54(8):1389-1399. [doi: 10.1249/MSS.0000000000002918] [Medline: 35320145]
17. Ouyoung LM, Swanson MS, Villegas BC, Damodar D, Kokot N, Sinha UK. ABCLOVE: Voice therapy outcomes for patients with head and neck cancer. Head Neck 2016 Apr 22;38 Suppl 1(S1):E1810-E1813. [doi: 10.1002/hed.24322] [Medline: 26689974]
18. Palm P, Zwisler AO, Svendsen JH, Thygesen LC, Giraldi A, Jensen KG, et al. Sexual rehabilitation for cardiac patients with erectile dysfunction: a randomised clinical trial. Heart 2019 May 31;105(10):775-782. [doi: 10.1136/heartjnl-2018-313778] [Medline: 30381319]
19. Zimmer P, Trebing S, Timmers-Trebing U, Schenk A, Paust R, Bloch W, et al. Eight-week, multimodal exercise counteracts a progress of chemotherapy-induced peripheral neuropathy and improves balance and strength in metastasized colorectal cancer patients: a randomized controlled trial. Support Care Cancer 2018 Feb 30;26(2):615-624. [doi: 10.1007/s00520-017-3875-5] [Medline: 28963591]
20. Seland M, Skrede K, Lindemann K, Skaali T, Blomhoff R, Bruheim K, et al. Distress, problems and unmet rehabilitation needs after treatment for gynecological cancer. Acta Obstet Gynecol Scand 2022 Mar 29;101(3):313-322 [FREE Full text] [doi: 10.1111/aogs.14310] [Medline: 34964982]
21. Stensvold E, Stadskleiv K, Myklebust T, Wesenberg F, Helseth E, Bechensteen AG, et al. Unmet rehabilitation needs in 86% of Norwegian paediatric embryonal brain tumour survivors. Acta Paediatr 2020 Sep 06;109(9):1875-1886. [doi: 10.1111/apa.15188] [Medline: 31977119]
22. Schmidt ME, Wiskemann J, Ulrich CM, Schneeweiss A, Steindorf K. Self-reported physical activity behavior of breast cancer survivors during and after adjuvant therapy: 12 months follow-up of two randomized exercise intervention trials. Acta Oncol 2017 Apr;56(4):618-627. [doi: 10.1080/0284186X.2016.1275776] [Medline: 28084890]

XSL•FO
RenderX

23. Wu M, Gong X, Huang F, Li X, Wang S. Research status of application of resistance training based on exercise rehabilitation in breast cancer related lymphedema. Chin J Rehabil Med 2023;38(01):115-119. [doi: 10.3969/j.issn.1001-1242.2023.01.023]
24. Yeung AWK, Tosevska A, Klager E, Eibensteiner F, Laxar D, Stoyanov J, et al. Virtual and Augmented Reality Applications in Medicine: Analysis of the Scientific Literature. J Med Internet Res 2021 Feb 10;23(2):e25499 [FREE Full text] [doi: 10.2196/25499] [Medline: 33565986]
25. Cipresso P, Giglioli IAC, Raya MA, Riva G. The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature. Front Psychol 2018;9:2086 [FREE Full text] [doi: 10.3389/fpsyg.2018.02086] [Medline: 30459681]
26. Faric N, Potts HWW, Hon A, Smith L, Newby K, Steptoe A, et al. What Players of Virtual Reality Exercise Games Want: Thematic Analysis of Web-Based Reviews. J Med Internet Res 2019 Sep 16;21(9):e13833 [FREE Full text] [doi: 10.2196/13833] [Medline: 31538951]
27. Canning CG, Allen NE, Nackaerts E, Paul SS, Nieuwboer A, Gilat M. Virtual reality in research and rehabilitation of gait and balance in Parkinson disease. Nat Rev Neurol 2020 Aug;16(8):409-425. [doi: 10.1038/s41582-020-0370-2] [Medline: 32591756]
28. Zhang B, Li D, Liu Y, Wang J, Xiao Q. Virtual reality for limb motor function, balance, gait, cognition and daily function of stroke patients: A systematic review and meta-analysis. J Adv Nurs 2021 Aug;77(8):3255-3273. [doi: 10.1111/jan.14800] [Medline: 33675076]
29. Bond S, Laddu DR, Ozemek C, Lavie CJ, Arena R. Exergaming and Virtual Reality for Health: Implications for Cardiac Rehabilitation. Curr Probl Cardiol 2021 Mar;46(3):100472. [doi: 10.1016/j.cpcardiol.2019.100472] [Medline: 31606141]
30. Abbas RL, Cooreman D, Al Sultan H, El Nayal M, Saab IM, El Khatib A. The Effect of Adding Virtual Reality Training on Traditional Exercise Program on Balance and Gait in Unilateral, Traumatic Lower Limb Amputee. Games Health J 2021 Feb 01;10(1):50-56. [doi: 10.1089/g4h.2020.0028] [Medline: 33533682]
31. Shahid Khan MU, Hussain HG, Farooq M, Shahnoor S, Humayun M, Baig E, et al. The potential of virtual reality for rehabilitation intervention in females with breast cancer. Int J Surg 2022 Oct;106:106940. [doi: 10.1016/j.ijsu.2022.106940] [Medline: 36152921]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Khalil H, Peters MD, Tricco AC, Pollock D, Alexander L, McInerney P, et al. Conducting high quality scoping reviews-challenges and solutions. J Clin Epidemiol 2021 Feb;130:156-160. [doi: 10.1016/j.jclinepi.2020.10.009] [Medline: 33122034]
35. Atef D, Elkeblawy MM, El-Sebaie A, Abouelnaga WAI. A quasi-randomized clinical trial: virtual reality versus proprioceptive neuromuscular facilitation for postmastectomy lymphedema. J Egypt Natl Canc Inst 2020 Jun 15;32(1):29. [doi: 10.1186/s43046-020-00041-5] [Medline: 32537717]
36. Basha MA, Aboelnour NH, Alsharidah AS, Kamel FH. Effect of exercise mode on physical function and quality of life in breast cancer-related lymphedema: a randomized trial. Support Care Cancer 2022 Mar;30(3):2101-2110. [doi: 10.1007/s00520-021-06559-1] [Medline: 34669036]
37. Benzing V, Spitzhüttl J, Siegwart V, Schmid J, Grotzer M, Heinks T, et al. Effects of Cognitive Training and Exergaming in Pediatric Cancer Survivors-A Randomized Clinical Trial. Med Sci Sports Exerc 2020 Nov;52(11):2293-2302 [FREE Full text] [doi: 10.1249/MSS.0000000000002386] [Medline: 33064404]
38. Chen X, Jin A, Zhu H, Zhang D. Application and effect of virtual cognitive rehabilitation training in breast cancer patients with cognitive impairment after chemotherapy. Chin J Nurs 2019;54(05):664-667. [doi: 10.3761/j.issn.0254-1769.2019.05.005]
39. da Silva Alves R, Iunes D, de Carvalho J, Menezes F, Silva A, Borges J, et al. Effects of Exergaming on Quality of Life in Cancer Patients. Games for Health Journal 2018 Dec;7(6):385-392. [doi: 10.1089/g4h.2017.0174]
40. da Silva Alves R, Iunes DH, Pereira IC, Borges JBC, Nogueira DA, Silva AM, et al. Influence of Exergaming on the Perception of Cancer-Related Fatigue. Games Health J 2017 Apr;6(2):119-126. [doi: 10.1089/g4h.2016.0051] [Medline: 28418751]
41. Feyzioğlu Ö, Dinçer S, Akan A, Algun ZC. Is Xbox 360 Kinect-based virtual reality training as effective as standard physiotherapy in patients undergoing breast cancer surgery? Support Care Cancer 2020 Sep;28(9):4295-4303. [doi: 10.1007/s00520-019-05287-x] [Medline: 31907649]
42. Hamari L, Järvelä L, Lähteenmäki P, Arola M, Axelin A, Vahlberg T, et al. The effect of an active video game intervention on physical activity, motor performance, and fatigue in children with cancer: a randomized controlled trial. BMC Res Notes 2019 Nov 29;12(1):784 [FREE Full text] [doi: 10.1186/s13104-019-4821-z] [Medline: 31783907]

XSL•FO
RenderX

43. Hoffman AJ, Brintnall RA, Brown JK, Eye AV, Jones LW, Alderink G, et al. Too sick not to exercise: using a 6-week, home-based exercise intervention for cancer-related fatigue self-management for postsurgical non-small cell lung cancer patients. Cancer Nurs 2013;36(3):175-188. [doi: 10.1097/NCC.0b013e31826c7763] [Medline: 23051872]
44. Hoffman AJ, Brintnall RA, Brown JK, von Eye A, Jones LW, Alderink G, et al. Virtual reality bringing a new reality to postthoracotomy lung cancer patients via a home-based exercise intervention targeting fatigue while undergoing adjuvant treatment. Cancer Nurs 2014;37(1):23-33. [doi: 10.1097/NCC.0b013e318278d52f] [Medline: 23348662]
45. House G, Burdea G, Grampurohit N, Polistico K, Roll D, Damiani F, et al. A feasibility study to determine the benefits of upper extremity virtual rehabilitation therapy for coping with chronic pain post-cancer surgery. Br J Pain 2016 Nov;10(4):186-197 [FREE Full text] [doi: 10.1177/2049463716664370] [Medline: 27867508]
46. Jin A, Chen X, Zhang X, Chen J. Construction and application of a VR platform based on the Omaha system for rehabilitation management of breast cancer patients. Chin Gen Pract 2018;21(24):2987-2993. [doi: 10.12114/j.issn.1007-9572.2018.00.026]
47. Jin A, Chen X, Zhang X, Chen J. Design and clinical application of rehabilitation VR system for breast cancer patients. Chin J Nurs 2018;53(2):8. [doi: 10.3761/j.issn.0254-1769.2018.02.008]
48. Lin H, Wang L, Lin H. Effect of virtual cognitive rehabilitation training on the prognosis of young breast cancer patients with immediate breast reconstruction after modiffed radical mastectomy with nipleareola preservation. Chin J Front Med Sci (Electron Version) 2021;13(09):71-75. [doi: 10.12037/YXQY.2021.09-09]
49. Park H, Nam KE, Lim J, Yeo SM, Lee JI, Hwang JH. Real-Time Interactive Digital Health Care System for Postoperative Breast Cancer Patients: A Randomized Controlled Trial. Telemed J E Health 2023 Jul;29(7):1057-1067. [doi: 10.1089/tmj.2022.0360] [Medline: 36454316]
50. Sabel M, Sjölund A, Broeren J, Arvidsson D, Saury J, Blomgren K, et al. Active video gaming improves body coordination in survivors of childhood brain tumours. Disabil Rehabil 2016 Oct;38(21):2073-2084. [doi: 10.3109/09638288.2015.1116619] [Medline: 26728453]
51. Schwenk M, Grewal GS, Holloway D, Muchna A, Garland L, Najafi B. Interactive Sensor-Based Balance Training in Older Cancer Patients with Chemotherapy-Induced Peripheral Neuropathy: A Randomized Controlled Trial. Gerontology 2016;62(5):553-563 [FREE Full text] [doi: 10.1159/000442253] [Medline: 26678611]
52. Tanriverdi M, Cakir E, Akkoyunlu ME, Cakir FB. Effect of virtual reality-based exercise intervention on sleep quality in children with acute lymphoblastic leukemia and healthy siblings: A randomized controlled trial. Palliat Support Care 2022 Aug;20(4):455-461. [doi: 10.1017/S1478951522000268] [Medline: 35289266]
53. Tsuda K, Sudo K, Goto G, Takai M, Itokawa T, Isshiki T, et al. A Feasibility Study of Virtual Reality Exercise in Elderly Patients with Hematologic Malignancies Receiving Chemotherapy. Intern Med 2016;55(4):347-352 [FREE Full text] [doi: 10.2169/internalmedicine.55.5275] [Medline: 26875958]
54. Villumsen BR, Jorgensen MG, Frystyk J, Hørdam B, Borre M. Home-based 'exergaming' was safe and significantly improved 6-min walking distance in patients with prostate cancer: a single-blinded randomised controlled trial. BJU Int 2019 Oct;124(4):600-608. [doi: 10.1111/bju.14782] [Medline: 31012238]
55. Yang S, Chun MH, Son YR. Effect of virtual reality on cognitive dysfunction in patients with brain tumor. Ann Rehabil Med 2014 Dec;38(6):726-733 [FREE Full text] [doi: 10.5535/arm.2014.38.6.726] [Medline: 25566470]
56. Yoon J, Chun MH, Lee SJ, Kim BR. Effect of virtual reality-based rehabilitation on upper-extremity function in patients with brain tumor: controlled trial. Am J Phys Med Rehabil 2015 Jun;94(6):449-459. [doi: 10.1097/PHM.0000000000000192] [Medline: 25251249]
57. Zeng Y, Zhen Y, Guo Q. Design and preliminary application of a Virtual Reality technology-based cancer-related cognitive impairment assessment and rehabilitation system. Chin Nurs Manag 2020;20(02):7. [doi: 10.3969/j.issn.1672-1756.2020.02.007]
58. Zhou Z, Li J, Wang H, Luan Z, Li Y, Peng X. Upper limb rehabilitation system based on virtual reality for breast cancer patients: Development and usability study. PLoS One 2021;16(12):e0261220 [FREE Full text] [doi: 10.1371/journal.pone.0261220] [Medline: 34910786]
59. Zhu L, Yu J, Li Q. Application of rehabilitation training virtual reality system in postoperative rehabilitation exercise for breast cancer patients. J Qilu Nurs 2019;25(18):36. [doi: 10.3969/j.issn.1006-7256.2019.18.036]
60. Yazdipour AB, Saeedi S, Bostan H, Masoorian H, Sajjadi H, Ghazisaeedi M. Opportunities and challenges of virtual reality-based interventions for patients with breast cancer: a systematic review. BMC Med Inform Decis Mak 2023 Jan 23;23(1):17 [FREE Full text] [doi: 10.1186/s12911-023-02108-4] [Medline: 36691014]
61. Saposnik G, Cohen LG, Mamdani M, Pooyania S, Ploughman M, Cheung D, Stroke Outcomes Research Canada. Efficacy and safety of non-immersive virtual reality exercising in stroke rehabilitation (EVREST): a randomised, multicentre, single-blind, controlled trial. Lancet Neurol 2016 Sep;15(10):1019-1027 [FREE Full text] [doi: 10.1016/S1474-4422(16)30121-1] [Medline: 27365261]
62. Gatica-Rojas V, Cartes-Velásquez R, Soto-Poblete A, Lizama LEC. Postural control telerehabilitation with a low-cost virtual reality protocol for children with cerebral palsy: Protocol for a clinical trial. PLoS One 2023;18(8):e0268163 [FREE Full text] [doi: 10.1371/journal.pone.0268163] [Medline: 37590295]

XSL•FO
RenderX

63. Tuena C, Pedroli E, Trimarchi PD, Gallucci A, Chiappini M, Goulene K, et al. Usability Issues of Clinical and Research Applications of Virtual Reality in Older People: A Systematic Review. Front Hum Neurosci 2020;14:93 [FREE Full text] [doi: 10.3389/fnhum.2020.00093] [Medline: 32322194]
64. Buche H, Michel A, Piccoli C, Blanc N. Contemplating or Acting? Which Immersive Modes Should Be Favored in Virtual Reality During Physiotherapy for Breast Cancer Rehabilitation. Front Psychol 2021;12:631186 [FREE Full text] [doi: 10.3389/fpsyg.2021.631186] [Medline: 33897539]
65. Melillo A, Chirico A, De Pietro G, Gallo L, Caggianese G, Barone D, et al. Virtual Reality Rehabilitation Systems for Cancer Survivors: A Narrative Review of the Literature. Cancers (Basel) 2022 Jun 28;14(13):3163 [FREE Full text] [doi: 10.3390/cancers14133163] [Medline: 35804933]
66. Lu H, Xie J, Gerido LH, Cheng Y, Chen Y, Sun L. Information Needs of Breast Cancer Patients: Theory-Generating Meta-Synthesis. J Med Internet Res 2020 Jul 28;22(7):e17907 [FREE Full text] [doi: 10.2196/17907] [Medline: 32720899]
67. Serweta-Pawlik A, Lachowicz M, Żurek A, Rosen B, Żurek G. Evaluating the Effectiveness of Visuospatial Memory Stimulation Using Virtual Reality in Head and Neck Cancer Patients-Pilot Study. Cancers (Basel) 2023 Mar 07;15(6):1639 [FREE Full text] [doi: 10.3390/cancers15061639] [Medline: 36980525]
68. Jenny C, Reuter C. Usability of Individualized Head-Related Transfer Functions in Virtual Reality: Empirical Study With Perceptual Attributes in Sagittal Plane Sound Localization. JMIR Serious Games 2020 Sep 08;8(3):e17576 [FREE Full text] [doi: 10.2196/17576] [Medline: 32897232]
69. Logan DE, Simons LE, Caruso TJ, Gold JI, Greenleaf W, Griffin A, et al. Leveraging Virtual Reality and Augmented Reality to Combat Chronic Pain in Youth: Position Paper From the Interdisciplinary Network on Virtual and Augmented Technologies for Pain Management. J Med Internet Res 2021 Apr 26;23(4):e25916 [FREE Full text] [doi: 10.2196/25916] [Medline: 33667177]
70. Sousa CV, Fernandez A, Hwang J, Lu AS. The Effect of Narrative on Physical Activity via Immersion During Active Video Game Play in Children: Mediation Analysis. J Med Internet Res 2020 Mar 31;22(3):e17994 [FREE Full text] [doi: 10.2196/17994] [Medline: 32229466]
71. Aguilar-Lazcano CA, Rechy-Ramirez EJ, Hu H, Rios-Figueroa HV, Marin-Hernandez A. Interaction Modalities Used on Serious Games for Upper Limb Rehabilitation: A Systematic Review. Games Health J 2019 Oct;8(5):313-325. [doi: 10.1089/g4h.2018.0129] [Medline: 31287734]
72. Inchusri T, Surangsrirat D, Kwanmuang P, Poomivanichakij P, Apiwatgaroon P, Ongprakobkul S, et al. Association of Generation and Group Size With the Usage of a Mobile Health App in Thailand: Secondary Analysis of the ThaiSook Cohort Study. J Med Internet Res 2023 Aug 17;25:e45374 [FREE Full text] [doi: 10.2196/45374] [Medline: 37590057]
73. Owen AM, Hampshire A, Grahn JA, Stenton R, Dajani S, Burns AS, et al. Putting brain training to the test. Nature 2010 Jun 10;465(7299):775-778 [FREE Full text] [doi: 10.1038/nature09042] [Medline: 20407435]
74. Ezenwankwo EF, Nnate DA, Usoro GD, Onyeso CP, Anieto IB, Ibeneme SC, et al. A scoping review examining the integration of exercise services in clinical oncology settings. BMC Health Serv Res 2022 Feb 21;22(1):236 [FREE Full text] [doi: 10.1186/s12913-022-07598-y] [Medline: 35189864]
75. Wu Y, Wang N, Zhang H, Sun X, Wang Y, Zhang Y. Effectiveness of Virtual Reality in Symptom Management of Cancer Patients: A Systematic Review and Meta-Analysis. J Pain Symptom Manage 2023 May;65(5):e467-e482 [FREE Full text] [doi: 10.1016/j.jpainsymman.2023.01.023] [Medline: 36758907]

Abbreviations

ADLs: activities of daily living
AR: augmented reality
CIPN: chemotherapy-induced peripheral neuropathy
CRCI: cancer-related cognitive impairment
CRD: cancer-related dysfunction
CRF: cancer-related fatigue
PICOS: Population, Intervention, Comparator, Outcome, Study design
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QoL: quality of life
RCT: randomized clinical trial
ROM: range of motion
VR: virtual reality
VRER: virtual reality–based exercise rehabilitation

XSL•FO
RenderX

Edited by A Mavragani; submitted 24.05.23; peer-reviewed by S Triscari, S Okita; comments to author 11.10.23; revised version received 31.10.23; accepted 30.01.24; published 26.02.24.

Please cite as:
Su Z, Zhang L, Lian X, Guan M
Virtual Reality–Based Exercise Rehabilitation in Cancer-Related Dysfunctions: Scoping Review
J Med Internet Res 2024;26:e49312
URL: https://www.jmir.org/2024/1/e49312
doi:10.2196/49312
PMID:38407951

©Zhenzhen Su, Liyan Zhang, Xuemin Lian, Miaomiao Guan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Virtual Reality–Based Training in Chronic Low Back Pain: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Ran Li[1,2], MD; Yinghao Li[3], MD; Youli Kong[1,2], MSc; Hanbin Li[1,2], BSc; Danrong Hu[1,2], PhD; Chenying Fu[4,5], PhD; Quan Wei[1,2], MD

[1]Rehabilitation Medicine Center and Institute of Rehabilitation Medicine, West China Hospital, Sichuan University, Chengdu, China
[2]Key Laboratory of Rehabilitation Medicine in Sichuan Province, Chengdu, China
[3]Department of Orthopedics, Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China
[4]National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, China
[5]Aging and Geriatric Mechanism Laboratory, West China Hospital, Sichuan University, Chengdu, China

Corresponding Author:
Quan Wei, MD
Rehabilitation Medicine Center and Institute of Rehabilitation Medicine
West China Hospital
Sichuan University
No.37 Guo Xue Alley
Chengdu, 610041
China
Phone: 86 18980606730
Email: weiquan@scu.edu.cn

Abstract

Background: Low back pain is one of the most prevalent pain conditions worldwide. Virtual reality–based training has been used for low back pain as a new treatment strategy. Present evidence indicated that the effectiveness of virtual reality–based training for people with chronic low back pain is inconclusive.

Objective: This study conducted a meta-analysis to evaluate the immediate- and short-term effects of virtual reality–based training on pain, pain-related fear, and disability in people with chronic low back pain.

Methods: We searched the PubMed, Embase, Web of Science, PEDro, CENTRAL, and CINAHL databases from inception until January 2024. Only randomized controlled trials assessing the effects of virtual reality–based training on individuals with chronic low back pain were selected. The outcomes were focused on pain, pain-related fear measured by the Tampa Scale of Kinesiophobia, and disability measured by the Oswestry Disability Index. The immediate term was defined as the immediate period after intervention, and the short term was defined as 3 to 6 months after intervention. The Cochrane Risk of Bias tool and the GRADE (Grading of Recommendations, Assessment, Development and Evaluation) approach were used to evaluate the quality of the methodology and evidence, respectively.

Results: In total, 20 randomized controlled trials involving 1059 patients were eligible for analysis. Virtual reality–based training showed significant improvements in pain (mean difference [MD] –1.43; 95% CI –1.86 to –1.00; I^2=95%; P<.001), pain-related fear using the Tampa Scale of Kinesiophobia (MD –5.46; 95% CI –9.40 to 1.52; I^2=90%; P=.007), and disability using the Oswestry Disability Index (MD –11.50; 95% CI –20.00 to –3.01; I^2=95%; P=.008) in individuals with chronic low back pain immediately after interventions. However, there were no significant differences observed in pain (P=.16), pain-related fear (P=.10), and disability (P=.43) in the short term.

Conclusions: These findings indicated that virtual reality–based training can be used effectively for individuals with chronic low back pain in the immediate term, especially to reduce pain, alleviate pain-related fear, and improve disability. However, the short-term benefits need more high-quality trials to be demonstrated.

Trial Registration: PROSPERO CRD42021292633; http://tinyurl.com/25mydxpz

(J Med Internet Res 2024;26:e45406) doi:10.2196/45406

XSL•FO
RenderX

KEYWORDS

virtual reality; low back pain; chronic; rehabilitation; exercise

Introduction

Low back pain is a common musculoskeletal symptom influenced by complex interactions among biological, psychological, and social factors [1-3]. With more than 568 million people experiencing low back pain globally, it has become the leading cause of years lived with disability worldwide [4,5]. Low back pain is defined by the location of pain, typically between the lower rib margins and the buttock creases [6]. Once the symptom persists for more than 3 months, it can be considered as chronic low back pain. Chronic low back pain may be accompanied by fear avoidance and dysfunction. Research has suggested that individuals who hold negative beliefs about their pain or their condition may experience an exaggerated fear of pain and the potential negative consequences of their symptoms. This fear can lead to a cycle of catastrophizing thoughts, pain-related fear, and avoidance of movements that they perceive as potentially painful or harmful [7]. Furthermore, fear avoidance and catastrophizing thoughts are considered catalysts for chronicity, resulting in prolonged recovery and increased disability rates [8].

Given the impaired physical function, quality of life, and even social participation from chronic low back pain [9], it is a global priority to establish an effective treatment [6,10,11]. Multicomponent exercises are frequently prescribed by physicians for chronic low back pain, as recommended by clinical practice guidelines and established by randomized controlled trials (RCTs) [12-16]. Exercises encompass a diverse set of components including specific activities, postures, or movements (or all) [17]. It is reported that exercises may benefit patients with chronic low back pain by improving muscle strength and movement and enhancing postural musculature, stability, and coordination, or a combination of these factors [18]. Among the various exercises available, virtual reality–based training is becoming increasingly popular for the treatment of chronic low back pain [19-22]. It refers to digital training via computer-generated realities implemented with stereoscopic displays [23,24]. In addition, virtual reality–based training has been shown to reduce the focus on pain by dividing attention to tasks [25,26] and increase the motivation of movement through progressive achievement [27-29].

Two systematic reviews suggested that virtual reality–based training had a positive effect on improving pain intensity [30] and fear avoidance [31] in people with low back pain, while other systematic reviews indicated that the effectiveness of virtual reality–based training was inconclusive [32,33]. To our knowledge, evidence on the immediate-, short-, and long-term benefits of virtual reality–based training in patients with chronic low back pain does not exist as well. The evidence on the potential effectiveness of virtual reality–based training for chronic low back pain is still controversial and deficient. Therefore, we conducted this systematic review and meta-analysis to evaluate the immediate-term and short-term efficacy of virtual reality–based training for chronic low back pain.

Methods

Design

This systematic review and meta-analysis followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [34] and PRISMA 2020 guidelines (Multimedia Appendix 1) [35] and was performed following a protocol registered in PROSPERO (CRD42021292633).

Search Strategy

Two reviewers (RL and YL) independently searched the PubMed, Embase, Web of Science, PEDro, CENTRAL, and CINAHL databases from inception until January 2024. Low back pain and virtual reality were searched as keywords. The full search strategy is described in Multimedia Appendix 2. The bibliographic references of the included studies and previous systematic reviews were also checked to identify additional trials.

Study Selection

Studies concerning the effects of virtual reality involving individuals with low back pain were included in this systematic review. The inclusion criteria were as follows: (1) RCTs of parallel groups, (2) participants experiencing chronic low back pain, (3) virtual reality as intervention, (4) studies assessed clinical outcomes (eg, pain, pain-related fear, or disability), and (5) duration of intervention ≥8 sessions. The exclusion criteria were as follows: (1) reviews, case reports, and conference abstracts; (2) studies without enough information for data analysis; and (3) studies not written in English. The detailed inclusion and exclusion criteria are shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. The inclusion and exclusion criteria.

Inclusion criteria

- Paper type: randomized controlled trials of parallel groups
- Study subjects: participants with chronic low back pain (3 months or more)
- Interventions: virtual reality–based training as intervention alone or in combination with physical therapy, irregular of virtual reality devices; the duration of intervention is ≥8 sessions
- Outcomes: at least 1 of the following outcome measurements: pain (Visual Analog Scale, Numerical Rating Scale, and Defense and Veterans Pain Rating Scale), pain-related fear (Tampa Scale of Kinesiophobia), or disability (Oswestry Disability Index)
- Language: written in English

Exclusion criteria

- Paper type: reviews, case reports, and conference abstracts
- Study subjects: participants with other pain or low back pain lasting for less than 3 months
- Interventions: no interventions involved virtual reality
- Outcomes: without the required outcomes or no available data and not enough information for analysis
- Language: non-English publications

Outcome Measures

The outcomes were focused on pain, pain-related fear, and disability. The pain intensity was measured by the Visual Analog Scale, Numerical Rating Scale, and Defense and Veterans Pain Rating Scale, which have demonstrated strong validity and reliability in clinical and research settings [36-39]. The pain-related fear was measured by the Tampa Scale of Kinesiophobia, which is frequently used in patients with back pain [40,41]. As for the functional disability, the Oswestry Disability Index (ODI) was used [42].

Data Extraction

Two reviewers (RL and YL) independently extracted the main information for the included studies using a standard extraction spreadsheet (Microsoft Excel, Microsoft Corp). A third reviewer (YK) was consulted if the initial reviewers (RL and YL) disagreed. The detailed characteristics of the selected studies were summarized, which included study characteristics (author, year of publication, country, sample size, and follow-up points), population characteristics (gender, age, and pain duration), intervention characteristics (type, frequency, and duration), and outcome measurements (pain, pain-related fear, and disability).

Quality Assessment

The quality of RCTs was assessed using the Cochrane Risk of Bias tool [43], which consists of 5 domains and an overall judgment. The 5 domains include randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Based on the answers to a series of signaling questions, the judgment options in each domain consist of “low risk of bias,” “some concerns,” and “high risk of bias.” Two independent researchers (RL and HL) assessed the quality of each eligible study. Disagreements in the data were settled by consensus.

We used the GRADE (Grading of Recommendations, Assessment, Development and Evaluation) approach to assess the quality of evidence, which is rated as high, moderate, low, and very low [44]. The quality of RCTs is initially considered high and then downgraded based on risk of bias, imprecision, inconsistency, indirectness, and publication bias [45-49].

Statistical Analysis

This meta-analysis was conducted using Review Manager (version 5.30; Cochrane Collaboration), and all extracted data were input and checked by the reviewers (RL, YL, and YK). Mean differences (MDs) were calculated with the 95% CIs using the inverse variance method, when the outcomes were evaluated with the same scale. The chi-square test and inconsistency (I^2) were used to calculate statistical heterogeneity. The random effect model was used when I^2>50%; otherwise, the fixed-effect model was used. Statistical differences by meta-analysis were identified as P<.05. Furthermore, the minimal clinically important difference (MCID) was taken into consideration in this study, and the standard was established according to a previous report.

Subgroup analyses comparing the efficacy of virtual reality on pain, pain-related fear, and disability were performed for immediate-term and short-term outcomes separately. The immediate term was defined as the immediate period after intervention, and the short term was defined as 3 to 6 months after intervention.

Results

Study Selection and Characteristics

The initial search procedure yielded 583 records in total, 294 of which were removed due to duplication. After the screening process, 20 studies were selected for this systematic review and meta-analysis (Figure 1). The 20 papers included 1059 individuals diagnosed with chronic low back pain. The characteristics (sample size, gender, interventions, exposures, and outcome measurements) of the involved studies were summarized (Table 1). All studies had an RCT design: 7 used

XSL•FO
RenderX

a 3-arm parallel-group design [21,50-55], and 1 used a 4-arm parallel-group design [56]. Of the included papers, 4 studies involved only female patients [20,22,57,58], and 6 studies involved only male patients [52-56,59].

Figure 1. Flowchart showing the study selection process. RCT: randomized controlled trial.

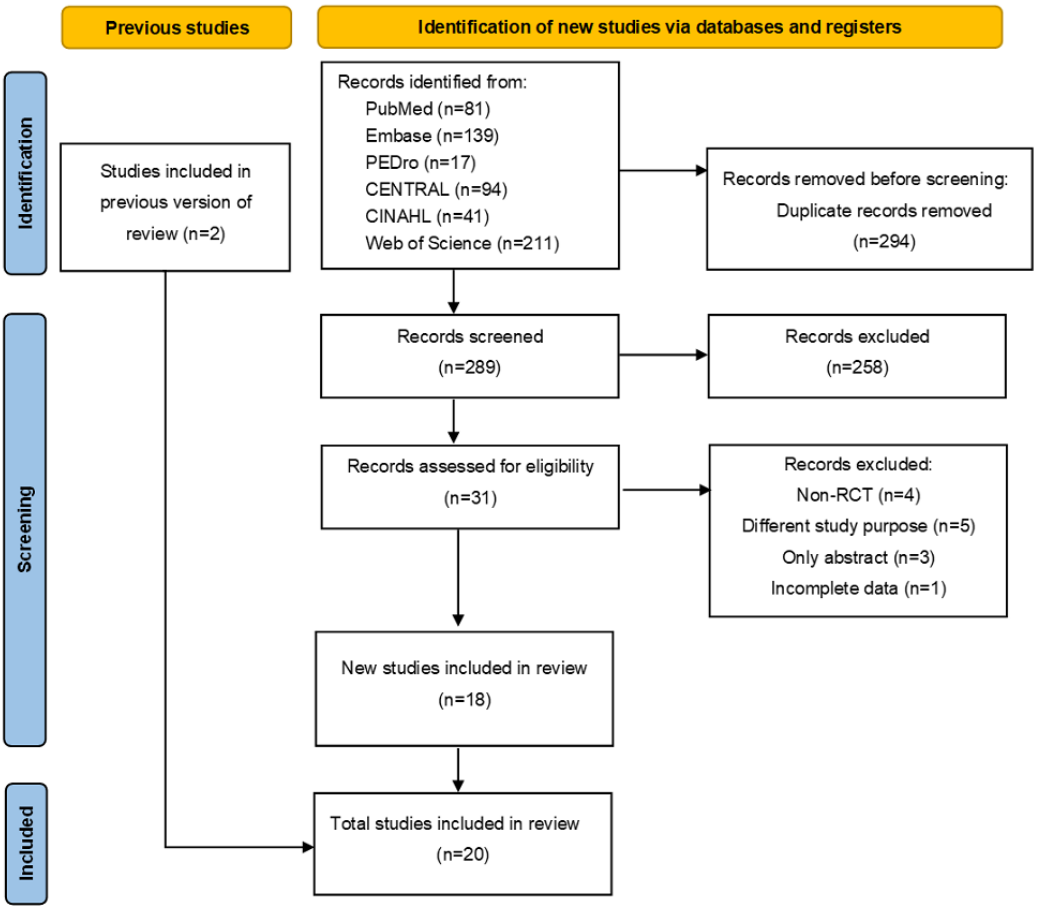

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study (year)	Country	Sample size	Gender	Mean age (years)	Pain duration (months)	Intervention	Exposure	Outcome measures	Follow-up points
Afzal et al (2022) [19]	Pakistan	• Exp[a]: 42 • Con[b]: 42	• Male: 28 • Female: 56	• Exp: 38.2 • Con: 37.5	—[c]	• Exp: Virtual reality–based exercises via kinetic exergames+physical therapy • Con: Back strengthening exercises+physical therapy	12 sessions; 3 times per week	VAS[d]; ODI[e]	4 weeks
Eccleston et al (2022) [50]	Finland	• Exp1: 14 • Exp2: 17 • Con: 11	• Male: 5 • Female: 37	• Exp1: 55.14 • Exp2: 52.76 • Con: 57.09	• >6	• Exp1: Virtual reality–based cognitive behavioral intervention via Oculus Quest and Touch virtual reality headset and handheld controllers • Exp2: Viewed text-based cognitive behavioral intervention via Oculus Quest and Touch virtual reality headset • Con: Standard care	30 sessions; 5 times per week; 15-60 minutes per session	NRS[f]; ODI; PROMIS[g] 6b; TSK[h]; pain medications; EuroQoL-5D-5L[i]; adverse events; PGIC[j]; Game Experience Questionnaire	9 weeks, 5 months
Garcia et al (2021) [60]	United States	• Exp: 94 • Con: 94	• Male: 43 • Female: 144	• Exp: 52.1 • Con: 51.3	• ≥6	• Exp: Virtual reality–based cognitive behavioral therapy via Pico G2 4K head-mounted virtual reality device (3D visual displays) • Con: Sham virtual reality–based cognitive behavioral therapy via Pico G2 4K head-mounted virtual reality device	56 sessions; 7 times per week; 2-16 minutes per session	DVPRS[k]; DVPRS-II; PSEQ[l]-2; PGIC; PROMIS 6b and 6a; PCS[m]; CPAQ-8[n]	8 weeks, 3 months, 6 months
Groenveld et al (2023) [61]	Netherlands	• Exp: 20 • Con: 20	• Male: 7 • Female: 33	• Exp: 51 • Con: 52	• Exp: 5 • Con: 4	• Exp: Virtual reality–based exercises via an Oculus Go head-mounted display • Con: Daily life routines	60 sessions; 3 times per day; 10-30 minutes per session	SF-12[o]; VAS; PCS; HADS[p]; ODI; PCCL[q]; NEADL[r]; BPI[s]	4 weeks, 4 months
Kim et al (2014) [20]	South Korea	• Exp: 15 • Con: 15	• Male: 0 • Female: 30	• Exp: 44.33 • Con: 50.46	—	• Exp: Virtual reality–based yoga program via Wii Fit activities • Con: Trunk stabilizing exercise+physical therapy program	12 sessions; 3 times per week; 30 minutes per session	VAS; pressure algometry; ODI; RMDQ[t]; FABQ[u]	4 weeks
Kim et al (2020) [62]	South Korea	• Exp: 24 • Con: 24	• Male: 26 • Female: 22	• Exp: 26.0 • Con: 28.79	• Exp: 58.22 • Con: 101.55	• Exp: Virtual reality–based simulated horseback riding • Con: Stabilization exercise via suspension	16 sessions; 2 times per week; 46 minutes per session	NRS; ODI; RMDQ; FABQ	4 weeks, 8 weeks, 6 months

XSL•FO
RenderX

Study (year)	Country	Sample size	Gender	Mean age (years)	Pain duration (months)	Intervention	Exposure	Outcome measures	Follow-up points
Li et al (2021) [51]	China	• Exp1: 11 • Exp2: 12 • Con: 11	• Male: 9 • Female: 25	• Exp1: 21.91 • Exp2: 23.75 • Con: 25.36	• Exp1: 30.18 • Exp2: 38.83 • Con: 49.82	• Exp1: Virtual reality–based Fruit Ninja game via Kinect Xbox 360+magnetic therapy • Exp2: Ultrasound-guided abdominal drawing-in maneuver training+magnetic therapy • Con: Conventional thermal magnetic therapy	10 sessions; 5 times per week; 30 minutes per session	VAS; ODI; sEMG[v]	2 weeks
Meinke et al (2022) [63]	Switzerland	• Exp: 13 • Con: 14	• Male: 10 • Female: 17	• Exp: 40.14 • Con: 40.85	—	• Exp: Virtual reality–based exergame via 2 inertial measurement units • Con: Daily life routines	9 sessions; 3 times per week; 20 minutes per session	NRS; TSK; RMDQ; WHOQOL-Bref[w]	3 weeks
Monteiro-Junior et al (2015) [57]	Brazil	• Exp: 16 • Con: 14	• Male: 0 • Female: 30	• 68	—	• Exp: Virtual reality–based physical training via Nintendo Wii-motion and WBBx+core and strength training • Con: Core and strength training	24 sessions; 3 times per week; 90 minutes per session	NRS; WBB; sit-to-stand test; POMS[y]	8 weeks
Nambi et al (2020) [54]	Saudi Arabia	• Exp1: 15 • Exp2: 15 • Con: 15	• Male: 45 • Female: 0	• Exp1: 21.25 • Exp2: 20.23 • Con: 20.78	• Exp1: 4.1 • Exp2: 4.1 • Con: 4.3	• Exp1: Virtual reality–based-balance training via Pro-Kin system PK 252 N • Exp2: Balance training via Swiss ball • Con: Conventional balance training via active isotonic and isometric exercise	20 sessions; 5 times per week; 30 minutes per session	VAS; player wellness; sprint performance; jump performance	4 weeks, 8 weeks, 6 months
Nambi et al (2021) [55]	Saudi Arabia	• Exp1: 20 • Exp2: 20 • Con: 20	• Male: 60 • Female: 0	• Exp1: 21.45 • Exp2: 21.39 • Con: 20.97	• Exp1: 4.8 • Exp2: 5.2 • Con: 4.9	• Exp1: Virtual reality training via firing game • Exp2: Core stabilization training via therapeutic ball • Con: Traditional active balance exercise	20 sessions; 5 times per week; 30 minutes per session	NRS; physical fitness index; sprint performance; jump performance	4 weeks, 8 weeks, 6 months
Nambi et al (2021) [53]	Saudi Arabia	• Exp1: 18 • Exp2: 18 • Con: 18	• Male: 54 • Female: 0	• Exp1: 22.3 • Exp2: 21.4 • Con: 21.9	• Exp1: 5.4 • Exp2: 5.3 • Con: 5.5	• Exp1: Virtual reality training+hot pack therapy+ultrasound • Exp2: Isokinetic training+hot pack therapy+ultrasound • Con: Conventional core training+hot pack therapy+ultrasound	20 sessions; 5 times per week; 30 minutes per session	VAS; TSK; blood serum level of stress hormones	4 weeks, 6 months
Nambi et al (2021) [52]	Saudi Arabia	• Exp1: 20 • Exp2: 20 • Con: 20	• Male: 60 • Female: 0	• Exp1: 23.2 • Exp2: 22.8 • Con: 23.3	• Exp1: 5.8 • Exp2: 5.2 • Con: 5.4		20 sessions; 5 times per week; 30 minutes per session	VAS; TSK; blood serum level of stress hormones	4 weeks, 6 months

XSL•FO
RenderX

Study (year)	Country	Sample size	Gender	Mean age (years)	Pain duration (months)	Intervention	Exposure	Outcome measures	Follow-up points
						• Exp1: Virtual reality training+hot pack therapy+ultrasound • Exp2: Isokinetic training+hot pack therapy+ultrasound • Con: Conventional core training+hot pack therapy+ultrasound			
Oh et al (2014) [56]	South Korea	• Exp1: 9 • Exp2: 9 • Exp3: 10 • Con: 9	• Male: 37 • Female: 0	• Exp1: 20.7 • Exp2: 20.56 • Exp3: 20.33 • Con: 20.44	• Exp1: 6.38 • Exp2: 6.21 • Exp3: 7.57 • Con: 6.75	• Exp1: Virtual reality training via horse simulator machine for 10 minutes • Exp2: Virtual reality training via horse simulator machine for 20 minutes • Exp3: Virtual reality training via horse simulator machine for 30 minutes • Con: Daily life routines	40 sessions; 5 times per week; 15-35 minutes per session	VAS; body composition; isokinetic trunk and hip extension or flexion and hip abduction or adduction	8 weeks
Park et al (2013) [21]	South Korea	• Exp1: 8 • Exp2: 8 • Con: 8	—	• Exp1: 44.12 • Exp2: 43.37 • Con: 44.12	• Exp1: 17.0 • Exp2: 16.0 • Con: 18.75	• Exp1: Virtual reality–based training via Nintendo Wii program+physical therapy • Exp2: Lumbar stabilization exercise+physical therapy • Con: Physical therapy (eg, hot pack, interferential current therapy, and deep heat with ultrasound)	24 sessions; 3 times per week; 80 minutes per session	VAS; isometric lifting strength for back strength; 1-legged Stand Test for balance ability; RAND-36[z]; SF-36[aa]	8 weeks
Park et al (2020) [58]	South Korea	• Exp: 40 • Con: 40	• Male: 0 • Female: 80	• Exp: 71.35 • Con: 72.05	• Exp: 23.61 • Con: 22.10	• Exp: Virtual reality–based training via equestrian simulator • Con: Sitting on the equestrian simulator	36 sessions; 3 times per week; 30 minutes per session	VAS; ODI; body composition; isokinetic trunk extension and flexion; spinal alignment	12 weeks
Yalfani et al (2022) [22]	Japan	• Exp: 13 • Con: 12	• Male: 0 • Female: 25	• Exp: 68 • Con: 67.08	• >6	• Exp: Virtual reality–based training via Xbox Kinect headset • Con: Daily life routines	24 sessions; 3 times per week; 30 minutes per session	VAS; SF-36; FRI[ab]; BBS[ac]	2 weeks
Yilmaz Yelvar et al (2017) [64]	Iran	• Exp: 22 • Con: 22	• Male: 16 • Female: 28	• Exp: 46.27 • Con: 52.81	• Exp: 5.27 • Con: 7.45	• Exp: Virtual walking task via Vita Digital Productions+physical therapy • Con: Physical therapy (eg, hot pack, TENSad, deep heat with ultrasound, and therapeutic exercises)	10 sessions; 5 times per week	VAS; TSK; ODI; Nottingham Health Profile; TUG[ae]; 6MWT[af]; single-leg balance test	2 weeks
Yoo et al (2014) [59]	South Korea								8 weeks

XSL•FO
RenderX

Study (year)	Country	Sample size	Gender	Mean age (years)	Pain duration (months)	Intervention	Exposure	Outcome measures	Follow-up points
		• Exp: 24 • Con: 23	• Male: 47 • Female: 0	• Exp: 20.44 • Con: 20.7	• Exp: 9.41 • Con: 8.35	• Exp: Virtual reality–based training via horse simulator • Con: Daily life routines	24 sessions; 3 times per week; 20-50 minutes per session	VAS; body composition; isokinetic trunk strength	
Zadro et al (2019) [65]	Turkey	• Exp: 30 • Con: 30	• Male: 29 • Female: 31	• Exp: 68.8 • Con: 67.8	• >3	• Exp: Video game home-based exercise via Nintendo Wii U console • Con: Daily life routines	24 sessions; 3 times per week; 60 minutes per session	NRS; TSK; PSEQ; 3-item questionnaire; Rapid Assessment of Physical Activity questionnaire; PSFS[ag]; RMDQ; 16-item Falls Efficacy Scale-International Questionnaire	8 weeks

[a]Exp: experimental group.
[b]Con: control group.
[c]Not available.
[d]VAS: Visual Analog Scale.
[e]ODI: Oswestry Disability Index.
[f]NRS: Numerical Rating Scale.
[g]PROMIS: Patient-Reported Outcomes Measurement Information System.
[h]TSK: Tampa Scale of Kinesiophobia.
[i]EuroQoL-5D-5L: European Quality of Life 5-dimension, 5-level scale.
[j]PGIC: Patient's Global Impression of Change.
[k]DVPRS: Defense and Veterans Pain Rating Scale.
[l]PSEQ: Pain Self-Efficacy Questionnaire.
[m]PCS: Pain Catastrophizing Scale.
[n]CPAQ-8: Chronic Pain Acceptance Questionnaire.
[o]SF-12: 12-item Short-Form Health Survey.
[p]HADS: Hospital Anxiety and Depression Scale.
[q]PCCL: Pain Coping and Cognition List.
[r]NEADL: Nottingham Extended Activities of Daily Living.
[s]BPI: Brief Pain Inventory.
[t]RMDQ: Roland Morris Disability Questionnaire.
[u]FABQ: Fear Avoidance Beliefs Questionnaire.
[v]sEMG: surface electromyography.
[w]WHOQOL-Brief: World Health Organization Quality of Life Questionnaire-short version.
[x]WBB: Wii Balance Board.
[y]POMS: Profile of Mood States.
[z]RAND-36: RAND-36 Health Status Inventory.
[aa]SF-36: 36-item Short-Form Health Survey.
[ab]FRI: Fall Risk Index.
[ac]BBS: Biodex Balance System.
[ad]TENS: transcutaneous electrical nerve stimulation.
[ae]TUG: Timed Up and Go Test.
[af]6MWT: 6-Minute Walk Test.

XSL•FO
RenderX

[ag]PSFS: Patient-Specific Functional Scale.

Risk of Bias

In total, 20 studies were considered as having some concerns in overall bias, 2 as having a high risk of bias [20,21], and 1 as having a low risk of bias [19] (Figure 2 [19-22,50-65]). Due to the nature and setting of the intervention, it was not possible to blind the patients or therapists delivering the intervention, leading to some concerns in the second domain (deviations from the intended intervention). There was a high risk of bias regarding the measurement of outcome because of insufficient information on blinded assessment [20,21]. Baseline between intervention groups had significant differences in 2 studies [64,65].

Figure 2. Risk-of-bias graph and summary.

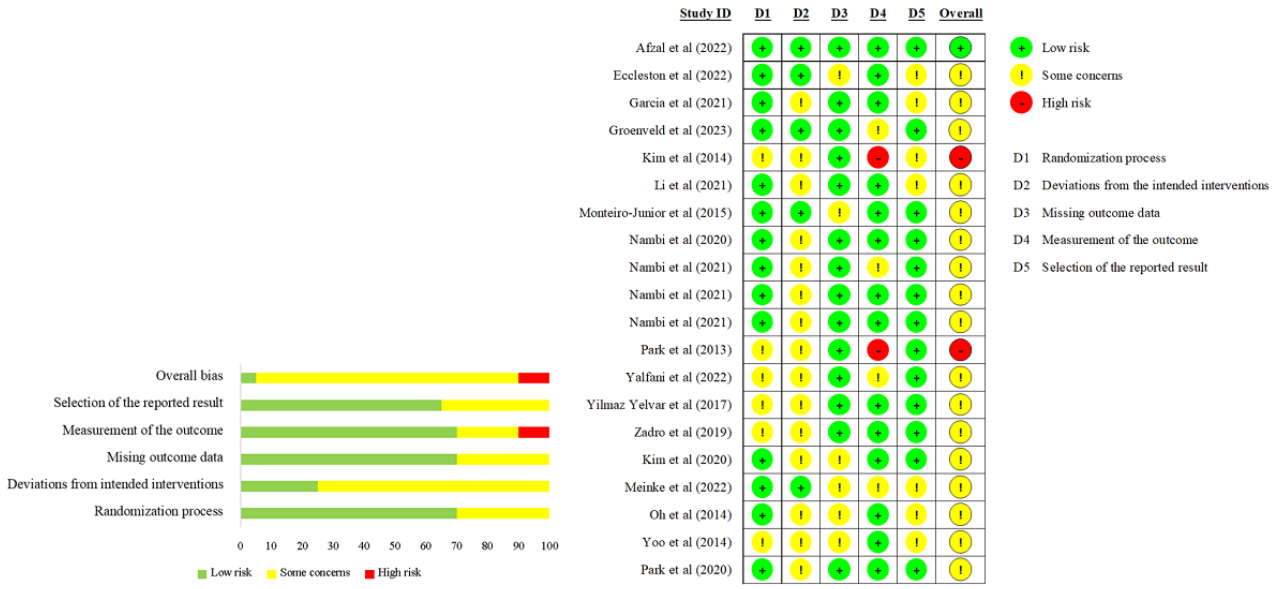

Outcomes

Pain

All studies investigated the efficacy of virtual reality on pain using the Visual Analog Scale, Numerical Rating Scale, and Defense and Veterans Pain Rating Scale on a scale from 0 to 10. In total, 19 RCTs provided available data that were pooled into a meta-analysis (Figure 3 [19-22,50-60,62-65]). There were statistically significant differences between virtual reality–based training and conventional treatments for pain reduction in the immediate term (MD –1.43; 95% CI –1.86 to –1.00; I^2=95%; P<.001) but not in the short term (MD –0.57; 95% CI –1.36 to 0.22; I^2=99%; P=.16). Both did not reach an MCID at the level of 2.5 [66,67].

XSL•FO
RenderX

Figure 3. Forest plots for virtual reality–based training compared with controls in pain. IV: inverse variance.

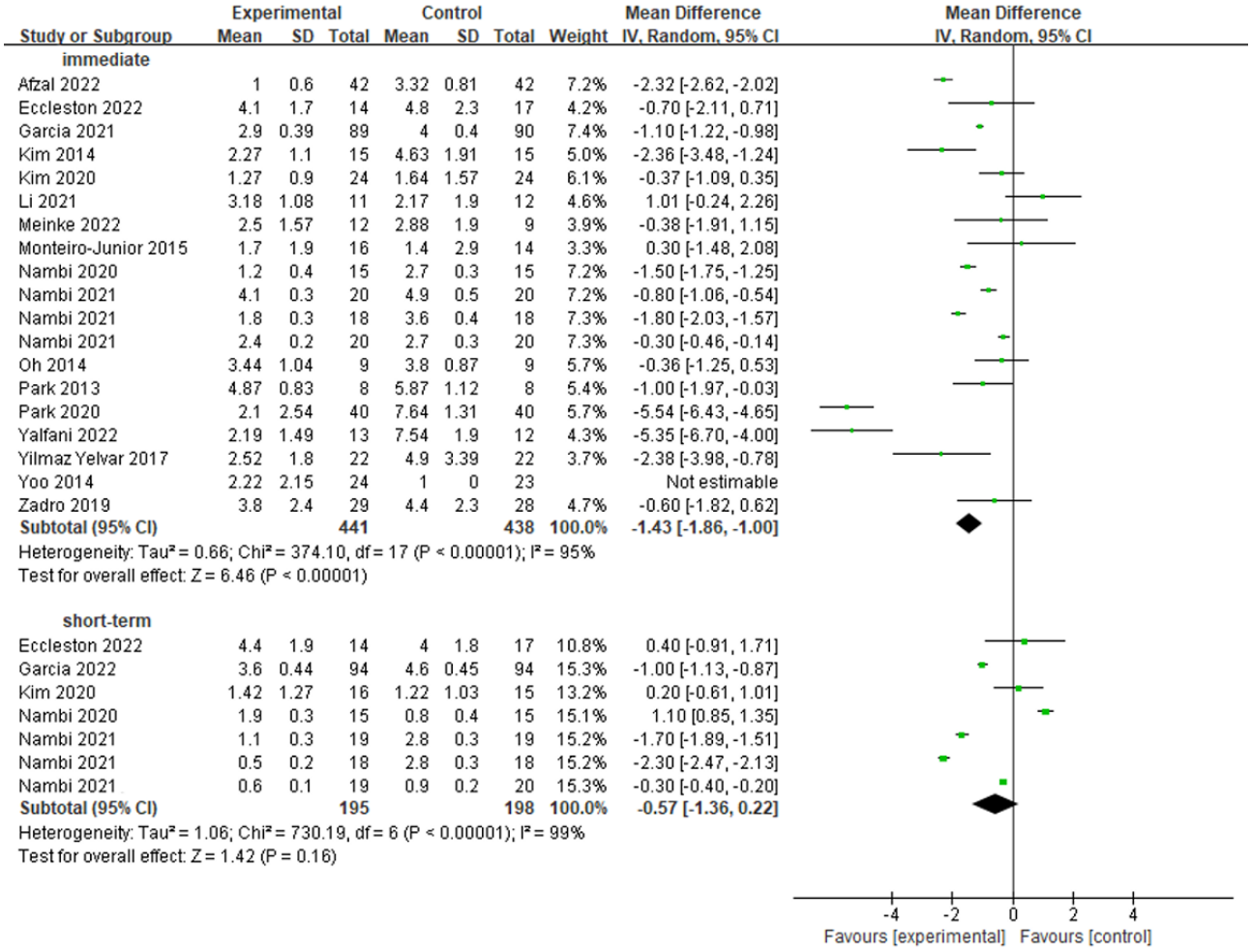

Pain-Related Fear

In total, 6 studies investigated the efficacy of virtual reality–based training on pain-related fear using the Tampa Scale of Kinesiophobia (Figure 4 [50,52,53,63-65]). Virtual reality–based training showed significant improvements in individuals with chronic low back pain in the immediate term (MD –5.46; 95% CI –9.40 to –1.52; I^2=90%; P=.007), while it did not show statistically significant differences in the short term (MD –5.66; 95% CI –12.34 to 1.01; I^2=96%; P=.10). The difference in the immediate- and short-term performance was extremely close to an MCID at the level of 5.5 [68].

XSL•FO
RenderX

Figure 4. Forest plots for virtual reality–based training compared with controls in pain-related fear. IV: inverse variance.

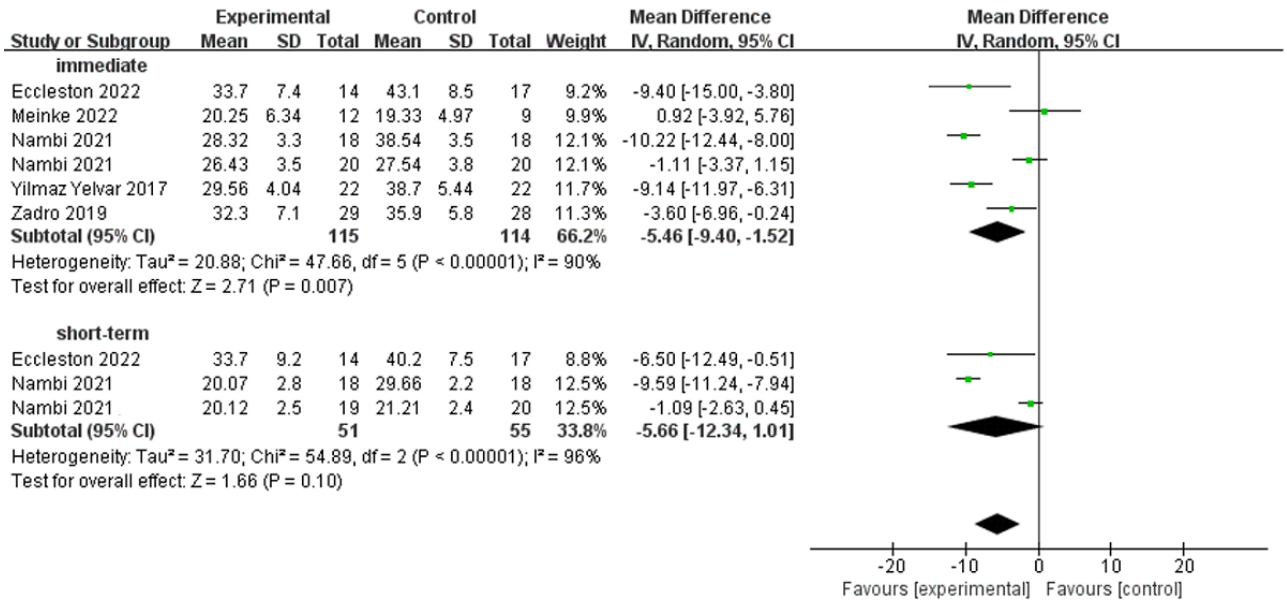

Disability

In total, 8 studies investigated the efficacy of virtual reality–based training on disability using the ODI (Figure 5 [19,20,50,51,58,61,62,64]). There were statistically significant differences between virtual reality–based training and conventional treatments in terms of the ODI in the immediate term (MD –11.50; 95% CI –20.00 to –3.01; I^2=95%; P=.008) but not in the short term (MD –1.28; 95% CI –4.47 to 1.90; I^2=0%; P=.43). Only the immediate-term outcome did reach an MCID at the level of 10 [66,69].

Figure 5. Forest plots for virtual reality–based training compared with controls in disability. IV: inverse variance.

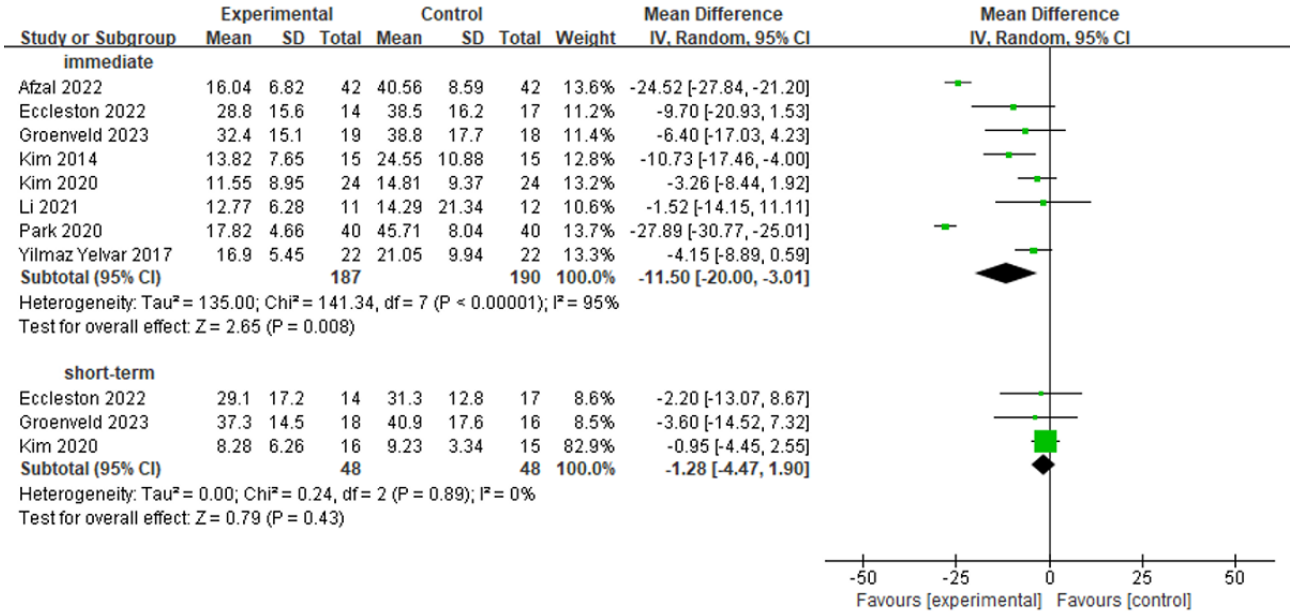

Quality of Evidence

The quality of evidence ranged from low to very low due to the risk of bias, imprecision, and inconsistency of included trials (Table 2). None were downgraded because of indirectness and publication bias.

XSL•FO
RenderX

Table 2. Summary of the findings and level of certainty using GRADE (Grading of Recommendations, Assessment, Development and Evaluation).

Time point and outcomes	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Participants, n	RCT[a], n	Mean difference (95% CI)	Certainty
Immediate term									
Pain	Serious[b]	Serious[c]	Not serious	Not serious	Not serious	879	19	–1.43 (–1.86 to –1.00)	Low
Pain-related fear	Serious[b]	Serious[c]	Not serious	Serious[d]	Not serious	229	6	–5.46 (–9.40 to –1.52)	Very low
Disability	Serious[b]	Serious[c]	Not serious	Serious[d]	Not serious	377	8	–11.50 (–20.00 to –3.01)	Very low
Short term									
Pain-related fear	Serious[b]	Serious[c]	Not serious	Serious[d]	Not serious	106	3	–5.66 (–12.34 to 1.01)	Very low
Pain	Serious[b]	Serious[c]	Not serious	Serious[d]	Not serious	393	7	–0.57 (–1.36 to 0.22)	Very low
Disability	Serious[b]	Not serious	Not serious	Serious[d]	Not serious	96	3	–1.28 (–4.47 to 1.90)	Low

[a]RCT: randomized controlled trial.

[b]Downgrade due to the majority of trials rated as some concerns.

[c]Downgrade due to I^2 statistics >50%.

[d]Downgrade due to pooled sample sizes <400.

Discussion

Principal Findings

This systematic review and meta-analysis, based on 20 RCTs, aimed to assess the available evidence on the efficacy of virtual reality–based training for people with chronic low back pain. The findings indicated that virtual reality–based training appeared to be an effective method of improving pain, pain-related fear, and disability immediately after intervention. However, current evidence failed to support the effectiveness of virtual reality–based training in the short term.

Virtual reality–based training has demonstrated efficacy in pain reduction for people with chronic low back pain [30,70]. However, our pooled results showed only a statistically significant reduction in pain immediately after interventions but not in the short term, which is consistent with previous studies [71-73]. In this study, there were no clinically significant differences between the immediate- and short-term performance. This may be explained by the fact that the virtual reality environment contributes to distraction from pain-related information [74,75]. However, once they gradually return to routines, the attention bias to pain might develop more often than being immersed in a virtual reality environment. This is supported by the finding that patients with chronic pain may less easily distract from the pain [76,77] due to higher levels of attention bias to pain [78].

Pain-related fear can activate the avoidance of movement, leading to the progression of disability [1]. In this study, we found that virtual reality–based training was superior to improve pain-related fear immediately after interventions, which is consistent with the effect on disability for individuals with chronic low back pain. It is worth noting that fear avoidance limits the opportunity to attune expectations to actual experiences, leading to long-term disability [79]. Consequently, with the alleviation of pain-related fear, individuals with chronic low back pain may be willing to pursue a positive experience, which is helpful to increase the level of physical activity. Although the large difference in the efficacy on pain-related fear in the short term indicated clinically important results but not statistically significant results, it should be interpreted cautiously due to the very low quality of evidence. Furthermore, the efficacy of virtual reality–based training on disability in the short term is also with low evidence. More high-quality RCTs are urgently needed for future research.

Strengths and Limitations

This systematic review and meta-analysis conducted a thorough screening and search strategy in 6 important databases. Additionally, this study focused on RCTs to reinforce the evidence of pooled results. We also used the Cochrane Risk of Bias tool and the GRADE approach to rate the quality of the included studies and evidence, respectively. However, the potential limitations in this study need to be mentioned. First, high heterogeneity was presented in this meta-analysis due to the variety of sample size, intervention items, frequency,

XSL•FO
RenderX

duration, and control items. Second, the study population varied from each other, especially age and gender. Third, there were only data on short-term outcomes (3 to 6 months after interventions), and no available data on the mid- or long-term (more than 6 months) effects of virtual reality–based training for participants with chronic low back pain. These findings indicated that more studies are required to strengthen the evidence.

Implications for Clinical Practice and Future Research

Despite the low level of certainty, these findings may provide important implications for health care professionals. In clinical practice, the virtual reality–based training may be recommended for patients with chronic low back pain in order to distract and reduce the focus on pain. However, the persistence of efficacy needs to be considered when applying virtual reality–based training. Based on the present evidence in this study, it may be even better for immediate effects. Furthermore, the virtual reality–based training should be tailored to the needs and characteristics of patients in order to optimize the performance and maintain the efficacy. Future research should focus on different types of virtual reality, intervention parameters (eg, types, frequency, intensity, and time), and different population. These issues require further refinement through larger sample sizes, longer intervention duration, and longer follow-ups. In addition, it is urgent to explore the feasibility of virtual reality–based training in different socioeconomic contexts to ensure broader and cost-effective access to this type of intervention.

Conclusions

In general, these findings support that virtual reality–based training is a promising treatment strategy for individuals with chronic low back pain immediately after interventions, especially to alleviate pain, pain-related fear, and disability. However, there is still not sufficient evidence to suggest that virtual reality–based training is effective in chronic low back pain in the short term. More high-quality RCTs are required to find short- and long-term benefits and to obtain more robust evidence.

Acknowledgments

All authors thank the authors who conducted the clinical trials included in our systematic review for their efforts. This work was supported by National Key R&D Program of China (grants 2020YFC2008500 and 2020YFC2008502); National Natural Science Foundation of China (grants 81572231 and 82172534); and 1·3·5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (grant ZYJC21038).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RL and QW contributed to the conception and design of the study. RL, YL, YK, and HL completed the search, study selection, and data extraction. RL, DH, and CF analyzed the data and drafted the paper. All authors have approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 41 KB - jmir_v26i1e45406_app1.docx]

Multimedia Appendix 2
Search strategy for PubMed, Embase, CINAHL, CENTRAL, and Web of Science.
[DOCX File , 23 KB - jmir_v26i1e45406_app2.docx]

References

1. Gatchel RJ, Peng YB, Peters ML, Fuchs PN, Turk DC. The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychol Bull 2007 Jul;133(4):581-624. [doi: 10.1037/0033-2909.133.4.581] [Medline: 17592957]
2. Mescouto K, Olson RE, Hodges PW, Setchell J. A critical review of the biopsychosocial model of low back pain care: time for a new approach? Disabil Rehabil 2022 Jun;44(13):3270-3284 [FREE Full text] [doi: 10.1080/09638288.2020.1851783] [Medline: 33284644]
3. Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJEM, Ostelo RWJG, Guzman J, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: Cochrane systematic review and meta-analysis. BMJ 2015 Feb 18;350:h444 [FREE Full text] [doi: 10.1136/bmj.h444] [Medline: 25694111]

XSL•FO
RenderX

4. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
5. Chen S, Chen M, Wu X, Lin S, Tao C, Cao H, et al. Global, regional and national burden of low back pain 1990-2019: a systematic analysis of the Global Burden of Disease study 2019. J Orthop Translat 2022 Jan;32:49-58 [FREE Full text] [doi: 10.1016/j.jot.2021.07.005] [Medline: 34934626]
6. Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, et al. What low back pain is and why we need to pay attention. Lancet 2018 Jun 09;391(10137):2356-2367 [FREE Full text] [doi: 10.1016/S0140-6736(18)30480-X] [Medline: 29573870]
7. Wertli MM, Eugster R, Held U, Steurer J, Kofmehl R, Weiser S. Catastrophizing—a prognostic factor for outcome in patients with low back pain: a systematic review. Spine J 2014 Nov 01;14(11):2639-2657. [doi: 10.1016/j.spinee.2014.03.003] [Medline: 24607845]
8. Synnott A, O'Keeffe M, Bunzli S, Dankaerts W, O'Sullivan P, O'Sullivan K. Physiotherapists may stigmatise or feel unprepared to treat people with low back pain and psychosocial factors that influence recovery: a systematic review. J Physiother 2015 Apr;61(2):68-76 [FREE Full text] [doi: 10.1016/j.jphys.2015.02.016] [Medline: 25812929]
9. O'Keeffe M, George SZ, O'Sullivan PB, O'Sullivan K. Psychosocial factors in low back pain: letting go of our misconceptions can help management. Br J Sports Med 2019 Jul;53(13):793-794. [doi: 10.1136/bjsports-2018-099816] [Medline: 30154205]
10. Buchbinder R, van Tulder M, Öberg B, Costa LM, Woolf A, Schoene M, et al. Low back pain: a call for action. Lancet 2018 Jun 09;391(10137):2384-2388 [FREE Full text] [doi: 10.1016/S0140-6736(18)30488-4] [Medline: 29573871]
11. Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, et al. Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet 2018 Jun 09;391(10137):2368-2383. [doi: 10.1016/S0140-6736(18)30489-6] [Medline: 29573872]
12. Bernstein IA, Malik Q, Carville S, Ward S. Low back pain and sciatica: summary of NICE guidance. BMJ 2017 Jan 06;356:i6748. [doi: 10.1136/bmj.i6748] [Medline: 28062522]
13. Chou R, Deyo R, Friedly J, Skelly A, Hashimoto R, Weimer M, et al. Nonpharmacologic therapies for low back pain: a systematic review for an American College of Physicians Clinical Practice Guideline. Ann Intern Med 2017 Apr 04;166(7):493-505 [FREE Full text] [doi: 10.7326/M16-2459] [Medline: 28192793]
14. Miyamoto GC, Franco KFM, van Dongen JM, Franco YRDS, de Oliveira NTB, Amaral DDV, et al. Different doses of Pilates-based exercise therapy for chronic low back pain: a randomised controlled trial with economic evaluation. Br J Sports Med 2018 Jul;52(13):859-868. [doi: 10.1136/bjsports-2017-098825] [Medline: 29525763]
15. Qaseem A, Wilt TJ, McLean RM, Forciea MA. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2017 Apr 04;166(7):514-530 [FREE Full text] [doi: 10.7326/M16-2367] [Medline: 28192789]
16. Saper RB, Lemaster C, Delitto A, Sherman KJ, Herman PM, Sadikova E, et al. Yoga, physical therapy, or education for chronic low back pain. Ann Intern Med 2017 Jun 20;167(2):85-94 [FREE Full text] [doi: 10.7326/M16-2579] [Medline: 28631003]
17. Hayden JA, Ellis J, Ogilvie R, Malmivaara A, van Tulder MW. Exercise therapy for chronic low back pain. Cochrane Database Syst Rev 2021 Sep 28;9(9):CD009790 [FREE Full text] [doi: 10.1002/14651858.CD009790.pub2] [Medline: 34580864]
18. Powell KE, Paluch AE, Blair SN. Physical activity for health: what kind? how much? how intense? on top of what? Annu Rev Public Health 2011;32:349-365. [doi: 10.1146/annurev-publhealth-031210-101151] [Medline: 21128761]
19. Afzal MW, Ahmad A, Mohseni Bandpei MA, Gilani SA, Hanif A, Waqas MS. Effects of virtual reality exercises and routine physical therapy on pain intensity and functional disability in patients with chronic low back pain. J Pak Med Assoc 2022 Mar;72(3):413-417 [FREE Full text] [doi: 10.47391/JPMA.3424] [Medline: 35320216]
20. Kim SS, Min WK, Kim JH, Lee BH. The effects of VR-based Wii Fit Yoga on physical function in middle-aged female LBP patients. J Phys Ther Sci 2014 Apr;26(4):549-552 [FREE Full text] [doi: 10.1589/jpts.26.549] [Medline: 24764631]
21. Park JH, Lee SH, Ko DS. The effects of the Nintendo Wii exercise program on chronic work-related low back pain in industrial workers. J Phys Ther Sci 2013 Aug;25(8):985-988 [FREE Full text] [doi: 10.1589/jpts.25.985] [Medline: 24259899]
22. Yalfani A, Abedi M, Raeisi Z. Effects of an 8-week virtual reality training program on pain, fall risk, and quality of life in elderly women with chronic low back pain: double-blind randomized clinical trial. Games Health J 2022 Apr;11(2):85-92. [doi: 10.1089/g4h.2021.0175] [Medline: 35290742]
23. Mirelman A, Maidan I, Deutsch JE. Virtual reality and motor imagery: promising tools for assessment and therapy in Parkinson's disease. Mov Disord 2013 Sep 15;28(11):1597-1608. [doi: 10.1002/mds.25670] [Medline: 24132848]
24. Wiederhold BK, Soomro A, Riva G, Wiederhold MD. Future directions: advances and implications of virtual environments designed for pain management. Cyberpsychol Behav Soc Netw 2014 Jun;17(6):414-422 [FREE Full text] [doi: 10.1089/cyber.2014.0197] [Medline: 24892206]
25. Matheve T, Bogaerts K, Timmermans A. Virtual reality distraction induces hypoalgesia in patients with chronic low back pain: a randomized controlled trial. J Neuroeng Rehabil 2020 Apr 22;17(1):55 [FREE Full text] [doi: 10.1186/s12984-020-00688-0] [Medline: 32321516]

XSL•FO
RenderX

26. Hoffman HG, Richards TL, Van Oostrom T, Coda BA, Jensen MP, Blough DK, et al. The analgesic effects of opioids and immersive virtual reality distraction: evidence from subjective and functional brain imaging assessments. Anesth Analg 2007 Dec;105(6):1776-1783 [FREE Full text] [doi: 10.1213/01.ane.0000270205.45146.db] [Medline: 18042882]
27. Kowatsch T, Lohse KM, Erb V, Schittenhelm L, Galliker H, Lehner R, et al. Hybrid ubiquitous coaching with a novel combination of mobile and holographic conversational agents targeting adherence to home exercises: four design and evaluation studies. J Med Internet Res 2021 Feb 22;23(2):e23612 [FREE Full text] [doi: 10.2196/23612] [Medline: 33461957]
28. Thomas JS, France CR, Applegate ME, Leitkam ST, Walkowski S. Feasibility and safety of a virtual reality dodgeball intervention for chronic low back pain: a randomized clinical trial. J Pain 2016 Dec;17(12):1302-1317 [FREE Full text] [doi: 10.1016/j.jpain.2016.08.011] [Medline: 27616607]
29. Georgescu R, Fodor LA, Dobrean A, Cristea IA. Psychological interventions using virtual reality for pain associated with medical procedures: a systematic review and meta-analysis. Psychol Med 2020 Aug;50(11):1795-1807. [doi: 10.1017/S0033291719001855] [Medline: 31456530]
30. Bordeleau M, Stamenkovic A, Tardif PA, Thomas J. The use of virtual reality in back pain rehabilitation: a systematic review and meta-analysis. J Pain 2022 Feb;23(2):175-195 [FREE Full text] [doi: 10.1016/j.jpain.2021.08.001] [Medline: 34425250]
31. Brea-Gómez B, Torres-Sánchez I, Ortiz-Rubio A, Calvache-Mateo A, Cabrera-Martos I, López-López L, et al. Virtual reality in the treatment of adults with chronic low back pain: a systematic review and meta-analysis of randomized clinical trials. Int J Environ Res Public Health 2021 Nov 11;18(22):11806 [FREE Full text] [doi: 10.3390/ijerph182211806] [Medline: 34831562]
32. Grassini S. Virtual reality assisted non-pharmacological treatments in chronic pain management: a systematic review and quantitative meta-analysis. Int J Environ Res Public Health 2022 Mar 29;19(7):4071 [FREE Full text] [doi: 10.3390/ijerph19074071] [Medline: 35409751]
33. Gumaa M, Rehan Youssef A. Is virtual reality effective in orthopedic rehabilitation? a systematic review and meta-analysis. Phys Ther 2019 Oct 28;99(10):1304-1325 [FREE Full text] [doi: 10.1093/ptj/pzz093] [Medline: 31343702]
34. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
35. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
36. Boonstra AM, Schiphorst Preuper HR, Reneman MF, Posthumus JB, Stewart RE. Reliability and validity of the visual analogue scale for disability in patients with chronic musculoskeletal pain. Int J Rehabil Res 2008 Jun;31(2):165-169. [doi: 10.1097/MRR.0b013e3282fc0f93] [Medline: 18467932]
37. Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine 2011 Oct 01;36(Suppl 21):S54-S68 [FREE Full text] [doi: 10.1097/BRS.0b013e31822ef74d] [Medline: 21952190]
38. Nassif TH, Hull A, Holliday SB, Sullivan P, Sandbrink F. Concurrent validity of the defense and veterans pain rating scale in VA outpatients. Pain Med 2015 Nov;16(11):2152-2161 [FREE Full text] [doi: 10.1111/pme.12866] [Medline: 26257151]
39. Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. J Clin Nurs 2005 Aug;14(7):798-804 [FREE Full text] [doi: 10.1111/j.1365-2702.2005.01121.x] [Medline: 16000093]
40. Lundberg MKE, Styf J, Carlsson SG. A psychometric evaluation of the Tampa Scale for Kinesiophobia—from a physiotherapeutic perspective. Physiother Theory Pract 2009 Jul 10;20(2):121-133. [doi: 10.1080/09593980490453002]
41. Woby SR, Roach NK, Urmston M, Watson PJ. Psychometric properties of the TSK-11: a shortened version of the Tampa Scale for Kinesiophobia. Pain 2005 Sep;117(1-2):137-144. [doi: 10.1016/j.pain.2005.05.029] [Medline: 16055269]
42. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine 2000 Nov 15;25(22):2940-2952; discussion 2952. [doi: 10.1097/00007632-200011150-00017] [Medline: 11074683]
43. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
44. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
45. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. rating the quality of evidence—imprecision. J Clin Epidemiol 2011 Dec;64(12):1283-1293 [FREE Full text] [doi: 10.1016/j.jclinepi.2011.01.012] [Medline: 21839614]
46. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011 Apr;64(4):407-415 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.07.017] [Medline: 21247734]

XSL•FO
RenderX

47. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. rating the quality of evidence—publication bias. J Clin Epidemiol 2011 Dec;64(12):1277-1282 [FREE Full text] [doi: 10.1016/j.jclinepi.2011.01.011] [Medline: 21802904]
48. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. rating the quality of evidence—inconsistency. J Clin Epidemiol 2011 Dec;64(12):1294-1302 [FREE Full text] [doi: 10.1016/j.jclinepi.2011.03.017] [Medline: 21803546]
49. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. rating the quality of evidence—indirectness. J Clin Epidemiol 2011 Dec;64(12):1303-1310 [FREE Full text] [doi: 10.1016/j.jclinepi.2011.04.014] [Medline: 21802903]
50. Eccleston C, Fisher E, Liikkanen S, Sarapohja T, Stenfors C, Jääskeläinen SK, et al. A prospective, double-blind, pilot, randomized, controlled trial of an "embodied" virtual reality intervention for adults with low back pain. Pain 2022 Sep 01;163(9):1700-1715 [FREE Full text] [doi: 10.1097/j.pain.0000000000002617] [Medline: 35324507]
51. Li Z, Yu Q, Luo H, Liang W, Li X, Ge L, et al. The effect of virtual reality training on anticipatory postural adjustments in patients with chronic nonspecific low back pain: a preliminary study. Neural Plast 2021;2021:9975862 [FREE Full text] [doi: 10.1155/2021/9975862] [Medline: 34367274]
52. Nambi G, Abdelbasset WK, Alrawaili SM, Alsubaie SF, Abodonya AM, Saleh AK. Virtual reality or isokinetic training; its effect on pain, kinesiophobia and serum stress hormones in chronic low back pain: a randomized controlled trial. Technol Health Care 2021;29(1):155-166. [doi: 10.3233/THC-202301] [Medline: 32831210]
53. Nambi G, Abdelbasset WK, Alsubaie SF, Saleh AK, Verma A, Abdelaziz MA, et al. Short-term psychological and hormonal effects of virtual reality training on chronic low back pain in soccer players. J Sport Rehabil 2021 Feb 16;30(6):884-893. [doi: 10.1123/jsr.2020-0075] [Medline: 33596538]
54. Nambi G, Abdelbasset WK, Elsayed SH, Alrawaili SM, Abodonya AM, Saleh AK, et al. Comparative effects of isokinetic training and virtual reality training on sports performances in university football players with chronic low back pain-randomized controlled study. Evid Based Complement Alternat Med 2020;2020:2981273 [FREE Full text] [doi: 10.1155/2020/2981273] [Medline: 32617104]
55. Nambi G, Abdelbasset WK, Elsayed SH, Verma A, George JS, Saleh AK. Eficiência clínica e física de jogos de realidade virtual em jogadores de futebol com dor lombar. Clinical and physical efficiency of virtual reality games in soccer players with low back pain. Rev Bras Med Esporte 2021;27(6):597 [FREE Full text] [doi: 10.1590/1517-86922021270620210034]
56. Oh HW, Lee MG, Jang JY, Jin JJ, Cha JY, Jin YY, et al. Time-effects of horse simulator exercise on psychophysiological responses in men with chronic low back pain. IES 2014 Jun 01;22(2):153-163. [doi: 10.3233/ies-140533]
57. Monteiro-Junior RS, de Souza CP, Lattari E, Rocha NBF, Mura G, Machado S, et al. Wii-workouts on chronic pain, physical capabilities and mood of older women: a randomized controlled double blind trial. CNS Neurol Disord Drug Targets 2015;14(9):1157-1164. [doi: 10.2174/1871527315666151111120131] [Medline: 26556092]
58. Park S, Park S, Min S, Kim CJ, Jee YS. A randomized controlled trial investigating the effects of equine simulator riding on low back pain, morphological changes, and trunk musculature in elderly women. Medicina (Kaunas) 2020 Nov 13;56(11):610 [FREE Full text] [doi: 10.3390/medicina56110610] [Medline: 33202928]
59. Yoo JH, Kim SE, Lee MG, Jin JJ, Hong J, Choi YT, et al. The effect of horse simulator riding on visual analogue scale, body composition and trunk strength in the patients with chronic low back pain. Int J Clin Pract 2014 Aug;68(8):941-949 [FREE Full text] [doi: 10.1111/ijcp.12414] [Medline: 25039929]
60. Garcia LM, Birckhead BJ, Krishnamurthy P, Sackman J, Mackey IG, Louis RG, et al. An 8-week self-administered at-home behavioral skills-based virtual reality program for chronic low back pain: double-blind, randomized, placebo-controlled trial conducted during COVID-19. J Med Internet Res 2021 Feb 22;23(2):e26292 [FREE Full text] [doi: 10.2196/26292] [Medline: 33484240]
61. Groenveld TD, Smits MLM, Knoop J, Kallewaard JW, Staal JB, de Vries M, et al. Effect of a behavioral therapy-based virtual reality application on quality of life in chronic low back pain. Clin J Pain 2023 Jun 01;39(6):278-285 [FREE Full text] [doi: 10.1097/AJP.0000000000001110] [Medline: 37002877]
62. Kim T, Lee J, Oh S, Kim S, Yoon B. Effectiveness of simulated horseback riding for patients with chronic low back pain: a randomized controlled trial. J Sport Rehabil 2020 Feb 01;29(2):179-185. [doi: 10.1123/jsr.2018-0252] [Medline: 30676224]
63. Meinke A, Peters R, Knols RH, Swanenburg J, Karlen W. Feedback on trunk movements from an electronic game to improve postural balance in people with nonspecific low back pain: pilot randomized controlled trial. JMIR Serious Games 2022 Jun 10;10(2):e31685 [FREE Full text] [doi: 10.2196/31685] [Medline: 35687390]
64. Yilmaz Yelvar GD, Çırak Y, Dalkılınç M, Parlak Demir Y, Guner Z, Boydak A. Is physiotherapy integrated virtual walking effective on pain, function, and kinesiophobia in patients with non-specific low-back pain? randomised controlled trial. Eur Spine J 2017 Feb;26(2):538-545. [doi: 10.1007/s00586-016-4892-7] [Medline: 27981455]
65. Zadro JR, Shirley D, Simic M, Mousavi SJ, Ceprnja D, Maka K, et al. Video-game-based exercises for older people with chronic low back pain: a randomized controlledtable trial (GAMEBACK). Phys Ther 2019 Jan 01;99(1):14-27 [FREE Full text] [doi: 10.1093/ptj/pzy112] [Medline: 30247715]
66. Ostelo RWJG, de Vet HCW. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 2005 Aug;19(4):593-607. [doi: 10.1016/j.berh.2005.03.003] [Medline: 15949778]

XSL•FO
RenderX

67. van der Roer N, Ostelo RWJG, Bekkering GE, van Tulder MW, de Vet HCW. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine 2006 Mar 01;31(5):578-582. [doi: 10.1097/01.brs.0000201293.57439.47] [Medline: 16508555]
68. Monticone M, Ambrosini E, Rocca B, Foti C, Ferrante S. Responsiveness of the Tampa Scale of Kinesiophobia in Italian subjects with chronic low back pain undergoing motor and cognitive rehabilitation. Eur Spine J 2016 Sep;25(9):2882-2888. [doi: 10.1007/s00586-016-4682-2] [Medline: 27356516]
69. Ostelo RWJG, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine 2008 Jan 01;33(1):90-94. [doi: 10.1097/BRS.0b013e31815e3a10] [Medline: 18165753]
70. Nagpal AS, Raghunandan A, Tata F, Kibler D, McGeary D. Virtual reality in the management of chronic low back pain: a scoping review. Front Pain Res (Lausanne) 2022;3:856935 [FREE Full text] [doi: 10.3389/fpain.2022.856935] [Medline: 35295809]
71. House G, Burdea G, Grampurohit N, Polistico K, Roll D, Damiani F, et al. A feasibility study to determine the benefits of upper extremity virtual rehabilitation therapy for coping with chronic pain post-cancer surgery. Br J Pain 2016 Nov;10(4):186-197 [FREE Full text] [doi: 10.1177/2049463716664370] [Medline: 27867508]
72. Sato K, Fukumori S, Matsusaki T, Maruo T, Ishikawa S, Nishie H, et al. Nonimmersive virtual reality mirror visual feedback therapy and its application for the treatment of complex regional pain syndrome: an open-label pilot study. Pain Med 2010 Apr;11(4):622-629 [FREE Full text] [doi: 10.1111/j.1526-4637.2010.00819.x] [Medline: 20202141]
73. Ortiz-Catalan M, Guðmundsdóttir RA, Kristoffersen MB, Zepeda-Echavarria A, Caine-Winterberger K, Kulbacka-Ortiz K, et al. Phantom motor execution facilitated by machine learning and augmented reality as treatment for phantom limb pain: a single group, clinical trial in patients with chronic intractable phantom limb pain. Lancet 2016 Dec 10;388(10062):2885-2894. [doi: 10.1016/S0140-6736(16)31598-7] [Medline: 27916234]
74. Kohl A, Rief W, Glombiewski JA. Acceptance, cognitive restructuring, and distraction as coping strategies for acute pain. J Pain 2013 Mar;14(3):305-315 [FREE Full text] [doi: 10.1016/j.jpain.2012.12.005] [Medline: 23352770]
75. Verhoeven K, Crombez G, Eccleston C, Van Ryckeghem DML, Morley S, Van Damme S. The role of motivation in distracting attention away from pain: an experimental study. Pain 2010 May;149(2):229-234 [FREE Full text] [doi: 10.1016/j.pain.2010.01.019] [Medline: 20188469]
76. Van Ryckeghem DM, Van Damme S, Eccleston C, Crombez G. The efficacy of attentional distraction and sensory monitoring in chronic pain patients: a meta-analysis. Clin Psychol Rev 2018 Feb;59:16-29. [doi: 10.1016/j.cpr.2017.10.008] [Medline: 29126746]
77. Vlaeyen JWS, Morley S, Crombez G. The experimental analysis of the interruptive, interfering, and identity-distorting effects of chronic pain. Behav Res Ther 2016 Nov;86:23-34 [FREE Full text] [doi: 10.1016/j.brat.2016.08.016] [Medline: 27614948]
78. Todd J, van Ryckeghem DML, Sharpe L, Crombez G. Attentional bias to pain-related information: a meta-analysis of dot-probe studies. Health Psychol Rev 2018 Dec;12(4):419-436. [doi: 10.1080/17437199.2018.1521729] [Medline: 30205757]
79. Crombez G, Eccleston C, Van Damme S, Vlaeyen JWS, Karoly P. Fear-avoidance model of chronic pain: the next generation. Clin J Pain 2012 Jul;28(6):475-483. [doi: 10.1097/AJP.0b013e3182385392] [Medline: 22673479]

Abbreviations

GRADE: Grading of Recommendations, Assessment, Development and Evaluation
MCID: minimal clinically important difference
MD: mean difference
ODI: Oswestry Disability Index
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by A Mavragani; submitted 29.12.22; peer-reviewed by A Ollevier, MR Underwood; comments to author 21.03.23; revised version received 16.04.23; accepted 29.01.24; published 26.02.24.

Please cite as:
Li R, Li Y, Kong Y, Li H, Hu D, Fu C, Wei Q
Virtual Reality–Based Training in Chronic Low Back Pain: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2024;26:e45406
URL: https://www.jmir.org/2024/1/e45406
doi:10.2196/45406
PMID:38407948

XSL•FO
RenderX

©Ran Li, Yinghao Li, Youli Kong, Hanbin Li, Danrong Hu, Chenying Fu, Quan Wei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Chatbots That Deliver Contraceptive Support: Systematic Review

Rhiana Mills[1*], BSc, MSc; Emily Rose Mangone[2*], MSc, PhD; Neal Lesh[3*], MPH, PhD; Gayatri Jayal[3*], BA; Diwakar Mohan[4*], MBBS, MPH, MD, DrPH; Paula Baraitser[1*], MBBS, MD

[1]SH24, London, United Kingdom
[2]The Bill & Melinda Gates Foundation, Seattle, WA, United States
[3]Dimagi, Cambridge, MA, United States
[4]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[*]all authors contributed equally

Corresponding Author:
Rhiana Mills, BSc, MSc
SH24
35A Westminster Bridge Road
London, SE1 7JB
United Kingdom
Phone: 44 7742932445
Email: rhiana@sh24.org.uk

Abstract

Background: A chatbot is a computer program that is designed to simulate conversation with humans. Chatbots may offer rapid, responsive, and private contraceptive information; counseling; and linkages to products and services, which could improve contraceptive knowledge, attitudes, and behaviors.

Objective: This review aimed to systematically collate and interpret evidence to determine whether and how chatbots improve contraceptive knowledge, attitudes, and behaviors. Contraceptive knowledge, attitudes, and behaviors include access to contraceptive information, understanding of contraceptive information, access to contraceptive services, contraceptive uptake, contraceptive continuation, and contraceptive communication or negotiation skills. A secondary aim of the review is to identify and summarize best practice recommendations for chatbot development to improve contraceptive outcomes, including the cost-effectiveness of chatbots where evidence is available.

Methods: We systematically searched peer-reviewed and gray literature (2010-2022) for papers that evaluated chatbots offering contraceptive information and services. Sources were included if they featured a chatbot and addressed an element of contraception, for example, uptake of hormonal contraceptives. Literature was assessed for methodological quality using appropriate quality assessment tools. Data were extracted from the included sources using a data extraction framework. A narrative synthesis approach was used to collate qualitative evidence as quantitative evidence was too sparse for a quantitative synthesis to be carried out.

Results: We identified 15 sources, including 8 original research papers and 7 gray literature papers. These sources included 16 unique chatbots. This review found the following evidence on the impact and efficacy of chatbots: a large, robust randomized controlled trial suggests that chatbots have no effect on intention to use contraception; a small, uncontrolled cohort study suggests increased uptake of contraception among adolescent girls; and a development report, using poor-quality methods, suggests no impact on improved access to services. There is also poor-quality evidence to suggest increased contraceptive knowledge from interacting with chatbot content. User engagement was mixed, with some chatbots reaching wide audiences and others reaching very small audiences. User feedback suggests that chatbots may be experienced as acceptable, convenient, anonymous, and private, but also as incompetent, inconvenient, and unsympathetic. The best practice guidance on the development of chatbots to improve contraceptive knowledge, attitudes, and behaviors is consistent with that in the literature on chatbots in other health care fields.

Conclusions: We found limited and conflicting evidence on chatbots to improve contraceptive knowledge, attitudes, and behaviors. Further research that examines the impact of chatbot interventions in comparison with alternative technologies, acknowledges the varied and changing nature of chatbot interventions, and seeks to identify key features associated with improved contraceptive outcomes is needed. The limitations of this review include the limited evidence available on this topic, the lack of formal evaluation of chatbots in this field, and the lack of standardized definition of what a chatbot is.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e46758) doi:10.2196/46758

KEYWORDS

chatbot; contraceptives; digital health; AI; systematic review; conversational agent; development best practices; development; counseling; communication; user feedback; users; feedback; attitudes; behavior

Introduction

Recent guidelines have identified digital technologies as "promising" for improving knowledge; influencing attitudes, beliefs, and expectations; and increasing self-efficacy in support of healthy sexual and reproductive health (SRH) behaviors [1-3]. Although these guidelines consider digital interventions in general, there is a growing interest in building a stronger evidence base on how SRH programs might leverage key technologies to improve their reach, engagement, and impact. One technology that has seen a substantial rise in use by SRH programs is chatbots [4,5].

A chatbot is a computer program that is designed to simulate conversation with humans [6]. Chatbots may use rule-based language applications or more advanced natural language processing (NLP) and artificial intelligence technology to automate conversations [7]. In the context of contraception, chatbots have the potential to provide rapid and responsive information, counseling, and linkages to products and services [8]. They may also serve as navigational agents and companions on a reproductive health journey or be gamified and serve as "edutainment" [9].

There is a small but expanding literature within health care on chatbot design, user experience, and the outcomes of chatbot use. For example, within mental health care, there have been several recent reviews on chatbot design [10,11], chatbots that enable service delivery [12], and the ability of chatbots to build relationships with humans [6,13]. There has been no attempt, to date, to consolidate and synthesize the literature on chatbots to improve contraceptive outcomes or to inform best practices for the design and development of chatbots for this purpose.

Methods

Objectives

This review aims to systematically collate and interpret evidence to determine whether and how chatbots improve contraceptive knowledge, attitudes, and behaviors. Contraceptive knowledge, attitudes, and behaviors include access to contraceptive information, understanding of contraceptive information, access to contraceptive services, contraceptive uptake, contraceptive continuation, and contraceptive communication or negotiation skills. A secondary aim of the review is to identify and summarize best practice recommendations for chatbot development to improve contraceptive outcomes including the cost-effectiveness of chatbots, where evidence is available.

To achieve these aims a PICO (Population, Intervention, Control, Outcome) framework was used to identify the components that should be included in the review process (Textbox 1).

Textbox 1. PICO framework for systematic review.

Population

- Users of an existing chatbot in any global context

Intervention

- Chatbots that improve access to contraceptive information, understanding of contraceptive information, access to contraceptive services, contraceptive uptake, contraceptive continuation, and contraceptive communication or negotiation skills

Control

- Any alternative intervention or no intervention

Outcomes

- Improved contraceptive knowledge, attitudes, or behaviors or best practice recommendations for developing chatbots to deliver these outcomes in an effective or cost-effective manner

Search Strategy

We systematically searched and collated peer-reviewed, original research, and gray literature sources published between January 2010 and September 2022, which reported outcome data and data collection methods. We chose to include gray literature because we anticipated very little published evidence available on chatbots that address contraceptive outcomes, as this technology is relatively new.

Searches were carried out in the following databases in October 2022: MEDLINE, Embase, EmCare, PubMed, Science Direct, Cochrane Library, Scopus, and Google Scholar. The search terms detailed in Textbox 2 were used to search the selected databases (see Multimedia Appendix 1). A search for gray literature was completed using the Google search engine using the following search terms: "chat bot" OR chatbot OR "conversational agent" AND contracept* OR "family planning." The first 100 Google search hits were screened.

XSL•FO
RenderX

Sources identified from the database and Google searches were imported into reference management software. Duplicates were removed, and the title and abstract of each source were screened against the inclusion and exclusion criteria detailed in Textbox 3 by RM. After this initial screening, a secondary screening of full-text resources was undertaken, again using the inclusion and exclusion criteria in Textbox 2. RM screened the sources, and PB screened 10% of the sources with any disagreements discussed to check for accuracy and consistency.

Textbox 2. Search terms.

Chatbot terms

- "chat bot*" or chat-bot* or chatbot* or "chatter bot*" or chatterbot* or "talk bot*" or talkbot* or talk-bot* or "interactive agent*" or "conversational agent*" or "artificial conversation* entit*" or "artificial intelligence" or AI or "human computer interaction" or "intelligent agent*" or "chat agent*" or "relational agent*" or "virtual agent*" or "virtual assistant*" or "virtual coach"

Sexual and reproductive health terms

- "sexual and reproductive health" or "reproductive and sexual health" or "sexual health" or "reproductive health" or "sexually transmitted infection*" or STI or STIs or "sexually transmitted disease" or STD or STDs or HIV of "human immunodeficiency virus" or chlamydia or gonorrhea or herpes or "herpes genitalis" or HPV or "human papillomavirus" or syphilis or condom or "cervical cancer" or "cervical screen" or "pap* test" or antenatal or prenatal or postnatal or perinatal or pregnan* or maternal or gynae* or birth or caesarean

Contraceptive terms

- Contracept* or "family planning" or LARC or "long acting reversible contraceptive" or "pill" or COC or POP or "progesterone only pill" or "combined oral contraception" or "inter-uterine device" or IUD or "inter-uterine system" or IUS or coil or "hormonal coil" or "copper coil" or "contracept* implant" or "injectable contracept*" or "self-injectable contracept* or "depo-provera" or "sayana Press" or "contraceptive decision making" or "family planning decision making"

Textbox 3. Inclusion and exclusion criteria.

Inclusion criteria

- The paper must be published between 2010 and September 2022
- The intervention must include an existing chatbot
- The intervention must aim to address an element of contraception—this need not be the main focus, but it must be included

Exclusion criteria

- The paper is not published between 2010 and September 2022
- The intervention does not include an existing chatbot
- The intervention does not aim to address an element of contraception
- The paper is not of reasonable research quality standard

Quality Assessment

Sources were critically appraised for research and reporting quality. The Critical Appraisal Skills Programme (CASP) Cohort Study appraisal tool, the CASP Qualitative Study appraisal tool, the CASP Randomized Control Trial (RCT) Study appraisal tool, and the AACODS (Authority, Accuracy, Coverage, Objectivity, Date, and Significance) checklist for gray literature were used to guide the critical appraisal process [14,15]. CASP tools use criteria to assess how methodologically sound a source is, how accurately and comprehensively results have been reported, and whether ethical considerations have been accounted for. The AACODS tool assesses the rigor and relevance of gray literature sources. PB appraised a randomly selected sample of the peer-reviewed journal articles to check for agreement on quality assessment. The authors RM, ERM, NL, GJ, and PB critically discussed the quality assessment of gray literature papers. A total of 9 sources were excluded because of not meeting quality standards (see Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) 2020 flow diagram for new systematic reviews that included searches of databases, registers, and other sources.

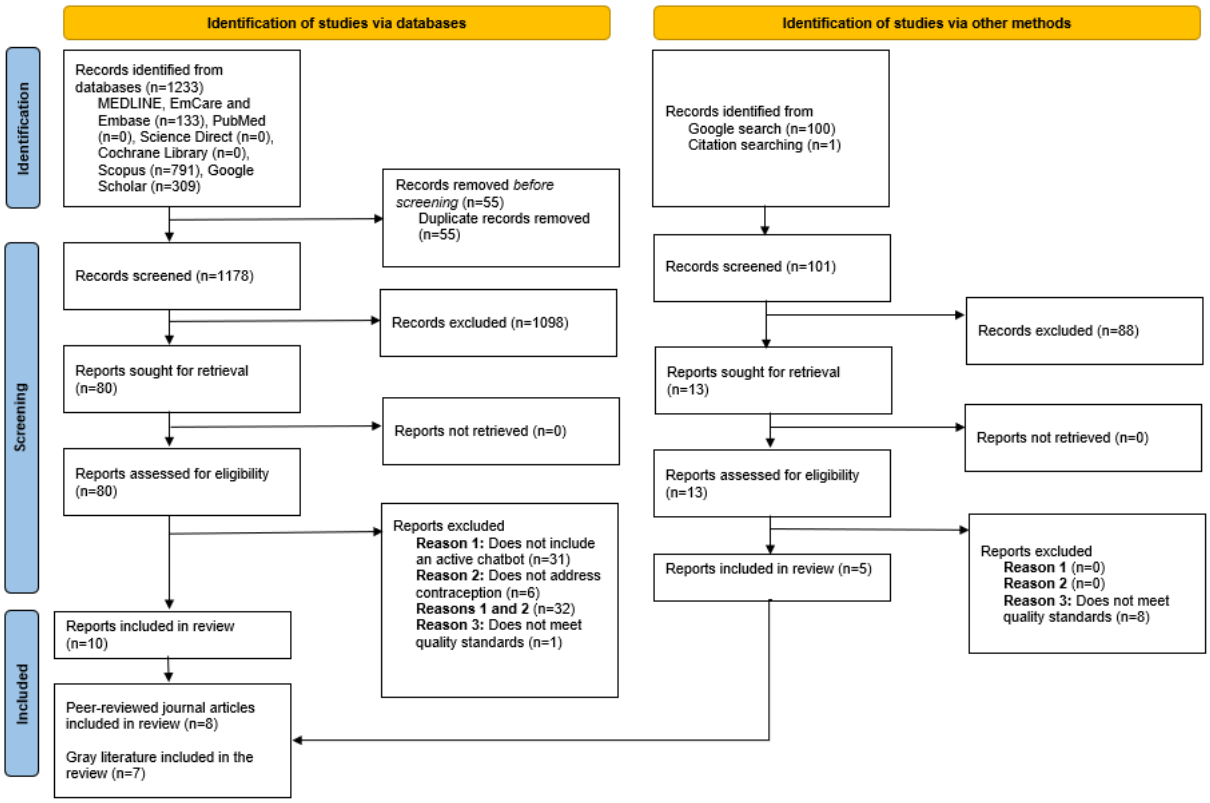

Data Extraction

Data were extracted from the included sources by RM using a standard data extraction template, which allowed for a standard set of variables to be extracted from each source (see Multimedia Appendix 2). We noted that standard variables were not reported in a source. Verbatim sections of text, which described best practice recommendations, were separately extracted from the sources that described the design and development and user feedback on chatbots and added to a spreadsheet. RM and PB regularly discussed data extraction to check for accuracy.

Synthesis

The data extracted from all sources were grouped under data extraction categories: use of theory to underpin chatbot development; efficacy and impact; improving knowledge; user engagement; number of users; length of engagement; demographics of those who engaged; user feedback; acceptability; convenience and accessibility; and anonymity, nonjudgment, and privacy.

Because of insufficient quantitative evidence for meta-analysis, quantitative findings were summarized under appropriate data extraction headings and are reported in summary in the Results section.

A narrative synthesis approach was used to collate and summarize qualitative evidence under each data extraction heading. The data under each heading were read and reread by RM to identify emerging themes and subthemes. Themes were considered emergent when multiple data points could be grouped in a theme or subtheme. RM regularly presented emergent themes and discussed the synthesis process with PB to check for accuracy and collaboratively develop themes. The synthesis process was shared with the authorship team on 2 occasions to check for accuracy.

Thematic analysis was also used to draw out best practices from the extracted data. RM grouped emerging themes that provided pragmatic information regarding best practice for chatbot development. These grouped themes were discussed with the authorship team to check for analytic accuracy and collaboratively label the best practices. Best practice guidance was only included if it was referenced across more than 1 evidence source.

Results

Overview

The PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) diagram (Figure 1) summarizes the results of the search process. Although we used a broader search strategy with regard to date, to identify all literature in this under-researched field, we did not find any literature that predated 2018. Only 15 sources were identified for evaluation, reflecting the nascent development of chatbots in this field.

Quality Assessment

The quality assessment process excluded 8 gray literature sources and 1 peer-reviewed journal article. The 8 gray literature sources were excluded based on the AACODS checklist, which

XSL•FO
RenderX

measures the authority of the source author and the content of the source in terms of accuracy, coverage, objectivity, date, and significance. Three sources were excluded because they did not come from a reputable source and lacked accuracy in the methods they used to evaluate chatbots and reporting of their methods (where methods were reported at all). Five sources were excluded because they lacked accuracy in their methods and reporting and because the content of the source was not considered significant. For example, some sources were website articles that described a chatbot for a lay audience and included very little or no detail on the evaluation of the chatbot or outcomes of its use.

The single peer-reviewed journal article that was excluded had a qualitative study design and was excluded because of unclear research aims and sparse description of data analysis, meaning analytic rigor could not be ascertained.

The Chatbots Identified

This review identified 15 sources, including 8 original research papers and 7 gray literature papers (see Tables 1 and 2). Seven chatbots in the review were implemented in high-income countries: the United Kingdom (n=2), United States (n=4), and Japan (n=1). Eight chatbots were implemented in low- or middle-income countries: Kenya (n=3), South Africa (n=1), Uganda (n=1), Nigeria (n=1), Bangladesh (n=1), and India (n=1). One chatbot had international reach. A total of 4 chatbots aimed specifically to engage the general population, and 6 sought to engage women and girls. Two of these specifically targeted young women: one targeted adolescent girls and one targeted young pregnant women. Three chatbots were aimed at adolescents of any gender. One chatbot was aimed at married couples, and one targeted those at risk of poor sexual health.

XSL•FO
RenderX

Table 1. Table of results.

Study	Date	Title	Source type	Source country	Study design	Chatbot name
Bonnevie et al [16]	2021	Layla's Got You: developing a tailored contraception chatbot for Black and Hispanic young women	Peer-reviewed journal article	United States	Development report	Layla's Got You
Brannock et al [17]	2019	Investigating attitudes and preferences towards a chatbot pregnancy guide within Facebook's social media platform amongst pregnant women in Kenya	Gray literature: thesis	Kenya	Development report	Mama Support
Chernick et al [18]	2021	The perspectives of women and their health-care providers regarding using an ECA to support mode of birth decisions	Peer-reviewed journal article	United States	Qualitative study	Dr Erica
Green et al [19]	2022	Predicting health care-seeking behavior based on stated readiness to act: development and validation of a prediction model.	Peer-reviewed journal article	Kenya	Model development report	Ask Nivi
Handforth and Bertermann [20]	2018	How Girl Effect built a chatbot	Gray literature: technical report	South Africa	Development report	Big Sis
Hussain et al [21]	2019	Mobile phone-based chatbot for family planning and contraceptive information	Gray literature: conference paper	United States	Feasibility study	Name not reported
Woo et al [22]	2020	Development of conversational artificial intelligence for pandemic healthcare query support	Peer-reviewed journal article	United Kingdom	Development report	Akira
Maeda et al [23]	2020	Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial	Peer-reviewed journal article	Japan	RCT[a]	Name not reported
Nadarzynski et al [24]	2021	Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis	Peer-reviewed journal article	United Kingdom	Qualitative interview study	Pat
Population Services International [25]	2019	Will access to sex-positive and reproductive health information through a chatbot lead to increased contraceptive use amongst Kenyan youth?	Website article	Kenya	Development report	Pleasure Bot
Rahman et al [26]	2021	AdolescentBot: Understanding opportunities for chatbots in combating adolescent sexual and reproductive health problems in Bangladesh	Peer-reviewed journal article	Bangladesh	Development report	AdolescentBot
Wang et al [27]	2022	An artificial intelligence chatbot for young people's sexual and reproductive health in India (SnehAI): instrumental case study	Peer-reviewed journal article	India	Instrumental case study	SnehAI
Wilson et al [28]	2017	In bed with Siri and Google Assistant: a comparison of sexual health advice	Gray literature: short report	Not context specific	Short report	Siri and Google Assistant
Winskell et al [29]	2021	Building a chatbot for health content? This is for you.	Website article	Nigeria and Uganda	N/A[b]	Tina
Work & Co [30]	2022	Planned parenthood case study	Website article	United States	Development report	Ask Roo

[a]RCT: randomized controlled trial.

[b]N/A: not applicable.

XSL•FO
RenderX

Table 2. Table of results, continued.

Study	Chatbot aim	Chatbot target group	Chatbot platform	Method of input by user	Chatbot persona	How does the chatbot address contraception
Bonnevie et al [16]	SRH[a] information dissemination	Black and Hispanic women in Onondaga County, New York	Website	Free text input	Female peer	The chatbot includes content on contraception, as part of wider SRH content
Brannock et al [17]	Pregnancy information dissemination	Pregnant women	Facebook Messenger	Menu selection input	Not reported	The chatbot includes content on postnatal contraception, as part of wider content on pregnancy
Chernick et al [18]	Contraception information dissemination	Adolescent girls presenting at an emergency department	SMS	Menu selection input	Health care worker	The chatbot disseminates information on contraception
Green et al [19]	SRH information dissemination and linkage to SRH services	General population	WhatsApp	Menu selection input	No persona	The chatbot includes content on SRH as part of wider SRH content, including contraception decision-making
Handforth and Bertermann [20]	SRH information dissemination and linkage to SRH services	Young women	WhatsApp	Menu selection input	Female relative	The chatbot includes content on contraception as part of wider SRH content
Hussain et al [21]	Contraception information dissemination	Married couples	SMS	Menu selection input	No persona	The chatbot disseminates information on contraception
Woo et al [22]	Pandemic setting health information dissemination	General population	Not reported	Free text input	Knowledgeable young woman	The chatbot includes content on contraception, as part of wider SRH content
Maeda et al [23]	Preconception care and fertility information dissemination	Women	Not reported	Free text input	Not reported	The chatbot includes content on oral contraceptives, as part of wider fertility and preconception care content
Nadarzynski et al [24]	SRH information dissemination	Those at higher risk of poor SRH	Website	Free text input	Neutral robot	The chatbot includes content on contraception, as part of wider SRH content
Population Services International [25]	Pleasure positive SRH information dissemination	General population	WhatsApp	Menu selection input	Knowledgeable young woman	The chatbot includes content on contraception, as part of wider SRH content
Rahman et al [26]	SRH information dissemination	Adolescents	Facebook Messenger	Free text input	Neutral robot	The chatbot includes content on contraception, as part of wider SRH content
Wang et al [27]	SRH information dissemination	Adolescents	Facebook Messenger	Free text input	Health care worker	The chatbot includes content on contraception, as part of wider SRH content
Wilson et al [27]	Information dissemination and digital personal assistant	General population	Google, Google Assistant, Siri	Free speech input	Knowledgeable young woman	The chatbot can answer SRH queries, as part of wider content
Winskell et al [29]	SRH information dissemination	Young women	WhatsApp	Menu selection input	Female peer	The chatbot can answer SRH queries, as part of wider content
Work & Co [30]	SRH information dissemination	Adolescents	Website	Free text input	Neutral robot	The chatbot includes content on contraception, as part of wider SRH content

[a]SRH: sexual and reproductive health.

Three sources focused solely on improving contraceptive outcomes. The Dr Erica chatbot aimed to improve contraceptive knowledge and uptake among adolescent girls presenting at an emergency department in the United States [18]. An analysis of the AskNivi chatbot in Kenya aimed to predict SRH service use after seeking contraceptive information from a chatbot [19].

XSL•FO
RenderX

Finally, a feasibility study evaluated a mobile phone chatbot providing contraceptive information for couples in the United States [21].

The majority of sources identified (n=11) aimed to improve SRH knowledge, attitudes, or behaviors, and contraceptive outcomes were a subset of these [16,17,20,23-30]. One chatbot answered health queries in a COVID-19 pandemic setting, including queries regarding contraception [22].

The chatbots described in our sources had a range of personas. Two had a female peer persona [16,29], 1 had a female relative persona [20], 3 had robot personas that were neutral in age and gender, 4 had knowledgeable female personas [22,25,28], 2 had female health care worker personas [18,27], 2 did not have a persona, and 2 sources did not report on the persona.

Seven of the chatbots in our sample allowed for free text to be inputted by the user and 2 allowed for free speech input. These chatbots use NLP to interpret the user's input, but none of the chatbots in our sample used NLP to generate responses. Seven of the chatbots in our sample used menus of queries and responses to allow users to interact with the chatbot.

Chatbots can be hosted on a range of platforms. In this sample, chatbots were hosted on a website (n=3), WhatsApp (n=4), and Facebook Messenger (n=3), and 2 interacted with the user via SMS text messages. Two chatbots were devices that responded to free speech input. Two sources did not report platform.

Use of Theory to Underpin Chatbot Development

Only 3 sources used a theoretical framework to underpin their analysis. The development of a contraceptive care-seeking predictive model used a "stages of change" model to predict whether women were "ready to act" to access contraceptives [19]. A case study of a national chatbot for adolescents applied Gibson's "theory of affordances," which seeks to understand how relational dynamics between users and environments can enable or prohibit certain actions [27]. In an analysis of an SMS text messaging chatbot, the Unified Theory of Acceptance and Use of Technology model was used to understand how effort, attitude, and user expectations impact behavioral intentions to use a chatbot [21].

Efficacy and Impact

There is little evidence to assess the effectiveness of chatbots in impacting contraceptive knowledge, attitudes, and behaviors. We identified 1 RCT, 1 small cohort study, 1 predictive model development report, 2 development reports, and 1 gray literature report on chatbot impacts on contraceptive behaviors. We found 1 informal report on the chatbot impact on contraceptive knowledge and no evidence on the impact on contraceptive attitudes.

The RCT recruited women planning a pregnancy in Japan and assessed effectiveness using pre- and postintervention surveys on fertility knowledge and behaviors including intention to use oral contraceptives (OCs) after birth [23]. In this double-blinded RCT, participants (n=927) were randomized into 3 groups. The intervention group interacted with the chatbot, control group 1 was invited to read a PDF containing the same fertility and preconception care information, and control group 2 was instructed to read a PDF about the National Pension System. Intention to use OCs was significantly higher in the intervention group than control group 2 ($P<.001$) but significantly lower in the intervention group compared with control group 1 ($P=.005$), suggesting that the chatbot did not impact on intention to use OCs compared with a leaflet with the same content [23].

The small cohort study [18] measured the uptake of hormonal contraception after interaction with the Dr Erica chatbot, an SMS text messaging chatbot providing contraception information. A convenience sample of adolescent girls (n=42) were recruited in an emergency department in the United States and received messages from Dr Erica over 10 weeks inviting the discussion of contraception. There was no control group. Of the 35 participants who completed follow-up through a web-based or telephone survey, 16 (46%) initiated hormonal contraceptives. Of the 30 participants who completed follow-up and were sexually active in the previous 3 months, 52% (15/29) had initiated hormonal contraceptives and 66% (20/30) had used any contraceptive at last intercourse. None of the participants became pregnant [18].

Other evidence on the efficacy of chatbots for contraceptive behavior outcomes is found in the evaluation of the Population Services International (PSI) chatbot in Kenya [25]. PSI piloted a chatbot that disseminated pleasure-positive SRH information including contraceptive information with 1000 referral vouchers for local SRH services issued to those who expressed interest in and intent to use these services. The study found that none of the vouchers were redeemed [25].

A third study reports on data from the users of contraceptive chatbot content within the AskNivi chatbot and predicted that 29% (233/817) of female users were "ready to act" to access contraceptive services [19]. Within 2 weeks of interaction with the chatbot, the chatbot sent up to 2 "check-in" messages and asked users to self-report whether they had sought contraceptive services. In total, 93% (760/817) of women classified by the model as "ready to act" reported accessing contraceptive services. The model was internally validated with a test set of AskNivi users. The model was found to slightly overpredict readiness to act [19].

Improving Knowledge

The only evidence found on the impact of chatbots on contraceptive knowledge comes from a technical report [20] on the development of a chatbot, Big Sis, in Kenya. This report includes the testing of user knowledge before and after engagement with an educational quiz within the chatbot. Average knowledge scores among users (n=419) improved by 36% ($P<.001$) after engaging with chatbot content. Big Sis users were also asked to complete a survey (11% response rate) through which they reported increased awareness and understanding of contraception. No comparison group is reported [20].

One source reports on the accuracy of SRH information provided by voice-activated and response chatbots [19]. This source compared 2 voice-activated and response chatbots (Google Assistant and Siri) for accuracy and relevance against written Google searches using 50 SRH queries. Google searches

XSL•FO
RenderX

outperformed both chatbots, providing 72% (36/50) of the best or equal best responses. Google searches also had the lowest outright failure rate, returning no completely inaccurate or irrelevant answers. Google Assistant performed better than Siri with 50% (25/50) of best or equal best responses versus 32% (16/50); *P*=.04 [19].

User Engagement

Eight sources report user engagement data through number of users, messages received, length of engagement, returning users, length of interaction, number of interactions with web-based promotional material, and the demographics of those who engaged.

Number of Users

Two chatbots reached large audiences, with SnehAI reaching 135,263 unique chatbot users who exchanged 8,170,879 messages with the chatbot over a 5-month period [27] and Ask Roo reporting 3.5 million conversations with users over a year but no data on number of unique users [30]. Other chatbots reached smaller audiences, with Tina reaching 7000 users within a 6-month period [29], Layla's Got You receiving 6164 messages over 21 months [16], and PSI's pleasure-positive chatbot reaching 4866 users over an unspecified time period [25]. Over 10 weeks, 83% (35/42) of Dr Erica users interacted with at least 1 message sent by the chatbot [18].

The number of people interacting with web-based promotional material is reported from the Leyla's Got You chatbot, which shows that a promotional campaign hosted across social media accounts had 2,483,683 impressions [16]. The web-based promotional campaign for Tina had 300,000 click-throughs to find out more about self-injectable contraceptives [29], and Big Sis web-based promotion recruited more than 2500 users for testing purposes [20].

Length of Engagement

Users engaged with SnehAI on average for 1.9 sessions for 7.6 minutes and exchanged 56.2 messages with the chatbot [27]. In total, 59% (1632/2767) of Big Sis users interacted with the chatbot for less than 1 hour, and reengagement rates were 23% to 43% after reminder messages [20]. PSI reports that 1866 users returned to the chatbot [25], with click-throughs to 2 or more branches of the chatbot content increasing from 67% to 83% with the introduction of pleasure content [25]. A Facebook Messenger chatbot, Mama Support, reported that in a 4-week period, a sample of 22 users interacted with the chatbot on average 19 times after reminders were sent [17].

The Demographics of Those Who Engaged

SnehAI engaged users who self-reported as 93% (125,795/135,263) male and 6.8% (9198/135,263) female, with the authors explaining this through the high levels of stigma and shame experienced by young women in this context around engaging with SRH content [27]. Ask Roo reports that 78% of its users are 13 to 19 years old, suggesting that they reached their target audience [30].

User Feedback

A total of 6 studies report user feedback on chatbots. Qualitative interviews were carried out with 40 participants who were at higher risk of poorer sexual health, aged between 18 and 50 years in the United Kingdom, and interacted with the Pat chatbot [24]. Interviews were also carried out with the users of Mama Support (n=22), who were pregnant women aged between 25 and 35 years in urban areas of Kenya [17]. A web-based survey of 256 participants was gathered from the users of AdolescentBot, who were adolescents aged between 10 and 19 years in Bangladesh [26]. Participants (n=278), women aged between 20 and 34 years in Japan, who interacted with the preconception care chatbot provided written feedback on chatbot use [23]. User feedback for Dr Erica was collected at follow-up via a web-based survey or telephone survey with female emergency department patients aged between 14 and 19 years in the United States [18]. The users of Akira (n=57), the pandemic health query chatbot, gave feedback through a user satisfaction survey and were aged between 18 and 20 years in the United Kingdom [22].

Acceptability

A wide range of users found chatbots acceptable. Adolescent girls found that Dr Erica was acceptable, with 94% (29/31) of participants liking the SMS text messages they interacted with and 83% (25/30) stating that they would recommend the chatbot to a friend [18]. Akira received an average satisfaction rating of 2.74 out of 5, although satisfaction decreased with each query that users asked [22]. Pregnant women who used Mama Support(n=22) said that they would refer the chatbot to a friend [17].

Users across different groups experienced chatbots as trustworthy [17,26]. Adolescents found chatbots trustworthy when the chatbots seemed to give accurate answers to their questions [26], and pregnant women trusted a chatbot when it was associated with a reputable organization [17]. Users experienced chatbots as incompetent because either the chatbots could not answer their questions or they had experienced technical difficulties with the platform or device [17,23,24]. Where chatbot content was limited to SRH, some users reported that they wanted to ask questions outside of these limits [17,26]. For example, some adolescents wanted AdolescentBot to provide other health content such as self-care tips for the common cold [26].

Feedback on chatbots included negative experiences of the chatbot's lack of humanness across user groups, including lack of empathy, limited interaction, and experience of the chatbot as "cold" and robot-like [23,24].

Convenience and Accessibility

A wide range of users felt that chatbots were convenient to use because they could be accessed at any time and from any location [23,24,26]. They valued fast and accurate responses to queries that did not require searching the internet, appraising internet sources, or reading long passages of text [17,24]. However, women in Japan felt that interacting with a chatbot was "burdensome" and that reading an accurate document would be more convenient [23]. The users of Pat thought that chatbots

XSL•FO
RenderX

were convenient sign posters to local and appropriate services [24]. The users of Mama Support disliked asking questions through a menu and would prefer to ask their own questions via free text [17].

Anonymity, Nonjudgment, and Privacy

A wide range of users appreciated the anonymity that talking to a chatbot afforded them [17,24,26], but some users had concerns about the privacy of the data provided to the chatbot [24]. Users suggested that they felt less shy talking to a chatbot and that chatbots were a good tool for those who would find it difficult to ask questions face to face because of embarrassment or stigma [17,23,24]. This was reported across user groups.

Cost-Effectiveness

There was no evidence available on the costs or cost-effectiveness of chatbots for impacting contraceptive knowledge, attitudes, and behaviors.

Best Practices

The secondary aim of this review was to establish best practice for developing chatbots to improve contraceptive outcomes. We reviewed all the evidence collected to identify best practice recommendations that came from more than 1 paper. A total of 7 best practices that emerged from the literature are listed in Table 3.

Table 3. Best practices.

Best practice	Quotation
It is best practice to design and develop chatbots with input from the target audience [16,20,29,30]	"The user experience is sacrosanct. All decisions need to be made with users at the center. This begins with understanding user needs and realities, including whether or not a chatbot is a channel with which they can engage" [20]
Chatbot use for contraceptive information is a new behavior; it is best practice for users to be given information on chatbots, how to use them, and why to use them before their first use [24,29]	"Using a chatbot to get information and support about health-related content (and specifically, sexual & reproductive health) is a completely new behavior. It isn't enough to tell consumers WHERE to find your chatbot, you also need to educate them on WHAT a chatbot is and WHY they should use it – the value it offers them as a platform" [29]
Using a chatbot should be intuitive and easy (frictionless) with no superfluous steps to maintain user engagement and avoid drop-offs [20,29]	"Our reading behavior online is to dip in and out of content in an 'on-demand' fashion. Long and complex menus can be overwhelming and turn people off. If you have directed people to your chatbot with a very specific call-to-action, don't make them work too hard to find the information they are interested in" [29]
Chatbots should be hosted on devices and platforms commonly used by the target audience, so that chatbot use is seamless with their routine digital behavior [20,26]	"27% of all users could not see Quick Reply buttons due to the application version or handset that they were using…These issues appeared to be most common amongst users with the Facebook Lite application, those using third-party applications, and those using older handsets. In response, we created an alternative to Quick Replies using numeric menus" [20]
Linking a chatbot with a reputable and well-known organization can build user trust [17,29]	"As I have said many times now, I saw that it [the chatbot] was related to Jacaranda Maternity. Of course, there are the reviews [of Jacaranda] and the reviews were so good. And I felt that by interacting with this chatbot, I was getting very relevant information from a very qualified source. So, I felt nothing fishy at all and I considered it a lot..." [FGD participant] [17]
The persona of a chatbot should resonate with the intended audience and reflect the intended relationship to the audience, for example, peer, health care worker, or neutral robot [16,20,22,25,30]	"A young woman of color was then hired as a copywriter to provide Layla with a 'personality' – rewriting responses using language that would resonate with the target audience" [16]
User privacy should be ensured and visible; this will increase user trust in the chatbot [20,24]	"I'm a bit worried especially due to data protection so I wouldn't do it. If I didn't have the certainty that my data protection absolutely complete itself so yes I don't feel comfortable talking about my sexual health online" [Interview participant] [24]
Some access to human support is important for those in need of urgent help, especially where the intended audience is young people [20,27]	"Humans are still important!... During the test phase, we had 243 triggers, with 61 being interactions that we would classify as a safeguarding alert. These interactions would be addressed via human intervention to connect girls with suitable support services" [20]

Discussion

Principal Findings

The evidence available to assess whether and how chatbots improve contraceptive knowledge, attitudes, and behaviors is sparse and conflicting. One RCT suggests no effect on intention to use contraception [23]. This RCT is well designed and has a large overall sample size of 927, split evenly between an intervention and 2 appropriate control groups. This study relies on self-reporting of intention to use OC rather than uptake of OC. The sample also had a higher prevalence of OC use than national data at baseline. Overall, the methodological quality here is high, and this evidence is considered to be the strongest evidence available on the impact of chatbots on contraceptive outcomes. A cohort study suggests increased uptake of contraception in adolescent girls [18]. However, this cohort study has a small sample size of 42 and did not include a control group. Without a control group, it is difficult to ascertain whether the uptake of contraception in adolescent girls is due

XSL•FO
RenderX

to interaction with the chatbot or due to other factors. A development report suggests no impact on access to services [25]. This was measured by counting the number of "referral vouchers" that were redeemed at local SRH services. This method of assessing the impact on access to services is suboptimal as users may have accessed the service but failed to redeem their voucher. There is poor-quality evidence to suggest increased contraceptive knowledge from interacting with chatbot content, as measured by in-chatbot quizzes [20]. This evidence is considered poor as, although the sample consists of 419 users, no control group was used. Also, this evidence measured user knowledge before and immediately after engagement with chatbot content. Longer-term knowledge retention is not measured. There are some concerns about the accuracy and relevance of the knowledge that personal assistant chatbots provide [28]. However, evidence from other fields suggests that chatbots could improve knowledge as they deliver information in a digestible and engaging format [31]. No evidence on the impact of chatbots on attitudes to contraception was found.

The chatbots included in this review showed limited user reach and engagement from over 100,000 users in 5 months [27] to under 10,000 in 6 months [29]. The reporting of engagement outcomes is highly variable, and the reporting of the marketing strategies used to drive engagement is limited, making interpretation of this evidence difficult.

The evidence base suggests that chatbots supporting contraceptive outcomes are acceptable, convenient, anonymous, and private, but that some users experience chatbots as incompetent, experience technical difficulties, find interaction burdensome, or experience chatbots as impersonal [17,18,23,24,26].

The best practice guidance from the contraception literature is consistent with that in the broader literature [6,10-13].

Research Recommendations

Further research is needed to understand the impact of chatbot technologies on contraceptive outcomes, and it is important that this research consider the following issues: (1) Impact at scale: there are simply no studies that demonstrate that chatbots can improve contraceptive knowledge, attitudes, and behaviors at scale; whether chatbots are more effective among specific populations; and whether chatbots provide efficiencies in cost or value for money. (2) Most chatbot evaluations do not offer comparisons with alternative technologies. These comparisons are essential to understand the impact of the dynamic nature of content delivery (ie, the chatbot), which could also be provided through simpler means, such as printed materials or websites. (3) Where engagement is reported, some description of the strategies used to promote engagement, as well as the size of the target audience, is required to help with the interpretation of data on engagement. This is because engagement may be dependent on factors unrelated to chatbot acceptability, such as the level of promotion. (4) A classification of levels of interactivity within chatbots is required to understand their impact. For example, chatbots that ask users to choose from a menu of questions are very different from those that interpret free-text questions. Research that understands the relative value of more and less complex chatbots is required. (5) The late 2022 launch of generative AI chatbots like ChatGPT and subsequent iterations opens up new possibilities for engaging users and supporting their contraceptive needs. Large language bots are likely the wave of the future, but will need thoughtful evaluation to determine whether they can effectively and cost-effectively improve contraceptive knowledge, attitudes, and behaviors, and to carefully assess and mitigate associated risks.

Limitations

First, we are aware of a number of chatbots that are active in the contraceptive space globally but have not been formally evaluated. To mitigate this, we conducted a thorough search of the gray literature, but it is still likely that this missed chatbots that have not been reported on at all and could therefore not be included in this review.

Second, a wide range of technologies are described as chatbots. These differ in their technical specifications, level of interactivity, method of response generation, target audiences, and the focus of their content. This makes collating the evidence challenging, and in this rapidly changing field, the published literature may not describe the most recent examples of this technology. This review used an up-to-date definition of a chatbot and included all technologies that met this definition regardless of whether they were described as a chatbot or otherwise.

Lastly, the lack of robust quantitative evaluations meant that it was not possible to complete a meta-analysis. Other methods of quantitative synthesis were explored, but ultimately the lack of evidence meant this was not possible. As more evidence becomes available in this field, a meta-analysis may be possible in future reviews.

Conclusions

We found limited and conflicting evidence on chatbots to improve contraceptive knowledge, attitudes, and behavior. Further research should aim to understand the impact of chatbot interventions in comparison with alternative technologies, acknowledge the variation in chatbot interventions, and seek to identify key features associated with their impact.

Acknowledgments

The work described in this paper was funded by the Bill & Melinda Gates Foundation through the Implementation Research and Innovation Support Framework Grant. This publication is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

XSL•FO
RenderX

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ERM, NL, and GJ led on the conceptualization of this project. The search strategy and methodology were developed by RM, PB, and DM, with substantive contributions from ERM, NL, and GJ. Database searching was carried out by RM. Search screening was carried out by RM and PB. Quality assessment of sources was carried out by RM and PB, with contributions from ERM, NL, DM, and GJ. Data extraction was carried out by RM and PB. Synthesis was carried out by RM and PB, with contributions from ERM, NL, DM, and GJ. RM and PB wrote an original draft of the published work. ERM, NL, DM, and GJ provided review and substantive editing in preparation of the final draft of the published work.

Conflicts of Interest

ERM is a program officer on the Family Planning team at the Bill & Melinda Gates Foundation. She developed the scope of work for this research. As a program officer, she contributed technical input but did not directly conduct the study. NL and GJ are affiliated with Dimagi. Dimagi builds and provides behaviour-change chatbots for organizations, including for Sexual and Reproductive Health.

Multimedia Appendix 1
Database search terms.
[DOCX File , 14 KB - jmir_v26i1e46758_app1.docx]

Multimedia Appendix 2
Data extraction framework.
[DOCX File , 13 KB - jmir_v26i1e46758_app2.docx]

Multimedia Appendix 3
PRISMA checklist.
[DOCX File , 32 KB - jmir_v26i1e46758_app3.docx]

References

1. Digital health to support family planning providers: improving knowledge, capacity, and service quality. High Impact Practices in Family Planning (HIPs). 2020. URL: https://www.fphighimpactpractices.org/briefs/digital-health-to-support-family-planning-providers/ [accessed 2022-10-01]
2. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2022-10-01]
3. IMAP statement on DHI for SRHR. International Planned Parenthood Federation (IPPF). 2022. URL: https://www.ippf.org/resource/imap-statement-dhi-srhr [accessed 2022-10-01]
4. Wasson EJ, Driver K, Hughes M, Bailey J. Sexual reproductive health chatbots: should we be so quick to throw artificial intelligence out with the bathwater? BMJ Sex Reprod Health 2021;47(1):73 [FREE Full text] [doi: 10.1136/bmjsrh-2020-200823] [Medline: 32883682]
5. Nadarzynski T, Bayley J, Llewellyn C, Kidsley S, Graham CA. Acceptability of artificial intelligence (AI)-enabled chatbots, video consultations and live webchats as online platforms for sexual health advice. BMJ Sex Reprod Health 2020;46(3):210-217 [FREE Full text] [doi: 10.1136/bmjsrh-2018-200271] [Medline: 31964779]
6. Rapp A, Curti L, Boldi A. The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. Int J Hum Comput 2021;151:102630 [FREE Full text] [doi: 10.1016/j.ijhcs.2021.102630]
7. Passanante A, Pertwee E, Lin L, Lee KY, Wu JT, Larson HJ. Conversational AI and vaccine communication: systematic review of the evidence. J Med Internet Res 2023;25:e42758 [FREE Full text] [doi: 10.2196/42758] [Medline: 37788057]
8. Mills R, Mangone ER, Lesh N, Mohan D, Baraitser P. Chatbots to improve sexual and reproductive health: realist synthesis. J Med Internet Res 2023;25:e46761 [FREE Full text] [doi: 10.2196/46761] [Medline: 37556194]
9. Nißen M, Selimi D, Janssen A, Cardona DR, Breitner MH, Kowatsch T, et al. See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. Comput Hum Behav 2022;127:107043 [FREE Full text] [doi: 10.1016/j.chb.2021.107043]
10. Dhinagaran DA, Martinengo L, Ho MHR, Joty S, Kowatsch T, Atun R, et al. Designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER): development of a conceptual framework. JMIR Mhealth Uhealth 2022;10(10):e38740 [FREE Full text] [doi: 10.2196/38740] [Medline: 36194462]

XSL•FO
RenderX

11. ter Stal S, Kramer LL, Tabak M, op den Akker H, Hermens H. Design features of embodied conversational agents in eHealth: a literature review. Int J Hum Comput 2020;138:102409 [FREE Full text] [doi: 10.1016/j.ijhcs.2020.102409]
12. Suhaili SM, Salim N, Jambli MN. Service chatbots: a systematic review. Expert Syst Appl 2021;184:115461 [FREE Full text] [doi: 10.1016/j.eswa.2021.115461]
13. Curtis RG, Bartel B, Ferguson T, Blake HT, Northcott C, Virgara R, et al. Improving user experience of virtual health assistants: scoping review. J Med Internet Res 2021;23(12):e31737 [FREE Full text] [doi: 10.2196/31737] [Medline: 34931997]
14. Critical Appraisal Skills Programme (CASP) randomised controlled trial and cohort checklists. Critical Appraisal Skills Programme. 2020. URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2022-10-01]
15. Tyndall J. AACODS checklist for appraising grey literature. Flinders University. 2010. URL: https://dspace.flinders.edu.au/jspui/bitstream/2328/3326/4/AACODS_Checklist.pdf [accessed 2022-10-01]
16. Bonnevie E, Lloyd TD, Rosenberg SD, Williams K, Goldbarg J, Smyser J. Developing a tailored contraception chatbot for Black and Hispanic young women. Health Educ J 2020;80(4):413-424 [FREE Full text] [doi: 10.1177/0017896920981122]
17. Brannock M. Investigating attitudes and preferences towards a chatbot pregnancy guide within Facebook's social media platform amongst pregnant women in Kenya. Master's thesis, Duke University. 2019. URL: https://hdl.handle.net/10161/18868 [accessed 2023-12-19]
18. Chernick LS, Stockwell MS, Gonzalez A, Mitchell J, Ehrhardt A, Bakken S, et al. A user-informed, theory-based pregnancy prevention intervention for adolescents in the emergency department: a prospective cohort study. J Adolesc Health 2021;68(4):705-712 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.07.020] [Medline: 32948403]
19. Green EP, Pradheep S, Heinzelman J, Nyanchoka A, Achieng D, Goyal S, et al. Predicting healthcare-seeking behavior based on stated readiness to act: development and validation of a prediction model. Transl Behav Med 2022;12(1):ibab096 [FREE Full text] [doi: 10.1093/tbm/ibab096] [Medline: 34283889]
20. Handforth C, Bertermann K. How Girl Effect built a chatbot. Girl Effect. 2018. URL: https://doi.org/10.13140/RG.2.2.26155.90404 [accessed 2022-10-01]
21. Hussain SA, Ogundimu F, Bhattarai S. Mobile phone-based chatbot for family planning and contraceptive information. In: Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Healthcare Applications. 2019 Presented at: 10th International Conference, DHM 2019, Held as Part of the 21st HCI International Conference, HCII 2019; July 26–31, 2019; Orlando, FL, USA URL: https://link.springer.com/book/10.1007/978-3-030-22219-2 [doi: 10.1007/978-3-030-22219-2_26]
22. Woo WL, Gao B, Al-Nima RRO, Ling WK. Development of conversational artificial intelligence for pandemic healthcare query support. Int J Autom Artif Intell Mach Learn 2020;1(1):54-79 [FREE Full text]
23. Maeda E, Miyata A, Boivin J, Nomura K, Kumazawa Y, Shirasawa H, et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. Reprod Biomed Online 2020;41(6):1133-1143 [FREE Full text] [doi: 10.1016/j.rbmo.2020.09.006] [Medline: 33039321]
24. Nadarzynski T, Puentes V, Pawlak I, Mendes T, Montgomery I, Bayley J, et al. Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis. Sex Health 2021;18(5):385-393 [FREE Full text] [doi: 10.1071/SH21123] [Medline: 34782055]
25. McPake R, Population Services International (PSI). Pilot story: will access to sex-positive and reproductive health information through a chatbot lead to increased contraceptive use amongst Kenyan youth? Frontier Technologies Hub. 2022. URL: https://medium.com/frontier-technologies-hub/pilot-story-will-access-to-sex-positive-and-reproductive-health-information-through-a-chatbot-d41738947d0c [accessed 2022-10-01]
26. Rahman R, Rahman MR, Tripto NI, ME A, Apon SH, Shahriyar R. Adolescentbot: understanding opportunities for chatbots in combating adolescent sexual and reproductive health problems in Bangladesh. New York: Association for Computing Machinery; 2021 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan URL: https://doi.org/10.1145/3411764.3445694 [doi: 10.1145/3411764.3445694]
27. Wang H, Gupta S, Singhal A, Muttreja P, Singh S, Sharma P, et al. An artificial intelligence chatbot for young people's sexual and reproductive health in India (SnehAI): instrumental case study. J Med Internet Res 2022;24(1):e29969 [FREE Full text] [doi: 10.2196/29969] [Medline: 34982034]
28. Wilson N, MacDonald EJ, Mansoor OD, Morgan J. In bed with Siri and Google Assistant: a comparison of sexual health advice. BMJ 2017;359:j5635 [FREE Full text] [doi: 10.1136/bmj.j5635] [Medline: 29237603]
29. Winskell A. Building chatbots for health content? This is for you. Population Services International. 2021. URL: https://www.psi.org/2021/12/building-chatbots-for-health-content-this-is-for-you/ [accessed 2022-10-01]
30. Planned parenthood: using AI to reach teens and close the sex education gap. Work & Co. URL: https://work.co/clients/planned-parenthood/ [accessed 2022-10-01]
31. Cho K, Foo YM, Dalziel B, Hu W. Chatbot-mediated learning of cardiac auscultation. Intern Med J 2022;52(12):2176-2180 [FREE Full text] [doi: 10.1111/imj.15971] [Medline: 37133367]

XSL•FO
RenderX

Abbreviations

AACODS: Authority, Accuracy, Coverage, Objectivity, Date, and Significance
CASP: Critical Appraisal Skills Programme
NLP: natural language processing
OC: oral contraceptive
PICO: Population, Intervention, Control, Outcome
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PSI: Population Services International
RCT: randomized controlled trial
SRH: sexual and reproductive health

Edited by G Eysenbach, T Leung; submitted 24.02.23; peer-reviewed by N Mungoli; comments to author 31.03.23; revised version received 25.04.23; accepted 16.11.23; published 27.02.24.

Please cite as:
Mills R, Mangone ER, Lesh N, Jayal G, Mohan D, Baraitser P
Chatbots That Deliver Contraceptive Support: Systematic Review
J Med Internet Res 2024;26:e46758
URL: https://www.jmir.org/2024/1/e46758
doi:10.2196/46758
PMID:38412028

©Rhiana Mills, Emily Rose Mangone, Neal Lesh, Gayatri Jayal, Diwakar Mohan, Paula Baraitser. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Attrition in Conversational Agent–Delivered Mental Health Interventions: Systematic Review and Meta-Analysis

Ahmad Ishqi Jabir[1,2], BSc, MSc; Xiaowen Lin[1], BA; Laura Martinengo[1], MD, PhD; Gemma Sharp[3], PhD; Yin-Leng Theng[4], PhD; Lorainne Tudor Car[1,5], MD, PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore, Singapore

[2]Future Health Technologies, Singapore-ETH Centre, Campus for Research Excellence And Technological Enterprise, Singapore, Singapore

[3]Department of Neuroscience, Monash University, Melbourne, Australia

[4]Centre for Healthy and Sustainable Cities, Wee Kim Wee School of Communication and Information, Nanyang Technological University Singapore, Singapore, Singapore

[5]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Lorainne Tudor Car, MD, PhD
Lee Kong Chian School of Medicine
Nanyang Technological University Singapore
11 Mandalay Road, Level 18
Singapore, 308232
Singapore
Phone: 65 69041258
Email: lorainne.tudor.car@ntu.edu.sg

Abstract

Background: Conversational agents (CAs) or chatbots are computer programs that mimic human conversation. They have the potential to improve access to mental health interventions through automated, scalable, and personalized delivery of psychotherapeutic content. However, digital health interventions, including those delivered by CAs, often have high attrition rates. Identifying the factors associated with attrition is critical to improving future clinical trials.

Objective: This review aims to estimate the overall and differential rates of attrition in CA-delivered mental health interventions (CA interventions), evaluate the impact of study design and intervention-related aspects on attrition, and describe study design features aimed at reducing or mitigating study attrition.

Methods: We searched PubMed, Embase (Ovid), PsycINFO (Ovid), Cochrane Central Register of Controlled Trials, and Web of Science, and conducted a gray literature search on Google Scholar in June 2022. We included randomized controlled trials that compared CA interventions against control groups and excluded studies that lasted for 1 session only and used Wizard of Oz interventions. We also assessed the risk of bias in the included studies using the Cochrane Risk of Bias Tool 2.0. Random-effects proportional meta-analysis was applied to calculate the pooled dropout rates in the intervention groups. Random-effects meta-analysis was used to compare the attrition rate in the intervention groups with that in the control groups. We used a narrative review to summarize the findings.

Results: The systematic search retrieved 4566 records from peer-reviewed databases and citation searches, of which 41 (0.90%) randomized controlled trials met the inclusion criteria. The meta-analytic overall attrition rate in the intervention group was 21.84% (95% CI 16.74%-27.36%; I^2=94%). Short-term studies that lasted ≤8 weeks showed a lower attrition rate (18.05%, 95% CI 9.91%- 27.76%; I^2=94.6%) than long-term studies that lasted >8 weeks (26.59%, 95% CI 20.09%-33.63%; I^2=93.89%). Intervention group participants were more likely to attrit than control group participants for short-term (log odds ratio 1.22, 95% CI 0.99-1.50; I^2=21.89%) and long-term studies (log odds ratio 1.33, 95% CI 1.08-1.65; I^2=49.43%). Intervention-related characteristics associated with higher attrition include stand-alone CA interventions without human support, not having a symptom tracker feature, no visual representation of the CA, and comparing CA interventions with waitlist controls. No participant-level factor reliably predicted attrition.

Conclusions: Our results indicated that approximately one-fifth of the participants will drop out from CA interventions in short-term studies. High heterogeneities made it difficult to generalize the findings. Our results suggested that future CA

XSL•FO
RenderX

interventions should adopt a blended design with human support, use symptom tracking, compare CA intervention groups against active controls rather than waitlist controls, and include a visual representation of the CA to reduce the attrition rate.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42022341415; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022341415

(J Med Internet Res 2024;26:e48168) doi:10.2196/48168

KEYWORDS

conversational agent; chatbot; mental health; mHealth; attrition; dropout; mobile phone; artificial intelligence; AI; systematic review; meta-analysis; digital health interventions

Introduction

Description of the Problem

Mental health disorders are among the largest contributors to the global disease burden, affecting 1 in every 8 people, or 970 million people around the world [1,2]. However, access to evidence-based interventions for the prevention and treatment of mental health disorders is limited [3,4]. This is due to various factors such as a lack of mental health services and professionals, poor mental health literacy, fear of stigma, and low perceived need for treatment [5-10]. There is a need for scalable and accessible mental health services. Digital technologies such as smartphones or websites are increasingly being used for the delivery of mental health interventions and have the potential to improve access to mental health care. Digital mental health interventions allow for the scalable delivery of diverse therapeutic approaches such as cognitive behavioral therapy and mindfulness for the treatment of mental health conditions such as depression, anxiety, substance abuse, and eating disorders [11-16].

Description of the Intervention

Conversational agents (CAs) or chatbots are a more recent type of digital intervention, and they are becoming a popular method to deliver mental health interventions. CAs can be defined as computer algorithms designed to simulate human conversations textually or via speech through an interface [17]. CA-delivered mental health interventions (CA interventions) combine the delivery of psychotherapeutic content with an automated dialogue system that simulates the interaction between a mental health expert and the user [18]. These interventions provide an alternative avenue of psychotherapy to individuals who are not able to access mental health services owing to issues regarding time, location, or availability of resources [19]. CAs can also be a useful addition to traditional in-person therapy [20,21]. The presence of a CA can further contribute to improved therapeutic alliances with users to enhance adherence to the intervention [22,23]. Evidence for the efficacy of CAs in delivering mental health support is growing steadily. A recent meta-analysis showed that CA-delivered psychotherapy in adults significantly improved depressive symptoms with a medium effect size [19]. Providing self-guided therapy remotely via CAs may help address barriers to mental health access such as cost, long waiting time, and stigma [24]. Although the impact of mental health interventions delivered by CAs seems promising, studies evaluating such interventions also suggest high study attrition among participants [19]. Attrition or dropout occurs when participants do not complete the randomized controlled trial (RCT) assessments or complete the research protocol.

Digital health interventions typically report rapid and high attrition [13,25]. The overall attrition rate quantifies the level of attrition for the whole sample in a clinical trial, and the differential attrition rate refers to the level of attrition in the intervention group compared with that in the comparison group [26]. Attrition in clinical trials may introduce bias by modifying the random composition of the trial groups, limiting the generalizability of the study, and reducing the study power owing to reduced sample size [13,27]. To improve the quality of future clinical trials on CA interventions, there is a need to determine the attrition rates and the factors contributing to attrition in CA interventions.

Why Is It Important to Conduct This Review?

There is scant evidence on the possible factors associated with attrition in CA interventions for mental health and health care in general. The review conducted by Lim et al [19] on the effectiveness of CA interventions for depression and anxiety symptoms indicated that almost a fifth of the participants (19%) attrited throughout the trials without exploring factors associated with the attrition. This was comparable with other reviews on digital health and digital mental health interventions reporting attrition rates that ranged from 24.1% to 47.8% after adjusting for publication bias [13,28]. In general, factors shown to be associated with attrition in trials of digital health interventions include poor user experience, a lack of perceived value, and privacy concerns [28,29]; for example, studies on mental health apps reported technical issues and errors that might affect users' overall experience [15,30]. Qualitative findings further suggested that factors such as a lack of human interactions in digital health interventions and users' technological competence also played a role in participants' attrition [31].

In addition, for smartphone-based mental health interventions, providing monetary compensation and reminders to engage were associated with significantly lower attrition rates [13]. Conversely, participants in the intervention condition were more likely to drop out than the waitlist participants [13,32]. These reviews focused only on smartphone-delivered interventions and included studies published before 2020, omitting several more recently published studies on CA interventions. To fully harness CA interventions, there is a need to better understand the factors associated with both overall attrition as well as differential attrition in these interventions.

To this end, we aimed to (1) estimate the overall and differential rates of attrition in CA interventions, (2) evaluate the impact of

XSL•FO
RenderX

the study on design- and intervention-related aspects on the overall and differential attrition rates in CA interventions, and (3) map and describe study design features aimed at reducing or mitigating study attrition in the trials.

Methods

Overview

We performed a systematic review of attrition rates in RCTs of CA health interventions in line with the Cochrane gold standard methodology [33] and the meta-analysis approach outlined by Linardon and Fuller-Tyszkiewicz [13]. We reported this review in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [34]. The PRISMA checklist is included in Multimedia Appendix 1.

Criteria for Study Selection

Types of Studies

Our inclusion criteria included RCTs, cluster RCTs, crossover RCTs, quasi-RCTs, and pilot RCTs in English. We decided to include these variations of RCTs because the field is still nascent, and findings from different forms of RCTs could be beneficial to understand the attrition rate in CA interventions. The publication types included peer-reviewed journals and conference papers that were published up to June 2022.

Types of Participants

Participants' characteristics included healthy participants and participants with subclinical or clinically diagnosed mental health disorders such as depression, anxiety, attention-deficit/hyperactivity disorder, and bipolar disorder. Participants of any age were included so long as they personally interacted with the CA.

Types of Interventions

We included studies reporting a synchronous 2-way exchange with the participants via a CA. We excluded studies where the CA dialogues were wholly or partially delivered by human operators (*Wizard of Oz*) and studies with asynchronous response systems.

The interventions included either the delivery of psychotherapeutic content or those that provided training to improve mental well-being, reduced the symptoms of mental health conditions, or both. This included studies that aimed to reduce the symptoms of depression for clinical and subclinical populations or studies implementing mindfulness training for healthy adults. Detailed inclusion and exclusion criteria are outlined in Multimedia Appendix 2 [13,17,33,35-39].

Types of Outcome Measures

The primary outcome was the reported attrition number and the attrition rate calculated by the weighted risk of attrition of participants against the sample size of the studies for participants assigned to the CA intervention who then discontinued the study. This included the total attrition rate and the differential attrition rate in the intervention and comparison groups.

Search Methods for the Identification of Studies

The search strategy included index terms and keywords that describe CAs, such as "conversational agent," "embodied conversational agent," "virtual assistant," and "virtual coach" (Multimedia Appendix 3). The search strategy was previously developed for our scoping reviews [40,41] and was updated to include studies from 2020 to 2022 with the assistance of a medical librarian. We conducted the searches in the following databases on June 6, 2022: PubMed, Embase (Ovid), PsycINFO (Ovid), Cochrane Central Register of Controlled Trials, and Web of Science. In addition, we conducted a gray literature search on the first 200 entries from Google Scholar as suggested by the literature for the most optimal combination of databases [42,43]. We did not include any filter terms in the search. We also performed citation chasing by searching for all records citing ≥1 of the included studies (forward citation chasing) and all records referenced in the included studies (backward citation chasing) using *Citationchaser* [44]. The tables of excluded studies are presented in Multimedia Appendix 4.

Data Collection and Analysis

Selection of Studies

On updating the searches from 2020 to 2022, we imported all identified references from the different electronic databases into a single file. The duplicated records were removed using *revtool* on R [35] and manually on Zotero (Corporation for Digital Scholarship). One reviewer (AIJ) performed the title and abstract screening using ASReview [36], an open-source machine learning software tool. The tool uses an active learning algorithm to actively sort and re-sort the records by prioritizing the most relevant records first based on the user's inclusion and exclusion decisions. The title and abstract screening steps are detailed in Multimedia Appendix 2.

Three reviewers (AIJ, XL, and LM) were engaged in the full-text review. One reviewer (AIJ) retrieved the full text of the studies, and 2 reviewers (AIJ and XL) assessed their eligibility independently and in parallel. Any disagreements were discussed and resolved between the reviewers and with a third reviewer (LM) acting as the arbiter. Studies that were identified in our previous reviews (up to 2020) [41] and met the inclusion criteria of this review were included based on discussions among the 3 reviewers (AIJ, XL, and LM).

Data Extraction and Management

Data were extracted using a data extraction form on Microsoft Excel. The data extraction form was piloted by 2 reviewers (AIJ and XL) on the same papers (5/41, 12%) and amended in line with the feedback. We also included additional fields as required from the data extraction form that we referenced from Linardon and Fuller-Tyszkiewicz [13]. Four reviewers (AIJ, XL, GS, and Nileshni Fernando) extracted the data independently and in parallel.

We extracted the year of publication; study design; the type of comparison group (active or inactive); the type of intervention; and details of the CAs, including the type of CA (rule based or artificial intelligence [AI] enhanced), the personality of the CA [17], the level of human support, the presence of a reminder

XSL•FO
RenderX

mechanism, and the input and output modalities of the CA. In addition, we extracted information on the study duration, compensation paid to study participants, and any other mechanism included specifically to increase user engagement. Any disagreements among the reviewers were resolved by discussion.

Assessment of the Risk of Bias in Included Studies

Four reviewers (AIJ, XL, GS, and Nileshni Fernando) independently assessed the risk of bias in the included studies using the Cochrane Risk of Bias Tool 2.0 [33] and visualized using *robvis* [45]. The risk of bias assessment was piloted with 10 (24%) of the 41 studies for consistency and clarity of judgment by 2 reviewers (AIJ and XL). The steps involved in the assessment of the risk of bias are detailed in Multimedia Appendix 2, and we have provided a summary, along with a table, in Multimedia Appendix 5. We requested clarification or more data from the authors of 1 (2%) of the 41 studies but did not receive any response even after sending a reminder 2 weeks later. The assessment of publication bias was reported via funnel plots and the Egger test for publication bias [37].

Data Analysis

The meta-analysis was conducted based on the approach outlined by Linardon and Fuller-Tyszkiewicz [13] and the Cochrane Handbook for Systematic Reviews of Interventions (version 6.3) [33]. We defined attrition as the number of participants who dropped out of the study during the intervention period by not completing the postintervention assessment. We did not include the follow-up period [13]. For crossover design studies, we only included the data before the crossover following the aforementioned definition. The second part of the crossover was not considered as the follow-up period.

The study's overall attrition rate was estimated by calculating the weighted pooled event rate using random-effect models based on a meta-proportional approach [33] using Freeman-Tukey double arcsine transformed proportion [38]. This indicated the relative risk of attrition against the sample size of the studies for participants assigned to the CA intervention group. The aim of this overall attrition analysis was to compute the overall rate of attrition in the intervention group after controlling for the different sample sizes in the included studies. Event rates were then converted to percentages of events per 100 participants and calculated separately for all included studies (short-term studies [≤8 wk from baseline] as well as longer-term studies [>8 wk from baseline]). We used short-term and long-term groupings to facilitate a comparison between our results and those of the previous study on attrition in smartphone-delivered interventions for mental health problems [13].

The differential attrition rate was calculated as the odds ratio (OR) of the likelihood to attrit between the CA intervention condition and the comparison condition. The aim of the differential attrition analysis was to understand the odds of attrition compared with the control group. The ORs were calculated using random-effect models separately for short-term and long-term studies weighted by their inverse variance. Studies with 0 events in both arms were weighted as 0, and a correction of 0.5 was added to the arm with 0 events as a continuity correction. A log OR of >1 indicated a higher likelihood of attrition in the CA intervention groups compared with the control groups. We also conducted subgroup analyses to explore the sources of heterogeneity in both overall and differential meta-analyses. The detailed meta-analysis procedure and subgroup analyses conducted are specified in Multimedia Appendix 2. We also performed post hoc sensitivity analyses for subgroups with <5 studies because the estimate of tau-square might be imprecise [39]. In addition, we conducted exploratory analyses of all included studies regardless of the intervention duration using the same set of prespecified subgroup analyses on the overall and differential meta-analyses.

Meta-analysis was not conducted on the participant-level factors and the predictors of attrition owing to variability in reporting. We also identified additional factors significantly associated with attrition (P<.05) in the included RCTs. We collated and narratively presented these factors associated with attrition as reported by the studies.

Results

Overview

The updated search strategy retrieved 2228 records from peer-reviewed databases and 2319 from citation searching. After removing duplicates, of the total 4547 (2228+2319) records, 4030 (1877+2319[citation searching]-147[duplicates in citation searching]-19[records from other methods]) (88.63%) were screened on ASReview. Of these 4030 studies, 179 (4.4%) were then considered for full-text screening. We included 11 (6.1%) of the 179 studies identified from the full-text screening. We further identified 2 systematic reviews on CA intervention and included 14 studies that were not identified from our search strategy [19,46]. These studies used the *Deprexis* and electronic problem-solving treatment (ePST) interventions that did not explicitly identify themselves as CAs; for instance, both *Deprexis* and ePST described themselves as simulated dialogue that tailored their responses based on users' input [47,48]. Subsequently, we followed up with an additional search on PubMed using "Deprexis OR ePST" as the search term and included 3 additional studies. We also included 13 studies identified in our previous review [41]. Thus, the total number of included studies is 41 (studies included in previous scoping review: n=13, 32%; new studies included from databases: n=11, 27%; and new studies included via other methods: n=17, 41%). Figure 1 presents the study selection process.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. CA: conversational agent; ePST: electronic problem-solving treatment; RCT: randomized controlled trial. [a][19]; [b][46].

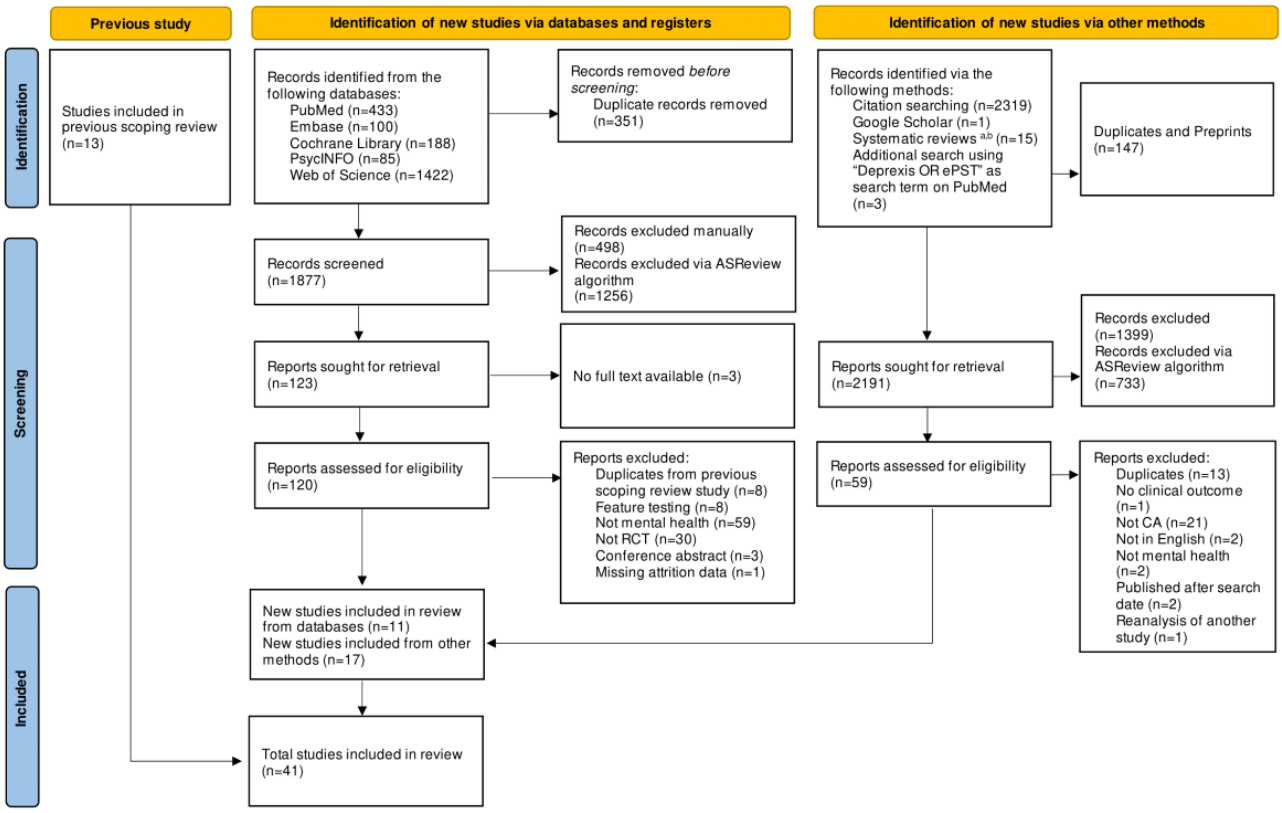

Study Characteristics

Of the 41 studies included in this review, 22 (54%) were published before 2020 [15,47-67], and 19 (46%) were published in 2020 or later [14,68-85] (Table 1). Most of the studies were conducted in the United States (13/41, 32%) [14,15,48,56,58,59,64-66,68,70,82,85] and Germany (13/41, 32%) [47,50,52-55,60-62,69,75], with some studies (2/41, 5%) conducted in multiple countries such as Switzerland and Germany [51] and Switzerland, Germany, and Austria [57]. Most of the interventions were short-term interventions and lasted ≤8 weeks (26/41, 63%) [14,15,48,49,52,56,58,62-68,70-74,76-81,84], whereas some of the studies (15/41, 37%) lasted >8 weeks [47,50,51,53-55,57,59-61,69,75,82,83,85].

XSL•FO
RenderX

Table 1. Characteristics of included studies (n=41).

Study characteristics	Values, n (%)
Year of publication	
Before 2020	22 (54)
2020 or later	19 (46)
Country	
United States	13 (32)
Germany[a,b]	13 (32)
South Korea	3 (7)
Switzerland[a,b]	2 (5)
United Kingdom	2 (5)
Other (European Union)[b,c]	5 (12)
Other[d]	6 (15)
Type of study design	
RCT[e]	29 (71)
Pilot RCT	8 (20)
Crossover RCT	4 (10)
Study duration	
≤8 wk	26 (63)
>8 wk	15 (37)
Sample population	
General population	11 (27)
Population considered at risk	18 (44)
Clinical population	12 (29)
Target clinical outcome	
Treatment and monitoring	17 (41)
Education and training	24 (59)
Target disorder	
Depression only	17 (41)
Mental well-being	9 (22)
Co-occurring depression and anxiety	5 (12)
Other[f]	10 (24)
Type of control	
Waitlist control	18 (44)
Active control	15 (37)
Treatment as usual	8 (20)
Enrollment method	
Remote enrollment option only	23 (56)
In-person enrollment option provided	16 (39)
Not specified	2 (5)
Session type	
Defined session length	29 (71)
User-determined session length	12 (29)

XSL•FO
RenderX

Study characteristics	Values, n (%)
Attrition range (%)	
0-10	13 (32)
11-20	6 (15)
21-30	11 (27)
31-40	2 (5)
41-50	3 (7)
>50	6 (15)

[a]Conducted in both Switzerland and Germany.

[b]Conducted in Switzerland, Germany, and Austria.

[c]Ireland, Sweden, Italy, and the Netherlands.

[d]Japan, Ukraine, Argentina, New Zealand, China, and Russia.

[e]RCT: randomized controlled trial.

[f]Anxiety only, panic disorder, height phobia, gambling disorder, substance abuse, attention-deficit/hyperactivity disorder, irritable bowel syndrome, eating disorder, and personality disorder.

Psychoeducation and training were the focus of 24 (59%) of the 41 studies [47-57,59-62,64,66,69,70,75,77-79,84]. In almost half of the studies (18/41, 44%), participants were screened for mental health symptoms before the start of the study [14,50,52-56,59,62,63,66,68,72-74,80,82,83], and more than half of the studies (23/41, 56%) enrolled participants remotely using the web or the telephone [14,15,47,49-53,56,57,62,64,65,68-70,72,75,77,81-84]. More than one-third of the studies (17/41, 41%) focused on depression as the target disorder [47,48,50-56,58-62,67,69,83]. Of the 41 studies, 18 (44%) used waitlist control group participants [14,48-54,56,58,62,63,66,68,72,77,78,82], and 15 (37%) used an active control that included information or self-help (n=10, 67%) [15,60,65,71,74-76,80,81,83], alternative or comparable treatments such as stress-management cognitive behavioral therapy without a digital component (n=3, 20%) [73,84,86], or symptoms rating only (n=2, 13%) [64,70].

In the intervention group, of the 41 studies, 13 (32%) reported attrition between 0% and 10% [15,48,49,51,58,59,63,65,70,71,73,79,80], 6 (15%) reported attrition between 11% and 20% [14,57,67,74,78,83], 11 (27%) reported attrition between 21% and 30% [47,52,54-56,60,61,66,69,72,85], 2 (5%) reported attrition between 31% and 40% [53,64], 3 (7%) reported attrition between 41% and 50% [68,75,81], and 6 (15%) reported >50% attrition [50,62,76,77,82,84].

Risk-of-Bias Assessment

The most common risk of bias in the included studies was the bias in the measurement of the outcomes because they were all self-reported outcomes (Multimedia Appendix 5). The second most common bias was due to the selection of the reported results because most of the studies (18/41, 44%) [15,48-51,58,59,62,64,66,70,71,74,77,80,81,83,84] did not cite the RCT protocol or statistical analysis plan. Most of the studies (29/41, 71%) reported an intention-to-treat analysis. Figure 2 shows the summary plot of the risk of bias.

Figure 2. Summary plot for the risk-of-bias assessment of the included studies.

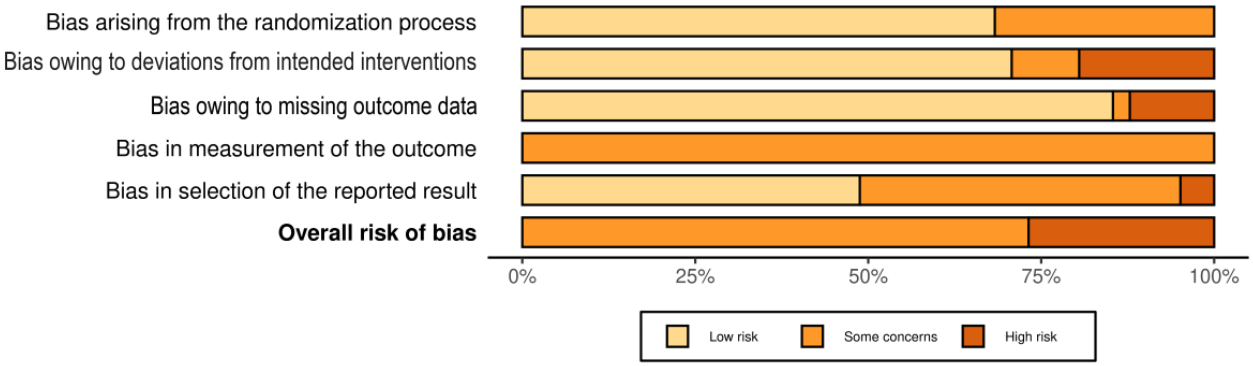

CA Characteristics

Most of the CAs were rule based (29/41, 71%), and 12 (29%) were AI enhanced using natural language processing or other AI-based algorithms (Table 2). More than one-third of the studies (15/41, 37%) did not describe any specific visual representation of the CA. These were mainly studies that included *Deprexis* or *Deprexis*-based interventions (14/15, 93%) because they did not specifically identify themselves as CAs but used dialogue-based interventions. Of the 41 studies, 14 (34%) included an avatar or a static visual representation of the CA and 8 (20%) represented the CA using an embodied CA

XSL•FO
RenderX

(ECA). With regard to the ECAs, of the 8 studies, 4 (50%) used relational agents, 3 (38%) used 3D-generated renders, and 2 (25%) used prerecorded videos. The CAs mostly presented a coach-like personality characterized by encouraging and nurturing personalities (19/41, 46%), followed by a factual personality characterized by being nonjudgmental and offering responses based on facts and observations (14/41, 34%). Of the 41 studies, in 5 (12%), the CA was designed to look like a physician or a health care professional, and in 3 (7%), the CA conversed with users using informal language characterized by exclamations, abbreviations, and emoticons in the dialogue (7%). More than one-third of the interventions were delivered via web-based applications (15/41, 37%), followed by those delivered by a stand-alone smartphone app (11/41, 27%).

Table 2. Characteristics of conversational agents (CAs; n=41).

Characteristics	Values, n (%)
Type of CA	
No avatar or no visual representation	15 (37)
Avatar only	14 (34)
ECA[a]	8 (20)
Not specified	4 (10)
Delivery channel	
Web-based application	15 (37)
Stand-alone smartphone app	11 (27)
Computer- or laptop computer- or tablet computer–embedded program	7 (17)
Messaging app based[b]	7 (17)
Not specified	1 (2)
Dialogue modality	
Rule based	29 (71)
AI[c] enhanced	12 (29)
Personality	
Coach like	19 (46)
Factual	14 (34)
Health care professional like	5 (12)
Informal	3 (7)

[a]ECA: embodied CA.

[b]Slack, Facebook Messenger, or WeChat.

[c]AI: artificial intelligence.

Study Attrition Rates

Overview

The overall weighted attrition rate for the intervention groups in all included studies was 21.84% (95% CI 16.74%-27.36%; I^2=94%), whereas the differential attrition rate differed from 0% (log OR 1.28, 95% CI 1.10-1.48; I^2=34.6%), indicating that the participants who received CA interventions were more likely to attrit than the control group participants. Figure 3 [14,15,47-85] shows the attrition rates for all included studies.

XSL•FO
RenderX

Figure 3. Overall attrition rates for the intervention group in conversational agent–delivered mental health interventions.

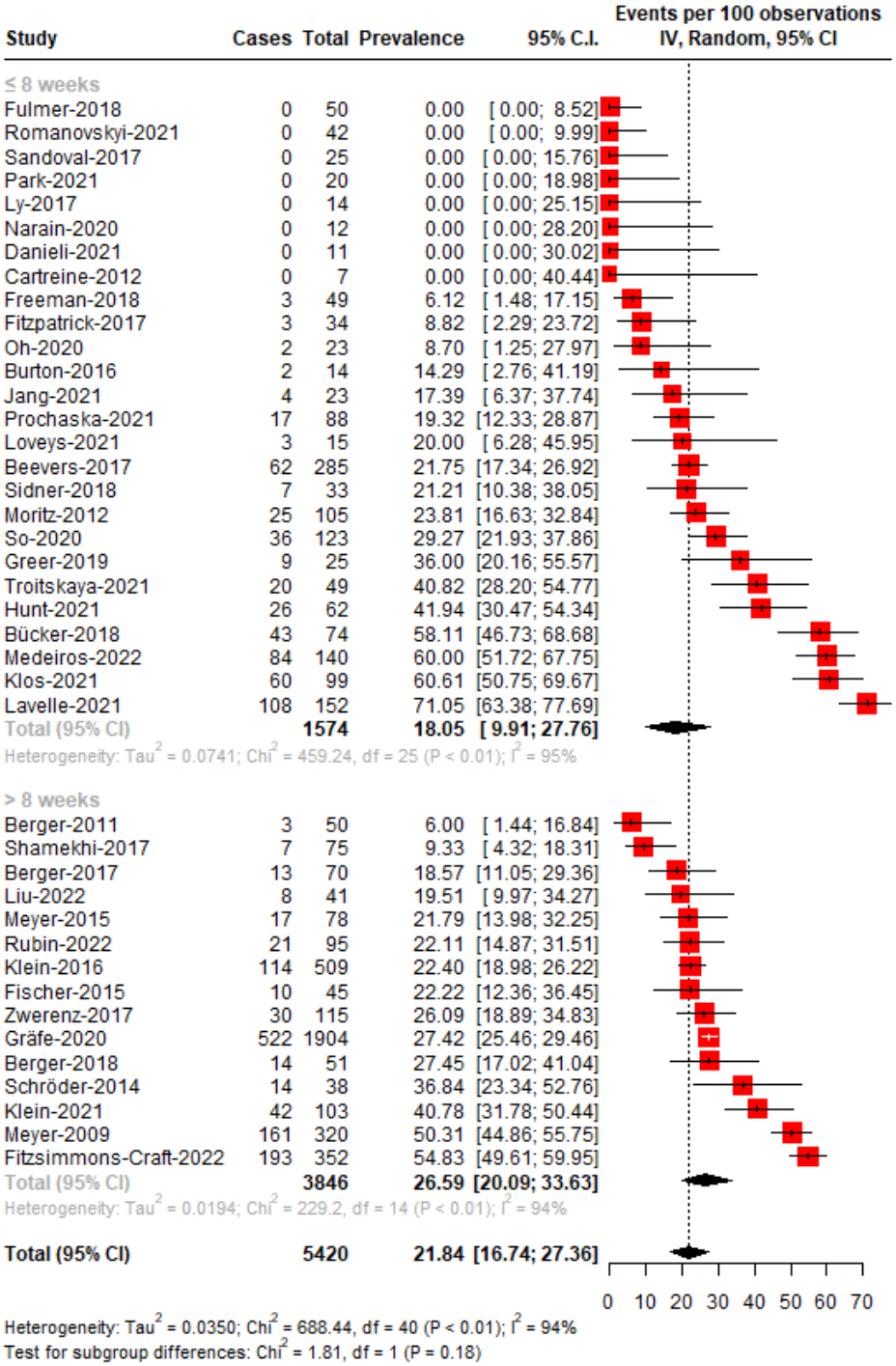

Short-Term Studies (≤8 Wk)

Overview

In the short-term studies, the overall weighted attrition rate in the intervention groups was 18.05% (95% CI 9.91%-27.76%), and there was evidence of high trial heterogeneity (I^2=94.6%, 95% CI 93.05%-95.74%). The high heterogeneity was due to high variations among the studies in terms of symptoms, types of interventions, and study populations. Of the 26 studies, 5 (19%) reported 0% attrition in the intervention group [48,49,70,73,86]. The lowest attrition rate was 6.12% (95% CI 1.48%-17.15%) [63], and the highest was 71.05% (95% CI 63.38%-77.69%) [77].

The differential attrition rate did not differ from 0% (log OR 1.22, 95% CI 0.99-1.50), indicating that the attrition rates were similar across the intervention and control groups.

XSL•FO
RenderX

The heterogeneity was low to moderate (I^2=21.89%, 95% CI 0%-54.6%). Of the 26 studies, 1 (4%) was identified as a potential outlier [15]. Removing this study from the model improved the I^2 value greatly (I^2=1.49%, 95% CI 0%-49.68%), and the differential attrition rate differed from 0% (log OR 1.27, 95% CI 1.04-1.54). This indicated that the attrition rate in the intervention group was larger than that in the control group after removing the outlying study. Multimedia Appendix 6 [14,15,47-85] shows the forest plot for the differential attrition meta-analysis [14,15,47-85].

Publication Bias for Short-Term Studies (≤8 Wk)

For the overall attrition rate, the Egger test was significant (intercept −4.70, 95% CI −8.12 to −1.28; *P*=.01), indicating possible publication bias. A closer look at the funnel plot showed missing studies toward the bottom right of the plot, which suggested possible publication bias for small sample–sized studies with high attrition rates (Figure 4A [14,15,48,49,52,56,58,62-68,70-74,76-81,84]). For the differential attrition rate, the funnel plot indicated evidence of plot symmetry, and the Egger test was not significant (intercept −4.85, 95% CI −1.56 to 0.58; *P*=.39; Figure 4B [14,15,48,49,52,56,58,62-68,70-74,76-81,84]).

Figure 4. Funnel plots and the Egger test for publication bias for (A) overall attrition and (B) differential attrition in meta-analyses of the short-term studies.

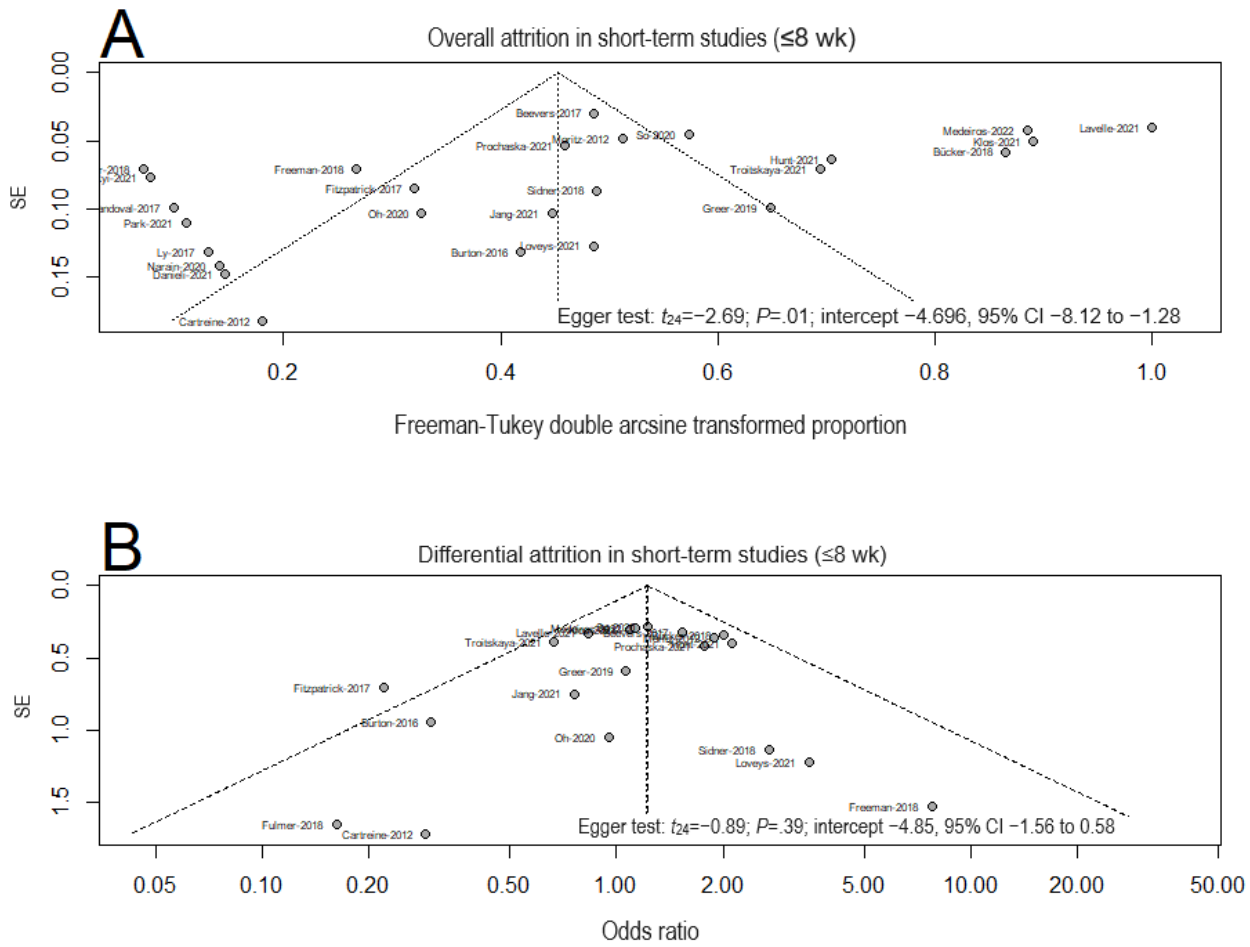

Subgroup Analyses of the Attrition Rates in Short-Term Studies (≤8 Wk)

Subgroup analyses were conducted for both overall attrition rate (Table 3) and differential attrition rate (Multimedia Appendix 6) in the CA-intervention groups compared with the control groups.

XSL•FO
RenderX

Table 3. Subgroup analyses for overall attrition rate in conversational agent (CA)–delivered mental health interventions.

Subgroups	Short-term studies (≤8 wk)				Long-term studies (>8 wk)			
	Interventions, n	Attrition rate, % (95% CI)	I^2 (%)	*P* value	Interventions, n	Attrition rate, % (95% CI)	I^2 (%)	*P* value
Risk of bias				.21				.22
High risk of bias	8	26.18 (15.69-38.13)	89.6		4	22.31 (17.85-27.10)	0	
Low risk of bias	18	14.76 (4.27-29.25)	95.6		11	28.35 (20.25-37.20)	95.4	
Funding source				.71				.38
Industry funding	9	20.43 (8.78-35.06)	92.4		4	32.36 (18.14-48.46)	97.0	
Public funding only	17	16.56 (6.35-29.79)	95.1		11	23.19 (15.92-33.65)	92.1	
Duration (wk)				.43				.68
0-4	17	15.57 (5.10-29.92)	95.3		N/A[a]	N/A	N/A	
5-8	9	23.15 (11.13-37.67)	93.4		N/A	N/A	N/A	
9-12	N/A	N/A	N/A		11	25.30 (19.35-31.75)	90.3	
>13	N/A	N/A	N/A		4	30.04 (10.75-53.88)	96.1	
Study design				.65				.19
RCT[b]	19	19.68 (10.34-30.95)	95.2		14	26.00 (19.33-33.27)	94.3	
Pilot RCT	7	12.73 (0-38.45)	91.8		1	N/A	N/A	
Type of disorder				.80				.32
Depression	6	17.89 (5.44-34.60)	91.5		11	24.03 (17.74-30.93)	91.4	
Depression and anxiety	5	7.68 (0-40.93)	97.1		N/A	N/A	N/A	
Mental well-being	9	24.29 (8.37-44.57)	94.2		N/A	N/A	N/A	
Other[c]	4	20.02 (10.61-31.36)	81.9		3	33.77 (17.01-52.90)	94.9	
With CBT[d]				.41				.65
Yes	17	21.35 (11.26-33.41)	95.1		12	27.51 (20.16-35.31)	94.5	
No	9	11.64 (0.21-33.29)	93.8		3	22.99 (7.94-42.72)	92	
With mindfulness component				.02				.27
Yes	12	30.24 (17.02-45.27)	95.5		11	23.89 (17.67-30.71)	91.5	
No	14	8.66 (0.89-21.20)	92.8		4	34.35 (18.00-52.84)	93.8	
Personalization				.63				.53
No personalization	6	17.04 (5.08-33.33)	84.7		2	38.10 (10.38-70.94)	97.1	
Minimal personalization	2	44.57 (2.50-92.50)	96.0		N/A	N/A	N/A	

XSL•FO
RenderX

Subgroups	Short-term studies (≤8 wk)				Long-term studies (>8 wk)			
	Interventions, n	Attrition rate, % (95% CI)	I^2 (%)	*P* value	Interventions, n	Attrition rate, % (95% CI)	I^2 (%)	*P* value
Substantial personalization	9	21.56 (12.52-32.12)	89.0		12	25.25 (19.16-31.85)	91.3	
Major personalization	9	10.98 (0-35.05)	96.5		1	19.51 (8.61-33.23)	N/A	
CA algorithm				.42				.49
Rule based	16	21.53 (11.72-33.12)	93.1		13	25.02 (19.30-31.19)	90.6	
AI[e] enhanced	10	13.25 (1.29-32.79)	96.2		2	37.16 (8.05-72.63)	95	
Type of CA				.14				.16
No avatar	3	33.36 (15.39-54.24)	94.36		12	29.21 (21.81-37.19)	94.6	
ECA[f]	6	8.03 (1.37-18.04)	59		2	15.35 (5.05-29.63)	80.3	
Avatar	13	20.13 (6.67-37.97)	96.4		1	19.51 (8.61-33.23)	N/A	
Not specified	4	15.15 (1.79-35.93)	86.4		N/A	N/A	N/A	
With rewards component				.74				<.001[g]
Yes	15	16.63 (9.93-27.41)	92.4		1	54.83 (49.60-60.00)	89.9	
No	11	20.07 (5.77;-39.39)	96		14	24.68 (19.20-30.59)		
Reminder				.20				.98
With reminder	14	23.96 (12.19-37.99)	94.8		7	26.42 (14.80-39.96)	95.7	
Without reminder	12	11.43 (1.76-26.27)	94.7		8	26.51 (17.85-36.16)	92.1	
Delivery channel				.02[g]				<.001[g]
Web based	4	30.74 (15.26-48.70)	91.5		11	26.96 (20.73-33.66)	90.8	
Computer- or laptop computer– or tablet computer–embedded program	6	4.72 (0.06-13.49)	61.5		1	9.33 (3.62-17.12)	N/A	
Smartphone app	10	14.32 (5.71-25.56)	87.5		1	54.83 (49.60-60.00)	N/A	
Messenging app based (meaning Slack, Facebook Messenger, or WeChat based)	6	33.27 (10.58-60.80)	96.9		1	19.51 (8.61-33.23)	N/A	
Not specified	N/A	N/A	N/A		1	22.11 (14.27-31.06)	N/A	
With blended component				.38				.02
Yes	3	19.12 (10.38-29.56)	95.1		6	18.65 (12.86-25.20)	75.4	
No	23	9.89 (0.35-26.31)	38.8		9	32.54 (23.01-42.84)	94.9	
Enrollment method				.15				.12

XSL•FO
RenderX

Subgroups	Short-term studies (≤8 wk)				Long-term studies (>8 wk)			
	Interventions, n	Attrition rate, % (95% CI)	I^2 (%)	*P* value	Interventions, n	Attrition rate, % (95% CI)	I^2 (%)	*P* value
Remote options only	14	26.63 (15.19-39.82)	95.4		9	30.18 (20.38-40.95)	95.6	
With in-person option	10	9.68 (0.15-27.14)	92		6	21.28 (16.69-26.26)	53.7	
Not specified	2	5.02 (0-33.08)	88		N/A	N/A	N/A	
Study population				.05[g]				.41
At risk	11	19.92 (1.53-29.81)	90.1		7	30.10 (17.09-44.93)	96.4	
Clinical	4	3.53 (0-13.05)	36		8	23.87 (18.55-29.61)	76.5	
General	11	22.05 (6.73-42.30)	96.1		N/A	N/A	N/A	
Session length				.61				.34
Defined session length	15	20.26 (9.93-32.79)	94.1		14	27.05 (20.28-34.38)	94.3	
User determined	11	15.31 (3.17-33.12)	95.4		1	19.51 (8.61-33.23)	N/A	
With symptom trackers component				.99				.003
Yes	14	17.77 (6.44-32.52)	93.9		6	16.36 (10.32-23.39)	64.1	
No	12	18.22 (7.21-32.44)	95.3		9	33.48 (24.91-42.62)	95.4	

[a]N/A: not applicable.

[b]RCT: randomized controlled trial.

[c]Anxiety only, panic disorder, height phobia, gambling disorder, substance abuse, attention-deficit/hyperactivity disorder, irritable bowel syndrome, eating disorder, and personality disorder.

[d]CBT: cognitive behavioral therapy.

[e]AI: artificial intelligence.

[f]ECA: embodied CA.

[g]Subgroup analyses were not significant after dropping subgroups with <5 studies.

For the overall attrition rate, there were significant differences in the attrition rates in short-term studies depending on the inclusion of mindfulness content (χ^2_1=5.1; *P*=.02). Interventions that included mindfulness content (n=12) showed a higher rate of attrition in the intervention group (30.24%, 95% CI 17.02%-42.27%) compared with interventions without mindfulness content (n=14; 8.66%, 95% CI 0.89%-21.2%). There were also significant differences depending on the population types, delivery channels, and types of disorders. However, these differences were not significant after excluding subgroups with <5 studies.

Subgroup analysis of the differential attrition rates showed that there were significantly different attrition rates in the intervention group compared with the control group depending on the subdurations (χ^2_1=5.8; *P*=.02). There were also significantly different attrition rates between study populations (χ^2_2=9.3; *P*=.009), and the types of controls (χ^2_1=4.7; *P*=.03). The relative risks of attrition for studies that lasted between 5 and 8 weeks were significantly different (n=9; log OR 1.61, 95% CI 1.22-2.13) compared with studies that lasted <5 weeks (n=17; log OR 0.99, 95% CI 0.75-1.31). Studies that recruited populations considered to be at risk (n=9) showed significantly higher attrition rates in the intervention group than in the control group (log OR 1.65, 95% CI 1.26-2.15) when compared with general populations (n=7; log OR 0.96, 95% CI 0.71-1.30) and clinical populations (n=3; log OR 0.47, 95% CI 0.13-1.66). The subgroup analysis was still significant when compared between the general population and the group considered to be at risk only (χ^2_1=6.9; *P*=.03). Finally, there were higher attrition rates in the intervention studies than in studies that used waitlist controls (n=11; log OR 1.52 95% CI 1.18-1.95) than those that used active controls (n=7; log OR 0.96, 95% CI 0.69-1.54). Only 1 (2%) of the 41 studies used treatment as usual as the control group [67]. No other comparisons were significant.

Long-Term Studies (>8 Wk)

Overview

The weighted attrition rate for the intervention group in long-term studies was 26.59% (95% CI 20.09%-33.63%), and

XSL•FO
RenderX

there was evidence of high trial heterogeneity (I^2=93.89%, 95% CI 91.43%-95.64%). The lowest relative attrition rate was 6% (95% CI 1.44%-16.84%) [51], and the highest was 54.83% (95% CI 49.61%-59.95%) [77].

The differential attrition rate differed from 0% (log OR 1.33, 95% CI 1.08-1.65), indicating that the attrition rates in the intervention group were higher than those in the control group. The heterogeneity was moderate (I^2=49.43%, 95% CI 0.083%-72.11%). However, 1 (2%) of the 41 studies was identified as a potential outlier [50]. Removing this study from the model improved the I^2 value greatly (I^2=24.22%, 95% CI 0%-59.80%), and the differential attrition rate still differed from 0% (log OR 1.22, 95% CI 1.05-1.42); again, this indicated that the attrition rates in the intervention group were significantly larger than those in the control group even after removing the outlying study. The outlying study [50] used a weighted randomization method in which 80% of the participants were allocated to the intervention group. The subgroup analyses were conducted without the outlier because the study seemed to explain >20% of the heterogeneity in the model. However, sensitivity analyses conducted with and without the outlying study did not change the results of the subgroup analysis.

Publication Bias in Long-Term Studies (>8 Wk)

For the overall attrition rate, the funnel plot indicated evidence of plot asymmetry, but the Egger test was not significant (intercept −0.79, 95% CI −4.34 to 2.75; *P*=.67), suggesting a low likelihood of publication bias. For the differential attrition rate, the funnel plot indicated evidence of plot symmetry, and the Egger test was not significant (intercept 0.46, 95% CI −0.66 to 1.58; *P*=.43; Figure 5 [14,15,47-85]).

Figure 5. Funnel plots and the Egger test for publication bias for (A) overall attrition and (B) differential attrition in meta-analyses of the long-term studies.

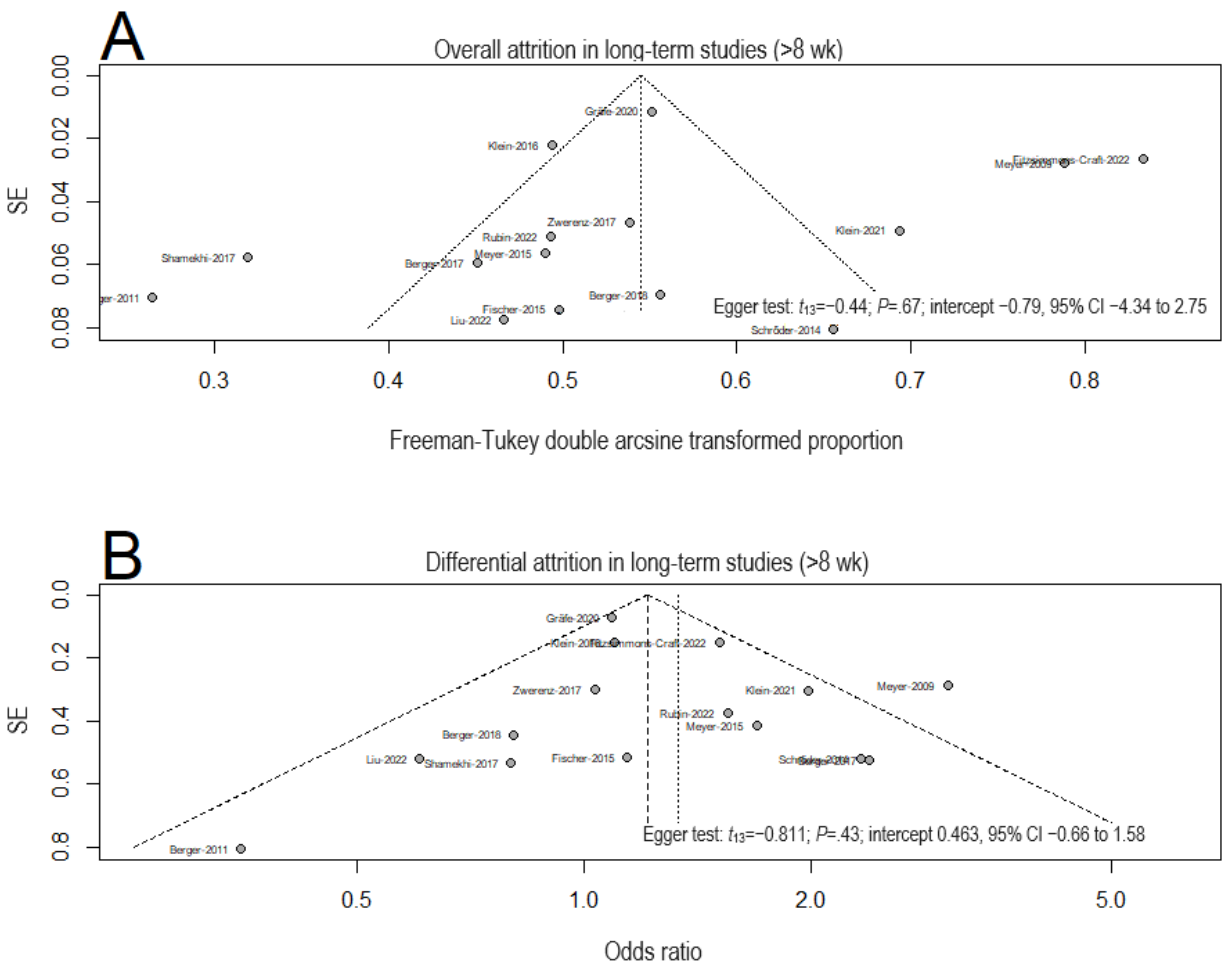

Subgroup Analyses of the Attrition Rates in Long-Term Studies (>8 Wk)

For overall attrition, there were significant differences in the attrition rates in the intervention groups of long-term studies that had a blended design (χ^2_1=4.7; *P*=.03) and included symptom trackers or mood monitoring (χ^2_1=9.0; *P*=.003). Interventions that included blended designs (n=6) showed a significantly lower attrition rates (20.41%, 95% CI 15.49%-25.81%) than those without (n=9; 33.3%, 95% CI 23.01%-44.44%). Interventions that included symptom trackers (n=6) showed a significantly lower attrition rates (16.36%, 95% CI 10.32%-23.39%) than those without (n=9; 33.48%, 95% CI

XSL•FO
RenderX

24.91%-42.62%; Table 3). No other comparisons were significant.

The differential attrition rates were significantly different across studies that included mindfulness content compared with those without (χ^2_1=5.0; *P*=.03) and studies that targeted depression symptoms compared with those that targeted other mental health symptoms (χ^2_1=8.6; *P*=.003). Studies without mindfulness intervention showed higher attrition rates in the intervention groups than in the controls (n=10; log OR 1.56, 95% CI 1.25%-1.96%) compared with studies that included mindfulness content (n=4; log OR 1.11, 95% CI 1.05%-1.42%). Studies that targeted depression symptoms specifically showed relatively similar attrition rates in both intervention and control groups (n=10; log OR 1.09, 95% CI 0.96%-1.22%) compared with studies that targeted other mental health symptoms such as gambling disorder and substance abuse (n=4; log OR 1.63, 95% CI 1.28%-2.08%). No other comparisons were significant.

Exploratory Subgroup Analyses

The overall weighted attrition rate for the intervention group in all included studies was 21.84% (95% CI 16.74%-27.36%; I^2=94%). Exploratory subgroup analyses using the prespecified subgroup analyses were conducted to explain the heterogeneity of the included studies regardless of the duration of the interventions. We have included the full findings in Multimedia Appendix 7 and only highlight findings that differed from our prespecified analysis here. For overall attrition, as in our prespecified analysis, there were significant differences in the attrition rates in the intervention groups depending on the inclusion of mindfulness content (χ^2_1=4.0; *P*=.05). However, we did not find significant differences in the inclusion of blended support compared with nonblended intervention and the inclusion of symptom tracker compared with intervention without symptom tracker. In addition, we found significant differences depending on the type of CA used (χ^2_3=13.1; *P*=.005), the CA delivery channel (χ^2_4=21.3; *P*<.001), and the study enrollment method (χ^2_2=7.4; *P*=.02). Studies that did not use any identifiable avatar reported the highest rate of attrition (n=15; 30%, 95% CI 23.44%-37.01%), followed by studies that did not specify the use of an avatar or a visual representation of the CA (n=14; 20.12%, 95% CI 7.29%-36.82%), studies that used a static avatar (n=4; 15.15%, 95% CI 1.79%-35.93%), and studies that used an ECA (n=8; 10.3%, 95% CI 4.29%-18.04%). Interventions that were delivered via messaging app (meaning "Slack, Facebook Messenger, or WeChat" based) showed the highest rate of attrition (n=7; 31.19%, 95% CI 10.68%-56.28%), followed by those delivered by web-based applications (n=15; 27.9%, 95% CI 22.35%-33.78%), and those delivered by stand-alone smartphone apps (n=11; 17.36%, 95% CI 6.54%-31.48%). CAs installed on a computer, a laptop computer, or a tablet computer showed the lowest rate of attrition (n=7; 5.61%, 95% CI 1.09%-12.3%). Finally, studies that offered remote onboarding only (n=23) showed a higher attrition rate (28.42%, 95% CI 21.30%-36.1%) than studies that offered an in-person onboarding process (n=16; 15.01%, 95% CI 8.46%-22.82%).

The differential attrition rate differed from 0% (log OR 1.28, 95% CI 1.10-1.48; I^2=34.6%), indicating that the participants who received CA interventions were more likely to attrit than the control group participants.

For differential attrition, our findings mostly concurred with our prespecified analysis. Unlike in the prespecified analysis, there was a significant difference for studies that included symptom trackers (χ^2_1=5.0; *P*=.02). Studies that included symptom trackers (n=17) showed relatively lower attrition in the intervention group than in the control group (log OR 1.02, 95% CI 0.81-1.29) compared with studies without symptom trackers (n=18; log OR 1.44, 95% CI 1.19-1.74).

Additional Factors Associated With Attrition

Of the 41 studies, 16 (39%) assessed the association of different study features on participants' attrition (short-term studies: n=8, 50%; long-term studies: n=8, 50%). We grouped the findings into two categories: (1) demographic predictors and (2) baseline measurement predictors or symptom severity.

The associations between participants' demographics and attrition were assessed and reported in 10 (63%) of the 16 studies (short-term studies: n=4, 40%; long-term studies: n=6, 60%). Only 1 (25%) of the 4 short-term studies found age to be significantly associated with attrition. Participants who dropped out were found to be significantly younger than those who completed the whole intervention [62]. Other demographics-related factors assessed in these 10 studies were not associated with attrition, including sex, years of education, marital status, employment, actively receiving therapy or medication, and current diagnosis, for both short-term and long-term studies.

Of the 41 studies, 12 (29%) explored the association between baseline predictors or symptom severity and attrition (short-term studies: n=6, 50%; long-term studies: n=6, 50%). More severe baseline symptoms were associated with attrition for some of the short-term studies (2/6, 33%) but not for the long-term studies. Higher anxiety-related symptoms measured using the General Anxiety Disorder-7 questionnaire [62] and the Visceral Sensitivity Index [68] were significantly related to attrition. Other factors found to be associated with higher attrition included lower quality of life measured using the World Health Organization Quality of Life Brief Questionnaire [52], higher Fear of Food Questionnaire score [68], higher severity of gambling pathology measured using the Pathological Gambling Adaptation of the Yale-Brown Obsessive-Compulsive Scale [62], and lower self-esteem [81]. Of the 10 studies, 1 (10%) reported that participants who attrited significantly greater positive affect compared to those who completed the study using the Positive and Negative Affect Schedule [77].

Discussion

Principal Findings

To our knowledge, this systematic review and meta-analysis is the first to examine attrition in RCTs of CA interventions for mental health. A total of 41 RCTs met the inclusion criteria. Our findings showed that approximately one-fifth of the

XSL•FO
RenderX

participants (18.05%) dropped out in short-term studies, and approximately a quarter (26.59%) dropped out in long-term studies. Participants who received CA interventions were more likely to attrit than the control group participants for both long-term and short-term studies. Several study-level moderators were identified. For short-term studies, higher overall attrition rates were found in intervention groups that included mindfulness content and those that included participants from the general population. Compared with the control group participants, participants in the short-term CA interventions were also more likely to attrit in studies that lasted >1 month, those that included participants considered to be at risk, and studies in which intervention group participants were compared against waitlist controls as opposed to alternative active controls. For long-term studies, higher overall attrition rates were found in interventions that did not include human support and studies that did not include a symptom tracker. Exploratory subgroup analyses conducted on all included studies regardless of the study duration largely supported the analysis except for the inclusion of blended support. In addition, exploratory analyses found that studies that used an ECA, delivered via a computer or smartphones, and provided in-person enrollment options were associated with lower attrition rates.

Comparison With Prior Research

Overall Attrition

The overall attrition rates for short-term and long-term studies in our review are lower than the attrition estimates of short-term and long-term studies on smartphone-delivered interventions (24.1% and 35.5%, respectively) [13] and electronic mental health treatments for eating disorders (21%) [87]. Our findings are comparable with attrition rates in studies evaluating face-to-face mental health treatments such as interpersonal psychotherapy (20.6%) [88], individual psychotherapy for depressive disorders (19.9%) [89], and generalized anxiety disorders (17%) [90]. When focusing only on studies evaluating smartphone-delivered CAs in our review (n=11), we found that only 14.32% of the participants dropped out of the short-term studies and 17.36% dropped out of all included studies; these rates are lower than the estimated attrition rate previously reported for smartphone-delivered mental health interventions [13]. Taken together, the delivery channel may indirectly influence study attrition [13]. Although we found lower attrition in studies that were delivered via programs installed on a computer or a laptop computer compared with other delivery channels, these studies [48,58,59,63,79] were conducted in a laboratory setting compared with the participants' environment, which might influence the retention rate.

Factors Associated With Attrition

CA interventions used as adjuvants of psychotherapy sessions or with human support may aid in retaining participants, particularly in long-term studies. A closer look at the long-term studies that included human support revealed that most of these studies (6/15, 40%) used the CA interventions as an adjunctive tool with no specific instruction given to the primary therapist on how to support participants' journey through the CA interventions. This suggests that the presence of the therapist alone could suffice to retain participation in the study over a longer period. Although participants may be staying for the primary therapist and not engaging with the CA intervention, a study with both therapist-guided and unguided groups found no significant differences in the time spent with the CA intervention [51]. It is also possible that participants may have consulted their primary therapist about the CA interventions, which was not reported by the studies. A recent scoping review reported that many studies that included human support did not consistently report the type of support provided by humans [91]. This was similarly observed in our study because none of the included studies mentioned whether the therapist discussed the CA intervention during the participants' routine sessions. Furthermore, it is also possible that participants within clinical settings are more likely to stay with the intervention, as was found in our results, unrelated to the blended support provided. Therefore, it is difficult to draw conclusions regarding the impact of using CA interventions adjunctively in terms of retaining participants within the study. However, this finding should be taken with caution because we found that there were no significant differences in the attrition rates between studies that included blended and nonblended approaches when we included all studies regardless of the study duration. This may be because fewer studies (9/41, 22%) included the blended approach overall.

In terms of the intervention content, short-term studies that included mindfulness content had higher overall attrition rates than those without mindfulness content. This was contrary to 2 previous meta-analyses that showed that the inclusion of the mindfulness component had no impact on attrition rates [13,28]. However, the attrition rate was similar to that in a systematic review of self-guided mindfulness interventions that reported an unweighted mean attrition rate of 31% (range 0%-79%) [92]. Future interventions may need to pay closer attention to participants' engagement when including mindfulness content as part of a CA intervention. Factors such as using symptom trackers and personalized feedback may increase the engagement rate [93]. This is aligned with our findings and prior meta-analyses that suggest that including feedback may improve the retention rate [28].

Interestingly, our results also found relatively lower differential attrition rates in the intervention groups of long-term studies that included mindfulness compared with studies without. However, this finding was not replicated when we analyzed all studies regardless of the study duration. A recent review suggested that longer mindfulness practice sessions may be associated with the development of mindfulness skills [92]. Therefore, this result should be interpreted with caution because the relationship between the frequency and the duration of mindfulness practice sessions is still unclear [92].

Our study also found that including any form of visual representation of a CA may be associated with lower attrition rates compared with no visualization at all. This is aligned with many studies on the design of CAs that stressed the importance of design to create positive perceptions of the CA [94]. However, a recent scoping review reported that visual representation of the CA showed mixed and no association with subjective user experience [95]. A closer look at the studies included in this review suggested that most of them (35/41,

XSL•FO
RenderX

85%) lasted only 1 session [95]. It is possible that not having any CA visualization could affect user experience over time as alluded to by our findings. Future studies should explore the relationship between CA visual representation and user experience and study adherence over a longer duration.

In terms of the study sample population, sample populations considered to be at risk were more likely to attrit than samples from general and clinical sample populations. However, other meta-analyses of digital interventions for mental health issues found no difference in attrition rates across sample populations [13,28]. This finding is difficult to interpret because there could be multiple factors that may affect this relationship, such as symptom severity and other demographic factors not included in our analysis. Future studies may explore this relationship further to better understand this association.

Factors Not Associated With Attrition

Providing monetary incentives did not affect the attrition rates significantly. This finding is similar to that of a previous meta-analysis focusing on smartphone apps specifically for depressive symptoms [28] but contrary to that of a study focusing on smartphone interventions for mental health in general [13]. Time spent within the study may be a greater driver for attrition, as can be seen in the higher attrition rates for longer-term studies found in our results. However, research on the impact of monetary incentives on participants' retention in digital health interventions is still in its infancy [96]. More needs to be done to understand how monetary incentives affect participants' retention as well as effective engagement in the intervention.

Finally, several features such as providing reminders and the level of personalization provided by the CA also did not affect attrition rates. This is noteworthy because a personalized intervention that is responsive to users' inputs is related to better engagement with the intervention [93]. There may be further nuances between attrition and effective engagement; for instance, factors that lower the risk of attrition might not be directly related to factors that promote study adherence [28].

Strengths and Limitations

We conducted a comprehensive literature search that included both peer-reviewed databases and gray literature sources; in addition, we conducted backward and forward citation searches. As this is a nascent area, we prioritized the sensitivity of our search to capture the various representations of CAs.

However, our study has some limitations. First, some unpublished literature presented at niche conferences and meetings may have been omitted. In addition, some studies might have escaped our search strategy, as evidenced by the inclusion of *Deprexis*, *Deprexis*-based interventions, and ePST intervention that did not explicitly mention terms related to *conversational agent* in the studies concerned. Second, the heterogeneity of the mental health conditions addressed by the CA intervention made it difficult to generalize the findings to a specific disorder. The recommendations provided here should be taken as a general suggestion to improve retention rates in CA interventions for mental health but not for a specific mental health disorder. Third, our results indicated possible publication bias in short-term studies. A closer look at the funnel plot suggested a lack of small sample–sized studies (n<20) with high attrition rates in the intervention groups. It is possible that these studies were not being published because they could be too experimental and small scaled. However, it is possible that the findings are reported elsewhere at niche conferences and internal sharing, which may have been omitted based on our search strategy. Fourth, our results may not directly translate into understanding factors that may increase engagement with the interventions. Although we recognize that engagement is interlinked with attrition [97], the lack of consensus and reporting of engagement metrics limits our understanding of this relationship. A recent review identified several patient-, intervention-, and system-level factors associated with engagement [98]. However, many of these associations were not consistent across different digital mental health interventions, and there was poor consistency in the reporting of the metrics. We echoed others' recommendations to include and standardize the reporting of engagement metrics to better understand the indicators of nonuse attrition [41,93,97,98]. Our findings can inform future researchers of the potential factors for attrition in CA interventions. These may serve as a basis to make informed decisions on the sample size required or to conduct further studies on the specific mechanisms that may or may not motivate attrition. Fifth and last, our subgroup analyses for all studies regardless of the intervention duration were exploratory post hoc analyses and should be interpreted as such.

Implications and Recommendations

Several implications and recommendations emerged from our findings. First, researchers may want to account for the attrition of approximately one-third of the participants when designing RCTs involving CA interventions. This number may need to be further adjusted depending on the sample population, delivery modes, and comparison group used in the intervention to minimize the potential threats to the external and internal validity of studies that evaluate the efficacy of CA interventions for mental health. Second, researchers may want to consider including active controls in RCTs. Our results and the findings from other similar reviews on attrition in digital health research [13,28] suggested that comparing digital interventions with waitlist controls might not be the ideal way because participants in the comparison group may be more motivated to remain in the study than those in the intervention group [13]. Control interventions consisting of periodic mood assessments via an app or a nonconversational version of the app may be more appropriate for the assessment of CA effectiveness. Third, the inclusion of a visual representation of the CA may help create a more positive perception of the CA and reduce attrition. A recent review suggested that design considerations such as having a humanlike representation and having medical attire for the CA may be helpful to reduce attrition [95]. Fourth, CA interventions should be adjunctive to ongoing therapy sessions. Although the association between attrition and the use of blended support may be inconclusive, the use of CA interventions may further enrich participants' experience between sessions and provide ongoing support to practice the skills learned during regular sessions. Fifth and last, clinicians interested in implementing CA interventions in their practice

XSL•FO
RenderX

should be aware of the high attrition rate and should closely monitor patients' progress within their practice.

Conclusions

According to our findings, at least one-fifth of the intervention group participants in RCTs of CA interventions will drop out of the trials within the intervention period. This attrition rate is comparable with that of face-to-face mental health interventions and less than that of other digital health interventions. Furthermore, intervention group participants were more likely to drop out than control group participants. Attrition was lower in shorter-term studies, studies that included participants considered to be at risk, and studies in which intervention group participants were compared against waitlist controls as opposed to alternative active controls. In addition, not including mindfulness content or symptom trackers was found to be associated with a smaller risk of attrition. Future studies may benefit from delivering CA interventions in a blended setting, with symptom screening; comparing the CA interventions against active controls such as symptom tracking only without the CA component; and including a visual representation of the CA to reduce attrition rates in the intervention group.

Acknowledgments

The authors would like to acknowledge Ms Yasmin Ally, Lee Kong Chian School of Medicine librarian, for her assistance in translating and executing the search strategy. The authors would also like to acknowledge Ms Nileshni Fernando for her assistance in the data extraction and risk-of-bias assessment. This research was supported by the Singapore Ministry of Education under the Singapore Ministry of Education Academic Research Fund Tier 1 (RG36/20). The research was conducted as part of the Future Health Technologies program, which was established collaboratively between ETH Zürich and the National Research Foundation, Singapore. This research was also supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise program.

Data Availability

This systematic review and meta-analysis does not contain primary data. The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request. The data will be made available by emailing the corresponding author.

Authors' Contributions

LTC conceptualized the study and provided supervision at all steps of the research. LTC and AIJ designed the study. AIJ, XL, and GS extracted the data. AIJ conducted the analysis and wrote the original manuscript. LTC, LM, XL, GS, and YLT provided a critical review of the manuscript. All authors approved the final version of the manuscript and take accountability for all aspects of the work.

Conflicts of Interest

LTC is an associate editor of JMIR Medical Education. The other authors reported no conflict of interest.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 198 KB - jmir_v26i1e48168_app1.pdf]

Multimedia Appendix 2
Criteria for study selection and extended meta-analysis method.
[DOCX File , 51 KB - jmir_v26i1e48168_app2.docx]

Multimedia Appendix 3
Search strategy for PubMed.
[DOCX File , 22 KB - jmir_v26i1e48168_app3.docx]

Multimedia Appendix 4
Tables of excluded studies.
[DOCX File , 184 KB - jmir_v26i1e48168_app4.docx]

Multimedia Appendix 5
Risk-of-bias assessment of the included studies.
[DOCX File , 242 KB - jmir_v26i1e48168_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Differential attrition analysis for the short-term and long-term studies.
[DOCX File , 87 KB - jmir_v26i1e48168_app6.docx]

Multimedia Appendix 7
Extended results for the exploratory subgroup analysis.
[DOCX File , 50 KB - jmir_v26i1e48168_app7.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Vigo D, Jones L, Atun R, Thornicroft G. The true global disease burden of mental illness: still elusive. Lancet Psychiatry 2022 Feb;9(2):98-100. [doi: 10.1016/S2215-0366(22)00002-5] [Medline: 35026138]
3. Silverman AL, Teachman BA. The relationship between access to mental health resources and use of preferred effective mental health treatment. J Clin Psychol 2022 Jun;78(6):1020-1045. [doi: 10.1002/jclp.23301] [Medline: 34993965]
4. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017 Feb;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
5. Chong SA, Abdin E, Vaingankar JA, Kwok KW, Subramaniam M. Where do people with mental disorders in Singapore go to for help? Ann Acad Med Singap 2012 Apr;41(4):154-160 [FREE Full text] [Medline: 22635279]
6. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
7. Ku BS, Li J, Lally C, Compton MT, Druss BG. Associations between mental health shortage areas and county-level suicide rates among adults aged 25 and older in the USA, 2010 to 2018. Gen Hosp Psychiatry 2021 May;70:44-50 [FREE Full text] [doi: 10.1016/j.genhosppsych.2021.02.001] [Medline: 33714795]
8. Oladeji BD, Gureje O. Brain drain: a challenge to global mental health. BJPsych Int 2016 Aug;13(3):61-63 [FREE Full text] [doi: 10.1192/s2056474000001240] [Medline: 29093905]
9. Clement S, Schauman O, Graham T, Maggioni F, Evans-Lacko S, Bezborodovs N, et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. Psychol Med 2015 Jan;45(1):11-27. [doi: 10.1017/S0033291714000129] [Medline: 24569086]
10. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019;2:118. [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
11. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
12. Linardon J, Hindle A, Brennan L. Dropout from cognitive-behavioral therapy for eating disorders: a meta-analysis of randomized, controlled trials. Int J Eat Disord 2018 May 01;51(5):381-391. [doi: 10.1002/eat.22850] [Medline: 29493805]
13. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
14. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
15. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
16. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
17. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
18. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie 2019 Aug 20;32(Suppl. 1):64-76. [doi: 10.1159/000501812]

XSL•FO
RenderX

19. Lim SM, Shiau CW, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
20. Beilharz F, Sukunesan S, Rossell SL, Kulkarni J, Sharp G. Development of a positive body image chatbot (KIT) with young people and parents/carers: qualitative focus group study. J Med Internet Res 2021 Jun 16;23(6):e27807 [FREE Full text] [doi: 10.2196/27807] [Medline: 34132644]
21. Salamanca-Sanabria A, Jabir AI, Lin X, Alattas A, Kocaballi AB, Lee J, et al. Exploring the perceptions of mHealth interventions for the prevention of common mental disorders in university students in Singapore: qualitative study. J Med Internet Res 2023 Mar 20;25:e44542 [FREE Full text] [doi: 10.2196/44542] [Medline: 36939808]
22. Scholten MR, Kelders SM, Van Gemert-Pijnen JE. Self-guided web-based interventions: scoping review on user needs and the potential of embodied conversational agents to address them. J Med Internet Res 2017 Nov 16;19(11):e383 [FREE Full text] [doi: 10.2196/jmir.7351] [Medline: 29146567]
23. Nißen M, Rüegger D, Stieger M, Flückiger C, Allemand M, V Wangenheim F, et al. The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: development of a design codebook and web-based study. J Med Internet Res 2022 Apr 27;24(4):e32630 [FREE Full text] [doi: 10.2196/32630] [Medline: 35475761]
24. Miner AS, Milstein A, Hancock JT. Talking to machines about personal mental health problems. JAMA 2017 Oct 03;318(13):1217-1218. [doi: 10.1001/jama.2017.14151] [Medline: 28973225]
25. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
26. What Works Clearinghouse™ standards handbook, version 4.1. Institute of Education Sciences. URL: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf [accessed 2023-12-07]
27. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. JAMA 2015 Nov 10;314(18):1966-1967 [FREE Full text] [doi: 10.1001/jama.2015.15281] [Medline: 26547468]
28. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
29. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
30. Bendig E, Erb B, Meißner D, Bauereiß N, Baumeister H. Feasibility of a software agent providing a brief intervention for self-help to uplift psychological wellbeing ("SISU"). A single-group pretest-posttest trial investigating the potential of SISU to act as therapeutic agent. Internet Interv 2021 Feb 24;24:100377 [FREE Full text] [doi: 10.1016/j.invent.2021.100377] [Medline: 33816127]
31. Sanders GJ, Cooke C, Gately P. Exploring reasons for attrition among vulnerable and under-served sub-groups across an online integrated healthy lifestyles service during COVID-19. SAGE Open Med 2021 Oct 22;9:20503121211054362 [FREE Full text] [doi: 10.1177/20503121211054362] [Medline: 34707867]
32. Goldberg SB, Bolt DM, Davidson RJ. Data missing not at random in mobile health research: assessment of the problem and a case for sensitivity analyses. J Med Internet Res 2021 Jun 15;23(6):e26749 [FREE Full text] [doi: 10.2196/26749] [Medline: 34128810]
33. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.3. London, UK: The Cochrane Collaboration; 2022.
34. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
35. Westgate MJ. revtools: an R package to support article screening for evidence synthesis. Res Synth Methods 2019 Dec;10(4):606-614. [doi: 10.1002/jrsm.1374] [Medline: 31355546]
36. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021 Feb 01;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
37. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
38. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-On Guide. Boca Raton, FL: CRC Press; Sep 14, 2021.
39. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta-Analysis. Hoboken, NJ: Wiley; Mar 12, 2009.
40. Martinengo L, Jabir AI, Goh WW, Lo NY, Ho MH, Kowatsch T, et al. Conversational agents in health care: scoping review of their behavior change techniques and underpinning theory. J Med Internet Res 2022 Oct 03;24(10):e39243 [FREE Full text] [doi: 10.2196/39243] [Medline: 36190749]
41. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548 [FREE Full text] [doi: 10.2196/44548] [Medline: 37074762]

XSL•FO
RenderX

42. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 06;6(1):245 [FREE Full text] [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]
43. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google scholar in evidence reviews and its applicability to grey literature searching. PLoS One 2015 Sep 17;10(9):e0138237 [FREE Full text] [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]
44. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: a tool for transparent and efficient forward and backward citation chasing in systematic searching. Res Synth Methods 2022 Jul;13(4):533-545. [doi: 10.1002/jrsm.1563] [Medline: 35472127]
45. McGuinness LA, Higgins JP. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
46. Twomey C, O'Reilly G, Meyer B. Effectiveness of an individually-tailored computerised CBT programme (Deprexis) for depression: a meta-analysis. Psychiatry Res 2017 Oct;256:371-377. [doi: 10.1016/j.psychres.2017.06.081] [Medline: 28686935]
47. Meyer B, Bierbrodt J, Schröder J, Berger T, Beevers CG, Weiss M, et al. Effects of an internet intervention (Deprexis) on severe depression symptoms: randomized controlled trial. Internet Interv 2015 Mar;2(1):48-59. [doi: 10.1016/j.invent.2014.12.003]
48. Cartreine JA, Locke SE, Buckey JC, Sandoval L, Hegel MT. Electronic problem-solving treatment: description and pilot study of an interactive media treatment for depression. JMIR Res Protoc 2012 Sep 25;1(2):e11 [FREE Full text] [doi: 10.2196/resprot.1925] [Medline: 23611902]
49. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
50. Meyer B, Berger T, Caspar F, Beevers CG, Andersson G, Weiss M. Effectiveness of a novel integrative online treatment for depression (Deprexis): randomized controlled trial. J Med Internet Res 2009 May 11;11(2):e15 [FREE Full text] [doi: 10.2196/jmir.1151] [Medline: 19632969]
51. Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther 2011;40(4):251-266. [doi: 10.1080/16506073.2011.616531] [Medline: 22060248]
52. Moritz S, Schilling L, Hauschildt M, Schröder J, Treszl A. A randomized controlled trial of internet-based therapy in depression. Behav Res Ther 2012 Aug;50(7-8):513-521. [doi: 10.1016/j.brat.2012.04.006] [Medline: 22677231]
53. Schröder J, Brückner K, Fischer A, Lindenau M, Köther U, Vettorazzi E, et al. Efficacy of a psychological online intervention for depression in people with epilepsy: a randomized controlled trial. Epilepsia 2014 Dec;55(12):2069-2076 [FREE Full text] [doi: 10.1111/epi.12833] [Medline: 25410633]
54. Fischer A, Schröder J, Vettorazzi E, Wolf OT, Pöttgen J, Lau S, et al. An online programme to reduce depression in patients with multiple sclerosis: a randomised controlled trial. Lancet Psychiatry 2015 Mar;2(3):217-223. [doi: 10.1016/S2215-0366(14)00049-2] [Medline: 26359900]
55. Klein JP, Berger T, Schröder J, Späth C, Meyer B, Caspar F, et al. Effects of a psychological internet intervention in the treatment of mild to moderate depressive symptoms: results of the EVIDENT study, a randomized controlled trial. Psychother Psychosom 2016;85(4):218-228 [FREE Full text] [doi: 10.1159/000445355] [Medline: 27230863]
56. Beevers CG, Pearson R, Hoffman JS, Foulser AA, Shumake J, Meyer B. Effectiveness of an internet intervention (Deprexis) for depression in a united states adult sample: a parallel-group pragmatic randomized controlled trial. J Consult Clin Psychol 2017 Apr;85(4):367-380. [doi: 10.1037/ccp0000171] [Medline: 28230390]
57. Berger T, Urech A, Krieger T, Stolz T, Schulz A, Vincent A, et al. Effects of a transdiagnostic unguided Internet intervention ('velibra') for anxiety disorders in primary care: results of a randomized controlled trial. Psychol Med 2017 Jan;47(1):67-80 [FREE Full text] [doi: 10.1017/S0033291716002270] [Medline: 27655039]
58. Sandoval LR, Buckey JC, Ainslie R, Tombari M, Stone W, Hegel MT. Randomized controlled trial of a computerized interactive media-based problem solving treatment for depression. Behav Ther 2017 May;48(3):413-425 [FREE Full text] [doi: 10.1016/j.beth.2016.04.001] [Medline: 28390503]
59. Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting group medical visits with conversational agents for stress management behavior change. In: Proceedings of the 12th International Conference, PERSUASIVE 2017. 2017 Presented at: 12th International Conference, PERSUASIVE 2017; April 4–6, 2017; Amsterdam, The Netherlands. [doi: 10.1007/978-3-319-55134-0_5]
60. Zwerenz R, Becker J, Knickenberg RJ, Siepmann M, Hagen K, Beutel ME. Online self-help as an add-on to inpatient psychotherapy: efficacy of a new blended treatment approach. Psychother Psychosom 2017;86(6):341-350. [doi: 10.1159/000481177] [Medline: 29131090]
61. Berger T, Krieger T, Sude K, Meyer B, Maercker A. Evaluating an e-mental health program ("deprexis") as adjunctive treatment tool in psychotherapy for depression: results of a pragmatic randomized controlled trial. J Affect Disord 2018 Feb;227:455-462. [doi: 10.1016/j.jad.2017.11.021] [Medline: 29154168]

XSL•FO
RenderX

62. Bücker L, Bierbrodt J, Hand I, Wittekind C, Moritz S. Effects of a depression-focused internet intervention in slot machine gamblers: a randomized controlled trial. PLoS One 2018 Jun 08;13(6):e0198859 [FREE Full text] [doi: 10.1371/journal.pone.0198859] [Medline: 29883479]
63. Freeman D, Reeve S, Robinson A, Ehlers A, Clark D, Spanlang B, et al. Virtual reality in the assessment, understanding, and treatment of mental health disorders. Psychol Med 2017 Oct;47(14):2393-2400 [FREE Full text] [doi: 10.1017/S003329171700040X] [Medline: 28325167]
64. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
65. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64. [doi: 10.2196/mental.9782] [Medline: 30545815]
66. Sidner CL, Bickmore T, Nooraie B, Rich C, Ring L, Shayganfar M, et al. Creating new technologies for companionable agents to support isolated older adults. ACM Trans Interact Intell Syst 2018 Jul 24;8(3):1-27. [doi: 10.1145/3213050]
67. Burton C, Szentagotai Tatar A, McKinstry B, Matheson C, Matu S, Moldovan R, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare 2016 Sep;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]
68. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152 [FREE Full text] [doi: 10.2196/26152] [Medline: 33872182]
69. Gräfe V, Moritz S, Greiner W. Health economic evaluation of an internet intervention for depression (deprexis), a randomized controlled trial. Health Econ Rev 2020 Jun 16;10(1):19 [FREE Full text] [doi: 10.1186/s13561-020-00273-0] [Medline: 32548724]
70. Narain J, Quach T, Davey M, Park HW, Breazeal C, Picard R. Promoting wellbeing with sunny, a chatbot that facilitates positive messages within social groups. In: Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI EA '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3334480.3383062]
71. Oh J, Jang S, Kim H, Kim JJ. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
72. So R, Furukawa TA, Matsushita S, Baba T, Matsuzaki T, Furuno S, et al. Unguided chatbot-delivered cognitive behavioural intervention for problem gamblers through messaging app: a randomised controlled trial. J Gambl Stud 2020 Dec;36(4):1391-1407. [doi: 10.1007/s10899-020-09935-4] [Medline: 32162075]
73. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. JMIR Form Res 2021 Dec 01;5(12):e30053. [doi: 10.2196/30053] [Medline: 34855607]
74. Jang S, Kim JJ, Kim SJ, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. Int J Med Inform 2021 Jun;150:104440. [doi: 10.1016/j.ijmedinf.2021.104440] [Medline: 33799055]
75. Klein JP, Hauer-von Mauschwitz A, Berger T, Fassbinder E, Mayer J, Borgwardt S, et al. Effectiveness and safety of the adjunctive use of an internet-based self-management intervention for borderline personality disorder in addition to care as usual: results from a randomised controlled trial. BMJ Open 2021 Sep 08;11(9):e047771 [FREE Full text] [doi: 10.1136/bmjopen-2020-047771] [Medline: 34497078]
76. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021 Aug 12;5(8):e20678 [FREE Full text] [doi: 10.2196/20678] [Medline: 34092548]
77. Lavelle J, Dunne N, Mulcahy HE, McHugh L. Chatbot-delivered cognitive defusion versus cognitive restructuring for negative self-referential thoughts: a pilot study. Psychol Rec 2021 Aug 24;72(2):247-261. [doi: 10.1007/s40732-021-00478-7]
78. Loveys K, Sagar M, Pickering I, Broadbent E. A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: randomized pilot trial. JMIR Ment Health 2021 Nov 08;8(11):e31586 [FREE Full text] [doi: 10.2196/31586] [Medline: 34596572]
79. Park S, Thieme A, Han J, Lee S, Rhee W, Suh B. “I wrote as if I were telling a story to someone I knew.”: designing chatbot interactions for expressive writing in mental health. In: Proceedings of the 2021 ACM Designing Interactive Systems Conference. 2021 Presented at: DIS '21; June 28-July 2, 2021; Virtual Event. [doi: 10.1145/3461778.3462143]
80. Romanovskyi O, Pidbutska N, Knysh A. Elomia chatbot: the effectiveness of artificial intelligence in the fight for mental health. In: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems. 2021 Presented at: COLINS-2021; April 22-23, 2021; Kharkiv, Ukraine URL: https://ceur-ws.org/Vol-2870/paper89.pdf
81. Troitskaya O, Batkhina A. Mobile application for couple relationships: results of a pilot effectiveness study. Fam Process 2022 Jun 14;61(2):625-642. [doi: 10.1111/famp.12733] [Medline: 34904235]

XSL•FO
RenderX

82. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh ML, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022 Mar;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
83. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022 Mar;27:100495 [FREE Full text] [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
84. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. IEEE Trans Hum Mach Syst 2022 Jun;52(3):343-353. [doi: 10.1109/thms.2021.3113643]
85. Rubin A, Livingston NA, Brady J, Hocking E, Bickmore T, Sawdy M, et al. Computerized relational agent to deliver alcohol brief intervention and referral to treatment in primary care: a randomized clinical trial. J Gen Intern Med 2022 Jan;37(1):70-77 [FREE Full text] [doi: 10.1007/s11606-021-06945-9] [Medline: 34145518]
86. Park S, Choi J, Lee S, Oh C, Kim C, La S, et al. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. J Med Internet Res 2019 Apr 16;21(4):e12231 [FREE Full text] [doi: 10.2196/12231] [Medline: 30990463]
87. Linardon J, Shatte A, Messer M, Firth J, Fuller-Tyszkiewicz M. E-mental health interventions for the treatment and prevention of eating disorders: an updated systematic review and meta-analysis. J Consult Clin Psychol 2020 Nov;88(11):994-1007. [doi: 10.1037/ccp0000575] [Medline: 32852971]
88. Linardon J, Fitzsimmons-Craft EE, Brennan L, Barillaro M, Wilfley DE. Dropout from interpersonal psychotherapy for mental health disorders: a systematic review and meta-analysis. Psychother Res 2019 Oct;29(7):870-881. [doi: 10.1080/10503307.2018.1497215] [Medline: 30005586]
89. Cooper AA, Conklin LR. Dropout from individual psychotherapy for major depression: a meta-analysis of randomized clinical trials. Clin Psychol Rev 2015 Aug;40:57-65. [doi: 10.1016/j.cpr.2015.05.001] [Medline: 26067572]
90. Gersh E, Hallford DJ, Rice SM, Kazantzis N, Gersh H, Gersh B, et al. Systematic review and meta-analysis of dropout rates in individual psychotherapy for generalized anxiety disorder. J Anxiety Disord 2017 Dec;52:25-33. [doi: 10.1016/j.janxdis.2017.10.001] [Medline: 29028610]
91. Bernstein EE, Weingarden H, Wolfe EC, Hall MD, Snorrason I, Wilhelm S. Human support in app-based cognitive behavioral therapies for emotional disorders: scoping review. J Med Internet Res 2022 Apr 08;24(4):e33307 [FREE Full text] [doi: 10.2196/33307] [Medline: 35394434]
92. Taylor H, Strauss C, Cavanagh K. Can a little bit of mindfulness do you good? A systematic review and meta-analyses of unguided mindfulness-based self-help interventions. Clin Psychol Rev 2021 Nov;89:102078. [doi: 10.1016/j.cpr.2021.102078] [Medline: 34537665]
93. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
94. Ter Stal S, Sloots J, Ramlal A, Op den Akker H, Lenferink A, Tabak M. An embodied conversational agent in an eHealth self-management intervention for chronic obstructive pulmonary disease and chronic heart failure: exploratory study in a real-life setting. JMIR Hum Factors 2021 Nov 04;8(4):e24110 [FREE Full text] [doi: 10.2196/24110] [Medline: 34734824]
95. Curtis RG, Bartel B, Ferguson T, Blake HT, Northcott C, Virgara R, et al. Improving user experience of virtual health assistants: scoping review. J Med Internet Res 2021 Dec 21;23(12):e31737 [FREE Full text] [doi: 10.2196/31737] [Medline: 34931997]
96. Boucher EM, Ward HE, Mounts AC, Parks AC. Engagement in digital mental health interventions: can monetary incentives help? Front Psychol 2021 Nov 18;12:746324 [FREE Full text] [doi: 10.3389/fpsyg.2021.746324] [Medline: 34867629]
97. Pham M, Do HM, Su Z, Bishop A, Sheng W. Negative emotion management using a smart shirt and a robot assistant. IEEE Robot Autom Lett 2021 Apr;6(2):4040-4047. [doi: 10.1109/lra.2021.3067867]
98. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psych 2023 Aug 25;10(3):119-135. [doi: 10.1007/s40501-023-00297-3]

Abbreviations

AI: artificial intelligence
CA: conversational agent
ECA: embodied conversational agent
ePST: electronic problem-solving treatment
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 25.04.23; peer-reviewed by X B, Y Xi, E Kim, C Muñoz; comments to author 24.07.23; revised version received 21.09.23; accepted 04.12.23; published 27.02.24.

Please cite as:
Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L
Attrition in Conversational Agent–Delivered Mental Health Interventions: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e48168
URL: https://www.jmir.org/2024/1/e48168
doi:10.2196/48168
PMID:38412023

©Ahmad Ishqi Jabir, Xiaowen Lin, Laura Martinengo, Gemma Sharp, Yin-Leng Theng, Lorainne Tudor Car. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effect of Digital Mental Health Literacy Interventions on Mental Health: Systematic Review and Meta-Analysis

GeckHong Yeo[1], PhD; Stephanie M Reich[2], PhD; Nicole A Liaw[3], BA; Elizabeth Yee Min Chia[3], MSc

[1]N.1 Institute for Health, National University of Singapore, Singapore, Singapore
[2]School of Education, University of California, Irvine, Irvine, CA, United States
[3]SHINE Children and Youth Services, Singapore, Singapore, Singapore

Corresponding Author:
Stephanie M Reich, PhD
School of Education
University of California, Irvine
3454 Education
Irvine, CA, 92697
United States
Phone: 1 (949) 824 5970
Email: smreich@uci.edu

Abstract

Background: Accelerated by technological advancements and the recent global pandemic, there is burgeoning interest in digital mental health literacy (DMHL) interventions that can positively affect mental health. However, existing work remains inconclusive regarding the effectiveness of DMHL interventions.

Objective: This systematic review and meta-analysis investigated the components and modes of DMHL interventions, their moderating factors, and their long-term impacts on mental health literacy and mental health.

Methods: We used a random-effects model to conduct meta-analyses and meta-regressions on moderating effects of DMHL interventions on mental health.

Results: Using 144 interventions with 206 effect sizes, we found a moderate effect of DMHL interventions in enhancing distal mental health outcomes (standardized mean difference=0.42, 95% CI −0.10 to 0.73; $P<.001$) and a large effect in increasing proximal mental health literacy outcomes (standardized mean difference=0.65, 95% CI 0.59-0.74; $P<.001$). Uptake of DMHL interventions was comparable with that of control conditions, and uptake of DMHL interventions did not moderate the effects on both proximal mental health literacy outcomes and distal mental health outcomes. DMHL interventions were as effective as face-to-face interventions and did not differ by platform type or dosage. *DMHL plus* interventions (DMHL psychoeducation coupled with other active treatment) produced large effects in bolstering mental health, were more effective than *DMHL only* interventions (self-help DMHL psychoeducation), and were comparable with non-DMHL interventions (treatment as usual). DMHL interventions demonstrated positive effects on mental health that were sustained over follow-up assessments and were most effective in enhancing the mental health of emerging and older adults.

Conclusions: For theory building, our review and meta-analysis found that DMHL interventions are as effective as face-to-face interventions. DMHL interventions confer optimal effects on mental health when DMHL psychoeducation is combined with informal, nonprofessional active treatment components such as skills training and peer support, which demonstrate comparable effectiveness with that of treatment as usual (client-professional interactions and therapies). These effects, which did not differ by platform type or dosage, were sustained over time. Additionally, most DMHL interventions are found in Western cultural contexts, especially in high-income countries (Global North) such as Australia, the United States, and the United Kingdom, and limited research is conducted in low-income countries in Asia and in South American and African countries. Most of the DMHL studies did not report information on the racial or ethnic makeup of the samples. Future work on DMHL interventions that target racial or ethnic minority groups, particularly the design, adoption, and evaluation of the effects of culturally adaptive DMHL interventions on uptake and mental health functioning, is needed. Such evidence can drive the adoption and implementation of DMHL interventions at scale, which represents a key foundation for practice-changing impact in the provision of mental health resources for individuals and the community.

XSL•FO
RenderX

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42023363995; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023363995

(J Med Internet Res 2024;26:e51268) doi:10.2196/51268

KEYWORDS

review and meta-analysis; digital mental health literacy; digital mental health interventions; mental health functioning

Introduction

Background

Worldwide, mental illness is projected to have an economic cost of approximately US $6 trillion owing to poor productivity and negative health functioning [1]. The World Health Organization and World Federation for Mental Health have advocated for increasing global awareness of mental health [1]. To increase global understanding of and address mental health problems, researchers, policy makers, and mental health practitioners have long recognized the significance of individual- and society-level mental health literacy (MHL) [2]. MHL refers to "knowledge and beliefs about mental disorders which aid their recognition, management, or prevention" [3]. Low MHL in the general public is a key impediment to seeking mental health treatment [2]. A multifaceted construct, MHL comprises (1) understanding how to prevent mental illness, (2) understanding when a disorder is developing, (3) awareness of support and treatments for mental illness, (4) ability to effectively address mild mental health problems, and (5) mental health first aid skills to support others [4]. In more recent work, MHL has been expanded to include knowledge about (1) obtaining and maintaining good mental health, (2) understanding mental illnesses and treatments, (3) reducing mental illness–related stigma, (4) enhancing help-seeking efficacy or behaviors, and (5) enhancing help-seeking attitudes or intentions [5,6].

Unlike the extensive body of work on health literacy and its relationship to positive health outcomes [2,5,7], research on MHL is more recent and less established in how it relates to or affects mental health functioning [2]. Accelerated by technological advancements and the recent global pandemic, there is burgeoning interest in digital mental health interventions that can be delivered at scale and translate to real-world benefits [8,9]. In particular, recent work has highlighted the use of digital platforms (eg, web-based platforms, apps, and social media) as key facilitators for building digital MHL (DMHL), especially among young people [10,11]. Web-based pedagogies; diverse educational content; and timely interactions with peer supporters, trainers, mental health professionals, and people with common lived experiences of mental illness afforded by digital platforms can effectively promote DMHL [10,11]. As a nascent field, research on DMHL has primarily focused on and been incorporated into mental health interventions, with few observational or field studies documenting DMHL in naturalistic settings in relation to mental health functioning [12,13].

Existing work on DMHL interventions is scattered, and findings about the effects of DMHL interventions on mental health are mixed [14]. These inconsistent findings on DMHL interventions stem from how components of DMHL can vary across studies [15,16] and whether interventions assess the impact on proximal literacy outcomes only (facets of DMHL) [17,18], distal mental health outcomes only (mental health symptoms and conditions) [19], or both [15,16]. In particular, some interventions focus on DMHL as the primary intervening component, in which self-help psychoeducation is used to access information and learn about mental health [16,20]. Other interventions combine DMHL psychoeducation with types of treatment such as skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; and informal, nonprofessional counselor interactions [15,21]. Digital mental health interventions that use DMHL as a primary or secondary component often include and are compared with other non-DMHL interventions. These non-DMHL interventions include treatment as usual with professional therapies (eg, cognitive behavioral therapy [CBT] and dialectical therapy) and skills training (eg, mindfulness [22,23]). Although receiving professional mental health treatment likely improves knowledge about mental health symptoms and management (eg, MHL), such interventions are not specifically focused on literacy. As such, they are categorized as non-DMHL interventions. Thus, it is unclear whether DMHL psychoeducation is sufficient or whether interventions need to incorporate DMHL with other active treatment components to improve individuals' mental health.

Importantly, interventions with different DMHL components can have differential impacts on proximal and distal mental health outcomes [24,25]. DMHL interventions with DMHL psychoeducation can increase knowledge and beliefs about recognition, management, or prevention of mental disorders (the 5 facets of DMHL described previously) and likely have a greater impact on proximal (eg, knowledge, attitudes, and beliefs) than on distal (eg, improved mental health) [16] mental health outcomes. DMHL interventions coupled with active treatment components, on the other hand, may enhance both proximal and distal mental health outcomes [15]. Meta-analyses have focused on digital mental health treatment interventions such as professional therapies and skill-based training (eg, CBT and mindfulness meditation, respectively) [26]. There are no studies that evaluate the impact of DMHL interventions on both proximal and distal mental health outcomes or that consider psychoeducation alone or in conjunction with treatment. Research effort is required to synthesize findings across studies to draw inferences about the impact of DMHL interventions on mental health—specifically, whether DMHL interventions are fundamentally effective (ie, pretest-posttest DMHL intervention comparisons) and more effective than well-controlled conditions (ie, waitlist control and non-DMHL interventions).

XSL•FO
RenderX

DMHL Versus Traditional Face-to-Face MHL Interventions

Scholars have argued that DMHL interventions combine ease of access and cost-efficiency with efficacy and are more effective on mental health [27,28]. DMHL interventions can overcome the shortcomings of traditional face-to-face MHL interventions, including low availability, a high threshold for participation, and substantial delivery costs [27,28]. DMHL interventions have several advantages [27,28]: (1) easy accessibility at any time and place; (2) assurance of anonymity to avoid stigmatization; (3) self-guidance for participants to work at their own pace and review materials as often as they want; (4) ability to reach individuals faster than traditional mental health services and prevent the onset of more severe mental health problems; and finally, (5) easy scalability, requiring only a small increase in resources to reach a greater proportion of the eligible population. Previous work has found that DMHL interventions combine ease of access and cost-efficiency with efficacy [27,28], whereas traditional face-to-face mental health treatment interventions that include DMHL components require close to 8 times more therapist time than digital ones [29].

DMHL interventions may also help populations that are not reached by existing traditional MHL approaches, particularly young people, whose lives are closely intertwined with digital media and are the primary users of web-based mental health resources [30]. DMHL interventions address 2 critical issues that traditional face-to-face MHL interventions encounter in targeting mental health, especially for youth mental health—concern for anonymity and limited reach [31]. For young people, the increased prevalence of mental health conditions, especially depression and anxiety [32,33], underscores the need to address major barriers to help-seeking behaviors, for instance, access, reach, and stigma [34]. Thus, DMHL interventions may play a particularly important role during adolescence and emerging adulthood as a cost-effective channel to intervene early in preventing mental illness and supporting mental health functioning [35].

An accumulating body of work indicates that DMHL interventions can reduce the public health burden by decreasing burnout among working professionals and distress among students [36,37]. Poor MHL is a primary factor that hinders the uptake of, engagement in, and adherence to mental health treatment prevention and intervention [2]. DMHL interventions have the potential to not only provide greater access and reach and reduce stigma but also improve MHL among the public [34]. Indeed, research has shown that improved MHL can address self-identifying mental health difficulties and enhance help-seeking intentions and behaviors, which are key to initial engagement in and subsequent adherence to treatment [2]. Thus, DMHL intervention is an upstream form of mental health prevention that can reduce the need for downstream intervention [1]. For instance, incorporating DMHL into mental health initiatives as a regular form of psychoeducation can target psychological readiness for enhancing mental health functioning [38]. DMHL may also function as a mental health resilience factor that protects individuals from and mediates the negative effects of adversity and risk factors for the development of psychopathology [39].

However, there is no conclusive evidence demonstrating the effectiveness of DMHL interventions compared with traditional face-to-face MHL interventions in bolstering mental health [40,41]. To establish evidence-based practices for DMHL interventions, an integrated and comprehensive investigation of DMHL features and affordances, as well as individual and contextual factors that amplify or attenuate the effectiveness of DMHL interventions on mental health, is needed [8,9]. The primary aim of our systematic review and meta-analysis was to evaluate the effectiveness of DMHL interventions on mental health by targeting several objectives. First, we addressed the lacunae in our current understanding of DMHL interventions by synthesizing findings across studies to ascertain how DMHL interventions compare with traditional face-to-face MHL interventions. Next, we assessed whether and how different DMHL components in DMHL interventions affect mental health outcomes. Third, we investigated the effectiveness of DMHL interventions on proximal MHL outcomes (eg, knowledge, attitudes, and beliefs) and distal mental health outcomes. Fourth, we established whether a strong inference can be drawn about the effectiveness of DMHL interventions—whether they are fundamentally effective (ie, pretest-posttest DMHL intervention comparisons) and more effective than well-controlled conditions. Finally, we amalgamated findings across studies to establish the long-term implications (carryover effects) of DMHL interventions for mental health functioning. A secondary aim was to identify moderating factors that amplify or attenuate the effectiveness of DMHL interventions, particularly DMHL features and affordances (new vs conventional platforms and dosages of intervention) and individual (sex, developmental differences, and severity of preexisting mental health conditions) and contextual (cultural contexts) factors.

Effectiveness of DMHL on Uptake and Mental Well-Being

With the burgeoning interest in digital mental health interventions, which combine scalability, translation to real-world benefits, and cost-efficiency with efficacy [8,9], substantive research has been devoted to understanding whether and how digital modes of delivering mental health services, including DMHL, are superior to or comparable with traditional face-to-face delivery [8,9]. Efforts to synthesize studies to compare the mental health impact of DMHL and traditional modes of MHL interventions are warranted to provide conclusive evidence on the effectiveness of DMHL interventions. In particular, research has focused on the implementation and clinical effectiveness of digital mental health interventions broadly (but not on DMHL specifically) by assessing uptake (ie, engagement and adherence) and mental health outcomes [26], respectively. Although there is evidence that digital mental health interventions increase uptake and engagement and that greater intervention uptake enhances mental health functioning [26], similar evidence specific to DMHL is lacking. DMHL studies are scattered in their efforts to understand the impact of DMHL interventions in terms of both uptake and intervening in mental health outcomes and whether intervention uptake affects mental health outcomes.

XSL•FO
RenderX

Drawing from reviews and meta-analyses on digital mental health interventions (in the absence of similar work on DMHL), findings indicate that most interventions are implemented and evaluated in high-income countries (Global North) involving White samples [42-44]. Given existing mental health disparities that are compounded with the mental health costs of marginalization, racial and ethnic minority groups face greater mental health risks, including limited access to and greater barriers to engaging in digital treatment interventions for mental health [42-44]. Such findings underscore the need to attend to the ethnic or racial composition of the samples to understand the overall effectiveness of digital mental health interventions, including DMHL, on uptake and mental health outcomes. In considering impact, scholars and practitioners have called for attention to the long-term mental health implications of DMHL—specifically, the carryover effects of DMHL interventions on mental health functioning [36,45]. For instance, some DMHL interventions demonstrate no carryover effects [46,47], whereas others yield 3, 6, or 12 weeks of carryover effects [48,49]. Such evidence can drive the adoption and implementation of evidence-based DMHL interventions at scale, which represents a key foundation for practice-changing impacts in the provision of mental health resources for individuals and the community.

Our review of the DMHL intervention literature found that DMHL is typically the primary component—specifically, self-help DMHL psychoeducation that involves acquiring knowledge and information on mental health (hereafter referred to as *DMHL only* [16,20])—or a secondary component, incorporating DMHL psychoeducation with other active treatment components such as skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; and informal, nonprofessional counselor interactions [15,21]. We refer to those interventions with DMHL as a secondary component as *DMHL plus*. To establish the effectiveness of DMHL interventions, it is necessary to consider the DMHL component of each intervention. DMHL interventions have predominantly focused on comparing mental health outcomes through the following DMHL components: (1) *DMHL only* (vs waitlist control [20]), (2) *DMHL plus* (vs waitlist control [15,16,50]), (3) *DMHL only* versus *DMHL plus* [21,51], and (4) *DMHL only* or *plus* versus non-DMHL (treatment as usual with professional therapies such as CBT and dialectical therapy [28,52]). Although receiving professional mental health treatment likely improves knowledge about mental health symptoms and management (eg, MHL), such interventions are not specifically focused on literacy. As such, they are categorized as non-DMHL interventions. Studies often compare the effectiveness of different intervention components on mental health outcomes with inconsistent findings. To ascertain whether DMHL interventions are fundamentally effective (pretest-posttest comparison) and whether stronger inferences about the effect of DMHL interventions can be drawn—that is, how they compare with control and other intervention conditions—we synthesized findings across three study designs: (1) pre- and postintervention comparison [53,54], (2) intervention group versus (waitlist) control group [49,54], and (3) DMHL intervention versus non-DMHL intervention [28,49].

Findings on the effectiveness of DMHL interventions on mental health outcomes are mixed, with some studies demonstrating positive effects [39,49] and others finding none [28,53]. A possible reason for these inconsistent findings is the contradictory ways of conceptualizing and operationalizing DMHL and mental health outcomes in DMHL interventions. Although most interventions include a DMHL component, few draw on the long-standing MHL paradigm [16,55]. Hence, for most interventions, it is often ambiguous which of the five facets of MHL were examined, including one or some combination of them [15,56]: (1) knowledge about obtaining and maintaining good mental health, (2) understanding mental illnesses and treatments, (3) reducing mental illness–related stigma, (4) enhancing help-seeking efficacy or behaviors, and (5) enhancing help-seeking attitudes or intentions [4].

Traditional face-to-face MHL interventions have found improvements in proximal mental health outcomes, particularly 1 or a combination of the 5 facets of MHL as well as distal mental health outcomes involving mental health conditions and functioning [56,57]. Extrapolating these findings to DMHL interventions, interventions targeting or incorporating different facets of MHL may have differential impacts on proximal and distal mental health outcomes [15,20]. However, there are insufficient DMHL studies that have investigated the 5 facets separately to distinguish their effects on mental health. Most studies have created composite variables of DMHL that comprise different combinations of the 5 facets [15,20]. Thus, with the available research on DMHL interventions, the synthesis of findings across studies is possible as a composite DMHL construct intervening in mental health to provide a conclusive understanding of the mental health impact of DMHL interventions.

Furthermore, DMHL interventions are inconsistent in whether they assess proximal outcomes only [17,18], distal outcomes only [19,54], or a combination of both increased literacy outcomes and better mental health conditions [21,28]. DMHL research often fails to acknowledge how changes in proximal DMHL outcomes factor in the effects of DMHL interventions on distal outcomes—mental health functioning [49,54]. Thus, research that distinguishes the effectiveness of DMHL interventions on both proximal DMHL outcomes and distal mental health outcomes is warranted. In addition, studies have obtained inconsistent findings on the effectiveness of DMHL interventions on the same mental health outcomes, such as depression [49,54], as well as across different mental health outcomes [58,59]. On the basis of the body of work on traditional face-to-face MHL interventions, researchers have argued for the importance of examining MHL with specific mental health outcomes (eg, eating disorders and depression) [2]. However, research on DMHL is in its infancy, and the existing literature contains insufficient estimates to test the impact of DMHL interventions on the range of mental health outcomes.

Moderating Role of DMHL Features and Individual and Contextual Factors

To build a strong economic case for investing in digital mental health interventions that have clinical effectiveness (ie, reducing

XSL•FO
RenderX

mental ill health and symptoms) broadly and on DMHL specifically, a key consideration is the dosage of the intervention [60]. Intervention durations that are too long or too short can affect its effectiveness [60], and typical digital mental health interventions that include DMHL components have a median dosage of 10 weeks [61,62]. However, it remains ambiguous whether this median of 10 weeks of intervention has a positive impact on mental health [60] and how it applies to the effectiveness of DMHL interventions. Hence, conclusive findings comparing different treatment dosages of DMHL interventions that demonstrate improvements or changes in mental health are necessary. Scholars have also argued that the features and affordances of digital platforms can factor into the effectiveness of digital mental health interventions by influencing users' initial uptake and sustained engagement [8,9]. Accumulating research indicates that technological affordances can vary across platforms, from linear, static websites to more interactive social media platforms and mobile apps [8,9]. Specifically, DMHL interventions are commonly delivered through new and more conventional platforms. New platforms involving mobile apps, web-based or internet applications, and social media afford greater interactivity [20,54], whereas more conventional platforms, including films, videos, multimedia, and emails, afford lower or limited interactivity [17,18]. Thus, this review considered how the duration of the intervention and platform of delivery affect outcomes.

Drawing from meta-analyses and systematic reviews of digital mental health interventions, individual factors related to preexisting physical and mental health conditions can modulate the effectiveness of interventions [8]. Physical illnesses such as rheumatoid arthritis can interact and generate or exacerbate mental health problems and reduce the effectiveness of interventions [63]. The severity of one's preexisting mental illness can limit intra- and interpersonal resources (ie, motivation and supportive relationships, respectively) for uptake of, engagement in, and adherence to digital mental health interventions, thus reducing their effectiveness [8]. Extending these arguments to DMHL interventions, some research has shown that individual factors, particularly being female and having a mental health problem, are associated with relatively high levels of MHL and reduced effects of DMHL interventions on mental health outcomes (a possible ceiling effect [64]). However, these moderating effects of sex and preexisting mental health conditions on the intervening effectiveness of DMHL interventions are tentative and not consistently documented across studies [64,65]. Thus, the characteristics of individuals need to be considered as important moderators.

Culture is a moderating contextual factor in the effectiveness of digital mental health interventions [66], but there is a dearth of research that investigates different cultures. Existing literature only allows for a comparison of the effects of DMHL interventions across Western and Eastern cultural contexts [67,68]. In more independence-oriented Western cultural contexts, values and norms are focused on developing a healthy sense of self, including positive mental health [67,68]. There is greater public awareness of mental health; less social stigma associated with help-seeking behaviors; and more concerted efforts to build mental health resiliency through the adoption and implementation of digital mental health interventions, including DMHL interventions [67,68]. Thus, in Western cultural contexts, greater MHL and a focus on mental health may amplify the effects of DMHL interventions. In contrast, most collectivistic, interdependence-oriented Asian societies are conservative in addressing mental health [67], and they promote MHL and positive mental health functioning in ways that differ from those of individualistic Western cultural contexts, for instance, cultivating interdependence relationships, relational harmony, and dialectical beliefs and emotions—a balanced state of opposites, including experiencing both positive and negative emotions, which are fused with and change into each other [68]. The increase in awareness of mental health issues is relatively recent, especially in contemporary Asian societies [69,70], which may be associated with weaker impacts of DMHL interventions on mental health [71]. The paucity of reviews and meta-analytic efforts that synthesize findings across international studies on digital mental health interventions makes findings about cultural context tentative and underscores the importance of this consideration in this study.

Aims of This Review and Meta-Analysis

To our knowledge, no systematic review or meta-analysis has evaluated the effectiveness of DMHL interventions on mental well-being or the moderating factors between these 2. To this end, our efforts to synthesize findings across studies focused on the main mental health impacts of DMHL interventions. First, we ascertained how DMHL interventions compare with traditional face-to-face MHL interventions in affecting mental health outcomes. Second, we examined the effects of DMHL interventions on the implementation outcome of uptake and the proximal DMHL outcomes and distal mental health outcomes. In addition, we examined whether uptake of DMHL interventions affects the impact on the proximal DMHL outcomes and distal mental health outcomes. Third, we investigated the impact of DMHL interventions on mental health through the following DMHL components: (1) *DMHL only* (self-help DMHL psychoeducation that entails acquiring knowledge and information), (2) *DMHL plus* (DMHL psychoeducation combined with skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; and informal, nonprofessional counselor interactions), (3) *DMHL only* versus non-DMHL (treatment as usual through professional therapies such as CBT and dialectical therapy), and (4) *DMHL plus* versus non-DMHL. Fourth, our review and meta-analytic efforts provide evidence for a stronger inference on the effectiveness of DMHL interventions on mental health functioning by combining and comparing results across three study designs: (1) pre- and postintervention comparison, (2) intervention group versus (waitlist) control group, and (3) DMHL intervention versus non-DMHL intervention. Fifth, we addressed the gap in the current literature regarding the long-lasting effectiveness of DMHL interventions on mental health by assessing carryover effects. Finally, we considered potential moderators of treatment effects, including dosage, platform, individual characteristics, and culture.

In addressing these main effects, our meta-analytic efforts conceptualized and operationalized DMHL as a composite construct to understand its mental health impacts as there are

XSL•FO
RenderX

limited studies that have examined all 5 facets of DMHL and mental well-being conjointly and separately. Thus, there are insufficient studies (and number of effect sizes) to determine how the 5 facets of DMHL compare in their impacts on mental health. Although scholars have argued for the study of digital mental health interventions for specific mental health outcomes (eg, depression vs anxiety vs eating disorders [8,9]), the limited number of studies in the literature renders it impossible to study such specific effects, particularly for DMHL interventions. On the basis of the literature reviewed previously, we examined the following research questions (RQs) and hypotheses:

1. What is the effect of DMHL interventions on uptake as an implementation outcome? (RQ 1a)
2. What are the effects of DMHL interventions on proximal literacy outcomes (ie, DMHL outcomes) and distal mental health outcomes? (RQ 1b)
3. Does the uptake of DMHL interventions moderate the effects on proximal literacy outcomes and distal mental health outcomes? (RQ 1c)
4. How do DMHL interventions with different DMHL components or conditions compare in their impact on mental health (*DMHL only*, *DMHL plus*, *DMHL only* vs non-DMHL, and *DMHL plus* vs non-DMHL)? (RQ 2a)
5. Do DMHL interventions demonstrate a stronger inference on intervening in mental health functioning? Specifically, what are the effects of DMHL interventions in pretest-posttest comparisons, (waitlist) control groups versus intervention groups, and DMHL interventions versus non-DMHL interventions? (RQ 2b)
6. Are there sustained or carryover effects of DMHL interventions on mental health? (RQ 3)
7. DMHL interventions are as effective as traditional face-to-face MHL interventions in bolstering mental health (hypothesis 1).
8. DMHL interventions administered through new platforms that afford greater interactivity (ie, mobile apps, web-based or internet platforms, and social media) than more conventional platforms (ie, films, videos, multimedia, and emails) have greater positive impacts on mental health (hypothesis 2).
9. DMHL interventions with a dosage of 10 weeks are most effective in increasing mental health functioning (hypothesis 3).
10. DMHL interventions demonstrate reduced effects on mental health in female participants compared with male participants (hypothesis 4a).
11. DMHL interventions have a lower impact on mental health in individuals with greater severity of preexisting mental health conditions (hypothesis 4b).
12. DMHL interventions have greater effects in adolescents than in participants at other developmental stages (ie, emerging adulthood and older adulthood; hypothesis 4c).
13. The intervening effectiveness of DMHL interventions on mental health is greater in Western than in Eastern cultural contexts (hypothesis 5).

Methods

This meta-analysis was conducted according to the guidelines from Quintana [72] and reported based on the latest version of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [73]. A protocol was registered a priori following the PRISMA guidelines (PROSPERO registration CRD42023363995).

Literature Search

With the assistance of a staff librarian at the first author's affiliated institution, 2 research assistants independently used 3 search strategies to systematically identify studies on DMHL and mental health. The 3 search strategies; search terms that were developed using the Population, Intervention, Comparison, and Outcome search strategy; and full search strings are provided in Tables S1 to S3 in Multimedia Appendix 1.

Study Selection

Study screening was conducted using the Covidence software (Veritas Health Innovation) [74]. Figure 1 presents a flowchart of the study selection process. Of the 42,014 records identified in the initial searches, 20,137 (47.93%) were duplicates. First-stage screening of the 21,879 (52.08%) remaining records entailed checking of titles and abstracts by 2 reviewers, which led to the elimination of 21,459 (98.08%) of the 21,879 records. From the 420 (1.92%) of the 2,1879 records sought for retrieval, 55 (13.1%) of the 420 records were not retrieved due to the lack of an English full-text pdf article. The remaining 367 (87.4%) records selected for full text review were independently screened by another 2 reviewers with any discrepancies resolved through consensus. A third reviewer was contacted if consensus could not be reached. Inclusion and exclusion criteria were established a priori and are included in Tables S1 to S3 Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

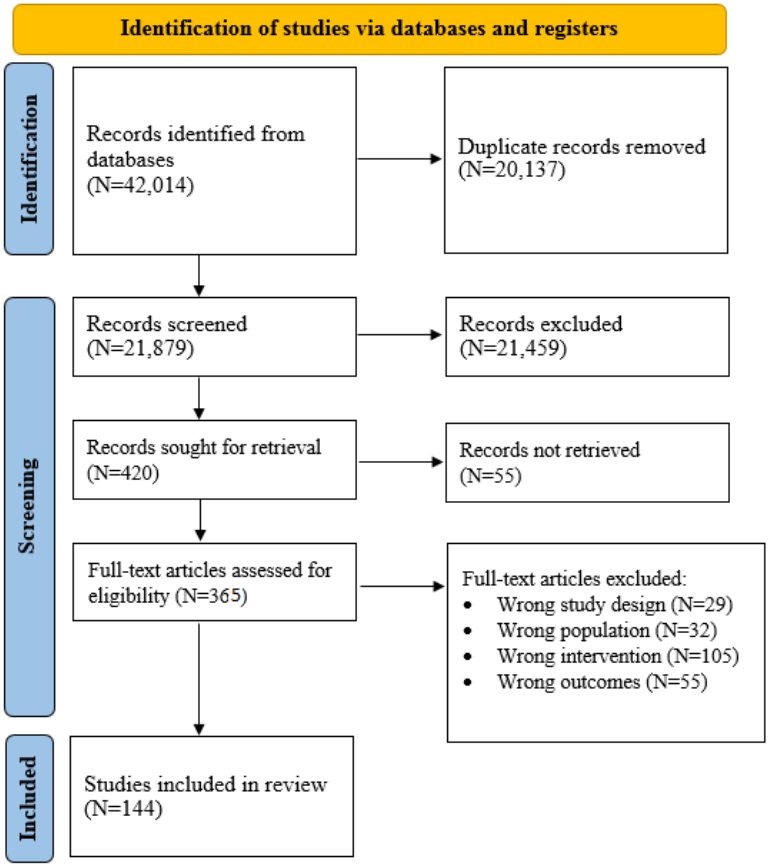

A total of 146 studies with 208 effect sizes were included in our meta-analyses, with 2 (1.4%) field studies reporting relevant correlation coefficients and 144 (98.6%) intervention studies providing sufficient statistics to compute the standardized mean difference (SMD; Cohen *d*), which contributed 2 and 206 effect sizes, respectively. Although DMHL interventions are examined in field studies and as interventions in experimental studies, it is methodologically feasible to examine both field and experimental studies within the same meta-analysis [75,76]. However, we excluded the field studies (2/146, 1.4%) as they contributed only 2 effect size estimates, which are insufficient for a comparison of the relationship between DMHL and mental health across field and intervention studies.

Data Extraction and Coding

In total, 2 independent reviewers (2 research assistants) extracted and coded data on multiple aspects (see Multimedia Appendix 2 [77-81] and Multimedia Appendix 3 [82,83] on study characteristics and summary and sample characteristics, respectively). The coders achieved 83% agreement on their codes. Any discrepancies in coding were discussed and resolved. Many of the interventions included in our meta-analysis (54/144, 37.5%) did not provide a measure of MHL. The 34% (49/144) of interventions that assessed DMHL outcomes primarily used the Mental Health Literacy Scale [84].

Risk-of-Bias (Quality) Assessment

Studies included in the meta-analysis were independently evaluated for quality by 2 reviewers, with differences discussed and resolved. DMHL interventions were assessed using the instrument by Downs and Black [85] and the Cochrane Collaboration risk-of-bias tool [86]. We report these results in Multimedia Appendix 2.

Multiple Dependent Effect Sizes and Computation of Effect Sizes

There were several instances in which the studies contributed multiple dependent effect sizes in our meta-analyses. Following the guidelines provided by Quintana [72], we dealt with this issue in 3 ways that we outlined in Figure S1 in Multimedia Appendix 4 [82,83].

Publication Bias Analysis

A total of 3 analyses were used to ascertain publication bias, and the results are reported in Figure S1 in Multimedia Appendix 5.

Data Analysis

Data were analyzed using RStudio (version 4.0.0; Post, PBC) with the *metafor* and *robumta* packages (version 3.02) [87-89].

Results

Overview

Summary and sample characteristics of the included DMHL intervention studies, as well as of traditional face-to-face MHL intervention studies, are presented in Tables S1 and S2 in Multimedia Appendix 2. We included a forest plot to visualize

XSL•FO
RenderX

the effect sizes and CIs from the included studies with a computed summary effect size (Multimedia Appendix 5). In the following sections, we present the results that address each hypothesis and RQ.

Effectiveness of DMHL Interventions in Enhancing Uptake and Mental Health

As hypothesized, DMHL interventions had similar effectiveness to that of traditional face-to-face MHL interventions in bolstering mental health ($Q_{between}$=4.12; P=.18; hypothesis 1; Table S3 in Multimedia Appendix 2). Addressing RQ 1a, DMHL interventions versus control conditions had comparable effects on uptake (odds ratio 0.998, 95% CI 0.91-1.03; P<.001; Table S4 in Multimedia Appendix 2). For RQ 1b, we found a high effect of DMHL interventions in increasing proximal literacy outcomes (ie, 5 facets of DMHL), with a pooled effect size of SMD=0.65 (95% CI 0.59-0.74; P<.001), and a moderate effect in enhancing distal mental health outcomes, with a pooled effect size of SMD=0.42 (95% CI −0.10 to 0.73; P<.001). Unfortunately, few studies (2/144, 1.4%) differentiated the 5 facets of DMHL, and thus, we were unable to assess the effects of each DMHL facet on the outcomes. For RQ 1c, we found that uptake of DMHL interventions did not moderate the effect of the interventions on proximal literacy outcomes ($Q_{between}$=1.07; P=.30). Subgroup comparisons of (1) DMHL interventions with baseline and completer samples that were similar in baseline and demographic characteristics, (2) DMHL interventions that did not provide information on baseline and completer samples' baseline and demographic characteristics, and (3) DMHL interventions with baseline and completer samples that were different in baseline and demographic characteristics indicated no significant difference in impacts on MHL outcomes ($Q_{between}$=1.19; P=.55). Similarly, our results revealed that the uptake of DMHL interventions did not moderate the effect of the interventions on distal mental health outcomes ($Q_{between}$=0.17; P=.68). Subgroup comparisons of (1) DMHL interventions with baseline and completer samples that were similar in baseline and demographic characteristics, (2) DMHL interventions that did not provide information on baseline and completer samples' baseline and demographic characteristics, and (3) DMHL interventions with baseline and completer samples that were different in baseline and demographic characteristics indicated no significant difference in impacts on mental health outcomes ($Q_{between}$=5.21; P=.08).

For RQ 2a, we compared DMHL interventions with DMHL alone or with additional aspects (ie, *DMHL only* vs waitlist control conditions, *DMHL plus* vs waitlist control conditions, *DMHL only* vs *DMHL plus*, and *DMHL plus* vs non-DMHL). Q statistics analyses comparing these 4 conditions indicated that they differed significantly ($Q_{between}$=9.45; P=.01). The effect size of *DMHL only* versus waitlist control was comparable with that of *DMHL plus* versus waitlist control but greater than that of *DMHL only* versus *DMHL plus* and *DMHL plus* versus non-DMHL (Table S1 in Multimedia Appendix 3). To address RQ 2b, we synthesized the effect sizes from 44.4% (64/144) of the DMHL interventions to elucidate whether we could draw stronger inferences about the effectiveness of DMHL interventions in increasing mental health. Q statistics analyses comparing the effect sizes of pretest-posttest DMHL intervention comparisons, waitlist control conditions versus DMHL interventions, and DMHL versus non-DMHL interventions revealed that they differed significantly ($Q_{between}$=12.09; P<.001), with higher effect sizes for pretest-posttest DMHL intervention comparisons and waitlist control conditions versus DMHL interventions than for DMHL versus non-DMHL interventions. For RQ 3, subgroup comparison of DMHL interventions that assessed mental health outcomes at the postintervention time point and those with follow-up assessments indicated that they did not differ significantly ($Q_{between}$=3.81; P=.06), with the effect sizes for interventions that assessed postintervention effects on mental health and those that assessed follow-up effects being comparable. Consistently, the carryover effects of DMHL interventions on mental health did not attenuate over time ($Q_{between}$=1.65; P=.20; Table S1 in Multimedia Appendix 3); the positive effects of DMHL interventions on mental health remained notwithstanding longer follow-up assessments.

Moderating Factors: DMHL Features, Individual Characteristics, and Cultural Contexts

We investigated potential moderators of DMHL interventions on mental health functioning (Table S1 in Multimedia Appendix 6 [90]) by performing subgroup analyses using meta-regressions to examine DMHL platform affordances (new interactive platforms, including mobile apps, web-based or internet platforms, and social media, vs conventional platforms, including films, videos, multimedia, and emails) and dosage of DMHL interventions as moderators. For dosage, we compared DMHL interventions of <10 weeks, 10 weeks, and >10 weeks. Contrary to hypothesis 2, DMHL interventions administered through new platforms that afford greater interactivity than more conventional platforms did not differ significantly in their impacts on mental health ($Q_{between}$=2.51; P=.31). In contrast to hypothesis 3, we found that the dosage of the DMHL interventions did not moderate their effectiveness on mental health outcomes ($Q_{between}$=2.13; P=.32), and similar positive intervening effects were found for all 3 dosage categories (Table S1 in Multimedia Appendix 3).

Developmental stages, including adolescence, emerging adulthood, and adulthood; sex, which was measured as the average proportion of participants in the sample who were female; and severity of mental health conditions at baseline were used as continuous predictors in testing their moderating effects. Contrary to hypothesis 4a, sex ($Q_{between}$=2.01; P=.34) did not moderate the impact of DMHL interventions on mental health. Contrary to our hypothesis, DMHL interventions were more effective in enhancing the mental health of emerging and older adults than that of adolescents (hypothesis 4b; $Q_{between}$=12.19; P=.001), and the severity of mental health conditions did not attenuate the effect of DMHL interventions on mental health (hypothesis 4c; $Q_{between}$=0.29; P=.45; Table S1 in Multimedia Appendix 6).

The studies included in the meta-analyses involved 23 countries, but there were not enough studies from each country to allow for a comparison of effect sizes among individual countries.

XSL•FO
RenderX

Instead, we followed the common approach of comparing Western and Eastern cultures [68,91]. Representatives of Western culture included Australia, Canada, Germany, the Netherlands, the United Kingdom, and the United States, whereas representatives of Eastern culture included China, Hong Kong, Pakistan, Singapore, and Taiwan. In contrast to hypothesis 5, we found that DMHL interventions conducted in Western and Eastern cultural contexts did not differ significantly in their effectiveness on mental health ($Q_{between}$=1.13; P=.64; Table S1 in Multimedia Appendix 6).

Discussion

Principal Findings

The rise in mental health issues worldwide accelerated during the pandemic, highlighting the need for upstream mental health prevention [1]. A key approach to mental health prevention is to build individuals' resiliency, which can protect them from and mediate the negative impacts of adversity and risk factors for the development of psychopathology [54]. Recent work on digital mental health interventions, particularly with DMHL, provides evidence for DMHL as a resilience factor that might mitigate stress and mental health conditions among working adults and students [36,37]. DMHL interventions combine ease of access with low cost, which has the potential to reduce public burden on a global scale by intervening in the mental health of individuals [27,28]. Overcoming the shortcomings of traditional face-to-face MHL interventions—specifically, low availability, high threshold for participation, and substantial delivery costs [27,28]—DMHL interventions may be more effective in addressing low levels of MHL in the public [2], which is an impediment to the uptake of, engagement in, and adherence to mental health prevention and intervention [2]. However, research on the effectiveness of DMHL interventions is scattered, and findings on their impact on mental health outcomes are inconclusive. This systematic review and meta-analysis is the first to provide empirical findings on the mental health implications of DMHL interventions.

First, we found that DMHL interventions are as effective as face-to-face interventions in improving MHL (SMD=0.64) and enhancing mental health functioning (SMD=0.42). DMHL interventions greatly improved pretest-posttest proximal DMHL outcomes, which involved various combinations of the 5 DMHL facets—knowledge about obtaining and maintaining good mental health; understanding mental illnesses and treatments; reducing stigma and enhancing help-seeking efficacy and attitudes; and more distal mental health conditions, such as anxiety, depression, loneliness, and distress; and bolstering well-being. Interestingly, we did not find greater uptake of and engagement with DMHL interventions (compared with control conditions), and uptake of DMHL interventions did not moderate the effects on both proximal MHL outcomes and distal mental health outcomes.

Second, our comparison of different DMHL intervention components and conditions, which included *DMHL only* (self-help DMHL psychoeducation) versus waitlist control conditions, *DMHL plus* (DMHL psychoeducation incorporated as a secondary component in other active treatments) versus waitlist control conditions, *DMHL only* versus *DMHL plus*, and *DMHL plus* versus non-DMHL (treatment as usual involving client-professional interactions and therapies such as CBT and dialectical therapy), found differential impacts on mental health outcomes. Compared with waitlist control conditions, *DMHL only* (SMD=0.59) and *DMHL plus* (SMD=0.45; P=.02) had similar positive mental health impacts. The effect size of *DMHL only* versus waitlist control was significantly greater than that of *DMHL only* versus *DMHL plus* (SMD=−0.35; P=.02), which indicates that the effectiveness of *DMHL only* interventions on mental health was significantly lower than that of *DMHL plus* interventions. On the other hand, the effect size of *DMHL plus* versus waitlist control was similar to that of *DMHL only* versus *DMHL plus* and *DMHL plus* versus non-DMHL (SMD=−0.33; P=.02). In other words, compared with waitlist control conditions, *DMHL only*, and non-DMHL, *DMHL plus* had similar positive effects on mental health.

Third, we elucidated whether stronger inferences can be drawn about the intervening effectiveness of DMHL interventions on mental health—whether they are fundamentally effective (pretest-posttest comparison) and more effective than (waitlist) control conditions or other (ie, non-DMHL) interventions. We found larger effect sizes for pretest-posttest DMHL intervention comparisons and waitlist control conditions versus DMHL interventions than for DMHL versus non-DMHL interventions. By comparing immediate and long-term effects, we found the benefits of DMHL interventions to be comparable at postintervention and follow-up assessments, with sustained effects on mental health regardless of longer follow-up assessments. This lack of fade-out of DMHL impacts is important and suggests that being more literate about mental health (recognition, prevention, and management) has long-term benefits for mental well-being. Finally, when considering how individual and contextual characteristics and dissemination methods might moderate the efficacy of DMHL interventions, we found no differences by sex, severity of preexisting mental health conditions, and cultural contexts and only slightly more efficacy in emerging and older adults than in adolescents. Our review and meta-analytic results indicate that digital platform interactivity and dosage of DMHL interventions do not enhance their efficacy. In particular, both new (involving mobile apps, web-based or internet platforms, and social media) and conventional (including films, videos, multimedia, and emails) platforms did not differ significantly in their impacts on mental health. The commonly administered dosage of 10 weeks for digital mental health interventions, including those with DMHL components, demonstrated a similar positive intervening effect on mental health to that of interventions with dosages below and above 10 weeks.

Implementation and Intervening Effectiveness of DMHL: Uptake and Mental Health

We found that DMHL interventions had a similar impact on mental health to that of traditional face-to-face MHL interventions. Engagement with DMHL interventions, which was assessed as the percentage of users who completed the intervention, ranged from 13.1% to 100% (dropout of 0% to 86.9%). However, participation in DMHL interventions did not differ from that in the control conditions, and uptake of DMHL

XSL•FO
RenderX

interventions did not moderate their impact on proximal (DMHL) or distal (mental health) outcomes. Studies of the facilitators of and barriers to the uptake of digital mental health interventions argue that engagement affects intervening effectiveness on mental health outcomes [8,26]. Our review and meta-analysis found limited work on the implementation effectiveness of DMHL interventions involving engagement and uptake that was focused primarily on attrition, indicating a need for future studies to consider other engagement indicators that demonstrate participation in interventions [8,26]. For instance, studies could consider the extent of content accessed (eg, number of modules completed) or engagement in intervention-related activities (eg, number of log-ins or visits, time spent, specific activities or exercises completed, and number of web-based interactions with therapists or peers) [8,26].

Unlike the well-established research on traditional forms of mental health interventions [2], research on digital mental health interventions that leverage technological advancement is nascent [9,41]. Therefore, our findings make important contributions to the growing body of evidence that the implementation and clinical effectiveness of digital mental health interventions are comparable with, if not superior to, those of their traditional face-to-face counterparts [40]. Our study highlights the need for future research to elucidate the development and deployment practices of DMHL interventions for mental health that are scalable and cost-effective and maximize reach [92,93], which can overcome the shortcomings of traditional face-to-face MHL interventions [27,28]. In addition, our meta-analytic results indicate that DMHL interventions have a moderate effect on enhancing mental health outcomes and a strong effect on literacy outcomes. These findings suggest a mechanism of change involving MHL as a mediator or proximal outcome that in turn affects the distal outcomes of mental health conditions [94]. Given that few studies have examined the 5 facets of DMHL conjointly and separately with mental health functioning [16,55], research is needed to elucidate whether and how the 5 facets of DMHL relate to mental health—in particular whether certain facets are essential (or not) or whether combinations of facets yield stronger impacts.

Our meta-analytic efforts to synthesize and compare findings on interventions with different DMHL components revealed aspects of DMHL interventions that were the most promising or effective in bolstering mental health functioning. In particular, *DMHL plus*, which incorporated DMHL as a secondary component with other active treatment components such as skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; and informal, nonprofessional counselor interactions [15,21], was the most effective. DMHL plus interventions were superior to waitlist control [15,50] and *DMHL only* conditions [21,51] and comparable with non-DMHL interventions in enhancing mental health functioning [28,52]. Such findings are not surprising given that *DMHL plus* includes some form of active treatment. However, basic forms of DMHL intervention that involve self-help DMHL psychoeducation that builds MHL can be more effective in bolstering mental health than the absence of such interventions [20,21]. This notion was supported by our findings—*DMHL only* interventions that involved self-help DMHL psychoeducation were more effective than waitlist control conditions in enhancing mental health functioning, with large effect sizes. Beyond fundamental effectiveness, in which mental health increases after as compared with before the intervention, DMHL interventions effectively enhanced mental health as compared with waitlist control conditions [21,51]. DMHL interventions had similar effectiveness to that of non-DMHL interventions such as treatment or care as usual that involved professional therapies [28,52].

Our review and meta-analytic results help reconcile the inconsistent findings on the effect of DMHL interventions on mental health documented in the existing literature that resulted from the comparisons of different DMHL components and study designs [15,16]. Although scholars argue that digital mental health interventions, especially DMHL interventions, have the potential to intervene in the mental health of individuals, the absence of consolidated evidence for DMHL interventions' (clinical) effectiveness presents a roadblock in promoting DMHL as a form of mental health prevention on a global scale [8,9]. Our meta-analytic efforts provide strong support for DMHL interventions achieving mental health effects. DMHL psychoeducation alone or coupled with active treatments is effective in bolstering mental health functioning compared with no intervention. However, for optimal mental health impact, self-help DMHL psychoeducation is not as effective as that provided with other active treatment components, such as skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; and informal, nonprofessional counselor interactions. Of note, DMHL interventions that incorporated DMHL into other active treatment components had similar effectiveness to that of treatment as usual and professional therapies (non-DMHL interventions). These findings lend credence to DMHL as a scalable upstream prevention that translates to real-world mental health benefits in bolstering individuals' mental health functioning [8,9].

Our synthesis of findings across studies on the overall effectiveness of DMHL interventions on mental health, combined with current work demonstrating the ease of access and cost-efficiency of DMHL interventions [27,28], supports the adoption and implementation of DMHL interventions at scale. Such efforts are beneficial at the simple level of self-help DMHL psychoeducation but are even more impactful when coupled with other active treatments. Interventions that combined DMHL with other treatments had the same mental health impact as treatment as usual and professional therapies. Such findings underscore the importance of rigorous intervention designs that examine different components. In particular, using randomized controlled trials (RCTs) to compare and distinguish the impacts of DMHL in combination with different active treatment components such as skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; informal, nonprofessional counselor interactions; and care as usual in changing mental health functioning is pivotal. Such a nuanced approach can ascertain how DMHL should be supplemented with active treatment as a necessary and sufficient intervening factor and to establish

XSL•FO
RenderX

the relationship to proximal literacy outcomes and distal mental health functioning as mechanisms of change [91].

A key concern regarding digital mental health interventions is their long-lasting impact [36,40]. Particularly for DMHL interventions, the focus on knowledge, beliefs, and attitudes surrounding mental health may render their effectiveness short-lived [46,47]. In contrast, digital mental health interventions such as internet-based CBT are more well established in targeting behavior change and long-term mental health impacts [26]. Our meta-analytic results provide evidence of sustained positive effects of DMHL interventions on mental health. Specifically, studies on DMHL interventions that evaluated immediate postintervention effects and those that assessed carryover effects for as long 34 weeks demonstrated comparable mental health outcomes, particularly in mitigating mental health problems such as depression, anxiety, loneliness, and internalizing and externalizing symptoms [24,25] and bolstering mental well-being, for instance, resilience, life satisfaction, and quality of life [55,95]. More importantly, the sustained effects of DMHL interventions on improved mental health were not attenuated with longer follow-up assessments.

The long-term effects of DMHL interventions on mental health, which averaged 18.2 (SD 3.49; range 4-34) weeks in this synthesis, suggest that the 5 facets of DMHL may be involved in a chain reaction or cascade effect on mental health [94]. Details as to which facets could not be obtained because of a dearth of research that examined all 5 facets of DMHL and mental well-being conjointly and separately. Thus, there are insufficient studies (and number of effect sizes) to determine how the 5 facets of DMHL compare regarding their impacts on mental health. By building knowledge, beliefs, and attitudes regarding mental health, DMHL interventions may have adaptive functions that influence other domains or levels of function (eg, behavior change in treatment uptake and adherence) that spread over time to promote positive mental health development [94]. Future research should unravel the role of the 5 facets of DMHL in well-timed and targeted interventions that promote positive mental health cascades.

DMHL Intervention Features, Individual Characteristics, and Cultural Contexts as Moderating Factors

The effect of DMHL interventions on mental health holds across new interactive platforms and more conventional, less interactive platforms. Although there are numerous studies examining affordances of digital platforms, such as asynchronicity, anonymity, and social interactions [9], we have a limited understanding of how specific platform affordances interact with mental health—whether they bolster positive mental health or generate and exacerbate mental health problems. Specific to our study, we investigated the interactivity afforded by new and conventional digital platforms in moderating the impact of DMHL interventions on mental health as there are limited studies on DMHL interventions that tap into the full range of digital modalities [9]. Studies investigating social media platforms note different types of affordances that vary across platforms [96]. Importantly, the additive and interactive effects of different affordances in a specific platform could amplify users' emotional experiences and expressions, resulting in heightened emotional lability and susceptibility to mental health conditions [96]. However, empirical work on whether and how the affordances of social media and other digital platforms affect the effectiveness of digital mental health interventions, including DMHL interventions, is lacking. Future work on DMHL interventions should consider specific platform affordances and their interactions when modulating intervention effectiveness.

Unexpectedly, our synthesis did not find dosage effect. Although other studies have found the median dosage or duration of interventions to be approximately 10 weeks [60], our meta-regressions indicated that DMHL interventions with a dosage of 10 weeks had similar impacts on enhancing mental health as dosages below and above 10 weeks. Our amalgamated findings suggest that the 10-week dosage that is commonly adopted by digital mental health interventions may not be optimal for DMHL interventions [61,62]. More research involving RCTs on digital mental health and DMHL interventions alike that examine the effects of varying dosages of interventions on mental health outcomes is warranted. However, our findings are promising given the challenges with mental health services and MHL intervention retention found in many countries [97,98]. Collectively, our meta-analysis and meta-regressions on the moderating role of DMHL intervention features provide important insights into best practices in DMHL that enhance mental health. Consistent with global efforts to raise mental health awareness and improve poor MHL in the public [1], this synthesis advances our understanding of DMHL by building a conceptual model of MHL [4] that could inform the design and implementation of evidence-based DMHL interventions [8].

DMHL intervention effects in this synthesis were comparable across sex and severity of mental health conditions. Other studies have found that female participants have higher levels of MHL [64,65], which can moderate the effects of DMHL interventions on mental health outcomes. When controlling for pre- and postintervention levels of MHL, we found that DMHL interventions had equivalent effects on increasing the mental health functioning of both male and female participants. Contrary to findings on the reduced effectiveness of digital mental health interventions in individuals with mental illness or with physical ill health [8], we found that the severity of mental health conditions did not attenuate the effectiveness of DMHL interventions on the mental health of healthy populations. We expected DMHL interventions to demonstrate greater effects on adolescents than on emerging and older adults given that adolescents are the primary users of digital platforms such as social media, mobile apps, and web-based mental health resources [30]. However, DMHL interventions appear to be most effective in emerging and older adults. The conceptual model of MHL describes the 5 facets of DMHL as positively associated with maturity, cognitive advancement, and educational attainment in adulthood [4]. As such, this helps explain why DMHL interventions had a greater impact on emerging and older adults' mental health. With increasing mental health concerns in adolescence [32,33], future research needs to illuminate whether and how some facets of DMHL are applicable to bolstering adolescent mental health and target

XSL•FO
RenderX

these facets in the design and implementation of DMHL interventions for youth mental health.

In contrast to our hypothesis, the impact of DMHL interventions did not differ between Western and Eastern cultural contexts. Most Asian societies with collectivistic cultural characteristics are conservative in the way in which they address mental health concerns and often promote MHL and positive mental health functioning in ways that differ from those of Western cultural contexts with individualistic cultural characteristics. For instance, cultivating interdependence relations, relational harmony, and dialectical beliefs and emotions may be more common in Eastern cultures [68]. Furthermore, the Asian community as a whole, including policy makers, mental health professionals, and the public, is only beginning to embrace the notion of digital mental health interventions [99,100]. On the other hand, studies on digital mental health interventions, especially DMHL interventions, are conducted predominantly in Western cultural contexts in high-income countries (Global North) [8,9], where there is greater public awareness of and less stigma associated with mental health conditions and greater focus on developing a healthy sense of independent self, including positive mental health [67,68].

Consistent with most mental health research [42-44], Western cultural contexts were overrepresented in the DMHL interventions, particularly countries in the Global North such as Australia, the United States, and the United Kingdom. Comparatively, there are limited studies conducted in the Global South that involve Eastern cultural contexts in Asian countries and scant research on DMHL interventions from South American and African countries, limiting assessments of DMHL intervention effectiveness in the Global South. Furthermore, most studies on DMHL interventions included in this review did not provide information on the ethnic or racial composition of their samples, and those that provided this information had samples that were predominantly White with limited racial or ethnic diversity. Our findings reinforce existing research on digital mental health interventions broadly that indicated that these interventions were designed for and evaluated on implementation and clinical effectiveness with largely homogeneous samples that are disproportionately White, from the Global North, and likely cissex [42-44]. Findings from our review and meta-analysis on DMHL, coupled with those on digital mental health interventions [42-44], highlight the need for more consideration of diversity in research but broadly support the utility of DMHL interventions.

In general, there is lower MHL, greater mental health stigma, and fewer services provided in the Global South [42-44]. DMHL interventions may be an important and accessible mental health resource that could not only provide greater access and reach and reduce stigma but also improve MHL among the public [42-44]. Our review found that DMHL interventions are implemented and evaluated mainly in Western cultural contexts in high-income countries [42-44] and some contemporary Asian societies [99,100]. Our results call for greater research attention to the design, implementation, and evaluation of culturally appropriate DMHL interventions in Asian, South American, and African countries, as well as more diversity in research in the Global North. Contextual stressors associated with racial and ethnic marginalization and mental health disparities, including lower access to and greater barriers to engaging in digital mental health interventions, underscore the need for future work on DMHL interventions specifically and digital mental health interventions broadly that target racial or ethnic minority groups [42-44].

Limitations

Our systematic review and meta-analysis highlights 4 major limitations in the broader literature on DMHL interventions. First, there is a lack of a clear conceptualization that distinguishes the DMHL components assessed in the interventions. Our review of existing literature and efforts to synthesize findings on the effectiveness of DMHL interventions suggest that, conceptually, *DMHL only* interventions refer to those that implement DMHL as a self-help psychoeducation component. In contrast, *DMHL plus* interventions incorporate DMHL as a secondary component with other active treatment components that are nonprofessional and informal in nature, for instance, skills training; peer support; group discussions and activities; exercises such as diary entries and reflection logs; and informal counselor interactions, and non-DMHL interventions refer to treatment as usual that involves client-professional visits, interactions, and therapies. Although receiving professional mental health treatment likely improves knowledge about mental health symptoms and management (eg, MHL), such interventions are not specifically focused on literacy. As such, they are categorized as non-DMHL interventions. These conceptual distinctions in DMHL interventional components have yet to be acknowledged in the current research despite the fact that all DMHL interventions included in our meta-analysis examined various conditions that could be classified as *DMHL only*, *DMHL plus*, and non-DMHL.

Second, DMHL interventions are not always clear about the facet of DMHL examined and rarely refer to the theoretical framework of DMHL that distinguishes five literacy facets [5,6]: (1) knowledge about obtaining and maintaining good mental health, (2) understanding mental illnesses and treatments, (3) reducing mental illness–related stigma, (4) enhancing help-seeking efficacy or behaviors, and (5) enhancing help-seeking attitudes or intentions. We recommend that future research recognize the conceptualization of DMHL facets and be clear about the facet of DMHL that demonstrates a positive relationship to or efficacy in enhancing mental health. Such an approach would facilitate a comprehensive and converged understanding of the conditions under which DMHL most strongly relates to and effectively increases mental health functioning. On a similar note, DMHL interventions do not always include both proximal literacy outcomes and distal mental health outcomes, which are important in establishing mechanisms of change with proximal literacy outcomes as possible mediators. As the field of DMHL interventions is nascent [8,9], studies have primarily focused on examining mental health outcomes (clinical outcomes) [8,9], and a major gap remains in understanding implementation effectiveness involving uptake and engagement. Drawing from work on digital mental health interventions, user uptake and engagement can vary across different indicators and affect intervening effectiveness on mental health outcomes [8,26]. Extant research

XSL•FO
RenderX

on digital mental health interventions, including DMHL, largely focuses on attrition as an indicator of uptake and engagement. Future DMHL interventions should consider other engagement indicators to demonstrate the extent of intervention use and potential key features for effectiveness.

Third, the features of DMHL interventions, such as dosage and platform affordances, warrant greater attention given that research on digital mental health interventions has demonstrated how these features can influence user engagement and the effectiveness of the intervention [8,9]. This synthesis was limited to sex comparisons involving cissex individuals even though transsex and nonbinary individuals are at greater risk of mental health issues. Future work needs to expand the consideration of sex beyond binary operationalizations. We did not find evidence for the role of the commonly endorsed dosage of 10 weeks of digital mental health interventions that included DMHL components and of platform interactivity affordances in amplifying the mental health impact of DMHL interventions. Future work should examine varying dosages of DMHL interventions and different affordances and their additive and interactive effects through empirical examination using an RCT study design.

Finally, most DMHL interventions (56/76, 74%) were found in Western cultural contexts, especially in high-income countries (Global North), and most studies on DMHL interventions (36/76, 47%) did not report information on the racial or ethnic composition of their samples. Among those that did, the samples were predominantly White with limited racial or ethnic diversity. As such, our findings may not be generalizable to other geographic regions or demographic groups. In light of contextual stressors associated with racial and ethnic marginalization and mental health disparities, future work on DMHL interventions that target racial or ethnic minority groups, particularly the design, adoption, and evaluation of the effects of culturally adaptive DMHL interventions on uptake and mental health functioning, is needed [42-44]. Collectively, our review and meta-analysis of DMHL and mental health makes important theoretical and practical contributions. For theory building on DMHL, we found evidence for the fundamental effectiveness of DMHL interventions that increased postintervention mental health and strong inferences for DMHL interventions' effectiveness in bolstering mental health as compared with waitlist control conditions. However, DMHL interventions confer optimal effects on mental health when DMHL psychoeducation is incorporated with informal, nonprofessional active treatment components such as skills training and peer support, which demonstrates comparable effectiveness with that of treatment as usual involving client-professional interactions and therapies. However, none of the interventions in this review considered the mechanism of action regarding how DMHL affects proximal outcomes that in turn affect distal mental health conditions. The large effect size of DMHL interventions on MHL outcomes indicates that literacy may serve as a mediating mechanism for enhancing mental health functioning. Thus, the MHL framework needs to unpack and incorporate different components of DMHL, especially in how they relate to the 5 DMHL facets, and consider their interplay with various active treatment components. More importantly, proximal and distal factors involved in the mechanism of action of DMHL and mental health need to be examined to build and expand the MHL framework.

Conclusions

Our review and meta-analysis found that DMHL interventions are as effective as face-to-face interventions. Basic DMHL interventions with self-help DMHL psychoeducation had similar effectiveness to that of interventions that incorporated DMHL as a secondary component with other active treatment components in bolstering mental health functioning. These findings are practically meaningful and underscore the feasibility and promise of digital modalities for improving mental health. DMHL interventions greatly increased literacy outcomes and moderately improved mental health functioning by reducing depression, anxiety, loneliness, and internalizing and externalizing symptoms and enhancing quality of life and resilience. Importantly, these effects, which did not differ by platform type or dosage, were sustained over time. Future research is needed to test our findings on the circumstances in which DMHL interventions are the most effective in enhancing mental health—specific DMHL components, dosage, extent of carryover effects, platform affordances, and individual and contextual factors—to aid policy makers, mental health professionals, and social services in establishing high-performance DMHL interventions that enhance mental health in the community.

Acknowledgments

The authors thank the following research interns for their help with the literature review and data extraction: Cameron Tan, Richelle Ho, and Fabian Fu.

Data Availability

The data sets generated during and/or analyzed during this study are available on the Open Science Framework [101].

Authors' Contributions

GHY conceptualized the study; conducted the literature review, data cleaning, coding, and analyses; and drafted and revised the manuscript. SMR drafted and revised the manuscript. EYMC and NAL conducted the risk-of-bias (quality) assessment and literature review.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search strings.
[DOCX File , 31 KB - jmir_v26i1e51268_app1.docx]

Multimedia Appendix 2
Characteristics of the included digital mental health literacy interventions and their traditional face-to-face counterparts.
[DOCX File , 142 KB - jmir_v26i1e51268_app2.docx]

Multimedia Appendix 3
Summary and sample characteristics of the digital mental health literacy interventions.
[DOCX File , 40 KB - jmir_v26i1e51268_app3.docx]

Multimedia Appendix 4
Forest plot of data investigating the effects of digital mental health literacy and traditional face-to-face mental health literacy interventions on mental health.
[DOCX File , 516 KB - jmir_v26i1e51268_app4.docx]

Multimedia Appendix 5
Funnel plot of the effect sizes of the digital mental health literacy and traditional face-to-face mental health literacy interventions on mental health.
[DOCX File , 515 KB - jmir_v26i1e51268_app5.docx]

Multimedia Appendix 6
Moderators of digital mental health literacy interventions and mental health.
[DOCX File , 21 KB - jmir_v26i1e51268_app6.docx]

Multimedia Appendix 7
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 120 KB - jmir_v26i1e51268_app7.pdf]

References

1. The Lancet Global Health. Mental health matters. Lancet Glob Health 2020 Nov;8(11):e1352 [FREE Full text] [doi: 10.1016/S2214-109X(20)30432-0] [Medline: 33069297]
2. Furnham A, Swami V. Mental health literacy: a review of what it is and why it matters. Int Perspect Psychol 2018 Oct;7(4):240-257 [FREE Full text] [doi: 10.1037/ipp0000094]
3. Jorm AF, Korten AE, Jacomb PA, Christensen H, Rodgers B, Pollitt P. "Mental health literacy": a survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. Med J Aust 1997 Feb 17;166(4):182-186. [doi: 10.5694/j.1326-5377.1997.tb140071.x] [Medline: 9066546]
4. Jorm AF. Mental health literacy: empowering the community to take action for better mental health. Am Psychol 2012 Apr;67(3):231-243. [doi: 10.1037/a0025957] [Medline: 22040221]
5. Kutcher S, Wei Y, Coniglio C. Mental health literacy: past, present, and future. Can J Psychiatry 2016 Mar;61(3):154-158 [FREE Full text] [doi: 10.1177/0706743715616609] [Medline: 27254090]
6. Chao HJ, Lien YJ, Kao YC, Tasi IC, Lin HS, Lien YY. Mental health literacy in healthcare students: an expansion of the mental health literacy scale. Int J Environ Res Public Health 2020 Feb 04;17(3):948 [FREE Full text] [doi: 10.3390/ijerph17030948] [Medline: 32033015]
7. Furnham A. Explaining health and illness: lay perceptions on current and future health, the causes of illness, and the nature of recovery. Soc Sci Med 1994 Sep;39(5):715-725. [doi: 10.1016/0277-9536(94)90026-4] [Medline: 7973869]
8. Borghouts J, Eikey E, Mark G, de Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
9. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]

XSL•FO
RenderX

10. Lai HJ, Lien YJ, Chen KR, Lin YK. The effectiveness of mental health literacy curriculum among undergraduate public health students. Int J Environ Res Public Health 2022 Apr 26;19(9):5269 [FREE Full text] [doi: 10.3390/ijerph19095269] [Medline: 35564671]
11. Mohammadi AS, Panahi S, Sayarifard A, Ashouri A. Identifying the prerequisites, facilitators, and barriers in improving adolescents' mental health literacy interventions: a systematic review. J Educ Health Promot 2020;9(1):322 [FREE Full text] [doi: 10.4103/jehp.jehp_623_20] [Medline: 33426126]
12. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
13. Wei Y, Kutcher S, Austen E, Comfort A, Gilham C, MacDougall C, et al. The impact of transitions, a mental health literacy intervention with embedded life skills for postsecondary students: preliminary findings from a naturalistic cohort study. Can J Psychiatry 2022 Jun 11;67(6):452-461. [doi: 10.1177/07067437211037131] [Medline: 34379024]
14. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
15. Bakker D, Kazantzis N, Rickwood D, Rickard N. A randomized controlled trial of three smartphone apps for enhancing public mental health. Behav Res Ther 2018 Oct;109:75-83. [doi: 10.1016/j.brat.2018.08.003] [Medline: 30125790]
16. Lintvedt OK, Griffiths KM, Sørensen K, Østvik AR, Wang CE, Eisemann M, et al. Evaluating the effectiveness and efficacy of unguided internet-based self-help intervention for the prevention of depression: a randomized controlled trial. Clin Psychol Psychother 2013;20(1):10-27. [doi: 10.1002/cpp.770] [Medline: 21887811]
17. Koike S, Yamaguchi S, Ojio Y, Ohta K, Shimada T, Watanabe K, et al. A randomised controlled trial of repeated filmed social contact on reducing mental illness-related stigma in young adults. Epidemiol Psychiatr Sci 2016 Dec 19;27(2):199-208. [doi: 10.1017/s2045796016001050]
18. Yamaguchi S, Ojio Y, Ando S, Bernick P, Ohta K, Watanabe KI, et al. Long-term effects of filmed social contact or internet-based self-study on mental health-related stigma: a 2-year follow-up of a randomised controlled trial. Soc Psychiatry Psychiatr Epidemiol 2019 Jan 12;54(1):33-42. [doi: 10.1007/s00127-018-1609-8] [Medline: 30315333]
19. Fogarty AS, Proudfoot J, Whittle EL, Clarke J, Player MJ, Christensen H, et al. Preliminary evaluation of a brief web and mobile phone intervention for men with depression: men's positive coping strategies and associated depression, resilience, and work and social functioning. JMIR Ment Health 2017 Aug 10;4(3):e33 [FREE Full text] [doi: 10.2196/mental.7769] [Medline: 28798009]
20. Hui A, Wong PW, Fu KW. Evaluation of an online campaign for promoting help-seeking attitudes for depression using a Facebook advertisement: an online randomized controlled experiment. JMIR Ment Health 2015;2(1):e5 [FREE Full text] [doi: 10.2196/mental.3649] [Medline: 26543911]
21. Lillevoll KR, Vangberg HC, Griffiths KM, Waterloo K, Eisemann MR. Uptake and adherence of a self-directed internet-based mental health intervention with tailored e-mail reminders in senior high schools in Norway. BMC Psychiatry 2014 Jan 21;14:14 [FREE Full text] [doi: 10.1186/1471-244X-14-14] [Medline: 24443820]
22. Carl JR, Miller CB, Henry AL, Davis ML, Stott R, Smits JA, et al. Efficacy of digital cognitive behavioral therapy for moderate-to-severe symptoms of generalized anxiety disorder: a randomized controlled trial. Depress Anxiety 2020 Dec 29;37(12):1168-1178. [doi: 10.1002/da.23079] [Medline: 32725848]
23. Mak WW, Tong AC, Yip SY, Lui WW, Chio FH, Chan AT, et al. Efficacy and moderation of mobile app-based programs for mindfulness-based training, self-compassion training, and cognitive behavioral psychoeducation on mental health: randomized controlled noninferiority trial. JMIR Ment Health 2018 Oct 11;5(4):e60 [FREE Full text] [doi: 10.2196/mental.8597] [Medline: 30309837]
24. Graham AK, Greene CJ, Kwasny MJ, Kaiser SM, Lieponis P, Powell T, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 01;77(9):906-914 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
25. Radovic A, Li Y, Landsittel D, Stein BD, Miller E. A social media website (supporting our valued adolescents) to support treatment uptake for adolescents with depression and/or anxiety and their parents: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2019 Jan 23;8(1):e12117 [FREE Full text] [doi: 10.2196/12117] [Medline: 30672741]
26. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
27. Ebert DD, Heber E, Berking M, Riper H, Cuijpers P, Funk B, et al. Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 2016 May;73(5):315-323. [doi: 10.1136/oemed-2015-103269] [Medline: 26884049]
28. Levin ME, Hayes SC, Pistorello J, Seeley JR. Web-based self-help for preventing mental health problems in universities: comparing acceptance and commitment training to mental health education. J Clin Psychol 2016 Mar;72(3):207-225. [doi: 10.1002/jclp.22254] [Medline: 26784010]

XSL•FO
RenderX

29. Axelsson E, Andersson E, Ljótsson B, Björkander D, Hedman-Lagerlöf M, Hedman-Lagerlöf E. Effect of internet vs face-to-face cognitive behavior therapy for health anxiety: a randomized noninferiority clinical trial. JAMA Psychiatry 2020 Sep 01;77(9):915-924 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.0940] [Medline: 32401286]
30. Wetterlin FM, Mar MY, Neilson EK, Werker GR, Krausz M. eMental health experiences and expectations: a survey of youths' Web-based resource preferences in Canada. J Med Internet Res 2014 Dec 17;16(12):e293 [FREE Full text] [doi: 10.2196/jmir.3526] [Medline: 25519847]
31. Burns JM, Davenport TA, Durkin LA, Luscombe GM, Hickie IB. The internet as a setting for mental health service utilisation by young people. Med J Aust 2010 Jun 07;192(S11):S22-S26. [doi: 10.5694/j.1326-5377.2010.tb03688.x] [Medline: 20528703]
32. Adolescent and young adult health. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions [accessed 2022-03-17]
33. Adolescent mental health. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2022-03-17]
34. van Doesum KT, Riebschleger J, Carroll J, Grové C, Lauritzen C, Mordoch E, et al. Successful recruitment strategies for prevention programs targeting children of parents with mental health challenges: an international study. Child Youth Serv 2016 Mar 30;37(2):156-174. [doi: 10.1080/0145935X.2016.1104075]
35. Reupert A, Maybery D, Bartholomew C, Cuff R, Foster K, Matar J, et al. The acceptability and effectiveness of an online intervention for youth with parents with a mental illness and/or substance use issue. J Adolesc Health 2020 May;66(5):551-558. [doi: 10.1016/j.jadohealth.2019.11.309] [Medline: 32001142]
36. Kurki M, Gilbert S, Mishina K, Lempinen L, Luntamo T, Hinkka-Yli-Salomäki S, et al. Digital mental health literacy -program for the first-year medical students' wellbeing: a one group quasi-experimental study. BMC Med Educ 2021 Nov 06;21(1):563 [FREE Full text] [doi: 10.1186/s12909-021-02990-4] [Medline: 34742258]
37. Lam LT, Lam MK, Reddy P, Wong P. Efficacy of a workplace intervention program with web-based online and offline modalities for improving workers' mental health. Front Psychiatry 2022 May 31;13:888157 [FREE Full text] [doi: 10.3389/fpsyt.2022.888157] [Medline: 35711597]
38. Evans-Lacko S, Hahn JS, Peter L, Schomerus G. The impact of digital interventions on help-seeking behaviour for mental health problems: a systematic literature review. Curr Opin Psychiatry 2022 May 01;35(3):207-218. [doi: 10.1097/YCO.0000000000000788] [Medline: 35579875]
39. Fraser E, Pakenham KI. Resilience in children of parents with mental illness: relations between mental health literacy, social connectedness and coping, and both adjustment and caregiving. Psychol Health Med 2009 Oct 19;14(5):573-584. [doi: 10.1080/13548500903193820] [Medline: 19844836]
40. O'Connell J, Shafran R, Pote H. A randomized controlled trial evaluating the effectiveness of face-to-face and digital training in improving child mental health literacy rates in frontline pediatric hospital staff. Front Psychiatry 2020 Feb 11;11:570125 [FREE Full text] [doi: 10.3389/fpsyt.2020.570125] [Medline: 33643077]
41. Peyton D, Goods M, Hiscock H. The effect of digital health interventions on parents' mental health literacy and help seeking for their child's mental health problem: systematic review. J Med Internet Res 2022 Feb 10;24(2):e28771 [FREE Full text] [doi: 10.2196/28771] [Medline: 35142623]
42. Ellis DM, Draheim AA, Anderson PL. Culturally adapted digital mental health interventions for ethnic/racial minorities: a systematic review and meta-analysis. J Consult Clin Psychol 2022 Oct;90(10):717-733. [doi: 10.1037/ccp0000759] [Medline: 36227330]
43. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
44. McGuire TG, Miranda J. New evidence regarding racial and ethnic disparities in mental health: policy implications. Health Aff (Millwood) 2008;27(2):393-403 [FREE Full text] [doi: 10.1377/hlthaff.27.2.393] [Medline: 18332495]
45. Brooks H, Irmansyah I, Syarif AK, Pedley R, Renwick L, Rahayu AP, et al. Evaluating a prototype digital mental health literacy intervention for children and young people aged 11-15 in Java, Indonesia: a mixed methods, multi-site case study evaluation. Child Adolesc Psychiatry Ment Health 2023 Jun 26;17(1):79 [FREE Full text] [doi: 10.1186/s13034-023-00608-9] [Medline: 37365594]
46. Andrews JL, Birrell L, Chapman C, Teesson M, Newton N, Allsop S, et al. Evaluating the effectiveness of a universal eHealth school-based prevention programme for depression and anxiety, and the moderating role of friendship network characteristics. Psychol Med 2023 Aug;53(11):5042-5051. [doi: 10.1017/S0033291722002033] [Medline: 35838377]
47. Moeini B, Bashirian S, Soltanian AR, Ghaleiha A, Taheri M. Examining the effectiveness of a web-based intervention for depressive symptoms in female adolescents: applying social cognitive theory. J Res Health Sci 2019 Aug 19;19(3):e00454 [FREE Full text] [Medline: 31586376]
48. Arjadi R, Nauta MH, Scholte WF, Hollon SD, Chowdhary N, Suryani AO, et al. Internet-based behavioural activation with lay counsellor support versus online minimal psychoeducation without support for treatment of depression: a randomised controlled trial in Indonesia. Lancet Psychiatry 2018 Sep;5(9):707-716. [doi: 10.1016/S2215-0366(18)30223-2] [Medline: 30006262]

XSL•FO
RenderX

49. Nguyen-Feng VN, Greer CS, Frazier P. Using online interventions to deliver college student mental health resources: evidence from randomized clinical trials. Psychol Serv 2017 Nov;14(4):481-489. [doi: 10.1037/ser0000154] [Medline: 29120206]
50. Moessner M, Minarik C, Özer F, Bauer S. Can an internet-based program for the prevention and early intervention in eating disorders facilitate access to conventional professional healthcare? J Ment Health 2016 Oct;25(5):441-447. [doi: 10.3109/09638237.2016.1139064] [Medline: 26850624]
51. Kawadler JM, Hemmings NR, Ponzo S, Morelli D, Bird G, Plans D. Effectiveness of a smartphone app (BioBase) for reducing anxiety and increasing mental well-being: pilot feasibility and acceptability study. JMIR Form Res 2020 Nov 10;4(11):e18067 [FREE Full text] [doi: 10.2196/18067] [Medline: 32969341]
52. Harrer M, Apolinário-Hagen J, Fritsche L, Salewski C, Zarski AC, Lehr D, et al. Effect of an internet- and app-based stress intervention compared to online psychoeducation in university students with depressive symptoms: results of a randomized controlled trial. Internet Interv 2021 Apr;24:100374 [FREE Full text] [doi: 10.1016/j.invent.2021.100374] [Medline: 33718001]
53. Moser C, Bachem R, Berger T, Maercker A. ZIEL: internet-based self-help for adjustment problems: results of a randomized controlled trial. J Clin Med 2019 Oct 11;8(10):1655 [FREE Full text] [doi: 10.3390/jcm8101655] [Medline: 31614528]
54. Tighe J, Shand F, Ridani R, Mackinnon A, de la Mata N, Christensen H. Ibobbly mobile health intervention for suicide prevention in Australian Indigenous youth: a pilot randomised controlled trial. BMJ Open 2017 Jan 27;7(1):e013518 [FREE Full text] [doi: 10.1136/bmjopen-2016-013518] [Medline: 28132007]
55. Alvarez-Jimenez M, Koval P, Schmaal L, Bendall S, O'Sullivan S, Cagliarini D, et al. The Horyzons project: a randomized controlled trial of a novel online social therapy to maintain treatment effects from specialist first-episode psychosis services. World Psychiatry 2021 Jun;20(2):233-243 [FREE Full text] [doi: 10.1002/wps.20858] [Medline: 34002511]
56. Kitchener BA, Jorm AF. Mental health first aid training in a workplace setting: a randomized controlled trial [ISRCTN13249129]. BMC Psychiatry 2004 Aug 15;4:23 [FREE Full text] [doi: 10.1186/1471-244X-4-23] [Medline: 15310395]
57. Perry Y, Petrie K, Buckley H, Cavanagh L, Clarke D, Winslade M, et al. Effects of a classroom-based educational resource on adolescent mental health literacy: a cluster randomized controlled trial. J Adolesc 2014 Oct;37(7):1143-1151. [doi: 10.1016/j.adolescence.2014.08.001] [Medline: 25151646]
58. Cieslak R, Benight CC, Rogala A, Smoktunowicz E, Kowalska M, Zukowska K, et al. Effects of internet-based self-efficacy intervention on secondary traumatic stress and secondary posttraumatic growth among health and human services professionals exposed to indirect trauma. Front Psychol 2016;7:1009 [FREE Full text] [doi: 10.3389/fpsyg.2016.01009] [Medline: 27458407]
59. Kuhn E, Kanuri N, Hoffman JE, Garvert DW, Ruzek JI, Taylor CB. A randomized controlled trial of a smartphone app for posttraumatic stress disorder symptoms. J Consult Clin Psychol 2017 Mar;85(3):267-273. [doi: 10.1037/ccp0000163] [Medline: 28221061]
60. Saleem M, Kühne L, de Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
61. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014 May 16;16(5):e130 [FREE Full text] [doi: 10.2196/jmir.3142] [Medline: 24836465]
62. Riadi I, Kervin L, Dhillon S, Teo K, Churchill R, Card KG, et al. Digital interventions for depression and anxiety in older adults: a systematic review of randomised controlled trials. Lancet Healthy Longev 2022 Aug;3(8):e558-e571. [doi: 10.1016/s2666-7568(22)00121-0]
63. Blaney C, Hitchon CA, Marrie RA, Mackenzie C, Holens P, El-Gabalawy R. Support for a non-therapist assisted, Internet-based cognitive-behavioral therapy (iCBT) intervention for mental health in rheumatoid arthritis patients. Internet Interv 2021 Apr;24:100385 [FREE Full text] [doi: 10.1016/j.invent.2021.100385] [Medline: 33912401]
64. Cormier E, Park H, Schluck G. College students' eMental health literacy and risk of diagnosis with mental health disorders. Healthcare (Basel) 2022 Nov 30;10(12):2406 [FREE Full text] [doi: 10.3390/healthcare10122406] [Medline: 36553930]
65. Bjørnsen HN, Ringdal R, Espnes GA, Eilertsen MB, Moksnes UK. Exploring MEST: a new universal teaching strategy for school health services to promote positive mental health literacy and mental wellbeing among Norwegian adolescents. BMC Health Serv Res 2018 Dec 29;18(1):1001 [FREE Full text] [doi: 10.1186/s12913-018-3829-8] [Medline: 30594201]
66. Baños R, Herrero R, Vara MD. What is the current and future status of digital mental health interventions? Span J Psychol 2022 Feb 02;25:e5. [doi: 10.1017/SJP.2022.2] [Medline: 35105398]
67. Cheon BK, Chiao JY. Cultural variation in implicit mental illness stigma. J Cross Cult Psychol 2012 Oct 01;43(7):1058-1062 [FREE Full text] [doi: 10.1177/0022022112455457] [Medline: 24311820]
68. Miyamoto Y, Ryff CD. Culture and health: recent developments and future directions. Jpn Psychol Res 2022 Apr 26;64(2):90-108 [FREE Full text] [doi: 10.1111/jpr.12378] [Medline: 35509718]

XSL•FO
RenderX

69. Subramaniam M, Abdin E, Vaingankar JA, Shafie S, Chua BY, Sambasivam R, et al. Tracking the mental health of a nation: prevalence and correlates of mental disorders in the second Singapore mental health study. Epidemiol Psychiatr Sci 2019 Apr 05;29:e29 [FREE Full text] [doi: 10.1017/S2045796019000179] [Medline: 30947763]
70. Lee YY, Ang S, Chua HC, Subramaniam M. Peer support in mental health: a growing movement in Singapore. Ann Acad Med Singap 2019 Mar;48(3):95-97 [FREE Full text] [Medline: 30997478]
71. Carter H, Araya R, Anjur K, Deng D, Naslund JA. The emergence of digital mental health in low-income and middle-income countries: a review of recent advances and implications for the treatment and prevention of mental disorders. J Psychiatr Res 2021 Jan;133:223-246 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.12.016] [Medline: 33360867]
72. Quintana DS. From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. Front Psychol 2015 Oct 08;6:1549 [FREE Full text] [doi: 10.3389/fpsyg.2015.01549] [Medline: 26500598]
73. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
74. Cochrane community. Covidence. URL: https://community.cochrane.org/help/tools-and-software/covidence [accessed 2023-07-01]
75. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. J R Stat Soc Ser A Stat Soc 2006 Jan;169(1):5-35 [FREE Full text] [doi: 10.1111/j.1467-985x.2005.00377.x]
76. Mak A, Cheung MW, Ho RC, Cheak AA, Lau CS. Bisphosphonates and atrial fibrillation: bayesian meta-analyses of randomized controlled trials and observational studies. BMC Musculoskelet Disord 2009 Sep 21;10(1):113 [FREE Full text] [doi: 10.1186/1471-2474-10-113] [Medline: 19772579]
77. Hay P, Mond J, Paxton S, Rodgers B, Darby A, Owen C. What are the effects of providing evidence-based information on eating disorders and their treatments? a randomized controlled trial in a symptomatic community sample. Early Interv Psychiatry 2007 Nov 27;1(4):316-324. [doi: 10.1111/j.1751-7893.2007.00044.x] [Medline: 21352119]
78. Imamura K, Kawakami N, Tsuno K, Tsuchiya M, Shimada K, Namba K. Effects of web-based stress and depression literacy intervention on improving symptoms and knowledge of depression among workers: a randomized controlled trial. J Affect Disord 2016 Oct;203:30-37. [doi: 10.1016/j.jad.2016.05.045] [Medline: 27280960]
79. Rodante DE, Kaplan MI, Olivera Fedi R, Gagliesi P, Pascali A, José Quintero PS, et al. CALMA, a mobile health application, as an accessory to therapy for reduction of suicidal and non-suicidal self-injured behaviors: a pilot cluster randomized controlled trial. Arch Suicide Res 2022 Oct 18;26(2):801-818. [doi: 10.1080/13811118.2020.1834476] [Medline: 33073745]
80. MacLean S, Corsi DJ, Litchfield S, Kucharski J, Genise K, Selaman Z, et al. Coach-facilitated web-based therapy compared with information about web-based resources in patients referred to secondary mental health care for depression: randomized controlled trial. J Med Internet Res 2020 Jun 09;22(6):e15001 [FREE Full text] [doi: 10.2196/15001] [Medline: 32515740]
81. Espie CA, Emsley R, Kyle SD, Gordon C, Drake CL, Siriwardena AN, et al. Effect of digital cognitive behavioral therapy for insomnia on health, psychological well-being, and sleep-related quality of life: a randomized clinical trial. JAMA Psychiatry 2019 Jan 01;76(1):21-30 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.2745] [Medline: 30264137]
82. Schmidt FL, Hunter JE. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Thousand Oaks, CA: Sage Publications; 2014.
83. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. Psychol Methods 2015 Sep;20(3):375-393. [doi: 10.1037/met0000011] [Medline: 24773356]
84. O'Connor M, Casey L. The Mental Health Literacy Scale (MHLS): a new scale-based measure of mental health literacy. Psychiatry Res 2015 Sep 30;229(1-2):511-516. [doi: 10.1016/j.psychres.2015.05.064] [Medline: 26228163]
85. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998 Jun;52(6):377-384 [FREE Full text] [doi: 10.1136/jech.52.6.377] [Medline: 9764259]
86. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
87. Fisher Z, Tipton E. Robumeta: robust variance meta-regression. Cran R. URL: http://CRAN.R-project.org/package=robumeta [accessed 2015-03-07]
88. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2024-01-29]
89. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Soft 2010;36(3):1-48 [FREE Full text] [doi: 10.18637/jss.v036.i03]
90. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
91. Krendl AC, Pescosolido BA. Countries and cultural differences in the stigma of mental illness: the east–west divide. J Cross Cult Psychol 2020 Feb 21;51(2):149-167. [doi: 10.1177/0022022119901297]

XSL•FO
RenderX

92. Bernecker SL, Zuromski KL, Curry JC, Kim JJ, Gutierrez PM, Joiner TE, et al. Economic evaluation of brief cognitive behavioral therapy vs treatment as usual for suicidal US army soldiers. JAMA Psychiatry 2020 Mar 01;77(3):256-264 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.3639] [Medline: 31774485]
93. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020 Jul 01;3(7):e2010918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
94. Domhardt M, Engler S, Nowak H, Lutsch A, Baumel A, Baumeister H. Mechanisms of change in digital health interventions for mental disorders in youth: systematic review. J Med Internet Res 2021 Nov 26;23(11):e29742 [FREE Full text] [doi: 10.2196/29742] [Medline: 34842543]
95. Eustis EH, Hayes-Skelton SA, Orsillo SM, Roemer L. Surviving and thriving during stress: a randomized clinical trial comparing a brief web-based therapist-assisted acceptance-based behavioral intervention versus waitlist control for college students. Behav Ther 2018 Nov;49(6):889-903. [doi: 10.1016/j.beth.2018.05.009] [Medline: 30316488]
96. Nesi J, Choukas-Bradley S, Prinstein MJ. Transformation of adolescent peer relations in the social media context: part 1-A theoretical framework and application to dyadic peer relationships. Clin Child Fam Psychol Rev 2018 Sep 7;21(3):267-294 [FREE Full text] [doi: 10.1007/s10567-018-0261-x] [Medline: 29627907]
97. Benjet C, Borges G, Orozco R, Aguilar-Gaxiola S, Andrade LH, Cia A, et al. Dropout from treatment for mental disorders in six countries of the Americas: a regional report from the World Mental Health Surveys. J Affect Disord 2022 Apr 15;303:168-179. [doi: 10.1016/j.jad.2022.02.019] [Medline: 35151675]
98. Marinucci A, Grové C, Allen K. A scoping review and analysis of mental health literacy interventions for children and youth. Sch Psychol Rev 2022 Feb 08;52(2):144-158. [doi: 10.1080/2372966x.2021.2018918]
99. Shorey S, Chua JY. Effectiveness of peer support interventions for adults with depressive symptoms: a systematic review and meta-analysis. J Ment Health 2023 Apr 07;32(2):465-479. [doi: 10.1080/09638237.2021.2022630] [Medline: 34994272]
100. Zhang X, Lewis S, Firth J, Chen X, Bucci S. Digital mental health in China: a systematic review. Psychol Med 2021 Sep 28;51(15):2552-2570. [doi: 10.1017/s0033291721003731]
101. Digital media use and psychological functioning among adolescents and college students. Open Science Framework. 2022 Apr 2. URL: https://osf.io/jzwsg/?view_only=7b37a58439974d9d8d652bb7c188daeb [accessed 2024-02-20]

Abbreviations

CBT: cognitive behavioral therapy
DMHL: digital mental health literacy
MHL: mental health literacy
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RQ: research question
SMD: standardized mean difference

Edited by A Mavragani; submitted 26.07.23; peer-reviewed by L Roemer, M Anderson; comments to author 05.09.23; revised version received 25.10.23; accepted 25.12.23; published 29.02.24.

Please cite as:
Yeo G, Reich SM, Liaw NA, Chia EYM
The Effect of Digital Mental Health Literacy Interventions on Mental Health: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e51268
URL: https://www.jmir.org/2024/1/e51268
doi:10.2196/51268
PMID:38421687

©GeckHong Yeo, Stephanie M Reich, Nicole A Liaw, Elizabeth Yee Min Chia. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Smartphone Apps for Food Purchase Choices: Scoping Review of Designs, Opportunities, and Challenges

Remco Benthem de Grave[1], MSc; Christopher N Bull[1], PhD; Diogo Monjardino de Souza Monteiro[2], PhD; Eleni Margariti[1], MSc; Gareth McMurchy[1], MSc; Joseph William Hutchinson[1], MSc; Jan David Smeddinck[1,3,4], PhD

[1]Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, United Kingdom

[2]Centre for Rural Economy, School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

[3]Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria

[4]Ludwig Maximilian University, Munich, Germany

Corresponding Author:
Remco Benthem de Grave, MSc
Open Lab
School of Computing
Newcastle University
Open Lab, Floor 1
Urban Sciences Building, Newcastle Helix
Newcastle upon Tyne, NE4 5TG
United Kingdom
Phone: 44 1912084630
Email: r.benthemdegrave2@newcastle.ac.uk

Abstract

Background: Smartphone apps can aid consumers in making healthier and more sustainable food purchases. However, there is still a limited understanding of the different app design approaches and their impact on food purchase choices. An overview of existing food purchase choice apps and an understanding of common challenges can help speed up effective future developments.

Objective: We examined the academic literature on food purchase choice apps and provided an overview of the design characteristics, opportunities, and challenges for effective implementation. Thus, we contribute to an understanding of how technologies can effectively improve food purchase choice behavior and provide recommendations for future design efforts.

Methods: Following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, we considered peer-reviewed literature on food purchase choice apps within IEEE Xplore, PubMed, Scopus, and ScienceDirect. We inductively coded and summarized design characteristics. Opportunities and challenges were addressed from both quantitative and qualitative perspectives. From the quantitative perspective, we coded and summarized outcomes of comparative evaluation trials. From the qualitative perspective, we performed a qualitative content analysis of commonly discussed opportunities and challenges.

Results: We retrieved 55 articles, identified 46 unique apps, and grouped them into 5 distinct app types. Each app type supports a specific purchase choice stage and shares a common functional design. Most apps support the product selection stage (selection apps; 27/46, 59%), commonly by scanning the barcode and displaying a nutritional rating. In total, 73% (8/11) of the evaluation trials reported significant findings and indicated the potential of food purchase choice apps to support behavior change. However, relatively few evaluations covered the selection app type, and these studies showed mixed results. We found a common opportunity in apps contributing to learning (knowledge gain), whereas infrequent engagement presents a common challenge. The latter was associated with perceived burden of use, trust, and performance as well as with learning. In addition, there were technical challenges in establishing comprehensive product information databases or achieving performance accuracy with advanced identification methods such as image recognition.

Conclusions: Our findings suggest that designs of food purchase choice apps do not encourage repeated use or long-term adoption, compromising the effectiveness of behavior change through nudging. However, we found that smartphone apps can enhance learning, which plays an important role in behavior change. Compared with nudging as a mechanism for behavior change, this mechanism is less dependent on continued use. We argue that designs that optimize for learning within each interaction have a better chance of achieving behavior change. This review concludes with design recommendations, suggesting that food purchase

XSL•FO
RenderX

choice app designers anticipate the possibility of early abandonment as part of their design process and design apps that optimize the learning experience.

(J Med Internet Res 2024;26:e45904) doi:10.2196/45904

KEYWORDS

behavior change; mobile apps; food choices; grocery shopping; sustainability; healthy eating; digital health; mobile phone

Introduction

Background

Recent reports suggest that current consumer patterns, particularly in Western societies, have a significant impact on health and pose a considerable burden on the natural environment [1,2]. There is increasing evidence that the food production systems and processes enabling current diets contribute to biodiversity losses, water scarcity, and climate change [2,3]. In addition, obesity and food-related diseases affect a significant proportion of the global population [4], leading to an increasing economic and social burden at both the individual and aggregate levels [1,5,6].

Although many people claim to have values that align with healthy and environmentally sustainable diets [7], actual consumption patterns are often inconsistent with such stated preferences [8,9]. Research has found that people often struggle to select suitable products to support a healthy or sustainable diet [10]. Diet choices are affected by factors such as broader values, preferences, prices, availability, social pressure, and socioeconomic constraints [10,11]. This frequently leads to inconsistencies and compromises between competing priorities [12]. Moreover, bounded rationality (ie, the limited capacity for rational decision-making) and heuristic biases (ie, rational errors in the way we make quick decisions) often lead to suboptimal choices.

Digital technology—particularly smartphone technology—can support people in making suitable food purchases [13,14] that align with their values. Smartphone technology (hereafter referred to as *apps* but not excluding any smartphone technology that is not strictly considered an app) can provide rapid access to information, summarize large amounts of data, and present these in personally meaningful ways. With smartphones being omnipresent, these capabilities are available at nearly any time and place, alleviating the need for retailers to make financial investments that could hold back the implementation of alternative systems [15,16]. In the last decade, several software and hardware mobile technologies have emerged to support consumers in making better (ie, healthier and sustainable) diet choices. This is particularly the case for healthy diet support apps, which have received much interest in recent years, particularly in the form of diet-tracking technologies [13,17-19].

Although a large body of research has examined digital technology for consumption tracking of different products [17-20], research on smartphone apps that support food purchase choices is still limited [13,14]. Responding to Flaherty et al [13] and Chan et al [14], who call for further research on the development of food purchase choice apps and to support the development of effective food purchase choice app designs, we conducted a comprehensive review of the research on such developments to date.

Related Reviews

Our scoping review complements previous reviews. Mauch et al [21] reviewed popular apps for healthy food provision available on the Android and iOS app stores. They downloaded 51 apps, assessed the app quality (following the Mobile App Rating Scale [22]), and identified the behavior change techniques (BCTs) present in the design. They classified most apps as recipe, meal-planning, or shopping list apps. In total, 2 were classified as food choice apps. They found that the apps covered relatively few BCTs and that, although the apps generally scored well on functionality, they scored poorly on engagement. They recommended that future developments use a range of features to simplify healthy food shopping and maximize the use of BCTs.

Chan et al [14] conducted a systematic review to assess the efficacy of point-of-sale nutritional information interventions. They included 26 papers that reported comparative evaluations, 5 of which used digital technology as an intervention medium. In total, 3 of the 5 digital technology intervention studies showed a positive health impact, and the authors concluded that digital point-of-sale information interventions could improve healthy food purchasing. However, they noted that, compared with shelf labels, the requirement to scan products to retrieve information may have posed a barrier that led to inconsistency in the results. They found no relationship between intervention effectiveness and the number of identified BCTs.

Objectives

To the best of our knowledge, the reviews by Mauch et al [21] and Chan et al [14] are the only ones of food purchase choice apps to date. However, they are incomplete. Mauch et al [21] provided a catalog of publicly available health and diet apps without much appraisal of their use. Chan et al [14] surveyed evaluations of food purchase choice apps but limited the scope of their review to 5 studies that followed a comparative evaluation design. Thus, our review aimed to complement their work by compiling a comprehensive overview of the academic literature on apps for food purchase choice published to date. Specifically, we aimed to (1) describe common design characteristics of the retrieved apps and (2) describe opportunities and challenges to effective implementation as they are observed in the literature.

To address these objectives, we chose to conduct a scoping review. Similar to meta-analyses, scoping reviews are systematic, yet they are less restrictive in scope. They are suitable for mapping the knowledge on a topic and the

XSL•FO
RenderX

characteristics of the evidence [23], which makes them well suited for our aims.

This work provides an overview of existing food purchase choice app characteristics, unpacking how different design approaches might affect food purchase choices. Moreover, it frames common challenges in the implementation of these apps. Finally, it frames design considerations that could support future effective food purchase choice app developments.

Methods

Overview

We conducted a scoping review of published literature on smartphone-compatible digital technologies for food purchase choices (food purchase choice apps). The review followed the guidelines covered by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Multimedia Appendix 1 [23]).

Information Sources

We searched for peer-reviewed articles in 4 libraries that were selected to provide coverage of digital health and well-being technologies from an empirical as well as technical perspective: IEEE Xplore, PubMed, Scopus—which covers the ACM Digital Library—and ScienceDirect. The most recent database search was conducted on July 10, 2023. Together, these libraries cover a broad range of literature, with research on mature as well as emerging technologies and publications from the fields of behavioral science, economics, and computer science. We further scanned the included studies for references to articles that would meet our inclusion criteria.

Search

We searched the 4 libraries using the following keyword combination: "(mobile OR smartphone OR app) AND (purchase OR shopping) AND (food OR grocery OR supermarket)." We limited our search to articles published from 2008 onward. We chose this date limit in consideration of the introduction of the iPhone—the first widely adopted consumer smartphone with a noteworthy ecosystem of apps—in January 2007 and the time required for app development to pick up.

Selection

For selecting the articles, we applied a stepwise process that involved 3 reviewers (RBG, JH, and GM). RBG exported the search results from the individual database searches, imported them into the article selection tool Rayyan (Rayyan Systems Inc) [24], and removed duplicates. We then proceeded with the selection of articles according to the eligibility criteria as described in Textbox 1. First, we filtered articles based on title and—if the title left uncertainty about inclusion or exclusion—abstract. RBG scanned the full set of articles. JH and GM scanned approximately half of the set each. This way, each article was scanned independently by at least 2 authors. We then compared the results and discussed any differences until we reached a mutual agreement. RBG then read the full texts of all the articles that remained for another round of filtering. JH and GM together read 17 full-text articles so that a subset of full-text articles was read and selected by multiple researchers. Any differences in selection were discussed until a mutual agreement was reached.

XSL•FO
RenderX

Textbox 1. Eligibility criteria for article selection.

Inclusion criteria (all must apply)

- App design characteristics:
  - Designed for or compatible with smartphones (such as smartphone apps, web applications, or SMS text messaging)
  - Designed to support grocery choices by considering noncommercial product information types such as nutritional value and social or environmental impact
- Article characteristics:
  - Primary works in peer-reviewed journals

Exclusion criteria (any of these resulted in exclusion)

- App design characteristics:
  - Does not address diet choices
  - Does not have an apparent influence on the decision process (eg, if the technology focuses on automatic billing)
  - Is exclusively designed for web-based grocery purchases (most purchases are still made in brick-and-mortar grocery stores [25], and interventions for brick-and-mortar and web stores involve rather different interaction challenges; designs for purchases in brick-and-mortar stores need to address transitions between the physical and digital spaces [26,27], whereas designs for web-based purchases do not need to address this transition)
  - Recipe recommenders (that recommend recipes rather than products)
  - Diet-tracking apps (unless specifically tracking purchases rather than consumption)
  - Designed specifically to provide support regarding physical disabilities (eg, mobility and eyesight disabilities)
  - Designed around exclusive programs (eg, exclusive voucher programs available only in restricted areas)
  - Technology purely designed and intended as a study support tool
- Article characteristics:
  - Literature reviews (in case relevant reviews were retrieved through the literature search, they were covered as related work in the Introduction section of this paper)
  - Not in English

Data Charting

The data charting process (also referred to as data extraction) was iterative, and the coding of several items was inductive.

Step 1: Open Coding

Open coding was performed by following an instruction manual and the codes were inserted into a Microsoft Excel (Microsoft Corp) spreadsheet that served as the charting form. The instruction manual was drafted by RBG and subsequently verified for comprehensiveness and clarity of instruction through duplicate coding of a test set of 7 articles by RBG, JH, and GM. For verification, the 3 coders met to compare and discuss differences in coding, and minor adjustments to the instruction manual were made. The 3 coders then continued to code the remaining articles. JH and GM each coded another 5 articles each, and RBG coded the remaining articles. Open coding at this stage aimed to cover relevant information at a high level of detail such that this could facilitate the inductive process. Variables that were coded at this stage were app variables (description of the app, including the user interaction, functions, required technical infrastructure, and considered product information), evaluation variables (study design, participants, findings, and reported challenges and opportunities), and article variables (publisher and publication year). In case the same app was discussed in multiple articles, the information for the app variables was combined.

Step 2: Collaborative Identification of Patterns

After completion of open coding, the coders held a workshop meeting to identify patterns in several of the data items from step 1 (description of the app, functions, findings, and reported challenges and opportunities) to determine app types, a harmonized collection of functions, and topics of opportunities and challenges. Miro (RealtimeBoard, Inc) [28], a digital whiteboard for collaborative note taking, was used for this workshop. Post-It notes with the initial codes for several of the data items (excluding the article items) were added to the whiteboard. A complete set of these notes was copied for each of the 3 coders, which they could reorganize to aid in the process of exploring and identifying patterns. Workshop participants also had access to a Microsoft PowerPoint (Microsoft Corp) slide deck with images and descriptions of each app and the spreadsheet with the complete coding from step 1. Pattern exploration was performed in 6 rounds. In each round, a subset of step 1 items was explored, first independently by each of the participants and then collectively.

XSL•FO
RenderX

Step 3: Final Coding

RBG updated the charting form with the outcomes of the workshop. Several data items were replaced such that the new data items aligned with the patterns from the workshop. In addition, the full texts of the articles were scanned for references to the patterns of opportunities and challenges that were identified in the workshop (also known as *indirect* opportunities and challenges).

Data Items

The final set of charted data items is outlined in the following sections.

Article Characteristics

We charted publication year and publisher. Classification of study types was coded according to the classification adapted from the study by Bardus et al [29]. We classified article types as (1) *design and development* (articles that described the design of an app), (2) *feasibility* (articles reporting procedural outcomes such as use, acceptance, and retention, as well as articles reporting outcomes of experiments that were performed under restrictive conditions [eg, laboratory setting or a predetermined selection of products to choose from]), (3) *evaluation* (reporting the effectiveness of a technology intervention trial on food purchase choice), and (4) process evaluations and causal-comparative studies (*process evaluations* in short; reporting effects on sociocognitive factors that are related to food purchase choice or reporting on comparisons with alternative interventions). A fifth type was reserved for *app store review*s. Combination types were also possible (eg, an article could cover both a feasibility and an evaluation study and would then be assigned both codes).

Design Characteristics

We recorded the *app name*. The variables *app type* and *functions* were charted in correspondence with the patterns that evolved from the workshop (see the *Data Charting* section, step 2). *Product information* represents the type of information that the app uses to support product choice. This covers nutritional information (eg, calories, allergens, and macro- and micronutrients), diet balance (or food group; information on food groups of purchases concerning recommendations for a balanced diet, such as MyPlate [30] or the Eatwell Guide [31]), environmental impact (eg, carbon footprint and food miles), and societal impact. *Retailer dependency* was classified as "dependent" in case the article mentioned the use of data infrastructure that is controlled by the retailer (eg, the retailer product database, loyalty card data, or beacons for indoor navigation), "independent" in case the article stated that only crowdsourced or open-access data were used, or "unknown" when this information was unavailable.

Study Characteristics (for Evaluation Studies Only)

We recorded the *study design* (ie, cohort or control group design), evaluation period (*duration*), number of participants in each study group (*participants*), and primary findings (*findings*). The topics of *opportunities and challenges* were coded in correspondence to the patterns from the workshop. Here, we distinguished between *direct* (reported as a challenge or opportunity in the article) and *indirect* (not reported as a challenge to the design and identified in the second pass of the article).

Synthesis of Results

We counted the frequencies at which the values were observed and provided narrative summaries of the findings. Specifically, to synthesize the findings on apps, we grouped observations (ie, apps) by app type and summarized the design characteristics for each group. We supplemented this with a narrative summary in which we focused on common patterns and salient exceptions to those patterns. Furthermore, we visualized the frequency of functions per app type using a bar chart.

Next, we investigated the opportunities and challenges from a quantitative and qualitative perspective. From the quantitative perspective, we counted significant and insignificant primary evaluation outcomes and provided a narrative summary of dominant patterns in the results. From the qualitative perspective, we provided a content analysis of the charted opportunities and challenges, meaning that we tabulated an overview of counted topics of opportunities and challenges and elaborated on these topics using a narrative summary that was supported by specific observations in the articles.

Results

Overview of the Included Articles

The database searches returned a total of 1353 articles. Another 11 articles were included from references. After screening, we retained 4.03% (55/1364) of the articles for this review. Figure 1 provides an overview of the number of articles excluded at each step of the selection process and the reasons.

XSL•FO
RenderX

Figure 1. Results of article search and screening. PI: product information.

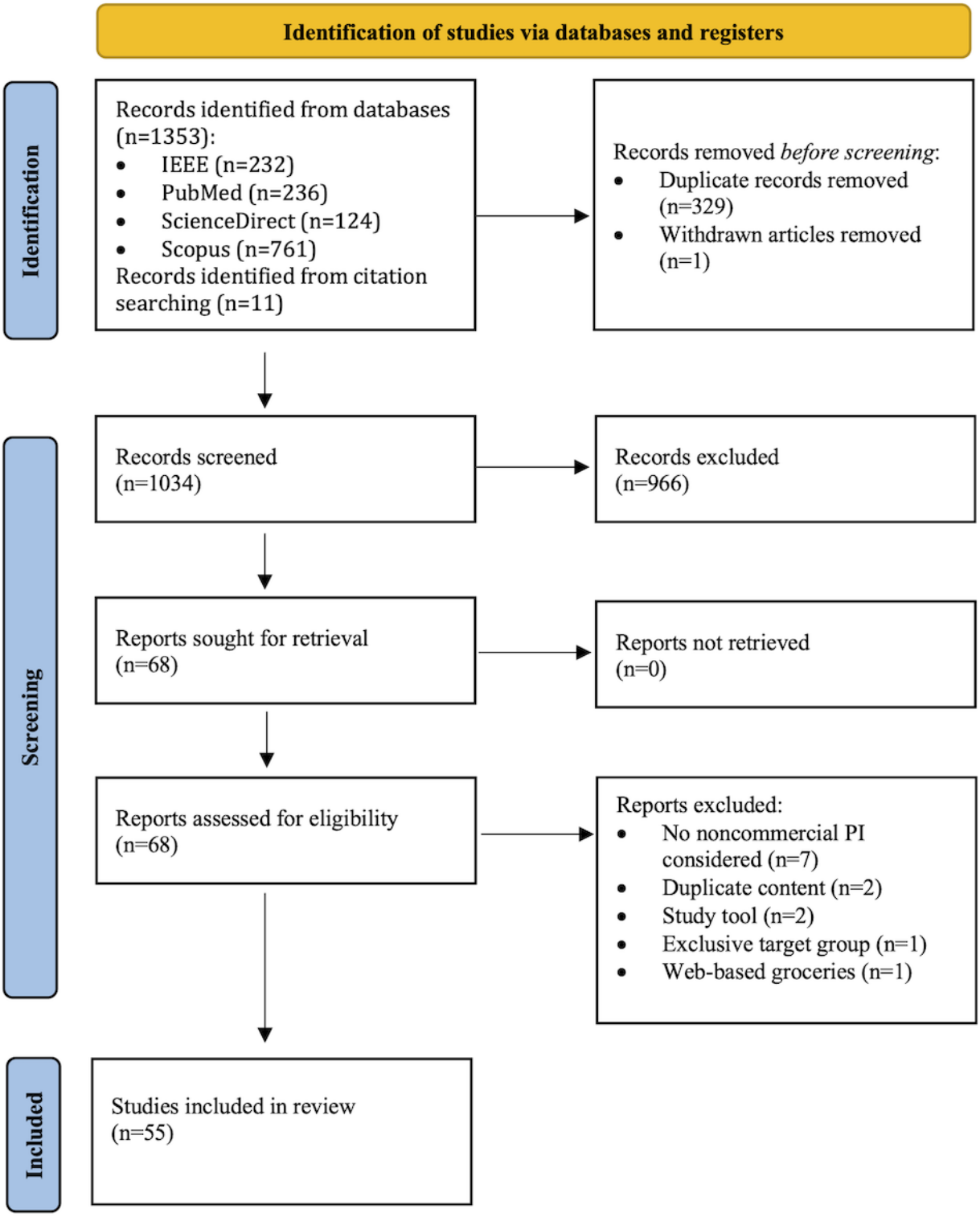

Among the 55 articles that remained after filtering, we identified 41 (75%) design and development studies, 32 (58%) feasibility studies, 13 (24%) evaluations, and 3 (5%) process evaluations. In total, 4% (2/55) of the articles were classified as app store reviews. These counts do not add up to 55 as articles often covered a combination of classifications. The dominant publication venues were *IEEE* (engineering; 13/55, 24%), *ACM* (computer science; 13/55, 24%), *Elsevier* (7/55, 13%), *JMIR* (5/55, 9%), and *Springer* (5/55, 9%). A complete overview of the included publications is available in Multimedia Appendix 2 [9,12,13,16,21,32-81].

Summary of Apps

Overview

We charted the characteristics of 46 unique food purchase choice apps (see Multimedia Appendix 3 [9,12,16,33-81] for a complete overview). A total of 5% (3/55) of the articles [13,21,32] provided insufficient information on their covered apps for us

XSL•FO
RenderX

to chart them. We identified 9 main functions. For 3 of these, we distinguished different implementations of the functions, creating a total of 21 different function labels (see Multimedia Appendix 4 [9,27,33-36,40,41,44,48,49,52,71,82] for descriptions). Most apps provided health-relevant information (37/46, 80%), with 70% (32/46) of the apps returning nutritional content and 13% (6/46) of the apps providing information on diet balance (1/46, 2% provided information on both). A total of 26% (12/46) of the apps provided information on environmental impact. Other product qualities included societal impact (3/46, 7%), product freshness (1/46, 2%), and product authenticity (whether the provided origin and product type were genuine; 1/46, 2%). Approximately one-quarter (12/46, 26%) of the apps were found to be dependent on the retailer.

We categorized the apps into 6 app types. We observed that all—except one [33]—of the covered apps (45/46, 98%) provided support for a distinct moment or stage regarding product choice. Therefore, we named each app type according to the stage that it addressed. We found apps that addressed (in chronological order) structured purchase *planning* (7/46, 15%), *contemplation* (2/46, 4%), *approaching the product* (6/46, 13%), physically *selecting the product* (27/46, 59%), and *reflecting* on the purchases (7/46, 15%). A total of 2% (1/46) of the apps addressed a combination of these stages (the planning, contemplation, selection, and reflection types. For calculations and graphical mappings, we did not count this toward the individual app types. In those cases, we included a separate multistage app type). In the following sections, we describe these app types along with common functions (Figure 2; an overview of function descriptions is provided in Multimedia Appendix 4) and examples.

Figure 2. Frequency of functions by app type. Separate panels for each app type are scaled to the total number of apps within the app type. NFC: near-field communication; RFID: radiofrequency identification.

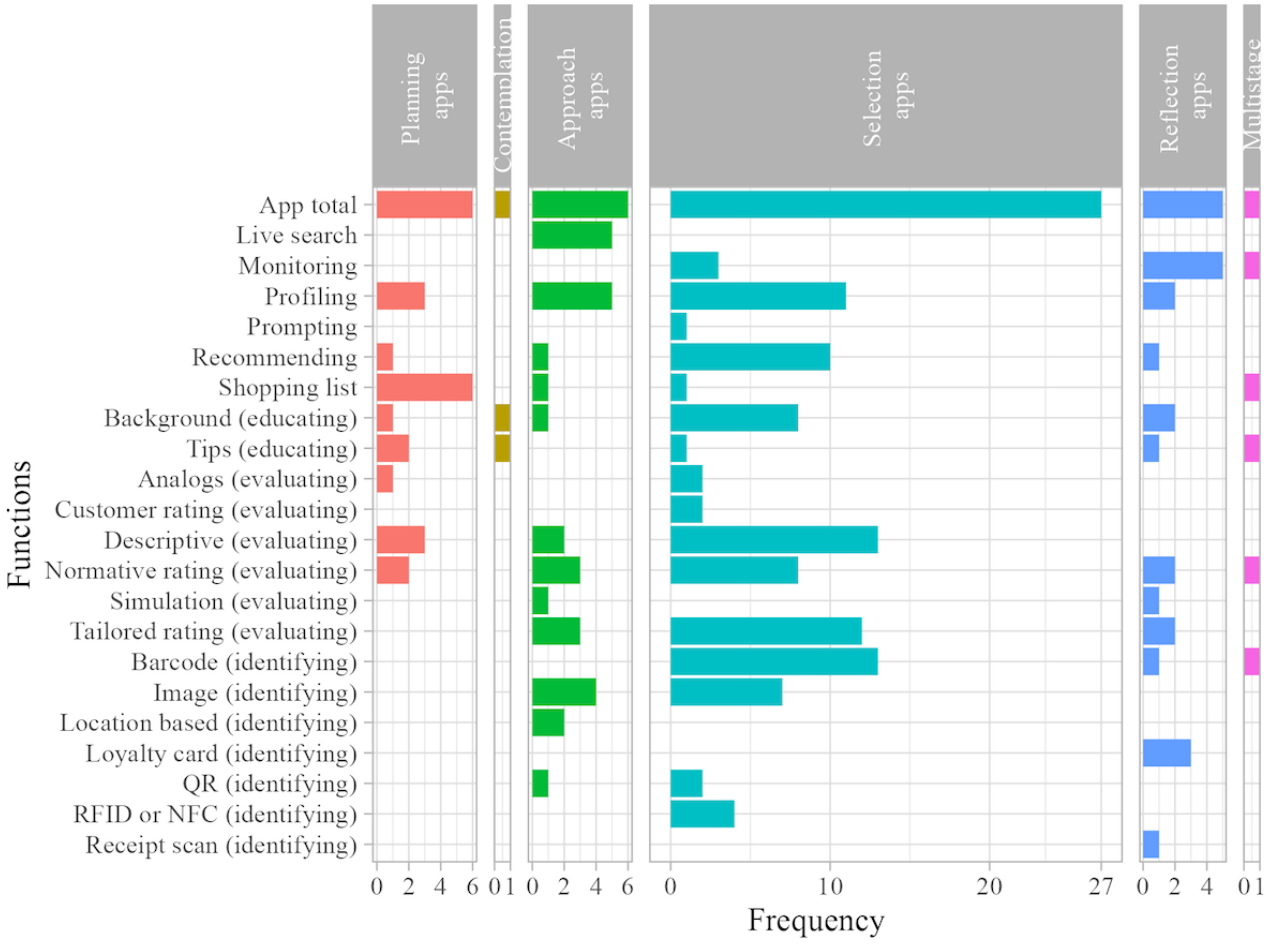

Purchase Planning Apps

These apps support the planning for the use of a shopping list. This can be done by automatically generating a shopping list [34], suggesting a goal for the shopping list content [33,35], or providing feedback about product characteristics within the shopping list [36,37].

For example, Hedin et al [36] used an autocomplete algorithm that lists food items from a carbon footprint database while the user types. The carbon footprint value per 100 g is then presented next to the list item.

Contemplation Apps

The contemplation apps have a mostly practical educational function and provide knowledge and tips that help the user make decisions about food purchases while shopping.

For example, Bangia et al [38] introduced a podcast series that informs the listener about the benefits of omega-3, the types of

XSL•FO
RenderX

products in which it is found, and practical considerations such as the affordability of different forms of omega-3. The podcast series has short episodes, one each covering products that are typically found in the same shopping aisle so that it can be played when visiting the specific aisle.

Approach Apps

Product approach apps use the live search function to provide information about nearby products. Options are visualized on the device screen together with suitability ratings for each product. These apps were found to provide nutritional and environmental but not diet balance information.

For example, the app described by Isley et al [39] uses augmented reality to project semitransparent traffic-light color frames and A-to-F tailored ratings over breakfast cereal boxes as the user views the respective shelf section through the phone camera.

Special cases were 4% (2/46) of the apps, which used indoor position data to provide the user with information about nearby products but relied on the retailer for these data.

Selection Apps

Selection apps provide the user with an evaluation of the product after it has been identified. *Barcode scanning* (13/46, 28%) is the most common type of identification. *Normative ratings* (8/46, 17%) and *tailored ratings* (12/46, 26%) are the most common types of evaluations. Many of these apps also recommend alternative products from the same product category (10/46, 22%; eg, the category of milk and dairy-free milk alternatives).

For example, upon barcode scanning, App 30 (FoodSwitch) [40] shows the product name and traffic-light nutrition label for the identified product and lists same-category alternatives ordered from more to less healthy below the identified product.

The following were special cases. Although selection apps rely on the active pursuit of product information, one of the identified selection apps makes an exception and allows the user to be passive in the information pursuit. Customers use the retailer-dependent App 22 (Dirk app) [41] to scan articles as they go around the store as part of a checkout-free automated payment process. In case the user scans an unhealthy item, the app provides a pop-up with information and an alternative.

In addition, one selection app, App 41 (Nutriflect Mobile) [42], is designed to be used in combination with a reflection app, App 40 (Nutriflect Home), that records the diet balance in food purchases. This is the only selection app that provides feedback on diet balance (see also the visual in Multimedia Appendix 5).

Reflection Apps

These apps rely on monitoring purchases to provide the user with feedback on an aggregated data visualization. Some of these apps are intended to be used in the store [16,33,35], tracking items as they are added to the shopping basket and providing the user with the opportunity to make adjustments before leaving the store. Other apps provide feedback based on a record of purchases (eg, from loyalty card data) and support intention formation for the next purchase.

For example, Bird et al [16] let the user scan the barcodes of products as they were added to the shopping basket to receive diet balance information. The app interface provided visual feedback on the content of the basket according to 5 food groups (grains, meat, dairy, fruits and vegetables, and treats) through a pie chart. The size of each slice of the pie chart was proportional to the recommended number of purchases for that food group, and each slice could be partly filled using, for example, light green to represent the remaining amount to be added to the basket and dark green to represent the amount already added.

Multistage Apps

One app was found to address multiple stages of the purchase process. Bomfim and Wallace [33] and Bomfim et al [35] proposed an app, App 44 (PBGA), that covers the planning, contemplation, selection, and reflection stages to support users in achieving a balanced diet. App 44 (PBGA) supports users in creating a shopping list in line with a healthy diet balance (planning), provides users with tips for identifying healthy options in various aisles (contemplation), provides nutrient information upon scanning the barcode of a product (selection), and monitors the scanned items to recommend purchases in each of 5 food groups.

Challenges and Opportunities

In the following sections, we explore the opportunities and challenges associated with the effective implementation of food purchase choice apps. For this purpose, we summarize the results of the evaluation studies and report a qualitative content analysis of common challenges.

Study Outcomes

Table 1 provides an overview of the 11 evaluation studies. These evaluation studies cover 22% (10/46) of the apps: 7% (2/27) of the selection apps (lowest) and all the reflection apps (7/7, 100%; highest). In total, 72% (8/11) of the studies reported a significant behavioral impact. However, mixed results were observed for the 2 evaluated selection apps. Otherwise, no clear patterns between app type and study outcome could be observed. A total of 27% (3/11) of the studies reported outcomes (findings) toward a diet balance, all showing significant results (see also Multimedia Appendix 6 [16,38,41-49] for grouping of evaluations by product information).

Table 2 provides an overview of the 6 process evaluations and causal-comparative studies. Knowledge gain was the most commonly reported sociocognitive outcome factor (4/6, 67%). All these studies reported significant knowledge gain.

XSL•FO
RenderX

Table 1. Summary of evaluation studies organized by app type.

App type and name	Design	Duration	Participants	Findings
Planning apps				
App 39 (MyNutriCart) [43]	Cohort	8 wk	27 participants	Significant healthier diet balance of purchases
Contemplation apps				
App 5 [38]	Cohort	6 mo	251 participants	Significant increase in omega-3–rich purchases
Approach apps				
App 8 [44]	Cohort	7 mo	69 participants in the intervention group and 323 participants in the control group	Significant increase in higher-rated purchases
Selection apps				
App 22 (Dirk app) [41]	3 UI[a] variations (1 × presenting the alternative without information about its healthiness and 2 × with variations of information that indicate that the presented item concerns a healthier alternative) vs control	5 wk	1783 scans[b]	Significantly healthier purchases when presenting alternatives without health information and nonsignificantly healthier when adding health information
App 30 (FoodSwitch) [45]	App vs no app	4 wk	33 participants in the intervention group and 33 participants in the control group	Significant reduction in sodium concentration in purchases
App 30 (FoodSwitch) [46]	App+reduced-sodium salt vs no app	12 wk	84 participants in the intervention group and 84 participants in the control group	Nonsignificant change in sodium intake
Reflection apps				
App 13 [47]	Cohort	2 wk	31 participants	Nonsignificant change in the healthiness of purchases
App 25 (EcoPanel) [48]	App vs no app	5 mo	65 participants in the intervention group and 2587 participants in the control group	Significant increase in organic purchases
App 32 (FutureMe) [49]	App+future-self simulation vs app without	12 wk	42 participants in the intervention group and 53 participants in the control group	Nonsignificant difference in the healthiness of purchases
App 35 (Healthy Shopping App) [16]	Cohort	4 wk	7 participants	Significantly healthier diet balance of purchases
App 40 (Nutriflect Home) [42]	Cohort	4 wk	21 participants	Significantly healthier diet balance of purchases

[a]UI: user interface.

[b]Participant counts are unknown. Only the number of scans that triggered information prompts was available.

XSL•FO
RenderX

Table 2. Summary of process evaluations and causal-comparative studies.

App type and name	Design	Duration	Participants	Findings
Planning apps				
App 33 (GreenCobra) [36]	Cohort	2 wk	30 participants	Significant knowledge gain
App 37 (MFG[a]) [35]	Cohort	3 wk	12 participants int eh control group	Significant knowledge gain
App 39 (MyNutriCart) [43]	App vs education	8 wk	27 participants in the intervention group and 24 participants in the control group	Nonsignificant difference in healthiness of purchases vs education
Selection apps				
App 26 (EDO app) [50]	Cohort	12 wk	143 participants	Significant knowledge gain and significant gain in self-efficacy
Multistage apps				
App 44 (PBGA[b]) [35]	Cohort[c]	3 wk	12 participants in the intervention group and 12 participants in the control group	Significant knowledge gain
App 44 (PBGA) [35]	App 44 (PBGA) vs App 37 (MFG)	3 wk	12 participants in the intervention group and 12 participants in the control group	Significant reduction in ultra-processed food purchases

[a]MFG: My Food Guide.

[b]PBGA: Pirate Bridgitte's Grocery Adventure.

[c]Knowledge gain was assessed using a pre- and posttest design, whereas behavior change was assessed by comparing the effects of different apps. Therefore, we list the results in separate table rows even though the 2 evaluations were performed in the same study.

Qualitative Content Analysis of Challenges and Opportunities

Overview

Table 3 provides the counts of common topics regarding challenges and opportunities as they were identified through a qualitative content analysis (Multimedia Appendix 7 [12,13,16,21,32,35-42,44-60,62-67,69,76,79,81]). We elaborate on these topics in the following sections.

Table 3. Common challenge topics found in the studies.

Topic	Direct[a]	Indirect[b]	Total
Behavior change	8	10	18
Barriers to change	5	4	9
Interaction	23	20	43
Engagement	12	2	14
Trust	5	7	12
Burden of use	5	4	9
Learning	1	7	8
Technical feasibility	13	7	20
Source data[c]	8	3	11
Performance	5	4	9

[a]The authors of the publication recorded the topic as a challenge to the implementation of the app.

[b]The authors addressed the topic but not directly as a challenge to the implementation of the app.

[c]The data that underlie the information that is viewed on the app. This is typically a database of product information.

XSL•FO
RenderX

Behavior Change Challenges

Barriers to Change

Behavior change is hard, and participants struggle to achieve diet goals [16,35]. Conflicting priorities complicate adoption of changes [12,36,51]. The user may not like the suggested products [51] or perceive the pursuit of adopting changes as too time-consuming [12,36]. Participants look for ways to prioritize their efforts but may find support for this lacking [36]. Zapico et al [48] argued that reflection on previous behavior helps people prioritize behavior change, and they developed a reflection app [48,52]. Apps can also support prioritization by providing salient reminders of past behavior as memories of past behavior may not be accurate [42]. Reitberger et al [42] supported this by linking a selection app with behavior-monitoring data from a reflection app. This allowed the selection app to assess the suitability of products from the perspective of past behavior.

Scope of Impact

Although a benefit of mobile apps lies in their scalability [13], the impact is limited to the users who choose to adopt the technology [9,40]. This is often only a small proportion of the population [40]. However, Reitberger et al [42] noted that this impact can be extended to household members by monitoring household purchases using a reflection app. Reflection on household behavior can spark conversation and lead to social facilitation.

Retailer dependency can also restrict the scope of impact. Information about purchases is available only at specific stores [52]. The app may have little value for people who purchase much of their food from other retailers [49]. The commercial interest of retailers may also lead them to restrict the information that is provided. For example, retailers may be interested in providing data that encourage the purchase of organic products but may be less inclined to provide data on other indicators of sustainability [52]. Retailers may also choose to withhold negative information or withhold information altogether as negative information was found to have a stronger impact on rejecting a product than that of positive information on choosing a product [53].

Interaction Challenges

Engagement

Engagement issues (limited use and abandonment) were frequently reported [12,13,21,32,45,46,49,51,54-57]. Engagement issues are seen as a barrier to behavioral impact [21,32,46,49,55], and 4% (2/55) of the studies attributed insignificant findings to this issue [46,49]. However, Zapico et al [48] did not observe a correlation between frequency of engagement and behavioral impact. They reported significant behavioral impact despite a low engagement frequency with their App 25 (EcoPanel) reflection app. Engagement issues may result from knowledge gain (learning) [50,58], a perceived burden of app use [46,57,58], performance [13,49], and trust issues [49], whereas tailored or personalized feedback can increase engagement [49,55]—although personalized feedback can negatively affect engagement when the data are not trusted [49]. Normark and Tholander [59] found that, when technology does not perform as expected, this leads to frustration and rejection of the technology.

Burden of Use

The use of the technology may be found effortful and disruptive to the shopping routine [42,44,60] or shopping list creation [12] and, therefore, less acceptable for frequent use [58,59]. Instead, studies indicated that technology is used as an infrequent lookup and to explore new products [12,44,58], which may not have been the intended use [12]. However, one study [58] was found to consider the likelihood of infrequent use as part of their design process.

Learning

Learning (knowledge gain) was framed as a challenge from the viewpoint of some authors as they perceived that it might hinder repeated use of the app and contribute toward its abandonment. Indeed, as people gain knowledge about products through the use of the app, they may perceive less need to use it again [50,58]. However, knowledge gain (ie, increasing food literacy) through digital interaction was frequently highlighted as an important mediator of behavior change [33,35,36,38,50]. This means that learning was viewed as a challenge for food purchase choice app adoption but not necessarily for achieving behavior change. Moreover, knowledge gain was generalized beyond the products that were evaluated using the app [58] and even the food group [50]. Some authors suggested that, because of rapid learning, initial app interactions may be the most important for guiding change [48,58], and in some cases, further engagements may have little incremental impact on further behavior change [48]. In addition, Bangia et al [38] observed a significant behavioral impact from a one-off intervention with their app 6 months after the intervention had taken place, pointing out that repetition is not an essential mechanism to achieve knowledge gain and mediate behavioral impact.

Trust

Issues with trust were also flagged as negatively affecting engagement with technology. Involvement of the retailer in the app source data was noted as a potential source of distrust [12,16,58], leading to efforts to develop technology that can work independently from the retailer [16]. Commercial interests may restrict transparency [52], involve advertisements in the form of recommendations to influence purchases [61], or restrict product information unfavorably [53]. Even when the motives of influence are genuine, external influence may result in reactance (the resistance against a persuasive intent that is nonetheless perceived as manipulation). van der Laan and Orcholska [41] showed how a persuasion attempt to choose a healthier alternative can be nullified when external motives are exposed.

Commercial interests may also make demands on privacy protection, particularly when personal data are shared [44]. As many of the apps are designed to be used in nonprivate environments, privacy concerns are also relevant for the mode of feedback, particularly when the feedback is not well concealed (eg, when using audio feedback [60]) or when the phone camera is used for recognition of products and personal identifiable features may be exposed [36].

XSL•FO
RenderX

Technical Feasibility Challenges

Source Data

Acquiring and maintaining (accurate) source data may also present feasibility issues, for example, when using location information for recognizing products [44,62,63]. Acquiring and updating source data to achieve the required performance was found to be very challenging [63,64]. In addition, some data, such as data on environmental sustainability and social impact, may be difficult to acquire [36,44,58,65]. This can pose significant implementation challenges.

Performance

Technical challenges may stand in the way of the practical implementation of an app. For example, performance issues may limit the app regarding approaches using image recognition. Studies that reported on recognition performance found that not all products could be identified [66,67], and no evaluation studies were found for apps that use image recognition to identify products. Indeed, Isley et al [39] noted that water bottles could not be detected by a camera, and QR labels on shelves were needed to assist their study. They also reported that cereal boxes were not always accurately detected or were not detected at all.

Field trials using near-field communication (NFC) or radiofrequency identification (RFID) as a detection method were not found. Although no performance issues were reported, it was noted that NFC or RFID labels were not (yet) provided for the grocery products in the study supermarket [42], which made practical implementation of this strategy unfeasible at the time of the study.

Even when a product is correctly scanned, product data may not be returned. We found various studies that noted issues with missing or faulty data for products, particularly in the case of retailer-independent designs [46,47,54,55], although faulty data could also be observed for retailer-dependent designs [12].

Discussion

Principal Findings

We systematically reviewed the academic literature on food purchase choice apps and inductively charted common design characteristics as well as opportunities and challenges for their effective implementation. We identified 5 distinct app types, each supporting a specific stage of food purchase choice. The most common app type (27/46, 59%) was selection apps. Most selection apps identify products through barcode scanning and evaluate the product by returning a (tailored or norm-based) rating of nutritional content without any dependence on a specific retailer. Other types of apps and product information were less frequent.

We investigated opportunities and challenges by charting quantitative results from evaluation trials and qualitatively exploring patterns in the authors' observations. Similarly to Chan et al [14], we observed that the results from 11 comparative evaluation trials indicate the potential for food purchase choice apps to support behavior change. However, the proportion of evaluated selection apps is notably small (2/26, 8% vs 9/20, 45% for other types); moreover, the results for this app type are mixed, suggesting potential efficacy challenges with this approach [83,84]. Notably, all comparative evaluations that provided diet balance information (3/11, 27%) showed a positive impact. Evaluation trials also indicated the potential of food purchase choice apps to contribute to knowledge gain, with all 4 evaluations reporting significant findings.

Qualitative findings based on content analysis of the authors' observations highlighted opportunities and challenges surrounding food purchase choice apps. In food purchase choice apps, engagement issues such as repeated use and long-term adoption present a common challenge. The findings indicate that engagement challenges frequently result from a perceived burden of use. These results are in line with those of the research on the abandonment of behavior change technology [85,86]. We also observed broader challenges relevant to the pursuit of behavior change (eg, people may struggle to identify which efforts to prioritize), trust in retailer motives, and technical feasibility. Learning was observed as a common opportunity derived from the use of food purchase choice apps. Opportunities for addressing various challenges were repeatedly found in apps that apply behavior monitoring (reflection apps).

Implications

Consideration of Mechanism of Technology-Mediated Change in Design Thinking

Selection apps address *the moment* of product selection and appeared particularly popular for supporting food purchase choices. Selection apps provide support just-in-time (at "exactly the right moment") [55] (see the "just-in-time" intervention literature [87,88]; ie, the moment a choice is translated into action in-store and a product is either added or not added to the shopping basket). There is no need to rely on—potentially inaccurate—memory [42]. Rather than relying on rational thinking, this approach allows the user to be guided by a gentle push (nudge) in the direction of healthy or sustainable choices on the spot [41,55,68] (see the studies by Caraban et al [89], Thaler and Sunstein [90], and Sunstein [91] for frameworks of nudging). As such, the approach aligns with a growing acknowledgment that most decisions during in-store grocery shopping are not made through deliberate, rational decision-making processes but rather through heuristics [69] ("mental shortcuts," or "snap decisions" based on less conscious evaluations of the available information [90,92]). It can be argued that conscious decision-making is not the design aim of a nudge-driven system, but it can be a side outcome. Along these lines, behavior change beyond a single purchase relies on the repeated exposure to these nudges and on the repeated use of the food purchase choice app that uses nudges to achieve impact.

Despite their claimed benefits, we observed frequent reporting of engagement challenges with selection apps and food purchase choice apps in general. Users perceive the use of selection apps as interruptive or burdensome [46,58,93], leading them to slowly abandon their use over time, which jeopardizes the intended behavior change goal of such apps. These findings align with those of Vhaduri and Prioleau [94], highlighting that infrequent

XSL•FO
RenderX

engagement with digital health technology jeopardizes the potential health benefits. Similarly, Ni Mhurchu et al [93] found that the healthiness of food purchases only increased for people who frequently used the selection apps.

Various researchers have attempted to characterize the causes that underlie these engagement issues with digital health applications [85,86,94,95] and proposed design modifications that may improve long-term engagement. Kalnikaité et al [27,92] and Todd et al [26] highlighted the importance of causing minimal disruption to the shopping routine and providing "just enough information, in the right form." A variety of product identification and evaluation approaches may be driven by the premise of minimal disruption in the selection of apps that we found through our literature search. However, these approaches sustain the focus on nudge-based systems. The questions remain open with regard to improving nudge-based systems enough to accomplish long-term use to achieve behavioral impact.

Related to this, we observed 2 important shortcomings in the attempt to improve behavior change impact by designing food purchase choice apps with the premise of repeated use. First, we found that engagement challenges are common across various domains of digital technology (eg, for just-in-time digital health interventions [96,97], web-based diet interventions [98], self-tracking apps [85,95], wearables [99], and Internet of Things [100]). The widespread presence of engagement issues suggests that they are hard to solve, and the results of efforts to achieve engagement for extended periods are likely to be modest at best. Short periods of engagement may be better understood as the rule rather than the exception [85,95].

Second, the proposal to improve nudge-based systems for improved engagement reflects the prioritization of quantity over quality of engagement. A focus on engagement quantity aligns with a nudge-mediated mechanism of behavior change (ie, the app redirects automatic behavior by changing the choice environment without significant involvement of conscious decision-making [90,92]). However, the findings of our study challenge this prioritization of quantity over quality and suggest that other important mechanisms are at play.

Our findings suggest that learning (knowledge gain) plays an important mediating role in the impact of food purchase choice apps on food purchase choice behavior. In their study with a reflection app, Flaherty et al [69] found that behavioral feedback can spark learning through critical reflection on products and one's beliefs about products and influence future purchases. Reitberger et al [42] and Zapico et al [48] observed that the use of a reflection app contributed to a positive change in food purchase choice. Also indicative of learning, Bangia et al [38] observed a positive change in food purchase choice 6 months after using an information-rich food purchase choice app during 1 shopping trip.

The mechanism of learning-mediated changes appears to depend more on the quality than on the quantity of food purchase choice app engagements. Few interactions may suffice for learning to happen [38,48,58]. A single interaction may even suffice for a significant impact on behavior [38,48]. In their research with selection apps, Samoggia and Riedel [50] and Herbig et al [58] noted that learning can even reduce the likelihood of future use as the perceived value of the information diminishes, findings that align with those of Epstein et al [95], who found that internalization of the learnings from the use of diet trackers can be a cause of abandonment of the technology as such a mechanism of learning-mediated change can intervene with a mechanism of nudge-mediated change.

Although the described mechanism of learning involves conscious reflection, it does not conflict with an appreciation for the role of heuristics in grocery shopping. Flaherty et al [69] observed that conscious reflection serves to update beliefs that drive heuristics-based food purchase choice. After a period of conscious reflection on purchases, routinized, heuristics-based shopping is continued.

In summary, we observed that there is an important role for learning in behavior change supported by food purchase choice apps and that—contrary to a mechanism that relies on nudging—learning is relatively robust against disengagement as it does not necessarily rely on continued engagement with the app. This has important implications for the design of food purchase choice apps. There are opportunities to design apps that can be effective despite potential abandonment that designers may want to pursue. In the next section, we provide some practical suggestions for designing with learning in mind, along with other suggestions for the future design of food purchase choice apps.

Suggestions for Future Design

The purpose of this section is to formulate briefly some practical suggestions for the design of food purchase choice apps that follow from the contents of this review.

Optimize for Learning

As discussed in the previous section, learning appears to play a key role in the potential behavioral impact of food purchase choice apps. We propose that the key elements of learning-oriented food purchase choice app design include information clarity, information contextualization (ie, providing informative context to support decisions), and an emphasis on rich information—elements that we observed in the use of functions such as *tips and tricks* and *background information*. Key elements also include design with a view for minimal interactions as opposed to repeated use, design to support reflection on information and rational thinking that extends beyond the app's use, and design to support user autonomy and gradual abandonment of the app.

This proposal aligns with our observations that behavior monitoring can benefit learning [47,48,69] and that the use of knowledge tips and background information is also associated with knowledge gain [38,58]. Importantly, the optimization of designs for a nudging effect [26,27] may conflict with efforts to optimize designs for learning.

Formalize Assumptions of App Use Over Time

Many food purchase choice app designs seem to be developed on the unspoken premise of continued engagement to achieve behavioral impact, although this is not a given. We observed that it is uncommon for designers to articulate expectations regarding the regularity and nature of engagement within the

XSL•FO
RenderX

design process (an exception is the study by Herbig et al [58]). However, our findings suggest that early abandonment appears to be common and can have a detrimental impact on the potential of an app to affect behavior. To mitigate unforeseen impact, we suggest that designers view abandonment as a key element of their design strategy and formulate engagement expectations early in the design process. It is recommended that designers include considerations about engagement over time and possible abandonment in design thinking. By bringing transparency into the matter, designers can then develop diverse mechanisms to address abandonment.

Overcome Biases and Norms When Selecting Design Characteristics

Although we observed a wide range of design characteristics (app types, functions, and product information), food purchase choice app designers seem biased toward selecting a set of characteristics (see, for instance, selection apps and the use of barcode scanning and normative or tailored ratings of nutritional information). We found that other app types and product information may have more potential. Therefore, we encourage designers to explore other options for the design of selection apps and, broadly, of food purchase choice apps. Some suggestions include the use of diet balance as product information and reflection apps as an app type, which are further addressed in the following sections.

Use Diet Balance as Product Information

This product information was commonly associated with significant behavioral impact (see the *Study Outcomes* section). Achieving a balanced diet can be framed as goal setting, which is considered an effective BCT [101-103]. In addition, the simplicity of classifying choices into 5 categories may be beneficial for facilitating behavior change. Goal setting requires a form of monitoring and is, therefore, suitable for reflection apps—for which monitoring is a key function. Goal setting is also compatible with planning apps [35] based on purchase intention monitoring. However, the need to define quantities for the intended purchases may present a barrier [12,36,70], which may make this a less suitable option.

Consider Behavior Monitoring (Reflection Apps)

We observed that various opportunities for addressing common challenges with food purchase choice apps are found in reflection apps framed around behavior monitoring. Behavior monitoring is considered one of the more impactful BCTs [101,102]. Behavior monitoring can help prioritize actions [48], which is a common barrier to achieving behavior change. Behavior monitoring can be resistant to engagement issues [48], particularly those related to missing product data [47], as it does not depend on the information of a single product (as opposed to many selection apps and nudging approaches). It also provides opportunities for feedback on diet balance, a type of product information that was commonly associated with significant behavioral impact. Moreover, as reflection is independent of the point of purchase (usually happens after the purchase), the interaction with reflection apps is likely to be less time-pressured than the interaction with those designed to be used during shopping (eg, selection apps and nudging systems), which can provide better learning opportunities.

Practical Limitations and Future Opportunities for Reflection Apps

Relevant to the aforementioned, it must be noted that most reflection apps included in this study used loyalty card data or receipt scanning, approaches that facilitated easy data recording [42,47,49,52]. However, practical implementation of these approaches can present difficulties. The applicability of loyalty card data is limited to the people who use loyalty cards, and the availability depends on the retailers that make these data available. Although Swiss law seems to require retailers to make these data available through an application programming interface (API) [49], such regulations may not apply in other countries, which can pose constraints to designers who want to pursue such approaches. Moreover, receipt scanning requires trained machine learning algorithms and may not work optimally [47,52]; however, given that alternatives such as barcode scanning are often perceived as too effortful [85], receipt scanning seems to be a promising avenue for future designs. These limitations can be framed as opportunities for further research for developing effective food purchase choice apps.

Limitations

We acknowledge several limitations in our decisions for executing and reporting this review. We report these in no specific order.

The use of digital technology for behavior change can raise important ethical concerns [104-106] as these technologies have the potential to be manipulative. This review did not systematically assess the ethical concerns that might be associated with the app designs or the approaches that were taken to mitigate manipulative potential (ie, design approaches that respect individual needs and priorities, such as participatory design [107-109] or user-centered design). Such an assessment was beyond the scope of this study. However, designers of technology must be respectful of individual needs and priorities and follow user-centered approaches.

We did not calculate a meta-statistic estimate of the effects of food purchase choice apps. Appropriate reporting of this statistic relies on the stringent review procedure of a meta-analysis. We provided a simple summary analysis of study findings that provided context for the opportunities and challenges that were reported in the literature. Calculating a meta-statistic would be beyond this means and not our objective, and following the requirements for a meta-analysis would have compromised our ability to report on the objectives that we did formulate. Moreover, with the current body of literature, we see little justification for conducting a meta-analysis to calculate the impact of food purchase choice apps. The variety in reported outcome measures, together with a modest number of studies reporting evaluation results, complicates the calculation of a meta-statistic and restricts the conclusions that can be drawn.

The final charting was performed by only 1 person. This limited the rigor of establishing exact counts of the observed codes. However, the main function of observation counts was to indicate the magnitude of common patterns in designs, opportunities, and challenges. Establishing exact counts for each observation was less important. Importantly, the

XSL•FO
RenderX

identification of the common patterns themselves played a superior role in addressing the study objectives. Multiple reviewers were involved in the identification of patterns in the literature.

We acknowledge that the apps could have been described along with additional characteristics. We listed the characteristics that appeared most salient in describing the variations in apps without much interpretation. The apps could have been characterized by mapping techniques from the BCT Taxonomy (BCTT) [110]. However, we had several reasons not to characterize apps using BCTs. First, mapping along the BCTT requires a firm understanding of the various techniques as well as a firm understanding of the intervention that applies the techniques [111]. The comprehensiveness and clarity of app descriptions in the articles varied considerably, and it was not feasible to download and investigate the apps as most were not deployed on app stores. This limited our ability to perform accurate mapping of BCTs, with the potential for misleading interpretations drawn from this mapping. Second, the value of mapping BCTs is expected to be moderate. BCTs have been associated with the efficacy of behavioral intervention [101-103], and counting BCTs in apps is used as a practice to estimate the potential of an app to affect behavior [14,21]. However, several studies failed to find a relationship between the number of applied BCTs and behavioral outcomes [14,112,113]. As a binary mapping approach, mapping along the BCTT does not consider the quality of the BCT implementation or its prominence in the app as a whole. For these reasons, we decided not to prioritize the mapping of BCTs.

Conclusions

This scoping review examined the design characteristics, opportunities, and challenges of food purchase choice apps. Most food purchase choice apps are designed to help users select healthier products by scanning the barcode and displaying a nutritional rating. The value of this design comes from its potential to influence (nudge) users' decisions at the point of purchase (just in time), with little demand for conscious decision-making. However, our findings suggest that this design approach does not encourage repeated use and long-term adoption, which limits opportunities for behavior change through nudging. Instead, our results indicate that learning plays an important role in behavior change and that this mechanism is less dependent on continued use. We argue that designs that optimize learning within each interaction will have a better chance of achieving behavior change. This review concludes with design recommendations, suggesting that food purchase choice app designers (1) anticipate the possibility of early abandonment as part of their design process and (2) design apps that optimize the learning experience.

Acknowledgments

RBG was funded by the Leverhulme Doctoral Scholarship in Behaviour Informatics (DS-2017-015). Financial support for open-access publication of this paper was provided by the Newcastle University Research England QR Enhancing Research Culture Fund. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank Irina Pavlovskaya for her contribution to screening of the initial records.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

As the main author, RBG designed the scoping review and took the lead in data collection and analysis. GM and JWH contributed to article selection, data charting, and data analysis. CNB, DMdSM, JDS, and EM contributed by editing the manuscript. All authors have reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Completed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 640 KB - jmir_v26i1e45904_app1.pdf]

Multimedia Appendix 2
Overview of the included articles.
[DOCX File , 19 KB - jmir_v26i1e45904_app2.docx]

Multimedia Appendix 3
Overview of apps.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v26i1e45904_app3.xlsx]

XSL•FO
RenderX

Multimedia Appendix 4
Function descriptions.
[DOCX File , 31 KB - jmir_v26i1e45904_app4.docx]

Multimedia Appendix 5
Sankey diagram of app designs grouped by app type, product information, and retailer dependency.
[PNG File , 100 KB - jmir_v26i1e45904_app5.png]

Multimedia Appendix 6
Overview of evaluation studies by product information.
[DOCX File , 15 KB - jmir_v26i1e45904_app6.docx]

Multimedia Appendix 7
Overview of challenges and opportunities.
[XLSX File (Microsoft Excel File), 15 KB - jmir_v26i1e45904_app7.xlsx]

References

1. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, et al. Food in the Anthropocene: the EAT-Lancet commission on healthy diets from sustainable food systems. Lancet 2019 Feb 02;393(10170):447-492. [doi: 10.1016/S0140-6736(18)31788-4] [Medline: 30660336]
2. Clark MA, Domingo NG, Colgan K, Thakrar SK, Tilman D, Lynch J, et al. Global food system emissions could preclude achieving the 1.5° and 2°C climate change targets. Science 2020 Nov 06;370(6517):705-708. [doi: 10.1126/science.aba7357] [Medline: 33154139]
3. Vermeulen SJ, Campbell BM, Ingram JS. Climate change and food systems. Annu Rev Environ Resour 2012 Nov 21;37(1):195-222 [FREE Full text] [doi: 10.1146/annurev-environ-020411-130608]
4. Global report on diabetes. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/204871 [accessed 2023-12-21]
5. Leonard T, McKillop C, Carson JA, Shuval K. Neighborhood effects on food consumption. J Behav Exp Econ 2014 Aug;51:99-113. [doi: 10.1016/j.socec.2014.04.002]
6. Althoff T, Nilforoshan H, Hua J, Leskovec J. Large-scale diet tracking data reveal disparate associations between food environment and diet. Nat Commun 2022 Jan 18;13(1):267 [FREE Full text] [doi: 10.1038/s41467-021-27522-y] [Medline: 35042849]
7. What's in our food and on our mind. Nielsen. 2016. URL: https://www.nielsen.com/insights/2016/whats-in-our-food-and-on-our-minds/ [accessed 2019-09-28]
8. Yamoah FA, Acquaye A. Unravelling the attitude-behaviour gap paradox for sustainable food consumption: Insight from the UK apple market. J Clean Prod 2019 Apr;217:172-184. [doi: 10.1016/j.jclepro.2019.01.094]
9. Katzeff C, Milestad R, Zapico JL, Bohné U. Encouraging organic food consumption through visualization of personal shopping data. Sustainability 2020 Apr 29;12(9):3599 [FREE Full text] [doi: 10.3390/su12093599]
10. O'Brien MC, McConnon A, Hollywood LE, Cuskelly GJ, Barnett J, Raats M, et al. Let's talk about health: shoppers' discourse regarding health while food shopping. Public Health Nutr 2015 Apr;18(6):1001-1010 [FREE Full text] [doi: 10.1017/S1368980014001116] [Medline: 24893209]
11. Bellisle F. The Factors That Influence Our Food Choices. Eufic. URL: https://tinyurl.com/kdcw6j5n [accessed 2022-06-08]
12. Lawo D, Neifer T, Esau M, Stevens G. Buying the 'right' thing: designing food recommender systems with critical consumers. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan p. 1-13 URL: https://dl.acm.org/doi/10.1145/3411764.3445264 [doi: 10.1145/3411764.3445264]
13. Flaherty SJ, McCarthy M, Collins A, McAuliffe F. Can existing mobile apps support healthier food purchasing behaviour? content analysis of nutrition content, behaviour change theory and user quality integration. Public Health Nutr 2018 Feb;21(2):288-298 [FREE Full text] [doi: 10.1017/S1368980017002889] [Medline: 29081322]
14. Chan J, McMahon E, Brimblecombe J. Point-of-sale nutrition information interventions in food retail stores to promote healthier food purchase and intake: a systematic review. Obes Rev 2021 Oct;22(10):e13311. [doi: 10.1111/obr.13311] [Medline: 34254422]
15. Narasimhan P, Gandhi R, Rossi D. Smartphone-based assistive technologies for the blind. In: Proceedings of the 2009 international conference on Compilers, architecture, and synthesis for embedded systems. 2009 Presented at: CASES '09; October 11-16, 2009; Grenoble, France p. 223-232 URL: https://dl.acm.org/doi/abs/10.1145/1629395.1629427 [doi: 10.1145/1629395.1629427]
16. Bird J, Fozzati D, Harrison D, Marshall P. Healthy shopping: a longitudinal study of a mobile app to encourage a balanced diet. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication. 2013

XSL•FO
RenderX

Presented at: UbiComp '13; September 8-12, 2013; Zurich, Switzerland p. 1047-1054 URL: https://dl.acm.org/doi/abs/10.1145/2494091.2496012 [doi: 10.1145/2494091.2496012]
17. Holzmann SL, Holzapfel C. A scientific overview of smartphone applications and electronic devices for weight management in adults. J Pers Med 2019 Jun 07;9(2):31 [FREE Full text] [doi: 10.3390/jpm9020031] [Medline: 31181705]
18. Maringer M, Van't Veer P, Klepacz N, Verain MC, Normann A, Ekman S, et al. User-documented food consumption data from publicly available apps: an analysis of opportunities and challenges for nutrition research. Nutr J 2018 Jun 09;17(1):59 [FREE Full text] [doi: 10.1186/s12937-018-0366-6] [Medline: 29885653]
19. Ferrara G, Kim J, Lin S, Hua J, Seto E. A focused review of smartphone diet-tracking apps: usability, functionality, coherence with behavior change theory, and comparative validity of nutrient intake and energy estimates. JMIR Mhealth Uhealth 2019 May 17;7(5):e9232 [FREE Full text] [doi: 10.2196/mhealth.9232] [Medline: 31102369]
20. Willems K, Smolders A, Brengman M, Luyten K, Schöning J. The path-to-purchase is paved with digital opportunities: an inventory of shopper-oriented retail technologies. Technol Forecast Soc Change 2017 Nov;124:228-242. [doi: 10.1016/j.techfore.2016.10.066]
21. Mauch CE, Wycherley TP, Laws RA, Johnson BJ, Bell LK, Golley RK. Mobile apps to support healthy family food provision: systematic assessment of popular, commercially available apps. JMIR Mhealth Uhealth 2018 Dec 21;6(12):e11867 [FREE Full text] [doi: 10.2196/11867] [Medline: 30578213]
22. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
23. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
25. Panzone LA, Larcom S, She PW. Estimating the impact of the first COVID-19 lockdown on UK food retailers and the restaurant sector. Glob Food Sec 2021 Mar;28:100495 [FREE Full text] [doi: 10.1016/j.gfs.2021.100495] [Medline: 36568027]
26. Todd PM, Rogers Y, Payne SJ. Nudging the cart in the supermarket: how much is enough information for food shoppers? CEUR Workshop Proceedings 2010. CEUR Workshop Proc 2010;690:23-26 [FREE Full text]
27. Kalnikaitė V, Bird J, Rogers Y. Decision-making in the aisles: informing, overwhelming or nudging supermarket shoppers? Pers Ubiquit Comput 2012 Aug 25;17(6):1247-1259. [doi: 10.1007/s00779-012-0589-z]
28. Miro online whiteboard. URL: https://miro.com/index/ [accessed 2023-11-20]
29. Bardus M, Smith JR, Samaha L, Abraham C. Mobile phone and web 2.0 technologies for weight management: a systematic scoping review. J Med Internet Res 2015 Nov 16;17(11):e259 [FREE Full text] [doi: 10.2196/jmir.5129] [Medline: 26573984]
30. U.S. Department of Agriculture. MyPlate. URL: https://www.myplate.gov/ [accessed 2023-11-23]
31. The Eatwell guide. National Health Service. URL: http://tinyurl.com/2rhz436f [accessed 2023-11-23]
32. Mauch CE, Laws RA, Prichard I, Maeder AJ, Wycherley TP, Golley RK. Commercially available apps to support healthy family meals: user testing of app utility, acceptability, and engagement. JMIR Mhealth Uhealth 2021 May 07;9(5):e22990 [FREE Full text] [doi: 10.2196/22990] [Medline: 33960951]
33. Bomfim MC, Wallace JR. Pirate Bri's grocery adventure: teaching food literacy through shopping. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI EA '18; April 21-26, 2018; Montreal, QC p. 1-6 URL: https://dl.acm.org/doi/10.1145/3170427.3188496 [doi: 10.1145/3170427.3188496]
34. Díaz-Hellín P, Fontecha J, Hervás R, Bravo J. NFC as a childhood obesity treatment tool. J Med Syst 2015 Sep;39(9):96. [doi: 10.1007/s10916-015-0299-y] [Medline: 26254253]
35. Bomfim MC, Kirkpatrick SI, Nacke LE, Wallace JR. Food literacy while shopping: motivating informed food purchasing behaviour with a situated Gameful app. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI p. 1-13 URL: https://dl.acm.org/doi/10.1145/3313831.3376801 [doi: 10.1145/3313831.3376801]
36. Hedin B, Grönborg L, Johansson G. Food carbon literacy: a definition and framework exemplified by designing and evaluating a digital grocery list for increasing food carbon literacy and changing behavior. Sustainability 2022 Sep 29;14(19):12442 [FREE Full text] [doi: 10.3390/su141912442]
37. Clear AK, Friday A. Designing a food ‘Qualculator’. School of Computing and CommunicationsLancaster University. 2012. URL: https://www.academia.edu/2608996/Designing_a_Food_Qualculator_ [accessed 2023-11-18]
38. Bangia D, Shaffner DW, Palmer-Keenan DM. A point-of-purchase intervention using grocery store tour podcasts about Omega-3s increases long-term purchases of Omega-3-rich food items. J Nutr Educ Behav 2017 Jun;49(6):475-80.e1. [doi: 10.1016/j.jneb.2017.02.008] [Medline: 28377095]
39. Isley SC, Ketcham R, Arent DJ. Using augmented reality to inform consumer choice and lower carbon footprints. Environ Res Lett 2017 May 23;12(6):064002. [doi: 10.1088/1748-9326/aa6def]

XSL•FO
RenderX

40. Dunford E, Trevena H, Goodsell C, Ng KH, Webster J, Millis A, et al. FoodSwitch: a mobile phone app to enable consumers to make healthier food choices and crowdsourcing of national food composition data. JMIR Mhealth Uhealth 2014 Aug 21;2(3):e37 [FREE Full text] [doi: 10.2196/mhealth.3230] [Medline: 25147135]
41. van der Laan LN, Orcholska O. Effects of digital just-in-time nudges on healthy food choice – a field experiment. Food Qual Prefer 2022 Jun;98:104535. [doi: 10.1016/j.foodqual.2022.104535]
42. Reitberger W, Spreicer W, Fitzpatrick G. Situated and mobile displays for reflection on shopping and nutritional choices. Pers Ubiquit Comput 2014 Jul 26;18(7):1721-1735. [doi: 10.1007/s00779-014-0781-4]
43. Palacios C, Torres M, López D, Trak-Fellermeier MA, Coccia C, Pérez CM. Effectiveness of the nutritional app "MyNutriCart" on food choices related to purchase and dietary behavior: a pilot randomized controlled trial. Nutrients 2018 Dec 12;10(12):1967 [FREE Full text] [doi: 10.3390/nu10121967] [Medline: 30545125]
44. Asikis T, Klinglmayr J, Helbing D, Pournaras E. How value-sensitive design can empower sustainable consumption. R Soc Open Sci 2021 Jan;8(1):201418 [FREE Full text] [doi: 10.1098/rsos.201418] [Medline: 33614080]
45. Eyles H, McLean R, Neal B, Jiang Y, Doughty RN, McLean R, et al. A salt-reduction smartphone app supports lower-salt food purchases for people with cardiovascular disease: findings from the SaltSwitch randomised controlled trial. Eur J Prev Cardiol 2017 Sep;24(13):1435-1444 [FREE Full text] [doi: 10.1177/2047487317715713] [Medline: 28631933]
46. Eyles H, Grey J, Jiang Y, Umali E, McLean R, Te Morenga L, et al. Effectiveness of a sodium-reduction smartphone app and reduced-sodium salt to lower sodium intake in adults with hypertension: findings from the salt alternatives randomized controlled trial. JMIR Mhealth Uhealth 2023 Mar 09;11:e43675 [FREE Full text] [doi: 10.2196/43675] [Medline: 36892914]
47. Lurz M, Esterbauer A, Böhm M, Krcmar H. Feedback on shopping receipt data through a mobile app: a pilot study. In: Proceedings of the 56th Hawaii International Conference on System Sciences. 2023 Presented at: HICSS '23; January 3-6, 2023; Maui, HI p. 2880-2889 URL: http://tinyurl.com/37b23tw4
48. Zapico JL, Katzeff C, Bohné U, Milestad R. Eco-feedback visualization for closing the gap of organic food consumption. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction. 2016 Presented at: NordiCHI '16; October 23-27, 2016; Gothenburg, Sweden p. 1-9 URL: https://dl.acm.org/doi/10.1145/2971485.2971507 [doi: 10.1145/2971485.2971507]
49. Mönninghoff A, Fuchs K, Wu J, Albert J, Mayer S. The effect of a future-self avatar mobile health intervention (FutureMe) on physical activity and food purchases: randomized controlled trial. J Med Internet Res 2022 Jul 07;24(7):e32487 [FREE Full text] [doi: 10.2196/32487] [Medline: 35797104]
50. Samoggia A, Riedel B. Assessment of nutrition-focused mobile apps' influence on consumers' healthy food behaviour and nutrition knowledge. Food Res Int 2020 Feb;128:108766. [doi: 10.1016/j.foodres.2019.108766] [Medline: 31955740]
51. López D, Torres M, Vélez J, Grullon J, Negrón E, Pérez CM, et al. Development and evaluation of a nutritional smartphone application for making smart and healthy choices in grocery shopping. Healthc Inform Res 2017 Jan;23(1):16-24 [FREE Full text] [doi: 10.4258/hir.2017.23.1.16] [Medline: 28261527]
52. Bohné U, Zapico JL, Katzeff C. The EcoPanel - designing for reflection on greener grocery shopping practices. Adv Comput Res 2015 [FREE Full text] [doi: 10.2991/ict4s-env-15.2015.26]
53. Hegen M. Exploring the impact of mobile product information systems on consumers' perceived food quality. In: Proceedings of the 20th Pacific Asia Conference on Information Systems. 2016 Presented at: PACIS '16; June 27-July 1, 2016; Chiayi, Taiwan p. 1 URL: https://aisel.aisnet.org/amcis2016/DigitalComm/Presentations/28
54. Vintsarevich V, Raikundalia G, Mathai ML, McAinch AJ, Battaglia C, Pereira S. Development of mobile-accessible nutritional system to improve healthy food choices. Res J Appl Sci Eng Technol 2011;3(9):1047.
55. Fuchs K, Barattin T, Haldimann M, Ilic A. Towards tailoring digital food labels: insights of a Smart-RCT on user-specific interpretation of food composition data. In: Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management. 2019 Presented at: MADiMa '19; October 21, 2019; Nice, France p. 67-75 URL: https://dl.acm.org/doi/10.1145/3347448.3357171 [doi: 10.1145/3347448.3357171]
56. Mahdi S, Chilcott J, Buckland NJ. Evaluating the Change4Life food scanner app in reducing children's sugar intake: a randomised pilot and feasibility study. Lancet 2022 Nov;400 Suppl 1:S13. [doi: 10.1016/S0140-6736(22)02223-1] [Medline: 36929955]
57. Tomlinson B. Prototyping a community-generated, mobile device-enabled database of environmental impact reviews of consumer products. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences. 2008 Presented at: HICSS '08; January 7-10, 2008; Waikoloa, HI p. 154 URL: https://ieeexplore.ieee.org/document/4438857 [doi: 10.1109/hicss.2008.365]
58. Herbig N, Kahl G, Krüger A. Design guidelines for assistance systems supporting sustainable purchase decisions. In: Proceedings of the 2018 Designing Interactive Systems Conference. 2018 Presented at: DIS '18; June 9-13, 2018; Hong Kong, China p. 1333-1344 URL: https://dl.acm.org/doi/10.1145/3196709.3196726 [doi: 10.1145/3196709.3196726]
59. Normark M, Tholander J. Performativity in sustainable interaction: the case of seasonal grocery shopping in ecofriends. In: Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 21-May 1, 2014; Toronto, ON p. 271-280 URL: https://dl.acm.org/doi/10.1145/2556288.2557318 [doi: 10.1145/2556288.2557318]

XSL•FO
RenderX

60. Broll G, Palleis H, Richter H, Wiethoff A. Exploring multimodal feedback for an NFC-based mobile shopping assistant. In: Proceedings of the 5th International Workshop on Near Field Communication. 2013 Presented at: NFC '13; February 5, 2013; Zurich, Switzerland p. 1-4 URL: https://ieeexplore.ieee.org/document/6482443 [doi: 10.1109/nfc.2013.6482443]
61. Jayananda PK, Seneviratne DH, Abeygunawardhana P, Dodampege L, Lakshani AM. Augmented reality based smart supermarket system with indoor navigation using beacon technology (easy shopping android mobile app). In: Proceedings of the 2018 IEEE International Conference on Information and Automation for Sustainability. 2018 Presented at: ICIAfS '18; December 21-22, 2018; Colombo, Sri Lanka p. 1-6 URL: https://ieeexplore.ieee.org/document/8913363 [doi: 10.1109/iciafs.2018.8913363]
62. Ahn J, Williamson J, Gartrell M, Han R, Lv Q, Mishra S. Supporting healthy grocery shopping via mobile augmented reality. ACM Trans Multimedia Comput Commun Appl 2015 Oct 21;12(1s):1-24 [FREE Full text] [doi: 10.1145/2808207]
63. Hormann LB, Putz V, Rudic B, Kastl C, Klinglmayr J, Pournaras E. Augmented shopping experience for sustainable consumption using the internet of thing. IEEE Internet Things Mag 2019 Sep;2(3):46-51 [FREE Full text] [doi: 10.1109/iotm.0001.1900047]
64. Tsai SM, Chuang ML, Yan HW, Yao YC, Chen GW, Chen LH, et al. Identification system of fish freshness based on deep learning. In: Proceedings of the 2021 IEEE International Conference on Consumer Electronics-Taiwan. 2021 Presented at: ICCE-TW '21; June 16-18, 2021; Penghu, Taiwan p. 1-2 URL: https://ieeexplore.ieee.org/document/9602879 [doi: 10.1109/icce-tw52618.2021.9602879]
65. Head M, Sevenster M, Odegard I, Krutwagen B, Croezen H, Bergsma G. Life cycle impacts of protein-rich foods: creating robust yet extensive life cycle models for use in a consumer app. J Clean Prod 2014 Jun;73:165-174. [doi: 10.1016/j.jclepro.2013.11.026]
66. Waltner G, Schwarz M, Ladstätter S, Weber A, Luley P, Bischof H, et al. MANGO - mobile augmented reality with functional eating guidance and food awareness. In: Proceedings of the 2015 International Workshops on New Trends in Image Analysis and Processing. 2015 Presented at: ICIAP '15; September 7-8, 2015; Genoa, Italy p. 425-432 URL: https://link.springer.com/chapter/10.1007/978-3-319-23222-5_52 [doi: 10.1007/978-3-319-23222-5_52]
67. Röddiger T, Doerner D, Beigl M. ARMart: AR-based shopping assistant to choose and find store items. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018 Presented at: UbiComp '18; October 8-12, 2018; Singapore, Singapore p. 440-443 URL: https://dl.acm.org/doi/abs/10.1145/3267305.3267631 [doi: 10.1145/3267305.3267631]
68. Fagerstrøm A, Eriksson N, Khamtanet S, Jitkuekul P, Sigurdsson V, Larsen NM. The relative impact of health communication conveyed via quick response codes: a conjoint experiment among young thai consumers doing grocery shopping. Health Mark Q 2023;40(2):206-225. [doi: 10.1080/07359683.2022.2085460] [Medline: 35758221]
69. Flaherty SJ, McCarthy MB, Collins AM, McCafferty C, McAuliffe FM. A phenomenological exploration of change towards healthier food purchasing behaviour in women from a lower socioeconomic background using a health app. Appetite 2020 Apr 01;147:104566. [doi: 10.1016/j.appet.2019.104566] [Medline: 31866355]
70. Kulpy A, Bekaroo G. Fruitify: nutritionally augmenting fruits through markerless-based augmented reality. In: Proceedings of the EEE 4th International Conference on Soft Computing & Machine Intelligence. 2017 Presented at: ISCMI '17; November 23-24, 2017; Port Louis, Mauritius p. 149-153 URL: https://ieeexplore.ieee.org/document/8279616 [doi: 10.1109/iscmi.2017.8279616]
71. Abao RP, Malabanan CV, Galido AP. Design and Development of FoodGo: A Mobile Application using Situated Analytics to Augment Product Information. Procedia Comput Sci 2018;135:186-193. [doi: 10.1016/j.procs.2018.08.165]
72. Akay K, Kandemir Ü, Dalkılıç F. Android-based personalized shopping assistant with navigation and augmented reality support. In: Proceedings of the 2022 Innovations in Intelligent Systems and Applications Conference. 2022 Presented at: ASYU '22; September 7-9, 2022; Virtual Event p. 1-6 URL: https://doi.org/10.1109/ASYU56188.2022.9925270 [doi: 10.1109/asyu56188.2022.9925270]
73. Dorman K, Yahyanejad M, Nahapetian A, Suh MK, Sarrafzadeh M, McCarthy W, et al. Nutrition monitor: a food purchase and consumption monitoring mobile system. In: Proceedings of the 1st International ICST Conference on Mobile Computing, Applications, and Services. 2010 Presented at: MobiCASE '09; October 26-29, 2009; San Diego, CA p. 1-11 URL: https://doi.org/10.1007/978-3-642-12607-9_1 [doi: 10.1007/978-3-642-12607-9_1]
74. Govoruhina A, Nikiforova A. Digital health shopping assistant with React Native: a simple technological solution to a complex health problem. In: Proceedings of the 2022 International Conference on Intelligent Data Science Technologies and Applications. 2022 Presented at: IDSTA '22; September 5-7, 2022; San Antonio, TX p. 34-40 URL: https://doi.org/10.1109/IDSTA55301.2022.9923047 [doi: 10.1109/idsta55301.2022.9923047]
75. Gutiérrez F, Verbert K, Htun NN. PHARA: an augmented reality grocery store assistant. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct. 2018 Presented at: MobileHCI '18; September 3-6, 2018; Barcelona, Spain p. 339-345 URL: https://doi.org/10.1145/3236112.3236161 [doi: 10.1145/3236112.3236161]
76. Harada T, Inoue K, Fukae K, Kobayashi T. Fruits traceability system. In: Proceedings of the IEEE 11th Global Conference on Consumer Electronics. 2022 Presented at: GCCE '22; October 18-21, 2022; Osaka, Japan p. 262-263 URL: https://doi.org/10.1109/GCCE56475.2022.10014042 [doi: 10.1109/gcce56475.2022.10014042]

XSL•FO
RenderX

77. Kassim M, Yahaya CK, Zaharuddin MH, Bakar ZA. A prototype of Halal product recognition system. In: Proceedings of the 2012 International Conference on Computer & Information Science. 2012 Presented at: ICCIS '12; August 3-5, 2012; Kuala Lumpur, Malaysia p. 990-994 URL: https://doi.org/10.1109/ICCISci.2012.6297170 [doi: 10.1109/iccisci.2012.6297170]
78. Serhani MA, Ahmad A, Alkhatri A, Almansoori M, Alkhyeli A. Non-invasive health control through guided shopping: a rule-based approach. In: Proceedings of the 3rd International Conference on Intelligent Computing in Data Sciences. 2019 Presented at: ICDS '19; October 28-30, 2019; Marrakech, Morocco p. 1-8 URL: https://doi.org/10.1109/ICDS47004.2019.8942238 [doi: 10.1109/icds47004.2019.8942238]
79. Waltner G, Schwarz M, Ladstätter S, Weber A, Luley P, Lindschinger M, et al. Personalized dietary self-management using mobile vision-based assistance. In: Proceedings of the 2017 International Workshops on New Trends in Image Analysis and Processing. 2017 Presented at: ICIAP '17; September 11-15, 2017; Catania, Italy p. 385-393 URL: https://doi.org/10.1007/978-3-319-70742-6_36 [doi: 10.1007/978-3-319-70742-6_36]
80. Wiley C, Saka E, Tauber S, Kang SR. Shopping cart interactive program (SCIP). In: Proceedings of the 14th International Conference, HCI International Conference on Human-Computer Interaction: Users and Applications. 2011 Presented at: HCI '11; July 9-14, 2011; Orlando, FL p. 285-293 URL: https://doi.org/10.1007/978-3-642-21619-0_37 [doi: 10.1007/978-3-642-21619-0_37]
81. Sackey DJ, Ullmann N. Visualizing data, encouraging change: technical interventions in food purchasing. In: Proceedings of the 2012 IEEE International Professional Communication Conference. 2012 Presented at: IPCC '12; October 8-10, 2012; Orlando,FL p. 1-5. [doi: 10.1109/IPCC.2012.6408654]
82. Julia C, Péneau S, Buscail C, Gonzalez R, Touvier M, Hercberg S, et al. Perception of different formats of front-of-pack nutrition labels according to sociodemographic, lifestyle and dietary factors in a French population: cross-sectional study among the NutriNet-Santé cohort participants. BMJ Open 2017 Jun 15;7(6):e016108 [FREE Full text] [doi: 10.1136/bmjopen-2017-016108] [Medline: 28619781]
83. Franco A, Malhotra N, Simonovits G. Social science. Publication bias in the social sciences: unlocking the file drawer. Science 2014 Sep 19;345(6203):1502-1505. [doi: 10.1126/science.1255484] [Medline: 25170047]
84. Chan AW, Pello A, Kitchen J, Axentiev A, Virtanen JI, Liu A, et al. Association of trial registration with reporting of primary outcomes in protocols and publications. JAMA 2017 Nov 07;318(17):1709-1711 [FREE Full text] [doi: 10.1001/jama.2017.13001] [Medline: 28892118]
85. Cordeiro F, Epstein DA, Thomaz E, Bales E, Jagannathan AK, Abowd GD, et al. Barriers and negative nudges: exploring challenges in food journaling. Proc SIGCHI Conf Hum Factor Comput Syst 2015 Apr;2015:1159-1162. [doi: 10.1145/2702123.2702155] [Medline: 26894233]
86. Lazar A, Koehler C, Tanenbaum TJ, Nguyen DH. Why we use and abandon smart devices. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 653-646 URL: https://dl.acm.org/doi/10.1145/2750858.2804288 [doi: 10.1145/2750858.2804288]
87. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219 [FREE Full text] [doi: 10.1037/hea0000306] [Medline: 26651462]
88. Goldstein SP, Evans BC, Flack D, Juarascio A, Manasse S, Zhang F, et al. Return of the JITAI: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors. Int J Behav Med 2017 Oct;24(5):673-682 [FREE Full text] [doi: 10.1007/s12529-016-9627-y] [Medline: 28083725]
89. Caraban AK, Karapanos E, Gonçalves DJ, Campos P. 23 ways to nudge: a review of technology-mediated nudging in human-computer interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4 -9, 2019; Glasgow, Scotland p. 1-15 URL: https://dl.acm.org/doi/10.1145/3290605.3300733 [doi: 10.1145/3290605.3300733]
90. Thaler RH, Sunstein CR. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale, MI: Yale University Press; 2008.
91. Sunstein CR. Nudging: a very short guide. J Consum Policy 2014 Oct 16;37(4):583-588. [doi: 10.1007/s10603-014-9273-1]
92. Kalnikaite V, Rogers Y, Bird J, Villar N, Bachour K, Payne S, et al. How to nudge in Situ: designing lambent devices to deliver salient information in supermarkets. In: Proceedings of the 13th International Conference on Ubiquitous Computing. 2011 Presented at: UbiComp '11; September 17-21, 2011; Beijing, China p. 11-20 URL: https://dl.acm.org/doi/10.1145/2030112.2030115 [doi: 10.1145/2030112.2030115]
93. Ni Mhurchu C, Volkova E, Jiang Y, Eyles H, Michie J, Neal B, et al. Effects of interpretive nutrition labels on consumer food purchases: the Starlight randomized controlled trial. Am J Clin Nutr 2017 Mar;105(3):695-704 [FREE Full text] [doi: 10.3945/ajcn.116.144956] [Medline: 28148503]
94. Vhaduri S, Prioleau T. Adherence to personal health devices: a case study in diabetes management. In: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2020 Presented at: PervasiveHealth '20; May 18-20, 2020; Atlanta, GA p. 62-72 URL: https://dl.acm.org/doi/abs/10.1145/3421937.3421977 [doi: 10.1145/3421937.3421977]

XSL•FO
RenderX

95. Epstein DA, Caraway M, Johnston C, Ping A, Fogarty J, Munson SA. Beyond abandonment to next steps: understanding and designing for life after personal informatics tool use. Proc SIGCHI Conf Hum Factor Comput Syst 2016 May;2016:1109-1113 [FREE Full text] [doi: 10.1145/2858036.2858045] [Medline: 28503678]
96. Park JS, Bernstein MS, Brewer RN, Kamar E, Morris MR. Understanding the representation and representativeness of age in AI data sets. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021 Presented at: AIES '21; May 19-21, 2021; Virtual Event p. 834-842 URL: https://dl.acm.org/doi/10.1145/3461702.3462590 [doi: 10.1145/3461702.3462590]
97. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
98. Young C, Campolonghi S, Ponsonby S, Dawson SL, O'Neil A, Kay-Lambkin F, et al. Supporting engagement, adherence, and behavior change in online dietary interventions. J Nutr Educ Behav 2019 Jun;51(6):719-739. [doi: 10.1016/j.jneb.2019.03.006] [Medline: 31036500]
99. Attig C, Franke T. Abandonment of personal quantification: a review and empirical study investigating reasons for wearable activity tracking attrition. Comput Human Behav 2020 Jan;102:223-237. [doi: 10.1016/j.chb.2019.08.025]
100. Melo B, Andrade R, Darin T. Longitudinal user experience studies in the IoT domain: a brief panorama and challenges to overcome. In: Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems. 2022 Presented at: IHC '22; October 17-21, 2022; Diamantina, Brazil p. 1-13 URL: https://dl.acm.org/doi/10.1145/3554364.3559135 [doi: 10.1145/3554364.3559135]
101. Humphreys G, Evans R, Makin H, Cooke R, Jones A. Identification of behavior change techniques from successful web-based interventions targeting alcohol consumption, binge eating, and gambling: systematic review. J Med Internet Res 2021 Feb 09;23(2):e22694 [FREE Full text] [doi: 10.2196/22694] [Medline: 33560243]
102. Aguiar M, Trujillo M, Chaves D, Álvarez R, Epelde G. mHealth apps using behavior change techniques to self-report data: systematic review. JMIR Mhealth Uhealth 2022 Sep 09;10(9):e33247 [FREE Full text] [doi: 10.2196/33247] [Medline: 36083606]
103. Lyons EJ, Lewis ZH, Mayrsohn BG, Rowland JL. Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis. J Med Internet Res 2014 Aug 15;16(8):e192 [FREE Full text] [doi: 10.2196/jmir.3469] [Medline: 25131661]
104. Kuyer P, Gordijn B. Nudge in perspective: a systematic literature review on the ethical issues with nudging. Ration Soc 2023 Jan 30;35(2):191-230. [doi: 10.1177/10434631231155005]
105. Busch J, Madsen EK, Fage-Butler AM, Kjær M, Ledderer L. Dilemmas of nudging in public health: an ethical analysis of a Danish pamphlet. Health Promot Int 2021 Aug 30;36(4):1140-1150. [doi: 10.1093/heapro/daaa146] [Medline: 33367635]
106. da Rocha CA, Hunziker MH. A behavior-analytic view on nudges: individual, technique, and ethics. Behav Soc Issues 2020 Nov 17;29(1):138-161 [FREE Full text] [doi: 10.1007/s42822-020-00037-9]
107. Tengland PA. Behavior change or empowerment: on the ethics of health-promotion strategies. Public Health Ethics 2012 Sep 21;5(2):140-153. [doi: 10.1093/phe/phs022]
108. Simonsen J, Robertson T. Routledge International Handbook of Participatory Design. London, UK: Routledge; 2012.
109. Davis J. Early experiences with participation in persuasive technology design. In: Proceedings of the 12th Participatory Design Conference: Research Papers - Volume 1. 2012 Presented at: PDC '12; August 12-16, 2012; Roskilde, Denmark p. 119-128 URL: https://dl.acm.org/doi/10.1145/2347635.2347653 [doi: 10.1145/2347635.2347653]
110. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
111. Wood CE, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. Applying the behaviour change technique (BCT) taxonomy v1: a study of coder training. Transl Behav Med 2015 Jun;5(2):134-148 [FREE Full text] [doi: 10.1007/s13142-014-0290-z] [Medline: 26029276]
112. Presseau J, Ivers NM, Newham JJ, Knittle K, Danko KJ, Grimshaw JM. Using a behaviour change techniques taxonomy to identify active ingredients within trials of implementation interventions for diabetes care. Implement Sci 2015 Apr 23;10:55 [FREE Full text] [doi: 10.1186/s13012-015-0248-7] [Medline: 25900104]
113. Dombrowski SU, Sniehotta FF, Avenell A, Johnston M, MacLennan G, Araújo-Soares V. Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. Health Psychology Review 2012 Mar;6(1):7-32. [doi: 10.1080/17437199.2010.513298]

Abbreviations

API: application programming interface
BCT: behavior change technique
BCTT: Behavior Change Technique Taxonomy

XSL•FO
RenderX

NFC: near-field communication
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RFID: radiofrequency identification

Edited by A Mavragani; submitted 14.02.23; peer-reviewed by S Flaherty, M Khokhar; comments to author 06.04.23; revised version received 30.11.23; accepted 20.12.23; published 06.03.24.

Please cite as:
Benthem de Grave R, Bull CN, Monjardino de Souza Monteiro D, Margariti E, McMurchy G, Hutchinson JW, Smeddinck JD
Smartphone Apps for Food Purchase Choices: Scoping Review of Designs, Opportunities, and Challenges
J Med Internet Res 2024;26:e45904
URL: https://www.jmir.org/2024/1/e45904
doi:10.2196/45904
PMID:38446500

©Remco Benthem de Grave, Christopher N Bull, Diogo Monjardino de Souza Monteiro, Eleni Margariti, Gareth McMurchy, Joseph William Hutchinson, Jan David Smeddinck. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Impact of Digital Hospitals on Patient and Clinician Experience: Systematic Review and Qualitative Evidence Synthesis

Oliver J Canfell[1,2,3,4*], PhD; Leanna Woods[1,2*], PhD; Yasaman Meshkat[5], MBBS; Jenna Krivit[5], BSc; Brinda Gunashanhar[5], BSc; Christine Slade[6], PhD; Andrew Burton-Jones[4], PhD; Clair Sullivan[1,2,7], MBBS, MD

[1]Centre for Health Services Research, Faculty of Medicine, The University of Queensland, Brisbane, Australia
[2]Queensland Digital Health Centre, Faculty of Medicine, The University of Queensland, Brisbane, Australia
[3]Digital Health Cooperative Research Centre, Australian Government, Sydney, Australia
[4]UQ Business School, Faculty of Business, Economics and Law, The University of Queensland, Brisbane, Australia
[5]School of Clinical Medicine, Faculty of Medicine, The University of Queensland, Brisbane, Australia
[6]Institute for Teaching and Learning Innovation, The University of Queensland, Brisbane, Australia
[7]Metro North Hospital and Health Service, Department of Health, Queensland Government, Brisbane, Australia
[*]these authors contributed equally

Corresponding Author:
Oliver J Canfell, PhD
Centre for Health Services Research
Faculty of Medicine
The University of Queensland
Level 5 Health Sciences Building
Central, Fig Tree Cres
Brisbane, 4006
Australia
Phone: 61 731765530
Email: o.canfell@uq.edu.au

Abstract

Background: The digital transformation of health care is advancing rapidly. A well-accepted framework for health care improvement is the Quadruple Aim: improved clinician experience, improved patient experience, improved population health, and reduced health care costs. Hospitals are attempting to improve care by using digital technologies, but the effectiveness of these technologies is often only measured against cost and quality indicators, and less is known about the clinician and patient experience.

Objective: This study aims to conduct a systematic review and qualitative evidence synthesis to assess the clinician and patient experience of digital hospitals.

Methods: The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and ENTREQ (Enhancing the Transparency in Reporting the Synthesis of Qualitative Research) guidelines were followed. The PubMed, Embase, Scopus, CINAHL, and PsycINFO databases were searched from January 2010 to June 2022. Studies that explored multidisciplinary clinician or adult inpatient experiences of digital hospitals (with a full electronic medical record) were included. Study quality was assessed using the Mixed Methods Appraisal Tool. Data synthesis was performed narratively for quantitative studies. Qualitative evidence synthesis was performed via (1) automated machine learning text analytics using Leximancer (Leximancer Pty Ltd) and (2) researcher-led inductive synthesis to generate themes.

Results: A total of 61 studies (n=39, 64% quantitative; n=15, 25% qualitative; and n=7, 11% mixed methods) were included. Most studies (55/61, 90%) investigated clinician experiences, whereas few (10/61, 16%) investigated patient experiences. The study populations ranged from 8 to 3610 clinicians, 11 to 34,425 patients, and 5 to 2836 hospitals. Quantitative outcomes indicated that clinicians had a positive overall satisfaction (17/24, 71% of the studies) with digital hospitals, and most studies (11/19, 58%) reported a positive sentiment toward usability. Data accessibility was reported positively, whereas adaptation, clinician-patient interaction, and workload burnout were reported negatively. The effects of digital hospitals on patient safety and clinicians' ability

XSL•FO
RenderX

to deliver patient care were mixed. The qualitative evidence synthesis of clinician experience studies (18/61, 30%) generated 7 themes: inefficient digital documentation, inconsistent data quality, disruptions to conventional health care relationships, acceptance, safety versus risk, reliance on hybrid (digital and paper) workflows, and patient data privacy. There was weak evidence of a positive association between digital hospitals and patient satisfaction scores.

Conclusions: Clinicians' experience of digital hospitals appears positive according to high-level indicators (eg, overall satisfaction and data accessibility), but the qualitative evidence synthesis revealed substantive tensions. There is insufficient evidence to draw a definitive conclusion on the patient experience within digital hospitals, but indications appear positive or agnostic. Future research must prioritize equitable investigation and definition of the digital clinician and patient experience to achieve the Quadruple Aim of health care.

(J Med Internet Res 2024;26:e47715) doi:10.2196/47715

KEYWORDS

electronic medical record; electronic health record; health care professionals; patients; patient satisfaction; hospitals; eHealth; attitude; perception; systematic; digital hospital; digital hospitals; experience; satisfaction

Introduction

Background

Investment in digital health is advancing rapidly. In 2020, the total global funding for digital health was the highest recorded at US $26.5 billion [1]. A global appetite for digital health, fueled recently by the COVID-19 pandemic and the associated rapid adoption of point-of-care technological solutions [2], including telehealth [3], has driven the digital disruption of health care. A core pillar of digital health investment is the digital transformation of hospitals [4].

The World Health Organization Global Strategy on Digital Health (2020-2025) recommends the implementation of a national digital health architecture, including digital hospitals [5]. Digital hospitals represent significant jurisdictional investments to improve the quality and safety of acute care [6]. A digital hospital uses a comprehensive electronic medical record (EMR) to achieve its clinical goals [7] and is becoming the predominant method of care delivery worldwide. These new digital hospital environments radically disrupt well-rehearsed clinical workflows and create unfamiliar environments for patients and clinicians, potentially affecting quality, safety, and experience of care [8-10].

It has been difficult to determine the value of digital hospital implementations as what is considered valuable changes over time and place and from person to person [11]. Previous studies evaluating digital health implementations have focused on three domains: (1) improving patient or hospital outcomes using quantitative evaluation [12]; (2) exploring patient [13] and clinician behavior, workflows, and attitudes toward EMRs or digital hospital transformations [10]; and (3) quantifying value using economic evaluations [4]. Evidence to date demonstrates conflicting impacts of EMRs on hospital practice, with positive indications for medication safety, guideline adherence, and decision support [12,14] and negative indications for physician-patient communication, staff attitude, and workflow disruption [15,16]. Focusing on the narrow aspects of digital health implementations has resulted in patchy assessments of the value of digital health technologies.

The Quadruple Aim is the overarching goal of a learning health care system of enhancing patient experience, improving population health, reducing health care costs, and improving the provider experience (Figure 1) [17]. The Quadruple Aim of health care has been proposed as a strategic compass to guide digital health investment planning, decision-making, and evaluation [11] and has been used in the health care workforce [18,19], innovation implementation [20], and COVID-19 pandemic [21] contexts to identify current trends and research gaps.

XSL•FO
RenderX

Figure 1. Quadruple Aim of health care [17].

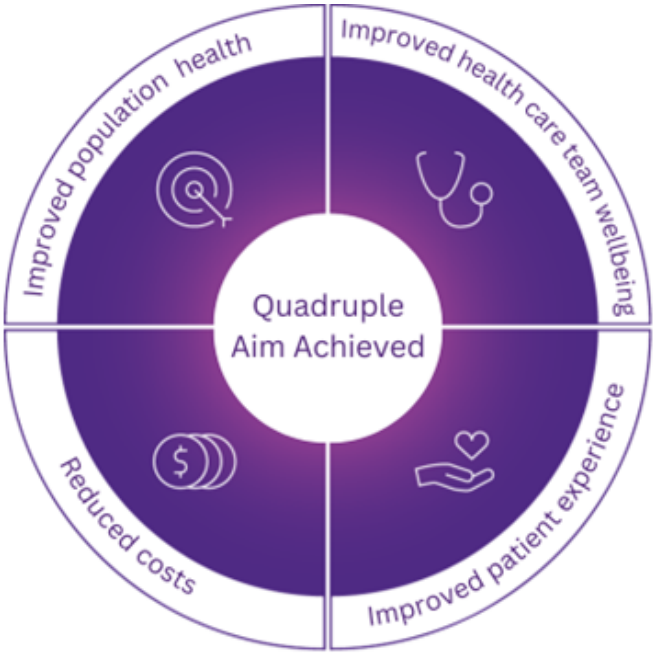

The experience of patients and clinicians has yet to be explored as important contributors to the Quadruple Aim. Previous evaluations of clinician experience have focused on individual retrospective recalls of attitudes, perceptions of EMR implementation [10], and observational time and motion studies [22]. Existing patient experience research has focused on bespoke digital systems for patient use (eg, internet-based care technologies, web-based patient platforms, and mobile health apps) or emerging trends (eg, COVID-19 impacts, effects of specific technologies, research methods, and new technologies) [23,24]. Traditional evaluations of technology in health care are selective in the outcomes they measure, with an overwhelming focus on clinical outcomes and efficiency.

Objectives

To address this research gap, our research question was as follows: *what is the clinician and patient experience of digital hospitals?* We hypothesized that clinicians and patients would report digital hospital experiences positively (eg, patient safety benefits because of digital safeguards [25]), negatively (eg, productivity loss because of documentation burden [15]), and ambivalently (eg, no observed impact of the digital environment on patient experience [10]). The study aim was to conduct a systematic literature review and qualitative evidence synthesis to assess the clinician and patient experience of digital hospitals.

Methods

Search Strategy and Identification of Included Articles

This review adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement checklist [26] (Multimedia Appendix 1 [26]) and the ENTREQ (Enhancing the Transparency in Reporting the Synthesis of Qualitative Research) guidelines [27] (Multimedia Appendix 2). The protocol for this review was registered in PROSPERO (CRD42021258719).

The PubMed, Embase, Scopus, CINAHL, and PsycINFO databases were searched twice—on June 24, 2021 (version [V1]), and subsequently on June 23, 2022 (version [V2])—using the same search strategy restricted by year of publication (2010 to the present) because of the relative novelty of the digital transformation of acute care.

The search strategy was initially developed in PubMed using the population, intervention, comparison, and outcome structure (Table 1) [28] and translated to other databases (Multimedia Appendix 3). We define experience as an individual's perception of events, incorporating themes of expectation and satisfaction. Thus, synonyms for "experience" were used to characterize the outcome. A combination of indexed terms (eg, Medical Subject Headings) and keywords identified after consultation with a research librarian and subject matter experts was used. Truncations, synonyms, and terminological variations (eg, "EMR" vs "EHR") were also used.

XSL•FO
RenderX

Table 1. Final search strategy executed in PubMed (version 1: June 24, 2021; version 2: June 23, 2022).

PICO[a] category	Domain	Search terms
Population	Patients and clinicians in a digital hospital	"Hospitals" (MeSH[b]) OR "hospital*" (tiab[c])
Intervention	Digital hospital (using an EMR[d] or EHR[e])	"electronic health records" (MeSH) OR "electronic health record*"(ti[f]) OR "electronic medical record*" (ti) OR "emr" (ti) OR "EHR" (ti) OR "digital hospital*" (tiab) OR "smart hospital*" (tiab) OR "digital hospital*" (tiab) OR "computerized medical record" (tiab)
Comparison	None	None
Outcome	Experience	"satisfaction" (tiab) OR "experience" (tiab) OR "attitude*" (tiab) OR "perception*" (tiab) OR "opinion*" (tiab) OR "behavior*" (tiab) OR "behaviour*" (tiab)

[a]PICO: population, intervention, comparison, and outcome.

[b]MeSH: Medical Subject Headings.

[c]tiab: title and abstract.

[d]EMR: electronic medical record.

[e]EHR: electronic health record.

[f]ti: title.

In total, 2 reviewers (YM and OJC) performed title and abstract screening. Full-text review was then performed based on the eligibility criteria. Backward citation tracking (snowballing) was used to identify additional articles in the reference lists of included articles and relevant reviews. Decision conflicts were resolved through internal discussion or by involving a third reviewer's opinion (C Slade) when required.

Eligibility Criteria

Studies were included if they described a quantitative, qualitative, or mixed methods investigation of clinician or adult inpatient "experience" in a digital hospital (Table 2). This review focused on multidisciplinary clinicians and adult inpatients as a first step in synthesizing evidence of digital hospital experience. Pediatric inpatients were excluded because of possible environmental factors that could confound the digital hospital experience (eg, patient entertainment systems). Our study setting prioritized EMRs—a real-time patient health record that collects, stores, and displays clinical information in a tertiary setting [29] (eg, Cerner Millenium and Epic)—as the foundation of a digital hospital [30] as opposed to an electronic health record (or personal health record), which displays summarized patient information to the consumer in the community and across multiple health care providers [29] (eg, My Health Record [Australia], the National Health Service app [United Kingdom], and personal health record [Ministry of National Guard Health Affairs, Kingdom of Saudi Arabia]). The terms "EMR" and "EHR" may be used interchangeably in some countries, so both terms were adopted in the search strategy. The stage of digital hospital implementation (eg, EMR maturity [31]) was not considered.

XSL•FO
RenderX

Table 2. Study eligibility criteria for this systematic review and qualitative evidence synthesis.

Domain	Inclusion criteria	Exclusion criteria
Population	• Multidisciplinary clinicians, including physicians, surgeons, nurses, allied health professionals and pharmacists, or adult inpatients	• Nonclinical hospital staff, including hospital executive and administration staff • Pediatric inpatients
Setting	• Defined as a digital hospital or defined as a digital hospital by our research team—an acute care (hospital) service that uses an EMRa or any terminological variation to contribute to the goals of the health service [30]	• Study not conducted in an acute hospital setting (eg, primary care, secondary care, internet-based health interventions, IoTb interventions, Hospital in the Home services, telehealth in primary care settings, mobile health, and public health) • Study not concerning an EMR or any terminological variation in a hospital setting
Design	• Qualitative methods for data collection (eg, interviews, focus groups, or ethnography) and data analysis (eg, thematic analysis or content analysis) • Quantitative methods (eg, survey) to collect and analyze data • Mixed methods studies in which either component fulfilled our eligibility criteria	• Qualitative studies that do not include verbatim quotes from participants (to enable evidence synthesis) • Articles not published as full-text empirical studies (ie, abstracts, conference proceedings, gray literature, dissertations, or theses; because of depth and quality variations) • Narrative or systematic literature reviews; however, their reference lists were checked for relevant articles before exclusion (as our review focused on individual empirical studies) • Articles published before 2010 (because of the rapid adoption of EMRs in hospitals since 2010) • Articles not published in English (because of limited funding available for large-scale translation)

[a]EMR: electronic medical record.

[b]IoT: Internet of Things.

Data Extraction

The Covidence software (Veritas Health Innovation) and Excel (Microsoft Corp) facilitated data extraction of study details (Multimedia Appendix 4). For studies eligible for the qualitative evidence synthesis (eg, presence of verbatim participant quotes), additional data that explained the qualitative findings were extracted, including the primary and main themes; secondary and subthemes; minor and unexpected themes; participant quotations; and any text labeled within the "results" or "findings" (ie, narrative) sections, including data-driven discoveries, judgments, or explanations the researchers offered about their phenomena [32,33]. All extracted data were cross-checked by a second reviewer for accuracy, and any discrepancies were resolved through discussion. Owing to the heterogeneity in study design, populations, and outcome measures for quantitative studies, a meta-analysis was inappropriate, and a narrative synthesis of quantitative study results was conducted.

Quality Assessment

The quality of the included studies was assessed using the Mixed Methods Appraisal Tool (MMAT) [34]. Each study's methodology was evaluated against 5 criteria (*Yes*, *No*, or *Can't tell*) that differed between study designs (qualitative, quantitative randomized controlled trial, quantitative nonrandomized, quantitative descriptive, and mixed methods). The included studies were divided equally among 5 reviewers (LW, YM, C Slade, JK, and OJC). An MMAT star rating was generated for each article (*Yes*=1 star; up to 5 stars in total), a method adapted from a recent systematic review by Freire et al [35], and the scores were cross-checked by a second reviewer. Discrepancies were discussed and resolved within the research team. No study was excluded from the review based on its MMAT score.

Data Synthesis

Data synthesis was conducted in 2 stages in accordance with our study aims.

Systematic Review: Narrative Synthesis

A narrative (qualitative) synthesis of quantitative studies was conducted to summarize and compare key findings [36]. We first developed a preliminary synthesis based on extracted data and then explored the relationships within and between studies to identify and explain any heterogeneity. Identified experience outcomes were inductively grouped together based on similarity (eg, ease of use and user-friendliness), and the group was given a descriptor (eg, usability) that would accurately reflect the experience outcomes within the group.

Qualitative Evidence Synthesis

The qualitative evidence synthesis [37] was conducted in 2 steps.

Step 1: Automated Text Analytics Using Machine Learning

Step 1 was undertaken using the text analytics tool *Leximancer* (version 4.5; Leximancer Pty Ltd) [38], an increasingly adopted [39] approach to qualitative analysis that is 74% effective at mapping complex concepts from matched qualitative data and >3 times faster than manual thematic analysis [40]. Leximancer applies an unsupervised machine learning algorithm and inbuilt thesaurus to uncover networks or patterns of word- and namelike terms in a body of text [41,42]. Leximancer then generates interconnections, structures, and patterns among terms to develop "concepts"—collections of words that are linked

XSL•FO
RenderX

together within the text—and group them into "themes"—concepts that are highly connected. The interrelationships between concepts and themes are visualized on a map. Advantages include expediting the early stages of qualitative analysis and providing a first impression of meaning within qualitative data that limits researcher bias.

After data extraction, qualitative data from the included articles were synthesized into 3 data sets—"themes" (primary, secondary, and minor), "quotes" (from participants), and "narrative" (any text under the "results" or "discussion" sections)—ready for separate Leximancer analysis. We chose to analyze each data set separately to account for any significant (but unknown at the time) heterogeneity across the studies. In total, 3 researchers were allocated 1 data set each using Leximancer to create an initial concept map without altering any settings. Initial concepts were reviewed for meaning, and redundant conversational words were removed where appropriate (eg, *study*, *doing*, and *participants*). Concept variations of EMR or electronic health record were removed as this was labeled as the independent variable and the target context already under analysis. Concept variations (eg, *patient* and *patients*) were merged where necessary. All other software settings were kept as the default values.

Step 2: Researcher-Led Thematic Analysis

The preliminary themes and concepts identified via text analytics underwent validation and researcher-led thematic analysis in accordance with a modified version of the method by Thomas and Harden [33]. First, the top 5 Leximancer-identified concepts (eg, "patient") were identified and connected with their 2 most related concepts (eg, "patient" AND "documentation" or "patient" AND "time") to create a concept grouping. In total, 3 researchers (OJC, LW, and JK) validated each Leximancer concept grouping by extracting relevant text and generating a preliminary interpretation of the meaning of each concept grouping, which was cross-checked between researchers. Researchers (OJC, LW, JK, and BG) then worked collaboratively across all 3 data sets to conduct a rapid thematic analysis using a cluster and name technique to generate a working thematic framework [43]. Through an iterative and interpretive process, researchers then grouped similar concepts into parent themes. Discrepancies were resolved through discussion until the final themes were decided and approved by consensus.

Results

Identification of Included Articles

In total, 2059 studies were identified from the first search (V1), and an additional 462 studies were identified from the second search (V2; Figure 2). Following duplicate removal and title and abstract screening, a total of 109 studies (V1: n=84, 77.1%; V2: n=25, 22.9%) remained and underwent full-text review. Of these 109 studies, 61 (56%) met our inclusion criteria and were included in this review, comprising quantitative (n=39, 64%), qualitative (n=15, 25%), and mixed methods (n=7, 11%) designs.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the systematic review and qualitative evidence synthesis. DH: digital hospital; EMR: electronic medical record.

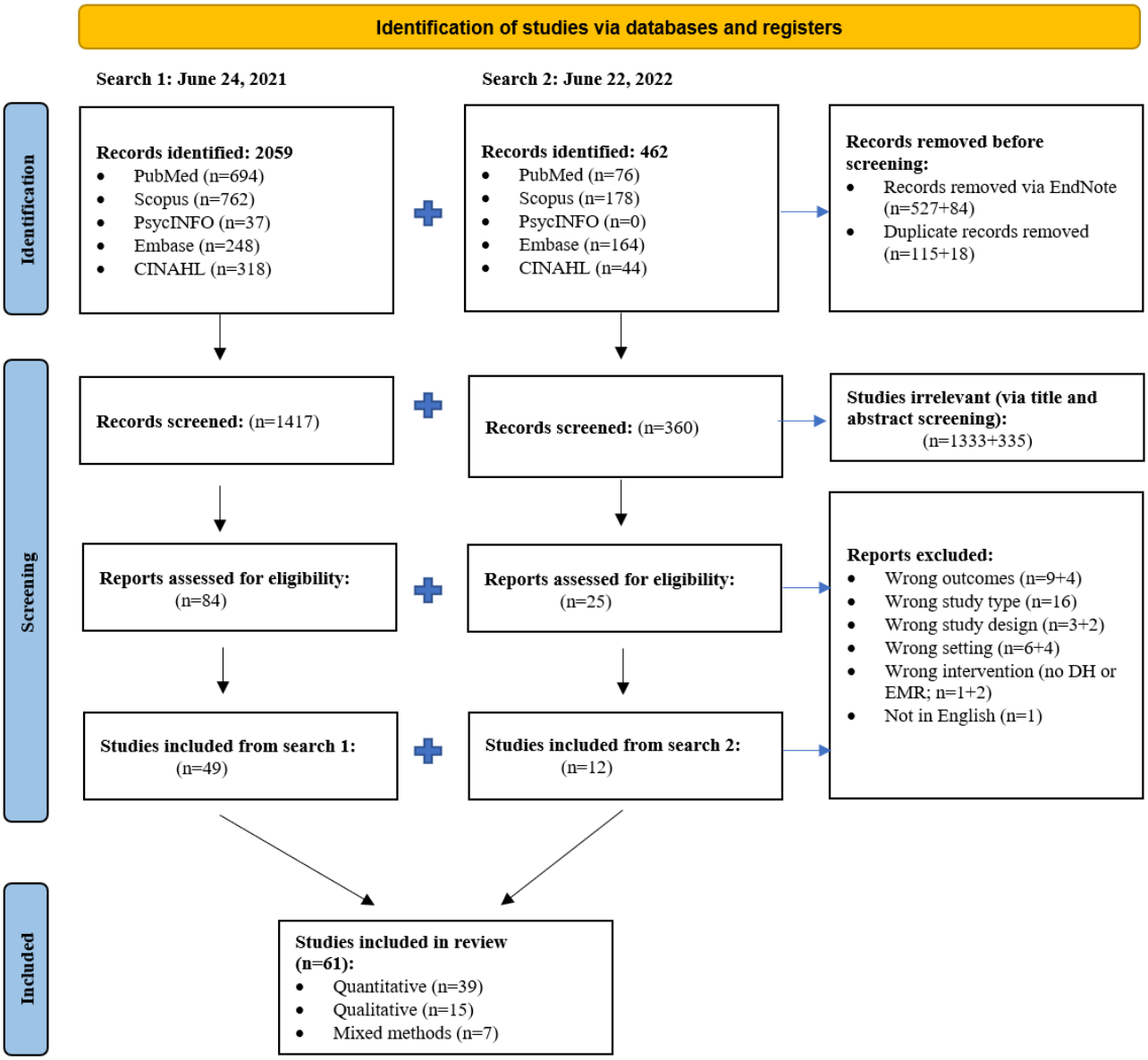

Quality Assessment

In total, 52% (32/61) of the included studies met all 5 MMAT quality criteria (Multimedia Appendix 5 [44-104]). An additional 20% (12/61) of the studies met 4 out of 5 of the quality criteria. Only 7% (4/61) of the studies met ≤2 of the 5 quality criteria, of which 75% (3/4) were mixed methods studies and 25% (1/4) had a quantitative descriptive design. For these studies, a score of ≤2 indicated inadequate sampling, and in the case of the mixed methods studies, the integration of and inconsistencies between quantitative and qualitative elements were not adequately described.

Characteristics of the Included Studies

Study Design

Of the 61 included articles (Multimedia Appendix 6 [44-104]), most (n=39, 64%) adopted quantitative methods to assess clinician and patient experiences. Most quantitative studies (31/39, 79%) conducted descriptive cross-sectional surveys to assess experience at one point. A total of 21% (8/39) of the studies assessed clinician and patient experience through quantitative nonrandomized methods in the pre– and post–EMR implementation periods. A minority of the included studies (15/61, 25%) qualitatively assessed experience through interviews, focus groups, and ethnographic observations. A total of 11% (7/61) of the studies used both quantitative and qualitative methods (mixed methods).

Setting

The most common country of study was the United States (21/61, 34%), followed by Australia (6/61, 10%), Saudi Arabia (5/61, 8%), and Canada (4/61, 7%). More than half (32/61, 52%) of the included studies were published after 2018. The settings included diversity across large tertiary academic hospitals and private hospitals in rural and metropolitan settings. The number of participating hospitals ranged from 1 to 2836.

XSL•FO
RenderX

Participants

In total, 90% (55/61) of the studies investigated the clinician experience using quantitative, qualitative, or mixed methods. Within the clinician experience group, 35% (19/55) of the studies included all EMR users, followed by nursing staff only (17/55, 31%) and physicians only (15/55, 27%). Study participation ranged from 8 to 3610 across the clinician experience studies. Only 16% (10/61) of the included studies investigated the patient experience. In total, 60% (6/10) of the studies focused exclusively on the patient experience with EMR, and 40% (4/10) included perspectives from both stakeholder groups. Patient participant counts ranged from 11 to 34,425.

Quantitative Results: Clinician and Patient Experience in Digital Hospitals

Overview

Table 3 reports the outcome measures of the digital hospital experience identified in the studies with quantitative components (46/61, 75%; 39/46, 85% quantitative and 7/46, 15% mixed methods).

XSL•FO
RenderX

Table 3. Measures of the digital hospital experience identified in the included studies (quantitative and mixed methods studies; n=46).

Experience outcome	Studies, n (%)	References
Patient experience (n=9; n=7 quantitative and n=2 mixed methods)		
Patient satisfaction	6 (67)	• Kazley et al [48] • Hu et al [84] • Jarvis et al [51] • Monturo et al [102] • Tian et al [96] • Burridge et al [69]
Recommended the hospital	2 (22)	• Kazley et al [48] • Hu et al [84]
Good discharge information	3 (33)	• Kazley et al [48] • Hu et al [84] • Burridge et al [69]
Provider-patient interaction (patient perspective)	3 (33)	• Migdal et al [57] • Monturo et al [102] • Burridge et al [69]
Clinician experience (n=41; n=34 quantitative and n=7 mixed methods)		
Overall EMR[a] satisfaction	24 (59)	• Tajirian et al [88] • Schwarz et al [87] • Schopf et al [77] • Williams et al [79] • Alobo et al [80] • Bani - issa et al [62] • Alharthi et al [54] • Harmon et al [59] • Kaipio et al [85] • Kutney-Lee et al [76] • Lakbala and Dindarloo [56] • Tilahun and Fritz [61] • Tubaishat [72] • Abu Raddaha [70] • Top and Gider [49] • Alsohime et al [73] • Al-Mujaini et al [46] • Shaker et al [60] • Claret et al [47] • Al Otaybi et al [98] • Cho et al [89] • Czernik et al [101] • Jedwab et al [97] • Welchen et al [104]
Usability	19 (46)	• De Groot et al [82] • Schwarz et al [87] • Schopf et al [77] • Aldosari et al [68] • Eden et al [83] • Alharthi et al [54] • Top et al [53] • Kutney-Lee et al [76] • Bossen et al [50] • Kaipio et al [66] • Tubaishat [72] • Abu Raddaha [70] • Top and Gider [49] • Strudwick et al [71] • Arikan et al [99] • Al Otaybi et al [98] • Cho et al [89] • Lloyd et al [93] • Welchen et al [104]

XSL•FO
RenderX

Experience outcome	Studies, n (%)	References
Adaptation	14 (34)	• Schopf et al [77] • Schenk et al [64] • Hung et al [75] • Lakbala and Dindarloo [56] • Bossen et al [50] • Kaipio et al [66] • Abu Raddaha [70] • Strudwick et al [71] • Shaker et al [60] • Claret et al [47] • Arikan et al [99] • Al Otaybi et al [98] • Czernik et al [101] • Lloyd et al [93]
Data accessibility	18 (44)	• De Groot et al [82] • Schwarz et al [87] • Aldosari et al [68] • Alobo et al [80] • Eden et al [83] • Alharthi et al [54] • Top and Gider [49] • Hung et al [75] • Kutney-Lee et al [76] • Bossen et al [50] • Kaipio et al [66] • Abu Raddaha [70] • Top et al [53] • Strudwick et al [71] • Arikan et al [99] • Al Otaybi et al [98] • Lloyd et al [93] • Welchen et al [104]
Provider-patient interaction (clinician perspective)	3 (7)	• Ratanawongsa et al [67] • Al Otaybi et al [98] • Czernik et al [101]
Workload and burnout	19 (46)	• Tajirian et al [88] • Schopf et al [77] • Kutney-Lee et al [92] • Alobo et al [80] • Eden et al [83] • Schenk et al [64] • Harmon et al [59] • Lakbala and Dindarloo [56] • Tilahun and Fritz [61] • Bossen et al [50] • Strudwick et al [71] • Shaker et al [60] • Claret et al [47] • Arikan et al [99] • Al Otaybi et al [98] • Czernik et al [101] • Heponiemi et al [90] • Jedwab et al [97] • Luyten and Marneffe [94]
Patient safety	14 (34)	

XSL•FO
RenderX

Experience outcome	Studies, n (%)	References
		• Tajirian et al [88] • Schwarz et al [87] • Williams et al [79] • Alobo et al [80] • Top et al [53] • Hung et al [75] • Kaipio et al [85] • Lakbala and Dindarloo [56] • Bossen et al [50] • Kaipio et al [66] • Alsohime et al [73] • Al Otaybi et al [98] • Lloyd et al [93] • Luyten and Marneffe [94]
Delivery of care	12 (29)	• De Groot et al [82] • Top et al [53] • Harmon et al [59] • Kutney-Lee et al [76] • Tilahun and Fritz [61] • Abu Raddaha [70] • Top and Gider [49] • Alsohime et al [73] • Al-Mujaini et al [46] • Claret et al [47] • Al Otaybi et al [98] • Czernik et al [101]

[a]EMR: electronic medical record.

Patient Experience in a Digital Hospital

Of the 9 quantitative or mixed methods studies reporting the patient perspective, 7 (78%) [48,51,57,69,84,96,102] used different survey methods (eg, the Hospital Consumer Assessment of Healthcare Providers and Systems [HCAHPS] survey) to quantify "patient experience" using various satisfaction metrics. Of these studies, 57% (4/7) reported a positive association [48,57,69,84] between EMR and patient satisfaction scores and 43% (3/7) reported no substantial change [51,96,102]. In total, 22% (2/9) of the studies [67,92] used hospital outcomes rather than patient feedback, which did not meet our definition of "experience."

A total of 33% (3/9) of the studies [57,96,102] reported patient experience before and after EMR implementation (or transition between EMR systems). Tian et al [96] surveyed 34,425 patients using the standardized HCAHPS survey and found a significant decreasing trend in patient experience scores for the 6 months after implementation followed by a return to baseline, with no significant changes overall. Monturo et al [102] surveyed 55 patients and found no significant changes in overall patient satisfaction.

Of the 9 studies, 3 (33%) cross-sectional studies [48,51,84] compared patient experiences in hospitals with and without an advanced EMR. Hu et al [84], using the HCAHPS survey at 1006 hospitals, and Kazley et al [48], using Hospital Compare data at 2836 hospitals, both found a positive association between EMR adoption and overall hospital rating and discharge information. Jarvis et al [51] found no significant difference in HCAHPS scores in advanced EMR versus non–advanced EMR hospitals.

Clinician Satisfaction With the EMR

Of the 41 quantitative studies that investigated the clinician experience, 24 (59%) included an *overall EMR satisfaction* metric, and 71% (17/24) of these studies reported a positive sentiment [47,49,59,60,62,70,72,73,76,79,80,85,87-89,98,101]. For instance, Kutney-Lee et al [76] used the registered nurse forecasting study (RN4CAST-US) nursing survey with 12,377 nurses across 353 hospitals and found a 74.9% "satisfaction with current EMR." In total, 25% (6/24) of the studies [64,68,72,75,89,94] used features of the technology acceptance model, and "perceived usefulness" or "perceived value" was equated to *overall satisfaction*. Evidence of increasing satisfaction with increased digitization was found in a study that stratified results by level of EMR adoption, including groups for basic EMR (71.3% satisfaction reported) and comprehensive EMR (78.4% satisfaction) [76].

Of the 24 studies that reported *overall EMR satisfaction* as an outcome of clinician experience, 7 (29%) [46,54,56,61,77,97,104] reported negative sentiment with the EMR. For instance, Tilahun and Fritz [61] surveyed 406 clinicians and found that 64.4% were dissatisfied with the use of the EMR system; however, only 22.8% strongly disagreed with the following statement: "I prefer EMR than the paper record." One study found that only 15.6% of respondents (n=141 physicians) felt that the EMR was an "effective tool" [46], and another found that only 38.9% of users (n=262 nurses and physicians using the National Usability-Focused Health Information System Scale) rated the EMR system as "high quality" [104].

XSL•FO
RenderX

Usability of the EMR by Clinicians

Of the 41 studies that included the EMR user perspective, 19 (46%) reported a *usability* metric, and 11 (58%) of these reported a positive sentiment [50,54,68,70,71,72,76,82,83,89,99]. Survey components that investigated usability used outcomes such as "ease of use," "user friendly," and "technical quality."

Positive sentiment was considered to be >3.5/5 on a 5-point Likert scale or >50% agreement with usability statements. Of those using a Likert scale, statements such as "the health record I am working with is user-friendly" scored 3.62 (n=667 nurses) [82], and "perceived ease of use" scored 3.7 (n=1539 nurses in 15 hospitals) [72] and 3.78 (n=223 nurses) [89]. Aldosari et al [68] surveyed 153 nurses and found a 79.7% agreement with the following statement—"It is easy to use the EMR"—and 70.5% agreement with the following statement: "I find the EMR system interface to be user friendly."

A negative sentiment regarding usability was found in 42% (8/19) of the studies [49,53,66,77,87,93,98,104]. In the study using the National Usability-Focused Health Information System Scale (n=3013 physicians), most participants (60.15%) disagreed with the following statement: "routine tasks can be performed in a straightforward manner without the need for extra steps using the systems" [66]. Comparatively fewer nurses (n=3560) in the same study disagreed with this statement [66].

Adaptation to New Systems

A total of 30% (14/41) of the studies discussed the experience of adapting existing workflows to integrate the new digital interface and transitioning to a digital environment on the wards [47,50,56,60,64,66,70,71,75,77,93,98,99,101]. Generally, the *adaptation* outcome had a negative sentiment from EMR users. One survey of 285 nurses 8 to 13 months after EMR implementation found that users felt that the EMR provided a "holistic view of the patient, but fragmentation and complexity introduce workflow challenges" [64]. Another study found that 35.1% of physicians (n=317) agreed with the following statement: "EMR does not disrupt workflow" [60]. A third study found that 48.7% of physicians (n=3013) and 62.3% of nurses (n=3560) disagreed with the following statement: "learning the EHR did not require a lot of training" [66].

Data Accessibility and Clinician-Patient Interaction

Data accessibility in digital hospitals was reported in 39% (18/41) of the quantitative studies [49,50,53,54,66,68,70,71,75,76,80,82,83,87,93,98,99,104]. Much of the sentiment was positive as clinicians agreed that the EMR allowed users to access information when and where they needed it. One survey of 153 nurses in a Saudi Arabian hospital found that 85.6% of respondents agreed with the following statement—"I have access to the information where I need it"—and 83.6% agreed with the following statement: "I have access to the information when I need it" [68]. One large cross-sectional study of 2684 clinicians found a 50.3% agreement with the following statement: "EMR provides precise information I need" [98]. Neutral sentiment was indicated in one study by an average of 3.5 (5-point Likert scale) in response to the following statement—"it is easy to find the information I need"—from 244 clinicians 2 months after EMR implementation [50]. The survey by Lloyd et al [93] found that 49% of physicians (n=224) and 59.4% of nurses (n=72) agreed that "it is easy to obtain necessary patient information using the EMR system."

A total of 7% (3/41) [67,98,101] of the studies reported on the impact of EMR on clinician-patient interaction, with all 3 studies agreeing that the EMR reduced this communication. Czernik et al [101] found that 39% of 126 physicians agreed (7-point Likert scale) with the statement that EMR causes "lack of proper patient-doctor communication." A total of 43% (3/7) [57,69,102] of patient experience studies reported on the impact of EMR on patient-provider interaction. Migdal et al [57] focused more specifically on physician-patient communication with their patient participants using a CICARE survey (17-question Likert scale) designed by the University of California, Los Angeles, Health system to assess resident physician performance. Of the 3417 patient surveys, Migdal et al [57] found that 9 of 16 relevant questions had statistically significant improvements after EMR implementation, suggesting improvement in communication between patients and providers after EMR implementation.

Workload and Burnout

Many studies (19/41, 46%) reported on the impact of EMR on clinical workload, including symptoms of burnout and subjective productivity. One cross-sectional study in Canada surveyed 208 physicians and found that 68.2% of respondents felt that the EMR "added to daily frustration"; 24.5% of respondents had one or more symptoms of burnout; and, of those with burnout, nearly 75% "identified EMR as contributor to burnout symptoms" [88]. Another study across 343 hospitals including 12,004 nurses compared EMR usability (as per the RN4CAST-US usability survey) with symptoms of burnout and found that lower EMR usability scores were associated with higher odds of burnout (odds ratio 1.41, 95% CI 1.21-1.64) [92]. Often, studies assessed workload in terms of productivity. In 5 low-resource hospitals 3 years after EMR implementation (n=405 physicians and nurses), 82.4% of physicians disagreed that "EMR improves productivity," whereas 61% of nurses agreed [61].

Patient Safety and Delivery of Care

There was a mixed sentiment across 34% (14/41) of the quantitative studies [50,53,56,66,73,75,79,80,85,87,88,93,94,98], which included survey items on the impact of the EMR on patient safety. In total, 43% (6/14) of these studies [53,56,66,73,93,98] included survey items about EMR preventing errors in patient care, especially mistakes associated with medications. One study in a large specialist hospital in Nigeria (n=35 health care workers) found that the EMR made clinicians "more prone to errors" [80]. Similarly, Kaipio et al [66] found that less than half of the surveyed physicians (44.7%) and nurses (40.2%) agreed with the following statement: "IT systems help in preventing errors and mistakes associated with medication." Conversely, Al Otaybi et al [98] (n=2684 health care workers) found that only 15.5% agreed with the following statement: "EMR increases the risk of making errors." One study investigated the change

XSL•FO
RenderX

in user experience over time and found that agreement with the statement that the EMR "improves prevention in errors and mistakes associated with medications" increased by 13% from 2010 to 2014 [85].

There was also a mixed sentiment across 29% (12/41) of the studies, which included outcomes on the impact of the EMR on clinicians' ability to deliver care to their patients. In a large study of >12,000 US nurses, more than half (55.4%) reported agreement with the statement that the EMR "systems interfere with the provision of care" [76], and in a smaller study, 84.2% of participants disagreed with the statement that the EMR "system has positive impact on quality of care" [61]. Conversely, in the Netherlands, nurses were more likely to agree with the following statement: "the information in the health records supports my activities during the provision of care" [82].

Qualitative Results: Clinician and Patient Experience in Digital Hospitals

A total of 18 studies (n=14, 78% qualitative and n=4, 22% mixed methods) with qualitative components were included in the qualitative evidence synthesis. Only 7% (1/15) of the qualitative studies in our review explored the patient experience in a digital hospital [52]; however, this study was excluded from the qualitative evidence synthesis and is reported narratively in the following paragraph. A total of 29% (2/7) of the mixed methods studies were also excluded for lacking direct participant quotes as per the exclusion criteria [50,75]; however, the quantitative results are reported in the previous sections.

Strauss [52] interviewed 11 patients about the dynamics with their nurses and the EMR. Similar to the qualitative evidence synthesis, participants described a positive perception of the EMR when the nurses acknowledged the participants before using the electronic device; however, many "expressed concerns [for] the privacy of their health record information." Interestingly, participants' expectations of the "clinical knowledge and competency of the nurse, within the technological arena, have increased with the implementation of the [EMR]."

Qualitative Evidence Synthesis: Clinician Experience Only

Step 1: Automated Text Analytics Using Leximancer

Multimedia Appendix 7 presents the results of the automated text analytics using Leximancer. Figures S1 to S3 in Multimedia Appendix 7 present the intertopic concept maps derived from the themes, quotes, and narrative qualitative data, respectively.

Table S1 in Multimedia Appendix 7 compares the top 5 concepts and their 2 most related concepts identified by Leximancer across each qualitative group. Owing to the relative homogeneity in the top 5 concepts identified among the themes, quotes, and narrative qualitative data, it was decided to perform researcher-led thematic analysis (step 2) collectively instead of individually for each data group. This decision was not a predetermined method and was made organically during data analysis.

Step 2: Researcher-Led Thematic Analysis

Overview

Textbox 1 presents 7 themes that describe the clinician experience in digital hospitals derived from the qualitative evidence synthesis (18/61, 30% of the studies).

Textbox 1. Final themes describing the clinician experience in digital hospitals from the qualitative evidence synthesis (n=18).

Theme 1

- Slow and inefficient digital documentation detracts from other clinical priorities.

Theme 2

- Inconsistent data quality and discoverability challenge clinician trust in making data-driven decisions.

Theme 3

- Digital technology creates new tensions that disrupt conventional health care relationships.

Theme 4

- Acceptance of digital hospitals is a value-based spectrum that changes over time.

Theme 5

- Clinicians value patient safety benefits while acknowledging concerns about new digital risks.

Theme 6

- Clinicians feel reliant on hybrid (digital and paper) workflows to maintain the standard of care.

Theme 7

- Clinicians worry about compromising patient data privacy to improve care efficiency.

XSL•FO
RenderX

Theme 1: Slow and Inefficient Digital Documentation Detracts From Other Clinical Priorities

Documentation was a time burden for clinicians with a slow and inefficient workflow [95] in which it was difficult to find information [95], there were "too many steps to accomplish simple tasks" [44], and users were required to re-enter the same data repeatedly [63].

Clinicians felt challenged by the requirement for accurate and complete documentation during the provision of patient care. Staff reported wanting to provide care but needing to complete medical records:

> *A lot of the time we're having to say [to patients], "Oh look, I'll have to come back to you, I've got to do my documentation." [81]*

The study by Schenk et al [64] concluded that "EHR implementation was disruptive to nursing care and adversely influenced nursing attitudes," reporting "too many steps to find and chart information, information that is fragmented and overly complex, leading to workflow challenges and interruptions of care." Shortcuts and quick orders save time [95] and were used as workarounds to improve efficiency.

Theme 2: Inconsistent Data Quality and Discoverability Challenge Trust in Decisions

Clinicians found that data quality improved in digital hospitals. Studies highlighted various benefits of digitization, including improved documentation of data [78], efficient display of information [95], improved data completeness [78], and improved documentation readability [63]. This was summarized by the following participant quote:

> *...the availability of data in the EHR is a good thing. [101]*

"Note bloat" was reported as a theme regarding difficulties finding information in the digital system [95]. Discoverability challenged clinicians and negatively affected their trust in making data-driven decisions. It was difficult to find information [95] and easy to miss information [95], as described by one participant:

> *...you've got to look through 100 documents to find the information you are looking for. [65]*

Inefficiencies in the EMR design may lead to inappropriate care:

> *A wrong decision happens based on the missing information. [45]*

Theme 3: Digital Technology Creates New Tensions That Disrupt Conventional Health Care Relationships

Reduced patient contact [55] from EMRs "inserted" between the patient and clinician deteriorated the personal relationship [44]. There was the potential to lose focus on the patient, undermine rapport [86], and communicate with the computer in lieu of direct bedside patient communication [81]. The effect of digital documentation on trust in the psychiatrist-patient relationship was noted, and required open communication between the psychiatrist and patient to promote transparency about what was documented [100].

Managing disrupted communication [81] to preserve the patient's personhood in the digital environment [81] created notable tension [69]. For example, the EMR can improve information accuracy [44] by viewing a patient history [103] or using a computer to facilitate conversation [86]. Preparing notes before consultations, minimizing screen use and explaining computer use, taking paper notes, sharing screens with patients, and viewing results digitally together "seems to counterbalance the negative effects of computer use" [86].

Interprofessional communication between clinicians was affected by digital notes "only when the data entered by different roles in the healthcare system are accurate, the clinicians can make timely and correct decisions" [103].

Theme 4: Acceptance of Digital Hospitals Is a Value-Based Spectrum That Changes Over Time

Individual beliefs about digital hospitals trended negatively (eg, threats to patient safety, waste of human resources, and perceived inefficiency), but objective experiences trended positively (eg, access to records, optimized treatment, efficiency and health system coordination [45], cost savings, improved productivity, and quality of care). Acceptance fluctuated between perception and reality—a mismatch between "work as imagined" and "work as done."

Negative experiences with "work as imagined" were related to perceived inefficiencies [45]. Digital hospitals created their own unique time pressures with improved productivity in some cases [45], with participants viewing technology as "both time saving and time consuming" [74].

Positive experiences with "work as done" coincided with a longer time since implementation, when adoption had progressed, disruption to their work processes had eased, and workflows had been integrated [58]. Initially, clinicians reported having a negative first impression of the EMR, especially perceived complexity and ease of use [91].

Theme 5: Clinicians Value Patient Safety Benefits While Acknowledging Concerns About New Digital Risks

Digital hospitals generate patient safety benefits while creating new ways to make errors and increase risks. New errors may negatively affect patient safety. These include wrong patient errors, alert fatigue, inappropriate alerts, data entry errors, technical problems [78], field auto-population or auto-refresh errors, and the absence of aids for dose calculations [95]. Concerns regarding clinician overreliance on the system, ignoring correct alerts, and prioritizing system compliance over clinical accuracy were raised in some cases [45].

Reduced medication errors were the primary reported patient safety benefit, including awareness of known adverse reactions to medications [103] and improved legibility [78]. Clinical decision support enabled by the EMR was seen as a safety benefit to alert staff for prompt intervention [78]:

> *There is a pop up in the system which questions are you supposed to give that medication right now? [103]*

The safety benefits of sharing medical information [100] and applying regulatory frameworks to the workflow [100] were

XSL•FO
RenderX

noted. Entering clinical data into structured mandatory fields and managerial-level review improved thoroughness and was considered to enhance patient safety [78].

Theme 6: Clinicians Feel Reliant on Hybrid (Digital and Paper) Workflows to Maintain Standard of Care

Digital transformation caused workflow disruption. New digital workflows were time consuming, with a one-size-only user interface and limited ability to adapt to individual patient characteristics or change information once documented [86]. The digital interface was insufficient to meet clinical workflow needs. Often, EMR workflows were supplemented with paper workflows [81]. Paper enabled total customization to fit with workflow conventions and was a "cognitive support" to supplement personal workflows to plan and prioritize in a flexible, convenient, comfortable, and trusted way [58]:

> *I go to my [paper] notes and I make little boxes, and if I do those tasks I tick them. [81]*

Theme 7: Clinicians Worry About Compromising Patient Data Privacy to Improve Care Efficiency

The improved documentation captured by the EMR created a perceived privacy risk for the patient. Perceptions on patient preferences to protect the disclosure of medical information involved elements such as diagnoses [100], mental health–related stigma [100], and patient distrust of the EMR system or its users [86]. Reported strategies enlisted by health care professionals were to only include clinical documentation that was general in nature, avoid labeling (eg, "mood disorder" instead of "depression") [100], prioritize based on clinical relevance [100], or limit types of information to critical information related to medication that is considered essential knowledge [86].

Discussion

Principal Findings and Comparison With Prior Work

Our findings revealed mixed and complex clinician and patient experiences of digital hospitals (Figure 3). Generally, clinicians reported positive overall satisfaction (17/24, 71% of the studies) with digital hospitals in quantitative measures; however, there were negative experiences for clinicians reported in qualitative studies, including compromised clinician-patient interactions, inefficient data workflows, and patient data privacy concerns. For example, acceptance of digital hospitals fluctuated over time and trended negatively if grounded in individual perceptions and beliefs (ie, "work as imagined") yet trended positively if based on objective measures of practice (ie, "work as done") [105]. These inconsistencies are likely reflective of the various contextual factors that influence experience, such as intervention design or stage of implementation. It is likely that the quantitative finding of mixed usability of EMRs explains clinician reliance on hybrid (digital and paper) workflows revealed in the qualitative evidence synthesis as clinicians seek to maintain their clinical workflow standard and validate data using additional sources of truth, such as paper [10].

Figure 3. Summary of key findings on the clinician and patient experience of digital hospitals. +: positive; ~: mixed; −: negative; EMR: electronic medical record; N/A: not available.

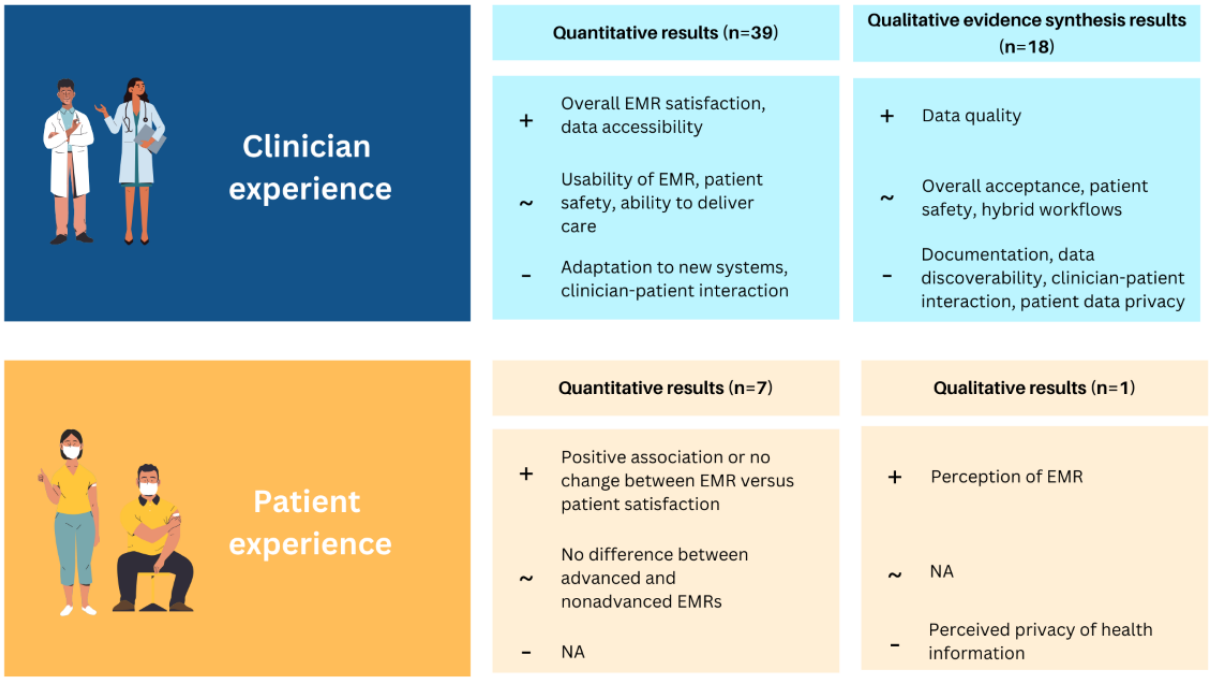

The effect of digital hospitals on clinician-reported patient safety and clinician ability to deliver care were mixed; there was acknowledgment that digitization primarily reduces medication error risk but creates new risks driven by questionable data quality. In the small proportion of studies that explored patient experience (10/61, 16%), there was weak evidence supporting a positive association between digital hospitals and patient satisfaction scores. To our knowledge, this is the first review to systematically evaluate clinician and patient experience in digital hospitals and use qualitative evidence synthesis with

XSL•FO
RenderX

machine learning (Leximancer) to consolidate identified themes in previous qualitative research into an empirical "umbrella" view of digital hospital experience.

Previous studies have evaluated the adoption of health ITs and identified enablers of and barriers to the routine use of EMRs in practice. The perceived value of the EMR to clinical workflows and data accessibility are key adoption facilitators, whereas cost and time consumption are barriers to adoption [106]. Our findings revealed that clinicians reported high satisfaction with digital hospitals and positively viewed data accessibility in quantitative measures; however, our qualitative evidence synthesis revealed themes of frustration with slow digital workflows and inconsistent data discoverability. Evidence of EMR adoption in low-income countries also highlights clinician perception of the EMR as a key facilitator and interoperability and clinician burnout as barriers, similar to our findings on the impact of EMR on workload and burnout symptoms [107].

Our review found that the patient experience of digital hospitals was reported disproportionately less frequently than the clinician experience. Evidence of patient satisfaction with EMRs was systematically reviewed in 2013, and it was found that a small number of studies (n=8) indicated positive patient satisfaction with the EMR in mixed settings across primary care, emergency, and outpatient departments [13]. This evidence is consistent with our findings of a positive or neutral association but remains grounded in cross-sectional methods that warrant rigorous trial evaluation. Beyond the clinician-facing EMR, patient-centered digital health records have emerged as a mechanism to engage and empower consumers living with chronic conditions. Patient portals within EMRs can contribute positively to health care quality and safety by improving medication adherence and clinician-patient communication [24] and have been shown to improve patient care navigation and disease knowledge without adverse effects and with high patient satisfaction [108]. Measuring satisfaction with digital hospitals using a simple quantitative scale is unlikely to capture the complexity and heterogeneity of digital hospital environments. There was dissonance in clinician perspective on data accessibility (or discoverability) between quantitative and qualitative studies. Clinicians reported objective satisfaction with data accessibility and positive attitudes toward data quality; however, they were dissatisfied with the inefficient workflows required to generate high-quality data (ie, input) and the ability to leverage these data for secondary use (ie, output).

Although patient experience is consistently positively associated with patient safety, clinical effectiveness, and self-rated and objectively measured health outcomes [109], there remains a paucity of empirical research that directly investigates patient experience in digital hospital environments. Our results revealed that, in a small number of studies, patient satisfaction scores were positively associated with digital hospitals (4/61, 7%) or remained unchanged (3/61, 5%). Patients and clinicians shared positive overall impressions of EMRs and negative attitudes toward data privacy in a digital hospital environment; however, this result must be considered in the disproportionate context of patient and clinician experience data reported in studies (10/61, 16% vs 55/61, 90%, respectively).

Patient experience in a digital hospital was sometimes inferred from surrogate measures, including hospital recommendations and discharge information quality, that do not capture the complexity of personal experience. For patients, crude quantitative measurement of satisfaction with digital hospitals neglects the complex nature of the digital patient experience [110] or emerging consumer digital health themes (eg, ethical implications, security, choice, privacy, transparency, accuracy, user-friendliness, and equity of access) [111,112]. The authors support the call by Viitanen et al [23] to develop a framework to describe the different aspects of patient experience and correlate them with appropriate methods for studying patient experience in this context.

The effect of digital hospitals on the clinician-patient relationship was consistently reported by clinicians as a negative outcome of digitization. These results align with the well-researched relationship between EMRs and the clinician-patient dynamic, with evidence supporting negative communication outcomes (eg, rapport, quality of interaction, and time) [113] from a clinician perspective. Patient perceptions of clinician-patient communication when using EMRs are relatively stable in previous systematic reviews despite objective studies describing potentially negative (eg, interrupted speech) and positive (eg, facilitating questions) effects [16]. Patient portals within EMRs can improve clinician-patient communication and should be considered a necessary infrastructure for health services implementing EMRs to mitigate potential negative effects [24].

Implications for Practice

"Improved patient experience" and "improved clinician experience" are 2 quadrants of the Quadruple Aim of health care that warrant significant health service investment amidst the widespread digital transformation of health care. Our findings highlight the need to address pervasive barriers to positive clinician and patient experiences in digital hospitals. To tackle the issues of clinician-patient interaction, inefficient documentation, workload, and burnout identified in this review, health services can invest in feasible and cost-effective solutions such as clinical education and training that are tailored to each clinical discipline as each discipline has unique EMR needs [10]. Prioritizing investment in patient-facing digital solutions such as patient portals can democratize clinical knowledge and empower patients on their unique health care journey [24].

Investment in optimizing EMR infrastructure will build a strong foundation for new clinical applications in descriptive, predictive (ie, artificial intelligence [AI]), and prescriptive (ie, causal AI and decision support) analytics that can benefit clinical workflows and patient outcomes [114]. COVID-19 initiated the rapid adoption of virtual health care [115], and clinical applications of AI are rapidly emerging as the future primary disruptors of health care [116]. Global health services are building machinery to shift from reactive (treat-manage) to proactive (predict-prevent) models of care, with evidence of success for acute clinical problems such as reducing mortality rate and organ failure in the early identification of sepsis [117]. Stakeholder perspectives on implementing clinical AI have been recently consolidated in a qualitative evidence synthesis [118];

XSL•FO
RenderX

however, similarly to our review, patients, carers, and consumers were an underrepresented group compared with clinicians (11.4% vs 70% of data, respectively).

Our approach of using AI (machine learning) via Leximancer to perform the qualitative evidence synthesis is a novel approach compared with recent manual evidence synthesis methods [118-120]. The use of semiautomated content analysis tools such as Leximancer can accelerate progress toward a learning health system. These tools offer an accelerated pipeline for analyzing "big" qualitative data that suffer from traditional yet burdensome manual analytic workflows. Key health care use cases of applying digital tools to routine analytical workflows are patient-reported experience measures [121], unstructured clinical notes in EMRs [122], and social media [123]. Natural language processing and machine learning have been tested to analyze free-text comments from patient experience feedback [124]. Applications of Leximancer can be pushed by investigating how its algorithms can be used in real-world health care to drive continuous cycles of quality improvement with greater speed and efficiency compared with a manual control. Leximancer offers an impartial starting point for content analysis by automating the identification of key concepts and themes that warrant further qualitative refinement by the research team.

Limitations

The scope of this review was limited to experiences in a digital hospital environment, and thus, the experiences of specific digital systems (eg, telehealth and patient portals) were not considered. The complex, interacting factors that influence experience; the stage of digital hospital implementation; and the differences among settings were not explored in this review and offer important foci for future research. By not including gray literature and articles not published in English, our search strategy may have missed informal evaluations of clinician and patient experience of digital hospitals (eg, within health service annual reports) and geographical variation in digital hospital evaluations. The heterogeneity of digital health environments is reflected in the heterogeneity of studies included in this review, meaning that it is difficult to draw definitive conclusions agnostic to time and place. One limitation of using Leximancer for the qualitative evidence synthesis is that it does not automatically identify emotive concepts; these are identified by researchers when interpreting the results. Although Leximancer reduces the potential for human bias when compared with manual analysis, researcher interpretation of Leximancer results remains a gateway for introducing bias [40]. A qualitative evidence synthesis for patient experience studies was not possible as we only identified one eligible study. The patient experience results should be interpreted with caution because of the relatively limited patient experience data in studies (10/61, 16%) compared with clinician experience data in studies (55/61, 90%).

Conclusions

The clinician experience of digital hospitals appeared positive according to high-level indicators (eg, overall satisfaction and data accessibility); however, the qualitative evidence synthesis revealed substantive tensions between digital hospitals and overall experience, such as weakening clinician-patient interaction, change burden, and inefficient data workflows. There is insufficient evidence to draw a definitive conclusion on the patient experience of digital hospitals, but quantitative indications of satisfaction appear positive or agnostic to digitization. Future research must prioritize investigating the patient experience in digital hospitals and measuring the link between exposure (digital hospital) and outcome (experience) in carefully designed pragmatic trials. Areas of interest include examining the interacting factors that influence experience, the stage of digital hospital implementation, and the differences among settings. Equitable investigation of the patient (including pediatric patients) and clinician digital hospital experience must be prioritized in future research. Worldwide, as digital health becomes inseparable from hospitals and general health care, understanding how to optimize the clinician and patient experience in digital hospital environments will be critical to achieving the Quadruple Aim of (digital) health care.

Acknowledgments

OJC and LW are funded by the Digital Health Cooperative Research Centre (DHCRC). DHCRC is funded under the Commonwealth Government Cooperative Research Centres Program, Australia. OJC was also funded by the Australian Research Council Linkage Program Grant (ARC LP170101154). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The data supporting the findings of this study are available from the corresponding author (OJC) upon reasonable request.

Authors' Contributions

OJC and YM conceptualized the study. YM designed the search strategy with input from OJC and C Slade. OJC and YM executed the search. OJC, YM, JK, and BG performed article screening. OJC, LW, JK, and BG performed data extraction and quality appraisal. OJC and LW validated the article screening and data extraction. C Slade assisted with designing the search strategy and article screening. OJC, LW, JK, and BG performed data analysis. OJC and LW wrote the first draft of the paper with input from JK and BG. C Sullivan and ABJ critically reviewed the manuscript and provided expert input. OJC, LW, and JK revised the manuscript. All authors have read and approved the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 25 KB - jmir_v26i1e47715_app1.docx]

Multimedia Appendix 2
ENTREQ (Enhancing the Transparency in Reporting the Synthesis of Qualitative Research) guideline checklist.
[DOCX File , 16 KB - jmir_v26i1e47715_app2.docx]

Multimedia Appendix 3
Final search strategies for all included databases.
[DOCX File , 14 KB - jmir_v26i1e47715_app3.docx]

Multimedia Appendix 4
Data extraction template.
[DOCX File , 14 KB - jmir_v26i1e47715_app4.docx]

Multimedia Appendix 5
Mixed Methods Appraisal Tool results.
[DOCX File , 17 KB - jmir_v26i1e47715_app5.docx]

Multimedia Appendix 6
Characteristics of the studies included in the systematic review and qualitative evidence synthesis on clinician and patient experience in digital hospitals (N=61).
[DOCX File , 23 KB - jmir_v26i1e47715_app6.docx]

Multimedia Appendix 7
Leximancer results—qualitative evidence synthesis.
[DOCX File , 1182 KB - jmir_v26i1e47715_app7.docx]

References

1. Digital health and the trends healthcare investors are following. Healthcare Information and Management Systems Society. URL: https://www.himss.org/resources/digital-health-and-trends-healthcare-investors-are-following [accessed 2024-11-01]
2. Salway RJ, Silvestri D, Wei EK, Bouton M. Using information technology to improve COVID-19 care at New York City health + hospitals. Health Aff (Millwood) 2020 Sep 01;39(9):1601-1604. [doi: 10.1377/hlthaff.2020.00930] [Medline: 32673131]
3. Thomas EE, Haydon HM, Mehrotra A, Caffery LJ, Snoswell CL, Banbury A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2020 Sep 26;28(4):301-308. [doi: 10.1177/1357633x20960638]
4. Nguyen KH, Wright C, Simpson D, Woods L, Comans T, Sullivan C. Economic evaluation and analyses of hospital-based electronic medical records (EMRs): a scoping review of international literature. NPJ Digit Med 2022 Mar 08;5(1):29 [FREE Full text] [doi: 10.1038/s41746-022-00565-1] [Medline: 35260765]
5. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://tinyurl.com/58d488cu [accessed 2024-01-29]
6. Barnett A, Winning M, Canaris S, Cleary M, Staib A, Sullivan C. Digital transformation of hospital quality and safety: real-time data for real-time action. Aust Health Rev 2019 Jan;43(6):656-661. [doi: 10.1071/AH18125] [Medline: 30384880]
7. Sullivan C, Staib A, Ayre S, Daly M, Collins R, Draheim M, et al. Pioneering digital disruption: Australia's first integrated digital tertiary hospital. Med J Aust 2016 Nov 07;205(9):386-389. [doi: 10.5694/mja16.00476] [Medline: 27809727]
8. Sullivan C, Staib A. Digital disruption 'syndromes' in a hospital: important considerations for the quality and safety of patient care during rapid digital transformation. Aust Health Rev 2018 Jun;42(3):294-298. [doi: 10.1071/AH16294] [Medline: 28514640]
9. Robertson ST, Rosbergen IC, Burton-Jones A, Grimley RS, Brauer SG. The effect of the electronic health record on interprofessional practice: a systematic review. Appl Clin Inform 2022 May 01;13(3):541-559 [FREE Full text] [doi: 10.1055/s-0042-1748855] [Medline: 35649501]

XSL•FO
RenderX

10. Canfell OJ, Meshkat Y, Kodiyattu Z, Engstrom T, Chan W, Mifsud J, et al. Understanding the digital disruption of health care: an ethnographic study of real-time multidisciplinary clinical behavior in a new digital hospital. Appl Clin Inform 2022 Oct 09;13(5):1079-1091 [FREE Full text] [doi: 10.1055/s-0042-1758482] [Medline: 36351558]
11. Woods L, Eden R, Canfell OJ, Nguyen KH, Comans T, Sullivan C. Show me the money: how do we justify spending health care dollars on digital health? Med J Aust 2023 Feb 06;218(2):53-57 [FREE Full text] [doi: 10.5694/mja2.51799] [Medline: 36502453]
12. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Aust Health Rev 2018 Sep;42(5):568-578. [doi: 10.1071/AH17255] [Medline: 29986809]
13. Liu J, Luo L, Zhang R, Huang T. Patient satisfaction with electronic medical/health record: a systematic review. Scand J Caring Sci 2013 Dec 26;27(4):785-791. [doi: 10.1111/scs.12015] [Medline: 23181421]
14. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016 Feb 30;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
15. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to electronic health record adoption: a systematic literature review. J Med Syst 2016 Dec;40(12):252 [FREE Full text] [doi: 10.1007/s10916-016-0628-9] [Medline: 27714560]
16. Alkureishi MA, Lee WW, Lyons M, Press VG, Imam S, Nkansah-Amankra A, et al. Impact of electronic medical record use on the patient-doctor relationship and communication: a systematic review. J Gen Intern Med 2016 May 19;31(5):548-560 [FREE Full text] [doi: 10.1007/s11606-015-3582-1] [Medline: 26786877]
17. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12(6):573-576 [FREE Full text] [doi: 10.1370/afm.1713] [Medline: 25384822]
18. Rathert C, Williams ES, Linhart H. Evidence for the quadruple aim: a systematic review of the literature on physician burnout and patient outcomes. Med Care 2018 Dec;56(12):976-984. [doi: 10.1097/MLR.0000000000000999] [Medline: 30339573]
19. Privitera MR. Addressing human factors in burnout and the delivery of healthcare: quality and safety imperative of the quadruple aim. Health 2018;10(05):629-644. [doi: 10.4236/health.2018.105049]
20. Liddy C, Keely E. Using the quadruple aim framework to measure impact of heath technology implementation: a case study of eConsult. J Am Board Fam Med 2018 May 09;31(3):445-455 [FREE Full text] [doi: 10.3122/jabfm.2018.03.170397] [Medline: 29743227]
21. Wong AH, Ahmed RA, Ray JM, Khan H, Hughes PG, McCoy CE, et al. Supporting the quadruple aim using simulation and human factors during COVID-19 care. Am J Med Qual 2021 Mar;36(2):73-83 [FREE Full text] [doi: 10.1097/01.JMQ.0000735432.16289.d2] [Medline: 33830094]
22. Bingham G, Tong E, Poole S, Ross P, Dooley M. A longitudinal time and motion study quantifying how implementation of an electronic medical record influences hospital nurses' care delivery. Int J Med Inform 2021 Sep;153:104537. [doi: 10.1016/j.ijmedinf.2021.104537] [Medline: 34343955]
23. Viitanen J, Valkonen P, Savolainen K, Karisalmi N, Hölsä S, Kujala S. Patient experience from an eHealth perspective: a scoping review of approaches and recent trends. Yearb Med Inform 2022 Aug 04;31(1):136-145 [FREE Full text] [doi: 10.1055/s-0042-1742515] [Medline: 36463871]
24. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019 Apr 11;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]
25. Bae J, Rask KJ, Becker ER. The impact of electronic medical records on hospital-acquired adverse safety events: differential effects between single-source and multiple-source systems. Am J Med Qual 2018 Apr 07;33(1):72-80. [doi: 10.1177/1062860617702453] [Medline: 28387525]
26. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
27. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012 Nov 27;12(1):181 [FREE Full text] [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
28. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007 Jun 15;7(1):16 [FREE Full text] [doi: 10.1186/1472-6947-7-16] [Medline: 17573961]
29. Almond H, Mather C. Digital Health: A Transformative Approach. Amsterdam, The Netherlands: Elsevier; 2023.
30. Lim HC, Austin JA, van der Vegt AH, Rahimi AK, Canfell OJ, Mifsud J, et al. Toward a learning health care system: a systematic review and evidence-based conceptual framework for implementation of clinical analytics in a digital hospital. Appl Clin Inform 2022 Mar;13(2):339-354 [FREE Full text] [doi: 10.1055/s-0042-1743243] [Medline: 35388447]
31. EMRAM: a strategic roadmap for effective EMR adoption and maturity. HIMSS Analytics. 2017. URL: https://www.himssanalytics.org/emram [accessed 2021-03-31]

XSL•FO
RenderX

32. Sandelowski M, Barroso J. Toward a metasynthesis of qualitative findings on motherhood in HIV-positive women. Res Nurs Health 2003 Apr 17;26(2):153-170. [doi: 10.1002/nur.10072] [Medline: 12652611]
33. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8(1):45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
34. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/efi-180221]
35. Freire K, Pope R, Jeffrey K, Andrews K, Nott M, Bowman T. Engaging with children and adolescents: a systematic review of participatory methods and approaches in research informing the development of health resources and interventions. Adolescent Res Rev 2022 Feb 03;7(3):335-354. [doi: 10.1007/s40894-022-00181-w]
36. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme version. ESRC Methods Programme. 2006. URL: https://tinyurl.com/5n8bj7wu [accessed 2024-01-29]
37. Noyes J, Harden A, Ames H, Booth A, Flemming K, France E, et al. Cochrane-Campbell handbook for qualitative evidence synthesis. Cochrane Training. 2023. URL: https://tinyurl.com/mr2j8rv3 [accessed 2023-10-31]
38. Smith AE, Humphreys MS. Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. Behav Res Methods 2006 May;38(2):262-279. [doi: 10.3758/bf03192778] [Medline: 16956103]
39. Sotiriadou P, Brouwers J, Le TA. Choosing a qualitative data analysis tool: a comparison of NVivo and Leximancer. Ann Leis Res 2014 Apr 22;17(2):218-234. [doi: 10.1080/11745398.2014.902292]
40. Engstrom T, Strong J, Sullivan C, Pole JD. A comparison of Leximancer semi-automated content analysis to manual content analysis: a healthcare exemplar using emotive transcripts of COVID-19 hospital staff interactive webcasts. Int J Qual Methods 2022 Aug 18;21:16094069221118993 [FREE Full text] [doi: 10.1177/16094069221118993] [Medline: 36032654]
41. Haynes E, Garside R, Green J, Kelly MP, Thomas J, Guell C. Semiautomated text analytics for qualitative data synthesis. Res Synth Methods 2019 Sep 09;10(3):452-464 [FREE Full text] [doi: 10.1002/jrsm.1361] [Medline: 31125493]
42. Leximancer user guide. Leximancer Pty Ltd. 2002. URL: https://tinyurl.com/3y5yc8j2 [accessed 2024-01-29]
43. Knapp J, Zeratsky J, Kowitz B. Sprint: How to Solve Big Problems and Test New Ideas in Just Five Days. New York City, NY: Simon & Schuster; 2016.
44. Boyer L, Samuelian JC, Fieschi M, Lancon C. Implementing electronic medical records in a psychiatric hospital: a qualitative study. Int J Psychiatry Clin Pract 2010 Sep 10;14(3):223-227. [doi: 10.3109/13651501003717243] [Medline: 24917324]
45. Holden RJ. Physicians' beliefs about using EMR and CPOE: in pursuit of a contextualized understanding of health IT use behavior. Int J Med Inform 2010 Feb;79(2):71-80 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.12.003] [Medline: 20071219]
46. Al-Mujaini A, Al-Farsi Y, Al-Maniri A, Ganesh A. Satisfaction and perceived quality of an electronic medical record system in a tertiary hospital in oman. Oman Med J 2011 Sep 20;26(5):324-328 [FREE Full text] [doi: 10.5001/omj.2011.81] [Medline: 22125726]
47. Claret PG, Sebbanne M, Bobbia X, Bonnec JM, Pommet S, Jebali C, et al. First medical contact and physicians' opinion after the implementation of an electronic record system. Am J Emerg Med 2012 Sep;30(7):1235-1240. [doi: 10.1016/j.ajem.2011.06.014] [Medline: 22030205]
48. Kazley AS, Diana ML, Ford EW, Menachemi N. Is electronic health record use associated with patient satisfaction in hospitals? Health Care Manage Rev 2012;37(1):23-30. [doi: 10.1097/HMR.0b013e3182307bd3] [Medline: 21918464]
49. Top M, Gider Ö. Nurses' views on electronic medical records (EMR) in Turkey: an analysis according to use, quality and user satisfaction. J Med Syst 2012 Jun 8;36(3):1979-1988. [doi: 10.1007/s10916-011-9657-6] [Medline: 21302133]
50. Bossen C, Jensen LG, Udsen FW. Evaluation of a comprehensive EHR based on the DeLone and McLean model for IS success: approach, results, and success factors. Int J Med Inform 2013 Oct;82(10):940-953. [doi: 10.1016/j.ijmedinf.2013.05.010] [Medline: 23827768]
51. Jarvis B, Johnson T, Butler P, O'Shaughnessy K, Fullam F, Tran L, et al. Assessing the impact of electronic health records as an enabler of hospital quality and patient satisfaction. Acad Med 2013 Oct;88(10):1471-1477. [doi: 10.1097/ACM.0b013e3182a36cab] [Medline: 23969372]
52. Strauss B. The patient perception of the nurse-patient relationship when nurses utilize an electronic health record within a hospital setting. Comput Inform Nurs 2013 Dec;31(12):596-604. [doi: 10.1097/CIN.0000000000000014] [Medline: 24365972]
53. Top M, Yilmaz A, Gider Ö. Electronic medical records (EMR) and nurses in Turkish hospitals. Syst Pract Action Res 2012 Oct 13;26(3):281-297. [doi: 10.1007/s11213-012-9251-y]
54. Alharthi H, Youssef A, Radwan S, Al-Muallim S, Zainab AT. Physician satisfaction with electronic medical records in a major Saudi Government hospital. J Taibah Univ Med Sci 2014 Sep;9(3):213-218 [FREE Full text] [doi: 10.1016/j.jtumed.2014.01.004]
55. Burgin A, O'Rourke R, Tully MP. Learning to work with electronic patient records and prescription charts: experiences and perceptions of hospital pharmacists. Res Social Adm Pharm 2014 Sep;10(5):741-755. [doi: 10.1016/j.sapharm.2013.11.005] [Medline: 24378236]

XSL•FO
RenderX

56. Lakbala P, Dindarloo K. Physicians' perception and attitude toward electronic medical record. Springerplus 2014 Feb 3;3(1):63 [FREE Full text] [doi: 10.1186/2193-1801-3-63] [Medline: 24516790]
57. Migdal CW, Namavar AA, Mosley VN, Afsar-manesh N. Impact of electronic health records on the patient experience in a hospital setting. J Hosp Med 2014 Oct 23;9(10):627-633. [doi: 10.1002/jhm.2240] [Medline: 25052463]
58. Dowding DW, Turley M, Garrido T. Nurses' use of an integrated electronic health record: results of a case site analysis. Inform Health Soc Care 2015 Dec;40(4):345-361. [doi: 10.3109/17538157.2014.948169] [Medline: 25122056]
59. Harmon C, Fogle M, Roussel L. Then and now: nurses' perceptions of the electronic health record. Online J Nurs Inform 2015;19(1) [FREE Full text]
60. Shaker HA, Farooq MU, Dhafar KO. Physicians' perception about electronic medical record system in Makkah Region, Saudi Arabia. Avicenna J Med 2015 Aug 09;5(1):1-5 [FREE Full text] [doi: 10.4103/2231-0770.148499] [Medline: 25625082]
61. Tilahun B, Fritz F. Comprehensive evaluation of electronic medical record system use and user satisfaction at five low-resource setting hospitals in ethiopia. JMIR Med Inform 2015 May 25;3(2):e22 [FREE Full text] [doi: 10.2196/medinform.4106] [Medline: 26007237]
62. Bani-Issa W, Al Yateem N, Al Makhzoomy IK, Ibrahim A. Satisfaction of health-care providers with electronic health records and perceived barriers to its implementation in the United Arab Emirates. Int J Nurs Pract 2016 Aug;22(4):408-416. [doi: 10.1111/ijn.12450] [Medline: 27481126]
63. Chang CP, Lee TT, Liu CH, Mills ME. Nurses' experiences of an initial and reimplemented electronic health record use. Comput Inform Nurs 2016 Apr;34(4):183-190. [doi: 10.1097/CIN.0000000000000222] [Medline: 26886680]
64. Schenk EC, Mayer DM, Ward-Barney E, Estill P, Goss L, Shreffler-Grant J. RN perceptions of a newly adopted electronic health record. J Nurs Adm 2016 Mar;46(3):139-145. [doi: 10.1097/NNA.0000000000000313] [Medline: 26906518]
65. Bardach SH, Real K, Bardach DR. Perspectives of healthcare practitioners: an exploration of interprofessional communication using electronic medical records. J Interprof Care 2017 May 02;31(3):300-306 [FREE Full text] [doi: 10.1080/13561820.2016.1269312] [Medline: 28151026]
66. Kaipio J, Lääveri T, Hyppönen H, Vainiomäki S, Reponen J, Kushniruk A, et al. Usability problems do not heal by themselves: national survey on physicians' experiences with EHRs in Finland. Int J Med Inform 2017 Jan;97:266-281 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.10.010] [Medline: 27919385]
67. Ratanawongsa N, Barton JL, Lyles CR, Wu M, Yelin EH, Martinez D, et al. Computer use, language, and literacy in safety net clinic communication. J Am Med Inform Assoc 2017 Jan;24(1):106-112 [FREE Full text] [doi: 10.1093/jamia/ocw062] [Medline: 27274017]
68. Aldosari B, Al-Mansour S, Aldosari H, Alanazi A. Assessment of factors influencing nurses acceptance of electronic medical record in a Saudi Arabia hospital. Inform Med Unlocked 2018;10:82-88. [doi: 10.1016/j.imu.2017.12.007]
69. Burridge L, Foster M, Jones R, Geraghty T, Atresh S. Person-centred care in a digital hospital: observations and perspectives from a specialist rehabilitation setting. Aust Health Rev 2018 Sep;42(5):529-535. [doi: 10.1071/AH17156] [Medline: 29081349]
70. Abu Raddaha AH. Nurses' perceptions about and confidence in using an electronic medical record system. Proc Singapore Healthc 2017 Sep 21;27(2):110-117. [doi: 10.1177/2010105817732585]
71. Strudwick G, McGillis Hall L, Nagle L, Trbovich P. Acute care nurses' perceptions of electronic health record use: a mixed method study. Nurs Open 2018 Oct 07;5(4):491-500 [FREE Full text] [doi: 10.1002/nop2.157] [Medline: 30338094]
72. Tubaishat A. Perceived usefulness and perceived ease of use of electronic health records among nurses: application of Technology Acceptance Model. Inform Health Soc Care 2018 Dec 18;43(4):379-389. [doi: 10.1080/17538157.2017.1363761] [Medline: 28920708]
73. Alsohime F, Temsah MH, Al-Eyadhy A, Bashiri FA, Househ M, Jamal A, et al. Satisfaction and perceived usefulness with newly-implemented Electronic Health Records System among pediatricians at a university hospital. Comput Methods Programs Biomed 2019 Feb;169:51-57. [doi: 10.1016/j.cmpb.2018.12.026] [Medline: 30638591]
74. Burkoski V, Yoon J, Hutchinson D, Solomon S, Collins BE. Experiences of nurses working in a fully digital hospital: a phenomenological study. Nurs Leadersh (Tor Ont) 2019 May 07;32(SP):72-85. [doi: 10.12927/cjnl.2019.25813] [Medline: 31099748]
75. Hung SY, Nakayama M, Chen CC, Tsai FL. Physician perceptions of electronic medical records: the impact of system service quality, and generation/experience gaps. Int J Healthc Technol Manag 2019;17(4):229-254. [doi: 10.1504/IJHTM.2019.10026703]
76. Kutney-Lee A, Sloane DM, Bowles KH, Burns LR, Aiken LH. Electronic health record adoption and nurse reports of usability and quality of care: the role of work environment. Appl Clin Inform 2019 Jan 20;10(1):129-139 [FREE Full text] [doi: 10.1055/s-0039-1678551] [Medline: 30786302]
77. Schopf TR, Nedrebø B, Hufthammer KO, Daphu IK, Lærum H. How well is the electronic health record supporting the clinical tasks of hospital physicians? A survey of physicians at three Norwegian hospitals. BMC Health Serv Res 2019 Dec 04;19(1):934 [FREE Full text] [doi: 10.1186/s12913-019-4763-0] [Medline: 31801518]
78. Tubaishat A. The effect of electronic health records on patient safety: a qualitative exploratory study. Inform Health Soc Care 2019 Jan 14;44(1):79-91. [doi: 10.1080/17538157.2017.1398753] [Medline: 29239662]

XSL•FO
RenderX

79. Williams DC, Warren RW, Ebeling M, Andrews AL, Teufel Ii RJ. Physician use of electronic health records: survey study assessing factors associated with provider reported satisfaction and perceived patient impact. JMIR Med Inform 2019 Apr 04;7(2):e10949 [FREE Full text] [doi: 10.2196/10949] [Medline: 30946023]
80. Alobo IG, Soyannwo T, Ukponwan G, Akogu S, Akpa AM, Ayankola K. Implementing electronic health system in Nigeria: perspective assessment in a specialist hospital. Afr Health Sci 2020 Jun 22;20(2):948-954 [FREE Full text] [doi: 10.4314/ahs.v20i2.50] [Medline: 33163063]
81. Burridge L, Foster M, Jones R, Geraghty T, Atresh S. Nurses' perspectives of person-centered spinal cord injury rehabilitation in a digital hospital. Rehabil Nurs 2020;45(5):263-270. [doi: 10.1097/rnj.0000000000000201] [Medline: 30664609]
82. De Groot K, De Veer AJ, Paans W, Francke AL. Use of electronic health records and standardized terminologies: a nationwide survey of nursing staff experiences. Int J Nurs Stud 2020 Apr;104:103523. [doi: 10.1016/j.ijnurstu.2020.103523] [Medline: 32086028]
83. Eden R, Burton-Jones A, Staib A, Sullivan C. Surveying perceptions of the early impacts of an integrated electronic medical record across a hospital and healthcare service. Aust Health Rev 2020;44(5):690. [doi: 10.1071/ah19157]
84. Hu X, Qu H, Houser SH, Ding J, Chen H, Zhang X, et al. Exploring association between certified EHRs adoption and patient experience in U.S. psychiatric hospitals. PLoS One 2020 Jun 17;15(6):e0234607 [FREE Full text] [doi: 10.1371/journal.pone.0234607] [Medline: 32555623]
85. Kaipio J, Kuusisto A, Hyppönen H, Heponiemi T, Lääveri T. Physicians' and nurses' experiences on EHR usability: comparison between the professional groups by employment sector and system brand. Int J Med Inform 2020 Feb;134:104018 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104018] [Medline: 31835158]
86. Moerenhout T, Fischer GS, Saelaert M, De Sutter A, Provoost V, Devisch I. Primary care physicians' perspectives on the ethical impact of the electronic medical record. J Am Board Fam Med 2020 Jan 06;33(1):106-117 [FREE Full text] [doi: 10.3122/jabfm.2020.01.190154] [Medline: 31907251]
87. Schwarz M, Coccetti A, Draheim M, Gordon G. Perceptions of allied health staff of the implementation of an integrated electronic medical record across regional and metropolitan settings. Aust Health Rev 2020;44(6):965. [doi: 10.1071/ah19024]
88. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020 Jul 15;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
89. Cho Y, Kim M, Choi M. Factors associated with nurses' user resistance to change of electronic health record systems. BMC Med Inform Decis Mak 2021 Jul 17;21(1):218 [FREE Full text] [doi: 10.1186/s12911-021-01581-z] [Medline: 34273990]
90. Heponiemi T, Gluschkoff K, Vehko T, Kaihlanen AM, Saranto K, Nissinen S, et al. Electronic health record implementations and insufficient training endanger nurses' well-being: cross-sectional survey study. J Med Internet Res 2021 Dec 23;23(12):e27096 [FREE Full text] [doi: 10.2196/27096] [Medline: 34941546]
91. Jung SY, Hwang H, Lee K, Lee D, Yoo S, Lim K, et al. User perspectives on barriers and facilitators to the implementation of electronic health records in behavioral hospitals: qualitative study. JMIR Form Res 2021 Apr 08;5(4):e18764 [FREE Full text] [doi: 10.2196/18764] [Medline: 33830061]
92. Kutney-Lee A, Brooks Carthon M, Sloane DM, Bowles KH, McHugh MD, Aiken LH. Electronic health record usability: associations with nurse and patient outcomes in hospitals. Med Care 2021 Jul 01;59(7):625-631 [FREE Full text] [doi: 10.1097/MLR.0000000000001536] [Medline: 33797506]
93. Lloyd S, Long K, Oshni Alvandi A, Di Donato J, Probst Y, Roach J, et al. A national survey of EMR usability: comparisons between medical and nursing professions in the hospital and primary care sectors in Australia and Finland. Int J Med Inform 2021 Oct;154:104535. [doi: 10.1016/j.ijmedinf.2021.104535] [Medline: 34425552]
94. Luyten J, Marneffe W. Examining the acceptance of an integrated electronic health records system: insights from a repeated cross-sectional design. Int J Med Inform 2021 Jun;150:104450. [doi: 10.1016/j.ijmedinf.2021.104450] [Medline: 33848941]
95. Pruitt ZM, Howe JL, Hettinger AZ, Ratwani RM. Emergency physician perceptions of electronic health record usability and safety. J Patient Saf 2021 Dec 01;17(8):e983-e987. [doi: 10.1097/PTS.0000000000000849] [Medline: 33871414]
96. Tian D, Hoehner CM, Woeltje KF, Luong L, Lane MA. Disrupted and restored patient experience with transition to new electronic health record system. J Patient Exp 2021 Aug 18;8:23743735211034064 [FREE Full text] [doi: 10.1177/23743735211034064] [Medline: 34423122]
97. Jedwab RM, Hutchinson AM, Manias E, Calvo RA, Dobroff N, Redley B. Change in nurses' psychosocial characteristics pre- and post-electronic medical record system implementation coinciding with the SARS-CoV-2 pandemic: pre- and post-cross-sectional surveys. Int J Med Inform 2022 Jul;163:104783 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104783] [Medline: 35512624]
98. Al Otaybi HF, Al-Raddadi R, Bakhamees FH. Performance, barriers, and satisfaction of healthcare workers toward electronic medical records in Saudi Arabia: a national multicenter study. Cureus 2022 Feb;14(2):e21899 [FREE Full text] [doi: 10.7759/cureus.21899] [Medline: 35155041]
99. Arikan F, Kara H, Erdogan E, Ulker F. Barriers to adoption of electronic health record systems from the perspective of nurses: a cross-sectional study. Comput Inform Nurs 2021 Nov 15;40(4):236-243. [doi: 10.1097/CIN.0000000000000848] [Medline: 34812779]

XSL•FO
RenderX

100. Chivilgina O, Elger BS, Benichou MM, Jotterand F. "What's the best way to document information concerning psychiatric patients? I just don't know"-a qualitative study about recording psychiatric patients notes in the era of electronic health records. PLoS One 2022 Mar 3;17(3):e0264255 [FREE Full text] [doi: 10.1371/journal.pone.0264255] [Medline: 35239698]
101. Czernik Z, Yu A, Pell J, Feinbloom D, Jones CD. Hospitalist perceptions of electronic health records: a multi-site survey. J Gen Intern Med 2022 Jan 21;37(1):269-271 [FREE Full text] [doi: 10.1007/s11606-020-06558-8] [Medline: 33479933]
102. Monturo C, Brockway C, Ginev A. Electronic health record transition: the patient experience. Comput Inform Nurs 2022 Jan 01;40(1):53-60. [doi: 10.1097/CIN.0000000000000805] [Medline: 34313622]
103. Upadhyay S, Hu HF. A qualitative analysis of the impact of electronic health records (EHR) on healthcare quality and safety: clinicians' lived experiences. Health Serv Insights 2022 Mar 03;15:11786329211070722 [FREE Full text] [doi: 10.1177/11786329211070722] [Medline: 35273449]
104. Welchen V, Matte J, Giacomello CP, Dalle Molle F, Camargo ME. Usability perception of the health information systems in Brazil: the view of hospital health professionals on the electronic health record. RAUSP Manag J 2022 Apr 07;57(3):264-279. [doi: 10.1108/rausp-02-2021-0023]
105. Catchpole K, Neyens DM, Abernathy J, Allison D, Joseph A, Reeves ST. Framework for direct observation of performance and safety in healthcare. BMJ Qual Saf 2017 Dec 28;26(12):1015-1021 [FREE Full text] [doi: 10.1136/bmjqs-2016-006407] [Medline: 28971880]
106. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption factors of the electronic health record: a systematic review. JMIR Med Inform 2016 Jun 01;4(2):e19 [FREE Full text] [doi: 10.2196/medinform.5525] [Medline: 27251559]
107. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun 12;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
108. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022 Dec 22;24(12):e43086 [FREE Full text] [doi: 10.2196/43086] [Medline: 36548034]
109. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013 Jan 03;3(1):e001570 [FREE Full text] [doi: 10.1136/bmjopen-2012-001570] [Medline: 23293244]
110. Wang T, Giunti G, Melles M, Goossens R. Digital patient experience: umbrella systematic review. J Med Internet Res 2022 Aug 04;24(8):e37952 [FREE Full text] [doi: 10.2196/37952] [Medline: 35925651]
111. Queensland digital health consumer charter. Health Consumers Queensland. URL: https://www.hcq.org.au/qdhcc/ [accessed 2024-01-29]
112. Kukafka R. Digital health consumers on the road to the future. J Med Internet Res 2019 Nov 21;21(11):e16359 [FREE Full text] [doi: 10.2196/16359] [Medline: 31750835]
113. Pelland KD, Baier RR, Gardner RL. "It's like texting at the dinner table": a qualitative analysis of the impact of electronic health records on patient-physician interaction in hospitals. J Innov Health Inform 2017 Jun 30;24(2):894 [FREE Full text] [doi: 10.14236/jhi.v24i2.894] [Medline: 28749316]
114. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci 2022 Aug 03;9(8):220638 [FREE Full text] [doi: 10.1098/rsos.220638] [Medline: 35950198]
115. Bouabida K, Lebouché B, Pomey MP. Telehealth and COVID-19 pandemic: an overview of the telehealth use, advantages, challenges, and opportunities during COVID-19 pandemic. Healthcare (Basel) 2022 Nov 16;10(11):2293 [FREE Full text] [doi: 10.3390/healthcare10112293] [Medline: 36421617]
116. Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics (MDPI) 2020 Sep;7(3):25 [FREE Full text] [doi: 10.3390/informatics7030025] [Medline: 33274178]
117. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul 21;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
118. Hogg HD, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
119. Odendaal WA, Anstey Watkins JA, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Mar 26;3(3):CD011942 [FREE Full text] [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
120. Ames HM, Glenton C, Lewin S, Tamrat T, Akama E, Leon N. Clients' perceptions and experiences of targeted digital communication accessible via mobile devices for reproductive, maternal, newborn, child, and adolescent health: a qualitative evidence synthesis. Cochrane Database Syst Rev 2019 Oct 14;10(10):CD013447 [FREE Full text] [doi: 10.1002/14651858.CD013447] [Medline: 31608981]

XSL•FO
RenderX

121. De Rosis S, Cerasuolo D, Nuti S. Using patient-reported measures to drive change in healthcare: the experience of the digital, continuous and systematic PREMs observatory in Italy. BMC Health Serv Res 2020 Apr 16;20(1):315 [FREE Full text] [doi: 10.1186/s12913-020-05099-4] [Medline: 32299440]
122. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
123. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearb Med Inform 2019 Aug 16;28(1):208-217 [FREE Full text] [doi: 10.1055/s-0039-1677918] [Medline: 31419834]
124. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar 02;28(1):e100262 [FREE Full text] [doi: 10.1136/bmjhci-2020-100262] [Medline: 33653690]

Abbreviations

AI: artificial intelligence
EMR: electronic medical record
ENTREQ: Enhancing the Transparency in Reporting the Synthesis of Qualitative Research
HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
V1: version 1
V2: version 2

Edited by T Leung; submitted 29.03.23; peer-reviewed by W Tam, E Pittman; comments to author 07.09.23; revised version received 08.11.23; accepted 31.01.24; published 11.03.24.

Please cite as:
Canfell OJ, Woods L, Meshkat Y, Krivit J, Gunashanhar B, Slade C, Burton-Jones A, Sullivan C
The Impact of Digital Hospitals on Patient and Clinician Experience: Systematic Review and Qualitative Evidence Synthesis
J Med Internet Res 2024;26:e47715
URL: https://www.jmir.org/2024/1/e47715
doi:10.2196/47715
PMID:38466978

©Oliver J Canfell, Leanna Woods, Yasaman Meshkat, Jenna Krivit, Brinda Gunashanhar, Christine Slade, Andrew Burton-Jones, Clair Sullivan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review

Karen O'Connor[1], MSc; Su Golder[2], PhD; Davy Weissenbacher[3], PhD; Ari Z Klein[1], PhD; Arjun Magge[1], PhD; Graciela Gonzalez-Hernandez[3], PhD

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]Department of Health Sciences, University of York, York, United Kingdom
[3]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States

Corresponding Author:
Karen O'Connor, MSc
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine
University of Pennsylvania
423 Guardian Dr
Philadelphia, PA, 19004
United States
Phone: 1 215 573 8089
Email: karoc@pennmedicine.upenn.edu

Abstract

Background: Patient health data collected from a variety of nontraditional resources, commonly referred to as *real-world data*, can be a key information source for health and social science research. Social media platforms, such as Twitter (Twitter, Inc), offer vast amounts of real-world data. An important aspect of incorporating social media data in scientific research is identifying the demographic characteristics of the users who posted those data. Age and gender are considered key demographics for assessing the representativeness of the sample and enable researchers to study subgroups and disparities effectively. However, deciphering the age and gender of social media users poses challenges.

Objective: This scoping review aims to summarize the existing literature on the prediction of the age and gender of Twitter users and provide an overview of the methods used.

Methods: We searched 15 electronic databases and carried out reference checking to identify relevant studies that met our inclusion criteria: studies that predicted the age or gender of Twitter users using computational methods. The screening process was performed independently by 2 researchers to ensure the accuracy and reliability of the included studies.

Results: Of the initial 684 studies retrieved, 74 (10.8%) studies met our inclusion criteria. Among these 74 studies, 42 (57%) focused on predicting gender, 8 (11%) focused on predicting age, and 24 (32%) predicted a combination of both age and gender. Gender prediction was predominantly approached as a binary classification task, with the reported performance of the methods ranging from 0.58 to 0.96 F_1-score or 0.51 to 0.97 accuracy. Age prediction approaches varied in terms of classification groups, with a higher range of reported performance, ranging from 0.31 to 0.94 F_1-score or 0.43 to 0.86 accuracy. The heterogeneous nature of the studies and the reporting of dissimilar performance metrics made it challenging to quantitatively synthesize results and draw definitive conclusions.

Conclusions: Our review found that although automated methods for predicting the age and gender of Twitter users have evolved to incorporate techniques such as deep neural networks, a significant proportion of the attempts rely on traditional machine learning methods, suggesting that there is potential to improve the performance of these tasks by using more advanced methods. Gender prediction has generally achieved a higher reported performance than age prediction. However, the lack of standardized reporting of performance metrics or standard annotated corpora to evaluate the methods used hinders any meaningful comparison of the approaches. Potential biases stemming from the collection and labeling of data used in the studies was identified as a problem, emphasizing the need for careful consideration and mitigation of biases in future studies. This scoping review provides valuable insights into the methods used for predicting the age and gender of Twitter users, along with the challenges and considerations associated with these methods.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47923) doi:10.2196/47923

KEYWORDS

social media; demographics; Twitter; age; gender; prediction; real-world data; neural network; machine learning; gender prediction; age prediction

Introduction

Background

Real-world data are data regarding patients' health collected outside randomized controlled trials from a variety of nontraditional resources such as electronic health records, medical claims data, or data generated by patients themselves such as social media data that may be used to support study design to develop real-world evidence [1]. Real-world data from social media have been increasingly recognized as a valuable resource for gaining knowledge about and insight into a variety of health-related research topics, including disease surveillance [2,3], pharmacovigilance [4,5], and mental health [6,7]. They can also be used for the identification of cohorts for potential recruitment into traditional studies [8,9]. In short, social media can readily provide abundant personal health information in real time.

The use of data from social media platforms, particularly Twitter (Twitter, Inc), for health-related research is subject to some inherent limitations in that demographic information (with the exception of location, which is available when the user has enabled the location feature) is not explicitly available through the application programming interface (API) [10]. Demographic traits, including age, gender, race or ethnicity, location, education, and income, hold significant value in health research. Few studies based on Twitter data incorporated an assessment of Twitter user demographics into their analysis [11]. Understanding the demographic traits of Twitter users provides significant value when using the data in health research. It not only facilitates sample representativeness, which is crucial for generalizing research findings and ensuring that the conclusions drawn from Twitter data can be extrapolated to broader populations [12], but also enables subgroup analysis. It allows for the comparison of health-related behaviors, attitudes, and outcomes across different groups and enables targeted interventions and tailored health care strategies [13,14]. Moreover, demographic information is actionable and can assist in designing public health interventions and policies for specific populations based on their needs and concerns as expressed on social media.

Predicting demographic traits is complex and challenging. A user's profile does not necessarily include such information, and researchers have used other features available in the data, such as names, content of the tweets, or the individual's network to make predictions. A 2018 systematic review assessed the use of social media to predict demographic traits, finding successful implementation for 14 traits, including gender and age [15,16]. Although the review provided a broad overview of the state of demographic prediction using social media, the details of the machine learning (ML) methods used were not reviewed. A recent review provided insights into the methods used for predicting the race and ethnicity of Twitter users [17].

Objectives

In this study, our objective was to present a scoping review of automated methods used for predicting the age and gender of Twitter users to provide an overview of the techniques published since 2017. We focused our review on studies that used Twitter, as it is the most commonly used social media platform for this research [15]. Twitter is an attractive platform to use in research, as the terms of use for this platform are well understood by both users and researchers, it includes an API, and the data on it are abundant for health-related research [18].

Although other demographic traits such as location, education, and income can provide valuable insights, the age and gender of Twitter users present distinct advantages and considerations for health research. Given the differences in disease presentation by gender, such as with acute coronary syndrome [19], and by age, such as with COVID-19 [20], identifying the age and gender of the users included in studies using Twitter data may elicit insights into disease prevalence, patterns, and variations across different subgroups in disease presentation or treatment response [21,22]. Age and gender also play crucial roles in shaping health behaviors and attitudes. For example, studying age and gender differences in smoking habits [23], physical activity levels [24], and adherence to medical treatments [25,26] can provide insights into effective interventions and health promotion campaigns for specific groups. Although Twitter users are generally representative of the population, there is a certain degree of skew in their demographics: there is an overrepresentation of individuals aged <30 years, whereas individuals aged >65 years are underrepresented when compared with the overall demographics of the US population [27,28]. Therefore, it is important to include the age and gender of Twitter users in a study to enable the accurate reporting of findings, making them specific to certain subgroups, or to make any necessary adjustments to account for potential biases that may arise from these demographic differences.

Although studies aimed at predicting Twitter users' gender began as early as 2011 [29-33] and efforts aimed at detecting the age of Twitter users have been made since 2013 [34-36], it is only since 2017 that the language processing community shifted its methods away from handcrafted rules and represented text documents with dense vectors to train deep neural networks (DNNs) [37,38], resulting in a noticeable increase in performance for many applications. We sought to examine whether these increases in performance were evident in the methods used for the prediction of the age and gender of Twitter users.

Methods

Overview

We report this review following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Extension for Scoping Reviews) [39] methodology. The completed PRISMA-ScR checklist is available in Table S1 in Multimedia Appendix 1. We searched several databases to identify studies on the prediction of Twitter users' age or gender or both. Our database search strategy combines 3 facets: facet 1 includes terms related to Twitter, facet 2 consists of terms for age or gender, and facet 3 consists of terms for methods of prediction such as ML. The search strategy was translated as appropriate for each database. The detailed search strategy is available in Multimedia Appendix 2. The ML term facet was expanded using terms from related reviews by Hinds and Joinson [15] and Umar et al [40]. The search criteria were limited to peer-reviewed journals, conference proceedings, books, and theses.

The following databases were searched with a publication date range of 2017 or later (Textbox 1).

Textbox 1. List of databases searched with the total number of combined facet results.

- ACL (Association for Computational Linguistics) Anthology: 5080, of which the first 50 records were screened
- ACM (Association for Computing Machinery) Digital Library: 23
- Cumulative Index to Nursing & Allied Health (CINAHL): 57
- Embase: 262
- Google Scholar: 767,000, of which the first 50 records were screened
- IEEE (Institute of Electrical and Electronics Engineers) Xplore: 23
- Library and Information Science Abstracts: 31
- Library, Information Science and Technology Abstracts: 48
- Proquest Dissertations and Theses—United Kingdom and Ireland: 58
- Ovid MEDLINE: 183
- PsycINFO: 104
- Science Citation Index, Social Science Citation Index, Conference Proceedings Citation Index—Science, and Conference Proceedings Citation Index—Social Science and Humanities: 131
- Zetoc: 61

Citations were exported to a shared EndNote (Clarivate) library for deduplication. Using the Population, Intervention, Comparison, Outcomes, and Study Design (PICOS) [41] framework, we developed a list of inclusion and exclusion criteria (refer to the *Inclusion and Exclusion Criteria* section), and 2 screeners from the research team screened the results independently, with disputes discussed after screening and a consensus decision reached. In addition, given that search engines and unmanageable data sources are recommended to be included as secondary data sources [42-44], the first 50 records from both ACL (Association for Computational Linguistics) Anthology and Google Scholar were screened using the aforementioned methods. We set a limit on the number of results screened, as the relevance of the results is ranked by the search engines, with the most relevant results listed first [45-48].

Inclusion and Exclusion Criteria

We framed our research question using the PICOS framework. Table 1 outlines our specific inclusion and exclusion criteria. As explained in the *Introduction* section, we restricted the date of our search to include only publications from 2017 and beyond. No language restrictions were applied to the inclusion criteria; however, financial and logistical restraints allowed us to include only studies written in English, Spanish, Chinese, or French.

XSL•FO
RenderX

Table 1. Inclusion and exclusion criteria, developed per the Population, Intervention, Comparison, Outcomes, and Study Design framework, for the scoping review.

Facet	Inclusion criteria	Exclusion criteria
Population	Any Twitter (Twitter, Inc) data on Twitter users, such as posts, profile details, photos, or avatars	Studies evaluating prediction from data on other social media platforms, such as Facebook (Meta Platforms, Inc) or Instagram (Meta Platforms, Inc)
Intervention	Methods for predicting the gender or age of Twitter users; articles that used machine learning, natural language processing, human in the loop, or other computationally assisted methods to predict the gender or age of the users	Studies that contained no computation methods
Comparator	Any or none; we included any studies irrespective of whether they had a comparator and, if they did have a comparator, irrespective of what that was	N/A[a]
Outcome	Gender or age prediction	Any other demographic trait prediction
Study design	Any type of peer-reviewed study reporting on the methods used to predict gender or age; such information must be the primary focus of the study or reported in enough detail to be reproducible	Discussion papers, commentaries, and letters
Date	2017 or later	Before 2017
Language	All	None

[a]N/A: not applicable.

Data Extraction

From each included paper, we extracted the following data: the year of publication, publication type (journal, conference paper, book chapter, or thesis), demographic predicted (gender, age, or both), language of tweets, size of the data set, collection method for the data set, details of prediction models, features used in the models (posts, profile, and images), performance of the models, name of any software used for prediction, measures used to assess the methods and results of any evaluation, and the availability of data or code. The included papers were distributed among the authors for data extraction. The extracted data were validated by another author (KO).

Results

Overview

Our database searches resulted in 981 studies, which were retrieved and entered into an EndNote library, where duplicates were removed, leaving 684 (69.7%) studies for sifting.

After the abstract review, 172 (25.1%) of 684 studies were deemed potentially relevant by either one of the independent sifters (SG and KO). The full texts of these studies were screened independently, and disagreements were discussed, resulting in the inclusion of 74 (43%) studies [49-122] and exclusion of 98 (57%) studies (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the included studies.

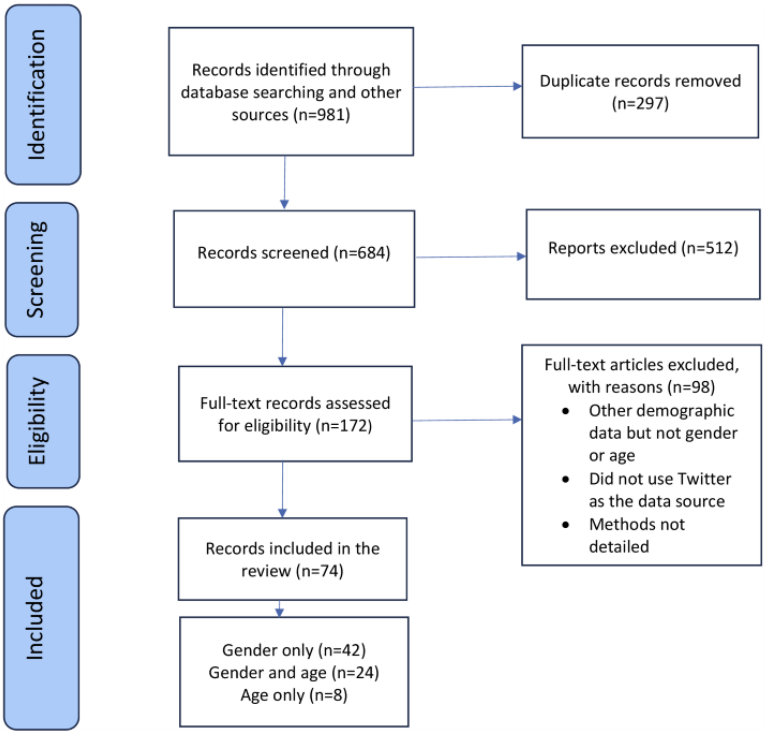

XSL•FO
RenderX

Characteristics of the Included Studies

Among the 74 included studies (Multimedia Appendices 3 [49-52,54-63,65,67-72,74-89,91-93,96-99,101-122] and 4 [51,53,55,56,59,60,63-67,70,73,74,77,80,83,84,87,90,94,95,99-101,108,110,112,116,118-120]), the majority (n=42, 57%) focused on predicting only the gender of the individual, 24 (32%) explored predicting both gender and age, and 8 (11%) focused solely on predicting age. Most of the studies were published in conference proceedings (44/74, 59%), followed by journal articles (28/74, 38%), theses (2/74, 3%), and a book chapter (1/74, 1%).

In 42 (57%) of the 74 studies, developing methods to predict Twitter users' age or gender or both was the primary purpose. In the remaining studies (32/74, 43%), the identification of the demographic characteristics of Twitter users was secondary. Within this last group, 9 (28%) studies developed ad-hoc methods to determine age, gender, or both, whereas the others used open-source models (13/32, 41%) or off-the-shelf software (10/32, 31%).

Studies Developing Ad-Hoc Methods for Gender and Age Prediction

Gender

Overview

Of the 74 studies, 44 (59%) developed ad-hoc methods to predict the Twitter users' gender. Of these 44 studies, 32 (73%) predicted the users' gender alone [49,50,52,54,57,58,68,69,71,72,75,76,79,81,82,85,86,89,92,93,96,102,104-107,111,113,115,117,121,122], and 12 (27%) predicted gender along with age [51,55,65,70,80,83,87,101,108,110,112,116].

Most studies that developed ad-hoc methods (41/44, 93%) approached the problem of gender prediction as a binary classification task, predicting whether the label male or female applies to each user account, whereas 4% (3/44) of studies [93,112,119] added the classification of organization or brand.

We found that approaches to predict gender included tweets written in multiple languages, including English [52,82,83,92,93,115,117], German [76], Slovenian [106], Italian [49], Japanese [89], Arabic and Egyptian [57,58,79], French, Dutch, Portuguese, and Spanish, and a multilingual study assessed tweets written in 28 languages and dialects [112].

Data Sets

For the training and validation of the ad-hoc approaches for gender detection, some studies (19/44, 43%) used previously created annotated corpora, whereas others (27/44, 61%) collected data directly from Twitter. Among the 19 studies that used previously annotated data sets, 9 (47%) [55,57,58,68,70,86,87,96,121] used corpora from the PAN-Conference and Labs of the Evaluation Forum (CLEF; PAN-CLEF) author profiling tasks [123-129], whereas 10 (53%) studies [72,75,83,85,93,104,105,115,117,122] relied on data sets from other studies [113,130-136].

In the 27 (61%) studies that collected data directly from Twitter, different components of Twitter accounts were used. These components were used either for manually or semiautomatically validating the gender of a user or for computing features describing the user to train a classifier (Multimedia Appendix 5 [49-122]). Despite data limitations from the Twitter API, it was the main source of data collection, with 22 (24%) studies [49-52,54,69,71,76,79,81,89,92,101,102,106-108,110,111,116,117,121] collecting data either as a random sample from the Twitter Streaming API or based on keywords or geographic location from the Twitter Search API. Of the 5 studies not using the Twitter API, 1 (20%) [82] collected data using a scraping tool, 3 (60%) [80,112,113] used a random sample from a collection of 10% of tweets from 2014 to 2017 or the Twitter archive, and 1 (20%) did not specify its data source [65].

The 24 studies that created a labeled data set (Multimedia Appendix 6 [49,51-54,63,64,66,69,71,73,76,77,80,82,89,90,92,106-108,110,112,113,116-118,120]) to train and test or to validate the performance of the system determined the gender of the users using multiple components of their Twitter accounts (Multimedia Appendix 5). A total of 11 (46%) studies labeled the data through manual annotation, where the annotators determined the gender using profile pictures [52,54], user names [71], profiles [89], or a combination of these [76,82,92,106,108,110,116]. There were 11 (46%) studies that automatically or semiautomatically labeled their data sets via the detection of self-reports or gender-identifying terms (eg, mother, son, and uncle) [69,80,108,110,112,117], the user's name [49,107,113], or declarations on other linked social media [116,117]. A total of 3 (13%) studies created their labeled data sets by using the accounts of famous social media influencers [65] or using an unspecified collection of users whose gender is known [51,79]. Of the 24 studies, only 8 (33%) reported data availability. Of the 8 studies, 6 (75%) stated availability *by request*, and 2 (25%) had working links to the whole corpus (Multimedia Appendix 6).

Nonpersonal Accounts

A Twitter account may not be authored by or represent a single person. There are organization or company accounts as well as *bot* accounts. A bot is an automatic or semiautomatic user account. Some bot accounts identify themselves as such and may be used to automatically amplify news or tweets related to a certain topic. Others may emulate human accounts and be used with a more malicious intent to sow discord, manipulate public opinion, or spread misinformation. There were 9 (12%) of the 74 included studies [49,76,92,93,96,103,104,106,112] that removed nonpersonal (organization) accounts when they manually annotated their collections. Some studies (11/74, 15%) implemented heuristics to explicitly detect and remove nonpersonal accounts [49,50,59,71,81,107,113,122], bot accounts [98], or both [79,137]. Others (39/74, 53%) used previously annotated data sets consisting of only personal accounts, labeled and removed nonpersonal accounts, or collected their data sets based on self-reports of age and gender or other identifiable personal information. The remaining (15/74, 20%) studies provided no details on how or whether these accounts were removed (Multimedia Appendix 5).

Features and Models

The reviewed studies used data labeled with the user's gender to build and evaluate classification models based on features

XSL•FO
RenderX

describing the tweets (such as n-grams, word embeddings, hashtags, and URLs) [57,58,65,68-71,75,79,82,86,87,92,96,104,109,113,121], features derived from the users' profile metadata (such as user names, bio, followers, and users followed) [49,51,52,72,80,85,112,115,122], features derived from a combination of their profile metadata and tweets [52,54,76,83,93,107,108,110,117] or images [52,80,108,112,116]. Of the 74 studies, 1 (3%) study from Japan included the user's geographic information under the assumption that, culturally, a person of a certain demographic is more likely to frequent specific places [89].

Among the systems that used handcrafted features (25/44, 57%), most (13/25, 52%) achieved their best results using a support vector machine (SVM) [49,54,65,72,82,85,86,104-106,113,116,138], whereas others (12/25, 48%) used logistic regression [87,107,110], naive Bayes [51,92], random forests [80], bag of trees [70], extreme gradient boosting [89], or ensemble approaches [76,79,107,122] (details are provided in Table 2). Other systems used deep learning methods (15/44, 34%) such as DNNs, convolutional neural networks, feed forward neural networks or recurrent neural networks [55,68,71,75,93,115,121], bidirectional long-term short-term memory [58], gated recurrent units [57], graph recursive neural networks [83], and multimodal deep learning networks [108,112].

One of the studies created a meta-classifier ensemble classifying users based on the predictions of multiple individual classifiers [117], including SVM, bidirectional encoder representations from transformers, and 2 existing models [112,139]. Another study created a DNN for learning with label proportion, a semisupervised approach [52]. The results of the best-performing deep learning model as reported in each study are presented in Table 3. Studies that used lexical matching (4/44, 9%) of the user's name to a curated name dictionary [50,81,101,102] to determine gender reported no validation or performance metrics.

XSL•FO
RenderX

Table 2. Top reported system performance for studies predicting the gender of Twitter users using traditional machine learning (ML) methods. Result metrics are reflected in this table as reported in the original publications and are not necessarily comparable to each other.

Study	Language	ML method	Reported performance	
			F_1-score	Accuracy
Cesare et al [122], 2017	English	Ensemble: lexical match and SVM[a] and DT[b]	0.84	0.83
Jurgens et al [80], 2017	English	RF[c] ensemble	0.78	0.80
Ljubešić et al [85], 2017	Portuguese, French, Dutch, Spanish, German, and Italian	SVM	N/A[d]	0.61-0.69
Markov et al [87], 2016	English, Spanish, Dutch, and Italian	LogR[e]	N/A	0.57-0.77
Mukherjee and Bala [92], 2016	English	NB[f]	0.75	0.71
Verhoeven et al [106], 2017	Slovenian	SVM	0.93	0.93
Volkova [110], 2015	English and Spanish	LogR	N/A	0.82
Xiang et al [116], 2017	English	SVM and PME[g]	N/A	0.76
Cheng et al [65], 2018	English, Filipino, and Taglish	SVC[h] with lasso	0.84	0.84
Emmery et al [69], 2017	English	fastText	N/A	0.76
Giannakopoulos et al [72], 2018	N/A	SVM PNN[i]	N/A	0.87
Khandelwal et al [82], 2018	Code-mixed Hindi-English	SVM	N/A	0.9
Miura et al [89], 2018	Japanese	XGBoost[j]	N/A	0.89
van der Goot et al [104], 2018	English, Dutch, French, Portuguese, and Spanish	SVM	N/A	0.66-0.72
Alessandra et al [49], 2019	Italian	Ensemble: lexical match and SVM	N/A	0.75
Hirt et al [76], 2019	German	Ensemble: binary classifiers	0.81	N/A
Hussein et al [79], 2019	Dialect Egyptian Arabic	Ensemble: RF and LinR[k]	NA	0.77-0.88
Vicente et al [107], 2018	English and Portuguese	Ensemble: Face++, LinR, and SVM	N/A	0.93-0.97
Arafat et al [51], 2020	Indonesian	Multinomial NB	N/A	0.75
Baxevanakis et al [54], 2020	Greek	SVM	N/A	0.7
Garcia-Guzman et al [70], 2020	English	Bag of trees	0.64	0.64
López-Monroy et al [86], 2020	English and Spanish	Bag of trees	0.64	0.64
Pizarro [96], 2020	English and Spanish	SVM	0.82-0.84	N/A
Vashisth and Meehan [105], 2020	English	LogR	N/A	0.57
Wong et al [113], 2020	English	SVM	0.58-0.62	0.60

[a]SVM: support vector machine.
[b]DT: decision tree.
[c]RF: random forest.
[d]N/A: not applicable.
[e]LogR: logistic regression.
[f]NB: naive Bayes.
[g]PME: projection matrix extraction.
[h]SVC: support vector classifier.
[i]PNN: probabilistic neural network.
[j]XGBoost: extreme gradient boosting.
[k]LinR: linear regression.

XSL•FO
RenderX

Table 3. Top reported system performance for studies predicting the gender of Twitter users using deep learning machine learning (ML) methods. Result metrics are reflected in this table as reported in the original publications and are not comparable to each other.

Study	Language	ML method	Reported performance	
			F_1-score	Accuracy
Ardehaly and Culotta [52], 2017	English	Deep LLP[a]	0.96	N/A[b]
Geng et al [71], 2017	English	Ensemble: LDA[c] and CNN[d]	N/A	0.87
Kim et al [83], 2017	English	GRNN[e]	N/A	0.68
Vijayaraghavan et al [108], 2017	English	DMT[f]	0.89	N/A
Wang et al [111], 2017	N/A	CNN	0.91	0.9
Bayot and Goncalves [55], 2017	English and Spanish	CNN	N/A	0.59-0.72
Bsir and Zrigui [57], 2018	Arabic	GRU[g]	N/A	0.79
Wood-Doughty et al [115], 2018	English	RNN[h]	0.84	0.84
Bsir and Zrigui [58], 2019	Arabic	BILSTM[i] with attention	N/A	0.82
Hashempour [75], 2019	Portuguese, French, Dutch, Spanish, German, and Italian	FFNN[j]	N/A	0.84-0.86
Wang et al [112], 2019	Multilingual	mmDNN[k]	0.92	N/A
ElSayed and Farouk [68], 2020	Egyptian and Arabic dialects	Multichannel CNN-biGRU[l]	N/A	0.84-0.91
Imuede et al [93], 2020	English	DNN[m]	N/A	0.68
Zhao et al [121], 2020	English	CNN	0.80	N/A
Yang et al [117], 2021	English	Ensemble: M3[n] and SVM[o]	0.95	0.94

[a]LLP: learning with label proportions.
[b]N/A: not applicable.
[c]LDA: latent Dirichlet allocation.
[d]CNN: convolutional neural network.
[e]GRNN: graph recurrent neural network.
[f]DMT: deep multimodal multitask.
[g]GRU: gated recurrent network.
[h]RNN: recurrent neural network.
[i]BILSTM: bidirectional long-term short-term memory.
[j]FFNN: feed forward neural network.
[k]mmDNN: multimodal deep neural network.
[l]biGRU: bidirectional gated recurrent unit.
[m]DNN: deep neural network.
[n]M3: multimodal, multilingual, and multi-attribute system.
[o]SVM: support vector machine.

Performance

Performance results from the traditional ML methods cannot be directly compared against the deep learning methods used, as they were evaluated against different gold-standard corpora, and they used nonstandardized reporting metrics. However, looking at the overall results in terms of F_1-score, the results of the studies using deep learning had a relatively narrower range of reported performance (0.84-0.96), with a higher minimum of 0.84 and higher maximum of 0.96, compared with the reported performance range for traditional ML methods, which spans from 0.64 to 0.93.

Age

Overview

We found 19 studies that developed ad-hoc methods to predict the Twitter user's age, among which 7 (37%) predicted age exclusively [53,64,66,73,90,94,95]. All but 1 (5%) of the studies [80] approached the detection of Twitter users' age as an automatic classification of predefined age groups. The number of age groups varied across the studies (Table 3), with the ages categorized into 2 [53,73,83,110,116], 3 [51,66,90,94,95,101,108], 4 [70,112], or more [55,64,65,87] groups. The range of ages within the groups also varied across the studies; for example, across the 5 studies that took a binary classification approach, Guimaraes et al [73] used 13 to 19 years and ≥20

XSL•FO
RenderX

years as the 2 age groups, Volkova et al [110] and Kim et al [83] used 18 to 23 years or ≥25 years, Xiang et al [116] used ≤30 years or >30 years, and Ardehaly and Culotta [53] used <25 years and ≥25 years.

Except for 2 (11%) studies that did not report the language of the tweets used [51,73], all studies used English language tweets. A total of 8 (42%) studies extended their systems to include additional languages, including Spanish [55,64,87,110], Dutch [87,94,95], Filipino [65], and multiple languages [112].

Data Sets

Most studies (9/19, 47%) that developed new algorithms prepared new data sets to evaluate them with data retrieved directly using Twitter's API [51,53,66,73,90,108] or used other sources of data for this purpose [64,80,112] (Multimedia Appendix 4). Several studies used data sets made available by other studies to train or evaluate their algorithms: among the 19 studies, 2 (11%) studies [94,95] combined data sets from Sloan et al [34], Nguyen et al [36], and Morgan-Lopez et al [90]; Kim et al [83] used the data set from Volkova et al [140]; and 3 (15%) studies [55,70,87] used data sets that were created for the PAN-CLEF author profiling shared tasks [124-126]. The studies that prepared new data sets (Multimedia Appendix 6) labeled users' age groups by automatically or semiautomatically searching (1) for tweets that self-reported birthday announcements or age [53,80,90,108,110,112], (2) for tweets in which a user was wished a happy birthday [90], (3) for profiles that self-reported age [64,66,108,112], (4) for profiles that mentioned age-related keywords (eg, *grandparent*) [66,112], or (5) for manual annotation based on images or profile metadata [112,116,140] or (6) by subjectively perceiving age groups based on the content of individual tweets [73]. In 1 (5%) study [51], a mixture of self-reported information and demographic information of known individuals was used to label the data. Similar to studies on gender, the reported availability of the corpora was scarce. Only 5 (26%) studies reported that their data sets were available, 2 (40%) by request, 1 (20%) provided a link to the whole data set, and 2 (40%) provided a link to a sample of the corpus (Multimedia Appendix 6).

Features and Models

The studies used labeled age groups to evaluate classification models based on the features of the users' profile metadata (eg, user names, bio, followers, and users followed) [51,53,64,80,112], a combination of their profile metadata and tweets (eg, n-grams, word embeddings, hashtags, and URLs) [73,83,90,94,95,108,110], tweet texts only [65,66,70,87], or images [80,108,112,116].

For automatic classification, most studies (12/19, 63%) used traditional supervised ML methods, including logistic regression [51,66,87,90,110], Bayesian probabilistic inference [64], random forests [80], bag of trees [70], or SVM [65,116], or a semisupervised approach, learning from label proportion [53]. Other studies (7/16, 37%) used deep learning methods such as convolutional neural networks [55,73,94,95], graph recursive neural networks [83], and multimodal deep learning networks [108,112]. The best-performing systems for each study are listed in Tables 4 and 5. Of the 19 studies, 1 (5%) [101] classified age based on a previously developed age lexicon and did not report any performance metrics.

XSL•FO
RenderX

Table 4. Top reported system performance for studies predicting the age of Twitter users using traditional machine learning (ML) methods. Result metrics are reflected in this table as reported in the original publications and are not directly comparable to each other. Reviews are ordered by the number of classification groups.

Study	Number of age groups	Age class detail (y)	Language	ML method	Reported performance	
					F_1-score	Accuracy
Jurgens et al [80], 2017	N/A[a]	Continuous	English	RF[b] regression	N/A	0.71
Volkova [110], 2017	2	18-23 and 25-30	English and Spanish	LogR[c]	N/A	0.77
Xiang et al [116], 2017	2	≤30 and >30	English	CPME[d]	N/A	0.74
Ardehaly and Culotta [53], 2018	2	<25 and >25	English	LLP[e]	N/A	0.78
Morgan-Lopez et al [90], 2017	3	13-17, 18-24, and >24	English	LogR	0.74	N/A
Arafat et al [51], 2020	3	≤24, 25-39, and ≥40	NR[f]	LogR	N/A	0.71
Cornelisse and Pillai [66], 2020	3	18-24, 25-54, and >55	English	LogR	0.78	N/A
Markov et al [87], 2017	5	18-24, 25-34, 35-49, 50-64, and >65	English, Spanish, Dutch, and Italian	LogR	N/A	0.56-0.65
Cheng et al [65], 2018	5	18-24, 25-34, 35-44, 45-54, and 55-64	English, Filipino, and Taglish	SVC[g]	0.61	0.86
Garcia-Guzman et al [70], 2020	4	18-24, 25-34, 35-49, and >50	English	Bag of trees	N/A	0.67
Chamberlain et al [64], 2017	10 (3 subgroups)	<12, 12-13, 14-15, 16-17, 18-24, 25-34, 35-44, 45-54, 55-64, and >64	English, Spanish, French, and Portuguese	Bayesian probability	0.31-0.86 (3 class)	N/A

[a]N/A: not applicable.

[b]RF: random forest.

[c]LogR: logistic regression.

[d]CPME: coupled projection matrix extraction.

[e]LLP: learning with label proportions.

[f]NR: not reported.

[g]SVC: support vector classifier.

XSL•FO
RenderX

Table 5. Top reported system performance for studies predicting the age of Twitter users using deep learning machine learning (ML) methods. Result metrics are reflected in this table as reported in the original publications and are not comparable to each other. Reviews are ordered by the number of classification groups.

Study	Number of age groups	Age class detail (y)	Language	ML method	Reported performance	
					F_1-score	Accuracy
Guimaraes et al [73], 2017	2	13-19 and >20	English	CNN[a]	0.94	N/A[b]
Kim et al [83], 2017	2	Young (18-23) and old (25-30)	English	GRNN[c]	N/A	0.81
Vijayaraghavan et al [108], 2017	3	<30, 30-60, and >60	English	DMT[d]	0.82	N/A
Pandya et al [94], 2018	3	Dutch: <20, 20-40, and >40; English 1: 13-17, 18-40, and >40; and English 2: 13-17, 18-24, and >25	English and Dutch	CNN	0.61-0.87	N/A
Pandya et al [95], 2020	3	Dutch: <20, 20-40, and >40; English 1: 13-17, 18-40, and >40; and English 2: 13-17, 18-24, and >25	English and Dutch	CNN	0.82-0.87	N/A
Wang et al [112], 2019	4	≤18, 18-30, 30-40, and 40-99	Multilingual—28	mmDNN[e]	0.52	N/A
Bayot and Goncalves [55], 2017	5	18-24, 25-34, 35-49, 50-64, and ≥65	English and Spanish	CNN	N/A	0.43-0.55

[a]CNN: convolutional neural network.

[b]N/A: not applicable.

[c]GRNN: graph recurrent neural network.

[d]DMT: deep multimodal multitask.

[e]mmDNN: multimodal deep neural network.

Performance

Assessing the performance differences between studies using traditional ML methods and those using deep learning or neural networks is challenging owing to variations in classification criteria (eg, different age groupings and different number of classification categories) and the variety of performance metrics reported. However, for both methods, higher performance was noted when the problem was framed as a binary or ternary classification than as a larger multinomial classification.

Studies Using Previously Developed Methods

Overview

Among the 74 included studies, there were 23 (31%) studies in which the detection of gender or age was secondary to their research, and previously developed methods were used to detect the demographic information of their cohort. Of the 23 studies, 13 (57%) used open-source models, and 10 (43%) used off-the-shelf software. More details about each study are given in the subsequent sections.

Open-Source Models

Of the 13 studies that used open-source models, 3 (4%) [74,99,100] drew upon an extant model [141] that used a predictive lexicon for the multiclass classification of age or gender for their applications. None of these studies created a validation corpus to assess the performance of the system, which was originally reported as 89.9% accuracy for gender and 0.84 Pearson correlation coefficient for age. One (1%) study [118] used the same text-based model [141] and an image model [142] to determine the age and gender of their cohort. When tested against their gold-standard corpus of self-reports from profile descriptions, they found that the imaging model performed best for gender (accuracy=90%-92%), whereas textual features gave the best results for age (accuracy=60%). A total of 3 (4%) studies [78,91,114] used demographer [115,139,143] for gender predictions, with 1 (33%) study [91] evaluating the performance against a set of users who had self-reported their gender in a survey, finding an F_1-score of 0.869 for women and 0.770 for men. A total of 2 (3%) studies [61,62] used an ensemble classifier of previously developed models, with a reported accuracy of 0.83 and an F_1-score of 0.83 [122]. Two (3%) other studies [67,120] used M3 [112] to detect gender and age, with 1 (50%) study validating the performance using a manually labeled data set, finding an accuracy of 95.9% and an F_1-score of 0.957 for gender and an accuracy of 77.6% and an F_1-score of 0.731 for age. One (1%) study [56] used Deep EXpectation of apparent age [144] for age and gender detection, which reported a validation error of 3.96 years for age and an 88% accuracy for gender. One (1%) study [98] used the rOPenSci *gender* package, and no assessment of performance was reported.

Off-the-Shelf Software

In the 10 studies that used off-the-shelf software, Face ++ was the most common software, being used in 6 (60%) studies [63,77,88,97,109,119]. The remaining studies used DemographicsPro [59,60], Microsoft Face API [84], and RapidMiner [103].

In 4 (40%) [88,97,103,109] of the 10 studies, no validation of performance was carried out, and a further 2 (20%) studies simply reported that DemographicsPro *requires* 95% confidence to make an estimation [59,60]. Other studies (n=4, 40%)

XSL•FO
RenderX

compared with manual annotation and identified an accuracy of 82.8% for age using Face ++ [77], 68% for strict age groups, or 83% if the age groupings were relaxed [63]. The performance for age using Microsoft Face API was measured at 0.895 Gwet agreement coefficient (AC) [84], when compared with manually labeled data sets.

For gender, the studies (2/10, 20%) that measured performance against their own gold-standard labeled set of users recorded accuracies of 94.4% [77] or 88% [63] using Face ++. Other studies (3/10, 30%) [88,97,109] reported a confidence level of 95% +0.015 or –0.015 using Face ++ for gender prediction.

Only 1 (10%) [119] of the 10 studies went beyond manual annotation to create a gold standard and used multiple search techniques to manually verify age and gender, including LinkedIn profiles, electoral roll listings, personal websites, Twitter descriptions, and Twitter profile images. In this study, Face++ accuracy for age was reported as 40.4%, and Face++ accuracy for gender was reported as 44.8% (with a valid image accuracy of 32.5% for age and 87.7% for gender), and crowdsourcing annotation accuracy for age was 60.8% and for gender was 86.4% (with valid image accuracy of 56.1% for age and 93.9% for gender).

Discussion

Principal Findings

Overview

In this review, we aimed to provide an overview of recent ML methods used to predict the gender and age of Twitter users, as these are key demographics for epidemiology. Our review indicates that both tasks have been popular, but the identification of gender has received more attention than the identification of age. However, no de facto standards for research (ie, data collection and evaluation) have emerged, resulting in a large number of heterogeneous studies that are not directly comparable. Thus, it is not straightforward to conclude where the state-of-the-art stands for these tasks.

Our review found evidence of potential bias that impacts the quality and representativeness of the data used in the studies. One prevalent source of bias lies in the data collection and labeling processes. For instance, some studies may introduce systemic bias through the use of imprecise labeling methods such as name matching for labeling Twitter users' gender. This approach can lead to mislabeling, especially for individuals with names that are culturally diverse or androgynous and introduce inaccuracies into the training data. Another problem is the introduction of sampling bias through the use of artificially balanced data sets, creating an unrepresentative sample of the Twitter population, which, in reality, has a skewed distribution, with certain age and gender groups being more prevalent than others.

It is important to address and limit these biases because when ML models are trained on biased data, they tend to replicate and amplify these biases in their predictions [145].

The prediction of demographic information is an important task to address to fully realize the potential advantages of using social media data, such as those of using Twitter data in health-related research. In the United States, the National Institute of Health has committed to including women participants in clinical studies and including sex as a biological variable, finding that the disaggregation of data by sex will allow for sex-based comparisons of results to identify any sex-based differences. A recent review [146] found that this disaggregation in the development of ML models led to the discovery of sex-based differences that improved the model performance for sex-specific cohorts. Age is also important, as it can correlate and be a factor in the course and progression of disease [146] or the effects of medication [147]. Given the significance of this information, accurate and reproducible models must be developed. One way to ensure the reproducibility of models is for researchers to make data and codes available, including annotation guidelines. In addition to model performance, studies that create annotated corpora should report annotator agreement measures to assess the quality of the corpora. Few of the included studies made their data or code available (Multimedia Appendices 3, 4, and 6).

A particular difficulty when comparing different systems comes from a lack of a *gold standard* that can be used to compare the systems. Some studies created their own corpora, collecting data randomly or based on keywords relevant to their studies. Others reused data sets from prior studies or shared tasks. Although outside the scope of this review, there have been shared tasks that aim to advance research through competition, focusing on gender and age prediction. A longstanding shared task focused on author profiling was hosted at the PAN workshop of CLEF [123-129]. More recently, Social Media Mining for Health (SMM4H), 2022, included 2 tasks for age detection, releasing new annotated corpora for the tasks [148]; several researchers reported using the corpora from these shared tasks. Testing and reporting performance metrics against these publicly available data sets, without alteration, would provide a comparable metric of different approaches. However, although reusing annotated corpora provides quick access to labeled data, it does have some limitations, including data loss over time as users delete their tweets, which not only reduces the size of the data but also can result in a data imbalance in the corpus.

A summary of our recommendations to reduce some potential bias in the data and improve the classification, reproducibility, and validation of the ML methods used can be found in Figure 2.

XSL•FO
RenderX

Figure 2. Summary of recommendations for best practices in the collection of training data, and the development and dissemination of age or gender machine learning prediction models.

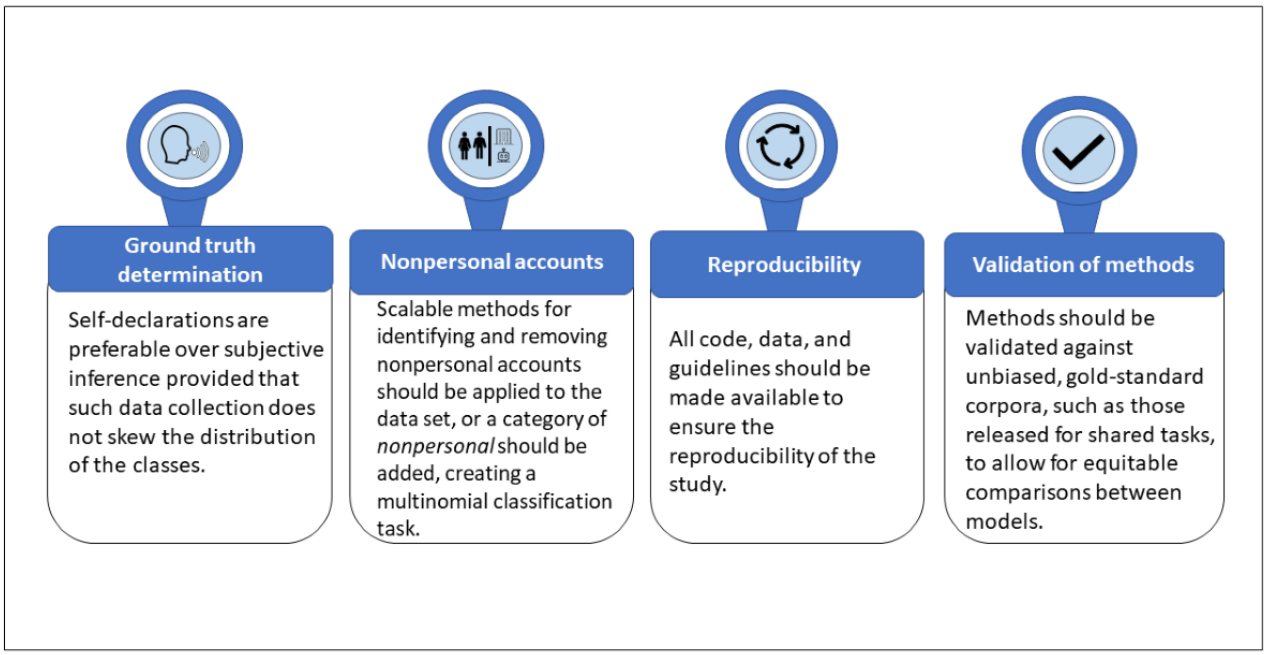

Gender Prediction

Almost all the included studies approached the gender prediction task as a binary classification task, identifying a user as either male or female. We note that even when focusing on binary gender classification, which is the prevalent approach, the task of gender prediction on Twitter could be better characterized as a multinomial classification task: given a user account, the classifier should return male, female, or *nonpersonal*. The last label (nonpersonal) can account for Twitter users representing organizations or bots. Although some studies attempted to identify and exclude nonpersonal accounts as a preprocessing step, other studies developed their systems using previously annotated data sets that were exclusively labeled as male or female users or removed nonpersonal accounts during annotation before training and testing. It is unknown how well these systems would perform when extended to unseen data that may contain nonpersonal accounts.

Excluding nonpersonal accounts, the ratio of male users to female users in the training data set is also important, as it should mimic the natural distribution of Twitter users, estimated to be 31.5% female users and 68.5% male users as of January 2021 [149]. However, some authors biased their collections using unconventional methods of collection or using artificially balanced data sets. The most conventional method to collect a set of Twitter accounts is to query for any tweet mentioning functional words without semantic meaning such as *of*, *the*, or *and* from the Twitter API. Whereas collecting Twitter users using functional or neutral keywords, a given language, or geographic areas resulted in a male:female ratio close to the ratio naturally observed on Twitter, other choices resulted in collections with different ratios. Such changes in ratios could have improved (or deteriorated) the training of the authors' classifiers and biased their evaluations, which did not reflect the performance of their approach on a random sample of Twitter users.

All studies treated gender as a binary determination of male or female. Although some referenced the limitation of this approach, they opted to use these designations given the need to align their data with outside resources, such as the US census or social security administration data. We note that gender, unlike biological sex, is not necessarily binary as it is a social construct and has been shown to influence a person's use of health care, interactions, therapeutic responses, disease perceptions, and decision-making [150]; this underlies the importance of expanding the efforts of classification beyond binary to improve accuracy and avoid misinterpreting results.

Age Prediction

The age prediction task generally had a lower performance than the gender prediction task. This was true for studies that developed their own models as well as those that used open-source or off-the-shelf software. This may be because most studies approached age prediction as a multiclass classification task. The proxies used, such as language, names, networks, or images, may have limited predictive value for age. In addition, the distribution of Twitter users means that any data set will be inherently imbalanced, providing few training examples for age groups at the tail end of the distribution. This data imbalance may lead to too few instances of the minority classes to effectively train the classifier. For classification models based on images, poor performance for age may be unsurprising given that it can be difficult for humans to discern age from a single image. In addition, photos may be subject to photo editing or enhancement or may not be a recent photograph of the user. Because of a lack of error analysis reports in the included studies, it is difficult to determine the source of the classification difficulty for age.

XSL•FO
RenderX

Performance aside, the fact that the number and range of age groups vary across studies suggests that a classification approach is not generalizable to all research applications. Identifying the exact age, rather than age groups, can generalize to applications that do not align with predefined groupings of binary or multiclass models; however, using high-precision rules to extract self-reports of exact age from the user's profile metadata had been shown not to scale. As we worked on this study, we noted that none of the reviewed systems opted for extracting the exact age. To test the feasibility and utility of a generalizable system that extracts the exact age from a tweet in a user's timeline using deep learning methods, separate from this study, our group developed a classification and extraction pipeline using the RoBERTa-Large model and a rule-based extraction model [151]. The system was trained and tested on 11,000 annotated tweets. The classification of tweets mentioning an age achieved an F_1-score of 0.93, and the extraction of age from these tweets achieved an F_1-score of 0.86. From a collection of 245,947 users, age was extracted for 54% using REPORTage. A shared task for the classification task ran at the SMM4H 2022 workshop, and we released the annotated data set. We did not include our approach in the scoping review, as there were no comparable systems published before the release of the exact age extraction approach as part of the SMM4H 2022 shared task.

Potential Bias of Differing Methods

The limitations of using names to distinguish between genders may promote bias, particularly if the names used for training do not represent the ethnic diversity of the population, and some cultures may have more unisex names than others, which cannot be used to distinguish genders. There can be a high degree of uncertainty for many users for whom gender cannot be classified by name; estimates by Sloan et al [152] are that 52% of users will be unclassified using this method. However, studies have suggested that the classifications made may be relatively accurate given that the data from UK Twitter demonstrates a high level of agreement with the UK census data [153]. Furthermore, when used alone, this heuristic may label some organization accounts, such as PAUL_BAKERY, as a person.

Relying on self-declarations may be prone to bias as well. For example, younger people are more likely to profess their age than older adults, as age may be more important to them. With respect to gender pronouns, these may be more likely to be declared by those in some occupations or age groups. Indeed, there may also be other biases to self-declarations of data based on culture, background, social class, or country of origin or residence.

Using users' profile images for gender and age identification is challenging. Not all Twitter users provide a picture of themselves, with many opting for pictures of their pets, objects, children, scenery, or even celebrities. Identifying the gender and age of even those with pictures of themselves can be problematic if the quality of the pictures is poor, the pictures contain more than 1 face, or the pictures are not recent, particularly for predicting age. A comparison of systems using images to predict demographics [154] measured not only the accuracy in identifying age and gender but also the percentage of images in which a face could be detected, finding that only approximately 30% of Twitter users had a single detectable face.

Methods to filter out organizations in the studies included removing accounts with a large number of followers [71] or explicitly searching for organizations by matching username terms linked to economic activities, such as restaurant and hotel [49]. These methods remove accounts that do not represent a single user. However, they do not remove bots. Although one of the studies created a classifier to detect bots, the filtering of bots was limited to those identified in manual annotation, by simple heuristics, or nonexistent in many studies (Multimedia Appendix 5).

Validation of Age and Gender Proxies

For cases where age or gender are estimated, it is necessary to conduct validation exercises whereby the data are compared with a *gold-standard data set* to establish accuracy levels. For example, 1 study [119] that used off-the-shelf software also created a manually annotated gold-standard data set for measuring accuracy. This study found that although the accuracy of crowdsourcing was higher than that of software, the accuracy was only approximately 60% for age. This puts into question the use of manual annotations alone as a gold standard.

The most reliable way of generating a *gold standard* is to obtain the information directly from the user. This may be done in the form of direct correspondence with the user, such as messaging via social media or, the other way around, requesting Twitter handles in surveys that collect demographic data. Other methods for validation, such as manual extraction, may be less rigorous. However, these methods can be improved by multiple independent annotators, using experienced teams.

External validation of the model is also a vital step to assess how the model will perform on unseen data [155,156]. In a validation on a second data set, Yang et al [117] found that performance dropped in all but 2 of their models, stressing the importance of benchmarking existing systems on a targeted corpus. This step is equally important when using existing systems, so a range of expected performances can be reported and used in any analysis of the output.

In addition to the potential biases reported earlier, predicting the age and gender of Twitter users has some potential limitations that should be considered and, when possible, addressed to limit their effects. As evidenced by the performance results of the included studies, determining the precise age or age group of Twitter users solely based on their Twitter profiles and tweet content can be challenging. Although methods to extract a user's self-reported age can be executed with high precision [151], predicting age, especially for more specific age groups, remains a complex task. Another limitation to consider is the potential for users to misrepresent their reported age or gender, which can introduce inaccuracies and affect the reliability of predictions based on user-supplied data. This phenomenon is not unique to Twitter and has been identified in other data sources such as surveys [157,158]. Many of the included studies used self-reported data to label their training data; therefore, any potential misrepresentations could be

XSL•FO
RenderX

approached as a noisy label problem. There are numerous methods that can be used to manage the effect of label noise on classification models, such as distance learning or ensemble methods [159,160]. Furthermore, it is important to effectively address potential noise and uncertainty when using the output data for secondary analysis. Statistical techniques that can handle imprecise or uncertain data, such as Bayesian inference or fuzzy logic, can be valuable in this context. Using these methods, the analysis can better account for uncertain predictions, leading to more robust and reliable results. Finally, users' age changes over time, and their profiles may not be updated accordingly, or the age tweet may be from an earlier year and not reflect their current age. Researchers should ensure that the users' labeled age is contemporaneous with the other data included in the prediction model. Predicting the age and gender of Twitter users provides valuable insights, and most identified limitations presented by the data can be mitigated.

Ethical Considerations

Several studies have shown that social media users generally do not have concerns about their data being used for research or even have favorable opinions about it [161,162]. However, the ethical frameworks for the use of these data are still being developed [163-165], and institutional review boards may deem the use of publicly available data, such as those collected from Twitter, as exempt from human participant research; however, it is incumbent on the researcher to consult with their institutional review boards or equivalent ethical committees to obtain such exemptions [165]. Although the data are publicly available, it is important to carefully consider potential ethical implications when predicting the age and gender of Twitter users. This process may raise privacy concerns, particularly when publishing data that may be considered sensitive, necessitating the protection of user identities and the anonymization of data to prevent reidentification [166]. Anonymizing the data may include removing user identifiers, modifying the tweet text, or generating synthetic tweets [165]. In addition, automated methods for predicting user age or gender have limitations and may result in misclassifications. Transparency regarding the limitations of the methods, algorithms, and data sources used in age and gender prediction are essential to report so that any use of these methods or data in secondary analysis can take such limitations into account. Although the prediction of age and gender may present some potential ethical concerns, it is important to recognize that there are also benefits to the use of these data for health research that can outweigh these concerns, such as eliciting insights into disease prevalence, patterns, and variations or distinguishing health behaviors and attitudes across different subgroups.

Limitations

It is unlikely that we have identified all studies using off-the-shelf software, as we did not search for specific named software, but part of our remit was to identify the array of software used. We did not limit our inclusion to only studies that developed their own software; therefore, we have included studies that used proprietary software. These software products do not publish their methodologies; therefore, we are unable to directly compare these approaches with others.

We also included studies for which the prediction of age and gender was secondary to the primary focus of their study. These studies either used proprietary software, previously developed methods, or developed limited methods to predict demographic information. In general, these studies did not report the performance of their prediction methods on their data sets. Although some reported the original performance metrics of the methods used, it cannot be assumed that these methods will perform similarly across all data.

Conclusions

The prediction of demographic data, such as age and gender, is an important step in increasing the value and application of social media data. Many methods have been reported in the literature with differing degrees of success. Although we sought to explore whether deep learning approaches would advance the performance for these tasks as they have been shown to do for other natural language processing tasks, many of the included studies used traditional ML methods. Although only explored by a handful of studies, deep learning methods appear to perform well for the prediction of a user's gender or age. However, direct comparison of the published methods was impossible, as different test sets were used in the studies. This highlights the need for recently developed, publicly available gold-standard corpora, such as those released for shared tasks such as SMM4H or PAN-CLEF, to have unbiased data and baseline metrics to compare different approaches going forward.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) under grant NIH-NLM R01LM011176. The NIH NLM funded this research but was not involved in the design or conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

The included studies are available on the web. The search strategy and extracted data on included studies are available in Multimedia Appendices 2-6.

XSL•FO
RenderX

Authors' Contributions

SG, KO, and GGH devised the study and identified data for extraction. SG created and executed the search strategy and created the initial draft of the manuscript. SG and KO were responsible for study selection. All the authors were responsible for data extraction, summarization, and discussion. KO synthesized all data and created all tables. All the authors commented on and edited the manuscript. KO provided the final version of the manuscript. All the authors contributed to the final draft of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 30 KB - jmir_v26i1e47923_app1.docx]

Multimedia Appendix 2
Search strategies and results for individual databases.
[DOCX File , 30 KB - jmir_v26i1e47923_app2.docx]

Multimedia Appendix 3
Extracted data from the included studies predicting gender.
[XLSX File (Microsoft Excel File), 44 KB - jmir_v26i1e47923_app3.xlsx]

Multimedia Appendix 4
Extracted data from the included studies predicting age.
[XLSX File (Microsoft Excel File), 27 KB - jmir_v26i1e47923_app4.xlsx]

Multimedia Appendix 5
Information on the identification and removal of nonpersonal or bot accounts from the data set. Features used for annotation or prediction of gender or age.
[XLSX File (Microsoft Excel File), 22 KB - jmir_v26i1e47923_app5.xlsx]

Multimedia Appendix 6
Details of corpora created in the included studies and their reported availability.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e47923_app6.xlsx]

References

1. Real-world evidence. U.S. Food and Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2023-03-30]
2. Alessa A, Faezipour M. A review of influenza detection and prediction through social networking sites. Theor Biol Med Model 2018 Feb 01;15(1):2 [FREE Full text] [doi: 10.1186/s12976-017-0074-5] [Medline: 29386017]
3. Bisanzio D, Kraemer MU, Bogoch II, Brewer T, Brownstein JS, Reithinger R. Use of Twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of COVID-19 at global scale. Geospat Health 2020 Jun 15;15(1) [FREE Full text] [doi: 10.4081/gh.2020.882] [Medline: 32575957]
4. Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. J Am Med Inform Assoc 2021 Sep 18;28(10):2184-2192 [FREE Full text] [doi: 10.1093/jamia/ocab114] [Medline: 34270701]
5. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681 [FREE Full text] [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
6. Guntuku SC, Sherman G, Stokes DC, Agarwal AK, Seltzer E, Merchant RM, et al. Tracking mental health and symptom mentions on twitter during COVID-19. J Gen Intern Med 2020 Sep;35(9):2798-2800 [FREE Full text] [doi: 10.1007/s11606-020-05988-8] [Medline: 32638321]
7. Ma L, Wang Y. Constructing a semantic graph with depression symptoms extraction from Twitter. In: Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Presented at: CIBCB 2019; July 9-11, 2019; Siena, Italy. [doi: 10.1109/cibcb.2019.8791452]
8. Bauermeister JA, Zimmerman MA, Johns MM, Glowacki P, Stoddard S, Volz E. Innovative recruitment using online networks: lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven

XSL•FO
RenderX

sampling (webRDS) strategy. J Stud Alcohol Drugs 2012 Sep;73(5):834-838 [FREE Full text] [doi: 10.15288/jsad.2012.73.834] [Medline: 22846248]
9. Weissenbacher D, Flores J, Wang Y, O'Connor K, Rawal S, Stevens R, et al. Automatic cohort determination from Twitter for HIV prevention amongst Black and Hispanic men. AMIA Jt Summits Transl Sci Proc 2022:504-513 [FREE Full text] [Medline: 35854738]
10. Twitter API. Twitter. URL: https://developer.twitter.com/en/docs/twitter-api [accessed 2023-03-30]
11. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
12. Gustafson DL, Woodworth CF. Methodological and ethical issues in research using social media: a metamethod of Human Papillomavirus vaccine studies. BMC Med Res Methodol 2014 Dec 02;14:127 [FREE Full text] [doi: 10.1186/1471-2288-14-127] [Medline: 25468265]
13. Beck C, McSweeney JC, Richards KC, Roberson PK, Tsai PF, Souder E. Challenges in tailored intervention research. Nurs Outlook 2010 Mar;58(2):104-110 [FREE Full text] [doi: 10.1016/j.outlook.2009.10.004] [Medline: 20362779]
14. Rimer BK, Kreuter MW. Advancing tailored health communication: a persuasion and message effects perspective. J Commun 2006 Aug;56(s1):S184-S201. [doi: 10.1111/j.1460-2466.2006.00289.x]
15. Hinds J, Joinson AN. What demographic attributes do our digital footprints reveal? A systematic review. PLoS One 2018 Nov 28;13(11):e0207112 [FREE Full text] [doi: 10.1371/journal.pone.0207112] [Medline: 30485305]
16. Dredze M. How social media will change public health. IEEE Intell Syst 2012 Jul;27(4):81-84 [FREE Full text] [doi: 10.1109/MIS.2012.76]
17. Golder S, Stevens R, O'Connor K, James R, Gonzalez-Hernandez G. Methods to establish race or ethnicity of twitter users: scoping review. J Med Internet Res 2022 Apr 29;24(4):e35788 [FREE Full text] [doi: 10.2196/35788] [Medline: 35486433]
18. Bour C, Ahne A, Schmitz S, Perchoux C, Dessenne C, Fagherazzi G. The use of social media for health research purposes: scoping review. J Med Internet Res 2021 May 27;23(5):e25736 [FREE Full text] [doi: 10.2196/25736] [Medline: 34042593]
19. van Oosterhout R, de Boer AR, Maas AH, Rutten FH, Bots ML, Peters SA. Sex differences in symptom presentation in acute coronary syndromes: a systematic review and meta-analysis. J Am Heart Assoc 2020 May 05;9(9):e014733 [FREE Full text] [doi: 10.1161/JAHA.119.014733] [Medline: 32363989]
20. Trevisan C, Noale M, Prinelli F, Maggi S, Sojic A, Di Bari M, EPICOVID19 Working Group. Age-related changes in clinical presentation of Covid-19: the EPICOVID19 web-based survey. Eur J Intern Med 2021 Apr;86:41-47 [FREE Full text] [doi: 10.1016/j.ejim.2021.01.028] [Medline: 33579579]
21. Brady E, Nielsen MW, Andersen JP, Oertelt-Prigione S. Lack of consideration of sex and gender in COVID-19 clinical studies. Nat Commun 2021 Jul 06;12(1):4015 [FREE Full text] [doi: 10.1038/s41467-021-24265-8] [Medline: 34230477]
22. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. Nature 2019 Nov;575(7781):137-146. [doi: 10.1038/s41586-019-1657-6] [Medline: 31695204]
23. Amiri P, Mohammadzadeh-Naziri K, Abbasi B, Cheraghi L, Jalali-Farahani S, Momenan AA, et al. Smoking habits and incidence of cardiovascular diseases in men and women: findings of a 12 year follow up among an urban Eastern-Mediterranean population. BMC Public Health 2019 Aug 05;19(1):1042 [FREE Full text] [doi: 10.1186/s12889-019-7390-0] [Medline: 31382950]
24. Rosselli M, Ermini E, Tosi B, Boddi M, Stefani L, Toncelli L, et al. Gender differences in barriers to physical activity among adolescents. Nutr Metab Cardiovasc Dis 2020 Aug 28;30(9):1582-1589. [doi: 10.1016/j.numecd.2020.05.005] [Medline: 32605880]
25. Chen SL, Lee WL, Liang T, Liao IC. Factors associated with gender differences in medication adherence: a longitudinal study. J Adv Nurs 2014 Sep;70(9):2031-2040. [doi: 10.1111/jan.12361] [Medline: 24506542]
26. Krueger K, Botermann L, Schorr SG, Griese-Mammen N, Laufs U, Schulz M. Age-related medication adherence in patients with chronic heart failure: a systematic literature review. Int J Cardiol 2015 Apr 01;184:728-735. [doi: 10.1016/j.ijcard.2015.03.042] [Medline: 25795085]
27. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 7. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2021-12-09]
28. World population prospects - population division - United Nations. United Nations Department of Economic and Social Affairs Population Division. URL: https://population.un.org/wpp/ [accessed 2023-07-28]
29. Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist J. Understanding the demographics of Twitter users. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;5(1):554-557. [doi: 10.1609/icwsm.v5i1.14168]
30. Fink C, Kopecky J, Morawski M. Inferring gender from the content of tweets: a region specific example. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;6(1):459-462. [doi: 10.1609/icwsm.v6i1.14320]
31. Alowibdi JS, Buy UA, Yu P. Language independent gender classification on Twitter. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Presented at: ASONAM '13; August 25-28, 2013; Niagara Falls, ON.
32. Chen L, Qian T, Zhu P, You Z. Learning user embedding representation for gender prediction. In: Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). Presented at: ICTAI 2016; November 6-8, 2016; San Jose, CA. [doi: 10.1109/ictai.2016.0048]

XSL•FO
RenderX

33. Culotta A, Ravi NK, Cutler J. Predicting the demographics of Twitter users from website traffic data. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Presented at: AAAI'15; January 25-30, 2015; Austin, TX URL: https://dl.acm.org/doi/abs/10.5555/2887007.2887018 [doi: 10.1609/aaai.v29i1.9204]
34. Sloan L, Morgan J, Burnap P, Williams M. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. PLoS One 2015;10(3):e0115545 [FREE Full text] [doi: 10.1371/journal.pone.0115545] [Medline: 25729900]
35. Oktay H, Firat A, Ertem Z. Demographic breakdown of Twitter users: an analysis based on names. In: Proceedings of the Academy of Science and Engineering (ASE). Presented at: ASE'14; September 15-19, 2014; Västerås, Sweden URL: https://www.researchgate.net/publication/315538705_Demographic_Breakdown_of_Twitter_Users_An_analysis_based_on_names
36. Nguyen D, Gravel R, Trieschnigg D, Meder T. "How old do you think I am?" A study of language and age in Twitter. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;7(1):439-448. [doi: 10.1609/icwsm.v7i1.14381]
37. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online January 16, 2013 . [doi: 10.48550/arXiv.1301.3781]
38. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, Nevada URL: https://dl.acm.org/doi/proceedings/10.5555/2999792
39. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
40. Umar A, Bashir SA, Abdullahi MB, Adebayo OS. Comparative study of various machine learning algorithms for tweet classification. i-Manager J Comput Sci 2019;6(4):12 [FREE Full text]
41. Amir-Behghadami M, Janati A. Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. Emerg Med J 2020 Jun;37(6):387. [doi: 10.1136/emermed-2020-209567] [Medline: 32253195]
42. Bramer WM, Giustini D, Kramer BM. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. Syst Rev 2016 Mar 01;5:39 [FREE Full text] [doi: 10.1186/s13643-016-0215-7] [Medline: 26932789]
43. Kugley S, Wade A, Thomas J, Mahood Q, Jørgensen AM, Hammerstrøm K, et al. Searching for studies: a guide to information retrieval for Campbell systematic reviews. Campbell Syst Rev 2017 Feb 13;13(1):1-73. [doi: 10.4073/cmg.2016.1]
44. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Technical Supplement to Chapter 4: searching for and selecting studies. In: Higgins JP, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 6.3. London, UK: The Cochrane Collaboration; 2022.
45. Booth A. How much searching is enough? Comprehensive versus optimal retrieval for technology assessments. Int J Technol Assess Health Care 2010 Oct;26(4):431-435. [doi: 10.1017/S0266462310000966] [Medline: 20923586]
46. Briscoe S, Nunns M, Shaw L. How do Cochrane authors conduct web searching to identify studies? Findings from a cross-sectional sample of Cochrane Reviews. Health Info Libr J 2020 Dec;37(4):293-318. [doi: 10.1111/hir.12313] [Medline: 32511888]
47. Stansfield C, Dickson K, Bangpan M. Exploring issues in the conduct of website searching and other online sources for systematic reviews: how can we be systematic? Syst Rev 2016 Nov 15;5(1):191 [FREE Full text] [doi: 10.1186/s13643-016-0371-9] [Medline: 27846867]
48. Godin K, Stapleton J, Kirkpatrick SI, Hanning RM, Leatherdale ST. Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. Syst Rev 2015 Oct 22;4:138 [FREE Full text] [doi: 10.1186/s13643-015-0125-0] [Medline: 26494010]
49. Alessandra R, Gentile MM, Bianco DM. Who tweets in Italian? Demographic characteristics of twitter users. In: Petrucci A, Racioppi F, Verde R, editors. New Statistical Developments in Data Science. Cham, Switzerland: Springer; Aug 21, 2019.
50. Alfayez A, Awwad Z, Kerr C, Alrashed N, Williams S, Al-Wabil A. Understanding gendered spaces using social media data. In: Meiselwitz G, editor. Social Computing and Social Media. Applications and Analytics. Cham, Switzerland: Springer; 2017.
51. Arafat TA, Budi I, Mahendra R, Salehah DA. Demographic analysis of candidates supporter in Twitter during Indonesian presidential election 2019. In: Proceedings of the International Conference on ICT for Smart Society (ICISS). Presented at: ICISS 2020; November 19-20, 2020; Bandung, Indonesia. [doi: 10.1109/iciss50791.2020.9307598]
52. Ardehaly EM, Culotta A. Co-training for demographic classification using deep learning from label proportions. In: Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW). Presented at: ICDMW 2017; November 18-21, 2017; New Orleans, LA URL: https://ieeexplore.ieee.org/document/8215778 [doi: 10.1109/icdmw.2017.144]
53. Ardehaly EM, Culotta A. Learning from noisy label proportions for classifying online social data. Soc Netw Anal Mining 2017 Nov 27;8(1). [doi: 10.1007/s13278-017-0478-6]

XSL•FO
RenderX

54. Baxevanakis S, Gavras S, Mouratidis D, Kermanidis KL. A machine learning approach for gender identification of Greek tweet authors. In: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments. Presented at: PETRA '20; June 30-July 3, 2020; Corfu, Greece URL: https://dl-acm-org.proxy.library.upenn.edu/doi/10.1145/3389189.3397992 [doi: 10.1145/3389189.3397992]
55. Bayot RK, Gonçalves T. Age and gender classification of Tweets using convolutional neural networks. In: Proceedings of the Third International Conference, MOD 2017. Presented at: MOD 2017; September 14-17, 2017; Volterra, Italy. [doi: 10.1007/978-3-319-72926-8_28]
56. Brandt J, Buckingham K, Buntain C, Anderson W, Ray S, Pool JR, et al. Identifying social media user demographics and topic diversity with computational social science: a case study of a major international policy forum. J Comput Soc Sci 2020 Jan 07;3:167-188. [doi: 10.1007/s42001-019-00061-9]
57. Bsir B, Zrigui M. Enhancing deep learning gender identification with gated recurrent units architecture in social text. Computacion y Sistemas 2018 Sep 30;22(3):757-766. [doi: 10.13053/CyS-22-3-3036]
58. Bsir B, Zrigui M. Document model with attention bidirectional recurrent network for gender identification. In: Proceedings of the 15th International Work-Conference on Artificial Neural Networks. Presented at: IWANN 2019; June 12-14, 2019; Gran Canaria, Spain. [doi: 10.1007/978-3-030-20521-8_51]
59. Cavazos-Rehg PA, Zewdie K, Krauss MJ, Sowles SJ. "No high like a brownie high": a content analysis of edible marijuana tweets. Am J Health Promot 2018 May;32(4):880-886. [doi: 10.1177/0890117116686574] [Medline: 29214836]
60. Cavazos-Rehg PA, Krauss MJ, Costello SJ, Kaiser N, Cahn ES, Fitzsimmons-Craft EE, et al. "I just want to be skinny.": a content analysis of tweets expressing eating disorder symptoms. PLoS One 2019 Jan 16;14(1):e0207506 [FREE Full text] [doi: 10.1371/journal.pone.0207506] [Medline: 30650072]
61. Cesare N, Dwivedi P, Nguyen Q, Nsoesie EO. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. Palgrave Commun 2019 Sep 17;5(1):106 [FREE Full text] [doi: 10.1057/s41599-019-0314-x] [Medline: 32661492]
62. Cesare N, Nguyen Q, Grant C, Nsoesie EO. Social media captures demographic and regional physical activity. BMJ Open Sport Exerc Med 2019 Jul 14;5(1):e000567 [FREE Full text] [doi: 10.1136/bmjsem-2019-000567] [Medline: 31423323]
63. Chakraborty A, Messias J, Benevenuto F, Ghosh S, Ganguly N, Gummadi K. Who makes trends? Understanding demographic biases in crowdsourced recommendations. In: Proceedings of the International AAAI Conference on Web and Social Media. Presented at: ICWSM-17; Montreal, Quebec; May 15-18, 2017. [doi: 10.1609/icwsm.v11i1.14894]
64. Chamberlain BP, Humby C, Deisenroth MP. Probabilistic inference of Twitter users' age based on what they follow. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2017). Presented at: ECML PKDD 2017; September 18-22, 2017; Skopje, Macedonia. [doi: 10.1007/978-3-319-71273-4_16]
65. Cheng JK, Fernandez A, Quindoza RG, Tan S, Cheng C. A model for age and gender profiling of social media accounts based on post contents. In: Neural Information Processing. New York City, NY: Springer International Publishing; 2018.
66. Cornelisse J, Pillai RG. Age inference on Twitter using SAGE and TF-IGM. In: Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval. 2020 Dec Presented at: NLPIR 2020; December 18-20, 2020; Seoul, Republic of Korea. [doi: 10.1145/3443279.3443300]
67. Duong V, Luo J, Pham P, Yang T, Wang Y. The ivory tower lost: how college students respond differently than the general public to the COVID-19 pandemic. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Presented at: ASONAM 2020; December 7-10, 2020; The Hague, The Netherlands. [doi: 10.1109/asonam49781.2020.9381379]
68. ElSayed S, Farouk M. Gender identification for Egyptian Arabic dialect in twitter using deep learning models. Egypt Inform J 2020 Sep;21(3):159-167. [doi: 10.1016/j.eij.2020.04.001]
69. Emmery C, Chrupała G, Daelemans W. Simple queries as distant labels for predicting gender on Twitter. In: Proceedings of the 3rd Workshop on Noisy User-generated Text. Presented at: NUT@EMNLP 2017; September 7, 2017; Copenhagen, Denmark URL: https://github.com/facebookresearch/ [doi: 10.18653/v1/w17-4407]
70. Garcia-Guzman R, Andrade-Ambriz YA, Ibarra-Manzano MA, Ledesma S, Gomez JC, Almanza-Ojeda DL. Trend-based categories recommendations and age-gender prediction for Pinterest and Twitter users. Appl Sci 2020 Aug 28;10(17):5957. [doi: 10.3390/app10175957]
71. Geng L, Zhang K, Wei X, Feng X. Soft biometrics in online social networks: a case study on Twitter user gender recognition. In: Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW). Presented at: WACVW 2017; March 24-31, 2017; Santa Rosa, CA. [doi: 10.1109/wacvw.2017.8]
72. Giannakopoulos O, Kalatzis N, Roussaki I, Papavassiliou S. Gender recognition based on social networks for multimedia production. In: Proceedings of the IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). Presented at: IVMSP 2018; June 10-12, 2018; Aristi Village, Greece. [doi: 10.1109/ivmspw.2018.8448788]
73. Guimaraes RG, Rosa RL, De Gaetano D, Rodriguez DZ, Bressan G. Age groups classification in social network using deep learning. IEEE Access 2017 May 23;5:10805-10816. [doi: 10.1109/access.2017.2706674]

XSL•FO
RenderX

74. Hasanuzzaman M, Dias G, Way A. Demographic word embeddings for racism detection on Twitter. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Presented at: IJCNLP 2017; November 28-30, 2017; Taipei, Taiwan.
75. Hashempour R. A deep learning approach to language-independent gender prediction on Twitter. In: Proceedings of the 2019 Workshop on Widening NLP. Presented at: 2019 Workshop on Widening NLP; July 28, 2019; Florence, Italy. [doi: 10.18653/v1/w17-2901]
76. Hirt R, Kühl N, Satzger G. Cognitive computing for customer profiling: meta classification for gender prediction. Electron Mark 2019 Feb 21;29:93-106. [doi: 10.1007/s12525-019-00336-z]
77. Huang X, Xing L, Dernoncourt F, Paul MJ. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. Presented at: LREC 2020; May 11-16, 2020; Marseille, France.
78. Huang X, Smith MC, Jamison AM, Broniatowski DA, Dredze M, Quinn SC, et al. Can online self-reports assist in real-time identification of influenza vaccination uptake? A cross-sectional study of influenza vaccine-related tweets in the USA, 2013-2017. BMJ Open 2019 Jan 15;9(1):e024018 [FREE Full text] [doi: 10.1136/bmjopen-2018-024018] [Medline: 30647040]
79. Hussein S, Farouk M, Hemayed E. Gender identification of Egyptian dialect in twitter. Egypt Inform J 2019 Jul;20(2):109-116. [doi: 10.1016/j.eij.2018.12.002]
80. Jurgens D, Tsvetkov Y, Jurafsky D. Writer profiling without the writer's text. In: Proceedings of the 9th International Conference, SocInfo 2017. Presented at: SocInfo 2017; September 13-15, 2017; Oxford, UK. [doi: 10.1007/978-3-319-67256-4_43]
81. Kang Y, Wang Y, Zhang D, Zhou L. The public's opinions on a new school meals policy for childhood obesity prevention in the U.S.: a social media analytics approach. Int J Med Inform 2017 Jul;103:83-88. [doi: 10.1016/j.ijmedinf.2017.04.013] [Medline: 28551006]
82. Khandelwal A, Swami S, Akhtar SS, Shrivastava M. Gender prediction in English-Hindi code-mixed social media content: corpus and baseline system. Computacion y Sistemas 2018;22(3) [FREE Full text] [doi: 10.13053/cys-22-4-3061]
83. Kim SM, Xu Q, Qu L, Wan S, Paris C. Demographic inference on Twitter using recursive neural networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Presented at: ACL 2017; July 30-August 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/p17-2075]
84. Kostakos P, Pandya A, Kyriakouli O, Oussalah M. Inferring demographic data of marginalized users in Twitter with computer vision APIs. In: Proceedings of the European Intelligence and Security Informatics Conference (EISIC). Presented at: EISIC 2018; October 24-25, 2018; Karlskrona, Sweden. [doi: 10.1109/eisic.2018.00022]
85. Ljubešić N, Fišer D, Erjavec T. Language-independent gender prediction on Twitter. In: Proceedings of the Second Workshop on NLP and Computational Social Science. Presented at: NLP+CSS 2017; August 3, 2017; Vancouver, Canada. [doi: 10.18653/v1/W17-2901]
86. López-Monroy AP, González FA, Solorio T. Early author profiling on Twitter using profile features with multi-resolution. Expert Syst Appl 2020 Feb;140:112909. [doi: 10.1016/j.eswa.2019.112909]
87. Markov I, Gómez-Adorno H, Posadas-Durán JP, Sidorov G, Gelbukh A. Author profiling with Doc2vec neural network-based document embeddings. In: Proceedings of the 15th Mexican International Conference on Artificial Intelligence. Presented at: MICAI 2016; October 23-28, 2016; Cancún, Mexico. [doi: 10.1007/978-3-319-62428-0_9]
88. Messias J, Vikatos P, Benevenuto F. White, man, and highly followed: gender and race inequalities in Twitter. In: Proceedings of the International Conference on Web Intelligence. Aug Presented at: WI '17; August 23-26, 2017; Leipzig, Germany. [doi: 10.1145/3106426.3106472]
89. Miura R, Hirota M, Kato D, Araki T, Endo M, Ishikawa H. Predicting user gender on social media sites using geographical information. In: Proceedings of the 10th International Conference on Management of Digital EcoSystems. Sep Presented at: MEDES '18; September 25-28, 2018; Tokyo, Japan. [doi: 10.1145/3281375.3281383]
90. Morgan-Lopez AA, Kim AE, Chew RF, Ruddle P. Predicting age groups of Twitter users based on language and metadata features. PLoS One 2017;12(8):e0183537 [FREE Full text] [doi: 10.1371/journal.pone.0183537] [Medline: 28850620]
91. Mueller A, Wood-Doughty Z, Amir S, Dredze M, Lynn Nobles A. Demographic representation and collective storytelling in the me too Twitter hashtag activism movement. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-28 [FREE Full text] [doi: 10.1145/3449181] [Medline: 35295189]
92. Mukherjee S, Bala PK. Gender classification of microblog text based on authorial style. Inf Syst E Bus Manage 2016 Mar 2;15(1):117-138. [doi: 10.1007/s10257-016-0312-0]
93. Imuede J, Raborife M, Ranchod P. Sentiment analysis as an indicator to evaluate gender disparity on sexual violence tweets in South Africa. In: Proceedings of the International SAUPEC/RobMech/PRASA Conference. Presented at: International SAUPEC/RobMech/PRASA Conference; January 29-31, 2020; Cape Town, South Africa. [doi: 10.1109/saupec/robmech/prasa48453.2020.9040955]
94. Pandya A, Oussalah M, Monachesi P, Kostakos P, Loven L. On the use of URLs and hashtags in age prediction of Twitter users. In: Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI). Presented at: IEEE IRI 2018; July 6-9, 2018; Salt Lake City, UT. [doi: 10.1109/iri.2018.00017]

XSL•FO
RenderX

95. Pandya A, Oussalah M, Monachesi P, Kostakos P. On the use of distributed semantics of tweet metadata for user age prediction. Future Gener Comput Syst 2020 Jan;102:437-452. [doi: 10.1016/j.future.2019.08.018]
96. Pizarro J. Profiling bots and fake news spreaders at PAN'19 and PAN'20 : bots and gender profiling 2019, profiling fake news spreaders on Twitter 2020. In: Proceedings of the IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Presented at: DSAA 2020; October 6-9, 2020; Sydney, Australia. [doi: 10.1109/dsaa49011.2020.00088]
97. Reis JC, Kwak H, An J, Messias J, Benevenuto F. Demographics of news sharing in the U.S. Twittersphere. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. Presented at: HT'17; July 4-7, 2017; Prague, Czech Republic. [doi: 10.1145/3078714.3078734]
98. Serfass DG. Assessing situations on social media: temporal, demographic, and personality influences on situation experience. Florida Atlantic University . 2016. URL: https://www.proquest.com/openview/6a5ca1b98806d5f3a2ce7e66e2f358cd/1?pq-origsite=gscholar&cbl=18750 [accessed 2023-03-30]
99. Stevens R, Bonett S, Bannon J, Chittamuru D, Slaff B, Browne SK, et al. Association between HIV-related Tweets and HIV incidence in the United States: infodemiology study. J Med Internet Res 2020 Jun 24;22(6):e17196 [FREE Full text] [doi: 10.2196/17196] [Medline: 32579119]
100. Stevens RC, Brawner BM, Kranzler E, Giorgi S, Lazarus E, Abera M, et al. Exploring substance use Tweets of youth in the United States: mixed methods study. JMIR Public Health Surveill 2020 Mar 26;6(1):e16191 [FREE Full text] [doi: 10.2196/16191] [Medline: 32213472]
101. Swain S, Seeja KR. TWEESENT: a web application on sentiment analysis. In: Proceedings of the Second International Conference on Smart Innovations in Communications and Computational Sciences. 2019 Presented at: ICSICCS-2018; April 28-29, 2018; Indore, India. [doi: 10.1007/978-981-13-2414-7_36]
102. Thelwall M, Thelwall S. Covid-19 tweeting in English: gender differences. Prof De La Inf 2020 May 04;29(3). [doi: 10.3145/epi.2020.may.01]
103. Udayakumar S, Senadeera DC, Yamunarani S, Cheon NJ. Demographics analysis of Twitter users who tweeted on psychological articles and tweets analysis. Procedia Comput Sci 2018;144:96-104. [doi: 10.1016/j.procs.2018.10.509]
104. van der Goot R, Ljubešić N, Matroos I, Nissim M, Plank B. Bleaching text: abstract features for cross-lingual gender prediction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Presented at: ACL 2018; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-2061]
105. Vashisth P, Meehan K. Gender classification using Twitter text data. In: Proceedings of the 31st Irish Signals and Systems Conference (ISSC). Presented at: ISSC 2020; June 11-12, 2020; Letterkenny, Ireland. [doi: 10.1109/issc49989.2020.9180161]
106. Verhoeven B, Škrjanec I, Pollak S. Gender profiling for Slovene Twitter communication: the influence of gender marking, content and style. In: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. Presented at: BSNLP 2017; April 4, 2017; Valencia, Spain. [doi: 10.18653/v1/w17-1418]
107. Vicente M, Batista F, Carvalho JP. Gender detection of Twitter users based on multiple information sources. In: Kóczy L, Medina-Moreno J, Ramírez-Poussa E, editors. Interactions Between Computational Intelligence and Mathematics Part 2. Cham, Switzerland: Springer; Nov 03, 2018.
108. Vijayaraghavan P, Vosoughi S, Roy D. Twitter demographic classification using deep multi-modal multi-task learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Presented at: ACL 2017; July 30-August 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/p17-2076]
109. Vikatos P, Messias J, Miranda M, Benevenuto F. Linguistic diversities of demographic groups in Twitter. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. Presented at: HT'17; July 4-7, 2017; Prague, Czech Republic. [doi: 10.1145/3078714.3078742]
110. Volkova S. Predicting demographics and affect in social networks. The Johns Hopkins University. 2015 Oct. URL: https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/39639/VOLKOVA-DISSERTATION-2015.pdf?sequence=1&isAllowed=y [accessed 2023-03-30]
111. Wang Y, Feng Y, Luo J. Gender politics in the 2016 U.S. Presidential election: a computer vision approach. In: Proceedings of the 10th International Conference, SBP-BRiMS 2017. Presented at: SBP-BRiMS 2017; July 5-8, 2017; Washington, DC. [doi: 10.1007/978-3-319-60240-0_4]
112. Wang Z, Hale S, Adelani DI, Grabowicz P, Hartman T, Flöck F, et al. Demographic inference and representative population estimates from multilingual social media data. In: Proceedings of the The World Wide Web Conference. Presented at: WWW '19; May 13-17, 2019; San Francisco, CA. [doi: 10.1145/3308558.3313684]
113. Wong SC, Teh PL, Cheng CB. How different genders use profanity on Twitter? In: Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis. Presented at: ICCDA 2020; March 9-12, 2020; Silicon Valley, CA. [doi: 10.1145/3388142.3388145]
114. Wood-Doughty Z, Smith M, Broniatowski D, Dredze M. How does Twitter user behavior vary across demographic groups? In: Proceedings of the Second Workshop on NLP and Computational Social Science. Presented at: NLP+CSS 2017; August 3, 2017; Vancouver, Canada. [doi: 10.18653/v1/w17-2912]
115. Wood-Doughty Z, Andrews N, Marvin R, Dredze M. Predicting Twitter user demographics from names alone. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. Presented at: PEOPLES 2018; June 6, 2018; New Orleans, LA. [doi: 10.18653/v1/w18-1114]

XSL•FO
RenderX

116. Xiang L, Sang J, Xu C. Demographic attribute inference from social multimedia behaviors: a cross-OSN approach. In: Proceedings of the 23rd International Conference, MMM 2017. Presented at: MMM 2017; January 4-6, 2017; Reykjavik, Iceland. [doi: 10.1007/978-3-319-51811-4_42]
117. Yang YC, Al-Garadi MA, Love JS, Perrone J, Sarker A. Automatic gender detection in Twitter profiles for health-related cohort studies. JAMIA Open 2021 Jun 23;4(2):ooab042 [FREE Full text] [doi: 10.1093/jamiaopen/ooab042] [Medline: 34169232]
118. Yazdavar AH, Mahdavinejad MS, Bajaj G, Romine W, Sheth A, Monadjemi AH, et al. Multimodal mental health analysis in social media. PLoS One 2020 Apr 10;15(4):e0226248 [FREE Full text] [doi: 10.1371/journal.pone.0226248] [Medline: 32275658]
119. Yildiz D, Munson J, Vitali A, Tinati R, Holland JA. Using Twitter data for demographic research. Demogr Res 2017 Nov 22;37:1477-1514. [doi: 10.4054/demres.2017.37.46]
120. Zhang C, Xu S, Li Z, Hu S. Understanding concerns, sentiments, and disparities among population groups during the COVID-19 pandemic via Twitter data mining: large-scale cross-sectional study. J Med Internet Res 2021 Mar 05;23(3):e26482 [FREE Full text] [doi: 10.2196/26482] [Medline: 33617460]
121. Zhao Y, Zhang H, Huo J, Guo Y, Wu Y, Prosperi M, et al. Mining Twitter to assess the determinants of health behavior towards palliative care in the United States. AMIA Jt Summits Transl Sci Proc 2020:730-739 [FREE Full text] [Medline: 32477696]
122. Cesare N, Grant C, Hawkins JB, Brownstein JS, Nsoesie EO. Demographics in social media data for public health research: does it matter? arXiv. Preprint posted online October 30, 2017 2023 [FREE Full text] [doi: 10.48550/arXiv.1710.11048]
123. Rangel F, Rosso P, Koppel M, Stamatatos E, Inches G. Overview of the author profiling task at PAN 2013. In: Proceedings of the CLEF Conference on Multilingual and Multimodal Information Access Evaluation. Presented at: CLEF 2013; September 23-26, 2013; Valencia, Spain URL: https://riunet.upv.es/handle/10251/46636
124. Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, et al. Overview of the 2nd author profiling task at pan 2014. In: Proceedings of the 2014 Computer Science Workshops. Presented at: CEUR-WS '14; March 18, 2014; Sheffield, UK URL: https://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-RangelEt2014.pdf
125. Rangel F, Celli F, Rosso P, Potthast M, Stein B, Daelemans W. Overview of the 3rd author profiling task at PAN 2015. In: Proceedings of the CLEF 2015 Conference and Labs of the Evaluation Forum. Presented at: CLEF 2015; September 8-11, 2015; Toulouse, France.
126. Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Proceedings of the Conference and Labs of the Evaluation Forum. Presented at: CLEF 2016; September 5-8, 2016; Evora, Portugal.
127. Rangel F, Rosso P, Potthast M, Stein B. Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter. In: Proceedings of the Conference and Labs of the Evaluation Forum. Presented at: CLEF 2017; September 11-14, 2017; Dublin, Ireland.
128. Rangel F, Rosso P, Montes-y-Gómez M, Potthast M, Stein B. Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in Twitter. In: Proceedings of the Conference and Labs of the Evaluation Forum. Presented at: CLEF 2018; September 10-14, 2018; Avignon, France.
129. Rangel F, Rosso P. Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in Twitter. In: Proceedings of the Conference and Labs of the Evaluation Forum. Presented at: CLEF 2019; September 9-12, 2019; Lugano, Switzerland.
130. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating gender on Twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Presented at: EMNLP 2011; July 27-31, 2011; Edinburgh, UK URL: https://aclanthology.org/D11-1120/
131. Volkova S, Yarowsky D. Improving gender prediction of social media users via weighted annotator rationales. Johns Hopkins University. URL: https://hltcoe.jhu.edu/wp-content/uploads/2016/11/17310_slides.pdf [accessed 2023-03-30]
132. Volkova S, Wilson T, Yarowsky D. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Presented at: EMNLP 2013; October 18-21, 2013; Seattle, WA.
133. Liu W, Ruths D. What's in a name? Using first names as features for gender inference in Twitter. In: Proceedings of the AAAI 2013 Spring Symposium Series. Presented at: AAAI 2013 Spring Symposium Series; March 25-27, 2013; Palo Alto, CA URL: https://aaai.org/papers/05744-whats-in-a-name-using-first-names-as-features-for-gender-inference-in-twitter/
134. Plank B, Hovy D. Personality traits on Twitter—or—how to get 1,500 personality tests in a week. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Presented at: WASSA 2015; September 17, 2015; Lisboa, Portugal. [doi: 10.18653/v1/w15-2913]
135. Verhoeven B, Daelemans W, Plank B. TwiSty: a multilingual twitter stylometry corpus for gender and personality profiling. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Presented at: LREC'16; May 23-28, 2016; Portorož, Slovenia URL: http://www.clips.uantwerpen.be/
136. Chauhan A. Gender classification dataset. Kaggle. URL: https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset [accessed 2022-10-17]

XSL•FO
RenderX

137. Radford J. Piloting a theory-based approach to inferring gender in big data. In: Proceedings of the IEEE International Conference on Big Data (Big Data). Presented at: IEEE BigData 2017; December 11-14, 2017; Boston, MA. [doi: 10.1109/bigdata.2017.8258555]
138. Pizarro J. Using N-grams to detect Bots on Twitter Notebook for PAN at CLEF 2019. In: Proceedings of the Conference and Labs of the Evaluation Forum. Presented at: CLEF 2019; September 9-12, 2019; Lugano, Switzerland.
139. Knowles R, Carroll J, Dredze M. Demographer: extremely simple name demographics. In: Proceedings of the First Workshop on NLP and Computational Social Science. Presented at: NLP+CSS 2016; November 5, 2016; Austin, Texas. [doi: 10.18653/v1/w16-5614]
140. Volkova S, Coppersmith G, Van Durme B. Inferring user political preferences from streaming communications. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Presented at: ACL 2014; June 22-27, 2014; Baltimore, MD. [doi: 10.3115/v1/p14-1018]
141. Sap M, Park G, Eichstaedt JC, Kern ML, Stillwell D, Kosinski M, et al. Developing age and gender predictive lexica over social media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Presented at: EMNLP 2014; October 25-29, 2014; Doha, Qatar URL: https://aclanthology.org/D14-1121.pdf [doi: 10.3115/v1/d14-1121]
142. Zhou E, Fan H, Cao Z, Jiang Y, Yin Q. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. Presented at: ICCV 2013; December 02-08, 2013; Sydney, Australia URL: https://ieeexplore.ieee.org/document/6755923 [doi: 10.1109/iccvw.2013.58]
143. Wood-Doughty Z, Xu P, Liu X, Dredze M. Using noisy self-reports to predict Twitter user demographics. arXiv. Preprint posted online May 1, 2020 2023 [FREE Full text] [doi: 10.18653/v1/2021.socialnlp-1.11]
144. Rothe R, Timofte R, Van Gool L. Deep expectation of real and apparent age from a single image without facial landmarks. Int J Comput Vis 2016 Aug 10;126(2-4):144-157. [doi: 10.1007/s11263-016-0940-3]
145. Wang R, Chaudhari P, Davatzikos C. Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies. Proc Natl Acad Sci U S A 2023 Feb 07;120(6):e2211613120 [FREE Full text] [doi: 10.1073/pnas.2211613120] [Medline: 36716365]
146. Geifman N, Cohen R, Rubin E. Redefining meaningful age groups in the context of disease. Age (Dordr) 2013 Dec;35(6):2357-2366 [FREE Full text] [doi: 10.1007/s11357-013-9510-6] [Medline: 23354682]
147. Sera LC, McPherson ML. Pharmacokinetics and pharmacodynamic changes associated with aging and implications for drug therapy. Clin Geriatr Med 2012 May;28(2):273-286. [doi: 10.1016/j.cger.2012.01.007] [Medline: 22500543]
148. Gonzalez-Hernandez G, Weissenbacher D. Proceedings of the seventh workshop on social media mining for health applications, workshop and shared task. In: Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. Presented at: SMM4H '22; October 12-17, 2022; Gyeongju, Republic of Korea URL: https://aclanthology.org/2022.smm4h-1.0/ [doi: 10.18653/v1/w18-5904]
149. Distribution of Twitter users worldwide as of January, 2021, by gender. Statista. URL: https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/
150. Mauvais-Jarvis F, Bairey Merz N, Barnes PJ, Brinton RD, Carrero J, DeMeo DL, et al. Sex and gender: modifiers of health, disease, and medicine. Lancet 2020 Aug;396(10250):565-582. [doi: 10.1016/s0140-6736(20)31561-0]
151. Klein AZ, Magge A, Gonzalez-Hernandez G. ReportAGE: automatically extracting the exact age of Twitter users based on self-reports in tweets. PLoS One 2022;17(1):e0262087 [FREE Full text] [doi: 10.1371/journal.pone.0262087] [Medline: 35077484]
152. Sloan L, Morgan J, Housley W, Williams M, Edwards A, Burnap P, et al. Knowing the tweeters: deriving sociologically relevant demographics from Twitter. Sociological Res Online 2013 Aug 31;18(3):74-84. [doi: 10.5153/sro.3001]
153. Sloan L. Who tweets in the United Kingdom? Profiling the Twitter population using the British social attitudes survey 2015. Soc Media Soc 2017 Mar 22;3(1):205630511769898. [doi: 10.1177/2056305117698981]
154. Jung S, An J, Kwak H, Salminen J, Jansen B. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. Proc Int AAAI Conf Web Soc Media 2018 Jun 15;12(1) [FREE Full text] [doi: 10.1609/icwsm.v12i1.15058]
155. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016 Jan;69:245-247 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
156. Siontis GC, Ioannidis JP. Response to letter by Forike et al.: more rigorous, not less, external validation is needed. J Clin Epidemiol 2016 Jan;69:250-251. [doi: 10.1016/j.jclinepi.2015.01.021] [Medline: 25724895]
157. Borkotoky K, Unisa S. Indicators to examine quality of large scale survey data: an example through district level household and facility survey. PLoS One 2014;9(3):e90113 [FREE Full text] [doi: 10.1371/journal.pone.0090113] [Medline: 24598760]
158. Basannar DR, Singh S, Yadav J, Yadav AK. Quantifying age heaping and age misreporting in a multicentric survey. Indian J Community Med 2022;47(1):104-106 [FREE Full text] [doi: 10.4103/ijcm.ijcm_1179_21] [Medline: 35368490]
159. Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: a systematic review. Procedia Comput Sci 2019;161:466-474. [doi: 10.1016/j.procs.2019.11.146]

XSL•FO
RenderX

160. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal 2020 Oct;65:101759 [FREE Full text] [doi: 10.1016/j.media.2020.101759] [Medline: 32623277]
161. Golder S, Scantlebury A, Christmas H. Understanding public attitudes toward researchers using social media for detecting and monitoring adverse events data: multi methods study. J Med Internet Res 2019 Aug 29;21(8):e7081 [FREE Full text] [doi: 10.2196/jmir.7081] [Medline: 31469079]
162. Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation. Sociology 2017 Dec 26;51(6):1149-1168 [FREE Full text] [doi: 10.1177/0038038517708140] [Medline: 29276313]
163. Singh L, Polyzou A, Wang Y, Farr J, Gresenz CR. Social media data - our ethical conundrum. Bulletin of the IEEE Computer Society Technical Committee on Data Engineerin. 2020. URL: http://sites.computer.org/debull/A20dec/p23.pdf [accessed 2023-03-30]
164. Valdez R, Keim-Malpass J. Ethics in health research using social media. In: Bian J, Guo Y, He Z, Hu X, editors. Social Web and Health Research. Cham, Switzerland: Springer; 2019.
165. Benton A, Coppersmith G, Dredze M. Ethical research protocols for social media health research. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Presented at: EthNLP@EACL; April 4, 2017; Valencia, Spain. [doi: 10.18653/v1/w17-1612]
166. McKee R. Ethical issues in using social media for health and health care research. Health Policy 2013 May;110(2-3):298-301. [doi: 10.1016/j.healthpol.2013.02.006] [Medline: 23477806]

Abbreviations

API: application programming interface
CLEF: Conference and Labs of the Evaluation Forum
DNN: deep neural network
ML: machine learning
PICOS: Population, Intervention, Comparison, Outcomes, and Study Design
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SMM4H: Social Media Mining for Health
SVM: support vector machine

Edited by A Mavragani; submitted 05.04.23; peer-reviewed by B Ru, C Ni; comments to author 05.06.23; revised version received 28.07.23; accepted 01.08.23; published 15.03.24.

Please cite as:
O'Connor K, Golder S, Weissenbacher D, Klein AZ, Magge A, Gonzalez-Hernandez G
Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review
J Med Internet Res 2024;26:e47923
URL: https://www.jmir.org/2024/1/e47923
doi:10.2196/47923
PMID:38488839

©Karen O'Connor, Su Golder, Davy Weissenbacher, Ari Z Klein, Arjun Magge, Graciela Gonzalez-Hernandez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Reducing Loneliness and Social Isolation of Older Adults Through Voice Assistants: Literature Review and Bibliometric Analysis

Rachele Alessandra Marziali[1*], MSc; Claudia Franceschetti[1*], MEng; Adrian Dinculescu[2*], PhD; Alexandru Nistorescu[2*], PhD; Dominic Mircea Kristály[3*], PhD; Adrian Alexandru Mo oi[4*], MSc; Ronny Broekx[5*], BSB; Mihaela Marin[2*], MSc; Cristian Vizitiu[2,3*], PhD; Sorin-Aurel Moraru[3*], PhD; Lorena Rossi[1*], MEng; Mirko Di Rosa[6*], PhD

[1]Centre for Innovative Models for Aging Care and Technology, IRCCS INRCA-National Institute of Health and Science on Aging, Ancona, Italy

[2]The Space Applications and Technologies Laboratory, Institute of Space Science – Subsidiary of INFLPR (National Institute for Laser, Plasma and Radiation Physics), Magurele, Romania

[3]Department of Automatics and Information Technology, Faculty of Electrical Engineering and Computer Science, Transilvania University of Brasov, Brasov, Romania

[4]Department of Psychology and Education Sciences, Faculty of Psychology and Education Sciences, Transilvania University of Brasov, Brasov, Romania

[5]Innovation Department, ePoint, Hamont, Belgium

[6]Centre for Biostatistics and Applied Geriatric Clinical Epidemiology, IRCCS INRCA-National Institute of Health and Science on Aging, Ancona, Italy

[*]all authors contributed equally

Corresponding Author:
Claudia Franceschetti, MEng
Centre for Innovative Models for Aging Care and Technology, IRCCS INRCA-National Institute of Health and Science on Aging
Via Santa Margherita 5
Ancona, 60124
Italy
Phone: 39 0718004788
Email: c.franceschetti@inrca.it

Abstract

Background: Loneliness and social isolation are major public health concerns for older adults, with severe mental and physical health consequences. New technologies may have a great impact in providing support to the daily lives of older adults and addressing the many challenges they face. In this scenario, technologies based on voice assistants (VAs) are of great interest and potential benefit in reducing loneliness and social isolation in this population, because they could overcome existing barriers with other digital technologies through easier and more natural human-computer interaction.

Objective: This study aims to investigate the use of VAs to reduce loneliness and social isolation of older adults by performing a systematic literature review and a bibliometric cluster mapping analysis.

Methods: We searched PubMed, Embase, and Scopus databases for articles that were published in the last 6 years, related to the following main topics: voice interface, VA, older adults, isolation, and loneliness. A total of 40 articles were found, of which 16 (40%) were included in this review. The included articles were then assessed through a qualitative scoring method and summarized. Finally, a bibliometric analysis was conducted using VOSviewer software (Leiden University's Centre for Science and Technology Studies).

Results: Of the 16 articles included in the review, only 2 (13%) were considered of poor methodological quality, whereas 9 (56%) were of medium quality and 5 (31%) were of high quality. Finally, through bibliometric analysis, 221 keywords were extracted, of which 36 (16%) were selected. The most important keywords, by number of occurrences and by total link strength; results of the analysis with the Association Strength normalization method; and default values were then presented. The final bibliometric network consisted of 36 selected keywords, which were grouped into 3 clusters related to 3 main topics (ie, VA use for social isolation among older adults, the significance of age in the context of loneliness, and the impact of sex factors on well-being). For most of the selected articles, the effect of VA on social isolation and loneliness of older adults was a minor theme. However, more investigations were done on user experience, obtaining preliminary positive results.

Conclusions: Most articles on the use of VAs by older adults to reduce social isolation and loneliness focus on usability, acceptability, or user experience. Nevertheless, studies directly addressing the impact that using a VA has on the social isolation

XSL•FO
RenderX

and loneliness of older adults find positive and promising results and provide important information for future research, interventions, and policy development in the field of geriatric care and technology.

(J Med Internet Res 2024;26:e50534) doi:10.2196/50534

KEYWORDS

voice assistant; loneliness; social isolation; older adults; literature review; bibliometric analysis; mobile phone

Introduction

Background

Nowadays, the aging of the population presents new challenges that requires consideration and response [1]. Among the major public health concerns regarding older adults, 2 significant concerns are loneliness and social isolation [2].

In fact, social networks seem to decrease with age and the prevalence of loneliness is estimated to increase as the population ages [2], to the extent that Valtorta and Hanratty [3] define loneliness and isolation as being "increasingly part of the experience of growing old."

Social isolation and loneliness have severe consequences for older adults' mental and physical health, including depressive symptoms [4], dementia [5], coronary heart disease and stroke [6], and mortality [7]. Moreover, social isolation and loneliness also have adverse outcomes concerning the use of health services, increasing emergency department and physician visits, hospital readmissions, and long-term care admissions [8].

New technologies may have a great impact on providing support in the daily lives of older people, especially in the areas of health monitoring, security, and comfort [9]. Therefore, they could be valuable tools to respond to the many challenges that older adults face.

In this scenario, technologies based on voice assistants (VAs) are of great interest and have potential benefits. VAs are systems based on artificial intelligence techniques that are programmed to be activated at a specific wake word to capture the user's voice, process and interpret the command via a server, and respond back with a voice response or completed task [10].

VA systems have the potential to support behavioral interventions using everyday life technologies such as smartphones, tablets, and smart speakers [9]. The strength behind the use of voice-based technology, having reached a significant stage of maturity, is strictly related to the concept of *ubiquitous computing* (Figure 1), introduced by Weiser in 1991 when thinking about a paradigm of technology able to adapt to the human environment that *vanish in the background* [11]. Indeed, VA technology is physically intangible; it does not force the user to be physically at a particular place to operate, and it provides interaction using natural language [9].

XSL•FO
RenderX

Figure 1. The functioning of ubiquitous computing. VR: virtual reality.

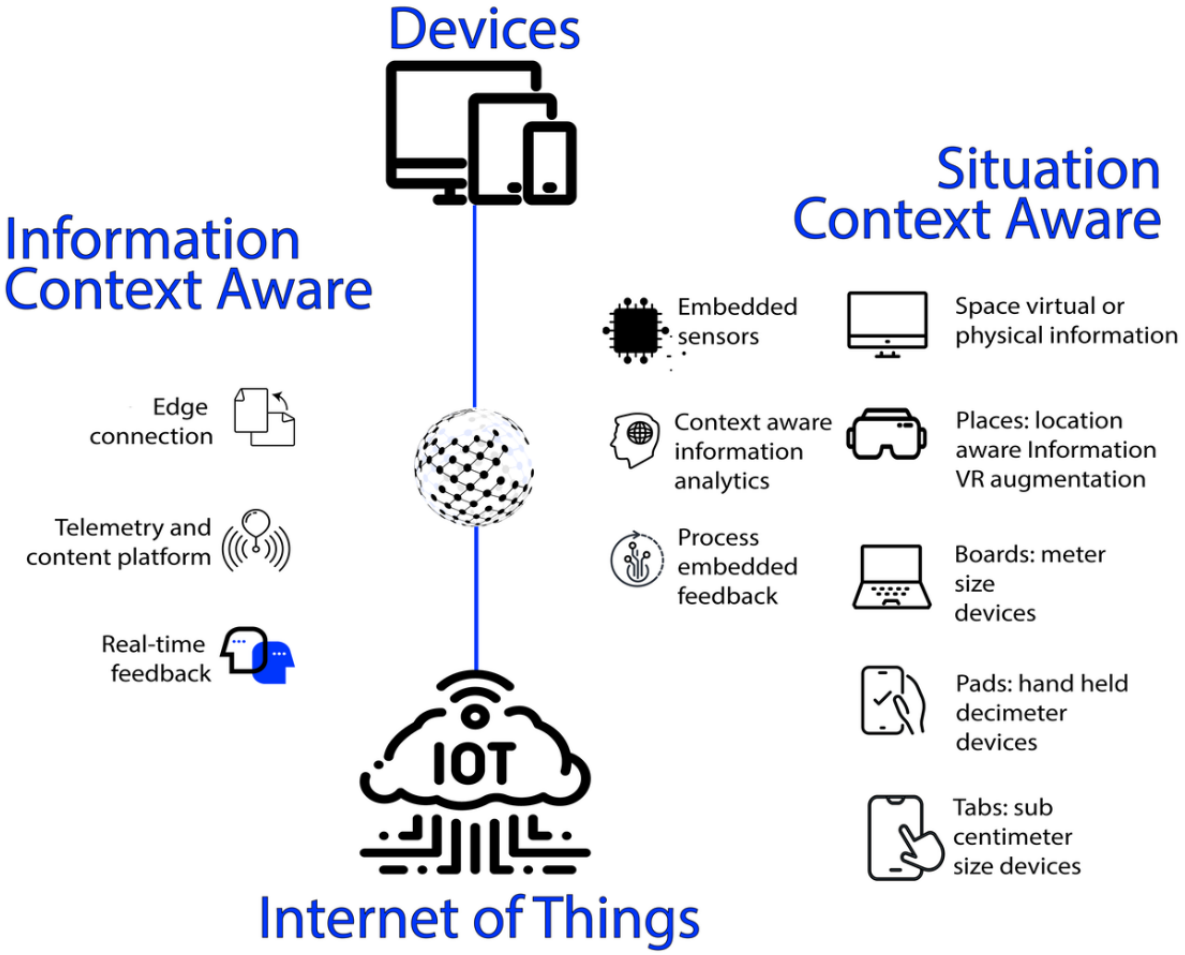

Concerning the application to older people, this easy and natural human-computer interaction gives VA systems the potential to overcome possible barriers existing with other digital technologies, which appears particularly promising and appropriate [9].

Objectives

In light of this, the objective of this study is to investigate the use of VAs to reduce loneliness and social isolation of older adults by performing a literature review and a bibliometric analysis.

Methods

Database Creation

A literature search of scientific articles published from January 1, 2018, to April 4, 2023, was conducted. Considering that VA technology had not reached a significant stage of maturity, especially in its application for social purposes, this time range was defined.

The PubMed, Embase, and Scopus databases were searched to extend the range of eligible articles. In particular, the search was performed by setting up the “Title/Abstract” field in PubMed, the “Title or Abstract” field in Embase, and the “Title, Abstract, Keywords” field in Scopus.

The search was performed using an appropriate sequence of keywords, based on the research objectives. The first part of the search string was focused on synonyms for VA, whereas the second part specified the application for isolation and loneliness in older adults. The search string used was as follows: ((voice interface) OR (voice assistant) OR (vocal interface) OR (vocal assistant) OR (speech agent) OR (vocal agent)) AND (olde* OR elder*) AND (isolation OR loneliness).

We collected a total of 40 publications: 34 from Scopus, 4 from PubMed, and 2 from Embase.

Study Selection

The selection of the eligible studies was performed according to the following principles:

1. Including only publications in English language: no documents were excluded.
2. Removal of overlaps between the different databases: 3 overlapping documents were identified.
3. Excluding papers in which the title and abstract were not relevant to the research question: 12 papers were excluded.
4. Removal of articles not retrieved: 1 article was excluded.
5. Excluding articles not pertinent to the research question: 8 documents were excluded.

The studies were assessed independently by 3 authors (CF, RAM, and AD). Any disagreement and uncertainties in the study selection were resolved by discussion. In particular, 2 authors conducted the first assessment, and another one solved the divergences.

XSL•FO
RenderX

Multimedia Appendix 1 [12-19] reports the list of excluded articles concerning eligibility assessment and details about the motivations for their exclusion.

The final database was composed of 40% (16/40) of the collected documents.

Figure 2 reports the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram [20], summarizing the identification, screening, and inclusion procedures performed.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. VA: voice assistant.

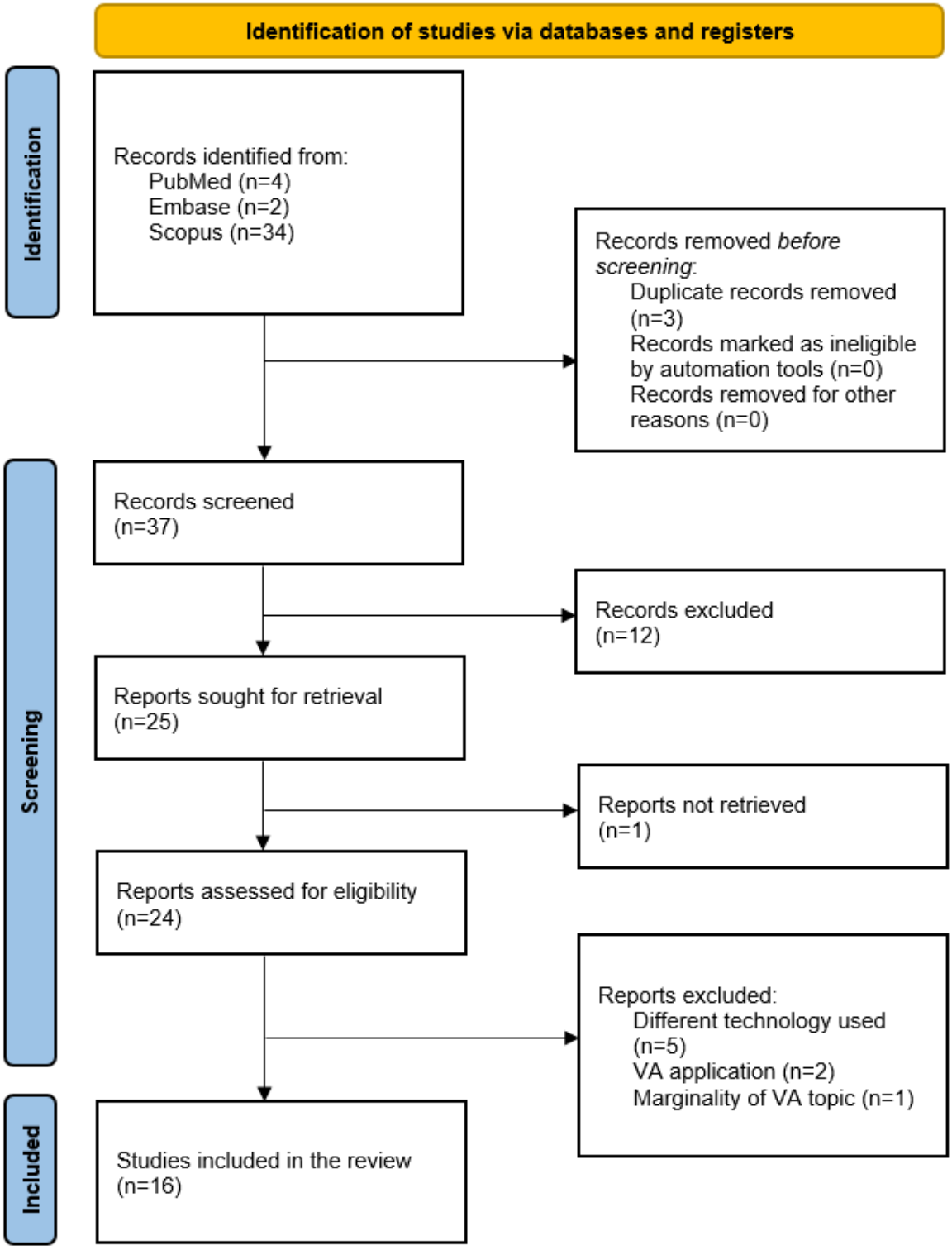

XSL•FO
RenderX

Quality Scoring

As systematic reviews are comprehensive and rigorous assessments of existing literature on a specific research question and they aim to synthesize the available evidence to provide a reliable and unbiased summary, the "Tool for Scoring Quality of Non-Empirical Data Sources" [21], owned by the Aerospace Medicine Systematic Review Group, was used to assess the quality of individual studies included in this review. In total, 2 authors (RAM and CF) performed this evaluation independently, solving any disagreements or doubts through discussion. It is important to note that the purpose of quality scoring in systematic reviews is not to exclude studies but rather to provide an evaluation of their methodological strengths and weaknesses. The scoring process helps reviewers assess the overall risk of bias in the body of evidence and inform their conclusions and recommendations.

Data Extraction

To perform the synthesis of findings, a data extraction from the 16 selected articles was conducted. The extraction consisted of a further evaluation of the full text of the articles. In total, 2 authors (MDR and CF) independently extracted information from the selected studies, including reference, population, technological solution, environment, study design, outcomes, and main results. The assessors made the information homogeneous and analyzed the articles together in the case of doubts or missing data. The data extracted were reported in the corresponding section of the synthesis of findings table (Table 1).

XSL•FO
RenderX

Table 1. Synthesis of findings.

Study, year	Population	Technological solution	Environment	Study design	Outcomes	Main results
Balasubramanian et al [22], 2021	A total of 44 adults (2 types of cohorts: one specifically focused on diabetes and the other on a range of long-term health conditions such as multiple sclerosis, dementia, and depression) and 7 informal carers, with age ranged from 50 to 90 y	VA[a] commercial device (Alexa Echo Show 8)	Home	Qualitative (service evaluation and market research for a pilot service redesign program)	Acceptance and user experience about assistive technology for health and social well-being	Positive impact on the health and social well-being of the users; many direct and indirect benefits were identified: reminders for medications and appointments; improved adherence and disease control; increased independence and productivity; and, for those living alone, the device helped combat their loneliness and low mood
Bravo et al [23], 2020	A total of 10 older adults (aged 60 to 77 y)	VA embedded in a mobile app	Retirement home or home with their family	Quantitative (evaluation test)	VA evaluation: naturalness, embodiment, interaction and affect, joy of use, ease of use, acceptability, and utility (5-point Likert scale)	Average satisfaction score of 3.98 on a scale of 1 (lowest) to 5 (highest; naturalness: 3.8, embodiment: 3.73, interaction and affect: 4.18, joy of use: 3.99, ease of use: 4.03, acceptability: 3.68, and utility 4.29)
Caselgrandi et al [24], 2021	A total of 39 adults aged >50 y (mean 63; 78% male adults) with PACS[b], previously hospitalized for severe COVID pneumonia	VA commercial device (Google Nest)	Home	Quantitative (pre-post study)	Assessment of clusters of PACS symptoms, frailty, and HRQoL[c]: depression, anxiety, and stress (DASS-21[d]); resilience (CD-RISC-25[e]), frailty (Reference Sites Network for Prevention and Care of Frailty and Chronic Conditions in community dwelling persons of EU Countries—SUNFRAIL), quality of life (EQ-5D-5L[f]), insomnia (ISI[g]), health (SF-36[h]); frailty phenotype; assessment for sarcopenia (hand grip measured with dynamometer); satisfaction with the VA tool; patients' empowerment: proportion of people acquiring a normal caloric and normal protein diet, reduction of sedentary life, and engagement with moderate and vigorous physical activity	PACS cluster symptoms, frailty, and HRQoL improved at 6 mo follow-up; 96% of the participants considered VA useful; 44% used VA for entertainment and to cope with loneliness; and 48% of the participants modified sedentary life
Corbett et al [25], 2021	Older adults	VAs Commercial devices (eg, Amazon Echo and Google Nest)	Home	Mini review	Social isolation and loneliness	VAs are perceived by many older adult users as "companions" and improve social connectedness and reduce loneliness

XSL•FO
RenderX

Study, year	Population	Technological solution	Environment	Study design	Outcomes	Main results
Farías-Barraza et al [26], 2022	A total of 1 adult (female; alpha tests) and 6 older adults aged >65 y (3 female adults and 3 male adults) with no experience with VA technology (beta tests)	VA with open-source software and available technologies (PC application)	Laboratory	Quantitative (development tests [alpha tests] and users test [beta tests])	Right answer ratio and percentage for all 20 commands (alpha tests). Functions success percentage; key indicators of success: language, right answers, app cost, and 20 basic existing commands (beta tests)	Alpha test: in all the commands, right answer ratio percentage >75% (lowest percentages given in "Timer" and "Conversation" functions). Beta test: language, app cost, 20 basic existing commands and right answers: achieved (average number of correct answers 86.9% and conversation function within the fewest values)
García-Méndez et al [27], 2021	A total of 31 older adults (20 female adults and 11 male adults; mean 75.5, SD 6.95 y; 10 had some basic technology skills [such as experience with Google or WhatsApp]) and 8 had hearing problems	VA developed as a chatbot (tablet) implementing Google Voice Android Software Development Kit	Laboratory	Mixed (evaluation test)	Users' experience: satisfaction, amazement, and chatbot-human likeliness (5-level Likert scale); users' ability to describe the content presented (NGD[i]) between news metadata and the words users explicitly chose to search for news; knowledge about the news in the dialogue between the chatbot and the users (NGD); and liked or disliked aspects	Users' satisfaction with the service close to 4 and the perception of chatbot-human likeliness close to 3 on average. Describing presented content, focused users with technology skills selected effective words (NGD=0.56), whereas stressed or confused users provided vague terms (NGD=0.84). Visual indications of the user's turn to speak, chatbot empathy, frontend avatar, and newscaster functionality were praised. Confused users were particularly baffled by chatbot interruptions when they paused for too long
Jones et al [28], 2021	A total of 16 older adults (69% female adults and 31% male adults) aged >75 y (mean 85.2, SD 5.02 y) with normative cognitive functioning	VA commercial device (Amazon Echo)	Independent living facility	Mixed (single-group quasi-experimental study)	Perceived loneliness (8-item University of California, Los Angeles Loneliness Scale) and anthropomorphic aspects of the interactions (relational greetings, comments and questions, polite interactions, and reaction)	Significant reductions in perceived loneliness after 4 wk of using the VA. Relational greetings to the Alexa VA predicted 4-wk loneliness reductions, whereas the number of reactions, polite interactions, or comments and questions did not significantly predict 4-wk loneliness reductions
O'Brien et al [29], 2022	A total of 16 adults (14 female adults and 2 male adults): 11 geriatric experts aged >21 y and 5 older adults aged >65 y	VA Commercial device (Google Home)	Home	Qualitative (evaluation test)	Experience with the device; codes and overarching themes	A total of 288 comments were received from which 8 major themes were identified as possible beneficial functions of VA: administrative, companionship, home control, education, emergencies, entertainment, health and well-being, and reminders

XSL•FO
RenderX

Study, year	Population	Technological solution	Environment	Study design	Outcomes	Main results
Pech et al [30], 2022	A total of 109 adults (86 female adults and 23 male adults) aged >50 (mean 81.2, SD 8.6) y with no severe visual or hearing impairment and no moderate to severe cognitive impairment	Digital intelligent platform available on smartphone, tablet, or computer and a VA	Home	Mixed (pre-post study)	System use and acceptability, service use and satisfaction, intervention global perception, system improvement, and operational team's feedback	A total of 39% used the services at least once; 63% had a positive opinion toward the system; 22% had a positive opinion on the intervention, 55% a mixed opinion and 23% a negative one; proposed improvements were easy access to trusted professionals, communication about city events, late-night pharmacy, activity propositions tailored to their needs, and videoconferencing option; and the team emphasized older adults' resistance to change, unplanned workload, and specific technological obstacles
Pérès et al [31], 2021	Older adults with a digital device	Digital intelligent platform available on smartphone, tablet, or computer and a VA	Home	Study protocol	Program impact and effectiveness; technical use; intervention mechanisms, transferability, and scalability conditions; health care consumption and outcomes; perceived social support; quality of life; loneliness; participation; sense of usefulness; self-esteem; frailty; activity limitation; program's impact on health and care trajectories	N/A[j]
Pradhan et al [32], 2019	A total of 7 older adults (6 female adults and 1 male) aged >65 (mean 71.7) y with low technology use	VA commercial device (Amazon Echo Dot) with a paired tablet (Amazon Fire tablet)	Home or older adult living community	Qualitative (pre-post study)	Initial perceptions of VA technology, device perceptions and use, technology desired use, daily diary entries, voice commands use logs, and participants' ontological categorization of VAs	Inconsistent use of personal pronouns for VA, users polite behaviors, users' perceptions of VA did not clearly classify into "humanlike" or "object-like," and VA role moving through different ontological categories
Razavi et al [33], 2022	A total of 19 older adults (13 female adults and 6 male adults) aged >60 (mean 71) y, with mild difficulties on social skills, depression and anxiety symptoms, and nonverbal impairment	Web-based automated version of a VA designed to improve communication skills	Home	Mixed (randomized controlled trial)	Verbal and nonverbal behavior in social communication (SSPA[k]); dialogue content	Significant improvement in eye contact and facial expressivity, users on average tend to provide longer responses as they proceed in a conversation, topic classes significantly affect users' response length, and user sentiment significantly more positive for some topics than others
Reis et al [34], 2018	Older adults	VA commercial devices (Google Assistant, Amazon Alexa, Apple Siri, and Microsoft Cortana)	Older adult care center	Quantitative (VAs test)	VAs performance on acknowledgment, engagement, effectiveness, usefulness, and follow-up in 4 types of interaction (basic greeting, email management, social media, and social games)	VAs obtained good results in the acknowledgment and engagement, mixed results in effectiveness and usefulness, and bad results in follow-up (except for social games)

XSL•FO
RenderX

Study, year	Population	Technological solution	Environment	Study design	Outcomes	Main results
Simpson et al [35], 2020	Medical community, 2 older adults, adults	VA embodied as a household potted flower	Home or retirement home	Conference speech on design-thinking approach	Physical and mental health challenges that a VA could help mitigate; older adults' everyday life, challenges, social interaction, thoughts on Vas, and their possible use; and device approachability and improvement	Development of device prototype
Striegl et al [36], 2021	A total of 9 older adults (2 female adults and 7 male adults) visiting day-care centers, without dementia (test with older adult participants); 4 professionals caregivers working in a day-care facility (2 female adults and 2 male adults) with age ranging from 33 to 61 y who had experience in caring for people with dementia	VA developed using the Amazon Alexa platform and Alexa Voice Services	Day-care facility	Mixed (usability test)	System usability, concept feasibility (5-point Likert scale), and participants technology affinity (test with older adults); task execution time, task number of interaction, task number of mistakes, system feasibility and usefulness (5-point Likert scale), opinion on new technology in the work environment, and technology affinity (test with caregiver)	High interest in talking to the VA and its functionality, high system feasibility to support people with dementia in ADL[l], and step-by-step instructions perceived as useful (test with older adults); high system feasibility to support people with dementia in ADL, step-by-step instructions perceived as useful, appreciation for personalization option, user interface perceived as effective and motivating, mixed results on efficiency, and high results on efficiency and effectiveness (test with caregiver)
Torres et al [37], 2018	Older adults	VA	Home	Conference speech on project	Definition of project objectives, scientific and technological goals, and actions	N/A

[a]VA: voice assistant.

[b]PACS: postacute COVID-19 syndrome.

[c]HRQoL: health-related quality of life.

[d]DASS-21: Depression Anxiety Scale-21.

[e]CD-RISC-25: Connor-Davidson Resilience Scale-25.

[f]EQ-5D-5L: EuroQol-5 Dimensions-5 Levels.

[g]ISI: Insomnia Severity Index.

[h]SF-36: 36-Item Short Form Health Survey.

[i]NGD: normalized Google distance.

[j]N/A: not applicable.

[k]SSPA: Social Skills Performance Assessment.

[l]ADL: activities of daily living.

Bibliometric Analysis

A bibliometric analysis was also conducted to construct a map of the selected articles using VOSviewer software (version 1.6.19; Leiden University's Centre for Science and Technology Studies). This tool represents one of the most popular programs for bibliometric cluster mapping [38].

To illustrate the keyword co-occurrence network, keywords were extracted from the list of the 16 included articles.

During the map creation, the authors choose the co-occurrence type of analysis on keywords and selected full counting as the counting method. The threshold of the minimum number of occurrences of a keyword was set at 2 keywords. All the keywords were illustrated regardless of the greatest total link strength. At the selected keywords' verification step, the authors considered it convenient to merge similar words by creating a thesaurus file. Thus, the thesaurus file included a column of similar keywords and another column with the keyword to be replaced with. Hence, in the final step, the selected keywords were analyzed using the Association Strength normalization method and default values. In addition, for clustering, the default values of resolution (ie, 1.00), minimum cluster size (ie, 1), and merge small cluster option were used.

Results

In the following sections, the synthesis of the findings and results of the bibliometric analysis and qualitative scoring of the 16 selected articles are presented.

XSL•FO
RenderX

Synthesis of Findings

The selected articles were assessed with regard to population, technological solution, environment, study design, outcomes, and main results. Table 1 presents a synthesis of the findings.

Population

In summary, the population most frequently involved in the selected studies is older adults. In some cases, informal caregivers [22], geriatric experts [29], the medical community, the general public [35], or formal caregivers working in a day-care facility with experience in caring for people with dementia [36] are also involved. All the articles detail the total number of people engaged, except for 31% (5/16) of the articles [26,31,32,34,35]. The remaining articles involve a minimum of 7 and a maximum of 109 older adults. Among the selected articles, the age of the population varies widely, including people aged >50 [22,24,26,30], >60 [23,33], >65 [27,29], and >75 years [28]. Naturally, professionals are younger, ranging from 21 [29] to 33 [36] years. However, for some articles [25,31,34-37], there is no information on the age of the population involved. Instead, the sex of the participants is only specified in 56% (9/16) of the articles [24,26-30,32,33,36], in which a majority of female users are included.

In addition, 25% (4/16) of the articles consider participants' familiarity with technology, involving only people with no experience with VA technology [26] and digital devices [31], involving only people with low technology use [32], or specifying people's technological abilities [27]. In addition, some studies consider clinical conditions: 6% (1/16) of the articles [22] included people with diabetes or long-term health conditions, whereas others include people with postacute COVID-19 syndrome [24]; with normative cognitive functioning [28]; with no severe visual or hearing impairment and no moderate to severe cognitive impairment [30]; with mild difficulties in social skills, depression and anxiety symptoms, and nonverbal impairment [33]; and without dementia [36].

Technological Solution

Regarding VA technology solutions, 44% (7/16) of the articles [22,24,25,28,29,32,34] report the use of commercially available VAs, for example, Google Assistant, Amazon Alexa, Apple Siri, and Microsoft Cortana. Some studies specify the design of new VA systems developed using the Amazon Alexa platform and Alexa Voice services [36] or implementing the Google Voice Android Software Development Kit on a tablet [27]. In other studies, the newly designed VA is embedded in a mobile app [23], a PC application [26], or even embodied as a household potted flower [35]. A total of 13% (2/16) of the articles [31,32] describe the design and the testing of a new VA-based digital intelligent platform. Finally, 1 (6%) article [33] presents a web-based automated version of a VA designed to improve communication skills, whereas another one [37] involves a personalized and expressive VA.

Environment

The environment in most of the articles [22,24,25,29-31,33,37] is the home, which is alternated, in the study by Pradhan et al [32], with the older adult living community and, in the studies by Bravo et al [23] and Simpson et al [35], with the retirement home. Instead, the environments in other articles are the laboratory [26,27], the independent living facility [28], the older adult care center [34], and the day-care facility [36]. Thus, the selected articles concerning the use of a VA for social isolation and loneliness address both older adults living independently at home and those living in a facility.

Study Design

Regarding the study design, among the 16 selected studies, 4 (25%) are quantitative, including 1 (6%) evaluation test [23], 1 (6%) pre-post study [24], 1 (6%) development and user test [26], and 1 (6%) VAs test [34]. Qualitative studies include 1 (6%) service evaluation [22], 1 (6%) evaluation test [29], and 1 (6%) pre- post study [32]. Then, there are 5 (31%) mixed studies, including both qualitative and quantitative methods, of which 1 (6%) is an evaluation test [27], 1 (6%) is a single-group quasi-experimental study [28], 1 (6%) is a pre-post study [30], 1 (6%) is a randomized controlled trial [33], and 1 (6%) was a usability study [36]. Finally, the remaining studies include 1 (6%) mini review [25], 2 (13%) conference speeches [35,37], and 1 (6%) study protocol [31]. More detailed information on the methodology results is presented in the *Quality Scoring* section.

Outcomes

Among the outcomes, only 31% (5/16) of the articles [22,25,28,31,35] consider loneliness or social isolation. Of these 16 studies, only 1 (6%) [28] uses a standardized instrument—the 8-item University of California, Los Angeles (UCLA) Loneliness Scale—to assess the perception of loneliness. Instead, most articles (9/16, 56%) [22-24,27,29,30,32,35,36] focus on topics related to the acceptability, user experience, satisfaction, and usability of the technological solution, whereas a smaller number (2/16, 13%) [26,34] focuses on its technical performance. To evaluate these aspects, 5-point Likert scales are used only by 19% (3/16) of the articles [23,27,36].

Further outcomes addressed are verbal and nonverbal behavior in social communication [33], definition of project objectives, scientific and technological goals and actions [37], program impact on health and care trajectories [31], codes and overarching themes [29], interaction anthropomorphic aspects [28], and psychological and physical aspects such as frailty and quality of life [24,31].

Main Results

Turning to the main results of using a VA, the impact on loneliness and social isolation is positive, leading to an improvement in users' perceptions. Specifically, the participants in 13% (2/16) of the studies [22,24] report that the VA helped them cope with loneliness, whereas another study (1/16, 6%) [28] finds a significant reduction in perceived loneliness after 4 weeks of use and that the relational greetings from the user to the VA predict this reduction. Moreover, the loneliness experienced by the person forecasts the number of greetings he or she makes to the VA. Finally, a mini review (1/16, 6%) [25] outlines that the use of VA in older adults improves social connectedness and reduces loneliness.

XSL•FO
RenderX

Other benefits obtained include a positive impact on health and social well-being [22]; improvement in postacute COVID-19 syndrome symptoms, frailty, and health-related quality of life at 6 months follow-up [24]; sedentary life changes [24]; and significant improvement in eye contact and facial expressivity [33].

Regarding the VA, it is considered useful [24], satisfying [23,27], and interesting [36], and it obtains good results in the acknowledgment (the ability to recognize user contextual information) and engagement (the ability to maintain a coherent conversation) performance [34]. In addition, among participants in the study by Pech et al [30], 63% have a positive opinion toward the system used, and in the study by Striegl et al [36], both older adults and formal caregivers describe that the VA used have a high feasibility to support people with dementia in activities of daily living.

The main results also include technical information about the VA. For example, in 1 (6%) study [26], the VA obtains, in all the commands, a right answer ratio percentage >75%; another (1/16, 6%) study [29] identifies 8 major themes as possible VA beneficial functions; and another (1/16, 6%) study [32] presents crucial information for VA development, whereas in another (1/16, 6%) study [35], the device prototype is developed. Finally, critical issues emerge: VA interruptions when the person pauses for too long [27], older adults' resistance to change, unplanned workload for a formal caregiver, specific technological obstacles [30], and bad results in the ability to suggest and perform some related activities at the end of the interaction [34]. Instead, the proposed improvements include facilitated access to professionals, communication at community events, late-night pharmacy service, customized activity proposals, and videoconferencing [30].

For 13% (2/16) of the articles [31,37], it is not applicable to define the main results.

Bibliometric Analysis

Along with the bibliometric analysis, the authors built a thesaurus file containing the words that can be replaced, considering their very close meaning. The thesaurus file is presented in Table 2.

The bibliometric analysis extracted 221 keywords from the included articles, of which 36 (16%) met the threshold of 2 occurrences. The keyword list is presented in Table 3, in descending order of occurrence, showing the number of occurrences and the total link strength.

As can be observed in Table 3, the most used keywords by occurrence were as follows: "social isolation" (n=8), "human" (n=6), "older adults" (n=6), "aged" (n=5), "covid-19" (n=5), "loneliness" (n=5), "human computer interaction" (n=4), and "voice assistant" (n=4).

The most used keywords by total link strength, as shown in Table 3, were as follows: "human" (n=53), "aged" (n=44), "loneliness" (n=44), "social isolation" (n=42), "covid-19" (n=42), "pandemics" (n=29), "very elderly" (n=29), "older adults" (n=28), "prospective study" (n=25), "quality of life" (n=25).

The bibliometric network is illustrated in Figure 3 and consists of 3 clusters of 36 keywords. The clusters are presented in more detail in Table 4, where each keyword from a cluster is shown in descending order by occurrence.

Table 2. Keywords merging using the thesaurus file.

Label	Replace by
human-computer interaction	human computer interaction
humans	human
pandemic	pandemics
Prospective studies	Prospective study
Voice assistants	Voice assistant

XSL•FO
RenderX

Table 3. The list of keywords, number of occurrences, and total link strength. Keywords are presented in descending order of occurrence.

Keyword	Occurrences	Total link strength
social isolation	8	42
human	6	53
older adults	6	28
aged	5	44
covid-19	5	42
loneliness	5	44
human computer interaction	4	16
voice assistant	4	18
aging	3	17
artificial intelligence	3	12
conversational agents	3	10
pandemics	3	29
prospective study	3	25
quality of life	3	25
very elderly	3	29
adult	2	21
aged, 80 and over	2	17
ambient assisted living	2	9
anthropomorphism	2	11
assisted living	2	9
assistive technology	2	13
clinical article	2	21
clinical assessment	2	17
conversational interface	2	8
digital divide	2	14
elderly people	2	7
female	2	21
health	2	12
healthy aging	2	14
male	2	21
middle aged	2	21
natural language generation	2	7
sars-cov-2	2	17
social interactions	2	9
user interfaces	2	8
well-being	2	17

XSL•FO
RenderX

Figure 3. Bibliometric network visualization.

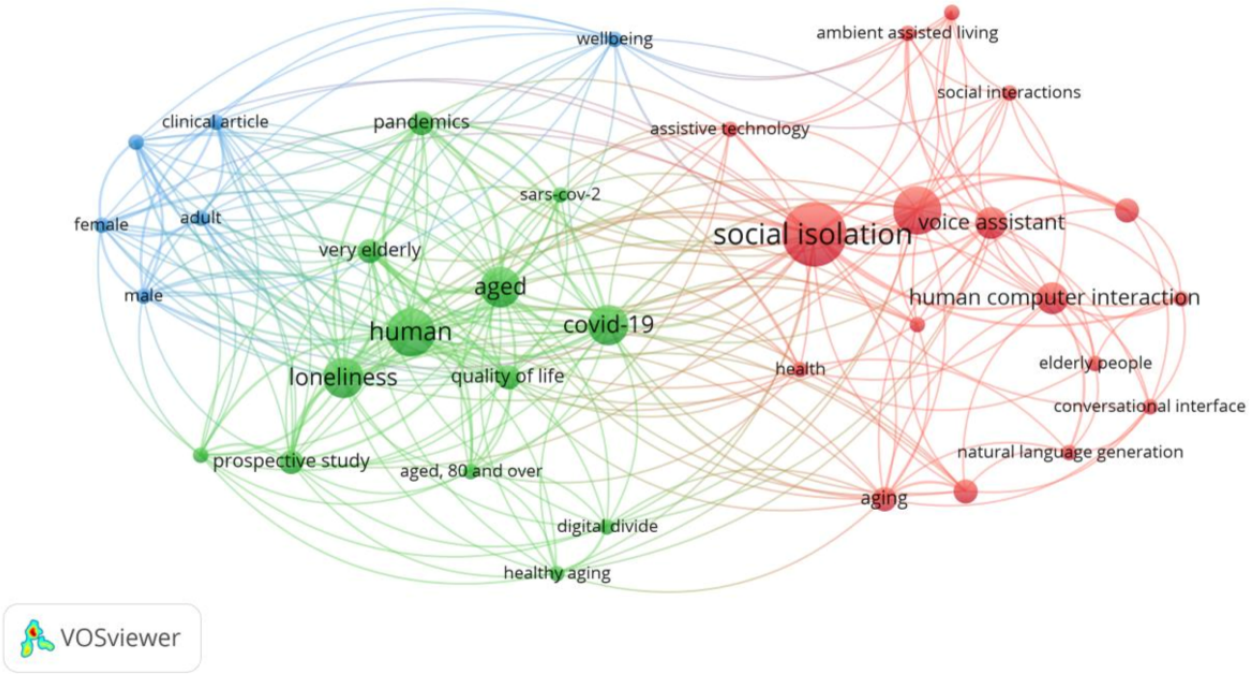

XSL•FO
RenderX

Table 4. Keywords clustering.

Cluster and keywords	Occurrences	Total link strength
Cluster 1 (red color; size: 17 items)		
social isolation	8	42
human	6	53
older adults	6	28
human computer interaction	4	16
voice assistant	4	18
aging	3	17
artificial intelligence	3	12
conversational agents	3	10
ambient assisted living	2	9
anthropomorphism	2	11
assisted living	2	9
assistive technology	2	13
conversational interface	2	8
elderly people	2	7
natural language generation	2	7
social interactions	2	9
user interfaces	2	8
Cluster 2 (green color; size: 13 items)		
aged	5	44
covid-19	5	42
loneliness	5	44
pandemics	3	29
prospective study	3	25
quality of life	3	25
very elderly	3	29
aged, 80 and over	2	17
clinical assessment	2	17
digital divide	2	14
health	2	12
healthy aging	2	14
sars-cov-2	2	17
Cluster 3 (blue color; size: 6 items)		
adult	2	21
clinical article	2	21
female	2	21
male	2	21
middle aged	2	21
well-being	2	17

XSL•FO
RenderX

Quality Scoring

According to the scoring tool, 13% (2/16) of the documents were assessed as being of poor quality in terms of the methodology. In the study by Simpson et al [35], it is unclear what the methodological information is based on, how it is presented, and if it is in line with other sources. The document is based on a conference speech on methods for the design-thinking approach. Instead, in the study by Torres et al [37], most of the information is not clearly sourced; it is unclear what the methodological information is based on and if it is in line with other sources. In addition, this paper is based on a speech at a conference on the objectives, goals, and actions of a research and innovation project.

A total of 56% (9/16) of the documents were considered medium quality. Specifically, 44% (7/16) articles [22,23,25-27,29,31] contain clear sources, methodological quality, and information value, presenting findings in line with the literature. Nevertheless, study designs were not of very high quality, representing mostly multiple case reports and case studies, whereas the study by Corbett et al [25] is a literature review.

A total of 13% (2/16) of the articles [24,34] have instead a more rigorous approach in the study design, representing a qualitative study and a single-group quasi-experimental study, respectively. However, the former is an abstract document lacking bibliographic references, while in the latter, it is unclear what the methodological information is based on. In both cases, the information presented is not clearly linked with the literature findings.

Finally, 31% (5/16) of the documents were deemed of high quality, considering that the information presented and the methodological information are clearly referenced. Among these, 1 (6%) article [33] is a randomized controlled study, while the remaining 25% (4/16) [28,30,32,36] are descriptive or observational studies.

Multimedia Appendix 2 [22-37] provides details of the quality scoring performed on the selected articles.

Discussion

Principal Findings

The purpose of this study is to synthesize knowledge about the use of VAs to reduce loneliness and social isolation among older adults.

Initially, after conducting the literature research, the quality of the selected articles is investigated, focusing on the strengths and weaknesses of the methodologies used. Of the 16 articles included in the review, only 2 (13%) articles [35,37] are considered poor quality, 9 (56%) articles [22-27,29,31,34] are medium quality, and 5 (31%) articles are high quality [28,30,32,33,36]. In summary, although recent publications in the literature on the use of VA by older adults for the reduction of loneliness and social isolation are not numerous, most of them are of medium to high methodological quality in terms of study design, authenticity, clear methodological quality, clear informational value, and representativeness of available primary sources.

After assessing the methodological quality of the selected articles, the findings are summarized, focusing on population, technological solution, environment, study design, outcomes, and main results for a more detailed overview. Among the 16 articles presented, most focus on the evaluation of acceptability, user experience, satisfaction, usability, or performance of the VA, while only 5 (31%) papers deepen the theme of social isolation and loneliness. Of these studies, 1 (6%) [31] has no available results, as it is a study protocol, and another (6%) [35] reached the development stage of a VA prototype. Therefore, 3 (19%) articles remain that investigate the possible effect of the use of a VA on social isolation and loneliness by older adults.

The first paper [22], a service evaluation study, found that using a VA for 2 months at home helped people with diabetes or other long-term health conditions (such as multiple sclerosis, dementia, and depression) combat loneliness. This is particularly relevant because it seems that social isolation increases the risk of mortality through physiological upregulation of chronic inflammation. This impact is significant even for middle-aged people, but is greater for older adults, particularly men [39]. Thus, the results obtained from the use of VAs are particularly relevant considering the population the study targeted but an assessment of loneliness would be needed to investigate the actual impact of the use on this dimension.

The second paper, a single-group quasi-experimental study [28], reported a significant reduction in perceived loneliness, assessed through the 8-item UCLA Loneliness Scale, after older adults living in an independent living facility used a VA for 4 weeks. Thus, loneliness among older adults living alone using a VA has decreased. Moreover, the loneliness perceived at the beginning of the intervention by participants predicts the number of greetings to the VA (such as “Good morning” or “Alexa, I’m home”), and, in addition, these relational greetings forecast loneliness reduction during the month of use. Therefore, according to the authors, VA anthropomorphization might have a role in combating loneliness in older adults.

Finally, the results of a mini review [25] suggest that the VA reduces loneliness among older adults and increases their connectedness. Older adults perceive the VA as a “companion,” especially those who live alone or have solitary lives for most of the day.

These studies show encouraging results about the potential of a VA in reducing social isolation and loneliness in older adults, in line with the suggestion from a systematic review [40] that new technologies can be promising opportunities to reduce social isolation and loneliness in this population. For example, 1 (6%) study found that the use of technology by older adults predicts less loneliness, which has in turn been associated with, on the one hand, better self-reported health and subjective well-being and, on the other hand, fewer chronic diseases and less depression [41]. Therefore, these are preliminary results suggesting that the association between technology use and physical and mental health may be mediated by loneliness.

VAs have the potential to be used by older adults to reduce their social isolation and loneliness, and the results presented go in

XSL•FO
RenderX

that direction; however, they cannot be exhaustive nor conclusive.

Finally, the bibliometric cluster mapping analysis provides valuable insights into the relationships between keywords in the included articles. The generated keyword co-occurrence network revealed 3 distinct clusters, each representing a specific theme or concept in the literature.

Cluster 1, represented by keywords such as “social isolation,” “elderly people,” “voice assistant,” and “human computer interaction,” highlights the relevance of VA technology in combating social isolation among older adults. This cluster emphasizes the relevance of the topic. A VA could be a promising tool for facilitating social interactions, promoting well-being, and addressing the challenges faced by older people regarding social isolation. The relevance of VAs in addressing social isolation among older adults aligns with the findings of Portet et al [9] on the design and evaluation of a smart home VA for older adults. This cluster also corresponds to the author’s focus on the use of quality scoring to evaluate the methodological strengths and weaknesses of the studies, as the inclusion of studies exploring the effectiveness of VAs in combating social isolation would be of particular interest. This cluster emphasizes the importance of designing user-friendly interfaces and incorporating natural language generation and recognition for effective human-computer interaction. This cluster aligns with the literature on ambient assisted living, assistive technology, and artificial intelligence, and it is supported by the work presented in 1 (6%) article [10] on VAs and their applications, as well as in another (1/16, 6%) article [8] that discusses technological solutions for addressing social isolation and loneliness in primary care.

Cluster 2 emphasizes the significance of age in the context of loneliness. Keywords such as “loneliness,” “human,” and “quality of life” indicate the importance of understanding the psychological and emotional aspects of loneliness, considering the diverse experiences of individuals across different demographics. This is supported by the works presented by Valtorta and Hanratty [3] and Holt-Lunstad et al [7], who discuss the association between loneliness, social isolation, and health outcomes in older adults, emphasizing the importance of considering demographic factors in understanding and addressing these issues. Cluster 2 is also relevant in the context of the COVID-19 pandemic, as it includes keywords such as “COVID-19,” “pandemics,” and “digital divide,” which illustrates the impact of the pandemic on social isolation and the need for technological solutions, such as VAs, to bridge the digital divide and ensure connectivity and support for older adults during times of crisis. A study [6] on the association between social isolation, loneliness, and health outcomes in the context of coronary heart disease and stroke further emphasizes the significance of addressing social isolation during pandemics.

Cluster 3 encapsulates a range of keywords related to sex, clinical research, and well-being. The presence of keywords, such as “adult,” “female,” and “male,” along with “clinical article” and “well-being” underscores the importance of understanding how sex-specific factors can significantly impact overall well-being. This cluster likely refers to studies and investigations that explore the intersection of sex-related variables with clinical research outcomes, shedding light on how these factors can influence health and well-being differently among various demographic groups. Moreover, Cluster 3 may offer valuable insights into the evolving landscape of clinical research and its focus on addressing sex-specific health concerns, thus promoting a more comprehensive approach to well-being across diverse populations.

These clusters shed light on important topics related to social isolation, loneliness, and the use of VAs in addressing these issues among older adults. The findings underlined here can inform future research, interventions, and policy development in the field of geriatric care and technology.

Strengths and Limitations

The study provides a comprehensive exploration of voice assistance systems used by older individuals, highlighting popular examples such as Amazon Alexa, Google Assistant, Apple Siri, Microsoft Cortana, Samsung Bixby, and Huawei HiVoice. The study examines the strengths and limitations of these systems.

One of the notable strengths of this study is its investigation into the use of VAs to alleviate loneliness and social isolation among older adults. This topic is fairly recent, but its relevance is growing in both the scientific and technological communities.

Moreover, this investigation is supported by both a literature review and a bibliometric analysis to gather as much knowledge as possible on the role of technology in combating loneliness and social isolation in older adults.

In addition, the selection of studies included in the article underwent an independent evaluation process by the authors, with any disagreements or uncertainties being resolved through discussion.

Another strength is the consideration of the scientific articles published in 2018. This choice was driven by the fact that VAs are relatively new and are continually advancing technological solutions. Furthermore, the application of such technology among older individuals is not yet widespread, resulting in a limited number of studies available on the topic. Despite this limitation, the potential benefits of VA solutions for older adults are highly intriguing, and this study aims to shed light on possible applications and the associated impact on older users.

This study also has limitations that need to be pointed out. First, the number of publications in the systematic review is reduced because the topic has only gained relevance recently. However, the authors decided to proceed with the bibliometric analysis to contribute in terms of interpretation, even though the number of papers on the use of VAs to reduce loneliness and social isolation among older adults is limited. Further limitations relate to the fact that 1 (6%) article [42] could not be retrieved and that the synthesis of findings is not comprehensive, as only the abstract was available for 1 article [24], nor complete, as it was not applicable to define the main results of 13% (2/16) articles [31,37]. Moreover, the selected studies had great heterogeneity, with only 6% (1/16) of studies [33] having a control group and 6% (1/16) of studies [28] having follow-up. Concerning the

XSL•FO
RenderX

information about the population, it is not specified if people involved in the studies live alone or not. This could limit considerations regarding social isolation and loneliness. Finally, most articles collected qualitative data without providing quantitative instruments to assess the actual impact of VA use.

Future Directions

On the basis of this literature review and bibliometric analysis, several priorities for future research can be identified. First, working with keywords from clusters 1 and 2, it is easy to see that "loneliness" and "social isolation" have a huge impact on older people [43]. On the basis of our literature review, authors are more interested in system use and acceptability [30], acceptance user experience [22], and system usability [36], which are just some examples. The main points are "loneliness" and "social isolation," and we only found 1 study [28] to reduce perceived loneliness in older adults. Thus, the topic of the use of VA for social isolation and loneliness among older adults seems to be underestimated in comparison to user experience aspects, which are more deeply investigated in the scientific literature.

Similarly, we encourage that researchers include questionnaires to measure loneliness in future studies, for example, the Revised UCLA Loneliness Scale [44], the De Jong Gierveld Loneliness Scale [45,46], the Steptoe Social Isolation Index for social isolation [44], and the Cornwell Perceived Isolation Scale for perceived isolation [47], for use with VA systems based on artificial intelligence techniques or other related systems to improve the life expectancy of older people. For other specific information about these questionnaires, refer to *Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System* [48]. Second, this work shows that the terms social isolation and loneliness are still often treated as interchangeable, although they are actually related but distinct concepts [3].

In fact, nowadays, the tendency is to refer to loneliness as a subjective negative feeling of perceiving a lack of social network or desired companion, whereas social isolation is the objective lack or scarcity of social contacts and interactions with family, friends, or community [3]. Therefore, it would be particularly relevant if future studies would clearly define which dimensions they measure, as mentioned in the preceding section. Third, future research should examine the large heterogeneity within the older adult population. Some of the selected articles described different characteristics of the population, but none delved into the possible different impacts of VA use in relation to these variables. Future studies should explore the effects of using a VA on the social isolation and loneliness of older adults, investigating possible differences in sex, socioeconomic background, and also familiarity with technology and living conditions.

Conclusions

This paper conducted a literature review and a bibliometric analysis of the use of VAs among older adults to reduce social isolation and loneliness. The findings indicate that most studies focus on the usability, acceptability, or user experience of the VA. However, studies directly addressing the impact that using a VA has on the social isolation and loneliness of older adults have positive results and provide important information for future research, interventions, and policy development in the field of geriatric care and technology.

Acknowledgments

This study has been developed within the framework of the EMILIO (Increase Self Management and Counteract Social Isolation Using a Voice Assistant Enabled Virtual Concierge) project (AAL-2021-8-120-CP), cofinanced under the Ambient Assisted Living Joint Programme of the European Commission [49] and the National Funding Agencies of Belgium, the Netherlands, Italy, and Switzerland.

The authors are grateful to all consortium partners: Italian National Institute of Health and Science on Aging (IRCCS INRCA), Solving Team SRL, ICT Factory GmbH, Erdmann Design AG, Magicview, ePoint, Vulpia VZW, Institute of Space Science, INFLPR Subsidiary, Transilvania University of Brasov.

The project website is available on the internet [50].

Authors' Contributions

RAM, CF, AD, AN, MM, and CV contributed to the methodology, investigation, writing of the original draft, and reviewing and editing. DMK, AAM, and S-AM were responsible for the investigation, writing of the original draft, reviewing, and editing. RB conducted reviewing and editing. LR was involved in conceptualization and funding acquisition, whereas MDR was involved in methodology, project administration, conceptualization, supervision, funding acquisition, reviewing, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excluded articles and motivations for the exclusion.
[DOCX File , 14 KB - jmir_v26i1e50534_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Quality scoring of selected articles.
[DOCX File , 14 KB - jmir_v26i1e50534_app2.docx]

References

1. Active ageing : a policy framework. World Health Organization. 2002. URL: https://apps.who.int/iris/handle/10665/67215 [accessed 2023-06-01]
2. Holt-Lunstad J. The potential public health relevance of social isolation and loneliness: prevalence, epidemiology, and risk factors. Public Policy Aging Rep 2017;27(4):127-130. [doi: 10.1093/ppar/prx030]
3. Valtorta N, Hanratty B. Loneliness, isolation and the health of older adults: do we need a new research agenda? J R Soc Med 2012 Dec;105(12):518-522 [FREE Full text] [doi: 10.1258/jrsm.2012.120128] [Medline: 23288086]
4. Van As BA, Imbimbo E, Franceschi A, Menesini E, Nocentini A. The longitudinal association between loneliness and depressive symptoms in the elderly: a systematic review. Int. Psychogeriatr 2021 Apr 14;34(7):657-669. [doi: 10.1017/s1041610221000399]
5. Kuiper JS, Zuidersma M, Oude Voshaar RC, Zuidema SU, van den Heuvel ER, Stolk RP, et al. Social relationships and risk of dementia: a systematic review and meta-analysis of longitudinal cohort studies. Ageing Res Rev 2015 Jul;22:39-57. [doi: 10.1016/j.arr.2015.04.006] [Medline: 25956016]
6. Valtorta NK, Kanaan M, Gilbody S, Hanratty B. Loneliness, social isolation and risk of cardiovascular disease in the English longitudinal study of ageing. Eur J Prev Cardiol 2018 Sep;25(13):1387-1396. [doi: 10.1177/2047487318792696] [Medline: 30068233]
7. Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. Perspect Psychol Sci 2015 Mar;10(2):227-237. [doi: 10.1177/1745691614568352] [Medline: 25910392]
8. Freedman A, Nicolle J. Social isolation and loneliness: the new geriatric giants: approach for primary care. Can Fam Physician 2020 Mar;66(3):176-182 [FREE Full text] [Medline: 32165464]
9. Portet F, Vacher M, Golanski C, Roux C, Meillon B. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers Ubiquit Comput 2011 Oct 2;17(1):127-144. [doi: 10.1007/s00779-011-0470-5]
10. Hoy MB. Alexa, Siri, Cortana, and more: an introduction to voice assistants. Med Ref Serv Q 2018 Jan 12;37(1):81-88. [doi: 10.1080/02763869.2018.1404391] [Medline: 29327988]
11. Weiser M. The computer for the 21st century. Sci Am 1991 Sep;265(3):94-104 [FREE Full text] [doi: 10.1038/scientificamerican0991-94]
12. Chen J, Yang YT, Zhu X, Zhu Z. Share and care: a senior-friendly family interaction application. In: Proceedings of the IEEE MIT Undergraduate Research Technology Conference (URTC). 2020 Presented at: URTC 2020; October 9-11, 2020; Cambridge, MA URL: https://ieeexplore.ieee.org/document/9668885 [doi: 10.1109/urtc51696.2020.9668885]
13. Eimontaite I, Voinescu A, Alford C, Caleb-Solly P, Morgan P. The impact of different human-machine interface feedback modalities on older participants' user experience of CAVs in a simulator environment. In: Proceedings of the International Conference on Human Factors in Transportation. 2019 Presented at: AHFE 2019; July 24-28, 2019; Washington, DC URL: https://link.springer.com/chapter/10.1007/978-3-030-20503-4_11 [doi: 10.1007/978-3-030-20503-4_11]
14. Eirale A, Martini M, Tagliavini L, Gandini D, Chiaberge M, Quaglia G. Marvin: an innovative omni-directional robotic assistant for domestic environments. Sensors (Basel) 2022 Jul 14;22(14):1-22 [FREE Full text] [doi: 10.3390/s22145261] [Medline: 35890940]
15. Martin-Hammond A, Vemireddy S, Rao K. Exploring older adults' beliefs about the use of intelligent assistants for consumer health information management: a participatory design study. JMIR Aging 2019 Dec 11;2(2):e15381 [FREE Full text] [doi: 10.2196/15381] [Medline: 31825322]
16. Méndez JI, Mata O, Ponce P, Meier A, Peffer T, Molina A. Multi-sensor system, gamification, and artificial intelligence for benefit elderly people. In: Ponce H, Martínez-Villaseñor L, Brieva J, Moya-Albor E, editors. Challenges and Trends in Multimodal Fall Detection for Healthcare. Cham, Switzerland: Springer; 2020.
17. Restyandito, Febryandi, Nugraha KA, Sebastian D. Mobile social media interface design for elderly in Indonesia. In: Proceedings of the HCI International 2020 – Late Breaking Posters. 2020 Presented at: HCII 2020; July 19-24, 2020; Copenhagen, Denmark URL: https://link.springer.com/chapter/10.1007/978-3-030-60703-6_10 [doi: 10.1007/978-3-030-60703-6_10]
18. Syeda MZ, Park M, Kim Y, Kwon YM. Tangible social content service system: making digital technology easier to use by elderly and its usability evaluation. In: Proceedings of the 12th International Conference on Complex, Intelligent, and Software Intensive Systems. 2018 Presented at: CISIS 2018; July 4-6, 2018; Matsue, Japan. [doi: 10.1007/978-3-319-93659-8_80]
19. Zhou D, Barakova EI, An P, Rauterberg M. Assistant robot enhances the perceived communication quality of people with dementia: a proof of concept. IEEE Trans Human Mach Syst 2022 Jun;52(3):332-342. [doi: 10.1109/thms.2021.3112957]

XSL•FO
RenderX

20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
21. Laws JM, Winnard A. Tool for scoring the quality of non-empirical data sources- E.G: technical reports. Aerospace Medicine and Rehabilitation Laboratory, Northumbria University. 2019. URL: https://www.researchgate.net/publication/331385312_Tool_for_Scoring_the_Quality_of_Non-Empirical_Data_Sources-_EG_Technical_Reports [accessed 2024-02-23]
22. Balasubramanian GV, Beaney P, Chambers R. Digital personal assistants are smart ways for assistive technology to aid the health and wellbeing of patients and carers. BMC Geriatr 2021 Nov 15;21(1):643 [FREE Full text] [doi: 10.1186/s12877-021-02436-y] [Medline: 34781881]
23. Bravo SL, Herrera CJ, Valdez EC, Poliquit KJ, Ureta J, Cu J, et al. CATE: an embodied conversational agent for the elderly. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART. 2020 Presented at: ICAART 2020; February 22-24, 2020; Valletta, Malta URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0009174009410948 [doi: 10.5220/0009174009410948]
24. Caselgrandi A, Milić J, Motta F, Belli M, Venuta M, Aprile E, et al. Voice assistance to develop a participatory research and action to improve health trajectories of people with PACS. Antivir Ther 2021 Dec 1;26(1_suppl):13-14. [doi: 10.1177/13596535211063242]
25. Corbett CF, Wright PJ, Jones K, Parmer M. Voice-activated virtual home assistant use and social isolation and loneliness among older adults: mini review. Front Public Health 2021;9:742012 [FREE Full text] [doi: 10.3389/fpubh.2021.742012] [Medline: 34708017]
26. Farías-Barraza B, Reyes-Rogget M, López FA, López-Martínez IN, Contreras-Bolton C, Linfati R. Low-cost voice assistant design and testing for older adults. In: Proceedings of the Computer Information Systems and Industrial Management. 2022 Presented at: CISIM 2022; July 15-17, 2022; Barranquilla, Colombia. [doi: 10.1007/978-3-031-10539-5_8]
27. Garcia-Mendez S, de Arriba-Perez F, Gonzalez-Castano FJ, Regueiro-Janeiro JA, Gil-Castineira F. Entertainment chatbot for the digital inclusion of elderly people without abstraction capabilities. IEEE Access 2021 May 17;9:75878-75891. [doi: 10.1109/access.2021.3080837]
28. Jones VK, Hanus M, Yan C, Shade MY, Blaskewicz Boron J, Maschieri Bicudo R. Reducing loneliness among aging adults: the roles of personal voice assistants and anthropomorphic interactions. Front Public Health 2021;9:750736 [FREE Full text] [doi: 10.3389/fpubh.2021.750736] [Medline: 34957013]
29. O'Brien K, Light SW, Bradley S, Lindquist L. Optimizing voice-controlled intelligent personal assistants for use by home-bound older adults. J Am Geriatr Soc 2022 May;70(5):1504-1509 [FREE Full text] [doi: 10.1111/jgs.17625] [Medline: 35029296]
30. Pech M, Gbessemehlan A, Dupuy L, Sauzéon H, Lafitte S, Bachelet P, et al. Lessons learned from the SoBeezy program for older adults during the COVID-19 pandemic: experimentation and evaluation. JMIR Form Res 2022 Nov 24;6(11):e39185 [FREE Full text] [doi: 10.2196/39185] [Medline: 36355629]
31. Pérès K, Zamudio-Rodriguez A, Dartigues JF, Amieva H, Lafitte S. Prospective pragmatic quasi-experimental study to assess the impact and effectiveness of an innovative large-scale public health intervention to foster healthy ageing in place: the SoBeezy program protocol. BMJ Open 2021 Apr 29;11(4):e043082 [FREE Full text] [doi: 10.1136/bmjopen-2020-043082] [Medline: 33926977]
32. Pradhan A, Findlater L, Lazar A. "Phantom friend" or "Just a box with information": personification and ontological categorization of smart speaker-based voice assistants by older adults. Proc ACM Hum Comput Interact 2019 Nov 07;3(CSCW):1-21. [doi: 10.1145/3359316]
33. Razavi SZ, Schubert LK, van Orden K, Ali MR, Kane B, Hoque E. Discourse behavior of older adults interacting with a dialogue agent competent in multiple topics. ACM Trans Interact Intell Syst 2022 Jul 23;12(2):1-21. [doi: 10.1145/3484510]
34. Reis A, Paulino D, Paredes H, Barroso I, Monteiro MJ, Rodrigues V, et al. Using intelligent personal assistants to assist the elderlies: an evaluation of Amazon Alexa, Google Assistant, Microsoft Cortana, and Apple Siri. In: Proceedings of the 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW). 2018 Presented at: TISHW; June 20-22, 2018; Thessaloniki, Greece URL: https://ieeexplore.ieee.org/document/8559503/authors#authors [doi: 10.1109/tishw.2018.8559503]
35. Simpson J, Gaiser F, MacÍk M, Breßgott T. Daisy: a friendly conversational agent for older adults. In: Proceedings of the 2nd Conference on Conversational User Interfaces. 2020 Presented at: CUI '20; July 22-24, 2020; Bilbao, Spain. [doi: 10.1145/3405755.3406166]
36. Striegl J, Gollasch D, Loitsch C, Weber G. Designing VUIs for social assistance robots for people with dementia. In: Proceedings of Mensch und Computer 2021. 2021 Presented at: MuC '21; September 5-8, 2021; Ingolstadt, Germany. [doi: 10.1145/3473856.3473887]
37. Torres MI, Chollet G, Montenegro C, Tenorio-Laranga J, Gordeeva O, Esposito A, et al. EMPATHIC, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderdy. 2018 Presented at: 4th International Conference on Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2018; 21-23 November 2018; Barcelona, Spain p. 172-173. [doi: 10.26342/2018-61-24]

XSL•FO
RenderX

38. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
39. Yang YC, McClintock MK, Kozloski M, Li T. Social isolation and adult mortality: the role of chronic inflammation and sex differences. J Health Soc Behav 2013 Jun;54(2):183-203 [FREE Full text] [doi: 10.1177/0022146513485244] [Medline: 23653312]
40. Poscia A, Stojanovic J, La Milia DI, Duplaga M, Grysztar M, Moscato U, et al. Interventions targeting loneliness and social isolation among the older people: an update systematic review. Exp Gerontol 2018 Feb;102:133-144. [doi: 10.1016/j.exger.2017.11.017] [Medline: 29199121]
41. Chopik WJ. The benefits of social technology use among older adults are mediated by reduced loneliness. Cyberpsychol Behav Soc Netw 2016 Sep;19(9):551-556 [FREE Full text] [doi: 10.1089/cyber.2016.0151] [Medline: 27541746]
42. Chen S, Nakamura M. Generating personalized dialogues based on conversation log summarization and sentiment analysis. In: Proceedings of the 23rd International Conference on Information Integration and Web Intelligence. 2021 Presented at: iiWAS2021; November 29-December 1, 2021; Linz, Austria. [doi: 10.1145/3487664.3487695]
43. O Súilleabháin PS, Gallagher S, Steptoe A. Loneliness, living alone, and all-cause mortality: the role of emotional and social loneliness in the elderly during 19 years of follow-up. Psychosom Med 2019;81(6):521-526 [FREE Full text] [doi: 10.1097/PSY.0000000000000710] [Medline: 31094903]
44. Steptoe A, Shankar A, Demakakos P, Wardle J. Social isolation, loneliness, and all-cause mortality in older men and women. Proc Natl Acad Sci U S A 2013 Apr 09;110(15):5797-5801 [FREE Full text] [doi: 10.1073/pnas.1219686110] [Medline: 23530191]
45. de Jong-Gierveld J, Kamphuls F. The development of a Rasch-type loneliness scale. Appl Psychol Meas 2016 Jul 27;9(3):289-299. [doi: 10.1177/014662168500900307]
46. Gierveld JD, Tilburg TV. A 6-item scale for overall, emotional, and social loneliness: confirmatory tests on survey data. Res Aging 2006;28(5):582-598. [doi: 10.1177/0164027506289723]
47. Cornwell EY, Waite LJ. Social disconnectedness, perceived isolation, and health among older adults. J Health Soc Behav 2009 Mar;50(1):31-48 [FREE Full text] [doi: 10.1177/002214650905000103] [Medline: 19413133]
48. National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Health and Medicine Division, Board on Behavioral, Cognitive, and Sensory Sciences, Board on Health Sciences Policy, Committee on the Health and Medical Dimensions of Social Isolation and Loneliness in Older Adults. Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System. Washington, DC: National Academies Press; 2020.
49. Ageing well in the digital world. Active Assisted Living Programme. URL: https://www.aal-europe.eu/ [accessed 2024-02-14]
50. Emilio–personal assistant. Active Assisted Living Programme. URL: https://www.emilio-aal.eu/ [accessed 2024-02-26]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
UCLA: University of California, Los Angeles
VA: voice assistant

Edited by T de Azevedo Cardoso; submitted 04.07.23; peer-reviewed by V Jones, F Yang; comments to author 26.09.23; revised version received 13.10.23; accepted 24.11.23; published 18.03.24.

Please cite as:
Marziali RA, Franceschetti C, Dinculescu A, Nistorescu A, Kristály DM, Mo oi AA, Broekx R, Marin M, Vizitiu C, Moraru SA, Rossi L, Di Rosa M
Reducing Loneliness and Social Isolation of Older Adults Through Voice Assistants: Literature Review and Bibliometric Analysis
J Med Internet Res 2024;26:e50534
URL: https://www.jmir.org/2024/1/e50534
doi:10.2196/50534
PMID:38498039

©Rachele Alessandra Marziali, Claudia Franceschetti, Adrian Dinculescu, Alexandru Nistorescu, Dominic Mircea Kristály, Adrian Alexandru Mo oi, Ronny Broekx, Mihaela Marin, Cristian Vizitiu, Sorin-Aurel Moraru, Lorena Rossi, Mirko Di Rosa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Use of Mobile Technologies to Promote Physical Activity and Reduce Sedentary Behaviors in the Middle East and North Africa Region: Systematic Review and Meta-Analysis

Huong Ly Tong[1,2], PhD; Aroub Alnasser[3], PhD; Najim Z Alshahrani[4], MD; Rowaedh A Bawaked[5], PhD; Reem AlAhmed[6], BSc; Reem F Alsukait[2], PhD; Severin Rakic[2], PhD; Volkan Cetinkaya[2], PhD; Hazzaa M Al-Hazzaa[7], PhD; Saleh A Alqahtani[8,9], PhD

[1]Cultural and Indigenous Research Centre Australia, Redfern, Australia
[2]The World Bank Group, Washington, DC, United States
[3]Department of Food Science and Nutrition, King Saud University, Riyadh, Saudi Arabia
[4]Department of Family and Community Medicine, Faculty of Medicine, University of Jeddah, Jeddah, Saudi Arabia
[5]Department of Public Health, Saudi Electronic University, Riyadh, Saudi Arabia
[6]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital & Research Center, Riyadh, Saudi Arabia
[7]Health Sciences Research Center, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[8]Organ Transplant Center of Excellence, King Faisal Specialist Hospital & Research Center, Riyadh, Saudi Arabia
[9]Division of Gastroenterology & Hepatology, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Huong Ly Tong, PhD
Cultural and Indigenous Research Centre Australia
16 Eveleight St
Redfern, 2016
Australia
Phone: 61 28585 ext 1311
Email: lytong.2208@gmail.com

Abstract

Background: The Middle East and North Africa (MENA) region faces unique challenges in promoting physical activity and reducing sedentary behaviors, as the prevalence of insufficient physical activity is higher than the global average. Mobile technologies present a promising approach to delivering behavioral interventions; however, little is known about the effectiveness and user perspectives on these technologies in the MENA region.

Objective: This study aims to evaluate the effectiveness of mobile interventions targeting physical activity and sedentary behaviors in the MENA region and explore users' perspectives on these interventions as well as any other outcomes that might influence users' adoption and use of mobile technologies (eg, appropriateness and cultural fit).

Methods: A systematic search of 5 databases (MEDLINE, Embase, CINAHL, Scopus, and Global Index Medicus) was performed. Any primary studies (participants of all ages regardless of medical condition) conducted in the MENA region that investigated the use of mobile technologies and reported any measures of physical activity, sedentary behaviors, or user perceptions were included. We conducted a narrative synthesis of all studies and a meta-analysis of randomized controlled trials (RCTs). The Cochrane risk-of-bias tool was used to assess the quality of the included RCTs; quality assessment of the rest of the included studies was completed using the relevant Joanna Briggs Institute critical appraisal tools.

Results: In total, 27 articles describing 22 interventions (n=10, 37% RCTs) and 4 (15%) nonexperimental studies were included (n=6141, 46% women). Half (11/22, 50%) of the interventions included mobile apps, whereas the other half examined SMS. The main app functions were goal setting and self-monitoring of activity, whereas SMS interventions were primarily used to deliver educational content. Users in experimental studies described several benefits of the interventions (eg, gaining knowledge and receiving reminders to be active). Engagement with the interventions was poorly reported; few studies (8/27, 30%) examined users' perspectives on the appropriateness or cultural fit of the interventions. Nonexperimental studies examined users' perspectives on mobile apps and fitness trackers, reporting several barriers to their use, such as perceived lack of usefulness, loss of interest, and technical issues. The meta-analysis of RCTs showed a positive effect of mobile interventions on physical activity outcomes

XSL•FO
RenderX

(standardized mean difference=0.45, 95% CI 0.17-0.73); several sensitivity analyses showed similar results. The trim-and-fill method showed possible publication bias. Only 20% (2/10) of the RCTs measured sedentary behaviors; both reported positive changes.

Conclusions: The use of mobile interventions for physical activity and sedentary behaviors in the MENA region is in its early stages, with preliminary evidence of effectiveness. Policy makers and researchers should invest in high-quality studies to evaluate long-term effectiveness, intervention engagement, and implementation outcomes, which can inform the design of culturally and socially appropriate interventions for countries in the MENA region.

Trial Registration: PROSPERO CRD42023392699; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=392699

(J Med Internet Res 2024;26:e53651) doi:10.2196/53651

KEYWORDS

mobile apps; fitness trackers; SMS; SMS text messaging; physical activity; exercise; sedentary behavior; Middle East; Africa, Northern; movement; physical inactivity; smartphone; mobile phone; mobile health; mHealth; digital health; behavior change; intervention

Introduction

Background

Chronic diseases were responsible for >70% of deaths worldwide in 2019, making them the leading cause of mortality and morbidity [1]. In the Middle East and North Africa (MENA) region [2], chronic conditions accounted for 79% of deaths in 2020 [3]. Although there are many factors contributing to the prevalence of chronic diseases, physical inactivity and sedentary behaviors are well-established risk factors [4-8]. Worldwide studies have revealed that the MENA region has higher rates of physical inactivity compared with the global average, with 32.8% of adults and 85% of adolescents considered insufficiently active compared with 28% of adults and 81% of adolescents worldwide [9,10].

There are specific factors unique to the MENA region that might influence physical inactivity and sedentary behaviors, such as environmental conditions or infrastructure. Specifically, extreme weather conditions, particularly during the hot summer months, can make outdoor exercise uncomfortable or even hazardous [11,12]. Limited access to sports facilities is another barrier to physical activity, and high urbanization associated with dependence on motor vehicles likely increases sedentary time [11,12]. Given these regional factors, interventions targeting physical activity and sedentary behaviors in the MENA region need to be tailored to a regional and population context.

Mobile interventions such as mobile apps, fitness trackers, and SMS text messages can be a powerful tool for promoting physical activity and reducing sedentary behaviors. The high mobile penetration rate in the MENA region [13,14] makes mobile technologies a possible solution for delivering large-scale real-time interventions. Moreover, the advanced capability of these technologies to automate and process data can allow interventions to be tailored according to the specific individual, context, and region [15-18]. Mobile apps or fitness trackers can also incorporate theory-based behavior change techniques that are known to be effective [19], such as automating self-monitoring of activity and providing feedback or allowing users to set goals.

Despite this potential, to date, little is known about the effectiveness of mobile technologies targeting physical activity and sedentary behaviors in the MENA region. Although some systematic reviews have reported a positive effect of mobile technologies on behavioral outcomes [18,20-31], none of these studies have focused on evidence from the MENA region. A systematic review explored physical activity interventions in 6 Arabian Gulf countries (ie, Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and United Arab Emirates) [32]—a subset of the MENA region—but did not solely focus on mobile interventions or explore interventions for sedentariness. Thus, it remains unclear whether mobile technologies are particularly effective in changing physical activity and sedentary behaviors in the MENA region. In addition, given the unique barriers that the MENA population faces, there is also a need to understand users' acceptability and the implementation outcomes of these mobile interventions (eg, the appropriateness or cultural fit of the interventions). Finally, it is worth noting that there is diversity among the countries in the MENA region regarding income level, economic and social stability [2], and mobile penetration rate [14]. Socioeconomic disparities have been linked to inequitable access to technologies and varying levels of digital literacy, creating a "digital divide" [33-35]. Understanding the geographical scope of mobile health research in the MENA region is crucial to gauge whether findings are applicable across the region and identify potential signs of a digital divide.

Objectives

The aim of this systematic review and meta-analysis was to summarize the characteristics and evaluate the effectiveness of mobile interventions targeting physical activity and sedentary behaviors in the MENA region. A secondary aim was to explore users' perspectives on these interventions as well as any other outcomes that might influence users' adoption and use of mobile technologies (eg, appropriateness and cultural fit).

Methods

This systematic review is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement (Multimedia Appendix 1 [36]). We followed the protocol registered in PROSPERO (CRD42023392699).

XSL•FO
RenderX

Search Strategy

A systematic search of the literature was conducted in MEDLINE, Embase, CINAHL, Scopus, and Global Index Medicus from database inception to January 19, 2023. The search strategy was developed after consultation with a research librarian and comprised three strings: (1) mobile technology terms, (2) physical activity and sedentary behavior terms, and (3) MENA countries (Multimedia Appendix 2). The reference lists of relevant articles and other reviews on similar topics were also screened to ensure that all eligible studies were captured. A gray literature search was performed using Google Scholar. We contacted the authors to obtain additional information when needed. We also contacted the authors to request the full text when it was not available. If the authors did not provide the full text, we excluded the studies.

Eligibility Criteria

We included studies that met the following population, intervention, comparator, outcome, and study design criteria: (1) participants of all ages regardless of medical condition, (2) interventions that used mobile technologies (ie, mobile apps, fitness trackers, or SMS text messages), (3) presence or absence of a comparator, (4) any outcomes related to physical activity or sedentary behaviors (eg, active minutes and sitting time) or to users' perceptions of mobile technologies, (5) conducted in MENA countries as listed by the World Bank [2], and (6) primary research studies (eg, randomized controlled trials [RCTs], quasi-experimental studies, or qualitative studies). Studies in all languages were included.

Studies were excluded if (1) the intervention did not have a mobile technology component (eg, web-based only) or (2) they only reported measures related to physical function (eg, sit-to-stand test). Multimedia Appendix 3 [2,20,37] provides a detailed outline of the eligibility criteria.

Screening and Data Extraction

The screening procedure was piloted before beginning. In total, 4 pairs of investigators independently conducted a 2-phase screening using Rayyan (Rayyan Systems, Inc) [38] for title and abstract and full text. Disagreements were resolved through discussion.

The data extraction form was developed in Microsoft Excel (Microsoft Corp) and piloted before extraction. The following data were collected for each study: author; year; country; study design; study population, sample size, participant demographics, and baseline characteristics; details of the intervention and control conditions; retention rates (ie, percentage that completed the follow-up assessment); outcomes and times of measurement; and source of funding and conflicts of interest. Data were extracted by one researcher and checked for accuracy by another. RCTs were assessed by 2 independent researchers using the Cochrane risk-of-bias tool [39], and disagreements were resolved through discussion. To assess outcome reporting bias, we compared the outcomes specified in the trial protocols with the outcomes reported in the corresponding trial publications; if the trial protocols were unavailable, we compared the outcomes reported in the methods and results sections of the trial publications. Quality assessment of the rest of the included studies was completed using the Joanna Briggs Institute (JBI) critical appraisal tools, including the checklists for quasi-experimental studies, cross-sectional studies, and qualitative research [40].

Strategies for Data Synthesis

A narrative synthesis was conducted of all included studies. A meta-analysis was conducted for RCTs in which the control group did not have a mobile technology component. This was because the study aimed to examine the effectiveness of mobile technologies on physical activity or sedentary behaviors compared with nonmobile interventions. Cluster randomized trials were also included in the meta-analysis by calculating the effective sample size using the Cochrane guideline [41] (Multimedia Appendix 4 [42,43]). We planned to perform 2 separate meta-analyses to combine outcome measures of physical activity and sedentary behaviors. However, only 2 trials measured sedentary behaviors; given this small number, a meta-analysis was used to combine physical activity outcomes only. Sedentary outcomes were summarized narratively.

Whenever a single study reported multiple outcomes for the same behavior (eg, reporting both daily step count and daily active minutes for physical activity), the outcome that was included in the meta-analysis was selected through the consensus of the authors favoring (1) the included studies' primary outcomes, (2) the most meaningful outcomes to intended users (eg, step count), and (3) the longest follow-up. All reported outcomes (from data collected after the intervention) were pooled, and all effect sizes were transformed into standardized mean differences. We used a random-effects model for all analyses; the restricted maximum likelihood estimator was used to calculate the heterogeneity variance (τ^2). To assess heterogeneity, I^2 was used.

The presence and potential impact of publication bias were explored using a funnel plot, the Egger test, and the trim-and-fill method by Duval and Tweedie [44]. In total, 2 sensitivity analyses were conducted including only (1) studies that were randomized at the individual level and (2) studies that had a low risk of bias in at least 3 out of 5 categories. Owing to an insufficient number of studies, a planned meta-regression was not conducted. We used the Grading of Recommendations Assessment, Development, and Evaluation system for grading the body of evidence [45]. All computations were conducted in R (version 4.2.3; R Foundation for Statistical Computing) [46]. The 2-tailed significance level for all statistical tests was set at $P<.05$.

Results

Study Selection

The search retrieved 2038 unique articles (Figure 1). After abstract and full-text screening, 23 articles were included. Multimedia Appendix 5 provides a list of excluded studies at full-text screening. In total, 2 articles were found through the reference lists of relevant articles, and 2 were included from the Google Scholar search. Finally, the systematic review included 27 articles: 11 (41%) articles describing 10 unique RCT studies (n=2, 20% cluster RCTs [42,43,47]) [42,43,47-55], 12 (44%)

XSL•FO
RenderX

quasi-experimental studies [56-67], 3 (11%) surveys [68-70], and 1 (4%) interview study [71] (Tables 1-3).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process. *Other sources include reference lists of relevant articles (n=2) and Google Scholar (n=2).

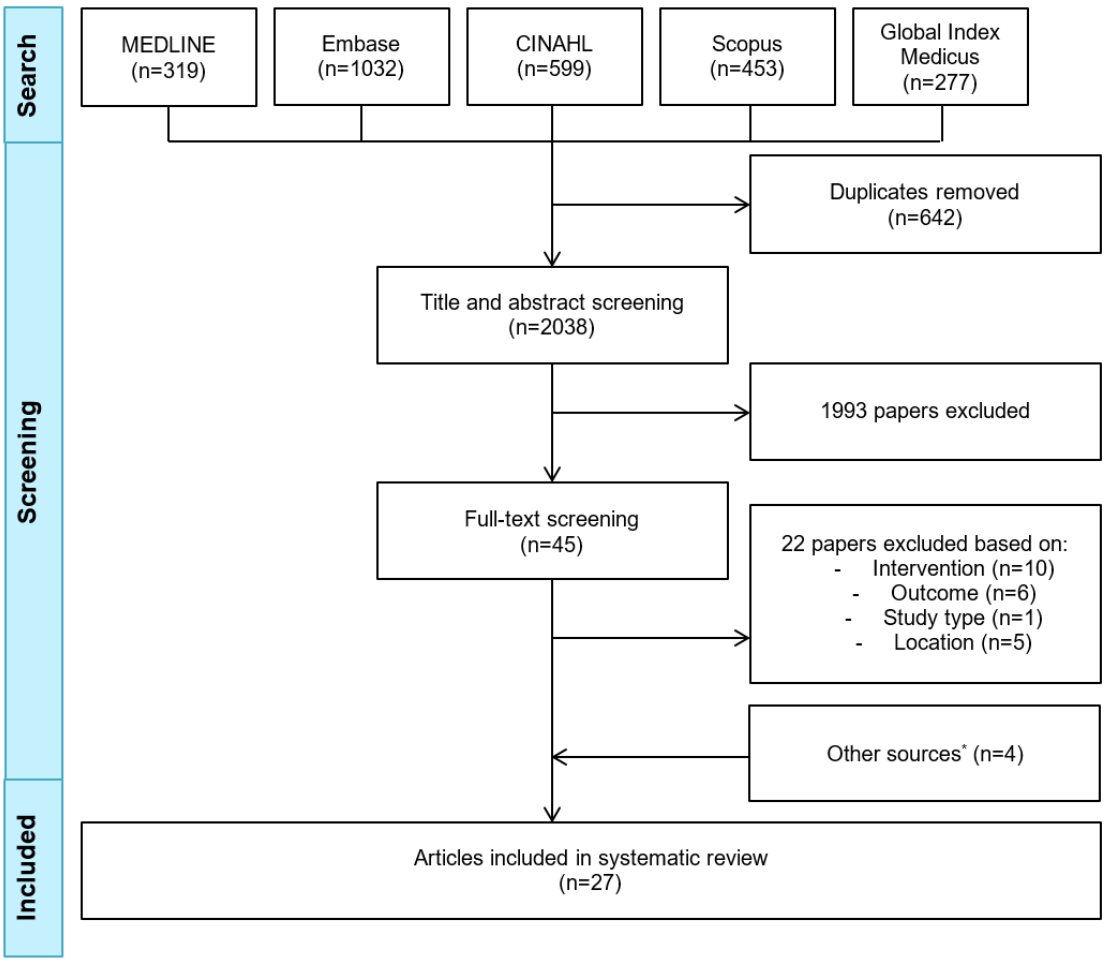

XSL•FO
RenderX

Table 1. Study and intervention characteristics of the included randomized controlled trials (RCTs).

Study, year, and country	Study population	Study length	Sample size, N	Female participants, n (%)	Age (y), mean	Details about the intervention	Details about the control
Ansari et al [55], 2022, Iran	Women aged 40-60 y	12 wk	IG[a]: 55; CG[b]: 55	110 (100)	52	The sample was divided into 2 groups: SMS text messaging or mobile social networking app; both received information about benefits and how to overcome potential barriers to physical activity.	No true control
Eslami et al [50], 2022, Iran	Pregnant women with overweight and obesity (ie, BMI >25 kg/m^2)	8 wk	IG: 70; CG: 70	140 (100)	NR[c]; >55% in the 20-30–y age group	SMS text messages about physical activity (recommendations of amount and intensity, benefits, safe exercises, and precautions during pregnancy), diet, and supplements; educational booklet; and 1 face-to-face training session	Usual care
Saleh et al [48], 2022, Jordan	Patients with heart failure	8 wk	IG: 76; CG: 76	52 (39)	60.8	2 apps: Samsung Health app (goal setting and self-monitoring of activity) and social media app (posts about the benefits of physical activity, exercise videos, and discussion forum with a health expert; participants were encouraged to post and comment in the group)	Usual care; participants also received the setup for the Samsung Health app without receiving details about its features
Alshahrani et al [49], 2021, Saudi Arabia	Female university students	10 wk	IG: 53; CG: 50	103 (100)	NR; 70% were aged ≥20 y	WhatsApp group: a 15-min orientation and 3-4 messages/wk about physical activity	No intervention
Abbaspoor et al [51], 2020, Iran	Pregnant women with prediabetes	12 wk	IG: 50; CG: 50	100 (100)	IG: 27.8; CG: 29	2 SMS text messages every second day about gestational diabetes and 4 face-to-face training sessions	4 face-to-face training sessions
Alghafri et al [43,47], 2020 and 2018, Oman	Inactive adults with type 2 diabetes	1 y	Cluster RCT; IG: 122; CG: 110	137 (59)	44	Monthly WhatsApp messages, 3 face-to-face 20-min consultations about physical activity, and pedometers	Usual care
Parandeh et al [52], 2019, Iran	Women aged 30-45 y	8 wk	IG: 63; CG: 63	126 (100)	IG: 36.1; CG: 35.8	Daily SMS text messages about osteoporosis prevention (including physical activity and diet)	Educational SMS text messages about cancers in women
Quronfulah [42], 2019, Saudi Arabia	Male office workers	12 wk	Cluster RCT; IG: 33; CG: 33	0 (0)	43.5	Weekly SMS text messages targeting self-efficacy and self-regulation of physical activity, workplace computer prompts to take a break, videos of short bursts of exercise, and group and 1-1 educational sessions	No intervention
Alsaleh et al [53], 2016, Jordan	Outpatients with coronary heart disease	26 wk	IG: 71; CG: 85	72 (46)	IG: 57.7; CG: 58	SMS text messages (2/wk for the first 3 mo and 1/wk for the last 3 mo) with reminders to be active and meet personal goals; 1 face-to-face consultation to discuss barriers and facilitators and increase self-efficacy for physical activity; 6 phone calls to provide tailored feedback, review goals, and discuss any arising barriers; and paper diary to self-monitor physical activity	Usual care
Goodarzi et al [54], 2012, Iran	Adults with type 2 diabetes aged >30 y	12 wk	IG: 43; CG: 38	63 (77)	IG: 51; CG: 56.7	4 SMS text messages/wk about exercise, diet, diabetes medication, and importance of self-monitoring blood glucose levels	No intervention

[a]IG: intervention group.

[b]CG: control group.

[c]NR: not reported.

XSL•FO
RenderX

Table 2. Study and intervention characteristics of the included quasi-experiments.

Study, year, and country	Study population	Study length	Sample size, N	Female participants, n (%)	Age (y), mean	Details about the intervention and control (if applicable)	Physical activity and other outcomes
Al-Daghri et al [56], 2022, Saudi Arabia	Adolescents aged 12-16 y	12 mo	1 arm: 2600	343 (53)	14.8	• Educational sessions (every 3 mo) about exercise or diet via Zoom or social media (WhatsApp, Telegram, Facebook, or Twitter), school-based educational sessions about type 2 diabetes and behavioral prevention, pamphlets, booklets, infographic videos, and gamification; participants with diabetes and prediabetes also received tailored advice via phone calls	• NS[a]: physical activity • Decrease in HbA_{1c}[b] in participants with diabetes and prediabetes (*P*<.001)
Ghofranipour et al [57], 2022, Iran	Mothers of children aged 3-5 y	6 wk	1 arm: 13	13 (100)	Mothers: 38; children: 4	• WhatsApp messages and educational videos on reducing sedentary behaviors, increasing physical activity, and promoting a healthy diet	• Increase in active minutes per day (*P*=.02) immediately and 3 mo after the intervention • NS: BMI z score, diet, and quality of life
Ali et al [67], 2021, United Arab Emirates	Female university students with overweight or obesity aged 18-35 y	16 wk	2 arms—website: 54; website+app: 111	246 (100)	22	• Website: self-monitoring of diet and physical activity, educational materials, web-based feedback and counseling, and news and weekly fitness challenges • Website+app: self-monitoring of physical activity (PACER) and diet (MyNetDiary) via app, meeting with nutritionists, and WhatsApp messages	• Website+app group: increase in number of days of moderate physical activity (*P*=.01) and minutes walked (*P*<.001), decrease in BMI (*P*=.04) and body fat (*P*<.001), and increase in nutritional knowledge (*P*<.001) • NS in these domains in the website group • Between-group significant decrease in waist circumference (*P*=.003) • Of 11 mothers, 9 (81.8%) stated that the videos were practical, and 8 (72.7%) found them useful. Participants preferred the materials to be in Arabic, shorter, and delivered via mobile apps only.
Alyousef [58], 2021, Saudi Arabia	Women with type 2 diabetes	8 wk	1 arm: 20	20 (100)	54.8	• WhatsApp messages with links to YouTube exercise videos and a phone call every 2 wk to encourage adherence	• Increase in step count (*P*=.04) • A total of 20 (100%) participants thought that the intervention was practical, they liked the videos, and they said that they would recommend it to family and friends. In total, 10% of the participants suggested a shorter video. Participants identified several barriers to following the YouTube exercises: lack of time, an appropriate place, and fast internet connection.

XSL•FO
RenderX

Study, year, and country	Study population	Study length	Sample size, N	Female participants, n (%)	Age (y), mean	Details about the intervention and control (if applicable)	Physical activity and other outcomes
Biglar Chopoghlo et al [59], 2021, Iran	Female adolescents aged 14-18 y with type 1 diabetes	12 wk	2 arms—IG[c]: 38; CG[d]: 38	76 (100)	15.9	• SMS text messages or MMS[e] with educational contents regarding physical activity, diet, and other diabetes information • CG: SMS text messages with nondiabetes content	• NS in between-group exercises • Between-group difference in self-efficacy (*P*<.001)
Khidir et al [60], 2021, Qatar	University staff and students	2 phases, 16 wk each	288 in phase 1 and 109 in phase 2	NR[f]; 56 women participated in both phases	NR	• Mobile app: self-monitoring of physical activity and calories and fat burned • Face-to-face presentation about benefits of physical activity, 4-mo walking competition, and pedometers	• NS change in daily step count
Yahia and Bayoumi [61], 2021, Egypt	Outpatients with type 2 diabetes	12 wk	1 arm: 150	63 (42)	40.3	• Mobile app (SOKARY): self-monitoring of physical activity, glucose level, and nutrition; medication reminders; and healthy lifestyle advice	• Increased activity level (*P*<.001) • Decreased HbA_{1c} (*P*=.03) and BMI (*P*=.006)
Jorvand et al [62], 2020, Iran	Health care workers	6 mo	2 arms—IG: 59; CG: 55	57 (50)	IG: 37.6; CG: 37.5	• Mobile app (Telegram): educational content about physical activity and reminder messages to exercise; participants could also send pictures of their own exercise to the group • CG: no intervention	• Between-group difference in daily and weekly minutes of exercise (*P*<.001)
Alnasser et al [63], 2019, Saudi Arabia	Women with overweight or obesity	16 wk	1 arm: 240	240 (100)	31	• Mobile app (Twazon): self-monitoring of step count and diet, recommendations for physical activity and diet, and setting of weight goals	• NS: physical activity, weight, waist circumference, and BMI • Usability score: within acceptable range
Lari et al [64], 2018, Iran	Outpatients with type 2 diabetes	12 wk	IG: 40; CG: 40	34 (47)	IG: 46.1; CG: 49.1	• SMS text messages (2-3/d for 2 wk and then 2/wk) with recommendations for physical activity for diabetes and how to overcome barriers and seek social support • CG: usual care	• Between-group difference in MET[g] min/wk (*P*<.001), perceived self-efficacy, barriers (*P*<.001), and family support (*P*=.046) • NS: perceived health status, benefits, and friend support
Peyman et al [65], 2018, Iran	Women	6 mo	2 arms—IG: 180; CG: 180	360 (100)	IG: 33.4; CG: 31.9	• Daily SMS text messages about the importance of physical activity; website with educational content and videos, assessment of physical activity and BMI, suggestions for women-only places for exercise, and chat room; and an educational CD • CG: no intervention	• Between-group difference in MET min/wk and knowledge and attitude about physical activity (*P*<.001)
		6 mo		100 (50)			

XSL•FO
RenderX

Study, year, and country	Study population	Study length	Sample size, N	Female participants, n (%)	Age (y), mean	Details about the intervention and control (if applicable)	Physical activity and other outcomes
Sani et al [66], 2018, Saudi Arabia	People who were diagnosed with type 2 diabetes within the last 5 y		2 arms—IG: 100; CG: 100		NR; 75% were in the 30-49–y age group	• SMS text messages and MMS in Arabic (2/wk) to encourage discussion; monthly meetings (input from specialist physicians, discussion, and peer group interactions); short presentations; video clips; and problem-based learning techniques about exercise, diet, and self-care • CG: usual care	• NS: physical activity, fasting blood glucose, diastolic blood pressure, triglycerides, and low-density lipoprotein • Between-group difference in HbA_{1c} (P<.001), BMI, and systolic blood pressure (P<.001)

[a]NS: not significant.

[b]HbA_{1c}: glycated hemoglobin.

[c]IG: intervention group.

[d]CG: control group.

[e]MMS: Multimedia Messaging Service.

[f]NR: not reported.

[g]MET: metabolic equivalent of task.

XSL•FO
RenderX

Table 3. Study information and summary of users' perspectives and experiences in nonexperimental studies.

Study, year, and country	Study population	Study design	Sample size, N	Female participants, n (%)	Age (y), mean	Mobile technologies examined	Main findings
Al Ansari et al [68], 2023, Saudi Arabia	People aged >15 y	Survey	195	122 (63)	NR[a]; 40% were aged >40 y	Mobile apps	• >50% agreed that the apps they used served all fitness levels. • >80% agreed that it was easy to learn how to use the mobile apps. • >70% agreed that mobile apps enhanced their knowledge of workouts and physical activity. • NS[b]: between male and female participants and between people aged <40 y and people aged >40 y with respect to perceived usefulness and ease of use, attitudes, experiences, and subjective quality. • Participants aged <40 y reported higher perceived ease of use than those aged >40 y.
Altabtabaei and Alhuwail [71], 2021, Kuwait	Students	Interviews	20	NR	NR	Fitness trackers	• Participants' main purposes of use were to lose weight and better understand and increase their physical activity levels. • Barriers to adoption included perceived lack of usefulness, lack of knowledge about potential benefits and how to use fitness trackers, and concerns about battery life and data inaccuracy. • Participants reported both positive (eg, more confidence) and negative (eg, guilt or stress when failing to achieve their activity goals) feelings associated with using fitness trackers.
Bardus et al [69], 2021, Lebanon	Student athletes	Survey	200	70 (35)	20	Mobile apps and fitness trackers	• 53% owned a fitness tracker; the main purpose of use was to facilitate self-monitoring. • Reasons for discontinued use were loss of interest or technical issues.
Zaman et al [70], 2021, Saudi Arabia	People with sleep problems	Survey	45	23 (51)	NR; 51% were aged 20-30 y	Mobile apps	• Participants reportedly improved their physical activity after using a health app (P=.009).

[a]NR: not reported.

[b]NS: not significant.

Description of All Studies

Most of the included studies were conducted in Iran [50-52,54,55,57,59,62,64,65] (10/27, 37%) or Saudi Arabia [42,49,56,58,63,66,68,70] (8/27, 30%). In total, 7% (2/27) of the studies were conducted in Jordan [48,53] and Oman [43,47], and 4% (1/27) of the studies were conducted in Egypt [61], Kuwait [71], Lebanon [69], Qatar [60], and the United Arab Emirates each [67]. The included studies were published between 2012 and 2023. The average duration of experimental studies was 20 (SD 14.4; range 6-52) weeks. The total number of participants was 6141 (1/27, 4% of the studies had 2600 participants). Among the study participants, 46% were women; 7% (2/27) of the studies did not report gender distribution [60,71]. A total of 33% (9/27) of the studies were conducted in populations with chronic conditions, including cardiovascular diseases [48,53,61], diabetes [54,58,59,64,66], and sleep problems [70]. The funding sources and conflicts of interest are summarized in Multimedia Appendix 6 [42,43,47-71].

Description of the Included Experimental Studies

Description of the Interventions

Half (11/22, 50%) of the interventions included a mobile app as a component [43,47-49,55-58,60-63,67]; 36% (4/11) of them used WhatsApp [43,47,49,57,58] (Tables 1 and 2). The other commercial apps used were Samsung Health [48] and Telegram [56,62]. In total, 19% (5/27) of the studies examined apps designed by the authors [55,60,61,63,67]. The main functions of mobile apps were to set goals, self-monitor activity, and receive educational information about physical activity [43,47-49,55-58,60-63,67]. Half (11/22, 50%) of the interventions used SMS text messaging to deliver educational content [42,50-55,59,64-66]; the frequency of delivery varied from 2 messages per day to 1 message per week.

In total, 36% (8/22) of the interventions also included a nonmobile component; the most popular were face-to-face sessions [42,43,47,50,51,53,60,66] (Tables 1 and 2). Other nonmobile components included websites [65,67], phone calls [53], computer prompts [42], pedometers [43,47,60], and an

XSL•FO
RenderX

educational booklet and CD [50,65]. In total, 37% (10/27) of the studies mentioned that the interventions were designed using behavior change theories [42,43,48,51,53,57,58,63,64,67], the most popular being social cognitive theory (4/10, 40%) [42,53,63,67]. Self-efficacy constructs [53,57] and the health promotion model [58,64] were used in 20% (2/10) of the studies each.

Description of the Control Groups

Of the 10 included RCTs, 3 (30%) had a true control group (ie, no intervention) [42,49,54] (Table 1). A total of 30% (3/10) had an active mobile control [48,52,55]. Specifically, in 10% (1/10) of the studies, the control group received usual care (ie, clinical consultation) as well as the Samsung Health app without being told about its features [48]; the intervention was guided by the theory of planned behavior. In another study, the intervention group received daily SMS text messages about physical activity, whereas the control group received educational SMS text messages about cancers; no theory was mentioned [52]. In the third study, the content of the intervention and comparator arms used the same behavior change techniques (ie, information about health consequences), the only difference being the delivery platform (ie, mobile app vs SMS text messaging); no theory was mentioned [55]. A total of 40% (4/10) of the RCTs had nonmobile controls [43,47,50,51,53]; all mentioned the use of a theory. In total, 75% (3/4) of these RCTs involved usual care in the control arm [43,47,50,53], and 25% (1/4) involved face-to-face education [51].

A total of 50% (6/12) of the quasi-experiments also had a control arm, of which 33% (2/6) were true controls (ie, no intervention [62,65]) and 33% (2/6) were usual care [64,66]. A total of 33% (2/6) of the studies involved an active digital control. In one study, the control group received SMS text messages that were not related to physical activity [59]; no theory was mentioned. In another study, the control group had access to a website with the self-monitoring, receiving feedback, goal setting, and information about health consequences behavior change techniques [67]; social cognitive theory was used to develop the intervention.

Engagement and Retention Metrics

None of the included RCTs reported metrics of engagement with the interventions. A total of 25% (3/12) of the quasi-experiments reported engagement metrics [57,60,63]. Specifically, one study reported that 23% (3/13) of mothers watched all videos delivered via WhatsApp (Multimedia Appendix 6) [57]. Another study reported a higher use of mobile apps than pedometers to monitor activity [60]. One study measured use rates and reported that, at 6 months, 55% (26/47) of the participants used the app at least once every 2 weeks [63].

The retention rate in the intervention groups ranged from 18% to 100%. The average retention rate for the intervention arms was 85%, with most falling within the 86% to 100% range. A total of 22% (6/27) of the studies had a 100% retention rate [50,55,57,61,62,65], and 7% (2/27) had a retention rate of <25% [56,63] (Multimedia Appendix 6).

Users' Perspectives in Experimental Studies

Of the 27 studies, 4 (15%) experimental studies (n=3, 75% RCTs [42,47,53] and n=1, 25% quasi-experiments [58]) examined users' perspectives and experiences with the mobile interventions (Table 2 and Multimedia Appendix 7 [42,43,47-67]). Most of the participants reported finding the interventions useful; some reported benefits such as gaining knowledge about how to change their behaviors, receiving reminders to be more active, and building a relationship with clinicians [53]. One study reported barriers to using the mobile intervention such as lack of reliable internet connection and lack of time and an appropriate place to exercise [58]. In one study, participants mentioned a preference for the intervention material to be linguistically adapted [67].

Users' Perspectives in Nonexperimental Studies

Of the 4 nonexperimental studies, 2 (50%) examined the role of mobile apps [68,70], 1 (25%) focused on fitness trackers [71], and 1 (25%) investigated both [69] (Table 3). A total of 50% (2/4) of the studies found that the main purpose of use was to self-monitor their activity [69,71]. Barriers to adoption and use included perceived lack of usefulness, lack of knowledge about potential benefits, loss of interest, and technical issues [69,71]. One survey found that people aged <40 years reported higher perceived ease of use than those aged >40 years [68]. A survey of people with sleep problems reported improved physical activity after using a health app [70].

Quality Assessment

The risk of bias was assessed as low in 3 out of 5 categories in half (5/10, 50%) of the included RCTs [43,48,50,53,55] (Table 4). Half (5/10, 50%) of the included RCTs described a low-risk randomization process [48,51-53,55]. In 20% (2/10) of the trials, the allocation sequence was not concealed until participants were enrolled [42,49]. Of the 10 RCTs, 2 (20%) cluster RCTs were assessed as low risk in an additional domain (ie, "timing of identification or recruitment of participants"). The risk of "deviations from intended interventions" was assessed as low in 40% (4/10) of the studies [43,50,53,55], there were "some concerns" in 40% (4/10) of the studies [42,48,51,52], and the risk was high in 20% (2/10) of the studies as an appropriate analysis (eg, intention-to-treat) was not used. Half (5/10, 50%) of the included RCTs had a low level of incomplete data [42,43,47,48,55], whereas the other half were assessed as having some concerns [49,51-54]. Most of the studies (7/10, 70%) were assessed as having some concerns regarding the measurement of the outcomes because of the self-report nature, and participants (ie, outcome assessors in this case) were aware of the intervention allocation [43,49-51,53-55]. More than half (6/10, 60%) of the included studies scored as having a low risk in the selection of the reported results [43,49-51,53,55], whereas 40% (4/10) of the studies were assessed as having some concerns because of the lack of details regarding a preplanned analysis [42,48,52,54].

XSL•FO
RenderX

Table 4. Risk-of-bias assessment of the included randomized controlled trials. Green: low risk; yellow: some concerns; red: high risk.

Study, year	Randomization process	Deviations from intended interventions	Missing outcome data	Measurement of the outcome	Selection of the reported result	Overall
Abbaspoor et al [51], 2020						
Alghafri et al [43], 2018						
Alsaleh et al [53], 2016						
Alshahrani et al [49], 2021						
Ansari et al [55], 2022						
Eslami et al [50], 2022						
Goodarzi et al [54], 2012						
Parandeh et al [52], 2019						
Quronfulah [42], 2019						
Saleh et al [48], 2022						

Table 5 provides a summary of the quality assessment of the quasi-experimental studies using the JBI checklist for quasi-experiments. All 12 quasi-experimental studies [56-67] made clear what was the cause versus the effect (question 1), had multiple measurements of the outcome both before and after the intervention (question 5), and measured the outcomes of participants in the comparison groups in the same way as those of participants in the intervention groups (question 7). In all except one study (11/12, 92%) where this was unclear [67], participants in the comparison group were similar to participants in the intervention group. In one study [65], it was unclear whether participants in the comparison group received similar treatment; the other 92% (11/12) of the studies met this criterion (question 3). In half (6/12, 50%) of the studies, there was no control group (question 4). Only 8% (1/12) of the studies provided information about whether follow-up was complete and how follow-up data were treated [62] (question 6). In total, 83% (10/12) of the studies measured outcomes in a reliable way; this information was unclear in 17% (2/12) of the studies [57,61] (question 8). Appropriate statistical analyses were conducted in 92% (11/12) of the studies; this was unclear in 8% (1/12) of the studies [61] (question 9).

The survey studies (3/27, 11%) [68-70] were assessed using the JBI checklist for cross-sectional studies (Table 6). The criteria for inclusion were not clearly defined in those 3 studies (question 1). The study participants and settings were described in detail in 67% (2/3) of the studies [68,69] (question 2). Owing to poor reporting, it was unclear in all 3 studies whether exposure and outcomes were measured in a valid and reliable way (questions 3 and 7) and whether confounding factors and strategies to deal with them were identified (questions 5 and 6). In 33% (1/3) of the studies, appropriate statistical analysis was used (question 9).

XSL•FO
RenderX

Table 5. Quality assessment of the included quasi-experimental studies.

Study, year	Q1[a]	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Al-Daghri et al [56], 2022	Y[b]	Y	Y	N[c]	Y	U[d]	Y	Y	Y
Ghofranipour et al [57], 2022	Y	Y	Y	N	Y	U	Y	U	Y
Ali et al [67], 2021	Y	U	Y	Y	Y	N	Y	Y	Y
Alyousef [58], 2021	Y	Y	Y	N	Y	U	Y	Y	Y
Biglar Chopoghlo et al [59], 2021	Y	Y	Y	Y	Y	N	Y	Y	Y
Khidir et al [60], 2021	Y	Y	Y	N	Y	U	Y	Y	Y
Yahia and Bayoumi [61], 2021	Y	Y	Y	N	Y	U	Y	U	U
Jorvand et al [62], 2020	Y	Y	Y	Y	Y	Y	Y	Y	Y
Alnasser et al [63], 2019	Y	Y	Y	N	Y	N	Y	Y	Y
Lari et al [64], 2018	Y	Y	Y	Y	Y	N	Y	Y	Y
Peyman et al [65], 2018	Y	Y	U	Y	Y	U	Y	Y	Y
Sani et al [66], 2018	Y	Y	Y	Y	Y	N	Y	Y	Y

[a]Q: question.
[b]Y: yes.
[c]N: no.
[d]U: unclear.

Table 6. Quality assessment of the included survey studies.

Study, year	Q1[a]	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Al Ansari et al [68], 2023	N[b]	Y[c]	U[d]	N/A[e]	U	U	U	U
Bardus et al [69], 2021	N	Y	U	N/A	U	U	U	Y
Zaman et al [70], 2021	N	N	U	N/A	U	U	U	U

[a]Q: question.
[b]N: no.
[c]Y: yes.
[d]U: unclear.
[e]N/A: not applicable.

The interview study [71] was assessed using the JBI checklist for qualitative research. It was assessed as meeting the criteria in 5 domains (ie, congruity between the research methodology and the research question, congruity between the research methodology and the methods to collect data, congruity between the research methodology and the representation and analysis of data, representation of participants and their voices, and ethics approval by an appropriate body). In total, 3 domains were unclear (ie, congruity between the stated philosophical perspective and the research methodology, congruity between the research methodology and the interpretation of results, and the relationship between the conclusions and the analysis or interpretation of the data). There was no statement locating the researcher culturally or theoretically or on the influence of the researcher on the study.

Meta-Analysis of RCTs and Outcomes in Quasi-Experiments

Of the 10 included RCTs, 7 (70%) were deemed eligible for inclusion in the meta-analysis of physical activity outcomes. Specifically, 20% (2/10) of the RCTs were excluded because they had a mobile component in the control arm [48,55], and 10% (1/10) were excluded for only reporting behavioral practices regarding diabetes care (combining physical activity and other health measures) [54]. In line with the strategies for data synthesis, for 71% (5/7) of the studies [43,49-52], the measures of metabolic equivalent of task minutes per week were included in the meta-analysis; for the other 29% (2/7) of the studies, the duration of moderate to vigorous physical activity was included [42,53].

The meta-analysis showed a positive effect of mobile interventions on physical activity outcomes (standardized mean difference=0.45, 95% CI 0.17-0.73; Figure 2 [42,43,49-53]). The I^2 was 74%, indicating a high heterogeneity. One study in particular [42] seemed to be the source of high heterogeneity. This was the only study that delivered an intervention in the workplace, with part of the intervention including regular computer prompts throughout the day. The high frequency of the computer prompts might have had an effect on behavior changes. The funnel plot appeared to indicate signs of publication bias (Multimedia Appendix 8). The Egger test had

XSL•FO
RenderX

an intercept of 3.65 (P=.02), indicating possible publication bias. The trim-and-fill method imputed an effect size of 0.30 (95% CI −0.65 to 1.25), indicating a possible nonsignificant effect. In total, 3 sensitivity analyses were conducted, including only (1) studies that were randomized at the individual level and (2) studies that had a low risk of bias in at least 3 categories and excluding the study by Quronfulah [42] as an outlier. All 3 sensitivity analyses showed a statistically significant positive result for physical activity (P<.05; Multimedia Appendix 9). Subgroup analyses were conducted among studies with patients who were chronically ill and studies with healthy individuals; both reported nonsignificant changes in physical activity (Multimedia Appendix 9). Only 20% (2/10) of the RCTs reported measures regarding sedentary behaviors; both reported a greater reduction in sedentary time in the intervention group than in the control group [42,43] (Multimedia Appendix 7).

Quasi-experiments were not included in the meta-analysis because of the generally lower quality and lack of a control group. Over half (7/12, 58%) of the quasi-experiments reported significant changes in physical activity [57,58,61,62,64,65,67]. Other anthropometric (eg, BMI and weight) or clinical (eg, glycated hemoglobin) outcomes were also reported (Table 2). It is worth noting that most experimental studies (19/22, 86%) measured physical activity outcomes using a validated questionnaire (eg, the International Physical Activity Questionnaire); only 20% (2/10) of the RCTs [42,43] and 8% (1/12) of the quasi-experiments [48] (3/27, 11% in total) used an objective method (eg, accelerometer).

Figure 2. Forest plot of effect sizes and 95% CIs of physical activity outcomes ordered by descending effect size of the individual studies. SMD: standardized mean difference.

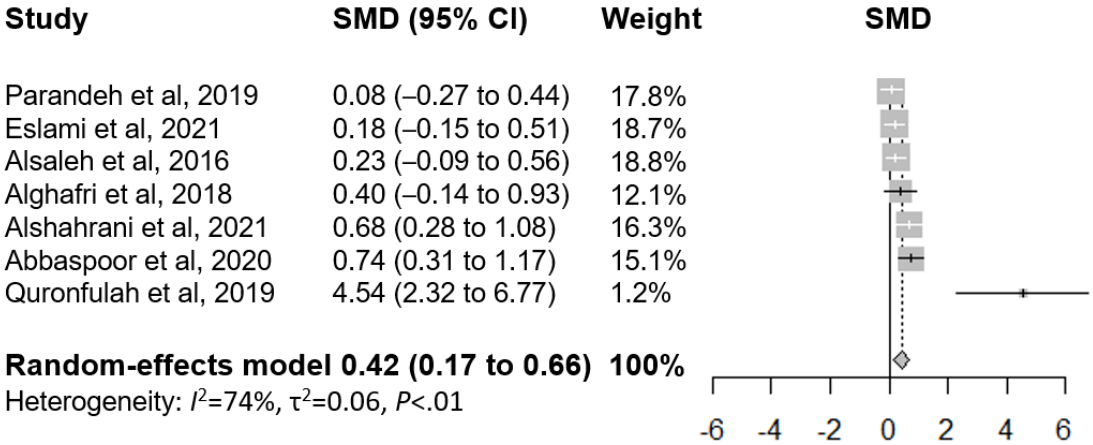

Discussion

Principal Findings

Our review revealed that the use of mobile technologies for physical activity and sedentary behaviors in the MENA region is at an early stage of research given the predominance of small and heterogeneous studies and very few RCTs. The meta-analysis provided preliminary evidence of a small to moderate positive effect of mobile interventions on physical activity in the short term; only 7% (2/27) of the studies measured sedentary behaviors, reporting positive outcomes from the interventions. It is important to note that this evidence is of low to moderate quality according to the Grading of Recommendations Assessment, Development, and Evaluation system [45] because of the high risk of bias in half (5/10, 50%) of the assessed RCTs, heterogeneity in intervention delivery and context, short study duration, possible small-study effects, and publication bias. The reporting standards of the included studies varied widely, and only a small number of studies (3/22, 14%) provided metrics of engagement with the mobile interventions. Nonexperimental studies revealed that users mainly used mobile apps and fitness trackers for self-monitoring of physical activity. Users also revealed several barriers to technology adoption; few studies (1/27, 4%) evaluated the implementation outcomes (eg, context-specific or cultural appropriateness) of the interventions.

Comparison With Existing Literature

This systematic review and meta-analysis focused on the MENA region and provided insights into the effectiveness of mobile technologies on physical activity and sedentary behaviors. Similar to the findings of our meta-analysis, a number of systematic reviews have found a small to moderate effect of mobile technologies on behavioral outcomes [18,20-31]; however, these reviews did not focus on a specific region, and thus, evidence applicable to the MENA region was limited. Furthermore, as these reviews were not MENA specific, it is unclear what would be the most promising interventions given the unique barriers that the MENA region faces. Our review found that particularly favorable interventions were those that conducted user needs assessments, considered daily lives and cultural contexts, and had multicomponent approaches. These findings are in line with those of a review focused on the Asian population that found that Asian apps are largely culturally adapted and multifunctional [72]. A few reviews have explored the use of technologies in specific regions, such as low- and middle-income countries or Asia [72,73], which also showed promising evidence of using these technologies for physical activity and other behavior changes. One systematic review explored physical activity interventions in only 6 Arabian Gulf countries [32] and did not exclusively examine mobile interventions or explore interventions for sedentariness. This review found limited evidence suggesting that pedometer-based interventions encouraging step counting and walking were

XSL•FO
RenderX

effective in promoting physical activity, which might suggest the potential of using mobile technologies that can automate self-monitoring of behavior [32]. Thus, our review expands on existing evidence by investigating the use of mobile interventions in MENA countries for both physical activity and sedentary behaviors and, hence, is able to assess whether mobile technologies were effective for this population and what the users' perspectives were. Combined with the findings from previous reviews, there is evidence suggesting that mobile technologies might be helpful in changing physical activity; however, these interventions likely need to be tailored to users' daily lives and cultural contexts.

Our meta-analysis suggests that mobile technologies may be promising for changing physical activity and sedentary behavior in the MENA region. Particularly favorable interventions were those that conducted user needs assessments, considered daily lives and cultural contexts, and had multicomponent approaches. For example, one study assessed users' preferred method of communication and intervention frequency to tailor the design accordingly, resulting in a high level of acceptability and perceived usefulness [42]. Interestingly, WhatsApp was identified as a preferred platform and used in several studies to deliver educational content or send reminders [43,47,49,57,58]. Conducting needs assessments during intervention development will allow researchers and policy makers to determine whether existing technologies can be leveraged or whether additional apps or features are required to meet users' needs. Incorporating cultural contexts, such as scheduling intervention messages during cultural or health events (eg, Ramadhan and World Hypertension Day) [43], was also a promising approach. In addition, successful interventions combined mobile technologies with face-to-face consultations, suggesting that mobile technologies can complement the role of periodic in-person consultations by delivering more frequent support in daily contexts. It is worth noting that, in the meta-analysis, studies that had a true control group also tended to report larger effect sizes and significant results. Overall, our findings emphasize the potential of mobile technologies to promote behavior changes in the MENA region, necessitating strategies that consider user needs, cultural fit, and multicomponent approaches.

Our review revealed several gaps involving country-specific contexts, targeted populations, and behaviors that should be addressed by researchers and policy makers. First, two-thirds of the included studies (18/27, 67%) were conducted in Saudi Arabia and Iran, and thus, evidence on the use of mobile technologies in other MENA countries was limited. It is likely that more research has been conducted in stable and high-income countries given the diversity of countries in the MENA region regarding income level, economic and social stability [2], and mobile penetration rate [14]. This finding flags the issues of equitable access to mobile technologies across the region, potentially worsening the digital divide and widening health gaps [33-35]. The benefits of mobile health will be limited if it can only reach people with a high socioeconomic status. Thus, concentrated efforts are essential to increase technology access across the region and promote broader research initiatives across countries and contexts.

Second, evidence on mobile interventions for children and adolescents remains limited, with only one included intervention targeting mothers with the ultimate aim of changing preschoolers' physical activity and diet. Given the importance of promoting healthy behaviors from an early age, policy makers should invest in the development and evaluation of mobile interventions for children and adolescents in the MENA region. Third, only 7% (2/27) of the studies intervened on and measured sedentary behaviors, highlighting the need for future research and investment from policy makers to address sedentariness so as to holistically address inactive lifestyles.

In addition, none of the experimental studies examined fitness trackers as part of the intervention despite interest in these devices (as reported in 2/4, 50% of the nonexperimental studies [69,71]) and existing evidence of their positive effects on physical activity [18,20,22,23,25]. Finally, most RCTs (7/10, 70%) measured behavioral outcomes using self-reporting methods (eg, validated questionnaires). Future research might consider using objective measures provided by mobile apps and fitness trackers, which can improve the accuracy and reliability of the data and provide more robust evidence for policy decision-making.

The reporting standards of the included studies in our review varied greatly, with few details provided regarding the intervention, study procedure, and methodology in some studies. Notably, few of the included experimental studies (3/22, 14%) assessed user engagement even though a large body of research has suggested that engagement with digital interventions is a precondition for effectiveness and highlighted issues of high dropout or nonuse attrition in mobile interventions. Future studies should adhere to reporting guidelines (eg, CONSORT [Consolidated Standards of Reporting Trials], STROBE [Strengthening the Reporting of Observational Studies in Epidemiology], and COREQ [Consolidated Criteria for Reporting Qualitative Research]) [74-76] to enable evidence synthesis and assess engagement metrics consistently to allow for future evaluation of the right "dose" of use of these mobile technologies for effectiveness.

Finally, our review also identified qualitative evidence on barriers to adoption and user preferences that policy makers and researchers should consider. Factors such as perceived lack of usefulness, loss of interest, and technical issues can hinder the effectiveness of mobile interventions [77,78]. In addition, although some participants mentioned a preference for intervention material being culturally and linguistically adapted, few studies (1/27, 4%) examined the context fit or cultural appropriateness of the interventions, highlighting the need for future evaluation of these implementation outcomes.

Strengths and Limitations

Our study has several strengths. We followed a prespecified protocol registered in the PROSPERO database. Our search included peer-reviewed and gray literature. We also hand searched related reviews to ensure that relevant studies were captured. Data extraction and risk-of-bias assessment were conducted by 2 reviewers, and the authors were contacted for additional information. Finally, we conducted several sensitivity analyses (which were consistent with our main results) and

XSL•FO
RenderX

assessed risk of bias and publication bias to better understand the limitations of our findings.

Our findings should be interpreted within the context of the study's limitations. The poor reporting of the included studies affected our synthesis capability. Owing to the small number of studies targeting sedentary behaviors, it was not possible to conduct a meta-analysis of this behavioral outcome. The meta-analysis findings were affected by the quality of the included studies, including a large proportion of self-reported outcomes, small sample sizes, and possible publication bias. Our data synthesis strategy selected one outcome from each study to be included in the meta-analysis; future research might consider performing a multilevel meta-analysis to include all reported outcomes of the studies [79]. In addition, as most of the included studies (18/27, 67%) were conducted in Saudi Arabia or Iran, it is important to acknowledge the potential limitations in generalizing the findings to other countries in the MENA region. However, the insights derived from this review can still serve as a valuable guide for future research endeavors and inform policy-making processes in other MENA countries.

Implications

Our findings have important implications for policy, practice, and future research in the MENA region. First, regarding policy implications, policy makers should support and fund RCTs with longer durations to determine the long-term effectiveness of mobile technologies on physical activity and sedentary behavior. Investment should also be made in technology infrastructure and research initiatives in countries with lower socioeconomic status to promote equitable access to mobile technologies across the MENA region. In addition, policy makers need to ensure that interventions are culturally sensitive and linguistically adapted to enhance their acceptability and effectiveness. Finally, it is important to recognize that mobile technologies alone are unlikely to address behavioral challenges in the MENA region given the variability in mobile ownership and social and economic conditions [2,14,80,81]. Thus, policy makers will need to direct efforts into designing and evaluating multifaceted interventions that can appropriately target physical activity and sedentary behaviors while considering the diverse country contexts.

The evidence of the preliminary effectiveness of mobile interventions found in this review needs to be supported by future rigorous evaluations, with the ultimate goal of assisting clinical practice. If there is sufficient, high-quality evidence, clinicians may consider discussing the use of mobile interventions with people who need to change their activity levels as part of a shared decision-making process [82,83]. Potentially, mobile technologies can enhance patient-clinician collaboration by capturing data to facilitate period review while also empowering individuals to manage their health more actively [84]. As research in the MENA region continues to evolve, evidence of the effectiveness of mobile technologies can be used to determine whether their use can become part of routine clinical care. It is important to note that, although technology prescription is a promising prospect, clinicians have reported several barriers to this practice—the most prominent concern being the lack of knowledge of prescribable technologies and lack of reliable sources to access this information [85,86]. Therefore, a nationally accessible repository of vetted and curated technologies for health care professionals is needed to promote the sustainability and scalability of mobile technologies in clinical practice [85,86].

Implications for future research encompass the need for well-designed RCTs, adherence to reporting standards, and assessment of implementation outcomes. First, well-designed and fully powered RCTs are needed to provide high-quality evidence on the effectiveness of mobile technologies on physical activity and sedentary behaviors, especially over a long-term follow-up. Second, it is crucial for future studies to adhere to existing reporting guidelines [74,75,87] to facilitate evidence synthesis on the most effective intervention, "dosage," or delivery channel. Finally, researchers should consistently assess and report intervention engagement, acceptability, and implementation outcomes (eg, cultural fit, sustainability, and cost-effectiveness) [76] to determine the viability of mobile technologies for behavior change and successful implementation in the MENA context.

Conclusions

Our systematic review and meta-analysis found that research in the MENA region on the use of mobile interventions for physical activity and sedentary behaviors is in its early stages, with preliminary evidence indicating their effectiveness. However, the studies varied greatly in terms of intervention components, methodology, and study quality, and therefore, the findings must be interpreted with caution. Policy makers and researchers need to invest in high-quality studies to evaluate the effectiveness, engagement, and implementation process of mobile interventions in the MENA region.

Acknowledgments

This work was supported by the King Faisal Specialist Hospital and Research Center and the World Bank. Rekha Menon, Practice Manager at the World Bank, and Issam Abousleiman, Country Director for the Gulf Cooperation Council Countries at the World Bank, provided input and support on the manuscript. The authors would like to thank Ms Isabelle Raison for her input on the search strategy. Funding was provided by the Ministry of Finance of Saudi Arabia under the World Bank Reimbursable Advisory Services program (P179873). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the King Faisal Specialist Hospital and Research Center or the World Bank.

XSL•FO
RenderX

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to study conceptualization and design. HLT, AA, NZA, RAB, and RA contributed to article screening. HLT, AA, RAB, and RA contributed to data extraction. HLT, AA, and NZA contributed to quality assessment. HLT contributed to the meta-analysis. HLT and NZA contributed to the first draft. All authors contributed to critical revision of the drafts for important intellectual content. All authors contributed to final approval of the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement.
[PDF File (Adobe PDF File), 412 KB - jmir_v26i1e53651_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 24 KB - jmir_v26i1e53651_app2.docx]

Multimedia Appendix 3
Eligibility criteria.
[DOCX File , 20 KB - jmir_v26i1e53651_app3.docx]

Multimedia Appendix 4
Calculating the effective sample size of cluster randomized controlled trials for the meta-analysis using Cochrane guidelines.
[DOCX File , 12 KB - jmir_v26i1e53651_app4.docx]

Multimedia Appendix 5
List of excluded studies after full-text review for not meeting the inclusion criteria regarding intervention or outcome.
[DOCX File , 15 KB - jmir_v26i1e53651_app5.docx]

Multimedia Appendix 6
Information about funding sources and conflicts of interest of the included studies.
[DOCX File , 13 KB - jmir_v26i1e53651_app6.docx]

Multimedia Appendix 7
Retention rates, engagement metrics, and other outcomes from the experimental studies.
[DOCX File , 19 KB - jmir_v26i1e53651_app7.docx]

Multimedia Appendix 8
Funnel plot of SE by standardized mean difference.
[DOCX File , 20 KB - jmir_v26i1e53651_app8.docx]

Multimedia Appendix 9
Sensitivity and subgroup analyses.
[DOCX File , 13 KB - jmir_v26i1e53651_app9.docx]

References

1. Global health estimates: leading causes of death. World Health Organization. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death [accessed 2023-04-19]
2. Middle East and North Africa. The World Bank. URL: https://www.worldbank.org/en/region/mena [accessed 2023-04-24]
3. Cause of death, by non-communicable diseases (% of total) - Middle East and North Africa. The World Bank. 2020. URL: https://data.worldbank.org/indicator/SH.DTH.NCOM.ZS?locations=ZQ [accessed 2023-04-24]

XSL•FO
RenderX

4. Katzmarzyk PT, Friedenreich C, Shiroma EJ, Lee IM. Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries. Br J Sports Med 2022 Jan 29;56(2):101-106 [FREE Full text] [doi: 10.1136/bjsports-2020-103640] [Medline: 33782046]
5. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet 2012 Jul 21;380(9838):219-229 [FREE Full text] [doi: 10.1016/S0140-6736(12)61031-9] [Medline: 22818936]
6. Park JH, Moon JH, Kim HJ, Kong MH, Oh YH. Sedentary lifestyle: overview of updated evidence of potential health risks. Korean J Fam Med 2020 Nov;41(6):365-373 [FREE Full text] [doi: 10.4082/kjfm.20.0165] [Medline: 33242381]
7. Li S, Lear SA, Rangarajan S, Hu B, Yin L, Bangdiwala SI, et al. Association of sitting time with mortality and cardiovascular events in high-income, middle-income, and low-income countries. JAMA Cardiol 2022 Aug 01;7(8):796-807 [FREE Full text] [doi: 10.1001/jamacardio.2022.1581] [Medline: 35704349]
8. Biswas A, Oh PI, Faulkner GE, Bajaj RR, Silver MA, Mitchell MS, et al. Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: a systematic review and meta-analysis. Ann Intern Med 2015 Jan 20;162(2):123-132. [doi: 10.7326/M14-1651] [Medline: 25599350]
9. Guthold R, Stevens GA, Riley LM, Bull FC. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1.6 million participants. Lancet Child Adolesc Health 2020 Jan;4(1):23-35. [doi: 10.1016/s2352-4642(19)30323-2]
10. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1.9 million participants. Lancet Glob Health 2018 Oct;6(10):e1077-e1086. [doi: 10.1016/s2214-109x(18)30357-7]
11. Chaabane S, Chaabna K, Abraham A, Mamtani R, Cheema S. Physical activity and sedentary behaviour in the Middle East and North Africa: an overview of systematic reviews and meta-analysis. Sci Rep 2020 Jun 09;10(1):9363 [FREE Full text] [doi: 10.1038/s41598-020-66163-x] [Medline: 32518254]
12. Chaabane S, Chaabna K, Doraiswamy S, Mamtani R, Cheema S. Barriers and facilitators associated with physical activity in the Middle East and North Africa region: a systematic overview. Int J Environ Res Public Health 2021 Feb 09;18(4):1647 [FREE Full text] [doi: 10.3390/ijerph18041647] [Medline: 33572229]
13. Mobile consumption in a post-growth world: global Mobile Consumer Survey 2019. Deloitte. URL: https://www2.deloitte.com/xe/en/pages/technology-media-and-telecommunications/articles/global-mobile-consumer-survey-2019.html [accessed 2023-04-24]
14. Mobile economy report Middle East and North Africa. GSM Association. 2023. URL: https://www.gsma.com/mobileeconomy/mena/ [accessed 2023-04-24]
15. Steinhubl SR, Muse ED, Topol EJ. Can mobile health technologies transform health care? JAMA 2013 Dec 11;310(22):2395-2396. [doi: 10.1001/jama.2013.281078] [Medline: 24158428]
16. Eapen ZJ, Peterson ED. Can mobile health applications facilitate meaningful behavior change?: time for answers. JAMA 2015 Sep 22;314(12):1236-1237. [doi: 10.1001/jama.2015.11067] [Medline: 26393844]
17. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
18. Tong HL, Quiroz JC, Kocaballi AB, Fat SC, Dao KP, Gehringer H, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021 Jul;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
19. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
20. Laranjo L, Ding D, Heleno B, Kocaballi B, Quiroz JC, Tong HL, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr 21;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
21. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019 Mar 19;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
22. Tang MS, Moore K, McGavigan A, Clark RA, Ganesan AN. Effectiveness of wearable trackers on physical activity in healthy adults: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e15576 [FREE Full text] [doi: 10.2196/15576] [Medline: 32706685]
23. Brickwood KJ, Watson G, O'Brien J, Williams AD. Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Apr 12;7(4):e11819 [FREE Full text] [doi: 10.2196/11819] [Medline: 30977740]
24. McDonough DJ, Su X, Gao Z. Health wearable devices for weight and BMI reduction in individuals with overweight/obesity and chronic comorbidities: systematic review and network meta-analysis. Br J Sports Med 2021 Aug;55(16):917-925 [FREE Full text] [doi: 10.1136/bjsports-2020-103594] [Medline: 33731385]

XSL•FO
RenderX

25. Larsen RT, Wagner V, Korfitsen CB, Keller C, Juhl CB, Langberg H, et al. Effectiveness of physical activity monitors in adults: systematic review and meta-analysis. BMJ 2022 Jan 26;376:e068047 [FREE Full text] [doi: 10.1136/bmj-2021-068047] [Medline: 35082116]
26. Shariful Islam SM, Farmer AJ, Bobrow K, Maddison R, Whittaker R, Pfaeffli Dale LA, et al. Mobile phone text-messaging interventions aimed to prevent cardiovascular diseases (Text2PreventCVD): systematic review and individual patient data meta-analysis. Open Heart 2019;6(2):e001017 [FREE Full text] [doi: 10.1136/openhrt-2019-001017] [Medline: 31673381]
27. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
28. Baumann H, Fiedler J, Wunsch K, Woll A, Wollesen B. mHealth interventions to reduce physical inactivity and sedentary behavior in children and adolescents: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 May 11;10(5):e35920 [FREE Full text] [doi: 10.2196/35920] [Medline: 35544294]
29. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699 [FREE Full text] [doi: 10.2196/26699] [Medline: 33811021]
30. Daryabeygi-Khotbehsara R, Shariful Islam SM, Dunstan D, McVicar J, Abdelrazek M, Maddison R. Smartphone-based interventions to reduce sedentary behavior and promote physical activity using integrated dynamic models: systematic review. J Med Internet Res 2021 Sep 13;23(9):e26315 [FREE Full text] [doi: 10.2196/26315] [Medline: 34515637]
31. Ludwig K, Arthur R, Sculthorpe N, Fountain H, Buchan DS. Text messaging interventions for improvement in physical activity and sedentary behavior in youth: systematic review. JMIR Mhealth Uhealth 2018 Sep 17;6(9):e10799 [FREE Full text] [doi: 10.2196/10799] [Medline: 30224335]
32. Nash EA, Critchley JA, Pearson F, Awad SF, Abu-Raddad LJ, Abu-Hijleh FM, et al. A systematic review of interventions to promote physical activity in six Gulf countries. PLoS One 2021 Oct 28;16(10):e0259058 [FREE Full text] [doi: 10.1371/journal.pone.0259058] [Medline: 34710147]
33. Western MJ, Armstrong ME, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Nov 09;18(1):148 [FREE Full text] [doi: 10.1186/s12966-021-01218-4] [Medline: 34753490]
34. Western MJ, Smit ES, Neter E, Busse H, Ainsworth B, König LM. Charting new territories in health psychology: a reflection on the EHPS 2022 'digital divide' hybrid roundtable by chairs, presenters, and participants. Eur Health Psychol 2023 May 11;23(2):1017.
35. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century Cures Act era. JAMA 2020 Jun 16;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
36. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906 [FREE Full text] [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
37. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
38. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
39. Higgins JP, Savović J, Page MJ, Elbers RG, Sterne JA. Assessing risk of bias in a randomized trial. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; Sep 20, 2019.
40. Critical appraisal tools. Joanna Briggs Institute. URL: https://jbi.global/critical-appraisal-tools [accessed 2023-11-26]
41. Higgins JP, Eldridge S, Li T. Including variants on randomized trials. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2019.
42. Quronfulah BS. Development and implementation of a health promotion intervention to reduce sedentary behaviour among male office workers in Saudi Arabia: the slim (sit less, impress and motivate) study. Queensland University of Technology. 2019. URL: https://eprints.qut.edu.au/134422/ [accessed 2024-02-15]
43. Alghafri TS, Alharthi SM, Al-Farsi Y, Alrawahi AH, Bannerman E, Craigie AM, et al. 'MOVEdiabetes': a cluster randomized controlled trial to increase physical activity in adults with type 2 diabetes in primary health in Oman. BMJ Open Diabetes Res Care 2018 Oct 31;6(1):e000605 [FREE Full text] [doi: 10.1136/bmjdrc-2018-000605] [Medline: 30487976]
44. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun;56(2):455-463 [FREE Full text] [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
45. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]

XSL•FO
RenderX

46. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2019. URL: https://www.R-project.org/ [accessed 2024-02-15]
47. Alghafri TS, Al Harthi SM, Al-Ajmi F, Al-Farsi Y, Craigie AM, Bannerman E, et al. Acceptability of the "MOVEdiabetes" physical activity intervention in diabetes primary care settings in Oman: findings from participants and practitioners. BMC Public Health 2020 Jun 08;20(1):887 [FREE Full text] [doi: 10.1186/s12889-020-09029-1] [Medline: 32513161]
48. Saleh ZT, Elshatarat RA, Elhefnawy KA, Helmi Elneblawi N, Abu Raddaha AH, Al-Za'areer MS, et al. Effect of a home-based mobile health app intervention on physical activity levels in patients with heart failure: a randomized controlled trial. J Cardiovasc Nurs 2023;38(2):128-139. [doi: 10.1097/JCN.0000000000000911] [Medline: 35389920]
49. Alshahrani A, Siddiqui A, Khalil S, Farag S, Alshahrani N, Alsabaani A, et al. WhatsApp-based intervention for promoting physical activity among female college students, Saudi Arabia: a randomized controlled trial. East Mediterr Health J 2021 Aug 26;27(8):782-789 [FREE Full text] [doi: 10.26719/emhj.21.012] [Medline: 34486714]
50. Eslami E, Mohammad Alizadeh Charandabi S, Farshbaf Khalili A, Asghari Jafarabadi M, Mirghafourvand M. The effect of a lifestyle training package on physical activity and nutritional status in obese and overweight pregnant women: a randomized controlled clinical trial. Int J Nurs Pract 2022 Dec 27;28(6):e12992. [doi: 10.1111/ijn.12992] [Medline: 34313366]
51. Abbaspoor Z, Amani A, Afshari P, Jafarirad S. The effect of education through mobile phone short message service on promoting self-care in pre-diabetic pregnant women: a randomized controlled trial. J Telemed Telecare 2018 Sep 07;26(4):200-206. [doi: 10.1177/1357633x18791419]
52. Parandeh L, Shafaie FS, Malakouti J, Mirghafourvand M, Asghari-Jafarabadi M. The effect of educational text message based on health belief model on osteoporosis preventive behaviors in women: a randomized controlled clinical trial. Women Health 2019 Apr 08;59(10):1128-1140. [doi: 10.1080/03630242.2019.1590495] [Medline: 30955478]
53. Alsaleh E, Windle R, Blake H. Behavioural intervention to increase physical activity in adults with coronary heart disease in Jordan. BMC Public Health 2016 Jul 26;16(1):643 [FREE Full text] [doi: 10.1186/s12889-016-3313-5] [Medline: 27460675]
54. Goodarzi M, Ebrahimzadeh I, Rabi A, Saedipoor B, Jafarabadi MA. Impact of distance education via mobile phone text messaging on knowledge, attitude, practice and self efficacy of patients with type 2 diabetes mellitus in Iran. J Diabetes Metab Disord 2012 Aug 31;11(1):10 [FREE Full text] [doi: 10.1186/2251-6581-11-10] [Medline: 23497632]
55. Ansari K, Afshari P, Abedi P, Haghighizadeh M. Comparing the effects of text messaging and mobile social networking on physical activity and anthropometric indices of middle-aged women: a randomized controlled trial. BMC Womens Health 2022 Jan 26;22(1):18 [FREE Full text] [doi: 10.1186/s12905-022-01598-0] [Medline: 35081935]
56. Al-Daghri NM, Amer OE, Hameidi A, Alfawaz H, Alharbi M, Khattak MN, et al. Effects of a 12-month hybrid (in-person + virtual) education program in the glycemic status of Arab youth. Nutrients 2022 Apr 22;14(9):1759 [FREE Full text] [doi: 10.3390/nu14091759] [Medline: 35565727]
57. Ghofranipour F, Hamzavi Zarghani N, Mohammadi E, Mehrizi AA, Tavousi M, De Craemer M, et al. An internet-based educational intervention for mothers targeting preschoolers' weight management promotion (PWMP): a pilot study. BMC Public Health 2022 Nov 29;22(1):2220 [FREE Full text] [doi: 10.1186/s12889-022-14543-5] [Medline: 36447163]
58. Alyousef ZM. Diabetes and physical activity among Saudi women. University of Texas at Arlington. 2021 Jul 14. URL: https://rc.library.uta.edu/uta-ir/handle/10106/29979 [accessed 2024-02-15]
59. Biglar Chopoghlo S, Hosseinkhani A, Khedmat L, Zaki-Nejad M, Puryaghoob M. The self-efficacy improvement in adolescent girls with type 1 diabetes mellitus with self-care education through mobile-based social networking. Int J Diabetes Dev Ctries 2021 Feb 10;41(4):676-682. [doi: 10.1007/s13410-021-00929-5]
60. Khidir ED, Stănescu M, Al Sayegh S. Using mobile technology to evaluate the active lifestyle of adults from campuses in Qatar. In: Proceedings of the International Scientific Conference eLearning and Software for Education. 2021 Presented at: eLSE 2021; April 22-23, 2021; Bucharest, Romania. [doi: 10.12753/2066-026x-21-194]
61. Yahia EA, Bayoumi MM. Effectiveness of using SOKARY mobile application on the compliance of patients with type II diabetes: a quasi-experimental study. Pak J Med Health Sci 2021 Jul;15(7):2001-2005 [FREE Full text] [doi: 10.53350/pjmhs211572001]
62. Jorvand R, Ghofranipour F, HaeriMehrizi A, Tavousi M. Evaluating the impact of HBM-based education on exercise among health care workers: the usage of mobile applications in Iran. BMC Public Health 2020 Apr 22;20(1):546 [FREE Full text] [doi: 10.1186/s12889-020-08668-8] [Medline: 32321484]
63. Alnasser A, Kyle J, Aloumi N, Al-Khalifa A, Marais D. The Twazon Arabic weight loss app: app-based intervention for Saudi women with obesity. JMIR Mhealth Uhealth 2019 May 28;7(5):e10923 [FREE Full text] [doi: 10.2196/10923] [Medline: 31140444]
64. Lari H, Noroozi A, Tahmasebi R. Impact of short message service (SMS) education based on a health promotion model on the physical activity of patients with type II diabetes. Malays J Med Sci 2018 May;25(3):67-77 [FREE Full text] [doi: 10.21315/mjms2018.25.3.7] [Medline: 30899188]
65. Peyman N, Rezai-Rad M, Tehrani H, Gholian-Aval M, Vahedian-Shahroodi M, Heidarian Miri H. Digital media-based health intervention on the promotion of women's physical activity: a quasi-experimental study. BMC Public Health 2018 Jan 15;18(1):134 [FREE Full text] [doi: 10.1186/s12889-018-5025-5] [Medline: 29334970]

XSL•FO
RenderX

66. Sani M, Makeen A, Albasheer OB, Solan YM, Mahfouz MS. Effect of telemedicine messages integrated with peer group support on glycemic control in type 2 diabetics, Kingdom of Saudi Arabia. Int J Diabetes Dev Ctries 2018 Jan 29;38(4):495-501. [doi: 10.1007/s13410-018-0608-3]
67. Ali HI, Attlee A, Alhebshi S, Elmi F, Al Dhaheri AS, Stojanovska L, et al. Feasibility study of a newly developed technology-mediated lifestyle intervention for overweight and obese young adults. Nutrients 2021 Jul 26;13(8):2547 [FREE Full text] [doi: 10.3390/nu13082547] [Medline: 34444707]
68. Al Ansari FS, Alfayez A, Alsalman D, Alanezi F, Alhodaib H, Al-Rayes S, et al. Using mobile health applications to enhance physical activity in Saudi Arabia: a cross-sectional study on users' perceptions. Int Health 2023 Jan 03;15(1):47-55 [FREE Full text] [doi: 10.1093/inthealth/ihac008] [Medline: 35348719]
69. Bardus M, Borgi C, El-Harakeh M, Gherbal T, Kharroubi S, Fares EJ. Exploring the use of mobile and wearable technology among university student athletes in Lebanon: a cross-sectional study. Sensors (Basel) 2021 Jun 30;21(13):4472 [FREE Full text] [doi: 10.3390/s21134472] [Medline: 34208798]
70. Zaman TU, Alqahtani F, Alsairafi A, Adetunji HA, Al Areefi M, Al Akhram N, et al. The role of mobile health applications in preventing sleep-related health problems – a pilot study in the Makkah city of Saudi Arabia. Int J Healthc Manag 2020 Apr 28;14(4):1230-1236. [doi: 10.1080/20479700.2020.1756103]
71. Altabtabaei R, Alhuwail D. Exploring the use and adoption of wearable physical activity trackers in oil-rich nations: a qualitative study of youth perspectives from Kuwait. Stud Health Technol Inform 2022 Jun 06;290:1100-1101. [doi: 10.3233/SHTI220287] [Medline: 35673225]
72. Ang SM, Chen J, Liew JH, Johal J, Dan YY, Allman-Farinelli M, et al. Efficacy of interventions that incorporate mobile apps in facilitating weight loss and health behavior change in the Asian population: systematic review and meta-analysis. J Med Internet Res 2021 Nov 16;23(11):e28185 [FREE Full text] [doi: 10.2196/28185] [Medline: 34783674]
73. Edney S, Chua XH, Müller AM, Kui KY, Müller-Riemenschneider F. mHealth interventions targeting movement behaviors in Asia: a scoping review. Obes Rev 2022 Apr;23(4):e13396 [FREE Full text] [doi: 10.1111/obr.13396] [Medline: 34927346]
74. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
75. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
76. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
77. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
78. Yang X, Ma L, Zhao X, Kankanhalli A. Factors influencing user's adherence to physical activity applications: a scoping literature review and future directions. Int J Med Inform 2020 Feb;134:104039. [doi: 10.1016/j.ijmedinf.2019.104039] [Medline: 31865054]
79. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw 2010;36(3):1-48 [FREE Full text] [doi: 10.18637/jss.v036.i03]
80. Fawcett L. The Middle East and COVID-19: time for collective action. Global Health 2021 Nov 22;17(1):133 [FREE Full text] [doi: 10.1186/s12992-021-00786-1] [Medline: 34809646]
81. Wang H, Yazbeck AS. Benchmarking health systems in Middle Eastern and North African countries. Health Syst Reform 2017 Jan 02;3(1):7-13. [doi: 10.1080/23288604.2016.1272983] [Medline: 31514711]
82. Barry MJ, Edgman-Levitan S. Shared decision making--pinnacle of patient-centered care. N Engl J Med 2012 Mar 01;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
83. Torous JB, Chan SR, Gipson SY, Kim JW, Nguyen TQ, Luo J, et al. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. Psychiatr Serv 2018 May 01;69(5):498-500. [doi: 10.1176/appi.ps.201700423] [Medline: 29446337]
84. Whitney RL, Ward DH, Marois MT, Schmid CH, Sim I, Kravitz RL. Patient perceptions of their own data in mHealth technology-enabled N-of-1 trials for chronic pain: qualitative study. JMIR Mhealth Uhealth 2018 Oct 11;6(10):e10291 [FREE Full text] [doi: 10.2196/10291] [Medline: 30309834]
85. Byambasuren O, Beller E, Glasziou P. Current knowledge and adoption of mobile health apps among Australian general practitioners: survey study. JMIR Mhealth Uhealth 2019 Jun 03;7(6):e13199 [FREE Full text] [doi: 10.2196/13199] [Medline: 31199343]
86. Byambasuren O, Beller E, Hoffmann T, Glasziou P. Barriers to and facilitators of the prescription of mHealth apps in Australian general practice: qualitative study. JMIR Mhealth Uhealth 2020 Jul 30;8(7):e17447 [FREE Full text] [doi: 10.2196/17447] [Medline: 32729839]

XSL•FO
RenderX

87. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010 Jun 01;152(11):726-732 [FREE Full text] [doi: 10.7326/0003-4819-152-11-201006010-00232] [Medline: 20335313]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
JBI: Joanna Briggs Institute
MENA: Middle East and North Africa
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by T de Azevedo Cardoso; submitted 13.10.23; peer-reviewed by Y Xie, J Fiedler; comments to author 14.11.23; revised version received 01.12.23; accepted 12.02.24; published 19.03.24.

Please cite as:
Tong HL, Alnasser A, Alshahrani NZ, Bawaked RA, AlAhmed R, Alsukait RF, Rakic S, Cetinkaya V, Al-Hazzaa HM, Alqahtani SA The Use of Mobile Technologies to Promote Physical Activity and Reduce Sedentary Behaviors in the Middle East and North Africa Region: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e53651
URL: https://www.jmir.org/2024/1/e53651
doi:10.2196/53651
PMID:38502160

©Huong Ly Tong, Aroub Alnasser, Najim Z Alshahrani, Rowaedh A Bawaked, Reem AlAhmed, Reem F Alsukait, Severin Rakic, Volkan Cetinkaya, Hazzaa M Al-Hazzaa, Saleh A Alqahtani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Clinical Decision Support System Used in Spinal Disorders: Scoping Review

Zheng An Toh[1], BSCN; Bjørnar Berg[2], PhD; Qin Yun Claudia Han[3], BSCN; Hwee Weng Dennis Hey[4,5], MBBS, MRCS, MMED, MCI; Minna Pikkarainen[6,7,8], PhD; Margreth Grotle[2,9], PhD; Hong-Gu He[10], MD, PhD

[1]National University Hospital, National University Health System, Singapore, Singapore
[2]Centre for Intelligent Musculoskeletal Health, Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway
[3]Department of Nursing, Tan Tock Seng Hospital, Singapore, Singapore
[4]Division of Orthopaedic Surgery, National University Hospital, National University Health System, Singapore, Singapore
[5]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[6]Department of Rehabilitation and Health Technology, Oslo Metropolitan University, Oslo, Norway
[7]Martti Ahtisaari Institute, Oulu Business School, Oulu University, Oulu, Finland
[8]Department of Product Design, Oslo Metropolitan University, Oslo, Norway
[9]Department of Research and Innovation, Division of Clinical Neuroscience, Oslo University Hospital, Oslo, Norway
[10]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Zheng An Toh, BSCN
National University Hospital
National University Health System
5 Lower Kent Ridge Road
Singapore, 119074
Singapore
Phone: 65 92289289
Email: E0191325@u.nus.edu

Abstract

Background: Spinal disorders are highly prevalent worldwide with high socioeconomic costs. This cost is associated with the demand for treatment and productivity loss, prompting the exploration of technologies to improve patient outcomes. Clinical decision support systems (CDSSs) are computerized systems that are increasingly used to facilitate safe and efficient health care. Their applications range in depth and can be found across health care specialties.

Objective: This scoping review aims to explore the use of CDSSs in patients with spinal disorders.

Methods: We used the Joanna Briggs Institute methodological guidance for this scoping review and reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) statement. Databases, including PubMed, Embase, Cochrane, CINAHL, Web of Science, Scopus, ProQuest, and PsycINFO, were searched from inception until October 11, 2022. The included studies examined the use of digitalized CDSSs in patients with spinal disorders.

Results: A total of 4 major CDSS functions were identified from 31 studies: preventing unnecessary imaging (n=8, 26%), aiding diagnosis (n=6, 19%), aiding prognosis (n=11, 35%), and recommending treatment options (n=6, 20%). Most studies used the knowledge-based system. Logistic regression was the most commonly used method, followed by decision tree algorithms. The use of CDSSs to aid in the management of spinal disorders was generally accepted over the threat to physicians' clinical decision-making autonomy.

Conclusions: Although the effectiveness was frequently evaluated by examining the agreement between the decisions made by the CDSSs and the health care providers, comparing the CDSS recommendations with actual clinical outcomes would be preferable. In addition, future studies on CDSS development should focus on system integration, considering end user's needs and preferences, and external validation and impact studies to assess effectiveness and generalizability.

Trial Registration: OSF Registries osf.io/dyz3f; https://osf.io/dyz3f

(J Med Internet Res 2024;26:e53951) doi:10.2196/53951

XSL•FO
RenderX

KEYWORDS

back pain; clinical decision support systems; CDSS; diagnosis; imaging; predictive; prognosis; spine

Introduction

Background

Spinal diseases are a group of conditions that affect the spinal column, leading to various symptoms ranging from pain to paralysis. The types of conditions may include spinal stenosis, herniated disc, scoliosis, osteoporosis, and degenerative disc disease, each with a unique etiology [1]. These conditions can be caused by various factors, such as genetic predisposition; age-related degeneration; trauma; infections; autoimmune and metabolic disorders; and lifestyle choices, including posture, exercise, and weight management [2]. Low back pain (LBP) is a significant health problem highly associated with spinal disorders [2], which affected an estimated 7.5% of the world's population in 2017, with approximately 568.4 million cases reported worldwide in 2019 [3]. It has prevailed as the leading cause of disability worldwide, contributing to 63.7 million years lived with disability as of 2019, influencing people of working age (from 20 to 65 years) and beyond [4]. In 2017, the cost of LBP topped the health care spending in the United States, estimated at US $134.5 billion [5]. Furthermore, LBP leads to wage and productivity losses, reflecting high costs to society [6-8]. Consequently, significant research efforts have been placed on spinal disorders, including technological patient management.

Presently, physicians are encouraged to deploy an evidence-based approach toward diagnosis and treatment by considering the best scientific (ie, matching symptoms and signs with relevant investigations and ensuring that the radiological features are concordant with the observed symptoms and signs) or research evidence and clinical experience while considering patients' values and preferences [9]. However, the overwhelming number of scientific publications makes it challenging for physicians to stay updated with the latest evidence. To address this issue, computer-based tools, such as clinical decision support systems (CDSSs), can be used.

CDSSs are computerized tools used in health care to provide personalized treatment recommendations, aid in clinical diagnosis, and predict patient-specific outcomes and prognoses [10]. These tools significantly enhance disease management in health care by improving diagnostic accuracy through timely information and narrowing down potential conditions [10]. It ensures that evidence-based treatment recommendations align with current medical guidelines, aiding medication management with alerts for interactions and allergies [11]. In personalized medicine, CDSSs use genetic data for tailored treatment plans [10]. They allow the optimization of health care workflows, reduces errors, and improves communication among professionals, thereby enhancing patient outcomes and efficient health care delivery [11]. The CDSSs can be broadly classified into knowledge-based and non–knowledge-based systems. Knowledge-based CDSSs use rules to match patient data with preset knowledge domains based on up-to-date, evidence-based clinical information, from which the best recommendations can be derived [11]. In contrast, non–knowledge-based systems use data-driven methods such as artificial intelligence (AI) or machine learning to make predictions or decisions. Although limited by their lack of transparency and auditing capability, non–knowledge-based systems can provide alternative perspectives and highlight potentially overlooked factors [10]. Recently, newer methods have been developed to interpret some AI findings, offering the possibility of greater acceptance of the non–knowledge-based methodology [12,13].

A systematic review and meta-analysis reported a 10% to 20% decrease in morbidity when CDSSs were used in patient care [14]. Physicians using CDSSs are more likely to order appropriate treatment or therapy and make fewer medication errors, thereby improving overall patient safety [10,15]. Despite these successes, research regarding the use of CDSSs in spinal disorders is still in its infancy, with much to be explored.

Objectives

Previous reviews have investigated the diagnostic and predictive performances of AI and machine learning [16-26]. However, no systematic or scoping review on the use of CDSSs in patients with spinal disorders has been identified. Therefore, this scoping review aimed to assess the extent of the literature in which CDSSs were implemented in clinical practice to assist health care professionals in offering personalized and meaningful care for patients with spinal disorders. The following review questions were answered: (1) Which CDSS tools can be identified in the current literature on spinal disorders? (2) What are the different purposes that the CDSS tools serve for spinal disorders? (3) How are these CDSS tools developed and assessed for effectiveness? and (4) What are the user's perceptions and experiences regarding the use of CDSS tools?

Methods

Overview

This review was conducted using the Joanna Briggs Institute (JBI) methodological guidance for scoping review and reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) statement [27,28]. The protocol for this review was registered in the Open Science Framework.

Eligibility Criteria

The following inclusion criteria were used to determine study inclusion: (1) the study examined the CDSS use in patients with spinal disorders affecting the spinal column, cord, nerves, discs, or vertebrae in the cervical, thoracic, lumbar, or sacral regions of the spine and those with back pain, neuropathic pain, numbness, abnormal sensation, or tension caused by spinal issues; (2) all types of CDSS were considered, including integrated or independent systems, with purposes including diagnosis, disease or treatment prognosis, and treatment management of spinal disorders; (3) all participants were considered, with no restrictions placed on their cultural or racial background, geographic location, sex, or clinical management

XSL•FO
RenderX

setting (acute or community); and (4) there were no restrictions placed on the study type, design, or source. The studies were excluded if they did not involve human participants, did not use a digitalized solution for ease of accessibility and use, were not applied in a clinical setting, or were reviews.

Search Strategy

Both published and unpublished studies were located through PubMed, Embase, Cochrane, CINAHL, Web of Science, Scopus, ProQuest, and PsycINFO databases from inception until October 11, 2022. A limited initial search of PubMed was conducted to identify related articles and gather relevant keywords to develop a complete search strategy. The search strategy (Multimedia Appendix 1) was formed using the main concepts, including *clinical decision support system* and *spinal disorders*, combined with Boolean operators of *AND* and *OR*. The keywords and index terms were adapted for each database, and the reference lists of the included sources were screened for additional relevant studies. No limitations were placed on the sources' language or date of publication to ensure that all relevant information on the topic was captured. In addition, sources of unpublished studies or gray literature, such as ClinicalTrials.gov, the International Standard Randomized Controlled Trial Number Register, the World Health Organization International Clinical Trials Registry Platform, and the Directory of Open Access Journals, were also searched.

Source of Evidence Selection

Potential records were collated and uploaded to EndNote 20 (Clarivate), with duplicates removed [29]. Two independent reviewers (ZAT and QYCH) screened the titles and abstracts based on the eligibility criteria. The full text of potentially relevant studies was retrieved and further assessed for eligibility by both reviewers. The studies that did not meet the inclusion criteria were recorded and reported in the scoping review. Any disagreements between the 2 reviewers at each stage of the selection process were resolved through discussion or involving an additional reviewer (BB).

Data Extraction and Synthesis

Data were extracted from the studies by 2 independent reviewers (ZAT and QYCH) using a data charting form adapted from the standardized data extraction tool of the JBI [27]. The extracted data included details about the participants, concept, context, study methods, and key findings relevant to the review questions. Iterative updates to the charting table allowed for the addition of valid unforeseen data [27]. We organized the research according to the applications examined and summarized the characteristics of each group, including the settings, participants, study designs, performance measures, and overall conclusions.

Results

Study Selection

A total of 26,828 records were identified from PubMed, Embase, Cochrane, CINAHL, Web of Science, Scopus, ProQuest, and PsycINFO databases. Of these, 73 (0.27%) full-text papers were retrieved after screening titles and abstracts and assessed against predetermined eligibility criteria (Figure 1); eventually, 31 (0.16%) studies were included for synthesis in this review, as summarized in Table 1. The studies were conducted in the United States (13/31, 42%), Australia (4/31, 13%), the Netherlands (3/31, 10%), Switzerland (2/31, 7%), Germany (3/31, 10%), Canada (1/31, 3%), Russia (1/31, 3%), Sweden (1/31, 3%), Ireland (1/31, 3%), South Korea (1/31, 3%), and the United Kingdom (1/31, 3%).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection. CDSS: clinical decision support system.

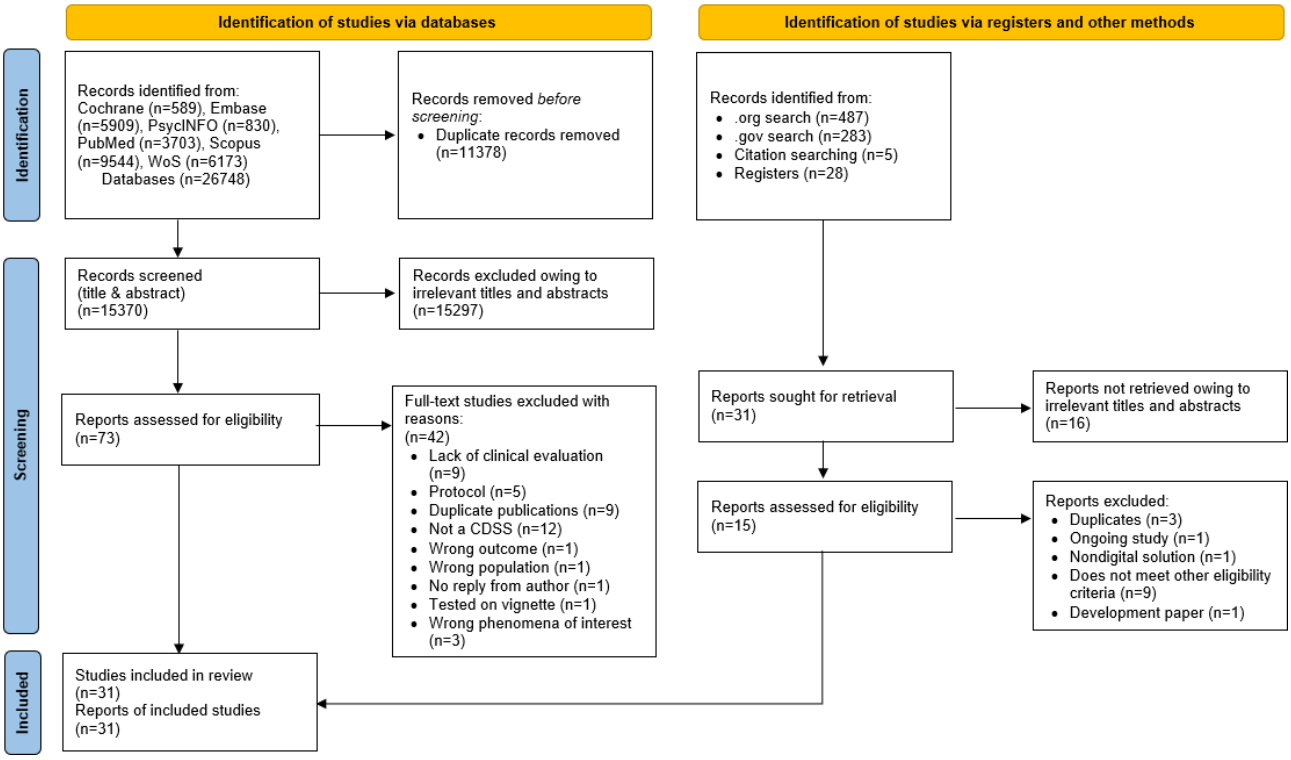

XSL•FO
RenderX

Table 1. A summary of the included CDSSs[a] presented according to their purpose and application.

CDSS name and study	Country and setting	Study design (date)	Population	Sample size, N	Female, n (%)	Age (y), mean (SD)
Preventing unnecessary imaging						
Choosing Wisely recommendation (Stanson Health), Chen et al [30], 2020	United States and single institution ambulatory clinic	CDSS testing: pre-post study (March 1, 2015, to April 30, 2017)	Patients with acute low back pain	Not reported	Not reported	Range 18 to 69
NEXUS[b] clinical decision rule (Medweb), Goergen et al [31], 2006	Australia and single institution emergency department	CDSS testing: prospective cohort study (October 2001 to September 2002) with historical controls (June 2000 to July 2001)	Patients with cervical spine trauma	Cohort: 353 and control: 403	Cohort: 156 (45) and control: 190 (47)	Cohort: 32 (23-45)[c] and control: 32 (24-49)[c]
Combined NEXUS criteria and CCSR[d] CDSS, Hynes et al [32], 2020	Ireland, and single institution emergency department	CDSS testing: prospective cohort study (March to April 2017) with historical controls (March to April 2016)	Patients with cervical spine trauma	Not reported	Not reported	Not reported
ACP[e] APS[f] guideline derived CDSS, Ip et al [33], 2014	United States, and primary care service in an integrated health system with a quaternary care hospital and outpatient network	CDSS testing: prospective cohort study (2007 to 2010) with control cohort derived from NAMCS[g]	Patients with low back pain	Cohort: 21,445 and control: 2240	Cohort: 14,950 (69.7) and control: 1283 (57.3)	Cohort: 53.0 (15.6) and control: 50.5 (15.8)
ACR[h] select tool (National Decision Support Company), Mallavarapu and Christiason [34], 2020	United States, and emergency department of a 204-bed community hospital	CDSS testing: interrupted time series from 12 months before and 10 months after modification	Patients with low back pain	Not reported	Not reported	Not reported
Choosing Wisely Canada CDSS, Min et al [35], 2017	Canada and single institution emergency department	CDSS testing: retrospective pre-post study from January 1, 2013, to May 31, 2016	Patients with acute low back pain	Not reported	Not reported	Not reported
ACR appropriateness criteria CDSS (Institute of Clinical Systems Improvement), Solberg et al [36], 2010	United States and multispecialty medical group primary care clinics	CDSS testing: retrospective pre-post study (2006 to 2007)	Patients requiring MRI[i] spine	Cohort: 148 cases and control: 151 cases	Overall: 62%	57.8
Zafar et al [37], 2019	United States and tertiary academic health system with 8 PCP[j] practices	CDSS testing: RCT[k] with varying intervention periods; baseline period (March 1, 2012, to October 4, 2012), intervention period 1 (February 6, 2013, to December 31, 2013), and intervention period 2 (January 14, 2014, to June 20, 2014, and September 4, 2014, to January 21, 2015)	Physicians ordering imaging for patients with low back pain	108 PCPs	Not reported	Not reported

XSL•FO
RenderX

CDSS name and study	Country and setting	Study design (date)	Population	Sample size, N	Female, n (%)	Age (y), mean (SD)
Diagnostic tool						
Benditz et al [38], 2019	Germany and single hospital orthopedic department	CDSS testing: cross-sectional correlational study	Patients with back pain	111	53 (47.7)	59.47 (15.81)
Benditz et al [39], 2021	Germany and single hospital orthopedic department	CDSS testing: cross-sectional study	Patients with back pain	86	40 (47)	51 (17)
Lin et al [40], 2006	United States and Europe: nationwide pain clinic in United States and clinics in Europe	CDSS development and testing: cross-sectional study	Patients with low back pain	180	Not reported	Not reported
Peiris et al [41], 2014	Australia and nationwide primary care clinics	CDSS development and testing: mixed methods study	Patients with back pain	Overall, not reported and 20 GPs[l] (recruited for qualitative portion)	7 (35)	8 (40) patients aged <50 y, 8 (40) patients aged 50 to 59 y, and 3 (15) patients aged >60 y[m]
Kim et al [42], 2022	South Korea and single institution hospital	CDSS testing: cross-sectional study	Patients with postural spinal deformity	140	81 (57.86)	24.94 (17.36)
The Vertebral Compression Fracture tool, Wang et al [43], 2011	United States	CDSS development: cross-sectional study (not reported)	Patients with vertebral compression fractures	128	Not reported	Not reported
Prognostic tool						
The Seattle Spine Score (Virginia Mason Medical Center), Buchlak et al [44], 2017	United States and single high-volume hospital	CDSS development and testing: retrospective predictive modeling study	Patients with spinal deformity and those who had undergone surgery	136	100 (73.5)	63.2 (11.2)
Simple Brace Predictor (University of Alberta Edmonton), Chalmers et al [45], 2015	United States, and single institution scoliosis clinic	CDSS development and testing: retrospective chart review	Patients with adolescent idiopathic scoliosis	Training data set: 62 and test data set: 28	75 (83.3)	13.5 (1.7)
The Dialogue Support (Swedish Society of Spinal Surgeons), Fritzell et al [46], 2022	Sweden and a nationwide study (data from Swespine)	CDSS development and testing: retrospective chart review	Patients with lumbar disc herniation, lumbar spinal stenosis, degenerative disc disease, and cervical radiculopathy and those who underwent surgery.	87,494	Not reported	Not reported
Subgroups for Targeted Treatment (STarT) Back screening tool (Keele University), Hill et al [47], 2008	United Kingdom and 8 primary care general practices	CDSS development and testing	Patients with nonspecific back pain	CDSS development sample: 131 and validation sample: 500	CDSS development sample: 77 (60) and validation sample: 293 (59)	CDSS development sample: 44 (10.0) and validation sample: 45 (9.7)
SCOAP-CERTAIN[n] tool (SCOAP-CERTAIN), Khor et al [48], 2018	United States and 15 Washington state hospitals	CDSS development and testing: prospective registry	Patients who have undergone lumbar spinal surgery	1583	944 (59.6)	61.3 (12.5)

XSL•FO
RenderX

CDSS name and study	Country and setting	Study design (date)	Population	Sample size, N	Female, n (%)	Age (y), mean (SD)
SpineSage (University of Washington), Lee et al [49], 2014	United States, 2 academic institutions	CDSS development and testing retrospective chart review	Patients who had undergone spina surgery	1476	634 (43)	49.4 (20.0)
Cleveland Lumbar Spine Surgery risk calculator (Cleveland Clinic), Lubelski et al [50], 2021	United States and single tertiary care institution	CDSS development: retrospective chart review	Patients who had undergone lumbar spine surgery	2996	1386 (46)	58.3 (15.0)
Dartmouth Back Treatment Outcomes Calculator (Dartmouth College), Moulton et al [51], 2018	United States and multidisciplinary spine centers and web-based consumer reports subscribers	CDSS testing; cross-sectional study	Web-based subscribers of consumer reports and patients presenting with IDH[o], SpS[p], or DS[q]	1256 consumer participants and 68 patient participants	Consumer: 336 (27) and patient: 30 (44)	Consumer: 67 (9.3) and patient: 59 (16)
Schulthess Klinik Prognostic tool (Schulthess Klinik), Müller et al [52], 2021	Switzerland and single institution hospital	CDSS development: prospective cohort study	Patients with thoracic, lumbar, or cervical spinal degenerative disease	8374	4471 (53.4)	63.9 (14.3)
SCOAP-CERTAIN tool (SCOAP-CERTAIN), Quddusi et al [53], 2020	Netherlands and Dutch specialist short-stay spine center	External validation of prediction model	Patients with transforaminal lumbar interbody fusion or posterior lumbar interbody fusion	100	49 (49)	50.4 (11.4)
FUSE-ML (Machine Intelligence in Clinical Neuroscience & MICrosurgical Neuroanatomy laboratory), Staartjes et al [54], 2022	Multinational and multicenter (Switzerland, Netherlands, Italy, South Korea, France, and Austria)	CDSS development and testing	Patients who had undergone lumbar spinal fusion for degenerative disease	CDSS development sample: 817 and validation sample: 298	CDSS development sample: 468 (57.3) and validation sample: 192 (64.4)	CDSS development sample: 61.19 (12.3) and validation sample: 59.73 (12.6)
Treatment recommendation						
Benditz et al [38], 2019	Germany and single hospital orthopedic department	CDSS testing: cross-sectional correlational study	Patients with back pain	111	53 (47.7)	59.47 (15.81)
Byvaltsev and Kalinin [55], 2021	Russia and single hospital	CDSS testing: prospective cohort study with retrospective controls	Patients who had undergone lumbar spinal surgery	59 prospective cohort and 196 retrospective controls	Prospective cohort: 21 (35.6) and retrospective control: 59 (30.1)	Not reported
Downie et al [56], 2020	Australia and community pharmacy setting	CDSS development: mixed methods cross-sectional study	Patients with lower back pain	5 practicing community pharmacists	Not reported	Not reported
Back-UP (Horizon 2020), Jansen-Kosterink et al [57], 2021	Netherlands and community setting	CDSS testing: mixed methods study	Patients with chronic lower back pain	98 PCPs	47 (48)	48 (12.2)
Subaxial Injury Classification (SLIC) CDSS (Kubben), Kubben et al [58], 2011	Netherlands and not specified	CDSS development: Descriptive study	Patients with subaxial cervical spinal injury	Not reported	Not reported	Not reported

XSL•FO
RenderX

CDSS name and study	Country and setting	Study design (date)	Population	Sample size, N	Female, n (%)	Age (y), mean (SD)
Peiris et al [41], 2014	Australia and nationwide primary care clinics	CDSS development and testing: mixed methods study	Patients with back pain	Overall, not reported; 20 GPs (recruited for qualitative portion)	7 (35)	8 (40) patients aged <50 y, 8 (40) patients aged 50 to 59 y, and 3 (15) patients aged >60 y

[a]CDSS: clinical decision support system.
[b]NEXUS: National Emergency X-Radiography Utilization Study Group.
[c]Median (IQR).
[d]CCSR: Canadian Cervical Spine Rule.
[e]ACP: American College of Physicians.
[f]APS: American Pain Society.
[g]NAMCS: National Ambulatory Medical Care Survey.
[h]ACR: American College of Radiology.
[i]MRI: magnetic resonance imaging.
[j]PCP: primary care provider.
[k]RCT: randomized controlled trial.
[l]GP: general practitioner.
[m]One response was missing for age value.
[n]SCOAP-CERTAIN: Surgical Care and Outcomes Assessment Programme-Comparative Effectiveness Translational Network.
[o]IDH: intervertebral disc herniations.
[p]SpS: spinal stenosis.
[q]DS: degenerative spondylolisthesis.

Study Characteristics

The use of CDSSs in spinal disorders is summarized into 4 major categories based on their primary purpose and application, as presented in Table 1: of 31 CDSSs, 8 (26%) were for the prevention of unnecessary imaging, 6 (19%) were for diagnostic applications, 11 (35%) were for prognostic applications, and 6 (19%) were for treatment recommendations. Only 5 (16%) of the 31 studies investigated user perceptions and experiences concerning the use of CDSSs [31,41,51,56,57].

CDSSs for Preventing Unnecessary Imaging

Of the 31 CDSS studies reviewed, the implementation and results of the 8 (26%) CDSSs used to determine if radiologic imaging was necessary for patients with lower back pathologies [30,33-35,37], patients with cervical spine trauma [31,32], and patients in general [36] are presented in Table 2. The CDSSs were mainly embedded into the electronic health record system or the computerized physician order entry, apart from the guidelines proposed by Goergen et al [31], which used a physical report card and independent software. These CDSSs were often implemented in health care settings, such as the emergency departments, where patients with back pain or cervical spine trauma were first seen by the physicians. They functioned as alerts to remind physicians to consider whether spinal imaging is necessary and can take different forms, including hard-stop, soft-stop, and passive alerts. Hard-stop alerts aim to prevent the physician from proceeding with imaging orders that do not meet the guideline requirements. In contrast, soft-stop alerts may allow the physician to continue with the ordered imaging but require them to provide a reason. Passive alerts only require acknowledgment and do not require further user interactions. Although some studies did not specify the type of alert used, the information provided in the studies allowed for inference that all studies used a soft-stop alert, excluding 1 study that used a passive alert function [37].

XSL•FO
RenderX

Table 2. Type, features, and results of the CDSSs[a] for preventing unnecessary imaging.

CDSS name and study	CDSS type	Features of the CDSS	Results

Choosing Wisely recommendation (Stanson Health), Chen et al [30], 2020	Knowledge based	• Pop-up and soft-stop alert:
		  • Provides best practice advise when a CT[b] scan, x-ray, or MRI[c] of lumbar spine is ordered for a female patient aged 18 to 49 y or a male patient aged 18 to 69 y
		  • Alert suppressed for comorbidities; complicated back pain owing to trauma, cauda equina syndrome, spondylitis, disc herniation, history of back surgery; and disciplines, including neurosurgery, orthopedics, trauma surgery, anesthesiology, rheumatology, physical medicine and rehabilitation, oncology, and neurology
			• Post-CDSS implementation:
			  • Overall imaging rate decreased from 5.8% to 5.2% (9.6% decrease; *P*=.02)
			  • MRI imaging rate decreased from 1.8% to 1.5% (16.7% decrease; *P*<.01)
			  • No statistically significant differences in the rates of x-ray (*P*=.39) or CT (*P*=.88) orders
			• Rationale for override
			  • A total of 64% used preset options: duration >6 weeks (37%), focal neurological deficit (14%), history of trauma (10%), previous spine surgery (1%), unexplained weight loss or insidious onset (1%), and unexplained fever or recent infection (1%)
			  • Free-text rationale (n=125, 36%); 56% were inappropriate

NEXUS[d] clinical decision rule (Medweb), Goergen et al [31], 2006	Knowledge based	• Guideline questionnaire based on NEXUS criteria and passive alert:
		  • Helps physicians to determine which patients to image and which imaging method (eg, plain radiography or helical CT) to use first
			• Compliance with CDSS and imaging guidelines:
			  • 40% (141/353) of patients were managed using the CDSS
			  • Of the 51 patients for whom the NEXUS guideline did not recommend imaging, 86% (43/51) did not receive any imaging
			• Cervical spine imaging ordered: CDSS intervention group: 63.8% and control group: 78.5% (*P*=.01)
			• Cervical spine imaging ordered (non-CDSS intervention):
			  • Non-CDSS intervention group[e]: 72.6% and control group: 78.5% (*P*=.11)

Combined NEXUS criteria and CCSR[f] CDSS, Hynes et al [32], 2020	Knowledge based	• Guideline questionnaire based on NEXUS and CCSR criteria and soft-stop alert
		  • Integrated in electronic imaging ordering system
		  • Helps physicians follow evidence-based guidelines when ordering cervical spine radiographs for patients who have experienced trauma
		  • Physicians asked to check boxes indicating which criteria the patient meets when ordering imaging
			• Cervical spine radiograph orders:
			  • Preintervention: 182
			  • Postintervention: 126 (*P*<.001)
			• Proportion of requests meeting NEXUS or CCSR criteria:
			  • Preintervention: 76.7%
			  • Postintervention: 99.2% (*P*<.001)

ACP[g] APS[h] guideline derived CDSS, Ip et al [33], 2014	Knowledge based	• Guideline questionnaire based on ACP or APS criteria and soft-stop alert
		  • Integrated into the CPOE[i] system
		  • Provides real-time decision support to physicians for imaging orders based on the patient's clinical history
			• Lumbar spine MRI orders:
			  • Preintervention: 5.3% (443/8437)
			  • Postintervention: 3.7% (477/13,008; *P*<.001)
			• Outpatient MRI orders 30 d after:
			  • Preintervention: 2.2%
			  • Postintervention: 2.7% (*P*=.03)
			• LBP[j]-related visits that resulted in an MRI within 30 d of the index visit, accounting for imaging that was ordered by specialists
			  • Preintervention: 8.9%
			  • Postintervention: 7.8% (*P*=.002)

XSL•FO
RenderX

CDSS name and study	CDSS type	Features of the CDSS	Results

ACR[k] select tool (National Decision Support Company), Mallavarapu and Christiason [34], 2020	Knowledge based	• Guideline questionnaire based on ACR criteria and soft-stop alert
		  • Integrated into the electronic medical health system
		  • The *free text* field, which allowed providers to bypass the ACR select tool within EHRs[l], has been removed to increase provider adherence to the tool
			• Preintervention: 13 scans/mo and postintervention: 11.6 scans/mo (*P*=.54)

Choosing wisely Canada CDSS, Min et al [35], 2017	Knowledge based	• Guideline questionnaire based on recommendations from the Canadian Association of Emergency Physicians, the College of Family Physicians of Canada, Occupational Medicine Specialists of Canada, the Canadian Association of Radiologists, and the Canadian Spine Society, and soft-stop alert
		  • Integrated with the CPOE
		  • Physicians must select a suspected diagnosis when ordering an imaging test for LBP
		  • If a physician selects *other* as the suspected diagnosis, they will need to provide an explanation for ordering the imaging test outside the established appropriateness criteria
			• Proportion of LBP patients with imaging order fell significantly compared with preimplementation baseline after CDSS implementation
			  • Median: 22% decreased to 17%
			  • Mean: 23% decreased to 18%; (*P*<.001)
			• Imaging ordering patterns
			  • A total of 60% (26/43) of the physicians reduced their ordering of imaging tests.
			• Imaging orders placed 1 to 30 d after LBP presentation
			  • Preintervention: 2.3% and postintervention: 2.2% (*P*=.97)
			• ED[m] revisit: preintervention 8.2% and postintervention 6.9% (*P*=.17)

ACR appropriateness criteria CDSS (Institute of Clinical Systems Improvement), Solberg et al [36], 2010	Knowledge based	• Guideline questionnaire based on ACR criteria and soft-stop alert
		  • Integrated within the EHR system and requires physicians to enter a reason for every order placed.
		  • No safeguards were in place to prevent orders from being placed even if they did not meet certain criteria. Physicians received little feedback on the outcomes of their orders.
			• Volume of spine MRI ordered decreased by 20%
			• Impact of CDSS on patient's health after spine MRI increased from 14% to 30% (*P*=.18)

Zafar et al [37], 2019	Knowledge based	• Guideline based on ACP and APS criteria and soft-stop or passive alert
		  • Embedded in CPOE, the CDS[n] algorithm screen the lumbar spine MRI orders for adherence to the guideline
		  • Intervention groups: periodic CDSS report cards vs real-time CDSS alerts vs both
			• Likelihood of placing lumbar spine MRI orders at the time of LBP presentation when compared with baseline
			  • CDSS report cards: 38% lower likelihood
			  • Real-time CDSS alerts: not associated with any change (*P*=.59)

[a]CDSS: clinical decision support system.
[b]CT: computed tomography.
[c]MRI: magnetic resonance imaging.
[d]NEXUS: National Emergency X-Radiography Utilization Study.
[e]Imaging guidelines given in a form of pocket card and posters, with small group teaching sessions.
[f]CCSR: Canadian Cervical Spine Rule.
[g]ACP: American College of Physicians.
[h]APS: American Pain Society.
[i]CPOE: Computerized provider order entry.
[j]LBP: low back pain.
[k]ACR: American College of Radiology.
[l]EHR: electronic health record.
[m]ED: emergency department.

XSL•FO
RenderX

[n]CDS: clinical decision support.

The included studies reported ≥1 of the following outcomes: change in the frequency of imaging order, change in the frequency of imaging order 1 to 30 days after LBP presentation, and adherence to order guidelines. All studies reported a decrease in imaging ordered on the initial presentation of LBP after the implementation of a CDSS, although the decrease was not clinically relevant in some studies [30,34]. Ip et al [33] reported a notable increase (22.7%; *P*=.03) from 2.2% (188/8437) to 2.7% (352/13,008) in the lumbar spine–magnetic resonance imaging (LS-MRI) ordered by outpatient specialists within 30 days of the patient's primary care visit. This increase may be explained by the fact that the CDSS intervention was implemented in the primary care setting but not in the outpatient setting. However, when considering the total percentage of the LS-MRI orders for LBP visits before and after CDSS implementation, there was a statistically significant decline (12%; *P*=.002) from 8.9% (753/8437) to 7.8% (1009/13,008) in imaging orders after adjusting for outpatient specialist orders.

Zafar et al [37] compared the outcomes of different CDSS deliveries for LS-MRI orders [37]. The CDSS report cards that were generated every 4 to 6 months led to fewer magnetic resonance imaging (MRI) orders (50/1739, 2.9%) for cases compared with immediate CDSS alerts (94/2021, 4.7%).

Furthermore, CDSSs, generally, were reported to improve adherence to imaging guidelines. For example, Hynes et al [32] reported a 99.2% adherence rate to the established imaging guidelines after CDSS implementation (125 indicated imaging out of 126 total imaging), an increase of 22.5% (76.7 to 99.2%) from preimplementation [32]. Similarly, Solberg et al [36] discovered a reduction of 20% in the volume of MRI spine orders and an increase in the appropriateness of MRI spine orders based on health impacts [36].

Diagnostic CDSS

Of the 31 studies reviewed, 6 (19%) explored diagnostic CDSSs (Table 3) and 3 (10%) examined the accuracy of CDSS compared with expert or *gold standard* diagnoses [38,39,42]. A moderate agreement was found between the CDSS and expert diagnoses for back pain (Cramer V=0.424) [38]. A higher agreement of 67% (58/86) of the cases between the CDSS and expert diagnosis (Cramer V=0.711) was found for patients with spinal disorders in general [39]. Another study by Lin et al [40] found that a CDSS performed a diagnosis comparable to that of experts and correctly recommended 75.82% of diagnoses based on gold-standard criteria [40]. In a recent study by Kim et al [42], the CDSS diagnosis demonstrated a 94% agreement with the gold-standard radiographic assessment for scoliosis, with higher agreement reported for patients within the normal and mild postural deformation range [42].

XSL•FO
RenderX

Table 3. Type, features, and results of the CDSSs[a] for diagnostic support.

CDSS name and study	CDSS type	Features of the CDSS	Results
Benditz et al [38], 2019	Knowledge based	• Questionnaire-based CDSS • A computerized tool with disease-specific algorithms cascading the next best questions leading to the most probable diagnosis and actions	• Diagnosis of CDSS compared with spinal surgeons: Cramer V=0.424[b]; *P*<.001

Benditz et al [39], 2021	Knowledge based	• Decision tree algorithm and app-based questionnaire
		  • Questionnaire will ask the patient to identify the location of their pain and present dichotomous questions to suggest a diagnosis.
		  • If the patient scores >65% on these questions, the diagnosis is confirmed.
		  • If not, the questionnaire will ask additional questions about the second most likely diagnosis.
		  • If the patient still scores <65% after these questions, they are advised to consult with a physician.
			• Diagnosis of CDSS compared with spinal surgeons: Cramer V=0.711[b]; *P*<.001
			• Concordance: 67.4%
			  • A total of 15.1% overestimated
			  • A total of 7% underestimated

Lin et al [40], 2006	Knowledge based	• Knowledge from 2 highly experienced physical therapists and web-based questionnaire
		  • Patients can start a self-diagnosis session with or without clinician's assistance.
		  • Questions regarding specific pain symptom or assessment will be presented through typically 13 to 15 web pages, depending on the number of follow-up questions triggered.
		  • After completion of the questions, a diagnosis that may consist of ≥1 parts, based on patient information, clinical evidence provided by the user, and system's rule activation, will be generated and the clinician can override any parts of the diagnosis.
		  • The explanatory panel can be activated by the clinician to review the system's reasoning process. The clinician can add, remove, or modify an existing rule in reference to its observation, decision outcome, or certainty level.
			• Not reported

Peiris et al [41], 2014	Knowledge based	• Recommendations from 15 guidelines for back pain management • After excluding serious pathology, the CDSS will continue to assess for the most probable diagnosis and treatment through a series of questions. A personalized information sheet will be printed.	• Not reported

XSL•FO
RenderX

CDSS name and study	CDSS type	Features of the CDSS	Results
Kim et al [42], 2022	Nonknowledge based	• Computer vision–based posture analysis system • The CDSS uses a Kinect sensor and specialized software to analyze a person's skeletal structure and gait. • The CDSS captures an image and records a moving video of the participant. The software then identifies the participant's joints and uses them to determine the skeletal structure and gait. • Furthermore, it uses a set of algorithms to judge the probability of scoliosis by analyzing the curvature of the participant's central coronal axis, which is determined by a line connecting the eyes, shoulders, and pelvis. The CDSS classifies scoliosis as normal (≤3 mm curvature), 20% scoliosis (3 mm to 10 mm curvature), or 50% scoliosis (>10 mm curvature).	• Postural deformations: assessed with 94% accuracy (comparable with radiographic assessments) • Normal or mild scoliosis: conformity assessment accuracy of 98.57% • CDSS's diagnostic accuracy for scoliosis was 0.94, with the most influential factors being spinal curvature and pelvis height, which accounted for 79.97% and 19.86% of the variance in the data, respectively
Vertebral Compression Fracture tool, Wang et al [43], 2011	Knowledge based	• Logistic regression and web-based checklist • Uses checklists for dichotomous and nondichotomous discrete variables based on MRI[c] features to generate a probability of malignancy and a text report. The model captures inputs from these variables to make its assessment.	• Not reported

[a]CDSS: clinical decision support system.

[b]Interpretation of Cramer V effect size measurement of association: effect size ≤0.2: weak association, <0.2 effect size ≤6: moderate association, and effect size >0.6: strong association.

[c]MRI: magnetic resonance imaging.

Prognostic CDSS

Of the 31 CDSS studies reviewed, 11 (35%) prognostic CDSS studies (Table 4) were knowledge based [44,45,47-54,59], with regression-based predictive algorithms. White-box models were used across all studies; most CDSSs were presented as web-based calculators, whereas others were presented as independent software. Prognostic CDSSs are used for various purposes, most commonly to predict the likelihood of complications, functional outcomes, pain, and quality of life following spinal surgery (8/11, 73%). Other purposes included predicting the outcome of brace treatment for adolescent idiopathic scoliosis (1/11, 9%), the risk of back pain chronicity (1/11, 9%), and treatment outcomes between surgical and nonsurgical options for spinal disorders (1/11, 9%). Regarding rigor, external validation was only available for 3 (27%) CDSS models (FUSE-ML, Surgical Care and Outcomes Assessment Programme-Comparative Effectiveness Translational Network Tool, and STarTBack), and an impact study was only performed for the StarTBack model.

A total of 2 key aspects, namely discrimination and calibration, are often measured to evaluate the performance of a model. Discrimination can be assessed using various measures such as area under the receiver operating characteristics, accuracy, sensitivity, positive predictive values, negative predictive values, R^2 measure or value, or any specific statistic measure, such as Nagelkerke, c-index, mean absolute error, and root mean square error. In contrast, calibration can be evaluated using techniques such as calibration plot, calibration intercept and slope, and the Hosmer-Lemeshow chi-square statistic.

The impact study was the only study that conducted a clinical impact testing follow-up, as reported by Foster et al [59]. This study developed an innovative web-based calculator that assesses patients' risk of developing chronic LBP and offers tailored treatment options for each risk stratum. Results from the impact study revealed small but significant improvements (P=.03) in Roland-Morris disability scores, with a mean difference of 0.71 (95% CI 0.06-1.36) compared with usual care after 6 months of implementation. Furthermore, the group with a higher risk of developing chronic LBP experienced a large and clinically significant improvement. Work absence was also reduced by 50% (4 days instead of 8 days; P=.03), and there was a 30% decrease in prescriptions for sickness certificates (45/368, 12.2% vs 40/554, 7.2% cases; P=.03).

XSL•FO
RenderX

Table 4. The predicted outcome, input variables, and internal and external validation of prognostic CDSSs[a].

CDSS name and study	Outcome	Input variables	Internal validation		External validation	
			Discrimination	Calibration	Discrimination	Calibration
Seattle Spine Score (Virginia Mason Medical Center), Buchlak et al [44], 2017	Percentage of likelihood of complications occurring within 30 d	Age, BMI, gender, smoking status, anemia, diabetes, and hypertension	• AUC[b]: 0.71 • Accuracy: 75%	• HLT[c]: 3.692; *P*=.88	Not reported	Not reported
Simple brace predictor (University of Alberta Edmonton), Chalmers et al [45], 2015	Scoliosis progression	In-brace correction and scoliometer measurements	• Accuracy: 75%	Not reported	Not reported	Not reported
The Dialogue support (Swedish Society of Spinal Surgeons), Fritzell et al [46], 2022	GA[d] pain and satisfaction	Diagnosis group, operated levels, clinical department type, age, gender, employment, disability or retirement pension, health profile, smoking history, previous spinal surgery, quality of life, comorbidity, back-specific information, walking distance, duration and severity of preoperative pain in legs and back, and ODI[e]	• Pain AUROC[f]: 0.67-0.68 • Satisfaction AUROC: 0.60-0.67	• Calibration plots: high degree of concordance	Not reported	Not reported
Subgroups for Targeted Treatment (STarT) Back screening tool (Keele University), Hill et al [47], 2008	Risk of chronicity	Referred leg pain, comorbid pain, disability, bothersomeness, catastrophizing, fear, anxiety, and depression	• AUROC: 0.74-0.92	Not reported	• Sensitivity: 80.1 • Specificity: 65.4 • Positive likelihood ratios: 2.32 • Negative likelihood ratios: 0.30	Not reported

XSL•FO
RenderX

CDSS name and study	Outcome	Input variables	Internal validation		External validation	
			Discrimination	Calibration	Discrimination	Calibration

SCOAP-CERTAIN[g] tool (SCOAP-CERTAIN), Khor et al [48], 2018, external validation, Quddusi et al [53], 2020	Functional outcome, back pain, and leg pain	Age, gender, insurance, race or ethnicity, ASA[h] score, smoking status, prior surgery, spondylolisthesis, disc herniation, postlaminectomy, failed back syndrome, stenosis, pseudarthrosis, radiculopathy, prescription opiate use, asthma, baseline ODI and NRS[i] score	• AUROC
			• ODI: 0.66
			• Back pain: 0.79
			• Leg pain: 0.69
				• Calibration plots	• ODI AUROC: 0.71
					• Sensitivity: 0.64
					• Specificity: 0.65
					• Accuracy: 0.65
					• PPV[j]: 0.84
					• NPV[k]: 0.4
					• F1-score: 0.49
					• NRS back pain
					  • AUROC: 0.72
					  • Sensitivity: 0.81
					  • Specificity: 0.48
					  • Accuracy: 0.73
					  • PPV: 0.84
					  • NPV: 0.42
					  • F1-score: 0.44
					• NRS-Leg pain
					  • AUROC: 0.83
					  • Sensitivity: 1.00
					  • Specificity: 0.38
					  • Accuracy: 0.85
					  • PPV: 0.84
					  • NPV: 1.00
					• F1-score: 0.54
						• ODI calibration intercept: 1.08
						• Calibration slope: 0.95
						• HLT: *P*=.002
						• Brier score: 0.22
						• NRS back pain
						  • Calibration intercept: 1.02
						  • Calibration slope: 0.74
						  • Brier score: 0.19
						• NRS-Leg pain
						  • Calibration intercept: 1.08
						  • Calibration slope: 0.95
						  • Brier score: 0.12

SpineSage (University of Washington), Lee et al [49], 2014	Occurrence of medical complications after spinal surgery	Age, gender, smoking status, alcohol use, diabetes, BMI, insurance status, surgical approach, revision surgery, region, diagnosis, surgical invasiveness, and medical comorbidity	• Any medical complications:
			  • AUROC: 0.76
			• Any major medical complications:
			  • AUROC: 0.81
				Not reported	Not reported	Not reported

Cleveland Lumbar Spine Surgery risk calculator (Cleveland Clinic), Lubelski et al [50], 2021	Postoperative ED[l] visit or readmission, and quality of life		• 30-d visits to the ED c-index[r]: 0.63 • 30-d readmission c-index: 0.66 • 90-d reoperation related to infection c-index: 0.73 • 1-y postoperative EQ-5D outcome c-index: 0.84	• Calibration plot	Not reported	Not reported

XSL•FO
RenderX

CDSS name and study	Outcome	Input variables	Internal validation		External validation	
			Discrimination	Calibration	Discrimination	Calibration
		Race, marital status, symptom duration, BMI, CCI[m], foraminal stenosis, disc herniation, spondylolisthesis, radiculopathy, procedures (eg, ALIF[n], PLIF[o], and TLIF[p], posterolateral lumbar fusion, and decompression), number of operated levels, preoperative Pain and Disability Questionnaire score, and EQ-5D[q]				
Dartmouth back treatment outcomes calculator (Dartmouth College), Moulton et al [51], 2018	Physical function, pain, sleep, sex life, and satisfaction with symptoms	Age, gender, height, weight, bothersomeness, back and leg pain, numbness, leg weakness, leg pain while sitting, activity level, employment status, smoking status, duration of sciatica worsening, work compensation, hypertension, depression, education level, physical therapy, sleep, and sex life	Not reported	Not reported	Not reported	Not reported
Schulthess Klinik Prognostic Tool (Schulthess Klinik), Müller et al [52], 2022	Back and leg pain, COMI[s], impairment, symptom-specific well-being, quality of life, social disability, and work disability	Preoperative axial and peripheral pain, catastrophizing, fear-avoidance beliefs, comorbidity, age, BMI, nationality, previous spinal surgery, type and spinal level of intervention, number of affected levels, and surgeon seniority	Not reported	Not reported	Not reported	Not reported

XSL•FO
RenderX

CDSS name and study	Outcome	Input variables	Internal validation		External validation	
			Discrimination	Calibration	Discrimination	Calibration
FUSE-ML (Machine Intelligence in Clinical Neuroscience & MICrosurgical Neuroanatomy laboratory), Staartjes et al [54], 2022	Functional outcome and back and leg pain	Age, gender, surgical indication, index level, height, weight, BMI, smoking status, ASA score, preoperative opioid use, bronchial asthma, prior thoracolumbar spinal surgery, race or ethnicity, surgical approach, pedicle screw insertion, ODI or COMI, and leg and back NRS	• ODI or COMI • AUROC: 0.75 • Accuracy: 0.70 • Sensitivity: 0.70 • Specificity: 0.70 • PPV: 0.88 • NPV: 0.43 • Back pain • AUROC: 0.71 • Accuracy: 0.68 • Sensitivity: 0.68 • Specificity: 0.63 • PPV: 0.91 • NPV: 0.26 • Leg pain • AUROC: 0.72 • Accuracy: 0.74 • Sensitivity: 0.77 • Specificity: 0.58 • PPV: 0.90 • NPV: 0.34	• ODI or COMI • Calibration intercept: 0.00 • Calibration slope: 0.89 • Back pain • Calibration intercept: 0.00 • Calibration slope: 0.86 • Leg pain • Calibration intercept: 0.00 • Calibration slope: 0.84	• ODI or COMI • AUROC: 0.67 • Accuracy: 0.61 • Sensitivity: 0.59 • Specificity: 0.66 • PPV: 0.81 • NPV: 0.39 • Back pain • AUROC: 0.72 • Accuracy: 0.70 • Sensitivity: 0.72 • Specificity: 0.64 • PPV: 0.90 • NPV: 0.34 • Leg pain • AUROC: 0.64 • Accuracy: 0.71 • Sensitivity: 0.76 • Specificity: 0.42 • PPV: 0.88 • NPV: 0.23	• ODI or COMI • Calibration intercept: −0.07 • Calibration slope: 0.63 • Back pain • Calibration intercept: −0.38 • Calibration slope: 1.10 • Leg pain • Calibration intercept: 0.14 • Calibration slope: 0.49

[a]CDSS: clinical decision support system.
[b]AUC: area under the curve.
[c]HLT: Hosmer-Lemeshow Test.
[d]GA: general anesthesia.
[e]ODI: Oswestry Disability Index.
[f]AUROC: area under the receiver operating characteristics.
[g]SCOAP-CERTAIN: Surgical Care and Outcomes Assessment Programme-Comparative Effectiveness Translational Network.
[h]ASA: American Society of Anesthesiologists.
[i]NRS: Numeric Rating Scale.
[j]PPV: positive predictive value.
[k]NPV: negative predictive value.
[l]ED: emergency department.
[m]CCI: Charlson Comorbidity Index.
[n]ALIF: anterior lumbar interbody fusion.
[o]PLIF: posterior lumbar interbody fusion.
[p]TLIF: transforaminal lumbar interbody fusion.
[q]EQ-5D: EuroQOL-5D.
[r]C-index: concordance.
[s]COMI: Core Outcome Measures Index.

XSL•FO
RenderX

CDSS for Treatment Recommendation

Of the 31 CDSS studies reviewed, studies exploring the use of CDSS for treatment recommendations for spinal disorders were divided into 2 categories based on their focus: 2 (6%) CDSSs for recommendations for spinal surgery [55,58] and 4 (13%) CDSSs for treatment of LBP [38,41,56,57] (Table 5). All CDSSs were knowledge based, except for 1, which was structured on medical ontology and fuzzy logic principles [55]. The system inputs required to generate personalized treatment recommendations include symptoms, clinical findings, and instrumental findings.

Byvaltsev and Kalinin [55] studied using a CDSS to recommend total disc replacement, minimally invasive rigid stabilization, and open rigid stabilization [55]. The researchers observed lower pain levels and improved functional status 6 months after surgery among those who received treatment recommendations using the CDSS [55]. Those who underwent minimally invasive rigid stabilization had better outcomes 3 months after surgery [55]. In the work of Benditz et al [38], although 49.6% (55/111 cases) of the treatment recommendations made by the CDSS were consistent with those of spinal surgeons, 36% (40/111) were overestimated and 14.4% (16/111) were underestimated [38]. In contrast, a study by Downie et al [56] revealed that CDSS recommendations were highly concordant with those made by pharmacists for cases involving self-care (18/20, 90%), medications (25/25, 100%), and referral advice (22/25, 88% [56]).

XSL•FO
RenderX

Table 5. Type, features, and outcomes measured for the CDSS[a] for treatment recommendation.

Study and CDSS name	CDSS type	Features of the CDSS	Outcomes measured	Results

Benditz et al [38], 2019	Knowledge based	• Questionnaire-based CDSS
		  • A computerized tool with disease-specific algorithms cascading the next best questions leading to the most probable diagnosis and actions
			• Association between the diagnoses and treatment recommendation of the tool and the physician's diagnosis	• Significant correlation with small-to-medium effect between the DSS[b] and the medical recommendation.
				  • Cramer V=0.293; *P*=.02
				• Concordance: 49.6%
				  • Overestimated: 36%
				  • Underestimated: 14.4%

Byvaltsev and Kalinin [55], 2021	Nonknowledge based	• Semantic network structured based on the medical ontology and fuzzy logic principles
		  • Computer-assisted electronic checklist and recommendations, which uses preoperative instrumental data on lumbar segments of patients with degenerative diseases
			• Pain using visual analog scale
			  • Lower limbs
			  • Lumbar spine
			• ODI[c]
				• For patients who underwent total disc replacement, pain syndrome level and functional status were comparable before surgery, on discharge and 3 mo after surgery (*P*>.05).
				• A total of 6 mo after the surgery, there was a decrease in pain intensity in the lower limbs (*P*=.02) and lumbar spine (*P*=.03) and an increase in functional status by ODI (*P*=.02) in the CDSS group.
				• In the CDSS patients group who underwent minimally invasive rigid stabilization, there was a decrease in pain intensity in the lower limbs (*P*=.01 for both 3 mo and 6 mo after surgery) and in the lumbar spine (*P*=.03 and *P*=.02 for 3 mo and 6 mo after surgery, respectively) and an increase in functional status by ODI (*P*=.01 and *P*=.03 for 3 mo and 6 mo after surgery, respectively).
				• For patients who underwent open rigid stabilization, pain syndrome level and functional status were comparable (*P*>.05) before surgery, on discharge and 3 mo after the surgery.
				• A total of 6 mo after surgery, there was a decrease in pain intensity in the lower limbs (*P*=.04) and lumbar spine (*P*=.03) and an increase in functional status by ODI (*P*=.01) in the group using CDSS.

Downie et al [56], 2020	Knowledge based	• Decision tree app–based CDSS
		• It consists of a knowledge base, reasoning engine, and interface. An advice report will be generated after the history and screening inputs. The pharmacist may add any key message or modify the advice.
			• Qualitative-based CDSS:
			  • Ease of use, consistency (of visual language or interaction model), system visibility, navigation or workflow, content, clarity, and acceptance
			  • System usability scale
			  • Level of acceptance of clinical reasoning and decision support
				• Ease of use: mostly negative sentiments (16/26, 62%)
				• Consistency: mostly neutral sentiments (7/13, 54%)
				• Visibility: mostly negative sentiments (7/16, 44%)
				• Navigation or workflow: mostly neutral sentiments (12/16, 75%)
				• Content: mostly positive sentiments (12/27, 44%)
				• Clarity: mostly neutral sentiments (9/15, 60%)
				• Acceptance: mostly positive sentiments (34/49, 69%)
				• System usability scale
				  • Overall system usability: excellent (mean 0.92, SD 6.5), with acceptance rated as good to excellent.
				• CDSS-pharmacists' agreement:
				  • Self-care recommendations: 90% (18/20)
				  • Medicines recommendations: 100% (25/25)
				  • Referral advice :88% (22/25)
				  • Pharmacists expressed uncertainty when screening for serious pathology in 40% (10/25) of the cases.
				• Pharmacists requested more direction from the CDSS in relation to automated prompts for user input and page navigation.

XSL•FO
RenderX

Study and CDSS name	CDSS type	Features of the CDSS	Outcomes measured	Results

Back-UP (Horizon 2020), Jansen-Kosterink et al [57], 2021	Knowledge based	• Binary logistic regression
		• Short questionnaires were completed by patients that stratified them into 1 of the 3 outcome groups. Targeted interventions were recommended for each outcome group.
			• Qualitative-based CDSS:
			• Factors that promote or hinder the acceptance of clinicians toward CDSS use
				• Reason to use a complex CDSS:
				  • Improve care of patients (assessment, n=20)
				  • Curiosity to test and use the CDSS, to see for themselves what the value of the system is (n=19)
				  • Expectation of increase in efficiency because of reduction of workload and time and allowing them to reorganize work (n=18)
				  • Support during decision-making (n=17)
				  • Patient empowerment (n=14)
				  • Work consistently with evidence-based medicine (n=8)
				  • Perceived the technology as friendly to use (n=3)
				• Barriers to using a complex CDSS:
				  • Worried about their own clinical practice and autonomy; physicians are reluctant to use a CDSS when it interferes too much with clinical practice (n=18)
				  • Do not want to use a CDSS when it comes at an increase in time and costs (n=18)
				  • The fear that the CDSS does not work correctly (n=17)
				  • A too generic approach (n=15)
				  • A lack of effectiveness and added value (n=11)
				  • Hampering personal contact with the patient (n=8)
				  • Data and security issues (n=8)
				  • Capitalizing on health care (n=4)
				  • Lack of trust (n=3)
				• If the use of CDSS is imposed by external parties, such as health care insurance companies (n=3)

SLIC[d] CDSS (Kubben), Kubben et al [58], 2011	Knowledge based	• Decision tree app–based CDSS • Offers evidence-based algorithms (eg, burst fractures, central cord syndrome, facet fracture dislocation, facet subluxation, and hyperextension injury) based on the Subaxial Injury Classification system	• Not reported	• Not reported

Peiris et al [41], 2014	Knowledge based	• Recommendations from 15 guidelines for back pain management
		  • After excluding serious pathology, the CDSS will continue to assess for the most probable diagnosis and treatment through a series of questions. A personalized information sheet will be printed.
			• Frequency of use of the web-based tool by physicians
			• Acceptability by physicians
				• Acceptability by physicians
				• Considered that back pain is easy to manage and the use of CDSS could insult their skills
				• Found CDSS useful for patient reassurance and minimizing complex tests or treatment.
				• Suggestions for improvement:
				  • Increase comprehensiveness of advice for complex pain management and referral and allow customization of advice
				  • Integration of software systems and easy navigation

[a]CDSS: clinical decision support system.

XSL•FO
RenderX

[b]DSS: decision support system.

[c]ODI: Oswestry Disability Index.

[d]SLIC: Subaxial Injury Classification.

User's Perception and Experience

Of the 31 CDSS studies reviewed, 5 (16%) studies examined the user acceptability of CDSS use and gathered feedback for improvement [31,41,56,57]. User perceptions were mixed, with the most receptive toward CDSS use [41,56,57] because it provides evidence-based content to support patient care and empowerment by involving patients in decision-making. Some perceived the use of CDSS as additional work [31], while others doubted the tool's accuracy owing to the complexity of LBP [41]. However, in cases where physicians felt that complex treatment or imaging was not recommended, CDSSs were found helpful in supporting their recommendations and reassuring patients about the decision [41]. Furthermore, the physicians were more likely to use CDSS if it lightened their workload or improved their efficiency [57].

Discussion

Principal Findings

We identified 4 major applications of the CDSS: preventing unnecessary imaging, aiding diagnosis, aiding prognosis, and suggesting treatment options. Only 2 studies used non–knowledge-based algorithms for diagnosis and treatment recommendations, whereas knowledge-based algorithms were the most commonly applied approach. Common input variables included age, gender, height, smoking status, education level, employment status, race or ethnicity, medical comorbidities, preoperative pain and disability, previous spinal surgery, symptom duration, surgical approach and intervention, BMI, American Society of Anesthesiologists score, and surgical diagnosis.

CDSS for Preventing Unnecessary Imaging

MRI detects soft tissue abnormalities [60], but the increased cost, time, and logistical demands compared with other imaging techniques make its use inconsistent with value-based care for nonspecific indications [61]. The National Emergency X-Radiography Utilization Study criteria, Canadian Cervical Spine Rule, and American College of Physicians and American Pain Society guidelines were created to direct the diagnosis and treatment of back pain and suspected spinal injury [62,63]. However, adherence to these guidelines is poor owing to *defensive medicine*, the continued use of unnecessary imaging to avoid missing serious pathologies [35,64].

Integrated CDSSs offer 2 benefits. First, they act as gatekeepers by adding an extra step before imaging is ordered [35]. Second, they educate or remind physicians of the existing guidelines, reducing the need to memorize multiple protocols [35]. Most studies have found that using the CDSS decreases the number of imaging tests ordered both at the time of the initial LBP visit and up to 30 days after the visit. However, other studies have not found a significant decrease in imaging orders, suggesting a potential mistrust of the system or a lack of awareness of imaging guidelines [34]. Furthermore, an insignificant decrease in imaging order may arise from the decision to use computerized tomography or x-ray instead of MRI, which could be more appropriate for some patients [30].

The use of alert-based CDSS raised concerns about alert fatigue, where repeated alerts may lead to physicians ignoring system prompts. Unnecessary imaging frequency was reduced when CDSS-generated report cards were distributed to physicians every 4 to 6 months compared with real-time alerts [37]. Furthermore, the ease of use of CDSS can hinder proper imaging if separate software is required, requiring the physician to toggle between the ordering and the CDSS system. In addition, the lack of real-time consequences for ignoring prompts may contribute to the continuation of unnecessary imaging practices.

Diagnostic CDSS

In general, diagnostic CDSSs operate through questionnaires that generate probable diagnoses. CDSS-generated diagnoses were found to be primarily concordant with expert or gold-standard recommendations, indicating potential feasible use. Despite its ability to provide reliable diagnoses, most studies still recommend using the diagnostic CDSS as an aid instead of a replacement for the expertise and judgment of trained and experienced health care professionals [38,39,43]. In addition, patient-provider interactions are essential, and a human connection is a part of building a healing and therapeutic relationship [65]. Health care providers can assess a patient's physical and emotional well-being better than a machine, which is only as good as its algorithm. As an aid, diagnostic CDSS could allow a brief initial assessment of the patient's condition and assist in triaging, allowing patients with critical spinal disorders to receive early attention [38,39].

To ensure generalizability and continued validity of the CDSS, it is crucial that regular updates with the latest evidence-based information be made available to the system [40]. Meanwhile, given the lack of non–knowledge-based CDSS for spinal diagnostic purposes, AI or machine learning algorithms should be explored. The potential of AI in the field of diagnosis remains to be fully tapped, especially in the areas of computer vision and image recognition. There are promising signs of the increased prominence of diagnostic CDSSs and their ability to produce faster and more accurate findings [66].

Prognostic CDSS

All the included prognostic CDSS studies used white-box models. This model allows for the adaptation and modification of variables to identify areas for optimization to improve the outcomes [67]. Traditional statistical methods for prognostic modeling use simpler computation methods that allow insight into causal effects [67]. In contrast, machine learning methods are often referred to as black-box models owing to the computational complexity that allows for fast and accurate predictions but at the cost of transparency. Previous research has shown that machine learning models may perform poorer than traditional statistical methods, suggesting that this tradeoff is not justified [68]. The poorer performance may have resulted

XSL•FO
RenderX

from using low-dimensional data; however, with the increasing availability of high-dimensional data and repositories of large data sets, such as biomarkers and imaging techniques, machine learning could have a competitive advantage over traditional statistics [54].

The prognostic CDSS systems are currently available as independent programs, as most are in the process of development or testing, and specialized sets of algorithms and flexibility for adjustments are required. Such an implementation could also be intentional to ease access for the users of a different electronic system, reduce the cost of integration, and ensure the confidentiality of data [10].

The prognostic CDSSs reviewed in our study were fragmented in their methodology, and none were ready for clinical implementation. The emergence of prognostic models employing AI and big data has been on the rise. However, reviews have identified poor standardization and quality of their development [69,70]. Previous reviews found that most prognostic model research ends with model development, with only a small number of studies performing external validation and even fewer conducting impact studies [70]. This aligns with the findings of our review, in which the included studies were found not to adhere well to standards, limiting the model's validity, generalizability, and application in real-world clinical settings. Only 2 (6%) of the 31 included studies [52,54] used a reporting guideline, namely the *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis* statement [71]. Future developments should adhere to the Prognosis Research Strategy prognostic model research framework, which emphasizes model development, external validation, impact testing [72], and reporting guidelines to ensure standardization and generalizability of the models.

The predictive ability of prognostic models is expected to weaken with time owing to changes in diagnostic and treatment approaches [72]. Therefore, it may be more beneficial to improve and recalibrate existing models instead of developing new models. In addition, including biomarkers and imagery data may improve model performance, but caution should be taken to address issues such as class imbalances, missing data, and the need for adequate validation [54]. Although adding more variables to a model can increase its predictive power, it can also make the model less user-friendly. To balance the tradeoff between accuracy and user-friendliness, parsimonious models that include only the most important or highly correlated predictors of the outcome are preferred. Techniques such as recursive feature elimination, principal component analysis, factor analysis, and multidimensional scaling can be used to identify key predictors [73].

CDSS for Treatment Recommendation

According to Benditz et al [38], only 49.6% of the treatment recommendations made by the CDSS agreed with those of the physicians [38]. Although this low level of concordance may be seen as a problem and may affect confidence in the use of the CDSS, it is important to note that concordance is not necessarily the best indicator of performance; instead, testing the clinical effects of treatment options recommended by the CDSS may be a more accurate method to assess its performance. Suggestions to improve the acceptance and usability of CDSSs include integrating them into the existing workflow and clinical decision-making processes [41]. This integration eases access to evidence-based information, encouraging use and adherence to the best practice guidelines [58].

Although the CDSS has been widely accepted for recommending treatment or management of spinal disorders, concerns and suggestions have been raised. The top barrier to CDSS use is interference with physician autonomy [57]. The physicians may feel threatened by CDSS recommendations and worry that they may eventually diminish their role in the care process [74], leading to questions about their competence [41]. In addition, ease of use is a common barrier; some physicians have negative sentiments toward the simplicity of their CDSS [56]. Furthermore, physicians are unwilling to use CDSS if it increases the time and cost [57]. Involving clinicians in the development of CDSS can improve system acceptance and adoption by ensuring that it meets the needs and preferences of users.

Strengths and Limitations

This review was conducted rigorously and adhered to established guidelines, including the JBI methodological guidance for scoping reviews and the PRISMA-ScR statement, ensuring transparency and credibility of the review [27,28]. In addition, 2 independent reviewers (ZAT and CQYH) were involved in the complete review process, which reduced potential biases. Furthermore, a systematic search was used to ensure a comprehensive coverage of the available literature.

Owing to the heterogeneous nature of the data included in this review, statistical analysis was not feasible, even among studies with similar objectives. Therefore, a rigorous and transparent scoping review was conducted to elucidate the mechanisms of action, effectiveness, and user acceptance of the CDSS for spinal disorders, with the hope of fostering interdisciplinary understanding and collaboration.

The methodology of this scoping review did not require a formal quality assessment of the included studies, and consequently, such an evaluation was not conducted. We recognize that the quality of the literature incorporated is crucial in shaping the outcomes of this study, thus constituting a limitation to the findings. During the screening process for study inclusion, interrater reliability was not systematically evaluated, representing another acknowledged limitation of this study. However, to address potential inconsistencies in judgment, we actively engaged in discussions and sought the input of a third reviewer (BB) to reach a consensus.

The current implementation of CDSSs for spinal disorders is fragmented and inconsistent, which poses a challenge to comprehending and advancing this field. The lack of a standardized reporting structure in the reviewed studies presents a limitation in quantifying the effectiveness of the CDSS. To better understand the impact of CDSS on health care delivery and optimize its use in clinical practice, further research with standardized reporting methods is needed.

Our recommendation for future work is to focus on assessing the quality of prediction models while adhering to transparent

XSL•FO
RenderX

reporting guidelines, such as the *Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis—Systematic Reviews and Meta-Analyses* [75]. Specifically, we suggest systematically evaluating models using validated tools, such as the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies to extract prognostic model studies and the Prediction Model Study Risk of Bias Assessment Tool to assess the quality of these models [76,77]. It is important to prioritize these efforts to ensure that the models are thoroughly evaluated and that their quality is properly assessed before application.

Conclusions

Previous studies assessing CDSS effectiveness typically focused on the concordance between CDSS recommendations and health care providers' decisions. A more favorable approach involves directly comparing CDSS suggestions with real clinical outcomes. To enhance CDSS development, future research should prioritize seamless system integration, considering end users' requirements. In addition, investigations into external validation and impact studies are essential for a thorough evaluation of the system's effectiveness across diverse health care settings. Emphasizing these factors will contribute to a more robust understanding of CDSS performance and its potential for broader implementation in the clinical practice for spinal disorders.

Acknowledgments

The authors would like to express their gratitude to the Research Council of Norway for funding and support throughout the course of this study. The funder played no role in the study design, data collection, analysis and interpretation of data, or writing of this manuscript.

Authors' Contributions

ZAT, BB, MP, MG, and HGH jointly conceived and designed this review. ZAT and QYCH were responsible for data collection, analysis, interpretation, and manuscript drafting. HWDH, MP, and MG provided valuable clinical and methodological insights. BB oversaw data interpretation and critically reviewed and revised the manuscript. HGH supervised the study and critically reviewed and revised the manuscript. All the authors made substantial contributions and approved the content of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for PubMed.
[DOCX File , 20 KB - jmir_v26i1e53951_app1.docx]

Multimedia Appendix 2
PRISMA ScR Checklist.
[PDF File (Adobe PDF File), 102 KB - jmir_v26i1e53951_app2.pdf]

References

1. Boos N, Aebi M. Spinal Disorders: Fundamentals of Diagnosis and Treatment. Berlin, Heidelberg: Springer; Sep 24, 2008.
2. Elfering A, Mannion AF. Epidemiology and risk factors of spinal disorders. In: Boos N, Aebi M, editors. Spinal Disorders. Berlin, Heidelberg: Springer; 2008.
3. Chen S, Chen M, Wu X, Lin S, Tao C, Cao H, et al. Global, regional and national burden of low back pain 1990-2019: a systematic analysis of the Global Burden of Disease study 2019. J Orthop Translat 2021 Sep 10;32:49-58 [FREE Full text] [doi: 10.1016/j.jot.2021.07.005] [Medline: 34934626]
4. Wu A, March L, Zheng X, Huang J, Wang X, Zhao J, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. Ann Transl Med 2020 Mar;8(6):299 [FREE Full text] [doi: 10.21037/atm.2020.02.175] [Medline: 32355743]
5. Dieleman JL, Cao J, Chapin A, Chen C, Li Z, Liu A, et al. US health care spending by payer and health condition, 1996-2016. JAMA 2020 Mar 03;323(9):863-884 [FREE Full text] [doi: 10.1001/jama.2020.0734] [Medline: 32125402]
6. Carregaro RL, Tottoli CR, Rodrigues DD, Bosmans JE, da Silva EN, van Tulder M. Low back pain should be considered a health and research priority in Brazil: lost productivity and healthcare costs between 2012 to 2016. PLoS One 2020 Apr 01;15(4):e0230902 [FREE Full text] [doi: 10.1371/journal.pone.0230902] [Medline: 32236113]
7. Gedin F, Alexanderson K, Zethraeus N, Karampampa K. Productivity losses among people with back pain and among population-based references: a register-based study in Sweden. BMJ Open 2020 Aug 13;10(8):e036638 [FREE Full text] [doi: 10.1136/bmjopen-2019-036638] [Medline: 32792439]

XSL•FO
RenderX

8. Serranheira F, Sousa-Uva M, Heranz F, Kovacs F, Sousa-Uva A. Low back pain (LBP), work and absenteeism. Work 2020;65(2):463-469. [doi: 10.3233/WOR-203073] [Medline: 32007966]
9. Siminoff LA. Incorporating patient and family preferences into evidence-based medicine. BMC Med Inform Decis Mak 2013;13 Suppl 3(Suppl 3):S6 [FREE Full text] [doi: 10.1186/1472-6947-13-S3-S6] [Medline: 24565268]
10. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
11. Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Berner E, editor. Clinical Decision Support Systems. Cham, Switzerland: Springer; 2016.
12. Shao Y, Cheng Y, Shah RU, Weir CR, Bray BE, Zeng-Treitler Q. Shedding light on the black box: explaining deep neural network prediction of clinical outcomes. J Med Syst 2021 Jan 04;45(1):5 [FREE Full text] [doi: 10.1007/s10916-020-01701-8] [Medline: 33404886]
13. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Ann Transl Med 2018 Jun;6(11):216 [FREE Full text] [doi: 10.21037/atm.2018.05.32] [Medline: 30023379]
14. Moja L, Kwag KH, Lytras T, Bertizzolo L, Brandt L, Pecoraro V, et al. Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. Am J Public Health 2014 Dec;104(12):e12-e22. [doi: 10.2105/AJPH.2014.302164] [Medline: 25322302]
15. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? a qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
16. Azimi P, Yazdanian T, Benzel EC, Aghaei HN, Azhari S, Sadeghi S, et al. A review on the use of artificial intelligence in spinal diseases. Asian Spine J 2020 Aug;14(4):543-571 [FREE Full text] [doi: 10.31616/asj.2020.0147] [Medline: 32326672]
17. Chang M, Canseco JA, Nicholson KJ, Patel N, Vaccaro AR. The role of machine learning in spine surgery: the future is now. Front Surg 2020 Aug 21;7:54 [FREE Full text] [doi: 10.3389/fsurg.2020.00054] [Medline: 32974382]
18. DelSole EM, Keck WL, Patel AA. The state of machine learning in spine surgery: a systematic review. Clin Spine Surg 2022 Mar 01;35(2):80-89. [doi: 10.1097/BSD.0000000000001208] [Medline: 34121074]
19. Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. JOR Spine 2019 Mar 05;2(1):e1044 [FREE Full text] [doi: 10.1002/jsp2.1044] [Medline: 31463458]
20. Gutman MJ, Schroeder GD, Murphy H, Flanders AE, Vaccaro AR. Artificial intelligence in spine care. Clin Spine Surg 2021 May 01;34(4):121-124. [doi: 10.1097/BSD.0000000000001082] [Medline: 32991359]
21. Joshi RS, Lau D, Ames CP. Machine learning in spine surgery: predictive analytics, imaging applications and next steps. Semin Spine Surg 2021 Jun;33(2):100878 [FREE Full text] [doi: 10.1016/j.semss.2021.100878]
22. Katsuura Y, Colón LF, Perez AA, Albert TJ, Qureshi SA. A primer on the use of artificial intelligence in spine surgery. Clin Spine Surg 2021 Nov 01;34(9):316-321. [doi: 10.1097/BSD.0000000000001211] [Medline: 34050043]
23. Khan O, Badhiwala JH, Grasso G, Fehlings MG. Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care. World Neurosurg 2020 Aug;140:512-518. [doi: 10.1016/j.wneu.2020.04.022] [Medline: 32797983]
24. Merali ZA, Colak E, Wilson JR. Applications of machine learning to imaging of spinal disorders: current status and future directions. Global Spine J 2021 Apr;11(1_suppl):23S-29S [FREE Full text] [doi: 10.1177/2192568220961353] [Medline: 33890805]
25. Rasouli JJ, Shao J, Neifert S, Gibbs WN, Habboub G, Steinmetz MP, et al. Artificial intelligence and robotics in spine surgery. Global Spine J 2021 May;11(4):556-564 [FREE Full text] [doi: 10.1177/2192568220915718] [Medline: 32875928]
26. Stephens ME, O'Neal CM, Westrup AM, Muhammad FY, McKenzie DM, Fagg AH, et al. Utility of machine learning algorithms in degenerative cervical and lumbar spine disease: a systematic review. Neurosurg Rev 2022 Apr;45(2):965-978. [doi: 10.1007/s10143-021-01624-z] [Medline: 34490539]
27. Aromataris E, Munn Z. JBI manual for evidence synthesis. Joanna Briggs Institute. 2020. URL: https://doi.org/10.46658/JBIMES-20-01 [accessed 2024-02-26]
28. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
29. EndNote. Clarivate. 2013. URL: https://endnote.com/product-details/ [accessed 2024-02-19]
30. Chen D, Bhambhvani HP, Hom J, Mahoney M, Wintermark M, Sharp C, et al. Effect of electronic clinical decision support on imaging for the evaluation of acute low back pain in the ambulatory care setting. World Neurosurg 2020 Feb;134:e874-e877. [doi: 10.1016/j.wneu.2019.11.031] [Medline: 31733384]
31. Goergen SK, Fong C, Dalziel K, Fennessy G. Can an evidence-based guideline reduce unnecessary imaging of road trauma patients with cervical spine injury in the emergency department? Australas Radiol 2006 Dec;50(6):563-569. [doi: 10.1111/j.1440-1673.2006.01655.x] [Medline: 17107528]

XSL•FO
RenderX

32. Hynes JP, Hunter K, Rochford M. Utilization and appropriateness in cervical spine trauma imaging: implementation of clinical decision support criteria. Ir J Med Sci 2020 Feb 10;189(1):333-336. [doi: 10.1007/s11845-019-02059-8] [Medline: 31292918]
33. Ip IK, Gershanik EF, Schneider LI, Raja AS, Mar W, Seltzer S, et al. Impact of IT-enabled intervention on MRI use for back pain. Am J Med 2014 Jun;127(6):512-8.e1 [FREE Full text] [doi: 10.1016/j.amjmed.2014.01.024] [Medline: 24513065]
34. Mallavarapu V, Christiason K. Effects of the elimination of the free text field in ACR select on ordering trends for low back pain imaging in the emergency department. In: Proceedings of the Society of General Internal Medicine (SGIM) Annual Meeting. Presented at: SGIM Annual Meeting; May 06-09, 2020; Birmingham, AL URL: https://europepmc.org/abstract/MED/32725471 [doi: 10.1007/s11606-020-05890-3]
35. Min A, Chan VW, Aristizabal R, Peramaki ER, Agulnik DB, Strydom N, et al. Clinical decision support decreases volume of imaging for low back pain in an urban emergency department. J Am Coll Radiol 2017 Jul;14(7):889-899. [doi: 10.1016/j.jacr.2017.03.005] [Medline: 28483544]
36. Solberg LI, Wei F, Butler JC, Palattao KJ, Vinz CA, Marshall MA. Effects of electronic decision support on high-tech diagnostic imaging orders and patients. Am J Manag Care 2010 Feb;16(2):102-106 [FREE Full text] [Medline: 20148614]
37. Zafar HM, Ip IK, Mills AM, Raja AS, Langlotz CP, Khorasani R. Effect of clinical decision support-generated report cards versus real-time alerts on primary care provider guideline adherence for low back pain outpatient lumbar spine MRI orders. AJR Am J Roentgenol 2019 Feb;212(2):386-394. [doi: 10.2214/AJR.18.19780] [Medline: 30476451]
38. Benditz A, Faber F, Wenk G, Fuchs T, Salak N, Grifka J, et al. The role of a decision support system in back pain diagnoses: a pilot study. Biomed Res Int 2019;2019:1314028 [FREE Full text] [doi: 10.1155/2019/1314028] [Medline: 31019964]
39. Benditz A, Pulido LC, Grifka J, Ripke F, Jansen P. A clinical decision support system in back pain helps to find the diagnosis: a prospective correlation study. Arch Orthop Trauma Surg 2023 Feb 04;143(2):621-625 [FREE Full text] [doi: 10.1007/s00402-021-04080-y] [Medline: 34347121]
40. Lin L, Hu PJ, Liu Sheng OR. A decision support system for lower back pain diagnosis: uncertainty management and clinical evaluations. Decis Support Syst 2006 Nov;42(2):1152-1169 [FREE Full text] [doi: 10.1016/j.dss.2005.10.007]
41. Peiris D, Williams C, Holbrook R, Lindner R, Reeve J, Das A, et al. A web-based clinical decision support tool for primary health care management of back pain: development and mixed methods evaluation. JMIR Res Protoc 2014 Apr 02;3(2):e17 [FREE Full text] [doi: 10.2196/resprot.3071] [Medline: 24694921]
42. Kim KH, Sohn MJ, Park CG. Conformity assessment of a computer vision-based posture analysis system for the screening of postural deformation. BMC Musculoskelet Disord 2022 Aug 22;23(1):799 [FREE Full text] [doi: 10.1186/s12891-022-05742-7] [Medline: 35996105]
43. Wang KC, Jeanmenne A, Weber GM, Thawait SK, Carrino JA. An online evidence-based decision support system for distinguishing benign from malignant vertebral compression fractures by magnetic resonance imaging feature analysis. J Digit Imaging 2011 Jun 3;24(3):507-515 [FREE Full text] [doi: 10.1007/s10278-010-9316-3] [Medline: 20680384]
44. Buchlak QD, Yanamadala V, Leveque JC, Edwards A, Nold K, Sethi R. The Seattle spine score: predicting 30-day complication risk in adult spinal deformity surgery. J Clin Neurosci 2017 Sep;43:247-255. [doi: 10.1016/j.jocn.2017.06.012] [Medline: 28676311]
45. Chalmers E, Westover L, Jacob J, Donauer A, Zhao VH, Parent EC, et al. Predicting success or failure of brace treatment for adolescents with idiopathic scoliosis. Med Biol Eng Comput 2015 Oct 23;53(10):1001-1009. [doi: 10.1007/s11517-015-1306-7] [Medline: 26002592]
46. Fritzell P, Mesterton J, Hagg O. Prediction of outcome after spinal surgery-using The Dialogue Support based on the Swedish national quality register. Eur Spine J 2022 Apr 27;31(4):889-900. [doi: 10.1007/s00586-021-07065-y] [Medline: 34837113]
47. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthritis Rheum 2008 May 15;59(5):632-641 [FREE Full text] [doi: 10.1002/art.23563] [Medline: 18438893]
48. Khor S, Lavallee D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, et al. Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 2018 Jul 01;153(7):634-642 [FREE Full text] [doi: 10.1001/jamasurg.2018.0072] [Medline: 29516096]
49. Lee MJ, Cizik AM, Hamilton D, Chapman JR. Predicting medical complications after spine surgery: a validated model using a prospective surgical registry. Spine J 2014 Feb 01;14(2):291-299 [FREE Full text] [doi: 10.1016/j.spinee.2013.10.043] [Medline: 24239799]
50. Lubelski D, Feghali J, Nowacki AS, Alentado VJ, Planchard R, Abdullah KG, et al. Patient-specific prediction model for clinical and quality-of-life outcomes after lumbar spine surgery. J Neurosurg Spine 2021 Jan 29;34(4):580-588. [doi: 10.3171/2020.8.SPINE20577] [Medline: 33528964]
51. Moulton H, Tosteson TD, Zhao W, Pearson L, Mycek K, Scherer E, et al. Considering spine surgery: a web-based calculator for communicating estimates of personalized treatment outcomes. Spine (Phila Pa 1976) 2018 Dec 15;43(24):1731-1738 [FREE Full text] [doi: 10.1097/BRS.0000000000002723] [Medline: 29877995]

XSL•FO
RenderX

52. Müller D, Haschtmann D, Fekete TF, Kleinstück F, Reitmeir R, Loibl M, et al. Development of a machine-learning based model for predicting multidimensional outcome after surgery for degenerative disorders of the spine. Eur Spine J 2022 Aug;31(8):2125-2136. [doi: 10.1007/s00586-022-07306-8] [Medline: 35834012]
53. Quddusi A, Eversdijk HA, Klukowska AM, de Wispelaere MP, Kernbach JM, Schröder ML, et al. External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. Eur Spine J 2020 Feb 22;29(2):374-383. [doi: 10.1007/s00586-019-06189-6] [Medline: 31641905]
54. Staartjes VE, Stumpo V, Ricciardi L, Maldaner N, Eversdijk HA, Vieli M, et al. FUSE-ML: development and external validation of a clinical prediction model for mid-term outcomes after lumbar spinal fusion for degenerative disease. Eur Spine J 2022 Oct;31(10):2629-2638 [FREE Full text] [doi: 10.1007/s00586-022-07135-9] [Medline: 35188587]
55. Byvaltsev V, Kalinin A. Assessment of clinical decision support system efficiency in spinal neurosurgery for personalized minimally invasive technologies used on lumbar spine. Sovrem Tekhnologii Med 2021 Oct;13(5):13-21 [FREE Full text] [doi: 10.17691/stm2021.13.5.02] [Medline: 35265345]
56. Downie AS, Hancock M, Abdel Shaheed C, McLachlan AJ, Kocaballi AB, Williams CM, et al. An electronic clinical decision support system for the management of low back pain in community pharmacy: development and mixed methods feasibility study. JMIR Med Inform 2020 May 11;8(5):e17203 [FREE Full text] [doi: 10.2196/17203] [Medline: 32390593]
57. Jansen-Kosterink S, van Velsen L, Cabrita M. Clinician acceptance of complex clinical decision support systems for treatment allocation of patients with chronic low back pain. BMC Med Inform Decis Mak 2021 Apr 27;21(1):137 [FREE Full text] [doi: 10.1186/s12911-021-01502-0] [Medline: 33906665]
58. Kubben PL, van Santbrink H, Cornips EM, Vaccaro AR, Dvorak MF, van Rhijn LW, et al. An evidence-based mobile decision support system for subaxial cervical spine injury treatment. Surg Neurol Int 2011 Mar 23;2:32 [FREE Full text] [doi: 10.4103/2152-7806.78238] [Medline: 21541200]
59. Foster NE, Mullis R, Hill JC, Lewis M, Whitehurst DG, Doyle C, et al. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. Ann Fam Med 2014;12(2):102-111 [FREE Full text] [doi: 10.1370/afm.1625] [Medline: 24615305]
60. Winegar BA, Kay MD, Taljanovic M. Magnetic resonance imaging of the spine. Pol J Radiol 2020;85:e550-e574 [FREE Full text] [doi: 10.5114/pjr.2020.99887] [Medline: 33101557]
61. Ashby K, Adams BN, Shetty M. Appropriate magnetic resonance imaging ordering. In: StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; Nov 14, 2022.
62. Stiell IG, Clement CM, Rowe BH, Schull MJ, Brison R, Cass D, et al. Comparison of the Canadian CT Head Rule and the New Orleans Criteria in patients with minor head injury. JAMA 2005 Sep 28;294(12):1511-1518. [doi: 10.1001/jama.294.12.1511] [Medline: 16189364]
63. Chou R, Qaseem A, Snow V, Casey D, Cross JTJ, Shekelle P, et al. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med 2007 Oct 02;147(7):478-491 [FREE Full text] [doi: 10.7326/0003-4819-147-7-200710020-00006] [Medline: 17909209]
64. Studdert DM, Mello MM, Sage WM, DesRoches CM, Peugh J, Zapert K, et al. Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. JAMA 2005 Jun 01;293(21):2609-2617. [doi: 10.1001/jama.293.21.2609] [Medline: 15928282]
65. Drossman DA, Chang L, Deutsch JK, Ford AC, Halpert A, Kroenke K, et al. A review of the evidence and recommendations on communication skills and the patient-provider relationship: a Rome Foundation Working Team report. Gastroenterology 2021 Nov;161(5):1670-88.e7 [FREE Full text] [doi: 10.1053/j.gastro.2021.07.037] [Medline: 34331912]
66. Malamateniou C, Knapp KM, Pergola M, Woznitza N, Hardy M. Artificial intelligence in radiography: where are we now and what does the future hold? Radiography (Lond) 2021 Oct;27 Suppl 1:S58-S62. [doi: 10.1016/j.radi.2021.07.015] [Medline: 34380589]
67. Rajula HS, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina (Kaunas) 2020 Sep 08;56(9):455 [FREE Full text] [doi: 10.3390/medicina56090455] [Medline: 32911665]
68. Ren G, Yu K, Xie Z, Wang P, Zhang W, Huang Y, et al. Current applications of machine learning in spine: from clinical view. Global Spine J 2022 Oct 10;12(8):1827-1840 [FREE Full text] [doi: 10.1177/21925682211035363] [Medline: 34628966]
69. Groot OQ, Ogink PT, Lans A, Twining PK, Kapoor ND, DiGiovanni W, et al. Machine learning prediction models in orthopedic surgery: a systematic review in transparent reporting. J Orthop Res 2022 Feb;40(2):475-483 [FREE Full text] [doi: 10.1002/jor.25036] [Medline: 33734466]
70. Lubelski D, Hersh A, Azad TD, Ehresman J, Pennington Z, Lehner K, et al. Prediction models in degenerative spine surgery: a systematic review. Global Spine J 2021 Apr;11(1_suppl):79S-88S [FREE Full text] [doi: 10.1177/2192568220959037] [Medline: 33890803]
71. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 07;350:g7594 [FREE Full text] [doi: 10.1136/bmj.g7594] [Medline: 25569120]

XSL•FO
RenderX

72. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381 [FREE Full text] [doi: 10.1371/journal.pmed.1001381] [Medline: 23393430]
73. Jia W, Sun M, Lian J, Hou S. Feature dimensionality reduction: a review. Complex Intell Syst 2022 Jan 21;8(3):2663-2693. [doi: 10.1007/s40747-021-00637-x]
74. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
75. Snell KI, Levis B, Damen JA, Dhiman P, Debray TP, Hooft L, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). BMJ 2023 May 03;381:e073538 [FREE Full text] [doi: 10.1136/bmj-2022-073538] [Medline: 37137496]
76. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744 [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
77. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
JBI: Joanna Briggs Institute
LBP: low back pain
LS-MRI: lumbar spine–magnetic resonance imaging
MRI: magnetic resonance imaging
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-analyses Extension for Scoping Reviews

Edited by G Tsafnat; submitted 28.10.23; peer-reviewed by L Yu, A Montazeri; comments to author 25.01.24; revised version received 29.01.24; accepted 10.02.24; published 19.03.24.

Please cite as:
Toh ZA, Berg B, Han QYC, Hey HWD, Pikkarainen M, Grotle M, He HG
Clinical Decision Support System Used in Spinal Disorders: Scoping Review
J Med Internet Res 2024;26:e53951
URL: https://www.jmir.org/2024/1/e53951
doi:10.2196/53951
PMID:38502157

©Zheng An Toh, Bjørnar Berg, Qin Yun Claudia Han, Hwee Weng Dennis Hey, Minna Pikkarainen, Margreth Grotle, Hong-Gu He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Methodological Frameworks and Dimensions to Be Considered in Digital Health Technology Assessment: Scoping Review and Thematic Analysis

Joan Segur-Ferrer[1], BSS, PT, MSc; Carolina Moltó-Puigmartí[1], BScPharm, PhD; Roland Pastells-Peiró[1], BA, MA, MsC; Rosa Maria Vivanco-Hidalgo[1], MD, MPH, PhD

Agency for Health Quality and Assessment of Catalonia, Barcelona, Spain

Corresponding Author:
Joan Segur-Ferrer, BSS, PT, MSc
Agency for Health Quality and Assessment of Catalonia
Roc Boronat Street, 81-95, 2nd Fl
Barcelona, 08005
Spain
Phone: 34 935 513 900
Fax: 34 935 517 510
Email: joan.segur@gencat.cat

Abstract

Background: Digital health technologies (dHTs) offer a unique opportunity to address some of the major challenges facing health care systems worldwide. However, the implementation of dHTs raises some concerns, such as the limited understanding of their real impact on health systems and people's well-being or the potential risks derived from their use. In this context, health technology assessment (HTA) is 1 of the main tools that health systems can use to appraise evidence and determine the value of a given dHT. Nevertheless, due to the nature of dHTs, experts highlight the need to reconsider the frameworks used in traditional HTA.

Objective: This scoping review (ScR) aimed to identify the methodological frameworks used worldwide for digital health technology assessment (dHTA); determine what domains are being considered; and generate, through a thematic analysis, a proposal for a methodological framework based on the most frequently described domains in the literature.

Methods: The ScR was performed in accordance with the guidelines established in the PRISMA-ScR guidelines. We searched 7 databases for peer reviews and gray literature published between January 2011 and December 2021. The retrieved studies were screened using Rayyan in a single-blind manner by 2 independent authors, and data were extracted using ATLAS.ti software. The same software was used for thematic analysis.

Results: The systematic search retrieved 3061 studies (n=2238, 73.1%, unique), of which 26 (0.8%) studies were included. From these, we identified 102 methodological frameworks designed for dHTA. These frameworks revealed great heterogeneity between them due to their different structures, approaches, and items to be considered in dHTA. In addition, we identified different wording used to refer to similar concepts. Through thematic analysis, we reduced this heterogeneity. In the first phase of the analysis, 176 provisional codes related to different assessment items emerged. In the second phase, these codes were clustered into 86 descriptive themes, which, in turn, were grouped in the third phase into 61 analytical themes and organized through a vertical hierarchy of 3 levels: level 1 formed by 13 domains, level 2 formed by 38 dimensions, and level 3 formed by 11 subdimensions. From these 61 analytical themes, we developed a proposal for a methodological framework for dHTA.

Conclusions: There is a need to adapt the existing frameworks used for dHTA or create new ones to more comprehensively assess different kinds of dHTs. Through this ScR, we identified 26 studies including 102 methodological frameworks and tools for dHTA. The thematic analysis of those 26 studies led to the definition of 12 domains, 38 dimensions, and 11 subdimensions that should be considered in dHTA.

(J Med Internet Res 2024;26:e48694) doi:10.2196/48694

KEYWORDS

digital health; eHealth; mHealth; mobile health; AI; artificial intelligence; framework; health technology assessment; scoping review; technology; health care system; methodological framework; thematic analysis

XSL•FO
RenderX

Introduction

Background

Digital health technologies (dHTs) are driving the transformation of health care systems. They are changing the way in which health services are delivered, and showing great potential to address some of the major challenges that European health systems, including the Spanish National Health System (SNS), are facing, such as the progressive aging of the population [1,2]; the growing demand for health and long-term care services [2]; the rise in health care costs, increasing financial pressures on health and welfare systems [1,3]; and the unequal distribution of health services across different geographical regions [4,5]. In addition, dHT can improve the accessibility, sustainability, efficiency, and quality of health care systems [6,7], leading to their becoming a determinant of health on their own [6,8].

However, the digital transformation of health care systems and the implementation of dHT (eg, artificial intelligence [AI]–based solutions, data-driven health care services, or the internet of things) are slow and unequal across different European regions [9,10]. Some of the reasons for this are (1) the immaturity of regulatory frameworks for the use of dHTs [9], (2) the lack of funding and investment for the implementation of dHTs [9], (3) the lack of sufficient and appropriate infrastructures and common standards for data management [6,9], (4) the absence of skills and expertise of professionals and users [10], and (5) the scarcity of strong evidence regarding the real benefits and effects of dHTs on health systems and people's well-being, as well as the cost-effectiveness of these technologies. This makes decision-making difficult, potentially leading to the development and reproduction of low-value and short-lived dHTs [6,11].

To overcome these challenges, harness the potential of dHTs, and avoid nonintended consequences, the World Health Organization (WHO) [4,11] states that dHTs should be developed under the principles of transparency, accessibility, scalability, privacy, security, and confidentiality. Their implementation should be led by robust strategies that bring together leadership, financial, organizational, human, and technological resources, and decisions should be guided by the best-available evidence [4,11].

Regarding this last aspect, health technology assessment (HTA), defined as a "multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its life cycle," is a widely accepted tool to inform decision-making and promote equitable, efficient, and high-quality health systems [12,13].

Generally, HTA is conducted according to specific methodological frameworks, such as the HTA Core Model of the European Network for Health Technology Assessment (EUnetHTA) [14] and the guidelines for the development and adaptation of rapid HTA reports of the Spanish Network of Agencies for Assessing National Health System Technologies and Performance (RedETS) [15]. These frameworks establish the methodologies to follow and the elements to evaluate. Although these frameworks are helpful instruments for evaluating various health technologies, they have certain limitations in comprehensively assessing dHTs. For this reason, in the past few years, different initiatives have emerged to adapt existing methodological frameworks or develop new ones. The objective is to consider additional domains (eg, interoperability, scalability) to cover the intrinsic characteristics of dHTs [16-18]. Examples of these initiatives are the Evidence Standard Framework (ESF) of National Institute for Health and Care Excellence (NICE) [19] or the Digi-HTA Framework of the Finnish Coordinating Center for Health Technology Assessment (FinCCHTA) [16]. Nonetheless, the majority of these frameworks have certain constraints, such as being designed for a particular socioeconomic or national setting, which restricts their transferability or suitability for use in other countries; the specificity or exclusion of certain dHTs, resulting in limitations in their application; or the limited evidence regarding their actual usefulness.

In this context, we performed a scoping review (ScR) with the aim of identifying the methodological frameworks that are used worldwide for the evaluation of dHTs; determining what dimensions and aspects are considered for each type of dHT; and generating, through a thematic analysis, a proposal for a methodological framework that is based on the most frequently described dimensions in the literature. This research focused mainly on mobile health (mHealth), non–face-to-face care models and medical devices that integrate AI, as these particular dHTs are the ones most frequently assessed by HTA agencies and units of RedETS.

Identifying Research Questions

This ScR followed by a thematic analysis answered the following research questions:

- What methodological frameworks currently exist for digital health technology assessment (dHTA)?
- What domains and dimensions are considered in dHTA?
- Do the different domains and dimensions considered depend on whether the dHT addressed is a non–face-to-face care model of health care provision, a mobile device (mHealth), or a device that incorporates AI?

Methods

Overview of Methods for Conducting the Scoping Review

We conducted an ScR of the literature and a thematic analysis of the studies included according to the published protocol [20]. The ScR aimed to answer the first research question, while the thematic analysis aimed to answer the second and third research questions. Spanish experts from various domains of HTA and dHT collaborated throughout the study design and development.

The ScR of the available scientific literature was carried out in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews) guidelines (Multimedia Appendix 1) [21] and following the recommendations of Peters et al [22] and Pollock et al [23].

XSL•FO
RenderX

Ethical Considerations

As this work was an ScR, no ethical board approval was required.

Search Strategy

The search strategy (Multimedia Appendix 2) was designed by an experienced information specialist (author RP-P) in accordance with the research questions and using the validated filter of Ayiku et al [24] for health apps, adding the terms for concepts related to mHealth, remote care models, AI, digital health, methodological frameworks, and HTA. The strategy was peer-reviewed according to the "Peer Review of Electronic Search Strategies Statement" [25] by authors JS-F and CM-P and was executed in the following 7 databases, considering the characteristics of each in terms of syntax, controlled vocabulary, and proximity operators: Medline (OVID), CINAHL Plus, Embase, Cochrane Library, Scopus, Web of Science, and TripDatabase. Note that no time, language, or other filters were used.

The identification of relevant studies was complemented with a manual search based on the references in the included studies, as well as the websites of the HTA agencies identified through the web pages of EUnetHTA, the International Network for Agencies for Health Technology Assessment (INAHTA), and Health Technology Assessment International (HTAi). Additionally, a search was conducted in Google Scholar, limiting the results to the first 250 items in order to guarantee the inclusion of all pertinent studies [26].

Inclusion and Exclusion Criteria

The inclusion criteria used in the reference-screening process were based on the previously detailed research questions and are outlined in Textbox 1 using the Population/Problem, Phenomenon of Interest, Context and Design (PICo-D) format [27,28]. The PICo-D format was used instead of the traditional Population/Problem, Intervention, Comparator, Outcomes, Design (PICO-D) format due to the qualitative nature of the research questions and the characteristics of the phenomenon of interest.

Studies were excluded if they were published before 2011, due to the rapid evolution of dHTs in the past few years, did not describe dimensions or evaluation criteria, or were based on methodological frameworks not intended for the assessment of dHTs (eg, EUnetHTA Core Model 3.0). Likewise, we excluded comments, editorials, letters, conference abstracts, frameworks, or tools focusing on the evaluation of dHTs by users (eg, User version of Mobile App Rating Scale [uMARS]) or documents in languages other than English, Spanish. or Catalan.

Textbox 1. Research questions in Population/Problem, Phenomenon of Interest, Context and Design (PICo-D) format.

Population/problem

Digital health technology assessment (dHTA)

Phenomenon of interest

Specific methodological frameworks for the evaluation of digital health (with special focus on mobile health [mHealth]: non–face-to-face care models and medical devices that integrate artificial intelligence [AI] due the type of technologies mostly assessed in the Spanish National Health System [SNS]) that describe the domains to be evaluated in dHTA

Context

Health technology assessment (HTA)

Design

Methodological guidelines and frameworks, scoping reviews (ScRs), systematic reviews (SRs), consensus documents, and qualitative studies

Reference Screening and Data Extraction

The screening of studies was carried out by authors CM-P and JS-F in 2 phases in accordance with the selection criteria detailed earlier (Textbox 1) and in a single-blind peer review manner. The first phase consisted of screening of the titles and abstracts of the studies identified in the bibliographic search. The second phase consisted of full-text screening of the studies included in the previous phase.

Data extraction was performed by 3 authors (CM-P, RP-P, and JS-F) using the web and desktop versions of ATLAS.ti version 22.0 (Scientific Software Development GmbH) [29] and the data extraction sheets designed ad hoc for this purpose following the recommendations of the *Cochrane Handbook for Systematic Reviews of Interventions* [30].

When disagreements emerged in either of the 2 processes, a consensus was reached between the 3 reviewers (CM-P, RP-P, and JS-F). When a consensus was not possible, a fourth reviewer (author RMV-H) was consulted.

Collecting, Summarizing, and Reporting the Results

A descriptive analysis was carried out to evaluate and report the existing methodological frameworks and their characteristics.

Overview of Methods for Thematic Analysis

The thematic analysis was performed following the recommendations and phases described by Thomas and Harden [31] to determine HTA dimensions for dHTs: (1) line-by-line text coding, (2) development of descriptive topics, and (3) generation of analytical themes. Both analyses were carried out by 3 authors (CM-P, RP-P, and JS-F) using the web and desktop versions of ATLAS.ti version 22.0 [29].

Dimensions identified from systematic reviews (SRs) that were derived from primary studies also identified in our systematic search were only counted once in order to avoid duplication of data and risk of bias. It is worth mentioning that the primary

XSL•FO
RenderX

studies included in the SRs were not directly analyzed but were analyzed through the findings reported in the SRs.

Results

Study Selection and Characteristics

A total of 3042 studies were retrieved throughout the systematic (n=3023, 99.4%) and the manual (n=19, 0.6%) search. Of these, 2238 (73.6%) studies were identified as unique after removing duplicates.

After title and abstract review, 81 (3.6%) studies were selected for full-text review, of which 26 (32.1%) were finally included in the analysis. The excluded studies and reasons for exclusion are detailed in Multimedia Appendix 3; in brief, the reasons for exclusion were phenomenon of interest (n=30, 37%), type of publication (n=15, 18.5%), purpose (n=6, 7.4%), language (n=2, 2.5%), and duplicated information (n=2, 2.5%). The study selection process is outlined in Figure 1 [32].

Of the 26 (32.1%) studies included in this ScR, 19 (73.1%) were designed as specific methodological frameworks for dHTA [16,17,33-47], 4 (15.4%) were SRs [48-51], 1 (3.9%) was a report from the European mHealth Hub's working group on mHealth assessment guidelines [52], 1 (3.9%) was a qualitative study [53], and 1 (3.9%) was a viewpoint [54]. In addition, 3 (11.5%) focused on the assessment of non–face-to-face care models [33-35], 8 (30.8%) on mHealth assessment [36-40,52,53,55], 2 (7.7%) on the assessment of AI technology [41,54], 4 (15.4%) on eHealth [42,43,48,50], and 9 (34.6%) on the overall assessment of digital health [16,17,44-47,49,51,56].

Figure 1. PRISMA 2020 flow diagram of the search and study selection process for new SRs, meta-analyses, and ScRs. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis; ScR: scoping review; SR: systematic review.

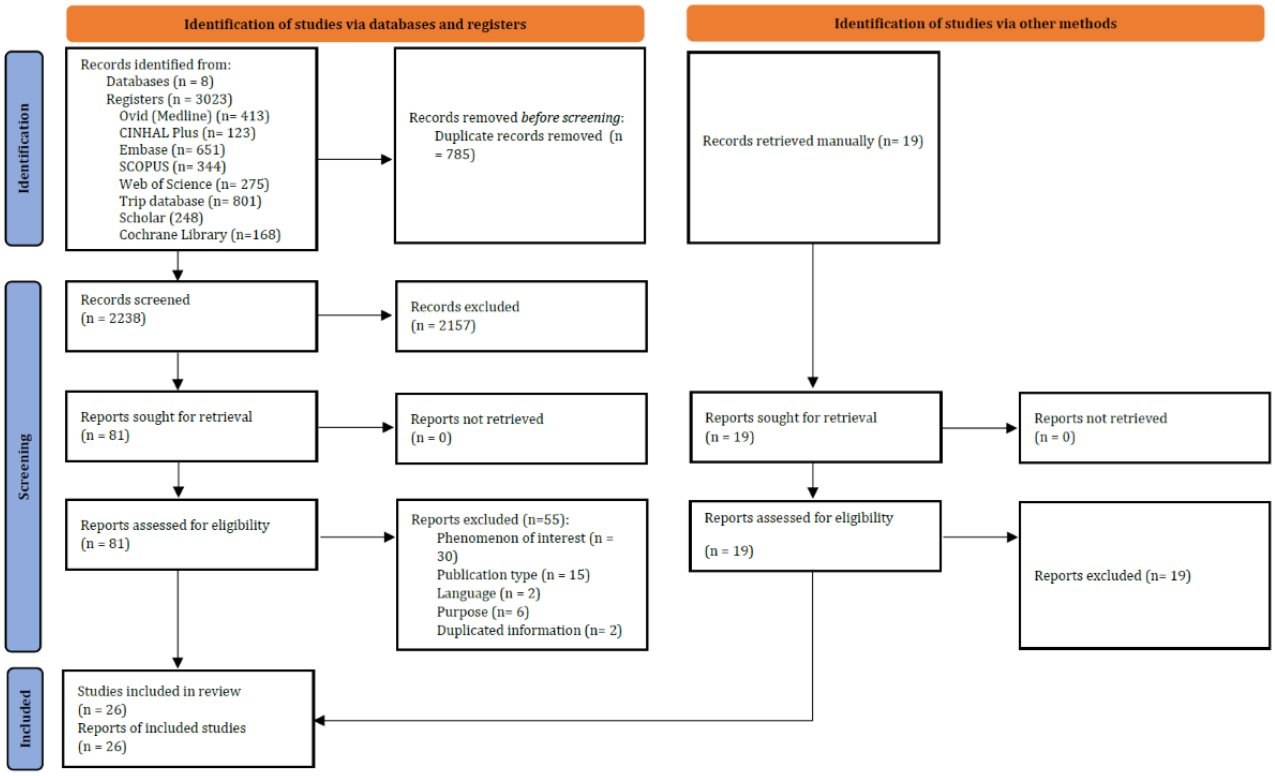

Research Question 1: Description of Identified Frameworks for dHTA

The 19 methodological frameworks for dHTA [16,17,33-47] were from various countries: The majority (n=5, 26.3%) originated in Australia [17,34,38,41,46], followed by 3 (15.8%) from the United States [43,45,56] and 2 (10.5%) from Switzerland [47,55]; the remaining 9 (47.4%) frameworks were developed in Afghanistan [42], Denmark [33], Scotland [35], Finland [16], Ireland [36], Israel [40], the United Kingdom [37], Spain [39], and Sweden [44].

The 19 methodological frameworks focused on evaluating various types of technologies. Specifically, 3 (15.8%) of them were designed for assessing non–face-to-face care models [33-35], 6 (31.6%) for mHealth [36-40], and 1 (5.3%) for AI solutions [41]. The other 9 (47.4%) frameworks addressed eHealth [42,43,56] or digital health in general [16,17,44-47], which encompasses non–face-to-face care models, mHealth, and occasionally AI-based solutions [18] within its scope. It is pertinent to mention that the differentiation between the methodological frameworks designed for the evaluation of eHealth and those designed for dHTA was based on the specific terminology and descriptions used by the authors of those frameworks.

The structures and characteristics of the analyzed methodological frameworks were considered heterogeneous in terms of evaluation specificity (whether they focused on a global evaluation that encompassed more than 1 domain or dimension or on a specific assessment that addressed only 1 domain or dimension), assessment approach (whether they adopted a

XSL•FO
RenderX

phased evaluation, a domain evaluation, or a hybrid of both), and number of domains included. Regarding evaluation specificity, 17 (89.5%) methodological frameworks were classified as global as they covered various aspects or domains within their scope [16,17,33-36,38-47,55,56], while 2 (10.5%) were classified as specific as they concentrated exclusively on 1 element or domain of assessment [37,46]. Regarding the assessment approach, 14 (73.7%) methodological frameworks proposed a domain-based evaluation [16, 17, 33, 35, 36, 38-40, 43, 44, 46, 55, 56], while 4 (21.1%) proposed a hybrid one (phased and domain based) [41,42,45,47]; the remaining methodological framework did not fit into any of the previous categories, as it was not structured by domains or phases but by types of risk [37]. Finally, the number of evaluation domains considered ranged from 1 to 14, with an average of 7. Table 1 outlines the primary features of the included methodological frameworks and provides a thorough breakdown of the domains and dimensions they address.

In contrast, from 3 (75%) [49-51] of the 4 SRs [48-51] and the report from the working group on guidelines for the evaluation of mHealth solutions from the European mHealth Hub [52], we identified other methodological frameworks and tools focusing on the assessment of dHTs. Specifically, we identified 16 methodological frameworks or tools focusing on the evaluation of non–face-to-face care models [57-72], along with 37 for the evaluation of mHealth [10,52,73-95], 11 for the evaluation of eHealth [96-107], and 17 for the evaluation of dHTs in general [108-124]. Additionally, 5 (26.3%) [33,34,36,37,42] of the 19 methodological frameworks included in this ScR were also identified and analyzed in 1 or more of the 4 literature synthesis documents [49-52]. It is important to note that the difference between the frameworks we retrieved through our systematic search and those identified in the 4 SRs is the result of the narrower perspective we adopted, focusing exclusively on frameworks directly relevant to the HTA field, in line with the aims of our study. In Multimedia Appendix 4, we provide a more detailed explanation of the methodological frameworks included in the studies mentioned earlier [19,49-52,57-73,75-135].

XSL•FO
RenderX

Table 1. Methodological frameworks (N=19) included in this ScR[a].

Methodological framework, year	Country	Assessment specificity	Assessment approach	Assessment domains, n
Methodological frameworks focusing on the assessment of non–face-to-face care models (n=3, 15.8%)				
Model for Assessment of Telemedicine Applications (MAST), 2012 [33]	Denmark	Overall evaluation	By domain	Health problem and application description; security; clinical effectiveness; patient perspective; economic aspects; organizational aspects; sociocultural, ethical, and legal aspects (n=7)
Scottish Centre for Telehealth & Telecare (SCTT) Toolkit [35]	Scotland	Overall evaluation	By domain	User benefits/costs; benefits/service costs; user experience; increased use of the technology platform available to support routine local services; confidence in the use and awareness of staff; awareness of telehealth and telecare as tools (n=6)
Telehealth framework, 2014 [34]	Australia	Overall evaluation	By domain	Health domain; health services; communication technologies; environment configuration; socioeconomic evaluation (n=5)
Methodological frameworks focusing on the evaluation of mHealth[b] technologies (n=6, 31.9%)				
Caulfield's evaluation framework, 2019 [36]	Ireland	Overall evaluation	By domain	Context information; cost information; normative compliance; scientific evidence; human factors; data collection and interpretation (n=6)
mHealth-based technology assessment for mobile apps, 2020 [39]	Spain	Overall evaluation	By domain	General information about the clinical condition and about the mHealth solution; privacy and security; technological aspects and interoperability; evidence and clinical effectiveness; user experience, usability, acceptability, ease of use, and aesthetics; costs and economic evaluation; impact on the organization (n=7)
Henson's app evaluation framework, 2019 [40]	Israel	Overall evaluation	By domain	Context information; privacy/security; scientific evidence; usability; data integration (n=5)
Lewis's assessment risk framework, 2014 [37]	United Kingdom	Risk/safety assessment	N/A[c]	Risk (n=1)
Mobile medical app evaluation module, 2020 [38]	Australia	Overall evaluation	By domain	Description and technical characteristics; current use of technology; effectiveness; security; effectivity cost; organizational aspects; ethical aspects; legal aspects; postmarket monitoring; social aspects (n=10)
Vokinger, 2020 [55]	Swiss	Overall evaluation	By domain	Purpose; usability; information accuracy; organizational reputation; transparency; privacy; self-determination or user control (n=7)
Methodological frameworks focusing on the assessment of solutions based on AI[d] (n=1, 5.3%)				
Translational Evaluation of Healthcare AI (TEHAI), 2021 [41]	Australia	Overall evaluation	Hybrid	Ability; utility; adoption (n=3)
Methodological frameworks focusing on the evaluation of eHealth technologies (n=3, 15.8%)				
Health Information Technology Evaluation Framework (HITREF), 2015 [43]	United States	Overall evaluation	By domain	Structural quality; quality of information logistics; unintended consequences/benefits; effects on quality-of-care outcomes; effects on process quality (n=5)
Heuristic evaluation of eHealth interventions, 2016 [56]	United States	Overall evaluation	By domain	Usability/ease of use/functionality; aesthetics; security; content; adherence; persuasive design; research evidence; owner credibility (n=8)
Khoja-Durrani-Sajwani (KDS) Framework, 2013 [42]	Afghanistan, Canada, Kenya, Pakistan	Overall evaluation	Hybrid	Health service results; technology results; economic results; sociotechnical and behavioral results; ethical results; preparation and change results; results of the regulation (n=7)
Methodological frameworks focusing on the evaluation of dHTs[e] (n=6, 31.6%)				
Deontic accountability framework, 2019 [46]	Australia	Ethical evaluation	By domain	Ethical principles (n=1)

XSL•FO
RenderX

Methodological framework, year	Country	Assessment specificity	Assessment approach	Assessment domains, n
Digi HTA, 2019 [16]	Finland	Overall evaluation	By domain	Company information; product information; technical stability; costs; effectiveness; clinical safety; data protection and security; usability and accessibility; interoperability; AI; robots (n=11)
Digital Health Scorecard, 2019 [45]	United States	Overall evaluation	Hybrid	Technical validation; clinical validation; usability; costs (n=4)
Framework for the design and evaluation of digital health interventions (DEDHI), 2019 [44]	Sweden	Overall evaluation	By domain	Easy to use; content quality; privacy and security; responsibility; adherence; aesthetics; perceived benefits; effectiveness; quality of service; personalization; perceived enjoyment; ethics; security (n=13)
Monitoring and Evaluating Digital Health Interventions Guide, 2016 [47]	Swiss	Overall evaluation	Hybrid	Costs; feasibility; usability; effectiveness; implementation science; efficiency; quality; use (n=8)
Precision Health Applications Evaluation Framework, 2021 [17]	Australia	Overall evaluation	By domain	Novelty; adaptability; information management; performance; clinical effectiveness; quality assurance (n=6)

[a]ScR: scoping review.

[b]mHealth: mobile health.

[c]N/A: not applicable.

[d]AI: artificial intelligence.

[e]dHT: digital health technology.

Research Question 2: Domains and Dimensions Being Considered in dHTA

The 26 (32.1%) studies included encompassed a broad range of items to consider in dHTA and often used diverse expressions for analogous concepts. We reduced this heterogeneity through our thematic analysis according to the recommendations and phases described by Thomas and Harden [31].

In this sense, in the first phase of thematic analysis, we identified and coded 176 units of meaning (coded as provisional codes) that represented different items (domains or dimensions) of the assessment. These units were then grouped into 86 descriptive themes (second phase), which were further refined into 61 analytical themes that captured the key concepts and relationships between them (third phase). Lastly, the 61 analytical themes were arranged in a 3-level vertical hierarchy based on the evidence: level 1 (12 domains), level 2 (38 dimensions), and level 3 (11 subdimensions). We used the term "domain" to refer to a distinct area or topic of evaluation that is integral to the assessment of the technology in question. A domain may encompass multiple related concepts or dimensions that are relevant to the evaluation. Each dimension, in turn, represents a specific aspect of evaluation that belongs to the domain and contributes to an understanding of its overall significance. Finally, a subdimension refers to a partial element of a dimension that facilitates its analysis. By using these terms, we aimed to provide a clear, rigorous, and comprehensive framework for conducting HTA.

Table 2 displays the 61 analytical themes in descending order of coding frequency, aligned with the hierarchy derived from the data analysis. Additionally, the table specifies the intervention modalities or dHTs that correspond to each code and lists the studies from which each code originated. The network of relationships among the codes can be found in Multimedia Appendix 5.

XSL•FO
RenderX

Table 2. Analytical themes of the thematic analysis presented in descending order of coding frequency and aligned with the hierarchy derived from the data analysis.

Domain (level 1) and dimensions (level 2)	Subdimension (level 3)	Type of dHT[a]
Description of the technology (n=19, 6.2%)		Non–face-to-face care models [33,34], mHealth[b] [36,38-40,55], AI[c] [41], eHealth [50], digital health [16,17,49,51]
Credibility and reputation (n=5, 1.6%)	—[d]	mHealth [40,55], digital health [56]
Scientific basis (n=5, 1.6%)	—	mHealth [36,39,40], digital health [17,56]
Technical evaluation and validation (n=3, 1.0%)	—	mHealth [36], digital health [45]
Adoption (n=2, 0.6%)	—	AI [41], digital health [47]
Adoption (n=2, 0.6%)	Usage (n=2, 0.6%)	AI [41], digital health [47]
Adoption (n=2, 0.6%)	Integration (n=1, 0.3%)	AI [41]
Information management (n=2, 0.6%)	—	eHealth [43], digital health [17]
Novelty (n=1, 0.7%)	—	Digital health [17]
Safety (n=19, 6.2%)		Non–face-to-face care models [33], mHealth [37-40,52], AI [41], eHealth [43,50], digital health [16,44,49,51,56]
Clinical safety (n=12, 3.9%)	—	Non–face-to-face care models [33], mHealth [37,38,52], AI [41], eHealth [43,50], digital health [16,44,49,51]
Technical safety (n=11, 3.6%)	—	Non–face-to-face care models [33], mHealth [37,39,52], digital health [16,44,49,51,56]
Clinical efficacy and effectiveness (n=17, 5.5%)		Non–face-to-face care models [33,35], mHealth [38,39,52,53], eHealth [43,50], digital health [16,44,45,47,49,51]
Economic aspects (n=16, 5.2%)		Non–face-to-face care models [33-35], mHealth [36,38,39], AI [54], eHealth [42,48,50], digital health [16,45,47,49,51]
Costs (n=10, 3.2%)	—	Non–face-to-face care models [33,35], mHealth [36,39], AI [54], digital health [16,45,47,49,51]
Economic evaluation (n=7, 2.3%)	—	Non–face-to-face care models [33,35], mHealth [38,39], eHealth [42,48], digital health [51]
Use of resources (n=4, 1.3%) and efficiency (n=1, 0.3%)	—	Non–face-to-face care models [35], mHealth [36], AI [54], digital health [47,49]
Ethical aspects (n=13, 4.2%)		Non–face-to-face care models [33], mHealth [38,55], AI [41,54], eHealth [42,48,50], digital health [44,46,49,51]
Equity (n=1, 0.3%)	—	Digital health [19]
User control and self-determination (n=1, 0.3%)	—	mHealth [55], digital health [46]
Responsibility (n=1, 0.3%)	—	Digital health [44,46]
Explainability (n=1, 0.3%)	—	Digital health [46]
Human and sociocultural aspects (n=13, 4.2%)		Non–face-to-face care models [33,35], mHealth [36,38,39], AI [54], eHealth [42,48], digital health [17,49,51]
User experience (n=7, 2.3%)	—	Non–face-to-face care models [35], mHealth [39,40,52], digital health [17,44,56]
Accessibility (n=3, 1.0%)	—	mHealth [52]
Acceptability (n=2, 0.6%)	—	mHealth [39], AI [41]
Engagement (n=2, 0.6%)	—	Digital health [44,56]
Perceived profit (n=1, 0.3%)	—	Digital health [44]
Organizational aspects (n=3.68%)		Non–face-to-face care models [33], mHealth [38,39], AI [41,54], eHealth [42,43,48,50], digital health [49,51]
Legal and regulatory aspects (n=10, 3.2%)		Non–face-to-face care models [33], mHealth [36,38], AI [54], eHealth [42,48,50], digital health [17,49,51]
Privacy (n=6, 1.9%)	—	mHealth [39,40,52,55], AI [41], digital health [44]

XSL•FO
RenderX

Domain (level 1) and dimensions (level 2)	Subdimension (level 3)	Type of dHT[a]
Transparency (n=4, 1.3%)	—	mHealth [52,55], AI [41,54]
Responsibility (n=1, 0.3%)	—	Digital health [44]
Description of health problem (n=8, 2.6%)		Non–face-to-face care models [33,34], mHealth [39], eHealth [50], digital health [49,51]
Content (n=5, 1.6%)		mHealth [55], eHealth [50], digital health [44,56]
Information adequacy (n=2, 0.6%)	—	mHealth [55], digital health [56]
Intervention adequacy (n=2, 0.6%)	—	Digital health [56]
Technical aspects (n=4, 1.3%)		AI [54], eHealth [42,48], digital health [17]
Usability (n=10, 3.2%)	—	mHealth [39,52,55], digital health [16,44,45,47,49,56]
Adaptability (n=8, 2.6%)	—	Digital health [17]
Adaptability (n=8, 2.6%)	Interoperability (n=4, 1.3%)	mHealth [39,52], digital health [16,49]
Adaptability (n=8, 2.6%)	Scalability (n=2, 0.6%)	mHealth [52], AI [41]
Adaptability (n=8, 2.6%)	Integration of data (n=1, 0.3%)	mHealth [40]
Adaptability (n=8, 2.6%%)	Transferability (n=1, 0.3%)	eHealth [48]
Quality (n=5, 1.6%)	—	eHealth [43], digital health [17,44,47]
Design (n=5, 1.6%)	—	Digital health [56]
Design (n=5, 1.6%)	Persuasive design (n=1, 0.3%)	Digital health [56]
Technical stability (n=4, 1.3%)	—	mHealth: [38,52], digital health [16,47,49]
Aesthetics (n=3, 1.0%)	—	mHealth [39], digital health [44,56]
Ease of use (n=3, 1.0%)	—	mHealth [39,40], digital health [56]
Accessibility (n=2, 0.6%)	—	mHealth [52], digital health [16]
Technical effectiveness (n=1, 0.3%) or performance (n=2, 0.6%)	—	Digital health [17,47]
Technical effectiveness (n=1, 0.3%) or performance (n=2, 0.6%)	Reliability (n=6, 1.9%)	mHealth [52,53], digital health [47]
Technical effectiveness (n=1, 0.3%) or performance (n=2, 0.6%)	Validity (n=5, 1.6%)	mHealth [52,53], AI [41]
Technical effectiveness (n=1, 0.3%) or performance (n=2, 0.6%)	Accuracy (n=2, 0.6%)	Digital health [19]
Technical effectiveness (n=1, 0.3%) or performance (n=2, 0.6%)	Sensitivity (n=1, 0.3%)	Digital health [17]
Feasibility (n=1, 0.3%)	—	Digital health [47]
Generalizability and reproducibility (n=1, 0.3%)	—	AI [54]
Interpretability (n=1, 0.3%)	—	AI [54]
Customization (n=1, 0.3%)	—	Digital health [44]
Postmarketing monitoring (n=3, 1%)		mHealth [38], digital health [47]

[a]dHT: digital health technology.

[b]mHealth: mobile health.

[c]AI: artificial intelligence.

[d]N/A: not applicable.

XSL•FO
RenderX

Research Question 3: Variability of Domains and Dimensions Among Technologies

Our thematic analysis revealed a significant degree of variability and heterogeneity in the number and type of domains and dimensions considered by the methodological frameworks.

In terms of numbers, the variability was quite pronounced when we compared frameworks addressing different types of dHTs. For instance, the thematic analysis of frameworks for assessing telemedicine only identified 9 (75%) domains and 6 (15.8%) dimensions; instead, in frameworks for assessing mHealth, we identified 10 (83.3%) domains, 20 (52.6%) dimensions, and 6 (54.5%) subdimensions, and in frameworks for assessing AI, we identified 8 (66.7%) different domains, 7 (18.4%) different dimensions, and 6 (54.5%) subdimensions.

In terms of the types of domains considered, certain dimensions and domains were identified as more distinctive for one kind of dHT than for another. For instance, clinical efficacy and effectiveness, technical safety, economic evaluation, and user experience were relevant for the evaluation of models of nonpresential health care and mHealth but not for AI. In contrast, there were specific dimensions and domains of mHealth that were not considered in the evaluation of non–face-to-face health care or AI, such as postmarketing monitoring, scientific basis, technical evaluation and validation, user control and self-determination, accessibility, content and adequacy of information, and data interoperability and integration. Finally, specific methodological frameworks for the evaluation of AI included dimensions such as technical aspects, adoption, use, integration, generalizability, reproducibility, and interpretability, which were not considered in the evaluation of telemedicine or mHealth. In conclusion, greater clarity and structuring in the presentation of these ideas are required to facilitate their understanding and assimilation.

Proposal for Domains, Dimensions, and Subdimensions for dHTA

These findings led to the development of a proposed methodological framework for dHTA, which comprises domains, dimensions, and subdimensions. These evaluation items were established objectively based on thematically analyzed evidence, without incorporating the researcher's perspective. Consequently, the proposal for domains, dimensions, and subdimensions emerged from the literature and represents the entirety of identified evaluation domains, dimensions, and subdimensions (n=61). Figure 2 presents a visual representation of the proposed framework comprising 12 domains, 38 dimensions, and their corresponding 11 subdimensions. Notably, the figure highlights certain domains, dimensions, and subdimensions that are particularly relevant to the evaluation of non–face-to-face care models, mHealth, and AI according to the evidence.

Figure 2. Proposed methodological framework for dHTA. aDimension identified as especially relevant for non–face-to-face care models; bdimension identified as especially relevant for mHealth; cdimension identified as especially relevant for AI; dHTA: digital health technology assessment. A higher-resolution version of this image is available as Multimedia Appendix 6.

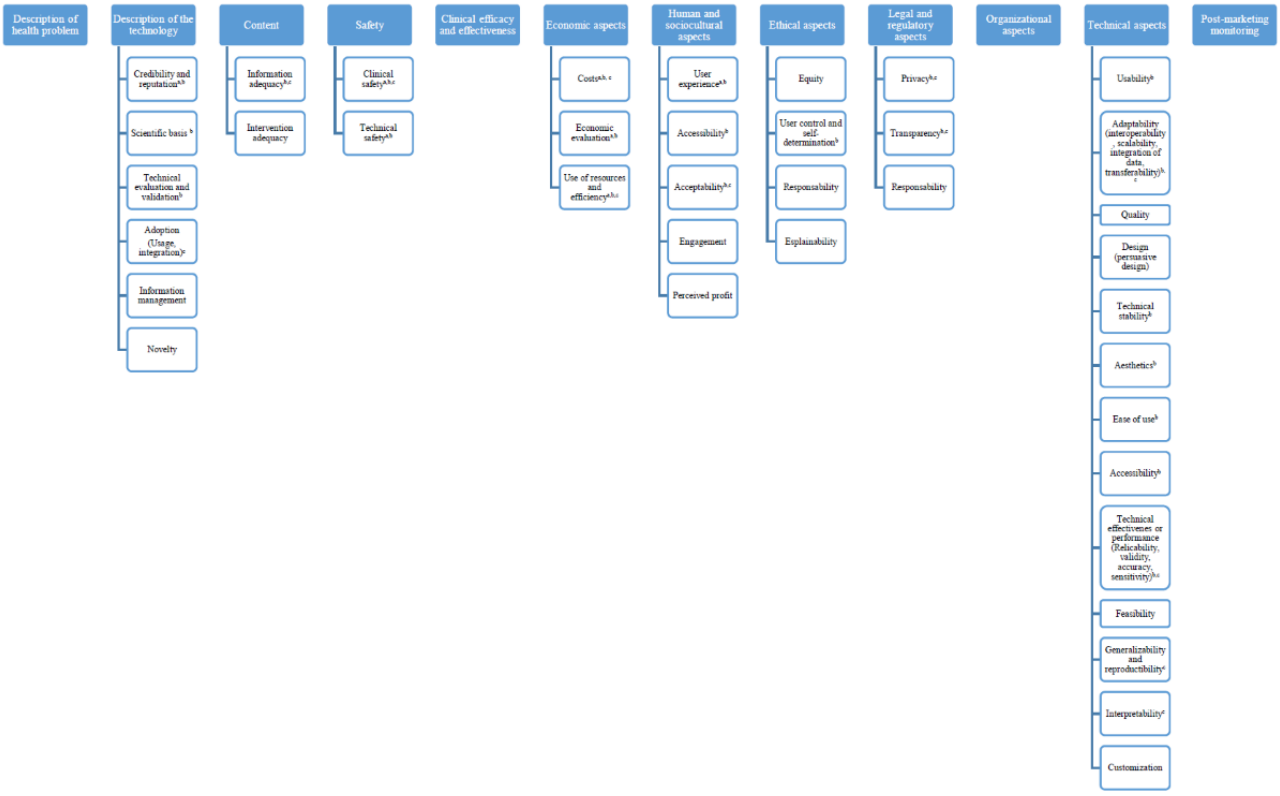

XSL•FO
RenderX

Discussion

Principal Findings

In recent years, the interest in digital health has increased significantly, giving rise to a myriad of available technologies. This has brought about a profound transformation in health care systems, fundamentally changing the provision and consumption of health care services [9]. However, despite these advancements, the shift toward digital health has been accompanied by challenges. One such challenge is the emergence of a plethora of short-lived implementations and an overwhelming diversity of digital tools, which has created a need for careful evaluation and analysis of the benefits and drawbacks of these technologies [4].

In this context, our ScR aimed to identify the methodological frameworks used worldwide for the assessment of dHTs; determine what domains are considered; and generate, through a thematic analysis, a proposal for a methodological framework based on the most frequently described domains in the literature.

Throughout the ScR, we identified a total of 95 methodological frameworks and tools, of which 19 [16,17,33-47] were directly identified through a systematic search and 75 were indirectly identified through 4 SRs [49-52]. The difference in the number of methodological frameworks identified through the ScR and the 4 evidence synthesis documents [49-52] is attributed to the inclusion of keywords related to the concept of HTA in the search syntax, the exclusion of methodological frameworks published prior to 2011 during the screening process, and the differences in perspectives used for the development of this paper compared to the 4 evidence synthesis documents mentioned earlier. In this sense, these 4 documents [49-52] have analyzed methodological frameworks and tools aimed at evaluating digital health that have not been developed from an HTA perspective despite the authors analyzing them as such. For example, von Huben et al. [51] included in their analysis the Consolidated Standards of Reporting Trials (CONSORT)-EHEALTH tool [97], which aims to describe the information that should be reported in papers and reports that focus on evaluating web- and mHealth-based interventions; Koladas et al [49] included the mobile health evidence reporting and assessment (mERA) checklist [73], which aims to determine the information that should be reported in trials evaluating mHealth solutions; and the European mHealth Hub document [52] includes the Isys Score, which is for cataloguing apps for smartphones.

However, as detailed in the *Results* section, some of the methodological frameworks identified through the ScR were characterized by the authors themselves as being specific for evaluating certain types of dHTs (eg, non–face-to-face care models, mHealth), presenting certain differences according to each typology. It is important to note that the differentiation among various types of dHTs, as described throughout this paper and commonly used in the field of digital health, cannot always be made in a precise and exclusive manner [136]. This is because a technology often can be classified in more than 1 category. For instance, an mHealth solution may use AI algorithms, while simultaneously being integrated into a non–face-to-face care model [137]. In this context, future research should consider using alternative taxonomies or classification methods that are based on the intended purpose of the technology, such as those proposed by NICE in the updated version of the Evidence Standards Framework [18] or the new digital health interventions system classification put forward by WHO [138].

After conducting a thematic analysis of the 26 included studies, we observed that various methodological frameworks include a set of evaluation items, referred to as domains, dimensions, or criteria. These items primarily focus on the safety; effectiveness; technical aspects; economic impact; and ethical, legal, and social consequences of dHTs. However, there is significant heterogeneity among these frameworks in terms of the way they refer to the evaluation items, the quantity and depth of their description, the degree of granularity, and the proposed evaluation methods, especially when comparing frameworks that focus on different types of dHTs. Despite this heterogeneity, most methodological frameworks consider evaluation items related to the 9 domains described by the HTA Core Model of EUnetHTA, while some frameworks propose additional evaluation elements, such as usability [16,44,45,47,49,56], privacy [39-41,44,52,55], and technical stability [16,38,47,49,52] among others. These findings are consistent with earlier research [50,51].

In addition, through the thematic analysis, the heterogeneity identified among the different methodological frameworks included in this ScR was reduced to a total of 61 analytical themes related to various evaluation elements that were arranged in a 3-level vertical hierarchy based on the evidence: level 1 (12 domains), level 2 (38 dimensions), and level 3 (11 subdimensions). At this point, it is pertinent to note that although from the researchers' perspective, some dimensions could have been classified under different domains (eg, responsibility under ethical aspects) or seen as essential for other kinds of dHTs, an effort was made to maintain the highest degree of objectivity possible. It is for this reason that privacy issues were not described as essential for non–face-to-face care models and why the dimension of accessibility was categorized within the domains of human and sociocultural aspects and technical aspects. This categorization was made because some of the methodological frameworks analyzed associated it with sociocultural elements (eg, evaluating whether users with functional diversity can access the technology and have sufficient ability to use it as expected), while others linked it to technical elements (eg, adequacy of the elements, options, or accessibility functionalities that the system incorporates according to the target audience) [16,52].

The ScR and thematic analysis conducted in this study led to a proposal for a methodological framework for dHTA. This framework was further developed using additional methodologies, such as consensus workshops by the Agency for Health Quality and Assessment of Catalonia (AQuAS), in collaboration with all agencies of RedETS, commissioned by the Ministry of Health of Spain. The final framework is a specific methodological tool for the assessment of dHTs, aimed at describing the domains and dimensions to be considered in dHTA and defining the evidence standards that such

XSL•FO
RenderX

technologies must meet based on their associated risk level. The proposed methodological framework enables the assessment of a wide range of dHTs, mainly those classified as medical devices according to the Regulation (EU) 2017/745 for medical devices [139] and Regulation (EU) 2017/746 for in vitro diagnostic medical devices, although it can be adapted to assess dHTs not classified as medical devices [140]. Unlike existing frameworks, it establishes a clear link between the identified domains and dimensions and the evidence standards required for dHTs to meet. This approach will enhance the transparency and consistency of dHTAs and support evidence-based decision-making. The final document was published from November 2023 onward and is available on the RedETS website as well as on the main web page of AQuAS in the Spanish language [141]. From the first week of February, the respective websites have hosted an English version of this document [141], which also is accessible in the INAHTA database. In addition, the Spanish and English versions of the document will be periodically reviewed and, if necessary, adapted to align with emerging technologies and changes in legislation.

Limitations

Although this ScR was conducted in accordance with the PRISMA-ScR guidelines (Multimedia Appendix 1) and following the recommendations of Peters et al [22] and Pollock et al [23], there were some limitations. First, the search incorporated a block of keywords related to the concept of HTA (see Multimedia Appendix 1) due to the perspective of our ScR, which may have limited the retrieval of some studies to meet the study objective. However, this limitation was compensated for by the analysis of the 3 SRs and the report of the working group on guidelines for the evaluation of mHealth solutions of the European mHealth Hub. Second, much of the literature related to HTA is gray literature and only published on the websites of the authoring agencies. Despite efforts to address this limitation through expert input and a comprehensive search of the websites of the world's leading agencies, it is possible that certain studies were not identified. Third, the quality and limitations of the analysis conducted by the authors of methodological frameworks and tools included in SRs may have had an impact on the indirect thematic analysis. Therefore, it is possible that some data could have been omitted or not considered during this process. Fourth, the focus on dHTs encompassed within the 3 previously mentioned categories (mHealth, non–face-to-face care models, and medical devices that integrate AI) may have influenced the outcomes of the thematic analysis conducted. Fifth, only methodological frameworks written in Catalan, Spanish, and English were included.

Comparison With Prior Work

To the best of our knowledge, this is the first ScR to examine the methodological frameworks for dHTA, followed by a thematic analysis with the aim of proposing a new comprehensive framework that incorporates the existing literature in an objective manner and enables the assessment of various technologies included under the concept of digital health. In this sense, existing SRs and other evidence synthesis documents have only analyzed the literature and reported the results in a descriptive manner [36,48,49,51,56,125,126]. Furthermore, this ScR also considered, in addition to scientific literature, gray literature identified by searching the websites of the agencies, thus covering some limitations of previous reviews [50]. Moreover, this review was carried out from the perspective of HTA, addressing a clear need expressed by HTA agencies [16].

Future research should aim to identify what domains and dimensions are relevant at the different stages of the technology life cycle, to establish or develop a standardized set of outcomes for assessing or reporting each domain, and to evaluate the effectiveness and usefulness of the existing methodological frameworks for the different intended users [50,142]. Moreover, future research should aim to determine the specific evaluation criteria that ought to be considered based on the level of risk associated with different types of technologies [51].

Conclusion

Our ScR revealed a total of 102 methodological frameworks and tools designed for evaluating dHTs, with 19 being directly identified through a systematic search and 83 through 4 evidence synthesis documents. Only 19 of all the identified frameworks were developed from the perspective of HTA. These frameworks vary in assessment items, structure, and specificity, and their proven usefulness in practice is scarce.

The thematic analysis of the 26 studies that met the inclusion criteria led to the identification and definition of 12 domains, 38 dimensions, and 11 subdimensions that should be considered when evaluating dHTs. Building on our results, a methodological framework for dHTA was proposed.

Acknowledgments

We acknowledge Benigno Rosón Calvo (Servicio Gallego de Salud [SERGAS]), Carme Carrion (Universitat Oberta de Catalunya [UOC]), Carlos A Molina Carrón (Dirección General de Salud Digital y Sistemas de Información para el SNS. Ministerio de Sanidad, Gobierno de España), Carme Pratdepadua (Fundació Tic Salut i Social [FTSS]), Celia Muñoz (Instituto Aragonés de Ciencias de la Salud [IACS]), David Pijoan (Biocat, BioRegió de Catalunya), Felip Miralles (Eurecat – Centre Tecnològic de Catalunya), Iñaki Guiterrez Ibarluzea (Osasun Teknologien Ebaluazioko Zerbitzua [Osteba]), Janet Puñal Riobóo (Unidad de Asesoramiento Científico-técnico [avalia-t], Agencia Gallega para la Gestión del Conocimiento en Salud [ACIS]), Jordi Piera-Jiménez (Àrea de Sistemes d'Informació del Servei Català de la Salut [CatSalut]), Juan Antonio Blasco (Evaluación de Tecnologías Sanitarias de Andalucía [AETSA]), Liliana Arroyo Moliner (Direcció General de Societat Digital, Departament d'Empresa i Treball de la Generalitat de Catalunya), Lilisbeth Perestelo-Perez (Servicio de Evaluación del Servicio Canario de la Salud [SESCS]), Lucía Prieto Remón (IACS), Marifé Lapeña (Dirección General de Salud Digital y Sistemas de Información

XSL•FO
RenderX

para el SNS. Ministerio de Sanidad, Gobierno de España), Mario Cárdaba (Insituto de Salud Carlos III [ISCIII]), Montserrat Daban (Biocat, BioRegió de Catalunya), Montserrat Moharra Frances (Agència de Qualitat i Avaluació Sanitàries de Catalunya), and Oscar Solans (CatSalut) for reviewing the protocol of this scoping review (ScR) and the ScR.

This research was framed within the budget of the work plan of the Spanish Network of Health Technology Assessment Agencies, commissioned by the General Directorate of Common Portfolio of Services of the National Health System and Pharmacy.

Authors' Contributions

JS-F and CM-P were responsible for conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, and visualization. RP-P handled conceptualization, methodology, formal analysis, investigation, resources, and writing—original draft. RMV-H handled conceptualization, writing—review and editing, supervision, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews (PRISMA-ScR) checklist [21].
[DOCX File , 25 KB - jmir_v26i1e48694_app1.docx]

Multimedia Appendix 2
Search strategies for each database.
[DOCX File , 34 KB - jmir_v26i1e48694_app2.docx]

Multimedia Appendix 3
References excluded at the full-text screening stage.
[DOCX File , 20 KB - jmir_v26i1e48694_app3.docx]

Multimedia Appendix 4
Methodological frameworks included in systematic reviews.
[DOCX File , 41 KB - jmir_v26i1e48694_app4.docx]

Multimedia Appendix 5
Network of relationships among the codes.
[PNG File , 399 KB - jmir_v26i1e48694_app5.png]

Multimedia Appendix 6
High-resolution image of Figure 2.
[PNG File , 121 KB - jmir_v26i1e48694_app6.png]

References

1. Avanzas P, Pascual I, Moris C. The great challenge of the public health system in Spain. J Thorac Dis 2017 May;9(Suppl 6):S430-S433 [FREE Full text] [doi: 10.21037/jtd.2017.04.59] [Medline: 28616336]
2. Grubanov BS, Ghio D, Goujon A, Kalantaryan S, Belmonte M, Scipioni M. Health and long-term care workforce: demographic challenges and the potential contribution of migration and digital technology. Luxembourg: European Commission; 2021:2045-2322.
3. World Health Organization (WHO). Health 2020: a European policy framework and strategy for the 21st century. Denmark: WHO; 2013.
4. World Health Organization (WHO). WHO guideline: recommendations on digital interventions for health system strengthening. Geneva: WHO; Jun 6, 2019.
5. Scholz N. Addressing health inequalities in the European Union. Luxembourg: European Parliamentary Research Service; 2020.
6. Kickbusch I, Agrawal A, Jack A, Lee N, Horton R. Governing health futures 2030: growing up in a digital world—a joint The Lancet and Financial Times Commission. Lancet 2019 Oct;394(10206):1309. [doi: 10.1016/s0140-6736(19)32181-6]
7. Reeves JJ, Ayers JW, Longhurst CA. Telehealth in the COVID-19 era: a balancing act to avoid harm. J Med Internet Res 2021 Feb 01;23(2):e24785 [FREE Full text] [doi: 10.2196/24785] [Medline: 33477104]

XSL•FO
RenderX

8. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
9. European Union. State of health in the EU: companion report 2021. Luxembourg: European Union; 2022.
10. LATITUD: Anàlisi comparativa de models d'atenció no presencial en l'àmbit de la salut. TIC Salut Social. 2020. URL: https://ticsalutsocial.cat/wp-content/uploads/2021/07/latitud.pdf [accessed 2024-03-28]
11. Global strategy on digital health 2020-2025. World Health Organization. 2021 Aug 18. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-03-28]
12. Ming J, He Y, Yang Y, Hu M, Zhao X, Liu J, et al. Health technology assessment of medical devices: current landscape, challenges, and a way forward. Cost Eff Resour Alloc 2022 Oct 05;20(1):54 [FREE Full text] [doi: 10.1186/s12962-022-00389-6] [Medline: 36199144]
13. O'Rourke B, Oortwijn W, Schuller T. The new definition of health technology assessment: a milestone in international collaboration. Int J Technol Assess Health Care 2020 May 13;36(3):187-190. [doi: 10.1017/s0266462320000215]
14. EUnetHTA Joint Action 2 Work Package 8. HTA Core Model ® version 3.0 2016. EUnetHTA. 2016. URL: https://www.eunethta.eu/wp-content/uploads/2018/03/HTACoreModel3.0-1.pdf [accessed 2024-03-28]
15. Puñal-Riobóo J, Baños, Varela LL, Castillo MM, Atienza MG, Ubago PR. Guía para la elaboración y adaptación de informes rápidos. Santiago de Compostela, Madrid: Agencia Gallega para la Gestión del Conocimientoto en Salud. Unidad de Asesoramiento Científico-técnico, avalia-t. Ministerio de Sanidad, Servicios Sociales e Igualdad; 2016.
16. Haverinen J, Keränen N, Falkenbach P, Maijala A, Kolehmainen T, Reponen J. Digi-HTA: health technology assessment framework for digital healthcare services. FinJeHeW 2019 Nov 02;11(4):326-341. [doi: 10.23996/fjhw.82538]
17. Hussain MS, Silvera-Tawil D, Farr-Wharton G. Technology assessment framework for precision health applications. Int J Technol Assess Health Care 2021 May 26;37(1):e67. [doi: 10.1017/s0266462321000350]
18. National Institute for Health and Care Excellence (NICE). Evidence standards framework (ESF) for digital health technologies. Contract no.: 30 de septiembre. London: NICE; 2022.
19. Evidence standards framework (ESF) for digital health technologies. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [accessed 2024-03-28]
20. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM. Methodological frameworks and dimensions to be taken into consideration in digital health technology assessment: protocol for a scoping review. JMIR Res Protoc 2022 Oct 11;11(10):e39905 [FREE Full text] [doi: 10.2196/39905] [Medline: 36222788]
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Peters M, Marnie C, Tricco A, Pollock D, Munn Z, Alexander L. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126. [doi: 10.11124/jbies-20-00167]
23. Pollock D, Davies EL, Peters MDJ, Tricco AC, Alexander L, McInerney P, et al. Undertaking a scoping review: a practical guide for nursing and midwifery students, clinicians, researchers, and academics. J Adv Nurs 2021 Apr 04;77(4):2102-2113 [FREE Full text] [doi: 10.1111/jan.14743] [Medline: 33543511]
24. Ayiku L, Hudson T, Glover S, Walsh N, Adams R, Deane J, et al. The NICE MEDLINE and Embase (Ovid) health apps search filters: development of validated filters to retrieve evidence about health apps. Int J Technol Assess Health Care 2020 Oct 27;37(1):e16. [doi: 10.1017/s026646232000080x]
25. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
26. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 06;6(1):245 [FREE Full text] [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]
27. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015;13(3):179-187. [doi: 10.1097/xeb.0000000000000062]
28. Systematic reviews - research guide. Defining your review question. Murdoch University. 2023. URL: https://libguides.murdoch.edu.au/systematic [accessed 2024-03-28]
29. Scientific Software Development GmbH. Atlas.ti qualitative data analysis. 22.0 ed. Berlin: Scientific Software Development GmbH; 2021.
30. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane handbook for systematic reviews of interventions version 6.3. London: Cochrane; 2022.
31. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8(1):45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
32. Page M, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021 Mar;18(3):e1003583 [FREE Full text] [doi: 10.1371/journal.pmed.1003583] [Medline: 33780438]

XSL•FO
RenderX

33. Kidholm K, Ekeland AG, Jensen LK, Rasmussen J, Pedersen CD, Bowes A, et al. A model for assessment of telemedicine applications: MAST. Int J Technol Assess Health Care 2012 Jan 23;28(1):44-51. [doi: 10.1017/s0266462311000638]
34. Nepal S, Li J, Jang-Jaccard J, Alem L. A framework for telehealth program evaluation. Telemed J E Health 2014 Apr;20(4):393-404. [doi: 10.1089/tmj.2013.0093] [Medline: 24621384]
35. Measuring the impact of telehealth and telecare: SCTT toolkit. Glasgow: SCTT; 2013.
36. Caulfield B, Reginatto B, Slevin P. Not all sensors are created equal: a framework for evaluating human performance measurement technologies. NPJ Digit Med 2019 Feb 14;2(1):7 [FREE Full text] [doi: 10.1038/s41746-019-0082-4] [Medline: 31304357]
37. Lewis TL, Wyatt JC. mHealth and mobile medical apps: a framework to assess risk and promote safer use. J Med Internet Res 2014 Sep 15;16(9):e210 [FREE Full text] [doi: 10.2196/jmir.3133] [Medline: 25223398]
38. Moshi MR, Tooher R, Merlin T. Development of a health technology assessment module for evaluating mobile medical applications. Int J Technol Assess Health Care 2020 May 18;36(3):252-261. [doi: 10.1017/s0266462320000288]
39. Puigdomènech E, Poses-Ferrer E, Espallargues CM, Blasco AJ, Varela LL, Paz VL. Evaluación de tecnología basada en mSalud para aplicaciones móviles. Madrid: Ministerio de Sanidad; 2020.
40. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digit Health 2019 Jun;1(2):e52-e54. [doi: 10.1016/s2589-7500(19)30013-5]
41. Reddy S, Rogers W, Makinen V, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021 Oct 12;28(1):e100444 [FREE Full text] [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
42. Khoja S, Durrani H, Scott RE, Sajwani A, Piryani U. Conceptual framework for development of comprehensive e-health evaluation tool. Telemed J E Health 2013 Jan;19(1):48-53. [doi: 10.1089/tmj.2012.0073] [Medline: 22957502]
43. Sockolow P, Bowles K, Rogers M. Health information technology evaluation framework (HITREF) comprehensiveness as assessed in electronic point-of-care documentation systems evaluations. Amsterdam: IOS Press; 2015.
44. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. it - Inf Technol 2019 Nov;61(5):253-263. [doi: 10.1515/itit-2019-0019]
45. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019 May 13;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
46. Milosevic Z. Ethics in digital health: a deontic accountability framework. 2019 Presented at: IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC); October 28-31, 2019; Paris, France. [doi: 10.1109/edoc.2019.00022]
47. World Health Organization (WHO). Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. Geneva: WHO; 2016.
48. Enam A, Torres-Bonilla J, Eriksson H. Evidence-based evaluation of eHealth interventions: systematic literature review. J Med Internet Res 2018 Nov 23;20(11):e10971 [FREE Full text] [doi: 10.2196/10971] [Medline: 30470678]
49. Kolasa K, Kozinski G. How to value digital health interventions? A systematic literature review. Int J Environ Res Public Health 2020 Mar 23;17(6):2119 [FREE Full text] [doi: 10.3390/ijerph17062119] [Medline: 32209988]
50. Vis C, Bührmann L, Riper H, Ossebaard HC. Health technology assessment frameworks for eHealth: a systematic review. Int J Technol Assess Health Care 2020 Apr 16;36(3):204-216. [doi: 10.1017/s026646232000015x]
51. von Huben A, Howell M, Howard K, Carrello J, Norris S. Health technology assessment for digital technologies that manage chronic disease: a systematic review. Int J Technol Assess Health Care 2021 May 26;37(1):e66. [doi: 10.1017/s0266462321000362]
52. Report of the Working Group on mhealth assessment guidelines. European Commission. 2021. URL: https://digital-strategy.ec.europa.eu/en/library/report-working-group-mhealth-assessment-guidelines [accessed 2024-03-28]
53. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mhealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
54. Alami H, Lehoux P, Auclair Y, de Guise M, Gagnon M, Shaw J, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 07;22(7):e17707 [FREE Full text] [doi: 10.2196/17707] [Medline: 32406850]
55. Vokinger KN, Nittas V, Witt CM, Fabrikant SI, von Wyl V. Digital health and the COVID-19 epidemic: an assessment framework for apps from an epidemiological and legal perspective. Swiss Med Wkly 2020 May 04;150(1920):w20282 [FREE Full text] [doi: 10.4414/smw.2020.20282] [Medline: 32418194]
56. Baumel A, Muench F. Heuristic evaluation of ehealth interventions: establishing standards that relate to the therapeutic process perspective. JMIR Ment Health 2016 Jan 13;3(1):e5 [FREE Full text] [doi: 10.2196/mental.4563] [Medline: 26764209]
57. Alfonzo A, Huerta M, Wong S, Passariello G, Diaz M, La CA, et al. Design of a methodology for assessing an electrocardiographic telemonitoring system. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:3729-3732. [doi: 10.1109/iembs.2007.4353142]

XSL•FO
RenderX

58. Bashshur R, Shannon G, Sapci H. Telemedicine evaluation. Telemed J E Health 2005 Jun;11(3):296-316 [FREE Full text] [doi: 10.1089/tmj.2005.11.296] [Medline: 16035928]
59. Beintner I, Vollert B, Zarski A, Bolinski F, Musiat P, Görlich D, et al. Adherence reporting in randomized controlled trials examining manualized multisession online interventions: systematic review of practices and proposal for reporting standards. J Med Internet Res 2019 Aug 15;21(8):e14181 [FREE Full text] [doi: 10.2196/14181] [Medline: 31414664]
60. Brear M. Evaluating telemedicine: lessons and challenges. Health Inf Manag 2006 Jul 21;35(2):23-31. [doi: 10.1177/183335830603500206] [Medline: 18209220]
61. DeChant HK, Tohme WG, Mun SK, Hayes WS, Schulman KA. Health systems evaluation of telemedicine: a staged approach. Telemed J 1996 Jan;2(4):303-312. [doi: 10.1089/tmj.1.1996.2.303] [Medline: 10165367]
62. Giansanti D, Morelli S, Macellari V. Telemedicine technology assessment part II: tools for a quality control system. Telemed J E Health 2007 Apr;13(2):130-140. [doi: 10.1089/tmj.2006.9988] [Medline: 17489699]
63. Giansanti D, Morelli S, Macellari V. Telemedicine technology assessment part I: setup and validation of a quality control system. Telemed J E Health 2007 Apr;13(2):118-129. [doi: 10.1089/tmj.2008.9963] [Medline: 17489698]
64. Grigsby J, Brega AG, Devore PA. The evaluation of telemedicine and health services research. Telemed J E Health 2005 Jun;11(3):317-328. [doi: 10.1089/tmj.2005.11.317] [Medline: 16035929]
65. Hailey D, Jacobs P, Simpson J, Doze S. An assessment framework for telemedicine applications. J Telemed Telecare 1999 Jun 23;5(3):162-170. [doi: 10.1258/1357633991933576] [Medline: 10628031]
66. Ohinmaa A, Hailey D, Roine R. Elements for assessment of telemedicine applications. Int J Technol Assess Health Care 2001 Jun 30;17(2):190-202. [doi: 10.1017/s0266462300105057] [Medline: 11446131]
67. Rajan B, Tezcan T, Seidmann A. Service systems with heterogeneous customers: investigating the effect of telemedicine on chronic care. Manag Sci 2019 Mar;65(3):1236-1267. [doi: 10.1287/mnsc.2017.2979]
68. Sisk JE, Sanders JH. A proposed framework for economic evaluation of telemedicine. Telemed J 1998 Jan;4(1):31-37. [doi: 10.1089/tmj.1.1998.4.31] [Medline: 9599071]
69. Zissman K, Lejbkowicz I, Miller A. Telemedicine for multiple sclerosis patients: assessment using Health Value Compass. Mult Scler 2012 Apr 30;18(4):472-480. [doi: 10.1177/1352458511421918] [Medline: 21965420]
70. Grustam AS, Vrijhoef HJM, Koymans R, Hukal P, Severens JL. Assessment of a business-to-consumer (B2C) model for telemonitoring patients with chronic heart failure (CHF). BMC Med Inform Decis Mak 2017 Oct 11;17(1):145 [FREE Full text] [doi: 10.1186/s12911-017-0541-2] [Medline: 29020993]
71. Hebert M. Telehealth success: evaluation framework development. Stud Health Technol Inform 2001;84(Pt 2):1145-1149. [Medline: 11604908]
72. Rojahn K, Laplante S, Sloand J, Main C, Ibrahim A, Wild J, et al. Remote monitoring of chronic diseases: a landscape assessment of policies in four European countries. PLoS One 2016 May 19;11(5):e0155738 [FREE Full text] [doi: 10.1371/journal.pone.0155738] [Medline: 27195764]
73. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]
74. Safety and quality strategy in mobile health apps. Complete list of recommendations on design, use and assessment of health apps. Agencia de Calidad Sanitaria de Andalucía. 2012. URL: http://www.calidadappsalud.com/en/listado-completo-recomendaciones-app-salud/ [accessed 2023-07-17]
75. The app evaluation model. American Psychiatric Association Initiative. 2022. URL: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/the-app-evaluation-model [accessed 2024-03-28]
76. Gdd AppStore. Association of Regional Public Health Services (GGD) and Regional Medical Emergency Preparedness and Planning (GHOR). 2016. URL: https://tinyurl.com/58th4p4w [accessed 2023-07-18]
77. AppQ: quality criteria core set for more quality transparency in digital health applications. Bertelsmann Stiftung. 2019 Oct 29. URL: https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/appq [accessed 2024-03-28]
78. Bradway M, Carrion C, Vallespin B, Saadatfard O, Puigdomènech E, Espallargues M, et al. mHealth assessment: conceptualization of a global framework. JMIR Mhealth Uhealth 2017 May 02;5(5):e60 [FREE Full text] [doi: 10.2196/mhealth.7291] [Medline: 28465282]
79. PAS 277:2015 - health and wellness apps. Quality criteria across the life cycle. British Standards Institution. URL: https://knowledge.bsigroup.com/products/health-and-wellness-apps-quality-criteria-across-the-life-cycle-code-of-practice/standard [accessed 2023-07-18]
80. MindApps. Centre for Telepsychiatry in the Region of Southern Denmark. 2017. URL: https://mindapps.dk/ [accessed 2023-07-18]
81. Dick S, O'Connor Y, Thompson MJ, O'Donoghue J, Hardy V, Wu TJ, et al. Considerations for improved mobile health evaluation: retrospective qualitative investigation. JMIR Mhealth Uhealth 2020 Jan 22;8(1):e12424 [FREE Full text] [doi: 10.2196/12424] [Medline: 32012085]
82. Federal Institute for Drugs and Medical Devices (BfArM). The fast-track process for digital health applications (diga) according to section 139e sgb v. A guide for manufacturers, service providers and users. Bonn, Germany: Federal Institute for Drugs and Medical Devices; 2020.

XSL•FO
RenderX

83. Gorski I, Bram JT, Sutermaster S, Eckman M, Mehta K. Value propositions of mHealth projects. J Med Eng Technol 2016 Aug 12;40(7-8):400-421. [doi: 10.1080/03091902.2016.1213907] [Medline: 27687907]
84. Hogaboam L, Daim T. Technology adoption potential of medical devices: the case of wearable sensor products for pervasive care in neurosurgery and orthopedics. Health Policy Technol 2018 Dec;7(4):409-419. [doi: 10.1016/j.hlpt.2018.10.011]
85. Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw Open 2019 Apr 05;2(4):e192542 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2542] [Medline: 31002321]
86. Maar MA, Yeates K, Perkins N, Boesch L, Hua-Stewart D, Liu P, et al. A framework for the study of complex mhealth interventions in diverse cultural settings. JMIR Mhealth Uhealth 2017 Apr 20;5(4):e47 [FREE Full text] [doi: 10.2196/mhealth.7044] [Medline: 28428165]
87. McMillan B, Hickey E, Patel MG, Mitchell C. Quality assessment of a sample of mobile app-based health behavior change interventions using a tool based on the National Institute of Health and Care Excellence behavior change guidance. Patient Educ Couns 2016 Mar;99(3):429-435 [FREE Full text] [doi: 10.1016/j.pec.2015.10.023] [Medline: 26607787]
88. Health and welness apps: new international guidelines to help to sort the best form the rest. NEN. URL: https://www.nen.nl/en/health-and-welness-apps [accessed 2024-03-28]
89. Continua design guidelines. Personal Connected Health Alliance. 2019. URL: https://www.pchalliance.org/continua-design-guidelines [accessed 2023-07-18]
90. Philpott D, Guergachi A, Keshavjee K. Design and validation of a platform to evaluate mhealth apps. Stud Health Technol Inform 2017;235:3-7. [Medline: 28423744]
91. Ruck A, Wagner BS, Lowe C. Second draft of guidelines. EU guidelines on assessment of the reliability of mobile health applications. European Commission, Directorate-General of Communications Networks, Content & Technology: Luxembourg; 2016.
92. Sax M, Helberger N, Bol N. Health as a means towards profitable ends: mhealth apps, user autonomy, and unfair commercial practices. J Consum Policy 2018 May 22;41(2):103-134. [doi: 10.1007/s10603-018-9374-3]
93. Wyatt JC. How can clinicians, specialty societies and others evaluate and improve the quality of apps for patient use? BMC Med 2018 Dec 03;16(1):225 [FREE Full text] [doi: 10.1186/s12916-018-1211-7] [Medline: 30501638]
94. IRBs could address ethical issues related to tracking devices: mobile devices raise new concerns. IRB Advisor. 2017 Nov 1. URL: https://www.reliasmedia.com/articles/141589-irbs-could-address-ethical-issues-related-to-tracking-devices [accessed 2024-03-28]
95. App check. Zentrum für Telematik und Telemedizin. URL: https://ztg-nrw.de/ [accessed 2024-03-28]
96. Bergmo TS. How to measure costs and benefits of ehealth interventions: an overview of methods and frameworks. J Med Internet Res 2015 Nov 09;17(11):e254 [FREE Full text] [doi: 10.2196/jmir.4521] [Medline: 26552360]
97. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
98. Shaw NT. 'CHEATS': a generic information communication technology (ICT) evaluation framework. Comput Biol Med 2002 May;32(3):209-220. [doi: 10.1016/s0010-4825(02)00016-1] [Medline: 11922936]
99. Brown M, Shaw N. Evaluation practices of a major Canadian telehealth provider: lessons and future directions for the field. Telemed J E Health 2008 Oct;14(8):769-774. [doi: 10.1089/tmj.2007.0126] [Medline: 18954246]
100. Casper GR, Kenron DA. A framework for technology assessment: approaches for the selection of a home technology device. Clin Nurse Spec 2005;19(4):170-174. [doi: 10.1097/00002800-200507000-00004] [Medline: 16027542]
101. Sitting D, Kahol K, Singh H. Sociotechnical evaluation of the safety and effectiveness of point-of-care mobile computing devices: a case study conducted in India. Stud Health Technol Inform 2013;192:515-519. [doi: 10.1201/b16306-16]
102. Haute Autorité de Santé. Good practice guidelines on health apps and smart devices (mobile health or mhealth). Paris: Haute Autorité de Santé; Nov 7, 2016.
103. Health Information and Quality Authority. International review of consent models for the collection, use and sharing of health information. Dublin: Health Information and Quality Authority; 2020.
104. Jurkeviciute M. Planning of a holistic summative ehealth evaluation: the interplay between standards and reality PhD Thesis. Gotherburg: Chalmers University of Technology; 2018.
105. Vimarlund V, Davoody N, Koch S. Steps to consider for effective decision making when selecting and prioritizing eHealth services. Stud Health Technol Inform 2013;192:239-243. [Medline: 23920552]
106. Currie WL. TEMPEST: an integrative model for health technology assessment. Health Policy Technol 2012 Mar;1(1):35-49. [doi: 10.1016/j.hlpt.2012.01.004]
107. Eivazzadeh S, Anderberg P, Larsson TC, Fricker SA, Berglund J. Evaluating health information systems using ontologies. JMIR Med Inform 2016 Jun 16;4(2):e20 [FREE Full text] [doi: 10.2196/medinform.5185] [Medline: 27311735]
108. Our data-driven future in healthcare: people and partnerships at the heart of health related technologies. Academy of Medical Sciences. 2018 Nov. URL: https://acmedsci.ac.uk/file-download/74634438 [accessed 2024-03-28]
109. Australian Commission on Safety and Quality in Healthcare. National safety and quality digital mental health standards - consultation draft. Australia: Australian Commission on Safety and Quality in Healthcare; 2020.

XSL•FO
RenderX

110. A guide to good practice for digital and data-driven health technologies. London: Department of Health & Social Care; 2021.
111. Drury P, Roth S, Jones T, Stahl M, Medeiros D. Guidance for investing in digital health. In: Sustainable Development Working Papers. Mandaluyong, the Philippines: Asian Development Bank; May 2018.
112. European Commission. Synospis report. Consultation: transformation health and care in the digital single market. Luxembourg: European Commission; 2018.
113. Federal Ministry of Health. Regulation on the procedure and requirements for testing the eligibility for reimbursement of digital health applications in the statutory public health insurance (Digital Health Applications Ordinance - DiGAV). Alemania: Federal Ministry of Health; 2020.
114. Haute Autorité de Santé. Guide to the specific features of clinical evaluation of a connected medical device (CMD) in view of its application for reimbursement. Paris: Haute Autorité de Santé; 2019.
115. How we assess health apps and digital tools. NHS Digital. 2019. URL: https://digital.nhs.uk/services/nhs-apps-library/guidance-forhealth-app-developers-commissioners-and-assessors/how-we-assess-healthapps-and-digital-tools [accessed 2023-07-17]
116. Lennon MR, Bouamrane M, Devlin AM, O'Connor S, O'Donnell C, Chetty U, et al. Readiness for delivering digital health at scale: lessons from a longitudinal qualitative evaluation of a national digital health innovation program in the United Kingdom. J Med Internet Res 2017 Feb 16;19(2):e42 [FREE Full text] [doi: 10.2196/jmir.6900] [Medline: 28209558]
117. McNamee P, Murray E, Kelly MP, Bojke L, Chilcott J, Fischer A, et al. Designing and undertaking a health economics study of digital health interventions. Am J Prev Med 2016 Nov;51(5):852-860 [FREE Full text] [doi: 10.1016/j.amepre.2016.05.007] [Medline: 27745685]
118. Haute Autorité de Santé. Medical Device and Health Technology Evaluation Committee (CNEDiMTS*). Paris: Haute Autorité de Santé; 2019.
119. Draft guidelines for preparing assessment reports for the Medical Services Advisory Committee: draft version 4.0. Australian Government Department of Health and Aged Care. 2020 Aug. URL: https://consultations.health.gov.au/technology-assessment-access-division/msac-guidelines-review-consultation/user_uploads/draft-msac-guidelines---clean-version---28-august-2020-3.pdf [accessed 2024-03-28]
120. Haute Autorité de Santé. Methodological choices for the clinical development of medical devices. Paris: Haute Autorité de Santé; 2013.
121. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
122. Mohr DC, Schueller SM, Riley WT, Brown CH, Cuijpers P, Duan N, et al. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. J Med Internet Res 2015 Jul 08;17(7):e166 [FREE Full text] [doi: 10.2196/jmir.4391] [Medline: 26155878]
123. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
124. Steventon A, Grieve R, Bardsley M. An approach to assess generalizability in comparative effectiveness research: a case study of the whole systems demonstrator cluster randomized trial comparing telehealth with usual care for patients with chronic health conditions. Med Decis Making 2015 May 18;35(8):1023-1036. [doi: 10.1177/0272989x15585131]
125. D2.1 Knowledge Tool 1. Health apps assessment frameworks. European mHealth Hub. 2020. URL: https://mhealth-hub.org/download/d2-1-knowledge-tool-1-health-apps-assessment-frameworks [accessed 2024-03-28]
126. Moshi MR, Tooher R, Merlin T. Suitability of current evaluation frameworks for use in the health technology assessment of mobile medical applications: a systematic review. Int J Technol Assess Health Care 2018 Sep 11;34(5):464-475. [doi: 10.1017/s026646231800051x]
127. Stensgaard T, Sørensen T. Telemedicine in Greenland — the creation of an evaluation plan. J Telemed Telecare 2016 Jun 22;7(1_suppl):37-38. [doi: 10.1177/1357633x010070s115]
128. HL7 Consumer Mobile Health Application Functional Framework (cMHAFF). HL7 International. 2018. URL: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=476 [accessed 2024-03-28]
129. mHealth. Kantonen K-uKvBu. 2017. URL: https://www.e-health-suisse.ch/gemeinschaften-umsetzung/ehealth-aktivitaeten/mhealth.html [accessed 2023-07-18]
130. mHealthBelgium. mHealthBELGIUM. URL: https://mhealthbelgium.be/ [accessed 2024-03-28]
131. Mookherji S, Mehl G, Kaonga N, Mechael P. Unmet need: improving mhealth evaluation rigor to build the evidence base. J Health Commun 2015 Jun 04;20(10):1224-1229. [doi: 10.1080/10810730.2015.1018624] [Medline: 26042660]
132. MySNS Selecção. Serviços Partilhados do Minisério da Saúde. URL: https://mysns.min-saude.pt/mysns-seleccao-processo-de-avaliacao/ [accessed 2023-07-18]
133. Nielsen S, Rimpiläinen S. Report on international practice on digital apps. Glasgow: Digital Health & Care Institute; 2018.

XSL•FO
RenderX

134. Accreditation of Digital Health solutions is a fundamental foundation for their safe adoption, equipping healthcare providers and practitioners with access to health apps assured to your standards Internet. Organisation for the Review of Care and Health Applications (ORCHA). URL: https://orchahealth.com/services/ [accessed 2023-07-18]
135. mHealth. TIC Salut Social. 2022. URL: https://ticsalutsocial.cat/projecte/mhealth/ [accessed 2024-03-28]
136. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022 Jun 28;24(6):e31921 [FREE Full text] [doi: 10.2196/31921] [Medline: 35763320]
137. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MDP, WARIFA Consortium. Quality, usability, and effectiveness of mhealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 04;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]
138. World Health Organization (WHO). Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health, 2nd ed. Geneva: WHO; Oct 24, 2023.
139. European Parliament, Council of the European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. EUR-Lex. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745 [accessed 2024-03-28]
140. European Parliament, Council of the European Union. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU. EUR-Lex. 2017. URL: https://eur-lex.europa.eu/eli/reg/2017/746/oj [accessed 2024-03-28]
141. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo R. Marco de evaluación de tecnologías sanitarias: adaptación para la evaluación de tecnologías de salud digital. Madrid, Barcelona: Ministerio de Sanidad, Agència de Qualitat i Avaluació Sanitàries de Catalunya; 2023.
142. Benedetto V, Filipe L, Harris C, Spencer J, Hickson C, Clegg A. Analytical frameworks and outcome measures in economic evaluations of digital health interventions: a methodological systematic review. Med Decis Making 2022 Oct 19;43(1):125-138. [doi: 10.1177/0272989x221132741]

Abbreviations

AI: artificial intelligence
dHT: digital health technology
dHTA: digital health technology assessment
EUnetHTA: European Network for Health Technology Assessment
HTA: health technology assessment
INAHTA: International Network for Agencies for Health Technology Assessment
mHealth: mobile health
NICE: National Institute for Health and Care Excellence
PICo-D: Population, Phenomenon of interest, Context and Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews
RedETS: Spanish Network of Agencies for Assessing National Health System Technologies and Performance
ScR: scoping review
SNS: Spanish National Health System
SR: systematic review
SNS: Spanish National Health System
WHO: World Health Organization

Edited by T Leung; submitted 03.05.23; peer-reviewed by R Gorantla, KL Mauco, M Aymerich, J Haverinen, M Behzadifar; comments to author 10.11.23; revised version received 01.12.23; accepted 20.02.24; published 10.04.24.

Please cite as:
Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM
Methodological Frameworks and Dimensions to Be Considered in Digital Health Technology Assessment: Scoping Review and Thematic Analysis
J Med Internet Res 2024;26:e48694
URL: https://www.jmir.org/2024/1/e48694
doi:10.2196/48694
PMID:38598288

XSL•FO
RenderX

©Joan Segur-Ferrer, Carolina Moltó-Puigmartí, Roland Pastells-Peiró, Rosa Maria Vivanco-Hidalgo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Interventions for Recreational Cannabis Use Among Young Adults: Systematic Review, Meta-Analysis, and Behavior Change Technique Analysis of Randomized Controlled Studies

José Côté[1,2,3], RN, PhD; Gabrielle Chicoine[3,4], RN, PhD; Billy Vinette[1,3], RN, MSN; Patricia Auger[2,3], MSc; Geneviève Rouleau[3,5,6], RN, PhD; Guillaume Fontaine[7,8,9], RN, PhD; Didier Jutras-Aswad[2,10], MSc, MD

[1]Faculty of Nursing, Université de Montréal, Montreal, QC, Canada
[2]Research Centre of the Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada
[3]Research Chair in Innovative Nursing Practices, Montreal, QC, Canada
[4]Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada
[5]Department of Nursing, Université du Québec en Outaouais, Saint-Jérôme, QC, Canada
[6]Women's College Hospital Institute for Health System Solutions and Virtual Care, Women's College Hospital, Toronto, ON, Canada
[7]Ingram School of Nursing, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[8]Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Sir Mortimer B. Davis Jewish General Hospital, Montreal, QC, Canada
[9]Kirby Institute, University of New South Wales, Sydney, Australia
[10]Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

Corresponding Author:
José Côté, RN, PhD
Research Centre of the Centre Hospitalier de l'Université de Montréal
850 Saint-Denis
Montreal, QC, H2X 0A9
Canada
Phone: 1 514 890 8000
Email: jose.cote@umontreal.ca

Abstract

Background: The high prevalence of cannabis use among young adults poses substantial global health concerns due to the associated acute and long-term health and psychosocial risks. Digital modalities, including websites, digital platforms, and mobile apps, have emerged as promising tools to enhance the accessibility and availability of evidence-based interventions for young adults for cannabis use. However, existing reviews do not consider young adults specifically, combine cannabis-related outcomes with those of many other substances in their meta-analytical results, and do not solely target interventions for cannabis use.

Objective: We aimed to evaluate the effectiveness and active ingredients of digital interventions designed specifically for cannabis use among young adults living in the community.

Methods: We conducted a systematic search of 7 databases for empirical studies published between database inception and February 13, 2023, assessing the following outcomes: cannabis use (frequency, quantity, or both) and cannabis-related negative consequences. The reference lists of included studies were consulted, and forward citation searching was also conducted. We included randomized studies assessing web- or mobile-based interventions that included a comparator or control group. Studies were excluded if they targeted other substance use (eg, alcohol), did not report cannabis use separately as an outcome, did not include young adults (aged 16-35 y), had unpublished data, were delivered via teleconference through mobile phones and computers or in a hospital-based setting, or involved people with mental health disorders or substance use disorders or dependence. Data were independently extracted by 2 reviewers using a pilot-tested extraction form. Authors were contacted to clarify study details and obtain additional data. The characteristics of the included studies, study participants, digital interventions, and their comparators were summarized. Meta-analysis results were combined using a random-effects model and pooled as standardized mean differences.

Results: Of 6606 unique records, 19 (0.29%) were included (n=6710 participants). Half (9/19, 47%) of these articles reported an intervention effect on cannabis use frequency. The digital interventions included in the review were mostly web-based. A total of 184 behavior change techniques were identified across the interventions (range 5-19), and *feedback on behavior* was the most frequently used (17/19, 89%). Digital interventions for young adults reduced cannabis use frequency at the 3-month follow-up

XSL•FO
RenderX

compared to control conditions (including passive and active controls) by −6.79 days of use in the previous month (95% CI −9.59 to −4.00; *P*<.001).

Conclusions: Our results indicate the potential of digital interventions to reduce cannabis use in young adults but raise important questions about what optimal exposure dose could be more effective, both in terms of intervention duration and frequency. Further high-quality research is still needed to investigate the effects of digital interventions on cannabis use among young adults.

Trial Registration: PROSPERO CRD42020196959; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=196959

(J Med Internet Res 2024;26:e55031) doi:10.2196/55031

KEYWORDS

cannabis; young adults; digital intervention; systematic review; meta-analysis; behavior change technique analysis; mobile phone

Introduction

Cannabis Use Among Young Adults Is Recognized as a Public Health Concern

Young adulthood (typically the ages of 18-30 y) is a critical developmental stage characterized by a peak prevalence of substance use [1,2]. Worldwide, cannabis is a substance frequently used for nonmedical purposes due in part to its high availability in some regions and enhanced product variety and potency [3,4]. The prevalence of cannabis use (CU) among young adults is high [5,6], and its rates have risen in recent decades [7]. In North America and Oceania, the estimated past-year prevalence of CU is ≥25% among young adults [8,9].

While the vast majority of cannabis users do not experience severe problems from their use [4], the high prevalence of CU among young adults poses substantial global health concerns due to the associated acute and long-term health and psychosocial risks [10,11]. These include impairment of cognitive function, memory, and psychomotor skills during acute intoxication; increased engagement in behaviors with a potential for injury and fatality (eg, driving under the influence); socioeconomic problems; and diminished social functioning [4,12-14]. Importantly, an extensive body of literature reveals that subgroups engaging in higher-risk use, such as intensive or repeated use, are more prone to severe and chronic consequences, including physical ailments (eg, respiratory illness and reproductive dysfunction), mental health disorders (eg, psychosis, depression, and suicidal ideation or attempts), and the potential development of CU disorder [4,15-17].

Interventions to Reduce Public Health Impact of Young Adult CU

Given the increased prevalence of lifetime and daily CU among young adults and the potential negative impact of higher-risk CU, various prevention and intervention programs have been implemented to help users reduce or cease their CU. These programs primarily target young adults regardless of their CU status [2,18]. In this context, many health care organizations and international expert panels have developed evidence-based lower-risk CU guidelines to promote safer CU and intervention options to help reduce risks of adverse health outcomes from nonmedical CU [4,16,17,19]. Lower-risk guidance-oriented interventions for CU are based on concepts of health promotion [20-22] and health behavior change [23-26] and on other similar harm reduction interventions implemented in other areas of population health (eg, lower-risk drinking guidelines, supervised consumption sites and services, and sexual health) [27,28]. These interventions primarily aim to raise awareness of negative mental, physical, and social cannabis-related consequences to modify individual-level behavior-related risk factors.

Meta-analyses have shown that face-to-face prevention and treatment interventions are generally effective in reducing CU in young adults [18,29-32]. However, as the proportion of professional help seeking for CU concerns among young adults remains low (approximately 15%) [33,34], alternative strategies that consider the limited capacities and access-related barriers of traditional face-to-face prevention and treatment facilities are needed. Digital interventions, including websites, digital platforms, and mobile apps, have emerged as promising tools to enhance the accessibility and availability of evidence-based programs for young adult cannabis users. These interventions address barriers such as long-distance travel, concerns about confidentiality, stigma associated with seeking treatment, and the cost of traditional treatments [35-37]. By overcoming these barriers, digital interventions have the potential to have a stronger public health impact [18,38].

State of Knowledge of Digital Interventions for CU and Young Adults

The literature regarding digital interventions for substance use has grown rapidly in the past decade, as evidenced by several systematic reviews and meta-analyses of randomized controlled trial (RCT) studies on the efficacy or effectiveness of these interventions in preventing or reducing harmful substance use [2,39-41]. However, these reviews do not focus on young adults specifically. In addition, they combine CU-related outcomes with those of many other substances in their meta-analytical results. Finally, they do not target CU interventions exclusively.

In total, 4 systematic reviews and meta-analyses of digital interventions for CU among young people have reported mixed results [42-45]. In their systematic review (10 studies of 5 prevention and 5 treatment interventions up to 2012), Tait et al [44] concluded that digital interventions effectively reduced CU among adolescents and adults at the posttreatment time point. Olmos et al [43] reached a similar conclusion in their meta-analysis of 9 RCT studies (2 prevention and 7 treatment interventions). In their review, Hoch et al [42] reported evidence of small effects at the 3-month follow-up based on 4 RCTs of brief motivational interventions and cognitive behavioral therapy (CBT) delivered on the web. In another systematic review and meta-analysis, Beneria et al [45] found that web-based CU

XSL•FO
RenderX

interventions did not significantly reduce consumption. However, these authors indicated that the programs tested varied significantly across the studies considered and that statistical heterogeneity was attributable to the inclusion of studies of programs targeting more than one substance (eg, alcohol and cannabis) and both adolescents and young adults. Beneria et al [45] recommend that future work "establish the effectiveness of the newer generation of interventions as well as the key ingredients" of effective digital interventions addressing CU by young people. This is of particular importance because behavior change interventions tend to be complex as they consist of multiple interactive components [46].

Behavior change interventions refer to "coordinated sets of activities designed to change specified behavior patterns" [47]. Their interacting active ingredients can be conceptualized as behavior change techniques (BCTs) [48]. BCTs are specific and irreducible. Each BCT has its own individual label and definition, which can be used when designing and reporting complex interventions and as a nomenclature system when coding interventions for their content [47]. The Behavior Change Technique Taxonomy version 1 (BCTTv1) [48,49] was developed to provide a shared, standardized terminology for characterizing complex behavior change interventions and their active ingredients. Several systematic reviews with meta-regressions that used the BCTTv1 have found interventions with certain BCTs to be more effective than those without [50-53]. A better understanding of the BCTs used in digital interventions for young adult cannabis users would help not only to establish the key ingredients of such interventions but also develop and evaluate effective interventions.

In the absence of any systematic review of the effectiveness and active ingredients of digital interventions designed specifically for CU among community-living young adults, we set out to achieve the following:

1. conduct a comprehensive review of digital interventions for preventing, reducing, or ceasing CU among community-living young adults,
2. describe the active ingredients (ie, BCTs) in these interventions from the perspective of behavior change science, and
3. analyze the effectiveness of these interventions on CU outcomes.

Methods

Protocol Registration

We followed the Cochrane Handbook for Systematic Reviews of Interventions [54] in designing this systematic review and meta-analysis and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines in reporting our findings (see Multimedia Appendix 1 [55] for the complete PRISMA checklist). This review was registered in PROSPERO (CRD42020196959).

Search Strategy

Overview

The search strategy was designed by a health information specialist together with the research team and peer reviewed by another senior information specialist before execution using Peer Review of Electronic Search Strategies for systematic reviews [56]. The search strategy revolved around three concepts:

1. CU (eg, "cannabis," "marijuana," and "hashish")
2. Digital interventions (eg, "telehealth," "website," "mobile applications," and "computer")
3. Young adults (eg, "emerging adults" and "students")

The strategy was initially implemented on March 18, 2020, and again on October 13, 2021, and February 13, 2023. The full, detailed search strategies for each database are presented in Multimedia Appendix 2.

Information Sources

We searched 7 electronic databases of published literature: CINAHL Complete, Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, Embase, MEDLINE, PubMed, and PsycINFO. No publication date filters or language restrictions were applied. A combination of free-text keywords and Medical Subject Headings was tailored to the conventions of each database for optimal electronic searching. The research team also manually screened the reference lists of the included articles and the bibliographies of existing systematic reviews [18,31,42-45] to identify additional relevant studies (snowballing). Finally, a forward citation tracking procedure (ie, searching for articles that cited the included studies) was carried out in Google Scholar.

Inclusion Criteria

The population, intervention, comparison, outcome, and study design process is presented in Multimedia Appendix 3. The inclusion criteria were as follows: (1) original research articles published in peer-reviewed journals; (2) use of an experimental study design (eg, RCT, cluster RCT, or pilot RCT); (3) studies evaluating the effectiveness (or efficacy) of digital interventions designed specifically to prevent, reduce, or cease CU as well as promote CU self-management or address cannabis-related harm and having CU as an outcome measure; (4) studies targeting young adults, including active and nonactive cannabis users; (5) cannabis users and nonusers not under substance use treatment used as controls in comparator, waitlist, or delayed-treatment groups offered another type of intervention (eg, pharmacotherapy or psychosocial) different from the one being investigated or participants assessed only for CU; and (6) quantitative CU outcomes (frequency and quantity) or cannabis abstinence. Given the availability of numerous CU screening and assessment tools with adequate psychometric properties and the absence of a gold standard in this regard [57], any instrument capturing aspects of CU was considered. CU outcome measures could be subjective (eg, self-reported number of CU days or joints in the previous 3 months) or objective (eg, drug screening test). CU had to be measured before the intervention (baseline) and at least once after.

XSL•FO
RenderX

Digital CU interventions were defined as web- or mobile-based interventions that included one or more activities (eg, self-directed or interactive psychoeducation or therapy, personalized feedback, peer-to-peer contact, and patient-to-expert communication) aimed at changing CU [58]. Mobile-based interventions were defined as interventions delivered via mobile phone through SMS text message, multimedia messaging service (ie, SMS text messages that include multimedia content, such as pictures, videos, or emojis), or mobile apps, whereas web-based interventions (eg, websites and digital platforms) were defined as interventions designed to be accessed on the web (ie, the internet), mainly via computers. Interventions could include self-directed and web-based interventions with human support. We defined young adults as aged 16 to 35 years and included students and nonstudents. While young adulthood is typically defined as covering the ages of 18 to 30 years [59], we broadened the range given that the age of majority and legal age to purchase cannabis differs across countries and jurisdictions. This was also in line with the age range targeted by several digital CU interventions (college or university students or emerging adults aged 15-24 years) [31,45]. Given the language expertise of the research team members and the available resources, only English- and French-language articles were retained.

Exclusion Criteria

Knowledge synthesis articles, study protocols, and discussion papers or editorials were excluded, as were articles with cross-sectional, cohort, case study or report, pretest-posttest, quasi-experimental, or qualitative designs. Mixed methods designs were included only if the quantitative component was an RCT. We excluded studies if (1) use of substances other than cannabis (eg, alcohol, opioids, or stimulants) was the focus of the digital intervention (though studies that included polysubstance users were retained if CU was assessed and reported separately); (2) CU was not reported separately as an outcome or only attitudes or beliefs regarding, knowledge of, intention to reduce, or readiness or motivation to change CU was measured; and (3) the data reported were unpublished (eg, conferences and dissertations). Studies of traditional face-to-face therapy delivered via teleconference on mobile phones and computers or in a hospital-based setting and informational campaigns (eg, web-based poster presentations or pamphlets) were excluded as well. Studies with samples with a maximum age of <15 years and a minimum age of >35 years were also excluded. Finally, we excluded studies that focused exclusively on people with a mental health disorder or substance use disorder or dependence or on adolescents owing to the particular health care needs of these populations, which may differ from those of young adults [1].

Data Collection

Selection of Studies

Duplicates were removed from the literature search results in EndNote (version X9.3.3; Clarivate Analytics) using the Bramer method for deduplication of database search results for systematic reviews [60]. The remaining records were uploaded to Covidence (Veritas Health Innovation), a web-based systematic review management system. A reviewer guide was developed that included screening questions and a detailed description of each inclusion and exclusion criterion based on PICO (population, intervention, comparator, and outcome), and a calibration exercise was performed before each stage of the selection process to maximize consistency between reviewers. Titles and abstracts of studies flagged for possible inclusion were screened first by 2 independent reviewers (GC, BV, PA, and GR; 2 per article) against the eligibility criteria (stage 1). Articles deemed eligible for full-text review were then retrieved and screened for inclusion (stage 2). Full texts were assessed in detail against the eligibility criteria again by 2 reviewers independently. Disagreements between reviewers were resolved through consensus or by consulting a third reviewer.

Data Extraction Process

In total, 2 reviewers (GC, BV, PA, GR, and GF; 2 per article) independently extracted relevant data (or informal evidence) using a data extraction form developed specifically for this review and integrated into Covidence. The form was pilot-tested on 2 randomly selected studies and refined accordingly. Data pertaining to the following domains were extracted from the included studies: (1) Study characteristics included information on the first and corresponding authors, publication year, country of origin, aims and hypotheses, study period, design (including details on randomization and blinding), follow-up times, data collection methods, and types of statistical analysis. (2) Participant characteristics included study target population, participant inclusion and exclusion criteria, sex or gender, mean age, and sample sizes at each data collection time point. (3) Intervention characteristics, for which the research team developed a matrix inspired by the template for intervention description and replication 12-item checklist [61] to extract informal evidence (ie, intervention descriptions) from the included studies under the headings name of intervention, purpose, underpinning theory of design elements, treatment approach, type of technology (ie, web or mobile) and software used, delivery format (ie, self-directed, human involvement, or both), provider characteristics (if applicable), intervention duration (ie, length of treatment and number of sessions or modules), material and procedures (ie, tools or activities offered, resources provided, and psychoeducational content), tailoring, and unplanned modifications. (4) Comparator characteristics were details of the control or comparison group or groups, including nature (passive vs active), number of groups or clusters (if applicable), type and length of the intervention (if applicable), and number of participants at each data collection time point. (5) Outcome variables, including the primary outcome variable examined in this systematic review, that is, the mean difference in CU frequency before and after the intervention and between the experimental and control or comparison groups. When possible, we examined continuous variables, including CU frequency means and SDs at the baseline and follow-up time points, and standardized regression coefficients (ie, β coefficients and associated 95% CIs). The secondary outcomes examined included other CU outcome variables (eg, quantity of cannabis used and abstinence) and cannabis-related negative consequences (or problems). Details on outcome variables (ie, definition, data time points, and

XSL•FO
RenderX

missing data) and measurements (ie, instruments, measurement units, and scales) were also extracted.

In addition, data on user engagement and use of the digital intervention and study attrition rates (ie, dropouts and loss to follow-up) were extracted. When articles had missing data, we contacted the corresponding authors via email (2 attempts were made over a 2-month period) to obtain missing information. Disagreements over the extracted data were limited and resolved through discussion.

Data Synthesis Methods

Descriptive Synthesis

The characteristics of the included studies, study participants, interventions, and comparators were summarized in narrative and table formats. The template for intervention description and replication 12-item checklist [61] was used to summarize and organize intervention characteristics and assess to what extent the interventions were appropriately described in the included articles. As not all studies had usable data for meta-analysis purposes and because of heterogeneity, we summarized the main findings (ie, intervention effects) of the included studies in narrative and table formats for each outcome of interest in this review.

BCT Coding

The BCTs used in the digital interventions were identified from the descriptions of the interventions (ie, experimental groups) provided in the articles as well as any supplementary material and previously published research protocols. A BCT was defined as “an observable, replicable, and irreducible component of an intervention designed to alter or redirect causal processes that regulate behavior” [48]. The target behavior in this review was the cessation or reduction of CU by young adults. BCTs were identified and coded using the BCTTv1 [48,49], a taxonomy of 93 BCTs organized into 16 hierarchical thematic clusters or categories. Applying the BCTTv1 in a systematic review allows for the comparison and synthesis of evidence across studies in a structured manner. This analysis allows for the identification of the explicit mechanisms underlying the reported behavior change induced by interventions, successful or not, and, thus, avoids making implicit assumptions about what works [62].

BCT coding was performed by 2 reviewers independently—BV coded all studies, and GC and GF coded a subset of the studies. All reviewers completed web-based training on the BCTTv1, and GF is an experienced implementation scientist who had used the BCTTv1 in prior work [63-65]. The descriptions of the interventions in the articles were read line by line and analyzed for the clear presence of BCTs using the guidelines developed by Michie et al [48]. For each article, the BCTs identified were documented and categorized using supporting textual evidence. They were coded only once per article regardless of how many times they came up in the text. Disagreements about including a BCT were resolved through discussion. If there was uncertainty about whether a BCT was present, it was coded as absent. Excel (Microsoft Corp) was used to compare the reviewers’ independent BCT coding and generate an overall descriptive synthesis of the BCTs identified. The BCTs were summarized by study and BCT cluster.

Statistical Analysis

Meta-analyses were conducted to estimate the size of the effect of the digital interventions for young adult CU on outcomes of interest at the posttreatment and follow-up assessments compared with control or alternative intervention conditions. The outcome variables considered were (1) CU frequency and other CU outcome variables (eg, quantity of cannabis used and abstinence) at baseline and the posttreatment time point or follow-up measured using standardized instruments of self-reported CU (eg, the timeline followback [TLFB] method) [66] and (2) cannabis-related negative consequences measured using standardized instruments (eg, the Marijuana Problems Scale) [67].

Under our systematic review protocol, ≥2 studies were needed for a meta-analysis. On the basis of previous systematic reviews and meta-analyses in the field of digital CU interventions [31,42-45], we expected between-study heterogeneity regarding outcome assessment. To minimize heterogeneity, we chose to pool studies with similar outcomes of interest based on four criteria: (1) definition of outcome (eg, CU frequency, quantity consumed, and abstinence), (2) type of outcome variable (eg, days of CU in the previous 90 days, days high per week in the previous 30 days, and number of CU events in the previous month) and measure (ie, instruments or scales), (3) use of validated instruments, and (4) posttreatment or follow-up time points (eg, 2 weeks or 1 month after the baseline or 3, 6, and 12 months after the baseline).

Only articles that reported sufficient statistics to compute a valid effect size with 95% CIs were included in the meta-analyses. In the case of articles that were not independent (ie, more than one published article reporting data from the same clinical trial), only 1 was included, and it was represented only once in the meta-analysis for a given outcome variable regardless of whether the data used to compute the effect size were extracted from the original paper or a secondary analysis paper. We made sure that the independence of the studies included in the meta-analysis of each outcome was respected. In the case of studies that had more than one comparator, we used the effect size for each comparison between the intervention and control groups.

Meta-analyses were conducted only for mean differences based on the change from baseline in CU frequency at 3 months after the baseline as measured using the number of self-reported days of use in the previous month. As the true value of the estimated effect size for outcome variables might vary across different trials and samples, we used a random-effects model given that the studies retained did not have identical target populations. The random-effects model incorporates between-study variation in the study weights and estimated effect size [68]. In addition, statistical heterogeneity across studies was assessed using I^2, which measures the proportion of heterogeneity to the total observed dispersion; 25% was considered low, 50% was considered moderate, and 75% was considered high [69]. Because only 3 studies were included in the meta-analysis [70-72], publication bias could not be assessed. All analyses were completed using Stata (version 18; StataCorp) [73].

XSL•FO
RenderX

Risk-of-Bias Assessment

The risk of bias (RoB) of the included RCTs was assessed using the Cochrane RoB 2 tool at the outcome level [74]. Each distinct risk domain (ie, randomization process, deviations from the intended intervention, missing outcome data, measurement of the outcome, and selection of the reported results) was assessed as "low," "some concerns," or "high" based on the RoB 2 criteria. In total, 2 reviewers (GC and BV) conducted the assessments independently. Disagreements were discussed, and if not resolved consensually by the 2, the matter was left for a third reviewer (GF) to settle. The assessments were summarized by risk domain and outcome and converted into figures using the RoB visualization tool *robvis* [75].

Results

Search Results

The database search generated a total of 13,232 citations, of which 7822 (59.11%) were from the initial search on March 18, 2020, and 2805 (21.2%) and 2605 (19.69%) were from the updates on October 13, 2021, and February 13, 2023, respectively. Figure 1 presents the PRISMA study flow diagram [76]. Of the 6606 unique records, 6484 (98.15%) were excluded based on title and abstract screening. Full texts of the remaining 1.85% (122/6606) of the records were examined, as were those of 25 more reports found through hand searching. Of these 147 records, 128 (87.1%) were excluded after 3 rounds of full-text screening. Of these 128 records, 39 (30.5%) were excluded for not being empirical research articles (eg, research protocols). Another 28.1% (36/128) were excluded for not meeting our definition of digital CU intervention. The remaining records were excluded for reasons that occurred with a frequency of ≤14%, including young adults not being the target population and the study not meeting our study design criteria (ie, RCT, cluster RCT, or pilot RCT). Excluded studies and reasons for exclusion are listed in Multimedia Appendix 4. Finally, 19 articles detailing the results of 19 original studies were included.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) study flow diagram.

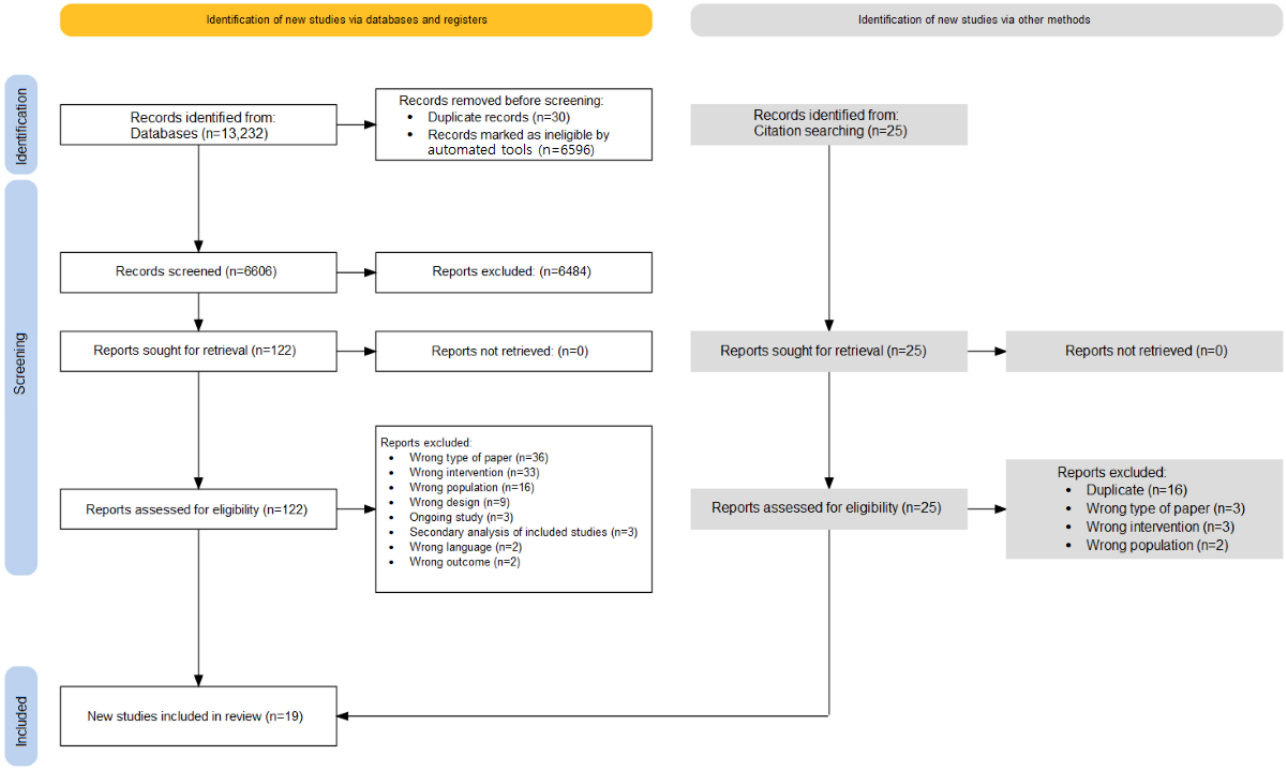

Description of Studies

Study Characteristics

Multimedia Appendix 5 [70-72,77-92] describes the general characteristics of the 19 included studies. The studies were published between 2010 and 2023, with 58% (11/19) published in 2018 or later. A total of 53% (10/19) of the studies were conducted in the United States [77-86], 11% (2/19) were conducted in Canada [87,88], 11% (2/19) were conducted in Australia [71,89], 11% (2/19) were conducted in Germany [72,90], 11% (2/19) were conducted in Switzerland [70,91], and 5% (1/19) were conducted in Sweden [92]. A total of 79% (15/19) were RCTs [70-72,77,79,81-83,86-92], and 21% (4/19) were pilot RCTs [78,80,84,85].

Participant Characteristics

The studies enrolled a total of 6710 participants—3229 (48.1%) in the experimental groups, 3358 (50%) in the control groups, and the remaining 123 (1.8%) from 1 study [82] where participant allocation to the intervention condition was not reported. Baseline sample sizes ranged from 49 [81] to 1292 [72] (mean 352.89, SD 289.50), as shown in Multimedia Appendix 5. Participant mean ages ranged from 18.03 (SD 0.31) [79] to 35.3 (SD 12.6) years [88], and the proportion of

XSL•FO
RenderX

participants who identified as female ranged from 24.7% [91] to 84.1% [80].

Of the 19 included studies, 10 (53%) targeted adults aged ≥18 years, of which 7 (70%) studies focused on adults who had engaged in past-month CU [70,71,80,84,85,90,91], 2 (20%) studies included adults who wished to reduce or cease CU [72,89], and 1 (10%) study focused on noncollege adults with a moderate risk associated with CU [88]. Sinadinovic et al [92] targeted young adults aged ≥16 years who had used cannabis at least once a week in the previous 6 months. The remaining 8 studies targeted college or university students (aged ≥17 y) specifically, of which 7 (88%) studies focused solely on students who reported using cannabis [78,79,81-83,86,87] and 1 (12%) study focused solely on students who did not report past-month CU (ie, abstainers) [77].

Intervention Characteristics

The 19 included studies assessed nine different digital interventions: (1) 5 (26%) evaluated Marijuana eCHECKUP TO GO (e-TOKE), a commercially available electronic intervention used at colleges throughout the United States and Canada [77,78,81-83]; (2) 2 (11%) examined the internationally known CANreduce program [70,91]; (3) 2 (11%) evaluated the German Quit the Shit program [72,90]; (4) 2 (11%) assessed a social media–delivered, physical activity–focused cannabis intervention [84,85]; (5) 1 (5%) investigated the Swedish *Cannabishjälpen* intervention [92]; (6) 1 (5%) evaluated the Australian Grassessment: Evaluate Your Use of Cannabis website program [89]; (7) 1 (5%) assessed the Canadian *Ma réussite, mon choix* intervention [87]; (8) 1 (5%) examined the Australian Reduce Your Use: How to Break the Cannabis Habit program [71]; and (9) 4 (21%) each evaluated a unique no-name intervention described as a personalized feedback intervention (PFI) [79,80,86,88]. Detailed information regarding the characteristics of all interventions as reported in each included study is provided in Multimedia Appendix 6 [70-72,77-113] and summarized in the following paragraphs.

In several studies (8/19, 42%), the interventions were designed to support cannabis users in reducing or ceasing their consumption [70,72,80,87,89-92]. In 37% (7/19) of the studies, the interventions aimed at reducing both CU and cannabis-related consequences [79,81-85,88]. Other interventions focused on helping college students think carefully about the decision to use cannabis [77,78] and on reducing either cannabis-related problems among undergraduate students [86] or symptoms associated with CU disorder in young adults [71].

In 26% (5/19) of the studies, theory was used to inform intervention design along with a clear rationale for theory use. Of these 5 articles, only 1 (20%) [87] reported using a single theory of behavior change, the theory of planned behavior [114]. A total of 21% (4/19) of the studies selected only constructs of theories (or models) for their intervention design. Of these 4 studies, 2 (50%) evaluated the same intervention [72,90], which focused on principles of self-regulation and self-control theory [93]; 1 (25%) [70] used the concept of adherence-focused guidance enhancement based on the supportive accountability model of guidance [94]; and 1 (25%) [71] reported that intervention design was guided by the concept of self-behavioral management.

The strategies (or approaches) used in the delivery of the digital interventions were discussed in greater detail in 84% (16/19) of the articles [70-72,79-81,83-92]. Many of these articles (9/19, 47%) reported using a combination of approaches based on CBT or motivational interviewing (MI) [70,71,79,83-85,90-92]. PFIs were also often mentioned as an approach to inform intervention delivery [7,71,79,86-88].

More than half (13/19, 68%) of all the digital interventions were asynchronous and based on a self-guided approach without support from a counselor or therapist. The study by Côté et al [87] evaluated the efficacy of a web-based tailored intervention focused on reinforcing a positive attitude toward and a sense of control over cannabis abstinence through psychoeducational messages delivered by a credible character in short video clips and personalized reinforcement messages. Lee et al [79] evaluated a brief, web-based personalized feedback selective intervention based on the PFI approach pioneered by Marlatt et al [95] for alcohol use prevention and on the MI approach described by Miller and Rollnick [96]. Similarly, Rooke et al [71] combined principles of MI and CBT to develop a web-based intervention delivered via web modules, which were informed by previous automated feedback interventions targeting substance use. The study by Copeland et al [89] assessed the short-term effectiveness of Grassessment: Evaluate Your Use of Cannabis, a brief web-based, self-complete intervention based on motivational enhancement therapy that included personalized feedback messages and psychoeducational material. In the studies by Buckner et al [80], Cunningham et al [88], and Walukevich-Dienst et al [86], experimental groups received a brief web-based PFI available via a computer. A total of 16% (3/19) of the studies [77,78,82] applied a program called the Marijuana eCHECKUP TO GO (e-TOKE) for Universities and Colleges, which was presented as a web-based, norm-correcting, brief preventive and intervention education program designed to prompt self-reflection on consequences and consideration of decreasing CU among students. Riggs et al [83] developed and evaluated an adapted version of e-TOKE that provided participants with university-specific personalized feedback and normative information based on protective behavioral strategies for CU [97]. Similarly, Goodness and Palfai [81] tested the efficacy of eCHECKUP TO GO-cannabis, a modified version of e-TOKE combining personalized feedback, norm correction, and a harm and frequency reduction strategy where a “booster” session was provided at 3 months to allow participants to receive repeated exposure to the intervention.

In the remaining 32% (6/19) of the studies, which examined 4 different interventions, the presence of a therapist guide was reported. The intervention evaluated by Sinadinovic et al [92] combined principles of psychoeducation, MI, and CBT organized into 13 web-based modules and a calendar involving therapist guidance, recommendations, and personal feedback. In total, 33% (2/6) of these studies evaluated a social media–delivered intervention with e-coaches that combined principles of MI and CBT and a harm reduction approach for risky CU [84,85]. Schaub et al [91] evaluated the efficacy of

XSL•FO
RenderX

CANreduce, a web-based self-help intervention based on both MI and CBT approaches, using automated motivational and feedback emails, chat with a counselor, and web-based psychoeducational modules. Similarly, Baumgartner et al [70] investigated the effectiveness of CANreduce 2.0, a modified version of CANreduce, using semiautomated motivational and adherence-focused guidance-based email feedback with or without a personal online coach. The studies by Tossman et al [72] and Jonas et al [90] used a solution-focused approach and MI to evaluate the effectiveness of the German Quit the Shit web-based program that involves weekly feedback provided by counselors.

In addition to using different intervention strategies or approaches, the interventions were diverse in terms of the duration and frequency of the program (eg, web-based activities, sessions, or modules). Of the 12 articles that provided details in this regard, 2 (17%) on the same intervention described it as a brief 20- to 45-minute web-based program [77,78], 2 (17%) on 2 different interventions reported including 1 or 2 modules per week for a duration of 6 weeks [71,92], and 7 (58%) on 4 different interventions described them as being available over a longer period ranging from 6 weeks to 3 months [70,72,79,84,85,87,90,91].

Comparator Types

A total of 42% (8/19) of the studies [72,77-80,85,87,92] used a passive comparator only, namely, a waitlist control group (Multimedia Appendix 5). A total of 26% (5/19) of the studies used an active comparator only where participants were provided with minimal general health feedback regarding recommended guidelines for sleep, exercise, and nutrition [81,82]; strategies for healthy stress management [83]; educational materials about risky CU [88]; or access to a website containing information about cannabis [71]. In another 21% (4/19) of the studies, which used an active comparator, participants received the same digital intervention minus a specific component: a personal web-based coach [70], extended personalized feedback [89], web-based chat counseling [91], or information on risks associated with CU [86]. A total of 21% (4/19) of the studies had more than one control group [70,84,90,91].

Outcome Variable Assessment and Summary of Main Findings of the Studies

Overview

The methodological characteristics and major findings of the included studies (N=19) are presented in Multimedia Appendix 7 [67,70-72,77-92,115-120] and summarized in the following sections for each outcome of interest in this review (ie, CU and cannabis-related consequences). Of the 19 studies, 11 (58%) were reported as efficacy trials [7,77,79,81-83,86-88,91,92], and 8 (42%) were reported as effectiveness trials [70-72,78,84,85,89,90].

Across all the included studies (19/19, 100%), participant attrition rates ranged from 1.6% at 1 month after the baseline [77,78] to 75.1% at the 3-month follow-up [70]. A total of 37% (7/19) of the studies assessed and reported results regarding user engagement [71,78,84,85,90-92] using different types of metrics. In one article on the Marijuana eCHECKUP TO GO (e-TOKE) web-based program [78], the authors briefly reported that participation was confirmed for 98.1% (158/161) of participants in the intervention group. In 11% (2/19) of the studies, which were on a similar social media–delivered intervention [84,85], user engagement was quantified by tallying the number of comments or posts and reactions (eg, likes and hearts) left by participants. In both studies [84,85], the intervention group, which involved a CU-related Facebook page, displayed greater interactions than the control groups, which involved a Facebook page unrelated to CU. One article [84] reported that 80% of participants in the intervention group posted at least once (range 0-60) and 50% posted at least weekly. In the other study [85], the results showed that intervention participants engaged (ie, posting or commenting or clicking reactions) on average 47.9 times each over 8 weeks. In total, 11% (2/19) of the studies [90,91] on 2 different web-based intervention programs, both consisting of web documentation accompanied by chat-based counseling, measured user engagement either by average duration or average number of chat sessions. Finally, 16% (3/19) of the studies [71,91,92], which involved 3 different web-based intervention programs, characterized user engagement by the mean number of web modules completed per participant. Overall, the mean number of web modules completed reported in these articles was quite similar: 3.9 out of 13 [92] and 3.2 [91] and 3.5 [71] out of 6.

Assessment of CU

As presented in Multimedia Appendix 7, the included studies differed in terms of how they assessed CU, although all used at least one self-reported measure of frequency. Most studies (16/19, 84%) measured frequency by days of use, including days of use in the preceding week [91] or 2 [80], days of use in the previous 30 [70-72,78,84-86,88-90] or 90 days [79,81,82], and days high per week [83]. Other self-reported measures of CU frequency included (1) number of CU events in the previous month [87,90], (2) cannabis initiation or use in the previous month (ie, yes or no) [77], and (3) days without CU in the previous 7 days [92]. In addition to measuring CU frequency, 42% (8/19) of the studies also assessed CU via self-reported measures of quantity used, including estimated grams consumed in the previous week [92] or 30 days [72,85,90] and the number of standard-sized joints consumed in the previous 7 days [91] or the previous month [70,71,89].

Of the 19 articles included, 10 (53%) [70-72,80,84-86,89,90,92] reported using a validated instrument to measure CU frequency or quantity, including the TLFB instrument [66] (n=9, 90% of the studies) and the Marijuana Use Form (n=1, 10% of the studies); 1 (10%) [79] reported using CU-related questions from an adaptation of the Global Appraisal of Individual Needs–Initial instrument [115]; and 30% (3/10) [81,82,91] reported using a questionnaire accompanied by a calendar or a diary of consumption. The 19 studies also differed with regard to their follow-up time measurements for assessing CU, ranging from 2 weeks after the baseline [80] to 12 months after randomization [90], although 12 (63%) of the studies included a 3-month follow-up assessment [70-72,79,81,82,84,85,88,90-92].

Of all studies assessing and reporting change in CU frequency from baseline to follow-up assessments (19/19, 100%), 47%

XSL•FO
RenderX

(9/19) found statistically significant differences between the experimental and control groups [70-72,80,81,83,85,87,91]. Importantly, 67% (6/9) of these studies showed that participants in the experimental groups exhibited greater decreases in CU frequency 3 months following the baseline assessment compared with participants in the control groups [70-72,81,85,91], 22% (2/9) of the studies showed greater decreases in CU frequency at 6 weeks after the baseline assessment [71,83], 22% (2/9) of the studies showed greater decreases in CU frequency at 6 months following the baseline assessment [81,85], 11% (1/9) of the studies showed greater decreases in CU frequency at 2 weeks after the baseline [80], and 11% (1/9) of the studies showed greater decreases in CU frequency at 2 months after treatment [87].

In the study by Baumgartner et al [70], a reduction in CU days was observed in all groups, but the authors reported that the difference was statistically significant only between the intervention group with the service team and the control group (the reduction in the intervention group with social presence was not significant). In the study by Bonar et al [85], the only statistically significant difference between the intervention and control groups at the 3- and 6-month follow-ups involved total days of cannabis vaping in the previous 30 days. Finally, in the study by Buckner et al [80], the intervention group had less CU than the control group 2 weeks after the baseline; however, this was statistically significant only for participants with moderate or high levels of social anxiety.

Assessment of Cannabis-Related Negative Consequences

A total of 53% (10/19) of the studies also assessed cannabis-related negative consequences [78-84,86,88,92]. Of these 10 articles, 8 (80%) reported using a validated self-report instrument: 4 (50%) [81,82,86,88] used the 19-item Marijuana Problems Scale [67], 2 (25%) [78,79] used the 18-item Rutgers Marijuana Problem Index [121,122], and 2 (25%) [80,84] used the Brief Marijuana Consequences Questionnaire [116]. Only 10% (1/10) of the studies [92] used a screening tool, the Cannabis Abuse Screening Test [117,118]. None of these 10 studies demonstrated a statistically significant difference between the intervention and control groups. Of note, Walukevich-Dienst et al [86] found that women (but not men) who received an web-based PFI with additional information on CU risks reported significantly fewer cannabis-related problems than did women in the control group at 1 month after the intervention (B=−1.941; P=.01).

Descriptive Summary of BCTs Used in Intervention Groups

After the 19 studies included in this review were coded, a total of 184 individual BCTs targeting CU in young adults were identified. Of these 184 BCTs, 133 (72.3%) were deemed to be present beyond a reasonable doubt, and 51 (27.7%) were deemed to be present in all probability. Multimedia Appendix 8 [48,70-72,77-92] presents all the BCTs coded for each included study summarized by individual BCT and BCT cluster.

The 184 individual BCTs coded covered 38% (35/93) of the BCTs listed in the BCTTv1 [48]. The number of individual BCTs identified per study ranged from 5 to 19, with two-thirds of the 19 studies (12/19, 63%) using ≤9 BCTs (mean 9.68). As Multimedia Appendix 8 shows, at least one BCT fell into 13 of the 16 possible BCT clusters. The most frequent clusters were *feedback monitoring*, *natural consequences*, *goal planning*, and *comparison of outcomes*.

The most frequently coded BCTs were (1) *feedback on behavior* (BCT 2.2; 17/19, 89% of the studies; eg, “Once a week, participants receive detailed feedback by their counselor on their entries in diary and exercises. Depending on the involvement of each participant, up to seven feedbacks are given” [90]), (2) *social support (unspecified)* (BCT 3.1; 15/19, 79% of the studies; eg, “The website also features [...] blogs from former cannabis users, quick assist links, and weekly automatically generated encouragement emails” [71]), and (3) *pros and cons* (BCT 9.2; 14/19, 74% of the studies; eg, “participants are encouraged to state their personal reasons for and against their cannabis consumption, which they can review at any time, so they may reflect on what they could gain by successfully completing the program” [70]). Other commonly identified BCTs included *social comparison* (BCT 6.2; 12/19, 63% of the studies) and *information about social and environmental consequences* (BCT 5.3; 11/19, 58% of the studies), followed by *problem solving* (BCT 2.1; 10/19, 53% of the studies) and *information about health consequences* (BCT 5.1; 10/19, 53% of the studies).

RoB Assessment

Figure 2 presents the overall assessment of risk in each domain for all the included studies, whereas Figure 3 [70-72,77-92] summarizes the assessment of each study at the outcome level for each domain in the Cochrane RoB 2 [74].

Figure 2 shows that, of the included studies, 93% (27/29) were rated as having a “low” RoB arising from the randomization process (ie, selection bias) and 83% (24/29) were rated as having a “low” RoB due to missing data (ie, attrition bias). For bias due to deviations from the intended intervention (ie, performance bias), 72% (21/29) were rated as having a “low” risk, and for selective reporting of results, 59% (17/29) were rated as having a “low” risk. In the remaining domain regarding bias in measurement of the outcome (ie, detection bias), 48% (14/29) of the studies were deemed to present “some concerns,” mainly owing to the outcome assessment not being blinded (eg, self-reported outcome measure of CU). Finally, 79% (15/19) of the included studies were deemed to present “some concerns” or were rated as having a “high” RoB at the outcome level (Figure 3 [70-72,77-92]). The RoB assessment for CU and cannabis consequences of each included study is presented in Multimedia Appendix 9 [70-72,77-92].

XSL•FO
RenderX

Figure 2. Risk-of-bias (RoB) assessment of the included studies (N=19) summarized by RoB domain.

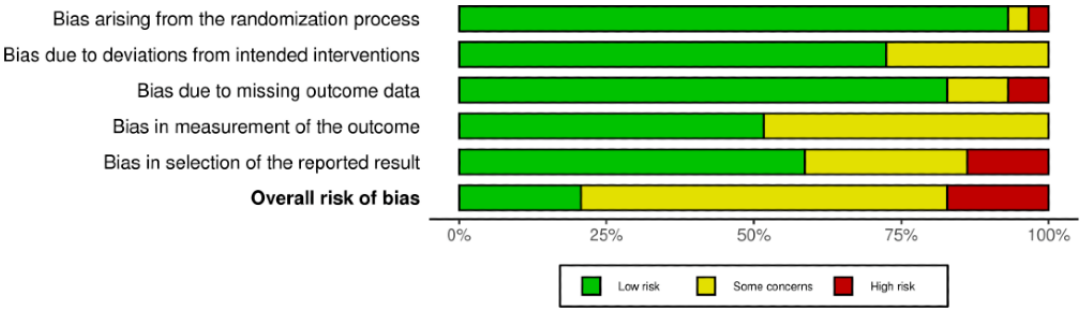

Figure 3. Summary of risk-of-bias (RoB) assessments of each included study (N=19) at the outcome level for each RoB domain.

XSL•FO
RenderX

Meta-Analysis Results

Due to several missing data points and despite contacting the authors, we were able to carry out only 1 meta-analysis of our primary outcome, CU frequency. Usable data were retrieved from only 16% (3/19) [70-72] of the studies included in this review. These 3 studies provided sufficient information to calculate an effect size, including mean differences based on change-from-baseline measurements and associated 95% CIs (or SE of the mean difference) and sample sizes per intervention and comparison conditions. The reasons for excluding the other 84% (16/19) of the studies included heterogeneity in outcome variables or measurements, inconsistent results, and missing data (Multimedia Appendix 10 [77-92]).

Figure 4 [70-72] illustrates the mean differences and associated 95% CIs of 3 unique RCTs [70-72] that provided sufficient information to allow for the measurement of CU frequency at 3 months after the baseline relative to a comparison condition in terms of the number of self-reported days of use in the previous month using the TLFB method. Overall, the synthesized effect of digital interventions for young adult cannabis users on CU frequency, as measured using days of use in the previous month, was −6.79 (95% CI −9.59 to −4.00). This suggests that digital CU interventions had a statistically significant effect (P<.001) on reducing CU frequency at the 3-month follow-up compared with the control conditions (both passive and active controls). The results of the meta-analysis also showed low between-study heterogeneity (I^2=48.3%; P=.12) across the 3 included studies.

Figure 4. Forest plot displaying random-effects meta-analysis of the effect of digital interventions for young adult cannabis users on cannabis use frequency at 3 months after the baseline relative to a comparison condition measured using the number of self-reported days of use in the previous month through the timeline followback method.

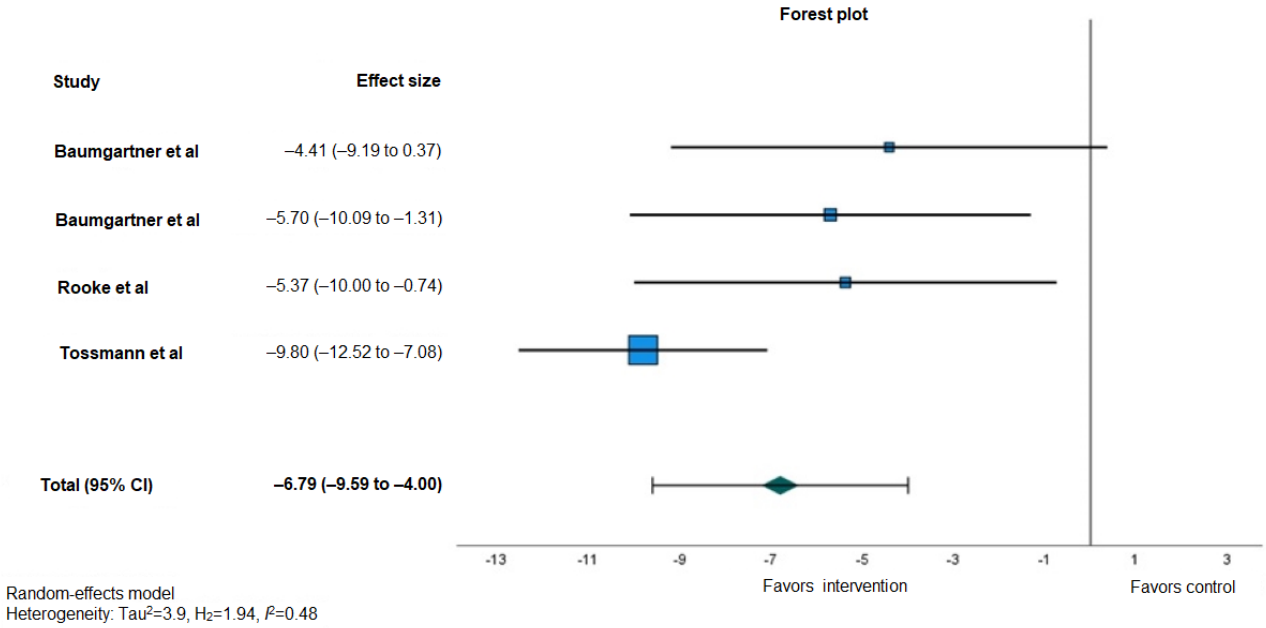

The samples of the 3 studies included in the meta-analysis varied in size from 225 to 1292 participants (mean 697.33, SD 444.11), and the mean age ranged from 24.7 to 31.88 years (mean 26.38, SD 3.58 years). These studies involved 3 different digital interventions and used different design approaches to assess intervention effectiveness. One study assessed the effectiveness of a web-based counseling program (ie, Quit the Shit) against a waitlist control [72], another examined the effectiveness of a fully self-guided web-based treatment program for CU and related problems (ie, Reduce Your Use: How to Break the Cannabis Habit) against a control condition website consisting of basic educational information on cannabis [71], and the third used a 3-arm RCT design to investigate whether the effectiveness of a minimally guided internet-based self-help intervention (ie, CANreduce 2.0) might be enhanced by implementing adherence-focused guidance and emphasizing the social presence factor of a personal e-coach [70].

Discussion

Summary of Principal Findings

The primary aim of this systematic review was to evaluate the effectiveness of digital interventions in addressing CU among community-living young adults. We included 19 randomized controlled studies representing 9 unique digital interventions aimed at preventing, reducing, or ceasing CU and evaluated the effects of 3 different digital interventions on CU. In summary, the 3 digital interventions included in the meta-analysis proved superior to control conditions in reducing the number of days of CU in the previous month at the 3-month follow-up.

Our findings are consistent with those of 2 previous meta-analyses by Olmos et al [43] and Tait et al [44] and with the findings of a recently published umbrella review of systematic reviews and meta-analyses of RCTs [123], all of which revealed a positive effect of internet- and computer-based interventions on CU. However, a recent systematic review and meta-analysis by Beneria et al [45] found that web-based CU

XSL•FO
RenderX

interventions did not significantly reduce CU. Beneria et al [45] included studies with different intervention programs that targeted diverse population groups (both adolescents and young adults) and use of more than one substance (eg, alcohol and cannabis). In our systematic review, a more conservative approach was taken—we focused specifically on young adults and considered interventions targeting CU only. Although our results indicate that digital interventions hold great promise in terms of effectiveness, an important question that remains unresolved is whether there is an optimal exposure dose in terms of both duration and frequency that might be more effective. Among the studies included in this systematic review, interventions varied considerably in terms of the number of psychoeducational modules offered (from 2 to 13), time spent reviewing the material, and duration (from a single session to a 12-week spread period). Our results suggest that an intervention duration of at least 6 weeks yields better results.

Another important finding of this review is that, although almost half (9/19, 47%) of the included studies observed an intervention effect on CU frequency, none reported a statistically significant improvement in cannabis-related negative consequences, which may be considered a more distal indicator. More than half (10/19, 53%) of the included studies investigated this outcome. It seems normal to expect to find an effect on CU frequency given that reducing CU is often the primary objective of interventions and because the motivation of users' is generally focused on changing consumption behavior. It is plausible to think that the change in behavior at the consumption level must be maintained over time before an effect on cannabis-related negative consequences can be observed. However, our results showed that, in all the included studies, cannabis-related negative consequences and change in behavior (CU frequency) were measured at the same time point, namely, 3 months after the baseline. Moreover, Grigsby et al [124] conducted a scoping review of risk and protective factors for CU and suggested that interventions to reduce negative CU consequences should prioritize multilevel methods or strategies "to attenuate the cumulative risk from a combination of psychological, contextual, and social influences."

A secondary objective of this systematic review was to describe the active ingredients used in digital interventions for CU among young adults. The vast majority of the interventions were based on either a theory or an intervention approach derived from theories such as CBT, MI, and personalized feedback. From these theories and approaches stem behavior change strategies or techniques, commonly known as BCTs. *Feedback on behavior*, included in the *feedback monitoring* BCT cluster, was the most common BCT used in the included studies. This specific BCT appears to be a core strategy in behavior change interventions [125,126]. In their systematic review of remotely delivered alcohol or substance misuse interventions for adults, Howlett et al [53] found that *feedback on behavior*, *problem solving*, and *goal setting* were the most frequently used BCTs in the included studies. In addition, this research group noted that the most promising BCTs for alcohol misuse were *avoidance/reducing exposure to cues for behavior*, *pros and cons*, and *self-monitoring of behavior*, whereas 2 very promising strategies for substance misuse in general were *problem solving* and *self-monitoring of behavior*. In our systematic review, in addition to *feedback on behavior*, the 6 most frequently used BCTs in the included studies were *social support*, *pros and cons*, *social comparison*, *problem solving*, *information about social and environmental consequences*, and *information about health consequences*. Although *pros and cons* and *problem solving* were present in all 3 studies of digital interventions included in our meta-analysis, *avoidance/reducing exposure to cues for behavior* was reported in only 5% (1/19) of the articles, and *feedback on behavior* was more frequently used than *self-monitoring of behavior*. However, it should be noted that the review by Howlett et al [53] examined digital interventions for participants with alcohol or substance misuse problems, whereas in this review, we focused on interventions that targeted CU from a harm reduction perspective. In this light, *avoidance/reducing exposure to cues for behavior* may be a BCT better suited to populations with substance misuse problems. Lending support to this, a meta-regression by Garnett et al [127] and a Cochrane systematic review by Kaner et al [128] both found interventions that used *behavior substitution* and *credible source* to be associated with greater reduction in excessive alcohol consumption compared with interventions that used other BCTs.

Beyond the number and types of BCTs used, reflecting on the extent to which each BCT in a given intervention suits (or does not suit) the targeted determinants (ie, behavioral and environmental causes) is crucial for planning intervention programs [26]. It is important when designing digital CU interventions not merely to pick a combination of BCTs that have been associated with effectiveness. Rather, the active ingredients must fit the determinants that the interventionists seek to influence. For example, *action planning* would be more relevant as a BCT for young adults highly motivated and ready to take action on their CU than would *pros and cons*, which aims instead to bolster motivation. Given that more than half of all digital interventions are asynchronous and based on a self-guided approach and do not offer counselor or therapist support, a great deal of motivation is required to engage in intervention and behavior change. Therefore, it is essential that developers consider the needs and characteristics of the targeted population to tailor intervention strategies (ie, BCTs) for successful behavior change (eg, tailored to the participant's stage of change). In most of the digital interventions included in this systematic review, personalization was achieved through feedback messages about CU regarding descriptive norms, motives, risks and consequences, and costs, among other things.

Despite the high number of recent studies conducted in the field of digital CU interventions, most of the included articles in our review (17/19, 89%) reported on the development and evaluation of web-based intervention programs. A new generation of health intervention modalities such as mobile apps and social media has drawn the attention of researchers in the past decade and is currently being evaluated. In this regard, the results from a recently published scoping review [129], which included 5 studies of mobile apps for nonmedical CU, suggested that these novel modes of intervention delivery demonstrated adequate feasibility and acceptability. Nevertheless, the internet remains a powerful and convenient medium for reaching young adults

XSL•FO
RenderX

with digital interventions intended to support safe CU behaviors [123,130].

Quality of Evidence

The GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach [131-133] was used to assess the quality of the evidence reviewed. It was deemed to be moderate for the primary outcome of this review, that is, CU frequency in terms of days of use in the previous month (see the summary of evidence in Multimedia Appendix 11 [70,72]). The direction of evidence was broadly consistent—in all 3 RCT studies [70-72] included in the meta-analysis, participants who received digital CU interventions reduced their consumption compared with those who received no or minimal interventions. The 3 RCTs were similar in that they all involved a web-based, multicomponent intervention program aimed at reducing or ceasing CU. However, the interventions did differ or vary in terms of several characteristics, including the strategies used, content, frequency, and duration. Given the small number of studies included in the meta-analysis, we could not conclude with certainty which intervention components, if any, contributed to the effect estimate observed.

Although inconsistency, indirectness, and imprecision were not major issues in the body of evidence, we downgraded the evidence from high to moderate quality on account of RoB assessments at the outcome level. The 3 RCT studies included in the meta-analysis were rated as having “some concerns” of RoB, mainly due to lack of blinding, which significantly reduced our certainty relative to subjective outcomes (ie, self-reported measures of CU frequency). A positive feature of these digital intervention trials is that most procedures are fully automated, and so there was typically a low RoB regarding randomization procedures, allocation to different conditions, and intervention delivery. It is impossible to blind participants to these types of behavior change interventions, and although some researchers have made attempts to counter the impact of this risk, performance bias is an inescapable issue in RCT studies of this kind. Blinding of intervention providers was not an issue in the 3 RCTs included in the meta-analysis because outcome data collection was automated. However, this same automated procedure made it very difficult to ensure follow‐up. Consequently, attrition was another source of bias in these RCT studies [70-72]. The participants lost to follow-up likely stopped using the intervention. However, there is no way of determining whether these people would have benefited more or less than the completers if they had seen the trial through.

The 3 RCTs included in the meta-analysis relied on subjective self-reported measures of CU at baseline and follow‐up, which are subject to recall and social desirability bias. However, all 3 studies used a well-validated instrument of measurement to determine frequency of CU, the TLFB [66]. This is a widely used, subjective self-report tool for measuring frequency (or quantity) of substance use (or abstinence). It is considered a reliable measure of CU [134,135]. Finally, it should be pointed out that any potential bias related to self‐reported CU frequency would have affected both the intervention and control groups (particularly in cases in which control groups received cannabis‐related information), and thus, it was unlikely to account for differential intervention effects. Moreover, we found RoB due to selective reporting in some studies owing mainly to the absence of any reference to a protocol. Ultimately, these limitations may have biased the results of the meta-analysis. Consequently, future research is likely to further undermine our confidence in the effect estimate we observed and report considerably different estimates.

Strengths and Limitations

Our systematic review and meta-analysis has a number of strengths: (1) we included only randomized controlled studies to ensure that the included studies possessed a rigorous research design, (2) we focused specifically on cannabis (rather than combining multiple substances), (3) we assessed the effectiveness of 3 different digital interventions on CU frequency among community-living young adults, and (4) we performed an exhaustive synthesis and comparison of the BCTs used in the 9 digital interventions examined in the 19 studies included in our review based on the BCTTv1.

Admittedly, this systematic review and meta-analysis has limitations that should be recognized. First, although we searched a range of bibliographic databases, the review was limited to articles published in peer-reviewed journals in English or French. This may have introduced publication bias given that articles reporting positive effects are more likely to be published than those with negative or equivocal results. Consequently, the studies included in this review may have overrepresented the statistically significant effects of digital CU interventions.

Second, only a small number of studies were included in the meta-analyses because many studies did not provide adequate statistical information for calculating and synthesizing effect sizes, although significant efforts were made to contact the authors in case of missing data. Because of the small sample size used in the meta-analysis, the effect size estimates may not be highly reflective of the true effects of digital interventions on CU frequency among young adults. Furthermore, synthesizing findings across studies that evaluated different modalities of web-based intervention programs (eg, fully self-guided vs with therapist guidance) and types of intervention approaches (eg, CBT, MI, and personalized feedback) may have introduced bias in the meta-analytical results due to the heterogeneity of the included studies, although heterogeneity was controlled for using a random-effects model and our results indicated low between-study heterogeneity.

Third, we took various measures to ensure that BCT coding was carried out rigorously throughout the data extraction and analysis procedures: (1) all coders received training on how to use the BCTTv1; (2) all the included articles were read line by line so that coders became familiar with intervention descriptions before initiating BCT coding; (3) the intervention description of each included article was double coded after a pilot calibration exercise with all coders, and any disagreements regarding the presence or absence of a BCT were discussed and resolved with a third party; and (4) we contacted the article authors when necessary and possible for further details on the BCTs they used. However, incomplete reporting of intervention content is a recognized issue [136], which may have resulted in our coding BCTs incorrectly as present or absent. Reliably

XSL•FO
RenderX

specifying the BCTs used in interventions allows their active ingredients to be identified, their evidence to be synthesized, and interventions to be replicated, thereby providing tangible guidance to programmers and researchers to develop more effective interventions.

Finally, although this review identified the BCTs used in digital interventions, our approach did not allow us to draw conclusions regarding their effectiveness. Coding BCTs simply as present or absent does not consider the frequency, intensity, and quality with which they were delivered. For example, it is unclear how many individuals should self - monitor their CU. In addition, the quality of BCT implementation may be critical in digital interventions where different graphics and interface designs and the usability of the BCTs used can have considerable influence on the level of user engagement [137]. In the future, it may be necessary to develop new methods to evaluate the dosage of individual BCTs in digital health interventions and characterize their implementation quality to assess their effectiveness [128,138]. Despite its limitations, this review suggests that digital interventions represent a promising avenue for preventing, reducing, or ceasing CU among community-living young adults.

Conclusions

The results of this systematic review and meta-analysis lend support to the promise of digital interventions as an effective means of reducing recreational CU frequency among young adults. Despite the advent and popularity of smartphones, web-based interventions remain the most common mode of delivery for digital interventions. The active ingredients of digital interventions are varied and encompass a number of clusters of the BCTTv1, but a significant number of BCTs remain underused. Additional research is needed to further investigate the effectiveness of these interventions on CU and key outcomes at later time points. Finally, a detailed assessment of user engagement with digital interventions for CU and understanding which intervention components are the most effective remain important research gaps.

Acknowledgments

The authors would like to thank Bénédicte Nauche, Miguel Chagnon, and Paul Di Biase for their valuable support with the search strategy development, statistical analysis, and linguistic revision, respectively. This work was supported by the Ministère de la Santé et des Services sociaux du Québec as part of a broader study aimed at developing and evaluating a digital intervention for young adult cannabis users. Additional funding was provided by the Research Chair in Innovative Nursing Practices. The views and opinions expressed in this manuscript do not necessarily reflect those of these funding entities.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JC contributed to conceptualization, methodology, formal analysis, writing—original draft, supervision, and funding acquisition. GC contributed to conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, visualization, and project administration. BV contributed to conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, and visualization. PA contributed to conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, visualization, and project administration. GR contributed to conceptualization, methodology, formal analysis, investigation, data curation, and writing—review and editing. GF contributed to conceptualization, methodology, formal analysis, investigation, data curation, and writing—review and editing. DJA contributed to conceptualization, methodology, formal analysis, writing—review and editing, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 237 KB - jmir_v26i1e55031_app1.pdf]

Multimedia Appendix 2
Detailed search strategies for each database.
[PDF File (Adobe PDF File), 1707 KB - jmir_v26i1e55031_app2.pdf]

Multimedia Appendix 3
Population, intervention, comparison, outcome, and study design strategy.
[PDF File (Adobe PDF File), 79 KB - jmir_v26i1e55031_app3.pdf]

XSL•FO
RenderX

Multimedia Appendix 4
Excluded studies and reasons for exclusion.
[PDF File (Adobe PDF File), 82 KB - jmir_v26i1e55031_app4.pdf]

Multimedia Appendix 5
Study and participant characteristics.
[PDF File (Adobe PDF File), 230 KB - jmir_v26i1e55031_app5.pdf]

Multimedia Appendix 6
Description of intervention characteristics in the included articles.
[PDF File (Adobe PDF File), 317 KB - jmir_v26i1e55031_app6.pdf]

Multimedia Appendix 7
Summary of methodological characteristics and major findings of the included studies categorized by intervention name.
[PDF File (Adobe PDF File), 452 KB - jmir_v26i1e55031_app7.pdf]

Multimedia Appendix 8
Behavior change techniques (BCTs) coded in each included study summarized by individual BCT and BCT cluster.
[XLSX File (Microsoft Excel File), 28 KB - jmir_v26i1e55031_app8.xlsx]

Multimedia Appendix 9
Risk-of-bias assessment of each included study for cannabis use and cannabis consequences.
[PDF File (Adobe PDF File), 601 KB - jmir_v26i1e55031_app9.pdf]

Multimedia Appendix 10
Excluded studies and reasons for exclusion from the meta-analysis.
[PDF File (Adobe PDF File), 37 KB - jmir_v26i1e55031_app10.pdf]

Multimedia Appendix 11
Summary of evidence according to the Grading of Recommendations Assessment, Development, and Evaluation tool.
[PDF File (Adobe PDF File), 161 KB - jmir_v26i1e55031_app11.pdf]

References

1. Arnett JJ. The developmental context of substance use in emerging adulthood. J Drug Issues 2005;35(2):235-254. [doi: 10.1177/002204260503500202]
2. Stockings E, Hall WD, Lynskey M, Morley KI, Reavley N, Strang J, et al. Prevention, early intervention, harm reduction, and treatment of substance use in young people. Lancet Psychiatry 2016 Mar;3(3):280-296. [doi: 10.1016/S2215-0366(16)00002-X] [Medline: 26905481]
3. ElSohly MA, Chandra S, Radwan M, Majumdar CG, Church JC. A comprehensive review of cannabis potency in the United States in the last decade. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Jun;6(6):603-606. [doi: 10.1016/j.bpsc.2020.12.016] [Medline: 33508497]
4. Fischer B, Robinson T, Bullen C, Curran V, Jutras-Aswad D, Medina-Mora ME, et al. Lower-Risk Cannabis Use Guidelines (LRCUG) for reducing health harms from non-medical cannabis use: a comprehensive evidence and recommendations update. Int J Drug Policy 2022 Jan;99:103381 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103381] [Medline: 34465496]
5. Rotermann M. What has changed since cannabis was legalized? Health Rep 2020 Feb 19;31(2):11-20 [FREE Full text] [doi: 10.25318/82-003-x202000200002-eng] [Medline: 32073644]
6. Degenhardt L, Stockings E, Patton G, Hall WD, Lynskey M. The increasing global health priority of substance use in young people. Lancet Psychiatry 2016 Mar;3(3):251-264. [doi: 10.1016/S2215-0366(15)00508-8] [Medline: 26905480]
7. Buckner JD, Bonn-Miller MO, Zvolensky MJ, Schmidt NB. Marijuana use motives and social anxiety among marijuana-using young adults. Addict Behav 2007 Oct;32(10):2238-2252 [FREE Full text] [doi: 10.1016/j.addbeh.2007.04.004] [Medline: 17478056]
8. Carliner H, Brown QL, Sarvet AL, Hasin DS. Cannabis use, attitudes, and legal status in the U.S.: a review. Prev Med 2017 Nov;104:13-23 [FREE Full text] [doi: 10.1016/j.ypmed.2017.07.008] [Medline: 28705601]
9. World drug report 2020. United Nations Office on Drugs and Crime. 2020. URL: https://wdr.unodc.org/wdr2020/index2020.html [accessed 2023-11-28]
10. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice, Committee on the Health Effects of Marijuana: An Evidence Review and Research Agenda.

XSL•FO
RenderX

The Health Effects of Cannabis and Cannabinoids: The Current State of Evidence and Recommendations for Research. Washington, DC: The National Academies Press; 2017.
11. Hall WD, Patton G, Stockings E, Weier M, Lynskey M, Morley KI, et al. Why young people's substance use matters for global health. Lancet Psychiatry 2016 Mar;3(3):265-279. [doi: 10.1016/S2215-0366(16)00013-4] [Medline: 26905482]
12. Cohen K, Weizman A, Weinstein A. Positive and negative effects of cannabis and cannabinoids on health. Clin Pharmacol Ther 2019 May;105(5):1139-1147. [doi: 10.1002/cpt.1381] [Medline: 30703255]
13. Memedovich KA, Dowsett LE, Spackman E, Noseworthy T, Clement F. The adverse health effects and harms related to marijuana use: an overview review. CMAJ Open 2018 Aug 16;6(3):E339-E346 [FREE Full text] [doi: 10.9778/cmajo.20180023] [Medline: 30115639]
14. Teeters JB, Armstrong NM, King SA, Hubbard SM. A randomized pilot trial of a mobile phone-based brief intervention with personalized feedback and interactive text messaging to reduce driving after cannabis use and riding with a cannabis impaired driver. J Subst Abuse Treat 2022 Nov;142:108867 [FREE Full text] [doi: 10.1016/j.jsat.2022.108867] [Medline: 36007434]
15. Chan GC, Becker D, Butterworth P, Hines L, Coffey C, Hall W, et al. Young-adult compared to adolescent onset of regular cannabis use: a 20-year prospective cohort study of later consequences. Drug Alcohol Rev 2021 May;40(4):627-636. [doi: 10.1111/dar.13239] [Medline: 33497516]
16. Hall W, Stjepanović D, Caulkins J, Lynskey M, Leung J, Campbell G, et al. Public health implications of legalising the production and sale of cannabis for medicinal and recreational use. Lancet 2019 Oct 26;394(10208):1580-1590. [doi: 10.1016/S0140-6736(19)31789-1] [Medline: 31657733]
17. The health and social effects of nonmedical cannabis use. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/251056 [accessed 2023-11-28]
18. Boumparis N, Loheide-Niesmann L, Blankers M, Ebert DD, Korf D, Schaub MP, et al. Short- and long-term effects of digital prevention and treatment interventions for cannabis use reduction: a systematic review and meta-analysis. Drug Alcohol Depend 2019 Jul 01;200:82-94 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.03.016] [Medline: 31112834]
19. Jutras-Aswad D, Le Foll B, Bruneau J, Wild TC, Wood E, Fischer B. Thinking beyond legalization: the case for expanding evidence-based options for cannabis use disorder treatment in Canada. Can J Psychiatry 2019 Feb;64(2):82-87 [FREE Full text] [doi: 10.1177/0706743718790955] [Medline: 30033757]
20. Garnett CV, Crane D, Brown J, Kaner EF, Beyer FR, Muirhead CR, et al. Behavior change techniques used in digital behavior change interventions to reduce excessive alcohol consumption: a meta-regression. Ann Behav Med 2018 May 18;52(6):530-543 [FREE Full text] [doi: 10.1093/abm/kax029] [Medline: 29788261]
21. Glanz K, Rimer BK, Viswanath K. Health Behavior: Theory, Research, and Practice, 5th Edition. Hoboken, NJ: Jossey-Bass; Jul 2015.
22. Prestwich A, Webb TL, Conner M. Using theory to develop and test interventions to promote changes in health behaviour: evidence, issues, and recommendations. Curr Opin Psychol 2015 Oct;5:1-5. [doi: 10.1016/j.copsyc.2015.02.011]
23. Webb TL, Sniehotta FF, Michie S. Using theories of behaviour change to inform interventions for addictive behaviours. Addiction 2010 Nov;105(11):1879-1892. [doi: 10.1111/j.1360-0443.2010.03028.x] [Medline: 20670346]
24. Cilliers F, Schuwirth L, van der Vleuten C. Health behaviour theories: a conceptual lens to explore behaviour change. In: Cleland J, Durning SJ, editors. Researching Medical Education. Hoboken, NJ: Wiley; 2015.
25. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344 [FREE Full text] [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
26. Eldredge LK, Markham CM, Ruiter RA, Fernández ME, Kok G, Parcel GS. Planning Health Promotion Programs: An Intervention Mapping Approach, 4th Edition. Hoboken, NJ: John Wiley & Sons; Feb 2016.
27. Marlatt GA, Blume AW, Parks GA. Integrating harm reduction therapy and traditional substance abuse treatment. J Psychoactive Drugs 2001;33(1):13-21. [doi: 10.1080/02791072.2001.10400463] [Medline: 11332996]
28. Adams A, Ferguson M, Greer AM, Burmeister C, Lock K, McDougall J, et al. Guideline development in harm reduction: considerations around the meaningful involvement of people who access services. Drug Alcohol Depend Rep 2022 Aug 12;4:100086 [FREE Full text] [doi: 10.1016/j.dadr.2022.100086] [Medline: 36846576]
29. Davis ML, Powers MB, Handelsman P, Medina JL, Zvolensky M, Smits JA. Behavioral therapies for treatment-seeking cannabis users: a meta-analysis of randomized controlled trials. Eval Health Prof 2015 Mar;38(1):94-114 [FREE Full text] [doi: 10.1177/0163278714529970] [Medline: 24695072]
30. Gates PJ, Sabioni P, Copeland J, Le Foll B, Gowing L. Psychosocial interventions for cannabis use disorder. Cochrane Database Syst Rev 2016 May 05;2016(5):CD005336 [FREE Full text] [doi: 10.1002/14651858.CD005336.pub4] [Medline: 27149547]
31. Halladay J, Scherer J, MacKillop J, Woock R, Petker T, Linton V, et al. Brief interventions for cannabis use in emerging adults: a systematic review, meta-analysis, and evidence map. Drug Alcohol Depend 2019 Nov 01;204:107565. [doi: 10.1016/j.drugalcdep.2019.107565] [Medline: 31751868]

XSL•FO
RenderX

32. Imtiaz S, Roerecke M, Kurdyak P, Samokhvalov AV, Hasan OS, Rehm J. Brief interventions for cannabis use in healthcare settings: systematic review and meta-analyses of randomized trials. J Addict Med 2020;14(1):78-88. [doi: 10.1097/ADM.0000000000000527] [Medline: 32012140]
33. Standeven LR, Scialli A, Chisolm MS, Terplan M. Trends in cannabis treatment admissions in adolescents/young adults: analysis of TEDS-A 1992 to 2016. J Addict Med 2020;14(4):e29-e36. [doi: 10.1097/ADM.0000000000000586] [Medline: 31985511]
34. Montanari L, Guarita B, Mounteney J, Zipfel N, Simon R. Cannabis use among people entering drug treatment in europe: a growing phenomenon? Eur Addict Res 2017;23(3):113-121 [FREE Full text] [doi: 10.1159/000475810] [Medline: 28511182]
35. Kerridge BT, Mauro PM, Chou SP, Saha TD, Pickering RP, Fan AZ, et al. Predictors of treatment utilization and barriers to treatment utilization among individuals with lifetime cannabis use disorder in the United States. Drug Alcohol Depend 2017 Dec 01;181:223-228 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.09.032] [Medline: 29107786]
36. Gates P, Copeland J, Swift W, Martin G. Barriers and facilitators to cannabis treatment. Drug Alcohol Rev 2012 May;31(3):311-319. [doi: 10.1111/j.1465-3362.2011.00313.x] [Medline: 21521384]
37. Hammarlund RA, Crapanzano KA, Luce L, Mulligan L, Ward KM. Review of the effects of self-stigma and perceived social stigma on the treatment-seeking decisions of individuals with drug- and alcohol-use disorders. Subst Abuse Rehabil 2018 Nov 23;9:115-136 [FREE Full text] [doi: 10.2147/SAR.S183256] [Medline: 30538599]
38. Bedrouni W. On the use of digital technologies to reduce the public health impacts of cannabis legalization in Canada. Can J Public Health 2018 Dec;109(5-6):748-751 [FREE Full text] [doi: 10.17269/s41997-018-0117-7] [Medline: 30097890]
39. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241 [FREE Full text] [doi: 10.1111/add.15687] [Medline: 34514668]
40. Kazemi DM, Borsari B, Levine MJ, Li S, Lamberson KA, Matta LA. A systematic review of the mHealth interventions to prevent alcohol and substance abuse. J Health Commun 2017 May;22(5):413-432 [FREE Full text] [doi: 10.1080/10810730.2017.1303556] [Medline: 28394729]
41. Nesvåg S, McKay JR. Feasibility and effects of digital interventions to support people in recovery from substance use disorders: systematic review. J Med Internet Res 2018 Aug 23;20(8):e255 [FREE Full text] [doi: 10.2196/jmir.9873] [Medline: 30139724]
42. Hoch E, Preuss UW, Ferri M, Simon R. Digital interventions for problematic cannabis users in non-clinical settings: findings from a systematic review and meta-analysis. Eur Addict Res 2016;22(5):233-242 [FREE Full text] [doi: 10.1159/000445716] [Medline: 27160333]
43. Olmos A, Tirado-Muñoz J, Farré M, Torrens M. The efficacy of computerized interventions to reduce cannabis use: a systematic review and meta-analysis. Addict Behav 2018 Apr;79:52-60. [doi: 10.1016/j.addbeh.2017.11.045] [Medline: 29248863]
44. Tait RJ, Spijkerman R, Riper H. Internet and computer based interventions for cannabis use: a meta-analysis. Drug Alcohol Depend 2013 Dec 01;133(2):295-304. [doi: 10.1016/j.drugalcdep.2013.05.012] [Medline: 23747236]
45. Beneria A, Santesteban-Echarri O, Daigre C, Tremain H, Ramos-Quiroga JA, McGorry PD, et al. Online interventions for cannabis use among adolescents and young adults: systematic review and meta-analysis. Early Interv Psychiatry 2022 Aug;16(8):821-844. [doi: 10.1111/eip.13226] [Medline: 34464502]
46. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Int J Nurs Stud 2013 May;50(5):587-592 [FREE Full text] [doi: 10.1016/j.ijnurstu.2012.09.010] [Medline: 23159157]
47. Michie S, Abraham C, Eccles MP, Francis JJ, Hardeman W, Johnston M. Strengthening evaluation and implementation by specifying components of behaviour change interventions: a study protocol. Implement Sci 2011 Feb 07;6:10 [FREE Full text] [doi: 10.1186/1748-5908-6-10] [Medline: 21299860]
48. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
49. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl Psychol 2008 Oct;57(4):660-680. [doi: 10.1111/j.1464-0597.2008.00341.x]
50. Scott C, de Barra M, Johnston M, de Bruin M, Scott N, Matheson C, et al. Using the behaviour change technique taxonomy v1 (BCTTv1) to identify the active ingredients of pharmacist interventions to improve non-hospitalised patient health outcomes. BMJ Open 2020 Sep 15;10(9):e036500 [FREE Full text] [doi: 10.1136/bmjopen-2019-036500] [Medline: 32933960]
51. Dombrowski SU, Sniehotta FF, Avenell A, Johnston M, MacLennan G, Araújo-Soares V. Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. Health Psychol Rev 2012;6(1):7-32. [doi: 10.1080/17437199.2010.513298]

XSL•FO
RenderX

52. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
53. Howlett N, García-Iglesias J, Bontoft C, Breslin G, Bartington S, Freethy I, et al. A systematic review and behaviour change technique analysis of remotely delivered alcohol and/or substance misuse interventions for adults. Drug Alcohol Depend 2022 Oct 01;239:109597 [FREE Full text] [doi: 10.1016/j.drugalcdep.2022.109597] [Medline: 35963209]
54. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.4. London, UK: The Cochrane Collaboration; 2023.
55. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
56. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
57. Halladay J, Petker T, Fein A, Munn C, MacKillop J. Brief interventions for cannabis use in emerging adults: protocol for a systematic review, meta-analysis, and evidence map. Syst Rev 2018 Jul 25;7(1):106 [FREE Full text] [doi: 10.1186/s13643-018-0772-z] [Medline: 30045752]
58. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
59. Arnett JJ. Emerging adulthood. A theory of development from the late teens through the twenties. Am Psychol 2000 May;55(5):469-480. [Medline: 10842426]
60. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc 2016 Jul;104(3):240-243 [FREE Full text] [doi: 10.3163/1536-5050.104.3.014] [Medline: 27366130]
61. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
62. Presseau J, Ivers NM, Newham JJ, Knittle K, Danko KJ, Grimshaw JM. Using a behaviour change techniques taxonomy to identify active ingredients within trials of implementation interventions for diabetes care. Implement Sci 2015 Apr 23;10:55 [FREE Full text] [doi: 10.1186/s13012-015-0248-7] [Medline: 25900104]
63. Fontaine G, Cossette S, Maheu-Cadotte MA, Deschênes MF, Rouleau G, Lavallée A, et al. Effect of implementation interventions on nurses' behaviour in clinical practice: a systematic review, meta-analysis and meta-regression protocol. Syst Rev 2019 Dec 05;8(1):305 [FREE Full text] [doi: 10.1186/s13643-019-1227-x] [Medline: 31806051]
64. Fontaine G, Cossette S. A theory-based adaptive E-learning program aimed at increasing intentions to provide brief behavior change counseling: randomized controlled trial. Nurse Educ Today 2021 Dec;107:105112. [doi: 10.1016/j.nedt.2021.105112] [Medline: 34455287]
65. Fontaine G, Cossette S. Development and design of E_MOTIV: a theory-based adaptive e-learning program to support nurses' provision of brief behavior change counseling. Comput Inform Nurs 2023 Mar 01;41(3):130-141. [doi: 10.1097/CIN.0000000000000942] [Medline: 35796716]
66. Sobell LC, Sobell MB. Timeline follow-back: a technique for assessing self-reported alcohol consumption. In: Litten RZ, Allen JP, editors. Measuring Alcohol Consumption. Totowa, NJ: Humana Press; 1992.
67. Stephens RS, Roffman RA, Simpson EE. Treating adult marijuana dependence: a test of the relapse prevention model. J Consult Clin Psychol 1994;62(1):92-99. [doi: 10.1037/0022-006x.62.1.92]
68. Harris RJ, Deeks JJ, Altman DG, Bradburn MJ, Harbord RM, Sterne JA. Metan: fixed- and random-effects meta-analysis. Stata J 2008;8(1):3-28. [doi: 10.1177/1536867x0800800102]
69. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
70. Baumgartner C, Schaub MP, Wenger A, Malischnig D, Augsburger M, Walter M, et al. CANreduce 2.0 adherence-focused guidance for internet self-help among cannabis users: three-arm randomized controlled trial. J Med Internet Res 2021 Apr 30;23(4):e27463 [FREE Full text] [doi: 10.2196/27463] [Medline: 33929333]
71. Rooke S, Copeland J, Norberg M, Hine D, McCambridge J. Effectiveness of a self-guided web-based cannabis treatment program: randomized controlled trial. J Med Internet Res 2013 Feb 15;15(2):e26 [FREE Full text] [doi: 10.2196/jmir.2256] [Medline: 23470329]
72. Tossmann HP, Jonas B, Tensil MD, Lang P, Strüber E. A controlled trial of an internet-based intervention program for cannabis users. Cyberpsychol Behav Soc Netw 2011 Nov;14(11):673-679. [doi: 10.1089/cyber.2010.0506] [Medline: 21651419]
73. StataCorp. Stata statistical software: release 18. StataCorp LLC. College Station, TX: StataCorp LLC; 2023. URL: https://www.stata.com/ [accessed 2023-11-28]
74. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]

XSL•FO
RenderX

75. McGuinness LA, Higgins JP. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
76. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022 Mar 27;18(2):e1230 [FREE Full text] [doi: 10.1002/cl2.1230] [Medline: 36911350]
77. Elliott JC, Carey KB. Correcting exaggerated marijuana use norms among college abstainers: a preliminary test of a preventive intervention. J Stud Alcohol Drugs 2012 Nov;73(6):976-980. [doi: 10.15288/jsad.2012.73.976] [Medline: 23036216]
78. Elliott JC, Carey KB, Vanable PA. A preliminary evaluation of a web-based intervention for college marijuana use. Psychol Addict Behav 2014 Mar;28(1):288-293. [doi: 10.1037/a0034995] [Medline: 24731118]
79. Lee CM, Neighbors C, Kilmer JR, Larimer ME. A brief, web-based personalized feedback selective intervention for college student marijuana use: a randomized clinical trial. Psychol Addict Behav 2010 Jun;24(2):265-273 [FREE Full text] [doi: 10.1037/a0018859] [Medline: 20565152]
80. Buckner JD, Zvolensky MJ, Lewis EM. On-line personalized feedback intervention for negative affect and cannabis: a pilot randomized controlled trial. Exp Clin Psychopharmacol 2020 Apr;28(2):143-149. [doi: 10.1037/pha0000304] [Medline: 31204824]
81. Goodness TM, Palfai TP. Electronic screening and brief intervention to reduce cannabis use and consequences among graduate students presenting to a student health center: a pilot study. Addict Behav 2020 Jul;106:106362. [doi: 10.1016/j.addbeh.2020.106362] [Medline: 32120198]
82. Palfai TP, Saitz R, Winter M, Brown TA, Kypri K, Goodness TM, et al. Web-based screening and brief intervention for student marijuana use in a university health center: pilot study to examine the implementation of eCHECKUP TO GO in different contexts. Addict Behav 2014 Sep;39(9):1346-1352 [FREE Full text] [doi: 10.1016/j.addbeh.2014.04.025] [Medline: 24845164]
83. Riggs NR, Conner BT, Parnes JE, Prince MA, Shillington AM, George MW. Marijuana eCHECKUPTO GO: effects of a personalized feedback plus protective behavioral strategies intervention for heavy marijuana-using college students. Drug Alcohol Depend 2018 Sep 01;190:13-19. [doi: 10.1016/j.drugalcdep.2018.05.020] [Medline: 29960918]
84. Bonar EE, Chapman L, Pagoto S, Tan CY, Duval ER, McAfee J, et al. Social media interventions addressing physical activity among emerging adults who use cannabis: a pilot trial of feasibility and acceptability. Drug Alcohol Depend 2023 Jan 01;242:109693. [doi: 10.1016/j.drugalcdep.2022.109693] [Medline: 36442441]
85. Bonar EE, Goldstick JE, Chapman L, Bauermeister JA, Young SD, McAfee J, et al. A social media intervention for cannabis use among emerging adults: randomized controlled trial. Drug Alcohol Depend 2022 Mar 01;232:109345 [FREE Full text] [doi: 10.1016/j.drugalcdep.2022.109345] [Medline: 35144238]
86. Walukevich-Dienst K, Neighbors C, Buckner JD. Online personalized feedback intervention for cannabis-using college students reduces cannabis-related problems among women. Addict Behav 2019 Nov;98:106040. [doi: 10.1016/j.addbeh.2019.106040] [Medline: 31302314]
87. Côté J, Tessier S, Gagnon H, April N, Rouleau G, Chagnon M. Efficacy of a web-based tailored intervention to reduce cannabis use among young people attending adult education centers in Quebec. Telemed J E Health 2018 Nov;24(11):853-860. [doi: 10.1089/tmj.2017.0144] [Medline: 29466093]
88. Cunningham JA, Schell C, Bertholet N, Wardell JD, Quilty LC, Agic B, et al. Online personalized feedback intervention to reduce risky cannabis use. Randomized controlled trial. Internet Interv 2021 Nov 14;26:100484 [FREE Full text] [doi: 10.1016/j.invent.2021.100484] [Medline: 34824985]
89. Copeland J, Rooke S, Rodriquez D, Norberg MM, Gibson L. Comparison of brief versus extended personalised feedback in an online intervention for cannabis users: short-term findings of a randomised trial. J Subst Abuse Treat 2017 May;76:43-48. [doi: 10.1016/j.jsat.2017.01.009] [Medline: 28162849]
90. Jonas B, Tensil MD, Tossmann P, Strüber E. Effects of treatment length and chat-based counseling in a web-based intervention for cannabis users: randomized factorial trial. J Med Internet Res 2018 May 08;20(5):e166 [FREE Full text] [doi: 10.2196/jmir.9579] [Medline: 29739738]
91. Schaub MP, Wenger A, Berg O, Beck T, Stark L, Buehler E, et al. A web-based self-help intervention with and without chat counseling to reduce cannabis use in problematic cannabis users: three-arm randomized controlled trial. J Med Internet Res 2015 Oct 13;17(10):e232 [FREE Full text] [doi: 10.2196/jmir.4860] [Medline: 26462848]
92. Sinadinovic K, Johansson M, Johansson AS, Lundqvist T, Lindner P, Hermansson U. Guided web-based treatment program for reducing cannabis use: a randomized controlled trial. Addict Sci Clin Pract 2020 Feb 18;15(1):9 [FREE Full text] [doi: 10.1186/s13722-020-00185-8] [Medline: 32070417]
93. Kanfer FH. Implications of a self-regulation model of therapy for treatment of addictive behaviors. In: Miller WR, Heather N, editors. Treating Addictive Behaviors. Boston, MA: Springer; 1986:29-47.
94. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]

XSL•FO
RenderX

95. Marlatt GA, Baer JS, Kivlahan DR, Dimeff LA, Larimer ME, Quigley LA, et al. Screening and brief intervention for high-risk college student drinkers: results from a 2-year follow-up assessment. J Consult Clin Psychol 1998 Aug;66(4):604-615. [doi: 10.1037/0022-006x.66.4.604]
96. Miller WR, Rollnick S. Motivational Interviewing: Preparing People for Change. New York, NY: Guilford Press; 2002.
97. Prince MA, Carey KB, Maisto SA. Protective behavioral strategies for reducing alcohol involvement: a review of the methodological issues. Addict Behav 2013 Jul;38(7):2343-2351. [doi: 10.1016/j.addbeh.2013.03.010] [Medline: 23584196]
98. Lundqvist TN. Cognitive Dysfunctions in Chronic Cannabis Users Observed During Treatment: An Integrative Approach. Stockholm, Sweden: Almqvist & Wiksell; 1997.
99. Kanter JW, Puspitasari AJ, Santos MM, Nagy GA. Behavioural activation: history, evidence and promise. Br J Psychiatry 2012 May;200(5):361-363. [doi: 10.1192/bjp.bp.111.103390] [Medline: 22550329]
100. Jaffee WB, D'Zurilla TJ. Personality, problem solving, and adolescent substance use. Behav Ther 2009 Mar;40(1):93-101. [doi: 10.1016/j.beth.2008.03.001] [Medline: 19187820]
101. Miller W, Rollnick S. Motivational Interviewing: Preparing People to Change Addictive Behavior. New York, NY: The Guilford Press; 1991.
102. Gordon JR, Marlatt GA. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. 2nd edition. New York, NY: The Guilford Press; 2005.
103. Platt JJ, Husband SD. An overview of problem-solving and social skills approaches in substance abuse treatment. Psychotherapy (Chic) 1993;30(2):276-283 [FREE Full text]
104. Steinberg KL, Roffman R, Carroll K, McRee B, Babor T, Miller M. Brief counseling for marijuana dependence: a manual for treating adults. Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration, US Department of Health and Human Services. URL: https://store.samhsa.gov/product/brief-counseling-marijuana-dependence-manual-treating-adults/sma15-4211 [accessed 2024-03-23]
105. de Shazer S, Dolan Y. More Than Miracles: The State of the Art of Solution-Focused Brief Therapy. Oxfordshire, UK: Routledge; 2007.
106. Copeland J, Swift W, Roffman R, Stephens R. A randomized controlled trial of brief cognitive-behavioral interventions for cannabis use disorder. J Subst Abuse Treat 2001 Sep;21(2):55-65. [doi: 10.1016/s0740-5472(01)00179-9] [Medline: 11551733]
107. Linke S, McCambridge J, Khadjesari Z, Wallace P, Murray E. Development of a psychologically enhanced interactive online intervention for hazardous drinking. Alcohol Alcohol 2008;43(6):669-674 [FREE Full text] [doi: 10.1093/alcalc/agn066] [Medline: 18693217]
108. Wang ML, Waring ME, Jake-Schoffman DE, Oleski JL, Michaels Z, Goetz JM, et al. Clinic versus online social network-delivered lifestyle interventions: protocol for the get social noninferiority randomized controlled trial. JMIR Res Protoc 2017 Dec 11;6(12):e243 [FREE Full text] [doi: 10.2196/resprot.8068] [Medline: 29229591]
109. Sepah SC, Jiang L, Peters AL. Translating the diabetes prevention program into an online social network: validation against CDC standards. Diabetes Educ 2014 Jul;40(4):435-443 [FREE Full text] [doi: 10.1177/0145721714531339] [Medline: 24723130]
110. Cunningham JA, van Mierlo T. The check your cannabis screener: a new online personalized feedback tool. Open Med Inform J 2009 May 07;3:27-31 [FREE Full text] [doi: 10.2174/1874431100903010027] [Medline: 19587808]
111. Bertholet N, Cunningham JA, Faouzi M, Gaume J, Gmel G, Burnand B, et al. Internet-based brief intervention for young men with unhealthy alcohol use: a randomized controlled trial in a general population sample. Addiction 2015 Nov;110(11):1735-1743 [FREE Full text] [doi: 10.1111/add.13051] [Medline: 26173842]
112. Walker DD, Roffman RA, Stephens RS, Wakana K, Berghuis J, Kim W. Motivational enhancement therapy for adolescent marijuana users: a preliminary randomized controlled trial. J Consult Clin Psychol 2006 Jun;74(3):628-632 [FREE Full text] [doi: 10.1037/0022-006X.74.3.628] [Medline: 16822119]
113. Miller MB, Leffingwell T, Claborn K, Meier E, Walters S, Neighbors C. Personalized feedback interventions for college alcohol misuse: an update of Walters and Neighbors (2005). Psychol Addict Behav 2013 Dec;27(4):909-920 [FREE Full text] [doi: 10.1037/a0031174] [Medline: 23276309]
114. Ajzen I. From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J, editors. Action Control. Berlin, Germany: Springer; 1985:11-39.
115. Dennis M, Titus JC, Diamond G, Donaldson J, Godley SH, Tims FM, et al. The Cannabis Youth Treatment (CYT) experiment: rationale, study design and analysis plans. Addiction 2002 Dec 11;97 Suppl 1(s1):16-34. [doi: 10.1046/j.1360-0443.97.s01.2.x] [Medline: 12460126]
116. Simons JS, Dvorak RD, Merrill JE, Read JP. Dimensions and severity of marijuana consequences: development and validation of the Marijuana Consequences Questionnaire (MACQ). Addict Behav 2012 May;37(5):613-621 [FREE Full text] [doi: 10.1016/j.addbeh.2012.01.008] [Medline: 22305645]
117. Legleye S. The Cannabis Abuse Screening Test and the DSM-5 in the general population: optimal thresholds and underlying common structure using multiple factor analysis. Int J Methods Psychiatr Res 2018 Jun 10;27(2):e1597 [FREE Full text] [doi: 10.1002/mpr.1597] [Medline: 29124816]

XSL•FO
RenderX

118. Legleye S, Karila LM, Beck F, Reynaud M. Validation of the CAST, a general population Cannabis Abuse Screening Test. J Subst Use 2009 Jul 12;12(4):233-242. [doi: 10.1080/14659890701476532]
119. Sobell LC, Sobell MB. Timeline follow back. In: Litten RZ, Allen JP, editors. Measuring Alcohol Consumption: Psychosocial and Biochemical Methods. Totowa, NJ: Humana Press; 1992:41-72.
120. White HR, Labouvie EW, Papadaratsakis V. Changes in substance use during the transition to adulthood: a comparison of college students and their noncollege age peers. J Drug Issues 2016 Aug 03;35(2):281-306. [doi: 10.1177/002204260503500204]
121. White HR, Labouvie EW. Towards the assessment of adolescent problem drinking. J Stud Alcohol 1989 Jan;50(1):30-37. [doi: 10.15288/jsa.1989.50.30] [Medline: 2927120]
122. Cloutier RM, Natesan Batley P, Kearns NT, Knapp AA. A psychometric evaluation of the Marijuana Problems Index among college students: confirmatory factor analysis and measurement invariance by gender. Exp Clin Psychopharmacol 2022 Dec;30(6):907-917 [FREE Full text] [doi: 10.1037/pha0000531] [Medline: 34735206]
123. Guo H, Yang H, Yuan G, Zhu Z, Zhang K, Zhang X, et al. Effectiveness of information and communication technology (ICT) for addictive behaviors: an umbrella review of systematic reviews and meta-analysis of randomized controlled trials. Comput Hum Behav 2023 Oct;147:107843. [doi: 10.1016/j.chb.2023.107843]
124. Grigsby TJ, Lopez A, Albers L, Rogers CJ, Forster M. A scoping review of risk and protective factors for negative cannabis use consequences. Subst Abuse 2023 Apr 07;17:11782218231166622 [FREE Full text] [doi: 10.1177/11782218231166622] [Medline: 37056398]
125. Harkin B, Webb TL, Chang BP, Prestwich A, Conner M, Kellar I, et al. Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. Psychol Bull 2016 Feb;142(2):198-229 [FREE Full text] [doi: 10.1037/bul0000025] [Medline: 26479070]
126. Samdal GB, Eide GE, Barth T, Williams G, Meland E. Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. Int J Behav Nutr Phys Act 2017 Mar 28;14(1):42 [FREE Full text] [doi: 10.1186/s12966-017-0494-y] [Medline: 28351367]
127. Garnett C, Crane D, Brown J, Kaner E, Beyer F, Muirhead C. Behavior Change Techniques Used in Digital Behavior Change Interventions to Reduce Excessive Alcohol Consumption: A Meta-regression. Ann Behav Med May 18 2018;52(6):A. [doi: 10.1093/abm/kax029]
128. Kaner EF, Beyer FR, Muirhead C, Campbell F, Pienaar ED, Bertholet N, et al. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database Syst Rev 2018 Feb 24;2(2):CD004148 [FREE Full text] [doi: 10.1002/14651858.CD004148.pub4] [Medline: 29476653]
129. Sedrati H, Belrhiti Z, Nejjari C, Ghazal H. Evaluation of mobile health apps for non-medical cannabis use: a scoping review. Procedia Comput Sci 2022;196:581-589. [doi: 10.1016/j.procs.2021.12.052]
130. Curtis BL, Ashford RD, Magnuson KI, Ryan-Pettes SR. Comparison of smartphone ownership, social media use, and willingness to use digital interventions between generation Z and millennials in the treatment of substance use: cross-sectional questionnaire study. J Med Internet Res 2019 Apr 17;21(4):e13050 [FREE Full text] [doi: 10.2196/13050] [Medline: 30994464]
131. Schünemann H, Brożek J, Guyatt G, Oxman A. The GRADE Handbook. London, UK: The Cochrane Collaboration; 2013.
132. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011 Apr;64(4):380-382. [doi: 10.1016/j.jclinepi.2010.09.011] [Medline: 21185693]
133. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
134. Hjorthøj CR, Hjorthøj AR, Nordentoft M. Validity of Timeline Follow-Back for self-reported use of cannabis and other illicit substances--systematic review and meta-analysis. Addict Behav 2012 Mar;37(3):225-233. [doi: 10.1016/j.addbeh.2011.11.025] [Medline: 22143002]
135. Robinson SM, Sobell LC, Sobell MB, Leo GI. Reliability of the Timeline Followback for cocaine, cannabis, and cigarette use. Psychol Addict Behav 2014 Mar;28(1):154-162. [doi: 10.1037/a0030992] [Medline: 23276315]
136. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
137. Garrett JJ. The Elements of User Experience: User-Centered Design for the Web and Beyond. London, UK: Pearson Education; 2010.
138. Lorencatto F, West R, Bruguera C, Brose LS, Michie S. Assessing the quality of goal setting in behavioural support for smoking cessation and its association with outcomes. Ann Behav Med 2016 Apr 24;50(2):310-318 [FREE Full text] [doi: 10.1007/s12160-015-9755-7] [Medline: 26603626]

Abbreviations

BCT: behavior change technique

XSL•FO
RenderX

BCTTv1: Behavior Change Technique Taxonomy version 1
CBT: cognitive behavioral therapy
CU: cannabis use
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MI: motivational interviewing
PICO: population, intervention, comparator, and outcome
PFI: personalized feedback intervention
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB: risk of bias
TLFB: Timeline Followback

Edited by T Leung, G Eysenbach; submitted 30.11.23; peer-reviewed by H Sedrati; comments to author 02.01.24; revised version received 09.01.24; accepted 08.03.24; published 17.04.24.

Please cite as:
Côté J, Chicoine G, Vinette B, Auger P, Rouleau G, Fontaine G, Jutras-Aswad D
Digital Interventions for Recreational Cannabis Use Among Young Adults: Systematic Review, Meta-Analysis, and Behavior Change Technique Analysis of Randomized Controlled Studies
J Med Internet Res 2024;26:e55031
URL: https://www.jmir.org/2024/1/e55031
doi:10.2196/55031
PMID:38630515

©José Côté, Gabrielle Chicoine, Billy Vinette, Patricia Auger, Geneviève Rouleau, Guillaume Fontaine, Didier Jutras-Aswad. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Mobile Apps to Support Mental Health Response in Natural Disasters: Scoping Review

Nwamaka Alexandra Ezeonu[1], MBBS, MSc, MBA; Attila J Hertelendy[2,3], BSc, MHS, MSc, PhD; Medard Kofi Adu[4], BSc, MSc; Janice Y Kung[5], BCom, LMIS; Ijeoma Uchenna Itanyi[1,6,7], MBBS, MPH; Raquel da Luz Dias[4], BSc, MSc, PhD; Belinda Agyapong[8], HDip, BSc, MEd; Petra Hertelendy[9], BS; Francis Ohanyido[10], MBBS, MBA, MPH; Vincent Israel Opoku Agyapong[4], BSc, PGD, MBChB, MSc, MD, PhD; Ejemai Eboreime[4], MBBS, MSc, PhD

[1]Center for Translation and Implementation Research, College of Medicine, University of Nigeria, Nsukka, Nigeria
[2]Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, FL, United States
[3]Department of Emergency Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, United States
[4]Department of Psychiatry, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada
[5]Geoffrey and Robyn Sperber Health Sciences Library, University of Alberta, Edmonton, AB, Canada
[6]Department of Community Medicine, University of Nigeria, Enugu, Nigeria
[7]Department of Public Health Sciences, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[8]Department of Psychiatry, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada
[9]Department of Psychology, Florida State University, Tallahassee, FL, United States
[10]West African Institute of Public Health, Abuja, Nigeria

Corresponding Author:
Ejemai Eboreime, MBBS, MSc, PhD
Department of Psychiatry
Faculty of Medicine
Dalhousie University
5909 Veterans' Memorial Lane
8th Floor Abbie J Lane Memorial Building, QEII Health Sciences Centre
Halifax, NS, B3H 2E2
Canada
Phone: 1 9024732479
Email: ejemai.eboreime@nshealth.ca

Abstract

Background: Disasters are becoming more frequent due to the impact of extreme weather events attributed to climate change, causing loss of lives, property, and psychological trauma. Mental health response to disasters emphasizes prevention and mitigation, and mobile health (mHealth) apps have been used for mental health promotion and treatment. However, little is known about their use in the mental health components of disaster management.

Objective: This scoping review was conducted to explore the use of mobile phone apps for mental health responses to natural disasters and to identify gaps in the literature.

Methods: We identified relevant keywords and subject headings and conducted comprehensive searches in 6 electronic databases. Studies in which participants were exposed to a man-made disaster were included if the sample also included some participants exposed to a natural hazard. Only full-text studies published in English were included. The initial titles and abstracts of the unique papers were screened by 2 independent review authors. Full texts of the selected papers that met the inclusion criteria were reviewed by the 2 independent reviewers. Data were extracted from each selected full-text paper and synthesized using a narrative approach based on the outcome measures, duration, frequency of use of the mobile phone apps, and the outcomes. This scoping review was reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews).

Results: Of the 1398 papers retrieved, 5 were included in this review. A total of 3 studies were conducted on participants exposed to psychological stress following a disaster while 2 were for disaster relief workers. The mobile phone apps for the interventions included Training for Life Skills, Sonoma Rises, Headspace, Psychological First Aid, and Substance Abuse and Mental Health Services Administration (SAMHSA) Behavioural Health Disaster Response Apps. The different studies assessed the effectiveness

XSL•FO
RenderX

or efficacy of the mobile app, feasibility, acceptability, and characteristics of app use or predictors of use. Different measures were used to assess the effectiveness of the apps' use as either the primary or secondary outcome.

Conclusions: A limited number of studies are exploring the use of mobile phone apps for mental health responses to disasters. The 5 studies included in this review showed promising results. Mobile apps have the potential to provide effective mental health support before, during, and after disasters. However, further research is needed to explore the potential of mobile phone apps in mental health responses to all hazards.

(J Med Internet Res 2024;26:e49929) doi:10.2196/49929

KEYWORDS

mental health; disasters; mobile health; mHealth; application; applications; app; apps; smartphone; stress; psychological; traumatic; disaster; disasters; hazard; hazards; emergency; psychological trauma; mobile apps; trauma; scoping; review methods; review methodology; mobile phone

Introduction

Rising global average temperatures and associated changes in weather patterns result in extreme weather events that include hazards such as heatwaves, wildfires, hurricanes, floods, and droughts [1]. These extreme events linked to climate change are resulting in overlapping and so-called cascading disasters leading to record numbers of "billion dollar" disasters with significant losses of lives and property [2,3]. In 2021 alone, approximately 10,000 fatalities caused by disasters were reported globally, while the economic loss was estimated at approximately US $343 billion [4]. Disasters are predicted to become more recurring as a result of the impact of human activities such as burning fossil fuels and deforestation, which release greenhouse gases into the atmosphere that trap heat and cause global temperatures to rise [5].

These catastrophes can adversely affect physical health, mental health, and well-being in both the short and long term as a result of changes due to the political and socioeconomic content, evacuations, social disruption, damage to health care facilities, and financial losses [6-10]. It is estimated that about 33% of people directly exposed to natural disasters will experience mental health sequelae such as posttraumatic stress disorders (PTSDs), anxiety, and depression, among others [11,12].

There is growing recognition of the importance of incorporating mental health into medical and emergency aspects of disaster response [12,13]. However, in contrast to most medical response strategies that are largely curative, mental health response to disasters is predicated on the principles of preventive medicine, thus, emphasizing health promotion, disaster prevention, preparedness, and mitigation [14]. The strategies of mental health response span across primary prevention (mitigating the risk of ill health before it develops), secondary prevention (early detection and intervention), and tertiary prevention (managing established ailment and averting further complications) [15].

Mobile health (mHealth) technology has shown great promise in mental health and has been applied across the 3 levels of prevention [16-20]. For example, SMS text messaging and mobile apps have been developed to promote mental health awareness among young people and older adults (primary prevention) [21]. Additionally, during the COVID-19 pandemic, mHealth was deployed at the population level in Canada to screen for symptoms of anxiety and depression (secondary prevention) [22]. In addition, mHealth interventions were deployed to support first responders and essential workers during the pandemic [23,24]. Further, the technology has been deployed for therapeutic purposes in patients diagnosed with mental health conditions while simultaneously providing support against complications such as suicidal ideation (tertiary prevention) [25].

Although videoconferencing and phone calls can be used for mental health conditions, mobile apps provide more mobility and accessibility, are interactive, more adaptable to users' routines, and can be used repeatedly [26,27]. While numerous academic studies have been conducted on the app of mHealth in the preventive and curative management of mental health conditions in clinical, community, and public health settings, including epidemic response and control, little is known about the use of mobile apps in the mental health components of natural disaster management. This scoping review aims to fill this gap in the literature by mapping where and how mobile apps have been used as part of natural disaster mental health response strategies.

Methods

Overview

This scoping review was reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [28]. The PRISMA-ScR checklist is available in Multimedia Appendix 1. The protocol was not registered.

Search Strategy

A medical librarian (JYK) collaborated with the research team to identify relevant keywords and subject headings for the review, such as mHealth or m-health; mobile health or mobile applications; public health emergency, disaster, or catastrophe; and flood, earthquake, or hurricane. Equipped with this knowledge, the librarian developed and executed comprehensive searches in 6 electronic databases, including Ovid MEDLINE, Ovid Embase, APA PsycInfo, CINAHL, Scopus, and Web of Science Core Collection. The search was conducted on June 30, 2022, and was limited to the English language. The full search strategies are available in Multimedia Appendix 2.

XSL•FO
RenderX

Inclusion and Exclusion Criteria

We included papers that applied mobile apps for mental health responses to disasters. Papers were included if the study participants were persons affected by a natural disaster (setting), the intervention included using a mobile phone app, and the outcome included the assessment of a mental health problem. Studies in which participants were exposed to a man-made disaster were included if the sample also included some participants exposed to a natural disaster. The mental health conditions included were stress, anxiety, depression, and PTSD. Only full-text studies published in English were included. Studies that did not include any intervention with a mobile app for mental health, those focused on videoconferencing or phone calls, and papers on protocols, trial registration, or review were excluded.

Selection of Studies

The search identified papers that were retrieved from the databases. After removing duplicates, the initial titles and abstracts of the unique papers were screened by 2 independent review authors based on the inclusion criteria in a web-based tool called Covidence (Veritas Health Innovation Ltd) [29]. Full texts of the selected papers that met the inclusion criteria were reviewed by the 2 independent reviewers. The research team resolved disagreements through discussion. The bibliographies from the included studies were also reviewed to identify additional studies for inclusion.

Data Extraction and Synthesis

Data from each selected full-text paper were extracted into a data extraction form developed by the research team. The data included the author and year of publication, country of study, study design, number of participants, type of natural disaster, name of the mobile app, duration of use of the app, outcome measures, and the study's findings. These data were synthesized using a narrative approach based on the outcome measures, the duration, frequency of use of the mobile apps, and the outcomes.

Results

Search Results

Of the 1532 papers retrieved from the searches, 976 unique papers had their titles and abstracts screened after deduplication. A total of 38 papers were moved to full-text screening, and data were extracted from 5 papers [30-34] (Figure 1). Table 1 shows the summary of the details of the papers.

XSL•FO
RenderX

Figure 1. PRISMA diagram. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

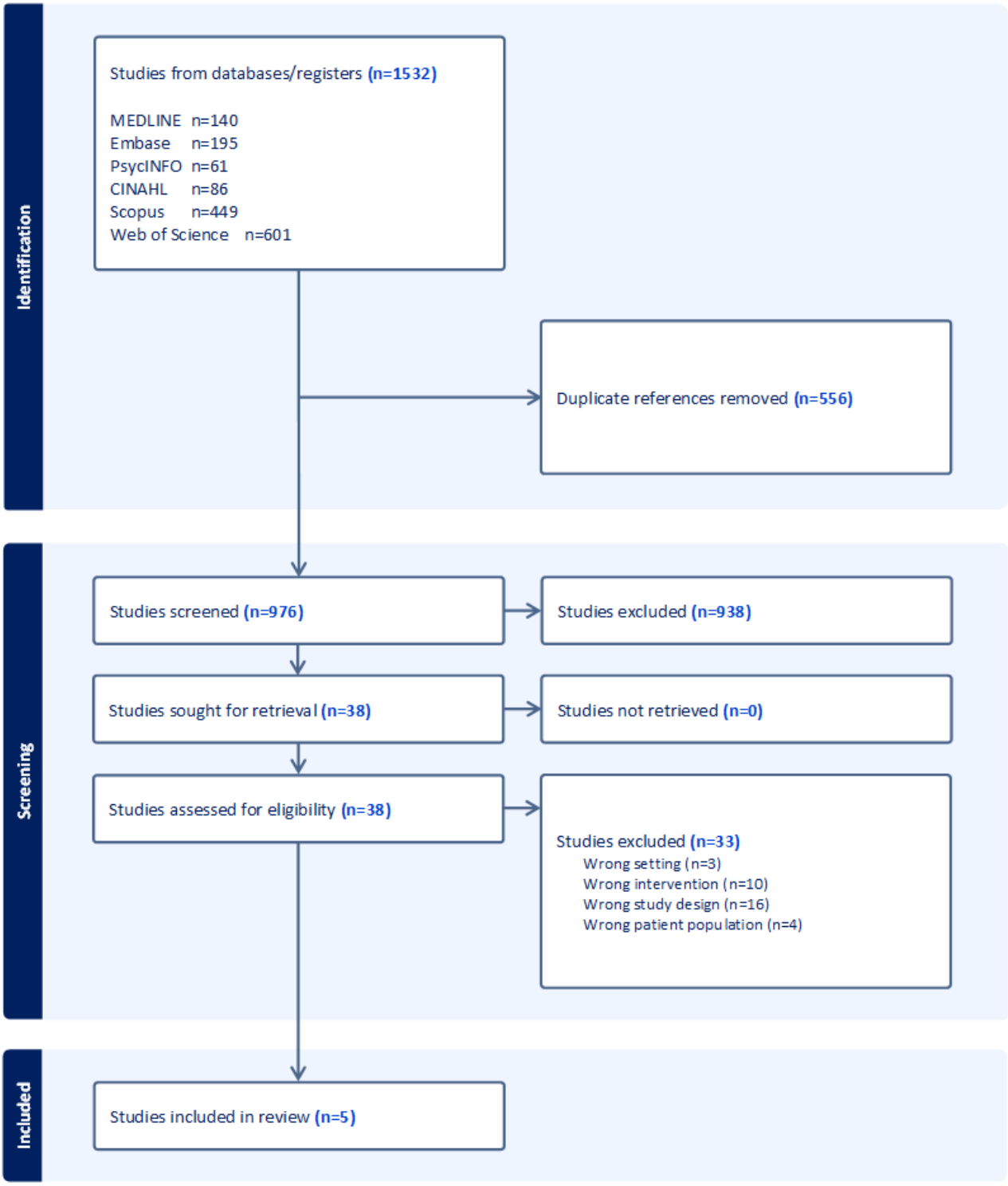

XSL•FO
RenderX

Table 1. Summary of studies using mobile phone apps for mental health response in disasters.

References	Country of study	Type of natural disaster	Mental health conditions targeted	Target population	Name of app (if available)	Intervention (app content)	Conclusion
Choi et al [32]	South Korea	Mixed: earthquakes—2 participants, fires—3 participants, hazardous chemical—4 participants, and traffic accidents—11 participants.	Post-traumatic stress, 1 year after a natural, social, or artificial disaster.	Disaster survivors	TLS[a] apps	• Information: self-assessment (PTSD[b], depression, anxiety, and sleep disorder), postdisaster responses, coping mechanisms (anger, sleep disorder, and addiction) and guidance on how to use information from specialized institutions. • Psychological healing: breathing, butterfly hug, meditation, positive affirmations, healing music, 108 bows, yoga, and writing. • Ventilation and diversion (mood change category): find the same picture game, fruit slot machine, bubble shot, coloring book, and Tetris.	The TLS mobile app was effective for increasing positive and decreasing negative psychological factors according to app usage time and is expected to provide psychological stability and to provide the appropriate conditions to enable effective self-control and management for disaster survivors.
Heinz et al [30]	United States	Wildfires	PTSD and post-disaster distress	Disaster survivors	Sonoma Rises	• The app includes 6 self-paced content sections, psychoeducation, and direct connections to free and local mental health care services. Participants completed daily ratings of anxiety and fear; weekly measures of post-traumatic stress symptoms; internalizing and externalizing symptoms; psychosocial functioning; and then measures of anxiety, depression, well-being, sleep, academic engagement, and perceived social support; and quantitative and qualitative measures of intervention satisfaction and feasibility.	Findings from this study suggest that adolescent disaster survivors are willing and able to use digital health tools and find them beneficial.
Rung et al [31]	United States	Hurricanes in 2005 and Deepwater Horizon oil spill in 2010	Stress prevention and management	Disaster survivors	The Headspace mobile mindfulness app	• A standardized meditation program on MBSR[c] used for at least 30 days, 10 minutes at a time.	The program was easy and cost-effective to implement and acceptable to those who participated, but few women elected to try it. Several short-term benefits of the program were identified, particularly for depression and sleep.

XSL•FO
RenderX

References	Country of study	Type of natural disaster	Mental health conditions targeted	Target population	Name of app (if available)	Intervention (app content)	Conclusion
Choi et al [33]	South Korea	Disaster simulation training for responders in flood, fire, or leakage of hazardous chemicals	Traumatic stress	Disaster relief workers	PFA[d] app	• PFA preparation: self-assessment for relief worker, precautions for general and vulnerable group, precautions by disaster types. • PFA implementation: forming therapeutic relationship; information gathering; and assessment of psychological status, skills for PFA, and linkage to community resources. • Psychological healing: response after disaster relief, stress management, and institutions for support	The use of PFA mobile app during disaster simulation training can function as a new framework for providing disaster relief teaching and methodical mental health services to survivors at the disaster site by disaster health care staff.
Seligman et al [34]	United States	Any disaster	Stress prevention and management	Disaster relief workers	SAMHSA[e] Behavioral Health Disaster Response App	The app provides access to trauma and disaster-related behavioral health resources available on the web or in hard copyright on a smartphone.	The SAMHSA disaster app assures that responders feel confident at having the best disaster behavior health resources.

[a]TLS: Training for Life Skills.

[b]PTSD: posttraumatic stress disorder.

[c]MBSR: Mindfulness-Based Stress Reduction.

[d]PFA: Psychological First Aid.

[e]SAMHSA: Substance Abuse and Mental Health Services Administration.

Characteristics of Included Studies

Of the 5 studies included in this review, 3 (60%) were conducted in the United States [30,31,34], while 2 (40%) were conducted in South Korea [32,33]. All studies used different study designs. A total of 3 studies used a quasi-experimental design—the first, a single group postexperiment with 22 participants [32]; the second, a multiple-baseline single case experimental design with 7 participants [30], while the third study used a 1-group pre- and posttest design with 318 participants [31]. The Training for Life Skills (TLS) app study had only a posttest following the use of the app [32]; the other 2 had baseline and follow-up measurements with the Sonoma Rises app study having, in addition, preintervention and postintervention measurements. The Psychological First Aid (PFA) study was designed as a qualitative study, while the Substance Abuse and Mental Health Services Administration (SAMHSA) study used a mixed methods descriptive design.

Characteristics of the Population

The TLS, Sonoma, and Headspace apps were designed for disaster survivors, while the PFA and SAMHA apps were designed to support disaster relief workers. The TLS app study was administered to adults with a median age of 32 years. Participants of the Sonoma Rises app study had a mean age of 16 (SD 0.98) years, while participants of the Headspace app study had a mean age of 46.1 (SD 10) years. The TLS app study focused on all types of disasters; the Sonoma Rises study focused on adolescents exposed to wildfires, while the Headspace app focused on women who experienced hurricanes and deep-water oil spillage. The PFA study involved 19 disaster health care workers who first underwent disaster simulation training using the mobile app.

Characteristics of the Mobile App Interventions

The included studies revealed several mobile phone apps used as interventions. The first, the TLS app, was used as a psychological first aid program for disaster survivors with content on information, psychological healing, and mood change [32]. The second was the Sonoma Rises app, a Health Insurance Portability and Accountability Act (HIPAA)–compliant, cloud-based mobile app with daily push notifications as reminders designed to help survivors of wildfires or other disasters to find their new routines, build resilience, and increase well-being. The app included 6 self-paced content sections, psychoeducation, and direct connections to free and local mental health care services. The third was the Headspace app for a mindfulness-based stress reduction program that included a series consisting of 10 sessions designed to be used for about 10 minutes per day. The SAMHSA Disaster App equips behavioral health providers to respond to all kinds of traumatic incidents by enabling them to readily access disaster-specific information and other important materials directly on their mobile devices [34]. The PFA mobile app provided evidence-based information and tools for disaster workers to prepare for, execute, and recover from providing psychological first aid during disasters. Accessibility via smartphones and the inclusion of multimedia interventions and assessments tailored

XSL•FO
RenderX

for disaster contexts were key features enabling its use integrated with the simulation training [33].

Frequency and Duration of App Use

The 3 survivor-based apps had variations in the duration of the intervention (app use), which were 8 weeks, at least 5 times a week, frequency of use per day not specified [32]; 4 weeks for 10 minutes per day [30]; and 6 weeks for 5-10 minutes per day [31]. Both the TLS app and the Sonoma Rises app studies had weekly follow-up assessments. The different interventions were applied at least a year following the disasters. Participants in the Sonoma Rises app study used the app on an average of 17 (SD 8.92) days and visited the app an average of 43.50 (SD 30.56) times, with an average session lasting 56.85 (SD 27.87) seconds. The mean time spent on the app was 35.77 (SD 30.03) minutes, while for the TLS app study, the median time spent on the app over the 8 weeks of use was 200-399 minutes. Participants used the Headspace app an average of 24 (SD 36) days and logged in an average of 36 (SD 80) times. There was no description of the frequency and duration of use for the relief worker apps.

Effectiveness Outcomes

Effectiveness outcomes refer to the effects or impact of an intervention or program on the intended outcomes or goals. Different measures were used to assess the effectiveness of the apps' use as either the primary or secondary outcome. Emotional quotients (emotional stability), basic rhythm quotients (brain stability), alpha-blocking rates (increased positive mood), and brain quotients assessed using electroencephalogram (EEG)–measured brainwave activities adjusted for self-reported app use time were used in the TLS app study [32]. The Headspace app study assessed effectiveness using a combination of measures such as trait mindfulness using a 15-item Mindful Attention Awareness Scale (MAAS)—trait version; depressive symptoms using the Center for Epidemiologic Studies Depression Scale-10 (CESD-10); perceived stress with the Perceived Stress Scale, 4-item version (PSS-4); and sleep quality using the Pittsburgh Sleep Quality Index (PSQI) [31]. The Sonoma Rises app study measured efficacy using daily ratings of anxiety and fear, weekly measures of post-traumatic stress symptoms using the Child PTSD Symptom Scale (CPSS-5) for *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*), internalizing and externalizing symptoms using the Behaviour and Feelings Survey (BFS), psychosocial functioning using the Ohio Scale for Youth—Functioning subscale (OSY), and measures of anxiety (Generalized Anxiety Disorder-7 [GAD-7]), depression (Patient Health Questionnaire-9 [PHQ-9]), well-being—Warwick-Edinburgh Mental Well-being Scale (WEMWBS), sleep (Insomnia—Severity Index [ISI]), academic engagement (Student Engagement Instrument [SEI]), and perceived social support (Wills' Social Support Scale [WSSS]) [30].

All 3 survivor-based apps were found to have positive benefits in addressing mental health issues among persons exposed to natural disasters. The TLS mobile app was shown to be effective in increasing positive and decreasing negative psychological factors according to app use time. The TLS mobile apps' use had a significant effect on the emotional quotients (β=.550; $P<.008$), explanatory power (EP) was 30%, had a significant positive effect on the basic rhythm quotient (left brain: β=.598; $P<.003$; EP 35; right brain: β=.451; $P<.035$; EP 20%). Additionally, it had a significant positive effect on the alpha-blocking rate (left brain: β=.510; $P<.015$; EP 26%; right brain: β=.463; $P<.035$, EP 21%); and a significant positive effect on the brain quotient (β=.451; $P<.035$; EP 20%) [16]. The Headspace app had a positive effect on depression (odds ratio [OR] 0.3, 95% CI 0.11-0.81), physical activity (OR 2.8, 95% CI 1.0-7.8), sleep latency (OR 0.3, 95% CI 0.11-0.81), sleep duration (OR 0.3, 95% CI 0.07-0.86), and sleep quality (OR 0.1, 95% CI 0.02-0.96); however, there was no change in mindfulness scores from baseline to follow-up. For the Sonoma Rises app, no significant effects were observed for the clinical and functional outcomes because the longitudinal part of the study was affected by limited statistical power as a result of small sample size and historical confounds that made the participants miss data submission. However, visual inspection of individual data following the intervention showed downward trends across the study phases for daily levels of anxiety, fearfulness, and individual posttraumatic stress symptom severity.

For the PFA app, the qualitative study explored disaster health workers' experiences with simulation training using focus group discussions. A total of 19 participants engaged in disaster scenarios with standardized patients, using a PFA app for guidance. Workers valued the practical educational approach, felt increased self-efficacy to support survivors, and identified areas for enhancing simulations and app tools to optimize effectiveness.

Implementation Outcomes

Implementation outcomes refer to the effects of an intervention or program implementation on various aspects of the implementation process, such as the fidelity of implementation, acceptability, adoption, feasibility, and maintainability. In the papers reviewed, feasibility was assessed using enrollment, program participation, and retention. Acceptability was measured using how well participants liked the app using a rating scale, how much of the app program was completed, the biggest barriers, and whether the app would be recommended to others. Data on characteristics of app use (engagement) were measured using the total number of log ins, average log ins per program completer, platform used (iOS, Android, or web-based), day of week of use (weekday vs weekend), and time of day of use (in 4-hour blocks) [30,31].

The Headspace app was reported to be cost-effective to implement and easy to use [31]. For engagement, only 14% (43/318) of the enrolled women used the app. The level of engagement with the app was high, with 72% (31/43) of participants completing some or all the sessions. Retention was also high with 74% (32/43) of the participants completing the follow-up survey. Lack of time was cited as the main barrier to using the app for 37% (16/43) of users and 49% (94/193) of nonusers. The majority of the users (32/43, 74%) reported high levels of satisfaction with the app. Acceptability was also high, with most participants (32/43, 74%) reporting that they liked the app and 86% (37/43) reporting that they would recommend

XSL•FO
RenderX

it to others. Characteristics of app use showed that of the 1530 log ins, most participants (n=1191, 78%) used the iOS platform, mainly on weekdays (n=1147, 75%) and at different times of day mostly from noon to 4 PM (n=375, 25%).

Sonoma Rises was found to be feasible in terms of engagement and satisfaction among teens with high levels of disaster-related posttraumatic stress symptoms [30]. The self-assessment and data visualization features of the Sonoma Rises app strongly appealed to all the participants, and they were willing to recommend the app to their friends. Self-satisfaction with the mobile app was rated as extremely high (mean 8.50, SD 0.58, on a scale of 0 to 10, with 10 as totally satisfied). The participants agreed or strongly agreed to recommend this intervention to a friend. The participants found the intervention helpful (mean 2, SD 0.82); had the content, functions, and capabilities they needed (mean 3, SD 1.12); and were satisfied with how easy it was to use the app (mean 2, SD 0), on a scale of 1 to 5 with 1 as strongly agree and 5 as strongly disagree. In the qualitative feedback, to make the use of the app better, the participants suggested more notifications to return to the app and the use of the app immediately after a disaster. Implementation outcome was not an objective of the TLS app, hence, none was reported.

Other Mobile Apps With Potential Use in Disasters

Some mobile apps not meeting the inclusion criteria showed promise for supporting mental health in disasters. PTSD Coach provides tools for managing PTSD symptoms [35]. Though not disaster-specific, its psychoeducation, symptom tracking, and coping strategies could aid survivors. Similarly, COVID Coach was designed to help manage pandemic-related stress and anxiety [36]. These apps are summarized in Table 2.

Table 2. Summary of other studies using mobile phone apps for mental health responses.

References	Country of study	Type of natural disaster	Mental health conditions targeted	Name of app (if available)	Intervention (app content)	Conclusion
Kuhn et al [35]	Australia, Canada, Netherlands, Germany, Sweden, and Denmark	None. Designed for general use in PTSD[a], but could have potential use in disaster response.	PTSD	PTSD Coach	• The learn section: psychoeducational information about, professional care, and PTSD and the family. • Track symptoms: PTSD checklist and interpretive feedback. • Manage symptoms: coping tools for acute distress related to PTSD symptoms. • Get support: access to crisis support resources.	PTSD Coach was found to be acceptable and perceived to be helpful for managing PTSD symptoms by veterans and community samples. It also demonstrated improvements in depression symptoms and psychosocial functioning.
Jaworski et al [36]	United States	None. Used in COVID-19 pandemic.	Stress and anxiety	COVID Coach	• Tools for coping with stress and challenging situations, psychoeducation, tracking of mental health symptoms, and access to networks and crisis support. • Tools for managing symptom adapted for life through the pandemic. • Psychoeducational topics about managing COVID-19–related issues.	The use of COVID Coach suggests that mobile apps might have the spread and accessibility needed to act as a useful medium for distributing mental health information and resources to persons enduring COVID-19 pandemic–related stress.

[a]PTSD: posttraumatic stress disorder.

Discussion

Principal Findings

This review sought to identify and map the use of mobile apps for the mental health component of natural disaster management. We found only 5 studies meeting the inclusion criteria. The scarcity of published literature in this area suggests that mobile apps have not been extensively used in mental health responses to natural disasters. Academic studies on the public's use of mobile technologies in disaster management are still nascent [37], but there has been increased interest in developing and deploying digital technology and mobile apps by governments and nonstate actors as part of disaster preparedness and response [38,39]. A recent systematic review found that there is a lack of mental health preparedness in most countries when it comes to disasters [40]. The 5 studies included in our scoping review confirmed this gap and further demonstrated that mobile apps can provide mental health support to disaster-affected individuals and communities. The studies found that the use of mobile apps was associated with improvements in mental health outcomes, such as decreased anxiety and depression symptoms and increased resilience. The reviewed studies also suggest that mobile apps can be effective in delivering psychoeducation and coping skills training to disaster-affected individuals. A 2017 scoping review found that mobile apps have been largely used for communication purposes in disaster management [37]. The

XSL•FO
RenderX

scope of use was classified into 5 categories which are not mutually exclusive. These categories are (1) crowdsourcing (organize and collect disaster-related data from the crowd), (2) collaborating platforms (serve as a platform for collaboration during disasters), (3) alerting and information (disseminate authorized information before and during disasters), (4) collating (gather, filter, and analyze data to build situation awareness), and (5) notifying (for users to notify others during disasters) [37].

Some authors classify disaster response into 3 phases: preparedness, response, and mitigation [41]. The studies included in this review exclusively examined the use of mobile apps during the recovery phase of disaster management. However, none of the studies explored the potential of mobile apps during the preparedness or response phases of disaster management. By addressing this gap, future research could help to provide more comprehensive and effective strategies for the use of mobile apps throughout all phases of disaster management. Examples of potential opportunities are demonstrated in Figure 2.

Figure 2. Potential opportunities for mobile apps use in disasters.

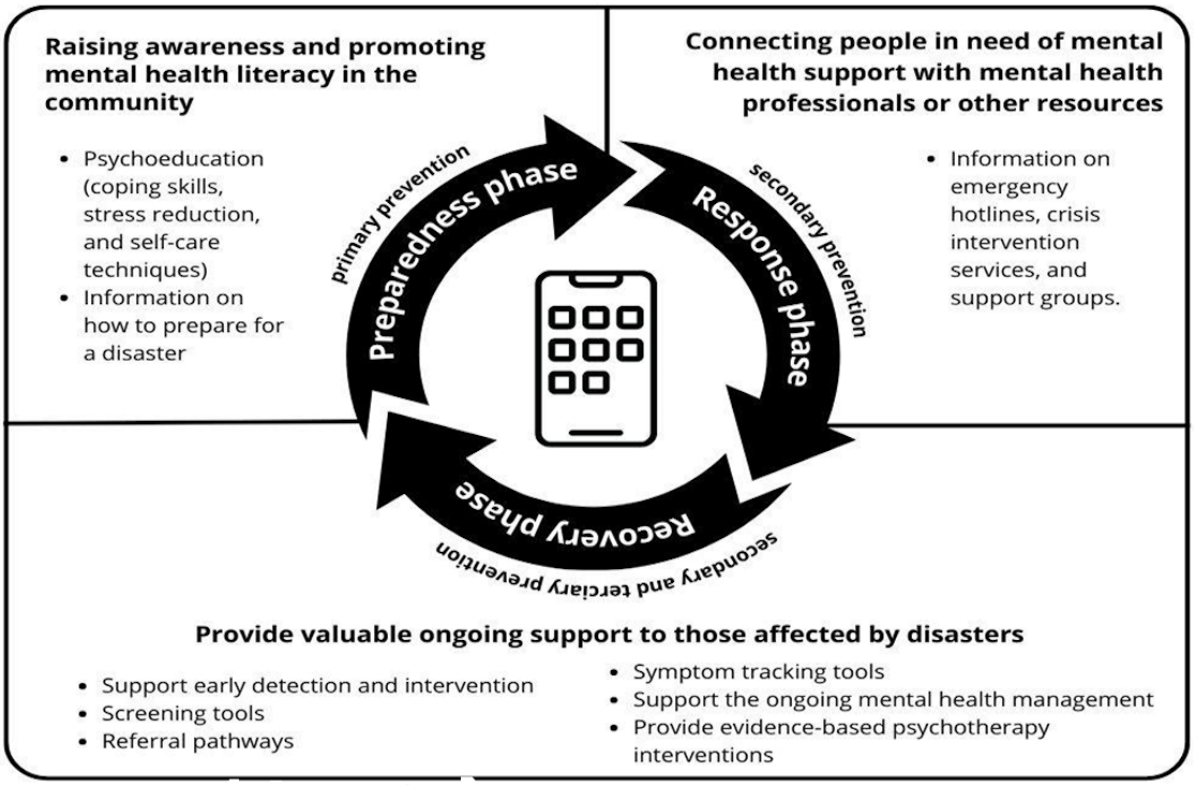

Preparedness Phase

Mobile apps can play a critical role as primary prevention interventions by raising awareness and promoting mental health literacy in the community in preparation for natural disasters. These apps can provide information on common mental health problems that may arise during and after disasters and offer tips on staying mentally healthy. For example, apps can include psychoeducation modules on coping skills, stress reduction, and self-care techniques, as well as information on how to prepare for a disaster and what steps to take to protect one's mental health during and after a disaster. The use and effectiveness of mobile apps in health literacy have been demonstrated in the literature [19], thus providing a foundation for adaptation in disaster management.

Response Phase

Mobile apps can be used to connect people in need of mental health support with mental health professionals or other resources. For example, apps can provide information on emergency hotlines, crisis intervention services, and support groups. This was demonstrated as effective during the COVID-19 pandemic [42]. Mobile apps can also provide coping strategies and techniques to manage stress and anxiety in response to other natural disasters [34]. In this scoping review, we found that 3 apps had positive benefits in addressing mental health issues among persons exposed to natural disasters.

Recovery Phase

As part of secondary and tertiary prevention strategies, mobile apps can provide valuable ongoing support to those affected by disasters. For secondary prevention, mobile apps can be designed to support early detection and intervention for mental health problems after a natural disaster. These apps can include screening tools to identify common mental health issues such as anxiety, depression, and PTSD and offer appropriate referral pathways [43]. Additionally, apps can provide symptom-tracking tools to help individuals monitor their mental health over time [43]. For tertiary prevention, mobile apps can support the ongoing management of established mental health problems after a natural disaster. For example, apps can provide evidence-based psychotherapy interventions, such as cognitive-behavioral therapy, to help individuals manage their

XSL•FO
RenderX

symptoms [44]. They can also connect individuals with support groups and peer-to-peer networks to provide additional emotional support and help individuals connect with others who have experienced similar challenges. Furthermore, mobile apps can offer self-help tools, such as meditation exercises and mood tracking, to help people cope with the ongoing mental health effects of the disaster. They can also provide information on local mental health services and support groups, helping individuals access the resources they need to manage their mental health.

General Mental Health Apps Show Promise for Disaster Response

While not specifically designed for disaster contexts, some mobile apps demonstrate strategies to support mental health that could aid disaster survivors. PTSD Coach delivers PTSD psychoeducation, symptom tracking tools, coping skills training, and crisis resource access—elements that could help survivors experiencing common postdisaster issues like trauma or loss [35]. Though it was tailored for veterans and civilians with PTSD, 1 study found it improved users' depression and functioning. Similarly, COVID Coach offered pandemic-related stress management through symptom tracking, healthy coping recommendations, and crisis line referrals [36]. By leveraging the scalability of mobile apps, COVID Coach reached many struggling during a global crisis. These examples illustrate that apps may provide accessible, far-reaching mediums for disseminating disaster mental health resources—even without disaster-specific tailoring. Research should further explore adapting evidence-based, general mental health apps for disaster contexts or incorporate elements of them into future disaster response tools. With mental health needs magnified during disasters, mobile apps with thoughtful design show promise in expanding access to psychosocial support.

There are several potential limitations when using mobile apps for mental health responses to disasters. One of the main concerns is the accessibility of these apps, as not all members of the affected communities may have access to smartphones or internet connectivity. Furthermore, language and cultural barriers may prevent effective use. Another potential limitation is the quality and accuracy of the information provided. Without proper oversight, some apps may provide misinformation or inaccurate advice, which could exacerbate mental health issues. In addition, privacy concerns around collecting and storing sensitive data must be addressed.

Barriers like lack of mobile devices and internet access can impede adoption, especially in marginalized areas. Apps not designed for low literacy users or that are only available in certain languages could also limit accessibility. Concerns around privacy and security may deter some individuals. However, smartphone ubiquity globally enables use by vulnerable groups. Government agencies and nongovernmental organizations (NGOs) can promote adoption by integrating vetted apps into disaster protocols and funding dissemination. Developing apps with stakeholders and prelaunch user testing also facilitate uptake. Monitoring user feedback allows for ongoing optimization and troubleshooting of barriers. Cultural tailoring to address stigma and use local beliefs further enables implementation success. Finally, limited evidence-based research into app effectiveness highlights the need for more rigorous evaluation and testing of mobile apps for disaster mental health response.

This scoping review has certain methodological limitations that should be considered while interpreting its results. First, the search was restricted to 6 electronic databases and only English-language papers were considered. We also searched MEDLINE and not PubMed, and these may have led to the omission of some relevant studies. Second, the study focused on mobile phone apps for mental health response to disasters, disregarding other types of technology that could also be used in disaster management such as telehealth, SMS text messaging, and emails. Moreover, since the study included only 5 papers, it may not offer a comprehensive overview of the use of mobile phone apps in disaster response strategies. There is the possibility of the existence of apps not yet published in academic literature. Fourth, the nonuse of a control group in the design of the studies makes it difficult to determine whether the observed effects were entirely due to the use of the apps or other characteristics of the participants that predisposed them to use the apps. Fifth, the small sample sizes for the studies mean they require caution with generalization. Despite these limitations, the review provides valuable insights into the use of mobile apps in disaster response and serves as a useful resource for developing contextually appropriate mobile apps for disaster management. Last, our study focused on natural disasters, further research should examine the role of apps in supporting mental health in conflict and complex emergencies such as wars, outbreaks of violence, and complex political conflict situations [45].

Conclusions

This scoping review found that mobile apps have not been extensively used in mental health responses to natural disasters, with only 5 studies meeting the inclusion criteria. However, the studies included in this review demonstrate that mobile apps can be useful in providing mental health support to disaster-affected individuals, as well as equip disaster responders. There is a critical gap identified in this study, as none of the studies investigated the use of mobile apps for potential victims in the preparedness or response phases of disaster management. We, therefore, recommend that mobile apps be integrated into the various phases of disaster management as part of mental health response. Additionally, it is important to ensure that these apps are accessible to all members of the community, taking into account cultural, linguistic, and other factors that may impact their effectiveness. Mobile apps have great potential to provide valuable ongoing support to those affected by disasters, and they can be a valuable resource in disaster management, helping people cope with the mental health effects of disasters and connecting with the necessary support services.

The findings from this scoping review have important implications for policy makers, disaster management professionals, and mental health practitioners. There is a clear need for policies and protocols that integrate evidence-based mobile apps into mental health disaster planning and response.

XSL•FO
RenderX

Disaster agencies should invest in developing, evaluating, and widely disseminating mobile apps specifically designed to mitigate psychological trauma before, during, and after catastrophic events. Mental health professionals can incorporate vetted mobile apps into their standard of care for at-risk disaster survivors. Going forward, a collaborative approach across these groups will be essential to leverage mobile technology in building community resilience and addressing the rising mental health burdens in an era defined by climate change–fueled natural disasters.

Acknowledgments

This work was funded by the Department of Psychiatry, Dalhousie University, Halifax, Canada. The funder was not involved in the conceptualization or implementation of the study, nor the decision to publish the findings.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The PRISMA-SCR checklist. PRISMA-SCR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.
[PDF File (Adobe PDF File), 497 KB - jmir_v26i1e49929_app1.pdf]

Multimedia Appendix 2
Detailed search strategy.
[DOCX File , 21 KB - jmir_v26i1e49929_app2.docx]

References

1. Birkmann J, Cardona OD, Carreño ML, Barbat AH, Pelling M, Schneiderbauer S, et al. Theoretical and conceptual framework for the assessment of vulnerability to natural hazards and climate change in Europe. In: Birkmann J, Kienberger S, Alexander DE, editors. Assessment of Vulnerability to Natural Hazards: A European Perspective. San Diego, California: Elsevier; 2014:1-19.
2. Bhola V, Hertelendy A, Hart A, Adnan SB, Ciottone G. Escalating costs of billion-dollar disasters in the US: climate change necessitates disaster risk reduction. J Clim Change Health 2023;10:100201. [doi: 10.1016/j.joclim.2022.100201]
3. Leppold C, Gibbs L, Block K, Reifels L, Quinn P. Public health implications of multiple disaster exposures. Lancet Public Health 2022;7(3):e274-e286 [FREE Full text] [doi: 10.1016/S2468-2667(21)00255-3] [Medline: 35065004]
4. Global SDG indicators data platform. United Nations. 2021. URL: https://unstats.un.org/sdgs/dataportal [accessed 2024-04-05]
5. Sloggy MR, Suter JF, Rad MR, Manning DT, Goemans C. Changing climate, changing minds? The effects of natural disasters on public perceptions of climate change. Clim Change 2021;168(3-4):25 [FREE Full text] [doi: 10.1007/s10584-021-03242-6] [Medline: 34720263]
6. Chen X, Bakal J, Whitten T, Waldie B, Ho C, Wright P, et al. Assessing the impact of COVID-19 pandemic on the health of residents and the healthcare system in Alberta, Canada: an observational study-the Alberta post-COVID follow-up study. BMJ Open 2023;13(2):e067449 [FREE Full text] [doi: 10.1136/bmjopen-2022-067449] [Medline: 36806075]
7. Pazderka H, Shalaby R, Eboreime E, Mao W, Obuobi-Donkor G, Agyapong B, et al. Isolation, economic precarity, and previous mental health issues as predictors of PTSD status in females living in Fort McMurray during COVID-19. Front Psychiatry 2022;13:837713 [FREE Full text] [doi: 10.3389/fpsyt.2022.837713] [Medline: 35370820]
8. To P, Eboreime E, Agyapong VIO. The impact of wildfires on mental health: a scoping review. Behav Sci (Basel) 2021;11(9):126 [FREE Full text] [doi: 10.3390/bs11090126] [Medline: 34562964]
9. Agyapong B, Shalaby R, Eboreime E, Obuobi-Donkor G, Owusu E, Adu MK, et al. Cumulative trauma from multiple natural disasters increases mental health burden on residents of Fort McMurray. Eur J Psychotraumatol 2022;13(1):2059999 [FREE Full text] [doi: 10.1080/20008198.2022.2059999] [Medline: 35599978]
10. Guerra O, Eboreime E. The impact of economic recessions on depression, anxiety, and trauma-related disorders and illness outcomes-a scoping review. Behav Sci (Basel) 2021;11(9):119 [FREE Full text] [doi: 10.3390/bs11090119] [Medline: 34562956]
11. Norris FH, Friedman MJ, Watson PJ, Byrne CM, Diaz E, Kaniasty K. 60,000 disaster victims speak: part I. An empirical review of the empirical literature, 1981-2001. Psychiatry 2002;65(3):207-239. [doi: 10.1521/psyc.65.3.207.20173] [Medline: 12405079]
12. North CS, Pfefferbaum B. Mental health response to community disasters: a systematic review. JAMA 2013;310(5):507-518. [doi: 10.1001/jama.2013.107799] [Medline: 23925621]

XSL•FO
RenderX

13. Pfefferbaum B, Flynn BW, Schonfeld D, Brown LM, Jacobs GA, Dodgen D, et al. The integration of mental and behavioral health into disaster preparedness, response, and recovery. Disaster Med Public Health Prep 2012;6(1):60-66. [doi: 10.1001/dmp.2012.1] [Medline: 22490938]
14. Math SB, Nirmala MC, Moirangthem S, Kumar NC. Disaster management: mental health perspective. Indian J Psychol Med 2015;37(3):261-271 [FREE Full text] [doi: 10.4103/0253-7176.162915] [Medline: 26664073]
15. Ali A, Katz DL. Disease prevention and health promotion: how integrative medicine fits. Am J Prev Med 2015;49(5 Suppl 3):S230-s240 [FREE Full text] [doi: 10.1016/j.amepre.2015.07.019] [Medline: 26477898]
16. Reddy MC, Paul SA, Abraham J, McNeese M, DeFlitch C, Yen J. Challenges to effective crisis management: using information and communication technologies to coordinate emergency medical services and emergency department teams. Int J Med Inform 2009;78(4):259-269. [doi: 10.1016/j.ijmedinf.2008.08.003] [Medline: 18835211]
17. Cheng DR, Coote A, South M. A digital approach in the rapid response to COVID-19—experience of a paediatric institution. Int J Med Inform 2021;149:104407 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104407] [Medline: 33588302]
18. Souza F, Kushchu I. Mobile disaster management system applications—current overview and future potential. In: IOP Conference Series: Materials Science and Engineering. 2005 Presented at: Proceedings EURO mGOV; 2005; NA p. 455-466 URL: https://iopscience.iop.org/article/10.1088/1757-899X/1009/1/012049
19. Chandrashekar P. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth 2018;4:6 [FREE Full text] [doi: 10.21037/mhealth.2018.03.02] [Medline: 29682510]
20. Hyde AM, Watt M, Carbonneau M, Eboreime EA, Abraldes JG, Tandon P. Understanding preferences toward virtual care: a pre-COVID mixed methods study exploring the perspectives of patients with chronic liver disease. Telemed J E Health 2022;28(3):407-414. [doi: 10.1089/tmj.2021.0099] [Medline: 34085869]
21. Eboreime E, Ohinmaa A, Rusak B, Cassidy KL, Morrison J, McGrath P, et al. The Text4HealthyAging program: an evidence-based text messaging innovation to support healthy urban aging in Canada and Australia. Gerontol Geriatr Med 2022;8:23337214221081378 [FREE Full text] [doi: 10.1177/23337214221081378] [Medline: 35252477]
22. Shalaby R, Adu MK, Andreychuk T, Eboreime E, Gusnowski A, Vuong W, et al. Prevalence, demographic, and clinical correlates of likely PTSD in subscribers of Text4Hope during the COVID-19 pandemic. Int J Environ Res Public Health 2021;18(12):6227 [FREE Full text] [doi: 10.3390/ijerph18126227] [Medline: 34207537]
23. Vilendrer S, Amano A, Johnson CGB, Favet M, Safaeinili N, Villasenor J, et al. An app-based intervention to support first responders and essential workers during the COVID-19 pandemic: needs assessment and mixed methods implementation study. J Med Internet Res 2021;23(5):e26573 [FREE Full text] [doi: 10.2196/26573] [Medline: 33878023]
24. Obuobi-Donkor G, Eboreime E, Bond J, Phung N, Eyben S, Hayward J, et al. An E-mental health solution to prevent and manage posttraumatic stress injuries among first responders in Alberta: protocol for the implementation and evaluation of text messaging services (Text4PTSI and Text4Wellbeing). JMIR Res Protoc 2022;11(4):e30680 [FREE Full text] [doi: 10.2196/30680] [Medline: 35468094]
25. Eboreime E, Shalaby R, Mao W, Owusu E, Vuong W, Surood S, et al. Reducing readmission rates for individuals discharged from acute psychiatric care in Alberta using peer and text message support: protocol for an innovative supportive program. BMC Health Serv Res 2022;22(1):332 [FREE Full text] [doi: 10.1186/s12913-022-07510-8] [Medline: 35279142]
26. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205. [doi: 10.1080/16506070903318960] [Medline: 20183695]
27. Nicholas J, Ringland KE, Graham AK, Knapp AA, Lattie EG, Kwasny MJ, et al. Stepping up: predictors of 'stepping' within an iCBT stepped-care intervention for depression. Int J Environ Res Public Health 2019;16(23):4689 [FREE Full text] [doi: 10.3390/ijerph16234689] [Medline: 31775297]
28. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
29. Covidence systematic review software. Veritas Health Innovation. Melbourne, Australia; 2022. URL: https://www.covidence.org/ [accessed 2024-04-05]
30. Heinz AJ, Wiltsey-Stirman S, Jaworski BK, Sharin T, Rhodes L, Steinmetz S, et al. Feasibility and preliminary efficacy of a public mobile app to reduce symptoms of postdisaster distress in adolescent wildfire survivors: Sonoma rises. Psychol Serv 2021;19(2):67-79. [doi: 10.1037/ser0000576] [Medline: 34780209]
31. Rung AL, Oral E, Berghammer L, Peters ES. Feasibility and acceptability of a mobile mindfulness meditation intervention among women: intervention study. JMIR Mhealth Uhealth 2020;8(6):e15943 [FREE Full text] [doi: 10.2196/15943] [Medline: 32442147]
32. Choi YJ, Ko EJ, Choi EJ, Um YJ. Managing traumatic stress using a mental health care mobile app: a pilot study. Int J Ment Health Promot 2021;23(3):385-393 [FREE Full text] [doi: 10.32604/IJMHP.2021.015018]
33. Choi YJ, Jung HS, Choi EJ, Ko E. Disaster healthcare workers' experience of using the psychological first aid mobile app during disaster simulation training. Disaster Med Public Health Prep 2021;17:e55. [doi: 10.1017/dmp.2021.308] [Medline: 34674792]

XSL•FO
RenderX

34. Seligman J, Felder SS, Robinson ME. Substance Abuse and Mental Health Services Administration (SAMHSA) behavioral health disaster response app. Disaster Med Public Health Prep 2015;9(5):516-518. [doi: 10.1017/dmp.2015.84] [Medline: 26165522]
35. Kuhn E, van der Meer C, Owen JE, Hoffman JE, Cash R, Carrese P, et al. PTSD coach around the world. Mhealth 2018;4:15 [FREE Full text] [doi: 10.21037/mhealth.2018.05.01] [Medline: 29963560]
36. Jaworski BK, Taylor K, Ramsey KM, Heinz A, Steinmetz S, Pagano I, et al. Exploring usage of COVID coach, a public mental health app designed for the COVID-19 pandemic: evaluation of analytics data. J Med Internet Res 2021;23(3):e26559 [FREE Full text] [doi: 10.2196/26559] [Medline: 33606656]
37. Tan ML, Prasanna R, Stock K, Hudson-Doyle E, Leonard G, Johnston D. Mobile applications in crisis informatics literature: a systematic review. Int J Disaster Risk Reduct 2017;24:297-311. [doi: 10.1016/j.ijdrr.2017.06.009]
38. Romano M, Onorati T, Aedo I, Diaz P. Designing mobile applications for emergency response: citizens acting as human sensors. Sensors (Basel) 2016;16(3):406 [FREE Full text] [doi: 10.3390/s16030406] [Medline: 27007375]
39. Schoning J, Rohs M, Krüger A, Stasch C. Improving the communication of spatial information in crisis response by combining paper maps and mobile devices. In: Mobile Response 2008. Lecture Notes in Computer Science, vol 5424. Berlin, Heidelberg: Springer; 2009:57-65.
40. Roudini J, Khankeh HR, Witruk E. Disaster mental health preparedness in the community: a systematic review study. Health Psychol Open 2017;4(1):2055102917711307 [FREE Full text] [doi: 10.1177/2055102917711307] [Medline: 28680695]
41. Colajanni G, Daniele P, Nagurney A, Nagurney LS, Sciacca D. A three-stage stochastic optimization model integrating 5G technology and UAVs for disaster management. J Glob Optim 2023;86:1-40 [FREE Full text] [doi: 10.1007/s10898-023-01274-z] [Medline: 36855677]
42. Singh HJL, Couch D, Yap K. Mobile health apps that help with COVID-19 management: scoping review. JMIR Nurs 2020;3(1):e20596 [FREE Full text] [doi: 10.2196/20596] [Medline: 32897271]
43. Funnell EL, Spadaro B, Martin-Key N, Metcalfe T, Bahn S. mHealth solutions for mental health screening and diagnosis: a review of app user perspectives using sentiment and thematic analysis. Front Psychiatry 2022;13:857304 [FREE Full text] [doi: 10.3389/fpsyt.2022.857304] [Medline: 35573342]
44. Murphy K, Coombes I, McMillan S, Wheeler AJ. Clozapine and shared care: the consumer experience. Aust J Prim Health 2018;24(6):455-462. [doi: 10.1071/PY18055] [Medline: 30452883]
45. Goniewicz K, Burkle FM, Hertelendy AJ, Khorram-Manesh A. The impact of war on emergency departments visits by Ukrainian refugees in Poland. Am J Emerg Med 2023;67:189-190. [doi: 10.1016/j.ajem.2023.02.040] [Medline: 36914486]

Abbreviations

BFS: Behaviour and Feelings Survey
CESD-10: Center for Epidemiologic Studies Depression Scale-10
CPSS-5: Child PTSD Symptom Scale for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
EEG: electroencephalogram
EP: explanatory power
GAD-7: Generalized Anxiety Disorder-7
HIPAA: Health Insurance Portability and Accountability Act
ISI: Insomnia—Severity Index
MAAS: Mindful Attention Awareness Scale
mHealth: mobile health
NGO: Non-Governmental Organization
OR: odds ratio
OSY: Ohio Scale for Youth—Functioning subscale
PFA: Psychological First Aid
PHQ-9: Patient Health Questionnaire-9
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PSQI: Pittsburgh Sleep Quality Index
PSS-4: Perceived Stress Scale, 4-item version
PTSD: posttraumatic stress disorder
SAMHSA: Substance Abuse and Mental Health Services Administration
SEI: Student Engagement Instrument
TLS: Training for Life Skills
WEMWBS: Warwick-Edinburgh Mental Well-being Scale
WSSS: Wills' Social Support Scale

XSL•FO
RenderX

Edited by G Eysenbach; submitted 13.06.23; peer-reviewed by T Benham, K Goniewicz, R Konu, J Ranse, P Moreno-Peral; comments to author 10.01.24; revised version received 25.02.24; accepted 23.03.24; published 17.04.24.

Please cite as:
Ezeonu NA, Hertelendy AJ, Adu MK, Kung JY, Itanyi IU, Dias RDL, Agyapong B, Hertelendy P, Ohanyido F, Agyapong VIO, Eboreime E
Mobile Apps to Support Mental Health Response in Natural Disasters: Scoping Review
J Med Internet Res 2024;26:e49929
URL: https://www.jmir.org/2024/1/e49929
doi:10.2196/49929
PMID:38520699

©Nwamaka Alexandra Ezeonu, Attila J Hertelendy, Medard Kofi Adu, Janice Y Kung, Ijeoma Uchenna Itanyi, Raquel da Luz Dias, Belinda Agyapong, Petra Hertelendy, Francis Ohanyido, Vincent Israel Opoku Agyapong, Ejemai Eboreime. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Investigating the Cost-Effectiveness of Telemonitoring Patients With Cardiac Implantable Electronic Devices: Systematic Review

Sarah Raes[1], MSc; Andrea Prezzi[1], MSc; Rik Willems[2], PhD; Hein Heidbuchel[3], PhD; Lieven Annemans[1], PhD

[1]Department of Public Health and Primary Care, Ghent University, Gent, Belgium

[2]Department of Cardiovascular Sciences, Universiteit Leuven, Leuven, Belgium

[3]Department of Genetics, Pharmacology and Physiopathology of Heart, Blood Vessels and Skeleton (GENCOR), Antwerp University, Antwerp, Belgium

Corresponding Author:
Sarah Raes, MSc
Department of Public Health and Primary Care
Ghent University
Corneel Heymanslaan 10
Gent, 9000
Belgium
Phone: 32 9 332 83 59
Email: Sarah.Raes@UGent.be

Abstract

Background: Telemonitoring patients with cardiac implantable electronic devices (CIEDs) can improve their care management. However, the results of cost-effectiveness studies are heterogeneous. Therefore, it is still a matter of debate whether telemonitoring is worth the investment.

Objective: This systematic review aims to investigate the cost-effectiveness of telemonitoring patients with CIEDs, focusing on its key drivers, and the impact of the varying perspectives.

Methods: A systematic review was performed in PubMed, Web of Science, Embase, and EconLit. The search was completed on July 7, 2022. Studies were included if they fulfilled the following criteria: patients had a CIED, comparison with standard care, and inclusion of health economic evaluations (eg, cost-effectiveness analyses and cost-utility analyses). Only complete and peer-reviewed studies were included, and no year limits were applied. The exclusion criteria included studies with partial economic evaluations, systematic reviews or reports, and studies without standard care as a control group. Besides general study characteristics, the following outcome measures were extracted: impact on total cost or income, cost or income drivers, cost or income drivers per patient, cost or income drivers as a percentage of the total cost impact, incremental cost-effectiveness ratios, or cost-utility ratios. Quality was assessed using the Consensus Health Economic Criteria checklist.

Results: Overall, 15 cost-effectiveness analyses were included. All studies were performed in Western countries, mainly Europe, and had primarily a male participant population. Of the 15 studies, 3 (20%) calculated the incremental cost-effectiveness ratio, 1 (7%) the cost-utility ratio, and 11 (73%) the health and cost impact of telemonitoring. In total, 73% (11/15) of the studies indicated that telemonitoring of patients with implantable cardioverter-defibrillators (ICDs) and cardiac resynchronization therapy ICDs was cost-effective and cost-saving, both from a health care and patient perspective. Cost-effectiveness results for telemonitoring of patients with pacemakers were inconclusive. The key drivers for cost reduction from a health care perspective were hospitalizations and scheduled in-office visits. Hospitalization costs were reduced by up to US $912 per patient per year. Scheduled in-office visits included up to 61% of the total cost reduction. Key drivers for cost reduction from a patient perspective were loss of income, cost for scheduled in-office visits and transport. Finally, of the 15 studies, 8 (52%) reported improved quality of life, with statistically significance in only 1 (13%) study (P=.03).

Conclusions: From a health care and patient perspective, telemonitoring of patients with an ICD or a cardiac resynchronization therapy ICD is a cost-effective and cost-saving alternative to standard care. Inconclusive results were found for patients with pacemakers. However, telemonitoring can lead to a decrease in providers' income, mainly due to a lack of reimbursement. Introducing appropriate reimbursement could make telemonitoring sustainable for providers while still being cost-effective from a health care payer perspective.

Trial Registration: PROSPERO CRD42022322334; https://tinyurl.com/puunapdr

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47616) doi:10.2196/47616

KEYWORDS

systematic review; cost-effectiveness; telemonitoring; cardiac device; implantable cardioverter-defibrillator; ICD; pacemaker; monitoring; patient management; effectiveness; cost; quality of life; cardiac implantable electronic device; cardiac

Introduction

Background

The implantation rates of cardiac implantable electronic devices (CIEDs), including pacemakers and implantable cardioverter-defibrillators (ICDs), have increased over the last decades due to expanded indications and a progressively aging population [1]. To evaluate the clinical status of the patient and device functioning, current guidelines recommend that older patients with pacemakers should be evaluated every 3 to 12 months and patients with ICDs should be evaluated every 3 to 6 months [2]. This regimen imposes a considerable burden on patients and physicians if the patient is required to be seen in person.

Telemonitoring, referring to the process of using telecommunication and information technology to monitor the health status of a patient and device function from a distance, can reduce this burden by replacing some in-office visits with transmissions from the patients' home [3]. Existing research indicated that telemonitoring is safe (eg, experiencing equal major adverse events to standard care) [4,5]. The advantages of telemonitoring include fewer inappropriate shocks for patients with ICDs [4,6] and fewer hospitalizations for patients with atrial arrhythmias and strokes [4,6,7]. Moreover, there is a rapid detection of cardiovascular events and device malfunction [5,7], leading to a time reduction between clinical decision and intervention [8].

Besides the effectiveness of telemonitoring, patient experience is essential in high-quality health care services. Overall, patients with pacemakers on telemonitoring reported positive experiences comparable to the experience of patients with in-hospital monitoring [9]. Telemonitored patients with pacemakers tended to receive less information about their diagnosis but no significant differences were found in other items, such as confidence in clinicians, treatment decision involvement, treatment satisfaction, and waiting time before admission [9]. Another study indicated that telemonitoring of patients with a cardiac resynchronization therapy defibrillator (CRT-D) was time-saving for both patients and physicians [10].

Cost-effectiveness analyses are important to quantify the value of new interventions, informing both medical decision-making and public policy [11]. However, cost-effectiveness analyses depend on the perspective considered. The different perspectives are the health care payer perspective (eg, Medicare or Medicaid and British National Health Service), the patient perspective, the provider perspective (eg, physician), and the society perspective. The health care payer and societal perspectives differ from each other as the societal perspective includes indirect nonmedical costs (eg, transport) [12].

Objectives

As cost-effectiveness analyses have shown heterogeneous results, it is still debatable whether telemonitoring is worth the investment relative to standard care. However, data on cost-effectiveness are important for health care payers to make decisions on the reimbursement of telemonitoring. Lack of reimbursement can be an important adoption barrier for new technology [13,14]. For these 2 reasons, this paper reviews the cost-effectiveness of telemonitoring, reviews how the results differ from different perspectives, and describes the key drivers of the cost-effectiveness of telemonitoring.

Methods

Overview

The review protocol was published by PROSPERO (International Prospective Register of Systematic Reviews; CRD42022322334). This systematic review was carried out in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guideline of 2020 [15], and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [16], which can be found in the Multimedia Appendix 1. Guidelines for preparing a systematic review of health economic evaluations were followed [17].

Literature Search

For this review, PubMed, Embase, EconLit, and Web of Science Core Collection were systematically searched. The last search was performed on July 7, 2022. No filters (eg, publication date or type of study) were applied. Search strategies for all electronic databases can be found in Multimedia Appendix 2.

Search strings were developed based on explorations of databases and previous reviews. The following key concepts were translated into strings: (1) CIEDs, (2) telemonitoring, and (3) economic evaluations (eg, cost-effectiveness analyses and cost-utility analyses). The latter was based on a validated search filter, designed to identify economic evaluations, and was broadened for this study to maximize sensitivity [18]. The search terms for CIEDs and telemonitoring were based on existing reviews [19-21].

Study Selection

Studies were included if their primary focus was on the cost-effectiveness of telemonitoring patients with a CIED. The eligibility criteria were defined a priori for study selection (Textbox 1). The population, intervention, comparator, and outcome strategy was applied to describe the criteria. Only complete and peer-reviewed studies were included. Specific exclusion criteria included partial economic evaluations, systematic reviews or reports, and studies without standard care as a control group. Only studies published in English, Dutch,

XSL•FO
RenderX

French, or German were eligible for inclusion. The reference lists of the included studies were searched manually to identify relevant studies. Two reviewers (SR and AP) independently screened the titles and abstracts of all records using Rayyan (Rayyan Systems Inc) [22]. After the initial screening, full texts were retrieved and screened for a second time. The second screening round was independently performed by 2 reviewers (SR and AP). Reasons for exclusion were documented (Figure 1). For both screening rounds, reviewers were blinded from each other's decision, and disagreements were resolved through discussion.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Intervention
  - Cardiac implantable electronic devices: pacemaker, implantable cardioverter-defibrillator, cardiac resynchronization therapy defibrillator, cardiac resynchronization therapy pacemaker, and loop recorder
- Comparator
  - Standard care
- Study design
  - Complete health economic evaluations (within-trial and model-based)
- Context
  - All settings
- Language
  - English, French, German, or Dutch

Exclusion criteria

- Intervention
  - Implantable pulmonary artery pressure monitor
- Study design
  - Partial health economic evaluations (outcomes related to costs or effectiveness only)
- Specific criteria
  - Systematic reviews, reports, commentaries, congress abstracts, protocols, and animal studies

XSL•FO
RenderX

Figure 1. Flowchart of the study selection.

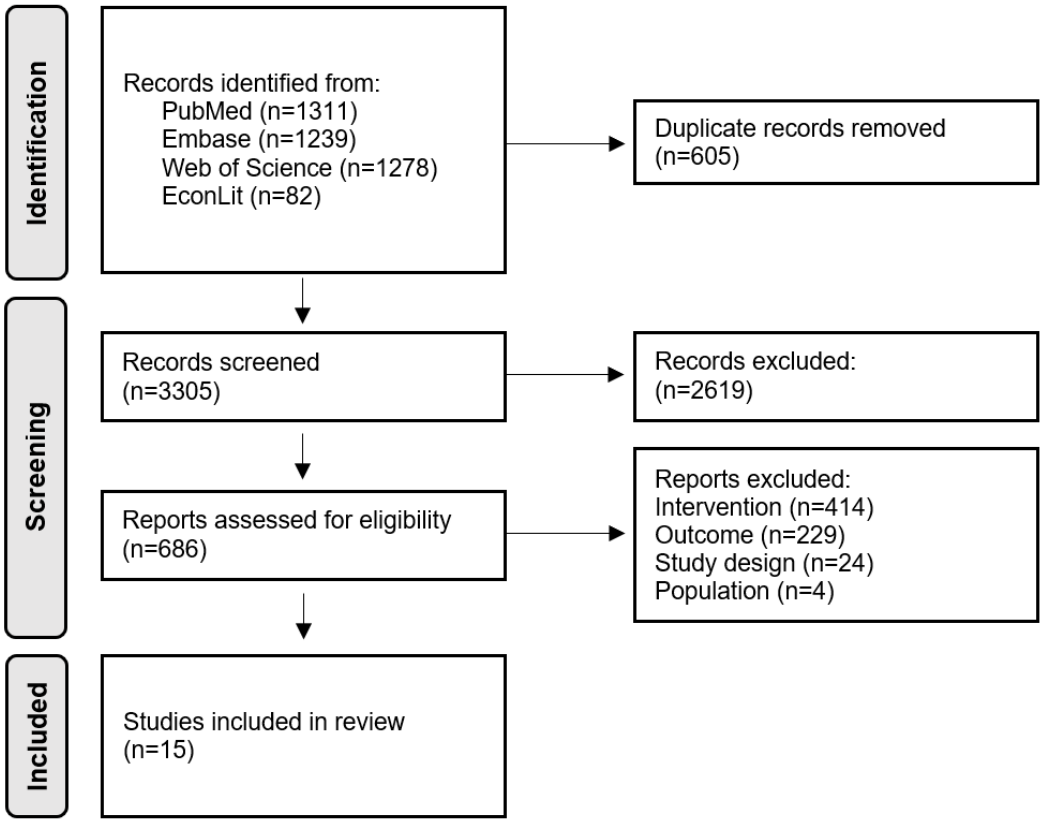

Quality Assessment

Two researchers (SR and AP) independently evaluated the original papers using the Consensus Health Economic Criteria (CHEC) checklist to assess the risk of bias [23]. The CHEC checklist included 19 items. Any disagreement was resolved by discussion and consensus. Interpretation of the CHEC list can be found in Multimedia Appendix 3. The included studies were classified into 4 quality categories: excellent (score of 100%), good quality (score between 75% and 100%), moderate quality (score between 50% and 75%), and low quality (score <50%) [24].

Synthesis of Results

The study characteristics and main outcomes of the original papers are presented in the *Results* section. SR extracted all data. A data extraction sheet was developed using an existing template [17]. The following information was extracted from the included studies: study identification, general study characteristics, results, and authors' conclusion. The principal outcome measures were health outcomes, cost or income outcomes (eg, the impact on total cost or income, cost or income drivers, cost or income drivers per patient, and cost or income drivers as a percentage of the total cost impact), and incremental cost-effectiveness ratios (ICERs) or cost-utility ratios.

To facilitate comparison across studies, the following adjustments and interpretations were made. First, the cost or income outcomes were presented per patient per year, and different currencies were converted to US Dollar (reference year: 2019 and reference country: United States) [25]. Second, perspectives were categorized into the health care payer perspective, patient perspective, provider perspective, and societal perspective. For the purpose of our study, the provider includes physicians who are directly involved in the care of patients with CIED.

Results

Overview

The selection process is shown in Figure 1. From a total of 3305 publications, 15 (0.45%) unique publications were reviewed. Studies were excluded because one of the following reasons: (1) intervention: the paper did not describe telemonitoring patients with a CIED; (2) outcome: the paper contained only a cost analysis and not a cost-effectiveness analysis; and (3) study design or publication: the paper was a partial health economic evaluation, congress abstract, protocol, systematic review, animal study, or with no peer review.

Population

Characteristics of the included studies can be found in Table 1. All 15 (100%) studies had a primarily male population, except for the Nordland study, which had an almost equal sex distribution (Table 1) [26]. The mean age of the population with pacemakers was between 75 (SD 24.64) and 81 (SD 6.47) years. The mean age of patients with an ICD or CRT-D was between 61 (SD 12.6) and 69 (SD not calculated) years, except for the PREDICT RM study, where >50% of the population was aged >75 years [27]. Furthermore, of the 15 studies, 1 (7%) included only older patients (with a mean age of 81 years) with pacemakers [28], and 2 (13%) ICD or CRT-D studies only included patients with heart failure [11,29].

XSL•FO
RenderX

Table 1. Main characteristics of the included studies.

Study	Author and year	Patients, n	Population characteristics	Age (years), mean (SD)	Male participant (%)	CIED[a] type
Poniente [28]	Bautista-Mesa et al [28], 2022	55	Mean age of 81 years	81 (6.47)	69	Pacemaker
PREDICT RM [27]	Hummel et al [27], 2019	15,254	N/A[b]	53% of participants aged ≥75 years[c]	72	ICD[d]
TARIFF[e] [30]	Ricci et al [30], 2016	209	N/A	69 (10.17)	85	ICD or CRT-D[f]
Nordland [26]	Lopez-Villegas et al [26], 2020	50	N/A	74.8 (24.64)	52	Pacemaker
EVOLVO[g] [11]	Zanaboni et al [11], 2013	200	Patients with heart failure	66-69 (SD not reported)	79	ICD or CRT-D
MORE-CARE[h] [29]	Boriani et al [29], 2016	865	Patients with heart failure	66 (10)	76	CRT-D
Burri et al [31]	Burri et al [31], 2013	N/A	Patients with biventricular CRT-D	65 (SD not reported)	N/A	ICD or CRT-D
Raatikainen et al [32]	Raatikainen et al [32], 2008	41	N/A	62 (10)	83	ICD
Al-Khatib et al [33]	Al-Khatib et al [33], 2009	151	N/A	63 (SD not reported)	72	ICD or CRT-D
CONNECT[i] [8]	Crossley et al [8], 2011	1997	N/A	65 (12.1)	71	ICD or CRT-D
ECOST[j] [34]	Guédon-Moreau et al [34], 2014	310	N/A	60.7 (12.6)	90	ICD
EuroEco[k] [13]	Heidbuchel et al [13], 2015	303	Patients with new or replacement VVI-ICD[l] or DDD-ICD[m]	62.4 (13.1)	81	ICD or CRT-D
SAVE-HM[n] trial [35]	Perl et al [35], 2013	115	Patients with dual chamber pacemaker	74 (9)	57	Pacemaker
SAVE-HM[n] trial [35]	Perl et al [35], 2013	36	Patients with ICD-implant due to primary prevention of sudden cardiac death	62.5 (10)	86	ICD
Chew et al [36]	Chew et al [36], 2020	1830	N/A	66 (SD not reported)	88	ICD or CRT-D
Dario et al [37]	Dario et al [37], 2016	1171	N/A	77.5 (9)	58	Pacemaker
Dario et al [37]	Dario et al [37], 2016	930	N/A	67.5 (12)	79	ICD

[a]CIED: cardiac implantable electronic device.

[b]N/A: not applicable.

[c]Age was a discrete variable in this study (higher of lower than 75 years old).

[d]ICD: implantable cardioverter-defibrillator.

[e]TARIFF: Health Economics Evaluation Registry for Remote Follow-Up.

[f]CRT-D: cardiac resynchronization therapy defibrillator.

[g]EVOLVO: Evolution of Management Strategies of Heart Failure Patients With Implantable Defibrillators.

[h]MORE-CARE: Monitoring Resynchronization Devices and Cardiac Patients.

[i]CONNECT: Clinical Evaluation of Remote Notification to Reduce Time to Clinical Decision.

[j]ECOST: Effectiveness and Cost of ICD Follow-Up Schedule With Telecardiology.

[k]EuroEco: European Health Economic Trial on Home Monitoring in ICD Patients.

[l]VVI-ICD: single-chamber ICD.

[m]DDD-ICD: dual-chamber ICD.

[n]SAVE-HM: Socio-Economic Effects and Cost Saving Potential of Remote Patient Monitoring.

XSL•FO
RenderX

Study Designs

Tables 2 and 3 show the summary table of results. Of 15 studies, 11 (73%) were conducted in Europe [11,26,28-32,34,35,37], 3 (20%) in the United States [27,33,38], and 1 (7%) in Canada [36]. Of the 15 studies, 3 (20%) calculated the ICER [26-28], 1 (7%) calculated the cost-utility ratio [11], and 11 (73%) calculated the cost impact of telemonitoring. All studies analyzed the health care payer perspective, with 33% (5/15) analyzing the patient perspective [11,28,30,32,34], 13% (2/15) analyzing the societal perspective [33,35], and 13% (2/15) analyzing the provider perspective [13,30].

XSL•FO
RenderX

Table 2. Summary of the main results.

Study	Country	Design	Time horizon	CIED[a] type	Effect	Cost-effectiveness in original currency and in reference year (in US $, 2019)	Conclusion
Pacemaker studies							
Poniente [28]	Spain	Non-RCT[b]	5 years	Pacemaker	QALY[c] difference: 0.27	ICER[d]: €301.16 per QALY (US $270.09 per QALY)	Cost-effective
Nordland [26]	Norway	RCT	1 year	Pacemaker	QALY difference: 0.03	ICER: €53,345 per QALY (US $59.746 per QALY)	Not cost-effective
SAVE-HM[e] trial [35]	Austria	RCT	17 months	Pacemaker	No adverse effects difference	N/A[f]	Cost-saving
Dario et al [37], 2016	Italy	Non-RCT	1 year	Pacemaker	Average time reduction to treat patients (−4.1 minutes/follow-up)	N/A	Cost-saving
ICD[g] or CRT-D[h] studies							
PREDICT RM [27]	United States	Real-world	Lifelong	ICD	QALY difference: 0.64	ICER: US $10,752 per QALY (US $12,069 per QALY)	Cost-effective
TARIFF[i] [30]	Italy	Non-RCT	12 months	ICD or CRT-D	QALY difference: 0.02	Not calculated because QALY difference was not significant (*P*=.53)	Cost-saving
EVOLVO[j] [11]	Italy	RCT	16 months	ICD or CRT-D	QALY difference: 0.066[k] (*P*=.03)	Cost-utility ratio <0	Dominant
MORE-CARE[l] [29]	Europe and Israel	RCT	2 years	CRT-D	QOL[m] difference:−1	N/A	Cost-saving
Burri et al [31], 2013	United Kingdom	Systematic review data	10 years	ICD or CRT-D	Inappropriate shocks: −51%; battery exhaustion: −7%	N/A	Cost-saving
Raatikainen et al [32], 2008	Finland	Non-RCT	18 months	ICD	Time burden for patients of −175 minutes and physician of −17 minutes/patient/follow-up	N/A	Cost-effective
Al-Khatib et al [33], 2009	United States	RCT	1 year	ICD or CRT-D	EuroQoL difference: 25%; no difference in satisfaction and mortality	N/A	Cost-saving
CONNECT[n] [8]	United States	RCT	15 months	ICD or CRT-D	Time from clinical event to clinical decision: 17.4 days[k] (*P*<.001)	N/A	Cost-saving
ECOST[o] [34]	France	RCT	27 months	ICD	Physical, psychological, and SF-36[p] QOL scores: not significant	N/A	Cost-saving

XSL•FO
RenderX

Study	Country	Design	Time horizon	CIED[a] type	Effect	Cost-effectiveness in original currency and in reference year (in US $, 2019)	Conclusion
EuroEco[q] [13]	Belgium, Finland, Germany, United Kingdom, Spain, and the Netherlands	RCT	2 years	ICD or CRT-D	SF-36 QOL score: not significant	N/A	Cost-saving
SAVE-HM trial [35]	Austria	RCT	26 months	ICD	No adverse effects difference	N/A	Cost-saving
Chew et al [36], 2022	Canada	Non-RCT	5 years	ICD or CRT-D	Risk of death (hazard ratio): 0.43[k] (P<.001)	N/A	Cost-saving
Dario et al [37], 2016	Italy	Non-RCT	1 year	ICD	Average time reduction to treat patient (−13.7 minute/follow-up)	N/A	Cost-saving

[a]CIED: cardiac implantable electronic device.
[b]RCT: randomized controlled trial.
[c]QALY: quality-adjusted life year.
[d]ICER: incremental cost-effectiveness ratio.
[e]SAVE-HM: Socio-Economic Effects and Cost Saving Potential of Remote Patient Monitoring.
[f]N/A: not applicable.
[g]ICD: implantable cardioverter-defibrillator.
[h]CRT-D: cardiac resynchronization therapy defibrillator.
[i]TARIFF: Health Economics Evaluation Registry for Remote Follow-Up.
[j]EVOLVO: Evolution of Management Strategies of Heart Failure Patients With Implantable Defibrillators.
[k]The values are statistically significant.
[l]MORE-CARE: Monitoring Resynchronization Devices and Cardiac Patients.
[m]QOL: quality of life.
[n]CONNECT: Clinical Evaluation of Remote Notification to Reduce Time to Clinical Decision.
[o]ECOST: Effectiveness and Cost of ICD Follow-Up Schedule With Telecardiology.
[p]SF-36: The 36-Item Short Form Survey.
[q]EuroEco: European Health Economic Trial on Home Monitoring in ICD Patients.

XSL•FO
RenderX

Table 3. Summary table of results related to perspectives and key cost or income drivers.

Study and perspective	Total cost or income[a] impact compared to standard care in original currency in reference year (US $ pp[b] per year, 2019) and cost or income[a] impact drivers	Cost or income[a] impact drivers pp (US $ pp per year, 2019)	Cost or income[a] impact drivers as a percentage of total cost or income impact (%)
Pacemaker studies			
Poniente [28]			
Health care payer perspective			
€, 2012: −€8 (−US $8.96)			
	Staff costs	−€3.7 (−US $ 4.56)	49
	Ambulance transport	−€3.2 (−US $3.9)	42
	Consultation room	−€0.8[c] (−US $0.9)	10
Patient perspective			
€, 2012: −€9 (−US $11.2)			
	Informal transport	−€5.1[c] (−US $ 6.2)	58
	Lost income	−€3.7 (−US $4.58)	42
Nordland [26]			
Health care payer perspective			
€, 2015: €1,808 (US $2,183)			
	Hospitalization	€1,808.31 (US $2,183)	100
	Ambulance transport	−€60 (−US $72.5)	−3
	Physician cost	€39.39[c](US $47.6)	2
	Consultation room	€20.17[c] (US $24.3)	1
SAVE-HM[d] trial [35]			
Societal perspective			
€, 2013: −€914 (−US $1,113)			
	Transport	−€911.3 (−US $1,020)	99.7
	Follow-up personnel cost	−€26.7 (−US $32.93)	2
Dario et al [37], 2016			
Health care payer perspective			
€, 2011: −€832[c] (−US $1,054)			
	Acute hospitalization	−€816[c] (−US $1,034)	98
	Pharmacy medication	−€26 (−US $32.93)	3
	ED[e] admission	−€11.89 (−US $15.01)	1
	Visits and procedure	€22.29 (US $28.22)	−3
ICD[f] or CRT-D[g] studies			
PREDICT RM [27]			
Health care payer perspective			
US $, 2006: −$566 (−US $635)			
	Hospitalization	−US $554 (−US $621.94)	98
	Nonhospital cost	−US $12 (−US $13.44)	2
TARIFF[h] [30]			
Health care payer perspective			

XSL•FO
RenderX

Study and perspective	Total cost or income[a] impact compared to standard care in original currency in reference year (US $ pp[b] per year, 2019) and cost or income[a] impact drivers	Cost or income[a] impact drivers pp (US $ pp per year, 2019)	Cost or income[a] impact drivers as a percentage of total cost or income impact (%)
	€, 2011: −€562 (−US $712)		
	Cardiovascular hospitalization	−€454[c] (−US $575.1)	80
	Scheduled visit, protocol based	−€64.24[c] (−US $81.4)	11
	Outpatient diagnostic test	−€36.93[c] (−US $46.82)	7
	Unscheduled visit	€12.27[c] (US $15.6)	−2
	Emergency visit costs	−€15.67 (−US $19.8)	0.03
	Outpatient clinical evaluation	−€3.12 (−US $3.9)	0.005
Patient perspective			
	€, 2011: −€68 (−US $86)		
	Patient loss of work	−€42.34[c] (−US $53.6)	62
	Traveling	−€25.86[c] (−US $32.8)	38
Provider perspective			
	€, 2011: −€55 (−US $69)		
	Scheduled visit, protocol based	−€64.24[c] (−US $81.4)	117
	Unscheduled visit	€12.27[c] (US $15.6)	−22
	Outpatient clinical evaluation	−€3.12 (−US $3.9)	6
EVOLVO[i] [11]			
Health care payer perspective			
	€, 2010: −€167 (−US $219.5)		
	Hospitalization	−€223 (−US $292.5)	134
	Scheduled visit, protocol based	−€33.66[c] (−US $44.1)	20
	ED[h] and urgent visit	−€8.81[c] (−US $11.5)	5
	Nonurgent in-office visit	€10.68 (−US $14)	−6
	Diagnostic examinations	−€0.56 (−US $0.78)	0
	Scheduled remote visit	€32.20[c] (−US $42.2)	−19
	Unscheduled remote visit	€56.42[c] (−US $74)	−34
Patient perspective			
	€, 2010: −€90 (−US $117)		
	Scheduled visit, protocol based	−€96.90[c] (−US $127.6)	110
	ED and urgent visit	−€23.81[c] (−US $31.2)	27
	Nonurgent visit	€30.74 (US $40.32)	−35
MORE-CARE[j] [29]			
Health care payer perspective			
	€, 2014, no reimbursement: −€62.5 (−US $76)		
	Cardiovascular hospitalization	−€44.3 (−US $53.76)	71
	Scheduled visit, protocol based	−€37.4 (−US $45.4)	61
	ED visits	−€0.5 (−US $0.56)	−1
	Unscheduled visit	€6.4 (US $7.8)	−10

XSL•FO
RenderX

Study and perspective	Total cost or income[a] impact compared to standard care in original currency in reference year (US $ pp[b] per year, 2019) and cost or income[a] impact drivers	Cost or income[a] impact drivers pp (US $ pp per year, 2019)	Cost or income[a] impact drivers as a percentage of total cost or income impact (%)
	Device hospitalization	€13.3 (US $16.1)	−11
	€, 2014, with reimbursement: −€44.3 (−US $18)		
	Cardiovascular hospitalization	−€44.3 (−US $53.8)	306
	Unscheduled remote check	Maximum −€29.4 (−US $35.7)	203
	Scheduled remote check	Maximum −€18.6 (−US $22.5)	128
	Scheduled visit, protocol based	−€37.4 (−US $45.4)	39
	ED visits	−€0.5 (−US $0.6)	−1
	Unscheduled visit	€6.4 (US $7.8)	−44
	Device hospitalization	€13.3 (US $16.1)	−93
Burri et al [31], 2013			
Health care payer perspective			
	£, 2007: −£3.3 (−US $6.7)		
	Initial investment and in-hospital follow-up visit	−£3.3 (−US $6.7)	100
Raatikainen et al [32], 2008			
Health care payer perspective			
	€, 2006: −€641 (−US $914)		
	In-office visit[k], only 1 visit is protocol based	−€560.0 (−US $798.1)	87
	Traveling[k]	−€198.7 (−US $283.1)	31
	Remote monitoring[k]	€146.7 (−US $208.9)	−23
	Accommodation[k]	−€1.3 (−US $1.9)	0
	Sickness allowance[k]	−€28.0 (−US $39.9)	4
Patient perspective			
	€, 2006:−€59 (−US $84)		
	Patient fee[k]	−€58.7 (−US $83.6)	100
Al-Khatib et al [33], 2009			
Societal perspective			
	US $, 2009: −US $254 (−US $245)		
	Patient loss of work	−US $383 (−US $370.3)	150
	Traveling	−US $19 (−US $18.4)	7
	Follow-up visit, only 1 visit is protocol based	US $148 (US $ 143)	−58
CONNECT[l] [8]			
Health care payer perspective			
	US$, 2008: −US $1,434 (−US $1,243)		
	Mean cost per hospitalization	−US $1,434.4 (−US $1,243)	100
ECOST[m] [34]			
Health care payer perspective			
	€, 2011: −€927 (−US $1,175)		
	Cardiovascular hospitalization	−€720 (−US $912.3)	78

XSL•FO
RenderX

Study and perspective	Total cost or income[a] impact compared to standard care in original currency in reference year (US $ pp[b] per year, 2019) and cost or income[a] impact drivers	Cost or income[a] impact drivers pp (US $ pp per year, 2019)	Cost or income[a] impact drivers as a percentage of total cost or income impact (%)
	Device cost	−€533 (−US $675)	58
	Nonhospital cost	−€227[c] (−US $287.6)	24
	Other nonhospital cost	−€182 (−US $230.6)	20
	Cardiovascular treatment	−€113 (−US $143.1)	12
	Device management cost	−€74[c] (−US $93.7)	8
	ICD ambulatory visit, 3 visits are protocol based	−€40[c] (−US $50.6)	4
	Traveling	−€50 (−US $63.4)	0.05
	Reimbursement	€1000 (US $1,266)	−108
Patient perspective			
€, 2011: −€9 (−US $11.2)			
	Traveling	−€9 (−US $11.4)	100
EuroEco[n] [13]			
Provider perspective			
€, 2013: −€0.5 (−US $1.12)			
	Follow-up visit, protocol based	−€0.5 (−US $0.56)	100
Health care payer perspective			
€, 2013: −€287 (−US $349)			
	Hospitalization	−€301 (€326.9)	105
	Follow-up visit, protocol based	−€0.5 (−US $0.56)	−0.1
	Examination	€1.5 (US $1.8)	−0.5
	Other physician visit	€12.5 (US $15.2)	−4
SAVE-HM trial[d] [35]			
Societal perspective			
€, 2013: −€804.9[c] (−US $981)			
	Transport	−€787 (−US $959)	98
	Follow-up personnel cost, protocol based	−€17 (−US $21.3)	2
Chew et al [36], 2022			
Health care payer perspective			
CAD $, 2019: −$709.6 (−US $535.4)			
	Hospitalization	−$682.4[c] (−US $514.4)	96
	ED[e] visits	−$10.2 (−US $7.68)	1
	In-office visits	−$17 (−US $12.8)	2
Dario et al [37], 2016			
Health care payer perspective			
€, 2011: −€338 (−US $428)			
	Hospitalization	−€295 (−US $374)	96
	Visits and procedure, protocol based	−€40 (−US $50.6)	12
	Pharmacy medication	−€31 (−US $39.3)	9

XSL•FO
RenderX

Study and perspective	Total cost or income[a] impact compared to standard care in original currency in reference year (US $ pp[b] per year, 2019) and cost or income[a] impact drivers	Cost or income[a] impact drivers pp (US $ pp per year, 2019)	Cost or income[a] impact drivers as a percentage of total cost or income impact (%)
	Hospital medication	€1.37[c] (US $1.8)	−0.45
	ED[e] admission	€27[c] (US $34.2)	−8

[a]If the perspective is health care system or patient, then *cost* and if the perspective is provider, then *income*.

[b]pp: per patient (in health care and patient perspectives) or per physician (in provider perspective).

[c]The values are statistically significant.

[d]SAVE-HM: Socio-Economic Effects and Cost Saving Potential of Remote Patient Monitoring.

[e]ED: emergency department.

[f]ICD: implantable cardioverter-defibrillator.

[g]CRT-D: cardiac resynchronization therapy defibrillator.

[h]TARIFF: Health Economics Evaluation Registry for Remote Follow-Up.

[i]EVOLVO: Evolution of Management Strategies of Heart Failure Patients With Implantable Defibrillators.

[j]MORE-CARE: Monitoring Resynchronization Devices and Cardiac Patients.

[k]Costs were recalculated per patient.

[l]CONNECT: Clinical Evaluation of Remote Notification to Reduce Time to Clinical Decision.

[m]ECOST: Effectiveness and Cost of ICD Follow-Up Schedule With Telecardiology.

[n]EuroEco: European Health Economic Trial on Home Monitoring in ICD Patients.

Intervention and Comparator

Telemonitoring entailed data transmission and data review. Table 4 shows the frequencies of data transmission, review, and in-office visits of the included studies. In 47% (7/15) of the studies, data were transmitted continuously or daily [26,28,30,31,34,35]; in 20% (3/15) studies, data were transmitted after a device alert [8,11,29]; and in 13% (2/15) studies, data were transmitted every 3 months [32,33]. In 20% (3/15) of the studies, data review was performed daily [28,34,35]; however, in 40% (6/15) of the studies, it was performed after a device alert was received [8,11,26,29-31]. Besides data transmission and review, telemonitoring included scheduled in-office visits. In 33% (5/15) of the studies, all scheduled in-office visits were based on the protocol [11,13,29,30,35]. In 7% (1/15) of the studies, at least 1 scheduled in-office visit was protocol based [37]. In 3 (20%) of the 15 studies, only 1 scheduled in-office visit was protocol based [32-34]. Protocol-based in-office visits are described in Table 4.

XSL•FO
RenderX

Table 4. Frequencies of data transmission, review, and in-office visits of included studies.

Study	Brand of CIED[a]	Frequency of data transmission	Frequency of data review	Frequency of in-office visits	
				Telemonitoring	Standard care
Poniente [28]	Medtronic	Continuously	Daily	Mean: 4.4 pp[b]	Mean: 7.5 pp
PREDICT RM [27]	Boston	Not specified	Not specified	Patient dependent	Not specified
TARIFF[c] [30]	St-Jude	Continuously	After alert and 3 monthly	After 12 months	3 monthly
Nordland [26]	Biotronik	Daily	After alert	Not specified	Not specified
EVOLVO[d] [11]	Medtronic	After alert and 8 monthly	After alert and 8 monthly	8 monthly	4 monthly
MORE-CARE[e] [29]	Medtronic	After alert and after 4, 12, and 20 months	After alert and after 4, 12, and 20 months	8 monthly	4 monthly
Burri et al [31]	Biotronik	Daily	After alert	12 monthly	4 monthly
Raatikainen et al [32]	Medtronic	After 3 and 6 months	After 3 and 6 months	After 9 months	6 monthly
Al-Khatib et al [33]	Medtronic	3 monthly	3 monthly	After 6 months through telephone and after 12 months	3 monthly
CONNECT[f] [8]	Medtronic	After alert and 3 monthly	After alert and 3 monthly	After 1 month and 12 months	3 monthly
ECOST[g] [34]	Biotronik	Daily	Daily	After 1 to 3 months and thereafter 12 monthly	After 1 month to 3 months and thereafter 6 monthly
EuroEco[h] [13]	Biotronik	Continuously	Depending on the researcher	After 6 weeks and thereafter 12 monthly	After 6 weeks, thereafter 12 monthly, and planned visits depending on the hospital
SAVE-HM[i] trial [35]	Biotronik pacemaker	Daily	Daily	No	12 monthly
SAVE-HM trial [35]	Biotronik ICD[j]	Daily	Daily	12 monthly	6 monthly
Chew et al [36]	N/A[k]	N/A	N/A	N/A	N/A
Dario et al [37]	Biotronik, Medtronic, Boston, St-Jude, and Sorin Group pacemaker	Continuously	Daily	Not unless necessary	12 monthly
Dario et al [37]	Biotronik, Medtronic, Boston, St-Jude, and Sorin Group ICD	Continuously	Daily	At least 1	6 monthly

[a]CIED: cardiac implantable electronic device.

[b]pp: per patient.

[c]TARIFF: Health Economics Evaluation Registry for Remote Follow-Up.

[d]EVOLVO: Evolution of Management Strategies of Heart Failure Patients With Implantable Defibrillators.

[e]MORE-CARE: Monitoring Resynchronization Devices and Cardiac Patients.

[g]ECOST: Effectiveness and Cost of ICD Follow-Up Schedule With Telecardiology.

[h]EuroEco: European Health Economic Trial on Home Monitoring in ICD Patients.

[i]Save-HM: Socio-Economic Effects and Cost Saving Potential of Remote Patient Monitoring.

[j]ICD: implantable-cardioverter defibrillator.

[k]N/A: not applicable.

XSL•FO
RenderX

Effectiveness

Effectiveness results of telemonitoring can be found in Table 2. Of the 15 studies, 9 (60%) investigated a quality-adjusted life year (QALY) or quality of life (QOL) difference [11,26-30,33,34]. A total of 53% (8/15) of studies reported an increase in QALY or QOL [11,26-28,30,33,34,39], but the QALY or QOL increase was only statistically significant in 1 (13%; P=.03) of the 8 studies [26-28,30,33,34]. In contrast, only 1 (11%) of the 9 studies investigating QOL or QALY reported a significant decrease in QOL [29]. Comparing all studies, QALY differences ranged from 0.03 to 0.27 in patients with pacemakers and ranged from −1 to 0.64 in patients with ICD or CRT-D.

Besides QALY or QOL, several studies reported other health outcomes. Chew et al [36] indicated that the risk of death was lower with telemonitoring. Al-Khatib et al [33] reported that mortality and general patient satisfaction with telemonitoring were equal to those of standard care. Crossley et al [8] reported that the time between the clinical event and the clinical decision was 17.4 days shorter in patients with an ICD or CRT-D on telemonitoring than in those on standard care (P<.001). Burri et al [31] indicated that telemonitoring patients with ICD or CRT-D led to fewer inappropriate shocks (−51%) and a reduction in battery exhaustion (−7%). Raatikainen et al [32] indicated that telemonitoring patients with an ICD reduced the average total time spent on device follow-up, with 17 minutes per patient per follow-up for physicians and 175 minutes per patient per follow-up for patients. Similarly, Dario et al [37] indicated that the time spent by physicians to treat the patient reduced by an average of 4.1 minutes per follow-up in patients with pacemakers and an average of 13.7 minutes per follow-up in patients with an ICD (SD was not reported).

Economic Impact

The results of the economic impact of telemonitoring are presented in Table 2. Of the 15 studies, 4 (27%) investigated the cost impact of telemonitoring in patients with pacemakers [26,28,35,37]. From a health care payer perspective, 1 (25%) of the 4 pacemaker studies indicated that telemonitoring increased costs with US $2183 per patient per year (not statistically significant) mainly because of increased hospitalization costs [26]. A total of 2 (50%) of the 4 pacemaker studies indicated that telemonitoring reduced costs by US $8.9 and US $1054 per patient per year mainly because of a reduction in hospitalization and staff costs, respectively [28,37]. Therefore, hospitalizations reduced costs in the study by Dario et al [37] but increased costs in the study by Lopez-Villegas et al [26]. From a patient and societal perspective, the results indicated that telemonitoring reduced costs by US $11 and US $1113 per patient per year, respectively, mainly because of lower transport costs [28,35].

Of the 15 studies, 13 (87%) investigated the cost or income impact of telemonitoring in patients with an ICD or CRT-D [8,11,13,27,29-37]. A total of 11 (85%) of the 13 ICD or CRT-D studies investigated the cost impact of telemonitoring from a health care payer perspective, all indicating that telemonitoring reduced costs for patients with an ICD or CRT-D [8,11,13,27,29-32,34,36,37]. A total of 9 (82%) of the 11 health care payer perspective studies indicated that hospitalization was the largest driver for cost reduction for patients with an ICD or CRT-D [8,11,13,27,29,30,34,36,37]. The hospitalization cost reduced by up to US $912.3 per patient per year [34]. In addition, scheduled in-office visits were reported as a driver for cost reduction in 5 (45%) of the 11 health care payer perspective studies, as up to 61% of the total cost reduction was due to a decrease in the number of scheduled in-office visits [11,29,30,32,34]. Besides cost drivers that reduced costs, there were also drivers that increased costs. In 3 (27%) of the 11 health care payer perspective studies, unscheduled visits increased the total cost impact of telemonitoring [11,13,29,30,33]. A total of 3 (20%) of the 15 studies indicated that the cost reduction for scheduled in-office visits outweighed the cost increase for unscheduled in-office visits (−US $81.4 vs US $15.6, −US $45.4 vs US $7.8, and −US $44.1 vs US $14/patient/year) [11,29,30].

The results of 4 (31%) of the 13 ICD or CRT-D studies that investigated the cost impact of telemonitoring from the patients' perspective [11,30,32,34] indicated that patient and caregiver loss of work or activity [30], scheduled in-office visits [11], and transport [34] were the largest drivers for cost reduction. The results of 2 (15%) of the 13 ICD or CRT-D studies that investigated the income impact of telemonitoring from a provider perspective indicated that the loss of reimbursed (scheduled) in-office visits was the most important factor for income loss due to telemonitoring [13,30], reducing income by up to €72.7 (US $77.21) per patient per year [30].

ICER and Cost-Utility Ratio

Results on ICER and the cost-utility ratio are presented in Table 2. Of the 15 studies, 3 (20%) calculated the ICER from a health care payer perspective [26-28] and 1 (7%) calculated the cost-utility ratio from a health care payer perspective [11]. Of the 15 studies, 2 (13%) calculating ICER were conducted with patients with pacemakers [26,28]. Notably, of the 2 studies, 1 (50%) indicated that telemonitoring was cost-effective (ICER: US $270.09/QALY) [28], and 1 (50%) indicated that telemonitoring was not cost-effective (ICER: US $64,410/QALY) [26]. For patients with an ICD or CRT-D, of the 2 studies, 1 (50%) indicated that telemonitoring was cost-effective (ICER: US $12,069/QALY) [27] and 1 (50%) indicated that telemonitoring was dominant [11].

Critical Appraisal

The critical appraisal of the individual studies is provided in Tables 5 and 6. Of the 15 studies, 1 (7%) was classified as excellent (score of 100%) [13], 8 (53%) had a good quality score (100%<score>75%) [26,28,30,31,33,34,36,37], and 6 (40%) had a moderate quality score (75%<score>50%) [8,11,27,29,32,35]. A total of 3 (20%) of the 15 studies scored the lowest, with 59% each [8,29,32]. More than 50% (>8/15) of the studies scored low for the items *cost valuation* (item 9) [11,27-32,34,35,37], *discounting* (item 14) [8,11,29,30,32,34,35,37], and *no conflict of interest* (item 18) [8,11,27,29,30,32,34,35]. All studies scored high on the items *study population* (item 1), *study design* (item 4), *time horizon* (item 10), *outcome identification* (item 11), *outcome measurement* (item 12), and *ethics* (item 19).

XSL•FO
RenderX

Table 5. Quality assessment of the first 8 studies.

	EuroEco[a] [13]	Nordland [26]	Poniente [28]	Burri et al [31]	Al-Khatib et al [33]	Chew et al [36]	TARIFF[b] [30]	ECOST[c] [34]
Study population	✓[d]	✓	✓	✓	✓	✓	✓	✓
Competing alternatives	✓	X[e]	✓	✓	✓	X	✓	✓
Research question	✓	✓	X	✓	X	✓	✓	✓
Study design	✓	✓	✓	✓	✓	✓	✓	✓
Time horizon	✓	✓	✓	✓	✓	✓	✓	✓
Perspective	✓	✓	✓	✓	✓	✓	✓	✓
Cost identification	✓	✓	✓	X	✓	✓	✓	✓
Cost measurement	✓	✓	✓	✓	✓	✓	✓	✓
Cost valuation	✓	✓	X	X	✓	✓	X	X
Outcome identification	✓	✓	✓	X	✓	✓	✓	✓
Outcome measurement	✓	✓	✓	✓	✓	✓	✓	✓
Outcome valuation	✓	✓	✓	N/A[f]	N/A	N/A	✓	N/A
Incremental analysis	N/A	✓	✓	N/A	N/A	N/A	N/A	N/A
Discounting	✓	N/A	✓	✓	✓	✓	X	X
Sensitivity analysis	✓	✓	✓	✓	X	✓	✓	✓
Conclusions	✓	✓	✓	✓	✓	✓	✓	✓
Generalizability	✓	✓	✓	✓	✓	X	✓	✓
No conflicts of interest	✓	✓	✓	✓	✓	✓	X	X
Ethics	✓	✓	✓	✓	✓	✓	✓	✓
Values, n (%)	18 (100)	16 (94)	15 (89)	15 (88)	15 (88)	15 (88)	15 (83)	14 (82)

[a]EuroEco: European Health Economic Trial on Home Monitoring in ICD patients.

[b]TARIFF: Health Economics Evaluation Registry for Remote Follow-Up.

[c]ECOST: Effectiveness and Cost of ICD Follow-Up Schedule With Telecardiology.

[d]Sufficient attention was given to this aspect.

[e]Insufficient attention was given to this aspect.

[f]N/A: not applicable.

XSL•FO
RenderX

Table 6. Quality assessment of the last 7 studies.

	Dario et al [37]	PREDICT RM [27]	SAVE-HM[a] [35]	EVOLVO[b] [11]	Raatikainen et al [32]	MORE-CARE[c] [29]	CONNECT[d] [8]
Study population	✓[e]	✓	✓	✓	✓	✓	✓
Competing alternatives	✓	X[f]	✓	✓	✓	✓	✓
Research question	✓	X	X	X	✓	X	X
Study design	✓	✓	✓	✓	✓	✓	✓
Time horizon	✓	✓	✓	✓	✓	✓	✓
Perspective	✓	✓	✓	✓	X	✓	X
Cost identification	✓	X	X	✓	X	X	X
Cost measurement	✓	✓	✓	X	✓	X	X
Cost valuation	X	X	X	X	X	X	✓
Outcome identification	✓	✓	✓	✓	✓	✓	✓
Outcome measurement	✓	✓	✓	✓	✓	✓	✓
Outcome valuation	N/A[g]	✓	N/A	✓	N/A	N/A	N/A
Incremental analysis	N/A	✓	N/A	N/A	N/A	N/A	N/A
Discounting	X	✓	X	X	X	X	X
Sensitivity analysis	X	✓	✓	✓	X	X	✓
Conclusions	✓	✓	✓	✓	X	✓	✓
Generalizability	✓	✓	✓	X	✓	✓	X
No conflicts of interest	✓	X	X	X	X	X	X
Ethics	✓	✓	✓	✓	✓	✓	✓
Values, n (%)	14 (82)	14 (74)	12 (71)	12 (67)	10 (59)	10 (59)	10 (59)

[a]Save-HM: Socio-Economic Effects and Cost Saving Potential of Remote Patient Monitoring.

[b]EVOLVO: Evolution of Management Strategies of Heart Failure Patients With Implantable Defibrillators.

[c]MORE-CARE: Monitoring Resynchronization Devices and Cardiac Patients.

[d]CONNECT: Clinical Evaluation of Remote Notification to Reduce Time to Clinical Decision.

[e]Sufficient attention is given to this aspect.

[f]Insufficient attention is given to this aspect.

[g]N/A: not applicable.

Discussion

Principal Findings and Comparison With Prior Work

The primary aim of this study was to investigate the cost-effectiveness of telemonitoring patients with an ICD or CRT-D and a pacemaker from different perspectives.

From a health care payer perspective, most studies indicated that telemonitoring was a cost-saving and effective alternative to standard care. The most important driver for cost reduction was hospitalizations, both in patients with a pacemaker and those with an ICD or CRT-D. The cost of hospitalizations was reduced by up to US $912.3 per patient per year [34]. Moreover, the reduction of scheduled in-office visits was the second most important cost-saving factor in most ICD or CRT-D studies, with up to 61% of the total cost reduction. Previous research indicated that up to 55% of the device follow-ups were routine checks with no actionable events or device programming [35,40,41]. Several researchers pointed out that most scheduled in-office visits could be replaced by telemonitoring without affecting the quality of care [7,34] and with potentially diagnosing >99.5% of arrhythmia and device problems [41]. Although scheduled in-office visits decreased, our results show that unscheduled in-office visits increased because of telemonitoring patients with an ICD or CRT-D, probably because of the possible faster detection of arrhythmia and device malfunction by telemonitoring [8]. However, in all studies analyzing both scheduled and unscheduled in-office visits, the cost reduction for scheduled in-office visits outweighed the cost increase for unscheduled in-office visits [11,29,30].

From a patient perspective, our results indicated that the reduction of professional activity, transport time, and costs due to scheduled in-office visits are the most important factors for cost reduction.

The provider perspective was investigated less frequently in the included studies, although it is very relevant. Owing to the reduction of scheduled in-office visits, providers will lose

XSL•FO
RenderX

income with telemonitoring if no reimbursement exists for telemonitoring but only for in-office visits. As a result, providers will be stimulated to maintain the classic follow-up instead of telemonitoring. Of the 15 studies, 1 (7%) observed that the total cost for insurance payers does not increase in countries where telemonitoring is reimbursed [13]. As telemonitoring decreases the overall costs from a health care payer perspective, there is room for proper compensation for providers to transition from in-office care to remote care. Hence, correct compensation (which is possible while still saving on the overall health care cost) will stimulate providers to switch to telemonitoring as the desired care path for patients with a CIED.

All studies reported the effectiveness of telemonitoring. Of the 15 studies, 9 (60%) indicated a QALY or QOL difference. Furthermore, 89% (8/9) of these studies indicated an increase in QALYs or QOL for telemonitoring patients with pacemakers or ICD or CRT-D, ranging from −1 to 0.64. Some studies (3/9, 33%) indicated this QALY or QOL increase was the result of the reduced routine in-office visits [7,34]. However, the QALY or QOL increase was only statistically significant (and positive) in 1 (11%) of the 9 studies [11]. Nevertheless, patient questionnaires have demonstrated a high acceptance of telemonitoring among patients with pacemakers and those with ICDs [39]. Moreover, telemonitoring is reported to lead to an increased sense of security [39]. Furthermore, the results indicated that telemonitoring leads to fewer inappropriate shocks, an important determinant of QALY, in patients with an ICD or CRT-D [31].

The cost-effectiveness analyses may be sensitive to the heterogeneity among the organization of telemonitoring in different hospitals. This may include different devices, the number of transmissions, the configuration of alerts, and hospital visit scheduling [26]. It seems reasonable to expect that the efficiency of telemonitoring not only depends on the technology but also on the organization of the service. If hospitals see telemonitoring as an additional service, on top of standard care, less cost-savings may be seen than if hospitals see telemonitoring as a substitute for standard care. A radical organizational change could lead to larger cost-savings, as suggested by an observational study by Facchin et al [42]. Moreover, such radical change may include a strategy involving other physicians, such as general practitioners, and referring cardiologists, that is, an integrated health care delivery [37].

Furthermore, the comparison between studies is challenged by differences in study design. The Poniente study by Bautista-Mesa et al [28] followed up patients with pacemakers for 12 months and indicated a QALY increase of 0.09 for telemonitoring. However, after 5 years of follow-up, the results indicated a QALY decrease of 0.20 for telemonitoring. Bautista-Mesa et al [28] indicated that some of the telemonitoring benefits (eg, reduction of in-office visits) may not be appreciated in the long term. Therefore, the evolution of utilities may be different depending on the follow-up time. In addition, the results indicated that hospitalizations reduced costs in the study by Dario et al [37] but increased costs in the study by Lopez-Villegas et al [26]. This discrepancy might be explained because significantly fewer patients were included in the study by Lopez-Villegas et al (50 vs 2101 patients). None of the 25 patients in the conventional follow-up group were hospitalized, whereas 12% (3/25) of the patients were hospitalized in the remotely monitored group (all for pacemaker problems) [26]. Furthermore, the included studies relied disproportionally on male participants, except for the Nordland study [26]. This may be explained by the significant sex disparity in ICD implantation rates, pointed out by Ingelaere et al [43]. Ingelaere et al [43] could not completely explain these differences by prevalence differences of cardiomyopathies and imply a possible undertreatment of women. Another study [44] observed an undertreatment of women with coronary heart disease, as they are less likely to undergo coronary angiography. Therefore, men may undergo more expensive treatments than women. This can explain why the included cost-effectiveness studies may present an overly positive result. In addition, time differences may impact the quality and cost-effectiveness of telemonitoring, as telemonitoring may evolve over time. However, our results did not provide meaningful insights in this respect.

The cost-effectiveness analyses may be sensitive to the heterogeneity among health care systems. From a provider perspective, our results indicated that telemonitoring generates lesser profit than standard care in the absence of reimbursement. Therefore, the lack of reimbursement is generally perceived as a major implementation barrier to telemonitoring, affecting 80% of the centers [45]. Consequently, providers tend to continue with standard care instead of telemonitoring. However, from a health care payer perspective, our results indicated that telemonitoring was still cost-saving even with reimbursement [13,34]. To stimulate providers to use telemonitoring, provider compensation should be provided based on overall health care cost-savings, making telemonitoring possible if it is preferred as the way to deliver CIED follow-up care.

Limitations

Because of the large discrepancies between health care systems' organization, costs, access, delivery, quality, and reimbursement of cardiac care, any generalization may be perceived as inaccurate [37,46]. For instance, the included studies were mainly performed in Western countries. The results may not be generalizable to non-Western countries. Therefore, the cost-effectiveness results are contingent on the context in which they were analyzed [46]. Another limitation of this research is that 40% (6/15) of the included studies are not randomized controlled trials. These studies may have unobserved confounding factors that cannot be controlled for. Finally, cost analyses were excluded in this study because of our research objective. However, future cost analyses could draw a lot of information from analyzing these excluded studies.

Conclusions

Telemonitoring patients with CIED may be a cost-effective alternative to standard follow-up. Moreover, telemonitoring may lead to a cost reduction from a health care and patient perspective, mainly by the reduction of hospitalizations and scheduled in-office visits. Owing to the reduction in scheduled in-office visits, providers' income tends to decrease when implementing telemonitoring without proper reimbursement. Introducing appropriate reimbursement could make

XSL•FO
RenderX

telemonitoring sustainable for providers, while still being cost-effective from a health care payer perspective.

Acknowledgments

The authors would like to thank Dr Ingrid Kremer of Maastricht University for her help in the manuscript review. This work was supported by the Fund for Scientific Research Flanders (Fonds Wetenschappelijk Onderzoek Vlaanderen, grant 1SC9322N, 2021). RW is supported as a postdoctoral clinical researcher by the Fund for Scientific Research Flanders (Fonds Wetenschappelijk Onderzoek Vlaanderen).

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices.

Authors' Contributions

SR and LA were responsible for the conceptualization of the manuscript. SR, LA, RW, and HH acquired the financial support necessary for this paper and developed the methodology. SR analyzed and investigated the data. AP, LA, RW, and HH validated the results. SR was responsible for the first and final drafts. LA, RW, and HH were involved in editing the drafts. All authors approved the final manuscript.

Conflicts of Interest

RW reports research funding from Abbott, Biotronik, Boston Scientific, and Medtronic and speakers and consultancy fees from Medtronic, Boston Scientific, Biotronik, and Abbott. None of these payments were personal; all were handled through the University of Leuven. HH received personal lecture and consultancy fees from Abbott, Biotronik, Daiichi-Sankyo, Pfizer-BMS, Medscape, and Springer Healthcare Limited. He received unconditional research grants through the University of Antwerp and the University of Hasselt from Abbott, Bayer, Biotronik, Biosense Webster, Boston Scientific, Boehringer Ingelheim, Daicchi-Sankyo, Fibricheck or Qompium, Medtronic, and Pfizer-BMS, all outside the scope of this work. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 108 KB - jmir_v26i1e47616_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 14 KB - jmir_v26i1e47616_app2.docx]

Multimedia Appendix 3
Interpretation Consensus Health Economic Criteria list.
[DOCX File , 18 KB - jmir_v26i1e47616_app3.docx]

References

1. Blomström-Lundqvist C, Traykov V, Erba PA, Burri H, Nielsen JC, Bongiorni MG, et al. European Heart Rhythm Association (EHRA) international consensus document on how to prevent, diagnose, and treat cardiac implantable electronic device infections-endorsed by the Heart Rhythm Society (HRS), the Asia Pacific Heart Rhythm Society (APHRS), the Latin American Heart Rhythm Society (LAHRS), International Society for Cardiovascular Infectious Diseases (ISCVID) and the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Europace 2020 Apr 01;22(4):515-549 [FREE Full text] [doi: 10.1093/europace/euz246] [Medline: 31702000]
2. Wilkoff BL, Auricchio A, Brugada J, Cowie M, Ellenbogen KA, Gillis AM, et al. HRS/EHRA expert consensus on the monitoring of cardiovascular implantable electronic devices (CIEDs): description of techniques, indications, personnel, frequency and ethical considerations. Heart Rhythm 2008 Jun;5(6):907-925. [doi: 10.1016/j.hrthm.2008.04.013] [Medline: 18551743]
3. de la Torre Díez I, Garcia-Zapirain B, Méndez-Zorrilla A, López-Coronado M. Monitoring and follow-up of chronic heart failure: a literature review of eHealth applications and systems. J Med Syst 2016 Jul;40(7):179. [doi: 10.1007/s10916-016-0537-y] [Medline: 27286984]

XSL•FO
RenderX

4. Guédon-Moreau L, Lacroix D, Sadoul N, Clémenty J, Kouakam C, Hermida J, et al. A randomized study of remote follow-up of implantable cardioverter defibrillators: safety and efficacy report of the ECOST trial. Eur Heart J 2013 Feb;34(8):605-614 [FREE Full text] [doi: 10.1093/eurheartj/ehs425] [Medline: 23242192]
5. Ricci RP, Morichelli L, Santini M. Remote control of implanted devices through Home Monitoring technology improves detection and clinical management of atrial fibrillation. Europace 2009 Jan;11(1):54-61 [FREE Full text] [doi: 10.1093/europace/eun303] [Medline: 19011260]
6. García-Fernández FJ, Osca Asensi J, Romero R, Fernández Lozano I, Larrazabal JM, Martínez Ferrer J, et al. Safety and efficiency of a common and simplified protocol for pacemaker and defibrillator surveillance based on remote monitoring only: a long-term randomized trial (RM-ALONE). Eur Heart J 2019 Jun 14;40(23):1837-1846 [FREE Full text] [doi: 10.1093/eurheartj/ehz067] [Medline: 30793735]
7. Varma N, Epstein AE, Irimpen A, Schweikert R, Love C. Efficacy and safety of automatic remote monitoring for implantable cardioverter-defibrillator follow-up: the Lumos-T Safely Reduces Routine Office Device Follow-up (TRUST) trial. Circulation 2010 Jul 27;122(4):325-332 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.110.937409] [Medline: 20625110]
8. Crossley GH, Boyle A, Vitense H, Chang Y, Mead RH. The CONNECT (Clinical Evaluation of Remote Notification to Reduce Time to Clinical Decision) trial: the value of wireless remote monitoring with automatic clinician alerts. J Am Coll Cardiol 2011 Mar 8;57(10):1181-1189 [FREE Full text] [doi: 10.1016/j.jacc.2010.12.012] [Medline: 21255955]
9. Catalan-Matamoros D, Lopez-Villegas A, Tore-Lappegard K, Lopez-Liria R. Patients' experiences of remote communication after pacemaker implant: The NORDLAND study. PLoS One 2019;14(6):e0218521 [FREE Full text] [doi: 10.1371/journal.pone.0218521] [Medline: 31220146]
10. Masella C, Zanaboni P, Di Stasi F, Gilardi S, Ponzi P, Valsecchi S. Assessment of a remote monitoring system for implantable cardioverter defibrillators. J Telemed Telecare 2008;14(6):290-294. [doi: 10.1258/jtt.2008.080202] [Medline: 18776073]
11. Zanaboni P, Landolina M, Marzegalli M, Lunati M, Perego GB, Guenzati G, et al. Cost-utility analysis of the EVOLVO study on remote monitoring for heart failure patients with implantable defibrillators: randomized controlled trial. J Med Internet Res 2013;15(5):e106 [FREE Full text] [doi: 10.2196/jmir.2587] [Medline: 23722666]
12. Annemans L. Health Economics for Non-economists: An Introduction to the Concepts, Methods and Pitfalls of Health Economic Evaluations. Kalmthout, Belgium: Pelckmans Pro; 2018.
13. Heidbuchel H, Hindricks G, Broadhurst P, Van Erven L, Fernandez-Lozano I, Rivero-Ayerza M, et al. EuroEco (European Health Economic Trial on Home Monitoring in ICD Patients): a provider perspective in five European countries on costs and net financial impact of follow-up with or without remote monitoring. Eur Heart J 2015 Jan 14;36(3):158-169 [FREE Full text] [doi: 10.1093/eurheartj/ehu339] [Medline: 25179766]
14. Klersy C, Boriani G, De Silvestri A, Mairesse GH, Braunschweig F, Scotti V, et al. Effect of telemonitoring of cardiac implantable electronic devices on healthcare utilization: a meta-analysis of randomized controlled trials in patients with heart failure. Eur J Heart Fail 2016 Feb;18(2):195-204 [FREE Full text] [doi: 10.1002/ejhf.470] [Medline: 26817628]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
17. Wijnen B, Van Mastrigt G, Redekop WK, Majoie H, De Kinderen R, Evers S. How to prepare a systematic review of economic evaluations for informing evidence-based healthcare decisions: data extraction, risk of bias, and transferability (part 3/3). Expert Rev Pharmacoecon Outcomes Res 2016 Dec;16(6):723-732. [doi: 10.1080/14737167.2016.1246961] [Medline: 27762640]
18. Search strategies: the centre for reviews and dissemination. University of York. 2018. URL: http://www.crd.york.ac.uk/crdweb/searchstrategies.asp [accessed 2024-07-01]
19. Auener SL, Remers TE, van Dulmen SA, Westert GP, Kool RB, Jeurissen PP. The effect of noninvasive telemonitoring for chronic heart failure on health care utilization: systematic review. J Med Internet Res 2021 Sep 29;23(9):e26744 [FREE Full text] [doi: 10.2196/26744] [Medline: 34586072]
20. Lopez-Villegas A, Leal-Costa C, Perez-Heredia M, Villegas-Tripiana I, Catalán-Matamoros D. Knowledge update on the economic evaluation of pacemaker telemonitoring systems. Int J Environ Res Public Health 2021 Nov 18;18(22):12120 [FREE Full text] [doi: 10.3390/ijerph182212120] [Medline: 34831876]
21. Vehmeijer JT, Brouwer TF, Limpens J, Knops RE, Bouma BJ, Mulder BJ, et al. Implantable cardioverter-defibrillators in adults with congenital heart disease: a systematic review and meta-analysis. Eur Heart J 2016 May 07;37(18):1439-1448 [FREE Full text] [doi: 10.1093/eurheartj/ehv735] [Medline: 26873095]
22. Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol 2021 Oct;138:80-94. [doi: 10.1016/j.jclinepi.2021.06.030] [Medline: 34242757]

XSL•FO
RenderX

23. Evers S, Goossens M, de VH, van TM, Ament A. Criteria list for assessment of methodological quality of economic evaluations: consensus on Health Economic Criteria. Int J Technol Assess Health Care 2005;21(2):240-245. [Medline: 15921065]
24. Jiang X, Ming W, You JH. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019 Jun 17;21(6):e13166 [FREE Full text] [doi: 10.2196/13166] [Medline: 31210136]
25. Consumptieprijsindex. Statbel. URL: https://statbel.fgov.be/nl/themas/consumptieprijsindex/consumptieprijsindex#figures [accessed 2022-07-08]
26. Lopez-Villegas A, Catalan-Matamoros D, Peiro S, Lappegard KT, Lopez-Liria R. Cost-utility analysis of telemonitoring versus conventional hospital-based follow-up of patients with pacemakers. The NORDLAND randomized clinical trial. PLoS One 2020;15(1):e0226188 [FREE Full text] [doi: 10.1371/journal.pone.0226188] [Medline: 31995558]
27. Hummel JP, Leipold RJ, Amorosi SL, Bao H, Deger KA, Jones PW, et al. Outcomes and costs of remote patient monitoring among patients with implanted cardiac defibrillators: an economic model based on the PREDICT RM database. J Cardiovasc Electrophysiol 2019 Jul;30(7):1066-1077 [FREE Full text] [doi: 10.1111/jce.13934] [Medline: 30938894]
28. Bautista-Mesa RJ, Lopez-Villegas A, Peiro S, Catalan-Matamoros D, Robles-Musso E, Lopez-Liria R, et al. Long-term cost-utility analysis of remote monitoring of older patients with pacemakers: the PONIENTE study. BMC Geriatr 2020 Nov 16;20(1):474 [FREE Full text] [doi: 10.1186/s12877-020-01883-3] [Medline: 33198629]
29. Boriani G, Da Costa A, Quesada A, Ricci RP, Favale S, Boscolo G, et al. Effects of remote monitoring on clinical outcomes and use of healthcare resources in heart failure patients with biventricular defibrillators: results of the MORE-CARE multicentre randomized controlled trial. Eur J Heart Fail 2017 Mar;19(3):416-425 [FREE Full text] [doi: 10.1002/ejhf.626] [Medline: 27568392]
30. Ricci RP, Vicentini A, D'Onofrio A, Sagone A, Rovaris G, Padeletti L, et al. Economic analysis of remote monitoring of cardiac implantable electronic devices: results of the health economics evaluation registry for remote follow-up (TARIFF) study. Heart Rhythm 2017 Jan;14(1):50-57 [FREE Full text] [doi: 10.1016/j.hrthm.2016.09.008] [Medline: 27614025]
31. Burri H, Sticherling C, Wright D, Makino K, Smala A, Tilden D. Cost-consequence analysis of daily continuous remote monitoring of implantable cardiac defibrillator and resynchronization devices in the UK. Europace 2013 Nov;15(11):1601-1608 [FREE Full text] [doi: 10.1093/europace/eut070] [Medline: 23599169]
32. Raatikainen MJ, Uusimaa P, van Ginneken MM, Janssen JP, Linnaluoto M. Remote monitoring of implantable cardioverter defibrillator patients: a safe, time-saving, and cost-effective means for follow-up. Europace 2008 Oct;10(10):1145-1151 [FREE Full text] [doi: 10.1093/europace/eun203] [Medline: 18703585]
33. Al-Khatib SM, Piccini JP, Knight D, Stewart M, Clapp-Channing N, Sanders GD. Remote monitoring of implantable cardioverter defibrillators versus quarterly device interrogations in clinic: results from a randomized pilot clinical trial. J Cardiovasc Electrophysiol 2010 May;21(5):545-550. [doi: 10.1111/j.1540-8167.2009.01659.x] [Medline: 20021522]
34. Guédon-Moreau L, Lacroix D, Sadoul N, Clémenty J, Kouakam C, Hermida J, et al. Costs of remote monitoring vs. ambulatory follow-ups of implanted cardioverter defibrillators in the randomized ECOST study. Europace 2014 Aug;16(8):1181-1188 [FREE Full text] [doi: 10.1093/europace/euu012] [Medline: 24614572]
35. Perl S, Stiegler P, Rotman B, Prenner G, Lercher P, Anelli-Monti M, et al. Socio-economic effects and cost saving potential of remote patient monitoring (SAVE-HM trial). Int J Cardiol 2013 Nov 30;169(6):402-407. [doi: 10.1016/j.ijcard.2013.10.019] [Medline: 24383121]
36. Chew DS, Zarrabi M, You I, Morton J, Low A, Reyes L, et al. Clinical and economic outcomes associated with remote monitoring for cardiac implantable electronic devices: a population-based analysis. Can J Cardiol 2022 Jun;38(6):736-744. [doi: 10.1016/j.cjca.2022.01.022] [Medline: 35093464]
37. Dario C, Delise P, Gubian L, Saccavini C, Brandolino G, Mancin S. Large controlled observational study on remote monitoring of pacemakers and implantable cardiac defibrillators: a clinical, economic, and organizational evaluation. Interact J Med Res 2016 Jan 13;5(1):e4 [FREE Full text] [doi: 10.2196/ijmr.4270] [Medline: 26764170]
38. Crossley G, Boyle A, Vitense H, Sherfesee L, Mead RH. Trial design of the clinical evaluation of remote notification to reduce time to clinical decision: the clinical evaluation Of remote NotificatioN to rEduCe Time to clinical decision (CONNECT) study. Am Heart J 2008 Nov;156(5):840-846. [doi: 10.1016/j.ahj.2008.06.028] [Medline: 19061696]
39. Ricci RP, Morichelli L, Quarta L, Sassi A, Porfili A, Laudadio MT, et al. Long-term patient acceptance of and satisfaction with implanted device remote monitoring. Europace 2010 May;12(5):674-679. [doi: 10.1093/europace/euq046] [Medline: 20200019]
40. Boriani G, Auricchio A, Klersy C, Kirchhof P, Brugada J, Morgan J, et al. Healthcare personnel resource burden related to in-clinic follow-up of cardiovascular implantable electronic devices: a European Heart Rhythm Association and Eucomed joint survey. Europace 2011 Aug;13(8):1166-1173. [doi: 10.1093/europace/eur026] [Medline: 21345922]
41. Heidbüchel H, Lioen P, Foulon S, Huybrechts W, Ector J, Willems R, et al. Potential role of remote monitoring for scheduled and unscheduled evaluations of patients with an implantable defibrillator. Europace 2008 Mar;10(3):351-357 [FREE Full text] [doi: 10.1093/europace/eun010] [Medline: 18245771]

XSL•FO
RenderX

42. Facchin D, Baccillieri MS, Gasparini G, Zoppo F, Allocca G, Brieda M, et al. Findings of an observational investigation of pure remote follow-up of pacemaker patients: is the in-clinic device check still needed? Int J Cardiol 2016 Oct 01;220:781-786. [doi: 10.1016/j.ijcard.2016.06.162] [Medline: 27394974]
43. Ingelaere S, Hoffmann R, Guler I, Vijgen J, Mairesse GH, Blankoff I, et al. Inequality between women and men in ICD implantation. Int J Cardiol Heart Vasc 2022 Aug;41:101075 [FREE Full text] [doi: 10.1016/j.ijcha.2022.101075] [Medline: 35782706]
44. Steingart RM, Packer M, Hamm P, Coglianese ME, Gersh B, Geltman EM, et al. Sex differences in the management of coronary artery disease. Survival and Ventricular Enlargement Investigators. N Engl J Med 1991 Jul 25;325(4):226-230. [doi: 10.1056/NEJM199107253250402] [Medline: 2057023]
45. Mairesse GH, Braunschweig F, Klersy K, Cowie MR, Leyva F. Implementation and reimbursement of remote monitoring for cardiac implantable electronic devices in Europe: a survey from the health economics committee of the European Heart Rhythm Association. Europace 2015 May;17(5):814-818. [doi: 10.1093/europace/euu390] [Medline: 25713012]
46. Lucà F, Cipolletta L, Di Fusco SA, Iorio A, Pozzi A, Rao CM, et al. Remote monitoring: doomed to let down or an attractive promise? IJC Heart Vasculature 2019 Sep;24:100380. [doi: 10.1016/j.ijcha.2019.100380]

Abbreviations

CHEC: Consensus Health Economic Criteria
CIED: cardiac implantable electronic device
CRT-D: cardiac resynchronization therapy defibrillator
ICD: implantable cardioverter-defibrillator
ICER: incremental cost-effectiveness ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PROSPERO: International Prospective Register of Systematic Reviews
QALY: quality-adjusted life year
QOL: quality of life

Edited by T Leung; submitted 28.03.23; peer-reviewed by P Jeurissen, B Dechert; comments to author 07.09.23; revised version received 13.09.23; accepted 13.02.24; published 19.04.24.

Please cite as:
Raes S, Prezzi A, Willems R, Heidbuchel H, Annemans L
Investigating the Cost-Effectiveness of Telemonitoring Patients With Cardiac Implantable Electronic Devices: Systematic Review
J Med Internet Res 2024;26:e47616
URL: https://www.jmir.org/2024/1/e47616
doi:10.2196/47616
PMID:38640471

©Sarah Raes, Andrea Prezzi, Rik Willems, Hein Heidbuchel, Lieven Annemans. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Patient and Staff Experience of Remote Patient Monitoring—What to Measure and How: Systematic Review

Valeria Pannunzio[1], PhD; Hosana Cristina Morales Ornelas[2], MSc; Pema Gurung[3], MSc; Robert van Kooten[4], MD, PhD; Dirk Snelders[1], PhD; Hendrikus van Os[5], MD, PhD; Michel Wouters[6], MD, PhD; Rob Tollenaar[4], MD, PhD; Douwe Atsma[7], MD, PhD; Maaike Kleinsmann[1], PhD

[1]Department of Design, Organisation and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands
[2]Department of Sustainable Design Engineering, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands
[3]Walaeus Library, Leiden University Medical Center, Leiden, Netherlands
[4]Department of Surgery, Leiden University Medical Center, Leiden, Netherlands
[5]National eHealth Living Lab, Department of Public Health & Primary Care, Leiden University Medical Center, Leiden, Netherlands
[6]Department of Surgery, Netherlands Cancer Institute – Antoni van Leeuwenhoek, Amsterdam, Netherlands
[7]Department of Cardiology, Leiden University Medical Center, Leiden, Netherlands

Corresponding Author:
Valeria Pannunzio, PhD
Department of Design, Organisation and Strategy
Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15
Delft, 2628 CE
Netherlands
Phone: 31 15 27 81460
Email: v.pannunzio@tudelft.nl

Abstract

Background: Patient and staff experience is a vital factor to consider in the evaluation of remote patient monitoring (RPM) interventions. However, no comprehensive overview of available RPM patient and staff experience–measuring methods and tools exists.

Objective: This review aimed at obtaining a comprehensive set of experience constructs and corresponding measuring instruments used in contemporary RPM research and at proposing an initial set of guidelines for improving methodological standardization in this domain.

Methods: Full-text papers reporting on instances of patient or staff experience measuring in RPM interventions, written in English, and published after January 1, 2011, were considered for eligibility. By “RPM interventions,” we referred to interventions including sensor-based patient monitoring used for clinical decision-making; papers reporting on other kinds of interventions were therefore excluded. Papers describing primary care interventions, involving participants under 18 years of age, or focusing on attitudes or technologies rather than specific interventions were also excluded. We searched 2 electronic databases, Medline (PubMed) and EMBASE, on February 12, 2021.We explored and structured the obtained corpus of data through correspondence analysis, a multivariate statistical technique.

Results: In total, 158 papers were included, covering RPM interventions in a variety of domains. From these studies, we reported 546 experience-measuring instances in RPM, covering the use of 160 unique experience-measuring instruments to measure 120 unique experience constructs. We found that the research landscape has seen a sizeable growth in the past decade, that it is affected by a relative lack of focus on the experience of staff, and that the overall corpus of collected experience measures can be organized in 4 main categories (service system related, care related, usage and adherence related, and health outcome related). In the light of the collected findings, we provided a set of 6 actionable recommendations to RPM patient and staff experience evaluators, in terms of both what to measure and how to measure it. Overall, we suggested that RPM researchers and practitioners include experience measuring as part of integrated, interdisciplinary data strategies for continuous RPM evaluation.

Conclusions: At present, there is a lack of consensus and standardization in the methods used to measure patient and staff experience in RPM, leading to a critical knowledge gap in our understanding of the impact of RPM interventions. This review

XSL•FO
RenderX

offers targeted support for RPM experience evaluators by providing a structured, comprehensive overview of contemporary patient and staff experience measures and a set of practical guidelines for improving research quality and standardization in this domain.

(J Med Internet Res 2024;26:e48463) doi:10.2196/48463

KEYWORDS

remote patient monitoring; telemonitoring; patient experience; staff experience; monitoring; intervention; health outcome; adherence; patient monitoring

Introduction

Background and Aim

This is a scenario from the daily life of a patient:

> *A beeping sound, and a message appears on the smartphone screen: "Reminder: check glucose before bedtime." Time to go to sleep, indeed, you think while putting down your book and reaching for the glucometer. As you wipe the drop of blood away, you make sure that both Bluetooth and Wi-Fi are on in your phone. Then, the reading is sent: you notice it seems to be rather far from your baseline. While you think of what you might have done differently, a slight agitation emerges: Is this why you feel so tired? The phone beeps again: "Your last glucose reading seems atypical. Could you please try again? Remember to follow these steps." Groaning, you unwrap another alcohol wipe, rub your finger with it, and test again: this time, the results are normal.*

Some patients will recognize certain aspects of this scenario, particularly the ones using a form of remote patient monitoring (RPM), sometimes referred to as remote patient management. RPM is a subset of digital health interventions that aim to improve patient care through digitally transmitted, health-related patient data [1]. Typically, RPM interventions include the use of 1 or more sensors (including monitoring devices, wearables, or implants), which collect patient data in or out of the hospital to be used for remote clinical decision-making. Partly due to a rapid expansion during the COVID-19 pandemic [2-5], the RPM domain has by now expanded to reach a broad range of medical specialties, sensing technologies, and clinical contexts [1,6,7].

RPM is presented as a strategy for enabling health care systems worldwide to face the pressing challenges posed by aging populations [8-10], including the dwindling availability of health care workers [11] and rising health care costs [12]. This is because deploying effective RPM solutions across health systems holds the potential to reduce health care resources use, while maintaining or improving care quality. However, evidence regarding RPM effectiveness at scale is mixed [13]. Few large-scale trials demonstrating a meaningful clinical impact of RPM have been conducted so far, and more research is urgently needed to clarify and address determinants of RPM effectiveness [7].

Among these determinants, we find the experience of patients and staff using RPM interventions. As noticeable in the introductory scenario, RPM introduces radical experiential changes compared to in-person care; patients might be asked to download and install software; pair, charge, and wear monitoring devices; submit personal data; or attend alerts or calls, all in the midst of everyday life contexts and activities. Similarly, clinical and especially nursing staff might be asked to carry out data analysis and administrative work and maintain remote contact with patients, often without a clear definition of roles and responsibilities and in addition to usual tasks [14].

Because of these changes, patient and staff experience constitutes a crucial aspect to consider when evaluating RPM interventions. Next to qualitative methods of experience evaluation, mixed and quantitative methods are fundamental, especially to capture information from large pools of users. However, the current RPM experience-measuring landscape suffers from a lack of methodological standardization, reflected in both what is measured and how it is measured. Regarding what is measured, it has been observed that a large number of constructs are used in the literature, often without a clear specification of their significance. This can be noticed even regarding popular constructs, such as satisfaction: Mair and Whitten [15], for instance, observe how the meaning of the satisfaction construct is seldom defined in patient surveys, leaving readers "unable to discern whether the participants said they were satisfied because telemedicine didn't kill them, or that it was 'OK,' or that it was a wonderful experience." Previous work also registers a broad diversity in the instruments used to measure a specific construct. For instance, in their review of RPM interventions for heart failure, Kraai et al [16] report that none of the papers they examined used the same survey to measure patient satisfaction, and only 1 was assessed on validity and reliability.

In this proliferation of constructs and instruments, no comprehensive overview exists of their application to measuring patient and staff experience in the RPM domain. The lack of such an overview negatively affects research in this domain in at least 2 ways. At the level of primary research, RPM practitioners and researchers have little guidance on how to include experience measuring in their study designs. At the level of secondary research, the lack of consistently used measures makes it hard to compare results between different studies and RPM solutions. Altogether, the lack of standardization in experience measuring constitutes a research gap that needs to be bridged in order for RPM to fully deliver on its promises.

In this review, this gap is addressed through an effort to provide a structured overview of patient and staff experience constructs and instruments used in RPM evaluation. First, we position the role of RPM-related patient and staff experience within the broader system of care using the Quadruple Aim framework.

XSL•FO
RenderX

Next, we describe the systematic review we performed of patient and staff experience–relevant constructs and instruments used in contemporary research aimed at evaluating RPM interventions. After presenting and discussing the results of this review, we propose a set of guidelines for RPM experience evaluators and indicate directions for further research.

The Role of Patient and Staff Experience in RPM

Many characterizations of patient and staff experience exist [17-19], some of which distinguish between determinants of experience and experience manifestations [20]. For our review, we maintained this distinction, as we aimed to focus on the broad spectrum of factors affecting and affected by patient and staff experience. To do so, we adopted the general conceptualization of patient and staff experience as characterized in the Quadruple Aim, a widely used framework for health system optimization centered around 4 overarching goals: improving the individual experience of care, improving the experience of providing care, improving the health of populations, and reducing the per capita cost of care [21]. Adopting a Quadruple Aim perspective allows health system researchers and innovators to recognize not only the importance of patient and staff experience in their own rights but also the inextricable relations of these 2 goals to the other dimensions of health system performance [22]. To clarify the nature of these relations in the RPM domain, we provide a schematic overview in Figure 1.

Figure 1. Schematic overview of the relations between patient and staff experience in RPM and the other components of the Quadruple Aim framework. Each arrow symbolizes a relation.

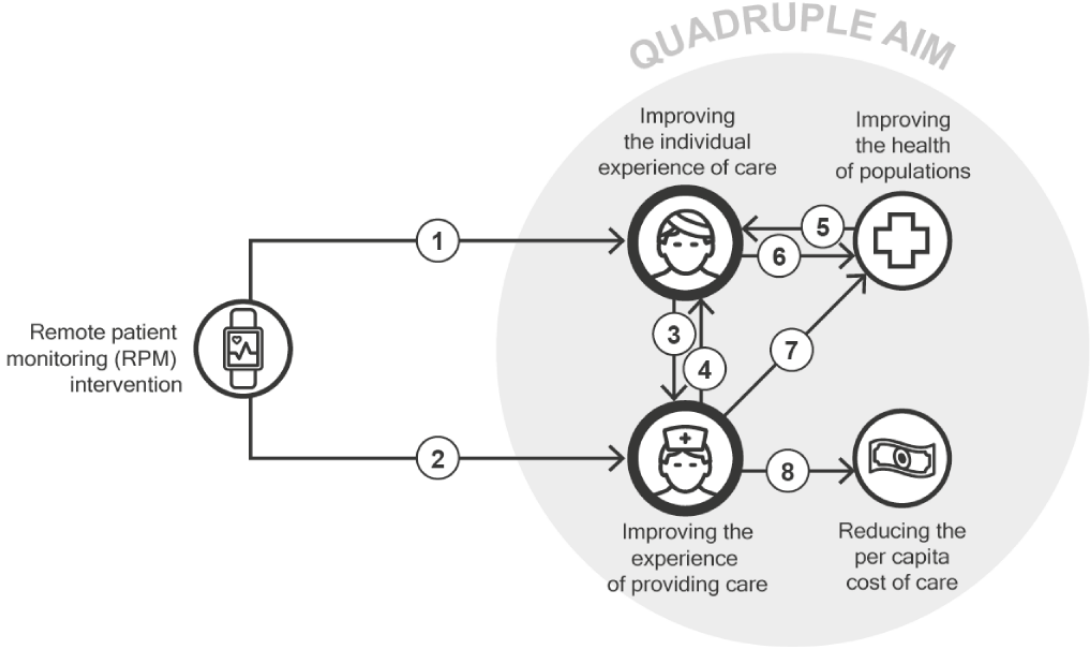

Next, we refer to the numbers in Figure 1 to touch upon prominent relationships between patient and staff experience in RPM within the Quadruple Aim framework and provide examples of experience constructs relevant to each relationship:

- Numbers 1 and 2: The characteristics of specific RPM interventions directly affect the patient and staff experience. Examples of experience constructs related to this mechanism are expressed in terms of *usability* or *wearability*, which are attributes of systems or products contributing to the care experience of patients and the work experience of staff.
- Numbers 3 and 4: Patient and staff experiences relate to each other through care delivery. Human connections, especially in the form of carer-patient relationships, represent a major factor in both patient and staff experience. An example of experience constructs related to this mechanism is expressed in terms of the *quality of the relationship*.
- Numbers 5 and 6: A major determinant of patient experience is represented by the health outcomes achieved as a result of the received care. An example of a measure of quality related to this mechanism is expressed in terms of the *quality of life*, which is an attribute of patient experience directly affected by a patient's health status. In contrast, patient experience itself is a determinant of the clinical effectiveness of RPM interventions. For example, the patient experience afforded by a given intervention is a determinant of both *adoption* of and *adherence* to that intervention, ultimately affecting its clinical impact. In a recent review, for instance, low patient adherence was identified as the main factor associated with ineffective RPM services [23].
- Number 7: Similarly, staff experience can be a determinant of clinical effectiveness. Experience-related issues, such as *alarm fatigue*, contribute to medical errors and lower the quality of care delivery [24].
- Number 8: Staff experience can also impact the cost of care. For example, the *time effort* required for the use of a given intervention can constitute a source of extra costs. More indirectly, low staff *satisfaction* and excessive *workload* can increase health care staff turnover, resulting in additional expenses at the level of the health system.

XSL•FO
RenderX

Overall, the overview in Figure 1 can help us grasp the nuances of the role of patient and staff experience on the overall impact of RPM interventions, as well as the importance of measuring experience factors, not only in isolation, but also in relation to other dimensions of care quality. In this review, we therefore covered a broad range of experience-relevant factors, including both experiential determinants (eg, usability) and manifestations (eg, adherence). Overall, this study aimed to obtain a comprehensive set of experience constructs and corresponding measurement instruments used in contemporary RPM research and to propose an initial set of guidelines for improving methodological standardization in this domain.

Methods

Protocol Registration and PRISMA Guidelines

The study protocol was registered in the PROSPERO (International Prospective Register of Systematic Reviews) database (CRD42021250707). This systematic review adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The PRISMA checklist is provided in Multimedia Appendix 1 [25].

Criteria for Study Eligibility

Our study population consisted of adult (≥18 years old) patients and staff members involved as participants in reported RPM evaluations. Full-text papers reporting instances of patient and staff experience measuring in RPM interventions, written in English, and published after January 1, 2011, were considered for eligibility.

For the scope of our review, we considered as RPM any intervention possessing the following characteristics:

- Sensor-based patient monitoring, intended as the use of at least 1 sensor to collect patient information at a distance. Therefore, we excluded interventions that were purely based on the collection of "sensor-less" self-reported measures from patients. This is because we believe the use of sensors constitutes a key element of RPM and one that strongly contributes to experiential aspects in this domain. However, we adopted a broad definition of "sensor," considering as such, for instance, smartphone cameras (eg, postoperative wound-monitoring apps) and analog scales or thermometers (eg, interventions relying on patients submitting manually entered weights or temperatures). By "at a distance," we meant not only cases in which data were transferred from nonclinical environments, such as home monitoring, but also cases such as tele–intensive care units (tele-ICUs), in which data were transferred from one clinical environment to another. Furthermore, we included interventions relying on both continuous and intermittent monitoring.
- Clinical decision-making as an intended use of remotely collected data. Therefore, we excluded interventions in which the collected data were meant to be used exclusively for research purposes and not as a stage of development of an RPM intervention to be adopted in patient care. For instance, we excluded cases in which the remotely collected patient data were only used to test research hypotheses unrelated to the objective of implementing RPM interventions (eg, for drug development purposes). This is because in this review we were interested in RPM as a tool for the provision of remote patient care, rather than as an instrument for research. We also excluded interventions in which patients themselves were the only recipients of the collected data and no health care professional was involved in the data analysis and use.

Furthermore, we excluded:

- Evaluations of attitudes, not interventions: contributions in which only general attitudes toward RPM in abstract were investigated, rather than 1 or more specific RPM interventions.
- Not reporting any evaluation: contributions not focusing on the evaluation of 1 or more specific RPM interventions, for instance, papers providing theoretical perspectives on the field (eg, research frameworks or theoretical models).
- Evaluation of technology, not interventions: contributions only focused on evaluating RPM-related technology, for instance, papers focused on testing sensors, software, or other service components in isolation rather than as a part of any specific RPM intervention.
- Not just RPM: contributions not specifically focused on RPM but including RPM interventions in their scope of research, for instance, papers reporting on surveys obtained from broad cohorts of patients (including RPM recipients) in a noncontrolled way. An example of such contributions would be represented by studies focusing on patient experience with mobile health apps in general, covering both interventions involving RPM and interventions not including any kind of patient monitoring, without a clear way to distinguish between the 2 kinds of interventions in the contribution results. This was chosen in order to maintain the review focus on RPM interventions. Instead, papers including both RPM and other forms of care provisions within the same intervention were included, as well as papers comparing RPM to non-RPM interventions in a controlled way.
- Primary care intervention only: interventions only involving general practitioners (GPs) and other primary care practitioners as health care providers of the RPM intervention. This is because we expected marked differences between the implementation of RPM in primary care and at other levels of care, due to deep dissimilarities in settings, workflows, and routines. Examples of RPM interventions *only* involving primary care providers included kiosk systems (for which a common measuring point was provided to many patients) or pharmacy-managed medication-monitoring programs. RPM interventions involving primary care providers *and* providers from higher levels of care, however, were included in the review.
- Staff-to-staff intervention: contributions reporting on interventions exclusively directed at staff, for instance, papers reporting on RPM methods aimed at monitoring stress levels of health care workers.
- Target group other than patient or staff: contributions aimed at collecting experience measures in target groups other than patients or staff, for instance, papers investigating the experience in RPM for informal caregivers.

XSL•FO
RenderX

Search Method

To identify relevant publications, the following electronic databases were searched: (1) Medline (PubMed) and (2) EMBASE. Search terms included controlled terms from Medical Subject Headings (MeSH) in PubMed and Emtree in EMBASE, as well as free-text terms. Query term selection and structuring were performed collaboratively by authors VP, HCMO, and PG (who is a clinical librarian at the Leiden University medical library). The full search strategies are reported in Multimedia Appendix 2. Because the aim of the review was to paint a contemporary picture of experience measures used in RPM, only studies published starting from January 1, 2011, were included.

Study Selection

Study selection was performed by VP and HCMO, who used Rayyan, an online research tool for managing review studies [26], to independently screen both titles and abstracts in the initial screening and full texts in the final screening. Discrepancies were solved by discussion. A flowchart of study selection is depicted in Figure 2.

Figure 2. Flowchart of study selection. RPM: remote patient monitoring.

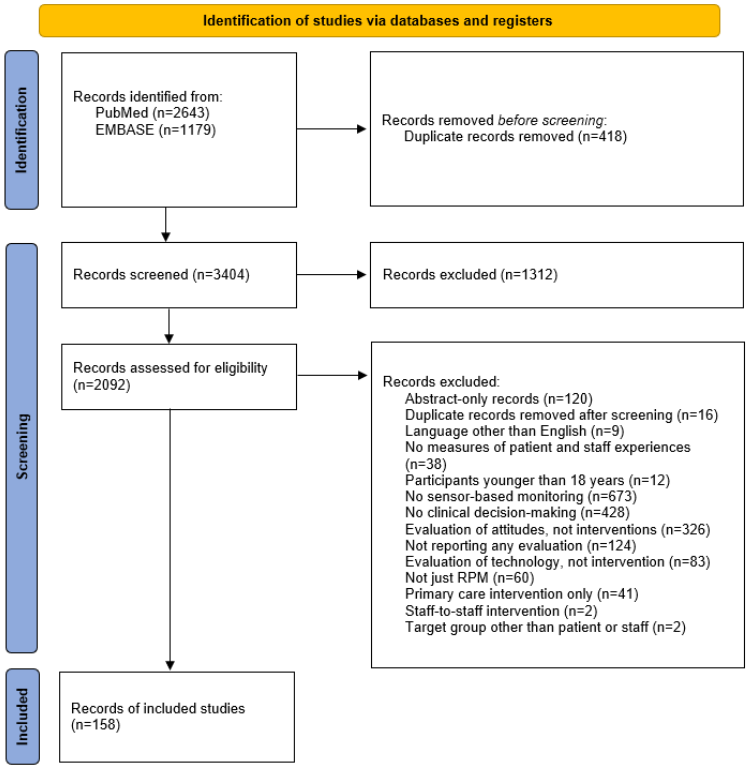

Quality Appraisal

The objective of this review was to provide a comprehensive overview of the relevant literature, rather than a synthesis of specific intervention outcomes. Therefore, no papers were excluded based on the quality appraisal, in alignment with similar studies [27].

Data Extraction and Management

Data extraction was performed independently by VP and HCMO. The extraction was performed in a predefined Microsoft Excel sheet designed by VP and HCMO. The sheet was first piloted in 15 included studies and iterated upon to optimize the data extraction process. The full text of all included studies was retrieved and uploaded in the Rayyan environment. Next, the full text of each included study was examined and relevant data were manually inputted in the predefined Excel sheet. Discrepancies were resolved by discussion. The following data types were extracted: (1) general study information (authors, title, year of publication, type of study, country or countries); (2) target disease(s), intervention, or clinical specialty; (3) used patient or staff experience evaluation instrument and measured experience construct; (4) evidence base, if indicated; and (5) number of involved staff or patient participants. By "construct," we referred to the "abstract idea, underlying theme, or subject matter that one wishes to measure using survey questions" [28]. To identify the measured experience construct, we used the definition provided in the source contribution, whenever available.

Data Analysis

First, we analyzed the collected data through building general overviews depicting the kind of target participants (patients or staff) of the collected experience measures and their use over time. To organize the diverse set of results collected through the systematic review, we then performed a correspondence analysis (CA) [29], a multivariate statistical technique used for exploring and displaying relationships between categorical data. CA transforms a 2-way table of frequencies between a row and a column variable into a visual representation of relatedness between the variables. This relatedness is expressed in terms of inertia, which represents "a measure of deviation from

XSL•FO
RenderX

independence" [30] between the row and column variables. Any deviations from the frequencies expected if the row and column variables were completely independent from each other contribute to the total inertia of the model. CA breaks down the inertia of the model by identifying mutually independent (orthogonal) dimensions on which the model inertia can be represented. Each successive dimension explains less and less of the total inertia of the model. On each dimension, relatedness is expressed in terms of the relative closeness of rows to each other, as well as the relative closeness of columns to each other. CA has been previously used to find patterns in systematic review data in the health care domain [31].

In our case, a 2-way table of frequencies was built based on how often any given instrument (eg, System Usability Scale [SUS]) was used to measure any given construct (eg, usability) in the included literature. Therefore, in our case, the total inertia of the model represented the amassed evidence base for relatedness between the collected experience constructs and measures, based on how they were used in the included literature.

To build the table of frequencies, the data extracted from the systematic review underwent a round of cleaning, in which the formulation of similar constructs was made more homogeneous: for instance, "time to review," "time to response," and "time for task" were merged under 1 label, "time effort." An overview of the merged construct formulations is provided in Multimedia Appendix 3. The result of the CA was a model where 2 dimensions contributed to more than 80% of the model's inertia (explaining 44.8% and 35.7%, respectively) and where none of the remaining 59 dimensions contributed more than 7.3% to the remaining inertia. This gap suggests the first 2 dimensions to express meaningful relationships that are not purely based on random variation. A 2D solution was thus chosen.

Results

General Observations

A total of 158 studies reporting at least 1 instance of patient or staff experience measuring in RPM were included in the review. The included studies covered a broad range of RPM interventions, most prominently diabetes care (n=30, 19%), implanted devices (n=12, 7.6%), and chronic obstructive pulmonary disease (COPD; n=10, 6.3%). From these studies, we reported 546 experience-measuring instances in RPM, covering 160 unique experience-measuring instruments used to measure 120 unique experience constructs.

Our results included 4 kinds of versatile (intended as nonspecific) experience-measuring instruments: the custom survey, log file analysis, protocol database analysis, and task analysis. All of them can be used for measuring disparate kinds of constructs:

- By "custom survey," we refer to survey instruments created to evaluate patient or staff experience in connection to 1 specific RPM study and only for that study.
- By "log file analysis," we refer to the set of experience assessment methods based on the automatic collection of data through the RPM digital infrastructures themselves [32]; examples are clicks, uploads, views, or other forms of interactions between users and the RPM digital system. This set of methods is typically used to estimate experience-relevant constructs, such as adherence and compliance.
- By "protocol database analysis," we refer to the set of experience assessment methods based on the manual collection of data performed by RPM researchers within a specific research protocol; an example of a construct measured with these instruments is the willingness to enroll.
- By "task analysis," we refer to the set of experience assessment methods based on the real-life observation of users interacting with the RPM system [33].

In addition to these 4 instruments, our results included a large number of specific instruments, such as standard indexes, surveys, and questionnaires. Overall, the most frequently reported instrument was, by far, the custom survey (reported in 155/546, 28.39%, instances), while the most frequently reported experience construct was satisfaction (85/546, 15.57%), closely followed by quality of life (71/546, 13%).

Target Participants and Timeline

We found large differences in the number of RPM-relevant experience constructs and instruments used for patients and for staff (see Figure 3). We also found instruments used for both patients and staff. Either these were broadly used instruments (eg, the SUS) that were administered to both patients and staff within the same study, or they were measures of interactions between patients and staff (eg, log file analysis instruments recording the number of remote contacts between patients and nursing assistants).

Figure 3. Count of mentioned instances of experience constructs organized by target participant: patient, staff, or both. Different shades of gray indicate different constructs.

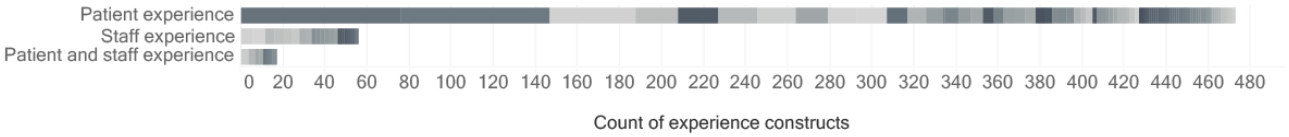

RPM research appears to focus much more on patient experience than on staff experience, which was investigated in only 20 (12.7%) of the 158 included papers. Although it is possible that our exclusion criteria contributed to the paucity of staff experience measures, only 2 (0.1%) of 2092 studies were excluded for reporting on interventions directed exclusively at

XSL•FO
RenderX

staff. Of the 41 (2%) studies we excluded for reporting on primary care interventions, we found 6 (15%) studies reporting on staff experience, a rate comparable to the one in the included sample. Furthermore, although our choice to exclude papers reporting on the RPM experience of informal caregivers might have contributed to a reduction in the number of collected constructs and measures, only 2 (0.1%) of 2092 studies were excluded for this reason, and the constructs used in these contributions were not dissimilar from the ones found in the included literature.

Among the included contributions that did investigate staff experience, we noticed that the number of participant staff members involved in the reported studies was only reported in a minority of cases (9/20, 45%).

Furthermore, a time-based overview of the collected results (Figure 4) provided us with an impression of the expansion of the field in the time frame of interest for both patient and staff experience measures.

Figure 4. Count of mentioned instances of experience constructs for patients (left) or for staff, and patients and staff (right) in the included literature from 2011 to 2020. Different shades of gray indicate different constructs.

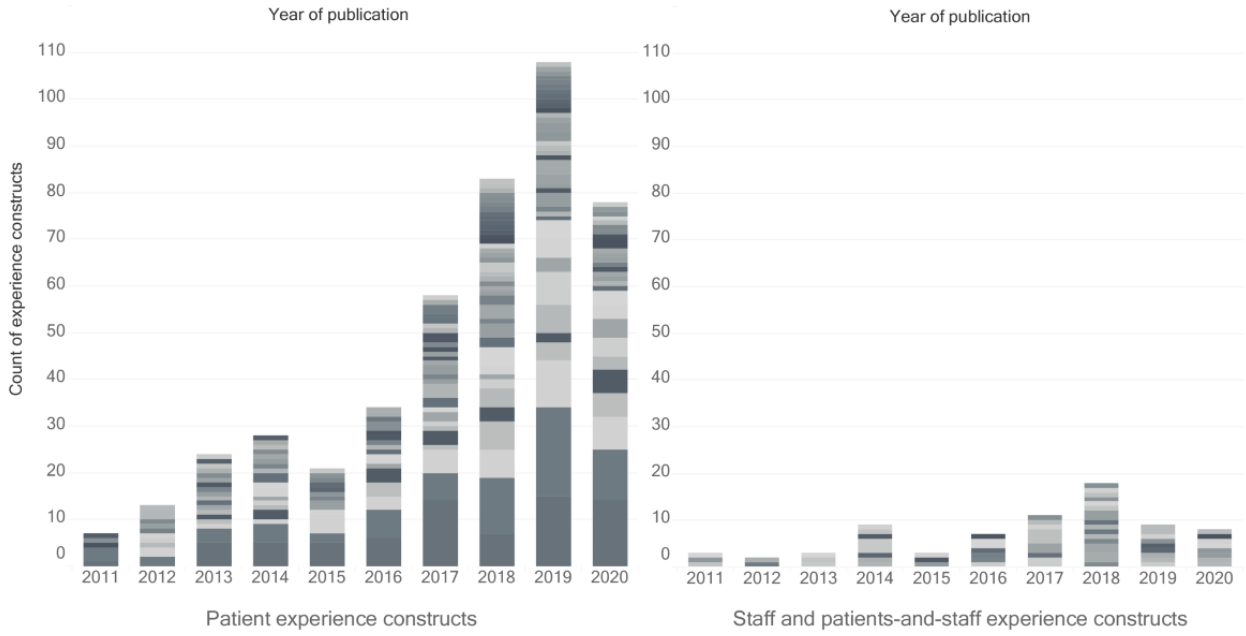

Correspondence Analysis

The plotted results of the CA of experience constructs are shown in Figure 5. Here, we discuss the outlook and interpretation of each dimension.

XSL•FO
RenderX

Figure 5. Graphical display of the results of the CA, visualized on the coordinates of dimensions 1 and 2. The labels (s), (p), and (s/p) refer to experience constructs used for staff, patients, and both. CA: correspondence analysis.

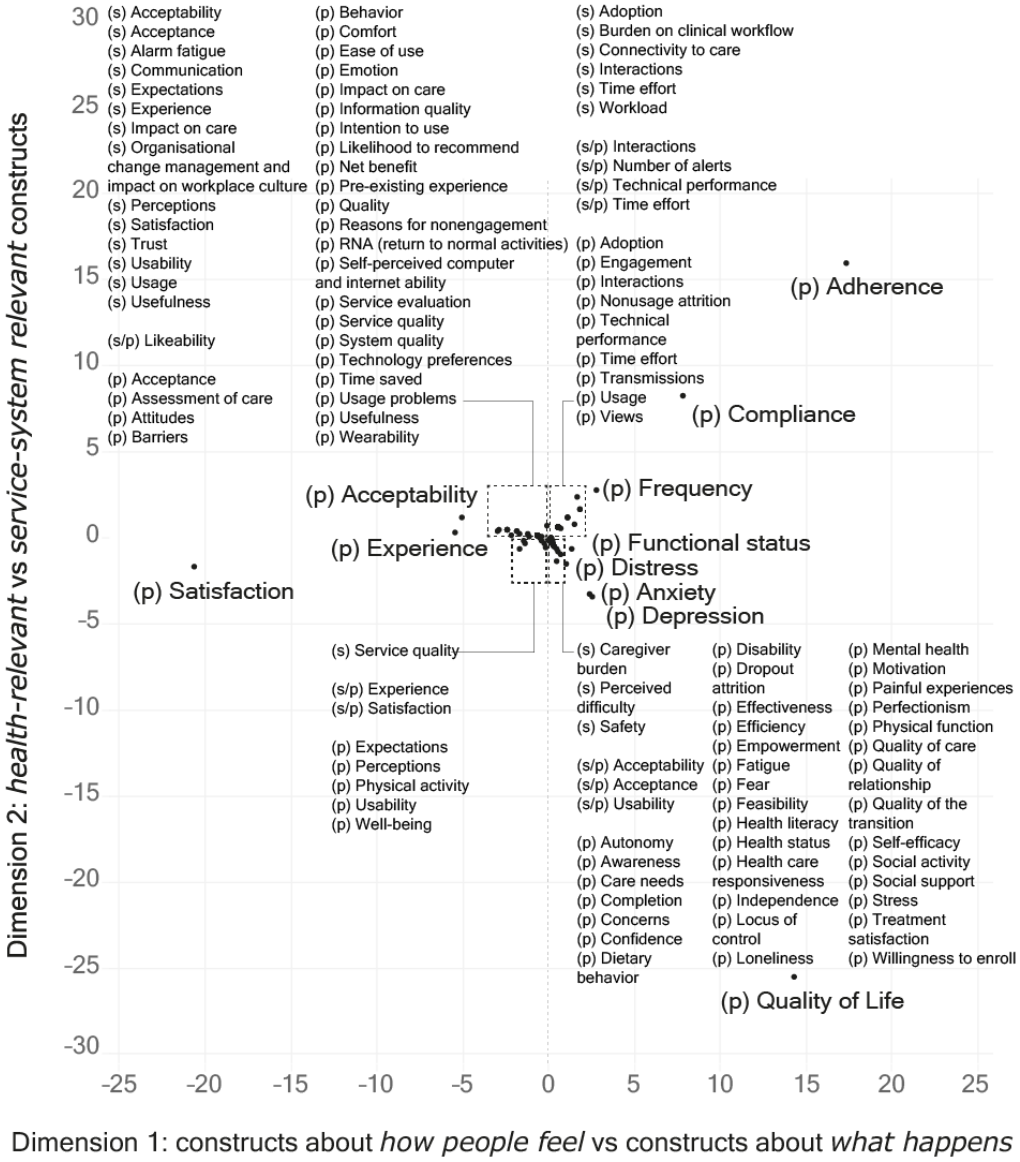

The first dimension explained more than 44% of the model's inertia. The contributions of this dimension showed which constructs had the most impact in determining its orientation: satisfaction (36%) and to a lesser extent adherence (26%) and quality of life (17%). On the negative (left) side of this dimension, we found constructs such as satisfaction, perceptions, and acceptability, which are associated with subjective measures of patient and staff experience and relate to how people feel or think in relation to RPM interventions. On the positive (right) side of this dimension, we found constructs such as adherence, compliance, and quality of life, which are associated with objectivized measures of patient and staff experience. By "objectivized measures," we referred to measures that are meant to capture phenomena in a factual manner, ideally independently from personal biases and subjective opinions. Adherence and compliance, particularly, are often measured through passive collection of system data (eg, log file analysis) that reflect objective measures of the way patients or staff interact with RPM propositions. Even in the case of (health-related) quality of life, which can include subjective connotations and components, measures usually aim at capturing an estimation of the factual impact of health status on a person's overall life quality.

In this sense, we attributed a distinction between *how people feel* versus *what happens* experience constructs to this first dimension. We noted that a similar distinction (between subjective vs objective measures of engagement in remote measurement studies) was previously proposed as a meaningful differentiation to structure "a field impeded by incoherent measures" [27].

The second dimension explained 35% of the model's inertia. The contributions of this dimension showed which constructs had the most impact in determining its orientation: quality of life (62%) and adherence (24%). On the negative (bottom) side of this dimension, we found constructs such as quality of life, depression, and anxiety, which are often used as experiential descriptors of health outcomes. On the positive (top) side of this dimension, we found adherence, compliance, and frequency, which are often used as descriptions of the interactions of

XSL•FO
RenderX

patients or staff with a specific (RPM) system. Thus, we attributed a distinction between *health-relevant* versus *system-relevant* experience constructs to this second dimension.

Based on the results of CA, we proposed a categorization of patient and staff experience–related constructs into 4 partly overlapping clusters. Coherent with the offered explanation of the 2 dimensions and in consideration of the constructs found in each area, we labeled these as service system–related experience measures, care-related experience measures, usage- and adherence-related experience measures, and health outcome–related experience measures. In Figure 6, we display the results of the CA labeled through this categorization. In this second visualization, we presented the results on a logarithmic scale to improve the visibility of constructs close to the center of the axes. Overall, this categorization of patient and staff experience constructs used in the RPM literature paints a landscape of the contemporary research in this field, which shows a mix of influences from clinical disciplines, health psychology, human factors engineering, service design, user research, systems engineering, and computer science.

Figure 6. Clustered results of the CA (displayed on a logarithmic scale). CA: correspondence analysis.

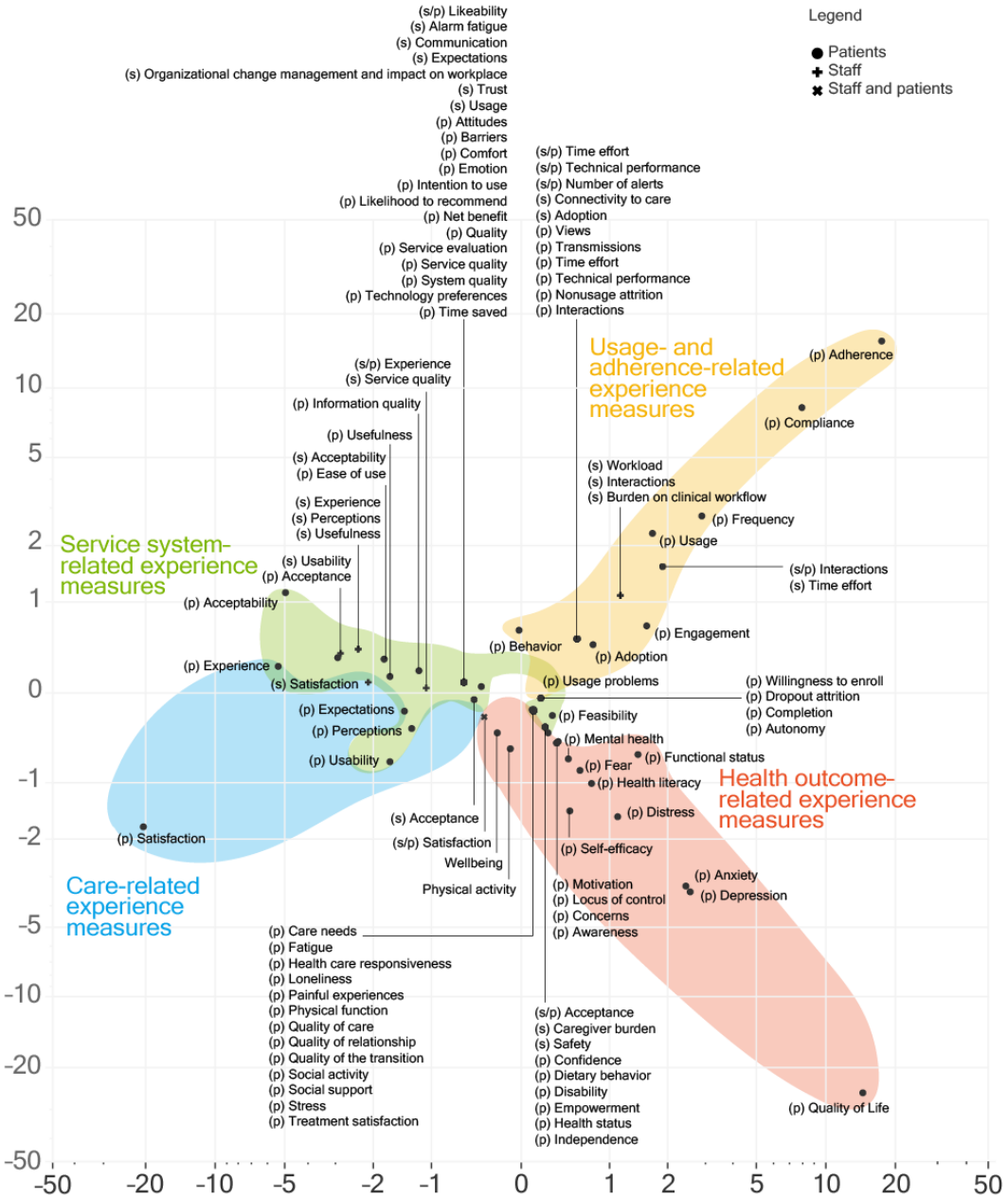

A visualization of the reported patient experience constructs and some of the related measuring instruments, organized by the categories identified in the CA, is available in Figure 7. A complete version of this visual can be found in Multimedia Appendix 4, and an interactive version can be found in [34]. In this figure, we can note the limited crossovers between constructs belonging to different categories, with the exception of versatile instruments, such as custom survey and log file analysis.

XSL•FO
RenderX

Figure 7. Reported patient experience constructs (left) and some associated measuring instruments (right). The thickness of each line refers to the number of instances each construct was used in the included studies. A complete version of this visual can be found in Multimedia Appendix 4, and an interactive version can be found in Ref [34].

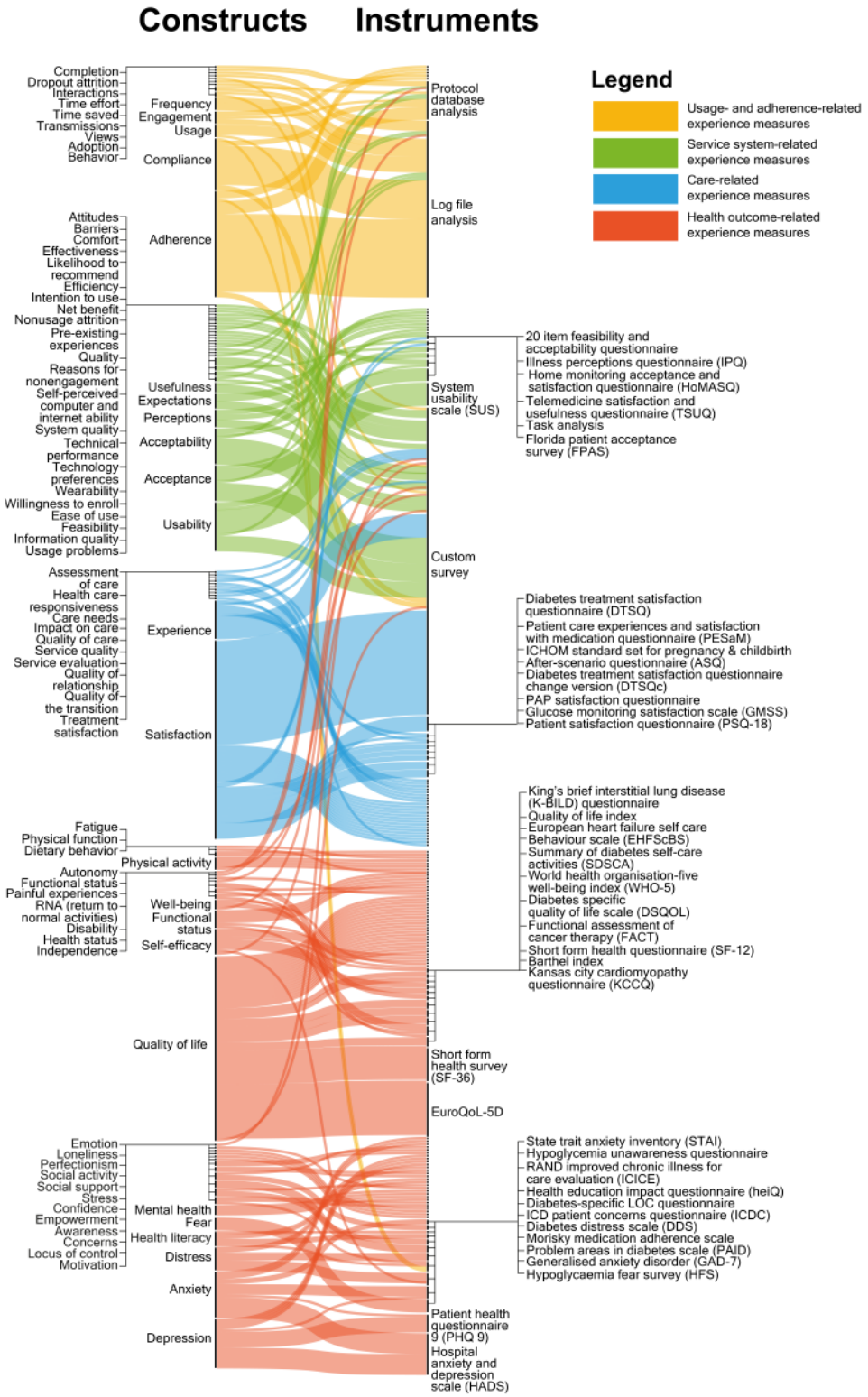

Recommendations

In the light of the collected findings, here we provide a set of recommendations to RPM patient and staff experience evaluators, in terms of both what to measure and how to measure it (Figure 8). Although these recommendations are functional to strengthen the quality of individual research protocols, they are also meant to stimulate increased standardization in the field as a whole.

XSL•FO
RenderX

Figure 8. Recommendations for patient and staff experience measuring in the RPM domain. RPM: remote patient monitoring.

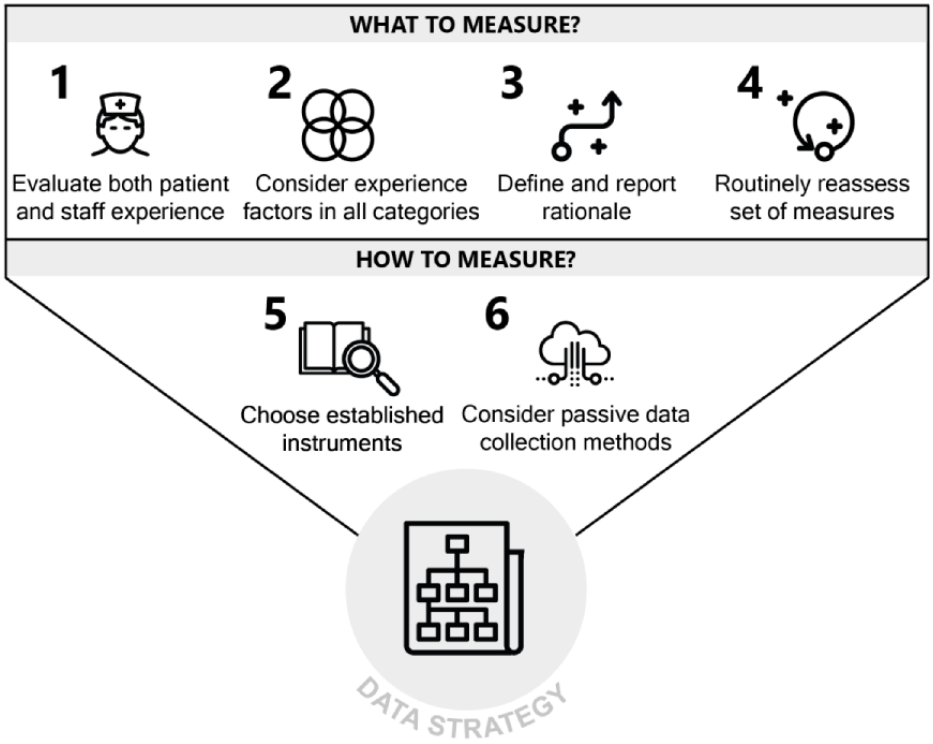

Regarding what to measure, we provide 4 main recommendations. The first is to conduct structured evaluations of staff experience next to patient experience. Failing to evaluate staff experience leads to risks, such as undetected staff nonadherence, misuse, and overworking. Although new competencies need to be developed in order for staff to unlock the untapped potential of RPM [35], seamless integration with existing clinical workflows should always be pursued and monitored.

The second recommendation is to consider experience constructs in all 4 clusters indicated in Figure 6, as these represent complementary facets of an overall experiential ensemble. Failing to do so exposes RPM evaluators to the risk of obtaining partial information (eg, only shedding light on *how people feel* but not on *what happens* in terms of patient and staff experience in RPM).

The third recommendation is to explicitly define and report a clear rationale regarding which aspects of patient and staff experience to prioritize in evaluations, depending on the goals and specificities of the RPM intervention. This rationale should ideally be informed by preliminary qualitative research and by a collaborative mapping of the expected relationships between patient and staff experience and other components of the Quadruple Aim framework for the RPM intervention at hand. Failing to follow this recommendation exposes RPM evaluators to the risk of obtaining results that are logically detached from each other and as such cannot inform organic improvement efforts. Virtuous examples of reporting a clear rationale were provided by Alonso-Solís et al [36] and den Bakker et al [37], who offered detailed accounts of the considerations used to guide the selection of included experience measures. Several existing frameworks and methods can be used to map such considerations, including the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [38] and the logical framework [39]. A relatively lightweight method to achieve such an overview can also be represented by the use of Figure 1 as a checklist to inventory possible Quadruple Aim relationships for a specific RPM intervention.

The fourth recommendation is to routinely reassess the chosen set of experience measures after each iteration of the RPM intervention design. Initial assumptions regarding relationships between experience factors and other dimensions of intervention quality should be verified once the relevant data are available, and new ones should be formulated, if necessary. If the RPM intervention transitions from research stages to implementation as the standard of care, it is recommended to keep on collecting at least some basic experience measures for system quality monitoring and continuous improvement. Failing to update the set of collected measures as the RPM intervention progresses through successive development stages exposes RPM evaluators to the risk of collecting outdated information, hindering iterative improvement processes.

Regarding how to measure RPM patient and staff experience, we provide 2 main recommendations. The first is to work with existing, validated and widely used instruments as much as possible, only creating new instruments after a convincing critique against current ones. Figure 7 can be used to find existing instruments measuring a broad range of experience-relevant constructs so as to reduce the need to create new ones.

For instance, researchers interested in evaluating certain experience constructs, ideally informed by preliminary qualitative research, might consult the full version of Figure 7 (available in Multimedia Appendix 4 or as an interactive map

XSL•FO
RenderX

in Ref. [34]) to find their construct of interest on the left side of the graph, follow the connecting lines to the existing relevant measures on the right, and identify the most frequently used ones. They can also use the visual to consider other possibly relevant constructs.

Alternatively, researchers can use the open access database of this review [40] and especially the "extracted data" Excel file to search for the construct of interest and find details of papers in the RPM domain in which the construct was previously measured.

Failing to follow this recommendation exposes RPM researchers to the risk of obtaining results that cannot be compared to meaningful benchmarks, compared to other RPM interventions, or be included in meta-analyses.

The second recommendation is to consider adopting automatic, "passive" methods of experience data collection, such as the ones we referred to in this review as log file analysis, so as to obtain actionable estimates of user behavior with a reduced need for patients and staff to fill tedious surveys [41] or otherwise provide active input. Failing to consider automatically collected log file data on patient and staff experience constitutes a missed opportunity in terms of both the quality and cost of evaluation data. We recognize such nascent data innovations as promising [42] but also in need of methodological definition, particularly in terms of an ethical evaluation of data privacy and access [43,44] in order to avoid exploitative forms of prosumption [45].

Discussion

Principal Findings

This study resulted in a structured overview of patient and staff experience measures used in contemporary RPM research. Through this effort, we found that the research landscape has seen a sizeable growth in the past 10 years, that it is affected by a relative lack of focus on staff experience, and that the overall corpus of collected measures can be organized in 4 main categories (service system–related, care-related, usage- and adherence-related, and health outcome–related experience measures). Little to no consensus or standardization was found in the adopted methods. Based on these findings, a set of 6 actionable recommendations for RPM experience evaluators was provided, with the aim of improving the quality and standardization of experience-related RPM research. The results of this review align with and expand on recent contributions in the field, with particular regard to the work of White et al [27].

Directions for Further Research

Fruitful future research opportunities have been recognized in various areas of RPM experience measuring. Among them, we stress the need for comparative studies investigating patient and staff experience factors across different RPM interventions; for studies clarifying the use, potential, and limitations of log file analysis in this domain; and (most importantly) for studies examining the complex relationships between experience factors, health outcomes, and cost-effectiveness in RPM.

Ultimately, we recognize the need for integrated data strategies for RPM, intended as processes and rules that define how to manage, analyze, and act upon RPM data, including continuously collected experience data, as well as clinical, technical, and administrative data. Data strategies can represent a way to operationalize a systems approach to health care innovation, described by Komashie et al [46] as "a way of addressing health delivery challenges that recognizes the multiplicity of elements interacting to impact an outcome of interest and implements processes or tools in a holistic way." As complex, adaptive, and partly automated systems, RPM interventions require sophisticated data strategies in order to function and improve [47]; continuous loops of system feedback need to be established and analyzed in order to monitor the impact of RPM systems and optimize their performance over time, while respecting patients' and staff's privacy. This is especially true in the case of RPM systems including artificial intelligence (AI) components, which require continuous monitoring and updating of algorithms [48-50]. We characterize the development of integrated, interdisciplinary data strategies as a paramount challenge in contemporary RPM research, which will require closer collaboration between digital health designers and health care professionals [51-53]. We hope to have provided a small contribution to this overall goal through our effort to structure the current landscape of RPM patient and staff experience evaluation.

Strengths and Limitations

We acknowledge both strengths and limitations of the chosen methodologies. The main strength of this review is its extensive focus, covering a large number of experience measures and RPM interventions. However, a limitation introduced by such a broad scope is the lack of differentiation by targeted condition, clinical specialty, RPM intervention characteristics, geographical area, or other relevant distinctions. Furthermore, limitations were introduced by choices, such as focusing exclusively on contributions in English and on nonprimary care and nonpediatric RPM interventions.

Conclusion

Contemporary patient and staff experience measuring in RPM is affected by a lack of consensus and standardization, affecting the quality of both primary and secondary research in this domain. This issue determines a critical knowledge gap in our understanding of the effectiveness of RPM interventions, which are known to bring about radical changes to the care experience of both patients and staff. Bridging this knowledge gap appears to be critical in a global context of urgent need for increased resource effectiveness across health care systems, including through the increased adoption of safe and effective RPM. In this context, this review offers support for RPM experience evaluators by providing a structured overview of contemporary patient and staff experience measures and a set of practical guidelines for improving research quality and standardization in this domain.

XSL•FO
RenderX

Acknowledgments

We gratefully acknowledge Jeroen Raijmakers, Francesca Marino, Lorena Hurtado Alvarez, Alexis Derumigny, and Laurens Schuurkamp for the help and advice provided in the context of this research.

Neither ChatGPT nor other generative language models were used in this research or in the manuscript preparation or review.

Data Availability

The data sets generated and analyzed during this review are available as open access in Ref. [40].

Authors' Contributions

VP conceived the study, performed the systematic review and data analysis, and was mainly responsible for the writing of the manuscript. HCMO collaborated on study design, performed independent screening of contributions, and collaborated on data analysis. RvK provided input to the study design and execution. PG supported query term selection and structuring. MK provided input on manuscript framing and positioning. DS provided input on the design, execution, and reporting of the correspondence analysis. All authors revised and made substantial contributions to the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 853 KB - jmir_v26i1e48463_app1.docx]

Multimedia Appendix 2
Full search strategies.
[DOCX File , 16 KB - jmir_v26i1e48463_app2.docx]

Multimedia Appendix 3
Overview of the merged construct formulations .
[DOCX File , 15 KB - jmir_v26i1e48463_app3.docx]

Multimedia Appendix 4
Reported patient experience constructs and associated measuring instruments (complete visual).
[PNG File , 2648 KB - jmir_v26i1e48463_app4.png]

References

1. da Farias FAC, Dagostini CM, Bicca YDA, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020 May 17;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
2. Taiwo O, Ezugwu AE. Smart healthcare support for remote patient monitoring during COVID-19 quarantine. Inform Med Unlocked 2020;20:100428 [FREE Full text] [doi: 10.1016/j.imu.2020.100428] [Medline: 32953970]
3. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284 [FREE Full text] [doi: 10.2196/19284] [Medline: 32501804]
4. Peek N, Sujan M, Scott P. Digital health and care in pandemic times: impact of COVID-19. BMJ Health Care Inform 2020 Jun 21;27(1):e100166 [FREE Full text] [doi: 10.1136/bmjhci-2020-100166] [Medline: 32565418]
5. Sust PP, Solans O, Fajardo JC, Peralta MM, Rodenas P, Gabaldà J, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 04;6(2):e19106 [FREE Full text] [doi: 10.2196/19106] [Medline: 32339998]
6. Vegesna A, Tran M, Angelaccio M, Arcona S. Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed J E Health 2017 Jan;23(1):3-17 [FREE Full text] [doi: 10.1089/tmj.2016.0051] [Medline: 27116181]
7. Noah B, Keller MS, Mosadeghi S, Stein L, Johl S, Delshad S, et al. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. NPJ Digit Med 2018 Jan 15;1(1):20172 [FREE Full text] [doi: 10.1038/s41746-017-0002-4] [Medline: 31304346]
8. Majumder S, Mondal T, Deen M. Wearable sensors for remote health monitoring. Sensors (Basel) 2017 Jan 12;17(1):130 [FREE Full text] [doi: 10.3390/s17010130] [Medline: 28085085]

XSL•FO
RenderX

9. Coye MJ, Haselkorn A, DeMello S. Remote patient management: technology-enabled innovation and evolving business models for chronic disease care. Health Aff (Millwood) 2009 Jan;28(1):126-135. [doi: 10.1377/hlthaff.28.1.126] [Medline: 19124862]
10. Schütz N, Knobel SEJ, Botros A, Single M, Pais B, Santschi V, et al. A systems approach towards remote health-monitoring in older adults: introducing a zero-interaction digital exhaust. NPJ Digit Med 2022 Aug 16;5(1):116 [FREE Full text] [doi: 10.1038/s41746-022-00657-y] [Medline: 35974156]
11. Drennan VM, Ross F. Global nurse shortages—the facts, the impact and action for change. Br Med Bull 2019 Jun 19;130(1):25-37. [doi: 10.1093/bmb/ldz014] [Medline: 31086957]
12. Global Burden of Disease Health Financing Collaborator Network. Past, present, and future of global health financing: a review of development assistance, government, out-of-pocket, and other private spending on health for 195 countries, 1995-2050. Lancet 2019 Jun 01;393(10187):2233-2260 [FREE Full text] [doi: 10.1016/S0140-6736(19)30841-4] [Medline: 31030984]
13. Mecklai K, Smith N, Stern AD, Kramer DB. Remote patient monitoring — overdue or overused? N Engl J Med 2021 Apr 15;384(15):1384-1386. [doi: 10.1056/nejmp2033275]
14. León MA, Pannunzio V, Kleinsmann M. The impact of perioperative remote patient monitoring on clinical staff workflows: scoping review. JMIR Hum Factors 2022 Jun 06;9(2):e37204 [FREE Full text] [doi: 10.2196/37204] [Medline: 35666574]
15. Mair F, Whitten P. Systematic review of studies of patient satisfaction with telemedicine. BMJ 2000 Jun 03;320(7248):1517-1520 [FREE Full text] [doi: 10.1136/bmj.320.7248.1517] [Medline: 10834899]
16. Kraai I, Luttik M, de Jong R, Jaarsma T, Hillege H. Heart failure patients monitored with telemedicine: patient satisfaction, a review of the literature. J Card Fail 2011 Aug;17(8):684-690. [doi: 10.1016/j.cardfail.2011.03.009] [Medline: 21807331]
17. Wolf JA, Niederhauser V, Marshburn D, LaVela SL. Reexamining “defining patient experience”: the human experience in healthcare. Patient Exp J 2021 Apr 28;8(1):16-29. [doi: 10.35680/2372-0247.1594]
18. Lavela S, Gallan A. Evaluation and measurement of patient experience. Patient Exp J 2014 Apr 1;1(1):28-36. [doi: 10.35680/2372-0247.1003]
19. Wang T, Giunti G, Melles M, Goossens R. Digital patient experience: umbrella systematic review. J Med Internet Res 2022 Aug 04;24(8):e37952 [FREE Full text] [doi: 10.2196/37952] [Medline: 35925651]
20. Zakkar M. Patient experience: determinants and manifestations. IJHG 2019 May 22;24(2):143-154. [doi: 10.1108/ijhg-09-2018-0046]
21. Sikka R, Morath JM, Leape L. The quadruple aim: care, health, cost and meaning in work. BMJ Qual Saf 2015 Oct 02;24(10):608-610. [doi: 10.1136/bmjqs-2015-004160] [Medline: 26038586]
22. Pannunzio V, Kleinsmann M, Snelders H. Design research, eHealth, and the convergence revolution. arXiv:1909.08398v1 [cs.HC] preprint posted online 2019. [doi: 10.48550/arXiv.1909.08398] . [doi: 10.48550/arXiv.1909.08398]
23. Thomas EE, Taylor ML, Banbury A, Snoswell CL, Haydon HM, Gallegos Rejas VM, et al. Factors influencing the effectiveness of remote patient monitoring interventions: a realist review. BMJ Open 2021 Aug 25;11(8):e051844 [FREE Full text] [doi: 10.1136/bmjopen-2021-051844] [Medline: 34433611]
24. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. AACN Adv Crit Care 2013;24(4):378-386. [doi: 10.1097/nci.0b013e3182a903f9]
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
27. White KM, Williamson C, Bergou N, Oetzmann C, de Angel V, Matcham F, et al. A systematic review of engagement reporting in remote measurement studies for health symptom tracking. NPJ Digit Med 2022 Jun 29;5(1):82 [FREE Full text] [doi: 10.1038/s41746-022-00624-7] [Medline: 35768544]
28. Dew D. Construct. In: Lavrakas PJ, editor. Encyclopedia of Survey Research Methods. Thousand Oaks, CA: SAGE Publications; 2008:134.
29. Greenacre MJ. Correspondence Analysis in the Social Sciences: Recent Developments and Applications. San Diego, CA: Academic Press; 1999.
30. Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunananthan S, et al. Erratum to “Correspondence analysis is a useful tool to uncover the relationships among categorical variables” [J Clin Epidemiol 2010;63:638-646]. J Clin Epidemiol 2010 Jul;63(7):809. [doi: 10.1016/j.jclinepi.2010.04.003]
31. Franceschi VB, Santos AS, Glaeser AB, Paiz JC, Caldana GD, Machado Lessa CL, et al. Population-based prevalence surveys during the COVID-19 pandemic: a systematic review. Rev Med Virol 2021 Jul 04;31(4):e2200 [FREE Full text] [doi: 10.1002/rmv.2200] [Medline: 34260777]
32. Huerta T, Fareed N, Hefner JL, Sieck CJ, Swoboda C, Taylor R, et al. Patient engagement as measured by inpatient portal use: methodology for log file analysis. J Med Internet Res 2019 Mar 25;21(3):e10957 [FREE Full text] [doi: 10.2196/10957] [Medline: 30907733]

XSL•FO
RenderX

33. Diaper D, Stanton N. The Handbook of Task Analysis for Human-Computer Interaction. Boca Raton, FL: CRC Press; 2003.
34. Interactive Sankey. Adobe. URL: https://indd.adobe.com/view/d66b2b4c-463c-4b39-8934-ac0282472224 [accessed 2024-03-25]
35. Hilty DM, Armstrong CM, Edwards-Stewart A, Gentry MT, Luxton DD, Krupinski EA. Sensor, wearable, and remote patient monitoring competencies for clinical care and training: scoping review. J Technol Behav Sci 2021;6(2):252-277 [FREE Full text] [doi: 10.1007/s41347-020-00190-3] [Medline: 33501372]
36. Alonso-Solís A, Rubinstein K, Corripio I, Jaaskelainen E, Seppälä A, Vella VA, m-Resist Group, et al. Mobile therapeutic attention for treatment-resistant schizophrenia (m-RESIST): a prospective multicentre feasibility study protocol in patients and their caregivers. BMJ Open 2018 Jul 16;8(7):e021346 [FREE Full text] [doi: 10.1136/bmjopen-2017-021346] [Medline: 30012788]
37. den Bakker CM, Schaafsma FG, van der Meij E, Meijerink WJ, van den Heuvel B, Baan AH, et al. Electronic health program to empower patients in returning to normal activities after general surgical and gynecological procedures: intervention mapping as a useful method for further development. J Med Internet Res 2019 Feb 06;21(2):e9938 [FREE Full text] [doi: 10.2196/jmir.9938] [Medline: 30724740]
38. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
39. Dey P, Hariharan S, Brookes N. Managing healthcare quality using logical framework analysis. Manag Serv Qual 2006 Mar 1;16(2):203-222. [doi: 10.1108/09604520610650655]
40. Pannunzio V, Ornelas HM. Data of article "Patient and staff experience evaluation in remote patient monitoring; what to measure and how? A systematic review". Version 1. Dataset. 4TU.ResearchData. URL: https://data.4tu.nl/articles/_/21930783/1 [accessed 2024-03-25]
41. de Koning R, Egiz A, Kotecha J, Ciuculete AC, Ooi SZY, Bankole NDA, et al. Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. Front Surg 2021 Aug 12;8:690680 [FREE Full text] [doi: 10.3389/fsurg.2021.690680] [Medline: 34458314]
42. Miriovsky BJ, Shulman LN, Abernethy AP. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. J Clin Oncol 2012 Dec 01;30(34):4243-4248. [doi: 10.1200/JCO.2012.42.8011] [Medline: 23071233]
43. Fernández-Alemán JL, Señor IC, Lozoya P, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013 Jun;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
44. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Privacy and security in mobile health apps: a review and recommendations. J Med Syst 2015 Jan;39(1):181. [doi: 10.1007/s10916-014-0181-3] [Medline: 25486895]
45. Lupton D. The commodification of patient opinion: the digital patient experience economy in the age of big data. Sociol Health Illn 2014 Jul 01;36(6):856-869. [doi: 10.1111/1467-9566.12109] [Medline: 24443847]
46. Komashie A, Ward J, Bashford T, Dickerson T, Kaya GK, Liu Y, et al. Systems approach to health service design, delivery and improvement: a systematic review and meta-analysis. BMJ Open 2021 Jan 19;11(1):e037667 [FREE Full text] [doi: 10.1136/bmjopen-2020-037667] [Medline: 33468455]
47. Abdolkhani R, Gray K, Borda A, DeSouza R. Patient-generated health data management and quality challenges in remote patient monitoring. JAMIA Open 2019 Dec;2(4):471-478 [FREE Full text] [doi: 10.1093/jamiaopen/ooz036] [Medline: 32025644]
48. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022 May 31;5(1):66 [FREE Full text] [doi: 10.1038/s41746-022-00611-y] [Medline: 35641814]
49. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. NPJ Digit Med 2020 Apr 07;3(1):53 [FREE Full text] [doi: 10.1038/s41746-020-0262-2] [Medline: 32285013]
50. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
51. Pannunzio V. Towards a convergent approach to the use of data in digital health design. Dissertation, Delft University of Technology. 2023. URL: https://tinyurl.com/4ah5tvw6 [accessed 2024-03-25]
52. Morales Ornelas HC, Kleinsmann M, Kortuem G. Exploring health and design evidence practices in eHealth systems' development. 2023 Presented at: ICED23: International Conference on Engineering Design; July 2023; Bordeaux, France p. 1795-1804. [doi: 10.1017/pds.2023.180]
53. Morales OH, Kleinsmann M, Kortuem G. Towards designing for health outcomes: implications for designers in eHealth design. In: Forthcoming. 2024 Presented at: DESIGN2024: International Design Conference; May 2024; Cavtat, Croatia.

XSL•FO
RenderX

Abbreviations

CA: correspondence analysis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses RPM: remote patient monitoring
SUS: System Usability Scale

Edited by T de Azevedo Cardoso; submitted 25.04.23; peer-reviewed by M Tai-Seale, C Nöthiger, M Gasmi ; comments to author 29.07.23; revised version received 25.08.23; accepted 20.02.24; published 22.04.24.

Please cite as:
Pannunzio V, Morales Ornelas HC, Gurung P, van Kooten R, Snelders D, van Os H, Wouters M, Tollenaar R, Atsma D, Kleinsmann M
Patient and Staff Experience of Remote Patient Monitoring—What to Measure and How: Systematic Review
J Med Internet Res 2024;26:e48463
URL: https://www.jmir.org/2024/1/e48463
doi:10.2196/48463
PMID:38648090

©Valeria Pannunzio, Hosana Cristina Morales Ornelas, Pema Gurung, Robert van Kooten, Dirk Snelders, Hendrikus van Os, Michel Wouters, Rob Tollenaar, Douwe Atsma, Maaike Kleinsmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Electronic Media Use and Sleep Quality: Updated Systematic Review and Meta-Analysis

Xiaoning Han[1*], PhD; Enze Zhou[1*], MA; Dong Liu[1*], PhD

School of Journalism and Communication, Renmin University of China, Beijing, China

[*]all authors contributed equally

Corresponding Author:
Dong Liu, PhD
School of Journalism and Communication
Renmin University of China
No. 59 Zhongguancun Street, Haidian District
Beijing, 100872
China
Phone: 86 13693388506
Email: bnuliudong@gmail.com

Abstract

Background: This paper explores the widely discussed relationship between electronic media use and sleep quality, indicating negative effects due to various factors. However, existing meta-analyses on the topic have some limitations.

Objective: The study aims to analyze and compare the impacts of different digital media types, such as smartphones, online games, and social media, on sleep quality.

Methods: Adhering to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, the study performed a systematic meta-analysis of literature across multiple databases, including Web of Science, MEDLINE, PsycINFO, PubMed, Science Direct, Scopus, and Google Scholar, from January 2018 to October 2023. Two trained coders coded the study characteristics independently. The effect sizes were calculated using the correlation coefficient as a standardized measure of the relationship between electronic media use and sleep quality across studies. The Comprehensive Meta-Analysis software (version 3.0) was used to perform the meta-analysis. Statistical methods such as funnel plots were used to assess the presence of asymmetry and a *p*-curve test to test the *p*-hacking problem, which can indicate publication bias.

Results: Following a thorough screening process, the study involved 55 papers (56 items) with 41,716 participants from over 20 countries, classifying electronic media use into “general use” and “problematic use.” The meta-analysis revealed that electronic media use was significantly linked with decreased sleep quality and increased sleep problems with varying effect sizes across subgroups. A significant cultural difference was also observed in these effects. General use was associated with a significant decrease in sleep quality ($P<.001$). The pooled effect size was 0.28 (95% CI 0.21-0.35; $k=20$). Problematic use was associated with a significant increase in sleep problems ($P\leq.001$). The pooled effect size was 0.33 (95% CI 0.28-0.38; $k=36$). The subgroup analysis indicated that the effect of general smartphone use and sleep problems was $r=0.33$ (95% CI 0.27-0.40), which was the highest among the general group. The effect of problematic internet use and sleep problems was $r=0.51$ (95% CI 0.43-0.59), which was the highest among the problematic groups. There were significant differences among these subgroups (general: $Q_{between}=14.46$, $P=.001$; problematic: $Q_{between}=27.37$, $P<.001$). The results of the meta-regression analysis using age, gender, and culture as moderators indicated that only cultural difference in the relationship between Eastern and Western culture was significant ($Q_{between}=6.69$; $P=.01$). All funnel plots and *p*-curve analyses showed no evidence of publication and selection bias.

Conclusions: Despite some variability, the study overall confirms the correlation between increased electronic media use and poorer sleep outcomes, which is notably more significant in Eastern cultures.

(J Med Internet Res 2024;26:e48356) doi:10.2196/48356

KEYWORDS

electronic media; sleep quality; meta-analysis; media types; cultural difference

XSL•FO
RenderX

Introduction

Sleep is vital to our health. Research has shown that high sleep quality can lead to improvements in a series of health outcomes, such as an improved immune system, better mood and mental health, enhanced physical performance, lower risk of chronic diseases, and a longer life span [1-5].

Electronic media refers to forms of media or communication that use electronic devices or technology to create, distribute, and display content. This can include various forms of digital media such as smartphones, tablets, instant messaging, phone calls, social media, online games, short video platforms, etc. Electronic media has permeated every aspect of our lives [6]. Many prefer to use smartphones or tablets before sleep, which can negatively affect sleep in many aspects, including delayed sleep onset, disrupted sleep patterns, shortened sleep duration, and poor sleep quality [7-10]. Furthermore, problematic use occurs when the behavior surpasses a certain limit. In this study, problematic use of electronic media is not solely determined by the amount of time spent on these platforms, but rather by behavioral indicators that suggest an unhealthy or harmful relationship with them.

Smartphones or tablet use can affect sleep quality in many ways. At first, the use of these devices may directly displace, delay, or interrupt sleep time, resulting in inadequate sleep quantity [11]. The sound of notifications and vibrations of these devices may interrupt sleep. Second, the screens of smartphones and tablets emit blue light, which can suppress the production of melatonin, the hormone responsible for regulating sleep-wake cycles [12]. Third, consuming emotionally charged content, such as news, suspenseful movies, or engaging in online arguments, can increase emotional arousal, making it harder to relax and fall asleep. This emotional arousal can also lead to disrupted sleep and nightmares [13]. Finally, the use of electronic devices before bedtime can lead to a delay in bedtime and a shortened sleep duration, as individuals may lose track of time while engaging with their devices. This can result in a disrupted sleep routine and decreased sleep quality [14].

Some studies have conducted meta-analyses on screen media use and sleep outcomes in 2016, 2019, and 2021 [15-17]. However, these studies had their own limitations. First, the sample size included in their meta-analyses was small (around 10). Second, these studies only focused on 1 aspect of the effect of digital media on sleep quality. For example, Carter et al [16] focused only on adolescents, and both Alimoradi et al [15] and Kristensen et al [17] only reviewed the relationship between problematic use of digital media or devices and sleep quality. Despite of the high heterogeneity found in the meta-analyses, none have compared the effects of different digital media or devices. This study aims to clarify and compare the effects of these different channels.

Methods

Literature Search

The research adhered to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Multimedia Appendix 1) and followed a predetermined protocol [18,19]. As the idea and scope of this study evolved over time, the meta-analysis was not preregistered. However, the methodology was defined a priori and strictly followed to reduce biases, and the possible influence of post hoc decisions was minimized. All relevant studies in English, published from January 1, 2018, to October 9, 2023, were searched. We searched the following databases: Web of Science, MEDLINE, PsycINFO, PubMed, Science Direct, Scopus, and Google Scholar. The abstracts were examined manually. The keywords used to search were the combination of the following words: “sleep” OR “sleep duration” OR “sleep quality” OR “sleep problems” AND “electronic media” OR “smartphone” OR “tablet” OR “social media” OR “Facebook” OR “Twitter” OR “online gaming” OR “internet” OR “addiction” OR “problematic” (Multimedia Appendix 2). Additionally, the reference lists of relevant studies were examined.

Two trained coders independently screened the titles and abstracts of the identified papers for eligibility, followed by a full-text review of the selected studies. Discrepancies between the coders were resolved through discussion until a consensus was reached. The reference lists of the included studies were also manually screened to identify any additional relevant studies. Through this rigorous process, we ensured a comprehensive and replicable literature search that could contribute to the robustness of our meta-analysis findings.

Inclusion or Exclusion Criteria

Titles and abstracts from search results were scrutinized for relevance, with duplicates removed. Full texts of pertinent papers were obtained, and their eligibility for inclusion was evaluated. We mainly included correlational studies that used both continuous measures of time spent using electronic media use and sleep quality. Studies must have been available in English. Four criteria were used to screen studies: (1) only peer-reviewed empirical studies, published in English, were considered for inclusion in the meta-analysis; (2) the studies should report quantitative statistics on electronic media use and sleep quality, including sample size and essential information to calculate the effect size, and review papers, qualitative studies, case studies, and conference abstracts were excluded; (3) studies on both general use and problematic use of electronic media or devices should be included; and (4) only studies that used correlation, regression, or odds ratio were included to ensure consistency.

Study Coding

Two trained coders were used to code the characteristics of the studies independently. Discrepancies were discussed with the first author of the paper to resolve. Sample size and characteristics of participants were coded: country, female ratio, average age, publication year, and electronic types. Effect sizes were either extracted directly from the original publications or manually calculated. If a study reported multiple dependent effects, the effects were merged into one. If a study reported multiple independent effects from different samples, the effects were included separately. Additionally, to evaluate the study quality, the papers were classified into 3 tiers (high, middle, and low) according to *Journal Citation Reports 2022*, a ranking

XSL•FO
RenderX

of journals based on their impact factor as reported in the Web of Science. The few unindexed papers were rated based on their citation counts as reported in Google Scholar.

Meta-Analysis and Moderator Analyses

The effect sizes were calculated using the correlation coefficient (*r*) as a standardized measure of the relationship between electronic media or device use and sleep quality across studies. When studies reported multiple effect sizes, we selected the one that best represented the overall association between electronic media use and sleep quality. If studies did not provide correlation coefficients, we converted other reported statistics (eg, standardized regression coefficients) into correlation coefficients using established formulas. Once calculated, the correlation coefficients were transformed into Fisher *z* scores to stabilize the variance and normalize the distribution.

Previous meta-studies have shown high levels of heterogeneity. Hence, the random effects model was adopted for all analyses. To explore potential factors contributing to the heterogeneity and to further understand the relationship between electronic media use and sleep quality, we conducted moderator analyses. The following categorical and continuous moderators were examined: media types (online gaming, social media, smartphone, or intent), participants' average age, culture, female ratio, and sleep quality assessment method. For categorical moderators, subgroup analyses were performed, while for continuous moderators, meta-regression analyses were conducted. All analyses were completed in the Comprehensive Meta-Analysis software (version 3.0; Biostat, Inc).

Publication Bias

We used statistical methods such as funnel plots to assess the presence of asymmetry and a *p*-curve test to test the *p*-hacking problem, which may indicate publication bias. In case of detected asymmetry, we applied techniques such as the trim-and-fill method to adjust the effect size estimates.

By addressing publication bias, we aimed to provide a more accurate and reliable synthesis of the available evidence, enhancing the validity and generalizability of our meta-analytic findings. Nevertheless, it is essential for readers to interpret the results cautiously, considering the potential limitations imposed by publication bias and other methodological concerns.

Results

Search Findings

A total of 98,806 studies were identified from databases, especially Scopus (n=49,643), Google Scholar (n=18,600), Science Direct (n=15,084), and Web of Science (n=11,689). Upon removing duplicate records and excluding studies that did not meet the inclusion criteria, 754 studies remained for the screening phase. After screening titles, abstracts, and full texts, 703 studies were excluded. A total of 4 additional studies were identified from the references of relevant reviews. Finally, 55 studies [20-74] were included in the meta-analysis. The flow diagram of the selection is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

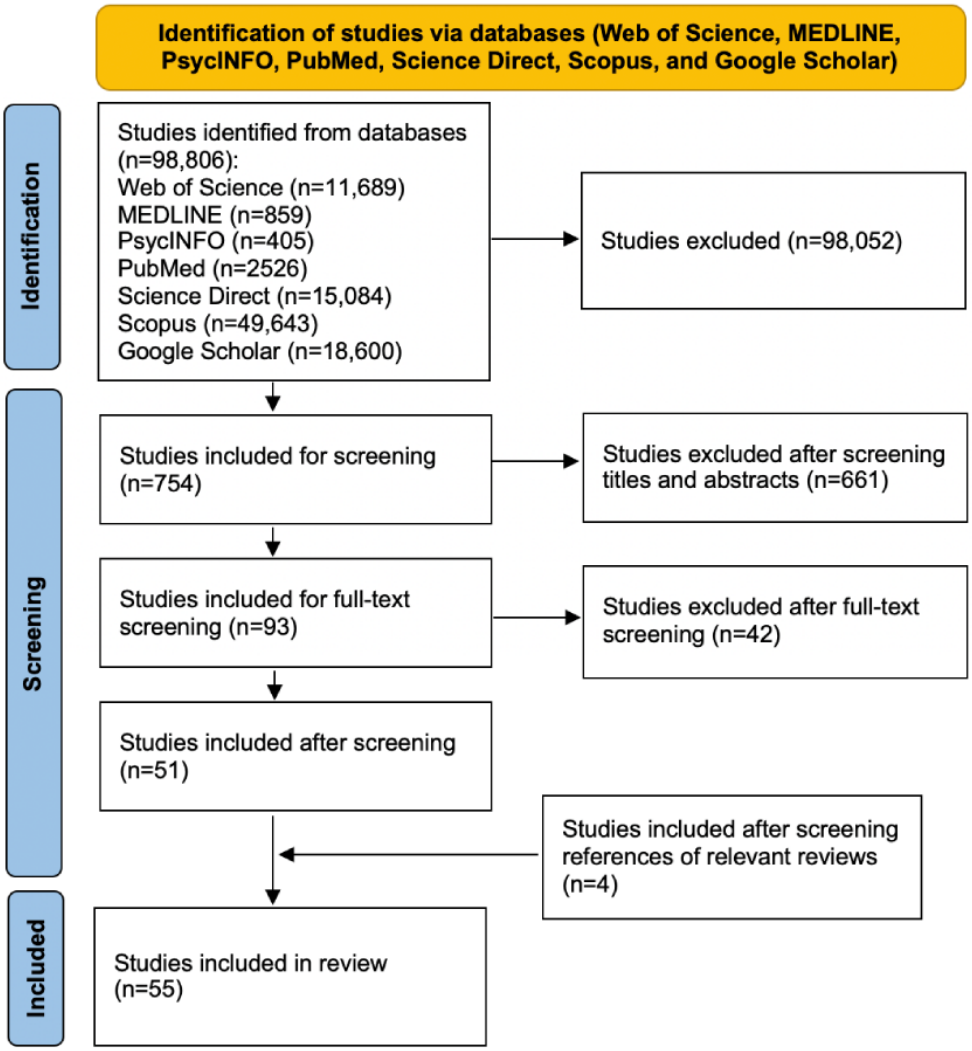

Characteristics of Included Studies

In 20 studies, 21,594 participants were included in the analysis of the general use of electronic media and sleep quality. The average age of the sample ranged from 9.9 to 44 years. The category of general online gaming and sleep quality included 4 studies, with 14,837 participants; the category of general smartphone use and sleep quality included 10 studies, with 5011 participants; and the category of general social media use and sleep quality included 6 studies, with 1746 participants.

These studies came from the following countries or areas: Germany, Serbia, Indonesia, India, China, Italy, Saudi Arabia, New Zealand, the United Kingdom, the United States, Spain, Qatar, Egypt, Argentina, and Portugal. The most frequently used measure of electronic media use was the time spent on it. The most frequently used measure of sleep was the Pittsburgh Sleep Quality Index.

In 35 studies, 20,122 participants were included in the analysis of the problematic use of electronic media and sleep quality. The average age of the sample ranged from 14.76 to 65.62 years. The category of problematic online gaming and sleep quality included 5 studies, with 1874 participants; the category of problematic internet use and sleep quality included 2 studies, with 774 participants; the category of problematic smartphone use and sleep quality included 18 studies, with 12,204 participants; and the category of problematic social media use and sleep quality included 11 studies, with 5270 participants. There was a study that focused on both social media and online gaming, which led to its inclusion in the analysis. These studies came from 14 countries or areas: Turkey, the United States, Indonesia, China, France, Taiwan, India, South Korea, Hong Kong, Iran, Poland, Israel, Hungary, and Saudi Arabia. The most frequently used measures of problematic electronic media use were the Internet Gaming Disorder Scale-Short Form, Smartphone Addiction Scale-Short Form, and Bergen Social Media Addiction Scale.

With respect to study quality, the 56 papers were published in 50 journals, 41 of which were indexed in *Journal Citation Reports 2022*, while the remaining 9 journals were rated based on their citation counts as reported in Google Scholar. As a result, of the 56 papers included in the study, 22 papers were assigned a high rating, 18 papers were assigned a middle rating, and 16 papers were assigned a low rating. More information about the included studies is listed in Multimedia Appendix 3 [20-74].

XSL•FO
RenderX

Meta-Analysis

The results of the meta-analysis of the relationship between general electronic media use and sleep quality showed that electronic media use was associated with a significant decrease in sleep quality (P<.001). The pooled effect size was 0.28 (95% CI 0.21-0.35; k=20), indicating that individuals who used electronic media more frequently were generally associated with more sleeping problems.

The second meta-analysis showed that problematic electronic media use was associated with a significant increase in sleep problems (P≤.001). The pooled effect size was 0.33 (95% CI 0.28-0.38; k=36), indicating that participants who used electronic media more frequently were more likely to have more sleep problems.

Moderator Analyses

At first, we conducted subgroup analyses for different media or devices. The results are shown in Tables 1 and 2. The effect of the relationship between general online gaming and sleep problems was r=0.14 (95% CI 0.06-0.22); the effect of the relationship between general smartphone use and sleep problems was r=0.33 (95% CI 0.27-0.40); and the effect of the relationship between general social media use and sleep problems was r=0.28 (95% CI 0.21-0.34). There are significant differences among these groups ($Q_{between}$=14.46; P=.001).

Table 1. The association of general use of electronic media use and sleep problems.

Media or device types	k	r (95% CI)	P value	I^2
Online gaming	4	0.142 (0.063-0.219)	<.001	90.953
Smartphone	10	0.334 (0.270-0.396)	<.001	82.437
Social media	6	0.277 (0.209-0.342)	<.001	47.067

Table 2. The association of problematic use of electronic media use and sleep problems.

Media or device types	k	r (95% CI)	P value	I^2
Online game	5	0.494 (0.227-0.691)	.001	97.491
Internet	2	0.514 (0.430-0.589)	<.001	38.784
Smartphone	18	0.250 (0.195-0.303)	<.001	88.754
Social media	11	0.347 (0.288-0.403)	<.001	80.835

The effect of the relationship between problematic gaming and sleep problems was r=0.49, 95% CI 0.23-0.69; the effect of the relationship between problematic internet use and sleep problems was r=0.51 (95% CI 0.43-0.59); the effect of the relationship between problematic smartphone use and sleep problems was r=0.25 (95% CI 0.20-0.30); and the effect of the relationship between problematic social media use and sleep problems was r=0.35 (95% CI 0.29-0.40). There are significant differences among these groups ($Q_{between}$=27.37; P<.001).

We also used age, gender, and culture as moderators to conduct meta-regression analyses. The results are shown in Tables 3 and 4. Only cultural difference in the relationship between Eastern and Western culture was significant ($Q_{between}$=6.694; P=.01). All other analyses were not significant.

Table 3. Culture analysis of general use of electronic media use and sleep problems.

Smartphone	k	r (95% CI)	P value
Eastern culture	1	0.211 (0.120-0.299)	__[a]
Western culture	5	0.322 (0.240-0.399)	.07

[a]Not applicable.

XSL•FO
RenderX

Table 4. Culture analysis of problematic use of electronic media use and sleep problems.

Media or device types	*k*	*r* (95% CI)	*P* value
Online game			
Eastern culture	1	0.249 (0.140-0.352)	—[a]
Western culture	3	0.461 (0.062-0.733)	.29
Smartphone			
Eastern culture	10	0.231 (0.164-0.295)	—
Western culture	6	0.287 (0.136-0.425)	.49
Social media			
Eastern culture	5	0.404 (0.333-0.470)	—
Western culture	4	0.276 (0.208-0.342)	.01

[a]Not applicable.

Publication Bias

All funnel plots of the analyses were symmetrical, showing no evidence of publication bias (Figures 2-5). We also conducted *p*-curve analyses to see whether there were any selection biases. The results also showed that there were no biases.

Figure 2. Publication bias of the effect between problematic social media use and sleep problems.

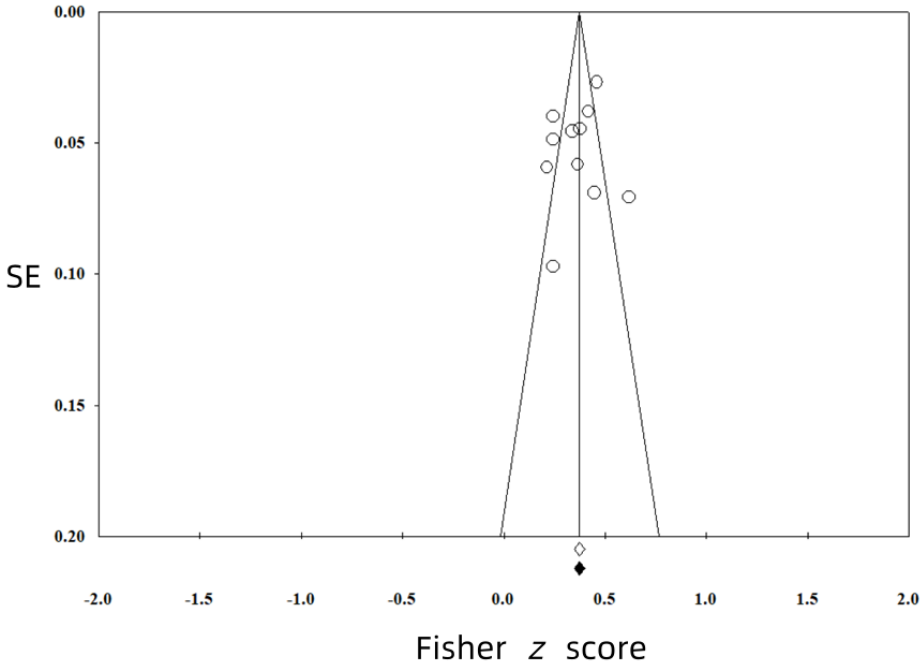

XSL•FO
RenderX

Figure 3. Publication bias of the effect between problematic smartphone use and sleep problems.

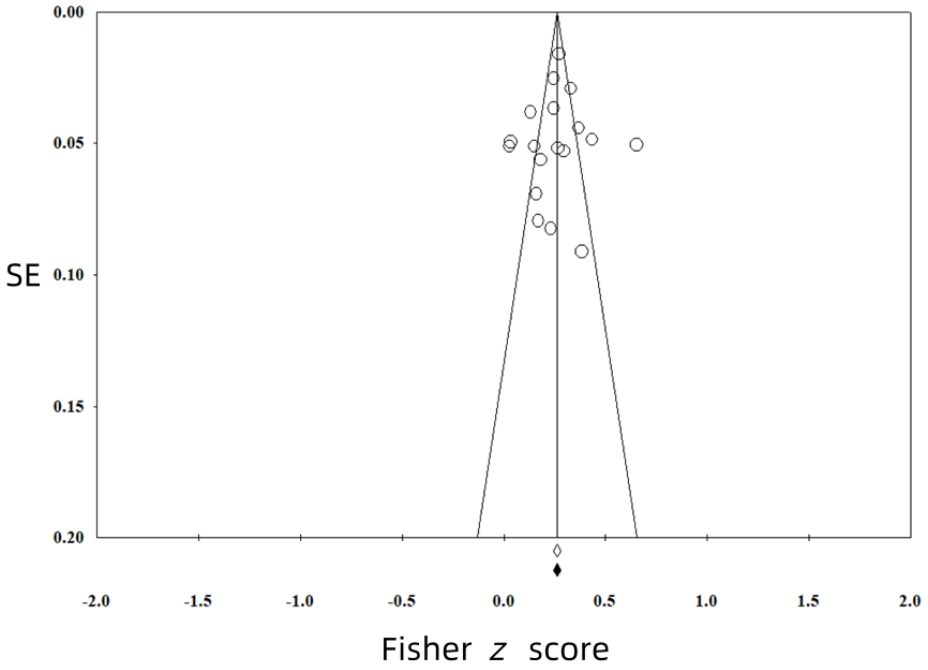

Figure 4. Publication bias of the effect between general smartphone use and sleep problems.

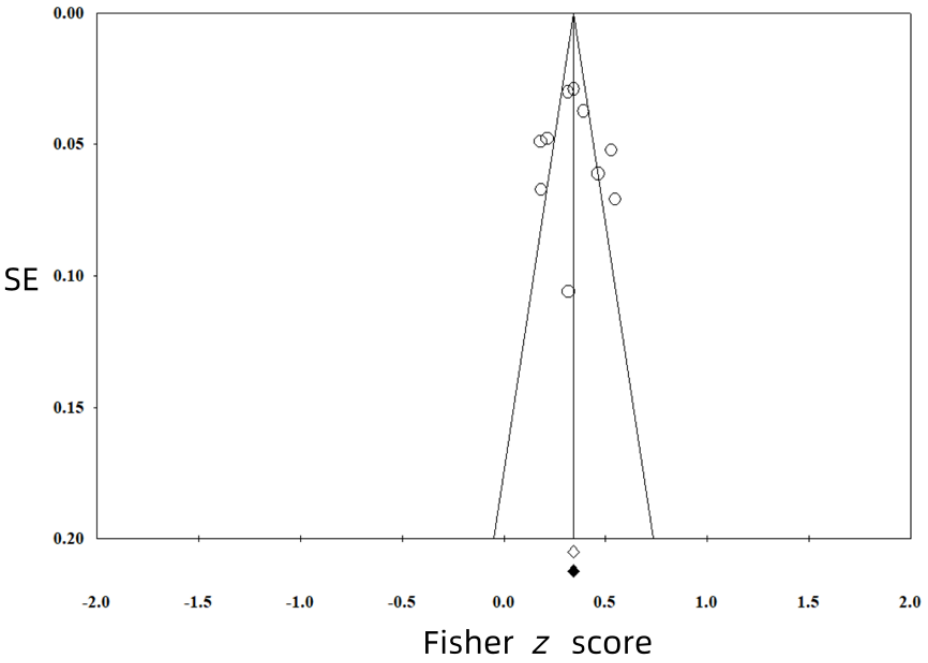

XSL•FO
RenderX

Figure 5. Publication bias of the effect between general social media use and sleep problems.

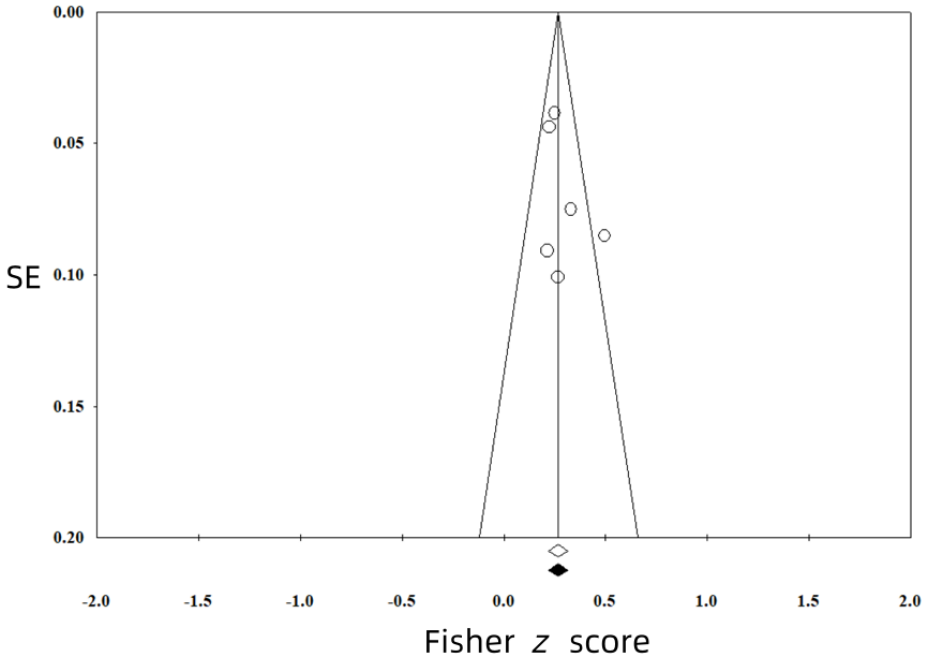

Discussion

Principal Findings

This study indicated that electronic media use was significantly linked with decreased sleep quality and increased sleep problems with varying effect sizes across subgroups. General use was associated with a significant decrease in sleep quality. Problematic use was associated with a significant increase in sleep problems. A significant cultural difference was also observed by the meta-regression analysis.

First, there is a distinction in the impact on sleep quality between problematic use and general use, with the former exhibiting a higher correlation strength. However, both have a positive correlation, suggesting that the deeper the level of use, the more sleep-related issues are observed. In addressing this research question, the way in which electronic media use is conceptualized and operationalized may have a bearing on the ultimate outcomes. Problematic use is measured through addiction scales, while general use is predominantly assessed by duration of use (time), leading to divergent results stemming from these distinct approaches. The key takeaway is that each measurement possesses unique strengths and weaknesses, and the pathways affecting sleep quality differ. Consequently, the selection of a measurement approach should be tailored to the specific research question at hand. The duration of general use reflects an individual's comprehensive involvement with electronic media, and its impact on sleep quality is evident in factors such as an extended time to fall asleep and reduced sleep duration. The addiction scale for problematic use illuminates an individual's preferences, dependencies, and other associations with electronic media. Its impact on sleep quality is evident through physiological and psychological responses, including anxiety, stress, and emotional reactions.

Second, notable variations exist in how different types of electronic media affect sleep quality. In general, the positive predictive effects of smartphone, social media, and online gaming use durations on sleep problems gradually decrease. In the problematic context, the intensity of addiction to the internet and online gaming has the most significant positive impact on sleep problems, followed by social media, while smartphones exert the least influence. On one hand, longitudinal comparisons within the same context reveal that the content and format of electronic media can have varying degrees of negative impact on sleep quality, irrespective of whether it involves general or problematic use. On the other hand, cross-context comparisons suggest that both general and problematic use play a role in moderating the impact of electronic media types on sleep quality. As an illustration, problematic use reinforces the positive impact of online gaming and social media on sleep problems, while mitigating the influence of smartphones. Considering smartphones as electronic media, an extended duration of general use is associated with lower sleep quality. However, during problematic use, smartphones serve as the platform for other electronic media such as games and social media, resulting in a weakened predictive effect on sleep quality. Put differently, in the context of problematic use, the specific type of electronic media an individual consumes on their smartphones becomes increasingly pivotal in shaping sleep quality.

Third, cultural differences were found to be significant moderators of the relationship between electronic media use and sleep problems in both our study and Carter et al [16]. Kristensen et al [17], however, did not specifically address the role of cultural differences but revealed that there was a strong and consistent association between bedtime media device use and sleep outcomes across the studies included. Our findings showed that the association between problematic social media use was significantly larger in Eastern culture. We speculate that the difference may be attributed to cultural differences in social media use patterns, perceptions of social norms and expectations, variations in bedtime routines and habits, and

XSL•FO
RenderX

diverse coping mechanisms for stress. These speculations warrant further investigation to understand better the underlying factors contributing to the observed cultural differences in the relationship between social media use and sleep quality.

Fourth, it was observed that gender and age had no significant impact on sleep quality. The negative effects of electronic media use are not only confined to the sleep quality of adults, and the association with gender differences remains unclear. Recent studies point out that electronic media use among preschoolers may result in a "time-shifting" process, disrupting their sleep patterns [75]. Similarly, children and adolescent sleep patterns have been reported to be adversely affected by electronic media use [76-78]. These findings underscore the necessity of considering age group variations in future research, as electronic media use may differently impact sleep quality across age demographics.

In conclusion, our study, Carter et al [16], and Kristensen et al [17] collectively emphasize the importance of understanding and addressing the negative impact of electronic media use, particularly problematic online gaming and smartphone use, on sleep quality and related issues. Further research is warranted to explore the underlying mechanisms and specific factors contributing to the relationship between electronic media use and sleep problems.

Strengths and Limitations

Our study, supplemented with research by Carter et al [16] and Kristensen et al [17], contributes to the growing evidence supporting a connection between electronic media use and sleep quality. We found that both general and problematic use of electronic media correlates with sleep issues, with the strength of the correlation varying based on the type of electronic media and cultural factors, with no significant relationship observed with age or gender.

Despite the vast amount of research on the relationship between electronic media use and sleep, several gaps and limitations still exist.

First, the inclusion criteria were restricted to English-language, peer-reviewed empirical studies published between January 2018 and October 2023. This may have led to the exclusion of relevant studies published in other languages or before 2018, potentially limiting the generalizability of our findings. Furthermore, the exclusion of non–peer-reviewed studies and conference abstracts may have introduced publication bias, as significant results are more likely to be published in peer-reviewed journals.

Second, although we used a comprehensive search strategy, the possibility remains that some relevant studies may have been missed. Additionally, the search strategies were not linked with Medical Subject Headings headers and may not have captured all possible electronic media types, resulting in an incomplete representation of the effects of electronic media use on sleep quality.

Third, the studies included in our meta-analysis exhibited considerable heterogeneity in sample characteristics, electronic media types, and measures of sleep quality. This heterogeneity might have contributed to the variability in effect sizes observed across studies. Although we conducted moderator analyses to explore potential sources of heterogeneity, other unexamined factors may still have influenced the relationship between electronic media use and sleep quality.

Fourth, our meta-analysis relied on the correlation coefficient (r) as the primary effect size measure, which may not fully capture the complex relationships between electronic media use and sleep quality. Moreover, the conversion of other reported statistics into correlation coefficients could introduce additional sources of error. The correlational nature of the included studies limited our ability to draw causal inferences between electronic media use and sleep quality. Experimental and longitudinal research designs would provide stronger evidence for the directionality of this relationship.

Given these limitations, future research should aim to include a more diverse range of studies, examine additional potential moderators, and use more robust research designs to better understand the complex relationship between electronic media use and sleep quality.

Conclusions

In conclusion, our updated meta-analysis affirms the consistent negative impact of electronic media use on sleep outcomes, with problematic online gaming and smartphone use being particularly impactful. Notably, the negative effect of problematic social media use on sleep quality appears more pronounced in Eastern cultures. This research emphasizes the need for public health initiatives to increase awareness of these impacts, particularly for adolescents. Further research, including experimental and longitudinal studies, is necessary to delve deeper into the complex relationship between electronic media use and sleep quality, considering potential moderators like cultural differences.

Acknowledgments

This research was supported by the Journalism and Marxism Research Center, Renmin University of China (MXG202215), and by funds for building world-class universities (disciplines) of Renmin University of China (23RXW195).

A statement on the use of ChatGPT in the process of writing this paper can be found in Multimedia Appendix 4.

Data Availability

The data sets analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 76 KB - jmir_v26i1e48356_app1.pdf]

Multimedia Appendix 2
Search strategies.
[DOC File , 29 KB - jmir_v26i1e48356_app2.doc]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File , 35 KB - jmir_v26i1e48356_app3.docx]

Multimedia Appendix 4
Large language model statement.
[PDF File (Adobe PDF File), 168 KB - jmir_v26i1e48356_app4.pdf]

References

1. Brink-Kjaer A, Leary EB, Sun H, Westover MB, Stone KL, Peppard PE, et al. Age estimation from sleep studies using deep learning predicts life expectancy. NPJ Digit Med 2022;5(1):103 [FREE Full text] [doi: 10.1038/s41746-022-00630-9] [Medline: 35869169]
2. Killgore WDS. Effects of sleep deprivation on cognition. Prog Brain Res 2010;185:105-129. [doi: 10.1016/B978-0-444-53702-7.00007-5] [Medline: 21075236]
3. Lee S, Mu CX, Wallace ML, Andel R, Almeida DM, Buxton OM, et al. Sleep health composites are associated with the risk of heart disease across sex and race. Sci Rep 2022;12(1):2023 [FREE Full text] [doi: 10.1038/s41598-022-05203-0] [Medline: 35132087]
4. Prather AA. Sleep, stress, and immunity. In: Grandner MA, editor. Sleep and Health, 1st Edition. Cambridge: Academic Press; 2019:319-330.
5. Scott AJ, Webb TL, Martyn-St James M, Rowse G, Weich S. Improving sleep quality leads to better mental health: a meta-analysis of randomised controlled trials. Sleep Med Rev 2021 Dec;60:101556 [FREE Full text] [doi: 10.1016/j.smrv.2021.101556] [Medline: 34607184]
6. Guttmann A. Statista. 2023. URL: https://www.statista.com/topics/1536/media-use/#topicOverview [accessed 2023-06-10]
7. Hysing M, Pallesen S, Stormark KM, Jakobsen R, Lundervold AJ, Sivertsen B. Sleep and use of electronic devices in adolescence: results from a large population-based study. BMJ Open 2015 Feb 02;5(1):e006748 [FREE Full text] [doi: 10.1136/bmjopen-2014-006748] [Medline: 25643702]
8. Lavender RM. Electronic media use and sleep quality. Undergrad J Psychol 2015;28(1):55-62 [FREE Full text]
9. Exelmans L, Van den Bulck J. Bedtime mobile phone use and sleep in adults. Soc Sci Med 2016;148:93-101. [doi: 10.1016/j.socscimed.2015.11.037] [Medline: 26688552]
10. Twenge JM, Krizan Z, Hisler G. Decreases in self-reported sleep duration among U.S. adolescents 2009-2015 and association with new media screen time. Sleep Med 2017;39:47-53. [doi: 10.1016/j.sleep.2017.08.013] [Medline: 29157587]
11. Exelmans L. Electronic media use and sleep: a self-control perspective. Curr Sleep Med Rep 2019;5:135-140. [doi: 10.1007/s40675-019-00147-w]
12. Jniene A, Errguig L, El Hangouche AJ, Rkain H, Aboudrar S, El Ftouh M, et al. Perception of sleep disturbances due to bedtime use of blue light-emitting devices and its impact on habits and sleep quality among young medical students. Biomed Res Int 2019;2019:7012350 [FREE Full text] [doi: 10.1155/2019/7012350] [Medline: 31950050]
13. Munezawa T, Kaneita Y, Osaki Y, Kanda H, Minowa M, Suzuki K, et al. The association between use of mobile phones after lights out and sleep disturbances among Japanese adolescents: a nationwide cross-sectional survey. Sleep 2011;34(8):1013-1020 [FREE Full text] [doi: 10.5665/SLEEP.1152] [Medline: 21804663]
14. Smith LJ, Gradisar M, King DL, Short M. Intrinsic and extrinsic predictors of video-gaming behaviour and adolescent bedtimes: the relationship between flow states, self-perceived risk-taking, device accessibility, parental regulation of media and bedtime. Sleep Med 2017;30:64-70. [doi: 10.1016/j.sleep.2016.01.009] [Medline: 28215265]
15. Alimoradi Z, Lin CY, Broström A, Bülow PH, Bajalan Z, Griffiths MD, et al. Internet addiction and sleep problems: a systematic review and meta-analysis. Sleep Med Rev 2019;47:51-61. [doi: 10.1016/j.smrv.2019.06.004] [Medline: 31336284]

XSL•FO
RenderX

16. Carter B, Rees P, Hale L, Bhattacharjee D, Paradkar MS. Association between portable screen-based media device access or use and sleep outcomes: a systematic review and meta-analysis. JAMA Pediatr 2016;170(12):1202-1208 [FREE Full text] [doi: 10.1001/jamapediatrics.2016.2341] [Medline: 27802500]
17. Kristensen JH, Pallesen S, King DL, Hysing M, Erevik EK. Problematic gaming and sleep: a systematic review and meta-analysis. Front Psychiatry 2021;12:675237 [FREE Full text] [doi: 10.3389/fpsyt.2021.675237] [Medline: 34163386]
18. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Akçay D, Akçay BD. The effect of computer game playing habits of university students on their sleep states. Perspect Psychiatr Care 2020;56(4):820-826. [doi: 10.1111/ppc.12497] [Medline: 32163182]
21. Alahdal WM, Alsaedi AA, Garrni AS, Alharbi FS. The impact of smartphone addiction on sleep quality among high school students in Makkah, Saudi Arabia. Cureus 2023;15(6):e40759 [FREE Full text] [doi: 10.7759/cureus.40759] [Medline: 37485102]
22. Alam A, Alshakhsi S, Al-Thani D, Ali R. The role of objectively recorded smartphone usage and personality traits in sleep quality. PeerJ Comput Sci 2023;9:e1261 [FREE Full text] [doi: 10.7717/peerj-cs.1261] [Medline: 37346703]
23. Almeida F, Marques DR, Gomes AA. A preliminary study on the association between social media at night and sleep quality: the relevance of FOMO, cognitive pre-sleep arousal, and maladaptive cognitive emotion regulation. Scand J Psychol 2023;64(2):123-132 [FREE Full text] [doi: 10.1111/sjop.12880] [Medline: 36256468]
24. Alshobaili FA, AlYousefi NA. The effect of smartphone usage at bedtime on sleep quality among Saudi non-medical staff at King Saud University Medical City. J Family Med Prim Care 2019;8(6):1953-1957 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_269_19] [Medline: 31334161]
25. Alsulami A, Bakhsh D, Baik M, Merdad M, Aboalfaraj N. Assessment of sleep quality and its relationship to social media use among medical students. Med Sci Educ 2019;29(1):157-161 [FREE Full text] [doi: 10.1007/s40670-018-00650-9] [Medline: 34457463]
26. Altintas E, Karaca Y, Hullaert T, Tassi P. Sleep quality and video game playing: effect of intensity of video game playing and mental health. Psychiatry Res 2019;273:487-492. [doi: 10.1016/j.psychres.2019.01.030] [Medline: 30685731]
27. Asbee J, Slavish D, Taylor DJ, Dietch JR. Using a frequentist and Bayesian approach to examine video game usage, substance use, and sleep among college students. J Sleep Res 2023;32(4):e13844. [doi: 10.1111/jsr.13844] [Medline: 36814416]
28. Bae ES, Kang HS, Lee HN. The mediating effect of sleep quality in the relationship between academic stress and social network service addiction tendency among adolescents. J Korean Acad Community Health Nurs 2020;31(3):290-299 [FREE Full text] [doi: 10.12799/jkachn.2020.31.3.290]
29. Chatterjee S, Kar SK. Smartphone addiction and quality of sleep among Indian medical students. Psychiatry 2021;84(2):182-191. [doi: 10.1080/00332747.2021.1907870] [Medline: 33856961]
30. Chung JE, Choi SA, Kim KT, Yee J, Kim JH, Seong JW, et al. Smartphone addiction risk and daytime sleepiness in Korean adolescents. J Paediatr Child Health 2018;54(7):800-806. [doi: 10.1111/jpc.13901] [Medline: 29626363]
31. Demir YP, Sumer MM. Effects of smartphone overuse on headache, sleep and quality of life in migraine patients. Neurosciences (Riyadh) 2019;24(2):115-121 [FREE Full text] [doi: 10.17712/nsj.2019.2.20180037] [Medline: 31056543]
32. Dewi RK, Efendi F, Has EMM, Gunawan J. Adolescents' smartphone use at night, sleep disturbance and depressive symptoms. Int J Adolesc Med Health 2018;33(2):20180095. [doi: 10.1515/ijamh-2018-0095] [Medline: 30447141]
33. Eden A, Ellithorpe ME, Meshi D, Ulusoy E, Grady SM. All night long: problematic media use is differentially associated with sleep quality and depression by medium. Commun Res Rep 2021;38(3):143-149. [doi: 10.1080/08824096.2021.1902798]
34. Ellithorpe ME, Meshi D, Tham SM. Problematic video gaming is associated with poor sleep quality, diet quality, and personal hygiene. Psychol Pop Media 2023;12(2):248-253. [doi: 10.1037/ppm0000397]
35. Elsheikh AA, Elsharkawy SA, Ahmed DS. Impact of smartphone use at bedtime on sleep quality and academic activities among medical students at Al -Azhar University at Cairo. J Public Health (Berl.) 2023 Jun 15:1-10 [FREE Full text] [doi: 10.1007/s10389-023-01964-8]
36. Gaya AR, Brum R, Brites K, Gaya A, de Borba Schneiders L, Duarte Junior MA, et al. Electronic device and social network use and sleep outcomes among adolescents: the EHDLA study. BMC Public Health 2023;23(1):919 [FREE Full text] [doi: 10.1186/s12889-023-15579-x] [Medline: 37208650]
37. Gezgin DM. Understanding patterns for smartphone addiction: age, sleep duration, social network use and fear of missing out. Cypriot J Educ Sci 2018;13(2):166-177. [doi: 10.18844/cjes.v13i2.2938]
38. Graham S, Mason A, Riordan B, Winter T, Scarf D. Taking a break from social media improves wellbeing through sleep quality. Cyberpsychol Behav Soc Netw 2021;24(6):421-425. [doi: 10.1089/cyber.2020.0217] [Medline: 33395534]
39. Guerrero MD, Barnes JD, Chaput JP, Tremblay MS. Screen time and problem behaviors in children: exploring the mediating role of sleep duration. Int J Behav Nutr Phys Act 2019;16(1):105 [FREE Full text] [doi: 10.1186/s12966-019-0862-x] [Medline: 31727084]

XSL•FO
RenderX

40. Hamvai C, Kiss H, Vörös H, Fitzpatrick KM, Vargha A, Pikó BF. Association between impulsivity and cognitive capacity decrease is mediated by smartphone addiction, academic procrastination, bedtime procrastination, sleep insufficiency and daytime fatigue among medical students: a path analysis. BMC Med Educ 2023;23(1):537 [FREE Full text] [doi: 10.1186/s12909-023-04522-8] [Medline: 37501113]
41. Herlache AD, Lang KM, Krizan Z. Withdrawn and wired: problematic internet use accounts for the link of neurotic withdrawal to sleep disturbances. Sleep Sci 2018;11(2):69-73 [FREE Full text] [doi: 10.5935/1984-0063.20180015] [Medline: 30083293]
42. Huang Q, Li Y, Huang S, Qi J, Shao T, Chen X, et al. Smartphone use and sleep quality in chinese college students: a preliminary study. Front Psychiatry 2020;11:352 [FREE Full text] [doi: 10.3389/fpsyt.2020.00352] [Medline: 32435208]
43. Hussain Z, Griffiths MD. The associations between problematic social networking site use and sleep quality, attention-deficit hyperactivity disorder, depression, anxiety and stress. Int J Ment Health Addict 2021;19:686-700 [FREE Full text] [doi: 10.1007/s11469-019-00175-1]
44. Imani V, Ahorsu DK, Taghizadeh N, Parsapour Z, Nejati B, Chen HP, et al. The mediating roles of anxiety, depression, sleepiness, insomnia, and sleep quality in the association between problematic social media use and quality of life among patients with cancer. Healthcare (Basel) 2022;10(9):1745 [FREE Full text] [doi: 10.3390/healthcare10091745] [Medline: 36141357]
45. Jeong CY, Seo YS, Cho EH. The effect of SNS addiction tendency on trait-anxiety and quality of sleep in university students'. J Korean Clin Health Sci 2018;6(2):1147-1155. [doi: 10.15205/KSCHS.2018.12.31.1147]
46. Karaş H, Küçükparlak İ, Özbek MG, Yılmaz T. Addictive smartphone use in the elderly: relationship with depression, anxiety and sleep quality. Psychogeriatrics 2023;23(1):116-125. [doi: 10.1111/psyg.12910] [Medline: 36366976]
47. Kater MJ, Schlarb AA. Smartphone usage in adolescents: motives and link to sleep disturbances, stress and sleep reactivity. Somnologie 2020;24(4):245-252. [doi: 10.1007/s11818-020-00272-7]
48. Kharisma AC, Fitryasari R, Rahmawati PD. Online games addiction and the decline in sleep quality of college student gamers in the online game communities in Surabaya, Indonesia. Int J Psychosoc Rehabil 2020;24(7):8987-8993 [FREE Full text] [doi: 10.37200/V24I7/19237]
49. Kumar VA, Chandrasekaran V, Brahadeeswari H. Prevalence of smartphone addiction and its effects on sleep quality: a cross-sectional study among medical students. Ind Psychiatry J 2019;28(1):82-85 [FREE Full text] [doi: 10.4103/ipj.ipj_56_19] [Medline: 31879452]
50. Lee Y, Blebea J, Janssen F, Domoff SE. The impact of smartphone and social media use on adolescent sleep quality and mental health during the COVID-19 pandemic. Hum Behav Emerg Technol 2023;2023:3277040 [FREE Full text] [doi: 10.1155/2023/3277040]
51. Li L, Griffiths MD, Mei S, Niu Z. Fear of missing out and smartphone addiction mediates the relationship between positive and negative affect and sleep quality among Chinese university students. Front Psychiatry 2020;11:877 [FREE Full text] [doi: 10.3389/fpsyt.2020.00877] [Medline: 33192635]
52. Li Y, Mu W, Sun C, Kwok SYCL. Surrounded by smartphones: relationship between peer phubbing, psychological distress, problematic smartphone use, daytime sleepiness, and subjective sleep quality. Appl Res Qual Life 2023;18:1099-1114. [doi: 10.1007/s11482-022-10136-x]
53. Luo X, Hu C. Loneliness and sleep disturbance among first-year college students: the sequential mediating effect of attachment anxiety and mobile social media dependence. Psychol Sch 2022;59(9):1776-1789. [doi: 10.1002/pits.22721]
54. Luqman A, Masood A, Shahzad F, Shahbaz M, Feng Y. Untangling the adverse effects of late-night usage of smartphone-based SNS among university students. Behav Inf Technol 2021;40(15):1671-1687. [doi: 10.1080/0144929x.2020.1773538]
55. Makhfudli, Aulia A, Pratiwi A. Relationship intensity of social media use with quality of sleep, social interaction, and self-esteem in urban adolescents in Surabaya. Sys Rev Pharm 2020;11(5):783-788. [doi: 10.31838/srp.2020.5.115]
56. Ozcan B, Acimis NM. Sleep quality in Pamukkale university students and its relationship with smartphone addiction. Pak J Med Sci 2021;37(1):206-211 [FREE Full text] [doi: 10.12669/pjms.37.1.3130] [Medline: 33437278]
57. Peltz JS, Bodenlos JS, Kingery JN, Abar C. Psychological processes linking problematic smartphone use to sleep disturbance in young adults. Sleep Health 2023;9(4):524-531. [doi: 10.1016/j.sleh.2023.05.005] [Medline: 37460376]
58. Pérez-Chada D, Bioch SA, Schönfeld D, Gozal D, Perez-Lloret S, Sleep in Adolescents Collaborative Study Group. Screen use, sleep duration, daytime somnolence, and academic failure in school-aged adolescents. PLoS One 2023;18(2):e0281379 [FREE Full text] [doi: 10.1371/journal.pone.0281379] [Medline: 36787301]
59. Przepiorka A, Blachnio A. The role of Facebook intrusion, depression, and future time perspective in sleep problems among adolescents. J Res Adolesc 2020;30(2):559-569. [doi: 10.1111/jora.12543] [Medline: 31868972]
60. Rudolf K, Bickmann P, Froböse I, Tholl C, Wechsler K, Grieben C. Demographics and health behavior of video game and eSports players in Germany: the eSports study 2019. Int J Environ Res Public Health 2020;17(6):1870 [FREE Full text] [doi: 10.3390/ijerph17061870] [Medline: 32183070]
61. Sami H, Danielle L, Lihi D, Elena S. The effect of sleep disturbances and internet addiction on suicidal ideation among adolescents in the presence of depressive symptoms. Psychiatry Res 2018;267:327-332. [doi: 10.1016/j.psychres.2018.03.067] [Medline: 29957549]

XSL•FO
RenderX

62. Scott H, Woods HC. Fear of missing out and sleep: cognitive behavioural factors in adolescents' nighttime social media use. J Adolesc 2018;68:61-65. [doi: 10.1016/j.adolescence.2018.07.009] [Medline: 30031979]
63. Spagnoli P, Balducci C, Fabbri M, Molinaro D, Barbato G. Workaholism, intensive smartphone use, and the sleep-wake cycle: a multiple mediation analysis. Int J Environ Res Public Health 2019;16(19):3517 [FREE Full text] [doi: 10.3390/ijerph16193517] [Medline: 31547191]
64. Stanković M, Nešić M, Čičević S, Shi Z. Association of smartphone use with depression, anxiety, stress, sleep quality, and internet addiction. empirical evidence from a smartphone application. Pers Individ Differ 2021 Jan;168:110342. [doi: 10.1016/j.paid.2020.110342]
65. Tandon A, Kaur P, Dhir A, Mäntymäki M. Sleepless due to social media? investigating problematic sleep due to social media and social media sleep hygiene. Comput Hum Behav 2020 Dec;113:106487 [FREE Full text] [doi: 10.1016/j.chb.2020.106487]
66. Wang PY, Chen KL, Yang SY, Lin PH. Relationship of sleep quality, smartphone dependence, and health-related behaviors in female junior college students. PLoS One 2019;14(4):e0214769 [FREE Full text] [doi: 10.1371/journal.pone.0214769] [Medline: 30943270]
67. Wang Q, Zhong Y, Zhao G, Song R, Zeng C. Relationship among content type of smartphone use, technostress, and sleep difficulty: a study of university students in China. Educ Inf Technol 2022 Aug 02;28(2):1697-1714. [doi: 10.1007/s10639-022-11222-1]
68. Wong HY, Mo HY, Potenza MN, Chan MNM, Lau WM, Chui TK, et al. Relationships between severity of internet gaming disorder, severity of problematic social media use, sleep quality and psychological distress. Int J Environ Res Public Health 2020;17(6):1879 [FREE Full text] [doi: 10.3390/ijerph17061879] [Medline: 32183188]
69. Xie X, Dong Y, Wang J. Sleep quality as a mediator of problematic smartphone use and clinical health symptoms. J Behav Addict 2018;7(2):466-472 [FREE Full text] [doi: 10.1556/2006.7.2018.40] [Medline: 29788754]
70. Yang SY, Chen KL, Lin PH, Wang PY. Relationships among health-related behaviors, smartphone dependence, and sleep duration in female junior college students. Soc Health Behav 2019;2(1):26-31 [FREE Full text] [doi: 10.4103/shb.shb_44_18]
71. Yıldırım M, Öztürk A, Solmaz F. Fear of COVID-19 and sleep problems in Turkish young adults: mediating roles of happiness and problematic social networking sites use. Psihologija 2023;56(4):497-515 [FREE Full text] [doi: 10.2298/PSI220412027Y]
72. Zhai X, Ye M, Wang C, Gu Q, Huang T, Wang K, et al. Associations among physical activity and smartphone use with perceived stress and sleep quality of Chinese college students. Mental Health and Physical Activity 2020 Mar;18:100323. [doi: 10.1016/j.mhpa.2020.100323]
73. Zhang MX, Wu AMS. Effects of smartphone addiction on sleep quality among Chinese university students: the mediating role of self-regulation and bedtime procrastination. Addict Behav 2020;111:106552. [doi: 10.1016/j.addbeh.2020.106552] [Medline: 32717501]
74. Zhang MX, Zhou H, Yang HM, Wu AMS. The prospective effect of problematic smartphone use and fear of missing out on sleep among Chinese adolescents. Curr Psychol 2021 May 24;42(7):5297-5305. [doi: 10.1007/s12144-021-01863-9]
75. Beyens I, Nathanson AI. Electronic media use and sleep among preschoolers: evidence for time-shifted and less consolidated sleep. Health Commun 2019;34(5):537-544 [FREE Full text] [doi: 10.1080/10410236.2017.1422102] [Medline: 29323943]
76. Mazurek MO, Engelhardt CR, Hilgard J, Sohl K. Bedtime electronic media use and sleep in children with autism spectrum disorder. J Dev Behav Pediatr 2016;37(7):525-531. [doi: 10.1097/DBP.0000000000000314] [Medline: 27355885]
77. King DL, Delfabbro PH, Zwaans T, Kaptsis D. Sleep interference effects of pathological electronic media use during adolescence. Int J Ment Health Addict 2014;12:21-35. [doi: 10.1007/s11469-013-9461-2]
78. Kubiszewski V, Fontaine R, Rusch E, Hazouard E. Association between electronic media use and sleep habits: an eight-day follow-up study. Int J Adolesc Youth 2013;19(3):395-407 [FREE Full text] [doi: 10.1080/02673843.2012.751039]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach, T Leung; submitted 20.04.23; peer-reviewed by M Behzadifar, F Estévez-López, R Prieto-Moreno; comments to author 18.05.23; revised version received 15.06.23; accepted 26.03.24; published 23.04.24.

Please cite as:
Han X, Zhou E, Liu D
Electronic Media Use and Sleep Quality: Updated Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e48356
URL: https://www.jmir.org/2024/1/e48356
doi:10.2196/48356
PMID:38533835

XSL•FO
RenderX

©Xiaoning Han, Enze Zhou, Dong Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Behavior Change Approaches in Digital Technology–Based Physical Rehabilitation Interventions Following Stroke: Scoping Review

Helen J Gooch[1], BSc; Kathryn A Jarvis[1], PhD; Rachel C Stockley[1], PhD

Stroke Research Team, School of Nursing and Midwifery, University of Central Lancashire, Preston, United Kingdom

Corresponding Author:
Helen J Gooch, BSc
Stroke Research Team
School of Nursing and Midwifery
University of Central Lancashire
BB247 Brook Building
Victoria Street
Preston, PR1 2HE
United Kingdom
Phone: 44 1772894956
Email: HJGooch@uclan.ac.uk

Abstract

Background: Digital health technologies (DHTs) are increasingly used in physical stroke rehabilitation to support individuals in successfully engaging with the frequent, intensive, and lengthy activities required to optimize recovery. Despite this, little is known about behavior change within these interventions.

Objective: This scoping review aimed to identify if and how behavior change approaches (ie, theories, models, frameworks, and techniques to influence behavior) are incorporated within physical stroke rehabilitation interventions that include a DHT.

Methods: Databases (Embase, MEDLINE, PsycINFO, CINAHL, Cochrane Library, and AMED) were searched using keywords relating to behavior change, DHT, physical rehabilitation, and stroke. The results were independently screened by 2 reviewers. Sources were included if they reported a completed primary research study in which a behavior change approach could be identified within a physical stroke rehabilitation intervention that included a DHT. Data, including the study design, DHT used, and behavior change approaches, were charted. Specific behavior change techniques were coded to the behavior change technique taxonomy version 1 (BCTTv1).

Results: From a total of 1973 identified sources, 103 (5%) studies were included for data charting. The most common reason for exclusion at full-text screening was the absence of an explicit approach to behavior change (165/245, 67%). Almost half (45/103, 44%) of the included studies were described as pilot or feasibility studies. Virtual reality was the most frequently identified DHT type (58/103, 56%), and almost two-thirds (65/103, 63%) of studies focused on upper limb rehabilitation. Only a limited number of studies (18/103, 17%) included a theory, model, or framework for behavior change. The most frequently used BCTTv1 clusters were feedback and monitoring (88/103, 85%), reward and threat (56/103, 54%), goals and planning (33/103, 32%), and shaping knowledge (33/103, 32%). Relationships between feedback and monitoring and reward and threat were identified using a relationship map, with prominent use of both of these clusters in interventions that included virtual reality.

Conclusions: Despite an assumption that DHTs can promote engagement in rehabilitation, this scoping review demonstrates that very few studies of physical stroke rehabilitation that include a DHT overtly used any form of behavior change approach. From those studies that did consider behavior change, most did not report a robust underpinning theory. Future development and research need to explicitly articulate how including DHTs within an intervention may support the behavior change required for optimal engagement in physical rehabilitation following stroke, as well as establish their effectiveness. This understanding is likely to support the realization of the transformative potential of DHTs in stroke rehabilitation.

(J Med Internet Res 2024;26:e48725) doi:10.2196/48725

XSL•FO
RenderX

KEYWORDS

behavior change; behavior therapy; digital health technology; digital health; digital technology; health behavior; physical and rehabilitation medicine; scoping review; stroke rehabilitation

Introduction

Background

Digital health technologies (DHTs) comprise apps, programs, or software used in the health and social care systems [1]. They are considered to have almost unlimited potential to transform health care interventions and delivery and empower people to take a greater role in their own care and well-being [2,3].

Stroke is one of the leading causes of acquired disability worldwide, with around 12 million people experiencing a stroke each year [4]. Rehabilitation is a complex, multifaceted process [5] that facilitates those with health conditions and disabilities to participate in and gain independence in meaningful life roles [6]. It is considered an essential aspect of health care provision following a stroke [7] as a means to address poststroke impairments, which can involve motor, sensory, and cognitive functions. Changes in the ability to move due to impairment of both movement and sensory function are commonly experienced by people following a stroke [8] and are addressed by physical rehabilitation comprising regular, intensive practice and repetition of movements and tasks [9,10]. Conventional physical rehabilitation often struggles to deliver the intensity required to optimize recovery [11], and over recent years, there has been significant interest in the use of DHTs, such as virtual reality (VR), telerehabilitation, robotics, and activity monitors [12-16], to enhance and increase the intensity of rehabilitation. DHTs can provide a whole intervention or be used as a component of a wider intervention; the term DHT-based intervention has been used within this review to refer to both situations.

For many people who survive a stroke, rehabilitation requires individuals to engage in regular and frequent rehabilitative activities to achieve improvements in function and realize their optimal recovery. This necessitates adjustments to an individual's behavior [17] over a sustained period of time. Changing behavior is a complex process and is underpinned by a variety of different theories, models, and frameworks [18], such as social cognitive theory [19] or the behavior change wheel framework [20]. Individual activities within a complex intervention that are designed to change behavior can be separated into replicable active components widely referred to as behavior change techniques (BCTs) [21]. Historically, labels applied to BCTs have lacked consensus, resulting in uncertainty and difficulty in comparing interventions. This has been addressed in the behavior change technique taxonomy version 1 (BCTTv1) [22], a classification system of 93 distinct BCTs clustered into 16 groups, which is a well-recognized tool to provide consistency with BCT reporting in interventions. DHTs provide an emerging opportunity to support the behavior change required within physical stroke rehabilitation interventions through facilitators that are embedded within the technology itself that aim to form, alter, or reinforce behaviors [23]. Understanding of this area is limited, with most literature exploring the use of DHTs to support behavior change focused on specific health-related behaviors such as physical activity or healthy eating [24] rather than as a core component of a type of rehabilitation intervention. Motivation is acknowledged to play an integral role in behavior change [25], and it is often assumed that DHTs provide motivation to engage with rehabilitation [26]. However, for this assumption to be realized, the DHTs must be able to support and deliver interventions that facilitate the vital changes in behavior needed to promote prolonged and sustained engagement in stroke rehabilitative activities. Imperative to this is understanding the theories, models, and frameworks that underpin interventions and the BCTs (active components) within the interventions [27-29]. The theories, models, and frameworks alongside the BCTs will be referred to hereinafter as approaches. Within the context of DHT-based physical stroke rehabilitation interventions, approaches to behavior change warrant further investigation.

Aim and Objectives

This scoping review aimed to identify if and how behavior change approaches are incorporated within DHT-based physical stroke rehabilitation interventions. Specifically, it sought to:

1. Establish if behavior change theories, models, and frameworks, or BCTs, are described when reporting on DHT-based interventions that have been developed or evaluated for use in poststroke physical rehabilitation.
2. Identify if behavior change theories, models, or frameworks underpin the interventions and which of these are being used.
3. Identify if the BCTTv1 is being used to report BCTs within interventions.
4. Determine which BCTs (based on the BCTTv1) can be identified within the interventions.
5. Explore whether the type of technology influences the techniques used to change behaviors.

Methods

Review Methodology

A scoping review was completed and reported following established guidelines [30,31] and the Preferred Reporting of Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR; Multimedia Appendix 1) [32]. The protocol was registered with the Open Science Framework [33].

Eligibility Criteria

Any published sources that reported a completed primary research study in which a behavior change approach could be identified within a DHT-based physical stroke rehabilitation intervention were included (Multimedia Appendix 2). Physical rehabilitation comprised interventions that addressed an impairment, or sequela of impairment, of sensory function and pain, neuromusculoskeletal and movement-related functions, or voice and speech, as defined by the International Classification of Functioning, Disability, and Health [34]. Completed primary research included all types of studies, both

XSL•FO
RenderX

quantitative and qualitative, and no minimum sample size or intervention length was set. The BCTTv1 [22] was used to support the identification of BCTs within the interventions.

Information Sources and Search Strategy

A systematic database search was conducted in Embase, MEDLINE, PsycINFO, CINAHL, Cochrane Database of Systematic Reviews, CENTRAL Register of Controlled Trials, and AMED on March 21, 2023. The search was completed in collaboration with an information specialist who provided support with the development of the free text and thesaurus search terms, created the final search, adjusted the searches for the different databases, and ran the search. It consisted of 4 distinct search streams: behavior change, DHT, physical rehabilitation, and stroke, which were then combined (Multimedia Appendix 3). Searches were restricted to the English language (due to review resources) and by date to search from 2001; the date restriction acknowledges the main time period of DHT growth [35], captures sources reported in systematic reviews of DHTs in stroke rehabilitation [12-16], and is reflected in other scoping literature exploring DHTs [24]. Additional sources were identified by hand searching, including scrutiny of the included source reference lists.

Selection of Sources of Evidence

The titles and abstracts of deduplicated sources from database searches and hand searches were independently screened by 2 reviewers, 1 of whom had completed the BCTTv1 web-based training package [36] to inform decisions made around the use of BCTs. Any conflicts were discussed, and if a consensus was not reached, the source was included for full-text screening. Attempts were made to locate a completed study publication from eligible conference abstracts, protocols, and trial registry entries. Full-text sources were screened independently by 2 reviewers, and disagreements were resolved by a third reviewer. Reasons for full-text exclusion were recorded. EndNote X9 software (Clarivate) and the Rayyan web tool software (Qatar Computing Research Institute) [37] were used to facilitate the source selection process.

Data Charting Process

A review-specific data charting tool was developed and initially piloted using 3 sources by 3 reviewers, and then further developed iteratively throughout the process [30]. Data charting was completed collectively by 2 reviewers. When several sources referred to a single study, these sources were grouped together for data charting, and if a source identified additional sources for further detail of the intervention (eg, a protocol or supplementary material), then this information was also used to support data charting.

Data Items

The data charting tool was developed with reference to the Template for Intervention Description and Replication (TIDieR) checklist [27] and with a focus on the DHT-based intervention and behavior change approaches (Multimedia Appendix 4 [14,38-40]). In the absence of a recognized predefined taxonomy for DHTs, the DHTs used in the sources were charted iteratively by the type of technology [41] from the information provided about the intervention. Over time, DHT categories emerged and were defined (Multimedia Appendix 4). Discrete BCTs were identified from the intervention detail provided using the BCTTv1 [22] (Multimedia Appendix 5 [42]). A pragmatic decision was made that the single reviewer who had completed the BCTTv1 web-based training package [36] would code the interventions to the BCTTv1. Any areas of uncertainty were discussed in detail among the review team.

Synthesis of Results

In accordance with the aims of a scoping review, formal assessments of methodological quality were not completed [30,31]. Findings were synthesized using descriptive statistics facilitated by SPSS Statistics 28.0.0.0 (IBM Corp) and Microsoft Excel (version 2208; Microsoft Corporation) and presented in text, table, and chart formats. The characteristics of the included sources, specifically participant numbers, age, and time since stroke, and intervention details, were summarized to provide contextual information for the review. Time since stroke was based on a published timeline framework [43], which describes the following phases: acute (1-7 days), early subacute (7 days to 3 months), late subacute (3-6 months), and chronic (greater than 6 months).

The behavior change theories, models, or frameworks underpinning the DHT-based interventions and sources where interventions had already been coded to the BCTTv1 were summarized. The use of individual BCTs, as coded by reviewers from intervention descriptions, was briefly summarized; however, the main focus of the BCT synthesis was completed by grouping the BCTs into the 16 BCTTv1 clusters, in order to provide an overview of their use across the sources and allow comparison with other reviews [44,45]. A cluster was only identified once per source, irrespective of the number of individual BCTs within that cluster. Relationships between BCTTv1 clusters and between DHT type and BCTTv1 clusters were descriptively explored. A relationship map was used to visually represent the strength of the connections between the BCTTv1 clusters, with a thicker line indicating that variables were more frequently reported together. No inferential statistical analysis was used.

Results

Selection of Sources of Evidence

From a total of 1973 sources screened, 357 full-text sources were assessed for eligibility, then after grouping sources that referred to a single study, 103 (5%) distinct sources were included in the review [46-148] (Figure 1). Of the 245 sources excluded at full-text screening, 165 (67%) were excluded due to a lack of a behavior change approach.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting of Systematic Reviews and Meta-Analyses) flow diagram of the source selection process.

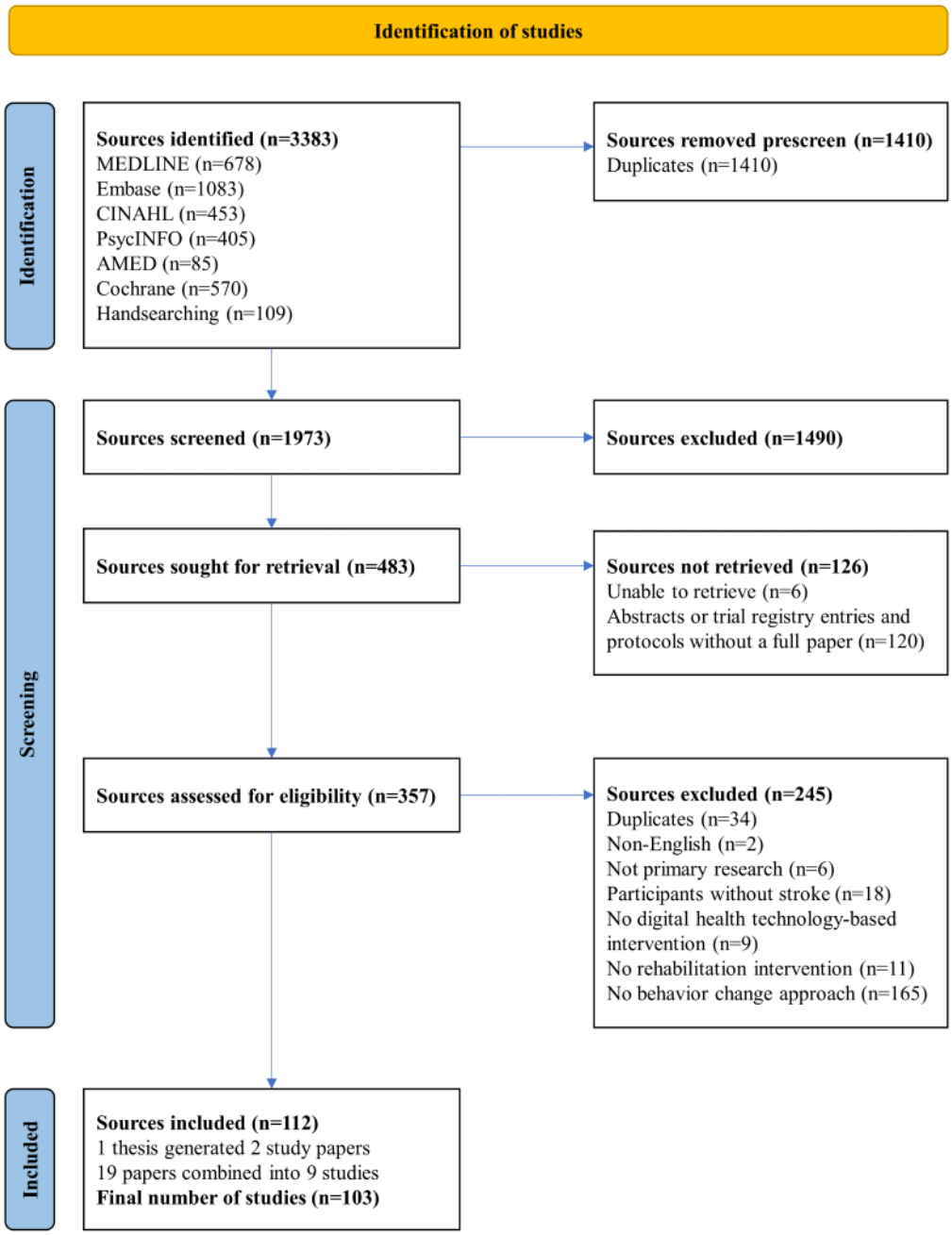

Characteristics of Sources of Evidence

General

All sources of evidence were studies and will be referred to as such hereinafter. The number of studies in this field has rapidly increased over time (Figure 2), from a single study in 2004 to 8 in 2022, with a peak of 15 in 2021. The majority (86/103, 83%) [47-51,53-56,58,59,61,63-68,71-86,89-95,97-105,107,109,111,112,114,115,117-126,128-136,138-148] were published in the past 10 years. Most studies took place in North America (41/103, 40%) [46-49,52,55,56,60,64-67,69,70,72,74,76-78,80,85-88,92,93,97,99,101,108-110,126-129,137,138,141,142,145] and Europe (35/103, 34%) [51,53,54,57,58,62,63,68,71,79,81-84,89,111,113-125,132,136,140,143,146,147], with the remainder in Asia (16/103, 16%) [50,59,61,91,94,95,98,100,102-104,107,135,139,144,148], Australasia (9/103, 9%) [75,96,105,106,112,130,131,133,134], Africa (1/103, 1%) [90], and a single multicontinental study (1/103, 1%) [73]. Almost half (45/103, 44%) the studies are reported as feasibility or pilot studies [49,56,58,64,66,68,69,72-74,76,77,79,82-84,89,90,92,93,95,97,100-104,106,108,114,116,117,119,122,124-126,131,134,136,138,139,141,143,147]. Other study designs included randomized controlled trials (20/103, 19%) [50,51,60,61,65,75,80,85,86,91,107,109,112,128-130,137,144,146,148], single session investigations (19/103, 18%) [47,52,57,59,71,78,87,88,98,110,115,118,120,123,127,132,133,135,142], nonrandomized experimental designs (13/103, 13%) [53-55,62,63,67,81,94,96,99,105, 113,145], case studies (4/103, 4%) [46,48,70,140], and realist evaluations (2/103, 2%) [111,121].

XSL•FO
RenderX

Figure 2. Number of studies by publication year (January 1, 2001, to March 21, 2023).

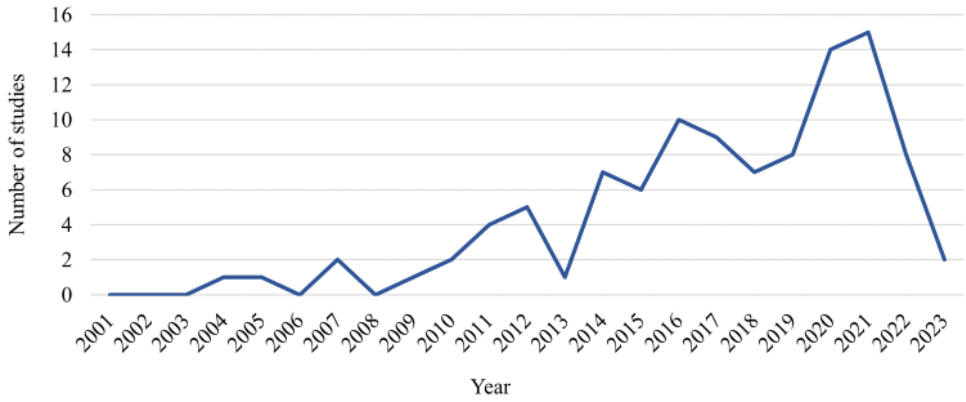

Participants

There were a total of 2825 participants in the 103 included studies. Studies tended to be small, with a median of 16 participants and a range of 1-188. Only half (55/103, 53%) the studies [46-48,50,56,57,59,61,64,67,69-72,78,79,82,87,88,92,93,95-99,101,102,105,106,108,111-121,123-127,134,138-140,142,143,145, 147] reported the minimum and maximum age of participants, which ranged from 17 to 99 years. Over three-quarters (83/103, 81%; 2508 participants) of studies reported the time since the onset of stroke. Of these 83 studies, 1 (1%; 48 participants) study [91] was conducted in the acute phase, 14 (17%; 504 participants) studies [60,61,68,74,79,92,100,102,109,114,133,144,146,148] were conducted in the early subacute phase, 11 (13%; 316 participants) studies [59,65,66,72,75,76, 81,104,107,121,134] were conducted in the late subacute phase, and 57 (69%; 1640 participants) studies [46,48,49,51,53,54,57,63,64,67,69,70,73,78,80,82,84,85,88,89,93-99,101,103,105,106,108,111-113,117-120,122-125,127-131,136-142,145,147] were conducted in the chronic phase [43].

Study Intervention

An overview of study intervention characteristics is provided (Table 1). Interventions were focused on upper limb rehabilitation in almost two-thirds (65/103, 63%) of the studies [46-49,51,54-59,62-65,68,71,72,74,75,77-81,85-88,92,95,96,99,101-103,105-108,110,112,113,116-118,121,123-125,127,128,132,133,135-137,139-142,144-147]. Nearly all interventions (96/103, 93%) [46-80,84-94,96-117,119-121,124-148] were delivered to individual participants, with over half (62/103, 60%) [46-50,53-58,60,61,64-70,72,74-77,79,80,82-86,89,90,93,94,96,97,99,101,105,111,112,116,117,119-122,126,129-131,134,136,138,139,141,143-145,147] delivered fully or partly in the participant's homes. Two-thirds (70/103, 68%) of studies [46-50,52-54,57,60,62,63,65-74,76-84,86-93,98,100,102,104,108,109,112-115,117,118,120,122-125,129-131, 135-138,140-142,144-146,148] included partial or full supervision of the intervention, with this predominately being provided face-to-face (48/70, 69%) [46,47,52,57,60,62,63,67,68,71,73,78,81-84,86-89,91,92,98,100,102,104,108,109,112-115,117,118,120,122-125,135-137,140,142,144-146,148]. Interventions lasted between a single session and 26 weeks.

Of the 103 studies, over half (n=57, 55%) of the studies [46,47,51-54,57,61,63,67,68,70,71,73,75-78,81,84-86,88-91,93,95,96,98,100,102-104,106,109,112,114,115,123-126,129,130,132,133,135-138,140,143-147] included 1 type of DHT, 30 (29%) studies [48,49,55,56,58-60,62,64,69,83,92,94,97,99,101,105,107,108,110,111,113,116,118,121,122,127,128, 139,142] included 2 types, and 16 (16%) studies [50,65,66,72,74,79, 80,82,87,117,119,120,131,134,141,148] included 3 types. VR was the most frequently used DHT (58/103, 56%) [46-49,51-53,57,59,62,63,65,66,69,71,72,74,77,78, 80,81,84-89,92,95,96,98,102-104,106,112,113,115,117-120,123-128,132,135-137,140,142,143,146-148] followed by apps (31/103, 30%) [50,55,56,58,61,64-66,72,74,75,79,82,83,94,97,99,101,105,108,111,114,116,119-122,131,134,139,141]. Further information on intervention characteristics with detail on associated citations is available (Multimedia Appendix 6 [46-148]).

XSL•FO
RenderX

Table 1. Intervention characteristics.

Intervention characteristics and details of characteristics	Studies (n=103), n (%)
Intervention focus	
Upper limb	65 (63)
Multifocus	13 (13)
Physical activity	10 (10)
Mobility	5 (5)
Activities of daily living	4 (4)
Lower limb	4 (4)
Balance	1 (1)
Speech	1 (1)
Individual or group intervention	
Individual	96 (93)
Group	2 (2)
Combination	4 (4)
Not reported	1 (1)
Intervention location	
Home	48 (47)
Home and another setting	14 (14)
Health care setting	19 (18)
Research setting	12 (12)
Not reported	10 (10)
Supervision	
Both supervised and unsupervised	38 (37)
Supervised	32 (31)
Unsupervised	23 (22)
Not reported	10 (10)
Supervision contact	
Face-to-face	48 (47)
Remote	12 (12)
Combination (F2F[a] and remote)	10 (10)
Not supervised or unreported	33 (32)
Number of DHT[b] types	
1	57 (55)
2	30 (29)
3	16 (16)
DHT type	
VR[c]	58 (56)
App	31 (30)
Sensor	17 (17)
Activity monitor	16 (16)
Audio-video platform	15 (15)
Robotics	13 (13)

XSL•FO
RenderX

Intervention characteristics and details of characteristics	Studies (n=103), n (%)
Messaging platform	11 (11)
Other	4 (4)

[a]F2F: face-to-face.

[b]DHT: digital health technology.

[c]VR: virtual reality.

Synthesis of Results

Behavior Change Theories, Models, and Frameworks

Most studies (93/103, 90%) [46-49,51-62,64-73,75-89,91-93,96-106,108-115,117-137,139, 140,142-148] endeavored to link the intervention to behavior change; however, in the majority of these studies (75/93, 81%) [46,51-56,58-62,64-69,71-73,75,77-89,91-93,96,97,99-101,103-106,108,110,112,114,115,117-120,123,124,127,128,131-137,139,140,142-144,146-148], this explanation was centered on the reporting of the techniques perceived to change behaviors without direct reference to use of the BCTTv1 or on the reporting of a component of the intervention or the whole of the intervention as motivating. These explanations lack detail on how or why this influences behavior change. Examples of this included "the app also provided performance feedback, allowing the user to compare their current performance against their score from the previous session" (Bhattacharjya et al [56]) and "games motivate patients to engage in enjoyable play behavior" (Cramer et al [66]). A limited number of studies (18/103, 17%) [47-49,57,70,76,98,102,109,111,113,121,122,125,126,129,130,145] articulated 1 or more theories, models, or frameworks of behavior change. While it is acknowledged that the BCTTv1 is a taxonomy framework rather than a theoretical framework, for the purpose of this review, it has been included as a framework for behavior change. A total of 13 different theories, models, or frameworks were identified within these 18 studies, with social cognitive theory being the most frequently reported (6/18, 33%) [76,109,111,121,129,130], followed by the behavior change technique taxonomy (4/18, 22%) [48,49, 122,129], game design theory (3/18, 17%) [47,57,125], operant conditioning (3/18, 17%) [47,98,121], and self-determination theory (3/18, 17%) [48,49,126]. Further information on behavior change theories, models, and frameworks, with details on associated citations, is available (Multimedia Appendix 7 [47-49,57,70,76,98,102,109,111,113,121,122,125,126,129,130,145]).

Behavior Change Techniques

Despite 4 studies acknowledging the BCTTv1, explicit BCTTv1 codes were only reported in 2 studies (2/103, 2%) [48,122]. However, a third study (1/103, 1%) mapped the techniques used to change behavior directly to the transtheoretical model [145]. There was a median of 3 (range 1-14) individual BCTs coded per study, with a total of 383 BCTs across the 103 studies. The most frequently identified individual BCTs were feedback on behavior and nonspecific reward (Multimedia Appendix 8).

There was also a median of 3 (range 1-8) BCTTv1 clusters per study, with a total of 288 clusters coded across the 103 studies. The most frequently used of the 16 possible clusters were feedback and monitoring (88/103, 85%) [46-60,62-69,71-74,76,78-80,82-92,94-106,108-113,116,117,119-129,134-146,148], reward and threat (56/103, 54%) [46-49,51-53,55-57,62,65,69,71,72,74,77,80,81,85,86,88,89,91,92,95,96,98,102,103,106-108,112,113,115,117-119,121-125,128,132,134-137,140,142,143,146-148], goals and planning (33/103, 32%) [49,58,60,65-68,70,72,74,76,79,80,82,83,90,91,93,94,97,100,109,111,112,121,122,126,129,130,134,138,141,145], and shaping knowledge (33/103, 32%) [46,48,50,53-56,58,60,61,64-72,74,75,86,94,97,101-103,108,111,113,114,120,129-131,139-141]. Other BCTTv1 clusters used were social support (24/103, 23%) [48,49,58,60,64,67,70,72,73,79,80,82,84,90,93,101,108,117,119,129-131,134,141], comparison of behavior (23/103, 22%) [46,50,53,54,60,61,64-66,74,75,81,86,101,104,111,114,118,122,123,125,131,139], associations (16/103, 15%) [58,60,65,66,68,75,80,83,87,90,110,120,131,133,139,144], repetition and substitution (6/103, 6%) [60,82,109, 122,129,130], scheduled consequences (3/103, 3%) [47, 80,88], natural consequences (2/103, 2%) [129,138], comparison of outcomes (2/103, 2%) [47,133], antecedents (1/103, 1%) [60], and self-belief (1/103, 1%) [70]. The clusters of regulation, identity, and covert learning were not identified. Within the context of the review, it was noted that the reward and threat cluster only included reward-based BCTs. A tabulated summary and graphical representation of the BCTTv1 clusters is available (Multimedia Appendix 9 [46-148]).

The exploration of clusters that were reported together in an intervention (Figure 3) identified the strongest relationship between the clusters of feedback and monitoring and reward and threat. Clear links were also identified between feedback and monitoring and 4 other clusters: goals and planning, shaping knowledge, social support, and comparison of behavior, and between the shaping knowledge and comparison of behavior clusters.

XSL•FO
RenderX

Figure 3. Relationships between behavior change technique taxonomy clusters.

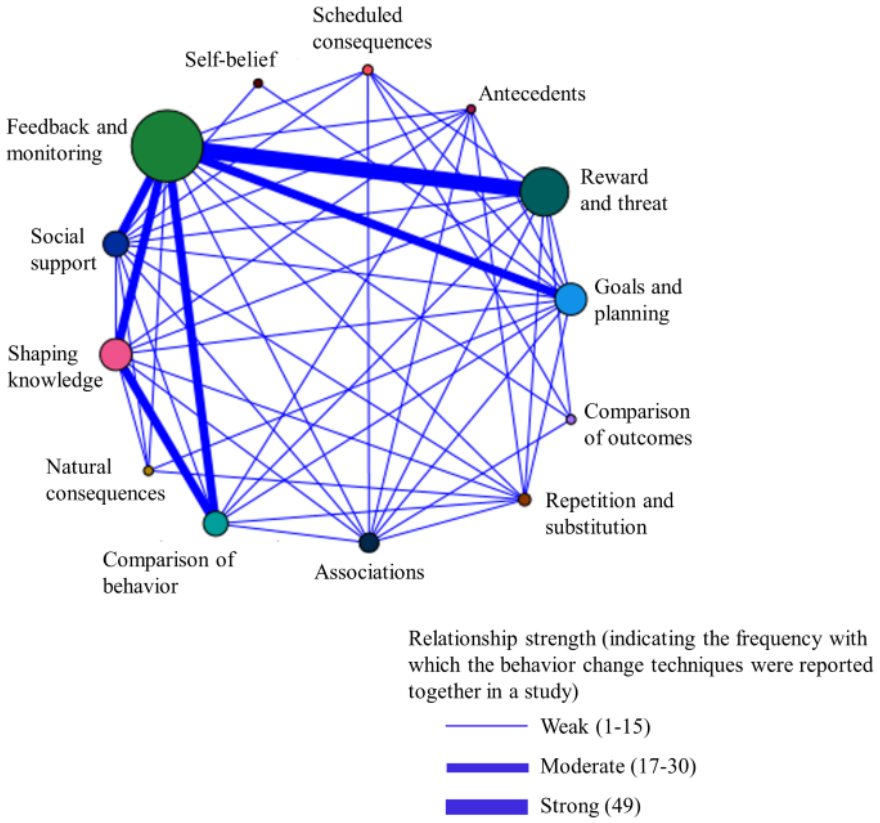

Behavior Change Techniques and Digital Health Technology

The feedback and monitoring cluster was reported most frequently for all types of DHT (Figure 4), with the greatest proportion of this cluster in robotics (11/25, 44%) [59,62,87,92,110,113,117,127, 128,142,148], VR (52/148, 35%) [46-49,51-53,57,59,62,63,65,66,69,71,72,74,78,80,84-89,92, 95,96,98,102-104,106,112,113,117,119,120,123- 135-137,140, 142,143,146,148], and sensors (17/48, 35%) [50,55,56,87,94,99, 101,105,108,110, 111,116,119-121,134,141]. Robotics and VR also often used the reward and threat cluster (9/25, 36% [62,92, 107,113,117,118,128,142,148] and 48/148, 32% [46-49,51-53, 57,62,65,69,71,72,74,77,80,81,85,86,88,89,92,95,96,98,102,103,106, 112,113,115,117-119,123-125,128,132,135-137,140,142,143,146-148], respectively), while the goals and planning cluster was a dominant second cluster in activity monitors (13/53, 25%) [67,68,76,79, 80,82,91,100, 109,122,129,138,145].

XSL•FO
RenderX

Figure 4. Percentage of behavior change technique taxonomy clusters use across the different types of digital health technology. BCTTv1: behavior change technique taxonomy version 1; DHT: digital health technology.

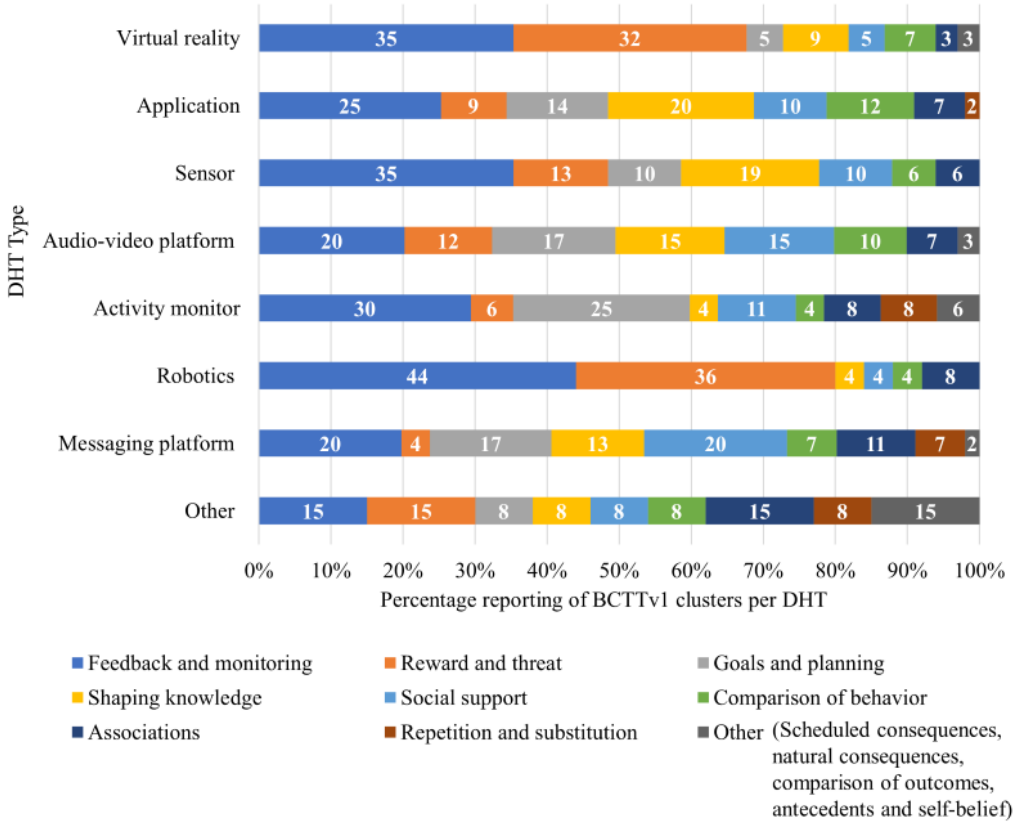

Discussion

Summary of Evidence

This scoping review provides a comprehensive overview of approaches used to support changes in behavior in DHT-based physical stroke rehabilitation interventions. Research in this field is in its infancy, with the predominance of studies in this review being described as pilot or feasibility studies with limited participants.

Despite using comprehensive behavior change search terms, only a limited number (103/1973, 6%) of screened sources were included. Over two-thirds of full-text sources were excluded as they did not describe or refer to any behavior change theories, models, or frameworks or BCTs, suggesting that in general, explicit behavior change approaches are not reported as being integral to DHT-based physical stroke rehabilitation.

Only 18 (17%) of the 103 included studies articulated a theory, model, or framework to underpin the intervention, which aimed to change behavior, despite widely published recommendations about the importance of overt use of theory when developing, evaluating, and reporting interventions [27,29], including those related to behavior change [28]. The proportion of studies articulating a behavior change theory, model, or framework in this work is significantly lower than review findings in non-rehabilitation DHT-based interventions that have sought to influence specific behaviors such as physical activity or weight control [24,44]. These reviews have identified up to two-thirds of sources reporting a theory, model, or framework. However, our findings mirror the relative absence of behavior change theories, models, and frameworks in rehabilitation interventions more generally, irrespective of whether they use digital technology [149] or not [45], and it is widely recognized that the complex nature of rehabilitation often results in the essential characteristics of interventions being poorly defined [150]. Consistent with our findings in these other reviews, a variety of theories, models, and frameworks were found to underpin interventions, with social cognitive theory being the most frequently reported [24,44,45,149]. The explicit description of BCTs using the BCTTv1 within DHT-based physical stroke rehabilitation interventions is also poorly reported (2%), despite a significant proportion of the sources being dated after the publication of the BCTTv1 in 2013 [22]. This lack of acknowledgment of behavior change approaches impedes the accumulation of knowledge within this field.

It is important that both the underpinning theory and BCTs are reported so the mechanisms by which the BCTs elicit change can be better understood [21]. The general assumption that the motivational and captivating aspects of DHTs will promote prolonged and repeated engagement with rehabilitative activities, in particular in those DHTs that incorporate game design [151], risks suboptimal outcomes for patients and wasted investment of time and money if the mechanisms by which the DHT elicits change are not considered.

When exploring which BCT clusters featured within the reviewed DHT-based interventions, the findings relating to the

XSL•FO
RenderX

commonly used clusters of feedback and monitoring, goals and planning, and shaping knowledge are consistent with findings from DHT-based interventions to change a specific behavior [44] and non-DHT–based rehabilitation [45]. However, a novel finding in our review was the frequent identification of the reward and threat cluster, although it was noted that all techniques related to reward and none to threat. A large number of studies in this review used VR technology, which frequently incorporates gamified tasks or gameplay. Reward is an integral part of game design theory alongside feedback [152], and so it is perhaps unsurprising that the feedback and monitoring, and reward and threat clusters dominated and an association between these 2 clusters was seen.

Limitations

Rehabilitation is a process that comprises multiple behaviors and so exploring approaches to change behavior within this context was complicated. There were challenges in searching and screening sources for inclusion as few studies explicitly reported approaches to change behavior, and there is a similarity in the vocabulary used within behavior change and other theoretical approaches (eg, "feedback," which is used within motor learning). Similarly, only a very small proportion of studies explicitly reported BCTs within interventions. The lack of clear reporting of behavior change introduces the risk that sources may be omitted during both the searching and screening process highlighting the difficulty of comprehensively reviewing this field of work. An inclusive approach to screening reduced the risk of erroneously excluding sources, but it is perhaps inevitable that the sources included reflect those studies that have reported a behavior change approach rather than all studies that have used one.

This lack of clear BCT reporting also posed challenges for intervention coding. The use of the BCTTv1 aimed to ensure the review used a generalizable nomenclature to describe BCTs, and the 1 reviewer who had completed BCTTv1 training coded all the interventions. It is acknowledged that decisions made in the application of the BCTTv1 within the context of the review will have introduced some subjectivity in intervention coding, which will ultimately influence the review findings. Although the coding process could have been made more robust by having a second reviewer trained in the BCTTv1 also code the interventions, regular and extensive discussions between all members of the review team took place with the aim of ensuring consistency with the coding process. Clear documentation as to how the BCTTv1 was used within this review (Multimedia Appendix 5) supports transparency as to the decisions made and the reproducibility of the review.

The absence of a recognized predefined taxonomy for DHTs posed a challenge when categorizing the DHT interventions, acknowledging that the distinction between the categories used to present the results is open to interpretation. A description of how the reviewers interpreted these categories is provided (Multimedia Appendix 4).

Implications for Research

Future studies aimed at developing and evaluating DHT-based rehabilitation interventions, including those relating to physical stroke rehabilitation, need to ensure there is explicit recognition and reporting of the specific approaches used to change behavior, articulating both the theory on which the intervention is based and how the intervention plans to deliver the change in behavior using universally recognized terminology. This should be reported as part of a program theory and potential mechanisms of action, which are key parts of developing and evaluating complex interventions [29]. This detailed reporting would further support an understanding of how changes in behavior could be best enabled by DHT-based rehabilitation interventions and how this contributes to changes in patient outcomes. It would also enable further evaluation of the optimal behavioral components of interventions, enabling patients to use and clinicians to deliver the most effective DHT-based rehabilitative interventions. More generally, as the use of DHTs expands, there is an urgent need for some form of taxonomy to categorize and clearly define the different types of DHTs to facilitate consistent reporting, replication, and comparison of DHT-based interventions.

Conclusion

This novel and original review is the first to explore if and how approaches to change behavior are incorporated within DHT-based physical stroke rehabilitation. It demonstrates that a minority of studies report using approaches to change behavior within this context, despite these changes in behavior being vital to meet the demands of rehabilitative activities. Those who do report behavior change often lack the underpinning detail as to how the DHT-based intervention will facilitate these changes. In order for DHT-based interventions to realize their potential within rehabilitation and their impact on patient outcomes, approaches to change behavior must be embedded in the intervention and appropriately reported.

Acknowledgments

The authors would like to thank Catherine Harris (Information Specialist, University of Central Lancashire) for her assistance in developing the search strategy and running the searches, and Rebekah Murray (Undergraduate Research Intern, University of Central Lancashire) for her support with aspects of the screening and data charting process.

This work was funded by a UK Research and Innovation Future Leaders Fellowship (grant MR/T022434/1).

Data Availability

All data supporting this study are openly available from the University of Central Lancashire repository [153].

XSL•FO
RenderX

Authors' Contributions

RCS conceived the review focus and oversaw the work. HJG developed the review design and search strategy. HJG, KAJ, and RCS completed the screening of the identified sources. HJG and KAJ piloted the data charting tool. HJG completed the data charting, data analysis, and the initial manuscript draft. KAJ and RCS reviewed and made substantial contributions to the manuscript. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting of Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) checklist.
[PDF File (Adobe PDF File), 47 KB - jmir_v26i1e48725_app1.pdf]

Multimedia Appendix 2
Inclusion and exclusion criteria.
[PDF File (Adobe PDF File), 39 KB - jmir_v26i1e48725_app2.pdf]

Multimedia Appendix 3
Full search strategy as used in Medline.
[PDF File (Adobe PDF File), 63 KB - jmir_v26i1e48725_app3.pdf]

Multimedia Appendix 4
Data charting tool.
[PDF File (Adobe PDF File), 149 KB - jmir_v26i1e48725_app4.pdf]

Multimedia Appendix 5
Review-specific behavior change technique taxonomy coding decisions.
[PDF File (Adobe PDF File), 96 KB - jmir_v26i1e48725_app5.pdf]

Multimedia Appendix 6
Intervention characteristics (with associated references).
[PDF File (Adobe PDF File), 42 KB - jmir_v26i1e48725_app6.pdf]

Multimedia Appendix 7
Behavior change theories, models, and frameworks reported (with associated references).
[PDF File (Adobe PDF File), 33 KB - jmir_v26i1e48725_app7.pdf]

Multimedia Appendix 8
Individual behavior change techniques coded.
[PDF File (Adobe PDF File), 50 KB - jmir_v26i1e48725_app8.pdf]

Multimedia Appendix 9
Behavior change technique taxonomy clusters identified (with associated references).
[PDF File (Adobe PDF File), 133 KB - jmir_v26i1e48725_app9.pdf]

References

1. Evidence standards framework for digital health technologies. National Institute of Clinical Excellence (NICE). 2018. URL: https://tinyurl.com/mpmrwrwx [accessed 2022-01-25]
2. Castle-Clark S. The NHS at 70: What will new technology mean for the NHS and its patients? Kings Fund. 2018. URL: https://tinyurl.com/44k9fmat [accessed 2024-04-05]
3. Topel E. The Topol review: preparing the healthcare workforce to deliver the digital future. Health Education England, NHS. 2019. URL: https://topol.hee.nhs.uk/ [accessed 2022-01-19]
4. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021;20(10):795-820 [FREE Full text] [doi: 10.1016/S1474-4422(21)00252-0] [Medline: 34487721]

XSL•FO
RenderX

5. Wade DT. What is rehabilitation? An empirical investigation leading to an evidence-based description. Clin Rehabil 2020;34(5):571-583 [FREE Full text] [doi: 10.1177/0269215520905112] [Medline: 32037876]
6. Rehabilitation. World Health Organisation. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/rehabilitation [accessed 2023-05-04]
7. Grefkes C, Fink GR. Recovery from stroke: current concepts and future perspectives. Neurol Res Pract 2020;2:17 [FREE Full text] [doi: 10.1186/s42466-020-00060-6] [Medline: 33324923]
8. Lawrence ES, Coshall C, Dundas R, Stewart J, Rudd AG, Howard R, et al. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. Stroke 2001;32(6):1279-1284 [FREE Full text] [doi: 10.1161/01.str.32.6.1279] [Medline: 11387487]
9. National clinical guideline for stroke for the United Kingdom and Ireland. Intercollegiate Stroke Working Party. 2023. URL: https://tinyurl.com/yxy6b9tp [accessed 2023-05-02]
10. Veerbeek JM, van Wegen E, van Peppen R, van der Wees PJ, Hendriks E, Rietberg M, et al. What is the evidence for physical therapy poststroke? a systematic review and meta-analysis. PLoS One 2014;9(2):e87987 [FREE Full text] [doi: 10.1371/journal.pone.0087987] [Medline: 24505342]
11. Lang CE, Macdonald JR, Reisman DS, Boyd L, Kimberley TJ, Schindler-Ivens SM, et al. Observation of amounts of movement practice provided during stroke rehabilitation. Arch Phys Med Rehabil 2009;90(10):1692-1698 [FREE Full text] [doi: 10.1016/j.apmr.2009.04.005] [Medline: 19801058]
12. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2017;11(11):CD008349 [FREE Full text] [doi: 10.1002/14651858.CD008349.pub4] [Medline: 29156493]
13. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020;1(1):CD010255 [FREE Full text] [doi: 10.1002/14651858.CD010255.pub3] [Medline: 32002991]
14. Lynch EA, Jones TM, Simpson DB, Fini NA, Kuys SS, Borschmann K, et al. Activity monitors for increasing physical activity in adult stroke survivors. Cochrane Database Syst Rev 2018;7(7):CD012543 [FREE Full text] [doi: 10.1002/14651858.CD012543.pub2] [Medline: 30051462]
15. Mehrholz J, Pohl M, Platz T, Kugler J, Elsner B. Electromechanical and robot-assisted arm training for improving activities of daily living, arm function, and arm muscle strength after stroke. Cochrane Database Syst Rev 2018;9(9):CD006876 [FREE Full text] [doi: 10.1002/14651858.CD006876.pub5] [Medline: 30175845]
16. Mehrholz J, Thomas S, Kugler J, Pohl M, Elsner B. Electromechanical-assisted training for walking after stroke. Cochrane Database Syst Rev 2020;10(10):CD006185 [FREE Full text] [doi: 10.1002/14651858.CD006185.pub5] [Medline: 33091160]
17. Jones F, Riazi A. Self-efficacy and self-management after stroke: a systematic review. Disabil Rehabil 2011;33(10):797-810. [doi: 10.3109/09638288.2010.511415] [Medline: 20795919]
18. Behaviour change: general approaches. National Institute of Clinical Excellence. UK: National Institute of Clinical Excellence (NICE); 2007. URL: https://tinyurl.com/yfthycch [accessed 2024-04-05]
19. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs, NJ: Prentice-Hall; 1986.
20. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
21. Carey RN, Connell LE, Johnston M, Rothman AJ, de Bruin M, Kelly MP, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2019;53(8):693-707 [FREE Full text] [doi: 10.1093/abm/kay078] [Medline: 30304386]
22. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
23. Oinas-Kukkonen H. A foundation for the study of behavior change support systems. Pers Ubiquit Comput 2012;17(6):1223-1235. [doi: 10.1007/s00779-012-0591-5]
24. Taj F, Klein MCA, van Halteren A. Digital health behavior change technology: bibliometric and scoping review of two decades of research. JMIR Mhealth Uhealth 2019;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
25. West R, Michie S. A Guide to Development and Evaluation of Digital Behaviour Change Interventions in Healthcare. London: Silverback Publishing; 2016.
26. Lewis GN, Rosie JA. Virtual reality games for movement rehabilitation in neurological conditions: how do we meet the needs and expectations of the users? Disabil Rehabil 2012;34(22):1880-1886. [doi: 10.3109/09638288.2012.670036] [Medline: 22480353]
27. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
28. Michie S, Prestwich A. Are interventions theory-based? development of a theory coding scheme. Health Psychol 2010;29(1):1-8 [FREE Full text] [doi: 10.1037/a0016939] [Medline: 20063930]

XSL•FO
RenderX

29. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
30. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146 [FREE Full text] [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
31. Peters MDJ, Godfrey CM, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping Reviews (2020 version). In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: JBI; 2020. [doi: 10.46658/JBIMES-20-12]
32. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
33. Gooch H, Stockley R, Jarvis K. Behaviour change approaches within digital health technology-based interventions in physical rehabilitation following stroke: a scoping review protocol Version 2 081122. OSF Registries. 2022. URL: https://osf.io/yjn5g [accessed 2022-11-15]
34. ICF checklist. World Health Organisation. 2003. URL: https://tinyurl.com/5fz6krew [accessed 2022-01-19]
35. Hillyer M. How has technology changed—and changed us—in the past 20 years? World Economic Forum. 2020. URL: https://tinyurl.com/5n74wsym [accessed 2023-10-18]
36. Online training. BCT Taxonomy v1. 2022. URL: http://www.bct-taxonomy.com/ [accessed 2022-03-08]
37. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
38. Wadley M, Bradbury M, Stockley R. Virtual reality. The Chartered Society of Physiotherapy. URL: https://www.csp.org.uk/professional-clinical/digital-physiotherapy/virtual-reality [accessed 2024-03-29]
39. Cambridge dictionary. Cambridge University Press and Assessment. URL: https://tinyurl.com/2f2mx6kz [accessed 2024-03-29]
40. Weber LM, Stein J. The use of robots in stroke rehabilitation: A narrative review. NeuroRehabilitation 2018;43(1):99-110. [doi: 10.3233/NRE-172408] [Medline: 30056437]
41. English C, Ceravolo MG, Dorsch S, Drummond A, Gandhi DB, Green JH, et al. Telehealth for rehabilitation and recovery after stroke: state of the evidence and future directions. Int J Stroke 2022;17(5):487-493. [doi: 10.1177/17474930211062480] [Medline: 34983266]
42. Taub E, Crago JE, Burgio LD, Groomes TE, Cook EW, DeLuca SC, et al. An operant approach to rehabilitation medicine: overcoming learned nonuse by shaping. J Exp Anal Behav 1994;61(2):281-293 [FREE Full text] [doi: 10.1901/jeab.1994.61-281] [Medline: 8169577]
43. Bernhardt J, Hayward KS, Kwakkel G, Ward NS, Wolf SL, Borschmann K, et al. Agreed definitions and a shared vision for new standards in stroke recovery research: the Stroke Recovery and Rehabilitation Roundtable taskforce. Int J Stroke 2017;12(5):444-450 [FREE Full text] [doi: 10.1177/1747493017711816] [Medline: 28697708]
44. Asbjørnsen RA, Smedsrød ML, Solberg Nes L, Wentzel J, Varsi C, Hjelmesæth J, et al. Persuasive system design principles and behavior change techniques to stimulate motivation and adherence in electronic health interventions to support weight loss maintenance: scoping review. J Med Internet Res 2019;21(6):e14265 [FREE Full text] [doi: 10.2196/14265] [Medline: 31228174]
45. Bayly J, Wakefield D, Hepgul N, Wilcock A, Higginson IJ, Maddocks M. Changing health behaviour with rehabilitation in thoracic cancer: a systematic review and synthesis. Psychooncology 2018;27(7):1675-1694. [doi: 10.1002/pon.4684] [Medline: 29476566]
46. Alankus G, Proffitt R, Kelleher C, Engsberg J. Stroke therapy through motion-based games: a case study. ACM Trans Access Comput 2011;4(1):1-35. [doi: 10.1145/2039339.2039342]
47. Alankus G, Kelleher C. Reducing compensatory motions in motion-based video games for stroke rehabilitation. Hum Comput Interact 2015;30(3-4):232-262. [doi: 10.1080/07370024.2014.985826]
48. Allegue DR, Kairy D, Higgins J, Archambault PS, Michaud F, Miller W, et al. A personalized home-based rehabilitation program using exergames combined with a telerehabilitation app in a chronic stroke survivor: mixed methods case study. JMIR Serious Games 2021;9(3):e26153 [FREE Full text] [doi: 10.2196/26153] [Medline: 34132649]
49. Allegue DR, Higgins J, Sweet SN, Archambault PS, Michaud F, Miller W, et al. Rehabilitation of upper extremity by telerehabilitation combined with exergames in survivors of chronic stroke: preliminary findings from a feasibility clinical trial. JMIR Rehabil Assist Technol 2022;9(2):e33745 [FREE Full text] [doi: 10.2196/33745] [Medline: 35731560]
50. Asano M, Tai BC, Yeo FY, Yen SC, Tay A, Ng YS, et al. Home-based tele-rehabilitation presents comparable positive impact on self-reported functional outcomes as usual care: the Singapore Tele-technology Aided Rehabilitation in Stroke (STARS) randomised controlled trial. J Telemed Telecare 2021;27(4):231-238. [doi: 10.1177/1357633X19868905] [Medline: 31462136]
51. Ballester BR, Maier M, Mozo RMSS, Castañeda V, Duff A, Verschure PFMJ. Counteracting learned non-use in chronic stroke patients with reinforcement-induced movement therapy. J Neuroeng Rehabil 2016;13(1):74 [FREE Full text] [doi: 10.1186/s12984-016-0178-x] [Medline: 27506203]

XSL•FO
RenderX

52. Bardack A, Bhandari P, Doggett J, Epstein M, Gagliolo N, Graff S, et al. EMG biofeedback videogame system for the gait rehabilitation of hemiparetic individuals. Digital Repository at the University of Maryland. 2010. URL: https://drum.lib.umd.edu/bitstream/handle/1903/10082/CHIP.pdf?sequence=1&isAllowed=y [accessed 2023-02-27]
53. Bellomo RG, Paolucci T, Saggino A, Pezzi L, Bramanti A, Cimino V, et al. The WeReha Project for an innovative home-based exercise training in chronic stroke patients: a clinical study. J Cent Nerv Syst Dis 2020;12:1179573520979866 [FREE Full text] [doi: 10.1177/1179573520979866] [Medline: 33402861]
54. Benvenuti F, Stuart M, Cappena V, Gabella S, Corsi S, Taviani A, et al. Community-based exercise for upper limb paresis: a controlled trial with telerehabilitation. Neurorehabil Neural Repair 2014;28(7):611-620 [FREE Full text] [doi: 10.1177/1545968314521003] [Medline: 24515928]
55. Bhattacharjya S, Stafford MC, Cavuoto LA, Yang Z, Song C, Subryan H, et al. Harnessing smartphone technology and three dimensional printing to create a mobile rehabilitation system, mRehab: assessment of usability and consistency in measurement. J Neuroeng Rehabil 2019;16(1):127 [FREE Full text] [doi: 10.1186/s12984-019-0592-y] [Medline: 31665036]
56. Bhattacharjya S, Cavuoto LA, Reilly B, Xu W, Subryan H, Langan J. Usability, usefulness, and acceptance of a novel, portable rehabilitation system (mRehab) using smartphone and 3D printing technology: mixed methods study. JMIR Hum Factors 2021;8(1):e21312 [FREE Full text] [doi: 10.2196/21312] [Medline: 33749608]
57. Burke JW, McNeill MDJ, Charles DK, Morrow PJ, Crosbie JH, McDonough SM. Optimising engagement for stroke rehabilitation using serious games. Vis Comput 2009;25(12):1085-1099. [doi: 10.1007/s00371-009-0387-4]
58. Burridge JH, Lee ACW, Turk R, Stokes M, Whitall J, Vaidyanathan R, et al. Telehealth, wearable sensors, and the internet: will they improve stroke outcomes through increased intensity of therapy, motivation, and adherence to rehabilitation programs? J Neurol Phys Ther 2017;41(Suppl 3):S32-S38. [doi: 10.1097/NPT.0000000000000183] [Medline: 28628594]
59. Cai S, Wei X, Su E, Wu W, Zheng H, Xie L. Online compensation detecting for real-time reduction of compensatory motions during reaching: a pilot study with stroke survivors. J Neuroeng Rehabil 2020;17(1):58 [FREE Full text] [doi: 10.1186/s12984-020-00687-1] [Medline: 32345335]
60. Chumbler NR, Quigley P, Li X, Morey M, Rose D, Sanford J, et al. Effects of telerehabilitation on physical function and disability for stroke patients: a randomized, controlled trial. Stroke 2012;43(8):2168-2174 [FREE Full text] [doi: 10.1161/STROKEAHA.111.646943] [Medline: 22627983]
61. Chung BPH, Chiang WKH, Lau H, Lau TFO, Lai CWK, Sit CSY, et al. Pilot study on comparisons between the effectiveness of mobile video-guided and paper-based home exercise programs on improving exercise adherence, self-efficacy for exercise and functional outcomes of patients with stroke with 3-month follow-up: a single-blind randomized controlled trial. Hong Kong Physiother J 2020;40(1):63-73 [FREE Full text] [doi: 10.1142/S1013702520500079] [Medline: 32489241]
62. Colombo R, Pisano F, Mazzone A, Delconte C, Micera S, Carrozza MC, et al. Design strategies to improve patient motivation during robot-aided rehabilitation. J Neuroeng Rehabil 2007;4:3 [FREE Full text] [doi: 10.1186/1743-0003-4-3] [Medline: 17309790]
63. Colomer C, Llorens R, Noé E, Alcañiz M. Effect of a mixed reality-based intervention on arm, hand, and finger function on chronic stroke. J Neuroeng Rehabil 2016;13(1):45 [FREE Full text] [doi: 10.1186/s12984-016-0153-6] [Medline: 27169462]
64. Conroy SS, Harcum S, Keldsen L, Bever CT. Novel use of existing technology: a preliminary study of patient portal use for telerehabilitation. J Telemed Telecare 2022;28(5):380-388. [doi: 10.1177/1357633X20950172] [Medline: 32869689]
65. Cramer SC, Dodakian L, Le V, See J, Augsburger R, McKenzie A, et al. Efficacy of home-based telerehabilitation vs in-clinic therapy for adults after stroke: a randomized clinical trial. JAMA Neurol 2019;76(9):1079-1087 [FREE Full text] [doi: 10.1001/jamaneurol.2019.1604] [Medline: 31233135]
66. Cramer SC, Dodakian L, Le V, McKenzie A, See J, Augsburger R, et al. A feasibility study of expanded home-based telerehabilitation after stroke. Front Neurol 2020;11:611453 [FREE Full text] [doi: 10.3389/fneur.2020.611453] [Medline: 33613417]
67. Danks KA, Roos MA, McCoy D, Reisman DS. A step activity monitoring program improves real world walking activity post stroke. Disabil Rehabil 2014;36(26):2233-2236 [FREE Full text] [doi: 10.3109/09638288.2014.903303] [Medline: 24670193]
68. Da-Silva RH, Moore SA, Rodgers H, Shaw L, Sutcliffe L, van Wijck F, et al. Wristband accelerometers to motivate arm exercises after Stroke (WAVES): a pilot randomized controlled trial. Clin Rehabil 2019;33(8):1391-1403. [doi: 10.1177/0269215519834720] [Medline: 30845829]
69. Deng H, Durfee WK, Nuckley DJ, Rheude BS, Severson AE, Skluzacek KM, et al. Complex versus simple ankle movement training in stroke using telerehabilitation: a randomized controlled trial. Phys Ther 2012;92(2):197-209 [FREE Full text] [doi: 10.2522/ptj.20110018] [Medline: 22095209]
70. Deutsch JE, Maidan I, Dickstein R. Patient-centered integrated motor imagery delivered in the home with telerehabilitation to improve walking after stroke. Phys Ther 2012;92(8):1065-1077 [FREE Full text] [doi: 10.2522/ptj.20110277] [Medline: 22499891]
71. Dias P, Silva R, Amorim P, Lains J, Roque E, Pereira ISF, et al. Using virtual reality to increase motivation in poststroke rehabilitation. IEEE Comput Graph Appl 2019;39(1):64-70. [doi: 10.1109/MCG.2018.2875630] [Medline: 30869599]

XSL•FO
RenderX

72. Dodakian L, McKenzie AL, Le V, See J, Pearson-Fuhrhop K, Quinlan EB, et al. A home-based telerehabilitation program for patients with stroke. Neurorehabil Neural Repair 2017;31(10-11):923-933 [FREE Full text] [doi: 10.1177/1545968317733818] [Medline: 29072556]
73. Dorsch AK, Thomas S, Xu X, Kaiser W, Dobkin BH, SIRRACT investigators. SIRRACT: an international randomized clinical trial of activity feedback during inpatient stroke rehabilitation enabled by wireless sensing. Neurorehabil Neural Repair 2015;29(5):407-415 [FREE Full text] [doi: 10.1177/1545968314550369] [Medline: 25261154]
74. Edwards D, Kumar S, Brinkman L, Ferreira IC, Esquenazi A, Nguyen T, et al. Telerehabilitation initiated early in post-stroke recovery: a feasibility study. Neurorehabil Neural Repair 2023;37(2-3):131-141 [FREE Full text] [doi: 10.1177/15459683231159660] [Medline: 36876946]
75. Emmerson KB, Harding KE, Taylor NF. Home exercise programmes supported by video and automated reminders compared with standard paper-based home exercise programmes in patients with stroke: a randomized controlled trial. Clin Rehabil 2017;31(8):1068-1077. [doi: 10.1177/0269215516680856] [Medline: 27920262]
76. Ezeugwu VE, Manns PJ. The feasibility and longitudinal effects of a home-based sedentary behavior change intervention after stroke. Arch Phys Med Rehabil 2018;99(12):2540-2547. [doi: 10.1016/j.apmr.2018.06.014] [Medline: 29981314]
77. Fluet GG, Qiu Q, Patel J, Cronce A, Merians AS, Adamovich SV. Autonomous use of the home virtual rehabilitation system: a feasibility and pilot study. Games Health J 2019;8(6):432-438 [FREE Full text] [doi: 10.1089/g4h.2019.0012] [Medline: 31769724]
78. Foreman MH, Engsberg JR. A virtual reality tool for measuring and shaping trunk compensation for persons with stroke: design and initial feasibility testing. J Rehabil Assist Technol Eng 2019;6:2055668318823673 [FREE Full text] [doi: 10.1177/2055668318823673] [Medline: 31245028]
79. Fusari G, Gibbs E, Hoskin L, Lawrence-Jones A, Dickens D, Crespo RF, et al. What is the feasibility and patient acceptability of a digital system for arm and hand rehabilitation after stroke? a mixed-methods, single-arm feasibility study of the 'OnTrack' intervention for hospital and home use. BMJ Open 2022;12(9):e062042 [FREE Full text] [doi: 10.1136/bmjopen-2022-062042] [Medline: 36171046]
80. Gauthier LV, Nichols-Larsen DS, Uswatte G, Strahl N, Simeo M, Proffitt R, et al. Video game rehabilitation for outpatient stroke (VIGoROUS): a multi-site randomized controlled trial of in-home, self-managed, upper-extremity therapy. EClinicalMedicine 2022;43:101239 [FREE Full text] [doi: 10.1016/j.eclinm.2021.101239] [Medline: 34977516]
81. Goršič M, Cikajlo I, Goljar N, Novak D. A multisession evaluation of an adaptive competitive arm rehabilitation game. J Neuroeng Rehabil 2017;14(1):128 [FREE Full text] [doi: 10.1186/s12984-017-0336-9] [Medline: 29208017]
82. Grau-Pellicer M, Lalanza JF, Jovell-Fernández E, Capdevila L. Impact of mHealth technology on adherence to healthy PA after stroke: a randomized study. Top Stroke Rehabil 2020;27(5):354-368. [doi: 10.1080/10749357.2019.1691816] [Medline: 31790639]
83. Guidetti S, Gustavsson M, Tham K, Andersson M, Fors U, Ytterberg C. F@ce: a team-based, person-centred intervention for rehabilitation after stroke supported by information and communication technology—a feasibility study. BMC Neurol 2020;20(1):387 [FREE Full text] [doi: 10.1186/s12883-020-01968-x] [Medline: 33096984]
84. Held JP, Ferrer B, Mainetti R, Steblin A, Hertler B, Moreno-Conde A, et al. Autonomous rehabilitation at stroke patients home for balance and gait: safety, usability and compliance of a virtual reality system. Eur J Phys Rehabil Med 2018;54(4):545-553 [FREE Full text] [doi: 10.23736/S1973-9087.17.04802-X] [Medline: 28949120]
85. Hernandez A, Bubyr L, Archambault PS, Higgins J, Levin MF, Kairy D. Virtual reality-based rehabilitation as a feasible and engaging tool for the management of chronic poststroke upper-extremity function recovery: randomized controlled trial. JMIR Serious Games 2022;10(3):e37506 [FREE Full text] [doi: 10.2196/37506] [Medline: 36166289]
86. Hung NT, Paul V, Prakash P, Kovach T, Tacy G, Tomic G, et al. Wearable myoelectric interface enables high-dose, home-based training in severely impaired chronic stroke survivors. Ann Clin Transl Neurol 2021;8(9):1895-1905 [FREE Full text] [doi: 10.1002/acn3.51442] [Medline: 34415114]
87. Johnson MJ, Shakya Y, Strachota E, Ahamed SI. Low-cost monitoring of patients during unsupervised robot/computer assisted motivating stroke rehabilitation. Biomed Tech (Berl) 2011;56(1):5-9. [doi: 10.1515/BMT.2010.050] [Medline: 21117891]
88. Johnson MJ, Van der Loos HFM, Burgar CG, Shor P, Leifer LJ. Experimental results using force-feedback cueing in robot-assisted stroke therapy. IEEE Trans Neural Syst Rehabil Eng 2005;13(3):335-348. [doi: 10.1109/TNSRE.2005.850428] [Medline: 16200757]
89. Jonsdottir J, Baglio F, Gindri P, Isernia S, Castiglioni C, Gramigna C, et al. Virtual reality for motor and cognitive rehabilitation from clinic to home: a pilot feasibility and efficacy study for persons with chronic stroke. Front Neurol 2021;12:601131 [FREE Full text] [doi: 10.3389/fneur.2021.601131] [Medline: 33897579]
90. Kamwesiga JT, Eriksson GM, Tham K, Fors U, Ndiwalana A, von Koch L, et al. A feasibility study of a mobile phone supported family-centred ADL intervention, F@ce, after stroke in Uganda. Global Health 2018;14(1):82 [FREE Full text] [doi: 10.1186/s12992-018-0400-7] [Medline: 30111333]
91. Kanai M, Izawa KP, Kobayashi M, Onishi A, Kubo H, Nozoe M, et al. Effect of accelerometer-based feedback on physical activity in hospitalized patients with ischemic stroke: a randomized controlled trial. Clin Rehabil 2018;32(8):1047-1056. [doi: 10.1177/0269215518755841] [Medline: 29400070]

XSL•FO
RenderX

92. Keeling AB, Piitz M, Semrau JA, Hill MD, Scott SH, Dukelow SP. Robot Enhanced Stroke Therapy Optimizes Rehabilitation (RESTORE): a pilot study. J Neuroeng Rehabil 2021;18(1):10 [FREE Full text] [doi: 10.1186/s12984-021-00804-8] [Medline: 33478563]
93. Kessler D, Anderson ND, Dawson DR. Occupational performance coaching for stroke survivors delivered via telerehabilitation using a single-case experimental design. Br J Occup Ther 2021;84(8):488-496 [FREE Full text] [doi: 10.1177/0308022620988471] [Medline: 34381289]
94. Kim DY, Kwon H, Nam KW, Lee Y, Kwon HM, Chung YS. Remote management of poststroke patients with a smartphone-based management system integrated in clinical care: prospective, nonrandomized, interventional study. J Med Internet Res 2020;22(2):e15377 [FREE Full text] [doi: 10.2196/15377] [Medline: 32130140]
95. Kim J, Lee M, Kim Y, Eun SD, Yoon B. Feasibility of an individually tailored virtual reality program for improving upper motor functions and activities of daily living in chronic stroke survivors: a case series. Eur J Integr Med 2016;8(5):731-737. [doi: 10.1016/j.eujim.2016.05.001]
96. King M, Hijmans JM, Sampson M, Satherley J, Hale L. Home-based stroke rehabilitation using computer gaming. N Z J Physiother 2012;40(3):128-134 [FREE Full text]
97. Kringle EA, Setiawan IMA, Golias K, Parmanto B, Skidmore ER. Feasibility of an iterative rehabilitation intervention for stroke delivered remotely using mobile health technology. Disabil Rehabil Assist Technol 2020;15(8):908-916 [FREE Full text] [doi: 10.1080/17483107.2019.1629113] [Medline: 31216917]
98. Kumar D, Sinha N, Dutta A, Lahiri U. Virtual reality-based balance training system augmented with operant conditioning paradigm. Biomed Eng Online 2019;18(1):90 [FREE Full text] [doi: 10.1186/s12938-019-0709-3] [Medline: 31455355]
99. Langan J, Bhattacharjya S, Subryan H, Xu W, Chen B, Li Z, et al. In-home rehabilitation using a smartphone app coupled with 3D printed functional objects: single-subject design study. JMIR Mhealth Uhealth 2020;8(7):e19582 [FREE Full text] [doi: 10.2196/19582] [Medline: 32706702]
100. Lawrie S, Dong Y, Steins D, Xia Z, Esser P, Sun S, et al. Evaluation of a smartwatch-based intervention providing feedback of daily activity within a research-naive stroke ward: a pilot randomised controlled trial. Pilot Feasibility Stud 2018;4:157 [FREE Full text] [doi: 10.1186/s40814-018-0345-x] [Medline: 30323946]
101. Lawson S, Tang Z, Feng J. Supporting stroke motor recovery through a mobile application: a pilot study. Am J Occup Ther 2017;71(3):7103350010p1-7103350010p5. [doi: 10.5014/ajot.2017.025023] [Medline: 28422641]
102. Lee M, Pyun SB, Chung J, Kim J, Eun SD, Yoon BC. A further step to develop patient-friendly implementation strategies for virtual reality-based rehabilitation in patients with acute stroke. Phys Ther 2016;96(10):1554-1564 [FREE Full text] [doi: 10.2522/ptj.20150271] [Medline: 27149961]
103. Lee M, Son J, Kim J, Pyun SB, Eun SD, Yoon BC. Comparison of individualized virtual reality- and group-based rehabilitation in older adults with chronic stroke in community settings: a pilot randomized controlled trial. Eur J Integr Med 2016;8(5):738-746. [doi: 10.1016/j.eujim.2016.08.166]
104. Lee MM, Shin DC, Song CH. Canoe game-based virtual reality training to improve trunk postural stability, balance, and upper limb motor function in subacute stroke patients: a randomized controlled pilot study. J Phys Ther Sci 2016;28(7):2019-2024 [FREE Full text] [doi: 10.1589/jpts.28.2019] [Medline: 27512255]
105. Levy T, Crotty M, Laver K, Lannin N, Killington M. Does the addition of concurrent visual feedback increase adherence to a home exercise program in people with stroke: a single-case series? BMC Res Notes 2020;13(1):361 [FREE Full text] [doi: 10.1186/s13104-020-05202-2] [Medline: 32727575]
106. Lewis GN, Woods C, Rosie JA, McPherson KM. Virtual reality games for rehabilitation of people with stroke: perspectives from the users. Disabil Rehabil Assist Technol 2011;6(5):453-463. [doi: 10.3109/17483107.2011.574310] [Medline: 21495917]
107. Li X, Wang L, Miao S, Yue Z, Tang Z, Su L, et al. Sensorimotor rhythm-brain computer interface with audio-cue, motor observation and multisensory feedback for upper-limb stroke rehabilitation: a controlled study. Front Neurosci 2022;16:808830 [FREE Full text] [doi: 10.3389/fnins.2022.808830] [Medline: 35360158]
108. Lum PS, Taub E, Schwandt D, Postman M, Hardin P, Uswatte G. Automated Constraint-Induced Therapy Extension (AutoCITE) for movement deficits after stroke. J Rehabil Res Dev 2004;41(3A):249-258 [FREE Full text] [doi: 10.1682/jrrd.2003.06.0092] [Medline: 15543442]
109. Mansfield A, Wong JS, Bryce J, Brunton K, Inness EL, Knorr S, et al. Use of accelerometer-based feedback of walking activity for appraising progress with walking-related goals in inpatient stroke rehabilitation: a randomized controlled trial. Neurorehabil Neural Repair 2015;29(9):847-857 [FREE Full text] [doi: 10.1177/1545968314567968] [Medline: 25605632]
110. Matarić MJ, Eriksson J, Feil-Seifer DJ, Winstein CJ. Socially assistive robotics for post-stroke rehabilitation. J Neuroeng Rehabil 2007;4:5 [FREE Full text] [doi: 10.1186/1743-0003-4-5] [Medline: 17309795]
111. Mawson S, Nasr N, Parker J, Davies R, Zheng H, Mountain G. A personalized self-management rehabilitation system with an intelligent shoe for stroke survivors: a realist evaluation. JMIR Rehabil Assist Technol 2016;3(1):e1 [FREE Full text] [doi: 10.2196/rehab.5079] [Medline: 28582250]
112. McNulty PA, Thompson-Butel AG, Faux SG, Lin G, Katrak PH, Harris LR, et al. The efficacy of Wii-based Movement Therapy for upper limb rehabilitation in the chronic poststroke period: a randomized controlled trial. Int J Stroke 2015;10(8):1253-1260. [doi: 10.1111/ijs.12594] [Medline: 26332338]

XSL•FO
RenderX

113. Mihelj M, Novak D, Milavec M, Ziherl J, Olenšek A, Munih M. Virtual rehabilitation environment using principles of intrinsic motivation and game design. Presence: Teleop Virt Environ 2012;21(1):1-15. [doi: 10.1162/pres_a_00078]
114. Mitchell C, Bowen A, Tyson S, Conroy P. A feasibility randomized controlled trial of ReaDySpeech for people with dysarthria after stroke. Clin Rehabil 2018;32(8):1037-1046 [FREE Full text] [doi: 10.1177/0269215517748453] [Medline: 29278019]
115. Moan ME, Vonstad EK, Su X, Vereijken B, Solbjør M, Skjæret-Maroni N. Experiences of stroke survivors and clinicians with a fully immersive virtual reality treadmill exergame for stroke rehabilitation: a qualitative pilot study. Front Aging Neurosci 2021;13:735251 [FREE Full text] [doi: 10.3389/fnagi.2021.735251] [Medline: 34795576]
116. Mountain G, Wilson S, Eccleston C, Mawson S, Hammerton J, Ware T, et al. Developing and testing a telerehabilitation system for people following stroke: issues of usability. J Eng Des 2010;21(2-3):223-236. [doi: 10.1080/09544820903333792]
117. Nijenhuis SM, Prange GB, Amirabdollahian F, Sale P, Infarinato F, Nasr N, et al. Feasibility study into self-administered training at home using an arm and hand device with motivational gaming environment in chronic stroke. J Neuroeng Rehabil 2015;12:89 [FREE Full text] [doi: 10.1186/s12984-015-0080-y] [Medline: 26452749]
118. Novak D, Nagle A, Keller U, Riener R. Increasing motivation in robot-aided arm rehabilitation with competitive and cooperative gameplay. J Neuroeng Rehabil 2014;11:64 [FREE Full text] [doi: 10.1186/1743-0003-11-64] [Medline: 24739255]
119. Olafsdottir SA, Jonsdottir H, Bjartmarz I, Magnusson C, Caltenco H, Kytö M, et al. Feasibility of ActivABLES to promote home-based exercise and physical activity of community-dwelling stroke survivors with support from caregivers: a mixed methods study. BMC Health Serv Res 2020;20(1):562 [FREE Full text] [doi: 10.1186/s12913-020-05432-x] [Medline: 32571316]
120. Olafsdottir SA, Jonsdottir H, Magnusson C, Caltenco H, Kytö M, Maye L, et al. Developing ActivABLES for community-dwelling stroke survivors using the Medical Research Council framework for complex interventions. BMC Health Serv Res 2020;20(1):463 [FREE Full text] [doi: 10.1186/s12913-020-05198-2] [Medline: 32450854]
121. Parker J, Mawson S, Mountain G, Nasr N, Zheng H. Stroke patients' utilisation of extrinsic feedback from computer-based technology in the home: a multiple case study realistic evaluation. BMC Med Inform Decis Mak 2014;14:46 [FREE Full text] [doi: 10.1186/1472-6947-14-46] [Medline: 24903401]
122. Paul L, Wyke S, Brewster S, Sattar N, Gill JMR, Alexander G, et al. Increasing physical activity in stroke survivors using STARFISH, an interactive mobile phone application: a pilot study. Top Stroke Rehabil 2016;23(3):170-177. [doi: 10.1080/10749357.2015.1122266] [Medline: 27077973]
123. Pereira F, Badia SBI, Jorge C, Cameirão MS. The use of game modes to promote engagement and social involvement in multi-user serious games: a within-person randomized trial with stroke survivors. J Neuroeng Rehabil 2021;18(1):62 [FREE Full text] [doi: 10.1186/s12984-021-00853-z] [Medline: 33853609]
124. Perez-Marcos D, Chevalley O, Schmidlin T, Garipelli G, Serino A, Vuadens P, et al. Increasing upper limb training intensity in chronic stroke using embodied virtual reality: a pilot study. J Neuroeng Rehabil 2017;14(1):119 [FREE Full text] [doi: 10.1186/s12984-017-0328-9] [Medline: 29149855]
125. Popović MD, Kostić MD, Rodić SZ, Konstantinović LM. Feedback-mediated upper extremities exercise: increasing patient motivation in poststroke rehabilitation. Biomed Res Int 2014;2014:520374 [FREE Full text] [doi: 10.1155/2014/520374] [Medline: 24991557]
126. Proffitt R, Lange B. Feasibility of a customized, in-home, game-based stroke exercise program using the Microsoft Kinect® sensor. Int J Telerehabil 2015;7(2):23-34 [FREE Full text] [doi: 10.5195/ijt.2015.6177] [Medline: 27563384]
127. Rosati G, Oscari F, Reinkensmeyer DJ, Secoli R, Avanzini F, Spagnol S, et al. Improving robotics for neurorehabilitation: enhancing engagement, performance, and learning with auditory feedback. IEEE Int Conf Rehabil Robot 2011;2011:5975373. [doi: 10.1109/ICORR.2011.5975373] [Medline: 22275577]
128. Rowe JB, Chan V, Ingemanson ML, Cramer SC, Wolbrecht ET, Reinkensmeyer DJ. Robotic assistance for training finger movement using a Hebbian model: a randomized controlled trial. Neurorehabil Neural Repair 2017;31(8):769-780 [FREE Full text] [doi: 10.1177/1545968317721975] [Medline: 28803535]
129. Sakakibara BM, Lear SA, Barr SI, Goldsmith CH, Schneeberg A, Silverberg ND, et al. Telehealth coaching to improve self-management for secondary prevention after stroke: a randomized controlled trial of Stroke Coach. Int J Stroke 2022;17(4):455-464 [FREE Full text] [doi: 10.1177/17474930211017699] [Medline: 33949270]
130. Saywell NL, Vandal AC, Mudge S, Hale L, Brown P, Feigin V, et al. Telerehabilitation after stroke using readily available technology: a randomized controlled trial. Neurorehabil Neural Repair 2021;35(1):88-97 [FREE Full text] [doi: 10.1177/1545968320971765] [Medline: 33190615]
131. Scrivener K, Sewastenko J, Bouvier-Farrell A, MacDonald K, Van Rijn T, Tezak J, et al. Feasibility of a self-managed, video-guided exercise program for community-dwelling people with stroke. Stroke Res Treat 2021;2021:5598100 [FREE Full text] [doi: 10.1155/2021/5598100] [Medline: 34035890]
132. Shah N, Amirabdollahian F, Basteris A. Designing motivational games for stroke rehabilitation. New Jersey: IEEE; 2014 Presented at: 7th International Conference on Human System Interactions (HSI); June 16-18, 2014; Costa da Caparica, Portugal p. 166-171. [doi: 10.1109/hsi.2014.6860468]

XSL•FO
RenderX

133. Signal NEJ, McLaren R, Rashid U, Vandal A, King M, Almesfer F, et al. Haptic nudges increase affected upper limb movement during inpatient stroke rehabilitation: multiple-period randomized crossover study. JMIR Mhealth Uhealth 2020;8(7):e17036 [FREE Full text] [doi: 10.2196/17036] [Medline: 32723718]
134. Simpson DB, Bird ML, English C, Gall SL, Breslin M, Smith S, et al. "Connecting patients and therapists remotely using technology is feasible and facilitates exercise adherence after stroke". Top Stroke Rehabil 2020;27(2):93-102. [doi: 10.1080/10749357.2019.1690779] [Medline: 31762412]
135. Song X, van de Ven SS, Chen S, Kang P, Gao Q, Jia J, et al. Proposal of a wearable multimodal sensing-based serious games approach for hand movement training after stroke. Front Physiol 2022;13:811950 [FREE Full text] [doi: 10.3389/fphys.2022.811950] [Medline: 35721546]
136. Standen PJ, Threapleton K, Richardson A, Connell L, Brown DJ, Battersby S, et al. A low cost virtual reality system for home based rehabilitation of the arm following stroke: a randomised controlled feasibility trial. Clin Rehabil 2017;31(3):340-350 [FREE Full text] [doi: 10.1177/0269215516640320] [Medline: 27029939]
137. Subramanian SK, Lourenço CB, Chilingaryan G, Sveistrup H, Levin MF. Arm motor recovery using a virtual reality intervention in chronic stroke: randomized control trial. Neurorehabil Neural Repair 2013;27(1):13-23 [FREE Full text] [doi: 10.1177/1545968312449695] [Medline: 22785001]
138. Sullivan JE, Espe LE, Kelly AM, Veilbig LE, Kwasny MJ. Feasibility and outcomes of a community-based, pedometer-monitored walking program in chronic stroke: a pilot study. Top Stroke Rehabil 2014;21(2):101-110. [doi: 10.1310/tsr2102-101] [Medline: 24710970]
139. Toh SFM, Gonzalez PC, Fong KNK. Usability of a wearable device for home-based upper limb telerehabilitation in persons with stroke: a mixed-methods study. Digit Health 2023;9:20552076231153737 [FREE Full text] [doi: 10.1177/20552076231153737] [Medline: 36776407]
140. Tsekleves E, Paraskevopoulos IT, Warland A, Kilbride C. Development and preliminary evaluation of a novel low cost VR-based upper limb stroke rehabilitation platform using Wii technology. Disabil Rehabil Assist Technol 2016;11(5):413-422. [doi: 10.3109/17483107.2014.981874] [Medline: 25391221]
141. Uswatte G, Taub E, Lum P, Brennan D, Barman J, Bowman MH, et al. Tele-rehabilitation of upper-extremity hemiparesis after stroke: proof-of-concept randomized controlled trial of in-home Constraint-Induced Movement therapy. Restor Neurol Neurosci 2021;39(4):303-318. [doi: 10.3233/RNN-201100] [Medline: 34459426]
142. Valdés BA, Van der Loos HFM. Biofeedback vs. game scores for reducing trunk compensation after stroke: a randomized crossover trial. Top Stroke Rehabil 2018;25(2):96-113. [doi: 10.1080/10749357.2017.1394633] [Medline: 29078743]
143. Wairagkar M, McCrindle R, Robson H, Meteyard L, Sperrin M, Smith A, et al. MaLT—combined motor and language therapy tool for brain injury patients using kinect. Methods Inf Med 2017;56(2):127-137 [FREE Full text] [doi: 10.3414/ME16-02-0015] [Medline: 28220928]
144. Wei WXJ, Fong KNK, Chung RCK, Cheung HKY, Chow ESL. "Remind-to-Move" for promoting upper extremity recovery using wearable devices in subacute stroke: a multi-center randomized controlled study. IEEE Trans Neural Syst Rehabil Eng 2019;27(1):51-59. [doi: 10.1109/TNSRE.2018.2882235] [Medline: 30475722]
145. Whitford M, Schearer E, Rowlett M. Effects of in home high dose accelerometer-based feedback on perceived and actual use in participants chronic post-stroke. Physiother Theory Pract 2020;36(7):799-809. [doi: 10.1080/09593985.2018.1493759] [Medline: 30004818]
146. Widmer M, Held JPO, Wittmann F, Valladares B, Lambercy O, Sturzenegger C, et al. Reward during arm training improves impairment and activity after stroke: a randomized controlled trial. Neurorehabil Neural Repair 2022;36(2):140-150 [FREE Full text] [doi: 10.1177/15459683211062898] [Medline: 34937456]
147. Wittmann F, Held JP, Lambercy O, Starkey ML, Curt A, Höver R, et al. Self-directed arm therapy at home after stroke with a sensor-based virtual reality training system. J Neuroeng Rehabil 2016;13(1):75 [FREE Full text] [doi: 10.1186/s12984-016-0182-1] [Medline: 27515583]
148. Yuan Z, Peng Y, Wang L, Song S, Chen S, Yang L, et al. Effect of BCI-controlled pedaling training system with multiple modalities of feedback on motor and cognitive function rehabilitation of early subacute stroke patients. IEEE Trans Neural Syst Rehabil Eng 2021;29:2569-2577 [FREE Full text] [doi: 10.1109/TNSRE.2021.3132944] [Medline: 34871175]
149. Chong MS, Sit JWH, Karthikesu K, Chair SY. Effectiveness of technology-assisted cardiac rehabilitation: a systematic review and meta-analysis. Int J Nurs Stud 2021;124:104087. [doi: 10.1016/j.ijnurstu.2021.104087] [Medline: 34562846]
150. Hart T, Dijkers MP, Whyte J, Turkstra LS, Zanca JM, Packel A, et al. A theory-driven system for the specification of rehabilitation treatments. Arch Phys Med Rehabil 2019;100(1):172-180. [doi: 10.1016/j.apmr.2018.09.109] [Medline: 30267669]
151. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining "gamification". New York, NY, US: Association for Computing Machinery; 2011 Presented at: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments; September 28-30, 2011; Tampere, Finland p. 9-15. [doi: 10.1145/2181037.2181040]
152. Barrett N, Swain I, Gatzidis C, Mecheraoui C. The use and effect of video game design theory in the creation of game-based systems for upper limb stroke rehabilitation. J Rehabil Assist Technol Eng 2016;3:2055668316643644 [FREE Full text] [doi: 10.1177/2055668316643644] [Medline: 31186903]

XSL•FO
RenderX

153. Welcome to UCLanData. University of Central Lancashire. URL: https://uclandata.uclan.ac.uk/383/ [accessed 2024-03-29]

Abbreviations

BCT: behavior change technique
BCTTv1: behavior change technique taxonomy version 1
DHT: digital health technology
PRISMA-ScR: Preferred Reporting of Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TIDieR: Template for Intervention Description and Replication
VR: virtual reality

Edited by A Mavragani; submitted 15.05.23; peer-reviewed by M Broderick, G Sweeney, E Crayton, D Pogrebnoy; comments to author 11.10.23; revised version received 14.11.23; accepted 26.12.23; published 24.04.24.

Please cite as:
Gooch HJ, Jarvis KA, Stockley RC
Behavior Change Approaches in Digital Technology–Based Physical Rehabilitation Interventions Following Stroke: Scoping Review
J Med Internet Res 2024;26:e48725
URL: https://www.jmir.org/2024/1/e48725
doi:10.2196/48725
PMID:

©Helen J Gooch, Kathryn A Jarvis, Rachel C Stockley. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Applications of Artificial Intelligence for Assessing Fall Risk: Systematic Review

Ana González-Castro[1], PT, MSc; Raquel Leirós-Rodríguez[2], PT, PhD; Camino Prada-García[3], MD, PhD; José Alberto Benítez-Andrades[4], PhD

[1]Nursing and Physical Therapy Department, Universidad de León, Ponferrada, Spain
[2]SALBIS Research Group, Nursing and Physical Therapy Department, Universidad de León, Ponferrada, Spain
[3]Department of Preventive Medicine and Public Health, Universidad de Valladolid, Valladolid, Spain
[4]SALBIS Research Group, Department of Electric, Systems and Automatics Engineering, Universidad de León, León, Spain

Corresponding Author:
Ana González-Castro, PT, MSc
Nursing and Physical Therapy Department
Universidad de León
Astorga Ave
Ponferrada, 24401
Spain
Phone: 34 987442000
Email: agonzc28@estudiantes.unileon.es

Abstract

Background: Falls and their consequences are a serious public health problem worldwide. Each year, 37.3 million falls requiring medical attention occur. Therefore, the analysis of fall risk is of great importance for prevention. Artificial intelligence (AI) represents an innovative tool for creating predictive statistical models of fall risk through data analysis.

Objective: The aim of this review was to analyze the available evidence on the applications of AI in the analysis of data related to postural control and fall risk.

Methods: A literature search was conducted in 6 databases with the following inclusion criteria: the articles had to be published within the last 5 years (from 2018 to 2024), they had to apply some method of AI, AI analyses had to be applied to data from samples consisting of humans, and the analyzed sample had to consist of individuals with independent walking with or without the assistance of external orthopedic devices.

Results: We obtained a total of 3858 articles, of which 22 were finally selected. Data extraction for subsequent analysis varied in the different studies: 82% (18/22) of them extracted data through tests or functional assessments, and the remaining 18% (4/22) of them extracted through existing medical records. Different AI techniques were used throughout the articles. All the research included in the review obtained accuracy values of >70% in the predictive models obtained through AI.

Conclusions: The use of AI proves to be a valuable tool for creating predictive models of fall risk. The use of this tool could have a significant socioeconomic impact as it enables the development of low-cost predictive models with a high level of accuracy.

Trial Registration: PROSPERO CRD42023443277; https://tinyurl.com/4sb72ssv

(J Med Internet Res 2024;26:e54934) doi:10.2196/54934

KEYWORDS

machine learning; accidental falls; public health; patient care; artificial intelligence; AI; fall risk

Introduction

Background

According to alarming figures reported by the World Health Organization in 2021, falls cause 37.3 million injuries annually that require medical attention and result in 684,000 deaths [1]. These figures indicate a significant impact of falls on the health care system and on society, both directly and indirectly [2,3].

Life expectancy has progressively increased over the years, leading to an aging population [4]. By 2050, it is estimated that 16% of the population will be >65 years of age. In this group, the incidence of falls has steadily risen, becoming the leading

XSL•FO
RenderX

cause of accidental injury and death (accounting for 55.8% of such deaths, according to some research) [5,6]. It is estimated that 30% of this population falls at least once a year, negatively impacting their physical and psychological well-being [7,8].

Physically, falls are often associated with severe complications that can lead to extended hospitalizations [9]. These hospitalizations are usually due to serious injuries, often cranioencephalic trauma, fractures, or soft tissue injuries [10,11]. Psychologically, falls among the older adult population tend to result in self-imposed limitations due to the fear of falling again [10,12]. These limitations lead to social isolation as individuals avoid participating in activities or even individual mobility [13]. Consequently, falls can lead to psychological conditions such as anxiety and depression [14,15]. Numerous research studies on the risk of falls are currently underway, with ongoing investigations into various innovations and intervention ideas [16-19]. These studies encompass the identification of fall risk factors [20,21], strategies for prevention [22,23], and the outcomes following rehabilitation [23,24].

In the health care field, artificial intelligence (AI) is characterized by data management and processing, offering new possibilities to the health care paradigm [24]. Some applications of AI in the health care domain include assessing tumor interaction processes [25], serving as a tool for image-based diagnostics [26,27], participating in virus detection [28], and, most importantly, as a statistical and predictive method [29-32].

Several publications have combined AI techniques to address health care issues [33-35]. Within the field of predictive models, it is important to understand certain differentiations. In AI, we have machine learning and deep learning [36-38]. Machine learning encompasses a set of techniques applied to data and can be done in a supervised or unsupervised manner [39,40]. On the other hand, deep learning is typically used to work with larger data sets compared to machine learning, and its computational cost is higher [41,42].

Some examples of AI techniques include the gradient boosting machine [43], learning method, and the long short-term memory (LSTM) [44] and the convolutional neural network (CNN) [45], all of them are deep learning methods.

Objectives

For all the reasons mentioned in the preceding section, it was considered necessary to conduct a systematic review to analyze the scientific evidence of AI applications in the analysis of data related to postural control and the risk of falls.

Methods

Data Sources and Searches

This systematic review and meta-analysis were prospectively registered on PROSPERO (ID CRD42023443277) and followed the Meta-Analyses of Observational Studies in Epidemiology checklist [46] and the recommendations of the Cochrane Collaboration [47].

The search was conducted in January 2024 on the following databases: PubMed, Scopus, ScienceDirect, Web of Science, CINAHL, and Cochrane Library. The Medical Subject Headings (MeSH) terms used for the search included *machine learning*, *artificial intelligent*, *accidental falls*, *rehabilitation*, and *physical therapy specialty*. The terms “predictive model” and “algorithms” were also used. These terms were combined using the Boolean operators AND and OR (Textbox 1).

XSL•FO
RenderX

Textbox 1. Search strategy according to the focused question.

PubMed

- (“machine learning”[MeSH Terms] OR “artificial intelligent”[MeSH Terms]) AND “accidental falls”[MeSH Terms]
- (“machine learning”[MeSH Terms] OR “artificial intelligent”) AND (“rehabilitation”[MeSH Terms] OR “physical therapy specialty”[MeSH Terms])
- “accidental falls” [Title/Abstract] AND “algorithms” [Title/Abstract]
- “accidental falls”[Title/Abstract] AND “predictive model” [Title/Abstract]

Scopus

- TITLE-ABS-KEY (“machine learning” OR “artificial intelligent”) AND TITLE-ABS-KEY (“accidental falls”)
- TITLE-ABS-KEY (“machine learning” OR “artificial intelligent”) AND TITLE-ABS-KEY (“rehabilitation” OR “physical therapy specialty”)
- TITLE-ABS-KEY (“accidental falls” AND “algorithms”)
- TITLE-ABS-KEY (“accidental falls” AND “predictive model”)

ScienceDirect

- Title, abstract, keywords: (“machine learning” OR “artificial intelligent”) AND “accidental falls”
- Title, abstract, keywords: (“machine learning” OR “artificial intelligent”) AND (“rehabilitation” OR “physical therapy specialty”)
- Title, abstract, keywords: (“accidental falls” AND “algorithms”)
- Title, abstract, keywords: (“accidental falls” AND “predictive model”)

Web of Science

- TS=(“machine learning” OR “artificial intelligent”) AND TS=“accidental falls”
- TS=(“machine learning” OR “artificial intelligent”) AND TS= (“rehabilitation” OR “physical therapy specialty”)
- AB= (“accidental falls” AND “algorithms”)
- AB= (“accidental falls” AND “predictive model”)

CINAHL

- (MH “machine learning” OR MH “artificial intelligent”) AND MH “accidental falls”
- (MH “machine learning” OR MH “artificial intelligent”) AND (MH “rehabilitation” OR MH “physical therapy specialty”)
- (AB “accidental falls”) AND (AB “algorithms”)
- (AB “accidental falls”) AND (AB “predictive model”)

Cochrane Library

- (“machine learning” OR “artificial intelligent”) in Title Abstract Keyword AND “accidental falls” in Title Abstract Keyword
- (“machine learning” OR “artificial intelligent”) in Title Abstract Keyword AND (“rehabilitation” OR “physical therapy specialty”) in Title Abstract Keyword
- “accidental falls” in Title Abstract Keyword AND “algorithms” in Title Abstract Keyword
- “accidental falls” in Title Abstract Keyword AND “predictive model” in Title Abstract Keyword

Study Selection

After removing duplicates, 2 reviewers (AGC and RLR) independently screened articles for eligibility. In the case of disagreement, a third reviewer (JABA) finally decided whether the study should be included or not. We calculated the κ coefficient and percentage agreement scores to assess reliability before any consensus and estimated the interrater reliability using κ. Interrater reliability was estimated using κ>0.7 indicating a high level of agreement between the reviewers, κ of 0.5 to 0.7 indicating a moderate level of agreement, and κ<0.5 indicating a low level of agreement [48].

For the selection of results, the inclusion criteria were established as follows: (1) articles should have been published in the last 5 years (from 2018 to the present); (2) they must apply some AI method; (3) AI analyses should be applied to data from samples of humans; and (4) the sample analyzed should consist of people with independent walking, with or without the use of external orthopedic devices.

After screening the data, extracting, obtaining, and screening the titles and abstracts for inclusion criteria, the selected abstracts were obtained in full texts. Titles and abstracts lacking sufficient information regarding inclusion criteria were also

XSL•FO
RenderX

obtained as full texts. Full-text articles were selected in case of compliance with inclusion criteria by the 2 reviewers using a data extraction form.

Data Extraction and Quality Assessment

The 2 reviewers mentioned independently extracting data from the included studies using a customized data extraction table in Excel (Microsoft Corporation). In case of disagreement, both reviewers debated until an agreement was reached.

The data extracted from the included articles for further analysis were: demographic information (title, authors, journal, and year), characteristics of the sample (age, inclusion and exclusion criteria, and number of participants), study-specific parameters (study type, AI techniques applied, and data analyzed), and the results obtained. Tables were used to describe both the studies' characteristics and the extracted data.

Assessment of Risk of Bias

The methodological quality of the selected articles was evaluated using the Critical Review Form for Quantitative Studies [49]. The ROBINS-E (Risk of Bias in Nonrandomized Studies of Exposures) tool was used to evaluate the risk of bias [50].

Results

Characteristics of the Selected Studies

A total of 3858 articles were initially retrieved, with 1563 duplicates removed. From the remaining 2295 articles, 2271 were excluded based on the initial selection criteria, leaving 24 articles for the subsequent analysis. In this second analysis, 2 articles were removed as they were systematic reviews, and 22 articles were finally selected [51-72] (Figure 1). After the first reading of all candidate full texts, the kappa score for inclusion of the results of reviewers 1 and 2 was 0.98, indicating a very high level of agreement.

The methodological quality of the 22 analyzed studies (Table S1 in Multimedia Appendix 1 [51,52,54,56,58,59,61,63,64,69,70,72]) ranged from 11 points in 2 (9.1%) studies [52,65] to 16 points in 7 (32%) studies [53,54,56,63,69-71].

XSL•FO
RenderX

Figure 1. Identification of studies via databases and registers.

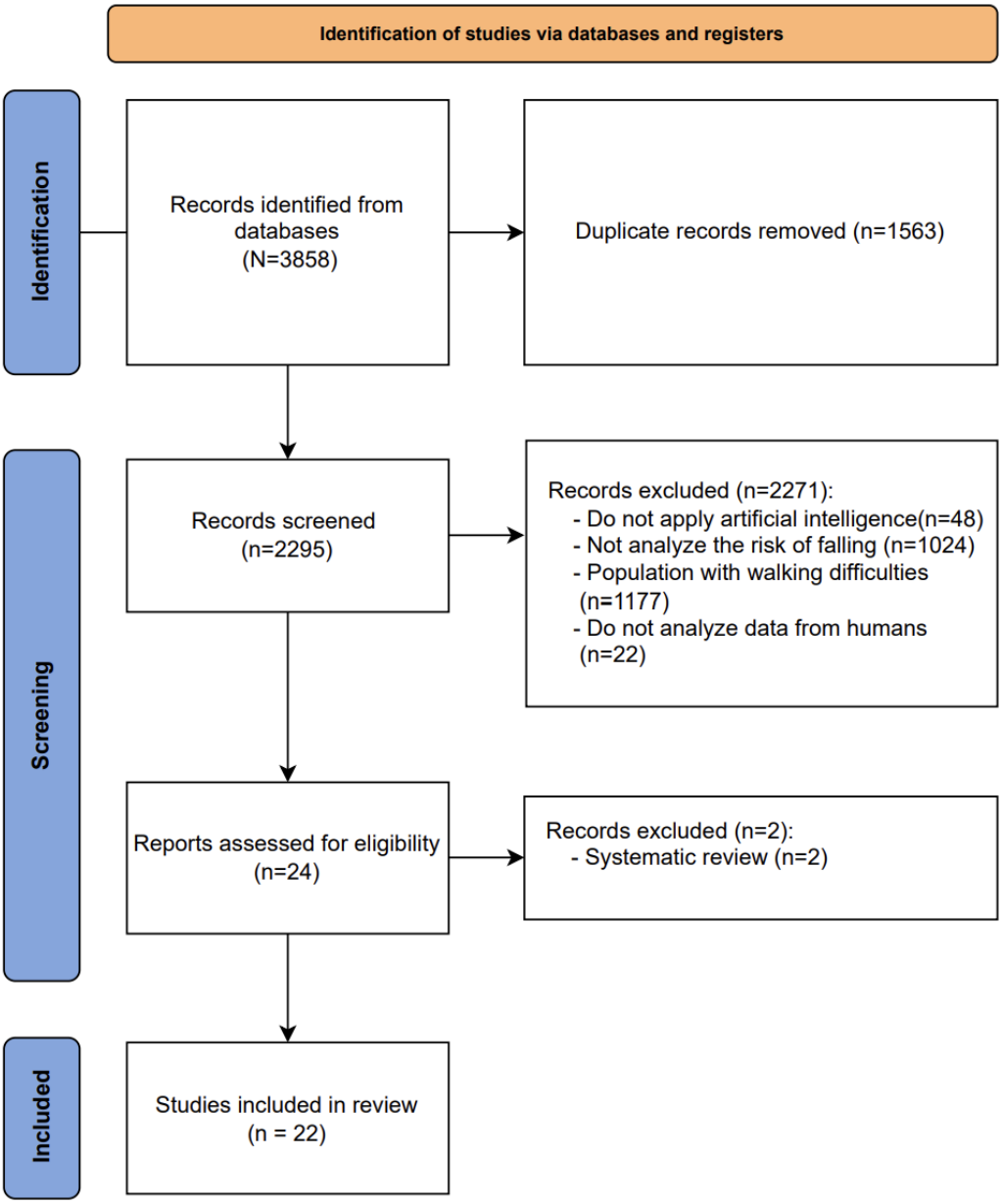

Study Characteristics and Risk of Bias

All the selected articles were cross-sectional observational studies (Table 1).

In total, 34 characteristics affecting the risk of falls were extracted and classified into high fall-risk and low fall-risk groups with the largest sample sizes significantly differing from the rest. Studies based on data collected from various health care systems had larger sample sizes, ranging from 22,515 to 265,225 participants [60,65,67]. In contrast, studies that applied some form of evaluation test had sample sizes ranging from 8 participants [56] to 746 participants [55].

It is worth noting the various studies conducted by Dubois et al [54,72], whose publications on fall risk and machine learning started in 2018 and progressed until 2021. A total of 9.1% (2/22) of the articles by this author were included in the final selection [54,72]. Both articles used samples with the same characteristics, even though the first one was composed of 43 participants [54] and the last one had 30 participants [72]. All 86.4% (19/22) of the articles used samples of individuals aged ≥65 years [51-60,62-65,68-72]. In the remaining 13.6% (3/22) of the articles, the ages ranged between 16 and 62 years [61,66,67].

Althobaiti et al [61] used a sample of participants between the ages of 19 and 35 years for their research, where these participants had to reproduce examples of falls for subsequent analysis. In 2022, Ladios-Martin et al [67] extracted medical data from participants aged >16 years for their research. Finally, in 2023, the study by Maray et al [66] used 3 types of samples, with ages ranging from 21 to 62 years. Among the 22 selected articles, only 1 (4.5%) of them did not describe the characteristics of its sample [52].

Finally, regarding the sex of the samples, 13.6% (3/22) of the articles specified in the characteristics of their samples that only female individuals were included among their participants [53,59,70].

XSL•FO
RenderX

Table 1. Methodological characteristics of results analyzed.

Study	Age range (mean; years)	Assessment procedures	AI[a] techniques	Results
Nait Aicha et al [51], (2018)	65-99 (75.3)	Over the following 6 months, they had monthly consultations to monitor the incidence of falls.	Classification of participants into fall and nonfall categories and creation of an ML[b] model that predicts falls from the accelerometer data.	A total of 34.1% of the participants experienced at least one fall. ML applied to the accelerometric data showed a predictive model of comparable accuracy to conventional fall risk assessment models.
Althobaiti et al [61], (2020)	19-34 (nd[c])	Participants performed 6 ADLs[d] and 9 falls on a foam landing mat. Accelerometer data were transmitted via Bluetooth and recorded continuously.	After extracting the signals from the accelerometers, the data were used to train ML models.	The ML models created can predict and detect falls with high accuracy. The supervised classification experiments achieved an F_1-score of 98.4% for distinguishing between falls and nonfalls, and an F_1-score of 88.1% for distinguishing between ADLs.
Dubois et al [54], (2018)	61-93 (83)	The TUG[e] test was administered 3 times, with 3-5 minutes of rest between attempts. The best result was used to categorize participants into low and high risk of falling.	The images were analyzed with a processing algorithm, and ML techniques were applied to predict the risk of falls.	All participants were classified using a combination of walking speed, stride length, and sitting speed. The algorithm used was used to correctly classify them according to their risk of falling and to assess their progress.
Dubois et al [72], (2021)	61-93 (83.3)	A first assessment, including the Tinetti and TUG tests, was conducted by a physiotherapist to identify patients who are "high risk." Daily activity was monitored for 8 hours.	Different parameters, such as walking speed, sitting speed, and total sitting time, were analyzed. The data were analyzed by combining statistics and ML algorithms.	The results obtained from the algorithms were compared with those obtained by physiotherapists using the Tinetti and TUG tests. Step length and sit-to-stand time could complement the assessment tools and improve the risk of falls.
Eichler et al [68], (2022)	65-nd (nd)	BBS[f] (consisting of 14 tasks). A total of 2 physiotherapists independently scored each participant on the 14 tasks. The number of participants varied across tasks due to differences in performance abilities.	ML to predict the risk of falling by categorizing it as "high," "medium," or "low." Initially, the movements of the participants were tracked using deep learning applied to the camera while they performed the BBS. Then, an algorithm predicted the BBS score, leading to the final risk classification.	The results obtained through deep learning were compared with the scores given by the physiotherapists, and the overall accuracy was 75.5% correct. The ML algorithm reduced the false-negative rate by adjusting the thresholds. The most predictive features were: 360° turn, alternating feet on the step, chair transfer, and reaching forward with the arm extended.
Ferrete et al [52], (2019)	nd	Participants were instructed to act out falls.	A total of 2 deep learning approaches were explored: the classifier based on ASM[g] and CNN[h].	The CNN demonstrated 92.7% accuracy in differentiating normal gait, prefall condition, and fall situation, considering the prefall step, and 100% when not considering the prefall step.
Gillain et al [64], (2019)	65-89 (71.3)	Participants underwent a physical and functional assessment, which included the evaluation of gait speed, stride length, frequency, symmetry, regularity, and minimum toe separation. These variables were recorded under conditions of comfortable walking, fast walking, and dual-task walking.	A supervised ML algorithm (J48) was applied to the recorded data.	A classification tree was obtained that correctly identifies 80% of future falls based on gait parameters, gender, leg length, and age. Accuracy was 84%, sensitivity was 80%, specificity was 87%, positive predictive value was 78%, and negative predictive value was 88%.
Greene et al [70], (2021)	nd (74.7)	The test lasted 30 seconds and was conducted with both eyes open and closed. A comprehensive geriatric assessment and the BBS were used to obtain risk factors for each participant. Data were collected using the Kinesis Balance FP[i] application, and a questionnaire on the socio-demographic data of the participants was administered through the app.	The ML algorithm assessed balance and the risk of falling based on previously collected data.	The use of ML models for balance and fall risk assessment could improve the approach to falls, including prescribing personalized prevention programs and monitoring progress. This would improve the remote care system with more targeted and digitized initiatives.

XSL•FO
RenderX

Study	Age range (mean; years)	Assessment procedures	AI[a] techniques	Results
Hauth et al [56], (2021)	69-82 (nd)	IMUs[j] measured linear and angular accelerations in the mediolateral, anteroposterior, and longitudinal axes. Participants wore a wrist-mounted voice recorder to indicate when the event occurred and to provide context descriptions.	Customized algorithms were used to integrate the data from the IMUs in 6 dimensions to obtain the position, orientation, and velocity of the foot. Model performance was evaluated in terms of AUROC[k] and AUPR[l].	Algorithms for data integration could be used retrospectively to track the occurrence of falls or loss of balance events. The model showed discriminatory ability to differentiate a fall from a normal walking event. The best model achieved an AUROC ≥0.87 for nonparticipants.
Hsu et al [58], (2020)	nd	All participants were assessed for fall risk using the MFS[m] upon admission.	The XGB[n] predictive model was applied, and its performance was compared with that of the MFS.	The XGB model is more sensitive than the MFS, and therefore it is more appropriate for assessing the risk of falls.
Hu et al [59], (2020)	nd	Participants underwent treadmill walking and MCT[o] with small, medium, and large perturbations. They were classified into a high-balance group and a low-balance group based on the MCT results.	ML analysis: 5-fold cross-validation where the algorithms were repeatedly randomly trained. Using the GBM[p] algorithm, participants were classified into high and low equilibrium through analysis of the accelerometric data.	The feasibility of ML was demonstrated by matching accelerometric data and predicting balance dysfunctions. This method was comparable to the data obtained from instrumented balance tests and was also less expensive.
Kim et al [63], (2019)	nd	All participants simulated various types of falls, resulting in a total of 282 scenes constituting the database. Falls, including forward, backward, and sideways motions, were replicated. Additionally, images without falls depicted actions such as sitting on a chair or lying on a bed.	The application of ML: The RF[q] system was used as a data management and predictive tool. A total of 200 decision trees were used, with 25 trees deep for each tree.	Improved detection of false fall data. The model combining accelerometric data and vision data gave promising results. The application of ML achieved good generalization of the obtained data. Accuracy of the proposed method was 90%.
Ladios-Martin et al [67], (2022)	16-106 (nd)	In total, 91 variables related to the risk of falls were selected. A total of 2 models were created: model A included the variable FP, while model B did not.	The performance of 10 algorithms was evaluated. Once the final models were created, they were compared with each other by evaluating their performances.	The performance of the model that included the FP variable performed better than that of the one that did not include it. Using AI techniques and the FP variable, a high-performance model is obtained that can be exported to other contexts.
Lo et al [65], (2019)	65-nd (nd)	Researchers established which patients were at risk of falls based on existing medical record data.	ML techniques (RF algorithm) were used to build a predictive model of fall risk.	The risk of falls is affected by a large number of factors (age, clinical diagnoses, daily habits, environment, and hygiene). ML makes it possible to predict the risk of falls and rank the importance of each risk factor.
Lockhart et al [57], (2021)	56-9 (74.3)	Participants conducted 10-meter walks. Out of 127 participants, 79.5%) did not experience falls, while 20.5%) did.	A total of 3 types of deep learning techniques were applied, and a predictive model was created with the data from the 127 participants (RF, RF with feature engineering, and RF with feature engineering and linear and nonlinear variables). The accuracy of the model was tested with the remaining 34.6%) participants with a 6-month follow-up of their fall history.	The IMU was a good tool for the identification of falls risk. Data analysis through AI obtained very good predictive results.
Maray et al [66], (2023)	21-65 (nd)	All participants performed fall representations.	Deep learning was used to match the data extracted from devices.	A deep learning model was built using data obtained from different handheld devices that gave better results than the initial creation of a new model.
Noh et al [55], (2021)	63-89 (73)	In total, 3 walking tests were conducted on a 20-meter straight walkway, involving walking at slower, faster, or the participant's preferred speed. Participants were categorized into low fall-risk and high fall-risk groups based on their fall history and fear of falling.	The ML XGB algorithm was applied to perform the predictive analysis. The predictive analysis sought to predict factors affecting the risk of falls. The parameters were spatiotemporal parameters extracted from the gait tests.	In total, 34 characteristics affecting the risk of falls were extracted and classified into high fall-risk and low fall-risk groups. The study demonstrated >70% accuracy. Stride length, speed, and posture accurately classified the risk of falling.

XSL•FO
RenderX

Study	Age range (mean; years)	Assessment procedures	AI[a] techniques	Results
Qiu et al [53], (2018)	nd	Participants underwent 7 tests: Sensory Integration Test, Limits of Stability Test, Sit-to-Stand test (5 times), TUG test, motor function test, Choice Reaction test, and Fear of Falling test.	A total of 38 algorithms were developed to extract meaningful measurements from the data extracted from the sensors. In total, 38 meaningful measures were obtained and incorporated into a supervised ML model. A prediction of "fallers" and "nonfallers" was made based on this data. The accuracy of the prediction was assessed through a 10-fold cross-validation.	The data obtained distinguished between people who fall and those who do not, with overall accuracy of 89.4% (92% sensitivity and 84.9% specificity). The model created from the inertial sensors in the test battery and the subsequent analysis of the data through AI are promising.
Roshdibenam et al [71], (2021)	65-95 (74.4)	A physiotherapist conducted various assessments, including the TUG test, 30-Second Stand Test, and 4-Stage Balance tests, and classified participants into "high" and "low" fall risk categories.	The CNN algorithm was used for the data for each individual sensor location. Algorithms were applied to evaluate the data obtained by the portable sensors.	Deep learning had high similarities to those obtained by the physiotherapist. Deep learning techniques combined with nonintrusive wearable sensors could provide a new tool for measuring risk factors and predicting fall risk.
Tang et al [69], (2022)	nd (66)	Participants were divided into 2 groups: "Normal walking" and "Abnormal gait." The "abnormal" group had a higher average age and a greater proportion of men compared to the "normal" group. The TUG test videos of all participants were observed.	ML with the AlphaPose algorithm and the naïve Bayes classifier to identify normal or abnormal gait.	The naïve Bayes classifier obtained a complete data accuracy of 90.1% and a LOOCV[r] accuracy of 89.1% for the detection of abnormal gait performance.
Tunca et al [62], (2020)	nd (76.8)	Participants were divided into 2 groups: "high fall risk" and "low fall risk." A total of 2 intervention groups were created: the validation set (30 low risk+30 high risk) and the test set (9 low risk+7 high risk). Participants were instructed to walk 8 m in a straight line at least 3 times.	The applicability of deep learning LSTM[s] networks were compared with traditional ML with respect to fall risk.	The LSTM achieved classification accuracy of 89% compared to accuracy of 82.8% (the highest accuracy achieved by the traditional ML models used).
Ye et al [60], (2020)	65-nd (nd)	A record of falls was compiled from April 2017 to March 2018. A total of 4361 falls were recorded.	The XGB algorithm was used to create the crash prediction model.	The fall prediction model achieved improved discriminatory ability. It may be able to identify risk factors and facilitate prevention interventions.

[a]AI: artificial intelligence.
[b]ML: machine learning.
[c]nd: none described.
[d]ADL: activities of daily living.
[e]TUG: Timed Up and Go.
[f]BBS: Berg Balance Scale.
[g]ASM: associative skill memories.
[h]CNN: convolutional neural network.
[i]FP: fall prevention.
[j]IMU: inertial measurement unit.
[k]AUROC: area under the receiver operating characteristic curve.
[l]AUPR: area under the precision-recall curve.
[m]MFS: Morse Fall Scale.
[n]XGB: extreme gradient boosting.
[o]MCT: motor control test.
[p]GBM: gradient boosting machine.
[q]RF: random forest.
[r]LOOCV: leave-one-out cross-validation.
[s]LSTM: long short-term memory.

XSL•FO
RenderX

Applied Assessment Procedures

All articles initially analyzed the characteristics of their samples to subsequently create a predictive model of the risk of falls. However, they did not all follow the same evaluation process.

Regarding the applied assessment procedures, 3 main options stood out: studies with tests or assessments accompanied by sensors or accelerometers [51-57,59,61-63,66,70-72], studies with tests or assessments accompanied by cameras [68,69], or studies based on medical records [58,60,65,67] (Figure 2). Guillan et al [64] performed a physical and functional evaluation of the participants. In their study, they evaluated parameters such as walking speed, stride frequency and length, and the minimum space between the toes. Afterward, they asked them to record the fall events they had during the past 2 years in a personal diary.

Figure 2. Infographic about the characteristics of the analyzed research.

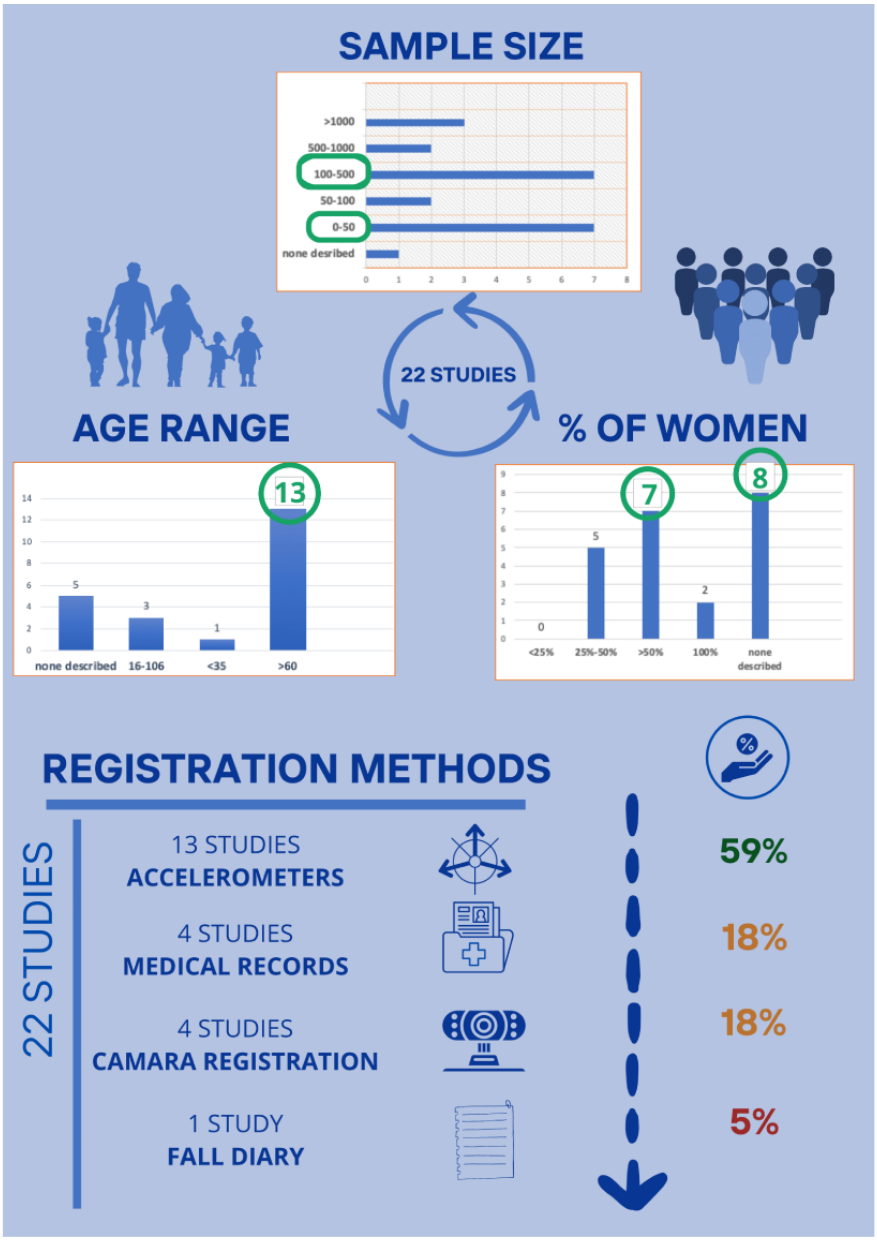

In total, 22.7% (5/22) of the studies used the Timed Up and Go test [53,54,69,71,72]. In 18.2% (4/22) of them, the participants performed the test while wearing a sensor to collect data [53,54,71,72]. In 1 (4.5%) study, the test was recorded with a camera for later analysis [69]. Another commonly used method in studies was to ask participants to perform everyday tasks or activities of daily living while a sensor collected data for subsequent analysis. Specifically, 18.2% (4/22) of the studies used this method to gather data [51,56,61,62].

A total of 22.7% (5/22) of the studies asked participants to simulate falls and nonfalls while a sensor collected data [52,61-63,66]. In this way, the data obtained were used to create the predictive model of falls. As for the tests used, Eichler et al [68] asked participants to perform the Berg Balance Scale while a camera recorded their performance.

Finally, other authors created their own battery of tests for data extraction [55,59,64,70]. Gillain et al [64] used gait records to analyze speed, stride length, frequency, symmetry, regularity, and foot separation. Hu et al [59] asked their participants to

XSL•FO
RenderX

perform normal walking, the postural reflexive response test, and the motor control test. In the study by Noh et al [55], gait tests were conducted, involving walking 20 m at different speeds. Finally, Greene et al [70] created a 12-question questionnaire and asked their participants to maintain balance while holding a mobile phone in their hand.

AI Techniques

The selected articles used various techniques within AI. They all had the same objective in applying these techniques, which was to achieve a predictive and classification model for the risk of falls [51-72].

In chronological order, in 2018, Nait Aicha et al [51] compared single-task learning models with multitask learning, obtaining better evaluation results through multitask learning. In the same year, Dubois et al [54] applied AI techniques that analyzed multiple parameters to classify the risk of falls in their sample. Qiu et al [53], also in the same year, used 6 machine learning models (logistic regression, naïve Bayes, decision tree, RF, boosted tree, and support vector machine) in their research.

In contrast, in 2019, Ferrete et al [52] compared the applicability of 2 different deep learning models: the classifier based on associative skill memories and a CNN classifier. In the same year, after confirming the applicability of AI as a predictive method for the risk of falls, various authors investigated through methods such as the RF to identify factors that can predict and quantify the risk of falls [63,65].

Among the selected articles, 5 (22.7%) were published in 2020 [58-62]. The research conducted by Tunca et al [62] compared the applicability of deep learning LSTM networks with traditional machine learning applied to the risk of falls. Hu et al [59] first used cross-validation, where algorithms were trained randomly, and then used the gradient boosting machine algorithm to classify participants as high or low risk. Ye et al [60] and Hsu et al [58] both used the extreme gradient boosting (XGBoost) algorithm based on machine learning to create their predictive model. In the same year, Althobaiti et al [61] trained machine learning models for their research.

In 2021, Lockhart et al [57] started using 3 deep learning techniques simultaneously with the same goal as before: to create a predictive model for the risk of falls. Specifically, they used the RF, RF with feature engineering, and RF with feature engineering and linear and nonlinear variables. Noh et al [55], again in the same year, used the XGBoost algorithm, while Roshdibenam et al [71], on the other hand, used the CNN algorithm for each location of the wearable sensors used in their research. Various machine learning techniques were used for classifying the risk of falls and for balance loss events in the research by Hauth et al [56]. Dubois et al [72] used the following algorithms: decision tree, adaptive boosting, neural net, naïve Bayes, k-nearest neighbors, linear support vector machine, radial basis function support vector machine, RF, and quadratic discriminant analysis. Hauth et al [56], on the other hand, used regularized logistic regression and bidirectional LSTM networks. In the research conducted by Greene et al [70], AI was used, but the specific procedure that they followed is not described.

Tang et al [69] published their research with innovation up to that point. In their study, they used a smart gait analyzer with the help of deep learning techniques to assess the diagnostic accuracy of fall risk through vision. Months later, in August 2022, Ladios-Martin et al [67] published their research, in which they compared 2 deep learning models to achieve the best results in terms of specificity and sensitivity in detecting fall risk. The first model used the Bayesian Point Machine algorithm with a fall prevention variable, and the second one did not use the variable. They obtained better results when using that variable, a mitigating factor defined as a set of care interventions carried out by professionals to prevent the patient from experiencing a fall during hospitalization. Particularly controversial, as its exclusion could obscure the model's performance. Eichler et al [68], on the other hand, used machine learning–based classifier training and later tested the performance of RFs in score predictions.

Finally, in January 2023, Maray et al [66] published their research, linking the previously mentioned terms (AI and fall risk) with 3 wearable devices that are commonly used today. They collected data through these devices and applied transfer learning to generalize the model across heterogeneous devices.

Findings

The results of the 22 articles provided promising data, and all of them agreed on the feasibility of applying various AI techniques as a method for predicting and classifying the risk of falls. Specifically, the accuracy values obtained in the studies exceed 70%. Noh et al [55] achieved the "lowest" accuracy among the studies conducted, with a 70% accuracy rate. Ribeiro et al [52] obtained an accuracy of 92.7% when using CNN to differentiate between normal gait and fall events. Hsu et al [58] further demonstrated that the XGBoost model is more sensitive than the Morse Fall Scale. Similarly, in their comparative study, Nait Aicha et al [51] also showed that a predictive model created from accelerometer data with AI is comparable to conventional models for assessing the risk of falls. More specifically, Dubois et al [54] concluded that using 1 gait-related parameter (excluding velocity) in combination with another parameter related to seated position allowed for the correct classification of individuals according to their risk of falls.

Discussion

Principal Findings

The aim of this research was to analyze the scientific evidence regarding the applications of AI in the analysis of data related to postural control and the risk of falls. On the basis of the analysis of results, it can be asserted that the following risk factors were identified in the analyzed studies: age [65], daily habits [65], clinical diagnoses [65], environmental and hygiene factors [65], sex [64], stride length [55,72], gait speed [55], and posture [55]. This aligns with other research that also identifies sex [73,74], age [73], and gait speed [75].

On the other hand, the "fear of falling" has been identified in various studies as a risk factor and a predictor of falls [73,76], but it was not identified in any of the studies included in this review.

XSL•FO
RenderX

As for the characteristics of the analyzed samples, only 9.1% (2/22) of the articles used a sample composed exclusively of women [53,59], and no article used a sample composed exclusively of men. This fact is incongruent with reality, as women have a longer life expectancy than men, and therefore, the number of women aged >65 years is greater than the number of men of the same age [77]. Furthermore, women experience more falls than men [78]. The connection between menopause and its consequences, including osteopenia, suggests a higher risk of falls among older women than among men of the same age [79,80].

Within the realm of analysis tools, the most frequently used devices to analyze participants were accelerometers [51-57,59,61-63,66,70-72]. However, only 36.4% (8/22) of the studies provided all the information regarding the characteristics of these devices [51,53,59,61,63,66,70,72]. On the other hand, 18.2% (4/22) of the studies used the term "inertial measurement unit" as the sole description of the devices used [55-57,71].

The fact that most of the analyzed procedures involved the use of inertial sensors reflects the current widespread use of these devices for postural control analysis. These sensors, in general (and triaxial accelerometers in particular), have demonstrated great diagnostic capacity for balance [81]. In addition, they exhibit good sensitivity and reliability, combined with their portability and low economic cost [82]. Another advantage of triaxial accelerometers is their versatility in both adult and pediatric populations [83-86], although the studies included in this review did not address the pediatric population.

The remaining studies extracted data from cameras [68,69], medical records [58,60,65,67], and other functional and clinical tests [59,64,70]. Regarding the AI techniques used, out of the 18.2% (4/22) of articles that used deep learning techniques [52,57,62,71], only 4.5% (1/22) did not provide a description of the sample characteristics [52]. In this case, the authors focused on the AI landscape, while the rest of the articles strike a balance between AI and health sciences.

Regarding the validity of the generated models, only 40.9% (9/22) of the articles assessed this characteristic [52,53,55,61-64,68,69]. The authors of these 9 (N=22, 40.9%) articles evaluated the validity of the models through accuracy. All the results obtained reflected accuracies exceeding 70%, with Ribeiro et al [52] achieving a notable accuracy of 92.7% and 100%. Specifically, they obtained a 92.7% accuracy through the CNN model for distinguishing normal gait, the prefall condition, and the falling situation, considering the step before the fall, and 100% when not considering it [52].

The positive results of sensitivity and specificity can only be compared between the studies of Qiu et al [53] and Gillain et al [64], as they were the only ones to take them into account, and in both investigations, they were very high. Similarly, in the case of the F_1-score, only Althobaiti et al [61] examined this validity measure. This measure is the result of combining precision and recall into a single figure, and the outcome obtained by these researchers was promising.

Despite these differences, the 22 studies obtained promising results in the health care field [51-72]. Specifically, their outcomes highlight the potential of AI integration into clinical settings. However, further research is necessary to explore how health care professionals can effectively use these predictive models. Consequently, future research should focus on studying the application and integration of the already-developed models. In this context, fall prevention plans could be implemented for the target populations identified by the predictive models. This approach would allow for a retrospective analysis to determine if the combination of predictive models with prevention programs effectively reduces the prevalence of falls in the population.

Limitations

Regarding limitations, the articles showed significant variation in the sample sizes selected. Moreover, even in the study with the largest sample size (with 265,225 participants [60]), the amount of data analyzed was relatively small. In addition, several of the databases used were not generated specifically for the published research but rather derived from existing medical records [58,60,65,67]. This could explain the significant variability in the variables analyzed across different studies.

Strengths

Despite the limitations, this research has strengths, such as being the first systematic review on the use of AI as a tool to analyze postural control and the risk of falls. Furthermore, a total of 6 databases were used for the literature search, and a comprehensive article selection process was carried out by 3 researchers. Finally, only cross-sectional observational studies were selected, and they shared the same objective.

Conclusions

The use of AI in the analysis of data related to postural control and the risk of falls proves to be a valuable tool for creating predictive models of fall risk. It has been identified that most AI studies analyze accelerometer data from sensors, with triaxial accelerometers being the most frequently used.

For future research, it would be beneficial to provide more detailed descriptions of the measurement procedures and the AI techniques used. In addition, exploring larger databases could lead to the development of more robust models.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quality scores of reviewed studies (Critical Review Form for Quantitative Studies tool results).

XSL•FO
RenderX

[DOCX File , 18 KB - jmir_v26i1e54934_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 156 KB - jmir_v26i1e54934_app2.pdf]

References

1. Step safely: strategies for preventing and managing falls across the life-course. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/978924002191-4 [accessed 2024-04-02]
2. Keall MD, Pierse N, Howden-Chapman P, Guria J, Cunningham CW, Baker MG. Cost-benefit analysis of fall injuries prevented by a programme of home modifications: a cluster randomised controlled trial. Inj Prev 2017 Feb;23(1):22-26. [doi: 10.1136/injuryprev-2015-041947] [Medline: 27312961]
3. Almada M, Brochado P, Portela D, Midão L, Costa E. Prevalence of falls and associated factors among community-dwelling older adults: a cross-sectional study. J Frailty Aging 2021;10(1):10-16. [doi: 10.14283/jfa.2020.44] [Medline: 33331616]
4. Menéndez-González L, Izaguirre-Riesgo A, Tranche-Iparraguirre S, Montero-Rodríguez Á, Orts-Cortés MI. [Prevalence and associated factors of frailty in adults over 70 years in the community]. Aten Primaria 2021 Dec;53(10):102128 [FREE Full text] [doi: 10.1016/j.aprim.2021.102128] [Medline: 34560377]
5. Guirguis-Blake JM, Michael YL, Perdue LA, Coppola EL, Beil TL. Interventions to prevent falls in older adults: updated evidence report and systematic review for the US preventive services task force. JAMA 2018 Apr 24;319(16):1705-1716. [doi: 10.1001/jama.2017.21962] [Medline: 29710140]
6. Pereira CB, Kanashiro AM. Falls in older adults: a practical approach. Arq Neuropsiquiatr 2022 May;80(5 Suppl 1):313-323 [FREE Full text] [doi: 10.1590/0004-282X-ANP-2022-S107] [Medline: 35976297]
7. Byun M, Kim J, Kim M. Physical and psychological factors affecting falls in older patients with arthritis. Int J Environ Res Public Health 2020 Feb 09;17(3):1098 [FREE Full text] [doi: 10.3390/ijerph17031098] [Medline: 32050468]
8. Goh HT, Nadarajah M, Hamzah NB, Varadan P, Tan MP. Falls and fear of falling after stroke: a case-control study. PM R 2016 Dec 04;8(12):1173-1180. [doi: 10.1016/j.pmrj.2016.05.012] [Medline: 27268565]
9. Alanazi FK, Lapkin S, Molloy L, Sim J. The impact of safety culture, quality of care, missed care and nurse staffing on patient falls: a multisource association study. J Clin Nurs 2023 Oct 12;32(19-20):7260-7272. [doi: 10.1111/jocn.16792] [Medline: 37309059]
10. Hossain A, Lall R, Ji C, Bruce J, Underwood M, Lamb SE. Comparison of different statistical models for the analysis of fracture events: findings from the Prevention of Falls Injury Trial (PreFIT). BMC Med Res Methodol 2023 Oct 02;23(1):216 [FREE Full text] [doi: 10.1186/s12874-023-02040-1] [Medline: 37784050]
11. Williams CT, Whyman J, Loewenthal J, Chahal K. Managing geriatric patients with falls and fractures. Orthop Clin North Am 2023 Jul;54(3S):e1-12. [doi: 10.1016/j.ocl.2023.04.001] [Medline: 37349065]
12. Gadhvi C, Bean D, Rice D. A systematic review of fear of falling and related constructs after hip fracture: prevalence, measurement, associations with physical function, and interventions. BMC Geriatr 2023 Jun 23;23(1):385 [FREE Full text] [doi: 10.1186/s12877-023-03855-9] [Medline: 37353752]
13. Lohman MC, Fallahi A, Mishio Bawa E, Wei J, Merchant AT. Social mediators of the association between depression and falls among older adults. J Aging Health 2023 Aug 12;35(7-8):593-603. [doi: 10.1177/08982643231152276] [Medline: 36633960]
14. Smith AD, Silva AO, Rodrigues RA, Moreira MA, Nogueira JD, Tura LF. Assessment of risk of falls in elderly living at home. Rev Lat Am Enfermagem 2017 Apr 06;25:e2754 [FREE Full text] [doi: 10.1590/1518-8345.0671.2754] [Medline: 28403333]
15. Koh V, Matchar DB, Chan A. Physical strength and mental health mediate the association between pain and falls (recurrent and/or injurious) among community-dwelling older adults in Singapore. Arch Gerontol Geriatr 2023 Sep;112:105015. [doi: 10.1016/j.archger.2023.105015] [Medline: 37060804]
16. Soh SE, Morgan PE, Hopmans R, Barker AL, Ackerman IN. The feasibility and acceptability of a falls prevention e-learning program for physiotherapists. Physiother Theory Pract 2023 Mar 18;39(3):631-640. [doi: 10.1080/09593985.2021.2023928] [Medline: 35040710]
17. Morat T, Snyders M, Kroeber P, De Luca A, Squeri V, Hochheim M, et al. Evaluation of a novel technology-supported fall prevention intervention - study protocol of a multi-centre randomised controlled trial in older adults at increased risk of falls. BMC Geriatr 2023 Feb 18;23(1):103 [FREE Full text] [doi: 10.1186/s12877-023-03810-8] [Medline: 36803459]
18. You T, Koren Y, Butts WJ, Moraes CA, Yeh GY, Wayne PM, et al. Pilot studies of recruitment and feasibility of remote Tai Chi in racially diverse older adults with multisite pain. Contemp Clin Trials 2023 May;128:107164. [doi: 10.1016/j.cct.2023.107164] [Medline: 36940813]
19. Aldana-Benítez D, Caicedo-Pareja MJ, Sánchez DP, Ordoñez-Mora LT. Dance as a neurorehabilitation strategy: a systematic review. J Bodyw Mov Ther 2023 Jul;35:348-363 [FREE Full text] [doi: 10.1016/j.jbmt.2023.04.046] [Medline: 37330793]

XSL•FO
RenderX

20. Jawad A, Baattaiah BA, Alharbi MD, Chevidikunnan MF, Khan F. Factors contributing to falls in people with multiple sclerosis: the exploration of the moderation and mediation effects. Mult Scler Relat Disord 2023 Aug;76:104838. [doi: 10.1016/j.msard.2023.104838] [Medline: 37390785]
21. Warren C, Rizo E, Decker E, Hasse A. A comprehensive analysis of risk factors associated with inpatient falls. J Patient Saf 2023 Oct 01;19(6):396-402. [doi: 10.1097/PTS.0000000000001123] [Medline: 37186671]
22. Gross M, Roigk P, Schoene D, Ritter Y, Pauly P, Becker C, Bundesinitiative Sturzprävention. [Update of the recommendations of the federal falls prevention initiative-identification and prevention of the risk of falling in older people living at home]. Z Gerontol Geriatr 2023 Oct 11;56(6):448-457. [doi: 10.1007/s00391-023-02229-5] [Medline: 37695366]
23. Li S, Li Y, Liang Q, Yang WJ, Zi R, Wu X, et al. Effects of tele-exercise rehabilitation intervention on women at high risk of osteoporotic fractures: study protocol for a randomised controlled trial. BMJ Open 2022 Nov 07;12(11):e064328 [FREE Full text] [doi: 10.1136/bmjopen-2022-064328] [Medline: 36344002]
24. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
25. Ye Y, Wu X, Wang H, Ye H, Zhao K, Yao S, et al. Artificial intelligence-assisted analysis for tumor-immune interaction within the invasive margin of colorectal cancer. Ann Med 2023 Dec;55(1):2215541 [FREE Full text] [doi: 10.1080/07853890.2023.2215541] [Medline: 37224471]
26. Kuwahara T, Hara K, Mizuno N, Haba S, Okuno N, Fukui T, et al. Current status of artificial intelligence analysis for the treatment of pancreaticobiliary diseases using endoscopic ultrasonography and endoscopic retrograde cholangiopancreatography. DEN Open 2024 Apr 30;4(1):e267 [FREE Full text] [doi: 10.1002/deo2.267] [Medline: 37397344]
27. Yokote A, Umeno J, Kawasaki K, Fujioka S, Fuyuno Y, Matsuno Y, et al. Small bowel capsule endoscopy examination and open access database with artificial intelligence: the SEE-artificial intelligence project. DEN Open 2024 Apr 22;4(1):e258 [FREE Full text] [doi: 10.1002/deo2.258] [Medline: 37359150]
28. Ramalingam M, Jaisankar A, Cheng L, Krishnan S, Lan L, Hassan A, et al. Impact of nanotechnology on conventional and artificial intelligence-based biosensing strategies for the detection of viruses. Discov Nano 2023 Dec 01;18(1):58 [FREE Full text] [doi: 10.1186/s11671-023-03842-4] [Medline: 37032711]
29. Yerukala Sathipati S, Tsai MJ, Shukla SK, Ho SY. Artificial intelligence-driven pan-cancer analysis reveals miRNA signatures for cancer stage prediction. HGG Adv 2023 Jul 13;4(3):100190 [FREE Full text] [doi: 10.1016/j.xhgg.2023.100190] [Medline: 37124139]
30. Liu J, Dan W, Liu X, Zhong X, Chen C, He Q, et al. Development and validation of predictive model based on deep learning method for classification of dyslipidemia in Chinese medicine. Health Inf Sci Syst 2023 Dec 06;11(1):21 [FREE Full text] [doi: 10.1007/s13755-023-00215-0] [Medline: 37035723]
31. Carou-Senra P, Ong JJ, Castro BM, Seoane-Viaño I, Rodríguez-Pombo L, Cabalar P, et al. Predicting pharmaceutical inkjet printing outcomes using machine learning. Int J Pharm X 2023 Dec;5:100181 [FREE Full text] [doi: 10.1016/j.ijpx.2023.100181] [Medline: 37143957]
32. Li X, Zhu Y, Zhao W, Shi R, Wang Z, Pan H, et al. Machine learning algorithm to predict the in-hospital mortality in critically ill patients with chronic kidney disease. Ren Fail 2023 Dec;45(1):2212790 [FREE Full text] [doi: 10.1080/0886022X.2023.2212790] [Medline: 37203863]
33. Bonnin M, Müller-Fouarge F, Estienne T, Bekadar S, Pouchy C, Ait Si Selmi T. Artificial intelligence radiographic analysis tool for total knee arthroplasty. J Arthroplasty 2023 Jul;38(7 Suppl 2):S199-207.e2. [doi: 10.1016/j.arth.2023.02.053] [Medline: 36858127]
34. Kao DP. Intelligent artificial intelligence: present considerations and future implications of machine learning applied to electrocardiogram interpretation. Circ Cardiovasc Qual Outcomes 2019 Sep;12(9):e006021 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.119.006021] [Medline: 31525077]
35. van der Stigchel B, van den Bosch K, van Diggelen J, Haselager P. Intelligent decision support in medical triage: are people robust to biased advice? J Public Health (Oxf) 2023 Aug 28;45(3):689-696 [FREE Full text] [doi: 10.1093/pubmed/fdad005] [Medline: 36947701]
36. Jakhar D, Kaur I. Artificial intelligence, machine learning and deep learning: definitions and differences. Clin Exp Dermatol 2020 Jan 09;45(1):131-132. [doi: 10.1111/ced.14029] [Medline: 31233628]
37. Ghosh M, Thirugnanam A. Introduction to artificial intelligence. In: Srinivasa KG, Siddesh GM, Sekhar SR, editors. Artificial Intelligence for Information Management: A Healthcare Perspective. Cham, Switzerland: Springer; 2021:88-44.
38. Taulli T. Artificial Intelligence Basics: A Non-Technical Introduction. Berkeley, CA: Apress Berkeley; 2019.
39. Patil S, Joda T, Soffe B, Awan KH, Fageeh HN, Tovani-Palone MR, et al. Efficacy of artificial intelligence in the detection of periodontal bone loss and classification of periodontal diseases: a systematic review. J Am Dent Assoc 2023 Sep;154(9):795-804.e1. [doi: 10.1016/j.adaj.2023.05.010] [Medline: 37452813]
40. Quek LJ, Heikkonen MR, Lau Y. Use of artificial intelligence techniques for detection of mild cognitive impairment: a systematic scoping review. J Clin Nurs 2023 Sep 10;32(17-18):5752-5762. [doi: 10.1111/jocn.16699] [Medline: 37032649]
41. Tan D, Mohd Nasir NF, Abdul Manan H, Yahya N. Prediction of toxicity outcomes following radiotherapy using deep learning-based models: a systematic review. Cancer Radiother 2023 Sep;27(5):398-406. [doi: 10.1016/j.canrad.2023.05.001] [Medline: 37482464]

XSL•FO
RenderX

42. Rabilloud N, Allaume P, Acosta O, De Crevoisier R, Bourgade R, Loussouarn D, et al. Deep learning methodologies applied to digital pathology in prostate cancer: a systematic review. Diagnostics (Basel) 2023 Aug 14;13(16):2676 [FREE Full text] [doi: 10.3390/diagnostics13162676] [Medline: 37627935]
43. Li K, Yao S, Zhang Z, Cao B, Wilson C, Kalos D, et al. Efficient gradient boosting for prognostic biomarker discovery. Bioinformatics 2022 Mar 04;38(6):1631-1638 [FREE Full text] [doi: 10.1093/bioinformatics/btab869] [Medline: 34978570]
44. Chen T, Chen Y, Li H, Gao T, Tu H, Li S. Driver intent-based intersection autonomous driving collision avoidance reinforcement learning algorithm. Sensors (Basel) 2022 Dec 16;22(24):9943 [FREE Full text] [doi: 10.3390/s22249943] [Medline: 36560308]
45. Huynh QT, Nguyen PH, Le HX, Ngo LT, Trinh NT, Tran MT, et al. Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. Diagnostics (Basel) 2022 Aug 03;12(8):1879 [FREE Full text] [doi: 10.3390/diagnostics12081879] [Medline: 36010229]
46. Brooke BS, Schwartz TA, Pawlik TM. MOOSE reporting guidelines for meta-analyses of observational studies. JAMA Surg 2021 Aug 01;156(8):787-788. [doi: 10.1001/jamasurg.2021.0522] [Medline: 33825847]
47. Scholten RJ, Clarke M, Hetherington J. The Cochrane collaboration. Eur J Clin Nutr 2005 Aug 28;59 Suppl 1(S1):S147-S196. [doi: 10.1038/sj.ejcn.1602188] [Medline: 16052183]
48. Warrens MJ. Kappa coefficients for dichotomous-nominal classifications. Adv Data Anal Classif 2020 Apr 07;15(1):193-208. [doi: 10.1007/s11634-020-00394-8]
49. Law M, Stewart D, Letts L, Pollock N, Bosch J. Guidelines for critical review of qualitative studies. McMaster University Occupational Therapy Evidence-Based Practice Research Group. URL: https://www.canchild.ca/system/tenon/assets/attachments/000/000/360/original/qualguide.pdf [accessed 2024-04-05]
50. Higgins JP, Morgan RL, Rooney AA, Taylor KW, Thayer KA, Silva RA, et al. Risk of bias in non-randomized studies - of exposure (ROBINS-E). ROBINS-E tool. URL: https://www.riskofbias.info/welcome/robins-e-tool [accessed 2024-04-02]
51. Nait Aicha A, Englebienne G, van Schooten KS, Pijnappels M, Kröse B. Deep learning to predict falls in older adults based on daily-life trunk accelerometry. Sensors (Basel) 2018 May 22;18(5):1654 [FREE Full text] [doi: 10.3390/s18051654] [Medline: 29786659]
52. Ribeiro NF, André J, Costa L, Santos CP. Development of a strategy to predict and detect falls using wearable sensors. J Med Syst 2019 Apr 04;43(5):134. [doi: 10.1007/s10916-019-1252-2] [Medline: 30949770]
53. Qiu H, Rehman RZ, Yu X, Xiong S. Application of wearable inertial sensors and a new test battery for distinguishing retrospective fallers from non-fallers among community-dwelling older people. Sci Rep 2018 Nov 05;8(1):16349 [FREE Full text] [doi: 10.1038/s41598-018-34671-6] [Medline: 30397282]
54. Dubois A, Bihl T, Bresciani JP. Automatic measurement of fall risk indicators in timed up and go test. Inform Health Soc Care 2019 Sep 13;44(3):237-245. [doi: 10.1080/17538157.2018.1496089] [Medline: 30102095]
55. Noh B, Youm C, Goh E, Lee M, Park H, Jeon H, et al. XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes. Sci Rep 2021 Jun 09;11(1):12183. [doi: 10.1038/s41598-021-91797-w] [Medline: 34108595]
56. Hauth J, Jabri S, Kamran F, Feleke EW, Nigusie K, Ojeda LV, et al. Automated loss-of-balance event identification in older adults at risk of falls during real-world walking using wearable inertial measurement units. Sensors (Basel) 2021 Jul 07;21(14):4661 [FREE Full text] [doi: 10.3390/s21144661] [Medline: 34300399]
57. Lockhart TE, Soangra R, Yoon H, Wu T, Frames CW, Weaver R. Prediction of fall risk among community-dwelling older adults using a wearable system. Sci Rep 2021;11(1):20976. [doi: 10.21203/rs.3.rs-142249/v1]
58. Hsu YC, Weng HH, Kuo CY, Chu TP, Tsai YH. Prediction of fall events during admission using eXtreme gradient boosting: a comparative validation study. Sci Rep 2020 Oct 08;10(1):16777 [FREE Full text] [doi: 10.1038/s41598-020-73776-9] [Medline: 33033326]
59. Hu Y, Bishnoi A, Kaur R, Sowers R, Hernandez ME. Exploration of machine learning to identify community dwelling older adults with balance dysfunction using short duration accelerometer data. In: Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2020 Presented at: EMBC '20; July 20-24, 2020; Montreal, QC p. 812-815 URL: https://ieeexplore.ieee.org/document/9175871 [doi: 10.1109/embc44109.2020.9175871]
60. Ye C, Li J, Hao S, Liu M, Jin H, Zheng L, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int J Med Inform 2020 May;137:104105 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104105] [Medline: 32193089]
61. Althobaiti T, Katsigiannis S, Ramzan N. Triaxial accelerometer-based falls and activities of daily life detection using machine learning. Sensors (Basel) 2020 Jul 06;20(13):3777 [FREE Full text] [doi: 10.3390/s20133777] [Medline: 32640526]
62. Tunca C, Salur G, Ersoy C. Deep learning for fall risk assessment with inertial sensors: utilizing domain knowledge in spatio-temporal gait parameters. IEEE J Biomed Health Inform 2020 Jul;24(7):1994-2005. [doi: 10.1109/jbhi.2019.2958879]
63. Kim K, Yun G, Park SK, Kim DH. Fall detection for the elderly based on 3-axis accelerometer and depth sensor fusion with random forest classifier. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2019 Presented at: EMBC '19; July 23-27, 2019; Berlin, Germany p. 4611-4614 URL: https://ieeexplore.ieee.org/document/8856698 [doi: 10.1109/embc.2019.8856698]

XSL•FO
RenderX

64. Gillain S, Boutaayamou M, Schwartz C, Brüls O, Bruyère O, Croisier JL, et al. Using supervised learning machine algorithm to identify future fallers based on gait patterns: a two-year longitudinal study. Exp Gerontol 2019 Nov;127:110730 [FREE Full text] [doi: 10.1016/j.exger.2019.110730] [Medline: 31520696]
65. Lo Y, Lynch SF, Urbanowicz RJ, Olson RS, Ritter AZ, Whitehouse CR, et al. Using machine learning on home health care assessments to predict fall risk. Stud Health Technol Inform 2019 Aug 21;264:684-688. [doi: 10.3233/SHTI190310] [Medline: 31438011]
66. Maray N, Ngu AH, Ni J, Debnath M, Wang L. Transfer learning on small datasets for improved fall detection. Sensors (Basel) 2023 Jan 18;23(3):1105 [FREE Full text] [doi: 10.3390/s23031105] [Medline: 36772148]
67. Ladios-Martin M, Cabañero-Martínez MJ, Fernández-de-Maya J, Ballesta-López FJ, Belso-Garzas A, Zamora-Aznar FM, et al. Development of a predictive inpatient falls risk model using machine learning. J Nurs Manag 2022 Nov 30;30(8):3777-3786. [doi: 10.1111/jonm.13760] [Medline: 35941786]
68. Eichler N, Raz S, Toledano-Shubi A, Livne D, Shimshoni I, Hel-Or H. Automatic and efficient fall risk assessment based on machine learning. Sensors (Basel) 2022 Feb 17;22(4):1557 [FREE Full text] [doi: 10.3390/s22041557] [Medline: 35214471]
69. Tang YM, Wang YH, Feng XY, Zou QS, Wang Q, Ding J, et al. Diagnostic value of a vision-based intelligent gait analyzer in screening for gait abnormalities. Gait Posture 2022 Jan;91:205-211. [doi: 10.1016/j.gaitpost.2021.10.028] [Medline: 34740057]
70. Greene BR, McManus K, Ader LG, Caulfield B. Unsupervised assessment of balance and falls risk using a smartphone and machine learning. Sensors (Basel) 2021 Jul 13;21(14):4770 [FREE Full text] [doi: 10.3390/s21144770] [Medline: 34300509]
71. Roshdibenam V, Jogerst GJ, Butler NR, Baek S. Machine learning prediction of fall risk in older adults using timed up and go test kinematics. Sensors (Basel) 2021 May 17;21(10):3481 [FREE Full text] [doi: 10.3390/s21103481] [Medline: 34067644]
72. Dubois A, Bihl T, Bresciani JP. Identifying fall risk predictors by monitoring daily activities at home using a depth sensor coupled to machine learning algorithms. Sensors (Basel) 2021 Mar 11;21(6):1957 [FREE Full text] [doi: 10.3390/s21061957] [Medline: 33799526]
73. Vo MT, Thonglor R, Moncatar TJ, Han TD, Tejativaddhana P, Nakamura K. Fear of falling and associated factors among older adults in Southeast Asia: a systematic review. Public Health 2023 Sep;222:215-228. [doi: 10.1016/j.puhe.2022.08.012] [Medline: 36229238]
74. Torun E, Az A, Akdemir T, Solakoğlu GA, Açiksari K, Güngörer B. Evaluation of the risk factors for falls in the geriatric population presenting to the emergency department. Ulus Travma Acil Cerrahi Derg 2023 Aug;29(8):897-903 [FREE Full text] [doi: 10.14744/tjtes.2023.07433] [Medline: 37563898]
75. Son NK, Ryu YU, Jeong HW, Jang YH, Kim HD. Comparison of 2 different exercise approaches: Tai Chi versus Otago, in community-dwelling older women. J Geriatr Phys Ther 2016;39(2):51-57. [doi: 10.1519/JPT.0000000000000042] [Medline: 25760277]
76. Sawa R, Doi T, Tsutsumimoto K, Nakakubo S, Kurita S, Kiuchi Y, et al. Overlapping status of frailty and fear of falling: an elevated risk of incident disability in community-dwelling older adults. Aging Clin Exp Res 2023 Sep 11;35(9):1937-1944 [FREE Full text] [doi: 10.1007/s40520-023-02474-z] [Medline: 37337077]
77. Calazans JA, Permanyer I. Levels, trends, and determinants of cause-of-death diversity in a global perspective: 1990-2019. BMC Public Health 2023 Apr 05;23(1):650 [FREE Full text] [doi: 10.1186/s12889-023-15502-4] [Medline: 37016370]
78. Kakara R, Bergen G, Burns E, Stevens M. Nonfatal and fatal falls among adults aged ≥65 years - United States, 2020-2021. MMWR Morb Mortal Wkly Rep 2023 Sep 01;72(35):938-943 [FREE Full text] [doi: 10.15585/mmwr.mm7235a1] [Medline: 37651272]
79. Dostan A, Dobson CA, Vanicek N. Relationship between stair ascent gait speed, bone density and gait characteristics of postmenopausal women. PLoS One 2023 Mar 22;18(3):e0283333 [FREE Full text] [doi: 10.1371/journal.pone.0283333] [Medline: 36947573]
80. Zheng Y, Wang X, Zhang ZK, Guo B, Dang L, He B, et al. Bushen Yijing Fang reduces fall risk in late postmenopausal women with osteopenia: a randomized double-blind and placebo-controlled trial. Sci Rep 2019 Feb 14;9(1):2089 [FREE Full text] [doi: 10.1038/s41598-018-38335-3] [Medline: 30765762]
81. Woelfle T, Bourguignon L, Lorscheider J, Kappos L, Naegelin Y, Jutzeler CR. Wearable sensor technologies to assess motor functions in people with multiple sclerosis: systematic scoping review and perspective. J Med Internet Res 2023 Jul 27;25:e44428 [FREE Full text] [doi: 10.2196/44428] [Medline: 37498655]
82. Abdollah V, Dief TN, Ralston J, Ho C, Rouhani H. Investigating the validity of a single tri-axial accelerometer mounted on the head for monitoring the activities of daily living and the timed-up and go test. Gait Posture 2021 Oct;90:137-140. [doi: 10.1016/j.gaitpost.2021.08.020] [Medline: 34481263]
83. Mielke GI, de Almeida Mendes M, Ekelund U, Rowlands AV, Reichert FF, Crochemore-Silva I. Absolute intensity thresholds for tri-axial wrist and waist accelerometer-measured movement behaviors in adults. Scand J Med Sci Sports 2023 Sep 12;33(9):1752-1764. [doi: 10.1111/sms.14416] [Medline: 37306308]

XSL•FO
RenderX

84. Löppönen A, Delecluse C, Suorsa K, Karavirta L, Leskinen T, Meulemans L, et al. Association of sit-to-stand capacity and free-living performance using Thigh-Worn accelerometers among 60- to 90-yr-old adults. Med Sci Sports Exerc 2023 Sep 01;55(9):1525-1532 [FREE Full text] [doi: 10.1249/MSS.0000000000003178] [Medline: 37005494]
85. García-Soidán JL, Leirós-Rodríguez R, Romo-Pérez V, García-Liñeira J. Accelerometric assessment of postural balance in children: a systematic review. Diagnostics (Basel) 2020 Dec 22;11(1):8 [FREE Full text] [doi: 10.3390/diagnostics11010008] [Medline: 33375206]
86. Leirós-Rodríguez R, García-Soidán JL, Romo-Pérez V. Analyzing the use of accelerometers as a method of early diagnosis of alterations in balance in elderly people: a systematic review. Sensors (Basel) 2019 Sep 09;19(18):3883 [FREE Full text] [doi: 10.3390/s19183883] [Medline: 31505828]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
LSTM: long short-term memory
MeSH: Medical Subject Headings
RF: random forest
ROBINS-E: Risk of Bias in Nonrandomized Studies of Exposures

Edited by A Mavragani; submitted 28.11.23; peer-reviewed by E Andrade, M Behzadifar, A Suárez; comments to author 09.01.24; revised version received 30.01.24; accepted 13.02.24; published 29.04.24.

Please cite as:
González-Castro A, Leirós-Rodríguez R, Prada-García C, Benítez-Andrades JA
The Applications of Artificial Intelligence for Assessing Fall Risk: Systematic Review
J Med Internet Res 2024;26:e54934
URL: https://www.jmir.org/2024/1/e54934
doi:10.2196/54934
PMID:38684088

©Ana González-Castro, Raquel Leirós-Rodríguez, Camino Prada-García, José Alberto Benítez-Andrades. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Suitability of the Current Health Technology Assessment of Innovative Artificial Intelligence-Based Medical Devices: Scoping Literature Review

Line Farah[1,2*], PharmD; Isabelle Borget[2,3,4], PharmB, PhD; Nicolas Martelli[2,5], PharmD, PhD; Alexandre Vallee[6*], MD, PhD

[1]Innovation Center for Medical Devices Department, Foch Hospital, Suresnes, France
[2]Groupe de Recherche et d'accueil en Droit et Economie de la Santé Department, University Paris-Saclay, Orsay, France
[3]Department of Biostatistics and Epidemiology, Gustave Roussy, University Paris-Saclay, Villejuif, France
[4]Oncostat U1018, Inserm, Équipe Labellisée Ligue Contre le Cancer, University Paris-Saclay, Villejuif, France
[5]Pharmacy Department, Georges Pompidou European Hospital, Paris, France
[6]Department of Epidemiology and Public Health, Foch Hospital, Suresnes, France
[*]these authors contributed equally

Corresponding Author:
Line Farah, PharmD
Innovation Center for Medical Devices Department
Foch Hospital
40 rue Worth
Suresnes, 92150
France
Phone: 33 952329655
Email: line.farah1@gmail.com

Abstract

Background: Artificial intelligence (AI)–based medical devices have garnered attention due to their ability to revolutionize medicine. Their health technology assessment framework is lacking.

Objective: This study aims to analyze the suitability of each health technology assessment (HTA) domain for the assessment of AI-based medical devices.

Methods: We conducted a scoping literature review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology. We searched databases (PubMed, Embase, and Cochrane Library), gray literature, and HTA agency websites.

Results: A total of 10.1% (78/775) of the references were included. Data quality and integration are vital aspects to consider when describing and assessing the technical characteristics of AI-based medical devices during an HTA process. When it comes to implementing specialized HTA for AI-based medical devices, several practical challenges and potential barriers could be highlighted and should be taken into account (AI technological evolution timeline, data requirements, complexity and transparency, clinical validation and safety requirements, regulatory and ethical considerations, and economic evaluation).

Conclusions: The adaptation of the HTA process through a methodological framework for AI-based medical devices enhances the comparability of results across different evaluations and jurisdictions. By defining the necessary expertise, the framework supports the development of a skilled workforce capable of conducting robust and reliable HTAs of AI-based medical devices. A comprehensive adapted HTA framework for AI-based medical devices can provide valuable insights into the effectiveness, cost-effectiveness, and societal impact of AI-based medical devices, guiding their responsible implementation and maximizing their benefits for patients and health care systems.

(J Med Internet Res 2024;26:e51514) doi:10.2196/51514

KEYWORDS

artificial intelligence; machine learning; health technology assessment; medical devices; evaluation

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) has emerged as a transformative technology with vast potential across various sectors, including health care [1,2]. In this field, AI-based medical devices have garnered significant attention due to their ability to revolutionize diagnosis, treatment, and patient monitoring [3]. These devices use advanced algorithms and machine learning techniques to analyze complex medical data sets, thereby providing valuable insights and support to health care professionals in their decision-making. To meet the ever-increasing demand for integration of AI-based medical devices into clinical practice, their efficient evaluation through adapted health technology assessment (HTA) is now crucial [4]. Several frameworks have been published showing an adaptation of items for reporting clinical trials related to AI-based medical devices (Table 1). However, there is still a lack of a much needed adaptation of the standard HTA framework to better suit the assessment of AI-based health care technologies.

Table 1. Summary of different frameworks on artificial intelligence (AI)–based health technologies.

Framework	Function	Definition and aim	Domains and items assessed	Reference
HTA[a] Core Model	Medical device assessment	A systematic evaluation of a medical technology's clinical, economic, ethical, and social implications to determine its overall value and impact on health care delivery; a methodological framework for production and sharing of HTA information	It evaluates nine domains: (1) health problem and current use of technology, (2) description and technical characteristics of the technology, (3) safety, (4) clinical effectiveness, (5) costs and economic evaluation, (6) ethical analysis, (7) organizational aspects, (8) patients and social aspects, and (9) legal aspects	European Network for Health Technology Assessment [5]
CONSORT-AI[b]	Trial reporting	A set of recommendations for clinical trial reports evaluating interventions with an AI component	It evaluates the 25 CONSORT[c] 2010 items+14 AI-specific extension items	Liu et al [6]
SPIRIT-AI[d]	Trial reporting	New reporting guidelines for clinical trial protocols evaluating interventions with an AI component	It evaluates the 33 SPIRIT[e] 2013 items+ 15 AI-specific extension items	Rivera et al [7]
TRIPOD-AI[f]	Research reporting	Reports of research or endeavors in which a multivariable prediction model is being developed (or updated) or validated (tested) using any (supervised) machine learning technique	It evaluates the 22 TRIPOD[g] 2013 items and the AI extension item definitions	Collins et al [8]

[a]HTA: health technology assessment.

[b]CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence.

[c]CONSORT: Consolidated Standards of Reporting Trials.

[d]SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence.

[e]SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials.

[f]TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Artificial Intelligence.

[g]TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

An HTA involves a systematic evaluation of a medical technology's clinical, economic, ethical, and social implications to determine its overall value and impact on health care delivery [9]. The European Network for Health Technology Assessment has designed an HTA Core Model that provides a methodological framework for production and sharing of HTA information [5]. It evaluates the following nine domains: (1) health problem and current use of technology, (2) description and technical characteristics of the technology, (3) safety, (4) clinical effectiveness, (5) costs and economic evaluation, (6) ethical analysis, (7) organizational aspects, (8) patients and social aspects, and (9) legal aspects (Multimedia Appendix 1).

A full understanding of the capabilities, limitations, and specificities of AI-based medical devices is paramount to complete these HTA domain assessments and, thereby, inform evidence-based decision-making and allow for policy development and the responsible integration of these technologies into health care systems [10,11]. Much uncertainty remains with regard to the reliability of AI-based medical devices, data issues, and regulatory processes, resulting in multiple challenges faced by HTA agencies assessing new technologies and delivering their approval [4]. Nevertheless, AI-based medical devices require strict regulations and specific legislations [12-14]. Over the last few decades, the evaluation of AI-based medical devices through an HTA process has received growing interest, as shown by the increase from 1 published article in 1990 to 94 in December 2023, with 484 articles in total during the period 1990-2023 (Table 2).

XSL•FO
RenderX

Table 2. Occurrences of "technology assessment" AND "medical device" AND "artificial intelligence" from inception to December 2023 in the PubMed literature.

Year	Articles (n=484), n (%)
1990-1999	5 (1%)
2000-2009	44 (9%)
2010-2019	151 (32%)
2020-2023	273 (58%)

Objectives

The objective of this review was to critically assess the comprehensive suitability of the current HTA process for AI-based medical devices. By evaluating the performance and capabilities of AI-based medical devices across multiple dimensions, this review aimed to inform health care professionals, policy makers, and researchers about the challenges and opportunities associated with these technologies. Ultimately, this review sought to facilitate evidence-based decision-making; promote responsible implementation; and maximize the potential benefits of AI-based medical devices in improving health care quality, accessibility, and outcomes.

To this end, we analyzed the suitability of each HTA domain for the assessment of AI-based medical devices and proposed an adapted HTA framework.

Methods

Search Strategies

A scoping literature review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology [15,16]. All articles related to HTA methods for AI-based medical devices were selected. Data extraction focused on assessment criteria, methodological evaluations, and results.

The search terms and strategy outlined in Textbox 1 were used for the scoping review.

Textbox 1. Search strategies and Medical Subject Heading (MeSH) terms used for the scoping literature review.

Database and search strategy

- PubMed
  - Health technology assessment
    - "technology assessment, biomedical" [MeSH term] OR ("technology" [all fields] AND "assessment" [all fields] AND "biomedical" [all fields]) OR "biomedical technology assessment" [all fields] OR ("health" [all fields] AND "technology" [all fields] AND "assessment" [all fields]) OR "health technology assessment" [all fields]
  - Medical device
    - "equipment and supplies" [MeSH term] OR ("equipment" [all fields] AND "supplies" [all fields]) OR "equipment and supplies" [all fields] OR ("medical" [all fields] AND "device" [all fields]) OR "medical device" [all fields]
  - Artificial intelligence
    - "artificial intelligence" [MeSH term] OR ("artificial" [all fields] AND "intelligence" [all fields]) OR "artificial intelligence" [all fields]
- Embase
  - ("artificial intelligence" OR "artificial intelligence-based medical device") AND "technology assessment"
- Cochrane Library
  - (technology assessment) and (artificial intelligence)

Databases

We searched multiple databases: PubMed, Embase, and the Cochrane Library. Additional articles were retrieved manually from the gray literature and from HTA agency websites (Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices in Germany], Haute Autorité de Santé [French Health Authority], National Health Service, National Institute for Health and Care Excellence, National Institute of Public Health, International Network of Agencies for Health Technology Assessment, and the European Network for Health Technology Assessment).

Study Selection Process

The studies were selected by 2 reviewers (LF and AV). After removing duplicates, both reviewers independently screened the abstracts to select eligible articles and then analyzed full-text reports for eligibility. A third party (NM) resolved the possible

XSL•FO
RenderX

discrepancies highlighted during the selection process should a consensus not be reached. An extraction database was used to list the selected articles meeting the inclusion criteria to ensure that all eligible articles were included.

The inclusion and exclusion criteria were related to (1) language (this review was limited to the English and French languages), (2) article type (reviews and primary research were included, whereas other article types were excluded, such as abstracts, commentaries, editorials, letters to the editor, case reports, case series, animal studies, phase-1 and phase-2 studies, pilot studies, duplicate studies, irrelevant studies, studies with the wrong aim, and studies available in abstract form only), (3) type of technology (only AI-based medical devices were eligible), and (4) type of evaluation (HTA articles were included).

Data Extraction

In total, 2 analysts (LF and AV) independently extracted data items from the selected articles. A third party was involved to resolve any discrepancies highlighted during the selection process. The following data were extracted from each article: (1) general characteristics of the studies (authors, journal, and publication date), (2) study objective, and (3) HTA assessment domain for AI-based medical devices related to the article.

Methodological Quality Appraisal

Neither the methodological quality nor the risk of bias of the included articles were assessed, consistent with scoping review guidance [15].

Application Example of the HTA of an AI-Based Medical Device

Concerning each HTA domain, we added a case study of the HTA of several AI-based medical devices in diabetes to illustrate our recommendations.

Results

Scoping Review Results

The literature search resulted in 775 citations summarized in the PRISMA flowchart (Figure 1). After removal of 32.6% (253/775) of duplicates, we carried out an initial screening of the remaining 522 publications, resulting in 129 (24.7%) potentially relevant full-text papers. After further screening, we assessed 77 reports for eligibility, based on which we excluded 11 (14%) for not referring to AI-based medical devices and 2 (3%) for not being HTAs. We then included a further 14 records from 59 potential citations available from HTA agency websites and gray literature.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process. AI: artificial intelligence; HTA: health technology assessment; MD: medical device.

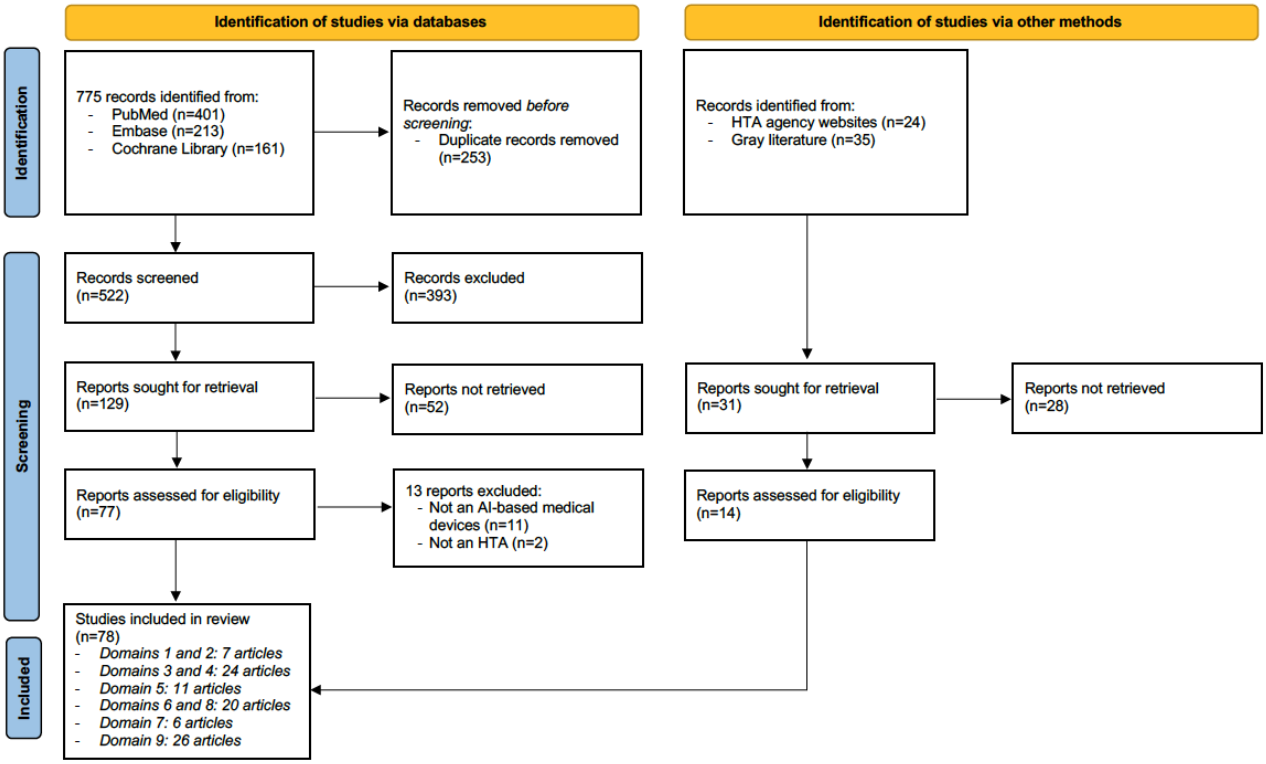

This gave a total of 78 included articles covering one or multiple HTA domains, the distribution of which is provided at the bottom of Figure 1.

We then summarized the data collection for each selected article in Multimedia Appendix 2 [1,2,4,9,10,17-89].

Suitability of Each HTA Domain for the Assessment of AI-Based Medical Devices

As domain 1, health problem and current use of technology, is suitable for any type of medical technology and systematically addressed these technologies, we focused on the suitability of the HTA domains (from 1 to 9).

XSL•FO
RenderX

Domains 1 and 2 (Health Problem, Current Use, Description, and Technical Characteristics of the Technology)

Data quality and integration are vital aspects to consider when describing and assessing the technical characteristics of AI-based medical devices during an HTA process [17]. These devices often rely on accessing and analyzing diverse health care data sources, including electronic health records, medical images, genetic data, and wearable device data. Therefore, it is essential to evaluate their ability to seamlessly integrate and exchange data with existing health care systems [18].

First, interoperability refers to the ability of AI-based medical devices to interact and communicate with other health care technologies and systems [10]. This includes the ability to access and use data from different sources, such as laboratory systems, imaging archives, and patient health records. The assessment should consider whether the devices adhere to relevant data standards and protocols, ensuring efficient data exchange and compatibility with existing health care infrastructures [19].

Second, data integration involves the ability of AI-based medical devices to aggregate and analyze data from multiple sources to provide comprehensive and accurate insights [20]. The HTA should assess whether the devices can handle different data types, formats, and resolutions and whether they can effectively integrate and harmonize data from disparate sources.

In addition, data privacy and security considerations are crucial when evaluating AI-based medical devices [21]. The assessment should examine whether the devices comply with relevant data protection regulations, use appropriate data anonymization and encryption techniques, and have robust security measures in place to protect patient information [22].

Therefore, a comprehensive HTA should address both the interoperability and data integration capabilities of AI-based medical devices, ensuring that they can seamlessly interact with existing health care systems and integrate data from multiple sources and that they adhere to data privacy and security standards. By evaluating all these aspects, the HTA could determine the devices' feasibility, scalability, and potential impact on health care delivery.

Domain 3 (Safety) and Domain 4 (Clinical Effectiveness)

The assessment of clinical effectiveness and the impact on patient outcomes is a crucial aspect when evaluating AI-based medical devices through a comprehensive HTA [4]. While accuracy and performance metrics are important, it is essential to determine how these devices translate into tangible benefits for patients and health care delivery [23].

Clinical utility refers to the extent to which the AI-based medical device improves clinical decision-making, patient outcomes, and health care processes [24]. The HTA should examine whether the device provides actionable and reliable information that potentially helps health care professionals make more accurate diagnoses or improve treatment plans or monitoring strategies [25,26]. It should also evaluate the potential impact of the device on patient outcomes by assessing, for example, improved survival rates, reduced complications, or enhanced quality of life [27].

To assess clinical utility, the HTA should consider the device's performance in relevant clinical scenarios and its ability to address specific clinical questions or challenges. This may involve evaluating the device's performance against established clinical guidelines or expert opinions as well as considering the device's potential to fill gaps in clinical practice or enhance existing diagnostic or treatment methods.

Furthermore, the HTA should examine the broader impact of AI-based medical devices on health care systems and resource allocation [21]. This includes evaluating the devices' potential to optimize resource use, reduce health care costs, or improve workflow efficiency. Economic evaluations such as cost-effectiveness analyses can provide insights into the value for money and long-term cost savings associated with the adoption of these devices [28,29].

A comprehensive HTA should thoroughly assess the clinical effectiveness and impact on patient outcomes of AI-based medical devices, examining their performance in relevant clinical scenarios, their alignment with clinical guidelines, and their potential to improve health care processes and resource allocation. By evaluating these aspects, the HTA can provide a holistic understanding of the AI devices' effectiveness and their potential to positively transform health care delivery [21].

AI-based medical devices must undergo rigorous clinical validation to assess their performance and reliability in real-world health care settings [30,31]. Clinical validation involves evaluating the device's accuracy, sensitivity, specificity, and overall diagnostic or prognostic performance. In addition, the devices should be tested across diverse patient populations and compared against gold-standard reference methods or expert opinions [32].

Furthermore, the use of real-world evidence (RWE) is crucial for a comprehensive HTA [33]. RWE involves gathering data from routine clinical practice, electronic health records, and other sources to evaluate the device's effectiveness and safety in real-world settings [34]. These data can help assess the device's performance in a broader patient population and identify any potential limitations or biases that may arise in specific clinical scenarios.

Robust clinical validation studies and the integration of RWE provide critical evidence for the evaluation of AI-based medical devices in an HTA process [4]. These studies should include a sufficient sample size, appropriate study design, and statistical analysis to ensure the validity and generalizability of the results [35-38]. By considering clinical validation and RWE adapted to AI-based medical devices, HTA can provide valuable insights into their clinical utility and impact in real-world health care settings, facilitating evidence-based decision-making for their adoption and use.

A lack of safety evaluation was highlighted in a systematic review, showing that only 9% of AI-based medical device studies evaluated safety criteria [4]. However, safety is crucial for the confidence in and adoption of AI-based technologies for both patients and health care professionals [39]. While these

XSL•FO
RenderX

devices have the potential to improve diagnosis and treatment outcomes, their integration into clinical practice must be accompanied by robust safety evaluations [2,40]. Identifying and mitigating potential risks, such as algorithmic bias, data privacy breaches, and algorithm failures, is essential to protect patient well-being and maintain trust in these innovative technologies [41].

To ensure the safety of AI-based medical devices, it is crucial to establish standardized evaluation frameworks and guidelines [39]. These frameworks should encompass rigorous testing methodologies, validation procedures, and continuous monitoring of AI performance. Collaboration among stakeholders, including manufacturers, regulatory agencies, health care providers, and researchers, is essential to develop and implement comprehensive safety evaluation protocols. Addressing the safety gap in AI-based medical devices not only ensures patient welfare but also instills confidence in health care professionals to embrace and use these technologies effectively [40]. By prioritizing safety evaluation, we can unleash the full potential of AI-based medical devices, leading to transformative advancements in health care delivery [1].

Domain 5 (Costs and Economic Evaluation)

Costs and economic evaluation play a crucial role in the comprehensive assessment of health technology adoption, particularly in the context of the HTA of AI-based medical devices [4,9,42]. While AI technologies show huge potential to improve health care delivery, reduce time to diagnosis, and save money, their economic impact requires careful consideration to ensure their successful integration and sustainability [23].

When assessing AI-based medical devices through the HTA process, economic evaluations involve analyzing the costs and benefits associated with the adoption and use of the technology [4]. These evaluations go beyond the upfront costs of acquiring the AI-based medical device; they encompass various factors, such as training, infrastructure modifications, maintenance, and ongoing operational costs [43]. Cost analysis also includes improved patient outcomes, reduced hospital readmissions, and shortened hospital stays. By conducting economic evaluations, decision makers can gauge the cost-effectiveness and cost utility to weigh the affordability of AI-based medical devices for their particular use against the potential mid- and long-term cost savings considered [44].

HTA evaluates the suitability of AI-based medical devices by considering their potential economic benefits, risks, and ethical implications [45,46]. By incorporating economic evaluation into the HTA process, decision makers can make evidence-based choices about the allocation of limited health care resources and prioritize interventions that provide the greatest value for money [4,47].

Finally, considering cost savings and conducting economic evaluations as part of the HTA process helps facilitate the adoption of AI-based medical devices [37]. It provides decision makers with the necessary information to determine the financial feasibility and potential return on investment associated with implementing these technologies. HTA ensures that AI-based medical devices are suitable for integration into the health care system, fostering confidence in their effectiveness, efficiency, and cost utility [4,43,48].

Domain 6 (Ethical Analysis) and Domain 8 (Patients and Social Aspects)

Ethical and societal implications are critical aspects that must be considered when evaluating the suitability of AI-based medical devices for a comprehensive HTA [45]. The integration of AI technologies into health care raises important ethical concerns that need to be addressed to ensure responsible and equitable use [49].

Data privacy and patient consent are primary ethical considerations [50]. The HTA should evaluate whether AI-based medical devices adhere to strict data protection regulations, maintain patient confidentiality, and obtain appropriate informed consent for data use. It is essential to assess whether the devices have mechanisms in place to handle sensitive patient information securely and protect it against unauthorized access or data breaches.

Algorithmic fairness and bias are additional ethical concerns [51]. AI algorithms can inadvertently perpetuate biases present in the data used for training, resulting in unequal treatment or access to health care resources [52]. The HTA should assess whether the devices have been evaluated for fairness and bias and consider the steps taken to mitigate any identified biases [53].

Moreover, the HTA should examine the impact of AI-based medical devices on health care disparities and access to care [54]. It is crucial to assess whether the devices have the potential to exacerbate existing inequalities or whether they can contribute to reducing disparities by improving health care access, particularly for underserved populations.

Accountability and transparency in AI decision-making processes are also important ethical considerations [55]. The HTA should evaluate whether the AI devices provide clear explanations for their outputs and ensure that health care professionals and patients can understand and challenge the device's recommendations [21].

A comprehensive HTA should thoroughly examine the ethical and societal implications of AI-based medical devices, ensuring that they prioritize patient privacy, fairness, and equitable access to care; fostering trust; and ensuring that these devices align with societal values and goals.

To ensure the adoption of AI-based medical devices through a truthfully AI concept, explainability, interpretability, and transparency are important considerations for their comprehensive HTA [56]. Indeed, these devices often use complex algorithms and machine and deep learning techniques, resulting in their operating as "black boxes" that reach decisions and recommendations that are challenging to understand [55]. However, explainability and transparency are essential to ensure trust, accountability, and acceptance of AI-based technologies in health care. First, Explainability refers to the ability to understand and interpret the reasoning behind the device's outputs [56]. It involves providing clinicians and users with

XSL•FO
RenderX

transparent explanations of how the AI algorithm processes input data and generates results. This enables health care professionals to trust the device's recommendations and make informed decisions based on the provided information. The HTA should evaluate the extent to which AI-based medical devices can provide interpretable and understandable explanations of their decision-making processes [17,46]. Second, Transparency involves disclosing important information about the AI-based medical devices, including the data used for training, algorithmic methodologies used, and potential limitations or biases [57]. Transparency promotes trust and allows stakeholders to assess the device's reliability and potential risks. The HTA should assess whether the device manufacturers provide clear documentation and information to health care professionals and patients about the device's capabilities, limitations, and potential errors [21,58,59]. Transparency is closely linked to regulatory considerations [60,61]. The HTA should consider whether the device complies with relevant regulatory standards and whether the manufacturers have provided the necessary documentation and evidence to support their ethical claims [21,62,63].

A comprehensive HTA process should address concerns regarding the explainability and transparency of AI-based medical devices, trust, accountability, and ethical implications related to the use of AI in health care [45,46].

Domain 7 (Organizational Aspects)

Organizational aspects play a crucial role in the effective integration of AI-based medical devices into health care systems [64]. To ensure consistency, transparency, and comparability in the evaluation process, there is a growing need for a robust methodological framework that provides standardized guidelines for assessing the organizational impact of the implementation of AI-based medical devices [65]. According to a descriptive analysis led in German hospitals, the main barriers to AI-based medical device adoption were lack of resources (staff, knowledge, and financial) [66].

Clear indicators are needed to measure the organizational readiness for and impact of AI-based medical devices [4,10,67]. Several criteria have been highlighted, such as (1) health care workplace readiness and stakeholder acceptance, (2) AI-based medical device organization alignment assessment, and (3) business plan (financing and investments) [64].

Domain 9 (Legal Aspects)

The roles and responsibilities of health care professionals are also impacted by these AI solutions [68]. It is necessary to evaluate whether they complement or replace health care professionals' expertise and whether additional training, supervision, or support is required for their optimal use [69,70]. As AI technologies could evolve and become more prevalent in health care, it is crucial to ensure that these devices comply with existing legal frameworks and regulations [49,71]. HTA plays a pivotal role in evaluating the legal implications of AI-based medical devices by assessing factors such as data privacy, security, liability, and regulatory compliance [72].

One key legal aspect to consider is data privacy and protection [73,74]. AI-based medical devices often rely on vast amounts of patient data for training and decision-making. Therefore, it is essential to evaluate whether these devices adhere to relevant data protection laws, such as the General Data Protection Regulation in the European Union [75,76]. HTA examines the measures taken by AI-based medical device manufacturers to safeguard patient privacy, including data anonymization, encryption, and secure data storage practices [37]. These issues are highlighted in the Artificial Intelligence Act published in June 2023 by the European Commission [77,78].

Cybersecurity is another critical consideration [21,79]. As AI-based medical devices handle sensitive patient data and make critical health care decisions, it is crucial to assess the security measures implemented to prevent unauthorized access, data breaches, or tampering [80]. HTA evaluates the robustness of the security protocols implemented by device manufacturers and their compliance with industry standards and regulations [21,81].

Liability is also a significant legal aspect to be addressed in the context of AI-based medical devices [82,83]. When errors or adverse events occur due to the use of these devices, determining liability can be complex [84]. HTA examines the legal frameworks and liability guidelines pertaining to AI technologies, including whether clear guidelines exist regarding the responsibility of manufacturers, health care providers, and users in case of device malfunctions or errors [85]. Assessing liability aspects within the HTA process helps establish accountability and ensures that legal frameworks adequately address potential risks [86].

Regulatory compliance is a crucial consideration when assessing the suitability of AI-based medical devices for HTA [87]. Depending on the jurisdiction, AI devices need to undergo regulatory approval processes before being introduced into the market [88]. HTA examines whether AI-based medical device manufacturers have obtained the necessary regulatory approvals, such as clearance from the relevant health authorities or certification from regulatory bodies such as the Food and Drug Administration (FDA) in the United States or notified bodies in Europe [89]. This evaluation ensures that AI-based medical devices comply with existing regulations and are fit for clinical use [61].

Discussion

Recommendations to Adapt the 9 HTA Domains for AI-Based Medical Devices

Taking into account the previous considerations, some recommendations could be proposed to adapt and personalize these standard HTA domains to AI-based medical devices' specificities. Therefore, we suggest 4 main recommendations by domain in Table 3.

XSL•FO
RenderX

Table 3. Recommendations to adapt and personalize the health technology assessment (HTA) of standard HTA domains to artificial intelligence (AI)–based medical device specificities.

Domains	4 recommendations per domain	Application examples
1 and 2—health problem, description, and technical characteristics of the technology	• Assess the technical performance and reliability of the AI-based medical device, especially for ML[a]- and DL[b]-based medical devices • Evaluate the usability and user interface of the AI-based medical device • Assess the interoperability and compatibility of the AI-based medical device with existing systems • Regularly update the assessment to reflect AI technological advancements • Update the assessment periodically to incorporate new information and advancements related to the AI-based medical device	Consider the example of an AI-based DMS[c] that assists patients with diabetes in managing their condition by providing blood glucose predictions based on their planned actions, thus contributing to better decision-making. The description and technical features involve defining patient data collection, data preprocessing, selecting relevant features, constructing an AI model, training the model & validating it using separate data, deploying it to provide blood glucose predictions to patients, and continuously improving it through real-time data collection.
3—safety	• Evaluate AI compliance with safety regulations and standards and AI risk management processes • Evaluate the AI's robustness and resilience to errors or failures • Assess the AI's impact on patient safety, user safety, and long-term safety • Consider the durability of the AI technology; regularly monitor and update safety information; and regularly monitor and reassess the AI's safety profile incorporating new information, evidence, and user experiences	The potential risk of patient injury of an insulin delivery AI-based medical device system should be taken into account in the risk management process. In the case of an evolutive DL–based medical device without continuous ongoing safety assessment, a wrong dosage administration due to an AI error could provoke a serious adverse event (ie acid ketosis coma for a patient with diabetes). A risk management plan should be available and regularly updated with AI changes impact on safety plan.
4—clinical effectiveness	• Compare the AI technology to existing alternatives or standard of care • Assess the generalizability of the clinical evidence and analyze the clinical outcomes in subpopulations • Consider patient-reported outcomes, quality-of-life measures, and impact on clinical workflow • Regularly update the clinical evidence base of the AI-based medical devices and monitor consistency during the AI lifetime	The FDA[d] in the United States has proposed a regulatory framework for modifications to evolutive AI- and ML-based software as a medical device with modification guidance focused on the risk to users and patients resulting from the AI changes. They have asked for an Algorithm Change Protocol with specific methods in place to achieve and control the risks of the anticipated types of modifications related to performance, use, or inputs. A continuous clinical effectiveness assessment could be an interesting approach for an AI-based medical device diabetes foot ulcer detection system for patients needing adaptative treatment modifications to prevent ulcer development.
5—costs and economic evaluation	• Assess the budget impact of adopting the AI-based medical device within the health care system • If needed, conduct a comprehensive cost-effectiveness analysis of the AI-based medical device compared to standard care and assess long-term impact on health care resource use (hospital admissions, length of stay, and time to diagnosis), health care disparities, and access to care in case of evolution of the AI's performance • Explore potential reimbursement strategies and work with payers to develop reimbursement models that align with the value and impact of the AI device (on clinical effectiveness, cost-effectiveness, and long-term sustainability) • Establish mechanisms for ongoing monitoring and reassessment of the AI device's economic value to ensure that it remains aligned with evolving health care priorities and resource allocation strategies	To evaluate a diabetic retinopathy screening AI-based medical device, clinical effectiveness is not sufficient. The cost-effectiveness of different types of AI diabetic retinopathy screening should be compared with no screening and ophthalmologist screening. A recent study demonstrated that AI-based screening was the most cost-effective, not only saving costs but also improving the quality of life of patients with diabetes. In this case, the long-term assessment of the economic impact of AI introduction in diabetic retinopathy screening highlighted the added value of this technology.

XSL•FO
RenderX

Domains	4 recommendations per domain	Application examples
7—organizational aspects	• Assess the organizational readiness and capacity for implementing AI-based medical devices in health care settings • Conduct a thorough analysis of the impact of implementing AI-based medical devices on health care workflows and processes • Establish clear guidelines and protocols for the appropriate use and integration of AI-based medical devices • Monitor and evaluate the impact of AI-based medical device implementation on patient outcomes, quality of care, and patient satisfaction • Regularly review and update implementation strategies based on the findings from monitoring AI	Decision makers and technology advocates need to address the complexities of AI more comprehensively and understand the systemic challenges that its adoption poses to health care organizations and systems. As an example, consider an AI tool used for diagnosing diabetic retinopathy in a primary care setting, such as by a family physician or nurse. In theory, this could lead to shorter waiting times for patients. However, if the health care organization faces challenges such as a shortage of specialized staff (eg, ophthalmologists), insufficient organization of care pathways, and lack of specialized facilities for proper management and follow-up after diagnosis, the introduction of AI might adversely affect the quality of care and patient experience.
Domains 6, 8, and 9—ethical analysis or patients and social aspects or legal aspects	• Conduct an ethical analysis and evaluate the social implications of implementing the AI-based medical devices • Assess the legal and regulatory aspects related to the AI-based medical device • Promote transparency, interpretability, and explainability of the AI-based medical devices • Monitor and address emerging ethical, social, and legal issues related to the AI-based medical devices • Establish mechanisms for ongoing surveillance and evaluation of the AI device's ethical, social, and legal implications	AI-based continuous glucose monitoring and insulin pumps should also be assessed on ethical and legal aspects. The risks of data theft and privacy breaches necessitate careful consideration of ethical and legal issues for patients. Although AI can aid in decision-making, it cannot wholly substitute a physician's expertise. Effective regulations and systems designed to ensure safety, reduce bias, and enhance transparency are essential.

[a]ML: machine learning.

[b]DL: deep learning.

[c]DMS: diabetes management system.

[d]FDA: Food and Drug Administration.

The adaptation of the HTA process through a methodological framework for AI-based medical devices enhances the comparability of results across different evaluations and jurisdictions [90-92]. It promotes consistency in the assessment methodologies, reporting formats, and presentation of findings, enabling decision makers to make informed choices based on reliable and comparable evidence. This standardization contributes to the overall credibility and acceptance of HTA outcomes related to AI-based medical devices [93]. In addition, a methodological framework would address the need for appropriate expertise and skills to conduct the HTA of AI-based medical devices [94-99]. It would outline the qualifications and competencies required for the individuals involved in the assessment, including knowledge of AI technologies [46]. By defining the necessary expertise, the framework supports the development of a skilled workforce capable of conducting robust and reliable HTAs of AI-based medical devices (Figure 2).

XSL•FO
RenderX

Figure 2. Suggested recommendations to adapt each standard health technology assessment (HTA) domain for the evaluation of artificial intelligence (AI)–based medical devices (MDs). RWE: real-world evidence.

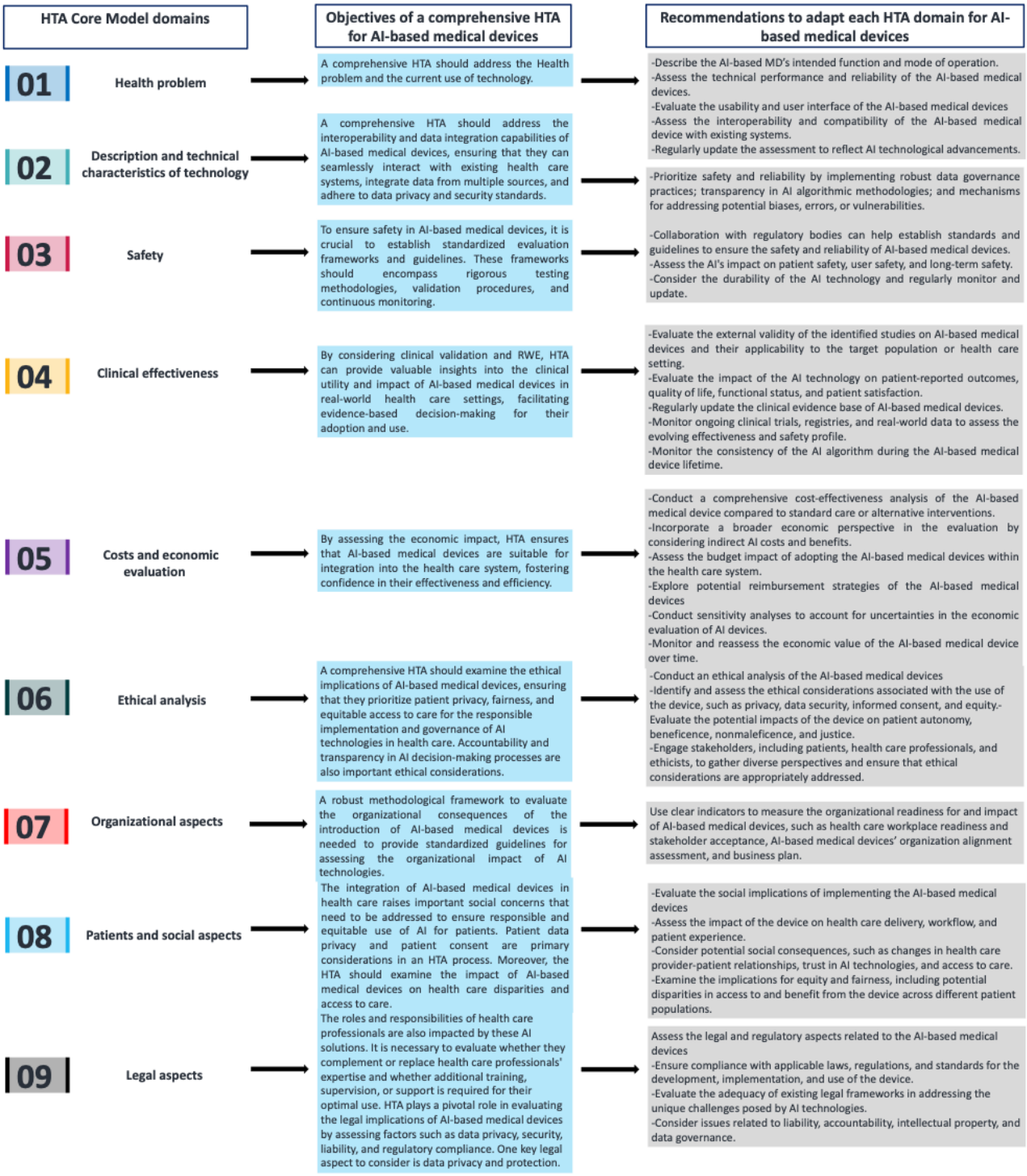

Evaluating the trade-offs and weighing different features of AI-based medical devices is indeed a complex and important task, especially in domains such as health care where the impact on human lives is significant. The acceptability of AI-based medical devices should be assessed on a case-by-case basis considering various factors, including, for instance, performance, accuracy, cost, explainability, and the specific context in which they are being used.

Trade-offs between accuracy and cost are common in AI. It may be acceptable for AI to increase follow-up care costs if it significantly improves accuracy and patient outcomes. For example, if an AI system can detect diseases at an earlier stage, it might lead to more effective treatment and, ultimately, lower overall health care costs in the long run.

The balance between explainability and performance is a critical consideration. While explainable algorithms are preferred for safety and transparency reasons, there may be cases in which a highly complex, unexplainable algorithm outperforms explainable ones. In such cases, the trade-off between transparency and performance should be carefully evaluated based on the specific use case and potential risks.

XSL•FO
RenderX

The minimum performance for adding this technology to the tools available to human clinicians depends on the specific task and the level of trust that patients, health care providers, and regulators have in the technology. Some key factors to consider include the following: (1) The complexity of the task—AI-based medical devices should excel in tasks that are well defined and data driven, but they may not replace human clinicians in tasks requiring complex decision-making, empathy, or ethical considerations. (2) Safety and reliability—AI-based medical devices should demonstrate a high level of safety and reliability, ideally surpassing the performance of human clinicians in terms of avoiding errors. (3) Ethical considerations—AI-based medical devices should adhere to ethical standards, including patient privacy, informed consent, and unbiased decision-making, which are often considered even more important than performance metrics. (4) Regulatory approval—regulatory bodies often establish performance thresholds for AI-based medical devices. Compliance with these thresholds is essential for market acceptance.

In general, AI should aim to complement and enhance the capabilities of human clinicians rather than completely replacing them. The specific threshold for acceptable performance will vary across applications and contexts, and it should be determined through a combination of rigorous testing, peer-reviewed studies, and input from health care professionals and patients.

It is important to note that ethical considerations, patient safety, and the potential for bias should always be at the forefront of these discussions, and AI-based medical devices should not be adopted solely for the sake of automation or cost reduction if they compromise these critical aspects of health care.

Use Case of the Application of the Aforementioned HTA Recommendations for AI: AI-Based Medical Devices in Pathways for Patients With Diabetes

Concerning the HTA domain 3 on the safety of an AI-based medical device, the potential risk of patient injury of an insulin delivery AI-based medical device system should be taken into account in the risk management process of algorithm development [96]. In the case of an evolutive deep learning–based medical device without continuous ongoing safety assessment, a wrong dosage administration due to an AI error, for instance, could provoke a serious adverse event such as an acid ketosis coma for a patient with diabetes.

A risk management plan should be available for users and regularly updated with the impact of the AI changes on the safety plan.

In relation to the HTA domain 4, which focuses on the effectiveness of an AI-based medical device, the FDA in the United States has proposed a regulatory framework for modifications to evolutive AI- and machine learning–based software as a medical device with modification guidance focused on the risk to users or patients resulting from the AI changes [100]. They have asked for an Algorithm Change Protocol with specific methods in place to achieve and control the risks of the anticipated types of modifications related to performance, use, or inputs. A continuous clinical effectiveness assessment could be an interesting approach for an AI-based medical device detection system for diabetes foot ulcer for patients needing adaptative treatment modifications to prevent ulcer development [101].

Concerning the HTA domain 5, to evaluate a diabetic retinopathy screening AI-based medical device, clinical effectiveness is not sufficient. The cost-effectiveness of different types of AI diabetic retinopathy screening should be compared with no screening and ophthalmologist screening. A recent study demonstrated that AI-based screening was the most cost-effective, not only saving costs but also improving the quality of life of patients with diabetes [44]. In this case, the long-term assessment of the economic impact of AI introduction in diabetic retinopathy screening highlighted the added value of this technology.

The HTA domain 7 on organizational impact has to assess how the AI-based medical device can be effectively integrated into the health care pathway and prevent wasteful spending. More thorough attention must be paid to the following aspects: (1) evaluating needs and determining the added value of the implementation of the AI-based medical device; (2) assessing workplace preparedness, including stakeholder acceptance of the introduction of the AI-based medical device and involvement; and (3) analyzing the alignment between AI technology and organizational structure [64]. Decision makers and technology advocates need to address the complexities of AI more comprehensively and understand the systemic challenges that its adoption poses to health care organizations and systems. As an example, consider an AI tool used for diagnosing diabetic retinopathy in a primary care setting, such as by a family physician or nurse [102]. In theory, this could lead to shorter waiting times for patients. However, if the health care organization faces challenges such as a shortage of specialized staff (eg, ophthalmologists), insufficient organization of care pathways, and lack of specialized facilities for proper management and follow-up after diagnosis, the introduction of AI might adversely affect the quality of care and patient experience. In such a scenario, the AI application might merely transfer the delay from primary to secondary care, failing to address the fundamental issue.

Finally, concerning the last domains (6, 8, and 9) about ethical, patient, social, and legal aspects, AI-based continuous glucose monitoring and insulin pumps should also be assessed on ethical and legal aspects [103]. While citizen juries have generally shown support for AI in research and treatment, concerns remain. The risks of data theft and privacy breaches necessitate careful consideration of ethical and legal issues for patients. Although AI can aid in decision-making, it cannot wholly substitute a physician's expertise. Effective regulations and systems designed to ensure safety, reduce bias, and enhance transparency are essential.

By implementing these recommendations, stakeholders can foster the responsible integration, regulation, and evaluation of AI-based medical devices. These measures can enhance the evidence base, address ethical concerns, and maximize the potential benefits of these AI technologies in improving health

XSL•FO
RenderX

care outcomes while protecting patient safety, privacy, and equity.

Practical Challenges and Potential Barriers to Implementing HTAs Specific for AI-Based Medical Devices

When it comes to implementing specialized HTAs for AI-based medical devices, several practical challenges and potential barriers could be encountered and should be taken into account: First, AI technological evolution timeline—AI technologies evolve at a much faster pace compared to traditional medical devices, making it challenging for HTA frameworks to keep up with the latest developments and assess their long-term impact effectively. AI-based medical devices have a short product lifetime, between 12 and 18 months, in contrast to drug products. This shorter life cycle highlights the need for evolutive and fast-track HTA processes for AI-based medical devices. Second, data requirements and quality—AI systems rely heavily on large data sets for training and validation. Ensuring the availability of high-quality, representative data is a significant challenge. There is also the issue of data privacy and security, which must be addressed. The availability and quality of data and evidence required for conducting HTAs on AI-based medical devices present a complex and evolving landscape. Assessing the feasibility and challenges associated with gathering such data is crucial for robust evaluations. First, the availability of data can vary significantly depending on the AI-based medical device. While some devices may have access to vast amounts of high-quality real-world patient data, others might face limitations due to the novelty of the technology or issues related to data privacy. Second, the quality of data is paramount as inaccurate or biased data can lead to flawed assessments. Ensuring data accuracy, representativeness, and relevance is a constant challenge in AI-based medical device evaluations [104]. The rapid pace of AI development can result in limited long-term data, making it difficult to assess the device's sustained performance and safety [105]. Balancing the need for robust evidence with the dynamic nature of AI technologies is a significant challenge that HTA organizations must address to provide valuable insights for informed decision-making in health care. Third, complexity and transparency—the complex algorithms used in AI-based medical devices can be difficult to understand and assess, leading to issues with transparency and explainability [17]. This complexity can pose a challenge for regulators and assessors in HTA processes. Fourth, clinical validation and safety requirements—generating robust clinical evidence to demonstrate the safety, efficacy, and effectiveness of AI-based medical devices can be challenging. This includes proving that these devices perform consistently across diverse patient populations. Moreover, to account for a life cycle that would include updates that may improve performance, one solution to investigate could be the proposition of the FDA in the United States in 2019 to experiment with a "dynamic certification" [100]. It allows for the re-evaluation of the AI-based medical device in case of a substantial modification of the indication of the medical device or the way to deliver the diagnosis, for instance. Fifth, regulatory and ethical considerations—adapting existing regulatory frameworks to accommodate AI-based medical devices, addressing ethical concerns such as bias, and ensuring equitable access are critical challenges [106]. Sixth, economic evaluation—determining the cost-effectiveness of AI-based medical devices, especially when benefits might be indirect or long term, poses a unique challenge for HTA [93]. Seventh, stakeholder engagement and trust—building trust among health care providers, patients, and policy makers regarding the reliability, trustworthiness, and usefulness of AI-based medical devices is crucial but challenging [107]. Eighth, integration into health care systems and interoperability—the integration of AI-based medical devices into existing health care workflows and systems can be complex and resource intensive [18]. Ninth, global and local applicability—ensuring that AI-based medical devices are effective and appropriate for use in different global and local contexts, considering varying health care systems and population needs, is another significant barrier.

Practically, one suggestion to implement such frameworks could be to implement these recommendations in the future European clinical joint assessment guidelines. As they are being currently discussed in a European project on a common HTA process for connected medical devices that includes AI-based medical devices, it could be the opportunity to tackle these challenges at the European level [108].

Addressing these challenges requires a concerted effort from regulators, health care providers, technology developers, and other stakeholders in the health care ecosystem.

Conclusions

AI-based medical devices have the potential to transform health care delivery, but the suitability of the current comprehensive HTA requires careful adaptation of the evaluation across the 9 dimensions. While these AI devices show promise in improving accuracy, safety, and efficiency, there is a need for robust clinical validation, integration into workflows, economic evaluation, and addressing of ethical and legal implications. A comprehensive adapted HTA framework for AI-based medical devices can provide valuable insights into their effectiveness, cost-effectiveness, and societal impact, guiding their responsible implementation and maximizing their benefits for patients and health care systems.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LF and AV contributed to the conceptualization. LF and AV contributed to the methodology. LF, AV, NM, and IB contributed to the article selection. LF and AV contributed to the writing and original draft preparation. NM and IB contributed to the writing

XSL•FO
RenderX

and review and editing. LF and AV contributed to the validation. The authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 9 health technology assessment (HTA) domains for the assessment of artificial intelligence–based medical devices adapted from the HTA Core Model proposed by the European Network for Health Technology Assessment.
[DOCX File , 13 KB - jmir_v26i1e51514_app1.docx]

Multimedia Appendix 2
Summary results of the scoping review and general characteristics of the studies included in the scoping review.
[DOCX File , 27 KB - jmir_v26i1e51514_app2.docx]

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
2. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. Cambridge, MA: Academic Press; 2020.
3. Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, et al. Artificial intelligence in health care: current applications and issues. J Korean Med Sci 2020 Nov 02;35(42):e379. [doi: 10.3346/jkms.2020.35.e379] [Medline: 33140591]
4. Farah L, Davaze-Schneider J, Martin T, Nguyen P, Borget I, Martelli N. Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review. Artif Intell Med 2023 Jun;140:102547. [doi: 10.1016/j.artmed.2023.102547] [Medline: 37210155]
5. HTA core model. European Network for Health Technology Assessment. URL: https://www.eunethta.eu/hta-core-model/ [accessed 2021-10-20]
6. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
7. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AICONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020 Sep 09;370:m3210. [doi: 10.1136/bmj.m3210] [Medline: 32907797]
8. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 09;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
9. Chen Y. Health technology assessment and economic evaluation: is it applicable for the traditional medicine? Integr Med Res 2022 Mar;11(1):100756. [doi: 10.1016/j.imr.2021.100756] [Medline: 34401322]
10. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
11. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Front Digit Health 2021 Jun 25;3:645232. [doi: 10.3389/fdgth.2021.645232] [Medline: 34713115]
12. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018 Nov 6;15(11):e1002689. [doi: 10.1371/journal.pmed.1002689] [Medline: 30399149]
13. Dzobo K, Adotey S, Thomford NE, Dzobo W. Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. OMICS 2020 May;24(5):247-263. [doi: 10.1089/omi.2019.0038] [Medline: 31313972]
14. Zawati M, Lang M. What's in the box?: uncertain accountability of machine learning applications in healthcare. Am J Bioeth 2020 Nov;20(11):37-40. [doi: 10.1080/15265161.2020.1820105] [Medline: 33103971]
15. Peters MD, Godfrey C, McInerney P, Khalil H, Larsen P, Marnie C, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022 Apr 01;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
16. Pollock D, Peters MD, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 01;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
17. Farah L, Murris JM, Borget I, Guilloux A, Martelli NM, Katsahian SI. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. Mayo Clin Proc Digit Health 2023 Jun;1(2):120-138. [doi: 10.1016/j.mcpdig.2023.02.004]

XSL•FO
RenderX

18. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019 Aug 20;2:79. [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
19. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020 Jul 22;20(1):170. [doi: 10.1186/s12911-020-01191-1] [Medline: 32698869]
20. Zia A, Aziz M, Popa I, Khan SA, Hamedani AF, Asif AR. Artificial intelligence-based medical data mining. J Pers Med 2022 Aug 24;12(9):1359. [doi: 10.3390/jpm12091359] [Medline: 36143144]
21. Ming J, He Y, Yang Y, Hu M, Zhao X, Liu J, et al. Health technology assessment of medical devices: current landscape, challenges, and a way forward. Cost Eff Resour Alloc 2022 Oct 05;20(1):54. [doi: 10.1186/s12962-022-00389-6] [Medline: 36199144]
22. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res 2016 Mar 15;8(3):1560-1580. [Medline: 27186282]
23. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
24. Sloane EJ, Silva RJ. Artificial intelligence in medical devices and clinical decision support systems. In: Clinical Engineering Handbook (Second Edition). Cambridge, MA: Academic Press; 2020.
25. Steuten LM. Early stage health technology assessment for precision biomarkers in oral health and systems medicine. OMICS 2016 Jan;20(1):30-35. [doi: 10.1089/omi.2015.0174] [Medline: 26760958]
26. Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S. Significance of machine learning in healthcare: features, pillars and applications. Int J Intell Networks 2022;3:58-73. [doi: 10.1016/j.ijin.2022.05.002]
27. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020 Jan 13;3:4. [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
28. Voets MM, Veltman J, Slump CH, Siesling S, Koffijberg H. Systematic review of health economic evaluations focused on artificial intelligence in healthcare: the tortoise and the cheetah. Value Health 2022 Mar;25(3):340-349. [doi: 10.1016/j.jval.2021.11.1362] [Medline: 35227444]
29. Kirisits A, Redekop WK. The economic evaluation of medical devices: challenges ahead. Appl Health Econ Health Policy 2013 Feb;11(1):15-26. [doi: 10.1007/s40258-012-0006-9] [Medline: 23329383]
30. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. Korean J Radiol 2021 Mar;22(3):442-453. [doi: 10.3348/kjr.2021.0048] [Medline: 33629545]
31. Tsopra R, Fernandez X, Luchinat C, Alberghina L, Lehrach H, Vanoni M, et al. A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. BMC Med Inform Decis Mak 2021 Oct 02;21(1):274. [doi: 10.1186/s12911-021-01634-3] [Medline: 34600518]
32. Bolboacă SD. Medical diagnostic tests: a review of test anatomy, phases, and statistical treatment of data. Comput Math Methods Med 2019 May 28;2019:1891569. [doi: 10.1155/2019/1891569] [Medline: 31275427]
33. Hogervorst MA, Pontén J, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Real world data in health technology assessment of complex health technologies. Front Pharmacol 2022 Feb 10;13:837302. [doi: 10.3389/fphar.2022.837302] [Medline: 35222045]
34. Simon GE, Bindman AB, Dreyer NA, Platt R, Watanabe JH, Horberg M, et al. When can we trust real-world data to evaluate new medical treatments? Clin Pharmacol Ther 2022 Jan;111(1):24-29. [doi: 10.1002/cpt.2252] [Medline: 33932030]
35. Pongiglione B, Torbica A, Blommestein H, de Groot S, Ciani O, Walker S, et al. Do existing real-world data sources generate suitable evidence for the HTA of medical devices in Europe? Mapping and critical appraisal. Int J Technol Assess Health Care 2021 Apr 26;37(1):e62. [doi: 10.1017/S0266462321000301] [Medline: 33896433]
36. Pongiglione B, Torbica A. How real can we get in generating real world evidence? Exploring the opportunities of routinely collected administrative data for evaluation of medical devices. Health Econ 2022 Sep;31 Suppl 1(Suppl 1):25-43. [doi: 10.1002/hec.4562] [Medline: 35762465]
37. Zemplényi A, Tachkov K, Balkanyi L, Németh B, Petykó ZI, Petrova G, et al. Recommendations to overcome barriers to the use of artificial intelligence-driven evidence in health technology assessment. Front Public Health 2023 Apr 26;11:1088121. [doi: 10.3389/fpubh.2023.1088121] [Medline: 37181704]
38. Daubner-Bendes R, Kovács S, Niewada M, Huic M, Drummond M, Ciani O, et al. Quo Vadis HTA for medical devices in central and eastern Europe? Recommendations to address methodological challenges. Front Public Health 2021 Jan 08;8:612410. [doi: 10.3389/fpubh.2020.612410] [Medline: 33490024]
39. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. J Am Coll Radiol 2021 Mar;18(3 Pt A):413-424. [doi: 10.1016/j.jacr.2020.09.060] [Medline: 33096088]
40. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599. [doi: 10.2196/18599] [Medline: 32706688]
41. Belenguer L. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI Ethics 2022;2(4):771-787. [doi: 10.1007/s43681-022-00138-8] [Medline: 35194591]

XSL•FO
RenderX

42. Binder L, Ghadban M, Sit C, Barnard K. Health technology assessment process for oncology drugs: impact of CADTH changes on public payer reimbursement recommendations. Curr Oncol 2022 Mar 01;29(3):1514-1526. [doi: 10.3390/curroncol29030127] [Medline: 35323327]
43. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res 2020 Feb 20;22(2):e16866. [doi: 10.2196/16866] [Medline: 32130134]
44. Abràmoff MD, Roehrenbeck C, Trujillo S, Goldstein J, Graves AS, Repka MX, et al. A reimbursement framework for artificial intelligence in healthcare. NPJ Digit Med 2022 Jun 09;5(1):72. [doi: 10.1038/s41746-022-00621-w] [Medline: 35681002]
45. Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? Front Artif Intell 2021 Nov 02;4:736697. [doi: 10.3389/frai.2021.736697] [Medline: 34796318]
46. Alami H, Lehoux P, Auclair Y, de Guise M, Gagnon MP, Shaw J, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 07;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
47. Love-Koh J, Peel A, Rejon-Parrilla JC, Ennis K, Lovett R, Manca A, et al. The future of precision medicine: potential impacts for health technology assessment. Pharmacoeconomics 2018 Dec;36(12):1439-1451. [doi: 10.1007/s40273-018-0686-6] [Medline: 30003435]
48. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of artificial intelligence as a decision-support system applied to the detection and grading of melanoma, dental caries, and diabetic retinopathy. JAMA Netw Open 2022 Mar 01;5(3):e220269. [doi: 10.1001/jamanetworkopen.2022.0269] [Medline: 35289862]
49. Naik N, Hameed BM, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg 2022 Mar 14;9:862322. [doi: 10.3389/fsurg.2022.862322] [Medline: 35360424]
50. Haynes CL, Cook GA, Jones MA. Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learnt from developing a disease register. J Med Ethics 2007 May;33(5):302-307. [doi: 10.1136/jme.2006.016907] [Medline: 17470509]
51. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020 May;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
52. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 01;178(11):1544-1547. [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
53. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Front Artif Intell 2021 Apr 15;3:561802. [doi: 10.3389/frai.2020.561802] [Medline: 33981989]
54. Tachkov K, Zemplenyi A, Kamusheva M, Dimitrova M, Siirtola P, Pontén J, et al. Barriers to use artificial intelligence methodologies in health technology assessment in central and east European countries. Front Public Health 2022 Jul 14;10:921226. [doi: 10.3389/fpubh.2022.921226] [Medline: 35910914]
55. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18 (forthcoming). [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
56. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
57. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Front Artif Intell 2022 May 30;5:879603. [doi: 10.3389/frai.2022.879603] [Medline: 35707765]
58. Baltaxe E, Hsieh HW, Roca J, Cano I. The assessment of medical device software supporting health care services for chronic patients in a tertiary hospital: overarching study. J Med Internet Res 2023 Jan 04;25:e40976. [doi: 10.2196/40976] [Medline: 36598817]
59. Garfield S, Polisena J, S Spinner D, Postulka A, Y Lu C, Tiwana SK, et al. Health technology assessment for molecular diagnostics: practices, challenges, and recommendations from the medical devices and diagnostics special interest group. Value Health 2016;19(5):577-587. [doi: 10.1016/j.jval.2016.02.012] [Medline: 27565275]
60. Fraser AG, Biasin E, Bijnens B, Bruining N, Caiani EG, Cobbaert K, et al. Artificial intelligence in medical device software and high-risk medical devices - a review of definitions, expert recommendations and regulatory initiatives. Expert Rev Med Devices 2023 Jun;20(6):467-491. [doi: 10.1080/17434440.2023.2184685] [Medline: 37157833]
61. Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. Phys Med 2021 Mar;83:1-8. [doi: 10.1016/j.ejmp.2021.02.011] [Medline: 33657513]
62. Melvin T, Torre M. New medical device regulations: the regulator's view. EFORT Open Rev 2019 Jun 03;4(6):351-356. [doi: 10.1302/2058-5241.4.180061] [Medline: 31312522]

XSL•FO
RenderX

63. Fleetcroft C, McCulloch P, Campbell B. IDEAL as a guide to designing clinical device studies consistent with the new European Medical Device Regulation. BMJ Surg Interv Health Technol 2021 Mar 04;3(1):e000066. [doi: 10.1136/bmjsit-2020-000066] [Medline: 35047800]
64. Alami H, Lehoux P, Denis JL, Motulsky A, Petitgand C, Savoldelli M, et al. Organizational readiness for artificial intelligence in health care: insights for decision-making and practice. J Health Organ Manag 2020 Dec 03 (forthcoming). [doi: 10.1108/JHOM-03-2020-0074] [Medline: 33258359]
65. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM. Methodological frameworks and dimensions to be taken into consideration in digital health technology assessment: protocol for a scoping review. JMIR Res Protoc 2022 Oct 11;11(10):e39905. [doi: 10.2196/39905] [Medline: 36222788]
66. Weinert L, Müller J, Svensson L, Heinze O. Perspective of information technology decision makers on factors influencing adoption and implementation of artificial intelligence technologies in 40 German hospitals: descriptive analysis. JMIR Med Inform 2022 Jun 15;10(6):e34678. [doi: 10.2196/34678] [Medline: 35704378]
67. de Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2. [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
68. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702. [doi: 10.7717/peerj.7702] [Medline: 31592346]
69. Widrig D, Tag B. HTA and its legal issues: a framework for identifying legal issues in health technology assessment. Int J Technol Assess Health Care 2014 Dec;30(6):587-594. [doi: 10.1017/S0266462314000683] [Medline: 25816824]
70. Vella Bonanno P, Bucsics A, Simoens S, Martin AP, Oortwijn W, Gulbinovič J, et al. Proposal for a regulation on health technology assessment in Europe - opinions of policy makers, payers and academics from the field of HTA. Expert Rev Pharmacoecon Outcomes Res 2019 Jun;19(3):251-261. [doi: 10.1080/14737167.2019.1575730] [Medline: 30696372]
71. McKee M, Wouters OJ. The challenges of regulating artificial intelligence in healthcare comment on "clinical decision support and new regulatory frameworks for medical devices: are we ready for it? - a viewpoint paper". Int J Health Policy Manag 2023;12:7261. [doi: 10.34172/ijhpm.2022.7261] [Medline: 36243948]
72. Hordern V. Data protection compliance in the age of digital health. Eur J Health Law 2016 Jun;23(3):248-264. [doi: 10.1163/15718093-12341393] [Medline: 27491248]
73. Stanberry B. The legal and ethical aspects of telemedicine. 2: data protection, security and European law. J Telemed Telecare 1998;4(1):18-24. [doi: 10.1258/1357633981931236] [Medline: 9640706]
74. Poullet Y. Legal aspects of data protection in medical informatics. Stud Health Technol Inform 1991;1:138-160. [Medline: 10164065]
75. Dove ES, Chen J. To what extent does the EU general data protection regulation (GDPR) apply to citizen scientist-led health research with mobile devices? J Law Med Ethics 2020 Mar;48(1_suppl):187-195. [doi: 10.1177/1073110520917046] [Medline: 32342746]
76. Marovic B, Curcin V. Impact of the European general data protection regulation (GDPR) on health data management in a European Union candidate country: a case study of Serbia. JMIR Med Inform 2020 Apr 17;8(4):e14604. [doi: 10.2196/14604] [Medline: 32301736]
77. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/ [accessed 2023-07-24]
78. Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. European Union. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 [accessed 2023-07-24]
79. Camara C, Peris-Lopez P, Tapiador JE. Security and privacy issues in implantable medical devices: a comprehensive survey. J Biomed Inform 2015 Jun;55:272-289. [doi: 10.1016/j.jbi.2015.04.007] [Medline: 25917056]
80. Migliore A, Ratti M, Cerbo M, Jefferson T. Health Technology Assessment: managing the introduction and use of medical devices in clinical practice in Italy. Expert Rev Med Devices 2009 May;6(3):251-257. [doi: 10.1586/erd.09.2] [Medline: 19419283]
81. Pisapia A, Banfi G, Tomaiuolo R. The novelties of the regulation on health technology assessment, a key achievement for the European Union health policies. Clin Chem Lab Med 2022 Jul 26;60(8):1160-1163. [doi: 10.1515/cclm-2022-0228] [Medline: 35624072]
82. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q 2021 Sep;99(3):629-647. [doi: 10.1111/1468-0009.12504] [Medline: 33822422]
83. Jassar S, Adams SJ, Zarzeczny A, Burbridge BE. The future of artificial intelligence in medicine: medical-legal considerations for health leaders. Healthc Manage Forum 2022 May;35(3):185-189. [doi: 10.1177/08404704221082069] [Medline: 35354409]
84. Samore MH, Evans RS, Lassen A, Gould P, Lloyd J, Gardner RM, et al. Surveillance of medical device-related hazards and adverse events in hospitalized patients. JAMA 2004 Jan 21;291(3):325-334. [doi: 10.1001/jama.291.3.325] [Medline: 14734595]
85. Bleher H, Braun M. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. AI Ethics 2022;2(4):747-761. [doi: 10.1007/s43681-022-00135-x] [Medline: 35098247]

XSL•FO
RenderX

86. Street J, Stafinski T, Lopes E, Menon D. Defining the role of the public in Health Technology Assessment (HTA) and HTA-informed decision-making processes. Int J Technol Assess Health Care 2020 Apr;36(2):87-95. [doi: 10.1017/S0266462320000094] [Medline: 32151290]
87. Massella M, Dri DA, Gramaglia D. Regulatory considerations on the use of machine learning based tools in clinical trials. Health Technol (Berl) 2022;12(6):1085-1096. [doi: 10.1007/s12553-022-00708-0] [Medline: 36373014]
88. Marcus HJ, Payne CJ, Hughes-Hallett A, Marcus AP, Yang GZ, Darzi A, et al. Regulatory approval of new medical devices: cross sectional study. BMJ 2016 May 20;353:i2587. [doi: 10.1136/bmj.i2587] [Medline: 27207165]
89. Milam ME, Koo CW. The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. Clin Radiol 2023 Feb;78(2):115-122. [doi: 10.1016/j.crad.2022.08.135] [Medline: 36180271]
90. Vreman RA, Mantel-Teeuwisse AK, Hövels AM, Leufkens HG, Goettsch WG. Differences in health technology assessment recommendations among European jurisdictions: the role of practice variations. Value Health 2020 Jan;23(1):10-16. [doi: 10.1016/j.jval.2019.07.017] [Medline: 31952664]
91. Allen N, Walker SR, Liberti L, Salek S. Health technology assessment (HTA) case studies: factors influencing divergent HTA reimbursement recommendations in Australia, Canada, England, and Scotland. Value Health 2017 Mar;20(3):320-328. [doi: 10.1016/j.jval.2016.10.014] [Medline: 28292476]
92. Agboola F, Wright AC, Herron-Smith S, Mathur D, Rind D. Evaluation of diversity of clinical trials informing health technology assessments in the United States: a 5-year analysis of institute for clinical and economic review assessments. Value Health 2023 Sep;26(9):1345-1352. [doi: 10.1016/j.jval.2023.05.014] [Medline: 37244417]
93. Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the economic value of clinical artificial intelligence: challenges and opportunities. Value Health 2022 Mar;25(3):331-339. [doi: 10.1016/j.jval.2021.08.015] [Medline: 35227443]
94. Nielsen CP, Lauritsen SW, Kristensen FB, Bistrup ML, Cecchetti A, Turk E. Involving stakeholders and developing a policy for stakeholder involvement in the European network for health technology assessment, EUnetHTA. Int J Technol Assess Health Care 2009 Dec;25 Suppl 2:84-91. [doi: 10.1017/S0266462309990729] [Medline: 20030895]
95. Scott AM, Wale JL, HTAi PatientCitizen Involvement in HTA Interest Group, Patient InvolvementEducation Working Group. Patient advocate perspectives on involvement in HTA: an international snapshot. Res Involv Engagem 2017 Jan 10;3:2. [doi: 10.1186/s40900-016-0052-9] [Medline: 29062527]
96. Gagnon MP, Desmartis M, Gagnon J, St-Pierre M, Gauvin FP, Rhainds M, et al. Introducing the patient's perspective in hospital health technology assessment (HTA): the views of HTA producers, hospital managers and patients. Health Expect 2014 Dec;17(6):888-900. [doi: 10.1111/hex.12010] [Medline: 23046439]
97. Gagnon M, Desmartis M, Gagnon J, St-Pierre M, Rhainds M, Coulombe M, et al. Framework for user involvement in health technology assessment at the local level: views of health managers, user representatives, and clinicians. Int J Technol Assess Health Care 2015 Jan;31(1-2):68-77. [doi: 10.1017/S0266462315000070] [Medline: 25952585]
98. Angelis A, Kanavos P. Value-based assessment of new medical technologies: towards a robust methodological framework for the application of multiple criteria decision analysis in the context of health technology assessment. Pharmacoeconomics 2016 May;34(5):435-446. [doi: 10.1007/s40273-015-0370-z] [Medline: 26739955]
99. Wale JL, Thomas S, Hamerlijnck D, Hollander R. Patients and public are important stakeholders in health technology assessment but the level of involvement is low - a call to action. Res Involv Engagem 2021 Jan 05;7(1):1. [doi: 10.1186/s40900-020-00248-9] [Medline: 33402216]
100. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food & Drug Administration. 2023 Oct 19. URL: https://tinyurl.com/m8ba5aw4 [accessed 2023-11-25]
101. Cassidy B, Hoon Yap M, Pappachan JM, Ahmad N, Haycocks S, O'Shea C, et al. Artificial intelligence for automated detection of diabetic foot ulcers: a real-world proof-of-concept clinical evaluation. Diabetes Res Clin Pract 2023 Nov;205:110951. [doi: 10.1016/j.diabres.2023.110951] [Medline: 37848163]
102. Mackenzie SC, Sainsbury CA, Wake DJ. Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges. Diabetologia 2024 Feb;67(2):223-235. [doi: 10.1007/s00125-023-06038-8] [Medline: 37979006]
103. Sridhar GR, Lakshmi G. Ethical issues of artificial intelligence in diabetes mellitus. Med Res Arch 2023 Aug;11(8):1-8. [doi: 10.18103/mra.v11i8.4287]
104. Zhang L, Zhang L. Artificial intelligence for remote sensing data analysis: a review of challenges and opportunities. IEEE Geosci Remote Sens Mag 2022 Apr 13;10(2):270-294 [FREE Full text] [doi: 10.1109/MGRS.2022.3145854]
105. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar;28(3):231-237. [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
106. Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med Exp 2019 Dec 10;7(1):70. [doi: 10.1186/s40635-019-0286-6] [Medline: 31823128]
107. Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bull World Health Organ 2020 Apr 01;98(4):257-262. [doi: 10.2471/BLT.19.237289] [Medline: 32284649]
108. Julian E, Gianfrate F, Sola-Morales O, Mol P, Bergmann JF, Salmonson T, et al. How can a joint European health technology assessment provide an 'additional benefit' over the current standard of national assessments? : insights generated from a

XSL•FO
RenderX

multi-stakeholder survey in hematology/oncology. Health Econ Rev 2022 Jun 02;12(1):30. [doi: 10.1186/s13561-022-00379-7] [Medline: 35652987]

Abbreviations

AI: artificial intelligence
FDA: Food and Drug Administration
HTA: health technology assessment
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RWE: real-world evidence

Edited by Y Zhuang; submitted 02.08.23; peer-reviewed by A de Hond, E Vashishtha, I Cano, N Hendrix; comments to author 28.10.23; revised version received 17.12.23; accepted 28.12.23; published 13.05.24.

Please cite as:
Farah L, Borget I, Martelli N, Vallee A
Suitability of the Current Health Technology Assessment of Innovative Artificial Intelligence-Based Medical Devices: Scoping Literature Review
J Med Internet Res 2024;26:e51514
URL: https://www.jmir.org/2024/1/e51514
doi:10.2196/51514
PMID:38739911

©Line Farah, Isabelle Borget, Nicolas Martelli, Alexandre Vallee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Person-Generated Health Data in Women's Health: Scoping Review

Jalisa Lynn Karim[1], BA, BMath; Rachel Wan[1], BSc, BSN, RN; Rhea S Tabet[2], BSc; Derek S Chiu[3], BSc, MSc; Aline Talhouk[1], BA, MSc, PhD

[1]Department of Obstetrics and Gynaecology, University of British Columbia, Vancouver, BC, Canada

[2]Department of Pharmacology and Therapeutics, McGill University, Montréal, QC, Canada

[3]Department of Molecular Oncology, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Aline Talhouk, BA, MSc, PhD
Department of Obstetrics and Gynaecology
University of British Columbia
593 - 828 West 10th Ave
Vancouver, BC, V5Z 1M9
Canada
Phone: 1 604 875 3111
Email: a.talhouk@ubc.ca

Abstract

Background: The increased pervasiveness of digital health technology is producing large amounts of person-generated health data (PGHD). These data can empower people to monitor their health to promote prevention and management of disease. Women make up one of the largest groups of consumers of digital self-tracking technology.

Objective: In this scoping review, we aimed to (1) identify the different areas of women's health monitored using PGHD from connected health devices, (2) explore personal metrics collected through these technologies, and (3) synthesize facilitators of and barriers to women's adoption and use of connected health devices.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for scoping reviews, we searched 5 databases for articles published between January 1, 2015, and February 29, 2020. Papers were included if they targeted women or female individuals and incorporated digital health tools that collected PGHD outside a clinical setting.

Results: We included a total of 406 papers in this review. Articles on the use of PGHD for women steadily increased from 2015 to 2020. The health areas that the articles focused on spanned several topics, with pregnancy and the postpartum period being the most prevalent followed by cancer. Types of digital health used to collect PGHD included mobile apps, wearables, websites, the Internet of Things or smart devices, 2-way messaging, interactive voice response, and implantable devices. A thematic analysis of 41.4% (168/406) of the papers revealed 6 themes regarding facilitators of and barriers to women's use of digital health technology for collecting PGHD: (1) accessibility and connectivity, (2) design and functionality, (3) accuracy and credibility, (4) audience and adoption, (5) impact on community and health service, and (6) impact on health and behavior.

Conclusions: Leading up to the COVID-19 pandemic, the adoption of digital health tools to address women's health concerns was on a steady rise. The prominence of tools related to pregnancy and the postpartum period reflects the strong focus on reproductive health in women's health research and highlights opportunities for digital technology development in other women's health topics. Digital health technology was most acceptable when it was relevant to the target audience, was seen as user-friendly, and considered women's personalization preferences while also ensuring accuracy of measurements and credibility of information. The integration of digital technologies into clinical care will continue to evolve, and factors such as liability and health care provider workload need to be considered. While acknowledging the diversity of individual needs, the use of PGHD can positively impact the self-care management of numerous women's health journeys. The COVID-19 pandemic has ushered in increased adoption and acceptance of digital health technology. This study could serve as a baseline comparison for how this field has evolved as a result.

International Registered Report Identifier (IRRID): RR2-10.2196/26110

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53327) doi:10.2196/53327

KEYWORDS

digital health; women's health; mobile health; health app; wearables; femtech; self-tracking; personalized health; person-generated health data; patient-generated health data; scoping review; mobile phone

Introduction

Background

The practice of keeping notes to monitor one's health is not a recent phenomenon. Individuals have long recognized the benefits of tracking various health aspects, including the ability to be more active participants in managing their health, gaining a more complete picture of their health, and reducing the frequency of in-person appointments; however, this tracking was previously done through paper logs [1]. Today, with the proliferation of digital tools, self-tracking has significantly evolved and become more prevalent. The increasing pervasiveness of technology, particularly mobile phones, has seamlessly integrated it into our daily lives, making self-tracking more accessible and convenient than ever before [2]. Connected digital health technologies such as smartphones, wearables (eg, smartwatches), sensors, the Internet of Things (eg, internet-enabled weight scales), and web-based applications have permeated society and are increasingly adopted to collect and track health data. In 2021, a total of 87% of Canadians owned a smartphone, up by 73% from 2009 [3]. With >350,000 digital health apps accessible via these smartphones [4], approximately two-thirds of Canadians digitally track at least one aspect of their health [5]; similar statistics have been reported in the United States [6]. Moreover, since the introduction and popularization of fitness trackers in 2010, sensors and wearable devices have increasingly become part of daily life [2]. During the global COVID-19 pandemic, self-tracking took on even greater significance [7,8]. With the heightened awareness of health and the need for proactive measures, individuals have turned to self-tracking to monitor their well-being and make informed decisions. With this transformation, self-tracking has transcended its previous boundaries, offering individuals new opportunities to optimize their well-being and ushering in a new era of personalized health care [9-11].

Digital health tools have revolutionized the active and passive collection of health data through various applications and wearable devices. These various digital health tools collect and generate an unprecedented amount of data that can be used to glean insights into one's health. Person-generated health data (PGHD), which are clinically relevant data captured outside traditional care settings [12], provide valuable insights that empower users to self-monitor and reflect on their health. PGHD can refer to any data collected from wearable and smart devices as well as self-input information into platforms such as mobile apps and websites. By leveraging digital technologies, individuals can collect and store their health data, enabling them to actively manage their own health and monitor chronic conditions. Furthermore, the integration of these data with research presents an opportunity to improve the patients' experience and enhance personalized medicine. The recognition of this opportunity has started to take shape with patient-reported outcome measures and patient-reported experience measures being increasingly recognized as essential information to assess quality of care and prioritize patient-centered approaches and with mandatory assessment as part of clinical trials [13]. Seamlessly linking PGHD that are captured outside traditional care settings with clinical data and disease models can unlock new possibilities for tailored treatments and predictive informatics. The integration of digital health tools not only facilitates patient-provider communication but also offers opportunities for education, increased awareness, self-tracking, and self-monitoring without burdening health care resources. By focusing on the individual's experience, personalization, and prevention, digital health tools contribute to a patient-centered care paradigm that aims to optimize health care outcomes and improve overall well-being while empowering patients to take charge of their health.

In recent years, the emergence of femtech, defined as technology-driven solutions specifically designed to address women's health needs and concerns, has revolutionized the landscape of self-tracking and health care for women [14]. Femtech encompasses a wide range of digital tools, such as period-tracking apps, fertility monitors, pregnancy trackers, and menopause management platforms. These innovative solutions empower women to track and manage their reproductive health, menstrual cycles, and overall well-being with greater accuracy and ease. Femtech has not only provided women with personalized insights into their bodies but has also helped break taboos and encouraged open conversations about topics that were once stigmatized or ignored. The rapid growth of femtech has promoted access to women's health information, greater autonomy in decision-making, and enhanced overall health care experiences for women worldwide. It has become an integral part of the self-tracking movement, demonstrating the transformative power of technology in promoting women's health and well-being.

Objectives

In this study, we reviewed the use of digital tools and PGHD in women's health research, focusing on articles published between January 1, 2015, and February 29, 2020, before the COVID-19 pandemic. Our review encompassed various connected health devices, which included both passive data collection devices such as wearable sensors and active input devices such as smartphone apps and websites. This review sought to accomplish the following:

1. Identify the different areas of women's health and health-related behaviors monitored using PGHD from connected health devices.
2. Explore personal metrics collected through these technologies.

XSL•FO
RenderX

3. Synthesize facilitators and barriers that impact women's adoption and use of connected health devices in managing their health.

Methods

Overview

This scoping review was conducted based on our previously published protocol [15]. We adopted the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [16]. The completed checklist is provided in Multimedia Appendix 1 [16].

Search Strategy

The search strategy was designed in close collaboration with a reference librarian with input from the authors (JLK and AT). We searched a total of 5 databases: MEDLINE, Embase, APA PsycINFO, CINAHL Complete, and Web of Science Core Collection. Initial searches were completed in early March 2020. Searches were limited to articles published in 2015 or later because publications with the keyword "digital health" started to emerge in the literature around that time [17], and with the fast evolution of the field, previous articles may not be relevant to the current landscape. Keywords and subject headings were designed to search the literature for the intersection of the following 4 topics: women, health, digital devices, and tracking. The full search strategy, including a full list of search terms, was published with the protocol [15] and is available in Multimedia Appendix 2.

Eligibility Criteria

We were interested in digital technologies and interventions targeting women and people assigned female at birth. To be included in the review, studies needed to specifically target women, focus on female-only health topics (eg, menstruation), or only include female participants. We included a variety of publication types but excluded conference abstracts and conference reviews, editorials, letters, and comments due to the limited details in such literature.

We excluded articles that presented digital health tools designed for health care providers as we were primarily interested in devices and apps that women can engage with outside a clinical setting. Articles only discussing the use of real-time consultations, whether through video, phone, or web-based chat, were excluded. We excluded articles that described digital health tools used solely for educational purposes; to maintain the focus of the review on tracking or monitoring one's data for health, devices must have allowed users to input personal health data.

The complete inclusion and exclusion criteria are presented in Textbox 1. We decided to retain the original inclusion end date of February 29, 2020, to maintain a focus on the literature before the COVID-19 pandemic and avoid potential complexities caused by pandemic-related disruptions in research and health care practices. Concentrating on prepandemic literature also established a clear baseline for future comparisons and allowed us to maintain feasibility of completion without compromising quality given the broad scope of the review.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Published between January 1, 2015, and February 29, 2020
- Refers to a health issue that pertains only to women or comprises only female participants of any age
- Includes the use of connected health tools for tracking or monitoring some aspect of health, which could include smartphone apps, wearable devices, the Internet of Things (eg, Bluetooth- or internet-enabled glucometers, blood pressure cuffs, and weight scales), and implantable devices
- Involves data collection from the user of the connected health tool (ie, the user either manually inputs data into the device or they are automatically uploaded)
- The user must be able to interact with the app or device on her own at home (outside a clinical setting)
- Available in English

Exclusion criteria

- Not available in English
- Conference abstracts, conference reviews, editorials, letters, or comments
- Study media releases and user reviews of specific applications
- Research conducted on animals
- Research involving male participants
- Tracking of infants and children unless tracking breastfeeding (because breastfeeding is directly related to the mother's health and body)
- Devices or apps that are meant for health care provider use or use in a clinical setting only or cannot be used independently without a health care provider present
- Digital health tools that are only for educational or informational purposes and do not allow the user to enter or track her own data (ie, no information exchange)
- Telemedicine services (eg, live video consultations with health care providers)

XSL•FO
RenderX

Study Selection

We imported the results from the database searches to the Covidence systematic review software (Veritas Health Innovation). Covidence detected records believed to be duplicates, and these were manually checked before removing them. In addition, some articles were manually recognized as duplicates during the screening process and were subsequently tagged as duplicates and removed. Screening was conducted independently by at least 2 reviewers (JLK, RST, and AT) at both the abstract screening stage and the full-text screening stage. We attempted to contact the corresponding authors of articles that passed abstract screening when we were unable to locate the full text. Conflicts at either stage were discussed and agreed upon among the 3 authors involved in the screening process.

Data Charting and Deviations From the Protocol

The final list of data charting elements is provided in Textbox 2. Data charting for all elements except for usability and acceptability was conducted using Google Sheets created by the study team. The categories for different data charting options were initially created based on a small subset of articles and were discussed among the authors involved in the charting process. The team met regularly throughout the data charting process to discuss and refine coding categories that best summarized the data. Starting with more granular categories and later combining them into broader concepts was necessary to summarize the number of articles included in this review. For each article included, data were charted by one reviewer (RW or RST) and verified for accuracy by a second reviewer (JLK). Data were summarized in bar graphs, maps, and tables (JLK, RST, and DSC), as presented in the following sections. For the locations, we recorded the countries from which the participants were recruited (if applicable). If an article did not describe recruiting participants, then the countries of the authors were recorded based on the authors' affiliations.

Textbox 2. Data charting elements.

Article information

- Title
- Authors
- Year of first publication

Study characteristics

- Country or countries in which the research was conducted
- Research study type

Contexts for women's connected health

- Health areas of focus

Digital device details

- Types of digital health
- Metrics collected by the devices

Usability and acceptability

- Facilitators of and barriers to the use of the technologies (coded into themes)

For the thematic analysis, articles that mentioned any aspect of usability, acceptability, facilitators, or barriers to the use of digital health tools were imported into NVivo (R1 2020; QSR International). Coding was done independently by 2 reviewers (JLK and RW) and then combined through discussions. As with the data charting process, we initially coded more granularly and then grouped the detailed codes together later in the analytic process. Decisions on how to group the codes into themes and subthemes were made through group consensus (JLK, RW, and AT).

In our protocol, we indicated that we would extract the name of the device or app used in each study. While we did complete this step in our data charting, we have not presented the results in this paper. Several articles either did not specify the brand name (eg, only specified that it was a mobile app) or had digital health tools named after the study, so we did not find this information useful to showcase in our results. There were no other deviations from the published protocol.

Results

Study Selection

The searches identified 14,629 records that were imported into the Covidence software for deduplication and screening. After deduplication, a total of 9102 articles were screened for relevance, and 8545 (93.88%) were excluded based on title and abstract. From reading the full texts of the remaining 557 records, an additional 151 (27.1%) were excluded. The most common reasons for exclusion were the inability of study participants to enter or track their own data (58/151, 38.4%) or

XSL•FO
RenderX

because the digital health technology was designed to be used by or with a health care provider (48/151, 31.8%). The remaining 406 publications were included in the scoping review. Some of the included publications reported on the same research project; in those cases, all of them were included. Our search did not encounter any articles that directly addressed or mentioned the inclusion of intersex, transgender, or nonbinary participants. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram detailing the full study selection process is shown in Figure 1. The list of included articles sorted by health areas of focus can be found in Multimedia Appendix 3 [18-58].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

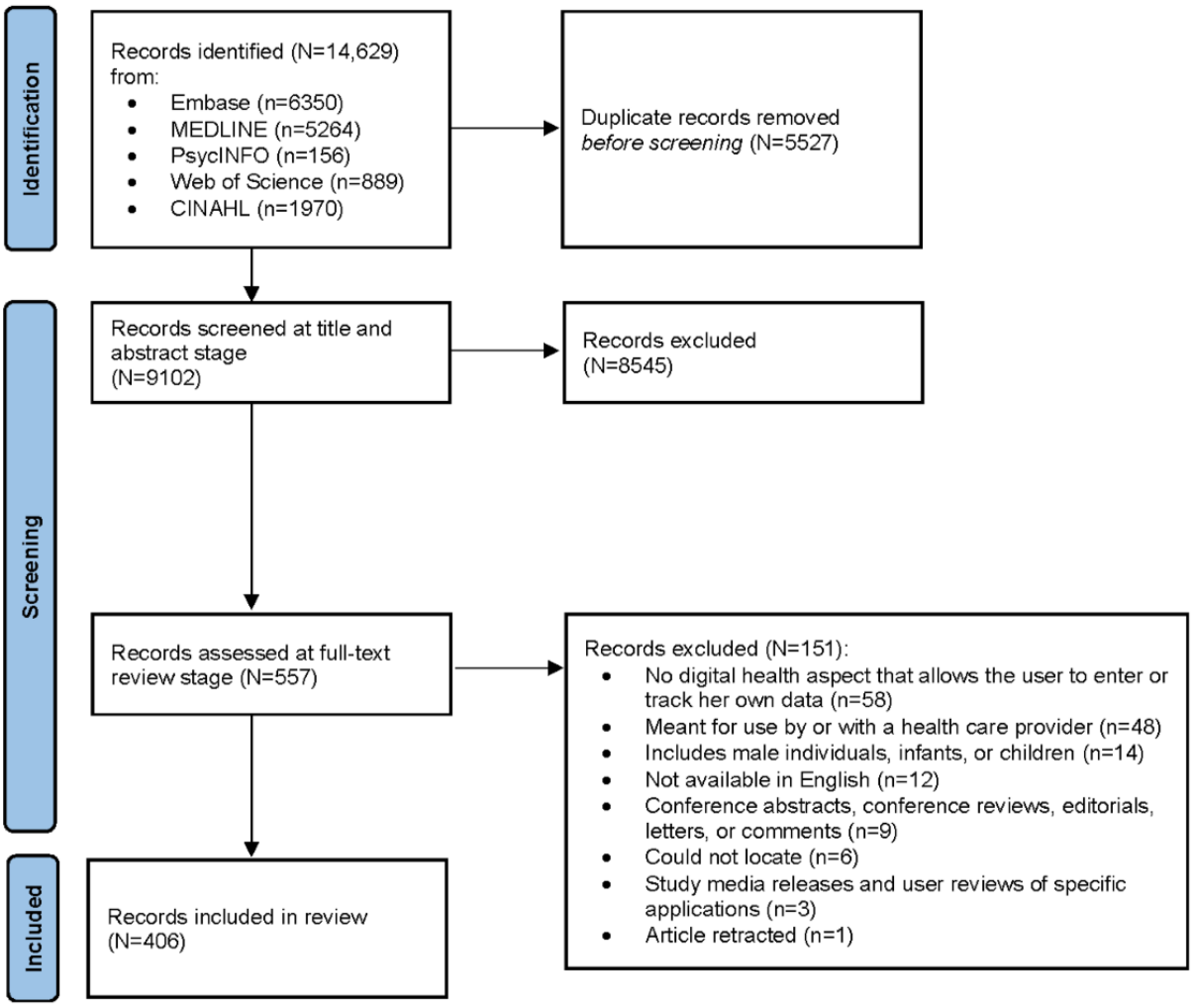

Year and Country

There was an increasing trend in number of publications per year, with 10.1% (41/406) of the articles published in 2015, a total of 13.3% (54/406) of the articles published in 2016, a total of 18% (73/406) of the articles published in 2017, a total of 26.4% (107/406) of the articles published in 2018, and 29.6% (120/406) of the articles published in 2019. Only 2.7% (11/406) of the publications were from 2020 because our cutoff date for inclusion was February 29, 2020.

Articles included in the review covered worldwide research, including every continent except Antarctica (Figure 2). As we only considered articles written in English, most of the articles were published in Western, English-speaking countries, primarily the United States (169/406, 41.6% of the articles), the United Kingdom (34/406, 8.4% of the articles), Australia (33/406, 8.1% of the articles), and Canada (19/406, 4.7% of the articles). Other countries where several included articles were published were China (13/406, 3.2% of the articles), the Netherlands (13/406, 3.2% of the articles), Spain (13/406, 3.2% of the articles), and Sweden (10/406, 2.5% of the articles).

Interestingly, of the 169 articles from the United States, 26 (15.4%) specifically focused on African American or Black, ethnic minority, or low-income women. One study from Singapore specifically included multiethnic women [18], and a study from Australia included Indigenous Australian women as their participants [19]. In addition, one review conducted by researchers in Australia looked specifically at studies with women from culturally and linguistically diverse backgrounds [20].

XSL•FO
RenderX

Figure 2. Number of publications by country.

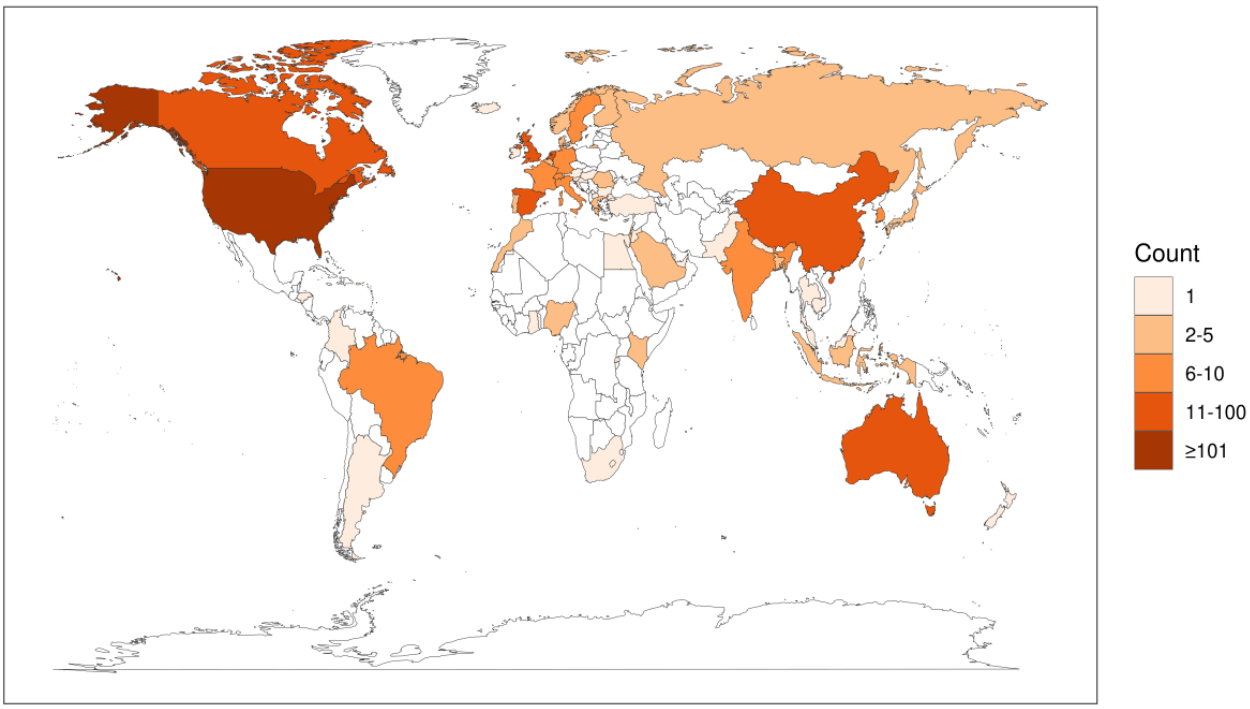

Study Types

The types of studies that used digital health tools in women's health research are reported in Figure 3 by year of publication (note that the articles could fall into more than one study category). The most common study type encountered was feasibility or acceptability studies (197/406, 48.5% of the articles, including 9/197, 4.6% protocols), followed by effectiveness studies (146/406, 36% of the articles, including 36/146, 24.7% protocols) and publications reporting on digital tool prototypes (73/406, 18% of the articles). Effectiveness studies reported on outcome measures of an intervention, including randomized and nonrandomized trials with one or more study arms. Reviews (of published literature, apps, or wearables), viewpoints, manuals, case studies, or analytical methods (56/406, 13.8% of the articles combined) were also encountered. Observational or correlative studies (44/406, 10.8% of the articles, including 3/44, 7% protocols) were studies that observed the health behaviors of individuals through digital health technologies without assessing the effectiveness of an intervention or analyzed associations between variables (eg, associations between heart rate and loss-of-control eating) [21]. Finally, measurement studies (23/406, 5.7% of the articles) reported on the validity, reliability, or accuracy of a digital health tool.

Figure 3. Study type by year of publication.

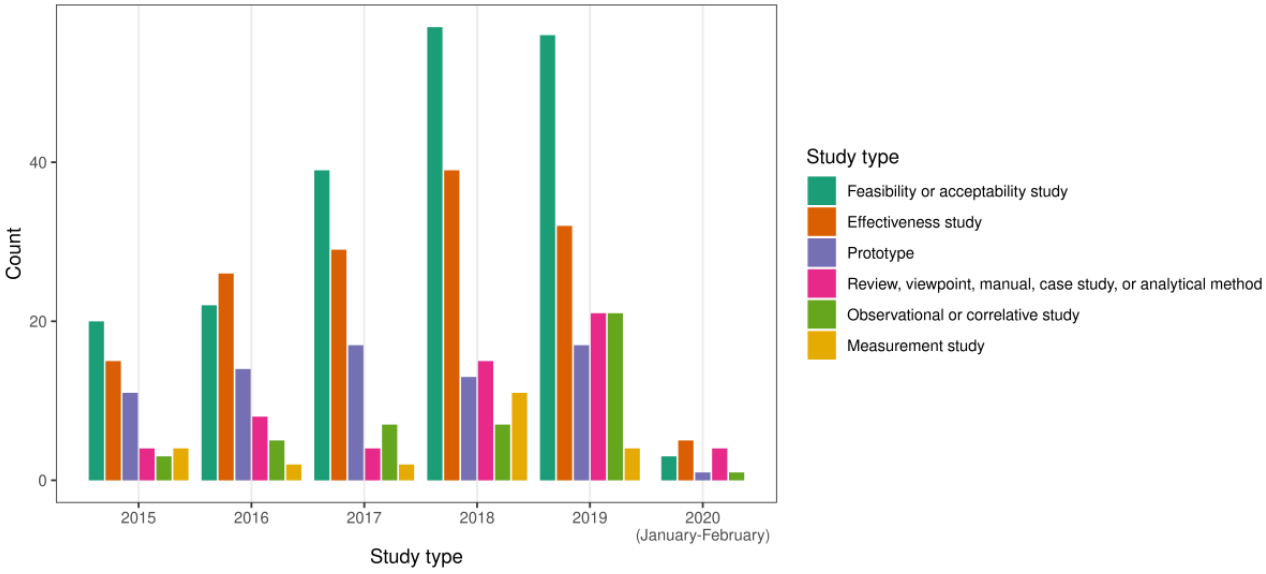

XSL•FO
RenderX

Health Areas of Focus

The analysis of the reviewed articles highlighted research in several recurring women's health areas of focus. A full breakdown of the health areas is reported in Table 1 (articles could fall into more than one health area). Pregnancy and the postpartum period emerged as the most prominent health area with 42.6% (173/406) of the articles. Within this category, there was a specific emphasis on general care and monitoring (45/173, 26% of the articles), physical activity and diet (34/173, 19.7% of the articles), and glucose monitoring (31/173, 17.9% of the articles). Cancer was identified as the second most common health area, with 19.5% (79/406) of the articles dedicated to its exploration. Specifically, a significant focus was observed on the relationship between cancer and cardiovascular health, with 47% (37/79) of the articles addressing this aspect. The impact of lifestyle on overall health and well-being was also addressed, with 14.3% (58/406) of the articles delving into physical activity, sedentary behavior, diet, weight, and obesity. Menstrual, sexual, and reproductive health were explored in 12.1% (49/406) of the articles to shed light on various aspects of women's reproductive health and associated concerns, with 76% (37/49) focusing on menstrual cycle tracking or fertility monitoring. Furthermore, 9.9% (40/406) of the articles were dedicated to chronic conditions (such as urinary incontinence, osteoporosis, and diabetes) with the aim of enhancing understanding and developing interventions for individuals living with chronic health conditions. To accommodate articles that did not fit within the primary health areas, an *Other* category comprising 6.4% (26/406) of the articles was established. This category included articles on athlete monitoring (10/26, 38% of the articles), such as heart rate monitoring during sports tournaments; mental health and quality of life (9/26, 35% of the articles); gender-based violence (3/26, 12% of the articles); and more. Finally, a small subset of 0.5% (2/406) of the articles did not align with any specific health area; these included a publication reporting results from a survey on African American women's willingness to participate in eHealth research [22] and a publication analyzing women's interactions with digital health technologies [23]. These articles were included because, although they did not discuss a specific health area, they still focused on women's use of digital health tools in general.

XSL•FO
RenderX

Table 1. Health areas of focus (N=406).

Health area of focus	Studies, n (%)
Pregnancy and the postpartum period (n=173)	
General care and monitoring	45 (26)
Physical activity, diet, gestational weight gain, postpartum weight retention, and weight management	34 (19.7)
Glucose monitoring and diabetes, including gestational diabetes	31 (17.9)
Blood pressure monitoring; pre-eclampsia; and hypertension, including gestational hypertension	22 (12.7)
Fetal monitoring and contraction monitoring	17 (9.8)
Depression, stress, self-efficacy, and mental well-being	11 (6.4)
Breastfeeding	7 (4)
Alcohol or tobacco use reduction	6 (3.5)
Cesarean section	2 (1.2)
HIV prevention and care	2 (1.2)
Sleep	2 (1.2)
Asthma	2 (1.2)
Vaccinations	2 (1.2)
Preterm delivery	1 (0.6)
Miscarriage	1 (0.6)
Postabortion care	1 (0.6)
Environmental exposures	1 (0.6)
Pregnancy research	1 (0.6)
Cancer (n=79)	
Cardiovascular or cardiometabolic health, physical activity, sedentary behavior, diet, weight, and obesity	37 (46.8)
Support, quality of life, self-efficacy, and other psychosocial measures	17 (21.5)
Prevention and early detection	9 (11.4)
Side effects from therapy and treatment and symptom management	7 (8.9)
Adjuvant endocrine therapy and medication adherence	6 (7.6)
Depression, stress, anxiety, and fear of cancer recurrence	4 (5.1)
Physical therapy and rehabilitation	3 (3.8)
Postoperative outcomes and care	3 (3.8)
Sleep	2 (2.5)
Cancer survivorship	1 (1.3)
Treatment-induced menopausal symptoms	1 (1.3)
Lifestyle (n=58)	
Physical activity, sedentary behavior, diet, weight, and obesity	50 (86.2)
Sleep	9 (15.5)
Alcohol dependency and problem drinking	2 (3.4)
Smoking cessation	2 (3.4)
Loss-of-control eating	1 (1.7)
Iron intake	1 (1.7)
Menstrual, sexual, and reproductive health (n=49)	
Menstrual cycle tracking and fertility monitoring	37 (75.5)
Pregnancy planning	4 (8.2)
Contraception	4 (8.2)

XSL•FO
RenderX

Health area of focus	Studies, n (%)
Menstrual pain, heavy menstrual bleeding, and bleeding disorders	3 (6.1)
Sexually transmitted infections	2 (4.1)
Menopause	2 (4.1)
PCOS[a]	2 (4.1)
Premenstrual symptoms	1 (2)
Chronic conditions (n=40)	
Urinary incontinence and pelvic muscle dysfunction	12 (30)
Osteoarthritis, osteoporosis, and bone health	8 (20)
CVD[b] or CVD risk	7 (17.5)
Diabetes	4 (10)
Asthma	2 (5)
COPD[c]	1 (2.5)
SLE[d]	1 (2.5)
IC[e] or BPS[f]	1 (2.5)
Epilepsy medication adherence	1 (2.5)
Chronic widespread pain	1 (2.5)
Atrial fibrillation	1 (2.5)
Prehypertension	1 (2.5)
ABL[g]	1 (2.5)
Other (n=26)	
Athlete monitoring	10 (38.5)
Depression, mood, stress, self-efficacy, and quality of life	9 (34.6)
Domestic violence, intimate partner violence, and gender-based violence	3 (11.5)
Environmental exposures and air pollution	2 (7.7)
Sweat analysis	1 (3.8)
Adolescent health research	1 (3.8)
Skin characteristics	1 (3.8)
No specific health area	2 (0.5)

[a]PCOS: polycystic ovary syndrome.
[b]CVD: cardiovascular disease.
[c]COPD: chronic obstructive pulmonary disease.
[d]SLE: systemic lupus erythematosus.
[e]IC: interstitial cystitis.
[f]BPS: bladder pain syndrome.
[g]ABL: accidental bowel leakage.

Figure 4 shows how the health areas of focus for women's use of digital health changed over the years that were included in the review (2015-2019 plus January 2020-February 2020). There was an increasing trend from 2015 to 2020 in the number of publications focusing on pregnancy and the postpartum period, as well as cancer and menstrual, sexual, and reproductive health. However, articles focused on women's use of digital health for lifestyle-related topics and chronic conditions did not see a notable increase over those years.

XSL•FO
RenderX

Figure 4. Health areas of focus by year of publication.

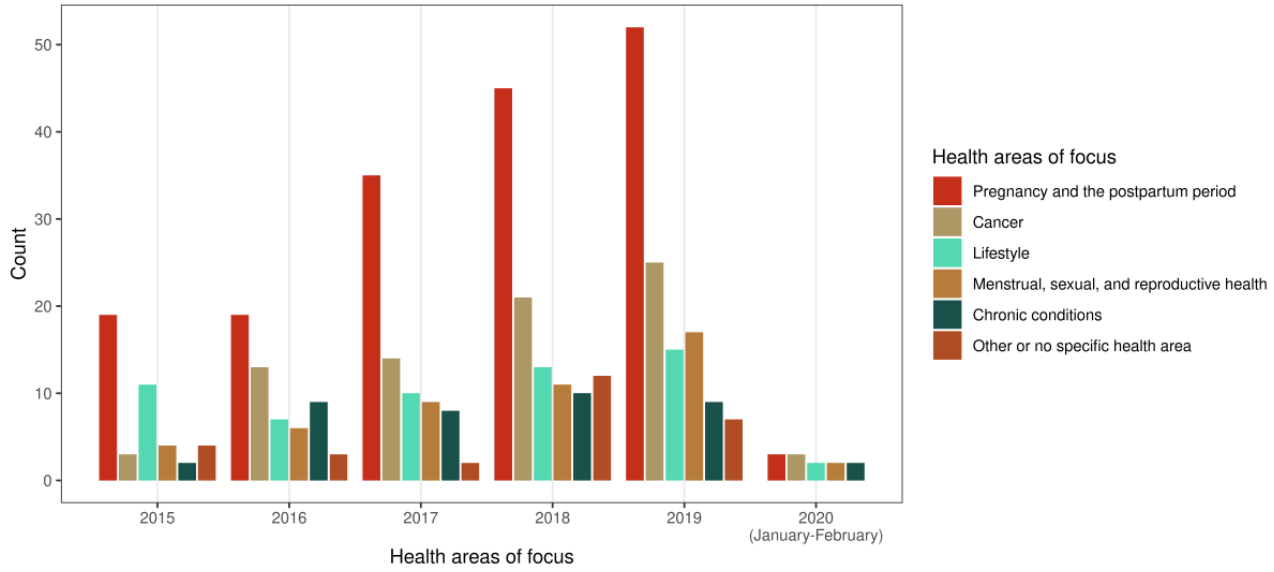

Type of Digital Health and Metrics Collected

Within the articles reviewed, smartphone, mobile, or tablet apps emerged as the most prevalent type of digital health (295/406, 72.7% of the articles), followed by wearable devices (165/406, 40.6% of the articles) and websites or patient portals (93/406, 22.9% of the articles). Other types of technology were not investigated as much. For example, 13.5% (55/406) of the articles addressed smart devices or the Internet of Things (referring to objects with sensors that connect to a network, such as Bluetooth-enabled glucometers and blood pressure machines). Finally, 7.4% (30/406) of the articles reported on 2-way messaging, 1% (4/406) of the articles reported on interactive voice response telephone calls, and only 0.5% (2/406) of the articles reported on implantable devices. With respect to the metrics collected, we found >250 metrics, such as heart rate, number of steps, mood, ovulation test results, and days of menstruation. A full list of the metrics is reported in Multimedia Appendix 4.

Thematic Analysis

Overview

Of the 406 articles included in this scoping review, 168 (41.4%) mentioned usability, acceptability, facilitators, or barriers to the use of digital health tools at least once. Our thematic analysis identified 6 themes: (1) accessibility and connectivity, (2) design and functionality, (3) accuracy and credibility, (4) audience and adoption, (5) impact on community and health service, and (6) impact on health and behavior. The themes are described in further detail in the following sections.

The thematic analysis detailed in the following sections is primarily based on the views of the participants in the studies we reviewed to provide a user perspective; however, one subsection in theme 5 focuses on the health care provider perspective.

Theme 1: Accessibility and Connectivity

The accessibility and connectivity of digital technologies emerged as an important theme with two subthemes: (1) cost and convenience and (2) connectivity, compatibility, and software issues.

Cost and Convenience

Our analysis revealed that the cost and convenience of digital tools collecting PGHD are important factors that can impact their adoption and use. On the one hand, digital health technologies can be seen as more affordable compared to traditional health care visits and more accessible to a wider range of people, including those of a lower socioeconomic status. On the other hand, they can also be perceived as too expensive and novelty items, and associated extra costs such as data plans can also be a barrier for some people. Because PGHD can be collected and entered throughout the day in real time, and because most people carry a phone around with them every day, these technologies offer greater convenience than traditional in-person health care encounters by providing anytime, anywhere virtual access and putting information at people's fingertips through smartphones and web platforms. One user spoke about an in-app treatment program:

> *That was what was so good about this, I can do this at home myself, no need to book an appointment, find the time and suit others, and you know, that process of booking a time.* [24]

Some inconvenient aspects of digital health technologies include uncomfortable wearables that are too bulky, difficulty of use, or not fitting into the users' lifestyles, as noted in one article:

> *Women also mentioned that the comfort of the wearable sensors was a barrier. Comfort became a barrier for some women during exercise and hot weather.* [25]

Devices with a short battery life and wearables that are not water resistant are also considered inconvenient as they require the user to frequently remember to charge the device or put the wearable back on after water-based activities. Certain restrictions, such as not being able to wear a device in a workplace, can also create inconvenient barriers for some users.

XSL•FO
RenderX

Connectivity, Compatibility, and Software Issues

Factors related to connectivity and other issues such as device synchronization, freezing, or disconnection can significantly impact the user experience and engagement with digital health tools. For example, the dependence on mobile and internet access can be a disadvantage. Cellphone and network coverage limitations can pose an important barrier in rural areas or during travel. Where mobile data or Wi-Fi connection are limited, people may struggle to use digital health tools that require internet connection; this can create disparities in access to health care resources, particularly for those of a lower socioeconomic status or living in remote communities with limited infrastructure. Incompatibility between operating systems such as Android and iOS, iPhone and iPad, or various browsers can also be an important barrier to accessing digital health technology.

Software issues can significantly impact the user experience of digital health technologies. Broken links can prevent users from accessing valuable information or features within apps or websites. App crashes can discourage users from engaging with the digital health tool altogether. In addition, slow loading times can negatively impact the user experience, making it challenging for users to access information or features quickly and efficiently.

Theme 2: Design and Functionality

The second theme centered on design and functionality and included four subthemes: (1) appearance and design; (2) functionality and features; (3) personalization; and (4) safety, privacy, and security.

Appearance and Design

Appearance and design play important roles in the success of connected health devices. In terms of app design, the color scheme and layout facilitate user-friendliness. Bad formatting can make it difficult for users to read or understand the content of an app or website. Font size that is too small can be challenging for those with visual impairments, and some color combinations can be difficult to read. The quality of the images used in digital health interventions can also impact user experience, with low-quality images potentially making it difficult for users to interpret the information being presented.

Apps that are visually appealing and easy to use are more likely to be successful. When it comes to wearables, women tend to prefer sleek, understated designs that are esthetically pleasing without being bulky. A sleek design can encourage use of the wearable. For example, some women consider their wearable to be a fashion item that sparks conversation, which encourages their continued use of the device, as illustrated in a participant quote:

> *Um, that it's, like, kind of stylish, like, I feel, like, cool that I wear one. A lot of people ask me, they're like, oh, which one is that, like, is that a Fitbit, is that an Apple watch? It has the interchangeable bands and stuff like that, so, you can, like, change the color of it and everything. It's like a conversation piece.* [26]

Other women prefer more discretion in the design of wearables and their size or in the app icon on their mobile device because they do not want to reveal the purpose of the device to others. People may feel self-conscious when wearing the device or using the app, especially if it reveals their medical condition. For example, the following quote is from a study that incorporated a sensor band worn on the wrist to help female undergraduate students with problematic drinking:

> *P310 noted that while in class, "my professor commented on it which made me feel awkward."* [27]

Functionality and Features

In terms of functionality, the availability of clinical interpretation of user data is deemed essential, and health warnings based on recorded PGHD are noted to be helpful. Moreover, notifications and reminders are also useful for improving adherence to self-tracking and maintaining goals, and users appreciate receiving automated SMS text messages and feedback on progress. Actionable advice is seen as very important, and women expressed a desire for more interaction and the ability to integrate with other apps. For example, users want the ability to access information from their health record and to be able to see graphical summaries of their data over time. Regarding the presentation of information, users appreciate concise information written in simple language. Choice of words is also perceived as especially important to ensure that the information is easy to understand. People enjoy the gamification of content, and the graphical presentation of results is found to be informative.

The ability to upload multimedia and the ability to customize the application's displays and notifications are noted as features that improve user engagement and satisfaction. The ability to record voice notes and consultations within apps is noted as a desirable feature, as well as having the option to book appointments directly through apps. Women also want the option to sync their desktop or phone calendar with apps to remind them of medical appointments and prescription requests, as some researchers noted:

> *Women could see the potential usefulness of being reminded to order their next prescription through the electronic alerts system. They found managing the monthly prescription requests challenging long-term and found setting up the reminder easy with the alert popping up on their phone or tablet.* [28]

> *The application also allows women to set appointment reminders to ensure she is not missing her appointments and developing gaps in her care [...] "It allows me to remain organized for my visits to my OB with concerns, questions, symptoms I have experienced since my last visit."* [29]

Issues that negatively impact user engagement and outcomes are the inability to edit information or unsubscribe from notifications, which are sometimes thought to be either inconvenient or intrusive, as well as the presence of advertisements within the app. Ease of use is essential as apps or websites that are difficult to navigate can discourage users from engaging with them. For example, a study including the use of a mobile phone app reported the following:

XSL•FO
RenderX

> *A hindrance and disliked aspect was the difficulty in navigating through the app (eg, no back button, clunkiness, and the inability of participants to edit their inputted daily goals) as well as a lack of color and visuals within the app, giving it a clinical appearance.* [30]

Some women are not comfortable answering questions that they consider intrusive, such as those related to sexual health. They rely on applications to provide trusted information about their condition and want suggestions for additional resources such as website links and local information.

Personalization

Women generally expressed a desire for greater personalization across several features within digital health tools. Messages and notifications that are personalized to the user's health and self-tracking history and goals are more motivational and less likely to be ignored or perceived as irritating. Even factors such as using a first name in messages from the app make women feel like the messages are more personal and supportive.

Users have individual preferences when it comes to the frequency and timing of notifications, and it is important for digital health apps to allow for the customization of these settings as they can greatly impact user engagement and adherence. Moreover, users expressed a desire for the ability to customize their goals and the metrics they tracked. For example, they may want to change their goals in an app when their life circumstances change (eg, moving, starting a new job, becoming pregnant, or sustaining an injury). The ability to customize the dashboard of an app or website according to the user's goals was also expressed as a desired feature. The ability to make these customizations will improve their adherence in the long run as their goals evolve.

When it comes to wearables, their placement on the body influences users' preference and adherence to their use. For example, some women may prefer a wrist-worn device, whereas others may prefer a chest strap, a ring, or a device worn on the waist or ankle. The type of activity being monitored may also influence placement preference. A wrist-worn device may be more appropriate for monitoring steps, whereas a chest strap may be better suited for monitoring heart rate during exercise. Furthermore, placement preference may also be influenced by factors such as comfort, convenience, and visibility. A user may prefer a wrist-worn device because it is more visible and easier to access, whereas another user may prefer a device worn on the waist because it is less obtrusive and more comfortable during exercise or sleep. For instance, one study found the following:

> *Eight of the participants (40%) reported at some point of the long study period that the smart wristbands were uncomfortable to wear, especially at night. The wristbands irritated the skin, possibly due to pregnancy-related swelling.* [31]

Finally, users have different preferences for how they want information to be presented in an app or website. Some people prefer to read content that is written out with citations and links to external websites. Others enjoy learning content from videos or audio recordings. When looking at their trends and progress, some users like to look at detailed graphs showing their daily progress, whereas others prefer to look at the data occasionally and only receive high-level information. The challenges concerning personalization were articulated by several authors:

> *It's a difficult one. Some women want the full picture to fully understand what they are taking. Others want a black and white sketch, but not the details. They just want to know enough. Others do not want to see the picture, they just want to get on with it without knowing too much. Catering for all is a challenge.* [28]

Safety, Privacy, and Security

Women are sometimes concerned about the physical safety of certain devices. For example, some mothers worried about their wearable wristbands scratching their babies [31]. Others worried about the effects of wearable devices on their skin, as expressed by a participant:

> *It's weird because it does have a little laser thing on it, and I wonder if that's, like, harming my skin (laughing). Like, I'll sleep in it, and when I wake up I'll have a red spot on my arm, it's itchy sometime or sensitive, and I think it's because of the laser thing, but I don't really know.* [26]

Some women are concerned about the privacy and security of digital health technologies and expect appropriate safeguards to be implemented in the tools they use. However, privacy and data security are not a concern for all women:

> *As I said, I'm very critical about patient data in general, especially in terms of data security...If you have a free app, it really depends on what happens to the private data. As a matter of fact, usually the information is stored on the app itself, and so other apps might gain access to the data easily.* [32]

> *The survey revealed a low level of concern about issues relating to privacy or security of personal data. This suggests that privacy concerns were secondary to the benefits offered by uploading personal details into apps to provide the type of customisation they seek.* [33]

Researchers also shared that some users perceived there to be more privacy when using an app as compared to traditional ways of communicating:

> *Some participants perceived the storage of their glucose levels on the smartphone as more secure than their current registration in a booklet.* [34]

> *Women, particularly those who worked outside of the home, also commented that they appreciated the added convenience and privacy of this [text-based] communication method over phone-based communication.* [34]

XSL•FO
RenderX

Theme 3: Accuracy and Credibility

In theme 3, we identified accuracy and credibility as important factors for acceptability considerations in digital health technologies.

The accuracy of digital health can impact user trust and adoption. Digital health tools enable users to keep track of their health, symptoms, and behaviors over time without relying on memory recall, which can be inaccurate or incomplete. Many studies reported that digital tracking can lead to more accurate data collection compared to paper-based methods. For example, at-home measurements of blood pressure and other vital signs have been found to be more accurate than those taken in a hospital or clinic setting. In some cases, apps are even able to accurately predict users' menstrual cycles and mood changes. In addition, food diaries and activity trackers are often found to be more accurate when tracked within the app compared to using traditional paper-based methods. As the following participant conveyed, digital health may also make it easier for patients to tell the truth about their habits or health concerns:

> *I like this principle because...I know exactly, that via tablet one would admit things you wouldn't necessarily tell the doctor or nurse. So, for starters, you can state it in the application. Of course, a conversation shouldn't be missed afterwards, but this might make it easier for you to overcome yourself.* [32]

However, accuracy can still be an issue in digital health. Different devices can produce different measurements, and some devices may miscount steps, the intensity of workouts, or the quantity and quality of sleep. For example, some women reported devices not tracking their steps while pushing a grocery cart or stroller, whereas others found that their steps were overcounted due to arm movements while they were seated. In addition, some users reported that food tracking options in apps were limited and did not include foods from their culture. Therefore, users may perceive digital health tools as not being representative of their true activity, which may lead them to discontinue the use of the devices. The following participant quote refers to a wrist-worn activity tracker:

> *Out paddling and we're huffing and puffing and barely breathing and this isn't even triggering anything. So it shows [...] that our 150 minute goal is like 60 or half of that. But we've actually put in the effort and then you just give up after a while. Like there's no way I can make this.* [35]

Women often prefer evidence-based health information (eg, explanations of conditions and symptoms and health advice) from a trustworthy source, such as an app curated from up-to-date and evidence-based research, over general internet searches. Users reported that the information provided in some apps was incomplete or inaccurate, with gaps in content or contradictory information that diminished their trustworthiness. In such cases, users may still prefer to talk to a health professional for more trustworthy information. Some women may also find it challenging to trust information that does not disclose sources as they are unsure of its reliability. Devices that are endorsed by, cite, and link to trustworthy health sources are more appealing to users. When sharing results from a web-based survey, the authors of one study reported the following:

> *Some respondents were specific about from where such advice should come, stating that they wanted expert, credible and up-to-date advice while others noted that they would like to see more Australian-specific or locally-based information in apps or apps that were not linked to the manufacturers of pregnancy or baby products.* [33]

Theme 4: Audience and Adoption

Our fourth theme concerns audience and adoption, which includes two subthemes: (1) demographics and inclusivity and (2) timing and circumstances.

Demographics and Inclusivity

One of the challenges with digital health is to avoid one-size-fits-all interventions and to strive to tailor interventions to address the specific needs of different populations. Digital health that targets specific demographic groups or specific health conditions may increase the adoption of digital tools in those populations. That said, even when targeting people with specific health conditions as the audience, attention must be paid to the language and content in apps and websites. Some researchers noted that women did not want to participate or continue in their study because they did not want to constantly be confronted with their disease. Too much of a focus on disease and ill health can deter women from engaging with the tools, as commented on by some authors:

> *All but one participant preferred text content that focused on health and physical activity rather than content explicit to cancer.* [36]

> *The women emphasized that less attention should be paid to chronic disease management and medication as the only treatment option. [...] it was important to explain the implications of the result of the scan and the risk of fractures in a way that will not place the women in a sickness role unnecessarily. [...] The knowledge base of osteoporosis should focus on osteoporosis as a common condition instead of a chronic bone disease.* [37]

Younger women are often more familiar with and more comfortable using digital technology and, therefore, are more likely to use and adhere to a digital health protocol. Users with low technology skills want more training on how to use the digital health tools properly. Little provision is made for those for whom English is not their primary language, which can limit the accessibility and usefulness of digital health interventions. Factors such as language barriers, cultural beliefs, or lack of access to technology may lead to less adoption by some people belonging to ethnic minority groups. The relevance and usefulness of digital health may also vary based on geographic location.

Digital health tools are negatively perceived by some users if not designed to be inclusive of attributes such as body type or gender. For example, users prefer applications that use pictures

XSL•FO
RenderX

or models that represent a diverse range of body sizes. Digital health technologies may not be gender inclusive and can conflate sex and gender. It is important to consider the unique health needs and experiences of individuals across the gender spectrum, as several researchers reported:

> *Participants commented on an exercise demonstration video and recommended that the model should have an "everyday-look" (e.g. plain clothes, jewellery). Also a choice of models of different ages to engage a wider range of patients and help them to relate or identify with the model was proposed.* [38]

> *[Participant quote]: Maybe the body image it presents...like on a lot of apps, the people doing it looked like they were athletes already. And maybe they should have more people that look normal.* [39]

> *Two women commented on the gendered design of most FTAs. FTA092 commented that "I chose Clue because it's the only app that wasn't pink." FTA051 also found the gendered design of her previous app insulting; "my last app had a pink flower and was called MyDays or something ...I felt like they were trying to lure me in with this kind of 'women's' approach" (FTA051). She subsequently stopped using that app and downloaded Clue.* [40]

Timing and Circumstances

Individuals are more motivated to use digital health tools during times of illness or when they have a specific health goal in mind. The introduction of technology at the appropriate time impacts the utility and effectiveness of digital health interventions, especially when they are integrated into existing health care systems and routines. Digital health apps need to account for existing medical conditions or medical history to ensure accurate and complete information. For example, technologies that do not provide an option to indicate current pregnancy are perceived as frustrating to users as the in-app goals or notifications can be irrelevant and inconsiderate of their current limitations. In a focus group, one mother shared the following:

> *I get frustrated with the Garmin [smartwatch] because I wear my watch during the night so it tracks my sleeping as well. Then it gives you like an insight—so a little note will pop up and you know whether your sleep has been really regular or you've had irregular sleep. I wish that there was a thing that during pregnancy where that I could put in and say I'm pregnant, because I got those notes that your sleep is really irregular, and I was like, "Because I'm pregnant!"* [23]

Users who are not experiencing symptoms or who perceive their health to be good are less likely to adopt digital health tools as they may not perceive any benefit from using them. Moreover, those who are already tracking their health using other methods (eg, paper-based tracking) are less interested in trying a new digital health tool. Similarly, regarding wearables, some people may already have a wearable and be less interested in having an additional wearable device.

Theme 5: Impact on Community and Health Service

This theme considers the impact of PGHD on community and health service, with three subthemes: (1) communication and community support, (2) clinical integration, and (3) health care provider perspective.

Communication and Community Support

One of the many perceived benefits by users of digital health interventions is the sense of community that these platforms enable. Even though some women reported feeling uncomfortable sharing personal information with strangers in a virtual group, most found that the ability to connect with others who shared similar experiences provided a sense of belonging and support that was motivating and reassuring, as shared by one woman:

> *What I did love about the apps is the forums. So if you have a weird pain or, you know, you have cramp in your legs at three a.m., you can get on your phone straight away, and you can get support by the women who are going through the same thing.* [41]

Discussion forums and social media platforms associated with digital health interventions are perceived as helpful for connecting with others, sharing personal stories, and receiving support. Digital health interventions can also help women elicit support from friends and family to stay motivated and achieve health goals. For example, researchers who reported on women's experiences of an app for stress urinary incontinence shared that some participants found it easier to talk to friends about an app for pelvic floor muscle training rather than talk about incontinence [24]. This can enable increased accountability and further encourage adherence to the intervention. One woman spoke about how her family supported her engagement with a digital health intervention for physical activity maintenance among female cancer survivors:

> *My husband's a good motivator. When I say I'm going for a walk, he'll go with me...with my sister-in-law and her kids, it's they want to go with me; so it's how many steps have you got today? Or, are we going to go for a walk. That kind of thing. And with my husband and my daughter it's, "how many steps did you get today, did you do your workout, let us get it going."* [36]

In addition to support from family, friends, and community members, these digital platforms can provide an alternative to speaking with a health care provider in person. Asynchronous communication with health care providers is helpful especially for those who may not have easy access to in-person visits or for those who are uncomfortable discussing sensitive information face-to-face. Records of PGHD can also improve the ability to gather and share details with health care providers about symptoms that are difficult to remember during an in-person visit.

Clinical Integration

Women are more willing to participate in digital health interventions if they perceive that they have a direct impact on their clinical care. They appreciate the idea that their health is being monitored and that someone is keeping an eye on their

XSL•FO
RenderX

data. Furthermore, women want to see more integration of their clinical test results within their digital health apps and websites. This increases their motivation to adhere to the interventions prescribed through the digital health application.

It was noted that physicians and other health care providers play a crucial role in promoting the use of digital health interventions among patients. As noted in the following participant quote, women enjoy being able to communicate with a health care provider through digital health:

> *I like it because you can tell the doctor what's going on and submit it to your doctor, that is the main reason I like it because you can talk directly to your doctor and tell them what is going on without going in or calling.* [42]

Women are more likely to adopt and use technology if it is recommended by their health care providers, family members, or friends. Women reported that digital health interventions were more effective when they were supported by a health care team. For example, having access to a health coach or counselor or receiving feedback from a health care provider on their progress increases their motivation to adhere to the interventions. This support also provides reassurance that they are on the right track toward achieving their health goals. However, some patients become frustrated when they receive conflicting advice from the digital health tool and their health care provider.

Health Care Provider Perspective

Some articles included thoughts from health care providers on digital health tools collecting PGHD [20,37,43-55]. From the health care provider perspective, digital health can offer several benefits, including the ability to monitor patients' adherence to treatment and interventions. This can be particularly helpful for patients with chronic conditions that require ongoing management. Providers can use digital health tools to track patients' progress and identify any potential issues that may require further attention, which can lead to improved clinical outcomes and reduce unnecessary consultations. For instance, one provider learned about their patient's anxiousness through a mobile health intervention:

> *I didn't know my patient was feeling anxious...But when she wrote it down, we could talk about it...* [43]

Some health care providers expressed that digital health tracking could give them a more accurate picture of their patients' activities and adherence to treatments. In a study about perspectives on a sensor attached to pills that can send data such as date and time of ingestion, a provider commented the following:

> *A positive would be data and getting a better grip on compliance. (...) I'm making sure the patient is adhering - assuming that the patient is taking everything inside of that blister, you can have confirmation of that.* [44]

In addition, digital health can improve the efficiency of care delivery by providing education and resources directly to patients. This can help patients better understand their condition, treatment options, and self-management strategies, which can lead to better health outcomes.

However, it was also noted that digital health interventions should not replace in-person visits but rather complement them. Some health care providers are concerned about overreliance on digital health tools as well as the potential for misinterpretation of the data they provide. There may be a lack of feedback on the correct use of interventions, such as interpretations of medical advice provided, and health care providers have raised concerns about the safety and trustworthiness of the medical advice generated by the digital health tools. Health care providers especially worry about medico-legal effects of having information from digital health tools taken out of context or without considering the full picture of the user's history and health, as demonstrated in the following quotes:

> *As a health care professional, I'm just mindful that if there was a video of me up there talking, if that was taken out of context or shared with another person where that information was not appropriate, that's a concern to me.* [45]

> *One anesthesiologist raised, "Who has access to the responses that I provide? Because if a patient receives information from me which they hold onto and is taken out of context, in a medical–legal situation, then that's a big issue as well."* [46]

Providers may also find that the abundance of information generated by digital health tools can be overwhelming and time-consuming to manage, adding to an already hectic workflow and blurring professional boundaries. Large volumes of alerts and notifications from digital health tools can be disruptive to health care providers, who expressed the need to set boundaries regarding how and when they engaged with digital health tools. In a study reporting on perspectives about digital health from key informants (health care providers and researchers), one participant shared their thoughts on the potential for digital health to increase workload and liability:

> *Sometimes the more information that we provide for them (doctors), the more work and liability we give them, right? So if they get so much information that becomes actionable but they are overwhelmed, now they would be obligated to do something with this patient, they are in a chain of distribution, a chain of liability.* [44]

Theme 6: Impact on Health and Behavior

Finally, our sixth theme describes the impact of PGHD on health and health behaviors.

Several studies reported that digital health interventions helped users stay motivated and, in turn, improved their health habits and behaviors, such as adherence to medication, physical activity, and healthy eating. The ability of users to look back at their data helps them identify patterns in their health and behaviors, which increases their awareness of their health and habits. The awareness then allows them to be more mindful of their habits and encourages self-reflection, thus promoting a deeper understanding of their health and well-being. The tracking of patterns in their health, combined with the educational component of some digital health tools, helps users

XSL•FO
RenderX

come up with better self-management strategies and feel more confident in their ability to reach their health goals, giving them a greater sense of self-efficacy and control over their health. In a digital health intervention aimed at treating lymphedema following breast cancer treatment, a participant spoke of changes in her awareness of symptoms and improvements:

> *It helped me realize that I had excess fluid. My arms got lighter each time I did the exercises. My arms began to feel less heavy. It noticed it in my clothes as well.* [56]

Digital health interventions are often reported to positively impact the mental health and well-being of individuals. Women reported improvements in their mood, emotional state, and coping abilities. They also reported a reduction in stress and anxiety levels, which can lead to improvements in overall health outcomes. The digital health tools provide users with a sense of support and accountability as well as feelings of accomplishment when meeting their goals.

However, it is important to note that, while digital health interventions can have many benefits, they may not be suitable for everyone and may even have negative effects on some individuals. For example, some users reported increased anxiety due to excessive monitoring or notifications, and others reported negative effects on their thoughts or worsening of symptoms related to health conditions. Some users found that self-tracking made them more attached to their phones, less likely to engage in social activities, and more isolated overall. Care should be taken to ensure that users do not become obsessive about self-tracking as this can be counterproductive or even harmful. Being hyperfocused on their symptoms or health condition could be distressing and even detrimental to their overall well-being. Therefore, it is important to carefully monitor the use of digital health interventions and adjust them as needed to ensure the best possible outcomes for each individual. One woman spoke about her overreliance on an app used to track breastfeeding:

> *I stopped using it because um I thought I'm being too anal about this...being too concerned about it, I just need to stress less, and just go with the flow and just be a bit more relaxed about it...so, that's why I stopped using it completely, and then I think the breastfeeding improved from there 'cause I was worrying about it less.* [57]

Table 2 provides a summary of the thematic analysis grouped into barriers and facilitators. It is worth noting that many things are both a barrier and a facilitator (eg, cost) depending on the individual. In addition, the presence of a specific feature may be a facilitator, whereas the absence of it may be a barrier.

XSL•FO
RenderX

Table 2. Summary of barriers to and facilitators of women's use of digital health.

Theme	Barriers	Facilitators
Accessibility and connectivity	• Extra costs • Uncomfortable physical design (for wearables) • Short battery life • Not compatible with daily activities • Network connectivity and device compatibility • Unreliable software	• Affordable • Convenient
Design and functionality	• Poor formatting or image quality • Too noticeable or medical • Inability to silence notifications and advertisements • Complex navigation • Intrusive questions • Physical safety • Trust, privacy, and security	• Streamlined user interface • Pleasant to look at • Decision support and clinical interpretation of data • Notifications or reminders and progress input • Graphical data summaries • Gamification • Feature and display customization • Personalization of content and output • Trust, privacy, and security
Accuracy and credibility	• Validity of measures • Limited data entry options • Failure to disclose information sources	• Accurate sensors and algorithms • Reliable data source • Evidence-based advice
Audience and adoption	• Too much focus on disease • Low digital literacy • Language and cultural barriers	• Designed for a specific user group • High digital literacy • Inclusive language and images • Focused health goals • Accounts for current health state
Impact on community and health service	• Overreliance on digital health • Potential for misinterpretation of data • Too much information output • Blur professional boundaries	• Creation of community • Facilitates communication • Social support from friends and family • Asynchronous or more accessible alternative to in-person visits • Supportive care providers • Integration with clinical care • Continuation of care • Education provision
Impact on health and behavior	• Stress caused by hyperfocus on tracking data • Decreased engagement in social activities	• Motivational tool • Supports self-awareness and self-reflection through pattern recognition • Increases self-efficacy • Improves symptoms • Reduces stress

Discussion

Principal Findings

In this scoping review, we summarized information from 406 articles on digital technologies collecting PGHD and how they have been used in women's health research. We found a steady increase in articles meeting our inclusion criteria from 2015 to 2020, indicating an increasing trend in the uptake and use of digital health tools in women's health research before the COVID-19 pandemic. Most included studies (310/406, 76.4%) were feasibility or acceptability studies, effectiveness studies, or reports of digital tool prototypes. Most studies (299/406, 73.6%) focused on tracking conditions related to pregnancy or the postpartum period, cancer survivorship, or menstrual, sexual, and reproductive health. Several types of digital health were represented, with the most common being apps, wearable devices, and websites or patient portals. Through our thematic analysis, we found several considerations of facilitators of and barriers to using digital health tools, including the accessibility and convenience of the tools, visual appearance, device functionality and ability to personalize the user experience, and accuracy of the algorithms and information provided. It is also important to consider the target audience to optimize the adoption of the tools. Engagement with digital health tools may help users improve their health and health-related behaviors and gather support from friends, family, and other digital health users. Women are more likely to use digital health if it is recommended by a health care provider, but there are both benefits and challenges that health care providers may face if considering integrating digital health technology into clinical practice.

A previously published scoping review focused on information and communications technologies as a tool for women's empowerment [59]. They reported that the concept of empowerment appeared in various ways with no clear consensus

XSL•FO
RenderX

on the definition, with some studies mentioning terms such as self-concept, self-esteem, self-worth, and self-efficacy. Our thematic analysis also found that some women's use of digital health tools increases their self-efficacy in managing their health. Another systematic review of 13 digital health interventions for midlife women found that many interventions did not use a specific behavior change theory [60]. Our scoping review did not examine the effectiveness of the interventions described, but those designing digital health tools and interventions may want to carefully consider behavioral theories in the design to increase adoption and retention rates and adherence to interventions.

Overall, digital health technology to collect PGHD has gained popularity over the past several years. The integration of wearables, smartphones, and digital health technologies has enabled the integration of passive data collection. This wealth of data provides valuable insights into various aspects of health, enabling informed decisions and the adoption of proactive measures to improve well-being. The uptake of this technology will usher in a new era in how we manage our health and well-being. This transformation has changed how we engage with our health and shifted our perception of health and the approach we take toward maintaining it.

Femtech, as a subset of digital health technology, has grown in popularity. This was evidenced by the large increase in the number of articles published between 2015 and 2020 that used digital health tools to track metrics during pregnancy and the postpartum period as well as metrics related to menstrual, sexual, and reproductive health. These technologies empower women and people assigned female at birth to take charge of their health. This is particularly relevant for people with conditions that are not diseases or health concerns per se but are nevertheless part of managing their overall health and well-being. In this way, femtech can provide a greater sense of control over reproductive health and choices, which can be precarious in many settings worldwide. However, in a previous scoping review, researchers reported that many mobile health apps do not follow data privacy, sharing, and security standards [61]. Issues related to the privacy and security of personal health data may be especially important when it comes to tracking reproductive health in settings where sexual and reproductive health rights are not guaranteed. This focus on pregnancy and reproductive health is consistent with the fact that women's health research has largely focused on reproductive health topics [62]. Researchers and digital health developers must address gaps in women's health regarding areas that are not strictly related to reproductive health. Women's health encompasses much more than obstetrics and gynecology; even for health conditions that affect men and women, there may be sex or gender differences in disease presentation, personal experiences, and treatment plans. While using gendered language and design in femtech has the potential to reinforce stereotypes regarding femininity that could cause harm [63], there is a need for apps to provide content relevant to female populations while being gender inclusive and conscious of biases in the language and advice presented.

When analyzing themes related to acceptability, personalization emerged as a key aspect influencing the adoption and sustained use of digital health tools. People respond positively and want to engage with tools that cater to their unique needs and preferences. The ability to customize elements such as the frequency of notifications, specific health measures tracked and displayed, goal-setting options, and the amount of health information provided enhanced user engagement and motivation. However, offering too many personalization options might overwhelm users, making apps or devices cumbersome to use and navigate. Simplicity and ease of use should not be compromised in the pursuit of personalization. Creating personalized experiences that are intuitive and user-friendly while integrating multiple functionalities into a given device is an important consideration. Recognizing that a "one-size-fits-all" approach is inadequate, digital intervention designers need to define their target audience clearly. Apps that cater to specific groups, such as those with certain chronic health conditions, may inherently provide a sense of personalization by addressing their unique requirements. We have also learned the importance of ensuring that the design is inclusive and accessible to everyone within the target audience. Our findings that some tools are not sensitive to certain circumstances such as pregnancy are consistent with those of a systematic review of digital health interventions for postpartum women, in which the authors reported that barriers related to postpartum status could make it more difficult to engage with the interventions [58]. Tools designed with these circumstances in mind may be more engaging for women during pregnancy and the postpartum period, leading to greater adoption and quality of the technologies. Attrition can be high among users of digital health interventions [64,65], but most participants were willing to self-track when motivated by a specific health condition.

An important finding of this review was the growing demand and expectation that PGHD are integrated with clinical care. As digital health continues evolving, patients seek more seamless interactions between digital health data and health care providers. Moreover, services delivered through digital health technologies were not expected to replace the role of health care professionals but rather to be a useful tool to support health care management. Maintaining the human touch during communication for health care delivery was seen as important, with technology complementing clinical care to enhance the overall experience for patients and providers.

One of the critical considerations in clinical integration is the accuracy of PGHD collected from digital health tools. Ensuring the reliability and validity of the data is essential for effective clinical decision-making. Striking a balance between patient empowerment and health care provider oversight is crucial to achieving the best possible outcomes. In general, it is important for health care providers to actively propose digital health during patient visits and encourage its use. While challenges and concerns associated with the use of digital health are noted from health care providers' perspective, such as concerns about medico-legal effects, maintaining professional boundaries, and not adding an abundance of work, the benefits of these tools in supporting patient care and improving outcomes are perceived as important.

XSL•FO
RenderX

Strengths, Limitations, and Future Directions

There are some limitations to this scoping review. Our inclusion criteria did not cover conference abstracts, conference reviews, editorials, letters, comments, or gray literature. Our review also did not include articles written in languages other than English. Therefore, there may be other uses of PGHD in women's health that were not captured in this review. The assessments of the quality of included articles, the effectiveness of the interventions, or the accuracy in validating PGHD were outside this review's scope and were not performed. Our aim was to provide a broad overview of PGHD in published women's health research literature rather than evaluating the quality of the digital technologies or intervention effectiveness. Another limitation is the rapid growth of digital health and femtech, especially during the COVID-19 pandemic. It is important to note that this scoping review only captures the use of PGHD in women's health before the emergence of the pandemic. We suggest that this review may provide a baseline for comparison in a future scoping review that captures articles published in March 2020 or later. The strengths of this review include the large number of publications analyzed and the data charting process conducted in duplicate by 2 reviewers. The broad scope of this review also helps provide an overall picture of digital health for women and highlights gaps in the research literature.

Future endeavors in this space should consider digital health tools for women for nonreproductive topics such as chronic health conditions that primarily affect women or conditions that have sex or gender differences in presentation and treatment. Within reproductive health, there was a large focus on pregnancy, but there is an unmet need for research and digital health tools appropriate for women in perimenopause and menopause. A previous literature review found <5 articles published between 2010 and 2020 about digital health technologies that meet the psychosocial needs of women experiencing menopause [66]. There may also be further opportunities for digital health tools geared toward specific racial or ethnic groups that are culturally sensitive and available in multiple languages. A systematic review found that barriers to the use of digital health among culturally and linguistically diverse populations include lower literacy levels and the use of complex medical terminology in some apps, lack of recognition of cultural concerns, stereotypes, and inaccurate portrayals of cultural groups [67]. Previous scoping reviews in the space of women's digital health have identified the need for femtech to pay more attention to cultural appropriateness and consider cultural contexts in their design [68,69].

Conclusions

In conclusion, the integration of wearables, smartphones, and other forms of digital health has revolutionized how we approach and engage with our health. Personalization, inclusivity, and integration with clinical care are vital aspects of developing effective digital health solutions. By understanding the needs of the target audience, providing meaningful personalization, and ensuring data accuracy, digital health can truly transform health care and empower individuals to take charge of their well-being while maintaining a collaborative relationship with health care professionals.

Acknowledgments

Thank you to Shannon Cheng, reference librarian, for her work in developing the search strategy and conducting the database searches. A big thank you to Dr Beth Payne for reviewing the manuscript draft and providing helpful feedback. AT is funded by a Michael Smith Health Research British Columbia Scholar award.

Data Availability

The data sets generated during this study are available in the OSF repository [70].

Authors' Contributions

JLK and AT conceived the study and designed the study protocol. JLK, RST, and AT conducted the article screening. JLK, RW, and RST conducted the data charting. JLK, RST, and DSC created the visualizations and tables of the charted data. The thematic analysis coding was conducted by JLK and RW, with additional discussions with AT in refining the themes. JLK, RST, and AT wrote the draft of the manuscript. All authors reviewed the manuscript before submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 106 KB - jmir_v26i1e53327_app1.pdf]

Multimedia Appendix 2
Full search strategy.
[PDF File (Adobe PDF File), 198 KB - jmir_v26i1e53327_app2.pdf]

XSL•FO
RenderX

Multimedia Appendix 3
List of included articles by health area.
[PDF File (Adobe PDF File), 390 KB - jmir_v26i1e53327_app3.pdf]

Multimedia Appendix 4
Metrics collected in the included studies.
[PDF File (Adobe PDF File), 176 KB - jmir_v26i1e53327_app4.pdf]

References

1. Conceptualizing a data infrastructure for the capture, use, and sharing of patient-generated health data in care delivery and research through 2024. The Office of the National Coordinator for Health Information Technology. 2018 Jan. URL: https://www.healthit.gov/sites/default/files/onc_pghd_final_white_paper.pdf [accessed 2024-05-03]
2. DeSilva J, Prensky-Pomeranz R, Zweig M. Digital health consumer adoption report 2020. RockHealth. 2021 Feb 26. URL: https://rockhealth.com/insights/digital-health-consumer-adoption-report-2020 [accessed 2024-05-03]
3. Taylor P. Penetration of mobile devices in Canada as share of the population from 2009 to 2021. Statista. URL: https://www.statista.com/statistics/462386/mobile-device-penetration-canada [accessed 2024-05-03]
4. Olsen E. Digital health apps balloon to more than 350,000 available on the market, according to IQVIA report. MobiHealthNews. 2021 Aug 4. URL: https://www.mobihealthnews.com/news/digital-health-apps-balloon-more-350000-available-market-according-iqvia-report [accessed 2024-05-03]
5. Paré G, Leaver C, Bourget C. Diffusion of the digital health self-tracking movement in Canada: results of a national survey. J Med Internet Res 2018 May 02;20(5):e177. [doi: 10.2196/jmir.9388] [Medline: 29720359]
6. Fox S, Duggan M. Tracking for health. Pew Research Center. 2013 Jan 28. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/media/Files/Reports/2013/PIP_TrackingforHealth-with-appendix.pdf [accessed 2024-05-03]
7. Peek N, Sujan M, Scott P. Digital health and care in pandemic times: impact of COVID-19. BMJ Health Care Inform 2020 Jun;27(1):e100166 [FREE Full text] [doi: 10.1136/bmjhci-2020-100166] [Medline: 32565418]
8. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
9. Bruining N. The post-pandemic legacy: the breakthrough of digital health and telemedicine. Cardiovasc Res 2021 Jul 27;117(9):e118-e119 [FREE Full text] [doi: 10.1093/cvr/cvab178] [Medline: 34143873]
10. Meskó B. COVID-19's impact on digital health adoption: the growing gap between a technological and a cultural transformation. JMIR Hum Factors 2022 Sep 19;9(3):e38926 [FREE Full text] [doi: 10.2196/38926] [Medline: 36121692]
11. Kim KK, Jalil S, Ngo V. Improving self-management and care coordination with person-generated health data and mobile health. In: Edmunds M, Hass C, Holve E, editors. Consumer Informatics and Digital Health. Cham, Switzerland: Springer; Jan 18, 2019.
12. Shapiro M, Johnston D, Wald J, Mon D. Patient-generated health data. White paper. RTI International. 2012. URL: https://www.rti.org/publication/patient-generated-health-data-white-paper [accessed 2024-05-03]
13. Vodicka E, Kim K, Devine EB, Gnanasakthy A, Scoggins JF, Patrick DL. Inclusion of patient-reported outcome measures in registered clinical trials: evidence from ClinicalTrials.gov (2007-2013). Contemp Clin Trials 2015 Jul;43:1-9. [doi: 10.1016/j.cct.2015.04.004] [Medline: 25896116]
14. Femtech—time for a digital revolution in the women's health market. Frost & Sullivan. 2018 Jan 31. URL: https://www.frost.com/frost-perspectives/femtechtime-digital-revolution-womens-health-market [accessed 2024-05-03]
15. Karim JL, Talhouk A. Person-generated health data in women's health: protocol for a scoping review. JMIR Res Protoc 2021 May 28;10(5):e26110 [FREE Full text] [doi: 10.2196/26110] [Medline: 34047708]
16. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
17. Ahmadvand A, Kavanagh D, Clark M, Drennan J, Nissen L. Trends and visibility of "digital health" as a keyword in articles by JMIR publications in the new millennium: bibliographic-bibliometric analysis. J Med Internet Res 2019 Dec 19;21(12):e10477 [FREE Full text] [doi: 10.2196/10477] [Medline: 31855190]
18. Lau Y, Cheng LJ, Chi C, Tsai C, Ong KW, Ho-Lim SS, et al. Development of a healthy lifestyle mobile app for overweight pregnant women: qualitative study. JMIR Mhealth Uhealth 2018 Apr 23;6(4):e91 [FREE Full text] [doi: 10.2196/mhealth.9718] [Medline: 29685868]
19. Maxwell H, O'Shea M, Stronach M, Pearce S. Empowerment through digital health trackers: an exploration of Indigenous Australian women and physical activity in leisure settings. Ann Leis Res 2019 Oct 04;24(1):150-167. [doi: 10.1080/11745398.2019.1674677]
20. Hughson JP, Daly JO, Woodward-Kron R, Hajek J, Story D. The rise of pregnancy apps and the implications for culturally and linguistically diverse women: narrative review. JMIR Mhealth Uhealth 2018 Nov 16;6(11):e189 [FREE Full text] [doi: 10.2196/mhealth.9119] [Medline: 30446483]

XSL•FO
RenderX

21. Ranzenhofer LM, Engel SG, Crosby RD, Haigney M, Anderson M, McCaffery JM, et al. Real-time assessment of heart rate variability and loss of control eating in adolescent girls: a pilot study. Int J Eat Disord 2016 Feb;49(2):197-201 [FREE Full text] [doi: 10.1002/eat.22464] [Medline: 26401652]
22. James DC, Harville C2, Whitehead N, Stellefson M, Dodani S, Sears C. Willingness of African American women to participate in e-Health/m-Health research. Telemed J E Health 2016 Mar;22(3):191-197 [FREE Full text] [doi: 10.1089/tmj.2015.0071] [Medline: 26313323]
23. Lupton D, Maslen S. The more-than-human sensorium: sensory engagements with digital self-tracking technologies. Senses Soc 2018 Jul 25;13(2):190-202. [doi: 10.1080/17458927.2018.1480177]
24. Asklund I, Samuelsson E, Hamberg K, Umefjord G, Sjöström M. User experience of an app-based treatment for stress urinary incontinence: qualitative interview study. J Med Internet Res 2019 Mar 14;21(3):e11296 [FREE Full text] [doi: 10.2196/11296] [Medline: 30869644]
25. Huberty J, Ehlers DK, Kurka J, Ainsworth B, Buman M. Feasibility of three wearable sensors for 24 hour monitoring in middle-aged women. BMC Womens Health 2015 Jul 30;15:55 [FREE Full text] [doi: 10.1186/s12905-015-0212-3] [Medline: 26223521]
26. Haney AC. Young female college millennials' intent for behavior change with wearable fitness technology. Walden University. 2018. URL: https://scholarworks.waldenu.edu/dissertations/5197/ [accessed 2020-04-11]
27. Leonard NR, Silverman M, Sherpa DP, Naegle MA, Kim H, Coffman DL, et al. Mobile health technology using a wearable sensorband for female college students with problem drinking: an acceptability and feasibility study. JMIR Mhealth Uhealth 2017 Jul 07;5(7):e90 [FREE Full text] [doi: 10.2196/mhealth.7399] [Medline: 28687533]
28. Brett J, Boulton M, Watson E. Development of an e-health app to support women prescribed adjuvant endocrine therapy after treatment for breast cancer. Patient Prefer Adherence 2018 Dec 11;12:2639-2647 [FREE Full text] [doi: 10.2147/PPA.S187692] [Medline: 30587936]
29. Chaudhry BM. Expecting great expectations when expecting. Mhealth 2018 Jan 10;4:2 [FREE Full text] [doi: 10.21037/mhealth.2017.12.01] [Medline: 29445731]
30. Mann D, Riddell L, Lim K, Byrne LK, Nowson C, Rigo M, et al. Mobile phone app aimed at improving iron intake and bioavailability in premenopausal women: a qualitative evaluation. JMIR Mhealth Uhealth 2015 Sep 28;3(3):e92 [FREE Full text] [doi: 10.2196/mhealth.4300] [Medline: 26416479]
31. Grym K, Niela-Vilén H, Ekholm E, Hamari L, Azimi I, Rahmani A, et al. Feasibility of smart wristbands for continuous monitoring during pregnancy and one month after birth. BMC Pregnancy Childbirth 2019 Jan 17;19(1):34 [FREE Full text] [doi: 10.1186/s12884-019-2187-9] [Medline: 30654747]
32. Goetz M, Müller M, Matthies LM, Hansen J, Doster A, Szabo A, et al. Perceptions of patient engagement applications during pregnancy: a qualitative assessment of the patient's perspective. JMIR Mhealth Uhealth 2017 May 26;5(5):e73 [FREE Full text] [doi: 10.2196/mhealth.7040] [Medline: 28550005]
33. Lupton D, Pedersen S. An Australian survey of women's use of pregnancy and parenting apps. Women Birth 2016 Aug;29(4):368-375. [doi: 10.1016/j.wombi.2016.01.008] [Medline: 26874938]
34. de Mooij MJ, Hodny RL, O'Neil DA, Gardner MR, Beaver M, Brown AT, et al. OB nest: reimagining low-risk prenatal care. Mayo Clin Proc 2018 Apr;93(4):458-466. [doi: 10.1016/j.mayocp.2018.01.022] [Medline: 29545005]
35. Kokts-Porietis RL, Stone CR, Friedenreich CM, Froese A, McDonough M, McNeil J. Breast cancer survivors' perspectives on a home-based physical activity intervention utilizing wearable technology. Support Care Cancer 2019 Aug;27(8):2885-2892. [doi: 10.1007/s00520-018-4581-7] [Medline: 30554276]
36. Gell NM, Tursi A, Grover KW, Dittus K. Female cancer survivor perspectives on remote intervention components to support physical activity maintenance. Support Care Cancer 2020 May;28(5):2185-2194. [doi: 10.1007/s00520-019-05038-y] [Medline: 31422476]
37. Ravn Jakobsen P, Hermann AP, Søndergaard J, Wiil UK, Clemensen J. Development of an mHealth application for women newly diagnosed with osteoporosis without preceding fractures: a participatory design approach. Int J Environ Res Public Health 2018 Feb 13;15(2):330 [FREE Full text] [doi: 10.3390/ijerph15020330] [Medline: 29438343]
38. Harder H, Holroyd P, Burkinshaw L, Watten P, Zammit C, Harris PR, et al. A user-centred approach to developing bWell, a mobile app for arm and shoulder exercises after breast cancer treatment. J Cancer Surviv 2017 Dec;11(6):732-742 [FREE Full text] [doi: 10.1007/s11764-017-0630-3] [Medline: 28741202]
39. Depper A, Howe PD. Are we fit yet? English adolescent girls' experiences of health and fitness apps. Health Sociol Rev 2016 Jul 04;26(1):98-112. [doi: 10.1080/14461242.2016.1196599]
40. Gambier-Ross K, McLernon DJ, Morgan HM. A mixed methods exploratory study of women's relationships with and uses of fertility tracking apps. Digit Health 2018 Jul 25;4:2055207618785077 [FREE Full text] [doi: 10.1177/2055207618785077] [Medline: 30225095]
41. Lupton D. 'It just gives me a bit of peace of mind': Australian women's use of digital media for pregnancy and early motherhood. Societies 2017 Sep 15;7(3):25. [doi: 10.3390/soc7030025]
42. Logsdon MC, Lauf A, Stikes R, Revels A, Vickers-Smith R. Partnering with new mothers to develop a smart phone app to prevent maternal mortality after hospital discharge: a pilot study. J Adv Nurs 2020 Jan;76(1):324-327. [doi: 10.1111/jan.14219] [Medline: 31588581]

XSL•FO
RenderX

43. Wright AA, Raman N, Staples P, Schonholz S, Cronin A, Carlson K, et al. The HOPE pilot study: harnessing patient-reported outcomes and biometric data to enhance cancer care. JCO Clin Cancer Inform 2018 Dec;2:1-12 [FREE Full text] [doi: 10.1200/CCI.17.00149] [Medline: 30652585]
44. de Mendoza AH, Cabling ML, Dilawari A, Turner JW, Fernández N, Henderson A, et al. Providers' perspectives on adherence to hormonal therapy in breast cancer survivors. Is there a role for the digital health feedback system? Health Technol (Berl) 2019 Mar;9(2):175-184 [FREE Full text] [doi: 10.1007/s12553-018-0267-x] [Medline: 31448187]
45. Willcox JC, van der Pligt P, Ball K, Wilkinson SA, Lappas M, McCarthy EA, et al. Views of women and health professionals on mHealth lifestyle interventions in pregnancy: a qualitative investigation. JMIR Mhealth Uhealth 2015 Oct 28;3(4):e99 [FREE Full text] [doi: 10.2196/mhealth.4869] [Medline: 26510886]
46. Ke JX, George RB, Wozney L, Chorney JL. Patient-centred perioperative mobile application in Cesarean delivery: needs assessment and development. Can J Anaesth 2019 Oct;66(10):1194-1201. [doi: 10.1007/s12630-019-01392-x] [Medline: 31087270]
47. Firet L, de Bree C, Verhoeks CM, Teunissen DA, Lagro-Janssen AL. Mixed feelings: general practitioners' attitudes towards eHealth for stress urinary incontinence - a qualitative study. BMC Fam Pract 2019 Jan 26;20(1):21 [FREE Full text] [doi: 10.1186/s12875-019-0907-x] [Medline: 30684962]
48. Garnweidner-Holme LM, Borgen I, Garitano I, Noll J, Lukasse M. Designing and developing a mobile smartphone application for women with gestational diabetes mellitus followed-up at diabetes outpatient clinics in Norway. Healthcare (Basel) 2015 May 21;3(2):310-323 [FREE Full text] [doi: 10.3390/healthcare3020310] [Medline: 27417764]
49. Grassl N, Nees J, Schramm K, Spratte J, Sohn C, Schott TC, et al. A web-based survey assessing the attitudes of health care professionals in Germany toward the use of telemedicine in pregnancy monitoring: cross-sectional study. JMIR Mhealth Uhealth 2018 Aug 08;6(8):e10063 [FREE Full text] [doi: 10.2196/10063] [Medline: 30089606]
50. Pais S, Parry D, Petrova K, Rowan J. Acceptance of using an ecosystem of mobile apps for use in diabetes clinic for self-management of gestational diabetes mellitus. Stud Health Technol Inform 2017;245:188-192. [Medline: 29295079]
51. Ragavan MI, Ferre V, Bair-Merritt M. Thrive: a novel health education mobile application for mothers who have experienced intimate partner violence. Health Promot Pract 2020 Mar;21(2):160-164. [doi: 10.1177/1524839919890870] [Medline: 31874566]
52. Runkle J, Sugg M, Boase D, Galvin SL, C Coulson C. Use of wearable sensors for pregnancy health and environmental monitoring: descriptive findings from the perspective of patients and providers. Digit Health 2019 Feb 06;5:2055207619828220 [FREE Full text] [doi: 10.1177/2055207619828220] [Medline: 30792878]
53. Sadigursky A. Move my mood: development and evaluation of a mobile mental health self-help app using behavioral activation for women with postpartum depression. Alliant International University. 2018. URL: https://books.google.co.in/books/about/Move_My_Mood.html?id=dYtP0AEACAAJ&redir_esc=y [accessed 2020-03-30]
54. Scherr CL, Feuston JL, Nixon DM, Cohen SA. A two-phase approach to developing SNAP: an iPhone application to support appointment scheduling and management for women with a BRCA mutation. J Genet Couns 2018 Apr;27(2):439-445. [doi: 10.1007/s10897-018-0222-z] [Medline: 29423568]
55. Tommasone G, Bazzani M, Solinas V, Serafini P. Midwifery e-health: from design to validation of “mammastyle — Gravidanza Fisiologica”. In: Proceedings of the IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom). 2016 Presented at: Healthcom 2016; September 14-16, 2016; Munich, Germany. [doi: 10.1109/healthcom.2016.7749499]
56. Fu MR, Axelrod D, Guth AA, Wang Y, Scagliola J, Hiotis K, et al. Usability and feasibility of health IT interventions to enhance Self-Care for Lymphedema Symptom Management in breast cancer survivors. Internet Interv 2016 Sep;5:56-64 [FREE Full text] [doi: 10.1016/j.invent.2016.08.001] [Medline: 28255542]
57. Dienelt K, Moores CJ, Miller J, Mehta K. An investigation into the use of infant feeding tracker apps by breastfeeding mothers. Health Informatics J 2020 Sep;26(3):1672-1683 [FREE Full text] [doi: 10.1177/1460458219888402] [Medline: 31793799]
58. Lim S, Tan A, Madden S, Hill B. Health professionals' and postpartum women's perspectives on digital health interventions for lifestyle management in the postpartum period: a systematic review of qualitative studies. Front Endocrinol (Lausanne) 2019 Nov 8;10:767 [FREE Full text] [doi: 10.3389/fendo.2019.00767] [Medline: 31787932]
59. Mackey A, Petrucka P. Technology as the key to women's empowerment: a scoping review. BMC Womens Health 2021 Feb 23;21(1):78 [FREE Full text] [doi: 10.1186/s12905-021-01225-4] [Medline: 33622306]
60. Sediva H, Cartwright T, Robertson C, Deb SK. Behavior change techniques in digital health interventions for midlife women: systematic review. JMIR Mhealth Uhealth 2022 Nov 09;10(11):e37234 [FREE Full text] [doi: 10.2196/37234] [Medline: 36350694]
61. Alfawzan N, Christen M, Spitale G, Biller-Andorno N. Privacy, data sharing, and data security policies of women's mHealth apps: scoping review and content analysis. JMIR Mhealth Uhealth 2022 May 06;10(5):e33735 [FREE Full text] [doi: 10.2196/33735] [Medline: 35522465]
62. Hallam L, Vassallo A, Pinho-Gomes AC, Carcel C, Woodward M. Does journal content in the field of women's health represent women's burden of disease? A review of publications in 2010 and 2020. J Womens Health (Larchmt) 2022 May;31(5):611-619. [doi: 10.1089/jwh.2021.0425] [Medline: 35333604]

XSL•FO
RenderX

63. Figueroa CA, Luo T, Aguilera A, Lyles CR. The need for feminist intersectionality in digital health. Lancet Digit Health 2021 Aug;3(8):e526-e533. [doi: 10.1016/s2589-7500(21)00118-7]
64. Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L. Attrition in conversational agent-delivered mental health interventions: systematic review and meta-analysis. J Med Internet Res 2024 Feb 27;26:e48168 [FREE Full text] [doi: 10.2196/48168] [Medline: 38412023]
65. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
66. Cronin C, Hungerford C, Wilson RL. Using digital health technologies to manage the psychosocial symptoms of menopause in the workplace: a narrative literature review. Issues Ment Health Nurs 2021 Jun;42(6):541-548 [FREE Full text] [doi: 10.1080/01612840.2020.1827101] [Medline: 33027594]
67. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719 [FREE Full text] [doi: 10.2196/42719] [Medline: 36853742]
68. Birati Y, Yefet E, Perlitz Y, Shehadeh N, Spitzer S. Cultural and digital health literacy appropriateness of app- and web-based systems designed for pregnant women with gestational diabetes mellitus: scoping review. J Med Internet Res 2022 Oct 14;24(10):e37844 [FREE Full text] [doi: 10.2196/37844] [Medline: 36240008]
69. Woodley SJ, Moller B, Clark AR, Bussey MD, Sangelaji B, Perry M, et al. Digital technologies for women's pelvic floor muscle training to manage urinary incontinence across their life course: scoping review. JMIR Mhealth Uhealth 2023 Jul 05;11:e44929 [FREE Full text] [doi: 10.2196/44929] [Medline: 37405818]
70. Person-generated health data in women's health: scoping review. OSF Home. 2024 Mar 7. URL: https://osf.io/3eync/ [accessed 2024-05-07]

Abbreviations

PGHD: person-generated health data
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by T de Azevedo Cardoso; submitted 04.10.23; peer-reviewed by D Liu, M Herron; comments to author 23.02.24; revised version received 15.03.24; accepted 26.03.24; published 16.05.24.

Please cite as:
Karim JL, Wan R, Tabet RS, Chiu DS, Talhouk A
Person-Generated Health Data in Women's Health: Scoping Review
J Med Internet Res 2024;26:e53327
URL: https://www.jmir.org/2024/1/e53327
doi:10.2196/53327
PMID:38754098

©Jalisa Lynn Karim, Rachel Wan, Rhea S Tabet, Derek S Chiu, Aline Talhouk. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Reporting of Ethical Considerations in Qualitative Research Utilizing Social Media Data on Public Health Care: Scoping Review

Yujie Zhang[1,2*], MD; Jiaqi Fu[1,2*], MD; Jie Lai[1,2*], MD; Shisi Deng[1,2], MD; Zihan Guo[1,2], MD; Chuhan Zhong[1,2], MD; Jianyao Tang[1,2], MD; Wenqiong Cao[1], BSc; Yanni Wu[1], PhD

[1]Nanfang Hospital, Southern Medical University, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yanni Wu, PhD
Nanfang Hospital
Southern Medical University
No 1838 Guangzhou Avenue North
Baiyun District, Guangdong Province
Guangzhou, 510515
China
Phone: 86 02061641192
Email: yanniwuSMU@126.com

Abstract

Background: The internet community has become a significant source for researchers to conduct qualitative studies analyzing users' views, attitudes, and experiences about public health. However, few studies have assessed the ethical issues in qualitative research using social media data.

Objective: This study aims to review the reportage of ethical considerations in qualitative research utilizing social media data on public health care.

Methods: We performed a scoping review of studies mining text from internet communities and published in peer-reviewed journals from 2010 to May 31, 2023. These studies, limited to the English language, were retrieved to evaluate the rates of reporting ethical approval, informed consent, and privacy issues. We searched 5 databases, that is, PubMed, Web of Science, CINAHL, Cochrane, and Embase. Gray literature was supplemented from Google Scholar and OpenGrey websites. Studies using qualitative methods mining text from the internet community focusing on health care topics were deemed eligible. Data extraction was performed using a standardized data extraction spreadsheet. Findings were reported using PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines.

Results: After 4674 titles, abstracts, and full texts were screened, 108 studies on mining text from the internet community were included. Nearly half of the studies were published in the United States, with more studies from 2019 to 2022. Only 59.3% (64/108) of the studies sought ethical approval, 45.3% (49/108) mentioned informed consent, and only 12.9% (14/108) of the studies explicitly obtained informed consent. Approximately 86% (12/14) of the studies that reported informed consent obtained digital informed consent from participants/administrators, while 14% (2/14) did not describe the method used to obtain informed consent. Notably, 70.3% (76/108) of the studies contained users' written content or posts: 68% (52/76) contained verbatim quotes, while 32% (24/76) paraphrased the quotes to prevent traceability. However, 16% (4/24) of the studies that paraphrased the quotes did not report the paraphrasing methods. Moreover, 18.5% (20/108) of the studies used aggregated data analysis to protect users' privacy. Furthermore, the rates of reporting ethical approval were different between different countries (P=.02) and between papers that contained users' written content (both direct and paraphrased quotes) and papers that did not contain users' written content (P<.001).

Conclusions: Our scoping review demonstrates that the reporting of ethical considerations is widely neglected in qualitative research studies using social media data; such studies should be more cautious in citing user quotes to maintain user privacy. Further, our review reveals the need for detailed information on the precautions of obtaining informed consent and paraphrasing to reduce the potential bias. A national consensus of ethical considerations such as ethical approval, informed consent, and privacy issues is needed for qualitative research of health care using social media data of internet communities.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51496) doi:10.2196/51496

KEYWORDS

qualitative research; informed consent; ethics approval; privacy; internet community

Introduction

Social media are web-based computer-mediated tools to collaborate, share, or exchange information, ideas, pictures, or videos in virtual communities and networks such as message boards, communities, chat rooms, forums, Twitter, and Facebook [1]. Moreover, patients and researchers can use internet communities to provide health care and disseminate health information [2,3]. Health care refers to the efforts made to improve or maintain physical, mental, or emotional well-being, including prevention, diagnosis, treatment, recovery, and other physical and mental impairments [4]. Currently, with 57% of the global population's access to social media, more than 40% of the patients and caregivers worldwide utilize the internet community for health care information needs [5]. With diverse populations accessing internet communities and sharing information about health care topics, researchers have the opportunity to collect and analyze text about health care from a diverse range of participants in the internet community, which was unavailable previously [6]. Usually, quantitative data are derived from information extraction, which can be analyzed statistically, and the summary results presented cannot be directly linked to individual participants. In contrast, qualitative research within internet community analysis posts and comments qualitatively or thematically involves a more detailed and in-depth analysis and understanding of the full written content [7]. However, a controversial ethical problem has been raised about conducting qualitative research containing internet users' verbatim quotes that could lead to traceability of the original post, thereby causing a threat to an individual's privacy [8]. Additionally, a previous study investigated public and patients' views regarding ethics in research using social media data and reported that internet users were aggrieved if they found any of their quotes cited in a medical research paper without obtaining their informed consent [9]. Further, besides the privacy breach caused by posts being traced, there is greater harm for special groups or vulnerable groups if we do not highlight the importance of the technical standards for text mining and privacy protection in health care. For instance, some unusual postings, abnormal pictures, and interactions that were expressed by individuals with mental disorders in social media can be detected by researchers by using text mining tools without obtaining their consent [10]. The publication of research on mental disorders, including quotes in posts, can result in a high risk of information harm, which can lead to personal information being revealed and further stigmatization of the condition or disease [11]. Since 2001, ethical concerns have been debated for decades about ethical approval, informed consent, and how to ensure anonymity and preserve data privacy and confidentiality in qualitative research in the internet community [12-14].

With the rapid development of social media and internet research, some ethical guidelines or standards have been published to ensure that research based on internet communities is conducted ethically. The Association of Internet Researchers (internet research ethical guidelines 2.0 and 3.0) showed that researchers working without the direct approval of ethics review boards would have additional challenges to face, and obtaining informed consent is obviously impracticable in several big data projects. However, with the ethical issues about privacy breaches and harms of risk of discrimination, the Association of Internet Researchers recommended reserving the acquisition of informed consent to the dissemination stage by asking for informed consent from specific participants before publication of their quotes [15,16]. Furthermore, researchers should take responsibility for information confidentiality and anonymity according to the internet research ethics criteria prepared by the National Committee for Research Ethics in the Social Sciences and the Humanities guidelines, which recommend a basic research ethic norm for the analyses, reports, and evaluations that apply to all research [17]. Moreover, the National Committee for Research Ethics in the Social Sciences and the Humanities guidelines contain more details about the demand for legal consent and privacy standards imposed by the European Union's General Data Protection Regulation. The General Data Protection Regulation is a European Union–wide regulation targeting the project of personal data processing. The General Data Protection Regulation defines personal data as any information relating to an identifiable person (data subject), including name, online identification number, location data, and other factors related to personal, physical, physiological, mental, or social identity [18]. The General Data Protection Regulation recommends using anonymous data and deleting identifiable information to ensure the confidentiality of the data. Consent should be obtained from the individual for use in scientific research [18,19]. The British Psychological Society guideline does not explicitly refer to the internet community but suggests that researchers may consider paraphrasing the verbatim quotes to reduce the risk of being traced or identified in qualitative research [20]. When paraphrasing, steps must be put into place to ensure that the original meaning of the message is maintained. Currently, there is no widespread consensus on ethical considerations by social media researchers.

Some researchers have tried to explore the reporting of existing ethical considerations in research papers using social media data. For instance, Sinnenberg et al [6] reported that only 32% and 12% of the papers mentioned acquiring ethical approval and informed consent, respectively, by utilizing multiple analysis methods, including surveillance, intervention, recruitment, engagement, content analysis, and network analysis with Twitter data before 2015. Thereafter, Takats et al [21] conducted an updated examination based on Sinnenberg et al's [6] study. They found that of 367 studies using different methodological approaches, including sentiment mining, surveillance, and thematic exploration of public health research using Twitter data between 2010 to 2019, 17% of the studies included verbatim tweets and identifiable information about the internet users [21]. Similarly, Lathan et al [22] reviewed papers,

XSL•FO
RenderX

including both qualitative and quantitative methods, by using Facebook data to explore public health issues and reported that only 48% and 10% of the papers obtained ethical approval and informed consent, respectively. Furthermore, in a study on research using YouTube data or comments, Tanner et al [23] found that only 26.1% of these studies sought ethical approval, only 1 paper (0.08%) sought informed consent, and 27.7% contained identifiable information. These findings indicate widespread neglect of ethical issues such as ethical approval, informed consent, and privacy issues in research papers using social media data.

Our study focuses on the ethical challenges of qualitative studies utilizing social media data. First, social media can be considered as sources for qualitative data collection because of the low cost, vast amount of available sources about health information, and users' health behaviors, experiences, and attitudes. Second, qualitative research is context-dependent and mainly contains quotations and written content to support the viewpoint. It is acknowledged that quote materials from social media would potentially be traced back to the original posts and threaten the users' privacy [24]. This is supported by findings reported by Ayers et al [25] who found that online searches of verbatim Twitter quotes in journal papers described as "content analyses" or "coded Twitter postings" can be traced back to individual internet users 84% of the time. Furthermore, Lathan et al [22] identified that 46% of the studies with verbatim or paraphrased quotes could be traced to the original posts in 10 minutes. Therefore, it is essential to investigate the extent to which ethical oversight is reported in qualitative studies using social media data. Moreover, qualitative research often involves personally sensitive data about health conditions and diseases; hence, anonymity and proper deidentification would be more important for researchers [26,27].

Previous studies have reviewed the ethical challenges and methodological use of social media platforms such as Twitter [6,21], Facebook [22], and YouTube [23] for health care research in both qualitative and quantitative studies. Although there is plenty of qualitative data pouring into social media such as blogs, Twitter, Facebook, and Weibo, evidence is lacking on the investigation of ethical considerations targeting qualitative data in different software and web-based discussion forums to provide a more comprehensive understanding of the ethical issues. To address the ethical considerations in qualitative research of different internet communities and draw the attention of researchers and publishers to ethical issues, we conducted this study to evaluate the ethical practices and ethical considerations of qualitative studies on health care by using data of internet communities. This review aims to (1) assess the rates of reporting institutional review board (IRB) approval and informed consent in studies focused on mining text in the internet community and social media, (2) compare these rates according to the year of publication, country conducting the research, website included in the study's analysis, and journal's guidelines about ethical approval for the type of study, and (3) describe whether the studies used anonymized/deidentified data.

Methods

Research Design

We conducted a scoping review to investigate how qualitative research mining social media data handles ethical approval, informed consent, and confidential issues. We performed this study according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. The completed PRISMA-ScR checklist is provided in Multimedia Appendix 1.

Search Strategy

All published qualitative studies from 2010 to March 31, 2023, focusing on mining text from online community and social media sources about health care in the following databases were included in this study: PubMed, Web of Science, CINAHL, Cochrane, and Embase. A standardized search string containing Medical Subject Headings (MeSH) and non-MeSH entry terms was used in the search strategy. In addition, the reference lists of the retrieved papers and citation tracking were manually searched as a supplement to database searches to improve comprehensiveness. Gray literature was also identified through internet searches in Google Scholar and OpenGrey websites. The search strategies are represented in Multimedia Appendix 2.

Inclusion and Exclusion Criteria

We divided the criteria into 2 parts. First, we limited the inclusion and exclusion criteria used at the title and abstract screening stage eligible for (1) studies mining existing text and posts from the internet community and social media data focusing on health care topics, (2) studies using qualitative methods or available qualitative parts in mixed methods studies to analyze data, and (3) studies only written in English. Ineligible studies were those related to investigating the use and dissemination of social media in health care, using social media or internet community as an intervention tool, and using social media to conduct web-based interviews, surveys, or focus groups. We also excluded studies published as reviews, case studies, conference abstracts, commentaries, policies, guidelines, and recommendations. Second, at the full screening stage, the specific eligible inclusion criteria were studies focused on mining text about health care topics with full-text papers. Studies that did not have the full text after contacting the authors and that were not originally in the English language were excluded.

Study Selection

All results of the searches were entered into the EndNote library, and duplicates were removed. Two researchers reviewed the titles and abstracts based on the inclusion and exclusion criteria independently. Those studies that were irrelevant to the study topic were discarded, and then the full text was screened to select eligible papers. Any disagreements were discussed and resolved by consensus or a third person.

Data Extraction

Data were extracted between April 2023 and May 2023. Two researchers independently read the full text carefully, and the results were extracted using a standardized data extraction

XSL•FO
RenderX

spreadsheet, including research type, first author, study objective, sample size, publication time, country where the research was conducted or country of the first author, website or internet community the studies focus on, type of data collected from social media, language of collected posts or data, privacy level of data (public or privacy posts), study design, research results, published journal, and information about the ethical considerations. Disagreements were resolved by consensus of a third person. The information about ethical considerations was analyzed to investigate the rates of reporting ethical approval, informed consent, and privacy issues: whether IRB review was reported (IRB approval, IRB exemption, unnecessary, not mentioned) and the reason for not requiring IRB approval; whether informed consent was obtained from participants or the websites' administrators, consent types (digitally informed consent or written informed consent, informed consent is not required, consent was waived by IRB), and the methods used to obtain consent in each study; and whether quoting a post in papers could lead to the identification of internet users in each study. The description of users' posts (verbatim quote, paraphrase) was recorded. We also analyzed if posts were paraphrased to maintain the original meaning, if actions were taken to deidentify the internet users, and if the posts contained other identifying information (ie, usernames, photos, links, hashtags) attached to the post. As every journal would provide publication ethical considerations and requirements, we also searched the submission guidelines and editorial policies of each journal submission website to check whether the journal contained any ethical guidance targeting studies using data from internet community and social media platforms. Additional information was included about the details of ethical approval, informed consent, and privacy, for example, whether individuals can withdraw their quotes if they want to be excluded from the study at any time without any reprisal and whether the quotations were tested for deidentification via search engines. There was excellent agreement on the primary outcome between the 2 researchers (k>.95 for all).

Data Analysis

Data were analyzed using SPSS software (IBM Corp). The chi-square test or Fisher exact tests (when cell size was less than 5) were used to test for differences between the rates of informed consent and ethical approval according to publication year, website, and different countries. All *P* values were 2-sided, and *P* values <.05 indicated significance.

Results

Study Selection for the Review

We reviewed 4674 papers after removing the duplicates. After screening the titles, abstracts, and full-texts, we reviewed 108 eligible papers (Figure 1). The full list of the included papers and all the extracted information are incorporated in Multimedia Appendix 3 [28-135]. Of the 108 studies reviewed, 73 (67.6%) were qualitative studies and 35 (32.4%) were mixed methods studies. All papers had text mined from internet communities or social media for qualitative analysis. The sample size ranged from 32 to 392,962. Approximately 82.4% (89/108) of the studies were published after 2018, and there was a sharp increase in the number of studies from 2019 to 2022. Moreover, nearly half of the studies (55/108, 50.9%) were published in the United States. Regarding the websites for mining text, the most widely used social media platform was Twitter (42/108, 38.9%), followed by Facebook (17/108, 15.7%).

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flow diagram of the study selection process.

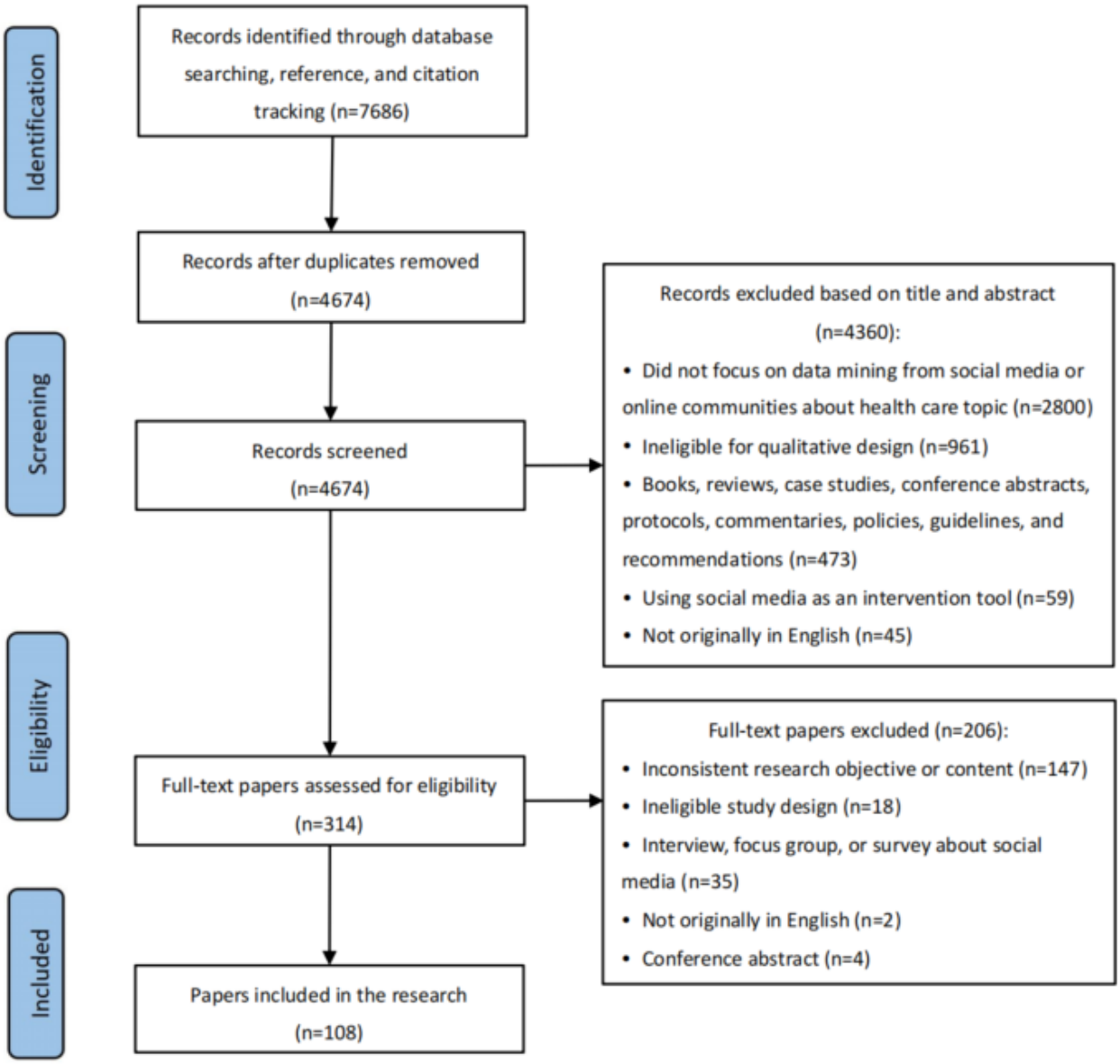

Ethics Approval in These Studies

Our results indicated that of the 108 studies, 78 (72.2%) reported ethics approval. Of the 78 studies, 31 (40%) explicitly stated that ethics approval was obtained before the study was undertaken, 33 (42%) reported that the ethics approval was granted through exemptions by the local IRB, and 14 (18%) explicitly demonstrated that approval by the ethics committee was not required because publicly available data were collected from internet communities and social media platforms. However, 30 (27.8%) of the 108 studies did not mention about obtaining IRB approval (Table 1).

XSL•FO
RenderX

Table 1. Ethical considerations in the qualitative studies using data of the internet community.

Ethical considerations	Values, n (%)
Institutional review board review sought (N=108)	
Yes (including exemption)	64 (59.3)
No	14 (12.9)
Not mentioned	30 (27.8)
Informed consent (N=108)	
Yes	14 (12.9)
No (not required/exemption)	35 (32.4)
Not mentioned	59 (54.7)
Anonymous data (N=108)	
Yes	104 (96.3)
No	4 (3.7)
Studies contain internet users' written content (n=76)	
Verbatim quote	52 (68)
Paraphrased	24 (32)
Identifiable information attached to the post (links, photos, screenshots) (n=76)	
Yes	14 (18)
No	62 (82)

Based on our exploration of the ethical guidelines of each journal to determine whether there were ethical requirements for studies mining social media data, only 36.1% (39/108) of the studies were published in journals that required ethical considerations for studies gathering data from social media platforms by using internet and digital technologies. Of the 39 studies published in 19 journals, 27 (69%) were published in the *Journal of Medical Internet Research* and its sister journals. The submission guidelines of the *Journal of Medical Internet Research* state that authors of manuscripts describing studies of internet, digital tools, and technologies are required to verify that they have adhered to local, national, regional, and international laws and regulations, and are required to verify that they complied with informed consent guidelines. Moreover, 2 journals also provided a specific requirement, that is, when researchers interact with individuals or obtain privacy information gathered from social media platforms, they should obtain ethics approval prior to conducting the study and informed consent from anyone who could potentially be identified. Surprisingly, there were no significant differences in the ethics approval reportage between journals with ethics approval guidelines and those that did not have ethics guidelines for researchers gathering data from social media platforms (P=.08). Notably, the rates of reporting ethics approval were different between different countries (P=.02). However, there were no statistically significant differences between the rates of reporting ethical approval and different websites or publication years (all P>.05) (Table 2).

XSL•FO
RenderX

Table 2. Reporting of ethical considerations in studies published in different publication years, countries, websites, and journals containing ethical requirements for research involving text mining and internet users' written content.

Items (total number of studies)	Ethical approval reported			Informed consent reported		
	Values, n (%)	Chi-square *(df)*	*P* value	Values, n (%)	Chi-square *(df)*	*P* value
Year		17.2 (13)	.11		12.1 (13)	.52
2010 (n=1)	1 (100)			0 (0)		
2011 (n=2)	2 (100)			1 (100)		
2012 (n=2)	1 (50)			1 (50)		
2013 (n=2)	0 (0)			0 (0)		
2014 (n=3)	1 (33)			2 (67)		
2015 (n=2)	1 (50)			1 (50)		
2016 (n=3)	3 (100)			2 (67)		
2017 (n=4)	4 (100)			3 (75)		
2018 (n=12)	7 (58)			3 (25)		
2019 (n=9)	5 (56)			4 (44)		
2020 (n=24)	16 (67)			9 (38)		
2021 (n=14)	11 (78)			5 (36)		
2022 (n=25)	22 (88)			16 (64)		
2023 (n=5)	4 (80)			2 (40)		
Country conducting the research		28.4 (20)	.02		17.8 (20)	.64
United States (n=55)	40 (73)			23 (43)		
Australia (n=12)	10 (83)			6 (50)		
United Kingdom (n=8)	8 (100)			5 (62)		
Canada (n=9)	7 (78)			5 (56)		
China (n=3)	0 (0)			0 (0)		
Netherlands (n=3)	2 (67)			2 (67)		
Turkey (n=2)	2 (100)			1 (50)		
United Arab Emirates (n=2)	1 (50)			1 (50)		
India (n=2)	0 (0)			0 (0)		
Sweden (n=1)	1 (100)			1 (100)		
Norway (n=1)	1 (100)			1 (100)		
Italy (n=1)	1 (100)			1 (100)		
Germany (n=1)	1 (100)			0 (0)		
France (n=1)	1 (100)			0 (0)		
Finland (n=1)	1 (100)			1 (100)		
Bangladesh (n=1)	1 (100)			1 (100)		
Austria (n=1)	1 (100)			1 (100)		
Thailand (n=1)	0 (0)			0 (0)		
Saudi Arabia (n=1)	0 (0)			0 (0)		
Singapore (n=1)	0 (0)			0 (0)		
Israel (n=1)	0 (0)			0 (0)		
Website cited in the research		14.7 (11)	.12		18.7 (11)	.07
Twitter (n=42)	26 (62)			14 (33)		
Facebook (n=17)	12 (70)			10 (59)		

XSL•FO
RenderX

Items (total number of studies)	Ethical approval reported			Informed consent reported		
	Values, n (%)	Chi-square *(df)*	*P* value	Values, n (%)	Chi-square *(df)*	*P* value
≥2 websites (n=14)	11 (79)			6 (43)		
Reddit (n=9)	8 (89)			3 (33)		
Specialist forums (n=7)	7 (100)			5 (57)		
Instagram (n=5)	4 (80)			4 (80)		
Blog (n=4)	4 (100)			4 (100)		
YouTube (n=4)	3 (75)			1 (25)		
Sina Weibo (n=3)	0 (0)			0 (0)		
Quora (n=1)	1 (100)			1 (100)		
STUMPPI (n=1)	1 (100)			1 (100)		
WhatsApp (n=1)	1 (100)			1 (100)		
Whether journals contained ethical requirements for research involving text mining from internet community and social media platforms		3.5 (1)	.08		2.2 (1)	.16
Yes (n=39)	24 (62)			14 (36)		
No (n=69)	54 (78)			35 (51)		
Whether studies had users' written content		12.9 (1)	<.001		2.2 (1)	.15
Yes (n=76)	60 (79)			38 (50)		
No (n=32)	14 (44)			21 (67)		

Informed Consent

Of the 108 studies, 59 (54.7%) showed that they did not include any information about informed consent and 49 (45.3%) mentioned informed consent. Of the 49 studies that mentioned informed consent, 14 (13%) demonstrated that informed consent was waived by local institutional boards, and 21 (19%) reported that informed consent was not required because this information is publicly available in websites or did not involve human participants. We interpreted this as not seeking informed consent. Only 14 (12.9%) of the 108 studies explicitly indicated that informed consent was obtained (Table 1). Among the 14 studies, 2 (14%) only provided a generic statement that informed consent was obtained but did not report the process of how the informed consent was obtained, while 12 (86%) received digital informed consent. Of the 12 studies that reported receiving digital informed consent, 6 reported that they sought permission from the communities' or groups' administrators and by posting a statement of the research objective on the group's wall, while 5 studies contacted the participants privately via email, commenting below the posts and software to gain consent, and 1 study reported that it had sent a digital version of the informed consent book. Furthermore, among the studies that had obtained informed consent, 7 studies included the statement that the individuals' posts would be removed if they wanted to be excluded from the study, and they could withdraw from the study whenever they wanted. In addition, the rates of reporting informed consent showed no statistical significance between publication years, different countries, and different websites (all $P>.05$) (Table 2).

Confidentiality of the Information

All data sources were obtained from anonymous websites or communities, and the majority (104/108, 96.3%) of the data sources did not contain usernames. Notably, only 3.7% (4/108) of the studies contained the participants' usernames or pseudonyms. One study reported that pseudonyms like Sasha had been used instead of the real name. The other 3 studies contained the expression for usernames but did not state whether pseudonyms were used. Except for 9 studies that used nonnative language quotes and 3 studies that were transcribed into text via video, among the 108 included studies, 76 (70.3%) quoted at least one native language post in their reports. Additionally, 20 studies presenting aggregated analysis or composite accounts did not include any quotation or written content. Of the 76 studies containing internet users' written content, 52 (68%) contained just verbatim-quoted participants' posts and 24 (32%) contained paraphrased posts (Table 1). Among the 52 studies containing direct and verbatim quotations, which are likely to be traced to the original posts from users, only 17 (33%) studies took measures to deidentify the users. The 17 studies mentioned that all names or usernames were removed and personal identifying information was removed to maintain privacy, while 42% (22/52) of the studies did not mention any measures that were taken to deidentify the users and maintain confidentiality. Approximately 32% (24/76) of the studies described that they paraphrased posts and removed any explicitly identified personal information to maintain confidentiality to reduce the likelihood of users being identified via search engines. Of the 24 studies, 20 (83%) reported that the quotations were slightly modified or summarized for readability, the symbol information was removed using "…", and key identifiable information was removed to protect privacy while maintaining the meaning of

XSL•FO
RenderX

posts. Four of the 24 (17%) studies did not report the methods and details of paraphrasing. Notably, only 3% (2/76) of the studies containing users' written content showed that researchers intentionally entered each quote into search engines to ensure that every quote did not lead to the original posts. Moreover, of the 76 studies containing written content, 62 (82%) did not contain other types of identity information attached to the posts, while 14 (18%) included other identifying data (hashtags, emojis, geolocation, photos, links, screenshots) attached to the original posts for analysis (Table 1). Of the 14 studies including other identifying information, 4 (29%) contained photos and screenshots associated with the website pages. Of the 52 studies that disclosed verbatim quotes and other identifiable information, 26 (46%) studies reported informed consent consideration, and only 8 (15%) obtained explicitly informed consent. Additionally, of the 77% (40/52) of the studies that mentioned IRB or ethical review, 38% (15/40) received IRB approval, and 63% (25/40) of the studies were granted exemption. The proportion of reporting ethical approval in studies containing users' written content was modestly higher than that in studies not containing users' written content (60/76, 79% vs 14/32, 44%; $P<.001$) (Table 2).

Discussion

Principal Findings and Comparison to Prior Work

In this scoping review, we included 108 studies (Multimedia Appendix 3; [28-135]) that focused on mining text from internet community and social media data for health care research, and we reviewed the ethical consideration reportage and outcome reports in these studies. We found that the rates of reporting IRB approval and informed consent in qualitative research on health care utilizing social media data were 59.3% (64/108) and 12.9% (14/108), respectively. Our findings demonstrate that the key ethical considerations for qualitative research in online communities are insufficiently discussed and described. However, the reporting rates of ethical considerations in the papers in our scoping review were much higher than those reported in systematic reviews including multiple analysis methodologies on only 1 social media platform. For example, ethics approval and informed consent were reported in 48% and 10% of research studies using only Facebook data [22], 32% and 0% from 2006 to 2019 [21], 40% and 0.9% (only 1 paper) from 2015 to 2016 in public health research using only Twitter data [25], and 26.1% and 0.8% (only 1 paper) in researches incorporating only YouTube data [23], respectively. In fact, previous studies were limited to only a few selected websites such as Twitter, Facebook, and YouTube. There is a lack of research that incorporates a variety of different social media data for comparisons. Differences in the reporting of ethical considerations may be attributed to the different methodologies adopted by studies. For example, Lathan et al [22] analyzed the ethical considerations in studies including predictive or model development, while our research focuses on the ethical considerations in qualitative studies.

Importantly, our findings indicate that there is a need to develop a standardized and apparent approach for the reporting of ethical considerations in qualitative research of data from social media and online communities. Our research demonstrates that the rates of reporting ethics approval are different in different countries ($P=.02$). Specifically, a wide variety of national research ethics governing bodies and over 1000 laws, regulations, and standards provide oversight for human subjects research in 130 countries. Obviously, a guideline is needed for best ethical practices for qualitative research involving posts from social media platforms. Surprisingly, there were no significant differences between the rates of reporting ethical approval and those of journals specifying ethical requirements for studies involving text mining ($P=.08$). This inconsistent result of publication guidelines and reports of ethical approval consent is similar to previous findings on the ethical standards in COVID-19 human studies [136]. Although there are journal publication guidelines for studies mining social media data, the reports of ethical approval and consent in the papers published in such journals do not exactly follow the guidelines. Consequently, this finding indicates that more ethical awareness is needed among researchers, editors, and reviewers for qualitative studies on data mining.

Besides the different legal and regulations in different countries, the inconsistency in the ethics approval in published papers may be because social media research is a highly interdisciplinary science, and computer science researchers may be less experienced or may pay less attention to the key ethical issues of protecting human subjects [137]. Medical and health science researchers may have considered some ethical concerns about gathering social media data but they may not be familiar with the relevant guidelines. For example, the Association of Internet Researchers has a detailed ethical guideline targeting social scientists conducting digital research, while it may be less popular and less well-known among medical and health care researchers. At the institute level, Ferretti et al [138] noticed that institutionalized review committees, especially the individual IRB institutes for universities and health care systems lack knowledge about the methodology, text mining technical standards, data security, and ethical harms for studies using big data and social media as sources. Because of this lack of knowledge, institutional ethics committees may have inconsistent ethical criteria and perspectives about web-based projects using social media data [139]. Therefore, some ethics review committees exclude research on internet communities from ethical oversights because their ethics standards are confined only to medical fields. Above all, it is additionally challenging for ethical approval institutions because of the continuous development and dynamic change of studies using social media data. Furthermore, it is necessary for ethics committee members to be trained about the ethical issues in studies mining text from social media. Inviting interdisciplinary researchers to join in the approval process would be an appropriate method to increase the awareness of ethical considerations [140,141].

Interestingly, the reporting rate of obtaining informed consent for mining social media data in qualitative studies was unexpected. The most influential ethical reports such as the Nuremberg Code [142], Declaration of Helsinki [143], and the Belmont Report [144] have demonstrated the principle of informed consent in research involving humans. Our review

XSL•FO
RenderX

shows that only 12.9% (14/108) of the studies explicitly obtained informed consent and 32.4% (35/108) of the studies reported that informed consent was exempted by IRB or was not required, as the information was available publicly in websites or did not involve human participants. Our results are similar to those of Wongkoblap et al [145] who reported that only 16.7% of the studies received informed consent from participants prior to data analysis on data mining of social network data on mental health disorders.

There are multiple reasons for the challenges in obtaining informed consent in an internet setting. First, it is impractical for researchers to gain individual informed consent from a large number of users in an internet community [146]. Second, members of ethics review boards lack consensus about the need for informed consent from an internet community for qualitative research under the current legal definition [147]. Moreover, there has been a debate on the criteria of human subject research in using social media data. The federal regulation recommends that if data in the studies are obtained from public social media websites, where data are identifiable and do not require interaction with individuals, such studies do not constitute human subject research, while studies involving the identification of private information or interaction with the individual can be considered as human subject research [148]. In contrast, some researchers believe that social media and big data research are not ethically exceptional and should be treated in the same manner and with the same rules as those for traditional forms of research [149]. There is ambiguity as to what is appropriate or should be standard practice for obtaining informed consent.

Currently, it is challenging to maintain privacy and protect the traceability of individuals posting content in the internet community. Our findings indicated that 70.3% (76/108) of the studies contained internet users' written content, of which 68% (52/76) included verbatim quotations of users' posts that could lead to identification, and 18% (14/76) of the studies included other identifiable information such as links, screenshots, and emojis linked to original posts, which are similar to the findings of Ayers et al [25] and Lathan et al [22]. Usha Lawrance et al [150] and Wilkinson and Thelwall [151] argued that using direct quotes to support findings would lead to the identification of users and breach users' confidentiality in internet community data. Moreover, quoting social media posts or disclosing usernames violate the International Committee of Medical Journal Editors' ethics standards, which state that identifying information such as written descriptions and photos should not be published unless the information is essential for scientific purposes and the participants give written informed consent for publication [152]. Furthermore, our study demonstrates that the proportion of studies containing users' written content (both direct and paraphrased quotations) is higher than that of studies that do not include any quotation or written content (60/76, 79% vs 14/32, 44%; $P<.001$)——a tentative explanation is that some researchers realized that ethical reportage should be stricter for qualitative papers with quotations from social media posts due to privacy and security issues. This is supported by Boyd and Crawford [153] who stated that rigorous thinking about the process of mining and anonymizing big data is required for ethics boards to ensure that people are protected. Our findings show that 32% (24/76) of the studies intentionally paraphrased the quotes to ensure that users could not locate them, and 20 studies used aggregated data interconnected with anonymity. Moreover, it is recommended by Wilkinson and Thelwall [151], Bond et al [154], and Markham et al [155] that researchers should not directly quote and work with aggregate data sets and separate texts from their original context, which is more acceptable to participants. In addition, the British Psychological Society guidelines recommend that researchers consider paraphrasing any verbatim quotes to reduce the risk of these being traced to the source [20]. Notably, 13 of the 25 papers in this study showed that they did not report the precautions taken for paraphrasing. This may be due to the lack of detailed methodology and consensus about paraphrasing quotes to reduce bias and maintaining the original meaning.

Limitations and Strengths

Our scoping review has several limitations. First, our research was limited to qualitative studies and the qualitative parts in mixed methods studies on text mining from social media, and it is unclear whether ethical considerations are critical in quantitative studies among internet communities. Second, we were restricted to studies published in English language and those with the full text available, and therefore, we could be underestimating the number of relevant papers published in other languages. Third, the rates of reporting ethical approval, informed consent, and privacy of this research relied on self-reported data. Thus, it is possible that although certain studies did not report the process of ethical considerations, such considerations may have been followed during the research. Conversely, some studies may have mentioned about the ethical considerations but may not have conducted them in practice. Hence, there is a bias because of the lack of accurate documentation that must be considered.

Conclusion

Social media text mining can be a useful tool for researchers to understand patient experiences of health conditions and health care. However, as illustrated by the absence of ethical discourse in publications, our analysis indicates significant gaps in the ethical considerations and governance of qualitative research of internet posts. Therefore, a complete and consistent consensus guideline of ethical considerations in qualitative research of internet posts is needed to protect users' data. With the continued advancing development of text-mining techniques, qualitative studies mining text from social media should be more cautious while using user quotations to maintain user privacy and protect the traceability of the internet users posting content. We suggest that authors should report their results by using aggregated findings or deidentified ways like paraphrasing instead of verbatim quotations, which can prevent internet users from being identified through search engines. In addition, authors should provide more detailed information about the precautions taken for obtaining informed consent and paraphrasing to reduce the potential bias. Furthermore, journals and editors should pay more attention to the reporting standards of ethical consideration and privacy issues in qualitative research involving social media data.

XSL•FO
RenderX

Acknowledgments

This project was funded by the National Natural Science Foundation of China (72304131) and the Outstanding Youths Development Scheme of Nanfang Hospital, Southern Medical University (2023J005). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript. We sincerely thank the funders of this study.

Data Availability

All data extracted and analyzed during this study are presented in this paper and in the multimedia appendices.

Authors' Contributions

YW was responsible for the protocol of the research and redrafted the paper critically. YZ and JF performed literature searches. YZ, JL, and WC performed study identification and screening. ZG, SD, CZ, and JT extracted and analyzed the data from the included journals. YZ and JL wrote the first draft of the paper. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 71 KB - jmir_v26i1e51496_app1.docx]

Multimedia Appendix 2
Search strategy for each database.
[DOCX File , 28 KB - jmir_v26i1e51496_app2.docx]

Multimedia Appendix 3
Summary of included literature.
[XLSX File (Microsoft Excel File), 65 KB - jmir_v26i1e51496_app3.xlsx]

References

1. Obar JA, Wildman S. Social media definition and the governance challenge: An introduction to the special issue. Telecommunications Policy 2015 Oct;39(9):745-750. [doi: 10.1016/j.telpol.2015.07.014]
2. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. Business Horizons 2010 Jan;53(1):59-68. [doi: 10.1016/j.bushor.2009.09.003]
3. VanDam C, Kanthawala S, Pratt W, Chai J, Huh J. Detecting clinically related content in online patient posts. J Biomed Inform 2017 Nov;75:96-106 [FREE Full text] [doi: 10.1016/j.jbi.2017.09.015] [Medline: 28986329]
4. Health care. Wikipedia. URL: https://en.wikipedia.org/wiki/Health_care [accessed 2023-06-01]
5. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
6. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: A systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
7. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
8. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Monitoring online discussions about suicide among Twitter users with schizophrenia: exploratory study. JMIR Ment Health 2018 Dec 13;5(4):e11483 [FREE Full text] [doi: 10.2196/11483] [Medline: 30545811]
9. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 06;19(6):e195 [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]
10. Athanasopoulou C, Sakellari E. Facebook and health information: content analysis of groups related to schizophrenia. Stud Health Technol Inform 2015;213:255-258. [Medline: 26153008]
11. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
12. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001 Nov 10;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]

XSL•FO
RenderX

13. Heilferty CM. Ethical considerations in the study of online illness narratives: a qualitative review. J Adv Nurs 2011 May;67(5):945-953. [doi: 10.1111/j.1365-2648.2010.05563.x] [Medline: 21231955]
14. Roberts LD. Ethical issues in conducting qualitative research in online communities. Qualitative Research in Psychology 2015 Apr 20;12(3):314-325. [doi: 10.1080/14780887.2015.1008909]
15. Franzke A, Bechmann A, Zimmer M, Ess C, the Association of Internet Researchers. Internet research: ethical guidelines 3.0. Association of Internet Researchers. 2020. URL: https://aoir.org/reports/ethics3.pdf [accessed 2023-05-30]
16. Markham A, Buchanan E. Ethical decision-making and internet research: recommendations from the AoIR ethics working committee (version 2.0). AoIR. 2012. URL: https://aoir.org/reports/ethics2.pdf [accessed 2023-06-05]
17. The NCFREITSSH(. Ethical guidelines for internet research. 2022. URL: https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/a-guide-to-internet-research-ethics [accessed 2023-06-13]
18. Chapter 3: Rights of the data subject. General Data Protection Regulation (GDPR). European Commission: Official Journal of the European Union; 2018. URL: https://gdpr-info.eu/chapter-3/ [accessed 2023-11-29]
19. Hand DJ. Aspects of data ethics in a changing world: where are we now? Big Data 2018 Sep 01;6(3):176-190 [FREE Full text] [doi: 10.1089/big.2018.0083] [Medline: 30283727]
20. British Psychological Society. Ethics guidelines for internet-mediated research. 2021. URL: https://explore.bps.org.uk/binary/bpsworks/64374754e0c1dd30/5a757d7c2d39a9837f0eedec1b9ba28fa5e9e38f0bffa9950b678ab727803959/rep155_2021.pdf [accessed 2023-05-30]
21. Takats C, Kwan A, Wormer R, Goldman D, Jones HE, Romero D. Ethical and methodological considerations of Twitter data for public health research: systematic review. J Med Internet Res 2022 Nov 29;24(11):e40380 [FREE Full text] [doi: 10.2196/40380] [Medline: 36445739]
22. Lathan HS, Kwan A, Takats C, Tanner JP, Wormer R, Romero D, et al. Ethical considerations and methodological uses of Facebook data in public health research: A systematic review. Soc Sci Med 2023 Apr;322:115807 [FREE Full text] [doi: 10.1016/j.socscimed.2023.115807] [Medline: 36889221]
23. Tanner J, Takats C, Lathan H, Kwan A, Wormer R, Romero D, et al. Approaches to research ethics in health research on YouTube: systematic review. J Med Internet Res 2023 Oct 04;25:e43060 [FREE Full text] [doi: 10.2196/43060] [Medline: 37792443]
24. Stockdale J, Cassell J, Ford E. "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. Wellcome Open Res 2018;3:6 [FREE Full text] [doi: 10.12688/wellcomeopenres.13531.2] [Medline: 30854470]
25. Ayers JW, Caputi TL, Nebeker C, Dredze M. Don't quote me: reverse identification of research participants in social media studies. NPJ Digit Med 2018;1:30 [FREE Full text] [doi: 10.1038/s41746-018-0036-2] [Medline: 31304312]
26. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73 [FREE Full text] [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]
27. Hunter RF, Gough A, O'Kane N, McKeown G, Fitzpatrick A, Walker T, et al. Ethical issues in social media research for public health. Am J Public Health 2018 Mar;108(3):343-348. [doi: 10.2105/ajph.2017.304249]
28. Ahmed OH, Sullivan SJ, Schneiders AG, McCrory P. iSupport: do social networking sites have a role to play in concussion awareness? Disabil Rehabil 2010;32(22):1877-1883. [doi: 10.3109/09638281003734409] [Medline: 20367328]
29. Bender JL, Jimenez-Marroquin M, Jadad AR. Seeking support on Facebook: a content analysis of breast cancer groups. J Med Internet Res 2011 Feb 04;13(1):e16 [FREE Full text] [doi: 10.2196/jmir.1560] [Medline: 21371990]
30. Gajaria A, Yeung E, Goodale T, Charach A. Beliefs about attention-deficit/hyperactivity disorder and response to stereotypes: youth postings in Facebook groups. J Adolesc Health 2011 Jul;49(1):15-20. [doi: 10.1016/j.jadohealth.2010.09.004] [Medline: 21700151]
31. Degroot JM. Maintaining relational continuity with the deceased on Facebook. Omega (Westport) 2012 Nov 01;65(3):195-212. [doi: 10.2190/om.65.3.c]
32. Donelle L, Booth RG. Health tweets: an exploration of health promotion on twitter. Online J Issues Nurs 2012 Sep 30;17(3):4 [FREE Full text] [Medline: 23036060]
33. Rui JR, Chen Y, Damiano A. Health organizations providing and seeking social support: a Twitter-based content analysis. Cyberpsychol Behav Soc Netw 2013 Sep;16(9):669-673. [doi: 10.1089/cyber.2012.0350] [Medline: 23786172]
34. Lyles CR, López A, Pasick R, Sarkar U. "5 mins of uncomfyness is better than dealing with cancer 4 a lifetime": an exploratory qualitative analysis of cervical and breast cancer screening dialogue on Twitter. J Cancer Educ 2013 Mar;28(1):127-133. [doi: 10.1007/s13187-012-0432-2] [Medline: 23132231]
35. Lee JL, DeCamp M, Dredze M, Chisolm MS, Berger ZD. What are health-related users tweeting? A qualitative content analysis of health-related users and their messages on twitter. J Med Internet Res 2014 Oct 15;16(10):e237 [FREE Full text] [doi: 10.2196/jmir.3765] [Medline: 25591063]
36. Ahlwardt K, Heaivilin N, Gibbs J, Page J, Gerbert B, Tsoh JY. Tweeting about pain: comparing self-reported toothache experiences with those of backaches, earaches and headaches. J Am Dent Assoc 2014 Jul;145(7):737-743 [FREE Full text] [doi: 10.14219/jada.2014.30] [Medline: 24982280]

XSL•FO
RenderX

37. Struik LL, Baskerville NB. The role of Facebook in Crush the Crave, a mobile- and social media-based smoking cessation intervention: qualitative framework analysis of posts. J Med Internet Res 2014 Jul 11;16(7):e170 [FREE Full text] [doi: 10.2196/jmir.3189] [Medline: 25016998]
38. Abdel-Razig S, Anglade P, Ibrahim H. Impact of the COVID-19 pandemic on a physician group's WhatsApp chat: qualitative content analysis. JMIR Form Res 2021 Dec 07;5(12):e31791 [FREE Full text] [doi: 10.2196/31791] [Medline: 34784291]
39. Reddy A. Skincare in social media: analyzing prominent themes in online dermatologic discussions. Cureus 2021 May 7:e14890. [doi: 10.7759/cureus.14890] [Medline: 34109079]
40. van der Pijl MSG, Hollander MH, van der Linden T, Verweij R, Holten L, Kingma E, et al. Left powerless: A qualitative social media content analysis of the Dutch #breakthesilence campaign on negative and traumatic experiences of labour and birth. PLoS One 2020;15(5):e0233114 [FREE Full text] [doi: 10.1371/journal.pone.0233114] [Medline: 32396552]
41. Rath L, Vijiaratnam N, Skibina O. Alemtuzumab in multiple sclerosis: lessons from social media in enhancing patient care. Int J MS Care 2017;19(6):323-328. [doi: 10.7224/1537-2073.2017-010] [Medline: 29270090]
42. Bloom R, Beck S, Chou W, Reblin M, Ellington L. In their own words: experiences of caregivers of adults with cancer as expressed on social media. ONF 2019 Sep 1;46(5):617-630. [doi: 10.1188/19.onf.617-630]
43. Grumme V, Gordon S. Social media use by transplant recipients for support and healing. Comput Inform Nurs 2016 Dec;34(12):570-577. [doi: 10.1097/CIN.0000000000000276] [Medline: 27366980]
44. Haug NA, Bielenberg J, Linder SH, Lembke A. Assessment of provider attitudes toward #naloxone on Twitter. Subst Abus 2016;37(1):35-41. [doi: 10.1080/08897077.2015.1129390] [Medline: 26860229]
45. Oren E, Martinez L, Hensley RE, Jain P, Ahmed T, Purnajo I, et al. Twitter communication during an outbreak of hepatitis A in San Diego, 2016–2018. Am J Public Health 2020 Oct;110(S3):S348-S355. [doi: 10.2105/ajph.2020.305900]
46. Parker C, Zomer E, Liew D, Ayton D. Characterising experiences with acute myeloid leukaemia using an Instagram content analysis. PLoS One 2021;16(5):e0250641 [FREE Full text] [doi: 10.1371/journal.pone.0250641] [Medline: 33939746]
47. Kurko T, Linden K, Kolstela M, Pietilä K, Airaksinen M. Is nicotine replacement therapy overvalued in smoking cessation? Analysis of smokers' and quitters' communication in social media. Health Expect 2015 Dec;18(6):2962-2977 [FREE Full text] [doi: 10.1111/hex.12280] [Medline: 25292017]
48. Reuter K, Lee D. Perspectives toward seeking treatment among patients with psoriasis: protocol for a Twitter content analysis. JMIR Res Protoc 2021 Feb 18;10(2):e13731 [FREE Full text] [doi: 10.2196/13731] [Medline: 33599620]
49. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, De Melker H, Paulussen T, et al. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. J Med Internet Res 2015 May 26;17(5):e128 [FREE Full text] [doi: 10.2196/jmir.3863] [Medline: 26013683]
50. O'Hagan ET, Traeger AC, Bunzli S, Leake HB, Schabrun SM, Wand BM, et al. What do people post on social media relative to low back pain? A content analysis of Australian data. Musculoskelet Sci Pract 2021 Aug;54:102402. [doi: 10.1016/j.msksp.2021.102402] [Medline: 34058728]
51. Zhou F, Zhang W, Cai H, Cao Y. Portrayals of 2v, 4v and 9vHPV vaccines on Chinese social media: a content analysis of hot posts on Sina Weibo. Hum Vaccin Immunother 2021 Nov 02;17(11):4433-4441 [FREE Full text] [doi: 10.1080/21645515.2021.1971016] [Medline: 34543155]
52. Golder S, Bach M, O'Connor K, Gross R, Hennessy S, Gonzalez Hernandez G. Public perspectives on anti-diabetic drugs: exploratory analysis of Twitter posts. JMIR Diabetes 2021 Jan 26;6(1):e24681 [FREE Full text] [doi: 10.2196/24681] [Medline: 33496671]
53. Mercier R, Senter K, Webster R, Henderson RA. Instagram users? experiences of miscarriage. Obstet Gynecol 2020 Jan;? 2020 Jan;135(1):166-173. [doi: 10.1097/aog.0000000000003621]
54. Alghamdi A, Abumelha K, Allarakia J, Al-Shehri A. Conversations and misconceptions about chemotherapy in Arabic tweets: content analysis. J Med Internet Res 2020 Jul 29;22(7):e13979 [FREE Full text] [doi: 10.2196/13979] [Medline: 32723724]
55. Golder S, O'Connor K, Hennessy S, Gross R, Gonzalez-Hernandez G. Assessment of beliefs and attitudes about statins posted on Twitter: a qualitative study. JAMA Netw Open 2020 Jun 01;3(6):e208953 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.8953] [Medline: 32584408]
56. Stekelenburg N, Horsham C, O'Hara M, Janda M. Using social media to determine the affective and cognitive components of tweets about sunburn. Dermatology 2020 Feb 27;236(2):75-80 [FREE Full text] [doi: 10.1159/000506102] [Medline: 32107345]
57. Meeking K. Patients' experiences of radiotherapy: Insights from Twitter. Radiography (Lond) 2020 Aug;26(3):e146-e151. [doi: 10.1016/j.radi.2020.01.001] [Medline: 32052751]
58. Shah S, Bradbury-Jones C, Taylor J. Using Facebook to tell stories of premature ageing and sexual and reproductive healthcare across the life course for women with cerebral palsy in the UK and USA. BMJ Open 2020 Feb 17;10(2):e032172 [FREE Full text] [doi: 10.1136/bmjopen-2019-032172] [Medline: 32071173]
59. Pretorius K, Choi E, Kang S, Mackert M. Sudden infant death syndrome on Facebook: qualitative descriptive content analysis to guide prevention efforts. J Med Internet Res 2020 Jul 30;22(7):e18474 [FREE Full text] [doi: 10.2196/18474] [Medline: 32729842]

XSL•FO
RenderX

60. Karmegam D, Mapillairaju B. What people share about the COVID-19 outbreak on Twitter? An exploratory analysis. BMJ Health Care Inform 2020 Nov;27(3):e100133 [FREE Full text] [doi: 10.1136/bmjhci-2020-100133] [Medline: 33214193]
61. Hairston TK, Links AR, Harris V, Tunkel DE, Walsh J, Beach MC, et al. Evaluation of parental perspectives and concerns about pediatric tonsillectomy in social media. JAMA Otolaryngol Head Neck Surg 2019 Jan 01;145(1):45-52 [FREE Full text] [doi: 10.1001/jamaoto.2018.2917] [Medline: 30452510]
62. Årsand E, Bradway M, Gabarron E. What are diabetes patients versus health care personnel discussing on social media? J Diabetes Sci Technol 2019 Mar;13(2):198-205 [FREE Full text] [doi: 10.1177/1932296818821649] [Medline: 30636445]
63. Jiang X, Jiang W, Cai J, Su Q, Zhou Z, He L, et al. Characterizing media content and effects of organ donation on a social media platform: content analysis. J Med Internet Res 2019 Mar 12;21(3):e13058 [FREE Full text] [doi: 10.2196/13058] [Medline: 30860489]
64. Oser TK, Minnehan KA, Wong G, Parascando J, McGinley E, Radico J, et al. Using social media to broaden understanding of the barriers and facilitators to exercise in adults with type 1 diabetes. J Diabetes Sci Technol 2019 May;13(3):457-465 [FREE Full text] [doi: 10.1177/1932296819835787] [Medline: 30862185]
65. Sutton J, Vos SC, Olson MK, Woods C, Cohen E, Gibson CB, et al. Lung cancer messages on Twitter: content analysis and evaluation. J Am Coll Radiol 2018 Jan;15(1 Pt B):210-217. [doi: 10.1016/j.jacr.2017.09.043] [Medline: 29154103]
66. Thomas J, Prabhu AV, Heron DE, Beriwal S. Twitter and brachytherapy: An analysis of "tweets" over six years by patients and health care professionals. Brachytherapy 2018;17(6):1004-1010. [doi: 10.1016/j.brachy.2018.07.015] [Medline: 30131255]
67. Kelly-Hedrick M, Grunberg PH, Brochu F, Zelkowitz P. "It's totally okay to be sad, but never lose hope": content analysis of infertility-related videos on YouTube in relation to viewer preferences. J Med Internet Res 2018 May 23;20(5):e10199 [FREE Full text] [doi: 10.2196/10199] [Medline: 29792296]
68. Gage-Bouchard EA, LaValley S, Mollica M, Beaupin LK. Cancer communication on social media: examining how cancer caregivers use Facebook for cancer-related communication. Cancer Nurs 2017;40(4):332-338. [doi: 10.1097/NCC.0000000000000418] [Medline: 27442210]
69. Rael CT, Pierre D, Frye V, Kessler D, Duffy L, Malos N, et al. Evaluating blood donor experiences and barriers/facilitators to blood donation in the United States using YouTube video content. Transfusion 2021 Sep;61(9):2650-2657 [FREE Full text] [doi: 10.1111/trf.16568] [Medline: 34196004]
70. Fisher S, Jehassi A, Ziv M. Hidradenitis suppurativa on Facebook: thematic and content analyses of patient support group. Arch Dermatol Res 2020 Aug;312(6):421-426. [doi: 10.1007/s00403-019-02027-3] [Medline: 31845011]
71. Abdoli S, Hessler D, Vora A, Smither B, Stuckey H. Descriptions of diabetes burnout from individuals with Type 1 diabetes: an analysis of YouTube videos. Diabet Med 2020 Aug;37(8):1344-1351. [doi: 10.1111/dme.14047] [Medline: 31168875]
72. Myneni S, Lewis B, Singh T, Paiva K, Kim SM, Cebula AV, et al. Diabetes self-management in the age of social media: large-scale analysis of peer interactions using semiautomated methods. JMIR Med Inform 2020 Jun 30;8(6):e18441 [FREE Full text] [doi: 10.2196/18441] [Medline: 32602843]
73. Charlie AM, Gao Y, Heller SL. What do patients want to know? questions and concerns regarding mammography expressed through social media. J Am Coll Radiol 2018 Oct;15(10):1478-1486. [doi: 10.1016/j.jacr.2017.09.020] [Medline: 29221997]
74. Watts G, Christou P, Antonarakis G. Experiences of individuals concerning combined orthodontic and orthognathic surgical treatment: a qualitative Twitter analysis. Med Princ Pract 2018;27(3):227-235 [FREE Full text] [doi: 10.1159/000487904] [Medline: 29642059]
75. Pai RR, Alathur S. Assessing mobile health applications with Twitter analytics. Int J Med Inform 2018 May;113:72-84. [doi: 10.1016/j.ijmedinf.2018.02.016] [Medline: 29602436]
76. Cheng TY, Liu L, Woo BK. Analyzing Twitter as a platform for Alzheimer-related dementia awareness: thematic analyses of tweets. JMIR Aging 2018 Dec 10;1(2):e11542 [FREE Full text] [doi: 10.2196/11542] [Medline: 31518232]
77. Bridges N, Howell G, Schmied V. Exploring breastfeeding support on social media. Int Breastfeed J 2018;13:22 [FREE Full text] [doi: 10.1186/s13006-018-0166-9] [Medline: 29983727]
78. Anderson JG, Hundt E, Dean M, Keim-Malpass J, Lopez RP. "The Church of Online Support": examining the use of blogs among family caregivers of persons with dementia. J Fam Nurs 2017 Feb;23(1):34-54. [doi: 10.1177/1074840716681289] [Medline: 27920340]
79. Kearney MD, Selvan P, Hauer MK, Leader AE, Massey PM. Characterizing HPV vaccine sentiments and content on Instagram. Health Educ Behav 2019 Dec;46(2_suppl):37-48. [doi: 10.1177/1090198119859412] [Medline: 31742459]
80. Davies SH, Langer MD, Klein A, Gonzalez-Hernandez G, Dowshen N. Adolescent perceptions of menstruation on Twitter: opportunities for advocacy and education. J Adolesc Health 2022 Jul;71(1):94-104. [doi: 10.1016/j.jadohealth.2022.01.224] [Medline: 35283044]
81. Van Diepen C, Rosales Valdes D. A content analysis on the perceptions of LGBTQ+ (centred) health care on Twitter. Health Expect 2022 Dec;25(6):3238-3245 [FREE Full text] [doi: 10.1111/hex.13631] [Medline: 36245307]
82. Pleasure ZH, Frohwirth LF, Li N, Polis CB. A content analysis of Reddit users' posts about challenges to contraceptive care-seeking during COVID-19-related restrictions in the United States. J Health Commun 2022 Oct 03;27(10):746-754. [doi: 10.1080/10810730.2022.2157911] [Medline: 36519832]

XSL•FO
RenderX

83. Du Y, Dennis B, Ramirez V, Li C, Wang J, Meireles CL. Experiences and disease self-management in individuals living with chronic kidney disease: qualitative analysis of the National Kidney Foundation's online community. BMC Nephrol 2022 Mar 04;23(1):88 [FREE Full text] [doi: 10.1186/s12882-022-02717-7] [Medline: 35246060]
84. Sadek Attalla S, Ow NL, McNarry M, De Simoni A. Experiences of exercise in patients with asthma: a qualitative analysis of discussions in a UK asthma online community. BJGP Open 2022 Apr 29;6(3):BJGPO.2021.0162. [doi: 10.3399/bjgpo.2021.0162]
85. Bartmess M, Talbot C, O'Dwyer ST, Lopez RP, Rose KM, Anderson JG. Using Twitter to understand perspectives and experiences of dementia and caregiving at the beginning of the COVID-19 pandemic. Dementia (London) 2022 Jul;21(5):1734-1752 [FREE Full text] [doi: 10.1177/14713012221096982] [Medline: 35549466]
86. Brewer G, Centifanti L, Caicedo JC, Huxley G, Peddie C, Stratton K, et al. Experiences of mental distress during COVID-19: thematic analysis of discussion forum posts for anxiety, depression, and obsessive-compulsive disorder. Illn Crises Loss 2022 Oct;30(4):795-811 [FREE Full text] [doi: 10.1177/10541373211023951] [Medline: 36199441]
87. Castillo LIR, Hadjistavropoulos T, Beahm J. Social media discussions about long-term care and the COVID-19 pandemic. J Aging Stud 2022 Dec;63:101076 [FREE Full text] [doi: 10.1016/j.jaging.2022.101076] [Medline: 36462920]
88. Colaceci S, Anderson G, Ricciuto V, Montinaro D, Alazraki G, Mena-Tudela D. Experiences of birth during COVID-19 pandemic in Italy and Spain: A thematic analysis. Int J Environ Res Public Health 2022 Jun 18;19(12):7488 [FREE Full text] [doi: 10.3390/ijerph19127488] [Medline: 35742741]
89. Tripathi SD, Parker PD, Prabhu AV, Thomas K, Rodriguez A. An examination of patients and caregivers on Reddit Navigating Brain Cancer: content analysis of the brain tumor Subreddit. JMIR Cancer 2022 Jun 22;8(2):e35324 [FREE Full text] [doi: 10.2196/35324] [Medline: 35731559]
90. Jina-Pettersen N. Fear, neglect, coercion, and dehumanization: is inpatient psychiatric trauma contributing to a public health crisis? J Patient Exp 2022;9:23743735221079138 [FREE Full text] [doi: 10.1177/23743735221079138] [Medline: 35968056]
91. Rossi NA, Devarajan K, Chokshi SN, Ochoa VJ, Benavidez M, Malaya LT, et al. Social media depictions of cochlear implants: An Instagram and TikTok analysis. Otol Neurotol 2023;44(1):e13-e21. [doi: 10.1097/mao.0000000000003744]
92. Singh GK, Rego J, Chambers S, Fox J. Health professionals' perspectives of the role of palliative care during COVID-19: content analysis of articles and blogs posted on Twitter. Am J Hosp Palliat Care 2022 Apr;39(4):487-493 [FREE Full text] [doi: 10.1177/10499091211024202] [Medline: 34189956]
93. Koly KN, Tasnim Z, Ahmed S, Saba J, Mahmood R, Farin FT, et al. Mental healthcare-seeking behavior of women in Bangladesh: content analysis of a social media platform. BMC Psychiatry 2022 Dec 19;22(1):797 [FREE Full text] [doi: 10.1186/s12888-022-04414-z] [Medline: 36529761]
94. Belcher R, Sim D, Meykler M, Owens-Walton J, Hassan N, Rubin R, et al. A qualitative analysis of female Reddit users' experiences with low libido: how do women perceive their changes in sexual desire? J Sex Med 2023 Feb 27;20(3):287-297. [doi: 10.1093/jsxmed/qdac045] [Medline: 36763942]
95. Culp F, Wu Y, Wu D, Ren Y, Raynor P, Hung P, et al. Understanding alcohol use discourse and stigma patterns in perinatal care on Twitter. Healthcare (Basel) 2022 Nov 26;10(12):2375 [FREE Full text] [doi: 10.3390/healthcare10122375] [Medline: 36553899]
96. Ferrey A, Ashworth G, Cabling M, Rundblad G, Ismail K. A thematic analysis of YouTube comments on a television documentary titled 'Diabulimia: The world's most dangerous eating disorder'. Diabet Med 2023 May;40(5):e15025. [doi: 10.1111/dme.15025] [Medline: 36508358]
97. Watt S, Salway T, Gómez-Ramírez O, Ablona A, Barton L, Chang H, et al. Rumination, risk, and response: a qualitative analysis of sexual health anxiety among online sexual health chat service users. Sex. Health 2022 May 23;19(3):182-191. [doi: 10.1071/sh21198]
98. Naganathan G, Bilgen I, Cleland J, Reel E, Cil T. #COVID19 and #Breastcancer: a qualitative analysis of tweets. Curr Oncol 2022 Nov 08;29(11):8483-8500 [FREE Full text] [doi: 10.3390/curroncol29110669] [Medline: 36354729]
99. Potgieter I, Hoare DJ, Fackrell K. Hyperacusis in children: a thematic analysis of discussions in online forums. Am J Audiol 2022 Mar 03;31(1):166-174. [doi: 10.1044/2021_aja-21-00137]
100. Manning Hutson M, Hosking SM, Mantalvanos S, Berk M, Pasco J, Dunning T. What injured workers with complex claims look for in online communities: netnographic analysis. J Med Internet Res 2022 Apr 07;24(4):e17180 [FREE Full text] [doi: 10.2196/17180] [Medline: 35389358]
101. Lawless MT, Hunter SC, Pinero de Plaza MA, Archibald MM, Kitson AL. "You Are By No Means Alone": a netnographic study of self-care support in an online community for older adults. Qual Health Res 2022 Nov;32(13):1935-1951. [doi: 10.1177/10497323221124979] [Medline: 36062369]
102. Keim-Malpass J, Mitchell EM, Sun E, Kennedy C. Using Twitter to understand public perceptions regarding the #HPV vaccine: opportunities for public health nurses to engage in social marketing. Public Health Nurs 2017 Jul;34(4):316-323. [doi: 10.1111/phn.12318] [Medline: 28261846]
103. Yamada R, Rasmussen KM, Felice JP. "What is 'enough,' and how do I make it?": a qualitative examination of questions mothers ask on social media about pumping and providing an adequate amount of milk for their infants. Breastfeed Med 2019;14(1):17-21 [FREE Full text] [doi: 10.1089/bfm.2018.0154] [Medline: 30431317]

XSL•FO
RenderX

104. Osadchiy V, Mills JN, Eleswarapu SV. Understanding patient anxieties in the social media era: qualitative analysis and natural language processing of an online male infertility community. J Med Internet Res 2020 Mar 10;22(3):e16728 [FREE Full text] [doi: 10.2196/16728] [Medline: 32154785]
105. Thomas TH, Nauth-Shelley K, Thompson MA, Attai DJ, Katz MS, Graham D, et al. The needs of women treated for ovarian cancer: results from a #gyncsm Twitter chat. J Patient Cent Res Rev 2018;5(2):149-157 [FREE Full text] [doi: 10.17294/2330-0698.1592] [Medline: 31413999]
106. Mehta N, Zhu L, Lam K, Stall NM, Savage R, Read SH, et al. Health forums and Twitter for dementia research: opportunities and considerations. J Am Geriatr Soc 2020 Dec;68(12):2881-2889. [doi: 10.1111/jgs.16790] [Medline: 32894780]
107. Gupta R, Ariefdjohan M. Mental illness on Instagram: a mixed method study to characterize public content, sentiments, and trends of antidepressant use. J Ment Health 2021 Aug;30(4):518-525. [doi: 10.1080/09638237.2020.1755021] [Medline: 32325006]
108. He L, He C, Reynolds T, Bai Q, Huang Y, Li C, et al. Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. J Am Med Inform Assoc 2021 Jul 14;28(7):1564-1573 [FREE Full text] [doi: 10.1093/jamia/ocab047] [Medline: 33690794]
109. Loeb S, Mihalcea R, Perez-Rosas V, Xu A, Taylor J, Byrne N, et al. Leveraging social media as a thermometer to gauge patient and caregiver concerns: COVID-19 and prostate cancer. Eur Urol Open Sci 2021 Mar;25:1-4 [FREE Full text] [doi: 10.1016/j.euros.2020.12.008] [Medline: 34337497]
110. Graf I, Gerwing H, Hoefer K, Ehlebracht D, Christ H, Braumann B. Social media and orthodontics: A mixed-methods analysis of orthodontic-related posts on Twitter and Instagram. Am J Orthod Dentofacial Orthop 2020 Aug;158(2):221-228. [doi: 10.1016/j.ajodo.2019.08.012] [Medline: 32576427]
111. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
112. Wang J, Zhou Y, Zhang W, Evans R, Zhu C. Concerns expressed by Chinese social media users during the COVID-19 pandemic: content analysis of Sina Weibo microblogging data. J Med Internet Res 2020 Nov 26;22(11):e22152 [FREE Full text] [doi: 10.2196/22152] [Medline: 33151894]
113. Wahbeh A, Nasralah T, Al-Ramahi M, El-Gayar O. Mining physicians' opinions on social media to obtain insights into COVID-19: mixed methods analysis. JMIR Public Health Surveill 2020 Jun 18;6(2):e19276 [FREE Full text] [doi: 10.2196/19276] [Medline: 32421686]
114. Çınar S, Boztepe H, Prof. The use of social media among parents of infants with cleft lip and/or palate. J Pediatr Nurs 2020;54:e91-e96. [doi: 10.1016/j.pedn.2020.05.007] [Medline: 32461012]
115. Jiang T, Osadchiy V, Mills JN, Eleswarapu SV. Is It all in my head? self-reported psychogenic erectile dysfunction and depression are common among young men seeking advice on social media. Urology 2020 Aug;142:133-140. [doi: 10.1016/j.urology.2020.04.100] [Medline: 32437776]
116. Dzubur E, Khalil C, Almario CV, Noah B, Minhas D, Ishimori M, et al. Patient concerns and perceptions regarding biologic therapies in ankylosing spondylitis: insights from a large-scale survey of social media platforms. Arthritis Care Res (Hoboken) 2019 Feb;71(2):323-330 [FREE Full text] [doi: 10.1002/acr.23600] [Medline: 29781587]
117. Gonzalez G, Vaculik K, Khalil C, Zektser Y, Arnold C, Almario CV, et al. Women's experience with stress urinary incontinence: insights from social media analytics. Journal of Urology 2020 May;203(5):962-968. [doi: 10.1097/ju.0000000000000706]
118. Jha SR, McDonagh J, Prichard R, Newton PJ, Hickman LD, Fung E, et al. #Frailty: A snapshot Twitter report on frailty knowledge translation. Australas J Ageing 2018 Dec;37(4):309-312. [doi: 10.1111/ajag.12540] [Medline: 29732721]
119. Litchman ML, Snider C, Edelman LS, Wawrzynski SE, Gee PM. Diabetes online community user perceptions of successful aging with diabetes: analysis of a #DSMA Tweet chat. JMIR Aging 2018 Jun 22;1(1):e10176 [FREE Full text] [doi: 10.2196/10176] [Medline: 31518231]
120. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel approach to cluster patient-generated data into actionable topics: case study of a web-based breast cancer forum. JMIR Med Inform 2018 Nov 29;6(4):e45 [FREE Full text] [doi: 10.2196/medinform.9162] [Medline: 30497991]
121. Gonzalez G, Vaculik K, Khalil C, Zektser Y, Arnold C, Almario CV, et al. Using large-scale social media analytics to understand patient perspectives about urinary tract infections: thematic analysis. J Med Internet Res 2022 Jan 25;24(1):e26781 [FREE Full text] [doi: 10.2196/26781] [Medline: 35076404]
122. Sormunen T, Westerbotn M, Aanesen A, Fossum B, Karlgren K. Social media in the infertile community-using a text analysis tool to identify the topics of discussion on the multitude of infertility blogs. Womens Health (Lond) 2021;17:17455065211063280 [FREE Full text] [doi: 10.1177/17455065211063280] [Medline: 34856827]
123. Berkovic D, Ackerman IN, Briggs AM, Ayton D. Tweets by people with arthritis during the COVID-19 pandemic: content and sentiment analysis. J Med Internet Res 2020 Dec 03;22(12):e24550 [FREE Full text] [doi: 10.2196/24550] [Medline: 33170802]
124. Della Rosa S, Sen F. Health topics on Facebook Groups: content analysis of posts in multiple sclerosis communities. Interact J Med Res 2019 Feb 11;8(1):e10146 [FREE Full text] [doi: 10.2196/10146] [Medline: 30741640]

XSL•FO
RenderX

125. Awofeso N, Imam SA, Ahmed A. Content analysis of media coverage of childhood obesity topics in UAE newspapers and popular social media platforms, 2014-2017. Int J Health Policy Manag 2019 Feb 01;8(2):81-89 [FREE Full text] [doi: 10.15171/ijhpm.2018.100] [Medline: 30980621]
126. Smith DJ, Mac VV, Hertzberg VS. Using Twitter for nursing research: a tweet analysis on heat illness and health. J Nurs Scholarsh 2021 May;53(3):343-350. [doi: 10.1111/jnu.12654] [Medline: 33757160]
127. Green BM, Van Horn KT, Gupte K, Evans M, Hayes S, Bhowmick A. Assessment of adaptive engagement and support model for people with chronic health conditions in online health communities: combined content analysis. J Med Internet Res 2020 Jul 07;22(7):e17338 [FREE Full text] [doi: 10.2196/17338] [Medline: 32492651]
128. Osakwe ZT, Cortés YI. Impact of COVID-19: a text mining analysis of Twitter data in Spanish language. Hisp Health Care Int 2021 Dec;19(4):239-245. [doi: 10.1177/15404153211020453] [Medline: 34323101]
129. Sümeyye Yorulmaz D, Karadeniz H. Vaccination refusal debate on social media in Turkey: a content analysis of the comments on Instagram blogs. Iran J Public Health 2022 Mar;51(3):615-623 [FREE Full text] [doi: 10.18502/ijph.v51i3.8938] [Medline: 35865064]
130. Choi E, Becker H, Kim S. A blog text analysis to explore psychosocial support in adolescents and young adults with cancer. Cancer Nurs 2022 Dec 11;46(2):143-151. [doi: 10.1097/ncc.0000000000001087]
131. Jun J, Wickersham K, Zain A, Ford R, Zhang N, Ciccarelli C, et al. Cancer and COVID-19 vaccines on Twitter: the voice and vaccine attitude of cancer community. J Health Commun 2023 Jan 02;28(1):1-14. [doi: 10.1080/10810730.2023.2168800] [Medline: 36755484]
132. Yashpal S, Raghunath A, Gencerliler N, Burns LE. Exploring public perceptions of dental care affordability in the United States: mixed method analysis via Twitter. JMIR Form Res 2022 Jul 01;6(7):e36315 [FREE Full text] [doi: 10.2196/36315] [Medline: 35658090]
133. Damier P, Henderson EJ, Romero-Imbroda J, Galimam L, Kronfeld N, Warnecke T. Impact of off-time on quality of life in Parkinson's patients and their caregivers: insights from social media. Parkinsons Dis 2022;2022:1800567 [FREE Full text] [doi: 10.1155/2022/1800567] [Medline: 36510568]
134. Miller WR, Malloy C, Mravec M, Sposato MF, Groves D. Nursing in the spotlight: Talk about nurses and the nursing profession on Twitter during the early COVID-19 pandemic. Nurs Outlook 2022;70(4):580-589 [FREE Full text] [doi: 10.1016/j.outlook.2022.02.009] [Medline: 35428483]
135. Hriberšek M, Eibensteiner F, Kapral L, Teufel A, Nawaz FA, Cenanovic M, et al. "Loved ones are not 'visitors' in a patient's life"-The importance of including loved ones in the patient's hospital stay: An international Twitter study of #HospitalsTalkToLovedOnes in times of COVID-19. Front Public Health 2023;11:1100280 [FREE Full text] [doi: 10.3389/fpubh.2023.1100280] [Medline: 36778575]
136. O'Sullivan L, Killeen RP, Doran P, Crowley RK. Adherence with reporting of ethical standards in COVID-19 human studies: a rapid review. BMC Med Ethics 2021 Jun 28;22(1):80 [FREE Full text] [doi: 10.1186/s12910-021-00649-9] [Medline: 34182962]
137. Denecke K, Bamidis P, Bond C, Gabarron E, Househ M, Lau AYS, et al. Ethical issues of social media usage in healthcare. Yearb Med Inform 2015 Aug 13;10(1):137-147 [FREE Full text] [doi: 10.15265/IY-2015-001] [Medline: 26293861]
138. Ferretti A, Ienca M, Sheehan M, Blasimme A, Dove ES, Farsides B, et al. Ethics review of big data research: What should stay and what should be reformed? BMC Med Ethics 2021 Apr 30;22(1):51 [FREE Full text] [doi: 10.1186/s12910-021-00616-4] [Medline: 33931049]
139. Kohn T, Shore C. The Ethics of University Ethics Committees: risk management and the research imagination. Death Public Univ Uncertain Futur High Educ Knowl Econ 2017 May:229-249. [doi: 10.2307/j.ctvw04bj2.17]
140. Friesen P, Redman B, Caplan A. Of straws, camels, research regulation, and IRBs. Ther Innov Regul Sci 2019 Jul;53(4):526-534. [doi: 10.1177/2168479018783740] [Medline: 30176736]
141. Sellers C, Samuel G, Derrick G. Reasoning "uncharted territory": notions of expertise within ethics review panels assessing research use of social media. J Empir Res Hum Res Ethics 2020;15(1-2):28-39 [FREE Full text] [doi: 10.1177/1556264619837088] [Medline: 31826689]
142. Kious B. The Nuremberg Code: its history and implications. Princet J Bioeth 2001;4:7-19. [Medline: 12166467]
143. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
144. Department of Health, Education,Welfare, National Commission for the Protection of Human Subjects of BiomedicalBehavioral Research. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. J Am Coll Dent 2014;81(3):4-13. [Medline: 25951677]
145. Wongkoblap A, Vadillo MA, Curcin V. Researching mental health disorders in the era of social media: systematic review. J Med Internet Res 2017 Jun 29;19(6):e228 [FREE Full text] [doi: 10.2196/jmir.7215] [Medline: 28663166]
146. Alim S. An initial exploration of ethical research practices regarding automated data extraction from online social media user profiles. First Monday 2014 Jul;19(7):105-127. [doi: 10.1136/pgmj.14.152.176] [Medline: 21313116]
147. Vitak J, Proferes N, Shilton K, Ashktorab Z. Ethics regulation in social computing research: examining the role of institutional review boards. J Empir Res Hum Res Ethics 2017 Dec;12(5):372-382. [doi: 10.1177/1556264617725200] [Medline: 28831844]

XSL•FO
RenderX

148. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013 Sep;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
149. Metcalf J, Crawford K. Where are human subjects in Big Data research? The emerging ethics divide. Big Data & Society 2016 Jun 01;3(1):205395171665021. [doi: 10.1177/2053951716650211]
150. Usha Lawrance J, Nayahi Jesudhasan JV. Privacy preserving parallel clustering based anonymization for big data using MapReduce framework. Applied Artificial Intelligence 2021 Oct 17;35(15):1587-1620. [doi: 10.1080/08839514.2021.1987709]
151. Wilkinson D, Thelwall M. Researching personal information on the public web. Social Science Computer Review 2010 Aug 17;29(4):387-401. [doi: 10.1177/0894439310378979]
152. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. Croat Med J 2003 Dec;44(6):770-783 [FREE Full text] [Medline: 14725274]
153. Boyd D, Crawford K. Critical questions for big data. Information, Communication & Society 2012 Jun;15(5):662-679. [doi: 10.1080/1369118X.2012.678878]
154. Bond CS, Ahmed OH, Hind M, Thomas B, Hewitt-Taylor J. The conceptual and practical ethical dilemmas of using health discussion board posts as research data. J Med Internet Res 2013 Jun 07;15(6):e112 [FREE Full text] [doi: 10.2196/jmir.2435] [Medline: 23748147]
155. Markham A, Tiidenberg K, Herman A. Ethics as methods: doing ethics in the era of big data research—introduction. Social Media + Society 2018 Jul 19;4(3):205630511878450-205630511878416. [doi: 10.1177/2056305118784502]

Abbreviations

IRB: institutional review board
MeSH: Medical Subject Headings
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by T Leung, T de Azevedo Cardoso; submitted 02.08.23; peer-reviewed by E Zibrowski, J Scheibner; comments to author 06.10.23; revised version received 29.11.23; accepted 16.04.24; published 17.05.24.

Please cite as:
Zhang Y, Fu J, Lai J, Deng S, Guo Z, Zhong C, Tang J, Cao W, Wu Y
Reporting of Ethical Considerations in Qualitative Research Utilizing Social Media Data on Public Health Care: Scoping Review
J Med Internet Res 2024;26:e51496
URL: https://www.jmir.org/2024/1/e51496
doi:10.2196/51496
PMID:38758590

©Yujie Zhang, Jiaqi Fu, Jie Lai, Shisi Deng, Zihan Guo, Chuhan Zhong, Jianyao Tang, Wenqiong Cao, Yanni Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Mobile Health Apps, Family Caregivers, and Care Planning: Scoping Review

Marjorie M Kelley[1], MS, RN, PhD; Tia Powell[2], MD; Djibril Camara[3], MPH, MD; Neha Shah[4], MSPH; Jenna M Norton[4], MPH, PhD; Chelsea Deitelzweig[5], BA; Nivedha Vaidy[4]; Chun-Ju Hsiao[6], PhD; Jing Wang[7], MPH, RN, PhD; Arlene S Bierman[5], MS, MD

[1]The Ohio State University College of Nursing, Columbus, OH, United States
[2]Montefiore Einstein Center for Bioethics, Albert Einstein College of Medicine, Bronx, NY, United States
[3]Credence Management Solution, USAID Global Health Technical Professionals, Washington, DC, United States
[4]National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, United States
[5]Agency for Health Care Research and Quality, Rockville, MD, United States
[6]Center for Evidence and Practice Improvement, Agency for Health Care Research and Quality, Rockville, MD, United States
[7]Florida State University College of Nursing, Tallahassee, FL, United States

Corresponding Author:
Arlene S Bierman, MS, MD
Agency for Health Care Research and Quality
5600 Fishers Lane
Rockville, MD, 20857
United States
Phone: 1 (301) 427 1104
Email: arlene.bierman@ahrq.hhs.gov

Abstract

Background: People living with multiple chronic conditions (MCCs) face substantial challenges in planning and coordinating increasingly complex care. Family caregivers provide important assistance for people with MCCs but lack sufficient support. Caregiver apps have the potential to help by enhancing care coordination and planning among the health care team, including patients, caregivers, and clinicians.

Objective: We aim to conduct a scoping review to assess the evidence on the development and use of caregiver apps that support care planning and coordination, as well as to identify key factors (ie, needs, barriers, and facilitators) related to their use and desired caregiver app functionalities.

Methods: Papers intersecting 2 major domains, mobile health (mHealth) apps and caregivers, that were in English and published from 2015 to 2021 were included in the initial search from 6 databases and gray literature and ancestry searches. As per JBI (Joanna Briggs Institute) Scoping Review guidelines and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews), 2 authors independently screened full texts with disagreements resolved by a third author. Working in pairs, the authors extracted data using a pilot-tested JBI extraction table and compared results for consensus.

Results: We identified 34 papers representing 25 individual studies, including 18 (53%) pilot and feasibility studies, 13 (38%) qualitative studies, and 2 experimental or quasi-experimental studies. None of the identified studies assessed an intervention of a caregiver app for care planning and coordination for people with MCCs. We identified important caregiver needs in terms of information, support, and care coordination related to both caregiving and self-care. We compiled desired functionalities and features enabling apps to meet the care planning and care coordination needs of caregivers, in particular, the integration of caregiver roles into the electronic health record.

Conclusions: Caregiver needs identified through this study can inform developers and researchers in the design and implementation of mHealth apps that integrate with the electronic health record to link caregivers, patients, and clinicians to support coordinated care for people with MCCs. In addition, this study highlights the need for more rigorous research on the use of mHealth apps to support caregivers in care planning and coordination.

(J Med Internet Res 2024;26:e46108) doi:10.2196/46108

XSL•FO
RenderX

KEYWORDS

caregivers; carers; informal caregivers; family caregivers; mHealth applications; telemedicine; mobile health; mHealth; eHealth; digital health; apps; chronic condition; caregiver; application; support; clinicians; development; electronic health record; implementation; mobile phone

Introduction

Background

In 2020, between 17.7 and 40 million Americans were family caregivers of adults aged 65 years or older [1], defined as unpaid relatives, partners, or friends who assist persons in daily activities due to disease, disability, or other conditions. The need for family caregivers is projected to increase by 2030 with the older adult population and complexity of care increasing [2]. Many care recipients have multiple chronic conditions (MCCs) defined as the presence of 2 or more chronic physical or mental health conditions [3]. Over a quarter of the US adult population (27.2%) struggles with MCCs, with the highest prevalence (76.9%) among adults with both Medicare and Medicaid [3]. People living with MCCs are high users of care, including outpatient, emergency, inpatient, postacute, home, and long-term care, as well as prescription drugs [4]. People with MCCs account for 64% of all clinician visits, 70% of all in-patient stays, 83% of all prescriptions, 71% of all health care spending, and 93% of Medicare spending [5].

Complex care routines are common among patients with MCCs and often difficult for people living with MCCs and their caregivers to maintain, leading to avoidable adverse events, poor health outcomes, increased health spending, duplication of services, and polypharmacy [6]. The many challenges associated with care complexity and care planning add to the physical, psychological, and financial burdens associated with caregiving [7]. In fact, 14.5% of American caregivers have reported that they experienced mental health decline for at least half the days in a month [2].

Poor caregiver health and unmet needs have been widely documented and include mental and physical health concerns [8], unmet need for information on medication and care management to support the care recipient [7], limited access to supportive services [7], issues with communication across the care continuum [9], and burdens associated with work, social isolation [7], and finances [10]. Importantly, assistance with care coordination and planning has been consistently noted as an unmet need for caregivers [11].

Care Planning and Care Coordination

Developing care plans and organizing care involves the marshaling of personnel and other resources needed to carry out essential patient care activities and requires the exchange of information among participants responsible for different aspects of care [12]. Care planning is a collaborative process focused on discussing patient and clinical goals of care, conducting shared decision-making to identify strategies for clinical and self-management to achieve these goals based on evidence and patient preference, clarifying roles for different members of the care team, and empowering patients and caregivers [13]. These processes link health professionals, caregivers, and patients in the tasks of designing and implementing care.

Developing a comprehensive care plan both requires and supports care coordination by aggregating and streamlining data on health and social concerns, goals, care management strategies, and health status. Effective care coordination entails the organization of patient care activities to facilitate the appropriate and timely delivery of health care services by multiple clinicians in multiple care settings [12]. Care coordination involves the patient, clinicians, health care teams including nurses, pharmacists, physical therapists, and social workers, and caregivers. Such care coordination has been shown to benefit multiple domains, including decreased symptoms and mortality, and increased quality of life [14].

Digital Solutions

Digital solutions offer an opportunity to alleviate some of the care planning and coordination burdens currently shouldered by caregivers and patients. Digital health solutions encompass a variety of information or communication technologies applied to health needs. Digital health is mobile health (mHealth) when implemented on mobile devices. Digital health apps—or programs designed to accomplish specific tasks—fall into the category of mHealth when they are designed to operate on a mobile device.

mHealth apps have the inherent capability of increasing the reach of interventions, and transcending geography and time. They are also often more broadly accessible in the United States, as the uptake of mobile devices is greater than desktop computers [11]. Furthermore, they can be explicitly tailored to individual needs. Recent advances in technology and software now allow apps to be linked to other digital devices and the electronic health record (EHR).

Several systematic reviews outlined challenges associated with existing apps for caregivers, especially insufficient scientific evidence to support the efficacy of these apps [15-20]. However, no review has focused either on care planning and coordination apps overall or on caregivers of people with MCCs. Moreover, no review focused on the importance of care planning and coordination between the caregiver, care recipient, and professional health care providers. We conducted a scoping review to examine the evidence on the development and use of caregiver apps designed to support care planning and coordination, identify key factors related to their use (ie, needs, barriers, and facilitators), and characterize desired functionality. This review was undertaken to inform the development of a comprehensive, interoperable electronic care plan with clinician-, patient-, and caregiver-facing components to enhance care planning and coordination, address fragmentation of health care, and enhance the collection and sharing of critical patient-centered data across community, clinical, and research settings for people living with MCCs. The Agency for Healthcare Research and Quality (AHRQ) and the National

XSL•FO
RenderX

Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), with support from the Assistant Secretary for Planning and Evaluation's Patient-Centered Outcomes Research Trust Fund, are working in partnership to develop an interoperable e-care plan.

Methods

We conducted a scoping literature review using JBI (Joanna Briggs Institute) Scoping Review guidelines [21] and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [22] to guide our methods and reporting. Papers published in English between January 2011 and June 2021 were included. Hence, our initial search activity specific to care planning and coordination revealed a dearth of papers, we broadened our search to include papers intersecting 2 major domains: mHealth apps and caregivers (Figure 1). We hoped to capture available information relevant to care planning and coordination from the perspective of the caregiver. We included mobile health apps like native apps (ie, residing on smartphones) as well as web-based apps designed for smartphone formats. We included all diseases and conditions and care settings (eg, ambulatory, hospital, home, hospice, and long-term care). Study types included pilot and feasibility and experimental and quasi-experimental study designs. Source documents included academic peer-reviewed journals, dissertations and theses, government policy documents, and white papers published by caregiver advocacy organizations (eg, AARP [American Association of Retired Persons] and National Alliance for Caregiving). Studies including paid caregivers or caregivers of patients aged younger than 18 years were excluded. Interventions delivered via social media, phone calls (including interactive voice response), video, telehealth, or text messaging alone were excluded. We also excluded interventions delivered in low- and middle-income countries given significant differences in information technology infrastructure and patterns of use [23]. As such, comparisons would be difficult. Research interventions involving assistive technologies (ie, motion sensors), non–health related, and health literacy alone were excluded. Source documents such as opinion or editorial papers, conference posters or abstracts, study protocols, blogs, and websites were excluded. Key search terms (Textbox 1) alone or in combination, were used to create our search protocols in 6 databases: PubMed, Cochrane, CINAHL, SCOPUS, Web of Science, and Embase. We conducted ancestry searches of caregiver app reviews and caregiver literature reviews and searched several domain-specific journal databases including the *Journal of the American Medical Informatics Association*, *Journal of Medical Internet Research*, *International Journal of Medical Informatics*, *Journal of the American Medical Association*, and *New England Journal of Medicine*.

Figure 1. PRISMA flow diagram (adapted from Tricco et al [22], with permission from PRISMA). PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

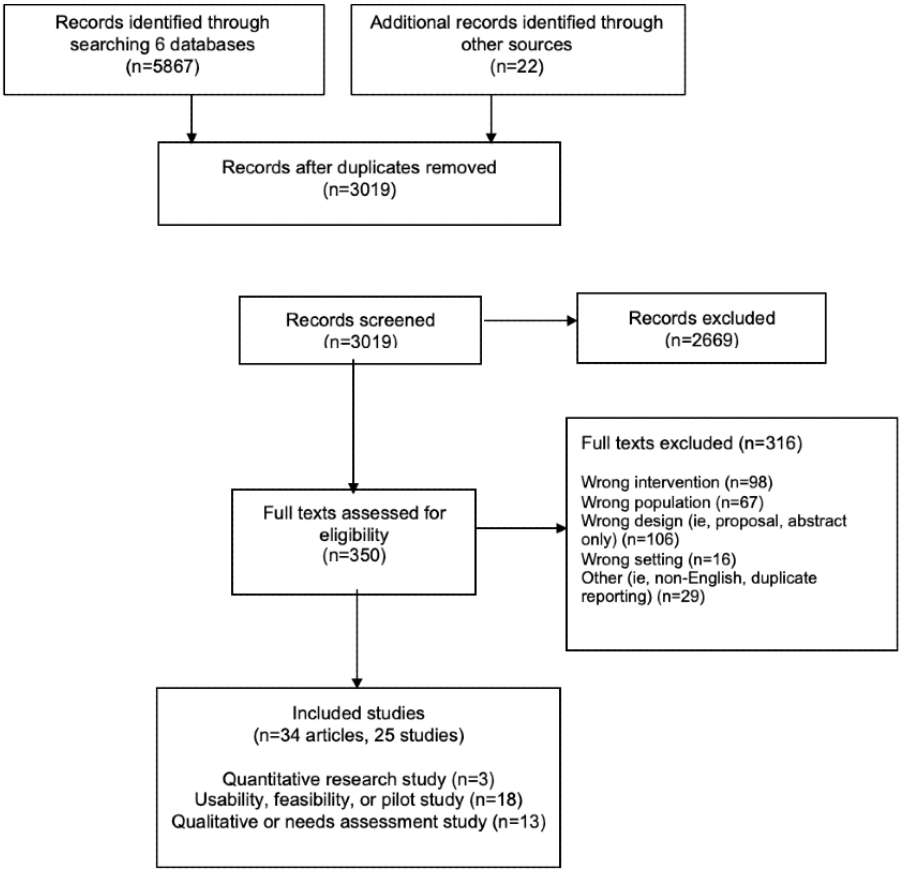

XSL•FO
RenderX

Textbox 1. Search terms.

Caregiver

- Caregiver; caretaker; care provider; carer; care

mHealth app

- mHealth; “mobile health” app; applications; “digital application”; eHealth; and smartphone
- Medical Subject Headings terms: telemedicine [encompasses mHealth]; mobile applications

We exported search results into EndNote, a reference management software platform to eliminate duplications, then uploaded them into Covidence, a web-based systematic review platform, to streamline evidence synthesis and author collaboration. Covidence allowed the research team to work collaboratively when screening papers at the title, abstract, and full-text level. In total, 2 authors independently screened titles and abstracts for eligibility with full-text screening conducted in the same manner. Screening disagreements were resolved through discussion or review by a third author. In keeping with scoping review methodological practices, critical appraisal, and risk of bias were not assessed.

Working in pairs, authors independently extracted data after adapting the JBI data extraction template and a previously used and pilot-tested data extraction table [24]. Then, each author compared results with the other for consensus about the extracted element. Data extraction elements included first author, publication date, health care domain of the care recipient, country, title, participant demographics, study purpose, study design, intervention description, app name and hyperlink if available, primary app users (ie, patient, caregiver, health care provider, and other), key or primary findings, app features and functionality—including desired functionality, how app supported care coordination, and how app supported caregivers (Multimedia Appendix 1 [25-45]). For qualitative studies, we extracted data elements associated with caregiver needs and desires. We used conventional content analysis methods, previously described by Hsieh and Shannon [46], to code and group categories as the phenomena of interest was new with little of the theoretical or literature available to guide the analysis. In keeping with conventional content analysis methods [47], we relied on inductive category development as categories and subcategories emerged from the literature, followed by deductive category and subcategory assignment.

Results

Overview

Of the 3019 nonduplicative records screened, 34 papers [25-45,48-60] representing 25 individual studies were included in this scoping literature review (Figure 1; Multimedia Appendix 1). Publication dates ranged from 2015 to 2021, with 29 (76%) papers published between January 2018 and August 2021. In total, 18 (53%) papers were feasibility, usability, or pilot studies [25-27,29-37,39-43,45] with qualitative or needs assessment papers representing 38% (n=13) [48-60]. Only 3 papers [28,38,44] reported using quantitative research methods to assess intervention efficacy (Textbox 2). Research was predominantly conducted in the United States (22 of 34). Further, 5 papers were from Australia, 3 from Spain, and one each from Canada, the United Kingdom, South Korea, and Turkey. In total, 14 papers focused on cancer caregiving, 7 on dementia caregiving, 6 on general caregiving, 2 each for stem cell transplant and mental health, and one each on heart failure, liver, mental health, and hospice. See Textbox 2 for details of the health care domain and paper type.

XSL•FO
RenderX

Textbox 2. Number of articles by health care domain and study type.

Cancer (n=14)

- Experimental and quasi-experimental (n=1)
- Pilot, feasibility, or usability (n=9)
- Qualitative (n=4)

Dementia (n=7)

- Experimental and quasi-experimental (n=1)
- Pilot, feasibility, or usability (n=2)
- Qualitative (n=4)

General caregiving (n=6)

- Experimental and quasi-experimental (n=1)
- Pilot, feasibility, or usability (n=4)
- Qualitative (n=1)

Mental health (n=2)

- Qualitative (n=2)

Stem cell transplant (n=2)

- Pilot, feasibility, or usability (n=1)
- Qualitative (n=1)

Heart failure (n=1)

- Qualitative (n=1)

Liver (n=1)

- Pilot, feasibility, or usability (n=1)

Hospice pain management (n=1)

- Pilot, feasibility, or usability (n=1)

Experimental and Quasi-Experimental Studies

Of the 3 quasi-experimental or experimental studies [28,38,44], Park and colleagues [38] developed an app for caregivers focused on knowledge of dementia, communication, and coping. Ferré-Grau and colleagues [28] conducted a randomized controlled trial of an app intervention designed to promote caregiver mental health. Finally, research conducted by Uysal et al [44], used an app for caregivers of patients with cancer focused on caregiver self-care and education. Overall, these studies, like many mHealth interventions for caregivers, addressed important caregiver needs including quality of life. However, none of these apps linked to information in the EHRs or leveraged data standards to support interoperability of data across the care team, nor did the apps provide enhanced communication among caregivers and the health care team. None of the studies investigated or measured care planning or coordination.

Pilot and Feasibility Studies

In total, 18 pilot and feasibility papers [25-27,29-37,39-43,45], representing 12 studies, were included in this review. The majority (n=14, 78%) of these studies used small convenience samples. Furthermore, 11 of the papers focused exclusively on caregiver mental health or included a component of caregiver mental health in the interventions [25-27,32-34,36,37,39,40,43]. In total, 5 reported on apps that included disease education or caregiving education [25,26,31,37,41]. Further, 3 focused on caregiver communications with family and friends [41,42,45] but did not assess care coordination or communication with health care providers. One included education on the skills necessary to communicate with health care professionals but did not assess care planning, coordination, or communication as an outcome as it was a feasibility study [45].

Most of the pilot and feasibility studies focused on the important goal of supporting caregivers' wellness but did not address care planning or coordination. For example, in one study—with results described in 3 papers [27,39,40]—the researchers conducted a 12-week feasibility study using a psycho-educational intervention delivered via video sessions with a goal of caregiver stress reduction. In another study [25,37], investigators used a mindfulness app and assessed cultural sensitivity and barriers to use as feasibility criteria.

XSL•FO
RenderX

Kubo and colleagues [32-34] evaluated a commercially available mindfulness app to assess the feasibility of use to improve caregivers' mental health. Similarly, Sikder and colleagues [43] pilot-tested an app focused on improving depression symptoms among caregivers.

In total, 7 papers included caregivers only as participants [27,31,36,39,40,42,43], while 9 papers included caregivers and care recipients as participants [25,29,30,32-35,37,41]. Only 2 feasibility studies, one conducted by Brown et al [26] and the other conducted by Wittenberg and colleagues [45], also included health professionals as participants. Brown and colleagues [26] examined the feasibility of an app for dementia caregivers, and included caregivers, homecare case managers, and primary health care providers as participants. The platform, CareHeros, was designed with the goal of bidirectional sharing of care recipients' information between caregivers and health care professionals. The platform did not communicate with EHRs, and bidirectional communication was only reported between case managers and primary care providers, exclusive of caregivers and care recipients. There was limited uptake of the app, with participants logging into CareHeros an average of only 2.18 times over the 11-week period of this study. Wittenberg and colleagues [45] demonstrated the feasibility of an mHealth app to support caregiver communication skills related to caregiving. The overall objectives of the app development included: (1) to improve caregiver communication skills related to caregiving, (2) to facilitate information sharing among family members, (3) to provide self-care resources for caregivers, and (4) to increase caregiver knowledge. The app was not designed to connect to the EHR, nor was it designed to increase or support communication between caregivers and health care professionals. Caregivers and health care professionals participated in the design and the development of the app as well as usability and acceptability testing. Both groups found the app to be usable and acceptable for helping caregivers with educational needs and communication skills related to caregiving.

While none of the 18 pilot and feasibility studies directly evaluated care planning or coordination as an aim or outcome, 2 [30,35] investigated apps that could assist in care delivery—with caregivers assessing care recipients' pain [35] and caregivers assessing care recipients' hepatic encephalopathy [30]. Ganapathy and colleagues [30] used the PatientBuddy app, which sent alerts with critical values regarding hepatic encephalopathy to dyads of patients and caregivers as well as clinicians to support care management, obtaining a positive impact reducing 30-day readmissions in a small cohort. Mayahara et al [35] conducted a pilot study using e-Pain Reporter, which assisted caregivers in assessing and managing the pain of family members in home hospice. The e-Pain Reporter was designed to provide information on patient pain and pain management to nurses in real time. However, this pilot study did not assess the communication aspect of the app.

In summary, among these pilot and feasibility studies, heterogeneity in study design, interventions, and outcomes preclude meta-analysis, generalization, and direct comparisons. Additionally, most failed to provide support for care planning or coordination and none linked with the EHR or leveraged interoperable data standards. As with most pilot and feasibility studies, these results were preliminary, not powered to identify statistically significant differences in outcomes, and were specific to the app under investigation. Still, a small number of promising studies [26,30,35,45] attempted to enhance communication or information sharing, a component of care planning and coordination.

Qualitative Studies

In total, 13 (38%) papers [48-60] included in this review were qualitative studies assessing caregiver needs associated with mHealth apps. These caregiver needs were synthesized into 3 broad categories: (1) needs associated with providing care, (2) needs associated with self-care, and (3) desired app features and functionality. In terms of providing care (category 1), caregivers needed information, support, and help with care coordination. For self-care (category 2), caregivers reported a need for information and support. A detailed list of desired mHealth app features and functionality (category 3) is provided in Textbox 3.

XSL•FO
RenderX

Textbox 3. Qualitative outcomes—caregiver needs.

Needs associated with providing care

- **Information**
  - Adjusting to a new role
  - Information on disease or condition of care recipient
  - Information on disease or condition common comorbidities
  - Symptom, behavior, or safety
  - When to seek help
  - Changing nature of caregiving
  - Financial and legal services (financial assistance, job help, and health care payment)
  - On-demand education and training
  - Community support links (transportation or community reintegration)
  - Content tailored to care recipients' needs
  - Simple—easy to understand
  - Up-to-date scientific evidence and mechanism for updating the information
  - Multimodal delivery of information: video, audio, text, or animations
  - Always accessible
- **Support**
  - Support for care recipients' physical and emotional needs
  - Support with rehabilitation and activities of daily living (oral, bathing, dressing, grooming, toileting, feeding and nutrition, transferring, and ambulation)
  - Decision-making support
  - Medication management
  - Tracking and monitoring of care recipient—mental, physical, emotional, and social (including symptoms, vital signs)
  - Content tailored to care recipient's needs
  - Family or personal relationships (asking for help, safety, and communication)
  - Always accessible
- **Care coordination**
  - Integrated app with health care system—care coordination
  - Ability to complete questionnaires at home, unrushed
  - Finding care equipment
  - List of important contacts and contact information for quick reference
  - Information and connection to support services (specialty care, first responders, advocacy organizations, and respite services)
  - Relationships with health care providers (personal contact)
  - Feedback from health care providers—instant
  - Automated data entry and reminders or prompts
  - "One-stop-shopping"—all information in 1 place

Needs associated with self-care

- **Information**
  - Information to help improve caregivers' health (stress management, peer support, and support groups)
  - Activities, programs, and therapy to improve mental, physical, and social support of caregivers
  - Content tailored to caregivers' needs
  - Family or personal relationship help (safety or asking others for help or support)

XSL•FO
RenderX

- Preventing social isolation

- **Support**
  - Tracking and monitoring of caregiver—mental, physical, emotional, or social (including symptoms, mental health, vital signs)
  - Content tailored to caregiver needs
  - Social media—“people like me” with expert moderator
  - Peer mentor, support, or coaching

Desired mHealth app features and functionality

- Easy to use
- Easy to learn
- Integrated with phone contacts and other apps (exercise and weight management)
- Ability to report care recipient status or symptoms to health care providers and get a response, feedback, or follow-up quickly
- Task reminders (appointments, medication management, etc)
- Integrate with other platforms or devices (electronic health records, smart watches, or pharmacy)
- Share information with family members
- Integrate music or other entertainment
- Track patient symptoms or issues over time
- Track caregiver issues over time
- Customizable
- App from a trusted source and evidence-based content
- Data secure
- Integrated across health care systems
- Not too much information—just in time with the right information
- Affordable
- Font or screen size readable—Americans with Disabilities Act Standards for Accessible Design compliant
- Sustainable
- Help for digital naïve
- Does not reduce time with physician
- Clear perceived benefit
- Ability to personalize features and functions
- Automated data entry

Discussion

Principal Findings

This scoping review synthesized the evidence on the development and use of caregiver apps designed to enable or support caregiver participation in care planning or care coordination. We identified key factors (ie, needs, barriers, and facilitators) related to care planning and coordination. We described important functionalities and features enabling caregiver apps to meet care planning and coordination needs and facilitate caregiving activities. This comprehensive summary of caregiver needs related to health apps and care coordination may be useful to developers and researchers as it relates to caregivers of those living with MCCs. A better understanding of usability and overall needs will enhance ongoing research efforts to improve e-care planning and care coordination among these populations.

Of the 34 papers, representing 25 individual studies included in this review, only 3 were experimental or quasi-experimental intervention studies [28,38,44]. None of the studies included in this review focused on care planning, care coordination, or care recipients with MCCs. This paucity of research precluded generalizations about caregivers’ apps, much less in care planning and coordination. Although most of the studies included in this review addressed important caregiver factors including caregiver education, coping, and self-care, these standalone interventions lacked components to reduce caregiver burdens associated with planning and coordinating complex care. An app designed to specifically improve care planning and coordination, thus reducing this burden, is

XSL•FO
RenderX

needed—particularly for the increasing number of care recipients with MCCs.

Most studies within this review were qualitative studies or pilot and feasibility studies. Yet, a few of these studies [26,30,35,45] identified elements important for care planning and coordination in mHealth apps. By definition, these studies are preliminary in nature thus precluding generalizations; they do not represent proven efficacy or settled science. However, they provide a foundation for future exploration of the role of mHealth interventions in promoting care planning and coordination.

Comparison to Prior Work

Our findings parallel and extend the results identified in a recent review focused on native apps for informal caregiving [61]. Native apps are apps residing on smartphones as opposed to web-based apps. The principal findings specific to native apps [61] align with our more comprehensive review (including both native and web-based apps) in that the nascent technology has not matured enough to make meaningful recommendations beyond that of caregiver needs and wants. More rigorous research is needed, specifically among caregivers of patients with MCCs.

In terms of caregiver needs associated with care planning and coordination, caregivers and care recipients in included studies identified several important areas of needs and wants including apps that delivered “one-stop-shopping” or all the information in 1 place. These needs and wants were similar to those identified by Margarido and colleagues [61] in their 2022 scoping review. The results from both indicated caregivers wanted apps that integrated with the health care system (including the EHR) and could allow them to complete questionnaires at home in an unrushed fashion. They wanted apps that could help them find care equipment and information about support services and support contacts. Relationships with health care providers and feedback from the providers were of key importance, as were timely reminders and prompts (eg, upcoming appointments and medication changes).

Future Directions

More research is needed as this scoping review did not identify any of the following: an app designed to provide access and enhance communication among caregivers, patients, and health care workers, with access for all 3 groups to the EHR; use of data standards in apps to promote interoperability of data across the care team, including caregivers and care recipients; a focus on care planning and coordination; a free and publicly available digital platform; or demonstration of successful usability, efficacy, and sustainability.

The potential exists for emerging mHealth apps to contribute to care coordination by linking caregivers, patients, and clinicians to information and resources that improve the ability of the entire care team to actively engage. Ongoing research focused on developing and evaluating [62-65] interventions to support caregiver engagement in health care through direct EHR access and other digital means could provide important insights. Today, mHealth app-facilitated care planning and coordination remains a possibility, not a reality. This scoping review provides further evidence that existing caregiver-facing mHealth apps are not sufficiently supported by research, with many studies focused on well-educated, tech-savvy female caregivers [30,37,50]. There is a need for app development to meet caregiver needs in diverse populations. Most such apps address the burdens of caregiving through interventions aimed at education, self-care, and stress reduction. Though these are helpful, they do not address the fundamental challenges related to care planning and coordination.

Current government federal policies encourage care planning and coordination. There is a federal mandate through the Office of the National Coordinator for Health Information Technology for third-party mHealth apps to integrate with the EHR. These technologies need to be implemented into current health care workflows, but data blocking and the inability to write back to the EHR present challenges. Current workforce shortages, especially for nurses, are well documented and may increase the difficulty of introducing new technologies and tasks, requiring both additional training and time from an already overburdened workforce. On the other hand, a well-designed app that facilitates information sharing, care planning, and communication could potentially reduce the burden.

Strengths and Limitations

We acknowledge several limitations of this study. First, papers included in the scoping review do not include work published after June 2021. It is possible our search terms failed to identify relevant papers in this rapidly developing field. Second, the review included only studies published in English. Though digital health literature is predominantly published in English, there is the possibility of missing important work in other languages. In keeping with guidelines for scoping reviews, we neither assessed the risk of bias nor methodologies in the included studies. Finally, the heterogeneity of included research precluded a meta-analysis of findings across all studies.

Summary

This scoping review synthesizes the current evidence on developing mHealth apps to support caregivers in care planning and coordination, providing insights to inform future mHealth app development to engage caregivers as members of the health care team, share critical information across the entire health care team, reduce the burdens caregivers experience in trying to coordinate care, as well as identifying the functionality caregivers desired. Few experimental studies involving apps with needed functionality were identified in the scoping review, even though use of digital technology for caregiver support is a growing interest. We found no studies focused on care planning or coordination, and a very small number of pilot and other preliminary studies addressing specific aspects of care coordination, such as communication. Given the limited number of studies and the preliminary nature of many, there is insufficient evidence on mHealth apps to support caregivers in care planning and coordination. However, the need and potential for further work to achieve these aims is substantial.

Conclusions

In sum, research and evidence on the effective use of mHealth apps to support caregivers involved in care planning and coordination for people living with MCCs is limited. Apps to

XSL•FO
RenderX

support caregivers have yet to be integrated into the EHRs. Multidirectional communication between caregivers, care recipients, and health care providers through the EHR holds great promise for relieving the burden on clinicians, patients, and their caregivers alike. The development and implementation of an mHealth app linking the 3 key stakeholder groups to work together to [65] enhance care planning and coordination, remains an unmet need. Prior work on the functionality desired by caregivers can inform this work.

Acknowledgments

No generative artificial intelligence was used. The findings and conclusions in this document are those of the authors, who are responsible for its content, and do not necessarily represent the views of the Agency for Healthcare Research and Quality (AHRQ), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), or the National Institutes of Health. No statement in this report should be construed as an official position of the AHRQ, NIDDK, National Institutes of Health, or the US Department of Health and Human Services. This project was conducted with support from the Assistant Secretary for Planning and Evaluation's Patient Centered Outcomes Research Trust Fund.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quantitative and pilot, feasibility, or acceptability study information.
[DOCX File , 33 KB - jmir_v26i1e46108_app1.docx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 498 KB - jmir_v26i1e46108_app2.pdf]

References

1. Committee on Family Caregiving for Older Adults, Board on Health Care Services, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine. In: Schulz R, Eden J, editors. Families Caring for an Aging America. Washington, DC: The National Academies Press; 2016.
2. Caregiving for family and friends—a public health issue. Center for Disease Control and Prevention. 2018. URL: https://www.cdc.gov/aging/caregiving/pdf/caregiver-brief-508.pdf [accessed 2024-03-14]
3. Boersma P, Black LI, Ward BW. Prevalence of multiple chronic conditions among US adults, 2018. Prev Chronic Dis 2020;17:E106 [FREE Full text] [doi: 10.5888/pcd17.200130] [Medline: 32945769]
4. Buttorff C, Ruder T, Bauman M. Multiple Chronic Conditions in the United States. Santa Monica, CA: RAND Corporation; 2017.
5. Gerteis J, Izrael D, Deitz D, LeRoy L, Ricciardi R, Miller T, et al. Multiple chronic conditions chartbook. In: Agency for Healthcare Research and Quality. Capitol Hill, NC: HHS; 2010.
6. Xu X, Mishra GD, Jones M. Evidence on multimorbidity from definition to intervention: an overview of systematic reviews. Ageing Res Rev 2017;37:53-68. [doi: 10.1016/j.arr.2017.05.003] [Medline: 28511964]
7. Price ML, Surr CA, Gough B, Ashley L. Experiences and support needs of informal caregivers of people with multimorbidity: a scoping literature review. Psychol Health 2020;35(1):36-69. [doi: 10.1080/08870446.2019.1626125] [Medline: 31321995]
8. Queluz FNFR, Kervin E, Wozney L, Fancey P, McGrath PJ, Keefe J. Understanding the needs of caregivers of persons with dementia: a scoping review. Int Psychogeriatr 2020;32(1):35-52 [FREE Full text] [doi: 10.1017/S1041610219000243] [Medline: 30967164]
9. Li J, Luo X, Cao Q, Lin Y, Xu Y, Li Q. Communication needs of cancer patients and/or caregivers: a critical literature review. J Oncol 2020;2020:7432849 [FREE Full text] [doi: 10.1155/2020/7432849] [Medline: 32454826]
10. Aung SHH, White K, Bloomfield J. The experiences and the needs of caregivers of patients with head and neck cancer: an integrative review. Cancer Nurs 2021;44(6):E361-E373. [doi: 10.1097/NCC.0000000000000833] [Medline: 32541207]
11. McGilton KS, Vellani S, Yeung L, Chishtie J, Commisso E, Ploeg J, et al. Identifying and understanding the health and social care needs of older adults with multiple chronic conditions and their caregivers: a scoping review. BMC Geriatr 2018;18(1):231 [FREE Full text] [doi: 10.1186/s12877-018-0925-x] [Medline: 30285641]
12. McDonald KM, Sundaram V, Bravata DM, Lewis R, Lin N, Kraft SA, et al. Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies (Vol. 7: Care Coordination). Rockville, MD: Agency for Healthcare Research and Quality; 2007.
13. Baker A, Cronin K, Conway PH, DeSalvo KB, Rajkumar R, Press MJ. Making the comprehensive shared care plan a reality. NEJM catalyst. 2016. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.16.0838 [accessed 2024-02-23]

XSL•FO
RenderX

14. Kastner M, Cardoso R, Lai Y, Treister V, Hamid JS, Hayden L, et al. Effectiveness of interventions for managing multiple high-burden chronic diseases in older adults: a systematic review and meta-analysis. CMAJ 2018;190(34):E1004-E1012 [FREE Full text] [doi: 10.1503/cmaj.171391] [Medline: 30150242]
15. Chelberg GR, Neuhaus M, Mothershaw A, Mahoney R, Caffery LJ. Mobile apps for dementia awareness, support, and prevention—review and evaluation. Disabil Rehabil 2022;44(17):4909-4920. [doi: 10.1080/09638288.2021.1914755] [Medline: 34034601]
16. Désormeaux-Moreau M, Michel CM, Vallières M, Racine M, Poulin-Paquet M, Lacasse D, et al. Mobile apps to support family caregivers of people with Alzheimer disease and related dementias in managing disruptive behaviors: qualitative study with users embedded in a scoping review. JMIR Aging 2021;4(2):e21808 [FREE Full text] [doi: 10.2196/21808] [Medline: 33861207]
17. Lorca-Cabrera J, Martí-Arques R, Albacar-Riobóo N, Raigal-Aran L, Roldan-Merino J, Ferré-Grau C. Mobile applications for caregivers of individuals with chronic conditions and/or diseases: quantitative content analysis. Int J Med Inform 2021;145:104310. [doi: 10.1016/j.ijmedinf.2020.104310] [Medline: 33161319]
18. Petrovic M, Gaggioli A. Digital mental health tools for caregivers of older adults-a scoping review. Front Public Health 2020;8:128 [FREE Full text] [doi: 10.3389/fpubh.2020.00128] [Medline: 32411643]
19. Sala-González M, Pérez-Jover V, Guilabert M, Mira JJ. Mobile apps for helping informal caregivers: a systematic review. Int J Environ Res Public Health 2021;18(4):1702 [FREE Full text] [doi: 10.3390/ijerph18041702] [Medline: 33578819]
20. Vandersman P, Perimal-Lewis L, Tieman J. Is there an app for that? A scoping review of apps for care workers in the aged care sector. In: Telehealth Innovations in Remote Healthcare Services Delivery. Amsterdam, Netherlands: IOS Press eBooks; 2021:124-134.
21. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126 [FREE Full text] [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
22. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
23. McCool J, Dobson R, Whittaker R, Paton C. Mobile health (mHealth) in low- and middle-income countries. Annu Rev Public Health 2022;43:525-539 [FREE Full text] [doi: 10.1146/annurev-publhealth-052620-093850] [Medline: 34648368]
24. Kelley MM, Kue J, Brophy L, Peabody AL, Foraker RE, Yen PY, et al. Mobile health applications, cancer survivors, and lifestyle modification: an integrative review. Comput Inform Nurs 2021;39(11):755-763 [FREE Full text] [doi: 10.1097/CIN.0000000000000781] [Medline: 34074873]
25. Beer JM, Smith KN, Kennedy T, Mois G, Acena D, Gallerani DG, et al. A focus group evaluation of breathe easier: a mindfulness-based mHealth app for survivors of lung cancer and their family members. Am J Health Promot 2020;34(7):770-778. [doi: 10.1177/0890117120924176] [Medline: 32406241]
26. Brown EL, Ruggiano N, Page TF, Roberts L, Hristidis V, Whiteman KL, et al. CareHeroes Web and Android™ apps for dementia caregivers: a feasibility study. Res Gerontol Nurs 2016;9(4):193-203 [FREE Full text] [doi: 10.3928/19404921-20160229-02] [Medline: 29977440]
27. Carr AL, Jones J, Gilbertson SM, Laudenslager ML, Kutner JS, Kilbourn K, et al. Impact of a Mobilized Stress Management Program (Pep-Pal) for caregivers of oncology patients: mixed-methods study. JMIR Cancer 2019;5(1):e11406 [FREE Full text] [doi: 10.2196/11406] [Medline: 31066678]
28. Ferré-Grau C, Raigal-Aran L, Lorca-Cabrera J, Lluch-Canut T, Ferré-Bergadà M, Lleixá-Fortuño M, et al. A mobile app-based intervention program for nonprofessional caregivers to promote positive mental health: randomized controlled trial. JMIR mHealth uHealth 2021;9(1):e21708 [FREE Full text] [doi: 10.2196/21708] [Medline: 33480852]
29. Frisbee KL. Variations in the use of mHealth tools: the VA mobile health study. JMIR mHealth uHealth 2016;4(3):e89 [FREE Full text] [doi: 10.2196/mhealth.3726] [Medline: 27436165]
30. Ganapathy D, Acharya C, Lachar J, Patidar K, Sterling RK, White MB, et al. The patient buddy app can potentially prevent hepatic encephalopathy-related readmissions. Liver Int 2017;37(12):1843-1851. [doi: 10.1111/liv.13494] [Medline: 28618192]
31. Heynsbergh N, Heckel L, Botti M, Livingston PM. A smartphone app to support carers of people living with cancer: a feasibility and usability study. JMIR Cancer 2019;5(1):e11779 [FREE Full text] [doi: 10.2196/11779] [Medline: 30702432]
32. Kubo A, Altschuler A, Kurtovich E, Hendlish S, Laurent CA, Kolevska T, et al. A pilot mobile-based mindfulness intervention for cancer patients and their informal caregivers. Mindfulness (N Y) 2018;9(6):1885-1894 [FREE Full text] [doi: 10.1007/s12671-018-0931-2] [Medline: 30740187]
33. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C, et al. A randomized controlled trial of mHealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. Integr Cancer Ther 2019;18:1534735419850634 [FREE Full text] [doi: 10.1177/1534735419850634] [Medline: 31092044]

XSL•FO
RenderX

34. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C, et al. Pilot pragmatic randomized trial of mHealth mindfulness-based intervention for advanced cancer patients and their informal caregivers. Psycho-Oncology 2024;33(2):e5557. [doi: 10.1002/pon.5557] [Medline: 32979294]
35. Mayahara M, Wilbur J, Fogg L, Breitenstein SM, Miller AM. Feasibility of e-Pain reporter: a digital pain management tool for informal caregivers in home hospice. J Hosp Palliat Nurs 2019;21(3):193-199. [doi: 10.1097/NJH.0000000000000548] [Medline: 31045994]
36. Otero P, Hita I, Torres, Vázquez FL. Brief psychological intervention through mobile app and conference calls for the prevention of depression in non-professional caregivers: a pilot study. Int J Environ Res Public Health 2020;17(12):4578 [FREE Full text] [doi: 10.3390/ijerph17124578] [Medline: 32630541]
37. Owens OL, Smith KN, Beer JM, Gallerani DG, McDonnell KK. A qualitative cultural sensitivity assessment of the breathe easier mobile application for lung cancer survivors and their families. Oncol Nurs Forum 2020;47(3):331-341. [doi: 10.1188/20.ONF.331-341] [Medline: 32301935]
38. Park E, Park H, Kim EK. The effect of a Comprehensive Mobile Application Program (CMAP) for family caregivers of home-dwelling patients with dementia: a preliminary research. Jpn J Nurs Sci 2020;17(4):e12343. [doi: 10.1111/jjns.12343] [Medline: 32363664]
39. Pensak NA, Carr AL, Jones J, Mikulich-Gilbertson SK, Kutner JS, Kilbourn K, et al. A pilot study of mobilized intervention to help caregivers of oncology patients manage distress. Psycho-Oncology 2021;30(4):520-528. [doi: 10.1002/pon.5597] [Medline: 33217070]
40. Pensak NA, Joshi T, Simoneau T, Kilbourn K, Carr A, Kutner J, et al. Development of a web-based intervention for addressing distress in caregivers of patients receiving stem cell transplants: formative evaluation with qualitative interviews and focus groups. JMIR Res Protoc 2017;6(6):e120 [FREE Full text] [doi: 10.2196/resprot.7075] [Medline: 28642213]
41. Quinn CC, Staub S, Barr E, Gruber-Baldini A. Mobile support for older adults and their caregivers: dyad usability study. JMIR Aging 2019;2(1):e12276 [FREE Full text] [doi: 10.2196/12276] [Medline: 31518271]
42. Quintana Y, Fahy D, Crotty B, Gorenberg M, Jain R, Kaldany E, et al. A usability evaluation of the InfoSAGE app for family-based medication management. Stud Health Technol Inform 2019;257:352-357 ITCH. [Medline: 30741222]
43. Sikder AT, Yang FC, Schafer R, Dowling GA, Traeger L, Jain FA. Mentalizing imagery therapy mobile app to enhance the mood of family dementia caregivers: feasibility and limited efficacy testing. JMIR Aging 2019;2(1):e12850 [FREE Full text] [doi: 10.2196/12850] [Medline: 31518275]
44. Uysal N, Bağçivan G, Özkaraman A, Eşer AK, Çetin F, Çalışkan BB, et al. Empowering caregivers in the radiotherapy process: the results of a randomized controlled trial. Support Care Cancer 2021;29(5):2395-2404. [doi: 10.1007/s00520-020-05743-z] [Medline: 32918607]
45. Wittenberg E, Xu J, Goldsmith J, Mendoza Y. Caregiver communication about cancer: development of a mHealth resource to support family caregiver communication burden. Psycho-Oncology 2019;28(2):365-371 [FREE Full text] [doi: 10.1002/pon.4950] [Medline: 30471236]
46. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
47. Mayring P. Qualitative content analysis: theoretical background and procedures. In: Bikner-Ahsbahs A, Knipping C, Presmeg N, editors. Approaches to Qualitative Research in Mathematics Education. Advances in Mathematics Education. Dordrecht, Germany: Springer; 2015.
48. Bamgboje-Ayodele A, Levesque JV, Gerges M, Girgis A. What male caregivers of women with breast cancer want in an online intervention: a qualitative study. J Patient Exp 2020;7(6):1771-1777 [FREE Full text] [doi: 10.1177/2374373520926098] [Medline: 33457642]
49. Boele FW, van Uden-Kraan CF, Hilverda K, Weimer J, Donovan HS, Drappatz J, et al. Neuro-oncology family caregivers' view on keeping track of care issues using eHealth systems: it's a question of time. J Neurooncol 2017;134(1):157-167 [FREE Full text] [doi: 10.1007/s11060-017-2504-y] [Medline: 28550505]
50. Chaar D, Shin JY, Mazzoli A, Vue R, Kedroske J, Chappell G, et al. A mobile health app (Roadmap 2.0) for patients undergoing hematopoietic stem cell transplant: qualitative study on family caregivers' perspectives and design considerations. JMIR mHealth uHealth 2019;7(10):e15775 [FREE Full text] [doi: 10.2196/15775] [Medline: 31651402]
51. Frisbee KL. The impact of mobile health (mHealth) technology on family caregiver's burden levels and an assessment of variations in mhealth tool use. The George Washington University. 2015. URL: https://scholarspace.library.gwu.edu/concern/gw_etds/1544bp27m?locale=de [accessed 2024-03-14]
52. Heynsbergh N, Botti M, Heckel L, Livingston PM. Caring for the person with cancer: information and support needs and the role of technology. Psychooncology 2018;27(6):1650-1655 [FREE Full text] [doi: 10.1002/pon.4722] [Medline: 29624783]
53. Heynsbergh N, Heckel L, Botti M, Livingston PM, Seung CO. Development of a smartphone app for informal carers of people with cancer: processes and learnings. JMIR Form Res 2019;3(2):e10990 [FREE Full text] [doi: 10.2196/10990] [Medline: 30973346]
54. Huerta-Ramos E, Escobar-Villegas MS, Rubinstein K, Unoka ZS, Grasa E, Hospedales M, et al. Measuring users' receptivity toward an integral intervention model based on mHealth solutions for patients with treatment-resistant schizophrenia

XSL•FO
RenderX

(m-RESIST): a qualitative study. JMIR mHealth uHealth 2016;4(3):e112 [FREE Full text] [doi: 10.2196/mhealth.5716] [Medline: 27682896]
55. Kales HC, Gitlin LN, Stanislawski B, Marx K, Turnwald M, Watkins DC, et al. WeCareAdvisor™: the development of a caregiver-focused, web-based program to assess and manage behavioral and psychological symptoms of dementia. Alzheimer Dis Assoc Disord 2017;31(3):263-270 [FREE Full text] [doi: 10.1097/WAD.0000000000000177] [Medline: 27849639]
56. Owen JE, Jaworski BK, Kuhn E, Hoffman JE, Schievelbein L, Chang A, et al. Development of a mobile app for family members of veterans with PTSD: identifying needs and modifiable factors associated with burden, depression, and anxiety. J Fam Stud 2017;26(2):286-307. [doi: 10.1080/13229400.2017.1377629]
57. Rathnayake S, Moyle W, Jones C, Calleja P. Co-design of an mHealth application for family caregivers of people with dementia to address functional disability care needs. Inform Health Soc Care 2021;46(1):1-17. [doi: 10.1080/17538157.2020.1793347] [Medline: 32706282]
58. Ruggiano N, Brown EL, Shaw S, Geldmacher D, Clarke P, Hristidis V, et al. The potential of information technology to navigate caregiving systems: perspectives from dementia caregivers. J Gerontol Soc Work 2019;62(4):432-450. [doi: 10.1080/01634372.2018.1546786] [Medline: 30422754]
59. Shreve J, Baier RR, Epstein-Lubow G, Gardner RL. Dementia caregivers' technology preferences: design insights from qualitative interviews. Gerontechnology 2016;14(2):89-96 [FREE Full text] [doi: 10.4017/gt.2016.14.2.004.00]
60. Wali S, Keshavjee K, Nguyen L, Mbuagbaw L, Demers C. Using an electronic app to promote home-based self-care in older patients with heart failure: qualitative study on patient and informal caregiver challenges. JMIR Cardio 2020;4(1):e15885 [FREE Full text] [doi: 10.2196/15885] [Medline: 33164901]
61. Margarido MG, Shah A, Seto E. Smartphone applications for informal caregivers of chronically ill patients: a scoping review. NPJ Digit Med 2022;5(1):33 [FREE Full text] [doi: 10.1038/s41746-022-00567-z] [Medline: 35314766]
62. Piette JD, Striplin D, Marinec N, Chen J, Trivedi RB, Aron DC, et al. A mobile health intervention supporting heart failure patients and their informal caregivers: a randomized comparative effectiveness trial. J Med Internet Res 2015;17(6):e142 [FREE Full text] [doi: 10.2196/jmir.4550] [Medline: 26063161]
63. Aikens JE, Trivedi R, Heapy A, Pfeiffer PN, Piette JD. Potential impact of incorporating a patient-selected support person into mHealth for depression. J Gen Intern Med 2015;30(6):797-803 [FREE Full text] [doi: 10.1007/s11606-015-3208-7] [Medline: 25666218]
64. Wolff JL, Darer JD, Berger A, Clarke D, Green JA, Stametz RA, et al. Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. J Am Med Inform Assoc 2017;24(e1):e166-e172 [FREE Full text] [doi: 10.1093/jamia/ocw108] [Medline: 27497795]
65. Piette JD, Marinec N, Janda K, Morgan E, Schantz K, Yujra ACA, et al. Structured caregiver feedback enhances engagement and impact of mobile health support: a randomized trial in a lower-middle-income country. Telemed J E Health 2016;22(4):261-268 [FREE Full text] [doi: 10.1089/tmj.2015.0099] [Medline: 26352854]

Abbreviations

AARP: American Association of Retired Persons
AHRQ: Agency for Healthcare Research and Quality
EHR: electronic health record
JBI: Joanna Briggs Institute
MCCs: multiple chronic conditions
mHealth: mobile health
NIDDK: National Institute of Diabetes and Digestive and Kidney Diseases
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by T Leung; submitted 30.01.23; peer-reviewed by J Wolff, Y Chu; comments to author 27.05.23; revised version received 28.09.23; accepted 01.03.24; published 23.05.24.

Please cite as:
Kelley MM, Powell T, Camara D, Shah N, Norton JM, Deitelzweig C, Vaidy N, Hsiao CJ, Wang J, Bierman AS
Mobile Health Apps, Family Caregivers, and Care Planning: Scoping Review
J Med Internet Res 2024;26:e46108
URL: https://www.jmir.org/2024/1/e46108
doi:10.2196/46108
PMID:38781588

XSL•FO
RenderX

©Marjorie M Kelley, Tia Powell, Djibril Camara, Neha Shah, Jenna M Norton, Chelsea Deitelzweig, Nivedha Vaidy, Chun-Ju Hsiao, Jing Wang, Arlene S Bierman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Behavior Change Intervention Designs for Habit Formation: Systematic Review

Yujie Zhu[1,2], MDes; Yonghao Long[1], PhD; Hailiang Wang[1], PhD; Kun Pyo Lee[1,2], PhD; Lie Zhang[3], PhD; Stephen Jia Wang[1,2], PhD

[1]School of Design, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[2]Laboratory for Artificial Intelligence in Design, Hong Kong Science Park, Hong Kong, China (Hong Kong)

[3]Academy of Arts & Design, Tsinghua University, Beijing, China

Corresponding Author:
Stephen Jia Wang, PhD
School of Design, The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom
Hong Kong
China (Hong Kong)
Phone: 852 2766 5478
Email: stephen.j.wang@polyu.edu.hk

Abstract

Background: With the development of emerging technologies, digital behavior change interventions (DBCIs) help to maintain regular physical activity in daily life.

Objective: To comprehensively understand the design implementations of habit formation techniques in current DBCIs, a systematic review was conducted to investigate the implementations of behavior change techniques, types of habit formation techniques, and design strategies in current DBCIs.

Methods: The process of this review followed the PRISMA (Preferred Reporting Item for Systematic Reviews and Meta-Analyses) guidelines. A total of 4 databases were systematically searched from 2012 to 2022, which included Web of Science, Scopus, ACM Digital Library, and PubMed. The inclusion criteria encompassed studies that used digital tools for physical activity, examined behavior change intervention techniques, and were written in English.

Results: A total of 41 identified research articles were included in this review. The results show that the most applied behavior change techniques were the self-monitoring of behavior, goal setting, and prompts and cues. Moreover, habit formation techniques were identified and developed based on intentions, cues, and positive reinforcement. Commonly used methods included automatic monitoring, descriptive feedback, general guidelines, self-set goals, time-based cues, and virtual rewards.

Conclusions: A total of 32 commonly design strategies of habit formation techniques were summarized and mapped to the proposed conceptual framework, which was categorized into target-mediated (generalization and personalization) and technology-mediated interactions (explicitness and implicitness). Most of the existing studies use the explicit interaction, aligning with the personalized habit formation techniques in the design strategies of DBCIs. However, implicit interaction design strategies are lacking in the reviewed studies. The proposed conceptual framework and potential solutions can serve as guidelines for designing strategies aimed at habit formation within DBCIs.

(J Med Internet Res 2024;26:e54375) doi:10.2196/54375

KEYWORDS

habit formation; digital health; digital behavior change interventions design; behavior change techniques; physical activity; mobile phone

XSL•FO
RenderX

Introduction

Background

Physical activity has been proved to play an essential role in lifestyle management [1]. Incorporating regular physical activity into daily routines to form habitual healthy behavior is an effective strategy for reducing risks of chronic diseases (eg, hypertension, stroke, or diabetes) and maintaining a healthy lifestyle [1]. Common types of physical activities in daily life include walking, bicycling, hiking, swimming, and participating in various sports. According to the guidelines of the World Health Organization and the American College of Cardiology [1,2], adopting 90 to 150 minutes of aerobic exercise during a week is recommended. However, >1.4 billion adults do not get sufficient physical activity, particularly in high-income countries [3].

Digital behavior change interventions (DBCIs) provide potential solutions to increase physical activity levels and maintain healthy lifestyle habits. With the development of ubiquitous computing and Internet of Things (IoT), smartphones are gradually becoming the center of mobile devices. Research on mobile health or eHealth has developed rapidly after 2010, whereas mobile apps are a commonly used medium in DBCIs [4]. Various sensors are embedded in smartphones or wearable devices, which have the ability to track a large amount of personal data and perceive changes in surrounding environments, such as routines, preferences of individuals, and context information [5-7]. Compared with traditional interventions, the real-time health status of patients could be captured and analyzed in real time because of the ability of big data [6,8]. Meanwhile, integrating artificial intelligence (AI) into health systems could benefit from predicting, identifying, and providing personalized diagnoses [9]. Therefore, research on technology-mediated behavior change interventions has gradually evolved from providing one-size-fits-all solutions to patient-centered approaches, which meet personalized requirements.

Despite the significant progress in DBCIs, prior studies have shown the challenges in maintaining long-term adoption [10,11]. On the one hand, although DBCI solutions have widely applied automatic tracking strategies, the interpretation and feedback of tracking data do not satisfy users' expectations [10,12,13]. Users may stop adopting DBCI solutions because they perceive uselessness. On the other hand, most DBCI solutions depend on users' initiative and consciousness to drive goals, which requires users' effort and frequent proactive interaction with devices [11,14]. However, changes in daily routine may result in additional workflow for tracking. The ability of mobile phones and smart wearable devices to implicitly sense surroundings to infer users' intentions remains underused. Users may discontinue because of the loss of motivation for long-term tracking.

Prior Studies of DBCIs for Habit Formation

To engage in long-term tracking of physical activity, habit formation is argued to play an essential role in the health behavior change domain [15,16]. Some research emphasized that habits develop from nonconscious and implicit associations between context and behavior [15,17]. In addition, behavior is not only facilitated by habits but also driven by consciousness [16,18], which corresponds with the dual process theory [19] that behavior results from both conscious and unconscious. Prior studies proposed the theoretical process of habit formation [15,16,20]. Three key elements were summarized to form habits: implementation intention, positive reinforcement, and cues [16,20,21]. Behavior initially acts on consciousness in reflective processes, which drive intention through goals and planning. Regarding maintaining long-term motivation to form habits, cues serve as triggers that are repeated to form associative memories and habitual execution. Positive reinforcements refer to internal and external rewards that are used to accelerate the process of cue-behavior association.

Regarding the integration of theoretical models and design practices in DBCIs, studies from human-computer interaction (HCI) focused on using psychological theories and capabilities of technology to design solutions for human behavior change. For instance, Pinder et al [20] proposed the Habit Alteration Model and a set of design guidelines to illustrate how reflective and automatic processes generate and influence habitual behavior. Meanwhile, prior studies have adopted various behavior change techniques (BCTs) in DBCIs for physical activity. BCTs refer to a series of replicable interventions with the goal of changing behavior [22], which have been widely adopted in DBCIs. Duff et al [23] conducted a systematic review to access the BCTs' application for chronic disease interventions, which demonstrated that the description of the BCTs' application was vague. Studies either fail to mention what type of BCTs were used in DBCIs' design process or link the BCTs types to the functionality of DBCIs. Nyman et al [24] summarized that goal setting, social support, and credible source have potential effects for improving physical activity levels among patients with dementia, although there is no evidence of the effectiveness of BCTs for long-term behavior change. In addition, Direito et al [25] argued that existing DBCIs had minor effects on physical activity for adults through a meta-analysis, which was potentially caused by some commonly applied BCTs only promoting behavior change in the short term instead of behavior maintenance. However, despite prior studies that have demonstrated the effects of DBCIs on behavior change in a short time, studies from the HCI domain emphasized that most existing DBCI solutions still rely on users' intention and proactive interaction with digital devices [20,26]. On the one hand, Adams et al [26] found that 94% of published DBCIs focused on goal-directed behaviors that aim to promote reflective thinking and motivate behavior change intention, which demands users' extra attention and effort to engage with technologies. Therefore, the unconscious behavior of users is induced by the subtle stimulus of the surrounding environment, and the automated process is ignored. On the other hand, limited studies have investigated how to design DBCI solutions using characteristics of these technologies, such as context awareness and inferring personal requirements [20,27]. According to a study analyzing apps from the Apple App Store, it has been found that the most common design patterns of DBCIs are setting reminders and determining goals of activity [28], which ignores the characteristics of these technologies in implicitly sensing the users' intention.

XSL•FO
RenderX

Objectives

To enhance understanding of the potential of technology and explore future possibilities in designing DBCIs to facilitate long-term adoption, this study aims to investigate the design strategies of habit formation in existing DBCIs for physical activity from the perspective of designing DBCI solutions. To tackle these objectives, a systematic review was conducted to summarize existing studies based on 2 main research questions (RQs):

- RQ1: What types of techniques are applied in DBCIs to facilitate habit formation?
- RQ2: What design strategies leveraging the capabilities and features of digital devices have been investigated in DBCIs for habit formation?

In this paper, habit refers to consistent and long-term performance of healthy behaviors in daily life. Habit formation techniques encompass design strategies and functional implementation solutions that use BCTs and characteristics of technology. As a result, this study contributes in the following ways:

1. Summarizing 32 commonly used habit formation techniques in existing DBCIs.
2. Proposing a conceptual framework that formulates these habit formation techniques from the perspectives of target-mediated and technology-mediated interactions.
3. Identifying the challenges faced by designing DBCIs in the habit formation process and suggesting potential solutions.

Methods

Overview

To investigate the limitations of existing studies and opportunities of DBCIs for long-term habit formation, specifically implementations of BCTs, a systematic review was conducted that followed the PRISMA (Preferred Reporting Item for Systematic Reviews and Meta-Analyses) guidelines [29].

Search Strategy

Web of Science, Scopus, ACM Digital Library, and PubMed were selected and searched for relevant studies published from 2012 to 2022. The search strategy combined concepts for health behavior type (physical activity), technology implementation, and behavior change to form habits. Hence, the following search terms were used: (physical activit* OR step* OR fitness OR exercise) AND (mobile* OR smart* OR wearable* OR digital* OR IoT OR internet of things) AND (behavior change OR habit* OR form habit OR formation of habits).

Selection Criteria

Articles were selected if they met the following inclusion criteria: (1) the publication is focused on behavior change interventions or habit formation for physical activity, (2) the publication should use digital tools (eg, smartphones, wearables, and IoT devices) for assisting behavior change, (3) the publication should include the entire process of study design and validation results, (4) the aim of the publication should be to use DBCI for personal health management, and (5) the publication is journal articles or conference papers in English. Studies were excluded if they were reviews, magazines, lecture notes, abstracts, books, and other situations that did not meet the inclusion criteria.

Data Extraction and Analysis

All initial search results were scanned with title and abstract by the first author and 2 research assistants, followed by a group meeting to discuss the disagreements. To gain a comprehensive understanding of the current state of DBCI for habit formation, the fourth, fifth, and sixth authors discussed and determined the following information to be extracted from selected articles:

- BCTs: adopted BCTs were summarized based on selected articles.
- Habit formation techniques: information was concluded based on 3 elements for forming habits, which were goal-directed behavior, positive reinforcement, and cues.
- Validation: information about the objectives of the selected articles was extracted.
- Outcome-related information: outcome measures and study duration were extracted.

The coding of the extracted information was conducted independently by the first 2 authors by using thematic analysis [30]. Subsequently, a group meeting involving the first 2 authors and the third author was held to discuss and review the extracted information. Any disagreements of coding were discussed within the group, and final decisions were made by the third author. The final coding was categorized and summarized into descriptive themes of DBCIs' implementations, which were reviewed by the senior researchers.

Results

Search Results

Figure 1 shows the entire selection process. A total of 3402 articles were identified from databases, of which 1198 (35.21%) duplicates were removed, leaving 2204 (64.79%) studies for title and abstract screening. Subsequently, of these 2204 studies, 311 (14.11%) were reviewed in full text, of which 270 (86.8%) were excluded with reasons, and 41 (13.2%) eligible publications were included.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Item for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

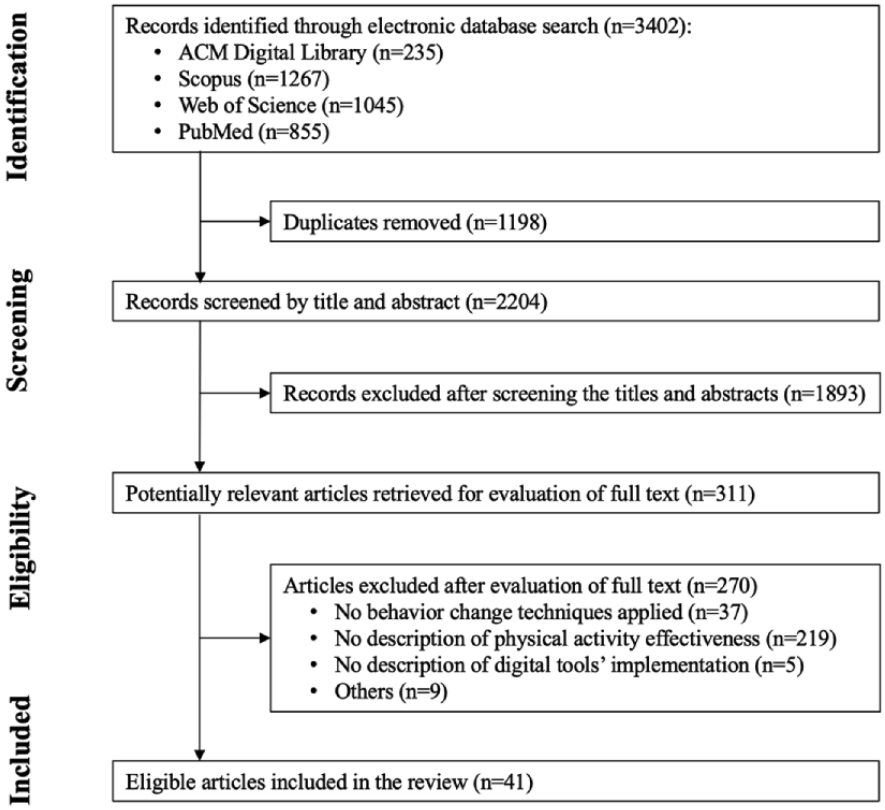

Frequency for Adopting BCTs

The most frequently applied BCT group used in the selected studies was feedback and monitoring (36/41, 88%), whereas the second and third most applied groups were associations (33/41, 80%) and goals and planning (30/41, 73%). This was followed by social support (27/41, 65%), reward and threat (27/41, 65%), comparison of behavior (22/41, 53%), natural consequences (11/41, 27%), repetition and substitution (11/41, 27%), and shaping knowledge (10/41, 24%). The 5 BCT groups that were less adopted in the selected studies were identity, comparison of outcomes, antecedents, scheduled consequences, and self-belief, with 27% (11/41) of the articles.

The frequency of adopting BCTs across the 41 studies was summarized. Three techniques were adopted in >60% (25/41) of the studies: prompts and cues (33/41, 80%), goal setting (27/41, 65%), and self-monitoring of behavior (25/41, 60%). In addition, 5 BCTs were applied in >30% (13/41) of the studies: feedback on behavior (23/41, 56%), social support (unspecified; 21/41, 51%), social comparisons (18/41, 43%), review outcome goals (16/41, 39%), and nonspecific rewards (16/41, 39%).

Validation of the Existing Studies

To visually represent the findings regarding the design objectives and strategies of the selected studies, we present Figures 2 and 3, which are inspired by Bae et al [31]. Figure 2 [32-72] summarizes the study duration, outcome measures, and validation information across 41 studies. Intervention duration ranged from 2 weeks to >2 years. Most articles (36/41, 88%) focused on short-term intervention, with 23 (56%) studies being <3 months and 13 (32%) studies being <1 month. A total of 5 studies focused on medium-term intervention (3-6 months), and only 1 study focused on long-term intervention (>6 months).

XSL•FO
RenderX

Figure 2. Summary of the study duration, outcome measures, and validation.

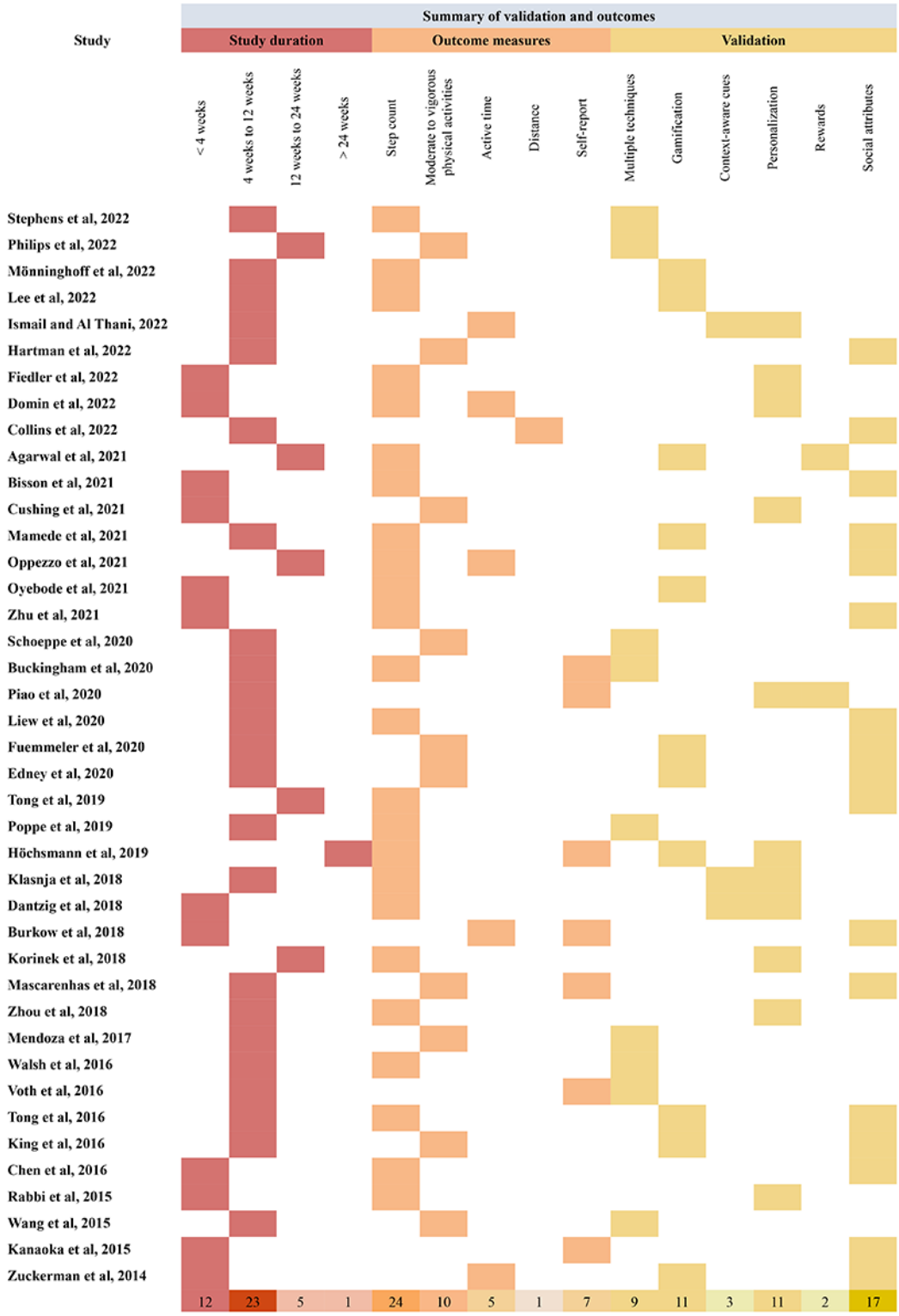

Regarding outcome measures, step count was the commonly applied parameter to evaluate the effectiveness of physical activity (26/41, 63%). This was followed by moderate to vigorous physical activities (10/41, 24%). In addition, most studies (34/41, 83%) automatically collected physical activity data from wearables and smartphones, whereas 17% (7/41) of the studies collected subjective data by self-repot.

Themes of validation were analyzed and summarized based on the objectives of the selected studies. A total of 9 articles combined and evaluated multiple BCTs. Besides, social influence was the most evaluated strategy across the selected studies (17/41, 41%), which refers to sharing behavior, teamwork, social support, and social comparison. The gamification theme refers to studies that validate gamified elements (eg, avatars, ranking, levels, and virtual rewards), which was second evaluated strategy in the selected articles (11/41, 27%). Similar to gamification, 11 (27%) articles validated the personalization strategy, including personalized

XSL•FO
RenderX

goals, tailored content based on individuals' performance and goals, and just-in-time reminders. In addition, only 3 (7%) studies investigated the effectiveness of context-aware cues, which refers to strategies that used context data to provide tailored interventions. A total of 2 (5%) articles focused on validating internal and external rewards effectiveness.

Design Strategies of Habit Formation Techniques

Overview

The codebook of habit formation techniques was based on the BCTs [22] and categorized them into 3 labels: implementations to facilitate the intention, cues, and positive reinforcements. The following paragraphs describe each technique. Figure 3 [32-72] presents the summary of the 3 labels of habit formation in the selected studies.

Figure 3. Summary of the design strategies of habit formation techniques.

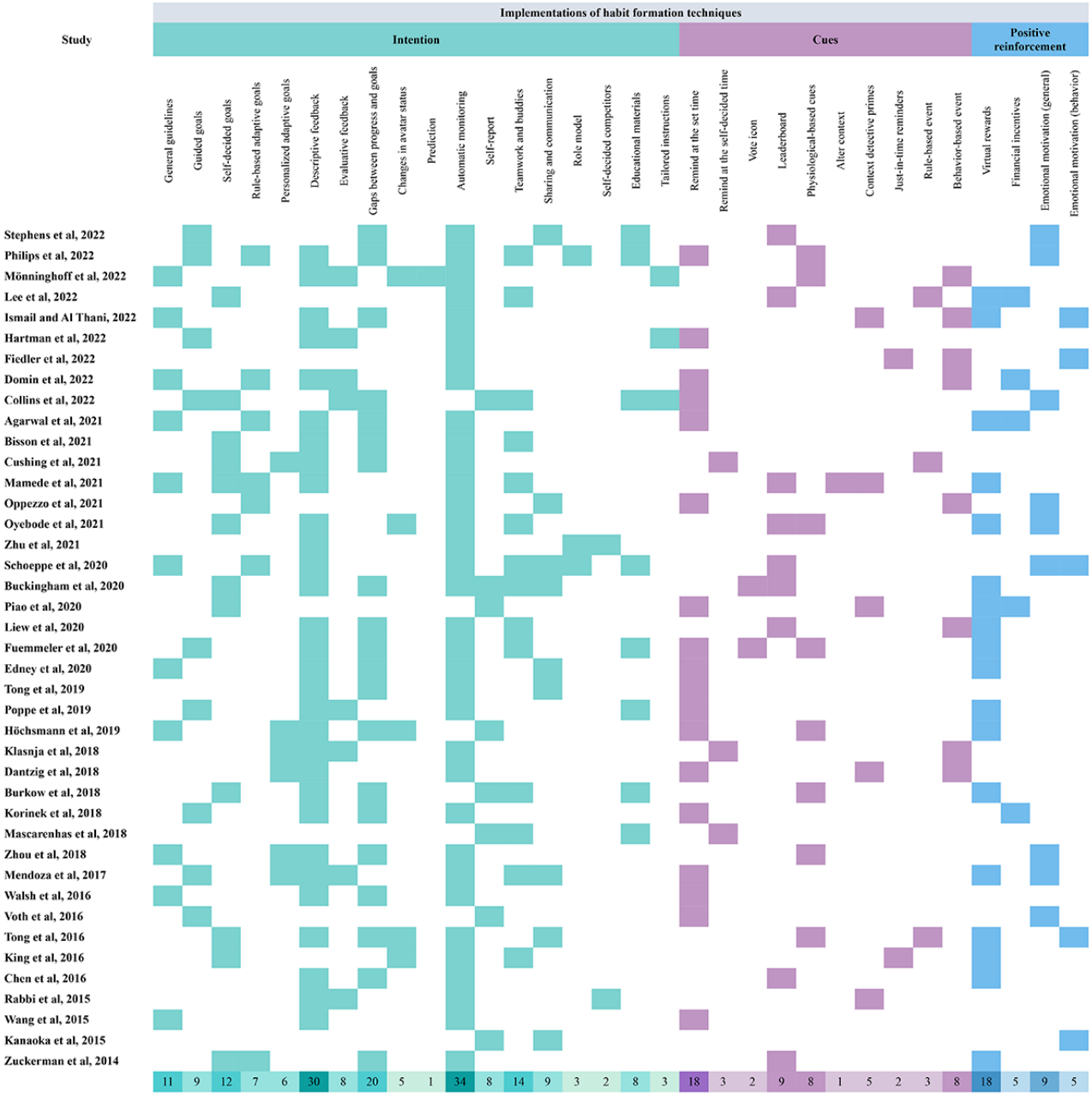

Implementations to Facilitate the Intention

These implementations aim to provide strategies that enhance reflective thinking and motivate intention to change behavior, which includes goal-directed behavior, feedback, self-monitoring, social influence, and shaping knowledge.

Goal-directed behavior was coded according to the performers and behavior status. Static goals refer to the fixed objectives that remain unchanged throughout interventions, which include *general guidelines* (eg, walking 10,000 steps a day), *self-decided goals*, and *goals guided by coaches or credible sources*. Most studies (27/41, 66%) adopted static goals across the selected articles (Figure 3). In specific, *self-decided goals* were influenced by information sources, which were divided into

XSL•FO
RenderX

self-decided goals based on system-provided plans [32], self-decided goals based on recommendations [33], and allowing to adjust goals during the intervention [34,35]. On the contrary, goals that adapt to changes based on individuals' behavior change and context information are adaptive goals. It is categorized into 2 types across the selected studies: *rule-based adaptive goals* and *personalized adaptive goals*. A total of 6 articles proposed *rule-based adaptive goals*, which implemented strategies to unlock more complex challenges [36] and goals increased weekly based on set rules [36-40]. In addition, 5 articles applied *personalized adaptive goals*, including human-guided adjustment of goals [41], tailored recommendations of goal setting generated by past behavior and context data [34,42,43], and tailored goals provided by the coaching system [44].

Other strategies also contribute to motivating intentions. First, *feedback and self-monitoring* were the most frequently adopted BCTs across 41 studies. Feedback contains various representations. *Descriptive feedback* refers to information displayed related to individual behavior data that are commonly presented as numerical or diagrams [39,45]. On the contrary, *evaluative feedback* provides comments about individuals' behavior and goal performances [46,47]. *Gaps between progress and goals* provide explicit information to motivate individuals [48-50]. For instance, Korinek et al [51] used a green bar chart to represent current step counts, and the daily goal was shown as a red line that users directly perceived the gap to the target. Moreover, *changes in avatar status* were commonly applied in gamified systems that linked avatar's appearances to individual behavior consequences to enhance reflective thinking [46,52,53]. Meanwhile, the *prediction* generated based on analyzing past behavior data provides the consequences of behavior to enhance awareness of maintaining health-enhancing behaviors or reducing health-impairing behaviors [54]. Regarding self-monitoring, most studies *automatically monitor* individual behaviors through smartphones or wearables, whereas some recorded physical activity by *self-report*, such as daily dairy [55-59].

Second, social support and comparison strategies boost behavior change by encouraging and competing with others. *Teamwork and buddies* motivated engagement in physical activity through support from relatives, friends, or team members with consistent goals [59-61]. *Sharing and communication* in the social network enhance behavior change through others' performances and stories [41,62,63]. Meanwhile, an individual performance develops as a *role model* that influences others' behavior, especially among family members [37]. In addition, allowing users to choose their competitors (*self-decided competitors*) generated by systems was argued to potentially motivate physical activity levels by comparing with those who performed better than themselves [64].

Finally, information related to shaping knowledge was applied in the selected studies. Health behavior knowledge is shaped through *educational materials* (eg, news) and *tailored instructions* [35-37].

Cues Implementations

According to prior studies, cues were classified into 5 categories in the selected articles [20,26]. The most applied strategies were time-based cues (22/41, 54%), such as push notifications at the set time (*remind at the set time*) or *remind at the self-decided time*. The second most applied strategy was social-based cues (10/41, 24%), which refers to cues that stimulate action based on others' support, the performance of team members, and comparison with others. For instance, *vote icons* (eg, "likes" or "thumbs-up") were applied to encourage others [65,66]. Displaying others' information (eg, profiles, status, or progress) was used to boost social comparison for physical activity [64]. Visualization of team members' performance (eg, *leaderboard*) and team goals were used to boost motivation for walking [39]. However, some instinctive and natural cues, such as physiological-based and context-aware cues, were rarely adopted. *Physiological-based cues* refer to priming by physiological signals and senses that already exist in humans, such as heart rate, smell, and sight. A total of 7 articles applied *physiological-based cues*, especially sight-based strategies. One strategy designed to develop cue-behavior associations involved mapping behavior consequences with virtual avatars. For example, a humanoid avatar was used to represent future health consequences based on current physical activity and nutrition, increasing users' awareness and motivating behavior change [53]. Another sight-based strategy used nudges to prompt physical activity by *altering context*. Mamede et al [39] used physical table signs in office spaces as motivational nudges to remind and boost physical activity levels and reduce sedentary behavior. In addition, *context detective primes* were argued to provide effective and personalized intervention by integrating context data [43,44,67]. Location, weather, and behavioral event are commonly context detective data that aim to deliver *just-in-time reminder* or recommendation for individuals. Despite context detection potentially prompting physical activity, limited studies focused on investigating the long-term effectiveness of context-aware cues [44,67]. Moreover, 2 articles set specific events by identifying users' behavior to prompt event-based cues, such as push notifications when users reach their goals or finish their lunch (*rule-based event*) [40,44]. Moreover, event-based cues were tailored solutions based on individuals' behavior (*behavior-based event*) instead of prompting reminders by setting a specific time.

Positive Reinforcement Implementations

Finally, providing internal and external rewards was the commonly adopted strategy for positive reinforcement. *Virtual rewards* (eg, points, badges, and trophies) were the most applied external rewards across the selected studies (16/41, 39%), which commonly used gamified elements to enhance engagement and motivation for physical activity in DBCIs [46,61,68-70]. On the contrary, although *financial incentives* were demonstrated to be more effective as external rewards compared to virtual rewards, their effects were argued to lack long-term sustainability and diminish over time [38,47]. In addition, *emotional motivation* was commonly applied as internal rewards across the selected studies (12/41, 29%). For instance, motivational text messages were provided when users reached the goal by a digital coach (chatbot) [33], whereas some

XSL•FO
RenderX

motivational messages that integrated behavior change theories provided general content to boost emotional motivation [35]. Emotional encouragement (eg, verbal thanks, high five, or messages) was used to reinforce behavior change [40,72].

Research Framework

Overview

To better understand and investigate how to leverage the capabilities of digital devices in DBCIs for habit formation, a research framework was developed and defined based on two levels: (1) the target degree, which focuses on the target of techniques; and (2) the technology-mediated interactions, which focus on the characteristics of these technologies in implicitly sensing.

Personalization Versus Generalization

DBCIs have gradually developed from providing generalized solutions to personalized interventions for individuals or a category of individuals. Theoretically, personalization represents different conceptualizations based on the research domains. According to the study by Fan and Poole [73], cognitive science defined personalization as "the understanding of the user, the user's tasks, and the context in which the user accomplishes tasks and goals." In the computer science domain, personalization is "...support individualized information inflow and outflow or presenting customers with services that are relevant to their current locations, activities, and surrounding environments."

Hence, in the context of DBCIs, the personalization-generalization dimension refers to the target population of BCTs, specific individuals, and a group of people who share common features or general populations. Meanwhile, personalization means those strategies that collect personal data to potentially impact individual behavior or affect the decision-making of changing behavior.

Implicitness Versus Explicitness

Another dimension is technology-mediated interactions. On the one hand, behavior is partly influenced by explicit BCTs, which prompt intention. At the same time, internal cues (eg, emotion, mood, belief, and attitudes) and the surrounding environment subtly influence and guide the association between cues and behavior without consciousness [74], which highlights the importance of the unconscious process and implicit cues. On the other hand, intelligent systems emphasize the ability to sense the environment and analyze the abundance of data for inference and prediction, which also highlights the importance of implicitness during the interaction between humans and intelligent systems.

Ju and Leifer [75] argued that traditional HCI was limited to command-based and graphical interface–based explicit interactions, whereas implicit interactions, defined as "those that occur without the explicit behest or awareness of the user," are common in day-to-day interactions. Developing on this, Serim and Jacucci [76] described implicit interaction as "...a system response to the user input does not rely on the user having conducted the input to intentionally achieve it." With the development of ubiquitous computing, big data, and AI, implicit interactions increasingly influenced users to interact with the intelligent system because these emerging technologies have the ability to sense the environment that does not rely on intentional input from users. Hence, in this context, implicitness refers to unintentionally taking action, reducing attention during input or output, action without awareness, and executing results without conscious processing [76]. The implicit-explicit interaction dimension refers to whether the design of BCTs requires intentional goals, users' attention, awareness of input, and conscious processing.

Framework of the DBCIs for Habit Formation

Overview

Therefore, we categorized habit formation techniques based on the following criteria:

- *Generalization*: the technique does not rely on user input and affects the general population.
- *Personalization*: the technique collects personal data that potentially influence the behavior or impact making behavior change decisions.
- *Explicitness*: the technique relates to intentionally achieving or avoiding a specific effect; the input and output processes demand attention and initiative to manipulate, and users are aware of the action outcome during the interaction.
- *Implicitness*: the technique does not rely on the user's intention to achieve a specific effect; the input or output processes reduce user attention, the action that uses physiological signals without awareness, and execute results without conscious processing during the interaction.

Habit formation techniques were positioned along the framework in 4 categories: *generalization-explicitness* (GE), *generalization-implicitness* (GI), *personalization-explicitness* (PE), and *personalization-implicitness* (PI). The frequency of habit formation techniques' application in each category is summarized in Figure 4. The detailed descriptions of each category are explained in the following subsections and summarized in Table 1.

XSL•FO
RenderX

Figure 4. Frequency of habit formation techniques in 4 categories. GE: generalization-explicitness; GI: generalization-implicitness; PE: personalization-explicitness; PI: personalization-implicitness.

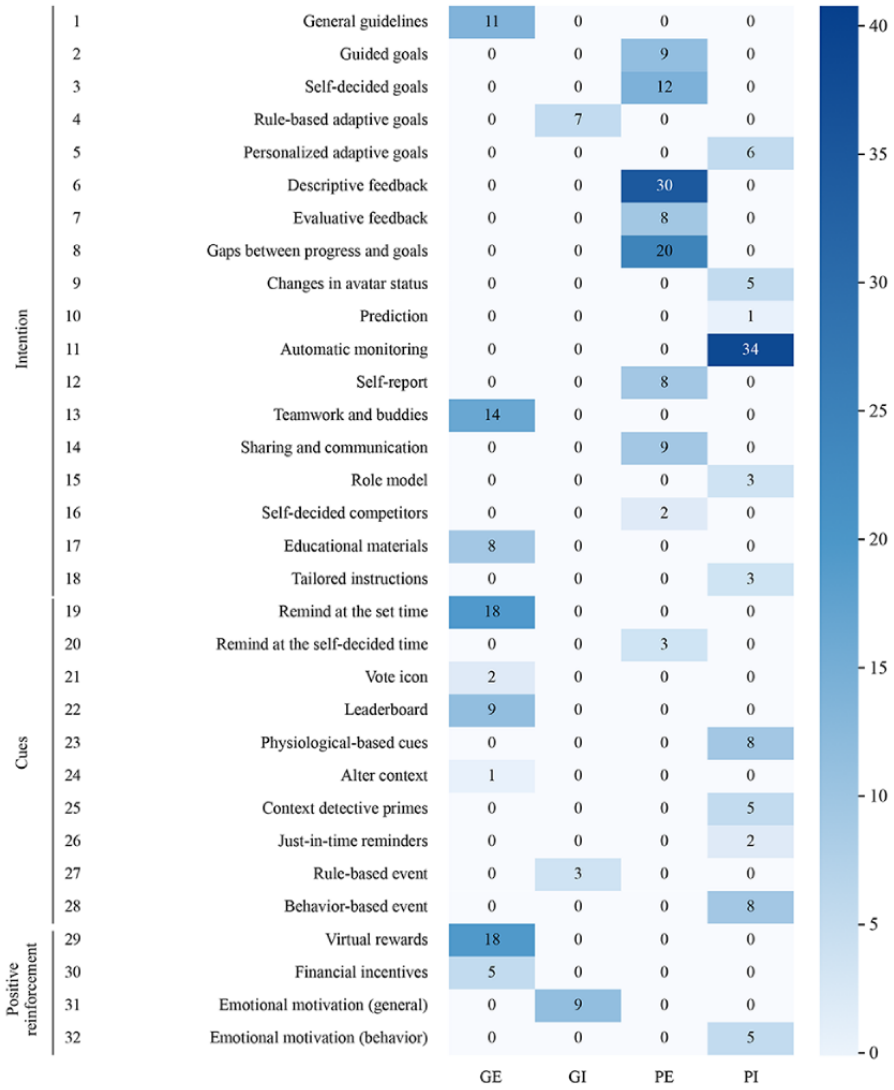

Table 1. Summary of the strengths and weaknesses of each category.

Category	Pros	Cons
Generalization-explicitness	• Well-proven effectiveness • Broad applicability	• It cannot be tailored based on individuals' situations and preferences • The effectiveness of the intervention cannot be maintained in the long term
Personalization-explicitness	• Increase implicit belief • Focus on background activities	• It cannot be adjusted based on individual situations
Generalization-implicitness	• Initiative • Reflection	• It ignores the potential impact from subtle cues that users may not be consciously perceive • Inaccurate data • The effectiveness of intervention cannot be maintained in the long term
Personalization-implicitness	• Unawareness • Context detection	• Not all health-enhancing behaviors are easily captured • Hard to define and identify cues in daily life • The effectiveness is questioned • Ethic issues

XSL•FO
RenderX

GE Category

The characteristics of the GE category are (1) not relying on the user input information, (2) effect works on the general population, and (3) the user intends to achieve goals through applying techniques. A total of 9 techniques are summarized into this category: goals with *general guidelines*, exercise with team members or fitness buddies (*teamwork and buddies*), *educational materials*, push notifications at the set time (*remind at the set time*), *altering context*, *leaderboard* display, voting for others (*vote icon*), and *virtual rewards* or *financial incentives*.

The advantages of GE techniques include well-proven effectiveness and broad applicability. On the one hand, a large number of studies have assessed the effectiveness of goal setting, social support by relatives and friends, notifications, and positive reinforcement for a long time [77-80]. On the other hand, several studies have assessed DBCIs across diverse target groups, demonstrating their broad applicability, including studies involving adults with hypertension [81], individuals with diabetes [82], older adults [83], and children with obesity [84]. Meanwhile, general guidelines serve as credible sources that are widely applied to interventions. Authorities (eg, the World Health Organization, American College of Cardiology, and Centers for Disease Control and Prevention) have published various reports on nonpharmacological intervention approaches to instruct patients with chronic diseases, physicians, and general populations, such as regular physical activity, reduce sodium intake, and weight control [2,85].

However, broad applicability means that the GE techniques cannot be tailored based on individuals' situations and preferences, while the effectiveness of intervention cannot be maintained in the long term [86]. According to the transtheoretical model, the individual performs different attitudes, skills, and abilities at different stages of the behavior change process, and intervention techniques should be adjusted based on the stage [87]. Therefore, tailoring is considered a vital technique for behavior change interventions [88]. In addition, personal interest and intrinsic motivation are argued to be beneficial to forming habits [89], which also emphasizes the importance of personalization.

GI Category

The category of GI contains the following attributes: (1) not relying on the user input information; (2) effect works on the general population; and (3) results in execution do not rely on users' intention, awareness, or consciousness. Only 3 techniques are included in this category, which are *rule-based adaptive goals*, *rule-based events*, and e*motional motivation (general)*.

The techniques in the GI category contribute to increasing implicit belief. According to Hagger [74], the development of implicit beliefs, attitudes, and affective reactions potentially enhances habit formation. Hence, techniques such as graded tasks are applied to motivate positive belief. Graded tasks refer to increasingly difficult but achievable small tasks that are divided by long-term goals [22]. Despite the difficulty of the tasks gradually increasing, a stronger belief (eg, "I can do it!") is formed when the user achieves the tasks. Meanwhile, some studies provide general motivational messages to enhance the development of belief by applying the Health Belief Model [90]. However, high cognition and awareness are required to develop and reinforce the belief, which is challenged by individual capabilities. Hence, the GI techniques assist in reducing the requirements of user capabilities by focusing on background activities. Background activities are defined as "interactions that elude the user's attention" [75]. The rule-based techniques are established to achieve behavior changes that do not rely on the user's intentionality or initiative. In other words, the rule-based techniques are developed according to the effect, which is not limited to individual cognitive processes. For instance, the next level of challenges was unlocked after achieving the event (eg, accomplishing previous levels' tasks) in gamified interventions [36]. Goals were adjusted weekly, increasing in difficulty by 10% according to predetermined rules [36-40].

However, rule-based techniques still cannot be adjusted based on individual situations. Although the GI techniques try to provide event-based influences, the individual behavior change process is not static and consistent, impacting the context. Meanwhile, the development of belief is also a highly individual process. Hence, generalized solutions have limited effects.

PE Category

The category of PE contains the following capabilities: (1) tracking personal data, (2) effect generated according to individual input, (3) techniques assist users in achieving their goals, and (4) outcomes aim to enhance reflection. A total of 9 techniques belong to the PE category, including *guided goals*, *self-decided goals*, *descriptive feedback*, *evaluative feedback*, *gaps between progress and goals*, *self-report*, *sharing and communication*, *self-decided competitors*, and push notifications at the self-decided time (*remind at the self-decided time*).

Initiative and reflection are strengths of the PE techniques. On the one hand, the user initiates behavior with high intentionality. Users actively make decisions, and effects are decided by what users intend to achieve. Hence, the PE techniques are applied to tailor for individuals. For example, self-decided goals provide autonomy that users can set or adjust the detailed plans according to their current situation, which is perceived as more persuasive by users [91]. On the other hand, personal information intends to help users' self-reflection to take action [92]. Explicit feedback is generated based on individual input that draws the individual's attention, which aims to reflect individual historical behavior to facilitate self-reflection.

However, the PE techniques rely too much on individual initiative and intentionality, which ignores the potential influence of subtle cues that users may not consciously perceive. Self-report data are argued to be inaccurate and influenced by individual cognitive bias [20]. Moreover, displaying individual information is argued to be useless for individual and cannot maintain long-term effectiveness [11]. Barriers such as inaccurate data, unclear feedback, and insufficient explanation are reasons for abandoned monitoring [93].

XSL•FO
RenderX

PI Category

The characteristics of the PI category are (1) tracking personal data; (2) effect generated according to individual input; (3) action without awareness; and (4) results in execution do not rely on users' intention, awareness, or consciousness. A total of 11 techniques are summarized in the PI category, which includes *personalized adaptive goals*, *changes in avatar status*, *prediction*, *automatic monitoring*, *role model*, *tailored instructions*, *behavior-based events*, *just-in-time reminders*, *context detective primes*, *physiological-based cues*, and *emotional motivation (behavior)*.

Unawareness and context detection are advantages of the PI category. Unawareness occurs during the whole process of DBCIs. First, users may not be aware of what health data will be collected. Second, user behavior may be constantly monitored without awareness through wearables, cameras, or other IoT devices. Third, collected information not only includes individual behavior data but also contains context data without intentionality. Multiple dimensions of data are integrated to identify problems and patterns, which are prepared for generating meaningful effects for individuals. Then, users may not be aware of using physiological signals (eg, heart rate, gaze, and smell) when they take action. Finally, effects may not rely on the user's cognitive process. Since the development of IoT, big data, and AI, techniques in the PI category could assist in achieving the above unawareness process. For instance, various sensors are embedded in IoT devices that support constant monitoring of individual behavior, physiological signals, and context data. A large amount of data contributes to recognizing patterns with similar conditions and real-time analysis of mood and context. In addition, the PI techniques are good at identifying implicit and subtle context cues in daily life, which users easily ignore. Context detection is argued to play a vital role in habit formation [20,94]. Ideally, habitual behaviors are repeatedly triggered by suitable context cues that can be detected stably and constantly in daily life [15]. Hence, identifying suitable cues for individuals could enhance the development of cues-behavior association to respond at the right time.

However, PI techniques have faced various challenges since adopting emerging technologies. First, not all health-enhancing behaviors are easily captured. For instance, accelerometers in smartphones or wearables could automatically track physical activity, such as steps or sedentary behavior [70,72]. On the contrary, information on eating behavior, such as calorie intake, is more difficult to capture [56]. Second, it is hard to define and identify cues in daily life [20,95]. Location and time cues are easy to track, so existing DBCI studies focus on time-based cues. Other types of cues (eg, mood or physiology signals) are rare to explore in terms of how they develop a connection with behaviors. Third, although existing DBCIs attempt to provide just-in-time reminders to impact user behavior, their effectiveness has been questioned [67,94]. Finally, ethical issues must be considered because of the implicitness and unawareness of PI techniques. For instance, privacy and trust issues are worrisome challenges because the reasoning process of intelligent systems is obscured, and the status of personal health data is untraced [96,97].

Discussion

Principal Findings

Habit formation is an iterative and cyclical long-term process. To gain a deeper understanding of how to adopt habit formation techniques in designing DBCIs, this study examines techniques from the design perspective of user target–mediated and technology-mediated interactions. The identified habit formation techniques are categorized into 4 main dimensions: *generalization*, *personalization*, *explicitness*, and *implicitness*. As shown in Figure 5, techniques of the GE and PE category have been widely implemented in DBCI solutions, such as *descriptive feedback*, *teamwork and buddies*, *remind at the set time*, and *virtual rewards*. These techniques demonstrate more applicability that could be adopted by a wide range of populations. Meanwhile, GE and PE techniques partly replace the role of traditional interventions, which reduces the burden of knowledge requirement and is beneficial to the implication of self-management in a low-cost manner [98]. Furthermore, there has been a growing use of GI and PI technologies after 2016, particularly automatic monitoring techniques. The development of wearable devices has allowed for the collection of an increasing amount of behavioral and physiological data. Studies focus on using emerging technologies (IoT, big data, and AI) to detect subtle cues and reduce user effort and attention [7,9,26], which indicates the trend of using techniques of PI category. On the basis of the reviewed literature, we summarize the existing habit formation techniques into the following categories: habit formation process, challenges, and potential solutions (Table 2). The following subsections provide detailed discussion of each category.

XSL•FO
RenderX

Figure 5. Relationship between habit formation techniques (left), categories (middle), and years (right). The bar widths indicate frequencies of each item. GE: generalization-explicitness; GI: generalization-implicitness; PE: personalization-explicitness; PI: personalization-implicitness.

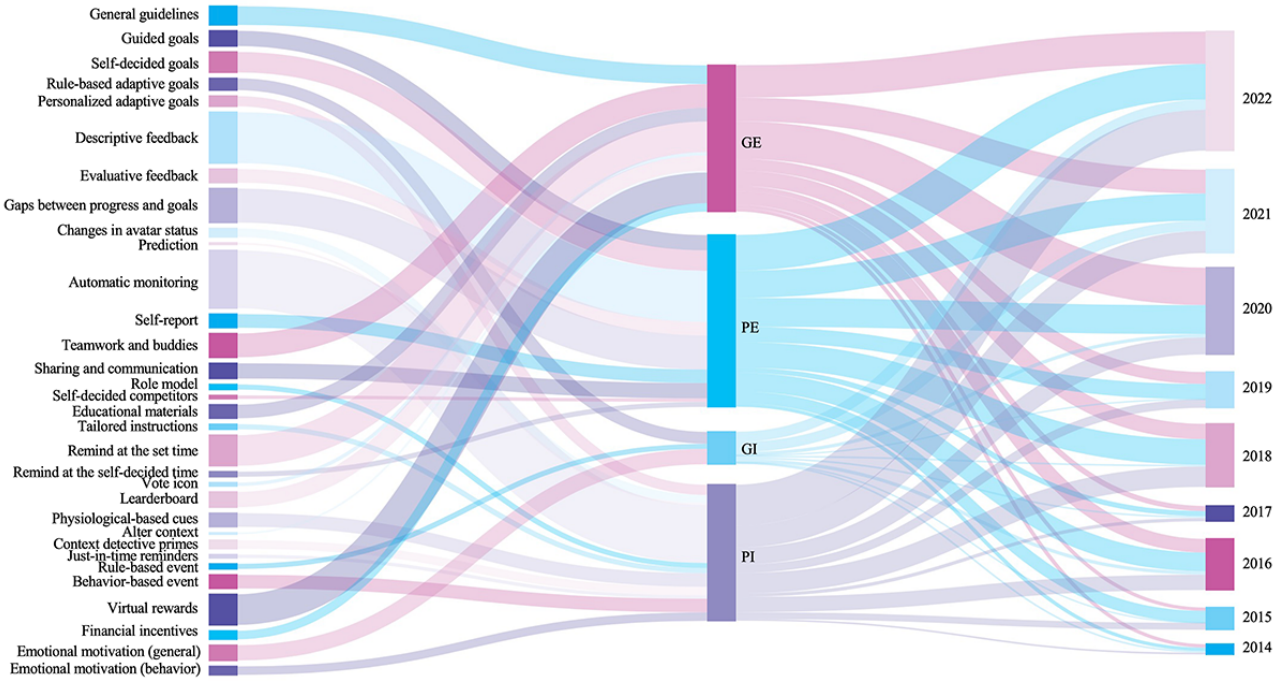

Table 2. Summary of techniques in the habit formation process, challenges, and potential solutions.

Habit formation	Challenges	Potential solutions
Intention and action control	• Most of the explicit techniques still rely on users' attention and frequently interact with the device, which demands high initiative	• Applying implicit techniques to reduce the demand for users' attention and motivation
	• Identifying and defining unexpected events	• Capturing contextualized data as historical records, which serve as a reference for identifying deviations from the norm or unusual data, enabling the inference of unexpected events
	• Providing coping planning to deal with unexpected events	• To achieve predictions and tailored recommendations, it is necessary to capture not only common data such as weather, location, time, and activity but also users' physiological data, emotions, social relationships, and other relevant data
Developing association between cues and behaviors	• Low user response to pushing notification	• Just-in-time adaptive interventions
	• The cue-behavior association is a fast, automatic, and subtle process	• Establishing a stable repetition of implicit cues in the context, such as incorporating other types of cues and leveraging preexisting associations
Providing personalized response	• Useful and meaningful data to enhance motivations and intentions to change behavior	• Identifying the specific data points that hold significance for users and effectively predict users' preferences is crucial for facilitating meaningful self-reflection and behavior change
Others	• Data privacy	• Developing effective techniques for tailored interventions and recommendations based on individuals' preferences • Addressing the need to present transparent information to raise awareness of data privacy
	• Perceived inaccuracy	• Improving sensor technology and processing algorithms to calibrate the accuracy of the tracking device • Transparency design methods that explain device abilities and limitations could enhance user trust and calibrate user expectations
	• Usability and comfort issues	• Ideally, users should feel physically comfortable wearing and intuitive to use the device

XSL•FO
RenderX

Forming Intention and Action Control

To form habits, people make decisions about goal achievement first. Subsequently, intentions are transferred into action through action planning [15,16,89]. Existing habit formation techniques, especially explicitness techniques (GE and PE), have been well-supported goal setting and action planning, such as *guided goals* and *self-decided goals*. Most of these explicit techniques still rely on users' attention and frequently interact with the device, which demands high initiative. Studies have applied implicit techniques to reduce the demand for users' attention and motivation. For instance, *rule-based adaptive goals* aim to automatically provide graded tasks to users without their awareness, which is based on time, progress, or users' performance.

However, the process of transferring intention to actions is influenced by unexpected events [99]. Unexpected events disrupt users' established routines, leading to an inability to consistently repeat behaviors in a stable environment. Few empirical studies have explored how to create and execute alternative goals and plans to deal with these unexpected events. One of the challenges is the identification and definition of unexpected events. Existing studies focus on capturing common contextualized data (eg, weather, location, temporal factors, and activity data) as historical records [99,100]. These records serve as a reference for identifying deviations from the norm or unusual data, enabling the inference of unexpected events. However, the influences of other implicit factors are disregarded in current studies, such as social activities and the chain reactions caused by unfinished tasks.

Another challenge is providing coping planning to deal with unexpected events. The concept of coping planning is proposed as a self-regulation method, which refers to identifying problems and obstacles that may impact planning and developing alternative plans to overcome these challenges [101]. Ideally, intelligent systems predict the point of unexpected events based on learning individual behavior and preferences. Meanwhile, multiple alternative plans could be suggested according to users' mental status and context information, such as suggesting alternative activity times or activity types that are suitable for the current context. To achieve predictions and tailored recommendations, it is necessary to capture not only common data such as weather, location, time, and activity but also users' physiological data, emotions, social relationships, and other relevant data.

Developing Cue-Behavior Associations

Regarding developing the association between cues and behavior, most of the existing studies are limited to static and fixed time-based cues (*remind at the set time*). This GE technique cannot adapt to individual situations to maintain long-term behavior change. Moreover, pushing notifications is facing the challenge of low user response [67]. Users might notice the notifications but may not respond; in other words, users may not act on reminders immediately. Integrating mobile sensing technology to deliver notifications at the right time based on personalized needs is argued to be one of the solutions. Prior studies have used the just-in-time adaptive interventions for various health behavior changes, such as weight management [71], dietary intake [56], and physical activity [67,102]. However, just-in-time adaptive interventions rely on factors such as type, time, and intensity, which ignores other context cues essential in developing cue-behavior associations for habit formation [20].

In addition, the cue-behavior association is a fast, automatic, and subtle process. Instead, *remind at the set time* is an explicit interaction that demands users' attention and initiative to respond, which is a slow and conscious goal-directed process. The development of association should pay more attention to instinctive signals and implicit cues [20,26]. It is crucial to explore potential solutions for establishing a stable repetition of implicit cues in the context. This can be achieved by incorporating other types of cues (eg, physiological signals, auditory responses, and visual cues) and leveraging preexisting associations (eg, red color associates with stop). By implementing these strategies, the effectiveness of implicit cues could be enhanced in influencing and guiding individuals' behaviors.

Personalized Response

Furthermore, personalized response plays a major role in DBCIs, which can stimulate users' self-reflection and motivation for changing behavior. As shown in Figure 4, studies have adopted various techniques to satisfy individual requirements, such as *descriptive feedback*, *gaps between progress and goals*, *just-in-time reminders*, and *role model*. However, most of the responses only present data through various types of visualized graphics [36-38,103,104]. For instance, similar types of user data are applied to facilitate individual reflection on their behavior [104]. Peaks and valleys are used to demonstrate the extreme value that may explain the reasons for their behavior. Visualizing the relationship between users' activity trajectory and the target progress could help enhance user awareness [105].

Users need to comprehend the meaning of these data and engage in self-reflection regarding their own actions. A major challenge lies in determining what data are useful and meaningful to individuals to enhance motivations and intentions to change behavior. Identifying the specific data points that hold significance for users and effectively predict users' preferences is crucial for facilitating meaningful self-reflection and behavior change. For instance, Zhu et al [64] proposed a model to personally recommend virtual fitness buddy profiles to users by capturing daily step data, which aimed to examine whether individuals reacted positively when they compared with those who performed better than themselves. However, their findings found the personalization paradox problem, which means that personalization models may inaccurately predict users' preferences because the model is developed based on daily steps. Various contexts and implicit information about individuals cannot be captured. Hence, individuals perceive the uselessness and inaccuracy of proposed solutions.

Other Challenges

Several research challenges should be taken into account in the future design of DBCIs. First, constant monitoring of the human body and surroundings is a critical feature of DBCIs. It raises concerns of data privacy violations, such as collecting data

XSL•FO
RenderX

without users' permission or unauthorized trading personal data with third parties. To address these privacy issues, regulations have been implemented to safeguard data privacy. In addition, studies focus on proposing transparency design principles to enhance users' awareness of data privacy [97]. Future studies should not only focus on developing effective techniques for tailored interventions and recommendations based on individuals' preferences but also address the need to present transparent information to raise awareness of data privacy. Second, the perception of inaccuracy in tracking devices is another challenge [11]. Users may question the reliability of the data collected by these devices, especially when the feedback from the tracking does not align with their subjective perceptions and expectations. Considering this discrepancy may lead to a loss of trust and confidence in the DBCIs, future studies should improve sensor technology and processing algorithms to calibrate the accuracy of the tracking device. Meanwhile, transparency design methods that explain device abilities and limitations could enhance user trust and calibrate user expectations. Finally, to achieve long-term behavior change to form habits, usability and comfort issues are important considerations to constantly adopting smart devices [11,14]. Ideally, users should find the device comfortable to wear and intuitive to use. Addressing these concerns will contribute to long-term engagement with these smart devices and enhance the effectiveness of DBCIs.

Limitations

Although we used the standard systematic review method to identify relevant articles, there are still limitations of the study, such as scope of the database and availability of data perspectives. As we aim to review the design implementations of DBCIs, it is possible to select the target databases from the HCI field (ACM Digital Library) and focus on the design process and techniques of interventions. Another limitation is that some articles might be missed because they did not include the validation processes. Meanwhile, some selected papers only had quantitative data in validation processes or conducted multiple techniques, which makes it hard to identify the effectiveness of implementations of DBCIs. Hence, future studies could also include the efficacy of a single technique or the combination of DBCIs in consideration.

Conclusions

This study presents a systematic review aimed at identifying the design strategies of habit formation techniques in existing DBCI studies for physical activity. The findings revealed the existence of gaps between behavior change theories, exploration of technology capabilities, and design strategies for long-term adoptions of DBCIs. By examining design approaches in current DBCIs for physical activity with a focus on design targets and technology capabilities, a conceptual framework was proposed to investigate the prospects for design strategies of habit formation.

The findings indicate that the most applied BCTs include self-monitoring of behavior, goal setting, and prompts and cues. Mapping to the design strategies of habit formation techniques in the conceptual framework, most studies used *automatic monitoring* and *descriptive feedback* techniques, which commonly apply various sensors in devices. Moreover, existing research emphasizes the use of explicit interaction in design strategies. For instance, *general guidelines* and *self-decided goals* were widely used to facilitate behavior intention. Time-based cues (*remind at the set time*) and *virtual rewards* were commonly applied cues and positive reinforcement. Meanwhile, there is a clear trend highlighting the significance of personalized habit formation techniques in DBCIs' design strategies. This not only necessitates exploring technological characteristics but also addresses the specific needs of individual users. However, implicit interaction design strategies of habit formation techniques are lacking in the reviewed studies. Potential challenges include difficulty in capturing all health-enhancing behaviors, defining and identifying various cues in daily life, and ethical considerations regarding the unawareness of automatic tracking behavior.

The outcome of the conceptual framework and potential solutions can serve as guidelines for further DBCIs' design strategies of habit formation. Future studies are suggested to investigate the implementations to facilitate the development of cue-behavior associations, especially those implicit interactions that adopt emerging technologies' features.

Acknowledgments

This research is funded by the Laboratory for Artificial Intelligence in Design (project code: RP2-4) under the InnoHK Research Clusters, Hong Kong Special Administrative Region Government. The work was substantially supported by the Projects of Strategic Importance (project ID: P0043542) and the Research Centre for Future (Caring) Mobility at The Hong Kong Polytechnic University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 33 KB - jmir_v26i1e54375_app1.docx]

References

XSL•FO
RenderX

1. World Health Organization. Adherence to long-term therapies: evidence for action. World Health Organization. 2003. URL: https://iris.who.int/handle/10665/42682 [accessed 2024-04-05]
2. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. J Am Coll Cardiol 2018 May 15;71(19):e127-e248 [FREE Full text] [doi: 10.1016/j.jacc.2017.11.006] [Medline: 29146535]
3. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018 Oct;6(10):e1077-e1086. [doi: 10.1016/s2214-109x(18)30357-7]
4. Taj F, Klein MC, van Halteren A. Digital health behavior change technology: bibliometric and scoping review of two decades of research. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
5. Röcker C, Ziefle M, Holzinger A. From computer innovation to human integration: current trends and challenges for pervasive healthtechnologies. In: Holzinger A, Ziefle M, Röcker C, editors. Pervasive Health: State-of-the-art and Beyond. London, UK: Springer; 2014:1-17.
6. Banerjee A, Chakraborty C, Kumar A, Biswas D. Emerging trends in IoT and big data analytics for biomedical and health care technologies. In: Balas VE, Solanki VK, Kumar R, Khari M, editors. Handbook of Data Science Approaches for Biomedical Engineering. Cham, Switzerland: Springer; 2020:121-152.
7. Tong HL, Quiroz JC, Kocaballi AB, Fat SC, Dao KP, Gehringer H, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021 Jul;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
8. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. IEEE J Biomed Health Inform 2015 Jul;19(4):1193-1208 [FREE Full text] [doi: 10.1109/JBHI.2015.2450362] [Medline: 26173222]
9. Kakhi K, Alizadehsani R, Kabir HM, Khosravi A, Nahavandi S, Acharya UR. The internet of medical things and artificial intelligence: trends, challenges, and opportunities. Biocybern Biomed Eng 2022 Jul;42(3):749-771. [doi: 10.1016/j.bbe.2022.05.008]
10. Clawson JR, Pater JA, Miller AD, Mynatt ED, Mamykina L. No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 647-658 URL: https://dl.acm.org/doi/10.1145/2750858.2807554 [doi: 10.1145/2750858.2807554]
11. Attig C, Franke T. Abandonment of personal quantification: a review and empirical study investigating reasons for wearable activity tracking attrition. Comput Human Behav 2020 Jan;102:223-237. [doi: 10.1016/j.chb.2019.08.025]
12. Lazar A, Koehler C, Tanenbaum TJ, Nguyen DH. Why we use and abandon smart devices. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 635-646 URL: https://dl.acm.org/doi/10.1145/2750858.2804288 [doi: 10.1145/2750858.2804288]
13. Gulotta R, Forlizzi J, Yang R, Newman MW. Fostering engagement with personal informatics systems. In: Proceedings of the 2016 ACM Conference on Designing Interactive Systems. 2016 Presented at: DIS '16; June 4-8, 2016; Brisbane, Australia p. 286-300 URL: https://dl.acm.org/doi/10.1145/2901790.2901803 [doi: 10.1145/2901790.2901803]
14. Coskun A. Design for long-term tracking: insights from a six-month field study exploring users' experiences with activity trackers. Des J 2019 Jul 09;22(5):665-686. [doi: 10.1080/14606925.2019.1634447]
15. Gardner B, Rebar AL. Habit formation and behavior change. In: Braddick O, editor. Oxford Research Encyclopedia of Psychology. Oxford, UK: Oxford University Press; 2019.
16. Hagger MS. Habit and physical activity: theoretical advances, practical implications, and agenda for future research. Psychol Sport Exerc 2019 May;42:118-129. [doi: 10.1016/j.psychsport.2018.12.007]
17. Wood W. Habit in personality and social psychology. Pers Soc Psychol Rev 2017 Nov 24;21(4):389-403. [doi: 10.1177/1088868317720362] [Medline: 28737111]
18. Gardner B, Rebar AL, Lally P. 'Habitually deciding' or 'habitually doing'? A response to Hagger (2019). Psychol Sport Exerc 2020 Mar;1(1):101539 [FREE Full text] [doi: 10.1016/j.psychsport.2019.05.008]
19. Evans JS. Dual-process theories of reasoning: contemporary issues and developmental applications. Dev Rev 2011 Sep;31(2-3):86-102. [doi: 10.1016/j.dr.2011.07.007]
20. Pinder C, Vermeulen J, Cowan BR, Beale R. Digital behaviour change interventions to break and form habits. ACM Trans Comput Hum Interact 2018 Jun 12;25(3):1-66 [FREE Full text] [doi: 10.1145/3196830]
21. Gardner B, Rebar AL, Lally P. Habit interventions. In: Hagger MS, Cameron LD, Hamilton K, Hankonen N, Lintunen T, editors. The Handbook of Behavior Change. Cambridge, UK: Cambridge University Press; 2020:599-616.
22. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

XSL•FO
RenderX

23. Duff OM, Walsh DM, Furlong BA, O'Connor NE, Moran KA, Woods CB. Behavior change techniques in physical activity eHealth interventions for people with cardiovascular disease: systematic review. J Med Internet Res 2017 Aug 02;19(8):e281 [FREE Full text] [doi: 10.2196/jmir.7782] [Medline: 28768610]
24. Nyman SR, Adamczewska N, Howlett N. Systematic review of behaviour change techniques to promote participation in physical activity among people with dementia. Br J Health Psychol 2018 Feb;23(1):148-170. [doi: 10.1111/bjhp.12279] [Medline: 28980370]
25. Direito A, Carraça E, Rawstorn J, Whittaker R, Maddison R. mHealth technologies to influence physical activity and sedentary behaviors: behavior change techniques, systematic review and meta-analysis of randomized controlled trials. Ann Behav Med 2017 Apr;51(2):226-239. [doi: 10.1007/s12160-016-9846-0] [Medline: 27757789]
26. Adams AT, Costa J, Jung MF, Choudhury T. Mindless computing: designing technologies to subtly influence behavior. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 719-730 URL: https://dl.acm.org/doi/10.1145/2750858.2805843 [doi: 10.1145/2750858.2805843]
27. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 03;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
28. Villalobos-Zúñiga G, Cherubini M. Apps that motivate: a taxonomy of app features based on self-determination theory. Int J Hum Comput Studies 2020 Aug;140:102449. [doi: 10.1016/j.ijhcs.2020.102449]
29. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Bae SS, Zheng C, West ME, Do EY, Huron S, Szafir DA. Making data tangible: a cross-disciplinary design space for data physicalization. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-18 URL: https://dl.acm.org/doi/abs/10.1145/3491102.3501939 [doi: 10.1145/3491102.3501939]
32. Bisson AN, Sorrentino V, Lachman ME. Walking and daily affect among sedentary older adults measured using the StepMATE app: pilot randomized controlled trial. JMIR Mhealth Uhealth 2021 Dec 01;9(12):e27208 [FREE Full text] [doi: 10.2196/27208] [Medline: 34855609]
33. Cushing CC, Bejarano CM, Ortega A, Sayre N, Fedele DA, Smyth JM. Adaptive mHealth intervention for adolescent physical activity promotion. J Pediatr Psychol 2021 Jun 03;46(5):536-546. [doi: 10.1093/jpepsy/jsaa125] [Medline: 33484137]
34. Höchsmann C, Infanger D, Klenk C, Königstein K, Walz SP, Schmidt-Trucksäss A. Effectiveness of a behavior change technique-based smartphone game to improve intrinsic motivation and physical activity adherence in patients with type 2 diabetes: randomized controlled trial. JMIR Serious Games 2019 Feb 13;7(1):e11444 [FREE Full text] [doi: 10.2196/11444] [Medline: 30758293]
35. Hartman SJ, Chen R, Tam RM, Narayan HK, Natarajan L, Liu L. Fitbit use and activity levels from intervention to 2 years after: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2022 Jun 30;10(6):e37086 [FREE Full text] [doi: 10.2196/37086] [Medline: 35771607]
36. Phillips SM, Penedo FJ, Collins LM, Solk P, Siddique J, Song J, et al. Optimization of a technology-supported physical activity promotion intervention for breast cancer survivors: results from Fit2Thrive. Cancer 2022 Mar 01;128(5):1122-1132 [FREE Full text] [doi: 10.1002/cncr.34012] [Medline: 34812521]
37. Schoeppe S, Salmon J, Williams SL, Power D, Alley S, Rebar AL, et al. Effects of an activity tracker and app intervention to increase physical activity in whole families-the step it up family feasibility study. Int J Environ Res Public Health 2020 Oct 20;17(20):7655 [FREE Full text] [doi: 10.3390/ijerph17207655] [Medline: 33092219]
38. Agarwal AK, Waddell KJ, Small DS, Evans C, Harrington TO, Djaraher R, et al. Effect of gamification with and without financial incentives to increase physical activity among veterans classified as having obesity or overweight: a randomized clinical trial. JAMA Netw Open 2021 Jul 01;4(7):e2116256 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.16256] [Medline: 34241628]
39. Mamede A, Noordzij G, Jongerling J, Snijders M, Schop-Etman A, Denktas S. Combining web-based gamification and physical nudges with an app (MoveMore) to promote walking breaks and reduce sedentary behavior of office workers: field study. J Med Internet Res 2021 Apr 12;23(4):e19875 [FREE Full text] [doi: 10.2196/19875] [Medline: 33843593]
40. Oppezzo MA, Tremmel JA, Kapphahn K, Desai M, Baiocchi M, Sanders M, et al. Feasibility, preliminary efficacy, and accessibility of a Twitter-based social support group vs Fitbit only to decrease sedentary behavior in women. Internet Interv 2021 Sep;25:100426 [FREE Full text] [doi: 10.1016/j.invent.2021.100426] [Medline: 34401385]
41. Mendoza JA, Baker KS, Moreno MA, Whitlock K, Abbey-Lambertz M, Waite A, et al. A Fitbit and Facebook mHealth intervention for promoting physical activity among adolescent and young adult childhood cancer survivors: a pilot study. Pediatr Blood Cancer 2017 Dec 15;64(12):e26660. [doi: 10.1002/pbc.26660] [Medline: 28618158]

XSL•FO
RenderX

42. Zhou M, Mintz Y, Fukuoka Y, Goldberg K, Flowers E, Kaminsky P, et al. Personalizing mobile fitness apps using reinforcement learning. CEUR Workshop Proc 2018 Mar 07;2068 [FREE Full text] [Medline: 32405286]
43. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 03;53(6):573-582 [FREE Full text] [doi: 10.1093/abm/kay067] [Medline: 30192907]
44. van Dantzig S, Bulut M, Krans M, van der Lans A, de Ruyter B. Enhancing physical activity through context-aware coaching. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: PervasiveHealth '18; May 21-24, 2018; New York, NY p. 187-190 URL: https://dl.acm.org/doi/10.1145/3240925.3240928 [doi: 10.1145/3240925.3240928]
45. Wang JB, Cadmus-Bertram LA, Natarajan L, White MM, Madanat H, Nichols JF, et al. Wearable sensor/device (Fitbit One) and SMS text-messaging prompts to increase physical activity in overweight and obese adults: a randomized controlled trial. Telemed J E Health 2015 Oct;21(10):782-792 [FREE Full text] [doi: 10.1089/tmj.2014.0176] [Medline: 26431257]
46. Tong X, Gromala D, Shaw CD, Choo A. A field study: evaluating gamification approaches for promoting physical activity with motivational models of behavior changes. In: Proceedings of the 18th International Conference, HCI International on Human-Computer Interaction. Novel User Experiences. 2016 Presented at: HCINUE '16; July 17-22, 2016; Toronto, ON p. 417-424 URL: http://link.springer.com/10.1007/978-3-319-39513-5_39 [doi: 10.1007/978-3-319-39513-5_39]
47. Domin A, Uslu A, Schulz A, Ouzzahra Y, Vögele C. A theory-informed, personalized mHealth intervention for adolescents (mobile app for physical activity): development and pilot study. JMIR Form Res 2022 Jun 10;6(6):e35118 [FREE Full text] [doi: 10.2196/35118] [Medline: 35687409]
48. Chen Y, Randriambelonoro M, Geissbuhler A, Pu P. Social incentives in pervasive fitness apps for obese and diabetic patients. In: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion. 2016 Presented at: CSCW '16; February 26-March 2, 2016; San Francisco, CA p. 245-248 URL: https://dl.acm.org/doi/10.1145/2818052.2869093 [doi: 10.1145/2818052.2869093]
49. Walsh JC, Corbett T, Hogan M, Duggan J, McNamara A. An mHealth intervention using a smartphone app to increase walking behavior in young adults: a pilot study. JMIR Mhealth Uhealth 2016 Sep 22;4(3):e109 [FREE Full text] [doi: 10.2196/mhealth.5227] [Medline: 27658677]
50. Burkow TM, Vognild LK, Johnsen E, Bratvold A, Risberg MJ. Promoting exercise training and physical activity in daily life: a feasibility study of a virtual group intervention for behaviour change in COPD. BMC Med Inform Decis Mak 2018 Dec 18;18(1):136 [FREE Full text] [doi: 10.1186/s12911-018-0721-8] [Medline: 30563507]
51. Korinek EV, Phatak SS, Martin CA, Freigoun MT, Rivera DE, Adams MA, et al. Adaptive step goals and rewards: a longitudinal growth model of daily steps for a smartphone-based walking intervention. J Behav Med 2018 Feb;41(1):74-86. [doi: 10.1007/s10865-017-9878-3] [Medline: 28918547]
52. Oyebode O, Ganesh A, Orji R. TreeCare: development and evaluation of a persuasive mobile game for promoting physical activity. In: Proceedings of the 2021 IEEE Conference on Games. 2021 Presented at: CoG '21; August 17-20, 2021; Copenhagen, Denmark p. 1-8 URL: https://ieeexplore.ieee.org/document/9619035 [doi: 10.1109/cog52621.2021.9619035]
53. Mönninghoff A, Fuchs K, Wu J, Albert J, Mayer S. The effect of a future-self avatar mobile health intervention (FutureMe) on physical activity and food purchases: randomized controlled trial. J Med Internet Res 2022 Jul 07;24(7):e32487 [FREE Full text] [doi: 10.2196/32487] [Medline: 35797104]
54. Poppe L, De Bourdeaudhuij I, Verloigne M, Shadid S, Van Cauwenberg J, Compernolle S, et al. Efficacy of a self-regulation-based electronic and mobile health intervention targeting an active lifestyle in adults having type 2 diabetes and in adults aged 50 years or older: two randomized controlled trials. J Med Internet Res 2019 Aug 02;21(8):e13363 [FREE Full text] [doi: 10.2196/13363] [Medline: 31376274]
55. Kanaoka T, Mutlu B. Designing a motivational agent for behavior change in physical activity. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. 2024 Presented at: CHI EA '15; April 18-23, 2015; Seoul, Republic of Korea p. 1445-1450 URL: https://dl.acm.org/doi/10.1145/2702613.2732924 [doi: 10.1145/2702613.2732924]
56. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth 2015 May 14;3(2):e42 [FREE Full text] [doi: 10.2196/mhealth.4160] [Medline: 25977197]
57. Voth EC, Oelke ND, Jung ME. A theory-based exercise app to enhance exercise adherence: a pilot study. JMIR Mhealth Uhealth 2016 Jun 15;4(2):e62 [FREE Full text] [doi: 10.2196/mhealth.4997] [Medline: 27307134]
58. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. JMIR Mhealth Uhealth 2020 May 19;8(5):e15085 [FREE Full text] [doi: 10.2196/15085] [Medline: 32427114]
59. Collins T, Geana M, Overton K, Benton M, Lu L, Khan F, et al. Use of a smartphone app versus motivational interviewing to increase walking distance and weight loss in overweight/obese adults with peripheral artery disease: pilot randomized trial. JMIR Form Res 2022 Feb 03;6(2):e30295 [FREE Full text] [doi: 10.2196/30295] [Medline: 35113020]

XSL•FO
RenderX

60. King AC, Hekler EB, Grieco LA, Winter SJ, Sheats JL, Buman MP, et al. Effects of three motivationally targeted mobile device applications on initial physical activity and sedentary behavior change in midlife and older adults: a randomized trial. PLoS One 2016;11(6):e0156370 [FREE Full text] [doi: 10.1371/journal.pone.0156370] [Medline: 27352250]
61. Liew SJ, Gorny AW, Tan CS, Müller-Riemenschneider F. A mobile health team challenge to promote stepping and stair climbing activities: exploratory feasibility study. JMIR Mhealth Uhealth 2020 Feb 04;8(2):e12665 [FREE Full text] [doi: 10.2196/12665] [Medline: 32014845]
62. Mascarenhas MN, Chan JM, Vittinghoff E, Van Blarigan EL, Hecht F. Increasing physical activity in mothers using video exercise groups and exercise mobile apps: randomized controlled trial. J Med Internet Res 2018 May 18;20(5):e179 [FREE Full text] [doi: 10.2196/jmir.9310] [Medline: 29776899]
63. Tong HL, Coiera E, Tong W, Wang Y, Quiroz JC, Martin P, et al. Efficacy of a mobile social networking intervention in promoting physical activity: quasi-experimental study. JMIR Mhealth Uhealth 2019 Mar 28;7(3):e12181 [FREE Full text] [doi: 10.2196/12181] [Medline: 30920379]
64. Zhu J, Dallal DH, Gray RC, Villareale J, Ontañón S, Forman EM, et al. Personalization paradox in behavior change apps: lessons from a social comparison-based personalized app for physical activity. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-21. [doi: 10.1145/3449190]
65. Buckingham SA, Morrissey K, Williams AJ, Price L, Harrison J. The physical activity wearables in the police force (PAW-force) study: acceptability and impact. BMC Public Health 2020 Nov 03;20(1):1645 [FREE Full text] [doi: 10.1186/s12889-020-09776-1] [Medline: 33143665]
66. Fuemmeler BF, Holzwarth E, Sheng Y, Do EK, Miller CA, Blatt J, et al. Mila blooms: a mobile phone application and behavioral intervention for promoting physical activity and a healthy diet among adolescent survivors of childhood cancer. Games Health J 2020 Aug;9(4):279-289 [FREE Full text] [doi: 10.1089/g4h.2019.0060] [Medline: 32391734]
67. Fiedler J, Seiferth C, Eckert T, Woll A, Wunsch K. A just-in-time adaptive intervention to enhance physical activity in the SMARTFAMILY2.0 trial. Sport Exerc Perform Psychol 2023 Feb;12(1):43-57. [doi: 10.1037/spy0000311]
68. Zuckerman O, Gal-Oz A. Deconstructing gamification: evaluating the effectiveness of continuous measurement, virtual rewards, and social comparison for promoting physical activity. Pers Ubiquit Comput 2014 Jul 5;18(7):1705-1719. [doi: 10.1007/s00779-014-0783-2]
69. Edney SM, Olds TS, Ryan JC, Vandelanotte C, Plotnikoff RC, Curtis RG, et al. A social networking and gamified app to increase physical activity: cluster RCT. Am J Prev Med 2020 Feb;58(2):e51-e62. [doi: 10.1016/j.amepre.2019.09.009] [Medline: 31959326]
70. Lee D, Frey GC, Cothran DJ, Harezlak J, Shih PC. Effects of a gamified, behavior change technique-based mobile app on increasing physical activity and reducing anxiety in adults with autism spectrum disorder: feasibility randomized controlled trial. JMIR Form Res 2022 Jul 28;6(7):e35701 [FREE Full text] [doi: 10.2196/35701] [Medline: 35900808]
71. Spruijt-Metz D, Wen CK, O'Reilly G, Li M, Lee S, Emken BA, et al. Innovations in the use of interactive technology to support weight management. Curr Obes Rep 2015 Dec;4(4):510-519 [FREE Full text] [doi: 10.1007/s13679-015-0183-6] [Medline: 26364308]
72. Stephens S, Schneiderman JE, Finlayson M, Berenbaum T, Motl RW, Yeh EA. Feasibility of a theory-informed mobile app for changing physical activity in youth with multiple sclerosis. Mult Scler Relat Disord 2022 Feb;58:103467. [doi: 10.1016/j.msard.2021.103467] [Medline: 34954651]
73. Fan H, Poole MS. What is personalization? perspectives on the design and implementation of personalization in information systems. J Organ Comput Electron Commer 2006 Jan;16(3-4):179-202. [doi: 10.1080/10919392.2006.9681199]
74. Hagger MS. Redefining habits and linking habits with other implicit processes. Psychol Sport Exerc 2020 Jan;46:101606. [doi: 10.1016/j.psychsport.2019.101606]
75. Ju W, Leifer L. The design of implicit interactions: making interactive systems less obnoxious. Design Issues 2008 Jul;24(3):72-84. [doi: 10.1162/desi.2008.24.3.72]
76. Serim B, Jacucci G. Explicating "implicit interaction": an examination of the concept and challenges for research. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK p. 1-16 URL: https://dl.acm.org/doi/10.1145/3290605.3300647 [doi: 10.1145/3290605.3300647]
77. Wexley KN, Nemeroff WF. Effectiveness of positive reinforcement and goal setting as methods of management development. J Appl Psychol 1975 Aug;60(4):446-450. [doi: 10.1037/h0076912]
78. Strozier AL. The effectiveness of support groups in increasing social support for kinship caregivers. Child Youth Serv Rev 2012 May;34(5):876-881. [doi: 10.1016/j.childyouth.2012.01.007]
79. McEwan D, Harden SM, Zumbo BD, Sylvester BD, Kaulius M, Ruissen GR, et al. The effectiveness of multi-component goal setting interventions for changing physical activity behaviour: a systematic review and meta-analysis. Health Psychol Rev 2016;10(1):67-88. [doi: 10.1080/17437199.2015.1104258] [Medline: 26445201]
80. Hernández-Reyes A, Cámara-Martos F, Molina Recio G, Molina-Luque R, Romero-Saldaña M, Moreno Rojas R. Push notifications from a mobile app to improve the body composition of overweight or obese women: randomized controlled trial. JMIR Mhealth Uhealth 2020 Feb 12;8(2):e13747 [FREE Full text] [doi: 10.2196/13747] [Medline: 32049065]

XSL•FO
RenderX

81. Li R, Liang N, Bu F, Hesketh T. The effectiveness of self-management of hypertension in adults using mobile health: systematic review and meta-analysis. JMIR Mhealth Uhealth 2020 Mar 27;8(3):e17776 [FREE Full text] [doi: 10.2196/17776] [Medline: 32217503]
82. Wang Y, Xue H, Huang Y, Huang L, Zhang D. A systematic review of application and effectiveness of mHealth interventions for obesity and diabetes treatment and self-management. Adv Nutr 2017 May 15;8(3):449-462 [FREE Full text] [doi: 10.3945/an.116.014100] [Medline: 28507010]
83. Muellmann S, Forberger S, Möllers T, Bröring E, Zeeb H, Pischke CR. Effectiveness of eHealth interventions for the promotion of physical activity in older adults: a systematic review. Prev Med 2018 Mar;108:93-110. [doi: 10.1016/j.ypmed.2017.12.026] [Medline: 29289643]
84. Azevedo LB, Stephenson J, Ells L, Adu-Ntiamoah S, DeSmet A, Giles EL, et al. The effectiveness of e-health interventions for the treatment of overweight or obesity in children and adolescents: a systematic review and meta-analysis. Obes Rev 2022 Feb 07;23(2):e13373. [doi: 10.1111/obr.13373] [Medline: 34747118]
85. Heart disease and stroke. national center for chronic disease prevention and health promotion. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm [accessed 2024-04-29]
86. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
87. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
88. Kreuter MW, Strecher VJ, Glassman B. One size does not fit all: the case for tailoring print materials. Ann Behav Med 1999 Dec;21(4):276-283. [doi: 10.1007/bf02895958]
89. Lally P, Gardner B. Promoting habit formation. Health Psychol Rev 2013 May;7(sup1):S137-S158. [doi: 10.1080/17437199.2011.603640]
90. Yuasa T, Harada F, Shimakawa H. Estimation of behavior change stage from walking information and improvement of walking volume by message intervention. Int J Environ Res Public Health 2022 Feb 01;19(3):1668 [FREE Full text] [doi: 10.3390/ijerph19031668] [Medline: 35162691]
91. Consolvo S, Klasnja P, McDonald DW, Landay JA. Goal-setting considerations for persuasive technologies that encourage physical activity. In: Proceedings of the 4th International Conference on Persuasive Technology. 2009 Presented at: Persuasive '09; April 26-29, 2009; Claremont, CA p. 1-8 URL: https://dl.acm.org/doi/10.1145/1541948.1541960 [doi: 10.1145/1541948.1541960]
92. Li I, Dey A, Forlizzi J. A stage-based model of personal informatics systems. In: Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems. 2024 Presented at: CHI '10; April 10-15, 2010; Atlanta, GA p. 557-566 URL: https://dl.acm.org/doi/10.1145/1753326.1753409 [doi: 10.1145/1753326.1753409]
93. Zhu Y, Lee K, Zhang L, Wang SJ. Meaningful smart health data: a design guide for transparent data to enhance self-reflection. In: Proceedings of the 8th International Conference on Human Interaction and Emerging Technologies. 2022 Presented at: IHIET '22; August 22-24, 2022; Nice, France p. 2022 URL: https://research.polyu.edu.hk/en/publications/meaningful-smart-health-data-a-design-guide-for-transparent-data- [doi: 10.54941/ahfe1002748]
94. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
95. Gardner B, Arden MA, Brown D, Eves FF, Green J, Hamilton K, et al. Developing habit-based health behaviour change interventions: twenty-one questions to guide future research. Psychol Health 2023 Apr 14;38(4):518-540 [FREE Full text] [doi: 10.1080/08870446.2021.2003362] [Medline: 34779335]
96. Kunkel J, Donkers T, Michael L, Barbu CM, Ziegler J. Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK p. 1-12 URL: https://dl.acm.org/doi/10.1145/3290605.3300717 [doi: 10.1145/3290605.3300717]
97. Long Y, Luo X, Zhu Y, Lee KP, Wang SJ. Data transparency design in internet of things: a systematic review. Int J Hum Comput Interact 2023 Jul 18:1-23 [FREE Full text] [doi: 10.1080/10447318.2023.2228997]
98. Cho J, Xu T, Zimmermann-Niefield A, Voida S. Reflection in theory and reflection in practice: an exploration of the gaps in reflection support among personal informatics apps. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-23 URL: https://dl.acm.org/doi/abs/10.1145/3491102.3501991 [doi: 10.1145/3491102.3501991]
99. Paruthi G, Raj S, Colabianchi N, Klasnja P, Newman MW. Finding the sweet spot(s): understanding context to support physical activity plans. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar 26;2(1):1-17. [doi: 10.1145/3191761]
100. Xu K, Yan X, Newman MW. Understanding people's experience for physical activity planning and exploring the impact of historical records on plan creation and execution. In: Proceedings of the 2022 CHI Conference on Human Factors in

XSL•FO
RenderX

Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-15 URL: https://dl.acm.org/doi/10.1145/3491102.3501997 [doi: 10.1145/3491102.3501997]
101. Sniehotta FF. Towards a theory of intentional behaviour change: Plans, planning, and self-regulation. British J Health Psychol 2010 Dec 24;14(2):261-273. [doi: 10.1348/135910708x389042]
102. Ismail T, Al Thani D. Design and evaluation of a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior at work: experimental study. JMIR Form Res 2022 Jan 26;6(1):e34309 [FREE Full text] [doi: 10.2196/34309] [Medline: 35080498]
103. Sommer B, Jia Wang S, Xu L, Chen M, Schreiber F. Hybrid-dimensional visualization and interaction - integrating 2D and 3D visualization with semi-immersive navigation techniques. In: Proceedings of the 2015 Big Data Visual Analytics. 2015 Presented at: BDVA '15; September 22-25, 2015; Hobart, Australia p. 1-8 URL: https://ieeexplore.ieee.org/document/7314295?reload=true [doi: 10.1109/bdva.2015.7314295]
104. Feustel C, Aggarwal S, Lee B, Wilcox L. People like me: designing for reflection on aggregate cohort data in personal informatics systems. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Sep 18;2(3):1-21. [doi: 10.1145/3264917]
105. Nim HT, Wang M, Zhu Y, Sommer B, Schreiber F, Boyd SE, et al. Communicating the effect of human behaviour on the great barrier reef via mixed reality visualisation. In: Proceedings of the 2016 Big Data Visual Analytics. 2016 Presented at: BDVA '16; November 22-25, 2016; Sydney, Australia p. 1-6 URL: https://ieeexplore.ieee.org/document/7787046 [doi: 10.1109/bdva.2016.7787046]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique
DBCI: digital behavior change intervention
GE: generalization-explicitness
GI: generalization-implicitness
HCI: human-computer interaction
IoT: Internet of Things
PE: personalization-explicitness
PI: personalization-implicitness
PRISMA: Preferred Reporting Item for Systematic Reviews and Meta-Analyses
RQ: research question

Edited by G Greco; submitted 08.11.23; peer-reviewed by KL Alfrey, B Gardner; comments to author 26.01.24; revised version received 22.03.24; accepted 08.04.24; published 24.05.24.

Please cite as:
Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ
Digital Behavior Change Intervention Designs for Habit Formation: Systematic Review
J Med Internet Res 2024;26:e54375
URL: https://www.jmir.org/2024/1/e54375
doi:10.2196/54375
PMID:38787601

©Yujie Zhu, Yonghao Long, Hailiang Wang, Kun Pyo Lee, Lie Zhang, Stephen Jia Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Advances in the Application of AI Robots in Critical Care: Scoping Review

Yun Li[1,2*], MD; Min Wang[1,2*], MPH; Lu Wang[1,2*], MD; Yuan Cao[3*], MPH; Yuyan Liu[1,2], MPH; Yan Zhao[2], MPH; Rui Yuan[1,2], MD; Mengmeng Yang[2], MPH; Siqian Lu[4], DSC; Zhichao Sun[4], DSC; Feihu Zhou[2], PhD; Zhirong Qian[4,5,6], PhD; Hongjun Kang[2], PhD

[1]Medical School of Chinese PLA, Beijing, China
[2]The First Medical Centre, Chinese PLA General Hospital, Beijing, China
[3]The Second Hospital, Hebei Medical University, Hebei, China
[4]Beidou Academic & Research Center, Beidou Life Science, Guangzhou, China
[5]Department of Radiation Oncology, Fujian Medical University Union Hospital, Fujian, China
[6]The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen, China
[*]these authors contributed equally

Corresponding Author:
Hongjun Kang, PhD
The First Medical Centre
Chinese PLA General Hospital
28 Fuxing Road
Haidian District,
Beijing, 100853
China
Phone: 86 13811989878
Email: doctorkang301@163.com

Abstract

Background: In recent epochs, the field of critical medicine has experienced significant advancements due to the integration of artificial intelligence (AI). Specifically, AI robots have evolved from theoretical concepts to being actively implemented in clinical trials and applications. The intensive care unit (ICU), known for its reliance on a vast amount of medical information, presents a promising avenue for the deployment of robotic AI, anticipated to bring substantial improvements to patient care.

Objective: This review aims to comprehensively summarize the current state of AI robots in the field of critical care by searching for previous studies, developments, and applications of AI robots related to ICU wards. In addition, it seeks to address the ethical challenges arising from their use, including concerns related to safety, patient privacy, responsibility delineation, and cost-benefit analysis.

Methods: Following the scoping review framework proposed by Arksey and O'Malley and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we conducted a scoping review to delineate the breadth of research in this field of AI robots in ICU and reported the findings. The literature search was carried out on May 1, 2023, across 3 databases: PubMed, Embase, and the IEEE Xplore Digital Library. Eligible publications were initially screened based on their titles and abstracts. Publications that passed the preliminary screening underwent a comprehensive review. Various research characteristics were extracted, summarized, and analyzed from the final publications.

Results: Of the 5908 publications screened, 77 (1.3%) underwent a full review. These studies collectively spanned 21 ICU robotics projects, encompassing their system development and testing, clinical trials, and approval processes. Upon an expert-reviewed classification framework, these were categorized into 5 main types: therapeutic assistance robots, nursing assistance robots, rehabilitation assistance robots, telepresence robots, and logistics and disinfection robots. Most of these are already widely deployed and commercialized in ICUs, although a select few remain under testing. All robotic systems and tools are engineered to deliver more personalized, convenient, and intelligent medical services to patients in the ICU, concurrently aiming to reduce the substantial workload on ICU medical staff and promote therapeutic and care procedures. This review further explored the prevailing challenges, particularly focusing on ethical and safety concerns, proposing viable solutions or methodologies, and illustrating the prospective capabilities and potential of AI-driven robotic technologies in the ICU environment. Ultimately,

XSL•FO
RenderX

we foresee a pivotal role for robots in a future scenario of a fully automated continuum from admission to discharge within the ICU.

Conclusions: This review highlights the potential of AI robots to transform ICU care by improving patient treatment, support, and rehabilitation processes. However, it also recognizes the ethical complexities and operational challenges that come with their implementation, offering possible solutions for future development and optimization.

(J Med Internet Res 2024;26:e54095) doi:10.2196/54095

KEYWORDS

critical care medicine; artificial intelligence; AI; robotics; intensive care unit; ICU

Introduction

Background

Artificial intelligence (AI) and robotics are 2 distinct yet interconnected concepts ubiquitous in contemporary media and digital platforms. The term *artificial intelligence* was first introduced as a Medical Subject Heading in the US National Library of Medicine's PubMed database in 1986, defined as "Theory and development of computer systems which perform tasks that normally require human intelligence" [1]. The hallmarks of AI encompass autonomous thinking, learning, recognition, reasoning, judgment, and inference. Medicine has long been considered a promising application field for AI, where it can augment clinical diagnostics and decision-making capacities [2]. Robotics, as defined by the Medical Subject Headings of the US National Library of Medicine, pertains to "the application of electronic, computerized control systems to mechanical devices designed to perform human functions" [3]. Presently, a standardized definition for "AI robots" remains elusive. However, they can be perceived as "physical devices inheriting electronic, computerized, and mechanical control systems, capable of perception, reasoning, learning, decision-making, and task execution without direct human control, able to mimic and execute various tasks of human intelligence" [4]. The crux of this review is the discussion of AI robot applications within the intensive care unit (ICU). While AI technology encompasses machine learning, deep learning, predictive modeling, and natural language processing, this review did not delve into these distinct technologies. Rather, it concentrated on the practical products and application cases where these AI technologies are integrated into robotic systems.

The past few decades have witnessed an exponential proliferation of research into AI robots, particularly in the health care domain. However, most of these developments have remained confined to the stages of product development and testing, with few achieving large-scale clinical implementation. The year 2020 sparked an unforeseen public health incident that catalyzed intense interest in AI robots. The outbreak of COVID-19 expedited the transformative revolution of "Healthcare + AI + Robotics" technologies and their applications [5,6]. The deployment of AI robots significantly curtailed the infection risk for medical personnel in contagion hot spots, and AI robots stood on the front lines in the battle against COVID-19 transmission [7]. Serving as a technology that improves performance, precision, and time efficiency as well as reducing costs, AI robots have promoted the upgrade and development of modern industry and are being rapidly adopted by many industries [8,9]. The applicability of robots in society is already evident and growing significantly [10].

Despite the rapid advancements in this field across military, security, transportation, and manufacturing sectors, our focus remains on intensive care scenarios such as COVID-19. ICUs are specialized settings designed to provide systematic, high-quality medical care and life-saving treatment to patients with single- or multiorgan dysfunction, life-threatening illnesses, or potential high-risk factors [11]. As scientific and technological advancements accelerate, the contradiction between the high demand for quality care for patients who are critically ill and the chronic shortage of medical resources becomes increasingly pronounced [12]. In 1995, Hanson and Marshall [13] postulated that AI could reduce care costs for patients in the ICU and improve their prognosis. "There are plenty of areas in critical care where it would be extremely helpful to have efficacious, fair, and transparent AI systems," notes Gary Weissman, professor of pulmonary and critical care medicine [14]. The full potential of AI will be realized once it becomes a trusted clinical assistant for intensivists. After all, ICUs, which routinely collect a significant volume of data, provide an ideal setting for the deployment of machine learning technologies [15].

Objectives

Currently, the application of AI in intensive care predominantly focuses on "assisting" health care professionals. This review targeted AI robots in ICUs, primarily discussing relevant advancements in recent years and presenting and analyzing challenges faced in ICUs and potential solutions, as well as strategies for health care professionals to handle AI robotic technologies, for the reference of health care professionals and system developers.

Methods

Design

We followed the scoping review methodology proposed by Arksey and O'Malley [16], which includes (1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data; and (5) collating, summarizing, and reporting the results. In addition, to ensure the rigor of the scoping review, we adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines (Multimedia Appendix 1 [17]).

XSL•FO
RenderX

Overview

Upon establishing the focus on the application of AI robots in ICUs, we embarked on a keyword search encompassing terms related to robots and intensive care. Due to the novelty and complexity of the research topic, the experimental designs and results presentation vary significantly across the literature, making traditional methods of literature quality assessment inadequate. Consequently, we relied on the profound knowledge and practical experience of domain experts to provide essential perspectives for understanding the complexities presented in the literature. We engaged 4 experts—2 from the IT sector and 2 medical professionals—to devise a search strategy and select appropriate databases. Given the interdisciplinary nature of our research, spanning medicine, robotics engineering, and human-computer interaction design, our search was not confined to medical databases; we also included databases from the engineering field. Our literature search encompassed 3 electronic databases: PubMed, Embase, and the IEEE Xplore Digital Library. We restricted our search to publications in English, imposing no limits on the year of publication. The search was conducted over a brief period, from May 1, 2023, to May 5, 2023. Textbox 1 provides detailed insights into our search methodology.

Textbox 1. Databases and search strings.

Search strings

- PubMed: (“Artificial Intelligence” [Medical Subject Heading (MeSH)] OR “AI” OR “Robotics” [MeSH] OR “Robots”) AND (“Intensive Care Units” [MeSH] OR “ICU” OR “Critical Care Units” OR “CCU”)
- Embase: (“artificial intelligence”/exp OR “ai” OR “robotics”/exp OR “robots”) AND (“intensive care unit”/exp OR “icu” OR “critical care unit” OR “ccu”)
- IEEE Xplore Digital Library: (“Artificial Intelligence” OR “AI” OR “Robotics” OR “Robots”) AND (“Intensive Care Units” OR “ICU” OR “Critical Care Units” OR “CCU”)

During our initial search across the 3 databases, we retrieved a total of 5908 articles. These search results were first imported into EndNote (Clarivate Analytics), a literature management tool, to facilitate the deduplication process. Subsequently, LW and YC independently conducted a screening of titles and abstracts to weed out articles that were clearly irrelevant. Given the market’s saturation with robot products of similar functionalities and the constraints on this review’s length, we opted to focus on the most representative or widely adopted technologies among robots offering the same functions. This selection process led to the inclusion of 77 articles in our review. These studies encompassed multiple phases, including the design, development, and validation of robots, collectively covering 21 different AI robotic products, detailing their development, evolution, and application examples. Figure 1 shows a comprehensive view of our study selection process.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the screening process.

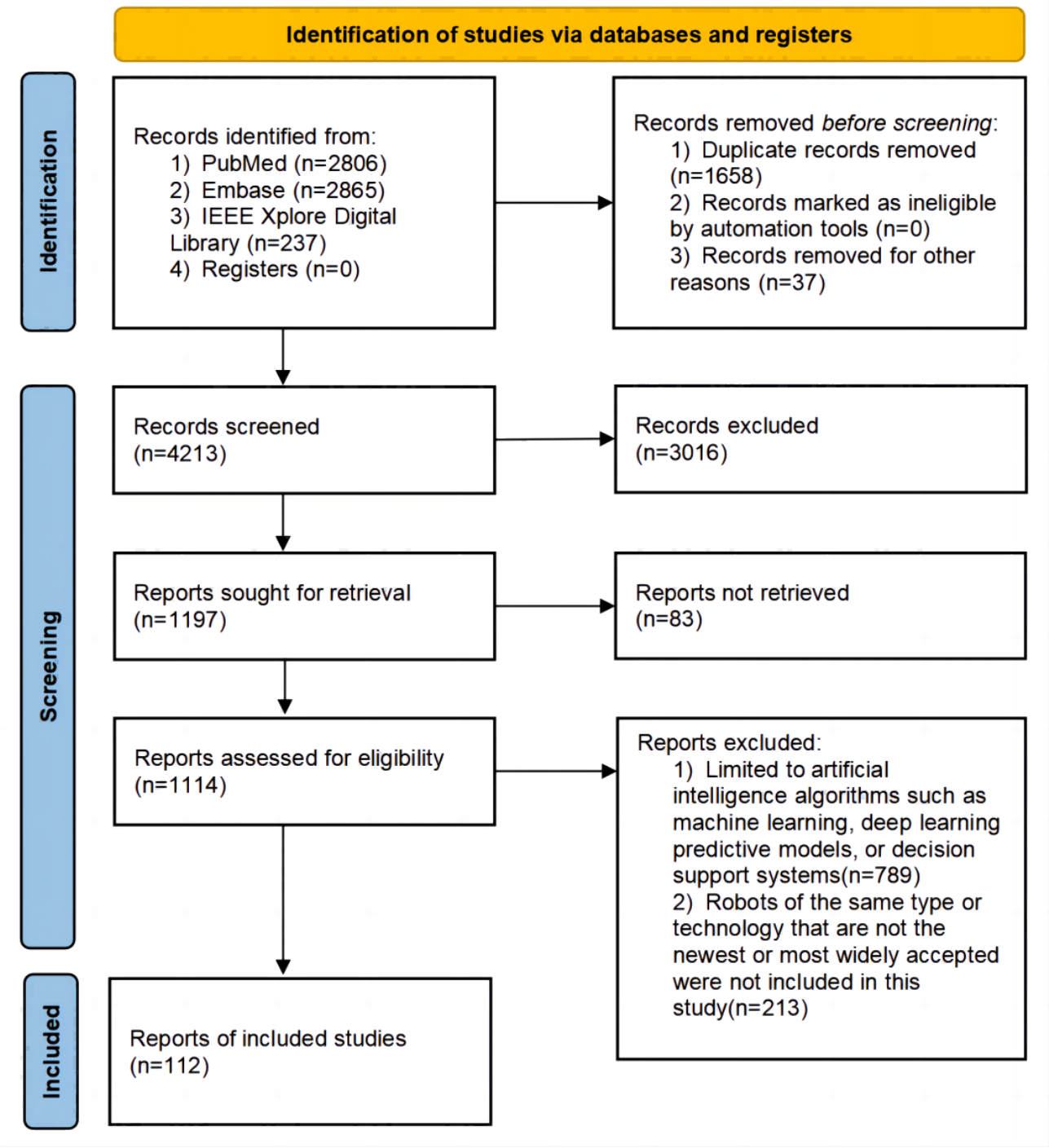

In our examination of the 77 articles selected for inclusion, we meticulously identified the key characteristics of each study. These characteristics encompassed the design purpose of the technology, the AI algorithms used, the anticipated application scenarios, and the target user groups. From this analysis, we developed an initial classification framework aimed at organizing technologies based on their primary functions and application contexts. For instance, technologies designed for nursing tasks were classified under nursing assistance robots, whereas those with therapeutic responsibilities were designated as therapeutic assistance robots. Similarly, devices integrated with ventilators capable of remote operation and achieving therapeutic goals were also classified as therapeutic assistance robots.

To validate the accuracy and logic behind our classification, we sought insights from additional experts in the domains of medicine and information engineering. Their feedback prompted adjustments and refinements to our framework, ensuring that it precisely represented the nuances and relationships among the various technologies.

Subsequently, applying the refined classification framework, we categorized the included studies into their respective groups, deriving organized results. This meticulous classification was undertaken with the aim of ensuring clarity and

XSL•FO
RenderX

comprehensibility for all potential readers, including professionals (physicians and engineers), patients, and their families. Our goal was to provide a clear understanding of the applications and potential of various technologies in the ICU, making the information accessible and valuable to a broad audience.

Inclusion and Exclusion Criteria

This review included studies that met two main criteria: (1) the studies needed to focus on actual products or application cases that conformed to the definition of AI robots, and (2) the application scenario of the study had to be the ICU or any place for patients discharged from the ICU.

We excluded studies that met any of the following conditions: (1) limited to AI algorithms such as machine learning, deep learning predictive models, or decision support systems; and (2) not published in English.

Results

Application of AI Robots in Intensive Care

Overview

Currently, there is a broad array of experimental studies and AI robots applied in ICUs. We categorized these into 5 main types based on their application scenarios and functions in ICUs: therapeutic assistance robots, nursing assistance robots, rehabilitation assistance robots, telepresence robots, and logistics and disinfection robots (Figure 2). The application of robotics in the medical field is extensive, encompassing myriad functions and scenarios, rendering their precise classification a formidable task. Multimedia Appendix 2 [6,18-90], which is neither exclusive nor exhaustive, summarizes the representative research groups and commercial suppliers to provide readers with a panoramic view of the field.

Figure 2. Types of robots available in an intelligent intensive care unit (ICU). CPR: cardiopulmonary resuscitation.

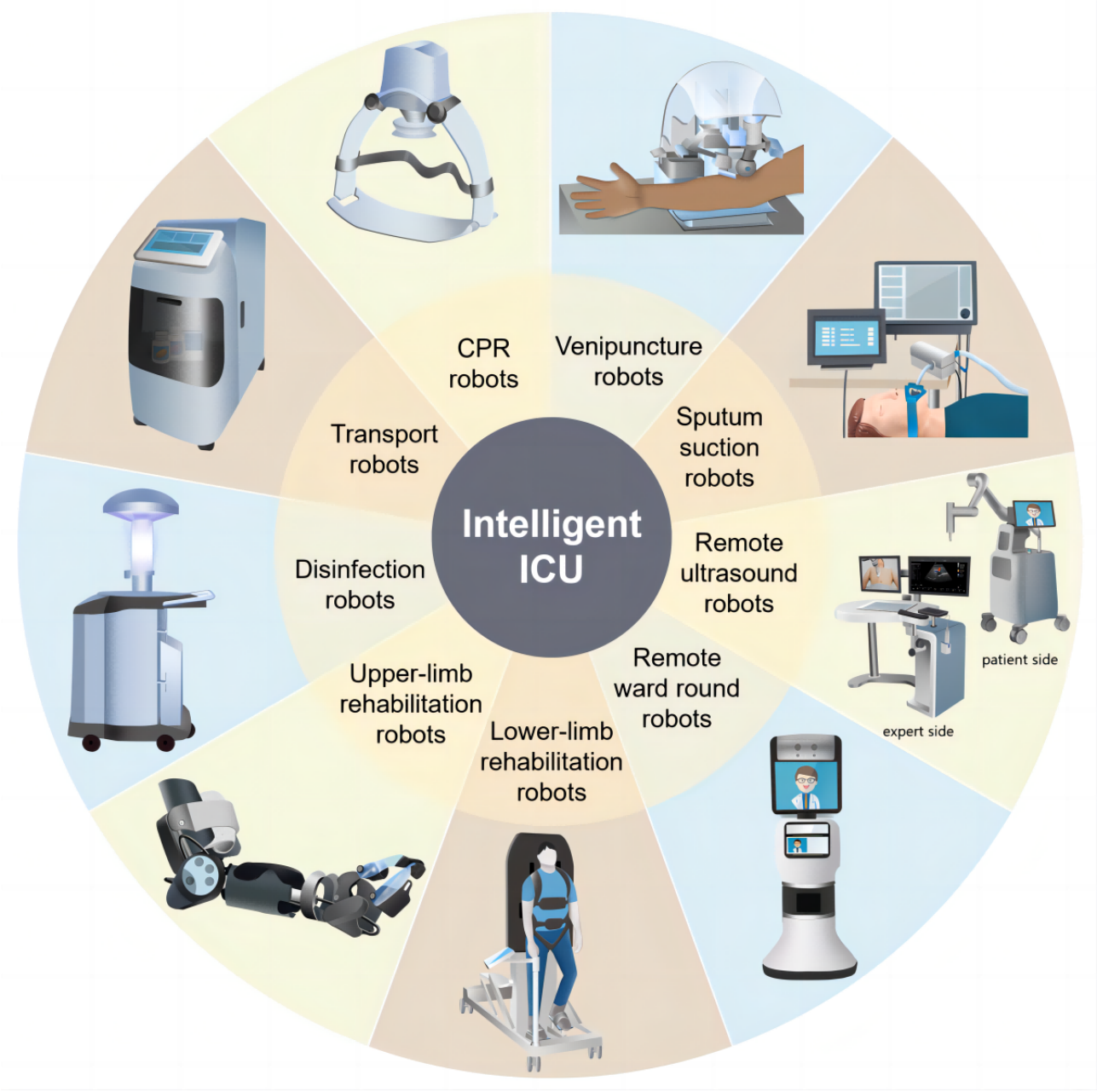

XSL•FO
RenderX

Therapeutic Assistance Robots

The ICU is a specialized department dedicated to the treatment of patients who are critically ill. Driven by the necessity to enhance therapeutic outcomes, mitigate medical risks, alleviate the workload of health care personnel, and provide personalized treatment regimens for patients, therapeutic assistance robots in the ICU have become indispensable [91].

Ventilator-Mounted Cartesian Robot

This can also be viewed as a type of telepresence robot. A teleoperated ventilator controller system has been developed, which consists of a custom robotic patient-side device and a touch-based master console. Computer vision tasks that enable an intuitive user interface and accurate robot control are executed on the master console. The system was initially developed and deployed for the most popular touch screen–controlled ventilator at Johns Hopkins Hospital, the Maquet Servo-u (Getinge AB) [18]. Different from the traditional ventilators in most ICUs, the robot performs remote control and monitoring outside the ICU room through the network. The actual ventilator settings are controlled and adjusted remotely via physical controls (such as buttons and knobs) and synchronized real-time image transmission video feedback [19]. It reduces the time staff spend entering the ICU performing simple tasks such as changing ventilator settings, reducing the risk of infection and stress on personal protective clothing resources. Following a qualitative assessment in clinical environments, feedback from respiratory therapists highlighted that the system could significantly empower the respiratory care team by liberating valuable resources. The design's Cartesian layout enables the robot to approach the operating table in as horizontal a manner as possible. This approach not only minimizes the installation burden but also ensures that the robot's operation remains closely aligned with conventional manual procedures. Such a configuration facilitates ease of use and integration into existing medical workflows, thereby enhancing efficiency without compromising the quality of patient care [18]. In a simulated ICU environment, a Cartesian robot reduced the total time required for a respiratory therapist to make typical setup adjustments to a traditional ventilator from 271 to 109 seconds, which was 2.49 times faster [18]. More recently, Song et al [20] developed an integrated telemonitoring or operation system with an accurate XY positioner and a 3-df end effector for accurate manipulation (maximum positioning error of 0.695 mm; repeatability is 0.348 mm).

In ICUs where a single brand of ventilator is predominantly used, configuring Cartesian robots proves to be more convenient. Nevertheless, to accommodate a broader array of ICU wards and equipment types such as infusion pumps, it is imperative to expand the variety and range of control interfaces. This expansion aims to enhance the robot system's capabilities for physical control interactions. However, a significant hurdle to the clinical application of these robotic systems is the challenge associated with their cleaning and disinfection. This issue could be mitigated in the future by enclosing the equipment within acrylic covers, thereby simplifying the process of maintaining hygiene and ensuring the equipment's safety for patient care [6]. This solution not only addresses infection control concerns but also aids in the seamless integration of robotic systems into the rigorous and cleanliness-focused environment of the ICU.

McSleepy

McSleepy (Intelligent Technology in Anesthesia research group laboratory, McGill University, Montreal, Quebec, Canada), a real robot for anesthesia, is able to autonomously control hypnosis, analgesia, and neuromuscular block at the same time with regard to induction, maintenance, and emergence [21]. The anesthesiologist begins by entering patient data, including height, weight, type of surgery, and medication history, on the touch screen. In fully automatic mode, the system will use remifentanil, propofol, and rocuronium or succinylcholine to induce anesthesia. McSleepy assists the anesthesiologist in the same way that automatic transmission assists people when driving. Through a sensor that measures muscle movement, it monitors the patient's depth of consciousness, pain severity, and muscle movement and injects the corresponding dose of medication intravenously according to a built-in algorithm based on the obtained data. It provides deep or peripheral muscle relaxation in different closed-loop models [22]. As such, anesthesiologists can focus more on other aspects of direct patient care. An additional feature is that the system can communicate with PDAs, making distant monitoring and anesthetic control possible [21,23,24].

In a pilot study in the Department of Anesthesiology and Critical Care of the Bordeaux University Hospital, McSleepy performed automatic anesthesia for cardiac surgery with isoproterenol, remifentanil, and rocuronium without manual control. Automatic cardiac anesthesia was successfully performed in 80% (97.5% CI 53%-95%) of cases. Hypnosis was monitored using the bispectral index, which was <20% of the bispectral index target of 45, showing better control over hypnosis and duration of anesthesia [25]. A randomized controlled trial investigating a novel closed-loop drug delivery system found that automated delivery achieves superior sedation control over manual methods. This improvement is credited to the closed-loop system's ability to frequently or continuously monitor control variables and adjust drug delivery rates more often, circumventing the fatigue that can impair manual administration [26]. To guarantee the safety of this automated system, numerous protective features have been implemented. These include preventing the administration of muscle relaxants when mask ventilation proves difficult and querying any manual actions that breach established safety protocols [27]. Recent advancements in research highlight the successful application of robotic technology in anesthesiology. Beyond the surgical anesthesia applications that we have detailed, delicate procedures such as endotracheal intubation are also seeing promising developments [28]. Anesthesiologists are encouraged to engage with robotic systems, leveraging their significant contributions to enhancing medical quality and efficacy, thereby ensuring the highest standard of patient care and treatment.

Cardiopulmonary Resuscitation Robots (Seoul University Medical College, South Korea)

Some reports indicate that failure of cardiopulmonary resuscitation (CPR) is an important limiting factor for

XSL•FO
RenderX

life-prolonging treatment of patients in the ICU [29,30]. Robots have enough power to achieve high-quality CPR, which could overcome the shortcomings of manual CPR and mechanical CPR devices. Recently, Jung et al [31] developed an automated robotic CPR system that performs CPR automatically, analyzes the patient's condition, and relays the information to the CPR system. In the initial state of the CPR process, the robot manipulator determines the optimal compression position by adjusting the point of pressure and, guided by end-tidal carbon dioxide levels, periodically and repeatedly delivers adequate speed and depth for performing CPR. The combined CPR system can accurately capture the compression condition of patients, overcome the blank periods when medical personnel alternately perform CPR, and increase the probability of cardiac resuscitation [31,32].

In a study using a porcine model of cardiac arrest, the use of a CPR robot did not significantly enhance the success rate of resuscitation efforts compared to manual CPR. However, there was a notable improvement in the neurological deficit scores observed 48 hours after resuscitation, suggesting a potential benefit in postresuscitation neurological outcomes [32]. Another study indicated that robot-assisted CPR outperformed traditional manual CPR methods, achieving higher resuscitation success rates in patients without specific injuries [31]. The anticipated benefits of incorporating robot-assisted CPR into clinical settings extend beyond merely improving resuscitation success rates. They also include the potential to reduce labor costs and minimize the instances of ICU staff congregating around a patient's bedside to alternate performing CPR. This technological advancement suggests a promising direction for enhancing the efficiency and effectiveness of resuscitation efforts, thereby improving patient outcomes while optimizing resource use within critical care environments.

Nursing Assistance Robots

Nursing care robots, designed to assist patients who are bed-ridden with simple long-term care services, have been widely used in hospitals to assist older adults and individuals with disabilities. The many invasive diagnostic procedures and heavy nursing workload (multiple multitasking) in the ICU make medical staff vulnerable and overloaded, with very limited ability to provide timely care to patients. Through the intervention of intelligent robots, heavy physical activity in nursing work is alleviated to a certain extent, and nurses are liberated from it so that they can put their energies into more professional and meticulous nursing work.

First, despite the proliferation of devices currently available on the market to assist with venipuncture, there remain certain limitations. Technologies such as ultrasound, which can easily penetrate human tissue, thereby enabling the high-resolution visualization of both shallow and deep tissue structures for vascular guidance, and devices such as the VeinViewer and AccuVein AV300 [33,34], which use near-infrared spectroscopy technology for vascular imaging, are fairly mature. However, these techniques are still constrained by their inability to provide depth of the vein as well as the fact that imaging technology does not directly aid the insertion of the needle. VenousPro (VascuLogic, LLC), an automated robotic venipuncture device, addresses these issues. It identifies vessels suitable for cannulation and robotically guides an attached needle toward the lumen center, enabling the safe drawing of blood from peripheral forearm veins [35,36]. The device uses 940-nm near-infrared light to enhance the contrast of the subcutaneous peripheral vein, selecting the appropriate vein through real-time imaging and mapping of the 3D spatial coordinates of the subcutaneous vein [36]. Chen et al [36] evaluated the system's cannulation accuracy on a vascular model through tracking, free-space localization, and use of a dark-skinned phlebotomy training model. Early versions of the device successfully demonstrated the feasibility of automatic venous access with a 100% success rate for venipuncture and placing of the needle in the desired position with high precision (mean positioning error 0.21 mm, SD 0.02 mm) [35]. This effectively reduced the risk of acupuncture injury [37].

Subsequently, the team developed an improved instrument based on a 9-df image-guided venipuncture robot, which increased radial rotation in response to rolling vein deformation and allowed for real-time adjustments to the pose and orientation of the needle [38]. Through robust optimization of near-infrared imaging coupled with image analysis and a robotic control system, this venipuncture robot realizes the automation of venipuncture, minimizing the likelihood of needlestick injuries and related bloodborne infections. This not only liberates nursing staff from the repetitive task of venipuncture but also protects the safety of practitioners [36]. In addition, this device may, in the future, be integrated with diagnostics, facilitating automated blood draws and rapid patient information synchronization. Moreover, the foundational imaging, computer vision, model recognition, and robotic technology developed for this device can be extended to arterial pathways, computer-assisted diagnostics, and miniature robotic surgery, potentially revolutionizing practice in ICUs and other departments.

Second, prompt and effective airway management can prove pivotal to the survival of patients who are critically ill. However, for patients with substantial secretions, sputum aspiration undertaken by nursing staff becomes a frequent necessity, significantly demanding the latter's time and effort. Therefore, a sputum suction robot (Xi'an Jiaotong University, China) has been developed, which is a simple 6-df manipulator with a steering gear. Exhibiting stable movement, the sputum suction robot is proficient in smoothly performing the clamping, insertion, back-off protection, and removal of the suction tube, thereby achieving effective sputum aspiration [39]. Naturally, the current iteration of this sputum suction robot faces certain challenges, such as the complexity of the mechanical arm structure and limited mobility that may result in some sputum being left unaspirated. The simplification of the driving mechanism and the optimization of mobility present promising directions for future enhancements. Such improvements aim to realize high-precision, high-quality automated sputum suction within the ICU.

Third, kangaroo care is a nursing method aimed at premature infants. It serves as a nonpharmacological intervention for treating surgical pain in infants. By promoting skin-to-skin contact between the infant and parents, it reduces the adverse effects of repeated surgical pain on the long-term development

XSL•FO
RenderX

of the nervous system [40]. However, a notable challenge within the neonatal ICU (NICU) setting is that parents cannot always be physically present to provide this essential care. This limitation calls for innovative solutions to ensure that premature infants still receive the benefits of kangaroo care, possibly through alternative methods or support systems that can simulate the presence and therapeutic effects of parental contact. Calmer (University of British Columbia, Vancouver, British Columbia, Canada) was developed to manage acute pain effectively for preterm infants in the NICU by simulating key pain-reducing components of human touch–based treatment. The human tactile pain treatment effect is achieved by placing the infant prone on the Calmer, which provides sensory intervention similar to parental skin contact while simulating the heartbeat and respiratory frequency of the parents with customized physiological signal processing software to mitigate the adverse neurodevelopmental effects of early pain exposure in preterm infants [41]. The effects of this robotic device on pain management during routine blood collection were studied in 10 infants. Calmer reduced physiological pain reactivity during and after painful blood collection procedures [34,40]. This approach not only aligns with the philosophy of nonpharmacological pain management but also potentially saves the NICU approximately US $380,000 per year in nursing time (United States). However, it is crucial to clarify that Calmer does not intend to replace the role of parents or neonatal care. It merely provides an alternative solution when such avenues are not available.

Fourth, older patients consistently represent a predominant segment of ICU admissions, manifesting unique characteristics and frequently presenting with cognitive impairments and mobility limitations. Physical assistive robots are strategically engineered to address the exigencies of daily living activities [92]. A robot named Paro has been extensively integrated within ICU environments and subsequent post-ICU discharge home care, offering socioemotional and psychological support to older patients in the ICU. With its lifelike appearance and behavior, Paro emulates genuine animal esthetics and movements, enabling it to perceive and respond to human vocalizations, touch, and actions. Recognized as one of the world's premier nursing assistance robots, Paro has been commercialized and operationalized across diverse care settings, particularly within ICU care, in multiple countries [42]. While it is designed to provide cognitive stimulation, no definitive evidence has surfaced indicating immediate or long-term cognitive improvement after Paro intervention. However, we cannot overlook its positive psychological impact in several settings. Its widespread application and accessibility also serve as a blueprint for the development of companion robots [43].

Beyond the 4 nursing assistance robots previously mentioned, a diverse range of technologies is under development aimed at enhancing nursing practice. These technologies are designed to assist with various tasks, including feeding patients, transporting patients, bathing, providing emotional support, administering medication, and facilitating other daily activities. Notable examples include the Care-O-bot, a compact, highly integrated service robot that primarily functions as a household assistant [44]. Another example is "RI-MAN," a soft-bodied robot designed to safely lift adults [45]. While these robots undertake certain nursing responsibilities and are potentially beneficial to the general population, they were not included in the scope of this review. This decision reflects our focus on specific categories of medical robot modules directly applicable to critical care and ICU settings, thereby delineating a clear boundary for our review's content and objectives.

The extensive research by Locsin [93] underscores a future where the integration of nursing with AI technology is not just a possibility but an inevitability. This projection implies that, as technology evolves, nurses will need to become proficient in advanced technological tools, ensuring that their participation in clinical practices is grounded in informed consent. In the nursing practices of the future, technology will serve to augment human nursing activities. This includes leveraging technology for predictive interventions to enhance nursing efficiency. However, it is critical to note that human cognition and emotional intelligence will remain at the heart of technological care. This integration suggests a model of care where technology and human elements complement each other, ensuring that patient care remains personalized, empathetic, and efficient, thereby reflecting the intrinsic values of nursing while embracing the advancements of AI technology.

Thus, the future of nursing practice is envisioned as a symbiotic relationship between human nurses and medical robots. This partnership aims not merely at task completion but also at leveraging sophisticated technology to deepen patient understanding and enhance care, thereby elevating the standard of nursing services provided. Human nurses are poised to maintain their pivotal role in health care, with person-centered care continuing to serve as the foundational principle of nursing. Through this fusion of technology and human touch, nursing will evolve, ensuring that care remains empathetic, responsive, and fundamentally human at its core.

Rehabilitation Assistance Robots

Most patients who are critically ill, confined to bed rest for extended periods, exhibit a high incidence of ICU-acquired weakness, significantly influencing patient prognoses [94]. Rehabilitation therapy, a potential solution to this issue, can notably enhance patient outcomes and quality of life. Numerous studies have attested to the pivotal role of early rehabilitation in improving patient prognoses and ensuring quality of life. Rehabilitation robots are devised to aid the restoration of impaired sensory, motor, and cognitive skills [95]. The deployment of these robots can relieve physicians of strenuous training tasks, allowing for the analysis of robot-derived data during training sessions to evaluate patient rehabilitation status. It has been established that intensive locomotor training can affect improvements in walking function for patients with movement disorders after stroke or spinal cord injury (SCI) [96-98]. Rehabilitation robots facilitate more extensive and intensive training compared to traditional therapeutic methods. Moreover, rehabilitation robots typically engage and challenge patients to meet objectives interactively. Although their application within the ICU is currently minimal, these robots hold promising potential in assisting early rehabilitation efforts for patients in the ICU. There are presently several commercially

XSL•FO
RenderX

available and experimental-stage noncommercial rehabilitation robots reported in the literature. Some cases are presented in the following paragraphs.

First, Lokomat (Hocoma, Inc) is a sophisticated robotic system designed for gait rehabilitation. It targets patients who exhibit locomotor anomalies induced by brain injury, spinal cord damage, neurological disorders, muscular injuries, and orthopedic diseases, facilitating improved mobility in patients who are neurologically compromised [46,47]. The Lokomat system comprises a structural frame affixed to a treadmill, featuring a load-bearing arrangement such as a gait orthosis. When attached to the patient, it modulates hip and knee joint movements to generate a predefined gait pattern [48]. Regular Lokomat-assisted training has been effectively used in rehabilitating patients who were comatose with cerebral hemorrhage. Following a 4-month therapeutic regimen, patients exhibited remarkable improvements in the severity of their comatose state (as per the Glasgow Coma Scale score of 4), an extension in walking duration from 15 to 32 minutes, and recuperation of eye and joint movement [49].

In a rigorous evaluation, Chillura et al [47] used Lokomat in a 6-month intensive conventional rehabilitation therapy. The outcome was a marked improvement in muscle strength (42/60), physical and mental independence (80/126), and 6-minute walk distance (47 m) in patients in the ICU with acquired weakness. A separate prospective study corroborates the potential of Lokomat in ameliorating patient rehabilitation after stroke [50]. A randomized single-blind, parallel-group clinical trial (40 participants per group) showed that training with the Lokomat system for 3 to 6 months after lesion can improve the walking ability of patients with incomplete SCI. This improvement was manifested in increased walking endurance and enhanced strength in the lower limbs [51]. Furthermore, 6 additional randomized controlled trials using the Lokomat system corroborated these findings [52-57], underscoring the consistency and reliability of the Lokomat as an effective rehabilitation tool for improving mobility and strength in individuals with incomplete SCI. This body of evidence strongly supports the Lokomat system's role in advancing the recovery process for patients with SCI, offering them a viable pathway to regain mobility and improve their quality of life.

Currently, advancements in robot-assisted lower-limb rehabilitation have led to the development of 3 main types of devices: exoskeleton, end effector, and portable powered robotic exoskeletons. End-effector devices interact directly with the patient's lower limbs, applying force, assisting movement, or guiding patients through specific movement patterns, exemplified by the "G-EO-Systems" [99] and "Haptic Walker" [100]. Exoskeleton-type devices, exemplified by "LOPES" from Delft University of Technology in the Netherlands [101], encase the patient's legs and provide support and movement assistance directly aligned with the limb's natural biomechanics. Among these, Lokomat stands out due to its superior customization capabilities, safety, comfort, and integration of virtual reality (VR). It also boasts features such as participation detection and motivational elements to engage users more effectively in their rehabilitation process. With 651 institutions worldwide adopting Lokomat, its widespread use underscores its significant advantages and effectiveness in facilitating the rehabilitation of patients with lower-limb impairments [58].

Such findings hint toward a broad applicability of Lokomat in ICU settings, enabling both early and post-ICU rehabilitation to be personalized and intelligent. As AI and machine learning continue to advance, the Lokomat system's ability to analyze patients' gait data and physiological indicators to offer precise rehabilitation strategies and real-time feedback is becoming a reality. Leveraging VR technology may render the rehabilitation training environment for patients in the ICU more captivating and web-based, fostering increased patient engagement and positivity during the recuperative process. Furthermore, integration with biometric sensors and remote monitoring technologies could feasibly enable real-time tracking and remote guidance of patients' rehabilitation progress, heralding the future direction of Lokomat's potential applications.

Second, aside from lower-limb or gait assistance rehabilitation robots, research into upper-limb rehabilitation robotics remains an active field of investigation. ArmeoSpring (Hocoma, Inc) is a rehabilitative exoskeleton that stabilizes the arm via a fixed frame. It is a passive exoskeleton that uses adjustable springs to deliver gravitational compensation for the patient, thus allowing the individual undergoing rehabilitation to concentrate solely on executing requisite tasks. Despite its passive nature, the ArmeoSpring exoskeleton is laden with sensors, enabling its application as an evaluation instrument for assessing the capacity and range of the user's arm [59]. These sensors also facilitate interactive training and integration with VR, enabling patients to simulate task-oriented motor exercises within a virtual learning environment on a computer screen, providing auditory and visual performance feedback during and after interaction.

A 5-year randomized controlled trial involving a subacute stroke program evaluated 215 patients with stroke with moderate to severe arm injuries who were undergoing rehabilitation therapy [60]. It appeared that there was no substantial difference in functional upper-extremity improvement with ArmeoSpring robotic intervention, although sensorimotor scores did demonstrate enhancement (13.32 vs 11.78). This equivocal result could be ascribed to a lack of sequencing of early testing interventions [61]. Conversely, another clinical trial involving a patient with mild to moderate hemiparesis demonstrated significant improvement in upper-extremity arm motion after 4 months of ArmeoSpring treatment [62]. It remains uncertain whether robotic-assisted rehabilitation definitively surpasses conventional physical therapy; rather, it appears to offer advantages to traditional treatments. In addition, in contrast to conventional physical therapy, careful patient selection is essential for robotic-assisted interventions. Factors such as the degree of functional impairment, age, disease duration, and cognitive levels play significant roles in this selection process. A study focusing on the evaluation of upper-limb movement parameters in patients after a stroke using the ArmeoSpring demonstrated its effectiveness in reliably and sensitively assessing motor impairments and the influence of therapeutic interventions on the motor learning process. The research highlighted the device's potential as a valuable assessment tool for quantifying sensorimotor disorders in the upper limb [59].

XSL•FO
RenderX

Another type of end-effector robot, the "MIT-MANUS" (Massachusetts Institute of Technology, United States) [63], trains the upper limb by applying force at a single point on the patient's arm. However, due to its linkage mechanism, the robot has a long mechanical arm and low torque output capacity, making it difficult to perform tasks requiring high-load resistance training. In addition, it has a large footprint. It lacks the ability to perform complex movements of the upper-limb joints, similar to the ArmeoSpring. Therefore, we did not include the MIT-MANUS in this review.

Third, patients in the ICU not only grapple with their primary illness but also frequently experience newly acquired long-term physical, psychological, and cognitive impairments, collectively referred to as post–intensive care syndrome [102]. Given the inconsistency in results from various screening tools, there is an urgent need for an objective, comprehensive, simplified, and unified assessment tool. The Kinesiological Instrument for Normal and Altered Reaching Movements (KINARM; Kingston, Ontario, Canada) is a robotic research tool specifically designed to execute quantitative neurological evaluations of sensorimotor, proprioceptive, and cognitive brain functions. It comprises a wheelchair and an upper-extremity exoskeleton tailored to patients based on their physical specifications. The KINARM permits researchers to gauge the coordination of limbs across multiple joints while also precisely measuring the joint-specific force exerted by the patient during task execution. The precision of this tool eliminates the subjectivity typically intrinsic to physiotherapeutic assessments of neurological status, such as muscle tone, spasticity, proprioception, and others [64].

A total of 104 patients in the ICU underwent sensorimotor and neurocognitive assessments using the KINARM 3 and 12 months after discharge. The team then performed a series of kinase evaluations on stroke survivors 3 and 12 months after discharge in patients who were critically ill and receiving acute renal replacement therapy and obtained a 0.3 correlation (90% strength in 89 patients) between regional cerebral oxygen saturation (a surrogate marker of cerebral autonomic regulation) and delirium in patients who were critically ill through KINARM scoring [65]. The tool has also been used to assess the correlation between brain tissue oxygenation (a surrogate marker of brain perfusion) during the acute phase of critical illness (ie, 24 hours) and long-term neurological dysfunction [66,103], as well as to evaluate sensorimotor deficits in patients with stroke and traumatic brain injury [67,104]. The KINARM provides objective and quantifiable data for sensorimotor and neurocognitive functions in ICU survivors. As a diagnostic and assessment tool, it aids rehabilitation and supports ICU survivors in regaining autonomy and independence in their daily lives. Despite its limitations, efforts must continue to enhance mobility and portability, broaden applicability and scope, and reduce cost and operational complexity.

Compared to traditional physiotherapy, robot-assisted rehabilitation based on AI and VR offers patients more intensive, systematic, repetitive, and task-oriented rehabilitation training, which plays a crucial role in promoting the process of functional recovery. Although many studies have shown that robot-assisted rehabilitation can effectively enhance the rehabilitation effect, a review of the clinical application of stroke rehabilitation points out that robot-assisted therapy has not shown obvious advantages in improving the motor function of patients with stroke [68]. Compared with traditional training or stand-alone training, its effect on the rehabilitation of patients with chronic stroke is still questionable. Similarly, the effectiveness of using exoskeleton devices for upper-limb motor function training also lacks sufficient evidence [105]. Therefore, at the current stage, robot-assisted rehabilitation therapy should be considered as a supplement to traditional physiotherapy, not a replacement. There is also no clear evidence to show that robot gait training can outperform traditional physical therapy when applied alone to patients with chronic stroke [106]. On the basis of the existing evidence, we can conclude that robot-assisted rehabilitation therapy can improve the motor function of patients needing rehabilitation and serve as an additional treatment intervention in combination with traditional rehabilitation therapy. However, with the further development of AI and machine learning technology in the future, we expect robot-assisted rehabilitation therapy to have greater development potential.

On the other hand, we must acknowledge that the current rehabilitative assistance robots face 3 core challenges: energy endurance, comfort assurance, and cost control. First, we need to change the existing endurance mode and adopt more effective energy supply methods. Second, we need to solve the comfort issues that may arise during the use of robots, such as blood circulation problems and muscle deformation that may be caused by wearing methods. Finally, we need to focus on controlling costs so that all patients who need rehabilitation can afford it.

Telepresence Robots

Telepresence represents a potential avenue for enhancing information accessibility for providers, encompassing aspects such as patients' visual and auditory feedback, bedside care, and vital sign data facilitated by remote monitoring or telechecking [107]. Given that most patients in the ICU are susceptible to unpredictable conditions, there is an acute need for swift identification and prompt response during emergent situations. The significance of telepresence robots lies in their ability to deliver expert health care services over distances, effectively mitigating the need for colocation of physicians and patients. This approach greatly augments the accessibility of health care services for patients in remote areas. Moreover, it potentially eradicates the likelihood of infectious disease transmission between patients and health care professionals [108]. Using AI and human-machine interaction, these telepresence robots supplement diagnostic and therapeutic processes via medical professionals' expertise, thereby enhancing the exchange of visual and electronic information between the patient and health care staff [109]. To illustrate this, we provide the following example.

InTouch Health Remote Presence-7 (RP-7; InTouch Health Systems) is a real-time audiovisual robotic telepresence system that provides communication among patients, hospital staff, and remote physicians [33,69,110]. Remote assessors used the RP-7 robot end point to conduct their clinical coma evaluations [69]. Compared with the total scores on the Glasgow Coma Scale or Full Outline of Unresponsiveness of the remote physician evaluators, the RP-7 robotic system had a similar

XSL•FO
RenderX

score (difference in scores of 0.25 and 0.40, respectively), and it can serve as a reliable scoring system to help evaluate patients in a coma [69]. In an ICU setting, the RP-7 assisted in the assessment of increased efficiency, care coordination, and throughput, with a decrease in patient ICU stay (–0.8 days), an increase in hospital discharges (+11%), and a significant decrease in the number of unexpected events (–1.2 days) [70,109]. Regarding the team cooperation ability (attitude, behavior, and cognition) of clinicians, the use of the RP-7 maintained the cooperation, trust, communication, and psychological safety of the team [107]. However, the RP-7 does not enhance collaboration between nurses and physicians in patient care decisions as compared to traditional telephone night checks [111].

A year later, the InTouch Vita, developed by InTouch Health and iRobot, was found to improve the independence of remote clinicians in managing patient care [71,112,113]. The Vita has an improved navigation system with an autopilot feature that enables remote service providers to directly control or automatically direct it to a predetermined location for improved efficiency [113]. In addition, the product provides real-time clinical access to patient data and has been cleared by the Food and Drug Administration for active patient monitoring that may be needed for immediate clinical action [72]. The Remote Presence Virtual + Independent Telemedicine Assistant (RP-VITA) also has an iPad interface that allows pilots to browse quickly and easily [114]. Unlike the RP-7, which must be powered using an active joystick, the RP-VITA only needs a mobile phone to log into the system and issue commands verbally to move the robot around to the designated location to complete the task [71]. To enable the robot to accurately follow the target character and maintain the corresponding safe distance and speed, Long et al [73] used the improved Gaussian filter algorithm to estimate and correct the centroid of the human body in real time, effectively improving the stability and safety of human tracking.

As a remote, virtual, and independent telemedicine assistant, the RP-VITA enables physicians to directly interact with patients from anywhere in the world. This technological advancement effectively transcends the traditional barriers of physical and biological constraints commonly encountered in ICU settings, where the immediacy of medical services is crucial yet often challenging to maintain. With the RP-VITA, physicians no longer need to undertake urgent commutes to the ward; instead, they can simply access the robot platform from the comfort of their home, offering a solution that is not only more expedient but also significantly more convenient [74]. This achieves truly meaningful health care services, delivering the right expertise to the right place at the right time to do the right thing at the right price. The RP-VITA exemplifies the future of health care delivery, emphasizing efficiency, accessibility, and the strategic allocation of medical expertise.

Second, Stevie has a stethoscope port and a high-definition pan-tilt-zoom camera, which can relay information during an examination of a patient and help physicians identify illnesses and diseases in the ICU [75]. It was developed by a research team from Trinity College, Dublin, Ireland, and is currently in use at Steve Biko Academic Hospital, South Africa, in the second design iteration [76]. It was designed to be neutral so as to avoid perceptions of gender, race, and age [115]. The upper body of the robot consists of a humanoid head (digital display), trunk, and 2 short arms [116]. A humanoid form makes robots more acceptable [117]. The control system gives the user the ability to operate the robot remotely, including controlling the robot voice and media volume as well as motion and motion. In addition, Stevie's humanoid facial expressions convey clear emotions [76]. Humanlike limbs can be used to form intuitive gestures, emphasize emotional states, and direct attention, conveying more information than facial expressions [116]. To provide static stability, the robot is equipped with wheels that support omnidirectional, allowing it to maneuver smoothly in any direction without shifting its base. This feature is essential for navigating complex environments and enhancing operational efficiency. Stevie is known as the "most popular baby" for the ICU team [115].

Third, MGI Ultrasound System-Remote 3 (MGIUS-R3; MGI Tech Co, Ltd), a robot-assisted teleultrasound diagnostic system, has considerable application value in the ICU [77,118]. It combines a robotic arm, an ultrasound imaging system, and audiovisual communication for remote manipulation, allowing physicians to manipulate the robotic arm and adjust parameters outside the ICU room for remote ultrasound workups via the fifth-generation (5G) network technology for real-time transmission of audio, video, and ultrasound images [78].

It has been used as a long-range ultrasound device to fight the COVID-19 outbreak in many places in China such as Wuhan [79]. Powerful cloud data transfer rates (up to 10 GB per second) with management mode enable high-definition image capture and high-quality information transfer for large data set sharing. At a distance of 700 or even 1479 km (between the isolation ward of Zhejiang Provincial People's Hospital and Central South Hospital of Wuhan University in Hubei Province and between Wuhan city in Hubei Province and Sanya city in Hainan Province), it can successfully conduct remote ultrasound examination of the lung and other parts of the patient with the quality meeting the requirements for clinical diagnosis [80,119]. A double-blind diagnostic trial of COVID-19 in 22 cases showed that the diagnostic accuracy of the MGIUS-R3 ultrasound for positive lesions was good (93%), which could replace the traditional scanning method to some extent (difference: P=.09) [77]. Another study also successfully performed ultrasound examination of the liver, gallbladder, pancreas, spleen, and kidney in 32 patients, and the high-quality image (average score of 4.73) met the requirements for remote ultrasound diagnosis [78]. Overall, the MGIUS-R3 is noninvasive and repeatable, reduces the risk of infection for patients and physicians, increases safety, and is highly feasible in the ICU. However, all remote ultrasound procedures and communications for the robot are based on 5G networks, which require stable network support throughout the procedure [119,120].

Through the network, remote ultrasound robots transmit ultrasound images of challenging cases from remote areas to tertiary hospitals, where consulting experts diagnose and analyze them, providing decision feedback. This provides an effective solution for situations with limited medical resources, lack of expertise, and high risk of infection.

XSL•FO
RenderX

Fourth, the intrinsic characteristics of the ICU, despite their provision of the highest level of medical care and reliance on advanced organ-support therapies, inherently limit the ability to facilitate real-time visits and communication between patients and their family members. This restriction not only affects the emotional well-being of patients due to a lack of verbal encouragement and emotional support from loved ones but also poses a challenge to patient recovery processes within the high-stress environment of the ICU. For the safety and emotional needs of patients and their families, as well as to achieve the goal of resource optimization, visiting robots have emerged. In 2021, the first 5G+ medical robot+VR visiting system was officially launched in the stroke ICU ward of West China Hospital of Sichuan University. The system successfully enables remote visits for family members at designated hospital locations equipped with matching VR glasses, allowing family members to interact with patients in the ICU in real-time through 2-way communication, facilitating an immersive visitation experience at the bedside [121]. The family members only need to wear VR glasses in the designated place of the hospital and operate the robot remotely through a computer, iPad, or mobile phone. The robot will automatically go to the patient bed designated to be visited after receiving the instruction. 5G networks with high-speed and low-latency transmission features (speeds of 10 to 30 Gbps) transmit back real-time 8K (resolutions of up to 7680 × 4320 pixels) ultra-high-definition full-motion video, enabling families to "physically" visit a patient's bedside [81,121].

The introduction of visiting robots represents a significant step forward in addressing a previously underappreciated aspect of health care—the impact of patient relationships, particularly with family members, on the process of disease recovery. By facilitating connections that were once hindered due to logistical, health, or institutional barriers, these robots enhance the humanization of health care services. They embody an innovative shift in the delivery models of health care services, ensuring that emotional support and the therapeutic benefits of family presence are not overlooked in patient care. This innovation underscores the importance of integrating technological advancements with the core values of empathy, care, and support, thereby enriching the patient experience.

Logistics and Disinfection Robots

Overview

The number of medical interventions in ICUs is larger than that in the general ward, and they are more invasive. In addition, the physiological condition of patients who are critical is often fragile, which makes patients in the ICU particularly vulnerable to iatrogenic injury [122,123]. At the same time, ICU health care workers face a high risk of infection [124]. The presence of logistics and disinfection robots avoids the spread of viruses, ensures clean areas for clinicians and patients, and minimizes the risk of infection for medical staff. Logistics and disinfection robots could have also worked in this application in a hospital setting during the COVID-19 pandemic, focusing on these same tasks and supporting patient management [125]. The common forms of logistics and disinfection robots are transport robots and infection control robots, which can solve the problem of ICU staff shortages and share heavy and tedious work. A brief description follows.

Transport Robot HelpMate (HelpMate Robotics Inc)

The transportation of goods and patients is an indispensable part of intensive care. However, a large amount of reciprocating work will consume most of the physical strength of nurses, long-term handling of patients will also cause serious lumbar muscle injury in nurses, and unarmed handling may cause secondary injuries to patients. HelpMate is a mature solution that provides autonomous transportation of materials and supplies, with users dispatching tasks through their console interface, and autonomous operations with unsupervised navigation technologies, such as proximity sensors for obstacle avoidance and path planning for navigation [48]. In the ICU, the use of automated transport vehicles will increase efficiency and avoid the potential for cross-infection, especially during the COVID-19 lockdown [82].

As early as the 1980s, the object transfer robot HelpMate was used to carry medical supplies, meals, and experimental samples, among other things. HelpMate mainly navigates based on prestored maps and has a certain obstacle avoidance ability [83]. In 2003, the University of Maryland Medical Center began a pilot program to determine the logistical capabilities and functional utility of the automated pharmacy system II Robot-Rx (McKesson) robotic technology in the delivery of medications from satellite pharmacies to ICU patient care units [84].

In view of the situation that it is not suitable to change a patient's posture during the process of transfer, Osaka General Medical Center in Japan found that more than half of nurses were willing to use a robot for patient transfer after using the robot Coupling-Parallel Adaption Merged (bilateral transfer bed, mainly through a conveyor belt to complete patient transfer between the bed and stretcher), and nearly half of the patients showed no discomfort during transfer [85].

Infection Control Robots

Infection control robots could make ICU wards safer and cleaner than ever before [126,127]. They are a major step up from traditional human cleaning methods, which take more time, are less effective, and often miss crevices that can hide nasty bugs. This is especially important for ICUs and clean rooms for patients who are immunocompromised, and it is possible that automated cleaning will soon become standard practice throughout the health care system. The existing disinfection robots can be mainly divided into 2 groups: UV robots and hydrogen peroxide vapor robots.

The Xenex robots use pulsed xenon to create intense bursts of broad-spectrum UV light that can cut bacterial contamination by a factor of 20 and kill 95% of deadly pathogens. More than 100 hospitals now use Xenex robots [86].

An "EPS" logistics disinfection robot (Ipsen Smart Health Tech [Shenzhen] Co, Ltd), which has both transport and disinfection functions, has been officially on duty in the ICU of Central South Hospital (Wuhan, China), undertaking the drug distribution work of nurses from the station to the ICU wards of patients with severe COVID-19 [87]. At the same time, it can customize the disinfection time and route for high-frequency

XSL•FO
RenderX

areas of physician-patient activity and can carry out automatic UV disinfection.

The hydrogen peroxide vapor disinfecting robot combines a hydrogen peroxide device with a robot (Shanghai Jiao Tong University and Lingzhi Technology joint research and development, China) [88]. The disinfecting system inside the robot generates disinfecting gas and can realize autonomous navigation and autonomous movement in an unmanned environment. At present, this disinfection robot has been effectively used in the prevention and control of the COVID-19 pandemic [89]. It is mainly used in ICUs, negative pressure isolation wards, infectious wards, and other closed spaces requiring sterilization.

Moreover, the literature has reported a novel robot system capable of real-time air pathogen monitoring in ICUs. The system comprises an automatic guided vehicle, an air sample collector, and a pathogen detection system. By autonomously patrolling and collecting air samples, the robot uses biosensor technology to perform real-time detection of airborne pathogens. This detection process includes lysis of pathogen particles, amplification of target sequences, and sensitive detection via the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associate system. This robot system allows for real-time monitoring of airborne pathogens within the ICU, aiding in the prevention of hospital-acquired infections and reducing the burden of ICU management. Endowed with a high degree of autonomy and real-time responsiveness, this system provides a new avenue for infection control within ICUs [90].

The ICU is a highly specific and complex area for monitoring patients who are critically ill, requiring not only 24-hour continuous care but also more intensive, timely, and coordinated interventions. Although the introduction of AI robots into ICUs is still in the initial stage, their autonomy, ease of training, and strong adaptability enhance their performance and provide substantial assistance to medical staff. In particular, the COVID-19 outbreak has increased the demand for AI robots that are not fatigued or infected, which may reduce the fatigue of ICU medical staff, reduce medical errors, and improve patient safety. However, for example, monitoring robots for single-organ life-support devices, comprehensive cardiopulmonary-monitoring robots for patients with multiple organ failure, and robots that are flexible in dealing with various unexpected tasks arising during patient care have not yet appeared. Therefore, the demand and development potential of AI robots in ICUs is huge.

Challenges and Potential Solutions Related to the Use of AI Robots in ICUs

Overview

Despite significant advancements and widespread applications of AI robots across all processes within the ICU, they face a host of challenges due to the complexity and uncertainty of the real world, algorithmic limitations, and ethical and moral considerations. These challenges encompass safety, dignity, privacy, and questions of liability. In addition, the implementation of advanced technologies requires considerable investment, making cost-benefit considerations indispensable. Our objective is to develop AI robots for ICU application that uphold human values, demonstrate ethically sound behavior, and draw robust conclusions [128]. It is imperative to accurately recognize these extant issues and propose effective solutions.

The “Asilomar AI Principles,” signed at the 2017 Asilomar conference held in Asilomar, California, United States, call on global AI professionals to adhere to these principles to safeguard the interests and safety of humanity in the future [129]. The principles emphasize ethical standards and values, including privacy, security, fairness, and transparency. Although these principles only provide an ethical framework and guiding principles for AI development and do not offer specific policies, regulations, or standards, they still serve as reference guidelines for the development of the AI field.

Security Issues

With the continuous evolution of AI robotic technology, its capabilities and functions are constantly being enhanced. However, there is a lack of clear standards to define the meaning of safety and accuracy and evaluate its specific programs. Narrow technical approaches are insufficient to ensure the safety of AI robots, which must be considered within the broader sociotechnical context in which they operate [130]. Therefore, systems with moral and social reasoning capabilities are becoming increasingly important. In some cases, human involvement can serve as a constraint on robot design, especially in decisions involving life and death. However, in the realm of high-speed decision-making, robots require built-in moral and social reasoning capabilities [131]. In addition, the lack of sample size, heterogeneity of diseases, and complex operations lead to biases in AI algorithms, which cannot ensure the safety and effectiveness of robot treatments. These challenges should be evaluated alongside the risks already familiar to humans, allowing us to set realistic expectations and foresee significant advancements in the realm of robot safety.

In an effort to confront the emerging safety challenges within the AI domain, various countries have embarked on proactive measures aimed at addressing these potential concerns. Initiatives such as the AI Foundation Model working group established by the United Kingdom in April 2023 [132] and the series of AI white papers updated annually by the Chinese Association for Artificial Intelligence are significant steps toward understanding and addressing these issues [133]. Moreover, the global AI Safety Summit held in November 2023 marked a pivotal moment, with 28 countries coming together to sign the “Bletchley Declaration.” This declaration represents a unified commitment to scrutinize the risks associated with the frontiers of AI technology, such as natural language processing, computer vision, and reinforcement learning, with a specific focus on the development of large language models by leading companies such as OpenAI, Meta, and Google [134]. These concerted efforts underline a global recognition of the complexities and challenges posed by AI, as well as a determined move toward collaborative solutions. By focusing on risk assessment, ethical standards, and safety protocols, these initiatives highlight an international resolve to navigate the advancements of AI technology in a manner that not only

XSL•FO
RenderX

benefits human society but also safeguards against potential hazards.

Dignity Issues

Medicine has always been a humanistic science where physicians are expected to not only adopt a scientific attitude toward patients but also resonate emotionally with them, embodying empathetic care. Patients in the ICU typically exhibit both physical and psychological fragility, necessitating humanistic care and emotional support from medical staff. This cannot be substituted by a robot due to its "mechanical"; "pre-programmed"; and, thus, "neutral" way of interacting with patients. Emotional recognition technology can be incorporated into AI robotic systems, providing corresponding emotional support by recognizing patient emotions, for instance, through voice, facial expressions, or body language. AI robots often have an unequal communication with patients, leaning more toward 1-way output from the robot. Before engaging in the discourse on equality between humans and AI robots, it is crucial to address a foundational question: should AI be classified as a tool or an agent? This distinction becomes especially pertinent in the context of conversational AI robots. When perceived merely as tools, there is a risk of undervaluing the anthropomorphic attributes and functionalities they embody. Conversely, viewing them as agents presents its own set of challenges as they inherently lack humanlike qualities such as empathy, intentionality, and the capacity to bear responsibility [135]. This debate is not new but remains central to the evolving conversation around AI's role in society. Recent technological developments suggest that, by integrating natural language processing and voice recognition technologies, robots can become more anthropomorphic and capable of responding to patients' language needs, alleviating feelings of loneliness and neglect in patients in the ICU. For example, the integration of the popular ChatGPT with AI robots may enhance their linguistic potential. Intelligent interactive AI robots combining ChatGPT's linguistic skills with the computer vision and tangible abilities of robots could revolutionize the way humans interact with technology [136]. They could be better at navigating the subtleties of human interaction, boasting superior natural language generation capabilities.

On the other hand, the narrow technical AI safety field lacks ideological and demographic diversity, leading to a lack of breadth and rigor in knowledge. Moreover, practitioners in this field often come from White, male, and underrepresented groups, which is insufficient to meet the broad participation and shared human care needed for technological development, potentially leading to racial, gender, and other biases in technology application [137,138]. For instance, using Framingham Heart Study data to predict cardiovascular event risk in people of color may lead to both overestimation and underestimation of risk [139]. Similar racial biases may inadvertently be built into health care algorithms. Therefore, it is necessary to expand the space for broader participation, pursuing equality and common development from the source of technology, to avoid defects in AI robot products [140].

Privacy Issues

To enhance the monitoring of patients who are critically ill, robots are often equipped with surveillance apparatuses to log pertinent data and wirelessly transmit information. Such actions may infringe upon the privacy rights of patients, jeopardizing their confidentiality. Nonetheless, these features also hold merit in ensuring patient safety [141]. As AI robots proliferate across various sectors, including health care, concerns over privacy safeguards have been a recurrent topic of critique. It is imperative that AI robots adhere to privacy regulations when handling medical data, such as the General Data Protection Regulation that is prevalent in Europe [142]. Encryption techniques should be used during data collection and storage, narrowing the scope of data acquisition and ensuring anonymization. Of paramount importance is the establishment of stringent access control mechanisms, ensuring that data are accessible only to authorized personnel. Concurrently, it is vital to instate ethical guidelines and standards focused on privacy protection, which will dictate the conduct of AI robots and their data-handling procedures. All these regulations must ensure that individuals retain a voice over the collection, storage, and use of their information [143].

Attribution of Liability Issues

The integration of AI into robotic systems has rendered questions of accountability more intricate. Endowing robots with autonomy and decision-making capabilities stands as a primary objective of AI integration into robotics. However, given the current trajectory of technological advancements, intelligent analytics still bear systemic decision-making risks. When malfunctions occur, attributing responsibility for the robot's actions becomes contentious given that robots lack comprehension of reprimand, sanctions, accountability, or remorse [144,145]. Consequently, the liability arising from erroneous decisions made by robots can pose significant legal conundrums [141]. Accountability in computer science encompasses a multifaceted domain, probing the responsibility attribution across its creation, dissemination, and use phases [146]. Thus, devising a legal framework pertinent to AI robots is essential, delineating the responsibilities and obligations of robot manufacturers, operators, and users spanning every facet of robot design, production, operation, and use. Regulatory bodies also play a pivotal role as rigorous supervisory protocols and technical benchmarks are needed to scrutinize and accredit AI robot designs and operations. Penalties and sanctions become indispensable components of this framework. Introducing a robotic liability insurance system might also serve as an efficacious remedy to mitigate the damages and risks incurred by robots. Manufacturers, operators, and users could opt for such liability insurance, necessitating clear demarcations of responsibility. Meanwhile, both the United States and the European Union advocate for a focus on algorithm transparency and accountability, aiming to make their decision-making processes more transparent and comprehensible. This will facilitate a better assessment of responsibility and serve as a warning against outsourcing moral responsibility to algorithms [147]. This calls for collaboration between manufacturers and developers, with the National Institute for Health and Care Excellence in the United States requiring developers to program

XSL•FO
RenderX

code in an environment of technological feasibility and respect for intellectual property rights, ensuring reproducibility and error checking [148].

Addressing the issue of robotic accountability necessitates a holistic perspective, incorporating technical, legal, ethical, and societal dimensions; fostering collaborative mechanisms with multistakeholder participation to collectively bear the decision-making risks and responsibilities; and ensuring the utmost protection of the interests and safety of robot users and beneficiaries.

The United Kingdom has introduced the Code of Conduct for Data-Driven Health and Care Technology, adopting a "Regulation as a Service" model. This innovative approach ensures that regulatory checks are embedded at various stages throughout the AI development cycle, aiming to uphold high standards of safety, efficacy, and ethics in the creation and implementation of AI technologies in health care [149]. In June 2021, the US Government Accountability Office released an AI accountability framework covering 4 aspects: governance (promoting accountability by establishing processes to manage, operate, and oversee implementation), data (ensuring quality, reliability, and representativeness of data sources and processing), performance (producing results consistent with program objectives), and monitoring (ensuring reliability and relevance over time). This accountability framework sets principles and directions for future legislation and policy making and also serves as a model for the advancement of accountability systems in other countries [150].

Cost-Benefit Issues

Economic viability remains a pivotal facet in the societal integration of any nascent technology, with the deployment of AI robots in ICUs presenting intricate cost-benefit deliberations. Being an avant-garde technology, the initial research and development expenditures for AI robots are considerable. Once successfully developed and incorporated into ICU settings, they are further subjected to financial burdens, encompassing but not confined to the costs of system enhancements and upgrades; resource use during staff training and acclimatization; and investments necessitated by data security, personal dignity, privacy safeguards, and responsibility allocation issues. However, when observed from a utility standpoint, the application of AI robots bears significant value, especially for populous nations such as China. In the face of scarce grassroots medical resources, AI robots serve as catalysts, facilitating the dispersion of health care provision to underserved and remote locales, meeting the medical needs of a vast populace—this aligns with the principal objective underpinning this technological advancement. In addition, the integration of AI robots can effectively bridge disparities in health care accessibility between high- and low-income nations, substantially augmenting the societal benefits of this technology [151]. Therefore, despite the steep initial investments and continual operational expenses, in the broader spectrum of health care service provision and societal equity realization, AI robots undoubtedly offer pronounced advantages and value.

Discussion

Principal Findings

AI robots have firmly established their significance within intensive care, with their integration into the ICU regimen continually deepening. This paper delineates 5 distinct application domains of AI robotic systems, be it experimental or commercial, within the ICU, addressing both technical impediments and prospective research avenues while proposing potential remedial strategies. The trajectory for AI robots within the ICU setting is promising. Currently at the nascent phase of AI robot technological deployment, there remains an extensive scope of endeavors to be pursued and myriad challenges to be surmounted. As the propagation of pertinent technologies ensues, health care professionals should welcome such intelligent implementations with optimism, recognizing the present-day confines of AI apparatuses synergistically amalgamating human and system intellect, thereby maximizing the data analysis, promptings, and recommendation proficiencies of intelligent robots. Today's robotic entities coalesce seamlessly with AI, where sophisticated robotic systems, characterized by safety and flexibility melded with augmented computational proficiencies, yield invaluable big data insights. Simultaneously, anthropomorphic designs provide patients in the ICU with a more comforting medical experience.

"There are plenty of areas in critical care where it would be extremely helpful to have efficacious, fair, and transparent AI systems," notes Gary Weissman, assistant professor in pulmonary and critical care medicine at the University of Pennsylvania Perelman School of Medicine [152]. Similarly, Dr Brijesh Patel, an intensivist at the Royal Brompton Hospital in London, emphasizes that "intensive care is a specialty with special prospects for AI. After all, the ICU is a space where a large amount of data is routinely collected, making it an ideal place for deploying machine learning techniques" [14]. Dr Patel, who dedicates a considerable portion of his ward rounds to adjusting ventilator settings, points out that the continuous advancement in AI technology could automate such repetitive tasks.

However, there is a consensus among experts that AI is not poised to replace physicians entirely. Instead, it is seen as a tool to streamline certain tasks, enhancing efficiency where it is most needed. Aldo Faisal, a professor of AI and neuroscience at Imperial College London, emphasizes a balanced perspective on AI's role within health care teams. He advocates for a realistic understanding of AI's capabilities and limitations, suggesting that neither undue fear nor excessive reverence is helpful [153].

This paragraph presents a futuristic scenario representing the potential evolution and future blueprint of AI robotics in the ICU ward. Envision a scenario where a patient who is critically injured is admitted to the ICU and greeted by "IntelliGreet"—an intelligent reception robot that collates the patient's fundamental data and medical history, seamlessly completing admission formalities. The therapeutic reins are taken over by "MediBot," a state-of-the-art medical robot that executes an array of treatment procedures (ranging from drug administration and

XSL•FO
RenderX

wound care to life support) predicated on the patient's condition and physician's directives. The patient's nursing needs are catered to by "CareCompanion," an omnipresent nursing robot that meticulously monitors vital signs, offering essential care services (such as cleaning, feeding, and movement) while concurrently assuming the responsibility of reporting to physicians. As the patient recuperates under their meticulous care and nears discharge, "DischargeDuty" steps in, formulating discharge plans and subsequent treatment regimens based on physician prescriptions, facilitating communication with both the patient and their kin, and overseeing the discharge processes. Finally, the domestic care robot "HomeCareHelper" persists in its caregiving, administering medication reminders, monitoring patient health, and even offering rudimentary domestic aid for solitary individuals. All AI robots in this envisioned setting are interlinked via a cloud data platform, enabling real-time sharing of patient medical data, achieving cohesive and synergized medical service delivery. Reflecting upon the advancements in AI robotic technology over recent decades, it is undeniable that this vision is poised to materialize.

Limitations of This Review

This review acknowledges a number of limitations that could affect the interpretation and applicability of its findings. One significant concern is the potential for publication bias, a common issue in scientific literature where studies with negative results are less likely to be published. This could lead to an overrepresentation of positive findings in this review. In addition, despite efforts to mitigate bias by involving interdisciplinary experts and employing dual reviewers during the literature search and data collection phases, subjective biases could still influence the selection and interpretation of the studies.

Another challenge is the absence of a universally accepted classification system for ICU robotic systems. In response, our classification framework was developed based on expert opinions and existing literature, striving for as comprehensive and rational an approach as possible. Nevertheless, the potential for omissions exists given the rapidly evolving nature of technology and the diverse applications of robotics in critical care settings. These limitations highlight the need for ongoing research and critical evaluation of emerging technologies in health care, emphasizing the importance of transparency and methodological rigor in scientific reviews.

Conclusions

This scoping review comprehensively covered AI robots in the ICU, detailing the most widely used or newly developed robotic devices on the market. Robots in ICU wards are becoming valuable assistants to physicians and nurses. Although ethical and safety concerns remain unresolved in this field, these challenges are inevitable in the development of new technologies, and experts and developers are focusing on addressing them. Future research should focus on developing policies and regulations to prevent or resolve these issues, making AI robots an integral part of ICUs and other hospital wards.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

HK was responsible for hypothesis generation. HK and ZQ were responsible for the conception of this study. Y Li, MW, LW, YC, Y Liu, YZ, RY, MY, SL, and ZS contributed substantially to all aspects of the work, including literature review and manuscript drafting. FZ revised the manuscript. All authors contributed to the manuscript and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 551 KB - jmir_v26i1e54095_app1.pdf]

Multimedia Appendix 2
Classification of intensive care unit robots.
[DOCX File , 24 KB - jmir_v26i1e54095_app2.docx]

References

1. Bobrow DG, Stefik MJ. Perspectives on artificial intelligence programming. Science 1986 Feb 28;231(4741):951-957. [doi: 10.1126/science.231.4741.951] [Medline: 17740291]
2. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
3. Georgopoulos AP, Schwartz AB, Kettner RE. Response: neuronal coding and robotics. Science 1987 Jul 17;237(4812):301. [doi: 10.1126/science.237.4812.301] [Medline: 17772058]

XSL•FO
RenderX

4. McKendrick M, Yang S, McLeod GA. The use of artificial intelligence and robotics in regional anaesthesia. Anaesthesia 2021 Jan;76 Suppl 1:171-181 [FREE Full text] [doi: 10.1111/anae.15274] [Medline: 33426667]
5. Javaid M, Haleem A, Vaishya R, Bahl S, Suman R, Vaish A. Industry 4.0 technologies and their applications in fighting COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):419-422 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.032] [Medline: 32344370]
6. Yang GZ, J Nelson B, Murphy RR, Choset H, Christensen H, H Collins S, et al. Combating COVID-19-the role of robotics in managing public health and infectious diseases. Sci Robot 2020 Mar 25;5(40):eabb5589. [doi: 10.1126/scirobotics.abb5589] [Medline: 33022599]
7. Fong SJ, Dey N, Chaki J. Artificial Intelligence for Coronavirus Outbreak. Singapore, Singapore: Springer; 1999.
8. Howard J. Artificial intelligence: implications for the future of work. Am J Ind Med 2019 Nov;62(11):917-926. [doi: 10.1002/ajim.23037] [Medline: 31436850]
9. Agarwal S, Punn NS, Sonbhadra SK, Tanveer M, Nagabhushan P, Pandian KK, et al. Unleashing the power of disruptive and emerging technologies amid COVID-19: a detailed review. arXiv. Preprint posted online on May 23, 2020.
10. Siciliano B, Khatib O. Robotics and the handbook. In: Siciliano B, Khatib O, editors. Springer Handbook of Robotics. Cham, Switzerland: Springer; 2016:1-6.
11. Marshall JC, Bosco L, Adhikari NK, Connolly B, Diaz JV, Dorman T, et al. What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine. J Crit Care 2017 Feb;37:270-276. [doi: 10.1016/j.jcrc.2016.07.015] [Medline: 27612678]
12. Clancy TR. Artificial intelligence and nursing: the future is now. J Nurs Adm 2020 Mar;50(3):125-127. [doi: 10.1097/NNA.0000000000000855] [Medline: 32068622]
13. Hanson CWIII, Marshall BE. Artificial intelligence applications in the intensive care unit. Crit Care Med 2001 Feb;29(2):427-435. [doi: 10.1097/00003246-200102000-00038] [Medline: 11269246]
14. Burki TK. Artificial intelligence hold promise in the ICU. Lancet Respir Med 2021 Aug;9(8):826-828. [doi: 10.1016/s2213-2600(21)00317-9]
15. Gutierrez G. Artificial intelligence in the intensive care unit. Crit Care 2020 Mar 24;24(1):101 [FREE Full text] [doi: 10.1186/s13054-020-2785-y] [Medline: 32204716]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Vagvolgyi BP, Khrenov M, Cope J, Deguet A, Kazanzides P, Manzoor S, et al. Telerobotic operation of intensive care unit ventilators. Front Robot AI 2021 Jun 24;8:612964 [FREE Full text] [doi: 10.3389/frobt.2021.612964] [Medline: 34250025]
19. Ehmke C, Kopp P, Karakikes M, Jahn R, Huang Y, Geiger J. A remote controller accessory for ventilators. GitHub. 2021. URL: https://github.com/claasehmke/Remote-Controller-Accessory-for-Ventilators [accessed 2024-04-29]
20. Song C, Yang G, Park S, Jang N, Jeon S, Oh SR, et al. On the design of integrated tele-monitoring/ operation system for therapeutic devices in isolation intensive care unit. IEEE Robot Autom Lett 2022 Oct;7(4):8705-8712. [doi: 10.1109/lra.2022.3188431]
21. Hemmerling TM, Arbeid E, Wehbe M, Cyr S, Taddei R, Zaouter C. Evaluation of a novel closed-loop total intravenous anaesthesia drug delivery system: a randomized controlled trial. Br J Anaesth 2013 Jun;110(6):1031-1039 [FREE Full text] [doi: 10.1093/bja/aet001] [Medline: 23427212]
22. Hemmerling TM. Automated anesthesia. Curr Opin Anaesthesiol 2009 Dec;22(6):757-763. [doi: 10.1097/ACO.0b013e328332c9b4] [Medline: 19812483]
23. Wehbe M, Arbeid E, Cyr S, Mathieu PA, Taddei R, Morse J, et al. A technical description of a novel pharmacological anesthesia robot. J Clin Monit Comput 2014 Feb;28(1):27-34. [doi: 10.1007/s10877-013-9451-8] [Medline: 23771652]
24. Naidoo S. Understanding artificial intelligence and machine learning in anaesthesia. School of Clinical Medicine, Discipline of Anaesthesiology and Critical Care. 2021. URL: https://anaesthetics.ukzn.ac.za/wp-content/uploads/2021/01/2021-15-Jan-Artificial-Intelligence-in-Anaesthesia-S-Naidoo.pdf [accessed 2024-04-29]
25. Zaouter C, Hemmerling TM, Lanchon R, Valoti E, Remy A, Leuillet S, et al. The feasibility of a completely automated total IV anesthesia drug delivery system for cardiac surgery. Anesth Analg 2016 Oct;123(4):885-893. [doi: 10.1213/ANE.0000000000001152] [Medline: 27644009]
26. Hemmerling TM, Charabati S, Zaouter C, Minardi C, Mathieu PA. A randomized controlled trial demonstrates that a novel closed-loop propofol system performs better hypnosis control than manual administration. Can J Anaesth 2010 Aug;57(8):725-735. [doi: 10.1007/s12630-010-9335-z] [Medline: 20533013]
27. Hemmerling TM, Taddei R, Wehbe M, Morse J, Cyr S, Zaouter C. Robotic anesthesia - a vision for the future of anesthesia. Transl Med UniSa 2011 Sep;1:1-20 [FREE Full text] [Medline: 23905028]
28. Tighe PJ, Badiyan SJ, Luria I, Lampotang S, Parekattil S. Robot-assisted airway support: a simulated case. Anesth Analg 2010 Oct;111(4):929-931 [FREE Full text] [doi: 10.1213/ANE.0b013e3181ef73ec] [Medline: 20870983]

XSL•FO
RenderX

29. Sprung CL, Ricou B, Hartog CS, Maia P, Mentzelopoulos SD, Weiss M, et al. Changes in end-of-life practices in European intensive care units from 1999 to 2016. JAMA 2019 Nov 05;322(17):1692-1704 [FREE Full text] [doi: 10.1001/jama.2019.14608] [Medline: 31577037]
30. Levitov A, Frankel HL, Blaivas M, Kirkpatrick AW, Su E, Evans D, et al. Guidelines for the appropriate use of bedside general and cardiac ultrasonography in the evaluation of critically ill patients-part II: cardiac ultrasonography. Crit Care Med 2016 Jun;44(6):1206-1227. [doi: 10.1097/CCM.0000000000001847] [Medline: 27182849]
31. Jung J, Kim J, Kim S, Kwon WY, Na SH, Kim KS, et al. Application of robot manipulator for cardiopulmonary resuscitation. In: Proceedings of the 2016 International Symposium on Experimental Robotics. 2016 Presented at: ISER '16; October 3-6, 2016; Tokyo, Japan p. 266-274 URL: https://link.springer.com/chapter/10.1007/978-3-319-50115-4_24 [doi: 10.1007/978-3-319-50115-4_24]
32. Suh GJ, Park J, Lee JC, Na SH, Kwon WY, Kim KS, et al. End-tidal CO-guided automated robot CPR system in the pig. Preliminary communication. Resuscitation 2018 Jun;127:119-124. [doi: 10.1016/j.resuscitation.2018.04.011] [Medline: 29665427]
33. Garingo A, Friedlich P, Tesoriero L, Patil S, Jackson P, Seri I. The use of mobile robotic telemedicine technology in the neonatal intensive care unit. J Perinatol 2012 Jan;32(1):55-63. [doi: 10.1038/jp.2011.72] [Medline: 21617643]
34. Williams N, MacLean K, Guan L, Collet JP, Holsti L. Pilot testing a robot for reducing pain in hospitalized preterm infants. OTJR (Thorofare N J) 2019 Apr;39(2):108-115. [doi: 10.1177/1539449218825436] [Medline: 30770034]
35. Balter ML, Chen AI, Maguire TJ, Yarmush ML. The system design and evaluation of a 7-DOF image-guided venipuncture robot. IEEE Trans Robot 2015 Aug;31(4):1044-1053 [FREE Full text] [doi: 10.1109/TRO.2015.2452776] [Medline: 26257588]
36. Chen A, Nikitczuk K, Nikitczuk J, Maguire T, Yarmush M. Portable robot for autonomous venipuncture using 3D near infrared image guidance. Technology (Singap World Sci) 2013 Sep;1(1):72-87 [FREE Full text] [doi: 10.1142/S2339547813500064] [Medline: 26120592]
37. Lippi G, Cadamuro J. Novel opportunities for improving the quality of preanalytical phase. A glimpse to the future? J Med Biochem 2017 Oct;36(4):293-300 [FREE Full text] [doi: 10.1515/jomb-2017-0029] [Medline: 30581325]
38. Chen AI, Balter ML, Maguire TJ, Yarmush ML. Real-time needle steering in response to rolling vein deformation by a 9-DOF image-guided autonomous venipuncture robot. Rep U S 2015;2015:2633-2638 [FREE Full text] [doi: 10.1109/IROS.2015.7353736] [Medline: 26779381]
39. Tan W, Shen C, Luo Y, Zhu H, Ma T, Dong D, et al. Development of motion unit of simulated intelligent endotracheal suctioning robot. Zhongguo Yi Liao Qi Xie Za Zhi 2019 Jan 30;43(1):17-20. [doi: 10.3969/j.issn.1671-7104.2019.01.0005] [Medline: 30770684]
40. Kostandy RR, Ludington-Hoe SM. The evolution of the science of kangaroo (mother) care (skin-to-skin contact). Birth Defects Res 2019 Sep 01;111(15):1032-1043. [doi: 10.1002/bdr2.1565] [Medline: 31419082]
41. Holsti L, MacLean K, Oberlander T, Synnes A, Brant R. Calmer: a robot for managing acute pain effectively in preterm infants in the neonatal intensive care unit. Pain Rep 2019;4(2):e727 [FREE Full text] [doi: 10.1097/PR9.0000000000000727] [Medline: 31041426]
42. Yu C, Sommerlad A, Sakure L, Livingston G. Socially assistive robots for people with dementia: systematic review and meta-analysis of feasibility, acceptability and the effect on cognition, neuropsychiatric symptoms and quality of life. Ageing Res Rev 2022 Jun;78:101633 [FREE Full text] [doi: 10.1016/j.arr.2022.101633] [Medline: 35462001]
43. Demange M, Pino M, Kerhervé H, Rigaud AS, Cantegreil-Kallen I. Management of acute pain in dementia: a feasibility study of a robot-assisted intervention. J Pain Res 2019;12:1833-1846 [FREE Full text] [doi: 10.2147/JPR.S179640] [Medline: 31289446]
44. Reiser U, Connette C, Fischer J, Kubacki J, Bubeck A, Weisshardt F. Care-O-bot® 3 - creating a product vision for service robot applications by integrating design and technology. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2009 Presented at: IROS '09; October 11-15, 2009; St. Louis, MO p. 1992-1998 URL: https://ieeexplore.ieee.org/document/5354526 [doi: 10.1109/iros.2009.5354526]
45. Odashima T, Onishi M, Tahara K, Takagi K, Asano F, Kato Y. A soft human-interactive robot RI-MAN. In: Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2006 Presented at: IROS '06; October 9-15, 2006; Beijing, China p. 1 URL: https://ieeexplore.ieee.org/abstract/document/4058352 [doi: 10.1109/iros.2006.282203]
46. Riener R, Lünenburger L, Maier I, Colombo G, Dietz V. Locomotor training in subjects with sensori-motor deficits: an overview of the Robotic Gait Orthosis Lokomat. J Healthc Eng 2010 Jun;1(2):197-216. [doi: 10.1260/2040-2295.1.2.197]
47. Chillura A, Bramanti A, Tartamella F, Pisano MF, Clemente E, Lo Scrudato M, et al. Advances in the rehabilitation of intensive care unit acquired weakness: a case report on the promising use of robotics and virtual reality coupled to physiotherapy. Medicine (Baltimore) 2020 Jul 10;99(28):e20939 [FREE Full text] [doi: 10.1097/MD.0000000000020939] [Medline: 32664093]
48. Dahl TS, Kamel Boulos MN. Robots in health and social care: a complementary technology to home care and telehealthcare? Robotics 2013 Dec 30;3(1):1-21. [doi: 10.3390/robotics3010001]
49. Rusek W, Adamczyk M, Baran J, Leszczak J, Pop T. Feasibility of Using Erigo and Lokomat in Rehabilitation for Patient in Vegetative State: A Case Report. J Clin Case Rep 2020;10(1318):1-4 [FREE Full text]

XSL•FO
RenderX

50. Uivarosan D, Tit DM, Iovan C, Nistor-Cseppento DC, Endres L, Lazar L, et al. Effects of combining modern recovery techniques with neurotrophic medication and standard treatment in stroke patients. Sci Total Environ 2019 Aug 20;679:80-87. [doi: 10.1016/j.scitotenv.2019.05.070] [Medline: 31100563]
51. Alcobendas-Maestro M, Esclarín-Ruz A, Casado-López RM, Muñoz-González A, Pérez-Mateos G, González-Valdizán E, et al. Lokomat robotic-assisted versus overground training within 3 to 6 months of incomplete spinal cord lesion: randomized controlled trial. Neurorehabil Neural Repair 2012;26(9):1058-1063. [doi: 10.1177/1545968312448232] [Medline: 22699827]
52. Esclarín-Ruz A, Alcobendas-Maestro M, Casado-Lopez R, Perez-Mateos G, Florido-Sanchez MA, Gonzalez-Valdizan E, et al. A comparison of robotic walking therapy and conventional walking therapy in individuals with upper versus lower motor neuron lesions: a randomized controlled trial. Arch Phys Med Rehabil 2014 Jun;95(6):1023-1031. [doi: 10.1016/j.apmr.2013.12.017] [Medline: 24393781]
53. Labruyère R, van Hedel HJ. Strength training versus robot-assisted gait training after incomplete spinal cord injury: a randomized pilot study in patients depending on walking assistance. J Neuroeng Rehabil 2014 Jan 09;11:4 [FREE Full text] [doi: 10.1186/1743-0003-11-4] [Medline: 24401143]
54. Tang Q, Huang Q, Hu C. Research on design theory and compliant control for underactuated lower-extremity rehabilitation robotic systems code: (51175368); 2012.01-2015.12. J Phys Ther Sci 2014 Oct;26(10):1597-1599 [FREE Full text] [doi: 10.1589/jpts.26.1597] [Medline: 25364122]
55. Shin JC, Kim JY, Park HK, Kim NY. Effect of robotic-assisted gait training in patients with incomplete spinal cord injury. Ann Rehabil Med 2014 Dec;38(6):719-725 [FREE Full text] [doi: 10.5535/arm.2014.38.6.719] [Medline: 25566469]
56. Duffell LD, Niu X, Brown G, Mirbagheri MM. Variability in responsiveness to interventions in people with spinal cord injury: do some respond better than others? Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:5872-5875. [doi: 10.1109/EMBC.2014.6944964] [Medline: 25571332]
57. Field-Fote EC, Lindley SD, Sherman AL. Locomotor training approaches for individuals with spinal cord injury: a preliminary report of walking-related outcomes. J Neurol Phys Ther 2005 Sep;29(3):127-137. [doi: 10.1097/01.npt.0000282245.31158.09] [Medline: 16398945]
58. Home page. Hocoma. URL: https://www.hocoma.com/us/solutions/lokomat/ [accessed 2023-03-05]
59. Brihmat N, Loubinoux I, Castel-Lacanal E, Marque P, Gasq D. Kinematic parameters obtained with the ArmeoSpring for upper-limb assessment after stroke: a reliability and learning effect study for guiding parameter use. J Neuroeng Rehabil 2020 Sep 29;17(1):130 [FREE Full text] [doi: 10.1186/s12984-020-00759-2] [Medline: 32993695]
60. Rémy-Néris O, Le Jeannic A, Dion A, Médée B, Nowak E, Poiroux É, REM Investigative Team*. Additional, mechanized upper limb self-rehabilitation in patients with subacute stroke: the REM-AVC randomized trial. Stroke 2021 Jun;52(6):1938-1947. [doi: 10.1161/STROKEAHA.120.032545] [Medline: 33910364]
61. Bernhardt J, Hayward KS. What is next after this well-conducted, but neutral, multisite trial testing self-rehabilitation approaches? Stroke 2021 Jun;52(6):1948-1950. [doi: 10.1161/STROKEAHA.121.034533] [Medline: 33910368]
62. Colomer C, Baldoví A, Torromé S, Navarro M, Moliner B, Ferri J, et al. Efficacy of Armeo® Spring during the chronic phase of stroke. Study in mild to moderate cases of hemiparesis. Neurologia 2013 Jun;28(5):261-267 [FREE Full text] [doi: 10.1016/j.nrl.2012.04.017] [Medline: 22727271]
63. Krebs HI, Ferraro M, Buerger SP, Newbery MJ, Makiyama A, Sandmann M, et al. Rehabilitation robotics: pilot trial of a spatial extension for MIT-Manus. J Neuroeng Rehabil 2004 Oct 26;1(1):5 [FREE Full text] [doi: 10.1186/1743-0003-1-5] [Medline: 15679916]
64. Wood MD, Maslove DM, Muscedere J, Scott SH, Boyd JG, Canadian Critical Care Trials Group. Robotic technology provides objective and quantifiable metrics of neurocognitive functioning in survivors of critical illness: a feasibility study. J Crit Care 2018 Dec;48:228-236. [doi: 10.1016/j.jcrc.2018.09.011] [Medline: 30243203]
65. Jawa NA, Holden RM, Silver SA, Scott SH, Day AG, Norman PA, et al. Identifying neurocognitive outcomes and cerebral oxygenation in critically ill adults on acute kidney replacement therapy in the intensive care unit: the INCOGNITO-AKI study protocol. BMJ Open 2021 Aug 17;11(8):e049250 [FREE Full text] [doi: 10.1136/bmjopen-2021-049250] [Medline: 34404711]
66. Wood M, Maslove D, Muscedere J, Scott S, Boyd J. Using robotic technology to quantify neurological deficits among survivors of critical illness: do they relate to brain tissue oxygen levels? a pilot study. Intensive Care Med Exp 2015 Oct 1;3(S1):1-2. [doi: 10.1186/2197-425x-3-s1-a730]
67. Debert CT, Herter TM, Scott SH, Dukelow S. Robotic assessment of sensorimotor deficits after traumatic brain injury. J Neurol Phys Ther 2012 Jun;36(2):58-67. [doi: 10.1097/NPT.0b013e318254bd4f] [Medline: 22592061]
68. Chang WH, Kim YH. Robot-assisted therapy in stroke rehabilitation. J Stroke 2013 Sep;15(3):174-181 [FREE Full text] [doi: 10.5853/jos.2013.15.3.174] [Medline: 24396811]
69. Adcock AK, Kosiorek H, Parikh P, Chauncey A, Wu Q, Demaerschalk BM. Reliability of robotic telemedicine for assessing critically ill patients with the full outline of UnResponsiveness Score and Glasgow Coma Scale. Telemed J E Health 2017 Jul;23(7):555-560 [FREE Full text] [doi: 10.1089/tmj.2016.0225] [Medline: 28085631]
70. McNelis J, Schwall GJ, Collins JF. Robotic remote presence technology in the surgical intensive care unit. J Trauma Acute Care Surg 2012 Feb;72(2):527-530. [doi: 10.1097/TA.0b013e31822f7d3b] [Medline: 22327990]

XSL•FO
RenderX

71. Pransky J. The Pransky interview: Dr Yulun Wang, founder and CEO of InTouch Health. Ind Robot 2015;42(5):381-385. [doi: 10.1108/ir-05-2015-0111]
72. Hand L. FDA clears telemedicine robot for use in hospitals. Medscape. 2013. URL: https://www.medscape.com/viewarticle/778672 [accessed 2023-05-19]
73. Long Y, Xu Y, Xiao Z, Shen Z. Kinect-based human body tracking system control of medical care service robot. In: Proceedings of the 2018 WRC Symposium on Advanced Robotics and Automation. 2018 Presented at: WRC SARA '18; August 15-19, 2018; Beijing, China p. 281-288 URL: https://ieeexplore.ieee.org/document/8584246 [doi: 10.1109/wrc-sara.2018.8584246]
74. Ackerman E. iRobot and InTouch health announce RP-VITA telemedicine robot. IEEE Spectrum. 2012. URL: https://spectrum.ieee.org/irobot-and-intouch-health-announce-rpvita-telemedicine-robot [accessed 2024-03-03]
75. Swanepoel L. Steve Biko academic hospital has robot for ICU. NuusFlits. 2021. URL: https://find-it.co.za/2021/07/07/steve-biko-academic-hospital-has-robot-for-icu/ [accessed 2024-02-28]
76. Taylor L, Downing A, Noury GA, Masala G, Palomino M, McGinn C. Exploring the applicability of the socially assistive robot Stevie in a day center for people with dementia. In: Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication. 2021 Presented at: RO-MAN '21; August 8-12, 2021; Vancouver, BC p. 957-962 URL: https://ieeexplore.ieee.org/document/9515423 [doi: 10.1109/ro-man50785.2021.9515423]
77. Wu S, Wu D, Ye R, Li K, Lu Y, Xu J, et al. Pilot study of robot-assisted teleultrasound based on 5G network: a new feasible strategy for early imaging assessment during COVID-19 pandemic. IEEE Trans Ultrason Ferroelectr Freq Control 2020 Nov;67(11):2241-2248. [doi: 10.1109/tuffc.2020.3020721]
78. Duan S, Liu L, Chen Y, Yang L, Zhang Y, Wang S, et al. A 5G-powered robot-assisted teleultrasound diagnostic system in an intensive care unit. Crit Care 2021 Apr 07;25(1):134 [FREE Full text] [doi: 10.1186/s13054-021-03563-z] [Medline: 33827638]
79. Li X, Guo L, Sun L, Yue PW, Xu H. Teleultrasound for the COVID-19 pandemic: a statement from China. Adv Ultrasound Diagn Ther 2020;4(2):50-56. [doi: 10.37015/AUDT.2020.200036]
80. Wang J, Peng C, Zhao Y, Ye R, Hong J, Huang H, et al. Application of a robotic tele-echography system for COVID-19 pneumonia. J Ultrasound Med 2021 Feb;40(2):385-390. [doi: 10.1002/jum.15406] [Medline: 32725833]
81. Li D. 5G and intelligence medicine-how the next generation of wireless technology will reconstruct healthcare? Precis Clin Med 2019 Dec;2(4):205-208 [FREE Full text] [doi: 10.1093/pcmedi/pbz020] [Medline: 31886033]
82. Khakharia A, Mehendale N. Robotic motion planning using Probabilistic Collision Hidden Markov Model (PCHMM) for contactless drug delivery. SSRN Journal. Preprint posted online on June 15, 2021 [FREE Full text] [doi: 10.2139/ssrn.3860266]
83. Evans J, Krishnamurthy B, Pong W, Croston R, Weiman C, Engelberger G. HelpMate™: a robotic materials transport system. Robot Auton Syst 1989 Nov;5(3):251-256. [doi: 10.1016/0921-8890(89)90049-3]
84. Summerfield MR, Seagull FJ, Vaidya N, Xiao Y. Use of pharmacy delivery robots in intensive care units. Am J Health Syst Pharm 2011 Jan 01;68(1):77-83. [doi: 10.2146/ajhp100012] [Medline: 21164071]
85. Wang H, Kasagami F. Careful-patient mover used for patient transfer in hospital. In: Proceedings of the 2007 IEEE/ICME International Conference on Complex Medical Engineering,. 2007 Presented at: IEEE/ICME '07; May 23-27, 2007; Beijing, China p. 20-26 URL: https://ieeexplore.ieee.org/document/4381684 [doi: 10.1109/iccme.2007.4381684]
86. Stibich M, Stachowiak J. The microbiological impact of pulsed xenon ultraviolet disinfection on resistant bacteria, bacterial spore and fungi and viruses. S Afr J Infect Dis 2016 Mar 31;31(1):12-15. [doi: 10.4102/sajid.v31i1.103]
87. Logistics Disinfection Robot of epsgroup Hospital Put into Anti-Epidemic Work in Central South Hospital of Wuhan University. Tencent QQ. 2020. URL: https://page.om.qq.com/page/OqCoH8_owiYNyNG8UvggcMHg0 [accessed 2024-02-29]
88. The "autonomous mobile disinfection robot" produced in Shanghai has arrived and has been used on the front line of the fight against the epidemic. China News Network. 2020. URL: https://baijiahao.baidu.com/s?id=1658226694285619350&wfr=spider&for=pc [accessed 2024-02-11]
89. Maclean M, McKenzie K, Moorhead S, Tomb RM, Coia JE, MacGregor SJ, et al. Decontamination of the hospital environment: new technologies for infection control. Curr Treat Options Infect Dis 2015 Jan 30;7(1):39-51. [doi: 10.1007/s40506-015-0037-5]
90. Yu LH. Real-time monitoring of air pathogens in the ICU with biosensors and robots. Crit Care 2022 Nov 14;26(1):353 [FREE Full text] [doi: 10.1186/s13054-022-04222-7] [Medline: 36376890]
91. Weil MH, Tang W. From intensive care to critical care medicine: a historical perspective. Am J Respir Crit Care Med 2011 Jun 01;183(11):1451-1453. [doi: 10.1164/rccm.201008-1341OE] [Medline: 21257788]
92. Site A, Lohan ES, Jolanki O, Valkama O, Hernandez RR, Latikka R, et al. Managing perceived loneliness and social-isolation levels for older adults: a survey with focus on wearables-based solutions. Sensors (Basel) 2022 Feb 01;22(3):1108 [FREE Full text] [doi: 10.3390/s22031108] [Medline: 35161852]
93. Locsin RC. The co-existence of technology and caring in the theory of technological competency as caring in nursing. J Med Invest 2017;64(1.2):160-164 [FREE Full text] [doi: 10.2152/jmi.64.160] [Medline: 28373615]

XSL•FO
RenderX

94. Hermans G, Van den Berghe G. Clinical review: intensive care unit acquired weakness. Crit Care 2015 Aug 05;19(1):274 [FREE Full text] [doi: 10.1186/s13054-015-0993-7] [Medline: 26242743]
95. Calabrò RS, Cacciola A, Bertè F, Manuli A, Leo A, Bramanti A, et al. Robotic gait rehabilitation and substitution devices in neurological disorders: where are we now? Neurol Sci 2016 Apr;37(4):503-514. [doi: 10.1007/s10072-016-2474-4] [Medline: 26781943]
96. van Kammen K, Boonstra AM, van der Woude LH, Visscher C, Reinders-Messelink HA, den Otter R. Lokomat guided gait in hemiparetic stroke patients: the effects of training parameters on muscle activity and temporal symmetry. Disabil Rehabil 2020 Oct;42(21):2977-2985. [doi: 10.1080/09638288.2019.1579259] [Medline: 30973764]
97. Baronchelli F, Zucchella C, Serrao M, Intiso D, Bartolo M. The effect of robotic assisted gait training with Lokomat® on balance control after stroke: systematic review and meta-analysis. Front Neurol 2021;12:661815 [FREE Full text] [doi: 10.3389/fneur.2021.661815] [Medline: 34295298]
98. Calafiore D, Negrini F, Tottoli N, Ferraro F, Ozyemisci-Taskiran O, de Sire A. Efficacy of robotic exoskeleton for gait rehabilitation in patients with subacute stroke : a systematic review. Eur J Phys Rehabil Med 2022 Feb;58(1):1-8. [doi: 10.23736/S1973-9087.21.06846-5] [Medline: 34247470]
99. Louie DR, Eng JJ, Lam T, Spinal Cord Injury Research Evidence (SCIRE) Research Team. Gait speed using powered robotic exoskeletons after spinal cord injury: a systematic review and correlational study. J Neuroeng Rehabil 2015 Oct 14;12:82 [FREE Full text] [doi: 10.1186/s12984-015-0074-9] [Medline: 26463355]
100. Mehrholz J, Pohl M. Electromechanical-assisted gait training after stroke: a systematic review comparing end-effector and exoskeleton devices. J Rehabil Med 2012 Mar;44(3):193-199 [FREE Full text] [doi: 10.2340/16501977-0943] [Medline: 22378603]
101. Veneman JF, Kruidhof R, Hekman EE, Ekkelenkamp R, Van Asseldonk EH, van der Kooij H. Design and evaluation of the LOPES exoskeleton robot for interactive gait rehabilitation. IEEE Trans Neural Syst Rehabil Eng 2007 Sep;15(3):379-386 [FREE Full text] [doi: 10.1109/tnsre.2007.903919] [Medline: 17894270]
102. Petrinec AB, Martin BR. Post-intensive care syndrome symptoms and health-related quality of life in family decision-makers of critically ill patients. Palliat Support Care 2018 Dec 26;16(6):719-724. [doi: 10.1017/S1478951517001043] [Medline: 29277171]
103. Wood MD, Maslove D, Muscedere J, Scott SH, Day A, Boyd JG. Assessing the relationship between brain tissue oxygenation and neurological dysfunction in critically ill patients: study protocol. Int J Clin Trials 2016 Aug 06;3(3):98-105. [doi: 10.18203/2349-3259.ijct20162792]
104. Dukelow SP, Herter TM, Moore KD, Demers MJ, Glasgow JI, Bagg SD, et al. Quantitative assessment of limb position sense following stroke. Neurorehabil Neural Repair 2010 Feb;24(2):178-187. [doi: 10.1177/1545968309345267] [Medline: 19794134]
105. Veerbeek JM, Langbroek-Amersfoort AC, van Wegen EE, Meskers CG, Kwakkel G. Effects of robot-assisted therapy for the upper limb after stroke. Neurorehabil Neural Repair 2017 Feb 24;31(2):107-121. [doi: 10.1177/1545968316666957] [Medline: 27597165]
106. Schwartz I, Meiner Z. Robotic-assisted gait training in neurological patients: who may benefit? Ann Biomed Eng 2015 May;43(5):1260-1269. [doi: 10.1007/s10439-015-1283-x] [Medline: 25724733]
107. Lazzara EH, Benishek LE, Patzer B, Gregory ME, Hughes AM, Heyne K, et al. Utilizing telemedicine in the trauma intensive care unit: does it impact teamwork? Telemed J E Health 2015 Aug;21(8):670-676. [doi: 10.1089/tmj.2014.0074] [Medline: 25885369]
108. Avgousti S, Christoforou EG, Panayides AS, Voskarides S, Novales C, Nouaille L, et al. Medical telerobotic systems: current status and future trends. Biomed Eng Online 2016 Aug 12;15(1):96 [FREE Full text] [doi: 10.1186/s12938-016-0217-7] [Medline: 27520552]
109. Vespa PM, Miller C, Hu X, Nenov V, Buxey F, Martin NA. Intensive care unit robotic telepresence facilitates rapid physician response to unstable patients and decreased cost in neurointensive care. Surg Neurol 2007 Apr;67(4):331-337. [doi: 10.1016/j.surneu.2006.12.042] [Medline: 17350395]
110. Chung KK, Grathwohl KW, Poropatich RK, Wolf SE, Holcomb JB. Robotic telepresence: past, present, and future. J Cardiothorac Vasc Anesth 2007 Aug;21(4):593-596. [doi: 10.1053/j.jvca.2007.04.013] [Medline: 17678796]
111. Bettinelli M, Lei Y, Beane M, Mackey C, Liesching TN. Does robotic telerounding enhance nurse-physician collaboration satisfaction about care decisions? Telemed J E Health 2015 Aug;21(8):637-643. [doi: 10.1089/tmj.2014.0162] [Medline: 25938296]
112. Jang SM, Hong YJ, Lee K, Kim S, Chi n BV, Kim J. Assessment of user needs for telemedicine robots in a developing nation hospital setting. Telemed J E Health 2021 Jun;27(6):670-678. [doi: 10.1089/tmj.2020.0215] [Medline: 33095109]
113. Bloss R. Unmanned vehicles while becoming smaller and smarter are addressing new applications in medical, agriculture, in addition to military and security. Ind Robot 2014;41(1):82-86. [doi: 10.1108/ir-10-2013-410]
114. Kristoffersson A, Coradeschi S, Loutfi A. A Review of Mobile Robotic Telepresence. Adv Hum Comput Interact 2013;2013:1-17. [doi: 10.1155/2013/902316]
115. Botha CF. Gender and humanoid robots: a somaesthetic analysis. Filos Theor J Afr Philos Cult Relig 2021 Dec 13;10(3):119-130. [doi: 10.4314/ft.v10i3.8]

XSL•FO
RenderX

116. McGinn C, Bourke E, Murtagh A, Donovan C, Lynch P, Cullinan MF, et al. Meet Stevie: a socially assistive robot developed through application of a 'design-thinking' approach. J Intell Robot Syst 2019 Jul 17;98(1):39-58. [doi: 10.1007/S10846-019-01051-9]
117. Păvăloiu IB, Vasilă eanu A, Popa R, Scurtu D, Hang A, Goga N. Healthcare robotic telepresence. In: Proceedings of the 13th International Conference on Electronics, Computers and Artificial Intelligence. 2021 Presented at: ECAI '21; July 1-3, 2021; Pitesti, Romania p. 1-6 URL: https://ieeexplore.ieee.org/document/9515025 [doi: 10.1109/ecai52376.2021.9515025]
118. Evans KD, Yang Q, Liu Y, Ye R, Peng C. Sonography of the lungs: diagnosis and surveillance of patients with COVID-19. J Diagn Med Sonogr 2020 Apr 21;36(4):370-376. [doi: 10.1177/8756479320917107]
119. Sarker S, Jamal L, Ahmed SF, Irtisam N. Robotics and artificial intelligence in healthcare during COVID-19 pandemic: a systematic review. Rob Auton Syst 2021 Dec;146:103902 [FREE Full text] [doi: 10.1016/j.robot.2021.103902] [Medline: 34629751]
120. Teng R, Ding Y, See KC. Use of robots in critical care: systematic review. J Med Internet Res 2022 May 16;24(5):e33380 [FREE Full text] [doi: 10.2196/33380] [Medline: 35576567]
121. Sichuan's first "5G+medical robot+VR" visitation system officially launched. China News Network. URL: https://www.chinanews.com.cn/jk/2021/02-08/9408104.shtml [accessed 2023-06-10]
122. Kelly FE, Fong K, Hirsch N, Nolan JP. Intensive care medicine is 60 years old: the history and future of the intensive care unit. Clin Med (Lond) 2014 Aug;14(4):376-379 [FREE Full text] [doi: 10.7861/clinmedicine.14-4-376] [Medline: 25099838]
123. Marshall JC. Critical illness is an iatrogenic disorder. Crit Care Med 2010 Oct;38(10 Suppl):S582-S589. [doi: 10.1097/CCM.0b013e3181f2002a] [Medline: 21164401]
124. Buetti N, Ruckly S, de Montmollin E, Reignier J, Terzi N, Cohen Y, et al. COVID-19 increased the risk of ICU-acquired bloodstream infections: a case-cohort study from the multicentric OUTCOMEREA network. Intensive Care Med 2021 Feb;47(2):180-187 [FREE Full text] [doi: 10.1007/s00134-021-06346-w] [Medline: 33506379]
125. Sierra Marín SD, Gomez-Vargas D, Céspedes N, Múnera M, Roberti F, Barria P, et al. Expectations and perceptions of healthcare professionals for robot deployment in hospital environments during the COVID-19 pandemic. Front Robot AI 2021;8:612746 [FREE Full text] [doi: 10.3389/frobt.2021.612746] [Medline: 34150856]
126. Di Lallo A, Murphy R, Krieger A, Zhu J, Taylor RH, Su H. Medical robots for infectious diseases: lessons and challenges from the COVID-19 pandemic. IEEE Robot Autom Mag 2021 Mar;28(1):18-27. [doi: 10.1109/mra.2020.3045671]
127. Tamantini C, di Luzio FS, Cordella F, Pascarella G, Agro FE, Zollo L. A Robotic Health-Care Assistant for COVID-19 Emergency: A Proposed Solution for Logistics and Disinfection in a Hospital Environment. IEEE Robot Autom Mag 2021 Mar;28(1):71-81. [doi: 10.1109/mra.2020.3044953]
128. Curtis N. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. Pediatr Infect Dis J 2023 Apr 01;42(4):275. [doi: 10.1097/INF.0000000000003852] [Medline: 36757192]
129. The FLI Team. A principled AI discussion in Asilomar. Future of Life Institute. 2017. URL: https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/ [accessed 2024-02-28]
130. Lazar S, Nelson A. AI safety on whose terms? Science 2023 Jul 14;381(6654):138. [doi: 10.1126/science.adi8982] [Medline: 37440644]
131. Yang GZ, Bellingham J, Dupont PE, Fischer P, Floridi L, Full R, et al. The grand challenges of science robotics. Sci Robot 2018 Jan 31;3(14):aar7650. [doi: 10.1126/scirobotics.aar7650] [Medline: 33141701]
132. Department for Science, Innovation and Technology, AI Safety Institute, Smith C, Sunak R. Tech entrepreneur Ian Hogarth to lead UK's AI foundation model taskforce. Government of UK. URL: https://www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce [accessed 2023-05-10]
133. Chinese artificial intelligence white paper series. Chinese Association for Artificial Intelligence. URL: https://www.caai.cn/index.php?s=/home/article/detail/id/3188.html [accessed 2024-02-29]
134. The Bletchley declaration by countries attending the AI safety summit. Department for Science IT, Government of UK. URL: https://tinyurl.com/2xbn9wcv [accessed 2023-05-10]
135. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023 May;23(5):4-13 [FREE Full text] [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
136. Tustumi F, Andreollo NA, Aguilar-Nascimento JE. Future of the language models in healthcare: the role of ChatGPT. Arq Bras Cir Dig 2023;36:e1727 [FREE Full text] [doi: 10.1590/0102-672020230002e1727] [Medline: 37162073]
137. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy augmented intelligence in health care. J Med Syst 2022 Jan 12;46(2):12 [FREE Full text] [doi: 10.1007/s10916-021-01790-z] [Medline: 35020064]
138. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. EBioMedicine 2021 May;67:103358 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103358] [Medline: 33962897]
139. Gijsberts CM, Groenewegen KA, Hoefer IE, Eijkemans MJ, Asselbergs FW, Anderson TJ, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. PLoS One 2015;10(7):e0132321 [FREE Full text] [doi: 10.1371/journal.pone.0132321] [Medline: 26134404]
140. Kostick-Quenet KM, Gerke S. AI in the hands of imperfect users. NPJ Digit Med 2022 Dec 28;5(1):197 [FREE Full text] [doi: 10.1038/s41746-022-00737-z] [Medline: 36577851]

XSL•FO
RenderX

141. Safdar NM, Banja JD, Meltzer CC. Ethical considerations in artificial intelligence. Eur J Radiol 2020 Jan;122:108768. [doi: 10.1016/j.ejrad.2019.108768] [Medline: 31786504]
142. Rumbold JM, Pierscionek BK. A critique of the regulation of data science in healthcare research in the European Union. BMC Med Ethics 2017 Apr 08;18(1):27 [FREE Full text] [doi: 10.1186/s12910-017-0184-y] [Medline: 28388916]
143. Sanderson K. GPT-4 is here: what scientists think. Nature 2023 Mar;615(7954):773. [doi: 10.1038/d41586-023-00816-5] [Medline: 36928404]
144. O'Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int J Med Robot 2019 Feb 09;15(1):e1968. [doi: 10.1002/rcs.1968] [Medline: 30397993]
145. Da Silva M, Horsley T, Singh D, Da Silva E, Ly V, Thomas B, et al. Legal concerns in health-related artificial intelligence: a scoping review protocol. Syst Rev 2022 Jun 17;11(1):123 [FREE Full text] [doi: 10.1186/s13643-022-01939-y] [Medline: 35715812]
146. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023 Mar;5(3):e105-e106 [FREE Full text] [doi: 10.1016/S2589-7500(23)00019-5] [Medline: 36754725]
147. Feijóo C, Kwon Y, Bauer JM, Bohlin E, Howell B, Jain R, et al. Harnessing artificial intelligence (AI) to increase wellbeing for all: the case for a new technology diplomacy. Telecomm Policy 2020 Jul;44(6):101988 [FREE Full text] [doi: 10.1016/j.telpol.2020.101988] [Medline: 32377031]
148. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
149. Greaves F, Joshi I, Campbell M, Roberts S, Patel N, Powell J. What is an appropriate level of evidence for a digital health intervention? Lancet 2019 Dec 22;392(10165):2665-2667. [doi: 10.1016/S0140-6736(18)33129-5] [Medline: 30545779]
150. Artificial intelligence: an accountability framework for federal agencies and other entities. U.S. Government Accountability Office. URL: https://www.gao.gov/products/gao-21-519sp [accessed 2023-05-10]
151. Sarfraz Z, Sarfraz A, Iftikar HM, Akhund R. Is COVID-19 pushing us to the fifth industrial revolution (society 5.0)? Pak J Med Sci 2021;37(2):591-594 [FREE Full text] [doi: 10.12669/pjms.37.2.3387] [Medline: 33679956]
152. Cosgriff CV, Stone DJ, Weissman G, Pirracchio R, Celi LA. The clinical artificial intelligence department: a prerequisite for success. BMJ Health Care Inform 2020 Jul;27(1):e100183 [FREE Full text] [doi: 10.1136/bmjhci-2020-100183] [Medline: 32675072]
153. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med 2019 Jan;25(1):16-18 [FREE Full text] [doi: 10.1038/s41591-018-0310-5] [Medline: 30617332]

Abbreviations

5G: fifth-generation
AI: artificial intelligence
CPR: cardiopulmonary resuscitation
CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
ICU: intensive care unit
KINARM: Kinesiological Instrument for Normal and Altered Reaching Movements
MGIUS-R3: MGI Ultrasound System-Remote 3
NICU: neonatal intensive care unit
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RP-7: Remote Presence-7
RP-VITA: Remote Presence Virtual + Independent Telemedicine Assistant
SCI: spinal cord injury
VR: virtual reality

XSL•FO
RenderX

Edited by S Ma, T Leung; submitted 29.10.23; peer-reviewed by S Wei, TAR Sure; comments to author 26.02.24; revised version received 07.03.24; accepted 22.04.24; published 27.05.24.

Please cite as:
Li Y, Wang M, Wang L, Cao Y, Liu Y, Zhao Y, Yuan R, Yang M, Lu S, Sun Z, Zhou F, Qian Z, Kang H
Advances in the Application of AI Robots in Critical Care: Scoping Review
J Med Internet Res 2024;26:e54095
URL: https://www.jmir.org/2024/1/e54095
doi:10.2196/54095
PMID:38801765

©Yun Li, Min Wang, Lu Wang, Yuan Cao, Yuyan Liu, Yan Zhao, Rui Yuan, Mengmeng Yang, Siqian Lu, Zhichao Sun, Feihu Zhou, Zhirong Qian, Hongjun Kang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of mHealth Apps for Maternal Health Care Delivery: Systematic Review of Systematic Reviews

Edward Kwabena Ameyaw[1,2], PhD; Padmore Adusei Amoah[1,2,3], PhD; Obidimma Ezezika[4], PhD

[1]Institute of Policy Studies, Lingnan University, Hong Kong, China (Hong Kong)
[2]School of Graduate Studies, Lingnan University, Hong Kong, China (Hong Kong)
[3]Department of Psychology, Lingnan University, Hong Kong, China (Hong Kong)
[4]Global Health & Innovation Lab, Faculty of Health Sciences, School of Health Studies, Western University, London, ON, Canada

Corresponding Author:
Edward Kwabena Ameyaw, PhD
Institute of Policy Studies
Lingnan University
Tuen Mun
Hong Kong
China (Hong Kong)
Phone: 852 55381568
Email: edmeyaw19@gmail.com

Abstract

Background: Globally, the use of mobile health (mHealth) apps or interventions has increased. Robust synthesis of existing systematic reviews on mHealth apps may offer useful insights to guide maternal health clinicians and policy makers.

Objective: This systematic review aims to assess the effectiveness or impact of mHealth apps on maternal health care delivery globally.

Methods: We systematically searched Scopus, Web of Science (Core Collection), MEDLINE or PubMed, CINAHL, and Cochrane Database of Systematic Reviews using a predeveloped search strategy. The quality of the reviews was independently assessed by 3 reviewers, while study selection was done by 2 independent raters. We presented a narrative synthesis of the findings, highlighting the specific mHealth apps, where they are implemented, and their effectiveness or outcomes toward various maternal conditions.

Results: A total of 2527 documents were retrieved, out of which 16 documents were included in the review. Most mHealth apps were implemented by sending SMS text messages with mobile phones. mHealth interventions were most effective in 5 areas: maternal anxiety and depression, diabetes in pregnancy, gestational weight management, maternal health care use, behavioral modification toward smoking cessation, and controlling substance use during pregnancy. We noted that mHealth interventions for maternal health care are skewed toward high-income countries (13/16, 81%).

Conclusions: The effectiveness of mHealth apps for maternity health care has drawn attention in research and practice recently. The study showed that research on mHealth apps and their use dominate in high-income countries. As a result, it is imperative that low- and middle-income countries intensify their commitment to these apps for maternal health care, in terms of use and research.

Trial Registration: PROSPERO CRD42022365179; https://tinyurl.com/e5yxyx77

(J Med Internet Res 2024;26:e49510) doi:10.2196/49510

KEYWORDS

mHealth; mobile health; maternal health; telemedicine; technology; health care; newborn; systematic review; database; mHealth impact; mHealth effectiveness; health care applications

Introduction

The use of mobile health (mHealth) apps has ascended following the proliferation of wearable devices, live audio-visual communication systems, SMS text messaging, and mobile phone app inter alia [1]. mHealth is defined by the World Health Organization (WHO) as a “medical and public health practice supported by mobile devices, such as mobile phones, patient

XSL•FO
RenderX

monitoring devices, personal digital assistants, and other wireless devices" [2]. mHealth has been identified as an essential public health tool for efficient health care delivery, especially in situations where the face-to-face model of care cannot be readily provided [3-5]. The grievous repercussions of the COVID-19 outbreak on the ailing health systems of low- and middle-income countries (LMICs) [6,7], and some high-income countries (HICs) [6-8], partly suffice the need to optimize the use of mHealth apps or interventions. A critical aspect of health care greatly impacted by COVID-19 was maternal health care delivery [9-11].

Nonetheless, mHealth could help subside some of the challenges confronting maternal health care delivery. Recent evidence has reverberated that the continuum of maternity health care has received its fair share of mHealth interventions globally, with increasing use and resultant positive outcomes [12-16]. So far, existing mHealth apps for maternal and newborn health care include the Mobile Technology for Community Health (MOTECH), behavioral change tools, and momConnect [17,18]. There are also mHealth apps for general health care delivery such as the Hospital Authority mobile app in Hong Kong [19]. mHealth aids to offset the human resource gap through diverse technologies designed to support treatment adherence, clinical diagnosis, and enhancement [2]. In this study, the terms "mHealth applications" and "mHealth interventions" are alternated. In addition, "newborn healthcare or care" is subsumed under "maternal healthcare or care."

It is anticipated that there will be approximately 5.6 billion mobile connections by 2025, with most of these being smartphones [20]. A considerable proportion of these mobile connections are earmarked to occur in LMICs, where worse maternal health conditions occur, as mobile connections are more readily accessible in some instances than clean water and electricity [21,22]. This could suggest that LMICs have an increased propensity to maximize the benefits of mHealth to truncate the alarming maternal and newborn morbidity and mortality, which at present is almost 95% of all global maternal deaths [23]. The utility of mHealth interventions can, therefore, be maximized to curtail the gloomy maternal health situation in some parts of the world.

There is a plethora of systematic reviews that have synthesized the relevance of mHealth interventions used in maternal health care [24-29]. For instance, Ambia and Mandela [26] realized that mobile phone–based interventions lead to a statistically significant rise in the uptake of early infant diagnosis of HIV while Bossman et al [29] highlighted the importance of SMS text messages and voice message reminders toward behavioral change among pregnant women by enhancing antenatal care (ANC) and prenatal care attendance, skill birth attendance, and vaccination uptake [29].

Meanwhile, these reviews have not been synthesized on a global scale to draw consistencies and inconsistencies in the evidence to guide maternal health care models and interventions, to maximize gains. A recent systematic review of systematic reviews rather focused on the effectiveness of mHealth on health issues such as diabetes, heart failure symptoms, hypertension, and other health conditions [30]. The prevailing evidence seems to converge that mHealth interventions are impactful and beneficial for maternal health care in both HICs and LMICs [31-34].

So far, no review has synthesized the existing systematic literature reviews to provide aggregate evidence to guide maternal health care practitioners, clinicians, and policy makers. Robust synthesis of existing systematic reviews on mHealth interventions may be useful in guiding clinicians and policy makers with relevant evidence and further pinpointing aspects requiring further evidence to minimize potential risks associated with the existing maternal health care models. Hence, this systematic review of systematic reviews explores the effectiveness or impact of mHealth apps or interventions on maternal health globally.

Methods

Overview

This paper is a systematic review of systematic reviews. The review addressed the following research questions: (1) What mHealth apps are used for maternal health care delivery globally? (2) What is the effectiveness of mHealth apps for maternal health care? and (3) What are the barriers and facilitators in the use of mHealth apps for maternal health care delivery?

We conducted this study in line with the updated guidelines for PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 1) statement and the methodological considerations when including existing systematic reviews [35].

Protocol and Registration

We developed a protocol to guide the conduct of the study. The protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO).

Search Strategy and Information Sources

Our search focused on published systematic reviews that focused on specific mHealth apps used in delivering any maternity health care (ie, prenatal, birth, or postnatal) and newborn care, subsumed under maternal care. A newborn refers to a neonate, thus, a child within 28 days of birth, as conceptualized by the WHO [36]. A systematic search for systematic reviews was executed in 5 databases: Scopus, Web of Science (Core Collection), MEDLINE or PubMed, CINAHL, and Cochrane Database of Systematic Reviews, using a predeveloped search strategy (Multimedia Appendix 2). The aforementioned databases were used due to their dominance in biomedical reviews. Searches were done on October 20, 2022. Besides, relevant references of retrieved papers were searched manually to retrieve other relevant papers.

Eligibility Criteria

This paper included systematic reviews involving primary studies of all study designs (eg, randomized controlled trials [RCTs], pre- and posttest designs, nonrandomized trials, and observational studies). We included reviews focusing on the effectiveness of all available mHealth apps used for maternal

XSL•FO
RenderX

and newborn health care delivery globally, for example, service delivery apps. A paper was considered a systematic review if it had these five characteristics: (1) clearly defined aim or research question, (2) eligibility criteria for included studies, (3) appropriate search strategy, (4) appraisal or quality assessment, and (5) analysis or synthesis [37]. Besides, the Population, Intervention, Comparison, and Outcomes framework was applied in determining the eligibility. Thus, the population of interest was pregnant women, women in labor or at birth, postnatal women, and neonates. The intervention of interest was mHealth as defined in the *Introduction* section. Where reported, the comparison was the women or neonates who were not treated with the assistance of any mHealth app, and the outcomes were the reported results that emerged after using mHealth app for any maternity or neonatal condition. Only studies in the English language were considered without year limits. Due to the language competency of authors, only papers published in the English language were considered as English is the working language of the authors. We also excluded papers focusing on nonmaternal health issues or reviews on interventions other than mHealth.

Study Selection

Reviews were imported to EndNote (Clarivate), and duplicates were removed, after which title and abstract screening were performed by 2 authors (EKA and PAA) using the inclusion and exclusion criteria. This was done while blinding each other's decision, and all discrepancies were discussed for resolution. Initial screening was on the titles and abstracts retrieved from the search. All suitable papers were retrieved for full-text evaluation. Afterward, we conducted a full-text assessment to determine the eligibility for inclusion. The journal, authorship, or years were not blinded in the process. Reference lists of included papers were searched for additional suitable papers, and we had no reason to contact authors as our results were exhaustive.

Quality Assessment, Data Extraction, and Synthesis

We extracted data following a standardized extraction form (EKA), after which the data were reviewed by a senior researcher (PAA). Specifically, we extracted the following data: author information and year of publication, setting or context, objective, time range of included papers, databases searched, number of included papers, design of included papers, and whether the meta-analysis was conducted or not (Table 1). In addition, we extracted data on the description of included mHealth apps, the target population, the effectiveness of the mHealth apps, and other relevant outcomes of interest as shown in Table 2. All discrepancies were discussed for resolution. By way of definition, (1) mHealth apps refer to the specific mHealth technique and tools that were used; (2) the target population could either be pregnant women, women in labor or childbirth, postnatal women, and neonates; and (3) effectiveness refers to the outcomes reported after using the reported mHealth technique or intervention. We presented a narrative synthesis of the findings by highlighting the specific mHealth apps, where they are implemented, and their effectiveness or outcomes toward various maternal conditions, such as the impact on maternal health care use. Each of the authors (EKA, PAA, and OE) independently assessed the quality of the reviews using the critical appraisal instrument for systematic reviews and research synthesis by the Joanna Briggs Institute (Multimedia Appendix 3 [38]). This is a 12-point critical appraisal tool for assessing the quality of a systematic review to determine if specific papers could be included or otherwise. Reporting was guided by the Preferred Reporting Items for Overviews of Reviews (PRIOR).

XSL•FO
RenderX

Table 1. Summary of included reviews.

Authors	Setting or context	Aim	Literature date	Databases	Included studies, n	Study design of included studies	Meta-analyzed
Bayrampour et al [39]	Sweden, Australia, United Kingdom, Switzerland, and Canada	To examine the effectiveness of eHealth interventions in reducing perinatal anxiety	2014-2017	MEDLINE, CINAHL, Embase, and PsycINFO	5	RCTs[a]	Yes
Chae and Kim [27]	United States, Sweden, Netherlands, Jordan, Norway, Australia, and Singapore	To investigate the effect of internet-based prenatal interventions among pregnant women	2017-2020	PubMed, CINAHL, Cochrane Library, Embase, and ERIC	15	RCTs	Yes
Daly et al [40]	Ireland, United States, and Australia	To determine the effects of mobile app interventions on healthy maternal behavior and perinatal health outcomes	2016	PubMed, Embase, Cochrane Library, CINAHL, WHO Global Health Library, POPLINE, and CABI Global Health	4	RCTs	No
Eberle et al [41]	Unspecified	To assess the clinical effectiveness of technologies for diabetes in pregnancy	2008-2020	MEDLINE, PubMed, Cochrane Library, Embase, CINAHL, and Web of Science Core Collection	22	RCTs, randomized crossover trials, cohort studies, and controlled clinical trials	No
Farzandipour et al [42]	Australia, United States, Norway, China, United Kingdom, and Europe	To investigate the effects and features of phone-based interventions to control gestational weight gain	2011-2016	MEDLINE (via PubMed), Scopus, and Cochrane Central Register of Controlled Trials	12	RCTs and nonrandomized controlled trials	No
Griffiths et al [43]	United States and United Kingdom	To explore whether digital interventions for pregnancy smoking cessation are effective, the impact of intervention platform on smoking cessation, the associations between specific BCTs[b] delivered in interventions and smoking cessation, and the association between the total number of BCTs delivered and smoking cessation	1997-2017	Academic Search Complete, ASSIA, CINAHL, The Cochrane Library, Embase, MEDLINE, PsycINFO, Scopus, and Web of Science	12	RCTs and quasi-RCTs	Yes

XSL•FO
RenderX

Authors	Setting or context	Aim	Literature date	Databases	Included studies, n	Study design of included studies	Meta-analyzed
Lau et al [34]	United States, Spain, Italy, and Ireland	To evaluate the efficacy of internet-based self-monitoring interventions in improving maternal and neonatal outcomes	2007-2015	Cochrane Central Register of Controlled Trials, PubMed, Embase, Cumulative Index to Nursing and Allied Health Literature, PsycINFO, Scopus, and ProQuest	9	RCTs and controlled clinical trials	Yes
Lee and Cho [44]	Switzerland, United States, and United Kingdom	To examine the development and delivery of technology-supported interventions for pregnant women and explored the effects of these interventions on the targeted outcomes	2005-2017	MEDLINE, CINAHL, and Scopus	11	RCTs, pilot RCTs, prospective mixed methods one-group pretest or posttest designs, case reports, pilot interventions, and mixed methods studies	No
Ming et al [45]	Italy, Poland, United States, Spain, and Ireland	To determine whether telemedicine solutions offer any advantages relative to standard care for women with diabetes in pregnancy	1996-2015	MEDLINE, Embase, Compendex, and PubMed	7	RCTs	Yes
Oh et al [46]	United States and Netherlands	To assess the effectiveness of digital interventions in preventing alcohol consumption during pregnancy or pregnancy-planning period, and effectiveness of alternative digital intervention platforms (eg, computers, mobiles, and text messaging services)	2012-2018	PubMed, Embase, CINAHL, and Web of Science	6	RCTs	Yes
Overdijkink et al [47]	Netherlands, United States, Australia, United Kingdom, and Japan	To evaluate the feasibility, acceptability, effectiveness of mHealth[c] lifestyle and medical apps to support health care during pregnancy	2007-2017	Embase, MEDLINE Epub (Ovid), Cochrane Library, Web of Science, and Google Scholar	29	RCTs, pilot studies, prospective or retrospective cohort studies, surveys, and qualitative health care research	Yes

XSL•FO
RenderX

Authors	Setting or context	Aim	Literature date	Databases	Included studies, n	Study design of included studies	Meta-analyzed
Rahman et al [48]	Kenya, Nigeria, Brazil, India, Ethiopia, Tanzania, and China	To summarize the effect of mHealth interventions on improving the uptake of ANC[d] visits, skilled birth attendance at the time of delivery, and facility delivery among pregnant women	2012-2018	APA PsycINFO, British Nursing Index, CINAHL Plus, Embase, MEDLINE, POPLINE, PubMed, The Cochrane Library, and Web of Science	9	RCTs and cluster RCTs	Yes
Silang et al [49]	United States and England	To evaluate the effectiveness of eHealth interventions in treating substance use during pregnancy	2009-2018	PsycINFO, Embase, CINAHL, and Cochrane CENTRAL	6	RCTs	Yes
Silang et al [50]	United States, Netherlands, China, Norway, Australia, United Kingdom, Sweden, Switzerland, and Thailand	To determine the effectiveness of eHealth interventions in preventing and treating depression, anxiety, and insomnia during pregnancy. Second, to identify demographic and intervention moderators of effectiveness	2008-2020	PsycINFO, MEDLINE, CINAHL, Embase, and Cochrane	17	RCTs	Yes
Sondaal et al [51]	Low- and middle-income countries (countries unspecified)	To assess the effect of mHealth interventions that support pregnant women during the antenatal, birth and postnatal period	2001-2014	Cochrane Database of Systematic Reviews, PubMed/MEDLINE, Embase, Global Health Library, and POPLINE	27	Intervention and descriptive	No
Wagnew et al [52]	Zanzibar, Thailand, Kenya, South Africa, Ethiopia, and India	To determine the effectiveness of SMS text messages on focused ANC visits and the attendance of skilled birth professionals	2008-2017	Cochrane Review, CINAHL, PsycINFO, PubMed, and Google Scholar	7	RCTs	Yes

[a]RCT: randomized controlled trial.

[b]BCT: behavior change technique.

[c]mHealth: mobile health.

[d]ANC: antenatal care.

XSL•FO
RenderX

Table 2. Summary of key outcomes or findings.

Authors	mHealth[a] app	Target population	Effectiveness or outcome	Other relevant outcomes
Bayrampour et al [39]	eHealth mental health intervention delivered through computer software, via web websites, or mobile apps	Pregnant and postpartum women	• Intervention group had significantly lower depression scores postintervention than the control group. • Relative to the control group, more women in the intervention group did not meet the diagnostic criteria for depression (63% vs 12%). • Compared with the control group, fewer women in the intervention group met the clinical diagnostic criteria for depression at the 12-week assessment (79% vs 18%; *P*=.001). • More women in the intervention group recovered on the Edinburgh Postnatal Depression Scale (62% vs 38%; *P*=.08).	• Satisfaction with intervention was high with an associated positive perception. • 80% of participants in 1 intervention reported enjoying the program and communicating with their coach. • Some participants were stressed about not keeping up with the treatment modules.
Chae et al [27]	Prenatal educational interventions delivered through any web, computer, mobile, eHealth, mHealth, telehealth, app, kiosk, or social networking service platform	Pregnant women	• Interventions had a small effect on depression and did not have a demonstrable effect on anxiety. • Meta-analysis showed an effect on postpartum depression (−0.16, 95% CI −0.26 to −0.05).	• Intervention influenced bonding, breastfeeding efficacy, social support, and quality of life. • No influence on nutrition, breastfeeding knowledge, parental self-efficacy, insomnia, work, social adjustment, and health status were observed.
Daly et al [40]	Mobile app–based interventions designed to influence maternal knowledge or behavior during pregnancy	Pregnant women	• More intervention participants transitioned to a "maintenance stage" of healthy lifestyle behaviors by 28 weeks of gestation, compared with the control participants (52.8% vs 32.7%; *P*=.004). • Intervention participants had an increased daily step at 12 weeks (1096, SD 1898 steps), compared with 259 (SD 1604) steps. • By 32 weeks gestation, participants using mobile app recorded information more frequently than the control group. • Intervention group had a higher proportion of participants with well-controlled asthma than the control group (82% vs 58%; *P*=.03) at 6 months from baseline.	• Clinically significant improvement in the asthma-related quality of life among the intervention group compared with usual care at 6 months from baseline.
Eberle et al [41]	Diabetes technologies: mHealth apps and others	Pregnant woman	• Overall, the intervention groups showed lower HbA_{1c}[b] values than the control groups. • There were significant differences in support of the intervention groups in patient compliance (defined as the ratio between actual and instructed blood glucose measurements × 100).	__[c]

XSL•FO
RenderX

Authors	mHealth[a] app	Target population	Effectiveness or outcome	Other relevant outcomes
Farzandipour et al [42]	Intervention using telephone, mobile phone, or smartphone on weight gain control or assessed the feasibility and pilot testing this type of interventions	Obese or overweight pregnant	• Most interventions had a statistically significant positive impact on the management of weight gain during pregnancy (7/12, 66%). • Meanwhile, some had no effect.	—
Griffiths et al [43]	Digital interventions delivered through computer, video or DVD, mobile telephone, or portable handheld device (eg, tablet or iPad)	Pregnant women who smoke cigarette	• The majority of studies reported 7-day point-prevalence smoking abstinence toward the end of pregnancy. • There was self-reported abstinence at 8 weeks after intervention.	—
Lau et al [34]	Technology-support self-monitoring intervention	Perinatal diabetic women	Maternal outcome: • The meta-analysis revealed that the intervention significantly improved HbA_{1c} levels (mean difference −0.12, 95% CI −0.22 to −0.02). • The interventions did not significantly improve cesarean delivery rate for overall effect (RR^d=0.84, 95% CI 0.68-1.05; z=1.55, *P*=.12). • The interventions significantly decreased the cesarean delivery rate among the mixed group (RR=0.73, z=2.23, *P*=.03). Neonatal outcome: • No significant differences were found between intervention and control groups in the gestational diabetes mellitus group (mean difference=92.21, z=1.47, *P*=.14) and the mixed group (mean difference=–36.42, z=0.59, *P*=.56). • The intervention group demonstrated no significant difference in overall effect (RR=1.09, z=0.24, *P*=.81) compared with the control group.	—
Lee et al [44]	Technology-supported intervention	Pregnant women	• Intervention was identified to be effective as conventional interventions. • Significant benefits in decreasing maternal stress, depression, enhancing pregnancy-related knowledge, lowering perceived barriers to being active, alcohol abstinence, and smoking cessation. • Women reported benefits in stress coping, reduction in anxiety, depression and enhanced bonding with babies.	• Suggestions for improvements to the intervention programs included embedding a tracking component for users' progress over time, not limiting the intervention content to that generated by a schedule and allowing for self-directed use, and expansion of the intervention to pregnant women's significant others.
Ming et al [45]				

XSL•FO
RenderX

Authors	mHealth[a] app	Target population	Effectiveness or outcome	Other relevant outcomes
	Modem transmission of blood glucose readings to a central hospital computer, websites accessible to patients and health care professionals, a telephone system that translated blood glucose readings into audio tones to transmit them to a computer database, SMS text message transmissions of blood glucose readings to a central database, and a telemedicine hub located in the women's home	Pregnant women diagnosed with gestational diabetes mellitus or with preexisting type 1 or type 2 diabetes	• Meta-analysis demonstrated a modest but statistically significant improvement in HbA_{1c} associated with the use of telemedicine technology. • 7.5% of women had either pregnancy-induced hypertension or preeclampsia and there was no difference in the risk ratio between the telemedicine or control groups. • Rates of cesarean section were high in both groups (50% in the telemedicine and 45% in the control). • Mean birth weight was 3363 (SD 115) g and 3302 (SD 121) g for telemedicine group and standard care group respectively. • Of the 18% of babies treated for hypoglycemia, no differences in intervention and control groups occurred.	• One trial reported mothers' satisfaction among the intervention group. • 90% (17/19) of women in the telemedicine group agreed or strongly agreed that they were satisfied with the (intervention) system and would use it again.
Oh et al [46]	Text4Baby and Text4Baby Pilot (text messaging service), CHOICES intervention (Automated internet intervention), electronic screening and brief intervention facilitating self-change, computer-delivered single session brief motivational intervention, and booster session	Pregnant women or women planning pregnancy	• All studies showed that digital interventions may decrease the odds of drinking during pregnancy relative to comparison groups. • The primary meta-analysis produced a sample-weighted odds ratio of 0.62 (95% CI 0.42-0.91; *P*=.02) in favor of digital interventions decreasing the risk of alcohol consumption during pregnancy when compared with controls. • Computer or internet-based interventions (odds ratio 0.59, 95% CI 0.38-0.93) were an effective platform for preventing alcohol consumption.	—
Overdijkink et al [47]	App or text message service during pregnancy	Pregnant women and their partners	• 10 studies reported positively on acceptability. • 4 out of 19 studies evaluating effectiveness showed significant results on weight gain restriction during pregnancy, intake of vegetables and fruits, and smoking cessation.	• One study on diabetic treatment reported on acceptability with a positive user satisfaction.
Rahman et al [48]	All types of mHealth interventions focusing on improving perinatal health care use, including skilled birth attendance and facility delivery	Healthy pregnant women		—

XSL•FO
RenderX

Authors	mHealth[a] app	Target population	Effectiveness or outcome	Other relevant outcomes
			• Positive effect of mHealth interventions on improving 4 or more ANC[e] visit irrespective of the direction of interventions. • Only 2-way mHealth interventions were effective in improving the use of skilled birth attendance during delivery but the effects were unclear for 1-way mHealth interventions when compared with standard care. • Interventions were effective for facility delivery in settings where fewer pregnant women used facility delivery, however, the effects were unclear in settings where most pregnant women already used facility delivery.	
Silang et al [49]	eHealth intervention (eg, video therapy sessions, telephone, SMS text messaging, and recorded therapy sessions) intending to reduce substance use	Pregnant women	• eHealth interventions significantly reduced substance use in pregnant women compared with controls. • Only 1 of the included studies showed a statistically significant benefit of eHealth interventions over the control group.	—
Silang et al [50]	eHealth interventions delivered in an electronic capacity (eg, video therapy sessions, telephone, SMS text messaging, self-help interventions, and recorded therapy sessions)	Pregnant women	• During pregnancy, eHealth interventions have small effect sizes for preventing and treating symptoms of anxiety and depression and a moderate effect size for treating symptoms of insomnia. • With the exception of intervention type for the outcome of depressive symptoms, where mindfulness interventions outperformed other intervention types, no significant moderators were detected.	—
Sondaal et al [51]	Medical and public health practices supported by mobile phones and tablets, using text, audio, images, video or coded data in the form of SMS text messages, voice SMS messages, applications accessible via GPRS[f], GPS, third- and fourth-generation mobile telecommunications, and Bluetooth	Women in pregnancy, labor and postnatal care up to 28 days post partum	• All studies addressing maternal and neonatal service use showed significant increases. • mHealth interventions increased maternal and neonatal service use: increased ANC attendance, facility-service use, skilled attendance at birth, and vaccination rates. • The total perinatal mortality rate based on stillbirth and neonatal mortality was 19 per 1000 births in the intervention group compared with 36 per 1000 births in the control group.	

XSL•FO
RenderX

Authors	mHealth[a] app	Target population	Effectiveness or outcome	Other relevant outcomes
				• Accessibility of mHealth interventions was enhanced by providing information in different local languages, using locally-based software. • Weaknesses of mHealth interventions included lack of mobile phone ownership, illiteracy and low accessibility to mobile phones among rural women. • Strengths of mHealth interventions on usability included ease of use, flexibility in use and adaptability to the time in pregnancy and development within the local context. • Weaknesses of mHealth interventions included uncertainty on whether a text is received and the limited length of messages. • Barrier to acceptance was recipient fatigue when too many messages were sent. • Privacy may also be compromised when the mobile phone is shared among family members.
Wagnew et al [52]	—	—	• Statistically significant associations in experimental group was noted as pregnant women who received text messaging had a 174% increase in focused ANC visits and 82% in skilled birth attendance.	—

[a]mHealth: mobile health.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]Not available.

[d]RR: relative ratio.

[e]ANC: antenatal care.

[f]GPRS: general packet radio service.

Results

Characteristics of Reviews

Our rigorous systematic search retrieved 2516 papers, with 168 duplicates thereby leaving 2348 papers (Figure 1). An additional 11 papers were identified through cross-referencing and free-hand search. As a result, we screened the titles and abstracts of 2359 papers, and 58 papers were eligible for full-text screening. Of these, 19 papers were deemed eligible for inclusion. However, 3 papers were excluded due to limited data, hence 16 (8.1%) reviews out of 198 studies were included in the final review (Figure 1), as no study was excluded based on quality assessment outcome. Of the 198 studies, 29 (14.6%) studies were included in more than 1 review and they were counted once. These reviews were published between 1996 and 2020 (Table 1), and 11 were meta-analyzed [27,34,39,43,45-50,52]. In all, 8 included papers on RCT designs only [27,39,40,45,46,49,50,52]; 7 had studies of RCTs and other designs such as controlled clinical trials, cohort studies, non-RCTs, quasi-RCTs, and pilot RCTs [34,41-44,47,48]; and 1 review had no RCT design [51].

Most of the reviews were conducted in HICs (13/16, 81%), predominantly from the United States (n=11) [27,34,40-47,49,50,52], United Kingdom (n=7) [39,41-43,46, 47,49], Australia (n=6) [27,39,40,42,47,49], the Netherlands (n=4) [27,46,47,50], and Norway (n=2) [41,49]. Only 3 of the reviews were conducted in LMICs (such as Tanzania, Thailand, Kenya, South Africa, Ethiopia, Nigeria, and India) [48,51,52]. In total, 11 (69%) of the 16 reviews focused on pregnant women alone [27,34,40-45,48-50], 2 focused on both pregnant and postpartum women [39,51], 1 was on both pregnant women and women intending to be pregnant [46], and another was focused on pregnant women and their partners [47] (Table 2).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for new systematic reviews which included searches of databases and registers only.

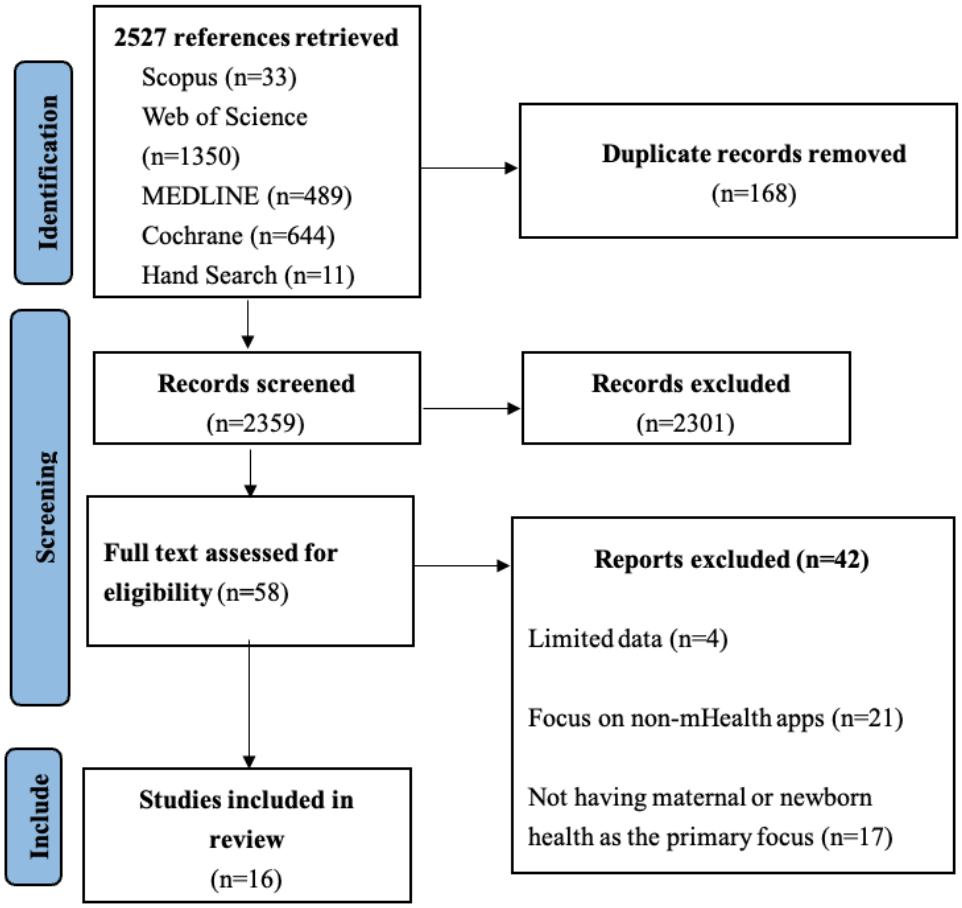

Characterizing mHealth Interventions and Conditions Supported

Different forms of mHealth interventions are implemented to augment the delivery of maternity health care globally. As illustrated in Table 2, the dominant channels or platforms by which mHealth interventions operate are mobile phones [42,43,51] usually in the form of SMS text messages [45-47,49-51]. Quite a significant proportion (3/16, 19%) of the mHealth interventions are delivered through websites [27,39,45]. With respect to the specific maternal conditions, a substantial proportion of the mHealth interventions were used to assist pregnant or postnatal women in overcoming anxiety and depression [27,39,44,50], assist with diabetes in pregnancy [41,45], and gestational weight [42].

Some were also used to help mitigate smoking cessation [43,47] and alcohol prevention or substance use in pregnancy [44,46,49] or to support ANC attendance and facility delivery [48,51,52]. One review focused on asthma in pregnancy [40]. For instance, in the review by Ming et al [45], they assessed modem transmission of blood glucose readings to a centralized health facility's telephone system that interpreted blood glucose readings into audio tones by transmitting them to a database. In the case of Farzandipour et al [42], a greater proportion of their studies used the telephone for prenatal weight monitoring, and provision of educational information. The review by Griffiths et al [43] involved interventions delivered by video or DVD, computer (either PC or laptop), mobile phone, or any portable handheld device [43].

Effectiveness of mHealth Interventions

Considering that different mHealth interventions targeted different conditions in pregnancy and postnatal phases, varied effectiveness outcomes emerged. Meanwhile, all the studies revealed that indeed mHealth interventions are effective for maternal health care, although at varying magnitudes or levels of effectiveness. On depression, there was a uniform agreement among all studies on this condition, demonstrating the extent to which mHealth interventions help mitigate the condition (Table 2). All the reviews on depression highlighted that mHealth interventions were helpful in subsiding depression in pregnancy and childbirth. In the review by Bayrampour et al [39], 4 out of 5 studies revealed positive impacts, with 1 reporting that only 18% of the pregnant and postpartum women met the clinical criteria for depression relative to 79% in the control group, after a 12-week assessment. Similar findings emerged from the review by Lee and Cho [44]. Meanwhile, 2 reviews concluded that mHealth intervention had minimal effect on depression among pregnant women [27,50].

mHealth interventions were reported to be effective in enhancing maternal health care use, as concluded by all the 3 studies focusing on maternal health use. These manifested in ANC enhancement [52], skilled birth attendance at birth leading to a decline in perinatal mortality [51], or both ANC and the use of health facilities for childbirth [48]. Relatedly, the mHealth

XSL•FO
RenderX

interventions focusing on alcohol or drinking and smoking showed effectiveness in diverse ways as indicated by all 4 reviews focusing on these [43,44,46,49]. Another area where significant impacts were recorded was obesity or overweight management, especially during pregnancy [42,47] and diabetes management in pregnancy and newborn outcomes [41,45]. For instance, 66% of studies included in the review by Farzandipour et al [42] revealed positive impacts on gestational weight gain. Some studies in the review by Eberle et al [41] also reported marginally higher neonatal birth weight in the intervention group relative to the nonintervention group.

In total, 3 reviews reported that women were satisfied with the mHealth interventions, partly due to the positive user experience and the benefits [39,45,47]. Among the 4 reviews, 1 highlighted dissatisfaction arising from text messaging–based mHealth interventions due to fatigue caused by multiple messages [51]. The same review indicated that the use of mHealth interventions is compromised by privacy issues in instances where the mobile device (eg, mobile phone) is coshared among family members, when women do not own mobile phones, or when there is illiteracy. Two reviews indicated that acceptability and use of mHealth is enhanced by providing information in varied local languages, using locally based software, and when the target is broadened to include women's significant others [44,51]. Meanwhile, low literacy rate, low accessibility to mobile phones among women in rural locations in LMICs, and uncertainty regarding whether a text is received by the women were some of the identified barriers [51].

Discussion

Principal Findings and Comparison With Other Studies

There is sufficient evidence that mHealth apps effectively mitigate diverse conditions that pregnant and postpartum women encounter. All 16 reviews converged that mHealth apps are useful for maternity health conditions. The specific areas where positive impacts manifested include increased use of maternal health care services (ANC, skilled birth, and postnatal care), reduction in perinatal deaths, weight management in pregnancy, alcohol or drinking and smoking mitigation, as well as diabetes management. These conditions represent some of the leading causes of maternal and newborn morbidity and mortality [53].

The positive impact of mHealth interventions on maternity health resonates with prevailing evidence on the implications of mHealth on overall health care such as the review by Marcolino et al [30], which also revealed positive impacts after synthesizing evidence from available systematic reviews. Additionally, consistent evidence has emerged from different parts of the world [54-56]. Continuous and effective use of mHealth interventions in maternal health care can, therefore, constitute a cornerstone strategy for achieving the first and second targets of the third sustainable development goal, which enjoins all countries to attain fewer than 70 maternal deaths per 100,000 live births and reduce neonatal mortality to no more than 12 deaths per 1000 live births by 2030.

Most of the reviews reported evidence from HICs, which is suggestive of limited use of mHealth interventions for maternal health care services in LMICs. As we did not restrict our search to HICs, it is unlikely that evidence from LMICs is missed due to eligibility issues. Hence, we can infer that there is limited use and evidence about mHealth and maternal health services across LMICs. The United States, United Kingdom, and Australia were the dominant locations where mHealth interventions were implemented and assessed. This situation is quite worrying as little evidence exists on the use and effectiveness of mHealth interventions in LMICs, where worse maternal and neonatal morbidities and mortalities occur. Estimates from the World Bank indicate that the United States has a maternal mortality ratio (MMR) of 19, while the United Kingdom and Australia have an MMR of 6 each, as of 2017 [57]. In the same year, African-based LMICs such as South Sudan, Chad, and Sierra Leone recorded MMRs of 1150, 1140, and 1120, respectively [57]. Hence, LMICs with alarming MMRs need to create a more conducive environment for mHealth to thrive. These may include conscious and continuous strengthening of internet connection and electrification, especially in deprived settings.

Several factors could account for the divide in mHealth use for maternal health care between HICs and LMICs. Mobile devices and the internet are essential prerequisites for mHealth activities to thrive. Meanwhile, mobile phone ownership and internet access are quite difficult in most LMICs. For instance, 1 of the 3 reviews that focused on LMICs reported that some women share mobile phones with other family members and this was a setback toward the maximization of the benefits of mHealth interventions [51]. A recent report on mobile phone surveys in 9 LMICs (Colombia, Ghana, India, Indonesia, Kenya, Mozambique, Nigeria, Rwanda, and South Africa) showed that 90% of people in LMICs do not have a decent internet connection [58]. The report also revealed that less than half of the people in LMICs have access to basic internet. The extent to which limited internet access compromises the operation of mHealth in LMICs has been echoed [59-61]. Undeniably, mobile phone access and use in HICs far exceed LMICs [62] and seem to have no relevance without the internet. All these factors might have culminated in the preponderance of mHealth interventions in HICs relative to LMICs.

Implications for Policy and Future Research Directions

mHealth apps are yielding enormous positive outcomes for maternal well-being. However, the use of mHealth is skewed toward HICs, relative to LMICs where worse maternal conditions emerge from. This is, however, not surprising given the disparity in internet access between LMICs and HICs [63]. Another contributory factor may be the relatively low government expenditure on health care in LMICs, manifesting in limited resource allocation to health care. As of 2019, LMICs averagely spent 5.32% of the gross domestic product on health care as compared with 12.49% average health care expenditure in HICs [64-66]. Hence, an increment in the overall allocation to health care could make it feasible to invest in the most cost-effective mHealth facilities that can offer women the opportunity to benefit from some mHealth interventions that are implementable given the context and acceptability. For

XSL•FO
RenderX

instance, simple text message–based mHealth interventions are implementable with relatively minimal resources relative to more advanced options like audio and video conference interventions [67,68].

The use of mHealth interventions typically requires some basic literacy skills [33].

The overall literacy rate of female individuals may, therefore, have to be enhanced for them to benefit from the mHealth interventions. Countries intending to enhance the benefits of mHealth, especially those in LMICs that do not currently have universal compulsory education, ought to give fair consideration to compulsory education or literacy skills training to ensure that all persons attain at least second-cycle education. This could better place female individuals and their partners to use current and future technological innovations not only in health care, but also in all spheres of life. We are not oblivious to the reality that some LMICs already have universal compulsory education systems [69,70]; for those countries, the fortification of such educational arrangements can enhance gains in the long run. More importantly, backing this with a formidable policy framework can guarantee sustainability and effective implementation of mHealth.

For instance, considering internet connectivity challenges in LMICs compared with HICs, text messaging mHealth interventions may be suitable and relatively easily implementable with political commitment backed by a rigorous policy framework. Second, as mHealth interventions require some level of literacy, there is a need for a conscious effort to reduce illiteracy rates. By this, all persons especially women will be well positioned to use mHealth interventions and ascertain the associated benefits. Globally, governments might have to build partnerships or intensify partnerships with the private sector, particularly the telecommunication companies to widen access, as this can offer some leverage for maximizing the use and impact of mHealth interventions for the well-being of women and newborns.

All countries intending to implement mHealth or maximize the benefits of mHealth could implement national mHealth policies that prioritize maternal issues and also invest in internet connectivity and electrification. These are the critical factors that can bridge the gap between HICs and LMICs as well as the rural and urban residents [71,72]. When women and their partners have uninterrupted internet and electricity and a generally suitable environment for mHealth use, minimal advocacy from ANC providers, nongovernmental agencies, and central government can translate into high use and acceptability. By so doing, the alarming MMR battling LMICs could be reduced, with the bridged disparity between HICs and LMICs, leading to a global decline in MMR.

Future research on this subject could focus on how mHealth interventions work out for special populations such as the people with physical disabilities; persons living with HIV in stereotyped settings; teenagers or adolescents; and persons identified with lesbian, gay, bisexual, transgender, and intersex sexual orientations. These populations are sometimes either not comfortable or have truncated access to services meant for the general population [73-76]. Identification of workable mHealth interventions for special populations will augment efforts to attain the global commitment encapsulated in the third sustainable development goal “ensure healthy lives and promote wellbeing for all at all ages” [77].

Strengths and Limitations

To our knowledge, this is the first systematic review of systematic reviews that has exclusively investigated the effectiveness of mHealth apps or interventions on maternal health globally. Besides, we included all available reviews irrespective of study design and year of publications. This aided in identifying a wide array of mHealth interventions and the myriad of maternity conditions they are used for. Notwithstanding the strengths, the study is not devoid of limitations. First, despite the comprehensive search, we only included reviews written in the English language, which happens to be the working language of the authors. Second, the search in PubMed did not include Medical Subject Headings terms. However, in addition to the database searches, we performed citation searches, specifically forward and backward citation searches, to identify additional reviews that might have been missed in the database search. This was done to identify additional reviews that might have been missed in the database search.

Conclusions

The effectiveness of mHealth apps in maternal health care is well established in the literature. This manifests in the positive impact toward management and reduction in anxiety and depression, diabetes in pregnancy, and gestational weight. Other aspects include smoking reduction or cessation during pregnancy and enhancement in maternal health care use. Meanwhile, mHealth apps are skewed toward HICs, likely due to the advancement in technology and other resources required to optimize the inherent utility of mHealth apps. Considering that mHealth apps dominate in HICs, LMICs must consider pragmatic approaches to improve their availability and use. Such approaches may include the adoption of workable mHealth policies and political commitment through increment in budget allocation for health care to offer cost-effective mHealth interventions for maternal health care delivery (eg, nonandroid text messages).

Acknowledgments

The authors are grateful to Prof Joshua Ka Ho Mok for his leadership and supervision. This study was funded by the Hong Kong Research Grants Council (RGC) through its Postdoctoral Fellowship Scheme. The authors are therefore thankful to the RGC and Lingnan University for providing financial assistance and resources.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article (and Multimedia Appendices 1-3).

Authors' Contributions

EKA conceived the study, developed the protocol, and contributed to the development of all sections of the manuscript. PAA contributed to screening, quality assessment, and review of drafts of the manuscript for important intellectual content. OE independently assessed the quality of all included reviews and reviewed drafts of the manuscript to ensure important intellectual content. All authors reviewed the final draft and approved it for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e49510_app1.docx]

Multimedia Appendix 2
Search Strategy.
[DOCX File , 23 KB - jmir_v26i1e49510_app2.docx]

Multimedia Appendix 3
Joanna Briggs Institute (JBI) assessment tool.
[DOCX File , 23 KB - jmir_v26i1e49510_app3.docx]

References

1. DeNicola N, Grossman D, Marko K, Sonalkar S, Butler Tobah YS, Ganju N, et al. Telehealth interventions to improve obstetric and gynecologic health outcomes: a systematic review. Obstet Gynecol 2020 Feb;135(2):371-382 [FREE Full text] [doi: 10.1097/AOG.0000000000003646] [Medline: 31977782]
2. World Health Organization. mHealth: New Horizons for Health Through Mobile Technologies: Second Global Survey on eHealth. Geneva: World Health Organization; 2011.
3. Kahn JG, Yang JS, Kahn JS. 'Mobile' health needs and opportunities in developing countries. Health Aff (Millwood) 2010 Feb;29(2):252-258. [doi: 10.1377/hlthaff.2009.0965] [Medline: 20348069]
4. Chib A, van Velthoven MH, Car J. mHealth adoption in low-resource environments: a review of the use of mobile healthcare in developing countries. J Health Commun 2015;20(1):4-34. [doi: 10.1080/10810730.2013.864735] [Medline: 24673171]
5. Colaci D, Chaudhri S, Vasan A. mHealth interventions in low-income countries to address maternal health: a systematic review. Ann Glob Health 2016;82(5):922-935 [FREE Full text] [doi: 10.1016/j.aogh.2016.09.001] [Medline: 28283147]
6. Walker PGT, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, et al. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. Science 2020;369(6502):413-422 [FREE Full text] [doi: 10.1126/science.abc0035] [Medline: 32532802]
7. Okereke M, Ukor NA, Adebisi YA, Ogunkola IO, Favour Iyagbaye E, Adiela Owhor G, et al. Impact of COVID-19 on access to healthcare in low- and middle-income countries: current evidence and future recommendations. Int J Health Plann Manage 2021 Jan;36(1):13-17. [doi: 10.1002/hpm.3067] [Medline: 32857892]
8. Blumenthal D, Fowler EJ, Abrams M, Collins SR. Covid-19 - implications for the health care system. N Engl J Med 2020 Oct 08;383(15):1483-1488 [FREE Full text] [doi: 10.1056/NEJMsb2021088] [Medline: 32706956]
9. Pant S, Koirala S, Subedi M. Access to maternal health services during COVID-19. Europasian J Med Sci 2020;2(2):46-50. [doi: 10.46405/ejms.v2i2.110]
10. Townsend R, Chmielewska B, Barratt I, Kalafat E, van der Meulen J, Gurol-Urganci I, et al. Global changes in maternity care provision during the COVID-19 pandemic: a systematic review and meta-analysis. EClinicalMedicine 2021 Jul;37:100947 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100947] [Medline: 34195576]
11. Ameyaw EK, Ahinkorah BO, Seidu AA, Njue C. Impact of COVID-19 on maternal healthcare in Africa and the way forward. Arch Public Health 2021 Dec 10;79(1):223 [FREE Full text] [doi: 10.1186/s13690-021-00746-6] [Medline: 34886893]
12. Evans K, Donelan J, Rennick-Egglestone S, Cox S, Kuipers Y. Review of mobile apps for women with anxiety in pregnancy: maternity care professionals' guide to locating and assessing anxiety apps. J Med Internet Res 2022 Mar 23;24(3):e31831 [FREE Full text] [doi: 10.2196/31831] [Medline: 35319482]

XSL•FO
RenderX

13. Feng YY, Korale-Liyanage S, Jarde A, McDonald SD. Psychological or educational eHealth interventions on depression, anxiety or stress following preterm birth: a systematic review. J Reprod Infant Psychol 2021 Apr;39(2):140-152. [doi: 10.1080/02646838.2020.1750576] [Medline: 32336122]
14. Haddad SM, Souza RT, Cecatti JG. Mobile technology in health (mHealth) and antenatal care-searching for apps and available solutions: a systematic review. Int J Med Inform 2019 Jul;127:1-8. [doi: 10.1016/j.ijmedinf.2019.04.008] [Medline: 31128820]
15. Lavender T, Richens Y, Milan SJ, Smyth RMD, Dowswell T. Telephone support for women during pregnancy and the first six weeks postpartum. Cochrane Database Syst Rev 2013 Jul 18;2013(7):CD009338 [FREE Full text] [doi: 10.1002/14651858.CD009338.pub2] [Medline: 23881662]
16. Linde DS, Korsholm M, Katanga J, Rasch V, Lundh A, Andersen MS. One-way SMS and healthcare outcomes in Africa: systematic review of randomised trials with meta-analysis. PLoS One 2019;14(6):e0217485 [FREE Full text] [doi: 10.1371/journal.pone.0217485] [Medline: 31170176]
17. Ezezika O, Varatharajan C, Racine S, Ameyaw EK. The implementation of a maternal mHealth project in South Africa: lessons for taking mHealth innovations to scale. Afr J Sci Technol Innov Dev 2021 Dec 20;14(7):1798-1812. [doi: 10.1080/20421338.2021.1985946]
18. Willcox M, Moorthy A, Mohan D, Romano K, Hutchful D, Mehl G, et al. Mobile technology for community health in Ghana: is maternal messaging and provider use of technology cost-effective in improving maternal and child health outcomes at scale? J Med Internet Res 2019 Feb 13;21(2):e11268 [FREE Full text] [doi: 10.2196/11268] [Medline: 30758296]
19. Hospital authority launches mobile application "HA Go". Press Releases. 2019. URL: https://www.info.gov.hk/gia/general/201912/12/P2019121200329.htm [accessed 2024-05-01]
20. Odea SSY. Global unique mobile subscribers from 2010-2025, by region (millions). Statista. 2020. URL: https://www.statista.com/statistics/740154/worldwide-unique-mobile-subscribers-by-region/ [accessed 2024-03-04]
21. Mccool J, Hill J, Dobson R, Whittaker R. Access to ICT in the Pacific Islands region: a brief report. Pac Health Dialog 2020;21(6):347-350. [doi: 10.26635/phd.2020.642]
22. Burki T. Developing countries in the digital revolution. Lancet 2018;391(10119):417. [doi: 10.1016/S0140-6736(18)30191-0] [Medline: 29407022]
23. Maternal mortality. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality [accessed 2024-03-04]
24. Lee SH, Nurmatov UB, Nwaru BI, Mukherjee M, Grant L, Pagliari C. Effectiveness of mHealth interventions for maternal, newborn and child health in low- and middle-income countries: systematic review and meta-analysis. J Glob Health 2016 Jun;6(1):010401 [FREE Full text] [doi: 10.7189/jogh.06.010401] [Medline: 26649177]
25. Almohanna AA, Win KT, Meedya S. Effectiveness of internet-based electronic technology interventions on breastfeeding outcomes: systematic review. J Med Internet Res 2020 May 29;22(5):e17361 [FREE Full text] [doi: 10.2196/17361] [Medline: 32469315]
26. Ambia J, Mandala J. A systematic review of interventions to improve prevention of mother-to-child HIV transmission service delivery and promote retention. J Int AIDS Soc 2016;19(1):20309 [FREE Full text] [doi: 10.7448/IAS.19.1.20309] [Medline: 27056361]
27. Chae J, Kim HK. Internet-based prenatal interventions for maternal health among pregnant women: a systematic review and meta-analysis. Children and Youth Services Review 2021 Aug;127:106079. [doi: 10.1016/j.childyouth.2021.106079]
28. Feldman N, Back D, Boland R, Torous J. A systematic review of mHealth application interventions for peripartum mood disorders: trends and evidence in academia and industry. Arch Womens Ment Health 2021 Dec;24(6):881-892 [FREE Full text] [doi: 10.1007/s00737-021-01138-z] [Medline: 33929636]
29. Bossman E, Johansen MA, Zanaboni P. mHealth interventions to reduce maternal and child mortality in Sub-Saharan Africa and Southern Asia: a systematic literature review. Front Glob Womens Health 2022;3:942146 [FREE Full text] [doi: 10.3389/fgwh.2022.942146] [Medline: 36090599]
30. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
31. McCool J, Dobson R, Whittaker R, Paton C. Mobile health (mHealth) in low- and middle-income countries. Annu Rev Public Health 2022 Apr 05;43:525-539 [FREE Full text] [doi: 10.1146/annurev-publhealth-052620-093850] [Medline: 34648368]
32. Whittaker R, Merry S, Dorey E, Maddison R. A development and evaluation process for mHealth interventions: examples from New Zealand. J Health Commun 2012;17 Suppl 1(Suppl 1):11-21 [FREE Full text] [doi: 10.1080/10810730.2011.649103] [Medline: 22548594]
33. Kreps GL. The relevance of health literacy to mHealth. ISU 2017 Jun 26;37(2):123-130 [FREE Full text] [doi: 10.3233/ISU-170828]
34. Lau Y, Htun TP, Wong SN, Tam WSW, Klainin-Yobas P. Efficacy of internet-based self-monitoring interventions on maternal and neonatal outcomes in perinatal diabetic women: a systematic review and meta-analysis. J Med Internet Res 2016 Aug 15;18(8):e220 [FREE Full text] [doi: 10.2196/jmir.6153] [Medline: 27526637]

XSL•FO
RenderX

35. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906 [FREE Full text] [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
36. Newborn health. World Health Organization. 2023. URL: https://www.who.int/westernpacific/health-topics/newborn-health [accessed 2024-05-02]
37. Krnic Martinic M, Pieper D, Glatt A, Puljak L. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. BMC Med Res Methodol 2019 Nov 04;19(1):203 [FREE Full text] [doi: 10.1186/s12874-019-0855-0] [Medline: 31684874]
38. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. Int J Evid Based Healthc 2015 Sep;13(3):132-140 [FREE Full text] [doi: 10.1097/XEB.0000000000000055] [Medline: 26360830]
39. Bayrampour H, Trieu J, Tharmaratnam T. Effectiveness of eHealth interventions to reduce perinatal anxiety: a systematic review and meta-analysis. J Clin Psychiatry 2019 Jan 22;80(1):18r12386 [FREE Full text] [doi: 10.4088/JCP.18r12386] [Medline: 30688418]
40. Daly LM, Horey D, Middleton PF, Boyle FM, Flenady V. The effect of mobile app interventions on influencing healthy maternal behavior and improving perinatal health outcomes: systematic review. JMIR Mhealth Uhealth 2018 Aug 09;6(8):e10012 [FREE Full text] [doi: 10.2196/10012] [Medline: 30093368]
41. Eberle C, Loehnert M, Stichling S. Clinical effectiveness of different technologies for diabetes in pregnancy: systematic literature review. J Med Internet Res 2021 Apr 28;23(4):e24982 [FREE Full text] [doi: 10.2196/24982] [Medline: 33908894]
42. Farzandipour M, Nabovati E, Anvari S, Vahedpoor Z, Sharif R. Phone-based interventions to control gestational weight gain: a systematic review on features and effects. Inform Health Soc Care 2020 Jan;45(1):15-30 [FREE Full text] [doi: 10.1080/17538157.2018.1540421] [Medline: 30479185]
43. Griffiths SE, Parsons J, Naughton F, Fulton EA, Tombor I, Brown KE. Are digital interventions for smoking cessation in pregnancy effective? a systematic review and meta-analysis. Health Psychol Rev 2018 Dec;12(4):333-356 [FREE Full text] [doi: 10.1080/17437199.2018.1488602] [Medline: 29912621]
44. Lee Y, Cho S. Technology-supported interventions for pregnant women: a systematic review. Comput Inform Nurs 2019 Oct;37(10):501-512 [FREE Full text] [doi: 10.1097/CIN.0000000000000535] [Medline: 31162146]
45. Ming WK, Mackillop LH, Farmer AJ, Loerup L, Bartlett K, Levy JC, et al. Telemedicine technologies for diabetes in pregnancy: a systematic review and meta-analysis. J Med Internet Res 2016 Nov 09;18(11):e290 [FREE Full text] [doi: 10.2196/jmir.6556] [Medline: 27829574]
46. Oh SS, Moon JY, Chon D, Mita C, Lawrence JA, Park EC, et al. Effectiveness of digital interventions for preventing alcohol consumption in pregnancy: systematic review and meta-analysis. J Med Internet Res 2022 Apr 11;24(4):e35554 [FREE Full text] [doi: 10.2196/35554] [Medline: 35404257]
47. Overdijkink SB, Velu AV, Rosman AN, van Beukering MD, Kok M, Steegers-Theunissen RP. The usability and effectiveness of mobile health technology-based lifestyle and medical intervention apps supporting health care during pregnancy: systematic review. JMIR Mhealth Uhealth 2018 Apr 24;6(4):e109 [FREE Full text] [doi: 10.2196/mhealth.8834] [Medline: 29691216]
48. Rahman MO, Yamaji N, Nagamatsu Y, Ota E. Effects of mHealth interventions on improving antenatal care visits and skilled delivery care in low- and middle-income countries: systematic review and meta-analysis. J Med Internet Res 2022 Apr 22;24(4):e34061 [FREE Full text] [doi: 10.2196/34061] [Medline: 35451987]
49. Silang K, Sanguino H, Sohal PR, Rioux C, Kim HS, Tomfohr-Madsen LM. eHealth interventions to treat substance use in pregnancy: a systematic review and meta-analysis. Int J Environ Res Public Health 2021 Sep 22;18(19):9952 [FREE Full text] [doi: 10.3390/ijerph18199952] [Medline: 34639252]
50. Silang KA, Sohal PR, Bright KS, Leason J, Roos L, Lebel C, et al. eHealth interventions for treatment and prevention of depression, anxiety, and insomnia during pregnancy: systematic review and meta-analysis. JMIR Ment Health 2022 Feb 21;9(2):e31116 [FREE Full text] [doi: 10.2196/31116] [Medline: 35188471]
51. Sondaal SFV, Browne JL, Amoakoh-Coleman M, Borgstein A, Miltenburg AS, Verwijs M, et al. Assessing the effect of mHealth interventions in improving maternal and neonatal care in low- and middle-income countries: a systematic review. PLoS One 2016;11(5):e0154664 [FREE Full text] [doi: 10.1371/journal.pone.0154664] [Medline: 27144393]
52. Wagnew F, Dessie G, Alebel A, Mulugeta H, Belay YA, Abajobir AA. Does short message service improve focused antenatal care visit and skilled birth attendance? a systematic review and meta-analysis of randomized clinical trials. Reprod Health 2018 Nov 22;15(1):191 [FREE Full text] [doi: 10.1186/s12978-018-0635-z] [Medline: 30466453]
53. What are examples and causes of maternal morbidity and mortality? US Department of health and human services. 2020. URL: https://www.nichd.nih.gov/health/topics/maternal-morbidity-mortality/conditioninfo/causes# [accessed 2024-03-01]
54. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020;3(1):4 [FREE Full text] [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
55. Santo K, Redfern J. The potential of mHealth applications in improving resistant hypertension self-assessment, treatment and control. Curr Hypertens Rep 2019;21(10):81 [FREE Full text] [doi: 10.1007/s11906-019-0986-z] [Medline: 31598792]

XSL•FO
RenderX

56. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. NPJ Digit Med 2018;1(1):12 [FREE Full text] [doi: 10.1038/s41746-018-0021-9] [Medline: 31304297]
57. WHO, UNICEF, UNFPA, World Bank Group, united nations population division. trends in maternal mortality: 2000 to 2017. Geneva, World Health Organization 2019. The World Bank Group. 2019. URL: https://data.worldbank.org/indicator/SH.STA.MMRT [accessed 2024-05-01]
58. Alliance for Affordable Internet. 2022. URL: https://tinyurl.com/2smwxswc [accessed 2024-03-01]
59. Aamir J, Ali SM, Kamel Boulos M, Anjum N, Ishaq M. Enablers and inhibitors: a review of the situation regarding mHealth adoption in low- and middle-income countries. Health Policy Technol 2018 Mar;7(1):88-97. [doi: 10.1016/j.hlpt.2017.11.005]
60. Feroz A, Kadir MM, Saleem S. Health systems readiness for adopting mhealth interventions for addressing non-communicable diseases in low- and middle-income countries: a current debate. Glob Health Action 2018;11(1):1496887 [FREE Full text] [doi: 10.1080/16549716.2018.1496887] [Medline: 30040605]
61. Leonard E, de Kock I, Bam W. Barriers and facilitators to implementing evidence-based health innovations in low- and middle-income countries: a systematic literature review. Eval Program Plann 2020 Oct;82:101832. [doi: 10.1016/j.evalprogplan.2020.101832] [Medline: 32585317]
62. Bastawrous A, Armstrong MJ. Mobile health use in low- and high-income countries: an overview of the peer-reviewed literature. J R Soc Med 2013 Apr;106(4):130-142 [FREE Full text] [doi: 10.1177/0141076812472620] [Medline: 23564897]
63. Billon M, Crespo J, Lera-López F. Internet, educational disparities, and economic growth: differences between low- middle and high-income countries. In: Marszk A, Lechman E, Kaur H, editors. Catalyzing Development Through ICT Adoption: The Developing World Experience. Cham: Springer International Publishing; 2017:51-68.
64. Current health expenditure (% of GDP) - high income. The World Bank. 2023. URL: https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?locations=XD [accessed 2024-05-01]
65. Global health expenditure database. World Health Organization. 2023. URL: https://apps.who.int/nha/database [accessed 2024-05-01]
66. Current health expenditure (% of GDP) - low & middle income. The World Bank. 2023. URL: https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?locations=XO [accessed 2024-03-01]
67. Hurt K, Walker RJ, Campbell JA, Egede LE. mHealth interventions in low and middle-income countries: a systematic review. Glob J Health Sci 2016;8(9):54429 [FREE Full text] [doi: 10.5539/gjhs.v8n9p183] [Medline: 27157176]
68. Khatun F, Heywood AE, Ray PK, Hanifi SMA, Bhuiya A, Liaw ST. Determinants of readiness to adopt mHealth in a rural community of Bangladesh. Int J Med Inform 2015 Oct;84(10):847-856. [doi: 10.1016/j.ijmedinf.2015.06.008] [Medline: 26194141]
69. Ekundayo OS. The right to free and compulsory primary education in Ghana: lessons for other African countries. J Law Policy Glob 2018;69:105.
70. Waydon EB, Ying L, Ketter BL. Free and compulsory primary education policy in Liberia: gap between promise and actual performance. Educ Res Int 2016;5(1):8-24.
71. Byambasuren O, Beller E, Hoffmann T, Glasziou P. Barriers to and facilitators of the prescription of mHealth apps in Australian general practice: qualitative study. JMIR Mhealth Uhealth 2020;8(7):e17447 [FREE Full text] [doi: 10.2196/17447] [Medline: 32729839]
72. Johnson A, Shukla N, Halley M, Nava V, Budaraju J, Zhang L, et al. Barriers and facilitators to mobile health and active surveillance use among older adults with skin disease. Health Expect 2021;24(5):1582-1592 [FREE Full text] [doi: 10.1111/hex.13229] [Medline: 34190397]
73. Lindberg C, Lewis-Spruill C, Crownover R. Barriers to sexual and reproductive health care: urban male adolescents speak out. Issues Compr Pediatr Nurs 2006;29(2):73-88 [FREE Full text] [doi: 10.1080/01460860600677577] [Medline: 16772237]
74. Kiplagat J, Mwangi A, Chasela C, Huschke S. Challenges with seeking HIV care services: perspectives of older adults infected with HIV in western Kenya. BMC Public Health 2019;19(1):929 [FREE Full text] [doi: 10.1186/s12889-019-7283-2] [Medline: 31296195]
75. Grut L, Mji G, Braathen SH, Ingstad B. Accessing community health services: challenges faced by poor people with disabilities in a rural community in South Africa. Afr J Disabil 2012;1(1):19 [FREE Full text] [doi: 10.4102/ajod.v1i1.19] [Medline: 28729977]
76. Karakaya S, Kutlu FY. LGBT individuals' opinions about their health care experiences: a qualitative research study. J Nurs Manag 2021;29(1):24-31 [FREE Full text] [doi: 10.1111/jonm.13199] [Medline: 33128281]
77. Transforming our world: the 2030 agenda for sustainable development. United Nations. 2015. URL: https://sdgs.un.org/2030agenda [accessed 2024-05-01]

Abbreviations

ANC: antenatal care
HIC: high-income country
LMIC: low- and middle-income country
mHealth: mobile health

XSL•FO
RenderX

MMR: maternal mortality ratio
MOTECH: Mobile Technology for Community Health
PRIOR: Preferred Reporting Items for Overviews of Review
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: Prospective Register of Systematic Review
RCT: randomized controlled trial
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 31.05.23; peer-reviewed by U Charantimath, K Matthias; comments to author 27.08.23; revised version received 29.10.23; accepted 16.04.24; published 29.05.24.

Please cite as:
Ameyaw EK, Amoah PA, Ezezika O
Effectiveness of mHealth Apps for Maternal Health Care Delivery: Systematic Review of Systematic Reviews
J Med Internet Res 2024;26:e49510
URL: https://www.jmir.org/2024/1/e49510
doi:10.2196/49510
PMID:38810250

©Edward Kwabena Ameyaw, Padmore Adusei Amoah, Obidimma Ezezika. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Methods, Indicators, and End-User Involvement in the Evaluation of Digital Health Interventions for the Public: Scoping Review

Vera Weirauch[1,2], MA; Clarissa Soehnchen[1], MLB; Anja Burmann[2], Dr rer med; Sven Meister[1,2], Prof Dr

[1]Health Informatics, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[2]Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Corresponding Author:
Vera Weirauch, MA
Fraunhofer Institute for Software and Systems Engineering
Speicherstraße 6
Dortmund, 44147
Germany
Phone: 49 023197677329
Email: vera.weirauch@isst.fraunhofer.de

Abstract

Background: Digital health interventions (DHIs) have the potential to enable public end users, such as citizens and patients, to manage and improve their health. Although the number of available DHIs is increasing, examples of successfully established DHIs in public health systems are limited. To counteract the nonuse of DHIs, they should be comprehensively evaluated while integrating end users. Unfortunately, there is a wide variability and heterogeneity according to the approaches of evaluation, which creates a methodological challenge.

Objective: This scoping review aims to provide an overview of the current established processes for evaluating DHIs, including methods, indicators, and end-user involvement. The review is not limited to a specific medical field or type of DHI but offers a holistic overview.

Methods: This scoping review was conducted following the JBI methodology for scoping reviews based on the framework by Arksey & O'Malley and complies with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Three scientific databases (PubMed, Scopus, and Science Direct) were searched in April 2023. English and German studies between 2008 and 2023 were considered when evaluating DHIs that explicitly address public end users. The process of study selection was carried out by several researchers to avoid reviewer bias.

Results: The search strategy identified 9618 publications, of which 160 were included. Among these included articles, 200 evaluations were derived and analyzed. The results showed that there is neither a consensus on the methods to evaluate DHIs nor a commonly agreed definition or usage of the evaluated indicators, which results in a broad variety of evaluation practices. This aligns with observations of the existing literature. It was found that there is a lack of references to existing frameworks for the evaluation of DHIs. The majority of the included studies referred to user-centered approaches and involved end users in the evaluation process. As assistance for people developing and evaluating DHIs and as a basis for thinking about appropriate ways to evaluate DHIs, a results matrix was created where the findings were combined per DHI cluster. Additionally, general recommendations for the evaluators of DHIs were formulated.

Conclusions: The findings of this scoping review offer a holistic overview of the variety and heterogeneity according to the approaches of evaluation of DHIs for public end users. Evaluators of these DHIs should be encouraged to reference established frameworks or measurements for justification. This would ease the transferability of the results among similar evaluation studies within the digital health sector, thereby enhancing the coherence and comparability of research in this area.

(J Med Internet Res 2024;26:e55714) doi:10.2196/55714

KEYWORDS

digital health; digital health intervention; public end user; evaluation methods; evaluation criteria; end-user involvement; scoping review

XSL•FO
RenderX

Introduction

Background

The potential of digital health interventions (DHIs) to improve care processes is widely recognized [1,2]. Particularly for public end users, DHIs have the potential to enable them to manage and improve their health by promoting health, supporting behavior change, personalizing health care delivery, and giving them the chance to individually organize health care [1-6]. Consequently, the development of DHIs is increasing, and a growing number of them are available in the market [3,4]. Nevertheless, examples of successfully established DHIs in public health systems are limited [7], and several challenges exist in this rapidly growing field, for example, methodological challenges [1,6,8,9].

Before elaborating on the challenges, it is useful to consider the term DHI from a definitional point of view. There are various terms for digital technologies in the health sector, such as eHealth, mobile health (mHealth), digital health services, and DHIs. To avoid conceptual ambiguity, this paper adopts the definitional approach of the World Health Organization (WHO), defining a DHI as "a discrete functionality of the digital technology to achieve health sector objectives" [10]. Following this definition, the growing field of digital public health interventions (DPHIs) can be regarded as a distinct subset of DHIs. DPHIs primarily focus on improving health and well-being at the population level, rather than at the individual level [11-13]. For this study, the definition of the key term "evaluation" has also been adopted according to the WHO, which has defined evaluation as "The systematic and objective assessment of an ongoing or completed intervention, with the aim of determining the fulfillment of objectives, efficiency, effectiveness, impact, and sustainability" [14]. Additionally, it is important to specify that in the context of this scoping review, "public end users" are defined as individuals, such as citizens and patients, who directly interact with digital health tools or services. Unlike health care professionals or caregivers, who may use DHIs as part of their job, public end users engage with these tools to meet their personal health needs. They are not limited to any specific patient group or demographic.

Status Quo on the Evaluation of DHIs

Overall, DHIs are characterized as complex interventions. This complexity is partly due to their interdisciplinary and multisectoral nature, involving a diverse mix of stakeholders, including patients, various health professionals, relatives, policymakers, and health insurers [1,3,8,15]. Additionally, DHIs consist of multiple interdependent components, both technical and nontechnical [4,8,15], and serve multiple aims, such as providing information, improving communication, facilitating data sharing, and enabling monitoring [1]. The value and impact of DHIs extend beyond clinical outcomes and also encompass organizational, behavioral, and technical dimensions [7]. These characteristics also apply to DPHIs [11-13]. Consequently, the evaluation of DHIs as well as DPHIs is a methodological challenge [1,6,8,9], and existing established methods like health technology assessment (HTA), evidence-based medicine (EBM), and randomized controlled trials (RCTs) are limited in their application [16-18]. HTA is a process for the systematic evaluation of medical procedures and technologies, with a focus on assessing the health benefits and costs associated with the use of therapeutics, medical products, and diagnostic procedures [17]. Therefore, this methodology is not immediately appropriate in the context of the evaluation of DHIs [17,19]. EBM considers that decisions about the care of patients should be based on the best available external clinical evidence from systematic research in combination with the clinical expertise of health professionals. Regarding DHIs, there are more aspects to evaluate than clinical aspects, so it requires other distinct approaches than the usual suggested process to gather evidence within EBM [20]. Although RCTs are considered as the gold standard [6], especially in the context of evaluating DHIs, they are a much-discussed topic. Due to the complexity of DHIs, the applicability of RCTs is widely criticized [3,6,9,15,16,18]. A comprehensive evaluation of DHIs is essential for generating robust evidence [14,20], contributing to their long-term successful implementation and aiding in the realization of the full benefits of DHIs [15,20]. This principle also applies to DPHIs, which should adhere to an evidence- and needs-based approach, incorporating a participatory user-targeted development design to enhance acceptance of the intervention within the population [11,13,21].

The literature describes considerable variability and heterogeneity in how DHIs should be evaluated, which is attributable to the absence of a standardized, established, and broadly applicable approach for evaluating DHIs [3,6,15,20,22-25]. One possible explanation for the absence of a standardized and broadly applicable guideline for the evaluation of DHIs is that the evaluation of these technologies is complex and complicated by various fundamental issues [17]. Various European authorities and scientists have addressed this issue by developing and publishing proposed frameworks for the evaluation of DHIs, such as the Monitoring and Evaluation Guideline of the WHO, the Evidence Standards framework for digital health technologies of the National Institute for Health and Care Excellence (NICE), the report of the Expert Panel on effective ways of investing in health (EXPH), the Swiss Evaluation framework by Kowatsch et al, and the approach of Murray et al [1,5,14,17,24]. In this review, there is an orientation toward the WHO and NICE frameworks, for instance, in categorizing the evaluated DHIs and determining the evaluation criteria.

Status Quo of End-User Involvement in the Evaluation of DHIs

For various reasons, such as counteracting the nonuse of DHIs as well as DPHIs, evidence generation and evaluation should be practice-oriented, necessitating the integration of end users in this process [3,11,13,15,21,25-28]. In recent years, approaches, such as user-centered design (UCD), participatory health research (PHR), and public and patient involvement (PPI), have gained increasing importance in the health care sector, each contributing to the overarching goal of creating patient-centered, accessible, and equitable health care solutions.

UCD or human-centered design is rooted in human-system interactions and can be seen as a set of principles and strategies

XSL•FO
RenderX

in the design and development of interactive digital health solutions, emphasizing the iterative research, design, and evaluation of services and systems by involving end users and stakeholders throughout the project life cycle [29-31]. PHR can be understood as a research paradigm rather than a research method, aiming to increase the participation of individuals whose life or work is the subject of research throughout the research process. The research process should be realized as a partnership among involved stakeholders (ie, with each other), instead of research on people as passive objects. Involved stakeholders could include academic researchers, health professionals, policymakers, and members of civil society [32]. INVOLVE, founded by the National Institute for Health Research (NIHR) and taken over by the NIHR Centre for Engagement and Dissemination in April 2020, defines PPI as research that is realized "with" or "by" members of the public, rather than "to," "about," or "for" them [33]. The term "public and patients" includes current, former, and potential patients; people who use health and social care services; people from organizations who represent other people using these services; and carers [34-36].

To sum up, regarding the definitional approaches of PHR and PPI, overlaps can be seen, especially regarding the statements that research should be performed with the research subjects rather than about them. UCD aligns with this principle but is more specifically focused on the design, development, and evaluation of interactive digital health solutions than on the scientific research context.

Through the involvement of end users in evaluation processes, acceptance and usability problems can be mitigated [15,26] and health interventions can be designed in a target group–specific and needs-based manner [27].

Objectives

This scoping review has been conducted to understand and provide an overview of the current established processes for evaluating DHIs for public end users, such as citizens and patients. Previous reviews have focused on investigating which aspects of DHIs were evaluated during different development phases [20], investigating evaluation methods regarding specific criteria [28,37,38] or systems [39], investigating evaluation methods in specific medical contexts [22], investigating concrete evaluation methods despite RCTs [16], investigating economic evaluations of preventive DPHIs [12], or investigating general methods to evaluate the effects of DHIs for citizens by performing a review about reviews [6]. This review differs from those mentioned earlier as it provides a holistic overview of evaluation processes for DHIs for public end users, which is not limited to a specific medical field. The objectives are to capture the (1) evaluation methods and (2) evaluation criteria that are currently used to evaluate DHIs. Additionally, there is a focus on investigating the involvement of public end users in the further development of DHIs specifically developed for them. Therefore, the review addresses the following research questions (RQs):

1. Which research methods are used to evaluate DHIs for individuals or public end users?
2. Which evaluation criteria can be identified? Which evaluation criteria have been investigated?
3. In which way are individuals or public end users involved in the evaluation process?

In summary, this scoping review aims to provide an overview of the currently established processes for evaluating DHIs, including methods, indicators, and end-user involvement.

Methods

Overview

Through a scoping review, broad topics can be explored and gaps in the evidence can be identified [40,41]. Due to the nature of scoping reviews, we did not formally assess the risk of bias or methodological quality of the included studies [40,41]. This review was conducted based on the methodological framework for scoping reviews of Arksey & O'Malley [42] and adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [43] incorporating the updates published by Peters et al [44] (Multimedia Appendix 1). The review protocol was registered a priori with the Center for Open Science (OSF) [45].

Search Strategy

The primary information sources for this scoping review were scientific databases, namely PubMed, Scopus, and ScienceDirect. Additionally, Google Scholar and the reference lists of included papers were manually screened. The search was conducted in April 2023. The search string, developed iteratively by 3 domain experts, was equally applied across the abovementioned databases, and is detailed in Table 1. The search was carried out without the assistance of librarians. The main search terms "evaluation," "digital health intervention," and "user-centered" along with their synonyms were combined using Boolean operators (Table 1).

Table 1. Search string.

Operator	Field	Context	Focus
AND (main search term)	Evaluation (A)	Digital health intervention (B)	User-centered (C)
OR (synonyms)	evaluat*, evaluation method*, formative evaluation, summative evaluation, assess*	digital health intervention*, digital health technology, digital public health intervention*, digital health service*, electronic health record, mHealth, eHealth, health information technology, health Information platform, health diary	user-oriented, user-centered

XSL•FO
RenderX

The search string for each of the 3 databases is listed in Multimedia Appendix 2. The results of each database were stored in Citavi (Swiss Academic Software) and exported into Microsoft Excel files.

Eligibility Criteria

Literature is eligible for inclusion if it describes evaluation methods or evaluation criteria for DHIs primarily aimed at public end users. Additionally, DHIs must be usable without the assistance of a health professional. The scoping review included all types of DHIs, including DPHIs without any exclusions, as long as the evaluation included an explorable version of the DHI. Moreover, literature that documents the integration of end users in the evaluation process is also eligible for inclusion. According to a previous manual search, it appears that the term DHI was more commonly used approximately 15 years ago. Additionally, the authors assumed that the introduction of the iPhone in 2007 led to an increase in mHealth evaluation studies. Therefore, literature published in the last 15 years (2008-2023) was included in the search. The eligibility criteria relevant to this scoping review are detailed in Textbox 1.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Targeted population: Primary end users of digital health interventions (DHIs) are individuals, such as patients and the public. The DHI can be used on its own. No limitations on the number of participants, their gender, or their origin.
- Study design: Original peer-reviewed studies, conference papers, book chapters, and grey literature, such as organizational reports.
- Context/field–DHI: DHIs can occur in the form of, for example, patient portals, platforms, web or mobile applications, or patient access to electronic health records. No exclusion of specific types of DHIs. There should be an explorable version of the DHI, which means at least a low-fidelity prototype, within the evaluation.
- Context/field–evaluation: The study proposes or describes the evaluation or assessment of a DHI. Iterative evaluations primarily used for requirement engineering or assessing basic needs are excluded or not focused.
- Context/field–end-user involvement: End users are kind of actively integrated into the evaluation process.
- Accessibility: Full text is freely available on the internet or after contacting the author.
- Language: English or German.
- Year: Published in the last 15 years (2008-2023).

Exclusion criteria

- Targeted population: End users are health professionals. The DHI can only be used with assistance from a health professional.
- Study design: Not peer-reviewed papers, preprints, reviews, comments, presentations, protocols, or posters.
- Context/field–DHI: There is no existing explorable version of the DHI.
- Context/field–evaluation: The paper describes a framework or a general overview of evaluation types and methods. The study only focuses on requirements of engineering processes, which means there is no explorable DHI (prototype or final DHI) available. Studies addressing only specific health issues designed to answer clinical research questions.
- Context/field–end-user involvement: The study did not include end users in the evaluation process. End users are only passive data objects, which means that they are not addressed with specific questions.
- Accessibility: Full text is not available on the internet or by contacting the author.
- Language: Other than English or German.
- Year: Literature older than 15 years.

Process of Study Selection

The electronic search results were stored in Microsoft Excel. An initial selection based on language and publication year was partially conducted within the scientific databases. To select the search results, duplications were first removed. Subsequently, titles and abstracts were screened, and literature not meeting the eligibility criteria was excluded. To avoid reviewer bias, the screening process was partially conducted by different researchers. A random sample of 400 titles (400/2896, 13.8%) was additionally screened by 2 independent researchers. Furthermore, the entire abstract screening was conducted by 2 independent researchers. In case of ambiguities regarding eligibility, discrepancies were discussed until a consensus was reached. Finally, full texts were screened against the eligibility criteria.

Data Extraction and Analysis

Data extraction was conducted to categorize the included papers for evidence synthesis. Data from the included sources were systematically extracted and organized into a predeveloped Excel spreadsheet. The data coding sheet was created within the research team and refined iteratively. The following data were extracted: bibliographic information (eg, author, year of publication, title, DOI), characteristics of the evaluated DHIs (eg, type of DHI classified according to the WHO and NICE, addressed medical issue, and intended use setting), evaluation methods (eg, study design, use of standardized approaches,

XSL•FO
RenderX

amount of different methods, methods such as surveys or questionnaires, interviews, task or scenario completion, thinking aloud, system data analysis, free-testing phase with the duration, focus groups, and others), evaluation criteria (eg, aspect of assessment, explanation or definition of the evaluated criteria or indicators, evaluated indicators such as clinical outcomes, user behavior change, user experience, technical performance, content performance, actual system usage, suggestions for improvement, and others), and type of end-user involvement (eg, passive data object, active data object, and qualitative data subject). As part of this phase, classification schemes for the DHIs and the evaluated indicators were formulated using inductive category formation, whereby the extracted data served as the basis. This aimed to cluster all extracted DHIs and indicators within the context of this scoping review. The key terms used are outlined in Tables S1-S3 in Multimedia Appendix 3, with consideration of the wide array of DHIs as well as evaluated indicators along with the lack of a universally used classification framework.

Results

Overview

A total of 9618 articles were identified in April 2023 from the search strategy across the 3 scientific databases used. Through initial selection from the databases and the removal of duplicates, 2896 records remained for title screening. After this phase, 560 records remained for abstract screening. A total of 239 records were selected for full-text screening. During this stage, 79 records were excluded. Finally, 160 full-text articles met the eligibility criteria, resulting in 200 evaluations being derived and analyzed. This is explained by the observation that several papers (38/160, 23.8%) included multiple rounds of evaluation, with each employing different methods. For a brief overview of the included papers, 160 articles were relevant, whereas for analysis purposes, the number of evaluations resulting from the included papers was relevant. The data extraction chart is presented in Multimedia Appendix 4. The screening process is shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. DHI: digital health intervention.

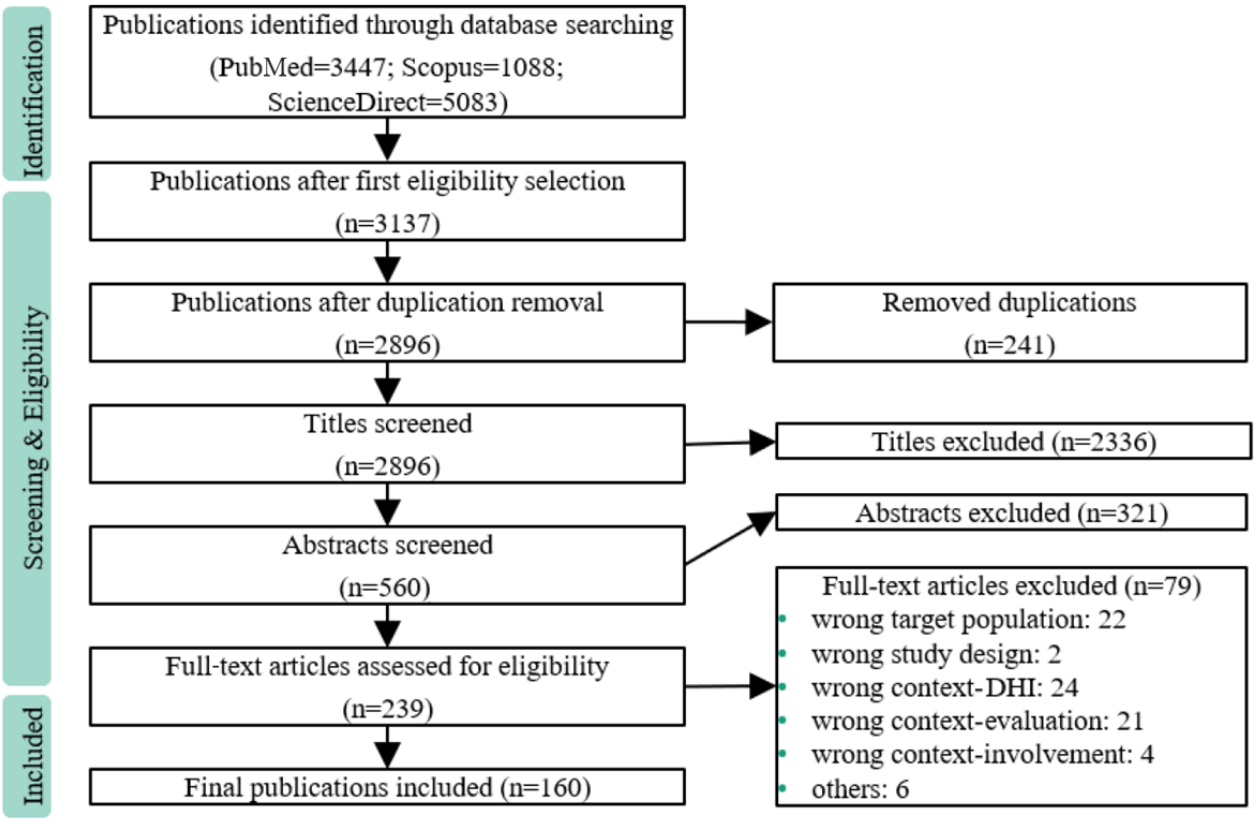

Characteristics of the Included Studies and DHIs Examined in These Studies

The included studies were published between 2010 and 2023, and the number of published studies increased over the years. Specifically, 59 studies (59/160, 36.9%) were published before 2020, and 101 studies (101/160, 63.1%) were published from 2020 to April 2023.

In order to categorize the DHIs investigated in the articles, it was decided to use the WHO and NICE classification schemes [5,10]. There were many different kinds of DHIs, ranging from simple websites for information purposes to complex multimodal interventions. However, these established classification schemes reached their limits, as DHIs often serve multiple functions. As a result, more than a quarter of the DHIs (44/160, 27.5%) could not be clearly categorized within either scheme. Therefore, a summarizing classification scheme to be used in the context of the scoping review was formulated by using inductive category formation whereby the extracted data served as the basis (Table S1 in Multimedia Appendix 3). According to this scheme, most DHIs (42/160, 23.8%) supported the self-management of health

XSL•FO
RenderX

and care, followed by DHIs that were used as a digital supportive component of a treatment (31/160, 19.4%).

In 66 articles (66/160, 41.2%), the primary focus was on the evaluation process. In 94 articles (94/160, 58.8%), the evaluation process was addressed secondarily, as part of a comprehensive description of the DHI development. Table 2 summarizes the characteristics of the included studies and the DHIs examined in these.

Table 2. Characteristics of the included articles.

Characteristic	Value (N=160), n (%)
Year of publication	
2010-2011	2 (1.3)
2012-2013	4 (2.5)
2014-2015	5 (3.1)
2016-2017	17 (10.6)
2018-2019	31 (19.4)
2020-2021	52 (32.5)
2022-2023	49 (30.6)
Focus of the article	
Primary focus on evaluation	66 (41.2)
Secondary focus on evaluation	94 (58.8)
DHI[a] categorization in established schemes	
Unclear or multiple	44 (27.5)
Categorized	116 (72.5)
DHI categorization in newly developed schemes	
Interaction with care provider: data transfer	2 (1.3)
Interaction with care provider: communication	5 (3.1)
Monitoring	22 (13.7)
Tailored information	22 (13.7)
Nontailored information	24 (15.0)
Digital supportive treatment component	31 (19.4)
Self-management	54 (33.8)
Addressed medical issues of DHIs	
Prevention or promotion	6 (3.8)
Generic	20 (12.5)
Mental	33 (20.6)
Somatic	101 (63.1)
Intended use setting of DHIs	
Nursing or retirement home setting	1 (0.6)
Rehabilitation setting	7 (4.4)
Clinical or stationary setting	8 (5.0)
Ambulant or primary care setting	40 (25.0)
Prevention or promotion	42 (26.2)
General health care setting	62 (38.8)

[a]DHI: digital health intervention.

XSL•FO
RenderX

Overview: Evaluation Methods

To answer RQ1 ("Which research methods are used to evaluate DHIs for individuals or public end users?"), the extracted data were analyzed to provide an overview of the used methods.

Analysis of the 200 evaluations revealed that a mixed-methods study design was most commonly used (121/200, 60.5%), followed by qualitative (42/200, 21.0%) and quantitative (37/200, 18.5%) study designs. Mostly, a combination of 2 (60/200, 30.0%) or 3 (62/200, 31.0%) different research methods was applied. Surveys were the most commonly used methods (150/200, 75.0%), followed by interviews (109/200, 54.5%) and testing phases (78/200, 39.0%). The duration of the testing phases predominantly exceeded 1 month (31/200, 15.5%), followed by durations between 1 week and 1 month (21/200, 10.5%) and those shorter than 1 day (13/200, 6.5%). Figure 2 provides a detailed overview of the specific research methods used.

Figure 2. Research methods.

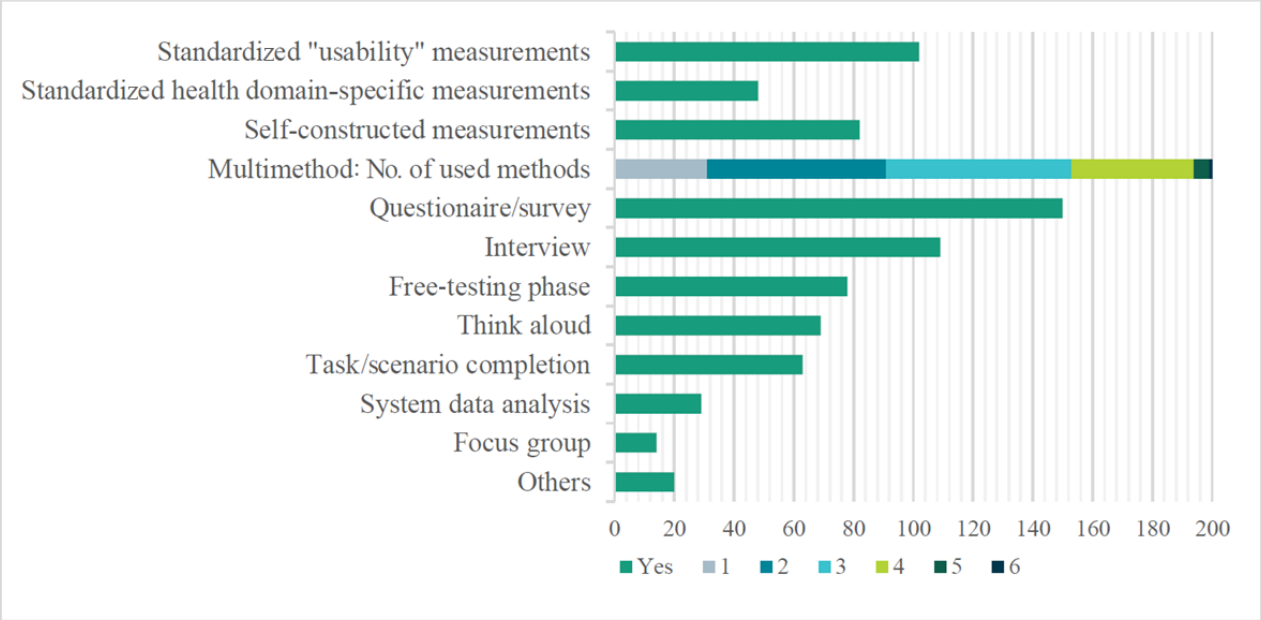

As depicted in Figure 2, in over half of the evaluations (102/200, 51.0%), standardized usability scales, measurements, or questionnaires were used. Self-constructed measurements were used in 41.0% (82/200) of evaluations, while standardized health domain–specific measurements were used in 24.0% (48/200) of evaluations.

To examine the commonly used standardized "usability" frameworks, Figure 3 visualizes most of the used measurements.

Figure 3. Overview of standardized usability measurements. For measurements that appeared in fewer than two evaluations, only health-specific ones have been listed.

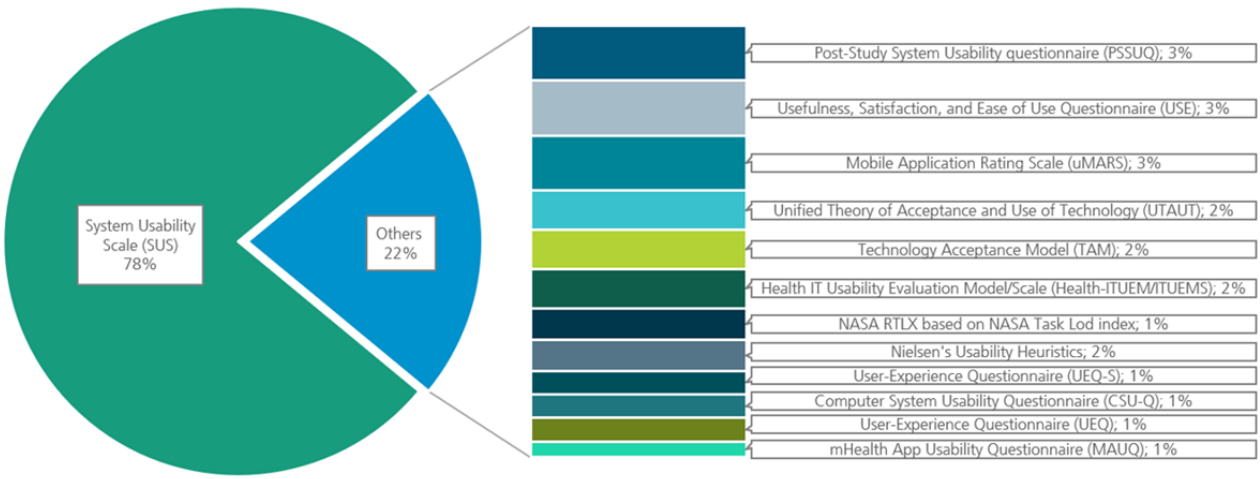

Overview: Evaluation Criteria

To answer RQ2 ("Which evaluation criteria can be identified? Which evaluation criteria have been investigated?"), extracted data were analyzed to create a holistic picture of the used evaluation criteria.

Two-thirds (132/200, 66.0%) of the generally outlined aspects of assessments were user-oriented, followed by multiple aspects

XSL•FO
RenderX

(51/200, 25.5%), clinical outcomes (13/200, 6.5%), and technical aspects (4/200, 2.0%). In order to categorize the evaluation criteria, it was intended to map the extracted evaluations into the established classification scheme of the WHO, which involves feasibility, usability, efficacy, effectiveness, and implementation research [14]. Similarly, when attempting to map evaluation criteria into the WHO classification scheme, limitations were encountered as there were multiple criteria mentioned by the authors or a classification was unclear because the terms were used differently. Moreover, the underlying descriptions or definitions of the terms were based on a variety of approaches and sources, and the used terms also varied. For example, in some evaluations, usability was defined and therefore measured by underlying indicators like effectiveness, efficiency, and satisfaction [46-49] or others like acceptance and feasibility [50,51]. In other cases, some of the previously listed underlying indicators were indicators of other criteria like feasibility [52-55] or were seen as independent criteria, for example, by the WHO or others [14,56-60]. Consequently, more than half of the evaluations (105/200, 52.5%) could not be clearly mapped to one of the WHO-described criteria, as shown in Figure 4.

Figure 4. Research criteria.

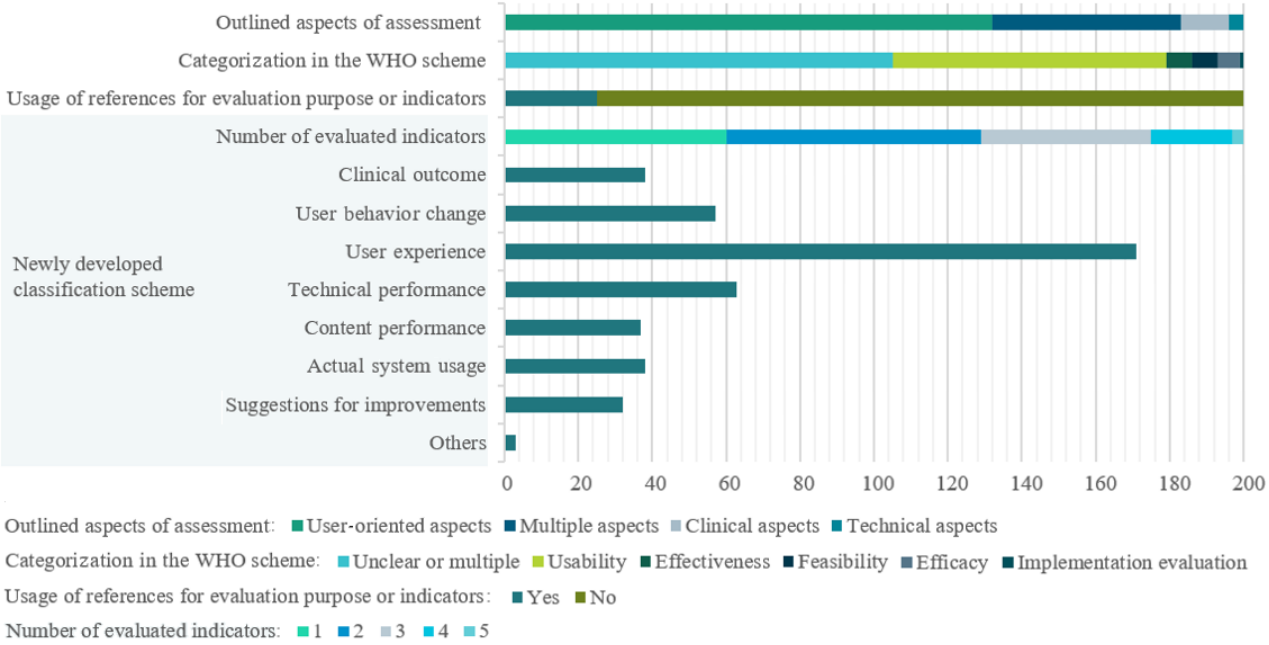

To further explore term usage, the sources or references referred to by the authors of the included articles were examined. It was apparent that in 87.5% (175/200) of evaluations, no information or literature references were provided to guide the terms used. Out of the 25 evaluations (25/200, 12.5%) that referred to literature to describe the used terms, ISO 9241-11 was mostly referred (12/200, 6.0%), followed by multiple authors (9/200, 4.5%), Nielsen (3/200, 1.5%), and other references used once like the Hix & Hartson or "Fit between Individuals, Task, and Technology" (FITT) framework.

As an alternative to the WHO classification scheme, a new scheme was formulated for use in this scoping review. Inductive category formation was used to create the new scheme, whereby all indicators extracted from the evaluations served as the basis of the data. This process led to the consolidation of 8 criteria dimensions, which are described in Table S2 in Multimedia Appendix 3, along with their associated indicators. Figure 4 visualizes the described findings and the classification according to the newly developed scheme.

Overview: End-User Involvement

To answer RQ3 ("In which way are individuals or public end users involved in the evaluation process?"), extracted data were analyzed to investigate end-user involvement in the evaluation of DHIs that are primarily used by them.

To investigate end-user involvement, the ways in which they were involved were categorized as follows: (1) passive data objects, where citizens or patients do not have an active role, such as in responding to surveys, and evidence is gathered, for example, by analyzing system data; (2) active data objects, characterized by consciously responding to surveys with predefined options or participating in task or scenario-based sessions; and (3) qualitative data objects, where patients or citizens provide individual responses, allowing them to express and explain their views and emotions. These groups are further described in Table S3 in Multimedia Appendix 3. The majority of evaluations actively involved public end users. In 83% (166/200) of evaluations, end users actively assessed the DHIs using predefined answer options, and in 73% (146/200), public end users and patients had the opportunity to individually express their views and emotions regarding the DHIs. The majority of evaluations referenced user-centered approaches (145/200, 72.5%), with UCD being the most commonly mentioned (133/200, 66.5%). Additionally, related approaches, such as participatory design, human-centered design, and co-design or co-creation, were mentioned alternatively or in

XSL•FO
RenderX

combination with UCD. In 25.5% (55/200) of evaluations, none of these user-centered approaches were mentioned.

Discussion

Principal Findings

This scoping review provides a holistic overview of the way DHIs for public end users are evaluated. A total of 160 studies resulting in 200 evaluations were included in this review. It focused on assessing the range of scientific literature concerning various methods, indicators, and types of public involvement. Research in the field of DHI evaluation seems to be in its early stages. Although studies from the last 15 years were eligible, the oldest eligible study was from 2010, and until 2016, the number of studies was in the single-digit range. The majority of eligible publications were from 2020 onwards (101/160, 63.13%), indicating the growing relevance of evaluating DHIs for public end users. Similar findings are reported in other literature [3,4,6,28].

It has been shown that there is a broad range of different DHIs with complex functions, and thus, some cannot be clearly categorized in established schemes like those of the WHO or NICE. Furthermore, there is a lack of methodological consistency, and references to existing frameworks, such as the Monitoring and Evaluation Guideline of the WHO, the Evidence Standards framework for digital health technologies of the NICE, the report of the EXPH, the Swiss Evaluation framework by Kowatsch et al, and the approach of Murray et al [1,5,14,17,24], are notably absent. The bare use of standardized frameworks was also concluded by another review [20]. Instead of these, standardized usability measurements were used in more than half of the evaluations, although most of these measurement instruments are not specially developed for the health care context (eg, System Usability Scale [SUS]). A major use of standardized measurement methods also resulted from a review of a specific medical field [22]. In other cases, standard measurements were considered insufficient or found to be too complex [60], and thus, self-constructed surveys are often developed. On the one hand, this raises questions about differences regarding measurement quality, because some self-constructed measurements appear to not be validated. On the other hand, the use of different measurement methods (some validated and some not validated) complicates the comparability of the results. Similar results were found in the literature or in former reviews [28]. In addition, an important question is why current frameworks that are developed by credible institutions or scientists and explicitly focus on the health care sector are not being considered.

In terms of the evaluation criteria, similarities to previously described observations can be seen. As with methodology, no references to existing frameworks, such as the Monitoring and Evaluation Guideline of the WHO, the Evidence Standards framework for digital health technologies of the NICE, the report of the EXPH, the Swiss Evaluation framework by Kowatsch et al, and the approach of Murray et al [1,5,14,17,24], were mentioned. In the few evaluations that referred to existing literature to explain the used terms, ISO 9241-11 was mostly mentioned, although this has not been specially developed for the health care context. Regarding the mapping of the extracted large amount of evaluation criteria into the established classification scheme of the WHO, a similar circumstance to DHI mapping is noticeable. A clear categorization in established schemes is not possible because of a missing commonly agreed definition and therefore differing usage of the evaluated indicators. The lack of a uniform definition and the variability in term usage raise concerns about whether this impedes the sharing of evidence among eHealth interventions due to insufficient comparability of the results. The same thoughts have been reported in the literature [7,20].

The literature calls, for example, for UCD processes to increase the involvement of different end-user groups [3]. The results showed that the majority of evaluations referred to user-centered approaches and actively involved public end users in the evaluation process.

In summary, the analysis revealed that there is neither a consensus on the methods for evaluating DHIs nor a commonly agreed definition or usage of the evaluated indicators, resulting in a broad variety of evaluation practices. This aligns with the findings of the existing literature [23,61]. Although several frameworks exist, the problem of heterogenicity and variability according to the evaluation of DHIs remains, and these circumstances appear to hamper gathering of reliable evidence through evaluation.

To address these challenges, practical suggestions and implications for further research have been identified and will be described subsequently.

Implications for Practice and Further Research

To consolidate the previously mentioned results on evaluation methods and criteria for each DHI, according to the newly established categorization scheme, Figure 5 presents a results matrix. Considering the observation that established frameworks seem to not be referred, this figure is not intended to function as a framework. Its purpose is to assist developers, evaluators, researchers, and others in this field with the decision-making process by providing an overview of how DHIs have been evaluated by other people. This matrix can address the methodological challenge by providing initial practical guidance or decision-support for those developing and evaluating DHIs for public end users, offering a foundation for considering appropriate ways to evaluate their own DHIs.

XSL•FO
RenderX

Figure 5. Results matrix method and criteria per digital health intervention. Others include cognitive walkthrough, observation, informal conversation, design workshop, and eye tracking.

	Survey	Interview	Free-Testing Phase	Think Aloud	Task/ Scenario Com-pletion	System Data Analysis	Focus Group	Others
Clinical Outcome	2.5% (5) 5.5% (11) 0.5% (1) 4% (8) 4.5% (9) 0.5% (1)	2.5% (5) 2.5% (5) 0.5% (1) 1.5% (3) 1.5% (3) 0.5% (1) 0.5% (1)	2% (4) 4.5% (9) 0.5% (3) 1.5% (3) 3% (6) 0.5% (1)	0.5% (1) 0.5% (1) 0.5% (1)	0.5% (1) 0.5% (1) 0.5% (1)	0.5% (1) 2.5% (5) 1.5% (3)		
User Behavior Change	5% (10) 5% (10) 2.5% (5) 5.5% (11) 3.5% (7) 0.5% (1)	5.5% (11) 2% (4) 2% (4) 2% (4) 3.5% (7) 0.5% (1)	4% (8) 0.5% (1) 1% (2) 3% (6) 3.5% (7) 0.5% (1)	2.5% (3) 1.5% (3) 1% (2) 1.5% (3) 0.5% (1)	2.5% (5) 2% (4) 1% (2) 0.5% (1) 0.5% (1)	2% (4) 2% (4) 2% (4)	0.5% (1) 1.5% (3) 0.5% (1)	1.5% (3) 0.5% (1) 1% (2)
User Experience	21.5% (43) 13% (26) 9% (18) 9% (18) 8.5% (17) 1.5% (3) 1% (2)	19.5% (39) 7.5% (15) 8.5% (17) 6% (12) 8% (16) 1% (2) 0.5% (1)	9% (18) 6% (12) 5% (10) 5.5% (11) 5% (10)	13% (26) 5.5% (11) 4% (8) 4% (8) 3.5% (7) 0.5% (1) 1% (1)	13% (26) 5% (10) 3.5% (7) 2% (4) 4% (8) 0.5% (1) 0.5% (1)	3.5% (7) 1.5% (3) 2.5% (59 3.5% (7) 2.5% (5)	2% (4) 2% (4) 2% (4) 0.5% (1)	3.5% (7) 1.5% (3) 1% (2) 1% (2) 1% (2) 0.5% (1) 0.5% (1)
Technical Performance	5% (10) 3% (6) 4.5% (9) 4.5% (9) 2% (4) 0.5% (1)	7.5% (15) 1% (2) 2.5% (5) 3.5% (7) 3.5% (7) 0.5% (1) 0.5% (1)	2.5% (5) 1% (2) 2% (4) 3% (6) 2% (4)	5% (10) 2% (4) 2% (4) 3.5% (7) 1% (2) 1% (2)	5% (10) 1.5% (3) 1.5% (3) 1% (2) 0.5% (1) 1.5% (3)	1% (2) 0.5% (1) 1% (2) 1.5% (3)	1.5% (3) 1% (2) 0.5% (1) 0.5% (1)	1% (2) 1% (2) 0.5% (1) 1% (2) 0.5% (1) 0.5% (1)
Content Performance	4% (8) 3.5% (7) 3.5% (7) 2.5% (5)	5% (10) 1% (2) 2% (4) 1.5% (3) 0.5% (1)	1% (2) 0.5% (1) 2% (4) 2% (4) 0.5% (1)	2% (4) 2% (4) 0.5% (3) 3% (6) 0.5% (1)	1.5% (3) 1.5% (3) 1.5% (3) 0.5% (1)	0.5% (1)	0.5% (1) 0.5% (1)	1.5% (3) 0.5% (1) 1% (2)
Actual System Usage	2.5% (5) 2% (4) 3% (6) 5.5% (11) 2.5% (5)	2.5% (5) 1% (2) 2% (4) 3.5% (7) 2.5% (5)	2.5% (5) 1.5% (3) 2.5% (5) 4% (8) 2.5% (5)	2% (4) 0.5% (1) 1% (2) 0.5% (1)	2.5% (5) 0.5% (1) 0.5% (1)	3% (6) 1.5% (3) 2.5% (5) 4% (8) 3% (6)	0.5% (1) 0.5% (1) 0.5% (1)	0.5% (1)
Suggestions for Improvement	2.5 (5) 2% (4) 3% (6) 5.5% (11) 2.5% (5)	2.5% (5) 1.5% (3) 2.5% (5) 0.5% (1) 2% (4)	1.5% (3) 1% (2) 2% (4) 1% (2)	2.5% (5) 1% (2) 1.5% (3) 0.5% (1) 1% (2)	3% (6) 1% (2) 0.5% (1) 1% (2)	0.5% (1) 0.5% (1) 0.5% (1) 0.5% (1)	0.5% (1) 0.5% (1) 0.5% (1)	0.5% (1) 0.5% (1)
Others	0.5% (1)			0.5% (1) 0.5% (1)	0.5% (1) 1% (2)		0.5% (1)	0.5% (1) 0.5% (1)

Self-management
Tailored information
Data transfer
Digital supportive treatment component
Monitoring
Nontailored information
Communication

As long as there is no commonly agreed and broadly applicable approach for evaluating DHIs, certain aspects derived from this review should be considered and described in evaluation studies to enhance the quality and comparability of the results, thereby creating evidence. Following the approach of Murray et al [1], a question-driven approach has been chosen for outlining these aspects.

What Type of DHI is at the Center of the Evaluation? What are its Main Characteristics and Intended Functions?

Describing the characteristics and key components of a DHI is crucial for comparability, and it forms the basis for selecting an appropriate methodology for the study. This description includes determining its primary functions, its user interface, the health domain it addresses, the specific medical issue it

XSL•FO
RenderX

targets, the population it is designed for, and the intended use setting. This aligns with suggestions from existing frameworks [1,5,17]. Moreover, evaluators should reference established classification systems, such as those provided by the WHO or NICE [5,10]. This approach can help in creating a shared understanding and definition of the DHI context, as recommended in the literature [9].

What are the Reasons Behind Choosing the Research Method? What Roles do Established Measurement Methods Play?

While the description of the chosen methods is standard in research, the underlying reasons or decision-making processes for their selection are often not detailed. Therefore, it is essential that evaluators not only carefully select methods suitable for DHI evaluation but also provide clear justifications for these choices [17]. This transparency in decision-making enhances the reliability and validity of the research. Established measurement methods should be given due consideration in the evaluation process, and when these are applicable and relevant, they should be integrated into the study. Literature shows different reviews exploring potential frameworks that can be applied to DHIs [18,24]. If established methods are deemed inappropriate or inapplicable for the specific context of the DHI being evaluated, it is important to briefly explain the reasons for the exclusion. This approach not only adheres to high research standards but also helps in understanding the unique aspects of the DHI that necessitate a deviation from conventional methods.

Which Criteria are Being Evaluated? Which Established Sources or References are Considered Regarding the Evaluated Criteria?

The evaluated criteria and underlying measured indicators should be briefly described by considering established references. Clarity in defining these criteria is essential for creating a standardized and shared understanding of the evaluation objectives. The selection of the criteria should be informed by established references and frameworks in the field of digital health. A possible source to refer to for justification in the context of digitalization and health is the WHO, as it lists and defines criteria, such as feasibility, usability, efficacy, and effectiveness [14]. Another valuable reference is the Swiss Evaluation framework by Kowatsch et al [24], which outlines various criteria specific to DHIs. Referring to established references helps in grounding the evaluation with recognized standards and contributes to the creation of a shared language and understanding in the field, as suggested in the literature [9]. This approach can facilitate comparability across studies and contribute to the broader discourse on DHI efficacy and impact.

In Which Way are End Users Involved in the Evaluation Process?

In accordance with the NICE framework, it is important to detail how representatives from intended end-user groups are incorporated into the evaluation [5]. This should include specifying their roles and the extent of their participation, whereby approaches, such as PHR and PPI, could be referred. Given the potential agile nature of DHI development, these approaches could reach their limits, and referring to the principles, strategies, and methods, UCD can be highly effective in ensuring meaningful end-user involvement.

The following implication for further research can be derived from the scoping review. During the scoping review, the question arose of why current established measurements or frameworks, which are developed by credible institutions or scientists and explicitly focused on the health care sector, are not being referenced. This potential research-practice gap could be addressed by further research, for example, by investigating the discrepancies between scientific best practices in user-centered evaluation strategies and the suggestions from established frameworks within the processes of real-world user-centered evaluation. In addition, DHIs are developed in not only academic settings but also commercial sectors. As stated in the limitations, the results of evaluation studies are not always published, especially those from the commercial sector. Therefore, it could be interesting to investigate evaluation methods, particularly in this field. Similar thoughts are reported in the literature [28]. Regarding the involvement of public end users, it would be interesting to investigate from a public end-user perspective which methods are suitable for providing close-to-reality feedback. This involves questions, such as “How do they like to explore DHIs?” “Are the measurement methods currently used by evaluators suitable for conveying the benefits to public end users in a realistic way, thus enabling them to assess the value?” and “From their point of view, which approaches are tangible and can be transferred to the reality of care?” Furthermore, it could be interesting to develop and investigate a reporting guideline that ensures an evidence-based minimum set of items for reporting evaluation processes on DHIs. Considering the incorporation of potential end users, further research could focus on reporting guidelines consisting of standardized components regarding the participation or involvement of end users in the development or evaluation of DHIs.

Limitations

This scoping review was conducted based on the methodological framework for scoping reviews by Arksey & O’Malley and complies with the PRISMA-ScR guidelines. To the best of our knowledge, the results provide the first holistic overview of scientific research on evaluation methods and indicators in the context of DHIs for public end users, as the scope is neither specific to medicine nor specific to criteria.

However, the review has some limitations. One limitation is associated with the way evaluation studies are published. Generally, evaluations of DHIs are published in one of the following two ways: different evaluations conducted during the development process of a DHI are published in separate papers or the entire development process is described in a single paper. In the latter case, there are often no detailed descriptions of the evaluation process, as it is not the main focus of the publication. This lack of detail impedes a comprehensive and thorough overview of the evaluation methods and indicators of DHIs throughout their development and actual use. Another limitation is that the results of evaluation studies, especially from the commercial sector, are not always published [20,23,28], which

XSL•FO
RenderX

can hamper a comprehensive overview. Additionally, in line with the nature of scoping reviews, there is neither a quality assessment of the included studies and evaluations nor a quality assessment of the extracted methods, criteria, and modes of public involvement. These aspects are presented as reported by the authors of the primary studies.

Conclusions

This scoping review provides a comprehensive overview of the current methods used in evaluating DHIs for public end users. The analysis revealed that there is neither a consensus on methods for evaluating DHIs nor a commonly agreed definition or usage of the evaluated indicators, resulting in a broad variety of evaluation practices. Although several frameworks exist, the problem of heterogenicity and variability according to the evaluation of DHIs remains, and these circumstances seem to hamper gathering reliable evidence through evaluation. Recommendations are derived from the findings in order to enhance the quality and comparability of the results of evaluation studies. It is important to note that the results are not intended to serve as a framework or as best-practice recommendations. Investigating these aspects could form a part of future research endeavors. We demonstrated that the research field is complex, heterogeneous, and broad, and our findings provide the first overview and have identified research gaps that could be addressed further. In conclusion, uniform use of terms, particularly regarding evaluation criteria and DHI classification, within the digital health sector could facilitate the transferability of results among similar evaluation studies. This standardization could significantly contribute to the cohesiveness and effectiveness of research in this evolving field.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 39 KB - jmir_v26i1e55714_app1.docx]

Multimedia Appendix 2
Search strategy for each database.
[DOCX File , 25 KB - jmir_v26i1e55714_app2.docx]

Multimedia Appendix 3
Key terms.
[DOCX File , 30 KB - jmir_v26i1e55714_app3.docx]

Multimedia Appendix 4
Data extraction chart.
[XLSX File (Microsoft Excel File), 259 KB - jmir_v26i1e55714_app4.xlsx]

References

1. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: Key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
2. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2023-06-29]
3. Fischer F. [Digital interventions in prevention and health promotion: What kind of evidence do we have and what is needed?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Jun;63(6):674-680. [doi: 10.1007/s00103-020-03143-6] [Medline: 32355991]
4. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: Methodological issues and recommendations for practice. Pharmacoeconomics 2022 Apr 08;40(4):367-378 [FREE Full text] [doi: 10.1007/s40273-022-01130-0] [Medline: 35132606]
5. Evidence standards framework (ESF) for digital health technologies. National Institute for Health and Care Excellence. URL: https://tinyurl.com/3dfvuv85 [accessed 2024-05-16]
6. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon M. Methods to evaluate the effects of internet-based digital health interventions for citizens: Systematic review of reviews. J Med Internet Res 2018 Jun 07;20(6):e10202 [FREE Full text] [doi: 10.2196/10202] [Medline: 29880470]
7. Kolasa K, Kozinski G. How to value digital health interventions? A systematic literature review. Int J Environ Res Public Health 2020 Mar 23;17(6):17062119 [FREE Full text] [doi: 10.3390/ijerph17062119] [Medline: 32209988]

XSL•FO
RenderX

8. Kernebeck S, Scheibe M, Sinha M, Fischer F, Knapp A, Timpel P, et al. [Development, evaluation and implementation of digital health interventions (Part 1) - Discussion paper of the digital health working group of the German Network for Health Services Research (DNVF)]. Gesundheitswesen 2023 Jan;85(1):58-64. [doi: 10.1055/a-1933-2779] [Medline: 36446615]
9. Kwasnicka D, Keller J, Perski O, Potthoff S, Ten Hoor GA, Ainsworth B, et al. White paper: Open digital health - accelerating transparent and scalable health promotion and treatment. Health Psychol Rev 2022 Dec;16(4):475-491 [FREE Full text] [doi: 10.1080/17437199.2022.2046482] [Medline: 35240931]
10. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. URL: https://apps.who.int/iris/handle/10665/260480 [accessed 2024-05-16]
11. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022 Jun 28;24(6):e31921 [FREE Full text] [doi: 10.2196/31921] [Medline: 35763320]
12. Lange O. Health economic evaluation of preventive digital public health interventions using decision-analytic modelling: a systematized review. BMC Health Serv Res 2023 Mar 17;23(1):268 [FREE Full text] [doi: 10.1186/s12913-023-09280-3] [Medline: 36932436]
13. Maaß L, Pan C, Freye M. Mapping digital public health interventions among existing digital technologies and internet-based interventions to maintain and improve population health in practice: Protocol for a scoping review. JMIR Res Protoc 2022 Mar 31;11(3):e33404 [FREE Full text] [doi: 10.2196/33404] [Medline: 35357321]
14. Monitoring and evaluating digital health interventions: A practical guide to conducting research and assessment. World Health Organization. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-05-16]
15. Hochmuth A, Exner A, Dockweiler C. [Implementation and participatory design of digital health interventions]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):145-152. [doi: 10.1007/s00103-019-03079-6] [Medline: 31938837]
16. Hrynyschyn R, Prediger C, Stock C, Helmer SM. Evaluation methods applied to digital health interventions: What is being used beyond randomised controlled trials?-A scoping review. Int J Environ Res Public Health 2022 Apr 25;19(9):19095221 [FREE Full text] [doi: 10.3390/ijerph19095221] [Medline: 35564616]
17. Opinion on assessing the impact of digital transformation of health services. European Commission, Directorate-General for Health and Food Safety. 2019. URL: https://data.europa.eu/doi/10.2875/644722 [accessed 2024-05-16]
18. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-A call for innovative evidence generation approaches. NPJ Digit Med 2020;3:110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
19. von Huben A, Howell M, Howard K, Carrello J, Norris S. Health technology assessment for digital technologies that manage chronic disease: a systematic review. Int J Technol Assess Health Care 2021 May 26;37(1):e66. [doi: 10.1017/S0266462321000362] [Medline: 34034851]
20. Enam A, Torres-Bonilla J, Eriksson H. Evidence-based evaluation of eHealth interventions: Systematic literature review. J Med Internet Res 2018 Nov 23;20(11):e10971 [FREE Full text] [doi: 10.2196/10971] [Medline: 30470678]
21. Jahnel T, Schüz B. [Participatory development of digital public health: tension between user perspectives and evidence]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):153-159. [doi: 10.1007/s00103-019-03082-x] [Medline: 31915866]
22. Holl F, Kircher J, Swoboda WJ, Schobel J. Methods used to evaluate mHealth applications for cardiovascular disease: A quasi-systematic scoping review. Int J Environ Res Public Health 2021 Nov 23;18(23):182312315 [FREE Full text] [doi: 10.3390/ijerph182312315] [Medline: 34886039]
23. Kernebeck S, Redaelli M, Kramer U, Vollmar H. Wie sollen digitale Gesundheitsanwendungen durch die Versorgungsforschung evaluiert werden? German Medical Science. 2019. URL: https://www.egms.de/static/en/meetings/dkvf2019/19dkvf029.shtml [accessed 2024-05-16]
24. Kowatsch T, Harperink S, Cotti A, Trémeaud R, Schneider J, Della Valle M. Evaluierungs-Framework und Kriterienkatalog für digitale Gesundheitsinterventionen in der Schweiz. ETH Library. URL: https://www.research-collection.ethz.ch/handle/20.500.11850/470919?show=full [accessed 2024-05-16]
25. Karpathakis K, Libow G, Potts HWW, Dixon S, Greaves F, Murray E. An evaluation service for digital public health interventions: User-centered design approach. J Med Internet Res 2021 Sep 08;23(9):e28356 [FREE Full text] [doi: 10.2196/28356] [Medline: 34494965]
26. Hastall MR, Dockweiler C, Mühlhaus J. Achieving end user acceptance: building blocks for an evidence-based user-centered framework for health technology development and assessment. In: Antona M, Stephanidis C, editors. Universal Access in Human–Computer Interaction. Human and Technological Environments. UAHCI 2017. Lecture Notes in Computer Science, vol 10279. Achieving End User Acceptance: Building Blocks for an Evidence-Based User-Centered Framework for Health Technology Development and Assessment 2017; 2017:13-25.
27. Dockweiler C, Razum O. [Digital health: New challenges for public health]. Gesundheitswesen 2016 Jan;78(1):5-7. [doi: 10.1055/s-0041-110679] [Medline: 26695540]

XSL•FO
RenderX

28. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: A scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
29. ISO 9241-210:2019 Ergonomics of human-system interactionPart 210: Human-centred design for interactive systems. ISO. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2024-05-16]
30. Cornet VP, Toscos T, Bolchini D, Rohani Ghahari R, Ahmed R, Daley C, et al. Untold stories in user-centered design of mobile health: Practical challenges and strategies learned from the design and evaluation of an app for older adults with heart failure. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17703 [FREE Full text] [doi: 10.2196/17703] [Medline: 32706745]
31. Dopp AR, Parisi KE, Munson SA, Lyon AR. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. Implement Sci Commun 2020;1:17 [FREE Full text] [doi: 10.1186/s43058-020-00020-w] [Medline: 32885179]
32. Position Paper 1: What is Participatory Health Research? International Collaboration for Participatory Health Research. 2013 May. URL: http://www.icphr.org/position-papers--discussion-papers/position-paper-no-1 [accessed 2024-05-16]
33. INVOLVE. National Institute for Health Research. 2017 Nov. URL: https://www.invo.org.uk/wp-content/uploads/2018/01/Induction-public-v3.pdf [accessed 2024-05-16]
34. What is Patient and Public Involvement and Public Engagement? National Institute for Health Research (NIHR) - School for Primary Care Research. URL: https://www.spcr.nihr.ac.uk/PPI/what-is-patient-and-public-involvement-and-engagement [accessed 2024-05-16]
35. Patient and public involvement, engagement and participation definitions. University of Oxford - Medical Sciences Division. URL: https://www.medsci.ox.ac.uk/research/patient-and-public-involvement/section-2-what-is-patient-and-public-involvement [accessed 2024-05-16]
36. Going the extra mile: Improving the nation's health and wellbeing through public involvement in research. National Institute for Health Research (NIHR). URL: https://tinyurl.com/prsbk2u4 [accessed 2024-05-16]
37. Jonsson M, Johansson S, Hussain D, Gulliksen J, Gustavsson C. Development and evaluation of eHealth services regarding accessibility: Scoping literature review. J Med Internet Res 2023 Aug 17;25:e45118 [FREE Full text] [doi: 10.2196/45118] [Medline: 37590050]
38. Scheckel B, Schmidt K, Stock S, Redaèlli M. Patient portals as facilitators of engagement in patients with diabetes and chronic heart disease: scoping review of usage and usability. J Med Internet Res 2023 Aug 25;25:e38447 [FREE Full text] [doi: 10.2196/38447] [Medline: 37624629]
39. Sun Y, Zhou J, Ji M, Pei L, Wang Z. Development and evaluation of health recommender systems: systematic scoping review and evidence mapping. J Med Internet Res 2023 Jan 19;25:e38184 [FREE Full text] [doi: 10.2196/38184] [Medline: 36656630]
40. von Elm E, Schreiber G, Haupt CC. Methodische Anleitung für Scoping Reviews (JBI-Methodologie). Z Evid Fortbild Qual Gesundhwes 2019 Jun;143:1-7. [doi: 10.1016/j.zefq.2019.05.004] [Medline: 31296451]
41. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
42. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
43. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
44. Peters M, Godfrey C, McInerney P, Munn Z, Tricco A, Khalil H. Chapter 11: Scoping Reviews (2020 version). In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Philadelphia, PA: Lippincott Williams & Wilkins; 2020.
45. Weirauch V. User-Centered Evaluation of Digital Health Intervention: A Scoping Review. Center for Open Science. 2023. URL: https://osf.io/x6gkf [accessed 2024-05-16]
46. Georgsson M, Staggers N. Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics. J Am Med Inform Assoc 2016 Jan;23(1):5-11 [FREE Full text] [doi: 10.1093/jamia/ocv099] [Medline: 26377990]
47. Marien S, Legrand D, Ramdoyal R, Nsenga J, Ospina G, Ramon V, et al. A web application to involve patients in the medication reconciliation process: a user-centered usability and usefulness study. J Am Med Inform Assoc 2018 Nov 01;25(11):1488-1500 [FREE Full text] [doi: 10.1093/jamia/ocy107] [Medline: 30137331]
48. Rochat J, Ehrler F, Siebert JN, Ricci A, Garretas Ruiz V, Lovis C. Usability testing of a patient-centered mobile health app for supporting and guiding the pediatric emergency department patient journey: mixed methods study. JMIR Pediatr Parent 2022 Mar 15;5(1):e25540 [FREE Full text] [doi: 10.2196/25540] [Medline: 35289754]
49. Sousa P, Fonseca H, Gaspar P, Gaspar F. Usability of an internet-based platform (Next.Step) for adolescent weight management. J Pediatr (Rio J) 2015;91(1):68-74 [FREE Full text] [doi: 10.1016/j.jped.2014.05.010] [Medline: 25245364]
50. Mendes-Santos C, Nunes F, Weiderpass E, Santana R, Andersson G. Development and evaluation of the usefulness, usability, and feasibility of INNOV breast cancer: Mixed methods study. JMIR Cancer 2022 Feb 15;8(1):e33550 [FREE Full text] [doi: 10.2196/33550] [Medline: 35166682]

XSL•FO
RenderX

51. Johnson AK, Haider S, Nikolajuk K, Kuhns LM, Ott E, Motley D, et al. An mHealth intervention to improve pre-exposure prophylaxis knowledge among young black women in family planning clinics: Development and usability study. JMIR Form Res 2022 Jul 28;6(7):e37738 [FREE Full text] [doi: 10.2196/37738] [Medline: 35900830]
52. Ebenfeld L, Kleine Stegemann S, Lehr D, Ebert DD, Funk B, Riper H, et al. A mobile application for panic disorder and agoraphobia: Insights from a multi-methods feasibility study. Internet Interv 2020 Mar;19:100296 [FREE Full text] [doi: 10.1016/j.invent.2019.100296] [Medline: 31890640]
53. Wiberg A, Ghaderi A, Danielsson HB, Safarzadeh K, Parling T, Carlbring P, et al. Internet-based cognitive behavior therapy for eating disorders - Development and feasibility evaluation. Internet Interv 2022 Dec;30:100570 [FREE Full text] [doi: 10.1016/j.invent.2022.100570] [Medline: 36110307]
54. Bendig J, Wolf A, Mark T, Frank A, Mathiebe J, Scheibe M, et al. Feasibility of a multimodal telemedical intervention for patients with Parkinson's disease: A pilot study. J Clin Med 2022 Feb 18;11(4):1074 [FREE Full text] [doi: 10.3390/jcm11041074] [Medline: 35207351]
55. Ahern DK, Parker D, Eaton C, Rafferty C, Wroblewski J, Goldman R. Patient-facing technology for identification of COPD in primary care. J Innov Health Inform 2016 Jul 15;23(2):824 [FREE Full text] [doi: 10.14236/jhi.v23i2.824] [Medline: 27869583]
56. van Doorn M, Nijhuis LA, Monsanto A, van Amelsvoort T, Popma A, Jaspers MWM, et al. Usability, feasibility, and effect of a biocueing intervention in addition to a moderated digital social therapy-platform in young people with emerging mental health problems: A mixed-method approach. Front Psychiatry 2022;13:871813 [FREE Full text] [doi: 10.3389/fpsyt.2022.871813] [Medline: 35693972]
57. Shrestha R, Altice FL, DiDomizio E, Sibilio B, Ranjit YS, Copenhaver MM. Feasibility and acceptability of an mHealth-based approach as an HIV prevention strategy among people who use drugs on pre-exposure prophylaxis. Patient Prefer Adherence 2020;14:107-118 [FREE Full text] [doi: 10.2147/PPA.S236794] [Medline: 32021122]
58. Peng ML, Wickersham JA, Altice FL, Shrestha R, Azwa I, Zhou X, et al. Formative evaluation of the acceptance of HIV prevention artificial intelligence chatbots by men who have sex with men in Malaysia: Focus group study. JMIR Form Res 2022 Oct 06;6(10):e42055 [FREE Full text] [doi: 10.2196/42055] [Medline: 36201390]
59. Slater H, Stinson JN, Jordan JE, Chua J, Low B, Lalloo C, et al. Evaluation of digital technologies tailored to support young people's self-management of musculoskeletal pain: mixed methods study. J Med Internet Res 2020 Jun 05;22(6):e18315 [FREE Full text] [doi: 10.2196/18315] [Medline: 32442143]
60. Floch J, Vilarinho T, Zettl A, Ibanez-Sanchez G, Calvo-Lerma J, Stav E, et al. Users' experiences of a mobile health self-management approach for the treatment of cystic fibrosis: Mixed methods study. JMIR Mhealth Uhealth 2020 Jul 08;8(7):e15896 [FREE Full text] [doi: 10.2196/15896] [Medline: 32673237]
61. Theodorou P, Tsiligkos K, Meliones A, Filios C. A training smartphone application for the simulation of outdoor blind pedestrian navigation: Usability, UX evaluation, sentiment analysis. Sensors (Basel) 2022 Dec 29;23(1):367 [FREE Full text] [doi: 10.3390/s23010367] [Medline: 36616964]

Abbreviations

DHI: digital health intervention
DPHI: digital public health intervention
EBM: evidence-based medicine
EXPH: Expert Panel on effective ways of investing in health
HTA: health technology assessment
mHealth: mobile health
NICE: National Institute for Health and Care Excellence
NIHR: National Institute for Health Research
PHR: participatory health research
PPI: public and patient involvement
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
RQ: research question
UCD: user-centered design
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Ma; submitted 22.12.23; peer-reviewed by W Evans, K Ali; comments to author 20.03.24; revised version received 28.03.24; accepted 24.04.24; published 31.05.24.

Please cite as:
Weirauch V, Soehnchen C, Burmann A, Meister S
Methods, Indicators, and End-User Involvement in the Evaluation of Digital Health Interventions for the Public: Scoping Review
J Med Internet Res 2024;26:e55714
URL: https://www.jmir.org/2024/1/e55714
doi:10.2196/55714
PMID:38819891

©Vera Weirauch, Clarissa Soehnchen, Anja Burmann, Sven Meister. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Vulnerability to Cyberattacks and Sociotechnical Solutions for Health Care Systems: Systematic Review

Pius Ewoh[1], MBA; Tero Vartiainen[1], PhD

School of Technology and Innovations, Information Systems Science, University of Vaasa, Vaasa, Finland

Corresponding Author:
Pius Ewoh, MBA
School of Technology and Innovations
Information Systems Science
University of Vaasa
Wolffintie 32
Vaasa, 65200
Finland
Phone: 358 414888477
Email: pius.ewoh@uwasa.fi

Abstract

Background: Health care organizations worldwide are faced with an increasing number of cyberattacks and threats to their critical infrastructure. These cyberattacks cause significant data breaches in digital health information systems, which threaten patient safety and privacy.

Objective: From a sociotechnical perspective, this paper explores why digital health care systems are vulnerable to cyberattacks and provides sociotechnical solutions through a systematic literature review (SLR).

Methods: An SLR using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) was conducted by searching 6 databases (PubMed, Web of Science, ScienceDirect, Scopus, Institute of Electrical and Electronics Engineers, and Springer) and a journal (*Management Information Systems Quarterly*) for articles published between 2012 and 2022 and indexed using the following keywords: "(cybersecurity OR cybercrime OR ransomware) AND (healthcare) OR (cybersecurity in healthcare)." Reports, review articles, and industry white papers that focused on cybersecurity and health care challenges and solutions were included. Only articles published in English were selected for the review.

Results: In total, 5 themes were identified: human error, lack of investment, complex network-connected end-point devices, old legacy systems, and technology advancement (digitalization). We also found that knowledge applications for solving vulnerabilities in health care systems between 2012 to 2022 were inconsistent.

Conclusions: This SLR provides a clear understanding of why health care systems are vulnerable to cyberattacks and proposes interventions from a new sociotechnical perspective. These solutions can serve as a guide for health care organizations in their efforts to prevent breaches and address vulnerabilities. To bridge the gap, we recommend that health care organizations, in partnership with educational institutions, develop and implement a cybersecurity curriculum for health care and intelligence information sharing through collaborations; training; awareness campaigns; and knowledge application areas such as secure design processes, phase-out of legacy systems, and improved investment. Additional studies are needed to create a sociotechnical framework that will support cybersecurity in health care systems and connect technology, people, and processes in an integrated manner.

(J Med Internet Res 2024;26:e46904) doi:10.2196/46904

KEYWORDS

health care systems; cybersecurity; sociotechnical; medical device; secure systems development; training; ransomware; data breaches; protected health information; patient safety

XSL•FO
RenderX

Introduction

Background

Cybersecurity in health care systems entails the safeguarding of electronic information and assets against unauthorized access, use, and disclosure [1]. The main objective of cybersecurity in health care systems is to protect the privacy, integrity, and accessibility of health information to provide secure health care services. Despite the digital transformation in health care delivery, health care organizations are facing increasing challenges and crises, which include data breaches of patient health information and vulnerability in their critical infrastructure [2]. Research has highlighted that health care systems are becoming more vulnerable to cyberattacks as technology advances [3]. Furthermore, the internet and its diverse nature and connection to the delivery of telehealth and continuous health care services create multiple points of access for cyberattacks [4,5].

In high-income countries such as Finland, the United States, and the United Kingdom, integrated technology is used to monitor and manage health care systems. For instance, at least 10 to 15 medical devices are linked to each patient's electronic bed in a public hospital [6]. These complexities increase the susceptibility of health care networks to cyberattacks [6,7]. Studies conducted through the simulation of medical devices have similarly revealed that pacemakers and pulse oximeters can be hacked and compromised without a physician's knowledge [8,9]. Ransomware is another type of man-made malware that can disrupt health care systems by infecting computer systems, locking people out of their files, and then demanding a ransom payment in exchange for access to those files [10,11]. Cyberattackers can publish the exposed health information to the web or sell it on the dark web [12]. This type of attack can result in breaches of patient privacy, subjecting health care organizations to fines that are consistent with human health service regulations and European General Data Protection Regulation (GDPR) policies for data breaches. For example, research has shown that, between 2012 and 2022, more than US $128,244,290 million in fines were paid in the United States alone for violations of Health Insurance Portability and Accountability Act laws on data breaches against health care organizations [13]. Although these fines were derived from no less than 111 health care organizations, many organizations have failed to report breaches.

Cybersecurity education is seriously lacking [14,15]. Moreover, a critical problem with cybersecurity in health care systems is the lack of involvement or recruitment of people with expertise and training in cybersecurity [16], resulting in considerable neglect of the cybersecurity infrastructure [17]. A systematic literature review (SLR) revealed that, between 2018 and 2019, more than 24% of the data breaches in all industries happened within the health care context [18,19].

Between 2009 and 2021, the US Department of Health and Human Services office reported 4419 health care data breaches, resulting in >314 million health care records being lost, stolen, or exposed [20]. In 2015, an estimated 113.27 million records were stolen and exposed, and in 2021 alone, the US Department of Health and Human Services also reported at least 2 health care data leaks daily [13]. The statistics clearly show an upward trend in health care data breaches over the past 10 years [21]. When considering this trend on a global scale, the number of health information breaches could potentially reach into the billions of health records. Organizations such as Vaastimo Oy Finland; National Health Service trusts in the United Kingdom; Anthem, Inc; Premera Blue Cross; and Excellus Health Plan have been victims of these threats and breaches of health information. Breaches and vulnerabilities in health care delivery, human safety, and protection of sensitive information are deeply disconcerting. However, it can be argued that research solutions are fragmented and sparse. There is a gap in the knowledge areas of health care cybersecurity in the literature and in practice regarding the vulnerability of health care systems and the reasons for cyberattacks. The argument and motivation are that a holistic approach to security is needed because humans are the weakest link in the cyberattack chain [11,22].

Coventry and Branley [6] have highlighted the need for resilience and changes in their studies on human behavior, technology, and processes as part of a holistic solution to the problem of health care system vulnerability. The information, technology, processes, objectivity and values, skills and knowledge, management systems and structure, and other resources dimensions by Heeks [23] also point out that avoiding security design reality gaps requires approaching the security functionality of a health information system as a sociotechnical system and not as a technical system. Security by design, or secure design, is an approach to cybersecurity that enables organizations to automate their data security controls and formalize the design of their infrastructure so that they can build security into their IT management processes [24,25].

In this study, a sociotechnical approach is defined as the interaction between humans and technology with the aim of creating technically efficient organizational information systems and user satisfaction [26]. Furthermore, conceptualizations of this approach are concerned with 3 primary dimensions: the social environment, technical environment, and organizational environment [27]. Sociotechnical design is identified as an approach to connect the integration of systems while ensuring that the multifaceted challenges and complexities in smart health care are well managed [28,29]. Smart health care can be defined as care that is equipped with smart IT, such as Internet of Medical Things (IoMT) devices that have the capabilities to anticipate and diagnose patient diseases; respond to treatments; guide, manage, and improve user comfort; and provide security and entertainment via hospital management systems. According to Coiera [30], "if healthcare is to evolve at a pace that will meet the needs of society, it will need to embrace the science of sociotechnical design." Therefore, the application of a sociotechnical perspective in health care cybersecurity in this study aimed at better understanding and mitigating the multifaceted challenges and poor uptake and performance of health care system security within health care organizations.

This existing gap in knowledge and practice was a major motivation for this SLR. It is necessary to connect the fragmented research and manage this knowledge gap regarding why health care systems are vulnerable to cyberattacks as the

XSL•FO
RenderX

study by Coventry and Branley [6] did not address this aspect in detail. An SLR was conducted to develop proactive cybersecurity strategies to mitigate threats and vulnerabilities that result in health care data breaches by proposing sociotechnical solutions and recommendations. Furthermore, to link human behavior, technology, and processes as highlighted by Coventry and Branley [6] and supported by the narrative review by Mohan et al [31] for further research, these 3 core areas can be interpreted as a sociotechnical framework [27]. It is essential to mitigate the increase in breaches of health information and protect health care from cybercrime and cyberattacks on critical health care infrastructure. However, none of these studies have examined why health care systems are vulnerable to attack through a sociotechnical lens. On the basis of this knowledge gap identified in the literature, the following research questions (RQs) were raised: (1) Why are health care systems vulnerable to cyberattacks? (RQ 1) (2) How can health care systems be protected? (RQ 2).

The objective of this review was to explore from a sociotechnical approach why digital health care systems are vulnerable to cyberattacks, provide sociotechnical solutions, and identify the areas of health care systems that need further improvement.

Previous Literature Review

Regarding the existing literature on health care cybersecurity, our previous SLR identified the following review themes: (1) cybersecurity threats and trends: studies that provide solutions and insights into threats and trends have been conducted to address cybersecurity threats and trends in health care systems [2,6,11,17,32,33]; (2) cybersecurity vulnerability: some studies have also investigated the cybersecurity vulnerability of health care systems to provide solutions and future directions for health care services [22,34-36]; and (3) cybersecurity interceptions in health care: studies have also investigated cybersecurity interceptions with health care systems to protect the security posture of these systems [12,19,37]—Coventry and Branley [6] have highlighted the need for further studies on human behavior, technology, and processes to further investigate why health care systems are vulnerable and provide a holistic solution to this problem.

Therefore, there is a need for further studies to identify the reasons behind the increase in health information breaches in health care systems. This area of study through a sociotechnical lens is lacking. Accordingly, our SLR critically investigated why health care systems are vulnerable to cyberattacks and expanded this area of study from a sociotechnical point of view. This review is significant given the lack of SLRs on the areas linking human behavior, technology, and processes using a holistic approach from a sociotechnical viewpoint in this context and as the studies by Coventry and Branley [6] and Mohan et al [31] were based on narrative reviews.

Methods

Protocol and Registration

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed to conduct our SLR using the checklist guide [38] (Multimedia Appendix 1). The aim of this review was to identify the reasons why health care systems are vulnerable to cyberattacks and provide sociotechnical solutions. In the planning stage of this review, a protocol for the sources of information, search strategies, study selection, criteria for eligibility, and data collection processes was created, and this review was not registered.

Eligibility Criteria

A paper was selected for inclusion if it was published in English and comprised a full-text version of the manuscript, review paper, conference proceeding paper, report, news article or website, or white paper published between 2012 and 2022. The introduction, abstract, results, and discussion sections of the paper were checked by the authors for conformity with the study objectives and critical appraisal using the checklist guidelines before inclusion. Research papers were excluded if they were not relevant to the research areas—cybersecurity, cybercrime, ransomware, and health care. These criteria are presented in Textbox 1.

Textbox 1. Article inclusion and exclusion criteria.

Inclusion criteria

- Study types: published peer-reviewed and original research papers (empirical and conceptual papers)
- Bibliometric study types: white papers and cybersecurity news reports in line with health care and cybersecurity
- Period: papers published between 2012 and 2022
- Language: English
- Subjects and domain: computer sciences, health care, and cybersecurity
- Requirements for paper inclusion: full-text papers.

Exclusion criteria

- Study types: unpublished work, editorial letters, textbooks, and research in progress
- Language: any other languages
- Subjects: studies outside the domain of cybersecurity and health care

XSL•FO
RenderX

Information Sources

To identify original research papers and review papers on cybersecurity in health care systems published between 2012 and 2022, a total of 6 databases (Web of Science, ScienceDirect, Scopus, PubMed, Springer, and the Institute of Electrical and Electronics Engineers) and a journal (*Management Information Systems Quarterly*) were searched. Furthermore, bibliometric records such as website reports, white paper reports, and magazine reports that supported cybersecurity in health care were also collected for the review. As a means of verifying the papers identified in our search, we searched Google Scholar using a search string.

Search Strategy

The following search string and keywords were used: (“cybersecurity” OR “cybercrime OR ransomware”) AND (“health care”) OR (“cybersecurity in healthcare”). Multimedia Appendix 2 provides more information.

Data Extraction

A total of 70 papers were extracted and recorded in a Microsoft Excel (Microsoft Corp) spreadsheet. The extracted data included information such as author or authors, year of publication, method, problem, and solution. The first author independently charted the data and updated the table to ensure the quality of the key findings drawn from the papers based on the recommendations of the second author. Critical appraisal was conducted to ensure the quality of evidence and the relevance of the articles. The data retrieved from the selected articles were analyzed.

Data Synthesis

The data from the literature were analyzed and synthesized using qualitative themes, which are presented in the following sections. The data were analyzed to identify the causes of vulnerabilities; solutions provided in the literature; and areas of classification based on sociotechnical, technical, and social perspectives in health care systems.

Results

Selection of Sources of Evidence

A total of 1257 papers were retrieved for the screening exercises. To determine whether the papers met our inclusion criteria regarding the topic domain, we began by scanning the abstracts and titles. The papers were reviewed by reading the full texts and determining their eligibility. Duplicated papers as well as those nonrelevant to cybersecurity, cybercrime, ransomware, and health care research were excluded. Furthermore, some papers were excluded after reading them in full and discovering that they were papers on research in progress. Finally, 70 papers were included in the analysis based on the eligibility criteria. Figure 1 illustrates the selection process.

The results of the SLR show the reasons why health care systems are vulnerable to cyberattacks and health care breaches. These reasons are the 5 vulnerability themes (Figure 2 and Table 1). Furthermore, the 5 vulnerability themes were classified into social, technical, and sociotechnical approaches.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for paper selection.

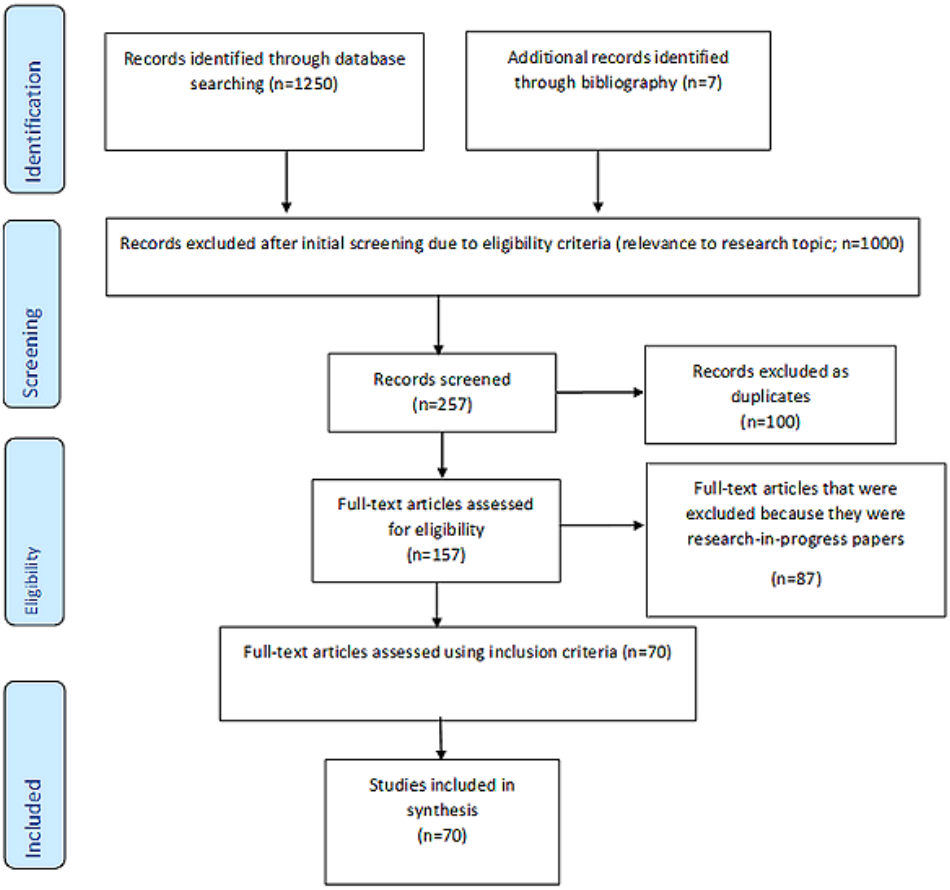

Figure 2. Results and insight into health care system vulnerability.

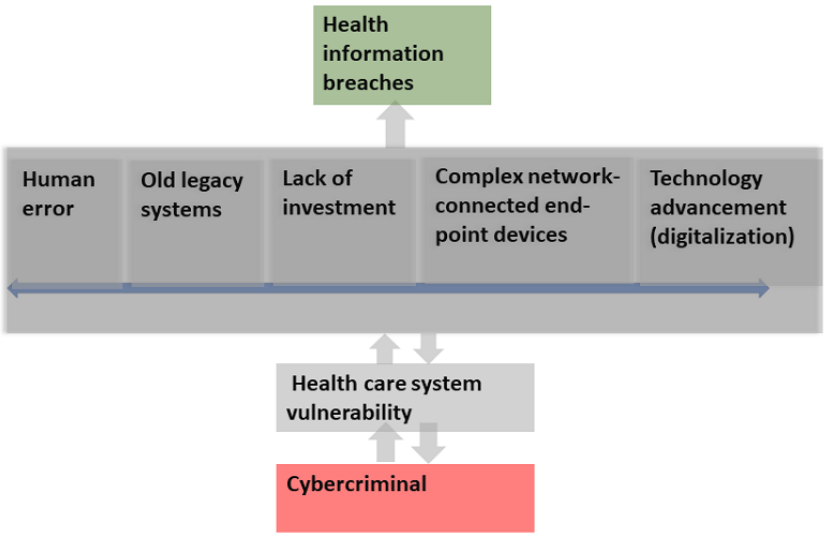

XSL•FO
RenderX

Table 1. Findings on health care system vulnerability categorized by themes and authors (N=70).

Vulnerabilities in health care	Type of approach	Studies, n (%)	References
Human error	Social	8 (11)	• Arndt [39] • Twitter [40] • Mukherjee [41] • Ponemon Institute [42] • IBM Security [43] • Scott and Wingfield [44] • Jalali et al [19] • He et al [36] • Gordon et al [45]
Old legacy systems	Sociotechnical	11 (16)	• Bouveret [46] • ECRI[a] Institute [47] • Sweeney [16] • Faruki et al [48] • Filkins [49] • Fu and Blum [50] • Offner et al [2] • McHugh [51] • Newman [52] • Scott and Wingfield [44] • Tully et al [53]
Lack of investment	Sociotechnical	15 (21)	• Argaw et al [11] • Emsisoft Malware Lab [54,55] • Branley-Bell et al [56] • Information Commissioner's Office, National Cyber Security Centre, and James M [57] • Kaspersky Inc [58] • PCEB[b] [59] • Rahman et al [60] • Gkioulos and Chowdhury [61] • Tully et al [53] • Williams and Woodward [34] • Coventry et al [62] • Jalali et al [19,33] • He et al [36] • Jalali and Kaiser [37]

XSL•FO
RenderX

Vulnerabilities in health care	Type of approach	Studies, n (%)	References
Complex network-connected end-point devices	Technical	36 (51)	• Burns et al [63] • Bouveret [46] • Chua [64] • Coventry et al [62] • Dameff et al [8] • Dienna et al [65] • ECRI Institute [47] • Filkins [49] • Francis [66] • Frost [3] • Twitter [40] • Giansanti [5] • Handa et al [67] • Offner et al [2] • Klonoff [9] • Lechner [68] • Lewis [69] • Lyon [70] • McHugh [51] • Mohan [71] • Newman [52] • Baranchuk et al [72] • Perakslis [73] • Peterson [74] • Sajedi and Rahbar Yaghobi [75] • Omotosho et al [76] • Singh et al [77] • Sittig and Singh [78] • Snell [79] • Tully et al [53] • Walker [7] • Williams and Woodward [34] • Jalali and Kaiser [37] • Jalali et al [19,33] • He et al [36]
Technology advancement (digitalization)	Technical	10 (14)	• Bhuyan et al [80] • Coventry and Branley [6] • Karambelas [4] • Kruse et al [17] • Raina MacIntyre et al [81] • Filkins et al [82] • PECB Insights [59] • Jalali et al [19,33] • Rodrigues et al [83]

[a]ECRI: Emergency Care Research Institute.

[b]PECB: Professional Evaluation and Certification Board.

The results also revealed that >24% of the data breaches from all industry clusters originated in the health care sector alone (Table 1) [19,21,84]. Other studies highlighted that organizations tend to spend more money on procuring new technology while committing only ≤5% of their budgets to the security of their critical health care systems [17,35]. Cybercriminals exploit health care systems due to the lack of investment, technology advancement as a result of digitalization, human error due to a lack of awareness and training, and old legacy systems, which enable cybercriminals to access valuable health information and sell it on the dark web for money and other gains [12]. The results reported a significant increase in data breaches and cyberattacks, with complex systems, IoMT devices, technology advancement, and network-connected end-point devices in complex connected heterogeneous health care systems identified as the major contributing factors.

The studies also identified a shortage of cybersecurity skills to contain cyberattacks or threats to health care organizations and systems [16]. The studies revealed that approximately 60% to 70% of health care organizations have witnessed breaches of health information without disclosure [85].

Human Error

Human error is a significant factor in the event of a cyberattack [11,22]. This shortcoming is one of the most crucial issues in health care systems as most cybercriminals use methods such as phishing to execute attacks with just a deceitful email. This is a social problem that can be addressed from a social approach. For example, human error posed a risk to the Geneva University

XSL•FO
RenderX

Hospitals [86]. Table 1 shows that 11% (8/70) of the studies acknowledged human error as the primary social reason for health care system vulnerability. Human error is attributed to a lack of skills and is a major trend in this ever-changing technological landscape, playing a role in several cybersecurity breaches [56]. From a technological point of view, a lack of expertise from humans and threats from human-related events are responsible for >70% of data fraud and breaches in business organizations (McCue, A, unpublished data, May 2008) [80] because of the value of health information on the dark web [6] and breaches in business organizations (McCue, A, unpublished data, May 2008) [80]. Furthermore, human-related threats have recently emerged as a growing concern.

Old Legacy Systems

Old legacy systems have been the basis of system development from the dawn of the medical device, operating system, and embedded mobile device era. Legacy operating systems such as Windows ME, Windows 2000, MS-DOS, UNIX, and firmware provide the foundation for system development. However, these systems pose a significant threat to health care sectors and organizations in our current era. Table 1 shows that 16% (11/70) of the studies acknowledged the vulnerability of health care systems to attacks due to old legacy systems. Such attacks occur from a sociotechnical approach, with cybercriminals exploiting humans and technology. Many data breaches, system incompatibilities, and security risks in health care systems and sectors are associated with legacy systems. Similarly, our SLR found that 85% of medical organizations use outdated operating systems or infrastructure [12,16]. Furthermore, Fu and Blum [50] raised concerns about organizations relying on unsupported software, alluding to medical devices that run on Windows XP operating systems with service packs but lack security updates. In addition, the case of the National Health Service 2017 WannaCry malware, which interrupted health care operations and shut down numerous hospitals by infecting thousands of computers, was caused by Windows XP software [87]. The authorities had been informed about the bugs but failed to act due to negligence. When a medical device is compromised, cybercriminals use it as a gateway to abuse hospitals, health care system networks, and health information or data. Perriello [88] and Meggitt [89] highlighted another issue, *Medijack*, referring to hackers hijacking medical devices to construct a back entrance into a hospital network. As a result, the use of a network of old legacy medical devices for administrative processes and care delivery increases the opportunities for an attacker or cybercriminal to easily intrude into hospital or health care organization networks and exploit and compromise the network of medical devices and health information. In this era of rapid medical technological advancement, health care systems also lack built-in security safeguards. Legacy systems do not support new technologies, and so the network of medical equipment in intensive care units, recovery rooms, operating rooms, and electronic health records (EHRs) will lack proper and secure communication and interoperability. Outdated legacy systems and unsupported operating systems are vulnerable to high-speed attacks. Furthermore, these problems are attributable to the lack of important updates to health care infrastructure. To support our point, health and human services should provide more guidance on applying the National Institute of Standards and Technology framework to the health care industry and consider appropriate incentives that would allow health care organizations to phase out old vulnerable legacy systems [16].

Lack of Investment

Investment in the health sector will yield better outcomes and quality health care delivery. According to our analysis and results, the health care sector suffers from underinvestment, and crucial infrastructure and training for health care cybersecurity are disregarded [6], which is one of the primary causes of the increase in sensitive health information breaches. Investment can be seen in social (human) and technical (technology) aspects. As shown in the analysis in Table 1, a total of 21% (15/70) of the studies acknowledged the lack of investment and advised both directly and indirectly regarding the necessity of cybersecurity investment in the health care industry [55,56]. The analysis acknowledged and revealed that the health care sector lagged more than other sectors in terms of health information protection and breaches. Furthermore, the findings of our SLR revealed that 80% to 85% of worldwide breaches occur in the health sector [4], whereas 45% to 90% of health care organizations have witnessed one or more threats or breaches [18,57]. Investment in critical infrastructure for health care and best practices in cyber hygiene will aid in the protection of health care systems from potential vulnerabilities. Proper investment will ensure the safeguarding of personal information and render health care systems more resilient to cyberattacks.

Complex Network-Connected End-Point Devices

Medical end-point devices have long served as a hospital's backbone for treatment, diagnosis, and precision-based technological applications to complement health care service operations and management. To fully exploit their potential, the medical device development pattern has shifted from traditional-based medical device system development to a network of wireless, connected end-point technological devices with built-in communications and remote connectivity. Complex network-connected end-point devices have increased the cyberattack surfaces in conjunction with their complexity and technological systems as heterogeneity in nature of medical technology has evolved. Complex network devices are classified as a technical challenge from the perspective of technical security system design. The analysis in Table 1 shows that 51% (36/70) of the studies acknowledged network-connected end-point medical devices as the most significant technical reason for health care systems' vulnerability to cyberattacks. The operational modes continue to evolve with more interconnections between new applications and devices such as cloud-based applications, third-party software, IoMT devices, and system networks in health care environments. Lechner [68] revealed that original equipment manufacturers are now creating interconnected medical devices without incorporating proper cybersecurity features into the development life cycle of medical and end-point device systems. The vulnerability of the end point requires urgent attention; otherwise, cybercriminals will continue to use the weakness of connected devices to access personal health information. According to research and cybersecurity

XSL•FO
RenderX

stakeholders, wearables, implanted devices, and sensors may become the new targets of future exploits [6,8]. As shown in Table 1, complex network-connected end-point medical devices also require medical technology security by design [72,90] as a solution strategy to protect critical health care infrastructure from breaches. In the past, medical device system development has primarily focused on critical performance and safety. Furthermore, the security aspects of these medical devices are not a factor during the planning and development process. The process indicates that developing traditional or stand-alone systems of noninterconnected devices was a suitable method for designing the traditional approach. These are the current legacy systems that lack interoperability, updates, security design, or compatibility. Furthermore, connected medical devices such as sensor-controlled drug infusion pumps, cardiac pacemakers, pulse oximeters, and network-connected x-ray machine components such as picture archiving and communication systems are vulnerable to cybersecurity threats and attacks [5]. To continue solving cybersecurity issues in medical devices, developers and actors must recognize the importance of the health care environment's complex operations. In addition, there should be incident reports, an audit trail in the device system database, and paper-based documentation of technical vulnerabilities [34]. Medical device manufacturers such as security experts or systems integrators must address this issue because, with a single cyber vulnerability, cybercriminals or hackers can exploit medical technology connected to the internet, compromising data integrity, wearable sensor readings, protected health information, patient safety, and care outcomes [2,50]. When cyberattackers manipulate systems or deposit a virus, this could cause medical device software or systems to malfunction, resulting in abnormal effects or different readings from the systems, such as implantable medical devices that take and display incorrect readings [5,8].

Technology Advancement (Digitalization)

Technology advancement has enabled unique access and benefits to revolutionize health care systems in terms of precision. Modern medical care now relies on health care delivery organizations, including hospitals and clinics, built on a backbone of connected computer-based infrastructure. Over the past 30 years, the expansive integration of new health care technology has changed the face of medicine [53]. However, the rapid digitalization in health care delivery, where medical devices are intertwined in a digital network setting and system to ensure the precision of health care delivery with the use of IoMT and digital devices, has created gateway access for cyberattacks, risks, and vulnerabilities [37,81]. Table 1 shows that 14% (10/70) of the studies acknowledged technology advancement due to digital transformation as the reason why health care systems are vulnerable to cyberattacks. This type of attack and vulnerability usually occur from the technical areas of cyberattacks, for example, a technology error such as glitches and design errors. One example of vulnerability is St. Joseph Hospital in California, where the health information of 31,800 patients was made public through a basic internet search engine for >1 year without anyone noticing. The underlying issue was that security settings on the medical devices were not correctly configured [91]. As technology continues to evolve, IoMT will become more inseparable in health care service delivery, which will create more vulnerabilities if health care organizations continue to disregard cybersecurity threats without proactive readiness to address them in this era of Industry 4.0. These vulnerabilities pose threats to the security and privacy of human and health information.

Studies have shown the health care sector to be unequipped and lacking in investment [11,92]. For example, the use of electronic health technology, motivated by acts such as the Meaningful Use program introduced by the US government, has compelled many health care organizations to increase the use of digital technology in health care, such as EHRs and electronic data exchange, and comply with enhanced health care delivery management. Organizations began to focus on adopting new technology and spending less on security, creating part of the problem [32]. Technological advancements and a federal policy mandate ultimatum are 2 of the causes noted in this SLR that have increased health care industry exposure to cyberattacks and breaches of health information [17]. Therefore, an organization should have proper planning; be proactive instead of reactive; and ensure the protection of health technology, information, patient privacy, and security when implementing or adopting advanced technology [17,80]. One such process is to ensure that a medical technology statement of disclosure and liability is included during the procurement, integration, and adoption of a technology. Support services and maintenance during and after procurement and installation should be part of the procurement process. Furthermore, the device manufacturer should also consider security in product development planning. Digital technology should also have the capability to monitor and collate threats and patterns and log these in a risk assessment register for analysis and improvement or threat containment.

Causes of Vulnerabilities in Health Care Systems

Figure 3 shows the causes of vulnerabilities in health care systems, which complement the findings regarding health care vulnerability, and categorizes them accordingly. The following sections address these vulnerabilities.

XSL•FO
RenderX

Figure 3. Causes of vulnerabilities in health care systems.

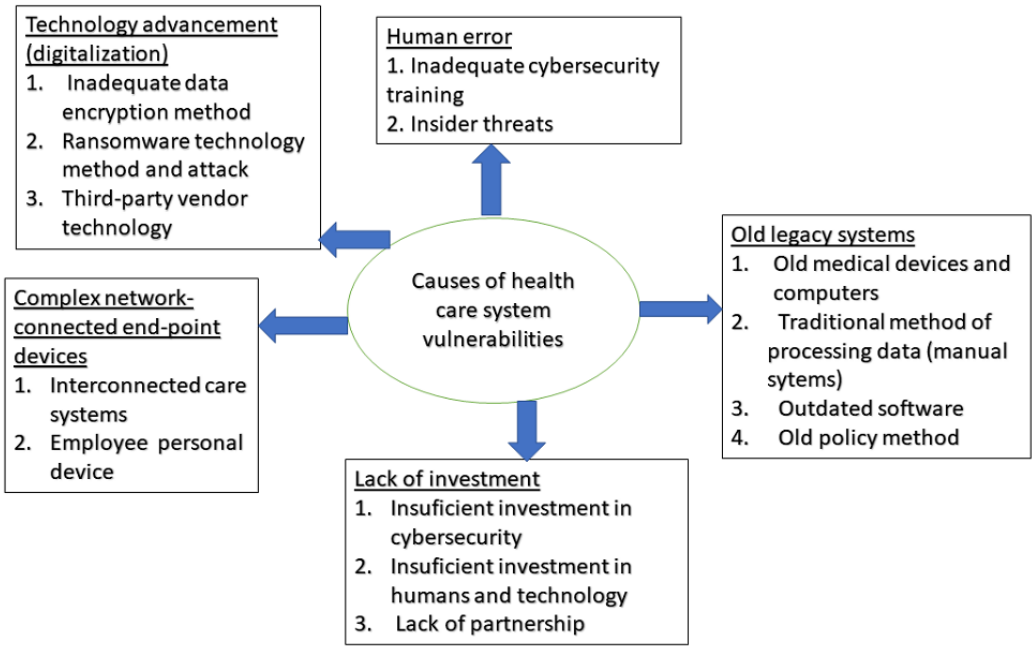

How Can Health Care Systems Be Protected?

Overview

This study summarizes how health care systems can be protected from cyber threats and cyberattacks and presented in Table 2.

XSL•FO
RenderX

Table 2. Health care system protection.

Health care vulnerability and description of challenges	Proposed solutions	References	Health care cybersecurity sociotechnical areas of application
Human error			Social approach
Information breaches and identity theft	• Inform human health office and owners of the data, train staff, learn to encrypt information, and have a backup plan and rollover system.	• Tuttle [93]	
Insecure behavior	• Implement training.	• Coventry et al [62]	
Cyber warfare	• Foster awareness and implementation of cyber hygiene. • Implement data encryption, network defense solutions, and protection of premises.	• Mukherjee [41]	
Employee negligence and error	• Implement training, invest in new skills for staff, and launch awareness campaign.	• He et al [36]	
Cybersecurity ethical issues, such as the disclosure and use of health information without consent	• Seek patient consent and balance privacy and autonomy for health information and usability.	• Loi et al [94] • Christen et al [95]	
Old legacy systems			Sociotechnical approach
Interoperability issues and incompatible device challenges	• Procure modern devices to enable seamless synchronization of devices and networks.	—[a]	
Interoperability issues	• Implement health policy, regulation compliance, and upgrades.	—	
Inability to update software and medical devices	• Phase out legacy systems.	• Sweeney [16]	
Lack of investment			Sociotechnical approach
Disregard of health care cyber critical infrastructure	• Invest in cyber critical systems.	• Kruse et al [17]	
Protect data, operations, and valuables	• Invest in cybersecurity protection mechanisms for sensitive activities.	—	
Design and device usability issues for processes and data security management	• Invest in human behavior, technology, and organizational processes.	• Coles-Kemp and Williams [96]	
Complex network-connected end-point devices			Technical approach
Cyberattack on hospital health care systems	• Defend the hospital with network security solutions. Have a backup and a roll-back system. Ensure that all standard policy and comprehensive guidelines are in place and always train staff to respond.	• Argaw et al [11]	
In case network-connected medical devices through the IoMT[b] are exposed	• Protect devices through assessment and extreme network defender solutions. Encrypt networks.	• Frost [3]	
Vulnerabilities due to sensor and IoT[c] devices	• Implement device simulation, security assessment, and extreme network defender solutions.	• Dameff et al [8]	
Vulnerability of end-point devices	• Develop network and device security protection solutions.	• Lewis [69] • Singh Rayat et al [77]	
Technology advancement (digitalization)			Technical approach

XSL•FO
RenderX

Health care vulnerability and description of challenges	Proposed solutions	References	Health care cybersecurity sociotechnical areas of application
Lack of security in medical devices and critical infrastructure	• Ensure that medical devices are designed with security before procurement and ensure that device manufacturers maintain and manage security.	• Lechner [68]	
Health care big data protection challenges	• Secure life cycle model and encryption through blockchain.	• Khaloufi et al [97]	
Health care system digitalization and medical device vulnerability	• Implement cyber hygiene and security in designing devices.	• Coventry and Branley [6]	
Digitalization and technology advancement vulnerability gap (digital dark alley) challenges	• Update firewall installations and use a secure design approach, cloud recovery planning, and backup.	• Karambelas [4]	

[a]Not applicable.

[b]IoMT: Internet of Medical Things.

[c]IoT: Internet of Things.

Human-Related Case Type and Challenges

The protection of health care systems from cyberattack-related vulnerabilities caused by human error, such as identity theft and health information breaches, requires by law that health care organizations inform the human health office, regulatory bodies, and data owners [93] to ensure compliance with ethical and privacy standard regulations [94,95]. A security compliance officer should also be employed to guide and ensure that proper cyber hygiene measures are in place to avoid such occurrences. It is important to ensure that health information is encrypted to assure that data are unusable and back up data offline and on the web. Furthermore, in cases in which a health care organization is saddled with challenges due to insecure human behavior, such as employee negligence, a lack of skills, and cyber warfare, the organization must ensure proper training of all staff [62] and implement awareness programs using a comprehensive guide to avert cyber threats [36,41]. This proposed solution requires a social approach in designing guidelines and training programs.

Old Legacy Systems Case Type and Challenges

Interoperability and compatibility challenges in medical devices stem from human-related activities within health care systems, potentially impacting the persistence of outdated legacy systems [50]. Therefore, to holistically protect health care systems, proposed solutions involve sociotechnical measures due to the old legacy in human work processes, organizational structures, and technology tasks, as mentioned by Offner et al [2]. Organizations should adhere to policies and standards linked to the old legacy, ensure proper updates and upgrades, and implement patches. Modern equipment that supports security and carries out updates must be procured to avert crises and phase out legacy systems [16].

Lack of Investment Case Type and Challenges

Investment in critical health care infrastructure is very important to ensure a health care ecosystem that is secure from cyberattacks and vulnerabilities. The susceptibility of health care to cyberattacks is a result of the underinvestment in and neglect of cybersecurity infrastructures. Kruse et al [17] also highlighted that a health organization invests ≤5% in cybersecurity but tends to focus on integrating and delivering care. It is important for a health care organization to invest in technology, human behavior, and processes [96] to protect sensitive and valuable health information from breaches and attacks.

Complex Network-Connected End-Point Devices Case Type and Challenges

The increase in health information breaches in hospitals is attributed to complex network-connected end-point devices, which are vulnerable to cyberattacks because sensor-based medical devices and system networks are interlinked and connected to the internet [8]. Internet of Things devices are vulnerable because they can be controlled through a media access control address and network. A proposed solution identified in this SLR highlighted that health care can be protected though proper encryption of data and installation of network defenders [3]. It is important that medical device simulation and assessment be performed through vulnerability analysis to ensure that devices are not tampered with or compromised [8].

Technology Advancement (Digitalization) Case Type and Challenges

Technology advancement has revolutionized the health care delivery process using digital technological processes. Manufactured medical devices enable patients to be diagnosed remotely, and physicians can administer care using telemedicine. However, technological advancements still lack security in the design of these devices because security is an afterthought during development, which makes them vulnerable to cyberattacks [5]. A proposed solution is that health care organizations must ensure that medical device security starts from the planning stage [68] and that device manufacturers maintain and manage security in the pre- and postmarket phases. This solution paradigm must be catalogued as a technical

XSL•FO
RenderX

measure. Hospitals with modern-day smart care should leverage comprehensive guidelines and compliance with standards such as those of the International Organization for Standardization or International Electrotechnical Commission 27001 or 27002, as well as cyber hygiene to enable effective and efficient care delivery processes [4,11]. Therefore, the implementation of solutions should always adopt a sociotechnical approach [96].

Intervention Application Areas and Domain Counts for 2012 to 2022

The selected studies from this SLR that discussed and presented knowledge interventions and solutions applied in some health care sectors between 2012 and 2022 are categorized and presented in Table 3.

XSL•FO
RenderX

Table 3. Intervention application areas and domain count for health care cybersecurity between 2012 and 2022 (N=70).

Vulnerability and knowledge application domain	Solution papers published in this domain between 2012 and 2022, n (%)	References
Human error		
Training	12 (17)	• Karambelas [4] • Giansanti [5] • Dameff et al [8] • Argaw et al [11] • Bhuyan et al [80] • Offner et al [2] • Holst et al [98] • Branley-Bell et al [56] • Chowdhury and Gkioulos [61] • Khando et al [99] • Coventry et al [62] • Information Commissioner's Office, National Cyber Security Centre, and James M [57]
Awareness	4 (6)	• Walker [7] • Filkins et al [82] • Kaspersky Inc [58] • PCEB[a] [59]
Education	2 (3)	• Rahman et al [60] • Francis [66]
Intelligence information sharing	5 (7)	• Bouveret [46] • Winton [100] • Dobuzinskis and Finkle [101] • Scott and Wingfield [44] • Lewis [69]
Old legacy systems		
Health policy and standards	25 (36)	• Sweeney [16] • Bouveret [46] • Newman [52] • Coles-Kemp and Williams [96] • Snell [79] • Emsisoft Malware Lab [54,55] • Kruse et al [17] • Rajamäki and Pirinen [90] • The HIPAA[b] Journal [13] • Hippa [13] • Khaloufi et al [97] • Tuttle [93] • Perakslis [73] • Ponemon Institute [42,85] • Tully et al [53] • Bhuyan et al [80] • Williams and Woodward [34] • Lechner [68] • McHugh [51] • Burns et al [63] • ECRI[c] Institute [47] • Loi et al [94] • Information Commissioner's Office, National Cyber Security Centre, and James M [57] • Kaspersky Inc [58] • PCEB [59]
Lack of investment		
Partnership	3 (4)	• Baranchuk et al [72] • Raina MacIntyre et al [81] • Chua [64]

XSL•FO
RenderX

Vulnerability and knowledge application domain	Solution papers published in this domain between 2012 and 2022, n (%)	References
Complex network-connected end-point devices		
Participatory design science (sociotechnical)	1 (1)	• Coles-Kemp and Williams [96]
Network security	16 (23)	• Frost [3] • Sittig and Singh [78] • Twitter [40] • Arndt [39] • Bickers et al [102] • Ponemon Institute [42,43] • Filkins [49] • Williams and Woodward [34] • Zorabedian [103] • Sajedi and Rahbar Yaghobi [75] • Omotosho et al [76,104] • ECRI Institute [47] • Djenna et al [65] • Mohan [71] • Baranchuk et al [72] • Singh et al [77]
Encryption	4 (6)	• Mukherjee [41] • Filkins [49] • Mohan [71] • Singh et al [77]
Technological advancement (digitalization)		
Machine learning	8 (11)	• Omotosho et al [76] • Zarour et al [12] • Khaloufi et al [97] • Reshmi [10] • Faruki et al [48] • Handa et al [67] • Chen et al [105] • Sajedi and Rahbar Yaghobi [75]
Blockchain	1 (1)	• Bhuyan et al [80]
Security by design	6 (9)	• Coventry and Branley [6] • Lyon [70] • Coles-Kemp and Williams [96] • Lechner [68] • Fu and Blum [50] • Andrea [74]

[a]PECB: Professional Evaluation and Certification Board.

[b]HIPAA: Health Insurance Portability and Accountability Act.

[c]ECRI: Emergency Care Research Institute.

Knowledge Application Domains and Vulnerabilities

The vulnerabilities listed in Table 3 reveal that human error was associated with interventions linked to one of the knowledge application domains of training, awareness, education, and intelligence information sharing.

Training

Employee training is important to avoid human factors or error challenges in health care. Table 3 shows the proposed solutions and interventions for training from 17% (12/70) of the studies. Figure 4 shows that training emerged in 2018 at 1% and increased to its peak between 2019 and 2021. However, this finding suggests the need for cybersecurity training in health care to manage human vulnerability challenges. This need is supported by the literature highlighting the importance of cybersecurity skills and education for health care professionals [16] and the need for investment in this area [17].

XSL•FO
RenderX

Figure 4. Knowledge application areas and domain count for health care cybersecurity between 2012 and 2022.

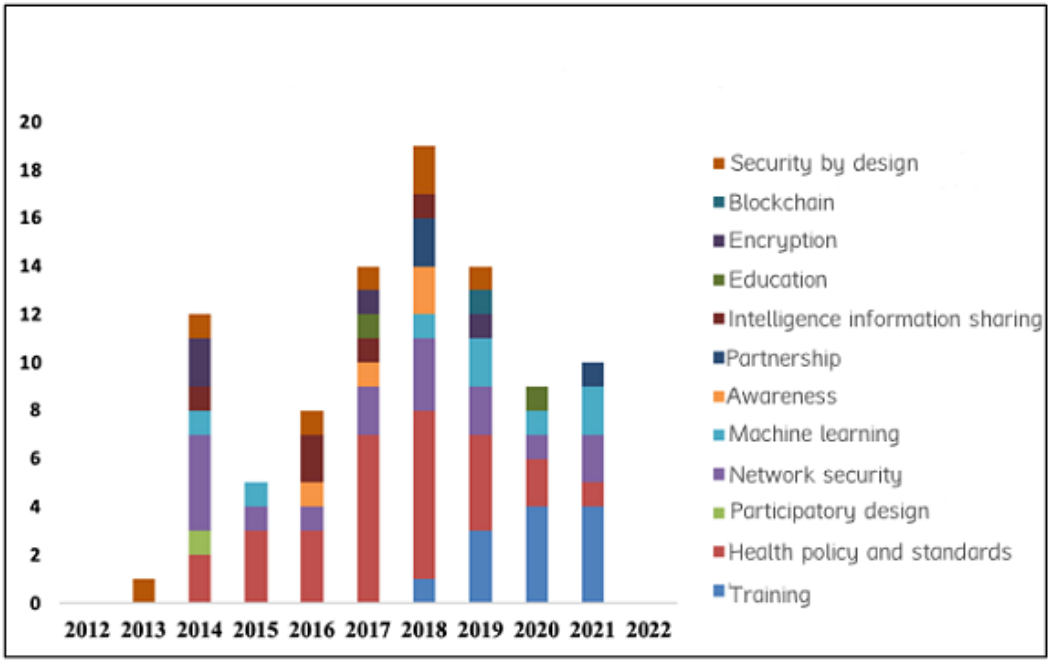

Education

The solutions presented regarding educational intervention were derived from 3% (2/70) of the studies (Table 3). Figure 4 shows that educational solutions emerged in 2017 and declined until 2020, when studies on educational intervention emerged. This finding is supported by research that shows a lack of educational skills [16]. Organizations must invest in educational training and skills to curb social and technical cybersecurity vulnerability in health care.

Awareness

A total of 6% (4/70) of the studies in Table 3 presented solutions on awareness to address the vulnerability of human errors. This small number of studies has shown a decline and a lack of cybersecurity awareness program in health care systems. Figure 4 similarly shows that cybersecurity awareness emerged in 2016 and reached its peak at 2 studies. This has been validated by previous studies that indicate a lack of awareness programs and training [45,62].

Intelligence Information Sharing

Table 3 also shows that intelligence information sharing was a solution investigated in 7% (5/70) of the studies. It can be seen that information sharing emerged in 2014 and declined in 2015 before re-emerging in 2017 and 2018 at the rate of 1 study each year. This finding also shows that health care organizations should collaborate in training and intelligence information sharing to address cybersecurity challenges in health care.

The vulnerabilities listed in Table 3 reveal that old legacy systems were associated with interventions linked to the knowledge application domain of health policy and standards.

Health Policy and Standards

The knowledge intervention analysis indicates that 36% (25/70) of the studies acknowledged and were linked to health policy and standards (Table 3). The analysis shows that governments and health care organizations have proposed more interventions or solutions regarding health policy and standards to regulate health care organizations. The policy studies shown in Figure 4 emerged in 2014 and continued to increase to their peak in 2018. Policies such as the Health Insurance Portability and Accountability Act, the GDPR, and the Health Information Technology for Economic and Clinical Health Act to engineer has helped to mitigate data breaches and vulnerabilities in health care organizations in addressing old legacy systems to avoid sanctions and fines in case of breaches. However, full implementation or enforcement of day-to-day monitoring in hospitals or health care organizations remains challenging.

The vulnerabilities listed in Table 1 reveal that a lack of investment was associated with interventions linked to the knowledge application domain of partnership.

Partnership

Partnership is key to sustaining and protecting health care systems from cybersecurity vulnerability [72]. When organizations fail to invest in critical cyber infrastructure, skills, and partnerships with governments and expert security organizations, it is likely that they will be vulnerable to cyberattacks and breaches of health information and lack the capability to protect health care systems from the vulnerability of underinvestment. Table 3 shows that partnership solutions were provided in 4% (3/70) of the studies, whereas Figure 4 shows that partnership emerged in 2018 and declined in 2021. There is a need for health care organizations to partner for better capability and structure to protect health care systems [64].

The vulnerabilities listed in Table 1 reveal that complex network-connected end-point devices were associated with interventions linked to the knowledge application domains of participatory design, network security, and encryption.

XSL•FO
RenderX

Participatory Design

Health care organizations and medical device manufacturers must jointly participate in designing processes and systems to avoid a sociotechnical design gap. This collaboration will help protect health care systems and increase the acceptability of organizational systems and productivity. Table 3 shows only 1 pertinent study in 2014. This infer that participatory design is one of the reasons for the vulnerabilities in complex network-connected end-point devices in health care systems. Health care systems comprise a complex environment that requires a sociotechnical and collaborative approach to addressing challenges [2].

Network Security

Network security solutions were covered in 23% (16/70) of the studies (Table 3). A number of intervention solution studies were conducted in this domain. As shown in Figure 4, the first increase was observed in 2014 with 4 studies, a decline to 2 studies was observed in 2017, and then the number of studies increased to 3 before a final decline to 2 studies in 2021. These studies still attest to the vulnerability of complex network-connected end-point devices, which require increased interventions to solve health care vulnerability challenges.

Encryption

The encryption technological solution in this review was mentioned in 6% (4/70) of the studies. There was a limited number of solutions regarding encryption intervention in this review (Figure 4). Encryption only emerged in 2014 with 2 studies, and there was a gap in studies until 2017 and 2018. This finding shows that health care organizations need to implement encryption technology to protect valuable health information from breaches and attacks [77].

The vulnerabilities listed in Table 1 reveal that technology advancement (digitalization) was associated with interventions linked to the knowledge application domains of machine learning, blockchain, and security design.

Machine Learning

Machine learning is a new area in which cybersecurity in health care systems is evolving. However, solutions were provided in only 11% (8/70) of the studies (Table 3). This technology surfaced in 2014 according to Figure 4. There was only 1 study in 2014 and 2015. No solutions were provided until 2018, and the number of interventions categorized under technology advancement increased from 2019 to 2021.

Blockchain

Blockchain technology is new and still lacking solutions according to this SLR, where only 1% (1/70) of the studies showed an effective intervention. Blockchain surfaced in 2019, as shown in Figure 4. Additional solutions and interventions are needed as this area is promising and can be categorized under technology advancement (digitalization) as the key to protecting smart health care systems.

Security by Design

Security by design is a strategy that demands that health care organizations implement auto-based technology to protect digital health care systems. Table 3 shows that 9% (6/70) of the studies acknowledged security by design as a solution for technology advancement to prevent vulnerability in digital systems. Figure 4 shows studies on secure design in 2013 to 2014. There were no studies in 2015, whereas in 2016 to 2019, some studies provided interventions. There is a need for more solutions in this area to protect technological advancement or digital health care systems from vulnerability [68].

Summary of the Knowledge Application Domains and Vulnerabilities

In summary, the findings of this SLR indicate that interventions provided for the containment of health care cybersecurity vulnerabilities were limited over the past 11 years. This SLR also revealed that interventions regarding the rate of technological advancements in addressing health care cybersecurity challenges were inconsistent and lagging between 2012 and 2022. Findings also indicates that interventions in some of the mapped variables were scarce between 2012 and 2022 (Table 3). Few or no solutions are provided to address the challenges in many domains regarding health care vulnerabilities.

Discussion

Brief Summary of Findings

This SLR provided a synthesis of literature on cybersecurity in health care and identified the reasons why health care systems are vulnerable to cyberattacks. This review analyzed 70 published studies and identified 5 vulnerability themes of cybersecurity in health care systems and also proposed sociotechnical solutions for health care organizations.

The findings indicate that the extensive vulnerability of health care systems is due to internet-connected devices and software applications. Health care organizations face significant challenges, such as medical end-point device complexities and saturated wireless medical technology resulting in its difficulty in securing an interconnected technological landscape.

Importantly, many cyberattacks occur within this interconnected network without the health care organization's awareness, contributing to health information breaches.

Our findings also underscore that the crucial role of investment in health care organizations is a key panacea for addressing cyberattacks and threats. Thus, lack of investment leverages the other vulnerabilities.

In addition, this study found that lack of adequate preparation for the potential threats or vulnerability in shifting to the digitalization of health care is also a contributing factor to most successful cyberattacks on health care organizations.

We found that human activity also played a major role in subjecting health care systems to cybercrimes. The decision of humans to develop medical devices, health software applications, management systems, and processes in an effective and secured manner is vital in safeguarding health information. However, there is a bit of disconnect in the human-centric design in health care system development, most importantly during

XSL•FO
RenderX

the planning of procurement of medical technology and systems and the integration between health care organizations and stakeholders such as medical device developers, health care professionals, cybersecurity compliance officers, and system integration experts. Generally, the findings revealed that health care organizations lack adequate cybersecurity preparations during transitions to digitalization.

The findings also revealed that the health care cybersecurity knowledge application domain areas in Figure 4 depict that more intervention studies over the past 11 years were focused on health policy and standards.

In Table 4, solutions are proposed from a sociotechnical perspective to counteract cybersecurity vulnerabilities in health care organizations.

Further findings on the vulnerabilities and implications for future research are discussed in the following sections.

Table 4 is an integrated table that is presented in a stand-alone view for health care system solutions from a sociotechnical viewpoint.

To protect health care systems from attacks and vulnerabilities, as shown in Table 4, through the intervention of effective and noneffective studies, it can be seen that sociotechnical intervention studies classified invention most often and were the most effective. There are patterns and convergences between technical solutions and sociotechnical solutions in their domain of applications and solutions, such as a lack of investment, complex network-connected end-point devices, old legacy systems, and technology advancement, which lean toward interventions.

While we can consider human errors in human-computer interactions and technology usability from a human perspective, design and management can be approached through a sociotechnical perspective [96]. This approach also considers the final users of digital health care systems. Organizations would benefit from leveraging the sociotechnical solutions and guide in Table 4 in the case of cyberattacks attributed to human error by training all staff to respond using a comprehensive guide to avert cyber threats [62]. Challenges of technology, such as network-connected end-point devices and technology advancement for digitalization, should be addressed through network and security solutions and encryptions [6,67].

Hospitals with modern-day smart care should leverage their comprehensive guidelines and standard International Organization for Standardization or International Electrotechnical Commission 27001 and 27002 compliances.

Health care organizations should ensure and implement proper cyber hygiene to enable effective and efficient health care delivery processes [4,11]. They should increase their budget for critical cyber systems to address the lack of investment [17] and phase out old legacy systems by increasing investment. These actions will enable resilience and preparedness for future response plans and mitigations.

XSL•FO
RenderX

Table 4. Health care system solutions from a sociotechnical viewpoint.

Vulnerability, knowledge application domain, and description of challenge or case type	Sociotechnical lens	Effective	Not effective
Human error			
Training			
• Ransomware or email phishing attack	• Sociotechnical solution	• Train and educate health care staff to use encrypted solutions for data and virus risk register; stay up-to-date on trends of virus attacks for health care systems [4,5,57]	Review cyberattacks against hospitals worldwide via training workshops through teleconferences with experts; incorrect training approach and method of delivery via teleconference [11]
• Cyberattack on critical medical infrastructure and device breaches • Ineptitude of employees regarding cybersecurity in managing health records	• Sociotechnical solution • Sociotechnical solution	• Train and educate clinicians through simulations of hacked medical devices for patient care to heighten their awareness [8,61] • Implement training for cybersecurity culture and proactive maturity resilience via human-computer interactions [2]	Review cyberattacks against hospitals worldwide via training workshops through teleconferences with experts; incorrect training approach and method of delivery via teleconference [11]
• Insecure behavior of staff	• Social solution	• Assess behavior of health care staff regarding cybersecurity (insecure behavior) Apply AIDE[a] behavior change techniques to ensure secure behavior [56,62]	—[b]
• Health information attacks and identity theft	• Sociotechnical solution	• Provide employees with ISA[c] content development material and enhance and analyze security behavior in public and private sectors • Apply gamification • Develop prototype game and behaviorism theory and mental model for private-sector training Apply real game and ANT[d] for public-sector training [99]	—
• Protection of health care system infrastructure	• Sociotechnical solution	• Implement cybersecurity planning and training using the CERT RMM[e] [79]	—
• Low digital literacy skills of employees	• Sociotechnical solution	• Implement essential and advanced digital literacy training via computers and smart devices [98]	—
Awareness			
• Inadequate cybersecurity awareness regarding the IoMT[f] devices • Lack of data protection compliance awareness	• Sociotechnical solution	• Apply cross-situational awareness model of IoMT devices for employees and management [7] • Provide awareness training on HIPPA[g] and GDPR[h] guidelines [7,59]	—
Education			
• Employee cyberbullying • Hacking and vulnerabilities of medical devices	• Sociotechnical solution	• Provide gamification education for web-based cyberbullies [60] • Provide awareness and educational programs on the vulnerabilities of medical devices [66]	Report on pacemaker hack that led to a disconnection based on a study; the study was generalized with speculation [66]

XSL•FO
RenderX

Vulnerability, knowledge application domain, and description of challenge or case type	Sociotechnical lens	Effective	Not effective
Intelligence information sharing			
• Notification alert of threat to critical infrastructure protection • Hospital management afraid to report data breach and cyberattack to protect their image	• Social solution	• Implement threat intelligence solution [58]. • Recruit and contact compliance officer and information sharing center to report breach [46,59,100].	—
Old legacy systems			
Health policy and standards			
• How can we manage cybersecurity vulnerability risks	• Sociotechnical solution	• Implement cybersecurity risk framework [46].	—
• Our devices lack updates	• Sociotechnical solution	• Provide updates and patches for legacy systems [57]	—
• What is the lasting solution for legacy systems	• Sociotechnical solution	• Phase out legacy systems and procure devices with a security update that supports aftersales	—
• Curtailing health care breaches	• Sociotechnical solution	• Implement GDPR and HITECH[i] policy for medical devices and data [13,42,57,58,85].	—
Lack of Investment			
Partnership			
• We are concerned with the threat alerts for implanted cardiovascular medical devices. • Lack of support to manage implantable devices such as pacemakers • Managing threats with stakeholders to protect patients	• Sociotechnical solution	• Ensure security in design from manufacturers and partners for aftersales support to ensure updates with remote monitoring or interrogation [72] • Ensure a partnership for a safer cardiovascular implantable device with the manufacturer's electronic device and follow FDA[j] and NIST-CSF[k] guidelines [72] • Health care organization should partner and implement HICP[l] guidance [64]	Developed new biosecurity risk methods and surveillance tools from traditional methods; they lack validation [81]
Complex network-connected end-point devices			
Participatory design science (sociotechnical)			
• Information security design gap challenges for health care systems	—	• Resolve information security design reality gap using the ITPOSOM[m] framework by Heeks [96] and through collaboration [65].	—
Network security			

XSL•FO
RenderX

Vulnerability, knowledge application domain, and description of challenge or case type	Sociotechnical lens	Effective	Not effective
• Insecurity of connected medical devices in protecting health information • Managing network security for IoMT devices	• Sociotechnical solution	• Install extreme network defenders to secure the network and manage IoMT devices [3]	Health record breaches in Australia are reportedly sold on the dark web; the study does not offer a solution [102]
• Attack on critical health care cyber infrastructure	—	• Develop a collaborative security approach and cybersecurity guidelines [65]	—
• Managing complex health care network access control and authentication	• Technical solution	• Implement the attribute trust framework for aggregation of user attributes in a reputation system [71]	—
• Protection of EHRs[n] for patient safety challenge	• Sociotechnical solution	• Apply the 3-phase e-PSG[o] framework [78]	—
Encryption			
• Protection of IoT[p] devices from breaches and being compromised	• Technical solution	• Secure IoT devices through FHSS[q] and RSSI[r] techniques [77]	Anthem's insurance health record breach report; investigation revealed that a foreign government was behind the attack, which is speculation without evidence-based facts [41]
• Managing cloud security concerns • Managing employee and patient devices on the health care network	• Sociotechnical solution	• Assure investment and compliance with regulatory standards and monitoring • Implement policy on BYOD[s] and apply all-layer multifactor protections for cloud systems [49]	—
• Protecting sensitive health care data and exchange between the EHR and the cloud-based database	• Technical solution	• Encrypt data using lightweight cryptographic protocols; store on the cloud-based PHR[t] [71]	—
Technology advancement (digitalization)			
Machine learning			
• Protecting health care systems from ransomware and other malware attacks • Managing health care big data challenges	• Technical solution	• Implement antimalware solutions using the dynamic method [10] • Implement a big data life cycle model using blockchain [80,97]	Adopting clusters to split the OCSVM[u] machine learning algorithm; however, the study does not offer a preventative solution [67]
Blockchain			
• How can we secure health information and personal identifiable information to enable privacy and security	• Technical solution	• Implement information-hiding algorithms using blockchain technology [80,97]	—
Secure design			

XSL•FO
RenderX

Vulnerability, knowledge application domain, and description of challenge or case type	Sociotechnical lens	Effective	Not effective
• Formidable medical device protection • Protecting health care ecosystems	• Sociotechnical solution	• Build in security from design planning and compliance [47,68,96] • Implement stakeholder collaborative design using sociotechnical behavior [65,96]	Security trade-off on safer medical devices for patients with diabetes; proposed improvement plans are not yet implemented [70]

[a]AIDE: Assess, Identify, Develop, and Evaluate.
[b]Not applicable.
[c]ISA: information security awareness.
[d]ANT: actor-network theory.
[e]CERT RMM: Computer Emergency Response Team Resilience Management Model.
[f]IoMT: Internet of Medical Things.
[g]HIPAA: Health Insurance Portability and Accountability Act.
[h]GDPR: General Data Protection Regulation.
[i]HITECH: Health Information Technology for Economic and Clinical Health.
[j]FDA: Food and Drug Administration.
[k]NIST-CSF: National Institute of Standards and Technology Cybersecurity Framework.
[l]HICP: Health Industry Cybersecurity Practices.
[m]ITPOSOM: information, technology, processes, objectivity and values, skills and knowledge, management systems and structure, and other resources.
[n]EHR: electronic health record.
[o]e-PSG: electronic health record–specific patient safety goals.
[p]IoT: Internet of Things.
[q]FHSS: frequency-hopping spread spectrum.
[r]RSSI: received signal strength indicator.
[s]BYOD: bring your own device.
[t]PHR: personal health record.
[u]OCSVM: one-class support vector machine.

Implications for Future Research

Overview

Health care sectors have improved with policies and measures developed to control health information breaches and vulnerabilities. However, further research is needed in social and technical interception design, namely, the human factor. Managing complex end-point devices and investment on addressing health care vulnerability and breaches should be considered from a sociotechnical design and sustainability perspective.

Protecting Complex Network-Connected End-Point Devices

The protection of complex network-connected end-point devices for health care organizations involves several key measures. The network of interconnected medical end-point devices and the software systems that connect to the internet are becoming vulnerable to attacks and breaches. This is a growing issue; health care organizations tend to procure medical device technology without proper equipment planning and guidelines in place. This implies that security is overlooked and is not a major focus area. Examples include hospital beds connected to >10 medical devices, such as pulse oximeters, syringe pumps, and patient care monitors, which are connected devices and vulnerable to attacks [2,6].

To address this technical challenge, organizations can concentrate on developing advanced threat detection and mitigation techniques, such as network defenders tailored to intricate network-connected end-point devices in health care and the integration of artificial intelligence using machine learning algorithms to effectively identify and respond to emerging threats. Furthermore, the health care industry must take a sociotechnical approach [96] toward implementing standard guidelines and technical solutions via the protection of health care networks through planning and integrating network security protection and segmentation. In addition, health information exchange over the network should undergo steganography and encryption as a solution using blockchain technology. Therefore, the integration of a complex end-point medical device should use built-in security with alert response and communication in processes to monitor health care cybersecurity ecosystems for a healthy security posture.

Health care organizations should collaborate with security experts and health care professionals and implement user education and incidence response to catalog cyber vulnerability incidences for further analysis. The implication is that, if networks and end-point medical devices are not properly secured, this will lead to breaches of health information through the network, which will cause patient information to be hijacked by cybercriminals for political gains. Sponsored state actors may use this weakness to seize networks and systems of care

XSL•FO
RenderX

delivery, demanding money from an organization before the latter can regain access. This approach will expose the health information of patients while they are receiving treatment and accessing health care services. This is an evolving challenge of the digital consequences of connected care. Building security through a design solution should be achieved from a sociotechnical approach as the human is the final user of systems of care.

Future research should focus on security by design before integrations of complex technology and design a simpler flow process with the disaggregation of complex network connections.

Increasing Investment in Cybersecurity

Investment in health care systems is critical to ensure the proper safeguarding of health care ecosystems from cyberattacks and vulnerabilities. To ensure efficient and secure health care, organizations should invest in human capital and technology to function effectively. An evaluation through research reveals that health care is lagging behind other sectors in terms of investment. This finding was confirmed by Kruse et al [17], who found that only 5% of health care investment is earmarked to protect health care, whereas a large percentage is allocated for health care delivery.

Insufficient investment in cybersecurity experts, awareness, and investment partnership plans will continue to subject health care employees to insecure behavior and result in a health care organization that is unprepared to mitigate cyber threats and other tactics used by attackers to disrupt evolving health care trends and patterns, particularly ransomware attacks.

Similarly, old legacy systems pose another security risk. Malicious actors can continue to exploit these systems to expose personal health information due to their limited capabilities and outdated organizational structure. Such vulnerability is worsened by a lack of investment in new cybersecurity infrastructure and computer devices to protect or process health information in a secure manner.

Health care organizations can engage in partnership with medical technology providers, application developers, and network solution integrators to develop strong systems and structures with seamless integration. Health care organizations should also develop and implement a framework for prioritizing cybersecurity investment based on risk assessments and threat intelligence. This approach can help identify the most critical areas of vulnerability within different departments, aiding organizations and policy makers in directing investments where they are most needed. Health care organizations should invest in humans and technology through training to ensure the development of necessary skills and investment in critical cyber infrastructure.

Awareness campaigns for patients and staff will help organizations recover from errors and breaches, whereas investment in technological security systems for health care will prepare health care organizations with the appropriate structure and system for resilience.

The findings presented in this paper are also highlighted in Table 4. Investment challenges in health care cybersecurity should focus on a sociotechnical approach that involves human behavior, technology, and organizational processes and should not be segregated as a separate technical or social problem. Future research should focus on security and investment in smart health care for attaining sustainability and resilience.

Managing Technological Advancement

Health care industries and organizations have improved over the years and are continuing to forge the development of new capabilities, technological advances, and processes to manage the multifaceted challenges of health care cybersecurity. Complexity in technology advancement and networks of digital systems increase the number of attack surfaces, where cybercriminals take advantage of the digital gateway access and execute malicious software programmed with code, such as malware to compromise digital technology and health care system networks. However, technological development necessitates a proactive approach to cybersecurity, particularly when considering security-by-design principles.

Future research projects must concentrate on important areas to protect networks, systems, and applications against vulnerabilities. Health care organizations should collaborate with medical device manufacturers as part of the planning phase of procurement requirements to ensure specifications needs before the development of medical devices technology for seamless integration. Implanted devices, for instance, should be built with security by design and continuously updated when necessary. A 2-factor authentication security for critical medical technology is also necessary. In addition, it is important that health care organizations quantify the risk, ensure that proper National Institute of Standards and Technology and GDPR standard guidelines are followed, and conduct threat modeling and simulation to evaluate the protectability of health care systems as a guideline in managing cybersecurity vulnerability.

Collaborative (sociotechnical) efforts among academia, industry, and policy makers are essential to drive this research agenda forward and create a safer digital landscape for the future.

The technology procurement requirement and collaboration should consider the integration of social and technical processes during digital technology development with health care delivery processes.

Health care organizations can adopt a blockchain technology solution for the protection of health information and other applications such as EHR systems from malicious use and insider threats.

Future research should examine the use of blockchain for health care big data protection and processes to manage cybersecurity vulnerability.

Containing Human Error in Cybersecurity

Humans are at the receiving end of the cyberattack chain. An example is the case of the WannaCry attack that affected 150,000 computers. It was attributed to human error because humans were warned of the attack on Windows server legacy systems but they ignored the warning by clicking on malicious

XSL•FO
RenderX

email links [38,43]. When an organization fails to train humans, cybercriminals take advantage of human weakness to exploit health care systems. Today, medical device manufacturers are building devices without considering humans as the final users or a participatory (sociotechnical) design approach. This is one factor of the clinical process and security dimension to protect critical infrastructure. Another factor is that, if a system is developed and does not start with security and support human usability, it becomes stressful for a human user to navigate the systems, which could cause them techno-stress, with the likelihood of mistakes. The health sector should use the Assess, Identify, Develop, and Evaluate technique to identify areas of human weakness, develop a new training method through simulations, and offer gamification training on issues such as phishing email deception and ransomware attacks. The implication is that, if humans are not trained, they will lead organizations to disaster because cybercriminals will continue to exploit the weakness of humans to cause more damage to health care systems. The consequences will include legal issues, fines, and possibly bankruptcy for health care organizations. Proper training and awareness campaigns should be implemented. Future research should focus on developing futuristic health care cybersecurity curriculums and training.

Practical Implications

Inadequate systems will cause health care systems and organizations to face increasing cyberattacks and setbacks in health information and patient safety. Moreover, a new trend reveals that, if implanted medical devices and technology are not protected, humans will be targeted by hackers seeking to make money or gain political power for ransom. However, implementation and adoption of the medical device security life cycle model [68] will protect medical devices, health information, patients, and organizations from harm and against future emerging threats. Thus, there is a need for the design of a cybersecurity sociotechnical framework toward sustaining smart health care systems.

Comparison With Prior Work

Previous narrative literature reviews by Coventry and Branley [6] and Mohan et al [31] highlight the need for an integrated approach in health care systems to address cybersecurity vulnerabilities. They emphasize the need for a comprehensive approach that connects human behavior, technology, and processes in a holistic way as a best strategy to tackle vulnerabilities, although the authors did not classify human behavior, technology, and processes from a sociotechnical lens. This systematic review supports their view by building and extending the literature on cybersecurity case challenge descriptions in all the tables in this paper to integrate human behavior, technology, and processes as a sociotechnical approach [2,23,26-28]. For example, an SLR conducted by Offner et al [2] reported that health care system vulnerability is a complex sociotechnical problem. Furthermore, for a health care organization to build resilience against cyberattacks and threats to avoid cybersecurity design gaps and vulnerabilities in the health care system, a strategic approach that integrates people, technology, and processes must be adopted [23,27,31]. The aforementioned aligns with the approach adopted in this study. Different schools of thought have highlighted the key importance of investment in technology and humans to protect health care systems from cyberattacks and threats [6,8,11,19,36,56]. This corroborates our findings that cybersecurity investment plays a main role in health care systems.

This study also revealed that complex network-connected end-point devices were mentioned several times by different schools of thought. Moreover, existing literature has opined that complex network-connected end-point devices were the most mentioned vulnerability [5,17,18,35,53].

Furthermore, technology advancement through a digital transformation evolution has created precision, and managed health care delivery [32,94]. However, more effort is still required in designing security features in health care technology. This study highlighted that security by design is required for medical device technology in health care systems [9,34,68].

Health care organizations must ensure that the design of technology evolves with a secure design approach from conception to avoid breaches of health information by external and internal attackers [24,32,68].

The sociotechnical solutions in Table 4 will aid health care organizations in being resilient in dealing with vulnerabilities and cybersecurity breaches in health care systems through a comprehensive and holistic approach. The sociotechnical perspective defines the meaning and constructs of technology, humans and processes [6,19,31,36,37]. This approach is promising and effective in dealing with health care system and cybersecurity vulnerabilities.

Limitations

For this study, non–English-language articles on cybersecurity and health care were not included. Closed-access articles directly related to cybersecurity and health care were also not included. Textbooks linked to cybersecurity and health care were excluded. In addition, as cybersecurity is a broad topic, more time was needed for data analysis.

Conclusions

This study conducted an SLR (PRISMA guidelines) to investigate the body of literature on cybersecurity in health care systems because of the exponential increase in health information breaches and vulnerability issues surrounding medical device technology and networks. This study also examined why health care systems are vulnerable to cyberattacks and threats.

In this review, sociotechnical solutions and mitigation strategies were proposed to protect patient health information, medical devices, and the critical cyber infrastructure of health care organizations from attacks and threats. We identified human error, lack of investment, complex network-connected end-point devices, old legacy systems, and technological advancement due to rapid digitalization as the causes of data breaches and the vulnerability of digital health care systems to attacks and threats. This study also revealed that research in the areas of education, awareness, training, collaborative partnerships, blockchain, and machine learning for health care cybersecurity

XSL•FO
RenderX

is underrepresented. In addition, there was inconsistency in the publication of intervention studies. There is a gap in intervention studies published between 2012 and 2013, as shown in this SLR, as well as breaks in research publications between 2012 and 2022, as illustrated in Table 3 and Figure 4.

As shown in Table 1, of the 70 papers published between 2012 and 2022 and reviewed in this study, only 8 (11%) carried out research in the areas of human error–related perspectives where health care systems are vulnerable to attacks. This finding clearly shows that considerably more studies are required on human factors. We also identified from this review that network-connected end-point devices are the most vulnerable challenge that causes health information breaches. However, stakeholders have rolled out interventions in the areas of health policy, health care system support (network security), and training. The support and training target operational activities and health care delivery while investment in cybersecurity critical infrastructure is disregarded. Rapid technology advancement has resulted to an increasing risk of cyberattacks and threats because most manufactured connected medical devices were not built with security in mind. With the possible sociotechnical solutions in Table 4, we form conclusions about how to protect health care systems as a sociotechnical solution in relation to the gap in research on technology, human behavior, and processes.

Health care organizations must concede that efficient and effective cybersecurity cannot be addressed with a technological process only but must also evolve beyond technological operation to a sociotechnical process that calls for a comprehensive knowledge of the human elements.

The profound implication of our findings steps further from just the concept of security. It deems it necessary for a major change in the approach to health care security by shifting from a reactive measure of patching and mitigation toward an approach of proactiveness and integration of detailed mechanisms that depend on complex sociotechnical dynamics at play in the design and development processes across the health care systems.

Our review emphasized the importance of a mandatory collaboration and cross-disciplinary engagement among stakeholders in health care, technology policy, and academia. The inclusion of a team-based effort from stakeholders will foster an integrated solution that responds to the challenges of cybersecurity vulnerabilities in health care systems.

In addition, our findings also give prominence to the great significance of investment in health care systems, such as in cybersecurity technology, medical devices, networks, health care professionals, and cybersecurity professionals, in advancing health care organizations. Furthermore, investment is imperative in cybersecurity education and training programs that will provide health care professionals and organizations with the updated knowledge and skills to navigate the complexities of cybersecurity vulnerabilities constructively. Governments should provide additional financial incentives for health care organizations to facilitate cybersecurity sustainability in health care systems. Future research should explore the application of blockchain technology for safeguarding health care system data. Blockchain offers a secure decentralized architecture. Therefore, system developers should consider a human-centric design approach when integrating blockchain technology into health care systems.

By strengthening awareness culture, intelligence information sharing, and accountability in health care systems, health care organizations can equip their operations and workforce to become active front-runners in safeguarding patient data and health care critical infrastructure and assuring the confidentiality, availability, and integrity of health care systems. Consequently, our SLR implores for an exhaustive procedure regarding cybersecurity in health care that affirms and entwines the sociotechnical nature of the vulnerabilities and challenges. By merging a technical approach with human-centric strategies, health care organizations can protect health care systems from vulnerabilities and cyber threats and advance a culture of resilience, trust, and innovation in health care service delivery. The implications of this review present a sociotechnical solution for establishing more secure and resilient health care ecosystems. This paper provides health care organizations with a better understanding of and resilience to cyberattacks, threats, and vulnerabilities.

Acknowledgments

The author is grateful to the Finnish Cultural Foundation and University of Vaasa in Finland for their support in funding this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist guide.
[PDF File (Adobe PDF File), 118 KB - jmir_v26i1e46904_app1.pdf]

Multimedia Appendix 2
Search strategy.

XSL•FO
RenderX

[DOCX File , 15 KB - jmir_v26i1e46904_app2.docx]

References

1. Cybersecurity in healthcare. Health Insurance Portability and Accountability Act. URL: https://www.himss.org/resources/cybersecurity-healthcare [accessed 2024-05-05]
2. Offner KL, Sitnikova E, Joiner K, MacIntyre CR. Towards understanding cybersecurity capability in Australian healthcare organisations: a systematic review of recent trends, threats and mitigation. Intell National Secur 2020 Apr 22;35(4):556-585 [FREE Full text] [doi: 10.1080/02684527.2020.1752459]
3. Frost. Medical Device and Network Security Coming to terms with the Internet of Medical Things (IoMT). - 2024:2019 [FREE Full text]
4. Karambelas C. Healthcare care technology: ransomware risk and protection. Am Bankruptcy Inst J 2020 May;39(5):30.
5. Giansanti D. Cybersecurity and the digital-health: the challenge of this millennium. Healthcare (Basel) 2021 Jan 11;9(1):62 [FREE Full text] [doi: 10.3390/healthcare9010062] [Medline: 33440612]
6. Coventry L, Branley D. Cybersecurity in healthcare: a narrative review of trends, threats and ways forward. Maturitas 2018 Jul;113:48-52 [FREE Full text] [doi: 10.1016/j.maturitas.2018.04.008] [Medline: 29903648]
7. Walker T. Interoperability a must for hospitals, but it comes with risks. Managed Healthcare Executive. 2017 Dec 10. URL: https://www.managedhealthcareexecutive.com/view/interoperability-must-hospitals-it-comes-risks [accessed 2024-05-06]
8. Dameff CJ, Selzer JA, Fisher J, Killeen JP, Tully JL. Clinical cybersecurity training through novel high-fidelity simulations. J Emerg Med 2019 Feb;56(2):233-238 [FREE Full text] [doi: 10.1016/j.jemermed.2018.10.029] [Medline: 30553562]
9. Klonoff DC. Cybersecurity for connected diabetes devices. J Diabetes Sci Technol 2015 Apr 16;9(5):1143-1147 [FREE Full text] [doi: 10.1177/1932296815583334] [Medline: 25883162]
10. Reshmi TR. Information security breaches due to ransomware attacks - a systematic literature review. Int J Inf Manag Data Insights 2021 Nov;1(2):100013 [FREE Full text] [doi: 10.1016/j.jjimei.2021.100013]
11. Argaw ST, Bempong NE, Eshaya-Chauvin B, Flahault A. The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review. BMC Med Inform Decis Mak 2019 Jan 11;19(1):10 [FREE Full text] [doi: 10.1186/s12911-018-0724-5] [Medline: 30634962]
12. Zarour M, Alenezi M, Ansari MT, Pandey AK, Ahmad M, Agrawal A, et al. Ensuring data integrity of healthcare information in the era of digital health. Healthc Technol Lett 2021 Jun;8(3):66-77 [FREE Full text] [doi: 10.1049/htl2.12008] [Medline: 34035927]
13. What are the penalties for HIPAA violations? The HIPAA Journal. URL: https://www.hipaajournal.com/what-are-the-penalties-for-hipaa-violations-7096/ [accessed 2024-05-06]
14. Mcnulty M, Kettani H. On cybersecurity education for non-technical learners. In: Proceedings of the 3rd International Conference on Information and Computer Technologies (ICICT). 2020 Presented at: ICICT 2020; March 9-12, 2020; San Jose, CA URL: https://doi.org/10.1109/ICICT50521.2020.00072 [doi: 10.1109/icict50521.2020.00072]
15. Ricci J, Breitinger F, Baggili I. Survey results on adults and cybersecurity education. Educ Inf Technol 2018 Jul 11;24(1):231-249 [FREE Full text] [doi: 10.1007/s10639-018-9765-8]
16. Sweeney E. Healthcare data breaches haven't slowed down in 2017, and insiders are mostly to blame. Fierce Healthcare. 2017 Aug 3. URL: https://tinyurl.com/yn2m49y8 [accessed 2024-05-06]
17. Kruse CS, Frederick B, Jacobson T, Monticone DK. Cybersecurity in healthcare: a systematic review of modern threats and trends. Technol Health Care 2017 Feb 21;25(1):1-10 [FREE Full text] [doi: 10.3233/thc-161263]
18. Wasserman L, Wasserman Y. Hospital cybersecurity risks and gaps: review (for the non-cyber professional). Front Digit Health 2022;4:862221 [FREE Full text] [doi: 10.3389/fdgth.2022.862221] [Medline: 36033634]
19. Jalali MS, Razak S, Gordon W, Perakslis E, Madnick S. Health care and cybersecurity: bibliometric analysis of the literature. J Med Internet Res 2019 Feb 15;21(2):e12644 [FREE Full text] [doi: 10.2196/12644] [Medline: 30767908]
20. Healthcare data breach statistics. The HIPAA Journal. URL: https://www.hipaajournal.com/healthcare-data-breach-statistics/#:~:text=2021%20was%20a%20bad%20year,stolen%2C%20or%20otherwise%20impermissibly%20disclosed [accessed 2024-05-06]
21. IBM report: cost of a data breach hits record high during pandemic. IBM. 2021 Jul 28. URL: https://tinyurl.com/euc26j9y [accessed 2024-05-06]
22. Nifakos S, Chandramouli K, Nikolaou CK, Papachristou P, Koch S, Panaousis E, et al. Influence of human factors on cyber security within healthcare organisations: a systematic review. Sensors (Basel) 2021 Jul 28;21(15):5119 [FREE Full text] [doi: 10.3390/s21155119] [Medline: 34372354]
23. Heeks R. Health information systems: failure, success and improvisation. Int J Med Inform 2006 Feb;75(2):125-137 [FREE Full text] [doi: 10.1016/j.ijmedinf.2005.07.024] [Medline: 16112893]
24. Casola V, De Benedictis A, Rak M, Villano U. Security-by-design in multi-cloud applications: an optimization approach. Inf Sci 2018 Jul;454-455:344-362 [FREE Full text] [doi: 10.1016/j.ins.2018.04.081]
25. Secure-by-design: shifting the balance of cybersecurity risk: principles and approaches for secure by design software. Cybersecurity and Infrastructure Security Agency. 2023 Oct 25. URL: https://www.cisa.gov/resources-tools/resources/secure-by-design [accessed 2024-05-06]

XSL•FO
RenderX

26. Mumford E. The story of socio - technical design: reflections on its successes, failures and potential. Inf Syst J 2006 Sep 04;16(4):317-342 [FREE Full text] [doi: 10.1111/j.1365-2575.2006.00221.x]
27. Palvia SC, Sharma RS, Conrath DW. A socio-technical framework for quality assessment of computer information systems. Ind Manag Data Syst 2001;101(5):237-251 [FREE Full text] [doi: 10.1108/02635570110394635]
28. Atkinson C, Eldabi T, Paul RJ, Pouloudi A. Investigating integrated socio-technical approaches to health informatics. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences. 2001 Presented at: HICSS 2001; January 6, 2001; Maui, HI URL: https://doi.org/10.1109/HICSS.2001.926578 [doi: 10.1109/hicss.2001.926578]
29. Altman R. Informatics in the care of patients: ten notable challenges. West J Med 1997 Feb;166(2):118-122 [FREE Full text] [Medline: 9109328]
30. Coiera E. Four rules for the reinvention of health care. BMJ 2004 May 15;328(7449):1197-1199 [FREE Full text] [doi: 10.1136/bmj.328.7449.1197] [Medline: 15142933]
31. Mohan DN, Gowda SS, Vikyath IS. Cyber security in health care. Int J Res Eng Sci Manag 2020;3(1):551-553. [doi: 10.47607/ijresm]
32. Luna R, Rhine E, Myhra M, Sullivan R, Kruse CS. Cyber threats to health information systems: a systematic review. Technol Health Care 2016 Jan 27;24(1):1-9 [FREE Full text] [doi: 10.3233/thc-151102]
33. Jalali MS, Russell B, Razak S, Gordon WJ. EARS to cyber incidents in health care. J Am Med Inform Assoc 2019 Jan 01;26(1):81-90 [FREE Full text] [doi: 10.1093/jamia/ocy148] [Medline: 30517701]
34. Williams P, Woodward A. Cybersecurity vulnerabilities in medical devices: a complex environment and multifaceted problem. Med Devices Evid Res 2015 Jul;8:305-316 [FREE Full text] [doi: 10.2147/mder.s50048]
35. Safavi S, Meer AM, Melanie EK, Shukur Z. Cyber vulnerabilities on smart healthcare, review and solutions. In: Proceedings of the Cyber Resilience Conference (CRC). 2018 Presented at: CR 2018; November 13-15, 2018; Putrajaya, Malaysia URL: https://doi.org/10.1109/CR.2018.8626826 [doi: 10.1109/cr.2018.8626826]
36. He Y, Aliyu A, Evans M, Luo C. Health care cybersecurity challenges and solutions under the climate of COVID-19: scoping review. J Med Internet Res 2021 Apr 20;23(4):e21747 [FREE Full text] [doi: 10.2196/21747] [Medline: 33764885]
37. Jalali MS, Kaiser JP. Cybersecurity in hospitals: a systematic, organizational perspective. J Med Internet Res 2018 May 28;20(5):e10059 [FREE Full text] [doi: 10.2196/10059] [Medline: 29807882]
38. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Arndt RZ. For epic, interoperability comes from within. Modern Healthcare. 2018 Jan 29. URL: https://www.modernhealthcare.com/article/20180130/NEWS/180139993/for-epic-interoperability-comes-from-within [accessed 2024-05-06]
40. Incident detection, email attacks continue to cause headaches for companies. Twitter. URL: https://twitter.com/SandraProske/status/967893399599796224 [accessed 2024-05-06]
41. Mukherjee SY. Anthem's historic 2015 health records breach was likely ordered by a foreign government. Fortune. 2017 Jan 10. URL: https://fortune.com/2017/01/09/anthem-cyber-attack-foreign-government/ [accessed 2024-05-06]
42. 2017 cost of data breach study: United States. Ponemon Institute. 2017 Jun 13. URL: https://www.ponemon.org/news-updates/blog/security/2017-cost-of-data-breach-study-united-states.html [accessed 2024-05-06]
43. Cost of a data breach report 2021. IBM Security. URL: https://info.techdata.com/rs/946-OMQ-360/images/Cost_of_a_Data_Breach_Report_2021.PDF [accessed 2024-05-06]
44. Scott M, Wingfield N. Hacking attack has security experts scrambling to contain fallout. New York Times. 2017 May 13. URL: https://tinyurl.com/4weatd6e [accessed 2024-05-06]
45. Gordon WJ, Wright A, Glynn RJ, Kadakia J, Mazzone C, Leinbach E, et al. Evaluation of a mandatory phishing training program for high-risk employees at a US healthcare system. J Am Med Inform Assoc 2019 Jun 01;26(6):547-552 [FREE Full text] [doi: 10.1093/jamia/ocz005] [Medline: 30861069]
46. Bouveret A. Cyber risk for the financial sector: a framework for quantitative assessment. SSRN Preprint posted online July 16, 2018 [FREE Full text] [doi: 10.2139/ssrn.3203026]
47. Top 10 health technology hazards for 2016. ECRI Institute. 2015 Nov. URL: https://www.ecri.org/Resources/Whitepapers_and_reports/2016_Top_10_Hazards_Executive_Brief.pdf [accessed 2024-05-06]
48. Faruki P, Bharmal A, Laxmi V, Ganmoor V, Singh Gaur M, Conti M, et al. Android security: a survey of issues, malware penetration, and defenses. IEEE Commun Surv Tutorials 2015;17(2):998-1022 [FREE Full text] [doi: 10.1109/comst.2014.2386139]
49. Filkins B. Health care cyberthreat report: widespread compromises detected, compliance nightmare on horizon. SANS Institute. 2014. URL: https://asprtracie.hhs.gov/technical-resources/resource/3381/health-care-cyberthreat-report-widespread-compromises-detected-compliance-nightmare-on-horizon [accessed 2024-05-06]
50. Fu K, Blum J. Controlling for cybersecurity risks of medical device software. Commun ACM 2013 Oct;56(10):35-37 [FREE Full text] [doi: 10.1145/2508701]
51. McHugh M. Medical device software and technology: the past, present and future. BEAI Spectrum, Biological and Clinical Engineers Association of Ireland. 2015. URL: https://arrow.tudublin.ie/scschcomart/38/ [accessed 2024-05-06]

XSL•FO
RenderX

52. Newman LH. Medical devices are the next security nightmare. WIRED. 2017 Mar 2. URL: https://www.wired.com/2017/03/medical-devices-next-security-nightmare/ [accessed 2024-05-06]
53. Tully J, Selzer J, Phillips JP, O'Connor P, Dameff C. Healthcare challenges in the era of cybersecurity. Health Secur 2020;18(3):228-231 [FREE Full text] [doi: 10.1089/hs.2019.0123] [Medline: 32559153]
54. The state of ransomware in the US: report and statistics 2019. Emsisoft Malware Lab. 2019 Dec 12. URL: https://tinyurl.com/ykx5zjce [accessed 2024-05-06]
55. The state of ransomware in the US: report and statistics 2020. Emsisoft Malware Lab. 2021 Jan 18. URL: https://www.emsisoft.com/en/blog/37314/the-state-of-ransomware-in-the-us-report-and-statistics-2020/ [accessed 2024-05-06]
56. Branley-Bell D, Coventry L, Sillence E, Magalini S, Mari P, Magkanaraki A, et al. Your hospital needs you: eliciting positive cybersecurity behaviours from healthcare staff. Ann Disaster Risk Sci 2020;3(1) [FREE Full text] [doi: 10.51381/adrs.v3i1.51]
57. Information Commissioner's Office, National Cyber Security Centre, James M. New figures show large numbers of businesses and charities suffer at least one cyber attack in the past year. United Kingdom Government. 2018 Apr 25. URL: https://www.gov.uk/government/news/new-figures-show-large-numbers-of-businesses-and-charities-suffer-at-least-one-cyber-attack-in-the-past-year [accessed 2024-05-06]
58. IT security in the era when everything can be hacked. Kaspersky Lab. URL: https://www.unodc.org/documents/organized-crime/cybercrime/cybercrime-april-2018/RUSSIAN_FED.pdf [accessed 2024-05-06]
59. GDPR: getting ready for the new EU data protection regulation. PECB Insights. 2018 Apr 27. URL: https://insights.pecb.com/gdpr-compliance-getting-ready/ [accessed 2024-05-06]
60. Rahman NA, Sairi IH, Zizi NA, Khalid F. The importance of cybersecurity education in school. Int J Inf Educ Technol 2020;10(5):378-382 [FREE Full text] [doi: 10.18178/ijiet.2020.10.5.1393]
61. Chowdhury N, Gkioulos V. Cyber security training for critical infrastructure protection: a literature review. Comput Sci Rev 2021 May;40:100361 [FREE Full text] [doi: 10.1016/j.cosrev.2021.100361]
62. Coventry L, Branley-Bell D, Sillence E, Magalini S, Mari P, Magkanaraki A, et al. Cyber-risk in healthcare: exploring facilitators and barriers to secure behaviour. In: Proceedings of the HCI for Cybersecurity, Privacy and Trust 2020. 2020 Presented at: HCI-CPT 2020; July 19-24, 2020; Copenhagen, Denmark URL: https://tinyurl.com/v2dbyrsx [doi: 10.1007/978-3-030-50309-3_8]
63. Burns AJ, Johnson ME, Honeyman P. A brief chronology of medical device security. Commun ACM 2016 Sep 22;59(10):66-72 [FREE Full text] [doi: 10.1145/2890488]
64. Chua JA. Cybersecurity in the healthcare industry - A collaborative approach. American Association for Physician Leadership. 2021 Jan 8. URL: https://www.physicianleaders.org/articles/cybersecurity-healthcare-industry-collaborative-approach [accessed 2024-05-06]
65. Djenna A, Harous S, Saidouni DE. Internet of things meet internet of threats: new concern cyber security issues of critical cyber infrastructure. Appl Sci 2021 May 17;11(10):4580 [FREE Full text] [doi: 10.3390/app11104580]
66. Francis R. Medical devices that could put you at security risk. IDG Communications. 2017 Apr 27. URL: https://www.csoonline.com/article/561347/medical-devices-that-could-put-you-at-security-risk.html [accessed 2024-05-06]
67. Handa A, Sharma A, Shukla SK. Machine learning in cybersecurity: a review. WIREs Data Min Knowl 2019 Feb 17;9(4):e1306 [FREE Full text] [doi: 10.1002/widm.1306]
68. Lechner NH. Developing a compliant cybersecurity process for medical devices. In: Proceedings of the Central European Conference on Information and Intelligent Systems. 2018 Presented at: CECIIS 2018; September 19-21, 2018; Varaždin, Croatia URL: https://www.proquest.com/openview/8a2a254a80f34ef64b55c71d5bac01d6/1?pq-origsite=gscholar&cbl=1986354
69. Lewis CJ. Cybersecurity in healthcare. Utica College. 2014. URL: https://tinyurl.com/3usz5jat [accessed 2024-05-16]
70. Lyon D. Making trade-offs for safe, effective, and secure patient care. J Diabetes Sci Technol 2017 Mar;11(2):213-215 [FREE Full text] [doi: 10.1177/1932296816676281] [Medline: 28264187]
71. Mohan A. Cyber security for personal medical devices internet of things. In: Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems. 2014 Presented at: DCOSS 2014; May 26-28, 2014; Marina Del Rey, CA URL: https://doi.org/10.1109/DCOSS.2014.49 [doi: 10.1109/dcoss.2014.49]
72. Baranchuk A, Refaat MM, Patton KK, Chung MK, Krishnan K, Kutyifa V, et al. Cybersecurity for cardiac implantable electronic devices: what should you know? J Am Coll Cardiol 2018 Mar 20;71(11):1284-1288 [FREE Full text] [doi: 10.1016/j.jacc.2018.01.023] [Medline: 29475627]
73. Perakslis E. Cybersecurity in health care. N Engl J Med 2014 Jul 31;371(5):395-397 [FREE Full text] [doi: 10.1056/nejmp1404358]
74. Peterson A. Yes, terrorists could have hacked Dick Cheney's heart. The Washington Post. 2013 Oct 21. URL: https://www.washingtonpost.com/news/the-switch/wp/2013/10/21/yes-terrorists-could-have-hacked-dick-cheneys-heart/ [accessed 2024-05-06]
75. Sajedi H, Rahbar Yaghobi S. Information hiding methods for E-Healthcare. Smart Health 2020 Mar;15:100104 [FREE Full text] [doi: 10.1016/j.smhl.2019.100104]

XSL•FO
RenderX

76. Omotosho A, Adegbola O, Mikail OO, Emuoyibofarhe J. A secure electronic prescription system using steganography with encryption key implementation. Int J Comput Inform Technol 2014 Sep;03(5):980-986. [doi: 10.48550/arXiv.1502.01264]
77. Singh Rayat A, Singh I, Singh K. Review on security challenges of data communication in IoT devices. Int J Electron Eng 2019;11(2):406-415 [FREE Full text]
78. Sittig DF, Singh H. Electronic health records and national patient-safety goals. N Engl J Med 2012 Nov 08;367(19):1854-1860 [FREE Full text] [doi: 10.1056/nejmsb1205420]
79. Snell E. Healthcare data breach costs highest for 7th straight year. Health IT Security. 2017. URL: https://healthitsecurity.com/news/healthcare-data-breach-costs-highestfor-7th-straight-year [accessed 2024-05-06]
80. Bhuyan SS, Kabir UY, Escareno JM, Ector K, Palakodeti S, Wyant D, et al. Transforming healthcare cybersecurity from reactive to proactive: current status and future recommendations. J Med Syst 2020 Apr 02;44(5):98 [FREE Full text] [doi: 10.1007/s10916-019-1507-y] [Medline: 32239357]
81. Raina MacIntyre C, Engells TE, Scotch M, Heslop DJ, Gumel AB, Poste G, et al. Converging and emerging threats to health security. Environ Syst Decis 2018;38(2):198-207 [FREE Full text] [doi: 10.1007/s10669-017-9667-0] [Medline: 32288980]
82. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res 2016;8(3):1560-1580 [FREE Full text] [Medline: 27186282]
83. Rodrigues JJ, de la Torre I, Fernández G, López-Coronado M. Analysis of the security and privacy requirements of cloud-based electronic health records systems. J Med Internet Res 2013 Aug 21;15(8):e186 [FREE Full text] [doi: 10.2196/jmir.2494] [Medline: 23965254]
84. -. Verizon: 2019 data breach investigations report. Comput Fraud Secur 2019 Jan;2019(6). [doi: 10.1016/S1361-3723(19)30060-0]
85. 2022 cost of insider threats global report. Ponemon Institute. 2022. URL: https://protectera.com.au/wp-content/uploads/2022/03/The-Cost-of-Insider-Threats-2022-Global-Report.pdf [accessed 2024-05-06]
86. Wagner S. The medical data of hundreds of HUG patients accessible on the internet. Ictjournal. 2019. URL: https://www.ictjournal.ch/news/2019-10-04/les-donnees-medicales-dune-centaines-de-patients-des-hug-accessibles-sur-internet [accessed 2019-10-04]
87. Arapi K. The healthcare industry: evolving cyber threats and risks. Utica College. 2018 May. URL: https://www.proquest.com/openview/6fb8d8f9984e83b682b5499fb1d36194/1?pq-origsite=gscholar&cbl=18750 [accessed 2024-05-06]
88. Perriello B. ‘Medjack:’ hackers threaten hospitals using medical devices as back doors. MassDevice. 2015 Jun 5. URL: https://www.massdevice.com/medjack-hackers-threaten-hospitals-using-medical-devices-as-back-doors/ [accessed 2024-05-06]
89. Meggitt S. MEDJACK attacks: the scariest part of the hospital. Tufts University. 2018 Dec 18. URL: https://www.cs.tufts.edu/comp/116/archive/fall2018/smeggitt.pdf [accessed 2024-05-06]
90. Rajamäki J, Pirinen R. Towards the cyber security paradigm of ehealth: resilience and design aspects. AIP Conf Proc 2017 Jun 5;1836(1) [FREE Full text] [doi: 10.1063/1.4981969]
91. Murphy S. Is cybersecurity possible in healthcare? National Cybersecur Institute J 2015;1(3):49 [FREE Full text]
92. Kioskli K, Fotis T, Mouratidis H. The landscape of cybersecurity vulnerabilities and challenges in healthcare: security standards and paradigm shift recommendations. In: Proceedings of the 16th International Conference on Availability, Reliability and Security. 2021 Presented at: ARES '21; August 17-20, 2021; Vienna, Austria URL: https://doi.org/10.1145/3465481.3470033 [doi: 10.1145/3465481.3470033]
93. Tuttle I. Cyberdisaster: how the government compromised our security. National Review. 2016 Sep 9. URL: https://www.nationalreview.com/2016/09/opm-hack-house-oversight-committee-report/ [accessed 2024-05-08]
94. Loi M, Christen M, Kleine N, Weber K. Cybersecurity in health – disentangling value tensions. J Inform Commun Ethics Soc 2019 May 13;17(2):229-245 [FREE Full text] [doi: 10.1108/jices-12-2018-0095]
95. Christen M, Gordijn B, Loi M. The ethics of cybersecurity. CrimRxiv. 2020. URL: https://www.crimrxiv.com/pub/s79bo1xu/release/1 [accessed 2024-05-06]
96. Coles-Kemp L, Williams PA. Changing places: the need to alter the start point for information security design. Electron J Health Inform 2014;8(2) [FREE Full text]
97. Khaloufi H, Abouelmehdi K, Beni-hssane A, Saadi M. Security model for big healthcare data lifecycle. Procedia Comput Sci 2018;141:294-301 [FREE Full text] [doi: 10.1016/j.procs.2018.10.199]
98. Holst C, Sukums F, Radovanovic D, Ngowi B, Noll J, Winkler AS. Sub-Saharan Africa—the new breeding ground for global digital health. Lancet Digit Health 2020 Apr;2(4):e160-e162 [FREE Full text] [doi: 10.1016/s2589-7500(20)30027-3]
99. Khando K, Gao S, Islam SM, Salman A. Enhancing employees information security awareness in private and public organisations: a systematic literature review. Comput Secur 2021 Jul;106:102267 [FREE Full text] [doi: 10.1016/j.cose.2021.102267]
100. Winton R. Hollywood hospital pays $17,000 in bitcoin to hackers; FBI investigating. Los Angeles Times. 2016 Feb 18. URL: https://tinyurl.com/5yae788s [accessed 2024-05-06]

XSL•FO
RenderX

101. Dobuzinskis A, Finkle J. California hospital makes rare admission of hack, ransom payment. Reuters. 2016 Feb 20. URL: https://www.reuters.com/article/idUSKCN0VS05M/ [accessed 2024-05-06]
102. Bickers S, Dunlevy S, Minear T. Hackers are offering to sell the medicare details of Australians on the dark web, government confirms. News Corp Australia Network. 2017 Jul 4. URL: https://tinyurl.com/4ryf66v8 [accessed 2024-05-06]
103. Zorabedian J. How malware works: anatomy of drive-by download web attack. Sophos News. 2014 Mar 26. URL: https://news.sophos.com/en-us/2014/03/26/how-malware-works-anatomy-of-a-drive-by-download-web-attack-infographic/ [accessed 2024-05-06]
104. Omotosho A, Asanga U, Fakorede A. Electronic prescription system for pediatricians. Eur Sci J 2017;13(18):426 [FREE Full text] [doi: 10.19044/esj.2017.v13n18p426]
105. Chen B, Ren Z, Yu C, Hussain I, Liu J. Adversarial examples for CNN-based malware detectors. IEEE Access 2019;7:54360-54371 [FREE Full text] [doi: 10.1109/access.2019.2913439]

Abbreviations

EHR: electronic health record
GDPR: General Data Protection Regulation
IoMT: Internet of Medical Things
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question
SLR: systematic literature review

Edited by A Mavragani; submitted 03.03.23; peer-reviewed by R Marshall, V Perez Jover; comments to author 27.07.23; revised version received 17.10.23; accepted 08.03.24; published 31.05.24.

Please cite as:
Ewoh P, Vartiainen T
Vulnerability to Cyberattacks and Sociotechnical Solutions for Health Care Systems: Systematic Review
J Med Internet Res 2024;26:e46904
URL: https://www.jmir.org/2024/1/e46904
doi:10.2196/46904
PMID:38820579

©Pius Ewoh, Tero Vartiainen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Use of Patient-Generated Health Data From Consumer-Grade Devices by Health Care Professionals in the Clinic: Systematic Review

Sharon Guardado[1], MSc; Maria Karampela[1], PhD; Minna Isomursu[1], PhD; Casandra Grundstrom[2], PhD

[1]Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland

[2]Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

Corresponding Author:
Sharon Guardado, MSc
Faculty of Information Technology and Electrical Engineering
University of Oulu
Pentti Kaiteran katu 1
Oulu, 90570
Finland
Phone: 358 504388396
Email: sharon.guardadomedina@oulu.fi

Abstract

Background: Mobile health (mHealth) uses mobile technologies to promote wellness and help disease management. Although mHealth solutions used in the clinical setting have typically been medical-grade devices, passive and active sensing capabilities of consumer-grade devices like smartphones and activity trackers have the potential to bridge information gaps regarding patients' behaviors, environment, lifestyle, and other ubiquitous data. Individuals are increasingly adopting mHealth solutions, which facilitate the collection of patient-generated health data (PGHD). Health care professionals (HCPs) could potentially use these data to support care of chronic conditions. However, there is limited research on real-life experiences of HPCs using PGHD from consumer-grade mHealth solutions in the clinical context.

Objective: This systematic review aims to analyze existing literature to identify how HCPs have used PGHD from consumer-grade mobile devices in the clinical setting. The objectives are to determine the types of PGHD used by HCPs, in which health conditions they use them, and to understand the motivations behind their willingness to use them.

Methods: A systematic literature review was the main research method to synthesize prior research. Eligible studies were identified through comprehensive searches in health, biomedicine, and computer science databases, and a complementary hand search was performed. The search strategy was constructed iteratively based on key topics related to PGHD, HCPs, and mobile technologies. The screening process involved 2 stages. Data extraction was performed using a predefined form. The extracted data were summarized using a combination of descriptive and narrative syntheses.

Results: The review included 16 studies. The studies spanned from 2015 to 2021, with a majority published in 2019 or later. Studies showed that HCPs have been reviewing PGHD through various channels, including solutions portals and patients' devices. PGHD about patients' behavior seem particularly useful for HCPs. Our findings suggest that PGHD are more commonly used by HCPs to treat conditions related to lifestyle, such as diabetes and obesity. Physicians were the most frequently reported users of PGHD, participating in more than 80% of the studies.

Conclusions: PGHD collection through mHealth solutions has proven beneficial for patients and can also support HCPs. PGHD have been particularly useful to treat conditions related to lifestyle, such as diabetes, cardiovascular diseases, and obesity, or in domains with high levels of uncertainty, such as infertility. Integrating PGHD into clinical care poses challenges related to privacy and accessibility. Some HCPs have identified that though PGHD from consumer devices might not be perfect or completely accurate, their perceived clinical value outweighs the alternative of having no data. Despite their perceived value, our findings reveal their use in clinical practice is still scarce.

International Registered Report Identifier (IRRID): RR2-10.2196/39389

(J Med Internet Res 2024;26:e49320) doi:10.2196/49320

XSL•FO
RenderX

KEYWORDS

patient-generated health data; mHealth; health care professionals; mobile technologies; self-management

Introduction

Background

The term "mobile health" (mHealth) has been in use for nearly 2 decades to refer to the application of mobile technologies in delivering health services and collecting data pertinent to disease diagnosis, prevention, and management [1,2]. In the last decade, the scope of mHealth has expanded to include consumer-grade devices, such as smartphones, wearable, sensors, and other quasi-medical devices, while it increasingly targets specific health conditions, in addition to wellness [2,3]. Whereas medical-grade mobile devices require clinical evidence for certification, often requiring years to bring a device to the market [4], consumer-grade mobile devices evolving at a rapid pace, and open numerous possibilities through their capacity for ubiquitous data collection [5]. mHealth solutions have become integral to many people's lives, serving as tools for tracking health and well-being. Research has found that mHealth solutions can benefit individuals in general by fostering moderate increases in physical activity [6] or by being a convenient tool for self-management of health issues [7]. For individuals with chronic diseases, mHealth solutions have been particularly effective in offering support for condition management, goal setting, and enhancing overall satisfaction [7,8]. In addition to supporting people's efforts to manage their health, mHealth solutions also enable the collection of electronic patient-generated health data (PGHD), which can be used in the clinical context. PGHD refer to health-related data created, recorded, and gathered by and from patients outside of the clinical settings [9,10]. PGHD encompasses a broad range of data types from both passive and active sensing [1,11]. Passive data collection usually involves sensors that are connected to a mobile device that may be worn or embedded, limiting the patient's participation to wearing, carrying, or activating the device [12]. Active data collection requires patients to manually enter information or interact with an external device such as a peak flowmeter, glucometer, or thermometer to generate information. These data are "patient-generated" since the patient has actively participated in collecting and recording [12]. It has been hypothesized that through both passively and actively collected PGHD, health care professionals (HCPs) could gain insights into patients' activities, lifestyle, and physical condition to inform care decisions and personalize care approaches [13].

In countries with medium or high levels of digitalization, more than 56% of people appear willing to share their personal health data, even if the purpose of sharing them is not directly related to the improvement of their health [14]. Similarly, 46.3% of individuals who owned a wearable medical device indicated having shared data with a health provider in 2019 [15]. With mHealth solutions becoming increasingly accessible, it can be expected that more people may be interested in sharing their health data with HCPs if they believe that it could help them improve health care. However, a recent study found that although providers of mHealth solutions for chronic condition self-management encourage data sharing with HCPs, few solutions are designed to facilitate HCPs' review of these data [4]. This issue, in combination with already known challenges such as interoperability, data privacy issues, data validity, and the added burden of reviewing [9,16], makes the use of PGHD in the clinic an unrealistic possibility for many HCPs.

HCPs might have different approaches and goals when deciding to ponder PGHD collected through nonmedical mobile devices. According to Nittas et al [17], when integrating PGHD into the care process, HCPs can take the supporter or the reviewer role. In the supporter role, they limit themselves to motivating patients to use mHealth, whereas in the reviewer role, HCPs assess PGHD to complement medical data. Taking the reviewer role implies an active stance, and though some might value PGHD's contribution to care, this type of data may still be a new and unfamiliar source of information for some HCPs [18]. For PGHD for mobile devices to be feasible as a complementary tool in the clinical setting, their use should benefit both patients and HCPs. Though the adoption of mHealth solutions by patients supports their well-being and enables the availability of PGHD, such availability does not automatically equate to usefulness for HCPs. Despite the acceptance and adoption of mHealth solutions by HCPs being one of the most influential factors regarding the success of those solutions [19,20], there has not been significant research on the role HCPs are expected to take in the use of mHealth solutions [4,17] or on the concrete experiences and motivators of those willing to review PGHD.

Objectives

The main objective of our review is to systematically analyze existing scientific literature to identify what types of PGHD and in what health conditions HCPs have been using PGHD from consumer-grade mobile devices, as well as further context information for their motivations to use these types of data as a complementary tool in the clinic.

To attain these objectives the proposed research questions for our review are as follows: (1) In what health conditions have PGHD from consumer-grade mobile been a suitable tool for HCP? (2) What types of PGHD have HCPs found useful in the care of chronic conditions? (3) What are the main motivations behind HCPs' decision to review PGHD from consumer-grade devices?

Methods

Study Design

A systematic literature review (SLR) was selected as our main research method to comprehensively synthesize evidence and prior research on HCPs' experiences reviewing PGHD from consumer-grade mobile devices to address our research questions. We wanted to follow a transparent and systematic method to inform further research on this topic. We adopted methodologies from the Guidelines for Performing Systematic Literature Reviews in Software Engineering [21] and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement) [22,23], both of which provide

XSL•FO
RenderX

reliable methodologies to perform SLRs in the fields of computer science and medicine, respectively (Multimedia Appendix 1 [24]). We deemed it pertinent to combine methodological traditions from both computer science and medicine, as our research topic combines technical and care viewpoints and is interdisciplinary by nature [24].

To perform this SLR, we adhered to a systematic review protocol that was prepared before starting the searches and screening process. The protocol has been published elsewhere [25] and provides an ample description of the methods used in the search strategy and the inclusion and exclusion criteria used. The review adhered closely to the original protocol, with no significant deviations.

Search Strategy

Eligible studies were identified through comprehensive literature searches we conducted in bibliographical databases on health and biomedicine and information technology domains. The searched databases included PubMed, ACM Digital Library (including the ACM Guide to Computing Literature), IEEE Xplore, and Scopus. The searches were carried out in May 2022.

To ensure the identification of relevant papers, we constructed the search strategy in an iterative way [25]. The search string used for each database is available in Multimedia Appendix 2. Based on the specific objectives of our review, and after conducting a pilot search in PubMed, we determined that the literature search should be constructed around 3 specific key topics: "patient-generated health data," "health personnel," and "mobile technologies." We used the corresponding Medical Subject Headings (MeSH) and their possible variants to construct the final search query. Once the query had been tested, it was validated by a research librarian from the University of Oulu. After completing the electronic searches, we performed a supplementary hand search of the citations found within other SLRs and scoping reviews that were retrieved during the literature searches.

Eligibility Criteria

The defined eligibility criteria (Table 1) aimed to include original papers that reported on the use of PGHD created via consumer-grade mHealth solutions by HCPs. PGHD reported in the studies should have been collected outside of the clinical setting, through either the patients' use of mobile health apps or the wearable devices such as smartwatches, smart rings, fitness trackers, and similar wearable trackers; studies reporting on PGHD collected by HCPs during appointments or inside the clinical settings were excluded. Studies were limited to those involving consumer-grade devices to focus on, excluding those solely focusing on PGHD from medical-grade devices. The included papers report on the experiences of HCPs who have experience using PGHD in their clinical practice, as part of a stand-alone mHealth solution, by personal initiative, or for any other reasons. We excluded papers that focus solely on the perceptions or perspectives of HCPs as potential users of PGHD. Eligible publications were restricted to those accepted in peer-reviewed journals and conference proceedings written in the English language.

Table 1. Eligibility criteria.

Criteria	Inclusion	Exclusion
PGHD[a] collection context	Outside of the clinic	In the clinical setting or during appointments with HCPs[b]
Type of mobile technologies for PGHD collection	Mobile apps and consumer-grade wearable devices	Medical-grade tracking or measuring devices
Use of PGHD by HCPs	HCPs had previously reviewed PGHD from mobile devices at the time of the study	HCPs had no previous experience reviewing PGHD from mobile devices before or during the study
Type of publication	Peer-reviewed journals and conference proceedings papers	Books, opinions, editorials, and other non–peer-reviewed sources
Language	English	Other than English

[a]PGHD: patient-generated health data.

[b]HCP: health care professional.

Though current consumer-grade mobile and wearable technologies started to become more accessible in the first half of the last decade, their impact on the health care scene started to become evident only years later. In 2013, it was acknowledged that only a few studies had assessed the impact of mobile apps in the health context, and all those studies referred to apps that had been created only for research purposes and were not available to the public at that time [26]. Therefore, we limited our search to papers with publication dates starting in 2013. Our search criteria did not delimit aspects such as the medical profile or specialties of the HCPs participating in the studies, or the health conditions treated, as we aimed to ascertain whether PGHD usage by HCPs would be more prevalent in the treatment of certain medical conditions or certain medical fields.

Selection Process

After the electronic search, the resulting papers were imported into Covidence (Veritas Health Innovation) for screening. The screening process was divided into 2 stages carried out independently by 2 researchers (SG and MK) with computer science backgrounds and previous research experience with mHealth and PGHD. Initially, the screening was limited to titles and abstracts. Before starting this stage, the reviewers completed a joint exercise to validate the review methodology and ensure that the inclusion and exclusion criteria were correctly

XSL•FO
RenderX

understood. The disagreements that arose during the initial stage were all discussed and resolved between the 2 reviewers before starting the second screening stage. The second screening round included the review of the full text of all the preliminarily included papers.

Data Extraction

The relevant information of the included papers was collected using a structured data extraction form constructed in Covidence (Multimedia Appendix 3). The most relevant data extracted for each paper included the professions of the participants; health conditions treated; mobile technologies used; the type of PGHD collected; and the channels used for visualization. In addition, to understand what motivated HCPs to review PGHD quotes related to their motivations and conclusions related to the topic were extracted from each study.

The data extraction task was completed by the 2 original reviewers (SG and MK) and 2 additional reviewers (CG and MI), all of whom have previous research experience with the topic of this review. Each paper was randomly assigned to be examined by 2 of the reviewers. Each reviewer performed the data extraction independently. Upon completion, the data extracted by both reviewers were compared. Discrepancies were resolved through discussion between the reviewers and a final consensus was reached in all cases.

Data Analysis and Synthesis

Quantitative and qualitative studies were included in this review. Due to the significant heterogeneity observed in the studies' design, types of health conditions, types of PGHD, and types of mHealth solutions, methods such as meta-analysis or meta-synthesis were not deemed the most appropriate approach for the data synthesis. The extracted data were summarized using a combination of descriptive and narrative syntheses [27]. The descriptive analysis was conducted to summarize data from the different studies. This involved classifying the studies based on the type of mobile technologies used, health conditions treated, and the types of PGHD reviewed. This approach arranged the studies into more homogenous subgroups, which aided in synthesizing different types of data. The data related to the motivations of HCPs were examined using a thematic analysis, from which different categories were derived. For the narrative synthesis, similarities, and differences between the findings of different studies were identified. The analysis and synthesis comprised three major steps: (1) organization of the included studies, (2) descriptive analysis of the findings within studies, and (3) a narrative synthesis aiming at exploring interconnections between the studies.

Quality Assessment

The quality of the included studies was assessed in parallel to the data extraction process. From the checklists, the quality of studies proposed by Kitchenham and Charters [21] were assessed. As the included studies were both qualitative and quantitative, we selected the questions that were most appropriate for our specific research questions that were present in both the qualitative and quantitative checklists.

Upon assessment, all reviewers agreed that the included studies had credible findings; proper data collection methods; clear and coherent reporting; and clear links between data, interpretation, and conclusions (Figure 1).

Figure 1. Quality assessment of included studies.

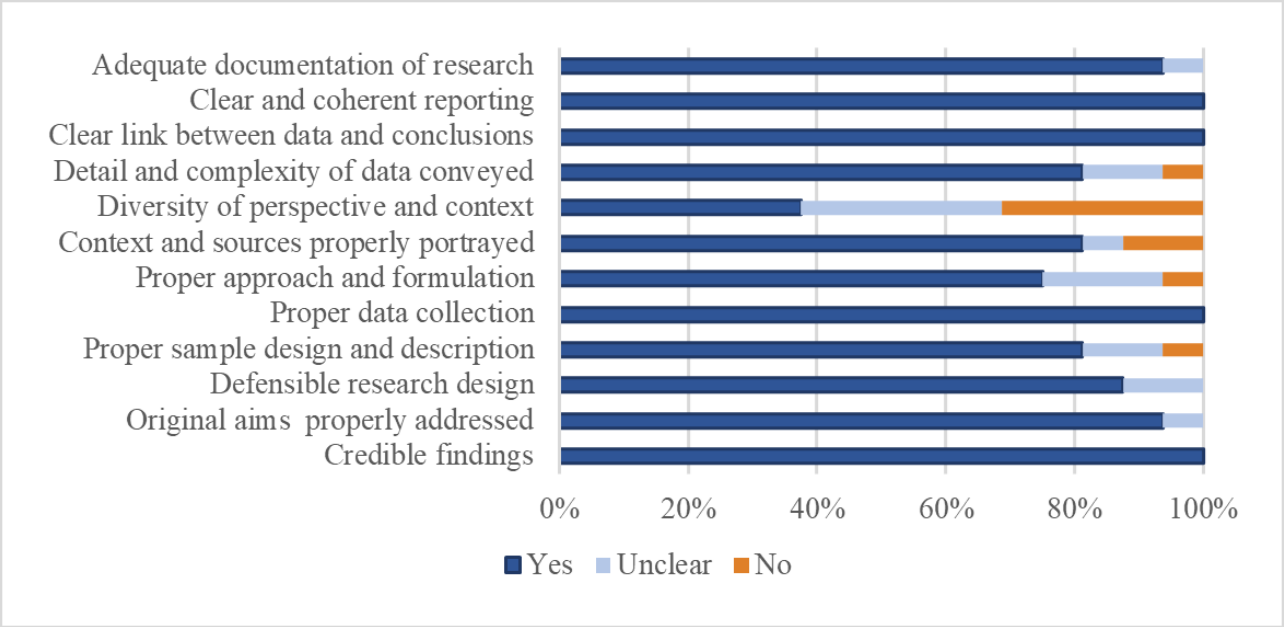

The single topic that produced some uncertainty during the quality assessment was the lack of clarity on whether some of the selected studies had explored enough diversity of perspective and context. This can likely be attributed to the fact that almost all included studies were performed in developed countries, predominantly in the United States or Europe, which is a typical setting for digital health studies. Hence, the findings of these investigations will provide the most accurate depiction of the state of health care systems in developed countries.

XSL•FO
RenderX

Ethical Considerations

The Ethics Committee of Human Sciences of the University of Oulu guidelines state that as no human or animal subjects were involved in the study, no separate ethics statement is required. However, the general ethical guidelines from the Finnish National Board on Research Integrity [28] guided the ethics of the study.

Results

Overview

Our search across electronic databases and supplementary hand searches identified 1696 papers. Covidence automatically removed 374 duplicates. We screened 1322 titles and abstracts, resulting in 86 papers for full-text screening. Following the completion of this second screening stage, 18 papers met all the inclusion criteria. However, upon closer examination, it was observed that 2 pairs of papers ([29,30] and [31,32]) had similar authors and identical samples and methodologies. Each pair was merged into a single study for analysis, resulting in the final inclusion of 16 studies for our SLR (Figure 2).

During full-text screening, papers were primarily excluded for focusing exclusively on PGHD from medical-grade devices (31/68, 46%); evaluating the usability of specific mHealth solutions, rather than PGHD use (27/68, 39%); lacking data collection from HCPs (9/68, 13%); and discussing potential rather than actual use of PGHD (1/68, 1%).

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. HCP: health care professional; PGHD: patient-generated health data; SLR: systematic literature review.

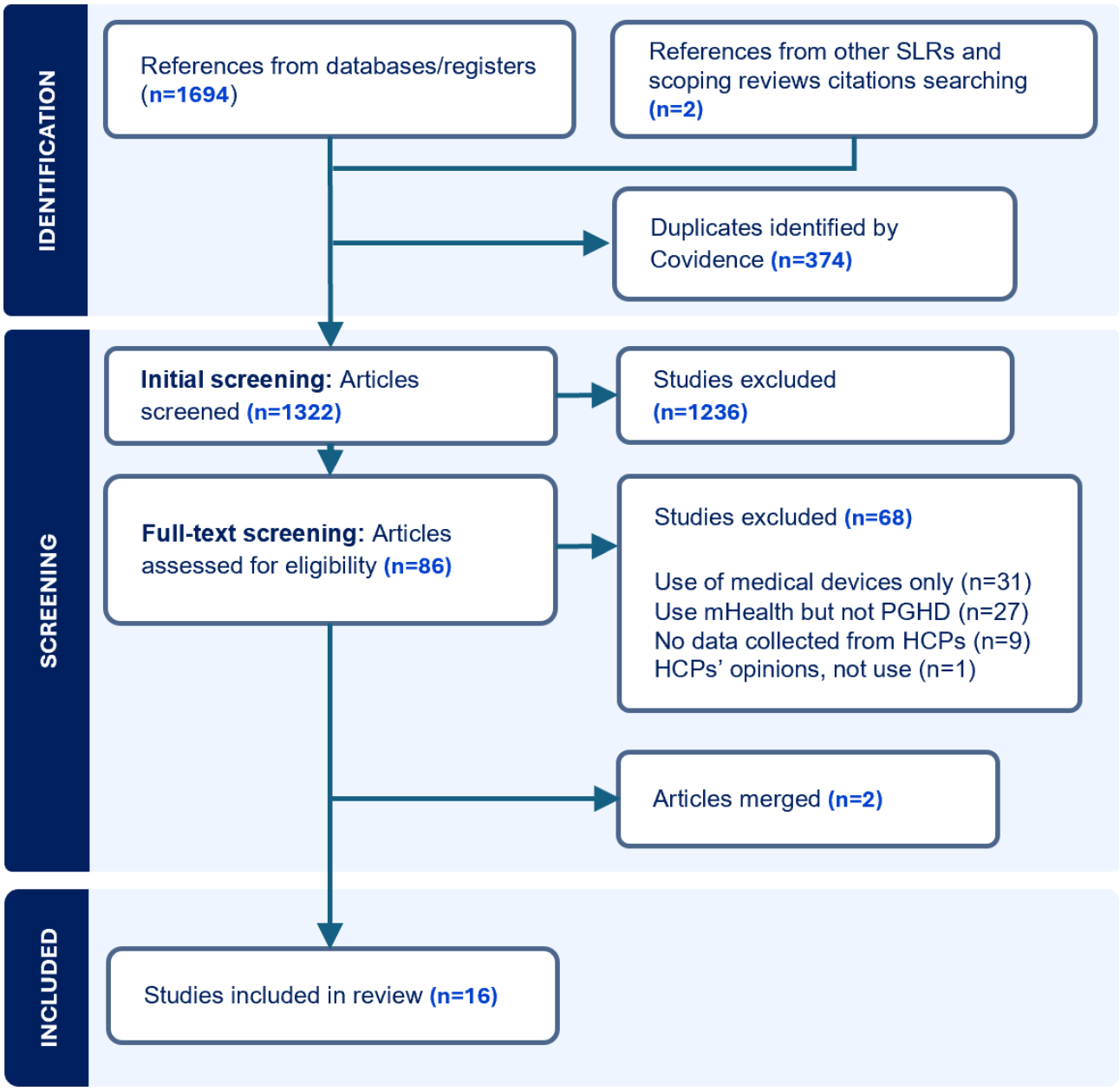

XSL•FO
RenderX

Characteristics of the Included Studies

We included studies spanning 2015-2021. Notably, more than two-thirds of the papers (11/16, 69%) were published in 2019 or later, indicating a growing interest in the topic both before and during the COVID-19 pandemic. The predominant location was North America (11/16, 69%), specifically the United States and Canada; within Europe (3/16, 19%), Sweden and the United Kingdom were the primary locations; and 1 study was conducted in Asia and 1 in a multicountry setting. The authors used diverse methodologies for data collection, with interviews (8/16, 50%) and mixed methods (4/16, 25%) being the most common. More comprehensive insights into the specific study designs and data collection methods are available in Table 2. A complete summary of the included studies can be found in Multimedia Appendix 4 [29,30,32-46].

XSL•FO
RenderX

Table 2. Summary of the included studies.

References	Location	Study design	Data collection methods	Diseases treated	Participants	Type of mobile technology used for data generation	Time frame of use
Abdolkhani et al [33]	United States, United Kingdom, Australia	Qualitative study	Interviews	Diabetes, cardiac arrhythmia, and sleep disorders	Physicians, nurses, allied health practitioners	Wearable devices	Information was collected from HCPs'[a] experiences with different wearables. The time frame was not specified
Alpert et al [29,30]	United States	Qualitative study	Interviews	Diverse conditions including musculoskeletal injuries, cardiovascular and respiratory conditions, and other nonspecific conditions	Physicians, nurses, geriatrics, physiotherapists, anesthesiology, and orthopedic surgery	Smartwatches	Information was collected from HCPs' experiences with data PGHD[b] from smartwatches. The time frame was not specified
Austin et al [34]	United States	Qualitative study	Surveys	Depression, substance use, blood pressure, diabetes, multiple sclerosis, cerebral palsy, perioperative care, osteoarthritis, and cognitive and gastrointestinal conditions	Physicians and researchers	Wearable devices, smartphones, medical grade wearable devices	Information was collected from HCPs' experiences with different interventions that ranged from months to 1 week
Cerna et al [32]	Sweden	Ethnographic study	Video and audio recordings of meetings and phone calls	Pelvic cancer	Nurses	Smartphones	Patients used a bespoke mobile app between appointments. The duration of the use was agreed between the patient and the nurse
Chung et al [35]	United States	Qualitative study	Interviews	Irritable bowel syndrome and obesity	Physicians, nurses, gastroenterologists, dietitians, and behavioral psychologists	Smartphones and physical activity trackers	Information was collected from HCPs' experiences with different solutions. The time frame was not specified
Cohen et al [36]	United States	Descriptive study	Interviews and immersion-crystallization approach	Asthma, cognitive decline, obesity, and Crohn disease	Physicians, nurses, and health coaches	Smartphones and other nonwearable devices	There were 5 different bespoke interventions, and their durations were varied
Costa Figueiredo et al [37]	United States	Qualitative study	Interviews	Infertility	Infertility specialists and endocrinologists	Smartphones	Information was collected from HCPs' experiences with different apps. The time frame was not specified
Gupta et al [38]	Canada	Mixed methods study	Workshops and interviews	Arthritis	Physiotherapists	Fitness tracking devices (Fitbit)	The duration of the intervention was agreed upon between the patient and the physiotherapist
Holtz et al [39]	United States	Cross-sectional study	Surveys	Not specified	Physicians and nurses	Smartphones, wearable devices, web-based apps	Information was collected from HCPs' experiences with different solutions. The time frame was not specified

XSL•FO
RenderX

References	Location	Study design	Data collection methods	Diseases treated	Participants	Type of mobile technology used for data generation	Time frame of use
Kim et al [40]	Canada	Mixed methods study	Focus groups, questionnaire	Care of older adults	Physicians, nurses and physiotherapists	Smartphones and wearable devices	Information was collected guided by a case study and from HCPs' experiences with different solutions. The time frame was not specified
Kim et al [41]	South Korea	Field study	Interviews and observation	Diabetes, hypertension, heart disease, obesity, hyperlipidemia, sleep apnea, cancer, chronic rhinosinusitis, and myoma uteri	Physicians, internal medicine, otorhinolaryngology, family medicine, obstetrics and gynecology, and rehabilitation	Smartphones and wearable devices	Patients used a bespoke mobile app between appointments. The duration of the use is not specified
Ng et al [42]	United States	Qualitative study	Interviews	Posttraumatic stress disorder	Nurses, nutritionists, and group and individual therapists	Fitness tracking devices (Fitbit)	The patient could determine the duration of the use
Tendedez et al [43]	England	Qualitative study	Focus groups and interviews	Chronic obstructive pulmonary disease	Physicians, nurses, assistant practitioners, and physiotherapists	Smartphone	Patients used an off-the-shelf mobile app for 4 weeks
West et al [44]	United Kingdom	Qualitative study	Interviews	Not specified	Cardiologists, mental health specialists, general practitioners, heart failure and oncology nurses, emergency doctors, junior surgeons, and audiologists	Smartphones, wearable devices	Information was collected from HCPs' experiences with different solutions. The time frame was not specified
Wu et al [45]	United States	Qualitative study	Interviews	Mental health	Psychiatrists and clinical psychologists	Smartphones, wearable devices	Information was collected from HCPs' experiences with different solutions. The time frame was not specified
Zhu et al [46]	United States	Qualitative study	Interviews	Not specified	Physicians, physiotherapists, internist, primary care psychologist, and pediatric nephrologists	Smartphones	Information was collected from HCPs' experiences with different apps. The time frame was not specified

[a]HCP: health care professional.

[b]PGHD: patient-generated health data.

Medical Profiles and Specialties

Although some of the studies examined data collected from various stakeholders such as patients, researchers, hospital managers, or solution providers, our focus centered on data collected from HCPs. Collectively, the studies in our review had 355 HCPs as participants. Among the represented professions, physicians accounted for the largest number of participants, present in 81% (13/16) of the studies. While approximately half of those studies referred to physicians using a general term, the other half provided clear information about the medical specialties of the physicians. Nurses were the second most represented profession, participating in 62% (10/16) of the studies. Physiotherapists were the third most represented, participating in 38% (6/16) of the studies. Other health professions present were psychologists (3/16, 19%) and surgeons, dietitians, health coaches, and assistant practitioners, each mentioned in 12% (2/16) of the studies (Table 2).

All the studies reported the medical specialties where PGHD was being used. Those specialties included geriatrics, anesthesiology, orthopedic surgery, gastroenterology, dietetics and nutrition, behavioral and clinical psychology, psychiatry, obstetrics and gynecology, infertility, endocrinology, internal medicine, family medicine, rehabilitation, pediatric nephrology, otorhinolaryngology, and audiology.

Health Conditions Treated

The studies examined a wide range of health conditions, classified according to the WHO *International Classification of Diseases, Eleventh Revision* (*ICD-11*), into categories such as endocrine, nutritional, or metabolic diseases; mental, behavioral, or neurodevelopmental disorders; diseases of the nervous, circulatory respiratory, and digestive systems, and

XSL•FO
RenderX

diseases of the musculoskeletal system or connective tissue. In addition, some studies reported the use of PGHD for other types of medical tasks including perioperative care and care of older adults.

The most cited health conditions for which PGHD from mobile devices were reviewed by HCPs were diabetes and obesity, each mentioned in at least 3 studies. A quarter of the studies did not address a specific health condition. In those cases, the contextual information provided was limited to medical specialties or professions (Table 2).

Types of mHealth Solutions

Among the 16 included studies, 5 mentioned specific mHealth solutions patients had been using to self-manage their health condition. The remaining studies mentioned commercial mHealth solutions in general. In half of the studies, HCPs reported using PGHD derived from a combination of diverse mHealth solutions, which included 1 or multiple mobile health apps and wearable devices. The remaining half of the studies addressed the experience of HCPs using PGHD exclusively generated through mobile health apps installed in patients' smartphones (4/16, 25%) or captured from wearable devices (4/16, 25%).

Types of PGHD

Various classifications of PGHD have been proposed in terms of purpose (self-use, behavior change, clinical use, and research), management of a condition (eg, diabetes, hypertension), data type (physiological, behavioral, or environmental), mode of data capture (using sensors, external devices, implanted devices, patient portals, web-based surveys, and manual entry), and whether the process is active, passive, or mixed [12]. In this study, we focused on classifying PGHD based on data types.

Physiological data were reviewed in all studies. In 7 of 16 studies, at least 3 different types of physiological data were collected. Weight was the most frequently mentioned physiological data, reported in 44% (7/16) of the studies, followed by mood (6/16, 38%) and vital signs (5/16, 31%). Other less commonly reviewed types of data were pain, blood glucose level, and other symptoms (Table 3).

XSL•FO
RenderX

Table 3. Types of PGHD[a] used by HCPs[b].

Classification of PGHD and type of PGHD	Studies, n
Physiological data types	
Weight	7
Mood	6
Vital signs	5
Symptoms	4
Pain	3
Glucose level	3
Fatigue	2
Cognition testing	2
Menstrual history	2
Defecation and urination frequencies	2
Stool and urinary leakage	1
Cough	1
Mobility	1
Breathlessness	1
Stress	1
Peak flow	1
Sputum production and characteristics	1
Paralinguistic	1
Self-perception of performance (mental and physical)	1
Behavioral data types	
Physical activity	12
Eating habits	9
Sleep	8
Medication adherence	6
Lifestyle	3
Sedentariness	2
Substance use	1
Technology use	1
Falls	1
Driving	1
Hydration	1
Goal setting	1
Environmental data types	
Location data	1
Environmental factors	1

[a]PGHD: patient-generated health data.

[b]HCP: health care professional.

Behavioral data constituted the most used category of PGHD. More than 80% (13/16) of the studies indicated that HCPs had reviewed some form of behavioral data, although always in combination with physiological data. Physical activity seems to be the most reviewed type of PGHD produced by consumer-grade devices, with 75% of the studies reporting its use, followed by food intake (9/16, 56%), sleep quality or quantity (8/16, 50%), and medication adherence (6/16, 38%).

XSL•FO
RenderX

Only 12% (2/16) of the studies reported the use of environmental data, which were primarily collected through passive sensing, using wearables, whereas physiological and behavioral types of data were reported to be collected through either passive or active sensing or by a combination of both. For instance, certain types of PGHD, such as sleep, physical activity, or sedentariness, were collected through active sensing in some studies and through passive sensing in others.

Access to PGHD

Diverse channels for PGHD access were presented. Notably, 19% (3/16) of the papers did not describe the precise channels HCPs used to access PGHD. Dashboards or solution portals were used in 56% (9/16) of the studies. The second most common channel was the patient's mobile device (5/16, 31%). In a few studies, HCPs accessed PGHD through integration with the electronic health record (EHR; 2/16, 12%), by email (2/16, 12%), or from patients' verbal summaries of data from their mobile devices (1/16, 6%).

Motivation for Reviewing PGHD

Although not all studies cited the reasons behind HCPs' willingness to review PGHD from consumer-grade devices, motivation for reviewing them centered into 3 main categories: benefits for the patient, supporting their clinical roles, and strengthening the patient-HCP relationship (Figure 3). Key motivations that showed how PGHD supported HCPs included topics such as accessing additional data types, identifying health patterns, and reducing data collection workload.

Figure 3. HCPs' motivations for using patient-generated health data. HCP: health care professional.

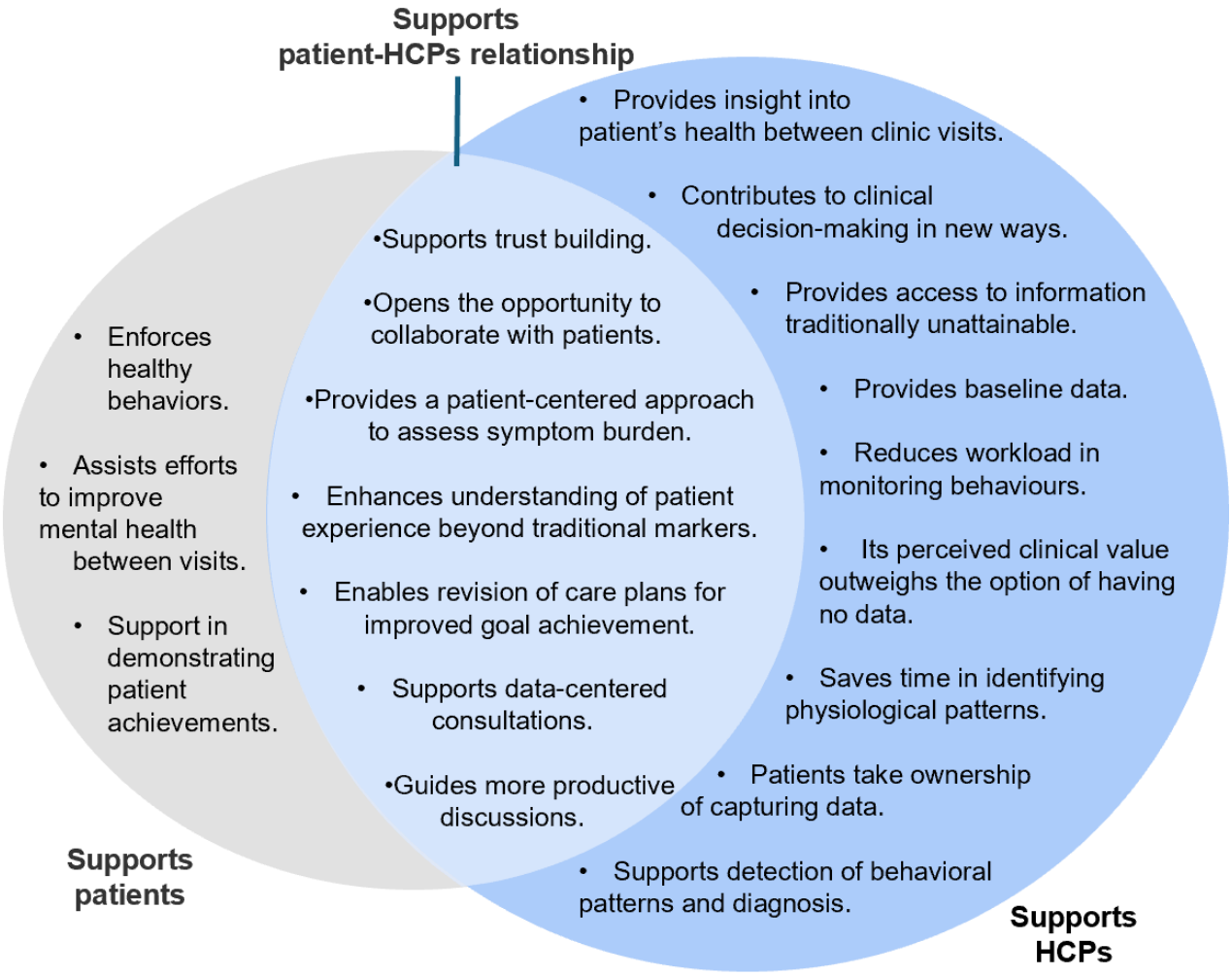

Discussion

Principal Findings

Our review underlines a growing interest in understanding the experiences of HCPs who are using PGHD in the clinic. We aimed to identify how PGHD from consumer-grade mobile devices have been used to assist them in clinical practice. HCPs, who were primarily physicians and nurses, shared their experience on the topic. The health conditions for which HCPs most resorted to PGHD were diabetes and obesity. We found that physiological data, such as weight, mood, and vital signs, and behavioral data, such as physical activity, food intake, and sleep quality, have been frequently used. HCPs had access to

XSL•FO
RenderX

PGHD through different channels, such as web portals provided by the mHealth solutions or through integration with the EHR.

Previous reviews have explored the role of PGHD in facilitating prevention and health promotion [17], their use in clinical practice [18], and their effect on patient-clinician relationships [47]. However, those studies have concentrated on PGHD from medical-grade devices, which tend to be more accurate and more accepted in the medical community. PGHD created through consumer-grade mHealth solutions, although praised for their potential to transform health care, have typically not been deemed reliable or accurate enough for the clinical context [3,48,49]. Despite concerns over PGHD accuracy and reliability, HCPs recognized that their clinical value outweighs the absence of data [40]. This value comes with a caveat, as recent studies indicate that PGHD must be curated by HCPs to ascribe actionable clinical value, but even then, they can be treated as supplementary to data collected through clinically recognized standards such as through laboratory tests [50,51]. PGHD from consumer-grade solutions have been used by HCPs in the treatment of a wide variety of health conditions, although it seems common only in the care of diabetes, cardiovascular diseases, and obesity (Table 2).

The most frequently used types of data (physical activity, food intake, sleep quantity, and weight) are highly associated with lifestyle health risks, implying that access to lifestyle-related data can provide valuable insights into the control of lifestyle-related diseases. Furthermore, patients having these conditions are more willing to share PGHD, therefore, fostering HCPs' familiarity with those types of data [15]. Our findings reveal that PGHD's use in clinical practice remains relatively scarce [29,36,42], pointing out a gap between their potential and their current use. This finding is in line with a recent study suggesting that in comparison with the expectations of policies related to the European Health Data Space, the prompting and reviewing of PGHD from consumer-grade devices seem still relatively rare [50]. It is plausible that these types of PGHD have been used by HCPs in practice, but research on the practicalities of this phenomenon has only increased in the last 5 years.

HCPs indicated that PGHD are useful in the identification of patterns, to support certain diagnoses, and for certain types of monitoring. For example, lifestyle diseases [36], irritable bowel syndrome [42], or infertility [37] requires long-term management or presents a high level of uncertainty. In these cases, PGHD can provide longitudinal insights into patients' health between clinic visits or even before they start treatment, saving time in identifying patterns. It is worth noting that, although HCPs in those studies acknowledged the value of PGHD, they also indicated engaging with PGHD infrequently and only with a few specific data types, in comparison with the substantial amount of data some patients want to share. For patients with chronic diseases, knowing that HCPs are reviewing their PGHD can be a comfortable way to know that they are being monitored and can provide data at the right time to facilitate decision-making and early intervention [41,52].

Multiple types of data were collected in all the studies, which signifies that as more data are collected, the need for analytical strategies that can support HCPs in reviewing and analyzing the potential relationships between different categories of data will be higher. Most existing mHealth solutions for self-monitoring lack standardized formats and mechanisms for patients to control and share PGHD [40,45]. Support for HCPs' data access and use requires standardization and, in some cases, EHR integration [44,45].

Limitations

We limited our inclusion to papers written in English. However, this approach may have excluded relevant papers from developing regions where English is not the primary language for scientific dissemination but where the interest and potential for mHealth solutions and PGHD are growing. Similarly, a gap in the current body of research regarding these topics in developing regions is highlighted, since all studies came from countries with high economic and digitalization levels.

PGHD is a relatively recent definition, and some relevant papers published prior to its official designation as a MeSH term may have employed alternative terminology to describe the same concept of PGHD used in our study.

The shift toward digital health solutions the COVID-19 pandemic potentiated may have modified HCPs' perceptions of PGHD use. However, no studies explicitly examining this relationship were identified in our prior searches or a later search. Therefore, future research could explore whether the shift toward digital health has catalyzed the adoption of consumer-grade technologies and PGHD in clinical settings.

Conclusions

Despite skepticism regarding the reliability and accuracy of PGHD and the multiple challenges that they convey, our study highlights a noticeable shift toward recognizing their practical value in health care, particularly in managing chronic conditions such as diabetes, obesity, and cardiovascular diseases. Yet, their impact in supporting the clinical practice is not clear from the literature. Many HCPs in the study, predominantly physicians and nurses, showed interest in using PGHD in the clinical workflows, albeit with a cautious approach that considers them as supplementary to traditional clinical data only. While they acknowledged the benefit of reviewing PGHD for the patient-HCP relationship, it was also noted that only certain types of PGHD are truly deemed useful and even then, they are not regularly used by HCPs. The findings call for continued research and innovation in mHealth, with a focus on enhancing the reliability, usability, and clinical relevance of PGHD, which in return can foster a culture of trust and collaboration between patients and HCPs.

XSL•FO
RenderX

Acknowledgments

We would like to acknowledge the More Stamina Project research group for supporting the development of this work. We acknowledge the use of ChatGPT version 4 and Grammarly to identify improvements in the organization of our text and to improve the writing style in the Introduction and Discussion sections.

Data Availability

This literature review synthesizes findings from peer-reviewed journal papers and conference papers. Given the nature of this review, it does not generate new primary data; instead, it compiled and analyzed existing publications on the use of PGHD from mobile technologies by HCPs. The reviewed papers are all available in public scientific databases. The data extraction form and the extracted data are available in the Multimedia Appendix section. These resources aim to ensure the reproducibility of our methods and facilitate future research in this area.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 105 KB - jmir_v26i1e49320_app1.pdf]

Multimedia Appendix 2
Search strategies for all searched databases.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v26i1e49320_app2.xlsx]

Multimedia Appendix 3
Data extraction form.
[PDF File (Adobe PDF File), 244 KB - jmir_v26i1e49320_app3.pdf]

Multimedia Appendix 4
Summary of the included studies.
[XLSX File (Microsoft Excel File), 19 KB - jmir_v26i1e49320_app4.xlsx]

References

1. Sim I. Mobile devices and health. N Engl J Med 2019;381(10):956-968. [doi: 10.1056/NEJMra1806949] [Medline: 31483966]
2. Istepanian RSH. Mobile health (m-Health) in retrospect: the known unknowns. Int J Environ Res Public Health 2022;19(7):3747 [FREE Full text] [doi: 10.3390/ijerph19073747] [Medline: 35409431]
3. Omoloja A, Vundavalli S. Patient generated health data: benefits and challenges. Curr Probl Pediatr Adolesc Health Care 2021;51(11):101103. [doi: 10.1016/j.cppeds.2021.101103] [Medline: 34799255]
4. Bradway M, Gabarron E, Johansen M, Zanaboni P, Jardim P, Joakimsen R, et al. Methods and measures used to evaluate patient-operated mobile health interventions: scoping literature review. JMIR Mhealth Uhealth 2020;8(4):e16814 [FREE Full text] [doi: 10.2196/16814] [Medline: 32352394]
5. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018;15(5):429-448 [FREE Full text] [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
6. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021;23(4):e26699 [FREE Full text] [doi: 10.2196/26699] [Medline: 33811021]
7. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019;7(7):e13817 [FREE Full text] [doi: 10.2196/13817] [Medline: 31293246]
8. Feng S, Mäntymäki M, Dhir A, Salmela H. How self-tracking and the quantified self promote health and well-being: systematic review. J Med Internet Res 2021;23(9):e25171 [FREE Full text] [doi: 10.2196/25171] [Medline: 34546176]
9. Lavallee DC, Lee JR, Austin E, Bloch R, Lawrence SO, McCall D, et al. mHealth and patient generated health data: stakeholder perspectives on opportunities and barriers for transforming healthcare. Mhealth 2020;6:8 [FREE Full text] [doi: 10.21037/mhealth.2019.09.17] [Medline: 32190619]
10. Figueiredo MC, Chen Y. Patient-generated health data: dimensions, challenges, and open questions. Found Trends Hum–Comput Interact 2020;13(3):165-297. [doi: 10.1561/1100000080]

XSL•FO
RenderX

11. Jim HSL, Hoogland AI, Brownstein NC, Barata A, Dicker AP, Knoop H, et al. Innovations in research and clinical care using patient-generated health data. CA Cancer J Clin 2020;70(3):182-199 [FREE Full text] [doi: 10.3322/caac.21608] [Medline: 32311776]
12. Mars M, Scott RE. Electronic patient-generated health data for healthcare. In: Linwood SL, editor. Digital Health. Brisbane, Australia: Exon Publications; 2022:1-16.
13. Li X, Dunn J, Salins D, Zhou G, Zhou W, Schüssler-Fiorenza Rose SM, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLoS Biol 2017;15(1):e2001402 [FREE Full text] [doi: 10.1371/journal.pbio.2001402] [Medline: 28081144]
14. Karampela M, Ouhbi S, Isomursu M. Connected health user willingness to share personal health data: questionnaire study. J Med Internet Res 2019;21(11):e14537 [FREE Full text] [doi: 10.2196/14537] [Medline: 31774410]
15. Turner K, Jo A, Wei G, Tabriz AA, Clary A, Jim HSL. Sharing patient-generated data with healthcare providers: findings from a 2019 national survey. J Am Med Inform Assoc 2021;28(2):371-376 [FREE Full text] [doi: 10.1093/jamia/ocaa272] [Medline: 33180896]
16. Gurupur VP, Wan TTH. Challenges in implementing mHealth interventions: a technical perspective. Mhealth 2017;3:32 [FREE Full text] [doi: 10.21037/mhealth.2017.07.05] [Medline: 28894742]
17. Nittas V, Lun P, Ehrler F, Puhan MA, Mütsch M. Electronic patient-generated health data to facilitate disease prevention and health promotion: scoping review. J Med Internet Res 2019;21(10):e13320 [FREE Full text] [doi: 10.2196/13320] [Medline: 31613225]
18. Demiris G, Iribarren SJ, Sward K, Lee S, Yang R. Patient generated health data use in clinical practice: a systematic review. Nurs Outlook 2019;67(4):311-330 [FREE Full text] [doi: 10.1016/j.outlook.2019.04.005] [Medline: 31277895]
19. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
20. Gagnon MP, Ngangue P, Payne-Gagnon J, Desmartis M. m-Health adoption by healthcare professionals: a systematic review. J Am Med Inform Assoc 2016;23(1):212-220 [FREE Full text] [doi: 10.1093/jamia/ocv052] [Medline: 26078410]
21. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Researchgate. 2007. URL: https://www.researchgate.net/profile/Barbara-Kitchenham/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering/links/61712932766c4a211c03a6f7/Guidelines-for-performing-Systematic-Literature-Reviews-in-Software-Engineering.pdf [accessed 2024-05-07]
22. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. BMJ 2009;339:b2535-b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Karampela M, Ouhbi S, Isomursu M. Personal health data: a systematic mapping study. Int J Med Inform 2018;118:86-98. [doi: 10.1016/j.ijmedinf.2018.08.006] [Medline: 30153927]
25. Medina SG, Isomursu M. The use of patient-generated health data from consumer-grade mobile devices in clinical workflows: protocol for a systematic review. JMIR Res Protoc 2023;12:e39389 [FREE Full text] [doi: 10.2196/39389] [Medline: 36848208]
26. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: a decade of evolution. J Med Internet Res 2013;15(5):e95 [FREE Full text] [doi: 10.2196/jmir.2430] [Medline: 23697600]
27. Evans D. Systematic reviews of interpretive research: interpretive data synthesis of processed data. Aust J Adv Nurs 2002;20(2):22-26 [FREE Full text]
28. Guidelines for ethical review in human sciences. Publications of the Finnish National Board on Research Integrity. 2019. URL: https://tenk.fi/sites/default/files/2023-11/RI_Guidelines_2023.pdf
29. Alpert JM, Manini T, Roberts M, Kota NSP, Mendoza TV, Solberg LM, et al. Secondary care provider attitudes towards patient generated health data from smartwatches. NPJ Digit Med 2020;3(1):27 [FREE Full text] [doi: 10.1038/s41746-020-0236-4] [Medline: 32140569]
30. Alpert JM, Kota NSP, Ranka S, Mendoza TV, Solberg LM, Rashidi P, et al. A simulated graphical interface for integrating patient-generated health data from smartwatches with electronic health records: usability study. JMIR Hum Factors 2020;7(4):e19769 [FREE Full text] [doi: 10.2196/19769] [Medline: 33124988]
31. Lindroth T, Islind AS, Steineck G, Lundin J. From narratives to numbers: data work and patient-generated health data in consultations. Stud Health Technol Inform 2018;247:491-495. [Medline: 29678009]
32. Cerna K, Grisot M, Islind AS, Lindroth T, Lundin J, Steineck G. Changing categorical work in healthcare: the use of patient-generated health data in cancer rehabilitation. Comput Supported Coop Work 2020;29(5):563-586 [FREE Full text] [doi: 10.1007/s10606-020-09383-z]

XSL•FO
RenderX

33. Abdolkhani R, Gray K, Borda A, DeSouza R. Patient-generated health data management and quality challenges in remote patient monitoring. JAMIA Open 2019;2(4):471-478 [FREE Full text] [doi: 10.1093/jamiaopen/ooz036] [Medline: 32025644]
34. Austin E, Lee JR, Amtmann D, Bloch R, Lawrence SO, McCall D, et al. Use of patient-generated health data across healthcare settings: implications for health systems. JAMIA Open 2020;3(1):70-76 [FREE Full text] [doi: 10.1093/jamiaopen/ooz065] [Medline: 32607489]
35. Chung CF, Cook J, Bales E, Zia J, Munson SA. More than telemonitoring: health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. J Med Internet Res 2015;17(8):e203 [FREE Full text] [doi: 10.2196/jmir.4364] [Medline: 26297627]
36. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project HealthDesign. JMIR Hum Factors 2016;3(2):e26 [FREE Full text] [doi: 10.2196/humanfactors.5919] [Medline: 27760726]
37. Costa Figueiredo M, Su HI, Chen Y. Using data to approach the unknown. Proc ACM Hum-Comput Interact 2021;4(CSCW3):1-35 [FREE Full text] [doi: 10.1145/3432926]
38. Gupta A, Heng T, Shaw C, Gromala D, Leese J, Li L. Oh, I didn't do a good job: how objective data affects physiotherapist-patient conversations for arthritis patients. In: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2020 Presented at: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 18-20, 2020; Atlanta, GA, USA p. 156-165. [doi: 10.1145/3421937.3421991]
39. Holtz B, Vasold K, Cotten S, Mackert M, Zhang M. Health care provider perceptions of consumer-grade devices and apps for tracking health: a pilot study. JMIR Mhealth Uhealth 2019;7(1):e9929 [FREE Full text] [doi: 10.2196/mhealth.9929] [Medline: 30668515]
40. Kim B, Ghasemi P, Stolee P, Lee J. Clinicians and older adults' perceptions of the utility of patient-generated health data in caring for older adults: exploratory mixed methods study. JMIR Aging 2021;4(4):e29788 [FREE Full text] [doi: 10.2196/29788] [Medline: 34738913]
41. Kim Y, Ji S, Lee H, Kim JW, Yoo S, Lee J. "My doctor is keeping an eye on me!": exploring the clinical applicability of a mobile food logger. 2016 Presented at: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, CA p. 5620-5631. [doi: 10.1145/2858036.2858145]
42. Ng A, Kornfield R, Schueller SM, Zalta AK, Brennan M, Reddy M. Provider perspectives on integrating sensor-captured patient-generated data in mental health care. Proc ACM Hum Comput Interact 2019;3(CSCW):1-25 [FREE Full text] [doi: 10.1145/3359217] [Medline: 33585802]
43. Tendedez H, Ferrario MA, McNaney R. 'The issue with that sort of data?': clinicians' accountability concerns around COPD self-monitoring tools. 2019 Presented at: Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing; November 9-13, 2019; Austin, TX p. 382-386. [doi: 10.1145/3311957.3359464]
44. West P, Van Kleek M, Giordano R, Weal MJ, Shadbolt N. Common barriers to the use of patient-generated data across clinical settings. 2018 Presented at: Proceedings of the 2018 CHI Conference on Human Factors in Computing System; April 21-26, 2018; Montreal, QC p. 1-13. [doi: 10.1145/3173574.3174058]
45. Wu DTY, Xin C, Bindhu S, Xu C, Sachdeva J, Brown JL, et al. Clinician perspectives and design implications in using patient-generated health data to improve mental health practices: mixed methods study. JMIR Form Res 2020;4(8):e18123 [FREE Full text] [doi: 10.2196/18123] [Medline: 32763884]
46. Zhu H, Colgan J, Reddy M, Choe EK. Sharing patient-generated data in clinical practices: an interview study. AMIA Annu Symp Proc 2016;2016:1303-1312 [FREE Full text] [Medline: 28269928]
47. Lordon RJ, Mikles SP, Kneale L, Evans HL, Munson SA, Backonja U, et al. How patient-generated health data and patient-reported outcomes affect patient-clinician relationships: a systematic review. Health Informatics J 2020;26(4):2689-2706 [FREE Full text] [doi: 10.1177/1460458220928184] [Medline: 32567460]
48. West P, Van Kleek M, Giordano R, Weal M, Shadbolt N. Information quality challenges of patient-generated data in clinical practice. Front Public Health 2017;5:284 [FREE Full text] [doi: 10.3389/fpubh.2017.00284] [Medline: 29209601]
49. Park JI, Lee HY, Kim H, Lee J, Shinn J, Kim HS. Lack of acceptance of digital healthcare in the medical market: addressing old problems raised by various clinical professionals and developing possible solutions. J Korean Med Sci 2021;36(37):e253 [FREE Full text] [doi: 10.3346/jkms.2021.36.e253] [Medline: 34581521]
50. Haase CB, Ajjawi R, Bearman M, Brodersen JB, Risor T, Hoeyer K. Data as symptom: doctors' responses to patient-provided data in general practice. Soc Stud Sci 2023;53(4):522-544 [FREE Full text] [doi: 10.1177/03063127231164345] [Medline: 37096688]
51. Reading MJ, Merrill JA. Converging and diverging needs between patients and providers who are collecting and using patient-generated health data: an integrative review. J Am Med Inform Assoc 2018;25(6):759-771 [FREE Full text] [doi: 10.1093/jamia/ocy006] [Medline: 29471330]
52. Gaveikaite V, Grundstrom C, Winter S, Schonenberg H, Isomursu M, Chouvarda I, et al. Challenges and opportunities for telehealth in the management of chronic obstructive pulmonary disease: a qualitative case study in Greece. BMC Med Inform Decis Mak 2020;20(1):216 [FREE Full text] [doi: 10.1186/s12911-020-01221-y] [Medline: 32912224]

XSL•FO
RenderX

Abbreviations

EHR: electronic health record
HCP: health care professional
mHealth: mobile health
PGHD: patient-generated health data
SLR: systematic literature review

Edited by A Mavragani; submitted 26.05.23; peer-reviewed by CM Chu, P Dunn, A Brigden, C Baxter; comments to author 08.02.24; revised version received 05.04.24; accepted 11.04.24; published 31.05.24.

Please cite as:
Guardado S, Karampela M, Isomursu M, Grundstrom C
Use of Patient-Generated Health Data From Consumer-Grade Devices by Health Care Professionals in the Clinic: Systematic Review
J Med Internet Res 2024;26:e49320
URL: https://www.jmir.org/2024/1/e49320
doi:10.2196/49320
PMID:38820580

©Sharon Guardado, Maria Karampela, Minna Isomursu, Casandra Grundstrom. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Examining the Effectiveness of Social Media for the Dissemination of Research Evidence for Health and Social Care Practitioners: Systematic Review and Meta-Analysis

Sarah Roberts-Lewis[1], PhD; Helen Baxter[2,3], PhD; Gill Mein[4], BSc; Sophia Quirke-McFarlane[5], MSc; Fiona J Leggat[1], PhD; Hannah Garner[6], BSc; Martha Powell[3], BSc; Sarah White[1], PhD; Lindsay Bearne[1,3], PhD

[1]Population Health Research Institute, St George's University of London, London, United Kingdom
[2]Bristol Population Health Science Institute, University of Bristol, Bristol, United Kingdom
[3]National Institute of Health and Care Research, London, United Kingdom
[4]Faculty of Health, Social Care and Education, St George's University of London, London, United Kingdom
[5]School of Psychology, University of Surrey, Guildford, United Kingdom
[6]Department of Physiotherapy, St George's University Hospitals NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Sarah Roberts-Lewis, PhD
Population Health Research Institute
St George's University of London
Cranmer Terrace
London, SW17 0RE
United Kingdom
Phone: 44 20 8725 0368
Email: sroberts@sgul.ac.uk

Abstract

Background: Social media use has potential to facilitate the rapid dissemination of research evidence to busy health and social care practitioners.

Objective: This study aims to quantitatively synthesize evidence of the between- and within-group effectiveness of social media for dissemination of research evidence to health and social care practitioners. It also compared effectiveness between different social media platforms, formats, and strategies.

Methods: We searched electronic databases for articles in English that were published between January 1, 2010, and January 10, 2023, and that evaluated social media interventions for disseminating research evidence to qualified, postregistration health and social care practitioners in measures of reach, engagement, direct dissemination, or impact. Screening, data extraction, and risk of bias assessments were carried out by at least 2 independent reviewers. Meta-analyses of standardized pooled effects were carried out for between- and within-group effectiveness of social media and comparisons between platforms, formats, and strategies. Certainty of evidence for outcomes was assessed using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) framework.

Results: In total, 50 mixed-quality articles that were heterogeneous in design and outcome were included (n=9, 18% were randomized controlled trials [RCTs]). Reach (measured in number of practitioners, impressions, or post views) was reported in 26 studies. Engagement (measured in likes or post interactions) was evaluated in 21 studies. Direct dissemination (measured in link clicks, article views, downloads, or altmetric attention score) was analyzed in 23 studies (8 RCTs). Impact (measured in citations or measures of thinking and practice) was reported in 13 studies. Included studies almost universally indicated effects in favor of social media interventions, although effect sizes varied. Cumulative evidence indicated moderate certainty of large and moderate between-group effects of social media interventions on direct dissemination (standardized mean difference [SMD] 0.88; P=.02) and impact (SMD 0.76; P<.001). After social media interventions, cumulative evidence showed moderate certainty of large within-group effects on reach (SMD 1.99; P<.001), engagement (SMD 3.74; P<.001), and direct dissemination (SMD 0.82; P=.004) and low certainty of a small within-group effect on impacting thinking or practice (SMD 0.45; P=.02). There was also evidence for the effectiveness of using multiple social media platforms (including Twitter, subsequently rebranded X; and Facebook), images (particularly infographics), and intensive social media strategies with frequent, daily posts and involving influential others. No included studies tested the dissemination of research evidence to social care practitioners.

XSL•FO
RenderX

Conclusions: Social media was effective for disseminating research evidence to health care practitioners. More intense social media campaigns using specific platforms, formats, and strategies may be more effective than less intense interventions. Implications include recommendations for effective dissemination of research evidence to health care practitioners and further RCTs in this field, particularly investigating the dissemination of social care research.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42022378793; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=378793

International Registered Report Identifier (IRRID): RR2-10.2196/45684

(J Med Internet Res 2024;26:e51418) doi:10.2196/51418

KEYWORDS

social media; dissemination; health care; social care; research evidence; practitioners; effectiveness; meta-analysis; systematic review; randomized controlled trial; RCT

Introduction

Background

It is essential that health and social care practitioners access contemporary, high-quality research evidence to help them deliver the best evidence-based clinical care and improve patient outcomes [1-5]. Rapid dissemination, by active approaches using specific channels and planned strategies, is recommended [6,7].

Social media may facilitate rapid dissemination to busy practitioners, allowing them to access and interpret research evidence efficiently [8-10]. Because social media are widely used and not limited in space and time [8,11], they have the potential to overcome barriers to dissemination, including reaching practitioners with limited professional opportunities or time constraints and filtering the exponentially increasing volume of research evidence produced every year [9,12-14]. Currently, closed social media channels, such as private and invitation-only groups, are often used by practitioners for day-to-day communications, clinical information sharing, and targeted clinical education, whereas open social media channels that can be accessed by everybody are used for reputation development; public health education; and, increasingly, research dissemination [9,10,15-26].

However, the effectiveness of open social media for the dissemination of research evidence to health and social care practitioners is largely unknown [27]. Existing reviews have narratively synthesized potential uses, benefits and risks, similarities and differences, and qualitative experiences of social media or provided commentaries on the mechanisms of research dissemination by social media [10,22,23,28]. No reviews have conducted a meta-analysis to quantitatively test the effectiveness of social media for the dissemination of research evidence to health and social care practitioners. To inform evidence-based recommendations, the evidence for using social media to disseminate research evidence must be investigated.

Objectives

The primary research question was as follows: “How effective is open social media as a way to disseminate research evidence to practitioners?” The objective of this systematic review was to quantitatively synthesize and meta-analyze evidence of the effectiveness of social media for the dissemination of research evidence to health and social care practitioners by evaluating both between-group comparisons of social media versus no social media and within-group comparisons of before-after social media campaigns. The social media platforms, formats, and strategies used were also identified, and their effectiveness was compared to understand the most effective social media intervention characteristics for the dissemination of research evidence to practitioners.

Methods

Design

The protocol was registered on the International Register of Systematic Review (PROSPERO; CRD42022378793) and published a priori [29]. It was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidance [30] (Multimedia Appendix 1).

Eligibility Criteria

Articles published between January 1, 2010, and January 10, 2023, were eligible for inclusion if they investigated research evidence targeted at health and social care practitioners that was shared using open social media. Articles were included if they quantitatively compared social media versus no social media (either between-group comparisons or before-after social media within-group comparisons) or if they compared social media platforms, formats, or strategies. Eligible study designs included randomized controlled trials (RCTs), case-controlled comparisons, crossover, nonrandomized group comparisons, before-after comparisons, cohort comparisons, and case reports. Eligible outcomes of interest included reach, engagement, direct dissemination, and impact. Definitions of the eligibility criteria terms are shown in Textbox 1.

Articles were not eligible for inclusion if they only compared social media effectiveness in terms of the topic or specialty of the research evidence–related social media post or posts. Excluded study designs included protocols, reviews, studies using only qualitative methods, opinion pieces, and conference abstracts with no linked full-text article. Articles were excluded if they preceded 2010 (refer to the protocol by Roberts-Lewis et al [29]), were not available in English, or did not feature research evidence–related social media of relevance to postregistration health or social care practitioners (eg, only targeting students, service users, or the public or featuring

XSL•FO
RenderX

non–health and social care research topics). Articles were also excluded if the social media campaign was targeted at practitioners for purposes other than the dissemination of research evidence (eg, delivering multisource clinical education courses, organizational information, administrative tasks, practical peer support, day-to-day interpersonal clinical communication, professional identity, or reputation promotion). Finally, articles that did not provide sufficient quantitative empirical data on reach, engagement, direct dissemination, or impact were excluded.

Textbox 1. Definitions of the eligibility criteria terms.

Definitions

- Research evidence: this was defined as published, peer-reviewed empirical human health and social care research findings that have met the publication standards of their specialty, presented as an original research article (primary research), a group of original research articles identified and synthesized systematically (secondary research), or evidence drawn together for evidence-based guidelines or clinical recommendations. Where research evidence was posted on social media, it either included a direct link to an open-access article or research information that had been summarized in the form of abstracts, microblogs, blogs, press articles, infographics, or educational videos.
- Targeted: by this we mean research evidence or social media posts that were professionally relevant to health and social care practitioners. Evidence was eligible if it was produced specifically for practitioners or when evidence was relevant to practitioners but other audiences such as the public also had access.
- Practitioners: these were postregistration health and social care professionals, collectively or as individual professions including but not limited to nurses, doctors, social workers, midwives, pharmacists, physiotherapists, occupational therapists, radiographers, and paramedics.
- Open social media: we defined open social media as internet-based social networking and media sharing platforms that allow any user to create and exchange user-generated content, making one-to-many posts and interacting by responding to others' posts. Our definition did not include mass media press articles, wikis, and blogs with no or limited facility for user interactions. Our definition also did not include purely communication-based apps, fee-paying, or closed, invite-only social media groups that could not be freely joined by any interested user. However, both noninteractional and closed social media groups were considered within our definition if they were highlighted on, or accessible via, open social media.
- Platforms: these were defined as open social networking and media sharing sites and apps, including but not limited to Facebook, YouTube, Instagram, WeChat, Tumblr, TikTok, Reddit, Twitter (subsequently rebranded X), and LinkedIn.
- Formats: these were a variety of media types, including but not limited to text, illustrative pictures, visual abstracts, infographics, videos, and podcasts.
- Strategies: these were the ways in which research evidence–related social media posts were delivered, including but not limited to a schedule of open sharing to the entire forum (frequency and timing), influencer endorsement, @mentions and #tagging, accessible special interest groups (eg, journal clubs), and live social media events (eg, tweet chats).
- Reach: this was defined as the number of practitioners reached by research evidence–related social media post or posts (eg, those following the social media account or participating in a social media event) or the social media analytics including the number of impressions (the number of times a post appears on social media feeds), views (the number of times a post is opened from social media feeds), or accesses (the number of times a post is accessed in any other way, eg, via a search engine).
- Engagement: this was measured by the number of positive responses (ie, likes) or interactions (such as shares, comments, reposting, or new posts) generated by a research evidence–related social media post.
- Direct dissemination: this was measured by the number of times an original piece of research evidence was accessed (eg, by link click from a social media post), viewed (eg, on an HTML web page), downloaded (eg, as a PDF document) or the altmetric attention score accumulated by original research articles.
- Impact: this included two discrete subcategories for the purposes of this review—(1) academic impact, the number of citations received by an original research evidence article or the journal impact factor, and (2) practical impact, measures of practitioners' changes in thinking or practice (eg, confidence, knowledge, or behavior change) after exposure to research evidence–related social media post or posts.

Information Sources

Six electronic databases were searched (MEDLINE [Ovid], PsycINFO [Ovid], CINAHL plus [EBSCO], and ERIC [EBSCO] as well as LISTA and OpenGrey). The date of the last search was January 10, 2023. Bibliographies of relevant reviews and included articles were searched for citations and PubMed, Elicit, and Google Scholar were used for reference harvesting.

Search Strategy

For full search strategies, refer to Multimedia Appendix 2 and the protocol by Roberts-Lewis et al [29]. Key search terms were grouped as follows:

- Practitioner groups, for example, health and social care staff and individual disciplines
- Research evidence, information, and knowledge
- Social media, network, web, sharing, and named platforms and formats
- Dissemination, reach, engagement, and impact
- Quantitative, evaluation, comparison, and named outcomes

Selection Process

Records from the electronic and citation searches were exported to EndNote Online (Clarivate) for deduplication and then imported to Rayyan software (Rayyan Systems) for title, abstract, and full-text screening.

XSL•FO
RenderX

Title and abstract screening were carried out by 2 independent reviewers (SRL and SQM). There was 92% agreement (κ=0.87) on eligibility decisions and 100% agreement after discussion.

Full-text screening was carried out by at least 2 of 5 independent reviewers (SRL, SQM, LB, HG, and FJL). There was 84% agreement (κ=0.83) on full-text inclusion decisions and 100% agreement after discussion.

Data Collection

Data from the included studies were extracted independently by at least 2 of 5 reviewers (see the *Selection Process* section) using a data extraction form developed a priori [29]. The accuracy of data extraction was confirmed by comparison between extraction forms, returning to the original article to resolve any disparity.

Data Items

The variables collected were study characteristics including the number and description of subjects; social media platforms, formats, and strategies; study design; comparisons; and outcomes. For each outcome of interest, means, SDs, and sample sizes were extracted for each comparison. When these data were missing, they were calculated from other reported statistics using recommended methods [31], where possible.

For studies that reported multiple outcome measures, only outcomes of interest were collected (reach, engagement, direct dissemination, and impact). Different measures for the same outcome were prioritized for inclusion in meta-analyses according to the a priori protocol [29]. Subsequent additions were made to the prioritization order to account for heterogeneous data reported in the included studies; these included aggregated total interactions, other types of post interactions, and the measurement time frame according to the most common time frames for each outcome of interest (Textbox 2).

Textbox 2. Summary of the outcomes of interest and their prioritization order for entry into meta-analyses.

Outcome and prioritizations

- Reach
  - Number of practitioners after 1 week
  - (1) Impressions, (2) views, and (3) accesses after 1 month
- Engagement
  - Number of positive responses (ie, likes) after 1 week
  - Number of post interactions—(1) total interactions (including shares, comments, and other interactions); (2) shares; (3) comments; (4) new posts, and (5) other post interactions—after 1 month
- Direct dissemination
  - (1) Link clicks and (2) article views after 1 month
  - Article downloads after 1 month
  - Altmetric attention score after 1 month
- Impact
  - (1) Citations and (2) impact factor after 1 year
  - Any measures of thinking or practice after any time frame

Risk of Bias Assessment

The 34-item (5-domain) Cochrane Risk of Bias 2.0 tool was used to rate the quality of the RCTs as lower risk of bias, some concerns, or higher risk of bias [32]. The Newcastle-Ottawa Scale (score range 0-9) was used to assess the risk of bias in nonrandomized designs [33]. A score of ≤3 was considered high risk of bias, scores between 4 and 6 were considered medium risk of bias, and a score of ≥7 was considered low risk of bias [33]. Risk of bias was assessed independently by at least 2 reviewers and data were checked for accuracy by a third reviewer.

Data Synthesis

The included studies were summarized narratively in text, tables, and figures. Quantitative comparisons were made using calculated standardized mean differences (SMDs), CIs, and *P* values for each comparison. SMD effect sizes were calculated using Hedges *g* to accommodate the heterogeneity of outcomes. Effect sizes of >0.8 were defined as large, ≥0.5 to 0.8 as moderate, and <0.5 as small [34]. Outcome effect sizes were presented as SMD, 95% CIs, *z*-test, and *P* value.

For outcomes where group means, SDs, and sample sizes were obtained from at least 2 studies, pooled effect sizes were calculated using random effects models in RevMan (version 5.3; The Cochrane Collaboration). The heterogeneity of pooled

XSL•FO
RenderX

data was assessed using I^2. Funnel plots were assessed visually for each meta-analysis to check for publication bias.

For pooled data with I^2>75%, subgroup analyses were planned; however, these were not possible due to the heterogeneous characteristics of the social media interventions in the included studies or an insufficient number of studies to achieve ≥80% statistical power [35]. Therefore, studies were ordered according to effect size, and the common characteristics of social media strategies in the studies with the largest effect sizes were narratively synthesized. Although no sensitivity analysis was planned a priori, evidence from RCTs and studies with low risk of bias was given greater weighting in the narrative synthesis than nonrandomized studies and those with high risk of bias.

Certainty Assessment

For each outcome, the certainty of the evidence base was evaluated based on the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) approach [36] and categorized as high, moderate, low, or very low [37,38].

Results

In total, 6461 records were identified, 555 full-text reports were screened, and 50 articles were included (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the process of study selection for a systematic review on the effectiveness of social media for dissemination of research evidence for health and social care practitioners, detailing a total of 6451 records identified (5896 excluded), 555 full-text reports screened (505 excluded), and 50 articles included.

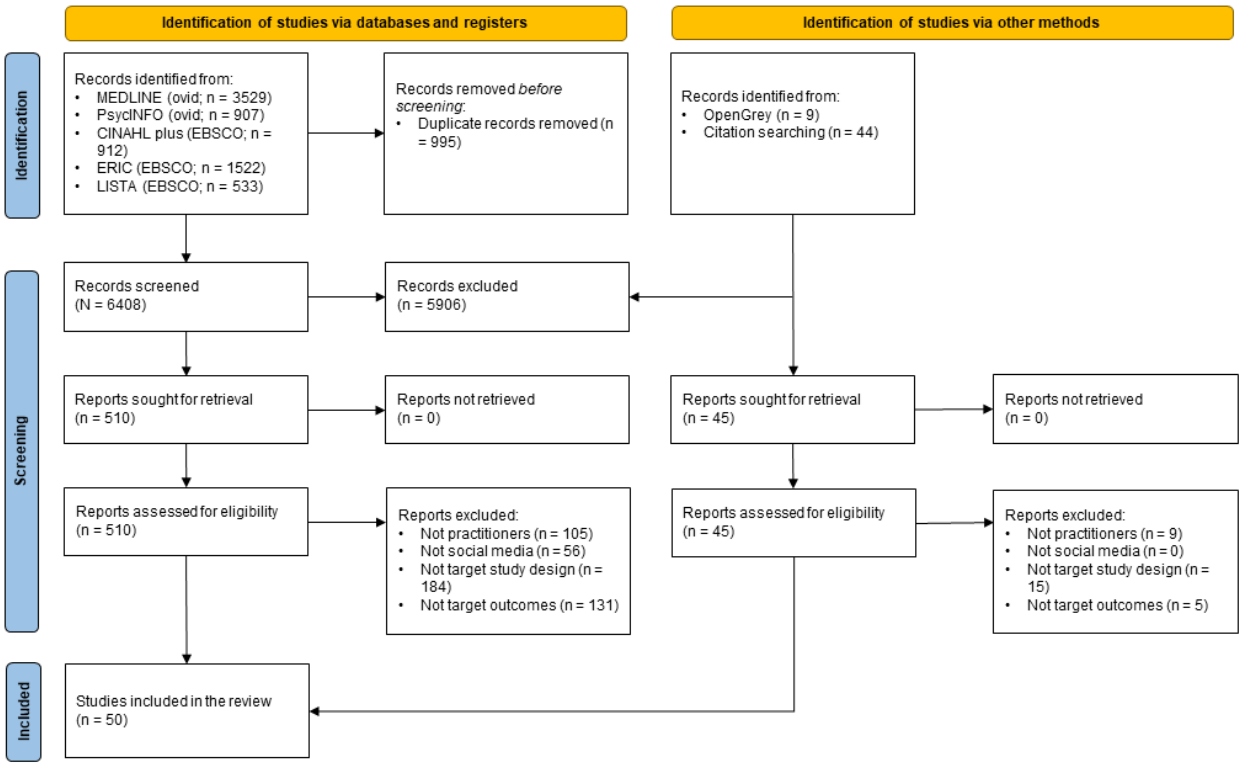

Included Studies

A total of 50 studies published between 2013 and 2022 were included; 9 were RCTs [39-48]. In total, 26 studies included nonrandomized comparisons [49-74], 12 studies were before-after comparisons [75-86], and 3 were case studies [87-89]. A total of 10 studies included both between-group comparisons and before-after analyses [40,44,45,47,51,53,60,63,70,74]. For study descriptions, refer to Multimedia Appendix 3 [39-89].

A total of 36 studies investigated the impact of social media on journal articles, with samples ranging from a single journal article [88] to 15,078 articles from multiple journals [54]. In total, 8 studies focused on research blogs [40,76,87] and microblogs [47,49,57,82,85], 4 studies examined research conference social media posts and hashtags [71,74,86,89], 2 studies investigated clinical guidelines [60,63], and 1 study tested research-related posts linked to health care hashtags [62].

Half of the studies explored multiple social media platforms and the other half of the studies examined a single platform. Twitter was used in all but 1 study [85], Facebook was used in 23 studies [40,42-44,47,48,53,57,60,63,64,67-70,76,77,80-82,85,89], LinkedIn was used in 8 studies [40,48,60,63,64,68,87,89], Instagram was used in 6 studies [57,62,67,68,76,87], and YouTube was used in 5 studies [58,60,63,64,87], whereas TikTok [87], Weibo [39], Google+ [64], Tumblr [81], and Spotify [87] were each used in 1 study. The most common media formats examined were text posts, which usually included links and images. Journal clubs or tweet chats were included in 9 studies [51,60,62,74,75,77,78,80,81], and video media were used in 6 studies [52,60,62,63,73,78].

XSL•FO
RenderX

The social media campaign duration ranged from 1 hour [51] to 5.5 years [54].

Outcomes measurement duration ranged from 3 days [86,89] to 4 years [64]. Typically, outcomes were measured after 1 month [42-44,48,52,53,56,60,63,67,75-81,83-85,87] or 1 to 2 weeks [39-41,45,47,51,68,74,82,88]. Citations were measured after ≥1 year, except in 1 study that reported citations after 6 months [61].

Risk of Bias

In total, 5 RCTs had low risk of bias [39,41-43,45,46], 3 RCTs had some concerns [40,44,48], and 1 RCT had high risk of bias [47] (Multimedia Appendix 3). The most common reasons for risk of bias included insufficient information provided about the allocation sequence, handling of missing data, or prioritization of multiple eligible outcome measurement time points. In total, 11 nonrandomized studies had low risk of bias [49,50,55-57,63,65,70,72,79,81]; 25 nonrandomized studies had moderate risk of bias [51-54,58,59,61,62,64,66-69,71,73-78,80,82-85]; and 5 nonrandomized studies had high risk of bias [60,86-89] (Multimedia Appendix 3). The most common reasons for risk of bias included targeted selection of studies for social media sharing and incomplete reporting of data handling. Funnel plots did not indicate a high risk of publication bias in pooled data.

Reach

Overview

In total, reach was evaluated in 26 studies (2 RCTs [41,45]). A total of 10 studies evaluated reach by reporting the numbers of practitioners receiving posts [41,45,66,67,75,78,79,84,86,89]. In total, 23 studies evaluated reach using social media analytics (17 in impressions [41,45,49,51,53,56,57,66,68,73,75,78,83,85,86,88,89]; 6 in views [62,63,76,79,81,87]; and none by reporting accesses).

Effects of Social Media Compared to No Social Media on Reach

There were insufficient studies comparing the reach of social media interventions versus no social media for pooled data analyses.

Evidence from individual studies included 1 RCT [45] with a low risk of bias that found a large between-group effect on the number of physicians reached by tweeted articles in a coordinated campaign, including a team with 12 social media accounts, 4 articles tweeted per day, and @mentions of authors and relevant institutions, compared to not tweeted articles (112 cardiothoracic surgery research articles; SMD 4.03, 95% CI 3.37-4.68; $P<.001$). Similarly, 1 nonrandomized study [63] with a low risk of bias reported a large between-group effect on views in favor of YouTube videos marketed by paid social media advertising on Facebook, Twitter, and LinkedIn, with relevant event hashtags, compared to video views in the absence of social media marketing (12 videos about tracheostomy safety; SMD 2.53, 95% CI 1.41-3.64; $P<.001$; Multimedia Appendix 3).

Within-Group Effects of Social Media on Reach

Pooled findings indicated large within-group effects on reach after social media interventions compared to before in both number of practitioners (SMD 2.03, 95% CI 0.97-3.10; $P<.001$; $I^2=53\%$; GRADE moderate) [75,78,79,84] and impressions or views (SMD 1.99, 95% CI 1.23-2.75; $P<.001$; $I^2=95\%$; GRADE moderate) [45,53,56,63,75,76,78,79,81,83,85]. The largest effects were reported in studies featuring social media marketing and scheduling tools [63,83] and multiple social media platforms (Twitter, Facebook, Instagram, and Tumblr) [63,76,81]; at least 1 post per day [76,79,81,83,84], including regular blogs [76,79,81] or microblogs [85]; posts coordinated with established live journal clubs, relevant events, hashtags, and @mentions [45,63,75,78,83]; and campaigns lasting 6 months to 4.5 years [76,81,85]. Smaller effects were reported by studies featuring one-off or less well-established tweet chats or events [75,78], 1 to 2 posts per month [53,56,75,78], and campaigns using a single social media platform (Twitter) [56,75,78,79] (Figure 2 [45,53,56,63,75,76,78,79,81,83-85]).

XSL•FO
RenderX

Figure 2. Meta-analyses of within-group effects after social media interventions on reach in (A) the number of practitioners and (B) impressions or post views.

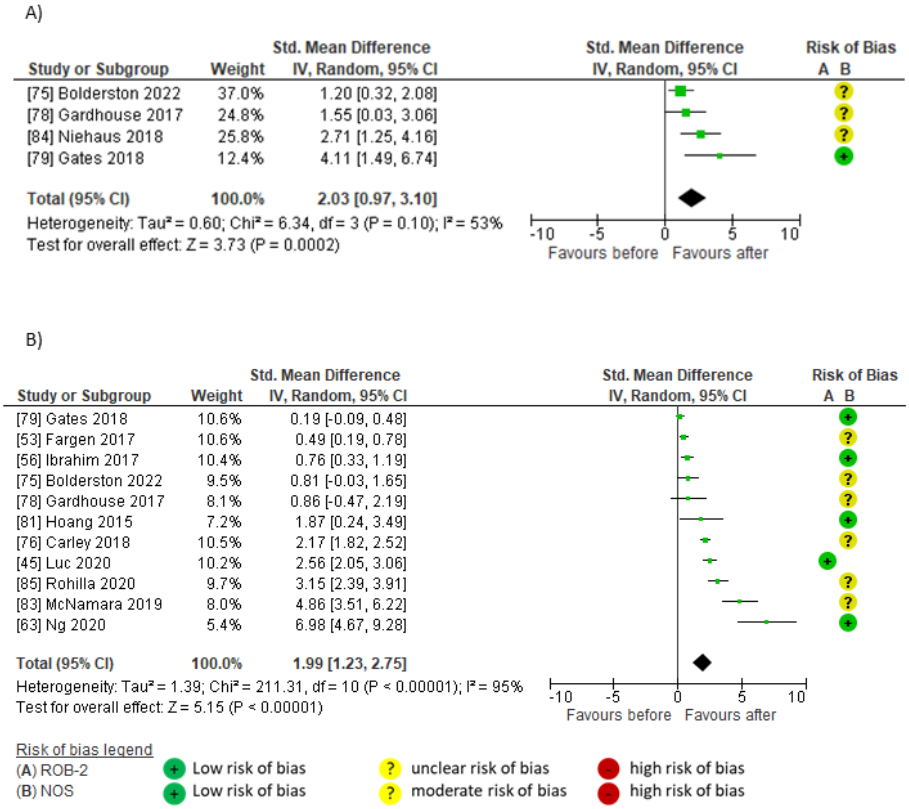

Between-Group Effects of Different Platforms, Formats, and Strategies on Reach

Pooled findings indicated a large between-group effect on impressions and views in favor of Twitter (vs Facebook and Instagram; SMD 1.87, 95% CI 1.54-2.21; P<.001; I^2=0%; GRADE low) and Facebook (vs Instagram; SMD 1.19, 95% CI 0.64-1.75; P<.001; I^2=46%; GRADE low) [57,68]. However, no effect was shown between platforms in the number of practitioners who were followers on Twitter, Facebook, and Instagram [67] (Multimedia Appendix 3).

Pooled findings showed a large effect on impressions in favor of posts with images, in particular, infographics, compared to no images (SMD 1.63, 95% CI 0.04-3.22; P=.04; I^2=95%; GRADE low; Multimedia Appendix 3) [41,45,49,56].

Pooled findings indicated a large effect on reach in favor of strategies using social media influencers or organizations compared to using standard social media user accounts (SMD 1.02, 95% CI 0.04-1.99; P=.04; I^2=100%; GRADE low) [66,73]. One RCT follow-up study [46] also reported that tweeting at 1 PM (EST, United States) generated the highest reach and tweeting at 9 PM generated the lowest reach to physicians (Multimedia Appendix 3).

Engagement

A total of 21 studies (including 3 RCTs [41,45,47]) evaluated engagement (6 studies examined likes [45,47,49,51,83,88]; 11 studies investigated total engagements, including shares, comments, and other interactions [41,45,49,51,53,54,67,71-73,78]; 7 studies assessed only post shares [47,49,52,56,68,83,84]; and 2 studies reported on other post interactions only [63,75]). No included studies evaluated comments or reposts alone.

Effects of Social Media Compared to No Social Media on Engagement

There were insufficient studies comparing the engagement of social media interventions versus no social media for pooled data analyses.

Evidence from individual studies included just 1 nonrandomized study [63] with a low risk of bias that reported a large between-group effect on interaction time spent watching YouTube videos marketed by paid social media advertising compared to video interaction time in the absence of social media marketing (12 videos about tracheostomy safety; SMD 2.36, 95% CI 1.27-3.44; P<.001).

Within-Group Effects of Social Media on Engagement

Pooled findings indicated large within-group effects on engagement after social media interventions compared to before. Effects were significant for post interactions (SMD 3.74, 95% CI 2.02-5.46; P<.001; I^2=96%; GRADE moderate) [45,53,63,75,78,83,84] but not for likes (SMD 3.18, 95% CI –0.25 to 6.62; P=.07; I^2=98%; GRADE low) [45,83,85]. The largest effects on engagement were evident in social media campaigns established over 3 to 18 months, usually featuring coordinated, paid social media strategies; daily posts; visually appealing formats; topical hashtags; and @mentions targeting relevant organizations, government resources, and events [53,63,83-85]. Large effects on engagement were also observed

XSL•FO
RenderX

in studies featuring a series of live journal clubs [75,78]. Smaller effects of social media on engagement were reported by 1 RCT [45] that used a 14-day Twitter campaign (Figure 3 [45,53,63,75,78,83-85]).

Figure 3. Meta-analyses of within-group effects after social media interventions on engagement in (A) positive responses (likes) and (B) post interactions (total interactions, shares [retweets], comments [replies], or other post interactions).

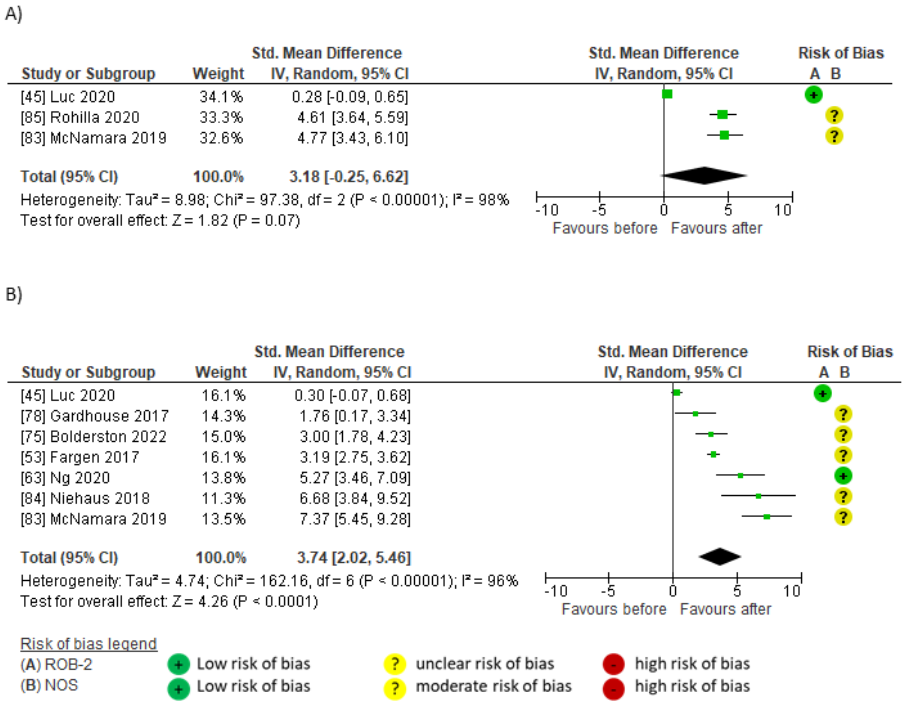

Between-Group Effects of Different Platforms, Formats, and Strategies on Engagement

Pooled findings indicated a large between-group effect on engagement in favor of Twitter (vs Facebook, Instagram, and LinkedIn; SMD 1.15, 95% CI 0.21-2.10; P=.02; I^2=79%; GRADE low; Multimedia Appendix 3) [47,67,68].

Pooled findings showed large between-group effects on engagement in favor of posts with images compared to no images. Effects were significant for interactions (SMD 1.24, 95% CI 0.53-1.96; P<.001; I^2=98%; GRADE low) but not for likes (SMD 0.87, 95% CI –0.40 to 2.14; P=.18; I^2=88%; GRADE low; Multimedia Appendix 3) [41,45,49,52,56,72].

Pooled findings indicated a small effect on post interactions of social media strategies with participation by influential others (including patients, authors, and non–peer-reviewed news sources; SMD 0.26, 95% CI 0.13-0.39; P<.001; I^2=82%; GRADE low) [54,71,73]. Evidence from individual studies also showed large effects of social media influencers with >1000 followers [52], morning and weekday posting [72], and hashtags [72] and a small effect of @mentions [52] (Multimedia Appendix 3).

Direct Dissemination

In total, 23 studies (including 8 RCTs [39-45,48]) evaluated direct dissemination (10 in link clicks [41,45,49,51,53,56,73,79,80,83], 13 studies reported article views [39,40,42-44,48,63,69,70,74,81,84,87], 8 measured PDF downloads [40,44,48,50,63,69,79,81], and 9 assessed the altmetric score [44,45,51,69,74,77,79,83,87]).

Effects of Social Media Compared to No Social Media on Direct Dissemination

Pooled data showed large between-group effects of social media on direct dissemination. Effects were significant for link clicks or article views (SMD 0.88, 95% CI 0.15-1.62; P=.02; I^2=95%; GRADE moderate) [39,40,42,43,45,48] and article downloads (SMD 1.25, 95% CI 0.86-1.65; P<.001; I^2=0%; GRADE high) [40,48,50] but not for the altmetric attention score [45,74] (SMD 1.48, 95% CI –1.00 to 3.96; P=.24; I^2=97%; GRADE low; Figure 4 [39,40,42,43,45,48,50,74]). Studies that reported the largest effects of social media on direct dissemination used campaigns including professional social media marketing and scheduling tools (Social Bro, Hootsuite, Sprinkler, and Spredfast); multiple posts per day; multiple platforms (Twitter, Weibo, Facebook, and LinkedIn); or multiple accounts on 1 platform, link, and blog [39,40,45,48,50]. The studies that showed the smallest effects of social media on direct dissemination used social media campaigns that posted less than once a day (approximately 0.36 [42] and 0.30 [43] posts per day or once a month publicizing a Twitter journal club [74]).

XSL•FO
RenderX

Figure 4. Meta-analyses of between-group effects of social media compared to no social media on direct dissemination in (A) link clicks and article views, (B) article downloads, and (C) the altmetric attention score.

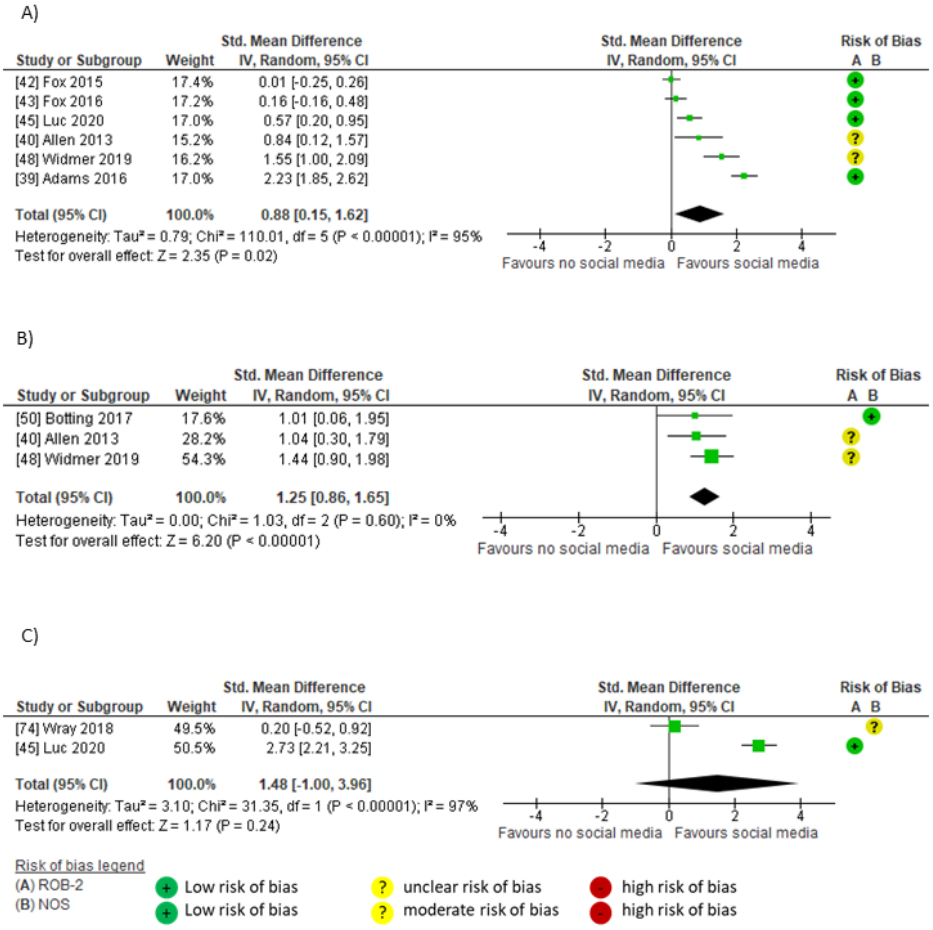

Within-Group Effects of Social Media on Direct Dissemination

Pooled findings indicated a large, significant within-group effect on direct dissemination after social media interventions compared to before in link clicks or views (SMD 1.93, 95% CI 1.23-2.62; P<.001; I^2=92%; GRADE high) [40,44,45,53,63,70,79-81,83,84], article downloads (SMD 0.82, 95% CI 0.26-1.37; P=.004; I^2=49%; GRADE moderate) [40,44,63,79,81], and altmetric attention score (SMD 1.92, 95% CI 0.75-3.09; P=.001; I^2=92%; GRADE moderate) [44,45,51,74,77,79,83]. Studies that reported the largest effects used campaigns that included coordinated or paid social media software [63,83,84], posting at least once a day [45,63,83,84], visually attractive elements and links [44,45,63,83,84], hashtags, @mentions [45,83,84], multiple platforms (including Twitter and Facebook) [40,44,63,80,81], multiple accounts [40,44,45], associated blogs [40,44,45,79,81], podcasts [81], and tweet chats [51,80] (Figure 5 [40,44,45,51,53,63,70,74,77,79-81,83,84]).

XSL•FO
RenderX

Figure 5. Meta-analyses of within-group effects after social media interventions on direct dissemination in (A) link clicks or article views, (B) article downloads, and (C) the altmetric attention score.

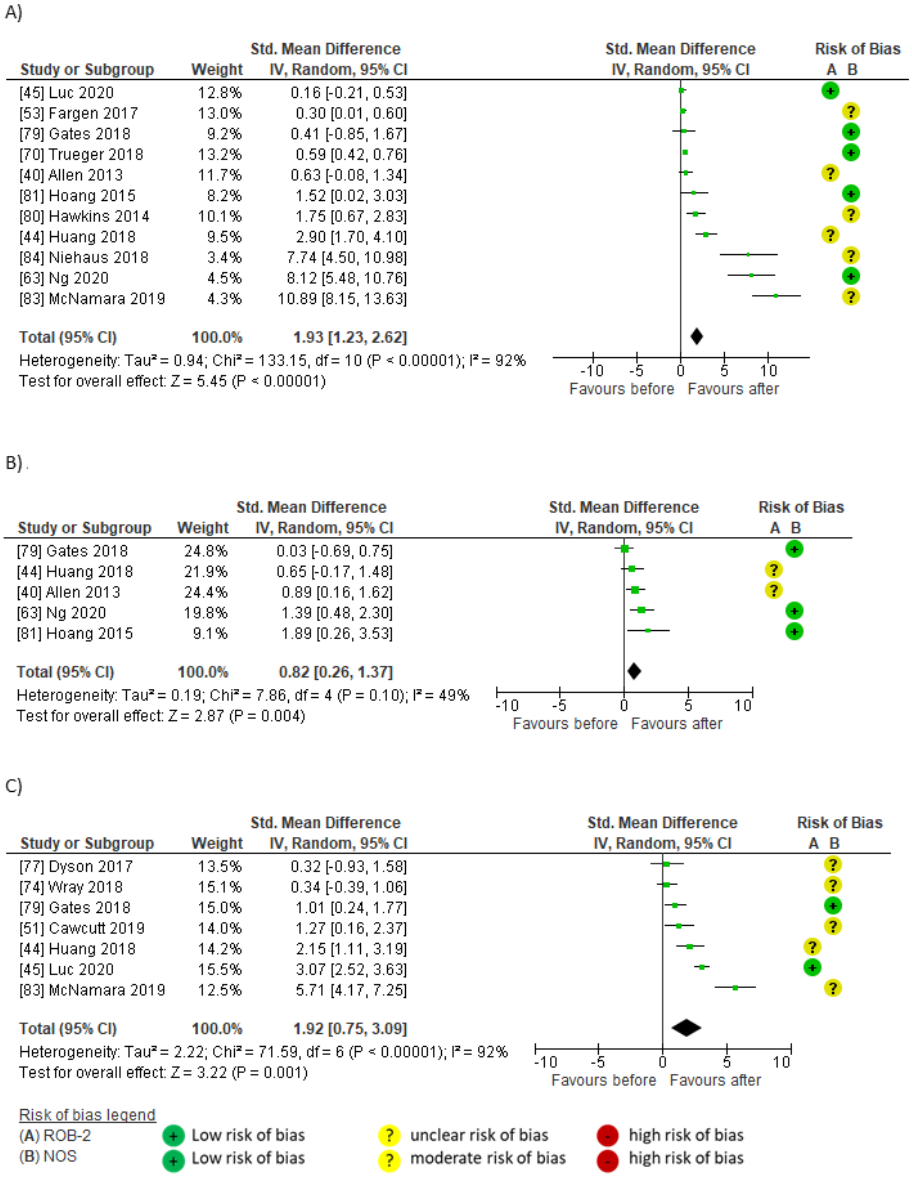

Between-Group Effects of Different Platforms, Formats, and Strategies on Direct Dissemination

Pooled findings showed no significant effect of the platform on direct dissemination (SMD 0.92, 95% CI –1.21 to 3.04; *P*=.40; I^2=99%; GRADE low); both Twitter and Facebook appeared effective [48,70] (Multimedia Appendix 3).

Pooled findings indicated large effects in favor of posts with images, particularly infographics, compared to no images. Effects were significant for link clicks or article views (SMD 1.18, 95% CI 0.27-2.10; *P*=.01; I^2=88%; GRADE low) [41,44,45,56,69] and for altmetric attention score (SMD 1.19, 95% CI 0.04-2.35; *P*=.04; I^2=83%; GRADE low) [44,45,69] but not for article downloads (SMD 0.26, 95% CI –0.32 to 0.83; *P*=.38; I^2=0%; GRADE low) [44,69]. Evidence from individual studies also highlighted large, significant effects of podcasts compared to infographics and standard posts on direct dissemination [69] and a positive effect on link clicks of posts with links compared to posts with infographics [49] (Multimedia Appendix 3).

There were insufficient studies comparing the effect of social media strategies on direct dissemination for pooled data analyses. Evidence from individual studies showed a large, significant effect on link clicks in favor of posting on Tuesdays, Wednesdays, and Saturdays compared to the other days of the week [53] and no effects of time of year [42,43] or non–peer-reviewed news source involvement [73] (Multimedia Appendix 3).

Impact

A total of 13 studies (including 3 RCTs [46-48]) evaluated impact (7 assessed article citations [46,48,54,55,58,59,65], 3 investigated impact factors [54,61,64], and 4 examined changes in thinking or behavior [47,57,60,82]).

XSL•FO
RenderX

Effects of Social Media Compared to No Social Media on Impact

Pooled findings indicated a moderate between-group effect of social media compared to no social media on citations (SMD 0.76, 95% CI 0.49-1.03; P<.001; I^2=79%; GRADE moderate) [46,48,55,58,59,61,64,65] and thinking and practice (SMD 0.65, 95% CI 0.37-0.93; P<.001; I^2=55%; GRADE low) [57,60]. The largest effects on impact were shown in studies that used social media interventions, including links and relevant @mentions [46,55,57-61] in campaigns that often had relatively short durations (14 days [46] to ≤12 months [57-61]). Larger effects on citations were shown in studies sharing articles on broad topics (such as urology or surgery) [46,55,58,59,61], whereas effects on knowledge and practice were evident in social media campaigns that were focused on a specialist topic (eg, persistent genital arousal disorder [57] and complementary and alternative medicine in multiple sclerosis [60]). Studies that showed smaller between-group effects of social media on impact either described social media intervention with infrequent posts on social media (7 times per month, approximately 0.3 posts per day) [48] or without indicating post frequency [64,65] (Figure 6 [46,48,55,57-61,64,65]).

Figure 6. Meta-analysis of the between-group effect of social media interventions compared to no social media interventions on impact in (A) citations and (B) thinking and practice.

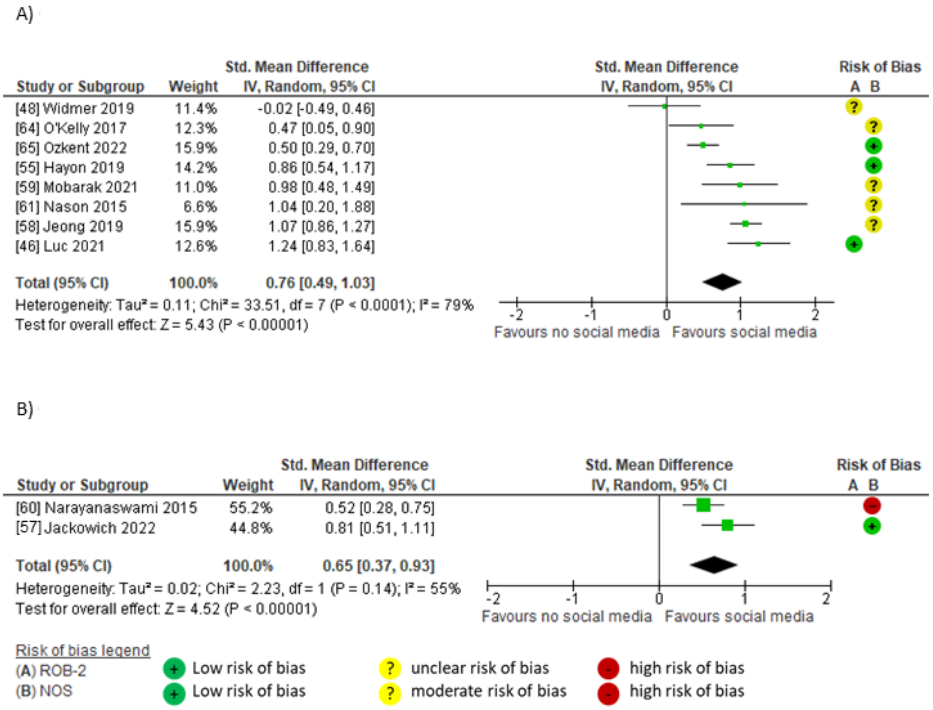

Within-Group Effects of Social Media on Impact

There were insufficient studies comparing the impact on citations after social media interventions compared to before; however, pooled findings showed a small within-group effect on thinking and practice after social media interventions (SMD 0.45, 95% CI 0.07-0.83; P=.02; I^2=91%; GRADE low) [47,60,82]. One RCT [47] and 1 nonrandomized study [82], both concerning tendinopathy practice points with links to research articles or evidence-based podcasts shared for 2 weeks on Twitter and Facebook, reported differing effect sizes on thinking and practice (small and nonsignificant in the RCT [47] but moderate and significant in the nonrandomized study [82]). Another nonrandomized study [60] also reported a moderate effect of a paid targeted social media advertising campaign on Twitter, Facebook, LinkedIn, and YouTube that included article links, images, videos, podcasts, and a live tweet chat with a prominent organization that impacted knowledge, attitudes, and behavior regarding complementary and alternative medicine in multiple sclerosis (Figure 7 [47,60,82]).

XSL•FO
RenderX

Figure 7. Meta-analysis of within-group effect after social media interventions on thinking and practice.

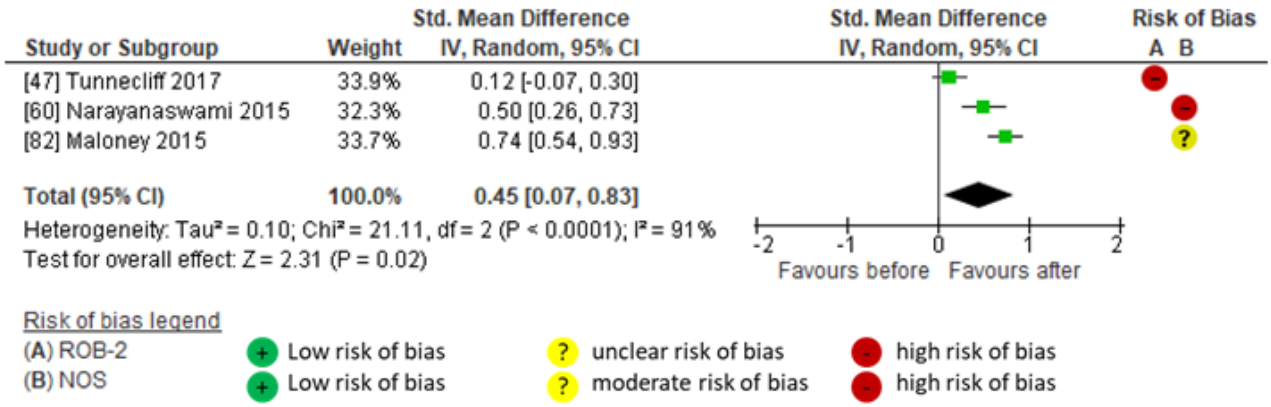

Between-Group Effects of Different Platforms, Formats, and Strategies on Impact

There were insufficient studies comparing the effect of social media platforms on impact for pooled data analyses. Evidence from individual studies showed a large, significant effect on the journal impact factor of posting on ≥3 social media platforms [64]; a positive association between the journal impact factor and the number of social media platforms used [59]; and a small, nonsignificant effect of platform in favor of Twitter (vs Facebook) on knowledge and practice changes [47] (Multimedia Appendix 3).

No included studies compared the effect of different formats on impact.

Pooled findings indicated a large, nonsignificant effect on citations of author tweets in addition to standard journal social media strategies (SMD 1.00, 95% CI –0.84 to 2.84; P=.29; I^2=98%; GRADE very low; Multimedia Appendix 3) [54,55].

Certainty

Considering all the evidence, there was high certainty that social media is effective for the dissemination of research to health care practitioners. The outcome with the highest certainty across all comparisons was direct dissemination. Evidence was insufficient to determine the size and certainty of between-group effects of social media compared to no social media on reach and engagement; however, there was moderate certainty of large and moderate effects of social media interventions on direct dissemination and impact, respectively. After social media exposure, there was moderate certainty of large, positive within-group effects on reach, engagement, and direct dissemination, whereas there was low certainty of a small effect on impact. Certainty was generally low regarding the size of the effects of platforms, formats, and strategies on each outcome. However, the direction of effects was consistently in favor of using multiple platforms (particularly Twitter and Facebook), using images (particularly infographics), and involving influential others in social media campaigns.

The level of certainty about the size of the effects in favor of social media was different depending on the outcome of interest and study characteristics. There was a tendency for smaller effect sizes in RCTs and studies with less-intensive social media interventions. The certainty ratings were lowered for all comparisons due to the variability of the included study designs, many of which were descriptive and not designed for rigorous quantitative evaluations, meaning most included studies had moderate risk of bias. Consistency and precision of effect size estimates were also threatened by the low number and heterogeneity of studies included in some comparisons. However, for some comparisons, certainty was uprated due to the large magnitude of estimated effects and potential dose-response gradients between the intensity of social media interventions and their effectiveness.

Discussion

Principal Findings

There was evidence of the effectiveness of social media for the dissemination of research evidence to health care practitioners. All the included studies reported some findings in favor of social media, although there was considerable heterogeneity in effect sizes, and study quality was mixed. Effect sizes of social media effectiveness were influenced by the frequency, intensity, and composition of social media interventions. Effectiveness was enhanced by the use of multiple social media platforms (including Twitter and Facebook); multiple social media accounts; ≥1 social media post per day; appealing formats (including infographics or other visual media, blogs, and links to articles); professional social media marketing and scheduling tools and involving relevant and influential people, organizations, and events in social media campaigns.

Our findings that social media was beneficial for the dissemination of research evidence to practitioners concurred with existing literature about the largely positive impact of social media on dissemination [10,21-23,28,65,90-93]. Quantitative analyses in other studies have revealed positive correlations between social media use and the dissemination and impact of health research evidence [65,91-93]. Narrative and qualitative reviews have highlighted benefits of social media for clinicians including connectedness and network accessibility to all (particularly with increasing use of communication technology and mobile apps in practice and day-to-day life), the large audience of practitioners and policy makers that uses

XSL•FO
RenderX

social media (particularly for information acquisition and educational purposes), and the effectiveness of social media to deliver clinical guidelines and research evidence–based information that could be implemented in practice [21-23,28,90]. Reviews also highlighted the challenges of synthesizing social media effectiveness for the dissemination of information due to the heterogeneity of how social media are studied and used [22,28,90].

Social media campaigns to disseminate research evidence to health care practitioners should consider target outcomes because specific features enhanced the 4 outcomes of interest differently (Table 1).

Our quantitative findings in favor of using multiple social media platforms, including Twitter (for engagement in particular) and Facebook, concurred with existing literature that recommends using a range of social media platforms [23,28,90,94] and highlights the prominence of Twitter for research dissemination [28] and Twitter and Facebook for e-professionalism [22]. New knowledge from our systematic review includes the potentially beneficial effect of using multiple social media accounts on the same platform. This may also link with the apparent dose-response relationship between effectiveness and the average number of social media posts per day.

There was consistent evidence to support the use of visual media, particularly infographics for better reach, engagement, and direct dissemination; our findings suggest that social media effectiveness may also be enhanced by other post formats including podcasts, blogs, questions, and *practice points* to disseminate research evidence on social media. This adds to previous literature, which recommended posting a range of appealing multimedia that is accessible, useful, relevant, authentic, and credible [10,23,28,90]. Our review also highlighted the importance of including links to original research; comprehensive infographics or practice points posted on social media might reduce the likelihood of viewing the original article by link click [49]. Nevertheless, including links offers the viewer the opportunity to check the authenticity and credibility of the information in a post. Including links may also facilitate the delivery of simple, clear, and practice-relevant messages without scientific language [88] because the main message from the research can be easily understood and accessible to all practitioners, while further details can be sought by accessing the link.

Our findings corroborate existing recommendations to identify and involve key influencers, organizations, events, communities, #hashtags, and @mentions and to use professional tools to plan sustained, scheduled, and regular posts to overcome the transient nature of social media [10,23,28,90]. Our systematic review extends existing recommendations by identifying that posting at least once a day on average (often achieved using multiple accounts, platforms or both) was more effective than less-intensive social media strategies, suggesting a dose-response relationship between post frequency and social media effectiveness. This dose-response relationship may explain the negligible effect of one journal's social media campaign tested in 2 RCTs included in this systematic review [42,43].

The optimal timing of social media posts and campaigns for the dissemination of research evidence to practitioners is contentious in existing literature [53,95,96]. Optimal timing may be outcome dependent; in this review, different times of the day and different days of the week were more effective depending on the outcome measured (Table 1). Similarly, while reach and engagement may be enhanced by established social media initiatives extending over months or years, impact may be best achieved in shorter, targeted social media campaigns over days or weeks.

Our findings suggested a tendency for greater impact on thinking and practice from social media featuring targeted topics or specialist areas of health care. This concurred with previous reviews that found that the dissemination of clinical guidelines or clear, evidence-based behavior change messages, such as *practice points*, may increase impact with health care practitioners [90,94]. This also resonates with recommendations in the existing literature to consider the target audience when selecting platforms, formats, and strategies that optimize content for the dissemination of research evidence [23,88,90,94]. For instance, the choice of platform and content about pediatric colorectal cancer should be guided by the understanding that, in this field, Twitter is typically used to share research evidence, Facebook is used for support offered by nonprofit organizations, and Instagram is used for sharing personal stories [67].

Our findings suggested a tendency that social media was more effective on impact (in citations) when used to share research articles about broad topics. Thus, there may be interactions between the content of research evidence posts and the effectiveness of dissemination on social media, regardless of platform, format, or strategy. Indeed, there are indications in the wider literature that research evidence source, topic, and post content may influence how effectively it can be disseminated on social media. For example, geographically, compared to the United States, authors from Europe, and UK clinical guidelines, achieve better altmetric attention scores and citation rates [58,97]. Published articles may achieve greater reach, research conference posts may receive greater engagement [73], and clinical guidelines may better influence thinking and practice [90]. Research evidence that aligns with "hot topics" on social media may also achieve better reach and engagement [68,98]. Engagement, direct dissemination, and citation rates may be better for open-access articles (especially reviews) that are recently published with shorter titles (which are provocative, interrogative, or declarative and free from methodological description), with a greater number of authors, and by higher impact factor journals (often with a larger social media presence in terms of followers and number of tweets per month) [54,58,59,65,99-101]. Depending on the health topic, posts with humor, shock value, inaccuracies, rumors, or emotional content might achieve better reach [66,102,103]; posts with practical guidance may be more likely to be shared [66,102,104,105]; and short videos with positive titles might receive more likes and comments [104].

XSL•FO
RenderX

Table 1. Summary of the social media characteristics that may enhance effectiveness depending on the outcome.

Outcomes	Reach	Engagement	Direct dissemination	Impact
Platforms	• Multiple platforms (including Twitter and Facebook)	• Twitter	• Multiple platforms (including Twitter and Facebook)	• ≥3 platforms
Format	• Images (particularly infographics) • Blogs • Microblogs	• Images (particularly infographics) • Appealing media (eg, videos, enticing statements, or questions)	• Images (particularly infographics) • Podcasts • Blogs • Links to articles	• Clinically relevant and useable posts • Microblogs • Links
Strategies	• Involvement of influencers and organizations	• Involvement of influencers, patients, authors, and organizations • Relevant @mentions and hashtags	• Intensive strategies • Relevant @mentions and hashtags	• Targeting specific audiences or specialism for practice • Broader topics for citations • Relevant @mentions
Timing	• ≥1 posts per day • Established campaigns sustained over months or years	• Scheduled posts (eg, weekday mornings) • Established campaigns coinciding with targeted events and government resources	• ≥1 posts per day • Any time of the year • Scheduled posts (eg, on Tuesdays, Wednesdays, and Saturdays)	• Frequent posts • Brief but focused campaigns (eg, a year or less)
Resources	• Multiple social media accounts • Social media marketing tools	• Regular established live events (eg, journal clubs and tweet chats) • Paid social media advertising	• Multiple social media accounts • Social media marketing and scheduling tools	• Paid social media advertising • Live events involving prominent organizations

Strengths and Limitations

The key strengths of this systematic review are the quantitative meta-analytical methods used that allow robust conclusions based on cumulative evidence about the effective use of social media for the dissemination of research evidence to practitioners. Methodological limitations have been discussed in the published protocol [29]. The search strategy was comprehensive, using multiple reviewers to ensure reliability and comprising a range of study designs including RCTs. However, no studies were identified investigating the effectiveness of social media for the dissemination of research evidence to social care practitioners. There were also relatively few RCTs, and the mixed quality of the included studies reduced the certainty of evidence about effect sizes for some outcomes. A wide range of social media interventions and research evidence content were represented by the included studies; this reduced the risk of confounding by topic, source, or content. However, time may have reduced the consistency of findings between included studies because the research and social media landscapes are rapidly evolving and have changed significantly from 2010 to 2023 [11,14]. Furthermore, the time frame of outcome measurement can influence potentially transient social media effectiveness [55]. Effectiveness was evaluated thoroughly in 4 outcome domains, comprising 9 outcomes of interest and multiple measures; this helped to accommodate the variability of reporting and design in the included studies. Although the diversity of the included studies threatened the consistency of effect sizes, the direction of effect in favor of social media was consistent. The heterogeneity of findings was ameliorated using a random effects model for more conservative estimates of effect size than a fixed effects model. Using 4 outcome domains added nuance to the existing understanding and facilitated the development of clear suggestions about how to optimize social media effectiveness for the dissemination of research evidence to health care practitioners. However, a cautious interpretation of a causal relationship between dissemination effectiveness and specific social media tactics is required. Effects may have been inflated by other confounders; for example, larger organizations with greater resources for public relations not only can post more frequently on social media but also may have greater reputational influence and share higher-quality research.

Conclusions

In conclusion, social media were effective for disseminating research evidence to health care practitioners. Large effects of and after social media interventions on measures of direct dissemination were particularly evident. There may be a dose-response relationship between the intensity of the social media campaign and its effectiveness. Selected social media intervention characteristics including platforms, formats, and strategies may enhance reach, engagement, direct dissemination, and impact of research evidence for practitioners. Future research directions include repetition of this review to keep up with the rapidly evolving use of social media for research dissemination; quantitative testing of the potential dose-response relationship between dissemination effectiveness and social media frequency and intensity; and further evaluation and exploration of how different practitioner groups, particularly

XSL•FO
RenderX

social care practitioners, use social media to access research evidence.

Acknowledgments

The authors would like to acknowledge the support of the National Institute of Health and Care Research (NIHR) and St George's University of London. This study presents independent research funded by NIHR evidence (2022/01). The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health.

Data Availability

The data sets generated and analyzed during this study will be available from the corresponding author on reasonable request. Findings from this systematic review will be disseminated at academic conferences and promoted using a range of social media strategies.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e51418_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 31 KB - jmir_v26i1e51418_app2.docx]

Multimedia Appendix 3
Description of the included articles and full details of meta-analyses.
[DOCX File , 628 KB - jmir_v26i1e51418_app3.docx]

References

1. NIHR launches new centre for engagement and dissemination. National Institutes for Health and Care Research. URL: https://www.nihr.ac.uk/news/nihr-launches-new-centre-for-engagement-and-dissemination/24576 [accessed 2024-04-03]
2. Embedding a research culture. National Institutes for Health and Care Research. URL: https://www.nihr.ac.uk/health-and-care-professionals/engagement-and-participation-in-research/embedding-a-research-culture.htm [accessed 2023-05-01]
3. Saving and improving lives: the future of UK clinical research delivery. Department of Health & Social Care, Government of UK. 2021 Mar. URL: https://www.gov.uk/government/publications/the-future-of-uk-clinical-research-delivery/saving-and-improving-lives-the-future-of-uk-clinical-research-delivery): [accessed 2024-04-03]
4. Hayes C. Approaches to continuing professional development: putting theory into practice. Br J Nurs 2016 Aug 11;25(15):860-864. [doi: 10.12968/bjon.2016.25.15.860] [Medline: 27523759]
5. What is evidence. Chartered Society of Physiotherapy. URL: https://www.csp.org.uk/professional-clinical/clinical-evidence/evidence-based-practice/what-evidence [accessed 2024-04-03]
6. Castro-Sánchez E, Holmes AH. Impact of organizations on healthcare-associated infections. J Hosp Infect 2015 Apr;89(4):346-350 [FREE Full text] [doi: 10.1016/j.jhin.2015.01.012] [Medline: 25726435]
7. How to disseminate your research. National Institute of Health and Care Research. 2019. URL: https://www.nihr.ac.uk/documents/how-to-disseminate-your-research/19951 [accessed 2022-11-23]
8. Henriksen D, Mishra P. Innovations in the dissemination of action research: rhetoric, media, and communication. In: Mertler CA, editor. The Wiley Handbook of Action Research in Education. Hoboken, NJ: John Wiley and Sons; 2019.
9. Millar BC, Lim M. The role of visual abstracts in the dissemination of medical research. Ulster Med J 2022 May;91(2):67-78 [FREE Full text] [Medline: 35722208]
10. Zhao J, Harvey G, Vandyk A, Huang M, Hu J, Modanloo S, et al. Understanding how and under what circumstances social media supports health care providers' knowledge use in clinical practice: a realist review. Telemed J E Health 2023 Apr;29(4):475-500 [FREE Full text] [doi: 10.1089/tmj.2022.0213] [Medline: 35994025]
11. Ortiz-Ospina E. The rise of social media. Our World in Data. 2019. URL: https://ourworldindata.org/rise-of-social-media [accessed 2024-04-03]
12. Galbraith K, Ward A, Heneghan C. A real-world approach to evidence-based medicine in general practice: a competency framework derived from a systematic review and Delphi process. BMC Med Educ 2017 May 03;17(1):78 [FREE Full text] [doi: 10.1186/s12909-017-0916-1] [Medline: 28468646]

XSL•FO
RenderX

13. Bwanga O. Barriers to continuing professional development (CPD) in radiography: a review of literature from Africa. Health Prof Educ 2020 Dec;6(4):472-480 [FREE Full text] [doi: 10.1016/j.hpe.2020.09.002]
14. Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit Soc Sci Commun 2021 Oct 07;8(1):224 [FREE Full text] [doi: 10.1057/s41599-021-00903-w]
15. Tunnecliff J, Ilic D, Morgan P, Keating J, Gaida JE, Clearihan L, et al. The acceptability among health researchers and clinicians of social media to translate research evidence to clinical practice: mixed-methods survey and interview study. J Med Internet Res 2015 May 20;17(5):e119 [FREE Full text] [doi: 10.2196/jmir.4347] [Medline: 25995192]
16. Zhu Z, Xing W, Hu Y, Zhou Y, Gu Y. Improving evidence dissemination and accessibility through a mobile-based resource platform. J Med Syst 2018 May 28;42(7):118. [doi: 10.1007/s10916-018-0969-7] [Medline: 29808443]
17. Using social media platforms to disseminate Cochrane evidence in China. Cochrane Community. URL: https://community.cochrane.org/news/using-social-media-platforms-disseminate-cochrane-evidence-china [accessed 2022-10-21]
18. Chambers CT, Dol J, Parker JA, Caes L, Birnie KA, Taddio A, et al. Implementation effectiveness of a parent-directed YouTube video ("it doesn't have to hurt") on evidence-based strategies to manage needle pain: descriptive survey study. JMIR Pediatr Parent 2020 Mar 04;3(1):e13552 [FREE Full text] [doi: 10.2196/13552] [Medline: 32130190]
19. Korda H, Itani Z. Harnessing social media for health promotion and behavior change. Health Promot Pract 2013 Jan;14(1):15-23. [doi: 10.1177/1524839911405850] [Medline: 21558472]
20. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
21. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
22. Giustini D, Ali SM, Fraser M, Kamel Boulos MN. Effective uses of social media in public health and medicine: a systematic review of systematic reviews. Online J Public Health Inform 2018;10(2):e215 [FREE Full text] [doi: 10.5210/ojphi.v10i2.8270] [Medline: 30349633]
23. Elliott SA, Dyson MP, Wilkes GV, Zimmermann GL, Chambers CT, Wittmeier KD, et al. Considerations for health researchers using social media for knowledge translation: multiple case study. J Med Internet Res 2020 Jul 23;22(7):e15121 [FREE Full text] [doi: 10.2196/15121] [Medline: 32706653]
24. Johannsson H, Selak T. Dissemination of medical publications on social media - is it the new standard? Anaesthesia 2020 Feb 23;75(2):155-157 [FREE Full text] [doi: 10.1111/anae.14780] [Medline: 31338831]
25. Farsi D. Social media and health care, part I: literature review of social media use by health care providers. J Med Internet Res 2021 Apr 05;23(4):e23205 [FREE Full text] [doi: 10.2196/23205] [Medline: 33664014]
26. Heaton-Shrestha C, Hanson K, Quirke-McFarlane S, Delaney N, Vandrevala T, Bearne L. Exploring how members of the public access and use health research and information: a scoping review. BMC Public Health 2023 Nov 07;23(1):2179 [FREE Full text] [doi: 10.1186/s12889-023-16918-8] [Medline: 37936117]
27. Dol J, Tutelman PR, Chambers CT, Barwick M, Drake EK, Parker JA, et al. Health researchers' use of social media: scoping review. J Med Internet Res 2019 Nov 13;21(11):e13687 [FREE Full text] [doi: 10.2196/13687] [Medline: 31719028]
28. Farris O, Todd A, Tomlin A, Badenoch D, Lamb D. #Dissemination: a systematic review of the role of social media in circulating healthcare research (pre-print submitted to but not published by JMIR). JMIR Preprints Preprint posted online December 24, 2021 [FREE Full text] [doi: 10.2196/preprints.19686]
29. Roberts-Lewis SF, Baxter HA, Mein G, Quirke-McFarlane S, Leggat FJ, Garner HM, et al. The use of social media for dissemination of research evidence to health and social care practitioners: protocol for a systematic review. JMIR Res Protoc 2023 May 12;12:e45684 [FREE Full text] [doi: 10.2196/45684] [Medline: 37171840]
30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Rev Esp Cardiol (Engl Ed) 2021 Sep;74(9):790-799 [FREE Full text] [doi: 10.1016/j.rec.2021.07.010] [Medline: 34446261]
31. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 03;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
32. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
33. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute. 2000. URL: https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [accessed 2022-10-20]
34. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Hillsdale, NJ: Laurence Erlbaum Publishers; 1988.
35. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015 Nov 04;351:h5651 [FREE Full text] [doi: 10.1136/bmj.h5651] [Medline: 26537915]

XSL•FO
RenderX

36. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
37. Schünemann H, Bro'ek J, Guyatt G, Oxman A. GRADE handbook. Grading of Recommendations, Assessment, Development and Evaluation. URL: https://gdt.gradepro.org/app/handbook/handbook.html#:~:text=GRADE%20handbook&text=The%20GRADE%20handbook%20describes%20the,www.gradeworkinggroup.org) [accessed 2023-10-01]
38. Siemieniuk R, Guyatt G. What is GRADE? Evidence-Based Medicine (EBM) Toolkit. URL: https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/ [accessed 2023-10-01]
39. Adams CE, Jayaram M, Bodart AY, Sampson S, Zhao S, Montgomery AA. Tweeting links to Cochrane schizophrenia group reviews: a randomised controlled trial. BMJ Open 2016 Mar 08;6(3):e010509 [FREE Full text] [doi: 10.1136/bmjopen-2015-010509] [Medline: 26956164]
40. Allen HG, Stanton TR, Di Pietro F, Moseley GL. Social media release increases dissemination of original articles in the clinical pain sciences. PLoS One 2013;8(7):e68914 [FREE Full text] [doi: 10.1371/journal.pone.0068914] [Medline: 23874810]
41. Chapman SJ, Grossman RC, FitzPatrick ME, Brady RR. Randomized controlled trial of plain English and visual abstracts for disseminating surgical research via social media. Br J Surg 2019 Oct 02:1611-1616 [FREE Full text] [doi: 10.1002/bjs.11307] [Medline: 31577372]
42. Fox CS, Bonaca MA, Ryan JJ, Massaro JM, Barry K, Loscalzo J. A randomized trial of social media from Circulation. Circulation 2015 Jan 06;131(1):28-33 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.013509] [Medline: 25406308]
43. Fox CS, Gurary EB, Ryan J, Bonaca M, Barry K, Loscalzo J, et al. Randomized controlled trial of social media: effect of increased intensity of the intervention. J Am Heart Assoc 2016 Apr 27;5(5):e003088 [FREE Full text] [doi: 10.1161/JAHA.115.003088] [Medline: 27121850]
44. Huang S, Martin LJ, Yeh CH, Chin A, Murray H, Sanderson WB, et al. The effect of an infographic promotion on research dissemination and readership: a randomized controlled trial. CJEM 2018 Nov;20(6):826-833. [doi: 10.1017/cem.2018.436] [Medline: 30289098]
45. Luc JG, Archer MA, Arora RC, Bender EM, Blitz A, Cooke DT, et al. Social media improves cardiothoracic surgery literature dissemination: results of a randomized trial. Ann Thorac Surg 2020 Feb;109(2):589-595. [doi: 10.1016/j.athoracsur.2019.06.062] [Medline: 31404547]
46. Luc JG, Archer MA, Arora RC, Bender EM, Blitz A, Cooke DT, et al. Does tweeting improve citations? One-year results from the TSSMN prospective randomized trial. Ann Thorac Surg 2021 Jan;111(1):296-300. [doi: 10.1016/j.athoracsur.2020.04.065] [Medline: 32504611]
47. Tunnecliff J, Weiner J, Gaida JE, Keating JL, Morgan P, Ilic D, et al. Translating evidence to practice in the health professions: a randomized trial of Twitter vs Facebook. J Am Med Inform Assoc 2017 Mar 01;24(2):403-408 [FREE Full text] [doi: 10.1093/jamia/ocw085] [Medline: 27357833]
48. Widmer RJ, Mandrekar J, Ward A, Aase LA, Lanier WL, Timimi FK, et al. Effect of promotion via social media on access of articles in an academic medical journal: a randomized controlled trial. Acad Med 2019 Oct;94(10):1546-1553. [doi: 10.1097/ACM.0000000000002811] [Medline: 31149923]
49. Barlow B, Barlow A, Webb A, Cain J. "Capturing your audience": analysis of Twitter engagements between tweets linked with an educational infographic or a peer-reviewed journal article. J Vis Commun Med 2020 Oct;43(4):177-183. [doi: 10.1080/17453054.2020.1809358] [Medline: 33028127]
50. Botting N, Dipper L, Hilari K. The effect of social media promotion on academic article uptake. J Assoc Inf Sci Technol 2016 Jun 11;68(3):795-800. [doi: 10.1002/asi.23704]
51. Cawcutt KA, Erdahl LM, Englander MJ, Radford DM, Oxentenko AS, Girgis L, et al. Use of a coordinated social media strategy to improve dissemination of research and collect solutions related to workforce gender equity. J Womens Health (Larchmt) 2019 Jun;28(6):849-862. [doi: 10.1089/jwh.2018.7515] [Medline: 30998087]
52. Cevik M, Ong DS, Mackenzie G. How scientists and physicians use Twitter during a medical congress. Clin Microbiol Infect 2019 Dec;25(12):1561.e7-1561.12 [FREE Full text] [doi: 10.1016/j.cmi.2019.04.030] [Medline: 31102781]
53. Fargen KM, Ducruet AF, Hyer M, Hirsch JA, Tarr RW. Expanding the social media presence of the Journal of Neurointerventional Surgery: editor's report. J Neurointerv Surg 2017 Feb;9(2):215-218. [doi: 10.1136/neurintsurg-2015-012251] [Medline: 26927814]
54. Gunaratne K, Haghbayan H, Coomes EA. Tweeting authors: impact on research publicity and downstream citations. J Gen Intern Med 2020 Jun;35(6):1926-1927 [FREE Full text] [doi: 10.1007/s11606-019-05454-0] [Medline: 31654356]
55. Hayon S, Tripathi H, Stormont IM, Dunne MM, Naslund MJ, Siddiqui MM. Twitter mentions and academic citations in the urologic literature. Urology 2019 Jan;123:28-33. [doi: 10.1016/j.urology.2018.08.041] [Medline: 30278190]
56. Ibrahim AM, Lillemoe KD, Klingensmith ME, Dimick JB. Visual abstracts to disseminate research on social media: a prospective, case-control crossover study. Ann Surg 2017 Dec;266(6):e46-e48. [doi: 10.1097/SLA.0000000000002277] [Medline: 28448382]

XSL•FO
RenderX

57. Jackowich RA, Young M, Poirier É, Pukall CF. #PGADFacts: Results from a 12-month knowledge translation campaign on persistent genital arousal disorder/genito-pelvic dysesthesia (PGAD/GPD). Can J Hum Sex 2022 Sep 01;31(2):242-252. [doi: 10.3138/cjhs.2022-0011]
58. Jeong JW, Kim MJ, Oh HK, Jeong S, Kim MH, Cho JR, et al. The impact of social media on citation rates in coloproctology. Colorectal Dis 2019 Oct;21(10):1175-1182. [doi: 10.1111/codi.14719] [Medline: 31124259]
59. Mobarak S, Stott MC, Lee W, Davé MS, Tarazi M, Macutkiewicz C. The importance of social media to the academic surgical literature: relationship between Twitter activity and readership metrics. Surgery 2021 Sep;170(3):650-656. [doi: 10.1016/j.surg.2021.01.003] [Medline: 33612291]
60. Narayanaswami P, Gronseth G, Dubinsky R, Penfold-Murray R, Cox J, Bever Jr C, et al. The impact of social media on dissemination and implementation of clinical practice guidelines: a longitudinal observational study. J Med Internet Res 2015 Aug 13;17(8):e193 [FREE Full text] [doi: 10.2196/jmir.4414] [Medline: 26272267]
61. Nason GJ, O'Kelly F, Kelly ME, Phelan N, Manecksha RP, Lawrentschuk N, et al. The emerging use of Twitter by urological journals. BJU Int 2015 Mar;115(3):486-490. [doi: 10.1111/bju.12840] [Medline: 24925047]
62. Navarro SM, Mazingi D, Keil E, Dube A, Dedeker C, Stewart KA, et al. Identifying new frontiers for social media engagement in global surgery: an observational study. World J Surg 2020 Sep;44(9):2881-2891 [FREE Full text] [doi: 10.1007/s00268-020-05553-8] [Medline: 32447417]
63. Ng FK, Wallace S, Coe B, Owen A, Lynch J, Bonvento B, et al. From smartphone to bed-side: exploring the use of social media to disseminate recommendations from the National Tracheostomy Safety Project to front-line clinical staff. Anaesthesia 2020 Feb;75(2):227-233 [FREE Full text] [doi: 10.1111/anae.14747] [Medline: 31250430]
64. O'Kelly F, Nason GJ, Manecksha RP, Cascio S, Quinn FJ, Leonard M, et al. The effect of social media (#SoMe) on journal impact factor and parental awareness in paediatric urology. J Pediatr Urol 2017 Oct;13(5):513.e1-513.e7. [doi: 10.1016/j.jpurol.2017.03.027] [Medline: 28483467]
65. Ozkent MS, Böcü K, Altintas E, Gul M. Correlation between Twitter mentions and academic citations in sexual medicine journals. Int J Impot Res 2022 Sep 12;34(6):593-598. [doi: 10.1038/s41443-021-00457-0] [Medline: 34253870]
66. Patel VR, Gereta S, Blanton CJ, Chu AL, Reddy NK, Mackert M, et al. #ColonCancer: social media discussions about colorectal cancer during the COVID-19 pandemic. JCO Clin Cancer Inform 2022 Jan;6:e2100180 [FREE Full text] [doi: 10.1200/CCI.21.00180] [Medline: 35025670]
67. Reppucci ML, De La Torre L, Peña A, Judd-Glossy L, Pickett K, Ketzer J, et al. The use of social media among the pediatric colorectal community. Pediatr Surg Int 2022 Jan;38(1):43-50. [doi: 10.1007/s00383-021-05022-w] [Medline: 34554295]
68. Rotolo SM, Jain S, Dhaon S, Dokhanchi JK, Kalata E, Shah T, et al. A coordinated strategy to develop and distribute infographics addressing COVID-19 vaccine hesitancy and misinformation. J Am Pharm Assoc (2003) 2022;62(1):224-231 [FREE Full text] [doi: 10.1016/j.japh.2021.08.016] [Medline: 34483058]
69. Thoma B, Murray H, Huang SY, Milne WK, Martin LJ, Bond CM, et al. The impact of social media promotion with infographics and podcasts on research dissemination and readership. CJEM 2018 Mar;20(2):300-306. [doi: 10.1017/cem.2017.394] [Medline: 28899440]
70. Trueger NS, Bokarius AV, Carroll S, April MD, Thoma B. Impact of a physician-led social media sharing program on a medical journal's web traffic. J Am Coll Radiol 2018 Jan;15(1 Pt B):184-189. [doi: 10.1016/j.jacr.2017.09.035] [Medline: 29122507]
71. Utengen A, Rouholiman D, Gamble JG, Grajales III FJ, Pradhan N, Staley AC, et al. Patient participation at health care conferences: engaged patients increase information flow, expand propagation, and deepen engagement in the conversation of tweets compared to physicians or researchers. J Med Internet Res 2017 Aug 17;19(8):e280 [FREE Full text] [doi: 10.2196/jmir.8049] [Medline: 28818821]
72. Wadhwa V, Latimer E, Chatterjee K, McCarty J, Fitzgerald RT. Maximizing the tweet engagement rate in academia: analysis of the AJNR Twitter feed. AJNR Am J Neuroradiol 2017 Oct;38(10):1866-1868 [FREE Full text] [doi: 10.3174/ajnr.A5283] [Medline: 28663265]
73. Widmer RJ, Engler NB, Geske JB, Klarich KW, Timimi FK. An academic healthcare Twitter account: the Mayo Clinic experience. Cyberpsychol Behav Soc Netw 2016 Jun;19(6):360-366. [doi: 10.1089/cyber.2015.0607] [Medline: 27327062]
74. Wray CM, Auerbach AD, Arora VM. The adoption of an online journal club to improve research dissemination and social media engagement among hospitalists. J Hosp Med 2018 Nov;13(11):764-769. [doi: 10.12788/jhm.2987] [Medline: 30484779]
75. Bolderston A, Meeking K, Snaith B, Watson J, Westerink A, Woznitza N. Five years of #MedRadJClub: an impact evaluation of an established Twitter journal club. J Med Radiat Sci 2022 Jun;69(2):165-173 [FREE Full text] [doi: 10.1002/jmrs.569] [Medline: 35143706]
76. Carley S, Beardsell I, May N, Crowe L, Baombe J, Grayson A, et al. Social-media-enabled learning in emergency medicine: a case study of the growth, engagement and impact of a free open access medical education blog. Postgrad Med J 2018 Feb;94(1108):92-96. [doi: 10.1136/postgradmedj-2017-135104] [Medline: 29054933]
77. Dyson MP, Newton AS, Shave K, Featherstone RM, Thomson D, Wingert A, et al. Social media for the dissemination of cochrane child health evidence: evaluation study. J Med Internet Res 2017 Sep 01;19(9):e308 [FREE Full text] [doi: 10.2196/jmir.7819] [Medline: 28864427]

XSL•FO
RenderX

78. Gardhouse AI, Budd L, Yang SY, Wong CL. #GeriMedJC: the twitter complement to the traditional-format geriatric medicine journal club. J Am Geriatr Soc 2017 Jun 22;65(6):1347-1351. [doi: 10.1111/jgs.14920] [Medline: 28432823]
79. Gates A, Featherstone R, Shave K, Scott SD, Hartling L. Dissemination of evidence in paediatric emergency medicine: a quantitative descriptive evaluation of a 16-week social media promotion. BMJ Open 2018 Jun 06;8(6):e022298 [FREE Full text] [doi: 10.1136/bmjopen-2018-022298] [Medline: 29880576]
80. Hawkins CM, Hillman BJ, Carlos RC, Rawson JV, Haines R, Duszak Jr R. The impact of social media on readership of a peer-reviewed medical journal. J Am Coll Radiol 2014 Nov;11(11):1038-1043. [doi: 10.1016/j.jacr.2014.07.029] [Medline: 25439618]
81. Hoang JK, McCall J, Dixon AF, Fitzgerald RT, Gaillard F. Using social media to share your radiology research: how effective is a blog post? J Am Coll Radiol 2015 Jul;12(7):760-765. [doi: 10.1016/j.jacr.2015.03.048] [Medline: 25959491]
82. Maloney S, Tunnecliff J, Morgan P, Gaida JE, Clearihan L, Sadasivan S, et al. Translating evidence into practice via social media: a mixed-methods study. J Med Internet Res 2015 Oct 26;17(10):e242 [FREE Full text] [doi: 10.2196/jmir.4763] [Medline: 26503129]
83. McNamara P, Usher K. Share or perish: social media and the International Journal of Mental Health Nursing. Int J Ment Health Nurs 2019 Aug 30;28(4):960-970. [doi: 10.1111/inm.12600] [Medline: 31257702]
84. Niehaus WN, Silver JK, Katz MS. The PM and R journal implements a social media strategy to disseminate research and track alternative metrics in physical medicine and rehabilitation. PM R 2018 May;10(5):538-543. [doi: 10.1016/j.pmrj.2017.12.003] [Medline: 29253533]
85. Rohilla R, Kumar MP, Patil AN, Panchal S, Arora A, Gupta M. Appropriateness evaluation of drug information center's Facebook page for evidence-based drug information dissemination. J Am Pharm Assoc (2003) 2020;60(6):e129-e132. [doi: 10.1016/j.japh.2020.03.018] [Medline: 32312670]
86. Segura Sampedro JJ, Morales Soriano R, Ramos Rodríguez JL, González-Argenté FJ, Mayol J. Twitter® use and its implications in Spanish association of surgeons meetings and congresses. Cir Esp (Engl Ed) 2018;96(6):352-356. [doi: 10.1016/j.ciresp.2018.02.012] [Medline: 29622342]
87. Buckarma EH, Thiels CA, Gas BL, Cabrera D, Bingener-Casey J, Farley DR. Influence of social media on the dissemination of a traditional surgical research article. J Surg Educ 2017;74(1):79-83. [doi: 10.1016/j.jsurg.2016.06.019] [Medline: 27993626]
88. Lindquist LA, Ramirez-Zohfeld V. Visual abstracts to disseminate geriatrics research through social media. J Am Geriatr Soc 2019 Jun;67(6):1128-1131. [doi: 10.1111/jgs.15853] [Medline: 30875100]
89. Schwenk ES, Jaremko KM, Gupta RK, Udani AD, McCartney CJ, Snively A, et al. Upgrading a social media strategy to increase Twitter engagement during the spring annual meeting of the American Society of Regional Anesthesia and Pain Medicine. Reg Anesth Pain Med 2017;42(3):283-288. [doi: 10.1097/AAP.0000000000000586] [Medline: 28267069]
90. Bhatt NR, Czarniecki SW, Borgmann H, van Oort IM, Esperto F, Pradere B, EAU Guidelines Office Dissemination Committee. A systematic review of the use of social media for dissemination of clinical practice guidelines. Eur Urol Focus 2021 Sep;7(5):1195-1204. [doi: 10.1016/j.euf.2020.10.008] [Medline: 33172773]
91. Eysenbach G. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. J Med Internet Res 2011 Dec 19;13(4):e123 [FREE Full text] [doi: 10.2196/jmir.2012] [Medline: 22173204]
92. Wong K, Piraquive J, Levi JR. Social media presence of otolaryngology journals: the past, present, and future. Laryngoscope 2018 Feb;128(2):363-368. [doi: 10.1002/lary.26727] [Medline: 28600839]
93. Azer SA, Azer S. Top-cited articles in medical professionalism: a bibliometric analysis versus altmetric scores. BMJ Open 2019 Jul 31;9(7):e029433 [FREE Full text] [doi: 10.1136/bmjopen-2019-029433] [Medline: 31371297]
94. Zhao J, Harvey G, Vandyk A, Gifford W. Social media for ImpLementing evidence (SMILE): conceptual framework. JMIR Form Res 2022 Mar 09;6(3):e29891 [FREE Full text] [doi: 10.2196/29891] [Medline: 35262488]
95. Alotaibi NM, Samuel N, Guha D, Nassiri F, Badhiwala JH, Tam J, et al. Social media for academic neurosurgical programs: the University of Toronto experience. World Neurosurg 2016 Sep;93:449-457. [doi: 10.1016/j.wneu.2016.06.134] [Medline: 27423202]
96. Sabato LA, Barone C, McKinney K. Use of social media to engage membership of a state health-system pharmacy organization. Am J Health Syst Pharm 2017 Jan 01;74(1):e72-e75. [doi: 10.2146/ajhp151042] [Medline: 28007724]
97. AlRyalat SA, Malkawi L, Alsaedi AA, Ashouri K, El Khatib O, Yaseen A, et al. Comparing the attention score between American and European cardiology guidelines. High Blood Press Cardiovasc Prev 2020 Jun;27(3):259-264. [doi: 10.1007/s40292-020-00379-w] [Medline: 32253743]
98. Lu Y, Wu Y, Liu J, Li J, Zhang P. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. J Med Internet Res 2017 Apr 07;19(4):e109 [FREE Full text] [doi: 10.2196/jmir.7087] [Medline: 28389418]
99. Araujo AC, Vanin AA, Nascimento DP, Gonzalez GZ, Costa LO. What are the variables associated with Altmetric scores? Syst Rev 2021 Jun 30;10(1):193 [FREE Full text] [doi: 10.1186/s13643-021-01735-0] [Medline: 34187573]

XSL•FO
RenderX

100. Araujo AC, Nascimento DP, Gonzalez GZ, Maher CG, Pena Costa LO. Impact of low back pain clinical trials measured by the Altmetric score: cross-sectional study. J Med Internet Res 2018 Apr 05;20(4):e86 [FREE Full text] [doi: 10.2196/jmir.9368] [Medline: 29622526]
101. Kunze KN, Manzi JE, Polce EM, Vadhera A, Bhandari M, Piuzzi NS. High social media attention scores are not reflective of study quality: an altmetrics-based content analysis. Intern Emerg Med 2022 Aug;17(5):1363-1374. [doi: 10.1007/s11739-022-02939-5] [Medline: 35137307]
102. Gough A, Hunter RF, Ajao O, Jurek A, McKeown G, Hong J, et al. Tweet for behavior change: using social media for the dissemination of public health messages. JMIR Public Health Surveill 2017 Mar 23;3(1):e14 [FREE Full text] [doi: 10.2196/publichealth.6313] [Medline: 28336503]
103. Biancovilli P, Makszin L, Csongor A. Breast cancer on social media: a quali-quantitative study on the credibility and content type of the most shared news stories. BMC Womens Health 2021 May 15;21(1):202 [FREE Full text] [doi: 10.1186/s12905-021-01352-y] [Medline: 33992111]
104. Chen Q, Min C, Zhang W, Ma X, Evans R. factors driving citizen engagement with government TikTok accounts during the COVID-19 pandemic: model development and analysis. J Med Internet Res 2021 Feb 04;23(2):e21463 [FREE Full text] [doi: 10.2196/21463] [Medline: 33481756]
105. Harris JK, Mart A, Moreland-Russell S, Caburnay CA. Diabetes topics associated with engagement on Twitter. Prev Chronic Dis 2015 May 07;12:E62 [FREE Full text] [doi: 10.5888/pcd12.140402] [Medline: 25950569]

Abbreviations

GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by T de Azevedo Cardoso; submitted 31.07.23; peer-reviewed by A Araujo, M Katz; comments to author 22.11.23; revised version received 15.12.23; accepted 18.03.24; published 05.06.24.

Please cite as:
Roberts-Lewis S, Baxter H, Mein G, Quirke-McFarlane S, Leggat FJ, Garner H, Powell M, White S, Bearne L
Examining the Effectiveness of Social Media for the Dissemination of Research Evidence for Health and Social Care Practitioners: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e51418
URL: https://www.jmir.org/2024/1/e51418
doi:10.2196/51418
PMID:38838330

©Sarah Roberts-Lewis, Helen Baxter, Gill Mein, Sophia Quirke-McFarlane, Fiona J Leggat, Hannah Garner, Martha Powell, Sarah White, Lindsay Bearne. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Physician Perspectives on Internet-Informed Patients: Systematic Review

Qianfeng Lu[1], MSc; Peter Johannes Schulz[1,2,3], PhD

[1]Faculty of Communication, Culture and Society, Università della Svizzera italiana, Lugano, Switzerland

[2]Department of Communication & Media, Ewha Womans University, Seoul, Republic of Korea

[3]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore

Corresponding Author:
Peter Johannes Schulz, PhD
Faculty of Communication, Culture and Society
Università della Svizzera italiana
Via Buffi 13
Lugano, 6900
Switzerland
Phone: 41 586664724
Email: schulzp@usi.ch

Abstract

Background: The internet has become a prevalent source of health information for patients. However, its accuracy and relevance are often questionable. While patients seek physicians' expertise in interpreting internet health information, physicians' perspectives on patients' information-seeking behavior are less explored.

Objective: This review aims to understand physicians' perceptions of patients' internet health information-seeking behavior as well as their communication strategies and the challenges and needs they face with internet-informed patients.

Methods: An initial search in PubMed, Scopus, CINAHL, Communication and Mass Media Complete, and PsycINFO was conducted to collect studies published from January 1990 to August 1, 2022. A subsequent search on December 24, 2023, targeted recent studies published after the initial search cutoff date. Two reviewers independently performed title, abstract, and full-text screening, adhering to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement guidelines. Thematic analysis was then used to identify key themes and systematically categorize evidence from both qualitative and quantitative studies under these themes.

Results: A total of 22 qualifying articles were identified after the search and screening process. Physicians were found to hold diverse views on patients' internet searches, which can be viewed as a continuous spectrum of opinions ranging from positive to negative. While some physicians leaned distinctly toward either positive or negative perspectives, a significant number expressed more balanced views. These physicians recognized both the benefits, such as increased patient health knowledge and informed decision-making, and the potential harms, including misinformation and the triggering of negative emotions, such as patient anxiety or confusion, associated with patients' internet health information seeking. Two communicative strategies were identified: the participative and defensive approaches. While the former seeks to guide internet-informed patients to use internet information with physicians' expertise, the latter aims to discourage patients from using the internet to seek health information. Physicians' perceptions were linked to their strategies: those holding positive views tended to adopt a participative approach, while those with negative views favored a defensive strategy. Some physicians claimed to shift between the 2 approaches depending on their interaction with a certain patient. We also identified several challenges and needs of physicians in dealing with internet-informed patients, including the time pressure to address internet-informed patient demands, a lack of structured training, and being uninformed about trustworthy internet sites that can be recommended to internet-informed patients.

Conclusions: This review highlights the diverse perceptions that physicians hold toward internet-informed patients, as well as the interplay between their perceptions, communication strategies, and their interactions with individual patients. Incorporating elements into the medical teaching curriculum that introduce physicians to reliable internet health resources for patient guidance, coupled with providing updates on technological advancements, could be instrumental in equipping physicians to more effectively manage internet-informed patients.

Trial Registration: PROSPERO CRD42022356317; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=356317

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47620) doi:10.2196/47620

KEYWORDS

internet-informed patients; physician-patient communication; health information–seeking; misinformation; digital health

Introduction

Background

The internet has become a common resource for people seeking health-related information [1,2]. It offers a vast array of health information that was previously accessible mainly through physicians. Over the past decade, attention to internet health information–seeking behavior among patients has grown. The idea of patients searching for information on the internet can be traced back to Ferguson [3], who coined the term *e-patients* in 2007 for those seeking health information on the internet.

People seeking health information on the internet may have different purposes [4]. Before seeing a physician, they might search on the internet to determine if a medical appointment is necessary. Once a consultation is scheduled, some patients look for background information to better prepare for their appointments [5]. After consultations, some turn to the internet to clarify and supplement the physician's information, such as details about medication indications and side effects [6]. These patients seek internet health information to better understand their diagnosed condition and manage their treatment [7]. Being equipped with knowledge from the internet, they are often referred to as internet-informed patients [8].

Some patients intentionally avoid information from the internet. For instance, information avoidance among cancer patients has been frequently explored by researchers [9,10]. However, several systematic reviews from patients' perspectives conclude that patients often feel more confident discussing their health with physicians and feel more empowered to manage their conditions after searching for information on the internet [11-13]. Internet health information allows patients to transition from being passive and uninformed recipients to empowered and informed consumers [14,15].

However, searching for health information on the internet poses challenges for both patients and the health care system. The vast amount of internet health information can be overwhelming for laypeople, who might find it difficult to sift through and identify relevant and accurate content. Consequently, patients may end up learning mis- and disinformation or become confused by conflicting information they encounter on the internet [16,17]. Seeking internet health information can exacerbate patients' anxiety about their health [18,19]. In contrast, physicians can provide accurate and personalized information, which can reassure patients. This becomes particularly vital in cases where patients, having been misinformed by internet sources, are taking active steps to manage their health [20].

Although the internet has become a popular source for accessing health information, consultations with physicians remain the primary influence on patients' medical decisions [21]. Patients have various strategies for addressing the internet health information they gather before a consultation. They might either implicitly or explicitly share this information, hoping the physician will consider it in their judgment [22]. However, some may withhold it, fearing it might upset their physician [13]. A significant barrier to such disclosure is the concern that the physician might perceive it as a challenge to their expertise [23]. Many patients are cautious about revealing internet information to their physicians to avoid causing offense [24]. As a result, they often assess the physician's potential reactions to such information before deciding whether to share it [25]. Research has shown that dialogue about internet information between patients and physicians can improve patient satisfaction and their relationships. This improvement is particularly noticeable when physicians acknowledge the patients' efforts and take their internet-sourced information seriously [26].

The physician-patient relationship is dyadic, meaning it is built on mutual trust and efficient communication at its core [27]. Patients often access health information through the internet and approach physicians with a desire to seek clarification and advice [28]. Discussing internet health information offers a new communication opportunity, allowing for physicians to better understand their patients' preferences and concerns and for patients to deepen their trust in their physicians. To maximize this communication opportunity, understanding the perspectives of both parties, that is, physicians and internet-informed patients, is crucial. However, while much of the literature focuses on internet-informed patients, there is less emphasis on physicians' perspectives [29]. Several reviews have explored patients' internet health information–seeking behavior and its impact on the physician-patient relationship [11,13,14,30,31]. Nonetheless, these reviews primarily focused on patients, with evidence collected mainly from the patient's viewpoint.

Objective

To the best of our knowledge, this is the first systematic review focusing on physicians' perspectives. We aimed to explore physicians' views on patients' internet information–seeking behavior and their experiences with internet-informed patients. Our research questions (RQs) are elaborated as follows:

- RQ1: how do physicians perceive patients' internet health information–seeking behavior?
- RQ2: what communicative strategies do physicians use in interacting with internet-informed patients? In this context, the term *strategies* refers to physicians' actions and responses to internet-informed patients.
- RQ3: what difficulties and needs do physicians encounter when dealing with internet-informed patients?

Methods

Search Strategy

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [32,33].

XSL•FO
RenderX

To identify the appropriate search algorithm, we undertook 3 steps. In the first step, we conducted an exploratory search using various combinations of terms such as *physician*, *communication strategies*, *internet-informed patients*, and *physician-patient communication*. Upon reviewing the reference lists of several primary articles identified from the exploratory search, we discerned 12 articles that were deemed relevant to this review. In the second step, we structured 3 blocks of keywords, each representing a facet of the RQs: physicians as the target population (block 1), patients' health information behavior (block 2), and physicians' perspectives or communication strategies with internet-informed patients (block 3; Textbox 1). Each block encompasses a set of synonyms and related Medical Subject Heading (MeSH) terms tailored to represent the aspect of the RQs under consideration. For instance, the patients' health information behavior (block 2) includes terms like *internet-informed patients*, *health-related internet use*, e-patients, *internet information*, *eHealth*, *health information search*, *internet health information*, and *online health information*. Within each block, synonymous terms and related MeSH terms were amalgamated using the OR operator. The 3 blocks were then combined with the AND operator. We used truncations to widen our search and set abstract or title limiters to circumvent the retrieval of an overwhelming number of results in several databases. This combination of terms was used across several platforms: PubMed, Scopus, CINAHL, Communication and Mass Media Complete, and PsycINFO.

Textbox 1. Search terms.

Block 1: physicians

physicians, doctors

Block 2: patients' health information behavior

internet-informed patients, health-related internet use, e-patients, internet information, eHealth, health information search, internet health information, online health information

Block 3: physicians' perspectives or communication strategies with internet-informed patients

doctor–patient relation, physician-patient relation, physician-patient communication, doctor–patient communication, experience, attitude, strategy, belief

The first search was conducted in August 2022, and we initially restricted our search to the time frame from January 1, 1990, to August 1, 2022. All search outcomes were cataloged in a data set for subsequent screening. Before this screening, we assessed the effectiveness of our search algorithm by seeking out the 12 articles initially pinpointed from the exploratory search. All 12 articles were successfully located via our search algorithm, affirming its adequacy. We then performed a second search on December 24, 2023, using the same combination of keywords in the 5 databases used for the first search. The second search was conducted to collect studies published after August 1, 2022, the cutoff date for the first search. Multimedia Appendix 1 provides a detailed overview of the search algorithm and the total number of results obtained from each database.

Inclusion and Exclusion Criteria

We included empirical qualitative, quantitative, and mixed methods studies that assessed physicians' perspectives on patients' internet health information–seeking behavior and their communication with internet-informed patients. A *physician* is defined as a medical doctor, either a specialist or a general practitioner (GP). This review focuses on physicians attending to physical health conditions. Studies had to be published in a peer-reviewed journal after 1990 and be written in English. We chose 1990 as the starting year because public access to the internet was not available before then. We did not consider the impact factor of the journals or their peer-review process. If a study covered multiple populations, such as physicians, nurses, and patients, we only considered those presenting the results for physicians separately from other groups. Textbox 2 lists the inclusion and exclusion criteria.

XSL•FO
RenderX

Textbox 2. Inclusion and exclusion criteria.

Inclusion criteria

- Study topic: physicians' perspectives regarding patients' internet health information–seeking behavior
- Study populations: medical doctors who are attending to physical health conditions
- Study type: empirical studies using qualitative, quantitative, or mixed methods
- Language: written in English
- Publication year: published after the year 1990

Exclusion criteria

- Study topic: physicians' perspectives in contexts unrelated to patients' internet health information seeking, such as telemedicine, digital devices, electronic health records, and physicians' own internet information–seeking behavior and social media use
- Study populations: nonmedical doctors or psychiatrists specializing in areas other than physical health
- Study type: reviews, evidence syntheses, study protocols, reports, book chapters, or any forms of gray literature
- Language: written in languages other than English
- Publication year: published before the year 1990

Studies Identification

The first author (QL) and a research assistant independently assessed the titles and abstracts of these articles based on the inclusion and exclusion criteria. Any disagreements in selection were discussed with another author until a consensus was reached. Full-text screenings were performed by 2 authors independently, and any discrepancies were resolved through discussion.

Data Analysis

A thematic analysis was conducted to identify prominent themes and to systematically organize the literature under these themes. This method permits the combination of qualitative and quantitative evidence [34] and has been used in previous reviews concerning physicians' internet health information needs and their role in cancer care [35,36]. Both authors (QL and PJS) of the present review analyzed data from the included studies and synthesized them into themes and subthemes. The qualitative results of the included articles were used to describe the themes by using specific quotations that offer direct evidence of physicians' experiences and perspectives. Quantitative results from survey studies were used to validate and support the identification of themes. For instance, if a theme emerged from the qualitative data suggesting that "physicians perceive that patients' internet health information seeking boosts patients' confidence in participating in their healthcare," quantitative data were used to indicate the percentage of physicians who echoed the perspective that internet health information seeking boosts patient confidence. Our analysis focused solely on the raw data and results of the included articles, extracting direct evidence without incorporating the interpretations or conclusions drawn by the authors of the included studies. We did not use specialized qualitative software for data analysis. Instead, we created 2 structured tables (Tables S1 and S2 in Multimedia Appendix 2 [22,29,37-56]) in Microsoft Word (Microsoft Corp) to systematically list, organize, and synthesize all themes and subthemes alongside their associated qualitative and quantitative evidence. This approach facilitated clear organization of qualitative quotations and the corresponding survey evidence.

Quality Assessment

Two tools were used by the first author to evaluate the risk of bias in the included studies. The Critical Appraisal Skills Program (CASP) checklist assessed the quality of qualitative research, while a tool developed by Hoy et al [57] evaluated quantitative research. Both tools comprise 10 questions that assess the study's objectives, methodology, sampling, ethical considerations, and other aspects. A quantitative study was considered to have a moderate risk of bias if it satisfied 4 or 5 out of the 10 criteria and a low risk of bias if it met 6 to 8 criteria. The CASP does not categorize qualitative research as "high," "medium," or "low" quality; instead, it appraises the strengths and limitations of its methodology. It has been widely used in health-related qualitative evidence syntheses [58].

Results

Overview of the Articles

Figure 1 illustrates the article selection process for this review in accordance with the PRISMA guidelines. Using the search algorithm described in the Search Strategy section, 2856 records were initially identified, which ultimately yielded 2159 (75.6%) unique articles after removing duplicates. After screening titles and abstracts, 32 (1.48%) articles remained for full-text assessment. In total, 10 (31%) articles that did not meet the inclusion criteria were excluded: 4 (40%) that focused on telemedicine and technological health devices; 3 (30%) with an incorrect sample (1 each targeting patients with cancer and caregivers, the general population, and nurses); 1 (10%) study that focused on physicians' social media use; 1 (10%) study on physicians' view of patient family members' information seeking; and 1 (10%) study for which the full article was inaccessible. Ultimately, 22 (69%) articles were included in this review.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the studies identified, screened, and assessed for eligibility.

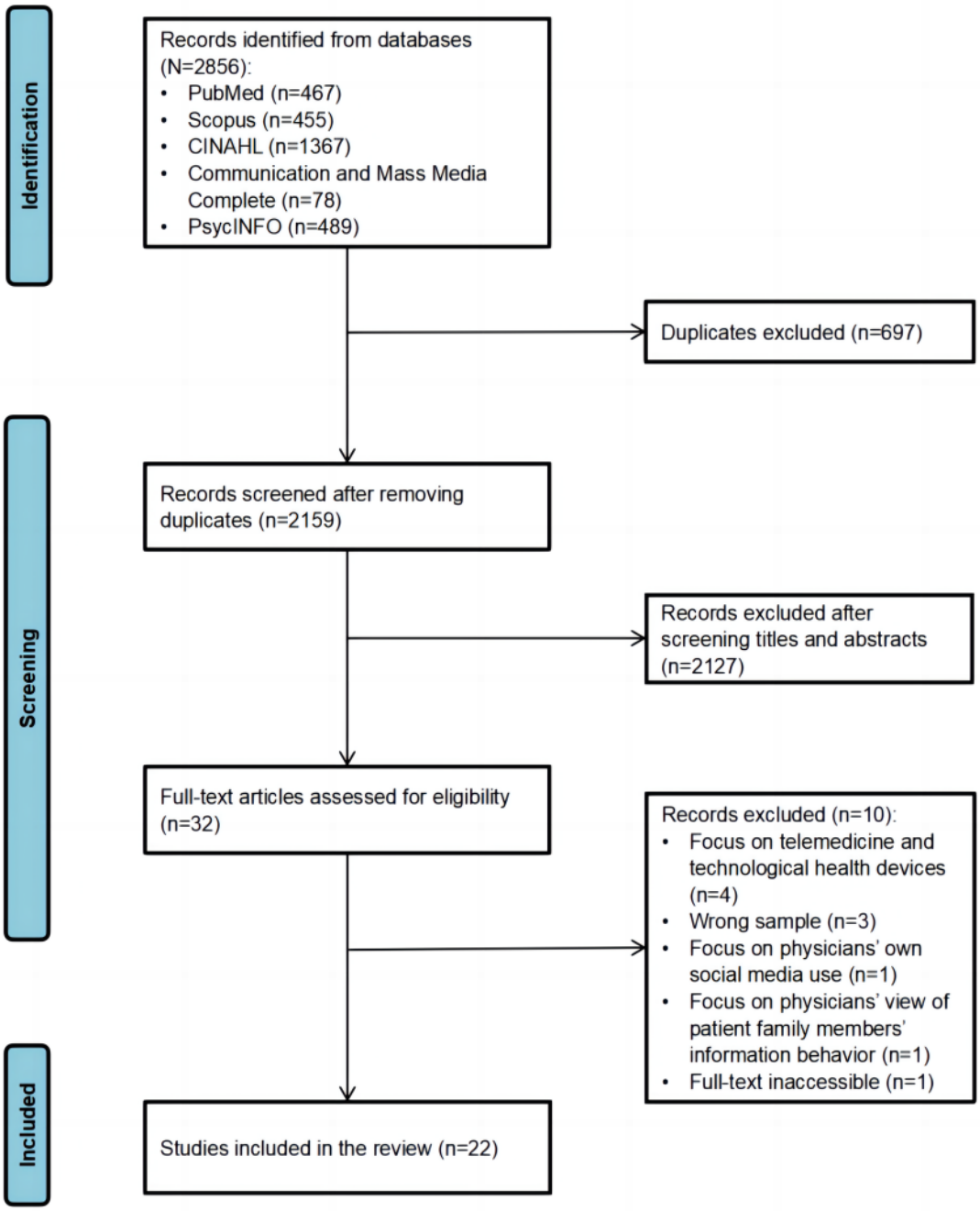

Of the 22 included articles, 11 (50%) used qualitative methods, including focus groups, interviews, and analysis of audio-recorded clinical consultations. Another 11 (50%) articles adopted a quantitative approach using survey data. Furthermore, 2 (18%) of the quantitative articles also incorporated a qualitative method through open-ended survey questions [37,38]. A total of 4 (18%) studies exclusively collected data from family physicians or GPs, 2 (9%) studies focused solely on oncologists, 1 (5%) study focused on dermatologists, while the remaining studies (n=15, 68%) encompassed physicians from various specialties. Most of the included studies were conducted in North America and Europe, encompassing the United States (n=4, 18%), Canada (n=2, 9%), Germany (n=4, 18%), the Netherlands (n=1, 5%), the United Kingdom (n=2, 9%), Switzerland (n=2, 9%), and Norway (n=1, 5%). Other countries featured in the studies included Israel (n=3, 14%), Oman (n=1, 5%), Brazil (n=1, 5%), and South Korea (n=1, 5%). One study interviewed physicians from 7 different countries: Australia, Israel, France, the United States, the Philippines, New Zealand, and Hungary. Table 1 provides an overview of the included articles.

XSL•FO
RenderX

Table 1. Overview of the included articles.

Study	Journal	Method	Data collection	Country	Participants
Wangler and Jansky [47], 2020	*Wiener Medizinische Wochenschrift*	Qualitative	Interviews	Germany	38 GPs[a]
Kim and Kim [42], 2009	*Informatics for Health and Social Care*	Quantitative	Web surveys	South Korea	430 specialists, 32 GPs, and 31 training doctors
Masters et al [49], 2020	*Journal of Medical Internet Research*	Quantitative	Surveys	Oman and Germany	108 surgeons
Ahmad et al [45], 2006	*Journal of Medical Internet Research*	Qualitative	Focus groups	Canada	48 family physicians
van Uden-Kraan et al [38], 2010	*Clinical Rheumatology*	Mixed methods	Surveys with open questions	The Netherlands	134 rheumatologists and 104 oncologists
Fujioka and Stewart [44], 2013	*Health Communication*	Quantitative	Web surveys	The United States	104 physicians
Murray et al [43], 2003	*Journal of Medical Internet Research*	Quantitative	Surveys	The United States	1050 physicians
Giveon et al [48], 2009	*Israel Medical Association Journal*	Quantitative	Surveys	Israel	118 family physicians
Caiata-Zufferey and Schulz [51], 2012	*Health Communication*	Qualitative	Interviews	Switzerland	17 physicians
Ahluwalia et al [40], 2010	*British Journal of General Practice*	Qualitative	Interviews	The United Kingdom	11 GPs
Moick and Terlutter [56], 2012	*Medicine 2.0*	Quantitative	Web surveys	Germany	287 physicians
Potts and Wyatt [54], 2002	*Journal of Medical Internet Research*	Quantitative	Web Surveys	The United Kingdom	748 physicians
Ohana and Barnoy [55], 2019	*Nursing Outlook*	Quantitative	Surveys	Israel	184 e-patients, 52 nurses, and 48 physicians[b]
Győrffy et al [39], 2020	*PLoS ONE*	Qualitative	Interviews	Australia, Israel, France, the United States, the Philippines, New Zealand, and Hungary	11 digitally engaged physicians
Shachar [29], 2022	*Social Science & Medicine*	Qualitative	Interviews	Israel	32 physicians
da Mota et al [50], 2018	*Revista da Associação Médica Brasileira*	Quantitative	Surveys	Brazil	183 physicians
Helft et al [37], 2003	*Journal of Clinical Oncology*	Mixed method	Surveys with open questions	The United States	266 oncologists
Sommerhalder et al [46], 2009	*Patient Education and Counseling*	Qualitative	Interviews	Switzerland	32 patients and 20 physicians[b]
Shen et al [22], 2015	*Psycho-Oncology*	Qualitative	21 audio-recorded clinical consultations	The United States	Oncologists (undefined number of physicians)
Fredriksen et al [41], 2018	*Health Care for Women International*	Qualitative	Interviews	Norway	13 Norwegian GPs, midwives, and physiotherapists[b]
MacDonald et al [52], 2018	*Journal of Medical Internet Research*	Qualitative	Interviews	Canada	12 physicians and nurses (3 GPs, 2 registered nurses, 1 nurse practitioner, 3 rheumatologists, 1 physician clinician-scientist, and 2 rheumatology fellows)[b]
Schick et al [53], 2023	*Journal of Medical Internet Research*	Qualitative	Interviews	Germany	16 patients and 12 dermatologists[b]

XSL•FO
RenderX

[a]GP: general practitioner.

[b]Only study results and quotations from physicians, including GPs and specialists, are considered.

In the quality assessment, all the qualitative studies (11/22, 50%) received "yes" responses to the first 5 questions on the CASP checklist. This indicates that these studies provided clear descriptions of their research aims and used appropriate methodologies, sampling procedures, and data collection techniques. Moreover, each study received positive responses to at least 8 out of the 10 questions. The primary area of concern, as denoted by negative responses, pertained to the following question on the checklist: "Has the relationship between the researcher and participants been adequately considered?" This suggests that many studies did not critically examine the researcher's role, its associated potential bias, and influence during the research process. Furthermore, 2 (9%) articles that used mixed methods were evaluated with the tool for quantitative studies, as most of their results were presented numerically. Most of the included quantitative studies (7/11, 64%) exhibited a low risk of bias, while the remainder (4/11, 36%) displayed a moderate risk. Multimedia Appendix 3 [22,29,37-56] details the quality check for each study.

Synthesis of the Studies

Overview of Categories

A total of 5 main categories were identified through the coding of evidence from the included articles. Two primary categories, "impacts on patients' health and health management" and "Impacts on physician-patient relations and health care services," represent physicians' perceptions of patients' internet health information seeking in terms of its effects on patients' health and health management, as well as on their relationships with physicians and on health care systems, respectively. These categories address RQ1. Two additional categories, "participative strategy" and "defensive strategy," summarize a series of communication actions physicians take in response to internet-informed patients, thereby addressing RQ2. The final category, "physicians' difficulties and needs," addresses RQ3 by delving into the challenges and needs that physicians perceive when dealing with internet-informed patients. Multimedia Appendix 2 provides a detailed list of all subthemes, accompanied by their qualitative quotations and supporting quantitative survey evidence.

In the following sections, we first offer a brief description of each primary category, accompanied by its corresponding subthemes highlighted in italics. We then delve deeper into physicians' perceptions, their choices regarding communicative strategies, and their identified difficulties and needs. Furthermore, we outline the antecedents of, or connections between, these categories. Direct qualitative quotations and survey evidence are provided to reinforce our observations.

Impacts on Patients' Health and Health Management

Physicians believe that seeking health information on the internet can *inform/educate patients* about their health conditions and *enhance patients' confidence* in participating in health care. The internet can also *provide social support* by connecting patients with their peers. However, physicians express concern that the internet might *misinform patients* with inaccurate information and some patients lack the capability to interpret internet information appropriately. Internet searches can potentially *trigger patients' negative emotions*, such as anxiety or confusion. Moreover, physicians view *patients' self-diagnosis and self-treatment* based on internet searches, conducted before medical consultations, in a negative light.

Impacts on Physician-Patient Relations and Health Care Services

Physicians believe that internet information seeking can *improve physician-patient relations*, as internet-informed patients are more empowered and informed in managing their health. Such searches also *encourage patients' participation in decision-making*. Moreover, internet-informed patients can *improve the efficiency of medical consultations*, potentially shortening the duration when patients have already informed themselves about their conditions. However, there have been reports from physicians about *nonadherent patients* who disregard medical advice due to their internet searches. Some physicians sense *feelings of distrust* from internet-informed patients and feel that their medical *authority is being challenged* by them. They also sometimes experience *negative emotions*, such as anxiety and uneasiness, when interacting with internet-informed patients. Furthermore, physicians note *increased time and information demands*, as well as *inappropriate medical requests* made by internet-informed patients.

Participative Strategy

Physicians *appreciate internet searches* when patients bring internet information to consultations. They help patients *examine the information* and *acknowledge their limited expertise* on specific topics. They clearly *explain the diagnosis and treatment* plan to patients and strive to *understand patients' emotional needs*, such as concerns about their health and motivations for conducting internet searches. Furthermore, physicians seek to *build ongoing relations with internet-informed patients* and *instruct patients on appropriate internet use*, including teaching them how to recognize appropriate information sources and recommending reliable internet sites.

Defensive Strategy

Physicians *decline to discuss* internet health information by displaying resistance or deferring patients to other specialists. They also *discredit the internet* as an unreliable information source and *devalue the internet health information* brought by patients.

Physicians' Difficulties and Needs

Physicians face *extra responsibility* as information interpreters or examiners beyond their traditional professional practices, adding pressure on an already overwhelmed health system. Physicians desire *training to manage internet-informed patients*, *stay updated with technology*, and be informed about *reliable internet sites* that they can recommend to patients.

XSL•FO
RenderX

Physicians' Perceptions of Internet-Informed Patients

Diverse views on internet-informed patients existed among physicians. These views can be seen as a continuous spectrum of opinions ranging from positive to negative. In qualitative studies, a group of digitally engaged physicians who were active on social media generally held positive perceptions toward internet health information, regarding its impact on physician-patient communication and patients' health [39]. In contrast, interviews with 11 GPs from the United Kingdom revealed generally negative perceptions [40]. Another interview study in Norway showed physicians with these 2 opposite opinions [41]. Between these 2 polarized views, a larger number of physicians expressed balanced perspectives on patients' internet searches. In survey studies, many physicians held neutral opinions regarding the impact of internet searches on physician-patient relations [42,43]. These physicians recognized both the beneficial and detrimental effects of internet information on patients' health [38,42,44]. Interview studies also showed that physicians' views on internet-informed patients varied based on their interactions with individual patients [41,45]. In other words, they did not approach all patients with a singular attitude; rather, they often held more nuanced or balanced perceptions.

Physicians with positive perceptions valued the more equal physician-patient relationship fostered by internet-informed patients [39]. For instance, an interview study from Switzerland found that physicians with positive perceptions believed that internet-informed patients, by searching on the internet, become more informed about their health conditions and thus feel encouraged and more confident to engage collaboratively with their physicians in health care [46]. Similarly, physicians from Israel mentioned in interviews that internet-informed patients had assisted them in making diagnoses and pinpointing referrals, thereby making medical encounters more efficient [29]. The following quotations exemplify this positive perspective:

> *Because of the Internet, social media and technology, my patients were coming to me with more information and they weren't looking to me to just solve a problem. They wanted to be involved in this problem. [39]* [Male physiotherapist aged 36 years]

> *I'd say spontaneously that it gives them [internet-informed patients] more right to have a say in a matter. They have, let's say, more empowerment to join in the conversation. They then already have an opinion, and don't come here thoroughly blank. [46]* [Male physician, obtained medical degree in 1983]

Physicians with negative perceptions expressed unpleasant feelings toward internet-informed patients. They felt that their authority was challenged by these patients [40]. Notably, 3 survey studies from South Korea, the United States, and the Netherlands showed that approximately 20% of physicians felt that internet-informed patients challenged or undermined their authority as medical professionals [38,42,43]. Furthermore, they perceived themselves as being devalued and distrusted by internet-informed patients [40,47]. Their views are expressed as follows:

> *For me that was the irritation, that the patient had far more trust in the computer and what they found on the web than in what I was trying to explain. [40]* [female GP]

> *I see a very big danger in the fact that the patient gets into a kind of tunnel through his/her constant search on the internet and then, in the end, is no longer receptive to the doctor's advice. Again and again, I experience those patients who constantly feel misunderstood and do doctor hopping. [47]* [female GP]

In contrast, many physicians hold mixed perceptions of internet-informed patients, which are influenced by their interactions with individual patients. For instance, 2 qualitative studies conducted in Canada and Norway have revealed that physicians differentiate between patients who use the internet for self-education and those who use it for self-diagnosis and self-treatment [41,45]. Physicians tend to view patients who use the internet for self-education in a favorable light. These patients often bring internet information to the physician for confirmation and remain receptive to the physicians' suggestions. In these cases, the internet serves as a helpful tool. In contrast, when patients use the internet to self-diagnose or self-treat, specifically, those who have already made up their minds before consulting a physician, they are often perceived as "challenging" patients. Physicians have reported feeling the need to defend their diagnosis or treatment plan, and this can evoke negative emotions, ranging from frustration to anger, when dealing with such self-diagnosing or self-treating patients [45]. Furthermore, physicians recognize the importance of established relationships with patients in shaping their perspectives. A positive prior relationship leads them to view the internet as beneficial [38]. Some physicians described it as follows:

> *I think there's one situation where the Internet is useful. If the person has the diagnosis, and they want to find out more, educate themselves,... I find that's actually helpful in cases where...it's not time-consuming for me.*

> *If they're, however, using it to diagnose, then I think that's where the problem lies... [45]* [A focus group of family physicians]

> *If the relationship is good, Internet use is not a problem. The biggest problem is with new patients with whom no relationship has yet been forged and who arrive with a certain assertivity or suspicion. [38]* [Not available]

Quantitative data also support the observation of physicians holding more neutral or balanced perceptions. A survey study of 406 US physicians who had previously encountered internet-informed patients during consultations found that 38% of the physicians consider the internet information brought by the internet-informed patients to benefit their relationship, 8% consider it harmful to the physician-patient relationship, and 54% provided neutral answers [43]. Another survey study involving 493 South Korean physicians found that 16.6% believed that discussing internet information with patients

XSL•FO
RenderX

positively impacted their relationships. In contrast, 25.6% indicated that it had negative impacts, while 42.6% chose a neutral response, indicating that they perceive it as having no significant impact on the physician-patient relationship either way [42]. In alignment with these 2 studies, the proportion of physicians with varying perceptions do not show significant differences across other survey studies from the Middle East, South America, and Europe [38,48-50]. Overall, most physicians hold neutral perceptions.

In addition, physicians' general perceptions of internet-informed patients appear to differ between different cultures, ages, and specializations. Focus groups involving 48 Canadian family physicians revealed that senior physicians felt a stronger sense of challenged authority than younger medical graduates [45]. In a survey of 108 surgeons, German surgeons were less content with internet information than their Omani counterparts [49]. Another survey from the Netherlands involving 238 oncologists and rheumatologists found that oncologists were less positive about internet use than rheumatologists [38]. However, these associations lack substantial evidence, as few studies have verified them.

Physicians' Communicative Strategies: How Do Physicians With Negative Perceptions Respond to Internet-Informed Patients?

Physicians' communicative strategies were mostly observed in qualitative studies. When physicians develop negative perceptions, they are likely to adopt specific actions or strategies aligning with these perceptions. For instance, an interview study with 38 GPs revealed that physicians who viewed themselves as the primary "decision-making and instructing authority" actively discouraged patients from internet searches, believing that other sources of information that might conflict with their own should be eliminated. In contrast, GPs who did not emphasize their authoritative role chose to collaborate with internet-informed patients, recommending reliable internet information sites and jointly examining internet information without seeking to prevent patients from further internet search [47].

Being identified in several studies [45,47], defensive actions have been termed by scholars as *resistance* [51] or *negative mediation* [44]. Defensive actions can be categorized into 2 types. The first type involves refusing to discuss internet information, which can manifest as resistance to discussing such information, terminating the physician-patient relationship by referring patients to other specialists, or even suggesting an additional charge for discussing internet-based information. The second type is about devaluing internet health information. This includes actions like discrediting the internet as a source, devaluing the health information that patients obtain from the internet, and correcting patients' misbeliefs with the implication that they should stop searching on the internet for health information [41,44,45,47,51]. Some physicians described this approach as follows:

> *When patients tell me, "yes, but on the Internet,"... I always cut short: "On the Web you find everything and its opposite, so forget it all and listen to what I'm saying, which is the standard." [51]* [Male gynecologist aged 63 years]

> *If they come in and it's too much and it's too specialized.... I let them slug it out with the specialist. They're paid very special money to do this kind of work. [45]* [A focus group of family physicians]

Physicians have claimed to take defensive actions to reduce the risk of internet health information for their patients [51]. A survey of 104 American physicians indeed showed that their negative assessment of the quality of internet information was positively correlated with more defensive actions [44]. However, another survey study of 1050 American physicians found that their perception of the potential harm of internet information to health was positively associated with their feelings of challenged authority [43]. Hence, the defensive strategy can be seen as coping mechanism that helps physicians defend themselves from the challenges and potential emotional unpleasantness posed by internet-informed patients.

Physicians' Communicative Strategies: How Do Physicians Collaborate With Internet-Informed Patients?

Physicians who have a positive perception of their patients' internet searches tend to work with their patients more collaboratively [44,51,52]. Their actions with internet-informed patients promote physician-patient relations by making the best use of the internet. These physicians believed that patients can also provide information and negotiate health care decisions, and they acknowledged the value of internet information brought by patients to the consultation [29]. As stated by the physicians, they would show interest in the information that patients bring, make patients feel respected and listened to, and examine patients' information. In addition, they believed that recommending reliable internet health information sites is an excellent way to guide patients in using internet information more effectively [39,52,53]. They also emphasized that the internet will never replace the human touch that physicians offer. They claimed to provide internet-informed patients with holistic care and to build ongoing relationships based on trust and familiarity [29].

While participative physicians acknowledged the risks associated with internet information, they regarded the internet as a powerful and legitimate health information source. Thus, they saw it as an opportunity for patients and believed that physicians should guide them in using this medium appropriately [51]. These actions can be categorized as the participative strategy, as opposed to the defensive one. Physicians take participative actions to "join in" with internet-informed patients in a way that allows them to guide and help patients navigate the internet with physicians' professional expertise. The participative approach is described as follows:

> *Yeah for me, for instance, the use of sites, I know patients when they come to you and you have to provide information they usually get shocked first to get a diagnosis and second to start treatment. And so I give them readings. I print some information for them and tell them if they have more questions to go*

XSL•FO
RenderX

> *to these sites and then you come back with me and we can discuss it if you want. [52]* [Male physician and scientist, 29 years in practice]

> *Often people find information you don't have. Why? We're doctors, we're constantly being updated, [but] we're not up to date on everything. We don't know everything about everything.... I actually like it when someone comes and tells me something I know nothing about, and I leave his room, and I go straight to the computer...and start reading about what they said. And if it's relevant, I can sometimes find myself incorporating [it] into my work. Definitely.... They come, they tell you something, I've never heard about it, I'm willing to check. And then after you check, you become a slightly better doctor. Because you know more. [29]* [Cardiologist]

Apart from physicians' self-claims, 1 study analyzed audio-recorded real clinical consultations between oncologists and patients with breast cancer in Switzerland [22]. The authors observed a series of participative actions with internet-informed patients, including physicians encouraging patients to use the internet, acknowledging the internet as an information tool, helping patients identify reliable internet information sources, providing detailed information to clarify patients' internet information, and admitting their limited knowledge on the topic being discussed. These actions align closely with the participative approaches that physicians claimed in other studies [22].

Similar to physicians' mixed perceptions of internet-informed patients, some physicians have reported shifting between different communication strategies. An interview study with 17 Swiss physicians revealed that they would shift between participative and defensive strategies depending on their interactions with individual patients [51]. They tended to be cautious toward internet information when the patients demonstrated a low level of health literacy, as judged by the patient's or the family member's education level and the quality of information that the patient had brought. Furthermore, patients' disrespect for physicians' expertise and their reluctance to consider physicians' suggestions also prompted defensive actions from physicians [51]. This characterization of patients who provoke a defensive response coincides with the description of self-diagnosing or self-treating patients given by physicians who maintain more neutral or balanced perceptions regarding patients' internet searches. Physicians' communication strategies, much like their perceptions, are not strictly limited to being participative or defensive. They adapt their communication approaches to suit individual patients. One physician illustrated his adaptability between different patients as follows:

> *You need to do the right thing with the right person. With some people you take the time to look at the information together, to evaluate it together. But there are also situations where you say no, I don't want to go into it. You have to consider, evaluate and grade, you need to weed some things out and to keep others... A few times, when I was really exasperated, I have said: "Time is up!" Those people were arrogant, and I have said "Time is up!" Two of them then left really pissed, but after the third repetition of the same thing.... Some patients do not want to understand: They have their idea and they want me to agree with it. [51]* [Male urologist, aged 53 years]

Physicians' Difficulties and Needs

With the introduction of internet information to medical consultations, physicians have frequently expressed in interview studies that their traditional roles have expanded to include responsibilities as internet information interpreters or examiners [39,41,45,46,52,53]. Two survey studies indicated that physicians believed patients brought internet information to consultations largely to seek physicians' opinions on that information [42,43]. Internet-informed patients are typically perceived as more demanding in terms of information and time compared with regular patients [49,54,55]. Furthermore, while physicians with negative views of internet-informed patients are often reluctant to assume this new responsibility [45], even proactive physicians have expressed time constraints when addressing internet-informed patients within an already burdened health system [39].

In addition, some physicians highlighted challenges in dealing with extreme cases [47]. Interviews with UK family physicians revealed feelings of being "left alone" when dealing with patients who exhibit intense health anxiety and excessive internet information searches. These physicians felt ill-equipped, both in terms of psychological expertise and time, to handle such situations. They favored the development of structured training that aids in identifying such extreme cases. Another group of digitally engaged physicians expressed a need for an updated medical curriculum to assist them in better establishing relations with internet-informed patients [39]. However, the studies included in this review provide limited details on the specific training needs of physicians, preventing a detailed exploration of their preferred curriculum.

Several studies highlighted the physicians' desire to stay updated with emerging technologies and trustworthy internet information sources for patient reference [38,45,47,48]. For example, a survey study of 118 family physicians revealed that 58% saw a need for training on internet use [48]. In another mixed methods study, 53% of physicians expressed difficulty in staying informed about credible internet health sites [38]. One physician illustrated it as follows:

> *It is imperative that doctors are trained in internet usage. I rarely know which website to recommend to patients. [38]* [Not available]

Discussion

Principal Findings

This review has examined physicians' perceptions of patients' internet health information–seeking behavior and their communication experiences with internet-informed patients. We included a mix of qualitative and quantitative studies, which enabled us to gain a deeper understanding from physicians'

XSL•FO
RenderX

perspectives and explore connections between physicians' perceptions and their communicative approaches.

We first identified a diversity of perceptions among physicians regarding patients' internet searches. These perceptions can be viewed as a continuous spectrum of opinions ranging from positive to negative. While some physicians hold more positive or negative views on patients' internet searches, a greater number of physicians expressed balanced views toward internet-informed patients. They acknowledge the benefits of internet health information but also raise concerns. The findings indicate that physicians, in general, are not entirely resistant to interacting with internet-informed patients. Their potential in educating and empowering patients to manage their health in this digital age should be further investigated and emphasized. For instance, scholars have drawn attention to the role of physicians in addressing internet misinformation. They have argued that although laypeople lack the medical expertise to distinguish the quality of internet material, the internet can become an extremely helpful tool with the collaboration of physicians [17].

Discrepancies exist when comparing the perceptions of physicians and patients. Previous reviews have concluded that patients tend to believe that internet searches make their consultations with physicians more effective [13]. However, our review found that physicians generally perceive their experiences with internet-informed patients as more time-consuming and demanding [42]. Physicians are also commonly concerned about the risk of misinformed patients [52,54]. Their concerns are not only about the poor quality of internet information but also about the patient's capability to process a large volume of information. The divergences between patients' and physicians' perspectives highlight the importance of understanding both parties' viewpoints. Future policy makers and researchers should consider these differences while investigating internet use and physician-patient communication.

We also identified 2 types of communicative strategies that physicians adopted to deal with internet-informed patients. These strategies include the defensive approach and the participative approach, each containing a series of specific communicative actions. Physicians' perceptions are linked to their strategies: positive perceptions toward internet-informed patients lead to a participative approach, while negative perceptions lead to a defensive approach. The 2 strategies were primarily identified through interview studies with physicians and are, therefore, based on their self-claims. Future research should build on our findings and delve into the interaction process between physicians and patients. This exploration will provide a better understanding of how physicians implement these strategies in actual medical consultations and how various strategies impact both the patients' use of internet material and the physician-patient relationship.

Furthermore, physicians shift between the 2 strategies based on their interaction with specific patients. The identification of neutral physicians and physicians' shift between the 2 communicative strategies emphasizes the dyadic nature of physician-patient relations. On the one hand, patients are cautious about disclosing internet-sourced information to their physicians and constantly evaluate the possible reactions of physicians to their information [23,25]. On the other hand, physicians adopt different attitudes and communicative approaches based on their interaction with that particular patient [51]. The communication process between the patient and the physician plays an essential role in determining whether patients will introduce internet information and how the information will be responded to by the physician during a consultation.

This study also identified physicians' difficulties and needs with internet-informed patients. The most common difficulty is the time pressure to address information and time demands from internet-informed patients. The medical systems in many countries are already overwhelmed [59]; therefore, the time issue could be difficult to address. Other common needs include training on new technologies and being informed of reliable internet sites that can be recommended to patients [38,48,53]. Some physicians also raised a need for training to help them address extreme cases of patient internet searches and form a more satisfying relationship with internet-informed patients [39,47]. This finding provides new insights into the medical teaching curriculum. Introducing special training that showcases trustworthy medical websites recommended for patients to reference could better equip physicians for medical encounters with internet-informed patients.

However, it is important to note that these training needs were primarily demonstrated in studies where physicians held more positive views of the internet [38,39,48]. Therefore, the expressed training needs primarily come from physicians who are receptive to internet information and tend to adopt a participative strategy toward internet-informed patients. Their motivation for such training programs likely revolves around enhancing their collaborative skills with internet-informed patients. Future research should take into account the perspectives of physicians with negative views, as they might have different concerns and might not be in favor of training specifically designed to enhance communication with internet-informed patients. Nonetheless, it is worth noting that newer medical graduates are more receptive to internet-informed patients compared with senior physicians [45]. Some medical educators have considered implementing communication training to prepare medical students for future consultations with internet-informed patients [60], and it has been observed that such training boosts the competency of medical students [61].

Limitations

Our review has certain limitations. First, the specialization of physicians may influence their perspectives on internet-informed patients; however, we could not draw concrete conclusions regarding the differences among medical specialists, as only a few of the included studies had examined them. In addition, some included studies focused solely on GPs or oncologists, which might skew the prevalence of their perspectives in this review and potentially diminish the applicability of our results to other specialists.

Second, most of the included studies are from Europe and North America, with few focusing on physicians from Eastern countries, such as South Korea and Israel. This makes it

XSL•FO
RenderX

challenging to draw further comparisons between different cultures. Particularly, in some East Asian cultures dominated by Confucianism, strong hierarchies between physicians and patients still exist [62]. The impact of internet information on changing physician-patient relationships in these cultures was found to be limited [24]. Physicians from such Eastern cultures might exhibit different perceptions and communicative actions toward internet-informed patients compared with their Western counterparts, where the concept of an equal physician-patient relationship is more common.

Third, the data used to observe communicative strategies primarily originated from qualitative interviews with physicians. Using quantitative data and observations from real-life consultations could strengthen our results, which are currently based on physicians' self-claims.

Conclusions

This review underscores the varied perceptions physicians hold toward internet-informed patients. Physicians' choice of communication strategies, whether adopting a participative or defensive approach, is intricately linked to their perceptions and their interactions with individual patients. Incorporating a medical teaching curriculum that introduces reliable internet sites to physicians for patient reference and provides updates on technology can potentially assist physicians in better coping with internet-informed patients.

Acknowledgments

The authors wish to extend their deepest appreciation to Professor Peter Salmon for his invaluable comments and insights on this review. The authors received no financial support for the research, authorship, or publication of this review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Keyword combinations and databases.
[DOCX File , 365 KB - jmir_v26i1e47620_app1.docx]

Multimedia Appendix 2
Description of themes and supporting evidence.
[DOCX File , 652 KB - jmir_v26i1e47620_app2.docx]

Multimedia Appendix 3
Quality assessment.
[DOCX File , 371 KB - jmir_v26i1e47620_app3.docx]

Multimedia Appendix 4
PRISMA checklist.
[DOCX File , 32 KB - jmir_v26i1e47620_app4.docx]

References

1. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175 [FREE Full text] [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
2. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. Cogent Soc Sci 2017 Mar 13;3(1):1302785 [FREE Full text] [doi: 10.1080/23311886.2017.1302785]
3. Ferguson T. e-patients: how they can help us heal healthcare. The e-Patient Scholars Working Group. 2007. URL: https://participatorymedicine.org/e-Patient_White_Paper_with_Afterword.pdf [accessed 2024-05-06]
4. Caiata-Zufferey M, Abraham A, Sommerhalder K, Schulz PJ. Online health information seeking in the context of the medical consultation in Switzerland. Qual Health Res 2010 Aug;20(8):1050-1061 [FREE Full text] [doi: 10.1177/1049732310368404] [Medline: 20442347]
5. Attfield SJ, Adams A, Blandford A. Patient information needs: pre- and post-consultation. Health Informatics J 2006 Jun;12(2):165-177 [FREE Full text] [doi: 10.1177/1460458206063811] [Medline: 17023406]
6. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 2014 Dec 02;16(12):e262 [FREE Full text] [doi: 10.2196/jmir.3706] [Medline: 25470306]

XSL•FO
RenderX

7. Bell RA, Hu X, Orrange SE, Kravitz RL. Lingering questions and doubts: online information-seeking of support forum members following their medical visits. Patient Educ Couns 2011 Dec;85(3):525-528. [doi: 10.1016/j.pec.2011.01.015] [Medline: 21315538]
8. Sjöström AE, Hörnsten Å, Hajdarevic S, Emmoth A, Isaksson U. Primary health care nurses' experiences of consultations with internet-informed patients: qualitative study. JMIR Nurs 2019;2(1):e14194 [FREE Full text] [doi: 10.2196/14194] [Medline: 34345770]
9. Miles A, Voorwinden S, Chapman S, Wardle J. Psychologic predictors of cancer information avoidance among older adults: the role of cancer fear and fatalism. Cancer Epidemiol Biomarkers Prev 2008 Aug;17(8):1872-1879 [FREE Full text] [doi: 10.1158/1055-9965.EPI-08-0074] [Medline: 18708374]
10. Chae J. Who avoids cancer information? Examining a psychological process leading to cancer information avoidance. J Health Commun 2016 Jul;21(7):837-844 [FREE Full text] [doi: 10.1080/10810730.2016.1177144] [Medline: 27337343]
11. Wang J, Ashvetiya T, Quaye E, Parakh K, Martin SS. Online health searches and their perceived effects on patients and patient-clinician relationships: a systematic review. Am J Med 2018 Oct;131(10):1250.e1-125.e10 [FREE Full text] [doi: 10.1016/j.amjmed.2018.04.019] [Medline: 29730361]
12. Luo A, Qin L, Yuan Y, Yang Z, Liu F, Huang P, et al. The effect of online health information seeking on physician-patient relationships: systematic review. J Med Internet Res 2022 Feb 10;24(2):e23354 [FREE Full text] [doi: 10.2196/23354] [Medline: 35142620]
13. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]
14. McMullan M. Patients using the internet to obtain health information: how this affects the patient-health professional relationship. Patient Educ Couns 2006 Oct;63(1-2):24-28 [FREE Full text] [doi: 10.1016/j.pec.2005.10.006] [Medline: 16406474]
15. Broom A. Virtually he@lthy: the impact of internet use on disease experience and the doctor-patient relationship. Qual Health Res 2005 Mar;15(3):325-345 [FREE Full text] [doi: 10.1177/1049732304272916] [Medline: 15761103]
16. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
17. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 02;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
18. Te Poel F, Baumgartner SE, Hartmann T, Tanis M. The curious case of cyberchondria: a longitudinal study on the reciprocal relationship between health anxiety and online health information seeking. J Anxiety Disord 2016 Oct;43:32-40 [FREE Full text] [doi: 10.1016/j.janxdis.2016.07.009] [Medline: 27497667]
19. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278 [FREE Full text] [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
20. Schulz PJ, Nakamoto K. The perils of misinformation: when health literacy goes awry. Nat Rev Nephrol 2022 Mar;18(3):135-136 [FREE Full text] [doi: 10.1038/s41581-021-00534-z] [Medline: 34997231]
21. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. The influence of online health information on health decisions: a systematic review. Patient Educ Couns 2021 Apr;104(4):770-784 [FREE Full text] [doi: 10.1016/j.pec.2020.11.016] [Medline: 33358253]
22. Shen MJ, Dyson RC, D'Agostino TA, Ostroff JS, Dickler MN, Heerdt AS, et al. Cancer-related internet information communication between oncologists and patients with breast cancer: a qualitative study. Psychooncology 2015 Nov;24(11):1439-1447 [FREE Full text] [doi: 10.1002/pon.3752] [Medline: 25631285]
23. Hay MC, Cadigan RJ, Khanna D, Strathmann C, Lieber E, Altman R, et al. Prepared patients: internet information seeking by new rheumatology patients. Arthritis Rheum 2008 Apr 15;59(4):575-582 [FREE Full text] [doi: 10.1002/art.23533] [Medline: 18383399]
24. Chiu YC. Probing, impelling, but not offending doctors: the role of the internet as an information source for patients' interactions with doctors. Qual Health Res 2011 Dec;21(12):1658-1666 [FREE Full text] [doi: 10.1177/1049732311417455] [Medline: 21799204]
25. Silver MP. Patient perspectives on online health information and communication with doctors: a qualitative study of patients 50 years old and over. J Med Internet Res 2015 Jan 13;17(1):e19 [FREE Full text] [doi: 10.2196/jmir.3588] [Medline: 25586865]
26. Bylund CL, Gueguen JA, Sabee CM, Imes RS, Li Y, Sanford AA. Provider-patient dialogue about internet health information: an exploration of strategies to improve the provider-patient relationship. Patient Educ Couns 2007 Jun;66(3):346-352 [FREE Full text] [doi: 10.1016/j.pec.2007.01.009] [Medline: 17329062]
27. Petrocchi S, Iannello P, Lecciso F, Levante A, Antonietti A, Schulz PJ. Interpersonal trust in doctor-patient relation: evidence from dyadic analysis and association with quality of dyadic communication. Soc Sci Med 2019 Aug;235:112391 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112391] [Medline: 31301438]

XSL•FO
RenderX

28. Bowes P, Stevenson F, Ahluwalia S, Murray E. ‘I need her to be a doctor’: patients’ experiences of presenting health information from the internet in GP consultations. Br J Gen Pract 2012 Nov 01;62(604):e732-e738 [FREE Full text] [doi: 10.3399/bjgp12x658250]
29. Shachar L. "You become a slightly better doctor": doctors adopting integrated medical expertise through interactions with E-patients. Soc Sci Med 2022 Jul;305:115038 [FREE Full text] [doi: 10.1016/j.socscimed.2022.115038] [Medline: 35598443]
30. Wald HS, Dube CE, Anthony DC. Untangling the web--the impact of internet use on health care and the physician-patient relationship. Patient Educ Couns 2007 Nov;68(3):218-224 [FREE Full text] [doi: 10.1016/j.pec.2007.05.016] [Medline: 17920226]
31. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res 2016 Aug 26;16(1):442 [FREE Full text] [doi: 10.1186/s12913-016-1691-0] [Medline: 27562728]
32. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336-341 [FREE Full text] [doi: 10.1016/j.ijsu.2010.02.007] [Medline: 20171303]
33. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
34. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: a review of possible methods. J Health Serv Res Policy 2005 Jan;10(1):45-53 [FREE Full text] [doi: 10.1177/135581960501000110] [Medline: 15667704]
35. Lawrence RA, McLoone JK, Wakefield CE, Cohn RJ. Primary care physicians' perspectives of their role in cancer care: a systematic review. J Gen Intern Med 2016 Oct;31(10):1222-1236 [FREE Full text] [doi: 10.1007/s11606-016-3746-7] [Medline: 27220499]
36. van der Keylen P, Tomandl J, Wollmann K, Möhler R, Sofroniou M, Maun A, et al. The online health information needs of family physicians: systematic review of qualitative and quantitative studies. J Med Internet Res 2020 Dec 30;22(12):e18816 [FREE Full text] [doi: 10.2196/18816] [Medline: 33377874]
37. Helft PR, Hlubocky F, Daugherty CK. American oncologists’ views of internet use by cancer patients: a mail survey of American Society of Clinical Oncology members. J Clin Oncol 2003 Mar 01;21(5):942-947. [doi: 10.1200/jco.2003.08.007]
38. van Uden-Kraan CF, Drossaert CH, Taal E, Smit WM, Seydel ER, van de Laar MA. Experiences and attitudes of Dutch rheumatologists and oncologists with regard to their patients' health-related internet use. Clin Rheumatol 2010 Nov;29(11):1229-1236 [FREE Full text] [doi: 10.1007/s10067-010-1435-1] [Medline: 20383731]
39. Győrffy Z, Radó N, Mesko B. Digitally engaged physicians about the digital health transition. PLoS One 2020;15(9):e0238658 [FREE Full text] [doi: 10.1371/journal.pone.0238658] [Medline: 32986733]
40. Ahluwalia S, Murray E, Stevenson F, Kerr C, Burns J. ‘A heartbeat moment’: qualitative study of GP views of patients bringing health information from the internet to a consultation. Br J Gen Pract 2010 Feb 01;60(571):88-94 [FREE Full text] [doi: 10.3399/bjgp10x483120]
41. Fredriksen EH, Moland KM, Harris J. How do health professionals acknowledge web-based knowledge in pregnancy consultations? Health Care Women Int 2018 Sep;39(9):955-967 [FREE Full text] [doi: 10.1080/07399332.2017.1423314] [Medline: 29313795]
42. Kim J, Kim S. Physicians' perception of the effects of internet health information on the doctor-patient relationship. Inform Health Soc Care 2009 Sep;34(3):136-148 [FREE Full text] [doi: 10.1080/17538150903102422] [Medline: 19670004]
43. Murray E, Lo B, Pollack L, Donelan K, Catania J, Lee K, et al. The impact of health information on the internet on health care and the physician-patient relationship: national U.S. survey among 1.050 U.S. physicians. J Med Internet Res 2003;5(3):e17 [FREE Full text] [doi: 10.2196/jmir.5.3.e17] [Medline: 14517108]
44. Fujioka Y, Stewart E. How do physicians discuss e-health with patients? The relationship of physicians' e-health beliefs to physician mediation styles. Health Commun 2013 May;28(4):317-328. [doi: 10.1080/10410236.2012.682971] [Medline: 22716050]
45. Ahmad F, Hudak PL, Bercovitz K, Hollenberg E, Levinson W. Are physicians ready for patients with internet-based health information? J Med Internet Res 2006 Sep 29;8(3):e22 [FREE Full text] [doi: 10.2196/jmir.8.3.e22] [Medline: 17032638]
46. Sommerhalder K, Abraham A, Zufferey MC, Barth J, Abel T. Internet information and medical consultations: experiences from patients' and physicians' perspectives. Patient Educ Couns 2009 Nov;77(2):266-271 [FREE Full text] [doi: 10.1016/j.pec.2009.03.028] [Medline: 19411157]
47. Wangler J, Jansky M. General practitioners' challenges and strategies in dealing with internet-related health anxieties-results of a qualitative study among primary care physicians in Germany. Wien Med Wochenschr 2020 Oct;170(13-14):329-339 [FREE Full text] [doi: 10.1007/s10354-020-00777-8] [Medline: 32767159]
48. Giveon S, Yaphe J, Hekselman I, Mahamid S, Hermoni D. The e-patient: a survey of israeli primary care physicians' responses to patients' use of online information during the consultation. Isr Med Assoc J 2009 Sep;11(9):537-541 [FREE Full text] [Medline: 19960847]

XSL•FO
RenderX

49. Masters K, Loda T, Johannink J, Al-Abri R, Herrmann-Werner A. Surgeons' interactions with and attitudes toward e-patients: questionnaire study in Germany and Oman. J Med Internet Res 2020 Mar 09;22(3):e14646 [FREE Full text] [doi: 10.2196/14646] [Medline: 32149714]
50. da Mota LR, Ferreira CC, Costa Neto HA, Falbo AR, Lorena SB. Is doctor-patient relationship influenced by health online information? Rev Assoc Med Bras (1992) 2018 Aug;64(8):692-699 [FREE Full text] [doi: 10.1590/1806-9282.64.08.692] [Medline: 30673039]
51. Caiata-Zufferey M, Schulz PJ. Physicians' communicative strategies in interacting with internet-informed patients: results from a qualitative study. Health Commun 2012;27(8):738-749 [FREE Full text] [doi: 10.1080/10410236.2011.636478] [Medline: 22260422]
52. Macdonald GG, Townsend AF, Adam P, Li LC, Kerr S, McDonald M, et al. eHealth technologies, multimorbidity, and the office visit: qualitative interview study on the perspectives of physicians and nurses. J Med Internet Res 2018 Jan 26;20(1):e31 [FREE Full text] [doi: 10.2196/jmir.8983] [Medline: 29374004]
53. Schick TS, Höllerl L, Biedermann T, Zink A, Ziehfreund S. Impact of digital media on the patient journey and patient-physician relationship among dermatologists and adult patients with skin diseases: qualitative interview study. J Med Internet Res 2023 Sep 22;25:e44129 [FREE Full text] [doi: 10.2196/44129] [Medline: 37738078]
54. Potts HW, Wyatt JC. Survey of doctors' experience of patients using the internet. J Med Internet Res 2002;4(1):e5 [FREE Full text] [doi: 10.2196/jmir.4.1.e5] [Medline: 11956037]
55. Ohana S, Barnoy S. Israeli e-patients' informational needs. Nurs Outlook 2019;67(2):190-198 [FREE Full text] [doi: 10.1016/j.outlook.2018.11.003] [Medline: 30558901]
56. Moick M, Terlutter R. Physicians' motives for professional internet use and differences in attitudes toward the internet-informed patient, physician-patient communication, and prescribing behavior. Med 2 0 2012;1(2):e2 [FREE Full text] [doi: 10.2196/med20.1996] [Medline: 25075230]
57. Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C, et al. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. J Clin Epidemiol 2012 Sep;65(9):934-939 [FREE Full text] [doi: 10.1016/j.jclinepi.2011.11.014] [Medline: 22742910]
58. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Res Method Med Health Sci 2020 Aug 06;1(1):31-42 [FREE Full text] [doi: 10.1177/2632084320947559]
59. Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open 2017 Nov 08;7(10):e017902 [FREE Full text] [doi: 10.1136/bmjopen-2017-017902] [Medline: 29118053]
60. Masters K. Preparing medical students for the e-patient. Med Teach 2017 May 22;39(7):681-685 [FREE Full text] [doi: 10.1080/0142159x.2017.1324142]
61. Herrmann-Werner A, Weber H, Loda T, Keifenheim KE, Erschens R, Mölbert SC, et al. “But Dr Google said…” – Training medical students how to communicate with e-patients. Med Teach 2019 Feb 01;41(12):1434-1440 [FREE Full text] [doi: 10.1080/0142159x.2018.1555639]
62. Pun JK, Chan EA, Wang S, Slade D. Health professional-patient communication practices in East Asia: an integrative review of an emerging field of research and practice in Hong Kong, South Korea, Japan, Taiwan, and Mainland China. Patient Educ Couns 2018 Jul;101(7):1193-1206 [FREE Full text] [doi: 10.1016/j.pec.2018.01.018] [Medline: 29422231]

Abbreviations

CASP: Critical Appraisal Skills Program
GP: general practitioner
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

Edited by A Mavragani; submitted 27.03.23; peer-reviewed by K Masters, R Poss-Doering, R McMullan; comments to author 09.09.23; revised version received 01.10.23; accepted 27.02.24; published 06.06.24.

Please cite as:
Lu Q, Schulz PJ
Physician Perspectives on Internet-Informed Patients: Systematic Review
J Med Internet Res 2024;26:e47620
URL: https://www.jmir.org/2024/1/e47620
doi:10.2196/47620
PMID:38842920

XSL•FO
RenderX

©Qianfeng Lu, Peter Johannes Schulz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Integrated Real-World Data Warehouses Across 7 Evolving Asian Health Care Systems: Scoping Review

Wen-Yi Shau[1*], MD, PhD; Handoko Santoso[1], MD; Vincent Jip[2], BE; Sajita Setia[2*], MD

[1]Regional Medical Affairs, Pfizer Corporation Hong Kong Limited, Hong Kong, Hong Kong, China (Hong Kong)

[2]Transform Medical Communications Limited, Wanganui, New Zealand

[*]these authors contributed equally

Corresponding Author:
Sajita Setia, MD
Transform Medical Communications Limited
184 Glasgow Street
Wanganui, 4500
New Zealand
Phone: 64 276175433
Email: sajita.setia@transform-medcomms.com

Abstract

Background: Asia consists of diverse nations with extremely variable health care systems. Integrated real-world data (RWD) research warehouses provide vast interconnected data sets that uphold statistical rigor. Yet, their intricate details remain underexplored, restricting their broader applications.

Objective: Building on our previous research that analyzed integrated RWD warehouses in India, Thailand, and Taiwan, this study extends the research to 7 distinct health care systems: Hong Kong, Indonesia, Malaysia, Pakistan, the Philippines, Singapore, and Vietnam. We aimed to map the evolving landscape of RWD, preferences for methodologies, and database use and archetype the health systems based on existing intrinsic capability for RWD generation.

Methods: A systematic scoping review methodology was used, centering on contemporary English literature on PubMed (search date: May 9, 2023). Rigorous screening as defined by eligibility criteria identified RWD studies from multiple health care facilities in at least 1 of the 7 target Asian nations. Point estimates and their associated errors were determined for the data collected from eligible studies.

Results: Of the 1483 real-world evidence citations identified on May 9, 2023, a total of 369 (24.9%) fulfilled the requirements for data extraction and subsequent analysis. Singapore, Hong Kong, and Malaysia contributed to ≥100 publications, with each country marked by a higher proportion of single-country studies at 51% (80/157), 66.2% (86/130), and 50% (50/100), respectively, and were classified as *solo scholars*. Indonesia, Pakistan, Vietnam, and the Philippines had fewer publications and a higher proportion of cross-country collaboration studies (CCCSs) at 79% (26/33), 58% (18/31), 74% (20/27), and 86% (19/22), respectively, and were classified as *global collaborators*. Collaboration with countries outside the 7 target nations appeared in 84.2% to 97.7% of the CCCSs of each nation. Among target nations, Singapore and Malaysia emerged as preferred research partners for other nations. From 2018 to 2023, most nations showed an increasing trend in study numbers, with Vietnam (24.5%) and Pakistan (21.2%) leading the growth; the only exception was the Philippines, which declined by –14.5%. Clinical registry databases were predominant across all CCCSs from every target nation. For single-country studies, Indonesia, Malaysia, and the Philippines favored clinical registries; Singapore had a balanced use of clinical registries and electronic medical or health records, whereas Hong Kong, Pakistan, and Vietnam leaned toward electronic medical or health records. Overall, 89.9% (310/345) of the studies took >2 years from completion to publication.

Conclusions: The observed variations in contemporary RWD publications across the 7 nations in Asia exemplify distinct research landscapes across nations that are partially explained by their diverse economic, clinical, and research settings. Nevertheless, recognizing these variations is pivotal for fostering tailored, synergistic strategies that amplify RWD's potential in guiding future health care research and policy decisions.

International Registered Report Identifier (IRRID): RR2-10.2196/43741

(J Med Internet Res 2024;26:e56686) doi:10.2196/56686

XSL•FO
RenderX

KEYWORDS

Asia; health care databases; cross-country comparison; electronic health records; electronic medical records; data warehousing; information storage and retrieval; real-world data; real-world evidence; registries; scoping review

Introduction

Background

Asia is a vast and diverse continent that also represents varied health care systems and socioeconomic challenges. Multiple evidence-driven approaches tailored to each nation's unique health care and research context are required to draw essential data to support the ambitious goals for universal health coverage in each country [1,2]. The strength and necessity of real-world data (RWD) and their concrete data analytical interference in terms of real-world evidence (RWE) are integral to this evidence generation. RWE has the potential to inform health technology assessments (HTAs), guide evidence-driven policies, and streamline service delivery [3]. However, as crucial as RWE is, the Asian health care landscape lacks a cohesive framework to harness its full potential despite its promise in pharmacoeconomics, pharmacovigilance, and pharmacoepidemiology [3,4].

The utility of RWD and RWE becomes even more apparent with large integrated research databases within health systems. The integrated warehouses offer vast connected data sets that sustain the statistical rigor and can assist in providing insights with minimal bias and confounding [5]. However, these data reservoirs have not been vastly studied across health care systems, which limits their broader utility in health care research; RWE data generation; and, consequently, universal health coverage [6].

Recognizing this potential, our previous research explored integrated RWD warehouses within 3 diverse health care pilots for Taiwan, India, and Thailand [7]. Our systematic research identified strong differences in the types of RWD and their warehouses in the 3 countries. Still, the results only partly reflected their divergent economic, social, and clinical settings. Hence, we continued to conduct similar research in many other diverse Asian health care systems in line with our published protocol [8].

Objectives

The literature on RWD practices and awareness of corresponding warehouses in certain Asian countries such as China, Japan, and South Korea is significant [5,9-17], partly because these countries also have recommendations on the utility of RWE by external regulators [3]. This study sought to understand the evolving landscape of RWD use and its implications across Hong Kong, Indonesia, Malaysia, Pakistan, the Philippines, Singapore, and Vietnam, where RWD practices are emerging or undergoing significant development [4]. Our selection of countries for this scoping review was strategically based on selecting a contrasting spectrum of HTA maturity across countries with evolving HTA systems, ranging from relatively mature systems in Singapore, Thailand, and Malaysia to emerging frameworks in Indonesia, the Philippines, and Vietnam and systems in the nascent stages in Pakistan [18,19]. Each nation, with its individualistic health care challenges and unique research capabilities, underscores the need for understanding recent patterns in RWD research and use of clinical research warehouses, especially in light of the marked underrepresentation of specific Asian demographics in traditional randomized clinical trials [20]. By systematically analyzing both single-country studies (SCSs) and cross-country collaboration studies (CCCSs), this research aimed to delineate the current state of RWE generation and collaborative research initiatives for RWE from integrated databases across different nations in Asia. Our objectives also included obtaining a comprehensive understanding of the preference for RWD methodologies by contrasting the emphasis on comparative effectiveness research (CER) with descriptive studies and discerning the preferred and popular real-world research databases.

The cyclical interplay between a nation's economic strength, health care infrastructure, and research capacity perpetuates disparities in RWD generation. We hypothesized that Asian countries with less extensively documented RWD research trends could be effectively clustered based on systematic patterns in RWD generation. This streamlined our objective to evaluate trends in RWD generation and shed light on targeted capacity-building strategies essential for informed health care policy making. Through this rigorous extended scoping research, we aimed to present insights that resonate with clinical stakeholders, medical researchers, and health policy makers, thereby guiding the formulation of strategies attuned to each nation's health care challenges and research diversities and complexities.

Methods

Research Approach

Our research approach was methodically aligned with the guidelines set forth by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [21]. Our published protocol specified a preliminary focus on 3 countries—India, Thailand, and Taiwan—as a representative pilot to explore the diversity of health care systems and RWD use in Asia [8]. The outcomes from our initial study covering Taiwan, India, and Thailand have been previously published [7]. Relevant insights from the latter publication were incorporated into the archetyping of the nations wherever applicable. In this study, we expanded our protocol to 7 other countries. However, we maintained consistency with the original protocol's methodological framework to ensure comparability across all countries studied. This expansion was aligned with our initial intent to potentially include more countries following the first study across 3 countries.

The search strategy is described in Table S1 in Multimedia Appendix 1. We filtered our search to include only English-language publications from the last 5 years, aiming to

XSL•FO
RenderX

highlight current and internationally relevant RWE or RWD. As the conversion of RWD to RWE emphasizes the stringent analytical processes necessary to yield valuable and credible findings, we intentionally chose to rely on PubMed as an exclusive source of relevant citations for screening. Our goal was to assess studies yielding robust RWD featured in esteemed, indexed, and peer-reviewed journals while reducing potential duplicates. By focusing solely on PubMed, we tried to identify research representing this standard and offering evidence of the utmost scientific integrity. This strategy aligns with the specifications outlined in our protocol [8].

Screening Eligible Studies for Data Analysis

All retrieved study abstracts were directly imported into the Covidence software (Veritas Health Innovation) for subsequent screening and data extraction. Studies were initially screened against predefined eligibility criteria to capture research from integrated RWD. The criteria encompassed 4 domains described in the original protocol: database type and requirement for research across >1 hospital or clinic, publication nature, RWD study type, and publication scope [8]. The scope of publication was adapted in this study to include citations on databases involving 1 of the target nations (Hong Kong, Indonesia, Malaysia, Pakistan, Philippines, Singapore, or Vietnam). The inclusion criteria also considered studies featuring nontarget countries as long as 1 of the 7 target nations was involved. Table S2 in Multimedia Appendix 1 provides a snapshot of the eligibility criteria used in this study.

Duplicate removal and a 2-step eligibility screening process were conducted in Covidence. The initial step (phase 1) assessed titles and abstracts, with relevant studies advancing to full-text evaluation in the second phase. Given the study volume, the screening for both phases was divided between 2 reviewers. An independent reviewer examined a random sample of 20% of the studies to maintain accuracy. Any ambiguities or discrepancies were collaboratively resolved, and another reviewer was consulted if needed. The final step involved data extraction and data analysis for eligible studies.

Data Extraction and Analysis

We used Covidence for data extraction through a custom template that covered the following:

1. Basic study details: Covidence identifier (ID) based on the first author's last name and publication year and title.
2. Presence of cross-country collaboration in the research (CCCS or SCS).
3. Nature of publication (clinical study or protocol).
4. Study categorization: CER versus descriptive study (non-CER), with CER definitions adapted from Medical Subject Headings. We expanded the criteria for CER to standardize its meaning in the context of this research as the "studies comparing interventions and strategies (including the comparison between active and nonactive interventions and strategies) to prevent, diagnose, treat, and monitor health conditions using validated methods for confounders elimination, e.g., matching, and statistical adjustments like stratification, weighting, regression, instrumental variable analysis etc" [7].
5. Research source database classification involving medical records, health insurance claims, clinical registries, pharmacy claims, or composite databases.
6. Disease specifics: name and area of the target disease under study (defined by primary diagnosis and pathophysiology or by prime medical specialty in charge if they intersected). The disease categories encompassed cardiology and metabolic disorders (CVM), oncology, inflammatory and autoimmune disorders, infectious diseases and vaccines (IDV), and others. These categories represent major research fields in clinical medicine with significant disease burdens, selected to provide pertinent insights into RWD and RWE applications within these critical domains.
7. Outcome types: clinical (treatment effect or safety), cost, or patient-reported outcomes (PROs), with PROs capturing direct patient responses.
8. Demographics (adults, children, or both), number of centers, study participants, and length of study and duration between last data collected and year of publication. The length or duration represented the span from the study's commencement to completion as specified by the authors.
9. The unique names for the databases used. When provided, the name of the specific database used in each study was collected and organized by target nation, database type, and disease area.

A total of 2 reviewers collaboratively managed data extraction, and all extractions underwent quality checks by another reviewer to ensure the accuracy and reliability of the extracted data. However, this process was not conducted independently or blinded to the other reviewer's decisions. Disagreements between the 2 reviewers were settled through discussions, and an additional reviewer was involved whenever there was a need for a consensus.

The final search was conducted on May 9, 2023, covering the preceding 5 years; to account for partial yearly data in 2018 and 2023, we calculated the equivalent of annual publication count using 365 multiplied by the average daily number of publications. We used linear regression, using the year as a continuous predictor variable, to understand the annual trend in nation study counts. This provided insights into the average annual trend in study numbers throughout the search period. To further even out year-to-year variations, a 2-year simple moving average (SMA) was applied to enhance the clarity of the data trends. This SMA approach was consistent with our previous research methodologies [7]. Given the study's descriptive nature, there was no a priori statistical hypothesis. Statistical analyses were conducted to calculate point estimates and their associated errors. Categorical data were presented as frequencies and percentages, whereas continuous data were presented as means and SDs. We used Microsoft Excel (Microsoft Corp) for all data analyses. Adobe Illustrator was used for crafting high-definition figures for the main manuscript.

Results

Eligible Studies

The search was conducted on May 9, 2023, and yielded 1483 studies with 1 duplicate. Of these 1483 studies, 553 (37.3%)

XSL•FO
RenderX

were included in phase-2 screening, and 369 (24.9%) studies were eligible for data extraction (Figure 1).

The vast majority of the publications (361/369, 97.8%) were original research, whereas the remaining 2.2% (8/369) were study protocols. The country-wise distribution of SCSs and CCCSs is illustrated in Multimedia Appendix 2.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the selection of eligible studies. *Search date: May 9, 2023. EHR: electronic health record; EMR: electronic medical record; PCT: pragmatic clinical trial; RCT: randomized clinical trial.

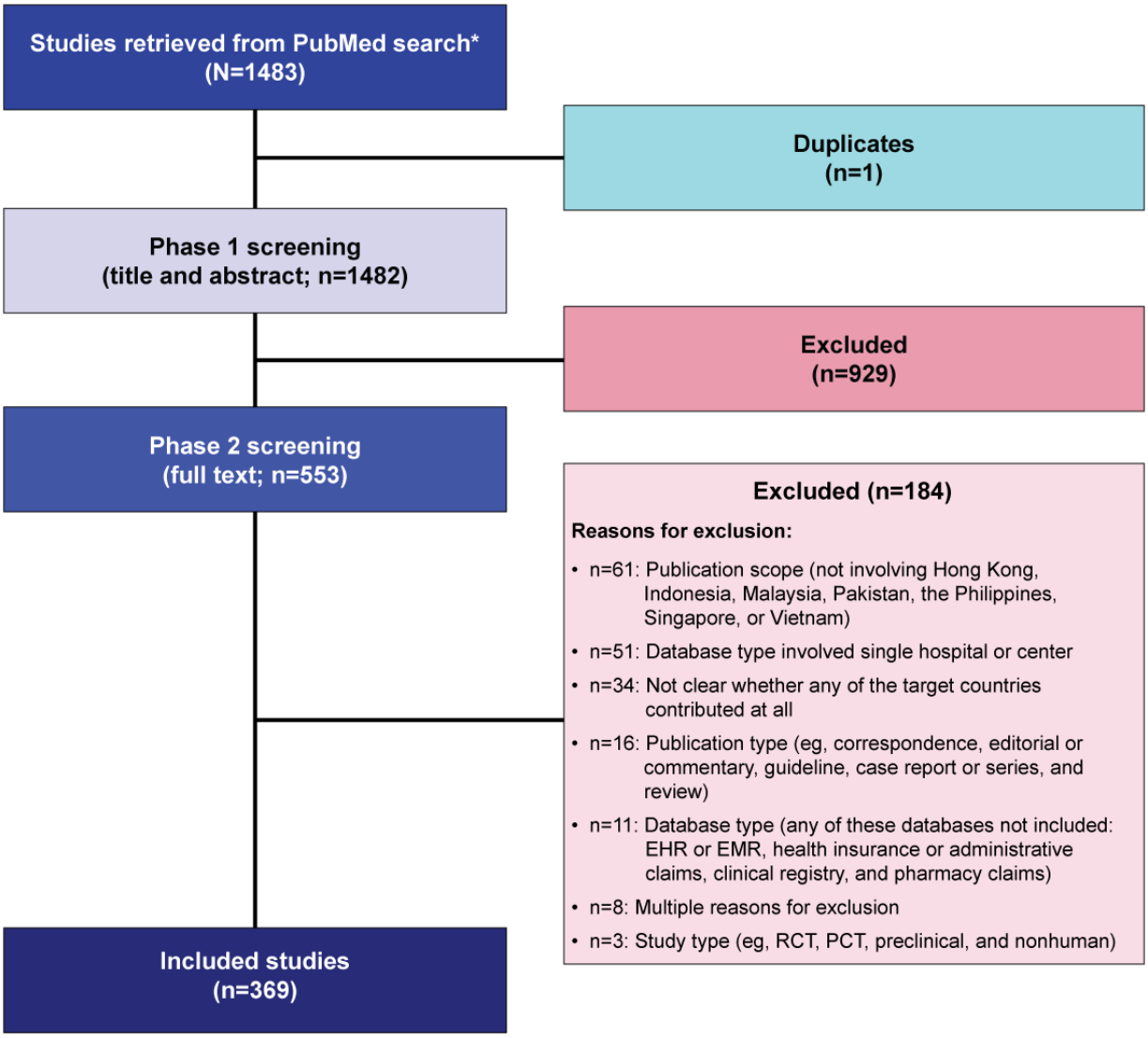

Geographic Distribution and Collaboration Relationship

Among the 369 studies that qualified for data extraction, Singapore, Hong Kong, and Malaysia each contributed to ≥100 publications, with respective counts of 157 (42.5%), 130 (35.2%), and 100 (27.1%). The other 4 nations—Indonesia, Pakistan, Vietnam, and the Philippines—were involved in fewer publications, with 8.9% (33/369), 8.4% (31/369), 7.3% (27/369), and 6% (22/369), respectively, and each had >50% of their studies classified as CCCSs (Table S3 in Multimedia Appendix 1). Given their lower overall study numbers and the predominance of CCCSs, these 4 nations were categorized as *global collaborators* in certain subsequent analyses (Figure 2 [7]).

XSL•FO
RenderX

Figure 2. Archetype by number of publications and percentage of cross-country collaboration studies (CCCSs) from real-world study databases in the last 5 years from 10 nations in Asia. Scatter plot archetyping for countries with a relatively high percentage of CCCSs but fewer total publications as global collaborators (highlighted within the dashed lines). Solo scholars cluster includes countries with a lower percentage of CCCSs (≤50%) and a higher number of publications (≥100 real-world data publications from integrated databases in the last 5 years). Countries such as Taiwan show a high number of total publications with a relatively low percentage of CCCSs, signifying a tendency to conduct independent research. The data for India, Taiwan, and Thailand were derived from our previous publication that used the same methodology as that used in this research [7].

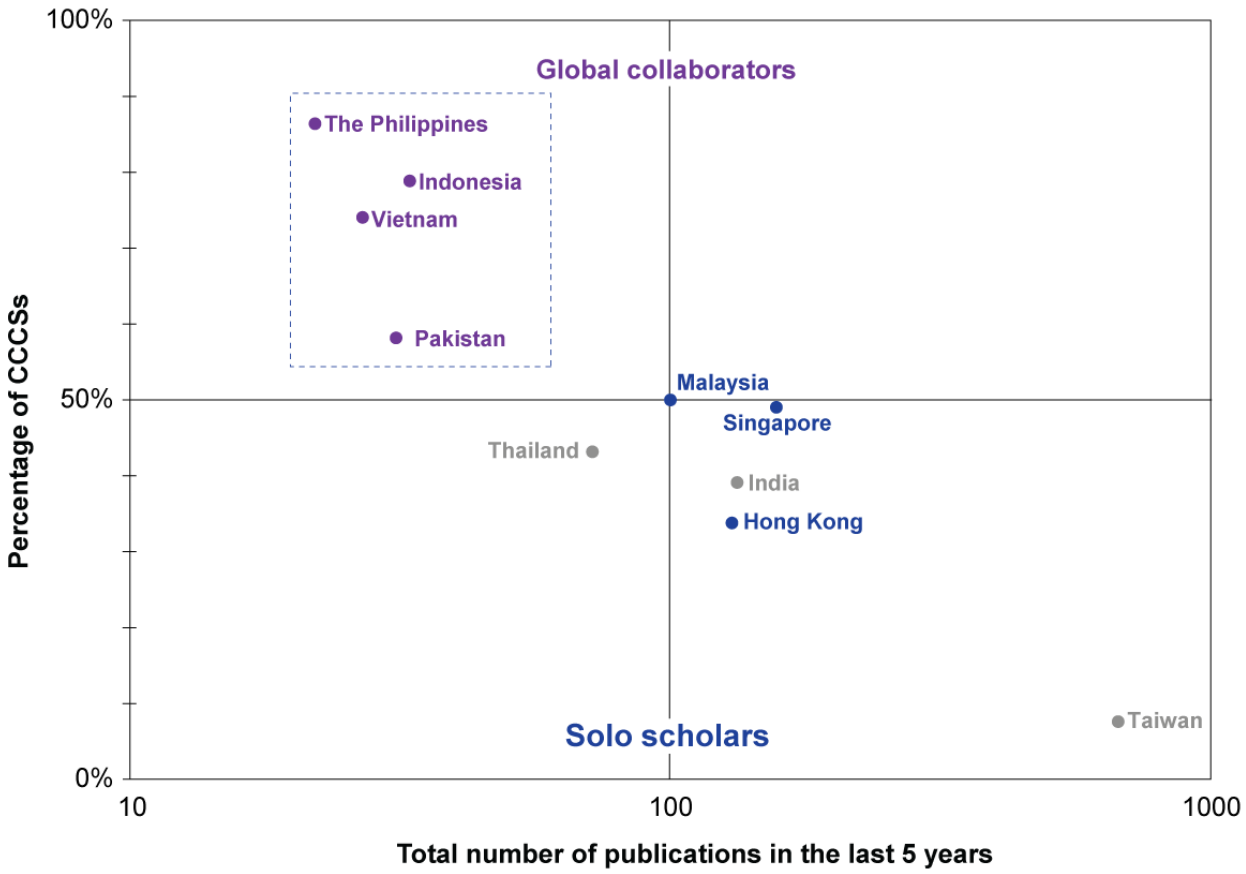

Countries beyond the 7 target nations of this study that were involved in collaborations were labeled as nontarget countries. The cross-country collaboration network across the 7 target nations and nontarget countries is described in Table S4 in Multimedia Appendix 1. The average number of collaborative countries (ANC) indicates the cross-country interconnection for research of a given nation. The ANC varied from 2.2 for Singapore to 4.1 for the Philippines. Despite having the lowest ANC, Singapore was involved in 30.3% (77/254) of the studies, making it the highest contributor to CCCSs. On the other hand, Malaysia, with a higher ANC of 3.1, participated in 50% (50/100) of CCCSs, making it the most engaged collaborator within the *solo scholars* cluster. Malaysia participated in 56% (10/18) of CCCSs from Pakistan, 75% (15/20) of CCCSs from Vietnam, 81% (21/26) of CCCSs from Indonesia, and 84% (16/19) of CCCSs from the Philippines.

Nontarget countries were common collaborators in CCCSs across all 7 target countries, with their involvement ranging from 84% (16/19) to 98% (43/44).

Time Trend

Figures 3 and 4 depict the yearly average counts and growth rates of SCSs and CCCSs across the 7 target nations from the *solo scholar* and *global collaborator* clusters, respectively. Between 2018 and 2023, every target nation except for the Philippines, which experienced a decline of –14.5%, exhibited an upward trend in the average number of studies published. Vietnam led with the steepest growth rate at 24.5%, trailed by Pakistan at 21.2%. Among solo scholars, there were growing trends for all 3 nations, with a growth rate of 6.2% for Hong Kong, 8.7% for Singapore, and 16.3% for Malaysia. Due to the small average number of studies by year in some of the individual global collaborator nations, growth trends are also presented collectively for global collaborators as a cluster in Figure 4. Duplicate studies within clusters were adjusted; thus, the sum of CCCSs from all 4 nations might be larger than the number of CCCSs of the cluster as whole. The growth rate was 21.2% in the global collaborators cluster after adjusting for duplicates.

XSL•FO
RenderX

Figure 3. Annual trends in average number of publications by geographical distribution for solo scholar nations. Malaysia showed the highest annual growth rate of 16.3% among the solo scholars cluster, followed by Singapore (8.7%) and Hong Kong (6.2%).

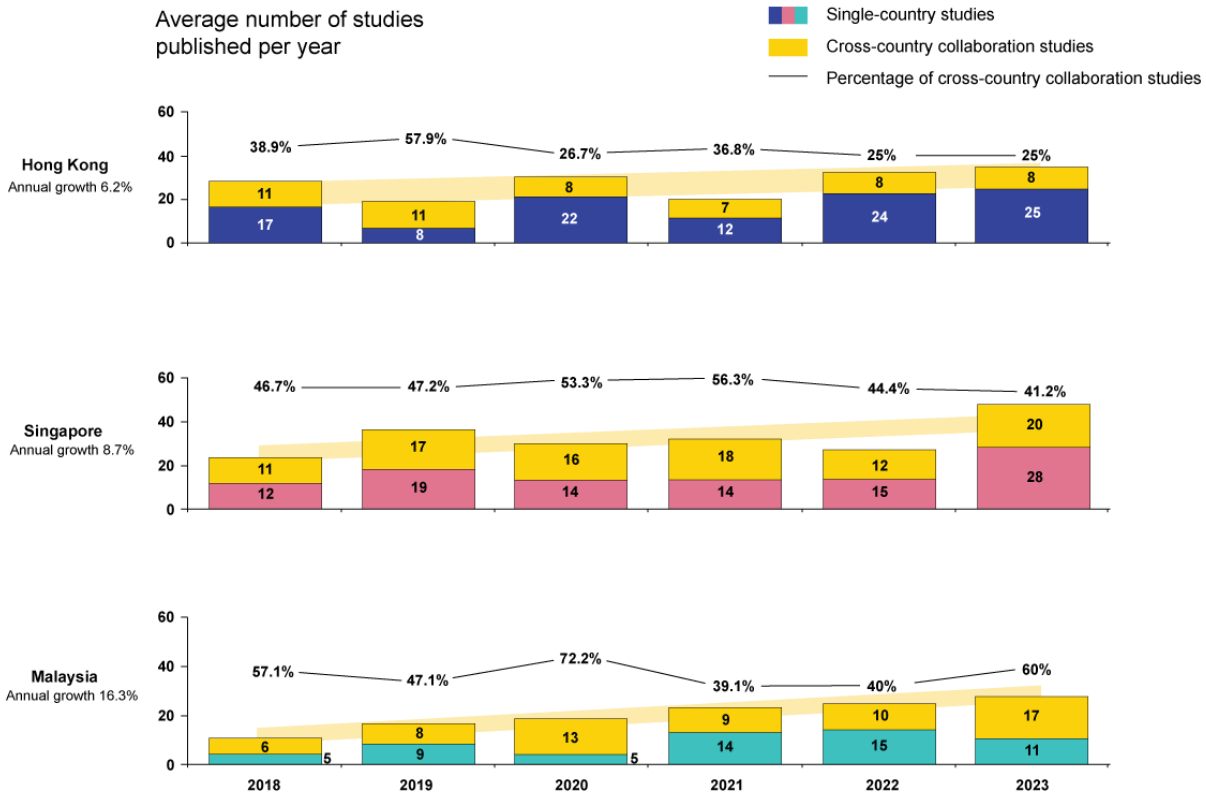

XSL•FO
RenderX

Figure 4. Annual trends in average number of publications by geographical distribution for global collaborator nations. Vietnam exhibited the most significant increase in total studies among the 7 target countries, followed by Pakistan and Malaysia. The Philippines was the only nation with a decline, whereas Indonesia and Hong Kong maintained a consistent study count. Vietnam and, to a lesser degree, Indonesia and Pakistan demonstrated a rising participation in cross-country collaborative research. Duplicate studies within the cluster were adjusted, and thus, the sum of cross-country collaboration studies (CCCSs) from all 4 nations might be larger than the number of CCCSs in the cluster as whole.

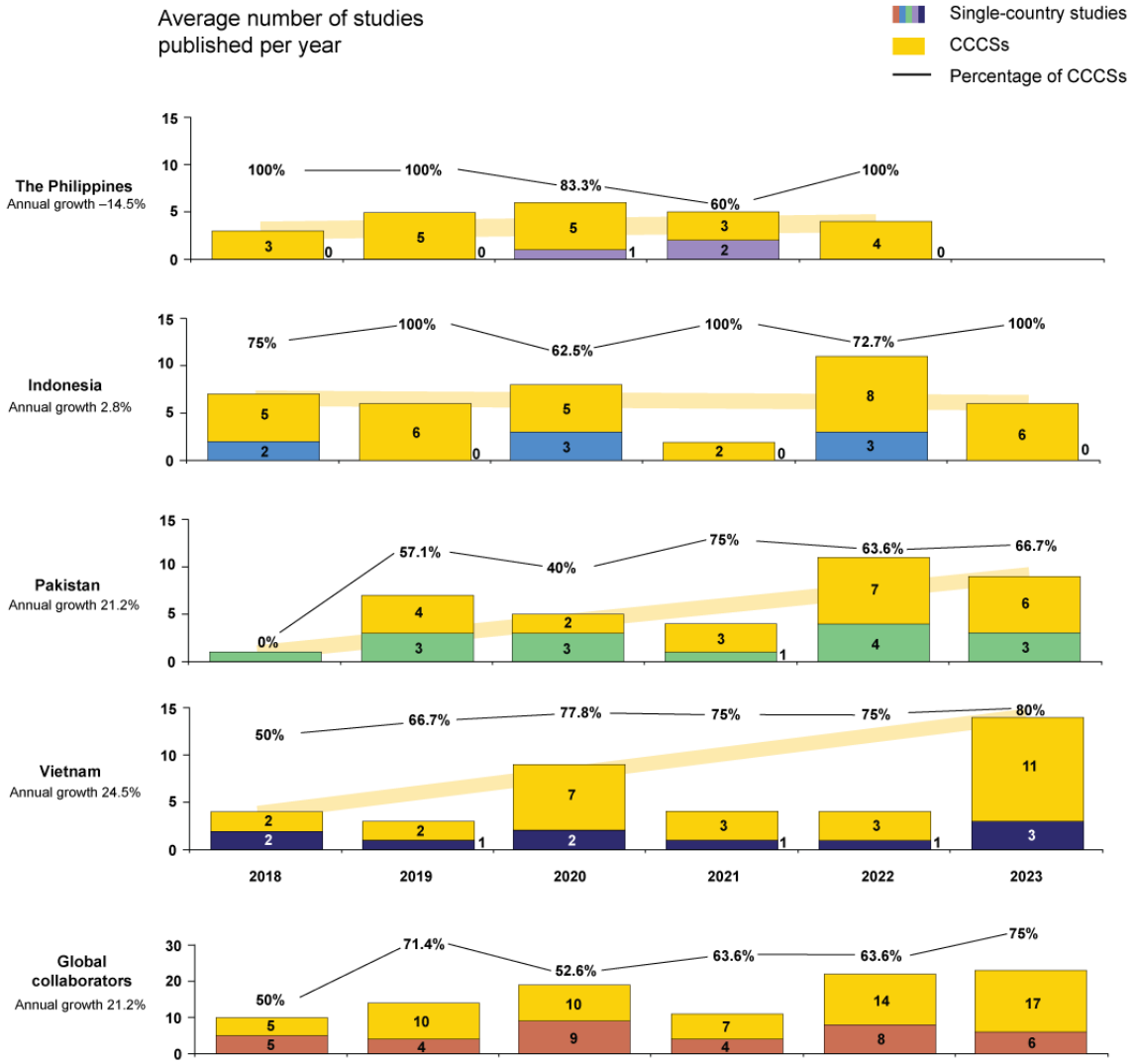

Overall Attributes

Overview

Tables 1 and 2 show the primary attributes of the eligible studies from each target nation by SCS and CCCS. The variables presented were study type, disease domain, data source, outcomes, participant demographics, length of study duration, time gap from last data collected to publication, sample size, and number of research centers.

XSL•FO
RenderX

Table 1. Study characteristics in the solo scholar cluster across the 7 target countries by single-country study (SCS) and cross-country collaboration study (CCCS; n=369)[a].

Study characteristics	Singapore (n=157)		Hong Kong (n=130)		Malaysia (n=100)	
	SCS (n=80)	CCCS (n=77)	SCS (n=86)	CCCS (n=44)	SCS (n=50)	CCCS (n=50)
Study type, n (%)						
CER[b]	54 (68)	44 (57)	57 (66)	31 (70)	28 (56)	27 (54)
Non-CER (descriptive)	26 (32)	33 (43)	29 (34)	13 (30)	22 (44)	23 (46)
Disease area, n (%)						
CVM[c]	33 (41)	32 (42)	36 (42)	20 (45)	13 (26)	25 (50)
IDV[d]	7 (9)	3 (4)	18 (21)	3 (7)	7 (14)	7 (14)
IAD[e]	2 (2)	3 (4)	6 (7)	1 (2)	6 (12)	2 (4)
Oncology	11 (14)	9 (12)	8 (9)	13 (30)	10 (20)	7 (14)
Others	27 (34)	30 (39)	18 (21)	7 (16)	14 (28)	9 (18)
Database type, n (%)						
EMR[f] or EHR[g]	43 (54)	19 (25)	69 (80)	12 (27)	11 (22)	12 (24)
Clinical registry	46 (58)	63 (82)	22 (26)	35 (80)	39 (78)	41 (82)
Health insurance and claims	4 (5)	1 (1)	0 (0)	1 (2)	0 (0)	0 (0)
Pharmacy claims	1 (1)	0 (0)	0 (0)	1 (2)	2 (4)	0 (0)
Multiple databases	12 (15)	5 (6)	5 (6)	3 (7)	2 (4)	3 (6)
Study outcome, n (%)						
Clinical	78 (98)	76 (99)	86 (100)	44 (100)	49 (98)	48 (96)
Cost	4 (5)	0 (0)	2 (2)	0 (0)	0 (0)	0 (0)
PROs[h]	0 (0)	5 (6)	1 (1)	2 (5)	1 (2)	7 (14)
Study population, n (%)						
Adults	63 (79)	61 (79)	67 (78)	37 (84)	32 (64)	37 (74)
Mixed	13 (16)	9 (12)	16 (19)	7 (16)	12 (24)	8 (16)
Pediatric	4 (5)	7 (9)	3 (3)	0 (0)	6 (12)	5 (10)
Study duration (y), mean (SD)	6.6 (4.8)	7.8 (7.4)	9.8 (7.3)	7.9 (6.7)	7.1 (4.2)	5.5 (7.4)
Lag period (y) from end of research to publication						
Overall, mean (SD)	5.8 (3.0)	4.8 (2.5)	5.1 (2.9)	5.0 (2.6)	4.8 (1.6)	4.6 (2.5)
<2, n (%)	3 (4)	7 (9)	11 (13)	4 (9)	1 (2)	7 (14)
2-5, n (%)	39 (49)	47 (61)	46 (53)	23 (52)	33 (66)	24 (48)
≥6, n (%)	34 (42)	16 (21)	26 (30)	10 (23)	16 (32)	9 (18)
Unknown, n (%)	4 (5)	7 (9)	3 (3)	7 (16)	0 (0)	10 (20)
Study size, mean (SD)						
Sample size (in thousands)	56.0 (180.9)	813.4 (4383.2)	205.0 (607.8)	4187.1 (23,979.1)	16.1 (30.7)	844.0 (4526.1)
Number of centers	4.4 (5.1)	89.7 (215.0)	42.7 (60.2)	66.8 (147.8)	19.5 (15.4)	92.4 (156.9)

[a]Study numbers for database types and study outcomes may appear as duplicates; hence, the total percentage may not add up to 100. CCCS numbers may appear as duplicates for studies conducted in multiple target countries. The percentages may add up to less or more than 100 because of rounding.

[b]CER: comparative effectiveness research.

[c]CVM: cardiology and metabolic disorders.

[d]IDV: infectious diseases and vaccines.

[e]IAD: inflammatory and autoimmune disorders.

[f]EMR: electronic medical record.

XSL•FO
RenderX

[g]EHR: electronic health record.

[h]PRO: patient-reported outcome.

XSL•FO
RenderX

Table 2. Study characteristics in the global collaborator cluster across the 7 target countries by single-country study (SCS) and cross-country collaboration study (CCCS; n=369)[a].

Study characteristics	Indonesia (n=33)		Pakistan (n=31)		Vietnam (n=27)		The Philippines (n=22)	
	SCS (n=7)	CCCS (n=26)	SCS (n=13)	CCCS (n=18)	SCS (n=7)	CCCS (n=20)	SCS (n=3)	CCCS (n=19)
Study type, n (%)								
CER[b]	3 (43)	14 (54)	1 (8)	12 (67)	5 (71)	11 (55)	1 (33)	10 (53)
Non-CER (descriptive)	4 (57)	12 (46)	12 (92)	6 (33)	2 (29)	9 (45)	2 (67)	9 (47)
Disease area, n (%)								
CVM[c]	1 (14)	13 (50)	2 (15)	5 (28)	2 (29)	9 (45)	1 (33)	8 (42)
IDV[d]	2 (29)	6 (23)	4 (31)	9 (50)	0 (0)	4 (20)	1 (33)	1 (5)
IAD[e]	2 (29)	2 (8)	0 (0)	1 (6)	3 (43)	1 (5)	0 (0)	2 (11)
Oncology	1 (14)	3 (12)	5 (38)	1 (6)	1 (14)	2 (10)	0 (0)	2 (11)
Others	1 (14)	2 (8)	2 (15)	2 (11)	1 (14)	4 (20)	1 (33)	6 (32)
Database type, n (%)								
EMR[f] or EHR[g]	2 (29)	12 (46)	9 (69)	9 (50)	4 (57)	10 (50)	1 (33)	7 (37)
Clinical registry	4 (57)	18 (69)	3 (23)	13 (72)	1 (14)	12 (60)	2 (67)	14 (74)
Health insurance and claims	1 (14)	0 (0)	0 (0)	0 (0)	2 (29)	0 (0)	0 (0)	0 (0)
Pharmacy claims	0 (0)	0 (0)	1 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Multiple databases	0 (0)	4 (15)	0 (0)	4 (22)	0 (0)	2 (10)	0 (0)	2 (11)
Study outcome, n (%)								
Clinical	7 (100)	26 (100)	13 (100)	16 (89)	6 (86)	20 (100)	2 (67)	19 (100)
Cost	2 (29)	0 (0)	0 (0)	0 (0)	2 (29)	0 (0)	1 (33)	0 (0)
PROs[h]	0 (0)	3 (12)	0 (0)	3 (17)	0 (0)	0 (0)	0 (0)	3 (16)
Study population, n (%)								
Adults	5 (71)	17 (65)	3 (23)	8 (44)	7 (100)	12 (60)	1 (33)	14 (74)
Mixed	2 (29)	6 (23)	9 (69)	8 (44)	0 (0)	5 (25)	2 (67)	4 (21)
Pediatric	0 (0)	3 (12)	1 (8)	2 (11)	0 (0)	3 (15)	0 (0)	1 (5)
Study duration (y), mean (SD)	3.1 (3.3)	7.1 (9.6)	2.1 (1.9)	9.1 (11.2)	4.8 (7.0)	8.2 (10.1)	12.0 (8.2)	8.4 (9.8)
Lag period (y) from end of research to publication								
Overall, mean (SD)	4.3 (1.8)	4.2 (2.4)	3.8 (1.6)	3.3 (1.8)	4.6 (1.3)	4.7 (2.6)	4.0 (1.7)	3.9 (2.3)
<2, n (%)	1 (14)	5 (19)	2 (15)	6 (33)	0 (0)	3 (15)	0 (0)	4 (21)
2-5, n (%)	4 (57)	16 (62)	8 (62)	6 (33)	6 (86)	11 (55)	2 (67)	13 (68)
≥6, n (%)	2 (29)	2 (8)	2 (15)	1 (6)	1 (14)	5 (25)	1 (33)	1 (5)
Unknown, n (%)	0 (0)	3 (12)	1 (8)	5 (28)	0 (0)	1 (5)	0 (0)	1 (5)
Study size, mean (SD)								
Sample size (in thousands)	122.6 (304.2)	1513.4 (6235.5)	5.9 (9.8)	2226.8 (7525.4)	202.7 (525.9)	1982.8 (7132.4)	161.9 (280.2)	2094.2 (7322.2)
Number of centers	5.0 (3.6)	88.9 (92.7)	19.9 (13.4)	35.7 (48.3)	11.7 (4.9)	181.9 (236.9)	3.0 (—[i])	81.6 (96.7)

[a]Study numbers for database types and study outcomes may appear as duplicates; hence, the total percentage may not add up to 100. CCCS numbers may appear as duplicates for studies conducted in multiple target countries. The percentages may add up to less or more than 100 because of rounding.

[b]CER: comparative effectiveness research.

[c]CVM: cardiology and metabolic disorders.

XSL•FO
RenderX

[d]IDV: infectious diseases and vaccines.

[e]IAD: inflammatory and autoimmune disorders.

[f]EMR: electronic medical record.

[g]EHR: electronic health record.

[h]PRO: patient-reported outcome.

[i]Not applicable.

Study Type: CER or Non-CER (Descriptive)

Of the 369 studies, 221 (59.9%) were CER studies, with the remaining 148 (40.1%) being non-CER or descriptive. The relative representation of CER versus non-CER for SCSs and CCCSs is illustrated in Multimedia Appendix 2. Singapore, Hong Kong, Malaysia, and Vietnam had a higher number of CER studies in both their SCSs and CCCSs. Vietnam's SCSs had the predominant CER representation at 71% (5/7), followed by Singapore at 68% (54/80) and Hong Kong at 66% (57/86). Among CCCSs, Hong Kong led with 70% (31/44) CER studies, followed by Pakistan with 67% (12/18) and Singapore with 57% (44/77). There were more descriptive non-CER studies in SCSs from Pakistan, the Philippines, and Indonesia, resulting in the CER study percentages being 8% (1/13), 33% (1/3), and 43% (3/7), respectively.

Figure 5 shows the yearly trends of CER percentages from 2018 to 2023 broken down by SCS and CCCS. The consistency in trends was more noticeable in SCSs across the 7 target nations compared to CCCSs. An upward trend in CER study percentage was observed in SCSs from Hong Kong and the global collaborators. Conversely, Malaysia's SCSs experienced a steady decrease in CER contribution over the same period.

Figure 5. Trends in comparative effectiveness research (CER) by single-country studies (SCSs) and cross-country collaboration studies (CCCSs). Trends in CCCSs across the 7 target countries showed more consistency compared to trends in SCSs. The SCSs of Hong Kong and the global collaborators showed an upward trend in CER percentages, whereas Malaysia's SCSs consistently decreased in CER contribution. The count of CCCSs represents the individual contributions of each country, leading to a total count across countries that exceeds the actual number of CCCSs due to some studies involving multiple collaborators.

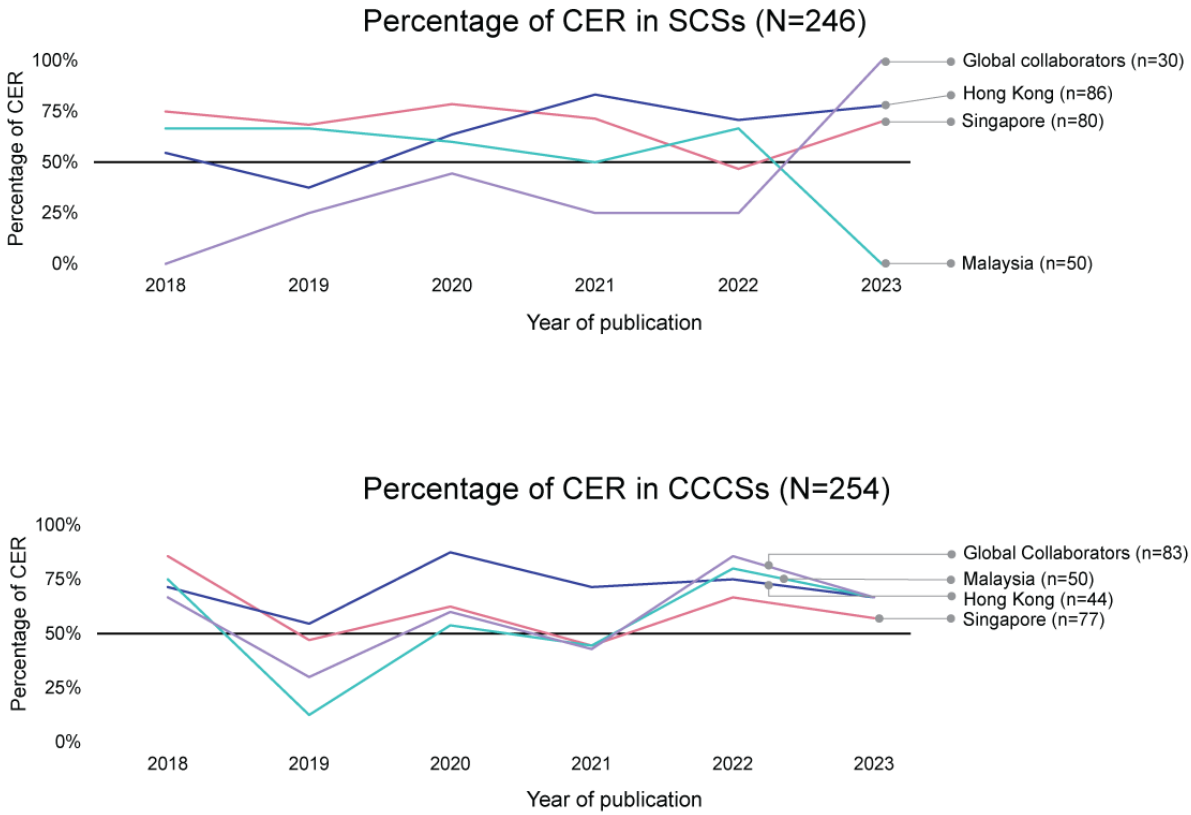

The 2-year SMA trends for CER and descriptive studies are illustrated in Multimedia Appendix 2 for the biennial average from 2018 to 2023. Hong Kong consistently increased its CER contributions in both SCSs and CCCSs, increasing from 47% (9/19) between 2018 and 2019 to 73% (24/33) between 2022 and 2023 for SCSs and similarly from 61% (11/18) between 2018 and 2019 to 73% (8/11) between 2022 and 2023 for CCCSs. Other notable rises in CER contributions in CCCSs were observed in Malaysia (from 4/12, 33% between 2018 and 2019 to 12/16, 75% between 2022 to 2023), Indonesia (from 3/9, 33% between 2018 to 2019 to 7/10, 70% between 2022 and 2023), Pakistan (from 1/4, 25% between 2018 and 2019 to 8/9, 89% between 2022 and 2023), and Vietnam (from 4/9, 44% between 2019 and 2020 to 6/7, 86% between 2022 and 2023).

XSL•FO
RenderX

Conversely, Malaysia's SCSs saw a consistent decline in CER contribution over the 5 years, dropping from 67% (8/12) between 2018 and 2019 to 53% (10/19) between 2022 and 2023. Furthermore, all of Pakistan's SCSs (12/12, 100%) were non-CER between 2018 and 2022.

Database Type

Of the 369 studies, 341 (92.4%) used a single database. Exclusive use of clinical registry databases was most common at 50.9% (188/369), followed by electronic medical records (EMRs) or electronic health records (EHRs) at 39.3% (145/369), health insurance and administrative claims at 1.4% (5/369), and pharmacy claims at 0.8% (3/369). The use of multiple databases was found in 7.6% (28/369) of the studies, primarily combining clinical registries and EMRs or EHRs (Multimedia Appendix 2). Use of EMR or EHR databases was more common for SCSs (120/246, 48.8%; Multimedia Appendix 2). On the other hand, the predominant exclusive database warehouse for CCCSs was clinical registries, used in 73.2% (9/123) of the studies. EMRs' or EHRs' contribution to CCCSs was lower, representing only 20.3% (25/123) of CCCSs, which is considerably lower than their share in SCSs (Tables 1 and 2 and Multimedia Appendix 2).

The use of the clinical registry database type consistently dominated across all CCCSs from all target nations, whether used on its own or in combination with other databases. For SCSs, (1) there were more clinical registries over EMRs or EHRs used in Indonesia, Malaysia, and the Philippines—the figures were 57% (4/7) versus 29% (2/7), 78% (39/50) versus 22% (11/50), and 67% (2/3) versus 33% (1/3), respectively; (2) Singapore's use was almost even, with 58% (46/80) of the studies using clinical registries and 54% (43/80) using EMRs or EHRs; and (3) conversely, Hong Kong, Pakistan, and Vietnam used more EMRs or EHRs than clinical registries—80% (69/86) versus 26% (22/86), 69% (9/13) versus 23% (3/13), and 57% (4/7) versus 14% (1/7), respectively.

Figure 6 reveals the evolution of EMR or EHR contributions, both exclusively and in combination with other databases, in the previous 5 years. In SCSs, Malaysia and the global collaborators experienced a consistent decline in EMR or EHR use, whereas Hong Kong exhibited an increase. Malaysia's EMR or EHR use in SCSs remained consistently at <50% during this period. In contrast to SCSs, where EMR or EHR use was predominant, EMR or EHR use in CCCSs from the target countries was always at <50% from 2018 to 2023, and no consistent time trend pattern was observed.

Figure 6. Trends in percentage of use of medical records by single-country studies (SCSs) and cross-country collaboration studies (CCCSs). A decline in exclusive or combined use of electronic medical records (EMRs) or electronic health records (EHRs) was observed for SCSs in Malaysia and the global collaborators, whereas Hong Kong saw an increase; in contrast, use of EMR or EHR databases in CCCSs was consistently below 50%, with no uniform trend emerging during this period. The count of CCCSs represents the individual contributions of each country, leading to a total count across countries that exceeds the actual number of CCCSs due to some studies involving multiple collaborators.

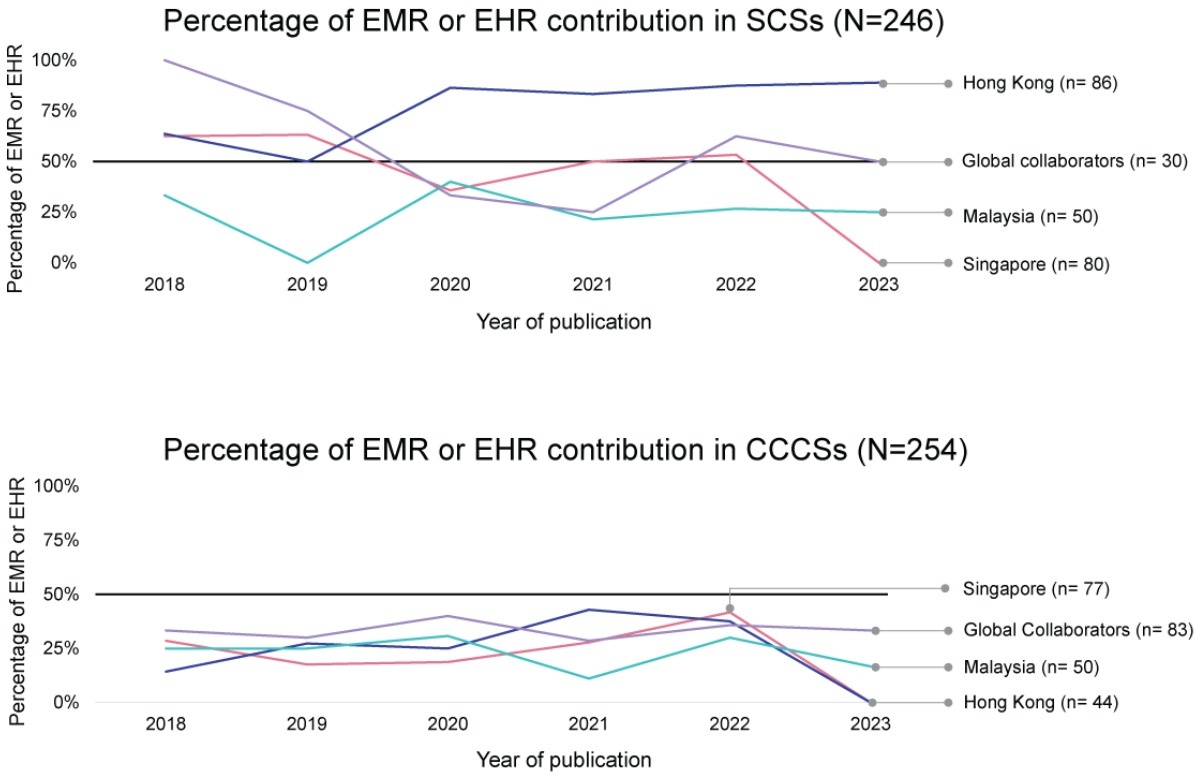

The 2-year SMA over the previous 5 years indicated that the exclusive use of EMRs or EHRs in SCSs from the 7 target nations increased from 46% (26/56) to 60% (49/82). In contrast, the reliance on clinical registry databases dipped from 50% (28/56) to 38% (31/82); Multimedia Appendix 2). For CCCSs, the distribution between clinical registries and EMRs or EHRs remained relatively steady, with clinical registries being the most common (Multimedia Appendix 2).

XSL•FO
RenderX

Disease Area

The leading medical research area was CVM, accounting for 36.9% (136/369) of the studies, trailed by oncology and IDV, each with 14.9% (55/369). Inflammatory and autoimmune disorders was the least prevalent area, representing 6.2% (23/369) of the studies. The remaining 27.1% (100/369) of the studies pertained to various other diseases. The proportion of CVM studies grew from 28% (11/39) in 2018 to 39% (14/36) in 2023, peaking in 2020 with 49% (36/73). Conversely, the share contributed by IDV medical area increased from 8% (6/73) in 2020 to 25% (21/84) in 2022 and 17% (6/36) in 2023, surpassing oncology as the second most common disease and therapeutics research area in recent years (Multimedia Appendix 2).

Study Outcomes

Most of the studies (348/369, 94.3%) presented clinical outcomes whether in terms of treatment effect or safety. There were 5.7% (21/369) of the studies that discussed cost outcomes, PROs, or a combination of these with clinical results. In the SCS category, every study from Pakistan focused on clinical outcomes, whereas cost outcomes were observed in SCSs from all countries except Malaysia and Pakistan. One study each from Hong Kong (1/86, 1%) and Malaysia (1/50, 2%) included PRO outcomes. In the CCCS category, none of the selected nations published studies focusing on cost outcomes (Tables 1 and 2).

Study Population

Of the 369 studies obtained, 273 (74%) investigated adults, 24 (6.5%) focused on the pediatric age group, and 72 (19.5%) encompassed both adult and pediatric participants. Notably, in the SCS category (Table 2), Pakistan (9/13, 69%) and the Philippines (2/3, 67%) reported higher proportions of mixed populations than of solely adult participants (Pakistan: 3/13, 23%; the Philippines: 1/3, 33%).

Pediatric representation in the CCCSs was 8.1% (10/123), slightly higher than in SCSs (14/246, 5.7%). Within CCCSs (Table 2), Vietnam led in pediatric-focused research with 15% (3/20) of the studies, followed by Indonesia (3/26, 12%) and Pakistan (2/18, 11%).

Study Duration

Information about study duration was reported for 94.6% (349/369) of the studies. The average duration was 7.4 (SD 6.3) years, ranging from 0.01 to 35.8 years. In total, 3.5% (13/369) of the studies had a duration of >20 years. The mean for SCSs was higher at 7.5 (SD 6.0) years compared to that for CCCSs at 7.1 (SD 6.8) years (Multimedia Appendix 2).

Among SCSs, the Philippines (Table 2) topped the list with the longest average study duration of 12.0 (SD 8.2) years. As shown in Table 1, Hong Kong followed closely with an average of 9.8 (SD 7.3) years. Conversely, Pakistan and Indonesia registered the shortest mean study durations with 2.1 (SD 1.9) years and 3.1 (SD 3.3) years, respectively. Over a 5-year span, based on a 2-year rolling average, the study duration in Indonesia showed an uptick, increasing from 0.6 years between 2018 and 2019 to 1.9 years between 2022 and 2023. Other target countries did not exhibit any consistent study duration trend patterns (Multimedia Appendix 2).

For CCCSs, Pakistan (Table 2) led with the longest mean study duration of 9.1 (SD 11.2) years, closely followed by the Philippines with 8.4 (SD 9.8) years. Malaysia (Table 1) recorded the shortest average study duration at 5.5 (SD 7.4) years. The study duration in Indonesia's CCCSs averaged 7.1 (SD 9.6) years, which was notably longer than that of its SCSs. Observing trends, there was a decline in the mean study duration of CCCSs in Malaysia, Pakistan, the Philippines, and Vietnam. Conversely, the average study duration in Singapore's CCCSs steadily rose, increasing from 5.6 years between 2018 and 2019 to 9.1 years between 2022 and 2023 (Multimedia Appendix 2).

Lag Between the Research Period and Publication

Of the eligible studies, most (205/369, 55.6%) were published between 2 and 5 years after the time of latest available data studied, 28.5% (105/369) were published after >6 years, and 9.5% (35/369) were published within 2 years. The remaining 6.5% (24/369) of the studies had unspecified year or years of research completion (Multimedia Appendix 2).

Eligible studies from all the target nations showed a similar trend, with most (205/345, 59.4%) being published within 2 to 5 years after the research period. However, in both Singapore and Hong Kong, the time taken from research completion to publication was notably longer for both SCSs and CCCSs, averaging 5.8 (SD 3.0) years and 5.1 (SD 2.9) years for SCSs and 4.8 (SD 2.5) years and 5.0 (SD 2.6) years for CCCSs, respectively (Table 1).

It is worth noting that 15.9% (17/107) of CCCSs were published within 2 years of the research period but only 7.6% (18/238) of SCSs were published within this time frame. We observed upward trends in the time to publication within 2 years from 2018 to 2023. For SCSs, it was from 3% (2/63) to 15% (12/83), and for CCCSs, it was from 16% (6/38) to 20% (5/25; Multimedia Appendix 2).

The publication time lag also varied according to the RWD source (Multimedia Appendix 2). Among studies that relied on a single database, the highest percentage of those published within 2 years after the research consistently used the EMR or EHR database type. This trend held true for both SCSs and CCCSs, with EMRs or EHRs dominating the quick turnaround for publications. Specifically, among SCSs, 12.8% (15/117) of studies using the EMR or EHR database type were published within this 2-year time frame, whereas only 2% (2/94) of those using clinical registry databases had the same publishing speed. CCCSs had a consistent pattern—26% (6/23) of the studies that used EMRs or EHRs were published within 2 years, in contrast to the 14% (11/78) of the studies that used clinical registry databases. Notably, no studies published within the 2-year window used the health insurance and medical claims database type or the pharmacy claims database type.

Study Size: Sample Size and Number of Centers

The sample size was specified in 98.1% (362/369) of the studies and varied considerably, ranging from as few as 16 to >154,500,000. The average sample size was 672,352 (SD

XSL•FO
RenderX

8,364,280). The average in CCCSs was much higher at 1,824,035 (SD 14,530,091) compared to 106,029 (SD 397,050) in SCSs (Multimedia Appendix 2). The 2-year SMA of sample size in SCSs indicated an increasing trend from 51,706 to 178,675 from 2018 to 2023. In contrast, a sharp decline was observed in CCCS sample sizes from 5,543,271 to 548,107 during the same period (Multimedia Appendix 2). Hong Kong had the highest average sample size in SCSs (205,006, SD 607,767) as well as in CCCSs (4,187,122, SD 23,979,119; Tables 1 and 2).

The number of participating centers was only specified in 44.4% (164/369) of the studies. The mean number of centers was 44.3 (SD 120.3), ranging from 2 to 1119. As noted in Tables 1 and 2, Hong Kong reported the highest average number of study centers at 42.7 (SD 60.2) for SCSs, followed by Pakistan at 19.9 (SD 13.4) and Malaysia at 19.5 (SD 15.4). For CCCSs, Vietnam had the highest mean number of study centers at 181.9 (SD 236.9), which was higher than twice the overall CCCS mean of 78.4 (SD 178.6).

The 2-year SMA for the number of study centers in SCSs initially rose from 81.8 between 2018 and 2019, reaching a peak of 131.9 between 2019 and 2020 before declining to 99.1; 37.1; and, finally, 30.8 in the subsequent years. In CCCSs, there was a downtrend starting from 16.1 between 2018 and 2019 to 10.7 centers at its lowest point between 2020 and 2021 before bouncing back to 33.9 centers between 2022 and 2023 (Multimedia Appendix 2).

Table S5 in Multimedia Appendix 1 provides and overview of the number of centers involved in RWD studies across various database types in the target countries. The Philippines reported the highest average number of centers (197.0) in studies using EMR or EHR databases, whereas Vietnam had the highest average (65.8) in studies associated with clinical registry databases. Information on the number of centers was not reported in many studies, particularly those using health insurance and claims databases.

Database Names

Multimedia Appendix 3 provides the specific name or names of the databases used in each study organized by target country, database type, and disease area.

Discussion

Principal Findings

This scoping review was based on our earlier research completed for Taiwan, India, and Thailand in Asia [7]. We have now expanded insights on integrated real-world study databases across 7 additional diverse Asian health care systems in Hong Kong, Indonesia, Malaysia, Pakistan, the Philippines, Singapore, and Vietnam. This has enabled us to provide a comprehensive perspective on the current landscape of RWE generation in these nations, thus aiding stakeholders in formulating informed research and policy decisions [22,23].

The archetyping of the target nations into 2 country clusters (ie, solo scholars and global collaborators) allowed us to uncover distinct patterns reflective of differing resources, priorities, and strategic objectives. Solo scholars tend to conduct independent studies, which exemplifies that these nations are equipped with robust research infrastructures. This autonomy allows for a deep dive into national health issues tailored to specific local contexts and country priorities [24]. On the other hand, global collaborators frequently engaged in international partnerships, a strategy that would be likely born out of necessity due to limited research funding and resources within the nation and, hence, a greater reliance on collaborative networks [25]. We found that >90% of CCCSs from global collaborators involving nontarget countries (76/83, 92%) also partnered with some of the 7 target nations. This collaborative pattern possibly indicated a stronger network of research collaboration within the neighboring regions of Asia. In contrast, among solo scholars such as Singapore and Hong Kong, despite the high number of CCCSs involving nontarget countries (>94%; 73/77, 95% in Singapore and 43/44, 98% in Hong Kong), there were approximately 40% of studies (33/77, 43%; 17/44, 39%) not partnered with other target nations. However, the emphasis on global or regional priorities could overshadow the unique health challenges and priorities of these nations. This imbalance can lead to a scarcity of data and insights directly applicable to domestic health care [26].

Among solo scholars, Hong Kong, Malaysia, and Singapore emerged as significant contributors to domestic RWD publications, showcasing their robust research infrastructure and commitment to harnessing RWD. In contrast, global collaborators, and especially the Philippines, had fewer studies, which could be attributed to various reasons, from funding sources to bureaucracies in research grant administration [27,28]. Singapore emerged as the predominant contributor to the CCCS pool, but on average, Malaysia collaborated with more countries. It is also particularly noteworthy that, from 2018 to 2023, most of our target nations (except the Philippines) manifested an increasing trend in the average number of studies published annually, with Vietnam leading in growth. Vietnam's health care system has been consistently advancing, and the nation is enhancing its research capabilities; this progress has been widely acknowledged in the literature [23,29].

Diving deeper into the nature of these studies, there was an evident leaning toward CER over descriptive studies in several nations, such as Vietnam, Singapore, and Hong Kong. The proportion of CER to non-CER studies offers insights into the nature of the questions that researchers in different regions were keen to address. Regions with a higher proportion of CER studies, such as Vietnam's SCSs, suggest an active interest in comparing the outcomes of different interventions and preventive and prognostic strategies, which can be crucial for policy decisions, including effective containment of COVID-19. Vietnam, along with Singapore and Hong Kong, has been extensively praised for effectively controlling the spread of COVID-19, especially during the early stages of the pandemic [30,31]. While we cannot assert a direct link with certainty, the possibility of a connection through RWD from CER also playing a role in fostering better-informed public health decisions cannot be denied [32]. Moreover, Singapore and Hong Kong's well-established reputation as quality research hubs in Asia

XSL•FO
RenderX

could further underscore the potential impact of robust research frameworks [33,34].

The predilection for certain database types, be it clinical registries or EMRs or EHRs, was due to a combination of availability and convenience and, hence, the ease of use of these databases in different regions. The 5-year trend showcased the evolving dynamics in RWD research. Nations leaning toward EMR or EHR databases, such as Hong Kong, might be signaling greater digitization of their health records or the perceived value in this data type [35,36], whereas nations such as Indonesia, Malaysia, and the Philippines primarily leveraged clinical registries. Apart from Hong Kong, Pakistan and Vietnam also displayed a marked inclination toward EMR or EHR use.

The variance in research duration offers a window into research efficiency and the possible administrative or infrastructural bottlenecks. Longer durations in nations such as the Philippines could indicate complex, long-term studies or challenges in study execution and continuity [27]. In addition, the expectation that RWD accelerates evidence generation is not reflected in our findings, where there was an average lag of approximately 2 to 5 years from research completion to publication. Notably, around 90% of the studies (310/345, 89.9%) extended beyond 2 years to reach publication, suggesting room for enhancing the efficiency of evidence generation, potentially through targeted support mechanisms. Interestingly, a comparatively higher proportion of CCCSs in contrast to SCSs (17/107, 15.9% vs 18/238, 7.6%) were published within 2 years, hinting at the efficiency benefits that international partnerships might offer in expediting research outputs. Larger study sample sizes and a greater number of centers, as observed in Hong Kong, reflect the ability to conduct expansive studies from territory-wide linked databases, indicating a propensity for large-scale, nationally representative research [37]. Similarly, the number of centers involved could also indicate the collaborative spirit within the research community or the need to pool resources.

The strategic application of RWE in health care research and policy formation is clearly evident on a global scale. Although there is an increasing reliance on HTA as a pivotal tool for informed health care decision-making, the nations categorized largely under the global collaborators cluster face a tremendous challenge in health care infrastructure and economic constraints to lead independent RWE that could shape local health care policy and reimbursement decision-making [38,39]. Hence, open data initiatives and international collaborations such as the Observational Health Data Sciences and Informatics and the HTAsiaLink, respectively, are crucial in this regard [40,41]. There is also the potential for other international networks (eg, the International Network of Agencies for Health Technology Assessment) in facilitating the alignment of health care policies with benchmark evidence-based practices [42].

Limitations

In our efforts to comprehensively assess the landscape of research in the field, we encountered challenges in data extraction, mainly due to inconsistencies in how various studies reported certain characteristics such as the number of centers or study periods. For instance, while some studies provided clear details on their duration, others only specified the enrollment phase, leaving us to speculate on the follow-up or observation period. In addition, the screening and extraction process involved multiple reviewers working under tight schedules. We acknowledge that this approach diverges from the ideal practice of having at least 2 independent investigators screen each title, abstract, and full text and subsequently extract data blinded to each other's decisions. While we implemented quality checks, including spot assessments and team discussions, the constraints may have inevitably introduced occasional inaccuracies.

Moreover, in some nations, the limited number of studies could make percentage-based analyses less reflective of the true study landscape. Still, these analyses offer a preliminary understanding of research trends in those regions. We must also acknowledge that our analysis, with a search conducted in May 2023, assumes a linear distribution for studies in 2019 and 2023, which might slightly deviate from the actual figures due to approximations.

We might have inadvertently overlooked some relevant studies, especially if abstracts failed to mention key terms such as RWD or RWE. Our decision to focus on English-language publications and rely solely on PubMed for citations, while strategic, may have missed a handful of non–English-language studies or those in other databases. Nonetheless, the vast number of studies that we analyzed offers valuable insights into using RWD to produce RWE in our target countries. In addition, we did not report study design and funding sources for the included studies. While this information may have been valuable, we faced challenges in the extraction of study design and funding information due to lack of clear and consistent reporting across publications. This necessitated the exclusion of these variables to prevent any subjective interpretations.

Despite these challenges, our findings underscore a need in the research community—a call for clearer, more standardized reporting on the databases used, study design, analysis methods, and important time points. A particular area that warrants attention is the clarity in detailing study duration (which should encompass the recruitment and observation periods), study design, and funding sources. While we recognize this study's limitations, we believe that it also paves the way for refining research methodologies in the future.

Conclusions: The Path Forward

Our comprehensive assessment of studies across the selected nations reveals intricate patterns that explain the diverse research landscape for RWD generation. Each nation's unique landscape for contemporary integrated RWD warehouses tells a narrative that is partially attributed to their economic, clinical, and research settings. Delving deeper into these patterns aids in formulating robust insights for future endeavors in health care research and policy making, including prioritization of competency building based on a nation's unique infrastructure, skill sets, and research strengths and weaknesses [4]. As the health care landscape evolves, there is an undeniable value in understanding and leveraging RWD [43]. Recognizing the diverse approaches and challenges across countries can lead to more collaborative and informed strategies [4]. The goal should be to address the present gaps and pave the way for future synergistic, impactful, and patient-centric research [3].

XSL•FO
RenderX

In conclusion, the observed variations across nations reiterate the essence of context in health care research. Every nation's unique story, as told by their data, accentuates the need for a tailored approach in using RWD—ensuring that they truly serve the multifaceted needs of health care research and decision-making.

Acknowledgments

This study was funded by Pfizer, and the research was conducted by Transform Medical Communications Limited (Transform Medcomms), New Zealand. The authors would like to thank Dr Veena Shetye and Dr Daniel Furtner for their assistance in screening and data extraction provided on behalf of Transform Medcomms. The views and opinions expressed herein are solely those of the authors and do not reflect the views or positions of their employers.

Data Availability

Data generated or analyzed during this study are included in Multimedia Appendices 1-3.

Authors' Contributions

All the authors were involved in the study conception, study design, and interpretation of the facts and data. WYS and VJ were involved in data analysis. SS led the manuscript writing, and all authors were engaged in revising it for scientific content and final approval before submission for publication.

Conflicts of Interest

WYS and HS declare that, while they are employees of Pfizer, there is no conflict of interest in relation to the work presented in this paper.

Multimedia Appendix 1
Supplementary tables.
[PDF File (Adobe PDF File), 200 KB - jmir_v26i1e56686_app1.pdf]

Multimedia Appendix 2
Supplementary figures.
[PDF File (Adobe PDF File), 298 KB - jmir_v26i1e56686_app2.pdf]

Multimedia Appendix 3
Identified databases.
[PDF File (Adobe PDF File), 508 KB - jmir_v26i1e56686_app3.pdf]

References

1. Takura T, Miura H. Socioeconomic determinants of universal health coverage in the Asian region. Int J Environ Res Public Health 2022 Feb 18;19(4):2376 [FREE Full text] [doi: 10.3390/ijerph19042376] [Medline: 35206562]
2. Ranabhat CL, Acharya SP, Adhikari C, Kim CB. Universal health coverage evolution, ongoing trend, and future challenge: a conceptual and historical policy review. Front Public Health 2023 Feb 3;11:1041459 [FREE Full text] [doi: 10.3389/fpubh.2023.1041459] [Medline: 36815156]
3. Shau WY, Setia S, Shinde S, Santoso H, Furtner D. Generating fit-for-purpose real-world evidence in Asia: how far are we from closing the gaps? Perspect Clin Res 2023;14(3):108-113 [FREE Full text] [doi: 10.4103/picr.picr_193_22] [Medline: 37554247]
4. Kc S, Lin LW, Bayani DB, Zemlyanska Y, Adler A, Ahn J, et al. What, where, and how to collect real-world data and generate real-world evidence to support drug reimbursement decision-making in Asia: a reflection into the past and a way forward. Int J Health Policy Manag 2023;12:6858 [FREE Full text] [doi: 10.34172/ijhpm.2023.6858] [Medline: 37579427]
5. Hiramatsu K, Barrett A, Miyata Y, PhRMA Japan Medical Affairs Committee Working Group 1. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs Real World Outcomes 2021 Dec 19;8(4):459-480 [FREE Full text] [doi: 10.1007/s40801-021-00266-3] [Medline: 34148219]
6. Graili P, Guertin JR, Chan KK, Tadrous M. Integration of real-world evidence from different data sources in health technology assessment. J Pharm Pharm Sci 2023 Jul 17;26:11460 [FREE Full text] [doi: 10.3389/jpps.2023.11460] [Medline: 37529633]
7. Shau WY, Setia S, Chen YJ, Ho TY, Prakash Shinde S, Santoso H, et al. Integrated real-world study databases in 3 diverse Asian health care systems in Taiwan, India, and Thailand: scoping review. J Med Internet Res 2023 Sep 11;25:e49593 [FREE Full text] [doi: 10.2196/49593] [Medline: 37615085]

XSL•FO
RenderX

8. Shau WY, Setia S, Shinde SP, Santoso H, Furtner D. Contemporary databases in real-world studies regarding the diverse health care systems of India, Thailand, and Taiwan: protocol for a scoping review. JMIR Res Protoc 2022 Dec 13;11(12):e43741 [FREE Full text] [doi: 10.2196/43741] [Medline: 36512386]
9. Kumamaru H, Togo K, Kimura T, Koide D, Iihara N, Tokumasu H, et al. Inventory of real-world data sources in Japan: annual survey conducted by the Japanese Society for Pharmacoepidemiology Task Force. Pharmacoepidemiol Drug Saf 2024 Jan;33(1):e5680. [doi: 10.1002/pds.5680] [Medline: 37650434]
10. Yoshida S, Takeuchi M, Tanaka-Mizuno S, Mizuno K, Nakashima M, Fukasawa T, et al. Clinical epidemiology and pharmacoepidemiology studies with real-world databases. Proc Jpn Acad Ser B Phys Biol Sci 2022;98(10):517-528 [FREE Full text] [doi: 10.2183/pjab.98.026] [Medline: 36504194]
11. Mazumder D, Bhandari H, Ranjan S, LoPresti M, Smulders M. Review of real-world databases in Japan and South Korea that could be used to inform healthcare resource use and cost inputs for health economic studies. Value Health 2018 Sep;21:S52-S53. [doi: 10.1016/j.jval.2018.07.398]
12. Ahn EK. A brief introduction to research based on real-world evidence: considering the Korean National Health Insurance Service database. Integr Med Res 2022 Jun;11(2):100797 [FREE Full text] [doi: 10.1016/j.imr.2021.100797] [Medline: 34754752]
13. Kim S, Kim MS, You SH, Jung SY. Conducting and reporting a clinical research using Korean healthcare claims database. Korean J Fam Med 2020 May;41(3):146-152 [FREE Full text] [doi: 10.4082/kjfm.20.0062] [Medline: 32456382]
14. Kim JW, Kim C, Kim KH, Lee Y, Yu DH, Yun J, et al. Scalable infrastructure supporting reproducible nationwide healthcare data analysis toward FAIR stewardship. Sci Data 2023 Oct 04;10(1):674 [FREE Full text] [doi: 10.1038/s41597-023-02580-7] [Medline: 37794003]
15. Xie J, Wu EQ, Wang S, Cheng T, Zhou Z, Zhong J, et al. Real-world data for healthcare research in China: call for actions. Value Health Reg Issues 2022;27:72-81 [FREE Full text] [doi: 10.1016/j.vhri.2021.05.002] [Medline: 34844062]
16. Li P, Wang S, Chen Y. Use of real-world evidence for drug regulatory decisions in China: current status and future directions. Ther Innov Regul Sci 2023 Nov;57(6):1167-1179 [FREE Full text] [doi: 10.1007/s43441-023-00555-9] [Medline: 37624556]
17. Zhong J, Zhang J, Fang H, Liu L, Xie J, Wu E. Advancing the development of real-world data for healthcare research in China: challenges and opportunities. BMJ Open 2022 Jul 29;12(7):e063139 [FREE Full text] [doi: 10.1136/bmjopen-2022-063139] [Medline: 35906059]
18. Sharma M, Teerawattananon Y, Dabak SV, Isaranuwatchai W, Pearce F, Pilasant S, et al. A landscape analysis of health technology assessment capacity in the Association of South-East Asian Nations region. Health Res Policy Syst 2021 Feb 11;19(1):19 [FREE Full text] [doi: 10.1186/s12961-020-00647-0] [Medline: 33573676]
19. Ghazinoory S, Majidi B, Nasri S, Zandi ME, Farrokhi H, Javedani M, et al. Differences between health technology assessment topics in high- and middle-income countries: a scoping review. Arch Public Health 2021 Dec 14;79(1):225 [FREE Full text] [doi: 10.1186/s13690-021-00754-6] [Medline: 34906239]
20. Lou J, Kc S, Toh KY, Dabak S, Adler A, Ahn J, et al. Real-world data for health technology assessment for reimbursement decisions in Asia: current landscape and a way forward. Int J Technol Assess Health Care 2020 Oct;36(5):474-480. [doi: 10.1017/S0266462320000628] [Medline: 32928330]
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest 2020 Feb 03;130(2):565-574 [FREE Full text] [doi: 10.1172/JCI129197] [Medline: 32011317]
23. Minh HV, Giang LM, Cashin C, Hinh ND. Health system research in Vietnam: generating policy-relevant knowledge. Glob Public Health 2015;10 Supppl 1(0 1):S1-S4 [FREE Full text] [doi: 10.1080/17441692.2014.986179] [Medline: 25622126]
24. Acharya KP, Pathak S. Applied research in low-income countries: why and how? Front Res Metr Anal 2019;4:3 [FREE Full text] [doi: 10.3389/frma.2019.00003] [Medline: 33870035]
25. Olopade CO, Olugbile S, Olopade OI. Chapter 47 - Issues and challenges for clinical research in international settings. In: Gallin JI, Ognibene FP, editors. Principles and Practice of Clinical Research Third Edition. New York, NY: Academic Press; 2012.
26. Alemayehu C, Mitchell G, Nikles J. Barriers for conducting clinical trials in developing countries- a systematic review. Int J Equity Health 2018 Mar 22;17(1):37 [FREE Full text] [doi: 10.1186/s12939-018-0748-6] [Medline: 29566721]
27. Sales RK, Reyes-Ramos GK, de Veyra C, Gascon G, Barraca V, Garcia G, et al. Advancing health through evidence assisted decisions with health policy and systems research program: a qualitative evaluation of a national health research grant management process in the Philippines. Health Res Policy Syst 2023 Jul 13;21(1):73 [FREE Full text] [doi: 10.1186/s12961-023-01021-6] [Medline: 37443116]
28. Feliciano KT, Lindo UC, Tanedo AN, Gaspar DC, Cortez JB. Cancer research funding and care in the Philippines: gaps and ways forward. Lancet Oncol 2023 Sep;24(9):e360. [doi: 10.1016/S1470-2045(23)00356-X] [Medline: 37657475]

XSL•FO
RenderX

29. Kagan J, Giang DD, Iademarco MF, Phung VT, Lau CY, Quang NN. Assessing clinical research capacity in Vietnam: a framework for strengthening capability for clinical trials in developing countries. Public Health Rep 2016;131(3):396-403 [FREE Full text] [doi: 10.1177/003335491613100305] [Medline: 27252559]
30. Tabish SA. COVID-19 pandemic: emerging perspectives and future trends. J Public Health Res 2020 Jun 04;9(1):1786 [FREE Full text] [doi: 10.4081/jphr.2020.1786] [Medline: 32550223]
31. Amul GG, Ang M, Kraybill D, Ong SE, Yoong J. Responses to COVID-19 in Southeast Asia: diverse paths and ongoing challenges. Asian Econ Policy Rev 2021 Aug 11;17(1):90-110. [doi: 10.1111/aepr.12362]
32. Price-Haywood EG. Clinical comparative effectiveness research through the lens of healthcare decisionmakers. Ochsner J 2015;15(2):154-161 [FREE Full text] [Medline: 26130978]
33. Loh T, Kang MS. Singapore: a diverse research hub of Asia. J Org Chem 2021 Jun 18;86(12):7893-7894. [doi: 10.1021/acs.joc.1c01114] [Medline: 34114455]
34. Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong Principles for assessing researchers: fostering research integrity. PLoS Biol 2020 Jul;18(7):e3000737 [FREE Full text] [doi: 10.1371/journal.pbio.3000737] [Medline: 32673304]
35. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1(Suppl 1):S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
36. Milcent C. From conventional healthcare to e-health: social and spatial transformation. Using a comparison between Hong Kong and Mainland China. J Clin Transl Res 2021 Oct 26;7(5):701-711 [FREE Full text] [Medline: 34786515]
37. Sek AC, Cheung NT, Choy KM, Wong WN, Tong AY, Fung VH, et al. A territory-wide electronic health record--from concept to practicality: the Hong Kong experience. Stud Health Technol Inform 2007;129(Pt 1):293-296. [Medline: 17911725]
38. Kim T, Sharma M, Teerawattananon Y, Oh C, Ong L, Hangoma P, et al. Addressing challenges in health technology assessment institutionalization for furtherance of universal health coverage through south-south knowledge exchange: lessons from Bhutan, Kenya, Thailand, and Zambia. Value Health Reg Issues 2021 May;24:187-192 [FREE Full text] [doi: 10.1016/j.vhri.2020.12.011] [Medline: 33838558]
39. Chavarina KK, Faradiba D, Teerawattananon Y. Navigating HTA implementation: a review of Indonesia's revised HTA guideline. Lancet Reg Health Southeast Asia 2023 Oct;17:100280 [FREE Full text] [doi: 10.1016/j.lansea.2023.100280] [Medline: 37727364]
40. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
41. Teerawattananon Y, Luz K, Yothasmutra C, Pwu RF, Ahn J, Shafie AA, et al. Historical development of the HTAsiaLINK network and its key determinants of success. Int J Technol Assess Health Care 2018 Jan;34(3):260-266 [FREE Full text] [doi: 10.1017/S0266462318000223] [Medline: 29911515]
42. O'Rourke B, Werkö SS, Merlin T, Huang LY, Schuller T. The 'top 10' challenges for health technology assessment: INAHTA viewpoint. Int J Technol Assess Health Care 2020;36(1):1-4. [doi: 10.1017/S0266462319000825] [Medline: 31775943]
43. Crane G, Lim JC, Gau CS, Xie J, Chu L. The challenges and opportunities in using real-world data to drive advances in healthcare in East Asia: expert panel recommendations. Curr Med Res Opin 2022 Sep;38(9):1543-1551. [doi: 10.1080/03007995.2022.2096354] [Medline: 35786170]

Abbreviations

ANC: average number of collaborative countries
CCCS: cross-country collaboration study
CER: comparative effectiveness research
CVM: cardiology and metabolic disorders
EHR: electronic health record
EMR: electronic medical record
HTA: health technology assessment
IDV: infectious diseases and vaccines
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PRO: patient-reported outcome
RWD: real-world data
RWE: real-world evidence
SCS: single-country study
SMA: simple moving average

XSL•FO
RenderX

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 23.01.24; peer-reviewed by A R, O Steichen, S Shao; comments to author 20.02.24; revised version received 11.03.24; accepted 07.05.24; published 11.06.24.

Please cite as:
Shau WY, Santoso H, Jip V, Setia S
Integrated Real-World Data Warehouses Across 7 Evolving Asian Health Care Systems: Scoping Review
J Med Internet Res 2024;26:e56686
URL: https://www.jmir.org/2024/1/e56686
doi:10.2196/56686
PMID:38749399

©Wen-Yi Shau, Handoko Santoso, Vincent Jip, Sajita Setia. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effects of mHealth Interventions on Quality of Life, Anxiety, and Depression in Patients With Coronary Heart Disease: Meta-Analysis of Randomized Controlled Trials

Qiao Ling Hou[1], BSN; Le Yang Liu[1], BSN; Ying Wu[1], PhD, Prof Dr

School of Nursing, Capital Medical University, Beijing, China

Corresponding Author:
Ying Wu, PhD, Prof Dr
School of Nursing
Capital Medical University
10 You-an-men Wai Xi-tou-tiao, Feng-tai District
Beijing, 100069
China
Phone: 86 10 8391 1766
Email: helenywu@vip.163.com

Abstract

Background: Coronary heart disease (CHD) is the leading cause of death globally. In addition, 20% to 40% of the patients with CHD have comorbid mental health issues such as anxiety or depression, affecting the prognosis and quality of life (QoL). Mobile health (mHealth) interventions have been developed and are widely used; however, the evidence for the effects of mHealth interventions on QoL, anxiety, and depression in patients with CHD is currently ambiguous.

Objective: In this study, we aimed to assess the effects of mHealth interventions on QoL, anxiety, and depression in patients with CHD.

Methods: We searched the Cochrane Library, PubMed, Embase, CINAHL, Web of Science, China National Knowledge Infrastructure, and Wanfang databases from inception to August 12, 2023. Eligible studies were randomized controlled trials that involved patients with CHD who received mHealth interventions and that reported on QoL, anxiety, or depression outcomes. We used the Cochrane risk-of-bias tool for randomized trials to evaluate the risk of bias in the studies, ensuring a rigorous and methodologically sound analysis. Review Manager (desktop version 5.4; The Cochrane Collaboration) and Stata MP (version 17.0; StataCorp LLC) were used to conduct the meta-analysis. The effect size was calculated using the standardized mean difference (SMD) and its 95% CI.

Results: The meta-analysis included 23 studies (5406 participants in total) and showed that mHealth interventions significantly improved QoL in patients with CHD (SMD 0.49, 95% CI 0.25-0.72; Z=4.07; $P<.001$) as well as relieved their anxiety (SMD −0.46, 95% CI −0.83 to −0.08; Z=2.38; $P=.02$) and depression (SMD −0.34, 95% CI −0.56 to −0.12; Z=3.00; $P=.003$) compared to usual care. The subgroup analyses indicated a significant effect favoring the mHealth intervention on reducing anxiety and depressive symptoms compared to usual care, especially when (1) the intervention duration was ≥6 months ($P=.04$ and $P=.001$), (2) the mHealth intervention was a simple one (only 1 mHealth intervention was used) ($P=.01$ and $P<.001$), (3) it was implemented during the COVID-19 pandemic ($P=.04$ and $P=.01$), (4) it was implemented in low- or middle-income countries ($P=.01$ and $P=.02$), (5) the intervention focused on mental health ($P=.01$ and $P=.007$), and (6) adherence rates were high (≥90%; $P=.03$ and $P=.002$). In addition, comparing mHealth interventions to usual care, there was an improvement in QoL when (1) the mHealth intervention was a simple one ($P<.001$), (2) it was implemented in low- or middle-income countries ($P<.001$), and (3) the intervention focused on mental health ($P<.001$).

Conclusions: On the basis of the existing evidence, mHealth interventions might be effective in improving QoL and reducing anxiety and depression in patients with CHD. However, large sample, high-quality, and rigorously designed randomized controlled trials are needed to provide further evidence.

Trial Registration: PROSPERO CRD42022383858; https://tinyurl.com/3ea2npxf

(J Med Internet Res 2024;26:e52341) doi:10.2196/52341

XSL•FO
RenderX

KEYWORDS

mobile health; coronary heart disease; quality of life; anxiety; depression; meta-analysis; mobile phone

Introduction

Background

Coronary heart disease (CHD) is the leading cause of death globally, taking an estimated 8.9 million lives each year, representing 16% of all deaths [1-4]. The prevalence of psychological disorders in patients with CHD is higher than that in the general population according to mounting evidence, and CHD morbidity and mortality are correlated with mental health status [5,6]. According to earlier research, 20% to 40% of the patients with CHD have concurrent mental health issues, such as anxiety and depression [7,8], which are not only risk factors for inducing or exacerbating CHD but also directly affect the prognosis, increase the risk of recurring major adverse cardiac events and death by reducing patients' compliance with treatment, and greatly reduce the clinical benefits of treatment [9]. Meanwhile, comorbid anxiety and depression are associated with diminished quality of life (QoL) [10,11].

A study involving 190 patients with CHD who were followed for 36 months found that the incidence of anxiety and depression increased over time, with anxiety increasing persistently from 42.6% to 51.1% and depression increasing from 33.3% to 43.7% [5]. Therefore, the European Society of Cardiology and the American Heart Association guidelines on the secondary prevention of CHD make explicit recommendations for alleviating psychological problems such as anxiety or depression in patients with CHD [3,12]. Furthermore, to improve well-being, supportive interventions are required to assist patients in dealing with challenges related to their QoL, anxiety, and depression.

Currently, guidelines state that patients with CHD with anxiety and depression should receive increased attention and assistance to enhance adherence to lifestyle changes, engage in psychotherapy, and receive pharmacological therapies [3]. Evidence suggests that healthy behaviors such as quitting smoking, increasing physical activity, and improving diet are beneficial and applicable to patients with CHD with mental issues [13-15]. Psychotherapy, such as psychological counseling and cognitive behavioral therapy, can improve patients' self-efficacy and lead to a healthy lifestyle, reducing the risks of CHD recurrence and mortality [16,17]. However, these methods are time consuming and labor intensive, and they lack sustainability, which makes it difficult to meet the needs of patients and promote well-being [18]. In addition, evidence indicates that patients with CHD with moderate to severe major depression should be considered for antidepressive treatment with a selective serotonin reuptake inhibitor, which can lower the rates of CHD readmission and all-cause mortality [19], but it can have some side effects [20].

Hence, there is a need for sustainable, time-saving, and effective ways to improve patients' QoL as well as reduce their anxiety and depression. The exponential growth and availability of mobile health (mHealth) technologies may offer a fresh stress management strategy [21]. The World Health Organization describes mHealth as a subfield of eHealth that involves the delivery of health services and information via mobile technology such as mobile phones and PDAs [22,23].

The Effectiveness of mHealth Interventions

There is rising evidence of the effectiveness of mHealth interventions on QoL, anxiety, and depression in patients with CHD. However, the evidence is often inconsistent or even conflicting. The studies by Yudi et al [24] and Houchen-Wolloff et al [25] found that, compared to usual care, mHealth interventions did not effectively relieve anxiety and depressive symptoms in patients with CHD. By contrast, the results of other studies showed that the mHealth group significantly outperformed the control group in terms of QoL and psychological distress [26,27]. The disparate conclusions may be due to the fact that there are still questions and doubts about the heterogeneity of the effects of mHealth interventions across various intervention contents, such as different intervention methods, durations, and adherence rates. This lack of clear understanding about how these factors influence outcomes limits the availability of definitive evidence regarding the psychological care of patients with CHD.

Hence, our meta-analysis aims to investigate the effectiveness of mHealth interventions on QoL, anxiety, and depression outcomes in patients with CHD.

Methods

Study Design

The study protocol was registered on PROSPERO (CRD42022383858), and we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [28,29] (Multimedia Appendix 1).

Search Strategy

We conducted a thorough literature search to identify randomized controlled trials (RCTs) related to mHealth interventions for patients with CHD. The databases searched were Cochrane Library, PubMed, Embase, CINAHL, Web of Science, China National Knowledge Infrastructure, and Wanfang (Multimedia Appendix 2). In addition, we manually reviewed the reference lists of the included full texts, performed a supplementary search on the ClinicalTrials.gov trial registration platform, and contacted the authors for further information if needed.

The search terms were constructed by combining subject words with free words. The key terms included *mobile health*, *mHealth*, *eHealth*, *telehealth*, *telemedicine*,*digital health*, *mobile phone*, *cellphone*, *smartphone*, *smartphone app**, *cellphone app**, *mobile app**, *smartphone-based*, *cellphone-based*, *mobile phone-based*, *portable electronic app**, *portable software app** and *coronary disease*, *coronary heart disease*, *coronary artery disease*, *myocardial ischemia*, *myocardial revascularization*, *acute coronary syndrome*, *coronary artery bypass*, *percutaneous coronary intervention*.

XSL•FO
RenderX

Eligibility Criteria

The eligibility criteria were structured according to population, intervention, comparison, outcomes, and study design (PICOS).

- Population: adults (aged ≥18 years) with CHD, including patients with myocardial infarction, revascularization (stent, coronary artery bypass grafting, or percutaneous transluminal coronary angioplasty), or CHD confirmed by angiography
- Intervention: mHealth interventions (eg, mobile based or web based monitoring devices, PDAs, or wireless devices; no restriction on duration, frequency, or type of intervention)
- Comparison: comparison group did not receive mHealth intervention; instead, it received a different intervention or placebo
- Outcomes: studies that reported QoL, anxiety, or depression as outcomes
- Study design: only RCTs (no restrictions on language; studies in languages other than English and Chinese would be translated using translation software)

Exclusion Criteria

Studies were excluded if (1) after contacting the authors, the complete text was unavailable, and important data could not be extracted; (2) the outcomes were unclear, and the data could not be combined, transformed, and used for analysis; (3) they were duplicate publications; or (4) they were pilot studies.

Study Selection

All studies retrieved from the database search were managed and organized by exporting the citations to EndNote X9 (Clarivate) and Rayyan (a web application used to screen a large number of records for rapid selection; Rayyan Systems Inc) [30]. Microsoft Excel was used for data extraction. Two reviewers (QLH and LYL) screened the titles, abstracts, and full texts of studies independently. Disagreements between the reviewers regarding the title and abstract screening, full-text review, and reasons for exclusion were resolved by discussion with the third reviewer (YW).

Data Extraction

A prespecified electronic data extraction table following the PRISMA guidelines [28,29] was designed to extract (1) study information (author, year, country of origin, sample size, intervention methods, frequency, duration, and adherence rates), (2) primary and secondary outcomes as well as measurement scales, and (3) key information for the risk-of-bias assessment. The data extraction table was pilot-tested by QLH with 3 studies (reviewed by LYL). Next, authors QLH and LYL retrieved data from the full-text papers, and conflicts were resolved by discussion and consultation with the third author (YW).

Risk of Bias Assessment

Two reviewers (QLH and LYL) independently assessed the quality of studies according to the Cochrane Handbook for Systematic Reviews of Interventions (version 6.3) [31], which includes 7 domains: (1) random sequence generation (selection bias), (2) allocation sequence concealment (selection bias), (3) blinding of participants and personnel (performance bias), (4) blinding of outcome assessment (detection bias), (5) incomplete outcome data (attrition bias), (6) selective outcome reporting (reporting bias), and (7) other sources of bias. The criteria for low, unclear, and high risk of bias within and across studies followed the Cochrane Handbook for Systematic Reviews of Interventions (version 6.3) [31]. A quality grade of *A* (low bias) was assigned if 7 domains were all low risk of bias, *B* (moderate bias) if ≥1 domains were unclear or high risk of bias, and *C* (high bias) if none of the 7 domains were satisfied. Disagreements between the 2 reviewers were settled by discussion with the third reviewer (YW). The ultimate decision was taken by consensus of the 3 reviewers.

Data Synthesis and Statistical Analysis

The meta-analysis was conducted using Review Manager (desktop version 5.4; The Cochrane Collaboration) and Stata MP (version 17.0; StataCorp LLC). We selected standardized mean difference (SMD) and 95% CI as the effect indicators because the QoL, anxiety, and depression were measured using different scales in different studies. Heterogeneity was evaluated using the I^2 statistic and defined as *not important* (I^2≤25%), *moderate* (I^2=26%-50%), *substantial* (I^2=51%-75%), and *considerable* (I^2>75%) [32]. Depending on the I^2 values, a random effects model (I^2≥50%) or a fixed effects model (I^2<50%) was selected.

To investigate the heterogeneity across several subgroups, subgroup analyses were performed for the different intervention durations (<6 months or ≥6 months), different intervention methods (simple intervention or complex intervention), the influence of the COVID-19 pandemic (before or during the pandemic), different country types (high-income country [HIC] or low- or middle-income country [LMIC]), different intervention contents (whether the mHealth intervention focused on mental health), and different levels of adherence (≥90% or <90%) to explore the heterogeneity among several subgroups. A simple mHealth group was defined as one that used only 1 mHealth intervention, such as a WeChat message or an SMS text message, while a complex mHealth group was defined as one that used multiple mHealth interventions, such as simultaneous use of smartphone apps and wearable devices. Different levels of adherence were defined as *higher* (intervention group adherence rate was ≥90%) and *lower* (intervention group adherence rate was <90%).

All statistical tests were 2-tailed, and P<.05 was considered statistically significant. We used funnel plots to evaluate publication bias, and plot asymmetry was tested using the Begg and Egger tests in Stata MP (version 17.0). The leave-one-out strategy was used to find outliers and analyze their influences. Moreover, the trim-and-fill method developed by Duval and Tweedie [33] was used to eliminate any tiny studies with substantial effect sizes from the positive side of the funnel plot to further clarify the potential impact of publication bias. In addition, a sensitivity analysis was carried out by redoing the meta-analysis after eliminating papers that had a high probability of bias [34]. The data used in the meta-analysis are all postintervention data.

XSL•FO
RenderX

Results

Overview

Figure 1 depicts the meta-analysis study selection process. On the basis of the search terms we defined, our initial search generated 6091 publications. Of these 6091 publications, we excluded 2783 (45.69%) duplicates and irrelevant publications. Of the remaining 3308 publications, we excluded 2982 (90.15%) after abstract screening. Of the 326 full-text reports obtained, we excluded 303 (92.9%) on the basis of the selection criteria and included 23 (7.1%) in the analysis. Of these 23 studies, 19 (83%) were in English and 4 (17%) in Chinese; they were all published between 2015 and 2023.

Figure 1. Flowchart of the literature search. The databases searched were the Cochrane Library (n=1110), Web of Science (n=1484), PubMed (n=448), Embase (n=555), CINAHL (n=1608), China National Knowledge Infrastructure (n=453), and Wanfang (n=432).

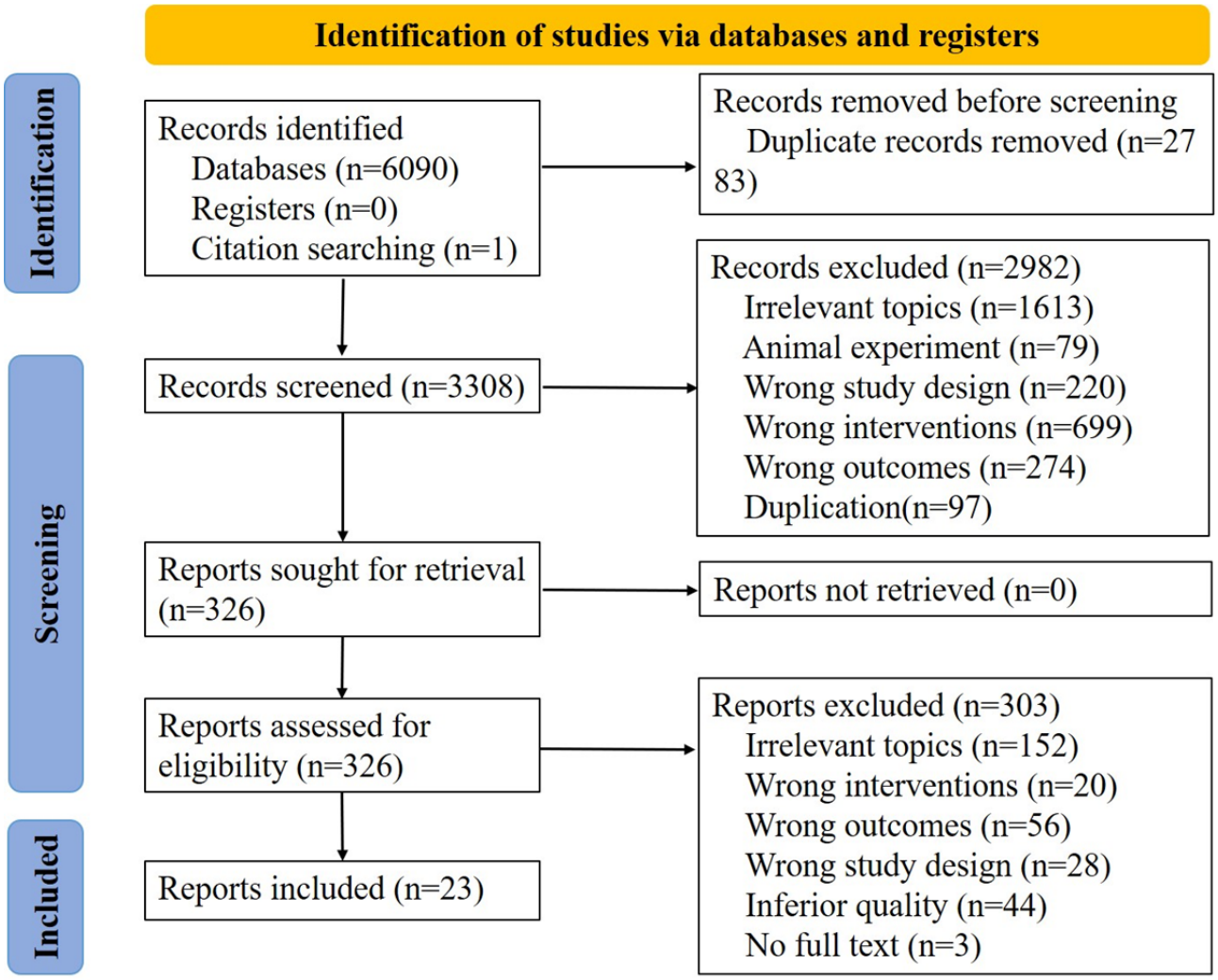

Characteristics of the Included Studies

Participants

The 23 RCTs selected [24-27,35-53] included 5406 participants (n=2708, 50.09% participants in the intervention groups and n=2698, 49.91% in the control groups), with sample sizes per study varying from 51 to 1424. All patients with CHD in the 23 RCTs were in the cardiac rehabilitation (CR) stage, including those who had undergone revascularization procedures (such as coronary artery bypass grafting or percutaneous transluminal coronary angioplasty) as well as patients with CHD confirmed through angiography. The RCTs were conducted in Asia (12/23, 52%) [27,37-41,43,45,48-51], Europe (6/23, 26%) [25,26,35,42,44,47], and Australia and New Zealand (5/23, 22%) [24,36,46,52,53]. Of the 23 studies, 19 (83%) reported QoL, 13 (57%) reported anxiety, and 17 (74%) reported depression (Tables 1 and 2).

XSL•FO
RenderX

Table 1. Characteristics of included studies (n=23).

Study and year	Country	Participants	Stage of disease	Sample	IG[a]	CG[b]	Retention rate, n/N (%)
Batalik et al [35], 2020	Czech Republic	Aged >18 years, diagnosed with CVD[c] (angina pectoris, myocardial infarction in the last 6 months, with left ventricular ejection fraction >45%)	Telerehabilitation	51 (IG: 25; CG: 26)	Wrist heart rate monitor	Usual care	IG: 25/28 (89.3); CG: 26/28 (92.9)
Cheung et al [53], 2023	Australia	CHD[d] was defined as previous myocardial infarction or documented >50% occlusion of a major coronary artery on coronary angiography	Previous myocardial infarction	453 (IG: 228; CG: 225)	Semipersonalized SMS text messaging service	Usual care	IG: 228/256 (89.1); CG: 225/250 (90)
Chow et al [36], 2022	Australia	Diagnosis of ACS[e]	Secondary prevention	1424 (IG: 716; CG: 708)	SMS text messaging (customized and personalized)	Usual care	IG: 697/716 (97.3); CG: 682/708 (96.3)
Dalli Peydró et al [26], 2022	Spain	Low-risk ACS, left ventricular ejection fraction ≥50%	Cardiac telerehabilitation	59 (IG: 31; CG: 28)	Monitoring of heart rate via participant smartphone and heart rate monitor	Usual care	IG: 31/33 (93.9); CG: 28/34 (82.4)
Dorje et al [37], 2019	China	Aged ≥18 years with documented CHD (including myocardial infarction and unstable or stable angina)	CR[f] and secondary prevention	312 (IG: 156; CG: 156)	Smartphone-based home-based CR (individualized exercise prescription and remote supervision)	Usual care	IG: 134/156 (85.9); CG: 131/156 (84)
Duan et al [38], 2018	China	Aged between 18 and 75 years, CHD	Home-based CR	114 (IG: 60; CG: 54)	8-week web-based intervention	Usual care	IG: 44/60 (73.3); CG: 39/54 (72.2)
Fang et al [39], 2019	China	Patients with low risk after PCI[g]	CR	67 (IG: 33; CG: 34)	Real-time physiological monitoring; home visits; telephone call	Usual care	IG: 33/40 (82.5); CG: 34/40 (85)
Hisam et al [40], 2022	Pakistan	Patients after ACS (ST-elevation myocardial infarction, non–ST-elevation myocardial infarction, and unstable angina)	CR	160 (IG: 80; CG: 80)	Diurnal mobile texting (standardized messages about healthy lifestyle changes through app)	Standard care	IG: 70/80 (87.5); CG: 49/80 (61.3)
Houchen-Wolloff et al [25], 2018	United Kingdom	Confirmed primary diagnosis of CHD (including angina, after myocardial infarction and after PCI	CR	60 (IG: 37; CG: 23)	Web-based CR program	Usual care	IG: 33/37 (89.2); CG: 21/23 (91.3)
Huang et al [41], 2017	China	Patients with unstable angina pectoris and undergoing their first elective PCI	CR	64 (IG: 31; CG: 33)	WeChat	Usual care	IG: 31/31 (100); CG: 33/33 (100)
Johnston et al [42], 2016	Sweden	Women or men aged >18 years, diagnosed as having an ST-elevation myocardial infarction or a non–ST-elevation myocardial infarction	Secondary prevention stage	151 (IG: 85; CG: 77)	Web-based application (interactive and personalized)	Simplified drug adherence diary	IG: 85/91 (93.4); CG: 77/83 (92.8)
Kang et al [43], 2021	Korea	Clinical ASCVD[h]	Secondary prevention stage	643 (IG: 322; CG: 321)	A smartphone app (personalized)	Usual care	IG: 322/333 (96.7); CG: 321/333 (96.4)
Kang G et al [27], 2023	China	Participants were required to be aged ≥18 years and to have a diagnosis of stable CAD[i]	Secondary prevention stage	200 (IG: 100; CG: 100)	WeChat	Usual care	IG: 98/100 (98); CG: 98/100 (98)

XSL•FO
RenderX

Study and year	Country	Participants	Stage of disease	Sample	IG[a]	CG[b]	Retention rate, n/N (%)
Kraal et al [44], 2017	Netherlands	Patients who entered CR after ACS (myocardial infarction or unstable angina) or a revascularization procedure (PCI or coronary artery bypass grafting)	Home-based CR	78 (IG: 37; CG: 41)	Heart rate monitor and web application	Usual care	IG: 37/45 (82.2); CG: 41/45 (91.1)
Pakrad et al [45], 2021	Iran	Patients who have undergone coronary artery bypass grafting	CR	88 (IG: 44; CG: 44)	CR was delivered in person in the first month; over the following 3 months, it was delivered via smartphone	Usual care	IG: 42/44 (95.5); CG: 39/44 (88.6)
Pfaeffli Dale et al [52], 2015	New Zealand	CHD (myocardial infarction, angina, or revascularization)	CR	123 (IG: 61; CG: 62)	Personalized SMS text messaging service, supporting website, and pedometer	Usual care	IG: 57/61 (93.4); CG: 59/62 (95.2)
Shariful Islam et al [46], 2019	Australia	CHD was defined as myocardial infarction, coronary artery bypass grafting, or PCI or proven angiographically	Secondary prevention	883 (IG: 333; CG: 350)	SMS text messaging	Usual care	IG: 333/352 (94.6); CG: 350/358 (97.8)
Snoek et al [47], 2021	Netherlands	ACS, PCI, or coronary artery bypass grafting within 3 months before the start of the CR program	CR	122 (IG: 61; CG: 61)	Smartphone and Bluetooth-connected heart rate belt	Usual care	IG: 59/61 (96.7); CG: 59/61 (96.7)
Su and Yu [48], 2021	China	Initial diagnosis of CHD based on angiography or the exacerbation of CHD in previously diagnosed cases	Nurse-led eHealth CR	146 (IG: 73; CG: 73)	Nurse-led eHealth CR platform and WeChat platform	Usual care	IG: 66/73 (90.4); CG: 58/73 (79.5)
Wang [49], 2018	China	Patients aged ≥18 years, and coronary artery stenosis indicated by first coronary angiography and successfully implanted one or more stents	Postdischarge stage	51 (IG: 23; CG: 28)	Mobile app	Usual care	IG: 23/32 (71.9); CG: 28/32 (87.5)
Yudi et al [24], 2021	Australia	Patient over 18 years old with a diagnosis of an ACS and documented coronary artery disease on angiography (coronary artery stenosis >50%)	Cardiac rehabilitation	168 (IG: 83; CG: 85)	Adjunctive smartphone-based CR program (SCRP)	Usual care	IG: 83/103 (80.6); CG: 85/103 (82.5)
Zhang et al [50], 2020	China	Patients who underwent a first and successful PCI procedure	Continuous care	88 (IG: 44; CG: 44)	Mobile app	Usual care	IG: 44/44 (100); CG: 44/44 (100)
Zheng [51], 2021	China	Patients undergoing initial PCI for CAD	Continuous care	90 (IG: 45; CG: 45)	Follow-up continuous nursing intervention	Usual care	IG: 45/45 (100); CG: 45/45 (100)

[a]IG: intervention group.

[b]CG: control group.

[c]CVD: cardiovascular disease.

[d]CHD: coronary heart disease.

[e]ACS: acute coronary syndrome.

[f]CR: cardiac rehabilitation.

[g]PCI: percutaneous coronary intervention.

[h]ASCVD: atherosclerotic cardiovascular disease.

[i]CAD: coronary artery disease.

XSL•FO
RenderX

Table 2. Characteristics of included studies (n=23).

Study and year	Intervention frequency	Intervention duration (months)	Adverse events	Outcomes	Scales
Batalik et al [35], 2020	3 times a week	3	Remotely monitored telerehabilitation seems to be safe, and no adverse cardiac events occurred during the intervention	A[a]	SF-36[b]
Cheung et al [53], 2023	4 messages a week	6	No adverse events or deaths were deemed related to the intervention	A and B[c]	SF-12[d] and PHQ-9[e]
Chow et al [36], 2022	4 messages a week in the first 6 months and 3 messages a week over the subsequent 6 months	12	The intervention was safe	A, B, and C[f]	SF-12, PHQ-9, and GAD-7[g]
Dalli Peydró et al [26] 2022	Daily	10	During the study period, 3 patients were readmitted, but the readmissions were not related to the intervention	A, B, and C	EQ-5D-5L and HADS[h]
Dorje et al [37], 2019	First and second month: 4 times per week; from third to sixth month: twice a week	6	No adverse events or SMART-CR/SP[i] program-related safety issues were recorded during the study	A, B, and C	SF-12 and PHQ-9
Duan et al [38], 2018	Once a week	2	Not mentioned	A and B	WHOQOL[j] and CES-D[k]
Fang et al [39], 2019	3 times a week	1.5	Not mentioned	A and B	SF-36 and CDS[l]
Hisam et al [40], 2022	Daily	6	Not mentioned, but the intervention in this study was a medically supervised cardiac rehabilitation program	A	SF-12 and Mac-New[m]
Houchen-Wolloff et al [25], 2018	Daily	6	There were 2 adverse events in the web group and 4 in the control group, but all were unrelated to the study	A, B, and C	MacNew and HADS
Huang et al [41], 2017	Every Monday and Thursday	3	Not mentioned	A	SAQ[n]
Johnston et al [42] 2016	Daily	6	No adverse device effects were reported during the study	A	EQ-5D VAS
Kang et al [43], 2021	Daily	6	Not mentioned	B	BDI[o]
Kang G et al [27], 2023	3 times a week	12	There were no obvious differences between the 2 groups in terms of adverse events	A, B, and C	SAQ, HAMA[p], and HAMD[q]
Kraal et al [44], 2017	At least 2 training sessions a week, and once a week, the patient received feedback	3	No serious adverse events were recorded during center-based and home-based training	A, B, and C	MacNew and HADS
Pakrad et al [45], 2021	8 in-person sessions and 4 group discussion sessions; in addition, each participant was contacted 24 times over 3 months through the mobile app	4	None of the intervention participants were hospitalized	A, B, and C	SF-36 and DASS-21[r]
Pfaeffli Dale et al [52], 2015	7 messages per week for 12 weeks, followed by 5 messages per week for the next 12 weeks	6	There were 13 serious adverse events reported (intervention: n=8; control: n=5), but none were study related	B and C	HADS
Shariful Islam et al [46], 2019	4 times a week	6	Not mentioned	A and B	SF-12 and PHQ-9
Snoek et al [47], 2021	5 days a week	6	No cardiovascular mortality and near sudden cardiac death was registered	A, B, and C	MacNew and HADS
Su and Yu [48], 2021	Daily	3	Not mentioned	A, B, and C	MacNew and DASS-21
Wang et al [49], 2018	Daily	6	Not mentioned	A	SAQ

XSL•FO
RenderX

Study and year	Intervention frequency	Intervention duration (months)	Adverse events	Outcomes	Scales
Yudi et al [24], 2021	Daily	2	Not mentioned	A, B, and C	SF-36, EQ-5D-5L, HADS, and CDS
Zhang et al [50], 2020	Daily	6	Not mentioned	B and C	SAS[s] and SDS[t]
Zheng [51], 2021	At 1 week and 1, 3, 6, and 12 months after discharge	12	Not mentioned	A, B, and C	SF-36, SAS, and SDS

[a]A: health-related quality of life.
[b]SF-36: Short Form Health Survey-36.
[c]B: depression.
[d]SF-12: Short Form Health Survey-12.
[e]PHQ-9: Patient Health Questionnaire-9.
[f]C: anxiety.
[g]GAD-7: General Anxiety Disorder-7.
[h]HADS: Hospital Anxiety and Depression Scale.
[i]SMART-CR/SP: smartphone- and social media–based cardiac rehabilitation and secondary prevention.
[j]WHOQOL: World Health Organization Quality of Life.
[k]CES-D: Center for Epidemiologic Studies Depression Scale.
[l]CDS: Cardiac Depression Scale.
[m]MacNew: MacNew Heart Disease Health-Related Quality of Life.
[n]SAQ: Seattle Angina Questionnaire.
[o]BDI: Beck Depression Inventory.
[p]HAMA: Hamilton Anxiety Scale.
[q]HAMD: Hamilton Depression Scale.
[r]DASS-21: Depression, Anxiety, and Stress Scale-21.
[s]SAS: Self-Rating Anxiety Scale.
[t]SDS: Self-Rating Depression Scale.

The Intervention Groups

In the 23 studies, the main mHealth intervention methods used were SMS text messaging (4/23, 17%), WeChat (4/23, 17%), mobile apps (11/23, 48%), wearable devices (5/23, 22%), and web-based interventions (5/23, 22%). The mHealth intervention durations ranged from 6 weeks to 12 months; of the 23 studies, 8 (35%) had intervention durations <6 months, and 15 (65%) had intervention durations ≥6 months. Of the 23 studies, 16 (70%) used only 1 mHealth intervention, whereas 7 (30%) used ≥2 mHealth interventions simultaneously. Of the 23 studies, 18 (78%) were conducted and completed before the outbreak of the COVID-19 pandemic, whereas 5 (22%) were conducted during the COVID-19 pandemic; furthermore, 12 (52%) were conducted in HICs and 11 (48%) in LMICs. In 9 (39%) of the 23 studies, the mHealth interventions focused on mental health, whereas in 14 (61%) studies, the mHealth interventions were less relevant to mental health. Of the 23 studies, 10 (43%) have lower (<90%) adherence rates, whereas 13 (57%) have *higher* (≥90%) adherence rates.

The Control Groups

The control group received usual care, including routine psychological counseling, health education, paper-based CHD educational booklets, and follow-up after discharge.

Outcomes

In sum, of the 23 studies, 19 (83%) reported QoL scores using the Short Form Health Survey-36 (n=5, 26%), the Short Form Health Survey-12 (n=4, 21%), the EQ-5D-5L (n=3, 16%), the MacNew Heart Disease Health-Related Quality of Life scale (n=4, 21%), and the Seattle Angina Questionnaire (n=3, 16%). Anxiety scores were reported by 13 (57%) of the 23 studies. Anxiety was measured with the Hospital Anxiety and Depression Scale (6/13, 46%); the Self-Rating Anxiety Scale (2/13, 15%); the Hamilton Anxiety Scale (1/13, 8%); the Depression, Anxiety, and Stress Scale-21 (2/13, 15%); and the Generalized Anxiety Disorder-7 (2/13, 15%). In addition, 17 (74%) of the 23 studies reported depression scores using the Hospital Anxiety and Depression Scale (n=5, 29%); the Self-Rating Depression Scale (n=2, 12%); the Hamilton Depression Scale (n=1, 6%); the Patient Health Questionnaire-9 (n=4, 24%); the Depression, Anxiety, and Stress Scale-21 (n=1, 6%); the Cardiac Depression Scale (n=2, 12%); the Center for Epidemiologic Studies Depression Scale (n=1, 6%); and the Beck Depression Inventory (n=1, 6%).

Risk of Bias in Studies

The Cochrane Handbook for Systematic Reviews of Interventions (version 6.3) was used to evaluate bias [54]. Of the 23 studies, 17 (74%) [25,27,35-37,40,41,43-50,52,53,55] used a computer-generated list of random numbers to allocate

XSL•FO
RenderX

individuals to the intervention or control groups, which was deemed minimal risk of bias, and 10 (43%) [26,36,37,40,44-46,48,52,53] concealed the allocation of participants, which was considered high risk of bias. Given the nature of the mHealth intervention, participants and personnel were blinded in only 4 (17%) of the 23 studies [37,39,46,47,55]. In 13 (57%) of the 23 studies [24,26,27,35-37,40,44-48,55], the outcome assessors were blinded. Of the 23 studies, 2 (9%) were grade *A* (low bias), and 21 (91%) were grade *B* (moderate bias; Figure 2; Figure S1 in Multimedia Appendix 2) [24-27,35-53].

Figure 2. Analysis of the risk of bias.

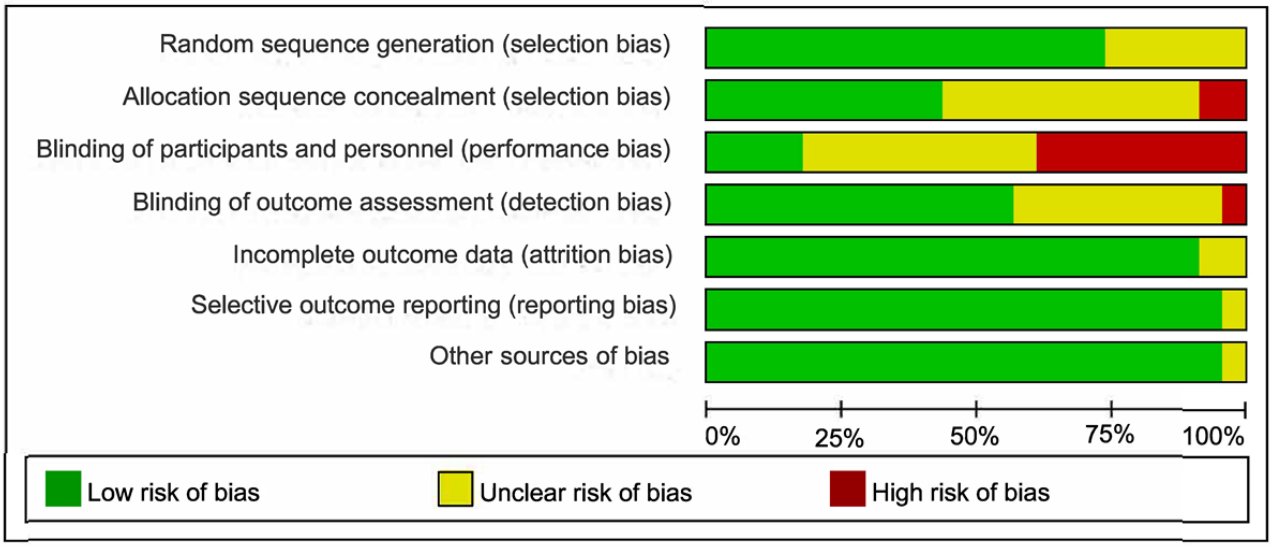

Analysis of Overall Effects

QoL Measurement

QoL was measured in 19 (83%) of the 23 studies. Figure 3 [24-27,35-42,44,45,47-49,51,53] displays the QoL scores from these 19 studies, illustrating that mHealth interventions significantly enhance QoL in patients with CHD compared to usual care (SMD 0.49, 95% CI 0.25-0.72; Z=4.07; *P*<.001). These studies showed considerable heterogeneity (I^2=90.4%; *P*<.001).

Figure 3. Forest plot of quality of life measured in 19 studies.

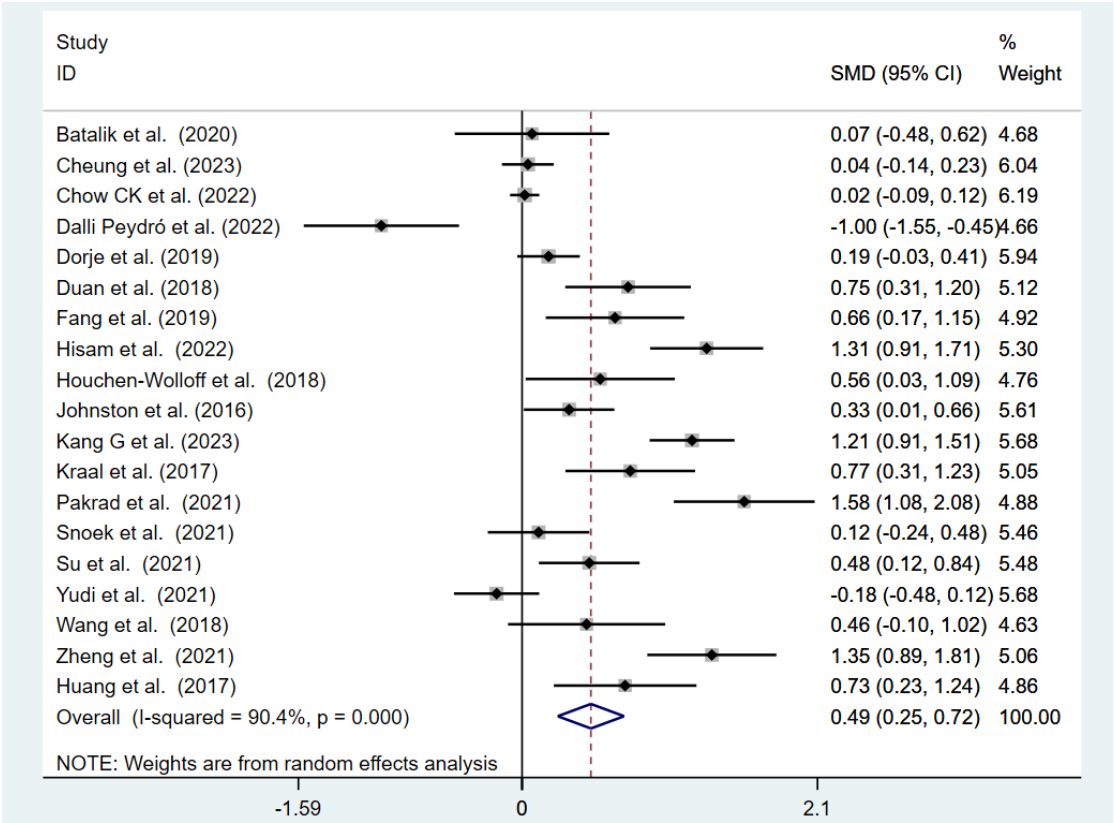

XSL•FO
RenderX

Anxiety

Anxiety was measured in 13 (57%) of the 23 studies. The anxiety scores displayed in Figure 4 [24-27,36,37,44,45,47,48,50-52] show that the mHealth interventions in these 13 studies relieved anxiety symptoms in patients with CHD compared to usual care (SMD −0.46, 95% CI −0.83 to −0.08; Z=2.38; *P*=.02). These studies revealed considerable heterogeneity (I^2=94.7%; *P*<.001).

Figure 4. Forest plot of anxiety measured in 13 studies.

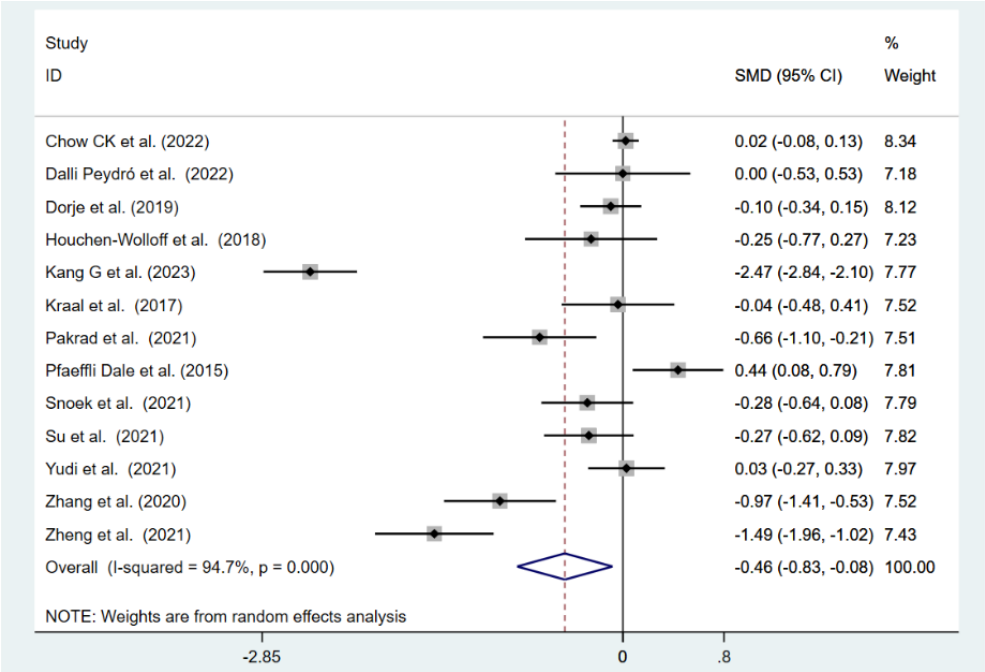

Depression

Depression was measured in 17 (74%) of the 23 studies. The depression scores displayed in Figure 5 [24-27,36-39,43-47,50-53] show that the mHealth intervention in these 17 studies relieved depressive symptoms in patients with CHD compared to usual care (SMD −0.34, 95% CI −0.56 to −0.12; Z=3.00; *P*=.003). These studies showed considerable heterogeneity (I^2=91.7%; *P*<.001).

Figure 5. Forest plot of depression measured in 17 studies.

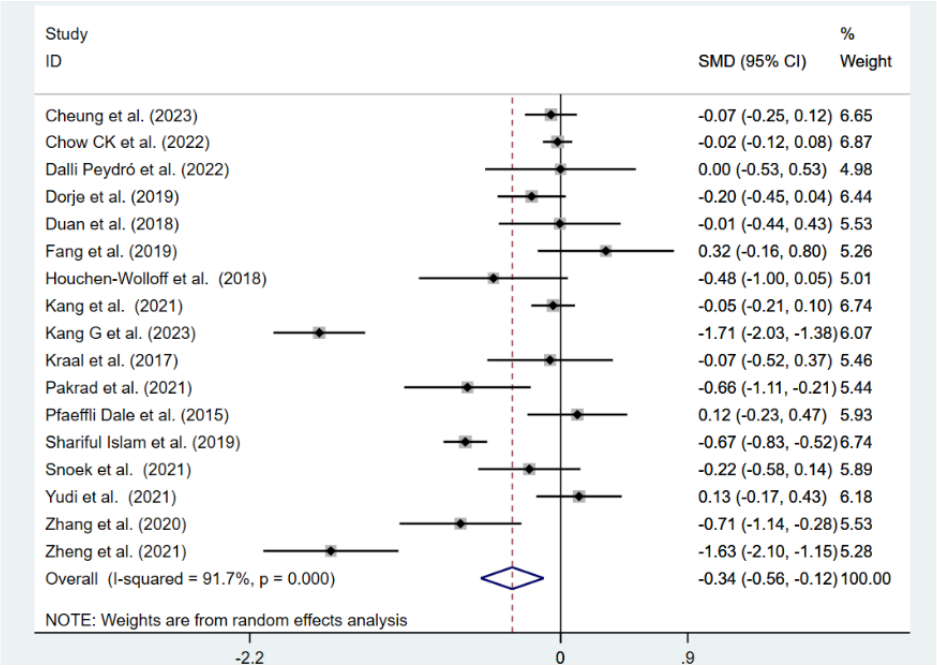

Publication Bias

The funnel plots for QoL (Egger test: *P*=.02; Begg test: *P*=.58) indicated that there is a certain publication bias. However, the trim-and-fill plots showed that the results were robust before (*P*<.001) and after (*P*<.001) the trim-and-fill analysis. That is, there is a publication bias regarding QoL, but it does not affect

XSL•FO
RenderX

the validity of the results (Figures S2 and S3 in Multimedia Appendix 2).

The funnel plots for anxiety (Egger test: P=.12; Begg test: P=.13) and depression (Egger test: P=.29; Begg test: P=.23) exhibited a generally symmetrical pattern. Furthermore, the trim-and-fill plots revealed that the effect sizes and P values did not vary before and after the trim-and-fill analysis, which was similar to the funnel plot results, showing that there was a low possibility of publication bias (Figures S4-S7 in Multimedia Appendix 2).

Sensitivity Analysis

A sensitivity analysis was conducted to test whether the results may be significantly influenced by a single study by removing 1 study at a time and analyzing the remaining studies. The results showed that the overall effect was not considerably affected, indicating that the conclusions of this meta-analysis were relatively robust [34] (Figures S8-S10 in Multimedia Appendix 2).

Subgroup Analysis

Different Durations of Interventions

The Effects on QoL

Both short-duration (<6 mo; SMD 0.60, 95% CI 0.21-0.98; P=.002; I^2=83.8%) and long-duration (≥6 mo; SMD 0.42, 95% CI 0.11-0.72; P=.008; I^2=92.5%) mHealth interventions were found to significantly enhance QoL among patients with CHD compared to usual care (Figure S11 in Multimedia Appendix 2).

The Effects on Anxiety and Depression

A significant effect favoring the mHealth intervention in mitigating anxiety (SMD −0.56, 95% CI −1.09 to −0.03; P=.04; I^2=96.3%; Figure S12 in Multimedia Appendix 2) and depressive symptoms (SMD −0.45, 95% CI −0.73 to −0.18; P=.001; I^2=93.8%; Figure S13 in Multimedia Appendix 2) compared to usual care was seen when only studies with long intervention durations (≥6 mo) were included, while this effect was not observed when only studies with short intervention durations (<6 mo) were included.

Different Intervention Methods: Simple Versus Complex Interventions

A significant effect favoring the mHealth intervention on enhancing QoL (SMD 0.57, 95% CI 0.28-0.86; P<.001; I^2=92.1%; Figure S14 in Multimedia Appendix 2) as well as alleviating anxiety (SMD −0.73, 95% CI −1.29 to −0.16; P=.01; I^2=96.7%; Figure S15 in Multimedia Appendix 2) and depressive symptoms (SMD −0.52, 95% CI −0.81 to −0.23; P<.001; I^2=94.5%; Figure S16 in Multimedia Appendix 2) compared to usual care was seen when only studies using simple mHealth interventions were included, while this effect was not significant when only studies using complex mHealth interventions were included.

Influence of the COVID-19 Pandemic

The Effects on QoL

A comparative analysis of the studies conducted before the COVID-19 pandemic and those conducted during the pandemic revealed significant findings. The comparison of the studies conducted before the pandemic (14/19, 74%; SMD 0.30, 95% CI 0.14-0.45; P<.001; I^2=70%) and those conducted during the pandemic (5/19, 26%; SMD 0.90, 95% CI 0.15-1.65; P=.02; I^2=93.5%) indicated that mHealth interventions significantly enhanced QoL in patients with CHD compared to usual care, irrespective of the COVID-19 pandemic (Figure S17 in Multimedia Appendix 2).

The Effects on Anxiety and Depression

A significant effect favoring the mHealth intervention on reducing anxiety (SMD −1.16, 95% CI −2.24 to −0.08; P=.04; I^2=95.7%; Figure S18 in Multimedia Appendix 2) and depressive symptoms (SMD −1.01, 95% CI −1.78 to −0.24; P=.01; I^2=92%; Figure S19 in Multimedia Appendix 2) compared to usual care was seen when only studies conducted during the COVID-19 pandemic were included, while this effect was not significant when only studies conducted before the pandemic were included.

Different Country Types: LMICs Versus HICs

The mHealth interventions implemented in LMICs demonstrated greater efficacy in improving QoL (SMD 0.87, 95% CI 0.54-1.19; P<.001; I^2=85.3%; Figure S20 in Multimedia Appendix 2) as well as alleviating anxiety (SMD −0.99, 95% CI −1.75 to −0.22; P=.01; I^2=96.1%; Figure S21 in Multimedia Appendix 2) and depressive symptoms (SMD −0.66, 95% CI −1.23 to −0.08; P=.02; I^2=93.6%; Figure S22 in Multimedia Appendix 2) in patients with CHD compared to usual care, while the mHealth interventions implemented in HICs showed limited effectiveness in enhancing these outcomes.

Different Intervention Contents: Focus on Mental Health

The mHealth interventions that focused on mental health proved significantly more effective in improving QoL (SMD 0.84, 95% CI 0.39-1.28; P<.001; I^2=94.9%; Figure S23 in Multimedia Appendix 2) as well as reducing anxiety (SMD −0.84, 95% CI −1.47 to −0.20; P=.01; I^2=97.1%; Figure S24 in Multimedia Appendix 2) and depression (SMD −0.81, 95% CI −1.40 to −0.22; P=.007; I^2=96.3%; Figure S25 in Multimedia Appendix 2) in patients with CHD compared to usual care. However, there was no significant difference between mHealth interventions that were less relevant to mental health and usual care in improving QoL, anxiety, and depression outcomes in patients with CHD.

Different Levels of Adherence

The Effects on QoL

As depicted in Figure S26 in Multimedia Appendix 2, of the 19 studies, 10 (53%) were included in the lower adherence rates (<90%) group (SMD 0.44, 95% CI 0.16-0.72; P=.002; I^2=83.4%), and 9 (47%) were included in the higher adherence

XSL•FO
RenderX

rates (≥90%) group (SMD 0.53, 95% CI 0.10-0.97; P=.01; I^2=94%). The findings indicated that mHealth interventions significantly enhanced QoL in patients with CHD, regardless of their level of adherence.

The Effects on Anxiety and Depression

Higher adherence rates (≥90%) were associated with superior outcomes in reducing anxiety (SMD −0.63, 95% CI −1.19 to −0.06; P=.03; I^2=96.4%; Figure S27 in Multimedia Appendix 2) and depression (SMD −0.68, 95% CI −1.10 to −0.26; P=.002; I^2=94.5%; Figure S28 in Multimedia Appendix 2) compared to usual care, while this effect was not observed when considering studies with lower adherence rates (<90%) alone.

Discussion

Principal Findings

We performed a meta-analysis of 23 RCTs (with 5406 participants in total) identified after a comprehensive database search to explore the effectiveness of mHealth interventions on QoL, anxiety, and depression in patients with CHD. Our meta-analysis revealed that mHealth interventions significantly enhanced QoL as well as reduced anxiety and depression in patients with CHD undergoing CR.

QoL is a comprehensive measurement index that encompasses various aspects of patient well-being, including physical, psychological, and social factors [56]. The relevant studies showed that there is a correlation between QoL, anxiety, and depression and other emotions of patients with CHD [57]. Furthermore, the incidence of cardiovascular disease has gradually increased in recent years, with high disability and mortality rates, making it the world's leading health care burden and a major contributor to reduced QoL [58]. In our meta-analysis, mHealth interventions significantly enhanced QoL in patients with CHD compared to usual care. The reason may be that the mHealth interventions could provide patients with long-term and personalized health management knowledge, effectively increasing their awareness of the disease, improving their ability to self-manage, and promoting health behavior change [52] (eg, increasing physical activity [59], consuming a healthy diet, quitting smoking [60,61], and maintaining drug adherence [42]), thus improving cardiovascular health and QoL in patients with CHD [62].

In our analysis, we found that the mHealth interventions were effective in relieving anxiety and depressive symptoms in patients with CHD compared to usual care. Epidemiological studies have revealed a bidirectional link between anxiety, depression, and CHD. Anxiety and depression may induce coronary artery restenosis and cause an increase in the number of adverse cardiovascular events in patients with CHD [63]. At the same time, chest pain, chest tightness, fatigue, palpitation, and other symptoms caused by myocardial ischemia make patients prone to anxiety and other psychological problems [64-67]. Thus, helping patients manage stress and psychosocial risk factors is crucial.

The increased effectiveness of mHealth interventions could stem from the fact that patients with anxiety or depression often tend to be withdrawn, exhibiting reluctance to engage with others and express their emotions, and mHealth, as an accessible strategy, may be more conducive to enabling them to talk about their emotions and, as a consequence, alleviate their symptoms. In addition, compared to traditional intervention approaches, mHealth interventions are more convenient, provide timely feedback on patient issues, address patient distress, and improve communication between medical staff and patients so that patients can still receive systematic and continuous care outside the hospital [21]. In the study by Ivanova et al [68], the authors found that interventions delivered via the internet and smartphone apps were effective in reducing the anxiety of participants. Firth et al [69] found that anxiety could be decreased by using smartphone-delivered psychological interventions. Deady et al [70] indicated that a smartphone app could alleviate the symptoms of depression and possibly even avoid cases of incident depression.

Although this meta-analysis observed that mHealth interventions improved QoL as well as reduced anxiety and depression in patients with CHD, considerable heterogeneity remains. Therefore, we conducted subgroup analyses based on the different intervention durations, different intervention methods, the influence of the COVID-19 pandemic, different country types, different intervention contents, and different mHealth intervention adherence rates.

Regarding intervention durations, both intervention duration <6 months and intervention duration ≥6 months substantially enhanced QoL in patients with CHD, but intervention duration ≥6 months was superior to intervention duration <6 months in reducing anxiety and depression in patients with CHD. The reason may be that patients are unfamiliar with mHealth apps or wearable devices when they first start using an mHealth intervention, which may decrease the stickiness of mHealth use and the effect of the intervention. However, with the elapse of intervention time, patients gradually become well-versed in mHealth, realize the convenience and benefits of mHealth, and their compliance improves. Therefore, future research should concentrate on the long-term efficacy of mHealth, emphasizing health behavior maintenance and psychological outcomes.

In addition, a subgroup analysis revealed that a simple mHealth intervention is more effective than a complex mHealth intervention in promoting QoL and reducing anxiety and depression. There might be 2 reasons for this. First, CHD often affects middle-aged and older adults [71]. Older people have decreased cognitive function and memory, less access to the internet and mobile devices, and limited online exposure, and according to a prior study, older persons may have challenges accepting and adapting to complex informatization programs because they are from a different technological generation [72]. Second, complex interventions refer to the use of ≥2 mHealth methods, which are more difficult to use than simple interventions and may decrease the stickiness of mHealth use. Therefore, the application of mHealth interventions should consider the receptivity and memory of middle-aged and older adults, enhance the development of mHealth software and hardware, increase the use of images and videos, enlarge the font size of pages, consider the individual needs of patients, and

XSL•FO
RenderX

motivate patients in an accessible and understandable way to actively participate in the mHealth program [73].

Furthermore, a subgroup analysis was performed according to whether the study was conducted before or during the COVID-19 pandemic and found that, during the pandemic, the use of mHealth was more conducive to improving anxiety and depressive symptoms in patients with CHD. The reasons for this are as follows. First, several surveys conducted during the COVID-19 pandemic showed a higher prevalence of anxiety and depression as well as lower well-being compared to previous estimates [74,75]. Second, the COVID-19 pandemic exerted great pressure on medical resources, and some medical institutions needed to prioritize the treatment of patients with COVID-19 infection, leaving other patients' medical needs unmet. mHealth technology can help patients with CHD to obtain timely medical services and psychological support and reduce anxiety and depression caused by the strain on medical resources [76]. Third, the COVID-19 pandemic has heightened people's awareness of health, and patients are more willing to self-manage and monitor their health using mHealth technology to minimize outdoor exposure. Through self-management, patients can better manage their condition and reduce anxiety and depression. Finally, during the COVID-19 pandemic, the use of mHealth increased, with a wider variety of applications offering more features that better met the personalized needs of patients and helped them relieve anxiety and depression.

Regarding different country types, the results showed that the use of mHealth was more conducive to improving QoL in patients with CHD in LMICs and reducing their anxiety and depressive symptoms. The reason may be that health care resources in LMICs are often scarcer than in HICs, making it difficult for patients to access timely medical care and mental health support. mHealth technology can compensate for the lack of health care resources by allowing patients to access medical information and mental health support through mobile phones or other mobile devices. In addition, mHealth technology can reduce the cost of health care services because patients do not have to pay for extra transportation and do not need to spend time waiting in line. This is particularly important for patients with CHD in LMICs because they often make up a group that is considered economically disadvantaged.

On the basis of our analysis, there are 2 possible reasons why mHealth interventions that focus on mental health could improve QoL and relieve anxiety and depressive symptoms in patients with CHD. First, mHealth interventions focusing on mental health can provide mental health education, psychological support, and other services, which can help relieve psychological problems such as anxiety and depression, thereby improving QoL in patients. Second, mHealth interventions focusing on mental health can help patients better self-monitor and self-manage, including self-monitoring physiological indicators and recording health behavior. This helps patients to better grasp their health status, improve their QoL, and relieve anxiety and depression.

Regarding different levels of adherence, we found that mHealth was more effective in relieving anxiety and depression in patients with CHD with higher adherence rates (≥90%) than those with lower adherence rates (<90%). There are several possible reasons why the adherence rates differed. First, as shown in our results, lower adherence may be caused by low engagement with the mHealth intervention, which is normal and in line with other research on health behavior change [77,78]. Thus, it is strongly advised to enhance the functionality of mHealth tools, such as apps and wearables, to increase adherence to mHealth interventions. This enhancement should focus on meeting the individual needs of patients with CHD and supporting their engagement. Su et al [79] underscore the significance of incorporating patients' preferences in the design of mHealth platforms and highlight the essential role of features such as self-monitoring and self-evaluation in improving user engagement and ensuring the effective use of these tools in real-world settings. Second, financial rewards may increase patient enthusiasm to complete interventions. Therefore, the incentive mechanism is indispensable in the intervention process. In addition, the mHealth intervention duration, intervention frequency, and intervention methods may also affect the compliance of patients.

Safety is a priority for all treatments. Despite the promising results of this study, there are general concerns about patient safety in unsupervised mHealth interventions, particularly with regard to physical activity. Notably, of the 23 RCTs, 12 (52%) addressed adverse safety events, and all 12 reported no intervention-related adverse events or deaths, suggesting a certain level of safety in mHealth interventions. In addition, a systematic review assessed the safety of home-based CR (HBCR) programs (digital or telehealth interventions) and concluded that the risk of adverse events occurring is low, thus reassuring the safety of such interventions [80]. Another review examined the incidence and severity of adverse events associated with HBCR and revealed a low rate of major exercise-related adverse events [81]. Furthermore, a systematic review that explored and evaluated the effectiveness of technology-assisted CR demonstrated that digital therapeutics–based CR is not only safe and feasible for patients with CHD but also enjoys a high level of satisfaction among patients [82].

In terms of the cost-effectiveness of mHealth, several reviews have indicated that mHealth interventions exhibit promising or comparable cost-effectiveness when compared to usual care [80,83-85]. A review conducted by McDonagh et al [80] concluded that the cost per patient for center-based CR and HBCR programs (digital or telehealth interventions) was similar. Another review highlighted the cost-effectiveness of exercise-based telehealth CR and recommended its use to reduce the economic burden of preventive health care and enhance CR accessibility [83]. In addition, Brouwers et al [86] evaluated the cost-effectiveness of cardiac telerehabilitation with relapse prevention compared with center-based CR among patients with coronary artery disease and indicated that, from a societal perspective, telerehabilitation was associated with lower costs compared with center-based CR (mean US $23,405, SE US $3142 vs mean US $27,843, SE US $4126, respectively).

There are 3 factors contributing to the cost-effectiveness of mHealth interventions. First, as demonstrated in an RCT, mHealth interventions could facilitate more frequent patient-specialist interactions, fostering enhanced interpersonal

XSL•FO
RenderX

engagement that could increase patient participation and self-efficacy in health management [87]. Second, mHealth interventions could offer responsive, personalized services, leading to increased patient adherence to CR protocols and enhanced rehabilitation outcomes. Third and last, mHealth has the potential to enhance medical service coverage, particularly benefiting individuals with limited access to CR, such as those residing in remote or rural areas.

Comparison With Prior Work

According to our literature review, there is a substantial body of research exploring the application of mHealth. However, to the best of our knowledge, no comprehensive study has been conducted on the impact of mHealth on QoL, anxiety, and depression outcomes in patients with CHD.

Among the published studies, Rintala et al [88] reviewed 11 studies and showed that mHealth applications are promising for enhancing QoL in stroke rehabilitation. Chin-Jung et al [89] found that mHealth interventions significantly improved life satisfaction and mental health in patients with type 1 diabetes. Dawes et al [90] indicated that postsurgery QoL improvements were expedited by mHealth. Ni et al [91] performed a comprehensive review of 19 RCTs, and the findings revealed that while mHealth reduced post–cardiac surgery depression and improved the physical aspect of QoL, it had no impact on the psychological aspect of QoL and anxiety in patients after cardiac surgery (subgroup analyses of the main outcomes were not performed).

Compared to previous studies, our study focused on QoL, anxiety, and depression in patients with CHD; we performed subgroup analyses based on the different intervention durations, different intervention methods, the influence of the COVID-19 pandemic, different country types, different intervention contents, and different levels of adherence to detect the source of heterogeneity; and we included a larger number of RCTs. Moreover, we followed the PRISMA guidelines for this meta-analysis to ensure that our methodology was robust and reliable [92].

Limitations

While we have attempted to perform a comprehensive meta-analysis of the effects of mHealth on QoL, anxiety, and depression in patients with CHD, some limitations need to be considered. First, some studies have methodological flaws (eg, allocation sequence concealment and a lack of blinding) that may influence the accuracy of this meta-analysis. Hence, the findings should be taken with caution, and future researchers are advised to carefully execute randomized controlled designs and provide more data following reporting criteria to ensure the study's quality. Second, although we considered the COVID-19 pandemic an important confounding factor and conducted subgroup analyses to address the influence of the pandemic on studies, we did not include a postpandemic subgroup for analysis. That is because, based on our database search period for this meta-analysis, we excluded studies published after August 12, 2023, by which time the COVID-19 pandemic was no longer considered a major public health event, and no RCT studies meeting the inclusion and exclusion criteria of this study had been retrieved.

Third, although we carried out subgroup analyses to determine effects within categories, the social determinants of health such as education level, income status, access to mHealth technologies, marital status, and so on, were difficult to consider because most of the included studies (19/23, 83%) did not report detailed information on the social determinants of health. Fourth, although mHealth interventions were used in all intervention groups, the wide range of intervention types, from SMS text messaging and mobile apps to wearables, may significantly contribute to the diversity in outcomes among the studies. Furthermore, because many studies (7/23, 30%) used >1 mHealth intervention, we only conducted a subgroup analysis for simple versus complex mHealth interventions, rather than different types of mHealth interventions specifically. Fifth, most of the studies (20/23, 87%) included in our meta-analysis only reported that the enrolled patients were patients with CHD; the specific subtype of CHD was not clarified. Consequently, while our findings indicate a positive effect of mHealth interventions on QoL and mental health (anxiety and depression) in patients with CHD, we are unable to determine the effectiveness across different CHD subtypes conclusively. Sixth and last, the sample sizes varied greatly among the included studies. Further research should focus on increasing the sample sizes and strengthening the randomization, allocation, and blinding methodologies to improve the quality of evidence.

Conclusions

mHealth interventions effectively improved QoL, anxiety, and depression in patients with CHD. Specifically, mHealth interventions effectively alleviated anxiety and depression among patients with CHD when (1) the intervention duration was ≥6 months, (2) the mHealth intervention was a simple one (only 1 mHealth intervention was used), (3) it was implemented during the COVID-19 pandemic, (4) it was implemented in LMICs, (5) the intervention focused on mental health, and (6) adherence rates were high (≥90%). Furthermore, a notable impact favoring mHealth interventions in enhancing QoL compared to usual care was evident in scenarios in which (1) the mHealth intervention was a simple one, (2) it was implemented in LMICs, and (3) the intervention focused on mental health. However, because most of the studies (21/23, 91%) we reviewed were of moderate quality, we must proceed with caution in applying the evidence. Large sample, high-quality, and rigorously designed RCTs are needed to provide more evidence.

Acknowledgments

This study is supported by a project grant from the National Natural Science Foundation of China (72034005).

XSL•FO
RenderX

Authors' Contributions

QLH and LYL collaborated on the database search, literature review, data extraction, and data analysis, and drafted the manuscript. YW provided guidance for the data analysis and writing of the manuscript, provided ideas and suggestions for revising and editing the manuscript, and oversaw the quality and validity of the manuscript. All authors discussed and helped to complete the final manuscript, and read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 84 KB - jmir_v26i1e52341_app1.pdf]

Multimedia Appendix 2
Sensitivity analyses, subgroup analyses results, and the search strategies used for different databases.
[PDF File (Adobe PDF File), 1669 KB - jmir_v26i1e52341_app2.pdf]

References

1. Cardiovascular diseases. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2021-12-12]
2. The top 10 causes of death. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2023-06-10]
3. Visseren FL, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J 2021 Sep 07;42(34):3227-3337. [doi: 10.1093/eurheartj/ehab484] [Medline: 34458905]
4. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
5. Wang D, Dai F, Liu W, Ge L, Qi X, Ma G. Longitudinal change and prognostic value of anxiety and depression in coronary heart disease patients. Ir J Med Sci 2021 Feb;190(1):107-116. [doi: 10.1007/s11845-020-02302-7] [Medline: 32661829]
6. Krittanawong C, Maitra NS, Qadeer YK, Wang Z, Fogg S, Storch EA, et al. Association of depression and cardiovascular disease. Am J Med 2023 Sep;136(9):881-895. [doi: 10.1016/j.amjmed.2023.04.036] [Medline: 37247751]
7. Celano CM, Huffman JC. Depression and cardiac disease: a review. Cardiol Rev 2011;19(3):130-142. [doi: 10.1097/CRD.0b013e31820e8106] [Medline: 21464641]
8. Thombs BD, Bass EB, Ford DE, Stewart KJ, Tsilidis KK, Patel U, et al. Prevalence of depression in survivors of acute myocardial infarction. J Gen Intern Med 2006 Jan;21(1):30-38 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.00269.x] [Medline: 16423120]
9. Cosh SM, Pinto R, Denson L, Tully PJ. Understandings and experiences of adherence to secondary prevention for patients with cardiovascular disease and comorbid depression or anxiety. Psychol Health Med 2023 Jul;28(6):1479-1486. [doi: 10.1080/13548506.2022.2060515] [Medline: 35356823]
10. Le J, Dorstyn DS, Mpofu E, Prior E, Tully PJ. Health-related quality of life in coronary heart disease: a systematic review and meta-analysis mapped against the international classification of functioning, disability and health. Qual Life Res 2018 Oct;27(10):2491-2503. [doi: 10.1007/s11136-018-1885-5] [Medline: 29779143]
11. Xia K, Wang LF, Yang XC, Jiang HY, Zhang LJ, Yao DK, et al. Comparing the effects of depression, anxiety, and comorbidity on quality-of-life, adverse outcomes, and medical expenditure in Chinese patients with acute coronary syndrome. Chin Med J (Engl) 2019 May 05;132(9):1045-1052 [FREE Full text] [doi: 10.1097/CM9.0000000000000215] [Medline: 30896567]
12. Virani SS, Newby LK, Arnold SV, Bittner V, Brewer LC, Demeter SH, Peer Review Committee Members. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA guideline for the management of patients with chronic coronary disease: a report of the American Heart Association/American College of Cardiology joint committee on clinical practice guidelines. Circulation 2023 Aug 29;148(9):e9-119 [FREE Full text] [doi: 10.1161/CIR.0000000000001168] [Medline: 37471501]
13. Barber S, Thornicroft G. Reducing the mortality gap in people with severe mental disorders: the role of lifestyle psychosocial interventions. Front Psychiatry 2018;9:463 [FREE Full text] [doi: 10.3389/fpsyt.2018.00463] [Medline: 30323773]
14. Palmer VJ, Lewis M, Stylianopolous V, Furler J. Primary care prevention of the cardiovascular health crisis for people with severe mental illnesses: the elephant in the room. Aust J Gen Pract 2018 Dec;47(12):846-850 [FREE Full text] [doi: 10.31128/AJGP-09-18-4692] [Medline: 31212402]

XSL•FO
RenderX

15. Osborn D, Burton A, Walters K, Atkins L, Barnes T, Blackburn R, et al. Primary care management of cardiovascular risk for people with severe mental illnesses: the Primrose research programme including cluster RCT. Programme Grants Appl Res 2019;7(2):1 [FREE Full text] [doi: 10.3310/pgfar07020]
16. Xu A, Wang Y, Wu X. Effectiveness of e-health based self-management to improve cancer-related fatigue, self-efficacy and quality of life in cancer patients: systematic review and meta-analysis. J Adv Nurs 2019 Dec;75(12):3434-3447. [doi: 10.1111/jan.14197] [Medline: 31566769]
17. Zhao M, Graham I, Cooney MT, Grobbee DE, Vaartjes I, Klipstein-Grobusch K. Determinants of coronary artery disease risk factor management across three world regions. Heart Asia 2019;11(1):e011112 [FREE Full text] [doi: 10.1136/heartasia-2018-011112] [Medline: 31031827]
18. Rautio N, Jokelainen J, Pölönen A, Oksa H, Peltonen M, Vanhala M, et al. Changes in lifestyle modestly reduce the estimated cardiovascular disease risk in one-year follow-up of the Finnish diabetes prevention program (FIN-D2D). Eur J Cardiovasc Nurs 2015 Apr;14(2):145-152. [doi: 10.1177/1474515114521713] [Medline: 24452452]
19. Kim JM, Stewart R, Lee YS, Lee HJ, Kim MC, Kim JW, et al. Effect of escitalopram vs placebo treatment for depression on long-term cardiac outcomes in patients with acute coronary syndrome: a randomized clinical trial. JAMA 2018 Jul 24;320(4):350-358 [FREE Full text] [doi: 10.1001/jama.2018.9422] [Medline: 30043065]
20. Honkola J, Hookana E, Malinen S, Kaikkonen KS, Junttila MJ, Isohanni M, et al. Psychotropic medications and the risk of sudden cardiac death during an acute coronary event. Eur Heart J 2012 Mar;33(6):745-751. [doi: 10.1093/eurheartj/ehr368] [Medline: 21920969]
21. Park YT. Emerging new era of mobile health technologies. Healthc Inform Res 2016 Oct;22(4):253-254 [FREE Full text] [doi: 10.4258/hir.2016.22.4.253] [Medline: 27895955]
22. mHealth: new horizons for health through mobile technologies. World Health Organization. URL: https://www.mobileworldlive.com/mhealth-new-horizons-forhealth-through-mobile-technologies [accessed 2023-06-01]
23. Ganesan AN, Louise J, Horsfall M, Bilsborough SA, Hendriks J, McGavigan AD, et al. International mobile-health intervention on physical activity, sitting, and weight: the Stepathlon cardiovascular health study. J Am Coll Cardiol 2016 May 31;67(21):2453-2463 [FREE Full text] [doi: 10.1016/j.jacc.2016.03.472] [Medline: 27050185]
24. Yudi MB, Clark DJ, Tsang D, Jelinek M, Kalten K, Joshi S, et al. SMARTphone-based, early cardiac REHABilitation in patients with acute coronary syndromes: a randomized controlled trial. Coron Artery Dis 2021 Aug 01;32(5):432-440. [doi: 10.1097/MCA.0000000000000938] [Medline: 32868661]
25. Houchen-Wolloff L, Gardiner N, Devi R, Robertson N, Jolly K, Marshall T, et al. Web-based cardiac RE habilitatio N alternative for those declining or dropping out of conventional rehabilitation: results of the WREN feasibility randomised controlled trial. Open Heart 2018;5(2):e000860 [FREE Full text] [doi: 10.1136/openhrt-2018-000860] [Medline: 30364405]
26. Dalli Peydró E, Sanz Sevilla N, Tuzón Segarra MT, Miró Palau V, Sánchez Torrijos J, Cosín Sales J. A randomized controlled clinical trial of cardiac telerehabilitation with a prolonged mobile care monitoring strategy after an acute coronary syndrome. Clin Cardiol 2022 Jan 24;45(1):31-41 [FREE Full text] [doi: 10.1002/clc.23757] [Medline: 34952989]
27. Kang G, Zhang H, Zhou J, Wan D. The WeChat platform-based health education improves health outcomes among patients with stable coronary artery disease. Patient Educ Couns 2023 Jun;111:107704. [doi: 10.1016/j.pec.2023.107704] [Medline: 36906932]
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
29. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336-341 [FREE Full text] [doi: 10.1016/j.ijsu.2010.02.007] [Medline: 20171303]
30. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
31. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3. Hoboken, NJ: Wiley-Blackwell; 2022.
32. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
33. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun;56(2):455-463. [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
34. Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. J R Stat Soc Ser C Appl Stat 2020 Nov 28;69(5):1091-1119 [FREE Full text] [doi: 10.1111/rssc.12440] [Medline: 33132447]
35. Batalik L, Dosbaba F, Hartman M, Batalikova K, Spinar J. Benefits and effectiveness of using a wrist heart rate monitor as a telerehabilitation device in cardiac patients: a randomized controlled trial. Medicine (Baltimore) 2020 Mar;99(11):e19556 [FREE Full text] [doi: 10.1097/MD.0000000000019556] [Medline: 32176113]
36. Chow CK, Klimis H, Thiagalingam A, Redfern J, Hillis GS, Brieger D, TEXTMEDS Investigators*. Text messages to improve medication adherence and secondary prevention after acute coronary syndrome: the TEXTMEDS randomized

XSL•FO
RenderX

clinical trial. Circulation 2022 May 10;145(19):1443-1455. [doi: 10.1161/CIRCULATIONAHA.121.056161] [Medline: 35533220]
37. Dorje T, Zhao G, Tso K, Wang J, Chen Y, Tsokey L, et al. Smartphone and social media-based cardiac rehabilitation and secondary prevention in China (SMART-CR/SP): a parallel-group, single-blind, randomised controlled trial. Lancet Digit Health 2019 Nov;1(7):e363-e374 [FREE Full text] [doi: 10.1016/S2589-7500(19)30151-7] [Medline: 33323210]
38. Duan YP, Liang W, Guo L, Wienert J, Si GY, Lippke S. Evaluation of a web-based intervention for multiple health behavior changes in patients with coronary heart disease in home-based rehabilitation: pilot randomized controlled trial. J Med Internet Res 2018 Nov 19;20(11):e12052 [FREE Full text] [doi: 10.2196/12052] [Medline: 30455167]
39. Fang J, Huang B, Xu D, Li J, Au WW. Innovative application of a home-based and remote sensing cardiac rehabilitation protocol in Chinese patients after percutaneous coronary intervention. Telemed J E Health 2019 Apr;25(4):288-293. [doi: 10.1089/tmj.2018.0064] [Medline: 30192210]
40. Hisam A, Haq ZU, Aziz S, Doherty P, Pell J. Effectiveness of mobile health augmented cardiac rehabilitation (MCard) on health-related quality of life among post-acute coronary syndrome patients: a randomized controlled trial. Pak J Med Sci 2022 Jan 14;38(3Part-I):716-723 [FREE Full text] [doi: 10.12669/pjms.38.3.4724] [Medline: 35480536]
41. Huang ST, Xu N, Chen LW, Chen Q. Application of the WeChat platform to implement continuous nursing for patients after percutaneous coronary intervention. J Nurse Train 2017;32(7):589-592 [FREE Full text]
42. Johnston N, Bodegard J, Jerström S, Åkesson J, Brorsson H, Alfredsson J, et al. Effects of interactive patient smartphone support app on drug adherence and lifestyle changes in myocardial infarction patients: a randomized study. Am Heart J 2016 Aug;178:85-94 [FREE Full text] [doi: 10.1016/j.ahj.2016.05.005] [Medline: 27502855]
43. Kang SH, Baek H, Cho J, Kim S, Hwang H, Lee W, et al. Management of cardiovascular disease using an mHealth tool: a randomized clinical trial. NPJ Digit Med 2021 Dec 03;4(1):165 [FREE Full text] [doi: 10.1038/s41746-021-00535-z] [Medline: 34862449]
44. Kraal JJ, Van den Akker-Van Marle ME, Abu-Hanna A, Stut W, Peek N, Kemps HM. Clinical and cost-effectiveness of home-based cardiac rehabilitation compared to conventional, centre-based cardiac rehabilitation: results of the FIT@Home study. Eur J Prev Cardiol 2017 Aug;24(12):1260-1273 [FREE Full text] [doi: 10.1177/2047487317710803] [Medline: 28534417]
45. Pakrad F, Ahmadi F, Grace SL, Oshvandi K, Kazemnejad A. Traditional vs extended hybrid cardiac rehabilitation based on the continuous care model for patients who have undergone coronary artery bypass surgery in a middle-income country: a randomized controlled trial. Arch Phys Med Rehabil 2021 Nov;102(11):2091-101.e3. [doi: 10.1016/j.apmr.2021.04.026] [Medline: 34175270]
46. Shariful Islam SM, Chow CK, Redfern J, Kok C, Rådholm K, Stepien S, et al. Effect of text messaging on depression in patients with coronary heart disease: a substudy analysis from the TEXT ME randomised controlled trial. BMJ Open 2019 Feb 20;9(2):e022637 [FREE Full text] [doi: 10.1136/bmjopen-2018-022637] [Medline: 30787075]
47. Snoek JA, Meindersma EP, Prins LF, Van't Hof AW, de Boer M, Hopman MT, et al. The sustained effects of extending cardiac rehabilitation with a six-month telemonitoring and telecoaching programme on fitness, quality of life, cardiovascular risk factors and care utilisation in CAD patients: the TeleCaRe study. J Telemed Telecare 2021 Sep;27(8):473-483. [doi: 10.1177/1357633X19885793] [Medline: 31760855]
48. Su JJ, Yu DS. Effects of a nurse-led eHealth cardiac rehabilitation programme on health outcomes of patients with coronary heart disease: a randomised controlled trial. Int J Nurs Stud 2021 Oct;122:104040. [doi: 10.1016/j.ijnurstu.2021.104040] [Medline: 34333211]
49. Wang WJ. Development of a mobile health application program to support health management for patients underwent percutaneous coronary intervention and its clinical pilot testing. Anhui China: Anhui Medical University. 2018. URL: https://xueshu.baidu.com/usercenter/paper/show?paperid=1q4h0a50vc0s0ru03h2v0r40q3755061&site=xueshu_se&hitarticle=1 [accessed 2024-04-29]
50. Zhang WQ, Song PS, Su YX. To explore the application of transitional care based on mobile health software in patients after percutaneous coronary intervention (PCI). J MuDanJiang Med Univ 2021;41(04):70-79. [doi: 10.13799/j.cnki.mdjyxyxb.2020.04.040]
51. Zheng X, Hu S, Jing S. Effect of telephone follow-up based continuous nursing on risk factor control and prognosis quality in patients with coronary heart disease after PCI. Int J Nurs 2021 Aug 8 [FREE Full text] [doi: 10.3760/cma.j.cn221370-20191106-00814]
52. Pfaeffli Dale L, Whittaker R, Jiang Y, Stewart R, Rolleston A, Maddison R. Text message and internet support for coronary heart disease self-management: results from the Text4Heart randomized controlled trial. J Med Internet Res 2015 Oct 21;17(10):e237 [FREE Full text] [doi: 10.2196/jmir.4944] [Medline: 26490012]
53. Cheung NW, Redfern J, Thiagalingam A, Hng T, Marschner S, Haider R, SupportMe Investigators. Effect of mobile phone text messaging self-management support for patients with diabetes or coronary heart disease in a chronic disease management program (SupportMe) on blood pressure: pragmatic randomized controlled trial. J Med Internet Res 2023 Jun 16;25:e38275 [FREE Full text] [doi: 10.2196/38275] [Medline: 37327024]

XSL•FO
RenderX

54. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
55. Widmer RJ, Allison TG, Lennon R, Lopez-Jimenez F, Lerman LO, Lerman A. Digital health intervention during cardiac rehabilitation: a randomized controlled trial. Am Heart J 2017 Jun;188:65-72. [doi: 10.1016/j.ahj.2017.02.016] [Medline: 28577682]
56. Rahkola-Soisalo P, Brummer T, Jalkanen J, Sjöberg J, Sintonen H, Roine RP, et al. Hysterectomy provides benefit in health-related quality of life: a 10-year follow-up study. J Minim Invasive Gynecol 2020;27(4):868-874. [doi: 10.1016/j.jmig.2019.08.009] [Medline: 31404711]
57. Tang L, Wan Y, Wang Y, Zhang F, Jiao J, Shen A. Research status of influencing factors of quality of life in cancer patients. J Nurse Contin Educ 2019;34(11):985-987 [FREE Full text] [doi: 10.16821/j.cnki.hsjx.2019.11.007]
58. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
59. Cyriac J, Jenkins S, Patten CA, Hayes SN, Jones C, Cooper LA, et al. Improvements in diet and physical activity-related psychosocial factors among African Americans using a mobile health lifestyle intervention to promote cardiovascular health: the FAITH! (fostering African American improvement in total health) app pilot study. JMIR Mhealth Uhealth 2021 Nov 12;9(11):e28024 [FREE Full text] [doi: 10.2196/28024] [Medline: 34766917]
60. Klimis H, Marschner S, Von Huben A, Thiagalingam A, Chow CK. Predictors of smoking cessation in a lifestyle-focused text-message support programme delivered to people with coronary heart disease: an analysis from the tobacco exercise and diet messages (TEXTME) randomised clinical trial. Tob Use Insights 2020 Jan 28;13:1179173X20901486 [FREE Full text] [doi: 10.1177/1179173X20901486] [Medline: 32063724]
61. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
62. Garcia-Lunar I, van der Ploeg HP, Fernández Alvira JM, van Nassau F, Castellano Vázquez JM, van der Beek AJ, et al. Effects of a comprehensive lifestyle intervention on cardiovascular health: the TANSNIP-PESA trial. Eur Heart J 2022 Oct 11;43(38):3732-3745 [FREE Full text] [doi: 10.1093/eurheartj/ehac378] [Medline: 35869885]
63. Peter RS, Meyer ML, Mons U, Schöttker B, Keller F, Schmucker R, et al. Long-term trajectories of anxiety and depression in patients with stable coronary heart disease and risk of subsequent cardiovascular events. Depress Anxiety 2020 Aug;37(8):784-792. [doi: 10.1002/da.23011] [Medline: 32237189]
64. Richards SH, Anderson L, Jenkinson CE, Whalley B, Rees K, Davies P, et al. Psychological interventions for coronary heart disease: Cochrane systematic review and meta-analysis. Eur J Prev Cardiol 2018 Feb;25(3):247-259 [FREE Full text] [doi: 10.1177/2047487317739978] [Medline: 29212370]
65. Liu Y, Zhao Y, Tian J, Tong T, Gao R, Liu Y. The association of depression following percutanous coronary intervention with adverse cardiovascular events: protocol for a systematic review and meta-analysis. Medicine (Baltimore) 2019 Jan;98(2):e13952 [FREE Full text] [doi: 10.1097/MD.0000000000013952] [Medline: 30633173]
66. Carney RM, Freedland KE. Depression and coronary heart disease. Nat Rev Cardiol 2017 Mar;14(3):145-155. [doi: 10.1038/nrcardio.2016.181] [Medline: 27853162]
67. Lichtman JH, Froelicher ES, Blumenthal JA, Carney RM, Doering LV, Frasure-Smith N, et al. Depression as a risk factor for poor prognosis among patients with acute coronary syndrome: systematic review and recommendations: a scientific statement from the American Heart Association. Circulation 2014 Mar 25;129(12):1350-1369. [doi: 10.1161/CIR.0000000000000019] [Medline: 24566200]
68. Ivanova E, Lindner P, Ly KH, Dahlin M, Vernmark K, Andersson G, et al. Guided and unguided acceptance and commitment therapy for social anxiety disorder and/or panic disorder provided via the internet and a smartphone application: a randomized controlled trial. J Anxiety Disord 2016 Dec;44:27-35. [doi: 10.1016/j.janxdis.2016.09.012] [Medline: 27721123]
69. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
70. Deady M, Glozier N, Calvo R, Johnston D, Mackinnon A, Milne D, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
71. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
72. van Boekel LC, Peek ST, Luijkx KG. Diversity in older adults' use of the internet: identifying subgroups through latent class analysis. J Med Internet Res 2017 May 24;19(5):e180 [FREE Full text] [doi: 10.2196/jmir.6853] [Medline: 28539302]

XSL•FO
RenderX

73. Chen Y, Ji M, Wu Y, Deng Y, Wu F, Lu Y. Individualized mobile health interventions for cardiovascular event prevention in patients with coronary heart disease: study protocol for the iCARE randomized controlled trial. BMC Cardiovasc Disord 2021 Jul 13;21(1):340 [FREE Full text] [doi: 10.1186/s12872-021-02153-9] [Medline: 34256698]
74. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
75. Bueno-Notivol J, Gracia-García P, Olaya B, Lasheras I, López-Antón R, Santabárbara J. Prevalence of depression during the COVID-19 outbreak: a meta-analysis of community-based studies. Int J Clin Health Psychol 2021;21(1):100196 [FREE Full text] [doi: 10.1016/j.ijchp.2020.07.007] [Medline: 32904715]
76. Takahashi EA, Schwamm LH, Adeoye OM, Alabi O, Jahangir E, Misra S, American Heart Association Council on Cardiovascular Radiology and Intervention, Council on Hypertension, Council on the Kidney in Cardiovascular Disease,Stroke Council. An overview of telehealth in the management of cardiovascular disease: a scientific statement from the American Heart Association. Circulation 2022 Dec 20;146(25):e558-e568. [doi: 10.1161/CIR.0000000000001107] [Medline: 36373541]
77. Hershman SG, Bot BM, Shcherbina A, Doerr M, Moayedi Y, Pavlovic A, et al. Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart Counts Study. Sci Data 2019 Apr 11;6(1):24. [doi: 10.1038/s41597-019-0016-7] [Medline: 30975992]
78. Ellingson LD, Lansing JE, DeShaw KJ, Peyer KL, Bai Y, Perez M, et al. Evaluating motivational interviewing and habit formation to enhance the effect of activity trackers on healthy adults' activity levels: randomized intervention. JMIR Mhealth Uhealth 2019 Feb 14;7(2):e10988 [FREE Full text] [doi: 10.2196/10988] [Medline: 30762582]
79. Su JJ, Paguio J, Baratedi WM, Abu-Odah H, Batalik L. Experience of coronary heart disease patients with a nurse-led eHealth cardiac rehabilitation: qualitative process evaluation of a randomized controlled trial. Heart Lung 2023;57:214-221. [doi: 10.1016/j.hrtlng.2022.10.005] [Medline: 36265371]
80. McDonagh ST, Dalal H, Moore S, Clark CE, Dean SG, Jolly K, et al. Home-based versus centre-based cardiac rehabilitation. Cochrane Database Syst Rev 2023 Oct 27;10(10):CD007130. [doi: 10.1002/14651858.CD007130.pub5] [Medline: 37888805]
81. Stefanakis M, Batalik L, Antoniou V, Pepera G. Safety of home-based cardiac rehabilitation: a systematic review. Heart Lung 2022;55:117-126 [FREE Full text] [doi: 10.1016/j.hrtlng.2022.04.016] [Medline: 35533492]
82. Chong MS, Sit JW, Karthikesu K, Chair SY. Effectiveness of technology-assisted cardiac rehabilitation: a systematic review and meta-analysis. Int J Nurs Stud 2021 Dec;124:104087. [doi: 10.1016/j.ijnurstu.2021.104087] [Medline: 34562846]
83. Batalik L, Filakova K, Sladeckova M, Dosbaba F, Su J, Pepera G. The cost-effectiveness of exercise-based cardiac telerehabilitation intervention: a systematic review. Eur J Phys Rehabil Med 2023 Apr;59(2):248-258 [FREE Full text] [doi: 10.23736/S1973-9087.23.07773-0] [Medline: 36692413]
84. Buckingham SA, Taylor RS, Jolly K, Zawada A, Dean SG, Cowie A, et al. Home-based versus centre-based cardiac rehabilitation: abridged Cochrane systematic review and meta-analysis. Open Heart 2016;3(2):e000463 [FREE Full text] [doi: 10.1136/openhrt-2016-000463] [Medline: 27738516]
85. Subedi N, Rawstorn JC, Gao L, Koorts H, Maddison R. Implementation of telerehabilitation interventions for the self-management of cardiovascular disease: systematic review. JMIR Mhealth Uhealth 2020 Nov 27;8(11):e17957 [FREE Full text] [doi: 10.2196/17957] [Medline: 33245286]
86. Brouwers RW, van der Poort EK, Kemps HM, van den Akker-van Marle ME, Kraal JJ. Cost-effectiveness of cardiac telerehabilitation with relapse prevention for the treatment of patients with coronary artery disease in the Netherlands. JAMA Netw Open 2021 Dec 01;4(12):e2136652 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.36652] [Medline: 34854907]
87. Maddison R, Rawstorn JC, Stewart RA, Benatar J, Whittaker R, Rolleston A, et al. Effects and costs of real-time cardiac telerehabilitation: randomised controlled non-inferiority trial. Heart 2019 Jan;105(2):122-129 [FREE Full text] [doi: 10.1136/heartjnl-2018-313189] [Medline: 30150328]
88. Rintala A, Kossi O, Bonnechère B, Evers L, Printemps E, Feys P. Mobile health applications for improving physical function, physical activity, and quality of life in stroke survivors: a systematic review. Disabil Rehabil 2023 Dec 02;45(24):4001-4015. [doi: 10.1080/09638288.2022.2140844] [Medline: 36325613]
89. Chin-Jung L, Hsiao-Yean C, Yeu-Hui C, Kuan-Chia L, Hui-Chuan H. Effects of mobile health interventions on improving glycemic stability and quality of life in patients with type 1 diabetes: a meta-analysis. Res Nurs Health 2021 Feb;44(1):187-200. [doi: 10.1002/nur.22094] [Medline: 33368403]
90. Dawes AJ, Lin AY, Varghese C, Russell MM, Lin AY. Mobile health technology for remote home monitoring after surgery: a meta-analysis. Br J Surg 2021 Nov 11;108(11):1304-1314. [doi: 10.1093/bjs/znab323] [Medline: 34661649]
91. Ni R, Liu M, Huang S, Yang J. Effects of eHealth interventions on quality of life and psychological outcomes in cardiac surgery patients: systematic review and meta-analysis. J Med Internet Res 2022 Aug 16;24(8):e40090 [FREE Full text] [doi: 10.2196/40090] [Medline: 35972792]
92. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

Abbreviations

CHD: coronary heart disease
CR: cardiac rehabilitation
HBCR: home-based cardiac rehabilitation
HIC: high-income country
LMIC: low- or middle-income country
mHealth: mobile health
PICOS: population, intervention, comparison, outcomes, and study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QoL: quality of life
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by T de Azevedo Cardoso; submitted 31.08.23; peer-reviewed by J Su, L Batalik; comments to author 12.12.23; revised version received 05.02.24; accepted 22.03.24; published 11.06.24.

Please cite as:
Hou QL, Liu LY, Wu Y
The Effects of mHealth Interventions on Quality of Life, Anxiety, and Depression in Patients With Coronary Heart Disease: Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2024;26:e52341
URL: https://www.jmir.org/2024/1/e52341
doi:10.2196/52341
PMID:38861710

©Qiao Ling Hou, Le Yang Liu, Ying Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Intensive Longitudinal Methods Among Adults With Breast or Lung Cancer: Scoping Review

Joran Geeraerts[1], MSc; Kim de Nooijer[1], MSc, PhD; Lara Pivodic[1], MSc, PhD; Mark De Ridder[2], MSc, PhD, MD; Lieve Van den Block[1], MSc, PhD

[1]End-of-Life Care Research Group, Vrije Universiteit Brussel, Brussels, Belgium

[2]Department of Radiotherapy, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel, Brussels, Belgium

Corresponding Author:
Joran Geeraerts, MSc
End-of-Life Care Research Group
Vrije Universiteit Brussel
Laarbeeklaan 103
Brussels, 1090
Belgium
Phone: 32 2 477 47 56
Email: joran.geeraerts@vub.be

Abstract

Background: Intensive longitudinal methods offer a powerful tool for capturing daily experiences of individuals. However, its feasibility, effectiveness, and optimal methodological approaches for studying or monitoring experiences of oncology patients remain uncertain.

Objective: This scoping review aims to describe to what extent intensive longitudinal methods with daily electronic assessments have been used among patients with breast or lung cancer and with which methodologies, associated outcomes, and influencing factors.

Methods: We searched the electronic databases (PubMed, Embase, and PsycINFO) up to January 2024 and included studies reporting on the use of these methods among adults with breast or lung cancer. Data were extracted on population characteristics, intensive monitoring methodologies used, study findings, and factors influencing the implementation of these methods in research and clinical practice.

Results: We identified 1311 articles and included 52 articles reporting on 41 studies. Study aims and intensive monitoring methodologies varied widely, but most studies focused on measuring physical and psychological symptom constructs, such as pain, anxiety, or depression. Compliance and attrition rates seemed acceptable for most studies, although complete methodological reporting was often lacking. Few studies specifically examined these methods among patients with advanced cancer. Factors influencing implementation were linked to both patient (eg, confidence with intensive monitoring system) and methodology (eg, option to use personal devices).

Conclusions: Intensive longitudinal methods with daily electronic assessments hold promise to provide unique insights into the daily lives of patients with cancer. Intensive longitudinal methods may be feasible among people with breast or lung cancer. Our findings encourage further research to determine optimal conditions for intensive monitoring, specifically in more advanced disease stages.

(J Med Internet Res 2024;26:e50224) doi:10.2196/50224

KEYWORDS

diary; ecological momentary assessment; neoplasms; quality of life; self-report; telemedicine; scoping review; longitudinal methods; breast cancer; lung cancer; patients with cancer; cancer; intensive monitoring; advanced disease stages; mobile phone

XSL•FO
RenderX

Introduction

Background

People diagnosed with cancer, among which breast and lung cancer are the most prevalent diagnoses globally [1], often experience various problems and concerns that affect their quality of life and well-being across physical, psychological, social, and spiritual domains [2-6]. Understanding the fluctuations, interactions, and contextual variations of the multidimensional problems and concerns in patients' daily lives is crucial to gain a comprehensive view of these patients' quality of life and to optimize patient-centered care. Such insights could lead to, among others, improvements in drug schedules and personalized treatment decision-making [7] and the identification of novel care intervention targets by identifying contexts or states that aggravate or buffer against certain problems and concerns [8].

An effective way to gather insights into the daily and within-day variability of patients' quality of life and well-being is the use of intensive longitudinal methods. Bolger and Laurenceau [9] defined intensive longitudinal methods as "an umbrella term to encompass data collection methods that employ enough repeated measurements to model a change process for each subject." The authors specify a minimum number of 5 sequential assessments, as it enables the estimation of linear models within each participant [9]. Examples of such methods are daily diaries and ecological momentary assessments (EMAs), also known as experience sampling methods (ESM). While predominantly developed in psychological research, these methods recently gained more attention in other fields and clinical practice, including oncology, due to advancements in handheld computer technologies that enable easier implementation than traditional pencil-and-paper approaches [9-13]. Despite easier implementation of these methods, researchers and clinicians in the field of oncology still lack a clear understanding of available options for intensive longitudinal monitoring, their opportunities, pitfalls, and feasibility in populations experiencing high symptom burden. This underscores the need for a structured overview of the use and capabilities of these methods.

Currently, no systematically conducted literature review exists on the use of intensive longitudinal methods in monitoring people with cancer. One systematic review [14] provided the most recent overview of the use of EMA in people with cancer across 42 studies (23 and 8 studies included people with breast and lung cancer, respectively) and found considerable heterogeneity in the methodologies used. However, due to its inclusion criteria focusing solely on EMAs, a large group of studies monitoring patients on a once-daily basis was left out [14]. Furthermore, the review did not report on the barriers and facilitators that were encountered during the implementation of ESM, which is crucial information for optimal use in practice [14].

Objective

We aimed to describe to what extent intensive longitudinal methods with daily electronic assessments have been used among patients with breast or lung cancer, along with the methodologies used, associated outcomes, and influencing factors. We limited the scope of this review to these patient groups with the most prevalent cancer diagnoses for feasibility reasons to provide a more nuanced picture for these methods among these groups and to inform our own ongoing ESM project among these patient groups [15]. More specifically, we described (1) the characteristics of the populations with breast or lung cancer among whom intensive longitudinal methods with daily electronic assessments have been used; (2) the objectives, design, and methods used; (3) the results obtained (including study findings and response-related results); and (4) the identified barriers and facilitators for implementing these methods in clinical and research practice.

Methods

Overview

We conducted a scoping review using a systematic search strategy to gain insight into the extent, range, and nature of current evidence on the use of intensive longitudinal methods with daily electronic assessments in people with breast or lung cancer, rather than providing evidence for a specific research question as in systematic reviews [16,17]. This manuscript adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [18].

Eligibility Criteria

We included articles that met the following criteria: articles that (1) performed in people diagnosed with breast or lung cancer through self-report or proxy responding; (2) included people aged ≥18 years; (3) used *active* intensive longitudinal methods, meaning the conscious reporting of experiences rather than passive data collection through wearables without conscious participant involvement [12]; (4) collected self-reports using electronic devices or allowed participants to choose between electronic and pen-and-paper self-reports, resulting in a partial sample that opted for electronic assessments; (5) applied a measurement period of >24 hours, with ≥5 planned assessments, including at least 1 assessment per day; and (6) included original full-text articles in English, Dutch, or French.

Articles were excluded if they met one or both of the following criteria: articles that (1) were conducted in people in complete cancer remission and (2) concerned reviews, meta-analyses, notes, letters to editors, conference abstracts, or study protocols.

Search Strategy

The initial literature search was conducted on April 7, 2022, and updated on January 19, 2024, both without restrictions for its time coverage. We searched 3 databases: PubMed, Embase, and PsycINFO. We consulted a librarian of the Vrije Universiteit Brussel for the development of the search strategy. Keywords included terms related to the population (eg, *cancer*) and methodology (eg, *ecological momentary assessment* and *daily diary*). The search strategy was validated in PubMed and translated to other databases. The full search strategy is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Study Selection

Figure 1 provides an overview of the selection procedure. Most duplicates were automatically detected and removed using EndNote (version 20; Clarivate) [19]. Screening followed a 2-step process. First, 2 researchers (JG and KdN) independently screened titles and abstracts and labeled them as relevant, irrelevant, or potentially relevant for inclusion. Additional duplicates not detected by EndNote were removed during this step. Second, both reviewers screened the full texts of relevant and potentially relevant studies for final inclusion. JG and KdN resolved discrepancies in both steps through discussion and consensus and consulted a third and fourth reviewer (LP and LVdB), if necessary. JG screened articles found during the updated search. We used Rayyan (Qatar Computing Research Institute) [20] for reference management and manual removal of duplicates.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram. ILM: intensive longitudinal method.

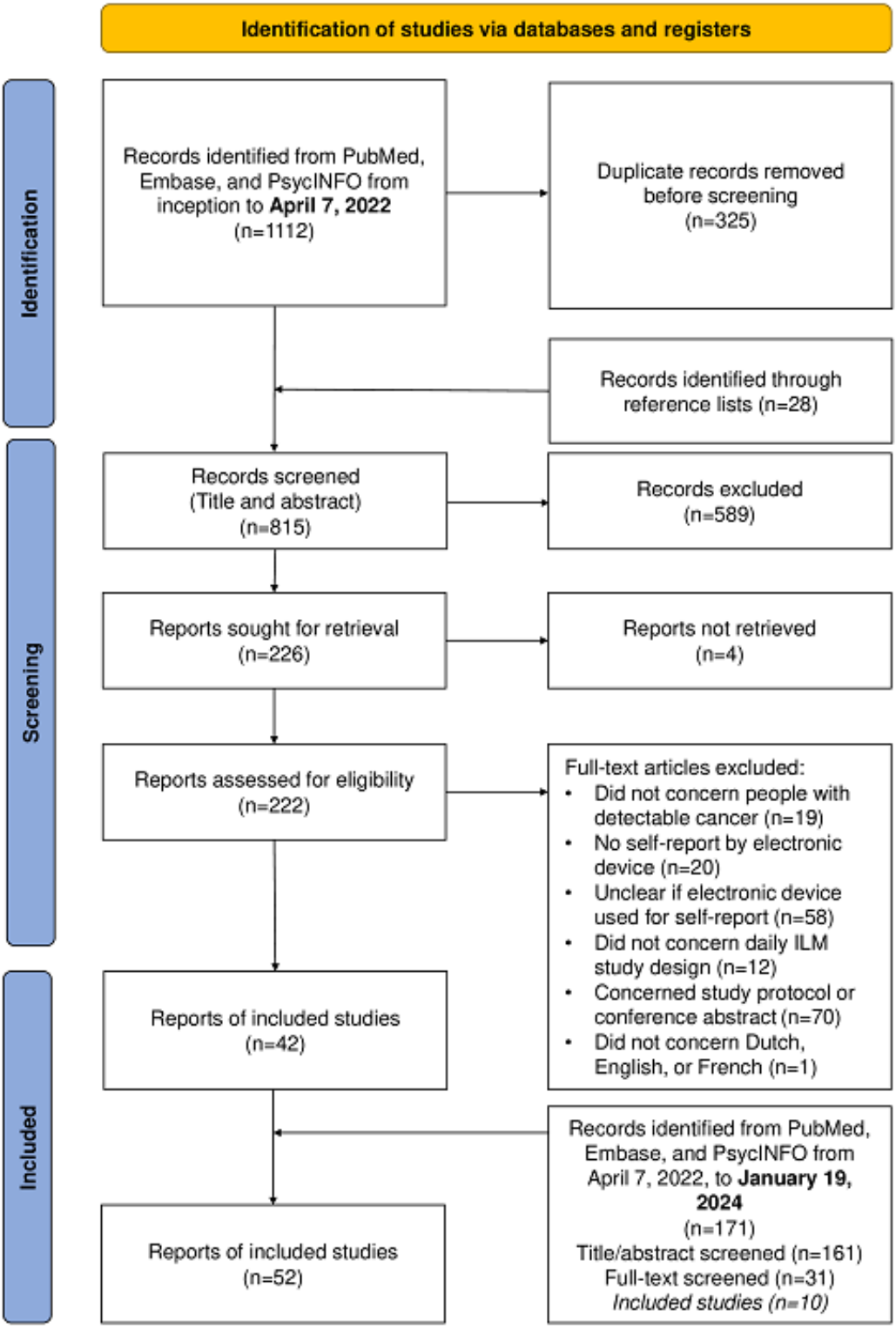

XSL•FO
RenderX

Data Extraction and Synthesis

JG extracted data into a precreated MS Excel (version 16; Microsoft Corp) spreadsheet. To ensure consistency and accuracy of the initial search, KdN independently completed the data extraction form for a random 10% sample of included articles. JG and KdN discussed and resolved discrepancies. JG extracted updated search results. The data extraction form was revised throughout the review process. It included study characteristics (ie, authors, year, country, and overarching study); sample characteristics (ie, sample size, site of primary tumor, stage of disease, mean age, proportion of female participants, and comparison group characteristics); study aims and design; system characteristics (ie, device, application, and operation system); daily questionnaire characteristics (ie, number of items, constructs measured daily, existing measurement instruments, or sources used); sampling schedule characteristics (ie, number of monitoring periods, duration of the monitoring periods, type of sampling scheme [ie, fixed or random signal-contingent, event-contingent, or interval-contingent] [9], daily prompt frequency, and approximate time interval between prompts); supportive features for participants; response-related results (ie, participation rate, attrition rate, proportion of completed prompts, and monetary incentives); and main study findings. We listed the barriers and facilitators for the implementation of the used method in research and clinical practice per study.

We have presented the study and sample characteristics, system and sampling schedule characteristics, and response-related results in the *Results* section, grouping articles reporting on the same study. We conducted content analysis on the extracted barriers and facilitators, inductively categorizing the content in themes and subthemes.

Results

Of the 1311 identified articles, we screened 253 (19.3%) full-text articles for eligibility. We included 52 articles, describing 41 unique studies (Figure 1).

Population Characteristics

All the 41 studies were conducted in high-income countries, except for 1 (2%) study in Türkiye [21] (all study and sample characteristics are listed in Multimedia Appendix 2) [21-72]. We included 21 (51%) studies [22-44] reporting on samples of people with mixed primary tumor sites (including breast and lung cancer), 16 (39%) studies [21,45-66] on samples of people with breast cancer only, and 4 (10%) studies [67-72] on a sample of lung cancer only (Table 1). A total of 7 (17%) studies included patients' partners [25,38,39,43,45-50,63]. While 26 (63%) studies were conducted in people at differing stages of disease, of which 11 (42%) included up to stage III [21,48-51,54-59,61,63,73] and 15 (58%) included up to stage IV [23,25,26,31,32,34,36-39,41-44,53,64,69,70], 6 (15%) studies [24,30,33,35,45-47,52] specifically focused on people with stage IV cancer. Sample sizes ranged from 4 [29] to 344 [42] participants, with a mean of 54.3 (SD 56.4). The mean ages were 51 (SD 4.7) years for patients with breast cancer, 65 (SD 2.8) years for patients with lung cancer, and 58 (SD 5.7) years for patients with mixed primary tumor sites. None of the studies used proxy responding.

XSL•FO
RenderX

Table 1. Study and sample characteristics of the included studies (n=41).

Characteristics	Frequency of studies, n (%)	Reference, year
Primary tumor sites		
Breast	16 (39)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Cai et al [51], 2020 • Carson et al [52], 2021 • Çınar et al [21], 2021 • Dasch et al [53], 2010 • Kim et al [60], 2016 • Lim et al [64], 2022 • Min et al [61], 2014 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • Xu et al [63], 2019
Lung	4 (10)	• Dunsmore et al [72], 2023 • Maguire et al [71], 2015 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Steffen et al [69], 2018; Steffen et al [70], 2020
Mixed	21 (51)	• Aigner et al [22], 2016 • Besse et al [34], 2016 • Chumbler et al [23], 2007 • Coolbrandt et al [40], 2022 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kearney et al [28], 2006 • Langer et al [25], 2018 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lee et al [41], 2023 • Maguire et al [29], 2005 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Schuler et al [43], 2023 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023 • Weaver et al [33], 2014 • Yap et al [37], 2013
Included patients and partners	7 (17)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018; • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Langer et al [25], 2018; LeBaron et al [38], 2022 • LeBaron et al [39], 2023 • Schuler et al [43], 2023 • Xu et al [63], 2019
Disease stage		
I to II	1 (2)	• Shiyko et al [68], 2014; Shiyko et al [67], 2019

XSL•FO
RenderX

Characteristics	Frequency of studies, n (%)	Reference, year
III to IV	9 (22)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Carson et al [52], 2021 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lim et al [64], 2022 • McCall et al [30], 2008 • Schuler et al [43], 2023 • Weaver et al [33], 2014
Mixed	20 (49)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Besse et al [34], 2016 • Cai et al [51], 2020 • Chumbler et al [23], 2007 • Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Dasch et al [53], 2010 • Langer et al [25], 2018 • McCann et al [31], 2009; Kearney et al [32], 2009 • Min et al [61], 2014 • Mooney et al [26], 2014 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [63], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019 • Yap et al [37], 2013
Not fully mentioned	11 (27)	• Aigner et al [22], 2016 • Dunsmore et al [72], 2023 • Kearney et al [28], 2006 • Kim et al [60], 2016 • Lee et al [41], 2023 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Sztachańska et al [62], 2019 • van den Berg et al [27], 2022
Sample size		

XSL•FO
RenderX

Characteristics	Frequency of studies, n (%)	Reference, year
4-20	9 (22)	• Besse et al [34], 2016 • Cai et al [51], 2020 • Hachizuka et al [35], 2010 • Kearney et al [28], 2006 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023
21-50	14 (34)	• Aigner et al [22], 2016 • Carson et al [52], 2021 • Chumbler et al [23], 2007 • Çınar et al [21], 2021 • Dunsmore et al [72], 2023 • Harper et al [24], 2012 • McCall et al [30], 2008 • Min et al [61], 2014 • Passardi et al [36], 2022 • Pinto et al [54], 2021 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Sztachańska et al [62], 2019 • van Roozendaal et al [44], 2023 • Weaver et al [33], 2014
51-100	13 (32)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Dasch et al [53], 2010 • Kim et al [60], 2016 • Lim et al [64], 2022 • McCann et al [31], 2009; Kearney et al [32], 2009 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Solk et al [56], 2020; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • Xu et al [63], 2019 • Yap et al [37], 2013
100-344	5 (12)	• Coolbrandt et al [40], 2022 • Langer et al [25], 2018 • Lee et al [41], 2023 • Mooney et al [26], 2014 • Nordhausen et al [42], 2022
Mean age (years)		

XSL•FO
RenderX

Characteristics	Frequency of studies, n (%)	Reference, year
40-50	10 (24)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Çınar et al [21], 2021 • Kim et al [60], 2016 • Langer et al [25], 2018 • Min et al [61], 2014 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Sztachańska et al [62], 2019 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019 • Yap et al [37], 2013
51-60	19 (46)	• Aigner et al [22], 2016 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Besse et al [34], 2016 • Cai et al [51], 2020 • Carson et al [52], 2021 • Coolbrandt et al [40], 2022 • Dasch et al [53], 2010 • Harper et al [24], 2012 • Lee et al [41], 2023 • Lim et al [64], 2022 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • Weaver et al [33], 2014
61-70	9 (22)	• Chumbler et al [23], 2007 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Maguire et al [71], 2015 • McCall et al [30], 2008 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Steffen et al [69], 2018; Steffen et al [70], 2020
Not mentioned	3 (7)	• Kearney et al [28], 2006 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Maguire et al [29], 2005

Study design as reported by study authors

XSL•FO
RenderX

Characteristics	Frequency of studies, n (%)	Reference, year
Observational	30 (73)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Chumbler et al [23], 2007 • Coolbrandt et al [40], 2022 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kearney et al [28], 2006 • Kim et al [60], 2016 • Langer et al [25], 2018 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lee et al [41], 2023 • Maguire et al [71], 2015 • McCall et al [30], 2008 • Min et al [61], 2014 • Nordhausen et al [42], 2022 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019 • Yap et al [37], 2013
Interventional	11 (27)	• Besse et al [34], 2016 • Cai et al [51], 2020 • Carson et al [52], 2021 • Çınar et al [21], 2021 • Lim et al [64], 2022 • Maguire et al [29], 2005 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Passardi et al [36], 2022 • Sztachańska et al [62], 2019 • Weaver et al [33], 2014
Study objectives		

XSL•FO
RenderX

Characteristics	Frequency of studies, n (%)	Reference, year
Feasibility, usability, or validity	19 (46)	• Besse et al [34], 2016 • Cai et al [51], 2020 • Chumbler et al [23], 2007 • Coolbrandt et al [40], 2022 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kearney et al [28], 2006 • Kim et al [60], 2016 • LeBaron et al [38], 2022 • Lee et al [41], 2023 • Maguire et al [29], 2005 • McCall et al [30], 2008 • Min et al [61], 2014 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Schuler et al [43], 2023 • Solk et al [56], 2019 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • Yap et al [37], 2013
Trajectory or relationship of variables	17 (41)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Langer et al [25], 2018 • LeBaron et al [39], 2023 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Shiyko et al [67], 2019 • Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • van Roozendaal et al [44], 2023 • Weaver et al [33], 2014 • Xu et al [63], 2019
Effectiveness of methods as intervention	4 (10)	• Çınar et al [21], 2021 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Maguire et al [71], 2015
Effectiveness of other interventions	3 (7)	• Carson et al [52], 2021 • Lim et al [64], 2022 • Sztachańska et al [62], 2019
Introduce statistical approach	1 (2)	• Shiyko et al [68], 2014

Study Design and Objectives

Of the 41 studies, 30 (73%) [22-25,27,28,30,37-50,53-61,63,65-67,69-72] used intensive methods in observational study designs, whereas 11 (27%) [21,26,29,31-34,36,51,52,62,64] used them in interventional studies. While 38% (20/52) of the articles [23,24,27-30,33,34,36-38,40-43,51,56,59-61] focused on the intensive method's feasibility, usability, or validity, other articles investigated the prevalence or trajectory of measured variables or relationships between those variables [22,25,39,44-50,53-55,57,58,63,65-67,69,70,72], the effectiveness of the intensive methods as an intervention [21,26,31,32,71], or the effectiveness of other interventions [52,62,64] or introduced a novel statistical approach [68].

Data Collection Methods

Daily Measured Constructs

Of the 41 studies, 30 (73%) [23-25,27-34,36,37,40-42,44-50,52-59,62,63,65-72] used items adapted from previous studies or scales (study details are listed in Multimedia Appendix 3) [21-72]. Some of the most frequently recurring questionnaires were the Common Toxicity Criteria Adverse Events grading system [29,31-33,37,41,74], EORTC

XSL•FO
RenderX

Core Quality of Life Questionnaire (EORTC-QLQ-C30) [24,42,69,70,75], and the Positive and Negative Affect Schedule-Expanded scale [48-50,53,69,70,76]. Measured constructs covered physical, psychological, and social domains; behaviors and intentions; daily events; sleep quality; and general quality of life. The physical domain was the most assessed domain, with the most frequently measured constructs being pain [22,23,26,27,30,34,35,38-41,45-47,52,54,56,57,59,65-68,70] and fatigue [23,26-32,35,40,41,44,54-57,59,65,66,70]. Anxiety [22,26,35,41,51,54,56,57,59-61,65,66,72] and depression [26,35,51,56,57,59,61,65,66] were the most frequently measured constructs in the psychological domain, and social support [25,45,48,50,62] and communication [25,38,39,46,49,50,63] were the most frequently measured constructs in the social domain. Frequently measured behavioral constructs included medication use [22,36,38-40,47,61,64] and physical activity [38,39,56,57,65,66].

Sampling Schedule Characteristics

Of the 41 studies, 23 (56%) [21-24,26-28,30,37-41,43,48,50-53,59,61,62,69,70,72] required patients to fill in the questionnaire once per day, while 6 (15%) studies [25,29,31-34,67,68] required 2 completions daily, and 7 (17%) studies [35,44-47,54-58,63,65,66] required 3-6 completions daily (Table 2). Moreover, 5 (12%) studies [36,42,60,64,71] did not report the specific amount.

XSL•FO
RenderX

Table 2. Data collection methods used in the included studies (n=41).

Data collection methods	Frequency of studies, n (%)	Reference, year
Sampling schedule		
Once daily	23 (56)	• Aigner et al [22], 2016 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Cai et al [51], 2020 • Carson et al [52], 2021 • Chumbler et al [23], 2007 • Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Harper et al [24], 2012 • Kearney et al [28], 2006 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lee et al [41], 2023 • McCall et al [30], 2008 • Min et al [61], 2014 • Mooney et al [26], 2014 • Schuler et al [43], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • van den Berg et al [27], 2022 • Yap et al [37], 2013
Twice daily	6 (15)	• Besse et al [34], 2016 • Langer et al [25], 2018 • Maguire et al [29], 2005 • McCann et al [31], 2009; Kearney et al [32], 2009 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Weaver et al [33], 2014
3-6 times daily	7 (17)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Hachizuka et al [35], 2010 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019
Not mentioned	6 (15)	• Kim et al [60], 2016 • Lim et al [64], 2022 • Maguire et al [71], 2015 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022
Sampling type[a]		
Fixed signal-contingent	15 (37)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Besse et al [34], 2016 • Cai et al [51], 2020 • Coolbrandt et al [40], 2022 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Langer et al [25], 2018 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Min et al [61], 2014 • Passardi et al [36], 2022 • Schuler et al [43], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Xu et al [63], 2019 • Yap et al [37], 2013

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
Random signal-contingent	7 (17)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Hachizuka et al [35], 2010 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2021; Welch et al [65], 2023; Whitaker et al [66], 2023 • van Roozendaal et al [44], 2023
Interval-contingent	6 (15)	• Çınar et al [21], 2021 • Dasch et al [53], 2010 • McCall et al [30], 2008 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • Weaver et al [33], 2014
Event-contingent	6 (15)	• Hachizuka et al [35], 2010 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Maguire et al [29], 2005 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Schuler et al [43], 2023
Not clearly mentioned	15 (37)	• Aigner et al [22], 2016 • Carson et al [52], 2021 • Chumbler et al [23], 2007 • Harper et al [24], 2012 • Kearney et al [28], 2006 • Kim et al [60], 2016 • Lee et al [41], 2023 • Lim et al [64], 2022 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • van den Berg et al [27], 2022
Data collection period length (days)		

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
5	1 (2)	• Yap et al [37], 2013
7	8 (20)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Cai et al [51], 2020 • Carson et al [52], 2021 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Kearney et al [28], 2006 • Pinto et al [54], 2021
8-13	3 (7)	• Otto et al [50], 2015 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Xu et al [63], 2019
14	7 (17)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Langer et al [25], 2018 • Maguire et al [29], 2005 • McCann et al [31], 2009; Kearney et al [32], 2009 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Sztachańska et al [62], 2019
>14	12 (29)	• Besse et al [34], 2016 • Çınar et al [21], 2021 • Lee et al [41], 2023 • Lim et al [64], 2022 • Maguire et al [71], 2015 • McCall et al [30], 2008 • Min et al [61], 2014 • Schuler et al [43], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • Weaver et al [33], 2014
Variable per person	10 (24)	• Chumbler et al [23], 2007 • Coolbrandt et al [40], 2022 • Harper et al [24], 2012 • Kim et al [60], 2016 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Mooney et al [26], 2014 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Ratcliff et al [55], 2014 • van Roozendaal et al [44], 2023
Data collection devices for self-report assessments		

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
Smartphone	11 (27)	• Cai et al [51], 2020 • Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Langer et al [25], 2018 • Min et al [61], 2014 • Pinto et al [54], 2021 • Schuler et al [43], 2023 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019
Smartwatch	2 (5)	• LeBaron et al [38], 2022; LeBaron et al [39], 2023
Handheld computer	8 (20)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kearney et al [28], 2006 • McCall et al [30], 2008 • Ratcliff et al [55], 2014 • Shiyko et al [68], 2014; Shiyko et al [67], 2019
Mobile device with telephone or SMS functionality	9 (22)	• Besse et al [34], 2016 • Carson et al [52], 2021 • Lee et al [41], 2023 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Weaver et al [33], 2014 • Yap et al [37], 2013
Device with internet functionality	5 (12)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016
Specifically developed device	2 (5)	• Chumbler et al [23], 2007 • Nordhausen et al [42], 2022
Not mentioned	5 (12)	• Otto et al [50], 2015 • Kim et al [60], 2016 • Lim et al [64], 2022 • Passardi et al [36], 2022 • Sztachańska et al [62], 2019
Device ownership		

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
Patient-owned	19 (46)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Besse et al [34], 2016 • Cai et al [51], 2020 • Carson et al [52], 2021 • Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Dasch et al [53], 2010 • Lee et al [41], 2023 • Min et al [61], 2014 • Mooney et al [26], 2014 • Pinto et al [54], 2021 • Schuler et al [43], 2023 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2021; Welch et al [65], 2023; Whitaker et al [66], 2023 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019 • Yap et al [37], 2013
Provided by researcher	12 (29)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Chumbler et al [23], 2007 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kearney et al [28], 2006 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Nordhausen et al [42], 2022 • Ratcliff et al [55], 2014 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Weaver et al [33], 2014
Option to choose between patient-owned and research device	2 (5)	• Langer et al [25], 2018 • Steffen et al [69], 2019; Steffen et al [70], 2020
Not mentioned	10 (24)	• Otto et al [50], 2015 • Dunsmore et al [72], 2023 • Kim et al [60], 2016 • Lim et al [64], 2022 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Passardi et al [36], 2022 • Sztachańska et al [62], 2019
Data collection software[a]		

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
Smartphone apps	9 (22)	• Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Kim et al [60], 2016 • Langer et al [25], 2018 • Min et al [61], 2014 • Pinto et al [54], 2021 • Schuler et al [43], 2023 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023
Browser-based surveys (sent via chat, mail, or SMS)	6 (15)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Xu et al [63], 2019
SMS	3 (7)	• Besse et al [34], 2016 • Cai et al [51], 2020 • Yap et al [37], 2013
Interactive voice responding systems	4 (10)	• Besse et al [34], 2016 • Carson et al [52], 2021 • Lee et al [41], 2023 • Mooney et al [26], 2014
Other specifically developed software	12 (29)	• Aigner et al [22], 2016 • Chumbler et al [23], 2007 • Kearney et al [28, 2006 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCann et al [31], 2009; Kearney et al [32], 2009 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Ratcliff et al [55], 2014 • Stone et al [59], 2016 • Weaver et al [33], 2014
Not mentioned	8 (20)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Otto et al [50], 2015 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Lim et al [64], 2022 • McCall et al [30], 2008 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Sztachańska et al [62], 2019
Used conditional questionnaire items	7 (17)	• Badr et al [45], 2010; Badr et al [46], 2013 Stephenson et al [47], 2018 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Coolbrandt et al [40], 2022 • Langer et al [25], 2018 • Mooney et al [26], 2014 • Shiyko et al [68], 2014; Shiyko et al [67], 2019
Used different questionnaire lengths depending on prompt timing	5 (12)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Langer et al [25], 2018 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
The number of questionnaire items		
1-20	20 (49)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Besse et al [34], 2016 • Carson et al [52], 2021 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kim et al [60], 2016 • Langer et al [25], 2018 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Min et al [61], 2014 • Mooney et al [26], 2014 • Nordhausen et al [42], 2022 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Stone et al [59], 2016 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023 • Yap et al [37], 2013
21-40	6 (15)	• Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Lee et al [41], 2023 • Pinto et al [54], 2021 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Sztachańska et al [62], 2019
41-84	2 (5)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015
Not clearly mentioned	13 (32)	• Cai et al [51], 2020 • Chumbler et al [23], 2007 • Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Kearney et al [28], 2006 • Lim et al [64], 2022 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Passardi et al [36], 2022 • Weaver et al [33], 2014 • Xu et al [63], 2019
Supportive features		

XSL•FO
RenderX

Data collection methods	Frequency of studies, n (%)	Reference, year
Automated self-care advice	9 (22)	• Chumbler et al [23], 2007 • Coolbrandt et al [40], 2022 • Kearney et al [28], 2006 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Weaver et al [33], 2014 • Yap et al [37], 2013
Clinician alerts	9 (22)	• Besse et al [34], 2016 • Coolbrandt et al [40], 2022 • Kearney et al [28], 2006 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Weaver et al [33], 2014 • Yap et al [37], 2013
Clinician could view summary of responses	5 (12)	• Coolbrandt et al [40], 2022 • Harper et al [24], 2012 • Kearney et al [28], 2006 • Min et al [61], 2014 • Nordhausen et al [42], 2022
Informational modules	2 (5)	• Çınar et al [21], 2021 • Passardi et al [36], 2022
Module allowing communication with clinicians	1 (2)	• Çınar et al [21], 2021
Patients received response summaries	2 (5)	• McCall et al [30], 2008 • Xu et al [63], 2019
Relaxation reminders	1 (2)	• Çınar et al [21], 2021
None mentioned	23 (56)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018; • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Carson et al [52], 2021 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Kim et al [60], 2016 • Langer et al [25], 2018 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lee et al [41], 2023 • Lim et al [64], 2022 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], (2023) • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • van den Berg et al [27], 2022 • van Roozendaal et al [44], 2023

[a]Multiple options possible per study.

XSL•FO
RenderX

Out of the 41 studies, 22 (54%) studies [25,34-40,43-51,54-58,61,63,65-70,72] applied signal-contingent sampling (ie, prompting respondents to complete the questionnaire) and 6 (15%) studies [21,30,33,53,59,62] applied interval-contingent sampling (ie, instructing respondents to complete the questionnaire at certain intervals), while 15 (37%) studies [22-24,26-29,31,32,36,41,52,60,64,71] did not specify the sampling method. Furthermore, 6 (15%) studies used event-contingent sampling on top of the other sampling methods; of these, 4 (67%) studies [29-32,38,39] instructed patients to complete the assessment when experiencing adverse events, 1 (17%) study [35] required the patients to assess when rescue medication was taken, and 1 (17%) study [43] prompted patients when a physiologically measured stress threshold was reached. Out of 22 signal-contingent sampling studies, 13 (59%) [25,34-37,43,48-51,61,63,69,70,72] prompted patients at fixed times, with times between prompts ranging from 3 to 24 hours. Moreover, 36% (8/22) of the studies [35,42,44-47,54-58,65-68] prompted patients at random times, of which 5 (62%) [35,44-47,55,67,68] randomly prompted within a fixed time block (eg, between 9 AM and midnight). Minimum time intervals between randomly timed prompts ranged from 30 minutes to 3 hours [45-47,54-58].

Of the 41 studies, 7 (17%) [31-33,52,54-58,65,66,71] had multiple data collection periods for each patient. While the most common data collection period lengths were 7 days [28,35,48-54,71,72] and 14 days [22,25,29,31,32,45-47,62,67,68], ranging from 1 to 336 days [42,77], 10 (24%) studies [23,24,26,36,38-40,42,44,55,60] mentioned differing study lengths for each patient (eg, based on patients' next chemotherapy visit) [55].

System Characteristics

Data collection devices and software varied substantively in the included studies (n=41), with 11 (27%) studies [21,25,27,40,43,44,51,54,56-58,61,63,65,66] using smartphones, 1 (2%) study using smartwatches [38,39], and 8 (20%) studies [22,24,28,30,35,45-47,55,67,68] using handheld computers for self-report assessments. Other studies used basic telephone and SMS text messaging functionality [26,34,37,41,51,52], internet functionality [48-50,53,56-59,65,66,69,70,72], and used a specifically developed device [23,38,39]. A total of 19 (46%) studies [21,26,27,34,37,40,41,43,44,48-54,56-59,61,63] used patients' devices, whereas 12 (29%) studies [22-24,28,30,33,35,38,39,42,45-47,55,67,68] provided devices to patients.

Different types of software were used, including smartphone apps [21,25,27,36,40,43,44,54,60,61,63,64], browser-based surveys [48-50,53,56-59,63,65,66,69,70,72], SMS text messaging [34,37,51], interactive voice responding systems [26,34,41,52], and other specifically developed software applications [22,23,28,29,31-33,38,39,42,55,71].

Questionnaire Length

Some studies (7/41, 17%) [25,26,38-40,43,45-50,65-68] used conditional items that were presented when a certain response was given to previous items and different questionnaires depending on the timing of the prompt (eg, the use of morning prompts to assess sleep quality [55,56]). Most studies (20/41, 49%) [21-25,27-29,31-35,37-40,42-47,51,52,55-61,63,65-68] had questionnaire lengths ranging between 1 and 20 items, with the longest being 84 items (including conditional items) [48-50]. Several studies (13/41, 32%) [21,23,28-33,36,40,51,63,64,71] did not provide complete information on the number of items.

Supportive Features

Of 41 studies, 17 (41%) [21,23,24,26,28-34,36,37,40,42,61,63,71] provided supportive features; 9 (22%) studies [23,28-33,37,40,71] offered automated self-care advice to patients based on their responses directly after response submission, for instance, offering advice for managing reported symptoms, with severe symptoms triggering advice to contact a health care professional [73]. Also, 9 (22%) studies [26,28,29,31-34,37,40,71] automatically contacted health care professionals based on symptom severity (ie, clinician alerts). Some studies (2/41, 5%) [31,32,71] differentiated between different severities to indicate varying levels of need for immediate intervention (eg, amber and red alerts). A total of 6 (15%) studies [28,29,31-33,37,71] combined automated self-care advice and clinician alerts. One study [26] alerted clinicians based on responses given on domains other than physical symptoms, namely psychological variables (ie, depressive mood and anxiety) and distress caused by symptoms. Other supportive features included providing the opportunity to clinicians to view a summary or visualization of responses given by the patient [24,28,40,42,61] and providing patients with informational modules [21,36], modules allowing communication with clinicians [21,36], response summaries [30,63], and relaxation reminders [21].

Study-Reported Findings

Findings Concerning Methodological Evaluations

Intensive longitudinal methods that sampled once daily [23,26-28,30,37,40,42] or multiple times per day [33-35,38,43,56] were deemed feasible and acceptable for patients. These findings applied to various system characteristics, such as interactive voice response and SMS text messaging systems [26,34,37] and smartphone apps [27,40,43]. Compliance decreased over time in a 90-day study [61], with higher compliance among unemployed women. Patients believed in the method's ability to improve symptoms [29], symptom management [28,71], and communication with clinicians [71]. Moreover, patients had positive views on the usability of the methods [26,30,34,35,56,71] and felt reassured by using them [29,33].

Health care professionals had a positive view of the methods [71] and found them reassuring for patients, especially during out of hours [33], and clinically useful [26,30,37]. In addition, health care professionals thought that the methods could be helpful aids in timely interventions [29] and for assessing [28] and managing symptoms [28,29]. However, one study [24] reported that quality of life data was not used for making treatment decisions, and other studies [26,42,64] reported that clinicians rarely contacted the patients after receiving clinical alerts or monitored their responses. In one study [71], health

XSL•FO
RenderX

care professionals mentioned that reduced complexity of the system was needed to promote its utility.

Some studies (5/41, 12%) [34,36,41,51,59] compared intensive longitudinal methods with other scales and found agreement between the methods, such as depression ratings and Patient Health Questionnaire-9 [60,78]. One study [27] found a lack of agreement between the intensive methods and the Short Form Health Survey [79], but this concordance improved with higher compliance rates.

Findings Concerning Prevalence and Covariability of Constructs

Several studies (16/41, 39%) examined the prevalence and covariability of constructs ranging across multiple topics. For instance, 7 studies [25,45,46,48-50,63] reported findings related to the social dynamics between patients and their partners. One study [45] found greater reports of relationship interference when patients experienced more pain and lower arousal mood. Moreover, partners were more likely to provide support when patients experienced more tiredness and less active mood resulting from pain [45]. Another study on this topic [48] found that partners' reports of support provision were positively associated with feelings of relationship intimacy reported by patients.

Overall, studies investigated various topics such as physical activity, affect, and physical symptoms. For instance, studies [54,65] showed associations between sedentary behavior, affective valence, and fatigue at different time points, analog to other studies [57,66] that found within-person associations between physical activity and same-day affect, fatigue, pain, and others.

Findings Concerning the Intensive Methods as an Intervention

Of the 41 studies, 7 (17%) [21,23,26,31,32,34,71] investigated the impact of intensive longitudinal methods as an intervention tool to improve symptoms, for instance, by providing automated self-care advice to patients or alerting clinicians when a certain symptom threshold was reached [71]. Patients in the intervention groups reported lower distress [21], lower fatigue, and higher levels of hand-foot syndrome [32] than those in the control groups. Patient-reported benefits included improved communication with health care professionals and symptom management and reassurance that symptoms were being monitored at home [31]. After the intervention, patients reported increased quality of life [21,23], lower anxiety and drowsiness, lower pain [34], and higher self-care efficacy [71] than at the baseline. One study using clinician alerts [26] found no improvements in symptom severity, explained by clinicians rarely contacting patients after alerts.

Response-Related Results

Of the 41 studies, 21 (51%) [22-26,33,34,37,40,43-50,52-54,56-58,63,65-67,69-71] reported participation rates ranging from 23.6% to 90.3% (mean 52.9, SD 3.4; Table 3; Multimedia Appendix 4) [21-72]. Overall, 17 (41%) studies [23,25,26,30,34,35,37-40,43,44,51,55,61,63,64,71] reported attrition rates, ranging from 0% to 56.9% (mean 19.7%, SD 17.7%). Furthermore, 19 (46%) studies [22,27,28,31-33,36,45-50,52-54,56-60,62,65-70,76] provided other attrition indicators, while 29 (71%) studies [22-27,33-35,40,42,43,45-47,50-67,69,70,72] reported compliance rates ranging from 44.2% to 98% (mean 74.9%, SD 16.4%).

XSL•FO
RenderX

Table 3. Response-related results of the included studies (n=41).

Results and characteristics	Frequency of studies, n (%)	Reference, year
Participation rate		
23%-25%	3 (7)	• Coolbrandt et al [73], 2021 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • van Roozendaal et al [44], 2023
26%-50%	8 (20)	• Aigner et al [22], 2016 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Carson et al [52], 2021 • Dasch et al [53], 2010 • Langer et al [25], 2018 • Maguire et al [71], 2015 • Xu et al [63], 2019 • Yap et al [37], 2013
51%-75%	4 (10)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018; • Pinto et al [54], 2021 • Schuler et al [43], 2023 • Weaver et al [33], 2014
76%-90%	6 (15)	• Besse et al [34], 2016 • Chumbler et al [23], 2007 • Harper et al [24], 2012 • Mooney et al [26], 2014 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Steffen et al [69], 2018; Steffen et al [70], 2020
Not mentioned	19 (46)	• Otto et al [50], 2015 • Cai et al [51], 2020 • Çınar et al [21], 2021 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Kearney et al [28], 2006 • Kim et al [60], 2016 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lee et al [41], 2023 • Lim et al [64], 2022 • Maguire et al [29], 2005 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Min et al [61], 2014 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Ratcliff et al [55], 2014 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • van den Berg et al [27], 2022
Attrition rate		
0%-25%	12 (29)	• Cai et al [51], 2020 • Coolbrandt et al [40], 2022 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Langer et al [25], 2018 • Min et al [61], 2014 • Mooney et al [26], 2014 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • van Roozendaal et al [44], 2023 • Xu et al [63], 2019 • Yap et al [37], 2013

XSL•FO
RenderX

Results and characteristics	Frequency of studies, n (%)	Reference, year
26%-57%	6 (15)	• Besse et al [34], 2016 • Chumbler et al [23], 2007 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lim et al [64], 2022 • Maguire et al [71], 2015 • McCall et al [30], 2008
Other indicators mentioned	18 (44)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018; • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Carson et al [52], 2021 • Dasch et al [53], 2010 • Kearney et al [28], 2006 • Kim et al [60], 2016 • Lee et al [41], 2023 • McCann et al [31], 2009; Kearney et al [32], 2009 • Passardi et al [36], 2022 • Pinto et al [54], 2021 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • van den Berg et al [27], 2022 • Weaver et al [33], 2014
None mentioned	5 (12)	• Otto et al [50], 2015 • Çınar et al [21], 2021 • Dunsmore et al [72], 2023 • Maguire et al [29], 2005 • Nordhausen et al [42], 2022
Compliance rate		

XSL•FO
RenderX

Results and characteristics	Frequency of studies, n (%)	Reference, year
44%-60%	6 (15)	• Otto et al [50], 2015 • Kim et al [60], 2016 • Min et al [61], 2014 • Ratcliff et al [55], 2014 • Schuler et al [43], 2023 • van den Berg et al [27], 2022
61%-80%	10 (24)	• Aigner et al [22], 2016 • Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Besse et al [34], 2016 • Carson et al [52], 2021 • Coolbrandt et al [40], 2022 • Dunsmore et al [72], 2023 • Mooney et al [26], 2014 • Pinto et al [54], 2021 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Xu et al [63], 2019
81%-100%	13 (32)	• Cai et al [51], 2020 • Chumbler et al [23], 2007 • Dasch et al [53], 2010 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Langer et al [25], 2018 • Lim et al [64], 2022 • Nordhausen et al [42], 2022 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023 • Steffen et al [69], 2018; Steffen et al [70], 2020 • Stone et al [59], 2016 • Sztachańska et al [62], 2019 • Weaver et al [33], 2014
Other indicators mentioned	6 (15)	• Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Lee et al [41], 2023 • Passardi et al [36], 2022 • van Roozendaal et al [44], 2023 • Yap et al [37], 2013
Not mentioned	6 (15)	• Çınar et al [21], 2021 • Kearney et al [28], 2006 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009
Monetary incentives		
Amount based on the number of completed assessments	6 (15)	• Badr et al [45], 2010; Badr et al [46], 2013; Stephenson et al [47], 2018 • Belcher et al [48], 2011; Pasipanodya et al [49], 2012 • Otto et al [50], 2015 • Langer et al [25], 2018 • Pinto et al [54], 2021 • Ratcliff et al [55], 2014
		Steffen et al [69], 2018; Steffen et al [70], 2020

XSL•FO
RenderX

Results and characteristics	Frequency of studies, n (%)	Reference, year
Fixed amount	5 (12)	• Cai et al [51], 2020 • Carson et al [52], 2021 • Stone et al [59], 2016 • LeBaron et al [38], 2022; LeBaron et al [39], 2023 • Solk et al [56], 2019; Phillips et al [57], 2020; Auster-Gussman et al [58], 2022; Welch et al [65], 2023; Whitaker et al [66], 2023
None provided	2 (5)	• Min et al [61], 2014 • van den Berg et al [27], 2022
Not specified	28 (68)	• Aigner et al [22], 2016 • Otto et al [50], 2015 • Besse et al [34], 2016 • Chumbler et al [23], 2007 • Çınar et al [21], 2021 • Coolbrandt et al [40], 2022 • Dasch et al [53], 2010 • Dunsmore et al [72], 2023 • Hachizuka et al [35], 2010 • Harper et al [24], 2012 • Kearney et al [28], 2006 • Kim et al [60], 2016 • Lee et al [41], 2023 • Lim et al [64], 2022 • Maguire et al [29], 2005 • Maguire et al [71], 2015 • McCall et al [30], 2008 • McCann et al [31], 2009; Kearney et al [32], 2009 • Mooney et al [26], 2014 • Nordhausen et al [42], 2022 • Passardi et al [36], 2022 • Schuler et al [43], 2023 • Shiyko et al [68], 2014; Shiyko et al [67], 2019 • Sztachańska et al [62], 2019 • van Roozendaal et al [44], 2023 • Weaver et al [33], 2014 • Xu et al [63], 2019 • Yap et al [37], 2013

Overall, 32% (13/41) of the studies provided monetary incentives, of which 8 (62%) studies [25,45-50,54,55,59,69,70] based attainable monetary amounts on the number of completed assessments, while 5 (38%) [38,39,51,52,56-58,65,66,72] provided patients with fixed amounts. Attainable monetary amounts ranged from US $40 to $200.

Barriers and Facilitators

Most studies reported the barriers and facilitators regarding the implementation of their methods in research or clinical practice (Table 4), either related to the person with cancer or the methods themselves. Some facilitating person-related factors included having confidence in using technology systems [31,56] and recognizing its clinical benefits [28,30,60]. Some person-related barriers were lack of smartphone ownership [40,61] and discomfort with technology [30,45-47,71]. However, inexperience with technology generally did not impact success with the study technologies [25,28,31,35]. However, smartphone users had higher compliance during an SMS protocol than basic phone users [37].

XSL•FO
RenderX

Table 4. Barriers and facilitators for the implementation of the method in practice and for research purposes, as stated by the papers' authors or extracted from the reported results.

Themes	Facilitators	Barriers
Factors related to the person with breast or lung cancer	• *Confidence* in their abilities to use technology systems [31,56] • Overall *preference for online diary* compared with paper diary [62] • Smartphone *users* had higher compliance than basic phone users [37] • *Recognize the clinical benefits* of using technology systems to report symptoms [28-30,42,60] and weigh these benefits against assessment burden [43] • *Willingness of patients* [30,42] • Patient *perceptions on the relevance* of the study to their needs [29] • *Sex, age, and diagnosis* did not impact compliance [42,43]; excluded participants appeared similar to the included participants [44] • *(Belief that) data are used* by clinicians [30,42,73]	• *Lack of interest or motivation* to participate can lead to small sample size [22] and lower compliance [42] • *Time constraints* affect participation rate and compliance [22,42,45,46] • *Symptoms and* side effects due to (advanced stage) illness and treatment may cause increased burden during study period, problems with pressing buttons, lower participation and compliance rates, and bias due to missing data [38,42,45-47,51,55,56,69] • Men were more likely to not use monitoring than women [26] • *Not owning a* smartphone prevents certain patients from using the monitoring system and thus participating in the study [61,73] • *Inexperience and discomfort about using the technology system* at start of the study period; particularly, *older adults* were less likely to participate [30,45-47,71] • *Caregiver status not easily verifiable through electronic health record*, disrupting eligibility screening [38] • Health care *professionals had doubts about the ability of patients* to complete electronic assessments [42] • Some patients barely wearing or averse to wearing the study device [38,43] • *Dyad studies require informed consent from patient and caregiver*, leading to logistical challenges [38] • Difficulties *remembering experiences with using the system* after the study period [31]

XSL•FO
RenderX

Themes	Facilitators	Barriers
Factors related to the method	• *Use of single items for constructs* to shorten questionnaire [39,48,58,69] reduces burden, improves adherence [39], and gives room for measurement of multiple constructs, possibly reducing reactivity to a single construct [69] • *Tailoring of sampling schedule to population of interest*, for example, limiting the frequency of assessments, to not overburden [67] or providing a broad enough window to respond in [53], possibly prompting the participant a second time if unanswered [55] • *Reminders or prompts*, including the option to tailor reminder schedules and contact by the researcher, might improve adherence [21,31,36,54,57,58,61] • Ability to use patients' *personal smartphones* [34,57], making the need for study visits to receive a specialized electronic study device obsolete [27,34,56] and providing a nonburdensome means to study individuals in their natural environment [27,34] • *Possibility to combine EMA*[a] *prompting with passive monitoring* through high-grade commercially available devices [43,57] • *Using electronic devices over paper-and-pencil* alternatives does not impact attrition [32] • *Portability* of mobile phones enables daily assessments [60], while smartwatches can enhance acceptability [38] • *Facial emotions scale* demands less cognitive effort, is less of a burden, and makes responding more enjoyable [60] • *"Unsure" response option* can improve data quality when patients are confused with a question [38] • *Simple questionnaire and system design* for an easier patient experience [29,31,42] • Option to report *additional information* after structured questionnaire for a better patient experience (eg, additional symptoms and having preexisting conditions) [31] • *More time explaining how to respond* correctly to SMS response system can improve the quality of responding when the response format is expected to be difficult [37] • *Standardized protocol checklist* for researchers to streamline deployment installation [38] • *Providing participants with handouts* before the study period, including frequently asked questions and contact information in case of difficulties in using system [25] • *Easy and fast access to PROMs*[b] *and gathered data*, for example, by the integration of monitoring system into the electronic patient, likely leads more uptake in clinical settings [30,42,73] and makes IT support crucial [42] • *Cloud services system* improves the ability to securely offload and store data in real time [38] • *Reducing time delays between consent and deployment* can mitigate attrition and accommodate the dynamic clinical status of patients [38] • *Iterative deployments* can improve setting up and removing the system [38] • *Personal support by research assistant* is appreciated by patients [42] and might improve adherence [44]	• *Single item constructs bring psychometric limitations* [39] • *Empty battery or low battery life*, possibly leading to device memory loss and missing data [38,45,46,51] • *Turned off phones or patients not wearing smartwatches* leading to missing data [38,51] • *Transmission or pairing errors* [33,38,42,51] can lead to frustrations [38] • Bugs in code *to monitor smartwatches* [38] • *Incompatibility issues* possible between smartphones' display specifications and the used app [61] • *Synchronization problems* related to automatic Android updates leading to inconsistent timing of EMA prompts [38] • *Poor reception* at home, for example, in rural areas [31,33], could cause necessity to switch SIM providers [33] • *Monitoring requires time and manpower* in a context with high clinician time constraints [37,42,71], possibly leading to fewer calls after clinician alerts [26], or lack of using monitoring results by clinical staff and trial investigators [42,64] • *Dependency of the implementation on* health care *professionals*, who are difficult to motivate to break the status quo [42] • *Vast amount of data* can be burdensome to clinicians [60] • *False-positive clinician alerts* due to errors in responding and transmission problems [33,37] • *Self-care information* not always read by patients [71] • Compliance to *time-blocked random signals* may be affected by participants waking up late or going to bed early [46] • *Developing EMA schemes can be challenging* when taking participant burden into account [39] • *Content irrelevant to patient* could cause dissatisfaction [37]; clinical monitoring measures should be tailored to their needs [42] • *24-hour recall may not be appropriate* to measure all symptoms [41] • *Unclear instructions* on when to complete event-contingent assessment can cause confusion among participants [38] • *Technical changes are complex and require time* to test and implement, but are often underestimated by clinical team [38] • In comparative trials, *electronic diary might bias patients* toward better self-management due to increased awareness and daily requirement to enter data [64] • *Interruption of monitoring assessment* (eg, due to diagnostics or therapy) [42] • *Rapid clinical staff turnover* [42]
Other factors	• COVID-19 *pandemic* [38,42]	__[c]

[a]EMA: ecological momentary assessment.

[b]PROM: patient-reported outcome measure.

[c]Not applicable.

Some facilitating method-related factors included the ability to tailor sampling schedules to the population of interest [53,55,67] and the option to use reminders [21,31,36,54,57,58,61]. Some barriers included technical issues such as empty batteries leading to memory loss and missing data [38,45,46,51] and false-positive clinician alerts due to faulty responding and transmission

XSL•FO
RenderX

problems [33,37]. All these factors were associated with improvements in participation and compliance rates, user-experience, patient burden, quality of responses, time requirements for researchers, and adoption in clinical settings [21,22,25-34,36-38,40,42-48,51,55-58,60,61,67,69,71].

Discussion

Principal Findings

Intensive longitudinal methods with daily electronic assessments have been used among people with breast or lung cancer at different disease stages. The methods involved 1-6 assessments per day to study a wide range of experiences in daily life, primarily physical and psychological symptoms. Some studies integrated supportive features within the longitudinal assessments. For most studies, compliance and attrition rates were acceptable, although many studies lacked complete methodological reporting. Few studies focused on patients in the advanced stage of disease. We identified the barriers and facilitators for using these methods, related to both the person with cancer and the method itself.

Our review highlights the promise of intensive longitudinal methods to provide unique insights into the daily lives of people living with cancer. Importantly, these methods generally seem feasible and acceptable among patients with breast or lung cancer, supported by positive patient and health care professional experiences, along with compliance and attrition rates indicating acceptable amounts of missing data. These findings were true for different methodological approaches, such as studies that assessed patients once or multiple times daily. Moreover, these methods demonstrate flexibility as they were used to address an array of objectives, such as exploring within-person symptom associations [55] or communication patterns in dyads [63].

Before widespread implementation of these intensive methods in oncology research and practice, several of our findings encourage further investigation into its feasibility and optimal study conditions. First of all, it is striking that response- and methodology-related reporting was often incomplete or reported in different ways (eg, compliance rates and amount of questionnaire items). Standardized reports of this information are critical to inform optimal methodological choices in future studies or clinical procedures, as poor choices can lead to additional patient burden and missing data. Due to the unstandardized reporting by many included studies, comparisons in response-related results between studies with different methodological features were not possible in this review. Yet, such comparisons are particularly important when using intensive sampling methods in populations who are already susceptible to increased disease-related burden. In addition, several identified factors need further exploration to enhance the implementation of intensive longitudinal methods with daily electronic assessments in research and practice, for example, participants' feelings of inexperience and discomfort with technology leading to a lower likelihood to participate in the study [30,45-47,71]. Finally, low participation rates of the included studies indicate participant recruitment to be difficult, and sample sizes were often small. This is a major barrier for research, as it could lead to sampling bias, for instance, through self-selected sampling of people more confident or experienced in using electronic systems. Subsequently, this could limit the validity of study findings.

Our review identified understudied areas that prevent gaining a complete understanding of people with breast or lung cancer and their daily experiences. First, several populations of people with breast or lung cancer are currently underrepresented in intensive longitudinal method studies, which significantly limits the generalizability of findings for these populations, including findings on the feasibility of these methods. For instance, of the 41 studies, only 4 (10%) were conducted in people with lung cancer specifically, 6 (15%) studies were conducted in people with stage IV cancer specifically, 1 (2%) study was conducted in a low-income country, and only 1 (2%) study included 1 male participant with breast cancer. Second, although the study objectives varied widely, studies predominantly focused on the aspects of physical health, such as pain, or had rather clinical views on psychological constructs by focusing on depression and anxiety. Only one included study [62] covered experiences from spiritual or existential quality of life domains, which is remarkable because these experiences generally have increasing value at the end of life [5,6]. Furthermore, although ESMs offer the potential for linking patient experiences with concurrent contexts (eg, where the patient is and what they are doing) [12], these contextual aspects remain understudied among people with breast or lung cancer. A broader focus encompassing different domains and contexts is needed to gain a more comprehensive understanding of patients' quality of life and well-being, ultimately enabling the improvement of patient-centered care.

Implications for Practice and Research

On the basis of our findings, we provide several recommendations for practice and research. First, applying existing reporting guidelines for EMAs, such as those synthesized by Liao et al [80], can improve transparency and consistency in reporting for intensive longitudinal studies in oncology. Their checklist serves as a starting point to fulfill recommended reporting criteria, such as reporting the use of prompts and complete questionnaire information [81]. This will allow future researchers to accurately explore the effects of study features on response-related results.

Second, addressing implementation factors highlighted in this review can be achieved through simple solutions, such as providing clear instructions, training on the use of the methods, and emphasizing the importance of the study to increase patient motivation and confidence [25,28-31,37,56,60]. Moreover, extensive pretesting such as conducting a pilot study is essential to uncover any technical issues that may arise.

Third, it is essential to determine optimal conditions for using intensive longitudinal methods with daily electronic assessments in people with cancer, such as ideal sampling schemes for the feasible measurement of specific constructs [82,83]. Studies should focus on populations at an increased risk for symptom burden, such as those with advanced stage cancer [84,85]. Furthermore, the use of supportive features such as automated feedback and clinician alerts needs more investigation to explore how it is optimally implemented in routine clinical practice for

XSL•FO
RenderX

the best possible outcomes. Moreover, it is recommended to develop measures to examine the quality of responses provided by patients [86], as these could be influenced by cancer and its treatment (eg, through cognitive impairment).

Fourth, future studies among patients with breast and lung cancers could broaden their focus to encompass more nonclinical psychological or spiritual-existential topics and contextual factors. This approach could yield novel insights into the interplay between physical functioning and other aspects of well-being and how they vary in different contexts [8]. Researchers could look to other populations of people living with or beyond cancer to further inform on the possibilities of these methods. For example, studies involving survivors of cancer could have a less clinical focus due to living past the treatment stage. Future literature reviews of the use of daily methods among such populations would be greatly beneficial.

Finally, studies should further explore how multiple daily measurements compare with the same constructs as measured by the more commonly used patient-reported outcome measures in oncology, in which patients are expected to aggregate experiences over ≥1 weeks [87,88]. Such research could examine the ecological validity of these commonly used patient-reported outcome measures [59] and provide valuable insights for oncology research and practice regarding which experiences are more accurately measured on a more frequent basis.

Strengths and Limitations

This scoping review followed a broad systematic search strategy in multiple databases, incorporating studies that used self-report methods to assess patients daily or multiple times a day. Consequently, it offers a comprehensive overview of the methods used to gain insight into the daily experiences of people with breast and lung cancers at various stages across different countries.

Nevertheless, this review has limitations. First, it is plausible that we missed studies that used different terms for their daily electronic self-report questionnaire than those used in our search string. However, the broadness of our search string minimized this risk, and we detected articles that reported on methods that could be classified as ESMs but were not identified by the previous review in 2019 [12]. Second, only 10% of data extraction was checked by a second reviewer, and none were compared during the updated search, introducing a slight possibility of inaccuracies. We consider this a minor risk, as we found no disagreements in the 10% data that we had checked.

Conclusions

Intensive longitudinal methods using daily electronic assessments hold promise and can be feasible to provide unique insights into the daily lives of patients with breast or lung cancer. However, our findings encourage further research on the feasibility of determining optimal conditions for intensive monitoring, specifically in more advanced disease stages, and better adherence to standardized reporting guidelines. Moreover, considering a more multidimensional approach to the topics studied, especially beyond physical and psychopathological symptoms, will enhance the value of these methods, ultimately aiding in the improvement of patient-centered care in oncology.

Acknowledgments

The authors would like to thank Lise Rosquin for her valuable comments on the manuscript. This work is part of a wider study funded by the Research Foundation-Flanders (FWO), under grant agreement (G0C8120N). LP is a senior postdoctoral fellow of the Research Foundation-Flanders. LVdB holds the Collen-Francqui Research Professor Chair “Aging, dementia and palliative care” (2020-2023).

Authors' Contributions

All authors contributed to the conception and design of the study. JG and KdN acquired and interpreted the data. JG drafted the manuscript. All authors reviewed the final draft and provided the final approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms.
[DOC File , 15 KB - jmir_v26i1e50224_app1.doc]

Multimedia Appendix 2
Study and sample characteristics.
[DOCX File , 33 KB - jmir_v26i1e50224_app2.docx]

Multimedia Appendix 3
Content and design characteristics.
[DOCX File , 37 KB - jmir_v26i1e50224_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Response-related characteristics.
[DOCX File , 22 KB - jmir_v26i1e50224_app4.docx]

Multimedia Appendix 5
PRISMA Scoping Review Checklist.
[PDF File (Adobe PDF File), 498 KB - jmir_v26i1e50224_app5.pdf]

References

1. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: an overview. Int J Cancer 2021 Apr 05 (forthcoming) [FREE Full text] [doi: 10.1002/ijc.33588] [Medline: 33818764]
2. Lam S, William L, Poon P. Acute cancer pain syndromes in palliative care emergencies. In: MacLeod R, van den Block L, editors. Textbook of Palliative Care. Cham, Switzerland: Springer International Publishing; 2018.
3. Henson LA, Maddocks M, Evans C, Davidson M, Hicks S, Higginson IJ. Palliative care and the management of common distressing symptoms in advanced cancer: pain, breathlessness, nausea and vomiting, and fatigue. J Clin Oncol 2020 Mar 20;38(9):905-914 [FREE Full text] [doi: 10.1200/jco.19.00470]
4. Caruso R, Nanni MG, Riba MB, Sabato S, Grassi L. The burden of psychosocial morbidity related to cancer: patient and family issues. Int Rev Psychiatry 2017 Oct;29(5):389-402 [FREE Full text] [doi: 10.1080/09540261.2017.1288090] [Medline: 28753076]
5. Selman LE, Brighton LJ, Sinclair S, Karvinen I, Egan R, Speck P, et al. Patients' and caregivers' needs, experiences, preferences and research priorities in spiritual care: a focus group study across nine countries. Palliat Med 2018 Jan 12;32(1):216-230 [FREE Full text] [doi: 10.1177/0269216317734954] [Medline: 29020846]
6. Cohen SR, Mount BM, Tomas JJ, Mount LF. Existential well-being is an important determinant of quality of life: evidence from the McGill quality of life questionnaire. Cancer 1996 Feb 1;77(3):576-586 [FREE Full text] [doi: 10.1002/(SICI)1097-0142(19960201)77:3<576::AID-CNCR22>3.0.CO;2-0]
7. Bos FM, Schoevers RA, aan het Rot M. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. Eur Neuropsychopharmacol 2015 Nov;25(11):1853-1864 [FREE Full text] [doi: 10.1016/j.euroneuro.2015.08.008] [Medline: 26336868]
8. van Knippenberg RJ, de Vugt ME, Ponds RW, Myin-Germeys I, van Twillert B, Verhey FR. Dealing with daily challenges in dementia (deal-id study): an experience sampling study to assess caregiver functioning in the flow of daily life. Int J Geriatr Psychiatry 2017 Sep;32(9):949-958 [FREE Full text] [doi: 10.1002/gps.4552] [Medline: 27464472]
9. Bolger N, Laurenceau JP. Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research. New York, NY: Guilford Press; 2013.
10. Aiello EJ, Taplin S, Reid R, Hobbs M, Seger D, Kamel H, et al. In a randomized controlled trial, patients preferred electronic data collection of breast cancer risk-factor information in a mammography setting. J Clin Epidemiol 2006 Jan;59(1):77-81 [FREE Full text] [doi: 10.1016/j.jclinepi.2005.07.007] [Medline: 16360564]
11. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. Control Clin Trial 2003 Apr;24(2):182-199 [FREE Full text] [doi: 10.1016/S0197-2456(02)00320-3]
12. Kampshoff CS, Verdonck-de Leeuw IM, van Oijen MG, Sprangers MA, Buffart LM. Ecological momentary assessments among patients with cancer: a scoping review. Eur J Cancer Care (Engl) 2019 May 14;28(3):e13095 [FREE Full text] [doi: 10.1111/ecc.13095] [Medline: 31090160]
13. Jim HS, Hoogland AI, Brownstein NC, Barata A, Dicker AP, Knoop H, et al. Innovations in research and clinical care using patient-generated health data. CA Cancer J Clin 2020 May;70(3):182-199 [FREE Full text] [doi: 10.3322/caac.21608] [Medline: 32311776]
14. Paterson C, Armitage L, Turner M. Current landscape of ecological momentary assessment (real-time data) methodology in cancer research: a systematic review. Semin Oncol Nurs 2023 Dec;39(6):151514 [FREE Full text] [doi: 10.1016/j.soncn.2023.151514] [Medline: 37865555]
15. Geeraerts J, Pivodic L, De Nooijer K, Naert E, Crombez G, De Ridder M, et al. Investigating experiences of people with advanced breast or lung cancer in their natural context: protocol for an experience sampling study. BMJ Open 2024 Feb 14;14(2):e075752 [FREE Full text] [doi: 10.1136/bmjopen-2023-075752] [Medline: 38355197]
16. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
17. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]

XSL•FO
RenderX

19. EndNote: Office 2013 compatibility. Clarivate. 2022 Jun 15. URL: https://support.clarivate.com/Endnote/s/article/EndNote-Office-20132016-Compatibility?language=en_US [accessed 2024-05-24]
20. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
21. Çınar D, Karadakovan A, Erdoğan AP. Effect of mobile phone app-based training on the quality of life for women with breast cancer. Eur J Oncol Nurs 2021 Jun;52:101960 [FREE Full text] [doi: 10.1016/j.ejon.2021.101960] [Medline: 33882446]
22. Aigner CJ, Cinciripini PM, Anderson KO, Baum GP, Gritz ER, Lam CY. The association of pain with smoking and quit attempts in an electronic diary study of cancer patients trying to quit. Nicotine Tob Res 2016 Jun;18(6):1449-1455 [FREE Full text] [doi: 10.1093/ntr/ntv118] [Medline: 26038362]
23. Chumbler NR, Mkanta WN, Richardson LC, Harris L, Darkins A, Kobb R, et al. Remote patient-provider communication and quality of life: empirical test of a dialogic model of cancer care. J Telemed Telecare 2007;13(1):20-25 [FREE Full text] [doi: 10.1258/135763307779701112] [Medline: 17288654]
24. Harper FW, Heath EI, Gleason ME, Penner L, Lorusso P, Wang D, et al. Physicians' use of patients' daily reports of quality of life to evaluate treatment response in phase I cancer trials. J Cancer Ther 2012 Oct;3(5):582-588 [FREE Full text] [doi: 10.4236/jct.2012.35074] [Medline: 24058742]
25. Langer SL, Romano JM, Todd M, Strauman TJ, Keefe FJ, Syrjala KL, et al. Links between communication and relationship satisfaction among patients with cancer and their spouses: results of a fourteen-day smartphone-based ecological momentary assessment study. Front Psychol 2018;9:1843 [FREE Full text] [doi: 10.3389/fpsyg.2018.01843] [Medline: 30364167]
26. Mooney KH, Beck SL, Friedman RH, Farzanfar R, Wong B. Automated monitoring of symptoms during ambulatory chemotherapy and oncology providers' use of the information: a randomized controlled clinical trial. Support Care Cancer 2014 Sep;22(9):2343-2350 [FREE Full text] [doi: 10.1007/s00520-014-2216-1] [Medline: 24687538]
27. van den Berg L, Brouwer P, Panda N, Hoogbergen M, Solsky I, Onnela JP, et al. Feasibility and performance of smartphone-based daily micro-surveys among patients recovering from cancer surgery. Qual Life Res 2022 Feb;31(2):579-587 [FREE Full text] [doi: 10.1007/s11136-021-02934-x] [Medline: 34283380]
28. Kearney N, Kidd L, Miller M, Sage M, Khorrami J, McGee M, et al. Utilising handheld computers to monitor and support patients receiving chemotherapy: results of a UK-based feasibility study. Support Care Cancer 2006 Jul;14(7):742-752 [FREE Full text] [doi: 10.1007/s00520-005-0002-9] [Medline: 16525792]
29. Maguire R, Miller M, Sage M, Norrie J, McCann L, Taylor L, et al. Results of a UK based pilot study of a mobile phone based advanced symptom management system (ASyMS) in the remote monitoring of chemotherapy related toxicity. Clin Eff Nurs 2005 Sep;9(3-4):202-210 [FREE Full text] [doi: 10.1016/j.cein.2006.08.013]
30. McCall K, Keen J, Farrer K, Maguire R, McCann L, Johnston B, et al. Perceptions of the use of a remote monitoring system in patients receiving palliative care at home. Int J Palliat Nurs 2008 Sep;14(9):426-431 [FREE Full text] [doi: 10.12968/ijpn.2008.14.9.31121] [Medline: 19060793]
31. McCann L, Maguire R, Miller M, Kearney N. Patients' perceptions and experiences of using a mobile phone-based advanced symptom management system (ASyMS) to monitor and manage chemotherapy related toxicity. Eur J Cancer Care (Engl) 2009 Mar;18(2):156-164 [FREE Full text] [doi: 10.1111/j.1365-2354.2008.00938.x] [Medline: 19267731]
32. Kearney N, McCann L, Norrie J, Taylor L, Gray P, McGee-Lennon M, et al. Evaluation of a mobile phone-based, advanced symptom management system (ASyMS) in the management of chemotherapy-related toxicity. Support Care Cancer 2009 Apr;17(4):437-444 [FREE Full text] [doi: 10.1007/s00520-008-0515-0] [Medline: 18953579]
33. Weaver A, Love SB, Larsen M, Shanyinde M, Waters R, Grainger L, et al. A pilot study: dose adaptation of capecitabine using mobile phone toxicity monitoring - supporting patients in their homes. Support Care Cancer 2014 Oct;22(10):2677-2685 [FREE Full text] [doi: 10.1007/s00520-014-2224-1] [Medline: 24771299]
34. Besse KT, Faber-te Boveldt ND, Janssen GH, Vernooij-Dassen M, Vissers KC, Engels Y. Pain assessment with short message service and interactive voice response in outpatients with cancer and pain: a feasibility study. Pain Pract 2016 Mar;16(3):320-326 [FREE Full text] [doi: 10.1111/papr.12278] [Medline: 25581306]
35. Hachizuka M, Yoshiuchi K, Yamamoto Y, Iwase S, Nakagawa K, Kawagoe K, et al. Development of a personal digital assistant (PDA) system to collect symptom information from home hospice patients. J Palliat Med 2010 Jun;13(6):647-651 [FREE Full text] [doi: 10.1089/jpm.2009.0350] [Medline: 20509795]
36. Passardi A, Foca F, Caffo O, Tondini CA, Zambelli A, Vespignani R, et al. A remote monitoring system to optimize the home management of oral anticancer therapies (ONCO-TreC): prospective training-validation trial. J Med Internet Res 2022 Jan 26;24(1):e27349 [FREE Full text] [doi: 10.2196/27349] [Medline: 35080505]
37. Yap KY, Low HX, Koh KS, Un M, Shih V, Chan A. Feasibility and acceptance of a pharmacist-run tele-oncology service for chemotherapy-induced nausea and vomiting in ambulatory cancer patients. Telemed J E Health 2013 May;19(5):387-395 [FREE Full text] [doi: 10.1089/tmj.2012.0136] [Medline: 23531095]
38. LeBaron V, Alam R, Bennett R, Blackhall L, Gordon K, Hayes J, et al. Deploying the behavioral and environmental sensing and intervention for cancer smart health system to support patients and family caregivers in managing pain: feasibility and acceptability study. JMIR Cancer 2022 Aug 09;8(3):e36879 [FREE Full text] [doi: 10.2196/36879] [Medline: 35943791]

XSL•FO
RenderX

39. LeBaron V, Homdee N, Ogunjirin E, Patel N, Blackhall L, Lach J. Describing and visualizing the patient and caregiver experience of cancer pain in the home context using ecological momentary assessments. Digit Health 2023;9:20552076231194936 [FREE Full text] [doi: 10.1177/20552076231194936] [Medline: 37654707]
40. Coolbrandt A, Muylaert K, Vandeneede E, Dooms C, Wildiers H. Remote system for daily symptom monitoring during systemic anticancer treatment. Cancer Nurs 2021 Aug 5;45(4):E758-E765 [FREE Full text] [doi: 10.1097/ncc.0000000000000999]
41. Lee MK, Basch E, Mitchell SA, Minasian LM, Langlais BT, Thanarajasingam G, et al. Reliability and validity of PRO-CTCAE® daily reporting with a 24-hour recall period. Qual Life Res 2023 Jul;32(7):2047-2058 [FREE Full text] [doi: 10.1007/s11136-023-03374-5] [Medline: 36897529]
42. Nordhausen T, Lampe K, Vordermark D, Holzner B, Al-Ali HK, Meyer G, et al. An implementation study of electronic assessment of patient-reported outcomes in inpatient radiation oncology. J Patient Rep Outcomes 2022 Jul 19;6(1):77 [FREE Full text] [doi: 10.1186/s41687-022-00478-3] [Medline: 35852715]
43. Schuler T, King C, Matsveru T, Back M, Clark K, Chin D, et al. Wearable-triggered ecological momentary assessments are feasible in people with advanced cancer and their family caregivers: feasibility study from an outpatient palliative care clinic at a cancer center. J Palliat Med 2023 Jul;26(7):980-985 [FREE Full text] [doi: 10.1089/jpm.2022.0535] [Medline: 37134212]
44. van Roozendaal AS, Schellekens MP, van Woezik RA, van der Lee ML. Exploring the dynamic interconnectedness of protective and perpetuating factors of cancer-related fatigue. Psychooncology 2023 Dec;32(12):1876-1884 [FREE Full text] [doi: 10.1002/pon.6235] [Medline: 37902175]
45. Badr H, Laurenceau JP, Schart L, Basen-Engquist K, Turk D. The daily impact of pain from metastatic breast cancer on spousal relationships: a dyadic electronic diary study. Pain 2010 Dec;151(3):644-654 [FREE Full text] [doi: 10.1016/j.pain.2010.08.022] [Medline: 20833474]
46. Badr H, Pasipanodya EC, Laurenceau JP. An electronic diary study of the effects of patient avoidance and partner social constraints on patient momentary affect in metastatic breast cancer. Ann Behav Med 2013 Apr;45(2):192-202 [FREE Full text] [doi: 10.1007/s12160-012-9436-8] [Medline: 23150234]
47. Stephenson E, DeLongis A, Bruel B, Badr H. Outpatient pain medication use: an electronic daily diary study in metastatic breast cancer. J Pain Symptom Manage 2018 Apr;55(4):1131-1137 [FREE Full text] [doi: 10.1016/j.jpainsymman.2017.11.032] [Medline: 29221847]
48. Belcher AJ, Laurenceau JP, Graber EC, Cohen LH, Dasch KB, Siegel SD. Daily support in couples coping with early stage breast cancer: maintaining intimacy during adversity. Health Psychol 2011 Nov;30(6):665-673 [FREE Full text] [doi: 10.1037/a0024705] [Medline: 21823795]
49. Pasipanodya EC, Parrish BP, Laurenceau JP, Cohen LH, Siegel SD, Graber EC, et al. Social constraints on disclosure predict daily well-being in couples coping with early-stage breast cancer. J Fam Psychol 2012 Aug;26(4):661-667 [FREE Full text] [doi: 10.1037/a0028655] [Medline: 22686265]
50. Otto AK, Laurenceau JP, Siegel SD, Belcher AJ. Capitalizing on everyday positive events uniquely predicts daily intimacy and well-being in couples coping with breast cancer. J Fam Psychol 2015 Feb;29(1):69-79 [FREE Full text] [doi: 10.1037/fam0000042] [Medline: 25528074]
51. Cai L, Boukhechba M, Gerber MS, Barnes LE, Showalter SL, Cohn WF, et al. An integrated framework for using mobile sensing to understand response to mobile interventions among breast cancer patients. Smart Health 2020 Mar;15:100086 [FREE Full text] [doi: 10.1016/j.smhl.2019.100086]
52. Carson JW, Carson KM, Olsen M, Sanders L, Westbrook K, Keefe FJ, et al. Yoga practice predicts improvements in day-to-day pain in women with metastatic breast cancer. J Pain Symptom Manage 2021 Jun;61(6):1227-1233 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.10.009] [Medline: 33065208]
53. Dasch KB, Cohen LH, Belcher A, Laurenceau JP, Kendall J, Siegel S, et al. Affective differentiation in breast cancer patients. J Behav Med 2010 Dec;33(6):441-453 [FREE Full text] [doi: 10.1007/s10865-010-9274-8] [Medline: 20585976]
54. Pinto BM, Kindred MD, Dunsiger SI, Williams DM. Sedentary behavior among breast cancer survivors: a longitudinal study using ecological momentary assessments. J Cancer Surviv 2021 Aug;15(4):546-553 [FREE Full text] [doi: 10.1007/s11764-020-00948-x] [Medline: 33029754]
55. Ratcliff CG, Lam CY, Arun B, Valero V, Cohen L. Ecological momentary assessment of sleep, symptoms, and mood during chemotherapy for breast cancer. Psychooncology 2014 Nov;23(11):1220-1228 [FREE Full text] [doi: 10.1002/pon.3525] [Medline: 24706506]
56. Solk P, Gavin K, Fanning J, Welch W, Lloyd G, Cottrell A, et al. Feasibility and acceptability of intensive longitudinal data collection of activity and patient-reported outcomes during chemotherapy for breast cancer. Qual Life Res 2019 Dec;28(12):3333-3346 [FREE Full text] [doi: 10.1007/s11136-019-02278-7] [Medline: 31493269]
57. Phillips SM, Welch WA, Fanning J, Santa-Maria CA, Gavin KL, Auster-Gussman LA, et al. Daily physical activity and symptom reporting in breast cancer patients undergoing chemotherapy: an intensive longitudinal examination. Cancer Epidemiol Biomarkers Prev 2020 Dec;29(12):2608-2616 [FREE Full text] [doi: 10.1158/1055-9965.EPI-20-0659] [Medline: 32994340]

XSL•FO
RenderX

58. Auster-Gussman LA, Gavin KL, Siddique J, Welch WA, Solk P, Whitaker M, et al. Social cognitive variables and physical activity during chemotherapy for breast cancer: an intensive longitudinal examination. Psychooncology 2022 Mar;31(3):425-435 [FREE Full text] [doi: 10.1002/pon.5820] [Medline: 34546611]
59. Stone AA, Broderick JE, Junghaenel DU, Schneider S, Schwartz JE. PROMIS fatigue, pain intensity, pain interference, pain behavior, physical function, depression, anxiety, and anger scales demonstrate ecological validity. J Clin Epidemiol 2016 Jun;74:194-206 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.08.029] [Medline: 26628334]
60. Kim J, Lim S, Min YH, Shin YW, Lee B, Sohn G, et al. Depression screening using daily mental-health ratings from a smartphone application for breast cancer patients. J Med Internet Res 2016 Aug 04;18(8):e216 [FREE Full text] [doi: 10.2196/jmir.5598] [Medline: 27492880]
61. Min YH, Lee JW, Shin YW, Jo MW, Sohn G, Lee JH, et al. Daily collection of self-reporting sleep disturbance data via a smartphone app in breast cancer patients receiving chemotherapy: a feasibility study. J Med Internet Res 2014 May 23;16(5):e135 [FREE Full text] [doi: 10.2196/jmir.3421] [Medline: 24860070]
62. Sztachańska J, Krejtz I, Nezlek JB. Using a gratitude intervention to improve the lives of women with breast cancer: a daily diary study. Front Psychol 2019;10:1365 [FREE Full text] [doi: 10.3389/fpsyg.2019.01365] [Medline: 31249544]
63. Xu W, Wang J, Schoebi D. The role of daily couple communication in the relationship between illness representation and fear of cancer recurrence in breast cancer survivors and their spouses. Psychooncology 2019 Jun;28(6):1301-1307 [FREE Full text] [doi: 10.1002/pon.5082] [Medline: 30946501]
64. Lim E, Boyle F, Okera M, Loi S, Goksu SS, van Hal G, et al. An open label, randomized phase 2 trial assessing the impact of food on the tolerability of abemaciclib in patients with advanced breast cancer. Breast Cancer Res Treat 2022 Oct;195(3):275-287 [FREE Full text] [doi: 10.1007/s10549-022-06690-5] [Medline: 35915198]
65. Welch WA, Solk P, Auster-Gussman L, Whitaker M, Siddique J, Fanning J, et al. Longitudinal sedentary time and symptoms in breast cancer patients during chemotherapy using ecological momentary assessment. Med Sci Sports Exerc 2023 May 01;55(5):966-974. [doi: 10.1249/MSS.0000000000003115] [Medline: 36574735]
66. Whitaker M, Welch WA, Fanning J, Santa-Maria CA, Auster-Gussman LA, Solk P, et al. Using ecological momentary assessment to understand associations between daily physical activity and symptoms in breast cancer patients undergoing chemotherapy. Support Care Cancer 2022 Aug;30(8):6613-6622 [FREE Full text] [doi: 10.1007/s00520-022-07071-w] [Medline: 35488902]
67. Shiyko MP, Siembor B, Greene PB, Smyth J, Burkhalter JE. Intra-individual study of mindfulness: ecological momentary perspective in post-surgical lung cancer patients. J Behav Med 2019 Feb;42(1):102-110 [FREE Full text] [doi: 10.1007/s10865-018-9942-7] [Medline: 29992367]
68. Shiyko MP, Burkhalter J, Li R, Park BJ. Modeling nonlinear time-dependent treatment effects: an application of the generalized time-varying effect model (TVEM). J Consult Clin Psychol 2014 Oct;82(5):760-772 [FREE Full text] [doi: 10.1037/a0035267] [Medline: 24364799]
69. Steffen LE, Vowles KE, Smith BW, Gan GN, Edelman MJ. Daily diary study of hope, stigma, and functioning in lung cancer patients. Health Psychol 2018 Mar;37(3):218-227 [FREE Full text] [doi: 10.1037/hea0000570] [Medline: 29172604]
70. Steffen LE, Cheavens JS, Vowles KE, Gabbard J, Nguyen H, Gan GN, et al. Hope-related goal cognitions and daily experiences of fatigue, pain, and functional concern among lung cancer patients. Support Care Cancer 2020 Feb;28(2):827-835 [FREE Full text] [doi: 10.1007/s00520-019-04878-y] [Medline: 31152302]
71. Maguire R, Ream E, Richardson A, Connaghan J, Johnston B, Kotronoulas G, et al. Development of a novel remote patient monitoring system: the advanced symptom management system for radiotherapy to improve the symptom experience of patients with lung cancer receiving radiotherapy. Cancer Nurs 2015;38(2):E37-E47 [FREE Full text] [doi: 10.1097/NCC.0000000000000150] [Medline: 24836956]
72. Dunsmore VJ, Neupert SD. Coping with 'scanxiety': within-person processes in lung cancer. Psychol Rep 2023 Mar 25:332941231164336 [FREE Full text] [doi: 10.1177/00332941231164336] [Medline: 36964680]
73. Coolbrandt A, Muylaert K, Vandeneede E, Dooms C, Wildiers H. Real-time symptom management in the context of a remote symptom-monitoring system: prospective process evaluation and cross-sectional survey to explore clinical relevance. Support Care Cancer 2021 Jun;29(6):3401-3408 [FREE Full text] [doi: 10.1007/s00520-021-06029-8] [Medline: 33515304]
74. Trotti A, Byhardt R, Stetz J, Gwede C, Corn B, Fu K, et al. Common toxicity criteria: version 2.0. an improved reference for grading the acute effects of cancer treatment: impact on radiotherapy. Int J Radiat Oncol Biol Phys 2000 Apr 01;47(1):13-47 [FREE Full text] [doi: 10.1016/s0360-3016(99)00559-3] [Medline: 10758303]
75. Fayers P, Bottomley A, EORTC Quality of Life Group, Quality of Life Unit. Quality of life research within the EORTC-the EORTC QLQ-C30. European Organisation for Research and Treatment of Cancer. Eur J Cancer 2002 Mar;38 Suppl 4:S125-S133 [FREE Full text] [doi: 10.1016/s0959-8049(01)00448-8] [Medline: 11858978]
76. Watson D, Clark LA. The PANAS-X: manual for the positive and negative affect schedule - expanded form. The University of Iowa. 1994. URL: https://iro.uiowa.edu/esploro/outputs/other/9983557488402771#file-0 [accessed 2024-05-25]
77. Kim J, Marcusson-Clavertz D, Yoshiuchi K, Smyth JM. Potential benefits of integrating ecological momentary assessment data into mHealth care systems. Biopsychosoc Med 2019;13:19 [FREE Full text] [doi: 10.1186/s13030-019-0160-5] [Medline: 31413726]

XSL•FO
RenderX

78. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care 2004 Dec;42(12):1194-1201 [FREE Full text] [doi: 10.1097/00005650-200412000-00006] [Medline: 15550799]
79. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992 Jun;30(6):473-483. [Medline: 1593914]
80. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
81. Trull TJ, Ebner-Priemer UW. Ambulatory assessment in psychopathology research: a review of recommended reporting guidelines and current practices. J Abnorm Psychol 2020 Jan;129(1):56-63 [FREE Full text] [doi: 10.1037/abn0000473] [Medline: 31868388]
82. Eisele G, Vachon H, Lafit G, Kuppens P, Houben M, Myin-Germeys I, et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. PsyArXiv Preprints. Preprint posted online February 20, 2020 [FREE Full text] [doi: 10.31234/osf.io/zf4nm]
83. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32 [FREE Full text] [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
84. Giesinger JM, Wintner LM, Oberguggenberger AS, Gamper EM, Fiegl M, Denz H, et al. Quality of life trajectory in patients with advanced cancer during the last year of life. J Palliat Med 2011 Aug;14(8):904-912 [FREE Full text] [doi: 10.1089/jpm.2011.0086] [Medline: 21711125]
85. Hwang SS, Chang VT, Fairclough DL, Cogswell J, Kasimis B. Longitudinal quality of life in advanced cancer patients: pilot study results from a VA medical cancer center. J Pain Symptom Manage 2003 Mar;25(3):225-235 [FREE Full text] [doi: 10.1016/s0885-3924(02)00641-3] [Medline: 12614957]
86. Geeraerts J, Kuppens P. Investigating careless responding detection techniques in experience sampling methods (ESM). OSF. URL: https://osf.io/8ymh5/download [accessed 2024-05-25]
87. van Roij J, Fransen H, van de Poll-Franse L, Zijlstra M, Raijmakers N. Measuring health-related quality of life in patients with advanced cancer: a systematic review of self-administered measurement instruments. Qual Life Res 2018 Aug;27(8):1937-1955 [FREE Full text] [doi: 10.1007/s11136-018-1809-4] [Medline: 29427216]
88. Miranda R, Raemdonck E, Deliens L, Kaasa S, Zimmermann C, Rodin G, et al. Do cancer centres and palliative care wards routinely measure patients' quality of life? An international cross-sectional survey study. Support Care Cancer 2023 Jul 31;31(8):499 [FREE Full text] [doi: 10.1007/s00520-023-07964-4] [Medline: 37523097]

Abbreviations

EMA: ecological momentary assessment
EORTC-QLQ-C30: EORTC Core Quality of Life Questionnaire
ESM: experience sampling methods
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by T de Azevedo Cardoso; submitted 23.06.23; peer-reviewed by C Ratcliff, M Schellekens, J Dekker; comments to author 17.01.24; revised version received 12.03.24; accepted 29.04.24; published 12.06.24.

Please cite as:
Geeraerts J, de Nooijer K, Pivodic L, De Ridder M, Van den Block L
Intensive Longitudinal Methods Among Adults With Breast or Lung Cancer: Scoping Review
J Med Internet Res 2024;26:e50224
URL: https://www.jmir.org/2024/1/e50224
doi:10.2196/50224
PMID:38865186

©Joran Geeraerts, Kim de Nooijer, Lara Pivodic, Mark De Ridder, Lieve Van den Block. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Characteristics and Effects of Home-Based Digital Health Interventions on Functional Outcomes in Older Patients With Hip Fractures After Surgery: Systematic Review and Meta-Analysis

Suphawita Pliannuom[1,2], MD; Kanokporn Pinyopornpanish[1,2], MD, PhD; Nida Buawangpong[1,2], MD; Nutchar Wiwatkunupakarn[1,2], MD; Poppy Alice Carson Mallinson[3], MSc, PhD; Wichuda Jiraporncharoen[1,2], MD, MSc; Chaisiri Angkurawaranon[1,2], MD, PhD

[1]Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

[2]Global Health and Chronic Conditions Research Group, Chiang Mai University, Chiang Mai, Thailand

[3]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom

Corresponding Author:
Kanokporn Pinyopornpanish, MD, PhD
Department of Family Medicine
Faculty of Medicine
Chiang Mai University
110 Inthawarorot Rd., Sriphum, Muang
Chiang Mai, 50200
Thailand
Phone: 66 53935462
Email: kanokporn.pinyopo@cmu.ac.th

Abstract

Background: Digital health interventions (DHIs) have been used to improve postoperative functional ability in older patients with hip fractures. However, there is limited information on the characteristics of home-based DHIs, and controversy exists regarding their impact on functional outcomes in this population.

Objective: This study aims to provide an overview of the characteristics and effects of home-based DHIs on functional outcomes in older patients with hip fractures after surgery.

Methods: We conducted a systematic review and meta-analysis following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Five electronic medical databases (PubMed, Embase, Cochrane, ProQuest, and CINAHL) were searched up until January 3, 2023. We included clinical trials or randomized controlled trials (RCTs) in English involving home-based DHIs for postoperative care among older patients with hip fractures. Excluded studies involved patients not hospitalized, not discharged to home, not directly using DHIs, or with inaccessible full text. The PROSPERO registration number is CRD42022370550. Two independent reviewers screened and extracted data (SP and NB). Disagreements were resolved through discussion and agreement with the third author (KP). Home-based DHIs were characterized in terms of purpose and content, mode of delivery, and health care provider. Functional outcomes assessed included Timed Up and Go (TUG) test, Short Physical Performance Battery (SPPB), and Functional Independence Measure (FIM). Summary measures were calculated using mean differences with 95% CIs. Risk of bias was assessed using the Risk-of-Bias 2 assessment tool for RCTs and ROBINS-I for non-RCTs. The quality of evidence was assessed using GRADE (Grading of Recommendations Assessment, Development and Evaluation).

Results: Of 2125 identified studies, 16 were included in the systematic review, involving 1467 participants. Six studies were included in the meta-analysis (4 for TUG, 4 for SPPB, and 2 for FIM). Home-based DHIs predominantly involved communication and feedback, education, and telerehabilitation. Telephone calls were the most common mode of delivery, followed by web-based software and mobile apps. Physical therapists were the main health care providers. The meta-analysis showed that home-based DHIs improved functional outcomes compared with usual care, with decreased TUG scores (mean difference=–7.89; 95% CI –10.34 to –5.45; $P<.001$), significantly increased SPPB scores (mean difference=1.11; 95% CI 0.51-1.72; $P<.001$), and increased FIM scores (mean difference=7.98; 95% CI 5.73-10.24; $P<.001$).

XSL•FO
RenderX

Conclusions: Home-based DHIs that integrate communication and feedback, education, and telerehabilitation have demonstrated effectiveness in enhancing functional outcomes among older patients recovering from hip fractures after surgery. These interventions are commonly administered by physical therapists, who play a crucial role in facilitating and guiding the rehabilitation process. However, while the existing evidence supports the efficacy of such interventions, further research is needed to enhance our understanding and optimize the implementation of home-based DHIs for this specific population.

(J Med Internet Res 2024;26:e49482) doi:10.2196/49482

KEYWORDS

home-based intervention; digital health interventions; postoperative care; older adults; hip fracture

Introduction

Hip fractures are a common and serious health issue [1-3], ranking among the top 10 causes of disability among older individuals [4]. The global prevalence is increasing, particularly among Asian people, who have the highest rate of hip fractures [4]. The mortality rate after they receive aggressive management in both surgery and rehabilitation is high, 10% in the first month [2,4] and 12%-36% in the first year [4,5]. Moreover, they have many consequences including functional decline [1], fear of falling [6], and reduced quality of life [2,4,5], and can lead to caregiver burden [1].

Apart from early surgery that can reduce mortality [7], the postoperative or transitional care after hospital discharge to home is also a crucial phase in the care process for patients with hip fracture [8]. The goal of treatment in this population is to improve functional outcomes, enhance independence, and reduce the risk of recurrent falls following surgery. Several home-based interventions, such as rehabilitation [5,9], dietary supplementation to prevent calorie and protein malnutrition [5], home hazard modification [5], fall prevention guidance [5,10], and timely and appropriate patient follow-up (including telephone calls and home visits) [5], as well as patient, family, and caregiver education and support [11], have been found beneficial in postoperative care among older patients with hip fractures. These interventions play a crucial role in improving patient functional ability, enhancing quality of life, reducing unplanned readmissions, minimizing complications and disabilities, and decreasing mortality [12].

Previous research found that combining postsurgical interventions for patients with fragility hip fractures across inpatient and outpatient settings may be able to improve physical function recovery and discovered that 63% of the studies were related to rehabilitation or medication or nutrition supplementation [13].

Over the past couple of decades, advancements in information and communication technologies have brought medical services to virtually all corners of the world [14]. Digital health interventions (DHIs) or the use of digital technology for health [15] has become a field of practice for using routine and innovative forms of information and communication technologies to address health needs [15]. DHIs have the potential to help address problems such as distance and access [15] and have been incrementally changing. DHIs do not only assist health care providers and patients receiving treatment but also benefit perfectly healthy people by providing health assessment [14] and health promotion. A recent review of home-based exercise programs delivered through DHIs for community-dwelling adults older than 65 years, involving a diverse study population with various health conditions, demonstrated improved physical function, particularly in terms of lower extremity strength [16]. Furthermore, many DHIs are specifically designed to detect, monitor, and provide care for specific health conditions [17], aiming to enhance various health aspects at home, even among older patients with hip fractures [18,19]. Previous studies have explored various types of home-based DHIs [18,19] with a focus on different interventions, such as rehabilitation, nutrition, and fall prevention, as well as different outcomes, including functional improvements, adherence to protocols, and the incidence of recurrent falls [9,19,20]. Typically, these previous studies have provided interventions to improve functional outcomes, and many of them have achieved success [16,19,21,22].

The scoping review elucidates the characteristics of DHIs for older adults who have undergone hip fracture surgery, with most studies focusing on physical therapy [23]. These DHIs were typically implemented in acute hospital settings and rehabilitation centers. However, there is a notable lack of information regarding DHIs specifically designed to improve functional outcomes in this population during transitional care settings. As a result, information on the characteristics of home-based DHIs in postoperative care among older patients with hip fractures is currently sparse, and there is a lack of information about specific home-based DHIs for this population [24]. Moreover, there is controversy over the conclusion that home-based DHIs have an impact on functional outcomes. Our systematic review and meta-analysis aimed to provide a general overview of the characteristics and effects of home-based DHIs in postoperative care on functional outcomes among older patients with hip fractures. This will lead to the development of home-based DHIs that can provide care for this population.

Methods

Ethical Considerations

We conducted a systematic review and meta-analysis and then reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Multimedia Appendix 1) [25]. This study was approved by the Chiang Mai University Ethics Committee (FAM-2565-09330) and we registered our protocol in PROSPERO (CRD42022370550).

XSL•FO
RenderX

Recruitment

Review Eligibility Criteria

We included the studies that related to (1) older adults (including patients aged 60 years or older) with hip fractures defined as any type of hip fracture, such as femoral neck, intertrochanteric, subtrochanteric fracture, and nonspecified type of hip fracture; (2) postoperative care (including transitional care, postdischarge care, and subacute care); and (3) home-based DHIs. In addition, the included studies that had (4) study design limited to clinical trials or randomized controlled trials (RCTs) and (5) compared between home-based DHIs (intervention) and usual care (comparator) in the English language with no date restriction. The following studies were excluded: patients enrolled in the emergency department but not admitted to the hospital, patients discharged from the hospital to a setting other than home, studies that involved DHIs that did not directly intervene with patients (including database management), and studies with inaccessible full-text or insufficient data for evaluating the characteristics of DHIs or quantitative synthesis.

Search Strategies

This study was systematically searched in 5 electronic medical databases, that is, PubMed, Embase, Cochrane, ProQuest, and CINAHL, with no date restriction. We searched until January 3, 2023. We searched in 3 heading terms including "hip fracture" AND "digital health" AND "post-operative" (Table S1 in Multimedia Appendix 2).

Review Selection

Duplicates were removed automatically by the EndNote X9 program (Clarivate Analytics) and the Rayyan website, as well as manually by SP and NB. Then, 2 authors (SP and NW) independently reviewed and screened all studies in title and abstract for study eligibility and 2 reviewers (SP and NB) independently screened the full paper for study selection. Disagreements were resolved through discussion and agreement with the third author (KP).

Data Extraction

Two authors (SP and NB) independently extracted data from the full original studies using a standard data collection form. Disagreements were resolved through discussion and agreement with the third author (KP). The data extracted included (1) the first author, (2) publication year, (3) study design, (4) study location, (5) population (including the number of patients, age, and sex), (6) the characteristic of home-based DHIs, (7) control group, (8) functional outcomes, and (9) other outcomes.

Characteristics of Home-Based DHIs

The characteristics of home-based DHIs focused on three dimensions, which included (1) purpose and content, (2) mode of delivery, and (3) health care provider. Two authors categorized subgroups in each dimension according to the definition (Table S2 in Multimedia Appendix 2) [17,26,27]. The categorizations possibly reported more than 1 category in each dimension. Purpose and content were categorized into 5 categories including education, telerehabilitation, communication and feedback, behavioral and lifestyle interventions or health coaching, and remote monitoring [17,26,27]. Modes of delivery were categorized into 4 categories including web-based software, telephone call, mobile apps, and sensor-based technology [17,26,27]. Health care providers were categorized into 6 categories including occupational therapists (OTs), physical therapists or physiotherapists (PTs), physicians, nurses, dieticians, and multidisciplinary teams. If the results were unable to be meta-analyzed, we conducted a narrative synthesis to provide an overview of the characteristics and effects of home-based DHIs on functional outcomes in this population.

Outcomes

In the meta-analysis, the studies that reported functional outcomes including Timed Up and Go (TUG) test, Short Physical Performance Battery (SPPB), and Functional Independence Measure (FIM) were considered. TUG is a tool that assesses functional walking ability and predicts fall risk in older adults. The shorter the time taken means better the balance function and walking ability, and the lower the risk of fall [28,29]. SPPB is a comprehensive assessment index that is used to evaluate the motor ability of the older adults and to predict the risk of falls. It consists of 3 parts that measure balance test, gait speed test, and 5 sit-to-stand tests, which assess static and dynamic balance, lower limb strength, and gait speed. The SPPB score runs from 0 to 12 and the higher the score mean the better the overall somatic ability and the lower the risk of fall [28]. FIM is a postdischarge follow-up evaluation test that consists of 18 items that measure patients' somatic, verbal, social, and cognitive abilities. The score ranges from 18 to 126. Higher scores mean higher levels of functional independence and lower levels of dependence [28].

Data Analysis

In meta-analysis, the outcomes included functional outcomes (such as TUG, SPPB, and FIM) reported as mean and SD at the end of the intervention for both the intervention group and the control group. Because the same methodology and range of outcomes were used in all studies to analyze the outcomes, the summary measure for meta-analysis was the mean difference with 95% CIs. Pooled-effect estimates of each functional outcome of interest were calculated by the random-effects DerSimonian-Laird model. Statistical heterogeneity among included studies was assessed using Cochran Q test and the percentage of total variability across studies due to heterogeneity (I^2=0%-40%, low; I^2=30%-60%, moderate; I^2=50%-90%, substantial, and I^2=75%-100%, considerable) [30]. Results of the meta-analysis and forest plot were performed using a standard software package (Stata, version 16.0; StataCorp LLC). P value of <.05 was considered statistically significant.

Risks of Bias

We assessed the quality of each included study using the Risk of Bias 2 assessment tool (RoB 2; The Cochrane Collaboration) for the 14 RCT studies. In addition, 1 non-RCT and 1 quasi-experimental study were evaluated using the ROBINS-I (The Cochrane Collaboration) [31]. The results were visualized using robvis, a visualization tool [32]. This assessment was conducted independently by 2 authors, SP and NB, with any

XSL•FO
RenderX

disagreements resolved through discussion and agreement with coauthor KP.

Grading the Quality of Evidence

We graded the quality of evidence using GRADE (Grading of Recommendations Assessment, Development and Evaluation) method [33]. This assessment was conducted independently by 2 authors, SP and NB, with any disagreements resolved through discussion and agreement with the third author, KP. The quality of evidence was categorized into 4 levels: very low, low, moderate, or high.

Results

During the initial literature search, a total of 2125 studies were identified. Among them, 244 studies were clinical trials or RCTs. After screening the titles and abstracts for eligibility criteria, 218 studies were retrieved and further assessed. Subsequently, 22 studies underwent full-text screening. The reasons for excluding studies were as follows: 2 studies involved a population without hip fractures [34,35], 2 study did not include home-based DHIs [36,37], 1 study had insufficient data [38], and 1 study used the same DHIs in both groups [39]. Finally, a total of 16 studies were included in the systematic reviews (Figure 1), with a subset of 6 studies included in the meta-analysis (4 studies for TUG, 4 studies for SPPB, and 2 studies for FIM).

The characteristics of 16 studies are shown in Table 1 and Table S3 in Multimedia Appendix 2 [28,40-51]; 13 studies were 2-arm RCTs (compared DHIs with usual care), 1 study was 3-arm RCT (compared with DHIs add on cognitive behavioral therapy [CBT], only CBT, and usual care) [19], 1 study was non-RCT [52], and 1 study was quasi-experimental study [53], published between 2005 and 2022 (Figure S1 in Multimedia Appendix 2). Studies were conducted in Europe (n=7), America (n=5), and Asia (n=4). The total number of participants in the included studies was 1467.

Figure 2 presents the characteristics of home-based DHIs. In terms of the purpose and content of home-based DHIs, most of the included studies (14 of 16) used multiple categories within each study. Two studies focused solely on communication and feedback. The common categories identified in this dimension were communication and feedback (n=15), education (n=12), telerehabilitation (n=7), health coaching (n=3), and remote monitoring (n=2). Regarding the mode of delivery, the majority of studies (13 of 16) used a single mode of delivery, while 3 studies used combined modes. The reported modes of delivery included telephone calls (n=10), web-based software platforms (n=5), mobile apps (n=3), and sensor-based technologies (n=1). The involvement of health care providers in DHIs varied. PTs were involved in 6 studies, OTs in 3 studies, physicians in 2 studies, nurses in 2 studies, dieticians in 2 studies, and a multidisciplinary team in 1 study.

The majority of research indicates that home-based DHIs have shown a beneficial effect on mobility and physical function, as evidenced by improvements in scores on tests such as TUG, SPPB, and FIM [28,40-43,52]. Furthermore, these studies suggest that home-based DHIs are feasible for enhancing various aspects of hip fractures after surgery, such as improving physical function (daily activities and instrumental activities of daily living) [19,44,53], improved adherence for self-care [45], enhancing nutritional intake and status [46,47], and potentially preventing urinary incontinence and reducing its frequency [48] (Table 1).

Ten of 16 studies had low risk of bias, 4 studies had some concerns, and 2 studies had high risk of bias (Figure 3 [19,28,40-45,47-51] and Figure 4 [52,53]). In the GRADE assessment of the quality of evidence across all outcomes, findings revealed moderate quality of evidence for TUG based on 4 studies, SPPB based on 4 studies, and for FIM based on 2 studies. (Table S4 in Multimedia Appendix 2). The forest plot analysis demonstrated that home-based DHIs resulted in a significant decrease in the TUG score (mean difference=–7.89; 95% CI –10.34 to –5.45; $P<.001$), improvement in SPPB score (mean difference=1.11; 95% CI 0.51-1.72; $P<.001$), and improvement in FIM score (mean difference=7.98; 95% CI 5.73-10.24; $P<.001$). Heterogeneity was low in both TUG ($I^2=0.00\%$; $P<.53$) and FIM scores ($I^2=12.62\%$; $P<.32$). However, substantial heterogeneity was observed in SPPB scores ($I^2=61.50\%$; $P<.05$). These findings are shown in Figures 5-7, respectively.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram. DHI: digital health intervention; RCT: randomized controlled trial.

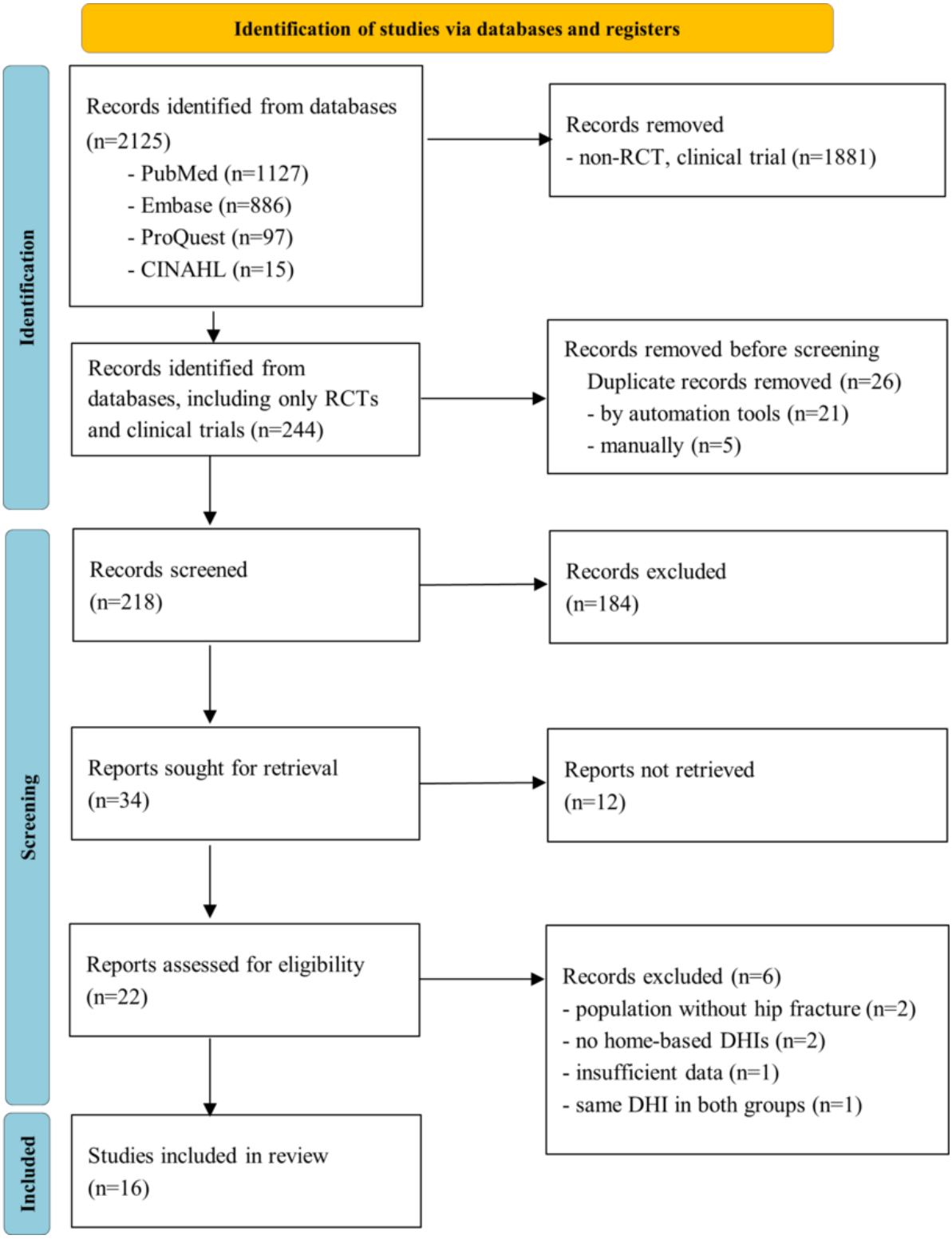

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

First author, year of publication	Population	Intervention	Purpose and content of DHIs[a]	Mode of delivery	Health care provider	Control	Functional outcomes	Conclusion
Gardner et al (2005) [49]	• 80: femoral or intertrochanteric fracture, mean age 82 years, female 78% • I[b]: 40 (end 36) • C[c]: 40 (end 36)	• Educational intervention with 5 discussion questions regarding osteoporosis treatment • Single telephone call to remind questions during a follow-up	Education, communication, and feedback	Telephone call	Primary medical physician	Brochure (methods for preventing falls)	N/A[d]	Patients provided vs not provided with information and questions were more likely to receive appropriate intervention
Krichbaum (2007) [44]	• 33: hip fracture average age 78.87 years (range 72-85 years), female 73%	• Nursing postacute care coordination intervention • First month: face-to-face then face-to-face or telephone until 6 months • Health assessment	Education, communication, and feedback	Telephone call	Gerontologic advanced practice nurse	Usual care	Intervention led to significantly improved function in most ADLs[e] and IADLs[f] measured using the functional status index score	I group had better function at 12 months on several activities and IADLs
Breedveld-Peters et al (2012) [46]	66: hip fracture, mean age 76 years (range 55-92 years), female 74%	• Dietetic counseling (telephone call 5 visits and face-to-face 5 visits) and ONS[g] for 3 months	Communication and feedback	Telephone call	Dieticians	Usual nutritional care	N/A	Face-to-face and telephone call nutritional counseling are feasible
Latham et al (2014) [41]	• 232: hip fracture, mean age 78 (SD 9.9) years, female 69% • I: 120 (end 100) • C: 112 (end 95)	• Functionally oriented exercises, home exercise program 3 times per week for 6 months • Monthly telephone call • DVD version of the program	Education, telerehabilitation, communication and feedback, and health coaching	Telephone call and web-based software	PT[h]	In-home and telephone-based cardiovascular nutrition education	I-group significantly improved in SPPB[i] between group difference 0.8 (95% CI 0.4-1.2; P<.01)	Home-based functionally oriented exercise program resulted in modest improvement in physical function at 6 months
Di Monaco et al (2015) [50]	• 169: hip fracture, mean age 78.7 (SD 7.2) years in I-group, C-group 79.3 (8.0) years • I: 84 (end 78) • C: 85 (end 75)	• Multidisciplinary rehabilitation program during hospitalization • Single telephone call to reinforce the targeted recommendations for fall prevention	Education, communication, and feedback	Telephone call	OT[j]	Multidisciplinary rehabilitation program during hospitalization	At least 1 fall during follow-up; relative risk 1.06 (95% CI 0.48-2.34)	This study did not support postdischarge single telephone call to reinforce the recommendations for fall prevention

XSL•FO
RenderX

First author, year of publication	Population	Intervention	Purpose and content of DHIs[a]	Mode of delivery	Health care provider	Control	Functional outcomes	Conclusion
Bedra and Finkelstein (2015) [53]	• 10: hip fracture, mean age 77 (SD 9) years, female 60% • Preintervention: 10 • Postintervention: 10	• Postcomprehensive telerehabilitation system to support individualized exercise program using home automated telemanagement • tailored feedback • multimedia education individualized to patient's need	Education, telerehabilitation, communication, and feedback	Web-based software	PT	Preintervention	*P* value pre- and postcomparing • Modified Barthel Index improved (*P*=.10) • LEFS[k] significantly improved (*P*=.03)	Home-based telerehabilitation may be a viable model for postacute hip fracture recovery
Langford et al (2015) [45]	• 30: hip fracture, mean age 81.5 (range 61-97) years, female 63.33% • I: 15 (end 11) • C: 15 (end 15)	• Educational (1 hour in-hospital session) plus 5 postdischarge telephone call coaching	Education, health coaching, communication and feedback	Telephone call	PT	Usual care plus 1 hour in-hospital educational session	No differences between groups for grip strength, gait speed	Telephone coaching for older adults after hip fracture to improve adherence to mobility recovery goals has feasibility
Kalron et al (2018) [40]	32: femoral neck fracture, mean age 66.2 (SD 11.6) years, female 62.5% • I: 20 (end 15) • C: 20 (end 17)	• Home-based telerehabilitation: video clips of common rehabilitation exercises focusing on the lower limbs, 40-50 minute/session	Education, tele-rehabilitation, communication, and feedback	Web-based software	PT (specialized in neurological rehabilitation)	Home-based exercise booklet	I-group greater improved significantly in 5/6 tests compare with C-group: TUG, 2-minute walk, sit to stand test, Walking speed, Mean step length	Telerehabilitation generated a positive effect on mobility in people following hip surgery
Wyers et al (2018) [47]	• 152: hip fracture, mean age 77 (SEM 1.2) years in I-group, C-group 76 (SEM 1.1) years, female 71% • I: 73 (end: 63) • C: 79 (end: 68)	• Intensive nutritional intervention • 10 counseling: 2 sessions in hospitalization and 8 sessions of weekly dietetic consultation (3 face-to-face, 5 telephone calls) • Energy-protein–enriched diet and ONS for 3 months	Communication and feedback	Telephone call	Dietician	Usual nutritional care	Does not affect functional and any outcome	Intensive nutritional intervention after hip fracture improved nutritional intake and status but not LOS[l] or clinical outcomes
Pol et al (2019) [19]	• 240: hip fracture, mean age 83.8 (SD 6.9) years, female 80%(3 arms) • CBT[m] on OT with sensor: 76 • CBT on OT: 87 • Usual 77	• CBT-based occupational therapy (weekly session coaching of skilled nursing facility, 4 home visits, 4 telephone consultation) • Sensor monitoring (physical activity monitor, motion sensors)	Remote monitoring, health coaching, communication and feedback	Sensor-based technology, telephone call	OT	• Usual care • CBT-based occupational therapy	Mean patient-reported daily functioning in the CBT-based OT with sensor monitoring was larger than usual group (difference 1.17 (95% CI 0.47-1.87; *P*=.001).	Rehabilitation program of sensor-monitoring-informed OT more effective in improving patient-reported daily functioning at 6 months than usual care

XSL•FO
RenderX

First author, year of publication	Population	Intervention	Purpose and content of DHIs[a]	Mode of delivery	Health care provider	Control	Functional outcomes	Conclusion
Pfeiffer et al (2020) [43]	• 115: hip and pelvic fracture, mean age 82.5 (SD 6.8) years, female 76% • I: 57 (end 46) • C: 58 (end 50)	• 8 individual sessions during inpatient rehabilitation • Additional support via 1 home visit, 4 telephone calls	Communication and feedback, education	Telephone call	PT or sports therapist	Geriatric inpatient rehabilitation	End point SPPB improved in favor of I-group from baseline to follow-up (2.03 vs 0.70; *P*=.002, *d*=0.58)	Intervention improved psychological and physical performance measures but did not increase daily walking duration
Ortiz-Piña et al (2021) [52]	• 71: hip fracture, mean age 75.86 (SD 5.79) years in I-group and 80.38 (SD 5.54) years in C-group, female 75.8% • I: 35 (end 28) • C: 36 (end 34)	• @ctivehip: multidisciplinary telerehabilitation program supervised by family caregivers, 50-60 minutes per session, 5 sessions per week in 2 we-based components: • 3 exercise sessions + 2 OT sessions • Videoconference (if need)	Education, telerehabilitation, communication, and feedback	Web-based software	Multidisciplinary (sport sciences professionals, PT, OT, and orthopedic surgeon consultants)	Home-based usual outpatient rehabilitation	• I-group higher FIM[n] (high effect size), better performance in TUG[o] (medium effect size) significantly compared with C-group • No statistically significant differences between group postintervention in SPPB	Telerehabilitation had better results in functional independence and physical condition (self-report and performance-based) than usual care
Córcoles-Jiménez et al, (2021) [48]	109: hip fracture, mean age 80.56 (SD 6.65) years, female 70.6% • I: 53 (end 51) • C: 56 (end 51)	• Urinary habit training (education): in-hospital stay • Telephonic reinforcement to remind the recommended activities, repeating list of activities	Communication and feedback, education	Telephone call	Nurse	Usual medical care	Mean of independence in ADL I-group 83.03 (SD 23.41), C-group 71.37 (SD 26.83) (*P*=.005)	Educational and remind system prevents the development of UI[p] and decreases the number of episodes
Zhang et al (2022) [28]	58: hip fracture, mean age 77.0 (SD 7.89) years in I-group, C-group 75.17 (SD 7.73) years, female 64.7% • I: 29 (end 27) • C: 29 (end 24)	• Home-based telerehabilitation. • Personalized rehabilitation programs, video, health knowledge • Remote-monitoring vital sign, assessment and guidance • Appointment for consultation	Education, telerehabilitation, communication and feedback, and remote monitoring	Mobile apps (patient sided), web-based software (physician side)	Physician	Former telephone follow-up at 2 weeks and 1, 2, 3, and months after discharge	• The Harris hip score, FIM of I-group higher than C-group at 1 and 3 months (*P*<.05) • TUG, SPPB I-group better than C-group at 3 months (*P*<.05)	Home-based telerehabilitation improves functional recovery of hip joint and enhances ability to perform ADL and somatic integration to a certain extent
				Mobile apps	PT			

XSL•FO
RenderX

First author, year of publication	Population	Intervention	Purpose and content of DHIs[a]	Mode of delivery	Health care provider	Control	Functional outcomes	Conclusion
Cheng et al (2022) [51]	39: hip fracture, mean age 77.4 years, female 49.7% • I: 19 • C: 20	Mobile apps in delivering home-based rehabilitation program. • Briefing session before discharge • Perform exercise along exercise video in app	Education, telerehabilitation, communication, and feedback			Home-based rehabilitation program using an exercise pamphlet	No difference in modified functional ambulatory category, elderly mobility scale, LEFS at 1 and 2 months	Mobile apps improved exercise adherence but did not improve physical performance, self-efficacy, and reduced exercise adherence
Li et al (2022) [42]	31: hip fracture, mean age 76.5 (SD 8.6) years in I-group, 82.1 (SD 9.7) years in C-group, female 80.6% • I: 15 • C: 16	Home program using the Caspar Health e-system. • Tailor-made telerehabilitation program through e-system calendar • Video, pictures, and written and verbal instructions shown on app • Patient feedback and OT update program	Education, telerehabilitation, communication and feedback	Mobile apps	OT	Home program paper sheets and log sheets	• I-group significant improvement in IADL (P=.010) at postintervention and follow-up • No significant differences in TUG, functional reach test, and modified Barthel index	Telerehabilitation via smartphone as alternative home program for use in OT practice with older after hip fracture surgery

[a]DHI: digital health intervention.
[b]I, I-group: intervention group.
[c]C, C-group: control group.
[d]N/A: not applicable.
[e]ADLs: activities of daily living.
[f]IADLs: instrument activities of daily living.
[g]ONS: oral nutritional supplement.
[h]PT: physical therapist or physiotherapist.
[i]SPPB: Short Physical Performance Battery.
[j]OT: occupational therapist.
[k]LEFS: Lower Extremity Functional Scale.
[l]LOS: length of stay.
[m]CBT: cognitive behavioral therapy.
[n]FIM: Functional Independence Measure.
[o]TUG: Timed Up and Go.
[p]UI: urinary incontinent.

XSL•FO
RenderX

Figure 2. The characteristics of home-based digital health interventions in 3 dimensions. Multidisciplinary team: 1 study (sport science, PT, OT, and orthopedic consultation). The categorizations possibly reported more than 1 category in each dimension. OT: occupational therapist; PT: physical therapist or physiotherapist.

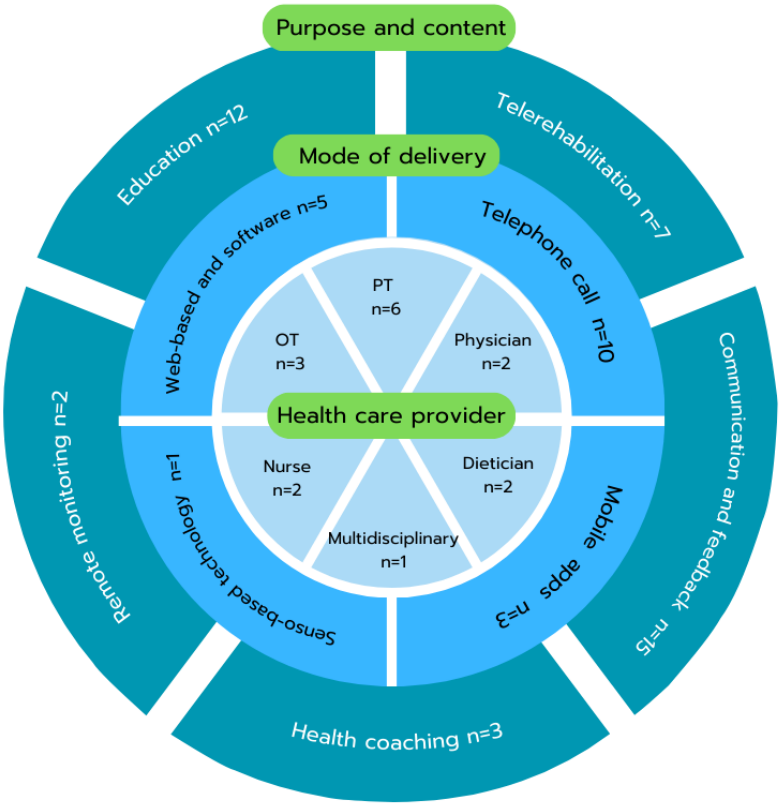

Figure 3. Risk of bias of included studies (14 randomized controlled trials).

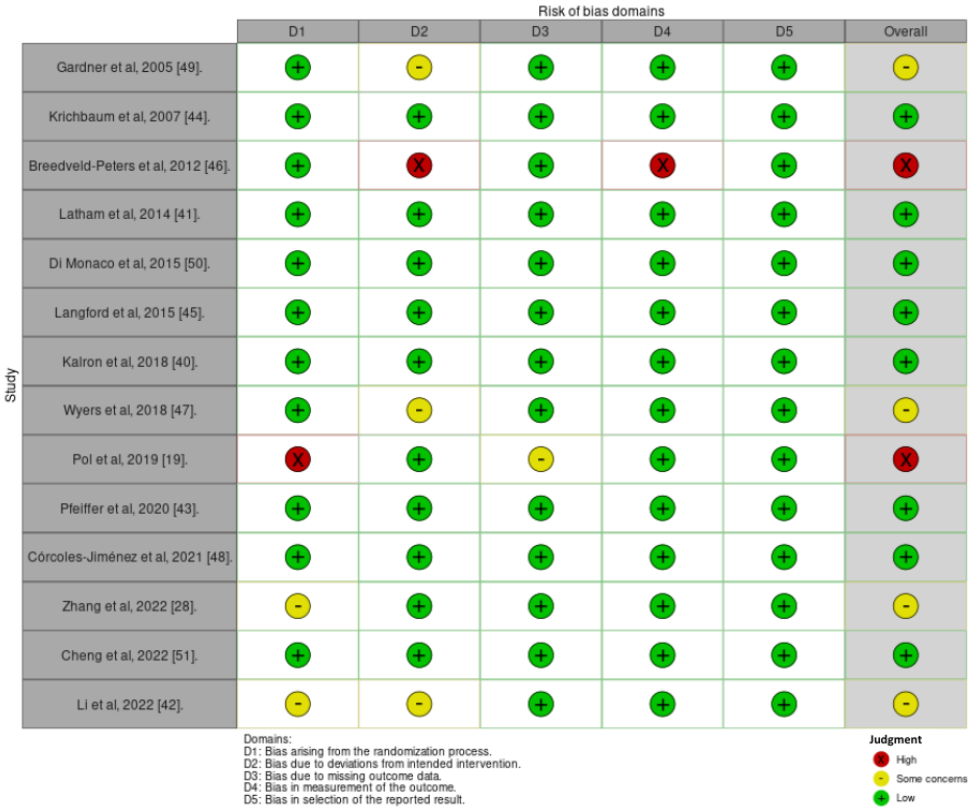

XSL•FO
RenderX

Figure 4. Risk of bias of included studies (2 non–randomized controlled trials).

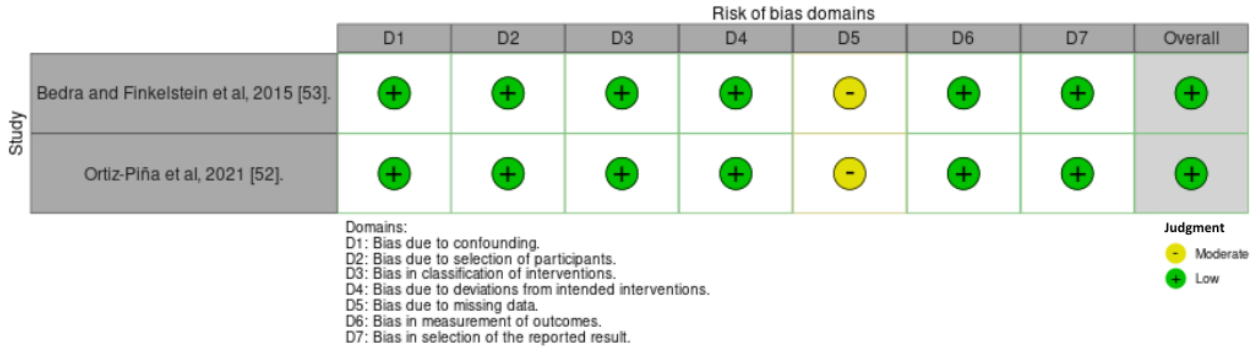

Figure 5. Summary measure Timed Up and Go for meta-analysis [28,40,42,52]. DHI: digital health intervention.

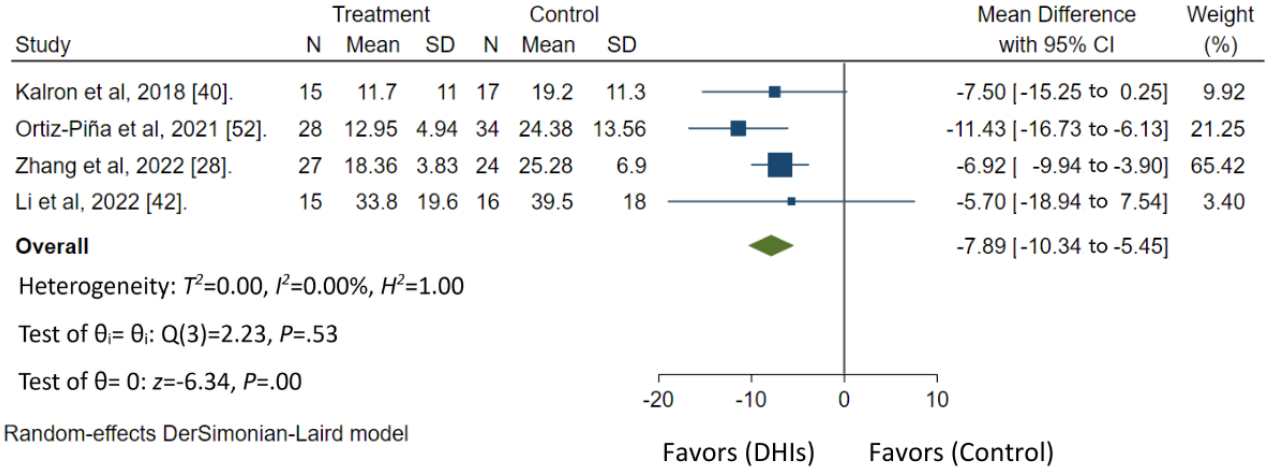

Figure 6. Summary measure Short Physical Performance Battery for meta-analysis [28,41,43,52]. DHI: digital health intervention.

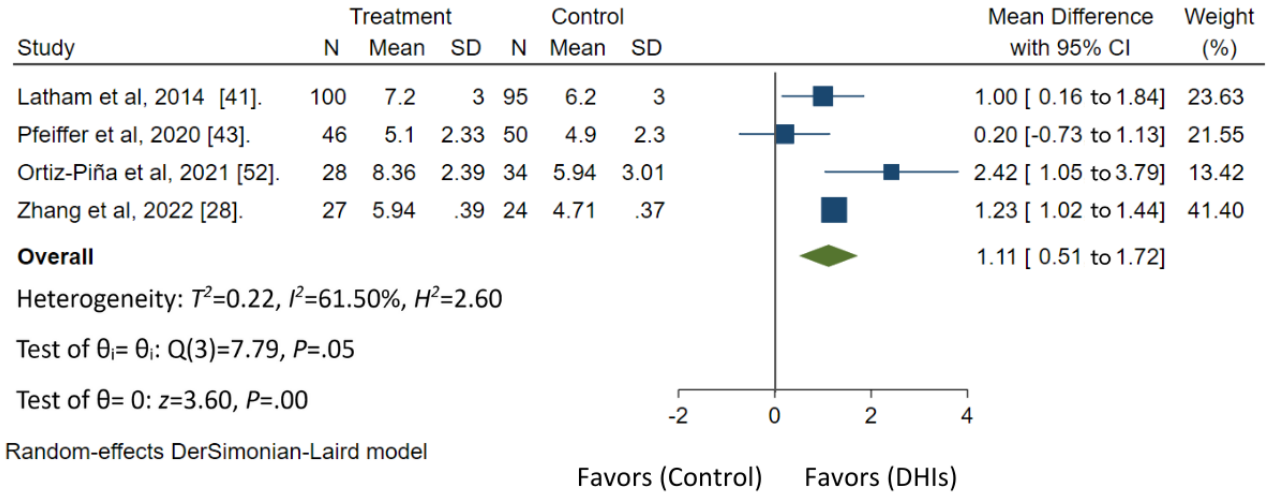

XSL•FO
RenderX

Figure 7. Summary measure Functional Independence Measure for meta-analysis [28,52]. DHI: digital health intervention.

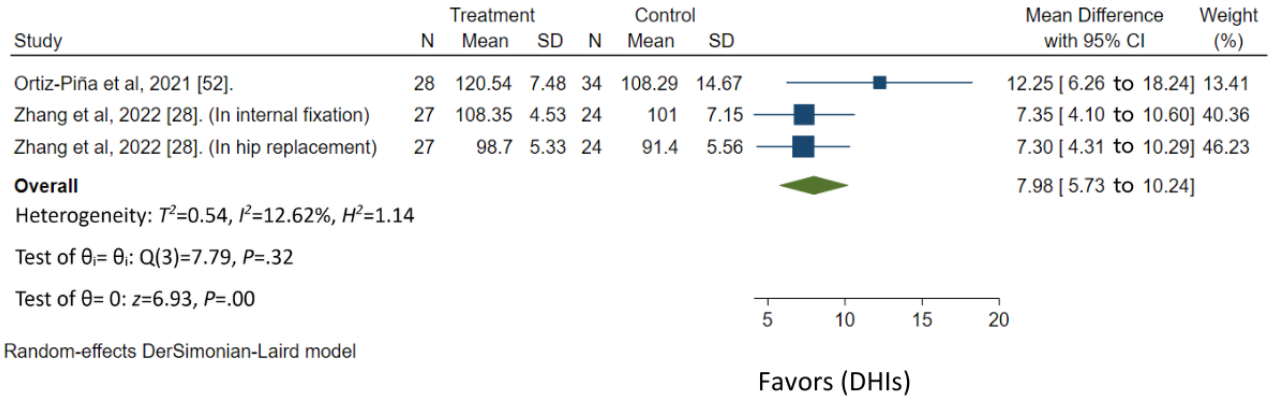

Discussion

Principal Findings

In this systematic review, we described the characteristics of home-based DHIs in postoperative care for older patients with hip fractures. Our findings revealed that the majority of home-based DHIs encompassed multiple components in terms of purpose and content. Furthermore, we observed that these interventions were typically delivered through a single mode of delivery and were provided by health care professionals. Importantly, our meta-analysis demonstrated that home-based DHIs were effective in improving functional outcomes for older patients during the postoperative phase of care for hip fractures.

Home-based DHIs in older patients with hip fractures after surgery were first described in 2005. This upward trend can be attributed to the use of DHIs in clinical care for various conditions, including frailty [17], diabetes mellitus [54], and chronic disease [55]. We found that the frequency of research on DHIs for older individuals with hip fractures increased due to the aging population on the continent and the development of DHIs following the COVID-19 pandemic [23]. Subsequently, policy briefs for DHIs were adopted and implemented, incorporating a highly developed health care infrastructure with advanced information technology systems, financial support, and health care provider training [56]. Looking ahead, it is predicted that the role of DHIs will continue to expand in line with the global strategy for digital health 2020-2024 [57] and the WHO guideline recommendations on digital interventions for health system strengthening in 2018 [15].

The majority of studies used multicomponent approaches, considering the purpose and content of home-based DHIs in postoperative care for older individuals with hip fractures. Freitag and Magaziner [24] recommended that postoperative care in this population may require multiple interventions provided individually or through coordinated management. In this study, the most common categories of purpose and content were communication and feedback, combined with education in areas such as rehabilitation [40-42,51-53], fall prevention [50], and osteoporosis [49]. A lack of knowledge and awareness was identified as an important barrier for self-care [58]. During the postoperative period, patients with hip fracture and their caregivers receive numerous recommendations, including early ambulation, exercise and rehabilitation programs, preventing postoperative complications, modifying home hazards, and taking nutrition supplements [24]. Therefore, effective communication skills with specific feedback are crucial for delivering accurate knowledge and information and ultimately leading to positive functional outcomes [24,59,60].

Telerehabilitation was the third most common content of home-based DHIs. It is essential for every patient with a hip fracture to receive rehabilitation in some setting, including outpatient rehabilitation, nursing homes, and home care. Rehabilitation is one of the most crucial aspects of postoperative care, as it contributes to recovering functional outcomes, patient ability, and overall quality of life [24]. Crotty et al [61] found that a home-based rehabilitation program yielded similar outcomes and placed less burden on caregivers compared with a hospital program. Following the COVID-19 pandemic, telerehabilitation has gained even greater significance in clinical care for patients with postoperative hip fracture. According to Seron et al [22], the most common interventions in telerehabilitation include therapeutic exercises, functional training, and education. Therefore, telerehabilitation is a vital option for rehabilitation as it reduces hospital visits, decreases the cost of hospital visits, and still enhances functional outcomes.

The telephone call was the most common mode of delivery in this study and previous studies due to its ease of use, availability in most households, and effectiveness in communicating and providing simple educational information to patients and their families [24,60]. This is especially true in the Thai context, where a high number of older adults do not use the internet in their everyday lives. Therefore, using telephone calls should enhance the chance for accessing care with DHIs. Multiple episodes of telephone calls may increase adherence and functional outcomes more than a single telephone call [43,50]. However, based on our findings, using the telephone alone was less likely to achieve the goal of improving functional outcomes than other modes of delivery for patients with postoperative hip fracture. The limitations of telephone calls include only voice evaluation and advice and the absence of video connection. Telephone calls are limited when health care providers need to assess patients and provide exercise programs through video calls [62]. Therefore, telerehabilitation should use web-based software or mobile apps as the mode of delivery. Both platforms can provide real-time or on-demand video calls and chat,

XSL•FO
RenderX

allowing for the delivery of rehabilitation programs, tracking of progress, and modification of the program based on the patient's status [24].

Most studies used only 1 health care provider who played an important role in this intervention. The top 2 providers were PT and OT, who addressed the population's rehabilitation needs during the recovery phase. They provided instructional materials and self-management guidance at home, leading to improvements in functional abilities, as shown by the results. However, a program that included only dieticians did not have evidence of improving the patients' status. This detail could be an insightful consideration for providers when developing care programs. While nutritional improvement is required following orthopedic surgery [63], focusing solely on nutrition may not be sufficient for patients to recover from their postoperative condition; a combination of nutrition and physical therapy interventions can maximize function [64]. Therefore, to provide home-based DHIs, health care providers should allow patients and families access to a specific scope of care that includes a rehabilitation program. In addition, given the complexity of hip fracture patients, a multidisciplinary approach is essential, typically involving early geriatric assessment [24]. This approach can enhance functional outcomes, reduce hospital stays, and decrease the future need for institutional care. Although the evidence regarding the effectiveness of multidisciplinary teams in this population is unclear [23], holistic care, such as holistic assessment, rehabilitation, and nutrition, remains important.

In the meta-analysis, it was found that home-based DHIs have a positive impact on enhancing functional outcomes by improving scores on tests such as TUG, SPPB, and FIM even after the interventions end [19,28,40-53]. The differences in population caused variances in the mean difference of TUG reduction. Although this study demonstrated that home-based DHIs have a positive impact on enhancing functional outcomes, further studies may be needed across other functional outcomes to provide a comprehensive summary of the results. The characteristics of home-based DHIs that were associated with increased functional outcomes included a combination of telerehabilitation, effective communication and feedback, and educational components. The modes of delivery commonly used were web-based software, and mobile apps, similar to previous literature [23]. Whether it involves a multidisciplinary team or a single health care provider, the inclusion of physical therapy content in home-based DHIs was found to improve functional outcomes. These findings highlight the importance of incorporating rehabilitation or physical therapy content through telehealth care. This may be particularly beneficial in the context of orthopedic surgery, where reduced limb usage can lead to muscle weakness and functional disability [65]. By enhancing movement through appropriate rehabilitation, improvements in range of motion, muscle strength, and overall performance can be achieved [65]. In addition, this study demonstrated that home-based DHIs potentially have a positive impact on improving physical function and activities of daily living, self-care, and nutritional status. Further studies may be needed to provide a comprehensive summary of these results. Future research should aim to include a larger number of studies reporting patient outcomes, ensure consistency in time points for outcome assessment, and explore the effectiveness of DHIs in diverse populations with hip fractures. For example, this could involve studying different underlying health conditions, various types of hip fractures or surgeries, or varying levels of functional impairment. Such research can help identify which subgroups benefit most from home-based DHIs. In addition, further research on cost-effectiveness of home-based DHIs compared with traditional interventions can inform health care decision makers and potentially lead to more widespread adoption of these approaches.

Limitations

In this study, we demonstrated the common characteristics of home-based DHIs that have shown benefits in improving functional outcomes. This finding holds promise for the development of home-based DHIs aimed at enhancing functional outcomes. However, it is important to acknowledge the limitations encountered during the study. First, only a limited number of studies (6 in total) reported the functional outcomes of interest, which may restrict the generalizability of our findings. Second, although a summary analysis of functional results was conducted, the time points at which these outcomes were measured varied among the included studies. The duration of the intervention itself could potentially impact the observed differences in outcomes. Third, most studies did not specify the type of hip fracture and the type of surgery. Their populations were characterized as patients with hip fractures. The specific type of hip fracture was presented only in 2 studies. Consequently, we were unable to conduct a subgroup analysis**.** Moreover, it is important to note that our study exclusively focused on DHIs used in the care of postoperative populations with hip fractures. This narrow scope may limit the applicability of our findings to nonsurgical hip fracture populations, where the intervention strategies and approaches may differ. Furthermore, our meta-analysis mainly focused on functional outcomes, but there are other factors that impact patient overall outcomes, such as fear of falling, pain control, and caregiver burden, which have not been conclusively examined in this study. These limitations should be taken into consideration when interpreting the results.

Conclusions

Overall, our findings suggest that home-based DHIs, particularly those incorporating telerehabilitation, offer a promising approach to enhance the recovery and functional abilities of older patients following hip fracture surgery. These interventions have the potential to improve the quality of care and patient outcomes in this population. Further research and implementation of home-based DHIs are warranted to fully leverage their benefits in postoperative care for older patients with hip fractures.

XSL•FO
RenderX

Acknowledgments

This work was partially supported by Chiang Mai University.

Data Availability

All data generated or analyzed during this study are included in the tables and figures of this published paper.

Authors' Contributions

Study conception and design were done by SP, KP, CA, PACM, and WJ. Data collection was performed by SP, NB, NW, and KP. Analysis and interpretation of results were done by SP, KP, NB, and CA. SP, KP, and CA contributed to drafting of the manuscript. All authors reviewed the results, revised the manuscript critically for important intellectual content, and approved the final version of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 275 KB - jmir_v26i1e49482_app1.pdf]

Multimedia Appendix 2
Searching and details of included articles.
[DOCX File , 111 KB - jmir_v26i1e49482_app2.docx]

References

1. Asif M, Cadel L, Kuluski K, Everall AC, Guilcher SJT. Patient and caregiver experiences on care transitions for adults with a hip fracture: a scoping review. Disabil Rehabil 2020;42(24):3549-3558. [doi: 10.1080/09638288.2019.1595181] [Medline: 31081400]
2. Parker M, Johansen A. Hip fracture. BMJ 2006;333(7557):27-30. [doi: 10.1136/bmj.333.7557.27] [Medline: 16809710]
3. Veronese N, Maggi S. Epidemiology and social costs of hip fracture. Injury 2018;49(8):1458-1460. [doi: 10.1016/j.injury.2018.04.015] [Medline: 29699731]
4. Bhandari M, Swiontkowski M. Management of acute hip fracture. N Engl J Med 2017;377(21):2053-2062. [doi: 10.1056/NEJMcp1611090] [Medline: 29166235]
5. LeBlanc KE, Muncie HL, LeBlanc LL. Hip fracture: diagnosis, treatment, and secondary prevention. Am Fam Physician 2014 Jun 15;89(12):945-951 [FREE Full text] [Medline: 25162161]
6. Rico CLV, Curcio CL. Fear of falling and environmental factors: a scoping review. Ann Geriatr Med Res 2022;26(2):83-93 [FREE Full text] [doi: 10.4235/agmr.22.0016] [Medline: 35673944]
7. Colais P, Di Martino M, Fusco D, Perucci CA, Davoli M. The effect of early surgery after hip fracture on 1-year mortality. BMC Geriatr 2015;15:141. [doi: 10.1186/s12877-015-0140-y] [Medline: 26510919]
8. Storm M, Siemsen IMD, Laugaland KA, Dyrstad DN, Aase K. Quality in transitional care of the elderly: key challenges and relevant improvement measures. Int J Integr Care 2014;14:e013 [FREE Full text] [doi: 10.5334/ijic.1194] [Medline: 24868196]
9. Mahran DG, Farouk O, Ismail MA, Alaa MM, Eisa A, Ragab II. Effectiveness of home based intervention program in reducing mortality of hip fracture patients: a non-randomized controlled trial. Arch Gerontol Geriatr 2019;81:8-17. [doi: 10.1016/j.archger.2018.11.007] [Medline: 30471472]
10. Fernandez MA, Griffin XL, Costa ML. Management of hip fracture. Br Med Bull 2015;115(1):165-172. [doi: 10.1093/bmb/ldv036] [Medline: 26311503]
11. Transitions of Care: Technical Series on Safer Primary Care. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252272 [accessed 2023-05-05]
12. Roper KL, Ballard J, Rankin W, Cardarelli R. Systematic review of ambulatory transitional care management (TCM) visits on hospital 30-day readmission Rates. Am J Med Qual 2017;32(1):19-26. [doi: 10.1177/1062860615615426] [Medline: 26625898]
13. Phang JK, Lim ZY, Yee WQ, Tan CYF, Kwan YH, Low LL. Post-surgery interventions for hip fracture: a systematic review of randomized controlled trials. BMC Musculoskelet Disord 2023;24(1):417 [FREE Full text] [doi: 10.1186/s12891-023-06512-9] [Medline: 37231406]
14. Fong B, Fong A, Li C. Telemedicine Technologies: Information Technologies in Medicine and Digital Health. Hoboken, NJ: Wiley; 2020.

XSL•FO
RenderX

15. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/311941/9789241550505-eng.pdf?ua=1 [accessed 2023-05-05]
16. Solis-Navarro L, Gismero A, Fernández-Jané C, Torres-Castro R, Solá-Madurell M, Bergé C, et al. Effectiveness of home-based exercise delivered by digital health in older adults: a systematic review and meta-analysis. Age Ageing 2022;51(11):afac243 [FREE Full text] [doi: 10.1093/ageing/afac243] [Medline: 36346736]
17. Linn N, Goetzinger C, Regnaux JP, Schmitz S, Dessenne C, Fagherazzi G, et al. Digital health interventions among people living with frailty: a scoping review. J Am Med Dir Assoc 2021;22(9):1802-1812 [FREE Full text] [doi: 10.1016/j.jamda.2021.04.012] [Medline: 34000266]
18. Gilboa Y, Maeir T, Karni S, Eisenberg ME, Liebergall M, Schwartz I, et al. Effectiveness of a tele-rehabilitation intervention to improve performance and reduce morbidity for people post hip fracture—study protocol for a randomized controlled trial. BMC Geriatr 2019;19(1):135. [doi: 10.1186/s12877-019-1141-z] [Medline: 31109289]
19. Pol MC, Riet GT, Hartingsveldt MV, Kröse B, Buurman BM. Effectiveness of sensor monitoring in a rehabilitation programme for older patients after hip fracture: a three-arm stepped wedge randomised trial. Age Ageing 2019;48(5):650-657. [doi: 10.1093/ageing/afz074] [Medline: 31204776]
20. Montero-Odasso MM, Kamkar N, Pieruccini-Faria F, Osman A, Sarquis-Adamson Y, Close J, Task Force on Global Guidelines for Falls in Older Adults. Evaluation of clinical practice guidelines on fall prevention and management for older adults: a systematic review. JAMA Netw Open 2021;4(12):e2138911 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.38911] [Medline: 34910151]
21. Kuijlaars IAR, Sweerts L, Nijhuis-van der Sanden MWG, van Balen R, Staal JB, van Meeteren NLU, et al. Effectiveness of supervised home-based exercise therapy compared to a control intervention on functions, activities, and participation in older patients after hip fracture: a systematic review and meta-analysis. Arch Phys Med Rehabil 2019;100(1):101-114. [doi: 10.1016/j.apmr.2018.05.006] [Medline: 29885311]
22. Seron P, Oliveros MJ, Gutierrez-Arias R, Fuentes-Aspe R, Torres-Castro RC, Merino-Osorio C, et al. Effectiveness of telerehabilitation in physical therapy: a rapid overview. Phys Ther 2021;101(6):pzab053 [FREE Full text] [doi: 10.1093/ptj/pzab053] [Medline: 33561280]
23. Zhang J, Yang M, Ge Y, Ivers R, Webster R, Tian M. The role of digital health for post-surgery care of older patients with hip fracture: a scoping review. Int J Med Inform 2022;160:104709. [doi: 10.1016/j.ijmedinf.2022.104709] [Medline: 35144101]
24. Freitag MH, Magaziner J. Post-operative considerations in hip fracture management. Curr Rheumatol Rep 2006;8(1):55-62. [doi: 10.1007/s11926-006-0027-1] [Medline: 16515767]
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GLD, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front. Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
27. Gagnon MP, Sasseville M, Leblanc A. Classification of digital mental health interventions: a rapid review and framework proposal. Stud Health Technol Inform 2022;294:629-633. [doi: 10.3233/SHTI220545] [Medline: 35612165]
28. Zhang YY, Zhang YG, Li Z, Li SH, Xu WG. Effect of home-based telerehabilitation on the postoperative rehabilitation outcome of hip fracture in the aging population. Orthop Surg 2022;14(8):1768-1777 [FREE Full text] [doi: 10.1111/os.13293] [Medline: 35819099]
29. Barry E, Galvin R, Keogh C, Horgan F, Fahey T. Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta-analysis. BMC Geriatr 2014;14:14 [FREE Full text] [doi: 10.1186/1471-2318-14-14] [Medline: 24484314]
30. Deeks JJ, Higgins JP, Altman DG. Analysing data and undertaking meta-analyses. Cochrane Collaboration. 2019. URL: https://training.cochrane.org/handbook/current/chapter-10 [accessed 2023-05-05]
31. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
32. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
33. Schünemann H, Higgins J, Vist G, Glasziou P, Akl E, Skoetz N. Completing 'Summary of findings' tables and grading the certainty of the evidence. Cochrane Handbook for Systematic Reviews of Interventions 2019 Sep 20:375-402. [doi: 10.1002/9781119536604.ch14]
34. Dallimore RK, Asinas-Tan ML, Chan D, Hussain S, Willett C, Zainuldin R. A randomised, double-blinded clinical study on the efficacy of multimedia presentation using an iPad for patient education of postoperative hip surgery patients in a public hospital in Singapore. Singapore Med J 2017 Sep;58(9):562-568 [FREE Full text] [doi: 10.11622/smedj.2016084] [Medline: 27121921]

XSL•FO
RenderX

35. Crawford DA, Lombardi AV, Berend KR, Huddleston III JI, Peters CL, DeHaan A, et al. Early outcomes of primary total hip arthroplasty with use of a smartphone-based care platform: a prospective randomized controlled trial. Bone Joint J 2021 Jul;103B(7 Supple B):91-97. [doi: 10.1302/0301-620X.103B7.BJJ-2020-2402.R1] [Medline: 34192907]
36. Jagos H, Pils K, Haller M, Wassermann C, Chhatwal C, Rafolt D, et al. Mobile gait analysis via eSHOEs instrumented shoe insoles: a pilot study for validation against the gold standard GAITRite. J Med Eng Technol 2017;41(5):375-386. [doi: 10.1080/03091902.2017.1320434] [Medline: 28573909]
37. Hou YJ, Zeng SH, Lin CC, Yang CT, Huang HL, Chen MC, et al. Smart clothes-assisted home-nursing care program for family caregivers of older persons with dementia and hip fracture: a mixed-methods study. BMC Geriatr 2022 Feb 05;22(1):104 [FREE Full text] [doi: 10.1186/s12877-022-02789-y] [Medline: 35123399]
38. Nahm ES, Resnick B, DeGrezia M, Brotemarkle R. Use of discussion boards in a theory-based health web site for older adults. Nurs Res 2009;58(6):419-426. [doi: 10.1097/nnr.0b013e3181bee6c4] [Medline: 19918152]
39. Lee H, Lee SH. Effectiveness of multicomponent home-based rehabilitation in elderly patients after hip fracture surgery: a randomized controlled trial. J Pers Med 2022;12(4):649 [FREE Full text] [doi: 10.3390/jpm12040649] [Medline: 35455765]
40. Kalron A, Tawil H, Peleg-Shani S, Vatine JJ. Effect of telerehabilitation on mobility in people after hip surgery: a pilot feasibility study. Int J Rehabil Res 2018;41(3):244-250. [doi: 10.1097/MRR.0000000000000296] [Medline: 29794545]
41. Latham NK, Harris BA, Bean JF, Heeren T, Goodyear C, Zawacki S, et al. Effect of a home-based exercise program on functional recovery following rehabilitation after hip fracture: a randomized clinical trial. JAMA 2014;311(7):700-708 [FREE Full text] [doi: 10.1001/jama.2014.469] [Medline: 24549550]
42. Li CT, Hung GK, Fong KN, Gonzalez PC, Wah SH, Tsang HW. Effects of home-based occupational therapy telerehabilitation via smartphone for outpatients after hip fracture surgery: a feasibility randomised controlled study. J Telemed Telecare 2022;28(4):239-247. [doi: 10.1177/1357633X20932434] [Medline: 32594832]
43. Pfeiffer K, Kampe K, Klenk J, Rapp K, Kohler M, Albrecht D, et al. Effects of an intervention to reduce fear of falling and increase physical activity during hip and pelvic fracture rehabilitation. Age Ageing 2020;49(5):771-778. [doi: 10.1093/ageing/afaa050] [Medline: 32832985]
44. Krichbaum K. GAPN postacute care coordination improves hip fracture outcomes. West J Nurs Res 2007;29(5):523-544. [doi: 10.1177/0193945906293817] [Medline: 17526868]
45. Langford DP, Fleig L, Brown KC, Cho NJ, Frost M, Ledoyen M, et al. Back to the future—feasibility of recruitment and retention to patient education and telephone follow-up after hip fracture: a pilot randomized controlled trial. Patient Prefer Adherence 2015;9:1343-1351 [FREE Full text] [doi: 10.2147/PPA.S86922] [Medline: 26491262]
46. Breedveld-Peters JJ, Reijven PLM, Wyers CE, van Helden S, Arts JJC, Meesters B, et al. Integrated nutritional intervention in the elderly after hip fracture. A process evaluation. Clin Nutr 2012;31(2):199-205. [doi: 10.1016/j.clnu.2011.10.004] [Medline: 22035956]
47. Wyers CE, Reijven PLM, Breedveld-Peters JJL, Denissen KFM, Schotanus MGM, van Dongen MCJM, et al. Efficacy of nutritional intervention in elderly after hip fracture: a multicenter randomized controlled trial. J Gerontol A Biol Sci Med Sci 2018;73(10):1429-1437 [FREE Full text] [doi: 10.1093/gerona/gly030] [Medline: 30204859]
48. Córcoles-Jiménez MP, Candel-Parra E, Del Egido-Fernández M, Villada-Munera A, Moreno-Moreno M, Piña-Martínez AJ, et al. Preventing functional urinary incontinence in hip-fractured older adults through patient education: a randomized controlled trial. J Appl Gerontol 2021;40(8):890-901. [doi: 10.1177/0733464820952608] [Medline: 32865102]
49. Gardner MJ, Brophy RH, Demetrakopoulos D, Koob J, Hong R, Rana A, et al. Interventions to improve osteoporosis treatment following hip fracture. A prospective, randomized trial. J Bone Joint Surg Am 2005;87(1):3-7. [doi: 10.2106/JBJS.D.02289] [Medline: 15634808]
50. Di Monaco M, De Toma E, Gardin L, Giordano S, Castiglioni C, Vallero F. A single postdischarge telephone call by an occupational therapist does not reduce the risk of falling in women after hip fracture: a randomized controlled trial. Eur J Phys Rehabil Med 2015 Feb;51(1):15-22 [FREE Full text] [Medline: 25184799]
51. Cheng KC, Lau KMK, Cheng ASK, Lau TSK, Lau FOT, Lau MCH, et al. Use of mobile app to enhance functional outcomes and adherence of home-based rehabilitation program for elderly with hip fracture: a randomized controlled trial. Hong Kong Physiother J 2022;42(2):99-110 [FREE Full text] [doi: 10.1142/S101370252250010X] [Medline: 37560168]
52. Ortiz-Piña M, Molina-Garcia P, Femia P, Ashe MC, Martín-Martín L, Salazar-Graván S, et al. Effects of tele-rehabilitation compared with home-based in-person rehabilitation for older adult's function after hip fracture. Int J Environ Res Public Health 2021;18(10):5493 [FREE Full text] [doi: 10.3390/ijerph18105493] [Medline: 34065523]
53. Bedra M, Finkelstein J. Feasibility of post-acute hip fracture telerehabilitation in older adults. Stud Health Technol Inform 2015;210:469-473. [Medline: 25991191]
54. Stevens S, Gallagher S, Andrews T, Ashall-Payne L, Humphreys L, Leigh S. The effectiveness of digital health technologies for patients with diabetes mellitus: a systematic review. Front Clin Diabetes Healthc 2022;3:936752 [FREE Full text] [doi: 10.3389/fcdhc.2022.936752] [Medline: 36992773]
55. von Huben A, Howell M, Howard K, Carrello J, Norris S. Health technology assessment for digital technologies that manage chronic disease: a systematic review. Int J Technol Assess Health Care 2021;37(1):e66. [doi: 10.1017/S0266462321000362] [Medline: 34034851]

XSL•FO
RenderX

56. Fahy N. Use of digital health tools in Europe: before, during and after COVID-19. European Observatory on Health Systems and Policies. 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK576970/ [accessed 2023-05-05]
57. Dhingra D, Dabas A. Global strategy on digital health. Indian Pediatr 2020;57(4):356-358 [FREE Full text] [doi: 10.1007/s13312-020-1789-7] [Medline: 32284477]
58. Kim MY, Lee EJ. Factors affecting self-care behavior levels among elderly patients with type 2 diabetes: a quantile regression approach. Medicina (Kaunas) 2019;55(7):340 [FREE Full text] [doi: 10.3390/medicina55070340] [Medline: 31277522]
59. Vincent C, Wegier P, Chien V, Kurahashi AM, Ginsburg S, Molla Ghanbari H, et al. Qualitative evaluation of a novel educational tool to communicate individualized hip fracture prognostic information to patients and surrogates: my hip fracture (my-HF). Geriatr Orthop Surg Rehabil 2021;12:21514593211050513 [FREE Full text] [doi: 10.1177/21514593211050513] [Medline: 34707914]
60. Yadav L, Haldar A, Jasper U, Taylor A, Visvanathan R, Chehade M, et al. Utilising digital health technology to support patient-healthcare provider communication in fragility fracture recovery: systematic review and meta-analysis. Int J Environ Res Public Health 2019;16(20):4047 [FREE Full text] [doi: 10.3390/ijerph16204047] [Medline: 31652597]
61. Crotty M, Whitehead C, Miller M, Gray S. Patient and caregiver outcomes 12 months after home-based therapy for hip fracture: a randomized controlled trial. Arch Phys Med Rehabil 2003;84(8):1237-1239. [doi: 10.1016/s0003-9993(03)00141-2] [Medline: 12917867]
62. Paul B, Saha I, Kumar S, Samim Ferdows SK, Ghose G. Mobile phones: time to rethink and limit usage. Indian J Public Health 2015;59(1):37-41 [FREE Full text] [doi: 10.4103/0019-557X.152856] [Medline: 25758729]
63. Hirsch KR, Wolfe RR, Ferrando AA. Pre- and post-surgical nutrition for preservation of muscle mass, strength, and functionality following orthopedic surgery. Nutrients 2021;13(5):1675 [FREE Full text] [doi: 10.3390/nu13051675] [Medline: 34063333]
64. Inoue T, Iida Y, Takahashi K, Shirado K, Nagano F, Miyazaki S, et al. Nutrition and physical therapy: a position paper by the physical therapist section of the Japanese Association of Rehabilitation Nutrition (Secondary Publication). JMA J 2022;5(2):243-251 [FREE Full text] [doi: 10.31662/jmaj.2021-0201] [Medline: 35611222]
65. Sanford JA, Griffiths PC, Richardson P, Hargraves K, Butterfield T, Hoenig H. The effects of in-home rehabilitation on task self-efficacy in mobility-impaired adults: a randomized clinical trial. J Am Geriatr Soc 2006;54(11):1641-1648. [doi: 10.1111/j.1532-5415.2006.00913.x] [Medline: 17087689]

Abbreviations

CBT: cognitive behavioral therapy
DHI: digital health intervention
FIM: Functional Independence Measure
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
OT: occupational therapist
PT: physical therapist or physiotherapist
RCT: randomized controlled trial
SPPB: Short Physical Performance Battery
TUG: Timed Up and Go

Edited by T de Azevedo Cardoso; submitted 06.06.23; peer-reviewed by SQ Yoong, C Zhang; comments to author 23.09.23; revised version received 07.10.23; accepted 29.04.24; published 12.06.24.

Please cite as:
Pliannuom S, Pinyopornpanish K, Buawangpong N, Wiwatkunupakarn N, Mallinson PAC, Jirapornchaoren W, Angkurawaranon C Characteristics and Effects of Home-Based Digital Health Interventions on Functional Outcomes in Older Patients With Hip Fractures After Surgery: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e49482
URL: https://www.jmir.org/2024/1/e49482
doi:10.2196/49482
PMID:38865706

©Suphawita Pliannuom, Kanokporn Pinyopornpanish, Nida Buawangpong, Nutchar Wiwatkunupakarn, Poppy Alice Carson Mallinson, Wichuda Jirapornchaoren, Chaisiri Angkurawaranon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Peer- or Professional-Led Support in Enhancing Adherence to Wearable Monitoring Devices Among Community-Dwelling Older Adults: Systematic Review of Randomized Controlled Trials

Colette Sze Wing Chan[1], BSc; Mandy Ming Pui Kan[1], MSc; Arkers Kwan Ching Wong[1], PhD

School of Nursing, The Hong Kong Polytechnic University, Hung Hom, China (Hong Kong)

Corresponding Author:
Arkers Kwan Ching Wong, PhD
School of Nursing
The Hong Kong Polytechnic University
GH 502, The Hong Kong Polytechnic University
Hung Hom, 00000
China (Hong Kong)
Phone: 852 34003805
Email: arkers.wong@polyu.edu.hk

Abstract

Background: Despite the well-documented health benefits associated with wearable monitoring devices (WMDs), adherence among community-dwelling older adults remains low. By providing guidance on the purpose and benefits of using WMDs, facilitating goal-setting aligned with the device's features, promoting comprehension of the health data captured by the device, and assisting in overcoming technological challenges, peers and health care professionals can potentially enhance older adults' adherence to WMDs. However, the effectiveness of such support mechanisms in promoting adherence to WMDs among older adults remains poorly understood.

Objective: The aims of this systematic review were to examine the effects of peer- or professional-led intervention programs designed to improve adherence to WMDs among community-dwelling older adults and to identify the intervention components that may positively influence the effects of the intervention.

Methods: We conducted a comprehensive search across 7 electronic databases (Cochrane Central Register of Controlled Trials [CENTRAL], PubMed, EMBASE, PsycINFO, British Nursing Index, Web of Science, and CINAHL) to identify articles published between January 1, 2010, and June 26, 2023. We specifically targeted randomized controlled trials that examined the impact of peer- or professional-led interventions on enhancing adherence to WMDs among individuals aged 60 years and older residing in the community. Two independent reviewers extracted data from the included studies and assessed the potential risk of bias in accordance with the Cochrane Risk of Bias tool for randomized trials, version 2.

Results: A total of 10,511 studies were identified through the database search. Eventually, we included 3 randomized controlled trials involving 154 community-dwelling older adults. The participants had a mean age of 65 years. Our review revealed that increasing awareness of being monitored and implementing the SystemCHANGE approach, a habit change tool focusing on personal goals and feedback, were effective strategies for enhancing adherence to WMDs among older adults. All of the included studies exhibited a low risk of bias.

Conclusions: By collaboratively designing specific goals related to WMDs with health care professionals, including nurses and physicians, older adults exhibited a higher likelihood of adhering to the prescribed use of WMDs. These goal-setting tools provided a framework for structure and motivation, facilitating the seamless integration of WMDs into their daily routines. Researchers should prioritize interventions that target awareness and goal-setting as effective approaches to enhance adherence to WMDs among older adults, thereby maximizing the realization of associated health benefits.

(J Med Internet Res 2024;26:e53607) doi:10.2196/53607

XSL•FO
RenderX

KEYWORDS

wearable monitoring device; older adults; adherence; systematic review; healthy aging; peer support; professional help; support; peers; peer; professionals; wearable; monitoring devices; monitoring device; community-dwelling; older adults; older adult; aging; aging; elderly

Introduction

Overview

The global population is ageing at an unprecedented speed, with the number of people older than 65 years expected to double from 1 billion in 2020 to 2.1 billion in 2050 [1]. The World Health Organization has transformed this challenge into an opportunity by putting forward the concept of "healthy ageing," which is about creating an environment that enables older adults to age with good health and well-being [2]. However, to put this concept into practice, older adults, who are usually living with multiple chronic health conditions, have to be competent in carrying out self-management activities, such as medication management, health monitoring, and fall prevention. This is more easily said than done, although it is not impossible with the help of emerging digital self-management technologies.

Wearable monitoring devices (WMDs) are a common digital technology specifically designed to help people take charge of their own health and develop an individualized care or treatment plan. WMDs, which include but are not limited to activity trackers, pedometers, and smartwatches, are useful for collecting a wide range of personal health data in real time, such as blood pressure and heart rate, and revealing an individual's physical movements, GPS location, and calorie expenditure [3]. Not only can WMDs ensure the safety of users by the early screening of symptoms, but they can also empower users to continuously measure their health and well-being independently without the help of family members or health care professionals. Previous studies have demonstrated that WMDs are particularly effective for older adults with chronic diseases. A systematic review of the effectiveness of WMDs reported that WMDs could bring about a noticeable increase in daily step count and moderate-to-vigorous physical activities, although its impact on reducing sedentary behavior was found to be insignificant [4].

Despite the benefits provided by WMDs, studies suggested that the rate of adherence to WMDs among older adults is low. A study showed that 1 in 3 adults abandoned the use of WMDs within 6 months after purchasing one, while those aged 70 years and older discontinued its use after only 2 weeks [5]. Nonadherence to WMDs affects the completeness of health data and impairs early detection of the aggravation of a disease, which will eventually have a negative impact on the health of users [6]. Previous studies indicated that the characteristics of a WMD, such as its ease of use, size, security and privacy setting, and maintenance cost, as well as the characteristics of older adults, such as their technological competence and perception of the usefulness of the WMD, are some factors that may influence the intention of users to keep using the WMD [5,7-9].

Peers or health care professionals are crucial in promoting the long-term use of WMDs among older adults [10-12]. The literature suggests that these people can help older adults gain insights into the purpose of using the WMD, facilitate the planning and setting of goals with regard to the features of the WMD, understand the health data captured in the WMD, and overcome technological difficulties [11,13,14]. However, to the best of our knowledge, there has been no systematic review evaluating the effects of these peer- or professional-led intervention programs on improving long-term adherence to WMDs among older adults or identifying the specific components in these programs that can lead to positive effects. Conducting such reviews could help researchers and policy makers design targeted strategies to promote the use of WMDs, ultimately leading to better health monitoring and improved health care outcomes.

Objectives

This review aims to explore the effects of peer- or professional-led intervention programs on enhancing adherence to WMDs among community-dwelling older adults and to disentangle the following research questions:

1. What existing interventions have been designed to improve adherence to WMDs among community-dwelling older adults?
2. What are the effects of these interventions on adherence to WMDs among community-dwelling older adults?
3. What intervention components may positively influence the effects of interventions aimed at enhancing adherence to WMDs?

Methods

Database Selection and Search Strategy

This review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [15] guidelines, and the protocol was registered on PROSPERO (CRD42023395459).

A total of 7 electronic databases (Cochrane Central Register of Controlled Trials [CENTRAL], PubMed, EMBASE, PsycINFO, British Nursing Index, Web of Science, and CINAHL) were searched to identify articles published between January 1, 2010, and June 26, 2023. A combination of search strings related to "wearable monitoring devices," "older adults," "intervention program," and "adherence" were used. A complete search strategy for each database is shown in Multimedia Appendix 1. A manual search of the reference lists of the included articles was performed to identify additional relevant articles.

Study Selection

This review intended to summarize evidence regarding the effectiveness of peer- or professional-led interventions on improving adherence to WMDs among community-dwelling older adults by comparing those who did and did not receive the intervention. For this review, our definition of peer-led and

XSL•FO
RenderX

professional-led interventions referred to how peers and health care professionals contributed to promoting adherence to WMDs. The eligibility criteria were developed using the PICO (population, intervention, comparison, and outcomes) process.

Inclusion and Exclusion Criteria

Articles on studies that adopted a parallel randomized controlled trial (RCT), cluster RCT, or crossover RCT were eligible for inclusion in the current review if the study participants were community-dwelling older adults (with or without chronic diseases) aged ≥60 years using a WMD (population); a peer-led or professional-led intervention that sought to improve WMD adherence was tested (intervention); a comparison was made with participants not exposed to the interventions (comparison); and their adherence to the WMD was measured (outcomes). Meanwhile, articles were excluded if they were in a language other than English or Chinese; involved institutionalized or hospitalized patients; involved patients who were cognitively impaired or physically limited, making them incapable of adhering to the WMD; and focused mainly on the clinical outcomes of using WMD.

Screening Process

Articles identified from databases were imported to EndNote 20 to remove duplicates. All titles and abstracts were first independently screened by 2 reviewers (MMPK and CSWC) based on the eligibility criteria to identify potential articles. The 2 reviewers then independently reviewed the full text of these articles to determine whether the articles satisfied the selection criteria. Disagreements were resolved by consensus with a third reviewer (AKCW).

Data Extraction and Appraisal of the Methodological Quality

The 2 independent reviewers extracted data from the included studies to a Microsoft Excel template sheet and assessed the potential risk of bias in the studies, in accordance with the Cochrane Risk of Bias tool version 2 (RoB 2) for randomized trials [16]. The corresponding authors of the included articles were contacted for missing data or clarification, if necessary.

The data extracted from each eligible article included the following study characteristics:

- Methods: country and periods during which the study was conducted, study design, and dropout rate
- Participants: characteristics of all the participants by groups (eg, the number of participants, their mean age, gender, ethnicity, educational level, and socioeconomic status)
- Features of the WMDs
- Interventions: description of the intervention type, provider, content, frequency, and duration
- Controls
- Outcomes: adherence, sustained usage, and time points

The RoB 2 tool addressed the biases that potentially affect the results of an RCT in five domains: (1) bias arising from the randomization process; (2) bias due to deviations from intended interventions; (3) bias due to missing outcome data; (4) bias in measuring the outcome; and (5) bias in selecting the reported result. The level of bias in each domain and the overall risk of a study were judged to be either a “low risk of bias,” “some concerns,” or a “high risk of bias” [16].

Data Synthesis

We have included qualitative studies in our research and presented a summary of the findings in a narrative format. We conducted both a thematic synthesis, following the narrative synthesis guidelines established by Popay et al [17], and a more theory-driven analysis to interpret the results.

Results

Selection

A total of 10,511 studies were identified from searching the databases. After removing 1624 duplicated studies, 8887 articles were screened based on their titles and abstracts. The full texts of 33 articles that were deemed eligible were then assessed. Finally, 3 articles were selected for the current review. No further articles were identified from the reference list of the included articles. The complete study selection process is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

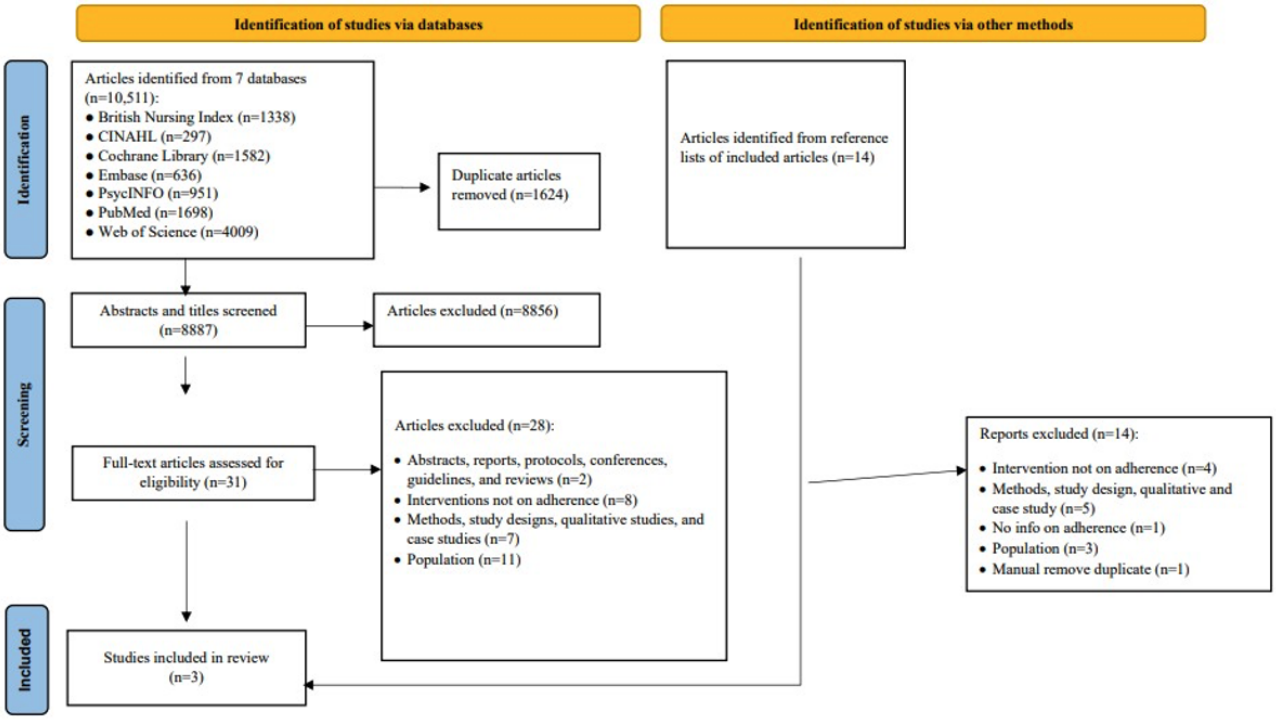

Study Characteristics

Of the 3 selected studies, 2 were conducted in the United States [18,19], and 1 was conducted in the Netherlands [20]. The study with the highest dropout rate (75.4%) was that by Lutjeboer et al [20]. All 3 articles used a 2-arm parallel-group RCT design [18-20].

Table 1 summarizes the 3 articles that were included in this study, which involved a total of 154 participants. The mean age of the participants was 65 years. The percentage of female participants ranged from 41.7% to 61.8%. Two studies focused on older adults who had undergone a kidney transplant [18,19], while the remaining study involved older adults with no major health problems [20]. The older adults in the included studies had a relatively high level of education, with more than half possessing an associate degree or higher [18,19]. The majority of participants in the two studies had an annual household income of more than US $20,000 [18,19]. This suggests that, in general, the participants had a fair to high socioeconomic status.

XSL•FO
RenderX

Table 1. Summary of the included articles (N=3).

Author; year; country	Design	Sample size, n	Dropout rate (%)	Mean age (year); female participants (%)	Study population; ethnicity	Education level	Information about the WMD[a]	Randomized ratio	Intervention arm 1	Comparison arm (intervention arm 2)	Outcomes measure and time points	Results	Intervention component
Lutjeboer et al (2020) [20]; Netherlands	Parallel RCT[b]	55	75.4	65.9; 1.8	Adults who received their first pair of orthopedic footwear for a health condition; NS[c]	NS	Custom-made orthopedic footwear with a temperature sensor to measure use and nonuse of orthopedic footwear	1:1	Aware of the temperature on the sensor that measures use and nonuse (n=25)	Only knew the shoe temperature was being measured (n=30)	Adherence to orthopedic footwear reflected by wearing time in weeks 6 to 12	Intervention group: 7.32 hours/day; control group: 6.11 hours/day	Conscious awareness of WMD adherence being monitored enhances adherence to WMD
O'Brien et al (2020) [19]; United States	Parallel RCT	53	11.7	65.4; 41.7	Older kidney transplant recipients; 98% non-Hispanic	58% of participants had an associate degree or higher; over 90% had an income of > US $20,000	Fitbit Charge 2 activity tracker measuring daily steps	1:1	Multicomponent physical activity intervention System-CHANGE (ie, personal system-based solutions support with visual feedback from the activity tracker; n=27)	Standard educational information about posttransplant (n=26)	Adherence to wearing the activity tracker at 6 months	Intervention group: 100%; control group: 92.3%	Tailored individual activity goals with visual feedback from WMD promotes adherence to WMD
O'Brien et al (2021) [18]; United States	Parallel RCT	46	23.3	65.4; 41.7	Older kidney transplant recipients; 87% non-Hispanic	52% of participants had an associate degree or higher; 77% had an income of >US $20,000	Fitbit Charge 2 activity tracker measuring daily steps	1:1	Multicomponent physical activity intervention System-CHANGE (ie, personal system-based solutions support with visual feedback from the activity tracker; n=23)	Standard educational information about posttransplant healthy lifestyle choices (n=23)	Adherence to wearing the activity tracker at 12 months	Intervention group: 96.5%; control group: 80.8%; odds ratio 6.6 (95% CI 2.1-21.2)	Tailored individual activity goals with visual feedback from WMD promotes adherence to WMD

[a]WMD: wearable monitoring device.

[b]RCT: randomized controlled trial.

[c]NS: not specified.

XSL•FO
RenderX

WMD Information and Adherence Outcomes

The two studies that involved Fitbit smartwatches stated the rate of adherence to WMDs as their outcome but did not specify how it was measured [18,19]. In the third study, which involved orthopedic footwear for people with a foot condition, adherence (both use and nonuse) was measured using a built-in temperature sensor [20].

Objective 1: Existing Interventions to Improve WMD Adherence

In the 3 studies, we identified 2 interventions aimed at enhancing adherence to WMDs. In the study conducted by Lutjeboer et al [20], the intervention used was awareness of being monitored for wearing time [20]. Specifically, the "no awareness" group only knew that their footwear contained WMDs, which in this case were temperature sensors, but did not know the purpose of the study. On the other hand, the "awareness" group received a letter explaining the purpose of the study, thus ensuring they were well informed of being monitored on the time they spent wearing the WMDs.

Another intervention was the SystemCHANGE approach, in which the focus was on changing habits by targeting personal goals and providing feedback through 4 steps [18,19]. The first step involved identifying individuals who had an impact on the participants' level of physical activity. These individuals, recognized by the participants as significant figures, could include peers, family members, and health care professionals. Then, the participants were asked to recognize their routine and the impact of that routine. Subsequently, the identified routine was graphically represented to help the participants understand how it affected their physical activity patterns. The final step entailed developing a customized solution for the participants, with the assistance of WMDs.

Objective 2: Effects of Interventions on WMD Adherence

Although the interventions in all 3 studies were not specifically designed to promote adherence or evaluate the components contributing to adherence, they did report adherence rates at single time points. The duration of the adherence that was measured varied across the studies, ranging from 12 weeks [20] to 12 months [18]. The results demonstrated that the participants' awareness of being monitored while using WMDs was effective in enhancing adherence. This was evident from the longer mean wearing time in the intervention group (7.32 hours/day) compared to the comparison group (6.11 hours/day) [20]. Similarly, the studies found that participants were more likely to adhere to the WMDs when using the SystemCHANGE approach, a habit change tool that targets personal goals and provides feedback [18,19]. Adherence among the intervention participants in these 2 studies was 100% at 6 months and 96.5% at 12 months, in contrast to 92.3% and 80.8%, respectively, for the control participants.

Objective 3: Components That Positively Influence the Intervention Effect

In Lutjeboer et al's study [20], physicians played a crucial role in promoting adherence to WMDs. The effectiveness of the intervention relied on physicians providing information about the purpose of the study to the participants. In the studies conducted by O'Brien et al [18,19], both nurses and participants worked together to achieve the intervention effect. The nurses were responsible for guiding the participants to reflect on their lifestyle during the monthly sessions, using the SystemCHANGE approach. Meanwhile, the participants actively engaged in formulating and implementing the action plan, taking an active role in the intervention process.

Risk of Bias Assessment

Figure 2 [21] indicates that in the RoB 2 bias assessment, there were "some concerns" about the 3 articles [18-20]. All of the included studies [18-20] demonstrated a low risk of bias in the domains of "randomization process" and "outcome measurements." In 2 studies [18,20], there was some concern regarding deviation from intended interventions, as the research staff may have been aware of the intervention during the trial. The remaining study was rated as being of "low risk" for this issue [19]. In terms of the measurement of the outcome, there were some concerns with all 3 articles [18-20], as knowledge of the intervention could have affected the study outcomes. Regarding the selection of the reported result, there were "some concerns" in 1 article [19], while the other 2 were considered as being of "low risk" [18,19].

Figure 2. Risk of bias of the included articles (N=3).

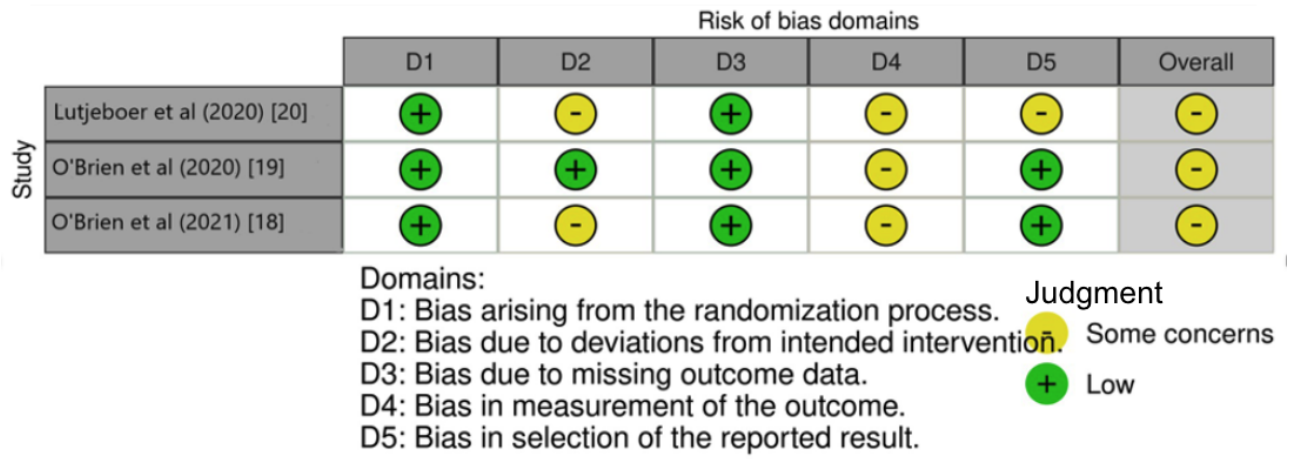

XSL•FO
RenderX

Discussion

Principal Findings

In summary, this systematic review identified 3 RCTs, published between 2020 and 2021, that reported adherence to WMDs. Our findings revealed that low-cost interventions aimed at increasing adherence to WMDs were effective. Specifically, the review highlighted 2 types of interventions that showed promise in enhancing adherence to WMDs among older adults. The first intervention involved raising awareness among older adults about the wearing time of the devices. This approach aimed to ensure that individuals were consistently wearing the devices for the recommended duration. The second intervention strategy involved the use of goal-setting tools. By setting specific goals related to wearing the WMDs, older adults were more likely to adhere to the prescribed use. These goal-setting tools could provide individuals with the structure and motivation to incorporate the use of WMDs into their daily routines.

It is surprising that none of the interventions in the included articles were primarily aimed at promoting adherence to WMDs but rather at examining the effectiveness of the WMDs in enhancing health outcomes. For instance, there was a study about the effectiveness of WMDs on increasing step count, but no data were presented on the wearing time of the WMDs [22]. Another study examined adherence to an intervention program [23]. This study focused on assessing how well participants adhered to a specific intervention or treatment program rather than on specifically evaluating adherence to WMDs. Furthermore, most of the cross-sectional studies identified in the systematic review concentrated on investigating the reasons and factors influencing the use or nonuse of WMDs by older adults [5,7,9,24-26]. These studies aimed to understand the barriers and facilitators associated with older adults' adoption and acceptance of WMDs but did not explore how older adults can sustain the use of WMDs over an extended period.

It is also noteworthy that the pilot study on organ transplant recipients [19] was related to another parent article [18], which theoretically followed up on the same population and intervention at two different time points. The decrease in the WMD adherence rate from 6 months (intervention: 100%; control: 92.3%) to 12 months (intervention: 96.5%; control: 80.8%) implies that interventions targeting adherence should be implemented to enhance the sustained use of WMDs for health monitoring. Previous studies found that the barriers contributing to the nonuse of WMDs included technological concerns and user experiences [25-27]. For instance, a lack of skills to use the WMDs, a feeling of uncomfortableness, and inaccuracies in the physiological data collected by WMDs reduced their use [27]. Additionally, inadequate support from health care providers to encourage the use of WMDs may reduce the patients' trust in the accuracy of the data, hence discouraging WMD use [24]. Both eliminating such barriers and identifying facilitators are necessary to enhance adherence to WMDs.

Our results from 2 studies [18,19] were consistent with the findings from other studies. For instance, previous research has shown that having health care professionals assist older adults in setting health-related goals and providing them with feedback is a crucial factor influencing adherence to WMDs [11,26]. This is because personalized goals and individualized interventions can motivate older adults to achieve desired health outcomes [11] and empower them to make choices related to their health [28]. For example, the SystemCHANGE approach, as illustrated in the 2 studies [18,19], involved guiding participants to collaboratively establish goals with health care professionals by modifying their current routines, thereby making the goals more attainable.

In addition, O'Brien et al [19] emphasized the importance of consistently following up and providing physical incentives, such as gift cards, to enhance motivation among older adults to sustain the use of WMDs. On the other hand, Lutjeboer et al [20] found that older adults who were aware of being tracked in their footwear use demonstrated improved adherence. This finding can be explained by the Hawthorne effect in psychology, whereby participants may alter their behavior when they are aware of being observed [29]. In the case of orthopedic footwear being prescribed by physicians, older adults may seek to establish a therapeutic relationship by meeting the expectations of their physician. Furthermore, older adults may exhibit a greater willingness to wear the footwear when they are pursuing health-related goals [25].

Concerning whether professional-led or peer-led intervention programs are more effective in enhancing WMD adherence, more information is needed before coming to a conclusion. Combining the results from the included studies, health care professionals were found to have played an important role in assisting the participants in improving WMD adherence. A probable explanation for this is that health care professionals, such as physicians and nurses, could monitor the progress of older adults and answer any questions they might have [30]. However, it is unclear whether the effect of peers is comparable to that of professionals. For example, in O'Brien et al's studies [18,19], monthly sessions were carried out in groups. This could imply that aside from health care professionals, peer sharing and support may help to enhance the WMD adherence of older adults. However, the report lacks detailed information on the group sessions, making it difficult to compare the effects of professional-led and peer-led intervention programs.

Our study has some limitations. Even though we planned to search for articles published during the past 13 years, the articles that matched our criteria for inclusion were only published in recent years. In addition, a meta-analysis on the topic could not be performed due to the small number of articles and their heterogeneity. This could be because WMDs have only become popular in recent years. Another limitation is linked to population bias. This report involved a small sample size of 154 older adults in total, and the population is assumed to be the same in 2 of the included studies. Meanwhile, all of the included studies targeted Western populations, so the results may not be generalizable to people in other countries because factors such as technological know-how could influence WMD use behavior.

Conclusions

Our findings revealed that low-cost interventions were effective in increasing adherence to WMDs. Several interventions have been identified as effective in promoting adherence to WMDs.

XSL•FO
RenderX

However, it is important to note that further studies are needed to validate the results and assess the effectiveness of these interventions. Researchers should also explore other potential behavior change interventions that could be beneficial in improving adherence to WMDs, with a particular focus on community-dwelling older adults. By expanding the range of interventions and conducting more comprehensive studies, we can enhance our understanding of the most effective strategies for improving adherence to WMDs in this specific population.

Acknowledgments

This work was supported by the Departmental General Research Fund, The Hong Kong Polytechnic University (G-UAQ2).

Authors' Contributions

AKCW contributed to the conception and design of the study. CSWC and MMPK conducted the search strategy. All authors prepared the draft and reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms in each database.
[DOCX File , 24 KB - jmir_v26i1e53607_app1.docx]

Multimedia Appendix 2
PRISMA checklist.
[PDF File (Adobe PDF File), 134 KB - jmir_v26i1e53607_app2.pdf]

References

1. Ageing and health. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2023-08-19]
2. Healthy ageing and functional ability. World Health Organization. 2020. URL: https://www.who.int/news-room/questions-and-answers/item/healthy-ageing-and-functional-ability [accessed 2023-08-19]
3. Yetisen AK, Martinez-Hurtado JL, Ünal B, Khademhosseini A, Butt H. Wearables in medicine. Adv Mater 2018 Jun 11;30(33):e1706910 [FREE Full text] [doi: 10.1002/adma.201706910] [Medline: 29893068]
4. Ringeval M, Wagner G, Denford J, Paré G, Kitsiou S. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. J Med Internet Res 2020 Oct 12;22(10):e23954 [FREE Full text] [doi: 10.2196/23954] [Medline: 33044175]
5. Ferguson C, Hickman LD, Turkmani S, Breen P, Gargiulo G, Inglis SC. "Wearables only work on patients that wear them": barriers and facilitators to the adoption of wearable cardiac monitoring technologies. Cardiovasc Digit Health J 2021 Apr;2(2):137-147 [FREE Full text] [doi: 10.1016/j.cvdhj.2021.02.001] [Medline: 35265900]
6. Huang Y, Upadhyay U, Dhar E, Kuo L, Syed-Abdul S. A scoping review to assess adherence to and clinical outcomes of wearable devices in the cancer population. Cancers (Basel) 2022 Sep 13;14(18):4437 [FREE Full text] [doi: 10.3390/cancers14184437] [Medline: 36139602]
7. Burford K, Golaszewski NM, Bartholomew J. "I shy away from them because they are very identifiable": a qualitative study exploring user and non-user's perceptions of wearable activity trackers. Digit Health 2021 Nov 20;7:20552076211054922 [FREE Full text] [doi: 10.1177/20552076211054922] [Medline: 34820134]
8. Creaser AV, Hall J, Costa S, Bingham DD, Clemes SA. Exploring families' acceptance of wearable activity trackers: a mixed-methods study. Int J Environ Res Public Health 2022 Mar 15;19(6):3472 [FREE Full text] [doi: 10.3390/ijerph19063472] [Medline: 35329166]
9. Lacey A, Whyte E, O'Keeffe S, O'Connor S, Moran K. A qualitative examination of the factors affecting the adoption of injury focused wearable technologies in recreational runners. PLoS One 2022 Jul 6;17(7):e0265475 [FREE Full text] [doi: 10.1371/journal.pone.0265475] [Medline: 35793284]
10. Lee AYL, Wong AKC, Hung TTM, Yan J, Yang S. Nurse-led telehealth intervention for rehabilitation (telerehabilitation) among community-dwelling patients with chronic diseases: systematic review and meta-analysis. J Med Internet Res 2022 Nov 02;24(11):e40364 [FREE Full text] [doi: 10.2196/40364] [Medline: 36322107]
11. Peng W, Li L, Kononova A, Cotten S, Kamp K, Bowen M. Habit formation in wearable activity tracker use among older adults: qualitative study. JMIR Mhealth Uhealth 2021 Jan 19;9(1):e22488 [FREE Full text] [doi: 10.2196/22488] [Medline: 33464216]

XSL•FO
RenderX

12. Wong KC, Wong FKY, Chang KKP. Health-social partnership intervention programme for community-dwelling older adults: a research protocol for a randomized controlled trial. J Adv Nurs 2015 Nov;71(11):2673-2685. [doi: 10.1111/jan.12700] [Medline: 26040946]
13. Wong AKC, Bayuo J, Wong FKY, Chow KKS, Wong SM, Lau ACK. The synergistic effect of nurse proactive phone calls with an mHealth app program on sustaining app usage: 3-arm randomized controlled trial. J Med Internet Res 2023 May 01;25:e43678 [FREE Full text] [doi: 10.2196/43678] [Medline: 37126378]
14. Wong AKC, Wong FKY, Chow KKS, Wong SM, Lee PH. Effect of a telecare case management program for older adults who are homebound during the COVID-19 pandemic: a pilot randomized clinical trial. JAMA Netw Open 2021 Sep 01;4(9):e2123453 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.23453] [Medline: 34499135]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
17. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC Methods Programme. Version 1. JECH 2005;59(Suppl 1):A7. [doi: 10.13140/2.1.1018.4643]
18. O'Brien T, Tan A, Rose K, Focht B, Daloul R. Maintenance phase of a physical activity intervention in older kidney transplant recipients: a 12-month follow-up. Geriatr Nurs 2021 Nov;42(6):1541-1546 [FREE Full text] [doi: 10.1016/j.gerinurse.2021.08.019] [Medline: 34741827]
19. O'Brien T, Russell CL, Tan A, Mion L, Rose K, Focht B, et al. A pilot randomized controlled trial using SystemCHANGE™ approach to increase physical activity in older kidney transplant recipients. Prog Transplant 2020 Dec 10;30(4):306-314 [FREE Full text] [doi: 10.1177/1526924820958148] [Medline: 32912051]
20. Lutjeboer T, van Netten JJ, Postema K, Hijmans J. Effect of awareness of being monitored on wearing of orthopaedic footwear. J Rehabil Med 2020 Nov 20;52(11):jrm00127 [FREE Full text] [doi: 10.2340/16501977-2759] [Medline: 33094825]
21. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions version 6.2 (updated February 2021). Cochrane Training. URL: https://training.cochrane.org/handbook [accessed 2023-11-01]
22. Kullgren JT, Harkins KA, Bellamy SL, Gonzales A, Tao Y, Zhu J, et al. A mixed-methods randomized controlled trial of financial incentives and peer networks to promote walking among older adults. Health Educ Behav 2014 Oct 01;41(1 Suppl):43S-50S [FREE Full text] [doi: 10.1177/1090198114540464] [Medline: 25274710]
23. Grossman J, Arigo D, Bachman JL. Meaningful weight loss in obese postmenopausal women: a pilot study of high-intensity interval training and wearable technology. Menopause 2018 Apr;25(4):465-470. [doi: 10.1097/GME.0000000000001013] [Medline: 29088015]
24. Godfrey A, Stuart S, editors. Digital Health: Exploring Use and Integration of Wearables. Amsterdam, Netherlands: Elsevier Science & Technology; 2021.
25. Li L, Peng W, Kononova A, Bowen M, Cotten SR. Factors associated with older adults' long-term use of wearable activity trackers. Telemed J E Health 2020 Jun 01;26(6):769-775. [doi: 10.1089/tmj.2019.0052] [Medline: 31553281]
26. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
27. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 09;8(11):e18907 [FREE Full text] [doi: 10.2196/18907] [Medline: 33164904]
28. Baker A, Cornwell P, Gustafsson L, Stewart C, Lannin NA. Developing tailored theoretically informed goal-setting interventions for rehabilitation services: a co-design approach. BMC Health Serv Res 2022 Jun 22;22(1):811 [FREE Full text] [doi: 10.1186/s12913-022-08047-6] [Medline: 35733190]
29. Adair JG. The Hawthorne effect: a reconsideration of the methodological artifact. J Applied Psychol 1984 May;69(2):334-345. [doi: 10.1037/0021-9010.69.2.334]
30. Atreja A, Bellam N, Levy SR. Strategies to enhance patient adherence: making it simple. MedGenMed 2005 Mar 16;7(1):4 [FREE Full text] [Medline: 16369309]

Abbreviations

CENTRAL: Cochrane Central Register of Controlled Trials
PICO: population, intervention, comparison, and outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB 2: Cochrane Risk of Bias tool version 2

XSL•FO
RenderX

WMD: wearable monitoring device

Edited by S Ma; submitted 12.10.23; peer-reviewed by SQ Yoong, E Fisher; comments to author 19.03.24; revised version received 25.03.24; accepted 23.04.24; published 20.06.24.

Please cite as:
Chan CSW, Kan MMP, Wong AKC
Effects of Peer- or Professional-Led Support in Enhancing Adherence to Wearable Monitoring Devices Among Community-Dwelling Older Adults: Systematic Review of Randomized Controlled Trials
J Med Internet Res 2024;26:e53607
URL: https://www.jmir.org/2024/1/e53607
doi:10.2196/53607
PMID:38900546

©Colette Sze Wing Chan, Mandy Ming Pui Kan, Arkers Kwan Ching Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Wearable Technologies for Detecting Burnout and Well-Being in Health Care Professionals: Scoping Review

Milica Barac[1*], BS; Samantha Scaletty[1*], BS; Leslie C Hassett[2], MLS; Ashley Stillwell[3], DO; Paul E Croarkin[4], DO, MS; Mohit Chauhan[5], MD; Sherry Chesak[6], PhD; William V Bobo[5], MD, MPH; Arjun P Athreya[1,4*], MS, PhD; Liselotte N Dyrbye[7*], MD, MPHE

[1]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN, United States
[2]Mayo Clinic Libraries, Mayo Clinic, Rochester, MN, United States
[3]Department of Family Medicine, Mayo Clinic, Phoenix, AZ, United States
[4]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, United States
[5]Department of Psychiatry and Psychology, Mayo Clinic, Jacksonville, FL, United States
[6]Department of Nursing, Mayo Clinic, Rochester, MN, United States
[7]Department of Medicine, University of Colorado School of Medicine, Aurora, CO, United States
[*]these authors contributed equally

Corresponding Author:
Liselotte N Dyrbye, MD, MPHE
Department of Medicine
University of Colorado School of Medicine
Mail Stop C290, Fitzsimons Bldg
13001 E 17th Pl. Rm #E1347
Aurora, CO, 80045
United States
Phone: 1 303 724 4982
Email: Liselotte.dyrbye@cuanschutz.edu

Abstract

Background: The occupational burnout epidemic is a growing issue, and in the United States, up to 60% of medical students, residents, physicians, and registered nurses experience symptoms. Wearable technologies may provide an opportunity to predict the onset of burnout and other forms of distress using physiological markers.

Objective: This study aims to identify physiological biomarkers of burnout, and establish what gaps are currently present in the use of wearable technologies for burnout prediction among health care professionals (HCPs).

Methods: A comprehensive search of several databases was performed on June 7, 2022. No date limits were set for the search. The databases were Ovid: MEDLINE(R), Embase, Healthstar, APA PsycInfo, Cochrane Central Register of Controlled Trials, Cochrane Database of Systematic Reviews, Web of Science Core Collection via Clarivate Analytics, Scopus via Elsevier, EBSCOhost: Academic Search Premier, CINAHL with Full Text, and Business Source Premier. Studies observing anxiety, burnout, stress, and depression using a wearable device worn by an HCP were included, with HCP defined as medical students, residents, physicians, and nurses. Bias was assessed using the Newcastle Ottawa Quality Assessment Form for Cohort Studies.

Results: The initial search yielded 505 papers, from which 10 (1.95%) studies were included in this review. The majority (n=9) used wrist-worn biosensors and described observational cohort studies (n=8), with a low risk of bias. While no physiological measures were reliably associated with burnout or anxiety, step count and time in bed were associated with depressive symptoms, and heart rate and heart rate variability were associated with acute stress. Studies were limited with long-term observations (eg, ≥12 months) and large sample sizes, with limited integration of wearable data with system-level information (eg, acuity) to predict burnout. Reporting standards were also insufficient, particularly in device adherence and sampling frequency used for physiological measurements.

Conclusions: With wearables offering promise for digital health assessments of human functioning, it is possible to see wearables as a frontier for predicting burnout. Future digital health studies exploring the utility of wearable technologies for burnout prediction should address the limitations of data standardization and strategies to improve adherence and inclusivity in study participation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50253) doi:10.2196/50253

KEYWORDS

wearable; healthcare professionals; burnout; digital health; mental health

Introduction

Burnout is an occupational syndrome characterized by emotional exhaustion, depersonalization, and feelings of reduced personal accomplishment caused by chronic, unmitigated high levels of job-related stress [1]. Burnout is common among health care professionals (HCPs, also referred to as health care workers), impacting an estimated 35% to 54% of nurses and physicians, and between 45% and 60% of medical students and resident physicians in the United States [2]. Several studies also reveal a high prevalence of depression and anxiety in HCPs that preceded the coronavirus pandemic [3-9]. Data further suggests that burnout and other forms of distress have increased among HCPs as a result of the COVID-19 pandemic [10-12].

This is concerning because the well-being of HCPs impacts the quality of patient care and patients' access to care. Several meta-analyses and systematic reviews have reported associations between burnout and negative impacts on the quality of care provided to patients, including increasing the risk of medical errors [13], malpractice claims [14], nosocomial infections [15], and mortality [16]. Additionally, other studies have found that HCPs who report experiencing burnout are more likely to reduce their time taking care of patients and quit, all of which negatively impact patient's access to care and add a burden to the global health care system [2]. The impacts of burnout go beyond the workplace, as HCPs with reported burnout are at increased risk of cardiovascular diseases [17,18], suicidal ideation [13,19], substance use disorders [20], uncontrolled stress [21], car accidents [22], and quality of life [23].

Contributors of burnout in HCPs are multifactorial and complex. While most factors contributing to burnout originate from system-level factors within the work environment, some risk factors originate from the personal domain or challenges in the personal-professional interface, such as work-home conflict (Figure 1). Due to the complexity of the factors involved, no model exists for predicting when an individual HCP or group of HCPs are at risk for developing burnout or other forms of distress. In response to the negative outcomes of burnout for HCPs and patients, the National Academies of Science, Engineering, and Medicine recommends health care organizations monitor (through frequent surveys) and respond to burnout. This approach is retrospective, as the time required for health care organizations to administer surveys, HCPs to complete them, and the additional time needed to analyze and interpret results all delay any response to burnout. A better approach would be a proactive one, where organizations or individual HCPs could predict and respond to high levels of job stress before the manifestation of burnout and associated personal and professional consequences result.

Figure 1. Wearable-augmented burnout management capturing the interplay of physiological and workplace factors.

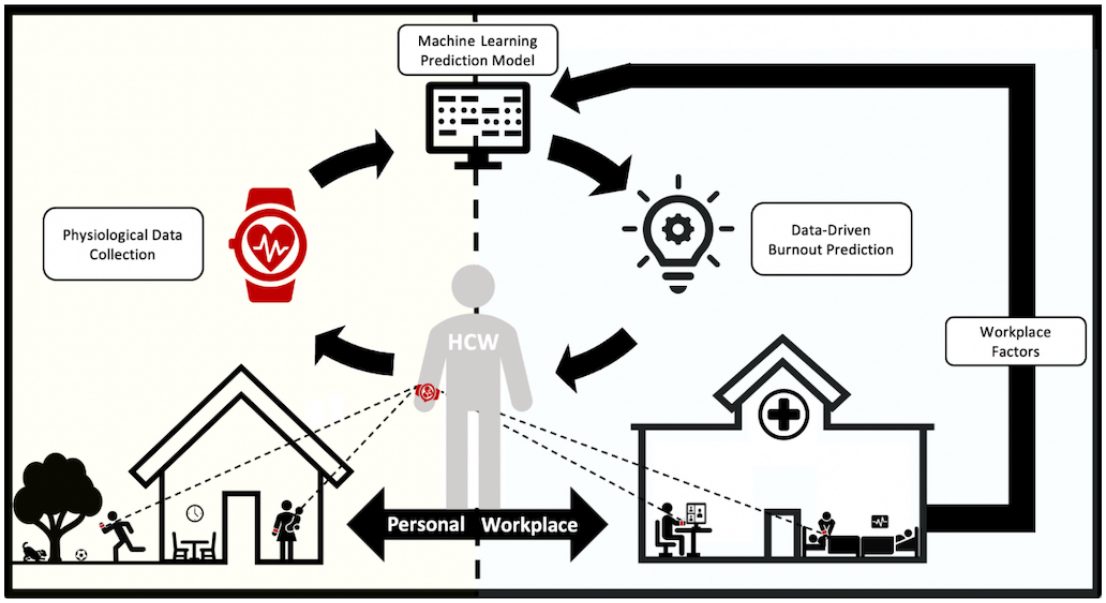

Previous studies and reviews suggest heart rate (HR) [24], heart rate variability (HRV) [24], sleep [25], and skin temperature [26] vary in response to stress. Additionally, sleep or fatigue also relates to the risk of burnout [27], depression [28], and other related conditions [29]. These types of data can be collected passively from wearable devices. Over the past 5 years, the adoption of wearable devices worldwide has more than doubled [30]. Therefore, data collected passively from wearable devices could potentially provide an avenue for detecting individuals at risk for high job stress, burnout, depression, and other related conditions. If predictive, such real-time information obtained passively from wearable devices could dramatically shift the current reactive paradigm to a proactive one, potentially leading to meaningful intervention before patients and HCPs experience adverse health consequences of burnout.

XSL•FO
RenderX

Previous systematic reviews suggest wearable devices may have some utility in predicting depression severity and stress levels [31]. To our knowledge, there is no review that investigates this relationship among HCPs or explores the ability of wearable devices to detect burnout risk. Hence, a scoping review was conducted to identify and summarize studies exploring associations between burnout, anxiety, depression, and stress, with data obtained from wearable devices in cohorts of HCPs.

Methods

Data Sources and Search Strategy

A comprehensive search of several databases was performed on June 7, 2022. No date limits were set for the search. The databases (and their coverage periods) were Ovid: MEDLINE (1946 to Present and Epub Ahead of Print, In-Process and Other Non-Indexed Citations and Daily), Embase (1974+), Healthstar (1966+), APA PsycInfo (1987+), Cochrane Central Register of Controlled Trials (1991+), Cochrane Database of Systematic Reviews (2005+), Web of Science Core Collection via Clarivate Analytics (1975+), Scopus via Elsevier (1788+), EBSCOhost: Academic Search Premier, CINAHL with Full Text (1981+), and Business Source Premier.

The search strategy was designed and conducted by a medical librarian (LCH) with input from the study's investigators (APA and LND). Controlled vocabulary supplemented with keywords was used. The actual strategies listing all search terms used and how they are combined are available in the Multimedia Appendix 1.

Review Strategy

The initial search yielded 505 papers. Two reviewers (MB and SS) independently identified and screened the titles and abstracts of potentially eligible papers. The inclusion criteria of the initial round of screening were as follows: the study must include a validated measure of burnout, stress, anxiety, or depression and the study must include only data from a wearable device worn by an HCP. For this work, we defined HCP as being a medical student, resident, practicing physician, or registered nurse in a hospital or outpatient clinical setting. The full-text reviews of the papers that resulted from the initial screening, data extraction, and quality assessment were also performed independently and in pairs by 2 reviewers (MB and SS). Papers were not excluded due to their calculated quality score. During this process, 475 papers were omitted because they did not satisfy the inclusion criteria (n=472) or were duplicates (n=3). After the initial screening, the full text of 30 papers was assessed for eligibility. Any disagreement was resolved by consensus with other senior reviewers (APA and LND) and the final source list was created, with senior reviewers blinded to reviews of each other and primary reviewers (MB and SS). The study selection process is illustrated in Figure 2. Tables 1 and 2 provide descriptions of the final 10 papers published from April 2017 to December 2021 included in this review.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram.

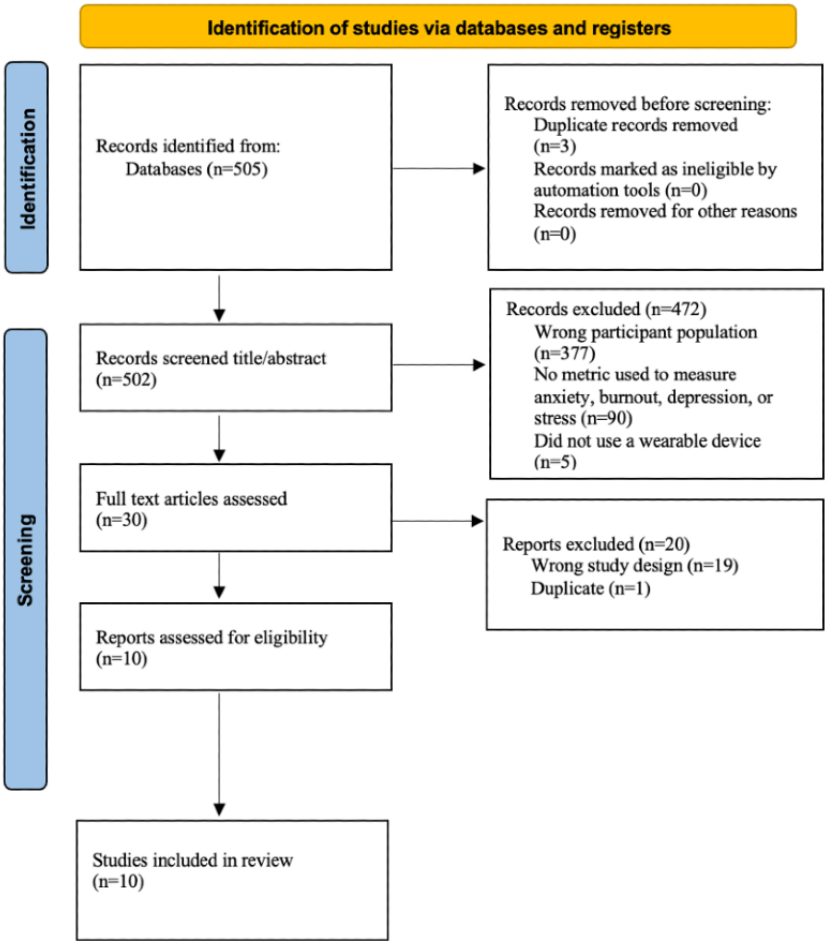

XSL•FO
RenderX

Table 1. Summary of included studies.

Author	Sample characteristics	Wearable-derived measurements	Validated anxiety, burnout, stress, or depression measures	Other measure included
Feng et al [32]	113 Nurses	HR[a], Sleep, and STC[b]	STAI[c]	Positive and Negative Affect Schedule, Satisfaction with Life Scale, Pittsburgh Sleep Quality Index, Affect EMA[d], Big Five Inventory-2, and Anxiety and Stress EMA[e]
Adler et al [33]	775 Residents	HR, Sleep, and STC	PHQ-9[f]	Mood EMA
Jevsevar et al [34]	21 Resident and Physicians	HRV[g], RHR[h], RR[i], and Sleep	MBI-Abbreviated	—[j]
Silva et al [35]	83 Medical students (19 had complete data)	HR and HRV	PSS-4[k]	—
Mendelsohn et al [36]	59 Residents	Sleep and STC	MBI-HSS[l]	Short-Form Health Survey, Epworth Sleepiness Scale, Satisfaction with Medicine Scale, and International Physical Activity Questionnaire
Marek et al [37]	28 Residents	RHR, Sleep, and STC	Single-item burnout measure	—
Sochacki et al [38]	21 Physicians	Sleep	MBI-HSS, PROMIS-29[m] (Depression and Anxiety)	—
Chaukos et al [39]	75 Residents (26 had complete data)	Activity level and Sleep	MBI–HSS, PSS-10, and PHQ-9	Functional Assessment of Chronic Illness Therapy-Fatigue, Penn State Worry Questionnaire, Revised Life Orientation Test, Interpersonal Reactivity Index Perspective-Taking subscale, Measure of Current Status-Part A, and Cognitive Affective Mindfulness Scale
de Looff et al [40]	114 Nurses	SC[n]	MBI–HSS (modified Dutch version)	—
Weenk et al [41]	20 Residents and Physicians	HR and HRV	STAI-short version	—

[a]HR: heart rate.
[b]STC: step count.
[c]STAI: State-Trait Anxiety Inventory.
[d]EDA: electrodermal activity.
[e]EMA: ecological momentary assessment.
[f]PHQ-9: Patient Health Questionnaire.
[g]HRV: heart rate variability.
[h]RHR: resting heart rate.
[i]RR: respiratory rate.
[j]Not available.
[k]PSS: Perceived Stress Scale.
[l]MBI-HSS: Maslach Burnout Inventory–Human Services Survey.
[m]PROMIS: Performance of the Patient-Reported Outcomes.
[n]SC: skin conductance.

XSL•FO
RenderX

Table 2. Primary findings and bias in studies using wearable devices in health care professionals.

Author	Device	Length of data collection	Primary findings	Newcastle Ottawa Scale Score
Feng et al [32]	Fitbit Charge 2	10 weeks	Baseline STAI[a] score did not relate to sensor-measured physical activity or sleep over the ensuing 10 weeks.	8
Adler et al [33]	Fitbit Charge 2	14 months	Quarterly measurements of change in depressive symptoms related to measured STC[b], sleep, and HR[c].	7
Jevsevar et al [34]	WHOOP	12 weeks	Being in the operating room related to the next day HRV[d]. Device reported sleep related to next-day HRV. Relationship between baseline burnout score and device measurements not reported.	8
Silva et al [35]	Microsoft Smart Band 2	2 weeks	Stress and HRV were both significantly different between the baseline and stress condition	8
Mendelsohn et al [36]	Fitbit Charge	14 days	Baseline burnout score did not relate to average daily sleep or STC over the ensuing 14 days.	7
Marek et al [37]	Fitbit Charge HR	16 weeks	Average daily sleep and activity level over a 2-4–week period did not relate to single-item burnout measure score. Average daily resting HR over a 2-4–week period was higher among residents with burnout versus those without burnout	8
Sochacki et al [38]	WHOOP	4 weeks	No significant association between weekly burnout score and device-measured hours of sleep over 4 weeks.	8
Chaukos et al [39]	Basis Health Tracker	6 months	No association between baseline depressive symptoms or stress levels and device-measured sleep or activity levels over 30 or 90 days of the study. No association between chronic burnout (burnout at 2 time points), never burned out, new burnout (burnout at 2nd but not 1st time point), and unknown burnout status (survey not completed) and devise measured sleep or activity level aggregated over first 30 days.	6
de Looff et al [40]	Empatica E4	1 day or night shift	Skin conductance collected over 1 shift among nursing staff did not correlate with burnout scores collected on questionnaires completed within 2 days of wearing the device (mean 2.4, SD 10 days; range 0-44 days).	8
Weenk et al [41]	HealthPatch	Up to 3 days (at least 2)	Stress measured by the patch increased during surgery, more so for less experienced trainees, but did not correlate with change in STAI score before or after surgery, perhaps due to small sample size or lack of sensitivity to change.	8

[a]STAI: State-Trait Anxiety Inventory.

[b]STC: step count.

[c]HR: heart rate.

[d]HRV: heart rate variability.

Extraction Strategy

Data extraction was mostly completed by a single researcher (MB). Other researchers (APA and SS) helped refine data extraction and review the tables. The following information was extracted from the papers and is included in Tables 1 and 2: sample population (size and occupation), anxiety, burnout, stress or depression assessment instrument, additional measurements used, wearable device used, measured physiological variable, study duration, primary findings, and the author-determined quality assessment score.

Quality Assessment

The methodological quality of nonrandomized or observational studies was assessed by 2 reviewers (MB and SS) using the Newcastle Ottawa Quality Assessment Form for Cohort Studies [42]. The Newcastle-Ottawa Scale is a validated scale of 8 items in 3 domains: selection, comparability, and outcome. Studies are rated from 0 to 9, with those studies rating 0-2 (poor quality), 3-5 (fair quality), and 6-9 (good or high quality). All 10 studies received a Newcastle-Ottawa Scale rating of good or high quality.

Results

Roles of Participating Health Care Professionals

Among the 10 reviewed studies, 8 were conducted in the United States, 1 study was conducted in Portugal [35], and another one was conducted in Canada [36]. Seven studies recruited either resident physicians (postgraduate medical trainees), practicing physicians, or a combination of both, primarily within the same specialty (eg, orthopedic surgery and emergency medicine). Two studies recruited registered nurses [32,40] and 1 study recruited medical students [35]. Sample sizes ranged from 20

XSL•FO
RenderX

to 775 participants per study (see Table 1). Only 3 studies had more than 100 participants [32,33,40].

Wearable Devices, Physiological Variables Collected, and Duration of Observation

Table 1 summarizes the sample population, sample size, physiological variables collected from wearable devices, and psychometrics used in the 10 studies. The devices used, length of data collection, and primary findings are listed in Table 2. Out of the 10 studies, 9 used wrist-worn biosensors, such as the Fitbit Charge (n=4) [32,33,35,40] WHOOP (n=2) [34,38], Basis B1 (n=1) [35], Empatica E4 (n=1) [40], and the Microsoft Smart Band 2 (n=1) [35]. Sensors embedded within wrist-worn biosensors included optical heart sensors, electrical heart sensors, accelerometers, and skin temperature sensors. The other device used was a HealthPatch, an adhesive patch with 2 ECG electrodes used to measure HR and HRV. A variety of physiological variables were collected, with sleep being the most common, measured in 7 studies. Studies ranged in length of data collection, from a single 12-hour shift to a 14-month period. Only 5 studies collected data for more than 10 weeks [32-34,37,39].

Methodological Wearable Data Reporting

Only 2 studies explicitly stated the sampling frequency used when processing data from the wearable device [33,39]. Four of the studies discussed how the data were processed; however, the level of detail varied [32,33,35,40]. Three of the studies indicated the cutoff values for physiological variables or explained how outliers were addressed [32,33,40]. Only 4 studies explicitly stated how much raw data were retrieved from the devices [32-34,36].

Reported Relationships Among Burnout, Depressive Symptoms, Stress, and Anxiety With Data Obtained From Wearable Devices

Burnout

Of the 10 included studies, 6 included a measure of burnout (Table 1) [34,36-40]. Four of these 6 studies used the Maslach Burnout Inventory–Human Services Survey (MBI-HSS) [43]. In a cross-sectional study of 114 nurses, no relationship was found between MBI-HSS score and skin conductance, a measure of autonomic nervous activity, collected through an Empatica E4, for 1 shift [40]. Another study investigated the relationship between MBI-HSS score, self-reported work hours, physical activity, and sleep, as measured by a Fitbit, in a cohort of 59 residents [36]. No relationship was found between the change in burnout score and data collected from the Fitbit over 2 weeks. In the third study, no relationship was found between MBI-HSS score and sleep, as measured by a WHOOP, over the course of 4 weeks [38]. Last, in a study of 75 medicine and psychiatry residents, no relationship was found between burnout score and sleep or activity levels, as measured by Basis B1 health-tracking device, during their first 6 months of residency [39].

Two studies measured burnout using scales other than the 22-item MBI-HSS (widely considered the gold standard) [34,37]. In a study of 21 orthopedic residents and surgeons, no association was found between baseline abbreviated MBI scores and WHOOP measures collected over 12 weeks [34]. The final study investigated the association between burnout, as measured by a commonly used single-item measure, and sleep and activity level, as measured by a Fitbit. In this study, of 28 emergency medicine residents, there was no association between burnout scores and sleep or activity levels over the course of the 16-week study [37].

Depressive Symptoms, Stress, and Anxiety

A 14-month study of 775 medical residents found a relationship between depressive symptoms, as measured by the 9-item Patient Health Questionnaire [44], and step count (STC) and sleep as measured by a Fitbit Charge 2 [33]. Medical residents whose depressive symptoms worsened over the period of the study had a significantly higher skew in their hourly STC distributions and spent less time in bed than those whose symptoms did not worsen. In a study of 83 medical students, Perceived Stress Scale-4 scores related to HR and HRV, were measured by a Microsoft Smartband 2, at baseline and during an examination [35].

In a 10-week study of 113 nurses led by Feng et al [32], no relationship was found between the level of anxiety, as measured by the State-Trait Anxiety Inventory (STAI) [45], and wearable sensor data (eg, sleep and HR) collected using Fitbit Charge 2 smartwatch. Weenk et al [41] conducted a study of 20 surgeons and surgical residents who completed an abbreviated version of the STAI before and after performing surgery, and wore a HealthPatch. This adhesive patch calculates stress using an HR and HRV-dependent algorithm for 48 to 72 hours [41]. There was no correlation found between the STAI score and HealthPatch data.

Device Use Compliance and Experience

Seven studies reported data on participant adherence or experience with wearable devices. Chaukos et al [39] reported that 25 (40%) of their participants wore their device for more than 50% of the time for the first 3 months of the study, while another 13 (21%) participants wore the device for more than 75% of the time for the first 3 months. Other studies, such as one conducted by Sochacki et al [38] reported that of the 26 participants, 5 did not complete the minimum WHOOP compliance (4 weeks). Surgeons involved in a study by Jevsevar et al [34] reported a high percentage of device compliance at 83.2% of the total collection window, similar to the 93% compliance rate reported by Mendelsohn et al [36] and Sochacki et al [38]. Weenk et al [41] reported that 6 of 20 individuals experienced problems with their HealthPatch, similar to Marek et al [37] who reported 1 of 30 participants dropped out due to fitness tracker intolerance. Problems included connection failure (n=2), loss of skin contact (n=2), and skin irritation (n=2). Feng et al [32] noted similar compliance between day-shift participants and night-shift participants (number of recordings day-shift: mean 44.6, SD 3.1 sessions; night-shift: mean 45, SD 20.2 sessions).

Risk of Bias

A risk of bias of assessment was completed for the 8 cohort studies and 1 cross-sectional study (Figure 3). While the risk

XSL•FO
RenderX

of bias was generally low across the studies, none included a comparison group of participants who did not wear a device.

Figure 3. Risk of bias assessment for observational cohort studies.

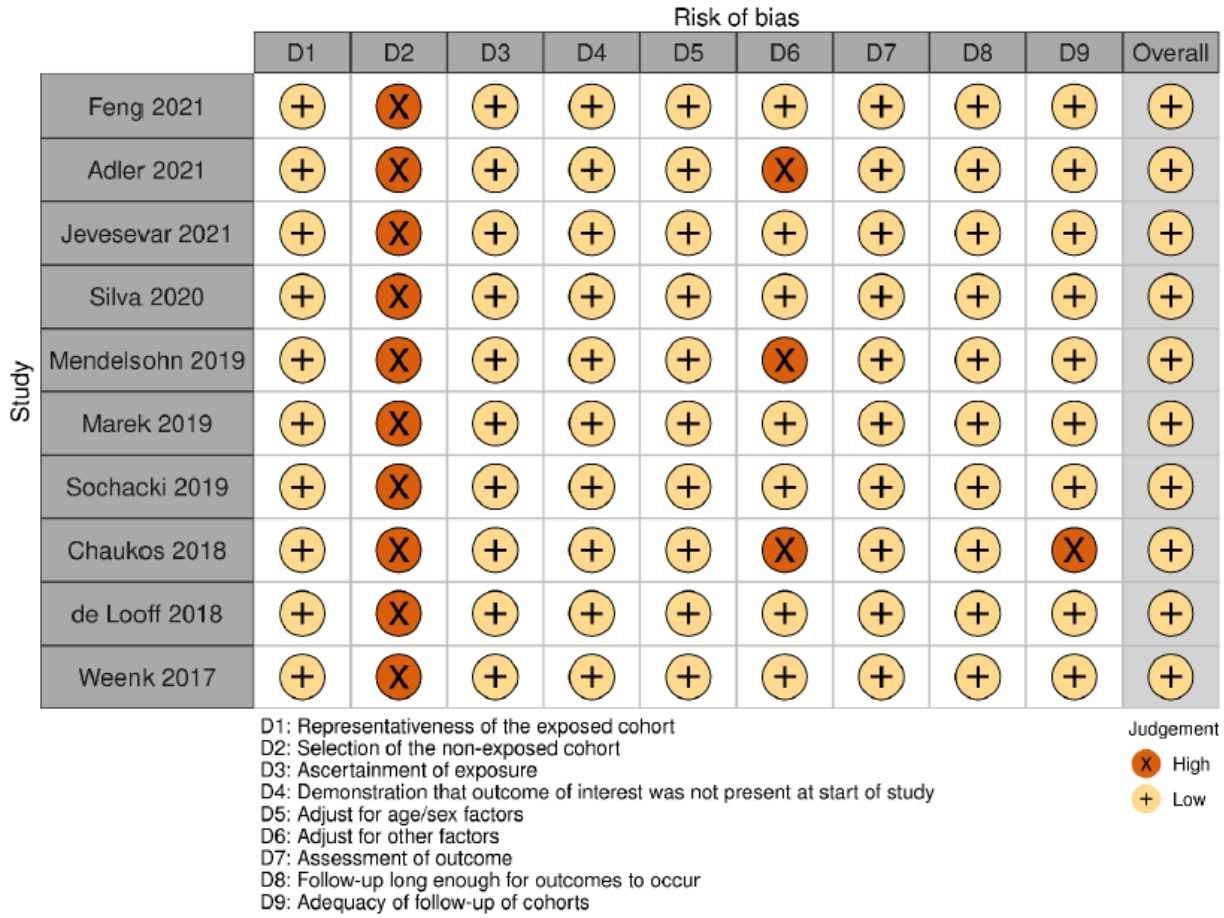

Discussion

To our knowledge, this is the first scoping review to investigate the use of wearable technologies for the prediction of burnout, anxiety, depression, and stress in HCPs. Among the 10 studies identified, a range of wearables collected data on HR, HRV, respiratory rate, skin temperature, sleep, and activity levels from a single shift of work and up to 14 months of data collection in relatively small samples of physicians, medical students, and nurses. In these studies, no relationships were found between collected physiological data from wearables and burnout or anxiety. One study reported a relationship between STC, time in bed, and depressive symptoms, and another between HR, HRV, and acute stress (during an examination). Identified studies had methodological limitations, including short duration which limits the capture of naturalistic variations in the workplace stressors.

In this review, 3 studies measured HRV [34,35,41] and only 1 found a significant relationship between HRV and acute stress. A previous systematic review involving non-HCPs identified 2 studies demonstrating relationships between HRV and acute stress-induced conditions and 1 study demonstrating a relationship between HRV and stress levels measured by catecholamine levels [31]. This previous systematic review also identified 1 study where in a setting of laboratory-induced stress, HRV parameters related to STAI score. These studies, however, differed substantially from the ones included in this review. For example, none of them collected physiological data longer than 24 minutes, stress was induced in a laboratory setting (vs occurring naturally in a work setting), and only 1 study compared physiological data with a self-reported stress measure (ie, STAI score).

Given these early findings, further research focusing on the following elements of rigor are warranted. First, the length of observation should be long enough (at least 2 or 3 consecutive quarters of a calendar year) to allow sufficient quanta of wearable data to capture fluctuations in and chronicity of workplace stress. Studies should systematically collect data using validated instruments measuring burnout (eg, MBI-HSS [43]), depression (eg, Center for Epidemiologic Studies Depression Scale [46] and Patient Health Questionnaire-9 [44]), and anxiety (eg, General Anxiety Disorder-7 [47]). Investigators may also want to consider designing cohorts comprising groups of HCPs defined by their type of medical specialty or practice location. For example, it is possible that workplace stressors, patient acuity, and job demand fluctuate between primary care and surgical specialties and between outpatient practices and hospital-based practices. Hence, the burnout biomarkers may vary between practices. Considering that burnout is defined as when job demands exceed job resources, it is possible that the workplace (eg, patient acuity and hospital bed size) and related staffing factors (eg, workload, shift length, and availability of

XSL•FO
RenderX

support staff) impact physiological biomarkers collected from wearables. Hence, future studies should consider collecting organizational variables to better understand the systemic contributors of burnout. Additionally, given the era of decentralized health care practice (eg, nontraditional shift days/hours and remote care with augmented reality), studies engaging with HCPs may benefit from no-contact passive monitoring and a digital app interface for survey collection (ie, decentralized trail). Finally, there is a bioethics component to understand how wearables can be successfully integrated into workforces' burnout management. Greater attention needs to be paid to participant engagement, including addressing comfort with wearing the device, resolving discrepancies in wearable-derived data versus self-reported data, and understanding factors that influence perceptions of fatigue but not recorded sleep [37,48,49].

The use of wearables to detect the functioning states of human beings is an active and rapidly evolving field. Several wearable-based studies have been shown to aid in the detection of mental health conditions or resilience in quality of life [50] through mindfulness practices including physical activity [51] and sleep [52-54] monitoring. Prior work has demonstrated that aspects of physical functioning when combined with data during the day could predict variations in aspects of QoL and mental well-being [55-58]. Work by Campbell et al [59-64] has demonstrated the ability of daily journaling, wearables, and mobile assessments to detect depressive symptoms and mental states in patients with schizophrenia. These prior efforts in the field of mental health and the work summarized in this scoping review demonstrate the promise of wearables in predicting states of one's functioning, including burnout. However, a consensus is lacking on the best approaches to collecting, processing, and reporting physiological data, much like CONSORT (Consolidated Standards of Reporting Trials) [65] for reporting randomized trials and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [66] guidelines for reporting observational studies. Standardization of variables should include the creation of a guideline for reporting the sampling frequency, device adherence, and other information regarding device parameters that impact data collection. Such standardization would assist with generalizing findings, validating predictive algorithms, informing meta-analysis, and the use of data for retraining predictive models regardless of the wearable's make and model. Additionally, there needs to be consensus around approaches to address bioethics, privacy, and confidentiality concerns of participants [67,68]. Predictive technologies, informed by personal biometric or physiologic data, may help improve work conditions but could also place individuals' privacy or perhaps even their job security at risk.

This study has limitations. Only studies that included physicians, resident physicians, medical students, and nurses and were published in English were included. Following the 2019 pandemic, physicians identifying as 2 or more races experienced the highest levels of burnout onset, according to a report by the American Medical Association [69]. Furthermore, there are known disparities in the access to, and the use of digital health technologies in underrepresented minorities [70,71]. Therefore, it is vital to understand the factors that cause burnout in these groups of professionals and remove barriers to access to personalized wellness technologies using wearables that may help understand and mitigate burnout. In the context of the use and access of digital health for burnout, 8 of the 10 studies reported the gender breakdown of participants, and only 1 study reported the race of their participants. With the urgent need to broaden access to digital health solutions to study and understand burnout, future efforts should (1) follow reporting guidelines (eg, set by National Institutes of Health in the Human Subjects sections) to report on participant characteristics by ethnicity, race, and gender, and (2) innovate study procedures (eg, decentralized protocols) that improve the recruitment and engagement of underrepresented minorities in digital health studies of burnout. Although we sought to include validated measures of burnout, stress, depression, and anxiety, the instruments used in the studies varied in their psychometric strengths. Finally, most studies lacked power calculations, making findings, effect sizes, or impact of dropouts difficult to interpret from the perspective of the generalizability of biomarkers.

Despite the popularity of wearable devices, only 10 studies were identified that explored relationships between physiological data and burnout, depressive symptoms, stress, or anxiety. Most of these studies had substantial methodological limitations, and nearly all reported limited data collection and processing information, participant experience with the wearable device, and device compliance. Standardizing study procedures, common data elements, and reporting of wearable data are needed to strengthen the rigor of digital health studies. Addressing these limitations will result in improvements in wearable device research, including data standardization and reporting, that will validate their use in providing early intervention for HCP wellness. Additional research is warranted to explore the potential of wearable devices, perhaps augmented with other system-level data (eg, work shift lengths and absenteeism), to predict burnout and other forms of distress, hopefully leading to meaningful action before it has an adverse impact on HCPs and patient care.

Acknowledgments

This study was partially supported by the Mayo Clinic Summer Undergraduate Research Fellowship, National Science Foundation (grant 2041339); National Institutes of Health (grant R01 NR020362); the Mayo Clinic Center for Individualized Medicine, and the Mayo Clinic Center for Clinical and Translational Science. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or National Institutes of Health.

XSL•FO
RenderX

Authors' Contributions

APA and LND contributed equally as co-Corresponding Authors. APA may be contacted at athreya.arjun@mayo.edu.

Conflicts of Interest

PC has received research support from the National Institutes of Health (NIH), National Science Foundation (NSF), Brain and Behavior Research Foundation, and the Mayo Clinic Foundation. PC has received research support from Pfizer, Inc. He has received equipment support from Neuronetics, Inc, and MagVenture, Inc. He received grant-in-kind supplies and genotyping from Assurex Health, Inc for an investigator-initiated study. He served as the primary investigator for a multicenter study funded by Neuronetics, Inc and a site primary investigator for a study funded by NeoSync, Inc. PC served as a paid consultant for Engrail Therapeutics, Sunovion, Procter and Gamble Company, Meta Platforms, Inc, and Myriad Neuroscience. PC is employed by the Mayo Clinic. LD is a coinventor of the Well-Being Index and its derivatives which Mayo Clinic has licensed. LD receives royalties. WB's research has been supported by the NIMH, NINR, NSF, the Blue Gator Foundation, the Watzinger Foundation, and the Mayo Foundation for Medical Education and Research. He has contributed chapters to UpToDate concerning the pharmacological management of patients with bipolar spectrum disorders. MCs research has been supported by NSF and the Mayo Foundation for Medical Education and Research.

Multimedia Appendix 1
Search strategy.
[DOCX File , 44 KB - jmir_v26i1e50253_app1.docx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 499 KB - jmir_v26i1e50253_app2.pdf]

References

1. ICD-11 for mortality and morbidity statistics. World Health Organization. 2023. URL: http://id.who.int/icd/entity/129180281 [accessed 2024-04-10]
2. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington, DC: National Academies Press; 2019.
3. Golonka K, Mojsa-Kaja J, Blukacz M, Gawłowska M, Marek T. Occupational burnout and its overlapping effect with depression and anxiety. Int J Occup Med Environ Health 2019;32(2):229-244 [FREE Full text] [doi: 10.13075/ijomeh.1896.01323] [Medline: 30855601]
4. Mata DA, Ramos MA, Bansal N, Khan R, Guille C, Di Angelantonio E, et al. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. JAMA 2015;314(22):2373-2383 [FREE Full text] [doi: 10.1001/jama.2015.15845] [Medline: 26647259]
5. Melnyk BM, Kelly SA, Stephens J, Dhakal K, McGovern C, Tucker S, et al. Interventions to improve mental health, well-being, physical health, and lifestyle behaviors in physicians and nurses: a systematic review. Am J Health Promot 2020;34(8):929-941 [FREE Full text] [doi: 10.1177/0890117120920451] [Medline: 32338522]
6. Ernst J, Jordan KD, Weilenmann S, Sazpinar O, Gehrke S, Paolercio F, et al. Burnout, depression and anxiety among Swiss medical students—a network analysis. J Psychiatr Res 2021;143:196-201 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.09.017] [Medline: 34500349]
7. Appiani FJ, Cairoli FR, Sarotto L, Yaryour C, Basile ME, Duarte JM. Prevalence of stress, burnout syndrome, anxiety and depression among physicians of a teaching hospital during the COVID-19 pandemic. [prevalencia de estrés, síndrome de desgaste profesional, ansiedad y depresión en médicos de un hospital universitario durante la pandemia de COVID-19]. Arch Argent Pediatr 2021;119(5):317-324 [FREE Full text] [doi: 10.5546/aap.2021.eng.317] [Medline: 34569739]
8. Colville GA, Smith JG, Brierley J, Citron K, Nguru NM, Shaunak PD, et al. Coping with staff burnout and work-related posttraumatic stress in intensive care. Pediatr Crit Care Med 2017;18(7):e267-e273. [doi: 10.1097/PCC.0000000000001179] [Medline: 28459762]
9. Misiolek-Marín A, Soto-Rubio A, Misiolek H, Gil-Monte PR. Influence of burnout and feelings of guilt on depression and health in anesthesiologists. Int J Environ Res Public Health 2020;17(24):9267 [FREE Full text] [doi: 10.3390/ijerph17249267] [Medline: 33322399]
10. Shanafelt TD, West CP, Dyrbye LN, Trockel M, Tutty M, Wang H, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022;97(12):2248-2258 [FREE Full text] [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]
11. Sexton JB, Adair KC, Proulx J, Profit J, Cui X, Bae J, et al. Emotional exhaustion among US health care workers before and during the COVID-19 pandemic, 2019-2021. JAMA Netw Open 2022;5(9):e2232748 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.32748] [Medline: 36129705]

XSL•FO
RenderX

12. Jaconia GD, Lynch LR, Miller LK, Hines RL, Pinyavat T. COVID-19 impact on resident mental health and well-being. J Neurosurg Anesthesiol 2022;34(1):122-126. [doi: 10.1097/ANA.0000000000000806] [Medline: 34870634]
13. Menon NK, Shanafelt TD, Sinsky CA, Linzer M, Carlasare L, Brady KJS, et al. Association of physician burnout with suicidal ideation and medical errors. JAMA Netw Open 2020;3(12):e2028780 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.28780] [Medline: 33295977]
14. Austin EE, Do V, Nullwala R, Pulido DF, Hibbert PD, Braithwaite J, et al. Systematic review of the factors and the key indicators that identify doctors at risk of complaints, malpractice claims or impaired performance. BMJ Open 2021;11(8):e050377 [FREE Full text] [doi: 10.1136/bmjopen-2021-050377] [Medline: 34429317]
15. Noah N, Potas N. Association between nursing work stress, burnout and nosocomial infection rate in a neonatal intensive care unit in Hargeisa, Somaliland. Trop Doct 2022;52(1):46-52 [FREE Full text] [doi: 10.1177/00494755211055250] [Medline: 34806491]
16. Schlak AE, Aiken LH, Chittams J, Poghosyan L, McHugh M. Leveraging the work environment to minimize the negative impact of nurse burnout on patient outcomes. Int J Environ Res Public Health 2021;18(2):610 [FREE Full text] [doi: 10.3390/ijerph18020610] [Medline: 33445764]
17. Melamed S, Shirom A, Toker S, Berliner S, Shapira I. Burnout and risk of cardiovascular disease: evidence, possible causal paths, and promising research directions. Psychol Bull 2006;132(3):327-353. [doi: 10.1037/0033-2909.132.3.327] [Medline: 16719565]
18. Melamed S, Kushnir T, Shirom A. Burnout and risk factors for cardiovascular diseases. Behav Med 1992;18(2):53-60. [doi: 10.1080/08964289.1992.9935172] [Medline: 1392214]
19. Kuhn CM, Flanagan EM. Self-care as a professional imperative: physician burnout, depression, and suicide. [prendre soin de soi, un impératif professionnel: l'épuisement professionnel, la dépression et le suicide chez les médecins]. Can J Anaesth 2017;64(2):158-168 [FREE Full text] [doi: 10.1007/s12630-016-0781-0] [Medline: 27910035]
20. Lacy BE, Chan JL. Physician burnout: the hidden health care crisis. Clin Gastroenterol Hepatol 2018;16(3):311-317 [FREE Full text] [doi: 10.1016/j.cgh.2017.06.043] [Medline: 28669661]
21. Arnsten AFT, Shanafelt T. Physician distress and burnout: the neurobiological perspective. Mayo Clin Proc 2021;96(3):763-769 [FREE Full text] [doi: 10.1016/j.mayocp.2020.12.027] [Medline: 33673923]
22. HaGani N, Hershler ME, Ben Shlush E. The relationship between burnout, commuting crashes and drowsy driving among hospital health care workers. Int Arch Occup Environ Health 2022;95(6):1357-1367 [FREE Full text] [doi: 10.1007/s00420-022-01855-7] [Medline: 35318536]
23. Pulcrano M, Evans SRT, Sosin M. Quality of life and burnout rates across surgical specialties: a systematic review. JAMA Surg 2016;151(10):970-978. [doi: 10.1001/jamasurg.2016.1647] [Medline: 27410167]
24. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018;15(3):235-245 [FREE Full text] [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
25. Hirotsu C, Tufik S, Andersen ML. Interactions between sleep, stress, and metabolism: from physiological to pathological conditions. Sleep Sci 2015;8(3):143-152 [FREE Full text] [doi: 10.1016/j.slsci.2015.09.002] [Medline: 26779321]
26. Herborn KA, Graves JL, Jerem P, Evans NP, Nager R, McCafferty DJ, et al. Skin temperature reveals the intensity of acute stress. Physiol Behav 2015;152(Pt A):225-230 [FREE Full text] [doi: 10.1016/j.physbeh.2015.09.032] [Medline: 26434785]
27. Membrive-Jiménez MJ, Gómez-Urquiza JL, Suleiman-Martos N, Velando-Soriano A, Ariza T, De la Fuente-Solana EI, et al. Relation between burnout and sleep problems in nurses: a systematic review with meta-analysis. Healthcare (Basel) 2022;10(5):954 [FREE Full text] [doi: 10.3390/healthcare10050954] [Medline: 35628091]
28. Dong L, Xie Y, Zou X. Association between sleep duration and depression in US adults: a cross-sectional study. J Affect Disord 2022;296:183-188 [FREE Full text] [doi: 10.1016/j.jad.2021.09.075] [Medline: 34607059]
29. Cox RC, Olatunji BO. A systematic review of sleep disturbance in anxiety and related disorders. J Anxiety Disord 2016;37:104-129. [doi: 10.1016/j.janxdis.2015.12.001] [Medline: 26745517]
30. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. The Networking REPORT. 2019. URL: https://networking.report/whitepapers/cisco-visual-networking-index-global-mobile-data-trafficforecast-update-2017%E2%80%932022 [accessed 2024-04-10]
31. Hickey BA, Chalmers T, Newton P, Lin CT, Sibbritt D, McLachlan CS, et al. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: a systematic review. Sensors (Basel) 2021;21(10):3461 [FREE Full text] [doi: 10.3390/s21103461] [Medline: 34065620]
32. Feng T, Booth BM, Baldwin-Rodríguez B, Osorno F, Narayanan S. A multimodal analysis of physical activity, sleep, and work shift in nurses with wearable sensor data. Sci Rep 2021;11(1):8693 [FREE Full text] [doi: 10.1038/s41598-021-87029-w] [Medline: 33888731]
33. Adler DA, Tseng VWS, Qi G, Scarpa J, Sen S, Choudhury T. Identifying mobile sensing indicators of stress-resilience. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021;5(2):51 [FREE Full text] [doi: 10.1145/3463528] [Medline: 35445162]
34. Jevsevar DS, Molloy IB, Gitajn IL, Werth PM. Orthopaedic surgeon physiological indicators of strain as measured by a wearable fitness device. J Am Acad Orthop Surg 2021;29(24):e1378-e1386. [doi: 10.5435/JAAOS-D-21-00078] [Medline: 33999882]

XSL•FO
RenderX

35. Silva E, Aguiar J, Reis LP, Sá JOE, Gonçalves J, Carvalho V. Stress among Portuguese medical students: the EuStress solution. J Med Syst 2020;44(2):45. [doi: 10.1007/s10916-019-1520-1] [Medline: 31897774]
36. Mendelsohn D, Despot I, Gooderham PA, Singhal A, Redekop GJ, Toyota BD. Impact of work hours and sleep on well-being and burnout for physicians-in-training: the resident activity tracker evaluation study. Med Educ 2019;53(3):306-315. [doi: 10.1111/medu.13757] [Medline: 30485496]
37. Marek AP, Nygaard RM, Liang ET, Roetker NS, DeLaquil M, Gregorich S, et al. The association between objectively-measured activity, sleep, call responsibilities, and burnout in a resident cohort. BMC Med Educ 2019;19(1):158 [FREE Full text] [doi: 10.1186/s12909-019-1592-0] [Medline: 31113435]
38. Sochacki KR, Dong D, Peterson L, McCulloch PC, Harris JD. The measurement of orthopaedic surgeon burnout using a validated wearable device. Arthrosc Sports Med Rehabil 2019;1(2):e115-e121 [FREE Full text] [doi: 10.1016/j.asmr.2019.09.004] [Medline: 32266348]
39. Chaukos D, Chad-Friedman E, Mehta DH, Byerly L, Celik A, McCoy TH, et al. SMART-R: a prospective cohort study of a resilience curriculum for residents by residents. Acad Psychiatry 2018;42(1):78-83. [doi: 10.1007/s40596-017-0808-z] [Medline: 29098597]
40. de Looff P, Nijman H, Didden R, Embregts P. Burnout symptoms in forensic psychiatric nurses and their associations with personality, emotional intelligence and client aggression: a cross-sectional study. J Psychiatr Ment Health Nurs 2018;25(8):506-516 [FREE Full text] [doi: 10.1111/jpm.12496] [Medline: 30199590]
41. Weenk M, Alken APB, Engelen LJLPG, Bredie SJH, van de Belt TH, van Goor H. Stress measurement in surgeons and residents using a smart patch. Am J Surg 2018;216(2):361-368. [doi: 10.1016/j.amjsurg.2017.05.015] [Medline: 28615137]
42. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-605. [doi: 10.1007/s10654-010-9491-z] [Medline: 20652370]
43. Maslach C, Jackson SE, Leiter MP. Maslach Burnout Inventory Manual, 3rd Edition. Palo Alto, CA: ConsultingPsychologists Press; 1996.
44. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
45. Spielberger CD. State-trait anxiety inventory for adults. American Psychological Association. 1983. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft06496-000 [accessed 2024-04-10]
46. Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): a systematic review with meta-analysis. PLoS One 2016;11(5):e0155431 [FREE Full text] [doi: 10.1371/journal.pone.0155431] [Medline: 27182821]
47. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
48. Sargent MC, Sotile W, Sotile MO, Rubash H, Barrack RL. Stress and coping among orthopaedic surgery residents and faculty. J Bone Joint Surg Am 2004;86(7):1579-1586. [doi: 10.2106/00004623-200407000-00032] [Medline: 15252111]
49. Sargent MC, Sotile W, Sotile MO, Rubash H, Barrack RL. Quality of life during orthopaedic training and academic practice. part 1: orthopaedic surgery residents and faculty. J Bone Joint Surg Am 2009;91(10):2395-2405. [doi: 10.2106/JBJS.H.00665] [Medline: 19797575]
50. Hirten RP, Suprun M, Danieletto M, Zweig M, Golden E, Pyzik R, et al. A machine learning approach to determine resilience utilizing wearable device data: analysis of an observational cohort. JAMIA Open 2023;6(2):ooad029 [FREE Full text] [doi: 10.1093/jamiaopen/ooad029] [Medline: 37143859]
51. Ghosh K, Nanda S, Hurt RT, Schroeder DR, West CP, Fischer KM, et al. Mindfulness using a wearable brain sensing device for health care professionals during a pandemic: a pilot program. J Prim Care Community Health 2023;14:21501319231162308 [FREE Full text] [doi: 10.1177/21501319231162308] [Medline: 36960553]
52. Huberty J, Green J, Glissmann C, Larkey L, Puzia M, Lee C. Efficacy of the mindfulness meditation mobile app "calm" to reduce stress among college students: randomized controlled trial. JMIR Mhealth Uhealth 2019;7(6):e14273 [FREE Full text] [doi: 10.2196/14273] [Medline: 31237569]
53. Pyrkov TV, Sokolov IS, Fedichev PO. Deep longitudinal phenotyping of wearable sensor data reveals independent markers of longevity, stress, and resilience. Aging (Albany NY) 2021;13(6):7900-7913 [FREE Full text] [doi: 10.18632/aging.202816] [Medline: 33735108]
54. Thota D. Evaluating the relationship between fitbit sleep data and self-reported mood, sleep, and environmental contextual factors in healthy adults: pilot observational cohort study. JMIR Form Res 2020;4(9):e18086 [FREE Full text] [doi: 10.2196/18086] [Medline: 32990631]
55. Fischer D, McHill AW, Sano A, Picard RW, Barger LK, Czeisler CA, et al. Irregular sleep and event schedules are associated with poorer self-reported well-being in US college students. Sleep 2020;43(6):zsz300 [FREE Full text] [doi: 10.1093/sleep/zsz300] [Medline: 31837266]
56. Sano A, Yu AZ, McHill AW, Phillips AJK, Taylor S, Jaques N, et al. Prediction of happy-sad mood from daily behaviors and previous sleep history. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:6796-6799 [FREE Full text] [doi: 10.1109/EMBC.2015.7319954] [Medline: 26737854]

XSL•FO
RenderX

57. Umematsu T, Sano A, Picard RW. Daytime data and LSTM can forecast tomorrow's stress, health, and happiness. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:2186-2190. [doi: 10.1109/EMBC.2019.8856862] [Medline: 31946335]
58. Umematsu T, Sano A, Taylor S, Tsujikawa M, Picard RW. Forecasting stress, mood, and health from daytime physiology in office workers and students. Annu Int Conf IEEE Eng Med Biol Soc 2020;2020:5953-5957. [doi: 10.1109/EMBC44109.2020.9176706] [Medline: 33019329]
59. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015;38(3):218-226 [FREE Full text] [doi: 10.1037/prj0000130] [Medline: 25844912]
60. Buck B, Hallgren KA, Campbell AT, Choudhury T, Kane JM, Ben-Zeev D. mHealth-assisted detection of precursors to relapse in schizophrenia. Front Psychiatry 2021;12:642200 [FREE Full text] [doi: 10.3389/fpsyt.2021.642200] [Medline: 34135781]
61. daSilva AW, Huckins JF, Wang W, Wang R, Campbell AT, Meyer ML. Daily perceived stress predicts less next day social interaction: evidence from a naturalistic mobile sensing study. Emotion 2021;21(8):1760-1770. [doi: 10.1037/emo0000994] [Medline: 35041440]
62. Huckins JF, DaSilva AW, Hedlund EL, Murphy EI, Rogers C, Wang W, et al. Causal factors of anxiety and depression in college students: longitudinal ecological momentary assessment and causal analysis using Peter and Clark momentary conditional independence. JMIR Ment Health 2020;7(6):e16684 [FREE Full text] [doi: 10.2196/16684] [Medline: 32519971]
63. Huckins JF, daSilva AW, Wang R, Wang W, Hedlund EL, Murphy EI, et al. Fusing mobile phone sensing and brain imaging to assess depression in college students. Front Neurosci 2019;13:248 [FREE Full text] [doi: 10.3389/fnins.2019.00248] [Medline: 30949024]
64. Huckins JF, daSilva AW, Wang W, Hedlund E, Rogers C, Nepal SK, et al. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 2020;22(6):e20185 [FREE Full text] [doi: 10.2196/20185] [Medline: 32519963]
65. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013;309(8):814-822 [FREE Full text] [doi: 10.1001/jama.2013.879] [Medline: 23443445]
66. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Int J Surg 2014;12(12):1495-1499 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.013] [Medline: 25046131]
67. Martinez-Martin N, Luo Z, Kaushal A, Adeli E, Haque A, Kelly SS, et al. Ethical issues in using ambient intelligence in health-care settings. Lancet Digit Health 2021;3(2):e115-e123 [FREE Full text] [doi: 10.1016/S2589-7500(20)30275-2] [Medline: 33358138]
68. Johansson JV, Bentzen HB, Mascalzoni D. What ethical approaches are used by scientists when sharing health data? an interview study. BMC Med Ethics 2022;23(1):41 [FREE Full text] [doi: 10.1186/s12910-022-00779-8] [Medline: 35410285]
69. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. Mayo Clin Proc 2019;94(9):1681-1694 [FREE Full text] [doi: 10.1016/j.mayocp.2018.10.023] [Medline: 30803733]
70. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
71. Busse TS, Nitsche J, Kernebeck S, Jux C, Weitz J, Ehlers JP, et al. Approaches to improvement of Digital Health Literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022;19(14):8309 [FREE Full text] [doi: 10.3390/ijerph19148309] [Medline: 35886158]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
HCP: health care professional
HR: heart rate
HRV: heart rate variability
MBI-HSS: Maslach Burnout Inventory–Human Services Survey
STAI: State-Trait Anxiety Inventory
STC: step count
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 24.06.23; peer-reviewed by T Pipe, P Punda; comments to author 01.12.23; revised version received 01.01.24; accepted 20.03.24; published 25.06.24.

Please cite as:
Barac M, Scaletty S, Hassett LC, Stillwell A, Croarkin PE, Chauhan M, Chesak S, Bobo WV, Athreya AP, Dyrbye LN
Wearable Technologies for Detecting Burnout and Well-Being in Health Care Professionals: Scoping Review
J Med Internet Res 2024;26:e50253
URL: https://www.jmir.org/2024/1/e50253
doi:10.2196/50253
PMID:38916948

©Milica Barac, Samantha Scaletty, Leslie C Hassett, Ashley Stillwell, Paul E Croarkin, Mohit Chauhan, Sherry Chesak, William V Bobo, Arjun P Athreya, Liselotte N Dyrbye. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Web-Based Mindfulness-Based Interventions for Patients With Cancer: Systematic Review and Meta-Analyses

Ting Wang[1*], MNS; Chulei Tang[1*], PhD; Xiaoman Jiang[1], MSc; Yinning Guo[1], MNS; Shuqin Zhu[1], PhD; Qin Xu[1], MM

School of Nursing, Nanjing Medical University, Nanjing, China

[*]these authors contributed equally

Corresponding Author:
Qin Xu, MM
School of Nursing, Nanjing Medical University
101 Longmian Avenue, Jiangning District
Nanjing, 211166
China
Phone: 86 13601587208
Email: qinxu@njmu.edu.cn

Abstract

Background: Cancer has emerged as a considerable global health concern, contributing substantially to both morbidity and mortality. Recognizing the urgent need to enhance the overall well-being and quality of life (QOL) of cancer patients, a growing number of researchers have started using online mindfulness-based interventions (MBIs) in oncology. However, the effectiveness and optimal implementation methods of these interventions remain unknown.

Objective: This study evaluates the effectiveness of online MBIs, encompassing both app- and website-based MBIs, for patients with cancer and provides insights into the potential implementation and sustainability of these interventions in real-world settings.

Methods: Searches were conducted across 8 electronic databases, including the Cochrane Library, Web of Science, PubMed, Embase, SinoMed, CINAHL Complete, Scopus, and PsycINFO, until December 30, 2022. Randomized controlled trials involving cancer patients aged ≥18 years and using app- and website-based MBIs compared to standard care were included. Nonrandomized studies, interventions targeting health professionals or caregivers, and studies lacking sufficient data were excluded. Two independent authors screened articles, extracted data using standardized forms, and assessed the risk of bias in the studies using the Cochrane Bias Risk Assessment Tool. Meta-analyses were performed using Review Manager (version 5.4; The Cochrane Collaboration) and the *meta* package in R (R Foundation for Statistical Computing). Standardized mean differences (SMDs) were used to determine the effects of interventions. The Reach, Effectiveness, Adoption, Implementation, and Maintenance framework was used to assess the potential implementation and sustainability of these interventions in real-world settings.

Results: Among 4349 articles screened, 15 (0.34%) were included. The total population comprised 1613 participants, of which 870 (53.9%) were in the experimental conditions and 743 (46.1%) were in the control conditions. The results of the meta-analysis showed that compared with the control group, the QOL (SMD 0.37, 95% CI 0.18-0.57; $P<.001$), sleep (SMD −0.36, 95% CI −0.71 to −0.01; $P=.04$), anxiety (SMD −0.48, 95% CI −0.75 to −0.20; $P<.001$), depression (SMD −0.36, 95% CI −0.61 to −0.11; $P=.005$), distress (SMD −0.50, 95% CI −0.75 to −0.26; $P<.001$), and perceived stress (SMD −0.89, 95% CI −1.33 to −0.45; $P=.003$) of the app- and website-based MBIs group in patients with cancer was significantly alleviated after the intervention. However, no significant differences were found in the fear of cancer recurrence (SMD −0.30, 95% CI −1.04 to 0.44; $P=.39$) and posttraumatic growth (SMD 0.08, 95% CI −0.26 to 0.42; $P=.66$). Most interventions were multicomponent, website-based health self-management programs, widely used by international and multilingual patients with cancer.

Conclusions: App- and website-based MBIs show promise for improving mental health and QOL outcomes in patients with cancer, and further research is needed to optimize and customize these interventions for individual physical and mental symptoms.

Trial Registration: PROSPERO CRD42022382219; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=382219

(J Med Internet Res 2024;26:e47704) doi:10.2196/47704

XSL•FO
RenderX

KEYWORDS

cancer; mindfulness-based interventions; mental health; randomized controlled trial; systematic review; meta-analysis; mindfulness; web-based intervention; oncology; delivery mode; efficacy; quality of life; program; adherence; mobile phone

Introduction

Background

The 2020 Global Cancer Statistics Report estimates that there are 19.3 million new cases of cancer worldwide and approximately 10 million cancer-related deaths [1]. The leading cause of disease and mortality among humans today is cancer [2,3]. The physical symptoms of patients with cancer have been alleviated because of the continuous advancement of medical technology, but the psychological problems of patients with cancer have not been adequately treated. The process of treating cancer is typically complex, with many patients experiencing negative side effects of cancer treatments, such as chemotherapy and radiation therapy, that may impact their mental health, quality of life (QOL), and sleep quality. Targeted interventions to address these cancer-related symptoms can reduce the psychological burden of cancer treatment and diagnosis, which is critical to improving patients' QOL and promoting their health [4]. With an increasing number of patients with cancer and a desire for physical and mental health, cancer care research is focusing on identifying the psychological problems of patients with cancer and developing and implementing patient-centered psychological care plans [5,6]. Rehabilitation for patients with cancer increasingly uses mental health as a therapeutic strategy; however, effective psychological intervention strategies are still urgently needed to satisfy the demands of patients with cancer [7].

Mindfulness-based interventions (MBIs) have emerged as promising intervention techniques for patients with cancer. Mindfulness can be defined as the ability to observe thoughts, bodily sensations, or feelings in the present moment with an open and accepting orientation toward one's experiences [8]. MBIs, which incorporate mindfulness practices into various therapies in mental health care, have been found to increase psychological flexibility and alleviate intense emotional states. MBIs can include additional mental training, such as mindfulness-based stress reduction (MBSR) [9], and acceptance and commitment therapy [10], which addresses psychological issues by increasing psychological flexibility [11]. Cognitive-behavioral therapy has been combined with MBSR, resulting in mindfulness-based cognitive therapy (MBCT) for preventing depression relapses [12]. Mindfulness-based cancer recovery (MBCR), an adaptation of MBSR, comprises contents tailored for patients with cancer [13]. Through facilitating awareness and nonjudgmental acceptance of moment-to-moment experiences, these MBIs are presumed to alleviate intense emotional states. Mindfulness interventions have been shown to improve the psychological status of patients with cancer [14,15].

The rapid development of information technologies has led to the delivery of MBIs via the internet, which is more practical than face-to-face interaction and can overcome time and geographic barriers, and it has been established that online MBIs are more suitable for people with psychological and physical symptoms [16]. Implementing psychological interventions through online or remote health can be a potential cost benefit for current referral pathways and treatment models [17]. online MBIs can be used as the adjunctive therapy in patients with cancer to manage cancer-related symptoms [18].

Despite the increasing popularity of online mindfulness-based therapies for patients with cancer and the growing number of randomized controlled trials (RCTs) examining such programs, there has not been a systematic review of these studies and their descriptions of the interventions regarding their characteristics (eg, delivery mode and approach). To date, only 2 systematic reviews addressing the impact of online interventions on health outcomes in patients with cancer have been published. However, these reviews have notable limitations. The first review [19] only searched 4 databases, potentially leading to publication bias and compromising the reliability of the findings. Furthermore, this systematic review did not conduct sensitivity, subgroup, or meta-analyses. The second review [20] evaluated the validity of online MBIs on only 4 health outcomes: anxiety, depression, QOL, and mindfulness. However, the restricted quantity of RCTs and papers within each subgroup analysis poses a challenge in reaching definitive conclusions. In addition, the external validity (eg, generalizability or applicability) based on the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework has not been examined in online MBIs for patients with cancer. Thus, attempts to synthesize the literature on the impact of online MBIs on the health of patients with cancer are limited, and there is a lack of analysis of the barriers and facilitators to the development of current online MBIs.

Objectives

This systematic review aims to synthesize the effectiveness of online MBIs, encompassing both app- and website-based MBIs, for patients with cancer, comprehensively assessing a wide range of outcomes, including psychological, physiological, and QOL aspects. We conducted a comprehensive search to evaluate the validity of app- and website-based MBIs on psychological outcomes in patients with cancer, using high-quality RCTs to assess many health outcomes before and after treatment. Moreover, this study aims to provide an overview of the outcomes related to the interventions, including their effectiveness and potential for implementation and sustainability in real-world settings. We used the RE-AIM framework [21] to evaluate the potential for implementation and sustainability of these interventions in real-world settings. Using this framework, we can provide a comprehensive evaluation of an intervention's potential impact and identify common traits of effective interventions. Overall, this study fills gaps in the literature by comprehensively evaluating the effectiveness and potential for implementation and sustainability of app- and website-based MBIs for patients with cancer.

XSL•FO
RenderX

Methods

Search Strategy

The protocol of this review was registered in PROSPERO (CRD42022382219) and written following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guideline. The methods outlined in the protocol were strictly adhered to throughout the experimental procedures. The databases were searched until December 30, 2022. To identify relevant studies for inclusion in our systematic reviews, we developed comprehensive search strategies and used 8 databases: Cochrane Library, Web of Science, PubMed, Embase, SinoMed, CINAHL Complete, Scopus, and PsycINFO. The literature search language was limited to Chinese and English. The search strategies used a combination of subject headings (eg, Medical Subject Headings in PubMed) and keywords for the following 5 concepts: mindfulness, carcinoma, intervention, telemedicine, and randomly. Multimedia Appendix 1 shows detailed database search strategies. Reference lists of included studies and relevant systematic reviews were also manually searched for additional relevant studies. Search results were captured using citation management software, and duplicates were removed.

Inclusion and Exclusion Criteria

Because of the explorative nature of this meta-analysis, we opted for rather broad inclusion criteria. The inclusion criteria were as follows: (1) studies that included patients with cancer (aged ≥18 years) with any cancer type and stage, including those receiving anticancer treatment, those in remission, those considered cured, and those in the terminal phases of the disease; (2) studies that used MBIs (including MBSR, MBCT, and MBCR) and administered the MBI via the internet (including websites, web conferencing, web-based games, and web-based video) or a smartphone app; (3) studies in which eligible controls were required to receive standard care or usual care; (4) studies were eligible if a mental health outcome (eg, fear of cancer recurrence [FCR] as measured with the Fear of Cancer Recurrence Inventory [FCRI] and posttraumatic growth [PTG] as measured with the Posttraumatic Growth Inventory), anxiety, depression, distress, stress, and sleep) or QOL was assessed; and (5) the RCT was published in English or Chinese.

Exclusion criteria were (1) other types of studies (eg, observational, review, protocol, and case report); (2) studies of health professionals, caregivers, or mixed populations in which outcomes for survivors of cancer could not be extracted; and (3) insufficient information to calculate an effect size or determine eligibility.

Screening and Data Extraction

Two reviewers independently screened all titles and abstracts; then, they independently screened full-text articles, and conflicts were resolved by consensus. Data were independently extracted by 2 reviewers using a data extraction form adapted from the Cochrane Handbook [22] and reported using PRISMA guidelines [23]. We extracted data from included trials using standardized data extraction forms. Study-level variables included the year of publication, country of study, age of participants, cancer diagnosis, delivery mode, reminder, cancer-adapted MBIs, primary and secondary outcomes, intervention and follow-up durations, intervention and control group details, outcome measurement metrics, and outcomes scores up to postintervention. Any discrepancy or uncertainties were resolved through regular meetings and discussion among the research team.

Risk-of-Bias Assessment

The risk of bias was independently assessed by 2 reviewers using the Cochrane Risk-of-Bias tool, with differences reconciled through discussion [24]. A total of 6 domains encompassed random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, and selective outcome reporting. Each domain was judged as low, high, or unclear risk. Discrepancies in assessments between the 2 reviewers were resolved by consensus or by a third reviewer as required.

Meta-Analytic Method

This study conducted a meta-analysis using Review Manager (version 5.4; The Cochrane Collaboration) and the *meta* package in R (R Foundation for Statistical Computing). The primary and secondary outcome mean and SD values at postintervention follow-up for intervention and control groups were converted to standardized mean difference (SMD), using Hedges *g*. The value of SMD <0.5 would be interpreted as small, SMD ≥0.5 as medium, and SMD ≥0.8 as large effect size [25]. Authors of studies with missing data were contacted through email. However, if no data were provided, a narrative synthesis would be conducted. The I^2 statistic was used to estimate the percentage of heterogeneity across the primary studies not attributable to random sample error alone. A value of 0% indicated no heterogeneity, and values of 25%, 50%, and 75% reflected low, moderate, and high degrees of heterogeneity, respectively [26]. Acknowledging differences across studies because of the varied population, length of intervention, and length of follow-up, meta-analyses were performed fitting random effects models [27]. In addition, subgroup analyses were conducted to examine effect sizes across different subgroups; the specific moderating variables included technology, sex, intervention type, intervention duration, study quality, and scale.

RE-AIM Framework

The RE-AIM framework is a valuable tool for evaluating interventions in health care [28]. Its 5 dimensions assess an intervention's potential for large-scale adoption, implementation, and sustainability, providing a comprehensive evaluation of its real-world efficacy and viability [29]. Reach refers to the extent of successfully targeting and engaging the intended audience, evaluated using the percentage of eligible patients enrolled in the study (n enrolled/n eligible). Efficacy measures the effect on outcomes such as mental health and QOL. Effect sizes (95% CIs) for the primary outcome were used to assess efficacy. Adoption measures the extent to which organizations or health care providers are willing and able to offer the intervention to their patients or clients, and barriers to adoption are evaluated by who recruited participants and where the intervention was

XSL•FO
RenderX

offered. Implementation evaluates how effectively the intervention is delivered and received by patients, including factors such as adherence and fidelity, and is evaluated by measures such as adherence to the intervention, percentage of dropouts of the most complex intervention (n postintervention follow-up/n baseline×100), intervention cost, and author-reported plans to upscale or implement. Maintenance measures the extent to which the intervention can be sustained over time and integrated into routine care, and it is evaluated by the duration of results and the author-reported availability of the intervention [30].

Results

Description of Studies

The systematic search revealed 4349 original articles, of which 54 (0.12%) were assessed at full-text level, and 15 (0.03%) studies were included in the final synthesis. Figure 1 displays the study flowchart of the search results and Table 1 presents the characteristics extracted from the included literature in the study. The total population comprised 1613 participants, of which 870 (53.94%) and 743 (46.06%) were in the experimental and control conditions, respectively. In most (13/15, 87%) studies, the majority of participants were women. Participants were aged from 41.84 to 66.45 years. Four studies were based on MBCR, 3 on MBCT, 2 on MBSR, and 6 on mindfulness-based programs. The 6 studies included interventions that were indeed rooted in mindfulness practices; however, they did not strictly adhere to the conventional frameworks of MBCT, MBCR, or MBSR. Instead, they used a variety of mindfulness-based approaches tailored to their respective study populations.

Furthermore, these studies did not specify the exact intervention methods used but were categorized as *mindfulness-based programs*. Because of the unique nature of these interventions, we cannot determine whether they belong to MBCT, MBCR, or MBSR interventions; we have categorized them as *mindfulness-based programs*, encompassing diverse methodologies beyond the traditional MBCT, MBCR, or MBSR frameworks. Trials used usual care (8 trials) and waitlists (7 trials) equally as comparators. Six studies had participants with breast cancer, 7 with mixed cancer types, and 2 with other cancer types. Five studies were conducted in China; 5 in the United States; and 1 each in the Netherlands, Denmark, Iran, Australia, and Canada.

XSL•FO
RenderX

Figure 1. Flow diagram of trial identification and selection. RCT: randomized controlled trial.

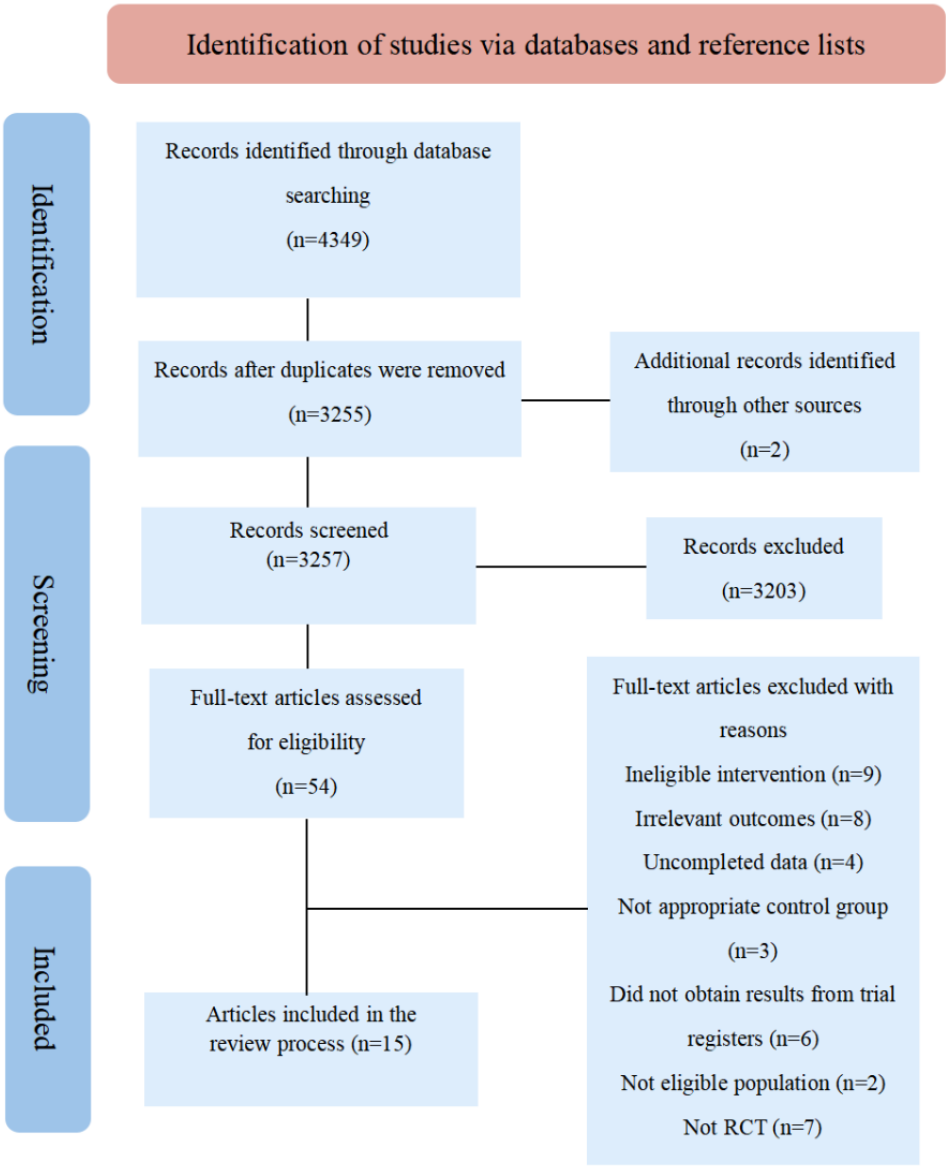

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study; country	Cancer type; age (y), mean (SD); gender (female; %)	Intervention (n); delivery mode	Reminders	Intervention duration, number of sessions; intervention dose	Cancer adapted	Technology	Control group (n)	Measurements	Outcomes: measure instrument
Chang et al [31], 2022; China	Breast; 49.6 (12.0); 100	MBSR[a] (41); web-based software and digital interactive whiteboard	—[b]	6 wk, 6 sessions; 2 h/wk	—	Website	Waitlist (26)	Pre and post	Depression, anxiety, and stress: DASS-21[c]
Compen et al [32], 2018; Netherlands	Mixed; 51.7 (10.7); 85	MBCT[d] (90); email, meditation audio file, and written feedback	—	8 wk; 40-45 min/d twice daily	—	Website	Usual care (78)	Pre and post	Distress: HADS[e], FCR[f]: FCRI[g], and QOL[h]: SF-12[i]
Kubo et al [33], 2019; US	Mixed; 58.2 (14.4); 68	MBP[j] (54); audio instruction, lecture videos, and foundation course	Study staff made phone calls if an intervention participant completed <3	8 wk, 2 h/wk	Specifically for individuals affected by cancer	App	Usual care (43)	Pre and post	Distress: NCCN[k], anxiety and depression: HADS, PTG[l]: PTGI[m], sleep: PROMIS[n], and QOL:FACT-G[o]
Kubo et al [34], 2018; US	Mixed; 66.5 (9.7); 69	MBP (52); online classroom and manual	The app can send reminders using push notifications; study staff made phone calls if an intervention participant completed <3	6 wk; 2 h/wk	Cancer pack, which was designed specifically for individuals affected by cancer	App	Usual care (51)	Pre and post	QOL: FACIT−Pal[p], distress: NCCN, and anxiety and depression: HADS
Liu et al [35], 2022; China	HCC[q]; 55.7 (8); 22	MBCT (61); WeChat audio and online platforms	Every day (texting)	6 wk, 20 min/d for 5 d/wk	Adoption of the main issues and needs of patients with HCC	App	Waitlist (61)	Pre, post, 1 mo FU[r], 3 mo FU	Distress: HADS, sleep: PSQI[s], QOL: FACT-hep[t], and stress: PSS[u]
Messer et al [36], 2020; US	Mixed; 51 (10.6); 76	MBSR (11); guided meditation audio clips and brief textual lessons	—	6 wk; mean duration of 12 min/session	—	Website	Usual care (10)	Pre and post	Distress: HADS, QOL: POMS-SF[v], and sleep: PSQI
Nissen et al [37], 2018; Denmark	Breast and prostate; 55.9 (12.1); 91	MBCT (104); website written material, audio exercises, writing tasks, and videos	—	10 wk, 10 sessions; 2 h/wk for 45 min/d	Program adjustments to meet the needs of survivors of cancer	Website	Waitlist (46)	Pre, post, and 6 mo FU	Anxiety: STAI - Y[w], depression: BDI - II[x], stress: PSS - 10, and sleep: ISI[y]

XSL•FO
RenderX

Study; country	Cancer type; age (y), mean (SD); gender (female; %)	Intervention (n); delivery mode	Reminders	Intervention duration, number of sessions; intervention dose	Cancer adapted	Technology	Control group (n)	Measurements	Outcomes: measure instrument
Peng et al [38], 2022; China	Breast; 41.8 (2.9); 100	MBP (30); website meeting 5P[z] medicine approach	—	6 wk, 6 sessions; 1.5 h/wk	On the basis of specific considerations for survivors of cancer	App	Usual care (30)	Pre, post, and 1 mo FU	FCR: FCRI-SF[aa] and QOL: Eortc-Qlq-C30[ab]
Rosen [39], 2017; US	Breast; 53 (10.3); 100	MBP (48); app-based courses include audio and video	General weekly check-in emails	9 wk	—	App	Waitlist (47)	Pre, wk 5, wk 9, and wk 4 FU	QOL: FACT-B[ac]
Rosen et al [40], 2018; US	Breast; 51.6 (10.3); 100	MBP (57); app-based audio and animated video	Weekly check-in email	9 wk	—	App	Waitlist (55)	Pre, post, and 1 mo FU	QOL: FACT - B
Russell et al [41], 2019; Australia	Melanoma; 53.4 (13.1); 54	MBP (46); embedded short videos, PDF transcript of the videos, and MP3 audio email	Automatically generated email reminders twice daily	6 wk	Survey to understand the knowledge of meditation among people with melanoma	Website	Waitlist (23)	Pre and post	FCR: FCRI and stress: PSS-10
Shen et al [42], 2021; China	Breast; 47.4 (7.5); 100	MBCR[ad] (37); online course, WeChat group, audio-video materials, and pictures	Every day (texting)	8 wk, 8 sessions; 15 min/d for 6 d/wk	Combine rich experience in rehabilitation psychotherapy of breast cancer	App	Usual care (40)	Pre and post	Stress: CPSS[ae] and anxiety: SAS[af]
Wang [43], 2022; China	Breast; 46.8 (7.9); 100	MBCR (51); web-based courses and intervention materials	—	4 wk, 4 sessions; 1.5 h/wk and 30 min daily	On the basis of the problems in the pilot study and participant feedback, adjusted internet-delivered MBCR program	Website	Usual care (52)	Pre and post	QOL: FACT-B
Yousefi et al [44], 2022; Iran	Colorectal and stomach; 54.9 (6.6); 42	MBCR (25); web-based session	An alert reminder message was sent 2 h before each session	9 wk, 9 sessions; 90 min/wk	Cancer-specific MBSR program was used in the study	Website	Usual care (25)	Pre, post, and 2 mo FU	Stress: DASS - 21 and sleep: ISI

XSL•FO
RenderX

Study; country	Cancer type; age (y), mean (SD); gender (female; %)	Intervention (n); delivery mode	Reminders	Intervention duration, number of sessions; intervention dose	Cancer adapted	Technology	Control group (n)	Measurements	Outcomes: measure instrument
Zernicke et al [45], 2014; Canada	Mixed; 58 (10.7); 72	MBCR (30); web-based classroom, guided meditation recordings, and videos	—	8 wk, 8 sessions; 45 min/d	Cancer-adapted MBSR	Website	Waitlist (32)	Pre and post	Depression and anxiety: POMS[ag], stress: CSOSI[ah], PTG: PTGI

[a]MBSR: mindfulness-based stress reduction.
[b]Not applicable.
[c]DASS-21: Depression, Anxiety, and Stress Scale-21.
[d]MBCT: mindfulness-based cognitive therapy.
[e]HADS: Hospital Anxiety and Depression Scale.
[f]FCR: fear of cancer recurrence.
[g]FCRI: Fear of Cancer Recurrence Inventory.
[h]QOL: quality of life.
[i]SF-12: 12-item Short-Form health survey.
[j]MBP: mindfulness-based program.
[k]NCCN: National Comprehensive Cancer Network Distress Thermometer.
[l]PTG: posttraumatic growth.
[m]PTGI: 21-item Posttraumatic Growth Inventory.
[n]PROMIS: 8-item PROMIS Sleep Disturbance scale.
[o]FACT-G: 27-item Functional Assessment of Cancer Therapy General Scale.
[p]FACIT - Pal: 46-item Functional Assessment of Chronic Illness Therapy—Palliative Care.
[q]HCC: hepatocellular carcinoma.
[r]FU: follow-up.
[s]PSQI: Pittsburgh Sleep Quality Index.
[t]FACT-Hep: Functional Assessment of Cancer Therapy-Hepatobiliary Carcinoma.
[u]PSS: Perceived Stress Scale.
[v]POMS-SF: Profile of Mood States-Short Form.
[w]STAI - Y: State-Trait Anxiety Inventory Y-Form.
[x]BDI - II: Beck Depression Inventory.
[y]ISI: Insomnia Severity Index.
[z]5P: The specific name of an application designed to promote mind and brain health and cultivate happiness.
[aa]FCRI-SF: Fear of Cancer Recurrence Inventory-Short Form.
[ab]Eortc-Qlq-C30: European Organization for Research and Treatment of Cancer questionnaire.
[ac]FACT-B: Functional Assessment of Cancer Therapy-Breast version 4.
[ad]MBCR: Mindfulness-based cancer recovery.
[ae]CPSS: Chinese version of the Perceived Stress Scale.
[af]SAS: Self-Rating Anxiety Scale.
[ag]POMS: Profile of Mood States.
[ah]CSOSI: Calgary Symptoms of Stress Inventory.

Risk of Bias

The risk-of-bias assessment is presented in Multimedia Appendix 2 [31-45]. Most studies (9/15, 60%) adequately generated and concealed allocation (Figure 2). In most studies (14/15, 93%), patient blinding was not possible because of the nature of online MBIs and was not considered to increase the risk of bias. However, of the 15 studies, 8 (53%) [31,33,37-40,42,44] presented insufficient information regarding researcher and outcome assessor blinding, whereas 7 (47%) reported blinding researchers [32,34-36,41,43,45] (low risk). A total of 14 studies reported complete outcome data (low risk), and 1 study had insufficient detail [44] (unclear risk). In 1 study [40], attrition was high and comparisons or reasons for attrition were not provided. Finally, 66% (10/15) of the studies did not reference a protocol or trial registration (unclear risk).

XSL•FO
RenderX

Figure 2. Cochrane risk-of-bias scores (percentage of low, unclear, and high risk) across bias domains (selective reporting, incomplete outcome data, blinding, allocation concealment, and random sequence generation) for the 15 included web-based mindfulness-based intervention studies for patients with cancer.

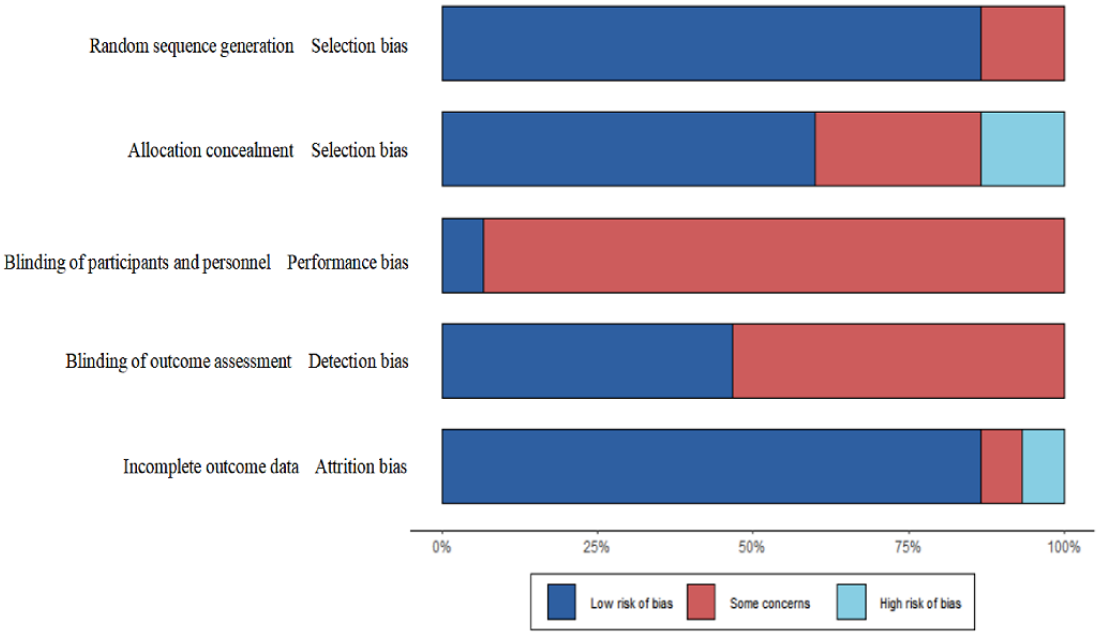

Meta-Analysis

Effects on QOL

A total of 8 studies reported the effects of app- and website-based MBIs on QOL among patients with cancer. To measure QOL in patients with cancer, 4 health-related QOL measures were used, including the Functional Assessment of Chronic Illness Therapy [34], the Functional Assessment of Cancer Therapy [33,35,39,40,43], the Short-Form 12 [32], and the European Organization for Research and Treatment of Cancer questionnaire [38], all of which have been validated in this patient population. Higher scores reflected a higher QOL. Because the physical and psychological components of the scale were measured separately and it was not possible to determine the overall change in the QOL, the data from 1 study [32] were not summarized. A total of 7 studies including 569 participants were evaluated in the meta-analysis. No significant heterogeneity was found between studies (I^2=26%; P=.23; Figure 3 [33-41,43-45]). The intervention group had a significant QOL improvement compared to the control group (SMD 0.37, 95% CI 0.18-0.57; P<.001). In addition, the exclusion of any single study at one time did not change the pooled results markedly.

XSL•FO
RenderX

Figure 3. Meta-analysis on (A) quality of life, (B) sleep, (C) fear of cancer recurrence, and (D) posttraumatic growth. SMD: standardized mean difference.

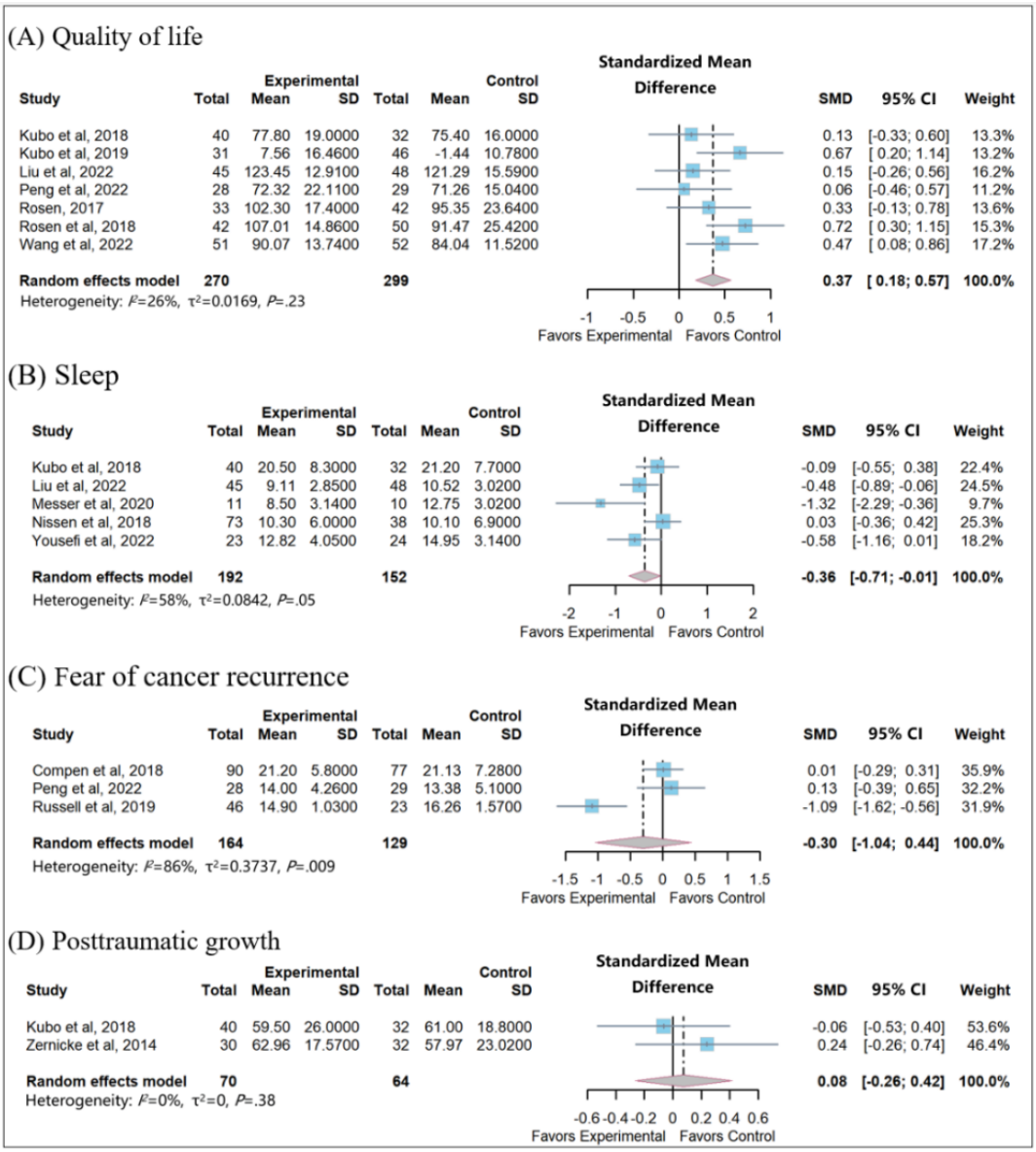

Effects on Sleep

Five studies investigated the impact of app- and website-based MBIs on sleep quality using 3 assessment tools: the 8-item PROMIS Sleep Disturbance scale [33], the Insomnia Severity Index [37,44], and the Pittsburgh Sleep Quality Index [35,36]. A higher score indicated a worse sleep quality. Moderate heterogeneity of effect sizes was observed (I^2=58%; P=.05; Figure 3). Grouping the studies by type of technology, scale, and intervention type did not resolve heterogeneity, so a random effects model was chosen to pool the results. The result revealed that app- and website-based MBIs could alleviate patients' sleep issues, with a statistical difference (SMD −0.36, 95% CI −0.72 to −0.01; P=.04). Only 1 outlier was detected [36]. After omitting the studies from the analysis, the effect size dropped to an SMD of −0.25 (95% CI −0.54 to 0.04; P=.09), and heterogeneity reduced substantially (I^2=38%). The possible reason for this change may be attributed to the fact that small sample sizes tend to yield more pronounced effects.

Effects on FCR

A total of 3 studies measured FCR; the pooled data included 224 participants. Two FCR measures were used: FCRI [32,41] and the Short-Form FCRI [38]. A higher score indicated a higher level of FCR. There is great heterogeneity among the studies (I^2=86%; P=.009; Figure 3). After the data of the study by Russell et al [41] are eliminated by the method of eliminating one by one, there is significantly lower heterogeneity (I^2=0%; P=.70). This may be due to Russell et al [41] presurveying patients with cancer so that the intervention on FCR was more effective. The results showed that the difference between the network-based MBIs and the control group was not statistically significant (SMD −0.30, 95% CI −1.04 to 0.44; P=.39).

Effects on PTG

Two studies examined the effect of app- and website-based MBIs on PTG, with a total of 134 participants. The measurement tool exclusively used across 2 studies to assess PTG was the Posttraumatic Growth Inventory [33,45]. Higher scores indicated

XSL•FO
RenderX

greater PTG. No significant heterogeneity was found between studies (I^2=0%; P=.38; Figure 3). We found that app- and website-based MBIs did not lead to a significant increase in PTG score (SMD 0.08, 95% CI −0.26 to 0.42; P=.66).

Effects on Anxiety

Anxiety levels were assessed in 6 studies using 5 validated scales. These scales include the Hospital Anxiety and Depression Scale (HADS) [33,34], the Depression Anxiety Stress Scale Depression Inventory [31], the State-Trait Anxiety Inventory Y-Form [37], the Self-Rating Anxiety Scale [42], and the Profile of Mood States [45]. Higher scores on these scales indicated elevated levels of anxiety. Meta-analysis showed that app- and website-based MBIs lead to a significant reduction in anxiety (SMD −0.48, 95% CI −0.75 to −0.20; P<.001; Figure 4 [31-37,41,42,44,45]). Moderate heterogeneity was found between studies (I^2=52%; P=.07). Grouping the studies by type of technology and intervention duration did not resolve heterogeneity (Table 2). Furthermore, when we examined subgroups based on sex, we found that studies including female participants had a significantly larger pooled effect size (SMD −0.67, 95% CI −1.01 to −0.33; P<.001) than the studies including both male and female participants (referred to as the mixed-gender subgroup; SMD −0.39, 95% CI −0.76 to −0.02; P=.04; Figure 4). The differences across these 2 subgroups were statistically nonsignificant (χ^2_1=1.2; P=.28).

Figure 4. Meta-analysis on (A) anxiety, (B) depression, (C) perceived stress, and (D) distress. SMD: standardized mean difference.

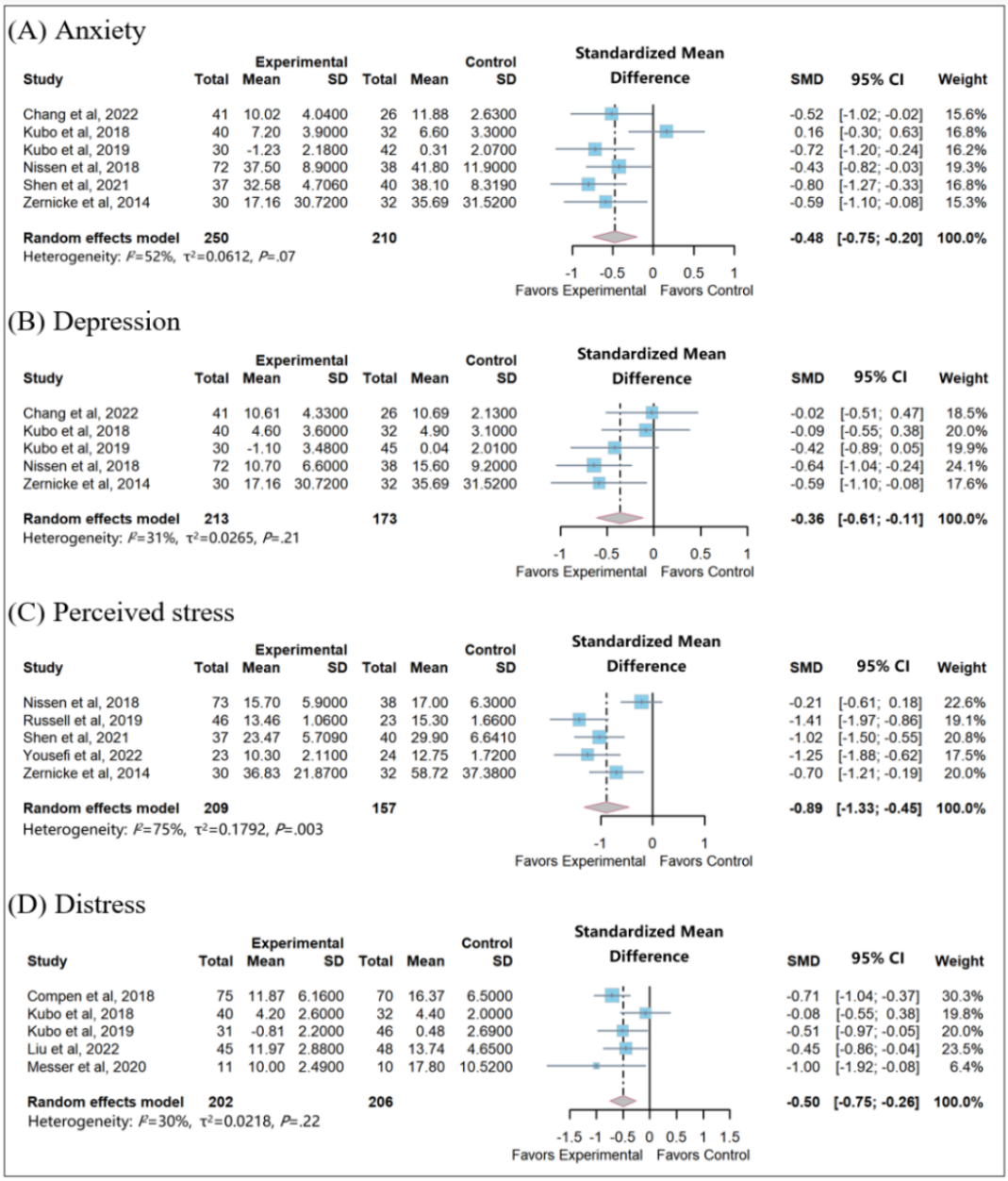

XSL•FO
RenderX

Table 2. Subgroup analyses of anxiety, perceived stress, and sleep.

Subgroup and stratification	Studies, n (%)	SMD[a] (95% CI)	*P* value for heterogeneity	I^2	*P* value for pooled results	*P* value for interaction
Anxiety (n=6)						
Technology						.54
Website	3 (50)	−0.57 (−0.82 to −0.31)	.47	0	.0001	
App	3 (50)	−0.38 (−0.93 to 0.18)	.02	74	.18	
Intervention duration						.43
<8 wk	2 (33)	−0.62 (−0.97 to−0.27)	.57	0	<.001	
≥8 wk	4 (67)	−0.41 (−0.81 to −0.01)	.03	67	.04	
Sex						.28
Female	2 (33)	−0.67 (−1.01 to −0.33)	.41	0	<.001	
Mixed	4 (67)	−0.39 (−0.76 to −0.02)	.05	62	.04	
Intervention type						.67
MBCR[b]	2 (33)	−0.70 (−0.05 to −0.36)	.55	0	<.001	
MBCT[c]	1 (17)	−0.43 (−0.82 to 0.03)	__[d]	—	.04	
MBIs[e]	2 (33)	−0.28 (−1.14 to 0.59)	.01	85	.53	
MBSR[f]	1 (17)	−0.52(−1.02 to −0.02)	—	—	.04	
Study quality						.07
Unclear risk[g]	5 (83)	−0.43 (−0.74 to −0.12)	.06	56	.007	
High risk[h]	1 (17)	−0.72 (−1.20 to −0.24)	—	—	.004	
Perceived stress (n=5)						
Technology						.68
Website	4 (80)	−0.87 (−1.44 to −0.29)	.002	80	.003	
App	1 (20)	−1.02 (−1.50 to −0.55)	—	—	<.001	
Intervention type						<.001
MBCR	3 (60)	−0.96 (−1.27 to−0.66)	.39	0	<.001	
MBCT	1 (20)	−0.21 (-0.61 to 0.18)	—	—	.29	
MBIs	1 (20)	−1.41 (−1.97 to −0.86)	—	—	<.001	
Sleep (n=5)						
Scale						.33
PROMs[i]	1 (20)	−0.09 (−0.55 to 0.38)	—	—	.72	
PSQI[j]	2 (40)	−0.78 (−1.58 to 0.02)	.11	60	.05	
ISI[k]	2 (40)	−0.23 (−0.83 to 0.36)	.09	65	.44	
Technology						.20
Website	3 (60)	−0.52 (−1.22 to 0.19)	.02	75	.15	
App	2 (40)	−0.02 (−0.32 to 0.28)	.70	0	.91	
Intervention duration						.16
<8 wk	2 (40)	−0.78 (−1.58 to −0.02)	.11	60	.05	
≥8 wk	3 (60)	−0.16 (−0.49 to 0.17)	.23	32	.35	
Study quality						.11
Unclear risk	4 (80)	−2.03 (−2.93 to −1.13)	.12	49	<.001	
High risk	1 (20)	0.20 (−2.93 to 2.79)	—	—	.88	

XSL•FO
RenderX

[a]SMD: standardized mean difference.
[b]MBCR: mindfulness-based cancer recovery.
[c]MBCT: mindfulness-based cognitive therapy.
[d]Not applicable.
[e]MBI: mindfulness-based intervention.
[f]MBSR: mindfulness-based stress reduction.
[g]Unclear risk: unclear risk of bias for one or more key domains.
[h]High risk: high risk of bias for one or more key domains.
[i]PROM: patient - reported outcome measure.
[j]PSQI: Pittsburgh Sleep Quality Index.
[k]ISI: Insomnia Severity Index.

Effects on Depression

Depression was assessed across 5 studies using various standardized instruments. These included the Depression Anxiety Stress Scale-21 [31], HADS [33,34], Beck Depression Inventory [37], and Profile of Mood States [45]. Elevated levels of depression were indicated by higher scores on these scales. The pooled data included 384 participants and showed a significant difference in improvement between the intervention and control groups (SMD −0.36, 95% CI −0.61 to −0.11; *P*=.005; Figure 4). Moderate heterogeneity of effect sizes was observed (I^2=31%; *P*=.21). In the sensitivity analysis using the one-study-out method, we found that the pooled estimates were not significantly altered when any 1 study was omitted in turn. The range of *P* values obtained varied from .0001 to .03, indicating that the summary effect size is robust.

Effects on Perceived Stress

A total of studies investigated the effects of app- and website-based MBIs on stress. Four distress measures were used: the Perceived Stress Scale [35,37,41], the Chinese version of the Perceived Stress Scale [42], the Depression and Stress Scale [44], and the Calgary Symptoms of Stress Inventory [45]. A total of 5 studies including 366 participants were evaluated in the meta-analysis. The data from 1 study was not pooled because the mean values and SD of outcomes were not reported [35], and between-study heterogeneity was found (I^2=75%; *P*=.003; Figure 4). This meta-analysis revealed a reduction in stress of −0.89 (95% CI −1.33 to −0.45) when comparing the intervention group to the control group at the postintervention stage.

To further explore the potential sources of heterogeneity, we conducted subgroup analyses by type of technology and intervention type (Table 2). The 2 studies using apps (SMD −1.02, 95% CI −1.50 to −0.55; I^2=0) were found to have low heterogeneity, whereas the 3 studies based on website-based technologies (SMD −0.87, 95% CI −1.44 to −0.29; *P*=.002) exhibited higher heterogeneity. After conducting sensitivity analysis and eliminating 1 study at a time, the exclusion of the study by Nissen et al [37] resulted in significantly lower heterogeneity (I^2=21%; *P*=.28). One possible reason is that the study by Nissen et al [37], which offered internet-delivered MBCT as a routine based on a screening procedure, may have included less motivated participants compared to studies with self-referral. In addition, Nissen et al [37] used a lower cutoff value for screening the study population, which could have resulted in a floor effect.

Effects on Distress

In the analysis of distress (involving 5 studies), HADS [32,35,36] and the National Comprehensive Cancer Network Distress Thermometer [33,34] were used to assess the current distress level. Low heterogeneity was found between studies (I^2=30%; *P*=.22; Figure 4), and the random effects model indicated that app- and website-based MBIs were associated with reduced distress levels in patients with cancer (SMD −0.50, 95% CI −0.75 to −0.26; *P*<.001).

Subgroup Analysis

Table 2 displays the results of subgroup analyses that were conducted to investigate the heterogeneity in the association between anxiety, perceived stress, and sleep in the context of MBIs. To explain the variability in the effects of mindfulness, we examined various moderating variables, such as technology, sex, intervention type, intervention duration, study quality, and scale. No statistically significant variables were found in the subgroup analysis of anxiety and sleep, whereas the type of intervention (*P*<.001) was a significant moderating variable for perceived stress

Publication Bias

Funnel plots and statistical tests were not performed as any of the outcomes had at least 10 studies to ensure sufficient power in detecting asymmetry [46]. However, we reduced the possibility of publication bias by conducting a thorough search across multiple databases to identify published studies [47].

RE-AIM Framework

Details of the RE-AIM Framework assessment are presented in Multimedia Appendix 3 [31-45]. Of the 15 studies, 14 (93%) reported 13% to 92% of eligible patients. Efficacy (effect size and 95% CI of primary outcome) was reported in 33% (5/15) of the studies [33,35,37,39,40] (Cohen *d* or η^2). For adoption barriers, health professionals or researchers conducted recruitment for all studies, and 53% (8/15) of the studies [35-38,41-44] recruited participants in person (hospital and cancer center). For implementation, intervention adherence ranged from 59% to 100% of participants completing all scheduled components. Dropouts of most complex interventions ranged from 0% to 48%, with 40% (6/15) of the studies [31,38-40,42,45] having <10% dropouts. The cost was reported in 4 studies [33,34,39,40], including the paid app (priced at US

XSL•FO
RenderX

$77 for 6 months and US $69.99 for 12 months) and the app already publicly available. In total, 46% (7/15) of the studies [35,37-40,43,44] reported maintenance of results, and 46% (7/15) of the studies [35,37-40,43,44] sustained results for 1 to 9 months. Four studies [33,34,39,40] explicitly reported on the potential for the interventions to remain accessible or whether there were plans for their continued implementation.

Discussion

Principal Findings

The objective of this study is to assess the effectiveness of MBIs in improving the mental health and QOL of patients with cancer. We discovered that patients' QOL can be greatly enhanced by app- and website-based MBIs, which also significantly lowers psychological distress, sleep problems, anxiety, depression, and perceived stress. This systematic review of meta-analyses and the RE-AIM framework demonstrate that app- and website-based interventions have a wide range of effects and are highly used by different (international and multilingual) patients with cancer. However, the use and accessibility of app- and website-based MBIs for patients with cancer have been constrained because of service fees and patient mobility limitations [48]; app- and website-based MBIs are mainly conducted in high-income countries. The possible explanation is the distinction between communication and economy; some high-income countries may have national health services in place to promote app- and website-based MBIs, whereas developing nations may not. Study shows that in many low- and middle-income countries, the accessibility of evidence-based mental health treatments remains limited [49]. The time commitment, teacher shortage, and high cost of classic mindfulness interventions may have hindered efforts to spread the associated benefits to individuals in developing countries [50]. For instance, Indonesia has yet to implement evidence-based internet-based mindfulness therapy, emphasizing the need for expanding evidence-based mental health interventions in resource-constrained settings.

The results of this study suggest that app- and website-based MBIs are effective in improving QOL and reducing anxiety and depressive symptoms in patients with cancer, which is consistent with previous meta-analyses [18,20]. A possible explanation for this is that app- and website-based MBIs can alleviate negative emotions, enhance positive emotions, and increase mindfulness skills among patients with cancer, as elaborated by previous research [51]. Moreover, the sleep quality of patients with cancer also improved after MBIs. This outcome may be attributed to the inclusion of techniques in the program that target sleep difficulties [7] and the nonjudging aspect of mindfulness, which can enhance sleep quality by mitigating stress and everyday tensions. Previous studies [52] have confirmed the moderate effect of mindfulness interventions on sleep quality, which suggests that the use of app- and website-based MBIs to manage QOL and sleep in patients with cancer should be further supported.

App- and website-based MBIs have shown potential in helping patients with cancer develop emotional regulation skills and cope with the distress associated with diagnosis and treatment [53]. It makes patients feel better emotionally and physically and helps patients with cancer reduce their psychological distress [54]. Incorporating MBIs into oncological treatment can promote emotional and physical well-being and alleviate psychological distress [55]. MBIs have been found to regulate biological variables associated with stress [56], such as immune function, hypothalamic-pituitary-adrenal regulation, and autonomic nervous system activity, thereby reducing pressure on patients. The data from this review showed that MBCR appeared to be particularly effective in reducing perceived stress, whereas MBCT was not effective in reducing stress after the intervention [51]. This finding was unexpected, given that many previous studies have suggested the effectiveness of MBCT in reducing stress [57]. However, because of the limited number of included studies, it is difficult to draw definitive conclusions regarding the comparative effectiveness of different MBIs.

However, although not statistically significant, app- and website-based MBIs can improve the level of PTG and FCR in patients with cancer. FCR is one of the most common problems of survivors of cancer, and it has been known that FCR can persist throughout the treatment and survival trajectory [58]; thus, specific intervention is needed for survivors of cancer who have clinically significant FCR. Previous meta-analysis showed that cognitive therapy and mindfulness exercises are very suitable for combating FCR [59]. Numerous psychological and behavioral mechanisms of change within mindfulness interventions have been suggested, encompassing acceptance, emotion regulation skills, and the reduction of ruminative thoughts [60]. The meta-analysis by Gu et al [61] provided empirical confirmation that rumination significantly mediates the impact of MBIs on mental health outcomes. In addition, the study by Butow et al [62] identified rumination as a crucial psychological mechanism associated with FCR. Therefore, the study suggests that the effectiveness of mindfulness interventions in addressing the FCR may be attributed to their potential to improve patients' levels of rumination. The improved PTG observed in this study may be explained by the systematic training in moment-by-moment awareness, and MBIs focus on viewing thoughts and feelings as mental events [63]. Such a *decentered* relationship enables a perception of mental events as aspects of experience moving through awareness, showing that mindfulness practice supports personal growth and transformation.

In this study, it was observed that short-term MBIs with a duration of <8 weeks exhibited a larger effect size concerning the outcomes of anxiety and sleep. In the study by Wang et al [43], short-term MBIs were found to be more effective in improving physical health compared to long-term MBIs, and interventions lasting <8 weeks demonstrated a greater effect size, possibly attributed to the increased participant engagement resulting from the shorter intervention duration and simplified intervention complexity. Shorter interventions may be more feasible and acceptable for patients with cancer who are dealing with a range of physical and emotional challenges [64]. Future research should aim to replicate and expand on these results, including investigating the optimal duration and timing of app- and website-based MBIs for patients with cancer.

XSL•FO
RenderX

Recommendations for Future Research

To the best of our knowledge, this study represents the first meta-analysis using the RE-AIM framework, systematically reviewing and synthesizing the effectiveness of MBIs for patients with cancer across various types of interventions. By accurately reporting the RE-AIM dimensions, this study seeks to enhance the replicability and universality of mindfulness interventions in oncology settings. Our assessment of app- and website-based MBIs for patients with cancer, conducted within the framework of RE-AIM, reveals that the participation rates of eligible patients range from 13% to 92%. The calculated median participation rate, at 67% (IQR 47.5%-82%), emphasizes the effectiveness of the interventions in reaching a substantial portion of the target population. However, only a minority of studies reported on efficacy, which limited our ability to draw conclusions on overall effectiveness. Recruitment was primarily conducted by health professionals or researchers, and more than half of the studies (8/15, 53%) recruited participants in person, potentially limiting generalizability. Intervention adherence was generally high, but dropout rates varied widely, indicating that certain interventions may be more challenging for some patients. Cost was reported in only a few studies (4/15, 27%), with implications for accessibility. Long-term effects were reported in more than half of the studies (7/15, 47%), highlighting the need for further research. This study underscores the importance of considering the RE-AIM framework in the implementation and evaluation of these interventions. Further research is needed to fully understand their potential benefits and limitations in real-world settings.

Internet-based interventions have previously been shown to be effective for anxiety disorders and fear-related disorders and have achieved the same effect as face-to-face treatment [65]. Consistent with the results of this study, delivery via the internet, group, or app is feasible and effective. Our results suggest that among forms of online MBIs for patients with cancer, the most widely studied type was website-based interventions. This observation is in line with an analysis conducted in recent years [66], which indicated that the most widely studied type of telehealth for patients with breast cancer was website-based interventions. Website-based MBIs may offer more content, functionality, and instruction than app-based interventions, which may enhance user engagement, learning, and practice of positive thinking skills [67,68]. Website-based MBIs had higher completion rates and lower attrition rates compared to app-based interventions, which may be due to factors such as convenience, accessibility, engagement, and personalization [69]. Finally, in our review, a website-based study [41] that greatly improved FCR and stress highlighted the sustainability and self-management of the intervention and enabled flexible navigation by accessing website content according to user preferences. Therefore, website-based MBIs may offer more opportunities for personalization and tailoring interventions to individual needs.

In our analysis, 53% (8/15) of the studies implemented a weekly or daily reminder system through various channels, such as email, text messages, apps, or smartphone notifications, to facilitate app- and website-based MBIs. However, the prevalence of reminder systems in the studies under review is relatively limited (7/15, 47%), a finding consistent with the investigation by Matis et al [19]. Matis et al conducted a systematic evaluation in this field, discussing the limited prevalence of reminder systems in reviewed studies and highlighting the current lack of direct comparisons between interventions with and without reminders. In addition, the study found that the frequency of reminders was positively associated with the magnitude of the intervention effect [70]. Consequently, to promote patient involvement in app- and website-based MBIs, it is vital to set reminders [67]. Some studies have also set up expert feedback, answers, and a variety of supervision methods to avoid reduced patient compliance. Therefore, app- and website-based MBIs can enhance engagement using features such as reminders, feedback, personalization, and facilitator-led components. However, it is important to note that the specific frequency, timing, and content of the reminders may vary depending on the individual and the context of the intervention. Our study results reveal heterogeneity in the types, frequencies, and content of reminder systems, preventing the establishment of specific standards for their effectiveness. Despite the evident practicality of reminder systems, a more comprehensive investigation into their types, frequencies, and effectiveness is imperative within the context of app- and website-based MBIs.

This systematic review found that most app- and website-based interventions have adopted online classrooms; application-based measures to implement mindfulness interventions; and multicomponent interventions that include audio, video, and documents. However, the study did not clarify which factors affect behavioral changes. Despite these differences, 67% (10/15) of the interventions are designed specifically for the cancer population and provide customizable interventions. For example, as demonstrated by Wang et al [43], a pilot website-based MBIs was conducted for patients with cancer; this is an adapted version of MBSR specifically tailored for individuals dealing with cancer-related stressors. The MBCR program retains the core principles and practices of MBSR while integrating specific intervention materials to address challenges associated with cancer, such as common experiences related to cancer, sleep issues, pain, and FCR, which is greatly beneficial for improving the physical and mental symptoms of patients with cancer. MBCR will provide a platform for patients with cancer to engage in discussions and address challenges related to cancer. Future app- and website-based MBIs should take into account the characteristics of patients and determine which intervention plan is most suitable for patients with cancer, emphasizing feedback sessions and communication with therapists to enable patients to learn self-management and make intervention plans sustainable.

Limitations

Although this review summarizes international RCTs for various outcomes, there are limitations. First, because the research results are measured by various tools, it may hinder the comparability of research outcomes. Second, in the 15 trials, there are differences in the personnel, duration, and methods of app- and website-based MBIs in various studies. Patients included in these studies have different characteristics. Third, The inability to access or adequately translate studies in languages other than English and Chinese may introduce bias

XSL•FO
RenderX

into the selection process, potentially limiting the comprehensiveness of the findings. Finally, in the subgroup analysis, the study of each subgroup is limited, which may reduce the ability to draw conclusions on the differences in the consistency of intervention effects between subgroups. The abovementioned factors may lead to heterogeneity between studies, which is closely related to the summary results, so these results need to be interpreted carefully. Nevertheless, the meta-analysis included RCTs only and used a random effect model to pool results to give the most conservative estimates. In addition, subgroup analysis and sensitivity analysis were conducted and showed that the pooled estimations were relatively robust.

Conclusions

This meta-analysis provides definite evidence regarding the efficacy of app- and website-based MBIs for patients with cancer. Our findings suggest that app- and website-based MBIs can be effective in improving OL, sleep, and mental health and can be integrated into stepped care in clinical practice. Future experiments should pay more attention to the development of intervention programs based on the wishes and characteristics of patients with cancer and study how to optimize interventions further and customize interventions based on individual physical and mental symptoms.

Acknowledgments

This study was supported by the Exploration of Trajectories and Intervention Program of Frailty for Gastric Cancer Survivors Based on the Health Ecology Theory project supported by the National Natural Science Foundation of China (number 82073407), the Studies on Construction of Core Competency Model and Development of Assessment Tool for Nurses of Hospice Care project supported by the National Natural Science Foundation of China (number 72004099), and the Basic Science (Natural Science) Foundation of Higher Education Institutions of Jiangsu Province (project number 22KJB320013).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

TW and XJ developed the key ideas for the manuscript and the hypotheses. TW developed the search strategy and conducted the literature searches. YG, XJ, and CT conducted screening and coding. QX, CT, and SZ contributed to manuscript review and editing. TW conducted the statistical analyses, summarized the findings, and prepared the initial draft of the manuscript. All authors contributed to and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOC File , 79 KB - jmir_v26i1e47704_app1.doc]

Multimedia Appendix 2
Summary of the risk of bias of studies included in the systematic review.
[DOC File , 63 KB - jmir_v26i1e47704_app2.doc]

Multimedia Appendix 3
RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework.
[DOC File , 51 KB - jmir_v26i1e47704_app3.doc]

Multimedia Appendix 4
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.
[PDF File (Adobe PDF File), 99 KB - jmir_v26i1e47704_app4.pdf]

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]

XSL•FO
RenderX

2. Choi HC, Lam KO, Pang HH, Tsang SK, Ngan RK, Lee AW. Global comparison of cancer outcomes: standardization and correlation with healthcare expenditures. BMC Public Health 2019 Aug 07;19(1):1065 [FREE Full text] [doi: 10.1186/s12889-019-7384-y] [Medline: 31391013]
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48 [FREE Full text] [doi: 10.3322/caac.21763] [Medline: 36633525]
4. Wei X, Yuan R, Yang J, Zheng W, Jin Y, Wang M, et al. Effects of Baduanjin exercise on cognitive function and cancer-related symptoms in women with breast cancer receiving chemotherapy: a randomized controlled trial. Support Care Cancer 2022 Jul;30(7):6079-6091. [doi: 10.1007/s00520-022-07015-4] [Medline: 35416502]
5. Mead KH, Raskin S, Willis A, Arem H, Murtaza S, Charney L, et al. Identifying patients' priorities for quality survivorship: conceptualizing a patient-centered approach to survivorship care. J Cancer Surviv 2020 Dec;14(6):939-958. [doi: 10.1007/s11764-020-00905-8] [Medline: 32607715]
6. Nekhlyudov L, Mollica MA, Jacobsen PB, Mayer DK, Shulman LN, Geiger AM. Developing a quality of cancer survivorship care framework: implications for clinical care, research, and policy. J Natl Cancer Inst 2019 Nov 01;111(11):1120-1130 [FREE Full text] [doi: 10.1093/jnci/djz089] [Medline: 31095326]
7. Schell LK, Monsef I, Wöckel A, Skoetz N. Mindfulness-based stress reduction for women diagnosed with breast cancer. Cochrane Database Syst Rev 2019 Mar 27;3(3):CD011518 [FREE Full text] [doi: 10.1002/14651858.CD011518.pub2] [Medline: 30916356]
8. Hadash Y, Bernstein A. Behavioral assessment of mindfulness: defining features, organizing framework, and review of emerging methods. Curr Opin Psychol 2019 Aug;28:229-237. [doi: 10.1016/j.copsyc.2019.01.008] [Medline: 30959378]
9. Forte P, Abate V, Bolognini I, Mazzoni O, Quagliariello V, Maurea N, et al. Mindfulness-based stress reduction in cancer patients: impact on overall survival, quality of life and risk factor. Eur Rev Med Pharmacol Sci 2023 Sep;27(17):8190-8197 [FREE Full text] [doi: 10.26355/eurrev_202309_33579] [Medline: 37750647]
10. Mathew A, Doorenbos AZ, Jang MK, Hershberger PE. Acceptance and commitment therapy in adult cancer survivors: a systematic review and conceptual model. J Cancer Surviv 2021 Jun;15(3):427-451 [FREE Full text] [doi: 10.1007/s11764-020-00938-z] [Medline: 32949353]
11. Zhang D, Lee EK, Mak EC, Ho CY, Wong SY. Mindfulness-based interventions: an overall review. Br Med Bull 2021 Jun 10;138(1):41-57 [FREE Full text] [doi: 10.1093/bmb/ldab005] [Medline: 33884400]
12. Segal Z, Dimidjian S, Vanderkruik R, Levy J. A maturing mindfulness-based cognitive therapy reflects on two critical issues. Curr Opin Psychol 2019 Aug;28:218-222 [FREE Full text] [doi: 10.1016/j.copsyc.2019.01.015] [Medline: 30798103]
13. Schellekens MP, Tamagawa R, Labelle LE, Speca M, Stephen J, Drysdale E, et al. Mindfulness-Based Cancer Recovery (MBCR) versus Supportive Expressive Group Therapy (SET) for distressed breast cancer survivors: evaluating mindfulness and social support as mediators. J Behav Med 2017 Jun;40(3):414-422 [FREE Full text] [doi: 10.1007/s10865-016-9799-6] [Medline: 27722908]
14. Johns SA, Tarver WL, Secinti E, Mosher CE, Stutz PV, Carnahan JL, et al. Effects of mindfulness-based interventions on fatigue in cancer survivors: a systematic review and meta-analysis of randomized controlled trials. Crit Rev Oncol Hematol 2021 Apr;160:103290 [FREE Full text] [doi: 10.1016/j.critrevonc.2021.103290] [Medline: 33675902]
15. Park S, Sato Y, Takita Y, Tamura N, Ninomiya A, Kosugi T, et al. Mindfulness-based cognitive therapy for psychological distress, fear of cancer recurrence, fatigue, spiritual well-being, and quality of life in patients with breast cancer-a randomized controlled trial. J Pain Symptom Manage 2020 Aug;60(2):381-389 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.02.017] [Medline: 32105790]
16. Toivonen KI, Zernicke K, Carlson LE. Web-based mindfulness interventions for people with physical health conditions: systematic review. J Med Internet Res 2017 Aug 31;19(8):e303 [FREE Full text] [doi: 10.2196/jmir.7487] [Medline: 28860106]
17. van Agteren J, Iasiello M, Lo L, Bartholomaeus J, Kopsaftis Z, Carey M, et al. A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. Nat Hum Behav 2021 May;5(5):631-652. [doi: 10.1038/s41562-021-01093-w] [Medline: 33875837]
18. Xunlin NG, Lau Y, Klainin-Yobas P. The effectiveness of mindfulness-based interventions among cancer patients and survivors: a systematic review and meta-analysis. Support Care Cancer 2020 Apr;28(4):1563-1578. [doi: 10.1007/s00520-019-05219-9] [Medline: 31834518]
19. Matis J, Svetlak M, Slezackova A, Svoboda M, Šumec R. Mindfulness-based programs for patients with cancer via ehealth and mobile health: systematic review and synthesis of quantitative research. J Med Internet Res 2020 Nov 16;22(11):e20709 [FREE Full text] [doi: 10.2196/20709] [Medline: 33196452]
20. Fung JY, Lim H, Vongsirimas N, Klainin-Yobas P. Effectiveness of eHealth mindfulness-based interventions on cancer-related symptoms among cancer patients and survivors: a systematic review and meta-analysis. J Telemed Telecare 2024 Apr;30(3):451-465. [doi: 10.1177/1357633X221078490] [Medline: 35212553]
21. Bu S, Smith A, Janssen A, Donnelly C, Dadich A, Mackenzie LJ, et al. Optimising implementation of telehealth in oncology: a systematic review examining barriers and enablers using the RE-AIM planning and evaluation framework. Crit Rev Oncol Hematol 2022 Dec;180:103869. [doi: 10.1016/j.critrevonc.2022.103869] [Medline: 36356701]

XSL•FO
RenderX

22. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 03;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Page MJ, Li T, Cumpston M, Welch VA, Higgins JP, Thomas J, et al. Cochrane Handbook for Systematic Reviews of Interventions, Second Edition. London, UK: The Cochrane Collaboration; 2019.
25. Andrade C. Mean difference, standardized mean difference (SMD), and their use in meta-analysis: as simple as it gets. J Clin Psychiatry 2020 Sep 22;81(5):20f13681 [FREE Full text] [doi: 10.4088/JCP.20f13681] [Medline: 32965803]
26. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
27. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
28. Hodgson W, Kirk A, Lennon M, Janssen X, Russell E, Wani C, et al. RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) evaluation of the use of activity trackers in the clinical care of adults diagnosed with a chronic disease: integrative systematic review. J Med Internet Res 2023 Nov 13;25:e44919 [FREE Full text] [doi: 10.2196/44919] [Medline: 37955960]
29. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar 29;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
30. Singleton AC, Raeside R, Hyun KK, Partridge SR, Di Tanna GL, Hafiz N, et al. Electronic health interventions for patients with breast cancer: systematic review and meta-analyses. J Clin Oncol 2022 Jul 10;40(20):2257-2270 [FREE Full text] [doi: 10.1200/JCO.21.01171] [Medline: 35500200]
31. Chang YC, Chiu CF, Wang CK, Wu CT, Liu LC, Wu YC. Short-term effect of internet-delivered mindfulness-based stress reduction on mental health, self-efficacy, and body image among women with breast cancer during the COVID-19 pandemic. Front Psychol 2022 Oct 25;13:949446 [FREE Full text] [doi: 10.3389/fpsyg.2022.949446] [Medline: 36389600]
32. Compen F, Bisseling E, Schellekens M, Donders R, Carlson L, van der Lee M, et al. Face-to-face and internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with cancer: a multicenter randomized controlled trial. J Clin Oncol 2018 Aug 10;36(23):2413-2421. [doi: 10.1200/jco.2017.76.5669]
33. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry CJ, et al. A randomized controlled trial of mHealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. Integr Cancer Ther 2019 May 16;18:1534735419850634 [FREE Full text] [doi: 10.1177/1534735419850634] [Medline: 31092044]
34. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry CJ, et al. Pilot pragmatic randomized trial of mHealth mindfulness-based intervention for advanced cancer patients and their informal caregivers. Psychooncology 2024 Feb;33(2):e5557. [doi: 10.1002/pon.5557] [Medline: 32979294]
35. Liu Z, Li M, Jia Y, Wang S, Zheng L, Wang C, et al. A randomized clinical trial of guided self-help intervention based on mindfulness for patients with hepatocellular carcinoma: effects and mechanisms. Jpn J Clin Oncol 2022 Mar 03;52(3):227-236. [doi: 10.1093/jjco/hyab198] [Medline: 35088079]
36. Messer D, Horan JJ, Larkey LK, Shanholtz CE. Effects of internet training in mindfulness meditation on variables related to cancer recovery. Mindfulness 2019 Jun 4;10(10):2143-2151. [doi: 10.1007/s12671-019-01182-y]
37. Nissen ER, O'Connor M, Kaldo V, Højris I, Borre M, Zachariae R, et al. Internet-delivered mindfulness-based cognitive therapy for anxiety and depression in cancer survivors: a randomized controlled trial. Psychooncology 2020 Jan 18;29(1):68-75 [FREE Full text] [doi: 10.1002/pon.5237] [Medline: 31600414]
38. Peng L, Yang Y, Chen M, Xu C, Chen Y, Liu R, et al. Effects of an online mindfulness-based intervention on Fear of Cancer Recurrence and quality of life among Chinese breast cancer survivors. Complement Ther Clin Pract 2022 Nov;49:101686. [doi: 10.1016/j.ctcp.2022.101686] [Medline: 36347151]
39. Rosen KD. Is there an app for that? An exploratory randomized controlled trial of app-based mindfulness training for women with breast cancer. The University of Texas at San Antonio. 2016. URL: https://www.proquest.com/openview/025a3306918540d35302bcee3e4ac77b/1?pq-origsite=gscholar&cbl=18750 [accessed 2024-06-04]
40. Rosen KD, Paniagua SM, Kazanis W, Jones S, Potter JS. Quality of life among women diagnosed with breast cancer: a randomized waitlist controlled trial of commercially available mobile app-delivered mindfulness training. Psychooncology 2018 Aug 01;27(8):2023-2030. [doi: 10.1002/pon.4764] [Medline: 29766596]
41. Russell L, Ugalde A, Orellana L, Milne D, Krishnasamy M, Chambers R, et al. A pilot randomised controlled trial of an online mindfulness-based program for people diagnosed with melanoma. Support Care Cancer 2019 Jul 30;27(7):2735-2746. [doi: 10.1007/s00520-018-4574-6] [Medline: 30506103]

XSL•FO
RenderX

42. Shen A, Qiang W, Wang Q, Zhao Z, Wang S. Effects of online mindfulness-based cancer rehabilitation on perceived stress, fatigue, and anxiety in transitional breast cancer survivors. J Nurs Train 2021;36(23):2123. [doi: 10.16821/j.cnki.hsjx.2021.23.001]
43. Wang L, Chen X, Peng Y, Zhang K, Ma J, Xu L, et al. Effect of a 4-week internet-delivered mindfulness-based cancer recovery intervention on the symptom burden and quality of life of patients with breast cancer: randomized controlled trial. J Med Internet Res 2022 Nov 22;24(11):e40059 [FREE Full text] [doi: 10.2196/40059] [Medline: 36413385]
44. Yousefi A, Arani AM, Ghadiany M, Jafari M, Lighvan MA, Asanjarani B, et al. The effectiveness of online mindfulness-based cancer recovery program on psychological variables of colorectal and stomach cancer patients: a randomized control trial. Med Sci 2022 Feb;26(120). [doi: 10.54905/dissi/v26i120/ms55e2021]
45. Zernicke KA, Campbell TS, Speca M, McCabe-Ruff K, Flowers S, Carlson LE. A randomized wait-list controlled trial of feasibility and efficacy of an online mindfulness-based cancer recovery program: the eTherapy for cancer applying mindfulness trial. Psychosom Med 2014 May;76(4):257-267. [doi: 10.1097/PSY.0000000000000053] [Medline: 24804884]
46. Drucker AM, Fleming P, Chan AW. Research techniques made simple: assessing risk of bias in systematic reviews. J Invest Dermatol 2016 Nov;136(11):e109-e114 [FREE Full text] [doi: 10.1016/j.jid.2016.08.021] [Medline: 27772550]
47. Oberoi S, Yang J, Woodgate RL, Niraula S, Banerji S, Israels SJ, et al. Association of mindfulness-based interventions with anxiety severity in adults with cancer: a systematic review and meta-analysis. JAMA Netw Open 2020 Aug 03;3(8):e2012598 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.12598] [Medline: 32766801]
48. Vrontaras N, Koulierakis G, Ntourou I, Karakatsoulis G, Sergentanis TN, Kyrou D, et al. Psychosocial interventions on the posttraumatic growth of adults with cancer: a systematic review and meta-analysis of clinical trials. Psychooncology 2023 Dec;32(12):1798-1826. [doi: 10.1002/pon.6241] [Medline: 37964424]
49. Listiyandini RA, Andriani A, Kusristanti C, Moulds M, Mahoney A, Newby JM. Culturally adapting an internet-delivered mindfulness intervention for Indonesian university students experiencing psychological distress: mixed methods study. JMIR Form Res 2023 Aug 31;7:e47126 [FREE Full text] [doi: 10.2196/47126] [Medline: 37651168]
50. Wu R, Liu LL, Zhu H, Su WJ, Cao ZY, Zhong SY, et al. Brief mindfulness meditation improves emotion processing. Front Neurosci 2019;13:1074 [FREE Full text] [doi: 10.3389/fnins.2019.01074] [Medline: 31649501]
51. Cillessen L, van de Ven MO, Compen FR, Bisseling EM, van der Lee ML, Speckens AE. Predictors and effects of usage of an online mindfulness intervention for distressed cancer patients: usability study. J Med Internet Res 2020 Oct 02;22(10):e17526 [FREE Full text] [doi: 10.2196/17526] [Medline: 33006567]
52. Burrowes SA, Goloubeva O, Stafford K, McArdle PF, Goyal M, Peterlin BL, et al. Enhanced mindfulness-based stress reduction in episodic migraine-effects on sleep quality, anxiety, stress, and depression: a secondary analysis of a randomized clinical trial. Pain 2022 Mar 01;163(3):436-444 [FREE Full text] [doi: 10.1097/j.pain.0000000000002372] [Medline: 34407032]
53. Glynn BA, Khoo EL, MacLeay HM, Duong A, Cantave R, Poulin PA. Exploring cancer patients' experiences of an online mindfulness-based program: a qualitative investigation. Mindfulness (N Y) 2020;11(7):1666-1677 [FREE Full text] [doi: 10.1007/s12671-020-01380-z] [Medline: 32670431]
54. King K. How ‘mindfulness’ can improve mental focus and help you feel better emotionally and physically. WTOP News. 2020 Jul 28. URL: https://wtop.com/health-fitness/2020/07/how-mindfulness-can-improve-mental-focus-and-help-you-feel-better-emotionally-and-physically/ [accessed 2024-06-04]
55. Cillessen L, Johannsen M, Speckens AE, Zachariae R. Mindfulness-based interventions for psychological and physical health outcomes in cancer patients and survivors: a systematic review and meta-analysis of randomized controlled trials. Psychooncology 2019 Dec;28(12):2257-2269 [FREE Full text] [doi: 10.1002/pon.5214] [Medline: 31464026]
56. Witek Janusek L, Tell D, Mathews HL. Mindfulness based stress reduction provides psychological benefit and restores immune function of women newly diagnosed with breast cancer: a randomized trial with active control. Brain Behav Immun 2019 Aug;80:358-373. [doi: 10.1016/j.bbi.2019.04.012] [Medline: 30953776]
57. Pedro J, Monteiro-Reis S, Carvalho-Maia C, Henrique R, Jerónimo C, Silva ER. Evidence of psychological and biological effects of structured Mindfulness-Based Interventions for cancer patients and survivors: a meta-review. Psychooncology 2021 Nov;30(11):1836-1848 [FREE Full text] [doi: 10.1002/pon.5771] [Medline: 34288218]
58. Tauber NM, O'Toole MS, Dinkel A, Galica J, Humphris G, Lebel S, et al. Effect of psychological intervention on fear of cancer recurrence: a systematic review and meta-analysis. J Clin Oncol 2019 Nov 01;37(31):2899-2915 [FREE Full text] [doi: 10.1200/JCO.19.00572] [Medline: 31532725]
59. Hall DL, Luberto CM, Philpotts LL, Song R, Park ER, Yeh GY. Mind-body interventions for fear of cancer recurrence: a systematic review and meta-analysis. Psychooncology 2018 Nov;27(11):2546-2558 [FREE Full text] [doi: 10.1002/pon.4757] [Medline: 29744965]
60. Jain S, Shapiro SL, Swanick S, Roesch SC, Mills PJ, Bell I, et al. A randomized controlled trial of mindfulness meditation versus relaxation training: effects on distress, positive states of mind, rumination, and distraction. Ann Behav Med 2007 Feb;33(1):11-21. [doi: 10.1207/s15324796abm3301_2] [Medline: 17291166]
61. Gu J, Strauss C, Bond R, Cavanagh K. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? A systematic review and meta-analysis of mediation studies. Clin Psychol Rev 2015 Apr;37:1-12. [doi: 10.1016/j.cpr.2015.01.006] [Medline: 25689576]

XSL•FO
RenderX

62. Butow PN, Turner J, Gilchrist J, Sharpe L, Smith AB, Fardell JE, et al. Randomized trial of ConquerFear: a novel, theoretically based psychosocial intervention for fear of cancer recurrence. J Clin Oncol 2017 Dec 20;35(36):4066-4077. [doi: 10.1200/jco.2017.73.1257]
63. Britton WB. Can mindfulness be too much of a good thing? The value of a middle way. Curr Opin Psychol 2019 Aug;28:159-165 [FREE Full text] [doi: 10.1016/j.copsyc.2018.12.011] [Medline: 30708288]
64. Li J, Li C, Puts M, Wu YC, Lyu MM, Yuan B, et al. Effectiveness of mindfulness-based interventions on anxiety, depression, and fatigue in people with lung cancer: a systematic review and meta-analysis. Int J Nurs Stud 2023 Apr;140:104447. [doi: 10.1016/j.ijnurstu.2023.104447] [Medline: 36796118]
65. Domhardt M, Geßlein H, von Rezori RE, Baumeister H. Internet- and mobile-based interventions for anxiety disorders: a meta-analytic review of intervention components. Depress Anxiety 2019 Mar;36(3):213-224. [doi: 10.1002/da.22860] [Medline: 30450811]
66. Chen YY, Guan BS, Li ZK, Li XY. Effect of telehealth intervention on breast cancer patients' quality of life and psychological outcomes: a meta-analysis. J Telemed Telecare 2018 Apr;24(3):157-167. [doi: 10.1177/1357633X16686777] [Medline: 28081664]
67. Winter N, Russell L, Ugalde A, White V, Livingston P. Engagement strategies to improve adherence and retention in web-based mindfulness programs: systematic review. J Med Internet Res 2022 Jan 12;24(1):e30026 [FREE Full text] [doi: 10.2196/30026] [Medline: 35019851]
68. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002 [FREE Full text] [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
69. Sylvia LG, Lunn MR, Obedin-Maliver J, McBurney RN, Nowell WB, Nosheny RL, et al. Web-based mindfulness-based interventions for well-being: randomized comparative effectiveness trial. J Med Internet Res 2022 Sep 12;24(9):e35620 [FREE Full text] [doi: 10.2196/35620] [Medline: 36094813]
70. Mrazek AJ, Mrazek MD, Cherolini CM, Cloughesy JN, Cynman DJ, Gougis LJ, et al. The future of mindfulness training is digital, and the future is now. Curr Opin Psychol 2019 Aug;28:81-86. [doi: 10.1016/j.copsyc.2018.11.012] [Medline: 30529975]

Abbreviations

FCR: fear of cancer recurrence
FCRI: Fear of Cancer Recurrence Inventory
HADS: Hospital Anxiety and Depression Scale
MBCT: mindfulness-based cognitive therapy
MBI: mindfulness-based intervention
MBSR: mindfulness-based stress reduction
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTG: posttraumatic growth
QOL: quality of life
RCT: randomized controlled trial
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
SMD: standardized mean difference

Edited by T de Azevedo Cardoso; submitted 30.03.23; peer-reviewed by J Luo, K Kershner, F Chio; comments to author 31.12.23; revised version received 26.03.24; accepted 23.04.24; published 25.06.24.

Please cite as:
Wang T, Tang C, Jiang X, Guo Y, Zhu S, Xu Q
Effectiveness of Web-Based Mindfulness-Based Interventions for Patients With Cancer: Systematic Review and Meta-Analyses
J Med Internet Res 2024;26:e47704
URL: https://www.jmir.org/2024/1/e47704
doi:10.2196/47704
PMID:

©Ting Wang, Chulei Tang, Xiaoman Jiang, Yinning Guo, Shuqin Zhu, Qin Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Design of Digital Mental Health Platforms for Family Member Cocompletion: Scoping Review

Ellen T Welsh[1], BSc, MEng; Jennifer E McIntosh[1], PhD; An Vuong[1], BPsy (Hons); Zoe C G Cloud[1], PhD; Eliza Hartley[1], DPsych; James H Boyd[2], PhD

[1]The Bouverie Centre, School of Psychology and Public Health, La Trobe University, Brunswick, Australia

[2]School of Psychology and Public Health, La Trobe University, Bundoora, Australia

Corresponding Author:
Ellen T Welsh, BSc, MEng
The Bouverie Centre
School of Psychology and Public Health
La Trobe University
8 Gardiner Street
Brunswick, 3056
Australia
Phone: 61 384814800
Email: e.welsh@latrobe.edu.au

Abstract

Background: The COVID-19 pandemic placed an additional mental health burden on individuals and families, resulting in widespread service access problems. Digital mental health interventions suggest promise for improved accessibility. Recent reviews have shown emerging evidence for individual use and early evidence for multiusers. However, attrition rates remain high for digital mental health interventions, and additional complexities exist when engaging multiple family members together.

Objective: As such, this scoping review aims to detail the reported evidence for digital mental health interventions designed for family use with a focus on the build and design characteristics that promote accessibility and engagement and enable cocompletion by families.

Methods: A systematic literature search of MEDLINE, Embase, PsycINFO, Web of Science, and CINAHL databases was conducted for articles published in the English language from January 2002 to March 2024. Eligible records included empirical studies of digital platforms containing some elements designed for cocompletion by related people as well as some components intended to be completed without therapist engagement. Platforms were included in cases in which clinical evidence had been documented.

Results: Of the 9527 papers reviewed, 85 (0.89%) met the eligibility criteria. A total of 24 unique platforms designed for co-use by related parties were identified. Relationships between participants included couples, parent-child dyads, family caregiver–care recipient dyads, and families. Common platform features included the delivery of content via structured interventions with no to minimal tailoring or personalization offered. Some interventions provided live contact with therapists. User engagement indicators and findings varied and included user experience, satisfaction, completion rates, and feasibility. Our findings are more remarkable for what was absent in the literature than what was present. Contrary to expectations, few studies reported any design and build characteristics that enabled coparticipation. No studies reported on platform features for enabling cocompletion or considerations for ensuring individual privacy and safety. None examined platform build or design characteristics as moderators of intervention effect, and none offered a formative evaluation of the platform itself.

Conclusions: In this early era of digital mental health platform design, this novel review demonstrates a striking absence of information about design elements associated with the successful engagement of multiple related users in any aspect of a therapeutic process. There remains a large gap in the literature detailing and evaluating platform design, highlighting a significant opportunity for future cross-disciplinary research. This review details the incentive for undertaking such research; suggests design considerations when building digital mental health platforms for use by families; and offers recommendations for future development, including platform co-design and formative evaluation.

(J Med Internet Res 2024;26:e49431) doi:10.2196/49431

XSL•FO
RenderX

KEYWORDS

family therapy; family; couples; eHealth; digital health; platform; platforms; e–mental health; internet interventions; psychosocial interventions; psychosocial; synthesis; review methods; review methodology; scoping; mental health; utility; design; family caregiver; caregiver; parent; child; development; cocompletion; access; accessibility; engagement; families; dyad; dyadic; user engagement; digital health; user experience; mobile phone

Introduction

Family Mental Health

Normatively, mental health disorders impacted >1 billion people worldwide in 2016 [1]. The COVID-19 pandemic brought further substantial impact on mental health, placing increased demand on mental health services [2]. Mental health is inherently relational [3,4], and family members and partners are inevitably impacted by an individual's mental health challenges [5]. During the COVID-19 pandemic, markers of heightened family stress included rising rates of family violence [6]; increased parenting stress [7]; and observed rates of maladaptive parenting practices, including neglectful, harsh, and coercive parenting [8-10].

There is a strong evidence base for family and systemic interventions for child- and adult-focused mental health challenges. Family participation supports members of the family to safely contribute to individual recovery and improved relationships [11-13] and can be more beneficial than individual work [14-16] and family educational interventions [17]. In addition, parent involvement in interventions for childhood behavioral [18] and adolescent anxiety disorders [19] has been shown to be beneficial and contributes to positive long-term outcomes.

Digital Mental Health

The World Health Organization has emphasized the significant potential of digital mental health interventions (DMHIs) in expanding reach and access to services [20]. Such DMHIs have shown promise in reaching underserved populations [21], leading to improved management of symptoms in individuals [22], particularly youth aged <25 years [23,24]. There is growing meta-analytic evidence for positive mental health outcomes of digitally delivered versus in-person individual treatment, for example, in the field of cognitive behavioral interventions [25]. With rapid developments in technology, research interest is expanding, with most of the literature so far focused on DMHIs for individuals. For example, a review of systematic reviews of digital interventions for mental health and well-being (with no limitations placed on population) conducted in 2021 identified 246 systematic reviews published between 2016 and 2021, all of which reviewed digitally delivered mental health interventions for individuals [26].

Beyond DMHIs designed for individuals, 2 first-generation reviews of dyadic (caregiver and care recipient) [27] and couple-targeted DMHIs [28] suggest that DMHIs can decrease barriers and improve timely access and outcomes for distressed relationships. However, research into DMHIs for *families to access together* is as yet undeveloped.

Despite growing evidence, and regardless of the population targeted, retention rates for DMHIs remain low, limiting their ultimate impact [29-32]. Among other factors, interface ease of use has been identified as a barrier to DMHI retention and engagement by individuals [25,33]. It is likely that similar (or possibly even greater) barriers for family engagement in the digital mental health space exist. Given the fundamental differences in the approach and focus for family and relational interventions when compared to interventions designed for individuals [3,34], it is likely that there are unique factors to consider when designing DMHIs for use by families. This might include considerations for individual user privacy and ways in which the platform allows multiple people to contribute to and especially cocomplete activities, such as shared goal setting. Thus, it would be ill-founded to extrapolate results from studies on DMHIs designed for use by individuals and assume similar platform interaction values for families. The need for further research specific to the design of DMHIs for family use is clear.

Design of DMHIs for Families

Therefore, the question arises about what an effective DMHI for family use might look like. Given that computers and tablets are designed for use by individuals, DMHIs intended for cocompletion by family members may use different platform and interface features to support and sustain family engagement. No review to date has examined evidence for design and build characteristics that promote cocompletion usability, including improved engagement and accessibility.

In that light, this review aimed to synthesize the available evidence regarding the build and design characteristics that enable cocompletion and discuss reported indicators of user engagement with platforms designed for such use, namely, usability, satisfaction, acceptability, and feasibility. In the digital mental health literature, these user engagement indicators measure the ability of a platform to engage and sustain users. However, there is a notable lack of agreement on both the definition and measurement of the construct of *engagement*, which can lead to inappropriate selection, presentation, and interpretation of user engagement indicators across studies [35]. As such, a scoping review was conducted, and we adopted the definition of user engagement as outlined by Perski et al [36]: "Engagement with [Digital Behaviour Change Interventions] is (1) the extent (e.g. amount, frequency, duration, depth) of usage and (2) a subjective experience characterised by attention, interest and affect."

In this scoping review, we differentiate the term "platform" from the term "intervention." We define "platform" as the tools, infrastructure, and technical foundation behind the delivery of an intervention, including interface characteristics such as the design, layout, and delivery mode. We define "intervention" as the mental health–related content that is delivered via the platform. This review sought to understand (1) the design and functionality characteristics that enable the effective engagement with and cocompletion of a family-oriented DMHI and (2)

XSL•FO
RenderX

whether these elements moderate the effect of the intervention on mental health or relational outcomes. To distinguish effective platform contributors to engagement from elements pertaining to intervention content, we selected only those platforms housing interventions of established clinical efficacy (which we defined as any intervention that had at least one study reporting a significant improvement in a mental health or relational outcome). In addition, it is expected that build characteristics may vary by population, and given that there is no uniform family composition, this review scoped platforms designed for cocompletion by any family relationship type, including couples, family subsystems, and whole families.

Methods

Search Strategy

To identify studies reviewing platforms delivering clinical interventions designed for cocompletion by families, a systematic search was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [37]. A comprehensive electronic literature search for articles published in English was conducted in the following databases: MEDLINE, Embase, and PsycINFO via the Ovid platform; CINAHL via the EBSCOhost platform, and Web of Science. In line with developments in digital technology, studies were included if they were published in or since 2002. The search was first conducted on June 24, 2022, and additional searches were conducted on November 24, 2022; April 21, 2023; and March 15, 2024.

Eligibility Criteria

As advised by the Joanna Briggs Institute's guidelines for conducting scoping reviews [38], the population, concept, and context framework was used to define eligibility. Textbox 1 shows the inclusion and exclusion criteria in line with the population, concept, and context framework and contains additional study elements relevant to the eligibility criteria.

Studies were not excluded when platforms contained additional components involving practitioner (sometimes referred to in the studies as a coach, professional, therapist, or staff member) engagement. Further to the inclusion and exclusion criteria outlined in Textbox 1, platforms offering interventions that had no evidence of clinical efficacy (ie, no identified studies that reported any significant improvements in mental health or relational outcomes) were excluded. Provided that at least 1 identified study established clinical efficacy for that platform, all studies on that intervention were then included regardless of whether they reported on clinical outcomes. Platforms that met all the other inclusion criteria but without established clinical efficacy are presented in Multimedia Appendix 1.

Textbox 1. Inclusion and exclusion criteria detailing the population, concept, and context framework for defining eligibility criteria for scoping reviews and additional study elements.

Inclusion criteria

- Population: Digital mental health interventions (DMHIs) designed for completion by at least 2 related people together
- Concept: Platform design elements of DMHIs (via a web or smartphone interface) containing some component that was intended to be completed without therapist or human intervention (ie, was self-directed by participants)
- Context: Open and included all care settings (eg, primary care and community) and all jurisdictions and geographic locations
- Study type and design: Empirical studies
- Publication date: from January 1, 2002, to March 15, 2024
- Publication language: English

Exclusion criteria

- Population: DMHIs designed for completion by individuals or designed for use by related people but with no activities completed together (ie, completed separately) and DMHIs where children were the focus and the parent's role was only in assisting their child to participate
- Concept: DMHIs in which the target condition was physical illness, physical activity, and weight management and programs delivered through virtual reality devices, wearable devices, DVD, or other non–web-based approaches
- Study type and design: Nonempirical studies and gray literature (ie, non–peer-reviewed or unpublished manuscripts)

Search and Data Extraction Methodology

A total of 3 key search constructs addressed the different elements of the research question: digital intervention, mental or relational health, and population. Results were combined using Boolean operators. The search strategies for each database can be found in Multimedia Appendix 2. The reference lists of relevant reviews were also screened for potentially relevant studies. Data extraction was completed by 2 researchers trained in systematic search methodology using a standardized template, and discrepancies were resolved through discussion between the 2 researchers. In cases in which it appeared that there could be cocompletion but it was not directly specified, the study authors were contacted, and websites were searched.

Screening and Selection Process

Search results were downloaded into EndNote (Clarivate Analytics) [39] and imported into Covidence (Veritas Health Innovation) [40]. Duplicates were first removed in EndNote and again following import into Covidence. In total, 2 researchers screened the identified studies at the title and abstract level, with 20% being double screened. Disagreements were resolved through discussion. A total of 2 researchers screened the articles at the full-text level with 20% double screening to

XSL•FO
RenderX

determine eligibility against the inclusion criteria outlined previously. Reasons for exclusion at the full-text level were recorded.

Data Synthesis

Data were synthesized using a narrative approach. Due to high variability in the reporting of outcomes and measurements across studies, a systematic or meta-analytic approach was not possible.

The included articles were grouped by the digital platform used. Information regarding the authors, the year of publication, the country where the study took place, the population, and associated user engagement indicators was extracted. Significant differences in mental health or relational outcomes following the DMHI were indicated. Details about the platforms were extracted into a separate table. Also detailed were the intervention target; the relationship between the participants; components designed to be completed in a self-paced manner, together, individually, or with a professional; tailored components; and any additional key features. Results were categorized and synthesized based on the targeted relationship for the intervention (eg, couples or families).

Results

Overview

The combined searches yielded 17,765 results. Following removal of 46.37% (8238/17,765) of duplicates in EndNote and Covidence, 9527 papers were screened at the title and abstract level, resulting in 9184 (96.4%) exclusions. A total of 343 full-text articles were reviewed for inclusion, with 263 (76.7%) exclusions. Reasons for exclusion included the platform being designed for use by individuals (154/263, 58.6%), nonempirical studies (55/263, 20.9%), the platform not containing any self-guided components (36/263, 13.7%), or wrong indication (eg, weight loss intervention; 18/263, 6.8%). A total of 80 studies were included for data extraction. An additional 5 studies were identified through reference scanning and included in data extraction, resulting in a total of 85 studies included in this review. Figure 1 shows the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) diagram [37].

The following sections first summarize the studies identified and then report on characteristics of and findings related to the included platforms.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) diagram showing the inclusion and exclusion of studies at each stage of the review process.

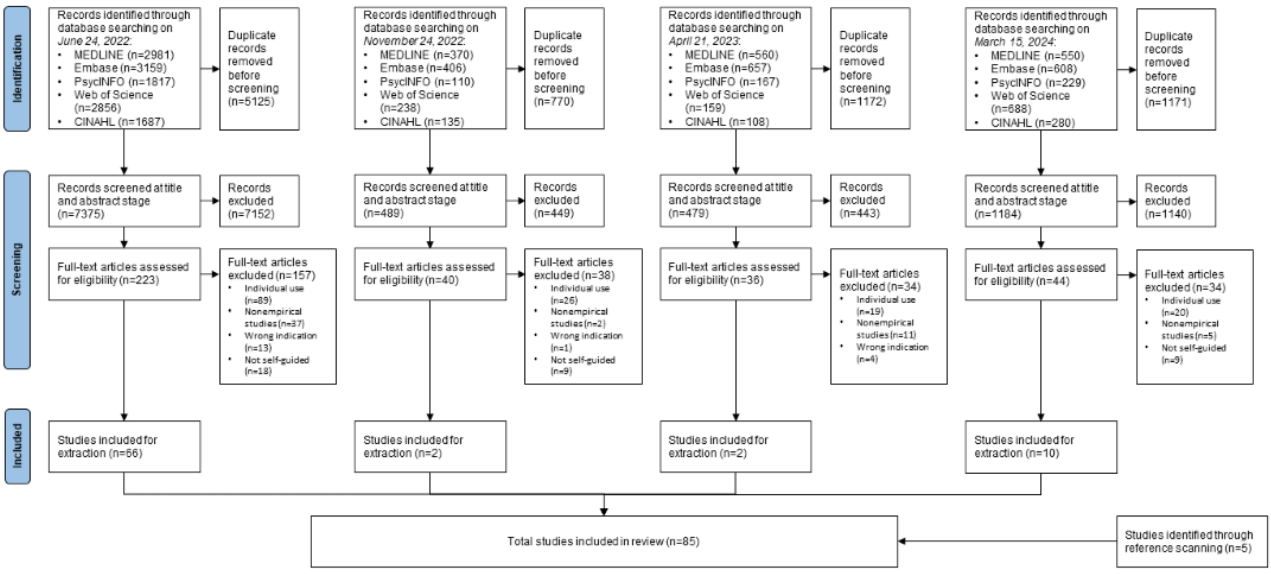

Included Studies

Table 1 details the characteristics of the 85 studies, including study type, their population and sample size, usability measures and findings, and an indication of clinical efficacy based on significant improvement in mental health or relational outcomes following completion of the intervention. Among the 85 included studies, data were collected during randomized controlled trials (n=63, 74%), pilot feasibility studies (n=14, 16%), single-arm studies (n=7, 8%), and nonrandomized quasi-experimental studies (n=1, 1%).

A total of 74% (63/85) of the studies were conducted in the United States; 12% (10/85) were conducted in Canada; 5% (4/85) were conducted in Australia; 2% (2/85) were conducted in the United Kingdom; 2% (2/85) were conducted in China; and 1% (1/85) were conducted each in the Netherlands, Sweden, Japan, and Korea. In total, 52% (44/85) of the included studies were published between January 2019 and March 2024, whereas 5% (4/85) of the studies were published in the first 5 years of the search period (2002-2006 inclusive) and the remainder (37/85, 43%) were published in between these periods.

XSL•FO
RenderX

Table 1. Characteristics of the studies meeting the eligibility criteria, including name of the platform examined; study and country; type of study and comparator (where applicable); population, sample size, and attrition rate; relational and individual constructs or outcomes; reported user engagement indicators; and corresponding findings.

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
4Cs:CRC[a] [41]; China	Pilot feasibility trial	Heterosexual couples where one member was experiencing colorectal cancer; 24 couples; 16.7%	Dyadic coping; cancer-related communication; self-efficacy; physical and mental health; positive and negative emotions	(1) Postintervention evaluation; (2) feasibility and acceptability	(1) Highly rated usefulness, ease of use, and satisfaction; all mean acceptability ratings >5.2/7; (2) 83.8% retention; 609 session views; mean 29 views per page; mean 3-7 page views per session per dyad
4Cs:CRC [42]; China	RCT[b]; web-based, face-to-face, blended, or control	Heterosexual couples where one member had colorectal cancer; 212 couples; 16%	Dyadic coping[c]; cancer-related communication[c]; marital satisfaction; self-efficacy[c]; physical and mental[c] health; positive[c] and negative[c] emotions	Not reported	Not reported
CA-CIFFTA[d] [43]; United Kingdom	RCT; no treatment	Hispanic (80%) and Black (20%) adolescents and their families; 80 parent-child dyads; 27% (intervention)	Family cohesion[c]; family conflict; parenting practices; adolescent behavioral problems[c]	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
C-MBI[e] for YBCSs[f] [44]; United States	RCT; MBI[g] completed by YBCSs only (I-MBI[h])	Female breast cancer survivors and their male partners; 117 couples; 26% (I-MBI) and 38% (C-MBI)	Couple functioning[c] (I-MBI only); individual-level functioning[c]	(1) Feasibility and acceptability of YBCSs (self-report); (2) feasibility and acceptability of partners (self-report)	(1) 39% requested more contact with peers; 63% would recommend it; 77% watched all videos; 90% used the supplemental material; 91% completed some or all of the assignments; rated most useful: mindfulness sessions (80%), yoga (14%), and partner interaction (7%); time constraints were the most cited reason for not recommending the intervention; (2) 93% had no desire to interact with peers; 69% would recommend it; 69% watched all videos; 89% used the supplemental materials; 92% completed some or all of the assignments; time constraints were the most cited reason for not recommending the intervention

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Cool Kids Online [45]; Australia	RCT; waitlist	Children (aged 7-12 years) with anxiety and their parents or caregivers; 95 dyads; 12% at posttreatment time point and 27% at 6-month follow-up	Anxiety diagnosis[c]; anxiety scale; life interference[c] (parent only); mood and feelings[c]; strengths and difficulties[c]	(1) Satisfaction; (2) completion	(1) 73% of parents were satisfied or very satisfied, 92% reported it as helpful, and 97% were moderately or very confident recommending the intervention; 64% of children were "happy" with the intervention, 89% reported it as helpful or very helpful, and 70% were moderately or very confident that it would help a friend; (2) 83% accessed all lessons (mean 7.52, SD 1.23; range 3-8); received a mean of 8.8/10 (SD 1.61; range 3-10) calls
Couple HOPES[i] [46]; Canada	Pilot feasibility trial; pretest-posttest	Couples where one member was a military member, veteran, or first responder with PTSD[j] symptoms; 10 couples; 30%	Relationship satisfaction[c] (partners only); conflict; PTSD symptoms[c]; partner's accommodations to PTSD symptoms[c]; anxiety, distress, and QoL[c,k]; AOD[l] use	Satisfaction (CSQ[m])	PTSD: mean 3.4/4 (SD 0.7); partner: mean 3.7/4 (SD 0.4)
Couple HOPES [47]; Canada	Single arm	Couples where one member was a military member, veteran, or first responder with PTSD symptoms; 17 couples; 35%	Relationship satisfaction[c] (ineffective arguing); PTSD self-report; partner report of PTSD symptoms; mental health; well-being[c] (perceived health); partner accommodations	Satisfaction (CSQ)	PTSD: mean 3.5/4 (SD 0.6); partner: mean 3.7/4 (SD 0.3)
Couple HOPES [48]; Canada	Single arm	Couples where one member was a military member, veteran, or first responder with PTSD symptoms; 27 couples; 33%	Relationship functioning[c] (ineffective arguing); mental health; well-being[c] (perceived health and QoL)	Satisfaction (CSQ)	Partner: mean 3.7/4 (SD 0.4)

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Couple HOPES [49]; United States	Pilot feasibility trial; pretest-posttest	Couples where one member was a military veteran with PTSD; 15 couples; 27%	Relationship satisfaction[c]; relationship quality[c] (negative relationship quality); PTSD symptoms[c]; depression; QoL; significant other's response to trauma	(1) Completion; (2) feedback	(1) Mean duration 7.20 (SD 5.56) weeks; n=11 completed; 4 noncompleters (n=2 completed 4/7 modules, n=1 completed 2/7, and n=1 completed 1/7); n=3 "treatment responders" completed it faster; (2) coach was helpful for processing information, thoughts, and feelings; feedback videos were unrealistic or "cheesy," others found them helpful for digesting and relating to the material
Couplelinks [50]; Canada	Pilot feasibility trial	Heterosexual couples where a member had a breast cancer diagnosis; 16 couples; 38%	Not reported	(1) Treatment satisfaction (TSQ[n]); (2) usability	(1) Mean 4/5 (SD 0.56); (2) mean 4/5 (SD 0.83)
Couplelinks [51]; Canada	Pilot feasibility trial	Heterosexual couples where a member had a breast cancer diagnosis; 6 couples; not reported	Not reported	Engagement promotion by therapist	Rational model of engagement promotion: friendly and positive yet firm approach, humanizing technology, and inclusive and empathic attitude; empirical model of engagement promotion: fostering couple-facilitator bond, fostering intervention adherence, and fostering within-couple bond
Couplelinks [52]; Canada	Pilot feasibility trial	Heterosexual couples where a member had a breast cancer diagnosis; 12 couples; not reported	Not reported	Types of engagement	Couple "types"—keen: completed with minimal engagement; compliant: met facilitator deadlines; apologetic: enjoyed it and were committed but had trouble staying on track; straggling: least engaged

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Couplelinks [53]; Canada	Pilot feasibility trial	Heterosexual couples where a member had a breast cancer diagnosis; 13 couples; not reported	Not reported	Perceived benefits and limitations	58% agreed or strongly agreed that it was beneficial; 35% said that it was somewhat beneficial
Couplelinks [54]; Canada	RCT; waitlist	Heterosexual couples where a member had a breast cancer diagnosis; 67 couples; 20.5% in the intervention group and 0% in the control group	Dyadic coping[c]; dyadic consensus, cohesion, and satisfaction; marital satisfaction; collective coping[c]; anxiety[c] and depression	Not reported	Not reported
Couplelinks [55]; Canada	RCT; waitlist	Heterosexual couples where a member had a breast cancer diagnosis; 57 participants; not reported	Not reported	Treatment satisfaction (TSQ)	Mean 4.3/5 (SD 0.54); female participants' satisfaction ratings were significantly higher (P=.01); medium effect size (0.57)

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
eMB[o] [56]; United States	RCT; control	Couples where one member was pregnant; 30 couples; 0%	Anxiety[c] (pregnant person's anxiety) and depression symptoms	(1) Satisfaction (CSQ-8[p]); (2) completion rates and adherence	(1) Excellent satisfaction: mean 3.42 (SD 0.55); pregnant: mean 3.42 (SD 0.59); and partner: mean 3.43 (SD 0.49); factors perceived to promote engagement included flexibility (independent and joint options and own pace) and focus on the self before talking to their partner; helpful elements included videos, web-based exercises, and activities; factors impacting engagement included video relatability, poor quality, outdated images, simplistic and low-technology visualizations, videos perceived as old or silly, extreme vignettes and illustrations, and videos being overly dramatized and unrelatable; (2) 50% used it alone, 9% used it together with their partner, and 27% were a combination of both; 14% did not engage; 0% completed 1 lesson per week as advised; 83 discrete log-ins; pregnant people visited more (mean 4.17 vs mean [partners] 3.44 visits to the intervention)
Embers the Dragon [57]; United Kingdom	Pilot feasibility trial; no treatment	Children aged 2-7 years and a parent; 129 families; 7.7% in the intervention group and 20.4% in the control group	Parental responses to childhood behaviors[c]	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
ePREP[q] (studies on ePREP and OurRelationship reported separately) [58]; United States	RCT; IRC[r]	Heterosexual couples in long-term relationships; 77 couples; 0%	Commitment attitudes[c]; communication[c]; relationship satisfaction[c]; psychological aggression and assault[c]; depression, dysphoria, and well-being[c]; anxiety[c]	Level of engagement as a moderator of clinical outcomes	

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
					Higher engagement (measured via results on quizzes): greater intervention effect for alternative monitoring (β=–.33; *P*=.04), constructive communication (β=.29; *P*=.07), self-reported physical assault (β=–.58; *P*=.11), male relationship satisfaction (β=.48; *P*=.02), and female depression (β=–.37; *P*=.10). Greater time spent completing homework assignments: greater intervention effect for reported couple physical assault (β=–.69; *P*=.06), severe psychological aggression for male (β=–.90; *P*=.02) and female (β=–.09; *P*=.01) individuals, and male-perpetrated physical assault (partner report; β=–1.10; *P*=.02) but an attenuation of the positive effect of ePREP on self-reported minor psychological aggression (male individuals: β=.40 and *P*=.11; female individuals: β=.43 and *P*=.12). Male individuals with higher engagement experienced attenuation of positive impact on anxiety (β=.35; *P*=.01), and female individuals who completed more homework assignments experienced attenuation of positive impact on depres-

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
					sion symptoms (β=.45; *P*=.03).
ePREP [59]; United States	RCT; IRC	Married couples; 52 couples; 4% after the intervention and 92% at the 1-year follow-up (8% in the intervention group and 7.6% in the control group)	Conflict resolution methods[c]; psychological aggression and assault[c]	Not reported	Not reported
FOCUS[s] [60]; United States	Single arm; repeated measures	Patient-caregiver dyads; 38 dyads; 14%	Communication; social support; emotional distress[c]; QoL[c]; appraisal[c]; coping resources; self-efficacy	(1) Satisfaction; (2) comfort and skill using computers and the internet; (3) feasibility	(1) Ease of use: mean 6.0/7 (SD 1.1); usefulness: mean 4.4/7 (SD 1.4); general satisfaction: mean 4.8 (SD 1.7); no adverse effects of completing the intervention together; (2) moderate skill level; (3) lower enrollment rate than previous in-person RCTs (51% compared with 68%-80%); retention rate was higher than in-person RCTs (86% compared with 62%-83%)
iCBT[t] [61]; Sweden	RCT; waitlist	Families where the child (aged 8-12 years) had a mental health diagnosis; 93 families (93 children and 182 parents); 2% in the intervention group and 4% in the control group	Anxiety[c] (parent reported); development and well-being; child depression; primary carer mental health	(1) Satisfaction; (2) compliance	(1) Child satisfaction: mean 3.67; parent satisfaction: mean 3.78; 86% of parents agreed or very much agreed that they would recommend it; 82% of children agreed or very much agreed that the treatment was effective; (2) completed modules: mean 9.7 (SD 1.8; range 4-11); 83% completed the first 9 modules; 4 families did not complete the modules intended for both children and parents

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Military Family Foundations [62]; United States	RCT; no treatment	Heterosexual couples expecting their first child where one member was in the military; 56 couples; 34.5% for mothers and 48.3% for fathers in the intervention group and 7.4% for mothers and 22.2% for fathers in the control group	Interparental relationship[c] (mothers only); parental adjustment[c]; parent report of child outcomes[c] (sadness)	Completion	Mean 3.93/8 completed modules
MindGuide Couple [63]; South Korea	Single arm	Korean heterosexual couples; 17 couples; 11%	Couple relationship satisfaction; family relationship[c]; mental health; positive and negative emotions; satisfaction with life[c]	(1) Satisfaction and acceptability; (2) recruitment, retention, and completion	(1) 100% reported that the content and tasks were helpful; 90% reported that the content was applicable to everyday activities; coaching was most helpful (90%), followed by video lectures (43%) and practical tasks (43%); reported benefits included flexible access (90%), being less burdensome than face-to-face interventions (86.3%), and no geographic limitations (76.7%); reported drawbacks included being too long (33.3%) and time burden (76.7%); 93.4% were satisfied; 100% were satisfied with the level of coaching; (2) 94.1% completed
MR[u] [64]; United States	RCT; MR plus PREP[v], PREP alone or waitlist	Veteran-partner dyads; 320 individuals (160 couples); 1.2% for MR, 2.5% for MR[c] plus PREP, 1.2% for PREP alone, and 0% for waitlist	Perceived social support; dyadic adjustment; stress[c]; depression[c]; PTSD symptoms[c]; self-compassion[c]; response to stressful experiences[c]; sleep quality; physical pain	(1) Intervention use; (2) satisfaction	(1) Mean 2.5 hours of use per week; at 16-week follow-up: mean 90 minutes per week; (2) likely to recommend: mean (veterans) 8.7/10 and mean (partners) 9.1/10

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Mother-daughter program [65]; United States	RCT; waitlist	Girls aged 10-13 years and their mothers; 202 dyads; 0% between pre- and posttest, and 2% in the intervention group and 1% in the control group lost between postintervention time point and follow-up	Mother communication[c]; conflict management[c]; daughter communication[c]; perceived rules[c]; parental monitoring; normative beliefs[c]; self-efficacy[c]; alcohol use[c]; drinking intention[c]; refusal skills; parental rules[c]; parental monitoring[c]	Anonymous program rating	Improved mother-daughter relationship: mean (girls) 4.14/5 (SD 0.35) and mean (mothers) 4.25/5 (SD 0.29); learned useful information: mean (girls) 4.16/5 (SD 0.38) and mean (mothers) 4.13/5 (SD 0.34); enjoyed the intervention: mean (girls) 4.07/5 (SD 0.39); mean (mothers) 4.20/5 (SD 0.34); found time to complete it together: mean (girls) 3.04/5 (SD 0.37); mean (mothers) 3.24/5 (SD 0.33)
Mother-daughter program [66]; United States	RCT; no treatment	Girls aged 11-13 years and their mothers; 591 dyads; 3.2% in the intervention group	Mother-daughter communication[c]; substance use[c]; family rules[c]; parental monitoring[c]; normative beliefs[c]; depression; problem-solving skills; body esteem; drug refusal self-efficacy[c]; intentions[c]	Not reported	Not reported
Mother-daughter program [67]; United States	RCT; no treatment	Girls aged 11-13 years and their mothers; 916 dyads; 5.7% from baseline to 1-year follow-up and 4.2% between 1- and 2-year follow-up	Communication[c]; mother-daughter closeness[c]; family rules[c]; parental monitoring[c]; body esteem; depression; coping ability[c]; normative beliefs[c]; refusal self-efficacy[c]; substance use[c]; intentions[c]; family rituals[c]	Not reported	Not reported
Mother-daughter program [68]; United States	RCT; no treatment	Asian American girls aged 11-14 years and their mothers; 108 dyads; 3.5% in the intervention group and 3.8% in the control group	Mother-daughter closeness[c]; mother-daughter communication[c]; substance use[c]; intentions; depression[c]; self-efficacy[c]; refusal skills[c]; parental monitoring[c]; family rules[c]	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Mother-daughter program [69]; United States	RCT; no treatment	Asian American girls aged 11-14 years and their mothers; 108 dyads; 89.2% completed the 2-year measure	Mother-daughter closeness[c] (girls only); mother-daughter communication[c]; parental monitoring[c] (girls only); family rules[c] (girls only); depressive symptoms; body esteem; self-efficacy[c]; refusal skills[c]; normative beliefs; substance use[c]; intentions[c]	Completion	96.4% completed the entire intervention; 94.6% completed the booster session; participants completed initial 9 sessions (mean 175, SD 68.9 days)
Mother-daughter program [70]; United States	RCT; no treatment	Black and Hispanic girls aged 10-13 years and their mothers; 564 dyads; 6.6% in the intervention group and 3.3% in the control group	Mother-daughter closeness; mother-daughter communication[c]; substance use[c]; normative beliefs[c]; intentions[c]; depression[c]; self-efficacy[c]; refusal skills; parental monitoring[c]; family rules[c]; body esteem	Not reported	Not reported
Mother-daughter program [71]; United States	RCT; no treatment	Mother-daughter dyads in public housing; 36 dyads; 3%	Mother-daughter closeness[c]; mother-daughter communication[c]; parental monitoring[c]; substance use; fruit and vegetable intake[c]; physical activity[c]; perceived stress[c]; drug refusal skills[c]	Fidelity	97% completed all 3 sessions
OFPS[w] [72]; United States	Pilot feasibility trial; pretest-posttest	Children (aged 5-16 years) with moderate to severe TBI[x] and families (all family members could participate; outcomes reported for one parent and child); 19 participants in 6 families; 0%	Child-parent relationship[c]; sibling relationship[c]; therapeutic alliance[c]	(1) Feasibility; (2) ease of use; (3) helpfulness and satisfaction (WEQ[y])	(1) All web sessions completed without therapist assistance; families completed a mean of 10.3 web sessions; (2) ease of use: mean 3.59/5; (3) website helpfulness: mean 4.12/5; videoconferencing helpfulness: mean 4.35/5; 94.7% would recommend the intervention to others
OFPS [73]; United States	Pilot feasibility trial; pretest-posttest	Children (aged 5-16 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child); 19 participants in 6 families; 0%	Injury-related family stress and burden[c]; therapeutic alliance[c]; parental distress, depression, and anxiety[c]; child adjustment[c]	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OFPS [74]; United States	RCT; usual care plus IRC	Children (aged 5-16 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child); 46 families; 12% in the intervention group and 0% in the IRC group	Family problem-solving, communication, and behavior management; parental problem-solving; parental distress, depression, and anxiety[c]	Website use and caregiver satisfaction (WEQ)	100% of parents indicated that they would recommend it to others; 33% indicated that they would prefer to meet in person; 94.4% reported that the website was moderately to extremely easy to use
OFPS [75]; United States	RCT; usual care plus IRC	Children (aged 5-16 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child); 46 families; 12% in the intervention group and 0% in the IRC group	Child adjustment[c] (self-control and compliance only)	(1) Child's self-reported website use; (2) satisfaction (WEQ)	(1) Strong negative correlations between number of sessions completed and child behavioral problems (–0.59) and parental distress (–0.60) at baseline, suggesting families with more problems at baseline completed fewer sessions; (2) 88% rated the website as at least moderately easy to use; 26% rated it as hardly or not easy to use relative to other sites; all children rated the website content as at least moderately helpful; 94% reported feeling at least moderate support and understanding when using the website; 31% reported feeling angry when using the website; 25% reported feeling moderately to extremely worried when using the website

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OFPS [76]; United States	RCT; usual care plus IRC	Children (aged 5-16 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child); 46 families; 12% in the intervention group and 0% in the IRC group	Therapeutic alliance (no moderation by previous technology use); parental depression (moderated by previous technology use) and anxiety	(1) Parents' self-reported website use; (2) satisfaction (WEQ); (3) previous computer use; (4) computer equipment comfort rating	(1) Both groups reported spending equivalent amounts of time on the website; (2) satisfaction did not differ by previous technology use; (3) significant effect of technology at home for improvements in depression (t_{22}=2.24; *P*=.04); trend in the same direction for anxiety; non–technology users more likely to miss sessions (mean 16.33 missed sessions, SD 11.29; t_{18}=2.43; *P*=.03); (4) technology users became more comfortable with the technology over time
OFPS [77]; United States	Pilot feasibility study; pretest-posttest	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 9 families; 0%	Family functioning[c]; adolescent adjustment; parental distress and depression[c]	Feasibility	All families completed the 10 core sessions; 6 families completed one or more supplemental sessions
OFPS [78]; United States	Pilot feasibility study; pretest-posttest	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 9 families; 0%	Not reported	(1) Self-reported website use; (2) satisfaction (WEQ and OSS[z])	(1) In addition to parents and teenagers, 9 siblings participated in at least some of the sessions; (2) father satisfaction was generally high; 4/9 teenagers and 2/7 mothers reported a preference for face-to-face meetings; feedback provided support for acceptability and helpfulness of the intervention

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OFPS [79]; United States	RCT; usual care plus IRC	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 41 families; 20% in the intervention group and 5% in the IRC group	Executive functioning[c] (teenagers with severe TBI)	Not reported	Not reported
OFPS [80]; United States	RCT; usual care plus IRC	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 41 families; 20% in the intervention group and 5% in the IRC group	Family conflict[c]; adolescent adjustment	Self-reported website use and satisfaction	Families completed an average of 10 sessions; 95% completed all 10 sessions; 87% of parents reported meeting their goals, learning ways to improve their child's behavior, and understanding their child better ($P<.05$ relative to IRC)
OFPS [81]; United States	RCT; usual care plus IRC	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 41 families; 20% in the intervention group and 5% in the IRC group	Parental distress and depression[c] (lower SES[aa] only); social problem-solving[c] (lower SES only)	Website use, ease of use, and satisfaction (WEQ and OSS)	93% rated it as moderately or extremely helpful compared to other sites; parents' suggestions for change included fewer questionnaires; 20% of parents agreed that the intervention was too short
OFPS [82]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 12.3% in the intervention group and 5.9% in the control group	Teenager executive function[c] (older adolescents)	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OFPS [83]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 12.3% in the intervention group and 5.9% in the control group	Caregiver depression and distress[c] (intention-to-treat analysis); caregiver efficacy	(1) Previous technology use; (2) completion	(1) Previous computer use did not moderate reductions in depression and distress; nonfrequent computer users in the intervention group reported significantly higher levels of caregiver efficacy (F_{41}=7.15; P=.01); (2) 43% of parents reported spending <30 minutes per week on CAPS[ab]; 50% reported spending 30 minutes-2 hours per week; 88% completed ≥4 sessions
OFPS [84]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 12.3% in the intervention group and 5.9% in the control group	Child behavioral outcomes[c] (older adolescents)	Completion	43% of parents reported spending <30 minutes per week on CAPS; 50% reported spending 30 minutes-2 hours per week; 88% completed ≥4 sessions; 93% rated the website as moderately to extremely helpful
OFPS [85]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 12.3% in the intervention group and 5.9% in the control group	Parent-teenager conflict; parent-teenager interactions; structural, organizational, and transactional characteristics of families	Completion	43% of parents reported spending <30 minutes per week on CAPS; 50% reported spending 30 minutes-2 hours per week; 88% completed ≥4 sessions

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OFPS [86]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 12.3% in the intervention group and 5.9% in the control group (final assessment: 13.4% in the intervention group and 11.4% in the control group)	Long-term caregiver depression and distress[c] (distress only); long-term perceived parenting efficacy	Not reported	Not reported
OFPS [87]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 12.3% in the intervention group and 5.9% in the control group (final assessment: 30.8% in the intervention group and 19.4% in the control group)	Long-term child behavioral outcomes[c] (internalizing behaviors of older adolescents)	Completion	Number of sessions completed unrelated to improvements in internalizing symptoms over time; those who completed more sessions reported less improvement in externalizing symptoms over time (P=.007)
OFPS [88]; United States	RCT; usual care plus IRC	Children (aged 12-17 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 132 children and their families; 25% in the intervention group and 21% in the control group	Adolescent emotional and behavioral functioning; adolescent mood and behavior (as a function of parent marital status)	Not reported	Not reported
OFPS [89]; United States	RCT; face-to-face F-PST[ac], therapist-guided F-PST, or self-guided web-based F-PST	Adolescents (aged 14-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 149 parents and caregivers; 18%	Parent depression[c] (therapist-guided group only); parent psychological distress[c] (therapist-guided group only)	Computer use before and during	Parents with less comfort with technology improved more with therapist-guided treatment when compared to self-guided treatment ($F_{1,107}$=3.80; P=.05)

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OFPS [90]; United States	RCT; face-to-face F-PST, therapist-guided F-PST, or self-guided web-based F-PST	Adolescents (aged 14-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 149 parents and caregivers; at the 9-month assessment: 35.3% in the face-to-face group, 21.5% in the therapist-guided group, and 20% in the self-guided group	Behavioral outcomes	(1) Patient-perceived preference for treatment (before the intervention); (2) adherence; (3) satisfaction; (4) computer use	

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
					(1) 71% of parents agreed or strongly agreed that self-guided F-PST was most convenient; 54% of parents agreed or strongly agreed that self-guided and therapist-guided web-based F-PST would be most beneficial; 55% of teenagers agreed or strongly agreed that self-guided F-PST was most convenient; (2) median 5 hours per week; parents assigned to their preferred group completed a mean of 5.29 sessions, and those assigned to their nonpreferred group completed a mean of 6.37 sessions; adolescents in their preferred group completed a mean of 6.12 sessions, and those in their nonpreferred group completed a mean of 5.17 sessions; adolescent treatment preference was significantly related to attrition (χ^2=4.2, 95% CI 1.03–5.44; *P*=.04); (3) parents in the face-to-face group rated the intervention more favorably than those in the therapist-guided (Cohen *d*=0.67, 95% CI 0.10-1.15; *t*=–2.49; *P*<.04) or self-guided (Cohen *d*=1.18, 95% CI 0.56-1.62; *t*=–4.36; *P*<.001) group;

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
					parents in the face-to-face group reported higher satisfaction than parents in the self-guided group (Cohen *d*=0.63, 95% CI 0.09-1.11; *t*=–2.51; *P*=.04); (4) no significant association with treatment preference
OFPS [91]; United States	RCT; face-to-face F-PST, therapist-guided F-PST, or self-guided web-based F-PST	Adolescents (aged 14-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 149 parents and caregivers; at the 9-month assessment: 35.3% in the face-to-face group, 21.5% in the therapist-guided group, and 20% in the self-guided group	Adolescent QoL[c]; brain injury symptoms	Not reported	Not reported
OFPS [92]; United States	RCT; TOPS[ad] with family, TOPS with teenagers only, or IRC	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 152 teenagers and their families; 31% in the TOPS with family group, 24% in the TOPS with teenagers only group, and 23% in the IRC group	Child behavioral outcomes[c] (TOPS with family)	Completion	Completion: mean sessions completed (TOPS with family) 8.00 (SD 2.90) and mean sessions completed (TOPS with teenagers only) 8.40 (SD 2.80); completed supplemental sessions: 14.29% for TOPS with family and 13.46% for TOPS with teenagers only
OFPS [93]; United States	RCT; TOPS with family, TOPS with teenagers only, or IRC	Teenagers (aged 11-18 years) with moderate to severe TBI and families (all family members could participate; outcomes reported for one parent and child only); 152 teenagers and their families; 31% in the TOPS with family group, 24% in the TOPS with teenagers only group, and 23% in the IRC group	Family functioning; family cohesion[c] (TOPS with family and 2-parent households); parent-adolescent conflict; parental psychological distress and depression[c] (TOPS with family and 2-parent households)	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship [94]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Relationship satisfaction[c]; positive and negative relationship quality[c] (reducing negative relationship quality); relationship confidence[c]; depression[c]; anxiety[c]; perceived health[c]; work functioning[c]; QoL[c]	(1) Evaluation (Client Evaluation of Services Questionnaire); (2) completion rates; (3) coach engagement	(1) Mean 26.81 (SD 4.44), nearly equivalent to in-person individual therapy (Cohen *d*=–0.07) and high-quality couple therapy (Cohen *d*=–0.18); 94% were mostly or very satisfied with the services received; 97% would recommend it to a friend; (2) 86% completed the entire intervention; an additional 5% completed up to the "Understand" phase; (3) coaches spent a mean of 51.32 (SD 17.11) minutes with the couples; individuals received a mean of 5.11 (SD 1.7) scripted chat reminders and no tailored chat messages
OurRelationship [95]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Relationship satisfaction[c] (no moderation by LIIPV[ae])	Not reported	Not reported
OurRelationship [96]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Relationship satisfaction; relationship confidence; positive and negative relationship quality[c] (moderated by rurality); depression; anxiety; perceived health[c] (moderated by race); work functioning; QoL	(1) Evaluation (Client Evaluation of Services Questionnaire); (2) participant predictors of completion	(1) Couples were generally satisfied with the intervention (mean 26.81, SD 4.44); service evaluation was not moderated by race, ethnicity, income, educational level, or rural status; (2) Hispanic couples (OR[af] 0.24; *P*=.009; Cohen *d*=0.79) and low-income couples (OR 0.21; *P*=.002; Cohen *d*=0.85) were more likely to drop out

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship [97]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Long term: relationship satisfaction; positive[c] and negative relationship quality; relationship confidence[c] (Hispanic couples); depression[c]; anxiety[c]; perceived health[c]; work functioning[c]; QoL[c]	Not reported	Not reported
OurRelationship [98]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Relationship satisfaction; coparenting conflict[c] (not maintained at follow-up); child functioning[c]	Not reported	Not reported
OurRelationship [99]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Relationship satisfaction[c]; communication[c]; emotional intimacy[c]; relationship problem confidence[c]; relationship problem acceptance[c]; self-protective orientation[c]	Not reported	Not reported
OurRelationship [100]; United States	RCT; waitlist	Heterosexual couples; 300 couples; 8%	Relationship satisfaction[c] (moderated by neuroticism); relationship confidence[c]; depression (moderated by neuroticism and conscientiousness); personality	Not reported	Not reported
OurRelationship [101]; United States	RCT; low coach support or high coach support	Heterosexual couples; 356 couples; 34% in the group with high coach support and 64% in the group with low coach support	Relationship satisfaction[c] (both groups); depression[c] (both groups); anxiety[c] (both groups; significantly greater in the high-support group)	Platform predictors of completion	Participants in the high-support group were significantly more likely to complete the entire intervention (66% vs 36%; χ^2_1=32.8, *P*<.001); participants in the high-support group were more likely to complete two-thirds of the intervention (69% vs 45%; χ^2_1=20.4, *P*<.001); no significant differences in first phase completion; completion did not differ by race, ethnicity, or household income

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship [102]; United States	RCT; low coach support, high coach support, or no coach support	Heterosexual couples; 529 couples; 93.9% in the group with no coach support, 34% in the group with high support, and 64% in the group with low support	Relationship satisfaction; relationship confidence; depression; anxiety	Platform predictors of completion	6.1% of participants in the group with no coach, 66.1% of participants in the high-support group, and 36% of participants in the low-support group completed the intervention; substantial and immediate dropout when compared with the high-support (b=–2.68; SE 0.35; t=–7.65; OR 0.07, 95% CI 0.04-0.14; P<.001) and low-support (b=–1.98; SE 0.34; t=–5.76; OR 0.14, 95% CI 0.07-0.27; P<.001; neither was significant) groups; Hispanic individuals were less likely to complete the intervention without a coach than non-Hispanic individuals (b=–3.99; P<.001); higher levels of depressive symptoms predicted less drop-off with no coach (b=0.08; P=.04)
OurRelationship [103]; United States	RCT; brief OurRelationship with coach, brief OurRelationship without coach, or waitlist	Heterosexual couples; 104 couples; 40.4% at midintervention, 25% at end of intervention, and 17.4% at follow-up in the arm with a coach and 56% at midintervention, 26% at end of intervention, and 26% at follow-up in the arm without a coach	Relationship satisfaction; positive and negative relationship quality[c] (positives); relationship confidence; communication; anxiety; depression; perceived health and QoL; work functioning	Platform predictors of completion	Dropout rate was 9.3% for the full OurRelationship and 28.8% for the brief OurRelationship with a coach (χ^2=12.1; P<.001); 71.2% completion in the coach condition and 42.3% completion in the no-coach condition (χ^2=8.8; P=.003)

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship [104]; United States	Pilot; pretest-posttest	Veterans and their partners; 13 couples; 15%	Relationship satisfaction and distress; relationship conflict; depression symptoms; probable PTSD; QoL	(1) Intervention satisfaction (CSQ-8); (2) completion	(1) Mean (veterans) 3.4/4 (SD 0.4) and mean (partners) 3.2/4 (SD 0.6); 91% were mostly or very satisfied; 96% would recommend it; positive qualitative feedback included structure, videos of similar couples, and reminder calls; negative qualitative feedback included repetition, length of some content, and technical and logistic frustrations; couples preferred the coach calls; (2) completion rate was 85%; median completion time was 52 (range 29-73) days; couples received clinical contact ranging from 52 to 95 minutes in total
OurRelationship [105]; United States	Single arm; pretest-posttest	Coparenting couples; 136 couples; 20%	Relationship satisfaction and distress[c]; coparenting satisfaction[c]; gatekeeping and gate closing behaviors[c]; perception of partner's gatekeeping and gate closing behaviors[c]	Not reported	Not reported
OurRelationship [106]; United States	RCT; OurRelationship or OurRelationship+ with greater therapist engagement	Couples; 314 couples; 64.3%	Individual use, joint use and perception of partner's [c] pornography use; arguments surrounding self-, joint, and partner's pornography consumption; individual pornography use[c]; problematic pornography use; lifestyle changes due to the COVID-19 pandemic	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship [107]; United States	RCT; full coach, automated coach, contingent coach, or waitlist	Couples; 740 couples; 30%	Relationship satisfaction[c] (comparable across all types of coach support)	Completion	Completion comparable across all conditions; posterior distributions indicated that the probability of full-coach couples having higher odds of completing phases 1, 2, and 3 relative to automated-coach couples was 28.4%, 43.9%, and 77.4%, respectively; probability of full-coach couples having higher odds of completing phases 1, 2, and 3 relative to contingent-coach couples was 65%, 70%, and 92.7%, respectively; probability of contingent-coach couples having higher odds of completing phases 1, 2, and 3 relative to automated-coach couples was 15.6%, 22.6%, and 21.7%, respectively
OurRelationship and ePREP [108]; United States	RCT; OurRelationship, ePREP, or waitlist	Romantic couples; 742 couples; 10.3% at posttreatment time point, 12.5% at 2-month follow-up, and 13% at 4-month follow-up	Relationship satisfaction[c]; communication conflict[c]; emotional support[c]; intimate partner violence[c]; breakup potential[c]	(1) Evaluation (Client Evaluation of Services Questionnaire); (2) completion	(1) Participants rated the intervention positively (mean 9.9/11); 96% would recommend it to a friend; 93% were satisfied; no significant difference between OurRelationship and ePREP in satisfaction (b=–0.058; SE 0.148; P=.70); (2) 69% in both ePREP and OurRelationship completed all content

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship and ePREP [109]; United States	RCT; OurRelationship, ePREP or waitlist	Romantic couples; 742 couples; 10.3% at posttreatment time point, 12.5% at 2-month follow-up, and 13% at 4-month follow-up	Relationship satisfaction[c]; breakup potential[c]; negative communication[c]; positive communication[c]; relationship problem intensity[c]; relationship problem confidence[c]; emotional support[c]	Not reported	Not reported
OurRelationship and ePREP [110]; United States	RCT; OurRelationship, ePREP, or waitlist	Romantic couples; 742 couples; 10.3% at posttreatment time point, 12.5% at 2-month follow-up, and 13% at 4-month follow-up	Psychological distress[c]; perceived stress[c]; anger[c]; problematic alcohol use[c]; perceived health[c]; insomnia[c]; exercise[c]	Not reported	Not reported
OurRelationship and ePREP [111]; United States	RCT; OurRelationship, ePREP, or waitlist	Romantic couples; 742 couples; 10.3% at posttreatment time point, 12.5% at 2-month follow-up, and 13% at 4-month follow-up	Cooperative parenting; parenting stress; parenting nurturance[c] (OurRelationship); physical and harsh verbal discipline[c] (OurRelationship)	Not reported	Not reported
OurRelationship and ePREP [112]; United States	RCT; OurRelationship, ePREP, or waitlist	Romantic couples; 742 couples; 10.3% at posttreatment time point, 12.5% at 2-month follow-up, 13% at 4-month follow-up, and 18.6% at 12-month follow-up	Long term: relationship satisfaction[c]; breakup potential[c]; positive communication[c]; communication conflict[c]; emotional support[c]; intimate partner violence; psychological distress[c]; perceived stress[c]; anger[c]; alcohol use[c]; perceived health[c]; insomnia[c]	Not reported	Not reported
OurRelationship and ePREP [113]; United States	RCT; OurRelationship, ePREP, or waitlist	Military and nonmilitary couples; 90 military couples; 43% for military couples	Relationship satisfaction[c]; communication conflict[c]; emotional support[c]; breakup potential[c]; intimate partner violence; psychological distress; perceived stress; anger; substance use; perceived health	(1) Evaluation (Client Evaluation of Services Questionnaire); (2) completion	(1) Evaluation ratings were similarly positive (b=0.470; P=.07); (2) 57% of military couples completed the entire intervention (compared with 71% of civilian couples), 8% completed two-thirds, 18% completed one-third, and 18% completed none

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
OurRelationship and ePREP [114]; United States	RCT; OurRelationship, ePREP, or waitlist	Low-income couples; 671 couples; 36% for OurRelationship and 31% for ePREP	Relationship satisfaction[c]; communication conflict[c]; emotional support[c]; intimate partner violence; breakup potential[c] (not maintained long term for ePREP)	(1) Evaluation (Client Evaluation of Services Questionnaire); (2) completion	(1) Participants' satisfaction: mean (OurRelationship) 9.51/11; mean (ePREP) 9.6/11; >95% of participants indicated that the intervention helped them; 97% indicated that they would recommend the intervention; 90% were satisfied with the intervention; no reliable differences in satisfaction between the 2 interventions (*B*=0.07, 95% CI –0.07 to 0.21); (2) 64% completed OurRelationship, and 69% completed ePREP
OurRelationship and ePREP [115]; United States	RCT, OurRelationship, ePREP, or waitlist	Low-income perinatal couples; 180 couples; 32.8% for OurRelationship and 36.1% for ePREP	Relationship satisfaction[c]; perceived likelihood of breakup[c]; communication conflict[c]; sexual intimacy[c]; emotional support[c]; experience of intimate partner violence; psychological distress[c]; perceived stress[c] (OurRelationship only)	Not reported	Not reported
OurRelationship and ePREP [116]; United States	RCT; OurRelationship, ePREP, or waitlist	Low-income couples; 659 couples; 16.8%	Relationship satisfaction[c]	Not reported	Not reported
OurRelationship and ePREP [117]; United States	RCT; OurRelationship, ePREP, or waitlist	Low-income couples; 615 couples; not reported	Perceived gratitude from partner[c]; relationship satisfaction[c]; relationship instability[c]; communication skills[c]; destructive communication[c]; partner emotional support[c]	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
ParentSTRONG [118]; United States	RCT; waitlist	Early adolescent male individuals and a parent or guardian; 119 dyads; 8.5%	Dating violence behaviors[c]; parent-child communication[c]; attitudes supporting dating violence; aggression; emotional regulation[c]	Acceptability and fidelity	90% of families completed all 6 modules; 87% of parents rated helpfulness as >4/5, and 99% of parents rated helpfulness as >3/5; 65% of teenagers rated helpfulness as >4/5, and 96% of teenagers rated helpfulness as >3/5; intervention did not allow participants to progress without completing all activities
PACT[ag] [119] Australia	RCT; waitlist	Parent-child dyads in which the child (aged 2-10 years) had cerebral palsy; 67 dyads; 24.4%	Emotional availability[c]; child involvement[c]; QoL[c]; parental mindfulness[c]; parental acceptance[c]; adjustment	Not reported	Not reported
ParentWorks [120]; Australia	Single arm; pretest-posttest measures	Parent or caregiver of a child aged 2-16 years; 388 families; 92.7% (nonstarters included)	Dysfunctional parenting[c]; interparental conflict[c]; child behavioral difficulties[c]; parental mental health[c]	Satisfaction (CSQ)	Mean 5.49 (SD 0.95); no significant sex differences (t_{452}=0.41; P>.05), indicating that mothers and fathers were equally satisfied

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
ParentWorks [121]; Australia	Single arm; pretest-posttest measures	Parent or caregiver of a child aged 2-16 years; 388 families; 92.7% (nonstarters included)	Parent and family functioning; parenting conflict; child behavioral difficulties; parental mental health	(1) Completion; (2) dropout characteristics; (3) participant predictors of completion	(1) For partial completers, mean 2.4/5 (SD 1.2) modules completed; for full completers, mean 5.58/6 (SD 0.76) modules completed (including 1 optional module); (2) mothers in the full completer and partial completer groups reported higher levels of conduct problems than nonstarters $F_{2,1749}$=3.99; P<.05); (3) relative to full completers, nonstarters were more likely to have older children, be married or in a de facto relationship, have higher levels of psychological difficulties, and have lower levels of child conduct problems; relative to full completers, partial completers were more likely to be married or in a de facto relationship and have higher levels of dysfunctional parenting

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
PERC[ah] [122]; United States	Single arm; pretest-posttest	Couples where one member had a prostate cancer diagnosis; 26 couples; 15%	Dyadic communication; relationship satisfaction; QoL[c]; symptom distress[c]; general symptoms[c]	(1) Feasibility and acceptability; (2) web activity; (3) ease of use	(1) 96% completed the intervention; (2) 37% of couples always logged in together, and 23% always logged in individually; mean 3.64 (SD 1.68) log-ins per couple; mean time spent on the platform per couple: 56.96 (SD 39.74) minutes; 83% used audio-enhanced slides; 94% visited the assignment and exercise section; (3) participants rated PERC as easy to use, engaging, and of high quality
Resilient Living [123]; the Netherlands	Pilot feasibility trial	Patients with stroke or brain tumor and their caregivers; 16 participants; 68.75%	Dyadic coping; resilience; stress; caregiver role overload; QoL; fatigue[c]; physical function[c]; anxiety[c]; sleep	(1) Intervention evaluation; (2) WiWi[ai]	(1) Mean 2.6/5 for "Do you think the skills you learned enhanced your resilience?" and mean 4.4/5 for "did you find the online intervention easy to use?"; remaining mean scores ranged between 3.3 and 4.2/5; length of modules and ability to complete them in their own time were identified as facilitators to use; finding time to complete them as a dyad was challenging; (2) 4/5 indicated that it was worthwhile participating in the study, 4/5 indicated that it was as expected, and 1 indicated it was better than expected
Web-based partnership support program [124]; Japan	Quasi-experimental design (nonrandomized); control	Infertile couples; 151 couples; 20.4%	QoL[c]; distress	Not reported	Not reported

XSL•FO
RenderX

Platform, study, and country	Study characteristics		Measures or outcomes		
	Design; comparator	Population; sample size; attrition	Relational and individual	User engagement indicators	Reported findings
Web-based PREP program [125]; United States	RCT; IRC	Heterosexual foster or adoptive couples; 32 couples; 35%	Negative communication; knowledge acquisition[c]; use of PREP skills[c]	Intervention feedback	Participants responded favorably to the intervention

[a]4Cs:CRC: Caring for Couples Coping With Colorectal Cancer.
[b]RCT: randomized controlled trial.
[c]Indicates significance, or that the intervention was superior to the comparator, at the postintervention time point for the outcome measure.
[d]CA-CIFFTA: Computer-Assisted, Culturally Informed, and Flexible Family-Based Treatment for Adolescents.
[e]C-MBI: couple mindfulness-based intervention.
[f]YBCS: young breast cancer survivor.
[g]MBI: mindfulness-based intervention.
[h]I-MBI: mindfulness-based intervention for individuals.
[i]HOPES: Helping Overcome Posttraumatic Stress Disorder and Enhance Satisfaction.
[j]PTSD: posttraumatic stress disorder.
[k]QoL: quality of life.
[l]AOD: alcohol and other drug.
[m]CSQ: Client Satisfaction Questionnaire.
[n]TSQ: Treatment Satisfaction Questionnaire.
[o]eMB: mothers and babies online course.
[p]CSQ-8: 8-item Client Satisfaction Questionnaire.
[q]ePREP: computer-based Prevention and Relationship Enhancement Program.
[r]IRC: internet resource comparison.
[s]FOCUS: family involvement, optimistic outlook, coping effectiveness, uncertainty reduction, and symptom management.
[t]iCBT: internet-delivered cognitive behavioral therapy.
[u]MR: Mission Reconnect.
[v]PREP: Prevention and Relationship Enhancement Program.
[w]OFPS: Online Family Problem-Solving Therapy.
[x]TBI: traumatic brain injury.
[y]WEQ: Website Evaluation Questionnaire.
[z]OSS: Online Satisfaction Survey.
[aa]SES: socioeconomic status.
[ab]CAPS: counselor-assisted problem-solving.
[ac]F-PST: family-problem-solving therapy.
[ad]TOPS: teen online problem-solving.
[ae]LI-IPV: low-intensity intimate partner violence.
[af]OR: odds ratio.
[ag]PACT: Parenting Acceptance and Commitment Therapy.
[ah]PERC: Prostate Cancer Education and Resources for Couples.
[ai]WiWi: Was It Worth It questionnaire.

The Platforms

Overview

A total of 24 unique platforms were identified from the 85 studies. Table 2 shows the characteristics of the 24 platforms, including the intervention target; relationship targeted; duration of intervention participation; components designed for cocompletion, individual completion, and therapist engagement; any tailoring offered; and additional reported features.

Most interventions (14/24, 58%) were designed for cocompletion by couples, with some identified interventions for parent-child dyads (6/24, 25%), families (2/24, 8%), and caregiver–care recipient dyads (2/24, 8%). Given that it was expected that build characteristics might differ according to the population (eg, number of participating family members and their ages), platform results are grouped and reported by the relationship structure targeted by the platform (ie, couples, parent-child dyads, families, and caregiver–care recipient dyads).

Data from Table 2 are synthesized based on the features of the platforms and detail reported user engagement indicators. As platforms were included only in cases in which at least one study had demonstrated clinical efficacy of the intervention, mental health and relational outcomes are not reported in this table (and are, instead, indicated in Table 1).

XSL•FO
RenderX

Table 2. Characteristics of the platforms identified in the included studies, including name of the platform, relationship between the participants, platform purpose, duration of intervention participation, components that were completed in a self-paced manner, components completed together (cocompletion) or by individuals alone, practitioner engagement components, any tailoring provided, and additional key features.

Platform	Target relationship; intervention target; intervention duration	Self-paced components	Cocompletion versus individual completion	Practitioner engagement components	Tailored platform components and additional key features
4Cs:CRC[a]	Couples; patient–partner coping with cancer; 6 weeks	6 intervention sections including dyadic learning sessions, health information, cancer news, web-based counseling, sharing circle, and personal center	Content intended to be completed by couples together	Face-to-face or web-based synchronous counseling sessions delivered biweekly to revisit content and provide additional support (some study conditions)	Weekly reminders to complete web-based sessions
CA-CIFFTA[b]	Parent-child; treat behavioral problems and family conflict in young minority adolescents and their families; 12 weeks	4-6 computer-based modules; links to academic websites	Parents watched videos independently first, then rewatched with the adolescent; individual log-ins; role-appropriate videos	6-10 face-to-face sessions; fortnightly phone calls; asynchronous communication	Modular format for families to select content most relevant to the family's clinical and cultural needs and preferences; custom links
Cool Kids Online	Parent-child; psychoeducation and CBT[c]-based anxiety management skills for children and their parents; 10 weeks	8 web-based lessons—first 6 released weekly and final 2 released biweekly	Web-based lessons completed together; parent trained as a "coach" for their child; additional web-based information provided to caregivers at the end of each lesson	Parents completed weekly phone calls with clinician—reinforce success, clarify questions, assist with barriers and skill implementation, reinforce practice, and normalize experience	Automated reminder emails—emails reinforced content, skill practice, and engagement
C-MBI[d] for YBCSs[e]	Couples; relationship distress for couples where one member is a breast cancer survivor; 8 weeks	8 weekly, prerecorded videos delivered via the web; video links and reminders emailed to participants weekly	All videos watched together	Participants encouraged to email or call research staff regarding questions or content during participation	None
Couple HOPES[f]	Couples; relationship functioning when one partner has PTSD[g]; 8 weeks	7 web-based modules containing videos, exercises, and practice assignments completed sequentially	Videos and module exercises completed together; partners had separate, linked accounts where they independently completed assignments; assignment entries and scores could be seen by both partners	4 scheduled calls with a coach after modules 1, 3, 5, and 7 plus 1 additional call as needed; engagement and adherence facilitated through platform messaging; coaches' role involved reviewing symptom change, reinforcing successes, enhancing motivation for engagement, and troubleshooting barriers	Automated feedback graph depicted reported symptom change over time; progress bar and module menu communicated and incentivized progress; web-based application and smartphone app
Couplelinks	Couples; relationship functioning after cancer diagnosis; 8 weeks	6 modules; each module begins with an informational component followed by instructions for interactive exercises; couples reflect after each module; additional articles and video resources available	Modules completed together	Asynchronous platform-based messaging; introductory telephone call and 2 brief "check-ins" to reinforce alliance and promote adherence	Additional noncompulsory content

XSL•FO
RenderX

Platform	Target relationship; intervention target; intervention duration	Self-paced components	Cocompletion versus individual completion	Practitioner engagement components	Tailored platform components and additional key features
eMB[h]	Couples; increase partner's understanding of perinatal mood and anxiety disorders and therapeutic approaches to managing associated symptoms; 8 weeks	Recommended completion of 1 lesson per week in any order, with revisits as needed; psychoeducational modules containing YouTube videos, vignettes, interactive quizzes, homework, guided meditation, and downloadable resources	Participants could choose whether to complete separately or together	None	Could be completed in any order
Embers the Dragon	Parent-child; supporting emotional development and parental responses to child behavior; 8 weeks	Two 6-minute animated episodes and accompanying videos and activities	Parent and child watch the episodes and complete postvideo activities together; following the episodes, parents watch explanation videos	None	None
ePREP[i]	Couples; preventative intervention to enhance relationship satisfaction and mental health; 6 weeks	6 hours of web-based modules and approximately 1-2 hours of homework	Couples completed modules and homework together	Four 15-minute appointments with coach practicing skills; weekly reminder emails to complete content and links to resources	Computer based, could be completed from mobile or tablet
FOCUS[j]	Caregiver–care recipient (family); psychosocial health of patients with cancer and their family caregivers; 6 weeks	3 sessions delivered sequentially, with time to practice skills learned in between	Dyads completed the sessions together	Asynchronous "help" function that generated an email to the project director	Tailored, app-generated messages provided web links addressing the dyad's specific concerns; offered a choice of tailored activities to complete between web sessions; tailoring based on baseline information provided
iCBT[k]	Families; family functioning when a child has an anxiety disorder diagnosis; 10 weeks	11 modules, including reading materials, film, animations, and illustrations	Parents worked on their modules first so that they could then work with the children; 7 modules aimed at parents only	Platform-based messages; tailored feedback after exercise completion; 3 telephone calls during treatment and additional ones as needed to clarify content, increase motivation, and solve problems	None
Military Family Foundations	Couples; military couples in the transition to parenthood; not specified	5 prenatal and 3 postnatal modules	Modules completed together	Email reminders sent to couples if they stopped engaging for >10 days	None
MindGuide Couples	Couples; preventative intervention centered on vulnerability to Korean middle adulthood depression, "Hwa-Byung," and couple relationships; 5-7 weeks	4 modules over 16 sessions, maximum 60 minutes each; sessions included audio-recorded mindfulness activities, video lectures, practical tasks, and case-based scenarios	Modules 3 and 4 were joint sessions, including creating a shared vision; performed practical tasks together; modules 1 and 2 were completed individually	Coaching sessions after each module to promote participation via reflective dialogue and provide feedback on participants' responses	None

XSL•FO
RenderX

Platform	Target relationship; intervention target; intervention duration	Self-paced components	Cocompletion versus individual completion	Practitioner engagement components	Tailored platform components and additional key features
MR[l]	Couples; relationship functioning when a member is a veteran with a history of deployment in a post-9/11 combat operation; 16 weeks	11 activities delivered via instructional videos, guided audio, and written manuals	Sessions on "Connecting with Partner" could be completed alone or together; the remaining sessions were completed independently	None	Accessible through website and mobile apps
Mother-daughter program	Parent-child; mother-daughter relationship quality and reduced risk of underage drinking; 10 weeks (4 weeks for the brief version)	9-14 modules; different adaptations were developed; animated characters portrayed the adolescent girl and her mother	Modules completed together; participants independently logged in to complete questions about content; participants could not advance until both mother and daughter had completed this	None	None
OFPS[m] (including CAPS[n] and TOPS[o])	Families; family functioning when the child, adolescent, or teenager has a TBI[p]; 6 months	7-11 sessions; core sessions and additional supplementary sessions provided based on identified need; web-based content included problem-solving skills, video clips, exercises, and assignments	Website used by multiple family members together	Initial face-to-face session completed in the family's home; telehealth session following web-based sessions to review exercises	Supplementary sessions provided based on personal need; family members selected their picture to indicate that they were present; when required, the platform would prompt particular family members to respond, and other times, the whole family was asked to respond together
OR[q]	Couples; relationship distress; 6 weeks (brief OR=2 weeks)	3 sections including video examples and psychoeducation	Content completed separately; couple completed guided conversation together at the end of each section	4 phone calls during the intervention; asynchronous chat feature	Tailored report on improvement provided; in some studies, automated tailored emails were provided
ParentSTRONG	Parent-child; adolescent boy domestic violence prevention intervention; 4 weeks	6 modules comprising 4-6 activities; parents and teenagers progress through alternate reality as avatars	After module 1, all modules are completed by the parent and child together; module 1 (introduction) completed by parents only	Staff could be contacted to troubleshoot technology	None
PACT[r]	Parent-child; emotional availability and parent and child adjustment when the child has cerebral palsy; 10 weeks (enforced break in the middle)	3 modules and a final review module after a short break	Some exercises were designed for individual completion	Fortnightly check-in (phone, SMS text message, or email) to monitor completion and check understanding of content	None
ParentWorks	Couples; father-inclusive parenting intervention; 4 weeks	5-8 modules	Participants accessed via a shared account; participants had the option to complete it independently	None	Feedback provided based on participant responses; formatted for mobile, laptop, and tablet viewing
PERC[s]	Couples; relationship distress following a prostate cancer diagnosis; 8 weeks	7 modules—5 core and 2 optional	Encouraged to view and complete everything together	None	Optional modules; users could select text- or audio-based slides depending on preference

XSL•FO
RenderX

Platform	Target relationship; intervention target; intervention duration	Self-paced components	Cocompletion versus individual completion	Practitioner engagement components	Tailored platform components and additional key features
Resilient Living	Caregiver–care recipient (family); building dyadic resilience skills for patients with stroke or brain tumor and their family caregivers; 8 weeks	4 web-based video modules and participant journal	Option to complete individually	Telehealth session before commencement of web-based modules	None
Web-based partnership support program	Couples; support intervention to prevent quality of life deterioration and reduce emotional distress in men undergoing fertility treatment; 2 weeks	30-minute self-paced content over 10 days	Watched information together; discussion between couples using the communication form; couples individually completed their communication form, which was subsequently used to guide their discussion	None	None
Web-based PREP[t] program	Couples; couple relationship education for foster or adoptive parents; 1 week	4 chapters plus additional resources	Entire intervention completed together	None	None

[a]4Cs:CRC: Caring for Couples Coping With Colorectal Cancer.
[b]CA-CIFFTA: Computer-Assisted, Culturally Informed, and Flexible Family-Based Treatment for Adolescents.
[c]CBT: cognitive behavioral therapy.
[d]C-MBI: couple mindfulness-based intervention.
[e]YBCS: young breast cancer survivor.
[f]HOPES: Helping Overcome Posttraumatic Stress Disorder and Enhance Satisfaction.
[g]PTSD: posttraumatic stress disorder.
[h]eMB: mothers and babies online course.
[i]ePREP: computer-based Prevention and Relationship Enhancement Program.
[j]FOCUS: family involvement, optimistic outlook, coping effectiveness, uncertainty reduction, and symptom management.
[k]iCBT: internet-delivered cognitive behavioral therapy.
[l]MR: Mission Reconnect.
[m]OFPS: Online Family Problem-Solving Therapy.
[n]CAPS: counselor-assisted problem-solving.
[o]TOPS: teen online problem-solving.
[p]TBI: traumatic brain injury.
[q]OR: OurRelationship.
[r]PACT: Parenting Acceptance and Commitment Therapy.
[s]PERC: Prostate Cancer Education and Resources for Couples.
[t]PREP: Prevention and Relationship Enhancement Program.

Couples

Features of Platforms for Couples

Overview

Of the platforms requiring cocompletion, platforms designed for couples were the most common. A total of 58% (14/24) of the identified platforms were for couples. The intervention targets included relationship distress when a member has a cancer diagnosis (2/14, 14%); relationship functioning when a member has a cancer diagnosis (2/14, 14%), has posttraumatic stress disorder (1/14, 7%), or is a veteran (1/14, 7%); parenting-focused interventions, including a father-inclusive parenting intervention (1/14, 7%), education for foster and adoptive parents (1/14, 7%), and an intervention for military couples transitioning to parenthood (1/14, 7%); partnership support interventions for cases in which the male partner is undergoing treatment for infertility (1/14, 7%) or a member is pregnant (1/14, 7%); general relational distress (1/14, 7%); and preventative interventions to enhance relationship satisfaction and mental health (1/14, 7%) and reduce vulnerability to middle adulthood depression (1/14, 7%).

Structure and Duration of Engagement

Duration of participation varied from 1 to 16 weeks, with the most common duration being 8 weeks (5/14, 36%) followed by 6 weeks (4/14, 29%), including 1 intervention described as taking 5 to 7 weeks. The intended duration of 7% (1/14) of the

XSL•FO
RenderX

interventions was not specified. One intervention offered a brief version that was completed by couples in 2 weeks as opposed to the 6-week full version. As per the inclusion criteria for this review, all interventions involved some web-based self-paced component completed on the platform. Most appeared to require at least weekly engagement, although it was not always specified or prescribed. One platform was designed such that participants could complete the intervention modules in any order but advised participants to access 1 module per week and complete all modules. For all the remaining interventions, it appeared that intervention content or modules were designed to be completed in a defined order and over a specified period.

Coparticipation and Contact With Practitioners

A total of 43% (6/14) of the interventions contained elements that were intended for individual completion (ranging from completing assessments to completion of entire sections of content), 50% (7/14) of the interventions required couples to cocomplete the whole intervention, and 7% (1/14) of the interventions gave participants the choice to complete some or all of the intervention together. In total, 57% (8/14) of the interventions included an element of practitioner engagement, including asynchronous platform-based messaging or scheduled synchronous counseling sessions.

Tailoring and Additional Features

Beyond personalization through contact with practitioners, 29% (4/14) of the platforms provided tailored content or options for personalization. A total of 14% (2/14) of the platforms provided supplementary content that could be accessed based on need, and 14% (2/14) of the platforms provided personalized feedback and reporting based on responses to questionnaires. In total, 29% (4/14) of the platforms specified that they were formatted for both web and mobile or tablet use, and 7% (1/14) of the platforms allowed participants to select either audio-enhanced or text-based presentation of content. Finally, 7% (1/14) of the platforms included an automated graph depicting reported symptom change over time and a progress bar to incentivize participation.

Reported User Engagement Indicators of Platforms for Couples

A total of 56% (48/85) of the studies examined the 14 couple-focused platforms. Of those 48 studies, 30 (62%) reported on user engagement indicators, including 23 (77%) studies that reported on satisfaction, feedback, usability, participant evaluation, feasibility, and acceptability and 18 (60%) that reported on completion rates and website use. The remaining 38% (18/48) of the studies did not report on any user engagement data or findings.

Measures used to collect participant satisfaction, feedback, usability, and evaluation varied. A total of 10% (5/48) of the studies administered the Client Evaluation of Services Questionnaire [126], and 15% (7/48) used the Client Satisfaction Questionnaire [127]. The remaining studies reported on satisfaction, feedback, and participant evaluation through nonvalidated measures. Satisfaction ratings were generally high across all studies.

The impact of video content on user engagement appeared mixed. Participants in 8% (4/48) of the studies provided feedback that the content and examples presented in the videos were helpful; however, in another 4% (2/48) of the studies, participants reported that the videos were unhelpful or that they negatively impacted engagement as they were not relatable, overly dramatized, or appeared outdated. In addition, participant qualitative feedback reported in another study suggested that outdated imagery and low-technology visualizations also negatively impacted engagement. Other factors that were reported to be important based on qualitative feedback included one study that reported on the structured nature of the intervention and reminder calls and another where participants reported that they were more likely to access audio-enhanced slides than text-based content. Feedback provided by participants in one study also noted that the flexibility of the web-based format facilitated engagement. However, in general, satisfaction, feedback, usability, and evaluation data were reported as average values on rating scales.

Reporting of completion rates and website use rates varied. They were reported as combinations of the following: the average number of participants who completed the entire intervention, the average number of modules or sessions completed by individuals or couples, the average time to completion, the number of discrete log-ins or page views, and the amount of time spent accessing the platform. Feasibility and acceptability data were reported similarly, with completion statistics often used as an indication of an intervention's feasibility or acceptability. In addition, 5 studies reported on predictors of noncompletion, including 3 (60%) studies that reported higher levels of support from a practitioner as predictors of completion. The remaining 40% (2/5) of the studies reported on participant baseline characteristics as predictors of noncompletion.

Finally, 4% (2/48) of the studies on the same platform identified different couple "types" with regard to their enthusiasm and engagement (eg, "keen completers" or "stragglers") and therapists' role in engagement promotion. One study reported that higher levels of engagement (measured using participants' correct responses to quiz questions) led to greater intervention effect on a number of clinical outcomes, and another found that those with the shortest time frame between commencement and completion (ie, completed the intervention faster) were more likely to be classified as "treatment responders" (identified by significant improvement on outcomes) at the postintervention assessment.

No studies of couple-based platforms identified build or design characteristics as moderators of intervention effect. No studies performed a formative evaluation of the platforms, and no studies reported design and build characteristics that enabled coparticipation beyond participant qualitative feedback.

Parent-Child Dyads

Features of Platforms for Parent-Child Dyads

Overview

Platforms designed for co-use by parent-child dyads were the second most common, accounting for 25% (6/24) of the

XSL•FO
RenderX

platforms identified in this review. The intervention targets included behavioral problems and conflict in young minority adolescents and their families (1/6, 17%), emotional development and parental responses to child behavior (1/6, 17%), mother-daughter relationship quality and risk of underage drinking (1/6, 17%), adolescent male domestic violence prevention (1/6, 17%), emotional availability and parent-child adjustment when a child has cerebral palsy (1/6, 17%), and anxiety management skills and psychoeducation for parents and children (1/6, 17%). A total of 67% (4/6) of the platforms were developed for adolescents and a parent, 17% (1/6) were for young children aged 2 to 7 years and a parent, and 17% (1/6) were for children aged 7 to 12 years and their parents. In all cases, only 1 parent was asked to participate. In the following sections, we summarize the reported features of the platforms as detailed in the included studies.

Structure and Duration of Engagement

Duration of intervention use varied from 4 to 12 weeks, with the most common duration being 10 weeks (3/6, 50%). One intervention of a 10-week duration in total enforced an extended break in the middle of intervention engagement, and another offered a brief version of only 4 weeks (compared with the 10-week full version). Participation varied from once a fortnight to 2 web-based sessions or modules a week. A total of 83% (5/6) of the interventions appeared to involve completion in a structured manner following a predetermined order. One platform presented intervention content in a modular format that allowed participants to select the content that was relevant to their needs and cultural preferences in any order. However, the intended duration and number of modules accessed appeared to be prescribed.

Coparticipation and Contact With Practitioners

The amount and method of coparticipation varied greatly. A total of 33% (2/6) of the platforms required parents to watch the intervention content or preparatory materials before engaging with their adolescent child. In total, 33% (2/6) of the platforms required the parent to complete explanation videos or additional content following cocompletion with their young child. A total of 17% (1/6) of the platforms involved cocompletion of all intervention modules and independent completion of questions about content, with both the parent and adolescent required to complete these questions before the dyad could progress to the next module. Finally, studies on 17% (1/6) of the platforms reported that “some exercises” were designed for cocompletion but did not specify the extent of cocompletion.

A total of 67% (4/6) of the interventions included contact with a practitioner, whereas 33% (2/6) were entirely self-guided. One of those offering contact with a practitioner only offered this to parents and not the participating child. A total of 33% (2/6) of the interventions included scheduled sessions with a practitioner to discuss content, with 17% (1/6) also supporting asynchronous communication with a practitioner via the platform. Finally, in 17% (1/6) of the interventions, participants could contact practitioners via the platform for technical troubleshooting as required.

Tailoring and Additional Features

In total, 17% (1/6) of the platforms allowed participants to select content based on their clinical and cultural needs. Content was selected from a list of available modules, although the process through which the dyads selected this content was not described. This same platform offered dyads links to external sources of information based on their responses to questionnaires.

Reported User Engagement Indicators of Platforms for Parent-Child Dyads

A total of 14% (12/85) of the studies evaluated 6 different interventions designed for use by parent-child dyads. Of those 12 studies, 5 (42%) reported on user engagement indicators, including completion or fidelity (4/5, 80%) and satisfaction or acceptability (3/5, 60%). The remaining 58% (7/12) of the studies reported on mental health or relational outcomes and did not report on user engagement indicators.

The 33% (4/12) of the studies reporting on completion or fidelity documented the number of participants who completed the entire intervention as prescribed. One study also reported on the average time it took participants to complete the intervention, and another reported on the number of dyads who accessed all sessions and received calls from a practitioner. No studies reported on participants’ interaction with the platform or any predictors of noncompletion.

All studies reporting on satisfaction and acceptability did so using nonvalidated measures. Mean satisfaction ratings were high. One study asked participants to indicate how easily they found time to complete the activities together, with a mean rating of 3.04/5 (SD 0.37) for daughters and a mean rating of 3.24/5 (SD 0.33) for mothers. In no study did the satisfaction and acceptability data distinguish between platform and intervention satisfaction.

No studies on parent-child interventions identified platform build or design characteristics as moderators of intervention effect. No studies performed formative evaluations of the platforms, and no studies reported on design and build characteristics that enabled coparticipation.

Families

Features of Platforms for Families

Overview

Among the 24 platforms, 2 (8%) designed for cocompletion by families were identified. The intervention targets included family functioning when a child has an anxiety diagnosis (1/2, 50%) and family functioning when a child, adolescent, or teenager has a traumatic brain injury (1/2, 50%). Both platforms were intended for use by a child, adolescent, or teenager with a presenting clinical concern and any family members, including parents and siblings. Though siblings and other family members were invited to participate, the studies detailed outcomes and engagement for a single parent and child only.

Structure and Duration of Engagement

Intervention participation on one platform extended for 10 weeks over 11 web-based chapters, and the other delivered 7 to 11

XSL•FO
RenderX

sessions over 6 months. Both were designed for sequential completion of module content.

Coparticipation and Contact With Practitioners

One platform asked family members to complete the entire intervention together. The other asked parents to complete sections themselves before working with their children on a small number of modules intended for cocompletion. Both included scheduled telehealth sessions with a practitioner during intervention participation. In addition, one platform also included a platform-based message system for contacting practitioners asynchronously. In this same platform, practitioners also provided reports to participants following exercise completion.

Tailoring and Additional Features

One platform provided no tailoring beyond engagement with and feedback provided by practitioners. The other platform included supplementary sessions that could be completed by families should they wish to. In addition, this platform supported cocompletion by asking family members to select their picture when they were present. The platform would then either prompt individual family members to respond or ask all family members to respond together.

Reported User Engagement Indicators of Platforms for Families

A total of 27% (23/85) of the studies examined the 2 family-based interventions. Of these 23 studies, 16 (70%) reported user engagement indicators including satisfaction and ease of use (n=9, 56%); completion rates, compliance, adherence, and website use (n=13, 81%); and feasibility (n=2, 12%). The remaining 30% (7/23) of the studies did not report on satisfaction, completion, or feasibility data or findings.

Of the 9 studies reporting satisfaction and ease of use, 6 (67%) used an adaptation of the Website Evaluation Questionnaire [128] to measure participant satisfaction with the intervention. The remaining 33% (3/9) of the studies administered nonvalidated measures developed for the studies. Satisfaction ratings were high across all studies. In 33% (2/6) of the studies in which the Website Evaluation Questionnaire was administered, participants were asked to rate the website's ease of use, generally reporting that the website was "moderately easy" to "easy" to use. Participants in one study reported a preference for meeting in person. Other than this, satisfaction ratings either were relevant to content or did not distinguish between platform and intervention satisfaction.

Completion rates, compliance, adherence, and website use were all reported as combinations of the following: the number of participants who completed the entire intervention, the average number of modules completed, time spent on the platform, and the number of families who completed supplemental sessions. Feasibility was reported similarly, with one study also reporting that families were able to complete all sessions without practitioner assistance. In addition, one study reported on number of sessions completed as a predictor of symptom change (with inconsistent effect), and another reported on participant characteristics at baseline as predictors of completion.

A total of 13% (3/23) of the studies also measured participants' technology use and comfort with technology before the commencement of the intervention and examined this as a predictor of intervention effect. Results were inconsistent. In addition, one study identified whether participants' preference for treatment modality before the intervention, that is, face-to-face, web self-paced, or therapist-guided modality, impacted treatment outcomes. It was found that adolescent treatment preference was significantly related to attrition, but there were no other links with treatment effect or satisfaction.

No studies on family-based platforms identified build or design characteristics as moderators of the intervention effect. No studies performed a formative evaluation of the platforms, and aside from one study describing how participants identified that they were present, no studies reported design and build characteristics that enabled coparticipation.

Caregiver–Care Recipient (Family) Dyads

Features of Platforms for Caregiver–Care Recipient Dyads

Among the 24 platforms, 2 (8%) were identified for family caregiver–care recipient dyads. The targets of the interventions on the platforms included dyadic resilience for patients with stroke or brain tumor and their family caregivers (1/2, 50%) and the psychosocial health of patients with cancer and their family caregivers (1/2, 50%).

Structure and Duration of Engagement

One of the 2 platforms involved intervention participation over 6 weeks, with 3 sessions delivered sequentially. The other platform contained 4 web-based modules and a participant journal completed over 8 weeks. While not explicitly reported, it appeared that this platform also required sequential completion of intervention content.

Coparticipation and Contact With Practitioners

Both platforms were designed to be completed by members of the dyad together; however, one had the option of completing the entire intervention independently if desired. One platform contained an asynchronous help function that generated an email to the project director. The other included a telehealth session before commencement of the web-based component.

Tailoring and Additional Features

One platform contained several tailored elements, whereas the other did not offer any personalization. Tailoring included platform-generated messages that provided web links addressing the dyad's concerns and supplementary activities offered between web sessions. Both were generated from self-reported baseline information.

Reported User Engagement Indicators of Platforms for Caregiver–Care Recipient Dyads

The 2 interventions were each evaluated in 1% (1/85) of the studies. Both studies reported on satisfaction, and one reported on feasibility, with both reporting high satisfaction ratings. One study reported that there were no adverse effects of participants completing the intervention on the web-based platform together, and the other identified the length of the modules and the ability to complete the intervention in the users' own time as facilitators

XSL•FO
RenderX

to use. In one study, participants noted that finding time to complete the intervention as a dyad was sometimes challenging. Where feasibility was reported, the study found lower enrollment rates than those for previous in-person randomized controlled trials but higher retention rates.

Neither study identified build or design characteristics as moderators of intervention effect. No study performed a formative evaluation of the platforms, and neither reported on design nor build characteristics that enabled coparticipation.

Discussion

Principal Findings

This review details build, design, and user engagement characteristics of platforms that enable cocompletion of clinical interventions by related people. To distinguish effective platform contributors to engagement from elements pertaining to intervention content, we selected only those platforms housing interventions of established clinical efficacy (ie, previously reported significant improvement of at least one mental health or relational outcome). Some common design features were identified; however, in contrast to expected findings, specific design characteristics enabling cocompletion were rarely reported, and evidence for engaging families was underexplored.

Common Platform Features

This review identified platform design features that were common across the included studies. Regardless of the relationship targeted, most platforms delivered a structured intervention that required engagement over a prescribed duration with content completed sequentially. A total of 8% (2/24) of the platforms allowed participants to access content in a nonsequential manner, and a handful (4/24, 17%) offered supplementary content based on identified need. Retention rates remain low for DMHIs [29], and there are further complexities when family members participate together [50]. As such, consideration might be given to ways in which families' time on the platform can be optimized.

Single Session Thinking is one process through which therapists treat each encounter as if it were the sole session, encouraging the participants to make the most of the time [129]. Adaptation to web delivery of family therapy sessions already holds promise [130], and digital single-session interventions have been trialed in college student settings with positive preliminary findings [131-133]. Therefore, there is emerging evidence suggesting that Single Session Thinking principles could be readily applied to DMHIs, mimicking single, stand-alone sessions that address the family's present needs as they identify them. Check-in prompts and invitations to return as needed could be automated from the platform to encourage return visits as required or desired by the family. A platform designed to deliver content in this way would likely reduce the burden on families and provide greater flexibility in how they access content.

Minimal tailoring was offered in 29% (7/24) of the DMHIs identified in this review, providing more or less the same intervention to all participants. A total of 67% (16/24) of the interventions included interaction (either synchronous or asynchronous) with a practitioner. Evidence for personalized mental health care is growing rapidly, acknowledging the complexity and diversity of individuals and families [134,135]. Higher levels of engagement are reported for guided interventions (ie, those where participants have some contact with a practitioner) than for self-guided interventions; however, incorporating human contact can be costly and can limit the flexibility and accessibility associated with DMHIs [136]. Research suggests that, compared with targeted or generic feedback, personalization can be used to improve engagement and subsidize personal contact and contributes to positive attitudes toward a DMHI [134,135]. Beyond this, several studies included in this review (9/85, 11%) identified baseline characteristics that moderated participants' responses to the intervention. These included characteristics such as age, relationship status, and previous comfort with technology. Understanding how baseline measures might impact participants' ability or desire to engage with platforms and providing options for personalization accordingly would likely result in greater engagement. A family-based platform might include tailored design options such as color and font choices, preferences for video- or text-based content, and preferences for receipt of prompts and reminders. In addition, if children are present, families could have the option to access content that has been adapted for younger readers. In a world where artificial intelligence is supporting personalization across the internet, it would be remiss not to consider personalization in family- and relational-based DMHIs.

Platform Features for Enabling Cocompletion

By their nature, computers and mobile devices are designed for use by individuals. Given obvious complexities involved in having multiple people participate in a web-based intervention together, it was expected that platforms designed for such use may contain features for enabling cocompletion across the life span. It was also expected that the way in which participants engage may differ from that in platforms designed for individual use. This could include considerations about privacy of individual participants' data, methods for encouraging participants to work together, and design choices to allow all participants to contribute to activities. One platform requested participants to select their image when they were in attendance, and this was then used to prompt individuals to respond and participate in activities. Other than this, no study identified platform characteristics that were included to specifically enable cocompletion. In general, studies detailed participants' engagement with the intervention but not with the platform. Reporting on platform engagement might include details on how participants navigated the interface, how they identified and accessed content, or the modes through which content was delivered. On the other hand, intervention reporting was found to delve into factors such as attrition rate and measurement completion. Crucially, it is important to distinguish between intervention trial attrition (ie, dropout or loss to follow-up) and platform disengagement (ie, nonuse attrition), as recommended in a previous review [137]. These 2 forms of attrition are influenced by distinct factors [138,139], and failure to differentiate between them could potentially lead to misinterpretation of platform engagement dynamics.

XSL•FO
RenderX

It was also expected that studies would provide insights into the build and design considerations related to individual user privacy and safety within a shared web-based space. This encompasses considerations such as determining when an individual's information can or should be shared with other members of the family and effectively identifying and responding to safety risks. From our perspective, these design aspects are essential considerations when developing a family-based DMHI. However, none of the studies identified in this review reported or discussed how they tackled or addressed these privacy and safety considerations. To further ensure the adequate addressing of not only these concerns and anticipate other potential considerations, rigorous co-design processes are essential. This co-design strategy would significantly contribute to the refinement of family-based DMHIs, ensuring that they meet the nuanced needs of users.

Engagement With Practitioners

The varied nature of engagement with guided tools (ie, involving interaction with practitioners, structured sessions, and feedback loops) stands in stark contrast to the self-guided use and consistent participation characterizing engagement with tools lacking contact with practitioners. Recognizing challenges intrinsic to self-guided tools, such as user motivation and adherence, becomes paramount, particularly given that the absence of practitioner involvement is likely to make the sustainability of user interest more demanding. The role of technology in promoting engagement with practitioners is multifaceted, encompassing communication facilitation through asynchronous methods and data-driven insights that enhance personalized interactions. Moreover, exploring hybrid models and incorporating periodic check-ins or teletherapy sessions within self-guided platforms presents a promising balance between autonomy and professional support.

Addressing challenges in technology engagement involves prioritizing user-centered design; integrating behavioral science principles; and leveraging feedback mechanisms, either automated or through clinician input, to ensure continuous support and guidance. Looking forward, suggested avenues for future research are many, including the long-term effectiveness of guided and self-guided tools, understanding the impact of different engagement strategies, and developing sophisticated technology-assisted therapeutic approaches.

Evidence for Enabling Cocompletion

We faced constraints in reporting evidence on platform features that engaged and enabled cocompletion by families because no study conducted a direct evaluation of the platform design. This limitation hindered our ability to provide comprehensive insights into the effectiveness of features promoting cocompletion among participating family members. In addition, while several studies (13/85, 15%) evaluated practitioner support, family member coparticipation, population characteristics, and baseline scores on mental health or relational measures as moderators of intervention outcomes, no study evaluated design features as potential moderators of intervention outcomes.

Of the 85 included studies, 66 (78%) reported on user engagement indicators. Of those, most (48/66, 73%) used custom, nonvalidated measures, and the remaining studies used validated measures that were intervention specific and gave no information about platform engagement. Given this measurement heterogeneity, little is possible by way of cross-study comparison. In addition, without evaluation of platform design strategies, no conclusions can be drawn about enabling or disabling features. The capacity for real-world translation and understanding of how to overcome known barriers is constrained.

A Need for Cohesive Platform Evaluation and Reporting

Platform user experience design, including ease of use, navigation, screen layout, readability, gamification, feedback, and attractiveness, plays a large role in a participant's perception of and engagement with a website and, ultimately, a site's usability [29,140,141]. In addition, individual participant characteristics such as age, literacy level, level of disability, and mental health conditions may impact their engagement with and ability to use a platform as designed. When a family presents on a web-based platform, more than one person's needs must be catered to.

There is a lack of consensus and shared understanding of how to usefully conceptualize and measure engagement with and accessibility of digital mental health platforms [35,36]. This variability is not unique to the context of family-based mental health platforms, with reviews of engagement in digital mental health reporting similar heterogeneity [35,134,142]. Studies tend to report on measures such as completion or attrition rates, usability, user satisfaction, acceptability, and feasibility as indicators of how well the application engaged users. Often, these data are self-reported. Given the high attrition rates for self-guided platforms [102] and additional complexities involved in requiring family members to cocomplete activities [50], understanding platform characteristics that enable co-use and promote engagement is vital to informing the future development of such platforms. There is limited direct evidence to support practitioners, developers, and designers in understanding why engagement levels remain low, and there remains a limited understanding of how to design a DMHI to optimize engagement for families.

Assessment of user engagement indicators such as completion data alone is likely insufficient to measure how well a platform engaged its users. For example, reporting on duration of participation and sessions completed neglects factors impacting how families navigate the website, such as interface design and organization, and user characteristics. Analysis of platform use patterns and baseline characteristics in addition to these completion statistics would provide greater insights into how families engage with a platform. Formative as opposed to summative evaluations of usability are conducted to inform the redesign and improvement of a web interface. Formative evaluations consider multiple factors and involve building a deep understanding of user perceptions and use patterns of platforms. In addition to self-reported measures and completion rates, formative evaluations often also consider website analytics such as bounce rate, pages per session, top exit pages, and the pathways that users take to get to pages where they ultimately

XSL•FO
RenderX

spend most of their time. It is a recommendation of this review that formative evaluations of web-based mental health platforms become common practice for DMHIs, particularly for novel and complex applications such as family-based platforms.

Finally, a systematic review of evaluations of usability of mobile mental health technologies [143] recommended closer collaboration between health care and computer science experts when evaluating DMHIs, suggesting that this would increase the quality of interpretation of the evaluation. A summary of learnings from the *ParentWorks* trial identified an expected benefit of having involved a web agency during the early stages of content translation to optimize user experience [144]. An interdisciplinary approach might enhance knowledge sharing, too, through detailed reporting of DMHI design decisions and their interactions with platform elements and clinical outcomes.

Clearly, there remains a need for coherent reporting and evaluation practices in the field to inform guidelines and policy on effective strategies for engaging families on the web in mental health–related interventions. Until rigorous co-design with families and an interdisciplinary approach between content experts and user experience designers is taken to formative evaluations, the growth and expansion of efficacious mental health platforms for family use will lag.

Study Strengths, Limitations, and Future Research Directions

This study represents the first of its kind. Using a replicable search strategy over 4 periods, we synthesized in this study the state of the published evidence regarding platform design and build characteristics enabling successful engagement of related parties with digitally delivered mental health interventions. Given that gray literature was not searched for this review, it is possible that emerging evidence for new multiuser digital platforms was missed. Our findings are limited by the technical reporting of the studies. Principally, many studies did not provide details about their platform build or the way in which participants engaged with the platform, including whether coparticipation was expected. Where this information was not provided, the study authors were contacted, and websites were searched to retrieve the relevant information. It is likely that examination of some relevant functionality was precluded when this information was not provided or was insufficient.

Many studies (24/85, 28%) explicitly excluded participants when those other than the identified person had a mental illness. Whether through caregiving burden, stigma, or familial shared conditions, it is rare for a family presenting for therapy to have only 1 member experiencing mental health stress or significant challenges [145,146]. Given the potential of these platforms to aid family therapy, further research with families in which multiple members experience stress or mental health challenges is needed. Until then, it is difficult to generalize the evidence reported in this review to the real-world experience of families who may present for family therapy.

In addition, diversity in populations was limited, with most studies including White, heterosexual, and middle-class participants. There was a lack of evidence from low- and middle-income countries (LMICs), with all studies conducted in more high-income countries. The technological experiences and needs of families in LMICs will likely vary significantly from those in more high-income countries given, among other factors, the varying degree of ease of access to technology. Digital interventions have the potential to expand reach and access to services; however, until participants from LMICs are included in studies of digital platforms for families, findings cannot be generalized to these populations and ultimate reach will be limited.

As this is a new and novel field, language and terminology are still being defined, and means of measuring and defining engagement and feasibility are not well established [29]. Of the included studies, 52% (44/85) were published in the last 5 years, reflecting rapid developments in technology and associated applications.

Conclusions

While there is emerging evidence suggesting that DMHIs are clinically effective, there remains a large evidence gap in the literature on the extent to which platform-specific design and build elements may also contribute to timely access, user experience, safe cocompletion by family members, and clinical outcomes. In the service of improved mental and relational health outcomes, our findings point to a significant opportunity for meaningful cross-disciplinary research, development, and evaluation of family-based mental health platforms. Findings from the next era of research will be central to enabling policy and practice advancements in equitable access to effective mental health care support for families.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excluded studies and platforms.
[DOCX File , 26 KB - jmir_v26i1e49431_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Search strategies.
[DOCX File , 22 KB - jmir_v26i1e49431_app2.docx]

References

1. Rehm J, Shield KD. Global burden of disease and the impact of mental and addictive disorders. Curr Psychiatry Rep 2019 Mar 07;21(2):10. [doi: 10.1007/s11920-019-0997-0] [Medline: 30729322]
2. Duden GS, Gersdorf S, Stengler K. Global impact of the COVID-19 pandemic on mental health services: a systematic review. J Psychiatr Res 2022 Oct;154:354-377 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.08.013] [Medline: 36055116]
3. Foster K, Maybery D, Reupert A, Gladstone B, Grant A, Ruud T, et al. Family-focused practice in mental health care: an integrative review. Child Youth Serv 2016 Mar 23;37(2):129-155 [FREE Full text] [doi: 10.1080/0145935x.2016.1104048]
4. Anderson H. Rethinking family therapy: a delicate balance. J Marital Fam Ther 2007 Jun 08;20(2):145-149 [FREE Full text] [doi: 10.1111/J.1752-0606.1994.tb01023.X]
5. Whisman MA. Marital distress and DSM-IV psychiatric disorders in a population-based national survey. J Abnorm Psychol 2007 Aug;116(3):638-643. [doi: 10.1037/0021-843X.116.3.638] [Medline: 17696721]
6. Department of Health and Human Services. Police-recorded crime trends in Victoria during the COVID-19 pandemic: update to end of December. Crime Statistics Agency. 2020. URL: https://www.crimestatistics.vic.gov.au/research-and-evaluation/publications/police-recorded-crime-trends-in-victoria-during-the-covid-19-1 [accessed 2024-04-29]
7. Painter FL, Booth AT, Letcher P, Olsson CA, McIntosh JE. The lived experience of stress for parents in the context of COVID-19–related disruption. Fam Relat 2023 Mar 07;72(4):1511-1531 [FREE Full text] [doi: 10.1111/fare.12867]
8. Chung G, Lanier P, Wong PY. Mediating effects of parental stress on harsh parenting and parent-child relationship during coronavirus (COVID-19) pandemic in Singapore. J Fam Violence 2022;37(5):801-812 [FREE Full text] [doi: 10.1007/s10896-020-00200-1] [Medline: 32895601]
9. Wolf JP, Freisthler B, Chadwick C. Stress, alcohol use, and punitive parenting during the COVID-19 pandemic. Child Abuse Negl 2021 Jul;117:105090 [FREE Full text] [doi: 10.1016/j.chiabu.2021.105090] [Medline: 33975257]
10. Connell CM, Strambler MJ. Experiences with COVID-19 stressors and parents' use of neglectful, harsh, and positive parenting practices in the Northeastern United States. Child Maltreat 2021 Aug;26(3):255-266 [FREE Full text] [doi: 10.1177/10775595211006465] [Medline: 33787377]
11. Bartle-Haring S, Slesnick N, Murnan A. Benefits to children who participate in family therapy with their substance-using mother. J Marital Fam Ther 2018 Oct;44(4):671-686 [FREE Full text] [doi: 10.1111/jmft.12280] [Medline: 28950404]
12. Horigian VE, Feaster DJ, Robbins MS, Brincks AM, Ucha J, Rohrbaugh MJ, et al. A cross-sectional assessment of the long term effects of brief strategic family therapy for adolescent substance use. Am J Addict 2015 Oct;24(7):637-645 [FREE Full text] [doi: 10.1111/ajad.12278] [Medline: 26359441]
13. Aldersey HM, Whitley R. Family influence in recovery from severe mental illness. Community Ment Health J 2015 May;51(4):467-476. [doi: 10.1007/s10597-014-9783-y] [Medline: 25492380]
14. Carr A. The effectiveness of family therapy and systemic interventions for adult‐focused problems. J Fam Ther 2008 Dec 23;31(1):46-74 [FREE Full text] [doi: 10.1111/j.1467-6427.2008.00452.x]
15. Carr A. The evidence base for family therapy and systemic interventions for child‐focused problems. J Fam Ther 2014 Jan 29;36(2):107-157 [FREE Full text] [doi: 10.1111/1467-6427.12032]
16. Carr A. How and why do family and systemic therapies work? Aust N Z J Fam Ther 2016 Apr 05;37(1):37-55 [FREE Full text] [doi: 10.1002/anzf.1135]
17. Hartmann M, Bäzner E, Wild B, Eisler I, Herzog W. Effects of interventions involving the family in the treatment of adult patients with chronic physical diseases: a meta-analysis. Psychother Psychosom 2010;79(3):136-148. [doi: 10.1159/000286958] [Medline: 20185970]
18. O'Brien M, Daley D. Self-help parenting interventions for childhood behaviour disorders: a review of the evidence. Child Care Health Dev 2011 Sep;37(5):623-637. [doi: 10.1111/j.1365-2214.2011.01231.x] [Medline: 21585419]
19. Cardy JL, Waite P, Cocks F, Creswell C. A systematic review of parental involvement in cognitive behavioural therapy for adolescent anxiety disorders. Clin Child Fam Psychol Rev 2020 Dec;23(4):483-509 [FREE Full text] [doi: 10.1007/s10567-020-00324-2] [Medline: 32862330]
20. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organisation. URL: https://iris.who.int/bitstream/handle/10665/311941/9789241550505-eng.pdf?sequence=31 [accessed 2024-04-29]
21. Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unützer J, Patel V, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. Lancet Psychiatry 2017 Jun;4(6):486-500 [FREE Full text] [doi: 10.1016/S2215-0366(17)30096-2] [Medline: 28433615]
22. Sherifali D, Ali MU, Ploeg J, Markle-Reid M, Valaitis R, Bartholomew A, et al. Impact of internet-based interventions on caregiver mental health: systematic review and meta-analysis. J Med Internet Res 2018 Jul 03;20(7):e10668 [FREE Full text] [doi: 10.2196/10668] [Medline: 29970358]

XSL•FO
RenderX

23. Babbage CM, Jackson GM, Davies EB, Nixon E. Self-help digital interventions targeted at improving psychological well-being in young people with perceived or clinically diagnosed reduced well-being: systematic review. JMIR Ment Health 2022 Aug 26;9(8):e25716 [FREE Full text] [doi: 10.2196/25716] [Medline: 36018675]
24. Richardson T, Stallard P, Velleman S. Computerised cognitive behavioural therapy for the prevention and treatment of depression and anxiety in children and adolescents: a systematic review. Clin Child Fam Psychol Rev 2010 Sep;13(3):275-290. [doi: 10.1007/s10567-010-0069-9] [Medline: 20532980]
25. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
26. De Witte NA, Joris S, Van Assche E, Van Daele T. Technological and digital interventions for mental health and wellbeing: an overview of systematic reviews. Front Digit Health 2021 Dec 23;3:754337 [FREE Full text] [doi: 10.3389/fdgth.2021.754337] [Medline: 35005695]
27. Shaffer KM, Tigershtrom A, Badr H, Benvengo S, Hernandez M, Ritterband LM. Dyadic psychosocial eHealth interventions: systematic scoping review. J Med Internet Res 2020 Mar 04;22(3):e15509 [FREE Full text] [doi: 10.2196/15509] [Medline: 32130143]
28. Knopp K, Schnitzer JS, Khalifian C, Grubbs K, Morland LA, Depp C. Digital interventions for couples: state of the field and future directions. Couple Family Psychol Res Pract 2023 Dec;12(4):201-217 [FREE Full text] [doi: 10.1037/cfp0000213]
29. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? Front Psychiatry 2022;13:900615 [FREE Full text] [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
30. Boardman S. Creating wellness apps with high patient engagement to close the intention-action gap. Int Psychogeriatr 2021 Jun;33(6):551-552. [doi: 10.1017/S1041610220003385] [Medline: 34173354]
31. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Mar 26;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
32. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and Somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 06;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
33. Usability 101: introduction to usability. Nielsen Norman Group. 2012. URL: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ [accessed 2023-04-23]
34. Falloon IR. Family interventions for mental disorders: efficacy and effectiveness. World Psychiatry 2003 Mar;2(1):20-28 [FREE Full text] [Medline: 16946881]
35. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
36. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun 13;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
37. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
38. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
39. Endnote Team. Endnote. Clarivate. 2013. URL: https://clarivate.com/innovation-exchange/solution/endnote/ [accessed 2024-04-29]
40. Babineau J. Product review: Covidence (systematic review software). J Can Health Libr Assoc 2014 Aug 01;35(2):68. [doi: 10.5596/C14-016]
41. Luo X, Li J, Cao Q, Sun L, Chen Y, Zhao J, et al. A feasibility study of an integrated couples-based supportive programme for Chinese couples living with colorectal cancer. Nurs Open 2021 Mar 26;8(2):920-926 [FREE Full text] [doi: 10.1002/nop2.700] [Medline: 33570269]
42. Wang Z, Chen M, Cao Q, Gong J, Zhao J, Lin C, et al. An integrated intervention programme for couples facing colorectal cancer: a randomized controlled trial. Clin Psychol Psychother (Forthcoming) 2023 Nov 10. [doi: 10.1002/cpp.2919] [Medline: 37947325]
43. Santisteban DA, Czaja SJ, Nair SN, Mena MP, Tulloch AR. Computer informed and flexible family-based treatment for adolescents: a randomized clinical trial for at-risk racial/ethnic minority adolescents. Behav Ther 2017 Jul;48(4):474-489. [doi: 10.1016/j.beth.2016.11.001] [Medline: 28577584]
44. Price-Blackshear MA, Pratscher SD, Oyler DL, Armer JM, Cheng A, Cheng MX, et al. Online couples mindfulness-based intervention for young breast cancer survivors and their partners: a randomized-control trial. J Psychosoc Oncol 2020;38(5):592-611. [doi: 10.1080/07347332.2020.1778150] [Medline: 32552446]

XSL•FO
RenderX

45. McLellan LF, Woon S, Hudson JL, Lyneham HJ, Karin E, Rapee RM. Treating child anxiety using family-based internet delivered cognitive behavior therapy with brief therapist guidance: a randomized controlled trial. J Anxiety Disord 2024 Jan;101:102802 [FREE Full text] [doi: 10.1016/j.janxdis.2023.102802] [Medline: 38071870]
46. Fitzpatrick S, Wagner AC, Crenshaw AO, Varma S, Whitfield KM, Valela R, et al. Initial outcomes of couple HOPES: a guided online couple intervention for PTSD and relationship enhancement. Internet Interv 2021 Sep;25:100423 [FREE Full text] [doi: 10.1016/j.invent.2021.100423] [Medline: 34401382]
47. Monson CM, Wagner AC, Crenshaw AO, Whitfield KM, Newnham CM, Valela R, et al. An uncontrolled trial of couple HOPES: a guided online couple intervention for PTSD and relationship enhancement. J Fam Psychol 2022 Sep;36(6):1036-1042. [doi: 10.1037/fam0000976] [Medline: 35266773]
48. Crenshaw AO, Whitfield KM, Collins A, Valela R, Varma S, Landy MS, et al. Partner outcomes from an uncontrolled trial of couple HOPES: a guided online couple intervention for posttraumatic stress disorder and relationship enhancement. J Trauma Stress 2023 Mar;36(1):230-238. [doi: 10.1002/jts.22878] [Medline: 36116104]
49. Morland LA, Wachsman T, Webster K, Fitzpatrick S, Valela R, Crenshaw AO, et al. A pilot of couple HOPES within the U.S. veterans affairs healthcare system: PTSD and relationship outcomes in veteran couples. Psychol Serv (Forthcoming) 2023 Jun 19. [doi: 10.1037/ser0000773] [Medline: 37338435]
50. Fergus KD, McLeod D, Carter W, Warner E, Gardner SL, Granek L, et al. Development and pilot testing of an online intervention to support young couples' coping and adjustment to breast cancer. Eur J Cancer Care (Engl) 2014 Jul;23(4):481-492. [doi: 10.1111/ecc.12162] [Medline: 24472013]
51. Ianakieva I, Fergus K, Ahmad S, Pos A, Pereira A. A model of engagement promotion in a professionally facilitated online intervention for couples affected by breast cancer. J Marital Fam Ther 2016 Oct;42(4):701-715. [doi: 10.1111/jmft.12172] [Medline: 27234009]
52. Ianakieva I, Fergus K, Ahmad S, Pereira A, Stephen J, McLeod D, et al. Varying levels and types of engagement in an online relationship enhancement program for couples following breast cancer. J Couple Relatsh Ther 2019 Jan 22;18(1):22-43 [FREE Full text] [doi: 10.1080/15332691.2018.1546248]
53. Fergus K, Ahmad S, Ianakieva I, McLeod D, Carter W. Metaphor and meaning in an online creative expression exercise to promote dyadic coping in young couples affected by breast cancer. Arts Health 2016 Aug 08;9(2):139-153 [FREE Full text] [doi: 10.1080/17533015.2016.1208667]
54. Fergus K, Ahmad S, Gardner S, Ianakieva I, McLeod D, Stephen J, et al. Couplelinks online intervention for young couples facing breast cancer: a randomised controlled trial. Psychooncology 2022 Mar;31(3):512-520. [doi: 10.1002/pon.5836] [Medline: 34669239]
55. Fergus K, Tanen A, Ahmad S, Gardner S, Warner E, McLeod D, et al. Treatment satisfaction with couplelinks online intervention to promote dyadic coping in young couples affected by breast cancer. Front Psychol 2022;13:862555 [FREE Full text] [doi: 10.3389/fpsyg.2022.862555] [Medline: 35783696]
56. Canfield SM, Canada KE, Rolbiecki AJ, Petroski GF. Feasibility and acceptability of an online mental health intervention for pregnant women and their partners: a mixed method study with a pilot randomized control trial. BMC Pregnancy Childbirth 2023 Oct 18;23(1):739 [FREE Full text] [doi: 10.1186/s12884-023-06031-4] [Medline: 37853333]
57. Selby E, Allabyrne C, Keenan JR. Delivering clinical evidence-based child-parent interventions for emotional development through a digital platform: a feasibility trial. Clin Child Psychol Psychiatry 2021 Oct;26(4):1271-1283 [FREE Full text] [doi: 10.1177/13591045211041580] [Medline: 34479455]
58. Braithwaite SR, Fincham FD. Computer-based dissemination: a randomized clinical trial of ePREP using the actor partner interdependence model. Behav Res Ther 2011 Mar;49(2):126-131. [doi: 10.1016/j.brat.2010.11.002] [Medline: 21159326]
59. Braithwaite SR, Fincham FD. Computer-based prevention of intimate partner violence in marriage. Behav Res Ther 2014 Mar;54:12-21 [FREE Full text] [doi: 10.1016/j.brat.2013.12.006] [Medline: 24463577]
60. Northouse L, Schafenacker A, Barr KL, Katapodi M, Yoon H, Brittain K, et al. A tailored Web-based psychoeducational intervention for cancer patients and their family caregivers. Cancer Nurs 2014;37(5):321-330 [FREE Full text] [doi: 10.1097/NCC.0000000000000159] [Medline: 24945270]
61. Vigerland S, Ljótsson B, Thulin U, Öst LG, Andersson G, Serlachius E. Internet-delivered cognitive behavioural therapy for children with anxiety disorders: a randomised controlled trial. Behav Res Ther 2016 Jan;76:47-56 [FREE Full text] [doi: 10.1016/j.brat.2015.11.006] [Medline: 26649465]
62. Feinberg ME, Boring J, Le Y, Hostetler ML, Karre J, Irvin J, et al. Supporting military family resilience at the transition to parenthood: a randomized pilot trial of an online version of family foundations. Family Relations 2019 Dec 10;69(1):109-124 [FREE Full text] [doi: 10.1111/fare.12415]
63. Gil M, Kim SS, Kim D, Kim S. Online coaching blended couple-oriented intervention for preventing depression among Korean middle adulthood: a feasibility study. Fam Process 2023 Dec;62(4):1478-1505. [doi: 10.1111/famp.12924] [Medline: 37641515]
64. Kahn JR, Collinge W, Soltysik R. Post-9/11 veterans and their partners improve mental health outcomes with a self-directed mobile and web-based wellness training program: a randomized controlled trial. J Med Internet Res 2016 Sep 27;18(9):e255 [FREE Full text] [doi: 10.2196/jmir.5800] [Medline: 27678169]

XSL•FO
RenderX

65. Schinke SP, Cole KC, Fang L. Gender-specific intervention to reduce underage drinking among early adolescent girls: a test of a computer-mediated, mother-daughter program. J Stud Alcohol Drugs 2009 Jan;70(1):70-77 [FREE Full text] [doi: 10.15288/jsad.2009.70.70] [Medline: 19118394]
66. Schinke SP, Fang L, Cole KC. Preventing substance use among adolescent girls: 1-year outcomes of a computerized, mother-daughter program. Addict Behav 2009 Dec;34(12):1060-1064 [FREE Full text] [doi: 10.1016/j.addbeh.2009.06.007] [Medline: 19632053]
67. Schinke SP, Fang L, Cole KC. Computer-delivered, parent-involvement intervention to prevent substance use among adolescent girls. Prev Med 2009 Nov;49(5):429-435 [FREE Full text] [doi: 10.1016/j.ypmed.2009.08.001] [Medline: 19682490]
68. Fang L, Schinke SP, Cole KC. Preventing substance use among early Asian-American adolescent girls: initial evaluation of a web-based, mother-daughter program. J Adolesc Health 2010 Nov;47(5):529-532 [FREE Full text] [doi: 10.1016/j.jadohealth.2010.03.011] [Medline: 20970090]
69. Fang L, Schinke SP. Two-year outcomes of a randomized, family-based substance use prevention trial for Asian American adolescent girls. Psychol Addict Behav 2013 Sep;27(3):788-798 [FREE Full text] [doi: 10.1037/a0030925] [Medline: 23276322]
70. Schinke SP, Fang L, Cole KC, Cohen-Cutler S. Preventing substance use among black and Hispanic adolescent girls: results from a computer-delivered, mother-daughter intervention approach. Subst Use Misuse 2011;46(1):35-45 [FREE Full text] [doi: 10.3109/10826084.2011.521074] [Medline: 21190404]
71. Schwinn TM, Schinke S, Fang L, Kandasamy S. A web-based, health promotion program for adolescent girls and their mothers who reside in public housing. Addict Behav 2014 Apr;39(4):757-760 [FREE Full text] [doi: 10.1016/j.addbeh.2013.11.029] [Medline: 24447886]
72. Wade SL, Wolfe CR, Pestian JP. A web-based family problem-solving intervention for families of children with traumatic brain injury. Behav Res Methods Instrum Comput 2004 May;36(2):261-269. [doi: 10.3758/bf03195572] [Medline: 15354692]
73. Wade SL, Wolfe C, Brown TM, Pestian JP. Putting the pieces together: preliminary efficacy of a web-based family intervention for children with traumatic brain injury. J Pediatr Psychol 2005;30(5):437-442. [doi: 10.1093/jpepsy/jsi067] [Medline: 15944171]
74. Wade SL, Carey J, Wolfe CR. An online family intervention to reduce parental distress following pediatric brain injury. J Consult Clin Psychol 2006 Jun;74(3):445-454. [doi: 10.1037/0022-006X.74.3.445] [Medline: 16822102]
75. Wade SL, Carey J, Wolfe CR. The efficacy of an online cognitive-behavioral family intervention in improving child behavior and social competence following pediatric brain injury. Rehabil Psychol 2006 Aug;51(3):179-189 [FREE Full text] [doi: 10.1037/0090-5550.51.3.179]
76. Carey JC, Wade SL, Wolfe CR. Lessons learned: the effect of prior technology use on web-based interventions. Cyberpsychol Behav 2008 Apr;11(2):188-195 [FREE Full text] [doi: 10.1089/cpb.2007.0025] [Medline: 18422412]
77. Wade SL, Walz NC, Carey JC, Williams KM. Preliminary efficacy of a web-based family problem-solving treatment program for adolescents with traumatic brain injury. J Head Trauma Rehabil 2008;23(6):369-377. [doi: 10.1097/01.HTR.0000341432.67251.48] [Medline: 19033829]
78. Wade SL, Walz NC, Carey JC, Williams KM. Brief report: description of feasibility and satisfaction findings from an innovative online family problem-solving intervention for adolescents following traumatic brain injury. J Pediatr Psychol 2009 Jun;34(5):517-522 [FREE Full text] [doi: 10.1093/jpepsy/jsn081] [Medline: 18667477]
79. Wade SL, Walz NC, Carey J, Williams KM, Cass J, Herren L, et al. A randomized trial of teen online problem solving for improving executive function deficits following pediatric traumatic brain injury. J Head Trauma Rehabil 2010;25(6):409-415. [doi: 10.1097/HTR.0b013e3181fb900d] [Medline: 21076241]
80. Wade SL, Walz NC, Carey J, McMullen KM, Cass J, Mark E, et al. Effect on behavior problems of teen online problem-solving for adolescent traumatic brain injury. Pediatrics 2011 Oct;128(4):e947-e953 [FREE Full text] [doi: 10.1542/peds.2010-3721] [Medline: 21890828]
81. Wade SL, Walz NC, Carey J, McMullen KM, Cass J, Mark E, et al. A randomized trial of teen online problem solving: efficacy in improving caregiver outcomes after brain injury. Health Psychol 2012 Nov;31(6):767-776. [doi: 10.1037/a0028440] [Medline: 22746261]
82. Kurowski BG, Wade SL, Kirkwood MW, Brown TM, Stancin TM, Taylor HG. Online problem-solving therapy for executive dysfunction after child traumatic brain injury. Pediatrics 2013 Jul;132(1):e158-e166 [FREE Full text] [doi: 10.1542/peds.2012-4040] [Medline: 23753094]
83. Wade SL, Karver CL, Taylor HG, Cassedy A, Stancin T, Kirkwood MW, et al. Counselor-assisted problem solving improves caregiver efficacy following adolescent brain injury. Rehabil Psychol 2014 Mar;59(1):1-9 [FREE Full text] [doi: 10.1037/a0034911] [Medline: 24611923]
84. Wade SL, Stancin T, Kirkwood M, Brown TM, McMullen KM, Taylor HG. Counselor-assisted problem solving (CAPS) improves behavioral outcomes in older adolescents with complicated mild to severe TBI. J Head Trauma Rehabil 2014;29(3):198-207 [FREE Full text] [doi: 10.1097/HTR.0b013e31828f9fe8] [Medline: 23640543]

XSL•FO
RenderX

85. Narad ME, Minich N, Taylor HG, Kirkwood MW, Brown TM, Stancin T, et al. Effects of a web-based intervention on family functioning following pediatric traumatic brain injury. J Dev Behav Pediatr 2015;36(9):700-707 [FREE Full text] [doi: 10.1097/DBP.0000000000000208] [Medline: 26461100]
86. Petranovich CL, Wade SL, Taylor HG, Cassedy A, Stancin T, Kirkwood MW, et al. Long-term caregiver mental health outcomes following a predominately online intervention for adolescents with complicated mild to severe traumatic brain injury. J Pediatr Psychol 2015 Aug;40(7):680-688 [FREE Full text] [doi: 10.1093/jpepsy/jsv001] [Medline: 25682211]
87. Wade SL, Taylor HG, Cassedy A, Zhang N, Kirkwood MW, Brown TM, et al. Long-term behavioral outcomes after a randomized, clinical trial of counselor-assisted problem solving for adolescents with complicated mild-to-severe traumatic brain injury. J Neurotrauma 2015 Jul 01;32(13):967-975 [FREE Full text] [doi: 10.1089/neu.2014.3684] [Medline: 25738891]
88. Raj SP, Zhang N, Kirkwood MW, Taylor HG, Stancin T, Brown TM, et al. Online family problem solving for pediatric traumatic brain injury: influences of parental marital status and participation on adolescent outcomes. J Head Trauma Rehabil 2018;33(3):158-166 [FREE Full text] [doi: 10.1097/HTR.0000000000000331] [Medline: 28731871]
89. Wade SL, Cassedy AE, McNally KA, Kurowski BG, Kirkwood MW, Stancin T, et al. A randomized comparative effectiveness trial of family-problem-solving treatment for adolescent brain injury: parent outcomes from the coping with head injury through problem solving (CHIPS) study. J Head Trauma Rehabil 2019;34(6):E1-E9. [doi: 10.1097/HTR.0000000000000487] [Medline: 31033747]
90. Wade SL, Cassedy AE, Sklut M, Taylor HG, McNally KA, Kirkwood MW, et al. The relationship of adolescent and parent preferences for treatment modality with satisfaction, attrition, adherence, and efficacy: the coping with head injury through problem-solving (CHIPS) study. J Pediatr Psychol 2019 Apr 01;44(3):388-401. [doi: 10.1093/jpepsy/jsy087] [Medline: 30452665]
91. Wade SL, Cassedy AE, Taylor HG, McNally KA, Kirkwood MW, Stancin T, et al. Adolescent quality of life following family problem-solving treatment for brain injury. J Consult Clin Psychol 2019 Nov;87(11):1043-1055. [doi: 10.1037/ccp0000440] [Medline: 31556652]
92. Wade SL, Taylor HG, Yeates KO, Kirkwood M, Zang H, McNally K, et al. Online problem solving for adolescent brain injury: a randomized trial of 2 approaches. J Dev Behav Pediatr 2018;39(2):154-162. [doi: 10.1097/DBP.0000000000000519] [Medline: 29076889]
93. Narad ME, Raj S, Yeates KO, Taylor HG, Kirkwood MW, Stancin T, et al. Randomized controlled trial of an online problem-solving intervention following adolescent traumatic brain injury: family outcomes. Arch Phys Med Rehabil 2019 May;100(5):811-820. [doi: 10.1016/j.apmr.2019.01.010] [Medline: 30738021]
94. Doss BD, Cicila LN, Georgia EJ, Roddy MK, Nowlan KM, Benson LA, et al. A randomized controlled trial of the web-based OurRelationship program: effects on relationship and individual functioning. J Consult Clin Psychol 2016 Apr;84(4):285-296 [FREE Full text] [doi: 10.1037/ccp0000063] [Medline: 26999504]
95. Roddy MK, Georgia EJ, Doss BD. Couples with intimate partner violence seeking relationship help: associations and implications for self-help and online interventions. Fam Process 2018 Jun;57(2):293-307. [doi: 10.1111/famp.12291] [Medline: 28425562]
96. Georgia Salivar EJ, Roddy MK, Nowlan KM, Doss BD. Effectiveness of the online OurRelationship program for underserved couples. Couple Family Psychol Res Pract 2018 Sep;7(3-4):212-226 [FREE Full text] [doi: 10.1037/cfp0000110]
97. Doss BD, Roddy MK, Nowlan KM, Rothman K, Christensen A. Maintenance of gains in relationship and individual functioning following the online OurRelationship program. Behav Ther 2019 Jan;50(1):73-86 [FREE Full text] [doi: 10.1016/j.beth.2018.03.011] [Medline: 30661568]
98. Doss BD, Roddy MK, Llabre MM, Georgia Salivar E, Jensen-Doss A. Improvements in coparenting conflict and child adjustment following an online program for relationship distress. J Fam Psychol 2020 Mar;34(1):68-78 [FREE Full text] [doi: 10.1037/fam0000582] [Medline: 31380688]
99. Roddy MK, Stamatis CA, Rothman K, Doss BD. Mechanisms of change in a brief, online relationship intervention. J Fam Psychol 2020 Mar;34(1):57-67 [FREE Full text] [doi: 10.1037/fam0000569] [Medline: 31380690]
100. Biesen J, Roddy MK, Doss BD. The role of five-factor model personality traits in a web-based relationship improvement program. Couple Family Psychol Res Pract 2022 Apr 04 [FREE Full text] [doi: 10.1037/cfp0000221]
101. Roddy MK, Rothman K, Doss BD. A randomized controlled trial of different levels of coach support in an online intervention for relationship distress. Behav Res Ther 2018 Nov;110:47-54. [doi: 10.1016/j.brat.2018.09.002] [Medline: 30240874]
102. Rothman K, Roddy MK, Doss BD. Completion of a stand‐alone versus coach‐supported trial of a web‐based program for distressed relationships. Fam Relat 2019 Aug 07;68(4):375-389 [FREE Full text] [doi: 10.1111/fare.12378]
103. Roddy MK, Nowlan KM, Doss BD. A randomized controlled trial of coach contact during a brief online intervention for distressed couples. Fam Process 2017 Dec;56(4):835-851. [doi: 10.1111/famp.12262] [Medline: 27859100]
104. Knopp K, Rashkovsky K, Khalifian CE, Grubbs KM, Doss BD, Depp CA, et al. Pilot open trial of the OurRelationship online couples' program in a Veterans Affairs Medical Center. Couple Fam Psychol Res Pract 2022 Mar;11(1):33-41 [FREE Full text] [doi: 10.1037/cfp0000196]
105. Le Y, Hatch SG, Goodman ZT, Doss BD. Does coparenting improve during the OurRelationship program? Explorations within a low-income sample during the COVID-19 pandemic. J Fam Psychol 2022 Sep;36(6):1030-1035 [FREE Full text] [doi: 10.1037/fam0000991] [Medline: 35482627]

XSL•FO
RenderX

106. Hatch SG, Goodman ZT, Hatch HD, Le Y, Guttman S, Doss BD. Web-based relationship education and pornography-related behaviors: a single-group design during the COVID-19 pandemic. Arch Sex Behav 2023 May 04;52(4):1841-1853 [FREE Full text] [doi: 10.1007/s10508-022-02518-x] [Medline: 36600001]
107. Le Y, Roddy MK, Rothman K, Salivar EG, Guttman S, Doss BD. A randomized controlled trial of the online OurRelationship program with varying levels of coach support. Internet Interv 2023 Dec;34:100661 [FREE Full text] [doi: 10.1016/j.invent.2023.100661] [Medline: 37674656]
108. Doss BD, Knopp K, Roddy MK, Rothman K, Hatch SG, Rhoades GK. Online programs improve relationship functioning for distressed low-income couples: results from a nationwide randomized controlled trial. J Consult Clin Psychol 2020 Apr;88(4):283-294. [doi: 10.1037/ccp0000479] [Medline: 32134290]
109. Le Y, Roddy MK, Hatch SG, Doss BD. Mechanisms of improvements and maintenance in online relationship programs for distressed low-income couples. J Consult Clin Psychol 2020 Dec;88(12):1091-1104. [doi: 10.1037/ccp0000616] [Medline: 33370133]
110. Roddy MK, Rhoades GK, Doss BD. Effects of ePREP and OurRelationship on low-income couples' mental health and health behaviors: a randomized controlled trial. Prev Sci 2020 Aug;21(6):861-871. [doi: 10.1007/s11121-020-01100-y] [Medline: 32108304]
111. Le Y, O'Reilly Treter M, Roddy MK, Doss BD. Coparenting and parenting outcomes of online relationship interventions for low-income couples. J Fam Psychol 2021 Oct;35(7):1033-1039. [doi: 10.1037/fam0000850] [Medline: 33764084]
112. Roddy MK, Knopp K, Georgia Salivar E, Doss BD. Maintenance of relationship and individual functioning gains following online relationship programs for low-income couples. Fam Process 2021 Mar 30;60(1):102-118 [FREE Full text] [doi: 10.1111/famp.12541] [Medline: 32352573]
113. Georgia Salivar E, Knopp K, Roddy MK, Morland LA, Doss BD. Effectiveness of online OurRelationship and ePREP programs for low-income military couples. J Consult Clin Psychol 2020 Oct;88(10):899-906. [doi: 10.1037/ccp0000606] [Medline: 33048570]
114. Hatch SG, Knopp K, Le Y, Allen MO, Rothman K, Rhoades GK, et al. Online relationship education for help-seeking low-income couples: a bayesian replication and extension of the OurRelationship and ePREP programs. Fam Process 2022 Sep;61(3):1045-1061. [doi: 10.1111/famp.12710] [Medline: 34383314]
115. Mitchell EA, Le Y, Hatch SG, Guttman S, Doss BD. Effects of online relationship programs for low-income couples during the perinatal period. Behav Res Ther 2023 Aug;167:104337. [doi: 10.1016/j.brat.2023.104337] [Medline: 37327534]
116. Le Y, Xia M, Roddy MK, Hatch SG, Doss BD. Profiles of low-income help-seeking couples and implications for intervention gains: a couple-centered approach. Behav Ther 2024 May;55(3):443-456. [doi: 10.1016/j.beth.2023.08.002] [Medline: 38670660]
117. Barton AW, Gong Q, Guttman S, Doss BD. Trajectories of perceived gratitude and change following relationship interventions: a randomized controlled trial with lower-income, help-seeking couples. Behav Ther 2024 Mar;55(2):401-411. [doi: 10.1016/j.beth.2023.07.014] [Medline: 38418049]
118. Rizzo CJ, Houck C, Barker D, Collibee C, Hood E, Bala K. Project STRONG: an online, parent-son intervention for the prevention of dating violence among early adolescent boys. Prev Sci 2021 Mar;22(2):193-204 [FREE Full text] [doi: 10.1007/s11121-020-01168-6] [Medline: 32940857]
119. Whittingham K, Sheffield J, Mak C, Wright A, Boyd RN. Parenting acceptance and commitment therapy: an RCT of an online course with families of children with CP. Behav Res Ther 2022 Aug;155:104129. [doi: 10.1016/j.brat.2022.104129] [Medline: 35662680]
120. Piotrowska PJ, Tully LA, Collins DA, Sawrikar V, Hawes D, Kimonis ER, et al. ParentWorks: evaluation of an online, father-inclusive, universal parenting intervention to reduce child conduct problems. Child Psychiatry Hum Dev 2020 Aug 24;51(4):503-513 [FREE Full text] [doi: 10.1007/s10578-019-00934-0] [Medline: 31650461]
121. Dadds MR, Sicouri G, Piotrowska PJ, Collins DA, Hawes DJ, Moul C, et al. Keeping parents involved: predicting attrition in a self-directed, online program for childhood conduct problems. J Clin Child Adolesc Psychol 2019;48(6):881-893. [doi: 10.1080/15374416.2018.1485109] [Medline: 30067388]
122. Song L, Rini C, Deal AM, Nielsen ME, Chang H, Kinneer P, et al. Improving couples' quality of life through a web-based prostate cancer education intervention. Oncol Nurs Forum 2015 Mar;42(2):183-192 [FREE Full text] [doi: 10.1188/15.ONF.183-192] [Medline: 25806885]
123. Rhudy LM, Hines EA, Farr EM, Esterov D, Chesak SS. Feasibility and acceptability of the resilient living program among persons with stroke or brain tumor and their family caregivers. NeuroRehabilitation 2023;52(1):123-135. [doi: 10.3233/NRE-220127] [Medline: 36617758]
124. Asazawa K, Jitsuzaki M, Mori A, Ichikawa T. Effectiveness of a web-based partnership support program for preventing decline in the quality of life of male patients undergoing infertility treatment: a quasi-experimental study. Jpn J Nurs Sci 2023 Jul;20(3):e12536. [doi: 10.1111/jjns.12536] [Medline: 37057602]
125. Loew B, Rhoades G, Markman H, Stanley S, Pacifici C, White L, et al. Internet delivery of PREP-based relationship education for at-risk couples. J Couple Relatsh Ther 2012 Oct;11(4):291-309 [FREE Full text] [doi: 10.1080/15332691.2012.718968]

XSL•FO
RenderX

126. Nguyen TD, Attkisson CC, Stegner BL. Assessment of patient satisfaction: development and refinement of a service evaluation questionnaire. Eval Program Plann 1983;6(3-4):299-313. [doi: 10.1016/0149-7189(83)90010-1] [Medline: 10267258]
127. Attkisson CC, Zwick R. The client satisfaction questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
128. Rotondi AJ, Sinkule J, Spring M. An interactive web-based intervention for persons with TBI and their families: use and evaluation by female significant others. J Head Trauma Rehabil 2005;20(2):173-185. [doi: 10.1097/00001199-200503000-00005] [Medline: 15803040]
129. Hoyt MF, Young J, Rycroft P. Single session thinking 2020. Aust N Z J Fam Ther 2020 Oct 07;41(3):218-230 [FREE Full text] [doi: 10.1002/anzf.1427]
130. Hartley E, Moore L, Knuckey A, von Doussa H, Painter F, Story K, et al. Walk-in together: a pilot study of a walk-in online family therapy intervention. Aust N Z J Fam Ther 2023 May 26;44(2):127-144. [doi: 10.1002/anzf.1534]
131. Osborn TL, Rodriguez M, Wasil AR, Venturo-Conerly KE, Gan J, Alemu RG, et al. Single-session digital intervention for adolescent depression, anxiety, and well-being: outcomes of a randomized controlled trial with Kenyan adolescents. J Consult Clin Psychol 2020 Jul;88(7):657-668. [doi: 10.1037/ccp0000505] [Medline: 32391709]
132. Wasil AR, Park SJ, Gillespie S, Shingleton R, Shinde S, Natu S, et al. Harnessing single-session interventions to improve adolescent mental health and well-being in India: development, adaptation, and pilot testing of online single-session interventions in Indian secondary schools. Asian J Psychiatr 2020 May;50:101980. [doi: 10.1016/j.ajp.2020.101980] [Medline: 32146337]
133. Wasil AR, Taylor ME, Franzen RE, Steinberg JS, DeRubeis RJ. Promoting graduate student mental health during COVID-19: acceptability, feasibility, and perceived utility of an online single-session intervention. Front Psychol 2021;12:569785 [FREE Full text] [doi: 10.3389/fpsyg.2021.569785] [Medline: 33897511]
134. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
135. Musiat P, Hoffmann L, Schmidt U. Personalised computerised feedback in E-mental health. J Ment Health 2012 Aug;21(4):346-354. [doi: 10.3109/09638237.2011.648347] [Medline: 22315961]
136. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
137. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016 Jan 08;18(1):e6 [FREE Full text] [doi: 10.2196/jmir.4790] [Medline: 26747176]
138. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
139. Murray E, White IR, Varagunam M, Godfrey C, Khadjesari Z, McCambridge J. Attrition revisited: adherence and retention in a web-based alcohol trial. J Med Internet Res 2013 Aug 30;15(8):e162 [FREE Full text] [doi: 10.2196/jmir.2336] [Medline: 23996958]
140. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
141. Garett R, Chiu J, Zhang L, Young SD. A literature review: website design and user engagement. Online J Commun Media Technol 2016 Jul;6(3):1-14 [FREE Full text] [Medline: 27499833]
142. Shim M, Mahaffey B, Bleidistel M, Gonzalez A. A scoping review of human-support factors in the context of internet-based psychological interventions (IPIs) for depression and anxiety disorders. Clin Psychol Rev 2017 Dec;57:129-140. [doi: 10.1016/j.cpr.2017.09.003] [Medline: 28934623]
143. Inal Y, Wake JD, Guribye F, Nordgreen T. Usability evaluations of mobile mental health technologies: systematic review. J Med Internet Res 2020 Jan 06;22(1):e15337 [FREE Full text] [doi: 10.2196/15337] [Medline: 31904579]
144. Collins DA, Tully LA, Piotrowska PJ, Hawes DJ, Dadds MR. Perspectives on ParentWorks: learnings from the development and national roll-out of a self-directed online parenting intervention. Internet Interv 2019 Mar;15:52-59 [FREE Full text] [doi: 10.1016/j.invent.2018.12.002] [Medline: 30656140]
145. Corrigan PW, Miller FE. Shame, blame, and contamination: a review of the impact of mental illness stigma on family members. J Ment Health 2009 Jul 06;13(6):537-548 [FREE Full text] [doi: 10.1080/09638230400017004]
146. Ennis E, Bunting BP. Family burden, family health and personal mental health. BMC Public Health 2013 Mar 21;13(1):255 [FREE Full text] [doi: 10.1186/1471-2458-13-255] [Medline: 23517472]

XSL•FO
RenderX

Abbreviations

DMHI: digital mental health intervention
LMICs: low- and middle-income countries
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Mavragani; submitted 29.05.23; peer-reviewed by M Ali, M Loades, F Tong; comments to author 01.12.23; revised version received 13.12.23; accepted 04.05.24; published 03.07.24.

Please cite as:
Welsh ET, McIntosh JE, Vuong A, Cloud ZCG, Hartley E, Boyd JH
Design of Digital Mental Health Platforms for Family Member Cocompletion: Scoping Review
J Med Internet Res 2024;26:e49431
URL: https://www.jmir.org/2024/1/e49431
doi:10.2196/49431
PMID:

©Ellen T Welsh, Jennifer E McIntosh, An Vuong, Zoe C G Cloud, Eliza Hartley, James H Boyd. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Cultural and Contextual Adaptation of Digital Health Interventions: Narrative Review

Aila Naderbagi[1], PhD; Victoria Loblay[1], BA, MA, PhD; Iqthyer Uddin Md Zahed[1], BA (Hons), MA, MBA, PhD; Mahalakshmi Ekambareshwar[1], BSc, Grad Dip (HSM), MHS (Hons), PhD; Adam Poulsen[1], PhD; Yun J C Song[1], PhD, JD; Laura Ospina-Pinillos[2], MD, PhD; Michael Krausz[3], MD, PhD; Mostafa Mamdouh Kamel[4], PhD; Ian B Hickie[1], AM, MD; Haley M LaMonica[1], BSc, MA, PhD

[1]Brain and Mind Centre, The University of Sydney, Sydney NSW, Australia

[2]Department of Psychiatry and Mental Health, Faculty of Medicine, Pontificia Universidad Javeriana, Bogota, Colombia

[3]Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada

[4]Department of Psychiatry, Tanta University, Tanta, Egypt

Corresponding Author:
Haley M LaMonica, BSc, MA, PhD
Brain and Mind Centre
The University of Sydney
94 Mallett Street
Sydney NSW, 2050
Australia
Phone: 61 0426955658
Email: haley.lamonica@sydney.edu.au

Abstract

Background: Emerging evidence suggests that positive impacts can be generated when digital health interventions are designed to be responsive to the cultural and socioeconomic context of their intended audiences.

Objective: This narrative review aims to synthesize the literature about the cultural adaptation of digital health interventions. It examines how concepts of culture and context feature in design and development processes, including the methods, models, and content of these interventions, with the aim of helping researchers to make informed decisions about how to approach cultural adaptation in digital health.

Methods: Literature searches for this narrative review were conducted across 4 databases. Following full-text article screening by 2 authors, 16 studies of interventions predominantly focused on the self-management of health were selected based on their detailed focus on the process of cultural adaptation. Key considerations for cultural adaptation were identified and synthesized through a qualitative narrative approach, enabling an integrative and in-depth understanding of cultural adaptation.

Results: The literature demonstrates varying approaches and levels of cultural adaptation across stages of intervention development, involving considerations such as the research ethos orienting researchers, the methodologies and models used, and the resultant content adaptations. In relation to the latter, culturally appropriate and accessible user interface design and translation can be seen as particularly important in shaping the level of adaptation.

Conclusions: Optimizing cultural adaptation involves linking culture with other contextual factors such as economic conditions and social systems to ensure accessibility and the sustained use of digital health interventions. Culturally humble approaches that use the involvement of a broad range of participants, experts, and other stakeholders are demonstrated to spark vital insights for content development, implementation, and evaluation.

(J Med Internet Res 2024;26:e55130) doi:10.2196/55130

KEYWORDS

cultural adaptation; digital health; context; translation; participatory research; mobile phone

XSL•FO
RenderX

Introduction

Background

Digital health interventions have increased exponentially in recent years as a means of expanding access to health information and services, improving cost-effectiveness, and empowering people to take control over their health [1]. Decades-long debates around cultural context in public health interventions [2] are now occurring within the digital health literature with the development of culturally appropriate web- and mobile-based interventions. The cultural adaptation of existing interventions is reported to be a cost-effective and efficient way of delivering evidence-based interventions at scale [3], with emerging evidence indicating that health outcomes are more favorable when interventions are responsive to the cultural and socioeconomic context of their intended audience [4-6]. Despite the demonstrated effectiveness, there remains a general lack of guidance about how to approach the cultural adaptation of digital health interventions, which can result in poor adaptations that are unable to effectively engage their intended audiences [7].

Cultural adaptation is often viewed as a systematic process whereby intervention components identified as *cultural* are isolated and tailored to target populations, while adherence to evidence-based components is uncompromised [8,9]. In complex system thinking, however, context is not a mere backdrop to intervention implementation but actively influences intervention success [10]. Adaptation must therefore be part of digital health intervention design from the outset and not a secondary process [11]. Resnicow et al [2] have distinguished between surface and deep structure adaptation to conceptualize the depths of cultural sensitivity in public health interventions, which are also applicable in the development of digital interventions. Cultural sensitivity at a surface structure level involves the integration of observable social and behavioral characteristics among a target population, such as common places, languages, clothing, and health behavior as well as appropriate channels or settings for the delivery of messages and programs [2]. The deep structure dimension of intervention sensitivity is about the factors influencing health behaviors among different racial and ethnic populations. Deep structure adaptation is recommended when an in-depth understanding is required around matters such as how members of the target population perceive the cause, course, and treatment of illnesses. This often involves qualitative research into the ways religion, family, society, economics, and the government influence health [2].

Objective

There are no 2 cases of cultural adaptation that are alike, and documentation and reporting differ widely. This narrative review collates diverse studies on the cultural adaptation of web- and mobile-based interventions, with a focus on case examples that provide a thorough record of the stages and processes involved. The aims of this review are to (1) analyze different approaches and levels of cultural adaptation in the digital health literature and (2) examine how the concepts of culture and context feature in the methods, models, and content of these interventions. To date, systematic reviews have emphasized the effectiveness of culturally adapted digital health interventions relative to those interventions that are left unadapted [5,6]. Therefore, through a qualitative narrative synthesis of learnings about major aspects of cultural adaptation, this review offers a distinctive and novel perspective that can be used to inform decision-making in the development of culturally appropriate web- and mobile-based interventions.

Methods

Search Strategy

The narrative review approach used here aims to provide a scholarly summary and interpretation of the digital health literature, with the purpose of synthesizing and deepening the understanding of cultural adaptation in digital health. The intention is not to present a hybrid systematic narrative review or evaluate the validity of a particular method or intervention [12]. We have prioritized relevant texts to highlight distinctive contributions and major themes, rather than presenting an exhaustive overview of the entire body of literature with strict inclusion and exclusion criteria [13]. By focusing on qualitative interpretation and synthesis of the literature, our narrative review differs from, but complements, a quantitative approach [14].

The eligibility criteria for this review included (1) studies of the design, development, or evaluation of a digital health intervention for the purposes of self-management of health and (2) studies that included a detailed description of how the digital intervention had been adapted to be culturally and contextually appropriate for target end users. We conducted literature searches across 4 databases, including CINAHL, PsycINFO, Scopus, and Ovid MEDLINE, as well as 2 separate hand searches between October and November 2022. The search was limited to peer-reviewed, English language papers. The search terms included (“digital health” OR “digital intervention” OR “mHealth” OR “mobile health” OR “eHealth” OR “telemedicine” OR “telehealth” OR “parenting app” OR “parent* app”) AND (“sociocultural factors” OR “cultural sensitivity” OR “cultural adaptation” OR “foreign language translation” OR “local context” OR “cross cultural test adaptation” OR “cultur* appropriate*”). No publication date limit was set, but given that digital health interventions have gained traction relatively recently, most articles were published from 2018 onward. Multimedia Appendix 1 provides the specific search strategy for each database.

Screening

Search queries varied slightly based on the constraints of the selected databases, yielding a total of 628 papers. Abstracts of these papers were discussed in team meetings (VL, ME, AN, and IUMZ), where we applied a hermeneutic approach to honing the literature search, classifying and critically assessing the papers for sorting and screening [15]. Importantly, a hermeneutic approach is an iterative process guided by judgment, reading, interpretation, and tacking back and forth between the literature and the search approach [15]. Of the interventions reviewed in the initial search cycle, we sorted those that focused on the self-management of health, screening out those that focused on health care service delivery. As we were interested in deepening our appreciation for how interventions approached cultural

XSL•FO
RenderX

adaptation, we also favored articles that presented thorough documentation of adaptation processes and considerations. A total of 113 studies were gathered in the initial screening and 2 authors (AN and IUMZ) engaged in a second cycle of iterative analysis, reading the texts in full and enriching their critical assessment of interventions using cultural adaptation. The 2 authors assessed the eligibility of the 113 studies independently by dividing the work evenly. This was carried out within the same workspace where they could readily discuss questions regarding eligibility as they arose. As the overarching criteria for paper selection were established prior, the minor questions that arose were resolved through discussion between the 2 authors with reference to the previously stated eligibility criteria. A final set of 16 studies were selected for full analysis, based on their meaningful engagement with contextual, methodological, and content-based factors in adaptation processes, with eligibility confirmed jointly by senior researcher authors (VL and HML).

Data Extraction

The studies were analyzed for information about intervention purpose, target population, context, adaptation processes, and modifications of intervention components. All data were extracted by the lead author (AN) with oversight from senior researcher authors (VL and HML). Our approach was grounded in a qualitative mode of analysis, which combines a case-specific, detailed view with a focus on interconnectivity across cases for a better picture of the ways culture and context are conceptualized. These considerations are presented through a narrative synthesis approach [14], which enables an integrative and in-depth understanding of cultural adaptation in digital health.

Results

Overview

The literature presented covers both digital health interventions that have been developed without a preexisting template and those adapted from an original template. There are diverse cultural contexts represented across the studies, with some targeting racial and ethnic minority groups within high-income countries such as the United States, and others adapted for low- and middle-income countries (Table 1). The interventions range from pregnancy and parenting programs to those aimed at the prevention and treatment of mental health disorders. Moreover, there is a high representation of health literacy, health information, and health promotion interventions. Most of these are mobile health (mHealth) apps used on smartphones or tablets (12/16, 75%), while a small number are web-based interventions (4/16, 25%). By presenting diverse examples of the in-depth processes underpinning digital health cultural adaptation, we have intended to show that this is an undertaking relevant for a wide range of such interventions.

XSL•FO
RenderX

Table 1. Interventions featured in the review.

Study, year	Intervention	Adapted content	Models and methods
Abuwalla et al [16], 2017	Project Competent Adult Transition with Cognitive-behavioural, Humanistic, and Interpersonal Training: a web-based depression prevention intervention adapted for Arab adolescents and young adults	3 cultural domains of identity, empowerment, and expectations applied to website content	The PEN-3 theoretical framework, using 3 domains of adaptation
Alsswey et al [17], 2020	(Unnamed) app for managing medication information and delivering health information across Arab countries	Images, colors, layout, and language	Analysis of app user interface design
Bailin and Bearman [18], 2022	Parenting A to Z: a digital single-session positive parenting intervention for low-income, Latinx families in the United States	Language, persons, context, content, and metaphors	A 2-phased, experimental mixed methods design with a cultural sensitivity framework
Bender et al [19], 2016	(Unnamed) physical activity mHealth[a] app for the Latino community in the United States	Cultural values in images and text and other culturally appropriate content; communicating information based on general literacy levels in the population	A 3-phased study exploring perceptions of images of the app; using findings to design and develop the app; testing the app for feasibility and efficacy
Black et al [20], 2018	Circle of Life: a web-based multimedia sexual health curriculum for American Indian and Alaska Native communities	Mode of intervention delivery and stories	Community-based participatory research
Doty et al [21], 2020	Padres Informados/Jovenes Preparados: mHealth app for Latinx parents of adolescents	Human-centered design principles applied to content design	Community-based participatory research approach, including surveys, interviews, and workshops across different stages
Gance-Cleveland et al [22], 2021	StartSmart: a perinatal outcomes mHealth app adapted for pregnant Hispanic women	Culturally appropriate translation of app from English to Spanish; consideration of conceptual validity of content	The Beaton process, involving 6 stages of translation
Garner et al [23], 2021	(Unnamed) A hypertension health literacy app for use in India	App content not discussed	Theories of cultural humility and social innovation collaboration
Hong et al [24], 2022	Wellness Enhancement for Caregivers: an mHealth program for Chinese American family caregivers of persons with dementia	Cultural values in translation and content; adapting for limited English proficiency and limited knowledge of health systems	Community-based, user-centered design principles
Kim et al [25], 2020	Health Club: a health promotion app for middle-aged, Korean Chinese migrant women in Korea	Content is not described	Living laboratories and intervention mapping methods
Mairs et al [26], 2020	Play Kindly: a cross-platform parenting app for Pacific and Māori families	App interface, scenario scripts, narration, and artwork; involvement of Pacific professionals in app development	Pacific and Māori research methodologies and qualitative methods to evaluate the app
Mueller et al [27], 2020	Maternal and Neonatal Technologies in Rural Areas: a neonatal game app for mothers in rural Nepal	Culturally appropriate images and audio	Co-design and cocreation
Abi Ramia et al [28], 2018	Step-by-Step: a digital mental health intervention adapted for Palestinian, Syrian, and Lebanese communities in Lebanon	Language of content; characters and illustrations; psychological concepts	Bottom-up process of adaptation using cognitive interviewing techniques
Ruvalcaba et al [29], 2019	Visit Planner: an English iPad tool for patient use before primary care visit, adapted for the Spanish-speaking Latino population in the United States	Culturally relevant concepts, images, and characters; adapting for literacy level; tackling stigma; use of peer-to-peer advising	3-step “transcreation” model (“translation” and “creation”)
Sit et al [30], 2020	Step-by-Step adapted for Chinese young adults	8 culturally sensitive elements were chosen as targets of adaptation: language, people, metaphors, content, concepts, goals, methods, and context	Ecological validity model
Smith et al [31], 2014	GetHealthyHarlem: a health literacy website for racial and ethnic minority groups in Harlem, New York	Adapting content for literacy level; short articles, interviews, and anecdotes from community members and local health professionals; features fostering local interactions	Community-based participatory research

[a]mHealth: mobile health.

XSL•FO
RenderX

Figure 1 depicts the layers of considerations and approaches to cultural adaptation in digital health. The first factors to consider relate to the research ethos, including how to navigate cultural differences within the research team as well as whether the use of culturally specific research methods is recommended. Next, we explore frameworks and models applied in adaptation, with particular attention to the cultural sensitivity framework, the ecological validity model, and the PEN-3 model. The third consideration is specific methodological strategies such as participatory research approaches used for accessing cultural insights and localizing global interventions. As the final aspect, we consider specific design-related aspects of adaptation, with a specific focus on language translation and user interface (UI) design.

Figure 1. Considerations and approaches to cultural adaptation in digital health.

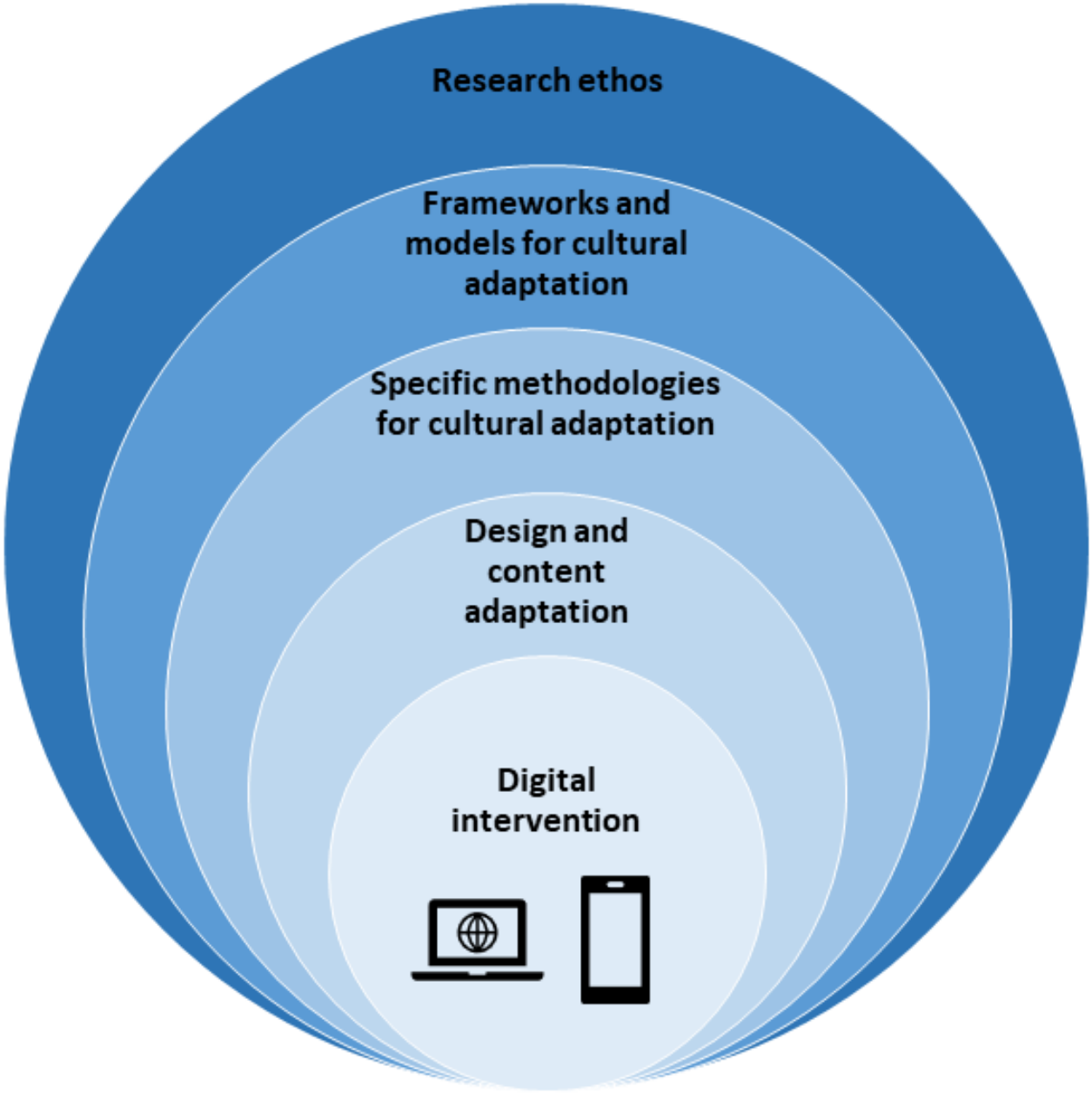

Research Ethos

Navigating Cultural Differences Within Research Teams

Social engagement and collaboration across cultural differences within research teams and partnerships are foundational aspects of cultural adaptation. Garner et al [23] drew on the theory of cultural humility in their evaluation of the development of a hypertension health literacy app for use in India. Cultural humility can be distinguished from cultural competence in terms of being an active commitment to self-awareness and ongoing learning on the part of practitioners, rather than an end point of mastering cultural differences as implied by the term competence [32]. Garner et al [23] showed that cultural humility can be sustained within research teams through continuous reflection of individual and group interactions, the acknowledgment and mitigation of challenges, willingness to compromise and adapt perceptions and processes, as well as

XSL•FO
RenderX

strengthening partnerships. Furthermore, cultural humility makes room for the identification of local cultural values that may differ from the developer's perspective. Recognition of the importance of communal decision-making in India resulted in app design changes including a 3D animation that depicts an individual talking about hypertension with family members and friends in a rural community setting [23]. The integration of such values as communal learning into intervention development will be further discussed in subsequent sections, as it is a distinct challenge for cultural adaptation given the individualized format of most digital interventions.

Culturally Specific Research Methods

The literature shows that some research contexts call for the use of culturally specific methods and protocols. This is evident in the evaluation of the first parenting app designed for a contemporary Pacific audience in New Zealand [26]. Play Kindly uses gamification to teach parenting strategies for everyday behavior problems such as fights over belongings, refusing to eat, and temper tantrums. Here, culturally appropriate research involved the inclusion of Pacific professionals in the creation of the app and also drew on Pacific and Māori research methods and guidelines (ie, Kaupapa Māori research methods ensures research is conducted "by Māori, for Māori, with Māori") [33] in the qualitative evaluation study of the app, including a welcome ceremony, food sharing, and setting the scene for the research [26]. These are framed around Pacific cultural values of communal relationships involving reciprocity, holism, and respect, while also integrating established concepts such as Teu le Vā—which involves identifying and understanding the "vā," or spaces, between different stakeholders in Pacific research—and sharing knowledge across these spaces [26]. Notably, the conduct of evaluation research within such a cultural framework remains unique in the cultural adaptation literature.

Models and Frameworks for Guiding Cultural Adaptation

Overview

Some studies in the literature use frameworks as a guide for the adaptation process, while others use adaptation models for a more systematic or staged approach for tailoring an intervention to cultural context. The extent of documented details, processes, methods, and content modifications can differ widely based on which frameworks or models are used. In subsequent sections we describe 3 of these frameworks as applied in the literature.

Cultural Sensitivity Framework

The cultural sensitivity framework was conceptualized by Resnicow et al [2] in the context of substance use prevention interventions. This framework is characterized by surface and deep structure adaptations and focuses on the involvement of stakeholders and end user populations to ensure cultural fit. Bailin and Bearman [18] applied this framework to enhance the customization, acceptability, and validity of evidence-based positive parenting practices presented in video format for low-income Latinx families in the United States. After conducting an initial needs assessment, researchers embedded themselves in the clinic to gather stakeholders' input and gain a basic understanding of the procedures involved on site. Interviews with caregivers were then conducted once a prototype was made to make surface structure adaptations, including changes to the language, persons, and metaphors used in the intervention [18].

Ecological Validity Model

Phased approaches to cultural adaptation are common with the use of adaptation models, with one widely recognized example being the ecological validity model. This addresses 8 culturally sensitive domains in an intervention: language, people, metaphor, content (eg, cultural knowledge about values and customs), concepts (eg, treatment constructs that resonate locally), treatment goals, methods and procedures used to accomplish the treatment goals, and context [30]. Following this model, the adaptation of Step-by-Step for the Chinese young adult population in mainland China and Macao was based on a 4-phased approach [30]. The adaptation process was shaped by the lived realities of Chinese young adults concerning accepted ways of social interaction, mental health challenges experienced at this life stage, societal norms around mental health such as stigma, and content preferences based on gender. As an example, in focus group discussions, experts and laypersons suggested that the intervention would be more effective framed as a stress management tool or an academic performance tool, rather than a mental health intervention, to reduce the stigma associated with using the latter [30]. Despite drawing on the systematic approach of the ecological validity model, cultural adaptation often requires multiple iterations to ensure interventions are applicable for and relevant to potential end users. In the example of the Chinese adaptation of Step-by-Step, it was noted that the voices of more diverse young adults, including those who live with higher levels of depression would likely improve intervention acceptability [30].

PEN-3 Model

The PEN-3 model was developed to address the omission of culture in explaining health outcomes in existing health behavior theories and models [16]. The model has three interrelated dimensions: (1) cultural identity, (2) relationships and expectations, and (3) cultural empowerment. Each includes 3 facets that form the acronym PEN: person, extended family, and neighborhood (cultural identity domain); perceptions, enablers, and nurturers (relationship and expectations domain); and positive, existential, and negative (cultural empowerment domain). In this way, the model centralizes culture in the study of health beliefs, behaviors, and outcomes, as well as in the development, implementation, and evaluation of public health interventions [34].

By focusing on the contextual factors shaping daily life and relationships in the Arab world, the PEN-3 model enabled deep adaptation of Competent Adulthood Transition with Cognitive-behavioral, Humanistic, and Interpersonal Training, a web-based depression prevention intervention for at-risk adolescents and young adults [16]. This included addressing how the therapeutic approach can be made to fit a family-centered worldview, rather than the Western therapeutic emphasis on the individual. It also allowed norms around social relationships to be incorporated into the intervention, for

XSL•FO
RenderX

example, by reducing stories that feature interactions between men and women. Moreover, the stigma around mental illness and its links to concepts of shame and religious beliefs were addressed through this model by including religious verses that promote seeking treatment for mental health disorders and discouraging labeling of patients with psychiatric conditions. The internet format is an advantage here as it offers a treatment option where no private information needs to be disclosed, thus serving the cultural value of protecting family honor and dignity [16]. This resonates with survey findings by Kamel et al [35] where psychiatrists indicated privacy, confidentiality, and security to be the most important barriers to the implementation of digital mental health interventions in Middle Eastern countries such as Egypt.

As the local adaptations of Step-by-Step and Competent Adulthood Transition with Cognitive-behavioral, Humanistic, and Interpersonal Training demonstrate, an original template can undergo substantial changes once the norms and preferences of local community members and experts are integrated through a cultural adaptation model. While not all interventions deploy a model, these may lead to more thorough adaptations of content, including language and UI design, which we focus on next [36].

Methodologies of Adaptation

Methods for Accessing Cultural Insights

Cultural adaptation is a response to community-based differences in the ways health issues manifest or get addressed, making engagement with target communities a common aspect of intervention development. This is approached through diverse methods in the literature, most of which are situated in the community-based participatory research (CBPR) approach. CBPR can be understood as a research paradigm where local communities are recognized as entities with knowledge, expertise, and resources and as partners in the research process [21].

The CBPR methodology can be used in digital health programs to address issues of access and trust in health for racial and ethnic minority communities [21]. The concept of trust is significant as it relates to both socioeconomic circumstances and manifests in culturally specific ways. This is evident in the study of an mHealth app based on an in-person parenting program aimed at reducing substance use among Latinx youth. The app was developed using iterative human-centered design processes integrating feedback from stakeholders and end users and was informed by The Centre for eHealth Research and Disease Management road map for designing digital health technology [21]. This approach aided in building trust through endorsing and centering the knowledge of community stakeholders, prevention program facilitators, and parents in app development and introducing the critical analysis of cultural needs [21]. Through the CBPR process, community facilitators of the program became cocreators of content for the mobile app version, which was then validated through interviews with parents. The conduct of participatory research in the language of participants here and in other examples ensures essential information is not lost in translation, while also saving time and resources [37]. The resulting design adaptations included a minimalist esthetic, clear navigation, structural consistency, a reduction in the length of videos, and inclusion of transcripts, as well as addressing internet connectivity issues, all of which are aimed at ensuring accessibility for the audience.

The multimethod approach of “living labs” is another way to understand and address the health needs of vulnerable communities [25]. In the development of a health promotion app for middle-aged, Korean Chinese migrant women, Kim et al [25] applied core principles derived from the living laboratories approach, namely user and stakeholder participation; basis in real-life settings; and cocreation to identify stakeholder needs, set objectives, identify relevant intervention strategies, and consider design options. The intervention was designed to be culturally appropriate by virtue of attending to the challenges faced by this demographic in Korea, including experiences of social isolation alongside language barriers and changes in socioeconomic status due to migration. Moreover, the phrase “cultural adaptation” was also thought to apply to users who were provided with information about life in Korea to encourage a greater sense of belonging [25]. In this sense, through the means of tailored health promotion, the intervention sought to address ways users can also culturally adapt.

Participatory Methods in Localizing Global Interventions

While the interventions discussed thus far were originally developed for a local context, others have been developed as scalable templates for a global audience and later adapted for local contexts using participatory methods. A major example of this is Step-by-Step, a transdiagnostic and guided digital mental health intervention that is supported by an app-based delivery modality, targeting common mental health disorders such as depression. The intervention was developed through a top-down process involving the World Health Organization; government bodies and other large organizations as a global public good with the intervention content (eg, stories, illustrations, and videos); guidance model; and mobile app being culturally adaptable [28,30,38-40]. CBPR methods such as cognitive interviewing [4] and focus group discussions have been used in different cultural adaptations of the intervention to make its content relevant for local communities. The involvement of local stakeholders in the adaptation process is crucial for identifying and reworking concepts derived from Western science used in the intervention, namely those around mental illness and mental health. Local knowledge of the influence of contextual factors on the issues explored in the intervention means the target audience is more likely to see their experiences reflected in the intervention, thus allowing for the relevant issues to be addressed more effectively.

Indeed, the Lebanese adaptation of Step-by-Step highlights how mood disorders are affected by social, economic, and health challenges in this content [28]. The explanation of depression symptoms was changed to reflect local understandings of how depression is experienced: maintaining daily activities and work but displaying higher levels of anger, irritation, and anxiety. Furthermore, clinical terms such as “mental illness” were not used in the intervention stories, rather terms such as “distress” or local idioms such as “tired psyche” were used instead [28]. The role of faith and religion in helping people deal with mental

XSL•FO
RenderX

difficulties was integrated into the activities after adaptation. Finally, gendered stereotypes were also tempered to depict cultural norms as recommended by local participants. Specifically, some changes were made to challenge the stereotypical depiction of women as submissive and expressing distress through crying or the depiction of men as expressing distress through anger; however, other gendered elements were maintained to ensure the content resonated with users. These modifications align with observations by Kamel et al [35] who stated that before the use of Western interventions in the Middle East, it is crucial to consider incompatibilities around gender-related norms, public awareness, and stigma of mental disorders, as well as language and presentation of distress.

Intervention Design and Content Adaptations

Culturally Adapted Translation

Language translation is a crucial component of cultural adaptation in digital interventions, as language discrepancy and cultural differences in understandings of health or illness often pose barriers to engagement. In cases where an intervention is adapted for a new cultural context, a literal translation can result in poorly constructed sentences, mistranslations, and failure to capture the metaphors and idioms within a language [36]. The literature demonstrates that culturally adapted translation can avoid these by involving processes such as forward and back translation, as well as reviews and syntheses of different versions. In Textbox 1, we provide an example of this by summarizing the Beaton process of translating and adapting StartSmart, a perinatal outcomes app helping pregnant women screen for risk and protective factors through a self-administered questionnaire [22].

Textbox 1. The Beaton process of culturally adapted translation.

Stage and process

- 1: Forward translation of the screening instrument and patient education materials by 2 bilingual translators.
- 2: Synthesis of translations through consensus by translators and the primary investigator.
- 3: Back translation into English by 2 different translators with English as their first language who did not see the original documents.
- 4: An expert committee review conducted with particular attention to capturing the nuances of the Spanish language. Consensus reached on the pretest version.
- 5: The app was pretested through 17 interviews with patients, and a focus group discussion with providers focused on usability and acceptability.
- 6: Appraisal of StartSmart carried out to determine whether the Beaton process was followed to translate and culturally adapt the system and consider refinements to the technology.

The “transcreation” process is another example of culturally adapted translation achieved through hiring research team members such as health educators and cultural brokers with appropriate linguistic and cultural expertise, careful assessment of intervention elements for their potential cultural significance, and iterative testing and assessment of the resulting changes [29]. This approach is particularly relevant when cultural adaptation predominantly centers on content wording and visual modifications. For example, a core cultural construct that impacts health-seeking behavior in the Latino context is “familismo” or family. This was embedded in the content of an mHealth intervention by changing the app’s menu wording, such as “Stress at Home or Work” to “Mi Familia” or my family [29].

Adaptation of UI Design

The UI is the point of contact and communication between users and digital interventions and is greatly emphasized by designers in relation to user experience and user acceptance [17]. UI includes the display screens and the way through which a user interacts with an app or website, such as voice- or touch-enabled features, multimedia graphics, and others. In this context, images are more than mere objects but rather cultural artifacts that represent values, ethos, or ideologies [17]. Indeed, the visual depiction of cultural values can be an effective way to increase engagement. To that end, familial cohesion and social support were centered visually in a low-text physical activity app designed for the Latino community, through the depiction of family members, friends, and peers exercising together during culturally relevant physical activities such as dancing and football [19]. Importantly, Alsswey et al [17] found a strong correlation between culturally adapted UI design and app use, in that users felt more confident when navigating technology that reflects familiar elements reflected in images, colors, language, and layout.

The principles of cocreation and co-design are vital for achieving these ends. Mueller et al [27] define cocreation as collective creativity and co-design as the application of this principle across the design process, which extends decision-making and stakes in the end product to a broader group. In the development of their gamified mHealth intervention for a low-literacy adult audience in rural Nepal, UI features such as pictograms were co-designed with in-country partners and artists and further refined through field evaluations with end users [27]. The attention to detail in the co-designed pictograms (eg, symbolic clothing colors representing marital status) was described as a crucial aspect of intervention localization [27].

Adapting Intervention Format

Establishing cultural relevance around values is not always clear cut, particularly when it comes to factors such as intervention format and the ways new knowledge is presented to users. In an evaluation study of the gamified intervention Play Kindly

XSL•FO
RenderX

discussed previously, Mairs et al [26] pointed out disagreements among participants about how culture specific the app should be based on how strongly the person identified with their culture. This can be characterized by differences between a participant who believes parenting is not shaped by culture and one who may identify cultural influences in parenting practices. While the Pacific theme of the UI and the real-life scenarios were widely accepted by users, the approach of providing parenting information via a virtual avatar coach (referred to as "the educator") stood out as a point of contention by leading some participants to feel as though they were making parenting mistakes. The gamified format of the app also faced mixed reactions given the rule of having to answer correctly to progress through the game, which similarly impacted users' view of themselves as parents [26]. These factors demonstrate that the mode of information delivery, and the overall format of an intervention can be major drawcards for user engagement if aligned with cultural beliefs.

Despite the advantages of digital health interventions regarding accessibility and cost-effectiveness, the impersonal technologies can be culturally inappropriate in cultures where knowledge has traditionally been transmitted in communal settings. This is particularly significant in the context of coming of age and sexual health, which are seen as a community responsibility among American Indian and Alaska Native people [20]. Community partners involved in the development of a web-based, multimedia sexual health curriculum highlighted the need for the interpersonal approach of traditional knowledge sharing to be reflected in the digital mode of intervention delivery [20]. As a response, a hybrid version of Circle of Life was created, combining web-based lessons with class-based activities and discussions. However, contextual factors related to poverty had direct consequences for engagement, such as logistical challenges, poor internet, and low class attendance. Therefore, the intervention had to be made flexible enough so that it was responsive to local circumstances according to the discretion of facilitators, demonstrating that cultural adaptation must also integrate socioeconomic context to ensure accessibility.

Discussion

Overview

The methodologies, models, and content adaptations we have examined demonstrate the wide range of approaches to cultural adaptation. It is evident that while an adaptation framework or model is not a necessary prerequisite for adaptation, an iterative process with multiple stages of intervention development, testing, and refinement is essential particularly when participatory research methods are used. The extent of community participation and the depth of cultural adaptation tends to be based on a range of factors such as intervention goals, research questions, processes, and the context of intended users. Some projects also include key stakeholders representing governmental and nongovernmental organizations, as well as other local experts as part of participatory research.

The involvement of a broad range of participants can lead to more successful adaptations by adding important insights for content development and implementation [4]. This is pertinent given that consensus does not always exist around cultural relevance within a community, as people may have different opinions on values and norms [26]. Appreciation of cultural heterogeneity within cultural groups is vital to avoid insensitive material in interventions. This can be ensured by incorporating multiple perspectives that appeal to a broad spectrum of the target population [2]. When a wide scope of coverage is intended, common factors must be deduced based on key themes and patterns identified in research, hence why the inclusion of cultural expertise can be vital. For instance, in cases where one language is shared across different settings (eg, Spanish), it may make sense to use common activities and examples and provide neutral recommendations to enable an intervention's applicability across different countries [41].

Surface and Deep Structure Adaptation

Our review highlights that adaptation occurs on a spectrum, ranging from surface to deep level adaptations. Specifically, surface level adaptations are often integrated in the UI, involving components such as labels and illustrations. As demonstrated in many interventions in this review, such components of the UI can affect the level of usability and engagement based on whether its elements resonate with the preferences of an audience as influenced by cultural context [17]. It is essential to integrate user feedback to ensure different components in the UI are culturally appropriate, and the layout is easy to navigate for different digital literacy levels. Deep adaptation goes beyond the visual aspects of an intervention or the literal translation of its content to include culturally specific perspectives on major concepts in an intervention. This is especially pertinent in relation to interventions based on culturally sensitive topics such as mental illness, as evidenced by the integration of local understandings of mental health disorders by Ramia et al [28].

Deep adaptation is necessary if case studies and stories are to reflect the cultural characteristics of relationships, such as gender norms or the importance of family and the local community for personal health [16,29]. Furthermore, collective knowledge sharing and learning is a recurring cultural value that can impact user engagement, as it challenges the conventional focus on the individual in digital interventions. Therefore, deep cultural adaptation also involves the integration of such values in intervention design. This approach to adaptation requires a deeper level of cultural sensitivity as it reframes intervention components that may have been previously assumed to be culturally neutral and not in need of adaptation.

When decisions must be made by researchers and developers between surface and deep structure adaptation, we recommend considering the distinction between literal translation and culturally appropriate translation as a guide. Spanhel et al [7] grouped all language translations as a surface structure adaptation, but we suggest that culturally adapted translation, which includes locally specific idiomatic expressions and metaphors, can be classified as deep structure adaptation. Culturally appropriate translation not only reduces the likelihood of mistranslations but also enables different perspectives and values around concepts to be presented in the content [36]. Deep translation draws on cultural expertise and can involve multiple

XSL•FO
RenderX

stages including back translation, synthesis of different translations, and consensus on the choice of translation [22]. This is a vital aspect of cultural adaptation given that digital interventions often rely on text to convey information. If text is convoluted or does not reflect local language norms, it can swiftly discourage user engagement.

Importantly, the notion of culture in adaptations of digital interventions is primarily about the familiarity, relatability, and accessibility of an intervention for users in both cases of surface and deep adaptation. This distinction around depth of cultural sensitivity and adaptation reveals that cultural adaptation cannot be prescribed beyond general guidelines, but rather must be based on contextual circumstances around an intervention.

The Intersection of Socioeconomic and Cultural Factors

Cultural adaptation means reconceiving how we think about interventions and their interactions with context. May et al [10] have conceptualized context as a *process* rather than a *place*, suggesting that the context in which intervention implementation occurs requires ongoing work to hold together and advance. Cultural adaptation enables analysis of the ways digital interventions interact with particular local contexts [40]. This involves not only actively integrating key elements of context into interventions but also exploring the ways these are implemented as users and stakeholders may themselves engage in processes of cultural adaptation once interventions are rolled out [42].

The studies examined in this review show that culture interacts dynamically with other contextual factors such as economic conditions and social systems that impact the accessibility and sustained use of interventions. Low literacy resulting from poverty or migration is a case in point. Many of the interventions featured have placed great emphasis on ensuring content is accessible by simplifying the UI design, reducing text length, and avoiding jargon words [19,21,24]. Other socioeconomic challenges such as isolation and stigma are also addressed as part of cultural adaptation, for example, by including features that encourage social interaction in the community [31], or offering alternative ways to discuss stigmatized topics [16]. Barriers to digital sources due to limited technological knowledge, mobile data costs, and the lack of devices are further contextual issues related to accessibility that are addressed alongside culture in these adaptations [41]. Given that digital interventions are often developed as a response to health challenges linked to socioeconomic factors, cultural adaptation is more likely to be effective when addressing culture outside of a vacuum. Detailed documentations of cultural adaptation create further opportunities to tease out the dynamic connections that constitute context.

Limitations

Given that this search had no specified date range, the literature surfaced during this review was very comprehensive in nature, enabling an in-depth exploration of the considerations and methods for cultural adaptation in digital health. It is important to note that the search was performed in November 2022 meaning that studies of cultural adaptation conducted more recently were not included in this narrative synthesis. However, in keeping with a hermeneutic approach described by Boell and Cecez-Kecmanovic [15], we reached saturation in terms of understanding how concepts of culture and context featured in design and develop processes and interventions; therefore, extending the search to the present would not add to the already rich themes.

Conclusions

Digital health interventions offer potential for direct, cost-efficient, and convenient ways to address local and global health challenges. Cultural adaptation is a key means of increasing the engagement of target communities so that this potential may be realized. While cultural adaptation has been critiqued as limiting who can engage with an intervention in terms of ethnicity, age, or income [18], this narrative review shows that adapted interventions are usually aimed at addressing issues in public health systems where preexisting interventions are failing to reach communities. This is true both globally in terms of the digital health gap in low- and middle-income countries and within countries with regard to marginalized communities and segments of the population such as youth and caregivers [43].

When embarking on cultural adaptation, decisions around the level of adaptation must consider the purpose of the intervention, complexity of its content, and the context of implementation. In this regard, we argue that wider project considerations such as budget and timelines should be tailored for adaptation, rather than determining adaptation decisions. Moreover, establishing what is culturally relevant is not straightforward. We suggest culturally humble approaches that use the involvement of a broad range of participants, experts, and other stakeholders as these can lead to more successful adaptations by sparking important insights for content development and implementation [4]. Where feasible, long-term research (eg, ethnographic fieldwork) is recommended as a way of gaining deeper understanding of local contextual meanings and conditions. The participatory engagement of stakeholders can be best harnessed when insights are iteratively fed back across all stages, from development to validation and evaluation phases.

Rigorous studies of cultural adaptation in the digital context remain scarce. The field of digital health would benefit from more studies documenting processes of cultural adaptation and implementation, including feasibility and effectiveness, as well as evaluation to establish impact. Similarly, more studies are welcomed that interrogate the relationships between culture and the broader context in which people interact with digital resources. This research is critical as the impacts of interventions are influenced by the contexts around implementation and use.

XSL•FO
RenderX

Acknowledgments

This research was conducted by the University of Sydney's Brain and Mind Centre pursuant to an agreement between the University and Minderoo Foundation Limited (Minderoo).

The authors would like to thank Mr Adam Yoon for his contributions to the initial literature search that helped inform the conceptualization of this narrative review.

IBH is supported by a National Health and Medical Research Council L3 Investigator Grant (GNT2016346).

Data Availability

Data sharing is not applicable to this paper as no data sets were generated or analyzed during this study.

Authors' Contributions

The initial literature search was conducted by AN and IUMZ. Screening criteria were established by AN, IUMZ, ME, and VL. Abstract and full-text reviews were conducted by AN and IUMZ, with oversight by ME and VL. Data extraction and synthesis were conducted by AN, with guidance and expert input from VL and HML. Additional expert perspectives were provided by AP, YJCS, LOP, MK, MMK, and IBH. Scientific oversight and guidance were provided by IBH to ensure all activities were conducted responsibly and in a culturally appropriate manner. All authors contributed to and have approved the final manuscript.

Conflicts of Interest

IBH is the codirector, health and policy at the Brain and Mind Centre, University of Sydney, Australia. The Brain and Mind Centre operates early-intervention youth services at Camperdown under contract to headspace. He is the chief scientific advisor to, and a 3.2% equity shareholder in, Innowell Pty Ltd, which aims to transform mental health services through the use of innovative technologies. VL is a board member for Matana Foundation, a philanthropic organization that provides funding to programs for disadvantaged young people in Australia. She does not receive any financial benefit for this role. HML is a section editor for JMIR Aging but does not receive preferential treatment in the peer review process for the JMIR suite of journals. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Database search strategy.
[DOCX File , 14 KB - jmir_v26i1e55130_app1.docx]

References

1. Koh A, Swanepoel DW, Ling A, Ho BL, Tan SY, Lim J. Digital health promotion: promise and peril. Health Promot Int 2021 Dec 13;36(Supplement_1):i70-i80 [FREE Full text] [doi: 10.1093/heapro/daab134] [Medline: 34897444]
2. Resnicow K, Soler R, Braithwaite RL, Ahluwalia JS, Butler J. Cultural sensitivity in substance use prevention. J Community Psychol 2000 May;28(3):271-290. [doi: 10.1002/(sici)1520-6629(200005)28:3<271::aid-jcop4>3.0.co;2-i]
3. Kumpfer K, Magalhães C, Xie J. Cultural adaptation and implementation of family evidence-based interventions with diverse populations. Prev Sci 2017 Aug 18;18(6):649-659. [doi: 10.1007/s11121-016-0719-3] [Medline: 27757773]
4. Naslund JA, Spagnolo J. Chapter 12 - Cultural adaptations of digital therapeutics. In: Jacobson N, Kowatsch T, Marsch L, editors. Digital Therapeutics for Mental Health and Addiction. Cambridge, MA: Academic Press; 2022:151-164.
5. Harper Shehadeh M, Heim E, Chowdhary N, Maercker A, Albanese E. Cultural adaptation of minimally guided interventions for common mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016 Sep 26;3(3):e44 [FREE Full text] [doi: 10.2196/mental.5776] [Medline: 27670598]
6. Hall GC, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016 Nov;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
7. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
8. Bernal G, Domenech Rodríguez MM. Advances in Latino family research: cultural adaptations of evidence-based interventions. Fam Process 2009 Jun;48(2):169-178. [doi: 10.1111/j.1545-5300.2009.01275.x] [Medline: 19579903]
9. Baumann AA, Powell BJ, Kohl PL, Tabak RG, Penalba V, Proctor EK, et al. Cultural adaptation and implementation of evidence-based parent-training: a systematic review and critique of guiding evidence. Child Youth Serv Rev 2015 Jun 01;53:113-120 [FREE Full text] [doi: 10.1016/j.childyouth.2015.03.025] [Medline: 25960585]
10. May CR, Johnson M, Finch T. Implementation, context and complexity. Implement Sci 2016 Oct 19;11(1):141 [FREE Full text] [doi: 10.1186/s13012-016-0506-3] [Medline: 27756414]

XSL•FO
RenderX

11. Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. J Epidemiol Community Health 2004 Sep;58(9):788-793 [FREE Full text] [doi: 10.1136/jech.2003.014415] [Medline: 15310806]
12. Lupton D. Young people's use of digital health technologies in the global north: narrative review. J Med Internet Res 2021 Jan 11;23(1):e18286 [FREE Full text] [doi: 10.2196/18286] [Medline: 33427684]
13. Loblay V, Freebairn L, Occhipinti JA. Conceptualising the value of simulation modelling for public engagement with policy: a critical literature review. Health Res Policy Syst 2023 Nov 27;21(1):123 [FREE Full text] [doi: 10.1186/s12961-023-01069-4] [Medline: 38012664]
14. Greenhalgh T, Thorne S, Malterud K. Time to challenge the spurious hierarchy of systematic over narrative reviews? Eur J Clin Invest 2018 Jun 16;48(6):e12931 [FREE Full text] [doi: 10.1111/eci.12931] [Medline: 29578574]
15. Boell SK, Cecez-Kecmanovic D. A hermeneutic approach for conducting literature reviews and literature searches. Commun Assoc Inf Syst 2014;34. [doi: 10.17705/1cais.03412]
16. Abuwalla Z, Kadhem Z, Gladstone T, Mikhael E, Bishay A, Van Voorhees BW. Proposed model for the cultural adaptation of an internet-based depression prevention intervention (CATCH-IT) for Arab adolescents. Int J Adolesc Med Health 2017 Jul 18;31(1):/j/ijamh.2019.31.issue-1/ijamh-2016-0147/ijamh-2016-0147.xml [FREE Full text] [doi: 10.1515/ijamh-2016-0147] [Medline: 28719365]
17. Alsswey AH, Al-Samarraie H, El-Qirem FA, Alzahrani AI, Alfarraj O. Culture in the design of mHealth UI: an effort to increase acceptance among culturally specific groups. Electron Libr 2020 Mar 27;38(2):257-272. [doi: 10.1108/el-04-2019-0097]
18. Bailin A, Bearman SK. Brief, digital, self-directed, and culturally adapted: developing a parenting intervention for primary care. Child Youth Serv Rev 2022 Jan;132:106314. [doi: 10.1016/j.childyouth.2021.106314]
19. Bender MS, Martinez S, Kennedy C. Designing a culturally appropriate visually enhanced low-text mobile health app promoting physical activity for Latinos: a qualitative study. J Transcult Nurs 2016 Jul;27(4):420-428. [doi: 10.1177/1043659614565249] [Medline: 25556208]
20. Black KJ, Morse B, Tuitt N, Big Crow C, Shangreau C, Kaufman CE. Beyond content: cultural perspectives on using the internet to deliver a sexual health intervention to American Indian youth. J Prim Prev 2018 Feb;39(1):59-70. [doi: 10.1007/s10935-017-0497-0] [Medline: 29344773]
21. Doty JL, Brady SS, Monardez Popelka J, Rietveld L, Garcia-Huidobro D, Doty MJ, et al. Designing a mobile app to enhance parenting skills of Latinx parents: a community-based participatory approach. JMIR Form Res 2020 Jan 24;4(1):e12618 [FREE Full text] [doi: 10.2196/12618] [Medline: 32012034]
22. Gance-Cleveland B, Leiferman J, Yates S, Williams AT, Amura CR, Roberts M, et al. Spanish translation of StartSmartTM using the beaton process to ensure tech equity. J Health Care Poor Underserved 2021;32(2S):85-105. [doi: 10.1353/hpu.2021.0052]
23. Garner SL, Koch H, George CE, Hitchcock J, Norman G, Green G, et al. Cross cultural team collaboration: integrating cultural humility in mHealth development and research. Inform Health Soc Care 2021 Dec 02;46(4):345-354. [doi: 10.1080/17538157.2021.1895168] [Medline: 33779468]
24. Hong YA, Shen K, Lu HK, Chen H, Gong Y, Ta Park V, et al. A social media-based intervention for Chinese American caregivers of persons with dementia: protocol development. JMIR Aging 2022 Sep 29;5(3):e40171 [FREE Full text] [doi: 10.2196/40171] [Medline: 36173667]
25. Kim Y, Lee H, Lee MK, Lee H, Jang H. Development of a living lab for a mobile-based health program for Korean-Chinese working women in South Korea: mixed methods study. JMIR Mhealth Uhealth 2020 Jan 08;8(1):e15359 [FREE Full text] [doi: 10.2196/15359] [Medline: 31913134]
26. Mairs RA, Bekker MJ, Patolo T, Hopkins SA, Cowley-Malcolm ET, Perese LM, et al. Gamifying parenting education using an app developed for pacific and other New Zealand families (play kindly): qualitative study. JMIR Serious Games 2020 Jun 10;8(2):e15647 [FREE Full text] [doi: 10.2196/15647] [Medline: 32519973]
27. Mueller S, Soriano D, Boscor A, Saville N, Arjyal A, Baral S, et al. MANTRA: development and localization of a mobile educational health game targeting low literacy players in low and middle income countries. BMC Public Health 2020 Jul 28;20(1):1171 [FREE Full text] [doi: 10.1186/s12889-020-09246-8] [Medline: 32723317]
28. Abi Ramia J, Harper Shehadeh M, Kheir W, Zoghbi E, Watts S, Heim E, et al. Community cognitive interviewing to inform local adaptations of an e-mental health intervention in Lebanon. Glob Ment Health (Camb) 2018 Nov 27;5:e39 [FREE Full text] [doi: 10.1017/gmh.2018.29] [Medline: 30637112]
29. Ruvalcaba D, Nagao Peck H, Lyles C, Uratsu CS, Escobar PR, Grant RW. Translating/creating a culturally responsive Spanish-language mobile app for visit preparation: case study of "trans-creation". JMIR Mhealth Uhealth 2019 Apr 05;7(4):e12457 [FREE Full text] [doi: 10.2196/12457] [Medline: 30950803]
30. Sit HF, Ling R, Lam AI, Chen W, Latkin CA, Hall BJ. The cultural adaptation of step-by-step: an intervention to address depression among Chinese young adults. Front Psychiatry 2020 Jul 7;11:650 [FREE Full text] [doi: 10.3389/fpsyt.2020.00650] [Medline: 32733296]

XSL•FO
RenderX

31. Smith M, Morita H, Mateo KF, Nye A, Hutchinson C, Cohall AT. Development of a culturally relevant consumer health information website for Harlem, New York. Health Promot Pract 2014 Sep 16;15(5):664-674. [doi: 10.1177/1524839914530401] [Medline: 24740963]
32. Tervalon M, Murray-García J. Cultural humility versus cultural competence: a critical distinction in defining physician training outcomes in multicultural education. J Health Care Poor Underserved 1998 May;9(2):117-125. [doi: 10.1353/hpu.2010.0233] [Medline: 10073197]
33. Wilson D, Mikahere-Hall A, Sherwood J. Using indigenous kaupapa Māori research methodology with constructivist grounded theory: generating a theoretical explanation of indigenous womens realities. Int J Soc Res Methodol 2021 Mar 15;25(3):375-390. [doi: 10.1080/13645579.2021.1897756]
34. Iwelunmor J, Newsome V, Airhihenbuwa CO. Framing the impact of culture on health: a systematic review of the PEN-3 cultural model and its application in public health research and interventions. Ethn Health 2014 Feb;19(1):20-46 [FREE Full text] [doi: 10.1080/13557858.2013.857768] [Medline: 24266638]
35. Kamel MM, Westenberg JN, Choi F, Tabi K, Badawy A, Ramy H, et al. Electronic mental health as an option for Egyptian psychiatry: cross-sectional study. JMIR Ment Health 2020 Aug 13;7(8):e19591 [FREE Full text] [doi: 10.2196/19591] [Medline: 32788155]
36. Havighurst SS, Choy R, Ulker A, Otterpohl N, Aghaie Meybodi F, Edrissi F, et al. A preliminary evaluation of the cultural appropriateness of the tuning in to kids parenting program in Germany, Turkey, Iran and China. Int J Environ Res Public Health 2022 Aug 19;19(16):10321 [FREE Full text] [doi: 10.3390/ijerph191610321] [Medline: 36011956]
37. Ospina-Pinillos L, Davenport T, Mendoza Diaz A, Navarro-Mancilla A, Scott EM, Hickie IB. Using participatory design methodologies to co-design and culturally adapt the Spanish version of the mental health eClinic: qualitative study. J Med Internet Res 2019 Aug 02;21(8):e14127 [FREE Full text] [doi: 10.2196/14127] [Medline: 31376271]
38. Burchert S, Alkneme MS, Bird M, Carswell K, Cuijpers P, Hansen P, et al. User-centered app adaptation of a low-intensity e-mental health intervention for Syrian refugees. Front Psychiatry 2018;9:663 [FREE Full text] [doi: 10.3389/fpsyt.2018.00663] [Medline: 30740065]
39. Garabiles MR, Harper Shehadeh M, Hall BJ. Cultural adaptation of a scalable World Health Organization e-mental health program for overseas Filipino workers. JMIR Form Res 2019 Mar 29;3(1):e11600 [FREE Full text] [doi: 10.2196/11600] [Medline: 30924784]
40. Shala M, Morina N, Burchert S, Cerga-Pashoja A, Knaevelsrud C, Maercker A, et al. Cultural adaptation of Hap-pas-Hapi, an internet and mobile-based intervention for the treatment of psychological distress among Albanian migrants in Switzerland and Germany. Internet Interv 2020 Sep;21:100339 [FREE Full text] [doi: 10.1016/j.invent.2020.100339] [Medline: 32983906]
41. Putcha V, Prakash S, Neuman MJ, Flam R, Roland M, Plaut R. Harnessing technology: a review of digital tools supporting parents and the early childhood workforce. Bernard van Leer Foundation. 2023. URL: https://researchconnections.org/childcare/resources/161281 [accessed 2023-07-27]
42. Loblay V, Ekambareshwar M, Naderbagi A, Song YJ, Ford M, Zahed I, et al. Enhancing equitable engagement for digital health promotion: lessons from evaluating a childrearing app in Indonesia. Digit Health 2023 Dec 25;9:20552076231222112 [FREE Full text] [doi: 10.1177/20552076231222112] [Medline: 38152442]
43. Ospina-Pinillos L, Davenport TA, Navarro-Mancilla AA, Cheng VW, Cardozo Alarcón AC, Rangel AM, et al. Involving end users in adapting a Spanish version of a web-based mental health clinic for young people in Colombia: exploratory study using participatory design methodologies. JMIR Ment Health 2020 Feb 08;7(2):e15914 [FREE Full text] [doi: 10.2196/15914] [Medline: 32027313]

Abbreviations

CBPR: community-based participatory research
mHealth: mobile health
UI: user interface

Edited by T de Azevedo Cardoso; submitted 04.12.23; peer-reviewed by CK Tan, S Hinder; comments to author 30.04.24; revised version received 12.05.24; accepted 21.05.24; published 09.07.24.

Please cite as:
Naderbagi A, Loblay V, Zahed IUM, Ekambareshwar M, Poulsen A, Song YJC, Ospina-Pinillos L, Krausz M, Mamdouh Kamel M, Hickie IB, LaMonica HM
Cultural and Contextual Adaptation of Digital Health Interventions: Narrative Review
J Med Internet Res 2024;26:e55130
URL: https://www.jmir.org/2024/1/e55130
doi:10.2196/55130
PMID:38980719

XSL•FO
RenderX

©Aila Naderbagi, Victoria Loblay, Iqthyer Uddin Md Zahed, Mahalakshmi Ekambareshwar, Adam Poulsen, Yun J C Song, Laura Ospina-Pinillos, Michael Krausz, Mostafa Mamdouh Kamel, Ian B Hickie, Haley M LaMonica. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Telemedicine in Improving Glycemic Control Among Children and Adolescents With Type 1 Diabetes Mellitus: Systematic Review and Meta-Analysis

Kun Zhang[1], MSN; Qiyuan Huang[1], MD; Qiaosong Wang[1,2], MSN; Chengyang Li[1], MSN; Qirong Zheng[1], MSN; Zhuoyue Li[3], MM; Dan Xu[4], MD; Cuiling Xie[1], MSN; Mingqi Zhang[1], MSN; Rongjin Lin[1,2], BM

[1]School of Nursing, Fujian Medical University, Fuzhou, China

[2]Department of Nursing, The First Affiliated Hospital of Fujian Medical University, Fujian Medical University, Fuzhou, China

[3]Department of Infectious diseases, Nanfang Hospital, Southern Medical University, Guangzhou, China

[4]Foreign Language Department, Xuzhou Medical University, Xuzhou, China

Corresponding Author:
Rongjin Lin, BM
School of Nursing
Fujian Medical University
No.1 Xuefu North Road, Minhou County
Fuzhou, 350005
China
Phone: 86 1 380 950 8580
Fax: 86 0 591 228 62526
Email: nplrj@126.com

Abstract

Background: Type 1 diabetes mellitus (T1DM) is the most common chronic autoimmune disease among children and adolescents. Telemedicine has been widely used in the field of chronic disease management and can benefit patients with T1DM. However, existing studies lack high-level evidence related to the effectiveness of telemedicine for glycemic control in children and adolescents with T1DM.

Objective: This study aims to systematically review the evidence on the effectiveness of telemedicine interventions compared with usual care on glycemic control among children and adolescents with T1DM.

Methods: In this systematic review and meta-analysis, we searched PubMed, Cochrane Library, Embase, Web of Science (all databases), and CINAHL Complete from database inception to May 2023. We included randomized controlled trials (RCTs) that evaluated the effectiveness of a telemedicine intervention on glycemic control in children and adolescents with T1DM. In total, 2 independent reviewers performed the study selection and data extraction. Study quality was assessed using the Cochrane Risk of Bias 2 tool. Our primary outcome was glycated hemoglobin (HbA_{1c}) levels. Secondary outcomes were quality of life, self-monitoring of blood glucose, the incidence of hypoglycemia, and cost-effectiveness. A random-effects model was used for this meta-analysis.

Results: Overall, 20 RCTs (1704 participants from 12 countries) were included in the meta-analysis. Only 5% (1/20) of the studies were at high risk of bias. Compared to usual care, telemedicine was found to reduce HbA_{1c} levels by 0.22 (95% CI –0.33 to –0.10; P<.001; I^2=35%). There was an improvement in self-monitoring of blood glucose (mean difference [MD] 0.54, 95% CI –0.72 to 1.80; P=.40; I^2=67.8%) and the incidence of hypoglycemia (MD –0.15, 95% CI –0.57 to 0.27; P=.49; I^2=70.7%), although this was not statistically significant. Moreover, telemedicine had no convincing effect on the Diabetes Quality of Life for Youth score (impact of diabetes: P=.59; worries about diabetes: P=.71; satisfaction with diabetes: P=.68), but there was a statistically significant improvement in non–youth-specific quality of life (MD –0.24, 95% CI –0.45 to –0.02; P=.04; I^2=0%). Subgroup analyses revealed that the effect of telemedicine on HbA_{1c} levels appeared to be greater in studies involving children (MD –0.41, 95% CI –0.62 to –0.20; P<.001), studies that lasted <6 months (MD –0.32, 95% CI –0.48 to –0.17; P<.001), studies where providers used smartphone apps to communicate with patients (MD –0.37, 95% CI –0.53 to –0.21; P<.001), and studies with medication dose adjustment (MD –0.25, 95% CI –0.37 to –0.12; P<.001).

XSL•FO
RenderX

Conclusions: Telemedicine can reduce HbA_{1c} levels and improve quality of life in children and adolescents with T1DM. Telemedicine should be regarded as a useful supplement to usual care to control HbA_{1c} levels and a potentially cost-effective mode. Meanwhile, researchers should develop higher-quality RCTs using large samples that focus on hard clinical outcomes, cost-effectiveness, and quality of life.

(J Med Internet Res 2024;26:e51538) doi:10.2196/51538

KEYWORDS

telemedicine; digital health; web-based; type 1 diabetes mellitus; children; adolescents; glycemic control; chronic disease management; randomized controlled trials; meta-analysis; mobile phone

Introduction

Background

Type 1 diabetes mellitus (T1DM) is the most common chronic autoimmune disease among children and adolescents, characterized by hyperglycemia and caused by an absolute deficiency of insulin [1,2]. More than 1.2 million children and adolescents worldwide currently have T1DM [3]. Adolescence is a period when glycemic control commonly deteriorates [4], and people with diabetes remain at high risk of serious complications, including diabetic cardiovascular disease and diabetic nephropathy [5-7]. T1DM has a serious impact on the life health of children and adolescents. It places a heavy medical burden on the families of those affected [8,9]. Therefore, there is an imperative to explore effective treatment together with management strategies to help children and adolescents maintain normoglycemia and promote their long-term health as well as their well-being.

In recent years, telemedicine has been widely used in the field of chronic disease management. Telemedicine (a subcomponent of eHealth) has been defined as "The delivery of health care services, where distance is a critical factor, by all health care professionals using information and communications technologies for the exchange of valid information for diagnosis, treatment and prevention of disease and injuries, research and evaluation, and the continuing education of health care workers, with the aim of advancing the health of individuals and communities" [10]. For patients with chronic diseases, the advantages of telemedicine can be reflected in improving access to services, ensuring continuity of care, and mitigating the costs of care delivery [11,12]. Although telemedicine may not be able to provide physicians with comprehensive diagnostic information about a patient in the same way that a physical examination can, it can assist physicians in monitoring and recording certain specific physiological indicators (eg, blood glucose, blood oxygen concentration, blood pressure, and heart rate) to help them observe the trajectory of a patient's health [13,14]. The current studies on telemedicine interventions for glycemic control in patients with diabetes focus on (1) telemonitoring (eg, a web-based telemedicine system was used to monitor patients with T1DM in the study by Ruiz de Adana et al [15]), (2) tele-education (eg, Molavynejad et al [16] delivered tele-education to patients with diabetes using remote video-based technology), and (3) teleconsultation and internet-based group appointments (eg, Bisno et al [17] provided both individual telehealth provider visits and internet-based group appointments for patients with T1DM through the CoYoT1 clinic). Moreover, previous meta-analyses have shown that the effectiveness of telemedicine in controlling blood glucose levels in patients with T1DM has been well validated [18-20]. It can be seen that telemedicine provides a huge advantage for diabetes glycemic control.

However, existing studies lack high-level evidence related to the effectiveness of telemedicine for glycemic control in children and adolescents with T1DM. Only a few studies have reported on the potential of telemedicine in the management of T1DM in children and adolescents. Moreover, the safety and applicability of telemedicine for children and adolescents with T1DM need to be further demonstrated. Therefore, we aimed to conduct a systematic review and meta-analysis of current randomized controlled trials (RCTs) to provide new evidence for clinical decision-making by comparing the effectiveness of telemedicine interventions with usual care in children and adolescents with T1DM.

Study Question

How does telemedicine compare with usual care in improving glycemic control among children and adolescents with T1DM? Which form of telemedicine intervention is more effective in improving glycemic control among children and adolescents with T1DM?

Study Objective

This meta-analysis aimed to comprehensively synthesize and evaluate evidence on the effectiveness of telemedicine on glycemic control among children and adolescents with T1DM.

Methods

Search Strategy

In total, 5 electronic databases covering the realms of biomedicine science, clinical medicine science, and general references were screened: PubMed, Cochrane Library, Embase, Web of Science (all databases), and CINAHL Complete. The dates searched were from establishment of each database to May 1, 2023. The search was conducted using the following keywords: ("Diabetes Mellitus, Type 1") AND ("Telemedicine" OR "Telemetry" OR "Telenursing" OR "Internet-Based Intervention") AND ("Child" OR "Adolescent"). Medical Subject Heading terms and their related terms were used. Multimedia Appendix 1 [21-40] shows the detailed search terms and search process. There were no restrictions in terms of participant age, year of publication, or region of study at this stage. The review protocol was reported according to the

XSL•FO
RenderX

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist (Multimedia Appendix 2).

Inclusion and Exclusion Criteria

The inclusion criteria were defined by population, intervention, comparison, outcome, and study design as follows:

1. Population: the target participants were children (aged ≤10 years) and adolescents (10 years<age≤19 years) [41] with T1DM.
2. Intervention: complete or partial telemedicine intervention. A complete telemedicine intervention was one in which there was no face-to-face contact between the participants and the health care providers throughout the trial period from baseline to the end of the intervention and the only telemedicine interventions were via telephone, web-based videoconferencing, a website, or a smartphone app (all treatments [including initial treatment] were delivered via telemedicine). *Partial telemedicine intervention* referred to treatments that combine telemedicine with nontelemedicine interventions (such as a follow-up visit in an outpatient clinic or a visit at home). These 2 broad categories of telemedicine interventions were further subdivided by the number of intervention forms. "Single" refers to the inclusion of only 1 form of telemedicine intervention, whereas "mixed" refers to the inclusion of ≥2 forms of telemedicine intervention. Complete telemedicine interventions were categorized as single and mixed complete telemedicine interventions; partial telemedicine interventions were categorized as single and mixed partial telemedicine interventions.
3. Comparison: containing a comparison group with usual care, including a nontelemedicine intervention and health guidance only before discharge treated as a blank control.
4. Outcome: we included all studies that reported serum glycated hemoglobin (HbA_{1c}) levels as either their primary or secondary outcomes.
5. Study design: only RCTs (parallel or crossover) were included.

The exclusion criteria were (1) studies using nonexperimental and quasi-experimental designs; (2) abstracts, brief reports, conference proceedings, conference papers, posters, and letters to editors; (3) studies on patients with gestational diabetes; and (4) studies published in languages other than English because of our lack of high-quality translational resources.

Study Screening

Throughout the screening processes, all studies included in the analysis were independently reviewed by 2 researchers (KZ and CL). First, we screened the titles and abstracts of all bibliographic records against the inclusion and exclusion criteria, and a label was created on a serial numbered sheet to add the reason for exclusion as a note. Second, we thoroughly read the full text of the study without exclusion labels to ensure that all inclusion and exclusion criteria were met. Disagreements between the researchers were resolved by meeting with a third reviewer (QH). Studies judged to be eligible at this stage were then included in the quality assessment where applicable.

Quality Assessment

We assessed the risk of bias using the Cochrane Risk of Bias 2 tool [42] to evaluate the randomization process, deviations from the intended interventions, missing outcome data, measurement of the outcome, and selection of the reported results. In total, 2 researchers (KZ and CL) assessed the trials independently and resolved any disagreements by meeting with a third reviewer (QH). The quality of evidence of each study was assessed by 2 reviewers (QH and QW) using the Grading of Recommendations, Assessment, Development, and Evaluations approach [43].

Outcome

The primary outcome was HbA_{1c} levels. Secondary outcomes were quality of life as measured using a validated instrument, daily frequency of self-monitoring of blood glucose (SMBG), the incidence of hypoglycemia, and cost-effectiveness.

Data Extraction

For each included study, 2 reviewers (KZ and CL) independently extracted the data for analysis. When data were missing or unclear, we contacted the authors. If the authors did not respond, the study was reassessed and excluded.

We extracted the following information from the selected studies: (1) study characteristics (study name, author, year of publication, country, study design, attrition rate, and sample size), (2) characteristics of the participants (age, gender, diabetes duration, baseline HbA_{1c} levels, total cholesterol levels, triglyceride levels, blood pressure, and BMI), (3) intervention details (duration, types of health care providers, frequency of feedback, characteristics of intervention content, communication forms between providers and patients, technology use modes, and telemedicine intervention forms; communication forms included modem, SMS text messaging, email, web conference, website—websites where patients upload blood glucose levels or other clinical data and share them with their health care providers—computer software, smart wearable devices—smart wearable devices are consumer-grade connected electronic devices that can be worn on the body as an accessory or embedded into clothing [44] —telephone, and smartphone or its apps), and (4) general information about outcomes (the mean and SD at baseline and at the end of the intervention, number of participants analyzed at the end of the intervention, and tools used for measurement; when several analyses were performed on the same outcome at the same time point, we extracted the data from the intention-to-treat analysis).

Data Analysis

Stata (version 17; StataCorp) and Review Manager (version 5.4; The Cochrane Collaboration) were used for all statistical analysis. For quantitative synthesis, we collected the difference between baseline and end-point values for both the intervention and control groups. In the absence of information, data were estimated from the mean and SD of baseline and end-point values using a correlation of 0.5 [45]. To ensure accuracy, different correlations, such as 0.4 and 0.6, were used for estimation data and sensitivity analysis. The final results showed that the estimated results obtained using the different

XSL•FO
RenderX

correlations remained stable after sensitivity analysis [45]. Data conversion tools were used to convert the median, maximum, and minimum values reported in the included studies into mean [46] and SD [47]. We reported the results of secondary outcomes when data from at least 2 studies could be merged. The magnitude of the overall effect size was calculated based on the pooled mean difference (MD) with 95% CI when the same measures were used in the studies. If outcomes were measured using different outcome measurement scales, the pooled standardized MD (SMD) with 95% CI was adopted. A P value of <.05 was considered statistically significant.

A random-effects or fixed-effects meta-analysis for continuous data was performed based on the results of the heterogeneity test. Study heterogeneity was determined using the Cochran Q test and Higgins I^2 test. I^2 values of 25%, 50%, and 75% indicated low, moderate, and high heterogeneity, respectively [45]. If P>.10 and I^2<50% were identified, fixed-effects models were used; otherwise, random-effects models were applied. To ensure the robustness of our results, a sensitivity analysis was performed by using leave-one-out analysis to assess the contribution of each study to the merged effect size.

Publication bias was assessed creating funnel plots, the Begg test, and performing the Egger regression test (considered significant at P<.05) by 2 reviewers (QW and CL), and agreement was reached through consensus [48,49]. For the primary outcome, we performed a series of subgroup analyses to quantify specific differences in the size of effects of particular telemedicine interventions based on study and intervention characteristics [50]. Moreover, we performed a univariable meta-regression analysis to investigate whether there was heterogeneity due to differences in study or intervention characteristics.

Protocol Deviation

First, the definition of the intervention group in the registration program as an Internet-based telemedicine intervention group is too broad and simplistic. After further research, we decided to categorize the interventions into complete and partial telemedicine interventions. These 2 broad categories of telemedicine interventions were further subdivided by the number of intervention forms. “Single” refers to the inclusion of only 1 form of telemedicine intervention, whereas “mixed” refers to the inclusion of ≥2 forms of telemedicine intervention. Complete telemedicine interventions were categorized as single and mixed complete telemedicine interventions; partial telemedicine interventions were categorized as single and mixed partial telemedicine interventions. These changes and clarifications help explain the impact of the “face-to-face contact between patient and healthcare provider” factor on telemedicine effectiveness during telemedicine interventions, which has important implications for the development of future telemedicine interventions.

Second, the definition of the control group (“usual care”) was also an oversimplification, so we illustrated 2 cases of “usual care” in this study, including a nontelemedicine intervention as well as health guidance only before discharge treated as a blank control.

Third, regarding secondary outcomes, initially, we identified secondary outcomes based on studies related to diabetes telemedicine in adults and other types of diabetes. However, during the literature reading, it was found that no studies analyzed blood pressure, weight, and patient satisfaction as study outcomes in telemedicine interventions on children and adolescents with T1DM. Some studies used only weight and blood pressure as baseline indicators and lacked postintervention data. Other studies only asked participants how satisfied they were with the telemedicine intervention through interviews at the end of the intervention, which prevented us from quantitatively assessing satisfaction. Therefore, secondary outcomes such as blood pressure, weight, and patient satisfaction were removed.

Finally, regarding the data synthesis strategy, we modified the section for missing data estimation. Because data for the primary and secondary outcomes were partially missing, we first used the commonly used correlation coefficient of 0.5 for data estimation according to the Cochrane Handbook for Systematic Reviews of Interventions [45]. However, as there is currently no clear specification for the use of correlation coefficients for data estimation (only a broad range of choices), to ensure that the effect sizes synthesized using the “estimated data” were sufficiently stable, we also used 0.4 and 0.6 as correlation coefficients for data estimation. (Our main purpose was to see whether the estimated effects using the new correlation coefficients would pass the sensitivity analyses after changing the correlation coefficient). The sensitivity analyses showed that the results synthesized after estimating the missing data using all 3 correlation coefficients were stable and reliable, but the data estimated using the more common correlation coefficient of 0.5 was used as the results of this study.

Results

Search Results

The phases of electronic search, identification, and screening for eligible studies are depicted in the PRISMA flowchart (Figure 1). A total of 546 studies were identified using the search strategy described previously. After removing duplicates and screening titles and abstracts, a total of 20 studies were retained for full-text evaluation. Finally, a total of 20 studies with 1704 participants were included in this systematic review and meta-analysis.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart depicting the main stages of the systematic review process. RCT: randomized controlled trial.

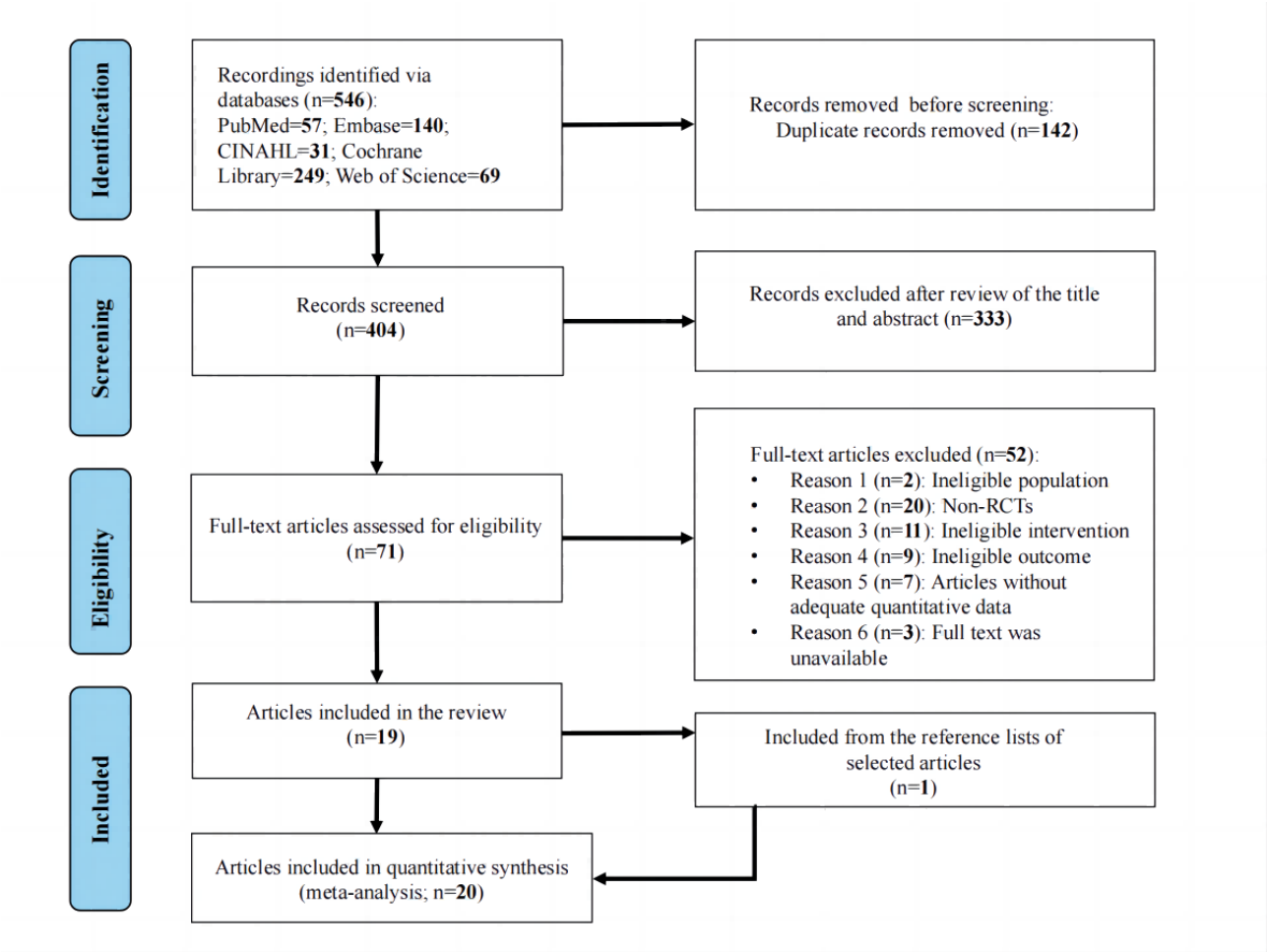

Study Characteristics

The characteristics of the studies are summarized in Table S1 in Multimedia Appendix 1. A total of 90% (18/20) of the studies were parallel-group RCTs [21-38], and 10% (2/20) were crossover studies [39,40]. Of the 20 included studies, 12 (60%) were published after 2015. In total, 40% (8/20) of the studies were published in North America [21-24,27,29,30,34], 35% (7/20) were published in Europe [25,26,33,35-37,39], 20% (4/20) were published in Asia [28,32,38,40], and 5% (1/20) were published in Oceania [31]. The sample sizes of the studies ranged from 20 to 240, with the intervention periods ranging from 3 to 60 months. All participants included in the studies were aged <20 years and had T1DM. The median mean age at baseline was 13.5 years, and the median mean diabetes duration at baseline was 6.2 years. A total of 90% (18/20) of the studies were performed in adolescents (mean age 13.6; range 10.8-17.3 years), and 10% (2/20) of the studies were performed in children (mean age 5.8; range 5.6-6.1 years). The proportion of female participants at baseline ranged from 42% to 62%. The floored threshold value of baseline HbA_{1c} levels in 35% (7/20) of the studies was ≥7.5%.

Intervention Characteristics

The telemedicine systems used in most studies were relatively simple to operate, having clear processes and including transmission of blood glucose data with feedback (15/20, 75%) [21-24,26,28,29,31,33-37,39,40] or blood glucose data only (5/20, 25%) [25,27,30,33,38]. A specialist diabetes care team, including a diabetologist, nurse, dietician, and psychologist, was reported in 45% (9/20) of the studies [22,24,29,31,33-35,37,40]. Feedback was provided monthly or less frequently in 50% (10/20) of the studies [23,26,31,33-37,39,40] and every 2 weeks or more frequently in 25% (5/20) of the studies [21,22,24,28,29], and the frequency of feedback was not specified in 25% (5/20) of the studies (Table 1) [25,27,30,32,38].

XSL•FO
RenderX

Table 1. Characteristics of the telemedicine interventions.

Study, year, and country or region	Health care provider	Communication form		Frequency of feedback	Intervention content					Technology use mode	Technology form	Telemedicine intervention form
		Provider to patient	Patient to provider		Internet-based follow-up	Medication adjustment	Diet guidance	Physical exercise	Basic health education			
Chase et al [21], 2003, United States	Nurse and physician	Telephone	Modem	Every 2 weeks	Yes	No	No	No	Yes	__[a]	Hardware	Complete telemedicine intervention (single)
Gandrud et al [22], 2018, United States	Diabetes educator, nurse, and physician	SMS text messaging and email	Smartphone app	Weekly	Yes	Yes	No	Yes	Yes	—	Software	Partial telemedicine intervention (mixed)
Goyal et al [23], 2017, Canada	Human factors specialist, nurse, and physician	Smartphone app and telephone	Smart wearable device	Every 3 months	Yes	No	No	No	No	Independently	Software	Partial telemedicine intervention (mixed)
Han et al [24], 2015, United States	Diabetes educator, nurse, and physician	SMS text messaging	Smartphone app and SMS text messaging	Every 2 days	Yes	Yes	No	No	Yes	Independently	Software	Complete telemedicine intervention (single)
Ibrahim et al [25], 2021, Europe	Diabetologist	SMS text messaging	—	—	No	No	No	No	No	Independently	Software	Partial telemedicine intervention (single)
Klee et al [39], 2018, Switzerland	Nurse and diabetologist	Smartphone app and website	Telephone and email	Monthly	Yes	Yes	Yes	No	No	—	Software	Partial telemedicine intervention (mixed)
Kowalska et al [26], 2017, Poland	Pediatrician and diabetologist	Computer software	Smart wearable device	Every 13 weeks	Yes	Yes	Yes	No	No	Parental assistance	Software	Partial telemedicine intervention (single)
Kumar et al [27], 2004, United States	Trained research assistant	Website	Modem and smart wearable device	—	No	Yes	Yes	No	No	Parental assistance	Software	Complete telemedicine intervention (single)
Landau et al [28], 2012, Israel	Dietitian and pediatric endocrinologist	Telephone	Website and smart wearable device	Every week	Yes	Yes	No	No	No	Independently	Software	Partial telemedicine intervention (mixed)
Marrero et al [29], 1995, United States	Pediatric diabetologist, nurse, social workers, and dietitians	Computer software and telephone	Modem	Every 2 weeks	Yes	Yes	Yes	No	No	—	Software	Partial telemedicine intervention (mixed)

XSL•FO
RenderX

Study, year, and country or region	Health care provider	Communication form		Frequency of feedback	Intervention content					Technology use mode	Technology form	Telemedicine intervention form
		Provider to patient	Patient to provider		Internet-based follow-up	Medication adjustment	Diet guidance	Physical exercise	Basic health education			
Mulvaney et al [30], 2010, United States	Diabetes professionals	Website	Website	—	No	No	No	No	Yes	—	Software	Complete telemedicine intervention (single)
Nunn et al [31], 2006, Australia	Pediatric endocrinologists, nurse, dietitian, and social worker	Telephone	Telephone	Every 2 months	Yes	Yes	Yes	Yes	Yes	Parental assistance	Hardware	Complete telemedicine intervention (single)
Raviteja et al [32], 2019, India	Consultant and physician	Smart wearable device	Smart wearable device and computer software	—	Yes	Yes	Yes	Yes	No	Independently	Hardware	Complete telemedicine intervention (mixed)
Schiaffini et al [33], 2016, Italy	Diabetologist, nurse, dietician, and psychologist	Web conference	Website and smart wearable device	Every month	Yes	No	Yes	Yes	Yes	Parental assistance	Software	Complete telemedicine intervention (mixed)
Shalitin et al [40], 2014, Israel	Diabetes care team	Website, email, and telephone	Smart wearable device and website	Every month	Yes	Yes	No	No	No	Parental assistance	Software	Complete telemedicine intervention (mixed)
Stanger et al [34], 2018, United States	Pediatric endocrinologist and diabetes care team	Web conference	Smart wearable device	Every month (last period)	Yes	Yes	No	No	No	Parental assistance	Software	Complete telemedicine intervention (single)
Von Sengbusch et al [35], 2020, Germany	Regular home diabetes team	Web conference	Smart wearable device and computer software	Every month	Yes	Yes	No	Yes	Yes	Parental assistance	Software	Partial telemedicine intervention (single)
Ware et al [36], 2022, United Kingdom	Nurse and physician	Smartphone app, telephone, and email	Smartphone app and smart wearable device	Every month	Yes	Yes	No	No	No	Parental assistance	Software	Complete telemedicine intervention (mixed)
Ware et al [37], 2022, United Kingdom	Research team and clinical team	Smartphone app, telephone, and email	Smartphone app and smart wearable device	Every month	Yes	Yes	Yes	No	Yes	Parental assistance	Software	Complete telemedicine intervention (mixed)

XSL•FO
RenderX

Study, year, and country or region	Health care provider	Communication form		Frequency of feedback	Intervention content					Technology use mode	Technology form	Telemedicine intervention form
		Provider to patient	Patient to provider		Internet-based follow-up	Medication adjustment	Diet guidance	Physical exercise	Basic health education			
Xu et al [38], 2021, China	Nurse and third-party health manager	Smartphone app	Smartphone app and smart wearable device	—	Yes	No	Yes	Yes	Yes	Independently	Software	Complete telemedicine intervention (mixed)

[a]Not reported.

The communication technologies used in the telemedicine interventions included in the studies took a variety of forms. Patients initiated communication with health care providers through different forms of telemedicine: smart wearable devices (6/20, 30%) [23,26,32,34,35,40], smartphone apps (5/20, 25%) [22,24,36-38], modem (3/20, 15%) [21,27,29], websites (3/20, 15%) [28,30,33], telephone (2/20, 10%) [31,39], and unclear (1/20, 5%) [25]. Health care providers initiated communication with patients through different forms of telemedicine: smartphone apps (5/20, 25%) [23,36-39], websites (3/20, 15%) [27,30,40], web conferences (3/20, 15%) [33-35], telephone (3/20, 15%) [21,28,31], SMS text messaging (3/20, 15%) [22,24,25], computer software (2/20, 10%) [26,29], or smart wearable devices (1/20, 5%) [32]. A total of 85% (17/20) of the studies mainly used various types of software [22-30,33-40], and 15% (3/20) of the studies used hardware [21,31,32].

In total, 45% (9/20) of the studies involved patients using telemedicine with parental assistance [26,27,31,33-37,40], and 30% (6/20) of the studies involved patients using telemedicine independently [23-25,28,32,38]. The form of intervention was complete telemedicine intervention in 60% (12/20) of the studies [21,24,27,30-34,36-38,40] and partial telemedicine intervention in 40% (8/20) of the studies [22,23,25,26,28,29,35,39]. The content of the telemedicine interventions in the studies included a variety of features: internet-based communication and follow-up (17/20, 85%) [21-24,26,28,29,31-40], medication dose adjustment (14/20, 70%) [22,24,26-29,31,32,34-37,39,40], basic health education (9/20, 45%) [21,22,24,30,31,33,35,37,38], diet guidance (9/20, 45%) [26,27,29,31-33,37-39], and physical exercise (6/20, 30%) [22,31-33,35,38]. A total of 55% (11/20) of the studies reported characteristics of the content of the intervention including at least 3 features [22,24,26,29,31-33,35,37-39]. No features of the content of the telemedicine interventions were reported in 5% (1/20) of the studies [25].

Risk of Bias

On the basis of the Cochrane Risk of Bias 2 tool, all studies except for 5% (1/20) with a high risk of bias [25] and 20% (4/20) with a low risk of bias [23,26,32,39] were found to have “some concerns” (Figures 2 and 3 [21-40]). The greatest bias was found in the randomization process. Randomization was reported to be implemented in all studies, among which only 25% (5/20) of the studies explicitly described the randomization strategies and properly applied allocation concealment [23,26,30,32,39]. The other study [25] was rated as high risk because of baseline differences between intervention groups. No preregistration was reported in 45% (9/20) of the studies [21,24,27-29,31,33,34,40], and the risk of bias regarding the choice of reporting outcomes was rated as “some concerns.” One of the domains with the highest proportion of low risk of bias was “bias from missing outcome data.”

XSL•FO
RenderX

Figure 2. Risk-of-bias graph of the included studies (part 1).

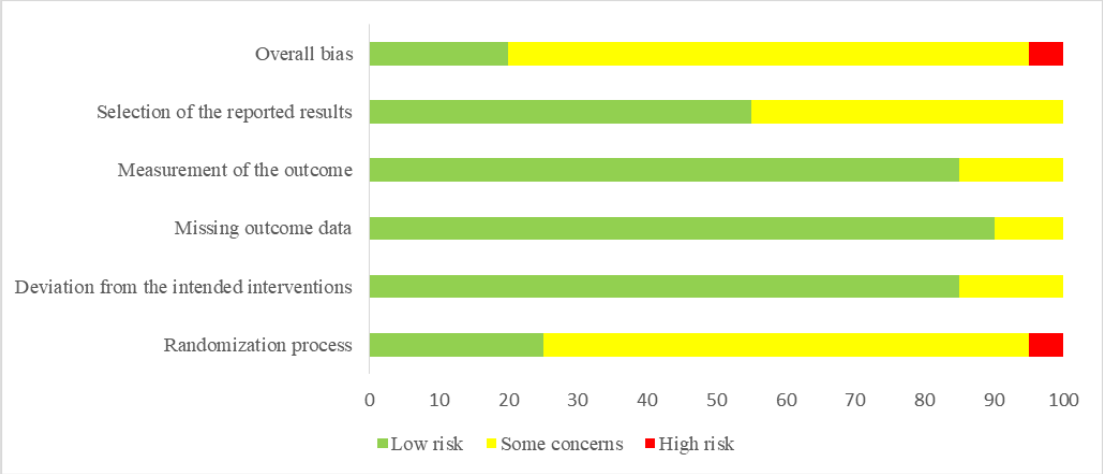

Figure 3. Risk-of-bias graph of the included studies (part 2).

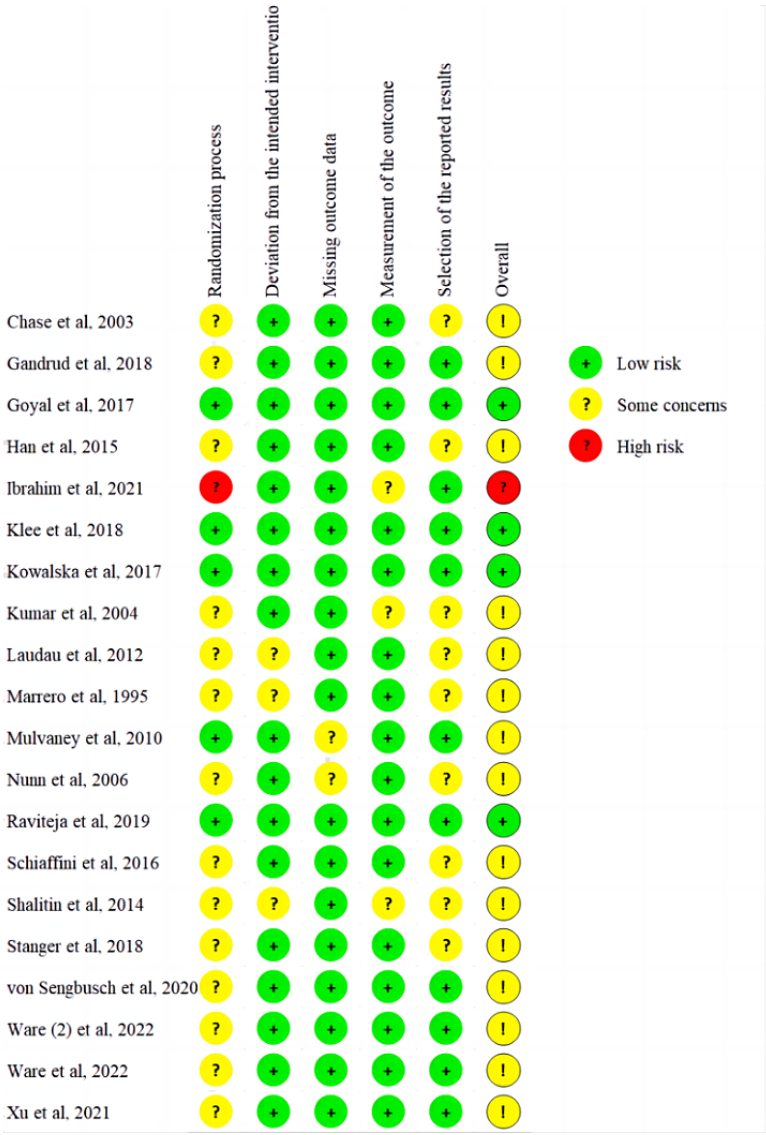

Meta-Analysis and Descriptive Analysis Results

A summary of the main results for the comparisons using the Grading of Recommendations, Assessment, Development, and Evaluations ratings is presented in Table 2. Detailed meta-analytic forest plots on all outcomes and subgroups are shown in Figure 4 and Figure S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Summary of findings—telemedicine compared to usual care for glycemic control in children and adolescents with type 1 diabetes mellitus.

Certainty assessment							Patients, n		Effect, absolute (95% CI; *P* value)	Certainty
Studies, n (%)	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Telemedicine	Usual care		
HbA_{1c}[a]										
20 (100)	Randomized trials	Serious[b]	Not serious[c]	Not serious	Serious[d]	None	822	822	MD[e] –0.22 (–0.33 to –0.10; <.001)	Low
DQOLY[f] (impact of diabetes)										
2 (10)	Randomized trials	Serious[g]	Not serious	Not serious	Very serious[h]	None	30	23	MD 1.27 (–3.31 to 5.86; .59)	Very low
DQOLY (worries about diabetes)										
2 (10)	Randomized trials	Serious[g]	Not serious	Not serious	Very serious[h]	None	30	23	MD 0.58 (–2.59 to 3.66; .71)	Very low
DQOLY (satisfaction with diabetes)										
2 (10)	Randomized trials	Serious[g]	Serious[i]	Not serious	Very serious[h]	None	30	23	MD 3.27 (–12.53 to 19.08; .68)	Very low
N-QOL[j]										
3 (15)	Randomized trials	Serious[k]	Not serious	Not serious	Serious[l]	None	165	160	SMD[m] –0.24 (–0.45 to –0.02; .04)	Low
SMBG[n]										
3 (15)	Randomized trials	Serious[o]	Serious[i]	Not serious	Very serious[h]	None	96	91	MD 0.54 (–0.72 to 1.8; .40)	Very low
Incidence of hypoglycemia										
4 (20)	Randomized trials	Serious[p]	Serious[i]	Not serious	Serious[l]	None	153	156	MD –0.22 (–0.66 to 0.23; .49)	Very low

[a]HbA_{1c}: glycated hemoglobin.

[b]Downgraded for unclear or inadequate randomization process (15/20, 75% of the included studies). In a large number of studies, allocation was not adequately concealed due to the nature of the intervention.

[c]Although the Cochran *Q* test and Higgins I^2 test suggested a low heterogeneity, we chose not to downgrade for inconsistency as this was fully explained by the inclusion of 1 study.

[d]A total of 65% (13/20) of the studies had sample sizes of <50 in both arms.

[e]MD: mean difference.

[f]DQOLY: Diabetes Quality of Life for Youth.

[g]One of the studies had some concerns (a moderate risk of bias).

[h]Sample sizes for each arm of the included studies were <50.

[i]Significant heterogeneity.

[j]N-QOL: non–youth-specific quality of life.

[k]All 3 studies had some concerns.

XSL•FO
RenderX

[l]There was at least 1 study with a sample size of <50 in both arms.

[m]SMD: standardized mean difference.

[n]SMBG: self-monitoring of blood glucose.

[o]A total of 10% (2/20) of the studies had some concerns.

[p]One study had some concerns.

Figure 4. Forest plot of the comparison of telemedicine interventions versus usual care. Outcome: glycated hemoglobin. DL: DerSimonian and Laird approach.

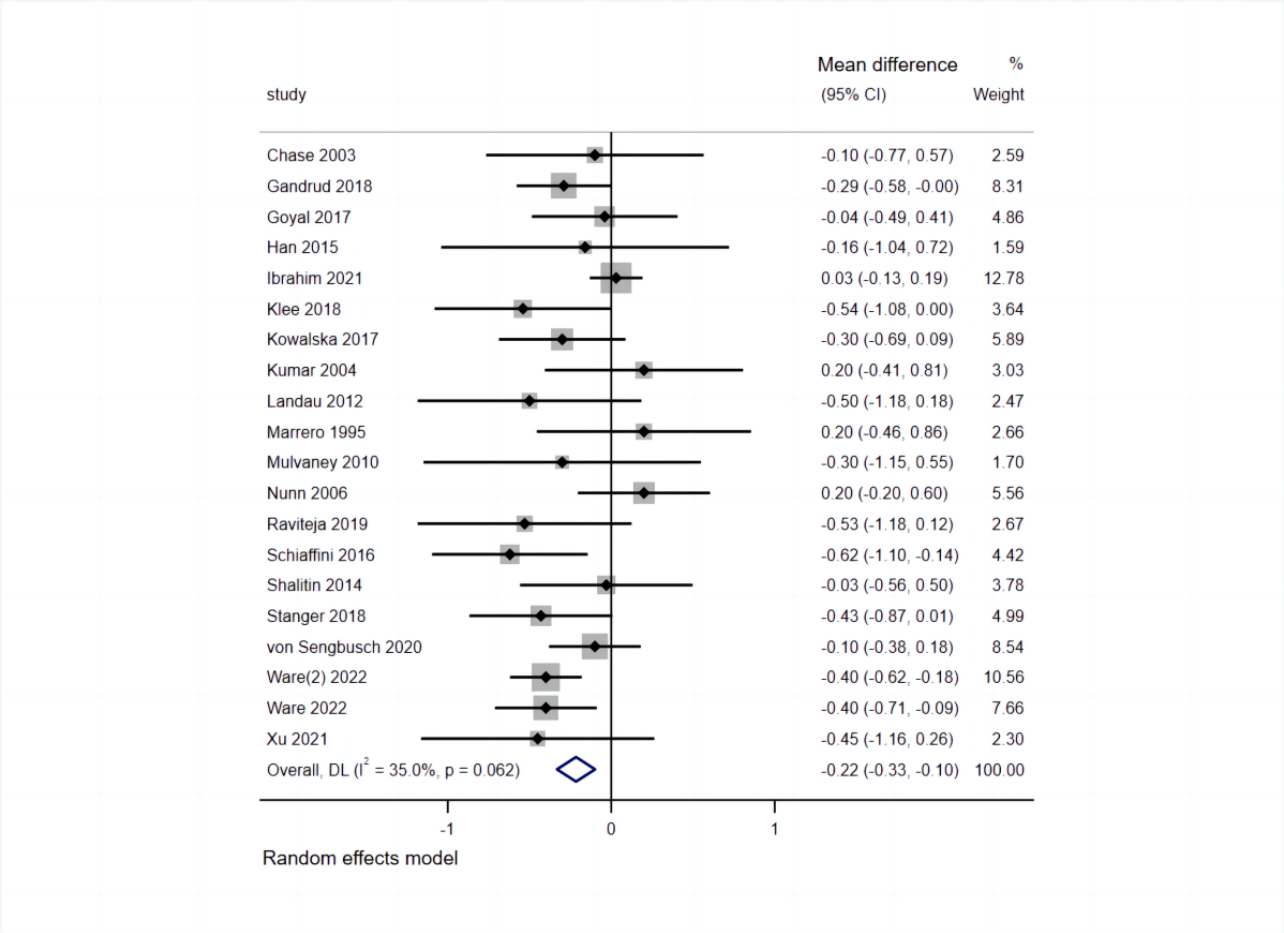

Effect of Telemedicine Interventions on HbA1c

The 20 studies, which reported HbA_{1c} levels at 3 to 50 months and examined 1704 participants, were included in the meta-analysis. Overall, telemedicine was found to reduce HbA_{1c} levels by 0.22 (95% CI –0.33 to –0.10; P<.001) at the end of the intervention. Furthermore, the heterogeneity of the effect size was confirmed as I^2 was 35% (Q_{19}=29.23; P=.06), suggesting heterogeneity of a low degree. Given the wide variety of technologies available for telemedicine, the heterogeneity of results is not surprising. No significant improvements were noted at the end of the 3- (MD –0.30, 95% CI –0.62 to 0.02; P=.07; n=4) or 12-month (MD –0.04, 95% CI –0.33 to 0.40; P=.85; n=2) follow-up; however, significant improvement was found at the end of the 6-month follow-up (MD –0.21, 95% CI –0.37 to –0.05; P=.01; n=8).

Effect of Telemedicine Interventions on Secondary Outcomes

We pooled the Diabetes Quality of Life for Youth (DQOLY) scores [51,52] from 10% (2/20) of the studies (n=53) [24,39], non–youth-specific quality of life (N-QOL; using diabetes quality of life [53] and health-related quality of life [54]) from 15% (3/20) of the studies (n=334) [35,38,40], daily frequency of SMBG from 15% (3/20) of the studies (n=187) [23,34,40], and incidence of hypoglycemia from 20% (4/20) of the studies (n=309) [23,26,32,38].

There was no significant effect size in secondary outcomes except for the N-QOL, with MD for DQOLY (DQOLY impact of diabetes subscale: MD 1.27, 95% CI –3.31 to 5.86, n=53, and I^2=32.2%; DQOLY worries about diabetes subscale: MD 0.58, 95% CI –2.49 to 3.66, n=53, and I^2=23.8%; DQOLY satisfaction with diabetes subscale: MD 3.27, 95% CI –12.53 to 19.08, n=53, and I^2=75.6%), an SMD of −0.24 for the N-QOL (95% CI –0.45 to –0.02; n=334; I^2=0%), an MD of 0.54 for daily frequency of SMBG (95% CI –0.72 to 1.80; n=187;

XSL•FO
RenderX

I^2=67.8%), and an SMD of −0.22 for incidence of hypoglycemia (95% CI –0.66 to 0.23; n=309; I^2=73.7%).

Only 5% (1/20) of the studies [21] reported economic data. The difference in cost-effectiveness of care between the 2 groups was significant. The average cost per patient in the intervention group for the 6 months was US $163. The control group spent an average of US $246 to visit the clinic. If additional costs (average US $59), such as mileage, parking, meals, hotel stays, and babysitting, were included, the average cost of a clinic visit increased to US $305. This result shows that the telemedicine intervention was cost-effective, at least in the United States.

Subgroup Analysis of HbA1c

Our subgroup analysis based on study and intervention characteristics revealed that the subgroup differences that yielded statistical significance were publication date, communication forms (from patient to provider), and internet-based follow-up (Table 3).

Regardless of age, intervention duration, and health care provider, HbA_{1c} levels significantly decreased in all studies after the telemedicine intervention.

XSL•FO
RenderX

Table 3. Summary of subgroup analysis based on glycated hemoglobin.

Characteristic and subgroup	Number of trials (number of participants)	Effect size, MD[a] (95% CI)	I^2 (%)	*P* value (*Q* test)	Heterogeneity between groups
Age					0.058
Children	2 (210)	–0.41 (–0.62 to –0.20)	0	.71	
Adolescents	18 (1434)	–0.18 (–0.30 to –0.06)	28.1	.13	
Publication date					0.010
2010 and before	5 (384)	0.11 (–0.15 to 0.37)	0	.80	
After 2010	15 (1260)	–0.27 (–0.40 to –0.15)	36.8	.08	
Intervention duration					0.199
<6 months	8 (495)	–0.32 (–0.48 to –0.17)	0	.58	
At least 6 months	12 (1149)	–0.18 (–0.33 to –0.03)	43.2	.06	
Health care provider					0.884
Professional diabetes care team	9 (850)	–0.21 (–0.38 to –0.04)	33.5	.15	
No professional diabetes care team	11 (794)	–0.23 (–0.40 to –0.06)	40.7	.08	
Feedback frequency					0.426
More than once a month	5 (364)	–0.23 (–0.449 to –0.002)	0	.63	
Less than or equal to once a month	10 (983)	–0.27 (–0.41 to –0.12)	35.6	.12	
Unclear	5 (297)	–0.08 (–0.32 to 0.16)	20.3	.29	
Communication form					
Provider to patient					0.259
Telephone	3 (244)	–0.06 (–0.46 to 0.35)	34.9	.22	
SMS text messaging	3 (229)	–0.10 (–0.35 to 0.15)	45.4	.16	
Smartphone app	5 (440)	–0.37 (–0.53 to –0.21)	0	.62	
Computer software	2 (211)	–0.12 (–0.59 to 0.35)	39.6	.20	
Website	3 (127)	0.00 (–0.36 to 0.36)	0	.63	
Web conference	3 (330)	–0.34 (–0.65 to –0.02)	49.9	.14	
Smart wearable device	1 (63)	−0.53 (−1.18 to 0.12)	__[b]	—	
Patient to provider					0.002
Modem	3 (209)	0.11 (–0.26 to 0.48)	0	.76	
Smartphone app	5 (453)	–0.37 (–0.51 to –0.22)	0	.96	
Smart wearable device	6 (595)	–0.20 (–0.37 to –0.03)	0	.60	
Telephone	2 (156)	–0.15 (–0.87 to 0.58)	78.3	.03	
Website	3 (139)	–0.53 (–0.89 to –0.18)	0	.81	
Not reported	1 (92)	0.03 (–0.13 to 0.19)	—	—	
Technology form					0.505
Hardware	3 (249)	–0.08 (–0.51 to 0.35)	43.2	.17	
Software	17 (1395)	–0.23 (–0.36 to –0.11)	34.8	.08	
Technology use mode					0.534
Independent use	6 (374)	–0.11 (–0.31 to 0.09)	0	.53	
Parental assistance	9 (899)	–0.24 (–0.41 to –0.07)	44.1	.07	
Unclear	5 (371)	–0.26 (–0.47 to –0.04)	0	.53	
Telemedicine intervention form					0.206
Complete telemedicine intervention	12 (802)	–0.28 (–0.43 to –0.13)	20.5	.24	

XSL•FO
RenderX

Characteristic and subgroup	Number of trials (number of participants)	Effect size, MD[a] (95% CI)	I^2 (%)	*P* value (*Q* test)	Heterogeneity between groups
Partial telemedicine intervention	8 (842)	–0.14 (–0.29 to 0.01)	29.4	.19	
Internet-based follow-up					0.002
With feature	17 (1460)	–0.27 (–0.38 to –0.17)	8.9	.35	
Without feature	3 (184)	0.03 (–0.12 to 0.18)	0	.64	
Medication adjustment					0.577
With feature	14 (1267)	–0.25 (–0.37 to –0.12)	20	.24	
Without feature	6 (377)	–0.17 (–0.41 to 0.07)	37.4	.16	
Physical exercise					0.823
With feature	6 (691)	–0.24 (–0.47 to 0.01)	44.6	.11	
Without feature	14 (953)	–0.21 (–0.35 to –0.07)	35.2	.09	

[a]MD: mean difference.

[b]Data synthesis is not possible with only one study.

Population of the Study

No statistically significant subgroup differences were identified in the subgroup analysis by age. A statistically significant decrease in HbA_{1c} levels was observed in subgroups of children (MD –0.41, 95% CI –0.62 to –0.20; *P*<.001) and adolescents (MD –0.18, 95% CI –0.30 to –0.06; *P*=.003). The children subgroup reported a higher MD than the adolescent subgroup.

Publication Date of the Studies

Subgroup analysis stratified by publication date demonstrated significant effectiveness of studies published after 2010 on glycemic control in children and adolescents with T1DM compared with those published before 2010. Moreover, a decrease in heterogeneity and statistically significant subgroup differences was found in the subgroup analysis based on publication date (*P*=.01), which can explain the heterogeneity in overall effect on HbA_{1c} levels.

Duration of Telemedicine Interventions

We created 2 subgroups: interventions lasting <6 months and interventions lasting at least 6 months. The results revealed that telemedicine interventions lasting <6 months demonstrated a more significant reduction in HbA_{1c} levels (MD –0.32, 95% CI –0.48 to –0.17; *P*<.001).

Health Care Provider of Telemedicine Interventions

Subgroup analysis based on health care provider demonstrated significant effectiveness with or without the professional diabetes care team, and similar MDs were reported between the 2 groups (with care team: MD –0.21, 95% CI –0.38 to –0.04, and *P*=.02; without care team: MD –0.23, 95% CI –0.40 to –0.06, and *P*=.01).

Feedback Frequency of Telemedicine Interventions

Contrary to the nonsignificant overall effect of –0.01 on HbA_{1c} levels in 25% (5/20) of the studies with feedback (not reported), the overall effect in 25% (5/20) of the studies with feedback (more than once a month) was –0.23 (95% CI –0.449 to –0.002; *P*=.048), and the overall effect in 50% (10/20) of the studies with feedback (less than or equal to once a month) was –0.27 (95% CI –0.41 to –0.12; *P*<.001); the results of the study were statistically significant.

Communication Forms Between Patients and Providers

The choice of provider-to-patient communication forms—smartphone apps (MD –0.37, 95% CI –0.53 to –0.21; *P*<.001) and web conferences (MD –0.34, 95% CI –0.65 to –0.02; *P*=.04)—significantly influenced the effect of telemedicine on HbA_{1c} levels. In addition, the choice of patient-to-provider communication in the form of smartphone apps (MD –0.37, 95% CI –0.51 to –0.22; *P*<.001), smart wearable devices (MD –0.20, 95% CI –0.37 to –0.03; *P*=.02), and websites (MD –0.53, 95% CI –0.89 to –0.18; *P*=.003) had a significant impact on the effect on HbA_{1c} levels. A statistically significant subgroup difference was found in the subgroup analysis based on patient-to-provider communication forms (*P*=.002).

Forms of Technology

Subgroup analysis by forms of technology showed that studies using software (MD –0.23, 95% CI –0.36 to –0.11; *P*<.001) had a significant effect on glycemic control in children and adolescents with T1DM compared with studies using only hardware (MD –0.08, 95% CI –0.51 to 0.35; *P*=.71).

Modes of Technology Use

The overall effect on HbA_{1c} levels in the 30% (6/20) of the studies with independent use of technology was –0.11 (95% CI –0.31 to 0.09; *P*=.27), whereas the overall effect on HbA_{1c} levels in the 45% (9/20) of the studies with parental assistance was –0.24 (95% CI –0.41 to –0.07; *P*<.001).

Forms of Telemedicine Interventions

Subgroup analysis based on the form of telemedicine intervention showed that complete telemedicine interventions (MD –0.28, 95% CI –0.43 to –0.13; *P*<.001) were better than partial telemedicine interventions (MD –0.14, 95% CI –0.29 to 0.01; *P*=.06).

XSL•FO
RenderX

Content of Telemedicine Interventions

Interventions with interactive communication and follow-up (MD –0.27, 95% CI –0.38 to –0.17; P<.001) and medication dose adjustment (MD –0.25, 95% CI –0.37 to –0.12; P<.001) were associated with a greater improvement in HbA_{1c} levels. However, interventions without a physical exercise feature also significantly influenced the effect of telemedicine on HbA_{1c} levels (MD –0.21, 95% CI –0.35 to –0.07; P=.004). Moreover, a decrease in heterogeneity and statistically significant subgroup differences was found in the subgroup analysis based on interactive communication and follow-up (P=.002), which can also explain the heterogeneity in the overall effect on HbA_{1c} levels.

Sensitivity Analysis

Leave-one-out analysis was performed by removing each study, and there was no significant change in the effect size (Figure S2 in Multimedia Appendix 1). Accordingly, no individual study had a statistically significant effect on the overall result. However, inspection of the effect size identified one outlier study [25] with an effect size larger than that of the other studies. The exclusion of this study did not materially affect our results for the primary outcome, but it did reduce heterogeneity (I^2=9%; Q_{18}=19.87; P=.34; fixed-effects model) and increase the impact of telemedicine (MD –0.26, 95% CI –0.36 to –0.17; P<.001).

Publication Bias

The contour funnel plot of HbA_{1c} levels was not obviously asymmetrical, consistent with publication bias (Figure S3 in Multimedia Appendix 1). We used the Egger regression test and Begg test to verify publication bias. The regression analysis bias estimate was insignificant (Egger test: bias=–1.02 and P=.32; Begg test: z=0.16 and P=.87).

Meta-Regression

The results of the meta-regression are presented in Table S2 in Multimedia Appendix 1. Meta-regression analysis showed that publication date (P=.04) and the “Interactive follow-up” intervention characteristic (P=.02) were moderating factors to explain the heterogeneity in this study.

Discussion

Principal Findings

In this systematic review and meta-analysis of RCTs comparing telemedicine with usual care, the difference in HbA_{1c} levels was in favor of telemedicine (MD –0.22; P<.001). Sensitivity analysis showed low heterogeneity (I^2=35%; P=.06) and stability of the outliers. Subgroup analyses revealed that studies published after 2010, studies with <6 months of follow-up, studies in children with T1DM, studies in the form of smartphone apps (provider to patient) and websites (patient to provider) for communication, and studies with medication dose adjustment reported significantly larger effects of telemedicine. We were delighted to find that smartphone apps may be a particularly effective way of connecting providers and patients and that telemedicine improves quality of life for children and adolescents with T1DM (SMD –0.24, 95% CI –0.45 to –0.02; P=.04; I^2=0%). However, there was no direct evidence that telemedicine could reduce the risk of hypoglycemia and improve SMBG. Our findings may help guide future clinical decision-making about the use of telemedicine for T1DM in children and adolescents.

Comparison With Prior Work

Our results showed that telemedicine interventions significantly reduced HbA_{1c} levels in children and adolescents with T1DM, which is similar to the results of previous meta-analyses in adults [20,55-57]. A recent study pointed out that a telemedicine intervention for HbA_{1c} in adults had a significant treatment effect [18]. In addition, Shulman et al [58] found no evidence for the effectiveness of telemedicine on HbA_{1c} levels in a 2010 meta-analysis specifically targeting T1DM in adolescents, which is consistent with the results of this study’s time-of-publication subgroup analysis. This suggested that telemedicine has evolved and improved rapidly over the past decade or so and is showing benefits for the treatment of children and adolescents with T1DM. In addition, the results of this study are contrary to the findings of the study by Lee et al [20], which did not find that telemedicine improved glycemic control in children and adolescents with T1DM by subgroup analysis.

Although improvements in the secondary outcomes of hypoglycemia risk and SMBG were not confirmed, it is encouraging to find that telemedicine improves quality of life in children and adolescents with T1DM. This is in contrast to previous studies with adolescents and children, where Shulman et al [58] did not find differences in quality of life between the telemedicine and control groups, and is also contrary to the results of previous studies [55,57] that did not restrict the type of diabetes and studies on T1DM [20] that did not restrict the population, which did not find a benefit of telemedicine in terms of quality of life.

In terms of studying the effect of follow-up time on HbA_{1c} levels, previous studies (not specifically for T1DM) [55,57,59] have shown that the effectiveness of telemedicine is higher when the intervention duration is at least 6 months. However, our findings are contrary to those presented in these studies. Our subgroup analysis showed a higher treatment effect in studies that lasted <6 months than in studies that lasted at least 6 months. This may be related to the “honeymoon” phase of T1DM. A “honeymoon” phase is a transient period of T1DM remission characterized by a significant reduction in insulin requirements and good glycemic control due to a temporary restoration of pancreatic β-cell function, which usually lasts for several months. The exact mechanisms are still uncertain, but one of the generally recognized mechanisms is that correction of “glucotoxicity” by exogenous insulin therapy leads to “β-cell rest” and β-cell recovery [60]. The concept of a “honeymoon” phase was first described by Jackson et al [61]. They observed a rapid decline in demand for exogenous insulin in children with diabetes after regular insulin treatment. In general, patients enter the “honeymoon” period approximately 3 months after starting insulin therapy, and it can last 6 to 9 months. Therefore, it is reasonable to speculate that, in T1DM studies with shorter intervention durations, patients are more likely to be influenced

XSL•FO
RenderX

by the "honeymoon" period and, thus, show a better intervention effect. Future RCTs in this area should carefully consider the duration of telemedicine interventions in their design, which should be >6 months if possible, especially if it is not sufficiently known whether the enrolled group is in or has passed the "honeymoon" period. This is to minimize the effects of the intervention being influenced by the "honeymoon" period and improve the realism and reliability of the effectiveness of telemedicine interventions. In addition, this may be related to the fact that this study targeted children and adolescents with T1DM. An alternative explanation might be that patients become less responsive to monitoring prompts as the potential novelty of telemedicine interventions diminishes. This explanation is well recognized in the related area of activity tracking via smart wearable devices [62].

Our subgroup analysis results suggested differences between children and adolescents. Telemedicine interventions had a greater effect in children compared with adolescents. This contrasts with the findings of the study by Shulman et al [58], which showed no difference in HbA_{1c} levels between the adolescent and child subgroups at the end of the intervention. It may also be due to the use of different criteria for defining children in this study from those used by Shulman et al [58]. The most recent age criteria for children and adolescents used in this study limit the age of children to less than or equal to 10 years; however, based on speculation about the publication date of the study by Schulman et al [58], they may have defined the age of the children as older. Thus based on the age criteria of the present study, we anticipate that more child-related studies in the future may make this difference more apparent. By conducting subgroup analyses, we preliminarily excluded the influence of technology forms and use modes on this result. A total of 10% (2/20) of the studies were conducted on children, one using a hardware device independently [32] and the other using software with parental assistance [36]. However, we found that the studies on children were all complete telemedicine interventions. Subgroup analysis based on intervention form showed that complete telemedicine interventions were better than partial telemedicine interventions, which could explain the observed results. This finding is supported by the study by Chen et al [63], which found that a mixed complete telemedicine intervention was superior to a partial telemedicine intervention in reducing the incidence of pressure injury in patients with spinal cord injury. Another plausible explanation is that children's blood glucose is more prone to fluctuations and a higher incidence of hypoglycemia compared to that of adolescents, which may lead to an exaggerated intervention effect. Although HbA_{1c} is the gold standard for long-term glycemic control, the use of HbA_{1c} alone to assess glycemic management in children can be misleading due to the magnitude of blood glucose fluctuations [64], and the pursuit of HbA_{1c} compliance can be accompanied by an increase in the frequency of hypoglycemia [65,66]. Hypoglycemia in children is a metabolic-endocrine emergency due to the potential for brain injury; permanent neurological sequelae; and, in rare cases, death [67]. Therefore, when assessing glycemic control in children, special attention should be paid to the incidence of hypoglycemia. We also found that telemedicine interventions with medication dose adjustment reported significant treatment effects in improving glycemic control in children and adolescents, consistent with the results of a study [55] on the effects of telemedicine on HbA_{1c} levels in patients with diabetes. Consequently, future well-designed studies should consider further enhancing insulin adjustment and monitoring in the intervention.

On the basis of the subgroup analysis by communication form, our results suggested that smartphone apps may be a very effective vehicle for linking intervention providers and patients, which can provide an intelligent management pathway for blood glucose in children and adolescents with T1DM. Nkhoma et al [68] also supported that smartphone apps improved glycemic control better than other tools. Moreover, the smartphone app studies included in this review (5/20, 25%) all evaluated the safety of apps and reported the incidence of adverse events such as hypoglycemia and diabetic ketoacidosis. Overall, smartphone apps are safe and do not increase the number of episodes of hypoglycemia [69]. Future studies could conduct an in-depth analysis of various types of smartphone apps in terms of core functionality (eg, health monitoring, smart health interventions and guidance, community interactions, and professional support), interface design and interaction experience, and dynamic sensing and self-adaptation (eg, automatically recommending personalized health plans based on the user's basic information, such as age, gender, and body weight) to further improve the telemedicine intervention's usability and effectiveness. This will enable children or adolescents with T1DM to benefit more from telemedicine.

Concerning cost-effectiveness, evidence is still lacking. Few studies included in this meta-analysis (1/20, 5%) discussed cost considerations, which is a common issue faced by telemedicine intervention studies. However, there are specific telemedicine cost analysis studies that may provide assistance with cost considerations. In a recently published study on the cost-effectiveness of telemedicine interventions, smartphone app, SMS text messaging, and website interventions were confirmed to be cost-effective without substantial differences among the different delivery modes [70]. A study by Elliott et al [71] showed that smart wearable devices increase short-term costs but their HbA_{1c}-lowering benefits will provide sufficient long-term health benefits and cost savings to justify the costs as long as the effects last into the medium term. The implementation of telemedicine services continues to be limited by cost and reimbursement barriers; future studies should increase transparency and conduct rigorous and in-depth cost-effectiveness analyses of the various types of telemedicine strategies to support T1DM management.

Practice, Policy, and Future Study

Our findings have potential ramifications for practice and policy. First, among studies evaluating the use of telemedicine interventions to improve care for children and adolescents with T1DM, we found that all (20/20, 100%) focused on HbA_{1c}, with only a small proportion of studies (9/20, 45%) reporting other outcomes such as quality of life and incidence of hypoglycemia. This prevents policy makers from considering the impact of interventions on outcomes other than HbA_{1c} when developing

XSL•FO
RenderX

and implementing telemedicine interventions for this population. This situation may result in the health care system failing to respond to the needs of children and adolescents with T1DM and creates difficulties in tailoring telemedicine interventions to this population [72]. Focusing only on HbA_{1c} may, in turn, compromise the continuity of managed care for patients with T1DM. Therefore, we suggest that future studies add the assessment of other important outcomes such as quality of life, incidence of hypoglycemia, SMBG, and cost-effectiveness.

However, the importance of HbA_{1c} is undisputed, with findings published by the UK Prospective Diabetes Study as early as 2000 showing that a 1% reduction in mean HbA_{1c} levels was associated with a 21% reduction in diabetes-related deaths, a 14% reduction in the risk of myocardial infarction, and a 37% reduction in microvascular complications in patients with type 2 diabetes mellitus [73]. Results of a recent cross-sectional study of 156,090 children and adolescents with T1DM showed that the probability of diabetic retinopathy increased with increasing HbA_{1c} levels (adjusted odds ratio per 1 mmol/mol increase in HbA_{1c} levels 1.03, 95% CI 1.03-1.03; $P<.001$) [74]. Therefore, if telemedicine could be implemented in all children and adolescents with T1DM, it would help reduce the risk of macrovascular and microvascular complications, improve glycemic control, and enhance quality of life.

In light of the aforementioned, our findings suggest a promising application of telemedicine in the management of the disease in children and adolescents with T1DM, especially after several decades of development, during which telemedicine has shown many benefits for children and adolescents with T1DM. Future studies should carefully consider the various forms of interventions as well as the age group of the target population when tailoring telemedicine interventions for T1DM in adolescents and children, particularly with regard to the need for self-monitoring and recognition of hypoglycemia. Although the results of this study suggest that smartphone apps may be the best way to improve patients' glycemic control, they may not be applicable to children aged <10 years. Taking China as an example, in addition to Chinese education policy discouraging the use of electronic devices in schools to minimize disruption and promote traditional teaching methods, children's weaker self-control and potential addiction to gaming and entertainment, difficulties in parental supervision, and adverse effects on children's face-to-face interactions and social skill development are important factors that make it difficult to apply this form of telemedicine.

Finally, this study also identified the lowest threshold of intervention duration intervals that may be able to safeguard the effectiveness of telemedicine interventions in children and adolescents with T1DM, making it necessary to conduct further studies with longer durations and larger cohort sizes in the future to determine the optimal intervals of intervention duration. Although this may be difficult; patients' ability to improve their self-management of glycemia through telemedicine is a gradual process involving multiple factors, including patients' learning ability, adaptability, acceptance of the technology, and the level of support from the health care team; and the time to achieve independent glycemic management may vary due to individual differences, the conduct of studies of longer durations is still very much appreciated.

Strengths and Limitations

This systematic review and meta-analysis has several strengths. To our knowledge, this is the first meta-analysis on telemedicine aimed at improving HbA_{1c} levels in children and adolescents. The substantial number of included RCTs and participants provided strong evidence for the clinical application of telemedicine for improving glycemic control in children and adolescents with T1DM. Second, we performed a relatively comprehensive subgroup analysis and confirmed that telemedicine may have the opposite effect in children and adolescents than in adults in terms of intervention duration. In addition, we undertook a comprehensive search of multiple databases and strictly adhered to methodological tools to report our research. Finally, we performed a leave-one-out sensitivity analysis, which allowed us to assess whether high-risk studies influenced the final results; however, excluding the high-risk study did not change the final results.

We also acknowledge that this meta-analysis has several limitations, mainly statistical assumptions such as deriving the mean and SD from the sample size, baseline, end point, and median, although these assumptions were robust in several sensitivity analyses. Second, data extraction could have included more baseline data from the study, such as medication use since diagnosis (total daily insulin dose, number of insulin injections per day, and insulin pump use), ethnicity, and nationality. Third, there was a certain degree of heterogeneity in the different types of telemedicine interventions. However, subgroup analysis should overcome this flaw. Fourth, only RCTs were included in this research; observational studies may yield pertinent insights for the correlation between telemedicine and HbA_{1c} levels. Fifth, most RCTs (15/20, 75%) did not explicitly report blinding or allocation concealment procedures because of intervention characteristic limitations, which would lead to performance and detection biases. Sixth, the precision of some secondary outcomes was relatively low because of the small number of relevant trials. More RCTs of high quality and with large sample sizes are needed for further validation. Finally, only articles published in English were reviewed, which would lead to potential selection bias, and therefore, the results' generalizability may be limited.

Conclusions

Our systematic review and meta-analysis has shown that telemedicine is an efficacious and safe intervention approach. It can reduce HbA_{1c} levels and improve quality of life in children and adolescents with T1DM. Complete telemedicine interventions are better than partial telemedicine interventions. However, in accordance with the idea of providing health care from a distance, telemedicine should not be regarded as a uniform approach to medication or as an alternative to usual care but rather as a useful supplement to usual care to control HbA_{1c} levels and a potentially cost-effective mode. Given the potential benefits of telemedicine, such as greater access for remote populations or people with ambulatory restrictions, these findings may encourage further implementation of eHealth

XSL•FO
RenderX

strategies for T1DM management, particularly as part of multifaceted interventions for integrated care of chronic diseases. The aforementioned conclusions need to be further verified in future studies. Meanwhile, researchers should develop higher-quality RCTs using large samples that focus on hard clinical outcomes, cost-effectiveness, and quality of life.

Acknowledgments

The authors are sincerely grateful for funding from Fujian Provincial Health Science and Technology Program project funding scheme for 2021 by the Fujian Provincial Health Commission (grant 2021CXB007). They thank all the individuals who contributed to this study.

Authors' Contributions

KZ and QW contributed to the study concept and design. KZ drafted the manuscript. QH helped draft the manuscript. KZ, CL, and QH assessed the risk of bias. QH and QW assessed the quality of evidence of each study. KZ and CL independently extracted the data for analysis. KZ, QZ, DX, CX, MZ, and RL were involved in discussing earlier versions of the text. All authors participated in the study design and read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy, funnel plot for primary outcome, forest plots of the subgroups and the secondary outcomes, summary of meta-regression results, and sensitivity analysis.
[DOCX File , 2062 KB - jmir_v26i1e51538_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 34 KB - jmir_v26i1e51538_app2.docx]

References

1. Sperling MA, Laffel LM. Current management of glycemia in children with type 1 diabetes mellitus. N Engl J Med 2022 Mar 24;386(12):1155-1164. [doi: 10.1056/NEJMcp2112175] [Medline: 35320645]
2. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. Lancet 2018 Jun 16;391(10138):2449-2462 [FREE Full text] [doi: 10.1016/S0140-6736(18)31320-5] [Medline: 29916386]
3. Magliano DJ, Boyko EJ. IDF diabetes atlas: diabetes around the world in 2021. International Diabetes Federation. URL: https://diabetesatlas.org/atlas/tenth-edition/ [accessed 2024-04-29]
4. Gregory JW, Cameron FJ, Joshi K, Eiswirth M, Garrett C, Garvey K, et al. ISPAD clinical practice consensus guidelines 2022: diabetes in adolescence. Pediatr Diabetes 2022 Nov;23(7):857-871 [FREE Full text] [doi: 10.1111/pedi.13408] [Medline: 36250644]
5. von Scholten BJ, Kreiner FF, Gough SC, von Herrath M. Current and future therapies for type 1 diabetes. Diabetologia 2021 May;64(5):1037-1048 [FREE Full text] [doi: 10.1007/s00125-021-05398-3] [Medline: 33595677]
6. Bjornstad P, Donaghue KC, Maahs DM. Macrovascular disease and risk factors in youth with type 1 diabetes: time to be more attentive to treatment? Lancet Diabetes Endocrinol 2018 Oct;6(10):809-820 [FREE Full text] [doi: 10.1016/S2213-8587(18)30035-4] [Medline: 29475800]
7. Marcovecchio ML, Dalton RN, Daneman D, Deanfield J, Jones TW, Neil HA, Adolescent type 1 Diabetes cardio-renal Intervention Trial (AdDIT) study group. A new strategy for vascular complications in young people with type 1 diabetes mellitus. Nat Rev Endocrinol 2019 Jul;15(7):429-435. [doi: 10.1038/s41574-019-0198-2] [Medline: 30996294]
8. Katsarou A, Gudbjörnsdottir S, Rawshani A, Dabelea D, Bonifacio E, Anderson BJ, et al. Type 1 diabetes mellitus. Nat Rev Dis Primers 2017 Mar 30;3:17016. [doi: 10.1038/nrdp.2017.16] [Medline: 28358037]
9. Teo E, Hassan N, Tam W, Koh S. Effectiveness of continuous glucose monitoring in maintaining glycaemic control among people with type 1 diabetes mellitus: a systematic review of randomised controlled trials and meta-analysis. Diabetologia 2022 Apr;65(4):604-619. [doi: 10.1007/s00125-021-05648-4] [Medline: 35141761]
10. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-04-29]
11. Chan RJ, Crichton M, Crawford-Williams F, Agbejule OA, Yu K, Hart NH, Multinational Association of Supportive Care in Cancer (MASCC) Survivorship Study Group. The efficacy, challenges, and facilitators of telemedicine in post-treatment cancer survivorship care: an overview of systematic reviews. Ann Oncol 2021 Dec;32(12):1552-1570 [FREE Full text] [doi: 10.1016/j.annonc.2021.09.001] [Medline: 34509615]

XSL•FO
RenderX

12. Cunha AS, Pedro AR, Cordeiro JV. Facilitators of and barriers to accessing hospital medical specialty telemedicine consultations during the COVID-19 pandemic: systematic review. J Med Internet Res 2023 Jul 10;25:e44188 [FREE Full text] [doi: 10.2196/44188] [Medline: 37262124]
13. Anderson K, Francis T, Ibanez-Carrasco F, Globerman J. Physician's perceptions of telemedicine in HIV care provision: a cross-sectional web-based survey. JMIR Public Health Surveill 2017 May 30;3(2):e31 [FREE Full text] [doi: 10.2196/publichealth.6896] [Medline: 28559226]
14. Sengupta A, Pettigrew S, Jenkins CR. Telemedicine in specialist outpatient care during COVID-19: a qualitative study. Intern Med J 2024 Jan;54(1):54-61. [doi: 10.1111/imj.16288] [Medline: 37926924]
15. Ruiz de Adana MS, Alhambra-Expósito MR, Muñoz-Garach A, Gonzalez-Molero I, Colomo N, Torres-Barea I, Diabetes Group of SAEDYN (Andalusian Society of Endocrinology, Diabetes, Nutrition). Randomized study to evaluate the impact of telemedicine care in patients with type 1 diabetes with multiple doses of insulin and suboptimal HbA in Andalusia (Spain): PLATEDIAN study. Diabetes Care 2020 Feb;43(2):337-342 [FREE Full text] [doi: 10.2337/dc19-0739] [Medline: 31831473]
16. Molavynejad S, Miladinia M, Jahangiri M. A randomized trial of comparing video telecare education vs. in-person education on dietary regimen compliance in patients with type 2 diabetes mellitus: a support for clinical telehealth Providers. BMC Endocr Disord 2022 May 02;22(1):116 [FREE Full text] [doi: 10.1186/s12902-022-01032-4] [Medline: 35501846]
17. Bisno DI, Reid MW, Fogel JL, Pyatak EA, Majidi S, Raymond JK. Virtual group appointments reduce distress and improve care management in young adults with type 1 diabetes. J Diabetes Sci Technol 2022 Nov;16(6):1419-1427 [FREE Full text] [doi: 10.1177/19322968211035768] [Medline: 34328029]
18. Udsen FW, Hangaard S, Bender C, Andersen J, Kronborg T, Vestergaard P, et al. The effectiveness of telemedicine solutions in type 1 diabetes management: a systematic review and meta-analysis. J Diabetes Sci Technol 2023 May;17(3):782-793 [FREE Full text] [doi: 10.1177/19322968221076874] [Medline: 35135365]
19. Eberle C, Stichling S. Clinical improvements by telemedicine interventions managing type 1 and type 2 diabetes: systematic meta-review. J Med Internet Res 2021 Feb 19;23(2):e23244 [FREE Full text] [doi: 10.2196/23244] [Medline: 33605889]
20. Lee SW, Ooi L, Lai YK. Telemedicine for the management of glycemic control and clinical outcomes of type 1 diabetes mellitus: a systematic review and meta-analysis of randomized controlled studies. Front Pharmacol 2017;8:330 [FREE Full text] [doi: 10.3389/fphar.2017.00330] [Medline: 28611672]
21. Chase HP, Pearson JA, Wightman C, Roberts MD, Oderberg AD, Garg SK. Modem transmission of glucose values reduces the costs and need for clinic visits. Diabetes Care 2003 May;26(5):1475-1479. [doi: 10.2337/diacare.26.5.1475] [Medline: 12716807]
22. Gandrud L, Altan A, Buzinec P, Hemphill J, Chatterton J, Kelley T, et al. Intensive remote monitoring versus conventional care in type 1 diabetes: a randomized controlled trial. Pediatr Diabetes (Forthcoming) 2018 Feb 21. [doi: 10.1111/pedi.12654] [Medline: 29464831]
23. Goyal S, Nunn CA, Rotondi M, Couperthwaite AB, Reiser S, Simone A, et al. A mobile app for the self-management of type 1 diabetes among adolescents: a randomized controlled trial. JMIR Mhealth Uhealth 2017 Jun 19;5(6):e82 [FREE Full text] [doi: 10.2196/mhealth.7336] [Medline: 28630037]
24. Han Y, Faulkner MS, Fritz H, Fadoju D, Muir A, Abowd GD, et al. A pilot randomized trial of text-messaging for symptom awareness and diabetes knowledge in adolescents with type 1 diabetes. J Pediatr Nurs 2015 Nov;30(6):850-861 [FREE Full text] [doi: 10.1016/j.pedn.2015.02.002] [Medline: 25720675]
25. Ibrahim N, Treluyer JM, Briand N, Godot C, Polak M, Beltrand J. Text message reminders for adolescents with poorly controlled type 1 diabetes: a randomized controlled trial. PLoS One 2021;16(3):e0248549 [FREE Full text] [doi: 10.1371/journal.pone.0248549] [Medline: 33720997]
26. Kowalska A, Piechowiak K, Ramotowska A, Szypowska A. Impact of ELKa, the electronic device for prandial insulin dose calculation, on metabolic control in children and adolescents with type 1 diabetes mellitus: a randomized controlled trial. J Diabetes Res 2017;2017:1708148 [FREE Full text] [doi: 10.1155/2017/1708148] [Medline: 28232949]
27. Kumar VS, Wentzell KJ, Mikkelsen T, Pentland A, Laffel LM. The DAILY (daily automated intensive log for youth) trial: a wireless, portable system to improve adherence and glycemic control in youth with diabetes. Diabetes Technol Ther 2004 Aug;6(4):445-453. [doi: 10.1089/1520915041705893] [Medline: 15320998]
28. Landau Z, Mazor-Aronovitch K, Boaz M, Blaychfeld-Magnazi M, Graph-Barel C, Levek-Motola N, et al. The effectiveness of internet-based blood glucose monitoring system on improving diabetes control in adolescents with type 1 diabetes. Pediatr Diabetes 2012 Mar;13(2):203-207. [doi: 10.1111/j.1399-5448.2011.00800.x] [Medline: 21848925]
29. Marrero DG, Vandagriff JL, Kronz K, Fineberg NS, Golden MP, Gray D, et al. Using telecommunication technology to manage children with diabetes: the computer-linked outpatient clinic (CLOC) study. Diabetes Educ 1995;21(4):313-319. [doi: 10.1177/014572179502100409] [Medline: 7621734]
30. Mulvaney SA, Rothman RL, Wallston KA, Lybarger C, Dietrich MS. An internet-based program to improve self-management in adolescents with type 1 diabetes. Diabetes Care 2010 Mar;33(3):602-604 [FREE Full text] [doi: 10.2337/dc09-1881] [Medline: 20032275]

XSL•FO
RenderX

31. Nunn E, King B, Smart C, Anderson D. A randomized controlled trial of telephone calls to young patients with poorly controlled type 1 diabetes. Pediatr Diabetes 2006 Oct;7(5):254-259. [doi: 10.1111/j.1399-5448.2006.00200.x] [Medline: 17054446]
32. Raviteja KV, Kumar R, Dayal D, Sachdeva N. Clinical efficacy of professional continuous glucose monitoring in improving glycemic control among children with type 1 diabetes mellitus: an open-label randomized control trial. Sci Rep 2019 Apr 16;9(1):6120 [FREE Full text] [doi: 10.1038/s41598-019-42555-6] [Medline: 30992480]
33. Schiaffini R, Tagliente I, Carducci C, Ullmann N, Ciampalini P, Lorubbio A, et al. Impact of long-term use of eHealth systems in adolescents with type 1 diabetes treated with sensor-augmented pump therapy. J Telemed Telecare 2016 Jul;22(5):277-281. [doi: 10.1177/1357633X15598425] [Medline: 26289613]
34. Stanger C, Lansing AH, Scherer E, Budney A, Christiano AS, Casella SJ. A web-delivered multicomponent intervention for adolescents with poorly controlled type 1 diabetes: a pilot randomized controlled trial. Ann Behav Med 2018 Nov 12;52(12):1010-1022 [FREE Full text] [doi: 10.1093/abm/kay005] [Medline: 30418521]
35. von Sengbusch S, Eisemann N, Mueller-Godeffroy E, Lange K, Doerdelmann J, Erdem A, et al. Outcomes of monthly video consultations as an add-on to regular care for children with type 1 diabetes: a 6-month quasi-randomized clinical trial followed by an extension phase. Pediatr Diabetes 2020 Dec;21(8):1502-1515. [doi: 10.1111/pedi.13133] [Medline: 33009690]
36. Ware J, Allen JM, Boughton CK, Wilinska ME, Hartnell S, Thankamony A, KidsAP Consortium. Randomized trial of closed-loop control in very young children with type 1 diabetes. N Engl J Med 2022 Jan 20;386(3):209-219. [doi: 10.1056/NEJMoa2111673] [Medline: 35045227]
37. Ware J, Boughton CK, Allen JM, Wilinska ME, Tauschmann M, Denvir L, DAN05 Consortium. Cambridge hybrid closed-loop algorithm in children and adolescents with type 1 diabetes: a multicentre 6-month randomised controlled trial. Lancet Digit Health 2022 Apr;4(4):e245-e255 [FREE Full text] [doi: 10.1016/S2589-7500(22)00020-6] [Medline: 35272971]
38. Xu Y, Xu L, Zhao W, Li Q, Li M, Lu W, et al. Effectiveness of a WeChat combined continuous flash glucose monitoring system on glycemic control in juvenile type 1 diabetes mellitus management: randomized controlled trial. Diabetes Metab Syndr Obes 2021;14:1085-1094 [FREE Full text] [doi: 10.2147/DMSO.S299070] [Medline: 33727842]
39. Klee P, Bussien C, Castellsague M, Combescure C, Dirlewanger M, Girardin C, et al. An intervention by a patient-designed do-it-yourself mobile device app reduces HbA1c in children and adolescents with type 1 diabetes: a randomized double-crossover study. Diabetes Technol Ther 2018 Dec;20(12):797-805. [doi: 10.1089/dia.2018.0255] [Medline: 30403495]
40. Shalitin S, Ben-Ari T, Yackobovitch-Gavan M, Tenenbaum A, Lebenthal Y, de Vries L, et al. Using the Internet-based upload blood glucose monitoring and therapy management system in patients with type 1 diabetes. Acta Diabetol 2014 Apr;51(2):247-256. [doi: 10.1007/s00592-013-0510-x] [Medline: 23982170]
41. Campbell F, Biggs K, Aldiss SK, O'Neill PM, Clowes M, McDonagh J, et al. Transition of care for adolescents from paediatric services to adult health services. Cochrane Database Syst Rev 2016 Apr 29;4(4):CD009794 [FREE Full text] [doi: 10.1002/14651858.CD009794.pub2] [Medline: 27128768]
42. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
43. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, GRADE Working Group. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol 2020 Jun;122:142-152. [doi: 10.1016/j.jclinepi.2019.12.021] [Medline: 32058069]
44. Bayoumy K, Gaber M, Elshafeey A, Mhaimeed O, Dineen EH, Marvel FA, et al. Smart wearable devices in cardiovascular care: where we are and how to move forward. Nat Rev Cardiol 2021 Aug;18(8):581-599 [FREE Full text] [doi: 10.1038/s41569-021-00522-7] [Medline: 33664502]
45. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3. Hoboken, NJ: John Wiley & Sons; 2019.
46. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 2018 Jun;27(6):1785-1805. [doi: 10.1177/0962280216669183] [Medline: 27683581]
47. Shi J, Luo D, Weng H, Zeng XT, Lin L, Chu H, et al. Optimally estimating the sample standard deviation from the five-number summary. Res Synth Methods 2020 Sep;11(5):641-654. [doi: 10.1002/jrsm.1429] [Medline: 32562361]
48. Littell JH, Corcoran J, Pillai V. Systematic Reviews and Meta-Analysis. New York, NY: Oxford University Press; 2008.
49. Mavridis D, Salanti G. How to assess publication bias: funnel plot, trim-and-fill method and selection models. Evid Based Ment Health 2014 Feb;17(1):30. [doi: 10.1136/eb-2013-101699] [Medline: 24477535]
50. Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the effective health care program. J Clin Epidemiol 2011 Nov;64(11):1187-1197. [doi: 10.1016/j.jclinepi.2010.08.010] [Medline: 21477993]
51. Ingersoll GM, Marrero DG. A modified quality-of-life measure for youths: psychometric properties. Diabetes Educ 1991;17(2):114-118. [doi: 10.1177/014572179101700219] [Medline: 1995281]

XSL•FO
RenderX

52. Skinner TC, Hoey H, McGee HM, Skovlund SE, Hvidøre Study Group on Childhood Diabetes. A short form of the diabetes quality of life for youth questionnaire: exploratory and confirmatory analysis in a sample of 2,077 young people with type 1 diabetes mellitus. Diabetologia 2006 Apr;49(4):621-628. [doi: 10.1007/s00125-005-0124-0] [Medline: 16525844]
53. The DCCT Research Group. Reliability and validity of a diabetes quality-of-life measure for the diabetes control and complications trial (DCCT). Diabetes Care 1988 Oct;11(9):725-732. [doi: 10.2337/diacare.11.9.725] [Medline: 3066604]
54. Ravens-Sieberer U, Bullinger M. Assessing health-related quality of life in chronically ill children with the German KINDL: first psychometric and content analytical results. Qual Life Res 1998 Jul;7(5):399-407. [doi: 10.1023/a:1008853819715] [Medline: 9691720]
55. Faruque LI, Wiebe N, Ehteshami-Afshar A, Liu Y, Dianati-Maleki N, Hemmelgarn BR, Alberta Kidney Disease Network. Effect of telemedicine on glycated hemoglobin in diabetes: a systematic review and meta-analysis of randomized trials. CMAJ 2017 Mar 06;189(9):E341-E364 [FREE Full text] [doi: 10.1503/cmaj.150885] [Medline: 27799615]
56. Wang X, Shu W, Du J, Du M, Wang P, Xue M, et al. Mobile health in the management of type 1 diabetes: a systematic review and meta-analysis. BMC Endocr Disord 2019 Feb 13;19(1):21 [FREE Full text] [doi: 10.1186/s12902-019-0347-6] [Medline: 30760280]
57. Correia JC, Meraj H, Teoh SH, Waqas A, Ahmad M, Lapão LV, et al. Telemedicine to deliver diabetes care in low- and middle-income countries: a systematic review and meta-analysis. Bull World Health Organ 2021 Mar 01;99(3):209-19B [FREE Full text] [doi: 10.2471/BLT.19.250068] [Medline: 33716343]
58. Shulman RM, O'Gorman CS, Palmert MR. The impact of telemedicine interventions involving routine transmission of blood glucose data with clinician feedback on metabolic control in youth with type 1 diabetes: a systematic review and meta-analysis. Int J Pediatr Endocrinol 2010;2010:536957 [FREE Full text] [doi: 10.1155/2010/536957] [Medline: 20886054]
59. Tchero H, Kangambega P, Briatte C, Brunet-Houdard S, Retali GR, Rusch E. Clinical effectiveness of telemedicine in diabetes mellitus: a meta-analysis of 42 randomized controlled trials. Telemed J E Health 2019 Jul;25(7):569-583. [doi: 10.1089/tmj.2018.0128] [Medline: 30124394]
60. Zhong T, Tang R, Gong S, Li J, Li X, Zhou Z. The remission phase in type 1 diabetes: changing epidemiology, definitions, and emerging immuno-metabolic mechanisms. Diabetes Metab Res Rev 2020 Feb;36(2):e3207. [doi: 10.1002/dmrr.3207] [Medline: 31343814]
61. Jackson RL, Boyd JB, Smith TE. Stabilization of the diabetic child. Arch Pediatr Adolesc Med 1940 Feb 01;59(2):332-341 [FREE Full text] [doi: 10.1001/archpedi.1940.01990130115008]
62. Shin G, Feng Y, Jarrahi MH, Gafinowitz N. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. JAMIA Open 2019 Apr;2(1):62-72 [FREE Full text] [doi: 10.1093/jamiaopen/ooy048] [Medline: 31984346]
63. Chen G, Wang T, Zhong L, He X, Huang C, Wang Y, et al. Telemedicine for preventing and treating pressure injury after spinal cord injury: systematic review and meta-analysis. J Med Internet Res 2022 Sep 07;24(9):e37618 [FREE Full text] [doi: 10.2196/37618] [Medline: 36069842]
64. Beck RW, Connor CG, Mullen DM, Wesley DM, Bergenstal RM. The fallacy of average: how using HbA (1c) alone to assess glycemic control can be misleading. Diabetes Care 2017 Aug;40(8):994-999 [FREE Full text] [doi: 10.2337/dc17-0636] [Medline: 28733374]
65. Gimenez M, Tannen AJ, Reddy M, Moscardo V, Conget I, Oliver N. Revisiting the relationships between measures of glycemic control and hypoglycemia in continuous glucose monitoring data sets. Diabetes Care 2018 Feb;41(2):326-332 [FREE Full text] [doi: 10.2337/dc17-1597] [Medline: 29191845]
66. Seyed Ahmadi S, Westman K, Pivodic A, Ólafsdóttir AF, Dahlqvist S, Hirsch IB, et al. The association between HbA (1c) and time in hypoglycemia during CGM and self-monitoring of blood glucose in people with type 1 diabetes and multiple daily insulin injections: a randomized clinical trial (GOLD-4). Diabetes Care 2020 Sep;43(9):2017-2024 [FREE Full text] [doi: 10.2337/dc19-2606] [Medline: 32641374]
67. Quarta A, Iannucci D, Guarino M, Blasetti A, Chiarelli F. Hypoglycemia in children: major endocrine-metabolic causes and novel therapeutic perspectives. Nutrients 2023 Aug 11;15(16):3544 [FREE Full text] [doi: 10.3390/nu15163544] [Medline: 37630734]
68. Nkhoma DE, Soko CJ, Bowrin P, Manga YB, Greenfield D, Househ M, et al. Digital interventions self-management education for type 1 and 2 diabetes: a systematic review and meta-analysis. Comput Methods Programs Biomed 2021 Oct;210:106370. [doi: 10.1016/j.cmpb.2021.106370] [Medline: 34492544]
69. Pi L, Shi X, Wang Z, Zhou Z. Effect of smartphone apps on glycemic control in young patients with type 1 diabetes: a meta-analysis. Front Public Health 2023;11:1074946 [FREE Full text] [doi: 10.3389/fpubh.2023.1074946] [Medline: 37064701]
70. Willems R, Annemans L, Siopis G, Moschonis G, Vedanthan R, Jung J, DigiCare 4You. Cost effectiveness review of text messaging, smartphone application, and website interventions targeting T2DM or hypertension. NPJ Digit Med 2023 Aug 18;6(1):150 [FREE Full text] [doi: 10.1038/s41746-023-00876-x] [Medline: 37596488]
71. Elliott RA, Rogers G, Evans ML, Neupane S, Rayman G, Lumley S, FLASH-UK Trial Study Group. Estimating the cost-effectiveness of intermittently scanned continuous glucose monitoring in adults with type 1 diabetes in England. Diabet Med 2024 Mar;41(3):e15232. [doi: 10.1111/dme.15232] [Medline: 37750427]

XSL•FO
RenderX

72. Zafra-Tanaka JH, Beran D, Bernabe-Ortiz A. Health system responses for type 1 diabetes: a scoping review. Diabet Med 2022 Jul;39(7):e14805 [FREE Full text] [doi: 10.1111/dme.14805] [Medline: 35124856]
73. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412 [FREE Full text] [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
74. Bratina N, Auzanneau M, Birkebaek N, de Beaufort C, Cherubini V, Craig ME, Australasian Diabetes Data Network (ADDN) Study Group, the Prospective Diabetes Follow-up Registry (DPV) initiative, Danish National Diabetes Registry (DanDiabKids), National Pediatric Diabetes Audit (NPDA), Region Marche Registry for Diabetes, Diabeter Diabetes Database, Slovenian Childhood Diabetes Registry, SEARCH for Diabetes in Youth Study. Differences in retinopathy prevalence and associated risk factors across 11 countries in three continents: a cross-sectional study of 156,090 children and adolescents with type 1 diabetes. Pediatr Diabetes 2022 Dec;23(8):1656-1664 [FREE Full text] [doi: 10.1111/pedi.13416] [Medline: 36097824]

Abbreviations

DQOLY: Diabetes Quality of Life for Youth
HbA_{1c}: glycated hemoglobin
MD: mean difference
N-QOL: non–youth-specific quality of life
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMBG: self-monitoring of blood glucose
SMD: standardized mean difference
T1DM: type 1 diabetes mellitus

Edited by T Leung, T de Azevedo Cardoso; submitted 02.08.23; peer-reviewed by A Cunha, X Shi, A Zhang; comments to author 02.02.24; revised version received 28.02.24; accepted 21.05.24; published 09.07.24.

Please cite as:
Zhang K, Huang Q, Wang Q, Li C, Zheng Q, Li Z, Xu D, Xie C, Zhang M, Lin R
Telemedicine in Improving Glycemic Control Among Children and Adolescents With Type 1 Diabetes Mellitus: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e51538
URL: https://www.jmir.org/2024/1/e51538
doi:10.2196/51538
PMID:

©Kun Zhang, Qiyuan Huang, Qiaosong Wang, Chengyang Li, Qirong Zheng, Zhuoyue Li, Dan Xu, Cuiling Xie, Mingqi Zhang, Rongjin Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Use of Telepsychiatry Services in Emergency Settings: Scoping Review

Ligat Shalev[1], PhD; Renana Eitan[2*], MD; Adam J Rose[1*], MSc, MD

[1]School of Public Health, Hebrew University, Jerusalem, Israel

[2]Psychiatric Division, Sourasky Medical Center, Tel Aviv-Yafo, Israel

[*]these authors contributed equally

Corresponding Author:
Ligat Shalev, PhD
School of Public Health, Hebrew University
Ein Kerem Campus
Jerusalem, 91120
Israel
Phone: 972 0507554025
Email: Ligat.Shalev@gmail.com

Abstract

Background: Telepsychiatry (TP), a live video meeting, has been implemented in many contexts and settings. It has a distinct advantage in the psychiatric emergency department (ED) setting, as it expedites expert assessments for psychiatric patients. However, limited knowledge exits for TP's effectiveness in the ED setting, as well as the process of implementing TP in this setting.

Objective: This scoping review aimed to review the existing evidence for the administrative and clinical outcomes for TP in the ED setting and to identify the barriers and facilitators to implementing TP in this setting.

Methods: The scoping review was conducted according to the guidelines for the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews). Three electronic databases were examined: PubMed, Embase, and Web of Science. The databases were searched from January 2013 to April 2023 for papers and their bibliography. A total of 2816 potentially relevant papers were retrieved from the initial search. Studies were screened and selected independently by 2 authors.

Results: A total of 11 articles were included. Ten papers reported on administrative and clinical outcomes of TP use in the ED setting and 1 on the barriers and facilitators of its implementation. TP is used in urban and rural areas and for settings with and with no on-site psychiatric services. Evidence shows that TP reduced waiting time for psychiatric evaluation, but in some studies, it was associated with prolonged total length of stay in the ED compared with in-person evaluation. Findings indicate lower admission rates in patients assessed with TP in the ED. Limited data were reported for TP costs, its use for involuntary commitment evaluations, and its use for particular subgroups of patients (eg, those with a particular diagnosis). A single paper examined TP implementation process in the ED, which explored the barriers and facilitators for implementation among patients and staff in a rural setting.

Conclusions: Based on the extant studies, TP seems to be generally feasible and acceptable to key stakeholders. However, this review detected a gap in the literature regarding TP's effectiveness and implementation process in the ED setting. Specific attention should be paid to the examination of this service for specific groups of patients, as well as its use to enable assessments for possible involuntary commitment.

(J Med Internet Res 2024;26:e51814) doi:10.2196/51814

KEYWORDS

implementation science; emergency department; telepsychiatry; organizational innovation; eHealth; mHealth; scoping review; implementation; psychiatric patient; clinical outcome; rural; feasible; acceptable; effectiveness; electronic database

XSL•FO
RenderX

Introduction

Telepsychiatry Use Over the Years

The history of telepsychiatry (TP) began with doubts about its use [1,2]. While there are still questions about TP, it has gained increasing acceptance in recent years, as reflected through changes in relevant regulations [3]. TP is used for psychiatric assessment, treatment, and follow-up [4]. So far, the most prevalent technologies used for TP are by telephone [5], email [6], or recorded or live videos and hybrid models [7]. TP is used in the private [8] and public sectors, including for primary care [9] and secondary care [10]. TP has also been delivered in clinical settings [10] and in patients' home environments [11]. TP has been used to treat different mental health conditions, and in different situations, including suicide attempts, self-harm, schizophrenia, and dual diagnosis of mental health conditions with substance abuse [12]. TP has been adapted to different treatment approaches, including for individuals [13] and group sessions [14]. TP has been used in both urban [15] and rural areas [11].

Current Evidence for Effectiveness of TP and Regarding Its Implementation

Various studies have examined the effectiveness of TP, often compared with face-to-face treatment approaches. In terms of accuracy of diagnosis and treatment decisions, TP has been shown to be as accurate as meeting with patients in person [16,17]. Using TP has been shown to reduce emergency department (ED) length of stay (LOS) by allowing more rapid access to psychiatric expertise [17]. For similar reasons, TP has been shown to reduce admission rates [18]. TP has been used to provide on-site psychiatric services to hospitals that previously did not have any [19]. Both patients [20] and providers [17] showed high satisfaction rates.

TP has also been examined using cost-effectiveness analyses [21], and at least some studies have found that it is cost saving compared with usual care [22,23]. Other studies have examined the process of implementing TP in different settings [9,24-26]. For example, some studies have detailed the experience of implementing remote mental health consultations during the COVID-19 [27], or reasons why some ED directors are avoiding the use of remote services, including TP [28].

Specific Challenges When Using TP for Psychiatric Emergencies

TP has advantages for general use, but it may have a particularly important role in addressing psychiatric emergencies. Most people find the ED an uncomfortable place to be [29]. However, for patients experiencing psychiatric emergencies, the ED may be even worse. The ED may exacerbate patients' agitation, which may put health care providers or other bystanders at risk for violence [29]. In addition to this immediate effect, the ED can also have a long-term effect on psychiatric patients. Faessler and colleagues [30] found that psychological distress could last up to 30 days after ED discharge for patients with psychiatric disorders. Considering these data, TP may be a highly useful solution for ED settings, if it can help minimize patients' time in the ED [31].

In the last few years, several reviews summarized the current evidence of TP services in the ED setting. One study reviewed the current data on acute situations but included not just psychiatric services but other practices and also included home-based services in addition to the ED setting [11]. A second review examined the barriers and facilitators of implementing TP, but most of the studies that were included did not focus on the use of TP for emergency settings [26]. A third study reviewed TP interventions in emergency and crisis situations, but this review included studies published more than a decade ago, when video-link technology was much less developed [32]. Thus, no updated published review of TP use for adult emergencies is available.

Objectives

TP may bridge critical gaps in mental health care access and quality, addressing key issues, such as prolonged waiting times or workforce shortages, that hinder timely and effective patient care. Thus, TP has considerable potential to help improve service delivery and outcomes in mental health. To our knowledge, reviews assessing the use of TP in emergency settings in the past 10 years have not been conducted. Given the limited evidence, the objectives of this review were to search the literature on psychiatric live video meeting in emergency settings, to assess the overall findings regarding clinical and implementation outcomes, and to enumerate the barriers and facilitators for successful implementation.

Methods

Overview

We conducted this scoping review following the methodological guidance proposed by Arksey and O'Malley [33], Levac et al [34], and The Joanna Briggs Institute Reviewers' Manual [35]. The 5 stages used in this scoping review were based on guidelines from Arksey and O'Malley: (1) identifying the research question; (2) identifying the relevant studies; (3) study selection; (4) charting the data; and (5) collating, summarizing, and reporting the results [33]. Our study focused on the current administrative and clinical evidence regarding the use of TP services in the ED setting, as well as the factors affecting their implementation in the ED setting. The reporting of this scoping review was guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for Scoping Reviews (PRISMA-ScR) checklist [36] (Multimedia Appendix 1).

Search Strategy

We searched 3 electronic databases, including PubMed, Embase, and Web of Science, using the following terms and combinations: (1) psychiatry, mental health, mental disorder, mental health care, and mental disease; (2) telepsychiatry, telemedicine, virtual medicine, tele health, eHealth, telecare, tele emergency, and digital mental health; and (3) video, video conference, videoconferencing, conference meeting, streaming, zoom, remote consultation, long distance consultation, distance counseling, eCounseling, and web-based counseling (Multimedia Appendix 2).

XSL•FO
RenderX

Inclusion and Exclusion Criteria

Studies were included based on the inclusion and exclusion criteria, and if they met the population, concept, and context categorization recommended by The Joanna Briggs Institute [35] (Table 1). We included studies focused on individuals 18 years of age and older who had a psychiatric session. Due to the different nature of the following patient groups, they were excluded from the search: children, couple or group sessions, and patients who have been arrested or convicted. While it was allowable for the clinician to address substance abuse as part of the service, we did not include studies solely addressing substance abuse issues. We also focused solely on 2-way video assessment and excluded other modes of communication such as telephone or asynchronous text messages.

We did make an exception to our rule about patient ages to include 3 especially important and broad-based studies. These studies were conducted nationwide [37,38] and statewide [15] and involved patients of all ages. We also focused on studies conducted within the past 10 years, due to significant advances in video-link technology around that time. We focused on the dates of data collection rather than publication, since a variable period may elapse between data collection and publication. Two studies started data collection on October 2012 for and finished collecting data years later; we decided to include these studies [39,40].

Table 1. Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Participants	
Patients who are 18 years of age and older or ED[a] staff that use and report about the telepsychiatry service based on direct experience	• Children and teenagers younger than 18 years
Individuals	• Couple, family, or group • Forensic cases, correctional facilities, or services focused primarily on delivering treatment for substance abuse • Not clinician-to-clinician telepsychiatry service (advice of psychiatrists to general clinicians about patients)
Concept	
Psychiatry services	• Other services than psychiatry (eg, psychology, neurology, or social work)
Use of live video communication	• Asynchrony video communication, telephone, text chat, email, app, or video game
Context	
Emergency settings	• Nonemergency settings or acute services in the community setting
Type of study	
Qualitative, quantitative, or mixed methods studies. Quantitative studies must describe results of at least 30 participants	—[b]
Empirical data with detailed methodology presented in journals, editorials, commentaries, letters to the editor, or scientific reports	• Nonempirical data or empirical data with insufficient description of study methodology. Also, conference abstracts, essays, book chapters, and books, and development of research tools without pilot-testing
Studies that use one of the following designs: observational and experimental, cross-sectional, or longitudinal; randomized controlled trials, nonrandomized or noncontrolled trials	• Case series or case studies
Data collected during the past 10 years (from 2013)	—
Language	
English	• Languages other than English

[a]ED: emergency department.

[b]Not applicable.

Screening and Selection of Studies

The initial search of the 3 databases yielded 2,686 results. The hand search of the selected papers' bibliography identified 130 additional records. After duplicates were removed, 1967 (69.8%) records were reviewed. The titles and abstracts of 69.8% (1967/2816) of the articles were screened, and 89.2% (1754/1967) of the articles were excluded as being not relevant. LS performed the initial screening to identify articles that were clearly not relevant, keeping articles that were questionably relevant or probably relevant. Then LS and AR performed independent full-text review of the 213 retained articles. A total

XSL•FO
RenderX

of 94.8% (202/213) articles were excluded based on the reasons shown in Figure 1. A total of 11 publications were ultimately included in the scoping review. In case of disagreement, LS and AR discussed the article until agreement was reached. The reasons for exclusion, as well as the entire selection procedure, are shown in the PRISMA flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. TP: telepsychiatry.

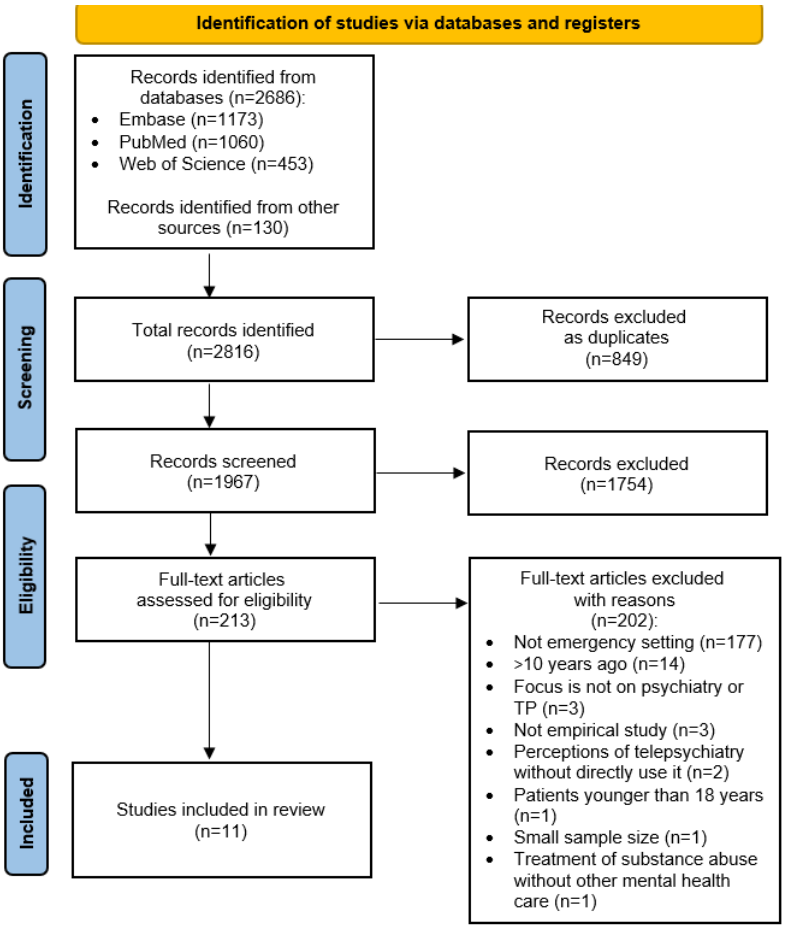

Charting the Data

The articles included in this scoping review were reviewed and results were recorded using a Microsoft Excel (Microsoft Corp) data charting table. The table included general information about the study characteristics (authors, publication year, title, data collection period, country, study purpose, study design, setting, and sample), a description of the TP service and usual care if applicable, outcome measures (administrative and clinical or process outcomes), and main study results.

Collating, Summarizing, and Reporting the Results

Included studies were examined thoroughly to understand similarities and differences. We had the following main categories of study outcomes: (1) *administrative outcomes* (eg, ED LOS, or mental illness spending); (2) *clinical outcomes* (eg, admission disposition or patient and providers satisfaction); and (3) *process outcomes* of TP implementation.

Results

A total of 11 articles were included for data extraction in this scoping review, as can be seen in the PRISMA flowchart (Figure 1).

Characteristics of the Studies

Although the search strategy was from January 2013 to April 2023, the studies that were identified and included collected data between October 2012 and 2023, and were published between 2015 and 2023. Of the 11 articles, 9 were published in the United States [15,19,37-43],1 in Australia [44], and 1 in Israel [45]. Of the 11 studies, 10 [15,19,37-43,45] were quantitative studies that examined administrative or clinical measurements of TP use, and 1 was a qualitative study that examined process outcomes [44]. Of the 10 quantitative studies, 3 were non–randomized controlled studies [40,41,45], 3 cross-sectional studies [15,19,38], 3 cohort studies [37,39,43], and 1 randomized controlled study [42] (Table 2).

In the next section, we will summarize the results of our review, organized by our 2 main research questions, namely (1) to review the existing evidence for both administrative and clinical outcomes of TP in the ED setting and (2) to identify the known barriers and facilitators to implementing TP in this setting.

> *First research question: what is the existing evidence for both administrative and clinical outcomes of TP in the ED setting?*

XSL•FO
RenderX

Table 2. Characteristics of studies included in the scoping review.

Authors (years)	Countries	Study objective	Setting	Study sample	Methods (study paradigm, design, and tools)	Main outcomes measures and special characteristics
Freeman et al (2023) [37]	United States	To investigate ED[a] TP[b] usage in the pre–COVID-19 era	• EDs • Nationwide • Rural and urban	• 69 EDs using TP from surveys conducted in 2017 and 2019	• Quantitative • Cohort study • Questionnaire	• Characteristics of EDs using TP • ED LOS[c] • Patient disposition
Bistre et al (2022) [45]	Israel	To evaluate the reliability and acceptability of TP assessments compared with in-person for involuntary admissions	• Psychiatric ED • Regional • Urban	• 38 patients were evaluated both by TP and in person	• Quantitative • Non-RCT[d] study • Interrater agreement	• TP reliability • Psychiatrists' certainty in TP evaluations, including need for admission • Psychiatrists' and patients' satisfaction of TP
Patel et al (2022) [19]	United States	To assess the impact of a TP service compared with in-person in EDs on admissions, mental illness spending, ED LOS, mental illness outpatient follow-up care, and mortality	• EDs across 22 states • Nationwide • Urban and rural	• 134 EDs using TP • 134 EDs not using TP	• Quantitative • Cross-sectional study • Medical records	• Characteristics of EDs with and with no TP • ED LOS for psychiatric patients • Patient disposition • Mental health spending
Saeed et al (2022) [39]	United States	To study the impact of a TP service in reducing hospitalizations and cost savings	• Hospital EDs • Statewide (North Carolina) • Remote	• 30 EDs using TP	• Quantitative • Cohort study • Medical records	• Involuntary admissions • Mental health spending
Zhong et al (2021) [15]	United States	To examine the impact of a TP service compared with in-person across EDs on visit dispositions	• EDs • Statewide (New York) • Urban and rural	• 18 EDs using TP • 115 EDs not using TP	• Quantitative • Cross-sectional study • Questionnaire and medical records	• Characteristics of EDs with and with no TP • Patient disposition
Brenner et al (2020) [41]	United States	To assess turnaround time with and with no TP and patients' satisfaction of TP	• Three general hospital EDs • Regional • Urban	• 206 TP visits • 186 in-person visits	• Quantitative • Non-RCT study • Medical records and questionnaire	• ED waiting time for psychiatric evaluation • Patient satisfaction
Freeman et al (2020) [38]	United States	To investigate the prevalence of TP use for mental health in general EDs	• EDs • Nationwide • Rural and urban	• 885 EDs using TP • 3525 EDs not using TP	• Quantitative • Cross-sectional study • Questionnaire and medical records	• Characteristics of EDs with and with no TP • ED LOS • Patient disposition
Kothadia et al (2020) [40]	United States	To examine differences in patient disposition for ED psychiatric patients with and with no TP service	• Hospital EDs • Statewide (North Carolina) • Remote	• 30 EDs with active and inactive periods of TP use • 44,857 TP visits • 42,074 in-person visits	• Quantitative • Non-RCT study • Medical records	• Patient disposition
Roberge et al (2020) [42]	United States	To assess whether TP use for mental health in the ED decreases hospitalization	• Six EDs • Regional • Urban	• 323 TP visits • 314 in-person visits • Randomized	• Quantitative • RCT study • Medical records and questionnaires	• Patient disposition • Suicide and self-harm diagnosis

XSL•FO
RenderX

Authors (years)	Countries	Study objective	Setting	Study sample	Methods (study paradigm, design, and tools)	Main outcomes measures and special characteristics
Fairchild et al (2019) [43]	United States	To determine the effects of a TP service on clinical, temporal, and cost outcomes for patients	• Four hospital EDs • Regional • Disadvantaged counties • Rural	• 287 TP visits • 153 in-person visits	• Quantitative • Cohort study • Medical records	• ED LOS and waiting time for psychiatric evaluation • Patient disposition • Suicide and self-harm diagnosis • Mental illness spending
Saurman et al (2015) [44]	Australia	To examine the experience of implementing and using TP for mental health in an ED setting,	• Seven EDs • Regional • Rural and remote	• 12 ED providers	• Qualitative • Interviews	• Process outcomes of TP implementation

[a]ED: emergency department.

[b]TP: telepsychiatry.

[c]LOS: length of stay.

[d]RCT: randomized controlled trial.

Scope and Location of Studies

The 10 studies that examined administrative and clinical outcomes in EDs varied in terms of the research population, study scope, and location. Three of the studies were conducted nationwide in the United States [19,37,38], 3 statewide (2 in North Carolina and 1 in New York) [15,39,40], and 4 regional in 1 or several local EDs [41-43,45]. Of the 10 studies, 3 provided TP services only in rural or remote areas [39,40,43], 3 in urban areas [41,42,45], and 4 in a mix of urban and rural areas [15,19,37,38]. Of the 4 studies with rural and urban EDs, 2 reported that most of the TP use occurred in urban areas [15,19] and 2 in rural areas [37,38]. It is worth noting that none of the included studies focused primarily on a comparison between the use of TP in rural and urban areas.

On-site Psychiatric Services and TP Services

Some papers reported whether the medical centers using TP in fact had on-site psychiatric services some of the time, as opposed to having none at all. Three studies reported that less than 20% of their study sites lacked on-site psychiatric service [15,37,38], while 1 study reported that 65% of their study sites lacked on-site psychiatric service [19], and only 1 study reported that all EDs included in their sample lacked on-site psychiatric services [41]. In addition, TP was reported to be the only form of emergency psychiatric services for more than half of the EDs that participated in 2 nationwide studies in the United States in 2017 and 2019 [37,38].

Effect of TP on Waiting Times in EDs

Two studies examined the effect of TP on waiting time from ED arrival until psychiatric assessment [41,43], while 4 examined the impact on ED LOS from arrival to discharge or admission [19,37,38,43]. Both studies that examined waiting time for psychiatric evaluation found it significantly lower for TP evaluation compared with in-person [41,43]. Of the 4 studies that examined total ED LOS, 2 showed a significant prolonged ED LOS for TP compared with in-person visits [19,43]; the other 2 studies showed similar prevalence in ED LOS for the same EDs in 2017 and in 2019 [37,38].

Effect of TP on Discharge, Admission, and Transfer to Another Facility

Six studies examined the impact of TP on discharge, admission, and transfer to another facility. Three studies found that TP was associated with significantly lower admission rates compared with in-person visits [19,40,43];1 study showed no significant differences in admission rates between TP and in-person evaluation (55% vs 63%; P=.06) [42]; and 1 study found that EDs that used TP had significantly more admissions than EDs without this service (14% vs 12%; P<.001) [15]. One study examined whether TP had an impact on the rate of transfers to another facility. The findings were nuanced; total rates of transfer were lower, but among patients with a LOS of 1-2 days the rate of transfer was higher than with in-person care [40]. Another study that examined patients' transfer to another facility did not find significant differences between TP patients (31%) compared with in-person (24%) [43].

TP Costs

Several studies reported on the costs involved with TP. Saeed and colleagues [39] had examined the cost impact of 19,383 TP visits to 30 EDs in North Carolina. Seventy percent of the visits were encounters for involuntary commitments, and of these, 34% were converted to voluntary hospitalizations sometime before the end of the hospital stay, through a TP encounter. The aggregate cost savings for these conversions of involuntary to voluntary hospitalizations were more than US $20 million [39].

A national study in 22 US EDs found a significant increase in admissions for TP visits compared with in-person visits, which resulted with a significant increase in spending in a 90-day follow-up analysis [19]. Another study examined 3 diagnosis groups and found that the significantly most expensive TP visits were for substance abuse cases (US $4556), followed by suicide and self-harm cases (US $3559), and anxiety, mood, and other health disorders case (US $3355) [43].

XSL•FO
RenderX

Data on Involuntary Commitment Cases

We found limited evidence regarding evaluations for involuntary commitment via TP. As mentioned earlier, Saeed et al [39] examined cost impact of using TP to enable staff to convert involuntary commitments into voluntary hospitalizations. In another study focused on examining the accuracy of TP compared with in-person evaluations, Bistre et al [45] evaluated the reliability of TP assessments compared with in-person assessments for involuntary admissions. An interrater agreement on recommended disposition and on indication for involuntary admission between raters was high [45]. Psychiatrists' perceived certainty rates were high for both TP and in-person evaluations. Participants reported a high level of satisfaction with both TP and in-person evaluations, which were not significantly different [45]. In a separate study, patients reported that they were highly satisfied with TP use in the ED, although it was not used to evaluate for involuntary commitment [41].

Special Behavioral Diagnostic Groups

Some studies have at least implied that TP may not be suitable for some groups of patients that require special attention. A study conducted in the United States found that TP was associated with a reduced wait time until psychiatric assessment, but a longer total ED LOS, compared with usual care. Interestingly, 36% (102/287) of the participants in the TP group were diagnosed as suicide and self-harm, compared with 22% (34/153) in the control group. This study also reported that the time from the end of TP assessment to disposition or discharge was significantly longer for patients with suicide and self-harm than for patients who were diagnosed with anxiety, mood, and other mental health disorders [43]. Those findings are implying that the poor TP performance may be related to the enlarged diagnosis group that requires more observation in the ED and not a result of TP use. In a national study that included patients with different diagnosis, TP was associated with longer ED LOS, more admissions, and greater costs. Yet, a nonsignificant higher rate of suicide and self-harm cases was found in the TP group (4925/35,861, 14%) compared with the in-person group (3734/34,982, 11%), suggesting, again, that the results may be affected by differences in patient characteristics between the TP group and the control group [19].

Having examined the existing evidence for TP's impact on administrative and clinical outcomes in EDs, we will now move to the second research question. In the following section, we will describe the current evidence about the barriers and facilitators to implementing TP in these settings, understanding that successful implementation hinges on navigating these elements.

> *Second research question: what are the known barriers and facilitators to implementing TP in ED setting?*

Our second research question revolves around the implementation process for TP in the ED setting. We found only 1 such study, which was conducted in a rural region of Australia, where the TP service was the main psychiatric service available [44]. The study was organized around the 6 concepts of the theory of access [46]. The following are the key findings, organized by these 6 concepts: (1) *Accessibility*: the staff were able to access mental health specialists for immediate assistance without transferring patients to another facility; (2) *Availability*: the service was valued as an available resource and eased the demands placed upon staff during emergency mental health presentations; (3) *Acceptability*: the service was acceptable to the providers and was a constant and easy resource; (4) *Affordability*: there were no direct costs borne by the providers or the hospitals to use TP to involve a psychiatrist, and it was free for the patients; (5) *Adequacy*: the 24-hour structure of the program was adequate to the clinical needs, particularly after-hours and on weekends; (6) *Awareness*: other than 1 provider, everyone else was aware of the service and had some experience using it [44]. In addition, the service provided a sense of security to the providers. They reported that before the service started, they had felt alone, unsupported, and lacking confidence when dealing with emergency mental health presentations [44]. All these improved due to the arrival of TP.

Discussion

Principal Findings

We performed a scoping review to examine the literature regarding the use of TP for adult emergencies. We summarized the evidence regarding (1) administrative and clinical outcomes for patients; and (2) process outcomes of implementing the TP service. Although TP is a known method for psychiatric evaluation, treatment, and follow-up, we found only 11 studies over the past decade to evaluate its application to the ED setting. Ten of these studies evaluated administrative and clinical outcomes, and only 1 study evaluated the implementation process.

Our review included articles that evaluated TP use in various settings and contexts. TP was acceptable and feasible nationally in the United States [15,19,37,38] and in a study of 7 Australian provinces [44]. We also found that TP was used in urban areas [41,42,45] and rural areas [40,47]. In some cases, TP was used in EDs as the only psychiatric service available [41]. The 1 study we found about the implementation process reported that TP was accepted and mostly appreciated by the ED staff, especially due to the lack of psychiatric expertise in their setting [44]. TP was also used for different sorts of patients, including those with anxiety and mood disorders and those with suicide or self-harm [43,45]. The identified lack of evidence regarding the use of TP in EDs significantly impacts our analysis, underscoring a crucial area where further research is needed to draw comprehensive conclusions. This gap highlights the limitations in our current understanding and emphasizes the necessity for targeted studies to elucidate the efficacy and implementation of TP in ED settings. Despite the limited number of studies we found, this diversity of settings and uses somewhat strengthens the argument that TP is broadly applicable across different ED settings and different patient groups.

Waiting Times

On the issue of ED waiting times, the existing evidence is mixed. Two studies showed that the ED waiting time from patients' arrival until psychiatric evaluation was significantly lower for TP visits than for in-person visits [41,43]. However, 2 studies found that the total ED LOS was significantly longer for TP

XSL•FO
RenderX

visits than for in-person visits [19,43]. Unfortunately, none of the included studies examined waiting times from the psychiatric evaluation until admission or discharge for TP compared with in-person visits, so this remains an unexamined issue. However, because TP is shortening waiting times for initial psychiatric assessment, this may contribute to putting the patient on a better path from the beginning [29]. This is supported by the main finding of this review that TP reduced admission rates [19,40,42,43]. In other words, perhaps the fast psychiatric evaluation by TP resulted in expert psychiatric input to the case sooner, which may partly explain the lower admission rate. Further studies will help clarify these points.

Patients' Characteristics

Another factor that remains relatively unexamined is whether TP is equally applicable to different sorts of patients. Most enrolled studies did not examine TP use through different patient characteristics, such as diagnosis groups or the need for direct observation. Patients who require direct observation usually have more severe presentations and thus a longer LOS [48]. Two studies did focus on the use of TP for patients seen for self-harm and suicide; these studies showed higher ED LOS [19,43]. Therefore, there is a need for further studies of patients with these more severe presentations to ensure that TP is applicable to them as well.

Lack of Findings Regarding Patients Evaluated for Involuntary Commitment

Patients requiring evaluation for involuntary commitment are a distinct group. As presented in the Results section, in 1 study, staff used TP to help evaluate which patients had improved enough to have their involuntary commitments converted into voluntary hospitalizations [39]. However, this does not speak to the initial decision to pursue an involuntary commitment. Given our group's experience, it may be easy to understand why relatively few studies have evaluated the use of TP for patients evaluated for involuntary commitment. Our ongoing study of this issue required special permission from the Israeli Ministry of Health after consultation with the Ministry of Justice and the Union of Psychiatrists [49]. Thus, it is easy to see why there have been relatively few studies regarding the use of TP for this special use case and certainly more are needed.

Patient Transfer

Another issue that was examined was the impact on interhospital transfers. One study found that the use of TP increased the number of transfers [43]. On the other hand, another study showed that among patients with an extended LOS, significantly less TP patients (29%) were transferred to a psychiatric hospital compared with in-person patients [40]. These divergent results may point to a complex and nuanced effect of TP on doctor-patient relationship. Technology in medicine holds the promise to contribute a more personalized style of care [50]. However, remote communication between psychiatrists and patients may affect doctor-patient engagement and lack personal touch compared with in-person encounters [51]. There is a possibility that it is easier for the psychiatrist using TP in emergency cases to decide on transfer rather than admission to the present facility. If there is an association between TP use in the ED and more patient transfers, this could lead to inconvenience for family members, as well as the cost of transport [52]. The impact of TP on the rate of interfacility transfers also requires further study.

Rural and Urban Areas

TP is perceived often as a critical solution for the lack of mental health services in rural and remote areas [31,53]. However, findings from this review indicate that TP is used in urban areas as well [41,42]. Several studies showed that TP was even more common at urban settings [15,19], even when they have existing on-site psychiatric services [15]. The demand to use TP even in urban areas may be driven by the fact that the attending physician is at home for more hours than not, and must drive to the hospital. However, rural areas may face special issues with TP use, including inadequate technology literacy [2], bad internet connectivity [54], or a general lack of resources [2,55]. Despite these challenges, there is a strong incentive to use TP in rural areas, so it may be worth the effort of addressing the challenges.

Strengths and Limitations

This scoping review had several strengths and limitations. A broad range of the main databases were searched, which allowed a comprehensive search. This review provides robust evidence of the included studies, provides a deeper understanding of the current evidence, and provides the needed data to broaden our understanding of TP in emergency settings.

This review also has some limitations. We examined only those studies published in English. All studies that we found were conducted in developed countries, which provides a limited perspective. In addition, the data we found about the use of TP for evaluations regarding possible involuntary commitment were particularly limited. This will be a key area for future research. We also did not find any studies that specifically compared TP use in urban versus rural settings, or that compared its use for specific groups of patients (eg, by diagnosis). Furthermore, we found only 1 interventional study; the others were observational. However, all our included studies had sample sizes larger than 30 participants and a detailed description of the study methodology. In part, we chose to do a scoping review as opposed to a systematic review, because the available literature was so limited.

Conclusions

TP has a strong evidence base for general use and is known to be acceptable, reliable, and effective. However, only a very few studies in the past decade (11 studies) evaluated its use in the ED. While these studies generally supported the idea that TP was feasible and highly acceptable, it is clear that further studies are needed. Further studies are needed for examining TP evaluations for involuntary commitments in the ED setting. In addition, there is a need for studies on the extent and trends of TP usage over time, including in the context of COVID-19. We also need more comprehensive assessments comparing the effectiveness of TP evaluations with in-person assessments and implementation science research to better understand the barriers, facilitators, and opportunities for adopting this practice in EDs. Special attention should be given to rural areas, which

XSL•FO
RenderX

usually have limited access to mental health services and yet may face special challenges in implementing them.

Acknowledgments

The authors would like to thank Tomer Ben-Shushan from the Berman Medical Library, Hebrew University, for conducting the database search for the paper.

Authors' Contributions

LS initiated the study and drafted the manuscript. LS and AJR led the scoping review, design, and analysis. All authors (LS, RE, and AJR) discussed the results, reviewed, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 32 KB - jmir_v26i1e51814_app1.pdf]

Multimedia Appendix 2
Web-based search strategy conducted on April 24, 2023.
[DOCX File , 21 KB - jmir_v26i1e51814_app2.docx]

References

1. Shore JH, Schneck C, Mishkind MC. Telepsychiatry and the coronavirus disease 2019 pandemic-current and future outcomes of the rapid virtualization of psychiatric care. JAMA Psychiatry 2020;77(12):1211-1212. [doi: 10.1001/jamapsychiatry.2020.1643] [Medline: 32391861]
2. Grady B. Promises and limitations of telepsychiatry in rural adult mental health care. World Psychiatry 2012;11(3):199-201 [FREE Full text] [doi: 10.1002/j.2051-5545.2012.tb00132.x] [Medline: 23024682]
3. Kinoshita S, Cortright K, Crawford A, Mizuno Y, Yoshida K, Hilty D, et al. Changes in telepsychiatry regulations during the Covid-19 pandemic: 17 countries and regions' approaches to an evolving healthcare landscape. Psychol Med 2022;52(13):2606-2613 [FREE Full text] [doi: 10.1017/S0033291720004584] [Medline: 33243311]
4. Shore JH. Telepsychiatry: videoconferencing in the delivery of psychiatric care. Am J Psychiatry 2013;170(3):256-262. [doi: 10.1176/appi.ajp.2012.12081064] [Medline: 23450286]
5. Peralta EA, Taveras M. Effectiveness of teleconsultation use in access to mental health services during the coronavirus disease 2019 pandemic in the Dominican Republic. Indian J Psychiatry 2020;62(Suppl 3):S492-S494 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_1047_20] [Medline: 33227069]
6. Stommel W, van der Houwen F. Complaining and the management of face in online counseling. Qual Health Res 2014;24(2):183-193. [doi: 10.1177/1049732313519706] [Medline: 24495987]
7. Shore J. The evolution and history of telepsychiatry and its impact on psychiatric care: current implications for psychiatrists and psychiatric organizations. Int Rev Psychiatry 2015;27(6):469-475. [doi: 10.3109/09540261.2015.1072086] [Medline: 26397182]
8. Glueck DA. Telepsychiatry in private practice. Child Adolesc Psychiatr Clin N Am 2011;20(1):1-11. [doi: 10.1016/j.chc.2010.08.006] [Medline: 21092908]
9. Hilty DM, Rabinowitz T, McCarron RM, Katzelnick DJ, Chang T, Bauer AM, et al. An update on telepsychiatry and how it can leverage collaborative, stepped, and integrated services to primary care. Psychosomatics 2018;59(3):227-250. [doi: 10.1016/j.psym.2017.12.005] [Medline: 29544663]
10. Zachrison KS, Boggs KM, M Hayden E, Espinola JA, Camargo CA. A national survey of telemedicine use by US emergency departments. J Telemed Telecare 2020;26(5):278-284. [doi: 10.1177/1357633X18816112] [Medline: 30558518]
11. Noble D, Haveland S, Islam M. Integrating telepsychiatry based care in rural acute community mental health services a systematic literature review. Asia Pac J Health Manage 2022;17(2):1-13 [FREE Full text] [doi: 10.24083/apjhm.v17i2.1105]
12. Sharma G, Devan K. The effectiveness of telepsychiatry: thematic review. BJPsych Bull 2023;47(2):82-89 [FREE Full text] [doi: 10.1192/bjb.2021.115] [Medline: 34915955]
13. Hughes M, Gorman J, Ren Y, Khalid S, Clayton C. Increasing access to rural mental health care using hybrid care that includes telepsychiatry. J Rural Mental Health 2019;43(1):30-37 [FREE Full text] [doi: 10.1037/rmh0000110]
14. Puspitasari AJ, Heredia D, Coombes B, Geske J, Gentry M, Moore W, et al. Feasibility and initial outcomes of a group-based teletherapy psychiatric day program for adults with serious mental illness: open, nonrandomized trial in the context of Covid-19. JMIR Ment Health 2021;8(3):e25542 [FREE Full text] [doi: 10.2196/25542] [Medline: 33651706]

XSL•FO
RenderX

15. Zhong C, Freeman R, Boggs K, Zachrison K, Gao J, Espinola J, et al. Receipt of telepsychiatry and emergency department visit outcomes in New York State. Psychiatr Q 2021;92(3):1109-1127 [FREE Full text] [doi: 10.1007/s11126-021-09886-y] [Medline: 33587257]
16. Seidel RW, Kilgus M. Agreement between telepsychiatry assessment and face-to-face assessment for emergency department psychiatry patients. J Telemed Telecare 2014;20(2):59-62. [doi: 10.1177/1357633X13519902] [Medline: 24414395]
17. Hubley S, Lynch S, Schneck C, Thomas M, Shore J. Review of key telepsychiatry outcomes. World J Psychiatry 2016;6(2):269-282 [FREE Full text] [doi: 10.5498/wjp.v6.i2.269] [Medline: 27354970]
18. Salmoiraghi A, Hussain S. A systematic review of the use of telepsychiatry in acute settings. J Psychiatr Pract 2015;21(5):389-393. [doi: 10.1097/PRA.0000000000000103] [Medline: 26348806]
19. Patel SY, Huskamp H, Barnett M, Zubizarreta J, Zachrison K, Busch A, et al. Association between telepsychiatry capability and treatment of patients with mental illness in the emergency department. Psychiatr Serv 2022;73(4):403-410 [FREE Full text] [doi: 10.1176/appi.ps.202100145] [Medline: 34407629]
20. Hulsbosch AM, Nugter M, Tamis P, Kroon H. Videoconferencing in a mental health service in The Netherlands: a randomized controlled trial on patient satisfaction and clinical outcomes for outpatients with severe mental illness. J Telemed Telecare 2017;23(5):513-520. [doi: 10.1177/1357633X16650096] [Medline: 27236703]
21. Hilty DM, Serhal E, Crawford A. A telehealth and telepsychiatry economic cost analysis framework: scoping review. Telemed J E Health 2023;29(1):23-37. [doi: 10.1089/tmj.2022.0016] [Medline: 35639444]
22. Narasimhan M, Druss BG, Hockenberry JM, Royer J, Weiss P, Glick G, et al. Impact of a telepsychiatry program at emergency departments statewide on the quality, utilization, and costs of mental health services. Psychiatr Serv 2015;66(11):1167-1172 [FREE Full text] [doi: 10.1176/appi.ps.201400122] [Medline: 26129992]
23. Naslund JA, Mitchell LM, Joshi U, Nagda D, Lu C. Economic evaluation and costs of telepsychiatry programmes: a systematic review. J Telemed Telecare 2022;28(5):311-330. [doi: 10.1177/1357633X20938919] [Medline: 32746762]
24. Cowan KE, McKean AJ, Gentry MT, Hilty DM. Barriers to use of telepsychiatry: clinicians as gatekeepers. Mayo Clin Proc 2019;94(12):2510-2523. [doi: 10.1016/j.mayocp.2019.04.018] [Medline: 31806104]
25. Jiménez-Rodríguez D, Santillán García A, Montoro Robles J, Rodríguez Salvador MDM, Muñoz Ronda FJ, Arrogante O. Increase in video consultations during the Covid-19 pandemic: healthcare professionals' perceptions about their implementation and adequate management. Int J Environ Res Public Health 2020;17(14):5112 [FREE Full text] [doi: 10.3390/ijerph17145112] [Medline: 32679848]
26. Ruiz-Cosignani D, Chen Y, Cheung G, Lawrence M, Lyndon M, Ma'u E, et al. Adaptation models, barriers, and facilitators for cultural safety in telepsychiatry: a systematic scoping review. J Telemed Telecare 2024;30(3):466-474 [FREE Full text] [doi: 10.1177/1357633X211069664] [Medline: 34989643]
27. Galvin E, Desselle S, Gavin B, Quigley E, Flear M, Kilbride K, et al. Stakeholder perspectives and experiences of the implementation of remote mental health consultations during the Covid-19 pandemic: a qualitative study. BMC Health Serv Res 2023;23(1):623 [FREE Full text] [doi: 10.1186/s12913-023-09529-x] [Medline: 37312119]
28. Zachrison KS, Boggs K, Hayden E, Espinola J, Camargo C. Understanding barriers to telemedicine implementation in rural emergency departments. Ann Emerg Med 2020;75(3):392-399. [doi: 10.1016/j.annemergmed.2019.06.026] [Medline: 31474481]
29. Rossi J, Swan M, Isaacs E. The violent or agitated patient. Emerg Med Clin North Am 2010;28(1):235-256. [doi: 10.1016/j.emc.2009.10.006] [Medline: 19945609]
30. Faessler L, Kutz A, Haubitz S, Mueller B, Perrig-Chiello P, Schuetz P. Psychological distress in medical patients 30 days following an emergency department admission: results from a prospective, observational study. BMC Emerg Med 2016;16(1):33 [FREE Full text] [doi: 10.1186/s12873-016-0097-y] [Medline: 27557531]
31. Tracy B, Leah B, Christine M, Farhad S. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. Int J Acad Med 2017;3(1):5-9 [FREE Full text]
32. Reinhardt I, Gouzoulis-Mayfrank E, Zielasek J. Use of telepsychiatry in emergency and crisis intervention: current evidence. Curr Psychiatry Rep 2019;21(8):63. [doi: 10.1007/s11920-019-1054-8] [Medline: 31263972]
33. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]
34. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
35. The Joanna Briggs Institute. The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews. Adelaide, Australia: The Joanna Briggs Institute; 2015. URL: https://reben.com.br/revista/wp-content/uploads/2020/10/Scoping.pdf [accessed 2024-06-28]
36. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
37. Freeman RE, Zhong C, Bahar P, Boggs KM, Faridi MK, Sullivan AF, et al. U.S. emergency department telepsychiatry use in 2019. Telemed J E Health 2023;29(3):366-375 [FREE Full text] [doi: 10.1089/tmj.2022.0191] [Medline: 35867053]

XSL•FO
RenderX

38. Freeman RE, Boggs KM, Zachrison KS, Freid RD, Sullivan AF, Espinola JA, et al. National study of telepsychiatry use in U.S. emergency departments. Psychiatr Serv 2020;71(6):540-546. [doi: 10.1176/appi.ps.201900237] [Medline: 32019430]
39. Saeed SA, Jones K, Muppavarapu K. The impact of NC statewide telepsychiatry program (NC-STeP) on cost savings by reducing unnecessary psychiatric hospitalizations during a 6½ year period. Psychiatr Q 2022;93(2):527-536. [doi: 10.1007/s11126-021-09967-y] [Medline: 34792742]
40. Kothadia RJ, Jones K, Saeed SA, Torres M. The impact of the North Carolina Statewide Telepsychiatry Program (NC-STeP) on patients' dispositions from emergency departments. Psychiatr Serv 2020;71(12):1239-1244. [doi: 10.1176/appi.ps.201900431] [Medline: 33019858]
41. Brenner R, Madhusoodanan S, Logiudice J, Castell G, MacKenzie T, O'Shaughnessy PM. A comparison study of the turnaround time for telepsychiatry versus face-to-face consultations in general hospital nonpsychiatric emergency rooms. Ann Clin Psychiatry 2020;32(1):12-16. [Medline: 31990965]
42. Roberge J, McWilliams A, Zhao J, Anderson WE, Hetherington T, Zazzaro C, et al. Effect of a virtual patient navigation program on behavioral health admissions in the emergency department: a randomized clinical trial. JAMA Netw Open 2020;3(1):e1919954 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.19954] [Medline: 31995214]
43. Fairchild RM, Ferng-Kuo SF, Laws S, Rahmouni H, Hardesty D. Telehealth decreases rural emergency department wait times for behavioral health patients in a group of critical access hospitals. Telemed J E Health 2019;25(12):1154-1164. [doi: 10.1089/tmj.2018.0227] [Medline: 30735100]
44. Saurman E, Kirby SE, Lyle D. No longer 'flying blind': how access has changed emergency mental health care in rural and remote emergency departments, a qualitative study. BMC Health Serv Res 2015;15:156 [FREE Full text] [doi: 10.1186/s12913-015-0839-7] [Medline: 25889260]
45. Bistre M, Juven-Wetzler A, Argo D, Barash I, Katz G, Teplitz R, et al. Comparable reliability and acceptability of telepsychiatry and face-to-face psychiatric assessments in the emergency room setting. Int J Psychiatry Clin Pract 2022;26(3):228-233. [doi: 10.1080/13651501.2021.1979586] [Medline: 34565277]
46. Penchansky R, Thomas J. The concept of access: definition and relationship to consumer satisfaction. Med Care 1981;19(2):127-140. [doi: 10.1097/00005650-198102000-00001] [Medline: 7206846]
47. Fairchild R, Ferng-Kuo SF, Rahmouni H, Hardesty D. An observational study of telemental care delivery and the context for involuntary commitment for mental health patients in a group of rural emergency departments. Telemed Rep 2020;1(1):22-35 [FREE Full text] [doi: 10.1089/tmr.2020.0005] [Medline: 33283206]
48. Barrett ML, Fingar KR, Owens PL, Stocks C, Sheng M. HCUP methods series identifying observation services in the Healthcare Cost and Utilization Project (HCUP) state databases. Report # 2015-05. 2015. URL: https://hcup-us.ahrq.gov/reports/methods/2015-05_public.pdf [accessed 2024-06-27]
49. Shalev L, Bistre M, Lubin G, Avirame K, Raskin S, Linkovski O, et al. Enabling expedited disposition of emergencies using telepsychiatry in israel: protocol for a hybrid implementation study. JMIR Res Protoc 2023;12:e49405 [FREE Full text] [doi: 10.2196/49405] [Medline: 37847548]
50. Institute of Medicine. Crossing the Quality Chasm. Washington, DC: National Academies Press; 2001.
51. Yellowlees P, Nafiz N. The psychiatrist-patient relationship of the future: anytime, anywhere? Harv Rev Psychiatry 2010;18(2):96-102. [doi: 10.3109/10673221003683952] [Medline: 20235774]
52. D'Souza R. Telemedicine for intensive support of psychiatric inpatients admitted to local hospitals. J Telemed Telecare 2000;6 Suppl 1:S26-S28. [doi: 10.1258/1357633001934834] [Medline: 10793963]
53. Saurman E, Lyle D, Kirby S, Roberts R. Use of a mental health emergency care-rural access programme in emergency departments. J Telemed Telecare 2014;20(6):324-329. [doi: 10.1177/1357633X14544422] [Medline: 25052388]
54. Vakkalanka JP, Nataliansyah MM, Merchant KAS, Mack LJ, Parsons S, Mohr NM, et al. Telepsychiatry services across an emergency department network: a mixed methods study of the implementation process. Am J Emerg Med 2022;59:79-84. [doi: 10.1016/j.ajem.2022.06.053] [Medline: 35810736]
55. Barrett A, Terry D, Lê Q, Hoang H. Factors influencing community nursing roles and health service provision in rural areas: a review of literature. Contemp Nurse 2016;52(1):119-135. [doi: 10.1080/10376178.2016.1198234] [Medline: 27264878]

Abbreviations

ED: emergency department
LOS: length of stay
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
TP: telepsychiatry

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 13.08.23; peer-reviewed by P Yellowlees, E Eboreime; comments to author 22.02.24; revised version received 25.02.24; accepted 29.04.24; published 15.07.24.

Please cite as:
Shalev L, Eitan R, Rose AJ
The Use of Telepsychiatry Services in Emergency Settings: Scoping Review
J Med Internet Res 2024;26:e51814
URL: https://www.jmir.org/2024/1/e51814
doi:10.2196/51814
PMID:

©Ligat Shalev, Renana Eitan, Adam J Rose. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of eHealth Interventions in Improving Medication Adherence Among Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis

Yiqun Miao[1*], PhD; Yuan Luo[1*], PhD; Yuhan Zhao[1], PhD; Mingxuan Liu[1], PhD; Huiying Wang[1], PhD; Ying Wu[1], PhD

School of Nursing, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ying Wu, PhD
School of Nursing
Capital Medical University
10 You-an-men Wai Xi-tou-tiao
Feng-tai District
Beijing, 100069
China
Phone: 86 10 8391 17
Email: helenywu@vip.163.com

Abstract

Background: Nonadherence to medication among patients with cardiovascular diseases undermines the desired therapeutic outcomes. eHealth interventions emerge as promising strategies to effectively tackle this issue.

Objective: The aim of this study was to conduct a network meta-analysis (NMA) to compare and rank the efficacy of various eHealth interventions in improving medication adherence among patients with cardiovascular diseases (CVDs).

Methods: A systematic search strategy was conducted in PubMed, Embase, Web of Science, Cochrane, China National Knowledge Infrastructure Library (CNKI), China Science and Technology Journal Database (Weipu), and WanFang databases to search for randomized controlled trials (RCTs) published from their inception on January 15, 2024. We carried out a frequentist NMA to compare the efficacy of various eHealth interventions. The quality of the literature was assessed using the risk of bias tool from the Cochrane Handbook (version 2.0), and extracted data were analyzed using Stata16.0 (StataCorp LLC) and RevMan5.4 software (Cochrane Collaboration). The certainty of evidence was evaluated using the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) approach.

Results: A total of 21 RCTs involving 3904 patients were enrolled. The NMA revealed that combined interventions (standardized mean difference [SMD] 0.89, 95% CI 0.22-1.57), telephone support (SMD 0.68, 95% CI 0.02-1.33), telemonitoring interventions (SMD 0.70, 95% CI 0.02-1.39), and mobile phone app interventions (SMD 0.65, 95% CI 0.01-1.30) were statistically superior to usual care. However, SMS compared to usual care showed no statistical difference. Notably, the combined intervention, with a surface under the cumulative ranking curve of 79.3%, appeared to be the most effective option for patients with CVDs. Regarding systolic blood pressure and diastolic blood pressure outcomes, the combined intervention also had the highest probability of being the best intervention.

Conclusions: The research indicates that the combined intervention (SMS text messaging and telephone support) has the greatest likelihood of being the most effective eHealth intervention to improve medication adherence in patients with CVDs, followed by telemonitoring, telephone support, and app interventions. The results of these network meta-analyses can provide crucial evidence-based support for health care providers to enhance patients' medication adherence. Given the differences in the design and implementation of eHealth interventions, further large-scale, well-designed multicenter trials are needed.

Trial Registration: INPLASY 2023120063; https://inplasy.com/inplasy-2023-12-0063/

(J Med Internet Res 2024;26:e58013) doi:10.2196/58013

XSL•FO
RenderX

KEYWORDS

eHealth; medication adherence; heart; cardiovascular; adherence; cardiology; digital health; cardiovascular diseases; network meta-analysis; synthesis; syntheses; review methods; review methodology; search; searches; searching; systematic; meta-analysis; meta-analyses; drug; drugs; pharmacy; pharmacies; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications

Introduction

Cardiovascular diseases (CVDs) encompass a range of disorders of the heart and blood vessels, including coronary heart disease, heart failure, aortic disease, cerebrovascular disease, and peripheral artery disease [1]. CVDs account for nearly one-third of global deaths, with approximately 17.9 million people dying annually [2]. Despite advancements in therapeutic modalities like angioplasty, pharmacological treatments remain the cornerstone for improving the prognosis and reducing the risk of major adverse cardiovascular events and mortality across all patient groups, regardless of whether they have undergone invasive interventions [3]. However, medication adherence among these patients is suboptimal, with about one-third not following prescribed regimens [4]. Poor medication adherence is concerning, as among self-reported nonadherent patients, the risk of mortality increases by 3.8 times [5]. In the United States, the preventable health care costs attributable to nonadherence exceed US $100 billion annually [6].

The challenges to medication adherence in patients with CVDs are multifaceted, primarily driven by forgetfulness, fear of side effects, communication gaps between patients and health care providers, socioeconomic factors, and lack of motivation [7]. In response to these challenges, eHealth interventions have gained attention as effective tools for enhancing preventive care across various settings [8]. eHealth encompasses diverse technological applications in health care, including digital information resources, remote monitoring, teleconsultations, and mobile-supported care [9]. These interventions facilitate real-time adjustments in health behavior to accommodate changes in individual health needs, interventional goals, and available resources [10]. Furthermore, eHealth technologies offer engaging, user-friendly, evidence-based solutions that can potentially reduce health care costs [11]. Previous studies have shown that eHealth technologies such as SMS text messaging, voice-enabled response systems, and phone calls have successfully provided feedback on medication adherence, significantly improving adherence rates among patients with chronic conditions [12,13]. Additionally, these interventions have proven effective in promoting lifestyle and behavioral changes necessary to manage blood pressure and cholesterol levels [14,15].

However, the concept of eHealth interventions is notably broad, encompassing a range of approaches that use various electronic devices. These interventions can vary greatly, from simple measures such as phone calls or emails to more comprehensive and integrated strategies that may include elements of automation and artificial intelligence [16]. Due to the diversity in the scope, components, and sophistication of eHealth interventions included in different studies, their effectiveness in enhancing medication adherence can also vary significantly [17]. Traditional meta-analyses or systematic reviews typically require that the interventions in the studies being compared are relatively consistent, which poses a challenge when attempting to compare the effects of different eHealth interventions [18,19]. Therefore, to address these variations and provide a more detailed understanding of how different eHealth interventions impact medication adherence among patients with CVDs, we conducted a network meta-analysis (NMA) on the existing evidence to rank the effects of various eHealth interventions on medication adherence among patients with CVDs, so as to provide evidence for clinical practice and health policy.

Methods

Protocol and Registration

The NMA adhered to the PRISMA-NMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses) guidelines (Table S1 in Multimedia Appendix 1) [20]. It has been registered with the INPLASY, with the registration number INPLASY2023120063.

Search Strategies

We screened several databases with no limitations on language or publication date, from inception to January 15, 2024. These databases included PubMed, Cochrane, Web of Science, Embase, China National Knowledge Infrastructure Library, Chinese Biomedical Literature Database, Chinese Scientific Journal database, and WanFang databases. To develop this comprehensive search strategy, we consulted with a librarian experienced in medical research and a cardiovascular expert to ensure a thorough search. The following Medical Subject Headings (MeSH) and keywords incorporating Boolean operators were applied: “cardiovascular disease,” “telemedicine,” “eHealth,” “smartphone,” “mobile applications,” “text messaging,” “medication adherence,” and “randomized controlled trials,” among others. Furthermore, manual searches in major international conference proceedings, systematic reviews, meta-analyses, and gray literature were carried out by 2 authors to systematically identify potential studies to prevent the omission of relevant studies, particularly those with only abstracts available. The search strategies are detailed in Textbox S1 in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Inclusion criteria were based on the PICOS (Participants, Interventions, Comparison, Outcomes, and Study Design) criterion. Studies that satisfied the following criteria were included:

- Participants: patients were diagnosed with CVDs, including hypertension, ischemic heart disease, myocardial infarction, acute coronary syndrome, heart failure, and peripheral arterial disease.
- Interventions: trials had to include at least 1 type of eHealth intervention, such as telephone support, mobile phone apps,

XSL•FO
RenderX

telemonitoring, SMS, or combined intervention (more than 2 types of eHealth interventions).

- Comparison: the comparative arms could include various interventions or usual care.
- Outcomes: the primary outcome was medication adherence, defined as the degree to which an individual's medication-taking behavior aligns with the recommendations provided by health care providers [21]. Studies using the validated scale for evaluation were included. Secondary outcomes included systolic and diastolic blood pressure.
- Study design: The study design was restricted to RCTs.

We excluded studies that were single-arm studies, conference abstracts, letters to the editor, or study protocols that lacked adequate data for extraction.

After reviewing the interventions and discussion, we classified the interventions into the following categories:

- App intervention: Defined as an intervention delivered by an app on a mobile phone operating on iOS or Android OS.
- SMS text messaging intervention: Defined as any intervention using SMS text messaging to deliver content to the patient, with the capability for back-and-forth communication between fixed telephone and mobile telephone equipment.
- Telemonitoring intervention: Defined as the use of remote devices for long-distance monitoring of patients, with collected data transmitted to health care professionals. Guidance and interventions are provided through remote communication methods upon the detection of issues or the need for adjustments to the treatment plan.
- Telephone support intervention: Defined as coaching, medication reminders, or education delivered via telephone. This includes voice interaction mode and telephone follow-up.
- Combined intervention: Defined as incorporating more than 2 types of eHealth interventions.
- Usual care: Defined as the standard treatment and nursing provided by health care professionals, typically including the prescription of cardiovascular medications and lifestyle advice. This does not include the information reminders offered by eHealth technologies.

Data Extraction

All initial search results were independently reviewed by 2 investigators. Duplicates and irrelevant articles were removed using EndnoteX9 (Thompson ISI Research Soft), based on titles and abstracts. Subsequently, full texts underwent comprehensive evaluation to ensure compliance with all inclusion and exclusion criteria. Any discrepancies were resolved through team discussions, with a third reviewer available for arbitration if necessary.

The data extracted included general information (authors, year of publication), participants' characteristics (gender proportion, age range), study characteristics (duration of intervention, recruiting area, number of patients, type of intervention), and methods of measuring medication adherence. We reached out to the authors for any missing data. Our primary focus was on medication adherence resulting from adopting eHealth interventions, using various validated scales (Table S2 in Multimedia Appendix 1 [22-26]). We extracted mean differences and SDs between baseline and the last observation in order to calculate the change score for medication adherence.

Risk of Bias and Certainty of Evidence Assessment

The methodological quality of each study using the Cochrane Risk of Bias tool 2.0 [27], categorizing studies into low, high, or unclear risk of bias, was independently assessed by 2 reviewers. The assessment covered several aspects, including random sequence generation, allocation concealment, blinding of participants and personnel, incomplete outcome data, selective reporting, and other biases. In evaluating the outcomes of the NMA, we used the GRADE framework to assess the evidence level [28]. This framework classifies the risk of bias, inconsistency, indirectness, imprecision, and publication bias for each paired comparison into 4 categories: "high," "moderate," "low," or "very low." Discrepancies were resolved by a third-party investigator.

Statistical Analysis

The NMA was performed using Stata (version 16.0; StataCorp LLC). A network plot visually represented the primary evidence for each eHealth intervention. Each node represented an eHealth intervention, with its size dependent on the number of patients directly involved. These nodes are connected by lines of varying thickness, which indicate whether there is a direct relationship between the interventions. Additionally, the thickness of the lines is weighted based on the direct evidence between them. We used standardized mean differences (SMDs) and 95% CIs to pool data from various adherence scales to ensure adequate comparability. The mean differences and SD were either directly extracted from the published data or computed using the available information [29]. Studies were evaluated for heterogeneity using the I^2 statistic, with low, medium, and high degrees of heterogeneity indicated by values of approximately 25%, 50%, and 75%, respectively [30]. Subgroup analyses were conducted to explore the effect of gender (males vs females), duration of intervention (≥3 months vs <3 months), and mode of delivery (theory-driven intervention vs interventions not based on theory).

It is crucial to consider network transitivity in NMA, as it serves as a foundational assumption that significantly impacts our subsequent analysis [31]. To ensure the validity of indirect inferences among the various treatment comparisons, we assessed the transitivity assumption by examining clinical and methodological characteristics, including patient profiles and study designs, across all studies included in our analysis. Surface under the cumulative ranking (SUCRA) probabilities were calculated to rank each intervention's efficacy compared to an ideal intervention that always performs the best without uncertainty. SUCRA scores range from 0 to 1, with a higher score indicating a higher likelihood that an intervention is the most effective [32]. We conducted node-splitting analysis to assess the inconsistency between direct and indirect evidence estimates for each intervention comparison [33]; in cases without significant inconsistency (P>.05), a consistency model was

XSL•FO
RenderX

applied. A comparison-adjusted funnel plot was created to identify potential sources of bias.

Results

Participants and Study Characteristics

The literature screening flowchart for our NMA is depicted in Figure 1. Initially, 3066 publications were identified through database searches, and 4 additional studies were found via manual searches. Of the 85 full-text articles evaluated for eligibility, 21 studies [34-54] were ultimately included. These studies involved a total of 3804 participants, divided into intervention (n=1937) and control (n=1867) groups. The intervention types were as follows: 4 studies used app interventions [38-41], 4 used telephone support [42-44], 4 used telemonitoring interventions [34-37], 6 used SMS text messaging interventions [45-50], and 4 used combined interventions [51-54]. Participants ranged in age from 35 to 80 years, with 51.3% being female. The duration of the interventions varied from 1 to 8 months, and the studies were conducted across various regions, including 7 studies in North America, 9 in Asia, and 5 in Europe and other areas. The characteristics of these studies are detailed in Table 1.

The risk of bias assessment is depicted in Figure S1 in Multimedia Appendix 1. Regarding random sequence generation, 20 RCTs exhibited low bias, and 19 RCTs showed low bias in allocation concealment. Performance bias was a concern in 1 RCT assessed as high risk and in 7 RCTs assessed with unclear risk levels. Detection bias was observed in 5 RCTs with high overall risks and in 6 with unclear risks. The GRADE assessment for outcomes is shown in Tables S3-S5 in Multimedia Appendix 1. Certainty-in-effect estimates for each comparison were displayed, and the reasons for downgrading were also shown.

Figure 1. Flowchart of study selection process.

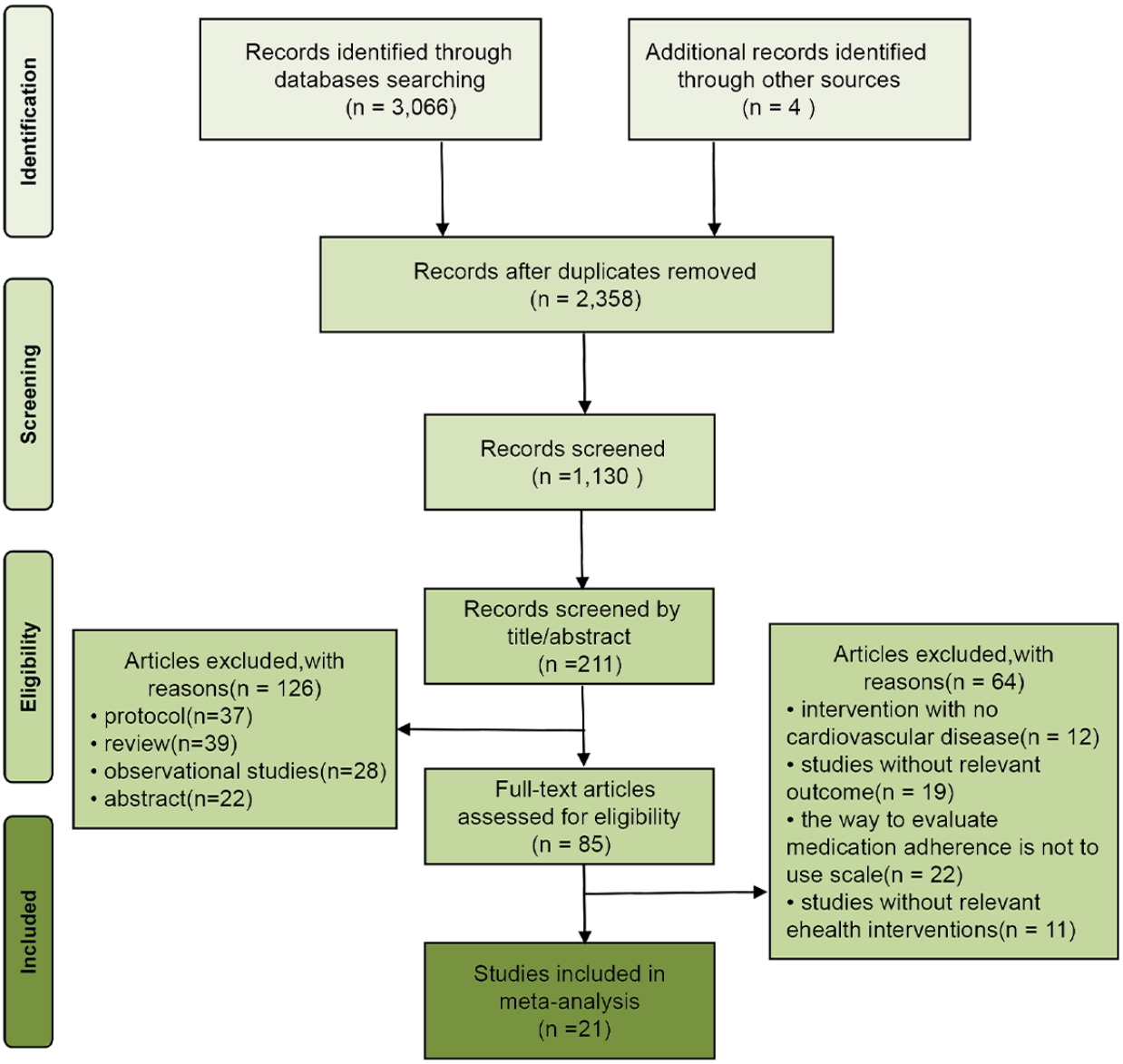

XSL•FO
RenderX

Table 1. Characteristics of included studies (N=3804).

Reference	Participants (n), intervention/control	Age (years), mean (SD)		Proportion of female participants (%)	Recruiting area	Study population	Intervention or control	Duration of intervention	Scale of medication adherence
		Intervention	Control						
He et al 2022 [38]	41/41	NR[a]	NR	51.2	China	Cardiovascular disease	App vs usual care	8 weeks	SEAMS[b]
Zha et al 2020 [34]	12/13	48.90 (8.00)	55.50 (5.20)	88.0	USA	Hypertension	Telemonitoring vs usual care	6 months	MASES[c]
Kim et al 2016 [35]	52/43	57.50 (8.60)	57.70 (8.70)	68.4	USA	Hypertension, cardiac arrhythmia	Telemonitoring vs usual care	6 months	MMAS-8[d]
Yanicelli et al 2020 [39]	15/15	NR	NR	20.0	Argentina	Heart failure	App vs usual care	90 days	MMAS-8
Sarfo et al 2018 [36]	30/30	54.30 (11.90)	55.90 (13.70)	35.0	Ghana	Stroke survivors	Telemonitoring vs SMS	3 months	MMAS-8
Chandler et al 2019 [37]	26/28	44.40 (7.20)	46.80 (8.10)	64.8	USA	Essential hypertension	Telemonitoring vs usual care	2 months	MMAS-8
Morawski et al 2018 [40]	209/202	51.70 (10.50)	52.40 (10.10)	60.1	USA	Hypertension	App vs usual care	12 weeks	MMAS-8
Santo et al, 2019 [41]	107/56	58.40 (9.04)	56.80 (8.64)	12.5	Australia	Coronary heart disease	App vs usual care	3 months	MMAS-8
Wan et al 2016 [42]	40/40	59.07 (12.36)	60.24 (12.57)	28.8	China	Ischemic stroke patients	Telephone support vs usual care	6 months	MMAS-4[e]
Migneault et al 2012 [43]	169/168	59.17 (12.63)	58.44 (12.59)	70.3	USA	Hypertension	Telephone support vs usual care	8 months	MMAS-7[f]
Najafi et al 2016 [44]	50/50	58.92 (9.64)	57.70 (10.64)	54.0	Iran	Myocardial infarction	Telephone support vs usual care	12 weeks	MMAS-8
Buis et al 2017 [45]	65/58	46.30 (8.00)	52.20 (7.60)	55.4	USA	Hypertension	SMS vs usual care	1 month	MMAS-8
Bermon et al 2021 [46]	462/468	64.00 (9.70)	63.10 (10.00)	21.6	Colombia	Cardiovascular disease	SMS vs usual care	52 weeks	MARS-5[g]
Bhandari et al 2022, [47]	100/100	49.20 (9.78)	51.70 (9.21)	44.5	Nepal	Hypertension	SMS vs usual care	12 weeks	MASES
Park et al 2014 [48]	28/28	58.20 (10.60)	61.60 (9.10)	30.4	USA	Coronary heart disease	SMS vs usual care	30 days	SEAMS
Kamal et al 2015 [49]	100/100	56.07 (1.50)	57.62 (1.30)	32.5	Pakistan	Stroke	SMS vs usual care	2 months	MMAS-8
Zhai et al 2020 [50]	192/192	68.50 (7.90)	69.40 (9.70)	69.0	China	Hypertension	SMS vs usual care	3 months	MMAS-8
Wan et al 2018 [51]	80/78	35.00-86.00[h]	35.00-86.00[h]	34.8	China	Hypertensive ischemic stroke	Combined intervention[i] vs telephone support	3 months	HPLP II[j]
Kes and Polat 2022 [52]	39/38	54.90 (6.60)	52.20 (6.20)	53.2	Turkey	Hypertension	Combined intervention[k] vs usual care	3 months	MASES

XSL•FO
RenderX

Reference	Participants (n), intervention/control	Age (years), mean (SD)		Proportion of female participants (%)	Recruiting area	Study population	Intervention or control	Duration of intervention	Scale of medication adherence
		Intervention	Control						
Kamal et al 2018 [53]	99/98	59.10 (11.60)	57.70 (11.10)	22.8	Pakistan	Stroke and heart attack survivors	Combined intervention[l] vs usual care	3 months	MMAS-8
Schoenthaler et al 2020 [54]	21/21	59.70 (10.70)	57.60 (11.10)	45.2	USA	Uncontrolled hypertension	Combined intervention[m] vs usual care	3 months	MMAS-8

[a]NR: not reported.
[b]SEAMS: Self-efficacy for appropriate medication use scale.
[c]MASES: Medication adherence self-efficacy scale.
[d]MMAS-8: Morisky Medication Adherence Scale–8.
[e]MMAS-4: Morisky Medication Adherence Scale–4.
[f]MMAS-7: Morisky Medication Adherence Scale–7.
[g]MARS-5: Medication Adherence Report Scale–5.
[h]Range.
[i]Combined intervention: SMS and telephone support.
[j]HPLP II: Health Promoting Lifestyle Profile II.
[k]Combined intervention: SMS and telephone support.
[l]Combined intervention: SMS and telephone support.
[m]Combined intervention: Tablet-based intervention, text message interactions, and video education.

Primary Outcome

The pairwise meta-analyses demonstrated that the efficacy of the combined intervention was superior to that of telephone support, app interventions, and telemonitoring interventions (Figure S2 in Multimedia Appendix 1). I^2 values reflected that our preliminary meta-analysis revealed a high degree of heterogeneity among all the included studies (I^2=78.4%, P<.001).

Figure 2 illustrates the network evidence for medication adherence outcomes. The analysis included 7 direct and 8 indirect comparisons among the eHealth interventions. The size of the nodes indicated that usual care involved the most participants (n=1759), while the combined intervention had the fewest (n=120). Figure 3 depicts the relative effect sizes of medication adherence post intervention. The following 4 eHealth interventions were statistically superior to usual care in enhancing medication adherence: the combined intervention (SMD 0.89, 95% CI 0.22-1.57), telemonitoring intervention (SMD 0.70, 95% CI 0.02-1.39), telephone support intervention (SMD 0.68, 95% CI 0.02-1.33), and app intervention (SMD 0.65, 95% CI 0.01-1.30). However, the SMS text messaging intervention did not show a significant difference from usual care (SMD 0.28, 95% CI –0.20 to 0.77; low to very low certainty). The hierarchy of each eHealth intervention was ranked according to SUCRA values. The combined intervention (SUCRA 79.3%) possessed the greatest likelihood of being the best intervention for medication adherence, along with the suboptimal telemonitoring intervention (SUCRA 64.6%), followed by telephone support (SUCRA 62.5%), app (SUCRA 60.7%), and SMS text messaging interventions (SUCRA 28.9%); usual care had the lowest SUCRA (3.9%), as shown in Figure S3 in Multimedia Appendix 1.

The node-splitting method revealed no local inconsistency between direct and indirect evidence among the eHealth interventions (P≥.05). The loop inconsistency plot formed a closed loop among different interventions, as shown in Figure S4 in Multimedia Appendix 1. Figure 4 displayed an inverted funnel plot with several scattered points asymmetrically distributed, suggesting a certain level of publication bias.

XSL•FO
RenderX

Figure 2. The network plot regarding medication adherence. App, mobile phone applications; SMS, short messaging service.

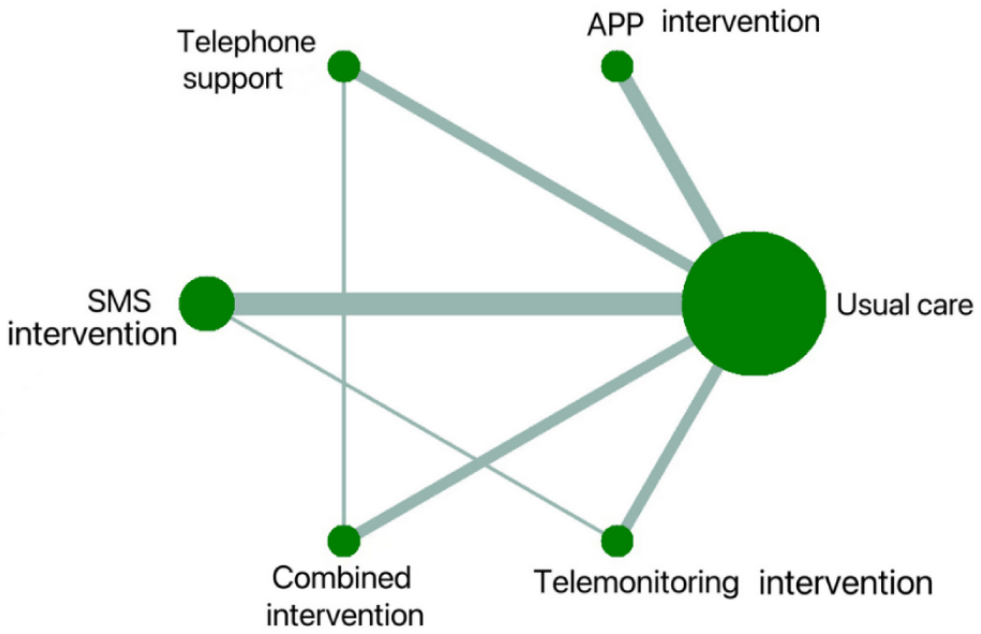

Figure 3. Relative effect sizes of medication adherence at postintervention according to network meta-analysis. App, mobile phone applications; SMS, short messaging service.

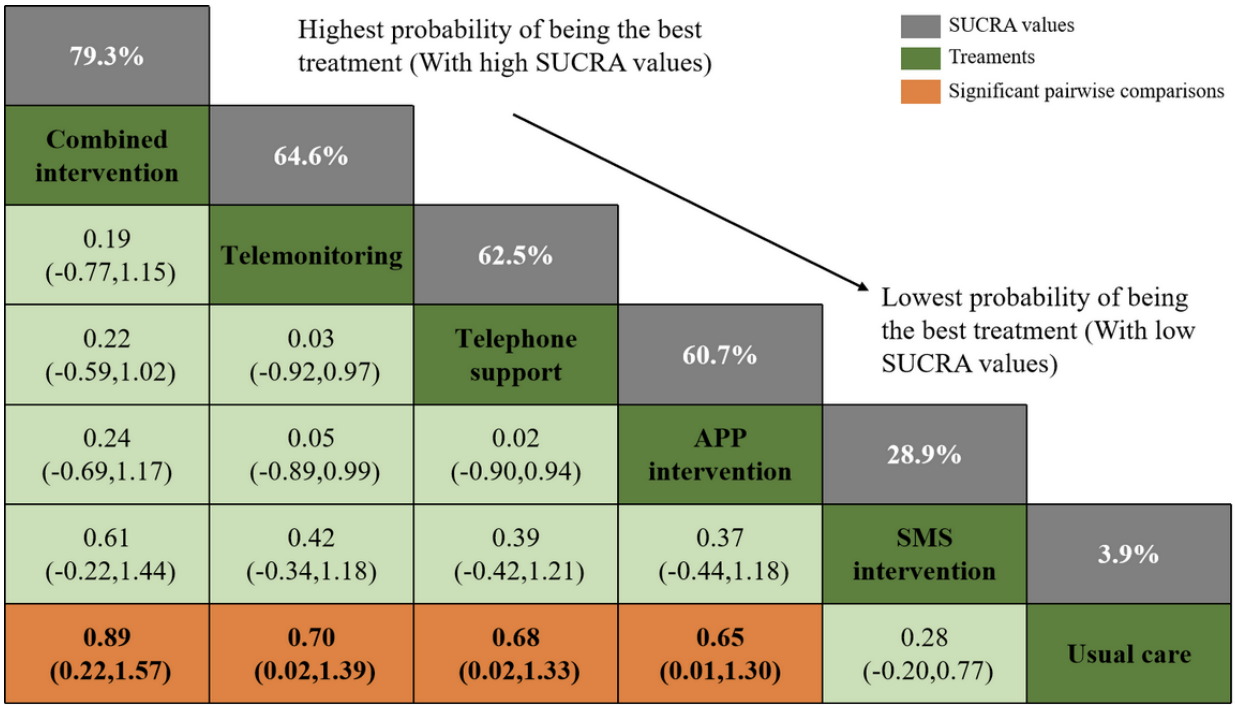

79.3%					
Combined intervention	64.6%				
0.19 (-0.77,1.15)	**Telemonitoring**	62.5%			
0.22 (-0.59,1.02)	0.03 (-0.92,0.97)	**Telephone support**	60.7%		
0.24 (-0.69,1.17)	0.05 (-0.89,0.99)	0.02 (-0.90,0.94)	**APP intervention**	28.9%	
0.61 (-0.22,1.44)	0.42 (-0.34,1.18)	0.39 (-0.42,1.21)	0.37 (-0.44,1.18)	**SMS intervention**	3.9%
0.89 (0.22,1.57)	**0.70 (0.02,1.39)**	**0.68 (0.02,1.33)**	**0.65 (0.01,1.30)**	0.28 (-0.20,0.77)	**Usual care**

XSL•FO
RenderX

Figure 4. Funnel plot of medication adherence. (A) Usual care. (B) Mobile phone applications intervention. (C) Telephone support intervention. (D) Short messaging service intervention. (E) Combined intervention. (F) Telemonitoring intervention.

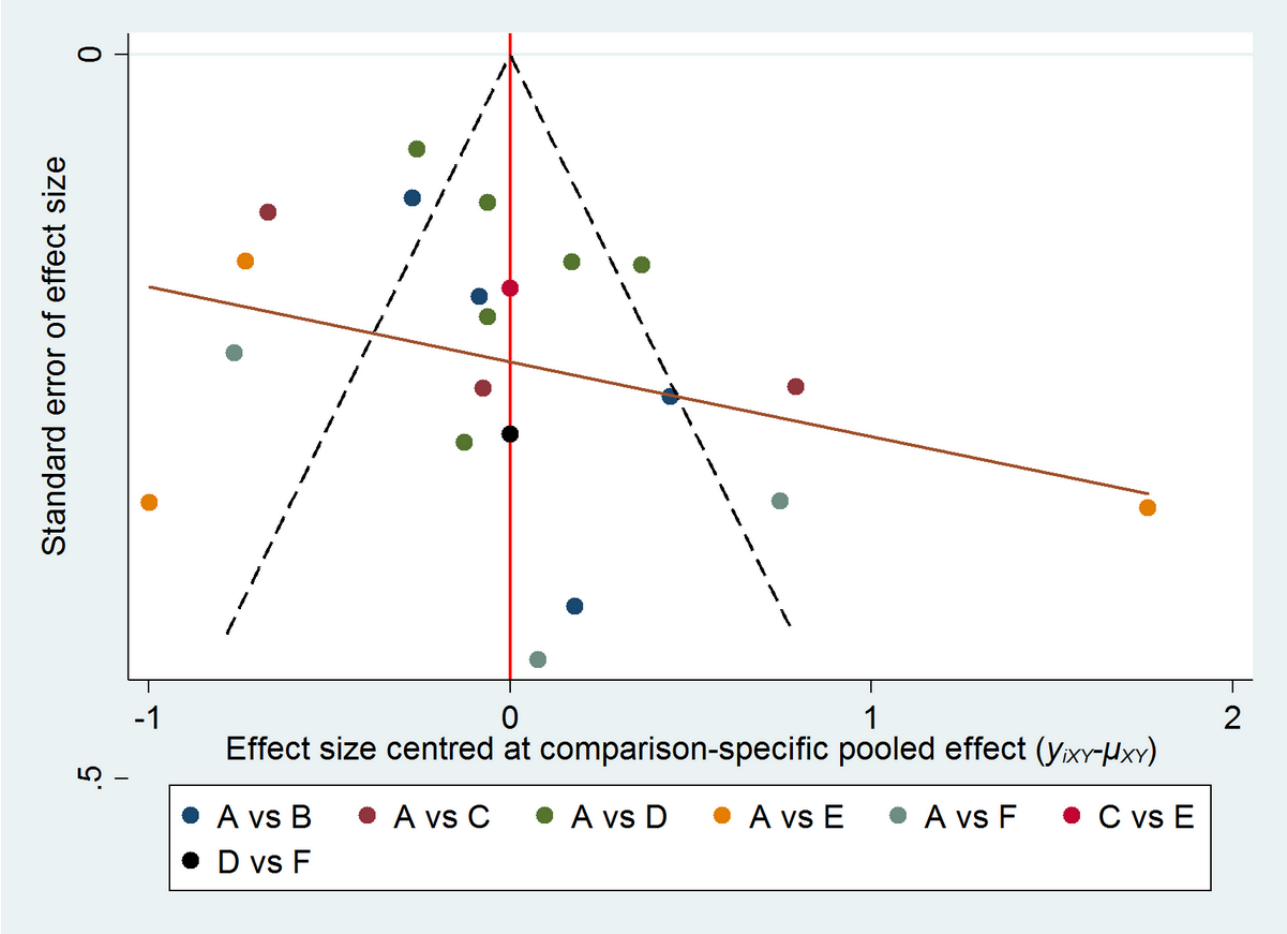

Secondary Outcomes

In terms of systolic blood pressure (SBP), 11 studies with 2608 participants were included in the analysis. The combined interventions emerged as the only method significantly more effective than usual care for lowering SBP. Specifically, it surpassed SMS text messaging interventions (SMD –1.08, 95% CI –2.15 to –0.01) and usual care (SMD –1.21, 95% CI –2.12 to –0.31). There was no statistical difference in app, telephone support, telemonitoring, and SMS text messaging interventions compared with the usual care group. The ranking of interventions based on their efficacy revealed that the combined interventions (89.4%) had the highest probability of being the most effective, followed by the app (67.0%), telephone support (61.1%), telemonitoring (35.3%), and SMS text messaging (30.0%) interventions, and usual care (17.2%; moderate to low certainty).

Regarding diastolic blood pressure (DBP), 10 studies with a total of 2197 participants were analyzed. The combined interventions were the only ones to outperform usual care statistically. It was more effective than SMS text messaging interventions (SMD 0.90, 95% CI –1.70 to –0.10), usual care (SMD –0.95, 95% CI –1.64 to –0.27), and telemonitoring interventions (SMD –1.13, 95% CI –2.11 to –0.16) in reducing DBP. There was no statistical difference in app, telephone support, telemonitoring, and SMS text messaging interventions compared with the usual care group. The results of intervention ranking showed that combined interventions (92.9%) had the highest probability of being the best intervention, followed by telephone support (70.9%), app (49.3%), SMS text messaging (37.2%), usual care (30.2%), and telemonitoring (19.2%; moderate to low certainty). The network plots for SBP and DBP were provided in Figure S5 in Multimedia Appendix 1. Table S6 in Multimedia Appendix 1 detailed the relative effect sizes of SBP and DBP at postintervention according to the NMA, and the SUCRA rankings were presented in Figure S6 in Multimedia Appendix 1. The node-splitting method found no significant inconsistencies between direct and indirect evidence among the eHealth interventions ($P \geq .05$). The funnel plot of SBP and DBP, shown in Figure S7A and B in Multimedia Appendix 1, indicated no evidence of publication bias.

Subgroup Analysis

In subgroups of patients with interventions lasting ≥3 months, the efficacy levels were similar to the main analysis, with the combined intervention significantly outperforming others for improving medication adherence. In contrast, for interventions lasting <3 months, the telephone support intervention emerged as potentially the most effective. The combined interventions demonstrated a significant benefit in medication adherence for males (SMD 2.60, 95% CI 1.06-4.13), while the app interventions showed significant benefits for females (SMD 0.76, 95% CI 0.29-1.23) compared to usual care. Subgroup analysis revealed that among theory-driven eHealth

XSL•FO
RenderX

interventions, the telemonitoring intervention is the most effective. For interventions not based on theory, the combined intervention was the most effective. The relative effect sizes of these subgroup analyses for duration of intervention, gender, and theory-based interventions post intervention according to the NMA are detailed in Table S7 A, B, and C in Multimedia Appendix 1.

Discussion

Principal Findings

To the best of our knowledge, this is the first meta-analysis to assess the comparative effectiveness of eHealth in improving medication adherence among patients with CVDs. The outcomes from direct and indirect comparisons were combined, and various eHealth interventions were quantitatively ranked to identify the optimal protocol. This process laid the foundation for choosing the appropriate eHealth interventions to enhance medication adherence. This systematic review determined that combined interventions, telephone support, telemonitoring interventions, and app interventions appear to be effective in improving medication adherence in individuals with CVDs. Notably, the combined interventions had a 79.3% probability of being the most effective for improving medication adherence. Additionally, combined interventions provided the greatest benefit for the outcome of SBP and DBP. The research outcomes underscore the importance of incorporating multimodal functionalities in the development of eHealth interventions.

In our study, the combined intervention proved to be the most effective method for enhancing medication adherence. This aligns with Ding et al's [55] findings, which highlight that medication adherence in patients is influenced by multiple factors, necessitating a combination of various approaches for a comprehensive intervention strategy. The effectiveness of combined interventions can be attributed to their tailored nature, offering personalized measures attuned to different disease stages. For example, employing telephone interventions for timely feedback during early stages and using automated SMS text messaging interventions for long-term reminders in later stages [51-53]. Additionally, the study of Schoenthaler et al [54] used tablet-based interventions to identify patients' most prominent adherence barriers, employing SMS text message interactions and video education to help alleviate these barriers. The findings reinforce the notion that in the design of electronic interventions, it is advisable to use multimodal functionalities. Such an approach enables the delivery of tailored, comprehensive treatment plans that are best suited to the unique requirements of each patient. Additionally, several studies [56,57] have explored the integration of smart pillboxes with telephones or apps for tracking whether the pillbox has been opened and providing reminders to patients, aiming to mitigate adherence issues associated with forgetfulness. However, these studies use internal algorithms to calculate adherence rates that are not standardized, making it difficult to combine and compare adherence data from different studies in our meta-analysis. Meanwhile, technological advancements have facilitated the development of complex customized algorithms. Ecological momentary assessments allow for real-time monitoring of patients' behaviors and experiences [58]. User personas provide a detailed understanding of patients' characteristics, preferences, and needs [59]. Studies have indicated that integrating these methods as auxiliary support in eHealth is a significant method for providing precise nursing to patients. However, most studies on these technologies focus on technology development and user acceptance, lacking RCTs to validate their efficacy, and were not included in our meta-analysis, highlighting them as potential areas for future research.

This study also highlights the differential effectiveness of various eHealth modalities. Telemonitoring interventions, probably due to their features like audiovisual and alarm support for medication adherence, appear superior to telephone support and app interventions, although telephone support and app interventions were also effective compared to usual care in this analysis. The underlying mechanisms are unclear, but multifaceted functions of the modality might contribute to its effectiveness in improving medication adherence. Audiovisual features can help patients better understand their condition and treatment plans, thereby enhancing their engagement and willingness to follow medical advice. With an alarm function, it can alert patients when medication nonadherence is detected, reducing the instances of forgetting or incorrectly taking medication and assisting in the formation of a habit of taking medication on schedule. Moreover, telemonitoring interventions focus on real-time biometric data tracking and remote clinical intervention, which facilitates rapid access to professional guidance for patients when needed [60]. In contrast, app interventions typically provide textual medication reminders and are more oriented toward encouraging self-management among patients with CVDs, thereby increasing the treatment burden on patients [12]. In addition, our research targets middle-aged and older individuals, with an average age of 56.47, who possess lower health information literacy and find it challenging to keep pace with technological advancements [61]. Coupled with the complex interfaces and small font sizes in app interventions, this renders it challenging for them to effectively engage with these apps [62]. Therefore, designers of eHealth services should aim to develop more user-friendly interfaces and establish clearer information retrieval models by incorporating features such as larger fonts, voice recognition functionality, and simplified navigation. Despite those disadvantages, telephone support and app interventions could also improve medication adherence when compared with usual care.

It should be noted that SMS text messaging interventions alone were not supported by our NMA as an effective tool in improving medication adherence. This limitation may be due to SMS text messaging interventions being a unidirectional form of communication. Among the 5 articles we included that focused on SMS text messaging interventions, only 1 study required participants to respond to SMS text messages within 6 hours to confirm whether they had taken their medication, while the others used automated, scheduled messages. As Yasmin et al [63] point out, communication between health care providers and patients can alter patients' attitudes and behaviors toward disease management. This reminds us that providers not only need to offer informational support but also emotional

XSL•FO
RenderX

support, thereby enhancing the patient's sense of participation and experience. The recommendation for a 2-way SMS text messaging system aligns with a global directive, yet the discreet nature of SMS text messaging should not be overlooked as it offers confidentiality and reduces stigma surrounding the receipt of treatment among patients [64,65]. It should also be noted that the influence of eHealth on patient medication adherence is not instantaneous but is gradually formed through improving their health information efficacy and health awareness [16].

This review shows that objective measures of blood pressure can be improved with combined eHealth interventions. A plausible reason may be that the enhancement of medication adherence via eHealth and the reduction of blood pressure are mutually reinforcing. A single explanation cannot fully capture the influence of combined eHealth interventions on blood pressure, as it is probable that they result from a combination of psychological aspects (such as self-efficacy, encouragement mechanisms) and physiological mechanisms. eHealth encourages patients to adopt regular medication routines and other positive lifestyle modifications, thereby enhancing the efficacy of blood pressure control and achieving the expected treatment outcomes. A previous review [66] also similarly indicated that eHealth in conjunction with other modes significantly reduced blood pressure. In addition, eHealth facilitates a transition from passive treatment to active prevention by continuously monitoring users' health status. This prevention-centric health care model contributes to reducing the frequency of outpatient visits, thereby saving on health care costs [18]. However, among the screened trials, there were no studies assessing the economic benefits of using eHealth interventions to improve medication adherence. This highlights the need for further research to understand their role in improving health outcomes while saving costs.

In the subgroup analysis, the combined interventions exhibited stronger efficacy in patients with an intervention time ≥3 months, which was similar to the results in the main analysis. Conversely, telephone support showed greater efficacy for intervention periods <3 months. A possible explanation is that combined interventions may provide patients with more comprehensive support, including informational education, reminder services, and behavioral incentives, which are particularly effective in long-term interventions. In contrast, for short-term interventions, the directness and convenience of telephone support may be more acceptable to patients [67].

It is worth noting that approximately 48% of the studies in this review reported the use of theory-driven design, such as self-determination theory and cognitive behavioral theory, to guide their eHealth interventions. In the subgroup of theory-driven eHealth interventions, telemonitoring intervention ranked first. Combined intervention became the most effective for patients among all interventions that are not based on theory. When interventions are theory-driven, telemonitoring interventions may be tailored more precisely and personalized, focusing on the behaviors and motivations of target users. In contrast, a diversified approach might be more adaptable to the needs and preferences of different patients, especially in situations lacking theory-driven guidance. Research indicates that employing theory-driven approaches to developing eHealth interventions may not only enhance our arsenal of effective strategies but also potentially enrich our theoretical and practical insights into changing health behaviors [68]. Therefore, to develop optimal eHealth interventions, it is essential to establish specific behavioral objectives and hypothesized mechanisms of action [69].

Limitations

Among the limitations of this study is the fact that, despite our careful consideration in classifying eHealth interventions, the classifications may not be entirely accurate due to the diversity of components and settings inherent in eHealth devices. Furthermore, we were unable to perform subgroup analyses based on specific types of CVDs many studies did not focus exclusively on a single condition. This limited our ability to conduct a more in-depth analysis of differences in medication adherence among patients with different CVDs. Moreover, our research did not account for differences in medication adherence among the various drugs included in the studies. Factors such as side effects and the palatability of medications could contribute to varying rates of nonadherence. Finally, given the lack of a gold standard for assessing medication adherence, our study excluded those that did not use validated scales to measure adherence. This exclusion may have led to the omission of some relevant studies, potentially affecting the comprehensiveness of our analysis.

Conclusions

To summarize, the research identifies combined interventions (SMS text messaging and telephone support) as the most effective eHealth method for improving medication adherence in patients with CVDs, followed in effectiveness by telemonitoring, telephone support, and app interventions. These findings could guide future practice-based interventions aimed at optimizing patient medication adherence. However, it is important to note that existing intervention protocols differ in terms of timing and specific strategies, necessitating further research to demonstrate the effectiveness of different eHealth interventions.

Acknowledgments

We affirm that the work submitted for publication is original and has not been submitted elsewhere for print or electronic publication consideration. We affirm that each person listed as an author participated in the work in a substantive manner and is prepared to take public responsibility for it. All authors consent to the investigation of any improprieties that may be alleged regarding the work. This research was supported by the Key Program of the National Natural Science Foundation of China (72034005).

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in the supplementary files accompanying this published article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 2142 KB - jmir_v26i1e58013_app1.docx]

References

1. Olvera LE, Ballard B, Jan A. Cardiovascular Disease. Treasure Island, FL: StatPearls Publishing LLC; 2023.
2. Simon ST, Kini V, Levy AE, Ho PM. Medication adherence in cardiovascular medicine. BMJ 2021;374:n1493. [doi: 10.1136/bmj.n1493] [Medline: 34380627]
3. Piña IL, Di Palo KE, Brown MT, Choudhry NK, Cvengros J, Whalen D, et al. Medication adherence: importance, issues and policy: a policy statement from the american heart association. Prog Cardiovasc Dis 2021;64:111-120. [doi: 10.1016/j.pcad.2020.08.003] [Medline: 32800791]
4. Chowdhury R, Khan H, Heydon E, Shroufi A, Fahimi S, Moore C, et al. Adherence to cardiovascular therapy: a meta-analysis of prevalence and clinical consequences. Eur Heart J 2013;34(38):2940-2948. [doi: 10.1093/eurheartj/eht295] [Medline: 23907142]
5. Naderi SH, Bestwick JP, Wald DS. Adherence to drugs that prevent cardiovascular disease: meta-analysis on 376,162 patients. Am J Med 2012;125(9):882-887. [doi: 10.1016/j.amjmed.2011.12.013] [Medline: 22748400]
6. Iuga AO, McGuire MJ. Adherence and health care costs. Risk Manag Healthc Policy 2014;7:35-44 [FREE Full text] [doi: 10.2147/RMHP.S19801] [Medline: 24591853]
7. Gandapur Y, Kianoush S, Kelli HM, Misra S, Urrea B, Blaha MJ, et al. The role of mHealth for improving medication adherence in patients with cardiovascular disease: a systematic review. Eur Heart J Qual Care Clin Outcomes 2016;2(4):237-244 [FREE Full text] [doi: 10.1093/ehjqcco/qcw018] [Medline: 29474713]
8. Kim K, Shin S, Kim S, Lee E. The relation between ehealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023;25:e40778 [FREE Full text] [doi: 10.2196/40778] [Medline: 36716080]
9. Pagoto S, Bennett GG. How behavioral science can advance digital health. Transl Behav Med 2013;3(3):271-276 [FREE Full text] [doi: 10.1007/s13142-013-0234-z] [Medline: 24073178]
10. Beatty AL, Fukuoka Y, Whooley MA. Using mobile technology for cardiac rehabilitation: a review and framework for development and evaluation. J Am Heart Assoc 2013;2(6):e000568 [FREE Full text] [doi: 10.1161/JAHA.113.000568] [Medline: 24185949]
11. Snoswell CL, Taylor ML, Comans TA, Smith AC, Gray LC, Caffery LJ. Determining if telehealth can reduce health system costs: scoping review. J Med Internet Res 2020;22(10):e17298 [FREE Full text] [doi: 10.2196/17298] [Medline: 33074157]
12. Akinosun AS, Polson R, Diaz-Skeete Y, De Kock JH, Carragher L, Leslie S, et al. Digital technology interventions for risk factor modification in patients with cardiovascular disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021;9(3):e21061 [FREE Full text] [doi: 10.2196/21061] [Medline: 33656444]
13. Al-Arkee S, Mason J, Lane DA, Fabritz L, Chua W, Haque MS, et al. Mobile apps to improve medication adherence in cardiovascular disease: systematic review and meta-analysis. J Med Internet Res 2021;23(5):e24190 [FREE Full text] [doi: 10.2196/24190] [Medline: 34032583]
14. Kuan PX, Chan WK, Fern Ying DK, Rahman MAA, Peariasamy KM, Lai NM, et al. Efficacy of telemedicine for the management of cardiovascular disease: a systematic review and meta-analysis. Lancet Digit Health 2022;4(9):e676-e691 [FREE Full text] [doi: 10.1016/S2589-7500(22)00124-8] [Medline: 36028290]
15. Fjellså HMH, Husebø AML, Storm M. eHealth in care coordination for older adults living at home: scoping review. J Med Internet Res 2022;24(10):e39584 [FREE Full text] [doi: 10.2196/39584] [Medline: 36256831]
16. Pouls BPH, Vriezekolk JE, Bekker CL, Linn AJ, van Onzenoort HAW, Vervloet M, et al. Effect of interactive eHealth interventions on improving medication adherence in adults with long-term medication: systematic review. J Med Internet Res 2021;23(1):e18901 [FREE Full text] [doi: 10.2196/18901] [Medline: 33416501]
17. Cavero-Redondo I, Saz-Lara A, Sequí-Dominguez I, Gómez-Guijarro MD, Ruiz-Grao MC, Martinez-Vizcaino V, et al. Comparative effect of eHealth interventions on hypertension management-related outcomes: a network meta-analysis. Int J Nurs Stud 2021;124:104085 [FREE Full text] [doi: 10.1016/j.ijnurstu.2021.104085] [Medline: 34601205]
18. Arshed M, Mahmud AB, Minhat HS, Ying LP, Umer MF. Effectiveness of mHealth interventions in medication adherence among patients with cardiovascular diseases: a systematic review. Diseases 2023;11(1):41 [FREE Full text] [doi: 10.3390/diseases11010041] [Medline: 36975590]

XSL•FO
RenderX

19. Bingham JM, Black M, Anderson EJ, Li Y, Toselli N, Fox S, et al. Impact of telehealth interventions on medication adherence for patients with type 2 diabetes, hypertension, and/or dyslipidemia: a systematic review. Ann Pharmacother 2021;55(5):637-649. [doi: 10.1177/1060028020950726] [Medline: 32815400]
20. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151(4):264-269 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
21. National CCFPC. Medicines adherence: involving patients in decisions about prescribed medicines and supporting adherence. In: National Institute for Health and Clinical Excellence: Guidance. London (UK): Royal College of General Practitioners; 2009.
22. Risser J, Jacobson TA, Kripalani S. Development and psychometric evaluation of the Self-efficacy for Appropriate Medication Use Scale (SEAMS) in low-literacy patients with chronic disease. J Nurs Meas 2007;15(3):203-219. [doi: 10.1891/106137407783095757] [Medline: 18232619]
23. Ogedegbe G, Mancuso CA, Allegrante JP, Charlson ME. Development and evaluation of a medication adherence self-efficacy scale in hypertensive African-American patients. J Clin Epidemiol 2003 Jun;56(6):520-529. [doi: 10.1016/s0895-4356(03)00053-2] [Medline: 12873646]
24. Morisky DE, Ang A, Krousel-Wood M, Ward HJ. Predictive validity of a medication adherence measure in an outpatient setting. J Clin Hypertens (Greenwich) 2008 May;10(5):348-354 [FREE Full text] [doi: 10.1111/j.1751-7176.2008.07572.x] [Medline: 18453793]
25. Kuster AE, Fong CM. Further psychometric evaluation of the Spanish language health-promoting lifestyle profile. Nurs Res 1993;42(5):266-269. [Medline: 8415038]
26. Stone JK, Shafer LA, Graff LA, Lix L, Witges K, Targownik LE, et al. Utility of the MARS-5 in assessing medication adherence in IBD. Inflamm Bowel Dis 2021 Feb 16;27(3):317-324 [FREE Full text] [doi: 10.1093/ibd/izaa056] [Medline: 32195533]
27. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
28. Puhan MA, Schünemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, GRADE Working Group. A grade working group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014;349:g5630 [FREE Full text] [doi: 10.1136/bmj.g5630] [Medline: 25252733]
29. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7(1):1-76 [FREE Full text] [Medline: 12583822]
30. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
31. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods 2012;3(2):80-97. [doi: 10.1002/jrsm.1037] [Medline: 26062083]
32. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med 2016;13(5):e1002028 [FREE Full text] [doi: 10.1371/journal.pmed.1002028] [Medline: 27218655]
33. van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Res Synth Methods 2016;7(1):80-93 [FREE Full text] [doi: 10.1002/jrsm.1167] [Medline: 26461181]
34. Zha P, Qureshi R, Porter S, Chao Y, Pacquiao D, Chase S, et al. Utilizing a mobile health intervention to manage hypertension in an underserved community. West J Nurs Res 2020;42(3):201-209. [doi: 10.1177/0193945919847937] [Medline: 31057081]
35. Kim JY, Wineinger NE, Steinhubl SR. The influence of wireless self-monitoring program on the relationship between patient activation and health behaviors, medication adherence, and blood pressure levels in hypertensive patients: a substudy of a randomized controlled trial. J Med Internet Res 2016;18(6):e116 [FREE Full text] [doi: 10.2196/jmir.5429] [Medline: 27334418]
36. Sarfo FS, Treiber F, Gebregziabher M, Adamu S, Nichols M, Singh A, PINGS Team. Phone-based intervention for blood pressure control among Ghanaian stroke survivors: a pilot randomized controlled trial. Int J Stroke 2018;14(6):630-638. [doi: 10.1177/1747493018816423] [Medline: 30465630]
37. Chandler J, Sox L, Kellam K, Feder L, Nemeth L, Treiber F. Impact of a culturally tailored mhealth medication regimen self-management program upon blood pressure among hypertensive hispanic adults. Int J Environ Res Public Health 2019;16(7):1226 [FREE Full text] [doi: 10.3390/ijerph16071226] [Medline: 30959858]
38. He J, Chen J, Li Q, Yang Z, Liang H, Wang L, et al. Application of family-involved smart medication management system in rural-dwelling middle-aged and older adult participants with chronic diseases: Management of chronic diseases in rural areas. Medicine (Baltimore) 2022;101(45):e31662 [FREE Full text] [doi: 10.1097/MD.0000000000031662] [Medline: 36397443]

XSL•FO
RenderX

39. Yanicelli LM, Goy CB, González VDC, Palacios GN, Martínez EC, Herrera MC. Non-invasive home telemonitoring system for heart failure patients: a randomized clinical trial. J Telemed Telecare 2021;27(9):553-561. [doi: 10.1177/1357633X19899261] [Medline: 31973633]
40. Morawski K, Ghazinouri R, Krumme A, Lauffenburger JC, Lu Z, Durfee E, et al. Association of a smartphone application with medication adherence and blood pressure control: the medisafe-bp randomized clinical trial. JAMA Intern Med 2018;178(6):802-809 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0447] [Medline: 29710289]
41. Santo K, Singleton A, Rogers K, Thiagalingam A, Chalmers J, Chow CK, et al. Medication reminder applications to improve adherence in coronary heart disease: a randomised clinical trial. Heart 2019;105(4):323-329. [doi: 10.1136/heartjnl-2018-313479] [Medline: 30150326]
42. Wan L, Zhang X, Mo M, Xiong X, Ou C, You L, et al. Effectiveness of goal-setting telephone follow-up on health behaviors of patients with ischemic stroke: a randomized controlled trial. J Stroke Cerebrovasc Dis 2016;25(9):2259-2270. [doi: 10.1016/j.jstrokecerebrovasdis.2016.05.010] [Medline: 27371106]
43. Migneault JP, Dedier JJ, Wright JA, Heeren T, Campbell MK, Morisky DE, et al. A culturally adapted telecommunication system to improve physical activity, diet quality, and medication adherence among hypertensive African-Americans: a randomized controlled trial. Ann Behav Med 2012;43(1):62-73 [FREE Full text] [doi: 10.1007/s12160-011-9319-4] [Medline: 22246660]
44. Najafi SS, Shaabani M, Momennassab M, Aghasadeghi K. The nurse-led telephone follow-up on medication and dietary adherence among patients after myocardial infarction: a randomized controlled clinical trial. Int J Community Based Nurs Midwifery 2016;4(3):199-208 [FREE Full text] [Medline: 27382586]
45. Buis L, Hirzel L, Dawood RM, Dawood KL, Nichols LP, Artinian NT, et al. Text messaging to improve hypertension medication adherence in african americans from primary care and emergency department settings: results from two randomized feasibility studies. JMIR Mhealth Uhealth 2017;5(2):e9 [FREE Full text] [doi: 10.2196/mhealth.6630] [Medline: 28148474]
46. Bermon A, Uribe AF, Pérez-Rivero PF, Prieto-Merino D, Saaibi JF, Silva FA, et al. Efficacy and safety of text messages targeting adherence to cardiovascular medications in secondary prevention: txt2heart colombia randomized controlled trial. JMIR Mhealth Uhealth 2021;9(7):e25548 [FREE Full text] [doi: 10.2196/25548] [Medline: 34319247]
47. Bhandari B, Narasimhan P, Jayasuriya R, Vaidya A, Schutte AE. Effectiveness and acceptability of a mobile phone text messaging intervention to improve blood pressure control (text4bp) among patients with hypertension in nepal: a feasibility randomised controlled trial. Glob Heart 2022;17(1):13 [FREE Full text] [doi: 10.5334/gh.1103] [Medline: 35342691]
48. Park LG, Howie-Esquivel J, Whooley MA, Dracup K. Psychosocial factors and medication adherence among patients with coronary heart disease: a text messaging intervention. Eur J Cardiovasc Nurs 2015;14(3):264-273. [doi: 10.1177/1474515114537024] [Medline: 24853566]
49. Kamal AK, Shaikh Q, Pasha O, Azam I, Islam M, Memon AA, et al. A randomized controlled behavioral intervention trial to improve medication adherence in adult stroke patients with prescription tailored Short Messaging Service (SMS)-SMS4Stroke study. BMC Neurol 2015;15:212 [FREE Full text] [doi: 10.1186/s12883-015-0471-5] [Medline: 26486857]
50. Zhai P, Hayat K, Ji W, Li Q, Shi L, Atif N, et al. Efficacy of text messaging and personal consultation by pharmacy students among adults with hypertension: randomized controlled trial. J Med Internet Res 2020;22(5):e16019 [FREE Full text] [doi: 10.2196/16019] [Medline: 32432556]
51. Wan L, Zhang X, You L, Ruan H, Chen S. The efficacy of a comprehensive reminder system to improve health behaviors and blood pressure control in hypertensive ischemic stroke patients: a randomized controlled trial. J Cardiovasc Nurs 2018;33(6):509-517. [doi: 10.1097/JCN.0000000000000496] [Medline: 29901484]
52. Kes D, Polat U. The effect of nurse-led telephone support on adherence to blood pressure control and drug treatment in individuals with primary hypertension: a randomized controlled study. Int J Nurs Pract 2022;28(3):e12995. [doi: 10.1111/ijn.12995] [Medline: 34318542]
53. Kamal AK, Khalid W, Muqeet A, Jamil A, Farhat K, Gillani SRA, et al. Making prescriptions "talk" to stroke and heart attack survivors to improve adherence: results of a randomized clinical trial (the talking Rx study). PLoS One 2018;13(12):e0197671 [FREE Full text] [doi: 10.1371/journal.pone.0197671] [Medline: 30571697]
54. Schoenthaler A, Leon M, Butler M, Steinhaeuser K, Wardzinski W. Development and evaluation of a tailored mobile health intervention to improve medication adherence in black patients with uncontrolled hypertension and type 2 diabetes: pilot randomized feasibility trial. JMIR Mhealth Uhealth 2020;8(9):e17135 [FREE Full text] [doi: 10.2196/17135] [Medline: 32965230]
55. Ding H, Chen SH, Edwards I, Jayasena R, Doecke J, Layland J, et al. Effects of different telemonitoring strategies on chronic heart failure care: systematic review and subgroup meta-analysis. J Med Internet Res 2020;22(11):e20032 [FREE Full text] [doi: 10.2196/20032] [Medline: 33185554]
56. Choudhry NK, Krumme AA, Ercole PM, Girdish C, Tong AY, Khan NF, et al. Effect of reminder devices on medication adherence: the remind randomized clinical trial. JAMA Intern Med 2017;177(5):624-631 [FREE Full text] [doi: 10.1001/jamainternmed.2016.9627] [Medline: 28241271]

XSL•FO
RenderX

57. Volpp KG, Troxel AB, Mehta SJ, Norton L, Zhu J, Lim R, et al. Effect of electronic reminders, financial incentives, and social support on outcomes after myocardial infarction: the heartstrong randomized clinical trial. JAMA Intern Med 2017;177(8):1093-1101 [FREE Full text] [doi: 10.1001/jamainternmed.2017.2449] [Medline: 28654972]
58. Giurgiu M, Niermann C, Ebner-Priemer U, Kanning M. Accuracy of sedentary behavior-triggered ecological momentary assessment for collecting contextual information: development and feasibility study. JMIR Mhealth Uhealth 2020;8(9):e17852 [FREE Full text] [doi: 10.2196/17852] [Medline: 32930668]
59. Haldane V, Koh JJK, Srivastava A, Teo KWQ, Tan YG, Cheng RX, et al. User preferences and persona design for an mhealth intervention to support adherence to cardiovascular disease medication in singapore: a multi-method study. JMIR Mhealth Uhealth 2019;7(5):e10465 [FREE Full text] [doi: 10.2196/10465] [Medline: 31140445]
60. Brons M, Koudstaal S, Asselbergs FW. Algorithms used in telemonitoring programmes for patients with chronic heart failure: a systematic review. Eur J Cardiovasc Nurs 2018;17(7):580-588 [FREE Full text] [doi: 10.1177/1474515118786838] [Medline: 29954184]
61. Oh SS, Kim K, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021;23(2):e26145 [FREE Full text] [doi: 10.2196/26145] [Medline: 33533727]
62. Peacock E, Craig LS, Krousel-Wood M. Electronic health strategies to improve medication adherence in patients with cardiometabolic disease: current status and future directions. Curr Opin Cardiol 2022;37(4):307-316 [FREE Full text] [doi: 10.1097/HCO.0000000000000971] [Medline: 35731675]
63. Yasmin F, Banu B, Zakir SM, Sauerborn R, Ali L, Souares A. Positive influence of short message service and voice call interventions on adherence and health outcomes in case of chronic disease care: a systematic review. BMC Med Inform Decis Mak 2016;16:46 [FREE Full text] [doi: 10.1186/s12911-016-0286-3] [Medline: 27106263]
64. Thakkar J, Kurup R, Laba T, Santo K, Thiagalingam A, Rodgers A, et al. Mobile telephone text messaging for medication adherence in chronic disease: a meta-analysis. JAMA Intern Med 2016;176(3):340-349. [doi: 10.1001/jamainternmed.2015.7667] [Medline: 26831740]
65. Aguilera A, Muñoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pr 2011;42(6):472-478 [FREE Full text] [doi: 10.1037/a0025499] [Medline: 25525292]
66. Tucker KL, Sheppard JP, Stevens R, Bosworth HB, Bove A, Bray EP, et al. Self-monitoring of blood pressure in hypertension: a systematic review and individual patient data meta-analysis. PLoS Med 2017;14(9):e1002389 [FREE Full text] [doi: 10.1371/journal.pmed.1002389] [Medline: 28926573]
67. Du S, Liu W, Cai S, Hu Y, Dong J. The efficacy of e-health in the self-management of chronic low back pain: a meta analysis. Int J Nurs Stud 2020;106:103507. [doi: 10.1016/j.ijnurstu.2019.103507] [Medline: 32320936]
68. Ribaut J, Leppla L, Teynor A, Valenta S, Dobbels F, Zullig LL, SMILe study team. Theory-driven development of a medication adherence intervention delivered by eHealth and transplant team in allogeneic stem cell transplantation: the SMILe implementation science project. BMC Health Serv Res 2020;20(1):827 [FREE Full text] [doi: 10.1186/s12913-020-05636-1] [Medline: 32878623]
69. Simoni JM, Ronen K, Aunon FM. Health behavior theory to enhance ehealth intervention research in HIV: rationale and review. Curr HIV/AIDS Rep 2018;15(6):423-430 [FREE Full text] [doi: 10.1007/s11904-018-0418-8] [Medline: 30511186]

Abbreviations

CNKI: China National Knowledge Infrastructure Library
CVD: cardiovascular disease
DBP: diastolic blood pressure
GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
PICOS: Participants, Interventions, Comparison, Outcomes, and Study Design
PRISMA-NMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses
RCT: randomized controlled trial
SBP: systolic blood pressure
SMD: standardized mean difference
SUCRA: surface under the cumulative ranking

XSL•FO
RenderX

Edited by G Eysenbach, N Cahill; submitted 03.03.24; peer-reviewed by D Xueying, W Jia, Q Li; comments to author 28.03.24; revised version received 02.04.24; accepted 30.05.24; published 15.07.24.

Please cite as:
Miao Y, Luo Y, Zhao Y, Liu M, Wang H, Wu Y
Effectiveness of eHealth Interventions in Improving Medication Adherence Among Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e58013
URL: https://www.jmir.org/2024/1/e58013
doi:10.2196/58013
PMID:

©Yiqun Miao, Yuan Luo, Yuhan Zhao, Mingxuan Liu, Huiying Wang, Ying Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Evaluating the Potential and Pitfalls of AI-Powered Conversational Agents as Humanlike Virtual Health Carers in the Remote Management of Noncommunicable Diseases: Scoping Review

Sadia Azmin Anisha[1*], BBusCom, MBusInfSys; Arkendu Sen[1*], MBBS, GCHE, DipOphthal, DipForensMed, MSurg, PhD; Chris Bain[2*], MBBS, MInfoTech, MACS, PhD

[1]Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia, Bandar Sunway, Malaysia

[2]Faculty of Information Technology, Data Future Institutes, Monash University, Clayton, Australia

[*]all authors contributed equally

Corresponding Author:
Arkendu Sen, MBBS, GCHE, DipOphthal, DipForensMed, MSurg, PhD
Jeffrey Cheah School of Medicine & Health Sciences
Monash University Malaysia
Jalan Lagoon Selatan
Bandar Sunway, 47500
Malaysia
Phone: 60 3 551 46325
Fax: 60 3 551 46323
Email: arkendu.sen@monash.edu

Abstract

Background: The rising prevalence of noncommunicable diseases (NCDs) worldwide and the high recent mortality rates (74.4%) associated with them, especially in low- and middle-income countries, is causing a substantial global burden of disease, necessitating innovative and sustainable long-term care solutions.

Objective: This scoping review aims to investigate the impact of artificial intelligence (AI)–based conversational agents (CAs)—including chatbots, voicebots, and anthropomorphic digital avatars—as human-like health caregivers in the remote management of NCDs as well as identify critical areas for future research and provide insights into how these technologies might be used effectively in health care to personalize NCD management strategies.

Methods: A broad literature search was conducted in July 2023 in 6 electronic databases—Ovid MEDLINE, Embase, PsycINFO, PubMed, CINAHL, and Web of Science—using the search terms "conversational agents," "artificial intelligence," and "noncommunicable diseases," including their associated synonyms. We also manually searched gray literature using sources such as ProQuest Central, ResearchGate, ACM Digital Library, and Google Scholar. We included empirical studies published in English from January 2010 to July 2023 focusing solely on health care–oriented applications of CAs used for remote management of NCDs. The narrative synthesis approach was used to collate and summarize the relevant information extracted from the included studies.

Results: The literature search yielded a total of 43 studies that matched the inclusion criteria. Our review unveiled four significant findings: (1) higher user acceptance and compliance with anthropomorphic and avatar-based CAs for remote care; (2) an existing gap in the development of personalized, empathetic, and contextually aware CAs for effective emotional and social interaction with users, along with limited consideration of ethical concerns such as data privacy and patient safety; (3) inadequate evidence of the efficacy of CAs in NCD self-management despite a moderate to high level of optimism among health care professionals regarding CAs' potential in remote health care; and (4) CAs primarily being used for supporting nonpharmacological interventions such as behavioral or lifestyle modifications and patient education for the self-management of NCDs.

Conclusions: This review makes a unique contribution to the field by not only providing a quantifiable impact analysis but also identifying the areas requiring imminent scholarly attention for the ethical, empathetic, and efficacious implementation of AI in NCD care. This serves as an academic cornerstone for future research in AI-assisted health care for NCD management.

Trial Registration: Open Science Framework; https://doi.org/10.17605/OSF.IO/GU5PX

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56114) doi:10.2196/56114

KEYWORDS

conversational agents; artificial intelligence; noncommunicable disease; self-management; remote monitoring; mobile phone

Introduction

Burden of Noncommunicable Diseases

Noncommunicable diseases (NCDs), also known as chronic diseases, are medical conditions that are not primarily caused by infectious agents (eg, viruses, bacteria, fungi, or parasites) and cannot be transmitted from one individual to another through close contact [1]. NCDs, such as cancer, cardiovascular diseases (CVDs), chronic obstructive pulmonary diseases, chronic respiratory diseases, chronic kidney diseases, cognitive disorders, metabolic syndrome, diabetes, and hypertension, are rising worldwide, with significantly higher rates in low- and middle-income countries (LMICs) [2-4]. In 2019, NCDs contributed to the highest proportion of total mortality (74.4%), accounting for 7.1 million additional deaths in 2019 as compared to 2009 [5]. Nearly half of the mortalities in Asia are attributable to NCDs, causing >80% of CVD and diabetes deaths, 90% of chronic obstructive pulmonary disease deaths, and two-thirds of all cancer deaths occurring in LMICs, resulting in 47% of the global burden of disease [3,6]. Major risk factors of NCDs include unhealthy dietary habits; physical inactivity; stress; and consumption of drugs, tobacco, and alcohol, which are generally modifiable due to lifestyle choices [5,7]. These chronic diseases are also major causes of long-term disabilities and prolonged costly treatment that may pose serious threats to a country's health care resources and expenditure, especially among lower-income countries where health systems are not sufficiently equipped to tackle the escalating challenges [3,4,8].

In addition to regular visits to health care centers, NCDs require challenging self-care management, including compliance with medications, lifestyle modifications, and constant symptom monitoring to prevent disease progression; nonetheless, adherence to these procedures is generally low, especially among adults with limited health literacy who struggle to comprehend and follow instructions from their health care professionals [9,10]. Moreover, the shortage of health care providers and their limited time availability are substantial causes of patients' being deprived of receiving adequate health education and support to make informed decisions required for the effective self-management of their chronic illnesses [11,12]. In addition, access to proper health services can be limited in many underdeveloped areas and rural communities due to poor health care infrastructure and mobility facilities [13].

The optimal prevention management strategy should incorporate elements of individual lifestyle management, societal health awareness management, national health policy decisions, and global health strategy [7]. However, the escalating prevalence of NCDs worldwide reveals that traditional disease management techniques are not sufficiently effective, thereby indicating an urgent necessity to develop effective supplementary management strategies to mitigate the substantial financial burden imposed by NCDs on many households, particularly in LMICs [2,3,14].

The Role of Telehealth in NCD Management

Telehealth applications have the potential to improve patient self-care and disease-specific knowledge as well as minimize hospitalizations and mortality [15]. There is evidence from several past studies suggesting the significant effectiveness of mobile-based telehealth apps in improving nutritional intake and physical activity with technology intervention, resulting in body weight loss and adoption of recommended lifestyle changes due to their convenience and accessibility [16]. This is because such apps, with the help of existing and emerging technologies, have the ability to assist patients in managing their chronic diseases more effectively by providing constant self-monitoring tools and promoting improved self-management of health problems [17]. Furthermore, the ever-growing ownership of mobile devices worldwide has greatly contributed to the shift toward digital health care services, including assessment, monitoring, and treatment of physical and mental health, thereby indicating the promising ability of mobile health apps in self-monitoring, assessment, and treatment of NCDs at a reduced cost [18-20]. Indeed, the COVID-19 pandemic has expedited the use of telehealth [21,22], which has been proposed as a cost-effective method of delivering better health care services to people with chronic illnesses in a more flexible, personalized, transparent, dynamic, and accessible way [21,23].

Potential Enhancement of Existing Telehealth Apps With Artificial Intelligence

Although mobile-based telehealth apps offer an ideal platform for systems designed to help patients manage their chronic illnesses due to the smartphones' computational power, connectivity, and consistent availability, many individuals struggle with complex user interfaces of existing digital health technologies [10]. While the ubiquity of these apps reduces some acceptance barriers, most apps still overlook many other barriers, such as lack of motivational, psychological, and emotional support [10,24]. Nonetheless, technological advancement involving artificial intelligence (AI) seems to have the potential to further upgrade existing mobile health apps with more user-friendly features that can support individual user needs [25]. For instance, many patients lack the ability to effectively navigate conventional telehealth apps due to limited health-related or computer literacy and disabilities such as visual impairment; hence, such intelligent dialogue systems, specifically termed *conversational agents* (CAs), may help overcome these limitations and improve usability by providing an oral presentation of the apps' contents in plain language [15]. AI-powered CAs are computer systems that can communicate with humans through text, voice, and images on mobile, web-based, or audio-based platforms using AI techniques such as machine learning (ML; a statistical method of training models using data for making predictions based on a variety of features) and natural language processing (machines' ability to detect and interpret humans' verbal and written languages) [26,27]. A CA may serve as an *empathetic* listener to understand patients'

XSL•FO
RenderX

problems as well as aid in monitoring a patient's health 24/7 and notify physicians about an anticipated medical emergency [13].

The popularity of CAs, especially those that use unrestrained natural language, has increased over the last decade as consumers can use their smartphones to interact with CAs for daily tasks [27]. When deployed on mobile devices, CAs have the potential to augment human intelligence and demonstrate multiple benefits such as delivering health education and behavior change for a range of chronic health conditions [10]. Moreover, these AI-based CAs provide additional communication channels that are particularly effective for developing trust and therapeutic alliance to encourage adherence among users [10]. The human-like conversational features of CAs due to advancements in natural language processing, voice recognition, and AI are increasingly substituting human employees in service encounters, including the health care industry to deliver personalized care and support to individuals with chronic health issues [28].

Types of CAs

In this paper, we broadly categorize intelligent CAs into 3 main types (Table 1): *chatbots* (text based or voice enabled—*voicebots*) without embodiment, computer-based embodied *digital avatars*, and physically embodied *humanoid robots* [29]. Although *robots* are beyond the scope of this paper, AI humanoid or social robots can also be classified as CAs that can be used as human-like health caregivers for managing NCDs. Despite the differences, all CAs, including humanoid robots, aim to enhance relational outcomes through human-like communication [29].

XSL•FO
RenderX

Table 1. Types of conversational agents (CAs).

AI[a]-based CA type (with examples)	Features and functionality	Interaction mode	Advantages	Challenges
Chatbots (eg, ChatGPT)	• Computer programs integrated into messaging platforms that interact with users via free text [30] • Evolved from preprogrammed, fixed scripted responses to AI-powered versions enabling more human-like conversations [31-33]	Text based	• Simple user interfaces • Easy and cheaper to design and develop	• Offers mostly chat-based interactions • Lack of personalized emotional tones [34]
Voicebots (or voice assistants; eg, Amazon Alexa and Apple's Siri)	• Voice-enabled intelligent chatbots that interact with users and respond to speaker commands primarily through voice [35,36]	Voice based (or text+voice enabled)	• Offers greater flexibility by allowing for hands-free conversations and multitasking [35] • Beneficial for users with limited digital expertise (eg, users with typing inabilities)	• Development of enhanced voice interfaces crucial for a more natural conversation [36] • Require some level of emotion awareness (a significant component of voice conversations) [36] • Prone to speech recognition errors [37]
Digital avatars (eg, Replika, Mitsuku, and Soul Machines)	• AI-created anthropomorphic representations of real-world characters in a computer-simulated environment, partially automated in some actions and movements, emulating human behavior [38,39] • Digital AI avatars can generate human-like interactions through a fusion of multimodal features [40,41] • Digital avatars evolved from 2D cartoonlike characters to visually realistic and interactive human faces with 3D imaging, leading to the emergence of highly realistic avatars ("Digital humans") [42] • Companies such as Soul Machines use CGI[b], AI, and NLP[c] to create digital human prototypes in metaverse spaces [43] • Digital humans can represent fictional characters or virtual replicas of real humans, requiring digital twin technology for data-driven personalized care [44-46]	Multimodal (text based, voice activated, and face-to-face)	• Ability to mimic natural human interactions by delivering highly personalized responses [40,41] • Capable of face-to-face conversations, displaying physical nonverbal behaviors (eg, facial expressions, hand gestures, nodding, and head and body postures) [47]	• Technical implementation is more complex, time-consuming, and resource intensive [47] • Real-time video interactions with digital AI avatars require high computing power and bandwidth connectivity [48] • Risk of perceived uncanniness due to increased realism, posing ethical concerns [25,49]

XSL•FO
RenderX

AI[a]-based CA type (with examples)	Features and functionality	Interaction mode	Advantages	Challenges
Humanoid (or anthropomorphic or social) robots (eg, Sophia and Ameca)	• Designed to mimic human-like characteristics both in behavior and physical appearance, including body structure and autonomous movement [50,51] • Often used for education, entertainment, assistance, and personal care through various sensor channels such as hearing, sight, and touch [52] • Similar to digital avatars, humanoid robots can exhibit human-like communicative behaviors, including social praise, head and torso movements, and nodding, to stimulate more natural conversations [29]	Physical embodiment	• Capability to perform difficult or dangerous tasks, provide companionship, and participate in social interactions, especially in circumstances where human interaction is limited [50,53] • Physical presence enables more distinct and natural interactions compared to chatbots or virtual avatars as HRI[d] differs from traditional HCI[e] by involving both linguistic and physical aspects [29,52]	• Higher development cost relative to other CAs • Lower ease of access for users compared to other CAs [52,54] • Higher risk of ethical and safety concerns

[a]AI: artificial intelligence.

[b]CGI: computer-generated imagery.

[c]NLP: natural language processing.

[d]HRI: human-robot interaction.

[e]HCI: human-computer interaction.

Ethical Concerns of AI Agents

While AI agents have the capability to provide constant health surveillance support, challenges and risks associated with using AI-based CAs in health care remain. These include ethical concerns regarding data collection and interpretability of results, patient safety risks, biases encoded in algorithms, and cybersecurity [55,56]. Furthermore, customers are often reluctant to engage with such AI-based CAs due to many factors such as trust, reliability, learning curve, usability, privacy, and data security that should be addressed in the design, deployment, and use of AI applications [12,29,57-59].

State-of-the-Art Summary

Our study is novel regarding CAs as human-like digital agents for managing NCDs remotely. As AI-based CAs are a relatively new area, limited research has been conducted on applying these emerging technologies in health care. While existing research on CAs as virtual caregivers for NCDs serves as a seminal foundation, such research has other critical limitations. To illustrate, most studies have primarily focused on applications of CAs for mental health conditions, overlooking broader NCDs such as CVDs, metabolic syndrome, or diabetes [60,61]. A recent review [61] that focused on a different set of research questions from ours concluded the following: “A future chatbot could be tailored to metabolic syndrome specifically, targeting all the areas covered in the literature, which would be novel.” Our research looks specifically into this gap.

Furthermore, despite some favorable anecdotal evidence, the effectiveness of CAs in NCD management is seldom explored in large-scale trials, particularly in older adults, who have the highest risk of developing NCDs [62-64]. Some recent previous studies have conducted systematic reviews [27,65,66] or scoping reviews [61,62,67] on CAs in chronic disease management; however, to the best of our knowledge, no review has been conducted yet on the application of AI-based CAs as human-like digital agents for managing NCDs remotely. Such limitations and gaps as highlighted previously present both challenges and opportunities for research to advance our understanding of how CAs can contribute to managing NCDs.

Aim

This scoping review aimed to provide an overview of the existing evidence and research on using assistive humanoid AI-based CAs in health and social care for managing NCDs. Our primary objective was to explore the impact of AI-based CAs, including embodied avatars, as human-like health carers for the self-management of chronic diseases. By examining the current literature on this topic, we hoped to identify key areas for future research and provide insights into how these technologies can be effectively used in health care and personalize NCD management strategies.

Research Objectives

Our research objectives were as follows: (1) to explore the current state of research on the use of AI-based CAs as human-like virtual health carers for managing NCDs, (2) to identify the potential benefits and challenges associated with the use of these technologies in the health care field, (3) to explore the efficacy of AI-based CAs in the remote management of NCDs, (4) to discover the specific target users primarily studied, and (5) to provide recommendations for future research in this area.

XSL•FO
RenderX

Research Questions

Our research questions were as follows: (1) what is the current state of research on using AI-based CAs as human-like health carers for managing NCDs? (2) What are the limitations or challenges associated with the use of these technologies in health care? (3) What are the potential benefits of using AI-based CAs in managing NCDs, and how can they be effectively used to improve health care delivery and reduce health care burden? (4) What is the efficacy of the CAs in the remote care of NCDs? (5) What are the frequently targeted user groups for such virtual agents (eg, specific age groups and individuals with special needs)?

Methods

Search Strategy

We followed the methodological frameworks proposed by Arksey and O'Malley [68] and Levac et al [69]. Initially, research objectives and questions were formulated, followed by a systematic literature search conducted on July 31, 2023. For primary searching, we used 6 electronic databases that are considered relevant to the research focus—Ovid MEDLINE, Embase, PsycINFO, PubMed, CINAHL, and Web of Science—applying the same set of keywords, such as "conversational agents," "artificial intelligence," and "noncommunicable diseases," including their associated synonyms (as shown in Textbox 1). The Boolean operators "*" and "OR" were used to expand and ensure that different word combinations were included. The operator "AND" was used for combining the main search terms to identify articles focusing on AI-based CAs (as health carers) only applicable in the health care field, particularly for managing NCDs or chronic diseases.

Additional studies were identified by hand searching the reference lists of included studies and relevant review articles. Furthermore, a supplementary manual search was conducted to identify specific articles from diverse sources, including ProQuest Central, ResearchGate, ACM Digital Library, and Google Scholar. Some of these articles were identified through pilot hand searching and preselected as they were closely relevant but not retrieved from the selected databases used for primary searching, whereas others were discovered through manual searches targeting specific authors well known in this field. The search strategy underwent refinement following expert recommendations, including insights from a coauthor with digital health expertise, who advised excluding "robots" from the search terms, recognizing it as a distinct field. These methodological steps collectively ensured a comprehensive exploration of the relevant literature.

Textbox 1. Search query example.

Example

- ("conversational agent*" OR "relational agent*" OR "virtual agent*" OR "dialogue agent*" OR "dialogue system*" OR "virtual assistant*" OR "chatbot*" OR "voice assistant*" OR "voicebot*" OR "voice-bot*" or "voice bot*" OR "humanoid *bot*" OR "social *bot*" OR "avatar*" OR "human-like avatar*" OR "anthropomorphic avatar*" OR "digital human*" OR "human digital twin*" OR "virtual human*") AND ("intelligent" OR "artificial intelligence" OR "AI" OR "AI-based") AND ("health" OR "healthcare" OR "caregiver" OR "self-management" OR "self-monitor*" OR "non-communicable disease*" OR "noncommunicable disease*" OR "chronic disease*")

Eligibility Criteria

Our scoping review used comprehensive inclusion and exclusion criteria to ensure the selection of pertinent studies. We did not impose any limitations based on gender or age groups in the selection of articles. The search scope was confined to scholarly articles in English published between January 2010 and July 2023, aligning with the substantial rise in CAs after 2010 [70], notably with the introduction of Apple's Siri in 2011 [71]. Evidently, most of the selected papers were on recent studies conducted within the last 5 years due to the accelerated technological advancements and the latest evolution of AI. Specifically, we focused on empirical studies exploring CAs applied exclusively to human interaction within the health care context, emphasizing their role in the remote management of NCDs, ideally within home environments.

The exclusion criteria comprised the exclusion of conference abstracts, posters, reviews, protocols, position papers or viewpoints, and certain types of studies such as those involving noninteractive robotic devices. We also excluded studies related to CAs designed for medical education and non–patient-centered applications in hospital settings and those addressing specific health care domains such as surgery, dentistry, pregnancy or maternity, addiction or substance use disorders, and communicable diseases. Furthermore, we excluded studies centered solely on medical history data storage, telephone monitoring, data set construction methods, or user evaluations of commercial CAs for health care without their practical applications in the remote management of NCDs. Our criteria aimed to streamline the focus on patient-centered CAs contributing to the self-management and remote monitoring of NCDs, eliminating studies with a primary emphasis on clinical interviews, disease prediction, or decision support without a social interaction element.

Screening and Selection

In total, 2 of the authors independently searched each database. Titles and abstracts were screened for inclusion according to the aforementioned criteria, followed by an exclusion of duplicates, unrelated studies, and articles that could not be retrieved. The abstract screening yielded 264 articles eligible for full-text screening, of which 70 (26.5%) were review papers comparing different types of CAs in the health care field and 156 (59.1%) were empirical studies that explored different types of CAs used in the prevention, treatment, or rehabilitation of chronic diseases involving consumers, caregivers, and health care professionals.

Subsequently, the authors screened full texts of the remaining articles independently, and 40 full-text articles that met the inclusion criteria were selected for review. In addition, 10 [10]

XSL•FO
RenderX

specific articles were obtained from hand searches of other sources (Google Scholar, ProQuest, ACM Digital Library, and ResearchGate), of which 3 (30%) [3] relevant ones were selected upon full-text screening. It was an iterative process; any discrepancies were discussed among the authors. The search and selection process is illustrated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the search and screening process. AI: artificial intelligence; CA: conversational agent; NCD: noncommunicable disease.

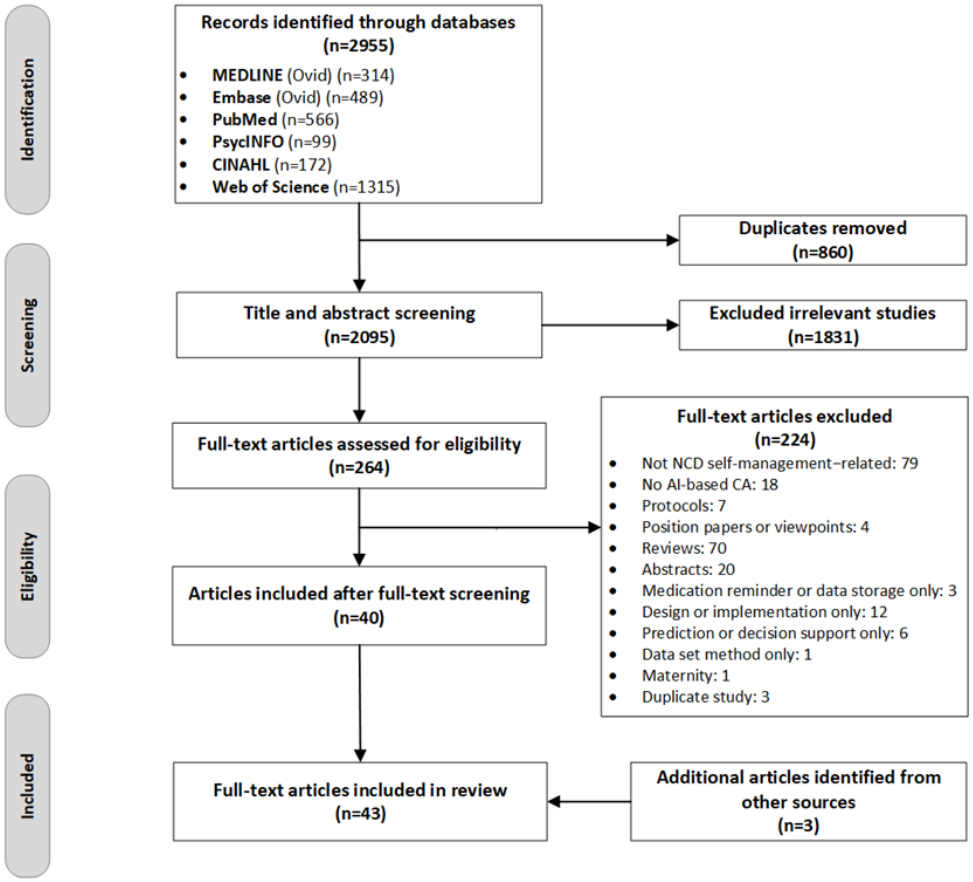

Data Collation and Reporting

An Excel (Microsoft Corp) spreadsheet was initially created to aid the screening and selection process. Following the screening process, a total of 43 articles were eventually selected for synthesis. Quantitative and qualitative data from the included studies were extracted and summarized in a tabular format, including information such as intervention, type of CA, target population, number of participants with their age, methods, study duration, location, measures and outcomes, and limitations. The relevant extracted information was collated and summarized using the narrative synthesis approach, which was deemed appropriate for capturing the breadth of evidence in scoping reviews, identifying themes aligned with the research questions as well as patterns observed across the included studies and relevant reviews. The results were reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Given the divergent methodologies of the included studies, predominantly using mixed methods and qualitative approaches, no quantitative synthesis or meta-analysis was conducted.

Results

Characteristics of the Included Studies

Most of the included studies (22/43, 51%) were feasibility, acceptability, and usability studies (Table 2). These studies mostly found positive results in terms of feasibility and usability of the proposed CAs. Some studies (4/43, 9%) used the System Usability Scale questionnaire to measure the system's usability score and predominantly found high usability, with System Usability Scale scores of >70 [72-75]. Alternatively, multiple studies (10/43, 23%) used the subjective ratings of most of the participants to evaluate the feasibility and acceptability of CAs by applying indicators such as perceived usefulness [76-79], ease of use [78,80], user satisfaction [80,81], engagement rate [80,82,83], perceived closeness [73], and Net Promoter Score [84]. However, 2% (1/43) of the studies reported negative usability and acceptability outcomes, indicating unsatisfactory reliability, usability, and goal structure, which hindered health care professionals' acceptance and trust [85].

XSL•FO
RenderX

Table 2. Characteristics of the included studies.

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
Watson et al [86], 2012	• Behavior change intervention for obesity self-management • Duration: 12 weeks (divided into four 3-week periods—period 1, period 2, period 3, and period 4)	• 2D-animated human-like avatar (rule-based AI[b]) • Platform: desktop and laptop computers (web based)	• Adults with overweight or obesity • N=62 • Participants aged 20-55 years (inclusive)	• Pretest-posttest observational study (quantitative)—2-arm RCT[c] • Location: United States	• Step count percentage: • No significant percentage change in step count between intervention and control arms from start to end (2.9% vs −12.8%, respectively; *P*=.07) • Significant difference between percentage change in step count across all study periods in the intervention vs control arms (*P*=.02) • Secondary outcomes: no significant changes in secondary outcomes—BMI, 7-day physical activity recall, physical activity stage of change, self-efficacy and exercise benefits and barriers, and program satisfaction (eg, mean decrease in BMI of 0.04 and 0.25 in control and intervention groups, respectively; *P*=.44) • Average step count: significant decrease in the average mean step count for the control group from period 1 (7174) to period 4 (6149; *P*=.01) but no significant change for the intervention group from period 1 (6943) to period 4 (7024; *P*=.85) • Mean activity level: significant percentage change in mean activity levels between period 1 and period 3 (*P*=.02) but no significant change between period 1 and period 2 (*P*=.12) and between period 1 and period 4 (*P*=.07) • 58% of the intervention participants agreed that the virtual coach influenced their increased activity	• Participants were primarily White, college-educated women, limiting generalizability of findings to a wider patient population with overweight or obesity • Lack of baseline step count data for participants, although survey results showed no substantial baseline activity level differences • Initial observed increase in step count after enrollment likely reflects a change from baseline, but without baseline data, this cannot be confirmed

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
Kimani et al [80], 2016	• Lifestyle intervention and patient counseling support for AF[d] self-management • Duration: not identified	• Animated human-like avatar (network-based AI with XML scripting language) • Platform: mobile devices or smartphones	• Patients with AF • N=16 (5 female and 11 male) • Participants aged ≥18 years (20-58 years)	• Feasibility assessment—mixed methods pilot study with self-report scale measures and a semistructured interview • Location: not identified	• High overall user satisfaction with the agent (mean 3.45 on a 4-point scale) and the AliveCor heart rhythm monitor (mean 3.54) • High ratings for ease of use (mean 3.54) • Interaction duration: participants reported a 7- to 10-minute–long interaction with the agent, whereas older participants reported longer interactions due to content relevancy • Agent's user satisfaction correlated with the participants' satisfaction with the use of the AliveCor heart rhythm monitor • Acceptable feasibility of delivering AF counseling via a smartphone-based humanoid avatar as a virtual agent	• Small sample size and limited patient diversity • Short-term study duration and no long-term use evaluation • No control group for comparison to provide additional insights into the effectiveness of the virtual agent • No assessment of the virtual agent's impact on patient outcomes or behavior change
Shamekhi et al [87], 2017	• Behavioral intervention for self-management of stress and depression • Duration: 9-21 weeks	• 3D-animated human-like avatar (rule-based AI) • Platform: touchscreen tablets	• Adults with chronic pain and depression • N=154 • Participants aged≥18 years	• Mixed methods observational study—2-arm RCT • Location: United States	• Impact on patient performance and satisfaction: significant positive stress management behaviors among the CA intervention participants compared to the control group after 9 weeks (t136=3.74; *P*<.001) • Avatar (Gabby) was found to be very useful for group visits, allowing for review of class lessons and detailed information at any time and location • Most participants found the meditation, yoga, and mindfulness sessions provided by Gabby to be very "useful" and "relaxing" • Participants were very likely to recommend Gabby to others	• Relatively small sample size from health centers in the Boston area, limiting generalizability of findings to other populations • Inability to isolate the effects of face-to-face weekly group visits from those of the home-based CA • Potential usability issues, especially for less computer-literate users; may affect Gabby's user satisfaction
Bickmore et al [10], 2018	• AF self-management through symptom monitoring and patient counseling • Duration: 30 days	• 3D-animated human-like avatar (network-based scripted AI) • Platform: mobile devices or smartphones	• Older adult patients with AF • N=120 • Participants aged≥60 years (mean 72.1, SD 9.10 years)	• Mixed methods pretest-posttest (observational) study—quasi-experimental demonstration of CA and 2-arm RCT for usability and effectiveness evaluation • Location: United States	• Change in AFEQT[e] score: significantly higher AFEQT score among the intervention group participants after 30 days compared to those in the control group (*P*<.05) • Over 89.7% of participants rated the CA positively, whereas 4.4% rated it negatively and 5.9% had mixed opinions • Participants praised the personalized interaction with the avatar	• Relatively small convenience samples with limited participant details and a singular focus on a specific chronic condition (AF) may impact generalizability • Lack of long-term, objectively measured health outcomes • Insufficient emphasis on the complexity of self-care management regimens, especially for patients with low health literacy

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
Cheng et al [58], 2018	• Personalized patient education and medication adherence intervention for improved self-management of T2DM[f] • Duration: 1 month	• AI voice assistant and voicebot (ML[g] based) • Platform: Google Home devices (API.AI platform)	• Older adult patients with T2DM • N=10 • Participant age not identified	• Feature-based comparison between the proposed voicebot and other similar mHealth[h] apps and usability evaluation—qualitative assessment • Location: not identified	• AADE[i] requirement evaluation: the feature-based comparison indicates that most similar mHealth apps fail to meet the AADE requirements for effective T2DM self-management compared to the proposed voicebot • Practical usability evaluation: • There were more satisfied users than unsatisfied ones, primarily due to the voicebot's speaker functionality and natural conversation flow • Unsatisfied users reported difficulty learning commands and limited answer choices • 80% of older adult participants would prefer using Healthy Coping on Google Home over a smartphone	• Voicebot's limitations • Narration speed concern: older adult users may struggle with application commands due to narration speed; a speed setting could allow users to adjust the pace • Accessibility limitations: Healthy Coping's voice interface may not fully support users with hearing or speech disabilities; integrating devices such as Bluetooth-enabled hearing aids could improve accessibility
Suganuma et al [88], 2018	• Internet-based cognitive behavioral intervention for self-monitoring of mental health • Duration: 1 month	• AI (rule-based) chatbot via SABORI (mental health) app • Platform: web based	• Adults with psychological distress and mental health problems • N=454 (intervention group: n=191; control group: n=263) • Participants' age: mean 38.07 (SD 10.75) years	• Feasibility and acceptability (pilot) study—nonrandomized prospective study with quantitative pretest-posttest intervention questionnaire • Location: Japan		• Nonrandomized comparison between control and experimental groups • Short-term study duration (1 month) for assessing behavioral activation effects • Recruitment of healthy participants in control group may have influenced outcomes, resulting in lack of effect for behavioral activation factors of avoidance and rumination

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
					• Japanese version of the WHO-5[j]: a 1-factor scale to measure positive mental health based on physical aspects using a 5-item, 6-point Likert scale • On the variable of positive mental health, a significant main effect (*P*=.02) and interaction for time (*P*=.02), including a significant simple main effect for the intervention group (*P*=.002) • K10[k]: a 1-factor scale to measure negative mental health based on physical aspects using a 10-item, 5-point Likert scale • On the variable of negative mental health, a significant main effect (*P*=.02) and interaction for time (*P*=.005), including a significant simple main effect for the intervention group (*P*=.001) • BADS[l]: a scale to measure 4 factors-"Activation" (BADS-AC[m]), "Avoidance/Rumination" (BADS-AR[n]), "Work/School Impairment" (BADS-WS[o]), and "Social Impairment" (BADS-SI[p]) • 2-way ANOVA test showed a significant trend for behavioral activation, indicating the possibility for a certain degree of effectiveness • BADS-AC: a significant main effect for time (*P*=.10), including a significant simple main effect trend in the intervention group (*P*=.06) • BADS-AR: no significant main or interaction effect	
Hussain and Athula [89], 2018		• AI-ML chatbot (VDMS[q]) • Platform: computer or mobile devices (web based)	• Patients with diabetes and their carers seeking diabetes-related information • N=10 • Participants aged 20-50 years		• Evaluation metrics (VDMS vs other sources): • Correct replies: VDMS=65%; search engine=80% • Satisfactory user satisfaction level of VDMS (very close to that of the search engine) due to VDMS's timely response and correct answers' quality and clarity	

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• To provide patient education for improved self-management of diabetes • Duration: not identified			• Quantitative and qualitative study—performance evaluation of VDMS compared to other information sources (eg, search and websites) • Location: not identified		• Small sample size (10 participants) may restrict diversity and evaluation scope • VDMS chatbot's implementation is incomplete and lacks full testing • Conversation is mostly controlled by the chatbot, limiting user input • Extracted data from Wikipedia may lack reliability as no published work has explored integrating Wikipedia knowledge with a diabetes management chatbot
Inkster et al [90], 2018	• To deliver positive psychology and mental well-being techniques for improved self-management of depression and anxiety • Duration: 8 weeks	• Commercial CA: text-based AI chatbot (Wysa) • Platform: mobile devices or smartphones	• Individuals with self-reported symptoms of depression • N=not identified • Participants' age not identified	• Quasi-experimental (pretest-posttest) design—mixed methods study using the Wysa app's inbuilt assessment questionnaire (for effectiveness evaluation) • Location: global (not specified)	• Impact (pre-post) analysis outcomes: • PHQ-9[r] score (within groups) measured using a Wilcoxon signed rank test: significant reduction in PHQ-9 score among both high users (P<.001) and low users (P=.01), indicating an improvement in depression symptoms from before to after the intervention • High users' group showed significantly higher average mood improvement compared with the low users' group (Mann-Whitney P=.03, with a moderate effect size of 0.63) • User engagement and feedback: 73.6% of users provided at least one response to the in-app feedback questions (indicating high user engagement), and 60.9% of them reported feeling better after app use	• Small sample size restricts investigation of user reactions to app design elements • Small and unbalanced comparison group sizes undermine the findings' generalizability • Limited detailed feedback on users' app experience hinders qualitative analysis • Lack of a randomized controlled environment may introduce biases • Absence of users' previous health information hampers comprehensive understanding • Quasi-experimental design used, slightly lower in design quality compared to interrupted time-series designs
Neerincx et al [85], 2019	• Behavioral intervention and patient education for self-management of T1DM[t] • Duration: 3 months		• Children with T1DM • N=49 • Participants aged 7-14 years	• Qualitative study—prototype design and usability evaluation using an iterative, incremental development process • Location: Italy and the Netherlands		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
		• Humanoid robot (PAL[u]) and its robotlike digital avatar version for home (MyPal; rule based+ML with cloud computing) • Platform (avatar): tablets			• Unsatisfactory reliability, usability, and goal structure of PAL system, which hindered the acceptance and trust of health care professionals • Improvement in children's diabetes knowledge when using the PAL system and increased motivation in performing diabetes-related activities due to enjoyable interaction with the PAL robot and avatar • Improvement suggestions: personalization enhancement required to establish patients' adherence • Innovative PAL functions (eg, experience-sharing function) were identified and tested with positive results • Reduced bonding effect with higher perceived similarity between the robot and its digital avatar • Most children stopped using the avatar (MyPAL app) some weeks after the study started	• Small sample size may affect the findings' generalizability • Limited study duration may not capture the PAL system's long-term effects on T1DM management • Comparison between physical robot and avatar function may introduce biases in user preference assessment • Insufficient detailed information on AI algorithms or ML techniques used restricts insights into the PAL system's technical aspects
Easton et al [91], 2019	• Self-management for people with comorbid long-term conditions and mental health problems • Duration: not identified	• AI chatbot with an animated human-like 2D avatar (Avachat) • Platform: computers, tablets, smartphones, and televisions	• Older adults with comorbid long-term conditions and mental health problems • N=10 • Participants aged ≥55 years (56-86 years)	• Co-design and acceptability testing involving stakeholders—qualitative study with snowball sampling and workshops for initial user requirement gathering of prototype and user feedback • Location: United Kingdom	• Acceptability and feasibility: • Patients found Avachat to be helpful, informative, and easy to use • Health care professionals are optimistic about the avatar's potential to improve patient outcomes and reduce the health care burden • Improvement suggestions: enhance personalization of contents and ensure accessibility for users with visual and hearing impairments	• Small sample of White, British, medically stable, and regional participants recruited may limit findings • No current mental health problems reported by participants, although past instances of low mood or worry were mentioned
Chaix et al [92], 2019			• Patients with breast cancer and their relatives • N=4737 • Participants' age not identified	• Quantitative study using user-chatbot conversational data • Location: not identified		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• To provide personalized patient education for improved quality of life and medication adherence in patients with breast cancer • Duration: 1 year	• Commercial CA—AI chatbot (Vik) • Platform: mobile devices or smartphones (iOS or Android) and web browser via Messenger app			• Users' interactivity level: average of 132,970 messages exchanged per month, with 147 total average interactions per question (for open-ended questions), resulting in 2.7 interactions per person per question • Overall user satisfaction rate: 93.9% (900/958), and 88% (843/958) found Vik helpful in following treatment effectively • Average medication adherence rate of patients improved by >20% (*P*=.04) after using Vik for 4 weeks	• Absence of a control group for comparison • Lack of long-term evaluation of the chatbot's impact on clinical outcomes • Reliance on users' self-reported data, introducing potential bias • No investigation of the viewpoints of health care providers or other stakeholders in patient care • Potential technical issues or limitations of the chatbot not addressed
Stephens et al [76], 2019	• Behavioral intervention for constant management of obesity and prediabetes symptoms • Duration: 6 months	• Commercial CA—AI behavioral coaching chatbot (Tess) • Platform: mobile devices or smartphones via messaging apps (eg, WhatsApp and Messenger); Google Home or Amazon Alexa (for voice conversations)	• Adolescent pediatric patients with obesity and prediabetes symptoms • N=23 • Participants aged 9-19 years	• Feasibility study—mixed methods approach with qualitative interviews, usability testing, and quantitative surveys (SUS[v] questionnaire) • Location: United States	• Positive progress toward goals was reported by the participants 81% of the time • 4123 messages were exchanged, with patients rating the chatbot's usefulness 96% of the time, indicating its high perceived usefulness and feasibility among adolescents	• Small sample size, causing limited generalizability • Gradual program adjustments may introduce variability and potential inconsistency • Lack of an experimental design to control for factors • Inability to ensure detection of a treatment effect
Balsa et al [72], 2020	• Behavioral intervention for self-management of T2DM • Duration: 8-10 days	• Animated human-like avatar (Vitória) in a 3D environment (rule-based AI) • Platform: mobile devices or smartphones	• Older adults with T2DM • N=20 • Participants aged ≥65 years (67-80 years)	• Usability study—qualitative assessment • Location: Portugal	• Usability evaluation for T2DM medication adherence and lifestyle improvement • Usability (SUS) score: 73.75 (SD 13.31; indicates significantly high usability of Vitória)	• Small sample size and skewed participant sample toward expert technology users • Reliance on previous experience and available resources for sample size estimation • Criticism of questionnaires for yielding only overall measures without addressing specific concerns • Potential bias in field-testing compared to laboratory setting for usability tests

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
Gong et al [24], 2020	• Behavioral intervention with personalized support and motivational coaching for remote self-management of T2DM • Duration: 12 months	• Human-like (text+voice enabled) avatar (Laura; BCT[w]-based prescripted AI) • Platform: mobile devices or smartphones	• Adults with T2DM • N=187 • Participants aged ≥18 years (mean 57, SD 10 years)	• Pretest-posttest (quantitative) observational study—2-arm, open-label RCT for adoption, usefulness, and effectiveness evaluation • Location: Australia	• Program adoption and use: • Number of valid chats with Laura completed per person: 18.4 (SD 15.0; range 1-53) • Total duration of valid chats per person: 242.7 (SD 212.3; range 0-1050) minutes • Number of glucose level uploads per person: 181.8 (SD 192.1; range 1-966) • Number of clinical alerts: total=297; average per month=13.7 (SD 8.8) • Number of technical alerts: total=179; average per month=8.3 (SD 6.5) • Number of posts on the web-based discussion forum: total=19; average per month=1.1 • Program effectiveness (in terms of coprimary and secondary outcomes)—coprimary outcomes: • Statistically significant between-arm difference at 12 months in the mean change in HRQoL[x] (AQoL-8D[y] utility value: 0.04, 95% CI 0.00-0.07; *P*=.04) • Reduction in HbA_{1c}[z] levels during the trial but no statistically significant between-arm difference at 6 months (0.06, 95% CI −0.35 to 0.47; *P*=.78) or 12 months (−0.04, 95% CI −0.45 to 0.36; *P*=.84) • Significant improvement in HRQoL from baseline to 12 months (mean estimated change in AQoL-8D score: 0.04, 95% CI 0.01-0.06; *P*=.007) • Increase in the score of the physical health and mental health subscales compared with baseline • Secondary outcomes: • Significant between-arm difference in the mean change in the HADS[aa] anxiety score at 6 months (–0.89, 95% CI –1.74 to –0.04; *P*=.04) but not at 12 months or for other secondary outcomes reported	• Small sample size may restrict the generalizability of the findings, and subgroup analyses require cautious interpretation • Absence of blinding of participants and their GPs[ab] to the study arm allocation might lead to potential self-report bias and Hawthorne effects • Control arm participants showed a higher rate of completed assessments, possibly due to their interest in program access or higher attrition in the intervention arm • Subgroup analyses were underpowered, with multiple testing increasing the risk of false positives
Issom et al [77], 2020						

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Health behavioral intervention for self-management of SCD[ac] • Duration: not identified	• AI chatbot • Platform: mobile devices or smartphones	• People with SCD • N=19 • Participants' age not specified	• Preliminary feasibility study—quantitative posttest survey • Location: France	• 88% of participants rated the following question—"The chatbot contains all the information I need"—with at least 3/4, and its total score was 54/68 • 58% rated the following question—"The chatbot encouraged me to be more active in order to improve my condition"—with at least 3/4, and its total score was 51/68 • Results indicate high perceived usefulness of the chatbot in promoting knowledge and motivation for improved self-care practices	• Small sample size: only 17 participants completed the evaluation, potentially limiting generalizability • 2 withdrawals due to smartphone issues may have impacted data collection and results • Reliance on self-reported data in the posttest survey could introduce bias or inaccuracies • Limited scope: evaluation focused solely on perceived usefulness of the chatbot, lacking assessment of long-term impacts on self-care or health outcomes • Lack of comparison with other support forms or interventions hinders assessment of relative effectiveness
Anastasiadou et al [93], 2020	• To offer continuous health education and interaction in English, Spanish, and Bulgarian for self-management of diabetes • Duration: 6 months	• Multilingual AI-ML chatbot (EVA[ad]) • Platform: mobile devices or smartphones	• People diagnosed with diabetes • N=not identified • Participants' age not identified	• Qualitative pilot study—validation and acceptance evaluation by integrating EVA into an mHealth app (CHRODIS PLUS Joint Action) to collect data on user queries and responses provided based on the educational content • Location: Greece	• Users sent a total of 940 unique messages to EVA • Users' common questions related to diabetes varied based on EVA's language: • English: self-measurement and understanding what diabetes is • Spanish: glucose self-monitoring • Bulgarian: insulin and high blood pressure • A comprehensive analysis of the effectiveness of EVA in improving diabetes management or patient outcomes was not provided	• Lack of detailed user demographic information may affect generalizability • Study duration limited to 6 months, possibly overlooking long-term user interactions and feedback • Absence of specified measures to address potential biases in user interactions or data collection • No discussion about potential technical limitations ad challenges during EVA system development and testing
Roca et al [94], 2021		• AI-ML–based chatbot • Platform: mobile devices or smartphones	• People with comorbid T2DM • N=13 • Participants aged ≥18 years	• Quantitative pilot trial—pretest-posttest observational study (1 arm) • Location: Spain		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Medication adherence intervention for improved self-management of T2DM and depressive disorder • Duration: 9 months				• Medical outcomes: • Significant improvement in the average HbA1c level and PHQ-9 scores from 7.6 (SD 0.7) to 7.3 (SD 0.8) and from 13.2 (SD 6.0) to 8.6 (3.6), respectively • Reduction in the number of physical medical consultations per month from 2.0 (SD 2.6) to 1.3 (SD 1.5) in 30.8% of the patients • Health care professionals participating in the study found the chatbot useful in improving medication adherence • Chatbot feedback: • Participants who used the chatbot daily found it useful in fulfilling their medication reminder needs • The chatbot's language and vocabulary were appropriate and easy to understand • 38% of participants had difficulty learning to use the chatbot, and 15.4% reported the chatbot's inability to understand the users' instructions • Almost 70% of the patients (9/13) expressed willingness to continue using the chatbot after the study	• Small sample size (13 participants) may limit generalizability of the study findings • Some patients did not update medication information after the initial configuration, leading to reminders being stopped • Patients require digital literacy or assistance for configuring the messaging platform • Reminder sounds may be accidentally disabled by patients, affecting the use of the chatbot
Yao et al [95], 2021	• Remote monitoring of patients after stroke and automation of suspected stroke screening • Duration: not identified	• Hyperrealistic human-like 3D avatar (iLAMA; ML based with computer vision algorithms) • Platform: mobile devices or smartphones and tablet	• Older adult patients after stroke • N=140 (videos) • Participants aged ≥60 years	• Quantitative beta testing of the prototype on 140 videos with volunteers via email • Location: not identified	• The app was able to recognize body parts and extract 68 facial landmarks from the facial videos • The app could provide accurate stroke screening results for neurologists and stroke specialists to review • No major technical issues or problems with the app during beta testing were reported • iLAMA has the potential to improve stroke assessment and care, particularly in areas with limited access to stroke specialists	• Participants were regular volunteers (friends and family) rather than patients after stroke • Beta testing involved a relatively small sample size of 140 videos • The app may struggle to differentiate minor mistakes from actual stroke signs, especially without supervision • Screening process takes approximately 5 minutes, which some users may find burdensome
Krishnakumar et al [82], 2021			• Adults with T2DM possessing Android smartphones • N=102 • Participants aged ≥18 years			

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Lifestyle intervention and self-monitoring of diet, exercise, weight, and blood glucose for T2DM management • Duration: 16 weeks	• AI chatbot via mHealth therapeutic app (Wellthy Care) • Platform: Android smartphones		• Retrospective study design—pretest-posttest (quantitative) observational study for effectiveness evaluation • Location: India	• Primary outcomes: • Statistically significant reduction in HbA1c levels from baseline to 16 weeks (P<.001) • Secondary outcomes: • Mean FBG[ae] and PPBG[af] levels decreased significantly from baseline to 16 weeks (P<.001) • Mean BMI and weight decreased significantly from baseline to 16 weeks (P<.001) • Average duration of interactions was 106 minutes with the app (87 participants; 95% CI 65-147 minutes) and 88 minutes with the AI chatbot (102 participants; 95% CI 66-110), indicating a significant positive association with program engagement	• Short program duration • No control group as real-world data were retrospectively analyzed • Selection bias from multiple selection approaches (physician recommended and voluntary approach) of participants • Reliance on self-reported disease biomarkers • Loss of data during follow-up • Variation in the number of male and female participants, introducing potential bias
Egede et al [96], 2021	• To improve mental health (depression and anxiety) assessments and remote counseling via VH[ag] mediation • Duration: not identified	• Human-like avatar (Greta; Wizard-of-Oz AI system using prescripted text and human-controlled responses) • Platform: not identified	• Adults with anxiety and depression • N=56 • Participants aged ≥18 years (18-45 years)	• Mixed methods comparison and UX[ah] evaluation study: • Behavioral expressiveness comparison between VH system and text-only system users • UX assessment and task ratings using UEQ[ai] and TAM[aj] questionnaires • Location: Japan		• Small sample size may limit depth of analysis, particularly in comparing depression severity classes • Combination of moderate and moderately severe depression groups may restrict findings; a 4-class comparison could yield more valuable data • Focus on VH mediation may have overshadowed other factors influencing user engagement and task performance • Long-term effects of VH-mediated tasks on user engagement and mental health outcomes were unexplored, limiting insights into intervention sustainability

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
					• Stronger visual cues produced by VH avatar guidance compared to non-VH (*P*=.04; 0.036 for 2 facial action units indicating a statistically significant difference in the frequency of activation of these facial action units between the VH and non-VH systems; *P*=.005 and 0.0032 for 2 head pose descriptors indicating a statistically significant difference in head pose characteristics between the VH and non-VH systems) • VH users make wider and broader head movements (indicating higher behavior expressiveness) compared to text-only users (*P*=.005 and 0.0032 for 2 head pose descriptors) • VH mediation improves users' inclination toward tasks, as indicated by the UEQ and TAM questionnaires • Significant differences (*P*=.05) found between the system modes in task ratings for facial expression mimicking (*P*=.04), emotion recall (*P*=.01), thematic apperception (*P*=.03), and mindfulness (*P*=.04) tasks, with VH mode's mean ratings significantly higher than those for the text-only mode except for the emotion recall task, where VH's ratings were lower • Groups with moderate depression exhibit more visible activity compared to groups with less depression (*P*=.04 and 0.004 for 2 facial action units)	
Romanovsky et al [97], 2021		• AI chatbot (Elomia; NLP[ak]+ML based) • Platform: not identified	• Young adults with depression, anxiety, and low mood • N=412 (202 female and 210 male) • Participants aged ≥18 years (19-23 years)		• Significant reduction in the high tendency toward depression (up to 28%) • Significant reduction in the high tendency toward anxiety (up to 31%) • Significant reduction in the high tendency toward negative effects (up to 15%) through the regular use of Elomia	

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Cognitive behavioral intervention for self-management of mental health (anxiety, depression, and low mood) • Duration: 4 weeks			• Pretest-posttest observational study—effectiveness evaluation using psychological quantitative research methods through a controlled experiment in 3 stages: formation of experimental and control samples, baseline testing, and final testing • Location: Ukraine		• Relatively small sample size and focus on specific student age group may limit generalizability of study findings to other populations • Short 4-week study duration may be insufficient to assess Elomia's long-term effectiveness • No follow-up assessment may restrict evaluation of Elomia's sustainability • Absence of qualitative analysis of user feedback may limit UX insights • Lack of comparison with other web-based psychological services or traditional face-to-face therapy may produce a less comprehensive evaluation of the chatbot's effectiveness
Anan et al [98], 2021	• Lifestyle intervention to support adherence to exercises for self-management and improvement in musculoskeletal symptoms • Duration: 12 weeks	• AI chatbot • Platform: mobile devices or smartphones via a messaging app (LINE)	• Employees with musculoskeletal symptoms (eg, neck and shoulder stiffness and low back pain) and smartphone users • N=121 (intervention group: n=61; control group: n=60) • Participants' age not specified	• Pretest-posttest (quantitative) observational study—2-arm RCT following CONSORT-EHEALTH[al] guidelines • Location: Japan		• Limited generalizability due to single-company study setting • No information collected on causes, diagnosis, and treatment status of neck and shoulder pain and low back pain • Pain improvement observed in the intervention group may be influenced by treatment changes and occupational factors, not only by the intervention. • Lack of long-term follow-up to assess intervention sustainability

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
					• Primary outcomes: • Statistically significant improvement (*P*<.001) in the average pain level (on a scale of 1 [highest] to 5 [lowest]) of the neck and shoulder stiffness or pain or low back pain in the intervention group (average pain level=3.0, SD 1.1), compared to the control group (average pain level=4.0, SD 0.8) • Proportion of participants with severe symptoms significantly decreased from 77% (37/48) to 33% (16/48) in the intervention group, whereas the decrease in the control group was from 76% (33/46) to 67% (31/46) • Significant improvements in the severity of the neck and shoulder pain and stiffness and low back pain in the intervention group compared to the control group (OR[am] 6.36, 95% CI 2.57-15.73; *P*<.001) • Subjective assessment: 75% in the intervention group showed symptom improvement, whereas only 3% in the control group showed improvement (*P*<.001) • Secondary outcomes: • The AI-based health intervention was effective in improving the EQ-5D-5L score and the RMDQ[an] score in the intervention group	
Zisis et al [99], 2021	• To provide patient education and support self-management of acute decompensated HFao • Duration: 12 weeks	• AI-based avatar (HF-Coach; NLP+ML based) • Platform: mobile devices or smartphones and tablets	• Patients with chronic acute decompensated HF • N=200 • Participants' age: mean 55 years	• RCT—quantitative study using patient medical records and oral interviews • Location: Australia	• No significant differences between the intervention and control groups in mood (GAD-7[s] and PHQ-9), cognition (MoCA[ap]), HF knowledge (DHFKS[aq]), or HRQoL • Enrolled participants had better self-care behavior • Barriers to implementation identified: lack of interest, inadequate technological and computer literacy, language limitations, anxiety caused by app questions, data retrieval issues, and disabilities (eg, visual, hearing, and cognitive impairments)	• Small enrollment numbers • Low patient engagement • High dropout from the HF app • Findings based on experiences from a single trial

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
Stara et al [73], 2021	• To support self-management of dementia in the home environment • Duration: 4 weeks	• 3D-animated human-like AI-based avatar (Anne) • Platform: mobile device or smartphones • Microsoft Surface Pro tablet	• Older adults with dementia and their caregivers • N=20 (30% male and 70% female) • Participants aged ≥65 years (mean 75.5, SD 4.2 years)	• Usability and acceptability evaluation study—mixed methods approach • Location: Italy	• Usability (SUS) score: 67.1 among older adults and 71.4 among caregivers, indicating acceptable usability • No significant change in quality of life in older adults with cognitive impairment (QoL-AD[ar]) before or after the study (mean change 0.4, SD 4.6) • 42% (8/20) of older adults' perceived closeness with Anne as feeling some overlaps, while 26% (6/20) reported feeling no overlap, and 11% (2/20) felt strong, equal or little overlap	• Small sample size and Italian national context and culture may limit result generalization • Short 4-week study duration may be insufficient to significantly evaluate acceptability and usability • Technical discomfort of participants with automatic speech recognition
Nguyen et al [74], 2021	• Self-management of diabetes through patient education on diabetes care, blood glucose monitoring, and managing diabetes complications • Duration: 8 weeks (52 days)	• Multilingual (text+voice enabled) AI bot (AMANDA; ML based using deep learning) • Platform: mobile devices or smartphones	• Patients with diabetes who need self-care • N=20 (nurses and clinicians) • Participants' age not identified	• Usability and quality evaluation of AMANDA—a quantitative survey questionnaire filled out by 20 judges after listening to the real and TTS[as]-generated audio samples • Location: Singapore	• Naturalness: 4.07 • Accent uniqueness: 3.98 • Information clarity: 3.88 • SUS score: 80.625 (above the average score of 68); 70% of the participants gave a score of ≥80 (indicating high usability of the interface)	• Small sample size and language restriction (availability in English only) may limit result generalization • Laboratory setting evaluation may not reflect real-world use • Impact of the CA on users' health outcomes (eg, blood glucose levels and medication adherence) and their emotional well-being or quality of life was not considered in the evaluation
Apergi et al [100], 2021	• Self-care and communication improvement between patients and health care providers • Duration: 90 days	• Commercial voicebot (Alexa) and a voice-enabled, animated avatar • Platform: • Alexa+: smart speaker device (Amazon) • Avatar: tablets	• Patients with HF • N=55 (Alexa+: n=28; avatar: n=27) • Participants' age not identified	• Quantitative pilot comparison study—demographic and technology survey and daily questionnaires • Location: not identified		• Small sample size and focus on specific patient populations may limit generalizability of the findings • Potential bias due to exclusion of possible important variables or unaccounted confounding factors in the model • Lack of normalization of data for control variables • Limited exploration of reasons for differences in technology engagement among Black patients

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
					• Positive correlation between patients' age and the technology use (coefficient=1.19; *P*=.004) • No statistically significant difference in engagement levels between the avatar and Alexa+ user groups • Decrease in use over time for both technologies, with a sharper decrease observed for Alexa+ participants • Black patients with similar characteristics used the technology 21 fewer times on average compared to non-Black patients (coefficient=–15.96; *P*=.08) • Results indicate that technology design may need to be better tailored for Black patients with HF	
Kataoka et al [81], 2021	• Personalized patient education for self-management of lung cancer symptoms • Duration: 1 month	• AI chatbot (rule-based AI using Google Cloud's Dialogflow and predetermined keywords) • Platform: mobile devices or smartphones and web browser (via LINE app)	• Patients with lung cancer and their caregivers • N=12 (11 patients and 1 caregiver) • Participants' age not identified	• Feasibility and usability study—sequential mixed methods approach through a web-based qualitative survey questionnaire and quantitative alpha and beta testing of chatbot with stakeholders • Location: Japan	• The chatbot was found to be feasible but inadequately used • Mean user satisfaction score was 2.7/5, indicating low user satisfaction • Chatbot was able to deliver appropriate responses to most FAQs[at] and also identify areas where additional responses were required • 82 categories of FAQs and formulated responses to these FAQs were identified, which were used to develop the chatbot • Potential of chatbot to improve patient knowledge of symptom management was demonstrated	• Small sample size and single-hospital setting may limit applicability of study findings • 8 questions did not match well with responses in phase 5, leading to patient dissatisfaction • Questions for nonexistent categories remained unmatched, indicating the need to add educational categories and responses through further discussion
Rathnayaka et al [101], 2022	• Behavioral intervention for remote management of mental health issues • Duration: 8 weeks	• AI chatbot (Bunji; rule based+ML) • Platform: mobile devices or smartphones	• Individuals with mental health issues • N=318 • Participants' age not identified	• Mixed methods pilot study—participatory evaluation through quasi-experimental design and qualitative feedback • Location: Australia	• Participatory evaluation (through 3 experimental pilot studies) • Study 1: positive impact in improving users' mood using Bunji's emotional support features • Study 2: positive impact of personalized conversation on improving the effectiveness of Bunji in providing mental health support to users • Study 3: positive feedback on the features of the chatbot and its effectiveness in providing remote mental health support to users	

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
						• Limited sample size may impact result generalizability • Short study duration (8 weeks) may be inadequate to capture long-term effects or user behaviors • Reliance on users' self-reported data such as mood scores and survey responses could introduce bias • Chatbot's performance could be affected by technical issues such as connectivity or device compatibility • Lack of a control group may limit the ability to attribute changes to the chatbot's intervention
Kannampallil et al [102], 2022	• To provide personalized and accessible mental health care with realistic and cognitively plausible verbal interaction • Duration: not identified	• AI voice-based coach and voicebot (Lumen) • Platform: Amazon Alexa	• Individuals with mild to moderate depression and anxiety seeking behavioral therapy • N=26 • Participants' age: mean 43.9 years	• Mixed methods observational study: • Efficacy evaluation—RCT • Acceptability and usability evaluation with WAI-TECH[au] survey • Location: not identified	• Task load (mental and physical effort required to complete a task): medium workload (higher temporal workload in session 2 [mean Task subscale 5.3, SD 0.9] than in session 1 [mean Task subscale 5.2, SD 0.9]) • Work alliance (collaborative relationship between therapist and client): moderately high, indicating well-aligned sessions with participants' needs, potential goals (session 1-mean Goal subscale 5.0, SD 0.9; session 2- mean Goal subscale 5.1, SD 0.9), and mutual liking between participants and voicebot (session 1-mean Bondscale 4.9, SD 1.0; session 2-mean Bondscale 4.7, SD 1.0) • Participants highlighted lack of personalization, depth, and emotional engagement in the conversations • Overall UX: positive evaluation (values of >0.8) for pragmatic, hedonic, and overall qualities related to UX with Lumen for both sessions	• Small sample size of users in a relatively controlled environment, potentially limiting applicability of findings • Evaluation limited to 2 sessions, not representing the full 8-session PST[av] program • Potential influence of research coordinator and note taker on participant responses and Lumen use • Inability to assess impact of various measures (task load and work alliance) over time due to study constraints • Technological limitations of current AI-based voice technology include challenges in understanding natural language and interpreting emotional cues in voice interactions
Rahmanti et al [78], 2022			• Individuals intending to lose or maintain weight • N=10 (100% female) • Participants aged 24-34 years		• Positive chatbot UX: ease of use; usefulness; and fun attributes such as use of emoticons, stickers, and GIF[aw] images • Negative chatbot UX: slow response time and irrelevant responses	

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Lifestyle intervention for weight management through empathetic persuasive conversation flows • Duration: 7 days	• AI (empathetic) chatbot (SlimMe; rule based+ML using Dialogflow) • Platform: mobile devices or smartphones		• Mixed methods usability study through simulation trial—UX questionnaire after the trial • Location: Taiwan		• Small sample size and language restriction (only English) may limit result generalization • Nutrition assessment methods rely on user self-reports, which may introduce bias • Self-reported anthropometric measurements may be less accurate compared to direct measurements
Alturaiki et al [103], 2022	• To improve self-management of β thalassemia and communication between patients and health care providers • Duration: not identified	• AI chatbot (rule based) • Platform: web browser on computers or mobile devices	• Patients with β thalassemia • N=34 • Participants' age not identified	• Prototype implementation and usability evaluation study through qualitative posttest survey • Location: not identified	• Usability and perceived utility assessment: • Most participants (72%) found the chatbot easy to use • Above 90% of participants considered the chatbot beneficial • Most participants agreed that the chatbot made managing β thalassemia easier and more efficient • Chatbot has the potential to save patients time and money that they usually spend on hospital visits	• Small sample size, which may limit generalizability of findings • Chatbot was tested in a live chat scenario, possibly not reflecting real-world use accurately • Limited data availability, with findings not publicly accessible due to privacy or ethical constraints • No specification of statistical analysis or validation methods for assessing the chatbot's effectiveness • Long-term usability and effectiveness of the chatbot in managing β thalassemia was unexplored
Zahedi et al [104], 2022		• Virtual human avatars (AI nurse avatar and digital patient avatars) in a virtual hospital system (Wepital) • Platform: computer (laptops and desktop computers) via the Second Life website	• Patients with GI problems • N=61 • Participants' age not identified	• Hypothesis experiment and usability study with real patients using the Wepital prototype—mixed methods approach • Location: not identified		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• To support constant ubiquitous medical care for GI[ax] disease management through an avatar-based telepresence system • Duration: not identified				• Significant positive impact of telepresence affordance on patient satisfaction with the Wepital system, with a path coefficient of 0.41 (P<.001) • Significant positive impact of convenience affordance on patient satisfaction with the Wepital system, with a path coefficient of 0.33 (P<.001) • Significant positive impact of trust affordance on patient satisfaction with the Wepital system, with a path coefficient of 0.43 (P<.001) • Significant positive impact of using real avatars with wearable sensors on patient satisfaction with the Wepital system, with a coefficient of 0.11 (P<.01) • Significant positive impact of patients' predispositions, such as flexibility regarding method of care and understanding information, on their perceptions of affordances in the Wepital system, with coefficients of 0.70, 0.63, and 0.29 (P<.001 in all cases), respectively • No significant influence on trust due to lack of privacy concern	• Small sample size due to limited patient participation because of sensitive context and strict recruitment rules • Researchers' restricted access to patients' medical records prevented the integration of real avatars with comprehensive medical information • Study focused on patients with GI in a single context, limiting generalizability of findings to other medical fields • Patients staying in the Wepital for persistent care were not involved in the study, limiting understanding of patient satisfaction
Eagle et al [37], 2022	• Evaluation of CA's mental health-related advice for improved self-management of depression and anxiety • Duration: not identified	• 6 commercial CAs—chatbots (Wysa and Replika) and voicebots (Google Assistant, Alexa, Cortana, and Siri) • Platform: • Chatbots: mobile devices or smartphones • Voicebots: smart speaker devices	• People experienced using CAs for mental health advice and management • N=141 • Participants' age not identified	• Mixed methods observational study—response quality evaluation using user-bot conversational data, PHQ-9 and GAD-7 survey questionnaires, and interviews with clinicians • Location: United States		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
					• Generally low quality of advice and recommendations provided by CAs for mental health questions • Wide variability in the quality of responses across different CAs—no significant advantage of voice agents over chatbots • Text-based chatbots slightly outperformed voice-based agents with improved responses regarding better treatment advice due to the chatbots' superior dialogue capabilities and empathy support, whereas voice assistants were prone to speech recognition errors and were more dependent on simple web searches in generating advice • CAs struggled to handle unexpected inputs • Anxiety-related responses were slightly better than the responses for depression and crisis situations • Extended dialogue context access improved the CAs' quality of responses • Design implications included providing clarification of subdialogues and access to extended dialogue context	• Limited sample size may restrict the generalizability of findings to a wider population • Assessment of CAs' responses limited to mental health questions on anxiety and depression • Standardized survey-based conversational probes may not have captured the full range of user queries and responses in real-world settings • Quality of responses assessed through subjective participant ratings, introducing potential biases • No investigation on the impact of user characteristics (eg, age, gender, or previous mental health service experience) on response quality • Long-term effects of CAs' advice on user outcomes such as mental health improvement or treatment adherence were not addressed • Ethical implications of using CAs for mental health advice, including privacy concerns and potential harm from inaccurate information, were unexplored • CAs' responses were not compared with those provided by human experts, which could have provided insights into effectiveness
Maharjan et al [83], 2022	• Self-report of depression or bipolar disorder • Duration: 4 weeks	• Voice assistant and voicebot (Sophia; rule-based AI using Dialogflow) • Platform: Google Nest Mini smart speakers	• Adults with depression or bipolar disorder • N=20 • Participants aged ≥18 years (18-34 years)			

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
				• Mixed methods in-the-wild study—acceptability and usability evaluation using user-agent engagement data, fortnightly completed WHO-5 health questionnaire, and semistructured interviews • Location: not identified	• Global average engagement rate=75%, indicating a high engagement level with the voicebot • Diverse UX with varying personified perceptions, social context, privacy and security concerns, and conversational features • Perceived helpfulness of self-reporting practice in daily reflection and through organization, and the process was considered meditative by some participants • The voicebot was perceived as usable and acceptable for self-reporting mental health well-being in a naturalistic setting	• Study conducted during the pandemic, potentially impacting participants' mental, physical, and social conditions, affecting result generalizability • Participants' interaction with the CA may be influenced by lockdowns, social distancing, and travel restrictions, affecting their typical context and experiences • Interview method (web-based or in person) could influence how participants conveyed their experiences, potentially influencing study outcomes • Privacy and data security concerns among participants, leading to self-censorship during self-reporting • Participants' engagement and experiences with the CA may be affected by their varying technology familiarity levels
Meheli et al [105], 2022	• Cognitive behavioral intervention for self-management of chronic pain and mental health • Duration: not identified	• Commercial CA—AI mental health chatbot (Wysa) • Platform: mobile devices or smartphones	• Individuals with self-reported chronic pain seeking digital mental health support • N=2194 (real-world data) • Participants' age not identified	• Mixed methods retrospective observational study—chatbot effectiveness evaluation through user-chatbot engagement data and pretest-posttest assessment questionnaires (PHQ-9 and GAD-7) • Location: not identified	• Identified themes related to health concerns, socioeconomic concerns, and pain management concerns among users with chronic pain • Users with chronic pain showed significantly higher app engagement (P<.001) compared to users without chronic pain, with a large effect size (Vargha and Delane A=0.76-0.80) • Significant reduction in anxiety and depression symptoms among users with chronic pain, as shown by the pretest-posttest assessments using PHQ-9 and GAD-7 scales (P<.001), with a medium effect size (Cohen d=0.60-0.61) • Chatbot was found to be useful in managing challenges, including anxiety, sleep, low energy, motivation, loss, and pain • Conversational data revealed perceived needs and experiences of individuals with chronic pain, including the need for personalized and flexible support	• Small samples for the third objective may further constrain generalizability of findings, underscoring the need for careful interpretation as preliminary outcomes • Nonrandomized sampling design may limit generalizability of findings • Data extraction keywords based on guidelines, literature, and clinical experience may have missed relevant pain terms • Repeated measurements for efficacy without a control group might risk increasing regression to the mean • Users were not mandated to complete assessments, limiting efficacy study to a small subset

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
Henson et al [106], 2022	• To provide appropriate patient education for self-management of GERD[ay] • Duration: not identified	• Commercial CA—AI chatbot (ChatGPT) • Platform: multiple devices—mobile devices or smartphones, laptops, and desktop computers (web browser and app)	• Patients with GERD • N=11 (8 patients and 3 gastroenterologists) • Participants' age not identified	• Feasibility study (quantitative)—potential utility evaluation of ChatGPT in GERD using 23 GERD management prompts • Location: not identified	• 63/69 (91.3%) appropriate responses to GERD management queries delivered by ChatGPT • Frequent inconsistency in responses to the same prompt, with 16/23 (70%) prompts resulting in variable appropriateness • Highest number of appropriate responses (39.4%) was to treatment-related prompts, whereas the highest number of inappropriate responses (14.3%) was to diagnosis and management–related prompts • Most responses (78.3%) contained some specific guidance, particularly for diagnosis-related prompts (93.3%) • Patients with different educational backgrounds considered the responses understandable and more useful than a search engine • ChatGPT failed to recommend consideration of Roux-en-Y gastric bypass for ongoing GERD symptoms with pathological acid exposure in obesity setting • Current limitations prevent the integration of ChatGPT into routine clinical practice at present	• Small sample size may limit the generalizability of findings • ChatGPT was not specifically trained on medical literature, and its ability to address repeated requests for clarification was not evaluated • Inappropriate responses with inconsistencies to the same prompt were observed, which may affect results • Limited specific guidance and content errors • PPI[az] risks were stated as fact, which lacked balanced consideration of benefits and context • ChatGPT often provided overly long responses, diluting the clinical impact of responses in many cases
Babington-Ashaye et al [75], 2023	• To improve patient knowledge and support symptom monitoring of hemophilia • Duration: not identified	• AI chatbot (Saytù Hemophilie; NLU[ba] and ML based) • Platform: mobile devices or smartphones (via messaging apps such as WhatsApp, Signal, and Telegram)	• People with hemophilia and their families in Senegal • N=30 (20 people with hemophilia and 10 family members) • Participants' age not identified	• Usability study—mixed methods approach with SUS survey • Location: Senegal (West Africa)	• Average usability (SUS) score: 81.7 (indicates high usability of the system) • 75.4% of participants expressed an overall high level of satisfaction with the French version of the chatbot • Most participants perceived the proposed AI-based chatbot as a potential solution to manage the symptoms at home while awaiting a physician's consultation	• Small sample size restricted the ability to conduct additional statistical tests, limiting result generalizability • Potential sample bias as 40.6% of participants recruited were students • While the SUS score was used, additional evaluation measures could be used to assess users' overall opinion of the chatbot
Alhmiedat and Alotaibi [79], 2023		• Humanoid robot (SARA) • Platform: physical presence	• Children with T1DM • N=5 (children) • Participants aged 5-9 years			

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Symptom monitoring and patient education for T1DM management • Duration: 5 days			• Pilot study—acceptability evaluation of SARA using quantitative questionnaires, observations, and experiments • Location: Saudi Arabia	• All participants succeeded in all the following measures with different outcomes: • Total interaction time with SARA • Productivity of education task • Productivity of the short test task • Productivity of listening to stories • Moderate worthiness level: 88.2%, indicating a high relative acceptability level • User feedback: positive feedback from the patients, their parents, and clinical staff	• Very small sample size of the pilot study, limiting generalizability of results • SARA framework is still under development, requiring further experiments before finalizing the robotic platform
Calvo et al [84], 2023	• To estimate the risk of an asthma exacerbation and provide recommendations for improved self-management of asthma • Duration: 28 days	• Text-based chatbot (rule-based AI) • Platform: mobile devices or smartphones (WhatsApp based)	• Adults with asthma • N=300 (40 for alpha testing and 260 for beta testing) • Participants aged ≥18 years	• Feasibility and usability study with pretest-posttest quantitative questionnaires • Location: United Kingdom	• Chatbot was found to be acceptable, usable, and satisfactory to participants as it was helpful in improving their asthma control • Task completion rate: 80% of participants completed the risk assessment process, whereas 20% dropped out • Net Promoter Score for the chatbot: 8.5 (indicating high likelihood of recommending it to others with asthma) • Participants reported that they were satisfied with the chatbot and that the risk measurement by the chatbot was useful • Participants suggested improving the chatbot by incorporating more personalized feedback and additional interactive features • Chatbot seemed to be effective in promoting motivational quality, engagement, and UX in relation to basic psychological needs • High level of alignment between expected and calculated risk, and the calculated risk was perceived as accurate by the user	• Limited sample size • No intention to find gender or age differences within the current sampling • No intention to measure significant asthma control improvements in 28 days • Potential for incorrect responses from individuals who incorrectly answered the screening questionnaire • Lack of validation for measurable health outcomes
Epalte et al [107], 2023		• AI chatbot (Vigo; rule based+ML) • Platform: smart devices (eg, Apple iPad tablet)	• Patients with stroke and their families • N=12 • Participants' age not identified	• Usability study—qualitative semistructured interviews after the intervention • Location: not identified		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Personalized exercise program with patient education and counseling support for self-management of post-stroke recovery • Duration: 1 month				• Participants considered the chatbot transparent, understandable, and handy • Overall design of Vigo was rated as good • Participants were mostly unsatisfied with the difficulty level and diversity of exercises • 5 themes related to the app's effectiveness were identified: flexibility, information, exercises, emotional support, and assessment • Chatbot could have a potential positive impact on post-stroke outcomes, particularly in emotional status, social interaction, improved mood, and motivation	• Strict inclusion criteria and chatbot's complexity limited recruitment of patients with greater physical limitations, potentially skewing insights • Patients with severe functional limitations, aphasia, cognitive deficits, and comorbidities may struggle to use the app independently, reducing the potential user base • Chatbot's availability only in Latvian may restrict use for Russian-speaking patients • Difficulty in adapting the program for patients due to their different health statuses, personal factors, and preferences
LeRouge et al [108], 2023	• Personalized lifestyle intervention for self-management of obesity • Duration: not identified	• AI-based computer-animated cartoon-like 2D avatars • Platform: mobile devices or smartphones and desktop computers	• Adolescents with obesity or overweight • N=70 (for phase 2) • Participants' age not identified	• Qualitative participatory research study in 2 phases: • Phase 1: focus groups with teenagers, provider interviews, and parent interviews • Phase 2: midrange prototype assessment by teenagers and providers • Location: United States		• Limited data sources may limit universal applicability of findings • Study scope was limited to a specific audience using qualitative methods, which may limit generalizability • Potential coding bias in the data analysis process despite minimization efforts

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
					• Effectiveness: avatars and virtual agents can be effective in engaging adolescents in weight management programs • Avatars and virtual agents can provide personalized support and motivation to adolescents • Users would be more motivated to use the virtual agents if they were more fun, creative, human-like, interactive, communicative, and indicative of progress or success • Design characteristics: • Personalize avatars matching user needs and preferences • Ability to stimulate 2-way interaction (preferably fun, creative, and more human-like) • Support multiple communication modes (eg, text, voice, and video) • Provide feedback on user progress and behavior • Challenges: • Implementation of avatars requires technical expertise and resources • User acceptance barriers due to trust issues and privacy concerns • Integration requirements of avatars with existing health care systems and workflows • Development of avatars and virtual agents can be expensive	
Park et al [109], 2023		• 2 AI chatbots (human-like vs machinelike; NLP using Dialogflow) • Platform: not identified	• Adults seeking web-based mental health support • N=385 • Participants' age not identified	• Quasi-experimental design—quantitative UX evaluation of a human-like vs machinelike chatbot for web-based mental health counseling • Location: United States		

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
	• Improvement of web-based mental health counseling efficacy through effective conversation between a user and a chatbot agent • Duration: not identified				• Human-like chatbot apparently yielded higher user intention to comply with health recommendations (mean 5.69, SD 0.61) than the machinelike one (mean 5.34, SD 0.93; t383=4.41; *P*<.001) • Human-like chatbot resulted in higher compliance with health recommendations mediated by psychological distance • Positive correlation between human representation and perceived trust and between perceived trust and compliance with health recommendations • Human-like chatbot led to psychological closeness and increased trust, thereby causing higher intention to comply with health recommendations	• Convenience sampling could limit generalizability • Study focused solely on chatbot's physical representation, excluding other important cues such as conversational and identity cues • User characteristics such as involvement and self-efficacy, which could moderate the impact of human-like cues, were unexplored • Effect of demographic variables (eg, age, gender, and ethnicity) was not investigated for both users and chatbot agents • Using a self-administered web-based questionnaire might introduce limitations such as social desirability bias and response bias while lacking control over participants' environments • Long-term effects of chatbot counseling on mental health outcomes were not investigated
Boggiss et al [110], 2023	• To deliver self-compassion coping lessons to support T1DM self-management and mental health • Duration: not identified	• AI self-compassion chatbot with customizable avatar (COMPASS) • Platform: mobile devices or smartphones	• Adolescents with T1DM • N=30 (19 adolescents with T1DM and 11 diabetes health care professionals) • Participants aged 12-16 years	• Qualitative study—focus groups and interviews • Location: New Zealand	• COMPASS chatbot was found to be acceptable to adolescents with T1DM and diabetes health care professionals • The chatbot apparently had the potential to support the T1DM self-management and mental health of adolescents during the COVID-19 pandemic • Participants emphasized the importance of personalization of both content and features of the chatbot (eg, games, apps, background color schemes, and hobbies) and cultural appropriateness to support users with different ethnic backgrounds	

XSL•FO
RenderX

Study, year	Intervention and study duration	CA[a] type and delivery platform	Target population, sample size, and participants' age	Methods and location	Measures and outcomes	Limitations
						• Small sample size may not be representative of the breadth of challenges faced by adolescents with T1DM or additional barriers to standard care access • Potential bias in user feedback due to existing relationships between the facilitator and participants • Feedback was limited as participants only viewed screen recordings of the chatbot, potentially reducing its richness • Primary assessment limited to acceptability, perceived clinical utility, and usability of chatbot features, potentially overlooking the intervention's overall impact • User feedback may be influenced by inequities in access to technology such as smartphones and continuous glucose monitors

[a]CA: conversational agent.
[b]AI: artificial intelligence.
[c]RCT: randomized controlled trial.
[d]AF: atrial fibrillation.
[e]AFEQT: AF Effect on Quality of Life questionnaire.
[f]T2DM: type 2 diabetes mellitus.
[g]ML: machine learning.
[h]mHealth: mobile health.
[i]AADE: American Association of Diabetes Educators.
[j]WHO-5: 5-item World Health Organization Well-Being Index.
[k]K10: Kessler Psychological Distress Scale.
[l]BADS: Behavioral Activation for Depression Scale.
[m]BADS-AC: Behavioral Activation for Depression Scale-Activation.
[n]BADS-AR: Behavioral Activation for Depression Scale-Avoidance/Rumination.
[o]BADS-WS: Behavioral Activation for Depression Scale-Work/School Impairment.
[p]BADS-SI: Behavioral Activation for Depression Scale-Social Impairment.
[q]VDMS: virtual diabetes management system.
[r]PHQ-9: 9-item Patient Health Questionnaire.
[s]GAD-7: 7-item Generalized Anxiety Disorder Scale.
[t]T1DM: type 1 diabetes mellitus.
[u]PAL: personal assistant for a healthy lifestyle.
[v]SUS: System Usability Scale.
[w]BCT: behavior change theory.
[x]HRQoL: health-related quality of life.
[y]AQoL-8D: Assessment of quality of life-8 dimensions.

XSL•FO
RenderX

[z]HbA_{1c}: glycated hemoglobin.
[aa]HADS: Hospital Anxiety and Depression Scale.
[ab]GP: general practitioner.
[ac]SCD: sickle cell disease.
[ad]EVA: Education virtual assistant.
[ae]FBG: fasting blood glucose.
[af]PPBG: postprandial blood glucose.
[ag]VH: virtual human.
[ah]UX: user experience.
[ai]UEQ: User Experience Questionnaire.
[aj]TAM: technology acceptance model.
[ak]NLP: natural language processing.
[al]CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth.
[am]OR: odds ratio.
[an]RMDQ: Roland-Morris Disability Questionnaire.
[ao]HF: heart failure.
[ap]MoCA: Montreal Cognitive Assessment.
[aq]DHFKS: Dutch Heart Failure Knowledge Scale.
[ar]QoL-AD: Quality of Life in Alzheimer Disease scale.
[as]TTS: Text-to-Speech.
[at]FAQ: frequently asked question.
[au]WAI-TECH: Working Alliance Inventory–Technology Version.
[av]PST: problem-solving treatment.
[aw]GIF: graphics interchange format.
[ax]GI: gastrointestinal.
[ay]GERD: gastrointestinal reflux disease.
[az]PPI: proton-pump inhibitor.
[ba]NLU: Natural Language Understanding.

A few studies (6/43, 14%) focused on evaluating the effectiveness of the CAs used for the self-management of certain chronic illnesses, such as atrial fibrillation [10], type 2 diabetes mellitus [24,82], and mental health conditions such as depression and anxiety [90,96,97]. All these studies found improvement in various parameters in management, such as significant reduction in glycated hemoglobin level ($P<.001$), as well as BMI, and weight ($P<.001$) in patients with type 2 diabetes mellitus [82]; notable improvement in quality of life measured using the Atrial Fibrillation Effect on Quality of Life questionnaire score ($P<.05$) or health-related quality of life utility score ($P=.007$) in patients with atrial fibrillation [10,24]; and substantial reduction in depression (9-item Patient Health Questionnaire score) and anxiety (7-item Generalized Anxiety Disorder Scale score) symptoms [90,97,105].

Types of AI-Based CAs in the Included Studies

Of the 47 AI-based CAs representing digital health assistants identified from the included studies, 9 (19%) were existing commercial CAs available as marketed products, which included voice assistants—Google Assistant, Amazon Alexa, Microsoft Cortana, and Apple's Siri, as well as social chatbots—ChatGPT and Replika (a voice-enabled, avatar-based chatbot), and health care chatbots—Vik, Wysa, and Tess. The remaining 38 CAs were proposed designs or prototypes, where 19 (50%) of them were chatbots, 2 (5%) were voice assistants or voicebots [58,102], 13 (34%) were human-like digital avatars, 1 (3%) was a digital avatar version of a humanoid social robot [85], 1 (3%) was a text-based voice-enabled bot [74], and 2 (5%) were avatar-based chatbots [91,110]. In addition, 5% (2/38) were CAs as humanoid social robots [79,85], which were included as they particularly emphasized managing NCDs such as diabetes (although the *robots* topic was beyond the scope of this study).

Intelligence Level of the AI-Based CAs

Of the 38 CA prototypes presented, 27 (71%) were found to have indicated their AI algorithm implementation methods to generate appropriate health-related responses. A total of 30% (8/27) of them were implemented using rule-based algorithms [72,81,83,84,86-88,103], 7% (2/27) of the CAs were developed using network-based scripted algorithms [10,80], and 4% (1/27) of the CAs were implemented using prescripted algorithms based on behavior change theory, whereas 33% (9/27) of the CAs were implemented using ML approaches [58,74,75,89,93-95,97,99] and 15% (4/27) of the CAs were developed using a combination of rule-based and ML techniques [78,85,101,107]. However, one of the CAs identified from a study was a Wizard-of-Oz human-like AI avatar with prescripted text and human-controlled responses to evaluate the potential effectiveness of visually embodied avatars in the assessment of anxiety and depression through users' behavioral expressiveness [96].

XSL•FO
RenderX

Types of Health Interventions and Study Duration

Most included studies (24/43, 56%) focused on behavioral or lifestyle interventions and patient education for improved self-management of physical NCDs such as *CVDs* [10,80,99,100], *diabetes* [24,72,74,79,82,85,89,93] or *prediabetes* [76], *cancer* [81,92], *gastrointestinal diseases* [104,106], and obesity [76,86,108].

A reasonable number of studies (13/43, 30%) focused on cognitive behavioral interventions related to the self-management of chronic cognitive impairments (eg, dementia and poststroke) [73,107], including chronic mental health conditions such as anxiety, depression, psychological distress, and bipolar disorders [83,87,88,90,96,97,101,102].

The duration of these studies was reasonably divergent, ranging from 5 days to 1 year. However, a substantial number of studies (14/43, 33%) could be considered long term, such as 8 weeks to 2 months [85,88,90,99,101], 12 weeks to 90 days [86,98,100], 16 weeks to 6 months [76,82,87,93], 9 months [94], and 12 months or 1 year [24,92]. A total of 26% (11/43) of the studies conducted trials for approximately ≤1 month [10,73,81,83,84,88,97,107] (ie, within 5-10 days) [72,78,79]. A total of 42% (18/43) of the studies did not mention the study duration.

Study Population and Location

Almost half (20/43, 47%) of the included studies did not specify the target population other than patients with a particular chronic disease that the health intervention was meant for. The studies that specified the target population or their age group (24/43, 56%) were relatively heterogeneous, with older adults [10,58,72,73,91,95] and middle-aged adults [24,99] being the most frequently targeted user group (8/43, 19%). However, some studies (6/43, 14%) also focused on younger adults [97], adolescents [76,108,110], and children [79,85].

The locations of the studies conducted were largely diverse, but most (27/43, 63%) were conducted in higher-income countries. A total of 9 studies were completed in Europe, where 2 (22%) of them were conducted in the United Kingdom [84,91], 2 (22%) studies were conducted in Italy and the Netherlands [73,85], 1 (11%) was conducted in Portugal [72], 1 (11%) was conducted in France [77], 1 (11%) was conducted in Greece [93], 1 (11%) was conducted in Spain [94], and 1 (11%) was conducted in Ukraine [97]. A total of 16% (7/43) of the studies were conducted in the United States [10,37,76,86,87,108,109]. In total, 4 of the studies were carried out in the Australasia region—3 (75%) in Australia [24,99,101] and 1 (25%) in New Zealand [110]. A total of 7 studies were conducted in Asia, where 4 (57%) of them were conducted in Japan [81,88,96,98], 1 (14%) was conducted in Singapore [74], 1 (14%) was conducted in Taiwan [78], and 1 (14%) was conducted in India [82]. A total of 5% (2/43) of the studies were conducted in Saudi Arabia [79] and Senegal in West Africa [75]. The remaining 33% (14/43) of the studies did not specify the study location.

Data Collection Methods of the Included Studies

The included studies mainly used quantitative (17/43, 40%) or mixed methods (17/43, 40%) approaches. A substantial proportion of these studies (11/43, 26%) were pretest-posttest intervention studies to observe the differences before and after the intervention within the same group of participants (single armed) or between intervention and control groups (2 armed). The study designs of the quantitative and mixed methods studies included observational experiments using 2-armed randomized controlled trials (7/34, 21%) [10,24,86,87,98,99,102], a nonrandomized prospective study (1/34, 3%) [88], retrospective studies (2/34, 6%) [82,105], quasi-experimental studies (4/34, 12%) [10,90,101,109], a single-armed pilot trial [94], and comparative evaluation studies (3/34, 9%) [89,96,100].

The qualitative studies (9/43, 21%) were mostly usability and acceptability assessments of prototypes through qualitative posttest surveys [103], postintervention interviews [107], focus groups [110], a participatory design approach with focus groups and interviews [108], and workshops [91]. A total of 22% (2/9) of the qualitative studies compared the proposed prototypes against similar existing applications [58,89].

Discussion

Principal Findings

The discussion of this scoping review synthesizes the findings in light of the overarching themes emergent from each research question, elucidating the interplay between the current state, challenges, benefits, and the targeted users of AI-based CAs within the scope of the examined literature.

Current State of AI-Powered CAs

Regarding the current state of research on AI-powered CAs as human-like health carers for managing NCDs, empirical studies have revealed the limited ability of existing commercial CAs to provide appropriate lifestyle advice and health recommendations [37,111]. Perhaps this is due to the lack of specified data training for these commercial tools in the medical context, which is currently restricting their integration in clinical and remote care [106].

Functionality-wise, commercial chatbots (eg, Wysa and Replika) tend to slightly outperform voice agents (eg, Google Assistant, Siri, Alexa, and Cortana) with improved response quality due to the former's better dialogue capabilities and empathy support. At the same time, the latter are more susceptible to speech recognition errors, including greater dependency on simple web searches to generate advice [37]. Nonetheless, voice-enabled CAs offer hands-free communication, making them preferable over text-based chatbots for older adult users, as suggested by a recent scoping review conducted by Even et al [70] focusing on the benefits and challenges of CAs in the older adult population (Table 1). In addition, another study revealed that older adults preferred voice-based communications over SMS text messages or free-text entry as voice is regarded as a powerful mode for promoting motivation [112]. In addition, voice agents may also be more accessible compared to digital avatars due to their lower bandwidth requirements, making them suitable for locations with lower bandwidth connectivity or devices with limited computing power [48]. Conversely, digital avatars are generally more likely to be favored by users over other types of CAs for their widely accepted face-to-face

XSL•FO
RenderX

conversation format [10] as they can support both text and audio outputs [47,113] and, therefore, can serve as a valuable alternative complement for remote care assistance [114]. This suggests that digital avatars remain a popular choice for their versatility and interactive capabilities, driving the rapid progression toward the anthropomorphism era as they represent realistic human-like traits in behaviors and appearances for enhanced interaction with the users [115].

Moreover, anthropomorphic CAs are more preferred over nonanthropomorphic CAs, as shown by Park et al [109], indicating that the human-like representations of a chatbot yielded a higher likelihood of intervention compliance than the machinelike chatbot because the former increased closeness and trust. Similarly, another study demonstrated that an augmented human-like embodiment of a digital avatar resulted in greater social closeness among the participants [116]. Correspondingly, strong social presence and emotional closeness are regarded as key facilitators of favorable attitudinal and behavioral outcomes among users [117]. Hence, anthropomorphism positively correlates to users' perceived trust and acceptance of a virtual agent [12,117]. Indeed, digital human faces as intelligent CAs are considerably promising for achieving a human-like presence to enhance human-machine interaction [42]. Consequently, the application of AI-based conversational avatars as digital human-like health carers has been gaining attention recently, leading patients to prefer digital human avatars as medical professionals over chatbots due to the former's perceived human-like and interactive nature, which is essential for maintaining the professionalism and emotional aspects in physician-patient relationships [39].

Nonetheless, this evolution may apparently raise perceived uncanniness, which is considered a genuine ethical concern with realistic anthropomorphic avatars, yet it does not appear to be a universal problem for the avatar appearance or limit users' engagement with the virtual environment [55]. Realistic humanoid avatars representing virtual carers can appear more trustworthy and increase compliance among users [55,109]. However, striking a balance between realistic and stylized avatar appearances can be worthwhile to avoid visual overloading and discomfort among users while representing the required elements accurately to maximize perceived trust, closeness, and acceptability [25].

Challenges and Limitations of CAs

The following challenges with CAs, including limitations in intelligence level, empathy, data privacy, and ensuring patient safety, highlight significant barriers to their adoption and effectiveness in health care.

Intelligence Level of CAs

Most of the proposed AI-based CAs were rule-based or scripted systems with a predetermined set of rules and conditions, which are easier to design and implement but have limited capability to handle complex conversations (unlike the automated ML-based systems) [33,56,118] and, therefore, cannot generate personalized responses in a more human-like manner, which the users would mostly prefer [84,91,108-110]. Furthermore, few studies proposed integrating personality-tailored behavior in CAs, precisely known as personality-adaptive CAs, which can automatically recognize and adapt to the users' distinctive personality traits accordingly to be able to support patients in a personalized and authentic human-like manner [119,120] as well as effectively persuade users to alter their attitudes and behaviors for healthy lifestyle habits [121].

Empathy of CAs

Studies have emphasized the importance of including the "empathy" element in the design of CAs [122,123], yet only one of the included studies notably aimed to incorporate "empathy" in the proposed chatbot by applying text-based emotion analysis to recognize the user's emotion and respond accordingly using emojis, stickers, and Graphics Interchange Format images [78]. Furthermore, a scoping review on AI-based CAs conducted by Kusal et al [56] has suggested that the major shortcoming of CAs is their lack of competence to make communication with humans seem natural with the aid of empathy and sentiments, mainly due to their failure to comprehend the human context and the users' emotions as their responses are restricted to the queries they are trained for, which can lead to a frustrating user experience. Indeed, a study conducted by Casas et al [124] determined the capability of an empathetic chatbot to surpass a benchmark bot or even humans in predicting the emotional state of textual messages. Therefore, CAs should be able to model their context and select important information to remember as communication depends on past messages [56].

Data Privacy

Surprisingly, the included studies demonstrated limited emphasis on the data privacy of the AI-based CAs while handling patient data, which is a common challenge faced in the design and evaluation of CAs [111,125]. Hence, it is recommended to prioritize data privacy and confidentiality factors during the development of such technologies to gain users' trust and acceptance [63].

Patient Safety

Another area that needs to be focused on is patient safety [126], which was also not highlighted in the included studies. For instance, the inability of digital health tools to recognize the urgency of the patient's health condition may generate inappropriate health advice or lead to delayed treatment, increasing the risk of further complications. In addition, limited validation and medical training for existing CAs (as mentioned previously) poses risk of inaccuracy and raises concerns about patient safety. Therefore, assessing the validation status of these digital health tools is crucial for patient safety and optimal health care outcomes.

Uses and Benefits of AI-Based CAs in Managing NCDs

This review suggests that CAs are primarily used for supporting nonpharmacological interventions such as behavioral or lifestyle medication for managing NCDs such as CVDs, diabetes, obesity, cancer, and gastrointestinal diseases, as well as preventive measures for healthy living, indicating their potential ability to complementarily assist in the long-term remote management of NCDs [114,127,128]. The ubiquitous self-care support through constant symptom monitoring, along with the

XSL•FO
RenderX

personalized behavioral modifications provided by CAs, can possibly rescue patients from minor health concerns and prevent the progression of chronic diseases [10,104]. Health care professionals are optimistic about CAs' potential ability to save patients time and money that they usually spend on physical clinic visits [103], thereby reducing the health care cost and burden and increasing accessibility [91,104]. Augmented reality applications could enhance the training of health care professionals and even patients in using these new technologies effectively [115,129,130].

Efficacy

The efficacy evaluation of CAs for remote care is still not thoroughly explored and well understood. According to Monaco et al [63], inadequate evidence regarding the efficacy of digital health tools may have hindered the potential beneficial outcomes of using such resources to control NCDs. Nonetheless, the few included studies that aimed to identify the effectiveness of CAs in managing the target NCDs found favorable outcomes [10,24,82,90,96,97,105]. Moreover, many reviews have emphasized the efficacy of CAs in health care support. For instance, a systematic review conducted by Milne-Ives et al [26] investigated the effectiveness and usability of AI-based CAs designed for health care and found them generally effective, with positive or mixed evidence, whereas a narrative review by Dingler et al [60] used context-aware voice assistants (eg, Google Home and Amazon Echo), demonstrating the utility of speech-based CAs in providing personalized remote health care support. However, a systematic review conducted by Hossain et al [131] exhibited low positive outcomes of digital health interventions among people with NCDs in India, requiring further exploration of advanced technologies for the equitable and sustainable development of digital health tools.

Targeted Users

Most of the target user groups were based on chronic disease conditions rather than any specific age group. A fair proportion of studies (6/43, 14%) focused on the application of CAs in the older adult population, exhibiting that this vulnerable user group may be most in need of such technologies as older adults have greater risk of developing serious NCDs [28,58,61] and many older adult patients lead an isolated, lonely life and, therefore, require constant remote monitoring as well as a virtual companion for healthy living [132,133]. Hence, in other words, the selected user groups were mostly based on targeted chronic diseases that may be more prevalent in older adults. However, we did not encounter any significant study that modified the CAs to enhance accessibility for older adults, but social networking sites could potentially enhance engagement with such CAs among this user group [134].

Moreover, it is crucial to recognize that the vast majority (41/43, 95%) of the included studies were conducted in high-income countries, underscoring the necessity of further research among underserved populations, particularly in LMICs where individuals may be more vulnerable to developing NCDs due to limited health awareness and access to health care services.

Limitations

Although this is an emerging field of research and we conducted an extensive literature search strategy, the possibility of noninclusion of relevant studies may still exist (eg, missing certain keywords or search terms, studies in languages other than English, studies not listed in the searched databases, and unidentified studies from gray literature). Similarly, our stringent inclusion and exclusion criteria were implemented to refine our focus on patient-centered AI-based CAs tailored for the remote management of NCDs, aiming to ensure the relevancy and applicability of our review to this particular domain of interest. However, this approach may have inadvertently narrowed the scope, potentially overlooking other relevant studies that could provide wider insights into our findings. Despite this, we strived to conduct a fair selection of studies meeting our inclusion criteria to mitigate selection bias and ensure a balanced representation of the literature. Nonetheless, it is important to acknowledge that other potential biases, such as publication bias, may still exist, which are beyond our control. In addition, new articles may have appeared after our search deadline (July 31, 2023), which may cause additional differentiations.

Furthermore, some common limitations were also identified within the included studies that may impact the overall findings of this review: (1) limited sample sizes (a vast majority of the studies had relatively small sample sizes, which could limit the generalizability of the findings), (2) short-term interventions or lack of long-term outcome measures (the duration of interventions in most studies was relatively short [80,82,88,101], potentially overlooking long-term user interactions and health outcomes essential for a comprehensive evaluation of CAs' ability to sustainably assist in NCD management [10,37,77,85,91-93,96-98,103,109]), (3) limited demographic data (some studies lacked detailed demographic information of the participants [37,93,109], which may potentially limit the analysis of how demographic factors can impact the feasibility or effectiveness of the CAs), (4) potential bias (many studies had potential biases resulting from self-reported user data [37,77,78,101,109], sample bias [75,82,105], coding bias [108], and exclusion of variables or unaccounted confounding factors in analyses [100], which could impact the validity of the results), and (5) unaddressed technical aspects (some studies did not discuss the technical aspects of the proposed CAs, including their potential technical limitations or challenges [85,92,93]). Addressing these issues could contribute to the enhancement of the design, development, and usability of CAs in the future [135].

Implications and Future Recommendations

As CAs strive to facilitate personalized and empathetic human-like conversations in the health care field [113], the necessity for context-oriented and domain-specific training emerges as a crucial consideration for improved accuracy and relevance [136]. Although advanced AI systems such as large language models trained on vast textual data (eg, the generative pretrained transformer series and LLaMA) have proven transformative in various sectors such as marketing, education, and customer service, their application in the health care sector remains underexplored, primarily due to a scarcity of relevant

XSL•FO
RenderX

high-quality medical data sets [137]. In enhancing the AI-based CAs for health care, only a few recent studies have aimed to investigate real-world interactions between patients and health care professionals [136-139]. While recognizing the dynamic and nonstandardized nature of historical human interactions in health care, our recommendation to study real-world communications between patients and health caregivers stems from the objective of training CAs to be contextually relevant, particularly in addressing the unique challenges associated with specific chronic diseases. This approach may seek to empower CAs to navigate the complexities of diverse health conditions with greater intelligence and empathy, thereby improving their adaptability through the use of high-quality data sets acquired from real-world scenarios [136] given the substantial amount of data and extensive knowledge base required for domain-specific contextual training [56]. However, to ensure effective training of AI agents in health care, it is essential to define their roles (eg, nurse, health assistant, nutritionist, and pharmacist), which can determine the appropriate level of training required, allowing for the selection of specific data sets tailored to the medical training and evaluation of CAs.

Findings from the literature suggest an overall higher user preference for realistic human-like representations in CAs, specifically in the form of anthropomorphic digital avatars, which can be accessed through mobile devices (unlike physical humanoid robots) [52,54], thus accommodating a broad user base including older adult users [58,140]. Therefore, future research should prioritize investigating the impact of anthropomorphic avatars or virtual humans in managing NCDs, including their potential ethical concerns (eg, uncanny valley) that should be addressed in the design considerations to increase acceptance of CAs for self-care [25,49,123].

Moreover, user preferences related to CA types and features may vary notably among different user groups. For instance, older adults and users with disabilities may require voice-activated CAs over text-based CAs due to their typing inabilities, and therefore, enhanced speech recognition functionality is required to incorporate into CAs to accommodate voice conversations in multiple languages [70]. In addition to language requirements, factors such as culture, ethnicity, and individual personality traits may also play a significant role in influencing the adoption of AI dialogue systems for remote care purposes. Indeed, the adoption of CAs can be impeded among populations traditionally experiencing health inequities due to the limited availability of such tools with localized and culturally tailored features. These populations, often at higher risk of NCDs, could greatly benefit from more culturally sensitive health care technologies, which could improve health care access and outcomes, thereby necessitating careful consideration in future work.

Conclusions

The review highlighted the promising acceptance of CAs by users for the self-management of chronic conditions, with feedback indicating helpfulness, satisfaction, and ease of use in most of the included studies. AI-based CAs present opportunities for communication and interaction in health care settings. However, understanding and optimizing the communication channels between humans and such CAs is crucial for enhancing their capabilities and potential benefits in health care. While our study confirmed the increasing role of CAs in augmenting self-care and potentially reducing health care costs, it also exposed critical limitations, particularly concerning conversational depth and emotional intelligence. This review also emphasized the lack of reliable and comparable evidence to determine the efficacy of AI-enabled CAs for chronic health conditions. Therefore, while user feedback and acceptance were positive, there is a need for more rigorous studies and standardized reporting to evaluate the effectiveness of AI-based CAs as human-like health carers for managing NCDs.

Acknowledgments

We acknowledge the article processing fee (APF) sponsorship provided by the Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia.

This study would not have been possible without the project's support provided by the Jeffrey Cheah School of Medicine and Health Sciences (Seed grant I-M010-SED-000158).

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this paper.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Noncommunicable diseases. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2024-04-29]

XSL•FO
RenderX

2. Boutayeb A, Boutayeb S. The burden of non communicable diseases in developing countries. Int J Equity Health 2005 Jan 14;4(1):2 [FREE Full text] [doi: 10.1186/1475-9276-4-2] [Medline: 15651987]
3. Islam SM, Purnat TD, Phuong NT, Mwingira U, Schacht K, Fröschl G. Non-communicable diseases (NCDs) in developing countries: a symposium report. Global Health 2014 Dec 11;10:81 [FREE Full text] [doi: 10.1186/s12992-014-0081-9] [Medline: 25498459]
4. Lee JS, Kang MA, Lee SK. Effects of the e-Motivate4Change program on metabolic syndrome in young adults using health apps and wearable devices: quasi-experimental study. J Med Internet Res 2020 Jul 30;22(7):e17031 [FREE Full text] [doi: 10.2196/17031] [Medline: 32729838]
5. Qiao J, Lin X, Wu Y, Huang X, Pan X, Xu J, et al. Global burden of non-communicable diseases attributable to dietary risks in 1990-2019. J Hum Nutr Diet 2022 Feb;35(1):202-213. [doi: 10.1111/jhn.12904] [Medline: 33834556]
6. Wagner KH, Brath H. A global view on the development of non communicable diseases. Prev Med 2012 May;54 Suppl:S38-S41. [doi: 10.1016/j.ypmed.2011.11.012] [Medline: 22178469]
7. Budreviciute A, Damiati S, Sabir DK, Onder K, Schuller-Goetzburg P, Plakys G, et al. Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors. Front Public Health 2020;8:574111 [FREE Full text] [doi: 10.3389/fpubh.2020.574111] [Medline: 33324597]
8. Mohamud WN, Ismail AA, Sharifuddin A, Ismail IS, Musa KI, Kadir KA, et al. Prevalence of metabolic syndrome and its risk factors in adult Malaysians: results of a nationwide survey. Diabetes Res Clin Pract 2011 Feb;91(2):239-245. [doi: 10.1016/j.diabres.2010.11.025] [Medline: 21146882]
9. Bickmore T. Relational agents for chronic disease self-management. In: Hayes BM, Aspray W, editors. Health Informatics: A Patient-Centered Approach to Diabetes. Cambridge, MA: MIT Pres; 2010:181-204.
10. Bickmore T, Pusateri A, Kimani EN, Paasche-Orlow MK, Trinh H, Magnani JW. Managing chronic conditions with a smartphone-based conversational virtual agent. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. 2018 Presented at: IVA '18; November 5-8, 2018; Sydney, Australia p. 119-124 URL: https://dl.acm.org/doi/10.1145/3267851.3267908 [doi: 10.1145/3267851.3267908]
11. Jennings A, Powell J, Armstrong N, Sturt J, Dale J. A virtual clinic for diabetes self-management: pilot study. J Med Internet Res 2009 Mar 30;11(1):e10 [FREE Full text] [doi: 10.2196/jmir.1111] [Medline: 21821504]
12. Moriuchi E. Leveraging the science to understand factors influencing the use of AI-powered avatar in healthcare services. J Technol Behav Sci 2022 Sep 09;7(4):588-602. [doi: 10.1007/s41347-022-00277-z]
13. Ahmed F. Beyond patient monitoring: conversational agents role in telemedicine and healthcare support for home-living elderly individuals. arXiv Preprint posted online March 3, 2018 [FREE Full text]
14. Kankeu HT, Saksena P, Xu K, Evans DB. The financial burden from non-communicable diseases in low- and middle-income countries: a literature review. Health Res Policy Syst 2013 Aug 16;11:31 [FREE Full text] [doi: 10.1186/1478-4505-11-31] [Medline: 23947294]
15. Gingele AJ, Amin H, Vaassen A, Schnur I, Pearl C, Brunner-La Rocca HP, et al. Integrating avatar technology into a telemedicine application in heart failure patients : a pilot study. Wien Klin Wochenschr 2023 Dec;135(23-24):680-684 [FREE Full text] [doi: 10.1007/s00508-022-02150-8] [Medline: 36732377]
16. Toro-Ramos T, Lee DH, Kim Y, Michaelides A, Oh TJ, Kim KM, et al. Effectiveness of a smartphone application for the management of metabolic syndrome components focusing on weight loss: a preliminary study. Metab Syndr Relat Disord 2017 Nov;15(9):465-473. [doi: 10.1089/met.2017.0062] [Medline: 29035677]
17. Jakobsen AS, Laursen LC, Rydahl-Hansen S, Østergaard B, Gerds TA, Emme C, et al. Home-based telehealth hospitalization for exacerbation of chronic obstructive pulmonary disease: findings from "the virtual hospital" trial. Telemed J E Health 2015 May;21(5):364-373 [FREE Full text] [doi: 10.1089/tmj.2014.0098] [Medline: 25654366]
18. Gan SK, Koshy C, Nguyen PV, Haw YX. An overview of clinically and healthcare related apps in Google and Apple app stores: connecting patients, drugs, and clinicians. Sci Phone Appl Mob Devices 2016 Jul 19;2(1). [doi: 10.1186/s41070-016-0012-7]
19. Karduck J, Chapman-Novakofski K. Results of the clinician apps survey, how clinicians working with patients with diabetes and obesity use mobile health apps. J Nutr Educ Behav 2018 Jan;50(1):62-9.e1. [doi: 10.1016/j.jneb.2017.06.004] [Medline: 29325664]
20. Marshall JM, Dunstan DA, Bartik W. Clinical or gimmickal: the use and effectiveness of mobile mental health apps for treating anxiety and depression. Aust N Z J Psychiatry 2020 Jan;54(1):20-28. [doi: 10.1177/0004867419876700] [Medline: 31552747]
21. Ladin K, Porteny T, Perugini JM, Gonzales KM, Aufort KE, Levine SK, et al. Perceptions of telehealth vs in-person visits among older adults with advanced kidney disease, care partners, and clinicians. JAMA Netw Open 2021 Dec 01;4(12):e2137193 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.37193] [Medline: 34870680]
22. Matamala-Gomez M, Bottiroli S, Realdon O, Riva G, Galvagni L, Platz T, et al. Telemedicine and virtual reality at time of COVID-19 pandemic: an overview for future perspectives in neurorehabilitation. Front Neurol 2021;12:646902 [FREE Full text] [doi: 10.3389/fneur.2021.646902] [Medline: 33841313]
23. Krausz M, Ward J, Ramsey D. From telehealth to an interactive virtual clinic. In: Mucic D, Hilty DM, editors. e-Mental Health. Cham, Switzerland: Springer; 2016:289-310.

XSL•FO
RenderX

24. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My diabetes coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 05;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]
25. Viana Ó, Terroso M, Serejo C, Vilaça JL. Ascertaining the influence of style on the credibility and appeal of a digital health avatar. In: Proceedings of the 6th International Conference on Design and Digital Communication. Cham: Springer; 2022 Oct 27 Presented at: DIGICOM '22; November 3–5, 2022; Barcelos, Portugal p. 27 URL: https://link.springer.com/chapter/10.1007/978-3-031-20364-0_6 [doi: 10.1007/978-3-031-20364-0_6]
26. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
27. Bin Sawad A, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, et al. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. Sensors (Basel) 2022 Mar 29;22(7):2625 [FREE Full text] [doi: 10.3390/s22072625] [Medline: 35408238]
28. Balsa J, Neves P, Félix I, Guerreiro MP, Alves P, Carmo MB, et al. Intelligent virtual assistant for promoting behaviour change in older people with T2D. In: Proceedings of the 19th EPIA Conference on Artificial Intelligence on Progress in Artificial Intelligence. 2019 Presented at: EPIA '19; September 3-6, 2019; Vila Real, Portugal p. 372-383 URL: https://link.springer.com/chapter/10.1007/978-3-030-30241-2_32 [doi: 10.1007/978-3-030-30241-2_32]
29. Van Pinxteren MM, Pluymaekers M, Lemmink JG. Human-like communication in conversational agents: a literature review and research agenda. J Serv Manag 2020 Jun 11;31(2):203-225. [doi: 10.1108/josm-06-2019-0175]
30. Potor M. Voice bot vs. chatbot: what's the difference, and why does it matter? Sinch Engage. URL: https://engage.sinch.com/blog/voice-bot-vs-chatbot-whats-the-difference/ [accessed 2024-04-29]
31. Singh S, Thakur HK. Survey of various AI chatbots based on technology used. In: Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions). 2020 Presented at: ICRITO '20; June 4-5, 2020; Noida, India p. 1074-1079 URL: https://ieeexplore.ieee.org/document/9197943 [doi: 10.1109/icrito48877.2020.9197943]
32. Daher K, Casas J, Khaled OA, Mugellini E. Empathic chatbot response for medical assistance. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. 2020 Presented at: IVA '20; October 20-22, 2020; Virtual Event p. 1-3 URL: https://dl.acm.org/doi/10.1145/3383652.3423864 [doi: 10.1145/3383652.3423864]
33. Singh S, Beniwal H. A survey on near-human conversational agents. J King Saud Univ Comput Inf Sci 2022 Nov;34(10):8852-8866. [doi: 10.1016/j.jksuci.2021.10.013]
34. Valtolina S, Hu L. Charlie: a chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In: Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter. 2021 Presented at: CHItaly '21; July 11-13, 2021; Bolzano, Italy p. 1-5 URL: https://dl.acm.org/doi/abs/10.1145/3464385.3464726 [doi: 10.1145/3464385.3464726]
35. Oruganti SC. Virtual bank assistance: an AI based voice bot for better banking. Int J Res 2020;9(1):177-183 [FREE Full text] [doi: 10.13140/RG.2.2.21535.10405]
36. Ma Y, Drewes H, Butz A. Fake moods: can users trick an emotion-aware VoiceBot? In: Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI EA '21; May 8-13, 2021; Yokohama, Japan p. 1-4 URL: https://dl.acm.org/doi/10.1145/3411763.3451744 [doi: 10.1145/3411763.3451744]
37. Eagle T, Blau C, Bales S, Desai N, Li V, Whittaker S. "I don't know what you mean by `I am anxious'": a new method for evaluating conversational agent responses to standardized mental health inputs for anxiety and depression. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-23. [doi: 10.1145/3488057]
38. Magyar G, Balsa J, Cláudio AP, Carmo MB, Neves P, Alves P, et al. Anthropomorphic virtual assistant to support self-care of type 2 diabetes in older people: a perspective on the role of artificial intelligence. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. 2019 Presented at: VISIGRAPP '19; February 25-27, 2019; Pragur, Czech Republic p. 323-331 URL: https://www.scitepress.org/Papers/2019/75724/75724.pdf [doi: 10.5220/0007572403230331]
39. Sestino A, D'Angelo A. My doctor is an avatar! the effect of anthropomorphism and emotional receptivity on individuals' intention to use digital-based healthcare services. Technol Forecast Soc Change 2023 Jun;191:122505. [doi: 10.1016/j.techfore.2023.122505]
40. Shaked NA. Avatars and virtual agents - relationship interfaces for the elderly. Healthc Technol Lett 2017 Jun;4(3):83-87 [FREE Full text] [doi: 10.1049/htl.2017.0009] [Medline: 28706725]
41. Jacob K. What is a digital avatar? DaveAI. URL: https://www.iamdave.ai/blog/what-is-a-digital-avatar/ [accessed 2024-04-29]
42. Seymour M, Yuan L, Dennis AR, Riemer K. Facing the artificial: understanding affinity, trustworthiness, and preference for more realistic digital humans. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020 Presented at: HICSS '20; January 7-10, 2020; Maui, HI p. 4673-4682 URL: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/07da3db0-4dec-40ae-9bd8-550de5636593/content [doi: 10.24251/hicss.2020.574]
43. Bellan R. Are lifelike digital humans the future of customer experience? Techcrunch. URL: https://techcrunch.com/2022/02/14/are-lifelike-digital-humans-the-future-of-customer-experience/ [accessed 2024-04-29]

XSL•FO
RenderX

44. Lukan E. Everything you need to know about digital humans. Firework. URL: https://www.synthesia.io/post/digital-humans [accessed 2024-04-29]
45. Okegbile SD, Cai J. Edge-assisted human-to-virtual twin connectivity scheme for human digital twin frameworks. In: Proceedings of the 95th Conference on Vehicular Technology Conference. 2022 Presented at: VTC '22; June 19-22, 2022; Helsinki, Finland p. 1-6 URL: https://ieeexplore.ieee.org/document/9860619 [doi: 10.1109/vtc2022-spring54318.2022.9860619]
46. Lauer-Schmaltz MW, Cash P, Hansen JP, Maier A. Designing human digital twins for behaviour-changing therapy and rehabilitation: a systematic review. Proc Des Soc 2022 May 26;2:1303-1312. [doi: 10.1017/pds.2022.132]
47. Shubham K, Mukherjee A, Jayagopi DB. Review of realistic behavior and appearance generation in embodied conversational agents: a comparison between traditional and modern approaches. In: Proceedings of the 2022 International Conference on Multimodal Interaction. 2022 Presented at: ICMI '22; November 7-11, 2022; Bengaluru, India p. 191-197 URL: https://dl.acm.org/doi/10.1145/3536221.3556592 [doi: 10.1145/3536221.3556592]
48. Yaqoob I, Salah K, Jayaraman R, Omar M. Metaverse applications in smart cities: enabling technologies, opportunities, challenges, and future directions. Internet Things 2023 Oct;23:100884. [doi: 10.1016/j.iot.2023.100884]
49. Lucas GM, Gratch J, King A, Morency LP. It's only a computer: virtual humans increase willingness to disclose. Comput Human Behav 2014 Aug;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
50. Tulsulkar G, Mishra N, Thalmann NM, Lim HE, Lee MP, Cheng SK. Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? A thorough study based on computer vision methods. Vis Comput 2021;37(12):3019-3038 [FREE Full text] [doi: 10.1007/s00371-021-02242-y] [Medline: 34345091]
51. Mukherjee S, Baral MM, Pal SK, Chittipaka V, Roy R, Alam K. Humanoid robot in healthcare: a systematic review and future research directions. In: Proceedings of the 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing. 2022 Presented at: COM-IT-CON '22; May 26-27, 2022; Faridabad, India p. 822-826 URL: https://ieeexplore.ieee.org/document/9850577 [doi: 10.1109/com-it-con54601.2022.9850577]
52. Linders GM, Vaitonytė J, Alimardani M, Mitev KO, Louwerse MM. A realistic, multimodal virtual agent for the healthcare domain. In: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents. 2022 Presented at: IVA '22; September 6-9, 2022; Faro, Portugal p. 1-3 URL: https://dl.acm.org/doi/abs/10.1145/3514197.3551250 [doi: 10.1145/3514197.3551250]
53. Adams B, Breazeal C, Brooks RA, Scassellati B. Humanoid robots: a new kind of tool. IEEE Intell Syst 2000 Jul;15(4):25-31. [doi: 10.1109/5254.867909]
54. Kaptein F, Kiefer B, Cully A, Celiktutan O, Bierman B, Rijgersberg-peters R, et al. A cloud-based robot system for long-term interaction: principles, implementation, lessons learned. J Hum Robot Interact 2021 Oct 18;11(1):1-27. [doi: 10.1145/3481585]
55. Garner TA, Powell WA, Carr V. Virtual carers for the elderly: a case study review of ethical responsibilities. Digit Health 2016;2:2055207616681173 [FREE Full text] [doi: 10.1177/2055207616681173] [Medline: 29942577]
56. Kusal S, Patil S, Choudrie J, Kotecha K, Mishra S, Abraham A. AI-based conversational agents: a scoping review from technologies to future directions. IEEE Access 2022;10:92337-92356. [doi: 10.1109/access.2022.3201144]
57. Yaghoubzadeh R, Kramer M, Pitsch K, Kopp S. Virtual agents as daily assistants for elderly or cognitively impaired people: studies on acceptance and interaction feasibility. In: Proceedings of the 13th International Conference on Intelligent Virtual Agents. 2013 Presented at: IVA '13; August 29-31, 2013; Edinburgh, UK p. 29-31 URL: https://link.springer.com/chapter/10.1007/978-3-642-40415-3_7 [doi: 10.1007/978-3-642-40415-3_7]
58. Cheng A, Raghavaraju V, Kanugo J, Handrianto YP, Yi S. Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes. In: Proceedings of the 15th IEEE Annual Consumer Communications & Networking Conference. 2018 Presented at: CCNC '18; January 12-15, 2018; Las Vegas, NV p. 1-5 URL: https://ieeexplore.ieee.org/document/8319283 [doi: 10.1109/ccnc.2018.8319283]
59. Thakur N, Han CY. An approach to analyze the social acceptance of virtual assistants by elderly people. In: Proceedings of the 8th International Conference on the Internet of Things. 2018 Presented at: IOT '18; October 15-18, 2018; Santa Barbara, CA p. 1-6 URL: https://dl.acm.org/doi/10.1145/3277593.3277616 [doi: 10.1145/3277593.3277616]
60. Dingler T, Kwasnicka D, Wei J, Gong E, Oldenburg B. The use and promise of conversational agents in digital health. Yearb Med Inform 2021 Aug;30(1):191-199 [FREE Full text] [doi: 10.1055/s-0041-1726510] [Medline: 34479391]
61. Lyzwinski LN, Elgendi M, Menon C. Conversational agents and avatars for cardiometabolic risk factors and lifestyle-related behaviors: scoping review. JMIR Mhealth Uhealth 2023 May 25;11:e39649 [FREE Full text] [doi: 10.2196/39649] [Medline: 37227765]
62. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
63. Monaco A, Palmer K, Holm Ravn Faber N, Kohler I, Silva M, Vatland A, et al. Digital health tools for managing noncommunicable diseases during and after the COVID-19 pandemic: perspectives of patients and caregivers. J Med Internet Res 2021 Jan 29;23(1):e25652 [FREE Full text] [doi: 10.2196/25652] [Medline: 33464206]

XSL•FO
RenderX

64. Guerreiro MP, Angelini L, Rafael Henriques H, El Kamali M, Baixinho C, Balsa J, et al. Conversational agents for health and well-being across the life course: protocol for an evidence map. JMIR Res Protoc 2021 Sep 17;10(9):e26680 [FREE Full text] [doi: 10.2196/26680] [Medline: 34533460]
65. Xing Z, Yu F, Qanir YA, Guan T, Walker J, Song L. Intelligent conversational agents in patient self-management: a systematic survey using multi data sources. Stud Health Technol Inform 2019 Aug 21;264:1813-1814. [doi: 10.3233/SHTI190661] [Medline: 31438357]
66. Griffin AC, Xing Z, Khairat S, Wang Y, Bailey S, Arguello J, et al. Conversational agents for chronic disease self-management: a systematic review. AMIA Annu Symp Proc 2020;2020:504-513 [FREE Full text] [Medline: 33936424]
67. Martinengo L, Jabir AI, Goh WW, Lo NY, Ho MR, Kowatsch T, et al. Conversational agents in health care: scoping review of their behavior change techniques and underpinning theory. J Med Internet Res 2022 Oct 03;24(10):e39243 [FREE Full text] [doi: 10.2196/39243] [Medline: 36190749]
68. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
69. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
70. Even C, Hammann T, Heyl V, Rietz C, Wahl HW, Zentel P, et al. Benefits and challenges of conversational agents in older adults : a scoping review. Z Gerontol Geriatr 2022 Aug;55(5):381-387. [doi: 10.1007/s00391-022-02085-9] [Medline: 35852588]
71. Grudin J. ChatGPT and chat history: challenges for the new wave. Computer 2023 May;56(5):94-100. [doi: 10.1109/mc.2023.3255279]
72. Balsa J, Félix I, Cláudio AP, Carmo MB, Silva IC, Guerreiro A, et al. Usability of an intelligent virtual assistant for promoting behavior change and self-care in older people with type 2 diabetes. J Med Syst 2020 Jun 13;44(7):130. [doi: 10.1007/s10916-020-01583-w] [Medline: 32533367]
73. Stara V, Vera B, Bolliger D, Rossi L, Felici E, Di Rosa M, et al. Usability and acceptance of the embodied conversational agent Anne by people with dementia and their caregivers: exploratory study in home environment settings. JMIR Mhealth Uhealth 2021 Jun 25;9(6):e25891 [FREE Full text] [doi: 10.2196/25891] [Medline: 34170256]
74. Nguyen TT, Sim K, Kuen AT, O'Donnell RR, Lim ST, Wang W, et al. Designing AI-based conversational agent for diabetes care in a multilingual context. arXiv Preprint posted online May 20, 2021 [FREE Full text]
75. Babington-Ashaye A, de Moerloose P, Diop S, Geissbuhler A. Design, development and usability of an educational AI chatbot for people with haemophilia in senegal. Haemophilia 2023 Jul;29(4):1063-1073. [doi: 10.1111/hae.14815] [Medline: 37347648]
76. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019 May 16;9(3):440-447. [doi: 10.1093/tbm/ibz043] [Medline: 31094445]
77. Issom DZ, Rochat J, Hartvigsen G, Lovis C. Preliminary evaluation of a mHealth coaching conversational artificial intelligence for the self-care management of people with sickle-cell disease. Stud Health Technol Inform 2020 Jun 16;270:1361-1362. [doi: 10.3233/SHTI200442] [Medline: 32570659]
78. Rahmanti AR, Yang HC, Bintoro BS, Nursetyo AA, Muhtar MS, Syed-Abdul S, et al. SlimMe, a chatbot with artificial empathy for personal weight management: system design and finding. Front Nutr 2022;9:870775 [FREE Full text] [doi: 10.3389/fnut.2022.870775] [Medline: 35811989]
79. Alhmiedat T, Alotaibi M. Employing social robots for managing diabetes among children: SARA. Wireless Pers Commun 2023 Mar 14;130(1):449-468. [doi: 10.1007/S11277-023-10293-8]
80. Kimani E, Bickmore T, Trinh H, Ring L, Paasche-Orlow MK, Magnani JW. A smartphone-based virtual agent for atrial fibrillation education and counseling. In: Proceedings of the 16th International Conference on Intelligent Virtual Agents. 2016 Presented at: IVA '16; September 20-23, 2016; Los Angeles, CA p. 120-127 URL: https://link.springer.com/chapter/10.1007/978-3-319-47665-0_11 [doi: 10.1007/978-3-319-47665-0_11]
81. Kataoka Y, Takemura T, Sasajima M, Katoh N. Development and early feasibility of chatbots for educating patients with lung cancer and their caregivers in Japan: mixed methods study. JMIR Cancer 2021 Mar 10;7(1):e26911 [FREE Full text] [doi: 10.2196/26911] [Medline: 33688839]
82. Krishnakumar A, Verma R, Chawla R, Sosale A, Saboo B, Joshi S, et al. Evaluating glycemic control in patients of South Asian origin with type 2 diabetes using a digital therapeutic platform: analysis of real-world data. J Med Internet Res 2021 Mar 25;23(3):e17908 [FREE Full text] [doi: 10.2196/17908] [Medline: 33764306]
83. Maharjan R, Doherty K, Rohani DA, Bækgaard P, Bardram JE. Experiences of a speech-enabled conversational agent for the self-report of well-being among people living with affective disorders: an in-the-wild study. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-29. [doi: 10.1145/3484508]
84. Calvo RA, Peters D, Moradbakhti L, Cook D, Rizos G, Schuller B, et al. Assessing the feasibility of a text-based conversational agent for asthma support: protocol for a mixed methods observational study. JMIR Res Protoc 2023 Feb 02;12:e42965 [FREE Full text] [doi: 10.2196/42965] [Medline: 36729586]

XSL•FO
RenderX

85. Neerincx MA, van Vught W, Blanson Henkemans O, Oleari E, Broekens J, Peters R, et al. Socio-cognitive engineering of a robotic partner for child's diabetes self-management. Front Robot AI 2019;6:118 [FREE Full text] [doi: 10.3389/frobt.2019.00118] [Medline: 33501133]
86. Watson A, Bickmore T, Cange A, Kulshreshtha A, Kvedar J. An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial. J Med Internet Res 2012 Jan 26;14(1):e1 [FREE Full text] [doi: 10.2196/jmir.1629] [Medline: 22281837]
87. Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting group medical visits with conversational agents for stress management behavior change. In: Proceedings of the 12th International Conference on Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors. 2017 Presented at: PERSUASIVE '17; April 4-6, 2017; Amsterdam, the Netherlands p. 55-67 URL: https://link.springer.com/chapter/10.1007/978-3-319-55134-0_5
88. Suganuma S, Sakamoto D, Shimoyama H. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. JMIR Ment Health 2018 Jul 31;5(3):e10454 [FREE Full text] [doi: 10.2196/10454] [Medline: 30064969]
89. Hussain S, Athula G. Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In: Proceedings of the 32nd International Conference on Advanced Information Networking and Applications Workshops. 2018 Presented at: WAINA '18; May 16-18, 2018; Krakow, Poland p. 698-703 URL: https://ieeexplore.ieee.org/document/8418155 [doi: 10.1109/waina.2018.00170]
90. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
91. Easton K, Potter S, Bec R, Bennion M, Christensen H, Grindell C, et al. A virtual agent to support individuals living with physical and mental comorbidities: co-design and acceptability testing. J Med Internet Res 2019 May 30;21(5):e12996 [FREE Full text] [doi: 10.2196/12996] [Medline: 31148545]
92. Chaix B, Bibault JE, Pienkowski A, Delamon G, Guillemassé A, Nectoux P, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer 2019 May 02;5(1):e12856 [FREE Full text] [doi: 10.2196/12856] [Medline: 31045505]
93. Anastasiadou M, Alexiadis A, Polychronidou E, Votis K, Tzovaras D. A prototype educational virtual assistant for diabetes management. In: Proceedings of the 20th International Conference on Bioinformatics and Bioengineering. 2020 Presented at: BIBE '20; October 26-28, 2020; Cincinnati, OH p. 999-1004 URL: https://ieeexplore.ieee.org/document/9288129 [doi: 10.1109/bibe50027.2020.00169]
94. Roca S, Lozano ML, García J, Alesanco Á. Validation of a virtual assistant for improving medication adherence in patients with comorbid type 2 diabetes mellitus and depressive disorder. Int J Environ Res Public Health 2021 Nov 17;18(22):12056 [FREE Full text] [doi: 10.3390/ijerph182212056] [Medline: 34831811]
95. Yao K, Wong KK, Yu X, Volpi J, Wong ST. An intelligent augmented lifelike avatar app for virtual physical examination of suspected strokes. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:1727-1730. [doi: 10.1109/EMBC46164.2021.9629720] [Medline: 34891620]
96. Egede JO, Price D, Krishnan DB, Jaiswal S, Elliott N, Morriss R. Design and evaluation of virtual human mediated tasks for assessment of depression and anxiety. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. 2021 Presented at: IVA '21; September 14-17, 2021; Virtual Event p. 52-59 URL: https://dl.acm.org/doi/10.1145/3472306.3478361 [doi: 10.1145/3472306.3478361]
97. Romanovskyi O, Pidbutska N, Knysh A. Elomia chatbot: the effectiveness of artificial intelligence in the fight for mental health. In: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems. 2021 Presented at: COLINS '21; April 22-23, 2021; Kharkiv, Ukraine p. 1-10 URL: https://ceur-ws.org/Vol-2870/paper89.pdf
98. Anan T, Kajiki S, Oka H, Fujii T, Kawamata K, Mori K, et al. Effects of an artificial intelligence-assisted health program on workers with neck/shoulder pain/stiffness and low back pain: randomized controlled trial. JMIR Mhealth Uhealth 2021 Sep 24;9(9):e27535 [FREE Full text] [doi: 10.2196/27535] [Medline: 34559054]
99. Zisis G, Carrington MJ, Oldenburg B, Whitmore K, Lay M, Huynh Q, et al. An m-Health intervention to improve education, self-management, and outcomes in patients admitted for acute decompensated heart failure: barriers to effective implementation. Eur Heart J Digit Health 2021 Dec;2(4):649-657 [FREE Full text] [doi: 10.1093/ehjdh/ztab085] [Medline: 36713108]
100. Apergi LA, Bjarnadottir MV, Baras JS, Golden BL, Anderson KM, Chou J, et al. Voice interface technology adoption by patients with heart failure: pilot comparison study. JMIR Mhealth Uhealth 2021 Apr 01;9(4):e24646 [FREE Full text] [doi: 10.2196/24646] [Medline: 33792556]
101. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel) 2022 May 11;22(10):3653 [FREE Full text] [doi: 10.3390/s22103653] [Medline: 35632061]

XSL•FO
RenderX

102. Kannampallil T, Ronneberg CR, Wittels NE, Kumar V, Lv N, Smyth JM, et al. Design and formative evaluation of a virtual voice-based coach for problem-solving treatment: observational study. JMIR Form Res 2022 Aug 12;6(8):e38092 [FREE Full text] [doi: 10.2196/38092] [Medline: 35969431]
103. Alturaiki AM, Banjar HR, Barefah AS, Alnajjar SA, Hindawi S. A smart chatbot for interactive management in beta thalassemia patients. Int J Telemed Appl 2022;2022:9734518 [FREE Full text] [doi: 10.1155/2022/9734518] [Medline: 35601050]
104. Zahedi FM, Zhao H, Sanvanson P, Walia N, Jain H, Shaker R. My real avatar has a doctor appointment in the Wepital: a system for persistent, efficient, and ubiquitous medical care. Inf Manag 2022 Dec;59(8):103706. [doi: 10.1016/j.im.2022.103706]
105. Meheli S, Sinha C, Kadaba M. Understanding people with chronic pain who use a cognitive behavioral therapy-based artificial intelligence mental health app (Wysa): mixed methods retrospective observational study. JMIR Hum Factors 2022 Apr 27;9(2):e35671 [FREE Full text] [doi: 10.2196/35671] [Medline: 35314422]
106. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA. Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management. Am J Gastroenterol 2023 Dec 01;118(12):2276-2279. [doi: 10.14309/ajg.0000000000002397] [Medline: 37410934]
107. Epalte K, Tomsone S, Vētra A, Bērziņa G. Patient experience using digital therapy "Vigo" for stroke patient recovery: a qualitative descriptive study. Disabil Rehabil Assist Technol 2023 Feb;18(2):175-184. [doi: 10.1080/17483107.2020.1839794] [Medline: 33155507]
108. LeRouge C, Dickhut K, Lisetti C, Sangameswaran S, Malasanos T. Engaging adolescents in a computer-based weight management program: avatars and virtual coaches could help. J Am Med Inform Assoc 2016 Jan;23(1):19-28 [FREE Full text] [doi: 10.1093/jamia/ocv078] [Medline: 26159466]
109. Park G, Chung J, Lee S. Human vs. machine-like representation in chatbot mental health counseling: the serial mediation of psychological distance and trust on compliance intention. Curr Psychol 2023 Apr 20:1-12 [FREE Full text] [doi: 10.1007/s12144-023-04653-7] [Medline: 37359642]
110. Boggiss A, Consedine N, Hopkins S, Silvester C, Jefferies C, Hofman P, et al. Improving the well-being of adolescents with type 1 diabetes during the COVID-19 pandemic: qualitative study exploring acceptability and clinical usability of a self-compassion chatbot. JMIR Diabetes 2023 May 05;8:e40641 [FREE Full text] [doi: 10.2196/40641] [Medline: 36939680]
111. Kocaballi AB, Quiroz JC, Rezazadegan D, Berkovsky S, Magrabi F, Coiera E, et al. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. J Med Internet Res 2020 Feb 09;22(2):e15823 [FREE Full text] [doi: 10.2196/15823] [Medline: 32039810]
112. Wiratunga N, Cooper K, Wijekoon A, Palihawadana C, Mendham V, Reiter E, et al. FitChat: conversational artificial intelligence interventions for encouraging physical activity in older adults. arXiv Preprint posted online April 29, 2020 [FREE Full text] [doi: 10.48550/arXiv.2004.14067]
113. Denecke K, Tschanz M, Dorner TL, May R. Intelligent conversational agents in healthcare: hype or hope? Stud Health Technol Inform 2019;259:77-84. [Medline: 30923277]
114. Kim K, Norouzi N, Losekamp T, Bruder G, Anderson M, Welch G. Effects of patient care assistant embodiment and computer mediation on user experience. In: Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality. 2019 Presented at: AIVR '19; December 9-11, 2019; San Diego, CA p. 17-177 URL: https://ieeexplore.ieee.org/document/8942267 [doi: 10.1109/aivr46125.2019.00013]
115. Sen A, Liew SH. Augmented reality and its use in education. In: Tatnall A, editor. Encyclopedia of Education and Information Technologies. Cham, Switzerland: Springer; 2018:1719-1726.
116. Kim K, Boelling L, Haesler S, Bailenson J, Bruder G, Welch GF. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In: Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality. 2018 Presented at: ISMAR '18; October 16-20, 2018; Munich, Germany p. 105-115 URL: https://ieeexplore.ieee.org/document/8613756 [doi: 10.1109/ISMAR.2018.00039]
117. Li Q, Luximon Y, Zhang J. The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: within-subject web-based experiment. J Med Internet Res 2023 Aug 10;25:e44479 [FREE Full text] [doi: 10.2196/44479] [Medline: 37561567]
118. Grunitz M. Rule-based AI vs machine learning: what's the difference? WeAreBrain. 2021 Sep. URL: https://wearebrain.com/blog/rule-based-ai-vs-machine-learning-whats-the-difference/ [accessed 2024-04-29]
119. Ahmad R, Siemon D, Gnewuch U, Robra-Bissantz S. The benefits and caveats of personality-adaptive conversational agents in mental health care. In: Proceedings of the 27th annual Americas Conference on Information Systems. 2021 Presented at: AMCIS '21; August 9-13, 2021; Virtual Event URL: https://web.archive.org/web/20220804131608id_/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1225&context=amcis2021
120. Siemon D, Ahmad R, Harms H, de Vreede T. Requirements and solution approaches to personality-adaptive conversational agents in mental health care. Sustainability 2022;14(7):3832 [FREE Full text]
121. Kang Y, Tan AH, Miao C. An adaptive computational model for personalized persuasion. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence. 2015 Presented at: IJCAI '15; July 25–31, 2015; Buenos Aires, Argentina p. 61-67 URL: https://www.ijcai.org/Proceedings/15/Papers/016.pdf

XSL•FO
RenderX

122. Faddoul G. The virtual diabetician: a virtual platform providing diabetes treatment information through storytelling and virtual companions. Center of Information Systems and Technology Claremont Graduate University. 2018. URL: https://www.proquest.com/openview/6b73f74fdd97109ee35723b289ea2b96/1?pq-origsite=gscholar&cbl=18750 [accessed 2024-04-29]
123. Tironi A, Mainetti R, Pezzera M, Borghese NA. An empathic virtual caregiver for assistance in exer-game-based rehabilitation therapie. In: Proceedings of the IEEE 7th International Conference on Serious Games and Applications for Health. 2019 Presented at: SeGAH '19; August 5-7, 2019; Kyoto, Japan p. 1-6 URL: https://ieeexplore.ieee.org/document/8882477
124. Casas J, Spring T, Daher K, Mugellini E, Khaled OA, Cudre-Mauroux P. Enhancing conversational agents with empathic abilities. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. 2021 Presented at: IVA '21; September 14-17, 2021; Virtual Event p. 41-47 URL: https://dl.acm.org/doi/10.1145/3472306.3478344 [doi: 10.1145/3472306.3478344]
125. May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Inform Health Soc Care 2022 Apr 03;47(2):194-210. [doi: 10.1080/17538157.2021.1983578] [Medline: 34617857]
126. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
127. Ranieri A, Ruggiero A. Complementary role of conversational agents in e-health services. In: Proceedings of the 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering. 2022 Presented at: MetroXRAINE '22; October 26-28, 2022; Rome, Italy p. 528-533 URL: https://ieeexplore.ieee.org/document/9967603 [doi: 10.7759/cureus.40922]
128. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. Cureus 2023 Jun;15(6):e40922 [FREE Full text] [doi: 10.7759/cureus.40922] [Medline: 37496532]
129. Sen A, Chuen CL, Hta AC. Toward smart learning environments: affordances and design architecture of augmented reality (AR) applications in medical education. In: Proceedings of 1st International Conference on Smart System, Innovations and Computing. 2019 Presented at: SSIC '17; April 14-16, 2017; Jaipur, India p. 843-861 URL: https://link.springer.com/chapter/10.1007/978-981-10-5828-8_80 [doi: 10.1177/2055207619896153]
130. Sen A, Leong C. Technology-enhanced learning. In: Tatnall A, editor. Encyclopedia of Education and Information Technologies. Cham, Switzerland: Springer; 2019:1-8.
131. Hossain MM, Tasnim S, Sharma R, Sultana A, Shaik AF, Faizah F, et al. Digital interventions for people living with non-communicable diseases in India: a systematic review of intervention studies and recommendations for future research and development. Digit Health 2019;5:2055207619896153 [FREE Full text] [doi: 10.1177/2055207619896153] [Medline: 31897307]
132. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. JMIR Mhealth Uhealth 2020 Jul 10;8(7):e17558 [FREE Full text] [doi: 10.2196/17558] [Medline: 32673246]
133. Jegundo AL, Dantas C, Quintas J, Dutra J, Almeida AL, Caravau H, et al. Perceived usefulness, satisfaction, ease of use and potential of a virtual companion to support the care provision for older adults. Technologies 2020 Jul 25;8(3):42 [FREE Full text] [doi: 10.3390/technologies8030042]
134. Nagarajan A, Sen A. Can Bloom's higher order thinking skills be achieved by gamified learning through social networking sites (SNS) like Facebook? Interact Des Archit 2022(53):144-160 [FREE Full text]
135. Sen A, Richardson S. Some controversies relating to the causes and preventive management of computer vision syndrome. In: Proceeding of the 2002 conference on CybErg. 2002 Presented at: CybErg '02; September 2-4, 2002; Johannesburg, South Africa p. 123-130 URL: https://research.monash.edu/en/publications/some-controversies-relating-to-the-causes-and-preventive-manageme [doi: 10.1109/icbir54589.2022.9786451]
136. Yang S, Zhao H, Zhu S, Zhou G, Xu H, Jia Y, et al. Zhongjing: enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. arXiv Preprint posted online August 7, 2023 [FREE Full text]
137. Wang J, Yao Z, Yang Z, Zhou H, Li R, Wang X, et al. NoteChat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. arXiv Preprint posted online October 24, 2023 [FREE Full text]
138. Kanzaki K, Isahara H. Analysis of dialogue between caregivers and recipients for a communication robot. In: Proceedings of the 7th International Conference on Business and Industrial Research. 2022 Presented at: ICBIR '22; May 19-20, 2022; Bangkok, Thailand p. 446-450 URL: https://ieeexplore.ieee.org/document/9786451 [doi: 10.1109/icbir54589.2022.9786451]
139. Huang CW, Wu BC, Nguyen PA, Wang HH, Kao CC, Lee PC, et al. Emotion recognition in doctor-patient interactions from real-world clinical video database: initial development of artificial empathy. Comput Methods Programs Biomed 2023 May;233:107480. [doi: 10.1016/j.cmpb.2023.107480] [Medline: 36965299]
140. Lewis L. Avatars and robots as social companions in healthcare: requirements, engineering, adoption and ethics. Int J Enterp Inf Syst 2014;10(2):21-39. [doi: 10.4018/ijeis.2014040102]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CA: conversational agent
CVD: cardiovascular disease
LMIC: low- and middle-income country
ML: machine learning
NCD: noncommunicable disease
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by T de Azevedo Cardoso; submitted 08.01.24; peer-reviewed by M Kim, T Su, E Hekler, DLT Steven; comments to author 05.02.24; revised version received 06.03.24; accepted 25.03.24; published 16.07.24.

Please cite as:
Anisha SA, Sen A, Bain C
Evaluating the Potential and Pitfalls of AI-Powered Conversational Agents as Humanlike Virtual Health Carers in the Remote Management of Noncommunicable Diseases: Scoping Review
J Med Internet Res 2024;26:e56114
URL: https://www.jmir.org/2024/1/e56114
doi:10.2196/56114
PMID:39012688

©Sadia Azmin Anisha, Arkendu Sen, Chris Bain. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Health Interventions for Chronic Wound Management: A Systematic Review and Meta-Analysis

Xinrui Bai[1*], BSN; Hongyan Zhang[2*], MSN; Yanxia Jiao[1], BSN; Chenlu Yuan[1], BSN; Yuxia Ma[1], MSN; Lin Han[1,2], PhD

[1]Evidence-Based Nursing Centre, School of Nursing, Lanzhou University, Lanzhou, Gansu, China

[2]Department of Nursing, Gansu Provincial Hospital, Lanzhou, Gansu, China

[*]these authors contributed equally

Corresponding Author:
Lin Han, PhD
Department of Nursing
Gansu Provincial Hospital
#204 Dong gang Road
Chengguan District
Lanzhou, Gansu, 730030
China
Phone: 86 0931 8281578
Email: LZU-hanlin@hotmail.com

Abstract

Background: Digital health interventions (DHIs) have shown promising results for the management of chronic wounds. However, its effectiveness compared to usual care and whether variability in the type of intervention affects wound outcomes are unclear.

Objective: The main objective was to determine the effectiveness of DHIs on wound healing outcomes in adult patients with chronic wounds. The secondary objectives were to assess if there was any variation in wound healing outcomes across the various types of DHIs.

Methods: In total, 9 databases were searched for the literature up to August 1, 2023. Randomized controlled trials (RCTs), cohort studies, and quasi-experimental studies comparing the efficacy of DHIs with controls in improving wound outcomes in adult patients with chronic wounds were included. Study selection, data extraction, and risk of bias assessment were conducted independently by 2 reviewers. We assessed the quality of each RCT, cohort study, and quasi-experimental study separately using the Cochrane risk of bias tool, ROBINS-I, and the Joanna Briggs Institute Critical Appraisal tools checklists. Relative risks (RRs) and 95% CIs were pooled using the random effects model, and heterogeneity was assessed by the I^2 statistic. Subgroup analysis and sensitivity analysis were also performed.

Results: A total of 25 studies with 8125 patients were included in this systematic review, while only 20 studies with 6535 patients were included in the meta-analysis. Efficacy outcomes in RCTs showed no significant differences between the DHIs and control groups in terms of wound healing (RR 1.02, 95% *CI* 0.93-1.12; *P*=.67) and all-cause mortality around 1 year (RR 1.08, 95% *CI* 0.55-2.12; *P*=.83). Compared with the control group, the use of DHIs was associated with significant changes in adverse events (RR 0.44, 95% *CI* 0.22-0.89; *P*=.02). Subgroup analysis suggested a positive effect of the digital platforms in improving wound healing (RR 2.19, 95% CI 1.35-3.56; *P*=.002). Although meta-analysis was not possible in terms of wound size, cost analysis, patient satisfaction, and wound reporting rates, most studies still demonstrated that DHIs were not inferior to usual care in managing chronic wounds.

Conclusions: The findings of our study demonstrate the viability of adopting DHIs to manage chronic wounds. However, more prominent, high-quality RCTs are needed to strengthen the evidence, and more detailed clinical efficacy research is required.

Trial Registration: PROSPERO CRD42023392415; https://tinyurl.com/4ybz6bs9

(J Med Internet Res 2024;26:e47904) doi:10.2196/47904

KEYWORDS

chronic wounds; digital health interventions; wound healing; meta-analysis; systematic review; digital technologies; mobile health; eHealth; telemedicine; telehealth

XSL•FO
RenderX

Introduction

Chronic wounds have a substantial impact on individual health, society, and health systems worldwide [1], with studies showing that the global prevalence of chronic wounds is estimated to be 1.67 per 1000 population [2]. A commonly used definition, labeling chronic wounds as wounds that "fail to proceed through an orderly and timely process to produce anatomic and functional integrity" [3]. Chronic wounds are classified by etiology and include, but are not limited to, lower extremity venous ulcers, neurological ulcers, diabetic foot ulcers, pressure injuries, and arterial ulcers [4], and these underlying pathologic factors often hinder or delay the healing process, resulting in significant negative impacts on the physical, emotional, and social well-being of patients. Many patients develop infections due to poor chronic wound management, experiencing increased pain, delayed wound healing, and even wound rupture and foul odors [5-7], imposing humanistic burdens (eg, health-related quality of life [8,9]) and economic costs (including direct expenditures, such as medical bills, and indirect lost productivity, such as sick leave and early retirement [10]) on both the patients themselves and society. It is crucial to highlight that there is no agreement on a specific healing period for chronic wounds, which means there is no set timeline for wound healing or when a wound becomes chronic; as a result, those suffering often require prolonged care [11].

Traditional wound care is mainly done in hospitals or specialized treatment facilities, and it is limited by time and location, treatment space, patients' financial status, and the scope of the medical center's services [10,12]. Patients, for example, frequently need to plan ahead of time for treatment, and those who reside in remote regions or have restricted mobility may face greater burdens [13]; so, many patients with chronic wounds remain without adequate wound management options. Consequently, even with the current advancements in in-hospital chronic wound management, it is still a vital challenge for the health care sector to address how to deliver a wound management program for patients with access challenges that is no less than the quality of in-hospital care without placing a greater financial burden on patients.

Digital health interventions (DHIs) have been recognized by the World Health Organization South-East Asia Regional Organization (WHO-SEARO) for their role in improving access to primary health care, may be a promising option for overcoming these barriers. According to the WHO definition, DHI is a discrete functionality of digital technology that is applied to achieve health objectives [14]. DHIs included, but were not limited to, devices used to deliver the intervention, such as mobile phones, mobile apps, portable tablets, web-based platforms, and activity trackers. DHIs are widely used for assessment, education, and symptom management in patients with a variety of disorders such as cancer, diabetes, stroke, and attention-deficit/hyperactivity disorder (ADHD) [15-19], with promising results in chronic wound management [20]. Remote consultation and follow-up via phone and email connect home-care nurses to wound experts, increasing the likelihood of wound healing [21-23]. The wound digital platform designed for inpatients has shown positive intervention outcomes [24,25]. Meanwhile, a novel study [26] developed the framework of the digital nursing quality management model and used digital wound care as an example in the conceptualization process, heralding the great potential of DHIs in the field of chronic wound management. Previous meta-analysis shows that DHIs show no inferiority in randomized trials compared with traditional face-to-face care. However, these meta-analyses have limitations: either only chronic wounds of 1 etiology have been considered [27,28], or they have looked at in control groups, such as community-based or nursing home–based interventions [29], whereas no studies have yet noted differences in DHIs.

Thus, an updated meta-analysis is warranted. The primary aim of this study is to assess the efficacy of DHIs for chronic wound management versus usual care. Our secondary aim is to explore whether and how modifiable types of DHIs (eg, digital platforms, telemedicine, or follow-up by telephone and email) affect chronic wound healing outcomes. These insights serve to inform existing or novel chronic wound-targeted treatment protocols and develop optimal treatment options for the benefit of patients.

Methods

This systematic review is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement (Multimedia Appendix 1) [30]. Before the start of the study, the review protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD42023392415).

Definitions and Categories of DHIs and Usual Care

The theoretical definition of DHI refers to a discrete functionality of digital technology that is applied to achieve health objectives. Telemedicine, digital platforms, and mobile phone and SMS text messaging follow-up are all examples of DHIs [31].

Although the definition of usual care has not been standardized, it can include the routine care received by patients for prevention or treatment of diseases [32]. In this study, in addition to regular hospital care, usual care forms consisted of the following 3 main categories and their collocated use: outpatient clinics, primary care, and home care. The operational definitions or meanings of several types of DHIs and usual care covered in this paper are explained in Table 1.

XSL•FO
RenderX

Table 1. Operational classification and definitions of digital health interventions and usual care.

Category	Operational definition or meaning
Digital Health Intervention[a]	
Telemedicine	The use of electronic technology for information and communication by health care professionals with patients (or caregivers), with the objective of providing and supporting medical care to patients when they are away from health care institutions [33].
Digital platform	Intermediaries that enable 2 or more customer or supplier and user groups to interact [34].
Follow-up by telephone and email	A method in which health care professionals evaluates a patient's health status along with offering periodic care using verbal descriptions over the phone or electronic information sent via mail.
Usual care	
Outpatient clinic	One of the most important departments of the hospital, where most elective care trajectories begin, with a consultation between a care provider and a patient [35].
Primary care	The provision of integrated, accessible health care services by clinicians who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients, and practicing in the context of family and community [36].
Home care	Medical and paramedical services delivered to patients at home [37].

[a]Only the digital health intervention categories addressed in this systematic review were explained.

Search Strategies

The 2 researchers XB and HZ performed an independent electronic search in PubMed, Web of Science, EMBASE, Cochrane Library, CINAHL, China Knowledge Resource Integrated Database, Wanfang Database, Weipu Database, and Chinese Biomedical from their inception to August 1, 2023. The search strategy was developed using the PICO search framework (ie, patient, intervention, comparison, outcome, and study design) and the search terms were divided into 3 categories: patients with chronic wounds (population), digital health (intervention), and wound status (outcome). Each category was combined with MeSH (Medical Subject Headings) terms and natural language (Multimedia Appendix 2).

Eligibility Criteria

The inclusion criteria of this study, which followed the PICOS (Population, Intervention, Comparison, Outcomes, and Study) design framework [38], were as follows: (1) Population: adults aged 18 years or older with chronic wounds of any type and severity. Chronic wound was defined as a wound that "fail to proceed through an orderly and timely process to produce anatomical and functional integrity" and follow the 4 major classifications of the Wound Healing Society, ie, pressure ulcers, diabetic foot ulcers, venous ulcers, and arterial insufficiency ulcers [6]; (2) Interventions: study interventions must have used a DHI to capture, assess, create, or communicate wound status in patients with chronic wounds. DHIs were defined as "health promotion approaches aided by various digital technologies," such as remote consultations, mobile applications, web-based platforms, mobile phones, wearables, SMS, and email; (3) Comparison: studies that assigned participants into either an experimental group or a control group including usual, routine, and conventional care, or waitlist as defined by the original research; (4) Outcomes: the primary outcome was the state of chronic wound healing (eg, wound healing, healing time, and change in wound size), and secondary outcomes include all-cause mortality, adverse events, patient satisfaction, and certain wound-specific scoring metrics; and (5) Study design: randomized controlled trials (RCTs), cohort studies, and quasi-experimental studies published in English and Chinese.

Articles were excluded if their main objective was to assess the acceptability of a newly developed DHI among patients or if the study did not contain a control group. Conference proceedings, magazines, news, electronic resources and reports, theses, dissertations, abstracts, editorials, and systematic reviews were also excluded.

Study Selection

Initially, search duplicates were removed using the reference manager tool EndNote (version X9.3.3; Clarivate). For final inclusion, each study was assessed independently by 2 researchers (XB and HZ), first by screening the title and abstract, and then through a full-text review. Disagreements on the selection of records between the 2 researchers were resolved by team discussion or by a third researcher (LH). In addition, we manually searched the reference lists of the included studies for additional studies. The PRISMA flow diagram was used to illustrate the study selection (refer to Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. CNKI: China Knowledge Resource Integrated Database; VIP: Weipu Database.

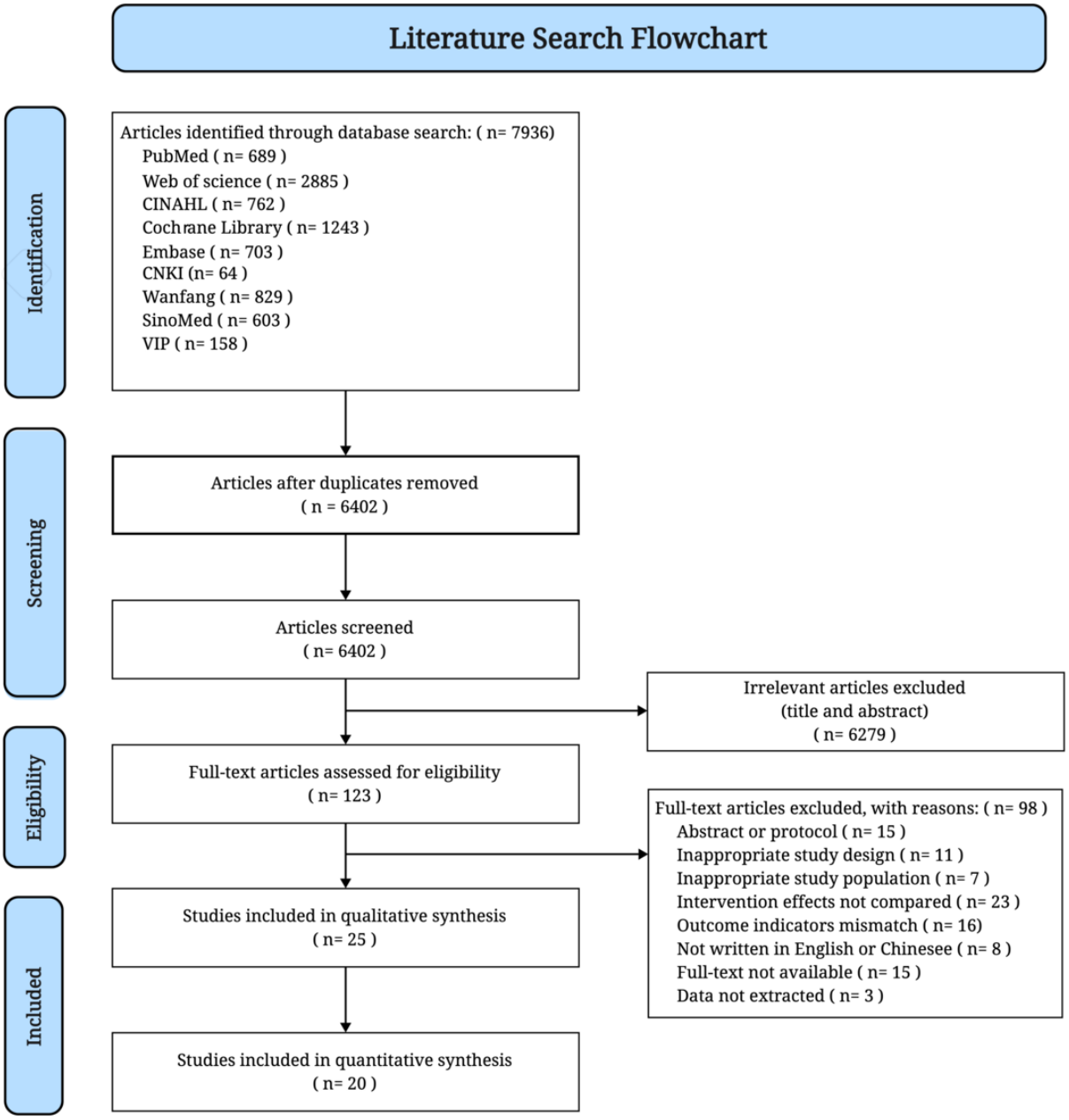

Data Extraction

XB and HZ independently extracted the data using a pre-designed form created with Excel. Disagreements were resolved through discussion or with assistance from a third reviewer (LH), if necessary. From each study, we extracted information about study details (eg, title, author, year, and country), study design (type of study, aims, sample methods, and inclusion or exclusion criteria), participants' characteristics (number of persons surveyed, population characteristics, ie, age, gender, and demographics), specific wound data and complications (diagnostic methods, ulcer specifications including stage), and outcomes (eg, wound healing, healing time, wound size change, adverse event, all-cause mortality, cost, and patients' satisfaction).

Assessment of Risk of Bias

The risk of bias for each study was assessed by 2 independent reviewers (XB and HZ). For RCTs, the Cochrane risk-of-bias tool [39] was used to assign a "high risk," "unclear risk," or "low risk" according to 5 domains: (1) sequence generation, (2) allocation concealment, (3) blinding, (4) incomplete outcome data, and (5) selective outcome reporting. We used ROBINS-I [40] for cohort studies, which assesses 7 types of bias from confounding variables, selection of participants, measurement classification of interventions, deviations from intended interventions, incomplete outcome data, outcome assessment, and selective outcome reporting. For quasi-experimental studies, the checklists of the Joanna Briggs Institute Critical Appraisal tools were used, comprising 9 items that can be rated yes, no, unclear, or not applicable.

XSL•FO
RenderX

Statistical Analysis

When possible, outcome data were analyzed quantitatively by calculating a pooled effect of different studies. A meta-analysis was performed when ≥2 studies with available data investigated the same outcome; otherwise, the outcomes were presented narratively. Considering the expected clinical heterogeneity among the included studies, the random-effects model was used to estimate the overall effect size and 95% *CI*. The pooled analysis was presented as a risk ratio (RR). The effect of the intervention on continuous outcomes was expressed as standardized mean difference (SMD) for outcomes reported with different measures and with 95% *CI* [41].

Heterogeneity among studies was estimated Cochran *Q* test and the I^2 statistic (I^2>50% indicated substantial heterogeneity) [42]. We also performed sensitivity analyses to examine the robustness of the results. Sensitivity analysis was carried out by removing a single study at a time to see how it impacted the overall estimate. For chronic wounds' outcomes, the forest plots were also constructed. When the number of included studies was more than 10 [43], the graphical symmetry of the funnel plot and the Egger test statistic were used to detect possible publication bias [44], because the power of the test is lower when the number of studies is small. In the absence of publication bias, the funnel plot is expected to be symmetrical, and the *P* values of Egger's test are >.05 [45,46].

In addition, RCT and observational data were analyzed separately, and exploratory subgroup analyses were carried out for different modalities of DHI. All analyses were performed with RevMan (version 5.4.1; The Cochrane Collaboration) and Stata (version 15.0; StataCorp; XB and HZ).

Results

Out of a total of 7936 potential articles identified in the literature search, 123 studies were selected for a full review. Finally, the systematic review included 25 studies involving a total of 8125 patients: 14 (56%) presented the results of RCTs [22,23,47-58], 6 (24%) cohort studies [21,59-63], and 5 (20%) quasi-experimental studies [24,25,64-66]. Figure 1 depicts the results of the study selection.

Study Characteristics

In total, 25 studies were published between 2004 and 2023, with 9 (36%) conducted in China [24,25,50,54-56,64-66], 3 (12%) in Norway and France [52,53,57,58,60,61], and 2 (8%) in the United States [48,59], Australia [47,51], and Denmark [21,49], respectively. In addition, the following countries had 1 (4%) of the studies each, the United Kingdom [22], Canada [23], Sweden [62], and Israel [63]. Sample sizes varied across studies, ranging from the smallest sample of 26 subjects in Vowden and Vowden [22] to the largest of 1988 subjects in Wickstrom et al [62]. A total of 40% (10/25) of the studies were conducted on patients with chronic wounds of various etiologies [22,47,48,53,55,56,60-62,66], 32% (8/25) focused on patients with pressure injuries [23-25,51,54,57,64,65], 16% (4/25) targeted patients with diabetic foot ulcers [49,52,58,59], 4% (1/25) recruited patients with lower extremity ulcers of various etiologies [63], 4% (1/25) included only patients with lower extremity venous ulcers [50], and a further 4% (1/25) excluded patients with pressure ulcers, surgical wounds, and cancer wounds [21].

Wound care in experimental groups varied, because of the clinical heterogeneity of DHIs between studies, with 15 (60%) studies collecting patients' wound data via telemedicine, such as web-based programs and video conference [21,22,47-50,53,54,57-59,61-63], 7 (28%) studies using the digital platform [24,25,55,56,64-66], and 3 (12%) studies using email and phone to facilitate the implementation. The follow-up periods were inconsistent among all studies and, where present, ranged from 3 to 35 months. In addition to hospitalization, the control group received wound usual care in a variety of settings, including outpatient clinics (14/25, 56%) [23-25,47-52,63-67], the home (3/25, 12%) [21,22,60], the community (2/25, 8%) [61,62], and home care combined with outpatient follow-up (1/25, 4%) [59]. With the exception of 2 (8%) studies, the remaining 23 (92%) studies were divided into 2 groups, where 3 groups were compared by Terry et al [48], group A (weekly visits with TM and wound care specialist consults), group B (weekly visits with weekly consults with WCS), and group C (routine care). Téot et al [53] divided the control group into group 1 (telemedicine), group 2a (home care), and group 2b (clinic care). Details of all 25 studies are summarized in Table S1 in Multimedia Appendix 3 [21-25,47-66].

Risk of Bias

Details of the assessment of the risk of bias are presented in Tables S2-S4 in Multimedia Appendix 4 [21-25,47-66]. Of the 14 RCTs, 4 (29%) RCTs were assessed as high risk in the "other bias" option, owing to unequal baseline characteristics [22,48,52,53], and 6 (43%) studies [49-51,55,57,58] demonstrated a low risk of bias, with bias in ≤1 domain.

The assessment result for cohort studies revealed that 2 (33%) studies [21,60] had a moderate risk of bias, while 4 (67%) studies [59,61-63] had a high risk of bias.

Finally, for the quasi-experimental study, all study items related to the integrity of follow-up were rated as not applicable, and the items were rated as unclear mostly because it was unknown whether the other measures received by the groups were identical.

Outcome Analysis

Wound Healing

Of the 25 studies, 10 (40%) studies [21-23,48,49,52,53,57,61,62] reported wound healing around 1 year, including 7 (28%) RCTs and 3 (12%) cohort studies. One (4%) RCT [48] was not included in the quantitative synthesis because of the uneven distribution of severity of wounds among groups. Pooled data in 9 (36%) studies revealed no significant difference in wound healing (RR 1.15, 95% CI 0.94-1.40; *P*=.17; I^2=85%; Figure 2), and the finding for the pooled is consistent in RCTs (RR 1.02, 95% CI 0.93-1.12; *P*=.67; I^2=12%) and cohort studies (RR 1.32, 95% CI 0.90-1.95; *P*=.15; I^2=81%).

A total of 12% (3/25) of the studies reported wound healing at 3 months, including 1 RCT and 2 cohort studies [50,59,60]. Pooling the data showed that wound healing at 3 months seems

XSL•FO
RenderX

superior in the DHIs group than the control group with usual care, but no statistically significant difference was observed (RR 1.44, 95% CI 0.51-4.05; P=.49; I^2=75%; Multimedia Appendix 5).

Additionally, one study [58] found that the intervention group's ulcer healing rate was 62.1%, which was greater than the control group's rate of 52.4%. Furthermore, the study by Wu and Fu [65] showed an increase in wound healing rate after the intervention; Santamaria et al [47] reported positive wound healing rates per week in the intervention group as well as negative rates per week in the control group; another study [63] defined a positive outcome as at least 50% ulcer closure. None of these 4 (16%) studies were appropriate for inclusion in the quantitative analysis.

The studies were separated into 3 subgroups: telemedicine, email and telephone follow-up, and digital platform, based on the clinical heterogeneity of the types of DHIs in the intervention groups. Providers also varied in the telemedicine model, with 73% (11/15) of the studies involved nurses as interveners [22,47,49,50,54,58-63], 20% of studies involved wound care specialists (3/15) [21,48,53] and a further 7% (1/15) with a multidisciplinary team responsible for the intervention [57]. Subgroup data demonstrated a statistically significant difference between the DHIs and control groups when the digital platform was used (RR 2.19, 95% CI 1.35-3.56; P=.002; I^2=82%; Figure 3). However, there was no significant change in wound healing in the groups with telemedicine (RR 1.15, 95% CI 0.91-1.45; P=.24; I^2=86%).

Figure 2. Forest plot of digital health interventions (DHIs) on wound healing around one year (different study types).

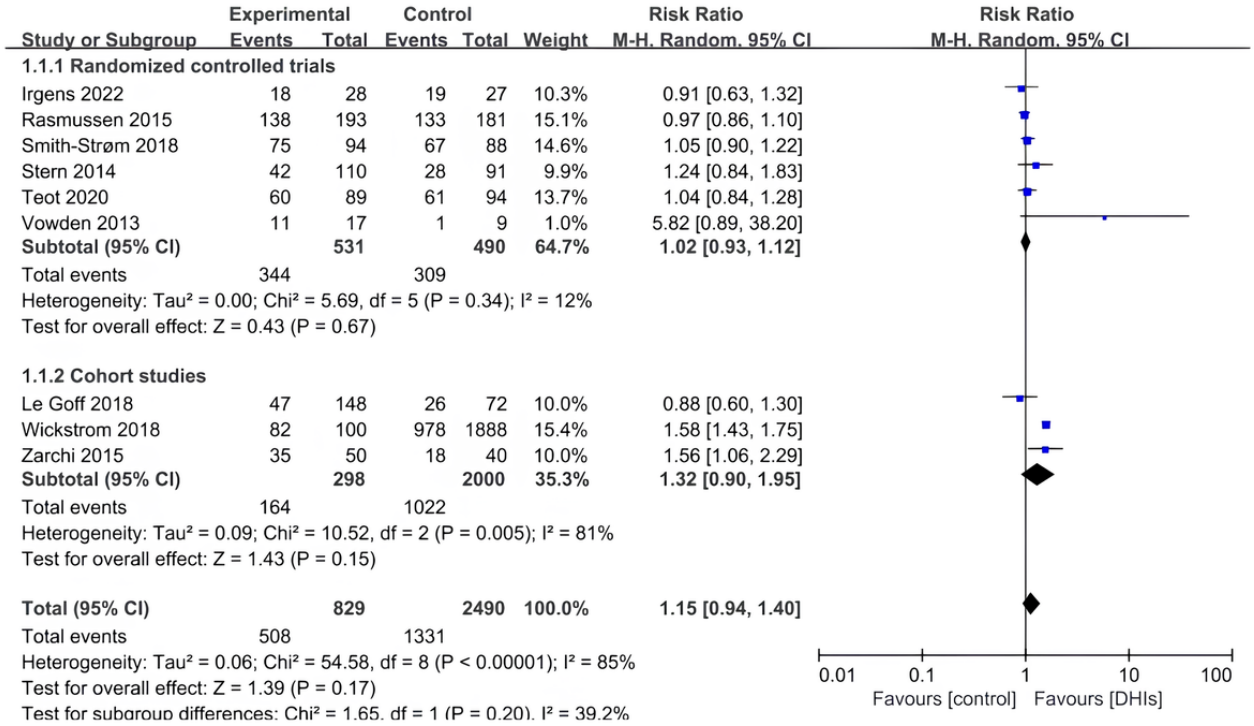

XSL•FO
RenderX

Figure 3. Forest plot of digital health interventions (DHIs) on wound healing around 1 year (subgroup of intervention types).

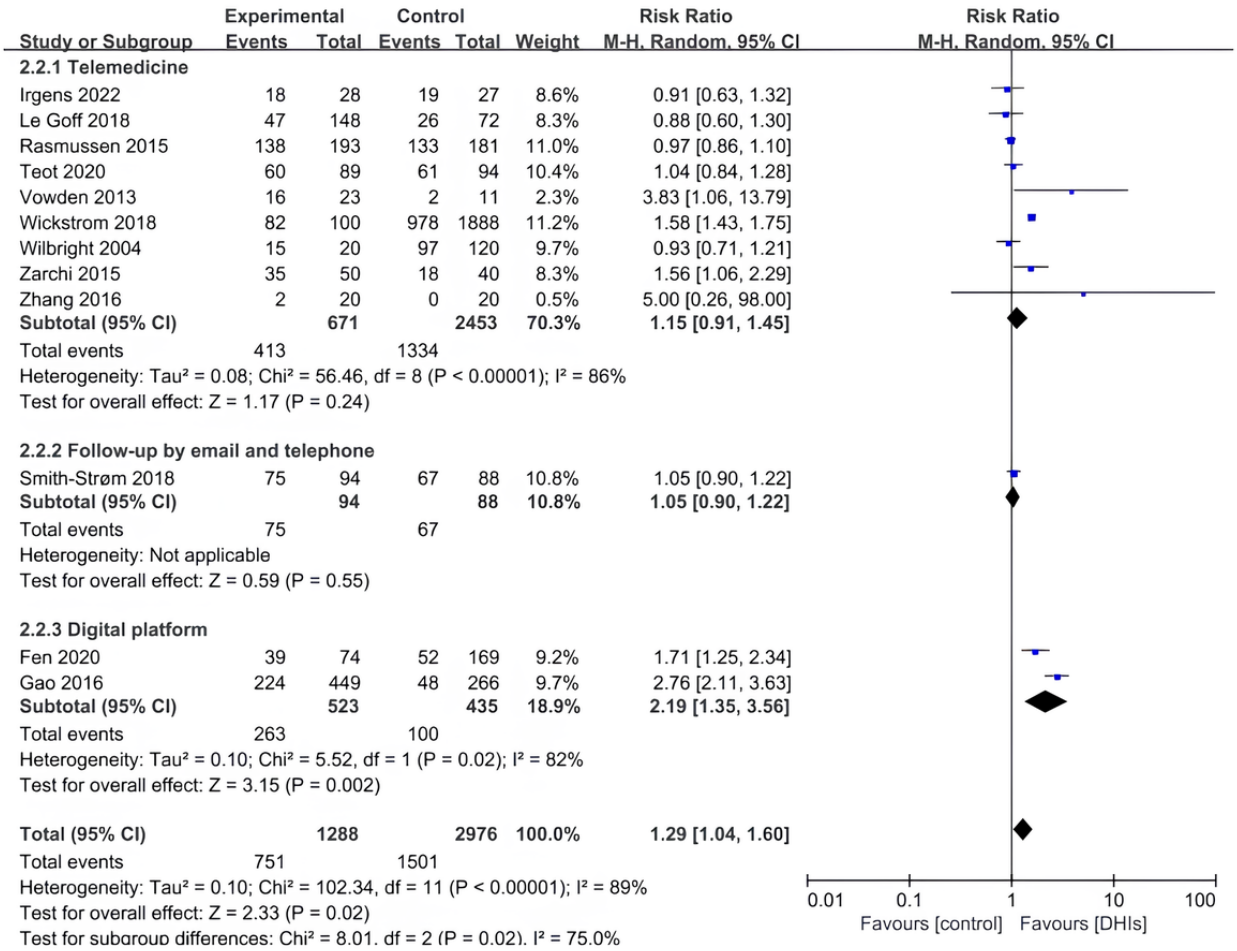

Wound Healing Time

Of the 25 studies, 8 (32%) studies reported different forms of wound healing time [52,53,55,57-59,61,62], 12% (3/25) of the studies, including 2 RCTs and 1 cohort study, reported the mean and SD of wound healing time frames [52,53,59], and a further 12% (3/25) of the studies provided median wound healing times [57,61,62], with 2 studies demonstrating considerable effectiveness in the DHIs group [61,62], and 1 study demonstrating a shorter wound healing time in the control group compared to the DHIs group (P=.56). In a study (1/25, 4%) that tracked the effect of the intervention according to the size of the patient's wounds, the results showed that the healing time of large, medium, and small wounds was shorter in the intervention group than in the control group (P<.05) [55]. It is also noteworthy that Dardari et al [58] found that wounds in the intervention group first showed improvement on day 21, much earlier than day 77 in the control group.

Wound Size

Of the 25 studies, 8% (2/25) of the studies consisting of in total of 168 patients reported a reduction of the wound area [51,60]. One of the studies [51] also noted that the DHIs group showed more considerable changes in wound depth. Another study [57] described the ulcer volume, indicating a mean reduction in ulcer volume in remote consultation was 79% versus 85% in usual care (P=.32).

Wound Reporting Rate

In total, 8% (2/25) of the studies reported wound reporting rate, all of which showed a significant increase following the use of a digital platform to monitor pressure injuries [24,65]. In addition, 1 pilot study (4%) noted in its results that the accuracy of screening and reporting of patients with pressure injuries from 97.26% to 100% after using the digital platform [64].

All-Cause Mortality

Of the 25 studies, 40% (10/25) of the studies reported all-cause mortality around 1 year [21-23,47,49,52,53,58,61,62]. There was no significant difference in all-cause mortality between the DHIs and control groups (RR 1.17,95% CI 0.70-1.96; P=.54; I^2=40%; Figure 4A). This finding is consistent in RCTs (RR 1.08, 95% CI 0.55-2.12; P=.83; I^2=42%) and cohort studies (RR 1.29, 95% CI 0.44-3.75; P=.64; I^2=51%). Subgroup analysis of RCTs reveals there was no statistically significant decreased risk of all-cause mortality in patients receiving telemedicine (RR 0.93, 95% CI 0.18-4.87; P=.93; I^2=63%) and follow-up by email and telephone (RR 1.48, 95% CI 0.78-2.82; P=.23; I^2=0%; Figure 4B).

XSL•FO
RenderX

Figure 4. (A) Forest plot of digital health interventions (DHIs) on all-cause mortality. (B) Forest plot of DHIs on all-cause mortality (subgroup of intervention types).

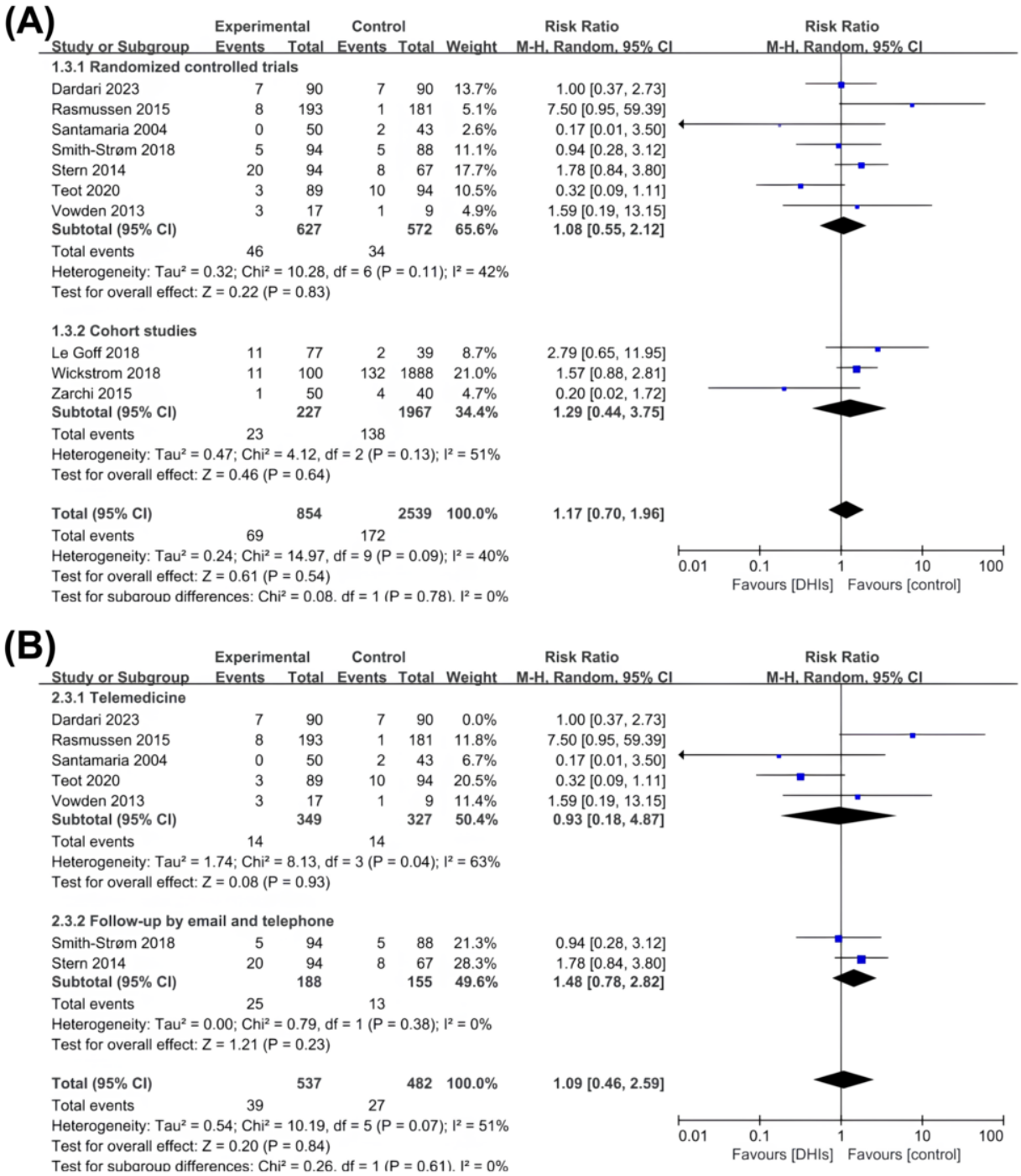

Adverse Event

A total of 6 (24%) studies [47,49,52,56,58,62], including 5 (20%) RCTs and 1 (4%) cohort study, reported the incidence of adverse events (infections and amputations). A meta-analysis of all RCTs found no statistically significant difference in the incidence of adverse events between the DHIs and the comparator group in patients with chronic wounds (RR 0.56, 95% CI 0.29-1.11; P=.10; I^2=77%) (Multimedia Appendix 6). Since 1 (4%) RCT clearly stated that none of the adverse events were attributed to the intervention [58], we attempted to remove this research from the meta-analysis and found that in the remaining 4 (16%) RCTs, the DHIs group had even fewer adverse events (RR 0.44, 95% CI 0.22-0.89; P=.02; I^2=41%) (Multimedia Appendix 7). Another cohort study (1/25, 4%) revealed that the DHIs group had a 2% amputation rate compared to 1% in the control group [62].

XSL•FO
RenderX

Cost Analysis

A cost analysis of DHIs against controls was addressed in 8 (32%) studies [23,48,49,51,53,58,60,61], with 5 (20%) studies revealed a reduction in total costs [23,49,58,60,61]. One study attributed the significant reduction in costs in the intervention group to a shorter hospital stay due to telemedicine [58], and another study investigated transport costs and found that the DHIs group spent an average of approximately US $650 less per person on transport for wound care compared with usual care [53]. However, 2 (8%) studies also concluded that the total costs were higher in the DHIs group [48,51], but for different reasons; the study by Terry et al [48] pointed out that the increased cost was due to larger and more severe wounds in the telemedicine group of patients, while Arora et al [51] attributed the increased cost to the use of new technology and assistive devices in the intervention group.

Patient Satisfaction

Patient satisfaction was reported in only a few (3/25, 12%) studies [52,56,66], of which 2 (8%) studies used a self-administered satisfaction questionnaire and found a substantial improvement in patient satisfaction after using DHIs [56,66], and the remaining study (1/25, 4%), measured using the GS-PEQ (The Generic Short Patient Experiences Questionnaire) scale [52], found no significant difference between the DHIs and the control group in patient satisfaction (MD 0.07, 95% CI 0.10-0.24). Due to heterogeneous outcome measures, a quantitative evaluation was not possible.

PUSH-Score

In addition to the typical result indicators for chronic wound evaluation outlined above, we observed significant changes in the PUSH-score (Pressure Ulcer Scale for Healing score), which is used to assess pressure injury only, in the DHIs group in 3 (12%) studies [51,54,66], including 2 (8%) RCTs and 1 (4%) quasi-experimental study. Although the variability of the research designs prohibited us from doing a meta-analysis, all studies revealed that using a digital platform improved the PUSH-score in patients with pressure injuries.

Publication Bias and Sensitivity Analysis

The funnel plot for all-cause mortality underlying the meta-analyses was symmetrical, which reflected a low risk of publication bias (Figure 5), the Egger test also showed consistent result (*P*=.37).

Sensitivity analysis was performed by omitting studies sequentially (Table 2). For wound healing around 1 year (Figure 6A), the pooled RR ranged from 1.05 (95% *CI* 0.93-1.18) to 1.19 (95% *CI* 0.95-1.47). For all-cause mortality (Figure 6B), the pooled RR ranged from 1.06 (95% *CI* 0.58-1.95) to 1.41 (95% *CI* 0.91-2.19). For adverse event (Figure 6C), the pooled RR ranged from 0.31 (95% *CI* 0.15-0.65) to 0.54 (95% *CI* 0.27-1.05). The results revealed that each outcome was relatively stable and would not change due to the elimination of a study; thus, the result of meta-analysis was robust.

Figure 5. Funnel figure of test for publication bias. RR: relative risk.

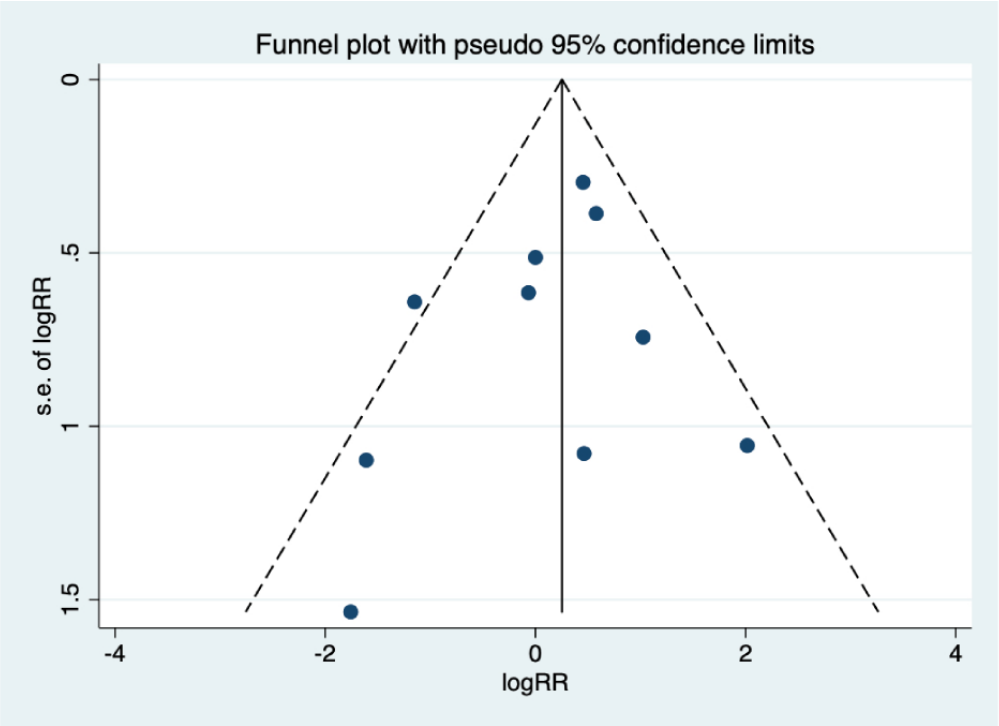

XSL•FO
RenderX

Table 2. Sensitivity analysis of included studies.

Study omitted	RR[a] (95% CI) for wound healing	RR (95% CI) for all-cause mortality	RR (95% CI) for adverse event
Irgens (2022)	1.18 (0.96-1.46)	__[b]	—
Rasmussen (2015)	1.19 (0.95-1.47)	1.09 (0.66-1.78)	0.31 (0.15-0.65)
Smith-Strøm (2018)	1.17 (0.93-1.48)	1.19 (0.67-2.12)	0.37 (0.12-1.17)
Stern (2014)	1.14 (0.92-1.42)	1.06 (0.58-1.95)	—
Teot (2020)	1.17 (0.94-1.46)	1.41 (0.91-2.19)	—
Vowden (2013)	1.13 (0.93-1.38)	1.14 (0.66-1.98)	—
Le Goff (2018)	1.19 (0.96-1.46)	1.08 (0.62-1.86)	—
Wickstrom (2018)	1.05 (0.93-1.18)	1.07 (0.57-2.03)	—
Zarchi (2015)	1.11 (0.90-1.38)	1.29 (0.79-2.09)	—
Dardari (2023)	—	1.19 (0.66-2.14)	—
Santamaria (2004)	—	1.24 (0.75-2.05)	0.51 (0.26-1.00)
Wu (2022)	—	—	0.54 (0.27-1.05)
Combined	1.15 (0.94-1.40)	1.17 (0.70-1.96)	0.44 (0.22-0.89)

[a]RR: relative risk.
[b]Not applicable.

XSL•FO
RenderX

Figure 6. (A) Sensitivity analysis of wound healing. (B) Sensitivity analysis of all-cause mortality. (C) Sensitivity analysis of adverse event.

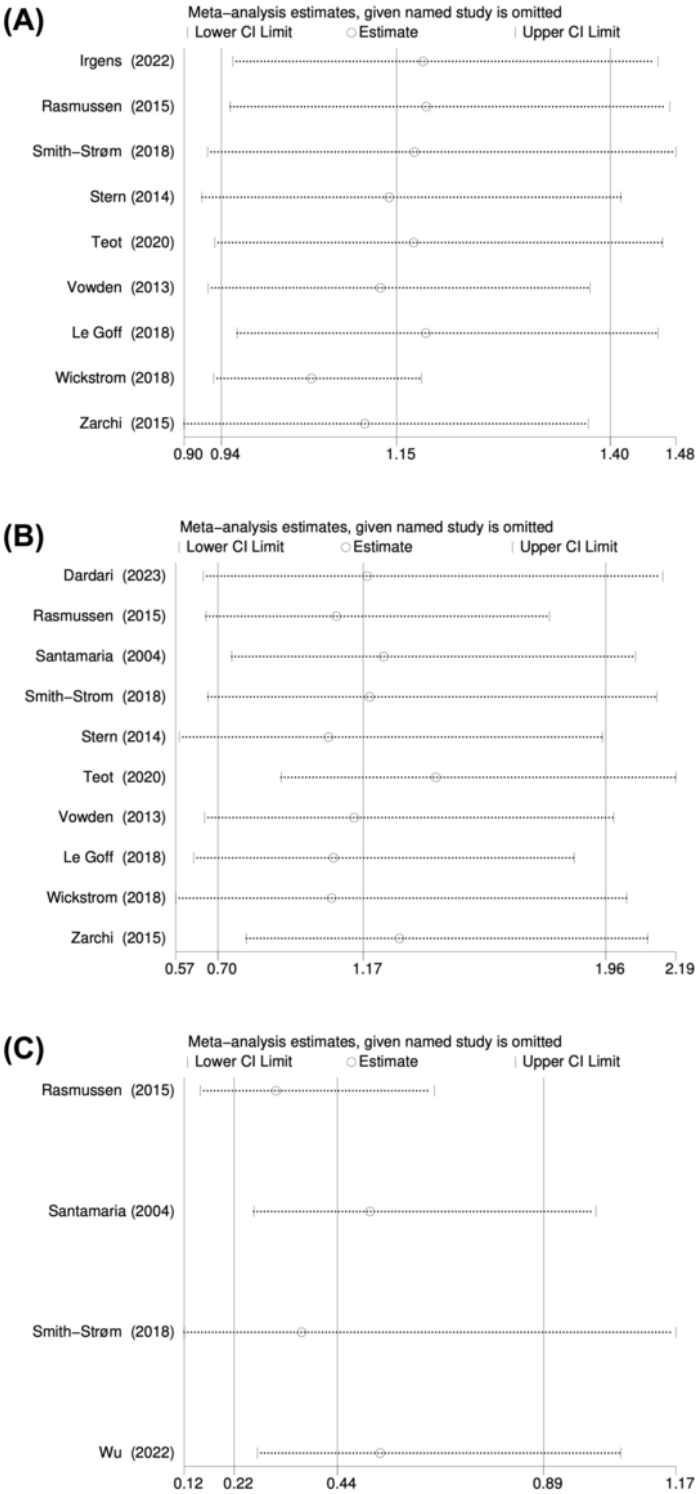

Discussion

Principal Results

This review provides a synthesis of high-quality evidence pertaining to the efficacy of DHIs for chronic wounds with different etiologies. Altogether, 12 RCTs, 5 cohort studies, and 3 quasi-experimental studies were identified and included in this meta-analysis. The magnitude of the intervention effect varied across studies, influenced by factors such as the nature of the intervention, assessment methods, and intervention duration.

The meta-analysis showed that DHIs did not significantly differ from usual care in terms of wound healing and all-cause mortality over around 1 year, which is consistent with findings of a previous meta-analysis targeting the effectiveness of telemedicine for chronic wound management [29]. Nevertheless, we believe it may be too early to conclude that DHIs are as

XSL•FO
RenderX

equally effective as usual care in improving wound healing due to the lack of a sufficient number of high-quality randomized controlled studies and the fact that digital technology is always evolving.

Our study also suggests that DHIs can reduce adverse events in chronic wound management, which is inconsistent with existing evidence. A previous meta-analysis showed no significant difference in the number of adverse events between the telemedicine and usual care groups [67]. This may be due to the inclusion of other forms of DHI in addition to telemedicine in our study, and it also lends support to the idea of exploratory subgroup analyses of intervention types.

Subgroup analyses demonstrated a beneficial effect of digital platforms for hospitalized patients on wound healing, implying differences in chronic wound management outcomes among intervention types, which may be linked to the characteristics and limitations of the technologies. Specifically, teleconsultation scenarios in the form of videoconferencing are excellent [68], but require high levels of equipment and patient finances [51,69]. Although telephone and email follow-ups are more concise and convenient [70], visual diagnosis of wounds is lacking and the effectiveness of the interaction is difficult to ensure; for example, nurses are frequently unable to validate that patients have read and understood the content of the email [71]. Furthermore, digital platforms used for in-hospital patient pressure injury evaluation and management have shown promise in improving patients' PUSH-score, but evidence from these studies is limited. Through further analysis of studies reporting this outcome metric, we discovered that the significant improvement in patients' PUSH-score is closely linked to the real-time tracking function of the digital platform. This feature effectively aligns with the PUSH scale's dynamic assessment of pressure injuries, enabling the quantification of dynamic changes in patients' wounds [72]. Additionally, the platform's built-in automated analytics function provides critical feedback, offering a reliable predictive basis for nurses to identify the risk of pressure injury development in patients [64,73]. However, whether digital platforms provide better intervention outcomes than other types of DHIs still need to be validated and explored in higher-quality studies.

Of note, few studies have considered the impact of patient age on the effectiveness of technological interventions, and the groups included in the studies were predominantly middle-aged and older people, but the fact is that digital health care is more accepted in younger age groups [74]. Cost is another component that must be considered. Several studies have combined different forms of digital health technology to positively impacted wound outcomes; yet, it remains to be seen whether the increased cost of care as a result of the combination of technologies reduces patient adherence to treatment, which in turn affects wound outcomes [74]. Similarly, unequal access, use, and knowledge of information and communication technology among patients may also affect the effectiveness of DHIs.

In addition to the issues mentioned above, a point of concern, however, was a lack of blinding in most studies. The nature of interventions made them difficult to blind participants or providers, as the conduct of interactions frequently necessitates that both parties understand what they are doing. Patients, for example, are required to prepare their own computers or visit certain specific websites to validate and register their personal identities when conducting real-time videoconferencing-supported wound diagnosis and follow-up, and it is difficult for doctors or nurses on the other end of the video to be unaware of the content and form of the intervention, which is closely related to their work experience and professionalism; when using mobile phones or SMS text messaging for interventions, patients and researchers frequently agree on the frequency of the intervention ahead of time and exchange contact information in case they miss or ignore the diagnostic content, care recommendations, or follow-up feedback. Future RCTs should address this aspect to strengthen the evidence on DHIs for the management of patients with chronic wounds.

Strengths and Limitations

To our knowledge, this is one of the most comprehensive and up - to - date reviews and meta - analyses to evaluate the effects of DHIs on chronic wound outcomes across a broad spectrum of the population, combining data from RCTs, cohort studies, and quasi-experimental studies. While there have been studies looking at the use of digital health technology in chronic wound management, this study is unique in that it notes the diversity of interventions and attempts to quantify the effects that different types of DHIs show in terms of improving wound outcomes. For example, an explorable link was discovered between digital platforms and changes in PUSH-score in patients with pressure injuries. This provides new ideas for future research on whether different types of digital intervention techniques can be coupled to various etiologic wound evaluation tools to improve intervention outcomes. Furthermore, studies included in this review cover a wide range of DHIs, which differ in terms of the persons engaged, the intervention management, and the technology used, and are applied to populations with various wound etiologies, allowing the results to be generalizable.

Despite the clinical implications and strength of our review, certain limitations need acknowledgment. First, few researchers have specified the wound staging of included patients, which prevents us from drawing firm conclusions about this. Additionally, differences in participants, intervention contents, methods, frequency, time, and measurements in the control group also resulted in heterogeneity. Moreover, due to the significant variation in intervention design, it was difficult to extract and classify interventions in a very standardized way. Even though we divided it into 3 types of intervention subgroups, the interventions are still not standardized within each subgroup, and many parameters are implicitly variable, which may influence the results of our review. Finally, it must be explained that we included other study types in the meta-analysis in addition to RCTs, which is not recommended by the official guidelines for meta-analysis. Therefore, to overcome this limitation, we have clearly categorized and described in detail the results of the RCTs and non-RCTs included in this paper to improve the clarity and reference value of the results.

XSL•FO
RenderX

Implication for Practice and Research

The results of the subgroup analyses point to the benefits of digital platforms for chronic wound management in hospitalized patients, but a limitation that cannot be ignored is that most of the studies were quasi-experimental and the platforms they used varied in terms of both structural design and technological quality. Therefore, future studies should be based on evidence-based practice, attempt to develop a digital platform that can be replicated on a large scale, and conduct more RCTs, considering the context and needs of the population, such as the acceptability of the technology, economic disparities, and the use of other services. At the same time, given the association between age and digital health literacy, it is necessary to provide interventions for each age group to clearly confirm effects in future studies.

In addition to these clinical implications, there are several possible directions for further research. Considering the heterogeneity of the interventions and the wound etiology, we recommend further research to investigate whether there are certain associations between the types of digital health interventions and patient characteristics to provide a valid reference for the clinical construction of a systematic digital wound management program, such as differences in the effectiveness of the same interventions for patients with wounds of different etiologies, and the relationship between patients' digital health literacy and the effectiveness of the interventions. Moreover, regarding the review process we mentioned in the principal results section, we found that only a few studies blinded wound therapists, nurses, or patients. We recommend that future studies consider using existing high-quality patient digital information collection programs or web-based data platforms and rationally using the automated analytical capabilities of the technology to conduct single-blind experiments in which patients with comparable baseline information are randomly grouped with the consent of the patient, and therapists are implemented blinding of assessors by performing software-based outcome measures for all primary and secondary outcomes and automatically storing patient self-reported data. Data review should be done in a blinded manner until analyses were performed, and data analysis was also done using blinded subgroups to improve the quality of evidence generated by the study.

Conclusions

In summary, this study suggests that DHIs are effective for chronic wound management, but the jury is still out on who is better between them and usual care. We also found indications that digital platforms can help with chronic wound management in hospitalized patients, warranting further investigations. Moreover, future high-quality research is needed, to identify factors contributing to improved patient-centered interventions with better care outcomes, as well as more careful consideration of individual patient characteristics.

Acknowledgments

This research received the support of National Research Training Program of Gansu Provincial Hospital grant (19SYPYA-4), Scientific Research Project of Health Industry of Gansu Province grant (GSWSHL-2021-003), and China Medical Board grant #20-374).

Data Availability

The authors confirm that the data supporting the findings of this study are available within the article (and Multimedia Appendices 1-7).

Authors' Contributions

XB, HZ, and LH contributed to study conceptualization and design. XB and HZ contributed to article screening. XB and HZ contributed to data extraction. YJ, CY, and YM contributed to quality assessment. XB contributed to the meta-analysis. XB and HZ contributed to the first draft. All authors contributed to critical revision of the drafts for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e47904_app1.docx]

Multimedia Appendix 2
Search strategy for all databases.
[DOCX File , 25 KB - jmir_v26i1e47904_app2.docx]

Multimedia Appendix 3
Characteristics of the included studies.

XSL•FO
RenderX

[DOCX File , 24 KB - jmir_v26i1e47904_app3.docx]

Multimedia Appendix 4
Risk of Bias.
[DOCX File , 24 KB - jmir_v26i1e47904_app4.docx]

Multimedia Appendix 5
Forest plot of digital health interventions on wound healing around three months (different study types).
[PDF File (Adobe PDF File), 189 KB - jmir_v26i1e47904_app5.pdf]

Multimedia Appendix 6
Forest plot of digital health interventions on adverse event (randomized controlled trials).
[PDF File (Adobe PDF File), 157 KB - jmir_v26i1e47904_app6.pdf]

Multimedia Appendix 7
Forest plot of digital health interventions on adverse event (adjusted version for randomized controlled trials).
[PDF File (Adobe PDF File), 122 KB - jmir_v26i1e47904_app7.pdf]

References

1. Ernstsen CV, Riishede A, Iversen AKS, Bay L, Bjarnsholt T, Nejsum LN. E-cadherin and aquaporin-3 are downregulated in wound edges of human chronic wounds. APMIS 2023 Aug;131(8):403-409. [doi: 10.1111/apm.13332] [Medline: 37267058]
2. Martinengo L, Olsson M, Bajpai R, Soljak M, Upton Z, Schmidtchen A, et al. Prevalence of chronic wounds in the general population: systematic review and meta-analysis of observational studies. Ann Epidemiol 2019 Jan;29:8-15 [FREE Full text] [doi: 10.1016/j.annepidem.2018.10.005] [Medline: 30497932]
3. Lazarus GS, Cooper DM, Knighton DR, Percoraro RE, Rodeheaver G, Robson MC. Definitions and guidelines for assessment of wounds and evaluation of healing. Wound Repair Regen 1994 Jul;2(3):165-170. [doi: 10.1046/j.1524-475X.1994.20305.x] [Medline: 17156107]
4. Dissemond J, Bültemann A, Gerber V, Jäger B, Kröger K, Münter C. Diagnosis and treatment of chronic wounds: current standards of Germany's Initiative for Chronic Wounds e. V. J Wound Care 2017 Dec 02;26(12):727-732. [doi: 10.12968/jowc.2017.26.12.727] [Medline: 29244967]
5. Falcone M, De Angelis B, Pea F, Scalise A, Stefani S, Tasinato R, et al. Challenges in the management of chronic wound infections. J Glob Antimicrob Resist 2021 Sep;26:140-147 [FREE Full text] [doi: 10.1016/j.jgar.2021.05.010] [Medline: 34144200]
6. Frykberg RG, Banks J. Challenges in the treatment of chronic wounds. Adv Wound Care (New Rochelle) 2015 Sep 01;4(9):560-582 [FREE Full text] [doi: 10.1089/wound.2015.0635] [Medline: 26339534]
7. Sen CK, Gordillo GM, Roy S, Kirsner R, Lambert L, Hunt TK, et al. Human skin wounds: a major and snowballing threat to public health and the economy. Wound Repair Regen 2009;17(6):763-771 [FREE Full text] [doi: 10.1111/j.1524-475X.2009.00543.x] [Medline: 19903300]
8. Riihimies R, Kosunen E, Koskela TH. Segmenting patients with diabetes with the navigator service in primary care and a description of the self-acting patient group: cross-sectional study. J Med Internet Res 2023;25:e40560 [FREE Full text] [doi: 10.2196/40560] [Medline: 37682585]
9. Price P, Fogh K, Glynn C, Krasner DL, Osterbrink J, Sibbald RG. Managing painful chronic wounds: the wound pain management model. Int Wound J 2007;4 Suppl 1(Suppl 1):4-15 [FREE Full text] [doi: 10.1111/j.1742-481X.2007.00311.x] [Medline: 17394625]
10. Nazeha N, Lee JY, Saffari SE, Meng L, Ho P, Ng YZ, et al. The burden of costs on health services from patients with venous leg ulcers in Singapore. Int Wound J 2023;20(3):845-852 [FREE Full text] [doi: 10.1111/iwj.13931] [Medline: 36098383]
11. Martinengo L, Yeo NJY, Markandran KD, Olsson M, Kyaw BM, Car LT. Digital health professions education on chronic wound management: a systematic review. Int J Nurs Stud 2020;104:103512. [doi: 10.1016/j.ijnurstu.2019.103512] [Medline: 32086027]
12. Gray TA, Rhodes S, Atkinson RA, Rothwell K, Wilson P, Dumville JC, et al. Opportunities for better value wound care: a multiservice, cross-sectional survey of complex wounds and their care in a UK community population. BMJ Open 2018 Mar 22;8(3):e019440 [FREE Full text] [doi: 10.1136/bmjopen-2017-019440] [Medline: 29572395]
13. Moyse T, Bates J, Bena JF, Morrison SL, Albert NM. Effect of a Hospital-Acquired Pressure Injury (HAPI) risk score on HAPI rates in patients with vascular diseases. J Wound Ostomy Continence Nurs 2023;50(1):13-18. [doi: 10.1097/WON.0000000000000944] [Medline: 36640159]

XSL•FO
RenderX

14. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organzation. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2023-11-18]
15. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: systematic literature review. J Med Internet Res 2023 Jul 26;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
16. Guo SH, Lin J, Hsing H, Lee C, Chuang S. The effect of mobile ehealth education to improve knowledge, skills, self-care, and mobile eHealth literacies among patients with diabetes: development and evaluation study. J Med Internet Res 2023;25:e42497 [FREE Full text] [doi: 10.2196/42497] [Medline: 38055321]
17. Ni Z, Wu B, Yang Q, Yan LL, Liu C, Shaw RJ. An mHealth intervention to improve medication adherence and health outcomes among patients with coronary heart disease: randomized controlled trial. J Med Internet Res 2022;24(3):e27202 [FREE Full text] [doi: 10.2196/27202] [Medline: 35262490]
18. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. Lancet Digit Health 2020;2(4):e168-e178 [FREE Full text] [doi: 10.1016/S2589-7500(20)30017-0] [Medline: 33334505]
19. Koehler J, Stengel A, Hofmann T, Wegscheider K, Koehler K, Sehner S, et al. Telemonitoring in patients with chronic heart failure and moderate depressed symptoms: results of the Telemedical Interventional Monitoring in Heart Failure (TIM-HF) study. Eur J Heart Fail 2021;23(1):186-194 [FREE Full text] [doi: 10.1002/ejhf.2025] [Medline: 33063412]
20. Eriksson E, Liu PY, Schultz GS, Martins-Green MM, Tanaka R, Weir D, et al. Chronic wounds: treatment consensus. Wound Repair Regen 2022;30(2):156-171 [FREE Full text] [doi: 10.1111/wrr.12994] [Medline: 35130362]
21. Zarchi K, Haugaard VB, Dufour DN, Jemec GBE. Expert advice provided through telemedicine improves healing of chronic wounds: prospective cluster controlled study. J Invest Dermatol 2015;135(3):895-900 [FREE Full text] [doi: 10.1038/jid.2014.441] [Medline: 25290685]
22. Vowden K, Vowden P. A pilot study on the potential of remote support to enhance wound care for nursing-home patients. J Wound Care 2013;22(9):481-488. [doi: 10.12968/jowc.2013.22.9.481] [Medline: 24005782]
23. Stern A, Mitsakakis N, Paulden M, Alibhai S, Wong J, Tomlinson G, et al. Pressure ulcer multidisciplinary teams via telemedicine: a pragmatic cluster randomized stepped wedge trial in long term care. BMC Health Serv Res 2014 Feb 24;14:83 [FREE Full text] [doi: 10.1186/1472-6963-14-83] [Medline: 24559218]
24. Gao R, Zhang Y, Wan L. Application and exploration of pressure ulcer professional tracking based on information platform. Chin J Mod Nurs 2016;22(7):1010-1009. [doi: 10.3760/cma.j.issn.1674-2907.2016.07.033]
25. Feng Y, Chen M, Yan Y, Hou Y, Sang N. 3-tier quality control management of pressure ulcers/injuries based on information system. Journal of Nursing Science 2020 Oct 25;35(20):67-70 [FREE Full text]
26. Han L, Bai X, Zhang H, An N. Digital nursing quality management: an evolutionary concept analysis. JMIR Preprints Preprint posted online Mar 25, 2024. [doi: 10.2196/preprints.58772]
27. Yammine K, Estephan M. Telemedicine and diabetic foot ulcer outcomes. A meta-analysis of controlled trials. Foot (Edinb) 2022;50:101872. [doi: 10.1016/j.foot.2021.101872] [Medline: 35219129]
28. Chen G, Wang T, Zhong L, He X, Huang C, Wang Y, et al. Telemedicine for preventing and treating pressure injury after spinal cord injury: systematic review and meta-analysis. J Med Internet Res 2022;24(9):e37618 [FREE Full text] [doi: 10.2196/37618] [Medline: 36069842]
29. Chen L, Cheng L, Gao W, Chen D, Wang C, Ran X. Telemedicine in chronic wound management: systematic review and meta-analysis. JMIR Mhealth Uhealth 2020;8(6):e15574 [FREE Full text] [doi: 10.2196/15574] [Medline: 32584259]
30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
31. Lewkowitz AK, Whelan AR, Ayala NK, Hardi A, Stoll C, Battle CL, et al. The effect of digital health interventions on postpartum depression or anxiety: a systematic review and meta-analysis of randomized controlled trials. Am J Obstet Gynecol 2024;230(1):12-43. [doi: 10.1016/j.ajog.2023.06.028] [Medline: 37330123]
32. Biesecker BB, Lillie SE, Amendola LM, Donohue KE, East KM, Foreman AKM, et al. A review and definition of 'usual care' in genetic counseling trials to standardize use in research. J Genet Couns 2021;30(1):42-50 [FREE Full text] [doi: 10.1002/jgc4.1363] [Medline: 33278053]
33. Ambrosino N, Vitacca M, Dreher M, Isetta V, Montserrat JM, Tonia T, ERS Tele-Monitoring of Ventilator-Dependent Patients Task Force. Tele-monitoring of ventilator-dependent patients: a European Respiratory Society Statement. Eur Respir J 2016;48(3):648-663 [FREE Full text] [doi: 10.1183/13993003.01721-2015] [Medline: 27390283]
34. Zhou X, Liu H, Li J, Zhang K, Lev B. Channel strategies when digital platforms emerge: a systematic literature review. Omega 2023 Oct;120:102919. [doi: 10.1016/j.omega.2023.102919]
35. Zonderland ME. Theoretical and Practical Aspects of Outpatient Clinic Optimization. In: Zonderland ME, Boucherie RJ, Hans EW, Kortbeek N, editors. Theoretical and practical aspects of outpatient clinic optimization. Germany: Springer, Cham Press; Mar 30, 2021:25-36.
36. Institute of Medicine. Defining Primary Care: An Interim Report. In: Donaldson M, Yordy K, Vanselow N, editors. Defining Primary Care: An Interim Report. Washington, DC: The National Academies Press; 1994:53.
37. Thomé B, Dykes A, Hallberg IR. Home care with regard to definition, care recipients, content and outcome: systematic literature review. J Clin Nurs 2003;12(6):860-872. [doi: 10.1046/j.1365-2702.2003.00803.x] [Medline: 14632979]

XSL•FO
RenderX

38. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
39. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
40. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
41. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 03;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
42. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
43. Peng J, Zeng J, Cai B, Yang H, Cohen MJ, Chen W, et al. Establishment of quantitative severity evaluation model for spinal cord injury by metabolomic fingerprinting. PLoS One 2014;9(4):e93736 [FREE Full text] [doi: 10.1371/journal.pone.0093736] [Medline: 24727691]
44. Morgan AJ, Reavley NJ, Ross A, Too LS, Jorm AF. Interventions to reduce stigma towards people with severe mental illness: systematic review and meta-analysis. J Psychiatr Res 2018 Aug;103:120-133. [doi: 10.1016/j.jpsychires.2018.05.017] [Medline: 29843003]
45. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
46. Zwetsloot P, Van Der Naald M, Sena ES, Howells DW, IntHout J, De Groot JA, et al. Standardized mean differences cause funnel plot distortion in publication bias assessments. Elife 2017 Sep 08;6:e24260 [FREE Full text] [doi: 10.7554/eLife.24260] [Medline: 28884685]
47. Santamaria N, Carville K, Ellis I, Prentice J. The effectiveness of digital imaging and remote expert wound consultation on healing rates in chronic lower leg ulcers in the Kimberley region of Western Australia. Primary Intention: The Australian Journal of Wound Management 2004;12(2):62-70 [FREE Full text]
48. Terry M, Halstead LS, O'Hare P, Gaskill C, Ho PS, Obecny J, et al. Feasibility study of home care wound management using telemedicine. Adv Skin Wound Care 2009 Aug;22(8):358-364. [doi: 10.1097/01.ASW.0000358638.38161.6b] [Medline: 19638799]
49. Rasmussen BSB, Froekjaer J, Bjerregaard MR, Lauritsen J, Hangaard J, Henriksen CW, et al. A randomized controlled trial comparing telemedical and standard outpatient monitoring of diabetic foot ulcers. Diabetes Care 2015 Sep;38(9):1723-1729. [doi: 10.2337/dc15-0332] [Medline: 26116717]
50. Zhang X, Li F, Liu B. The efficacy of teleconsultation model applied to the treatment of patients with venous ulcers of the lower extremities in nursing homes. West China Medical Journal 2016:00. [doi: 10.1097/tin.0000000000000054]
51. Arora M, Harvey LA, Glinsky JV, Chhabra HS, Hossain MS, Arumugam N, et al. Cost-effectiveness analysis of telephone-based support for the management of pressure ulcers in people with spinal cord injury in India and Bangladesh. Spinal Cord 2017 Dec;55(12):1071-1078. [doi: 10.1038/sc.2017.87] [Medline: 28809389]
52. Smith-Strøm H, Igland J, Østbye T, Tell GS, Hausken MF, Graue M, et al. The effect of telemedicine follow-up care on diabetes-related foot ulcers: a cluster-randomized controlled noninferiority trial. Diabetes Care 2018 Jan;41(1):96-103. [doi: 10.2337/dc17-1025] [Medline: 29187423]
53. Téot L, Geri C, Lano J, Cabrol M, Linet C, Mercier G. Complex wound healing outcomes for outpatients receiving care via telemedicine, home health, or wound clinic: a randomized controlled trial. Int J Low Extrem Wounds 2020 Jun;19(2):197-204. [doi: 10.1177/1534734619894485] [Medline: 31852312]
54. Zhou H, Qu H, Zhou H, Ma Y, Ruan Y, Zhou J. Application effect of nursing intervention based on telemedicine technology in patients with stress injury in community. Clinical Research and Practice 2021;6(06):120-122 [FREE Full text] [doi: 10.19347/j.cnki.2096-1413.202106043]
55. Shen H, Sun R, Yang Y, Bai H. Practice of internet information platform in home nursing service of difficult healing wounds. Clinical Research 2022;30(10):185-188 [FREE Full text]
56. Wu W, Hao Z, Tian Q, Li X, Ju F. Effect evaluation of urban-rural medical specialty cooperation based on telemedicine platform. Smart Healthcare 2022;8(02):173-176 [FREE Full text] [doi: 10.19335/j.cnki.2096-1219.2022.02.055]
57. Irgens I, Midelfart-Hoff J, Jelnes R, Alexander M, Stanghelle JK, Thoresen M, et al. Videoconferencing in pressure injury: randomized controlled telemedicine trial in patients with spinal cord injury. JMIR Form Res 2022 Apr 19;6(4):e27692 [FREE Full text] [doi: 10.2196/27692] [Medline: 35438645]
58. Dardari D, Franc S, Charpentier G, Orlando L, Bobony E, Bouly M, et al. Hospital stays and costs of telemedical monitoring versus standard follow-up for diabetic foot ulcer: an open-label randomised controlled study. Lancet Reg Health Eur 2023 Sep;32:100686 [FREE Full text] [doi: 10.1016/j.lanepe.2023.100686] [Medline: 37520145]

XSL•FO
RenderX

59. Wilbright WA, Birke JA, Patout CA, Varnado M, Horswell R. The use of telemedicine in the management of diabetes-related foot ulceration: a pilot study. Adv Skin Wound Care 2004 Jun;17(5 Pt 1):232-238. [doi: 10.1097/00129334-200406000-00012] [Medline: 15192491]
60. Bergersen TK, Storheim E, Gundersen S, Kleven L, Johnson M, Sandvik L, et al. Improved clinical efficacy with wound support network between hospital and home care service. Adv Skin Wound Care 2016 Nov;29(11):511-517. [doi: 10.1097/01.ASW.0000499714.97688.4b] [Medline: 27755050]
61. Le Goff-Pronost M, Mourgeon B, Blanchère JP, Teot L, Benateau H, Dompmartin A. Real-world clinical evaluation and costs of telemedicine for chronic wound management. Int J Technol Assess Health Care 2018;34(6):567-575. [doi: 10.1017/S0266462318000685] [Medline: 30369340]
62. Wickström HL, Öien RF, Fagerström C, Anderberg P, Jakobsson U, Midlöv PJ. Comparing video consultation with inperson assessment for Swedish patients with hard-to-heal ulcers: registry-based studies of healing time and of waiting time. BMJ Open 2018 Feb 15;8(2):e017623 [FREE Full text] [doi: 10.1136/bmjopen-2017-017623] [Medline: 29449288]
63. Gamus A, Keren E, Kaufman H, Chodick G. Synchronous video telemedicine in lower extremities ulcers treatment: a real-world data study. Int J Med Inform 2019 Apr;124:31-36. [doi: 10.1016/j.ijmedinf.2019.01.009] [Medline: 30784424]
64. Hu Q, Wu X, Qin Y, Yang Y, Xi H, Qin X. Construction of pressure injury information system for patients with surgery based on chain-type process. Chinese Journal of Modern Nursing 2018;24(25):3016-3021 [FREE Full text] [doi: 10.3760/cma.j.issn.1674-2907.2018.25.009]
65. Wu X, Fu J. Evaluation of application effect of pressure ulcer information management platform. Journal of Cardiovascular Surgery(Electronic Edition) 2019;8(1):63-65 [FREE Full text]
66. Xie P, Zhang C, Zhang Z, Zhou H. Construction and application of holistic informationalized chronic wound management system based on "Internet +". Journal of Clinical Medicine in Practice 2019;23(21):125-128 [FREE Full text]
67. Huang Z, Wu S, Yu T, Hu A. Efficacy of telemedicine for patients with chronic wounds: a meta-analysis of randomized controlled trials. Adv Wound Care (New Rochelle) 2021 Feb;10(2):103-112. [doi: 10.1089/wound.2020.1169] [Medline: 32320355]
68. Xiao Z, Han X. Evaluation of the effectiveness of telehealth chronic disease management system: systematic review and meta-analysis. J Med Internet Res 2023 Apr 27;25:e44256 [FREE Full text] [doi: 10.2196/44256] [Medline: 37103993]
69. Kamilu Sulaiman S, Wong AYL, Liangchi Li L, Fordjour Antwi-Afari M, Ou H, Wh Tsang H. The use of mobile health technology in the management of osteoarthritis: a scoping review with scientometric analyses. Int J Med Inform 2023 Feb;170:104937. [doi: 10.1016/j.ijmedinf.2022.104937] [Medline: 36493537]
70. Safavi AH, Lovas M, Liu ZA, Melwani S, Truong T, Devonish S, et al. Virtual care and electronic patient communication during COVID-19: cross-sectional study of inequities across a Canadian tertiary cancer center. J Med Internet Res 2022 Nov 04;24(11):e39728 [FREE Full text] [doi: 10.2196/39728] [Medline: 36331536]
71. Reiter A, Tov EY, Hochberg I. Do patients read emails from their physician containing tips on improving lifestyle habits? A pilot study. Int J Med Inform 2023 Feb;170:104967. [doi: 10.1016/j.ijmedinf.2022.104967] [Medline: 36587533]
72. Qiao H, Sun W, Liu M, Wang R, Chen S, Jin Y. Design of the mobile data application for pressure injuries and its application in clinical management of primary hospitals. Chinese Nursing Management 2022;22(10):1473-1477 [FREE Full text]
73. Xia D, Shi T, Lu W, Lu Y, Chen J, Fu Q. Development and application of clinical decision support system for pressure injury. Chinese Journal of Nursing 2020;55(01):50-54 [FREE Full text] [doi: 10.3761/j.issn.0254-1769.2020.01.008]
74. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52 [FREE Full text] [doi: 10.2196/jmir.3951] [Medline: 25803266]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
DHI: digital health intervention
MeSH: Medical Subject Headings
PICOS: Population, Intervention, Comparison, Outcomes, and Study
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized control trial
RR: relative risks
SMD: standardized mean difference
WHO-SEARO: World Health Organization South-East Asia Regional Organization

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, S Ma; submitted 05.04.23; peer-reviewed by L Martinengo, MMAZ Asyura; comments to author 28.07.23; revised version received 16.08.23; accepted 24.05.24; published 16.07.24.

Please cite as:
Bai X, Zhang H, Jiao Y, Yuan C, Ma Y, Han L
Digital Health Interventions for Chronic Wound Management: A Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e47904
URL: https://www.jmir.org/2024/1/e47904
doi:10.2196/47904
PMID:39012684

©Xinrui Bai, Hongyan Zhang, Yanxia Jiao, Chenlu Yuan, Yuxia Ma, Lin Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Mapping Digital Public Health Interventions Among Existing Digital Technologies and Internet-Based Interventions to Maintain and Improve Population Health in Practice: Scoping Review

Laura Maaß[1,2,3], MA; Konstantinos Angoumis[4,5], MSci; Merle Freye[2,6], PhD; Chen-Chia Pan[2,5,7], MA

[1]University of Bremen, SOCIUM Research Center on Inequality and Social Policy, Bremen, Germany
[2]Leibniz ScienceCampus Digital Public Health Bremen, Bremen, Germany
[3]Digital Health Section, European Public Health Association - EUPHA, Utrecht, Netherlands
[4]University of Bielefeld, Bielefeld, Germany
[5]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany
[6]University of Bremen, Institute for Information, Health and Medical Law - IGMR, Bremen, Germany
[7]University of Bremen, Institute for Public Health and Nursing Research – IPP, Bremen, Germany

Corresponding Author:
Laura Maaß, MA
University of Bremen, SOCIUM Research Center on Inequality and Social Policy
Mary-Somerville-Straße 3
Bremen, 28359
Germany
Phone: 49 421 218 58610
Email: laura.maass@uni-bremen.de

Abstract

Background: The rapid progression and integration of digital technologies into public health have reshaped the global landscape of health care delivery and disease prevention. In pursuit of better population health and health care accessibility, many countries have integrated digital interventions into their health care systems, such as web-based consultations, electronic health records, and telemedicine. Despite the increasing prevalence and relevance of digital technologies in public health and their varying definitions, there has been a shortage of studies examining whether these technologies align with the established definition and core characteristics of digital public health (DiPH) interventions. Hence, the imperative need for a scoping review emerges to explore the breadth of literature dedicated to this subject.

Objective: This scoping review aims to outline DiPH interventions from different implementation stages for health promotion, primary to tertiary prevention, including health care and disease surveillance and monitoring. In addition, we aim to map the reported intervention characteristics, including their technical features and nontechnical elements.

Methods: Original studies or reports of DiPH intervention focused on population health were eligible for this review. PubMed, Web of Science, CENTRAL, IEEE Xplore, and the ACM Full-Text Collection were searched for relevant literature (last updated on October 5, 2022). Intervention characteristics of each identified DiPH intervention, such as target groups, level of prevention or health care, digital health functions, intervention types, and public health functions, were extracted and used to map DiPH interventions. MAXQDA 2022.7 (VERBI GmbH) was used for qualitative data analysis of such interventions' technical functions and nontechnical characteristics.

Results: In total, we identified and screened 15,701 records, of which 1562 (9.94%) full texts were considered relevant and were assessed for eligibility. Finally, we included 185 (11.84%) publications, which reported 179 different DiPH interventions. Our analysis revealed a diverse landscape of interventions, with telemedical services, health apps, and electronic health records as dominant types. These interventions targeted a wide range of populations and settings, demonstrating their adaptability. The analysis highlighted the multifaceted nature of digital interventions, necessitating precise definitions and standardized terminologies for effective collaboration and evaluation.

Conclusions: Although this scoping review was able to map characteristics and technical functions among 13 intervention types in DiPH, emerging technologies such as artificial intelligence might have been underrepresented in our study. This review underscores the diversity of DiPH interventions among and within intervention groups. Moreover, it highlights the importance of precise terminology for effective planning and evaluation. This review promotes cross-disciplinary collaboration by emphasizing

XSL•FO
RenderX

the need for clear definitions, distinct technological functions, and well-defined use cases. It lays the foundation for international benchmarks and comparability within DiPH systems. Further research is needed to map intervention characteristics in this still-evolving field continuously.

Trial Registration: PROSPERO CRD42021265562; https://tinyurl.com/43jksb3k

International Registered Report Identifier (IRRID): RR2-10.2196/33404

(J Med Internet Res 2024;26:e53927) doi:10.2196/53927

KEYWORDS

digital public health; digital health; public health; telemedicine; electronic health records; e-prescription; e-referral; e-consultation; e-surveillance; e-vaccination registries; scoping review

Introduction

Background

Digital technologies have become ubiquitous in all facets of life and work [1]. Within health care, digital solutions offer a promising opportunity to make processes and operational methods more efficient and effective [2,3]. This advancement in health care and public health will empower individuals to engage in self-management and actively manage their overall well-being and health [4]. Similarly, digital technologies can facilitate the cost-efficient implementation of health promotion, prevention, monitoring, and surveillance measures, making them easy to access [5-11]. These technologies also foster novel care possibilities, enabling sustainable and equitable access to health services across entire populations [12-16]. Therefore, adopting digital health solutions presents a valuable opportunity to strengthen national health care systems. Eventually, these solutions will transform traditional clinical care and public health structures, fundamentally reshaping health systems on a global scale.

However, to our knowledge, no study has yet attempted to comprehensively map out the heterogeneous landscape of digital public health (DiPH) and digital public health interventions (DiPHIs). Despite existing reviews focusing on selected intervention categories [17], specific target groups such as adolescents [18,19] or patients who are chronically ill with specific medical conditions [20], a comprehensive mapping of interventions in digital health is still missing. In addition, in all existing reviews, the studies analyzed digital health interventions but not DiPHIs, thereby limiting these interventions to their medical and clinical usability. Therefore, this scoping review aims to fill the research gap and comprehensively map out the heterogeneous landscape of DiPHIs. The development, implementation, integration, and evaluation of evidence- and needs-based DiPH interventions require a clear and shared understanding of the specific characteristics of digital health technologies for public health purposes [21]. Conversely, the lack of a standardized lexicon creates communication challenges between the subdisciplines of DiPH [22,23], while concurrently hindering comparability between national health systems.

Definition of DiPH

For this review, DiPH is defined as using information and communication technologies (ICTs) such as digital technologies, services, or tools to achieve traditional public health goals [24] with a population health impact [25]. According to Winslow [26], public health aims at "preventing disease, prolonging life, and promoting physical and mental health and efficiency through organized community efforts..." This definition is still used nowadays by the World Health Organization (WHO) to define essential public health functions. This places public health as the leading discipline in health governance, financing, health information systems, communication, coverage, care, education, and health regulations to protect and promote the health of populations [27].

Hence, we define DiPHI as interventions that address "at least one essential public health function through digital means." [26]. To enhance acceptance among the population, DiPHI should consistently follow a needs- and evidence-based design with a participatory and human-centric approach [21,25,28,29]. The impetus behind their development should not stem solely from technological capabilities but from DiPHIs potential to mitigate disparities and improve access to health services among different groups of a population [22,25,27].

Unlike the domain of *digital health*, which involves the application of health ICT for personalized medicine and telemedicine [30] and its subfield *mobile health*, which primarily focuses on patient monitoring and the use of personal digital assistants in the clinical setting [31], DiPHIs encompass a broader spectrum of services and interventions. These include those in health promotion, such as wearables to promote physical activity; in prevention, such as web-based vaccination registries; in health care, such as telemedicine or electronic health records (EHRs); and in surveillance systems, such as dashboards or tracing systems for infectious disease monitoring on a national level. The goal of DiPHIs is to improve the health and well-being of groups considered vulnerable rather than individuals [25].

By combining the foundational goals of public health with the capabilities of ICT applications, DiPHIs have the potential to substantially strengthen population health, promote well-being, and advance public health objectives. This impact can be realized at the individual, community, and national levels [32].

Study Aims and Objective

This scoping review primarily aims to outline a comprehensive range of proposed to implemented DiPHI, including different levels of prevention, health care, and public health research initiatives. Following the aforementioned definition of DiPHI, the second objective is to map the landscape of the existing DiPHI, shedding light on their self-reported digital health

XSL•FO
RenderX

functions (based on the evidence standards framework [ESF] for digital health technologies from the National Institute for Health and Care Excellence [NICE] [33]) and the addressed essential public health functions as defined by the WHO [27]. Our scoping review will guide future research to address underexplored DiPHI types by identifying research gaps (eg, underrepresented intervention types). The multinational and interdisciplinary comparative analysis of interventions will generate a data set on various interventions, their characteristics, functions, and use cases, which can support researchers or policy makers to make informed decisions for future DiPHIs planning and upscaling.

It is vital to acknowledge that this scoping review, while intended to illustrate the distinctive characteristics of DiPHI, does not seek to provide exhaustive analyses on cost-effectiveness, efficacy, or barriers and facilitators of implementation. The inherent heterogeneity of DiPHI renders it impractical to address these multifaceted aspects within a single literature review [13,14].

Methods

Following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, we registered our review protocol with the PROSPERO database (CRD42021265562) on August 2, 2021, and details and rationale of the protocol were published in 2022 [34].

Eligibility Criteria

The standards for eligibility were constructed in the participants, intervention, and study design format. An overview of the inclusion and exclusion criteria is presented in Textbox 1. Accompanying these criteria, the general requirements of full-text accessibility and the publication language were also considered. The review expanded the publication language limitation to English, German, and Chinese based on the language skills of the reviewer team. Given the scope of our study, which focuses on mapping the characteristics of DiPHI rather than comparing the effectiveness or costs, the review process refrains from specifically incorporating reported outcome variables.

XSL•FO
RenderX

Textbox 1. Eligibility criteria for the scoping review.

Inclusion criteria

- Population
  - The study focuses on the community level or above (ie, regional or national) of the general population.
- Intervention
  - The paper describes a concrete digital public health intervention (DiPHI) that addresses at least 1 essential public health function through digital means as defined by the World Health Organization (WHO).
  - The DiPHI matches our definition of digital public health.
  - The DiPHI addresses at least 1 essential public health function as defined by the WHO.
  - The DiPHI is paid or reimbursed by the government or health insurance.
  - The DiPHI uses the internet or Bluetooth to provide its core function or service.
  - The DiPHI is the central research object of the paper.
- Study design
  - All original peer-reviewed studies, reports, books, book chapters, or peer-reviewed conference papers that have a description of a DiPHI as their primary intervention component.
- Full text
  - Full text was available.

Exclusion criteria

- Population
  - The study population consists of veterans, armed forces, prisoners, inmates, refugees, or asylum seekers.
- Intervention
  - The intervention needs to be privately bought by the user without reimbursement by the government or health insurance.
  - The intervention focuses on background management processes.
- Study design
  - Study protocols, editorials, commentaries, conference proceedings, or reviews (ie, narrative reviews, scoping reviews, systematic reviews, or meta-analyses).
- Full text
  - Full text was not available.

Literature Search

The initial search followed a comprehensive search strategy conducted on February 19, 2021, in CENTRAL, PubMed, and Web of Science. We ran an additional search following the same search strategy on December 1, 2021, for 2 additional databases: ACM Full-Text Collection and IEEE Xplore. The search results for all 5 databases were updated on October 5, 2022. We applied a consistent search string to all databases with adaption according to index and syntax specifications differences. Given the substantial volume of publications aligning with our inclusion criteria, supplementary manual searches were not undertaken. However, we conducted reference list searches for reviews and meta-analyses describing DiPHIs. No additional references were identified through this approach, as all references were already identified through our systematic search.

The search strategy for all databases was structured around 3 themes connected through “AND.” Terms within the themes are combined through “OR” (refer to Figure 1 for the general overview and Multimedia Appendix 1 for the complete search strategy for all databases, including results per term of the initial search). We decided against searching for concrete interventions such as EHRs, specific health or medical apps, or other interventions to reduce the risk of confirmation bias.

We additionally used Medical Subject Headings wherever appropriate during database searches. All other terms were limited to title, abstract, or keywords. No search limits were applied.

XSL•FO
RenderX

Figure 1. The fundamental concept of the search strategy.

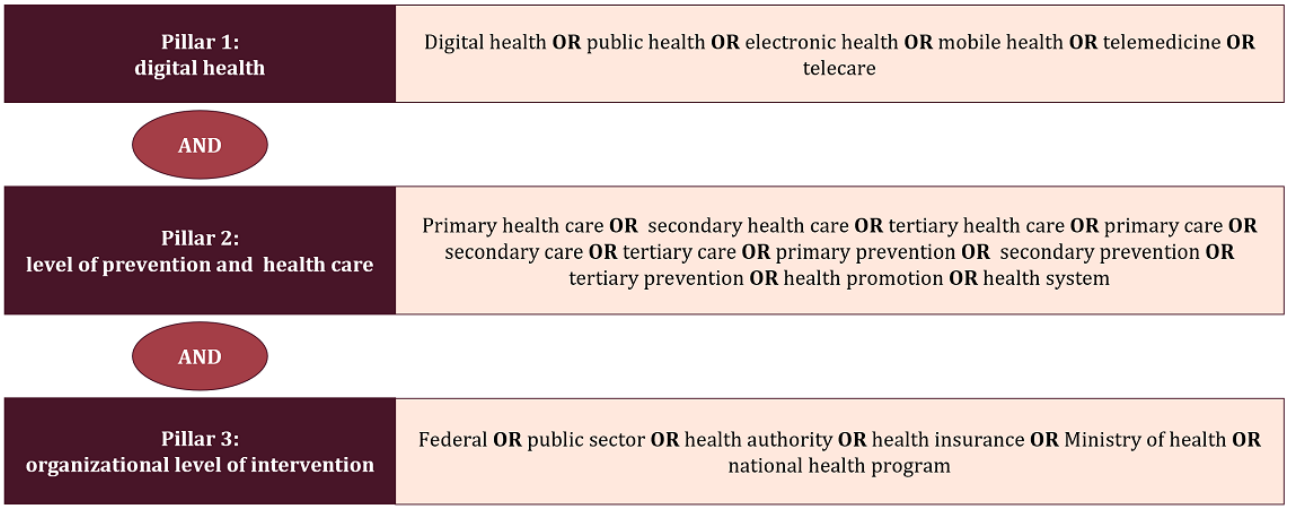

Screening Process

We applied a 3-stage screening strategy (ie, titles, abstracts, and full texts) to identify suitable publications (Figure 2). To reduce the list of reasons for full-text exclusion, the first 3 exclusion criteria mentioned in Textbox 1 were summarized under "wrong setting." Three authors (LM, MF, and KA) were involved in the screening process for eligibility, and each stage had 2 authors partake in the screening independently. After each screening stage, another author (CCP) resolved any disagreements on study selection with independent decisions. The screening was conducted through the free web and mobile app Rayyan (Rayyan Systems Inc) [35]. References published in another language than English, Chinese, or German were excluded during full-text screening. Chinese references were screened only by CCP. Cohen κ for full-text screening was 0.77 (identifying a substantial agreement rate according to the study by Landis and Koch [36]). Full texts were included if the article was published open access, if it was published in a subscription journal licensed by our institution, or if the corresponding author provided a copy. An overview of all excluded full texts, along with the reasons for their exclusion, is included in Multimedia Appendix 2 [37-39].

Figure 2. Flowchart of the search and screening process.

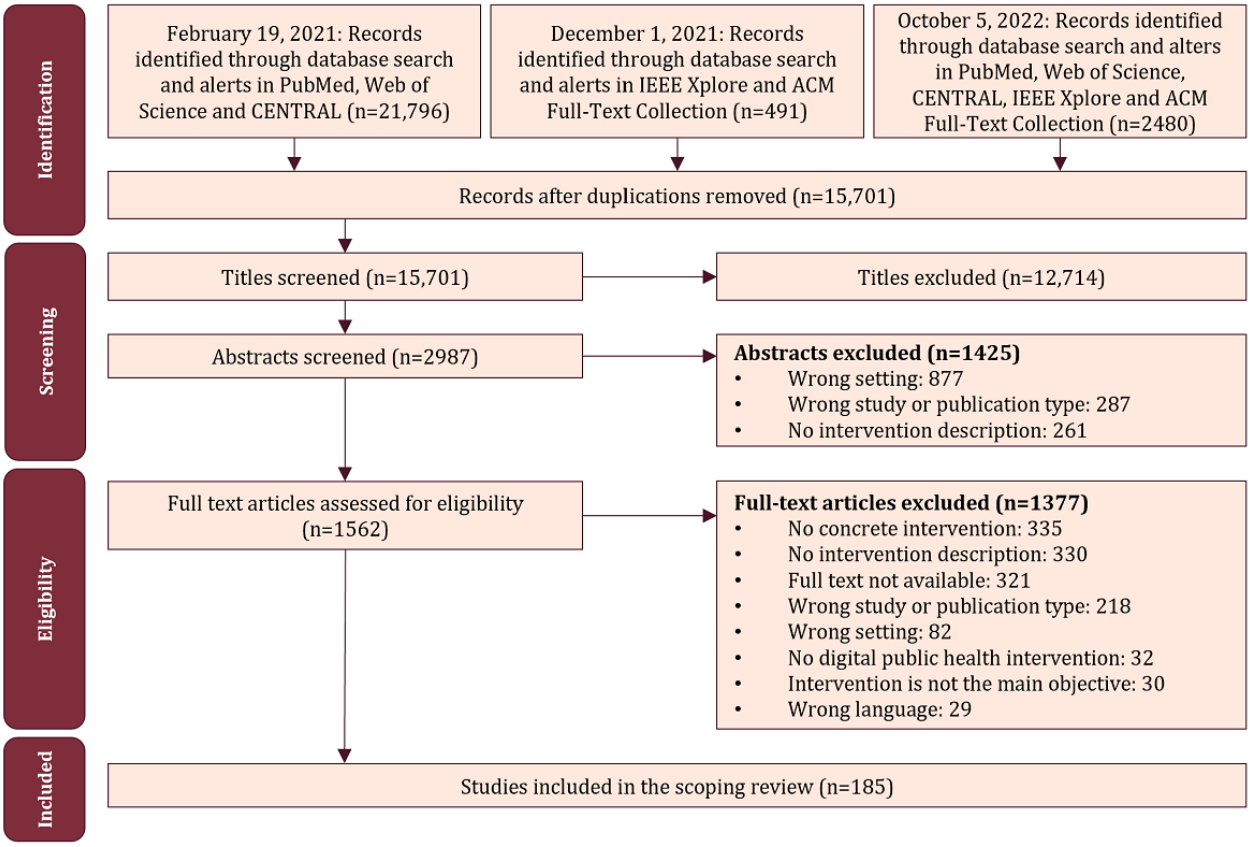

XSL•FO
RenderX

Data Extraction and Qualitative Content Analysis

Two authors (LM and KA) independently extracted data in an Excel 2019 spreadsheet (Microsoft Corp). The extraction sheet was piloted with the first 10 included publications. The extraction for all data except for the technical functions and nontechnical intervention characteristics (which were extracted as free texts) followed a predefined coding table, which was extended if necessary (Table 1). The extraction for all data was based on qualitative description directly provided by the included texts. Discrepancies in data extraction were resolved through discussion between the 2 authors. Missing data for the included publications or data where only a vaguely formulated description was given in the text were requested from authors via email. Where multiple interventions were described within 1 publication, data for each intervention was extracted in a new row. Target groups were defined and extracted based on the digital health application directory by Lantzsch et al [37] due to their precise differentiation between healthy users and patient users. Additional target groups (eg, policy makers, health professionals, or researchers) were added in an iterative process. The digital health functions and the intervention types followed the 2019 edition of the NICE ESF for digital health technologies [33]. The most updated ESF from 2022 reduced its number of individual technical functions and was changed toward a more clinical setting. However, the 2019 framework still included 10 different functions, for example, intervention types covering health promotion, health care, communication, education, and system services. Therefore, we deemed the 2019 version more suitable for our mapping approach. In addition, these functions were supported by the work of Lantzsch et al [37]. The public health functions were based on the 11 WHO functions [27]. Building on our previous research, ESF and ESPHF are merged to map DiPH for this study [29]. Extracting authors selected the option “other” where multiple target groups, functions, or prevention or health care levels appealed.

Target groups were defined and extracted based on the digital health application directory by Lantzsch et al [37] due to their precise differentiation between healthy users and patient users. Extracting authors selected the option “other” where multiple functions or target groups appealed. This decision displays the only deviation from the registered study protocol for this review [34] where we did not define different categories within target groups (the study by Lantzsch et al [37] was published after our study protocol, which underlines the rapidly evolving nature of DiPH). The definitions for prevention and health care (ie, last column in Table 2) were based on the WHO definitions (Table 2).

In addition to the previously mentioned topics, we also extracted information on the study design, publication year and type, intervention name, technical features, and nontechnical characteristics. We did not extract data on outcome measures or type of data collection in the included publications because this review aims to conduct an intervention mapping focusing on their characteristics and not on their effectiveness or costs.

XSL•FO
RenderX

Table 1. Coding table for data extraction.

Study type	Country	Type of intervention	Implementation status	Primary target group	Digital health function	Essential public health function	Level of prevention, health care, and research
Case report	Afghanistan	Health or medical app	Proposed intervention	Health insurances	Active monitoring	Governance	Primary prevention
Case-control study	Albania	Disease surveillance system	Planned intervention	Health professionals	Simple monitoring	Financing	Secondary prevention
Longitudinal study	Algeria	Early warning software	Pilot study	Researchers	Calculate	Human resources	Tertiary prevention
Cross-sectional study	Andorra	Electronic health record	Locally implemented	Policy makers	Communicate	Health information system	Primary health care
Economical study	Angola	Electronic consultation	Regionally implemented	Healthy without known risk factors	Diagnose	Research	Secondary health care
Mixed methods study	Antigua and Barbuda	Electronic medication plan	Nationally implemented	Healthy with known risk factors	Inform	Social participation and health communication	Tertiary health care
Trial or Experimental study	Argentina	Electronic prescribing	Internationally implemented	Acute ill not life threatening	Preventative behavior change	Health promotion	Research
Qualitative study	Armenia	Electronic referral	Website	Chronically ill stable	Self-manage	Health protection	—
__[a]	Austria	Electronic disease registry	—	Highly vulnerable or unstable health status	System service	Disease prevention	—
—	Azerbaijan	Electronic vaccination registry	—	Disaster management professionals	Treatment	Preparedness for public health emergencies	—
—	Bahamas	Health information system	—	—	—	health care	—
—	Bahrain	Implants	—	—	—	—	—
—	Bangladesh	Information website	—	—	—	—	—
—	Barbados	Patient portal	—	—	—	—	—
—	Belarus	Patient to provider communication portal	—	—	—	—	—
—	Belgium	Provider to provider communication portal	—	—	—	—	—
—	Belize	Telemedicine	—	—	—	—	—
—	Benin	Wearable	—	—	—	—	—

[a]Not applicable.

XSL•FO
RenderX

Table 2. Applied definitions for prevention and health care.

Level of prevention or health care	Definition	Source
Primary prevention	Avoiding new diseases, disabilities, or injuries	WHO[a] [40] and Baumann and Ylinen [41]
Secondary prevention	Early detection of diseases, disabilities, or injuries	WHO [40,42] and Baumann and Ylinen [41]
Tertiary prevention	Reducing complication risk of existing diseases, disabilities, or injuries	Baumann and Ylinen [41]
Primary health care	Care for acute mild illness, injuries, or medical problems at a primary care provider with a whole-of-society approach (eg, through a general practitioner)	WHO [40,42], Hopayian [43], and Bodenheimer and Grumbach [44]
Secondary health care	Specialist or emergency medical care through specialized physicians	Hopayian [43] and Bodenheimer and Grumbach [44]
Tertiary health care	Highly specialized medical health care over an extended period in stationary settings, usually in referral hospitals	Hopayian [43] and Bodenheimer and Grumbach [44]

[a]WHO: World Health Organization.

For the geographic analysis of interventions, we counted every intervention multiple times if it was implemented internationally. For instance, Adler et al [45] described an intervention that was implemented in 7 countries. Weighting these interventions per country by the number of countries they were implemented in (here 1/7) would have weakened internationally planned or implemented interventions compared to interventions only targeting 1 country. Therefore, we decided to give full weight to each intervention and count the intervention as 1 for every targeted country. While this approach resulted in an artificially higher number of included interventions, it also ensured that international interventions were not discriminated against based on the geographic analysis.

For the qualitative analysis, we grouped the identified references by intervention types based on our study protocol. We then summarized the extracted data for each intervention type, as illustrated in Table 2. Additional intervention types were added in an inductive procedure where needed. In addition, we developed an inductive category system based on 5 randomly selected publications per intervention type to describe the technical functions and nontechnical characteristics of different intervention types. The qualitative content analysis through iterative coding was then conducted for all references of 1 intervention type in MAXQDA 2022.7 (VERBI GmbH) by LM. The procedure followed the inductive content analysis procedure described by Vears and Gillam [46].

Results

Overview

In total, we identified 15,701 different publications published until October 5, 2022 (24,767 before deduplication), through our systematic search in 5 scientific databases. Of these, 2987 (19.02%) were assessed for abstract screening and 1609 (10.25%) for full-text screening. The process excluded 1414 records, primarily due to unavailable full-text content and a lack of specificity in describing interventions. Of the screened full texts, the remaining 11.5% (185/1609) were considered eligible for this review. Some of these publications described multiple interventions, whereas, in other cases, the same intervention was presented by >1 publication. Of the 185 publications, 96 (51.9%) were case reports that focused on a detailed intervention design or implementation description, while the aim of the remaining 89 (48.1%) publications lay on the intervention evaluation, most often evaluated through cross-sectional studies (n=37, 20%), trials (n=19, 10.3%), or longitudinal studies (n=18, 9.7%). The characteristics and extracted information of all included 185 references are displayed in Multimedia Appendix 3 [38,45,47-229] instead of the main manuscript. As sometimes multiple references are reported on the same intervention, this scoping review mapped the characteristics of 179 different DiPHI for the included 185 publications.

Distribution of Interventions Globally

Figure 3 displays the distribution of the included interventions by country. For this analysis, all interventions were weighted equally regardless of the number of countries they were implemented in. As several interventions were implemented in >1 country, this analysis artificially increased the number of observed interventions to 199. Although most interventions (36/199, 18.1%) were from the United States, continent -wise, Europe took the lead with 38.2% (76/199) reported interventions. Ranked third was Asia with 17.6% (35/199), followed by Africa (26/199, 13.1%), Oceania and Australia (9/199, 4.5%), and South America (6/199, 3%). From a gross domestic product point of view (ie, grouped in high, middle, or low-income countries as classified by the World Bank [230]), most interventions were planned for or implemented in high-income countries (129/199, 64.8%) and were set-up in middle-income countries (58/199, 29.1%). Of the 10 interventions in low-income countries, 3 (30%) targeted Malawi [73-76,82,218], and 2 (20%) each were applied in Afghanistan and Ethiopia [73-76,91,175,222]. Furthermore, of the 10 interventions, 1 (10%) each was planned or implemented in Mali, Sierra Leone, and Uganda [56,111,190].

XSL•FO
RenderX

Figure 3. World map for the number of included interventions by country, continent, and high-income to low-income countries.

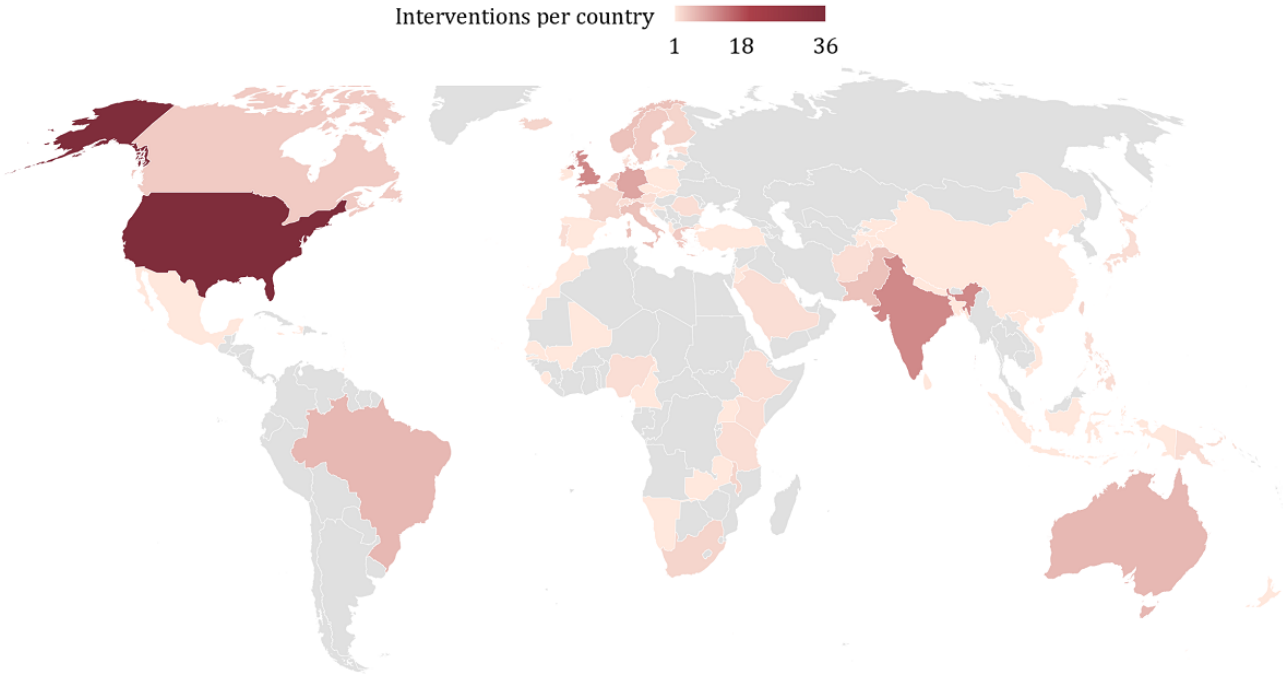

Intervention Types and Implementation Status at the Time of Publication

Under the classification of intervention by NICE ESF, most interventions (49/179, 27.4%) were centered on telemedical services, including telestroke, telecare, and telemonitoring. According to Timpel et al [231], these interventions are characterized using ICT to cover a geographic distance in health care delivery from a health professional to a patient or group of patients. The second most frequently counted interventions were health apps and medical apps (28/179, 15.6%), followed by EHR (23/179, 12.8%). The complete overview of intervention types is displayed in Figure 4, together with their implementation status at the reported time from the included references. As 4 publications reported on the same intervention but at different time phases, all publications were included for this display (eg, Austria's national EHR *ELGA* was only planned when Ströher and Honekamp [199] reported about it, but was already nationally implemented when Herbek et al [113] and Schaller et al [177] described the platform). This results in a total of 181 displayed interventions. Interventions in >1 country were categorized as internationally implemented.

XSL•FO
RenderX

Figure 4. Overview of intervention types and implementation status. Some interventions were described during earlier phases and then again in later periods. Therefore, they are listed once for every reported intervention life cycle in this figure.

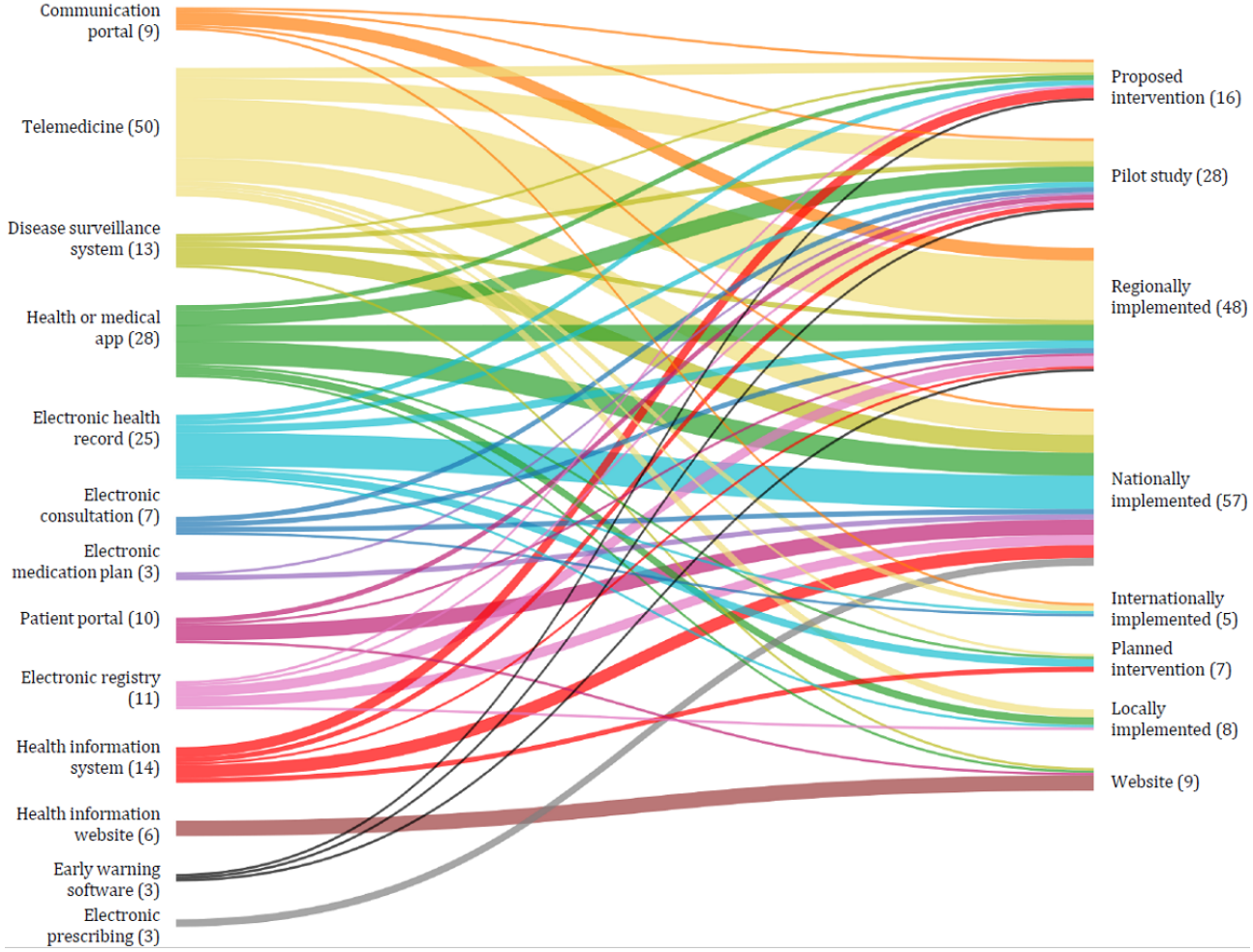

Target Population and Intervention Setting

Our analysis showed that DiPHIs tend to focus on an average of 3 distinct groups within the population. These include health professionals, policy makers, public insurance executives, researchers, disaster management professionals, patients, and healthy individuals. Unsurprisingly, health professionals were the most often targeted group, with 140 (78%) of 179 interventions. They were followed by patients who were chronically ill but stable (101/179, 56.4%). Of all interventions, 51.9% (93/179) were aimed at healthy people with or without risk factors for certain health conditions.

Parallel to the observation of the target population, each intervention affected, on average, 2 different settings, that is, level of prevention, health care, or research as defined previously. Most interventions were health care applications (ie, 100/179, 55.9% for primary, 89/179, 49.7% for secondary, and 55/179, 30.7% for tertiary health care). Out of the 179 interventions, while 16 (8.9%) interventions included a research component [57,84,93,104,128,139,154,163,180,220,221,225], only 3 (1.7%) focused exclusively on research: the Lone Star Stroke Consortium Telestroke Registry, the SAI Databank for electronic health research and evaluation, and the German COVID-19 data donation app (ie, Corona Datenspende) [57,93,128]. For prevention, of the 179 interventions, 13 (7%) exclusively targeted primary prevention (ie, through vaccination registries, behavior change programs, or health education in healthy individuals) [55,62,96,98,101,106,126,140,143,188,208,223,226]. Of the 179 interventions, 8 (4%) focused only on secondary prevention, as digital screening and surveillance programs [48,61,153,161,171,194,201,219], and 3 (2%) improved the health of patients who were chronically ill on a tertiary prevention level [103,147,202]. In contrast, of all included 179 interventions, 10 (6%) were established for primary health care purposes only [49,67,79,109,124,172,178,186,210,212,228,229], 17 (9%) were designed for secondary health care through specialists or emergency use cases [52,54,60,65,78,83,88,95,134,135,173-175,179,189,193,216], and 5 (3%) aimed at tertiary health care in hospital settings [70,109,132,157,200]. Figure 5 gives an overview of the relative size of each target group and the addressed level of prevention, health care, or research per intervention setting.

XSL•FO
RenderX

Figure 5. Addressed target groups and level of prevention, health care, or research in relative distribution per intervention type.

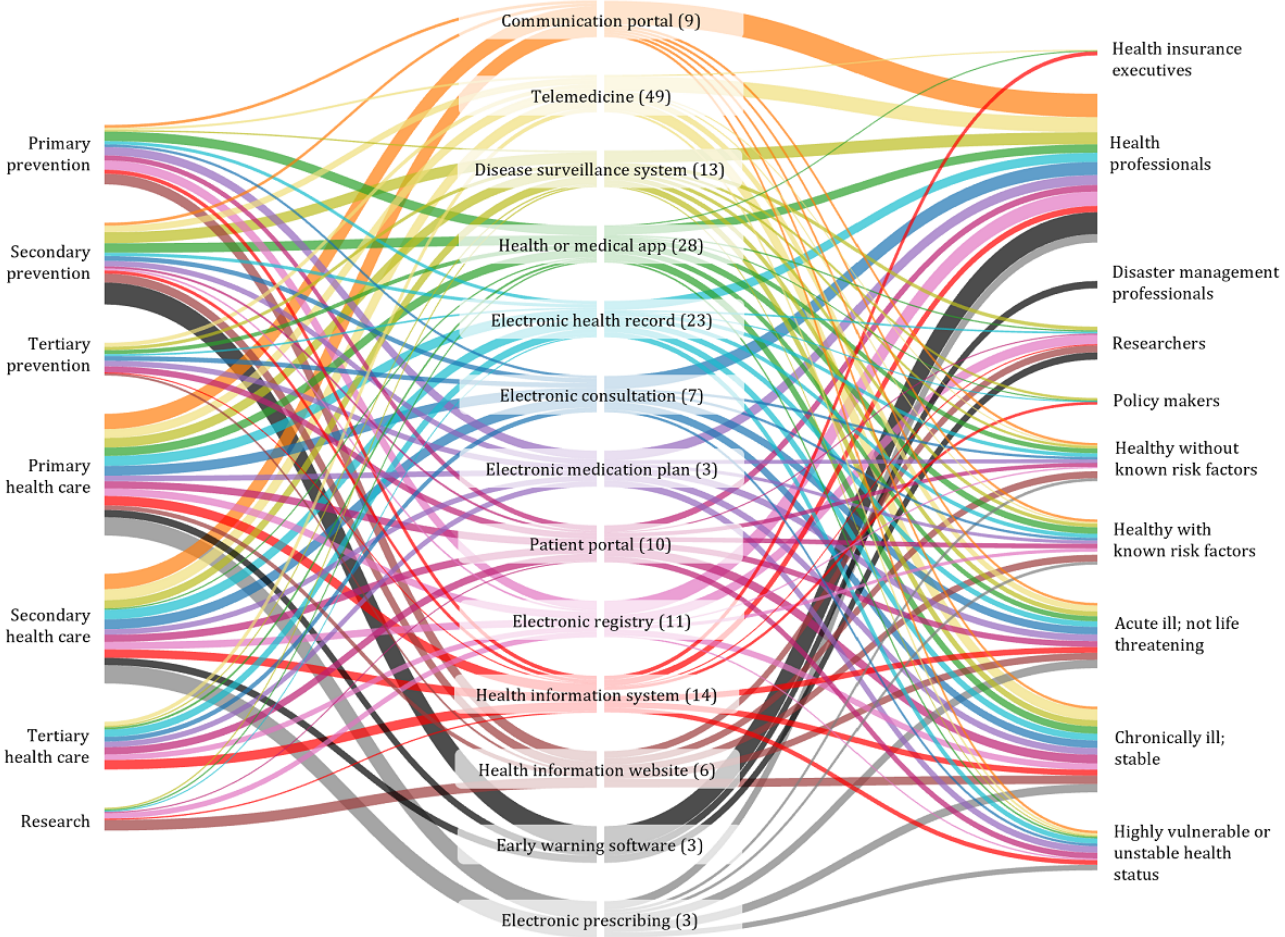

The Intersection of Digital Health Technologies and Essential Public Health Functions in DiPHIs

The analysis of the distribution of EPHFs and digital health technology functions matches the aforementioned focus on health care in DiPHIs. The heat map, displayed in Figure 6, highlights that most interventions were found in the cross-section of health care and diagnostic (64/179, 35.8%), treatment (63/179, 35.2%), communication (61/179, 34.1%), and information (57/179, 31.8%) or as system services in health care (42/179, 23.5%). Relatively few interventions, in contrast, were used for calculation (11/179, 6.1%), preventive behavior change (40/179, 22.3%), or public health research (48/179, 26.8%). A subgroup analysis for all intervention types with at least 20 included references is included in Multimedia Appendix 4. This analysis covers telemedicine, EHR, and health or medical apps, offering an overview of the diversity of topics addressed even within the same intervention category.

XSL•FO
RenderX

Figure 6. Heat map of the essential public health functions and digital health technologies to map digital public health interventions.

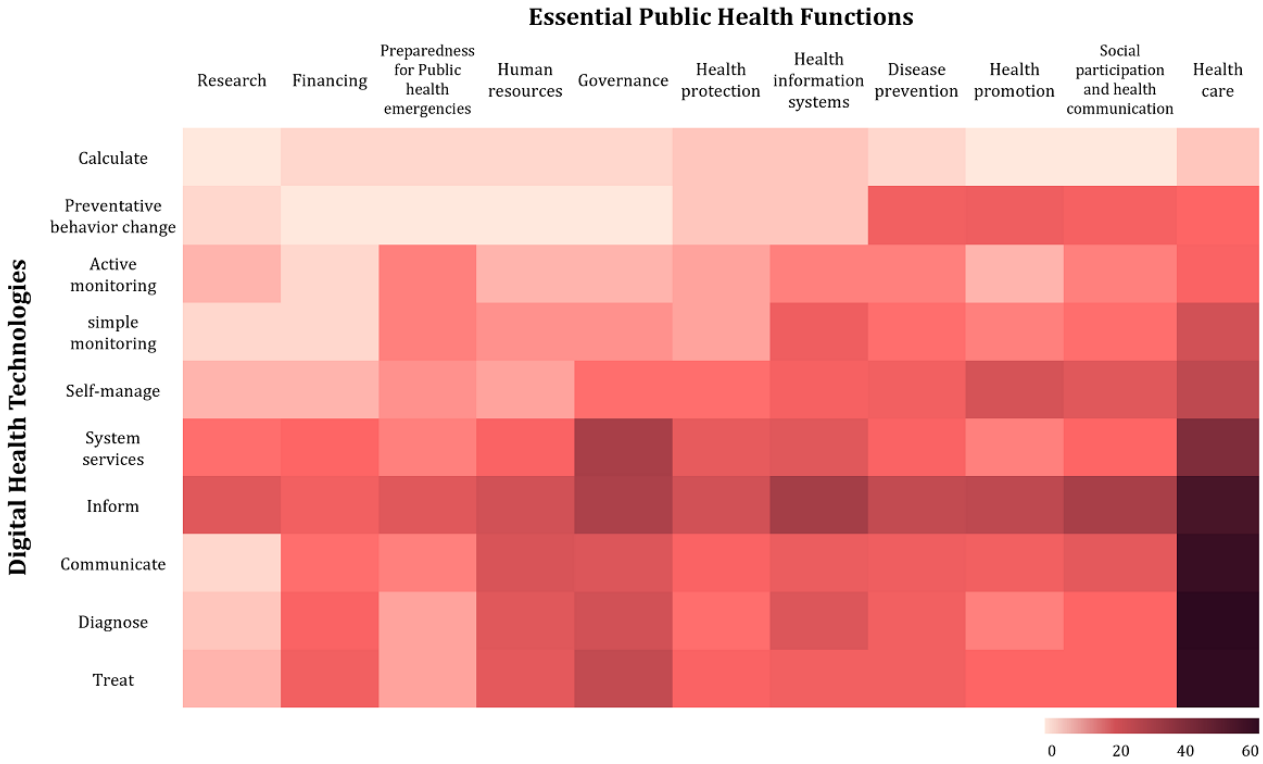

Technological Functions and Nontechnical Intervention Characteristics

Given the sample size, we only provide details on the technical functions and nontechnical characteristics of telemedical interventions and EHRs. We chose not to include health apps or medical apps in this analysis due to the considerable heterogeneity in the wide range of purposes and functions of these apps, as discussed in a previous article [232].

Telemedicine

According to a review by Sood et al [233], which included 104 publications, telemedicine "...uses communications networks for delivery of health care services and medical education from one geographical location to another, primarily to address challenges like uneven distribution and shortage of infrastructural and human resources." The category system includes 5 mayor categories and subcategories for which we assessed all 49 telemedical interventions included in our review [45,47,50,51,60,61,63,65,66,71,85,92,100,101,109,110,115,116,118,119,124,129,134,141,142,145,146,149,151,155,156,159,164,165,174,176,178,179,184,189,193,200,204,206,209-212,227]. The results are displayed in Table 3. Each intervention may exhibit a combination of these features, showcasing how telemedicine is applied in different settings. The most dominant characteristic for included interventions was that the software program was supported by hardware to facilitate clinical care. Besides computers and webcams, this included high-quality cameras, sensors (eg, for blood pressure or blood sugar), ultrasounds, or digital stethoscopes. Nearly half of the interventions (25/49, 51%) described how they used the internet or satellite connections to process and store data. The share of interventions that included a provider-to-provider communication option was slightly lower (17/49, 34%) than those that allowed patient-to-provider communication (23/49, 47%). Nevertheless, 20 (41%) of 49 interventions were characterized by their focus on individual patients, sometimes even through personalized intervention interfaces. However, telemedical interventions may not only focus on clinical care. As stated in the definition, they are also applied to the education context (ie, for patients and health care workers), health promotion, and improving self-efficacy in patients who were chronically ill.

In summary, telemedical interventions exhibit core characteristics that include the use of supporting hardware during consultations, data storage through databases, internet connectivity, bidirectional communication between users, educational modules, user interfaces, and the ability to plan and adjust clinical care plans when needed.

XSL•FO
RenderX

Table 3. Technical functions and nontechnical characteristics of the telemedical interventions (n=49).

Category	Interventions, n (%)	References
Technical infrastructure		
Software supported through hardware tools	41 (84)	[45,47,50,60,61,63,65,66,71,85,92,100,101,109,115,116,118,124,129,134,141,145,146,151,155,156,159,164,174,176,178,184,189,193, 200,204,209-212,227][a-c]
Data acquisition, transmission, and storage in a database	33 (67)	[45,50,51,60,61,63,65,71,85,109,110,116,118,119,124,134,135,145,146,151,155,156,164,174,176, 179,184,189,193,200,206,210][a]
Internet network and connectivity	25 (51)	[47,50,60,61,71,85,100,101,115,118,124,146,151,159,174,178,179,184,189,193,200,206,209,210,212]
Web-based user interfaces	22 (45)	[47,50,51,60,65,85,92,101,110,119,145,149,151,156,159,174,178,184,200,206,210,212]
Data security procedures	18 (37)	[45,47,60,61,66,85,100,101,149,155,159,174,179,189,193,204,206,211,212,227][a]
Centralized server application	16 (33)	[47,50,51,61,85,92,100,109,110,116,119,142,155,156,176,210][b]
Multiaccess, multiprofile web application	16 (33)	[60,66,85,92,100,101,110,118,135,151,156,174,178,184,204,211,212,227][a]
Clinician modules and units	14 (29)	[47,50,51,60,66,92,100,116,124,145,176,189,193,209,211,227][a]
Patient modules and units	6 (12)	[85,92,109,179,209,210][b]
Patient-provider interaction		
Bidirectional communication (patient-provider)	23 (47)	[60,63,92,100,109,116,119,124,135,142,145,146,149,151,156,174,176,178,184,189,200,204,212][c]
Clinical care treatment planning and adjustments	22 (45)	[51,60,61,63,66,85,92,100,109,110,116,118,142,146,149,151,176,189,193,200,204,211,227][a-c]
Bidirectional communication (provider-provider)	17 (35)	[50,71,85,92,116,118,119,124,129,134,145,149,156,176,189,193,210]
Remote monitoring of vital signs and clinical status through monitoring kits and sensors	17 (35)	[50,71,85,92,116,118,119,124,129,134,145,149,156,176,189,193,210]
Compliance and adherence monitoring	11 (22)	[92,100,116,118,124,145,146,149,178,189,204]
Custom and predefined questionnaires	6 (12)	[85,116,145,156,176,204]
User-centered features		
Patient-centered interactions possible	20 (41)	[45,51,60,85,92,109,110,142,145,149,151,155,156,159,164,165,174,204,212][b,c]
Self-monitoring and tracking	10 (20)	[92,101,116,119,124,145,149,156,176,210]
Personalization (eg, avatars, screen names, communication preferences, reminders, and alerts)	8 (16)	[65,85,92,101,149,151,204,210]
Alarm detection and notification	6 (12)	[85,92,116,119,184,206]
Use of patient data		
Data interpretation and analysis	12 (24)	[50,51,60,61,63,65,85,92,100,145,151,156]
Data visualization tool	13 (27)	[45,50,51,66,85,92,101,110,151,176,204,209-211,227][a]
Data- and guideline-driven decision support	11 (22)	[50,66,71,85,92,145,149,151,156,176,193,211,227][a]
Support and engagement		
Education and counseling modules	22 (45)	[45,51,71,92,101,109,116,118,134,135,145,146,149,156,159,164,165,178,189,200,204,206][c]
Counseling and motivation module	10 (20)	[45,92,116,118,141,146,149,155,189,204]
Patient empowerment and self-efficacy	5 (10)	[85,141,145,146,189]

[a]3 studies [66,211,227] described the same intervention.

[b]Intervention 1 by Hartvigsen et al [109].

[c]Intervention 2 by Hartvigsen et al [109].

XSL•FO
RenderX

EHRs

Following the broadly accepted definition by the International Organization for Standardization (2019), an EHR is "a data repository regarding the health and health care of a subject of care where all information is stored on electronic media" [234]. According to this terminology, we developed a category system with 4 major topics and subcategories, which was applied to the 26 publications that described 23 different EHR systems [67,69,81,83,87,89-91,104,105,107,113,114,123,125,162,166,177,181,182,185,195,199,205,215,221,224]. The results are displayed in Table 4. Nearly all interventions (18/23, 78%) allowed health care professionals access to the EHR after the patient consented or an active relationship (eg, due to treatment of chronic diseases) existed between the 2 parties. Although not as often described, most EHR systems (14/23, 61%) stated that patients were able to access their data. Data sharing was possible in 61% (14/23) of the interventions, and another 61% (14/23) described data protective procedures for accessing and sharing data. Although fewer times, interoperability with other health care systems was described (11/23, 48%). We observed differences regarding the stored and accessible data. Unsurprisingly, nearly all interventions (21/23, 91%) reported on storing health and clinical data. However, only 30% (7/23) of the interventions included personal data on the record holder or their immunization records. Only 17% (4/23) of the interventions allowed clinicians to upload their notes to the system.

XSL•FO
RenderX

Table 4. Technical functions and nontechnical characteristics of the electronic health records (EHRs; n=23).

Category	Interventions, n (%)	References
Technical infrastructure		
Integration and interoperability with external health care systems	11 (48)	[69,87,91,104,107,113,123,162,177,181,195,199,205,215,221][a,b]
Mobile platform	7 (30)	[87,90,113,166,177,185,199,205,221][b,c]
Centralized server	6 (26)	[81,104,113,114,162,177,182,195,199,215][a,b]
Communication protocols	6 (26)	[69,81,113,177,181,182,199,221][b,c]
Cloud-based system	4 (17)	[87,90,185,205,221][c]
Data synchronization	3 (13)	[81,185,205,221]
Data transfer and privacy		
Data sharing possible	14 (61)	[67,69,81,87,91,104,105,107,113,123,162,177,181,182,199,215,221][a-c]
Data security protocols embedded	14 (61)	[69,87,89,90,104,107,113,123,125,162,166,177,181,195,199,205,215,221][a-c]
Opt-out approach	4 (17)	[67,104,113,162,177,199,215,221][a-c]
Opt-in approach	4 (17)	[69,83,182,205,221][c]
Stored data		
Medical history (eg, diagnoses, medical procedures, laboratory results, and treatments)	21 (91)	[69,81,83,87,89,91,104,105,107,113,114,123,125,162,166,177,181,182,185,195,199,205,215,221,224][a-c]
Medication plan	14 (61)	[69,83,89,91,105,107,113,114,123,166,177,181,182,199,205,221][b,c]
Patient demographic data	7 (30)	[69,89,91,113,125,166,177,195,199][b]
Immunization records	7 (30)	[69,91,105,125,185,195,205,221][c]
Financial and administrative data related to health care services	5 (22)	[91,114,205,221,224][c]
Clinician notes	4 (17)	[105,107,224]
EHR system features and modules		
Health professional access to patient data	18 (78)	[67,69,87,89,104,105,107,113,123,125,162,166,177,181,182,185,195,199,205,215,221,224][a-c]
Patient access to data	14 (61)	[69,87,89-91,104,105,113,162,166,177,181,182,199,205,215,221,224][a-c]
User interfaces	11 (48)	[69,87,89,90,113,123,166,177,181,185,195,199,205,221][b,c]
Clinical case–finding features	8 (35)	[67,104,107,113,114,123,162,177,181,195,199,215][a,b]
Patient management modules	3 (13)	[114,166,195]
Data used for real-time surveillance and monitoring	3 (13)	[81,87,195]

[a]3 studies [104,162,215] described the same intervention.

[b]3 studies [113,177,199] described the same intervention.

[c]2 studies [205,221] described the same intervention.

Discussion

Principal Findings

Our scoping review has identified a substantial number of publications, underscoring the substantial interest and activity in DiPH. The wealth of publications illustrates the broad spectrum of interventions being explored and implemented worldwide. This diversity reflects the evolving landscape of health care delivery and disease prevention in the digital age.

The distribution of DiPHIs across countries and continents showcases the worldwide interest in leveraging technology to enhance health care but also points out significant differences in implementation adaptation between high-income and low-income countries. A reason for this imbalance negatively impacting the global digital divide might lie in the inequality regarding the digital advance among developed countries compared to developing countries [235,236]. The prevalence of interventions in high-income countries may partly be

XSL•FO
RenderX

attributed to more significant resources and infrastructure [237]. This divide is highlighted by a lack of funding, inadequate computer availability, and limited internet skills that limit the expansion and applicability of ICT and digital (public) health [238]. Still, interventions in low-income countries signal the potential for digital technologies to bridge health care gaps in resource-constrained settings and aging societies [237]. Indeed, the Global Digital Health Monitor covers our findings by displaying overall scores for countries regarding their digital health maturity. Developing countries (especially in Africa) rank lower than their developed counterparts in Europe or North America [239]. More rigorous public health research and practice efforts are needed to empower developing countries to apply digital tools for health-related purposes. The WHO's Digital Health Atlas can serve as such an effort where project leads can voluntarily register their digital health intervention free of charge [240]. Maps such as these will support public health policy, research, and practice in better resource allocation (eg, the map can display already existing subnational initiatives that could be scaled up to the national or international level without creating double-infrastructures), benchmarking between interventions and further developing intervention types. However, the explicit distinction between medically focused digital health interventions and population-centric DiPHIs will be needed for future versions of the atlas.

Our analysis shows that DiPHIs often target multiple population groups and settings. Health professionals are a primary target, reflecting technology integration into health care practice and education. The emphasis on patients who are chronically ill but stable underscores the potential of digital interventions to improve care for this population. Moreover, interventions in diverse prevention, health care, and research settings illustrate their versatility across the health care continuum. This observation was supported by the heat- map representation of EPHFs and ESFs, highlighting a significant concentration of digital interventions in health care–related functions such as diagnostics, treatment, communication, and information. This reflects the predominant role of digital interventions in enhancing clinical care and health care management. However, there is room for further exploration of interventions addressing prevention and public health research functions.

Health apps, telemedical services, and EHRs emerged as the dominant intervention types in our review. Health apps are well known for their heterogeneity [232], which provides the potential to address all areas of public health, including health promotion, clinical care, and rehabilitation [241]. Telemedical services stood out as a substantial focus of our included references, aligning with their capacity to extend universal health coverage reach across geographic boundaries and increase health care delivery speed while lowering treatment costs and response times to control disease outbreaks [242]. EHRs, in contrast, can be essential tools for public health surveillance in providing data on the population's health status and the health care system's performance [39,114,243]. These findings highlight the multifaceted nature of DiPHIs and their adaptability to various health care needs and contexts in public health.

Overall, the analysis underscores the heterogeneity within DiPHIs, even within the same intervention category, as we have displayed through a qualitative analysis of included publications for telemedical interventions and EHRs. This diversity necessitates precise definitions and standardized terminologies to facilitate effective communication, collaboration, and evaluation. Clear definitions can also aid in benchmarking and comparing interventions across different health care systems. The detailed analysis of technological functions and nontechnical characteristics of both intervention types revealed the complexity and variation within each intervention type. Identifying core characteristics and added features can guide intervention development and enhance our understanding of their capabilities. In 2019, WHO published their recommendations on digital interventions for the clinical setting, which provides a definition for each intervention (eg, client-to-provider telemedicine), synonyms for these tools, and recommendations regarding their use cases [244]. While this overview serves as a good starting point, it requires extensions for nonclinical settings and an overview of core functions and technical features to allow benchmarking. The heterogeneity in DiPHI forces us to rethink how national health system assessments should be conducted to allow for comparable results. Instead of asking, *Does your country have an EHR system?"* it is critical to ask for the availability of an intervention with specific characteristics: "*Does your country have an intervention with the following characteristics, targeting these population groups, and being applied for at least one of the following settings?"*

As the field continues to evolve, it is essential to prioritize rigorous research and evaluation to assess these interventions' effectiveness, safety, and impact on population health and health care accessibility. This review serves as a valuable resource for policy makers, researchers, and health care practitioners seeking to navigate the dynamic and diverse field of DiPHIs.

Strengths and Limitations

Originally conceived as a systematic review, our research protocol was registered with PROSPERO to conduct a comprehensive synthesis of the existing literature on DiPHIs. The systematic approach is typically characterized by a rigorous process involving reviewing evidence to answer narrowly defined research questions. However, as our study progressed, we recognized that our research question and the evolving DiPH landscape were better suited for a scoping review methodology. While systematic reviews focus on specific research questions and outcomes, the strengths of scoping reviews for effectively mapping and characterizing interdisciplinary and comprehensive topics (such as the diverse spectrum of DiPHIs in our case, including their characteristics, settings, and target populations) [245] were more applicable for our study aim. This change allowed us to adapt to the dynamic nature of the field. It provided a broader perspective essential for understanding the multifaceted nature of digital interventions. This transition underscores the importance of selecting the most suitable methodology to achieve comprehensive insights into the subject matter. The selection of our information sources demonstrated the interdisciplinarity of DiPH by combining literature databases from public health and computer science. This rigorous strategy mitigates the risk of missing essential publications. The broader inclusion of publications from 3 languages (ie, English, German,

XSL•FO
RenderX

and Chinese) is another strength of our study, given the rising number of publications on digital health and DiPH from German-speaking and Chinese-speaking countries. Throughout the process, a minimum of 2 researchers independently screened all references, adhering to the 4-eye principle. Any disagreements in decisions regarding inclusion or exclusion were resolved by another author (CCP) to ensure the reliability of the results. This study followed the PRISMA-ScR [245] to provide high-quality, transparent research results. The filled checklist is provided in Multimedia Appendix 5.

Nevertheless, while this scoping review exhibits numerous strengths, it also presents certain limitations. We did not create publication alerts for the selected databases, limiting our results to references published before October 6, 2022. However, we conducted manual reference list checks for reviews and meta-analyses to identify additional publications that were potentially missed in our search. As no further relevant findings were added from the manual search, we are confident that our scoping review draws a holistic picture of the landscape of DiPHIs globally. Furthermore, we are convinced that the included references are sufficient for an initial intervention mapping of DiPHIs. Nonetheless, it is worth noting that certain intervention types might be underrepresented, such as early warning software, electronic medication plans, electronic prescribing, or emerging intervention types that were not yet reported during our search (eg, artificial intelligence tools). Here, we invite other researchers to build upon our first results and continue the mapping process. Following the PRISMA-ScR [245], assessing the quality of the included publications was not applicable, which could lead to questions regarding the reporting accuracy of the included publications. Nevertheless, as we focused solely on the functions and use cases of the interventions and not on their outcomes, this limitation remains minor.

Furthermore, the challenge of intervention specificity emerged as a notable consideration. The varying levels of detail in the descriptions of digital interventions in the literature posed difficulties in discerning their precise nature. This lack of specificity hindered the ability to categorize and analyze interventions effectively, as it often left critical questions unanswered. For instance, understanding the core technological functions, nontechnical attributes, determined use cases, and defined user groups of a given digital intervention is essential for its accurate classification and evaluation. Although in our case, data extraction was conducted by 2 authors (LM and KA) independently and authors were contacted in case of uncertainty, the absence of such information generally can impede efforts to compare interventions, assess their impact, and make informed decisions regarding their implementation or scalability. This is why we advocate for clearer reporting standards for digital health and DiPH interventions. Finally, we must point out that although the volume of included publications was substantial, a distinct portion could not be included in our review due to factors such as unavailable full-text content and a lack of specificity in intervention descriptions. This highlights challenges in accessing relevant literature and emphasizes the need for more detailed reporting in published studies, particularly regarding the specifics of digital interventions. In an era where timely access to information is critical, the inaccessibility of full-text articles poses an obstacle to researchers and stakeholders seeking to explore and build upon existing knowledge in DiPHIs.

Further Research

Due to the heterogeneity of DiPHIs, it was not feasible for this scoping review to report detailed information for every intervention type. Therefore, we encourage other researchers to uptake our methodology, conduct reviews on specific intervention types, and refine our initial intervention mapping results. More critically, a shared and multidisciplinary understanding of key terminology in digital health and DiPH is needed. Consequently, consensus finding methods (eg, Delphi approaches) need to be applied to internationally create a shared understanding of DiPHI.

In addition, assessments of the sustainability of DiPHIs are crucial to map the implementation status. As Iwelunmor et al [246] and Kaboré et al [247] already highlighted, the sustainability of DiPHIs is an essential component of the whole life cycle of implemented interventions. Therefore, it is crucial to consider whether an intervention has been sustainably implemented or stopped after the initial pilot funding phase has ended. Further research should be conducted to analyze, through a longitudinal design or pre-post comparison, how sustainably DiPHIs are implemented and whether differences between intervention types exist.

Conclusions

This scoping review has shown that DiPHIs are distinctly diverse in their use cases between individual intervention groups and within such groups. Therefore, using precise terminology when planning, implementing, or evaluating DiPHIs is crucial. For instance, instead of phrasing an intervention as an EHR, one should ask for specific technological functions and nontechnical characteristics, determine use cases (from a digital health and public health perspective), and define user groups. This approach facilitates cross-disciplinary collaboration in intervention development and research and fosters international benchmarks and comparability across global DiPH systems. These insights offer valuable guidance for policy makers, researchers, and health care practitioners interested in using the potential of digital interventions to improve population health and health care accessibility.

Acknowledgments

The authors gratefully acknowledge the support of the Leibniz ScienceCampus Bremen Digital Public Health. The study is part of the Leibniz ScienceCampus Bremen Digital Public Health, which is jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen, and the Leibniz Institute for Prevention Research and Epidemiology (BIPS).

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LM conceptualized and designed the study, including its methodology, with input from CCP. LM served as the principal investigator and project administrator for the entire project. LM, KA, and MF conducted the literature review (data collection), with CCP settling disagreements during the screening process. LM and KA independently conducted the data analysis, with support from CCP. LM and CCP conducted the interpretation of the study data. LM drafted the manuscript, and this paper was revised critically for important intellectual content by all authors. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete search strategies for the PubMed, Web of Science, CENTRAL, IEEE Xplore, and ACM Full-Text Collection databases.
[PDF File (Adobe PDF File), 272 KB - jmir_v26i1e53927_app1.pdf]

Multimedia Appendix 2
Publications excluded during full-text screening.
[PDF File (Adobe PDF File), 825 KB - jmir_v26i1e53927_app2.pdf]

Multimedia Appendix 3
Overview of the included interventions.
[XLSX File (Microsoft Excel File), 42 KB - jmir_v26i1e53927_app3.xlsx]

Multimedia Appendix 4
Heat maps of the selected intervention types.
[PDF File (Adobe PDF File), 154 KB - jmir_v26i1e53927_app4.pdf]

Multimedia Appendix 5
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for this manuscript.
[PDF File (Adobe PDF File), 221 KB - jmir_v26i1e53927_app5.pdf]

References

1. Susskind R, Susskind D. The Future of the Professions: How Technology Will Transform the Work of Human Experts. Oxford, UK: Oxford University Press; 2015.
2. Moller AC, Merchant G, Conroy DE, West R, Hekler E, Kugler KC, et al. Applying and advancing behavior change theories and techniques in the context of a digital health revolution: proposals for more effectively realizing untapped potential. J Behav Med 2017 Feb;40(1):85-98 [FREE Full text] [doi: 10.1007/s10865-016-9818-7] [Medline: 28058516]
3. Németh O, Simon F, Benhamida A, Kivovics M, Gaál P. eHealth, teledentistry and health workforce challenges: results of a pilot project. BMC Oral Health 2022 Dec 01;22(1):552 [FREE Full text] [doi: 10.1186/s12903-022-02603-6] [Medline: 36456948]
4. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. J Diabetes Sci Technol 2017 Sep;11(5):1015-1027 [FREE Full text] [doi: 10.1177/1932296817713506] [Medline: 28560898]
5. Strain T, Wijndaele K, Brage S. Physical activity surveillance through smartphone apps and wearable trackers: examining the UK potential for nationally representative sampling. JMIR Mhealth Uhealth 2019 Jan 29;7(1):e11898 [FREE Full text] [doi: 10.2196/11898] [Medline: 30694198]
6. Zeeb H, Pigeot I, Schüz B. [Digital public health-rapid technological progress, but many open public health questions]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):135-136 [FREE Full text] [doi: 10.1007/s00103-020-03092-0] [Medline: 31960072]
7. Bailey JV, Murray E, Rait G, Mercer CH, Morris RW, Peacock R, et al. Interactive computer-based interventions for sexual health promotion. Cochrane Database Syst Rev 2010 Sep 08(9):CD006483 [FREE Full text] [doi: 10.1002/14651858.CD006483.pub2] [Medline: 20824850]

XSL•FO
RenderX

8. Murray E, Burns J, See TS, Lai R, Nazareth I. Interactive health communication applications for people with chronic disease. Cochrane Database Syst Rev 2005 Oct 19(4):CD004274. [doi: 10.1002/14651858.CD004274.pub4] [Medline: 16235356]
9. Dalton CB. Enablers of innovation in digital public health surveillance: lessons from Flutracking. Int Health 2017 May 01;9(3):145-147 [FREE Full text] [doi: 10.1093/inthealth/ihx009] [Medline: 28582555]
10. Desai AN, Kraemer MU, Bhatia S, Cori A, Nouvellet P, Herringer M, et al. Real-time epidemic forecasting: challenges and opportunities. Health Secur 2019;17(4):268-275 [FREE Full text] [doi: 10.1089/hs.2019.0022] [Medline: 31433279]
11. Shakeri Hossein Abad Z, Kline A, Sultana M, Noaeen M, Nurmambetova E, Lucini F, et al. Digital public health surveillance: a systematic scoping review. NPJ Digit Med 2021 Mar 03;4(1):41 [FREE Full text] [doi: 10.1038/s41746-021-00407-6] [Medline: 33658681]
12. Alwashmi M, Hawboldt J, Davis E, Marra C, Gamble JM, Abu Ashour W. The effect of smartphone interventions on patients with chronic obstructive pulmonary disease exacerbations: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016 Sep 01;4(3):e105 [FREE Full text] [doi: 10.2196/mhealth.5921] [Medline: 27589898]
13. Gunasekeran DV, Tseng RM, Tham YC, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021 Feb 26;4(1):40 [FREE Full text] [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
14. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019 Oct 01;29(Supplement_3):13-17 [FREE Full text] [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
15. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug 07;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
16. Wynn M, Garwood-Cross L, Vasilica C, Griffiths M, Heaslip V, Phillips N. Digitizing nursing: a theoretical and holistic exploration to understand the adoption and use of digital technologies by nurses. J Adv Nurs 2023 Oct 02;79(10):3737-3747. [doi: 10.1111/jan.15810] [Medline: 37530425]
17. Xue J, Zhang B, Zhao Y, Zhang Q, Zheng C, Jiang J, et al. Evaluation of the current state of chatbots for digital health: scoping review. J Med Internet Res 2023 Dec 19;25:e47217 [FREE Full text] [doi: 10.2196/47217] [Medline: 38113097]
18. Kouvari M, Karipidou M, Tsiampalis T, Mamalaki E, Poulimeneas D, Bathrellou E, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022 Feb 14;24(2):e30675 [FREE Full text] [doi: 10.2196/30675] [Medline: 35156934]
19. Yun J, Shin J, Lee H, Kim DJ, Choi IY, Kim M. Characteristics and potential challenges of digital-based interventions for children and young people: scoping review. J Med Internet Res 2023 Apr 14;25:e45465 [FREE Full text] [doi: 10.2196/45465] [Medline: 37058340]
20. Fournier V, Duprez C, Grynberg D, Antoine P, Lamore K. Are digital health interventions valuable to support patients with cancer and caregivers? An umbrella review of web-based and app-based supportive care interventions. Cancer Med 2023 Dec 08;12(23):21436-21451 [FREE Full text] [doi: 10.1002/cam4.6695] [Medline: 37937812]
21. Zeeb H, Pigeot I, Schüz B, Leibniz-WissenschaftsCampus Digital Public Health Bremen. [Digital public health-an overview]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb 09;63(2):137-144. [doi: 10.1007/s00103-019-03078-7] [Medline: 31919531]
22. Darmann-Finck I, Rothgang H, Zeeb H. [Digitalization and health sciences - white paper digital public health]. Gesundheitswesen 2020 Jul 22;82(7):620-622 [FREE Full text] [doi: 10.1055/a-1191-4344] [Medline: 32698204]
23. Taj F, Klein MC, van Halteren A. Digital health behavior change technology: bibliometric and scoping review of two decades of research. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
24. Odone A, Buttigieg S, Ricciardi W, Azzopardi-Muscat N, Staines A. Public health digitalization in Europe: EUPHA vision, action and role in digital public health. Eur J Public Health 2019 Oct 01;29(Supplement_3):28-35 [FREE Full text] [doi: 10.1093/eurpub/ckz161] [Medline: 31738441]
25. Wong BL, Maaß L, Vodden A, van Kessel R, Sorbello S, Buttigieg S, European Public Health Association (EUPHA) Digital Health Section. The dawn of digital public health in Europe: implications for public health policy and practice. Lancet Reg Health Eur 2022 Mar;14:100316 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100316] [Medline: 35132399]
26. Winslow CE. The untilled fields of public health. Science 1920 Jan 09;51(1306):23-33. [doi: 10.1126/science.51.1306.23] [Medline: 17838891]
27. Essential public health functions, health systems and health security. Developing conceptual clarity and a WHO roadmap for action. World Health Organization. 2018. URL: https://iris.who.int/bitstream/handle/10665/272597/9789241514088-eng.pdf?sequence=1 [accessed 2024-04-29]
28. van Kessel R, Hrzic R, O'Nuallain E, Weir E, Wong BL, Anderson M, et al. Digital health paradox: international policy perspectives to address increased health inequalities for people living with disabilities. J Med Internet Res 2022 Feb 22;24(2):e33819 [FREE Full text] [doi: 10.2196/33819] [Medline: 35191848]

XSL•FO
RenderX

29. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022 Jun 28;24(6):e31921 [FREE Full text] [doi: 10.2196/31921] [Medline: 35763320]
30. What is digital health? U.S. Food & Drug Administration. 2020 Sep 22. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2024-06-09]
31. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. WHO Global Observatory for eHealth. 2011. URL: https://iris.who.int/bitstream/handle/10665/44607/9789241564250_eng.pdf?sequence=1&isAllowed=y [accessed 2024-04-29]
32. Murray CJ, Alamro NM, Hwang H, Lee U. Digital public health and COVID-19. Lancet Public Health 2020 Sep;5(9):e469-e470. [doi: 10.1016/s2468-2667(20)30187-0]
33. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence (NICE). 2019. URL: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf [accessed 2024-06-09]
34. Maaß L, Pan CC, Freye M. Mapping digital public health interventions among existing digital technologies and internet-based interventions to maintain and improve population health in practice: protocol for a scoping review. JMIR Res Protoc 2022 Mar 31;11(3):e33404 [FREE Full text] [doi: 10.2196/33404] [Medline: 35357321]
35. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
37. Lantzsch H, Panteli D, Martino F, Stephani V, Seißler D, Püschel C, et al. Benefit assessment and reimbursement of digital health applications: concepts for setting up a new system for public coverage. Front Public Health 2022 Apr 21;10:832870 [FREE Full text] [doi: 10.3389/fpubh.2022.832870] [Medline: 35530738]
38. Andriuskevicius E, Dobravolskas E, Punys V, Sinciene V, Valentinavicius A. Architecture for national e-health infrastructure in Lithuania. In: Hasman A, Haux R, VanderLei J, DeClercq E, France FH, editors. Ubiquity: Technologies for Better Health in Aging Societies. Amsterdam, the Netherlands: Ios Press; 2006:421.
39. Elliott AF, Davidson A, Lum F, Chiang MF, Saaddine JB, Zhang X, et al. Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions in the United States. Am J Ophthalmol 2012 Dec;154(6 Suppl):S63-S70 [FREE Full text] [doi: 10.1016/j.ajo.2011.10.002] [Medline: 23158225]
40. Health promotion and disease prevention through population-based interventions, including action to address social determinants and health inequity. World Health Organization. 2023. URL: https://www.emro.who.int/about-who/public-health-functions/health-promotion-disease-prevention.html [accessed 2024-06-09]
41. Baumann LC, Ylinen A. Prevention: primary, secondary, tertiary. In: Gellman MD, editor. Encyclopedia of Behavioral Medicine. New York, NY: Springer New York; 2017:1-3.
42. Primary health care. World Health Organization. 2023. URL: https://www.who.int/health-topics/primary-health-care#tab=tab_1 [accessed 2024-06-09]
43. Hopayian K. Health care systems: primary, secondary, tertiary and quaternary care. J Fam Med Med Sci Res 2022;11(4):1000133. [doi: 10.37532/2327-4972.22.11.133]
44. Bodenheimer T, Grumbach K. How health care is organized—I: primary, secondary, and tertiary care. In: Bodenheimer T, Grumbach K, editors. Understanding Health Policy: A Clinical Approach. New York, NY: McGraw Hill; 2020.
45. Adler E, Alexis C, Ali Z, Allen U, Bartels U, Bick C, et al. Bridging the distance in the Caribbean: telemedicine as a means to build capacity for care in paediatric cancer and blood disorders. Stud Health Technol Inform 2015;209:1-8. [Medline: 25980698]
46. Vears DF, Gillam L. Inductive content analysis: a guide for beginning qualitative researchers. Focus Health Prof Educ 2022 Mar 31;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
47. Agarwal S, Day DJ, Sibson L, Barry PJ, Collas D, Metcalf K, et al. Thrombolysis delivery by a regional telestroke network--experience from the U.K. National Health Service. J Am Heart Assoc 2014 Feb 26;3(1):e000408 [FREE Full text] [doi: 10.1161/JAHA.113.000408] [Medline: 24572251]
48. Akintoye S, Ogoh G, Krokida Z, Nnadi J, Eke D. Understanding the perceptions of UK COVID-19 contact tracing app in the BAME community in Leicester. J Inf Commun Ethics Soc 2021 Dec 07;19(4):521-536. [doi: 10.1108/jices-06-2021-0071]
49. Alharbi A. Knowledge, attitude and practice toward the mHealth app Mawid: a cross-sectional study. Int Health 2023 May 02;15(3):342-350 [FREE Full text] [doi: 10.1093/inthealth/ihac062] [Medline: 36107975]
50. Alis C, del Rosario C, Buenaobra B, Mar Blanca C. Lifelink: 3G-based mobile telemedicine system. Telemed J E Health 2009 Apr;15(3):241-247. [doi: 10.1089/tmj.2008.0098] [Medline: 19382861]
51. Alkmim MB, Silva CB, Figueira RM, Santos DV, Ribeiro LB, da Paixão MC, et al. Brazilian national service of telediagnosis in electrocardiography. Stud Health Technol Inform 2019 Aug 21;264:1635-1636. [doi: 10.3233/SHTI190571] [Medline: 31438267]
52. Anderson D, Villagra V, Coman EN, Zlateva I, Hutchinson A, Villagra J, et al. A cost-effectiveness analysis of cardiology eConsults for medicaid patients. Am J Manag Care 2018 Jan 01;24(1):e9-16 [FREE Full text] [Medline: 29350511]

XSL•FO
RenderX

53. Ando T, Mori R, Takehara K, Asukata M, Ito S, Oka A. Correction: effectiveness of pediatric teleconsultation to prevent skin conditions in infants and reduce parenting stress in mothers: randomized controlled trial. JMIR Pediatr Parent 2022 Apr 12;5(2):e38059 [FREE Full text] [doi: 10.2196/38059] [Medline: 35412999]
54. Angileri FF, Cardali S, Conti A, Raffa G, Tomasello F. Telemedicine-assisted treatment of patients with intracerebral hemorrhage. Neurosurg Focus 2012 Apr;32(4):E6. [doi: 10.3171/2012.1.FOCUS11356] [Medline: 22463116]
55. Angula N, Dlodlo N. A mobile application for health information dissemination: a Namibian context. In: Proceedings of the 2016 International Conference on Advances in Computing and Communication Engineering. 2016 Presented at: ICACCE '16; November 28-29, 2016; Durban, South Africa p. 461-466 URL: https://ieeexplore.ieee.org/document/8073792
56. Arnold K, Mugisha GA, Uzoka FN, Imanirakiza S, Muhumuza C, Bukenya JN. Development of an e-health system for improving health-care access in developing countries. In: Proceedings of the 2021 Conference on Future Technologies Conference. 2021 Presented at: FTC '21; October 28-29, 2021; Vancouver, BC p. 607-606 URL: https://link.springer.com/chapter/10.1007/978-3-030-89880-9_45 [doi: 10.1007/978-3-030-89880-9_45]
57. Astudillo C, Ankrom C, Trevino A, Malazarte RM, Bambhroliya AB, Savitz S, et al. Rationale and design of a statewide telestroke registry: lone star stroke consortium telestroke registry (LeSteR). BMJ Open 2019 Sep 04;9(9):e026496 [FREE Full text] [doi: 10.1136/bmjopen-2018-026496] [Medline: 31488463]
58. Balco P, Linhardt P, Drahosova M. The e-health solution in Slovak Republic: cloud concept for centralised national solution. In: Proceedings of the 2017 IEEE International Symposium on Signal Processing and Information Technology. 2017 Presented at: ISSPIT '17; December 18-20, 2017; Bilbao, Spain p. 434-438 URL: https://ieeexplore.ieee.org/abstract/document/8388682 [doi: 10.1109/isspit.2017.8388682]
59. Barnett ML, Yee Jr HF, Mehrotra A, Giboney P. Los Angeles safety-net program eConsult system was rapidly adopted and decreased wait times to see specialists. Health Aff (Millwood) 2017 Mar 01;36(3):492-499. [doi: 10.1377/hlthaff.2016.1283] [Medline: 28264951]
60. Bartnik SE, Copeland SP, Aicken AJ, Turner AW. Optometry-facilitated teleophthalmology: an audit of the first year in Western Australia. Clin Exp Optom 2018 Sep;101(5):700-703. [doi: 10.1111/cxo.12658] [Medline: 29444552]
61. Bastos de Carvalho A, Ware SL, Lei F, Bush HM, Sprang R, Higgins EB. Implementation and sustainment of a statewide telemedicine diabetic retinopathy screening network for federally designated safety-net clinics. PLoS One 2020;15(11):e0241767 [FREE Full text] [doi: 10.1371/journal.pone.0241767] [Medline: 33147290]
62. Becker ER, Shegog R, Savas LS, Frost EL, Coan SP, Healy CM, et al. Parents' experience with a mobile health intervention to influence human papillomavirus vaccination decision making: mixed methods study. JMIR Pediatr Parent 2022 Feb 21;5(1):e30340 [FREE Full text] [doi: 10.2196/30340] [Medline: 35188469]
63. Bladin CF, Kim J, Bagot KL, Vu M, Moloczij N, Denisenko S, et al. Improving acute stroke care in regional hospitals: clinical evaluation of the Victorian Stroke Telemedicine program. Med J Aust 2020 May;212(8):371-377. [doi: 10.5694/mja2.50570] [Medline: 32255520]
64. Bracale M, Cesarelli M, Pepino A, Bifulco P. Telemedicine-islands project: cost-effectiveness and cost-comparison analysis. In: Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2000 Presented at: EMBS '00; July 23-28, 2000; Chicago, IL p. 1364-1367 URL: https://ieeexplore.ieee.org/document/897992 [doi: 10.1109/iembs.2000.897992]
65. Brauchli K, O'Mahony D, Banach L, Oberholzer M. iPath - a telemedicine platform to support health providers in low resource settings. In: Marsh A, Bos L, Laxminarayan S, editors. Medical and Care Compunetics 2. Amsterdam, the Netherlands: Ios Press; 2005:2-7.
66. Bruno A, Lanning KM, Gross H, Hess DC, Nichols FT, Switzer JA. Timeliness of intravenous thrombolysis via telestroke in Georgia. Stroke 2013 Sep;44(9):2620-2622. [doi: 10.1161/STROKEAHA.113.001898] [Medline: 23839507]
67. Buendia O, Shankar S, Mahon H, Toal C, Menzies L, Ravichandran P, et al. Is it possible to implement a rare disease case-finding tool in primary care? A UK-based pilot study. Orphanet J Rare Dis 2022 Feb 16;17(1):54 [FREE Full text] [doi: 10.1186/s13023-022-02216-w] [Medline: 35172857]
68. Bunnell BE, Davidson TM, Dewey D, Price M, Ruggiero KJ. Rural and urban/suburban families' use of a web-based mental health intervention. Telemed J E Health 2017 May;23(5):390-396 [FREE Full text] [doi: 10.1089/tmj.2016.0153] [Medline: 27753542]
69. Cabrnoch M, Hasić B. Electronic health book—a unique Czech solution for eHealth. Health Technol 2011 Jul 20;1(2-4):57-69. [doi: 10.1007/s12553-011-0006-z]
70. Camara G, Diallo AH, Lo M, Tendeng JN, Lo S. A national medical information system for Senegal: architecture and services. In: Hoerbst A, Hackl WO, DeKeizer N, Prokosch HU, HercigonjaSzekeres M, DeLusignan S, editors. Exploring Complexity in Health: An Interdisciplinary Systems Approach. Amsterdam, the Netherlands: Ios Press; 2016:43-47.
71. Cheung ST, Davies RF, Smith K, Marsh R, Sherrard H, Keon WJ. The Ottawa telehealth project. Telemed J 1998 Jan;4(3):259-266. [doi: 10.1089/tmj.1.1998.4.259] [Medline: 9831750]
72. Conway NT, Wotayan RA, Alkuzam A, Al-Refaei FF, Badawi D, Barake R, et al. The Kuwait-Scotland eHealth Innovation Network (KSeHIN): a sustainable approach to quality improvement in healthcare. Qual Prim Care 2014;22(1):43-51. [Medline: 24589150]

XSL•FO
RenderX

73. Cunningham M, Cunningham PM. mHealth4Afrika pilot validation in healthcare facilities in Ethiopia, Kenya and Malawi. In: Proceedings of the 2019 IEEE Global Humanitarian Technology Conference. 2019 Presented at: GHTC '19; October 17-20, 2019; Seattle, WA p. 297-304 URL: https://ieeexplore.ieee.org/document/9033052 [doi: 10.1109/ghtc46095.2019.9033052]
74. Cunningham P, Cunningham M. mHealth4Afrika – challenges when co-designing a cross-border primary healthcare solution. In: Proceedings of the 2018 IEEE Conference on International Symposium on Technology and Society. 2018 Presented at: ISTAS '18; November 13-14, 2018; Washington, DC p. 32-36 URL: https://ieeexplore.ieee.org/document/8638279 [doi: 10.1109/istas.2018.8638279]
75. Cunningham P, Cunningham M. mHealth4Afrika - lessons learnt and good practices from validation of a comprehensive healthcare solution for resource constrained environments. In: Proceedings of the 2019 IEEE International Symposium on Technology and Society. 2019 Presented at: ISTAS '19; November 15-16, 2019; Medford, MA p. 1-8 URL: https://ieeexplore.ieee.org/document/8937991 [doi: 10.1109/istas48451.2019.8937991]
76. Cunningham PM, Cunningham M. mHealth4Afrika-Co-designing a standards based solution for use in resource constrained primary healthcare facilities. In: Proceedings of the 41st Annual International Conference of the Ieee Engineering in Medicine and Biology Society. 2019 Presented at: EMBC '19; July 23–27, 2019; Berlin, Germany p. 4289-4292 URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8856860 [doi: 10.1109/embc.2019.8856860]
77. Dash S, Aarthy R, Mohan V. Telemedicine during COVID-19 in India-a new policy and its challenges. J Public Health Policy 2021 Sep;42(3):501-509 [FREE Full text] [doi: 10.1057/s41271-021-00287-w] [Medline: 34012012]
78. de Bustos EM, Berthier E, Chavot D, Bouamra B, Moulin T. Evaluation of a French regional telemedicine network dedicated to neurological emergencies: a 14-year study. Telemed J E Health 2018 Feb;24(2):155-160. [doi: 10.1089/tmj.2017.0035] [Medline: 29346039]
79. de Melo MD, Nunes MV, Resende RF, Figueiredo RR, Ruas SS, Dos Santos AF, et al. Belo Horizonte telehealth: incorporation of teleconsultations in a health primary care system. Telemed J E Health 2018 Aug;24(8):631-638. [doi: 10.1089/tmj.2017.0165] [Medline: 29271706]
80. de Wit LM, van Uden-Kraan CF, Lissenberg-Witte BI, Melissant HC, Fleuren MA, Cuijpers P, et al. Adoption and implementation of a web-based self-management application "Oncokompas" in routine cancer care: a national pilot study. Support Care Cancer 2019 Aug 18;27(8):2911-2920 [FREE Full text] [doi: 10.1007/s00520-018-4591-5] [Medline: 30564933]
81. deRiel E, Puttkammer N, Hyppolite N, Diallo J, Wagner S, Honoré JG, et al. Success factors for implementing and sustaining a mature electronic medical record in a low-resource setting: a case study of iSanté in Haiti. Health Policy Plan 2018 Mar 01;33(2):237-246. [doi: 10.1093/heapol/czx171] [Medline: 29253138]
82. Dharmayat KI, Tran T, Hardy V, Chirambo BG, Thompson MJ, Ide N, et al. Sustainability of 'mHealth' interventions in sub- Saharan Africa: a stakeholder analysis of an electronic community case management project in Malawi. Malawi Med J 2019 Sep 03;31(3):177-183 [FREE Full text] [doi: 10.4314/mmj.v31i3.3] [Medline: 31839886]
83. Dittrich T. Die neue elektronische „Notfallpatientenakte“. Notf Rett 2020 Nov 13;24(3):239-241. [doi: 10.1007/s10049-020-00802-8]
84. Dixon BE, Tao GY, Wang JE, Tu WZ, Hoover S, Zhang ZY, et al. An integrated surveillance system to examine testing, services, and outcomes for sexually transmitted diseases. Stud Health Technol Inform 2017;245:361-365. [Medline: 29295116]
85. Donati M, Celli A, Ruiu A, Saponara S, Fanucci L. A telemedicine service system exploiting BT/BLE wireless sensors for remote management of chronic patients. Technologies 2019 Jan 18;7(1):13. [doi: 10.3390/technologies7010013]
86. Draganescu D, Lupuleasa D, Dumitrescu IB, Parvu CE, Ciolan DF. Evidence on e-prescribing systems worldwide. First Romanian results. Farmacia 2013 Apr;61(2):353-360 [FREE Full text]
87. Dubovitskaya A, Baig F, Xu ZG, Shukla R, Zambani PS, Swaminathan A, et al. ACTION-EHR: patient-centric blockchain-based electronic health record data management for cancer care. J Med Internet Res 2020 Aug 21;22(8):e13598 [FREE Full text] [doi: 10.2196/13598] [Medline: 32821064]
88. Dugonik A, Kalač Pandurovič M, Vok M, Dugonik B. Development and significance of the E-surveillance system for contact allergies in Slovenia. Acta Dermatovenerol Croat 2017 Jul;25(2):91-98. [Medline: 28871922]
89. Dyb K, Warth LL. The Norwegian national summary care record: a qualitative analysis of doctors' use of and trust in shared patient information. BMC Health Serv Res 2018 Apr 06;18(1):252 [FREE Full text] [doi: 10.1186/s12913-018-3069-y] [Medline: 29625587]
90. Estuar MR, Batangan D, Coronel A, Castro FE, Amarra AC, Caliso RA, et al. eHealth TABLET: a developing country perspective in managing the development and deployment of a mobile - cloud electronic medical record for local government units. In: Proceedings of the 2014 IEEE 15th International Conference on Mobile Data Management. 2014 Presented at: MDM '14; July 14-18, 2014; Brisbane, Australia p. 313-316 URL: https://ieeexplore.ieee.org/document/6916935 [doi: 10.1109/mdm.2014.44]
91. Fanta GB, Pretorius L, Erasmus L. Hospitals' readiness to implement sustainable SmartCare systems in Addis Ababa, Ethiopia. In: Proceedings of the 2019 Portland International Conference on Management of Engineering and Technology.

XSL•FO
RenderX

2019 Presented at: PICMET '19; August 25-29, 2019; Portland, OR p. 1-9 URL: https://ieeexplore.ieee.org/document/8893824 [doi: 10.23919/picmet.2019.8893824]
92. Finkelstein J, Joshi A, Arora M. Home automated telemanagement in hypertension. In: Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems. 2004 Presented at: CBMS '04; June 25, 2004; Bethesda, MD p. 452-459 URL: https://ieeexplore.ieee.org/document/1311756
93. Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res 2009 Sep 04;9:157 [FREE Full text] [doi: 10.1186/1472-6963-9-157] [Medline: 19732426]
94. France FR. eHealth in Belgium, a new "secure" federal network: role of patients, health professions and social security services. Int J Med Inform 2011 Feb;80(2):e12-e16. [doi: 10.1016/j.ijmedinf.2010.10.005] [Medline: 21035383]
95. Galli R, Keith JC, McKenzie K, Hall GS, Henderson K. TelEmergency: a novel system for delivering emergency care to rural hospitals. Ann Emerg Med 2008 Mar;51(3):275-284. [doi: 10.1016/j.annemergmed.2007.04.025] [Medline: 17764784]
96. Ganesan D. Human rights implications of the digital revolution in health care in India. Health Hum Rights 2022 Jun;24(1):5-19 [FREE Full text] [Medline: 35747286]
97. Gavrilov G, Vlahu-Gjorgievska E, Trajkovik V. Analysis of introducing e-services: a case study of health insurance fund of Macedonia. J Health Organ Manag 2016;30(3):354-371. [doi: 10.1108/jhom-10-2014-0171]
98. Gilbert SS, Bulula N, Yohana E, Thompson J, Beylerian E, Werner L, et al. The impact of an integrated electronic immunization registry and logistics management information system (EIR-eLMIS) on vaccine availability in three regions in Tanzania: a pre-post and time-series analysis. Vaccine 2020 Jan 16;38(3):562-569 [FREE Full text] [doi: 10.1016/j.vaccine.2019.10.059] [Medline: 31706808]
99. Gios L, Crema Falceri G, Micocci S, Patil L, Testa S, Sforzin S, et al. Use of eHealth platforms and apps to support monitoring and management of home-quarantined patients with COVID-19 in the province of Trento, Italy: app development and implementation. JMIR Form Res 2021 May 31;5(5):e25713 [FREE Full text] [doi: 10.2196/25713] [Medline: 33909586]
100. Gong Y, Zhou JY. What are the demands of telegeriatrics medical services for elderly patients during the COVID-19 pandemic. Front Public Health 2022 Aug 8;10:935684 [FREE Full text] [doi: 10.3389/fpubh.2022.935684] [Medline: 36003633]
101. Graham ML, Uesugi KH, Niederdeppe J, Gay GK, Olson CM. The theory, development, and implementation of an e-intervention to prevent excessive gestational weight gain: e-Moms Roc. Telemed J E Health 2014 Dec;20(12):1135-1142 [FREE Full text] [doi: 10.1089/tmj.2013.0354] [Medline: 25354350]
102. Grisot M, Vassilakopoulou P, Aanestad M. The Norwegian eHealth platform: development through cultivation strategiesincremental changes. In: Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P, editors. Information Infrastructures within European Health Care: Working with the Installed Base. Cham, Switzerland: Springer; 2017:193-208.
103. Guerra GM, Wen CL, Motta RA, Vieira MM, Fistarol IR, de Oliveira JC, et al. Developmentimplementation of the e-care portal health education for hypertension. In: Proceedings of the 11th International Technology, Education and Development Conference. 2017 Presented at: INTED '17; Valencia, Spain; March 6-8, 2017 p. 7392-7403 URL: https://observatorio.fm.usp.br/entities/publication/51369631-b540-4ff4-988c-d975bb4d2e2e [doi: 10.21125/inted.2017.1712]
104. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. J Med Internet Res 2005 Mar 14;7(1):e3 [FREE Full text] [doi: 10.2196/jmir.7.1.e3] [Medline: 15829475]
105. Hägglund M, Scandurra I. Patients' online access to electronic health records: current status and experiences from the implementation in Sweden. Stud Health Technol Inform 2017;245:723-727. [Medline: 29295193]
106. Hamaya R, Fukuda H, Takebayashi M, Mori M, Matsushima R, Nakano K, et al. Effects of an mHealth app (Kencom) with integrated functions for healthy lifestyles on physical activity levels and cardiovascular risk biomarkers: observational study of 12,602 users. J Med Internet Res 2021 Apr 26;23(4):e21622 [FREE Full text] [doi: 10.2196/21622] [Medline: 33900203]
107. Hardardottir GA, Thoroddsen A. National eHealth Implementation: Country Experience. Amsterdam, the Netherlands: Ios Press; 2016.
108. Harno K, Kauppinen-Mäkelin R, Syrjäläinen J. Managing diabetes care using an integrated regional e-health approach. J Telemed Telecare 2006 Dec 02;12 Suppl 1(1_suppl):13-15. [doi: 10.1258/135763306777978380] [Medline: 16884565]
109. Hartvigsen G, Johansen MA, Hasvold P, Bellika JG, Arsand E, Arild E, et al. Challenges in telemedicine and eHealth: lessons learned from 20 years with telemedicine in Tromsø. Stud Health Technol Inform 2007;129(Pt 1):82-86. [Medline: 17911683]
110. Harzheim E, Gonçalves MR, Umpierre RN, da Silva Siqueira AC, Katz N, Agostinho MR, et al. Telehealth in Rio Grande do Sul, Brazil: bridging the gaps. Telemed J E Health 2016 Nov;22(11):938-944. [doi: 10.1089/tmj.2015.0210] [Medline: 27096384]
111. Hebert E, Ferguson W, McCullough S, Chan M, Drobakha A, Ritter S, et al. mBody health: digitizing disabilities in Sierra Leone. In: Proceedings of the 2016 IEEE Global Humanitarian Technology Conference. 2016 Presented at: GHTC '16; October 13-16, 2016; Seattle, WA p. 717-723 URL: https://ieeexplore.ieee.org/document/7857357 [doi: 10.1109/ghtc.2016.7857357]

XSL•FO
RenderX

112. Heider AR, Maloney NA, Satchidanand N, Allen GM, Mueller R, Gangloff S, et al. Developing a communitywide electronic health record disease registry in primary care practices: lessons learned from the Western new york beacon community. EGEMS (Wash DC) 2014;2(3):1089 [FREE Full text] [doi: 10.13063/2327-9214.1089] [Medline: 25848616]
113. Herbek S, Eisl HA, Hurch M, Schator A, Sabutsch S, Rauchegger G, et al. The electronic health record in Austria: a strong network between health care and patients. Eur Surg 2012 Jun;44(3):155-163. [doi: 10.1007/s10353-012-0092-9]
114. Hernández-Ávila JE, Palacio-Mejía LS, Lara-Esqueda A, Silvestre E, Agudelo-Botero M, Diana ML, et al. Assessing the process of designing and implementing electronic health records in a statewide public health system: the case of Colima, Mexico. J Am Med Inform Assoc 2013;20(2):238-244 [FREE Full text] [doi: 10.1136/amiajnl-2012-000907] [Medline: 23019239]
115. Herrington G, Zardins Y, Hamilton A. A pilot trial of emergency telemedicine in regional Western Australia. J Telemed Telecare 2013 Oct;19(7):430-433. [doi: 10.1177/1357633X13506531] [Medline: 24218359]
116. Hofer F, Schreyögg J, Stargardt T. Effectiveness of a home telemonitoring program for patients with chronic obstructive pulmonary disease in Germany: evidence from the first three years. PLoS One 2022;17(5):e0267952 [FREE Full text] [doi: 10.1371/journal.pone.0267952] [Medline: 35551546]
117. Hostenkamp G. The introduction of the electronic medication card: what Germany can learn from Denmark. Gesundh ökon Qual manag 2016 Aug 23;22(01):54-60. [doi: 10.1055/s-0042-113151]
118. Hoth AB, Shafer C, Dillon DB, Mayer R, Walton G, Ohl ME. Iowa TelePrEP: a public-health-partnered telehealth model for human immunodeficiency virus preexposure prophylaxis delivery in a rural state. Sex Transm Dis 2019 Aug;46(8):507-512. [doi: 10.1097/OLQ.0000000000001017] [Medline: 31295217]
119. Hsu M, Chu T, Yen J, Chiu W, Yeh G, Chen T, et al. Development and implementation of a national telehealth project for long-term care: a preliminary study. Comput Methods Programs Biomed 2010 Mar;97(3):286-292. [doi: 10.1016/j.cmpb.2009.12.008] [Medline: 20092907]
120. Idachaba FE, Idachaba EM. Robust e-Health communication architecture for rural communities in developing countries. Eng Technol Appl Sci Res 2012 Jun 04;2(3):237-240. [doi: 10.48084/etasr.45]
121. Ionescu B, Gadea C, Solomon B, Ionescu D, Stoicu-Tivadar V, Trifan M. A cloud based real-time collaborative platform for eHealth. In: Cornet R, StoicuTivadar L, Horbst A, Calderon CL, Andersen SK, HercigonjaSzekeres M, editors. Digital Healthcare Empowering Europeans. Amsterdam, the Netherlands: Ios Press; 2015:919-923.
122. Janzek-Hawlat S, Ammenwerth E, Dorda W, Duftschmid G, Hackl W, Hörbst A, et al. The Austrian e-Medikation pilot evaluation: lessons learned from a national medication list. Stud Health Technol Inform 2013;192:347-351. [Medline: 23920574]
123. Jensen TB, Thorseng AA. Building national healthcare infrastructure: the case of the Danish e-Health portal. In: Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P, editors. Information Infrastructures within European Health Care: Working with the Installed Base. Cham, Switzerland: Springer; 2017:209-224.
124. Jonsson F, Carson DB, Goicolea I, Hurtig AK. Strengthening community health systems through novel eHealth initiatives? Commencing a realist study of the virtual health rooms in rural Northern Sweden. Int J Health Policy Manag 2022 Jan 01;11(1):39-48 [FREE Full text] [doi: 10.34172/ijhpm.2021.08] [Medline: 33619935]
125. Kassari P, Papaioannou P, Billiris A, Karanikas H, Eleftheriou S, Thireos E, et al. Electronic registry for the management of childhood obesity in Greece. Eur J Clin Invest 2018 Mar 11;48(3). [doi: 10.1111/eci.12887] [Medline: 29327339]
126. Khader Y, Alyahya M, El-Khatib Z, Batieha A, Al-Sheyab N, Shattnawi K. The Jordan Stillbirth and Neonatal Mortality Surveillance (JSANDS) system: evaluation study. J Med Internet Res 2021 Jul 21;23(7):e29143 [FREE Full text] [doi: 10.2196/29143] [Medline: 34287214]
127. Fidel Kinori SG, Carot-Sans G, Cuartero A, Valero-Bover D, Roma Monfa R, Garcia E, et al. A web-based app for emotional management during the COVID-19 pandemic: platform development and retrospective analysis of its use throughout two waves of the outbreak in Spain. JMIR Form Res 2022 Mar 31;6(3):e27402 [FREE Full text] [doi: 10.2196/27402] [Medline: 35142638]
128. Klingwort J, Schnell R. Critical limitations of digital epidemiology: why COVID-19 apps are useless. Surv Res Methods 2020;14(2):95-100. [doi: 10.18148/srm/2020.v14i2.7726]
129. K??hler F, Nettlau H, Schweizer T, Waller T, Anker SD. The research project of the German Federal Ministry of Economics and Technology: 'partnership for the heart' - a new approach in telemedicine. Dis Manag Health Out 2006;14(Suppl 1):37-41. [doi: 10.2165/00115677-200614001-00010]
130. Kose I, Akpinar N, Gurel M, Arslan Y, Ozer H, Yurt N, et al. Turkey's national health information system (NHIS). In: Cunningham P, Cunningham M, editors. Collaboration and the Knowledge Economy: Issues, Applications, Case Studies. Amsterdam, the Netherlands: Ios Press; 2008:1-7.
131. Kouroubali A, Kondylakis H, Kavlentakis G, Logothetides F, Stathiakis N, Petrakis Y, et al. An eHealth platform for the holistic management of COVID-19. Stud Health Technol Inform 2020 Sep 04;273:182-188. [doi: 10.3233/SHTI200636] [Medline: 33087610]
132. Lai L, Sato R, He SH, Ouchi K, Leiter R, de Lima Thomas J, et al. Usage patterns of a web-based palliative care content platform (PalliCOVID) during the COVID-19 pandemic. J Pain Symptom Manage 2020 Oct;60(4):e20-e27 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.07.016] [Medline: 32730951]

XSL•FO
RenderX

133. Lämsä E, Timonen J, Ahonen R. Pharmacy customers' experiences with electronic prescriptions: cross-sectional survey on nationwide implementation in Finland. J Med Internet Res 2018 Feb 23;20(2):e68 [FREE Full text] [doi: 10.2196/jmir.9367] [Medline: 29475826]
134. Latifi R, Dasho E, Merrell RC, Lopes M, Azevedo V, Bekteshi F, et al. Cabo Verde telemedicine program: initial results of nationwide implementation. Telemed J E Health 2014 Nov;20(11):1027-1034. [doi: 10.1089/tmj.2014.0026] [Medline: 25083737]
135. Latifi R, Dasho E, Shatri Z, Tilley E, Osmani KL, Doarn CR, et al. Telemedicine as an innovative model for rebuilding medical systems in developing countries through multipartnership collaboration: the case of Albania. Telemed J E Health 2015 Jun;21(6):503-509. [doi: 10.1089/tmj.2014.0138] [Medline: 25347524]
136. Lazaridis C, DeSantis SM, Jauch EC, Adams RJ. Telestroke in South Carolina. J Stroke Cerebrovasc Dis 2013 Oct;22(7):946-950. [doi: 10.1016/j.jstrokecerebrovasdis.2011.11.008] [Medline: 22196873]
137. Le Pape MA, Suárez JC, Mhayi A, Haazen D, Özaltin E. Developing an HMIS architecture framework to support a national health care eHealth strategy reform: a case study from Morocco. Health Syst Reform 2017 Jan 02;3(1):56-67. [doi: 10.1080/23288604.2017.1265041] [Medline: 31514712]
138. Lee MS, Ray KN, Mehrotra A, Giboney P, Yee Jr HF, Barnett ML. Primary care practitioners' perceptions of electronic consult systems: a qualitative analysis. JAMA Intern Med 2018 Jun 01;178(6):782-789 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0738] [Medline: 29801079]
139. Lehmann NJ, Karagülle MU, Spielmann F, George B, Zick B, Heuer J. mHealthAtlas - an expert-based multi-sided platform for the evaluation of mHealth applications. In: Proceedings of the 2021 IEEE 9th International Conference on Healthcare Informatics. 2021 Presented at: ICHI '21; August 9-12, 2021; Victoria, BC p. 449-450 URL: https://ieeexplore.ieee.org/document/9565707 [doi: 10.1109/ichi52183.2021.00079]
140. Lelu iu-Weinberger C, Manu M, Ionescu F, Dogaru B, Kovacs T, Dorobăn escu C, et al. An mHealth intervention to improve young gay and bisexual men's sexual, behavioral, and mental health in a structurally stigmatizing national context. JMIR Mhealth Uhealth 2018 Nov 14;6(11):e183 [FREE Full text] [doi: 10.2196/mhealth.9283] [Medline: 30429117]
141. Lewiecki EM, Boyle JF, Arora S, Bouchonville 2nd MF, Chafey DH. Telementoring: a novel approach to reducing the osteoporosis treatment gap. Osteoporos Int 2017 Jan;28(1):407-411. [doi: 10.1007/s00198-016-3708-2] [Medline: 27439373]
142. Li SE, Hossain M, Gilman B, Forrow LV, Lee KM, Brown R. Effects of a nursing home telehealth program on spending and utilization for Medicare residents. Health Serv Res 2022 Oct;57(5):1191-1200 [FREE Full text] [doi: 10.1111/1475-6773.13936] [Medline: 35043397]
143. Mahdi S, Michalik-Denny EK, Buckland NJ. An assessment of behavior change techniques in two versions of a dietary mobile application: the change4life food scanner. Front Public Health 2022 Feb 23;10:803152 [FREE Full text] [doi: 10.3389/fpubh.2022.803152] [Medline: 35284376]
144. Makai P, Perry M, Robben SH, Schers H, Heinen M, Olde Rikkert MG, et al. Which frail older patients use online health communities and why? A mixed methods process evaluation of use of the Health and Welfare portal. J Med Internet Res 2014 Dec 17;16(12):e278 [FREE Full text] [doi: 10.2196/jmir.3609] [Medline: 25519769]
145. Mammen JR, Schoonmaker JD, Java J, Halterman J, Berliant MN, Crowley A, et al. Going mobile with primary care: smartphone-telemedicine for asthma management in young urban adults (TEAMS). J Asthma 2022 Jan;59(1):132-144. [doi: 10.1080/02770903.2020.1830413] [Medline: 33064038]
146. Markwick L, McConnochie K, Wood N. Expanding telemedicine to include primary care for the urban adult. J Health Care Poor Underserved 2015 Aug;26(3):771-776. [doi: 10.1353/hpu.2015.0078] [Medline: 26320911]
147. Mash R, Schouw D, Fischer AE. Evaluating the implementation of the GREAT4Diabetes WhatsApp chatbot to educate people with type 2 diabetes during the COVID-19 pandemic: convergent mixed methods study. JMIR Diabetes 2022 Jun 24;7(2):e37882 [FREE Full text] [doi: 10.2196/37882] [Medline: 35537057]
148. Mathew D. Information technology and public health management of disasters--a model for South Asian countries. Prehosp Disaster Med 2005 Jun 28;20(1):54-60. [doi: 10.1017/s1049023x00002156] [Medline: 15748016]
149. Mochari-Greenberger H, Peters A, Vue L, Pande RL. A nationally scaled telebehavioral health program for chronic pain: characteristics, goals, and psychological outcomes. Telemed J E Health 2017 Aug;23(8):640-648. [doi: 10.1089/tmj.2016.0188] [Medline: 28157442]
150. Mochari-Greenberger H, Vue L, Luka A, Peters A, Pande RL. A tele-behavioral health intervention to reduce depression, anxiety, and stress and improve diabetes self-management. Telemed J E Health 2016 Aug;22(8):624-630. [doi: 10.1089/tmj.2015.0231] [Medline: 26954880]
151. Modi D, Dholakia N, Gopalan R, Venkatraman S, Dave K, Shah S, et al. mHealth intervention "ImTeCHO" to improve delivery of maternal, neonatal, and child care services-a cluster-randomized trial in tribal areas of Gujarat, India. PLoS Med 2019 Oct 24;16(10):e1002939 [FREE Full text] [doi: 10.1371/journal.pmed.1002939] [Medline: 31647821]
152. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 01;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]

XSL•FO
RenderX

153. Monamele CG, Messanga Essengue LL, Ripa Njankouo M, Munshili Njifon HL, Tchatchueng J, Tejiokem MC, et al. Evaluation of a mobile health approach to improve the early warning system of influenza surveillance in Cameroon. Influenza Other Respir Viruses 2020 Sep 14;14(5):491-498 [FREE Full text] [doi: 10.1111/irv.12747] [Medline: 32410384]
154. Müller-Wieland D, Ickrath M. The electronic health record for diabetes (eDA) of the German diabetes association. Diabetologe 2021 Mar 17;17(3):260-264. [doi: 10.1007/S11428-021-00746-8]
155. Nalleballe K, Sharma R, Brown A, Joiner R, Kapoor N, Morgan T, et al. Ideal telestroke time targets: telestroke-based treatment times in the United States stroke belt. J Telemed Telecare 2020 Apr 23;26(3):174-179. [doi: 10.1177/1357633X18805661] [Medline: 30352525]
156. Nascimento BR, Brant LC, Castro AC, Froes LE, Ribeiro AL, Cruz LV, et al. Impact of a large-scale telemedicine network on emergency visits and hospital admissions during the coronavirus disease 2019 pandemic in Brazil: data from the UNIMED-BH system. J Telemed Telecare 2023 Feb;29(2):103-110 [FREE Full text] [doi: 10.1177/1357633X20969529] [Medline: 33100183]
157. Ngoma M, Mushi B, Morse RS, Ngoma T, Mahuna H, Lambden K, et al. mPalliative care link: examination of a mobile solution to palliative care coordination among Tanzanian patients with cancer. JCO Glob Oncol 2021 Aug;7:1306-1315 [FREE Full text] [doi: 10.1200/GO.21.00122] [Medline: 34406856]
158. Nurmansyah MI, Rosidati C, Yustiyani Y, Nasir NM. Measuring the success of PeduliLindungi application use for supporting COVID-19 prevention: a case study among college students in Jakarta, Indonesia. Kesmas Natl Public Health J 2022 Aug 19;17(sp1):1. [doi: 10.21109/kesmas.v17isp1.6057]
159. Oliveira TC, Bayer S, Gonçalves L, Barlow J. Telemedicine in Alentejo. Telemed J E Health 2014 Jan;20(1):90-93 [FREE Full text] [doi: 10.1089/tmj.2012.0308] [Medline: 24180419]
160. Otu AA, Effa EE, Onwusaka O, Omoyele C, Arakelyan S, Okuzu O, et al. mHealth guideline training for non-communicable diseases in primary care facilities in Nigeria: a mixed methods pilot study. BMJ Open 2022 Aug 26;12(8):e060304 [FREE Full text] [doi: 10.1136/bmjopen-2021-060304] [Medline: 36028271]
161. Pal Bhowmick I, Chutia D, Chouhan A, Nishant N, Raju PL, Narain K, et al. Validation of a mobile health technology platform (FeverTracker) for malaria surveillance in India: development and usability study. JMIR Form Res 2021 Nov 10;5(11):e28951 [FREE Full text] [doi: 10.2196/28951] [Medline: 34757321]
162. Pearce C, Bainbridge M. A personally controlled electronic health record for Australia. J Am Med Inform Assoc 2014;21(4):707-713 [FREE Full text] [doi: 10.1136/amiajnl-2013-002068] [Medline: 24650635]
163. Perrotta D, Bella A, Rizzo C, Paolotti D. Participatory online surveillance as a supplementary tool to sentinel doctors for influenza-like illness surveillance in Italy. PLoS One 2017;12(1):e0169801 [FREE Full text] [doi: 10.1371/journal.pone.0169801] [Medline: 28076411]
164. Petersen MJ, Baune DA, Huggins IG, LaMarche DJ. Telemedicine in Utah: the rural Utah telemedicine pilot project. In: Proceedings of Frontiers in Education 26th Annual Conference on Technology-Based Re-Engineering Engineering Education. 1996 Presented at: FIE '96; November 6-9, 1996; Salt Lake City, UT p. 986-989 URL: https://ieeexplore.ieee.org/document/573152 [doi: 10.1109/fie.1996.573152]
165. Pollack TM, Nhung VT, Vinh DT, Hao DT, Trang LT, Duc PA, et al. Building HIV healthcare worker capacity through telehealth in Vietnam. BMJ Glob Health 2020;5(4):e002166 [FREE Full text] [doi: 10.1136/bmjgh-2019-002166] [Medline: 32337087]
166. Power K, McCrea Z, White M, Breen A, Dunleavy B, O'Donoghue S, et al. The development of an epilepsy electronic patient portal: facilitating both patient empowerment and remote clinician-patient interaction in a post-COVID-19 world. Epilepsia 2020 Sep 10;61(9):1894-1905 [FREE Full text] [doi: 10.1111/epi.16627] [Medline: 32668026]
167. Prinja S, Gupta A, Bahuguna P, Nimesh R. Cost analysis of implementing mHealth intervention for maternal, newborn and child health care through community health workers: assessment of ReMIND program in Uttar Pradesh, India. BMC Pregnancy Childbirth 2018 Oct 03;18(1):390 [FREE Full text] [doi: 10.1186/s12884-018-2019-3] [Medline: 30285669]
168. Prinja S, Nimesh R, Gupta A, Bahuguna P, Gupta M, Thakur JS. Impact of m-health application used by community health volunteers on improving utilisation of maternal, new-born and child health care services in a rural area of Uttar Pradesh, India. Trop Med Int Health 2017 Jul 13;22(7):895-907 [FREE Full text] [doi: 10.1111/tmi.12895] [Medline: 28510997]
169. Ray G, Altschuler A, Karmali R, Binswanger I, Glanz JM, Clarke CL, et al. Development and implementation of a prescription opioid registry across diverse health systems. JAMIA Open 2022 Jul;5(2):ooac030 [FREE Full text] [doi: 10.1093/jamiaopen/ooac030] [Medline: 35651523]
170. Rosewell A, Makita L, Muscatello D, John LN, Bieb S, Hutton R, et al. Health information system strengthening and malaria elimination in Papua New Guinea. Malar J 2017 Jul 05;16(1):278 [FREE Full text] [doi: 10.1186/s12936-017-1910-0] [Medline: 28679421]
171. Rosewell A, Ropa B, Randall H, Dagina R, Hurim S, Bieb S, et al. Mobile phone-based syndromic surveillance system, Papua New Guinea. Emerg Infect Dis 2013 Nov;19(11):1811-1818 [FREE Full text] [doi: 10.3201/eid1911.121843] [Medline: 24188144]
172. Ruas SS, Assunção A. Teleconsultations by primary care physicians of Belo Horizonte: challenges in the diffusion of innovation. Telemed J E Health 2013 May;19(5):409-414. [doi: 10.1089/tmj.2012.0095] [Medline: 23600411]

XSL•FO
RenderX

173. Sá-Sousa A, Fonseca JA, Pereira AM, Ferreira A, Arrobas A, Mendes A, et al. The Portuguese severe asthma registry: development, features, and data sharing policies. Biomed Res Int 2018;2018:1495039 [FREE Full text] [doi: 10.1155/2018/1495039] [Medline: 30584531]
174. Saeed SA. Tower of babel problem in telehealth: addressing the health information exchange needs of the North Carolina statewide telepsychiatry program (NC-STeP). Psychiatr Q 2018 Jun 14;89(2):489-495. [doi: 10.1007/s11126-017-9551-6] [Medline: 29238901]
175. Sajwani A, Qureshi K, Shaikh T, Sayani S. eHealth for remote regions: findings from Central Asia health systems strengthening project. Stud Health Technol Inform 2015;209:128-134. [Medline: 25980715]
176. Salzsieder E, Augstein P. The Karlsburg Diabetes Management System: translation from research to eHealth application. J Diabetes Sci Technol 2011 Jan 01;5(1):13-22 [FREE Full text] [doi: 10.1177/193229681100500103] [Medline: 21303620]
177. Schaller M, Dornauer V, Hackl WO, Lechleitner G, Überegger M, Ammenwerth E. Implementing national electronic health records in nursing homes in Tyrol: a nursing management perspective. Stud Health Technol Inform 2020 Jun 23;271:240-247. [doi: 10.3233/SHTI200102] [Medline: 32578569]
178. Segal E, Alwan L, Pitney C, Taketa C, Indorf A, Held L, et al. Establishing clinical pharmacist telehealth services during the COVID-19 pandemic. Am J Health Syst Pharm 2020 Aug 20;77(17):1403-1408 [FREE Full text] [doi: 10.1093/ajhp/zxaa184] [Medline: 34279578]
179. Seiwerth S, Danilovic Z. The telepathology and teleradiology network in Croatia. Anal Cell Pathol 2000;21(3-4):223-228 [FREE Full text] [doi: 10.1155/2000/602182] [Medline: 11339571]
180. Sellberg N, Eltes J. The Swedish patient portal and its relation to the national reference architecture and the overall eHealth infrastructure. In: Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P, editors. Information Infrastructures within European Health Care: Working with the Installed Base. Cham, Switzerland: Springer; 2017:225-244.
181. Sepper R, Ross P, Tiik M. Nationwide health data management system: a novel approach for integrating biomarker measurements with comprehensive health records in large populations studies. J Proteome Res 2011 Jan 07;10(1):97-100. [doi: 10.1021/pr1007784] [Medline: 21080730]
182. Séroussi B, Bouaud J. Use of a nationwide personally controlled electronic health record by healthcare professionals and patients: a case study with the French DMP. In: Cornet R, Randell R, McCowan C, Peek N, Scott PJ, editors. Informatics for Health: Connected Citizen-Led Wellness and Population Health. Amsterdam, the Netherlands: Ios Press; 2017:333-337.
183. Séroussi B, Bouaud J. The (Re)-relaunching of the DMP, the French shared medical record: new features to improve uptake and use. In: Ugon A, Karlsson D, Klein GO, Moen A, editors. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created Ehealth. Amsterdam, the Netherlands: Ios Press; 2018:256-260.
184. Shaikh A, AlReshan MS, Asiri Y, Sulaiman A, Alshahrani H. Tele-COVID: a telemedicine SOA-based architectural design for COVID-19 patients. CMC Comput Mater Con 2021;67(1):549-576. [doi: 10.32604/cmc.2021.014813]
185. Shilpa DM, Naik PR, Shewade HD, Sudarshan H. Assessing the implementation of a mobile App-based electronic health record: a mixed-method study from South India. J Educ Health Promot 2020;9:102 [FREE Full text] [doi: 10.4103/jehp.jehp_749_19] [Medline: 32509910]
186. Shinde S, Shinde R, Shanbhag S, Solanki M, Sable P, Kimbahune S. mHEALTH-PHC - application design for rural health care. In: Proceedings of the 2014 IEEE Canada International Humanitarian Technology Conference. 2014 Presented at: IHTC '14; June 1-4, 2014; Montreal, QC p. 1-5 URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7147514 [doi: 10.1109/ihtc.2014.7147514]
187. Siddiqui AA, Raza-us-Samad M, Siddiqui FA, Shaikh ZA. Community health information system focusing Pakistan. In: Proceedings of the 5th International Conference on Information and Communication Technologies. 2013 Presented at: ICICT '13; December 14-15, 2013; Karachi, Pakistan p. 1-9 URL: https://ieeexplore.ieee.org/document/6732779 [doi: 10.1109/icict.2013.6732779]
188. Silva BS, de Azevedo Guimarães EA, de Oliveira VC, Cavalcante RB, Pinheiro MM, Gontijo TL, et al. National immunization program information system: implementation context assessment. BMC Health Serv Res 2020 Apr 21;20(1):333 [FREE Full text] [doi: 10.1186/s12913-020-05175-9] [Medline: 32316947]
189. Siminerio L, Ruppert K, Huber K, Toledo FG. Telemedicine for reach, education, access, and treatment (TREAT): linking telemedicine with diabetes self-management education to improve care in rural communities. Diabetes Educ 2014;40(6):797-805. [doi: 10.1177/0145721714551993] [Medline: 25253624]
190. Simonyan D, Gagnon MP, Duchesne T, Roos-Weil A. Effects of a telehealth programme using mobile data transmission on primary healthcare utilisation among children in Bamako, Mali. J Telemed Telecare 2013 Sep;19(6):302-306. [doi: 10.1177/1357633X13503429] [Medline: 24163292]
191. Singh Sodha T, Grønbæk A, Bhandari A, Mary B, Sudke A, Smith LT. mHealth learning tool for skilled birth attendants: scaling the Safe Delivery App in India. BMJ Open Qual 2022 Aug;11(Suppl 1):e001928 [FREE Full text] [doi: 10.1136/bmjoq-2022-001928] [Medline: 35977730]
192. Singh Y, Jackson D, Bhardwaj S, Titus N, Goga A. National surveillance using mobile systems for health monitoring: complexity, functionality and feasibility. BMC Infect Dis 2019 Sep 16;19(Suppl 1):786 [FREE Full text] [doi: 10.1186/s12879-019-4338-z] [Medline: 31526387]

XSL•FO
RenderX

193. Skorning M, Bergrath S, Rörtgen D, Brokmann JC, Beckers SK, Protogerakis M, et al. [E-health in emergency medicine - the research project Med-on-@ix]. Anaesthesist 2009 Mar;58(3):285-292. [doi: 10.1007/s00101-008-1502-z] [Medline: 19221700]
194. Smith GE, Elliot AJ, Lake I, Edeghere O, Morbey R, Catchpole M, Public Health England Real-time Syndromic Surveillance Team. Syndromic surveillance: two decades experience of sustainable systems - its people not just data!. Epidemiol Infect 2019 Jan 22;147:e101 [FREE Full text] [doi: 10.1017/S0950268819000074] [Medline: 30869042]
195. Soler D, Baldacchino D, Amato-Gauci A. Pilot of a child health electronic surveillance system (CHESS) in Malta. J Public Health 2007 May 01;15(3):199-209. [doi: 10.1007/S10389-007-0096-2]
196. Spielberg F, Levy V, Lensing S, Chattopadhyay I, Venkatasubramanian L, Acevedo N, et al. Fully integrated e-services for prevention, diagnosis, and treatment of sexually transmitted infections: results of a 4-county study in California. Am J Public Health 2014 Dec;104(12):2313-2320. [doi: 10.2105/ajph.2014.302302]
197. Stanimirovic D, Savic D. Two Years of ePrescription in Slovenia - applications and potentials. Stud Health Technol Inform 2018;247:261-265. [Medline: 29677963]
198. Stocks J, Choi Y, Ibrahim S, Huchko M. Iterative development of a mobile phone app to support community health volunteers during cervical cancer screening in Western Kenya: qualitative study. JMIR Form Res 2022 Feb 24;6(2):e27501 [FREE Full text] [doi: 10.2196/27501] [Medline: 35200151]
199. Ströher A, Honekamp W. [ELGA--the electronic health record in the light of data protection and data security]. Wien Med Wochenschr 2011 Jul;161(13-14):341-346. [doi: 10.1007/s10354-011-0011-x] [Medline: 21858632]
200. Sudhamony S, Nandakumar K, Binu PJ, Issac Niwas SI. Telemedicine and tele-health services for cancer-care delivery in India. IET Commun 2008;2(2):231. [doi: 10.1049/iet-com:20060701]
201. Syred J, Holdsworth G, Howroyd C, Spelman K, Baraitser P. Choose to test: self-selected testing for sexually transmitted infections within an online service. Sex Transm Infect 2019 May;95(3):171-174 [FREE Full text] [doi: 10.1136/sextrans-2018-053796] [Medline: 30842345]
202. Talboom-Kamp EP, Holstege MS, Chavannes NH, Kasteleyn MJ. Effects of use of an eHealth platform e-Vita for COPD patients on disease specific quality of life domains. Respir Res 2019 Jul 10;20(1):146 [FREE Full text] [doi: 10.1186/s12931-019-1110-2] [Medline: 31291945]
203. Thompson S, Mercer MA, Hofstee M, Stover B, Vasconcelos P, Meyanathan S. Connecting mothers to care: effectiveness and scale-up of an mHealth program in Timor-Leste. J Glob Health 2019 Dec;9(2):020428 [FREE Full text] [doi: 10.7189/jogh.09.020428] [Medline: 31673341]
204. Titov N, Dear BF, Nielssen O, Wootton B, Kayrouz R, Karin E, et al. User characteristics and outcomes from a national digital mental health service: an observational study of registrants of the Australian MindSpot Clinic. Lancet Digit Health 2020 Nov;2(11):e582-e593 [FREE Full text] [doi: 10.1016/S2589-7500(20)30224-7] [Medline: 33103097]
205. Tsai WH, Wu YS, Cheng CS, Kuo MH, Chang YT, Hu FK, et al. A technology acceptance model for deploying masks to combat the COVID-19 pandemic in Taiwan (My Health Bank): web-based cross-sectional survey study. J Med Internet Res 2021 Apr 21;23(4):e27069 [FREE Full text] [doi: 10.2196/27069] [Medline: 33819168]
206. van der Heijden JP, de Keizer NF, Witkamp L, Spuls PI. Evaluation of a tertiary teledermatology service between peripheral and academic dermatologists in the Netherlands. Telemed J E Health 2014 Apr;20(4):332-337. [doi: 10.1089/tmj.2013.0197] [Medline: 24684477]
207. Velinov G, Jakimovski B, Lesovski D, Ivanova Panova D, Frtunik D, Kon-Popovska M. EHR system MojTermin: implementation and initial data analysis. Stud Health Technol Inform 2015;210:872-876. [Medline: 25991280]
208. Vilardaga R, Rizo J, Ries RK, Kientz JA, Ziedonis DM, Hernandez K, et al. Formative, multimethod case studies of learn to quit, an acceptance and commitment therapy smoking cessation app designed for people with serious mental illness. Transl Behav Med 2019 Nov 25;9(6):1076-1086 [FREE Full text] [doi: 10.1093/tbm/iby097] [Medline: 30445507]
209. Vinekar A, Jayadev C, Mangalesh S, Shetty B, Vidyasagar D. Role of tele-medicine in retinopathy of prematurity screening in rural outreach centers in India - a report of 20,214 imaging sessions in the KIDROP program. Semin Fetal Neonatal Med 2015 Oct;20(5):335-345. [doi: 10.1016/j.siny.2015.05.002] [Medline: 26092301]
210. Wac K, Bults R, van Beijnum B, Widya I, Jones V, Konstantas D, et al. Mobile patient monitoring: the MobiHealth system. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:1238-1241. [doi: 10.1109/IEMBS.2009.5333477] [Medline: 19963997]
211. Wang S, Lee SB, Pardue C, Ramsingh D, Waller J, Gross H, et al. Remote evaluation of acute ischemic stroke: reliability of National Institutes of Health Stroke Scale via telestroke. Stroke 2003 Oct;34(10):e188-e191. [doi: 10.1161/01.STR.0000091847.82140.9D] [Medline: 14500929]
212. Wherton J, Greenhalgh T, Shaw SE. Expanding video consultation services at pace and scale in Scotland during the COVID-19 pandemic: national mixed methods case study. J Med Internet Res 2021 Oct 07;23(10):e31374 [FREE Full text] [doi: 10.2196/31374] [Medline: 34516389]
213. Widmer RJ, Allison TG, Lerman LO, Lerman A. Digital health intervention as an adjunct to cardiac rehabilitation reduces cardiovascular risk factors and rehospitalizations. J Cardiovasc Transl Res 2015 Jul;8(5):283-292 [FREE Full text] [doi: 10.1007/s12265-015-9629-1] [Medline: 25946990]

XSL•FO
RenderX

214. Williams A, Turer C, Smith J, Nievera I, McCulloch L, Wareg N, et al. Adoption of an electronic medical record tool for childhood obesity by primary care providers. Appl Clin Inform 2020 Mar 18;11(2):210-217 [FREE Full text] [doi: 10.1055/s-0040-1705106] [Medline: 32187633]
215. Wolf G, Mendelson D. The my health record system: potential to undermine the paradigm of patient confidentiality? Univ N S W Law J 2019 Jun;42(2):619-651. [doi: 10.53637/dded3771]
216. Wootton R, Wu WI, Bonnardot L. Nucleating the development of telemedicine to support healthcare workers in resource-limited settings: a new approach. J Telemed Telecare 2013 Oct;19(7):411-417. [doi: 10.1177/1357633X13506511] [Medline: 24218356]
217. Wrzosek N, Zimmermann A, Balwicki L. Survey on Patients' opinions and preferences concerning the use of the internet patient account in Poland: a preliminary report. Acta Pol Pharm Drug Res 2022 Jun 18;79(2):289-299. [doi: 10.32383/appdr/149254]
218. Joseph Wu TS, Kagoli M, Kaasbøll JJ, Bjune GA. Integrated disease surveillance and response (IDSR) in Malawi: implementation gaps and challenges for timely alert. PLoS One 2018 Nov 29;13(11):e0200858 [FREE Full text] [doi: 10.1371/journal.pone.0200858] [Medline: 30496177]
219. Yang CY, Chen RJ, Chou WL, Lee YJ, Lo YS. An integrated influenza surveillance framework based on national influenza-like illness incidence and multiple hospital electronic medical records for early prediction of influenza epidemics: design and evaluation. J Med Internet Res 2019 Feb 01;21(2):e12341 [FREE Full text] [doi: 10.2196/12341] [Medline: 30707099]
220. Yaqoob T, Mir F, Abbas H, Shahid WB, Shafqat N, Amjad MF. Feasibility analysis for deploying national healthcare information system (NHIS) for Pakistan. In: Proceedings of the 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services. 2017 Presented at: Healthcom '17; October 12-15, 2017; Dalian, China p. 1-6 URL: https://ieeexplore.ieee.org/document/8210836 [doi: 10.1109/healthcom.2017.8210836]
221. Yeh MJ, Saltman RB. Creating online personal medical accounts: recent experience in two developed countries. Health Policy Technol 2019 Jun;8(2):171-178. [doi: 10.1016/j.hlpt.2019.05.004]
222. Zaidi S, Kazi AM, Riaz A, Ali A, Najmi R, Jabeen R, et al. Operability, usefulness, and task-technology fit of an mHealth app for delivering primary health care services by community health workers in underserved areas of Pakistan and Afghanistan: qualitative study. J Med Internet Res 2020 Sep 17;22(9):e18414 [FREE Full text] [doi: 10.2196/18414] [Medline: 32940612]
223. Zaidi S, Shaikh SA, Sayani S, Kazi AM, Khoja A, Hussain SS, et al. Operability, acceptability, and usefulness of a mobile app to track routine immunization performance in rural Pakistan: interview study among vaccinators and key informants. JMIR Mhealth Uhealth 2020 Feb 13;8(2):e16081 [FREE Full text] [doi: 10.2196/16081] [Medline: 32053115]
224. Zanaboni P, Kummervold PE, Sørensen T, Johansen MA. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020 Feb 07;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
225. Zens M, Shajanian Zarneh Y, Dolle J, De Bock F. [Digital public health-leverage for community capacity building in health promotion : current situation, developmental issues and TEAviisari as a model implementation]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Jun 25;63(6):729-740. [doi: 10.1007/s00103-020-03148-1] [Medline: 32451594]
226. Zeretzke-Bien C, McCall J, Wylie T, Chowdhury MA, Balakrishnan M, Hendry P, et al. Using an online vaccination registry to confirm tetanus status in children with tetanus-prone wounds. West J Emerg Med 2020 Aug 21;21(5):1140-1146 [FREE Full text] [doi: 10.5811/westjem.2020.6.46582] [Medline: 32970567]
227. Zhang D, Shi L, Ido MS, Green DE, Li Y, Su D, et al. Impact of participation in a telestroke network on clinical outcomes. Circ Cardiovasc Qual Outcomes 2019 Jan;12(1):e005147. [doi: 10.1161/CIRCOUTCOMES.118.005147] [Medline: 30606053]
228. Zhou YY, Garrido T, Chin HL, Wiesenthal AM, Liang LL. Patient access to an electronic health record with secure messaging: impact on primary care utilization. Am J Manag Care 2007 Jul;13(7):418-424 [FREE Full text] [Medline: 17620037]
229. Zhou YY, Kanter MH, Wang JJ, Garrido T. Improved quality at Kaiser Permanente through e-mail between physicians and patients. Health Aff (Millwood) 2010 Jul;29(7):1370-1375. [doi: 10.1377/hlthaff.2010.0048] [Medline: 20606190]
230. GDP (current US$) - high income, low income, upper middle income, lower middle income. The World Bank. 2022. URL: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2022&locations=XD-XM-XT-XN&name_desc=true&start=2000 [accessed 2024-06-09]
231. Timpel P, Oswald S, Schwarz PE, Harst L. Mapping the evidence on the effectiveness of telemedicine interventions in diabetes, dyslipidemia, and hypertension: an umbrella review of systematic reviews and meta-analyses. J Med Internet Res 2020 Mar 18;22(3):e16791 [FREE Full text] [doi: 10.2196/16791] [Medline: 32186516]
232. Maaß L, Freye M, Pan CC, Dassow HH, Niess J, Jahnel T. The definitions of health apps and medical apps from the perspective of public health and law: qualitative analysis of an interdisciplinary literature overview. JMIR Mhealth Uhealth 2022 Oct 31;10(10):e37980 [FREE Full text] [doi: 10.2196/37980] [Medline: 36315221]

XSL•FO
RenderX

233. Sood S, Mbarika V, Jugoo S, Dookhy R, Doarn CR, Prakash N, et al. What is telemedicine? A collection of 104 peer-reviewed perspectives and theoretical underpinnings. Telemed J E Health 2007 Oct;13(5):573-590. [doi: 10.1089/tmj.2006.0073] [Medline: 17999619]
234. SO 13606-1:2019: health informatics — electronic health record communication: part 1: reference model. International Organization for Standardization (ISO). 2019. URL: https://www.iso.org/standard/67868.html [accessed 2024-05-11]
235. Sheikh M. Digital health information system in Africaâ€™s resource poor countries: current challenges and opportunities. J Health Inform Dev Ctries 2014;8(1):78-87.
236. Acilar A. Exploring the aspects of digital divide in a developing country. Inform Sci Inf Technol 2011;8:231-244. [doi: 10.28945/1415]
237. Socha-Dietrich K. Empowering the health workforce to make the most of the digital revolution. Organisation for Economic Co-operation and Development. URL: https://www.oecd-ilibrary.org/docserver/37ff0eaa-en.pdf?expires=1717739288&id=id&accname=guest&checksum=484AA5943C75F20A98800EA9F187ED9B [accessed 2024-04-29]
238. Brooks S, Donovan P, Rumble C. Developing nations, the digital divide and research databases. Ser Rev 2005 Dec;31(4):270-278. [doi: 10.1016/j.serrev.2005.09.002]
239. State of digital health around the world today. Global Digital Health Monitor (GDHM). URL: https://monitor.digitalhealthmonitor.org/indicators_info [accessed 2024-01-24]
240. Digital health atlas. World Health Organization (WHO). URL: https://digitalhealthatlas.org/en/-/ [accessed 2024-01-25]
241. Naszay M, Stockinger A, Jungwirth D, Haluza D. Digital age and the Public eHealth perspective: prevailing health app use among Austrian internet users. Inform Health Soc Care 2018 Dec;43(4):390-400. [doi: 10.1080/17538157.2017.1399131] [Medline: 29256715]
242. Maroju RG, Choudhari SG, Shaikh MK, Borkar SK, Mendhe H. Role of telemedicine and digital technology in public health in India: a narrative review. Cureus 2023 Mar;15(3):e35986 [FREE Full text] [doi: 10.7759/cureus.35986] [Medline: 37050980]
243. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan 2009 May;24(3):217-228 [FREE Full text] [doi: 10.1093/heapol/czp010] [Medline: 19304786]
244. Recommendations on digital interventions for health system strengthening. World Health Organization (WHO). 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2024-04-29]
245. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
246. Iwelunmor J, Blackstone S, Veira D, Nwaozuru U, Airhihenbuwa C, Munodawafa D, et al. Toward the sustainability of health interventions implemented in sub-Saharan Africa: a systematic review and conceptual framework. Implement Sci 2016 Mar 23;11:43 [FREE Full text] [doi: 10.1186/s13012-016-0392-8] [Medline: 27005280]
247. Kaboré SS, Ngangue P, Soubeiga D, Barro A, Pilabré AH, Bationo N, et al. Barriers and facilitators for the sustainability of digital health interventions in low and middle-income countries: a systematic review. Front Digit Health 2022;4:1014375 [FREE Full text] [doi: 10.3389/fdgth.2022.1014375] [Medline: 36518563]

Abbreviations

DiPH: digital public health
DiPHI: digital public health intervention
EHR: electronic health record
ESF: evidence standards framework
ICT: information and communication technology
NICE: National Institute for Health and Care Excellence
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
WHO: World Health Organization

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 24.10.23; peer-reviewed by J Köberlein-Neu, P Velmovitsky; comments to author 09.01.24; revised version received 31.01.24; accepted 15.05.24; published 17.07.24.

Please cite as:
Maaß L, Angoumis K, Freye M, Pan CC
Mapping Digital Public Health Interventions Among Existing Digital Technologies and Internet-Based Interventions to Maintain and Improve Population Health in Practice: Scoping Review
J Med Internet Res 2024;26:e53927
URL: https://www.jmir.org/2024/1/e53927
doi:10.2196/53927
PMID:39018096

©Laura Maaß, Konstantinos Angoumis, Merle Freye, Chen-Chia Pan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Interventions for Self-Management of Type 2 Diabetes Mellitus: Systematic Literature Review and Meta-Analysis

David Kerr[1], MBChB, DM; David Ahn[2], MD; Kayo Waki[3], MD, MPH, PhD; Jing Wang[4], MPH, PhD; Boris Breznen[5], PhD; David C Klonoff[6], MD

[1]Center for Health Systems Research, Sutter Health, Walnut Creek, CA, United States
[2]Mary & Dick Allen Diabetes Center at Hoag, Newport Beach, CA, United States
[3]The University of Tokyo, Tokyo, Japan
[4]Florida State University College of Nursing, Tallahassee, FL, United States
[5]Evidinno Outcomes Research Inc, Vancouver, BC, Canada
[6]Diabetes Research Institute, Mills-Peninsula Medical Center, San Mateo, CA, United States

Corresponding Author:
David Kerr, MBChB, DM
Center for Health Systems Research
Sutter Health
2121 N. California Ave, Suite 309
Walnut Creek, CA, 94596
United States
Phone: 1 8056248688
Email: david.kerr@sutterhealth.org

Abstract

Background: The proliferation of digital technology has the potential to transform diabetes management. One of the critical aspects of modern diabetes management remains the achievement of glycemic targets to avoid acute and long-term complications.

Objective: This study aims to describe the landscape of evidence pertaining to the relative effectiveness or efficacy and safety of various digital interventions for the self-management of type 2 diabetes mellitus (T2DM), with a primary focus on reducing glycated hemoglobin A_{1c} (HbA_{1c}) levels.

Methods: A systematic literature review (SLR) was conducted by searching Embase, MEDLINE, and CENTRAL on April 5, 2022. Study selection, data extraction, and quality assessment were performed by 2 independent reviewers. Eligibility criteria for the SLR included randomized controlled trials (RCTs) and comparative observational studies evaluating interventions containing both human (eg, coaching) and digital components (eg, glucose meter) in adult patients with T2DM. The primary meta-analysis was restricted to studies that reported laboratory-measured HbA_{1c}. In secondary analyses, meta-regression was performed with the intensity of coaching in the digital intervention as a categorical covariate.

Results: In total, 28 studies were included in this analysis. Most studies (23/28, 82%) used the reduction of HbA_{1c} levels as the primary end point, either directly or as a part of a multicomponent outcome. In total, 21 studies reported statistically significant results with this primary end point. When stratified into 3 intervention categories by the intensity of the intervention supporting the digital health technology (analyzing all 28 studies), the success rate appeared to be proportional to the coaching intensity (ie, higher-intensity studies reported higher success rates). When the analysis was restricted to RCTs using the comparative improvement of HbA_{1c} levels, the effectiveness of the interventions was less clear. Only half (12/23, 52%) of the included RCTs reported statistically significant results. The meta-analyses were broadly aligned with the results of the SLR. The primary analysis estimated a greater reduction in HbA_{1c} associated with digital interventions compared with usual care (–0.31%, 95% CI –0.45% to –0.16%; $P<.001$). Meta-regression estimated reductions of –0.45% (95% CI –0.81% to –0.09%; $P=.02$), –0.29% (95% CI –0.48% to –0.11%; $P=.003$), and –0.28% (95% CI –0.65% to 0.09%; $P=.20$) associated with high-, medium-, and low-intensity interventions, respectively.

Conclusions: These findings suggest that reducing HbA_{1c} levels in individuals with T2DM with the help of digital interventions is feasible, effective, and acceptable. One common feature of effective digital health interventions was the availability of timely and responsive personalized coaching by a dedicated health care professional.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55757) doi:10.2196/55757

KEYWORDS

coaching; digital health; eHealth; meta-analysis; patient empowerment; patient engagement; self-care; systematic review; telemedicine; type 2 diabetes; digital interventions; self-management; systematic literature review; effectiveness; efficacy; safety; meta-regression

Introduction

Digital health and telemedicine acceptance is rapidly growing, accelerated by the COVID-19 pandemic restrictions. Although it is difficult to estimate the acceptance of digital health in people with type 2 diabetes mellitus (T2DM) for methodological reasons [1], patients have access to a growing number of digital health technologies to support self-management of their condition. A recent study in Italy found that more than 70% of participants use continuous glucose systems [2]. The concept of self-management as an important part of long-term management of chronic diseases is gaining acceptance, and it is now considered essential for achieving long-term improvement in health outcomes and quality of life [3]. Compared with traditional approaches focused on managing a specific disease condition, the new paradigm is based on a patient-provider partnership involving collaborative care and education in chronic disease self-management [3]. Transition to this new paradigm has been increasingly important for patients with T2DM. A recent survey showed that the standard of care in T2DM, although generally acceptable, cannot meet the variety of patients' needs in terms of accessibility and timeliness of psychological, emotional, and behavioral support [4]. This unmet need can be alleviated by a wider use of digital technologies designed to help patients with their lifestyle and health-related decisions by making accessible critical data and on-demand consultations [5]. The technologies for managing T2DM include medical devices such as glucose meters, insulin pumps, continuous glucose monitors, and connected insulin pens; digital interventions including mobile apps, SMS text messaging, electronic communications, and videoconference platforms; and wearable technologies for monitoring health, such as activity trackers, sleep trackers, and smartwatches [6,7]. Digital health technologies can also support digital health care services, outside of a clinic or office, by using remotely collected data and communication capabilities of mobile phone devices and the internet [8]. The specific form of remote care can vary significantly: from occasional automated text messages to real-time teleconferencing with a dedicated health care professional (HCP). The intensity of remote care is therefore one of the factors that may impact the success of the interventions.

One of the critical goals of modern diabetes management remains the achievement of acceptable levels of glycemia to avoid the acute and long-term complications associated with T2DM [9]. Unfortunately, many individuals do not achieve their preferred glycemic targets or experience unwanted glycemic variability [10]. It has been suggested that digital technology has the potential to support people living with T2DM in their efforts toward achieving their glycemic goals [11,12]. A core need within diabetes self-management is to provide actionable information based on measured glucose levels [13]. This can be accomplished with timely information and possibly additional support from HCPs [8]. The advantages of digital technologies in managing glucose levels from patient's perspective were recently summarized in 3 essential concepts: competence, autonomy, and connectivity [14]. Competence refers to the understanding of the blood glucose levels with the help of supporting apps, autonomy means that the digital interventions allow for independent and timely decisions, and connectivity means that an HCP is always available through text messages or email. Digital interventions provide for all 3 components mentioned here, and thus, they are empowering the patients in their effort to cope with the disease. However, research suggests that people with T2DM may need more than knowledge about healthy eating, exercise, and self-monitoring of blood glucose [15]. They also need assistance in building insights into their daily health-related behaviors and routines [16,17].

The aim of this systematic literature review (SLR) and meta-analysis was to analyze digital health interventions for diabetes stratified by the levels of intensity of the intervention to determine whether (1) digital health interventions for diabetes are associated with improved outcomes and (2) whether the intensity of the intervention affects the degree of improvement. Additional outcomes of interest included user engagement measured by adherence or persistence, retention, and study withdrawal rates.

Methods

Eligibility Criteria

An SLR was undertaken following the standard methodologies for conducting and reporting systematic reviews as recommended by the Cochrane Handbook for Systematic Reviews of Interventions and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [18,19]. Study eligibility criteria were defined using the PICO (Population, Intervention, Comparator, and Outcomes) framework (Multimedia Appendix 1). Briefly, eligibility criteria for the SLR included randomized controlled trials (RCTs) and comparative observational studies evaluating interventions containing both human (eg, coaching) and digital components (eg, glucose meter) in adults (>18 years) with T2DM.

Two independent reviewers were responsible for reviewing all records, inclusive of conference proceedings and gray literature sources, at the title and abstract stage according to the predefined selection criteria. The eligible studies identified during the title and abstract screening proceeded to the full-text screening stage, where they were assessed for eligibility by the same reviewers. During each of the previous 2 screening stages, reviewers reconciled differences between their decisions, and in scenarios

XSL•FO
RenderX

of unresolved discrepancies, a third reviewer intervened to reach a consensus.

Studies that matched the PICO criteria following the full-text screening were included for data extraction. A standardized data extraction table was generated to define the study characteristics, including participant characteristics, intervention characteristics, and outcomes from eligible studies. Two independent reviewers extracted all relevant data from the final list of included studies. The reviewers reconciled discrepancies between their data extraction, and in scenarios of unresolved discrepancies, a third reviewer intervened to reach a consensus.

Information Sources

Relevant studies were identified by searching the following databases on April 5, 2022: Embase (Multimedia Appendix 2), MEDLINE (Multimedia Appendix 3), and CENTRAL (Multimedia Appendix 4). Abstracts from relevant conferences held between 2018 and 2022 were also searched via Embase or their respective websites. Additionally, selected company websites (Dario, Lark, Livongo, Omada, Onduo, OneDrop, Vida, Virta, and Welldoc) and the US clinical trials registry (ClinicalTrials.gov) were searched.

Risk of Bias

For quality control, 2 independent reviewers assessed the quality of the included studies using the Cochrane risk of bias tool for RCTs and the Risk of Bias in Nonrandomized Studies–Interventions tools. A third investigator intervened to reach a consensus if there were any unresolved conflicts. Results were summarized in a narrative form.

Synthesis of Results (Qualitative)

Following the study selection, the results were summarized by grouping the interventions into 3 broad categories as described below. The 2 main components present in all interventions were the technological (devices and software) and the human (coaching). The coaching sessions varied markedly in terms of their frequency (how often the HCPs communicated with the individual with T2DM), duration (both duration of the individual sessions and overall duration of coaching), mode of communication (in person, videoconferencing, phone calls, and SMS text messaging), and the content (personalized vs generic). Categories were created by considering the features and intensity of the coaching component, and for categorization, the intervention had to meet most of the following criteria:

- High intensity: Participant data are automatically uploaded to the cloud at regular intervals. The coaching includes personalized motivational and goal-setting components based on the most recent data and is delivered by dedicated HCP staff. Communication happens regularly, either in person or remotely, at least once per week. Education includes specific modules explaining disease, behavioral strategies, and psychological coping.
- Medium intensity: Participant data are manually uploaded to the clinic database. Coaching includes personalized advice based on individual data but does not include behavioral advice in terms of motivational and goal-setting components. The communication is ad hoc and initiated by the HCPs. Education includes general information about the disease and technical information about the use of the devices.
- Low intensity: Participant data sharing is limited (eg, patients brought the glucose meters to the center, or a nurse visited patients), and the feedback is generic, often using preexisting templates. The communication is asynchronous or delayed (eg, email or follow-up phone call). There is limited or no education.

In addition to separating the studies into the 3 categories of intervention intensity, 3 additional features of the coaching were identified as potentially relevant to the success of an intervention:

- Communication mode—synchronous versus asynchronous: Synchronous mode meant that participants were in direct contact in real time with the HCP or the coach (eg, a telephone call or a teleconference) [20]. Asynchronous communications usually involved web-based portals, emails, or SMS text messages [21].
- Frequency of communication: this varied considerably across the studies, and therefore, the final binary classification was chosen to be unlimited communications [22], or restricted or scheduled communications [23].
- Qualification of the coaches: diabetes specialists (eg, certified diabetes educators [24] or diabetes nurse educators [25]) versus general HCPs (such as general practitioners [26] or study nurses [27]).

Synthesis of Results (Meta-Analysis)

Nine random effects meta-analyses were conducted on the mean difference (MD) in the change in glycated hemoglobin A_{1c} (HbA_{1c}). The primary analysis comprised all the included RCTs with laboratory-measured HbA_{1c} levels and was performed both with and without intervention intensities as categorical covariates via a meta-regression. Subgroup analyses were conducted for high-intensity (number of studies, k=4), medium-intensity (k=12), high- and medium-intensity (k=16), and low-intensity (k=4) interventions. Sensitivity analyses were conducted (1) including studies with nonlaboratory measured HbA_{1c}, (2) excluding studies with continuous glucose monitoring (CGM), and (3) excluding studies identified as posing a high risk of bias using the Cochrane risk of bias tool.

All meta-analyses were conducted in R (version 4.1.1, The R Foundation for Statistical Computing) using the *metafor* (version 3.0-2) package. The restricted maximum likelihood estimator was used to measure the between-study variance (τ^2) as heterogeneity due to variation in intervention design, follow-up time, and clinical population across the evidence base was anticipated. We also report the estimated heterogeneity using Q and I^2. If more than 1 usual care arm was present within a single study, then they were pooled into 1 sample size weighted mean prior to the meta-analysis. In cases where a study reported multiple timepoints, the final time point was chosen for analysis.

XSL•FO
RenderX

Results

SLR Results

Study Selection and Study Characteristics

In total, 6288 papers were identified from the SLR, including 6275 papers via Embase, MEDLINE, and CENTRAL and 13 additional papers through conference proceedings and company websites. After title and abstract screening and full-text screening, a total of 28 studies [20-47] were included in the SLR (Figure 1).

Of the 28 included studies, 23 (82%) were RCTs [20-30,32-34,36-43,45]; 2 (7%) were nonrandomized comparative [31,46]; and the remaining 3 (11%) were cross-sectional [44], prospective cohort [47], and retrospective cohort [35] studies. There were 9 countries where the studies were conducted: the United States (n=12); South Korea (n=6); the United Kingdom (n=3); China (n=2); and 1 each in Belgium, Canada, France, India, and Malaysia. The studies were published between 2003 and 2021. The study population ranged from a minimum of 17 [42] to 772 [35], with an average population size of 202 and a median of 143 participants.

Follow-up durations ranged from 1 month [26] to 24 months [31], with an average follow-up duration of 7.9 months and a median of 6 months (1 abstract did not include information on the follow-up duration [44]).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram.

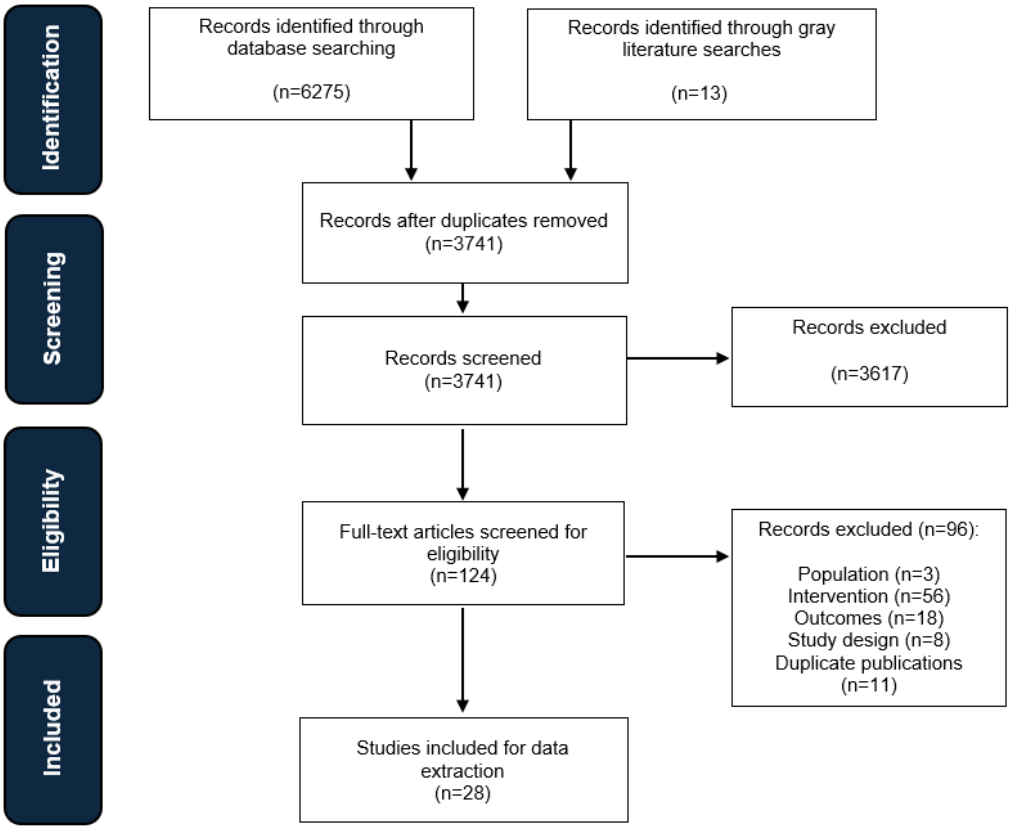

Participant Characteristics

All studies enrolled populations with T2DM, but some studies targeted subpopulations of individuals with T2DM that met specific criteria. A total of 4 studies enrolled only participants taking glucose-lowering prescription medications [24,37,39,42]. Individuals with suboptimally controlled diabetes were investigated by 5 studies [27,29,35,41,44]. The definition of "suboptimally controlled" varied. In some cases, the criterion was set by quantitative thresholds: 2 consecutive HbA_{1c} recordings greater than 8% in the previous 12 months [29,41] or HbA_{1c}≥7.5% to ≤13% [27]. However, in 2 studies, the definition was only descriptive with no quantitative data [35,44]. Four studies targeted populations with low income or low socioeconomic status [20,22,32,38]. Out of the 4 studies, 2 (50%) studies enrolled only individuals who were overweight (BMI≥25) [31] or obese (BMI≥41) [46]. The 2 studies selected participants from a pool of insured patients [30,45]. One study enrolled only women with T2DM [36], and another study enrolled self-described physically inactive individuals [34]. The mean age of the participants ranged from 47.3 [40] to 64 years [24], with an overall mean age of 55.7 years, a median age of 54.3 years, and 51% (range 29%-100%) being female. Ethnicities and races in the studies included Black, Chinese, Korean, and White. A total of 11 studies reported the baseline average disease duration from 2.6 [42] to 14 years [23], with an overall mean of 7.9 years and a median of 8 years.

All included studies reported the average (or median) level of HbA_{1c} at baseline, and all studies tested disparities between intervention groups. None of them reported significant differences in baseline HbA_{1c} levels. The overall average

XSL•FO
RenderX

baseline HbA_{1c} levels varied from 6.8% [40] to 10.9% [39], with an overall mean of 8.6% and median of 8.5%. The distribution of the mean baseline HbA_{1c} across the studies was as follows (Multimedia Appendix 5): ≤7% (2 studies), >7% and ≤8% (7 studies), >8% and ≤9% (10 studies), and >9% (9 studies).

Intervention Characteristics

The technology component of the interventions included technology for measuring glucose: either CGM (5 studies) [26,35,36,43,47], self-monitoring of blood glucose (21 studies) [20-24,27-33,37-42,44-46], or both [25]. An exception was a study measuring blood glucose in a clinical setting [34]. In addition to glucose-monitoring devices, several studies used connected scales [31,46] for weight monitoring or accelerometers to monitor physical activity [21,25,30,34,36,41]. The glucose data were usually uploaded to a central server and then used by HCPs to adjust the treatment regimen and to coach or advise the participants on appropriate actions. The information about the measured glucose levels was generally available to participants either directly through the device display or sometimes through a visualization software application or a dedicated website. Additional details on the digital interventions, the usual care groups, and coaching components of the included studies are summarized in Multimedia Appendix 6 [20-47].

Outcome Characteristics

Most studies reported improvement in glycemic control in patients with T2DM using HbA_{1c} levels as the primary end point either directly or as a part of a multicomponent outcome. A few studies used feasibility, acceptability, and self-efficacy of the intervention as their end point [36,43,47], and 1 study used a physical activity end point [34]. The breakdown of study end points and the respective number of significant results is summarized in Table 1.

The classification of the studies into the three categories of interventions as outlined in the *Methods* section yielded the following stratification: (1) a total of 7 studies in the high-intensity category [20,22,26,35,41,45,46] (out of which 5 were RCTs) [20,22,26,41,45], (2) a total of 16 studies in the medium-intensity category [21,23,25,27,29-33,37-40,42-44] (out of which 14 were RCTs) [21,23,25,27,29,30,32,33,37-40,42,43], and (3) a total of 5 studies in the low-intensity category [24,28,34,36,47] (out of which 4 were RCTs) [24,28,34,36].

Table 2 shows reported outcomes across the 3 categories as measured by the number of significant primary end points (across all 28 studies and across the 23 RCTs, respectively).

Table 3 summarizes the reported successes in the comparative reduction of HbA_{1c} across the 3 intervention categories within the included RCTs. The data show the number of studies reporting a statistically significant difference in HbA_{1c} reduction between the intervention arm and the comparator arm.

Table 4 shows the summary of the successes in the comparative reduction of HbA_{1c} separated into the categories outlined earlier. Only results from RCTs are included.

Participant engagement and satisfaction were investigated in 3 studies based on their involvement in the counseling and educational sessions [36], reported measurement of the burden [47], or using validated questionnaires targeting self-care and self-efficacy [43]. The tools by which the studies measured some aspects of participant satisfaction consisted of standardized questionnaires and exit interviews. In the studies investigating user engagement, significant differences between the intervention and usual care groups in terms of changes in self-care behaviors were observed. Overall, digital interventions were well received with high completion rates (most of the studies had dropout rates below 20%) and acceptable additional burden to the patients. In the studies investigating satisfaction using the Diabetes Treatment Satisfaction Questionnaire (DTSQ), 2 (67%) out of 3 studies reported significant improvement in DTSQ scores in the digital intervention groups, compared with usual care. One study reported DTSQ improvement from 31.9 (SD 10.1) to 42.0 (SD 3.8) points (P<.001) in the intervention group, and from 34.3 (SD 8.5) to 36.4 (SD 8.9) points in the control group (P=.10). This difference between the 2 groups was significant (P=.01) [39]. In the second study, DTSQs showed a significant rise only in the intervention group, resulting in a 2.21-point increase in the intervention group compared with the control group at 3 months (P=.01) [33].

Another aspect of user engagement can be inferred from dropout rates. A total of 25 studies reported dropout rates (defined as the number of participants enrolling in the program but not finishing for any reason). Most studies reported ≤20% dropout, with rates of >30% reported in only 3 studies [20,27,45].

A total of 11 (39%) of the 28 included studies reported on adverse events. Seven of these studies reported no intervention-related adverse events [22,25,27,37,38,45,46]. Four studies reported the occurrence of adverse events without commenting on their relationship to digital interventions [30,31,33,42]. The most common adverse event reported in these 4 studies was hypoglycemia, followed by cardiovascular events, cancer, and metabolic disruptions. Adverse events were reported in both the intervention groups and the usual care groups. The authors of these studies were agnostic about the causal relationship between digital interventions and adverse events.

XSL•FO
RenderX

Table 1. Study end points and reported results.

Primary end point	Studies (N=28), n (%)	Studies with significant results[a] (n=21), n (%)
Change in HbA_{1c}[b]	19 (68)	14 (67)
Multicomponent outcomes including HbA_{1c}[c]	4 (14)	4 (19)
HOMA2-IR[d]	1 (4)	1 (5)
Feasibility, acceptance, and self-efficacy	3 (11)	2 (10)
Physical activity	1 (4)	0 (0)
Total	28 (100)	21 (100)

[a]Number of studies achieving statistically significant results in primary end point.
[b]HbA_{1c}: glycated hemoglobin A_{1c}.
[c]Multiple primary end points: HbA_{1c}, glycemic control (HOMA2-IR, glycemic variability, fasting blood glucose, and postprandial 2-hour blood glucose), medication use, BMI, weight control, and retention rate.
[d]HOMA2-IR: Homeostatic Model Assessment of Insulin Resistance.

Table 2. Success rate in achieving its predetermined primary end point across the 3 intervention categories.

Intervention category	Studies (N=28), n (%)	Significant end points[a], n/N (%)	RCTs[b] (n=23), n (%)	Significant end points[c], n/N (%)
High intensity	7 (25)	6/7 (86)	5 (22)	4/5 (80)
Medium intensity	16 (57)	12/16 (75)	14 (61)	10/14 (71)
Low intensity	5 (18)	3/5 (60)	4 (17)	3/4 (75)

[a]Number of studies achieving statistically significant results in primary end point.
[b]RCT: randomized controlled trial.
[c]Number of RCTs achieving statistically significant results in primary end point.

Table 3. Success rate of randomized controlled trials in achieving a reduction of glycated hemoglobin A_{1c} across the 3 intervention categories.

Intervention category	RCTs[a] (n=23), n (%)	Significant results[b], n/N (%)
High intensity	5 (22)	2/5 (40)
Medium intensity	14 (61)	8/14 (57)
Low intensity	4 (17)	2/4 (50)

[a]RCT: randomized controlled trial.
[b]Number of studies achieving statistically significant results in comparative reduction of HbA_{1c}.

XSL•FO
RenderX

Table 4. Significant comparative reduction of glycated hemoglobin A_{1c} by intervention features.

Intervention feature	RCTs[a] (n=23), n (%)	Significant results[b], n/N (%)
Communication mode		
Synchronous	12 (52)	4/12 (33)
Asynchronous	11 (48)	8/11 (73)
Frequency of communications		
Unlimited	7 (30)	4/7 (57)
Restricted	16 (70)	8/16 (50)
Qualification of coaches		
Diabetes specialists	11 (48)	6/11 (55)
General HCPs[c]	12 (52)	6/12 (50)

[a]RCT: randomized controlled trial.

[b]Number of studies achieving statistically significant results in comparative reduction of glycated hemoglobin A_{1c}.

[c]HCP: health care professional.

Study Quality Assessment and Risk of Bias

The Cochrane risk of bias tool [48] was used to assess the 23 RCTs (Multimedia Appendix 7). One (4%) study had a low risk of bias, 17 (74%) studies had some concerns regarding the overall risk of bias, and 5 (22%) studies had a high risk of bias overall (3 studies did not blind investigators, participants, or interventionists to group assignment [20,23,36]; 1 study used an inappropriate method of measuring HbA_{1c} (self-reported via a questionnaire) [26]; and 1 did not report how outcome data were collected) [37]. The Risk of Bias in Nonrandomized Studies–Interventions assessment tool [49] was used to evaluate the 5 nonrandomized interventional studies (Multimedia Appendix 8). Overall, 3 (60%) studies had a low risk of bias, and 2 (40%) studies were considered to contain insufficient information on which to base a judgment about the risk of bias, mainly because of missing information in one or more key domains [35,44].

Meta-Analysis

Primary Analyses

In the primary analysis, the random effects meta-analysis (20 studies) [20-25,27,28,30,33,34,36-43,45] comparing the change in HbA_{1c} for intervention (number of patients across all studies, n=1637) versus usual care (n=1389) estimated an MD of –0.31% (95% CI –0.45% to –0.16%; P<.001). Heterogeneity was statistically significant (Q=57.64; df=19, P<.001), with an estimated τ of 0.21 (95% CI 0.12-0.61) and I^2 of 67.54% (95% CI 41.19%-94.48%).

In the meta-regression of studies that measured HbA_{1c} in a laboratory with intervention intensity as a categorical covariate (Figure 2), random effects meta-analysis (20 studies) [20-25,27,28,30,33,34,36-43,45] of change in HbA_{1c} on intensity for intervention (n=1637) versus usual care (n=1389) was estimated using the following equation:

where is the predicted MD, “intensity=low” is 1 if the intervention is low intensity and 0 otherwise; and “intensity=medium” is 1 if the intervention is medium intensity and 0 otherwise. This predicts an MD of –0.45% (95% CI –0.81% to –0.09%; P=.02) for high-intensity interventions, –0.29% (95% CI –0.48% to –0.11%; P=.003) for medium-intensity interventions, and –0.28% (95% CI –0.65% to 0.09%; P=.20) for low-intensity interventions. The low-intensity (P=.51) and medium-intensity (P=.45) coefficients were not statistically significant. Heterogeneity was statistically significant (Q=50.84, df=17, P<.001), with an estimated τ of 0.23 (95% CI 0.13-0.68) and I^2 of 67.74% (95% CI 40.87%-94.72%).

XSL•FO
RenderX

Figure 2. Forest plot of the MD and CI of change in HbA_{1c} with meta-regression on intervention intensity. * indicate statistical significance. HbA_{1c}: glycated hemoglobin A_{1c}; MD: mean difference.

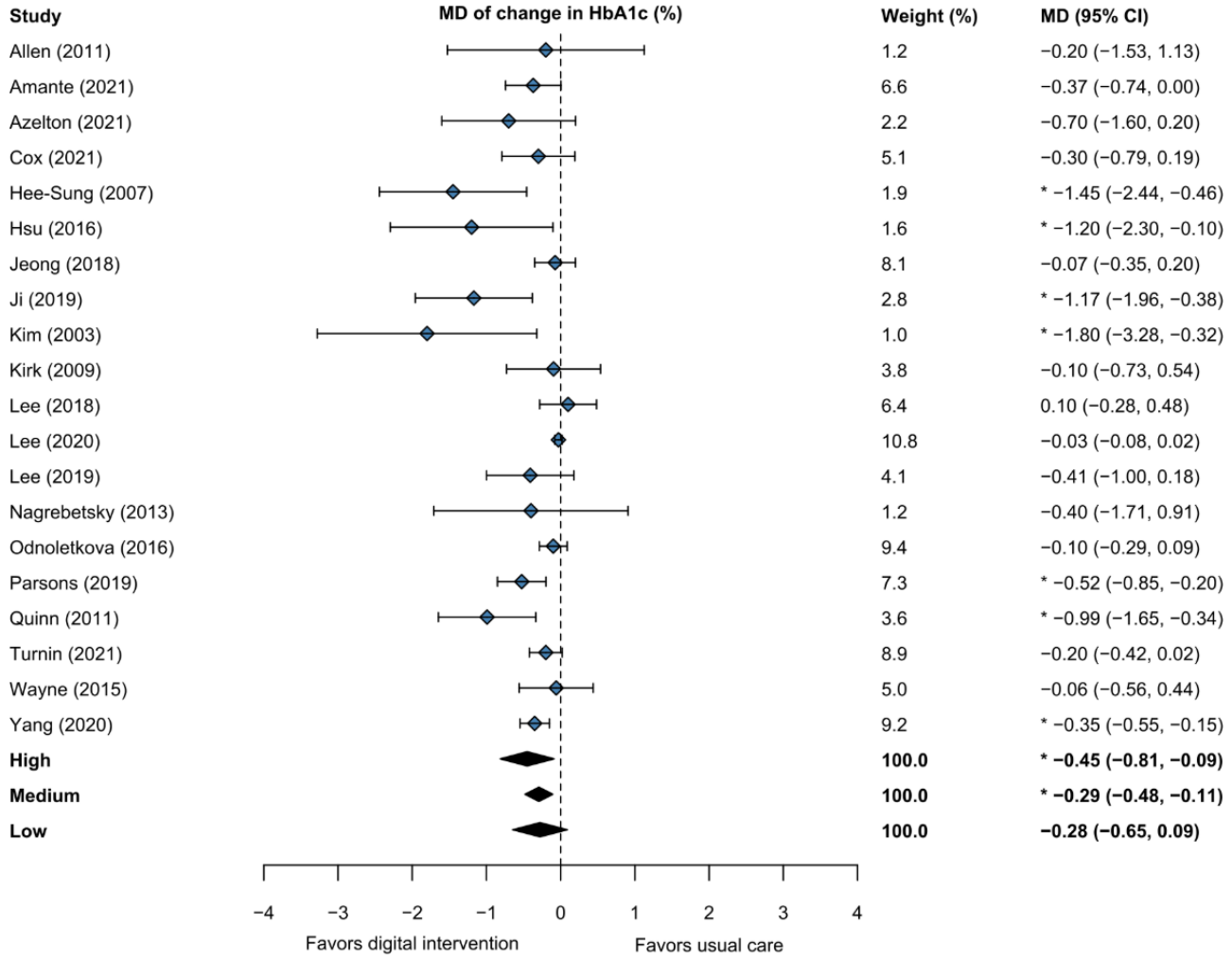

Subgroup Analyses

For the high-intensity interventions, the random effects meta-analysis (4 studies) [20,22,41,45] of change in HbA_{1c} for intervention (n=253) versus control (n=201) estimated an MD of –0.46% (95% CI –0.84% to –0.07%; P<.02). Heterogeneity was statistically significant (Q=5.35, df=3, P<.20), with an estimated τ of 0.27 (95% CI 0.00-1.47) and I^2 of 46.77% (95% CI 0.00%-96.43%).

For the medium-intensity interventions, the random effects meta-analysis (12 studies) [21,23,25,27,30,33,37-40,42,43] of change in HbA_{1c} for intervention (n=989) versus control (n=847) estimated an MD of –0.28% (95% CI –0.45% to –0.11%; P<.002). Heterogeneity was statistically significant (Q=38.76, df=11, P<.001), with an estimated τ of 0.20 (95% CI 0.11-0.87) and I^2 of 68.18% (95% CI 39.74%-97.57%).

For the combined high- and medium-intensity interventions, the random effects meta-analysis (16 studies) [20-23,25,27,30,33,36-43,45] of change in HbA_{1c} for intervention (n=1242) versus usual care (n=1048) estimated an MD of –0.32% (95% CI –0.47% to –0.16%; P<.001). Heterogeneity was statistically significant (Q=50.34, df=15, P<.001), with an estimated τ of 0.21 (95% CI 0.12-0.68) and I^2 of 67.49% (95% CI 38.55%-95.46%).

For the low-intensity interventions, random effects meta-analysis (4 studies) [24,28,34,36] of change in HbA_{1c} for intervention (n=395) versus usual care (n=341) estimated an MD of –0.34% (95% CI –0.83 to 0.16; P<.2). Heterogeneity was not statistically significant (P<.09; Q=6.73, df=3), with an estimated τ of 0.38 (95% CI 0.00-1.90) and I^2 of 60.64% (95% CI 0.00 to 97.50).

Sensitivity Analyses

Three sensitivity analyses were performed using (1) inclusion of studies using nonlaboratory-based HbA_{1c} measurement, (2) exclusion of studies that used CGM, and (3) exclusion of studies identified as high risk of bias according to the Cochrane risk of bias assessment. The results of all 3 analyses were in line with the primary analysis yielding a significant MD effect in favor of the digital intervention as compared with usual care (Table 5).

XSL•FO
RenderX

Table 5. Summary results of all conducted analyses.

Analysis set	Studies, n	Pooled mean difference estimate (%; 95% CI)	Significant heterogeneity
Primary			
Studies with laboratory-measured HbA_{1c}[a]	20	–0.31 (–0.45 to –0.16)[b]	Yes
Studies with laboratory-measured HbA_{1c}, meta-regression			Yes
High	20	–0.45 (–0.81 to –0.09)[b]	
Medium	20	–0.29% (–0.48 to –0.11)[b]	
Low	20	–0.28% (–0.65 to 0.09)[c]	
Subgroups			
High-intensity interventions	4	–0.43 (–0.78 to –0.09)[b]	No
Medium-intensity interventions	12	–0.28 (–0.45 to, –0.11)[b]	Yes
High- and medium-intensity interventions	16	–0.32 (–0.47 to –0.16)[b]	Yes
Low-intensity interventions	4	–0.34 (–0.83 to 0.16)[c]	No
Sensitivity			
Including nonlaboratory-measured HbA_{1c} studies	23	–0.40 (–0.56 to –0.24)[b]	Yes
Excluding continuous glucose monitoring studies	18	–0.31 (–0.47 to –0.15)[b]	Yes
Excluding high-risk studies	16	–0.31 (–0.46 to –0.15)[b]	Yes

[a]HbA_{1c}: glycated hemoglobin A_{1c}.

[b]Statistically significant.

[c]Nonsignificant.

Discussion

Principal Findings

In this SLR and meta-analysis, the currently available evidence suggests that the use of digital health interventions, compared with usual care, is associated with clinically significant improvement in HbA_{1c} levels for individuals with T2DM. Furthermore, the intensity of support provided by HCPs also appears to impact the HbA_{1c} levels. Here, intensity included the types and frequency of interactions between professionals and people with T2DM as well as the qualifications of the professional.

Although most (17/23, 74%) of the RCTs reported their primary end points as defined by the study protocol, achieving a significant comparative reduction of HbA_{1c} between digital health intervention and usual care appears to be challenging. Only half (12/23, 52%) of the RCTs reported that the digital health interventions, compared with usual care, were successful in achieving a statistically significant difference in HbA_{1c} reduction. In addition, there was variability in the performance of different digital interventions. Based on this analysis, 2 essential components of each intervention—technology and coaching—seem to independently influence the outcomes. Information about the self-measured glucose levels was available to participants in all included studies. With 1 exception, the intervention arm included either self-monitoring of blood glucose or CGM devices. The devices provided on-demand glucose data to the person using the device as well as to the supervising health care team. Therefore, the availability of raw glucose data does not account for the observed outcome differences between the studies since the glucose data were available to both the intervention group and the control group in all studies. Even the presence or absence of additional devices (such as connected weight scales [46] or accelerometers [25,34]) did not appear to make a difference. Consequently, easy access to self-measured data alone did not seem to be a sufficient condition for improved glycemic control.

In addition to the data provided by the devices, the other aspect of the intervention was coaching. Coaching can be stratified into 2 components: education and counseling. All studies provided educational sessions to the participants, albeit the extent and quality of the education varied. Some studies provided only basic forms of education usually based on preexisting materials published by outside sources (such as the American Diabetes Association guidelines [37,45] or Diabetes UK [34]). The educational content was restricted to general diabetes information and to the technical aspects of blood glucose monitoring [24,27,30,38]. The educational sessions were led by nurses and the participants were mostly receiving preprinted materials and watching prerecorded videos. In higher-tier interventions (as defined in the *Methods* section), the educational materials were usually produced in house by the institution conducting the study [20,29,35,36] and tailored

XSL•FO
RenderX

to the needs of the target population. The educational sessions were led by specialists (such as diabetes educators [20,36], dietitians [29,31], or pharmacists [38]) and often in small groups or in one-on-one settings. The sessions had the active participation of both the health care staff and the participants.

The coaching element was the most distinctive feature which differentiated the more intensive interventions from the lower-tier interventions (Table 2). The top-tier interventions provided regular, individualized coaching sessions with trained diabetes educators, using graphical visualization tools to go over an individual's data with them and advise on the best course of action. Sessions were in person [36] or remote via videoconferencing [35,41,45], voice call [26], SMS text messaging [46], or a mixture of these. The content of the coaching sessions was tailored to the specific goals of the digital intervention. Three examples of such coaching are problem-solving [36], where participants were asked to discuss their specific barriers in implementing the intervention; development of an individualized plan to improve problem-solving skills and self-care [41], implemented within the diabetes program and targeting 7 self-care behaviors; and motivational interviewing, goal-setting, and confidence building [20].

Across the high-intensity studies (including the 3 examples listed earlier), coaching was frequent, available on-demand, and tailored for the individual. The medium for communication did not appear to be important.

Further, additional features of the interventions such as mode of communication, frequency of the communications, and qualification of the coaches were also important. The most pronounced difference was between studies using synchronous compared with asynchronous communication, with the advantage favoring the use of an asynchronous mode of communication. This finding appears counterintuitive, as direct human-to-human contact is the most common way of coaching (health or otherwise). One possible explanation may come from the fact that the unlimited frequency of communication also seems to have a slight advantage over restricted or scheduled communications. The availability of the coaches for direct contact is constrained by the patient-to-coach ratios and the limited amount of time that each coach will be available in real time. On the other hand, asynchronous communication modalities such as text messages and emails allow for near-real-time communication without the logistic constraints of direct interactions.

This finding is novel and not yet supported by other studies. A systematic review of longitudinal management of chronic conditions by telehealth interventions [8] reported no difference between asynchronous and in-person (synchronous) reviews of patient data. However, the asynchronous mode was represented by a dedicated web page rather than text messaging, so the comparison may not be quite relevant. In an umbrella review of technology-enabled diabetes self-management [50], a new taxonomy for digital interventions was proposed. This taxonomy includes a distinction between synchronous and asynchronous feedback modes, but the authors noted poor reporting on this issue in the reviewed studies. Here, the qualifications of the coaches were the least significant factor in this analysis. However, all included RCTs were driven by a protocol outlining the important details of the intervention. All staff participating in those interventions were therefore instructed before the beginning of the trial in the proper method of coaching and patient interactions. Even the nondiabetes specialists were given specialized instructions on how to approach the patients which may have contributed to blurring the distinction between diabetes specialists and general HCPs.

When analyzing the comparative reduction of HbA_{1c} within the context of RCTs, the meta-analysis confirmed the findings from the SLR. A statistically significant reduction in HbA_{1c} relative to usual care was observed globally for high- and medium-intensity interventions, but not for low-intensity interventions. Meta-regression coefficients were not statistically significant, and hence no support was found for a difference in efficacy according to intervention intensity, but this finding was limited by the relatively small number of high- and low-intensity interventions. The relatively modest, although statistically significant, effect size observed in the comparative reduction of HbA_{1c} levels may be explained by several factors present in all included RCTs. First, the participants in the comparator arms of the studies were receiving usual diabetes care and the results show that this level of care also reduced their HbA_{1c}. In this context, digital interventions can be viewed as an adjunct therapy. Second, the improved efficacy relative to the usual care might not be the only advantage of digital interventions. Finally, the effect size observed in this meta-analysis is similar to the one reported elsewhere investigating mobile health efficacy in diabetes treatment and management across developing and developed countries [51]. In addition, digital intervention can induce behavior change via coaching sessions that include problem-solving and identifying barriers, so the effects of HbA_{1c} improvement caused by this behavior change could be sustainable compared with usual care.

The sensitivity analysis including studies with non–laboratory-measured HbA_{1c} levels resulted in a more favorable result for digital interventions. This suggests that the decision to exclude those studies led to a more conservative result. The other 2 sensitivity analyses did not change the result, and so we conclude that the inclusion of CGM and high-risk of bias studies was not a determining factor in the results. One caveat to the observation that the inclusion of CGM did not change the results was the fact that the number of studies with CGM was 2. In this analysis, the evidence base had studies spanning a broad range of years including early years when CGM was less prevalent.

In the studies from this review investigating engagement, the results showed that the digital intervention led to a significant difference between the intervention and usual care groups in terms of changes in self-care behaviors, and that digital interventions are well received with high completion rates and no additional burden. In the 3 studies investigating patient satisfaction using DTSQ questionnaires, all reported significant improvement in DTSQ scores in the digital intervention groups, suggesting a high satisfaction with the treatment. A challenge facing the reports on engagement is the lack of clear

XSL•FO
RenderX

differentiation between patient participant adherence and engagement. Those 2 concepts are often used interchangeably; however, they refer to different aspects of an individual's behavior [52]. The usual definition of patient adherence includes a willingness to follow the study protocol in all aspects. Engagement includes an individual's initiative to actively seek improvement in their disease management. This distinction is important for differentiating between passive following of instructions and self-initiated activities.

A total of 11 (39%) of the 28 included studies reported on adverse events. Seven of these studies reported no intervention-related adverse events and 4 studies did report the occurrence of adverse events without stating whether those were related to the intervention or not. Given the nature of these events and the fact that the participants remained on their previously prescribed medication regimen, the link between the digital interventions and the adverse events is difficult to establish. Overall, the reporting on adverse events in the included studies was poor and this constitutes an unmet need in the domain of digital interventions.

Limitations of the Study

There are multiple limitations to our review and meta-analysis. First, the apparent relationships and conclusions regarding the intensity of the digital intervention must be tested in a prospective manner to see whether they prove to be valid. Next, the heterogeneous nature of the featured interventions makes it difficult to generalize the findings. This has been a consistent theme and conclusion in other systematic reviews with or without meta-analyses in digital health. The heterogeneity can be seen in a variety of study settings across multiple countries with different cultures that may influence the acceptance of the intervention. This aspect was not addressed in the individual studies. Our approach in this study was to stratify the interventions based on the intensity of the coaching as the most distinctive pattern among the interventions. However, because of a wide variety of coaching strategies, clear boundaries between the categories were not easy to draw. This was also true for selecting the 3 additional features of the interventions (communication mode, frequency of communications, and qualification of the coaches) used to further investigate the factors contributing to efficacy.

The meta-regression and subgroup analyses were limited by the small number of high- and low-intensity interventions, which resulted in low power to detect differences in HbA_{1c} reduction according to intervention intensity. The effect of time of follow-up was also not investigated.

As well, only RCTs were included in the meta-analysis. Although this reduced the introduction of bias associated with nonrandomized studies, it does limit the generalizability of the findings in the real world [53]. The risk-of-bias analysis revealed that all RCTs in the evidence base contained some degree of bias. Although no difference in the degree of bias between 1 intervention and another was found, the presence of bias confounds the results of the meta-analysis.

With respect to the scalability of the interventions in the real world, the need for the dedicated staff of HCPs to support the higher-intensity studies adds additional economic and logistic burden. Some of the solutions may require a dedicated database; communication infrastructure; customized user software; and trained, professional staff. After adding the necessary maintenance expenditures, the overall cost of these solutions may be out of the reach of certain clinics. Finally, some digital interventions are intended to deliver behavior change, but there was a paucity of clear evidence that behavior changed because the behavior change aspects of the intervention were not measured appropriately.

Recommendations for Future Research

Based on the findings, recommendations for future research in digital health include the following. (1) An agreed definition of engagement in digital health as an end point may help with improved targeting of interventions. (2) Reporting should standardize digital health data into meaningful outcomes by therapeutic area (and then beyond), such as sensor data and patient-reported outcome measures, so that future systematic reviews and meta-analyses can be less heterogeneous. (3) Studies in digital interventions should strive for a clear reporting of adverse events, especially in terms of the relationship between the digital health product and the adverse events. (4) Digital health studies that include coaching should systematically record multiple dimensions of the intervention, including frequency, duration, asynchronous versus synchronous, coaching or behavior change techniques deployed, human coach qualification (if relevant), and guidance and introduction to patients. For methodological purposes, a newly developed scoring system for the classification of the intensity of coaching would help future analyses of digital interventions. (5) Consider different designs and methodologies to study digital health interventions, especially those that are intended to deliver behavior change, so that meaningful patient engagement in the digital solution and outcome measures aligned with intended use can be assessed. To eliminate bias, perhaps cluster randomization (or some other method for eliminating bias) should be used in future digital health interventions.

Conclusions

Reducing HbA_{1c} levels in patients with T2DM with the use of digital interventions, in addition to usual care, is feasible and acceptable to people with T2DM, as consistently demonstrated by a large number of studies of various populations, goals, and methods of interventions. When analyzing the comparative efficacy of digital interventions within the context of RCTs, the advantage of digital interventions becomes less pronounced. Some forms of intervention perform better than others, but it is difficult to identify the exact reasons for this difference given the variety of methodologies featured in the studies. However, a broadly defined intensity of coaching seems to play an important role. A common feature of successful studies was the availability of timely and responsive personalized coaching. Therefore, the relevance and the content of the coaching are more important than the communication medium used to deliver the messages. Scaling up the personalized, on-demand coaching featured in some of the studies may lead to logistical and economic roadblocks. Overcoming these roadblocks will largely determine the success of digital interventions in real-world

XSL•FO
RenderX

clinical practice. In conclusion, digital health interventions for diabetes appear to be a useful tool for improving outcomes.

Acknowledgments

The authors would like to thank Mir-Masoud Pourrahmat of Evidinno Outcomes Research Inc for medical writing and editorial assistance, which was funded by Sanofi. The authors would like to thank Edward Han-Burgess for study conceptualization and design; Adee Kennedy for managing the study; and Felix Lee for study conceptualization, design, results interpretation, and critical review of the manuscript (Sanofi). The authors would also like to thank Donald Nelinson (formerly of Sanofi) for coordinating the development, facilitating author discussions, and critical review of this manuscript. Publication tracking by Evidence (part of the Envision Pharma Group) was funded by Sanofi.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors fulfilled authorship requirements according to ICMJE guidelines. All authors contributed to extensive critical revision and interpretation and approved the final draft for submission.

Conflicts of Interest

DK is a consultant to Sanofi and Better Therapeutics. DA has received speaker's fees from Abbott, Ascensia, Eli Lilly, Mannkind, Novo Nordisk, and Xeris Pharmaceuticals. He has received consulting or advisory fees from Eli Lilly, Ascensia, and Mannkind. KW and JW report no conflict of interest. BB reports a contract with Evidinno Outcomes Research Inc. DCK is a consultant to Atropos Health, Better Therapeutics, Eoflow, Integrity, Lifecare, Nevro, Novo, Sanofi, and Thirdwayv.

Multimedia Appendix 1
PICO (Population, Intervention, Comparator, and Outcomes) criteria used for study selection.
[DOCX File , 32 KB - jmir_v26i1e55757_app1.docx]

Multimedia Appendix 2
Search strategy for Embase via OvidSP.
[DOCX File , 42 KB - jmir_v26i1e55757_app2.docx]

Multimedia Appendix 3
Search strategy for MEDLINE via OvidSP.
[DOCX File , 33 KB - jmir_v26i1e55757_app3.docx]

Multimedia Appendix 4
Search strategy for CENTRAL via OvidSP.
[DOCX File , 33 KB - jmir_v26i1e55757_app4.docx]

Multimedia Appendix 5
Distribution of the mean baseline HbA_{1c} levels across included studies reporting. HbA_{1c}: glycated hemoglobin A_{1c}.
[PNG File , 50 KB - jmir_v26i1e55757_app5.png]

Multimedia Appendix 6
Categorization of the studies into the intervention categories.
[DOCX File , 97 KB - jmir_v26i1e55757_app6.docx]

Multimedia Appendix 7
Risk of bias as percentage across randomized controlled trials (intention to treat).
[PNG File , 102 KB - jmir_v26i1e55757_app7.png]

Multimedia Appendix 8
Risk of bias in nonrandomized studies.
[PNG File , 139 KB - jmir_v26i1e55757_app8.png]

XSL•FO
RenderX

References

1. Silberman J, Wicks P, Patel S, Sarlati S, Park S, Korolev IO, et al. Rigorous and rapid evidence assessment in digital health with the evidence DEFINED framework. NPJ Digit Med 2023;6(1):101 [FREE Full text] [doi: 10.1038/s41746-023-00836-5] [Medline: 37258851]
2. Fontecha J, González I, Barragán A, Lim T. Use and trends of diabetes self-management technologies: a correlation-based study. J Diabetes Res 2022;2022:5962001 [FREE Full text] [doi: 10.1155/2022/5962001] [Medline: 35712029]
3. Allegrante JP, Wells MT, Peterson JC. Interventions to support behavioral self-management of chronic diseases. Annu Rev Public Health 2019;40:127-146 [FREE Full text] [doi: 10.1146/annurev-publhealth-040218-044008] [Medline: 30601717]
4. Pal K, Dack C, Ross J, Michie S, May C, Stevenson F, et al. Digital health interventions for adults with type 2 diabetes: qualitative study of patient perspectives on diabetes self-management education and support. J Med Internet Res 2018;20(2):e40 [FREE Full text] [doi: 10.2196/jmir.8439] [Medline: 29463488]
5. Jarl F, Davelid A, Hedin K, Stomby A, Petersson C. Overcoming the struggle of living with type 2 diabetes – diabetes specialist nurses' and patients' perspectives on digital interventions. BMC Health Serv Res 2023;23(1):313 [FREE Full text] [doi: 10.1186/s12913-023-09277-y] [Medline: 36998038]
6. Bults M, van Leersum CM, Olthuis TJJ, Bekhuis REM, den Ouden MEM. Mobile health apps for the control and self-management of type 2 diabetes mellitus: qualitative study on users' acceptability and acceptance. JMIR Diabetes 2023;8:e41076 [FREE Full text] [doi: 10.2196/41076] [Medline: 36692927]
7. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 7. Diabetes technology: standards of care in diabetes–2023. Diabetes Care 2023;46:S111-S127 [FREE Full text] [doi: 10.2337/dc23-S007] [Medline: 36507635]
8. Lewinski AA, Walsh C, Rushton S, Soliman D, Carlson SM, Luedke MW, et al. Telehealth for the longitudinal management of chronic conditions: systematic review. J Med Internet Res 2022;24(8):e37100 [FREE Full text] [doi: 10.2196/37100] [Medline: 36018711]
9. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 6. Glycemic targets: standards of care in diabetes–2023. Diabetes Care 2023;46:S97-S110 [FREE Full text] [doi: 10.2337/dc23-S006] [Medline: 36507646]
10. Bin Rakhis SA, AlDuwayhis NM, Aleid N, AlBarrak AN, Aloraini AA. Glycemic control for type 2 diabetes mellitus patients: a systematic review. Cureus 2022;14(6):e26180 [FREE Full text] [doi: 10.7759/cureus.26180] [Medline: 35891859]
11. Eberle C, Löhnert M, Stichling S. Effectiveness of disease-specific mhealth apps in patients with diabetes mellitus: scoping review. JMIR Mhealth Uhealth 2021 15;9(2):e23477 [FREE Full text] [doi: 10.2196/23477] [Medline: 33587045]
12. Eberle C, Stichling S. Effect of telemetric interventions on glycated hemoglobin a1c and management of type 2 diabetes mellitus: systematic meta-review. J Med Internet Res 2021;23(2):e23252 [FREE Full text] [doi: 10.2196/23252] [Medline: 33595447]
13. Kerr D, Messing R, Resch A. Actionable self-monitoring of blood glucose: redefining the role for patients using multiple daily injection therapy. J Diabetes Sci Technol 2011;5(6):1584-1590 [FREE Full text] [doi: 10.1177/193229681100500635] [Medline: 22226281]
14. Fu HNC, Wyman JF, Peden-McAlpine CJ, Draucker CB, Schleyer T, Adam TJ. App design features important for diabetes self-management as determined by the self-determination theory on motivation: content analysis of survey responses from adults requiring insulin therapy. JMIR Diabetes 2023;8:e38592 [FREE Full text] [doi: 10.2196/38592] [Medline: 36826987]
15. Fundoiano-Hershcovitz Y, Hirsch A, Dar S, Feniger E, Goldstein P. Role of digital engagement in diabetes care beyond measurement: retrospective cohort study. JMIR Diabetes 2021;6(1):e24030 [FREE Full text] [doi: 10.2196/24030] [Medline: 33599618]
16. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes–2023. Diabetes Care 2023;46:S68-S96 [FREE Full text] [doi: 10.2337/dc23-S005] [Medline: 36507648]
17. Hessler D, Strycker L, Fisher L. Reductions in management distress following a randomized distress intervention are associated with improved diabetes behavioral and glycemic outcomes over time. Diabetes Care 2021;44(7):1472-1479 [FREE Full text] [doi: 10.2337/dc20-2724] [Medline: 33990376]
18. Higgins JT, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester: John Wiley & Sons; 2019.
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Azelton KR, Crowley AP, Vence N, Underwood K, Morris G, Kelly J, et al. Digital health coaching for type 2 diabetes: randomized controlled trial of healthy at home. Front Digit Health 2021;3:764735 [FREE Full text] [doi: 10.3389/fdgth.2021.764735] [Medline: 34901926]
21. Turnin M, Gourdy P, Martini J, Buisson J, Chauchard M, Delaunay J, et al. Impact of a remote monitoring programme including lifestyle education software in type 2 diabetes: results of the Educ@dom randomised multicentre study. Diabetes Ther 2021;12(7):2059-2075 [FREE Full text] [doi: 10.1007/s13300-021-01095-x] [Medline: 34160791]
22. Wayne N, Perez DF, Kaplan DM, Ritvo P. Health coaching reduces HbA1c in type 2 diabetic patients from a lower-socioeconomic status community: a randomized controlled trial. J Med Internet Res 2015;17(10):e224 [FREE Full text] [doi: 10.2196/jmir.4871] [Medline: 26441467]

XSL•FO
RenderX

23. Kim HS, Oh JA. Adherence to diabetes control recommendations: impact of nurse telephone calls. J Adv Nurs 2003;44(3):256-261. [doi: 10.1046/j.1365-2648.2003.02800.x] [Medline: 14641395]
24. Odnoletkova I, Goderis G, Nobels F, Fieuws S, Aertgeerts B, Annemans L, et al. Optimizing diabetes control in people with type 2 diabetes through nurse-led telecoaching. Diabet Med 2016;33(6):777-785. [doi: 10.1111/dme.13092] [Medline: 26872105]
25. Cox DJ, Oser T, Moncrief M, Conaway M, McCall A. Long-term follow-up of a randomized clinical trial comparing glycemic excursion minimization (GEM) to weight loss (WL) in the management of type 2 diabetes. BMJ Open Diabetes Res Care 2021;9(2):e002403 [FREE Full text] [doi: 10.1136/bmjdrc-2021-002403] [Medline: 34845062]
26. Guo M, Meng F, Guo Q, Bai T, Hong Y, Song F, et al. Effectiveness of mHealth management with an implantable glucose sensor and a mobile application among Chinese adults with type 2 diabetes. J Telemed Telecare 2023;29(8):632-640. [doi: 10.1177/1357633X211020261] [Medline: 34152238]
27. Parsons SN, Luzio SD, Harvey JN, Bain SC, Cheung WY, Watkins A, et al. Effect of structured self-monitoring of blood glucose, with and without additional TeleCare support, on overall glycaemic control in non-insulin treated type 2 diabetes: the SMBG study, a 12-month randomized controlled trial. Diabet Med 2019;36(5):578-590 [FREE Full text] [doi: 10.1111/dme.13899] [Medline: 30653704]
28. Ji H, Chen R, Huang Y, Li W, Shi C, Zhou J. Effect of simulation education and case management on glycemic control in type 2 diabetes. Diabetes Metab Res Rev 2019;35(3):e3112 [FREE Full text] [doi: 10.1002/dmrr.3112] [Medline: 30520255]
29. Pimazoni-Netto A, Rodbard D, Zanella MT, Diabetes EducationControl Group. Rapid improvement of glycemic control in type 2 diabetes using weekly intensive multifactorial interventions: structured glucose monitoring, patient education, and adjustment of therapy–a randomized controlled trial. Diabetes Technol Ther 2011;13(10):997-1004 [FREE Full text] [doi: 10.1089/dia.2011.0054] [Medline: 21751888]
30. Lee DY, Park J, Choi D, Ahn H, Park S, Park C. The effectiveness, reproducibility, and durability of tailored mobile coaching on diabetes management in policyholders: a randomized, controlled, open-label study. Sci Rep 2018;8(1):3642 [FREE Full text] [doi: 10.1038/s41598-018-22034-0] [Medline: 29483559]
31. Athinarayanan SJ, Adams RN, Hallberg SJ, McKenzie AL, Bhanpuri NH, Campbell WW, et al. Long-term effects of a novel continuous remote care intervention including nutritional ketosis for the management of type 2 diabetes: a 2-year non-randomized clinical trial. Front Endocrinol 2019;10:348 [FREE Full text] [doi: 10.3389/fendo.2019.00348] [Medline: 31231311]
32. Welch G, Zagarins SE, Santiago-Kelly P, Rodriguez Z, Bursell SE, Rosal MC, et al. An internet-based diabetes management platform improves team care and outcomes in an urban Latino population. Diabetes Care 2015;38(4):561-567 [FREE Full text] [doi: 10.2337/dc14-1412] [Medline: 25633661]
33. Yang Y, Lee EY, Kim HS, Lee SH, Yoon KH, Cho JH. Effect of a mobile phone–based glucose-monitoring and feedback system for type 2 diabetes management in multiple primary care clinic settings: cluster randomized controlled trial. JMIR Mhealth Uhealth 2020;8(2):e16266 [FREE Full text] [doi: 10.2196/16266] [Medline: 32130172]
34. Kirk A, Barnett J, Leese G, Mutrie N. A randomized trial investigating the 12-month changes in physical activity and health outcomes following a physical activity consultation delivered by a person or in written form in type 2 diabetes: Time2Act. Diabet Med 2009;26(3):293-301. [doi: 10.1111/j.1464-5491.2009.02675.x] [Medline: 19317825]
35. Layne JE, Bergenstal RM, Barleen NA, Dixon RF, Zisser H. 597-P: long-term A1C outcomes with and without intermittent CGM use in adults with T2D participating in the Onduo program. Diabetes 2021;70:597. [doi: 10.2337/db21-597-p]
36. Allen N, Whittemore R, Melkus G. A continuous glucose monitoring and problem-solving intervention to change physical activity behavior in women with type 2 diabetes: a pilot study. Diabetes Technol Ther 2011;13(11):1091-1099. [doi: 10.1089/dia.2011.0088] [Medline: 21919735]
37. Jeong JY, Jeon JH, Bae KH, Choi YK, Park KG, Kim JG, et al. Smart care based on telemonitoring and telemedicine for type 2 diabetes care: multi-center randomized controlled trial. Telemed J E Health 2018;24(8):604-613. [doi: 10.1089/tmj.2017.0203] [Medline: 29341843]
38. Lee JY, Chan CKY, Chua SS, Ng CJ, Paraidathathu T, Lee KKC, et al. Telemonitoring and team-based management of glycemic control on people with type 2 diabetes: a cluster-randomized controlled trial. J Gen Intern Med 2020;35(1):87-94 [FREE Full text] [doi: 10.1007/s11606-019-05316-9] [Medline: 31512187]
39. Hsu WC, Lau KHK, Huang R, Ghiloni S, Le H, Gilroy S, et al. Utilization of a cloud-based diabetes management program for insulin initiation and titration enables collaborative decision making between healthcare providers and patients. Diabetes Technol Ther 2016;18(2):59-67 [FREE Full text] [doi: 10.1089/dia.2015.0160] [Medline: 26645932]
40. Hee-Sung K. Impact of web-based nurse's education on glycosylated haemoglobin in type 2 diabetic patients. J Clin Nurs 2007;16(7):1361-1366. [doi: 10.1111/j.1365-2702.2007.01506.x] [Medline: 17584355]
41. Amante DJ, Harlan DM, Lemon SC, McManus DD, Olaitan OO, Pagoto SL, et al. Evaluation of a diabetes remote monitoring program facilitated by connected glucose meters for patients with poorly controlled type 2 diabetes: randomized crossover trial. JMIR Diabetes 2021;6(1):e25574 [FREE Full text] [doi: 10.2196/25574] [Medline: 33704077]
42. Nagrebetsky A, Larsen M, Craven A, Turner J, McRobert N, Murray E, et al. Stepwise self-titration of oral glucose-lowering medication using a mobile telephone-based telehealth platform in type 2 diabetes: a feasibility trial in primary care. J Diabetes Sci Technol 2013;7(1):123-134 [FREE Full text] [doi: 10.1177/193229681300700115] [Medline: 23439168]

XSL•FO
RenderX

43. Lee SK, Shin DH, Kim YH, Lee KS. Effect of diabetes education through pattern management on self-care and self-efficacy in patients with type 2 diabetes. Int J Environ Res Public Health 2019;16(18):3323 [FREE Full text] [doi: 10.3390/ijerph16183323] [Medline: 31505892]
44. Das S, Manikandan R, Boochandran T, Ahmed A, Dwarakanath C, Jaganmohan B. Comprehensive Apollo sugar diabetes management programs: adherence for better clinical outcomes. Indian J Endocrinol Metab 2018;22 [FREE Full text]
45. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. Diabetes Care 2011;34(9):1934-1942 [FREE Full text] [doi: 10.2337/dc11-0366] [Medline: 21788632]
46. McKenzie AL, Hallberg SJ, Creighton BC, Volk BM, Link TM, Abner MK, et al. A novel intervention including individualized nutritional recommendations reduces hemoglobin A1c level, medication use, and weight in type 2 diabetes. JMIR Diabetes 2017;2(1):e5 [FREE Full text] [doi: 10.2196/diabetes.6981] [Medline: 30291062]
47. Fantasia KL, Stockman MC, Ju Z, Ortega P, Crable EL, Drainoni ML, et al. Professional continuous glucose monitoring and endocrinology econsult for adults with type 2 diabetes in primary care: results of a clinical pilot program. J Clin Transl Endocrinol 2021;24:100254 [FREE Full text] [doi: 10.1016/j.jcte.2021.100254] [Medline: 33898271]
48. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
49. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
50. Greenwood DA, Litchman ML, Isaacs D, Blanchette JE, Dickinson JK, Hughes A, et al. A new taxonomy for technology-enabled diabetes self-management interventions: results of an umbrella review. J Diabetes Sci Technol 2022;16(4):812-824 [FREE Full text] [doi: 10.1177/19322968211036430] [Medline: 34378424]
51. Mao Y, Lin W, Wen J, Chen G. Impact and efficacy of mobile health intervention in the management of diabetes and hypertension: a systematic review and meta-analysis. BMJ Open Diabetes Res Care 2020;8(1):e001225 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001225] [Medline: 32988849]
52. Marzban S, Najafi M, Agolli A, Ashrafi E. Impact of patient engagement on healthcare quality: a scoping review. J Patient Exp 2022;9:23743735221125439 [FREE Full text] [doi: 10.1177/23743735221125439] [Medline: 36134145]
53. Klonoff DC, Gutierrez A, Fleming A, Kerr D. Real-world evidence should be used in regulatory decisions about new pharmaceutical and medical device products for diabetes. J Diabetes Sci Technol 2019;13(6):995-1000 [FREE Full text] [doi: 10.1177/1932296819839996] [Medline: 30943790]

Abbreviations

CGM: continuous glucose monitoring
DTSQ: Diabetes Treatment Satisfaction Questionnaire
HbA_{1c}: glycated hemoglobin A_{1c}
HCP: health care professional
MD: mean difference
PICO: Population, Intervention, Comparator, and Outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SLR: systematic literature review
T2DM: type 2 diabetes mellitus

Edited by T de Azevedo Cardoso; submitted 22.12.23; peer-reviewed by R Valdes Ramos, T Mao; comments to author 30.04.24; revised version received 13.05.24; accepted 16.05.24; published 22.07.24.

Please cite as:
Kerr D, Ahn D, Waki K, Wang J, Breznen B, Klonoff DC
Digital Interventions for Self-Management of Type 2 Diabetes Mellitus: Systematic Literature Review and Meta-Analysis
J Med Internet Res 2024;26:e55757
URL: https://www.jmir.org/2024/1/e55757
doi:10.2196/55757
PMID:39037772

©David Kerr, David Ahn, Kayo Waki, Jing Wang, Boris Breznen, David C Klonoff. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.07.2024. This is an open-access article distributed under the terms of the

XSL•FO
RenderX

Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review

Moustafa Laymouna[1,2,3], MD, MSc; Yuanchao Ma[2,3,4,5], MScA; David Lessard[2,3,4], PhD; Tibor Schuster[1], PhD; Kim Engler[2,3,4], PhD; Bertrand Lebouché[1,2,3,4], MD, PhD

[1]Department of Family Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[2]Centre for Outcomes Research and Evaluation, Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[3]Infectious Diseases and Immunity in Global Health Program, Research Institute of McGill University Health Centre, Montreal, QC, Canada
[4]Chronic and Viral Illness Service, Division of Infectious Disease, Department of Medicine, McGill University Health Centre, Montreal, QC, Canada
[5]Department of Biomedical Engineering, Polytechnique Montréal, Montreal, QC, Canada

Corresponding Author:
Bertrand Lebouché, MD, PhD
Centre for Outcomes Research and Evaluation
Research Institute of the McGill University Health Centre
D02.4110 – Glen Site, 1001 Decarie Blvd
Montreal, QC, H4A 3J1
Canada
Phone: 1 514 843 2090
Email: bertrand.lebouche@mcgill.ca

Abstract

Background: Chatbots, or *conversational agents*, have emerged as significant tools in health care, driven by advancements in artificial intelligence and digital technology. These programs are designed to simulate human conversations, addressing various health care needs. However, no comprehensive synthesis of health care chatbots' roles, users, benefits, and limitations is available to inform future research and application in the field.

Objective: This review aims to describe health care chatbots' characteristics, focusing on their diverse roles in the health care pathway, user groups, benefits, and limitations.

Methods: A rapid review of published literature from 2017 to 2023 was performed with a search strategy developed in collaboration with a health sciences librarian and implemented in the MEDLINE and Embase databases. Primary research studies reporting on chatbot roles or benefits in health care were included. Two reviewers dual-screened the search results. Extracted data on chatbot roles, users, benefits, and limitations were subjected to content analysis.

Results: The review categorized chatbot roles into 2 themes: *delivery of remote health services*, including patient support, care management, education, skills building, and health behavior promotion, and *provision of administrative assistance to health care providers*. User groups spanned across patients with chronic conditions as well as patients with cancer; individuals focused on lifestyle improvements; and various demographic groups such as women, families, and older adults. Professionals and students in health care also emerged as significant users, alongside groups seeking mental health support, behavioral change, and educational enhancement. The benefits of health care chatbots were also classified into 2 themes: *improvement of health care quality* and *efficiency and cost-effectiveness in health care delivery*. The identified limitations encompassed ethical challenges, medicolegal and safety concerns, technical difficulties, user experience issues, and societal and economic impacts.

Conclusions: Health care chatbots offer a wide spectrum of applications, potentially impacting various aspects of health care. While they are promising tools for improving health care efficiency and quality, their integration into the health care system must be approached with consideration of their limitations to ensure optimal, safe, and equitable use.

(J Med Internet Res 2024;26:e56930) doi:10.2196/56930

KEYWORDS

chatbot; conversational agent; conversational assistant; user-computer interface; digital health; mobile health; electronic health; telehealth; artificial intelligence; AI; health information technology

XSL•FO
RenderX

Introduction

Background

In the dynamic landscape of IT and digital communication, chatbots—known as *conversational agents*—stand at the forefront, revolutionizing interactions between technology and human users. Chatbots are computer programs designed to simulate conversation through text, image, audio, or video messaging with human users on platforms such as websites, smartphone apps, or stand-alone computer software [1-47]. Originating from the concept *ChatterBot*, coined in 1994 [48], chatbots have undergone substantial evolution in their functionality and application.

The evolution of chatbots represents a significant technological leap, transitioning from reliance on predefined, rule-based scripted conversations to the sophisticated use of natural language processing and artificial intelligence (AI). By leveraging natural language processing and AI, chatbots have become capable of understanding and appropriately responding to user requests [49,50]. Their versatility has facilitated applications in a variety of sectors such as education, e-commerce, finance, news, health care, and entertainment. Popular instances of these applications include Amazon's Alexa [51], Apple's Siri [52], Google Assistant [53], Microsoft's Cortana [54], and Samsung's Bixby [55].

A notable advancement in the field of chatbots has been the integration of generative AI and large language models (LLMs) such as ChatGPT [56-58]. They have the capability to generate human-like text, enabling more natural and informative interactions [56-58]. However, their application in health care is still emerging. The risk of misinformation and errors is a significant concern [59,60], particularly in health care where accuracy is critical. The *one-size-fits-all* approach of LLMs may not align well with the nuanced needs of patient-centered care in the health sector [59].

The promise of chatbots in health care is considerable, offering potential for more efficient, cost-effective, and high-quality care [61-65], as well as their broad spectrum of uses and acceptability [66,67]. The use of chatbots to access and deliver health care services seems to be on the rise [23,68-70], granting them multiple potential roles in prevention, diagnosis, and support with care and treatment, with possible impacts on the whole health care system.

Despite the potential benefits, health care chatbots face unique challenges [71-74]. The need for highly specialized and context-sensitive advice is paramount. Generic responses from current chatbot models often overlook individual health profiles and local health contexts, which are crucial for patient care [75].

While a wide range of health care chatbot reviews have been conducted—demonstrating the versatility of chatbots in areas such as genetic cancer risk assessment [44]; oncological care [9,11,24,25]; sexual and reproductive health [35,45]; preconception, pregnancy, and postpartum health [36]; support for smoking cessation [38]; management of weight [39] and chronic conditions [6,9,20,40]; vaccine communication [26]; and broader health care acceptability [27]—these reviews often exhibit significant limitations in scope and depth. They tend to concentrate narrowly on specific applications such as rehabilitation for neurological conditions [28], mental health support [4,8,12-17,29,30,41,42], health behavior change [31-33,37], the language used in health communication by chatbots [43], and the use of chatbots in the COVID-19 public health response [44], leading to a fragmented understanding of chatbots' roles in health care; for instance, while some reviews [3,7] offer insights, they do not encompass a comprehensive evaluation of the broader implications of chatbots, particularly in diverse contexts. By contrast, other reviews [5,30] concentrate extensively on technical aspects and AI algorithms [24,25,75,76]; yet, this focus tends to overshadow a detailed exploration of the impact these technologies have on health care outcomes.

Objectives

This approach has left significant gaps in the literature. There is an evident need for an integrative overview that thoroughly analyzes the varied roles of chatbots across different health care applications, capturing new trends and advancements. Furthermore, the interactions and benefits of health care chatbots for diverse demographic groups, especially those who are underrepresented, are underexplored. There is also a conspicuous absence of a deeper understanding of the potential benefits and practical limitations of health care chatbots in various contexts.

Therefore, the objectives of this review are to bridge these existing knowledge gaps. Our review aims to provide a comprehensive exploration of chatbots' functional roles, analyze the specific populations they serve, and examine in detail their potential and reported benefits, as well as the limitations of these innovative tools in health care. This endeavor will offer a more holistic and nuanced understanding of chatbots in the health care sector, addressing critical areas overlooked in previous studies.

Methods

Design and Search Strategy

This study is a rapid review, which refers to an accelerated, resource-efficient process of knowledge synthesis through streamlining or omitting specific methods associated with more traditional review processes [77-79]. Hence, a rapid review assesses what is already known in a given area within a relatively short period.

Our search strategy, detailed in Textbox 1, was developed in collaboration with a health sciences librarian and performed within the MEDLINE and Embase databases on February 5, 2022. Recognizing the dynamic nature of our study field, we conducted 2 subsequent updates to our search: the first on April 22, 2022, and the second on October 30, 2023. The strategy also included searches within reference lists and websites (eg, Google Scholar) for relevant material. We exported our search records to EndNote (Clarivate).

Our search was limited to records published in English, as suggested by the Cochrane rapid reviews guide [80], from 2017 to 2023. This time frame was chosen based on preliminary searches that indicated that the largest number of relevant

XSL•FO
RenderX

articles was published during this period [81]. Furthermore, it allowed us to focus on chatbots incorporating more recent technological advancements. No limitations were set based on the study population.

Our rapid review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, as depicted in Figure 1 [82].

Textbox 1. Search strategy for MEDLINE and Embase.

Search terms

- user-computer interface/or (Chatbot* or chat bot* or User-Computer Interface* or (conversational adj2 (agent* or assistant*)).mp
- Limit 1 to yr = "2017 - Current"
- Limit 2 to English

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the number of studies identified, screened, assessed for eligibility, and included in the final analysis.

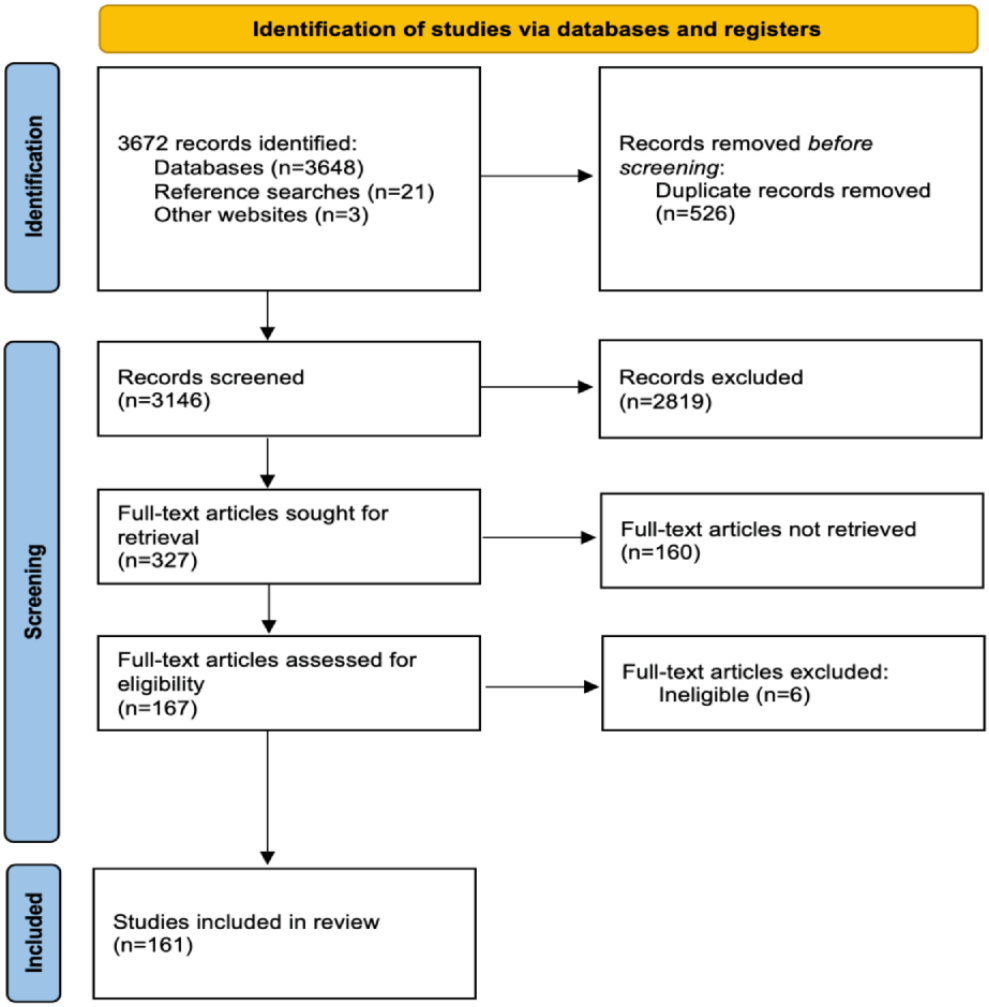

Study Selection

We included primary research studies that used text- or voice-based tailored chatbots as interventions within the health care system or as a means to deliver interventions. These studies report original data on the roles and benefits of chatbots in the health care setting.

Studies not meeting the inclusion criteria were excluded, as were studies reporting any of the following: engineering or computer science data, preintervention data about future initiatives such as protocols, and studies in the preintervention or predevelopment phase. We also excluded interventions based solely on nonbehavioral actions such as gestures and facial expressions without text or voice interaction, interactions with an actual robot (as opposed to a conversational interface), and virtual reality chatbots. In addition, abstracts lacking sufficient details were excluded.

Data Extraction and Synthesis

Two reviewers (ML and YM) dual-screened 15% of the titles and abstracts and full texts to calculate the percentage agreement and interrater reliability, using Cohen κ [83]. Any discrepancies

XSL•FO
RenderX

were resolved through discussion. ML conducted all remaining screenings. Data extraction was performed using Microsoft Office 365 (Excel and Word), capturing key study characteristics, including title, authors, month and year of publication, journal, study design, chatbot users, the chatbot's medical specialty, whether the chatbot uses AI or is animated, and country of origin. In addition, we extracted information about the roles of chatbots, their benefits to health care, and their limitations. We categorized the source data into empirical and indicative data. This distinction reflects the 2-fold impact of contributions to the field: the actual findings demonstrate concrete evidence about the roles, users, benefits, and limitations of existing chatbots, while the authors' discussion extends the conversation beyond current applications, providing perspective on the potential impacts, challenges, and future directions of health care chatbots, thus more comprehensively rounding out our assessment.

To synthesize these diverse pieces of information, relevant data underwent content analysis to generate subcategories, categories, and overarching themes [84].

While our research centers on chatbots, we have chosen to use the number of studies, rather than the chatbots themselves, as the basis for presenting most of our results. This approach accounts for the diverse adaptations to the identified chatbots across different contexts. Many of the chatbots we studied were modified to serve varied roles; cater to different user groups; and, in some cases, were given entirely different names in separate studies, as indicated in the *Results* section. Importantly, we noticed that a given study could contribute to multiple categories, indicating the flexible and interconnected characteristics of chatbot roles, users, benefits, or limitations. By focusing on the individual studies, we capture a more detailed and context-specific understanding of each chatbot's functionality and versatility, which would be obscured if we merely counted each chatbot once, regardless of its various adaptations.

Results

Database Searches

Our search yielded 3672 records (databases: n=3146, 85.68%; reference searches: n=21, 0.57%; and other websites: n=3, 0.08%). After removing 526 (14.32%) duplicates from the 3672 records, 3122 (85.02%) records remained for title and abstract screening. During this screening phase, we achieved a 97% agreement rate and a Cohen κ value of 0.85, indicating substantial agreement beyond chance. Subsequently, of the 3146 records, 327 (10.39%) full texts were reviewed [85-245] (Figure 1), with 94% agreement and a Cohen κ value of 0.88 among the reviewers. Interrater reliability between the 2 reviewers, covering both the screening and final study inclusion as well as the data extraction process, ranged from 64% to 81%, indicating strong agreement [83]. This ensures the reliability and validity of the study selection and data extraction phases of our review.

After reviewing the 327 full texts, we ultimately included 161 (49.2%) studies that reported the roles and benefits of chatbots. All 161 studies reported on the roles of chatbots, 157 (97.5%) mentioned their benefits, and 157 (97.5%) addressed their limitations. Each study also reported on the user group or groups of focus that the chatbot was designed to assist.

Origins of the Included Studies

More than a quarter of the studies originated from the United States (46/161, 28.6%; Figure 2). China (15/161, 9.3%), Australia (10/161, 6.2%), Japan (9/161, 5.6%), and Spain (7/161, 4.3%) followed. Of the 161 studies, Italy, Switzerland, the United Kingdom, Singapore, Brazil, and South Korea each contributed 6 (3.7%), France and the Netherlands each contributed 4 (2.5%), while New Zealand, Greece, Russia, Norway, Malaysia, India, Senegal, Peru, Portugal, Canada, Latvia, South Africa, Indonesia, Argentina, Thailand, Saudi Arabia, Germany, and Austria each contributed 1 (0.6%) study. Notably, some studies were multinational; for instance, 1 (0.6%) of the 161 studies included Switzerland, Austria, and Germany; another included Northern Ireland, the Republic of Ireland, Scotland, Sweden, and Finland; yet another included Thailand, China, and Singapore; another study included India, North America, and the United Kingdom; a study included Finland, Denmark, and the Netherlands; another included Norway and Switzerland; and an additional study included the Netherlands and Scotland. Collectively, these 7 multinational studies account for 4.3% of the 161 included studies.

In our review of 161 studies, certain chatbots were the focus of multiple studies, particularly in the United States, Australia, South Korea, Switzerland, New Zealand, and Singapore; for instance, 2 specific chatbots were each the subject of 4 (2.5%) of the 161 studies (Gabby [94,99,101,114] and Woebot [86,92,119,173]). In addition, 11 chatbots were each studied twice (Todaki [90,102], GAMBOT [97,133], Laura [98,121], Vik [108,186], Termbot [151,195], ChatPal [158,168], a chatbot in a virtual ward [180,194], Corowa-kun [181,197], Dokbot [189,192], BotMaria [193,205], and COUCH [236,237]). Among these, a unique situation was observed in 5 (3.1%) of the 161 studies where the same original chatbot was presented under 5 different names [89,104,107,124,240]. These studies often shared several coauthors, indicating a common origin but with adaptations for different populations and roles. However, it is important to note that not all studies with mutual coauthors clearly indicated a shared origin of the chatbots.

XSL•FO
RenderX

Figure 2. Map showing the countries that contributed the studies.

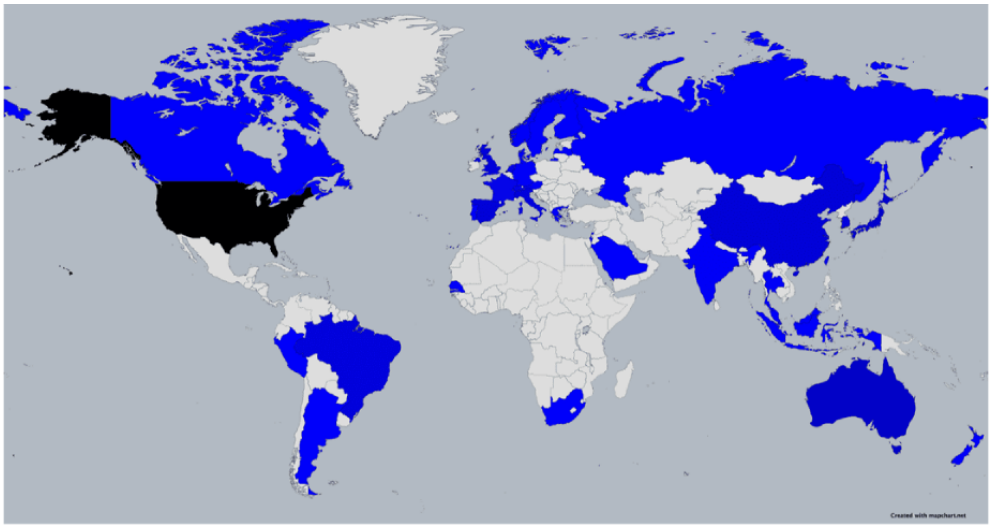

Chatbot Roles

Overview

All studies stated the role or roles of the chatbot used, with at least 1 role per study. Our analysis yielded 14 subcategories of primary roles (presented in italics), grouped into 5 categories, which were organized into 2 overarching themes, as summarized in Table 1.

Table 1. Health care chatbot roles (n=161).

Theme, category, and subcategory	Studies, n (%)
Delivery of remote health services	
Patient support and care management	
Mental health support	46 (28.6)
Counseling and treatment advice	26 (16.1)
Self-management and monitoring for chronic conditions	22 (13.7)
Triaging, screening, risk assessment, and referral	14 (8.7)
Self-care and monitoring for COVID-19 symptoms	8 (5)
Rehabilitation guidance	8 (5)
Reminders	7 (4.3)
Education and skills building	
Health literacy	23 (14.3)
Medical education and clinical skills for health care professionals and medical students	12 (7.5)
Psychoeducation	5 (3)
Health behavior promotion	
Healthy lifestyle behavior	30 (18.6)
Self-monitoring for health behavior change	6 (3.7)
Provision of administrative assistance to health care providers	
Health-related administrative tasks	
Data collection and storage in patient electronic medical records	6 (3.7)
Research purposes	
Recruitment and data collection	3 (1.9)

XSL•FO
RenderX

Theme 1: Delivery of Remote Health Services

This theme refers to health services offered at a distance as an alternative or complement to the usual on-site modes of care delivery. It includes 3 categories and 7 subcategories of roles, with 158 (98.1%) of the 161 studies contributing to this theme.

Patient Support and Care Management

This category refers to the facilitation of medical consultations or the delivery of advice or support by providing counseling or treatment advice, triaging patients' complaints, and fostering self-management and monitoring.

Overall, 103 (65.2%) of the 158 studies contributed to this category. Of these 103 studies, 46 (44.7%) mentioned using chatbots for *mental health support*, 26 (25.2%) reported providing *counseling and treatment advice* through chatbots, while 22 (21.4%) included chatbot use for improving *self-management or monitoring for chronic conditions*. Furthermore, of the 103 studies, 14 (13.6%) described chatbot use for *triaging, screening, risk assessment, and referral*; 8 (7.8%) studies each reported chatbot use for *self-care and monitoring for COVID-19 symptoms* and *rehabilitation guidance*; whereas 7 (6.8%) studies used chatbots to provide *reminders*.

Education and Skills Building

This category included the dissemination of educational material or medical information or skills development material (eg, exercising and using a medical device) for users, including patients, health care providers, or nursing and medical students.

In all, 41 (25.9%) of the 158 studies contributed to this category. Of these 41 studies, 23 (56%) reported promoting *health literacy* of the targeted population with the chatbot, 12 (29%) reported using chatbots in *medical education and clinical skills for health care professionals and medical students*, and *psychoeducation* was reported by 5 (12%) studies to enhance mental well-being.

Health Behavior Promotion

This category included the promotion of healthy lifestyles such as physical activity, a healthy diet, or stress management. Of the 158 studies, 39 (24.7%) contributed to this category. Of these 39 studies, *healthy lifestyle behavior* was encouraged through the chatbot in 30 (77%), while 6 (15%) reported *self-monitoring for health behavior change* as a chatbot role.

Theme 2: Provision of Administrative Assistance to Health Care Providers

This theme refers to all types of administrative work carried out by the chatbots, grouped within 2 categories—health-related administrative tasks and research purposes—with 9 (5.6%) of the 161 studies contributing to this theme.

Health-Related Administrative Tasks

This category included the completion of health care providers' routine administrative work, such as data collection (eg, medical history taking), data entry, or transferring data to patients' medical records. Of the 9 studies, 6 (67%) reported using the chatbot for *data collection and storage in patient electronic medical records* and charts, as well as for patient-reported outcome data, which could be captured by chatbots to replace collection by health care providers.

Research Purposes

This category refers to chatbot use for the completion of research-related work such as participant recruitment, the consent process, or data collection through surveys. Of the 9 studies, 3 (33%) contributed to this category, reporting the use of chatbots for participants' *recruitment and data collection* through a self-administered questionnaire, in addition to obtaining electronic consent from individuals to participate in the study.

Chatbot Users

Overview

All 161 studies specified the intended chatbot user population. The content analysis yielded 21 subcategories of chatbot users (presented in italics), grouped into 8 broader categories of users, as summarized in Table 2.

XSL•FO
RenderX

Table 2. Intended health care chatbot users (n=161).

Category and subcategory	Studies, n (%)
Health condition–focused groups	
Individual seekers of mental health support	23 (14.3)
Patients with chronic conditions	10 (6.2)
Patients with cancer	7 (4.3)
Recovering patients	6 (3.7)
Lifestyle and general well-being enthusiasts	
Healthy adults	44 (27.3)
General public	16 (9.9)
Lifestyle-improvement seekers	9 (5.6)
Demographic and family-centric groups	
Women	14 (8.7)
Parents and children	7 (4.3)
Families	4 (2.5)
Age-based user groups	
Older adults	11 (6.8)
Young seekers of mental health support	8 (5)
Children	4 (2.5)
Underserved populations	
Culturally diverse groups	14 (8.7)
Individuals with disabilities	8 (5)
Health care professionals and students	
Medical and nursing students	8 (5)
Health care professionals	7 (4.3)
Health-related–behavior-change seekers	
Behavioral change seekers	8 (5)
Individuals in addiction recovery	7 (4.3)
Educational and skills enhancement seekers	
Nonmedical professionals	8 (5)
Health care training users	7 (4.3)

Lifestyle and General Well-Being Enthusiasts

This category, with 69 (42.9%) of the 161 studies, addressed individuals aiming to improve or maintain their health and well-being. Of these 69 studies, 44 (64%) focused on *healthy adults* (adults who are in good health, without any significant or chronic medical conditions). *General public* (16/69, 23%) targeted the broader and more inclusive population that encompasses all segments of the population, regardless of their health status. *Lifestyle-improvement seekers*, encompassing 9 (13%) of the 69 studies, included individuals motivated to change their lifestyle.

Health Condition–Focused Groups

This category, comprising 46 (28.6%) of the 161 studies, included patients with specific health conditions across 4 subcategories. Of these 46 studies, individuals seeking mental health support, the largest subcategory with 23 (50%) studies, referred to adults with conditions such as attention-deficit and panic symptoms. *Patients with chronic conditions* (10/46, 22%) focused on individuals with conditions such as irritable bowel syndrome and hypertension. *Patients with cancer* (7/46, 15%) targeted those with breast cancer and those at risk for hereditary cancer. *Recovering patients* (6/46, 13%) focused on patients in various stages of recovery.

Demographic and Family-Centric Groups

Addressing specific demographic groups and family dynamics, this category comprised 15.5% (25/161) of the included studies. *Women* (14/25, 56%) focused on women's health issues. *Parents and children* (7/25, 28%) centered on the health issues of children and adolescents. *Families* (4/25, 16%) looked at family dynamics and health.

XSL•FO
RenderX

Unlike age-based groups that are defined solely by the age of individuals, demographic and family-centric groups consider a wider range of factors, including gender, family roles, and the interplay of relationships within a family unit.

Age-Based User Groups

With 23 (14.3%) of the 161 studies, this category targeted specific age groups or life stages. *Older adults* (11/23, 48%) focused on older adults and age-related health concerns. Young seekers of mental health support (8/23, 35%) focused on mental health support for young adults. *Children* (4/23, 17%) targeted health issues specific to children.

Underserved Populations

With 22 (13.7%) of the 161 studies, this category focused on inclusive and accessible health care. *Culturally diverse groups* (14/22, 64%) targeted ethnic and cultural groups. *Individuals with disabilities* (8/22, 36%) focused on the unique health care needs of people with disabilities.

Health Care Professionals and Students

Encompassing 15 (9.3%) of the 161 studies, this category targeted health care professionals and students. *Medical and nursing students* (8/15, 53%) covered educational aspects for students in medical and nursing fields. Health care professionals (7/15, 47%) focused on training and professional development with this group.

Health-Related–Behavior-Change Seekers

Comprising 15 (9.3%) of the 161 studies, this category focused on behavioral health and lifestyle changes. *Behavioral change seekers* (8/15, 53%) included studies on individuals seeking to change health-related behaviors. *Individuals in addiction recovery* (7/15, 47%) targeted those dealing with addictions.

Educational and Skills Enhancement Seekers

Comprising 15 (9.3%) of the 161 studies, this category involved the use of chatbots for educational purposes. *Nonmedical professionals* (8/15, 53%) focused on skills enhancement for various professionals. *Health care training users* (7/15, 47%) were concerned about chatbots being used to train health care professionals.

While the *health care professionals* subcategory within the *health care professionals and students* category focuses on the professional development and training of individuals in the health care field, the *educational and skills enhancement seekers* category addresses a broader spectrum of users, including nonmedical professionals, and emphasizes the role of chatbots as a tool for educational purposes across various sectors.

Health Care Chatbot Benefits

Overview

Most of the studies (157/161, 97.5%) described the benefits of using chatbots in health care. The content analysis yielded 7 different subcategories of benefits (presented in italics), grouped into 5 categories, which were organized into 2 broad themes, as summarized in Table 3.

Table 3. Reported health care chatbot benefits (n=157).

Theme, category, and subcategory	Studies, n (%)
Improvement of health care quality	
Improvement in health outcomes and patient management	
Improved mental health and well-being	42 (26.8)
Enhanced self-management	15 (9.6)
Improved physical health	8 (5.1)
Promotion of patient-centered care and health equity	
Increased accessibility and reach of health care	60 (38.2)
Engaged and satisfied users	16 (10.2)
Supported groups considered vulnerable and reduced biases in health care delivery	4 (2.5)
Efficiency and cost-effectiveness in health care delivery	
Optimization of resources	
Saved time and cost of health interventions	75 (47.8)
Scalability of health interventions	4 (2.5)
Data quality and research support	4 (2.5)

Theme 1: Improvement of Health Care Quality

This theme refers to the processes of enhancing the standards, personalization, and accessibility of health care services delivered to the targeted chatbot users. It included 6 subcategories grouped into 2 categories of benefits, with 121 (77.1%) of the 157 studies contributing to the overarching theme.

Improvement in Health Outcomes and Patient Management

Of the 121 studies in this category, 65 (53.7%) addressed the benefits of chatbots to improve health outcomes and patient

XSL•FO
RenderX

management. Of these 65 studies, 42 (65%) reported on *improved mental health and well-being*, 15 (23%) reported on *enhanced self-management*, and 8 (12.3%) reported on *improved physical health* as outcomes of using chatbots.

Personalization Through Patient-Centered and Equitable Care

Of the 121 studies, 62 (51.2%) reported on promoting personalization through patient-centered and equitable care. Chatbot personalization refers to customizing its interactions, content, and functionalities to suit individual needs and preferences, ensuring that it provides relevant, user-specific advice and support, enhancing its effectiveness and user experience. Health equity refers to minimizing disparities and inequality based on the social determinants of health, including differences between groups in terms of socioeconomic factors, gender, and ethnicity [246]. Patient-centered care addresses patients' specific health care needs and concerns, improving the quality of personal, professional, and organizational relationships and aiding patients to actively participate in their own care [247,248].

Of the 62 studies, 60 (97%) discussed chatbot use benefits in terms of *increased accessibility and reach of health care* by helping engage diverse populations to access health services for minor health concerns that do not require emergency visits, with convenience and 24/7 availability.

Moreover, 16 (26%) of the 62 studies discussed using a chatbot to achieve *engaged and satisfied users*. In these studies, user acceptance was assessed by measuring the users' positive feedback and their willingness to use the chatbot. This was often gauged through surveys or user feedback sessions after the interaction. The studies also highlighted that friendly interactions facilitated by the chatbot could enhance self-disclosure, further contributing to user satisfaction and engagement.

Of the 62 studies, 4 (6%) described chatbot use benefits for *supported groups considered vulnerable and reduced biases in health care delivery*, particularly for groups considered marginalized (eg, Black women and older users) facing stigma in health care settings and for people with low technological literacy.

Theme 2: Efficiency and Cost-Effectiveness in Health Care Delivery

Overview

This theme refers to chatbot use as favoring efficient care for targeted users. Providing efficient care means producing desired results with minimal or no waste of time, costs, materials, or personnel [249]. Three categories of benefits contributed to this overarching theme.

Optimization of Resources

In all, 75 (47.8%) of the 157 studies indicated reduced administrative or financial burdens for the health care system through chatbots because they can help relieve the burden of managing chronic health conditions, staffing shortage, and overwhelmed primary care settings. These studies indicated that chatbots could provide *saved time and cost of health interventions*, especially compared to other routine interventions.

Scalability of Health Interventions

Of the 157 studies, 4 (2.5%) indicated the feasibility of using chatbots for the implementation of large-scale health interventions to capture and assess large-scale public health situations, providing evidence for researchers and policy makers. The studies also addressed the significance of user data collected during the COVID-19 pandemic to evaluate the public health situation and aid decision-making by policy makers, public health authorities, and researchers.

Data Quality and Research Support

Of the 157 studies, 4 (2.5%) pointed out the benefits of enhancing data collection and clinical research quality by chatbots, providing timely, consistent, and standardized data collection, reducing human error, increasing patient engagement, and assisting in recruiting a diverse participant pool.

Health Care Chatbot Limitations

Overview

Most of the studies (157/161, 97.5%) identified specific limitations of chatbots in health care, presented as 12 subcategories grouped into 5 categories, as summarized in Table 4.

XSL•FO
RenderX

Table 4. Reported health care chatbot limitations (n=157).

Category and subcategory	Studies, n (%)
Challenges in user experience and overreliance	
Overconfidence and overreliance	154 (98.1)
Usability and accessibility issues	135 (86)
Technical challenges	
Complexity of effective language and communication processing	24 (15.3)
Limitations in empathy and personal connection	17 (10.8)
Challenges with resource allocation and cost efficiency	2 (1.3)
Medicolegal and safety concerns	
Regulatory and legal issues	3 (1.9)
Concerns about content and information quality	2 (1.3)
Challenges in emergency response and expertise	2 (1.3)
Societal and economic challenges	
Social, economic, and political challenges	5 (3.2)
Issues of inequality in accessibility	4 (2.5)
Ethical challenges	
Privacy and confidentiality concerns	2 (1.3)
Ethical and safety concerns	2 (1.3)

Challenges in User Experience and Overreliance

A total of 157 (97.5%) of the 161 studies contributed to this category, addressing the tendency of *overconfidence and overreliance* among users who overestimate the capabilities of chatbots or rely excessively on them for health care needs, as noted in 154 (98.1%) studies. Overconfidence in chatbots can lead to users substituting professional medical advice with chatbot suggestions, while overreliance might result in users neglecting other essential aspects of health care or disregarding the need for human health care professional intervention. This subcategory highlights the importance of maintaining a balanced perspective on the capabilities and limitations of chatbots in health care contexts.

In addition, this category encompasses the *usability and accessibility issues* related to the ease with which users can interact with chatbots and the extent to which these chatbots are accessible to a diverse range of users, as referred to in most of the studies (135/157, 86%). It includes considerations of user interface design, the intuitiveness of chatbot interactions, the chatbots' adaptability to different user needs, and their accessibility to individuals with varying levels of technology savviness or disabilities. Challenges in this category can lead to user dissatisfaction, reduced effectiveness of the chatbot, and potentially lower engagement with the health care service it provides.

Technical Challenges

This category refers to the broad spectrum of technological difficulties encountered in the design, development, and implementation of these systems, with 32 (20.1%) of the 157 studies contributing to it. This category underscores the need for sophisticated technology that can handle the nuances of health care communication and patient interaction while being accessible and practical for real-world application.

It includes the *complexity of effective language and communication processing*, as noted in 24 (75%) of the 32 studies, to ensure accurate and relevant medical information, as well as the chatbot's ability to understand and respond to a range of user inputs, including those related to emotional states and complex health care queries.

The limitations extend to *challenges in empathy and personal connection*, which refer to the difficulties chatbots face in simulating human conversations and establishing rapport with users. This is a critical aspect in health care settings where patient trust and comfort are paramount, as highlighted in 17 (53%) of the 32 studies.

In addition, this category involves considering the *challenges with resource allocation and cost efficiency* of developing and maintaining these systems to ensure that they are not only technologically advanced but also financially viable and sustainable, as indicated in 2 (6%) of the 32 studies.

Medicolegal and Safety Concerns

With 6 (3.8%) of the 157 contributing studies, this category includes *regulatory and legal issues* encompassing the implications of chatbot advice and overall patient safety, as highlighted in 3 (50%) studies. These issues include chatbots' compliance with health care regulations and patient privacy laws, liability for misdiagnosis or inadequate advice, and the need for specific regulatory guidelines for their development and application.

XSL•FO
RenderX

Furthermore, challenges extend to *concerns about content and information quality*, such as the medical accuracy of information provided by chatbots (eg, the potential for misdiagnosis) and the reliability of medical content. It also concerns limitations tied to the chatbot's *challenges in emergency response and expertise* capabilities. Each of these subcategories was noted in 2 (33%) of the 6 studies.

Societal and Economic Challenges

This category refers to the wider implications of health care chatbots on the broader societal context and the economy, with 5 (3.2%) of the 157 contributing studies. It covers the influence of social, political, and economic factors on the adoption and effectiveness of chatbots in different communities.

It includes *social, economic, and political challenges* and considerations, as noted in all 5 studies. This subcategory scrutinizes the challenges arising from the integration of chatbots into the health care system, such as potential shifts in social norms, and the influence on economic policies and political decision-making in health care.

This category also includes *issues of inequality in accessibility*, as highlighted in 4 (80%) of the 5 studies. This subcategory delves into the challenges related to unequal access to chatbot technology. It focuses on how chatbots might inadvertently exacerbate existing disparities in health care, particularly for groups considered underprivileged, thereby highlighting the need for equitable distribution and accessibility of these technologies.

Ethical Challenges

This category deals with the ethical implications of using chatbots in health care, with 3 (1.9%) of the 157 studies contributing to it. It includes patient *privacy and confidentiality concerns* related to the use of patient data. This category also includes *ethical and safety concerns* encompassing the need to maintain transparency with users about the chatbot being a nonhuman agent and ensuring ethical standards in patient interactions. Each of these 2 subcategories was discussed in 2 (67%) of the 3 studies.

Discussion

Principal Findings

This rapid review revealed that chatbot roles in health care are diverse, ranging from patient support to administrative tasks, and they show great promise in improving health care accessibility, especially for groups considered marginalized. It also highlighted critical gaps in the literature, which are addressed in the following subsections.

Global Trends in Chatbot Research Indicate Its Predominance in Higher-Income Countries and Opportunities in Lower-Income Regions

With 35 countries represented by the studies in this review, the topic is clearly of global interest. However, more than a quarter of the included studies (46/161, 28.6%) originated from the United States, with the remainder conducted in high- or upper–middle-income countries across North America, Europe, and parts of Asia [250]. The concentration of chatbot research in high-income countries reflects underlying disparities with low- or lower–middle-income countries, particularly in parts of Africa, South America, and certain regions in Asia, in terms of technology access and health care investment. This gap highlights the need for more research focused on these regions, considering their unique digital infrastructure and resource challenges, to democratize health technology and address chronic conditions and health literacy [20,251-254].

Chatbots Have Varied Roles in the Enhancement of Health Care Delivery and User-Centric Services

Our review underscores the transformative roles of chatbots in health care, particularly in delivering remote health services and enhancing patient support, care management, and mental health support. Consistent with previous literature [254-257], our findings affirm chatbots' potential to improve health care accessibility and patient management. The findings' emphasis on education and skills building, particularly to enhance health literacy (which aligned with past literature [255,258]) and to support behavioral change (also highlighted by past research [255]), aligns with the growing need for patient empowerment in health care. The administrative efficiency of chatbots, noted in our review, resonates with previous findings [23,35,255,258] on the importance of resource optimization in health care settings.

Our findings indicate that chatbots also play a key role in facilitating clinical research, consistent with past work [259], a potential that needs further exploration, especially considering AI's evolving role in health care [72,259-262].

The Diverse User Base of Chatbots Shows Their Potential to Support Equity and Bridge the Access Gap in Health Care Services

Our analysis indicates a broad and diverse user base for health care chatbots. From individuals focused on general well-being to those with specific health conditions, chatbots have been designed to cater to a wide array of needs. Notably, their use by demographic and family-centric groups and their accessibility to underserved populations underline the inclusive capacity of chatbots and their role in enhancing health care access and equity, especially for groups considered marginalized, in line with existing research [12,67,255,263-265].

In addition, our findings show the significant use of chatbots in mental health support for various age groups, reflecting the pressing need for accessible mental health services highlighted by others [4,8,12-17,29,30].

Furthermore, chatbots have emerged as tools for reducing stigma [12,265], linking users to health services [266-268], and protecting sensitive information [269]. Their empathetic and multilingual capabilities, as seen in our results [107,111,112,120,122,126-128,132] and past literature [270-276], are vital to reach diverse populations. They are particularly critical in light of the digital divide and the need for inclusive and accessible health care solutions [254,258,263,277,278].

XSL•FO
RenderX

The Use of AI in Chatbots Is a Promising but Still Evolving Field

The studies included in our review show a substantial number of AI-based chatbots, with fewer relying on non-AI platforms. AI in health care is recognized for its potential to improve health outcomes and the quality of life globally [260]. Given advances in machine learning and AI, expanding the scope of chatbots is expected to cause a mutation in their role in the health care system to assist clinicians and potentially take over some of their duties [72,261,262]. The synergy between big data and AI, coupled with the increasing availability of data in health care, suggests that AI-based chatbots could effectively use extensive health care data [259,279]. This aligns with 1 (0.6%) of the 161 included studies [94], which discusses the use of collected data as a key benefit of chatbots. However, ethical considerations such as data privacy and algorithmic biases must be addressed for responsible AI deployment, crucial for maintaining trust and fairness [73].

Studies included in this review indicate that using avatars in these chatbots to simulate social behaviors can enhance user engagement and trust. This form of chatbot technology is particularly appealing in patient interactions and medical education to establish trust and therapeutic alliances between health care professionals and patients and to improve the communication skills of medical students and health care professionals [118,123,130,131,280].

Balancing AI's benefits to enhance data use and user interactions with its ethical concerns, including data privacy and algorithmic bias, is crucial for its implementation, shaping the future of patient care and medical education in an innovative and ethically sound way.

Despite the Potential Revolutionary Roles of Chatbots in Health Care, Critical Challenges and Limitations Exist

This review stresses that despite chatbots' roles and benefits, their use comes with various challenges, including ethical, technical, medicolegal, and user experience concerns, as also discussed in past literature [3-5,23,25,30,72,74,95,281,282].

While the studies included in our review have highlighted chatbot use to address minor health concerns and provide off-hour information, there is a noticeable gap in evaluating their technical limitations, especially in complex health care scenarios, as underscored by past literature [3-5,23,25,30,72,74,95,281,282]. This raises concerns about patient safety and the accuracy of health management, emphasizing the need for comprehensive assessment and iterative improvement of chatbot technologies [22,25,68,72,95,254,283].

The findings in our review indicate the regulatory and ethical landscape for chatbots as another area of concern. This agrees with past studies highlighting the need for ethical use, data privacy, and transparent communication about chatbots' capabilities and limitations [4,73,74,254,281,284,285]. The absence of specific laws and regulations addressing health care chatbot use introduces risks around liability and medicolegal issues [72,286,287]. These challenges are further complicated by ethical dilemmas, such as privacy and confidentiality in nonanonymous interactions [71,72,288,289] and safety concerns in medical emergencies due to limited chatbot expertise [72].

Technical issues identified by this review, including difficulty in language processing and a lack of empathic response, can lead to trust issues and increased clinical workload and align with past literature [3-5,68,72,73,280,290]. Overreliance on chatbots for self-diagnosis and health care decisions may lead to misjudgments, potentially exacerbating health issues [4,68,73]. In addition, the financial motives of private companies in the health sector raise ethical concerns about the primary purpose and application of health chatbots [73]. The requirement for sophisticated AI technology also implies increased demands on human resource expertise and storage services, potentially escalating costs [73,287].

Our results indicate that chatbots serve a wide range of populations from various groups in terms of age, gender, ethnicity, and socioeconomic and educational status due to their promising acceptability and usability [291]. However, the digital divide [292-294], algorithmic ethical concerns [295], and the potential misuse of chatbots in replacing established health services [296] present risks. These factors, along with social, economic, and political influences [297], could inadvertently widen health disparities, highlighting the importance of inclusive and equitable chatbot development and deployment.

The discussion on health care chatbots is fundamentally about their potential and promise, grounded in our exploration of current studies and developments. These digital tools could significantly enhance health care access, service quality, and efficiency. However, realizing their full potential hinges on addressing challenges such as ethical AI use, data privacy, and integration with health care systems.

Efforts moving forward should concentrate on incorporating AI responsibly and designing chatbots that cater to all user demographics, ensuring equitable health care access. Collaboration across technology, health care, and policy sectors is crucial to establish ethical guidelines and confirm chatbots' efficacy and safety. Successfully navigating these challenges will enable chatbots to fulfill their promising role in health care, contributing to a more accessible and patient-focused system.

Limitations

This review, while insightful, is not without its limitations. Although rapid and systematic reviews are often considered comparable in their conclusions, each methodology has its own set of constraints [289,298]. Specifically, this rapid review was limited by a noncomprehensive search strategy that included only 2 databases. In addition, the inclusion criteria were restricted by date and language, which potentially led to the exclusion of some pertinent studies. Another limitation was the concentration of screening and analysis tasks on a single reviewer (ML), which might have introduced bias or overlooked nuances in the data. Moreover, a formal quality appraisal of the included studies was not conducted due to the descriptive nature of this review. Consequently, this limitation may affect the depth of understanding and the strength of the conclusions drawn.

XSL•FO
RenderX

One critical aspect of our methodology was the combination of empirical findings and opinion-based data from the discussions in the included studies. We did not distinguish between these 2 types of data but rather treated them as a unified source of information. This approach, while allowing for a comprehensive overview of chatbots in health care, might have led to a potential bias in favor of chatbot benefits because both empirical results and positive speculative insights were reported together. However, this potential bias is somewhat mitigated by our consistent reporting of the challenges associated with chatbots, as identified in the included studies. By presenting both the potential benefits and the challenges, we aimed to offer a balanced view, reducing the likelihood of a 1-sided interpretation favoring chatbot benefits.

In addition, this review might have overestimated the results due to the dependence on the discussion sections of each study, which may have overcounted the results and miscounted those that may have disagreed or contradicted the results of these included studies. However, this did not significantly impact the study's aim to provide an exploratory and descriptive overview of health care chatbots, mapping out the landscape of their applications in health care. In such a context, a broad, inclusive approach that captures diverse opinions and trends is more important than precise quantification.

Moreover, one of the potential limitations of this review is the exclusion of generative AI and LLMs such as ChatGPT. However, among the studies we reviewed, a standout comparison involved a health care chatbot, specialized in medical terminology, and ChatGPT. This unique comparison serves to highlight the advanced capabilities of LLMs such as ChatGPT in enhancing the delivery and accuracy of remote health services [59,75]. Nonetheless, a significant challenge persists in guaranteeing the contextual relevance and appropriateness of chatbot responses, particularly in intricate medical scenarios [59,60]. In addition, the personalization of health care interactions and the precision of information provided by these AI-driven systems are critical areas necessitating extensive future research and rigorous evaluation of their outputs [59,60,299].

Finally, the results were presented solely as a narrative summary [77], which might limit the breadth of perspectives and interpretations that a more diverse methodological approach could have provided. Nevertheless, the inclusion of both benefits and challenges in our reporting suggests that the review may not be significantly biased toward a positive portrayal of chatbots, providing a more nuanced understanding of their role in health care.

Conclusions

This review underscores the significant potential of chatbots in health care, evident in their diverse roles, benefits, and user populations. In addition, it explores the current limitations and challenges of chatbot development and implementation in health care. Finally, it underscores significant research gaps in the field. As such, this review aims to contribute to academic discourse on this important topic and offer insights into the effective design, implementation, and investigation of chatbots in health care.

Acknowledgments

BL is supported by 2 career awards—a Senior Salary Award from Fonds de recherche du Québec–Santé (311200) and the Lettre d'Entente (LE) 250 from Québec's Ministry of Health for researchers in family medicine—and holds a Canadian Institutes for Health Research Strategy for Patient-Oriented Research Mentorship Chair in Innovative Clinical Trials. ML is supported by a Graduate Excellence Fellowship Award from the Department of Family Medicine at McGill University and a McGill Centre for Viral Diseases studentship. YM was supported by a doctoral scholarship from the Fonds de recherche du Québec–Nature et Technologies given in partnership with the Strategy for Patient-Oriented Research Support Unit of Québec. YM is supported by a postgraduate scholarship–doctoral program given through the Natural Sciences and Engineering Research Council of Canada.

Authors' Contributions

ML and YM dual-screened 15% of the titles and abstracts and the full texts. ML performed all other screenings. ML analyzed and interpreted the results and wrote the manuscript. DL reviewed the results. All authors reviewed and edited the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

1. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The personalization of conversational agents in health care: systematic review. J Med Internet Res 2019 Nov 07;21(11):e15360 [FREE Full text] [doi: 10.2196/15360] [Medline: 31697237]
2. Luo TC, Aguilera A, Lyles CR, Figueroa CA. Promoting physical activity through conversational agents: mixed methods systematic review. J Med Internet Res 2021 Sep 14;23(9):e25486 [FREE Full text] [doi: 10.2196/25486] [Medline: 34519653]
3. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]

XSL•FO
RenderX

4. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul 21;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
5. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
6. Bin Sawad A, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, et al. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. Sensors (Basel) 2022 Mar 29;22(7):2625 [FREE Full text] [doi: 10.3390/s22072625] [Medline: 35408238]
7. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
8. Vaidyam AN, Linggonegoro D, Torous J. Changes to the psychiatric chatbot landscape: a systematic review of conversational agents in serious mental illness: changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse. Can J Psychiatry 2021 Apr 16;66(4):339-348 [FREE Full text] [doi: 10.1177/0706743720966429] [Medline: 33063526]
9. Geoghegan L, Scarborough A, Wormald JC, Harrison CJ, Collins D, Gardiner M, et al. Automated conversational agents for post-intervention follow-up: a systematic review. BJS Open 2021 Jul 06;5(4):zrab070 [FREE Full text] [doi: 10.1093/bjsopen/zrab070] [Medline: 34323916]
10. Allouch M, Azaria A, Azoulay R. Conversational agents: goals, technologies, vision and challenges. Sensors (Basel) 2021 Dec 17;21(24):8448 [FREE Full text] [doi: 10.3390/s21248448] [Medline: 34960538]
11. Bibault JE, Chaix B, Nectoux P, Pienkowsky A, Guillemasse A, Brouard B. Healthcare ex Machina: are conversational agents ready for prime time in oncology? Clin Transl Radiat Oncol 2019 May;16:55-59 [FREE Full text] [doi: 10.1016/j.ctro.2019.04.002] [Medline: 31008379]
12. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform 2019 Dec;132:103978 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
13. Pacheco-Lorenzo MR, Valladares-Rodríguez SM, Anido-Rifón LE, Fernández-Iglesias MJ. Smart conversational agents for the detection of neuropsychiatric disorders: a systematic review. J Biomed Inform 2021 Jan;113:103632 [FREE Full text] [doi: 10.1016/j.jbi.2020.103632] [Medline: 33276112]
14. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 09;19(5):e151 [FREE Full text] [doi: 10.2196/jmir.6553] [Medline: 28487267]
15. Rampioni M, Stara V, Felici E, Rossi L, Paolini S. Embodied conversational agents for patients with dementia: thematic literature analysis. JMIR Mhealth Uhealth 2021 Jul 16;9(7):e25381 [FREE Full text] [doi: 10.2196/25381] [Medline: 34269686]
16. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
17. Bérubé C, Schachner T, Keller R, Fleisch E, V Wangenheim F, Barata F, et al. Voice-based conversational agents for the prevention and management of chronic and mental health conditions: systematic literature review. J Med Internet Res 2021 Mar 29;23(3):e25933 [FREE Full text] [doi: 10.2196/25933] [Medline: 33658174]
18. Chew HS. The use of artificial intelligence-based conversational agents (Chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022 Apr 13;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
19. Kramer LL, Ter Stal S, Mulder BC, de Vet E, van Velsen L. Developing embodied conversational agents for coaching people in a healthy lifestyle: scoping review. J Med Internet Res 2020 Feb 06;22(2):e14058 [FREE Full text] [doi: 10.2196/14058] [Medline: 32022693]
20. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
21. Reger GM, Norr AM, Gramlich MA, Buchman JM. Virtual standardized patients for mental health education. Curr Psychiatry Rep 2021 Jul 15;23(9):57. [doi: 10.1007/s11920-021-01273-5] [Medline: 34268633]
22. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: scoping review. J Med Internet Res 2020 Dec 18;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
23. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. J Med Internet Res 2020 Jun 05;22(6):e18301 [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]
24. Wang A, Qian Z, Briggs L, Cole AP, Reis LO, Trinh QD. The use of chatbots in oncological care: a narrative review. Int J Gen Med 2023 May 1;16:1591-1602 [FREE Full text] [doi: 10.2147/IJGM.S408208] [Medline: 37152273]

XSL•FO
RenderX

25. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021 Nov 29;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
26. Passanante A, Pertwee E, Lin L, Lee KY, Wu JT, Larson HJ. Conversational AI and vaccine communication: systematic review of the evidence. J Med Internet Res 2023 Oct 03;25:e42758 [FREE Full text] [doi: 10.2196/42758] [Medline: 37788057]
27. Wutz M, Hermes M, Winter V, Köberlein-Neu J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. J Med Internet Res 2023 Sep 26;25:e46548 [FREE Full text] [doi: 10.2196/46548] [Medline: 37751279]
28. Hocking J, Oster C, Maeder A, Lange B. Design, development, and use of conversational agents in rehabilitation for adults with brain-related neurological conditions: a scoping review. JBI Evid Synth 2023 Feb 01;21(2):326-372. [doi: 10.11124/JBIES-22-00025] [Medline: 35976047]
29. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862 [FREE Full text] [doi: 10.2196/43862] [Medline: 37115595]
30. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021 [FREE Full text] [doi: 10.2196/16021] [Medline: 32673216]
31. Lin X, Martinengo L, Jabir AI, Ho AH, Car J, Atun R, et al. Scope, characteristics, behavior change techniques, and quality of conversational agents for mental health and well-being: systematic assessment of apps. J Med Internet Res 2023 Jul 18;25:e45984 [FREE Full text] [doi: 10.2196/45984] [Medline: 37463036]
32. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
33. Martinengo L, Jabir AI, Goh WW, Lo NY, Ho MH, Kowatsch T, et al. Conversational agents in health care: scoping review of their behavior change techniques and underpinning theory. J Med Internet Res 2022 Oct 03;24(10):e39243 [FREE Full text] [doi: 10.2196/39243] [Medline: 36190749]
34. Webster EM, Ahsan MD, Perez L, Levi SR, Thomas C, Christos P, et al. Chatbot artificial intelligence for genetic cancer risk assessment and counseling: a systematic review and meta-analysis. JCO Clin Cancer Inform 2023 Sep;7:e2300123 [FREE Full text] [doi: 10.1200/CCI.23.00123] [Medline: 37934933]
35. Mills R, Mangone ER, Lesh N, Mohan D, Baraitser P. Chatbots to improve sexual and reproductive health: realist synthesis. J Med Internet Res 2023 Aug 09;25:e46761 [FREE Full text] [doi: 10.2196/46761] [Medline: 37556194]
36. Chua JY, Choolani M, Chee CY, Chan YH, Lalor JG, Chong YS, et al. Insights of parents and parents-to-be in using chatbots to improve their preconception, pregnancy, and postpartum health: a mixed studies review. J Midwifery Womens Health 2023 Feb 03;68(4):480-489. [doi: 10.1111/jmwh.13472] [Medline: 36734375]
37. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023 Jun 23;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
38. He L, Balaji D, Wiers RW, Antheunis ML, Krahmer E. Effectiveness and acceptability of conversational agents for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Jun 09;25(7):1241-1250 [FREE Full text] [doi: 10.1093/ntr/ntac281] [Medline: 36507916]
39. Noh E, Won J, Jo S, Hahm DH, Lee H. Conversational agents for body weight management: systematic review. J Med Internet Res 2023 May 26;25:e42238 [FREE Full text] [doi: 10.2196/42238] [Medline: 37234029]
40. Thoumrungroje P, Chainarong A, Namwaing P, Pitiruangsit L, Sittichanbuncha Y, Ngamjarus C, et al. Chatbot intervention in asthma and obstructive sleep apnea: a systematic review. J Med Assoc Thai 2023;106:S134 [FREE Full text] [doi: 10.35755/jmedassocthai.2023.S01.13741]
41. Lim SM, Shiau CW, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
42. Ogilvie L, Prescott J, Carson J. The use of chatbots as supportive agents for people seeking help with substance use disorder: a systematic review. Eur Addict Res 2022 Aug 30;28(6):405-418 [FREE Full text] [doi: 10.1159/000525959] [Medline: 36041418]
43. Shan Y, Ji M, Xie W, Qian X, Li R, Zhang X, et al. Language use in conversational agent-based health communication: systematic review. J Med Internet Res 2022 Jul 08;24(7):e37403 [FREE Full text] [doi: 10.2196/37403] [Medline: 35802407]
44. Amiri P, Karahanna E. Chatbot use cases in the COVID-19 public health response. J Am Med Inform Assoc 2022 Apr 13;29(5):1000-1010 [FREE Full text] [doi: 10.1093/jamia/ocac014] [Medline: 35137107]
45. Balaji D, He L, Giani S, Bosse T, Wiers R, de Bruijn GJ. Effectiveness and acceptability of conversational agents for sexual health promotion: a systematic review and meta-analysis. Sex Health 2022 Oct;19(5):391-405 [FREE Full text] [doi: 10.1071/SH22016] [Medline: 35863761]

XSL•FO
RenderX

46. Christopherjames JE, Saravanan M, Thiyam DB, Surendhar S PA, Sahib MY, Ganapathi MV, et al. Natural language processing based human assistive health conversational agent for multi-users. In: Proceedings of the 2021 2nd International Conference on Electronics and Sustainable Communication Systems. 2021 Presented at: ICESC '21; August 4-6, 2021; Coimbatore, India p. 1414-1420 URL: https://ieeexplore.ieee.org/document/9532913 [doi: 10.1109/icesc51422.2021.9532913]
47. Imtiaz MT, Kennington C. Incremental unit networks for distributed, symbolic multimodal processing and representation. In: Proceedings of the 13th International Conference, DHM 2022, Held as Part of the 24th HCI International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management for Health, Operations Management, and Design. 2022 Presented at: HCII '22; June 26-July 1, 2022; Virtual Event p. 344-363 URL: https://link.springer.com/chapter/10.1007/978-3-031-06018-2_24 [doi: 10.1007/978-3-031-06018-2_24]
48. Mauldin ML. Chatterbots, Tinymuds, and the Turing test: entering the Loebner prize competition. In: Proceedings of the 25th AAAI National Conference on Artificial Intelligence. 1994 Presented at: AAAI '94; August 1-4, 1994; Seattle, WA p. 16-21 URL: https://dl.acm.org/doi/abs/10.5555/2891730.2891733
49. Nuruzzaman M, Hussain OK. A survey on chatbot implementation in customer service industry through deep neural networks. In: Proceedings of the 2018 IEEE 15th International Conference on e-Business Engineering. 2018 Presented at: ICEBE '18; October 12-14, 2018; Xi'an, China p. 54-61 URL: https://ieeexplore.ieee.org/document/8592630 [doi: 10.1109/icebe.2018.00019]
50. Kumar VM, Keerthana A, Madhumitha M, Valliammai S, Vinithasri V. Sanative chatbot for health seekers. Int J Eng Comput Sci 2016 Mar 29;5(03):16022-16025 [FREE Full text] [doi: 10.18535/ijecs/v5i3.28]
51. Alexa developer portal. Amazon. 2014. URL: https://www.developer.amazon.com/en-US/alexa/ [accessed 2022-12-22]
52. Siri. Apple. 2010. URL: https://www.apple.com/siri/ [accessed 2022-12-22]
53. Google Assistant, your own personal Google. Google. 2016. URL: https://assistant.google.com/ [accessed 2022-12-22]
54. Cortana- your personal productivity assistant. Microsoft. 2014. URL: https://www.microsoft.com/en-us/cortana/ [accessed 2022-12-22]
55. Bixby: apps and services. Samsung. 2017. URL: https://www.samsung.com/in/apps/bixby/ [accessed 2022-12-22]
56. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2022-12-22]
57. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv Preprint posted online March 4, 2022 [FREE Full text]
58. OpenAI. GPT-4 technical report. arXiv. Preprint posted online March 15, 2023 2023 [FREE Full text]
59. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
60. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
61. Huisman L, van Duijn SM, Silva N, van Doeveren R, Michuki J, Kuria M, et al. A digital mobile health platform increasing efficiency and transparency towards universal health coverage in low- and middle-income countries. Digit Health 2022 Apr 11;8:20552076221092213 [FREE Full text] [doi: 10.1177/20552076221092213] [Medline: 35433018]
62. Osipov VS, Skryl TV. Impact of digital technologies on the efficiency of healthcare delivery. In: Marques G, Bhoi AK, de Albuquerque VH, Hareesha KS, editors. IoT in Healthcare and Ambient Assisted Living. Singapore, Singapore: Springer; 2021:243-261.
63. Jones SP, Patel V, Saxena S, Radcliffe N, Ali Al-Marri S, Darzi A. How Google's 'ten things we know to be true' could guide the development of mental health mobile apps. Health Aff (Millwood) 2014 Sep;33(9):1603-1611. [doi: 10.1377/hlthaff.2014.0380] [Medline: 25201665]
64. Chandrashekar P. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth 2018 Mar;4:6 [FREE Full text] [doi: 10.21037/mhealth.2018.03.02] [Medline: 29682510]
65. Tanielian T, Jaycox LH. Invisible Wounds of War: Psychological and Cognitive Injuries, Their Consequences, and Services to Assist Recovery. Santa Monica, CA: RAND Corporation; 2008.
66. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019 Aug 21;5:2055207619871808 [FREE Full text] [doi: 10.1177/2055207619871808] [Medline: 31467682]
67. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
68. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory Group. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (Chatbots) in mental health support. Biomed Inform Insights 2019 Mar 05;11:1178222619829083 [FREE Full text] [doi: 10.1177/1178222619829083] [Medline: 30858710]

XSL•FO
RenderX

69. Cheng Y, Jiang H. AI‐Powered mental health chatbots: examining users' motivations, active communicative action and engagement after mass‐shooting disasters. J Conting Crisis Man 2020 Sep 29;28(3):339-354. [doi: 10.1111/1468-5973.12319]
70. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec 31;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
71. Luxton DD, Anderson SL, Anderson M. Ethical issues and artificial intelligence technologies in behavioral and mental health care. In: Luxton DD, editor. Artificial Intelligence in Behavioral and Mental Health Care. Washington, DC: Academic Press; 2016:255-276.
72. Denecke K, Abd-Alrazaq A, Househ M. Artificial intelligence for chatbots in mental health: opportunities and challenges. In: Household M, Borycki E, Kushniruk A, editors. Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges. Cham, Switzerland: Springer; 2021:115-128.
73. Parviainen J, Rantala J. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos 2022 Mar 04;25(1):61-71 [FREE Full text] [doi: 10.1007/s11019-021-10049-w] [Medline: 34480711]
74. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019 Apr 05;21(4):e12887 [FREE Full text] [doi: 10.2196/12887] [Medline: 30950796]
75. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol 2024 Mar 29;15(3):1-45. [doi: 10.1145/3641289]
76. Mohanta B, Das P, Patnaik S. Healthcare 5.0: a paradigm shift in digital healthcare system using artificial intelligence, IOT and 5G communication. In: Proceedings of the 2019 International Conference on Applied Machine Learning. 2019 Presented at: ICAML '19; May 25-26, 2019; Bhubaneswar, India p. 191-196 URL: https://ieeexplore.ieee.org/document/8989306 [doi: 10.1109/icaml48257.2019.00044]
77. Tricco AC, Antony J, Zarin W, Strifler L, Ghassemi M, Ivory J, et al. A scoping review of rapid review methods. BMC Med 2015 Sep 16;13(1):224 [FREE Full text] [doi: 10.1186/s12916-015-0465-6] [Medline: 26377409]
78. Hamel C, Michaud A, Thuku M, Skidmore B, Stevens A, Nussbaumer-Streit B, et al. Defining Rapid Reviews: a systematic scoping review and thematic analysis of definitions and defining characteristics of rapid reviews. J Clin Epidemiol 2021 Jan;129:74-85. [doi: 10.1016/j.jclinepi.2020.09.041] [Medline: 33038541]
79. Tricco AC, Langlois EV, Straus SE. Rapid reviews to strengthen health policy and systems: a practical guide. World Health Organization. 2017. URL: https://iris.who.int/bitstream/handle/10665/258698/9789241512763-eng.pdf?sequence=1 [accessed 2024-04-29]
80. Garritty C, Gartlehner G, Nussbaumer-Streit B, King VJ, Hamel C, Kamel C, et al. Cochrane rapid reviews methods group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol 2021 Feb;130:13-22 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.10.007] [Medline: 33068715]
81. Denecke K, May R. Developing a technical-oriented taxonomy to define archetypes of conversational agents in health care: literature review and cluster analysis. J Med Internet Res 2023 Jan 30;25:e41583 [FREE Full text] [doi: 10.2196/41583] [Medline: 36716093]
82. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
83. McHugh ML. Interrater reliability: the Kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
84. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr 18;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
85. Nazareth S, Hayward L, Simmons E, Snir M, Hatchell KE, Rojahn S, et al. Hereditary cancer risk using a genetic chatbot before routine care visits. Obstet Gynecol 2021 Dec 01;138(6):860-870 [FREE Full text] [doi: 10.1097/AOG.0000000000004596] [Medline: 34735417]
86. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend 2021 Oct 01;227:108986 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108986] [Medline: 34507061]
87. Söderström A, Shatte A, Fuller-Tyszkiewicz M. Can intelligent agents improve data quality in online questiosnnaires? A pilot study. Behav Res Methods 2021 Oct 05;53(5):2238-2251. [doi: 10.3758/s13428-021-01574-w] [Medline: 33821454]
88. Anan T, Kajiki S, Oka H, Fujii T, Kawamata K, Mori K, et al. Effects of an artificial intelligence-assisted health program on workers with neck/shoulder pain/stiffness and low back pain: randomized controlled trial. JMIR Mhealth Uhealth 2021 Sep 24;9(9):e27535 [FREE Full text] [doi: 10.2196/27535] [Medline: 34559054]
89. Stasinaki A, Büchter D, Shih CH, Heldt K, Güsewell S, Brogle B, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 09;21(1):308 [FREE Full text] [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]

XSL•FO
RenderX

90. Jang S, Kim JJ, Kim SJ, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. Int J Med Inform 2021 Jun;150:104440. [doi: 10.1016/j.ijmedinf.2021.104440] [Medline: 33799055]
91. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152 [FREE Full text] [doi: 10.2196/26152] [Medline: 33872182]
92. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
93. Echeazarra L, Pereira J, Saracho R. TensioBot: a chatbot assistant for self-managed in-house blood pressure checking. J Med Syst 2021 Mar 15;45(4):54. [doi: 10.1007/s10916-021-01730-x] [Medline: 33723721]
94. Gardiner P, Bickmore T, Yinusa-Nyahkoon L, Reichert M, Julce C, Sidduri N, et al. Using health information technology to engage African American women on nutrition and supplement use during the preconception period. Front Endocrinol (Lausanne) 2020 Jan 19;11:571705 [FREE Full text] [doi: 10.3389/fendo.2020.571705] [Medline: 33584534]
95. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. J Med Internet Res 2021 Jan 06;23(1):e19928 [FREE Full text] [doi: 10.2196/19928] [Medline: 33404508]
96. Maeda E, Miyata A, Boivin J, Nomura K, Kumazawa Y, Shirasawa H, et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. Reprod Biomed Online 2020 Dec;41(6):1133-1143. [doi: 10.1016/j.rbmo.2020.09.006] [Medline: 33039321]
97. So R, Furukawa TA, Matsushita S, Baba T, Matsuzaki T, Furuno S, et al. Unguided ChatBot-delivered cognitive behavioural intervention for problem gamblers through messaging app: a randomised controlled trial. J Gambl Stud 2020 Dec 11;36(4):1391-1407. [doi: 10.1007/s10899-020-09935-4] [Medline: 32162075]
98. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My diabetes coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 05;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]
99. Jack BW, Bickmore T, Yinusa-Nyahkoon L, Reichert M, Julce C, Sidduri N, et al. Improving the health of young African American women in the preconception period using health information technology: a randomised controlled trial. Lancet Digit Health 2020 Sep;2(9):e475-e485 [FREE Full text] [doi: 10.1016/S2589-7500(20)30189-8] [Medline: 33328115]
100. Yoneoka D, Kawashima T, Tanoue Y, Nomura S, Ejima K, Shi S, et al. Early SNS-based monitoring system for the COVID-19 outbreak in Japan: a population-level observational study. J Epidemiol 2020 Aug 05;30(8):362-370 [FREE Full text] [doi: 10.2188/jea.JE20200150] [Medline: 32475884]
101. Bickmore T, Zhang Z, Reichert M, Julce C, Jack B. Promotion of preconception care among adolescents and young adults by conversational agent. J Adolesc Health 2020 Aug;67(2S):S45-S51 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.09.006] [Medline: 32718515]
102. Oh J, Jang S, Kim H, Kim JJ. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
103. Anthony CA, Rojas EO, Keffala V, Glass NA, Shah AS, Miller BJ, et al. Acceptance and commitment therapy delivered via a mobile phone messaging robot to decrease postoperative opioid use in patients with orthopedic trauma: randomized controlled trial. J Med Internet Res 2020 Jul 29;22(7):e17750 [FREE Full text] [doi: 10.2196/17750] [Medline: 32723723]
104. Kramer JN, Künzler F, Mishra V, Smith SN, Kotz D, Scholz U, et al. Which components of a smartphone walking app help users to reach personalized step goals? Results from an optimization trial. Ann Behav Med 2020 Jun 12;54(7):518-528 [FREE Full text] [doi: 10.1093/abm/kaaa002] [Medline: 32182353]
105. Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, acceptability, and effectiveness of web-based conversational agents to facilitate problem solving in older adults: controlled study. J Med Internet Res 2020 May 27;22(5):e16794 [FREE Full text] [doi: 10.2196/16794] [Medline: 32384055]
106. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching ChatBot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. JMIR Mhealth Uhealth 2020 May 19;8(5):e15085 [FREE Full text] [doi: 10.2196/15085] [Medline: 32427114]
107. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 03;8(4):e15806 [FREE Full text] [doi: 10.2196/15806] [Medline: 32242820]
108. Bibault JE, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M, et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. J Med Internet Res 2019 Nov 27;21(11):e15787 [FREE Full text] [doi: 10.2196/15787] [Medline: 31774408]
109. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
110. Fadhil A, Wang Y, Reiterer H. Assistive conversational agent for health coaching: a validation study. Methods Inf Med 2019 Jun 22;58(1):9-23. [doi: 10.1055/s-0039-1688757] [Medline: 31117129]

XSL•FO
RenderX

111. Abdullah AS, Gaehde S, Bickmore T. A tablet based embodied conversational agent to promote smoking cessation among veterans: a feasibility study. J Epidemiol Glob Health 2018 Dec;8(3-4):225-230 [FREE Full text] [doi: 10.2991/j.jegh.2018.08.104] [Medline: 30864768]
112. Auriacombe M, Moriceau S, Serre F, Denis C, Micoulaud-Franchi JA, de Sevin E, et al. Development and validation of a virtual agent to screen tobacco and alcohol use disorders. Drug Alcohol Depend 2018 Dec 01;193:1-6. [doi: 10.1016/j.drugalcdep.2018.08.025] [Medline: 30321739]
113. King AC, Campero I, Sheats JL, Castro Sweet CM, Garcia D, Chazaro A, et al. Testing the comparative effects of physical activity advice by humans vs. computers in underserved populations: the COMPASS trial design, methods, and baseline characteristics. Contemp Clin Trials 2017 Oct;61:115-125 [FREE Full text] [doi: 10.1016/j.cct.2017.07.020] [Medline: 28739541]
114. Gardiner PM, McCue KD, Negash LM, Cheng T, White LF, Yinusa-Nyahkoon L, et al. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Educ Couns 2017 Sep;100(9):1720-1729 [FREE Full text] [doi: 10.1016/j.pec.2017.04.015] [Medline: 28495391]
115. Hajna S, Sharp SJ, Cooper AJ, Williams KM, van Sluijs EM, Brage S, et al. Effectiveness of minimal contact interventions: an RCT. Am J Prev Med 2021 Mar;60(3):e111-e121 [FREE Full text] [doi: 10.1016/j.amepre.2020.10.010] [Medline: 33612170]
116. Ali R, Hoque E, Duberstein P, Schubert L, Razavi SZ, Kane B, et al. Aging and engaging: a pilot randomized controlled trial of an online conversational skills coach for older adults. Am J Geriatr Psychiatry 2021 Aug;29(8):804-815 [FREE Full text] [doi: 10.1016/j.jagp.2020.11.004] [Medline: 33308893]
117. Reger GM, Norr AM, Rizzo AS, Sylvers P, Peltan J, Fischer D, et al. Virtual standardized patients vs academic training for learning motivational interviewing skills in the US Department of Veterans Affairs and the US Military: a randomized trial. JAMA Netw Open 2020 Oct 01;3(10):e2017348 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.17348] [Medline: 33057643]
118. Lorenz N, Heim E, Roetger A, Birrer E, Maercker A. Randomized controlled trial to test the efficacy of an unguided online intervention with automated feedback for the treatment of insomnia. Behav Cogn Psychother 2019 May;47(3):287-302. [doi: 10.1017/S1352465818000486] [Medline: 30185239]
119. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
120. Abujarad F, Ulrich D, Edwards C, Choo E, Pantalon MV, Jubanyik K, et al. Development and usability evaluation of VOICES: a digital health tool to identify elder mistreatment. J Am Geriatr Soc 2021 Jun 21;69(6):1469-1478 [FREE Full text] [doi: 10.1111/jgs.17068] [Medline: 33615433]
121. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. Acceptability of an embodied conversational agent for type 2 diabetes self-management education and support via a smartphone app: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e17038 [FREE Full text] [doi: 10.2196/17038] [Medline: 32706734]
122. Hirsh AT, Miller MM, Hollingshead NA, Anastas T, Carnell ST, Lok BC, et al. A randomized controlled trial testing a virtual perspective-taking intervention to reduce race and socioeconomic status disparities in pain care. Pain 2019 Oct 10;160(10):2229-2240 [FREE Full text] [doi: 10.1097/j.pain.0000000000001634] [Medline: 31568099]
123. Kim YJ, DeLisa JA, Chung YC, Shapiro NL, Kolar Rajanna SK, Barbour E, et al. Recruitment in a research study via chatbot versus telephone outreach: a randomized trial at a minority-serving institution. J Am Med Inform Assoc 2021 Dec 28;29(1):149-154 [FREE Full text] [doi: 10.1093/jamia/ocab240] [Medline: 34741513]
124. Kowatsch T, Schachner T, Harperink S, Barata F, Dittler U, Xiao G, et al. Conversational agents as mediating social actors in chronic disease management involving health care professionals, patients, and family members: multisite single-arm feasibility study. J Med Internet Res 2021 Feb 17;23(2):e25060 [FREE Full text] [doi: 10.2196/25060] [Medline: 33484114]
125. Loveys K, Sagar M, Broadbent E. The effect of multimodal emotional expression on responses to a digital human during a self-disclosure conversation: a computational analysis of user language. J Med Syst 2020 Jul 22;44(9):143. [doi: 10.1007/s10916-020-01624-4] [Medline: 32700060]
126. Loveys K, Sagar M, Pickering I, Broadbent E. A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: randomized pilot trial. JMIR Ment Health 2021 Nov 08;8(11):e31586 [FREE Full text] [doi: 10.2196/31586] [Medline: 34596572]
127. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
128. Morse KE, Ostberg NP, Jones VG, Chan AS. Use characteristics and triage acuity of a digital symptom checker in a large integrated health system: population-based descriptive study. J Med Internet Res 2020 Nov 30;22(11):e20549 [FREE Full text] [doi: 10.2196/20549] [Medline: 33170799]
129. O'Rourke SR, Branford KR, Brooks TL, Ives LT, Nagendran A, Compton SN. The emotional and behavioral impact of delivering bad news to virtual versus real standardized patients: a pilot study. Teach Learn Med 2020 Aug 22;32(2):139-149. [doi: 10.1080/10401334.2019.1652180] [Medline: 31437006]

XSL•FO
RenderX

130. Schario ME, Bahner CA, Widenhofer TV, Rajaballey JI, Thatcher EJ. Chatbot-assisted care management. Prof Case Manag 2022;27(1):19-25. [doi: 10.1097/NCM.0000000000000504] [Medline: 34846321]
131. Dhinagaran DA, Sathish T, Soong A, Theng YL, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. JMIR Form Res 2021 Dec 03;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]
132. Troitskaya O, Batkhina A. Mobile application for couple relationships: results of a pilot effectiveness study. Fam Process 2022 Jun 14;61(2):625-642. [doi: 10.1111/famp.12733] [Medline: 34904235]
133. So R, Emura N, Okazaki K, Takeda S, Sunami T, Kitagawa K, et al. Guided versus unguided chatbot-delivered cognitive behavioral intervention for individuals with moderate-risk and problem gambling: a randomized controlled trial (GAMBOT2 study). Addict Behav 2024 Feb;149:107889. [doi: 10.1016/j.addbeh.2023.107889] [Medline: 37857043]
134. An J, Ferrante JM, Macenat M, Ganesan S, Hudson SV, Omene C, et al. Promoting informed approaches in precision oncology and clinical trial participation for Black patients with cancer: community-engaged development and pilot testing of a digital intervention. Cancer (Forthcoming) 2023 Oct 13. [doi: 10.1002/cncr.35049] [Medline: 37837177]
135. Coelho J, Pecune F, Levavasseur Y, De Sevin E, D'incau E, Sagaspe P, et al. From improved sleep regularity to reduced sleep complaints and mental health conditions: a population-based interventional study using a smartphone-based virtual agent. Sleep 2023 Sep 08;46(9):zsad165. [doi: 10.1093/sleep/zsad165] [Medline: 37282717]
136. Siglen E, Vetti HH, Augestad M, Steen VM, Lunde Å, Bjorvatn C. Evaluation of the rosa chatbot providing genetic information to patients at risk of hereditary breast and ovarian cancer: qualitative interview study. J Med Internet Res 2023 Sep 01;25:e46571 [FREE Full text] [doi: 10.2196/46571] [Medline: 37656502]
137. Griffin AC, Khairat S, Bailey SC, Chung AE. A chatbot for hypertension self-management support: user-centered design, development, and usability testing. JAMIA Open 2023 Oct;6(3):ooad073 [FREE Full text] [doi: 10.1093/jamiaopen/ooad073] [Medline: 37693367]
138. Nadarzynski T, Lunt A, Knights N, Bayley J, Llewellyn C. "But can chatbots understand sex?" Attitudes towards artificial intelligence chatbots amongst sexual and reproductive health professionals: an exploratory mixed-methods study. Int J STD AIDS 2023 Oct 03;34(11):809-816 [FREE Full text] [doi: 10.1177/09564624231180777] [Medline: 37269292]
139. Alkoudmani RM, Ooi GS, Tan ML. Implementing a chatbot on Facebook to reach and collect data from thousands of health care providers: PharmindBot as a case. J Am Pharm Assoc (2003) 2023 Sep;63(5):1634-42.e3. [doi: 10.1016/j.japh.2023.06.007] [Medline: 37327997]
140. Morato JE, do Nascimento JW, Roque GD, de Souza RR, Santos IC. Development, validation, and usability of the Chatbot ESTOMABOT to promote self-care of people with intestinal ostomy. Comput Inform Nurs 2023 Dec 01;41(12):1037-1045. [doi: 10.1097/CIN.0000000000001075] [Medline: 37725781]
141. Al-Hilli Z, Noss R, Dickard J, Wei W, Chichura A, Wu V, et al. A randomized trial comparing the effectiveness of pre-test genetic counseling using an artificial intelligence automated chatbot and traditional in-person genetic counseling in women newly diagnosed with breast cancer. Ann Surg Oncol 2023 Oct 11;30(10):5990-5996. [doi: 10.1245/s10434-023-13888-4] [Medline: 37567976]
142. Hsu MH, Chen YH. Personalized medical terminology learning game: guess the term. Games Health J 2024 Apr 01;13(2):84-92. [doi: 10.1089/g4h.2023.0054] [Medline: 37699207]
143. Escobar-Viera CG, Porta G, Coulter RW, Martina J, Goldbach J, Rollman BL. A chatbot-delivered intervention for optimizing social media use and reducing perceived isolation among rural-living LGBTQ+ youth: development, acceptability, usability, satisfaction, and utility. Internet Interv 2023 Dec;34:100668 [FREE Full text] [doi: 10.1016/j.invent.2023.100668] [Medline: 37746640]
144. Cheng CI, Lin WJ, Liu HT, Chen YT, Chiang CK, Hung KY. Implementation of artificial intelligence chatbot in peritoneal dialysis nursing care: experience from a Taiwan medical center. Nephrology (Carlton) 2023 Dec 12;28(12):655-662. [doi: 10.1111/nep.14239] [Medline: 37698229]
145. Wang Z, Chan PS, Fang Y, Yu FY, Ye D, Zhang Q, et al. Chatbot-delivered online intervention to promote seasonal influenza vaccination during the COVID-19 pandemic: a randomized clinical trial. JAMA Netw Open 2023 Sep 05;6(9):e2332568 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.32568] [Medline: 37695585]
146. de Queiroz DA, Passarello RS, de Moura Fé VV, Rossini A, da Silveira EF, de Queiroz EA, et al. A wearable chatbot-based model for monitoring colorectal cancer patients in the active phase of treatment. Healthc Anal 2023 Dec;4:100257 [FREE Full text] [doi: 10.1016/j.health.2023.100257]
147. Chow JS, Blight V, Brown M, Glynn V, Lane B, Larkin A, et al. Curious thing, an artificial intelligence (AI)-based conversational agent for COVID-19 patient management. Aust J Prim Health 2023 Aug;29(4):312-318. [doi: 10.1071/PY22045] [Medline: 36683166]
148. Marrs JC, Orlando ST, Saseen JJ, Novins-Montague S, Sandy LC, Waughtal J, et al. Description of patient questions received by clinical pharmacists in the Nudge Study. Am J Health Syst Pharm 2023 Sep 07;80(18):1247-1254 [FREE Full text] [doi: 10.1093/ajhp/zxad139] [Medline: 37353220]
149. Rainey JP, Blackburn BE, McCutcheon CL, Kenyon CM, Campbell KJ, Anderson LA, et al. A multilingual chatbot can effectively engage arthroplasty patients who have limited english proficiency. J Arthroplasty 2023 Jul;38(7 Suppl 2):S78-S83. [doi: 10.1016/j.arth.2023.04.014] [Medline: 37068567]

XSL•FO
RenderX

150. El Ayadi AM, Singh P, Duggal M, Kumar V, Kaur J, Sharma P, et al. Feasibility and acceptability of Saheli, a WhatsApp Chatbot, on COVID-19 vaccination among pregnant and breastfeeding women in rural North India. BMJ Innov 2023 Aug 22;9(4):195-206. [doi: 10.1136/bmjinnov-2022-001012]
151. Hsu MH. Mastering medical terminology with ChatGPT and Termbot. Health Educ J 2023 Sep 04 [FREE Full text] [doi: 10.1177/00178969231197371]
152. Mahlknecht A, Engl A, Piccoliori G, Wiedermann CJ. Supporting primary care through symptom checking artificial intelligence: a study of patient and physician attitudes in Italian general practice. BMC Prim Care 2023 Sep 04;24(1):174 [FREE Full text] [doi: 10.1186/s12875-023-02143-0] [Medline: 37661285]
153. Xu J, Yang L, Guo M. Designing and evaluating an emotionally responsive virtual patient simulation. Simul Healthc (Forthcoming) 2023 May 18. [doi: 10.1097/SIH.0000000000000730] [Medline: 37651599]
154. Gupta S, Gupta SS, McMath K, Sugandh S. Enhancing complex wound care by leveraging artificial intelligence: an artificial intelligence chatbot software study. Wounds 2023 Aug;35(8):E265-E267 [FREE Full text] [doi: 10.25270/wnds/23073] [Medline: 37643453]
155. Beavers J, Schell RF, VanCleave H, Dillon RC, Simmons A, Chen H, et al. Evaluation of inpatient medication guidance from an artificial intelligence chatbot. Am J Health Syst Pharm 2023 Dec 05;80(24):1822-1829. [doi: 10.1093/ajhp/zxad193] [Medline: 37611187]
156. Larizza C, Bosoni P, Quaglini S, Chasseur M, Bevolo V, Zuccotti G, et al. V-care: An application to support lifestyle improvement in children with obesity. Int J Med Inform 2023 Sep;177:105140. [doi: 10.1016/j.ijmedinf.2023.105140] [Medline: 37463558]
157. Suharwardy S, Ramachandran M, Leonard SA, Gunaseelan A, Lyell DJ, Darcy A, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep 2023 Aug;3(3):100165 [FREE Full text] [doi: 10.1016/j.xagr.2023.100165] [Medline: 37560011]
158. Potts C, Lindström F, Bond R, Mulvenna M, Booth F, Ennis E, et al. A multilingual digital mental health and well-being chatbot (ChatPal): pre-post multicenter intervention study. J Med Internet Res 2023 Jul 06;25:e43051 [FREE Full text] [doi: 10.2196/43051] [Medline: 37410537]
159. Babington-Ashaye A, de Moerloose P, Diop S, Geissbuhler A. Design, development and usability of an educational AI chatbot for people with haemophilia in Senegal. Haemophilia 2023 Jul 22;29(4):1063-1073. [doi: 10.1111/hae.14815] [Medline: 37347648]
160. Taylor JJ, Subramanian A, Freitas A, Ferreira DM, Dickinson CM. What do individuals with visual impairment need and want from a dialogue-based digital assistant? Clin Exp Optom 2023 Aug 29;106(6):656-665. [doi: 10.1080/08164622.2022.2159791] [Medline: 36709512]
161. Santa-Cruz J, Moran L, Tovar M, Peinado J, Cutipe Y, Ramos L, et al. Mobilizing digital technology to implement a population-based psychological support response during the COVID-19 pandemic in Lima, Peru. Glob Ment Health (Camb) 2022 Jul 28;9:355-365 [FREE Full text] [doi: 10.1017/gmh.2022.36] [Medline: 36618717]
162. Chun-Hung L, Guan-Hsiung L, Wu-Chuan Y, Yu-Hsin L. Chatbot-assisted therapy for patients with methamphetamine use disorder: a preliminary randomized controlled trial. Front Psychiatry 2023 Jul 7;14:1159399 [FREE Full text] [doi: 10.3389/fpsyt.2023.1159399] [Medline: 37484677]
163. Mane HY, Channell Doig A, Marin Gutierrez FX, Jasczynski M, Yue X, Srikanth NP, et al. Practical guidance for the development of Rosie, a health education question-and-answer chatbot for new mothers. J Public Health Manag Pract 2023;29(5):663-670 [FREE Full text] [doi: 10.1097/PHH.0000000000001781] [Medline: 37478093]
164. Cho K, Foo YM, Dalziel B, Hu W. Chatbot-mediated learning of cardiac auscultation. Intern Med J 2022 Dec 17;52(12):2176-2180. [doi: 10.1111/imj.15971] [Medline: 37133367]
165. Bruijnes M, Kesteloo M, Brinkman WP. Reducing social diabetes distress with a conversational agent support system: a three-week technology feasibility evaluation. Front Digit Health 2023 Jun 13;5:1149374 [FREE Full text] [doi: 10.3389/fdgth.2023.1149374] [Medline: 37383944]
166. van der Schyff EL, Ridout B, Amon KL, Forsyth R, Campbell AJ. Providing self-led mental health support through an artificial intelligence-powered chat bot (Leora) to meet the demand of mental health care. J Med Internet Res 2023 Jun 19;25:e46448 [FREE Full text] [doi: 10.2196/46448] [Medline: 37335608]
167. Guo N, Luk TT, Wu YS, Guo Z, Chu JC, Cheung YT, et al. Effect of mobile interventions with nicotine replacement therapy sampling on long-term smoking cessation in community smokers: a pragmatic randomized clinical trial. Tob Induc Dis 2023 Mar 24;21(March):44-13 [FREE Full text] [doi: 10.18332/tid/160168] [Medline: 36969982]
168. Booth F, Potts C, Bond R, Mulvenna M, Kostenius C, Dhanapala I, et al. A mental health and well-being chatbot: user event log analysis. JMIR Mhealth Uhealth 2023 Jul 06;11:e43052 [FREE Full text] [doi: 10.2196/43052] [Medline: 37410539]
169. Anmella G, Sanabra M, Primé-Tous M, Segú X, Cavero M, Morilla I, et al. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: development, feasibility, and potential effectiveness studies. J Med Internet Res 2023 Apr 03;25:e43293 [FREE Full text] [doi: 10.2196/43293] [Medline: 36719325]

XSL•FO
RenderX

170. Fabian KE, Foster KT, Chwastiak L, Turner M, Wagenaar BH. Adapting a transdiagnostic digital mental health intervention for use among immigrant and refugee youth in Seattle: a human-centered design approach. Transl Behav Med 2023 Nov 05;13(11):867-875. [doi: 10.1093/tbm/ibad041] [Medline: 37418614]
171. Massa P, de Souza Ferraz DA, Magno L, Silva AP, Greco M, Dourado I, et al. A transgender chatbot (Amanda Selfie) to create pre-exposure prophylaxis demand among adolescents in Brazil: assessment of acceptability, functionality, usability, and results. J Med Internet Res 2023 Jun 23;25:e41881 [FREE Full text] [doi: 10.2196/41881] [Medline: 37351920]
172. Tapolin FM, Liaskos J, Zoulias E, Mantas J. A conversational web-based chatbot to disseminate COVID-19 advisory information. Stud Health Technol Inform 2023 Jun 29;305:483-486. [doi: 10.3233/SHTI230538] [Medline: 37387072]
173. Durden E, Pirner MC, Rapoport SJ, Williams A, Robinson A, Forman-Hoffman VL. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot". Internet Interv 2023 Sep;33:100637 [FREE Full text] [doi: 10.1016/j.invent.2023.100637] [Medline: 37635948]
174. Pereira DS, Falcão F, Nunes A, Santos N, Costa P, Pêgo JM. Designing and building OSCEBot® for virtual OSCE - performance evaluation. Med Educ Online 2023 Dec 22;28(1):2228550 [FREE Full text] [doi: 10.1080/10872981.2023.2228550] [Medline: 37347808]
175. Brinsley J, Singh B, Maher CA. A digital lifestyle program for psychological distress, wellbeing and return-to-work: a proof-of-concept study. Arch Phys Med Rehabil 2023 Nov;104(11):1903-1912. [doi: 10.1016/j.apmr.2023.04.023] [Medline: 37209933]
176. Matheson EL, Smith HG, Amaral AC, Meireles JF, Almeida MC, Linardon J, et al. Using chatbot technology to improve Brazilian adolescents' body image and mental health at scale: randomized controlled trial. JMIR Mhealth Uhealth 2023 Jun 19;11:e39934 [FREE Full text] [doi: 10.2196/39934] [Medline: 37335604]
177. Walters NL, Lindsey-Mills ZT, Brangan A, Savage SK, Schmidlen TJ, Morgan KM, et al. Facilitating family communication of familial hypercholesterolemia genetic risk: assessing engagement with innovative chatbot technology from the IMPACT-FH study. PEC Innov 2023 Dec;2:100134 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100134] [Medline: 37214500]
178. Jackson-Triche M, Vetal D, Turner EM, Dahiya P, Mangurian C. Meeting the behavioral health needs of health care workers during COVID-19 by leveraging chatbot technology: development and usability study. J Med Internet Res 2023 Jun 08;25:e40635 [FREE Full text] [doi: 10.2196/40635] [Medline: 37146178]
179. Nehme M, Schneider F, Perrin A, Sum Yu W, Schmitt S, Violot G, et al. The development of a chatbot technology to disseminate post-COVID-19 information: descriptive implementation study. J Med Internet Res 2023 Jun 05;25:e43113 [FREE Full text] [doi: 10.2196/43113] [Medline: 37195688]
180. Ko SQ, Chua CM, Koh SH, Lim YW, Shorey S. Experiences of patients and their caregivers in a virtual ward in Singapore: a descriptive qualitative study. Int J Med Inform 2023 Sep;177:105111 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105111] [Medline: 37307721]
181. Kobayashi T, Tomoi H, Nishina Y, Harada K, Tanaka K, Sasaki S, et al. Effect of a mobile app chatbot and an interactive small-group webinar on COVID-19 vaccine intention and confidence in Japan: a randomised controlled trial. BMJ Glob Health 2023 May 29;8(5):e010370 [FREE Full text] [doi: 10.1136/bmjgh-2022-010370] [Medline: 37247873]
182. Lee KY, Dabak SV, Kong VH, Park M, Kwok SL, Silzle M, et al. Effectiveness of chatbots on COVID vaccine confidence and acceptance in Thailand, Hong Kong, and Singapore. NPJ Digit Med 2023 May 25;6(1):96 [FREE Full text] [doi: 10.1038/s41746-023-00843-6] [Medline: 37231110]
183. Sabour S, Zhang W, Xiao X, Zhang Y, Zheng Y, Wen J, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023 May 4;5:1133987 [FREE Full text] [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
184. Wang Q, Peng S, Zha Z, Han X, Deng C, Hu L, et al. Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. Front Psychiatry 2023 Apr 17;14:1148534 [FREE Full text] [doi: 10.3389/fpsyt.2023.1148534] [Medline: 37139323]
185. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022 Apr 11;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
186. Bézie A, Morisseau V, Rolland R, Guillemassé A, Brouard B, Chaix B. Using a chatbot to study medication overuse among patients suffering from headaches. Front Digit Health 2022 Mar 17;4:801782 [FREE Full text] [doi: 10.3389/fdgth.2022.801782] [Medline: 35373183]
187. Moilanen J, van Berkel N, Visuri A, Gadiraju U, van der Maden W, Hosio S. Supporting mental health self-care discovery through a chatbot. Front Digit Health 2023 Mar 7;5:1034724 [FREE Full text] [doi: 10.3389/fdgth.2023.1034724] [Medline: 36960179]
188. Burger F, Neerincx MA, Brinkman WP. Using a conversational agent for thought recording as a cognitive therapy task: feasibility, content, and feedback. Front Digit Health 2022 Jul 19;4:930874 [FREE Full text] [doi: 10.3389/fdgth.2022.930874] [Medline: 35928046]
189. Soni H, Ivanova J, Wilczewski H, Bailey A, Ong T, Narma A, et al. Virtual conversational agents versus online forms: patient experience and preferences for health data collection. Front Digit Health 2022 Oct 13;4:954069 [FREE Full text] [doi: 10.3389/fdgth.2022.954069] [Medline: 36310920]

XSL•FO
RenderX

190. Crovari P, Pidò S, Pinoli P, Bernasconi A, Canakoglu A, Garzotto F, et al. GeCoAgent: a conversational agent for empowering genomic data extraction and analysis. ACM Trans Comput Healthc 2021 Oct 15;3(1):1-29. [doi: 10.1145/3464383]
191. Dwyer T, Hoit G, Burns D, Higgins J, Chang J, Whelan D, et al. Use of an artificial intelligence conversational agent (Chatbot) for hip arthroscopy patients following surgery. Arthrosc Sports Med Rehabil 2023 Apr;5(2):e495-e505 [FREE Full text] [doi: 10.1016/j.asmr.2023.01.020] [Medline: 37101866]
192. Wilczewski H, Soni H, Ivanova J, Ong T, Barrera JF, Bunnell BE, et al. Older adults' experience with virtual conversational agents for health data collection. Front Digit Health 2023 Mar 15;5:1125926 [FREE Full text] [doi: 10.3389/fdgth.2023.1125926] [Medline: 37006821]
193. Montenegro JL, da Costa CA, Janssen LP. Evaluating the use of chatbot during pregnancy: a usability study. Healthc Anal 2022 Nov;2:100072. [doi: 10.1016/j.health.2022.100072]
194. Ko SQ, Kumar SK, Jacob J, Hooi BM, Soo M, Nashi N, et al. Technology-enabled virtual ward for COVID management of the elderly and immunocompromised in Singapore: a descriptive cohort. BMC Infect Dis 2023 Feb 21;23(1):102 [FREE Full text] [doi: 10.1186/s12879-023-08040-2] [Medline: 36809977]
195. Hsu MH, Chan TM, Yu CS. Termbot: a chatbot-based crossword game for gamified medical terminology learning. Int J Environ Res Public Health 2023 Feb 26;20(5):4185 [FREE Full text] [doi: 10.3390/ijerph20054185] [Medline: 36901193]
196. Epalte K, Tomsone S, Vētra A, Bērziņa G. Patient experience using digital therapy "Vigo" for stroke patient recovery: a qualitative descriptive study. Disabil Rehabil Assist Technol 2023 Feb 06;18(2):175-184. [doi: 10.1080/17483107.2020.1839794] [Medline: 33155507]
197. Kobayashi T, Nishina Y, Tomoi H, Harada K, Tanaka K, Matsumoto E, et al. Corowa-kun: a messenger app chatbot delivers COVID-19 vaccine information, Japan 2021. Vaccine 2022 Jul 30;40(32):4654-4662 [FREE Full text] [doi: 10.1016/j.vaccine.2022.06.007] [Medline: 35750541]
198. Chin H, Lima G, Shin M, Zhunis A, Cha C, Choi J, et al. User-chatbot conversations during the COVID-19 pandemic: study based on topic modeling and sentiment analysis. J Med Internet Res 2023 Jan 27;25:e40922 [FREE Full text] [doi: 10.2196/40922] [Medline: 36596214]
199. Albers N, Hizli B, Scheltinga BL, Meijer E, Brinkman WP. Setting physical activity goals with a virtual coach: vicarious experiences, personalization and acceptance. J Med Syst 2023 Jan 30;47(1):15 [FREE Full text] [doi: 10.1007/s10916-022-01899-9] [Medline: 36710276]
200. Kaywan P, Ahmed K, Ibaida A, Miao Y, Gu B. Early detection of depression using a conversational AI bot: a non-clinical trial. PLoS One 2023 Feb 3;18(2):e0279743 [FREE Full text] [doi: 10.1371/journal.pone.0279743] [Medline: 36735701]
201. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719 [FREE Full text] [doi: 10.2196/40719] [Medline: 36355633]
202. Castilla E, Escobar JJ, Villalonga C, Banos O. HIGEA: an intelligent conversational agent to detect caregiver burden. Int J Environ Res Public Health 2022 Nov 30;19(23):16019 [FREE Full text] [doi: 10.3390/ijerph192316019] [Medline: 36498092]
203. Chen T, Chen L, Li S, Du J, Su H, Jiang H, et al. Virtual digital psychotherapist app-based treatment in patients with methamphetamine use disorder (Echo-APP): single-arm pilot feasibility and efficacy study. JMIR Mhealth Uhealth 2023 Jan 31;11:e40373 [FREE Full text] [doi: 10.2196/40373] [Medline: 36719727]
204. Dosovitsky G, Bunge E. Development of a chatbot for depression: adolescent perceptions and recommendations. Child Adolesc Ment Health 2023 Feb 11;28(1):124-127. [doi: 10.1111/camh.12627] [Medline: 36507594]
205. Montenegro JL, da Costa CA, da Rosa Righi R, Farias ER, Matté LB. Development and validation of conversational agent to pregnancy safe-education. J Med Syst 2023 Jan 10;47(1):7. [doi: 10.1007/s10916-022-01903-2] [Medline: 36626106]
206. Ntinga X, Musiello F, Keter AK, Barnabas R, van Heerden A. The feasibility and acceptability of an mHealth conversational agent designed to support HIV self-testing in South Africa: cross-sectional study. J Med Internet Res 2022 Dec 12;24(12):e39816 [FREE Full text] [doi: 10.2196/39816] [Medline: 36508248]
207. Rodriguez-Arrastia M, Martinez-Ortigosa A, Ruiz-Gonzalez C, Ropero-Padilla C, Roman P, Sanchez-Labraca N. Experiences and perceptions of final-year nursing students of using a chatbot in a simulated emergency situation: a qualitative study. J Nurs Manag 2022 Nov 18;30(8):3874-3884 [FREE Full text] [doi: 10.1111/jonm.13630] [Medline: 35411629]
208. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ 2022 Dec 01;22(1):830 [FREE Full text] [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
209. Singh E, Bompelli A, Wan R, Bian J, Pakhomov S, Zhang R. A conversational agent system for dietary supplements use. BMC Med Inform Decis Mak 2022 Jul 07;22(Suppl 1):153 [FREE Full text] [doi: 10.1186/s12911-022-01888-5] [Medline: 35799177]
210. Andriani SP, Adhyanacarira P, Fuad A, Pertiwi AA. Comparison of non-AI and AI-enabled mHealth platforms for COVID-19 self screening in Indonesia. Stud Health Technol Inform 2022 Jun 29;295:226-229. [doi: 10.3233/SHTI220703] [Medline: 35773849]
211. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano ME, Martinez-Suberviola FJ, Dejal@ Group. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter,

XSL•FO
RenderX

controlled, randomized clinical trial in primary care. JMIR Mhealth Uhealth 2022 Jun 27;10(6):e34273 [FREE Full text] [doi: 10.2196/34273] [Medline: 35759328]
212. Rizzato Lede DA, Inda D, Rosa JM, Zin Y, Tentoni N, Médici MM, et al. Tana, a healthcare chatbot to help patients during the COVID-19 pandemic at a university hospital in Argentina. Stud Health Technol Inform 2022 Jun 06;290:301-303. [doi: 10.3233/SHTI220083] [Medline: 35673022]
213. Denecke K, Lombardo P, Nairz K. Digital medical interview assistant for radiology: opportunities and challenges. Stud Health Technol Inform 2022 May 16;293:39-46 [FREE Full text] [doi: 10.3233/SHTI220345] [Medline: 35592958]
214. Ben-Shabat N, Sharvit G, Meimis B, Joya DB, Sloma A, Kiderman D, et al. Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study. Int J Med Inform 2022 Dec;168:104897 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104897] [Medline: 36306653]
215. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh ML, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022 Mar 28;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
216. Pithpornchaiyakul S, Naorungroj S, Pupong K, Hunsrisakhun J. Using a chatbot as an alternative approach for in-person toothbrushing training during the COVID-19 pandemic: comparative study. J Med Internet Res 2022 Oct 21;24(10):e39218 [FREE Full text] [doi: 10.2196/39218] [Medline: 36179147]
217. Yang LW, Ng WY, Lei X, Tan SC, Wang Z, Yan M, et al. Development and testing of a multi-lingual natural language processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: a multi-center study. Front Public Health 2023 Feb 13;11:1063466 [FREE Full text] [doi: 10.3389/fpubh.2023.1063466] [Medline: 36860378]
218. Suárez A, Adanero A, Díaz-Flores García V, Freire Y, Algar J. Using a virtual patient via an artificial intelligence chatbot to develop dental students' diagnostic skills. Int J Environ Res Public Health 2022 Jul 18;19(14):8735 [FREE Full text] [doi: 10.3390/ijerph19148735] [Medline: 35886584]
219. Ogawa M, Oyama G, Morito K, Kobayashi M, Yamada Y, Shinkawa K, et al. Can AI make people happy? The effect of AI-based chatbot on smile and speech in Parkinson's disease. Parkinsonism Relat Disord 2022 Jun;99:43-46 [FREE Full text] [doi: 10.1016/j.parkreldis.2022.04.018] [Medline: 35596975]
220. Vasileiou MV, Maglogiannis IG. The health chatbots in telemedicine: intelligent dialog system for remote support. J Healthc Eng 2022 Oct 6;2022:1-12. [doi: 10.1155/2022/4876512]
221. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022 Sep 23;9(9):e38067 [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
222. Luk TT, Lui JH, Wang MP. Efficacy, usability, and acceptability of a chatbot for promoting COVID-19 vaccination in unvaccinated or booster-hesitant young adults: pre-post pilot study. J Med Internet Res 2022 Oct 04;24(10):e39063 [FREE Full text] [doi: 10.2196/39063] [Medline: 36179132]
223. Martinengo L, Lum E, Car J. Evaluation of chatbot-delivered interventions for self-management of depression: content analysis. J Affect Disord 2022 Dec 15;319:598-607. [doi: 10.1016/j.jad.2022.09.028] [Medline: 36150405]
224. Shah J, DePietro B, D'Adamo L, Firebaugh ML, Laing O, Fowler LA, et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. Int J Eat Disord 2022 Sep 18;55(9):1229-1244 [FREE Full text] [doi: 10.1002/eat.23798] [Medline: 36056648]
225. De Marchi F, Serioli M, Collo A, Belotti EG, Alloatti F, Biroli G, et al. A telehealth intervention for nutritional counseling in amyotrophic lateral sclerosis patients. J Clin Med 2022 Jul 23;11(15):4286 [FREE Full text] [doi: 10.3390/jcm11154286] [Medline: 35893377]
226. Rabinowitz AR, Collier G, Vaccaro M, Wingfield R. Development of RehaBot-a conversational agent for promoting rewarding activities in users with traumatic brain injury. J Head Trauma Rehabil 2022;37(3):144-151. [doi: 10.1097/HTR.0000000000000770] [Medline: 35293365]
227. Alturaiki AM, Banjar HR, Barefah AS, Alnajjar SA, Hindawi S. A smart chatbot for interactive management in beta thalassemia patients. Int J Telemed Appl 2022 May 11;2022:9734518-9734513 [FREE Full text] [doi: 10.1155/2022/9734518] [Medline: 35601050]
228. Terblanche N, Molyn J, de Haan E, Nilsson VO. Comparing artificial intelligence and human coaching goal attainment efficacy. PLoS One 2022 Jun 21;17(6):e0270255 [FREE Full text] [doi: 10.1371/journal.pone.0270255] [Medline: 35727801]
229. Nam KH, Kim DY, Kim DH, Lee JH, Lee JI, Kim MJ, et al. Conversational artificial intelligence for spinal pain questionnaire: validation and user satisfaction. Neurospine 2022 Jun;19(2):348-356 [FREE Full text] [doi: 10.14245/ns.2143080.540] [Medline: 35577340]
230. Nißen M, Rüegger D, Stieger M, Flückiger C, Allemand M, V Wangenheim F, et al. The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: development of a design codebook and web-based study. J Med Internet Res 2022 Apr 27;24(4):e32630 [FREE Full text] [doi: 10.2196/32630] [Medline: 35475761]
231. Larbi D, Denecke K, Gabarron E. Usability testing of a social media chatbot for increasing physical activity behavior. J Pers Med 2022 May 20;12(5):828 [FREE Full text] [doi: 10.3390/jpm12050828] [Medline: 35629252]

XSL•FO
RenderX

232. Ritchie JB, Frey LJ, Lamy JB, Bellcross C, Morrison H, Schiffman JD, et al. Automated clinical practice guideline recommendations for hereditary cancer risk using chatbots and ontologies: system description. JMIR Cancer 2022 Jan 31;8(1):e29289 [FREE Full text] [doi: 10.2196/29289] [Medline: 35099392]
233. Hope DL, Grant GD, Rogers GD, King MA. Virtualized gamified pharmacy simulation during COVID-19. Pharmacy (Basel) 2022 Mar 26;10(2):41 [FREE Full text] [doi: 10.3390/pharmacy10020041] [Medline: 35448700]
234. Munsch N, Gruarin S, Nateqi J, Lutz T, Binder M, Aberle JH, et al. Symptoms associated with a COVID-19 infection among a non-hospitalized cohort in Vienna. Wien Klin Wochenschr 2022 May 13;134(9-10):344-350 [FREE Full text] [doi: 10.1007/s00508-022-02028-9] [Medline: 35416543]
235. Minutolo A, Damiano E, De Pietro G, Fujita H, Esposito M. A conversational agent for querying Italian patient information leaflets and improving health literacy. Comput Biol Med 2022 Feb;141:105004. [doi: 10.1016/j.compbiomed.2021.105004] [Medline: 34774337]
236. Hurmuz MZ, Jansen-Kosterink SM, Beinema T, Fischer K, Op den Akker H, Hermens HJ. Evaluation of a virtual coaching system eHealth intervention: a mixed methods observational cohort study in the Netherlands. Internet Interv 2022 Mar;27:100501 [FREE Full text] [doi: 10.1016/j.invent.2022.100501] [Medline: 35198411]
237. Beinema T, Op den Akker H, Hurmuz M, Jansen-Kosterink S, Hermens H. Automatic topic selection for long-term interaction with embodied conversational agents in health coaching: a micro-randomized trial. Internet Interv 2022 Mar;27:100502 [FREE Full text] [doi: 10.1016/j.invent.2022.100502] [Medline: 35198412]
238. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022 Mar;27:100495 [FREE Full text] [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
239. Morgan KM, Hamilton JG, Symecko H, Kamara D, Jenkins C, Lester J, et al. Targeted BRCA1/2 population screening among Ashkenazi Jewish individuals using a web-enabled medical model: an observational cohort study. Genet Med 2022 Mar;24(3):564-575 [FREE Full text] [doi: 10.1016/j.gim.2021.10.016] [Medline: 34906490]
240. Dhinagaran DA, Martinengo L, Ho MH, Joty S, Kowatsch T, Atun R, et al. Designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER): development of a conceptual framework. JMIR Mhealth Uhealth 2022 Oct 04;10(10):e38740 [FREE Full text] [doi: 10.2196/38740] [Medline: 36194462]
241. Schmidlen T, Jones CL, Campbell-Salome G, McCormick CZ, Vanenkevort E, Sturm AC. Use of a chatbot to increase uptake of cascade genetic testing. J Genet Couns 2022 Oct 26;31(5):1219-1230. [doi: 10.1002/jgc4.1592] [Medline: 35616645]
242. Wang WT, Tan N, Hanson JA, Crubaugh CA, Hara AK. Initial experience with a COVID-19 screening chatbot before radiology appointments. J Digit Imaging 2022 Oct 13;35(5):1303-1307 [FREE Full text] [doi: 10.1007/s10278-022-00650-7] [Medline: 35562634]
243. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel) 2022 May 11;22(10):3653 [FREE Full text] [doi: 10.3390/s22103653] [Medline: 35632061]
244. Miura C, Chen S, Saiki S, Nakamura M, Yasuda K. Assisting personalized healthcare of elderly people: developing a rule-based virtual caregiver system using mobile chatbot. Sensors (Basel) 2022 May 18;22(10):3829 [FREE Full text] [doi: 10.3390/s22103829] [Medline: 35632238]
245. Asensio-Cuesta S, Blanes-Selva V, Conejero A, Portolés M, García-Gómez M. A user-centered chatbot to identify and interconnect individual, social and environmental risk factors related to overweight and obesity. Inform Health Soc Care 2022 Jan 02;47(1):38-52. [doi: 10.1080/17538157.2021.1923501] [Medline: 34032537]
246. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health 2006 Apr 01;27(1):167-194. [doi: 10.1146/annurev.publhealth.27.021405.102103] [Medline: 16533114]
247. Epstein RM, Street RL. The values and value of patient-centered care. Ann Fam Med 2011 Mar 14;9(2):100-103 [FREE Full text] [doi: 10.1370/afm.1239] [Medline: 21403134]
248. Shaller D. Patient-centered care: what does it take? The Commonwealth Fund. 2007. URL: https://www.commonwealthfund.org/publications/fund-reports/2007/oct/patient-centered-care-what-does-it-take [accessed 2024-04-29]
249. Efficient definition and meaning. Merriam-Webster. 2022. URL: https://www.merriam-webster.com/dictionary/efficient [accessed 2024-04-29]
250. The world by income and region. The World Bank. 2022. URL: https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html [accessed 2024-04-29]
251. Primary health care. World Health Organization. URL: https://www.who.int/health-topics/primary-health-care [accessed 2024-04-29]
252. Frangoudes F, Hadjiaros M, Schiza EC, Matsangidou M, Tsivitanidou O, Neokleous K. An overview of the use of chatbots in medical and healthcare education. In: Proceedings of 8th International Conference, LCT 2021, Held as Part of the 23rd HCI International Conference, HCII 2021 on Learning and Collaboration Technologies: Games and Virtual Environments for Learning. 2021 Presented at: HCII '22; July 24-29, 2021; Virtual Event p. 170-184 URL: https://link.springer.com/chapter/10.1007/978-3-030-77943-6_11 [doi: 10.1007/978-3-030-77943-6_11]

XSL•FO
RenderX

253. Thakre K, Rothe PR, Kukade S, Shinde P, Madame K. Health care chatbot using NLP and flask. Ijraset Int J Res Appl Sci Eng Technol 2022 May 31;10(5):2632-2654. [doi: 10.22214/ijraset.2022.42854]
254. Sheth A, Yip HY, Shekarpour S. Extending patient-chatbot experience with internet-of-things and background knowledge: case studies with healthcare applications. IEEE Intell Syst 2019 Jul 1;34(4):24-30 [FREE Full text] [doi: 10.1109/mis.2019.2905748] [Medline: 34690576]
255. Roca S, Sancho J, García J, Alesanco Á. Microservice chatbot architecture for chronic patient support. J Biomed Inform 2020 Feb;102:103305 [FREE Full text] [doi: 10.1016/j.jbi.2019.103305] [Medline: 31622802]
256. Vryoni V, Bailey S. Chatbots in healthcare: connecting patients to information: emerging health technologies. Canadian Agency for Drugs and Technologies in Health. 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK602381/ [accessed 2024-04-29]
257. Jadczyk T, Wojakowski W, Tendera M, Henry TD, Egnaczyk G, Shreenivas S. Artificial intelligence can improve patient management at the time of a pandemic: the role of voice technology. J Med Internet Res 2021 May 25;23(5):e22959 [FREE Full text] [doi: 10.2196/22959] [Medline: 33999834]
258. Fadhil A. Beyond patient monitoring: conversational agents role in telemedicine and healthcare support for home-living elderly individuals. arXiv Preprint posted online March 3, 2018 [FREE Full text]
259. Bidve V, Virkar A, Raut P, Velapurkar S. NOVA-a virtual nursing assistant. Indones J Electr Eng Comput Sci 2023 Apr 01;30(1):307-315. [doi: 10.11591/ijeecs.v30.i1.pp307-315]
260. Stone P, Brooks R, Brynjolfsson E, Calo R, Etzioni O, Hager G, et al. Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. arXiv Preprint posted online October 31, 2022 [FREE Full text] [doi: 10.48550/arXiv.2211.06318]
261. Nishida T, Nakazawa A, Ohmoto Y, Mohammad Y. Conversational Informatics: A Data-Intensive Approach with Emphasis on Nonverbal Communication. Cham, Switzerland: Springer; 2014.
262. Wolters MK, Kelly F, Kilgour J. Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. Health Informatics J 2016 Dec 26;22(4):854-866 [FREE Full text] [doi: 10.1177/1460458215593329] [Medline: 26276794]
263. Tzelios C, Contreras C, Istenes B, Astupillo A, Lecca L, Ramos K, et al. Using digital chatbots to close gaps in healthcare access during the COVID-19 pandemic. Public Health Action 2022 Dec 21;12(4):180-185 [FREE Full text] [doi: 10.5588/pha.22.0046] [Medline: 36561900]
264. Vijayarani M, Balamurugan G. Chatbot in mental health care. Indian J Psychiatry Nurse 2019;16(2):126-128. [doi: 10.4103/iopn.iopn_34_19]
265. Peng ML, Wickersham JA, Altice FL, Shrestha R, Azwa I, Zhou X, et al. Formative evaluation of the acceptance of HIV prevention artificial intelligence chatbots by men who have sex with men in Malaysia: focus group study. JMIR Form Res 2022 Oct 06;6(10):e42055 [FREE Full text] [doi: 10.2196/42055] [Medline: 36201390]
266. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-39. [doi: 10.1145/3485874]
267. Ahmed A, Hassan A, Aziz S, Abd-Alrazaq AA, Ali N, Alzubaidi M, et al. Chatbot features for anxiety and depression: a scoping review. Health Informatics J 2023 Jan 24;29(1):14604582221146719 [FREE Full text] [doi: 10.1177/14604582221146719] [Medline: 36693014]
268. Mehta A, Virkar S, Khatri J, Thakur R, Dalvi A. Artificial intelligence powered chatbot for mental healthcare based on sentiment analysis. In: Proceedings of the 2022 5th International Conference on Advances in Science and Technology. 2022 Presented at: ICAST '22; December 2-3, 2022; Mumbai, India p. 185-189 URL: https://ieeexplore.ieee.org/document/10039548 [doi: 10.1109/icast55766.2022.10039548]
269. Lucas GM, Gratch J, King A, Morency LP. It's only a computer: virtual humans increase willingness to disclose. Comput Human Behav 2014 Aug;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
270. Trost MJ, Chrysilla G, Gold JI, Matarić M. Socially-assistive robots using empathy to reduce pain and distress during peripheral IV placement in children. Pain Res Manag 2020 Apr 09;2020:7935215-7935217 [FREE Full text] [doi: 10.1155/2020/7935215] [Medline: 32351642]
271. Suzuki Y, Galli L, Ikeda A, Itakura S, Kitazaki M. Measuring empathy for human and robot hand pain using electroencephalography. Sci Rep 2015 Nov 03;5:15924. [doi: 10.1038/srep15924] [Medline: 26525705]
272. Chita-Tegmark M, Ackerman JM, Scheutz M. Effects of assistive robot behavior on impressions of patient psychological attributes: vignette-based human-robot interaction study. J Med Internet Res 2019 May 19;21(6):e13729 [FREE Full text] [doi: 10.2196/13729] [Medline: 31199297]
273. Bulla C, Parushetti C, Teli A, Aski S, Koppad S. A review of AI based medical assistant chatbot: research and applications of web development and design. Res Appl Web Develop Des 2020;3(2):1-4 [FREE Full text]
274. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):625-636. [doi: 10.1089/cyber.2018.0110] [Medline: 30334655]
275. Kao CH, Chen CC, Tsai YT. Model of multi-turn dialogue in emotional chatbot. In: Proceedings of the 2019 International Conference on Technologies and Applications of Artificial Intelligence. 2019 Presented at: TAAi '19; November 21-23, 2019; Kaohsiung, Taiwan p. 1-5 URL: https://ieeexplore.ieee.org/document/8959855 [doi: 10.1109/taai48200.2019.8959855]

XSL•FO
RenderX

276. Purushotham K, Priya KS, Jayasree K, Sekhar GS, Jaswanth N, Kumar KK. Automated conversation chatbot for multiple languages for hospitals. Int J Adv Res Sci Commun Technol 2023 Jan 10;3(1):209-217 [FREE Full text] [doi: 10.48175/ijarsct-7859]
277. Hernández-Neuta I, Neumann F, Brightmeyer J, Ba Tis T, Madaboosi N, Wei Q, et al. Smartphone-based clinical diagnostics: towards democratization of evidence-based health care. J Intern Med 2019 Jan 12;285(1):19-39 [FREE Full text] [doi: 10.1111/joim.12820] [Medline: 30079527]
278. Iancu I, Iancu B. Interacting with chatbots later in life: a technology acceptance perspective in COVID-19 pandemic situation. Front Psychol 2022 Jan 16;13:1111003 [FREE Full text] [doi: 10.3389/fpsyg.2022.1111003] [Medline: 36726494]
279. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec 21;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
280. Brown JE, Halpern J. AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. SSM Mental Health 2021 Dec;1:100017. [doi: 10.1016/j.ssmmh.2021.100017]
281. Whitby B. The ethical implications of non-human agency in health care. In: Proceedings of the 2014 Machine Ethics in the Context of Medical and Care Agents. 2014 Presented at: MEMCA '14; April 1-4, 2014; London, UK p. 1-4 URL: https://doc.gold.ac.uk/aisb50/AISB50-S17/AISB50-S17-Whitby-Paper.pdf
282. Chaix B, Guillemassé A, Nectoux P, Delamon G, Brouard B. Vik: a chatbot to support patients with chronic diseases. Health 2020;12(07):804-810. [doi: 10.4236/health.2020.127058]
283. Srilakshmi V, Prathima V. Disease prediction and diagnosis using machine learning. Int J Res Appl Sci Eng Technol 2022 Jul 31;10(7):1594-1597. [doi: 10.22214/ijraset.2022.45450]
284. Jovanovic M, Baez M, Casati F. Chatbots as conversational healthcare services. IEEE Internet Comput 2021 May;25(3):44-51. [doi: 10.1109/mic.2020.3037151]
285. Murtarelli G, Gregory A, Romenti S. A conversation-based perspective for shaping ethical human–machine interactions: the particular challenge of chatbots. J Bus Res 2021 May;129:927-935. [doi: 10.1016/j.jbusres.2020.09.018]
286. Følstad A, Araujo T, Law EL, Brandtzaeg PB, Papadopoulos S, Reis L, et al. Future directions for chatbot research: an interdisciplinary research agenda. Computing 2021 Oct 19;103(12):2915-2942. [doi: 10.1007/s00607-021-01016-7]
287. Cameron G, Cameron D, Megaw G, Bond RR, Mulvenna MD, O'Neill S, et al. Best practices for designing chatbots in mental healthcare: a case study on iHelpr. In: Proceedings of the 32nd International BCS Human Computer Interaction Conference. 2018 Presented at: HCI '18; July 4-6, 2018; Belfast, UK p. 1-5 URL: https://dl.acm.org/doi/10.14236/ewic/HCI2018.129 [doi: 10.14236/ewic/hci2018.129]
288. Stiefel S. 'The chatbot will see you now': mental health confidentiality concerns in software therapy. SSRN Journal Preprint posted online July 19, 2018 [FREE Full text] [doi: 10.2139/ssrn.3166640]
289. Watt A, Cameron A, Sturm L, Lathlean T, Babidge W, Blamey S, et al. Rapid versus full systematic reviews: validity in clinical practice? ANZ J Surg 2008 Nov 21;78(11):1037-1040. [doi: 10.1111/j.1445-2197.2008.04730.x] [Medline: 18959712]
290. Kandpal P, Jasnani K, Raut R, Bhorge S. Contextual chatbot for healthcare purposes (using deep learning). In: Proceedings of the 2020 4th World Conference on Smart Trends in Systems, Security and Sustainability. 2020 Presented at: WorldS4 '20; July 27-28, 2020; London, UK p. 625-634 URL: https://ieeexplore.ieee.org/document/9210351 [doi: 10.1109/worlds450073.2020.9210351]
291. Pernencar C, Saboia I, Dias JC. How far can conversational agents contribute to IBD patient health care-a review of the literature. Front Public Health 2022 Jun 30;10:862432 [FREE Full text] [doi: 10.3389/fpubh.2022.862432] [Medline: 35844879]
292. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 01;27(7):1147-1148 [FREE Full text] [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
293. Lorence DP, Park H, Fox S. Racial disparities in health information access: resilience of the Digital Divide. J Med Syst 2006 Aug 7;30(4):241-249. [doi: 10.1007/s10916-005-9003-y] [Medline: 16978003]
294. Cornejo Müller A, Wachtler B, Lampert T. Digital divide-social inequalities in the utilisation of digital healthcare. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb 08;63(2):185-191 [FREE Full text] [doi: 10.1007/s00103-019-03081-y] [Medline: 31915863]
295. Tett G. When algorithms reinforce inequality. Financial Times. 2018. URL: https://www.ft.com/content/fb583548-0b93-11e8-839d-41ca06376bf2 [accessed 2024-04-29]
296. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
297. Kim HW, Kankanhalli A. Investigating user resistance to information systems implementation: a status quo bias perspective. MIS Q 2009;33(3):567-582. [doi: 10.2307/20650309]
298. Best L, Stevens A, Colin-Jones D. Rapid and responsive health technology assessment: the development and evaluation process in the South and West region of England. J Clin Eff 1997;2(2):51-56. [doi: 10.1108/eb020865]

XSL•FO
RenderX

299. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clin Proc Digit Health 2023 Sep;1(3):226-234 [FREE Full text] [doi: 10.1016/j.mcpdig.2023.05.004]

Abbreviations

AI: artificial intelligence
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by YH Lin; submitted 02.02.24; peer-reviewed by J Xie, R Forsyth; comments to author 18.03.24; revised version received 07.04.24; accepted 12.04.24; published 23.07.24.

Please cite as:
Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B
Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review
J Med Internet Res 2024;26:e56930
URL: https://www.jmir.org/2024/1/e56930
doi:10.2196/56930
PMID:

©Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, Bertrand Lebouché. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Supportive eHealth Technologies and Their Effects on Physical Functioning and Quality of Life for People With Lung Cancer: Systematic Review

Suriya Kirkpatrick[1], BA, MSc; Zoe Davey[1*], PhD; Peter Richard Wright[1*], PhD; Catherine Henshall[1*], PhD

School of Nursing, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Suriya Kirkpatrick, BA, MSc
School of Nursing
Faculty of Health and Life Sciences
Oxford Brookes University
Headington Rd, Headington
Oxford, OX3 0BP
United Kingdom
Phone: 44 07909921833
Email: 19228607@brookes.ac.uk

Abstract

Background: Despite advancements in treatment and early diagnosis, people with lung cancer are not living as long as those with other cancers. The more common symptoms of lung cancer, such as breathlessness, fatigue, and depression, can be alleviated by improving patients' physical functioning. Therefore, good symptom management and improved health-related quality of life (HRQoL) are priorities in this patient group. However, current health care services have limited capacity to provide this support. One way to address this issue of health care resources is to empower patients to self-manage their condition using eHealth technologies.

Objective: The purpose of this review was to identify and assess available research on technologies that support persons with lung cancer to improve or maintain their physical functioning, HRQoL, or both.

Methods: Six databases—PubMed, Web of Science, CINAHL, MEDLINE, SPORTDiscus, and PsycINFO—were searched from January 1, 1990, to April 30, 2023. Studies were suitable for inclusion if the participants included people with lung cancer aged >18 years who had been exposed to a physical activity, exercise, or training intervention that was delivered via an electronic or web-based application with or without a comparator. Furthermore, the study had to report on the impact of the intervention on physical functioning and HRQoL. Studies that focused on telemedicine without a digital intervention were excluded. The Grading of Recommendations Assessment, Development, and Evaluation system was used to assess the quality of the included papers. Due to the heterogeneity of the studies, a narrative synthesis was undertaken.

Results: This review is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. A total of 794 papers were initially identified through our search, of which, after screening, 8 (1%) were confirmed suitable for inclusion in the review. As 2 (25%) of the 8 papers reported on different stages of the same study, we included only 7 studies in our analysis. The studies were undertaken between 2010 and 2018 across multiple countries and aimed to develop a technology and test its feasibility or acceptance. The 7 technologies identified included web-based applications, mobile apps, and gaming consoles. The studies demonstrated impact on walking distance, muscle strength, balance, dyspnea symptoms, and cancer-related fatigue. HRQoL scores also showed improvement.

Conclusions: The findings indicate that eHealth technologies can positively impact physical functioning and well-being for people with lung cancer, but there are limited studies that demonstrate the impact of these digital interventions over longer periods. None of the studies reported on the implementation or adoption of a mobile health or eHealth intervention in routine clinical practice, highlighting the need for further research in this area.

Trial Registration: PROSPERO CRD42023414094; https://tinyurl.com/39hhbwyx

(J Med Internet Res 2024;26:e53015) doi:10.2196/53015

XSL•FO
RenderX

KEYWORDS

lung cancer; physical activity; exercise; physical functioning; mobile technology; smartphone apps; digital health; mobile phone

Introduction

Background

Lung cancer is the most prevalent cancer globally, with 2.21 million new cases being diagnosed in 2020; this is anticipated to increase to 3.8 million by 2050 [1]. Lung cancer also accounts for the highest number of cancer-related deaths across all cancer types [2].

In the United Kingdom, approximately 48,500 new lung cancer cases are diagnosed per year [3]. The incidence of lung cancer strongly correlates with age, peaking among older individuals. In the United Kingdom, from 2016 to 2018, >44% of new cases annually were in those aged ≥75 years. Age-specific incidence rates rise sharply from around ages 45 to 49 years, reaching a peak in women aged 75 to 79 years and men aged 85 to 89 years, and then decline in older age groups. Women typically have lower incidence rates than men, particularly evident at age ≥90 years, where the rate in women is half that of men [3]. One-year survival rates have almost doubled since the 1970s due to early diagnosis and improved treatments. However, lung cancer survival rates at 5 and 10 years have not improved as much as those for other cancers [3].

For people living with lung cancer, it is imperative that supportive care needs, which are central to patient-centered care [4], are addressed promptly because their condition is associated with a high symptom burden and high levels of unmet needs throughout the disease trajectory [5]. In addition, approximately two-thirds of people with lung cancer have at least 1 other preexisting health condition, and up to half have ≥2 preexisting health conditions [6]. Addressing their supportive care needs will contribute to efficient use of health care resources and minimize hospital admissions. If not managed effectively, this could negatively impact patient outcomes, including physical functioning, psychological well-being, and health-related quality of life (HRQoL) [4].

Common symptoms of lung cancer include fatigue, breathlessness, and depression—all of which can be alleviated by exercise interventions [7]. More generally, other positive implications for people with cancer undertaking physical activity include improvements in HRQoL [8-11], lung function [8,10], sleep [8,12], immune function and markers [8,13], mood [8,9], bone strength [8,14], and muscle mass [8,15], as well as decreased cancer cell proliferation [8,13]. However, less than one-third of people with lung cancer meet recommended exercise guidelines to reduce time spent sedentary, increase strength- and balance-building activities, and undertake 150 minutes of aerobic activity per week [16].

Self-management practices, including those with an exercise component, can help patients with cancer to regain health and fitness, reduce side effects from treatment and symptoms of the disease, relax the mind and body, enhance HRQoL, and regain a sense of normality [17]. More recently, the National Institute for Health and Care Excellence has recognized exercise as a first-line treatment within health care and holistic disease management [18]. In the absence of a robust national rehabilitation system, there is pressure for self-management support to be integrated into routine cancer care [19]. However, patient adherence to rehabilitation programs delivered at hospital outpatient centers can be low due to the required travel and associated socioeconomic factors [20]. Furthermore, studies have demonstrated that home-based rehabilitation improves patient adherence and HRQoL [21].

Web-based interventions have grown in popularity in recent years. These interventions enable the user to independently navigate a recommended online program that is operated via a website with the aim to create positive changes in health and well-being [22]. Government organizations are actively trying to transition in-person activities to web-based platforms [23]. After the COVID-19–related lockdown, this gained renewed prominence because the advantages of digital technology became increasingly evident. Furthermore, earlier reviews have identified several mobile and electronic apps designed to support various stages of the cancer continuum ranging from prevention to survivorship [24]. The use of mobile health (mHealth) and eHealth technologies, such as wearables and activity trackers as well as apps and web-based programs that can be accessed via smartphones and tablets, provide new methods for educating, monitoring, and supporting patients with chronic conditions and cancer. The World Health Organization recognizes the potential of mHealth and eHealth interventions to support health care delivery [25]. These interventions can assist patients in self-managing their health behaviors and are considered feasible, acceptable, and effective approaches to providing supportive care [26,27].

There is a growing body of evidence to support technology interventions in health care, and this is supported within the UK National Health Service Long Term Plan [28]. Nevertheless, evidence-based mHealth and eHealth interventions to enhance exercise and physical activity for people with lung cancer remain uncommon. Furthermore, of the cancer-related apps that are available, a limited number adopt a personalized approach to physical activity and exercise that accommodates patients' preferences.

Aim

This study aims to identify the mHealth and eHealth technologies that have been developed to support people with lung cancer to improve or maintain physical functioning and enhance their HRQoL.

Objectives

The primary objective of the review was to determine whether any of the mHealth or eHealth technologies identified impacted the physical functioning and HRQoL of people with lung cancer.

The secondary objectives were to assess the demand on clinician time; evaluate the acceptability of the intervention to patients, carers, and health care professionals; investigate user satisfaction with the technology; identify security features (clinical safety,

XSL•FO
RenderX

data protection, and technical security); and examine the cost impact of the mHealth or eHealth app or intervention.

Methods

Design

This review was prospectively registered with PROSPERO (CRD42023414094) and is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to improve the quality of the review and ensure transparency at all stages.

Inclusion Criteria

Types of Studies

We included all primary research studies, without study design or publication status limitations or language or geographic area restrictions. Case studies were also included. Reference lists of systematic reviews were cross-checked to identify any potential studies for inclusion. We limited the search to studies published after January 1, 1990, because internet interventions did not exist before this date [29].

Population

We included studies that were undertaken with adults (aged >18 y) who were diagnosed with lung cancer, regardless of the stage of their disease, treatment allocation, sex, or where they received care.

Intervention

Study participants in the included studies must have been exposed to a physical activity, exercise, or training intervention that was delivered via an electronic or web-based application with or without a comparator.

Outcome Measures

Studies were included if they reported on the impact of the digital intervention on physical function or HRQoL or both physical functioning and HRQoL using any validated measure. We included studies that measured impact at ≥1 time points.

Exclusion Criteria

The following studies were excluded: (1) cancer studies in which the total number of participants with lung cancer accounted for <50% of the study population, (2) studies that focused on telehealth care only and did not include an electronic or web-based intervention (eg, studies that evaluated remote sessions with a clinician via video link), and (3) studies in which apps that were used to track activity could be linked to a wearable device but did not provide any other support.

Search Strategy

A search of 6 databases—PubMed, Web of Science, CINAHL, MEDLINE, SPORTDiscus, and PsycINFO—was carried out on April 17, 2023, via EBSCO, using a list of key terms focusing on 3 distinct categories: the intervention characteristics (eg, web-based, internet, app or application, remote, and digital), physical functioning (eg, activity, exercise, training, movement, and athletics), and the population of interest (eg, patients with lung cancer or survivors of lung cancer and patients with cancer). These were amalgamated using Boolean operators to formulate a comprehensive search string. The full list of search terms is presented in Textbox 1. Other sources, such as references of included records, were also searched.

Textbox 1. Groups of keywords used in the search strategy.

Search terms

*lung cancer patient** or *lung cancer surviv** or *lung canceror lung neoplasm** or *lung tumor** or *lung tumour** or *lung adenocarcinoma* AND *physical activity* or *exercise* or *training* or *physical function** or *mobility* or *rehabilitation* or *prehabilitation* or *physical fitness training* or *physical rehabilitation* or *physical recovery* or *mobility training* AND *mobile applications* or *mobile apps* or *mobile phone apps* or *phone apps* or *smartphone apps* or *smartphone applications* or *web apps* or *web applications* or *mhealth* or *m-health* or *ehealth* or *e-health* or *online support system* or *web-based technology* or *app* or *apps* or *software app* or *cell phone apps* or *cellular phone apps* or *mobile technologies* or *mobile devices* or *smartphones* or *technology-enabled care services* or *interactive apps* or *telemedicine* or *virtual medicine* or *interactive consultative services* or *Web based tool* or *activity tracker* or *fitness tracker* or *physical fitness tracker* or *technology enabled care services*

Data Collection and Analysis

Selection of Studies

Of the 794 papers identified through the database search, 213 (26.8%) duplicates were removed, leaving 581 (73.2%) papers. Two reviewers (SK, PRW, CH, or ZD) screened the titles and abstracts of these 581 papers, excluding 544 (93.6%) in the process and retaining 37 (6.4%). Of these 37 papers, a further 7 (23%) were excluded because they were conference abstracts. A full-text screening of the remaining 30 papers was then conducted, and 22 (73%) were excluded, leaving 8 (27%) for final inclusion. Any discrepancies identified by the reviewers during the screening process were resolved by discussion with a third member of the review team. The screening process is outlined in the PRISMA diagram (Multimedia Appendix 1).

Data Extraction

Data were independently extracted from the included papers by the lead author (SK) using a data extraction template developed for the purpose of this review. The extracted data were reviewed by the coauthors (CH, ZD, and PRW), and discrepancies were resolved. In case of missing study data, we attempted to contact the corresponding authors to obtain the required information. Three authors were contacted [30-32], but none replied. However, these papers were not excluded because they still met the inclusion criteria and reported on some of our objectives. We extracted the data shown in Textbox 2.

XSL•FO
RenderX

Textbox 2. Extracted data.

Study characteristics

Authors, year of publication, title, country of study, year of study, study objective, overall study design, recruitment method, sample size, participant age range, sex, and study duration

Intervention characteristics

Technology product used, setting, intervention details, and exercise details (type, frequency, intensity, and duration)

Outcome measures of interest

Impact on physical functioning, impact on HRQoL, and user acceptability

Other outcomes of interest

Impact on clinician time, user acceptability or satisfaction, safety features, and cost impact

Critical Appraisal

Risk of Bias

Two reviewers (SK and ZD) independently assessed the risk of bias of each included study and confirmed agreement. The JBI checklists [33] were used to assess the methodological and reporting quality of the included studies.

Quality Appraisal

The Grading of Recommendations Assessment, Development, and Evaluation approach was adopted to assess the quality of the evidence used to support the synthesized findings [34,35].

Data Synthesis and Analysis

Due to the methodological heterogeneity of the included studies, we were unable to undertake a meta-analysis. Instead, the Synthesis Without Meta-Analysis reporting items checklist was used to aid transparency in the reporting process [36]. This enabled us to report on the key features of the included studies, group the studies, and explain our findings. A narrative synthesis [37] was also undertaken in accordance with the study objectives stated earlier. This allowed us to provide a comprehensive summary of the impact and effectiveness of the interventions identified in the included studies.

Results

Overview

Eight papers were identified as suitable for inclusion. However, 2 (25%) of these 8 papers [30,38] reported on different stages of the same study; therefore, only 7 studies were included in the review. The screening process is outlined in the PRISMA diagram presented in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

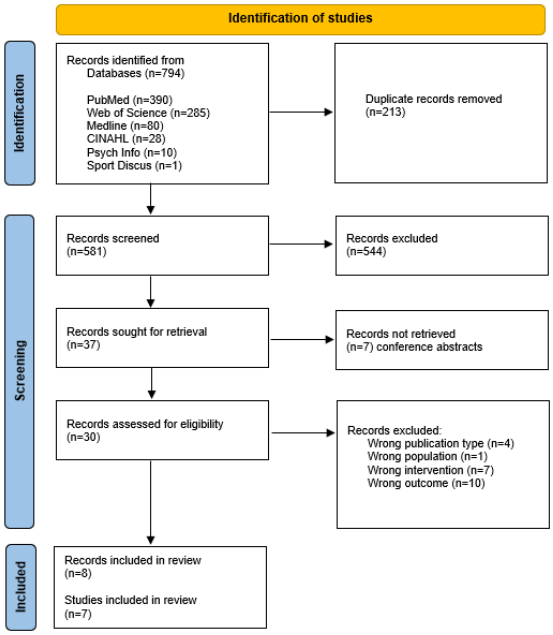

Study Characteristics

The included studies were published between 2013 and 2022 across international settings, including the United States [30,38], the Netherlands [39], South Korea [32,40,41], the United Kingdom [31], and Canada [42].

The studies were undertaken between 2010 [30] and 2018 [40]. The aim of the included studies was to develop a technology [31] and test its usability [39], feasibility [31,39,41,42], acceptance [30,31], and efficacy [30,32,38,41]. The studies were primarily quasi-experimental and nonrandomized experimental studies [30,32,39,41,42]. Only 1 randomized

XSL•FO
RenderX

controlled trial [40] and 1 cohort study [31] were included. The research methods used were predominantly quantitative (5/7, 71%) [30,32,40-42], with the remaining studies using mixed methods (2/7, 29%) [31,39].

Participants were identified from secondary care settings. The number of study participants ranged from 5 [42] to 100 [41], and their mean ages ranged from 55.1 (SD 8.7) years [41] to 64.6 (SD 6.5) years [30,38]. Overall, there were more female participants (195/340, 57.4%) than male participants (145/340, 42.6%) across 4 (57%) of the 7 studies [30,39-41]; the study by Kadiri et al [31] did not report the sex breakdown of the participants. The intervention duration ranged from 6 to 12 weeks; the study by Kadiri et al [31] did not report the intervention duration, but it was clear that the intervention was delivered pre- and postoperatively with a study duration of ≤18 months. The characteristics of the included studies are detailed in Table 1.

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Authors, year; country	Title	Year of study	Study objectives	Study design	Recruitment method	Sample size (n)	Age (y), mean (SD)	Sex (n)	Lung cancer stage	Intervention duration	Risk of bias	Quality appraisal score
Hoffman et al [38], 2013; United States	Too Sick Not to Exercise Using a 6-Week, Home-Based Exercise Intervention for Cancer-Related Fatigue Self-Management for Postsurgical Non–Small Cell Lung Cancer Patients	2010	To evaluate the feasibility, acceptability, safety, and changes in study end points of a home-based exercise intervention to enhance perceived self-efficacy for cancer-related fatigue self-management for persons after thoracotomy for NSCLC[a] transitioning from hospital to home	Quasi-experimental non randomized experimental study; first 6 weeks after discharge (quantitative)	Potential participants were identified during clinical appointments while undergoing diagnostics to confirm a potential diagnosis of NSCLC	7	64.6 (6.5)	F[b]=5, M[c]=2	I, II, or III	6 wk	Low	Moderate
Hoffman et al [30], 2014; United States	Virtual Reality Bringing a New Reality to Postthoracotomy Lung Cancer Patients via a Home-Based Exercise Intervention Targeting Fatigue While Undergoing Adjuvant Treatment	2010-2011	To investigate the feasibility, acceptability, and preliminary efficacy of an exercise intervention for postthoracotomy patients with NSCLC to include those initiating and completing adjuvant therapy	Quasi-experimental non-randomized experimental study; weeks 7 to 16 after discharge (quantitative)	Participants from phase 1 were asked whether they would like to participate in phase 2	__[d]	—	—	—	10 wk	Low	Moderate
Groen et al [39], 2017; Netherlands	Supporting Lung Cancer Patients With an Interactive Patient Portal: Feasibility Study	2014	To evaluate the feasibility and usability of the patient portal and generate preliminary evidence on its impact	Quasi-experimental non-randomized experimental study (mixed methods)	Patients were approached by letter, followed by a telephone call from the researchers to discuss participation and check further eligibility criteria	37	59.6 (8.4)	F=16, M=21	I, II, or III	4 mo	Low	Moderate

XSL•FO
RenderX

Authors, year; country	Title	Year of study	Study objectives	Study design	Recruitment method	Sample size (n)	Age (y), mean (SD)	Sex (n)	Lung cancer stage	Intervention duration	Risk of bias	Quality appraisal score
Ji et al [40], 2019; South Korea	Mobile Health Management Platform–Based Pulmonary Rehabilitation for Patients With Non–Small Cell Lung Cancer: Prospective Clinical Trial	2017-2018	To examine the outcome of home-based pulmonary rehabilitation regarding exercise capacity, dyspnea symptoms, and QoL[e] in adult patients being treated for NSCLC; primary end points were pulmonary function parameters, and the secondary end point was QoL	Randomized controlled trial (quantitative)	Participants identified from the outpatient department of a single tertiary hospital	64	60.50 (9.80) in fixed-interactive exercise group; 57.97 (10.10) in fixed exercise group	Fixed exercise group: F=45, M=21; fixed-interactive exercise group: F=51, M=24	I-IV	12 wk	Low	Moderate
Kadiri et al [31], 2019; United Kingdom	Fit 4 Surgery, a Bespoke App With Biofeedback Delivers Rehabilitation at Home Before and After Elective Lung Resection	NR[f]	To develop a bespoke pulmonary rehabilitation app and test its feasibility and acceptability to patients undergoing lung resection surgery	Cohort study (mixed methods)	Patients deemed eligible for curative lung cancer surgery, based on British Thoracic Society guidelines, were referred by the multidisciplinary teams to the regional thoracic surgery unit, where potential patients were identified	31	64 (12)	NR	NR	Unclear	Low	Moderate
Park et al [41], 2019; South Korea	Mobile Phone App–Based Pulmonary Rehabilitation for Chemotherapy-Treated Patients With Advanced Lung Cancer: Pilot Study	2016	To determine the feasibility and efficacy of smartphone app–based pulmonary rehabilitation on exercise capacity, symptom management, and QoL in patients with advanced lung cancer undergoing chemotherapy	Quasi-experimental nonrandomized experimental study (quantitative)	Consecutive patients with histologically diagnosed advanced NSCLC were identified	100	55.1 (8.7)	F=54, M=46	II-IV	12 wk	Low	Moderate

XSL•FO
RenderX

Authors, year; country	Title	Year of study	Study objectives	Study design	Recruitment method	Sample size (n)	Age (y), mean (SD)	Sex (n)	Lung cancer stage	Intervention duration	Risk of bias	Quality appraisal score
Coats et al [42], 2020; Canada	Feasibility of an Eight-Week Telerehabilitation Intervention for Patients With Unresectable Thoracic Neoplasia Receiving Chemotherapy: A Pilot Study	2014	To investigate the feasibility, adherence, and satisfaction of a home-based telerehabilitation program with acquisition of real-time physiological parameters in patients with unresectable thoracic neoplasia receiving chemotherapy and to explore its effects on patients' functional capacity	Quasi-experimental nonrandomized experimental study (quantitative)	5 consecutive eligible patients diagnosed with unresectable thoracic neoplasia and receiving chemotherapy were recruited	5	62 (7)	F=2, M=3	NR	8 wk	Low	Moderate
Yang et al [32], 2022; South Korea	Evaluation of a Smart After-Care Program for Patients With Lung Cancer: A Prospective, Single-Arm Pilot Study	2015	To evaluate the efficacy of a remote health care program for patients with lung cancer: the Smart After-Care Program	Quasi-experimental nonrandomized experimental study (quantitative)	NR	50	58.3 (11.9)	F=22, M=28	I-IV	12 wk	Low	Moderate

[a]NSCLC: non–small cell lung cancer.

[b]F: female.

[c]M: male.

[d]Not applicable.

[e]QoL: quality of life.

[f]NR: not reported.

Quality Appraisal

The risk of bias for all studies was low, and quality appraisal scores were moderate (Table 1).

Intervention Characteristics

The review identified 7 technologies that had been studied in people with lung cancer: 4 (57%) mobile apps [31,32,40,41] and 3 (43%) web applications [30,38,39,42]. Of the 3 web applications, 2 (67%) used a gaming console to deliver part of the exercise prescription [30,38,42].

The interventions were delivered or accessed from various settings: web based [31,32,39,40], home based [30,38,42], or a combination of outpatient department and home based [41].

The frequency, intensity, time, and type formula [43] was used to extract key components of the exercise prescription of each study. Only 5 (62%) of the 8 papers reported the full details of the exercise prescription according to the frequency, intensity, time, and type formula [22,30,31,38,41].

The papers that did not provide a detailed exercise prescription provided more general information and recommendations regarding physical activity [32,39]; alternatively, the interactive app would support the participant to edit the frequency, intensity, and duration of the exercise [40]. Where the technology included an interactive patient portal, it was noted that the physical activity support program that was incorporated in the portal was only used by one-third of the participants [39].

The intervention details are summarized in Table 2.

XSL•FO
RenderX

Table 2. Intervention characteristics.

Authors, year	Technology product used	Setting	Intervention details	Exercise details			
				Type	Frequency	Intensity	Time
Hoffman et al [30,38], 2014 and 2013	Nintendo Wii Fit Plus fitness game	Home based	Light-intensity exercise intervention using a game console	Light-intensity walking and balance exercises	Walking: daily for 5 days in week 1, then every day; balance exercises: 5 days per week	Walking: comfortable and self-paced; balance exercises: <3.0 METs[a]	Walking: started at 5 min/d for 5 days during week 1 and was anticipated to increase by 5 min/d each week with the goal of walking 30 min/d during week 6; balance exercises: duration not reported; based on a gaming format and scoring system
Groen et al [39], 2017	MyAVL interactive portal	Online	Web-based patient portal that included physical activity advice	__[b]	—	—	Used a computerized system that provided advice depending on nutritional status; possible contraindications for physical activity; treatment phase; tumor type (breast or lung cancer); whether the patient is participating in a supervised exercise program, and if yes, whether additional information on physical activity is desired
Ji et al [40], 2019	efil breath app	Online	Personalized pulmonary rehabilitation program using mobile apps: 1 app included fixed exercises, and another app included interactive exercises; a patient monitoring website was also used	Walking and exercises	—	—	The fixed exercise group used the fixed exercise program for 12 weeks; there were 6 levels of walking distance: 600 m, 1200 m, 1800 m, 2400 m, 3000 m, and 3600 m; when the user achieved a fixed walking distance within a day and 14 times in total, the app increased the walking distance to the next level; the interactive exercise group used the fixed exercise program for 6 weeks and then switched to the app with interactive exercises for the next 6 weeks; the interactive exercise used 12 levels; initial walking intensity was set to 80% of the maximum walking speed; once initiated, a metronome in the app was used to help guide the walking speed of the patient; the level of exercise was then adjusted according to the modified Borg scale
Kadiri et al [31], 2019	Fit 4 Surgery app	Web based	Mobile exercise app that included 10 exercises based on the lung cancer "Rehabilitation for Operated Lung Cancer" surgery program	10 exercises: upper and lower limb, aerobic and strength based	Patients' discretion	Target HR[c] >60% of maximum HR	3 min/exercise=30 min in total
Park et al [41], 2019	Smart Aftercare app	Outpatient department and home based	Pulmonary rehabilitation program using a smartphone app	Stretching exercise, aerobic exercise, and muscle strengthening	3 times per week	Target HR 70% of HR reserve plus resting HR; oxygen saturation >88%	30-60 min

XSL•FO
RenderX

Authors, year	Technology product used	Setting	Intervention details	Exercise details			
				Type	Frequency	Intensity	Time
Coats et al [42], 2020	$TELE_{RP}$ (telerehabilitation program)	Home based	Telerehabilitation program using the eChez-Soi telerehabilitation platform	Exercise ball and elastic bands to exercise upper limbs, as well as wall squats and lunges for lower limbs; cardiovascular exercise with Xbox dance mat and Wii balance board	3 sessions per week, each lasting 75 min, for 8 wks; 15 supervised and 9 unsupervised	SpO_2[d] >88%; cardiovascular exercise at moderate intensity; target HR 60%-80% of the VO_{2peak}[e]	Repetitions increased according to patients' tolerance until 2 sets of 15 repetitions; 20 min of cardiovascular exercise
Yang et al [32], 2022	Smart AfterCare app	Web based	Mobile app that provided information about rehabilitation exercises and a healthy diet for patients with lung cancer	Muscle strength using elastic bands, stretching, and breathing	NR[f]	NR	NR

[a]MET: metabolic equivalent of task.

[b]Not applicable.

[c]HR: heart rate.

[d]SpO_2: peripheral oxygen saturation.

[e]VO_{2peak}: peak oxygen uptake.

[f]NR: not reported.

Outcome Measures

Primary Outcome Measures

Physical Functioning

All included studies demonstrated some improvement in physical functioning, but the methods of assessment varied. The most common assessment was walking time and walking distance. In the studies by Ji et al [40] and Coats et al [42], participants demonstrated an improvement in the 6-minute walk test and the 6-minute walk distance test, while Yang et al [32] noted an improvement in the 2-minute walk distance test. An increase in walking duration and step count was also observed by Hoffman et al [30], while Kadiri et al [31] noted an increase in the distance covered in the shuttle walk test.

No improvement in muscle strength was noted by Coats et al [42], while Yang et al [32] noted an improvement in lower extremity muscle strength but not in upper extremity muscle strength. Only the study by Hoffman et al [30,38] showed an improvement in balance and cancer-related fatigue, while the study by Ji et al [40] showed an improvement in dyspnea grade.

The study by Groen et al [39] identified no significant improvement in physical activity, but an improvement in vigorous activity over time was noted. Finally, the study by Park et al [41] found that although exercise capacity improved in stable patients, this was not the case in patients with progressive disease.

Impact on HRQoL

Almost all studies (6/7, 86%) included in our review reported the impact of the digital health intervention on HRQoL [31,32,39-42].

The tool most commonly used to assess HRQoL was the European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30. Improvements in symptom scales were observed by Kadiri et al [31] and Park et al [41], improvements in functional scales were observed by Yang et al [32] and Park et al [41], while Coats et al [42] did not observe any significant changes.

Ji et al [40] used the EuroQol visual analog scale score to demonstrate a significant improvement between visit 1 and visit 3, but the EQ-5D scores did not differ between the same time points. The Short Form Health Survey-36 was used in the study by Groen et al [39], but no significant changes in the scores were noted.

Secondary Outcome Measures

Demand on Clinician Time

The technologies varied with regard to the level of clinician involvement required for implementation and ongoing patient management. The authors attempted to assess the impact on clinician time across the included studies (Table 3). Only the study by Kadiri et al [31] explicitly reported the impact of adopting the intervention on health care professional time: 60 minutes.

XSL•FO
RenderX

Table 3. Study outcome measures.

Authors, year	Impact on physical functioning	Impact on HRQoL[a]	Demand on clinician time	Acceptability of technology	User satisfaction	Safety features	Costs
Hoffman et al [30,38], 2014 and 2013	Reduced cancer-related fatigue, improved balance and walking duration, and increased step count	Not assessed	Preoperative teaching, postdischarge call, home visit to set up equipment, follow-up call 24 hours later to assess progress and address queries, home visit at 2 weeks, and weekly calls until week 6; ongoing nurse access via telephone during exercise, and nurse would make home visit if there were safety concerns	All participants strongly agreed that exercising at home was convenient, that the nurse interactions during the telephone calls were helpful, and that they would recommend the program to others undergoing similar surgery	Participants agreed strongly that they had a high level of satisfaction with the exercise intervention, giving it a mean score of 5.8 out of 6, with 6 indicating *agreed strongly*, which exceeded the goal of 4 out of 6	Clinical safety: participants required telephone access while exercising should they need assistance; the nurse was available by telephone and could make a home visit if required; and participants were taught how to maintain a light-intensity dose of exercise and also provided with heart rate monitors	Not assessed
Groen et al [39], 2017	Levels of physical activity did not change significantly, but vigorous physical activity tended to increase over time from median 0 (IQR 0-840) MET[b] min/wk to median 240 (IQR 0-1140) MET min/wk	SF-36[c]: no significant changes	Recruitment procedures and onboarding	93% (25/27) of the patients found the app easy to use, 56% (15/27) reported that it contributed to a sense of control over their health, and 69% (18/26) indicated that it was a valuable addition to their health care experience	Most of the patients (22/27, 81%) were satisfied with the MyAVL portal, and 77% (20/26) intended to continue using it	Authorization procedures (username, password, and SMS text message authentication)	Not assessed
Ji et al [40], 2019	Comparison between pre- and postintervention results; for all participants in both groups, the 6MWD[d] test performance improved significantly from visit 1 to visit 3 (from mean 433.43, SD 65.60 m to mean 471.25, SD 75.69 m; *P*=.001); subjective dyspnea grade measured using the Modified Medical Research Council Dyspnea Scale improved from visit 1 to visit 3 (from mean 0.94, SD 0.66 to mean 0.61, SD 0.82; *P*=.02); no statistical differences were noted between the fixed and fixed-interactive exercise groups	Comparison between pre- and postintervention results: EQ-VAS[e] score improved significantly from mean 76.05 (SD 12.37) at visit 1 to mean 82.09, (SD 13.67) at visit 3 (*P*=.002); the mean EQ-5D scores were not significantly different between the time points; no statistical differences were noted between the fixed and fixed-interactive exercise groups	Use a central monitoring website to store records and access summary of compliance and patient health status; measure heart rate and SpO_2[f] during exercise, as well as the 6MWT[g] results	Not assessed	PGA[h] scores measured at visit 3 showed significant improvement over PGA scores at visit 2 (from mean 13.77, SD 3.68 to mean 15.08, SD 3.99; *P*=.01)	The apps were linked to a wearable pulse oximeter via Bluetooth, and activity-related data were sent to the monitoring website; this is a secure database, which ensures that each participating hospital can only access its own patient data	Not assessed

XSL•FO
RenderX

Authors, year	Impact on physical functioning	Impact on HRQoL[a]	Demand on clinician time	Acceptability of technology	User satisfaction	Safety features	Costs
Kadiri et al [31], 2019	Improved shuttle walk test performance	EORTC QLQ-C30[i] score improvement noted in fatigue, pain, and dyspnea; the global health score at 5 months for the app significantly increased and returned to baseline level	60 min of health care professional time	Considered acceptable by the researchers because patients in the app group managed more sessions during the pre- and postoperative periods compared to those in the rehabilitation group	Participants found the app easy to use, the ability to see oxygen levels and heart rate was motivational, and they liked the variety of exercises; the novelty factor of using the app for exercise was appealing to some patients, and even patients who had good levels of fitness before using the app found it beneficial	Clinical safety: the app collected baseline measurements of SpO_2 levels and heart rate for safety; a safety notification screen prevented patients with poor compliance from continuing on the app	£16-£34 (US $20.80-$44.20) per patient
Park et al [41], 2019	Significant difference between before and after the patients in the baseline 6MWD test performance according to baseline ECOG-PS[j] scores; the mean distance covered was 416.8 (SD 55.4) m in patients with ECOG-PS 0, 369.8 (SD 80.3) m in those with ECOG-PS 1, and 305.7 (SD 89.1) m in patients with ECOG-PS 2 (P=.04); after pulmonary rehabilitation, the 6MWD test performance had improved significantly: from 380.1 (SD 74.1) m at baseline and 429.1 (SD 58.6) m (P<.001) at 6 weeks to 448.1 (SD 50.0) m (P<.001) at 12 weeks; patients with stable disease showed significantly improved 6MWD test performance: from 384.2 (SD 74.6) m at baseline and 426.1 (SD 6.5) m (P<.001) at 6 weeks to 447.4 (SD 50.4) m (P<.001 at 12 weeks; the dyspnea scale, evaluated using the EORTC QLQ-C30, did not show any significant improvement in the patients overall, but patients with stable disease tended to show improvement	Global health score and HRQoL tended to improve in patients, but these results were not statistically significant	Exercise duration and intensity was prescribed by a physician and adjusted at every clinic visit	NR[k]	90% of the participants were satisfied with the service, and 88% would recommend the program to others	NR	Not assessed

XSL•FO
RenderX

Authors, year	Impact on physical functioning	Impact on HRQoL[a]	Demand on clinician time	Acceptability of technology	User satisfaction	Safety features	Costs
Coats et al [42], 2020	No impact on weight, exercise tolerance, or functional capacity; quadriceps muscle function, peak torque, total work, and fatigue index did not change significantly; the 6MWT and the TST[l] performance were significantly improved (mean 40, SD 20 m; *P*=.01; and mean –3.0, SD 0.2 s; *P*=.05, respectively)	EORTC QLQ-C30 and EORTC QLQ-LC13[m]: no significant changes in scores	15 supervised sessions (supervised by a clinical exercise physiologist or cancer exercise trainer certified by the American College of Sports Medicine) and 9 unsupervised sessions; 85 hours of kinesiologists' time and 36 hours spent on installation and uninstallation (technicians and engineers) of the telerehabilitation program	NR	Quebec User Evaluation of Satisfaction with Assistive Technology questionnaire used: participants' ratings ranged from *quite satisfied* to *very satisfied*	A complete clinical evaluation was made at baseline and after the intervention (within 1 week after the last exercise session); continuous data acquisition and recording from physiological sensors during rehabilitation sessions	Not assessed
Yang et al [32], 2022	No significant difference was observed in upper extremity muscle strength, but significant improvements in lower extremity muscle strength were observed, with repetitions increasing from 18 to 22 for the 30-second chair stand test (*P*=.01); a significant improvement was also noted in walking distance in the 2-minute walk distance test (from 185.7 to 195.0 m; *P*=.03)	EORTC QLQ-C30 functional scales: significant improvements were observed in all subsections (*P*<.05); no significant improvements were seen in the symptom scales; no significant differences were observed in the EORTC QLQ-LC13 scores	Exercise instructions and training from rehabilitation specialist at baseline, 6 weeks, and 12 weeks; in addition, at baseline, training was provided on devices, app, and equipment	Acceptable for, and supportive of, patients with reduced pulmonary function after lung cancer treatment; the Smart After-Care Program was found to be particularly useful for patients living far from the hospital (>80 km)	Assessed at 6 and 12 weeks; in the final satisfaction survey, 88% of the patients rated overall satisfaction as *very good* or *good*	NR	Not assessed

[a]HRQoL: health-related quality of life.
[b]MET: metabolic equivalent of task.
[c]SF-36: Short Form Health Survey-36.
[d]6MWD: 6-minute walk distance.
[e]EQ-VAS: EuroQol visual analog scale.
[f]SpO_2: peripheral oxygen saturation.
[g]6MWT: 6-minute walk test.
[h]PGA: Patient Global Assessment.
[i]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30.
[j]ECOG-PS: Eastern Cooperative Oncology Group-Performance Status.
[k]NR: not reported.
[l]TST: timed stair test.
[m]EORTC QLQ-LC13: EORTC QLQ–Lung Cancer 13.

Acceptability and User Satisfaction

All included papers reported either user satisfaction [40-42] or both acceptability and user satisfaction [30-32,39], as summarized in Table 3, and demonstrated high levels of acceptability and user satisfaction.

Safety Features

The studies reported primarily on clinical safety features of the technology (eg, access to a mobile phone while exercising to contact the study nurse [30,38], heart rate monitoring [30,31,38,40], and monitoring of peripheral oxygen saturation levels [31], as noted in Table 3). Of the 8 studies, 2 (25%) also reported on the technology's security features to ensure limited authorized access to the app [40] or database [39].

Cost-Effectiveness

None of the studies evaluated the cost-effectiveness of the digital health interventions in terms of clinical outcomes within a health economic context. Only the single-center cohort study [31] reported on the cost of implementing the technology, which

XSL•FO
RenderX

was estimated to be between £16 (US $20.8) and £34 (US $44.2) per patient.

Discussion

eHealth: Benefits and Concerns

eHealth, which encompasses information and communication technology, facilitates remote care delivery and health information transmission, promoting patient involvement, improving care quality, and enhancing accessibility, particularly in remote areas. Despite benefits such as convenience and reduced travel, concerns exist, such as privacy issues, fears of diminished human interaction, and lack of access to technology and internet infrastructure [44]. eHealth holds promise in supporting patient-centered care and empowers patients by enhancing their involvement in their health care [45].

Principal Findings

To our knowledge, this is the first review that has looked specifically at digital technologies developed for people with lung cancer and their impact on physical functioning and HRQoL. Despite searching across a wide range of databases, we were only able to identify 7 studies for inclusion in our review.

The findings of this review suggest that digital health interventions with an exercise component are acceptable to people with lung cancer because the interventions help them play a more active role in their health care and can positively impact their physical functioning and HRQoL. Various tools were used in the studies to assess HRQoL. The EQ-5D, a questionnaire conventionally used for economic cost analysis, was used in the study by Coats et al [42] to assess HRQoL, and the authors noted that, despite an observed increase in the EuroQol visual analog scale score, the EQ-5D score did not improve. The evidence to support the use of digital exercise interventions remains sparse, and we have drawn on the findings of only a few studies that aimed primarily to develop and assess the feasibility of the intervention used.

Nevertheless, the quality of the studies was high, thereby strengthening the credibility of the review findings. Further research is required to observe the effects of the interventions over a longer period and to explore the potential cost savings associated with using a remote health management mobile app, including a reduction in clinician time or the number of consultations, improved disease management, and the costs of implementing the technologies in routine care, as well as the safety and security risks of the technologies.

Results in Context

In recent years, there has been a growth in the number of digital health technologies, including mHealth and eHealth, as well as wearable devices that are becoming an integral part of modern health care [46]. The promotion of patient self-management practices facilitated by digital technology has gained attention because it can improve patient engagement and health care delivery [47]. This has become particularly relevant and has intensified in the wake of the pressures placed on the health service during the COVID-19 pandemic [22]. It is anticipated that strategies to encourage patients to self-manage their health behaviors will continue to be integrated into care pathways in the future [48].

In the context of lung cancer, approximately two-thirds of the people with lung cancer have at least 1 other preexisting health condition, and up to half have ≥2 preexisting health conditions [6], making this group more complex to treat because they require a more tailored approach to rehabilitation [49]. We note that the technologies identified in our review either produced an exercise prescription relevant to the individual participant's capability or gave the patient guidance on exercises they might be able to undertake. As the participants' exercise tolerance improved, the prescription was updated, affording a bespoke approach to rehabilitation that might be more relevant to people with lung cancer.

This review highlights the benefits of a digital health intervention that included an exercise component and demonstrated improvements in cancer-related fatigue and balance [30,38], walking duration or distance or step count [30-32,38,42], general activity levels, dyspnea [40], and muscle strength for people living with lung cancer [32]. Nevertheless, there are still areas for development; for example, the study by Yang et al [32] found that while there was an increase in lower extremity muscle strength, no change in upper extremity muscle strength was noted, and, in another study [41], for participants with progressive disease, no improvement was observed in exercise capacity or dyspnea scale score. Furthermore, none of the included studies provided a bespoke exercise prescription and demonstrated a positive impact on exercise endurance, muscle endurance, or dyspnea symptoms. This reinforces the need to produce an exercise prescription that is relevant to the participants' disease status and individual exercise goals. More consideration is required to develop technologies with intelligent algorithms or artificial intelligence that can support persons with lung cancer to input data on their current condition and capability before a safe, bespoke exercise prescription is recommended. This needs to be factored into app development and further research.

Improving HRQoL is a priority goal in supporting people with lung cancer [49]. All studies that assessed the impact of HRQoL reported a positive impact. The direction of the association between HRQoL and improved physical functioning was consistent across most of the studies (n=5, 71%) that reported on both outcomes [31,32,40-42]. Improvements were primarily noted in the global health score as well as the symptom and functional scales, but it is unclear whether these improvements are a direct result of the exercise intervention or improved disease management and control.

This review highlighted the high levels of acceptability and participant satisfaction with the digital intervention. Having a home-based exercise intervention or web-based intervention, such as the technologies highlighted in this review, is advantageous because participants can continue to exercise during chemotherapy treatment when they are usually advised to limit their contact with other people to reduce the risk of infection. Furthermore, the studies that used a game console as part of their exercise program incentivized their participants to

XSL•FO
RenderX

use the technology, work out more often, and increase the intensity of the exercise [30,38,42]. Other useful features of the technologies that have been highlighted include the use of a virtual environment for exercising [30,38], dance-based workouts [42], dietary advice [32], and the value of incentivizing participants [30,38,42]. Hence, it is important for future app developers to consider what incentives can be built into the intervention to improve adherence, reduce clinician workload, and empower patients to be more independent regarding their disease management.

It is anticipated that health technologies, such as the ones identified in this review, will enable improved disease management, reduce hospital admissions, and save the UK National Health Service money [31]. However, limited cost analyses have been conducted to date to demonstrate the cost impact on health care services, the health care workforce, and people living with lung cancer. Our review identified only 1 cohort study [31] undertaken at a single research center that reported the cost of the intervention to the health care provider and the estimated time impact on health care professionals.

Limitations

We reported on evidence available within the published papers. We faced challenges in gathering details regarding missing study data. We made attempts to reach out to the corresponding authors to obtain the necessary information, but we did not receive any response. Furthermore, it is unclear whether any development information not included in the reports of the studies was excluded [30-32]; in other words, important details could have been left out of the study reports, and the reasons for their possible omission are not known.

The studies included in our review were primarily quasi-experimental and nonrandomized studies [30,32,39,41,42]. Only 1 randomized controlled trial [40] and 1 cohort study [31] were included. The studies focused on technology development and feasibility assessment, most frequently for patients undergoing surgery [30-32,39,40]. In addition, the small cell subtype of lung cancer has poorer outcomes than non–small cell subtypes, with a primary treatment focus being to improve HRQoL; this population was not well represented in the included studies. Furthermore, various tools were used to assess physical functioning and HRQoL, making it difficult to draw comparisons across the studies due to their heterogeneity.

Conclusions

There seems to be a consistent relationship between the use of digital technologies by people with lung cancer and a positive impact on physical functioning and HRQoL. The technologies reported in this review demonstrated high levels of acceptability or user satisfaction and have the potential to support people with lung cancer to manage their health more independently. However, most of the studies (n=5,71%) identified here report on the development and feasibility of the technology. Further multicenter, large-scale research studies using a randomized controlled trial design over an extended period and including a cost-effectiveness outcome measure are required to fully assess the true benefit of adopting eHealth technologies in standard care services for health care providers and people with lung cancer.

Acknowledgments

The authors would like to thank the academic librarian at Oxford Brookes University, Richard Persaud, for his support with conducting the searches across the scientific databases. CH acknowledges the support of the National Institute for Health and Care Research Oxford Health Clinical Research Facility. The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research, the UK National Health Service, or the UK Department of Health and Social Care.

Authors' Contributions

SK drafted the manuscript. CH, ZD, and PRW supported the development of the concept for the review. SK and 1 other author (CH, ZD, or PRW) screened all abstracts and full texts. All data were extracted by SK and subsequently reviewed by CH, ZD, and PRW. SK conducted the quality assessment in collaboration with ZD and CH. All authors contributed to the methods for the review, multiple versions of the manuscript, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 549 KB - jmir_v26i1e53015_app1.pdf]

References

1. Sharma R. Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050. Int J Clin Oncol 2022 Apr;27(4):665-675 [FREE Full text] [doi: 10.1007/s10147-021-02108-2] [Medline: 35020103]

XSL•FO
RenderX

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
3. Lung cancer statistics. Cancer Research UK. URL: https://tinyurl.com/mrewzd4d [accessed 2024-04-29]
4. Personalised care. National Health Service. 2023. URL: https://www.longtermplan.nhs.uk/areas-of-work/personalised-care/ [accessed 2024-04-29]
5. Shin JY, Chaar D, Kedroske J, Vue R, Chappell G, Mazzoli A, et al. Harnessing mobile health technology to support long-term chronic illness management: exploring family caregiver support needs in the outpatient setting. JAMIA Open 2020 Dec;3(4):593-601 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa053] [Medline: 33758797]
6. Fowler H, Belot A, Ellis L, Maringe C, Luque-Fernandez MA, Njagi EN, et al. Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. BMC Cancer 2020 Jan 28;20(1):2 [FREE Full text] [doi: 10.1186/s12885-019-6472-9] [Medline: 31987032]
7. Avancini A, Sartori G, Gkountakos A, Casali M, Trestini I, Tregnago D, et al. Physical activity and exercise in lung cancer care: will promises be fulfilled? Oncologist 2020 Mar;25(3):e555-e569 [FREE Full text] [doi: 10.1634/theoncologist.2019-0463] [Medline: 32162811]
8. Curry J, Lind M, Short CE, Vandelanotte C, Evans HE, Pearson M, et al. Evaluating a web-based computer-tailored physical activity intervention for those living with and beyond lung cancer (ExerciseGuide UK): protocol for a single group feasibility and acceptability study. Pilot Feasibility Stud 2022 Aug 13;8(1):182 [FREE Full text] [doi: 10.1186/s40814-022-01129-6] [Medline: 35964141]
9. Henshall CL, Allin L, Aveyard H. A systematic review and narrative synthesis to explore the effectiveness of exercise-based interventions in improving fatigue, dyspnea, and depression in lung cancer survivors. Cancer Nurs 2019;42(4):295-306. [doi: 10.1097/NCC.0000000000000605] [Medline: 29787385]
10. Bade BC, Thomas DD, Scott JB, Silvestri GA. Increasing physical activity and exercise in lung cancer: reviewing safety, benefits, and application. J Thorac Oncol 2015 Jun;10(6):861-871 [FREE Full text] [doi: 10.1097/JTO.0000000000000536] [Medline: 25831230]
11. Yoo JS, Yang HC, Lee JM, Kim MS, Park EC, Chung SH. The association of physical function and quality of life on physical activity for non-small cell lung cancer survivors. Support Care Cancer 2020 Oct;28(10):4847-4856 [FREE Full text] [doi: 10.1007/s00520-020-05302-6] [Medline: 31980896]
12. Humpel ND, Iverson DC. Sleep quality, fatigue and physical activity following a cancer diagnosis. Eur J Cancer Care (Engl) 2010 Nov;19(6):761-768. [doi: 10.1111/j.1365-2354.2009.01126.x] [Medline: 19832890]
13. van Doorslaer de Ten Ryen S, Deldicque L. The regulation of the metastatic cascade by physical activity: a narrative review. Cancers (Basel) 2020 Jan 08;12(1):153 [FREE Full text] [doi: 10.3390/cancers12010153] [Medline: 31936342]
14. Singh B, Toohey K. The effect of exercise for improving bone health in cancer survivors - a systematic review and meta-analysis. J Sci Med Sport 2022 Jan;25(1):31-40. [doi: 10.1016/j.jsams.2021.08.008] [Medline: 34465518]
15. Stene GB, Helbostad JL, Balstad TR, Riphagen II, Kaasa S, Oldervoll LM. Effect of physical exercise on muscle mass and strength in cancer patients during treatment--a systematic review. Crit Rev Oncol Hematol 2013 Dec;88(3):573-593. [doi: 10.1016/j.critrevonc.2013.07.001] [Medline: 23932804]
16. Bower JE, Bak K, Berger A, Breitbart W, Escalante CP, Ganz PA, American Society of Clinical Oncology. Screening, assessment, and management of fatigue in adult survivors of cancer: an American Society of Clinical oncology clinical practice guideline adaptation. J Clin Oncol 2014 Jun 10;32(17):1840-1850 [FREE Full text] [doi: 10.1200/JCO.2013.53.4495] [Medline: 24733803]
17. Henshall C, Greenfield S, Gale N. The role of self-management practices as mechanisms for re-establishing normality in cancer survivors. Qual Health Res 2017 Mar;27(4):520-533. [doi: 10.1177/1049732316651252] [Medline: 27222037]
18. NICE creates new menu of treatment options for those suffering from depression. National Institute for Health and Care Research. 2021. URL: https://wwwniceorguk/news/nice-creates-new-menu-of-treatment-options-for-those-suffering-from-depression [accessed 2024-04-29]
19. Howell D, Richardson A, May C, Calman L, Fazelzad R, Moradian S, et al. Implementation of self-management support in cancer care and normalization into routine practice: a systematic scoping literature review protocol. Syst Rev 2019 Jan 31;8(1):37 [FREE Full text] [doi: 10.1186/s13643-019-0952-5] [Medline: 30704509]
20. Oates GR, Hamby BW, Stepanikova I, Knight SJ, Bhatt SP, Hitchcock J, et al. Social determinants of adherence to pulmonary rehabilitation for chronic obstructive pulmonary disease. COPD 2017 Dec;14(6):610-617 [FREE Full text] [doi: 10.1080/15412555.2017.1379070] [Medline: 29020525]
21. Holland AE, Mahal A, Hill CJ, Lee AL, Burge AT, Cox NS, et al. Home-based rehabilitation for COPD using minimal resources: a randomised, controlled equivalence trial. Thorax 2017 Jan;72(1):57-65 [FREE Full text] [doi: 10.1136/thoraxjnl-2016-208514] [Medline: 27672116]
22. Barak A, Klein B, Proudfoot JG. Defining internet-supported therapeutic interventions. Ann Behav Med 2009 Aug;38(1):4-17. [doi: 10.1007/s12160-009-9130-7] [Medline: 19787305]
23. Next steps on the NHS five year forward view. National Health Service. 2017. URL: https://www.england.nhs.uk/publication/next-steps-on-the-nhs-five-year-forward-view/2021 [accessed 2024-04-29]

XSL•FO
RenderX

24. Davis SW, Oakley-Girvan I. mHealth education applications along the cancer continuum. J Cancer Educ 2015 Jun;30(2):388-394. [doi: 10.1007/s13187-014-0761-4] [Medline: 25482319]
25. mHealth: new horizons for health through mobile technologies. World Health Organization. 2011. URL: https://iris.who.int/bitstream/handle/10665/44607/9789241564250_eng.pdf?sequence=1 [accessed 2024-04-29]
26. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
27. Jee H. Review of researches on smartphone applications for physical activity promotion in healthy adults. J Exerc Rehabil 2017 Feb;13(1):3-11 [FREE Full text] [doi: 10.12965/jer.1732928.464] [Medline: 28349027]
28. NHS England long-term plan - key implications for physiotherapy. National Health Service. URL: https://wwwcsporguk/campaigns-influencing/shaping-healthcare/nhs-long-term-plan-england/nhs-england-long-term-plan-key [accessed 2024-04-29]
29. Andersson G. Internet interventions: past, present and future. Internet Interv 2018 Jun;12:181-188 [FREE Full text] [doi: 10.1016/j.invent.2018.03.008] [Medline: 30135782]
30. Hoffman AJ, Brintnall RA, Brown JK, von Eye A, Jones LW, Alderink G, et al. Virtual reality bringing a new reality to postthoracotomy lung cancer patients via a home-based exercise intervention targeting fatigue while undergoing adjuvant treatment. Cancer Nurs 2014;37(1):23-33. [doi: 10.1097/NCC.0b013e318278d52f] [Medline: 23348662]
31. Kadiri SB, Kerr AP, Oswald NK, Budacan AM, Flanagan S, Golby C, et al. Fit 4 surgery, a bespoke app with biofeedback delivers rehabilitation at home before and after elective lung resection. J Cardiothorac Surg 2019 Jul 05;14(1):132 [FREE Full text] [doi: 10.1186/s13019-019-0951-6] [Medline: 31277671]
32. Yang HC, Chung SH, Yoo JS, Park B, Kim MS, Lee JM. Evaluation of a smart after-care program for patients with lung cancer: a prospective, single-arm pilot study. J Chest Surg 2022 Apr 05;55(2):108-117 [FREE Full text] [doi: 10.5090/jcs.21.138] [Medline: 35135903]
33. Critical appraisal tools. Joanna Briggs Institute. URL: https://jbi.global/critical-appraisal-tools [accessed 2024-04-29]
34. GRADE handbook. GRADEpro. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2024-04-29]
35. Goldet G, Howick J. Understanding GRADE: an introduction. J Evid Based Med 2013 Feb;6(1):50-54. [doi: 10.1111/jebm.12018] [Medline: 23557528]
36. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890 [FREE Full text] [doi: 10.1136/bmj.l6890] [Medline: 31948937]
37. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme. Economic social and cultural rights (ESCR). 2006. URL: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf [accessed 2024-04-29]
38. Hoffman AJ, Brintnall RA, Brown JK, von Eye A, Jones LW, Alderink G, et al. Too sick not to exercise: using a 6-week, home-based exercise intervention for cancer-related fatigue self-management for postsurgical non-small cell lung cancer patients. Cancer Nurs 2013;36(3):175-188. [doi: 10.1097/NCC.0b013e31826c7763] [Medline: 23051872]
39. Groen WG, Kuijpers W, Oldenburg HS, Wouters MW, Aaronson NK, van Harten WH. Supporting lung cancer patients with an interactive patient portal: feasibility study. JMIR Cancer 2017 Aug 08;3(2):e10 [FREE Full text] [doi: 10.2196/cancer.7443] [Medline: 28790025]
40. Ji W, Kwon H, Lee S, Kim S, Hong JS, Park YR, et al. Mobile health management platform-based pulmonary rehabilitation for patients with non-small cell lung cancer: prospective clinical trial. JMIR Mhealth Uhealth 2019 Jun 21;7(6):e12645 [FREE Full text] [doi: 10.2196/12645] [Medline: 31228180]
41. Park S, Kim JY, Lee JC, Kim HR, Song S, Kwon H, et al. Mobile phone app-based pulmonary rehabilitation for chemotherapy-treated patients with advanced lung cancer: pilot study. JMIR Mhealth Uhealth 2019 Feb 04;7(2):e11094 [FREE Full text] [doi: 10.2196/11094] [Medline: 30714943]
42. Coats V, Moffet H, Vincent C, Simard S, Tremblay L, Maltais F, et al. Feasibility of an eight-week telerehabilitation intervention for patients with unresectable thoracic neoplasia receiving chemotherapy: A pilot study. Can J Respir Crit Care Sleep Med 2019 Apr 02;4(1):14-24. [doi: 10.1080/24745332.2019.1575703]
43. Shambhu PA, Jarugool T, Rubee D, Emily E, Nistha S, Cheryl K. FITT-CORRECT: updated dynamic and evidence-based principle of exercise prescription. J Nov Physiother Rehabil 2021 Feb 15;5(1):5-9. [doi: 10.29328/journal.jnpr.1001039]
44. Widberg C, Wiklund B, Klarare A. Patients' experiences of eHealth in palliative care: an integrative review. BMC Palliat Care 2020 Oct 14;19(1):158 [FREE Full text] [doi: 10.1186/s12904-020-00667-1] [Medline: 33054746]
45. Groen WG, Kuijpers W, Oldenburg HS, Wouters MW, Aaronson NK, van Harten WH. Empowerment of cancer survivors through information technology: an integrative review. J Med Internet Res 2015 Nov 27;17(11):e270 [FREE Full text] [doi: 10.2196/jmir.4818] [Medline: 26614438]
46. The growing value of digital health: evidence and impact on human health and the healthcare system. IQVIA Inc. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/the-growing-value-of-digital-health [accessed 2024-04-29]

XSL•FO
RenderX

47. Lupton D. The digitally engaged patient: self-monitoring and self-care in the digital health era. Soc Theory Health 2013 Jun 19;11(3):256-270. [doi: 10.1057/sth.2013.10]
48. The Forgotten "C"?: the impact of COVID-19 on cancer care. Macmillan Cancer Support. URL: https://www.macmillan.org.uk/dfsmedia/1a6f23537f7f4519bb0cf14c45b2a629/9601-10061/the-forgotten-c-the-impact-of-covid-on-cancer-care [accessed 2024-04-29]
49. Henshall C, Davey Z. Development of an app for lung cancer survivors (iEXHALE) to increase exercise activity and improve symptoms of fatigue, breathlessness and depression. Psychooncology 2020 Jan;29(1):139-147. [doi: 10.1002/pon.5252] [Medline: 31773808]

Abbreviations

HRQoL: health-related quality of life
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by K Williams; submitted 22.09.23; peer-reviewed by S Harding, C Holländer-Mieritz; comments to author 08.03.24; revised version received 01.05.24; accepted 18.05.24; published 26.07.24.

Please cite as:
Kirkpatrick S, Davey Z, Wright PR, Henshall C
Supportive eHealth Technologies and Their Effects on Physical Functioning and Quality of Life for People With Lung Cancer: Systematic Review
J Med Internet Res 2024;26:e53015
URL: https://www.jmir.org/2024/1/e53015
doi:10.2196/53015
PMID:

©Suriya Kirkpatrick, Zoe Davey, Peter Richard Wright, Catherine Henshall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Learning Outcomes of e-Learning in Psychotherapy Training and Comparison With Conventional Training Methods: Systematic Review

Kasperi Mikkonen[1,2], MA; Eeva-Eerika Helminen[1,2], MD, PhD; Samuli I Saarni[2,3], MD, PhD; Suoma E Saarni[2,3,4], MD, PhD

[1]Faculty of Medicine, University of Helsinki, Helsinki, Finland
[2]Department of Psychiatry, Helsinki University Hospital, Helsinki, Finland
[3]Faculty of Medicine and Health Technology, University of Tampere, Tampere, Finland
[4]Psychiatry, Wellbeing Services County of Päijät-Häme, Lahti, Finland

Corresponding Author:
Kasperi Mikkonen, MA
Department of Psychiatry
Helsinki University Hospital
Välskärinkatu 12
Helsinki, 00029 HUS
Finland
Phone: 358 405561374
Email: kasperi.mikkonen@gmail.com

Abstract

Background: Mental disorders pose a major public health problem in most western countries. The demand for services for common mental health disorders has been on the rise despite the widespread accessibility of medication. Especially, the supply and demand for evidence-based psychotherapy do not align. Large-scale increase of modern psychotherapy is difficult with current methods of training which are often expensive, time consuming, and dependent on a small number of top-level professionals as trainers. E-learning has been proposed to enhance psychotherapy training accessibility, quality, and scalability.

Objective: This systematic review aims to provide an overview of the current evidence regarding e-learning in psychotherapy training. In particular, the review examines the usability, acceptability, and learning outcomes associated with e-learning. Learning outcomes are assessed in different modalities including trainee experiences, knowledge acquisition, skill acquisition, and application of trained content in daily practice. Furthermore, the equivalence of web-based training and conventional training methods is evaluated.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a search from Ovid, MEDLINE, PsycINFO, and Scopus databases between 2008 and June 2022 was conducted. Inclusion criteria required studies to describe e-learning systems for psychotherapy training and assess acceptability, feasibility, or learning outcomes. The risk of bias was evaluated for both randomized and nonrandomized studies. Learning outcomes were categorized using the Kirkpatrick model. Effect sizes comparing e-learning and traditional methods were calculated.

Results: The search yielded 3380 publications, of which 34 fulfilled the inclusion criteria. Positive learning outcomes are generally associated with various e-learning programs in psychotherapy training including trainee satisfaction, knowledge, and skill acquisition, and in application of trained content in clinical practice. Learning outcomes generally show equivalence between e-learning and conventional training methods. The overall effect size, indicating this disparity, was 0.01, suggesting no significant difference. This literature displays a high level of heterogeneity in e-learning solutions and assessment methods.

Conclusions: e-Learning seems to have good potential to enhance psychotherapy training by increasing access, scalability, and cost-effectiveness while maintaining quality in terms of learning outcomes. Results are congruent with findings related to e-learning in health education in general where e-learning as a pedagogy is linked to an opportunity to carry out learner-centric practices. Recommendations for conducting psychotherapy training programs in blended settings supported by activating learning methods are presented. However, due to the heterogeneity and limitations in the existing literature, further research is necessary to replicate these findings and to establish global standards for e-learning, as well as for the assessment of training outcomes in psychotherapy

XSL•FO
RenderX

education. Research is especially needed on the effects of training on patient outcomes and optimal ways to combine e-learning and conventional training methods in blended learning settings.

(J Med Internet Res 2024;26:e54473) doi:10.2196/54473

KEYWORDS

mental health; psychotherapy; digital learning; Kirkpatrick model; e-learning; online health; psychotherapy training; learning outcome; learning outcomes; systematic review; training methods; mental disorders; mental disorder; accessibility; evidence-based; scalability; cost-effectiveness; internet; education

Introduction

Mental disorders pose a global public health problem [1]. Psychotherapy has been found to be effective across a variety of mental disorders [2-4]. However, access to evidence-based psychotherapy in public health care settings is often overly difficult and the supply and demand of evidence-based methods do not align [5]. One widely accepted strategy in meeting the existing and increasing demand of mental health services is to train health care professionals in various psychotherapies. The need for training encompasses both the training of new practitioners and updating the competences of existing professionals following the approach of continuing medical education.

Closing the gap between the demand for services and available resources through conventional training methods can be challenging. The main barriers in the dissemination of psychotherapy training include cost and time [6]. Conventionally, psychotherapy training is labor-intensive and relies heavily on the time provided by top-level professionals carrying out lectures and workshops that often lead to expensive and time-consuming programs and problems when striving to scale up. While psychotherapy training is linked to positive learning outcomes, the evidence on the impact of training on patient outcomes is mixed [7]. Digital solutions and e-learning have been acknowledged as a potential tool in expanding the reach of training, increasing the quality of training programs, and improving the dissemination of treatment [8-10]. With the comprehensive educational need, using digital solutions may be necessary.

e-Learning, which is often used synonymously with “digital learning,” “online learning,” and “web-based learning,” can be conceptualized as a framework that combines educational theories and technology or a practical application of educational technology that evolves in tandem with the advancements in technology and the internet [11,12]. The terminology is overlapping, and different labels give rise to diverse expectations [13]. One detail in terminology to highlight is the emphasis on “distance” between the trainer and the trainee, which is not a defining characteristic of e-learning [14]. It is tied to the trainee engagement and whether the training is held synchronously (eg, Zoom lectures) or asynchronously (eg, self-directed and individually paced learning in a web-based environment with multimedia materials and assignments).

Asynchronous e-learning with an emphasis on the active role of the trainee offers a possible solution for many current problems in psychotherapy training. First, e-learning offers a platform to carry out pedagogical principles that prioritize trainee engagement, as seen in approaches such as the flipped classroom, which involves a transition from trainer-led lectures to individual studying which happens before the synchronous learning event and which has been associated with improvements in learning outcomes compared with conventional training methods in health profession education [15,16]. Second, asynchronous e-learning offers flexibility in terms of time and place, which can remove some of the constraints involved in conventional training methods. Third, e-learning content can be scaled up more easily than conventional training methods. Fourth, while the current training programs are mostly rooted on various therapeutic traditions (or “schools”), which often hinders evidence-based development of psychotherapy, an adequate use of e-learning can be a way to create transparent training content which is easy to update and which fosters therapeutic competence [17-19]. Moreover, while the evidence regarding the cost-effectiveness of e-learning in health education is still in its infancy, e-learning is often associated as a means to increase the cost-effectiveness of training programs in the long term [20].

e-Learning is a widely used method for training in various industries and is generally associated with positive learning outcomes that are either comparable or superior to conventional training methods [21-23]. The literature examining e-learning in psychotherapy training is emerging, and its promise has been recognized [24]. However, an overview of the current evidence-base remains undetermined, and it remains unspecified how e-learning should be precisely used. For example, what is the role of e-learning in knowledge and skill acquisition, and is the value of e-learning most apparent when implemented as a blended learning curriculum, which combines conventional training methods with e-learning [19]?

The goals of this systematic literature review are to (1) provide an overview and assess the quality of the current literature; (2) determine the acceptability, usability, and achievable learning outcomes of e-learning in the context of psychotherapy training; and (3) assess how does e-learning compared with conventional training methods. Based on our results, we aim to formulate the current maturity and limitations of e-learning in this domain and discuss what kind of recommendations can be made regarding e-learning and psychotherapy.

Methods

Design

A systematic review was conducted to extract recently published scientific papers that dealt with e-learning in psychotherapy

XSL•FO
RenderX

training. We focused on the last 15 years (from 2008 onward) since rapid software and hardware development has occurred during that period although internet delivered and computerized e-learning as a concept emerged as early as the 1990s [25]. This review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [26].

Search Method

The search was conducted within Ovid MEDLINE, PsycINFO, and Scopus databases. The search strategy included a process where overlapping labels depicting psychotherapy and e-learning were combined. The search was formulated to cover both the training of larger treatment modalities and the specific therapeutic competences. Thus, the search strings were as follows: (psychotherap* or psycholog* or therap* or counsel*) adjacent to (skill* or train* or educat* or compentenc*) OR (interpersonal) adjacent to (skill* or train* or educat* or compentenc*) OR (psychotherap* or psychosoc* or mental health or behavioral health or behavioural health) adjancent to (intervent* or treat* or therap* or evidence-based intervent* or evidence-based treat* or evidence-based therapy* AND (digital* or mobile or web-based or online or distance) adjancent to (learn* or train* or educat* or program* or self-administered learn* or self-administered train* or self-administered educat* or self-administered program*) OR (e-learn* or elearn* or e-educat*). The searches were conducted on all databases in June 2022. To ensure optimal coverage considering the field's unestablished terminology, a comprehensive array of search terms was used.

Selection Criteria

Inclusion

To be included in this review, the studies should involve (1) a description of an asynchronous e-learning system (ie, a digital learning environment, eg, a website designed for self-directed learning) in the context of training psychotherapy competences and (2) measures of acceptability, feasibility, or learning outcomes.

Exclusion

Studies were excluded if they (1) represented an unpublished thesis or dissertation, (2) were not peer-reviewed, (3) were not experimental studies, (4) were unavailable in English, and (5) defined e-learning solely as synchronous distance learning (eg, Zoom lectures).

Study Selection

The eligibility of the search results was assessed by the first (KM) and last (SES) authors through three sequential steps: (1) initially, KM meticulously examined the titles and abstracts of all search results, systematically eliminating studies that distinctly met the specified exclusion criteria; (2) in cases where the title or abstract lacked sufficient detail to determine eligibility, KM proceeded to evaluate the full text of the study, applying the same exclusion benchmarks as in the initial phase; and (3) for the final step, KM scrutinized the full texts of all remaining studies, while SES independently reviewed a randomly selected subset of these studies to ensure representativeness. Any differences in opinion were subsequently resolved through a collaborative discussion targeting a unanimous decision. This stepwise approach ensured that studies were disqualified only if they unambiguously satisfied the predefined exclusion criteria. Given our anticipatory search terms' breadth, a substantial volume of studies not aligning with our inclusion criteria, such as those investigating digital treatment interventions rather than the training of professionals, was anticipated. Hence, to optimize efficiency, the first 2 screening stages were singularly executed by 1 reviewer (KM).

Data Collection

Examination of the titles and abstracts of the publications and the assessment for potential inclusion were carried out by first (KM) and last author (SES). Data extraction was carried out by the first author and included (1) target group of the training; (2) description of training content; (3) description of e-learning methods; (4) comparison between training modalities; (5) sample sizes; (6) primary outcome; and (7) summary of results. Furthermore, to calculate a measure of effect size for comparing differences between conventional training methods and e-learning, the means and standard deviations of the learning outcomes reported in the studies were extracted.

Risk of Bias

The risk of bias was assessed with the latest Cochrane risk-of-bias tools: RoB 2—tool for randomized trials and ROBINS-I (Risk of Bias in Non-randomized Studies of Interventions) for nonrandomized trials [27,28]. The risk of bias analysis was systematically conducted by the first (KM) and last (SES) authors in an independent manner. Subsequent to this individual assessment, any emergent discrepancies were collaboratively addressed through discussions focused on achieving consensus.

Framework for Classifying Results

We classified the results of the primary outcomes of the included studies using the Kirkpatrick model of training evaluation [29], which evaluates training outcomes on 4 levels. The first level ("reaction") evaluates the training participants' immediate reaction and feedback regarding the training program including satisfaction, engagement, and relevance. The second ("learning") evaluates the learning outcomes of the training program including change in attitudes, confidence, knowledge, and skills. The third level ("behavior") evaluates how much the trainee participants apply the learned knowledge and skills when treating patients in real life. The final level ("results") evaluates how much the patient outcomes improve because of the training.

Comparison Between Training Methods

In order to quantify the magnitude of the difference observed in these results, Hedges g was calculated as a measure of effect size with 95% CIs [30]. Hedges g is comparable with Cohen d but corrects small sample bias. Therefore, Hedges g of 0.2 represents a small effect; 0.5, a moderate effect; and 0.8, a large effect [31]. The effect size was calculated from the studies that included the figures necessary for the calculations. For studies with missing information, the first author of the publication was contacted, and they were asked for the data. An overall effect

XSL•FO
RenderX

size describing the equivalence between training methods was formulated as an arithmetic mean of all calculated effect sizes.

Results

Search Outcome

The first screening from the databases identified 3380 articles. Duplicates (n=697) were removed using the "Merge items" tool in the "Duplicate items"—folder in "Zotero," which the authors used as a reference management software. After removing the duplicates, 2683 articles were identified and the eligibility was assessed by KM based on their title, and abstract if necessary, and 2530 articles were excluded as clearly fulfilling the exclusion criteria mentioned in the "Selection Criteria" section of the article (see Figure 1). The remaining 153 articles were read in full-text versions by KM and 79 articles were excluded according to the same criteria. KM reviewed the remaining 74 articles in full, and SES reviewed in full, a random sample of those articles (n=45). The size of random sample required was calculated using Cohen κ for interrater agreement level with the assumptions that 30%-40% of the remaining articles would be included in the final sample, and the interrater agreement is moderate at the least [32]. The calculations were conducted with R with the "CIBinary"—function within the "kappaSize"—library. Based on these calculations, the last author reviewed the random sample of 45 articles. The authors worked blind to each other. The results were compared, and the authors were in substantial agreement on which articles to include and exclude (Cohen κ=0.73). All discrepancies were discussed to a consensus (eg, a detailed discussion whether the target of a training described in a specific study was related to psychotherapy skills or more general interaction skills used in health care). Thirty-four studies were included in the final review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

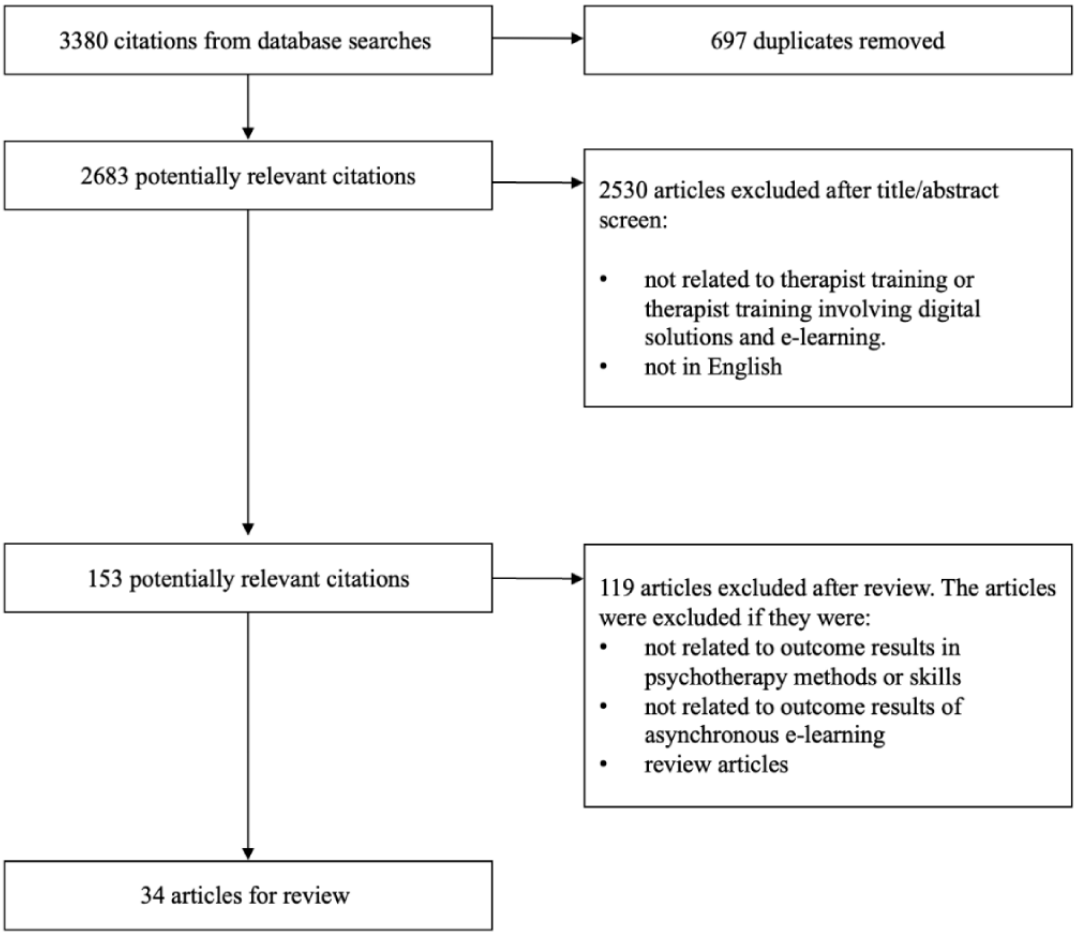

The articles included in the study are listed in Table 1 for randomized controlled trials (n=17) and Table 2 for nonrandomized (n=17) studies. The tables include the target group of the training, the content of the training, a depiction of the e-learning system used, a comparison between training modalities if applicable, the primary outcomes of the studies, and a summary of the main findings. In all studies, training was targeted to health care providers. The training content was most often related to some form of cognitive behavioral therapy (CBT) and the length of the e-learning programs ranged from 15 minutes to 52 hours. In the studies included, the effects of e-learning are evaluated in within-trainee designs and comparisons with waiting for training and other active training methods. The data were extracted by the first author as they were presented in the original studies.

XSL•FO
RenderX

Table 1. Randomized controlled trials included in the review.

Study	Target group	Training content	Description of e-learning	Comparison and sample sizes	Primary outcome	Summary of results
Shinohara et al (2022) [33]	Midwives and public health nurses	Perinatal mental health assessment and empathic communication	1-day course with short video scripts	Intervention group which received training (n=58) vs no-training control group (n=57)	Improvement of empathic communication	The intervention group was at a significantly higher level for empathic communication skills compared with the control group (*P*<.001)
Nisar et al (2022) [34]	Nursing school students	Evidence-based psychosocial intervention for perinatal depression	A tablet-based, multimedia training program including narrative scripts	e-Learning intervention group (n=42) vs conventional specialist–delivered training (n=47)	Competence in the trained intervention posttraining and at 3-month follow-up	e-Learning was not inferior to specialist-led training (*P*=.14)
Taylor et al (2021) [35]	Licensed behavioral health providers	Cognitive Behavioral Therapy for Insomnia	A website including video instructions and demonstrations, descriptions of intervention techniques, handouts, and exercises	e-Learning group (n=21) vs in-person workshop control (n=23)	Knowledge acquisition in CBT-I [a]	Knowledge acquisition was equivalent in both groups (*P*=.16)
Sansen et al (2020) [36]	Psychotherapists and psychotherapist trainees	Trauma-focused cognitive behavioral therapy	52-hour web-based training course including written information, animations, video clips, various assignments, and simulated cases	e-Learning group (n=247) vs wait-list control group (n=252)	Subjective ratings of knowledge, confidence, competencies, and willingness to use trauma-focused intervention and performance in a knowledge test	Significantly greater improved subjective ratings of knowledge, confidence, competencies, and willingness, as well as performance in the knowledge test in the e-learning group compared with the control group (all *P* values <.01)
Sansen et al (2019) [37]	Psychotherapists and psychotherapist trainees	Trauma-focused cognitive behavioral therapy	52-hour web-based training course including written information, animations, video clips, various assignments, and simulated cases	e-Learning group (n=247) vs wait-list control group (n=252)	Fears and reservation toward conducting trauma-focused methods	A significant reduction of fears and reservations during e-learning (*P*<.01)
Kullberg et al (2020) [38]	Undergraduate psychology students	Suicide prevention practice	1-hour web-based training including videos of professionals interacting with various patients with suicidal tendencies	e-Learning group (n=211) vs wait-list control group (n=187)	Self-reported guideline adherence, knowledge acquisition, and provider confidence	Significantly greater improvements in self-reported knowledge, confidence, and guideline adherence in the e-learning group compared with the control group (all *P* values <.01). Learning effects were maintained in 3-month follow-up
Rahman et al (2019) [39]	Nonspecialist health workers	Psychosocial management of perinatal depression	A tablet-based training course including narrative scripts, reflective assignments, and video role-plays	e-Learning group (n=30) vs face-to-face training control group (n=30)	Competence in intervention	No differences between e-learning and face-to-face training in competence scores post assessment and at 3-month follow-up (*P*>.05)

XSL•FO
RenderX

Study	Target group	Training content	Description of e-learning	Comparison and sample sizes	Primary outcome	Summary of results
Puspitasari et al (2017) [40]	Mental health clinicians	Behavioral activation	4 × 40-minute web-based audio-guided presentations of core BA[b] skills	e-Learning intervention group (n=40) vs trainer-led training (n=37)	BA skills competence	Significantly greater increases in all BA skills in trainer-led training compared with web-based training (*P*<.0005)
Stein et al (2015) [41]	Mental health clinicians	Interpersonal and social rhythm therapy for bipolar disorder	12-hour self-paced web-based training	e-Learning group + 3-6 months of monthly telephone supervision (n=16) vs control group in 2-day workshop and weekly supervision (n=20)	Use of IPSRT[c] techniques during treatment	A significant increase in the use of IPSRT techniques in posttraining which increased in 12-month follow-up (*P*<.01). The increase was comparable in both groups
Dimeff et al (2015) [42]	Mental health clinicians	Two core strategies in DBT[d]: chain analysis and validation	Interactive, media-rich web-based training modules with expert commentaries, exercises, simulations, printable handouts, and knowledge checks	Web-based training (e-learning) (n=55) vs ILT[e] (n=55) vs TM[f] (n=62)	Satisfaction, self-efficacy, motivation, knowledge, clinical proficiency, and clinical use	ILT was superior compared with web-based training (e-learning) and TM in satisfaction (*P*<.0001), self-efficacy (*P*<.001), and motivation (*P*<.05). Web-based training (e-learning) was the most effective method for increasing knowledge (*P*<.01). There were no differences between observer-rated proficiency (*P*=.76) and self-reported clinical use (*P*=.48), which increased during training
Harned et al (2014) [43]	Mental health providers	Exposure therapy for anxiety disorders	10-hour course incorporating gaming technology and evidence-based instructional design strategies. The learner could choose whether to engage with media-rich didactic materials or simulated clinical scenarios	Web-based training (e-learning) (n=60) vs web-based training (e-learning) + brief computerized motivational enhancement intervention (web-based training [e-learning] + ME[g]) (n=60) vs web-based training (e-learning) + ME + a web-based learning community (web-based training (e-learning) + ME + LC[h]) (n=61)	Knowledge acquisition, attitudes toward exposure therapy, self-efficacy, clinical use, and clinical proficiency	All training methods comparably improved self-efficacy and clinical use of exposure therapy. The increase in attitudes in the web-based training (e-learning) + ME + LC group at 6-week follow-up was significantly larger than in the web-based training (e-learning) group (*P*=.01) as well as in clinical proficiency at 12-week follow-up (*P*=.04). Web-based training (e-learning)–ME-LC condition was also superior in knowledge acquisition (*P* values =0.004-0.05)

XSL•FO
RenderX

Study	Target group	Training content	Description of e-learning	Comparison and sample sizes	Primary outcome	Summary of results
Ruzek et al (2014) [44]	Mental health clinicians with responsibilities in treating veterans with PTSD[i]	CBT[j] skills for traumatic stress: motivation enhancement, goal setting, and behavioral task assignment	3-module web-based training with interactive exercises, skills demonstrations, downloadable materials, and bibliographies	e-Learning group (n=46) vs e-learning + telephone consultation (n=42) vs no-training or training as usual (free to take part in any training activities they would otherwise receive) (n=51)	Skills acquisition	Skills improved significantly in both active training groups compared with the control group in motivation enhancement and goal setting (*P*<.001, *P*<.005). Web-based training plus supervision was superior compared with web-based training alone in the motivation enhancement module (*P*=.001) with no significant differences in the 2 other modules.
Larson et al (2013) [45]	Addiction counsellors and supervisors	Cognitive behavioral therapy	Web-based training including various exercises, audio-vignettes and role-played treatment sessions, and an assessment between modules	e-Learning (n=47) vs training manual (n=52)	Adherence to CBT practice	No statistically significant difference in pass rates and adequacy scores in adherence to CBT practice between groups (*P*=.19). The posttraining gains in CBT skill adequacy were small
Bennett-Levy et al (2012) [46]	Practicing in a counseling and mental health role with a relevant degree or students closely completing a degree	Cognitive behavioral therapy	30-module web-based training with didactive text and video demonstrations with case examples and quizzes	e-Learning group (n=23) vs supported training group with the possibility of six 15-minute support calls (n=24)	Completion rate, knowledge acquisition, self-reported skills acquisition, confidence in CBT, and skills use in practice	Both groups demonstrated similar improvements in objective knowledge test and on self-reported measures of knowledge, skills, confidence, and use of skills at follow-up (improvement all significant at *P*<.001). Supported training group had a significantly higher completion rate (*P*<.05).
Dimeff et al (2011) [47]	Community mental health providers	DBT distress tolerance skills	A 5-module web-based training with media-rich content, expert insights, practice exercises, and knowledge checks with downloadable handouts	e-Learning (n=47) vs training manual (n=43) vs placebo e-learning (n=42)	Knowledge acquisition, self-reported self-efficacy, and motivation to use learned skills	Web-based training and the TM outperformed the control condition in knowledge acquisition (*P*<.05) and self-efficacy (*P*<.05) but not in motivation to learn and use the treatment. By the 15-week follow-up, the web-based training outperformed manual in knowledge acquisition (*P*<.05)

XSL•FO
RenderX

Study	Target group	Training content	Description of e-learning	Comparison and sample sizes	Primary outcome	Summary of results
Dimeff et al (2009) [48]	Community mental health providers	Dialectical behavior therapy	A 20-hour 5-module web-based training with media-rich content, expert insights, clinical vignettes, practice exercises, and knowledge checks	e-Learning (n=54) vs training manual (n=49) vs ILT (2-day workshop) (n=47)	Satisfaction, knowledge acquisition, confidence, self-efficacy, and application of trained content. Performance in structured role-play	Higher satisfaction in web-based training and ILT than in training-manual (*P*<.001), but no differences between web-based training and ILT (*P*=.33). Web-based training outperformed the other conditions in knowledge acquisition (*P*<.05) while there were no differences between ILT and training manual (*P* values >.90). At posttraining, web-based training and ILT conditions reported more elevated self-efficacy than the manual condition (*P*<.01) but the differences diminished in follow-up. All training methods resulted in comparable increase in the application of trained content in clinical situations (*P*=.68). Adherence and competence in role-play scenario improved significantly in all conditions but there was no difference between conditions (*P* values >.20)
Weingard et al (2006) [49]	Substance abuse counsellors	A cognitive behavioral approach in treating cocaine addiction	A 60-minute web-based training module with pre- and posttests	e-Learning (n=52) vs face-to-face workshop (n=55) vs delayed training (n=59)	Knowledge acquisition	Improvements in knowledge were significant comparable in both web-based training and face-to-face workshops

[a]CBT-I: Cognitive Behavioral Therapy for Insomnia.

[b]BA: behavioral activation.

[c]IPSRT: interpersonal and social rhythm therapy for bipolar disorder.

[d]DBT: dialectical behavior therapy.

[e]ILT: instructor-led training.

[f]TM: treatment manual.

[g]ME: motivational enhancement.

[h]LC: learning community.

[i]PTSD: posttraumatic stress disorder.

[j]CBT: cognitive behavioral therapy.

XSL•FO
RenderX

Table 2. Nonrandomized trials included in the review.

Study	Target group	Training content	Description of e-learning	Comparison and sample size(s)	Primary outcome	Summary of results
Wilkerson et al (2022) [50]	Health and mental health professionals, educators, and trainees	Cognitive Behavioral Therapy for Insomnia	6- to 10-hour 10-module web-based training with didactic videos, therapy vignettes, and resources regarding the specific modules	N=2586	Training completion rate and knowledge acquisition	1203/2586 (46.52%) of registered users started the training and 570/2586 completed the training with a 3-month period with both pre-and posttests complete. A significant increase in knowledge was reported (*P*<.001)
Seidler et al (2022) [51]	Mental health practitioners	Engaging men in therapy and dealing with mental distress and suicidality experiences by males	8-hour web-based media-rich asynchronous training program including written information, downloadable worksheets, and videos	N=192	Improvement in self-reported competencies in engaging male clients	Significant increase in self-reported competencies (*P*<.001)
Soll et al (2021) [52]	Postgraduate health care professionals	Cognitive behavioral therapy for psychosis	Prerecorded lectures, exercises, web-based role-plays, and web-based conferences	e-Learning (n=85), in-person training (n=142)	Satisfaction and acceptance	Participants' satisfaction and acceptance toward the web-based training were noninferior compared with in-person training in all respects except for "room for active participation" and "professional benefit"
Applebaum et al (2021) [53]	Health care professionals working in psychosocial oncology	Acute cancer cognitive therapy	6-hour web-based training course including written information and various assignments	N=46	Knowledge acquisition	A significant increase in knowledge (*P*<.001)
Gonzalez Salas Duhne et al (2020) [54]	Professionals delivering psychological intervention	Implementation intentions ("if-then"—planning)	A 15-minute video-based training program with pre- and postquestionnaires	N=87	Knowledge acquisition, acquisition of practical knowledge, and self-reported use of implementation intentions	Significant increase in knowledge both posttraining and in 1-month follow-up (*P*<.001). No significant increase in practical knowledge. A significant increase in use of implementation intentions (*P*=.018)
Graber (2019) [55]	Psychiatric mental health nursing students	Therapeutic crisis management techniques	A 30-minute presentation in a web-based educational environment	e-Learning (n=47), in-person training (n=63)	Knowledge acquisition and student satisfaction	No differences between e-learning and in-person training in knowledge acquisition (*P*=.483) or student satisfaction (*P*=.87)

XSL•FO
RenderX

Study	Target group	Training content	Description of e-learning	Comparison and sample size(s)	Primary outcome	Summary of results
Cramer et al (2019) [56]	Mental health professionals	Core competencies in suicide prevention	20-hour self-paced narrated presentations with associated resources and handouts, case descriptions, and various assignments	N=43	Knowledge acquisition, attitudes toward self-harm patients, and skills performance	Improvement in knowledge (*P*<.001), perceived skill mastery (*P*<.001), capacity to work with suicidal patients (*P*<.001), objective risk judgment accuracy (*P*<.001), and reduction of negative feelings about self-harming patients (*P*<.007)
Meredith et al (2018) [57]	Nurses and occupational therapists in mental health settings	Sensory modulation approaches in mental health	A 2-hour, 3 self-paced modules including interactive activities, case scenarios, opportunities of self-reflection, and links to further reading	N=121	Knowledge acquisition and change in perceived level of knowledge confidence and attitudes	Significant increase in knowledge acquisition, confidence, and attitude toward sensory modulation approaches (all *P* values <.001)
German et al (2018) [58]	Mental health clinicians	Cognitive behavioral therapy	A self-paced web-based training including videotaped role-playing, on-screen activities, and quizzes	e-Learning with peer-led consultation (n=148), in-person training with expert-led consultation (n=214)	CBT[a] competency, knowledge acquisition, and learner retention	e-Learning with peer-led consultation was not inferior to in-person training with expert-led consultation in increasing clinical competency. No differences between knowledge acquisition between groups. Trainees in the web-based training program were less likely to complete the training (*P*<.001)
Kuhn and Hugo (2017) [59]	Psychiatry residents	Prolonged exposure to posttraumatic stress disorder	A 7- to 8-hour 13 web-based modules including step-by-step instructions, demo videos of expert clinicians, printable scripts, pre- and posttests, and a “common questions” section	N=12	Trainee impressions	The web-based training used in a blended training program may help address the common problems in traditional training models. The first impressions of the program support its viability, practicality, and efficiency
Kobak et al (2017) [60]	Mental health clinicians	Interpersonal therapy	A 4-hour web-based training with animations, graphical illustrations, clinical vignettes, and various exercises	N=26	Knowledge acquisition	Significant increase in knowledge of interpersonal therapy (*P*<.001)

XSL•FO
RenderX

Study	Target group	Training content	Description of e-learning	Comparison and sample size(s)	Primary outcome	Summary of results
Fairburn et al (2017) [61]	Mental health clinicians	Cognitive Behavioral Therapy for Eating Disorders	A 9-hour web-based training providing information on how to implement CBT-E[b] with videos, handouts, learning exercises, and tests of knowledge with feedback. A separate digital library of detailed materials of therapy implementation with various subgroups	N=80	Competence in CBT-E	A significant increase in competence scores after completing the training (*P*<.001). Of the trainees 42.5% reached a predetermined cutoff point indicating good competence
Curran et al (2015) [62]	Counsellors working with individuals with substance use disorders	Manualized CBT group treatment for depression	A 12- to 16-hour media-rich and interactive digital training	N=7	Feasibility of self-paced digital training in replacement of in-person training. Observed barriers and facilitators, recommendations for improvement, and concerns of implementation	The main barrier concerning the completion of the training was the lack of protected time. The main facilitators included positive attitudes toward the training, supervisor support, counselor dedication, and beliefs supporting providing services for depression
Hickey et al (2015) [63]	Psychiatry residents	Two core topics in IS-TDP[c]	Two asynchronous e-learning modules of "the metapsychology of the unconscious" and "Davanloo's technique of IS-TDP"	e-Learning (n=19), in-person training (n=19)	Knowledge acquisition	Significant improvement in knowledge levels during training in both topics (*P*<.01). No differences between knowledge acquisition between web-based training and lectures (*P*<.05)
Samuelson et al (2014) [64]	Primary care providers	Detection, assessment, initial management, and referral of patients with PTSD[d]	70-minute web-based training including multimedia didactic content, case presentations, and clinical vignettes	N=73	Knowledge acquisition, comfort regarding PTSD-related skills, and usage in clinical practice	Significant increase in PTSD-related knowledge (*P*<.001) and significantly greater comfort regarding skills compared with pretraining (*P*<.001). 47% reported using the training content in clinical practice
Kobak et al (2013) [65]	Social workers, psychologists and social work, and psychology students	CBT for anxiety disorders	A 5-hour 10-module web-based training with interactive exercises, animations, graphical illustrations, and expert demonstration videos	N=39	Knowledge acquisition and satisfaction	A significant increase in knowledge in CBT (*P*<.001) and trainees were satisfied in both the content and the technical aspects of the web-based training

XSL•FO
RenderX

Study	Target group	Training content	Description of e-learning	Comparison and sample size(s)	Primary outcome	Summary of results
Martino et al (2011) [66]	Counsellors working with substance use disorders	Motivational interviewing	Four 1-hour modules with instructions for self-practice with reflections on a discussion board managed by a course instructor	N=26	Adherence and competence in motivational interviewing	The web-based training was a part of a stepwise training program with workshops and supervision after the web-based training which were offered only if the learner showed inadequate competence after the web-based training. The participants who performed adequately after the web-based training did not improve in subsequent training. Participants who showed inadequate competence after the web-based training continued to improve in subsequent training

[a]CBT: cognitive behavioral therapy.

[b]CBT-E: Cognitive Behavioral Therapy for Eating Disorders.

[c]IS-TDP: intensive short-term dynamic psychotherapy.

[d]PTSD: posttraumatic stress disorder.

Risk of Bias

Risk of bias was assessed both for the randomized and nonrandomized studies included in the review using the Cochrane risk of bias tools (RoB 2 and ROBINS-I). The risk of bias analysis was carried out by the first and last authors (KM, SES). The authors agreed upon the used tools and familiarized themselves with the logic behind them. The first author did the risk of bias analysis and classification. Then the last author examined the classification to identify potential discrepancies. No discrepancies occurred. In the review, 13 out of the 17 randomized studies, which corresponds to 76%, have been categorized as exhibiting a low risk of bias. One of the studies (6%) showed some concerns regarding the risk of bias. Three of the studies (17%) were flagged for having a high risk of bias. One of these studies had a source of bias from the randomization process and the 2 others have a potential source of bias from a large amount of missing outcome data. These results can be viewed in Figure 2. The ROBINS-I classification for risk of bias in nonrandomized studies included in the review can be viewed in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Risk of bias assessment for the randomized trials included in the review using the Cochrane risk of bias tool (RoB 2) [33-49].

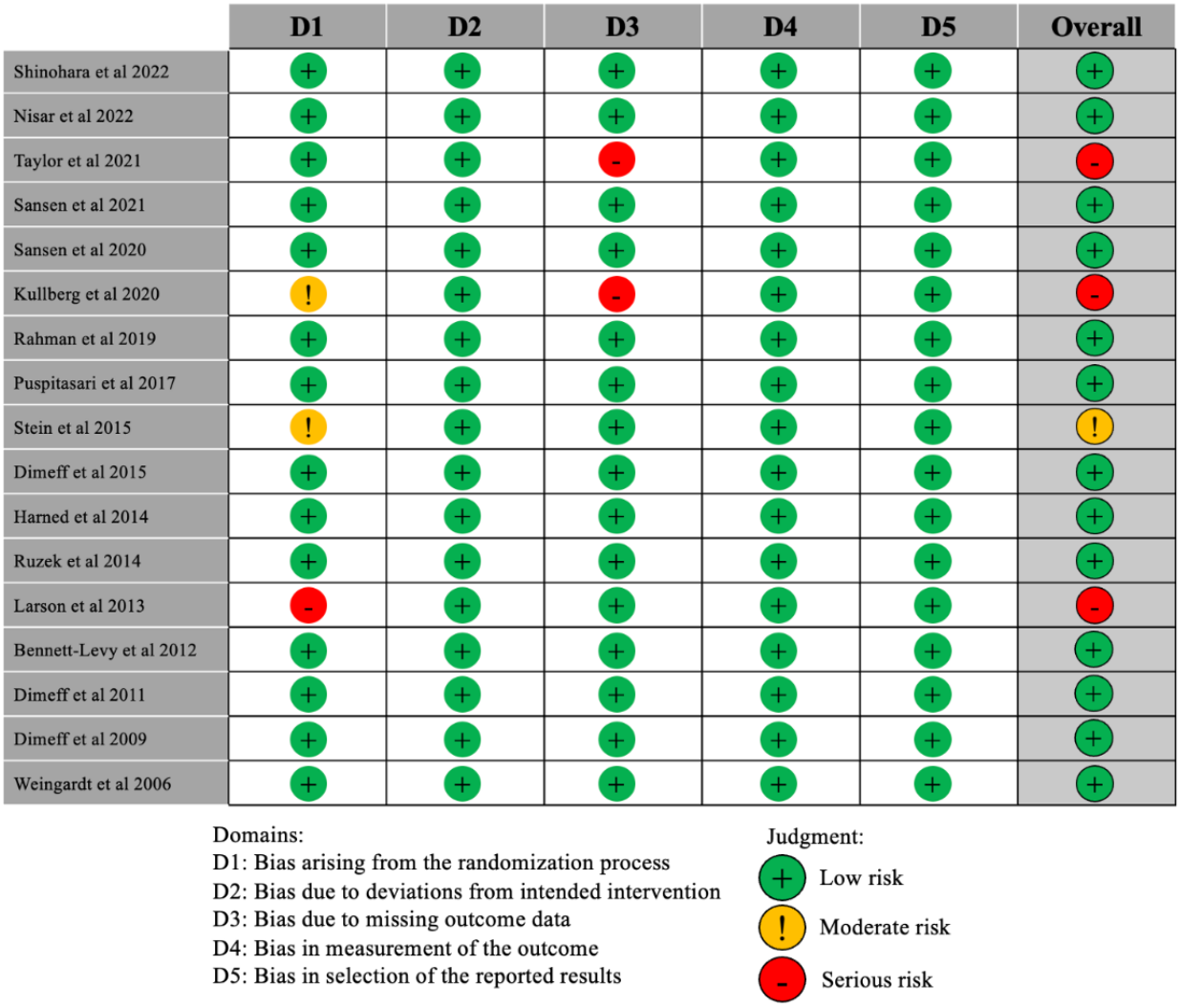

Among the nonrandomized studies included in the review, 16 out of 17 (94%) were flagged for having either a serious or critical risk of bias. The source of bias in all these studies was, at the very least, inadequate control of potential confounding variables. One study [58] was flagged at being only at moderate risk of bias.

Learning Outcomes According to the Kirkpatrick Model Subcategories

Kirkpatrick: Reaction

Out of the 34 articles included in the review, 7 (21%) presented findings related to the "Reaction" level of the Kirkpatrick model. These findings included such concepts as satisfaction, acceptance, relevance, first impressions, viability, practicality, and feasibility conceptualized as observed barriers and facilitators regarding implementation of the trained content. E-Learning was associated with positive changes in these outcomes in training programs for dialectical behavior therapy (DBT) [48], CBT for psychosis [52], therapeutic crisis management [55], CBT for depression [62], prolonged exposure to posttraumatic stress disorder (PTSD) [59], and CBT for anxiety disorders [65].

Kirkpatrick: Learning

Out of the studies included in the review, 30 out of 34 (88%) presented findings related to the "Learning" level of the Kirkpatrick model. These included concepts related to acquisition of knowledge and skills, as well changes in attitudes and the level of self-efficacy. E-Learning was associated with positive changes in these outcomes in training programs for perinatal mental health assessment and empathic communication [33], evidence-based intervention in perinatal depression [34], CBT for insomnia [35], trauma-focused CBT [36], suicide prevention practices [38], behavioral activation [40], DBT [42,47,48], exposure therapy for anxiety disorders [43], CBT skills for traumatic stress [44], CBT [45,46,58], CBT for addictions [49], skills in engaging men in therapy [51], CBT addressed for patient with acute cancer [53], implementation intentions [54], therapeutic crisis management techniques [55], suicide prevention [67], sensory modulation approaches [57], interpersonal therapy [60], CBT for eating disorders [61], short-term psychodynamic therapy [63], management of PTSD [55], and motivational interviewing [66].

Kirkpatrick: Behavior

Out of the studies included in the review, 7 out of 34 (21%) reported results that can be classified to the "Behavior" level

XSL•FO
RenderX

of the Kirkpatrick model which in all cases was reported as application of the trained content in clinical practice. E-Learning was generally associated with positive changes in these outcomes in training programs for interpersonal and social rhythm therapy for bipolar disorder [41], DBT [42], exposure therapy for anxiety disorders [43], CBT [46], and implementation intentions [54].

Kirkpatrick: Results

No studies included in this review display results that could be classified in the "Results" level of the Kirkpatrick model.

Comparison Between e-Learning and Conventional Training Methods

Thirteen out of 34 (38%) studies, in which 9 were randomized trials [34,35,39-42,46,48,49] and 4 were nonrandomized [52,55,58,63], reported a comparison between a web-based training and training carried out by conventional methods. Of the 13 studies examined, 2 (15%) [42,48] reported results suggesting superiority (related to knowledge acquisition outcomes) for e-learning compared with conventional training. Meanwhile, 3 (23%) [40,42,52] indicated inferiority (concerning outcomes related to skills acquisition, satisfaction, self-efficacy, motivation, opportunity for active participation, and trainee-assessed professional benefit). In addition, 12 (92%) [34,35,39,41,42,46,48,49,52,55,58,63] found no significant difference between e-learning and conventional training methods, reporting equivalence (in terms of outcomes related to knowledge acquisition, skills acquisition, clinical application of the methods trained, self-efficacy, satisfaction, and acceptance rates) at posttraining evaluations. It is worth noting that studies reporting multiple outcomes could be included in more than 1 category.

Eight out of these 13 studies included the necessary figures to calculate Hedges *g*. The first authors of the remaining studies were contacted, but none provided the figures or raw data within an acceptable time window of 1 month. The overall mean effect size from all comparisons within the 8 studies (n=30) was 0.01. A forest plot displaying effect sizes and their corresponding 95% CIs is shown in Figure 3. The results are organized and presented in accordance with the Kirkpatrick model categorization.

Figure 3. Forest plot of Hedges g effects sizes depicting the difference between e-learning and conventional training methods in Kirkpatrick levels 1-3 [34,39,42,46,48,52,55,58]. CBT: cognitive behavioral therapy; DBT: dialectical behavior therapy; EBM: evidence-based method.

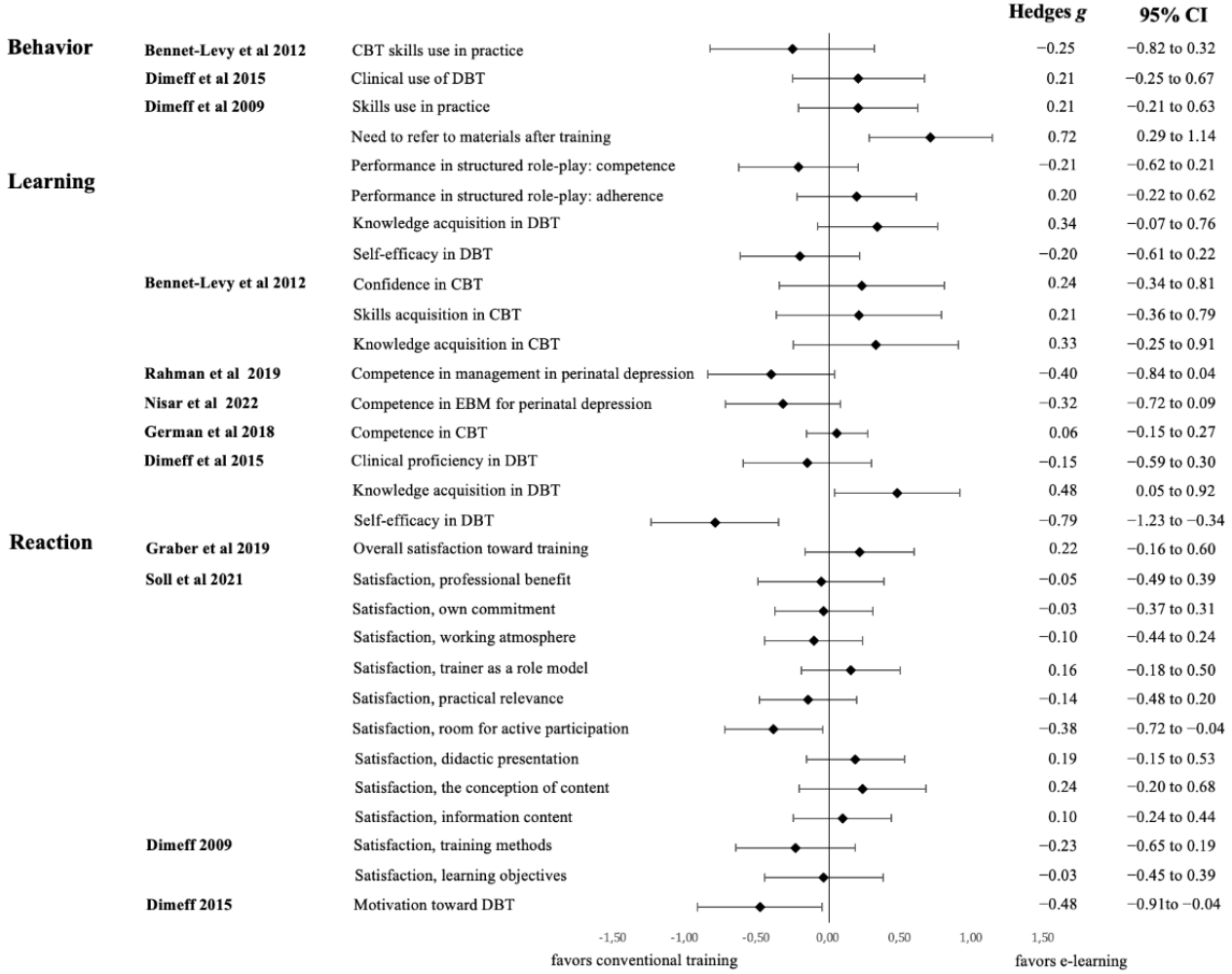

XSL•FO
RenderX

Discussion

Principal Results

A systematic literature review of e-learning in the context of psychotherapy training was conducted following the PRISMA guidelines. E-Learning was positively associated with learning outcomes in terms of knowledge and skills acquisition as well as implementing learned skills to practice according to the Kirkpatrick levels 1-3. There were no apparent differences between e-learning and conventional training methods. None of the studies examined the impact of training on patient outcomes, and the research demonstrated significant heterogeneity.

This is the first study to exclusively assess the level of research regarding e-learning within this domain. The review included 34 studies conducted between 2008 and 2022, encompassing a mix of both randomized and nonrandomized research. Up to this point, findings on training outcomes have been reported across various outcomes and the level of evidence is still emerging. The target group, training content, a description of e-learning, sample sizes, primary outcomes, and a summary of results were provided from each article included in the review. Most of the randomized studies were linked with a low risk of bias, whereas nonrandomized studies exhibited a severe risk of bias. The risk of bias in the majority of the nonrandomized studies stemmed from limited control of confounding variables. The execution method of e-learning, the measurement methods used in training outcome assessments, and the trainees were highly heterogenous across the included studies. The high amount of heterogeneity in various aspects of the included literature and small sample sizes in many of the included studies create challenges in synthesizing the findings and drawing clear conclusions. Therefore, it is necessary to process the literature using a higher-lever categorization in learning outcomes (such as the Kirkpatrick model).

e-Learning within psychotherapy training showed positive associations with several aspects: trainee satisfaction, enhanced learning outcomes—including knowledge and skills acquisition, and the application of trained methods into practice, as observed in within-trainee designs. These findings align with levels 1-3 of the Kirkpatrick model. No outcomes related to patient outcomes were reported (Kirkpatrick level 4). The results of this review point in the direction of equivalence between e-learning and conventional training methods such as lectures and workshops. Training outcomes including trainee satisfaction, knowledge, and skill acquisition, and application of trained content in clinical work were comparable between e-learning and conventional training. The overall effect size signifying this difference was 0.01, indicating no difference.

In the only study that reported superiority of conventional training methods in skills acquisition e-learning involved solely watching previously recorded lectures rather than actively interacting with content specifically designed as an e-learning program [40]. In the other studies, where the inferiority of e-learning was presented as a result with a more modern approach on e-learning, the inferiority was in the context of trainee satisfaction, self-efficacy, and motivation [42,52]. E-Learning was reported to be superior compared with conventional training only in the context of knowledge acquisition when training DBT [42,48].

The trainees and the clinical contexts of the trainees varied heavily between included studies. Furthermore, the operationalization of the learning outcomes and the associated measurement methods were not consistent and included both subjective self-ratings of knowledge and competence and more objective measures (eg, quizzes and expert-rated competence scores). In most of the studies, the training content included CBT-based methods, and in some studies interpersonal therapy and short-term psychodynamic therapy. Most of the e-learning solutions described were similar and included case vignettes, demonstrations, multimedia didactic content, and various assignments.

Strengths and Limitations

This review has many strengths. These include following the PRISMA guidelines, giving an exhaustive look in the current literature, and that the results of research comparing e-learning and conventional training methods were uniformly aggregated behind a single statistic. As anticipated, studies identified in this review had high heterogeneity of learning methods, therapy content, and outcomes, making systematic comparison difficult. We tried to control this heterogeneity by using the Kirkpatrick model of training evaluation, summarizing the results by calculating effect sizes, and requesting original data from authors to do that. However, it seems that heterogeneity is a feature of this field of study and cannot be easily overcome with post hoc methods; a more uniform research methodology would be needed in future studies.

This systematic review also has limitations. First, the review did not include gray literature (materials and research produced outside of academic publishing) or literature beyond the scope of the conducted search, which may include information relevant to this study. Second, in addition to the risk of bias assessment, the quality of the research included was not systematically assessed (eg, publication bias). Third, the search method yielded a substantial amount of literature not relevant to the study, which makes the replication of this review challenging. This is mainly attributed to the overlap in vocabulary between literature related to training and psychosocial interventions themselves. This issue also touches upon a possible methodological limitation in our screening process, namely, that the initial screening phases were conducted by a sole reviewer. The purpose of this preliminary screening was to promptly identify and eliminate studies that were unmistakably irrelevant to our review, with a large proportion of such studies investigating digital psychosocial interventions (eg, internet-based cognitive behavioral therapy), a testament to the prevalent overlap in terminologies used. To ensure a balanced review, the authors applied a statistical approach to select a representative sample of studies for evaluation by the second author. We posit that the inclusion of 2 independent raters from the outset, as opposed to 1, during the initial screening phases might have had a negligible effect on the outcomes, considering that all studies with any element of ambiguity were carried forward for full review. Nonetheless, we acknowledge that our approach may

XSL•FO
RenderX

be more susceptible to bias compared with methodologies that use multiple independent reviewers for the assessment of all studies. We consider this as a potential limitation and agree that the use of dual raters throughout the process might augment the robustness against bias.

Conclusions

Based on this literature review, we conclude that positive learning outcomes are generally associated with various e-learning programs in psychotherapy training including trainee satisfaction, knowledge and skill acquisition, and in application of trained content in clinical practice. Learning outcomes are generally equivalent between e-learning and conventional training methods. This equivalence is interpreted as a supportive finding for e-learning, considering the other benefits of e-learning, such as scalability and consistency of quality. While more research is warranted, and research methodology needs to be more consistent in the description of e-learning solutions and in learning outcome assessment, it seems advisable to use e-learning as a part of psychotherapy training curriculums.

e-Learning creates an opportunity to scale up training in large and even in nationwide settings. This, in turn, can significantly improve equal and sufficient access to psychotherapy in public health care settings, which are impossible with conventional and unscalable training methods [61]. This can be seen, for example, within the study by Wilkerson et al [50], where it was estimated that 2500 individuals will annually complete an e-learning program in Cognitive Behavioral Therapy for Insomnia, a scale the authors declared as “massive dissemination.” With centralized digital solutions, e-learning might prove itself as the more cost-effective solution than conventional training methods while ensuring quality in training as many learning outcomes are comparable between these methods.

As the literature is still emerging and current corpus is displaying a high amount of heterogeneity, small sample sizes, and inconsistent description of e-learning solutions and measurement methods in learning outcomes, it would be advisable to use e-learning as a part of blended training programs. By comparison with e-learning results in health education in general, a blended training model might even be the most optimal way to implement e-learning. E-Learning is generally recognized as an opportunity to improve the availability of training programs by eliminating constraints set by time and geographical location and also a way to increase the quality of training programs by shifting the focus of the surrounding pedagogics toward more learning-centric direction. Knowledge acquisition can be supported through e-learning, but refining skills and ensuring competencies may require synchronous interaction with a trainer to reach optimal learning outcomes [67].

In psychotherapy training, a shift toward minimizing the didactic approach of trainers by facilitating the active role of the trainee in knowledge acquisition is warranted. E-Learning viewed as a pedagogy rather than technology may be an optimal solution in implementing flipped classroom and other activating learning methods. Within these methods, synchronous learning events are held with trainees who have already familiarized themselves with the necessary materials and focus on honing therapeutic and reflective skills and addressing student motivation. Future research needs to focus on replicating the effects of e-learning in trainee satisfaction, learning and behavior, and on studying the effectiveness and cost-effectiveness of training (e-learning or conventional methods) on patient outcomes. Research is especially needed to further determine the optimal way to combine e-learning and conventional training methods. Furthermore, to reach the full potential of e-learning in the context of psychotherapy training, international collaboration is warranted to bring consistency on how e-learning systems are developed, described, and assessed.

Acknowledgments

KM conceived and designed the study with the advice from SES. SES, EH, and SIS reviewed, analyzed, and interpreted the data and contributed to drafting, revising, and finalizing the manuscript. All authors read and approved the final version of this manuscript.

Data Availability

The data extracted from the included studies are available upon request from the first author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ROBINS-I (Risk of Bias in Non-randomized Studies of Interventions) classification for risk of bias in nonrandomized studies included in the review.
[DOCX File , 255 KB - jmir_v26i1e54473_app1.docx]

Multimedia Appendix 2
PRSIMA 2020 checklist.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 128 KB - jmir_v26i1e54473_app2.pdf]

References

1. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
2. Barth J, Munder T, Gerger H, Nüesch E, Trelle S, Znoj H, et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. Focus (Am Psychiatr Publ) 2016;14(2):229-243 [FREE Full text] [doi: 10.1176/appi.focus.140201] [Medline: 31997951]
3. Cuijpers P, Miguel C, Harrer M, Plessen CY, Ciharova M, Ebert D, et al. Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: a comprehensive meta-analysis including 409 trials with 52,702 patients. World Psychiatry 2023;22(1):105-115 [FREE Full text] [doi: 10.1002/wps.21069] [Medline: 36640411]
4. van Dis EAM, van Veen SC, Hagenaars MA, Batelaan NM, Bockting CLH, van den Heuvel RM, et al. Long-term outcomes of cognitive behavioral therapy for anxiety-related disorders: a systematic review and meta-analysis. JAMA Psychiatry 2020;77(3):265-273 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.3986] [Medline: 31758858]
5. Harvey AG, Gumport NB. Evidence-based psychological treatments for mental disorders: modifiable barriers to access and possible solutions. Behav Res Ther 2015;68:1-12 [FREE Full text] [doi: 10.1016/j.brat.2015.02.004] [Medline: 25768982]
6. Stewart RE, Chambless DL, Baron J. Theoretical and practical barriers to practitioners' willingness to seek training in empirically supported treatments. J Clin Psychol 2012;68(1):8-23 [FREE Full text] [doi: 10.1002/jclp.20832] [Medline: 21901749]
7. Knox S, Hill C. Training and supervision in psychotherapy: what we know and where we need to go. In: Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. Hoboken, NJ: John Wiley & Sons Inc; 2021:327-349.
8. Bennett-Levy J, Perry H. The promise of online cognitive behavioural therapy training for rural and remote mental health professionals. Australas Psychiatry 2009;17 Suppl 1:S121-S124. [doi: 10.1080/10398560902948126] [Medline: 19579124]
9. Khanna MS, Kendall PC. Bringing technology to training: web-based therapist training to promote the development of competent cognitive-behavioral therapists. Cogn Behav Pract 2015;22(3):291-301. [doi: 10.1016/j.cbpra.2015.02.002]
10. Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. Behav Res Ther 2017;88:19-25 [FREE Full text] [doi: 10.1016/j.brat.2016.08.012] [Medline: 28110672]
11. Guri-Rosenblit S, Gros B. E-learning: confusing terminology, research gaps and inherent challenges. IJEDE Internet. 2011 Mar 3. URL: https://www.ijede.ca/index.php/jde/article/view/729 [accessed 2023-10-17]
12. Choudhury S, Pattnaik S. Emerging themes in e-learning: a review from the stakeholders' perspective. Comput Educ 2020;144:103657. [doi: 10.1016/j.compedu.2019.103657]
13. Moore JL, Dickson-Deane C, Galyen K. e-Learning, online learning, and distance learning environments: Are they the same? Internet Higher Educ 2011;14(2):129-135. [doi: 10.1016/j.iheduc.2010.10.001]
14. Guri-Rosenblit S. ‘Distance education’ and ‘e-learning’: not the same thing. High Educ 2005;49(4):467-493. [doi: 10.1007/s10734-004-0040-0]
15. Hew KF, Lo CK. Flipped classroom improves student learning in health professions education: a meta-analysis. BMC Med Educ 2018;18(1):38-41 [FREE Full text] [doi: 10.1186/s12909-018-1144-z] [Medline: 29544495]
16. Telford M, Senior E. Healthcare students' experiences when integrating e-learning and flipped classroom instructional approaches. Br J Nurs 2017;26(11):617-622. [doi: 10.12968/bjon.2017.26.11.617] [Medline: 28594611]
17. Ruggeri K, Farrington C, Brayne C. A global model for effective use and evaluation of e-learning in health. Telemed J E Health 2013;19(4):312-321 [FREE Full text] [doi: 10.1089/tmj.2012.0175] [Medline: 23472702]
18. Rief W. Moving from tradition-based to competence-based psychotherapy. Evid Based Ment Health 2021;24(3):115-120 [FREE Full text] [doi: 10.1136/ebmental-2020-300219] [Medline: 33468517]
19. Hickey C, McAleer S. Competence in psychotherapy: the role of e-learning. Acad Psychiatry 2017;41(1):20-23. [doi: 10.1007/s40596-015-0443-5] [Medline: 26608063]
20. Meinert E, Eerens J, Banks C, Maloney S, Rivers G, Ilic D, et al. Exploring the cost of elearning in health professions education: scoping review. JMIR Med Educ 2021;7(1):e13681 [FREE Full text] [doi: 10.2196/13681] [Medline: 33704073]
21. Means B, Toyama Y, Murphy R, Baki M. The effectiveness of online and blended learning: a meta-analysis of the empirical literature. Teachers College Record 2013;115(3):1-47. [doi: 10.1177/016146811311500307]
22. Yuwono KT, Sujono HD. The effectiveness of e-learning: a meta-analysis. J Phys Conf Ser 2018;1140(1):012024.
23. Fontaine G, Cossette S, Maheu-Cadotte M, Mailhot T, Deschênes MF, Mathieu-Dupuis G, et al. Efficacy of adaptive e-learning for health professionals and students: a systematic review and meta-analysis. BMJ Open 2019;9(8):e025252 [FREE Full text] [doi: 10.1136/bmjopen-2018-025252] [Medline: 31467045]
24. Jackson CB, Quetsch LB, Brabson LA, Herschell AD. Web-Based training methods for behavioral health providers: a systematic review. Adm Policy Ment Health 2018;45(4):587-610 [FREE Full text] [doi: 10.1007/s10488-018-0847-0] [Medline: 29352459]
25. Li K. The evolution of open learning: a review of the transition from pre-e-learning to the era of e-learning. Knowledge Manag E-Learn 2018;10(4):408-425.

XSL•FO
RenderX

26. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151(4):264-269 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
27. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
28. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
29. Smidt A, Balandin S, Sigafoos J, Reed VA. The Kirkpatrick model: a useful tool for evaluating training outcomes. J Intellect Dev Disabil 2009;34(3):266-274. [doi: 10.1080/13668250903093125] [Medline: 19681007]
30. Hedges L. Distribution theory for Glass's estimator of effect size and related estimators. J Educ Stat 1981;6(2):107-128 [FREE Full text] [doi: 10.2307/1164588]
31. Cohen J. Statistical Power Analysis for the Behavioral Sciences, Second Edition. Cambridge, MA: Academic Press; 1989.
32. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
33. Shinohara E, Ohashi Y, Hada A, Usui Y. Effects of 1-day e-learning education on perinatal psychological support skills among midwives and perinatal healthcare workers in Japan: a randomised controlled study. BMC Psychol 2022;10(1):133 [FREE Full text] [doi: 10.1186/s40359-022-00832-6] [Medline: 35606868]
34. Nisar A, Yin J, Nan Y, Luo H, Han D, Yang L, et al. Standardising training of nurses in an evidence-based psychosocial intervention for perinatal depression: randomized trial of electronic vs. face-to-face training in China. Int J Environ Res Public Health 2022;19(7):4094 [FREE Full text] [doi: 10.3390/ijerph19074094] [Medline: 35409775]
35. Taylor D, Dietch J, Pruiksma K, Calhoun C, Milanak M, Wardle-Pinkston S, et al. Developing and testing a web-based provider training for cognitive behavioral therapy of insomnia. Mil Med 2021;186(Suppl 1):230-238. [doi: 10.1093/milmed/usaa359] [Medline: 33499510]
36. Sansen LM, Saupe LB, Steidl A, Fegert JM, Hoffmann U, Neuner F. Development and randomized-controlled evaluation of a web-based training in evidence-based trauma therapy. Prof Psych Res Pract 2020;51(2):115-124. [doi: 10.1037/pro0000262]
37. Sansen LM, Saupe LB, Steidl A, Fegert JM, Hoffmann U, Neuner F. Daring to process the trauma: using a web-based training to reduce psychotherapists' fears and reservations around implementing trauma-focused therapy. Eur J Psychotraumatol 2019;10(1):1696590 [FREE Full text] [doi: 10.1080/20008198.2019.1696590] [Medline: 31853333]
38. Kullberg MJ, Mouthaan J, Schoorl M, de Beurs D, Kenter RMF, Kerkhof AJ. E-Learning to improve suicide prevention practice skills among undergraduate psychology students: randomized controlled trial. JMIR Ment Health 2020;7(1):e14623 [FREE Full text] [doi: 10.2196/14623] [Medline: 32012076]
39. Rahman A, Akhtar P, Hamdani SU, Atif N, Nazir H, Uddin I, et al. Using technology to scale-up training and supervision of community health workers in the psychosocial management of perinatal depression: a non-inferiority, randomized controlled trial. Glob Ment Health (Camb) 2019;6:e8 [FREE Full text] [doi: 10.1017/gmh.2019.7] [Medline: 31157115]
40. Puspitasari AJ, Kanter JW, Busch AM, Leonard R, Dunsiger S, Cahill S, et al. A randomized controlled trial of an online, modular, active learning training program for behavioral activation for depression. J Consult Clin Psychol 2017;85(8):814-825. [doi: 10.1037/ccp0000223] [Medline: 28726481]
41. Stein BD, Celedonia KL, Swartz HA, DeRosier ME, Sorbero MJ, Brindley RA, et al. Implementing a web-based intervention to train community clinicians in an evidence-based psychotherapy: a pilot study. Psychiatr Serv 2015;66(9):988-991 [FREE Full text] [doi: 10.1176/appi.ps.201400318] [Medline: 25930041]
42. Dimeff LA, Harned MS, Woodcock EA, Skutch JM, Koerner K, Linehan MM. Investigating bang for your training buck: a randomized controlled trial comparing three methods of training clinicians in two core strategies of dialectical behavior therapy. Behav Ther 2015;46(3):283-295. [doi: 10.1016/j.beth.2015.01.001] [Medline: 25892165]
43. Harned MS, Dimeff LA, Woodcock EA, Kelly T, Zavertnik J, Contreras I, et al. Exposing clinicians to exposure: a randomized controlled dissemination trial of exposure therapy for anxiety disorders. Behav Ther 2014;45(6):731-744 [FREE Full text] [doi: 10.1016/j.beth.2014.04.005] [Medline: 25311284]
44. Ruzek JI, Rosen RC, Garvert DW, Smith LD, Sears KC, Marceau L, et al. Online self-administered training of PTSD treatment providers in cognitive-behavioral intervention skills: results of a randomized controlled trial. J Trauma Stress 2014;27(6):703-711. [doi: 10.1002/jts.21977] [Medline: 25522731]
45. Larson MJ, Amodeo M, Locastro JS, Muroff J, Smith L, Gerstenberger E. Randomized trial of web-based training to promote counselor use of cognitive behavioral therapy skills in client sessions. Subst Abus 2013;34(2):179-187 [FREE Full text] [doi: 10.1080/08897077.2012.746255] [Medline: 23577913]
46. Bennett‐levy J, Hawkins R, Perry H, Cromarty P, Mills J. Online cognitive behavioural therapy training for therapists: outcomes, acceptability, and impact of support. Aust Psychol 2020;47(3):174-182. [doi: 10.1111/j.1742-9544.2012.00089.x]
47. Dimeff LA, Woodcock EA, Harned MS, Beadnell B. Can dialectical behavior therapy be learned in highly structured learning environments? Results from a randomized controlled dissemination trial. Behav Ther 2011;42(2):263-275. [doi: 10.1016/j.beth.2010.06.004] [Medline: 21496511]

XSL•FO
RenderX

48. Dimeff LA, Koerner K, Woodcock EA, Beadnell B, Brown MZ, Skutch JM, et al. Which training method works best? A randomized controlled trial comparing three methods of training clinicians in dialectical behavior therapy skills. Behav Res Ther 2009;47(11):921-930. [doi: 10.1016/j.brat.2009.07.011] [Medline: 19695562]
49. Weingardt KR, Villafranca SW, Levin C. Technology-based training in cognitive behavioral therapy for substance abuse counselors. Subst Abus 2006;27(3):19-25. [doi: 10.1300/J465v27n03_04] [Medline: 17135177]
50. Wilkerson AK, Wardle-Pinkston S, Dietch JR, Pruiksma KE, Simmons RO, Bunnell BE, et al. Web-based provider training of cognitive behavioral therapy of insomnia: engagement rates, knowledge acquisition, and provider acceptability. Cogn Behav Ther 2022;51(4):343-352. [doi: 10.1080/16506073.2021.1996453] [Medline: 35099360]
51. Seidler ZE, Wilson MJ, Toogood N, Oliffe JL, Kealy D, Ogrodniczuk JS, et al. Pilot evaluation of the men in mind training program for mental health practitioners. Psychology Men Masculinities 2022;23(2):257-264. [doi: 10.1037/men0000383]
52. Soll D, Fuchs R, Mehl S. Teaching cognitive behavior therapy to postgraduate health care professionals in times of Covid 19—an asynchronous blended learning environment proved to be non-inferior to in-person training. Front Psychol 2021;12:657234 [FREE Full text] [doi: 10.3389/fpsyg.2021.657234] [Medline: 34646190]
53. Applebaum A, Walsh LE, Polacek LC, Benvengo S, Levin T. Acute cancer cognitive therapy online training program: feasibility and impact on clinician knowledge uptake. J Cancer Educ 2021;36(5):1081-1085 [FREE Full text] [doi: 10.1007/s13187-020-01738-w] [Medline: 32253724]
54. Gonzalez Salas Duhne P, Horan AJ, Ross C, Webb TL, Hardy GE. Assessing and promoting the use of implementation intentions in clinical practice. Soc Sci Med 2020;265:113490. [doi: 10.1016/j.socscimed.2020.113490] [Medline: 33261903]
55. Graber J. Comparison of mental health nursing student academic achievement and satisfaction: classroom versus online education in teaching therapeutic crisis management techniques. Issues Ment Health Nurs 2019;40(3):247-251. [doi: 10.1080/01612840.2018.1505985] [Medline: 30412433]
56. Cramer RJ, Long MM, Gordon E, Zapf PA. Preliminary effectiveness of an online-mediated competency-based suicide prevention training program. Prof Psychol Res Pract 2019;50(6):395-406. [doi: 10.1037/pro0000261]
57. Meredith P, Yeates H, Greaves A, Taylor M, Slattery M, Charters M, et al. Preparing mental health professionals for new directions in mental health practice: Evaluating the sensory approaches e-learning training package. Int J Ment Health Nurs 2018;27(1):106-115. [doi: 10.1111/inm.12299] [Medline: 28042908]
58. German RE, Adler A, Frankel SA, Stirman SW, Pinedo P, Evans AC, et al. Testing a web-based, trained-peer model to build capacity for evidence-based practices in community mental health systems. Psychiatr Serv 2018;69(3):286-292. [doi: 10.1176/appi.ps.201700029] [Medline: 29137558]
59. Kuhn E, Hugo E. Technology-based blended learning to facilitate psychiatry resident training in prolonged exposure (PE) therapy for PTSD. Acad Psychiatry 2017;41(1):121-124. [doi: 10.1007/s40596-016-0648-2] [Medline: 27921266]
60. Kobak KA, Lipsitz JD, Markowitz JC, Bleiberg KL. Web-based therapist training in interpersonal psychotherapy for depression: pilot study. J Med Internet Res 2017;19(7):e257 [FREE Full text] [doi: 10.2196/jmir.7966] [Medline: 28716769]
61. Fairburn CG, Allen E, Bailey-Straebler S, O'Connor ME, Cooper Z. Scaling up psychological treatments: a countrywide test of the online training of therapists. J Med Internet Res 2017;19(6):e214 [FREE Full text] [doi: 10.2196/jmir.7864] [Medline: 28623184]
62. Curran GM, Woo SM, Hepner KA, Lai WP, Kramer TL, Drummond KL, et al. Training substance use disorder counselors in cognitive behavioral therapy for depression: development and initial exploration of an online training program. J Subst Abuse Treat 2015;58:33-42. [doi: 10.1016/j.jsat.2015.05.008] [Medline: 26219680]
63. Hickey C, McAleer S, Khalili D. E-learning and traditional approaches in psychotherapy education: comparison. Arch Psychiatry Psychother 2015;4:48-52 [FREE Full text] [doi: 10.12740/app/60644]
64. Samuelson KW, Koenig CJ, McCamish N, Choucroun G, Tarasovsky G, Bertenthal D, et al. Web-based PTSD training for primary care providers: a pilot study. Psychol Serv 2014;11(2):153-161. [doi: 10.1037/a0034855] [Medline: 24364595]
65. Kobak KA, Craske MG, Rose RD, Wolitsky-Taylor K. Web-based therapist training on cognitive behavior therapy for anxiety disorders: a pilot study. Psychotherapy (Chic) 2013;50(2):235-247 [FREE Full text] [doi: 10.1037/a0030568] [Medline: 23398031]
66. Martino S, Canning-Ball M, Carroll KM, Rounsaville BJ. A criterion-based stepwise approach for training counselors in motivational interviewing. J Subst Abuse Treat 2011;40(4):357-365 [FREE Full text] [doi: 10.1016/j.jsat.2010.12.004] [Medline: 21316183]
67. Regmi K, Jones L. A systematic review of the factors—enablers and barriers—affecting e-learning in health sciences education. BMC Med Educ 2020;20(1):91 [FREE Full text] [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]

Abbreviations

CBT: cognitive behavioral therapy
DBT: dialectical behavior therapy
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTSD: posttraumatic stress disorder
ROBINS-I: Risk of Bias in Non-randomized Studies of Interventions

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 11.11.23; peer-reviewed by R Philippe, S Payne; comments to author 10.04.24; revised version received 22.04.24; accepted 21.05.24; published 29.07.24.

Please cite as:
Mikkonen K, Helminen EE, Saarni SI, Saarni SE
Learning Outcomes of e-Learning in Psychotherapy Training and Comparison With Conventional Training Methods: Systematic Review
J Med Internet Res 2024;26:e54473
URL: https://www.jmir.org/2024/1/e54473
doi:10.2196/54473
PMID:39073862

©Kasperi Mikkonen, Eeva-Eerika Helminen, Samuli I Saarni, Suoma E Saarni. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

eHealth-Integrated Psychosocial and Physical Interventions for Chronic Pain in Older Adults: Scoping Review

Annalisa De Lucia[1], MSc; Cinzia Perlini[1], PhD; Alessandro Chiarotto[2], PhD; Sara Pachera[3], MSc; Ilenia Pasini[1], MSc; Lidia Del Piccolo[1], PhD; Valeria Donisi[1], PhD

[1]Section of Clinical Psychology, Department of Neuroscience, Biomedicine and Movement Sciences, University of Verona, Verona, Italy

[2]Department of General Practice, Erasmus University Medical Center, Rotterdam, Netherlands

[3]Fondazione Casa di Riposo S Giuseppe Nonprofit Organization of Social Utility, San Martino Buon Albergo (Verona), Italy

Corresponding Author:
Cinzia Perlini, PhD
Section of Clinical Psychology
Department of Neuroscience, Biomedicine and Movement Sciences
University of Verona
Piazzale Ludovico Antonio Scuro 10
Verona, 37134
Italy
Phone: 39 0458124038
Email: cinzia.perlini@univr.it

Abstract

Background: Chronic noncancer pain (CNCP) is highly present among older adults, affecting their physical, psychological, and social functioning. A biopsychosocial multimodal approach to CNCP management is currently extensively suggested by international clinical practice guidelines. Recently, the growing development and application of eHealth within pain management has yielded encouraging results in terms of effectiveness and feasibility; however, its use among the older population remains underexamined.

Objective: The overall aim of this scoping review was to systematically map existing literature about eHealth multimodal interventions (including both physical and psychosocial components) targeting older adults with CNCP.

Methods: This review adhered to the JBI methodology, a protocol was a priori registered as a preprint on the medRxiv platform, and the results were reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Four electronic databases (PubMed, Cochrane Central Register of Controlled Trials, Web of Science, and PsycINFO) were systematically searched for relevant articles. Studies were included if they reported on multimodal interventions (including both physical and psychosocial components) delivered through any eHealth modality to an older population with any type of CNCP. Two reviewers selected the studies: first by screening titles and abstracts and second by screening full-text articles. The quality of the included studies was evaluated using the Quality Assessment Tool for Studies with Diverse Designs. The results of the studies were summarized narratively.

Results: A total of 9 studies (n=6, 67% published between 2021 and 2023) with quality rated as medium to high were included, of which 7 (78%) were randomized controlled trials (n=5, 71% were pilot and feasibility studies). All the included studies evaluated self-management interventions, most of them (n=7, 78%) specifically designed for older adults. The participants were aged between 65 and 75 years on average (mean 68.5, SD 3.5 y) and had been diagnosed with different types of CNCP (eg, osteoarthritis and chronic low back pain). Most of the included studies (5/9, 56%) involved the use of multiple eHealth modalities, with a higher use of web-based programs and video consulting. Only 1 (11%) of the 9 studies involved a virtual reality–based intervention. The evaluated interventions showed signs of effectiveness in the targeted biopsychosocial outcomes, and the participants' engagement and ratings of satisfaction were generally positive. However, several research gaps were identified and discussed.

Conclusions: Overall, of late, there has been a growing interest in the potential that eHealth multimodal interventions offer in terms of improving pain, physical, and psychosocial outcomes in older adults with CNCP. However, existing literature on this topic still seems scarce and highly heterogeneous, with few proper randomized controlled trials, precluding robust conclusions. Several gaps emerged in terms of the older population considered and the lack of evaluation of comorbidities.

International Registered Report Identifier (IRRID): RR2-10.1101/2023.07.27.23293235

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55366) doi:10.2196/55366

KEYWORDS

chronic pain; older adults; eHealth; scoping review; psychological intervention; physical intervention; multimodal intervention; biopsychosocial model for chronic pain; self-management; mobile phone

Introduction

Background

Chronic pain represents a substantial burden to health care systems and society [1]. According to recent estimates, its prevalence is especially high in older adults [2]. In fact, during recent decades, the prolongation of life expectancy has led to a substantial demographic revolution. According to global estimates [3], the number of people living beyond age 60 years will double to 2.1 billion by 2050 worldwide, and it is a trend that is bound to continue. Simultaneously, this process has entailed an increase in the prevalence of health conditions and symptoms usually associated with aging, such as chronic pain [2], which often results in significant physical disability, psychological disorders, cognitive decline, and sleep impairment [4,5]. Chronic pain negatively affects the quality of life of older adults [6], and, conversely, according to older people themselves, the absence of pain contributes to their psychological well-being [7]. Evidence suggests that between 25% and 50% of community-dwelling older adults and approximately 80% of older adults who are institutionalized experience pain, thus reflecting a contemporary clinical issue [8]. Unfortunately, there is a common disbelief, often shared by older adults themselves and even by health care professionals, that chronic pain is a natural and inevitable part of the aging process [9]. This can lead older adults to skip necessary care or to minimize or underestimate the presence of pain, with negative effects in terms of physical and mental health.

The assessment and treatment of pain in the older population are complex and challenging for health care providers because the pain typically occurs in the setting of physiological changes, multiple comorbidities, and polypharmacy, often limiting some therapy options, such as nonsteroidal anti-inflammatory drugs [10,11]. As a consequence, the focus of clinicians and researchers in recent years has been on a biopsychosocial model of chronic pain, according to which the experience of pain is determined by the complex interaction between biological, psychological, and social aspects [12-15]. The adoption of a multimodal approach is even recommended for likely a more effective management of persistent pain in older adults through combinations of both pharmacological and nonpharmacological therapies (such as psychological interventions, exercise therapy, and occupational therapy) [16-20]. To improve the accessibility to effective treatment programs for chronic pain, an important step forward has been made with the digital transformation that, in recent decades, has impacted various areas of society, including health care [21]. The term “eHealth” (or digital health) refers to “the use of information and communications technology in support of health and health-related fields” [22]. As for chronic pain treatment, different solutions have been developed and implemented with encouraging results, such as web-based programs, mobile health (mHealth), virtual reality (VR) systems, and video consulting [23,24].

The use of digital technologies in the field of chronic pain management among older adults is still little explored, especially regarding the potential effects of multimodal interventions (including both physical and psychosocial components) delivered by digital devices. In this regard, it should be noted that older adults, especially those aged ≥85 years, usually show concern and anxiety when using digital technologies due to many barriers to accessing and understanding these tools in light of the existing digital gap compared to the new generations [25-27]. Nevertheless, the adoption of digital technologies by older adults has significantly increased in the last 10 years; for example, the results of a recent survey in the United States showed that 96% of people aged 50 to 64 years and 75% of those aged ≥65 years were using the internet [28].

Given that the topic of eHealth for older adults with chronic pain has not yet been extensively reviewed and in light of the heterogeneous nature of the body of knowledge in this area, a scoping review of the literature is the most appropriate methodology to investigate this emerging field of research.

Objectives

This scoping review aims to systematically map existing literature on eHealth multimodal interventions (including both physical and psychosocial components) designed for older adults with chronic noncancer pain (CNCP) to answer the following research questions (RQs):

- RQ1: What is the body of evidence (eg, in terms of the number and quality of studies)?
- RQ2: What are the gaps in current literature?
- RQ3: What kind of populations have been considered (eg, in terms of specific chronic pain characteristics and underlying conditions, age subgroups, and comorbidities)?
- RQ4: What are the main characteristics of the eHealth multimodal interventions (including physical and psychosocial components) used in older adults with CNCP?
- RQ5: What are the main outcomes and promising results of these interventions?

Methods

Overview

This scoping review was guided by the JBI methodology [29,30] and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines and checklist [31]. The corresponding protocol has been posted on the medRxiv preprint platform [32].

XSL•FO
RenderX

Search Strategy and Screening Procedure

Four electronic databases were searched for relevant articles from their inception up to August 2023: PubMed, Cochrane Central Register of Controlled Trials, Web of Science, and PsycINFO.

A combination of key terms related to the following four main topics was used as a search strategy: (1) eHealth (eg, telemedicine, telehealth, and mHealth), (2) psychosocial and physical (eg, psychotherapy, psychoeducation, physiotherapy, and physical activity) or multimodal (eg, multicomponent or multifactorial and mind-body therapy) interventions, (3) older people, and (4) chronic pain. The full search strategies used for the databases are presented in Multimedia Appendix 1.

An additional manual search of the references was carried out by screening reviews identified through the search strategy. Furthermore, forward citation tracking on the included articles via Google Scholar was performed on November 9, 2023, to broaden the findings and update the search.

All identified articles were exported to Rayyan (Rayyan Systems Inc), a web-based application designed to facilitate the systematic review process [33], and duplicates were removed. Two reviewers (ADL and SP) independently assessed titles and abstracts for eligibility. Full texts of the potentially eligible articles were retrieved and assessed against eligibility criteria by the same 2 independent reviewers. Doubts were discussed, and, where necessary, a third reviewer (VD or CP) was involved.

Inclusion and Exclusion Criteria

Eligibility criteria were defined according to the population, concept, context framework [29] (Textbox 1). As for the age criterion, although some conventional "old age thresholds" exist (eg, 65 years old) [34-36], different age limits or specific age ranges have been used in the literature [37]. Thus, in this scoping review, we included all studies targeting specifically the older population without setting an age limit. However, as this is a scoping review, studies were included even when they had a sample with a mean age of ≥65 years (without any upper limit) of either sex.

Textbox 1. Eligibility criteria based on the population, concept, context framework.

Population

1. Studies aimed at the older population, or the population included in the study had a mean age of ≥65 years, of either sex
2. Presence of chronic noncancer pain, with no limitations in terms of underlying clinical conditions

Concept

1. Multimodal interventions, involving both physical (eg, therapeutic exercise and functional training) and psychosocial (any intervention targeting ≥1 emotional, cognitive, behavioral, or interpersonal aspects) components, delivered (even partially) through any type of eHealth tool and targeting the following main outcomes: pain (eg, severity), emotional functioning (eg, emotional distress, anxiety and depressive symptoms, catastrophizing, kinesiophobia, and perceived self-efficacy), physical functioning (eg, functional mobility and endurance), and integrated outcomes (eg, health-related quality of life, well-being, general functioning, and disability)

Context

1. Any care setting (eg, primary care, outpatient, community, secondary care, or tertiary care)

Despite the lack of a shared definition in the literature, the umbrella term "multimodal" in the context of chronic pain management generally refers to "the combination of multiple therapeutic components, not necessarily provided by different operators" [17]. Specifically, in the context of this review, in accordance with the recent NICE (National Institute for Health and Care Excellence) guideline for chronic pain [19], we included eHealth multimodal interventions that have ≥2 components, including at least 1 physical (nonpharmacological) component and 1 psychosocial component.

Studies were excluded if they (1) were not aimed at the older population or, in the case of a study not explicitly aimed at the older population, the population included in the study had a mean age of <65 years; (2) focused on acute or cancer-related pain; (3) described a single-component intervention (ie, only physical or only psychosocial); and (4) did not use an eHealth modality to deliver (even partially) the intervention.

Journal articles were included if they were written in English or Italian. No limitations with regard to the year of publication were imposed. Both qualitative and quantitative studies were considered eligible for this review. The following types of studies were excluded: systematic reviews, narrative reviews, meta-analyses, bibliometric analyses, letters, case studies, books or book chapters, comments, editorials, congress abstracts or symposia proceedings, poster presentations, and dissertations.

In the cases of different papers presenting the same intervention by the same research groups, the article describing the main results of the eHealth intervention (often the more recent one) was included in the review and considered a reference paper in the Results section, while the previous or secondary papers (referred to as *secondary papers*) were excluded in the flowchart and whenever we reported quantitative information on the number of papers in the text. However, when useful data were reported in these secondary papers, contents were added to the text and in the corresponding tables.

Data Extraction and Synthesis

Data were extracted from the selected articles by 2 independent reviewers (ADL and SP) using a data-charting table in Microsoft Excel that was approved by the research team. Any disagreement was resolved through discussion until a consensus was reached.

XSL•FO
RenderX

The extracted data included the following items: study characteristics (ie, authors, year of publication, and study design); information related to the sampled population (ie, age, sex, type of chronic pain, pain intensity and duration, main comorbidities, and setting); the intervention (ie, its conceptual basis, structure, main components, duration and format, specifying whether it is guided or unguided, and the providers involved); the type of eHealth tool used to deliver the intervention; the targeted outcomes (including older people's experiences and perceptions of eHealth interventions, when applicable), the measurement tools, and the main results; the follow-up duration, when applicable; and, when involved, the type of control group. The extracted information was presented in tabular form along with a narrative summary that is in line with the scoping review's objective.

Assessment of Methodological Quality

Two independent reviewers (ADL and IP) evaluated all eligible studies using the Quality Assessment Tool for Studies with Diverse Designs (QATSDD) [38]. Any potential disagreement was discussed, with a third rater (VD or CP) adjudicating. The tool has shown good reliability and validity [38], and it allows the quality of research papers to be rated on a scale ranging from 0=*not at all* to 3=*complete* across 16 criteria. These criteria apply to studies that adopt different methodologies (ie, qualitative, quantitative, and mixed methods). For qualitative or quantitative research, the highest score is 42; and for mixed methods studies, the maximum score is 48. For each included article, the score attributed to each indicator and the articles' overall quality score (ie, resulting from the sum of individual scores for each item) were provided. Moreover, to identify the items with higher and lower values, each item's mean and SD were calculated in addition to the average quality score for all studies.

Results

Study Selection

The electronic literature search yielded 4823 records, from which 427 (8.85%) duplicates were removed. The remaining 4396 records were screened by title and abstract, and 4371 (99.46%) were excluded based on the inclusion and exclusion criteria. Finally, 25 records were selected for the full-text analysis, of which 16 (64%) were excluded for various reasons (refer to Figure 1 for details, and refer to Multimedia Appendix 2 [39-58] for the list of the full texts excluded and the reasons for exclusion). Thus, 9 articles from the initial database search were included. Of the 16 excluded articles, 6 (38%) were excluded because they focused on the same eHealth intervention described in a previous or secondary paper by the same research group. In addition, 309 papers were identified from the forward citation tracking methodology of included articles and then screened by title and abstract according to the inclusion and exclusion criteria. Of the 5 papers selected for the full-text analysis, 4 (80%) were excluded. Finally, 9 articles were included in this scoping review [59-67] (a preliminary study was excluded from the 9 articles identified from the initial database search and replaced with the study identified from the forward citation tracking).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram showing the identification and selection of studies.

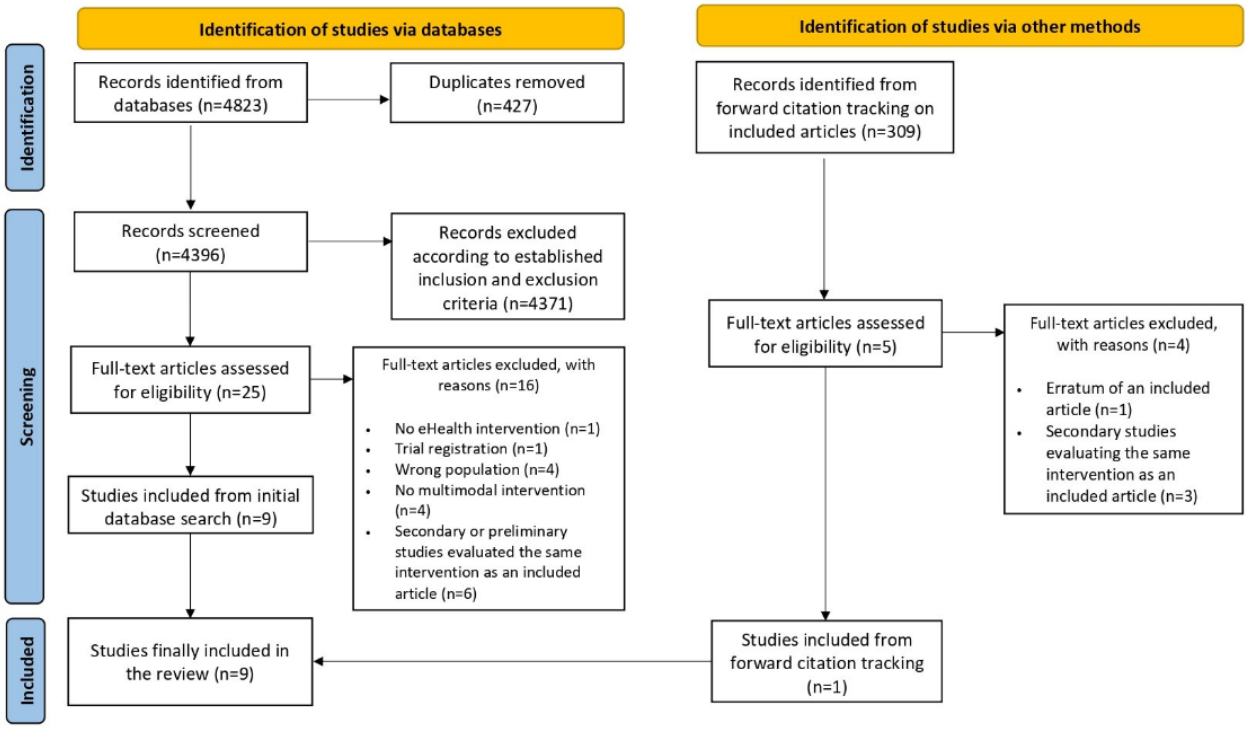

XSL•FO
RenderX

Characteristics of the Included Studies

Of the 9 included studies, 6 (67%) were published between 2021 and 2023 [59,62-66], whereas 1 (11%) was published in 2009 [61], 1 (11%) in 2015 [60], and 1 (11%) in 2016 [67]. Of the 9 studies, 4 (44%) were carried out in the United States [61-64]; 2 (22%) were conducted in Europe, specifically in Germany [65] and the United Kingdom [67]; and 1 (11%) study each was conducted in Canada [66], Australia [59], and Thailand [60]. Of the 9 studies, 7 (78%) were randomized controlled trials (RCTs) [59-65], most of which (n=5, 71%) were pilot and feasibility studies [61-65], while 2 (22%) were mixed methods studies that reported both qualitative and quantitative data [66,67].

Population Characteristics of the Included Studies

Population characteristics of the included studies are presented in Table 1. Of the 9 studies, 7 (78%) evaluated interventions designed and proposed specifically for older adults [60-65,67], while the remaining 2 (22%) described interventions targeting a population with a mean age of ≥65 years [59,66]. In all the included studies, the average age of the population was between 65 and 75 years (Table 1). Where provided, the percentage of female patients in the studies ranged between 54.8% and 92.5%. In most of the included studies (7/9, 78%), participants were recruited from community-based settings (eg, older adult centers, services providers, and religious congregations) [61]; private health insurers [59]; research databases; and, via word of mouth, community locations serving older adults [64]. Conversely, in 2 (22%) of the 9 studies, research participants were outpatients at a general hospital [62] and an orthopedic clinic [66].

XSL•FO
RenderX

Table 1. Population characteristics (sociodemographic and clinical) of the included studies (n=9).

Study	Sample size (n); age (y), mean (SD), sex (%)	Pain: type; intensity[a] (*measure*[b]); duration	Excluded health conditions	Other concurrent health characteristics	Setting
Bennell et al [59]	414; CG[c]: 65.3 (8.7), F[d]=67.2; IG[e]1: 65.4 (8.2), F=54.1; IG2: 64.1 (8.1), F=50.9	Knee OA[f]; moderate (*numeric rating scale*); IG1: median 5.0 (IQR 2.0-10.0) y, IG2: median 5.0 (IQR 3.0-10.0) y, and CG: median 4.0 (IQR 2.0-10.0) y	Conditions where the dietary intervention necessitated medical monitoring (eg, BMI ≥41 kg/m^2 and type 1 or type 2 diabetes requiring medication except metformin, warfarin use); stroke or cardiac event in the past 6 months; unstable heart condition; fluid intake restriction; inflammatory arthritis and knee surgery within the past 6 months	Overweight or obesity as well as cardiac and respiratory issues; bladder or prostate or urologic, psychological, gastrointestinal, type 2 diabetes, and osteoporosis conditions	Community
Saraboon et al [60]	80; IG: 67.3 (6.3), F=92.5; CG: 67.5 (7.3), F=92.5	Knee OA; mild to moderate (*Knee Severity Scale*); NR[g]	Cognitive deficits; history of hip or knee surgery or treatment plans for knee surgery; knee joint injections in the 3 months before enrollment; serious illnesses or conditions	Overweight; comorbid conditions (%): IG: 52 and CG: 55	Community
Berman et al [61]	78; 65.8 (SD NR), F=87.2	A variety of underlying conditions; moderate (*Brief Pain Inventory–Short Form*); NR	NR	NR	Community
Doorley et al [62]	20; IG: 68.9 (5.9), F=66.7; CG: 70.8 (8.4), F=63.6	Any type of musculoskeletal chronic pain; moderate (*numeric rating scale*); >5 y=30%, 6-10 y=25%, and ≥11 y=45%	Medical illness expected to worsen in the next 6 months; serious mental illness; inability to walk without assistance	Cognitive impairment	General hospital
Fanning et al [63]	44; 68.8 (7.9), F=75	Chronic multisite pain; moderate (*PROMIS*[h]*3-itemversion*); NR	Unstable weight; >1 fall in the previous year; impaired vision or cognition; severe medical illness; orthopedic surgery or joint replacement in the previous 6 months or planned to have such a surgery in the coming 6 months	Obesity	Community
Janevic et al [64]	46; 72.1 (7.2), F=89	Chronic musculoskeletal pain; moderate *(PROMIS43-itemversion*); ≥3 mo	Serious acute illness or hospitalization in the last month or planned major surgery in the next 3 months; significant memory difficulties that interfered with daily activities and with participation in the program	Underserved community; comorbid conditions: mean 6.0 (SD 2.2)	Community
Stamm et al [65]	22; IG: 75 (5.8), F=72.7; CG: 75.5 (4.4), F=54.5	Chronic back pain; mild (*numeric rating scale*); IG: mean 15.8 (SD 12.7) y and CG: mean 26.4 (SD 16.6) y	Intervertebral disc surgery in medical history; immobile or mobility possible only with assistance; sensory and motor failure; spinal malignancies, spondylitis, or spondylodiscitis; severe vestibular restrictions affecting the ability to balance	NR	Community
Godziuk et al [66]	53; 65 (7), F=71.7	Knee OA; moderate to severe (*3-point scale*); >5 y=66.1%	BMI <25 kg/m^2; prior joint arthroplasty	Overweight or obesity	Outpatients at orthopedic centers
Pearson et al [67]	83; 67 (SD NR), F=67	OA; NR; <1 y: 6%, 1-5 y: 33%, 5-10 y: 31%, >10 y: 28%, and unknown: 2%	NR	NR	Community

[a]Pain level reported by the authors of the included studies.

[b]In this column, the pain intensity measures used in each study are italicized.

[c]CG: control group.

[d]F: female.

[e]IG: intervention group.

[f]OA: osteoarthritis.

[g]NR: not reported.

[h]PROMIS: Patient-Reported Outcomes Measurement Information System.

XSL•FO
RenderX

With regard to the patients' pain characteristics, the reported information was heterogeneous and varied significantly among the included studies; however, all participants experienced chronic pain with different levels of severity. Of the 9 studies, 4 (44%) included participants with knee osteoarthritis [59,60,66] or with osteoarthritis affecting various body sites (eg, hip and back) [67]. The remaining studies (5/9, 56%) recruited participants with different types of CNCP. More in detail, in the study by Berman et al [61], the most common causes of chronic pain reported by participants were arthritis, spinal stenosis or degenerative disc problems, previous injuries or surgery, and sciatica. Other types of pain included were chronic back pain [65], chronic musculoskeletal pain [62,64], and chronic pain in at least 1 of 3 areas (back, hip, or knees) [63].

Main Characteristics of the Interventions in the Included Studies

The main characteristics of the eHealth interventions in the included studies are presented in Multimedia Appendix 3 [59-67]. Of the 9 interventions, 3 (33%; ie, Active Brains–Fitbit [62], Mobile Intervention to Reduce Pain and Improve Health II [MORPH-II] [63], and Positive Seniors Using Technology to Engage in Pain Self-Management [64]) were investigated and slightly adapted in various studies by the same research group. Herein, we describe the intervention presented in the main study (ie, the one included in the review), even reporting any relevant adjustment to the intervention made in secondary studies (ie, changes reported in italics in Multimedia Appendix 3).

Overall, all included studies assessed multimodal self-management interventions aimed at promoting more effective self-management of pain symptoms and improved health-related quality of life (HRQoL) through health behavior modification techniques and strategies. These programs are diverse in terms of theoretical frameworks, contents, specific outcomes, and targeted populations. Despite the differences, these interventions are generally structured in different modules, characterized by both psychosocial and physical components and activities, the content of which covers to a varying extent the following three areas:

1. Health behavior education about pain physiology, the effects of pain, and treatment options; the benefits of physical activity; nutrition; sleep hygiene; stress management; and healthy lifestyles counseling
2. Self-management cognitive behavioral skills, such as mindfulness practices, relaxation exercises, activity pacing, goal setting, and positive thinking
3. Physical activity programs and muscle-strengthening exercises [60,65-67] that often involve the adoption of wearable activity monitors (eg, Fitbit devices, accelerometers, and actigraph units) [59,62-64]

In addition to this content, some of the studies (3/9, 33%) also provided nutritional and weight loss programs, aimed at promoting a healthy and active lifestyle in older patients with chronic pain and overweight or obesity [60,61,63]. In some of the cases (2/9, 22%), video consulting with dietitians or nutritionists was offered as part of an individualized treatment plan [59,66]. Intervention duration varied across the studies, ranging from 4 weeks [65] to 6 months [59].

The interventions were administered individually in two-thirds of the studies (6/9, 67%) [59,61,64-67], in a group setting in 1 (11%) of the 9 studies [62], and through a mixed format in the remaining studies (2/9, 22%) [60,63]. More than two-thirds of the interventions (7/9, 78%) were delivered fully remotely [59,61-64,66,67]; the remaining interventions (2/9, 22%) adopted a blended modality [60,65].

eHealth Modalities Adopted for the Interventions

Regarding the type of eHealth modalities, those used in the included studies to deliver the intervention have been grouped into 4 categories. Specifically, of the 9 studies, 5 (56%) used web-based modalities (ie, self-guided or therapist-assisted programs to improve knowledge and provide support, care, or treatment to a diverse population with a range of health problems by using websites or web applications [23,68]); 4 adopted mHealth modalities (ie, "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, Personal Digital Assistants [PDAs], and other wireless devices" [69]); 5 (56%) used teleconsulting, such as telephone consulting (ie, "telephone support from health practitioners" [70]) and video consulting (ie, "the utilization of online internet networks to access real-time, high-quality video and audio connections" [71]); and 2 (22%) studies used "other devices," including VR (ie, "a three-dimensional virtual world that may be explored, interacted with, and manipulated by the person" [72,73]; Table 2).

XSL•FO
RenderX

Table 2. eHealth modalities used to deliver the interventions in the included studies.

Study	Modality			
	Web	mHealth[a]	Video or teleconsulting	Other devices
Bennell et al [59]	Website	Wearable activity monitor (Fitbit)	Videoconferencing platform[b]	—[c]
Saraboon et al [60]	—	—	—	Video CDs
Berman et al [61]	Website	—	—	—
Doorley et al [62]	—	Wearable activity monitor (Fitbit; ActiGraph)	Videoconferencing platform[b]	—
Fanning et al [63]	—	Tablet and smartphone app; wearable activity monitor (Fitbit); wireless weight scale	Videoconferencing platform[b]	—
Janevic et al [64]	Web-based videos[b] (website)	Wearable activity monitor	Telephone calls	—
Stamm et al [65]	—	—	—	VR[d] system
Godziuk et al [66]	Emails[b]; web platform	—	Webinar; telephone and video consulting	—
Pearson et al [67]	Website	—	—	—

[a]mHealth: mobile health.

[b]Main eHealth modality.

[c]Not present.

[d]VR: virtual reality.

These eHealth modalities were not mutually exclusive, and most of the studies (5/9, 56%) used a combination of some of these modalities to deliver the intervention (such as a combination of web-based modality and video consulting; mHealth and video consulting; or web-based modality, video consulting, and wearable activity monitor) [59,62-64,66]. In some cases (2/9, 22%), the intervention's core contents (eg, videos and booklets) were sent directly to participants via email, providing additional resources delivered through web platforms, video consulting [66], or telephone sessions [64]. In other cases (3/9, 33%), the main components of the program were administered during web-based sessions through the use of videoconferencing platforms, including the support of supplemental digital tools, such as wearable activity trackers, websites, and smartphone apps [59,62,63]. Of the 4 studies that used only 1 modality, in 2 (50%), participants were asked to access a website to complete the intervention's modules [61,67]; in 1 (25%), the treatment was provided through a VR system in a research setting [65]; and, finally, in 1 (25%) study, the use of video CDs was planned as part of an intervention program delivered during in-person group meetings [60].

Psychosocial, Physical, and Integrated Outcomes Targeted by the eHealth Interventions

Multimedia Appendix 4 [59-67] presents the outcomes considered in the included studies, which were classified into 5 categories: *pain*, *psychological*, *physical*, *integrated* (including perceived biopsychosocial outcomes), and *other outcomes*. For each study, the outcome measures used to assess each variable and the related main significant or not significant results are also provided.

Among the pain outcomes, pain intensity was investigated in all studies. The targeted psychological outcomes were quite heterogeneous among the studies, but, overall, they consisted of emotional distress (eg, anxiety and depressive symptoms) and pain-related psychological variables. As for the former, 3 (33%) of the 9 studies assessed anxiety and depressive symptoms [59,61,62]. The pain-related psychological outcomes considered in the studies were perceived self-efficacy [61,62,64,66], pain catastrophizing [62], pain acceptance [62], kinesiophobia [62,65], illness representation [60], and self-awareness of responses to pain [61]. Other psychological outcomes assessed were cognitive functioning, mindfulness and self-compassion skills [62], coping, resilience [64], and self-determinative (autonomy, competence, and relatedness) needs satisfaction and frustration [63]. The evaluated physical outcomes were physical function and motor skills [59,60,62-64]. In the studies involving participants who were obese or overweight, weight was also monitored [59,60,63]. Among the integrated outcomes, perceived health status and HRQoL [65,66], pain interference [61,63,64], functional capacities [62,65], well-being [66], social participation [64] and social functioning [62], and health behavior related to pain [60] were included. Other variables concerned satisfaction with, and usefulness of, the intervention; acceptability and use of the eHealth tool; engagement; and participation [62-67].

Signs of Effectiveness of the eHealth Interventions

Among the included articles, the RCTs (7/9, 78%) compared the eHealth interventions with other control conditions such as information only, waiting list, conventional therapy, and so on [59-65] (Multimedia Appendix 4). Only the study by Bennell et al [59] included a long-term follow-up assessment. The results are described herein in the following order: first, the RCTs (2/9, 22%); next, the pilot or feasibility RCTs (5/9, 56%); and, finally, the mixed methods studies (2/9, 22%).

In the study by Bennell et al [59], looking at the between-group mean difference, both the exercise and diet+exercise programs

XSL•FO
RenderX

revealed greater increases in pain intensity reduction (diet+exercise: $P<.001$; exercise: $P=.01$) and physical function (diet+exercise: $P<.001$; exercise: $P<.001$) compared with an information control group at 6 months after treatment; the diet+exercise program was also superior to the exercise-only program for these outcomes ($P=.005$). Up to 79.4% of the participants in the diet+exercise group and 58.2% in the information control group achieved the minimal clinically important difference (MCID) in overall knee pain (1.8 units on the numeric rating scale [NRS]; $P<.001$); and 74.6% of the participants in the diet+exercise program, 66.2% in the exercise program, and 34.5% in the control group achieved the MCID in physical function (6 units on the Western Ontario and McMaster Universities Osteoarthritis Index; $P<.001$). Similar results were found at 12-month follow-up; in addition, 78.2% of the participants in the diet+exercise group and 66% in the exercise group achieved the MCID on the NRS ($P=.02$), while 78.1% of the participants in the diet+exercise group and 63.8% in the exercise group achieved the MCID on the Western Ontario and McMaster Universities Osteoarthritis Index ($P=.045$).

In the study by Saraboon et al [60], the multifactorial intervention program showed a statistically significant increase in osteoarthritis knowledge, illness representation, health behavior, movement ability, and joint range of motion, as well as a reduction in knee pain and body weight ($P<.001$). All these variables were found to have significant differences between the intervention and control (information only) groups ($P<.001$).

Looking at the pilot or feasibility RCTs [61-65], Berman et al [61] found statistically significant improvements in pain intensity (worst pain: $P=.01$; least pain: $P=.05$ for the intervention group and $P=.01$ for the control group; average pain: $P=.01$ for the intervention group and $P=.05$ for the control group; average intensity: $P=.01$ for the intervention group and $P<.001$ for the control group) and pain interference ($P=.01$) in both the web-based intervention and waiting list control groups as well as increases in "confidence with using nonmedical self-care techniques to manage pain" ($P<.001$) in the intervention group.

In the study by Doorley et al [62], preliminary signs of improvement for the Active Brains–Fitbit web-based intervention group were found in pain intensity (pain intensity with activity: $P=.05$; Cohen $d=1.0$) as well as cognitive ($P=.05$; Cohen $d=0.8$) and physical (average steps: $P=.04$; Cohen $d=0.4$; physical activity intensity: $P=.04$; Cohen $d=1.2$) functioning. Conversely, the health enhancement program control group showed a significant decrease in physical functioning (average steps: $P=.02$; Cohen $d=0.3$).

Consistent with promising results from a series of earlier trials [52,54], Fanning et al [63] observed a large effect on competence satisfaction ($P<.01$; $\eta^2=0.22$), average daily steps ($P=.02$; $\eta^2=0.23$), and postural shifts ($P=.02$; $\eta^2=0.24$) in the intervention group but not in the control group.

Janevic et al [64] found greater significant improvements in the web-based intervention group versus the waiting list control group in measures of pain self-efficacy ($P=.007$; $\eta_p^2=0.155$) and pain interference ($P<.001$; $\eta_p^2=0.166$). Moreover, significantly more participants in the intervention group compared with participants in the control group reported "better" or "much better" global functioning ($P<.001$) and pain ($P=.003$) after treatment. The minimally important difference of 2.5 points on the PROMIS (Patient-Reported Outcomes Measurement Information System) pain interference subscale was achieved by 53% of the intervention group participants and 17% of the control group participants ($P=.02$).

In the study by Stamm et al [65], only the VR intervention was associated with a significant increase in functional capacities ($P=.03$; $r=0.67$), and only the conventional control condition showed a significant increase in general mental health ($P=.01$; $r=0.81$). Although not statistically significant, both intervention and control groups showed a reduction in pain intensity after treatment ($P=.50$ and $P=.07$, respectively). However, considering that, on average, a reduction of approximately 30% on the NRS represents a clinically important difference for chronic low back pain, the intervention group showed a reduction in pain intensity of only 18.02% on the NRS, whereas the control group achieved a reduction of 43.64%.

Finally, although the study by Godziuk et al [66] presents specific methodological limitations (ie, it is a single-arm study), the multimodal intervention in this study was associated with an increase in perceived self-efficacy for managing symptoms ($P=.003$) and daily activities ($P<.001$), HRQoL components of pain ($P=.02$) and physical functioning ($P<.001$), and an understanding of arthritis symptoms ($P<.001$), as well as with a decrease in interest in total knee arthroplasty ($P<.001$).

Feasibility and Acceptability of the eHealth Interventions

Feasibility, acceptability, or participants' satisfaction with the intervention was evaluated in 8 (89%) of the 9 included studies [59,61-67] (Multimedia Appendix 4).

In the study by Bennell et al [59], participants in the exercise group and those in the diet+exercise group (92.7% and 97%, respectively) were more satisfied with their care compared to the participants in the control group (16.7%) at 6 months after treatment ($P<.001$), with similar results even at 12 months.

In the study by Berman et al [61], most of the participants found the web-based intervention to be highly useful (81.4%) and user-friendly (88.4%).

In the study by Doorley et al [62], the Active Brains–Fitbit group met the criteria for "excellent" on all benchmarks set a priori by the authors (eg, feasibility of recruitment, program acceptability, and credibility), except for treatment expectancy, and was rated both feasible and satisfactory in the qualitative results.

In the study by Fanning et al [63], both qualitative and quantitative data supported the feasibility and acceptability of the MORPH-II program. More in detail, support for the feasibility of the program included a high attendance rate for the remote sessions (82.5% on average) and a high retention rate for follow-up testing (90.9%). Technological aspects of the MORPH-II program were evaluated as acceptable and usable

XSL•FO
RenderX

(with an average score on the System Usability Scale of 77, which is categorized as "good to excellent," and a median score of 85, which is classified as "best imaginable" usability). Qualitative interviews revealed overall positive feedback on the program, which was described as "beneficial" and "life changing."

In the study by Janevic et al [64], participants provided overall positive feedback about the Positive Seniors Using Technology to Engage in Pain Self-Management intervention, with 95% of the participants strongly agreeing or agreeing that they increased their understanding of pain management and that the program helped them reach their pain management goals. Moreover, the retention rate was 90%, while, in terms of engagement, the results showed a mean of 5.7 completed sessions (out of 7), with 95% of the participants reported to have watched all program videos.

In the study by Stamm et al [65], users of the VR system experienced a "higher degree of immersion" (average score of 19.09 points on the Technology Usage Inventory) and rated the VR program as at least above average on most of the evaluation criteria (eg, attractiveness, perspicuity, efficiency, dependability, and stimulation) and as mostly good and excellent in attractiveness and perspicuity.

In the study by Godziuk et al [66], the intervention's resources—especially email content—were positively perceived as appropriate and tailored for patients with knee osteoarthritis. However, concerns related to privacy and entering personal information on web platforms emerged. Finally, the mixed methods study by Pearson et al [67] aimed at co-designing a web-based version of a multimodal intervention for chronic pain relying on qualitative data from surveys, focus groups, and semistructured interviews. Briefly, according to these data, older people seemed more likely to accept a web-based program as a supplement to, rather than as a replacement for, a face-to-face intervention.

Assessment of the Methodological Quality of the Included Studies

Overall, the QATSDD score ranged between 71.4% (mean raw score=30) [60] and 92.9% (mean raw score=39) [62] (Multimedia Appendix 5 [59-67]). The average quality score for all studies was 82.1% (raw score=34.5). Variations in quality among the studies mainly concerned the following items: evidence of sample size considered in terms of analysis, the presence of a statistical assessment of the reliability and validity of measurement tools, and a critical discussion of the strengths and limitations. The lowest QATSDD single-item score related to user involvement in the study design, which was considered only in 5 (56%) of the 9 studies (item score: mean 1.4, SD 1.4) [62-65,67]. In addition, the statistical assessment of the reliability and validity of measurement tools was limited in most of the studies (6/9, 67%; item score: mean 1.6, SD 1.1).

Discussion

Principal Findings

This scoping review systematically mapped existing literature regarding eHealth multimodal interventions (including both physical and psychosocial components) in older adults with CNCP. The population involved, potential gaps in the literature, and the main interventions' characteristics, as well as the results in terms of the signs of effectiveness, feasibility, and acceptability, were explored and described. Nine studies, mainly of recent publication (6/9, 67% were published between 2021 and 2023), were included in this scoping review. Although the implementation of digital solutions is increasingly occurring in several health care settings [74], the use of such technologies specifically targeting multimodal interventions for CNCP management in the older population still seems to be relatively rare and very recent.

The dearth of literature on eHealth solutions for CNCP dedicated to the older population might be explained by several factors. First, the negative stereotypes that portray older adults as a homogeneous group mostly characterized by vulnerability, inactivity, and cognitive decline may deter their inclusion in research, especially when it concerns digital technologies [75]. Second, the fact that the older population is overall less digitally literate than younger cohorts [76] might have discouraged researchers and investors from financing the research and development of digital health solutions specifically targeted at older people [77]. It is also noteworthy that, in the last decade, we have been facing a technological revolution with the rapid growth of digital health solutions which, on the one hand, brings major benefits and positive changes for society and the health care system but, on the other hand, poses a great challenge in terms of economic, ethical, and regulatory issues [78]. In particular, the recent development of therapeutic options being classified as a medical device (MD; ie, "any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more...specific medical purposes" [79]) or a digital therapeutic (DT; ie, "health software intended to treat or alleviate a disease, disorder, condition, or injury by generating and delivering a medical intervention that has a demonstrable positive therapeutic impact on a patient's health" [80]) has been forcing countries to adapt their regulatory systems in terms of definitions, terminology, requested evidence, and payment and reimbursement rules, with tremendous differences among countries [81,82]. Although some solutions have been recently introduced for different pain conditions [83,84], no MD or DT specifically developed for the management of chronic pain in the older population exists, to the best of our knowledge. Despite this, although not specific for the target population of this scoping review, such a complex scenario, together with the lengthy development process for these new devices, may have impacted the ease of conducting studies in this field, especially when added to the unique challenges linked to the older population (eg, negative stereotypes and digital literacy).

XSL•FO
RenderX

The quality of the studies included in this scoping review, in general, was rated as medium to high according to the QATSDD tool. The main strengths were the clear description of the research aims or objectives, setting, procedure for data collection, recruitment data, and the fit between the stated RQ and the methods of data collection and analysis selected. The main limitation of the studies was the very limited statistical assessment of the reliability and validity of the measurement tools. Moreover, the user involvement in the research or intervention design was somewhat considered only in just over half of the included studies (5/9, 55%). Specifically, in those studies meeting this criterion, qualitative methods (eg, surveys, focus groups, and semistructured interviews) were used as a first step to gain a better understanding of older adults' preferences and needs and to consequently develop tailored programs. This is in line with the relevance of using participatory approaches when developing interventions for chronicity [85] and with the recommendation to adopt a person-centered approach by including end users in the design process of the intervention from the early stages to improve feasibility, acceptability, and satisfaction with the innovative digital technology [86].

A few issues regarding the characteristics of the populations covered by the studies should be taken into account. First, although our review was intentionally broad in considering the older population, with no age limits established, it is interesting to note that the average age of the participants included in the studies was between 65 and 75 years [87]. Therefore, other subgroups, including those aged 75 to 84 years and especially those aged ≥85 years, seem poorly represented in this field, suggesting an area of further research. Indeed, such cohorts may vary deeply in terms of physical and psychosocial (eg, needs, goals, and expectations) characteristics [88] as well as age-related differences with respect to the determinants of psychological well-being [89]; consequently, this potential uniqueness needs to be taken into account when designing tailored interventions. This becomes even more important when considering that the number of digitally literate people aged ≥75 years is expected to significantly increase in the future as the world's population ages, and technology continues to advance. Therefore, it would be worthwhile to further focus on the diversity within the older population in this field, broadening the scope of interest of research dedicated to the older population to include people aged ≥75 years, who are more likely to be excluded from the evaluation of new technologies intended to meet the needs of the older population [7,90].

Moreover, in the included studies targeting the older population, the lower age limit varied substantially. Indeed, although the eHealth intervention was specifically designed for older adults, some of the studies considered populations with a minimum age of 50 years [60,67] or 55 years [61,63]. This is coherent with the lack of an established chronological age to define the older population and the high variability in the lower age limit, even after accounting for the differences among countries [37]; however, this poses a challenge for comparability among studies and underscores the need for studies on the older population to provide a more detailed description of results based on age distribution.

Second, the samples were composed primarily of female individuals, although this is consistent with CNCP epidemiology [91]. Third, the included studies were highly heterogeneous with respect to the types of CNCP included and also when considering the comorbid conditions. As regards the distribution of underlying conditions, 44% (4/9) of studies targeted participants with a specific diagnosis of osteoarthritis (ie, mainly of the knee and to a lesser extent of other body sites); in the remaining studies (5/9, 56%), different types of CNCP (eg, chronic musculoskeletal pain and chronic low back pain) were considered. On average, participants with mild to moderate levels of pain were recruited, except for a study in which approximately 70% of the participants self-identified their pain as severe [66]. All studies necessarily used self-reported methods to evaluate pain, pain being a subjective experience. Although it is necessary to respect patient pain reports, it has to be considered that complex factors influence how people report and interpret numerical pain ratings (eg, due to the "peak-end phenomenon" [92]), thus potentially introducing bias into the pain assessment, which should be accurately considered when interpreting the results [93,94].

With regard to comorbidities, it is particularly interesting to note that the chronic pain and cognitive decline comorbidity was addressed by only 1 (11%) of the 9 included studies [62]. This is surprising, considering that cognitive impairment has been found to be significantly and bidirectionally associated with CNCP in older adults [95] and that both of these health conditions show a high prevalence among this age group [95]. Moreover, considering that the world's population continues to age, it can be assumed that an increasing number of people will experience such conditions in the future [95]. As for the study by Doorley et al [62], the authors developed and adapted a mind-body activity program to the unique needs of older people with both chronic pain and cognitive impairment; for example, they introduced cognitive functioning skills aimed at developing compensatory strategies for cognitive deficits and promoting intellectual stimulation. Overall, the aforementioned factors, as well as the large heterogeneity of research methodologies among the included studies, make it challenging to generalize the findings to the entire older population with CNCP. Finally, it is noteworthy to observe that none of the included studies carried out any assessment of frailty, although this state of vulnerability increases steadily with aging, and it is significantly associated with pain conditions [96].

To conclude, the aforementioned factors (eg, common comorbidities, the presence of frailty, and possible needs related to decreased digital literacy) should be considered when tailoring interventions for the older population to encourage the participation of older people with different characteristics.

Regarding the evaluated eHealth interventions, these programs differ in terms of theoretical basis, contents, specific outcomes, and targeted populations, but they are inspired by a common biopsychosocial approach consistent with the current multidimensional conception of CNCP management [17,19,97]. This common framework finds expression through the implementation of both physical and psychological activities, broadly aimed at supporting more effective chronic pain self-management skills and ultimately at promoting an increased

XSL•FO
RenderX

well-being and a better quality of life. In Figure 2, we present a graphical representation to illustrate the biopsychosocial components and strategies considered in the eHealth multimodal interventions included in this scoping review. The recent and increasingly widespread integrated pain team models promote the adoption of both pharmacological and nonpharmacological treatment options within patient-centered care settings [98], highlighting the importance of the provider-patient relationship, where the patients should be encouraged and empowered to effectively manage their health care. Regarding this, in 5 (56%) of the 9 studies included in this scoping review, a training and orientation phase on digital health tools was provided [61-64,66]. Moreover, in 7 (78%) of the 9 studies, at least 1 health care professional was available to offer video consultations, supervision, prompts, and support for the entire duration of the intervention, especially with respect to any technology-related questions. These aspects are in line with recent evidence, according to which older adults' engagement with eHealth interventions might be optimized by providing social support to overcome potential technical barriers [99]; thus, considering these aspects might improve the practical application of eHealth solutions in the clinical context. Another factor that seems to contribute to older people's commitment to a digital health intervention is the support from family members or caregivers, especially in assisting them during their initial exposure to these technologies [99]. However, only 1 (11%) of the 9 included studies considered the active involvement of relatives in supporting older adults to navigate technology-related challenges [62].

Figure 2. Graphical illustration of the main biopsychosocial components and strategies that emerged in the included eHealth interventions.

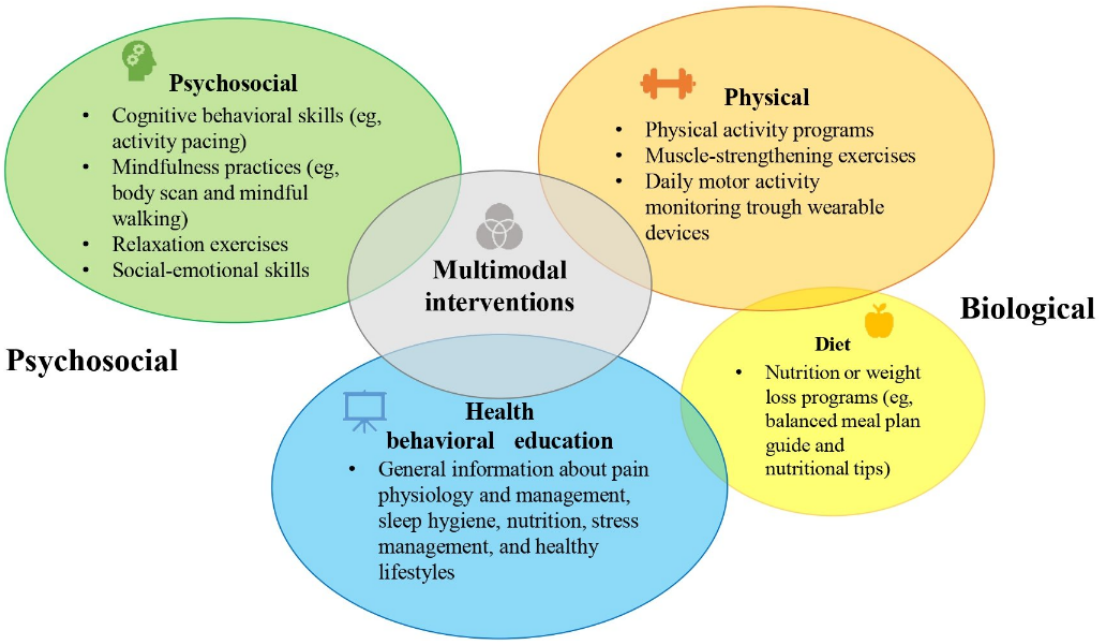

As for the interventions' delivery formats, they ranged from fully remote programs (ie, completely based on an eHealth modality) to programs in which the digital health tool was a component of broader in-person treatment (eg, the use of VR or video CDs as part of in-person meetings). More in detail, the most common eHealth modality types used for the interventions were web-based programs (eg, websites and web platforms) and video consulting, followed by VR, which was applied in only 1 (11%) of the 9 studies. With respect to this last point, it should be noted that there is a lack of literature investigating VR-based multimodal interventions in the specific context of chronic pain and especially in the older population [100]. However, most of the included studies (5/9, 56%) involved multiple modalities (eg, by providing additional digital devices, such as wearable activity monitors, or smartphone apps). Some recent evidence suggests that older adults may be more prone to engage in eHealth interventions as long as these modalities do not completely replace face-to-face usual care [99]. Considering these data, video consulting may be a particularly acceptable digital health alternative for older people to be used in clinical practice because it allows for direct and synchronous contact with health care providers. Although the studies included in this scoping review did not consider these aspects, it would be interesting to consider the attitudes of health care professionals with respect to integrating some of these digital tools with usual treatments in clinical practice. In this regard, there is evidence of how the negative and ageist attitudes of health care providers may influence how the older adults are perceived in relation to the use of digital tools and consequently impact the actual adoption of these tools in clinical practice [75].

Overall, the evaluated interventions showed signs of effectiveness in the targeted pain, psychological, physical, and integrated outcomes. However, on the one hand, the limits of the methodological designs of the studies and, on the other hand, the heterogeneity of the outcomes, measures used to evaluate the outcomes, populations, and interventions considered in the studies hinder us from drawing conclusive results. In particular, with regard to the design of the 9 studies, it should be noted that only 2 (22%) were proper RCTs, whereas 5 (56%) were pilot and feasibility RCTs. In particular, the results of the 2 RCTs, both regarding chronic knee osteoarthritis pain (with overweight or obesity), suggested that multimodal education

XSL•FO
RenderX

and exercise programs lead to improvement in diverse outcomes, including pain [59,60]. However, in the study by Bennell et al [59], the eHealth programs were completely delivered in a telehealth format, while Saraboon et al [60] only partially introduced the use of remote digital tools.

In line with these results, preliminary signs of effectiveness emerged from the RCTs; for example, summarizing the results for pain outcomes, a web-based intervention demonstrated positive effects on at least 1 pain measure in the studies by Doorley et al [62] and Janevic et al [64]; in addition, an improvement in confidence in managing pain was observed in the study by Berman et al [61]. In the study by Fanning et al [63], although there was no improvement in pain, the MORPH-II intervention led to increased perceived competence and physical activity. However, due to the limits intrinsic to their nature, these pilot and feasibility studies do not provide conclusive evidence on the effectiveness of the provided eHealth interventions, but they might serve as a guide in the implementation of future full-scale effectiveness studies (as mentioned in 3/9, 33% of the included studies [62-64]).

By contrast, these studies, together with the mixed methods studies, might be considered useful to examine the feasibility of the studied approach. As regards the feasibility of the interventions, when evaluated, the participants' engagement, satisfaction, and feedback were generally positive, thus supporting the potential feasibility and acceptability of eHealth multimodal programs in older adults with CNCP.

Finally, only 3 (33%) of the 9 studies [59,64,65] considered the clinical relevance of the interventions, using diverse cutoffs for the MCID in pain (and in n=1, 33%, even function), and in 2 (67%) of these 3 studies, suggestions for improvement emerged [59,64]. In general, although the evaluation of clinical relevance is central for clinical guideline development and the interpretation of results, the introduction of the MCID in chronic pain research seems still underused, and more attention should be given to its appropriate methodological use [101]. As regards the use of VR for older adults with chronic back pain, it is interesting to note that only the control group achieved clinical relevance on this score in the study by Stamm et al [65], with interesting insights provided by the authors on the potential limits of this eHealth tool, which should be better considered to improve the effective application of VR in this context.

To sum up, not only are studies targeting eHealth interventions for chronic pain among the older population scarce, but the number of confirmatory RCTs is also especially limited, suggesting gaps to be covered in future research.

Strengths and Limitations

This scoping review followed a systematic and structured theoretical framework to map the existing evidence regarding eHealth multimodal interventions for CNCP in older adults, in line with the JBI methodology [29,30]. In addition, it was reported in line with the PRISMA-ScR guidelines [31]. The review highlighted several gaps in this research field, suggesting areas that require more in-depth attention with regard to the implementation of eHealth multimodal interventions for CNCP in older adults. Finally, it provided a systematic synthesis of findings from both technical and clinical perspectives through the collaboration of a multidisciplinary team of researchers and clinicians.

However, this review has some limitations. First, because gray literature was not searched, any unpublished or ongoing studies are likely to have been missed. Although, to the best of our knowledge, MDs or DTs targeting chronic pain in the older population do not exist, other eHealth tools may have been developed by companies and investors but not tested in published studies and therefore not found within scientific articles. Second, only articles published in English or Italian were selected. In this regard, it should be noted that there is some evidence showing that excluding non-English literature from evidence syntheses does not alter conclusions [102]. Third, the adoption of selective criteria, such as the focus on the older population (refer to the Principal Findings section) and on multimodal interventions with regard to the inclusion of the studies, may have contributed to the paucity of studies identified. Regarding this last aspect, considering the relevance of adopting a multimodal approach to achieve optimal results in CNCP management, our aim was to explore the combined use of physical and psychosocial components through eHealth interventions. However, it is important to note that physical and psychosocial interventions may be proposed as stand-alone interventions and that our review did not include studies discussing only physical or only psychosocial interventions, which might be considered in a future exploration of the literature, even discussing how they might be integrated with each other and comparing the benefits of a single-component intervention versus a multimodal intervention.

Conclusions

To conclude, growing interest has been shown in the potential that eHealth multimodal interventions offer in terms of improving CNCP management in older adults, although there is a dearth of studies in this field. The studies included in this scoping review mainly involved people aged 65 to 75 years, while those aged ≥75 years seem to be vastly underrepresented in this field. Similarly, only a few studies (6/9, 67%) considered the presence of comorbidities, particularly cognitive decline, and no study conducted an assessment of frailty. Moreover, existing literature seems to be limited by the small number of RCTs evaluating eHealth interventions for chronic pain among the older population and highly heterogeneous in terms of study designs, contents, measurement instruments, outcomes, and targeted populations, heavily limiting the possibility to draw robust conclusions. Nonetheless, our review indicates the potential of eHealth multimodal interventions in improving several pain-related outcomes in the older population. As this is an emerging and evolving field of research, several gaps and unresolved research issues remain that need to be fully and deeply addressed by future research.

XSL•FO
RenderX

Acknowledgments

This paper was produced during ADL's attendance at the PhD program (Cycle XXXVIII) in neuroscience, psychological and psychiatric sciences, and movement sciences at the University of Verona, Italy, with the support of a scholarship financed by ministerial decree 351 of April 9, 2022, based on the PNRR (National Recovery and Resilience Plan), funded by the European Union (NextGenerationEU program), Mission 4 ("Education and Research"), component 1 ("Enhancement of the offer of educational services: from nurseries to universities") as part of the following investments: investment 3.4 ("Teaching and advanced academic skills"; project M4C1: CUP B31I22000190006) and investment 4.1 ("Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage"; project M4C1: CUP 31I22000200006).

Authors' Contributions

VD and CP conceptualized the study. ADL, VD, CP, AC, IP, and SP curated the data. ADL, VD, CP, and AC prepared the original draft. LDP, IP, and SP reviewed and edited the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File (Adobe PDF File), 89 KB - jmir_v26i1e55366_app1.pdf]

Multimedia Appendix 2
Full texts excluded and reasons for exclusion.
[PDF File (Adobe PDF File), 95 KB - jmir_v26i1e55366_app2.pdf]

Multimedia Appendix 3
Main characteristics of the included eHealth interventions.
[PDF File (Adobe PDF File), 135 KB - jmir_v26i1e55366_app3.pdf]

Multimedia Appendix 4
Targeted outcomes, measures, and main results of the included studies.
[PDF File (Adobe PDF File), 225 KB - jmir_v26i1e55366_app4.pdf]

Multimedia Appendix 5
Assessment of methodological quality.
[PDF File (Adobe PDF File), 119 KB - jmir_v26i1e55366_app5.pdf]

Multimedia Appendix 6
PRISMA-ScR-Checklist.
[PDF File (Adobe PDF File), 270 KB - jmir_v26i1e55366_app6.pdf]

References

1. Dieleman JL, Cao J, Chapin A, Chen C, Li Z, Liu A, et al. US health care spending by payer and health condition, 1996-2016. JAMA 2020 Mar 03;323(9):863-884 [FREE Full text] [doi: 10.1001/jama.2020.0734] [Medline: 32125402]
2. GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990-2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Jun;5(6):e316-e329 [FREE Full text] [doi: 10.1016/S2665-9913(23)00098-X] [Medline: 37273833]
3. World population prospects: the 2017 revision, key findings and advance tables. United Nations, Department of Economic and Social Affairs, Population Division. 2017. URL: https://population.un.org/wpp/publications/files/wpp2017_keyfindings.pdf [accessed 2024-04-29]
4. Domenichiello AF, Ramsden CE. The silent epidemic of chronic pain in older adults. Prog Neuropsychopharmacol Biol Psychiatry 2019 Jul 13;93:284-290 [FREE Full text] [doi: 10.1016/j.pnpbp.2019.04.006] [Medline: 31004724]
5. Dagnino AP, Campos MM. Chronic pain in the elderly: mechanisms and perspectives. Front Hum Neurosci 2022;16:736688 [FREE Full text] [doi: 10.3389/fnhum.2022.736688] [Medline: 35308613]
6. Ferretti F, Castanha AC, Padoan ER, Lutinski J, da Silva MR. Quality of life in the elderly with and without chronic pain. Br J Pain 2018;1(2):111-115 [FREE Full text] [doi: 10.5935/2595-0118.20180022]

XSL•FO
RenderX

7. Lara E, Martín-María N, Forsman AK, Cresswell-Smith J, Donisi V, Ådnanes M, et al. Understanding the multi-dimensional mental well-being in late life: evidence from the perspective of the oldest old population. J Happiness Stud 2019 Feb 23;21(2):465-484. [doi: 10.1007/s10902-019-00090-1]
8. Cravello L, Di Santo S, Varrassi G, Benincasa D, Marchettini P, de Tommaso M, et al. Chronic pain in the elderly with cognitive decline: a narrative review. Pain Ther 2019 Jun;8(1):53-65 [FREE Full text] [doi: 10.1007/s40122-019-0111-7] [Medline: 30666612]
9. Thielke S, Sale J, Reid MC. Aging: are these 4 pain myths complicating care? J Fam Pract 2012 Nov;61(11):666-670 [FREE Full text] [Medline: 23256096]
10. Reid MC, Eccleston C, Pillemer K. Management of chronic pain in older adults. BMJ 2015 Feb 13;350:h532 [FREE Full text] [doi: 10.1136/bmj.h532] [Medline: 25680884]
11. Niknejad B, Bolier R, Henderson CR, Delgado D, Kozlov E, Löckenhoff CE, et al. Association between psychological interventions and chronic pain outcomes in older adults: a systematic review and meta-analysis. JAMA Intern Med 2018 Jun 01;178(6):830-839 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0756] [Medline: 29801109]
12. Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJ, Ostelo RW, Guzman J, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: Cochran systematic review and meta-analysis. BMJ 2015 Feb 18;350:h444 [FREE Full text] [doi: 10.1136/bmj.h444] [Medline: 25694111]
13. Perlini C, Donisi V, Del Piccolo L. From research to clinical practice: a systematic review of the implementation of psychological interventions for chronic headache in adults. BMC Health Serv Res 2020 May 25;20(1):459 [FREE Full text] [doi: 10.1186/s12913-020-05172-y] [Medline: 32450871]
14. Miaskowski C, Blyth F, Nicosia F, Haan M, Keefe F, Smith A, et al. A biopsychosocial model of chronic pain for older adults. Pain Med 2020 Sep 01;21(9):1793-1805. [doi: 10.1093/pm/pnz329] [Medline: 31846035]
15. Gandolfi M, Donisi V, Battista S, Picelli A, Valè N, Del Piccolo L, et al. Health-related quality of life and psychological features in post-stroke patients with chronic pain: a cross-sectional study in the neuro-rehabilitation context of care. Int J Environ Res Public Health 2021 Mar 17;18(6):3089 [FREE Full text] [doi: 10.3390/ijerph18063089] [Medline: 33802829]
16. Cheng ST, Chan KL, Lam RW, Mok MH, Chen PP, Chow YF, et al. A multicomponent intervention for the management of chronic pain in older adults: study protocol for a randomized controlled trial. Trials 2017 Nov 09;18(1):528 [FREE Full text] [doi: 10.1186/s13063-017-2270-3] [Medline: 29121961]
17. Giusti EM, Castelnuovo G, Molinari E. Differences in multidisciplinary and interdisciplinary treatment programs for fibromyalgia: a mapping review. Pain Res Manag 2017;2017:7261468 [FREE Full text] [doi: 10.1155/2017/7261468] [Medline: 28620267]
18. Schwan J, Sclafani J, Tawfik VL. Chronic pain management in the elderly. Anesthesiol Clin 2019 Sep;37(3):547-560 [FREE Full text] [doi: 10.1016/j.anclin.2019.04.012] [Medline: 31337484]
19. Chronic pain (primary and secondary) in over 16s: assessment of all chronic pain and management of chronic primary pain. National Institutes for Health and Care Excellence. 2021. URL: https://www.nice.org.uk/guidance/ng193/resources/chronic-pain-primary-and-secondary-in-over-16s-assessment-of-all-chronic-pain-and-management-of-chronic-primary-pain-pdf-66142080468421 [accessed 2024-03-07]
20. Kechichian A, Lafrance S, Matifat E, Dubé F, Lussier D, Benhaim P, et al. Multimodal interventions including rehabilitation exercise for older adults with chronic musculoskeletal pain: a systematic review and meta-analyses of randomized controlled trials. J Geriatr Phys Ther 2022;45(1):34-49. [doi: 10.1519/JPT.0000000000000279] [Medline: 34928239]
21. van de Vijver S, Tensen P, Asiki G, Requena-Méndez A, Heidenrijk M, Stronks K, et al. Digital health for all: how digital health could reduce inequality and increase universal health coverage. Digit Health 2023;9:20552076231185434 [FREE Full text] [doi: 10.1177/20552076231185434] [Medline: 37434727]
22. Recommendations on digital interventions for health system strengthening – executive summary. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-RHR-19.8 [accessed 2024-04-29]
23. Slattery BW, Haugh S, O'Connor L, Francis K, Dwyer CP, O'Higgins S, et al. An evaluation of the effectiveness of the modalities used to deliver electronic health interventions for chronic pain: systematic review with network meta-analysis. J Med Internet Res 2019 Jul 17;21(7):e11086 [FREE Full text] [doi: 10.2196/11086] [Medline: 31317869]
24. Donisi V, De Lucia A, Pasini I, Gandolfi M, Schweiger V, Del Piccolo L, et al. e-Health interventions targeting pain-related psychological variables in fibromyalgia: a systematic review. Healthcare (Basel) 2023 Jun 25;11(13):1845 [FREE Full text] [doi: 10.3390/healthcare11131845] [Medline: 37444679]
25. Seifert A. The digital exclusion of older adults during the COVID-19 pandemic. J Gerontol Soc Work 2020;63(6-7):674-676. [doi: 10.1080/01634372.2020.1764687] [Medline: 32401181]
26. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
27. O'Reilly PM, Harney OM, Hogan MJ, Mitchell C, McGuire BE, Slattery B. Chronic pain self-management in middle-aged and older adults: a collective intelligence approach to identifying barriers and user needs in eHealth interventions. Digit Health 2022;8:20552076221105484 [FREE Full text] [doi: 10.1177/20552076221105484] [Medline: 35694121]

XSL•FO
RenderX

28. Faverio M. Share of those 65 and older who are tech users has grown in the past decade. Pew Research Center. 2022. URL: https://www.pewresearch.org/fact-tank/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/ [accessed 2024-04-29]
29. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Implement 2021 Mar;19(1):3-10. [doi: 10.1097/XEB.0000000000000277] [Medline: 33570328]
30. Peters MD, Godfrey C, McInerney P, Khalil H, Larsen P, Marnie C, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022 Apr 01;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
32. De Lucia A, Chiarotto A, Pasini I, Pachera S, Del Piccolo L, Perlini C, et al. e-Health multimodal interventions for older adults with chronic non-cancer pain: a scoping review protocol. medRxiv. Preprint posted online July 28, 2023 [FREE Full text] [doi: 10.1101/2023.07.27.23293235]
33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
34. Singh S, Bajorek B. Defining 'elderly' in clinical practice guidelines for pharmacotherapy. Pharm Pract (Granada) 2014 Oct;12(4):489 [FREE Full text] [doi: 10.4321/s1886-36552014000400007] [Medline: 25580172]
35. Definition of an older or elderly person. World Health Organization. 2010. URL: http://www.who.int/healthinfo/survey/ageingdefnolder/en/index.html [accessed 2024-04-29]
36. Orimo H, Ito H, Suzuki T, Araki A, Hosoi T, Sawabe M. Reviewing the definition of “elderly”. Geriatr Gerontol Int 2006 Aug 16;6(3):149-158. [doi: 10.1111/j.1447-0594.2006.00341.x]
37. Shenkin SD, Harrison JK, Wilkinson T, Dodds RM, Ioannidis JP. Systematic reviews: guidance relevant for studies of older people. Age Ageing 2017 Sep 01;46(5):722-728 [FREE Full text] [doi: 10.1093/ageing/afx105] [Medline: 28655142]
38. Sirriyeh R, Lawton R, Gardner P, Armitage G. Reviewing studies with diverse designs: the development and evaluation of a new tool. J Eval Clin Pract 2012 Aug;18(4):746-752. [doi: 10.1111/j.1365-2753.2011.01662.x] [Medline: 21410846]
39. Heleno E, Andias R, Neto M, Silva AG. A feasibility study of pain neuroscience education and exercise for community-dwelling older adults with chronic pain. J Geriatr Phys Ther 2023;46(1):26-35. [doi: 10.1519/JPT.0000000000000327] [Medline: 34417415]
40. Stamm O. ViRST- evaluation of the system for virtual reality pain therapy in geriatric patients with chronic backpain. The Federal Institute for Drugs and Medical Devices. URL: https://drks.de/search/en/trial/DRKS00020576 [accessed 2024-04-29]
41. Rybarczyk B, DeMarco G, DeLaCruz M, Lapidos S. Comparing mind-body wellness interventions for older adults with chronic illness: classroom versus home instruction. Behav Med 1999;24(4):181-190. [doi: 10.1080/08964289.1999.11879274] [Medline: 10023496]
42. Kelechi TJ, Hernandez-Tejada MA, Balasubramanian S, Bian J, Madisetti M, Nagel A. Addressing physical, functional, and physiological outcomes in older adults using an integrated mHealth intervention "active for life": a pilot randomized controlled trial. Int J Nurs Health Care Res (Lisle) 2022;5(3):1285 [FREE Full text] [doi: 10.29011/2688-9501.101285] [Medline: 36189445]
43. Hernandez-Tejada MA, Nagel A, Madisetti M, Balasubramanian S, Kelechi T. Feasibility trial of an integrated treatment "activate for life" for physical and mental well-being in older adults. Pilot Feasibility Stud 2022 Feb 11;8(1):38 [FREE Full text] [doi: 10.1186/s40814-022-01000-8] [Medline: 35148798]
44. Suman A, Schaafsma FG, van de Ven PM, Slottje P, Buchbinder R, van Tulder MW, et al. Effectiveness of a multifaceted implementation strategy compared to usual care on low back pain guideline adherence among general practitioners. BMC Health Serv Res 2018 May 11;18(1):358 [FREE Full text] [doi: 10.1186/s12913-018-3166-y] [Medline: 29751801]
45. Rini C, Porter LS, Somers TJ, McKee DC, DeVellis RF, Smith M, et al. Automated Internet-based pain coping skills training to manage osteoarthritis pain: a randomized controlled trial. Pain 2015 May;156(5):837-848 [FREE Full text] [doi: 10.1097/j.pain.0000000000000121] [Medline: 25734997]
46. Madill ES, Samuels R, Newman DP, Boudreaux-Kelley M, Weiner DK. Development of an evaluative, educational, and communication-facilitating app for older adults with chronic low back pain: patient perceptions of usability and utility. Pain Med 2019 Nov 01;20(11):2120-2128. [doi: 10.1093/pm/pnz088] [Medline: 31329964]
47. Allen K, Vu MB, Callahan LF, Cleveland RJ, Gilbert AL, Golightly YM, et al. Osteoarthritis physical activity care pathway (OA-PCP): results of a feasibility trial. BMC Musculoskelet Disord 2020 May 16;21(1):308 [FREE Full text] [doi: 10.1186/s12891-020-03339-6] [Medline: 32416715]
48. Wang L, Xie S, Bao T, Zhu S, Liang Q, Wang X, et al. Exercise and education for community-dwelling older participants with knee osteoarthritis: a video-linked programme protocol based on a randomised controlled trial. BMC Musculoskelet Disord 2021 May 22;22(1):470 [FREE Full text] [doi: 10.1186/s12891-021-04331-4] [Medline: 34022825]
49. Mace RA, Doorley JD, Popok PJ, Vranceanu AM. Live video adaptations to a mind-body activity program for chronic pain and cognitive decline: protocol for the virtual active brains study. JMIR Res Protoc 2021 Jan 04;10(1):e25351 [FREE Full text] [doi: 10.2196/25351] [Medline: 33208301]

XSL•FO
RenderX

50. Janevic M, Robinson-Lane SG, Murphy SL, Courser R, Piette JD. A pilot study of a chronic pain self-management program delivered by community health workers to underserved African American older adults. Pain Med 2022 Dec 01;23(12):1965-1978 [FREE Full text] [doi: 10.1093/pm/pnaa468] [Medline: 33779759]
51. Fanning J, Brooks AK, Ip E, Nicklas BJ, Rejeski WJ. A mobile health intervention to reduce pain and improve health (MORPH) in older adults with obesity: protocol for the MORPH trial. JMIR Res Protoc 2018 May 14;7(5):e128 [FREE Full text] [doi: 10.2196/resprot.9712] [Medline: 29759957]
52. Fanning J, Brooks AK, Ip E, Nicklas BJ, Rejeski WJ, Nesbit B, et al. A mobile health behavior intervention to reduce pain and improve health in older adults with obesity and chronic pain: the MORPH pilot trial. Front Digit Health 2020 Dec;2:598456 [FREE Full text] [doi: 10.3389/fdgth.2020.598456] [Medline: 33817686]
53. Fanning J, Brooks AK, Hsieh KL, Kershner K, Furlipa J, Nicklas BJ, et al. Building on lessons learned in a mobile intervention to reduce pain and improve health (MORPH): protocol for the MORPH-II trial. JMIR Res Protoc 2021 Jul 19;10(7):e29013 [FREE Full text] [doi: 10.2196/29013] [Medline: 34279241]
54. Fanning J, Brooks AK, Hsieh KL, Kershner K, Furlipa J, Nicklas BJ, et al. The effects of a pain management-focused mobile health behavior intervention on older adults' self-efficacy, satisfaction with functioning, and quality of life: a randomized pilot trial. Int J Behav Med 2022 Apr;29(2):240-246 [FREE Full text] [doi: 10.1007/s12529-021-10003-3] [Medline: 34018138]
55. Frontiers Production Office. Erratum: a remote group-mediated daylong physical activity intervention for older adults with chronic pain: results of the MORPH-II randomized pilot trial. Front Digit Health 2023;5:1298782 [FREE Full text] [doi: 10.3389/fdgth.2023.1298782] [Medline: 37877126]
56. Mace RA, Gates MV, Bullard B, Lester EG, Silverman IH, Quiroz YT, et al. Development of a novel mind-body activity and pain management program for older adults with cognitive decline. Gerontologist 2021 Apr 03;61(3):449-459 [FREE Full text] [doi: 10.1093/geront/gnaa084] [Medline: 32601670]
57. Vranceanu AM, Choukas NR, Rochon EA, Duarte B, Pietrzykowski MO, McDermott K, Active Brains Project, et al. Addressing the chronic pain-early cognitive decline comorbidity among older adults: protocol for the active brains remote efficacy trial. JMIR Res Protoc 2023 Sep 28;12:e47319 [FREE Full text] [doi: 10.2196/47319] [Medline: 37768713]
58. Harris A, Hinman RS, Lawford BJ, Egerton T, Keating C, Brown C, et al. Cost-effectiveness of telehealth-delivered exercise and dietary weight loss programs for knee osteoarthritis within a twelve-month randomized trial. Arthritis Care Res (Hoboken) 2023 Jun;75(6):1311-1319 [FREE Full text] [doi: 10.1002/acr.25022] [Medline: 36106928]
59. Bennell KL, Lawford BJ, Keating C, Brown C, Kasza J, Mackenzie D, et al. Comparing video-based, telehealth-delivered exercise and weight loss programs with online education on outcomes of knee osteoarthritis : a randomized trial. Ann Intern Med 2022 Feb;175(2):198-209. [doi: 10.7326/M21-2388] [Medline: 34843383]
60. Saraboon Y, Aree-Ue S, Maruo SJ. The effect of multifactorial intervention programs on health behavior and symptom control among community-dwelling overweight older adults with knee osteoarthritis. Orthop Nurs 2015;34(5):296-308. [doi: 10.1097/NOR.0000000000000180] [Medline: 26375841]
61. Berman RL, Iris MA, Bode R, Drengenberg C. The effectiveness of an online mind-body intervention for older adults with chronic pain. J Pain 2009 Jan;10(1):68-79 [FREE Full text] [doi: 10.1016/j.jpain.2008.07.006] [Medline: 18774342]
62. Doorley JD, Mace RA, Popok PJ, Grunberg VA, Ragnhildstveit A, Vranceanu AM. Feasibility randomized controlled trial of a mind-body activity program for older adults with chronic pain and cognitive decline: the virtual "active brains" study. Gerontologist 2022 Aug 12;62(7):1082-1094 [FREE Full text] [doi: 10.1093/geront/gnab135] [Medline: 34487167]
63. Fanning J, Brooks AK, Ford S, Robison JT, Irby MB, Rejeski WJ. A remote group-mediated daylong physical activity intervention for older adults with chronic pain: results of the MORPH-II randomized pilot trial. Front Digit Health 2022;4:1040867 [FREE Full text] [doi: 10.3389/fdgth.2022.1040867] [Medline: 36405417]
64. Janevic M, Robinson-Lane SG, Courser R, Brines E, Hassett AL. A community health worker-led positive psychology intervention for African American older adults with chronic pain. Gerontologist 2022 Oct 19;62(9):1369-1380 [FREE Full text] [doi: 10.1093/geront/gnac010] [Medline: 35394525]
65. Stamm O, Dahms R, Reithinger N, Ruß A, Müller-Werdan U. Virtual reality exergame for supplementing multimodal pain therapy in older adults with chronic back pain: a randomized controlled pilot study. Virtual Real 2022;26(4):1291-1305 [FREE Full text] [doi: 10.1007/s10055-022-00629-3] [Medline: 35194374]
66. Godziuk K, Prado CM, Quintanilha M, Forhan M. Acceptability and preliminary effectiveness of a single-arm 12-week digital behavioral health intervention in patients with knee osteoarthritis. BMC Musculoskelet Disord 2023 Feb 17;24(1):129 [FREE Full text] [doi: 10.1186/s12891-023-06238-8] [Medline: 36797720]
67. Pearson J, Walsh N, Carter D, Koskela S, Hurley M. Developing a web-based version of an exercise-based rehabilitation program for people with chronic knee and hip pain: a mixed methods study. JMIR Res Protoc 2016 May 19;5(2):e67 [FREE Full text] [doi: 10.2196/resprot.5446] [Medline: 27197702]
68. Khan K, Hall CL, Davies EB, Hollis C, Glazebrook C. The effectiveness of web-based interventions delivered to children and young people with neurodevelopmental disorders: systematic review and meta-analysis. J Med Internet Res 2019 Nov 01;21(11):e13478 [FREE Full text] [doi: 10.2196/13478] [Medline: 31682573]
69. m-Health: new horizons for health through mobile technologies. World Health Organization. 2011. URL: https://iris.who.int/bitstream/handle/10665/44607/9789241564250_eng.pdf?sequence=1 [accessed 2024-04-29]

XSL•FO
RenderX

70. Andersson G, Lundström P, Ström L. Internet-based treatment of headache: does telephone contact add anything? Headache 2003 Apr;43(4):353-361. [doi: 10.1046/j.1526-4610.2003.03070.x] [Medline: 12656706]
71. Herbert MS, Afari N, Liu L, Heppner P, Rutledge T, Williams K, et al. Telehealth versus in-person acceptance and commitment therapy for chronic pain: a randomized noninferiority trial. J Pain 2017 Feb;18(2):200-211 [FREE Full text] [doi: 10.1016/j.jpain.2016.10.014] [Medline: 27838498]
72. Botella C, Garcia-Palacios A, Vizcaíno Y, Herrero R, Baños RM, Belmonte MA. Virtual reality in the treatment of fibromyalgia: a pilot study. Cyberpsychol Behav Soc Netw 2013 Mar;16(3):215-223. [doi: 10.1089/cyber.2012.1572] [Medline: 23496678]
73. Garcia-Palacios A, Herrero R, Vizcaíno Y, Belmonte MA, Castilla D, Molinari G, et al. Integrating virtual reality with activity management for the treatment of fibromyalgia: acceptability and preliminary efficacy. Clin J Pain 2015 Jun;31(6):564-572. [doi: 10.1097/AJP.0000000000000196] [Medline: 25551475]
74. Haleem A, Javaid M, Singh RP, Suman R. Medical 4.0 technologies for healthcare: features, capabilities, and applications. Internet Things Cyber Phys Syst 2022;2:12-30. [doi: 10.1016/j.iotcps.2022.04.001]
75. Mannheim I, Schwartz E, Xi W, Buttigieg SC, McDonnell-Naughton M, Wouters EJ, et al. Inclusion of older adults in the research and design of digital technology. Int J Environ Res Public Health 2019 Oct 02;16(19):3718 [FREE Full text] [doi: 10.3390/ijerph16193718] [Medline: 31581632]
76. van Deursen AJ. Digital inequality during a pandemic: quantitative study of differences in COVID-19-related internet uses and outcomes among the general population. J Med Internet Res 2020 Aug 20;22(8):e20073 [FREE Full text] [doi: 10.2196/20073] [Medline: 32750005]
77. Frishammar J, Essén A, Bergström F, Ekman T. Digital health platforms for the elderly? Key adoption and usage barriers and ways to address them. Technol Forecast Soc Change 2023 Apr;189:122319. [doi: 10.1016/j.techfore.2023.122319]
78. Recchia G, Maria Capuano D, Mistri N, Verna R. Digital therapeutics-what they are, what they will be. Act Scie Medic Sci 2020 Feb 26;4(3):1-9. [doi: 10.31080/asms.2020.04.0575]
79. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC (text with EEA relevance.). European Union. 2023 Mar. URL: http://data.europa.eu/eli/reg/2017/745/oj [accessed 2024-04-29]
80. ISO/TR 11147:2023(E). Health informatics: personalized digital health: digital therapeutics health software systems. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/83767.html [accessed 2024-04-29]
81. Mantovani A, Leopaldi C, Nighswander CM, Di Bidino R. Access and reimbursement pathways for digital health solutions and diagnostic devices: current scenario and challenges. Front Med Technol 2023;5:1101476 [FREE Full text] [doi: 10.3389/fmedt.2023.1101476] [Medline: 36891483]
82. Trancart A, Riche VP, Disset A, Camus D, Josseran A, Bécache P, et al. Evaluation of digital medical devices: how to take into account the specificities of these solutions? Therapie 2024 Jan;79(1):137-150. [doi: 10.1016/j.therap.2024.01.001] [Medline: 38307754]
83. Lin J, Paganini S, Sander L, Lüking M, Ebert DD, Buhrman M, et al. An internet-based intervention for chronic pain. Dtsch Arztebl Int 2017 Oct 13;114(41):681-688 [FREE Full text] [doi: 10.3238/arztebl.2017.0681] [Medline: 29082858]
84. Garcia LM, Birckhead BJ, Krishnamurthy P, Sackman J, Mackey IG, Louis RG, et al. An 8-week self-administered at-home behavioral skills-based virtual reality program for chronic low back pain: double-blind, randomized, placebo-controlled trial conducted during COVID-19. J Med Internet Res 2021 Feb 22;23(2):e26292 [FREE Full text] [doi: 10.2196/26292] [Medline: 33484240]
85. Donisi V, Poli S, Mazzi MA, Gobbin F, Schena F, Del Piccolo L, et al. Promoting participatory research in chronicity: the ESPRIMO biopsychosocial intervention for young adults with multiple sclerosis. Front Psychol 2022;13:1042234 [FREE Full text] [doi: 10.3389/fpsyg.2022.1042234] [Medline: 36405126]
86. Gilbert RM. Reimagining digital healthcare with a patient-centric approach: the role of user experience (UX) research. Front Digit Health 2022;4:899976 [FREE Full text] [doi: 10.3389/fdgth.2022.899976] [Medline: 36016600]
87. Zizza CA, Ellison KJ, Wernette CM. Total water intakes of community-living middle-old and oldest-old adults. J Gerontol A Biol Sci Med Sci 2009 Apr;64(4):481-486 [FREE Full text] [doi: 10.1093/gerona/gln045] [Medline: 19213852]
88. Escourrou E, Durrieu F, Chicoulaa B, Dupouy J, Oustric S, Andrieu S, et al. Cognitive, functional, physical, and nutritional status of the oldest old encountered in primary care: a systematic review. BMC Fam Pract 2020 Mar 27;21(1):58 [FREE Full text] [doi: 10.1186/s12875-020-01128-7] [Medline: 32220228]
89. Donisi V, Tedeschi F, Gonzalez-Caballero JL, Cresswell-Smith J, Lara E, Miret M, et al. Is mental well-being in the oldest old different from that in younger age groups? Exploring the mental well-being of the oldest-old population in Europe. J Happiness Stud 2020 Jul 14;22(4):1693-1717. [doi: 10.1007/s10902-020-00292-y]
90. Poli A, Kelfve S, Berg K, Motel-Klingebiel A. Old-age diversity is underrepresented in digital health research: findings from the evaluation of a mobile phone system for post-operative progress monitoring in Sweden. Ageing Soc 2021 Nov 12;43(10):2264-2286. [doi: 10.1017/s0144686x21001641]
91. Borra C, Hardy R. Differences in chronic pain prevalence between men and women at mid-life: a systematic review protocol. BMJ Open 2023 Apr 28;13(4):e065497 [FREE Full text] [doi: 10.1136/bmjopen-2022-065497] [Medline: 37116997]

XSL•FO
RenderX

92. Fillingim RB, Loeser JD, Baron R, Edwards RR. Assessment of chronic pain: domains, methods, and mechanisms. J Pain 2016 Sep;17(9 Suppl):T10-T20 [FREE Full text] [doi: 10.1016/j.jpain.2015.08.010] [Medline: 27586827]
93. Boring BL, Walsh KT, Nanavaty N, Ng BW, Mathur VA. How and why patient concerns influence pain reporting: a qualitative analysis of personal accounts and perceptions of others' use of numerical pain scales. Front Psychol 2021;12:663890 [FREE Full text] [doi: 10.3389/fpsyg.2021.663890] [Medline: 34282355]
94. Boring BL, Ng BW, Nanavaty N, Mathur VA. Over-rating pain is overrated: a fundamental self-other bias in pain reporting behavior. J Pain 2022 Oct;23(10):1779-1789 [FREE Full text] [doi: 10.1016/j.jpain.2022.06.002] [Medline: 35724938]
95. Chen J, Wang X, Xu Z. The relationship between chronic pain and cognitive impairment in the elderly: a review of current evidence. J Pain Res 2023;16:2309-2319 [FREE Full text] [doi: 10.2147/JPR.S416253] [Medline: 37441267]
96. Lam N, Green J, Hallas S, Forster A, Crocker TF, Andre D, et al. Mapping review of pain management programmes and psychological therapies for community-dwelling older people living with pain. Eur Geriatr Med 2024 Feb;15(1):33-45 [FREE Full text] [doi: 10.1007/s41999-023-00871-1] [Medline: 37853269]
97. Raja SN, Carr DB, Cohen M, Finnerup NB, Flor H, Gibson S, et al. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. Pain 2020 Sep 01;161(9):1976-1982 [FREE Full text] [doi: 10.1097/j.pain.0000000000001939] [Medline: 32694387]
98. Gibson CJ, Grasso J, Li Y, Purcell N, Tighe J, Zamora K, et al. An integrated pain team model: impact on pain-related outcomes and opioid misuse in patients with chronic pain. Pain Med 2020 Sep 01;21(9):1977-1984. [doi: 10.1093/pm/pnaa003] [Medline: 32100002]
99. van Acker J, Maenhout L, Compernolle S. Older adults' user engagement with mobile health: a systematic review of qualitative and mixed-methods studies. Innov Aging 2023;7(2):igad007 [FREE Full text] [doi: 10.1093/geroni/igad007] [Medline: 37007638]
100. Stamm O, Dahms R, Müller-Werdan U. Virtual reality in pain therapy: a requirements analysis for older adults with chronic back pain. J Neuroeng Rehabil 2020 Sep 29;17(1):129 [FREE Full text] [doi: 10.1186/s12984-020-00753-8] [Medline: 32993678]
101. Olsen MF, Bjerre E, Hansen MD, Tendal B, Hilden J, Hróbjartsson A. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: systematic review of empirical studies. J Clin Epidemiol 2018 Sep;101:87-106.e2. [doi: 10.1016/j.jclinepi.2018.05.007] [Medline: 29793007]
102. Nussbaumer-Streit B, Klerings I, Dobrescu AI, Persad E, Stevens A, Garritty C, et al. Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. J Clin Epidemiol 2020 Feb;118:42-54. [doi: 10.1016/j.jclinepi.2019.10.011] [Medline: 31698064]

Abbreviations

CNCP: chronic noncancer pain
DT: digital therapeutic
HRQoL: health-related quality of life
MCID: minimal clinically important difference
MD: medical device
mHealth: mobile health
MORPH-II: Mobile Intervention to Reduce Pain and Improve Health II
NICE: National Institute for Health and Care Excellence
NRS: numeric rating scale
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PROMIS: Patient-Reported Outcomes Measurement Information System
QATSDD: Quality Assessment Tool for Studies with Diverse Designs
RCT: randomized controlled trial
RQ: research question
VR: virtual reality

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 11.12.23; peer-reviewed by L Benning, D Singh; comments to author 21.02.24; revised version received 11.03.24; accepted 30.04.24; published 29.07.24.

Please cite as:
De Lucia A, Perlini C, Chiarotto A, Pachera S, Pasini I, Del Piccolo L, Donisi V
eHealth-Integrated Psychosocial and Physical Interventions for Chronic Pain in Older Adults: Scoping Review
J Med Internet Res 2024;26:e55366
URL: https://www.jmir.org/2024/1/e55366
doi:10.2196/55366
PMID:39073865

©Annalisa De Lucia, Cinzia Perlini, Alessandro Chiarotto, Sara Pachera, Ilenia Pasini, Lidia Del Piccolo, Valeria Donisi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Online Medical Consultation Service–Oriented Recommendations: Systematic Review

Hongxun Jiang[1], PhD; Ziyue Mi[1], MS; Wei Xu[1,2], PhD

[1]School of Information, Renmin University of China, Beijing, China

[2]School of Smart Governance, Renmin University of China, Suzhou, China

Corresponding Author:
Wei Xu, PhD
School of Information, Renmin University of China
59 Zhongguancun Street
Haidian District
Beijing, 100872
China
Phone: 86 01082500904
Email: weixu@ruc.edu.cn

Abstract

Background: Online health communities have given rise to a new e-service known as online medical consultations (OMCs), enabling remote interactions between physicians and patients. To address challenges, such as patient information overload and uneven distribution of physician visits, online health communities should develop OMC-oriented recommenders.

Objective: We aimed to comprehensively investigate what paradigms lead to the success of OMC-oriented recommendations.

Methods: A literature search was conducted through e-databases, including PubMed, ACM Digital Library, Springer, and ScienceDirect, from January 2011 to December 2023. This review included all papers directly and indirectly related to the topic of health care–related recommendations for online services.

Results: The search identified 611 articles, of which 26 (4.3%) met the inclusion criteria. Despite the growing academic interest in OMC recommendations, there remains a lack of consensus among researchers on the definition of e-service–oriented recommenders. The discussion highlighted 3 key factors influencing recommender success: features, algorithms, and metrics. It advocated for moving beyond traditional e-commerce–oriented recommenders to establish an innovative theoretical framework for e-service–oriented recommenders and addresses critical technical issues regarding 2-sided personalized recommendations.

Conclusions: This review underscores the essence of e-services, particularly in knowledge- and labor-intensive domains such as OMCs, where patients seek interpretable recommendations due to their lack of domain knowledge and physicians must balance their energy levels to avoid overworking. Our study's findings shed light on the importance of customizing e-service–oriented personalized recommendations to meet the distinct expectations of 2-sided users considering their cognitive abilities, decision-making perspectives, and preferences. To achieve this, a paradigm shift is essential to develop unique attributes and explore distinct content tailored for both parties involved.

(J Med Internet Res 2024;26:e46073) doi:10.2196/46073

KEYWORDS

online health community; online medical consultation; personalized recommendations; 2-sided matching; load balancing

Introduction

Background

Technology innovations have brought the medical industry into the digital, networked, and intelligent era of the medical internet. Combined with the impact of the pandemic, telemedicine increasingly prevails. A growing number of patients have started to seek contactless counseling from physicians through online health communities (OHCs). Meanwhile, physicians are providing the public with health care posts, free medical consultations, and even paid customized service [1-3]. HaoDF, one of the leading OHC platforms in China, has gathered >890,000 physicians from 10,000 hospitals across the country up to March 2023. It offers patients the service of telehealth or web-based live chat (ie, online medical consultation [OMC]). Telehealth offers greater convenience to patients than the

XSL•FO
RenderX

in-person visits previously available. However, it worsens the problem of information overload, as there are too many candidates for users to choose from, which exacerbates the level of hesitation [4]. Patients face challenges in selecting suitable physicians due to limited medical knowledge and cognitive abilities. An OMC-oriented recommendation system is crucial to provide patients with professional, accurate, and responsible referrals, ensuring that they connect with qualified and suitable physicians.

Most existing studies of physician recommendations are in the wrong direction regardless of their diverse methodologies, such as collaborative filtering (CF), demographic statistics, or association rules. Previous research has overlooked the fact that OHCs serve both patients and physicians (ie, a 2-sided market scenario). Figure 1 illustrates an OHC jointly formed by patients and physicians. When a market is 2-sided, there are cross-network externalities, which means that the number of users on one side will affect the number of users on the other side and the overall transaction volume on both sides [5]. An OMC recommendation is a service that an OHC offers to both sides of users. A type of online service such as OMC (ie, an e-service) is an emerging field of internet business under the knowledge economy. As opposed to e-commerce, an e-service is composed of consultees and consultants rather than users and commodities. The offered item is an intangible service rather than a tangible one, but it has to meet the different needs, expectations, and preferences of 2-sided users. Furthermore, medical consultations are knowledge- and labor-intensive services that demand high levels of professionalism and energy investment [6]. The energy limits of physicians vary, and each physician can receive consultations only to a certain extent. In addition, patients lack the professional knowledge to distinguish the candidates, so they need recommendations that can be interpreted. Thus, it is impossible to transplant an e-commerce recommendation model to solve OMC recommendation cases. Research on recommendation systems suffers from a "blind side" that is the lack of research focusing on service-oriented applications, requiring academicians to develop new attributes and research new content. OMC services demonstrate the typical characteristics of online knowledge services, which represent the emerging trend of the "Internet+" economy. In the context of the knowledge economy, research on service recommendations is particularly pertinent, and now is an excellent time to start. As far as we know, no comprehensive research has been conducted in academic circles on service recommendations. Personalized service recommendation is a new topic yet to be clearly defined and fully explored.

Figure 1. Online health communities function as 2-sided markets involving both patients and physicians. In such markets, the user base on one side of the platform affects the user base on the other side, leading to cross-network externalities that influence the overall transaction volume across both sides. Recommendation systems must consider both patients' and physicians' needs and preferences.

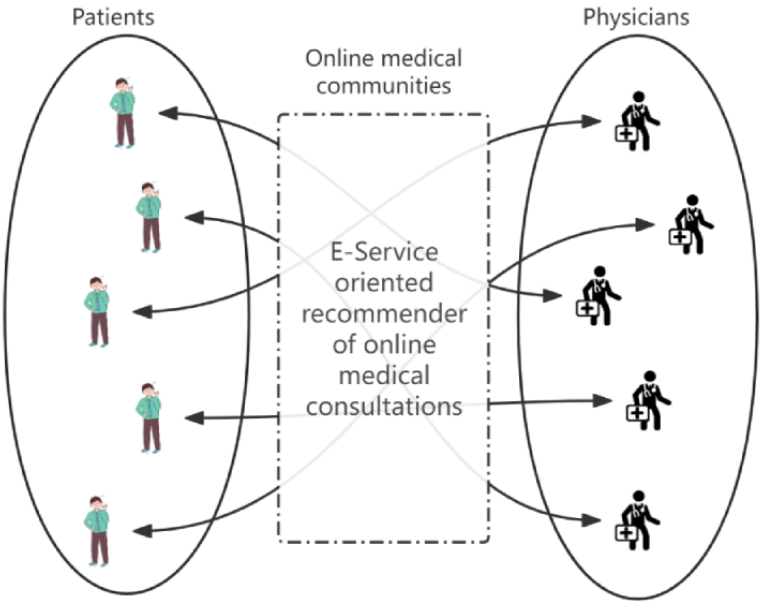

Objectives

Despite several reviews on health care recommender systems focusing on patient interests [7-9], there remains a gap in service-oriented recommendations. Our systematic review aimed to fill this void by concentrating solely on 2-sided recommendations. By providing the latest review of this domain, we aimed to gather comprehensive evidence for evaluating current studies, identifying successful paradigms and approaches in service-oriented recommendations, and informing public health interventions and policy making. This will leverage 2-sided recommendation technologies to enhance the well-being of both patients and physicians in the emerging OMC service industry.

Methods

This review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1 [10]) [11].

Search Strategy

As the OMC service–oriented recommendation system spans multiple disciplines such as health care, business information systems, and computer science, the authors conducted separate

XSL•FO
RenderX

literature searches in databases from each field. These included 1 medical-focused database (PubMed); 1 computer-focused database (ACM Digital Library); and 2 multidisciplinary full-text databases (SpringerLink and ScienceDirect) from 2 leading publishing groups, Springer and Elsevier, respectively. Our search was tailored to the review topic followed by an analysis of text words found in titles, abstracts, and keywords used in retrieved papers. The electronic search was conducted in December 2023 using keyword combinations in the title, abstract, and keywords fields to ensure comprehensive coverage. Keywords were selected and classified into 4 categories: OMC (subject of the study), recommendation (objective of the study), OHC (fields of the study), and excluded keywords—queries 1 to 4. To emphasize the recent advancements, an additional query 5 set a time limit from January 1, 2001, to November 1, 2023. The overall search strategy was 1 AND 2 AND 3 AND (NOT 4) AND 5. Table 1 presents the hierarchical search query and all keywords.

Following the keyword search, a reference list search (ie, backward reference search) and a cited reference search (ie, forward reference search) were conducted on the full-text articles that met the study selection criteria. Using the results of the backward and forward reference searches, the same study selection criteria were applied to further screen and evaluate articles. We repeated these procedures on all newly identified articles until no additional relevant articles were found.

Table 1. Literature search strategy.

Search number	Search	Keywords
1	Title	*(doctor OR physician OR consultation OR treatment OR e-health OR m-health OR telehealth OR remote health OR digital health OR online medical service OR web-based, health OR internet-based, health)*
2	Title	*(recommendation OR recommender OR recommending OR matching OR rating OR choosing OR selection)*
3	Title/abstract/keywords	*((Health OR Healthcare) AND (Communities OR Forums OR Platforms))*
4	Title/abstract/keywords	*(Qualitative research OR Practice)*
5	Time range	*January 1, 2011, to November 1, 2023*

Eligibility Criteria

The titles and abstracts of identified articles were independently screened by 2 researchers (HJ and ZM) to determine inclusion in the full review. Figure 2 illustrates the paper selection process. If either or both reviewers selected the paper for further evaluation, it was included for full assessment. Articles were considered for analysis if they met at least one of the following criteria: (1) OHC-oriented physician recommendations, (2) coding or documenting of patient preferences, (3) motivations or perceptions of physicians involved in OHCs, (4) implementation of a recommender system for medical services, and (5) recommendation acceptance and interface design in the domain of medical-related recommender systems. Disagreements were resolved with a third reviewer (WX) until consensus was reached. In addition, articles must have met the following three criteria to be considered for analysis: (1) published in peer-reviewed journals or conference proceedings, excluding research articles without detailed research designs or results; (2) written in English; and (3) published between 2011 and 2023 to align with the recent emergence of OHCs over the last decade.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the study.

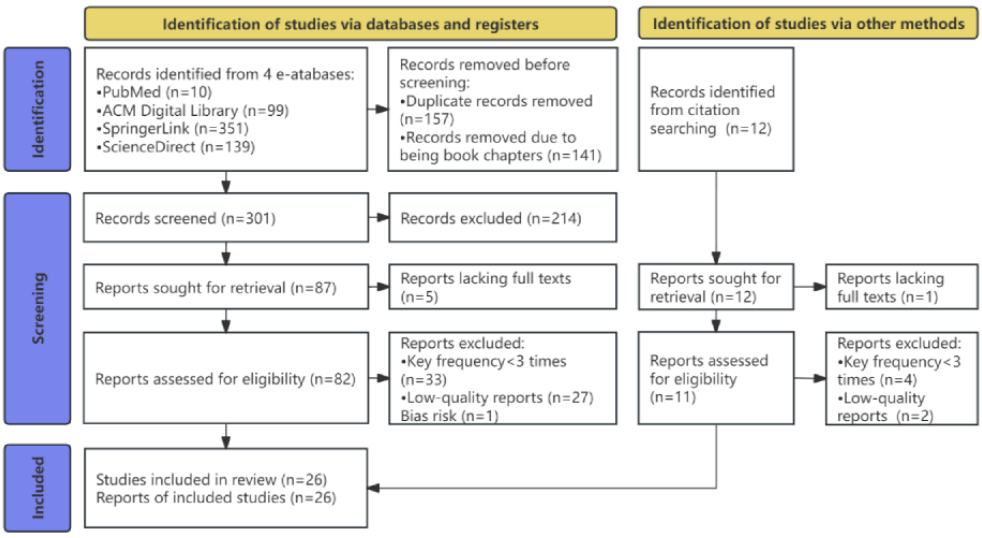

XSL•FO
RenderX

Quality Assessment

To ensure the quality of the articles, we applied a GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) framework. The purpose of this initiative is to help individuals make informed decisions using evidence systematically and transparently [12]. The GRADE Evidence to Decision frameworks have been illustrated and are useful in making and using health-related recommendations and decisions [13]. They identify 4 levels of evidence for each study: very low, low, moderate, and high. GRADE criteria examine the risk of bias, imprecision, inconsistency, indirectness, and publication bias in evaluating the quality of evidence of a study [14]. In the scenario of OMC recommendation design, those studies start at high quality of evidence if they use both offline data sets and online data streams for randomized controlled experiments. In contrast, observational studies begin at a lower quality of evidence due to residual confounding. Referring to previous studies [15], only moderate and high-quality articles were selected to avoid low-quality articles.

Results

Search Results

The electronic database search yielded 599 studies, with an additional 12 studies identified from Google Scholar through reference list and cited reference searches for each obtained study. After removing book chapters and deduplicating entries, 49.3% (301/611) of the studies remained. The review process involved excluding 71.1% (214/301) of ineligible studies after screening titles and abstracts as they did not meet the criteria specific to the theme of physician recommendations, the OMC domain, peer-reviewed sources, or English language. This left 87 studies for full-text review, of which 5 (6%) reports were found to have no full texts and were subsequently removed. The full texts of the remaining 94% (82/87) of the studies were assessed for bias risks and qualitatively analyzed. Ultimately, 21 high-quality studies with adequate outcome data were selected for quantitative analysis. In addition, 5 studies were identified through citation searching and included in the quantitative analysis.

Study Characteristics

Table 2 summarizes the features and discoveries of each of the 26 studies. The number of publications increased over time. This indicates that this field is receiving more and more attention from scholars and practitioners due to the global prosperity of OHCs. This review comprised studies from 6 countries across Asia, Europe, and North America, most of which were low-income countries. The largest source of articles was China, followed by Portugal and India. This reflects that, in countries with less developed offline health networks, online medical services are of particular benefit, as illustrated by the growth of OMCs in China. There were various recommendation algorithms used: analytic hierarchy process, CF, content-based filtering, decision tree, neural network, matrix factorization, and regression analysis. Although research has been conducted using a variety of methods, CF, matrix factorization, and analytic hierarchy process are the top 3 most commonly used ones, with 38% (10/26) of the studies involving their use entirely or in part. However, over the last 5 years, the research methods of graph-based deep learning have become increasingly popular. The most widely used method of data engineering in the field is text analysis, such as latent Dirichlet allocation (LDA) and word2vec, followed by knowledge graphs (KGs).

Of the 26 studies included, 13 (50%) used data directly from OHCs as their source, whereas 7 (27%) used data indirectly from the official websites of hospitals or health care centers. Depending on the sources, the data sets could be classified as online or offline. There are numerous OHCs that produce massive amounts of heterogeneous, multimodal, and high-dimensional raw data continuously [16,17]. Online data generated by these OHCs support medical diagnosis and decision-making. Offline data are usually collected from various medical institutions and typically stored in health care information systems. They are exported once permission has been granted [18-22]. In addition, many studies collected primary data through questionnaires directed at patients or physicians [23]. Depending on the objects described, the data sets could be classified as patient, physician, and institution data, as shown in

Table 3. The data quality of physician profiles, such as educational background, professional experience, disciplines, and expertise, was high and well defined. The patients were OHC users and service consumers, but they were nonexperts or laypersons in the medical field who usually chat on the web without any restrictions or limitations, which results in poor-quality data from their consultations. This makes data processing and feature extraction quite complicated and challenging.

XSL•FO
RenderX

Table 2. The bibliographic characteristics of the included studies.

Study	Date	Country	Study aim	Method	Data sources
Huang et al [18]	December 2012	China	Using patient preferences and physician performance to recommend doctors	CF[a] and AHP[b]	Official appointment platform for the Shanghai Medical League
Jiang and Xu [24]	December 2014	China	Combining the relevance and quality of doctors in an integrated recommender system	Semantic similarity computing and AHP	OHCs[c]: HaoDF, XYWY, Ask39, and 51daifu
Gong et al [19]	September 2015	China	Using medical social networks and a medical data set to recommend doctors	Time-constraint probability factor graph and random walk with restart	Clinic experiments at the Chinese Academy of Sciences
Narducci et al [25]	May 2015	Italy	Delivering a semantic recommender system based on social networks	Similarity computing and CF	__[d]
Guo et al [26]	July 2016	China	Identifying KOLs[e] using health care data mining for any specific disease	Unsupervised aggregation approach	Medical journal papers
Zhang et al [27]	January 2017	China and United States	Using topic modeling and emotional offset to recommend doctors	Matrix factorization, LDA[f], and sentiment analysis	Yelp
Sridevi and Rajeshwara [28]	August 2018	India	A personalized physician recommender	Similarity computation and combined ratings	—
Han et al [20]	October 2018	Portugal	Establishing a mechanism for matching patients with family doctors	Hybrid matrix factorization and latent representation	Consultation histories of a leading European health care provider
Waqar et al [23]	January 2019	Pakistan	Combining content-based and collaborative and demographic filtering to create a hybrid physician recommender	Content-based filtering, CF, similarity measure, and AHP	Survey data from 3 hospitals in Islamabad, Pakistan
Pan et al [29]	January 2019	China	Personalizing physician selections based on patient preferences and illness conditions	Dynamic assortment planning and upper confidence bound	Simulation data
Xu et al [30]	June 2019	China	Recommendations based on doctors' reputation scores and similarities with patients' demands	Truth discovery, modified Paillier cryptosystem, and Dirichlet distribution	Simulation data
Ye et al [31]	August 2019	China	Picking doctors using signaling theory with patient needs	Binary long short-term memory, LDA, regression, and AHP	OHCs: HaoDF and XYWY
Yang et al [32]	February 2020	China	Enhancing physician recommendations based on patient preferences	Intuitionistic fuzzy sets and Bonferroni mean	OHC: HaoDF
Wen et al [33]	April 2020	China	Providing real-time personalized recommendations by optimizing limited physician resources	Adjusted exponential inventory balancing	Simulation data
Mondal et al [21]	October 2020	India	Modeling patient-physician relationships to recommend doctors	Multilayer graph data model	Records from health centers and hospitals
Yan et al [34]	October 2020	China	Fusing review text and physician information to improve medical consultation recommendations	Convolutional neural network and probabilistic matrix factorization	OHC: HaoDF
Meng and Xiong [35]	January 2021	China	To propose a hybrid physician recommendation model based on OHCs	Eigenvector, word2vec, and LDA	OHC: Chunyu
Peito and Han [36]	January 2021	Portugal	Developing a content-based matchmaking system for patients and doctors	Pretrained Poincaré embeddings and transfer learning	A data set of a European private health network
Wang et al [37]	January 2021	China	Proposing a diversity-enhanced hierarchical physician recommendation approach	Matrix factorization and heuristics	OHC: HaoDF

XSL•FO
RenderX

Study	Date	Country	Study aim	Method	Data sources
Ju and Zhang [38]	August 2021	China	Ontology-based recommendation of doctors based on disease text mining	Ontology and text mining	OHC: GuaHao
Yuan and Deng [4]	February 2022	China	Using knowledge graphs and deep learning to recommend doctors based on OHCs	Knowledge graph and deep learning	OHC: HaoDF
Lu et al [39]	May 2022	China	Recommending doctors through expertise learning in OHCs	Multi-head attention and pretrained BERT[g]	OHC: Chunyu
Chen et al [40]	July 2022	China	Considering patients' risk preference in a probabilistic linguistic environment to recommend doctors	Probabilistic linguistic term set, TF-IDF[h], and word2vec	OHC: HaoDF
Wang et al [41]	August 2022	China	Developing a model to predict patients' preferences regarding medical consultations based on physician characteristics	LASSO[i], multilayer perceptron, decision tree, and Shapley Additive Explanations	OHC: HaoDF
Wu et al [42]	February 2023	China	Making a decision-making method for online physician selection that considers correlation	Choquet integral, BERT, and 2-additive fuzzy measure	OHC: DXY
Valdeira et al [22]	August 2023	Portugal	Physician recommendation with implicit feedback and limited patient information	Deep extreme classification with label features	Consultations of a European private health network

[a]CF: collaborative filtering.
[b]AHP: analytic hierarchy process.
[c]OHC: online health community.
[d]Not applicable.
[e]KOL: key opinion leader.
[f]LDA: latent Dirichlet allocation.
[g]BERT: Bidirectional Encoder Representations from Transformers.
[h]TF-IDF: term frequency–inverse document frequency.
[i]LASSO: least absolute shrinkage and selection operator.

Table 3. The online health community data set contains various categories and features, including information related to physicians, patients, and hospitals.

Category and features	Description
Physician profiles	
ID, name, age, gender, geographic location, hospital, and department	Physician's personal information
Specialties, number of patients, and professional title	Professional experience and expertise
Academic background, research achievements, and academic titles	Academic background
Patient ratings, patient reviews, and patient satisfaction	Online and offline word of mouth
Number of popular science articles	Physicians' willingness to engage in science popularization.
Historical records	Physicians' historical consultations
Patient profiles	
ID, gender, age, and location	Patients' personal basic information
Disease description and medical history	Disease information provided by the patient in advance
Consulting records	Records of patient consultations with physicians
Hospital information	
Hospital grade and ranking	Hospital reputation

XSL•FO
RenderX

Data and Feature Engineering

An accurate acquisition of features enables an effective recommendation system, and feature engineering forms the foundation of personalized recommendation systems. Data engineering begins with raw data preprocessing. Duplicate or missing values can be handled by deleting them or using average values. Semistructured data, such as the demographics of physicians or patients, need to be converted into structured data by recognizing named entities and extracting information. When analyzing unstructured data, such as physician-patient consultation records, the content may be nonstandard, repetitive, short, and straightforward. Pycorrector, a third-party open-source library developed by Python, can be used to correct some common errors in oral expression [38]. Afterward, word separation, deactivation removal, normalization, and other procedures are performed. A word separation process extracts and vectorizes text features. Considering the specificity and professional nature of the medical field, the consultation records contain many medical professional words, and synonymous disease names must be substituted (eg, the term "trisomy 21" indicates a pediatric Down syndrome disorder). To ensure that professional terms are recognized during word segmentation, it is recommended to develop a dictionary based on medical ontologies. Furthermore, medical experts can be consulted to refine the dictionary by deleting terms outside the required domain [4]. Afterward, stop words should be removed to eliminate meaningless words or characters and reduce noise. For word segmentation, the most commonly used tools are Jieba and WordNet Lemmatizer in the Natural Language Toolkit library; for removing stop words, the most commonly used lexicons include the Harbin Institute of Technology stop word list, the Baidu stop word list, and stop words in the Natural Language Toolkit library.

OMC recommendations also face data sparsity challenges. Domain specialization leads to data sparsity. An OMC is not a domain of fast-moving consumer goods but a professional service. Most people do not consult physicians regularly but rather initiate consultations only when they need one, such as when a condition arises. In most cases, patients will consult only 1 physician for a condition or disease. Once cured, they will not revisit the same physician; otherwise, they will try another physician. In other words, it is rare for a physician and patient to have multiple records of the same condition or disease. Despite OHCs having an extensive collection of physicians, most of those physicians are considered "silent" in the communities as, in most cases, patients pay attention only to those physicians who are well-known and highly regarded. It was only possible for patients to rate or write reviews for physicians they had consulted rather than for other physicians. All these factors contribute to data sparsity.

To alleviate data sparsity, either the model should be improved or more features should be mined. According to the literature [30], patients' uncertain characteristics and preferences could be revealed through uncertainty languages, and fuzzy analysis could be used to improve recommender systems' sparsity problem. KGs have been introduced to represent physician-patient interaction features in the physician recommendation problem, thereby alleviating data sparsity [4]. A sociosemantic approach was used to address the problem of data sparsity caused by user-based CF [43]. Son and Choi [44] used ordinal and binary ratings of experts to refine user opinions and mitigated data sparsity in hand-edited expert recommendations. Wang et al [37] proposed a matrix decomposition to handle sparse data and improve prediction accuracy.

Medicine is a very specialized field of science. Often, because of cognitive limitations, patients cannot express their conditions and medical histories in consultation content, and some are unable to even express their personal needs. Using topic models, unstructured texts are analyzed for their content to retrieve, classify, cluster, summarize, and find topics that have similarities or relevance. The most common topic modeling method, LDA, uses an unsupervised probabilistic model to generate topics [31]. Typically, LDA is used to extract topics from large data sets of documents by mining potential semantic relationships between them. Meng and Xiong [35] used all physician consultations as a corpus for LDA, as shown in Figure 3, and each physician's text-topic distribution was then used to train a model to retrieve the corresponding physician for a specific topic. Zhang et al [27] applied LDA to extract patients' potential preferences and the characteristics of the physicians they consulted from patient reviews. LDA has some shortcomings. First, LDA lacks semantic contextual information when processing text because the commonly used bag-of-words model ignores it [34]. Second, LDA models perform poorly when text topics are too sparse to represent potential features; training LDAs tends to overfit if there are too many topics, so a fair number of topics must be selected to strike a balance between the degree of fit and simplicity. Finally, LDA models cannot handle labeled data on documents, causing uninterpretable topics to be generated. From various perspectives, scholars have proposed solutions to the aforementioned drawbacks. Ye et al [31] reduced the time complexity of LDA via Gibbs sampling and determined the optimal number of LDA topics based on the confusion level. Because the patient's "initial inquiry" text is usually short and the corresponding topic vector representation is sparse, Liu [45] used a short text aggregation algorithm to represent the topic vector. Pan and Ni [46] used a labeled LDA model to generate probability distributions for health questions and topics and topics or words based on the text set of physicians' answers to health questions.

Sentiment analysis identifies users' attitudes and opinions on commodities or services from their review texts. In addition to medical topics, consultations and patient evaluations on OHCs include patient emotions and feelings. Using sentiment mining techniques, sentiment information can be extracted from text data. Text sentiment analysis can be divided into 2 main types: lexicon based and deep learning based. Sentiment dictionaries are the traditional tool for analyzing words and short texts' sentiment tendencies [31]. These dictionaries describe not only the positive and negative sentiment attributes of words in static dictionaries but also the offsets of sentiment information of words in sentence frameworks. The China National Knowledge Infrastructure, the Information Retrieval Laboratory at Dalian University of Technology, and the Natural Language Processing and Social Humanities Computing Laboratory at Tsinghua

XSL•FO
RenderX

University are 3 dictionaries commonly used for sentiment analysis of Chinese texts. On the basis of sentiment dictionaries, Zhang et al [27] used unsupervised learning methods to calculate the offset between patients' comments and their sentiments and correct the original patient ratings. There is evidence that deep learning is superior in the analysis of long texts containing complex sentiments. To analyze positive and negative sentiments in patient reviews, Ye et al [31] used the binary long short-term memory method, which achieved better results than sentiment dictionary analysis. For sentiment polarity analysis in review texts, Wu and Sun [47] used the Bidirectional Encoder Representations from Transformers model, and for recommendation results, they applied the Wilson interval method. Due to the subjective nature of patient comments and the unreliability of sentiment ratings, sentiment mining methods have limitations. Data sources of uneven quality can also affect the accuracy of sentiment evaluations. The fuzzy analysis of the text can help address the uncertainty of text description [48]. The fuzzy analysis mainly applies fuzzy mathematical or fuzzy linguistic methods, which allow recommender systems to express uncertainty and obtain personalized features from patient comments. Intuitionistic fuzzy numbers serve as effective tools for dealing with fuzzy information (ie, describing the degree of neutrality in uncertain situations). Yang et al [32] converted raw data into intuitionistic fuzzy numbers to describe uncertainty information by combining the patient's disease description with comments. Xu et al [48] examined data based on hesitant fuzzy language multi-criteria preference analysis to enhance patient preferences for physician recommendations.

Figure 3. Identifying each physician's specialty involved analyzing their historical consultation texts using medical terminology recognition and topic classification mining.

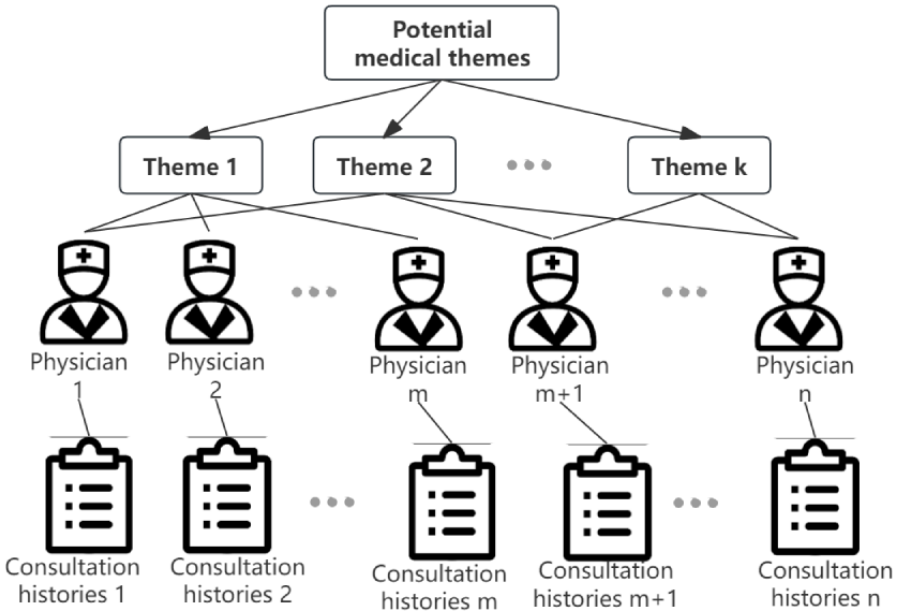

Recommendation Algorithms

Overview

OMCs recommend a service with suitable physicians according to the patients' needs, an application scenario differing from that of item recommendation in e-commerce and rather resembling expert discovery in online question and answer (Q&A) communities or academic peer review. These recommendations have one thing in common: the recommended subject is not a product but rather a human, a competent and knowledgeable professional. A physician's expertise can be inferred from their educational and professional background as well as historical consultations, similar to the history of expert responses in Q&A communities or the list of academic papers of a scholar. Patient comments and ratings for a physician are similar to the number of likes for a Q&A expert or citations of a scholar. It can be compared to assigning a competent academic reviewer to a new topic, finding a suitable expert to answer a new question, or recommending an appropriate physician based on graphic descriptions of the patient's consultation.

Knowledge-Oriented Recommendations

Knowledge-intensive service recommendations are determined by matching large amounts of textual information between patients' inquiries and physicians' skill sets. In general, the better the information match, the more likely it is that the service recommendation will be successful. As part of content-based recommendations, physicians' backgrounds and historical data are gathered, and textual topic techniques are used to mine their expertise, such as LDA, probabilistic latent semantic analysis, and so forth. Pan and Ni [46] modeled the textual topics of historical consultations and physician responses under each section, mined physician expertise using labeled LDA, and completed physician recommendations based on candidate physician expertise and pending inquiries.

Social network–based expert recommendations have grown in popularity and are derived from a classical algorithm of information retrieval (ie, PageRank). For expert recommendations, Wang et al [49] proposed a convolutional neural network for answering online expert questions that effectively reduces waiting time for the questioner and improves the quality of the answer. To alleviate the cold-start problem

XSL•FO
RenderX

for new-coming patients, physician recommendation–related studies should consider patients with similar conditions in the OHC who exchange information and provide emotional support, as illustrated in Figure 4. Recently, expert recommendation research has increasingly incorporated integrated models that combine features such as social networks and knowledge content. Xu et al [50] proposed a scholarly recommendation framework that integrates social network analysis and conceptual semantic analysis in 2 dimensions: social relationships among scholars and information about their expertise. Yang et al [51] used information about research relevance, personal social networks, and institutional connections to identify the most appropriate experts for collaboration on research. Xu et al [52] proposed a methodology for a collaborative recommendation that integrates expert expertise and social information in a complex heterogeneous network using heterogeneous network mining. It identifies valuable meta-paths through information gain and uses regularized optimization to generate personalized recommendations tailored to each scholar's needs. Different recommendation algorithms have different strengths in comparison. Expert recommendations based on knowledge content are better suited for use in enterprises with high levels of information quality and clearly defined knowledge hierarchies. Information quality in OHCs is significantly lower than that in general organizations, and expert recommendations are greatly influenced by the structure of social networks [53]. Both of these features are present in the OMC service recommendations studied in this review.

A KG is a structured semantic knowledge base that integrates heterogeneous information from multiple sources and represents rich entity relationships using complex networks, which facilitates the storage, processing, and communication of complex real-world knowledge. Medicine is a specialized scientific field, and vector representations of KGs enable algorithms to obtain embeddings of concepts, class hierarchies, entities, and relationships and, in turn, graph structures, paths, and subgraphs. Algorithms can achieve logical reasoning in vector space with the help of ontology embedding and rule learning. For the OHC platform to be credible, physicians must provide their real names, educational backgrounds, professional experience, and expertise so that their profiles can be verified. Using document clustering analysis, LDA topic segmentation, and feature extraction from physician historical consultations, a KG describing physician specialty and expertise can be constructed. Yuan and Deng [4] produced a more accurate and interpretable recommendation scheme based on the KG to overcome the problem of sparse data. It is common practice for existing studies to extract entities based on physician historical consultations; however, these data alone are not sufficient to represent physician professional specialties. For example, if an otolaryngologist has only received consultations related to the ear and nose for various reasons, then the system only measures their expertise in the ear and nose. However, in practice, they also have excellent expertise in laryngology, which the system cannot calculate. An appropriate recommendation system should be designed to recognize the differences between specific diseases and the expertise of various physicians within the same department. As shown in Figure 5, the original scope of historical consultations should be extended to include new entity nodes such as specialized disciplines, physicians, and consultations. To optimize the network structure of the KG, we should analyze the semantic connotation of keywords, determine the semantic similarity between consulting cases and their attribution to specialized disciplines, and examine the professional areas of physicians and their evolution trends.

KG-based physician recommendations are a new trend in OMC service recommendations. By using logistic regression, plain Bayesian classification, and noise-immune gate Bayesian networks, Rotmensch et al [54] constructed a KG, and from the parameter training, a disease-symptom topological relationship graph was generated. Liu [45] used the k-means algorithm to cluster physicians and generalized goodness-of-fit metrics to evaluate and adjust the clustering results. By comparing the patient's consultation content with the physician clustering center and the individual physician information in each category, a physician category and physician object that are more closely matched could be recommended. Xu et al [52] proposed a collaborative recommendation method for scholars based on heterogeneous network mining, combining expert expertise with social information, identifying valuable meta-paths through information gain, and providing personalized recommendations for each scholar through canonical optimization. On the basis of the similarity of consultation texts, Meng and Xiong [35] constructed a co-occurrence label network of physicians and calculated the centrality of the feature vector to recommend the most important physicians. Gong et al [19] proposed a hybrid multilayer architecture, iBole, of physician recommendations, mining physician-patient relationships using a time-constrained probabilistic factor graph model and recommending physicians based on random wandering. KG-based physician recommendations also have drawbacks. The OMC service faces more complicated application scenarios involving multiple entities and interentity relationships that reflect a physician's knowledge or disease-symptom connection. It is difficult to integrate different attributes and relationships between attributes in traditional recommendation methods, and it is nearly impossible to visualize the relationship between each knowledge attribute and physicians. KG-based OMC service recommendations should use multisource heterogeneous information to mine physicians' comprehensive expertise, take their profiles as basic professional descriptions, mine all their published articles using text semantics, and then combine their historical consultations with multimodal data to extract features using multimodal mining and LDA topic segmentation.

XSL•FO
RenderX

Figure 4. Social network–based recommendations leverage patient-friend relationships to recommend a physician to a new patient with a similar condition, aiming to gain the trust of the new patient.

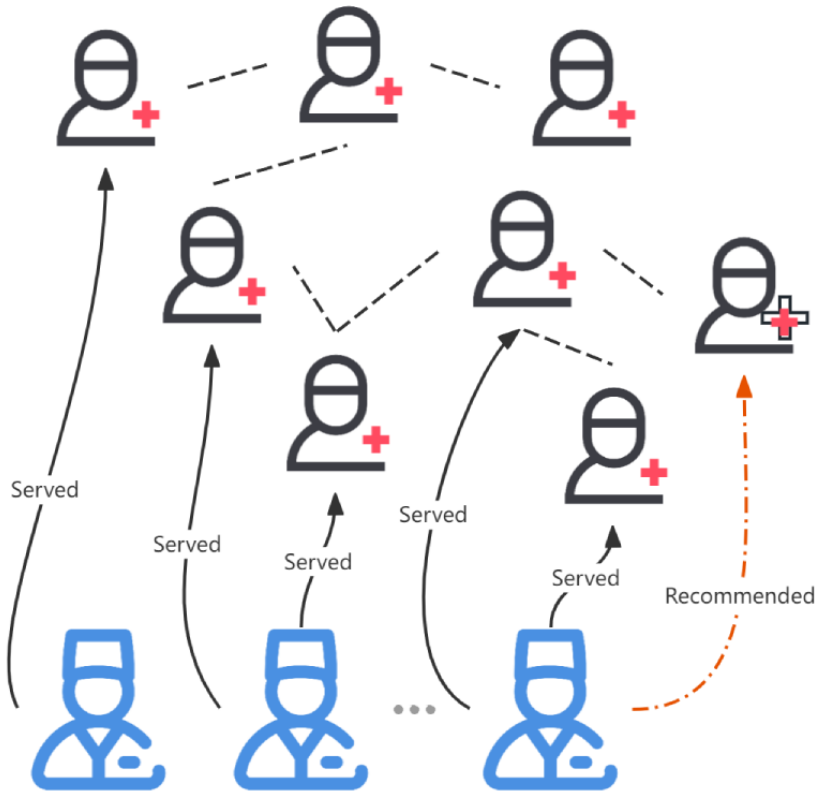

Figure 5. Constructing a physicians' knowledge graph—an example in cardiovascular medicine. The knowledge graph comprises various medical entity nodes, such as specialized disciplines, diseases, symptoms, and pharmaceuticals. These entities are extracted from the historical consultation records of physicians. ECG: electrocardiogram; PIP: medical problem indicates or reveals aspects of another medical problem; TeCP: test given to investigate a medical problem; TeRP: test reveals a medical problem; TrAP: treatment administered for medical problem, but outcome is not mentioned in the sentence; TrIP: treatment improves or cures medical problem.

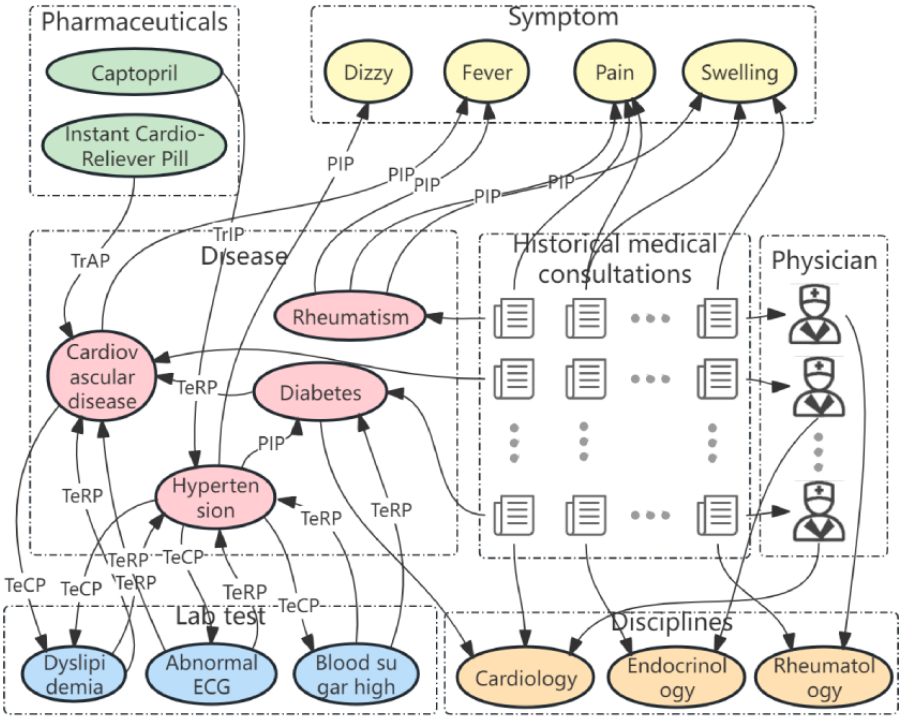

XSL•FO
RenderX

Interpretable Recommendations

As medicine is such a specialized field of science, recommendations must be interpreted according to the patients' cognitive capacity. It is difficult for patients to make autonomous judgments about the recommendations using their knowledge because they lack theories and relevant experience. Most of the existing research on recommendation systems is devoted to the professional accuracy of recommendation results. It casually ignores the interpretability of recommendation schemes and the lack of transparency in the system computation process [4]. In other words, the recommendation process and logic are not adequately explained to patients considering their cognitive capabilities. It is very critical for the recommendation system to be interpretable as it directly correlates with the level of trust of patients [55]. To provide patients with a reference for decision-making, we believe that a good recommendation system for OMC services must incorporate an interpretable and user-friendly recommendation algorithm. As a result, patient acceptance and recognition of the recommendation results will be enhanced, which will ultimately result in a higher acceptance rate of the recommended solution of the system. As a result of their limited cognitive abilities, many patients, in addition to not judging the recommendations, struggle to make their inquiries clear and complete, and in a few cases, they even cannot accurately articulate their personal needs. As an alternative to solving such difficult problems, multimodal data mining techniques may be considered, such as multimodal graphical topic modeling for patient description and consultation needs. Not only can key information from patient consultations be explored and labels can be extracted, but it is also possible to avoid creating too sparse input text variables by avoiding personalized verbal expressions and symptoms. Machine learning algorithms can easily process clustered documents when they are converted into vector distributions.

Recommendation algorithms can be interpreted in light of the rich semantic connections between physicians and patients in the KG [4]. Some studies have demonstrated that interpretable recommendation algorithms based on KGs enhance the level of patient trust. Using KG-based disease diagnosis algorithms, Wu and Sun [47] obtained initial disease alternative sets by querying the KG and using the KG embedding model; the KG-structured information was used to enrich the disease alternative set, enhancing the recommendation accuracy and facilitating the recommendation of potential diseases to the user. To identify the different roles of physician-patient interaction characteristics and individual physician characteristics in physician recommendations, Yuan and Deng [4] developed a deep learning model that can provide accurate and interpretable physician recommendation information by combining layer-by-layer association propagation techniques with deep neural networks. Considering the accuracy, diversity, and interpretability of KG-based recommendations resulting from information such as rich semantic relationships and item links within a network, we propose that interpretable recommendations should be built based on KG path inferences. The algorithm should adopt a knowledge-aware path recurrent network model, which generates path representations by combining the semantics of entities and relations, reasoning by using sequential dependency in paths to infer interaction between users and items, and incorporating a weighted pool into the process of inferring user preferences to differentiate between different contributions from different paths to provide interpretable recommendations.

Evaluation

Physician recommendations can be evaluated online or offline. Online evaluation involves measuring the effectiveness of the recommendation system by obtaining the target users' evaluation of the recommended object, namely, the rating of the recommended physician by patients. Guo et al [26] asked 3 faculty members and 3 graduate students with medical backgrounds to judge candidate physicians based on their perception of their professional activities and reputation and use the mean of the ratings to rank them. Ye et al [31] recruited 18 students with experience in helping relatives choose a physician to consult on the web and asked them to assess the relevance of the physician in response to a given consultation question. Wu and Sun [47] used a questionnaire to assess the accuracy of a physician's recommendation and validate the proposed recommendation algorithm, including whether the respondents had had a particular disease, had been treated in the area, and had approved of the physician. An online evaluation has several shortcomings, including a high implementation cost and the difficulty of excluding the characteristics of the group surveyed as well as personal subjective factors from the results. An offline evaluation involves feeding training set data into the system for training the recommendation model and calculating the recommendation results based on test data to measure the performance of the recommendation system. In most cases, machine learning models are trained through supervised learning, which means that the predicted output of the recommendation model is compared with the true value, and based on the difference, model training methods can be altered and parameters can be adjusted to facilitate the continuous optimization of the model [53]. There are different measurement criteria for the difference between the predicted output of the model and the true value. Offline evaluations are predominantly based on accuracy, which includes classification accuracy, prediction accuracy, and ranking accuracy.

The diversity and coverage of recommended physicians have also been used to evaluate the performance of recommendation algorithms. According to the literature [32], recommending only similar physicians results in a limited choice for patients and an imbalance in physician use. Patients will be more likely to engage with the recommender system if there is more diversity of recommended physicians. A measure of coverage refers to the proportion of recommended physicians to all physicians [56]. A low level of coverage indicates that a limited number of physicians are available to patients. Patients are likely to be less satisfied with a recommender system if the candidate pool is limited. However, diversity and coverage metrics are not currently heavily used for evaluating physician recommendation systems. Physician recommendations differ from traditional e-commerce recommendations in some respects. Patients should be recommended physicians with similar expertise or experience that matches their disease conditions

XSL•FO
RenderX

rather than a greater variety and number of physicians. Increasing the diversity and coverage of physician recommendations is unfavorable to patient outcomes, thereby affecting the application of these 2 metrics in physician recommendations.

Discussion

Principal Findings

Personalized recommendation studies have previously focused on commodity recommendations based on "users versus items" and rarely considered service recommendations based on "users versus users." This paper focused on human carriers who deliver OMC services, particularly when recommending professional services. The OMC service represents a new form of e-business under the knowledge economy as well as a new direction for the development of e-services. Figure 6 shows that a knowledge service–oriented recommendation differs from a traditional commodity-oriented recommendation from a system thinking perspective.

An independent service-oriented recommendation system requires a novel theoretical framework and its key techniques. Table 4 illustrates the comparison between e-commerce and e-service recommendations. First, earlier studies only considered the interests and preferences of the user, not the feelings of the providers recommended; the adoption of an OMC recommendation depends not only on the opinion of the consumer but also on the preference of the service provider. It is impossible to achieve even the so-called "best" recommendation scheme by focusing only on the needs of consumers and ignoring the individual preferences of service providers. Moreover, as the physician is more aware than the patient, they should have a higher priority in terms of decision-making [57]. Existing personalized recommendation systems have obvious flaws and weaknesses both theoretically and algorithmically even when designed specifically for consulting services. Although the recommended subjects in some expert recommendation system research, such as thesis review, project approval, and other scenarios, are also humans, the recommendation algorithm still focuses on the personalized characteristics of the demand side, analyzing only the professional skills of the experts rather than considering their preferences. These experts are just "tool men." In the case of e-service recommendation applications such as OMCs, such a research perspective and research conclusions are not applicable. Due to the existence of intrinsic and extrinsic needs of 2-sided users, it is apparent that a new paradigm of personalized recommendation research must be based on a service-oriented approach.

The professional characteristics of the service require that the system provide consumers with explainable recommendations according to their cognitive levels. Medical diagnosis and treatment is a very specialized field. Most patients do not have a very clear understanding of it. The model should be capable of explaining the recommendation schemes so that patients can make informed decisions [4]. In the case of e-commerce–oriented recommendations, interpretability is not required as users understand the utility of the items and what they desire. Thus, the system simply needs to fully exploit the hidden needs and interests of users. Algorithms focus primarily on collecting users' side information to identify their potential needs and respond to their individualized preferences [58]. Due to the consideration of medical privacy in the OMC scenario, the system is unable to extract patients' hidden medical histories or other information from their historical treatment records [30]. Furthermore, patients generally lack medical knowledge and are unable to make independent judgments about the recommended results. Having interpretable algorithms improves not only the transparency of the recommendations but also the trust and acceptance of patients, which improves postevent satisfaction with physicians [4].

OMCs' particularity is also reflected in its knowledge- and labor-intensive nature. OMCs are professional consultations and intellectually demanding services that involve bilateral interactions between physicians and patients [59], so physician workload must be carefully considered. Traditional e-commerce–oriented recommendation algorithms typically produce "popular products" or "superstars," which do not consider the overwork of physicians. In reality, it is impossible to achieve an overloaded recommender scheme regardless of how well the patient's condition matches the physician's specialty. A few studies have addressed the "diversity" or "coverage" of recommendations; however, they only increase the total number of item types without considering the frequency of recommendations for a single item. Whenever a human-based service recommendation system is used, the workload problem must be considered, yet it has rarely been taken into account in previous studies.

Data about users are not always valuable. Whether user reviews contribute to the formulation of recommendations is also a difference between OMC scenarios and those of other applications. Several previous studies have attempted to obtain useful information from patient reviews, but these efforts have proven unsuccessful [60,61]. In general, patients are attracted to "popular" physicians with many positive reviews and few moderate and poor reviews, whereas young or unknown physicians are underrepresented, with few respondents and a lack of adequate review data. In total, 3 factors contribute to this phenomenon: patients are unprofessional, physicians are uncooperative, and evaluation of services is difficult. The first challenge is that patients are incapable of evaluating the effectiveness of professional services, and no significant correlation has been found between the online reviews of patients and their clinical outcomes [3]. Second, physicians will vigorously resist unprofessional, emotional, and malicious reviews that can harm their professional reputation [62] and may even "vote with their feet" to force the platform to block complaints. Third, the success of OMC services is dependent not only on physicians' professionalism but also on patients' perceptions and expectations. In addition, the patient's experience also depends on whether the medical institution where the doctor works can provide advanced medical equipment and a convenient medical environment. Even the ease of use, stability, and privacy security of OHC platforms may have an impact on patients' evaluations [63]. Until it has been established what techniques and methods are being used

XSL•FO
RenderX

to extract key elements from subjective, ambiguous, and complex patient reviews, e-service–oriented personalized recommendation systems should be cautious about using comments and ratings.

Figure 6. The distinction between commerce-oriented and service-oriented recommendations. Commerce-oriented recommendations focus solely on user preferences, disregarding item recommendation limits, whereas service-oriented recommendations must consider the needs and preferences of both parties involved as well as the providers' capacity constraints. (A) e-Commerce oriented personalized recommendation system; (B) e-Service oriented personalized recommendation system.

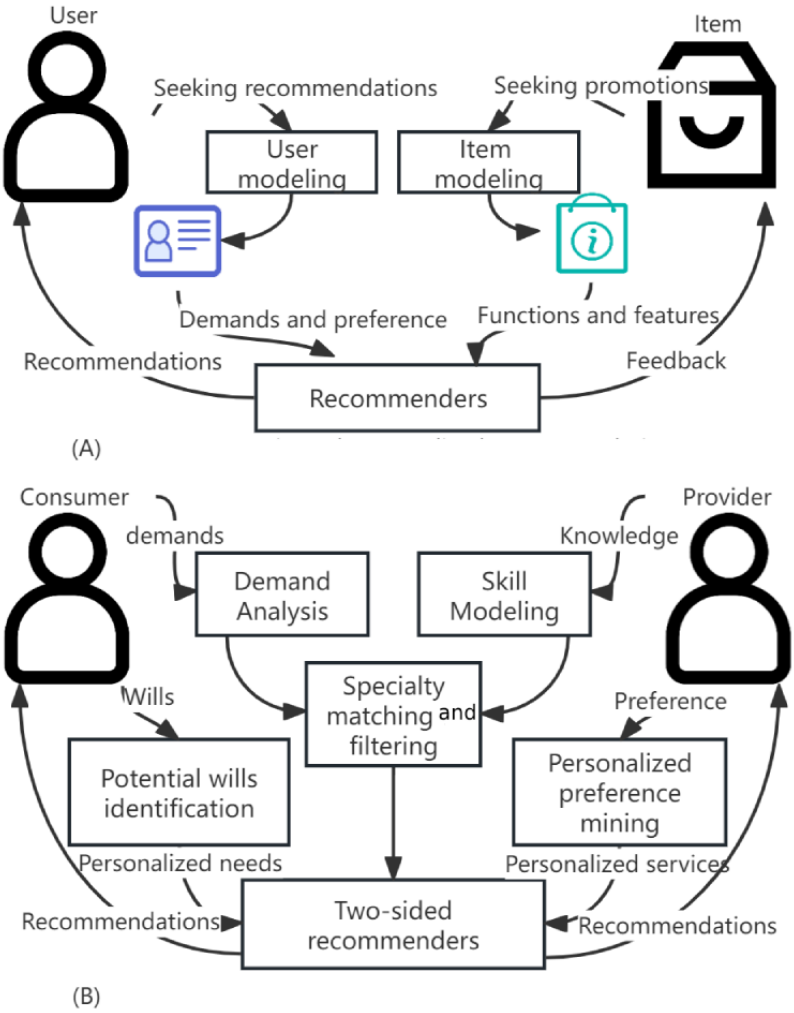

Table 4. Comparison of commodity-oriented versus service-oriented personalized recommendations.

	e-Commerce oriented	e-Service oriented
Components	Users vs commodities	Users vs users
Recommended items	Commodities	Services
Decision makers	Only users	2-sided users
Personal preferences	Only users	Both patients and physicians
Workload	__[a]	Physicians
Reviews and ratings	Important features	Useful but needs caution
Interpretability	Optional	Required

[a]Not applicable.

2-Sided Preferences

Overview

Personalized recommendations are based on user preferences, and acquiring accurate user preferences is key to ensuring their quality [29,32]. In contrast to other recommendations, OMC recommendations need to consider the preferences of both consumers and providers as an OHC is a 2-sided market constituted by both patients and physicians, each with independent and stable preferences. Physician preferences have

XSL•FO
RenderX

regrettably been ignored in previous recommendation systems, which has resulted in infeasible recommendations. Malgonde et al [5] proposed a 2-sided recommendation framework for digital platforms to mitigate user emergence as a commercially complex adaptive system with differential and evolving goals, preferences, and constraints for both sides of a 2-sided market. The patients' personal preferences influence their selection behavior and, thus, their satisfaction with the recommendations [23,33,38]. In turn, the physicians' preferences influence their willingness to receive consultations, and in turn, the physicians' onboarding and retention determine the continuity and development of the OHC [64]. Due to the differences in scale and quality of data between the 2 types of users, patients and physicians should have independent approaches to the extraction of features and the mining of behavioral patterns.

Patient Preferences

Overview

Patients' preferences and needs have been relatively adequately explored in existing studies on physician recommendations. As shown in Table 5, when choosing a physician, patients typically consider the physician's disciplinary background, professional competence, and institutional reputation, as well as other factors such as distance, cost, and follow-up care. To provide patients with personalized recommendations, Pan et al [29] proposed a user preference–learning algorithm to learn patient preferences. Jiang and Xu [24] proposed an integrated recommendation method that uses hierarchical analysis to screen candidate physicians based on 3 dimensions: semantic matching of physician-patient professional texts, objective evaluation of physician authority, and subjective evaluation of physician online word of mouth. Ye et al [65] used SPSS to screen patient decision factors and recommend physicians based on their composite scores. Wang et al [37] even directly used the number of visits as an important determining factor for how patients viewed the standard of care provided by physicians. Xu et al [30] investigated the privacy issues of patients and provided a multi-indicator group decision–ranking system of physicians.

XSL•FO
RenderX

Table 5. Various factors influencing patients' choice of physician. The existing literature examines expertise, reputation, communication skills, location convenience, appointment availability, insurance acceptance, cost, recommendations, online reviews, and cultural and language preferences.

Study	Reputation						Service (experience)			Affordability (costs)			Others (cares)	
	Affiliation		Reputation		Word of mouth									
	Education	Organization	Position or title	Achievements	Online ratings	User evaluation	Expertise	Practices	Histories	Distance	Expenses	Following costs	Privacy	Discrimination
Jiang and Xu [24]	✓	✓	✓	✓	✓	✓	✓		✓					
Liu et al [66]		✓	✓		✓									
Deng et al [67]			✓			✓		✓						
Li et al [68]				✓	✓	✓		✓						
Li et al [69]		✓	✓		✓			✓						
Li and Hubner [70]					✓									
Xu et al [30]				✓		✓			✓					
Xu et al [52]							✓		✓				✓	
Gong et al [71]		✓	✓		✓	✓		✓						✓
Ju and Zhang [38]					✓				✓	✓				
Wang et al [37]			✓				✓	✓	✓			✓		
Yuan and Deng [4]		✓				✓	✓		✓				✓	

Reputation

"Worshipping famous physicians" has become a very common phenomenon among patients. No matter the severity of the patient's disease, most patients prefer senior physicians from bigger institutions and more reputable practices [37]. The reputation of a physician is one of their most valuable attributes and plays an important role in patients' decision-making process [67]. Generally, physician reputation can be divided into 2 categories: offline and online reputation [66]. The former is determined by the hospital's rank, academic title, professional level, the number of years in the field, and the popularity of the physician, and the latter depends on patient evaluations and ratings as well as the number of votes received, acknowledgment letters, online gifts, and other factors. Patients' cult of famous physicians is largely based on physicians' offline reputations. Liu et al [66] found that the ranking of the hospital and the title of the physician had a direct impact on patients' choices. The higher the title and ranking, the more popular the individual was. Deng et al [67] also concluded that the title of the physician had a significant impact on the choice of the patient. Patients favored the chief or deputy chief physician over the regular resident physician. In addition, offline reputation can moderate the impact of online reviews on patient choice. Li et al [69] demonstrated that hospital rank and physician professional credentials negatively moderate the effect of physician online ratings and activity on patient choice. Huang et al [72] revealed that a physician's high title negatively moderated the effect on physician service ratings while positively moderating the number of service reviews. Word of mouth in OHCs determines physicians' online reputation. The experiences of previous

XSL•FO
RenderX

patients, reviews, and recommendations are important decision-making aids for newcomers. Deng et al [67] revealed that the number of views and votes received on physicians' home pages positively influenced patients' choice of physician. Gong et al [71] examined the impact of online reviews and online ratings of physicians on patient decisions from the perspective of trust theory. Li et al [69] found that positive physician reviews were positively related to a patient's choice of physician, whereas negative physician reviews played the opposite role, and that negative reviews had a greater impact on a patient's choice of a physician than positive reviews. Li and Hubner [70] demonstrated that patients preferred physicians with higher technical skills over those with higher interpersonal skills based on the different dimensions of physician ratings.

Serviceability

With regard to social exchange theory, physicians' participation in OHCs is a social exchange behavior, and services such as publishing scientific articles, providing OMC services, and offering appointment registration can bring physicians financial and social rewards [6,72]. The quality of a physician's services is reflected in patients' online ratings and postevaluations, which in turn influence the decision to choose a physician made by potential patients in the future. Physician service quality in OHCs can be measured by the level of platform activity, engagement, responsiveness, and frequency of updating popular articles. Deng et al [67] asserted that physicians' behaviors, such as regular updating of medical information, publication of scientific articles, and answering patients' questions, can enhance their community reputation, which in turn can attract more patients. Gong et al [71] noted that updating physicians' information frequently and providing quality online services were critical to building trust between physicians and patients. Using the number of physician publications of popular articles in OHCs, Li et al [69] found that physician activeness was positively associated with patient selection.

Affordability

It is also important for patients to consider the time and financial expense of visiting their physician when selecting a physician, preferring an appointment time and location that is convenient for them as well as cost-effective treatment options [29,32]. One of the factors that patients consider when choosing a physician is the location of the physician. Typically, patients consult on the web before consulting offline, and the location of the OMC-receiving physician is related to the convenience of future offline consultations. Ju and Zhang [38] considered the location of the patient to improve the convenience of combining online consultation with offline treatment. Deveugele et al [73] analyzed questionnaire data from 6 European countries and studied video recordings of consultations and found that the location of a physician's hospital affected the length of the online consultation. Compared to geographical location, consultation costs have relatively little impact on patients' choice of OMC services. Khairat et al [74] reported that costs were one of the primary factors determining patients' choice between mobile health and telemedicine. Fletcher et al [75] also argued that the cost of providing mental health treatment via video at home was significantly lower than the cost of providing in-person care assuming that patients can make use of existing personal technology.

Others

The personal characteristics of a physician, such as their appearance and gender, can also influence patients' choices. Ouyang and Wang [76] found that a serious and stable physician appearance image contributes to patients' trust in physicians, which in turn influences their medical choices. In addition, patients have some stereotypes about physicians' gender. The gender difference in physicians also extends to the distinction between different departments and medical specialties. Bertakis [77] found that male and female physicians practice in different ways, with female physicians providing more psychological counseling and preventive services and male physicians focusing more on technical practices such as physical examination. A physician's gender also influences patient choice. Gong et al [71] found that physician gender influenced physician ratings and patient choice and that patient choice was enhanced when the physician was male.

Physician Preferences

Overview

Continual physician involvement is crucial to the survival, growth, and prosperity of OHCs [64]. Although patients are consumers of OHCs and physicians are merely providers, the latter are of greater significance and influence. Patients who participate in OHCs seek out famous physicians, and existing OHCs are essentially physician-driven organizations [57]. In comparison with their counterparts, physicians possess a higher level of cognition and more logical behavior. There is a relatively large amount of data on physicians in current OHCs. By mining behavioral data, it is possible to gain a better understanding of their motivations and expectations. Unfortunately, most previous studies have been primarily concerned with physicians' fitness from a professional perspective rather than with their willingness and preferences from a drive and reward perspective. Current paradigms of research, which ignore the individualized preferences of the recommended population, are not adequate to meet the growing need for human-based, knowledge-based service recommendations. According to physician motivation theory, we propose a research paradigm to examine how perceptions of personal benefits and costs, satisfaction with individual needs, and cultural differences influence physicians' OMC decisions. Few studies have examined physicians' preferences, and more have discussed physicians' motivation to participate, which influences physicians' performance in OHCs. Physicians who join OHCs and provide OMC services face both costs and rewards [6]. A rational decision is based on weighing the costs and benefits. Physicians incur cognitive costs, which include fatigue, pain, and irritability generated by providing knowledge- and labor-intensive services, and implementation costs, which include time, material, and financial costs. Physicians receive a variety of rewards, including both social and economic rewards. The former describes that a physician is respected and valued by their patients for the services they provide in OHCs as well as for fulfilling their own needs and realizing their self-worth, and the latter represents that a physician receives

XSL•FO
RenderX

both direct financial gains from OMCs as well as virtual gifts and bonuses from their patients. Financial and social rewards are significant factors influencing physicians' engagement in OHCs and OMCs. Physicians' expectations also play a role in the extent of their influence. Figure 7 illustrates how data mining of physicians who participate in OHCs and determining their motivation to participate in OMCs can be carried out. Data collected include but are not limited to academic titles, educational background, career experience, scientific research accomplishments, and case characteristics associated with their historical consultations. The objective of mining these data is to develop a multidimensional preference index system for material motivation, career motivation, and social capital motivation. This will enable us to improve the adoption rate of recommendations and promote a personalized physician recommendation system.

Figure 7. Mining physician motivations and personalized preferences. The collected data, encompassing academic titles, educational background, career experience, research accomplishments, and consultation case characteristics, inform a multidimensional preference index system. This system addresses material, career, and social capital motivations, enhancing recommendation adoption rates and personalized physician recommendations. OHC: online health community.

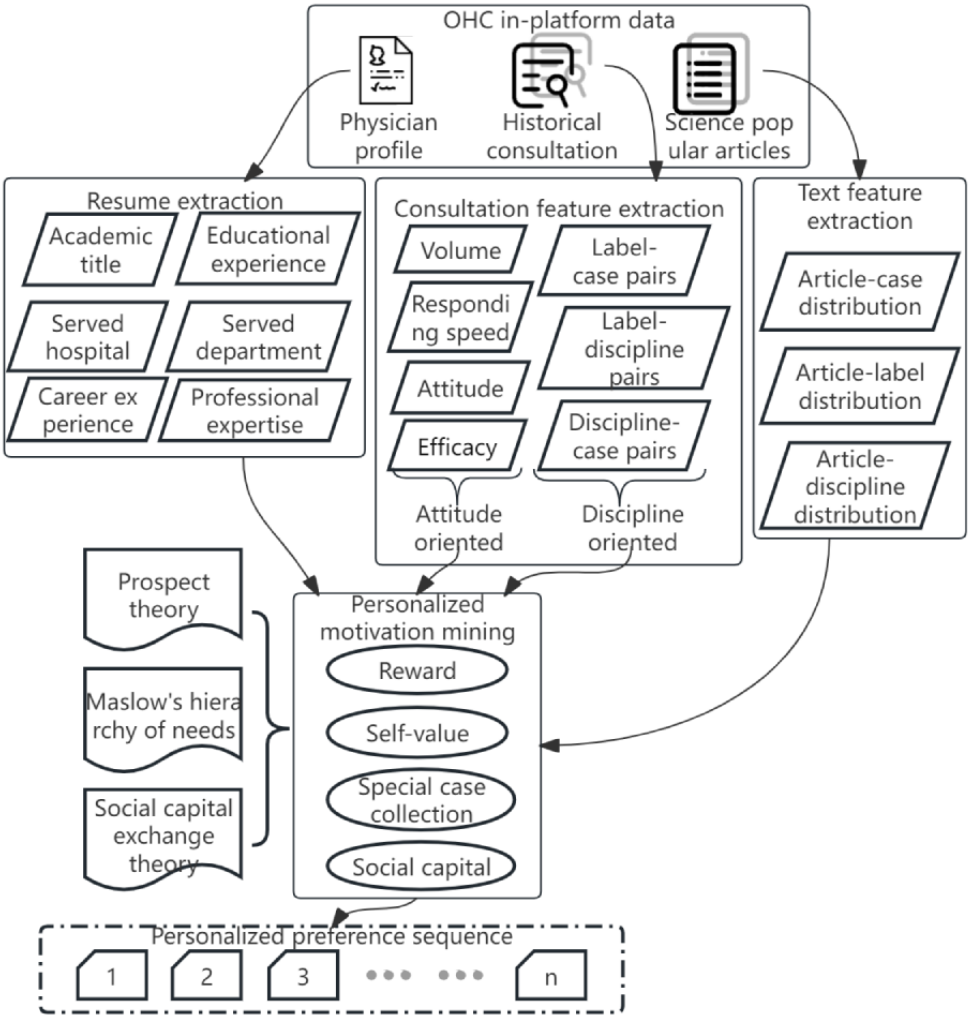

Motivation

Physicians' motivations for joining OHCs are remarkable in their diversity. Physicians are concerned not only with financial rewards but also with career planning, professional reputation, and social capital. These considerations include the need for self-worth realization, prestige, social support, and personal branding [64,78]. The needs theory by Maslow [79] suggests that prestige contributes to self-realization. Social exchange theory also reveals that self-realization, prestige, and social support positively influence physicians' willingness to provide online services, whereas executive costs negatively impact their willingness to do so. Using expectancy theory, Chen et al [64] found that both external motivation (eg, external rewards and expected relationships) and intrinsic motivation (ie, a sense of self-worth) positively influenced physicians' willingness to provide consultation services, whereas consultation time, as a major cost, negatively moderated the relationship between physicians' willingness to serve and their behavior. Zhou et al [80] combined mental health–related OHCs with motivation theory and demonstrated that both intrinsic (technical competence) and extrinsic (network reputation and financial rewards) motivations positively influenced psychologists' voluntary behaviors. Yang et al [81] suggested that physicians' contributions to OHCs were positively influenced by both personal and social motivations and physicians' professional titles moderated this effect, with physicians with high titles emphasizing reputation and physicians with low titles

XSL•FO
RenderX

emphasizing monetary rewards. Zhang et al [3] found that, when physicians reach an advanced level of expertise and knowledge, their material motivation declines and their professional motivation increases. Some physicians place great emphasis on personal branding, and their online services are designed to support their brand positioning and identity. Zhang et al [82] indicated that the OHC environment impacts brand performance, including trust and reputation, which become more significant factors in determining whether physicians participate in a consultation.

Economic Returns

Most physicians provide OMC services for financial reasons. OHCs need to understand how to improve financial rewards for physicians to retain good physicians. Ren and Ma [17] investigated the factors influencing physicians' economic income in OHCs in the context of the pandemic. They found that service quality had a significant positive effect on physicians' economic returns. In addition, they found that physician teams increase income with disease privacy and physicians who established a team were more likely to earn more money. On OHCs, physicians share articles about health and medicine as well as providing paid OMC services. According to the literature [3], physicians share free messages due to both material and professional motivation, with the role of material motivation diminishing as physicians gain more expertise. Zhang et al [78] reported that mutual aid and altruism can positively influence the willingness of health experts to share knowledge. In addition, reputation and self-efficacy can play a greater role than regular users in health experts' willingness to share knowledge. Yang et al [81] demonstrated that physicians are motivated to share paid messages for a variety of reasons. External motivation, enjoyment motivation, and professional motivation are all important factors.

Social Rewards

According to the literature [16], social rewards have less influence on physician motivation than financial rewards. A combination of psychological and material rewards increases physician motivation to participate in OHCs. Material rewards are usually more useful than psychological rewards, but extreme rewards are less effective than moderate rewards. To increase physician retention, OHCs often include gamification elements such as badges, points, and leaderboards. Liu et al [83] observed that including gamification elements in medical communities can encourage continued participation and increase physician incomes, but on the other hand, gamification elements can also lead to greater income disparities among physicians.

2-Sided Matching

Unlike previous studies, this paper focused on the personalized service recommendation system for 2-sided users. It is not just about providing patients with a list of physicians but also about exploring the overall combination solutions with optimal mutual benefits for both patients and physicians, shown in Figure 8. Several important issues need to be addressed by researchers in this field, including the adoption of appropriate decision methods that effectively match the interests and preferences of both physicians and patients, improve the adoption rate of recommended solutions, and enhance the satisfaction of 2-sided users [84]. Xi and Juan [84] addressed the real problem of matching the supply and demand of health care services under an intelligent platform and proposed a decision-making method that considers both provider's and consumer's expectations as well as the psychological characteristics of hesitation and uncertainty. Gao et al [85] analyzed the problem of matching decisions for medical services in OHCs and constructed a matching decision model that is both satisfactory and stable. Zhong and Bai [86] analyzed the patient-physician preference matrix and constructed a 2-way matching model for specialty outpatient appointments oriented toward satisfying patients and physicians. Yang et al [87] used the 2-sided matching theory to design a patient-specialist paired appointment system in which the appointment process and the one-to-many appointment-matching algorithm were described. Chen et al [88] developed an innovative multi-attribute decision-making method for 2-sided matching considering the psychological behaviors of matching bodies as well as values of aspiration levels and evaluations.

The future research direction of the physician-patient 2-sided matching recommendation system should take into account the decision-making environment of realistic situations. As an example, due to the complexity of medicine and the ambiguity of human thinking, most patients are unable to express clear preference sequences due to their cognitive limitations. By mining consultation text and behavioral characteristics of OHC users, the OMC recommendation system should be able to capture customized preference sequences. Even for physicians, who have higher cognitive levels, more logical behavior, and clearer motivation, there are still situations in which expectation evolution and multiple preferences cannot be ordered. Therefore, the recommendation system must accommodate their intuitive fuzzy preferences. Using intuitive fuzzy preferences, biased order relations can be expressed and preference strengths can be differentiated. Figure 9 illustrates how an intuitionistic fuzzy set matrix is transformed into a satisfaction matrix. The system should then construct a multi-objective optimized, stable 2-sided matching model based on intuitionistic fuzzy number information with the objective of maximizing physician-patient matches, stability, and satisfaction with the matching results.

XSL•FO
RenderX

Figure 8. Service-oriented recommendations aim for more than just listing physicians for patients; they seek comprehensive solutions that benefit both parties. This concept is illustrated in a physician-patient matching diagram.

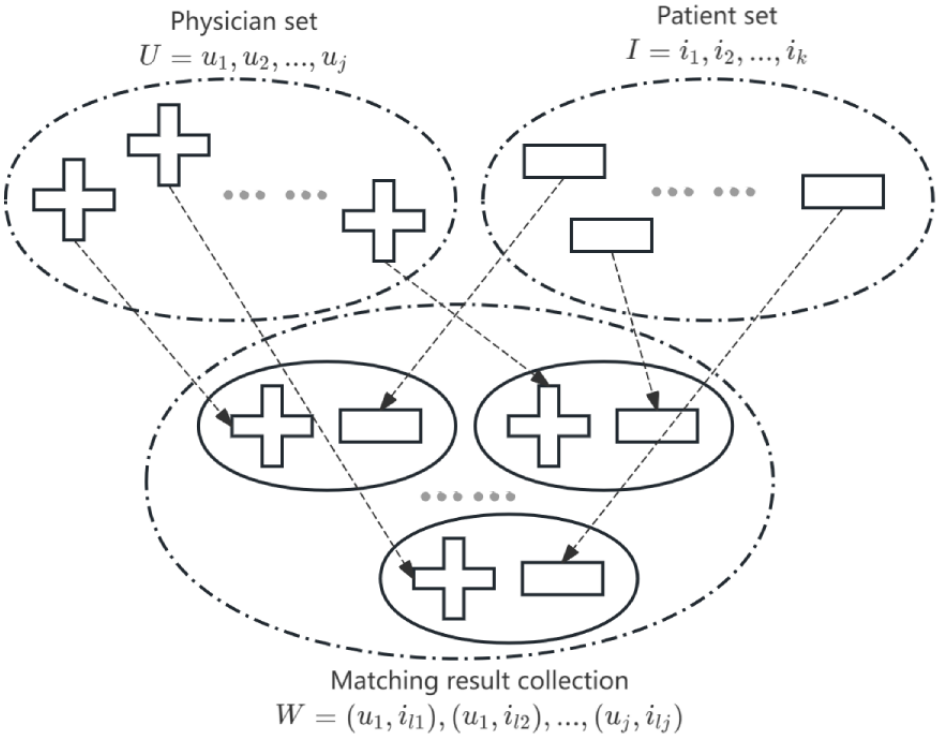

Figure 9. Modeling of physician-patient personalized preference order maximization. The recommendation system transforms an intuitionistic fuzzy set matrix into a satisfaction matrix and constructs a multi-objective matching model. It aims to maximize physician-patient matches while ensuring stability and satisfaction.

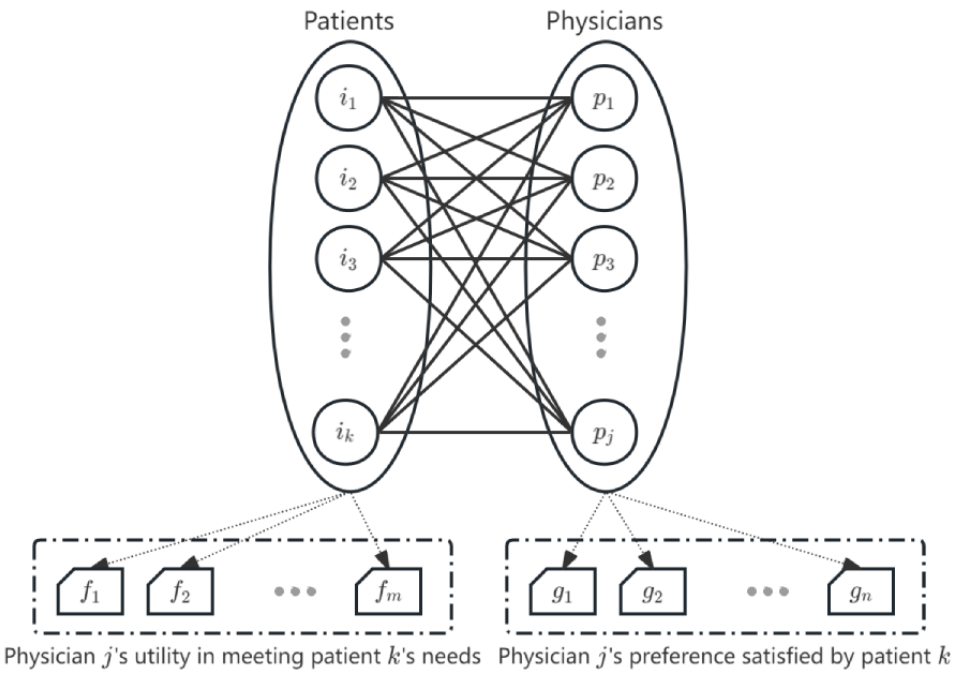

Workload Balancing

Physicians, as humans, have not only individual drivers and preferences but also variability in load tolerance. The fact that the recommended physicians represent a limited human resource has generally been overlooked in previous studies. Physicians should not overwork, and they should not be overused for an extended period [37]. Physician overload affects physician fatigue and consultative quality as well as patient waiting time, which deteriorates the comprehensive evaluation of the recommendation system [29]. Currently, very few studies have explicitly considered the workload of recommended physicians in recommender systems. To address the problem of unbalanced use among physicians, Pan et al [29] added a balanced use

XSL•FO
RenderX

approach (use balancing) to a preference-learning algorithm that included a negative penalty term for physicians whose current use exceeded the mean value. To balance patient preferences and hospital staff workload, Wang et al [37] developed a utility-diversity trade-off model based on physician capacity, patient preference, and outpatient workload, which had the effect of reducing the workload for highly regarded hospitals and physicians. Yuan and Deng [4] suggested that limiting the number of times that a physician is recommended could balance the workload while exposing more people to new physicians who could also share the workload. In addition to reducing the workload of chief physicians, Yang et al [32] increased the number of recommendations to new physicians, which translates to saving time and money for patients. The system could also be used to identify the activity of each physician's intake based on historical consultations obtained from OHCs, which, we believe, represents a difference in the upper limit of the workload of individual physicians, which is influenced by the physician's age, specialty department, and the number of offline consultations they have received.

The load balancing of OMC service recommendations is similar to personalized reranking, which generally refers to ranking items in the recommendation result list based on the user's preference. On the basis of the recommendation results list, load balancing attempts to determine the workload of each recommended physician, adjust the list order, or replace the candidate physicians according to their predefined individual thresholds so that the recommendations are achieved as efficiently as possible. The reranking algorithms typically use 2 categories of indicators. First, they integrate the reranking indicators directly into the recommendation algorithm to train a multi-objective model. Second, heuristics are used to optimize the reranking indicators using a 2-stage approach of filtering and reranking followed by optimization of the load balancing. Among the integrated algorithms, Adomavicius [56] presented heuristic neighborhood techniques and matrix decomposition techniques to generate a more diverse set of recommendations with a lower workload for each physician. Pedronette and Torres [89] proposed a method for reordering image content retrieval systems that combined recommendations with clustering and encoding context through ranking lists. Among the 2-stage algorithms, Yu et al [90] investigated the relationship between recommendation accuracy and diversity and proposed an adaptive trust-aware recommendation model to improve cold-start and long-tail items. In the literature [33], a dynamic exponential inventory-balancing algorithm for recommendations is presented based on the condition that physician resources are limited in a dynamic environment and based on real-time remaining resources. Wang et al [37] developed 2 heuristic algorithms for balancing patient preferences and hospital staff workload as well as updating physician rankings without changing physician capabilities so that patients can access more skilled physicians in more hospitals. In summary, the algorithms differ depending on the application scenario. On the basis of mining historical data, we can determine physicians' work tolerance levels; to optimize recommendation results, we can personalize constraints on physicians' upper limit of workload and dynamically optimize between patients' needs and physician energy so that the results are maximized while maintaining the quality of recommendations and reducing the workload of physicians. Using these ideas can reduce the waiting time for patients and ease the strain on physician resources.

Privacy Protection Issues

National legislation to protect user privacy in the health care sector is among the most stringent [30,80]. OMC service recommendations can only use anonymized, scrambled, encrypted, and other technically processed historical data. Consequently, it is difficult to obtain an individual identifier for each patient in the data set, which limits the algorithmic mining of patient features. Furthermore, national regulations regarding the prevention of leakage and misuse of personal information are becoming increasingly strict, and all personalized recommendation systems must and can only conduct legitimate research following user privacy protection [63]. Technically, CF models are not suitable for OMC recommendation scenarios regardless of whether they are user-based CF or term-based CF. A user is unlikely to seek help on the web unless they are ill or experiencing certain symptoms. Therefore, the specialty of physicians that patients seek is not determined by their explicit or implicit interests but rather by their medical needs at that time. The concept of "inferring future needs from patients' historical data" is not logical in the context of the OMC service scenario. Unfortunately, some existing studies continue to attempt to mine peripheral information and even private information from patients, which is both illegal and ineffective. Simply reusing CF from e-commerce recommendations and recommending physicians based on historical patient data regardless of medical privacy will ruin personalized e-service recommendations. Xu et al [30] proposed an effective and privacy-preserving medical service recommendation scheme that identifies patients' demands with physicians' information along with their reputation score, and it is considered the first study to develop a physician recommendation scheme that ensures computational efficiency. Similarly, to ensure patient privacy, Narducci et al [25] constructed a semantic recommendation system that does not link the health data entered by patients to their true identities. As user information is protected by regulations, patient consultations contain only isolated texts and graphics related to disease descriptions. Additional information is lacking, potential preferences are unclear, and invisible needs are not addressed comprehensively. As a means of achieving intelligent recommendations under privacy protection, the system must "dance with shackles on." To guide personalized preference mining, engineering psychology theories would be better applied, followed by natural semantic processing tools, topic models to refine patient descriptions, and semantic mining to quantify qualitative indicators. Patients' social networks and multimodal interaction sessions in OHCs would be better collected through this system, as well as identifying potential preferences, qualitative indicators, quantitative indicators, and perceptions of patients through natural language processing, multimodal data analysis, and heterogeneous dynamic network mining.

XSL•FO
RenderX

Contributions and Limitations

Theoretical Contributions

This review highlights a significant gap in research regarding service-oriented recommendations within OHCs. While OMCs are widely used on the internet, there is a notable scarcity of corresponding research on service recommendations within these environments. Traditionally, research on OMC recommendation systems has followed the conventional e-commerce model, focusing on recommending "items" to "users" rather than customizing e-service recommendations, such as recommending "users" to "users." This lack of focus on personalized service recommendations limits the potential for enhancing user experience within OHCs. Moreover, existing recommendation algorithms primarily focus on mining, modeling, and matching expert knowledge, neglecting the consideration of 2-sided user preferences and the workload of service providers. This oversight can result in recommendations that do not effectively cater to the needs and preferences of both service providers and consumers within OHCs.

Another crucial aspect highlighted in this review is the limited consideration given to the cognitive capabilities of service consumers. Current recommendation algorithms often fail to adequately address the issue that service consumers may lack professional cognitive capabilities. Adopting interpretable recommendation algorithms could help bridge this gap and improve the effectiveness of service recommendations within OHCs. Furthermore, this review emphasizes the importance of using consumer comments judiciously in the recommendation process. While consumer comments can provide valuable insights, they should be analyzed with caution to ensure the reliability and relevance of the recommendations generated.

In summary, research on personalized recommendations for online knowledge services within OHCs is still in its early stages, facing challenges such as the "cold start" problem and the lack of a theoretical framework or algorithm. Addressing these challenges is crucial for advancing the field and enhancing the quality of service recommendations within OHCs.

Practical Enlightenment

The practical implications of the review findings are 2-fold and can greatly benefit stakeholders within OHCs. First, the insights provided by this review can aid OHC stakeholders, including platform administrators and policy makers, in evaluating and optimizing the design of recommender systems. By understanding that service-oriented recommendation systems should function as 2-sided matching systems rather than just expertise retrieval systems, stakeholders can make informed decisions about system design and implementation. This understanding can lead to the promotion of policies that prioritize the consideration of 2-sided preferences, thereby enhancing user satisfaction and engagement within OHCs.

Second, the review findings can assist developers in prioritizing their work and implementing measures to address key challenges faced by OHCs. For instance, developers can focus on enhancing workload balancing for physicians by optimizing recommendation algorithms to consider both the workload of service providers and the preferences of service consumers. In addition, developers can implement measures to protect patient privacy while still providing personalized recommendations, thereby fostering trust and confidence among users.

Overall, the practical value of the review findings lies in their ability to guide stakeholders and developers in optimizing the design and functionality of recommender systems within OHCs, ultimately leading to improved user experiences and outcomes.

Limitations and Future Work

The primary limitation is the relatively small number of included studies, leading to less robust synthesized results. Despite a growing body of research on physician recommendations, there remains a scarcity of strictly designed OMC-oriented recommender systems. Notably, while online medical applications are widely used in China, this review excluded papers published in Chinese due to language constraints.

Conclusions

Recent years have seen an explosion of interest in physician recommendations, largely driven by the spread of OHCs and the success of artificial intelligence in other fields. As a result of the emergence of OMCs, an online service, physician recommendations have moved into a new age. These new-generation recommendation systems are service oriented rather than commodity oriented and build on the concept of 2-sided markets. This synergizes both patients and physicians with their needs and preferences individually, inspiring e-service recommendation thinking, vision, paradigms, approaches, and practices. This study has a distinctive pioneering character, and it is expected to open up a new branch of recommendation system theory. The e-service–oriented recommendations demonstrate their transformational, transdisciplinary, and translational features in terms of thinking, paradigms, methodologies, technologies, engineering, and practices. The paradigm shifts and directions are discussed in this paper. Unlike traditional e-commerce recommendations, e-service recommendations emphasize big-picture, outside-the-box thinking as well as data-driven, model-based, 2-sided hypotheses that pursue foundational and original recommendation thinking, theories, and practices from the essence of knowledge- and labor-intensive services inherent in the knowledge economy.

Acknowledgments

This project was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (grant 23XNL017). This work was supported by the National Natural Science Foundation of China (72271233, 72071203), Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), and Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101).

XSL•FO
RenderX

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HJ and WX contributed to study conception and design. HJ and ZM collected, analyzed, and interpreted the data. HJ drafted the manuscript, and WX was responsible for its critical revision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 32 KB - jmir_v26i1e46073_app1.docx]

References

1. Tan H, Zhang X, Yang Y. Satisfaction or gratitude? Exploring the disparate effects of physicians' knowledge sharing on patients' service evaluation in online medical consultations. Inf Syst J 2023 May 05;33(5):1186-1211. [doi: 10.1111/isj.12440]
2. Wang Q, Qiu L, Xu W. Informal payments and doctor engagement in an online health community: an empirical investigation using generalized synthetic control. Inf Syst Res 2023 Jun 01. [doi: 10.1287/isre.2020.475]
3. Zhang X, Guo F, Xu T, Li Y. What motivates physicians to share free health information on online health platforms? Inf Process Manag 2020 Mar;57(2):102166. [doi: 10.1016/j.ipm.2019.102166]
4. Yuan H, Deng W. Doctor recommendation on healthcare consultation platforms: an integrated framework of knowledge graph and deep learning. Internet Res 2021 Jun 08;32(2):454-476. [doi: 10.1108/intr-07-2020-0379]
5. Malgonde O, Zhang H, Padmanabhan B, Limayem M. Taming the complexity in search matching: two-sided recommender systems on digital platforms. MIS Q 2020 Jan 01;44(1):48-84. [doi: 10.25300/misq/2020/14424]
6. Guo S, Guo X, Fang Y, Vogel D. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. J Manag Inf Syst 2017 Aug 17;34(2):487-519. [doi: 10.1080/07421222.2017.1334480]
7. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health recommender systems: systematic review. J Med Internet Res 2021 Jun 29;23(6):e18035 [FREE Full text] [doi: 10.2196/18035] [Medline: 34185014]
8. Etemadi M, Bazzaz Abkenar S, Ahmadzadeh A, Haghi Kashani M, Asghari P, Akbari M, et al. A systematic review of healthcare recommender systems: open issues, challenges, and techniques. Expert Syst Appl 2023 Mar;213:118823. [doi: 10.1016/j.eswa.2022.118823]
9. Sun Y, Zhou J, Ji MM, Pei LS, Wang ZW. development and evaluation of health recommender systems: systematic scoping review and evidence mapping. J Med Internet Res 2023 Jan 19;25:e38184 [FREE Full text] [doi: 10.2196/38184] [Medline: 36656630]
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
11. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
12. Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, GRADE Working Group. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: introduction. BMJ 2016 Jun 28;353:i2016. [doi: 10.1136/bmj.i2016] [Medline: 27353417]
13. Moberg J, Oxman AD, Rosenbaum S, Schünemann HJ, Guyatt G, Flottorp S, GRADE Working Group. The GRADE Evidence to Decision (EtD) framework for health system and public health decisions. Health Res Policy Syst 2018 May 29;16(1):45 [FREE Full text] [doi: 10.1186/s12961-018-0320-2] [Medline: 29843743]
14. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
15. Seidler AL, Aberoumand M, Hunter KE, Barba A, Libesman S, Williams JG, iCOMP Collaborators. Deferred cord clamping, cord milking, and immediate cord clamping at preterm birth: a systematic review and individual participant data meta-analysis. Lancet 2023 Dec 09;402(10418):2209-2222. [doi: 10.1016/S0140-6736(23)02468-6] [Medline: 37977169]
16. Wang JN, Chiu YL, Yu H, Hsu YT. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: longitudinal study. J Med Internet Res 2017 Dec 21;19(12):e427 [FREE Full text] [doi: 10.2196/jmir.9082] [Medline: 29269344]

XSL•FO
RenderX

17. Ren D, Ma B. Effectiveness of interactive tools in online health care communities: social exchange theory perspective. J Med Internet Res 2021 Mar 12;23(3):e21892 [FREE Full text] [doi: 10.2196/21892] [Medline: 33709940]
18. Huang YF, Liu P, Pan Q, Lin JS. A doctor recommendation algorithm based on doctor performances and patient preferences. In: Proceedings of the 2012 International Conference on Wavelet Active Media Technology and Information Processing. 2012 Presented at: ICWAMTIP '12; December 17-19, 2012; Chengdu, China p. 92-95 URL: https://ieeexplore.ieee.org/document/6413447 [doi: 10.1109/icwamtip.2012.6413447]
19. Gong JB, Wang LL, Sun ST, Peng SW. iBole: a hybrid multi-layer architecture for doctor recommendation in medical social networks. J Comput Sci Technol 2015 Sep 14;30(5):1073-1081. [doi: 10.1007/s11390-015-1583-5]
20. Han Q, Ji M, de Rituerto de Troya IM, Gaur M, Zejnilovic L. A hybrid recommender system for patient-doctor matchmaking in primary care. In: Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. 2018 Presented at: DSAA '18; October 1-3, 2018; Turin, Italy p. 481-490 URL: https://ieeexplore.ieee.org/document/8631410 [doi: 10.1109/dsaa.2018.00062]
21. Mondal S, Basu A, Mukherjee N. Building a trust-based doctor recommendation system on top of multilayer graph database. J Biomed Inform 2020 Oct;110:103549 [FREE Full text] [doi: 10.1016/j.jbi.2020.103549] [Medline: 32871286]
22. Valdeira F, Racković S, Danalachi V, Han Q, Soares C. Extreme multilabel classification for specialist doctor recommendation with implicit feedback and limited patient metadata. arXiv Preprint posted online August 21, 2023 [FREE Full text] [doi: 10.48550/arXiv.2308.11022]
23. Waqar M, Majeed N, Dawood H, Daud A, Aljohani NR. An adaptive doctor-recommender system. Behav Inf Technol 2019 Jun 12;38(9):959-973. [doi: 10.1080/0144929x.2019.1625441]
24. Jiang H, Xu W. How to find your appropriate doctor: an integrated recommendation framework in big data context. In: Proceedings of the 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health. 2014 Presented at: CICARE '14; December 9-12, 2014; Orlando, FL p. 154-158 URL: https://ieeexplore.ieee.org/document/7007848 [doi: 10.1109/cicare.2014.7007848]
25. Narducci F, Musto C, Polignano M, de Gemmis M, Lops P, Semeraro G. A recommender system for connecting patients to the right doctors in the healthnet social network. In: Proceedings of the 24th International Conference on World Wide Web. 2015 Presented at: WWW '15; May 18-22, 2015; Florence, Italy p. 81-82 URL: https://dl.acm.org/doi/10.1145/2740908.2742748#:~:text=The%20recommender%20system%20is%20the,data%20shared%20by%20the%20community [doi: 10.1145/2740908.2742748]
26. Guo L, Jin B, Yao C, Yang H, Huang D, Wang F. Which doctor to trust: a recommender system for identifying the right doctors. J Med Internet Res 2016 Jul 07;18(7):e186 [FREE Full text] [doi: 10.2196/jmir.6015] [Medline: 27390219]
27. Zhang Y, Chen M, Huang D, Wu D, Li Y. iDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization. Future Gener Comput Syst 2017 Jan;66:30-35. [doi: 10.1016/j.future.2015.12.001]
28. Sridevi M, Rajeshwara RR. Finding right doctors and hospitals: a personalized health recommender. In: Proceedings of 3rd International Conference on Information and Communication Technology for Competitive Strategies. 2017 Presented at: ICTCS '17; December 16-17, 2017; Udaipur, India p. 709-719 URL: https://link.springer.com/chapter/10.1007/978-981-13-0586-3_69 [doi: 10.1007/978-981-13-0586-3_69]
29. Pan X, Song J, Zhang F. Dynamic recommendation of physician assortment with patient preference learning. IEEE Trans Autom Sci Eng 2019 Jan;16(1):115-126. [doi: 10.1109/tase.2018.2839651]
30. Xu C, Wang J, Zhu L, Zhang C, Sharif K. PPMR: a privacy-preserving online medical service recommendation scheme in eHealthcare system. IEEE Internet Things J 2019 Jun;6(3):5665-5673. [doi: 10.1109/jiot.2019.2904728]
31. Ye Y, Zhao Y, Shang J, Zhang L. A hybrid IT framework for identifying high-quality physicians using big data analytics. Int J Inf Manage 2019 Aug;47:65-75. [doi: 10.1016/j.ijinfomgt.2019.01.005]
32. Yang Y, Hu J, Liu Y, Chen X. Doctor recommendation based on an intuitionistic normal cloud model considering patient preferences. Cogn Comput 2018 Dec 6;12(2):460-478. [doi: 10.1007/s12559-018-9616-3]
33. Wen H, Song J, Pan X. Physician recommendation on healthcare appointment platforms considering patient choice. IEEE Trans Automat Sci Eng 2020 Apr;17(2):886-899. [doi: 10.1109/tase.2019.2950724]
34. Yan Y, Yu G, Yan X. Online doctor recommendation with convolutional neural network and sparse inputs. Comput Intell Neurosci 2020 Oct 15;2020:8826557-8826510 [FREE Full text] [doi: 10.1155/2020/8826557] [Medline: 33123187]
35. Meng QQ, Xiong HX. A doctor recommendation based on graph computing and LDA topic model. Int J Comput Intell Syst 2021;14(1):808. [doi: 10.2991/ijcis.d.210205.002]
36. Peito J, Han Q. Incorporating domain knowledge into health recommender systems using hyperbolic embeddings. In: Proceedings of the 9th International Conference on Complex Networks and Their Applications. 2020 Presented at: COMPLEX NETWORKS '20; December 1-3, 2020; Madrid, Spain p. 130-141 URL: https://link.springer.com/chapter/10.1007/978-3-030-65351-4_11 [doi: 10.1007/978-3-030-65351-4_11]
37. Wang H, Ding S, Li Y, Li X, Zhang Y. Hierarchical physician recommendation via diversity-enhanced matrix factorization. ACM Trans Knowl Discov Data 2020 Dec 07;15(1):1-17. [doi: 10.1145/3418227]
38. Ju C, Zhang S. Doctor recommendation model based on ontology characteristics and disease text mining perspective. Biomed Res Int 2021;2021:7431199 [FREE Full text] [doi: 10.1155/2021/7431199] [Medline: 34426788]

XSL•FO
RenderX

39. Lu X, Zhang Y, Li J, Zong S. Doctor recommendation in online health forums via expertise learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022 Presented at: Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2022;1; May 22-27, 2022; Dublin, Ireland p. 1111-1123 URL: https://aclanthology.org/2022.acl-long.79.pdf [doi: 10.18653/v1/2022.acl-long.79]
40. Chen X, Wang H, Li X. Doctor recommendation under probabilistic linguistic environment considering patient's risk preference. Ann Oper Res 2022 Jul 12. [doi: 10.1007/s10479-022-04843-9]
41. Wang Y, Zhao Y, Song J, Liu H. What drives patients to choose a physician online? A study based on tree models and SHAP values. In: Proceedings of the IEEE 18th International Conference on Automation Science and Engineering. 2022 Presented at: CASE '22; August 20-24, 2022; Mexico City, Mexico p. 1676-1683 URL: https://dl.acm.org/doi/abs/10.1109/CASE49997.2022.9926467 [doi: 10.1109/case49997.2022.9926467]
42. Wu J, Zhang G, Xing Y, Liu Y, Zhang Z, Dong Y, et al. A sentiment analysis driven method based on public and personal preferences with correlated attributes to select online doctors. Appl Intell (Dordr) 2023 Feb 20;53(16):1-22 [FREE Full text] [doi: 10.1007/s10489-023-04485-9] [Medline: 36844914]
43. Zaman N, Li J. Semantics-enhanced recommendation system for social healthcare. In: Proceedings of the IEEE 28th International Conference on Advanced Information Networking and Applications. 2014 Presented at: AINA '14; May 13-16, 2014; Victoria, BC p. 765-770 URL: https://ieeexplore.ieee.org/document/6838741/similar#similar [doi: 10.1109/aina.2014.93]
44. Son Y, Choi Y. Improving matrix factorization based expert recommendation for manuscript editing services by refining user opinions with binary ratings. Appl Sci 2020 May 14;10(10):3395. [doi: 10.3390/app10103395]
45. Liu T. An application research of automatic physician matching algorithm based on online healthcare consultation records. Inf Sci Theory Appl 2018;41(06):143-148. [doi: 10.16353/j.cnki.1000-7490.2018.06.024]
46. Pan Y, Ni X. Recommending online medical experts with Labeled-LDA model. Data Anal Knowl Discov 2020;4(4):34-43 [FREE Full text] [doi: 10.11925/infotech.2096-3467.2019.0815]
47. Wu J, Sun Y. Recommendation system for medical consultation integrating knowledge graph and deep learning methods. J Front Comput Sci Technol 2021;15(8):1432. [doi: 10.3778/j.issn.1673-9418.2101029]
48. Xu Z, Chen H, He Y. A recommender system based on hesitant fuzzy linguistic information with MAPPACC approach. Stud Inform Control 2020 Jul 02;29(2):145-158. [doi: 10.24846/v29i2y202001]
49. Wang J, Sun J, Lin H, Dong H, Zhang S. Convolutional neural networks for expert recommendation in community question answering. Sci China Inf Sci 2017 Oct 13;60(11). [doi: 10.1007/s11432-016-9197-0]
50. Xu Y, Guo X, Hao J, Ma J, Lau RY, Xu W. Combining social network and semantic concept analysis for personalized academic researcher recommendation. Decis Support Syst 2012 Dec;54(1):564-573. [doi: 10.1016/j.dss.2012.08.003]
51. Yang C, Ma J, Silva T, Liu X, Hua Z. A multilevel information mining approach for expert recommendation in online scientific communities. Comput J 2014 May 09;58(9):1921-1936. [doi: 10.1093/comjnl/bxu033]
52. Xu Y, Zhou D, Ma J. Scholar-friend recommendation in online academic communities: an approach based on heterogeneous network. Decis Support Syst 2019 Apr;119:1-13. [doi: 10.1016/j.dss.2019.01.004]
53. Nikzad–Khasmakhi N, Balafar M, Feizi–Derakhshi MR. The state-of-the-art in expert recommendation systems. Eng Appl Artif Intell 2019 Jun;82:126-147. [doi: 10.1016/j.engappai.2019.03.020]
54. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
55. Khedkar S, Gandhi P, Shinde G, Subramanian V. Deep learning and explainable AI in healthcare using EHR. In: Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A, editors. Deep Learning Techniques for Biomedical and Health Informatics. Cham, Switzerland: Springer; 2020:129-148.
56. Adomavicius G, Kwon Y. Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans Knowl Data Eng 2012;24(5):896-911. [doi: 10.1111/j.1708-8240.2003.tb00281.x] [Medline: 12638772]
57. Liu QB, Liu X, Guo X. The effects of participating in a physician-driven online health community in managing chronic disease: evidence from two natural experiments. MIS Q 2020 Jan 01;44(1):391-419. [doi: 10.25300/misq/2020/15102]
58. Qian X, Feng H, Zhao G, Mei T. Personalized recommendation combining user interest and social circle. IEEE Trans Knowl Data Eng 2014 Jul;26(7):1763-1777. [doi: 10.1109/tkde.2013.168]
59. Wan Y, Peng Z, Wang Y, Zhang Y, Gao J, Ma B. Influencing factors and mechanism of doctor consultation volume on online medical consultation platforms based on physician review analysis. Internet Res 2021 Jun 24;31(6):2055-2075. [doi: 10.1108/intr-10-2020-0589]
60. Widmer RJ, Maurer MJ, Nayar VR, Aase LA, Wald JT, Kotsenas AL, et al. Online physician reviews do not reflect patient satisfaction survey responses. Mayo Clin Proc 2018 Apr;93(4):453-457 [FREE Full text] [doi: 10.1016/j.mayocp.2018.01.021] [Medline: 29622095]
61. Daskivich TJ, Houman J, Fuller G, Black JT, Kim HL, Spiegel B. Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. J Am Med Inform Assoc 2018 Apr 01;25(4):401-407 [FREE Full text] [doi: 10.1093/jamia/ocx083] [Medline: 29025145]
62. Menon AV. Do online reviews diminish physician authority? The case of cosmetic surgery in the U.S. Soc Sci Med 2017 May;181:1-8. [doi: 10.1016/j.socscimed.2017.03.046] [Medline: 28363107]

XSL•FO
RenderX

63. Lu SF, Rui H. Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. Manag Sci 2018 Jun;64(6):2557-2573. [doi: 10.1287/mnsc.2017.2741]
64. Chen J, Lan YC, Chang YW, Chang PY. Exploring doctors' willingness to provide online counseling services: the roles of motivations and costs. Int J Environ Res Public Health 2019 Dec 22;17(1):110 [FREE Full text] [doi: 10.3390/ijerph17010110] [Medline: 31877861]
65. Ye J, Xiong H, Jiang W. A physician recommendation algorithm integrating inquiries and decisions of patients. Data Anal Knowl Discov 2020;4(2/3):153-164. [doi: 10.11925/infotech.2096-3467.2019.0633]
66. Liu X, Guo X, Wu H, Wu T. The impact of individual and organizational reputation on physicians' appointments online. Int J Electron Commer 2016 Jun 20;20(4):551-577. [doi: 10.1080/10864415.2016.1171977]
67. Deng Z, Hong Z, Zhang W, Evans R, Chen Y. The effect of online effort and reputation of physicians on patients' choice: 3-wave data analysis of China's good doctor website. J Med Internet Res 2019 Mar 08;21(3):e10170 [FREE Full text] [doi: 10.2196/10170] [Medline: 30848726]
68. Li S, Lee-Won RJ, McKnight J. Effects of online physician reviews and physician gender on perceptions of physician skills and primary care physician (PCP) selection. Health Commun 2019 Oct 24;34(11):1250-1258. [doi: 10.1080/10410236.2018.1475192] [Medline: 29792519]
69. Li Y, Ma X, Song J, Yang Y, Ju X. Exploring the effects of online rating and the activeness of physicians on the number of patients in an online health community. Telemed J E Health 2019 Nov 23;25(11):1090-1098. [doi: 10.1089/tmj.2018.0192] [Medline: 30676279]
70. Li S, Hubner A. The impact of web-based ratings on patient choice of a primary care physician versus a specialist: randomized controlled experiment. J Med Internet Res 2019 Jun 28;21(6):e11188 [FREE Full text] [doi: 10.2196/11188] [Medline: 31254337]
71. Gong Y, Wang H, Xia Q, Zheng L, Shi Y. Factors that determine a patient's willingness to physician selection in online healthcare communities: a trust theory perspective. Technol Soc 2021 Feb;64:101510 [FREE Full text] [doi: 10.1016/j.techsoc.2020.101510] [Medline: 36033357]
72. Huang N, Yan Z, Yin H. Effects of online–offline service integration on e‐healthcare providers: a quasi‐natural experiment. Prod Oper Manag 2021 Aug 01;30(8):2359-2378. [doi: 10.1111/poms.13381]
73. Deveugele M, Derese A, van den Brink-Muinen A, Bensing J, De Maeseneer J. Consultation length in general practice: cross sectional study in six European countries. BMJ 2002 Aug 31;325(7362):472 [FREE Full text] [doi: 10.1136/bmj.325.7362.472] [Medline: 12202329]
74. Khairat S, Liu S, Zaman T, Edson B, Gianforcaro R. Factors determining patients' choice between mobile health and telemedicine: predictive analytics assessment. JMIR Mhealth Uhealth 2019 Jun 08;7(6):e13772 [FREE Full text] [doi: 10.2196/13772] [Medline: 31199332]
75. Fletcher TL, Hogan JB, Keegan F, Davis ML, Wassef M, Day S, et al. Recent advances in delivering mental health treatment via video to home. Curr Psychiatry Rep 2018 Jul 21;20(8):56. [doi: 10.1007/s11920-018-0922-y] [Medline: 30032337]
76. Ouyang P, Wang J. Physician's online image and patient's choice in the online health community. Internet Res 2022 Mar 14;32(6):1952-1977. [doi: 10.1108/intr-04-2021-0251]
77. Bertakis KD. The influence of gender on the doctor-patient interaction. Patient Educ Couns 2009 Sep;76(3):356-360. [doi: 10.1016/j.pec.2009.07.022] [Medline: 19647968]
78. Zhang X, Liu S, Deng Z, Chen X. Knowledge sharing motivations in online health communities: a comparative study of health professionals and normal users. Comput Human Behav 2017 Oct;75:797-810. [doi: 10.1016/j.chb.2017.06.028]
79. Maslow AH. A theory of human motivation. Psychol Rev 1943 Jul;50(4):370-396. [doi: 10.1037/H0054346]
80. Zhou J, Zuo M, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019 Oct;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
81. Yang Y, Zhu X, Song R, Zhang X, Guo F. Not just for the money? An examination of the motives behind physicians' sharing of paid health information. J Inf Sci 2021 Feb 14;49(1):145-163. [doi: 10.1177/0165551521991029]
82. Zhang T, Yan X, Wang WY, Chen Q. Unveiling physicians' personal branding strategies in online healthcare service platforms. Technol Forecast Soc Change 2021 Oct;171:120964. [doi: 10.1016/j.techfore.2021.120964]
83. Liu J, Zhang X, Meng F, Lai KH. Deploying gamification to engage physicians in an online health community: an operational paradox. Int J Prod Econ 2020 Oct;228:107847. [doi: 10.1016/j.ijpe.2020.107847]
84. Xi C, Juan W. Matching method for medical service supply and demand considering bodies? Psychological behavior based on intelligent platform. Oper Manag Res Sci 2018;27(10):125-132. [doi: 10.12005/orms.2018.0240]
85. Gao YX, Du YP, Sun BZ, Wang R. Matching method for medical service considering the personalized demand of patients. Oper Manag Res Sci 2019;28(4):17. [doi: 10.12005/orms.2019.0075]
86. Zhong L, Bai Y. Equivalence of two-sided stable matching. J Comb Optim 2018 Jun 20;36(4):1380-1387. [doi: 10.1007/s10878-018-0308-4]
87. Yang Y, Luo S, Fan J, Zhou X, Fu C, Tang G. Study on specialist outpatient matching appointment and the balance matching model. J Comb Optim 2017 Nov 21;37(1):20-39. [doi: 10.1007/s10878-017-0208-z]

XSL•FO
RenderX

88. Chen X, Wang J, Liang H, Han J. Hesitant multi-attribute two-sided matching: a perspective based on prospect theory. J Intell Fuzzy Syst 2019 Jun 11;36(6):6343-6358. [doi: 10.3233/jifs-182627]
89. Pedronette DC, Torres RD. Exploiting pairwise recommendation and clustering strategies for image re-ranking. Inf Sci 2012 Nov;207:19-34. [doi: 10.1016/j.ins.2012.04.032]
90. Yu T, Guo J, Li W, Wang HJ, Fan L. Recommendation with diversity: an adaptive trust-aware model. Decis Support Syst 2019 Aug;123:113073. [doi: 10.1016/j.dss.2019.113073]

Abbreviations

CF: collaborative filtering
GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
KG: knowledge graph
LDA: latent Dirichlet allocation
OHC: online health community
OMC: online medical consultation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
Q&A: question and answer

Edited by A Mavragani; submitted 10.03.23; peer-reviewed by M Mars, S Pesälä; comments to author 01.11.23; revised version received 22.12.23; accepted 21.05.24; published 30.07.24.

Please cite as:
Jiang H, Mi Z, Xu W
Online Medical Consultation Service–Oriented Recommendations: Systematic Review
J Med Internet Res 2024;26:e46073
URL: https://www.jmir.org/2024/1/e46073
doi:10.2196/46073
PMID:38777810

©Hongxun Jiang, Ziyue Mi, Wei Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Examining the Effectiveness of Electronic Patient-Reported Outcomes in People With Cancer: Systematic Review and Meta-Analysis

Melissa Betty Perry[1], BSc, MPH; Sally Taylor[1,2], PhD; Binish Khatoon[1,2], PhD; Amy Vercell[1], MSc; Corinne Faivre-Finn[3,4], PhD; Galina Velikova[5,6], PhD; Antonia Marsden[7], PhD; Calvin Heal[7], MSc; Janelle Yorke[1,2,8], PhD

[1]Christie Patient Centred Research, The Christie NHS Foundation Trust, Manchester, United Kingdom

[2]Division of Nursing Midwifery and Social Work, School of Health Sciences, Faculty of Biology, Medicine & Health, The University of Manchester, Manchester, United Kingdom

[3]Division of Cancer Science, The University of Manchester, Manchester, United Kingdom

[4]Clinical Oncology Department, The Christie NHS Foundation Trust, Manchester, United Kingdom

[5]Leeds Institute of Medical Research, University of Leeds, Leeds, United Kingdom

[6]St James's Institute of Oncology, St James's University Hospital, Leeds, United Kingdom

[7]Division of Population Health, Health Services Research & Primary Care, The University of Manchester, Manchester, United Kingdom

[8]School of Nursing, The Hong Kong Polytechnic University, Kowloon, China (Hong Kong)

Corresponding Author:
Janelle Yorke, PhD
School of Nursing
The Hong Kong Polytechnic University
Room GH507a
11 Yuk Choi Rd, Hung Hom
Kowloon, 852
China (Hong Kong)
Phone: 852 92570178
Email: janelle.yorke@polyu.edu.hk

Abstract

Background: Electronic patient-reported outcomes (ePROs) are commonly used in oncology clinical practice and have shown benefits for patients and health resource use.

Objective: The aim of this study was to compare the isolated effect of administering ePROs to patients with cancer versus a control condition.

Methods: The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed. Randomized controlled trials evaluating ePRO interventions that aimed to improve health-related outcomes among patients with cancer were included. The primary outcome was health-related quality of life (HRQOL), and the secondary outcomes were symptoms, hospital admissions, unplanned visits, chemotherapy completion, survival, and satisfaction with care. The effect sizes of ePROs on health-related outcomes were analyzed as standardized mean differences (SMDs) with 95% CIs using a random effects model.

Results: The search identified 10,965 papers, of which 19 (0.17%) from 15 studies were included. The meta-analysis showed an improvement in HRQOL at 3 months, measured by the Functional Assessment of Cancer Therapy–General (SMD 0.28, 95% CI –1.22 to 1.78), and at 6 months, assessed using various HRQOL measures (SMD 0.07, 95% CI –1.24 to 1.39). The results should be interpreted with caution, given the wide 95% CIs. Of the 15 studies, 9 (60%) reported a positive signal on HRQOL, with two-thirds of the studies (n=6, 67%) including tailored patient advice and two-thirds (n=6, 67%) using clinician alert systems.

Conclusions: The meta-analysis showed a potential improvement in HRQOL at 6 months and in Functional Assessment of Cancer Therapy–General scores at 3 months for studies that included tailored advice and clinician alerts, suggesting that these elements may improve ePRO effectiveness. The findings will provide guidance for future use and help health care professionals choose the most suitable ePRO features for their patients.

Trial Registration: PROSPERO CRD42020175007; https://tinyurl.com/5cwmy3j6

XSL•FO
RenderX

(J Med Internet Res 2024;26:e49089) doi:10.2196/49089

KEYWORDS

telemedicine; patient-reported outcome measure; neoplasms; quality of life; systematic review; meta-analysis; randomized controlled trial

Introduction

Background

A patient-reported outcome (PRO) has been defined as "a measurement of any aspect of a patient's health status that comes directly from the patient" [1]. PROs can be more sensitive and reliable than outcomes reported by clinicians, particularly when reporting adverse events [2,3]. Electronic PROs (ePROs) refer to an electronic administration of the PROs. ePRO systems can present summary reports of patients' health-related quality of life (HRQOL) and symptoms to the clinician in real time to enhance communication [4] and improve patient management and outcomes.

Many different types of ePRO systems have been developed to monitor and manage diseases, treatments, and symptoms [5] in a variety of patient groups. Guidelines published by the European Society for Medical Oncology recommend using ePROs for symptom monitoring in routine oncology care during systemic cancer treatment due to evidence of benefits for communication, satisfaction of care, treatment adherence, symptom control, HRQOL, hospital admissions and visits, and survival [6,7]. ePROs are being increasingly integrated as part of routine oncology clinical care in the United Kingdom, the United States, and Australia [8-11]. One study exploring the use of ePROs in clinical practice has shown that symptom and quality-of-life scores reported by patients are clinically meaningful and relevant: cough and mobility scores were lower for people with poorer performance status, and patients undergoing chemotherapy and radiotherapy showed improvements in some symptoms [12].

Previous reviews have examined the features of ePRO systems (eg, exploring the integration of ePROs into clinical practice and identifying features that may be associated with patient engagement and patient-centered outcomes [13,14]). Reviews have also been conducted to examine the impact of ePROs on quality of patient care [15].

Objectives

To our knowledge, no existing reviews have grouped randomized controlled trials (RCTs) to estimate an effect size through a meta-analysis to establish clinical benefit. Some meta-analyses evaluated the capacity of telehealth or eHealth interventions to enhance HRQOL in patients with cancer [16,17]. However, many of these studies included complex interventions with various components, not just the ePRO alone. Similarly, other reviews have not specifically explored interventions that provide results to clinicians, and they have not explored the specific components of the ePRO interventions [18,19]. The primary objective of our review was to examine the effect of administering ePROs to patients with cancer on HRQOL compared to usual care. The secondary objectives included the comparison of survival, symptoms, psychological well-being, health care use, and satisfaction with care between participants receiving the ePRO intervention and those receiving usual care.

Methods

Search Strategy

The review protocol was registered with PROSPERO (CRD42020175007). Subsequent protocol changes included the requirement that ePRO results be fed back to clinicians for review. The reporting of this review was guided by the standards of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. We used the PRISMA guidelines [20] to ensure that all relevant information detailing the processes we followed as well as the findings of this review were included within the manuscript. The completed PRISMA checklist can be found in Multimedia Appendix 1 [20]. Seven databases (PubMed, MEDLINE, PsycInfo, Cochrane Central Register of Controlled Trials, Web of Science, Embase, and CINAHL) were searched systematically on July 16, 2020 (search 1), and June 20, 2022 (search 2). Similar keywords were used across the databases, adapting Boolean operators and Medical Subject Headings vocabulary. The search terms related to the electronic dimension of the "ePRO" system, "patient reported outcomes," "cancer," and "randomized controlled trials."

Inclusion Criteria

Full-text research papers in English published in peer-reviewed journals from January 2000 to June 2022 were included in the review. Given the technological advances in this field over recent years, studies published before 2000 were excluded because the ePRO systems they described would not be comparable to currently available ePRO systems. The inclusion criteria are listed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion criteria.

Inclusion criteria and determinants

- Population: those with any type of cancer diagnosis or cancer stage, aged ≥16 years
- Intervention: electronic patient-reported outcome interventions where participants report outcomes electronically (web-based, computer, mobile phone, tablet, etc), and responses are subsequently made available to clinical teams
- Comparison: usual care or other control conditions
- Outcomes: validated questionnaires measuring health-related quality of life (primary outcome), symptoms, psychological well-being, satisfaction with care, health care use, survival, and progression-free survival (secondary outcomes)

Note: The PICO (Population, Intervention, Comparison, Outcomes) framework to identify determinants [21] was used for the inclusion criteria.

Exclusion Criteria

We excluded studies based on the criteria presented in Textbox 2.

Textbox 2. The criteria used to exclude studies.

Exclusion criteria

- Book chapters, conference abstracts, commentaries, opinion articles, reviews, meta-analyses, unpublished data, and so on
- Not an electronic patient-reported outcome (ePRO) intervention
- Not focusing on patients with cancer and patients aged ≥16 years
- Not a randomized controlled trial (eg, nonrandomized trial, correlational study, or case study)
- Conditions differ except for the ePRO intervention (eg, the ePRO group received another intervention that was not received by the control group)
- Data published elsewhere
- No control group or an inappropriate control group (eg, where the control condition received an ePRO intervention or another intervention not received by the experimental group)
- Not a health-related outcome measured using validated questionnaires: no health-related quality of life, physical symptoms (eg, nausea, vomiting, pain, breathlessness, and fatigue), psychological symptoms (eg, anxiety and depression), satisfaction with care, health care use outcomes (hospital admission, emergency department visit, and chemotherapy completion), survival, or progression-free survival
- ePRO results were not fed back to clinicians
- Full text not available

Screening

The identified papers were collated and duplicates removed. The screening of titles and abstracts was conducted independently by 2 reviewers (BK and MBP). Full texts were located for any papers meeting the inclusion criteria and again reviewed by 2 reviewers (BK and MBP). Authors were contacted if full texts could not be obtained. A third reviewer (ST) was consulted in case of any disagreements. Backward and forward reference searching was used to identify additional papers.

Data Extraction

Data were extracted and recorded by 2 researchers (BK and MBP) and included sociodemographic and clinical information, type of ePRO system, feature included in the ePRO (according to the taxonomy of system features [14]), study design and characteristics, type and validity of outcome measure, dropouts, sample sizes, and data used to compute the effect size.

Risk-of-Bias Assessment

The risk of bias was assessed by 3 researchers independently (BK, MBP, and AV) according to the revised Cochrane risk-of-bias tool for randomized trials [22] before making a final collaborative decision. Studies were categorized as *low risk of bias*, *some concern*, or *high risk of bias*.

Data Analysis

The outcomes used and the time points of assessments across the studies were assessed for consistency. If a sufficient number of studies reported common outcomes but measured them in different ways, data were combined as standardized mean differences (SMDs). This is a commonly used summary statistic in meta-analysis that expresses the magnitude of the effect in each study compared to the variability observed. It is calculated by taking the difference in mean outcomes between the groups and dividing it by the SD of the outcome among participants [23]. Where outcomes used the same measurement scale, we combined data as the mean difference. We completed 2 separate meta-analyses: one looking at any HRQOL measure closest to 6 months and the other examining Functional Assessment of Cancer Therapy–General (FACT-G) scores at 3 months; we looked at this particular measure and time point specifically because it was the most commonly used. For all analyses, a

XSL•FO
RenderX

maximum likelihood random effects meta-analysis was performed.

The I^2 statistic was used to assess the presence of heterogeneity. There was no observed heterogeneity (I^2=0%), but the 95% CIs were wide (95% CI 0%-64.8% and 95% CI 0%-79.2%), suggesting that the *true* heterogeneity could plausibly be high; therefore, random effects were chosen. The causes of heterogeneity were not explored, although, as sensitivity analyses, we repeated the meta-analyses with fixed effects. Forest plots were used to visually present the results. Stata (version 14; StataCorp LLC) was used for all analyses, specifically the *metaan* command [18].

Results

Results of the Search

The search was conducted in 2 stages. A total of 10,965 papers were identified across the 2 searches (2020 and 2022; search 1: n=7281, 66.4%; search 2: n=3684, 33.6%). The full texts of 37 papers were reviewed (search 1: n=19, 51%; search 2: n=18, 49%). Of these 37 papers, 13 (35%) met the inclusion criteria and were included in the review (Figure 1). References from the included papers were reviewed, and this led to 6 additional articles being identified (search 1: n=5, 83%; search 2: n=1, 17%). Thus, overall, 19 papers were deemed eligible after full-text review.

Figure 1. Illustration of the flow of studies through the systematic review process. ePRO: electronic patient-reported outcome.

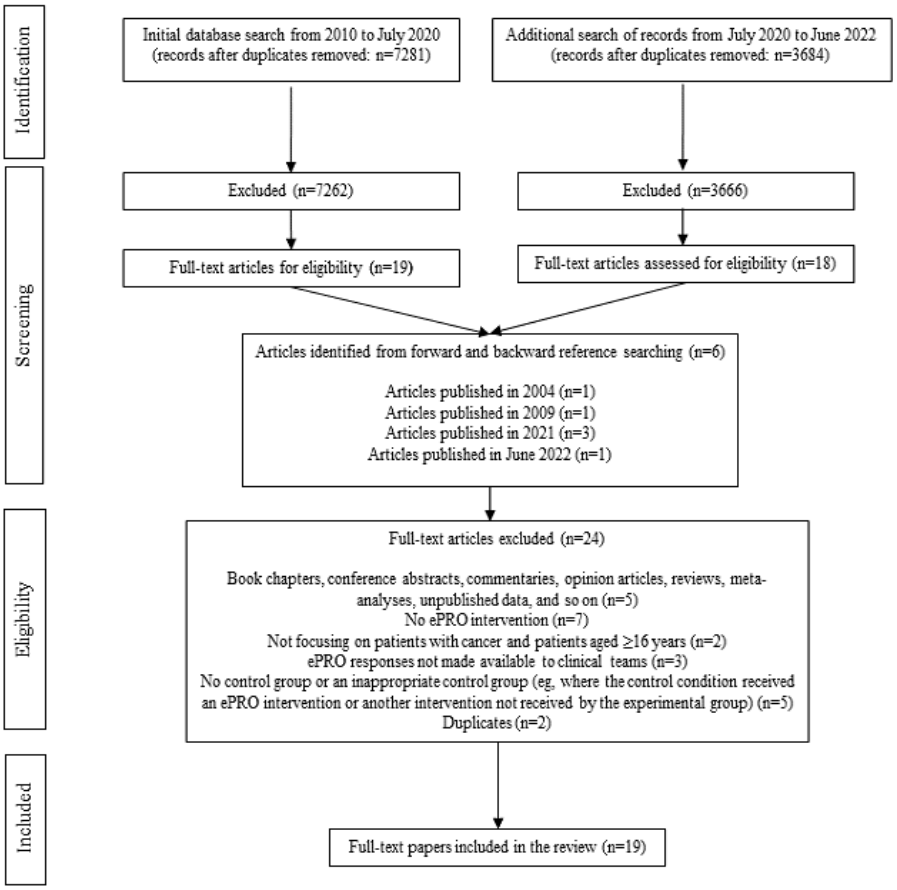

Included Studies

Nineteen papers from 15 RCTs were identified. The second paper of a study was only used if it provided some additional data not presented in the first paper. Reasons for exclusion are summarized in the flow diagram (Figure 1). The characteristics of the 19 included papers are presented in Table 1, and a summary of the studies is presented in Textbox 3. All patients received systemic anticancer therapy. Each study included in the review compared patient reporting of symptoms to the clinical team via an ePRO to a control group. The majority of the studies (14/15, 93%) compared just 2 groups; however, 1 (7%) of the 15 studies included a third *attention-control* group where participants completed the intervention, but the results were not fed back to the clinical team [24]. The control comparison group was defined in the majority of papers as *standard* or *usual* care. Standard or usual care generally consisted of regular appointments with oncology specialists to assess symptoms, followed by appropriate management. Patients were also encouraged to contact the clinical team by telephone if they experienced any problems between appointments. Of the 15 studies, 2 (13%) [25,26] did not provide a definition of standard or usual care.

XSL•FO
RenderX

Table 1. The included papers categorized by study (year and country), methods, sample size, and participant characteristics.

Study, year; country	Methods	Sample size	Participant characteristics
Absolom et al [27], 2021; United Kingdom	RCT[a]; 2 groups; random permuted blocks via an automated 24-hour system	Eligible=782; randomized=508; I[b]=256, C[c]=252; DO: I=42, C=19	• Age (y) mean: I=55.9 (SD 12.2), C=56 (SD 11.3) • Sex: M[d]=102, F[e]=406 • Diagnosis and stage of disease: colorectal, breast, or gynecologic cancers; cancer of any stage
Basch et al [28], 2016; United States	RCT; 2 groups each, with 2 subgroups based on level of prior computer use; computer system using randomly permuted blocks	Eligible=1107; randomized=766; CE[f] participants: I=286, C=253; CI[g] participants: I=155, C=72; DO: I=69, C=89 (CE participants); I=46, C=26 (CI participants)	• Age (y), median: I=61 (IQR 30-91), C=62 (26-88) • Sex: M=322, F=444 • Diagnosis and stage of disease: breast, genitourinary, gynecologic, or lung cancers; metastatic
Basch et al [29], 2017; United States	RCT; 2 groups each, with 2 subgroups based on level of prior computer use; computer system using randomly permuted blocks	Eligible: not stated; randomized=766; I=539, C=227; DO (overall): 249	• Age (y; overall), median: 61 (IQR 26-91) • Sex: M=322, F=444 • Diagnosis and stage of disease: breast, genitourinary, gynecologic, or lung cancers; metastatic
Basch et al [30], 2022; United States	RCT; 2 groups; cluster design	Eligible=1444; randomized=1191; I=593, C=598; DO: I=278, C=230	• Age (y), median: I=64 (IQR 29-89), C=62 (IQR 28-93) • Sex: M=496, F=694; 1197 randomized and 1191 analyzed; however, data on sex are only listed for 1190 patients • Diagnosis and stage of disease: cancer of any type; metastatic
Denis et al [31], 2017; France	RCT; 2 groups; minimization program	Eligible: not stated; randomized=133; I=67, C=66; DO: I=7, C=5	• Age (y), median: I=65.2 (IQR 35.7-86.9), C=64.3 (IQR 42.7-88.1) • Sex: M=81, F=40; 133 randomized and 121 analyzed • Diagnosis and stage of disease: nonprogressive small cell lung cancer or non–small cell lung cancer; advanced (at least cTxN1 or pTxpN1 to TxNxM+ at least stage II cancer)
Greer et al [32], 2020; United States	RCT; 2 groups; computer-generated randomization scheme stratified by cancer type	Eligible=500; randomized=181; I=91, C=90; DO: I=11, C=1	• Age (y), mean: I=52.85 (SD 13.74), C=53.76 (SD 12.08) • Sex: M=84, F=97 • Diagnosis and stage of disease: diverse malignancies; not stated
Kearney et al [33], 2009; United Kingdom	RCT; 2 groups; automated interactive voice response telephone randomization system	Eligible: not available due to incomplete data; randomized=112; I=56, C=56; DO: I=27, C=27	• Age (y), mean: I=55.1 (SD 10.6), C=56.9 (SD 10.5) • Sex: M=26, F=86 • Diagnosis and stage of disease: breast, lung, or colorectal cancer; irrespective of stage
Maguire et al [34], 2021; Austria, Greece, Ireland, Norway, and the United Kingdom	RCT; 2 groups; repeated measures parallel group evaluator-masked stratified trial	Eligible=1222; randomized=840; I=422, C=418; DO: I=134, C=122	• Age (y), mean: I=51.9 (SD 12.4), C=52.9 (SD 12.1) • Sex: M=151, F=678; 840 randomized and 829 analyzable • Diagnosis and stage of disease: breast cancer, colorectal cancer, Hodgkin disease, or non-Hodgkin lymphoma; stages 0 to IV (not metastatic breast or colorectal cancer)
Pappot et al [25], 2021; Denmark	RCT; 2 groups; cluster randomization	Eligible=682; randomized=682; I=347, C=335; DO: I=11, C=22	• Age (y), median: I=53 and C=53 (range 21-82) • Sex: M=0, F=682 • Diagnosis and stage of disease: breast cancer; not stated

XSL•FO
RenderX

Study, year; country	Methods	Sample size	Participant characteristics
Post et al [26], 2013; United States	RCT; 2 groups; method of randomization not described	Eligible=93; randomized=60; I=31, C=29; DO: I=4, C=6	• Age (y), mean: I=49.5 (SD 10.7), C=52.1 (SD 8.5) • Sex: M=0, F=60 • Diagnosis and stage of disease: breast cancer; primary (stages I-III)
Riis et al [35], 2020; Denmark	RCT; 2 groups; computer-generated sequence	Eligible=177; randomized=134; I=65, C=69; DO: I=5, C=7	• Age (y), mean: I=64.4, C=64.2 • Sex: M=0, F=129 • Diagnosis and stage of disease: breast cancer; primary (early breast cancer, stages I-III)
Riis et al [36], 2021; Denmark	RCT; 2 groups; computer-generated sequence	Eligible=177; randomized=134; I=65, C=69; DO: I=5, C=7	• Age (y; overall), mean: 64.3 • Sex: M=0, F=129 • Diagnosis and stage of disease: breast cancer; primary (early breast cancer, stages I-III)
Tolstrup et al [37,38], 2020, 2022; Denmark	RCT; 2 groups; open-label, computer-randomized trial	Eligible=200; randomized=146; I=73, C=73; DO: I=6, C=2	• Age (y), median: I=66, C=66 • Sex: M=78, F=68 • Diagnosis and stage of disease: melanoma; metastatic (stages III-IV)
Velikova et al [24], 2004; United Kingdom	RCT; 3 groups; random permuted blocks by telephone	Eligible=439; randomized: 286; I=144, attention-C=70, C=72; DO: I=60, attention-C=35, C=27	• Age (y), mean: I=55.1 SD (13.02), attention-C=54.8 (SD 12.4), C=54.7 (SD 11.67) • Sex: M=76, F=210 • Diagnosis and stage of disease: cancer of any type; mixed stage
Velikova et al [39], 2010; United Kingdom	RCT; 3 groups; telephone by the research office	Eligible=439; randomized=286; I=144; attention-C=70, C=72; DO: I=59, attention-C=36, C=25	• Age (y), mean: I=54.8 (SD 12.9), attention-C=55.2 (SD 11.79), C=54.9 (SD 11.76) • Sex: M=64, F=194 • Diagnosis and stage of disease: cancer of any type; mixed stage
Wheelock et al [40], 2015; United States	RCT; 2 groups; block design	Eligible=102; randomized=100; I=59, C=41; DO: I=9, C=6	• Age (y), mean: I=54.78 (SD 8.66), C=3.3 (SD 10.79) • Sex: M=0, F=100 • Diagnosis and stage of disease: breast cancer; primary (stages I-III)
Yang et al [41], 2019; China	RCT; 2 groups; scheme generated by independent statistical personnel using a computer	Eligible=58; randomized=58; I=31, C=27; DO: 0	• Age (y), mean: I=51.1 (SD 8.98), C=53.96 (SD 8.58) • Sex: M=38, F=20 • Diagnosis and stage of disease: cancer of any type; not stated
Zhang et al [42], 2022; China	RCT; 2 groups; open-label trial	Eligible=364; randomized=300; I=150, C=150; DO: I=9, C=13	• Age (y), mean: I=57.6 (SD 12.6), C=60.1 (SD 12.7) • Sex: M=206, F=72 • Diagnosis and stage of disease: cancer of any type; life expectancy was at least 6 months; not stated

[a]RCT: randomized controlled trial.

[b]I: intervention.

[c]C: control.

[d]M: male.

[e]F: female.

[f]CE: computer-experienced.

[g]CI: computer-inexperienced.

XSL•FO
RenderX

Textbox 3. Summary of included studies.

Studies summarized

- Countries: United States (5/15, 33%), United Kingdom (3/15, 20%), Denmark (3/15, 20%), China (2/15, 13%), France (1/15, 7%), and multiple European countries (1/15, 7%)
- Randomized controlled trials: 2 groups (14/15, 93%) and 3 groups (1/15, 7%)
- Study size: ranging from 58 to 1191 patients, with a total of 5446 patients; 5497 patients randomized
- Sex: 83.6% (4553/5446) were female
- Stage of disease: primary cancer (3/15, 20%), metastatic cancer (4/15, 27%), and any stage or not specified (8/15, 53%)

Study Quality

The risk-of-bias assessment for each study is summarized in Table 2.

XSL•FO
RenderX

Table 2. Results from the risk-of-bias assessment performed using the revised Cochrane risk-of-bias tool for randomized trials.

Study, year	Randomization and allocation (selection bias)	Blinding of participants (performance bias)	Missing outcome data (attrition bias)	Blinding of outcome assessment (detection bias)	Selective reporting (reporting bias)
Absolom et al [27], 2021	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Basch et al [28,29], 2016 and 2017	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Basch et al [30], 2022	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Denis et al [31], 2017	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Greer et al [32], 2020	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Kearney et al [33], 2009	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Maguire et al [34], 2021	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias
Pappot et al [25], 2021	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Post et al [26], 2013	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Some concerns (did not include a prespecified analysis plan)
Riis et al [35,36], 2020 and 2021	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Tolstrup et al [37,38], 2020 and 2022	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Velikova et al [24,39], 2004 and 2010	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias
Wheelock et al [40], 2015	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Some concerns (provided minimal details of a prespecified analysis plan)
Yang et al [41], 2019	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias
Zhang et al [42], 2022	Low risk of bias	Low risk of bias	Low risk of bias	Some concerns (outcome assessors were aware of the intervention received by study participants)	Low risk of bias

XSL•FO
RenderX

ePRO Intervention Components

Overview

Table 3 describes the components of the included interventions.

XSL•FO
RenderX

Table 3. Intervention components and the effect on review primary outcome and secondary outcomes.

Study, year	Intervention components					Review primary outcome, secondary outcomes, and effect						
	Symptom monitoring	Symptom management	Communication	Alert management	Timing of alerts	Quality of life (primary outcome)	Patient survival	Symptoms	Hospital admissions	Emergency department visits	Chemotherapy completion	Satisfaction with care
Absolom et al [27], 2021	Patients and clinicians (linked to electronic patient records)	Tailored advice for patients; reports sent to clinicians	NR[a]	Clinical team shared email address; monitored by nurses	Real time	+[b]	—[c]	+	–[d]	—	–	—
Basch et al [28,29], 2016 and 2017	Patients and clinicians	Reports sent to clinicians	NR	Email sent to nurses; not monitored 24 hours	Printed at each clinic visit	+	+	+	–	+	+	—
Basch et al [30], 2022	Patients and clinicians	Tailored advice for patients; reports sent to clinicians	NR	Email sent to designated admin staff who forwarded it to relevant nurse	Real time; reports at clinic visits	+	—	+	—	—	—	—
Denis et al [31], 2017	Patients and clinicians	Reports sent to clinicians	NR	Email sent to oncologist	Real time	+	+	—	—	—	—	—
Greer et al [32], 2020	Patients and clinicians	Generic advice for patients; reports sent to clinicians	NR	Email sent to clinician	Not specified	–	—	–	–	–	—	–
Kearney et al [33], 2009	Patients and clinicians	Tailored advice for patients; reports sent to clinicians	NR	Dedicated 24-hour pager system; clinicians should contact patients within 1 hour for severe symptoms	Real time	—	—	+	—	—	—	—
Maguire et al [34], 2021	Patients and clinicians	Tailored advice for patients; reports sent to clinicians	NR	Alerts sent to clinicians on dedicated handsets	Real time	+	—	+	—	—	—	—
Pappot et al [25], 2021	Patients and clinicians	NR	NR	NR	Shown to clinicians after completion at each visit	–	—	–	–	—	–	—
Post et al [26], 2013	Patients and clinicians	Tailored advice for patients	NR	NR	Printed at each clinic visit	–	—	–	—	—	—	—

XSL•FO
RenderX

Study, year	Intervention components					Review primary outcome, secondary outcomes, and effect						
	Symptom monitoring	Symptom management	Communication	Alert management	Timing of alerts	Quality of life (primary outcome)	Patient survival	Symptoms	Hospital admissions	Emergency department visits	Chemotherapy completion	Satisfaction with care
Riis et al [35,36], 2020 and 2021	Patients and clinicians	Reports sent to clinicians	The patient could request a consultation through the ePROM[e] system	Principal investigator monitored questionnaire and emails	Not specified	–	—	–	—	—	–	–
Tolstrup et al [37,38], 2020 and 2022	Patients and clinicians	Patients advised to contact clinical team for severe symptoms	NR	NR	Log in to system to view before consultation	+	—	–	—	—	—	—
Velikova et al [24,39], 2004 and 2010	Patients and clinicians	NR	NR	NR	Printed at each clinic visit	+	—	+	—	—	—	–
Wheelock et al [40], 2015	Patients and clinicians	Tailored advice for patients; reports sent to clinicians	Free text to report concerns and ask questions	Monitored by designated nurse practitioner	Real time	—	—	–	—	—	—	—
Yang et al [41], 2019	Patients and clinicians	Patients advised to follow medication plan in case of severe pain	Real-time consultation	NR	Viewed when patients request a consultation	+	—	+	—	—	—	—
Zhang et al [42], 2022	Patients and clinicians	Tailored advice for patients; reports sent to clinical team	Consult team via app at any time	An oncology specialist and 2 nurses from each center	Viewed before consultation	+	–	+	—	+	–	—

[a]NR: not reported.

[b]Statistically significant effect ($P<.05$).

[c]Did not measure this outcome.

[d]No statistically significant effect ($P>.05$).

[e]ePROM: electronic patient-reported outcome measure.

Symptom Monitoring

All identified ePROs required patients to monitor and report symptoms and included the facility for a clinician to view results. The majority of the ePROs (14/15, 93%) collected data on a wide range of symptoms, whereas 7% (1/15) focused specifically on pain [41].

Symptom Management

Although all ePROs gave the clinical team access to patient reports, 7 (37%) of the 19 studies [24-26,37-39,41] did not actively send reports to the clinical team. Of the 14 systems, 9 (64%) incorporated a facility to alert clinicians if patients reported severe symptoms or a change in symptoms over time

XSL•FO
RenderX

[27-31,33-36,40,42]. Greer et al [32] sent all reports (not just those reporting severe symptoms) to the clinical team. Of the 15 interventions, 9 (60%) provided tailored self-management advice for patients based on the problems reported and their severity [26,27,30,33,34,40,42]. In some instances (2/15, 13%), the advice was to contact the clinical team [28,29,40], whereas others (5/15, 33%) provided links to self-management techniques and advice [26,30,33,34,41]. Some (2/15, 13%) used algorithms based on symptom severity to indicate whether patients should receive self-management advice or be advised to contact the clinical team [27,32].

Communication

Of the 15 ePROs, 4 (27%) facilitated patient communication with the clinical team [35,36,40-42]. Communication facilities included the ability to contact the team at any time through the app, initiate or request the need for a consultation, and use free text to report concerns and ask questions.

Alert Management

The majority of the ePROs (6/15, 40%) sent alerts by email; however, in 2 (13%) of the 15 studies, dedicated handsets or pager systems were used [33,34]. Only 1 (7%) of the 15 studies specified 24-hour alert monitoring [33]. Reports were usually sent to designated clinicians; however, in 1 (7%) of the 15 studies, reports were sent to an administrative team who then directed them to an appropriate member of the nursing team [30]. ePRO questionnaire responses were only integrated into electronic patient record systems in 1 (7%) of the 15 studies [27]. The remaining studies (2/15, 13%) used stand-alone web-based systems that required the clinical team to log in to view ePRO responses.

Timing of Alerts

The timing of delivery of ePRO reports to clinicians varied: in 6 (40%) of the 15 studies, electronic reports or alerts were provided in real time [27,30,31,33,34,40]; in 6 (40%) of the 15 studies, reports were reviewed before consultations [24-26,28,29,37-39,42]; and in other studies, reports were reviewed weekly (1/15, 7%) [32], only if patients requested a consultation (1/15, 7%) [41], or if no timing was specified (2/15, 13%) [32,35,36].

Delivery of the Intervention

The frequency of expected ePRO completion varied across the studies. Some of the studies requested reports at specified time intervals: daily (1/15, 7%) [41], weekly (7/15, 47%) [26,27,30-32,37,38,42], or every third month (2/15, 13%) [35,36,40]. Other studies based ePRO completion around clinical time points: before each clinic visit (2/15, 13%) [24,28,29,39], before each cycle of chemotherapy (2/15, 13%) [25,34], or on days 1 to 14 of each chemotherapy cycle (2/15, 13%) [33]. None of the included studies provided data detailing the fidelity of intervention delivery. Of the 15 studies, 2 (13%) mentioned administrative errors where patient data were not collected due to questionnaires not being given, but this referred to outcome data only and not to the ePRO intervention [32,38].

Patient Adherence to Allocated Intervention

Data on patient adherence to the trial interventions were available in 10 (67%) of the 15 studies [24-30,33-40]. Patient adherence was not standardized across the studies; rather, it was assessed and reported in various ways. Of the 19 papers, 4 (21%) [25,30,34,40] reported the percentage of intervention completions versus expected intervention completions across the whole study; 4 (21%) [27,33,35,36] reported completion rates by time point, either for individual patients or as an average; 4 (21%) [28,29,37,38] reported the percentage of patients across the whole study who completed the intervention as per protocol; and 1 (5%) [26] reported the percentage of participants who completed reports. Each paper reported an individual adherence rate, and the figures reported were between 50% and 100%, with only 11% (2/19) reporting figures <70% [27,40]. None of the studies reported adherence in terms of whether the interventions were fully or partially completed. Of the 4 papers that presented adherence by time point, Kearney et al [33] and Absolom et al [27] reported a decrease over time (from 100% to 73% and from 72% to 58%, respectively), whereas Riis et al [35,36] reported no significant change over time (P=.37).

Primary Outcome: Quality of Life

Table 4 highlights the intervention focus for the included papers along with the primary outcome and secondary outcomes.

XSL•FO
RenderX

Table 4. Intervention types along with the study primary outcome and secondary outcomes.

Study, year	Intervention type	Study primary outcome	Study secondary outcomes
Absolom et al [27], 2021	eRAPID[a], an online eHealth system for patients to self-report symptoms	Symptom control	Impacts on hospital services (process of care measures) and cost-effectiveness
Basch et al [28,29], 2016 and 2017	Reporting of 12 common symptoms via STAR[b], a web-based interface	HRQOL[c]	ED[d] visits, hospitalizations, overall survival, and survival at 1 year
Basch et al [30], 2022	Electronic symptom monitoring with PRO[e] surveys	Overall survival	Physical function, symptom control, and HRQOL
Denis et al [31], 2017	Web-mediated follow-up of symptoms	Overall survival	Performance status at first relapse, progression-free survival, and HRQOL
Greer et al [32], 2020	Smartphone mobile app	Adherence, symptom burden, and quality of life	Patient satisfaction with treatment and health care use
Kearney et al [33], 2009	Mobile phone–based remote monitoring ASyMS[f]	Symptom scores and the occurrence of 6 symptoms that are components of the total symptom score	__[g]
Maguire et al [34], 2021	Remote monitoring via the ASyMS	Symptom burden	HRQOL, supportive care needs, anxiety, self-efficacy, and work limitations
Pappot et al [25], 2021	ePRO[h] questionnaire of symptom toxicities	Number of patients with ≥1 treatment adjustments	Number of patients with ≥1 hospitalizations, ≥1 events of febrile neutropenia, number of patients with treatment postponed >7 days, as well as completion of the scheduled 6 cycles of chemotherapy was registered, and compliance to ePRO
Post et al [26], 2013	PDA-delivered symptom communication	Effects on pain, depression, and fatigue symptoms	Study feasibility, patient and clinician responses to study participation, and intervention effects on HRQOL and communication self-efficacy
Riis et al [35,36], 2020 and 2021	A patient-initiated follow-up program customized to the needs of the individual	Satisfaction with care and unmet needs	Use of consultations, adherence to treatment, and quality of life; number of in-person, telephone, and email consultations; and patient satisfaction
Tolstrup et al [37,38], 2020 and 2022	Web-based symptom reporting using AmbuFlex	Number of severe adverse events (grades 3-4)	Service use (eg, number of telephone consultations as well as outpatient and inpatient visits), length of time toxicities experienced, and length of time steroids required; HRQOL; and associations between toxicity severity and HRQOL
Velikova et al [24,39], 2004 and 2010	Touch screen HRQOL questionnaires	HRQOL, physician-patient communication, and clinical management	Process measures (tests, drugs, and medical records), continuity of care, and patient satisfaction
Wheelock et al [40], 2015	An online health questionnaire with a component of remote follow-up called SIS.NET[i]	Quantify the time between symptom reporting and remote evaluation of symptoms	Compare use of health care resources (breast cancer–related visits, total number of medical appointments, and laboratory and imaging studies)
Yang et al [41], 2019	A mobile phone app (Pain Guard)	Remission rate of pain	Medication adherence, improvements in HRQOL, frequency of breakthrough cancer pain, incidence of adverse reactions, and patient satisfaction
Zhang et al [42], 2022	ePRO follow-up mobile app	Incidence of serious (grades 3-4) immune-related adverse events, ED visits, HRQOL, time spent implementing the ePRO model, rate of treatment discontinuation, and death	—

[a]eRAPID: electronic patient self-reporting of adverse events: patient information and advice.

[b]STAR: Symptom Tracking and Reporting.

[c]HRQOL: health-related quality of life.

XSL•FO
RenderX

[d]ED: emergency department.

[e]PRO: patient-reported outcome.

[f]ASyMS: advanced symptom management system.

[g]Not applicable.

[h]ePRO: electronic patient-reported outcome.

[i]SIS.NET: system for individualized survivorship care, based on patient self-reported data, with review by nurse practitioners, targeted education, and triage.

Of the 15 studies, 13 (87%) measured HRQOL, of which 9 (69%) found a statistically significant effect. Only 5 (38%) of these 13 studies [24,26,32,34,40] used a specific tool to measure psychological well-being; in most cases (n=8, 62%), this was measured with a general HRQOL tool. Tolstrup et al [38] found that, at 48 weeks, the intervention group patients had higher mean scores than the control group patients (mean difference 0.06, 95% CI −0.00 to 0.13; *P*=.05). Yang et al [41] found that global HRQOL scores for the ePRO group were significantly higher than those for the control group (*P*<.001). Basch et al [28] found statistically significant improvements for the intervention arm at 6 months for 3 EQ-5D subdomains (mobility: *P*=.02, self-care: *P*=.01, and anxiety and depression: *P*=.01). Improvements in FACT-G scores were reported by Maguire et al [34] (SMD 4.06, 95% CI 2.65-5.46; *P*<.001) and Velikova et al [24,39] (SE 2.84, 95% CI 13.64-2.37; *P*=.006). Velikova et al [24] found statistically significant changes in FACT-G physical well-being and FACT-G functional well-being subscale scores in particular (*P*=.03). Zhang et al [42] found higher total mean scores for HRQOL in the intervention group at 6 months (mean 74.2, SD 15.1, 95% CI 71.7-76.9 vs mean 64.7, SD 28.5, 95% CI 61.0-68.4; *P*=.01), particularly physical function (mean 84.9, SD 10.5, 95% CI 82.9-88.5 vs mean 68.8, SD 20.7, 95% CI 65.8-72.5; *P*=.001). Basch et al [30] found that patients in the ePRO group had significantly greater HRQOL than those in the usual care group (odds ratio 1.41, 95% CI 1.10-1.81; *P*=.006). Absolom et al [27] found that participants in the intervention group reported better overall health on the EQ-5D visual analog scale at 18 weeks (mean 75.6, SD 18.0 vs mean 68.7, SD 20.4; mean 4.48, 95% CI 1.11-7.86; *P*=.009) and 12 weeks (mean 74.0, SD 16.6 vs mean 71.4, SD 19.5; mean 3.50, 95% CI 0.35-6.66; *P*=.03), but there was no difference at 6 weeks (mean 74.0, SD 17.3 vs mean 71.4, SD 19.5; mean 1.36, 95% CI 21.66-4.39; *P*=.38). Denis et al [31] reported that HRQOL at 6 months was stable or that it improved more in the experimental arm (81% vs 59%; *P*=.04). Of the 15 studies, 4 (27%) [25,26,32,35,36] found no statistically significant differences in relation to HRQOL between groups, and 2 (13%) did not examine HRQOL [28,29,40].

Meta-Analysis

For RCTs with >1 paper (eg, Basch et al [28,29]), only 1 paper was included in the meta-analysis. Of the 15 studies, 8 (53%) were included in the meta-analysis exploring the effect of any HRQOL measure closest to 6 months (Figure 2 [24,27,28,31,32,34,35,38]). Overall, treatment at 6 months demonstrated an average small improvement (SMD 0.07, 95% CI –1.24 to 1.39), although with a wide 95% CI, suggesting that an effect in either direction is possible. There was relatively little variability in reported effect sizes, which ranged from –0.22 to 0.56, although the 95% CIs surrounding these values were often wide. Of the 15 studies, 5 (33%) were included in a meta-analysis of FACT-G scores at 3 months (Figure 3 [24,27,31,32,34]). Here too, the effect of treatment on FACT-G scores at 3 months showed a small average improvement with a wide 95% CI (SMD 0.28, 95% CI –1.22 to 1.78), again suggesting that the true effect of treatment could be positive, negative, or null. Repeating these analyses using fixed effects did not change the results.

XSL•FO
RenderX

Figure 2. Forest plot of the meta-analysis exploring the effect of any health-related quality of life measure at 6 months.

Restricted Maximum Likelihood (REML) method selection

Study	Effect	[95% Confidence Interval]		% Weight
Absolom et al.	0.070	-2.520	2.660	25.72
Basch et al. 2016	0.312	-2.633	3.257	19.89
Greer et al.	-0.223	-3.369	2.923	17.43
Riis et al. 2020	0.000	-2.726	2.726	23.21
Denis et al.	0.561	-6.577	7.698	3.39
Velikova et al. 2004	0.121	-5.875	6.117	4.80
Tolstrup et al. 2022	0.180	-8.808	9.169	2.14
Maguire et al. 2021	0.114	-6.970	7.199	3.44

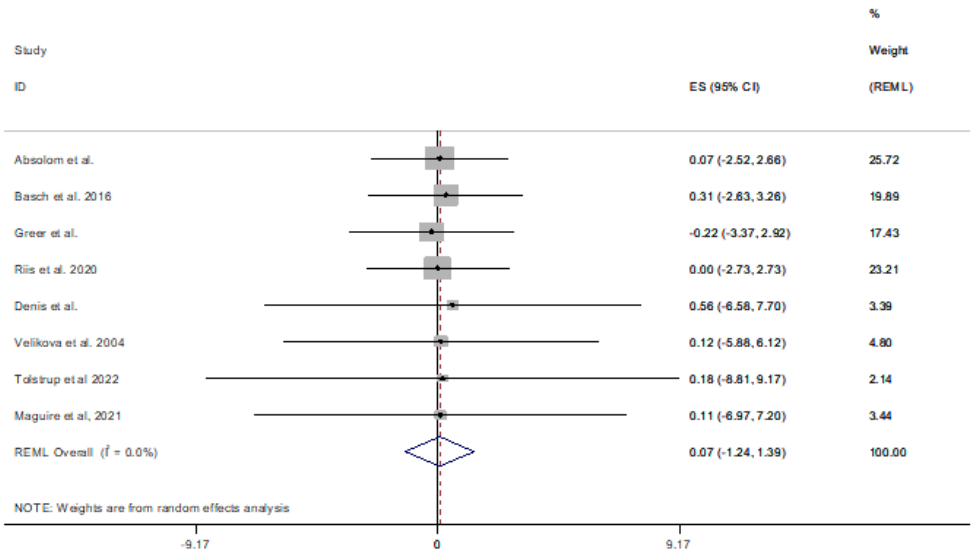

XSL•FO
RenderX

Figure 3. Forest plot of the meta-analysis exploring the effect of treatment on Functional Assessment of Cancer Therapy–General scores at 3 months. ES: effect size.

Restricted Maximum Likelihood (REML) method selection

Study	Effect	[95% Confidence Interval]		% Weight
Denis et al.	0.239	-6.202	6.680	5.41
Absolom et al.	0.312	-2.786	3.410	23.36
Greer et al.	0.224	-2.922	3.369	22.67
Velikova et al. 2004	0.394	-5.418	6.207	6.64
Maguire et al. 2021	0.275	-2.038	2.587	41.92
Overall effect (REML)	0.278	-1.220	1.775	100.00

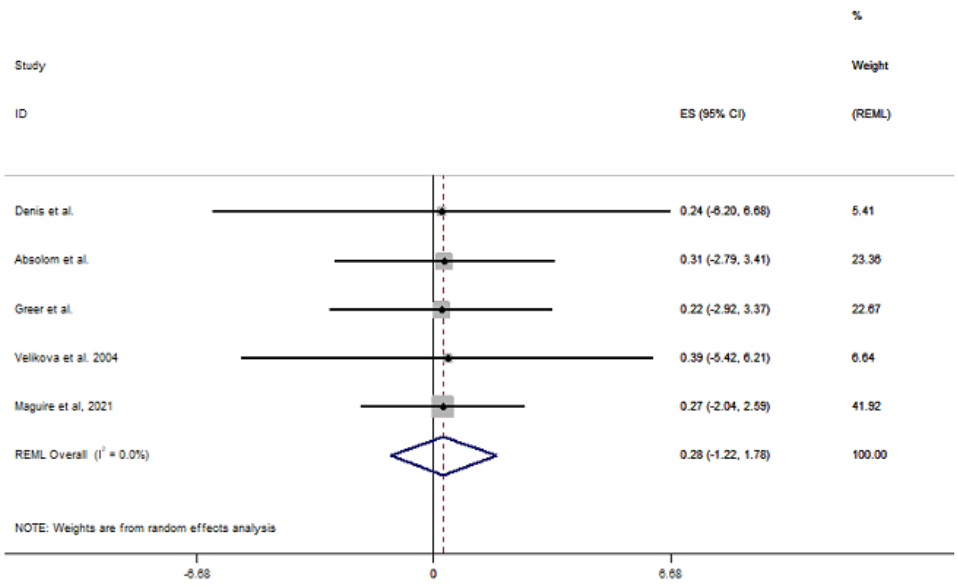

Secondary Outcomes

Patients' Survival

Of the 19 papers, 4 (21%) reported survival data [28,31,42], with Basch et al [29] performing a follow-up survival analysis. Basch et al [28] found a statistically significant difference in patients' overall survival at 1 year (75% vs 69%; *P*=.05) and in quality-adjusted survival (mean 8.7 vs 8.0 mo; *P*=.004). Basch et al [29] also explored longer-term survival and reported a significant difference between groups (mean 31.2 months for the ePRO group compared to mean 26.0 months for the usual care group; *P*=.03). Denis et al [31] reported an improvement of 7 months in overall survival for the intervention group (hazard ratio 0.32, 95% CI 0.15 to 0.67; *P*=.002). Zhang et al [42] found no significant differences in survival between the 2 groups (hazard ratio 0.38, 95% CI 0.07-1.99; *P*=.28). The methodologies used by the studies to calculate survival rates were different and included logistic regression [28], the Kaplan-Meier method with log rank tests [29,31], and chi-square tests [42]. Due to the inconsistent methods of comparison, we compared these descriptively rather than in a meta-analysis.

Patient Symptoms

Of the 19 papers, 14 (74%) reported the effect of the intervention on patients' symptoms, as shown in Table 3. Of these 14 papers, 7 (50%) reported a positive effect on symptoms at 5 to 6 months or after 6 cycles of chemotherapy [24,28-30,34,39,42]. Velikova et al [24] found statistically significant changes in FACT-G physical well-being subscale scores in the intervention group compared to the control group (*P*=.006). Maguire et al [34] found that symptoms were better controlled in the intervention group, remaining at pretreatment levels, whereas the symptoms of control group participants increased initially (least squares absolute mean difference –0.15, 95% CI –0.19 to –0.12; *P*<.001).

Other studies reported the impact of the ePRO on symptoms at 3 months. Absolom et al [27] found that FACT-G physical well-being subscale scores showed that there was better symptom control for participants in the electronic patient self-reporting of adverse events: patient information and advice arm at 6 weeks (difference of least squares means 1.08, SE 0.49, 95% CI 0.12 to 2.05; *P*=.03) and 12 weeks (difference of least squares means 1.01, SE 0.49, 95% CI 0.05-1.98; *P*=.04), but there was no difference at 18 weeks (difference of least squares means 0.2, SE 0.51, 95% CI 20.81-1.20; *P*=.70). Basch et al [30] found a significant improvement in symptom control in the intervention group compared to the control group (odds ratio 1.50, 95% CI 1.15-1.95; *P*=.003). Kearney et al [33] found that patients who had received the intervention reported lower levels of fatigue than those in the control group (67% vs 81%; *P*=.04).

Yang et al [41] reported a positive effect on symptoms at 2 weeks and 4 weeks. Yang et al [41] found significant improvements in various symptoms (nausea and vomiting:

XSL•FO
RenderX

W=272; *P*=.01; constipation: W=261; *P*=.008; fatigue: W=211.5; *P*=.001; and pain: W=177; *P*<.001) in the intervention group compared to the control group. Of the 19 papers, 6 (32%) reported that the ePRO interventions had no statistically significant effect on symptom control [25,26,32,35,37,40], whereas 1 (5%) paper did not report this outcome [31].

Hospital Admissions

Of the 19 papers, 4 (21%) reported the effect of the intervention on hospital admissions. Basch et al [28] reported that the patients in the intervention group were less frequently hospitalized at 1 year (incidence rate: 45% vs 49%; *P*=.08). The remaining papers (3/4, 75%) reported no significant difference in hospital admissions [25,27,32]. All studies assessed the statistical significance differently and at different time points; for example, Absolom et al [27] used incidence rate ratio over 18 weeks, whereas Pappot et al [25] used odds ratio over 6 cycles of treatment.

Emergency Department Visits

Of the 19 papers, 3 (16%) reported a positive effect of the ePRO intervention on emergency department (ED) visits. Basch et al [28] found that patients receiving the intervention were less frequently admitted to the ED (34% vs 41%; *P*=.02), as did Zhang et al [42] (23% vs 41%; hazard ratio 0.46, 95% CI 0.26-0.81; *P*=.01). Greer et al [32] reported fewer ED visits resulting in hospitalization (*P*=.05). All studies assessed the statistical significance differently and at varying time points, using incidence at 1 year [28], hazard ratio after the intervention (6 mo) [42], and mean (SE) during the 12-week study period [32].

Chemotherapy Completion

Of the 19 papers, 5 (26%) reported the effect of the intervention on chemotherapy completion. Basch et al [28] found that the intervention group remained on chemotherapy longer (mean 8.2 vs 6.3 mo; *P*=.002). Pappot et al [25] found that the 6 scheduled cycles of chemotherapy were completed with treatment adjustments in 34% (ePRO arm) and 40.6% (usual care arm) of the participants (*P*=.10), but this result was not significant. Absolom et al [27], Riis et al [35], and Zhang et al [42] found no significant difference in treatment adherence at the end of the study.

Patient Satisfaction With Care

Of the 19 papers, 4 (21%) reported the effect of the intervention on patient satisfaction with care [24,32,35,39]. None of the papers reported a statistically significant difference between the groups. Velikova et al [39] found that between 79% and 89% of the patients, regardless of the study arm, rated their quality of care as *very good*/*excellent*. Similarly, Riis et al [35] found that participants in both groups were highly satisfied with their follow-up care.

Discussion

Principal Findings

To our knowledge, this is the first meta-analysis to explore the effects of ePROs on health-related outcomes in adults with cancer. All interventions involved patients reporting their symptoms, which clinicians could view and monitor. The presentation of patients' ePRO responses to clinicians varied across the studies. Of the 14 systems, only 1 (7%) embedded reports within the electronic patient record system [27]. Some of the studies (6/15, 40%) provided results to clinicians in real time [27,30,31,33,34,40], whereas others (6/15, 40%) distributed results before consultations [24-26,28,29,37-39,42]. Of the 8 systems that included clinician alerts and reported HRQOL, 6 (75%) had a positive effect [27-31,34,42]. Of the 9 systems that alerted clinicians and reported symptoms, 6 (67%) had a positive effect [27-30,33,34,42]. This suggests that alerting clinicians is a key component of ePROs and likely to lead to positive effects on symptoms and HRQOL. None of the studies explored whether the intervention was more effective for particular patient groups, such as a particular sex or cancer type.

The meta-analysis showed an improvement in HRQOL at 6 months, although caution should be taken when interpreting the results due to the wide 95% CI. FACT-G scores at 3 months showed a small average improvement but again with a wide 95% CI. Due to the heterogeneity of the studies, specifically the different outcome measures and the different data collection time points, not all studies were included in the meta-analysis. Only 5 (33%) of the 15 studies were included in the FACT-G 3-month meta-analysis, with the weighting predominantly spread across 3 (60%) studies [27,32,34]. These 3 studies all provided advice for patients and sent reports to clinicians. Of the 15 studies, 8 (53%) were included in the HRQOL 6-month meta-analysis, with the majority of the weighting spread across 4 (50%) studies [27-29,32,35,36]. Of the 15 studies, 9 (60%) reported a statistically significant effect on HRQOL. The measures used to assess HRQOL, the specific HRQOL elements that improved, and the time points when significant differences were recorded varied across the studies.

Symptoms were the most commonly reported secondary outcome, and a positive effect was reported in 8 (57%) of the 14 studies where this was measured. The symptoms tracked by ePROs varied across each of the studies however all studies covered common treatment side effects such as fatigue, pain, and nausea. A variety of different tools were used to measure symptoms; therefore, it is difficult for comparisons to be made. Some symptom assessment tools covered a wide range of symptoms, whereas others focused on single symptoms. A narrative review has highlighted the potential benefits of ePRO use in outpatient care in terms of symptom monitoring and facilitating more timely interventions [6]. However, the review does not explore individual components of the interventions, which may significantly impact effectiveness.

Only 3 (20%) of the 15 studies measured survival [28,31,42], of which 2 (67%) reported improved survival in the intervention group [28,31], whereas 1 (33%) found no difference [42]. Many of the other studies only included participants with early disease treated with curative intent; therefore, including survival is not appropriate. Studies exploring the use of ePRO interventions in patients with advanced disease should consider including survival in their outcome measures. A meta-analysis exploring the prognostic values of PROs for survival in cancer suggests that there is evidence of a relationship between PRO data,

XSL•FO
RenderX

particularly physical functioning as measured by the European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30 [43]. Another meta-analysis of 6 studies tentatively suggested that patient monitoring via PROs could have a positive effect on survival [44], but only two-thirds of the studies included ePROs, and only one-third were RCTs. In terms of the impact on care and hospital use, only 2 (13%) of the 15 studies [28,42] reported a significant effect on ED visits. The study by Basch et al [28] also reported a significant impact on hospital admissions and chemotherapy completion. In Basch et al [28,29], patients in the intervention arm remained on chemotherapy for significantly longer. The intervention is similar to other interventions in that symptom reports are available to clinicians, alerts are sent to the nursing team, and results are printed at each clinic visit. The study by Basch et al [28] discusses the direct actions taken after patients report problems. These actions included symptom management counseling, supportive medication changes, referral to the ED or hospital, chemotherapy dose modification, and imaging or test orders. All these actions may have had an impact on chemotherapy tolerance and continued adherence.

Of the 15 studies, 14 (93%) had some methodological concerns, which could indicate a potential risk of bias. However, no study was categorized as *high risk of bias* in any of the categories. Blinding was not possible, given the nature of the intervention.

Limitations

There are several limitations to this review, which make it difficult to draw firm conclusions regarding the effectiveness of ePRO interventions in clinical practice. In some of the studies (7/15, 47%), the data presented were insufficient for use in the meta-analysis. Authors were contacted where possible, but the required data were not always available. The ability to combine and compare results was further weakened by the varied outcome measures collected at different time points. Many of the studies (12/15, 80%) included a higher proportion of female participants. Male participants may not be as engaged with ePRO interventions or may not openly discuss health problems; therefore, ePRO interventions may not be as effective for male participants.

Comparison With Prior Work

ePROs have been successfully implemented in clinical practice in other diagnostic groups. In rheumatology, ePROs were found to be feasible for use and provided a unique insight into patient experience [45]. Similarly, in diabetes, the completion of preconsultation ePROs was reported to be feasible and acceptable [46]. The AmbuFlex system was successfully used in follow-up across 9 diagnostic groups, including heart disease, epilepsy, asthma, and some cancers. ePROs were completed and the results used to indicate whether a follow-up appointment was necessary. In line with the studies identified in this review, all these systems included clinician review of symptoms to direct patient management and clinical decisions.

New and more effective oncology treatments are being introduced, meaning that the patient population and clinical workload are increasing. The European Society for Medical Oncology has published guidelines on ePRO use in oncology, providing further evidence on the effectiveness of this approach [7]. Symptom reporting could be useful in helping clinicians to manage a larger number of patients for a longer period of time. The findings from this review can highlight which interventions (and components) may be most useful to support clinical practice and improve patient care. Including a facility to send alerts to clinicians is an important feature that should be included in ePROs aiming to improve HRQOL. Regardless of the features included, any ePRO system will only be a success if patients and clinicians fully engage with it [47]. Patients need to regularly complete PROs, and clinicians need to review results and respond.

Conclusions

The aim of this systematic review and meta-analysis was to synthesize all current evidence on the effectiveness of ePROs in RCTs. In total, 19 papers of 15 RCTs were identified. Nearly two-thirds (9/15, 67%) of the interventions showed positive effects on HRQOL and symptoms in adults with cancer. However, caution should be taken in interpreting the results of this review due to the heterogeneity in the interventions, outcome measures, and data collection time points, as well as the wide 95% CIs. This systematic review should act as a driver for further RCTs to be conducted to explore the effectiveness of ePROs using validated outcome measures. This will help to orientate health care professionals toward the most suitable setting and ePRO features to develop and propose optimal care to patients with cancer.

Acknowledgments

The authors thank members of the Christie Patient Centred Research team for their continued support and the Christie Library and Knowledge Service for their guidance on literature searches.

Data Availability

Data sharing is not applicable to this paper as no data sets were generated or analyzed during this study.

Authors' Contributions

JY, ST, and CF-F conceptualized the study and designed the review. MBP, BK, and AV were involved in the risk-of-bias analysis and in preparing the first draft of the manuscript. MBP and ST extracted and interpreted the data and wrote the final draft of this

XSL•FO
RenderX

manuscript. CF-F and GV randomly checked the accuracy of extracted data. AM and CH performed the meta-analysis. All authors read and approved the final manuscript.

Conflicts of Interest

GV has received honoraria from Pfizer, Novartis, Eisai, Lilly; consultancy fees from AstraZeneca, Roche, Novartis, Pfizer, Seagen, Eisai, Sanofi; and an institutional grant from Pfizer, all unrelated to this work.

Multimedia Appendix 1
Completed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 72 KB - jmir_v26i1e49089_app1.pdf]

References

1. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006 Oct 11;4:79 [FREE Full text] [doi: 10.1186/1477-7525-4-79] [Medline: 17034633]
2. Basch E, Jia X, Heller G, Barz A, Sit L, Fruscione M, et al. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. J Natl Cancer Inst 2009 Dec 02;101(23):1624-1632 [FREE Full text] [doi: 10.1093/jnci/djp386] [Medline: 19920223]
3. Fromme EK, Eilers KM, Mori M, Hsieh YC, Beer TM. How accurate is clinician reporting of chemotherapy adverse effects? A comparison with patient-reported symptoms from the Quality-of-Life Questionnaire C30. J Clin Oncol 2004 Sep 01;22(17):3485-3490. [doi: 10.1200/JCO.2004.03.025] [Medline: 15337796]
4. Carrasco S, Symes L. Patient use of electronic methods to self-report symptoms: an integrative literature review. Oncol Nurs Forum 2018 May 01;45(3):399-416. [doi: 10.1188/18.ONF.399-416] [Medline: 29683117]
5. Campbell N, Ali F, Finlay AY, Salek SS. Equivalence of electronic and paper-based patient-reported outcome measures. Qual Life Res 2015 Aug;24(8):1949-1961. [doi: 10.1007/s11136-015-0937-3] [Medline: 25702266]
6. Aiyegbusi OL, Hughes SE, Peipert JD, Schougaard LM, Wilson R, Calvert MJ. Reducing the pressures of outpatient care: the potential role of patient-reported outcomes. J R Soc Med 2023 Feb;116(2):44-64 [FREE Full text] [doi: 10.1177/01410768231152222] [Medline: 36758615]
7. Di Maio M, Basch E, Denis F, Fallowfield LJ, Ganz PA, Howell D, ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. The role of patient-reported outcome measures in the continuum of cancer clinical care: ESMO clinical practice guideline. Ann Oncol 2022 Sep;33(9):878-892 [FREE Full text] [doi: 10.1016/j.annonc.2022.04.007] [Medline: 35462007]
8. LeBlanc TW, Abernethy AP. Patient-reported outcomes in cancer care - hearing the patient voice at greater volume. Nat Rev Clin Oncol 2017 Dec;14(12):763-772. [doi: 10.1038/nrclinonc.2017.153] [Medline: 28975931]
9. Crockett C, Gomes F, Faivre-Finn C, Howell S, Kasipandian V, Smith E, et al. The routine clinical implementation of electronic patient-reported outcome measures (ePROMs) at the christie NHS foundation trust. Clin Oncol (R Coll Radiol) 2021 Dec;33(12):761-764. [doi: 10.1016/j.clon.2021.06.004] [Medline: 34229926]
10. Crockett C, Faivre-Finn C, Yorke J, Price J. Initial clinical experience of routine electronic patient-reported outcome measures at the christie NHS foundation trust. Clin Oncol (R Coll Radiol) 2022 Apr;34(4):e168. [doi: 10.1016/j.clon.2021.12.011] [Medline: 34991932]
11. Payne A, Horne A, Bayman N, Blackhall F, Bostock L, Chan C, et al. Patient and clinician-reported experiences of using electronic patient reported outcome measures (ePROMs) as part of routine cancer care. J Patient Rep Outcomes 2023 May 04;7(1):42 [FREE Full text] [doi: 10.1186/s41687-023-00544-4] [Medline: 37140730]
12. Crockett C, Price J, Pham M, Abdulwahid D, Bayman N, Blackhall F, et al. Experience with the routine use of electronic patient-reported outcome measures for patients with lung cancer. JCO Clin Cancer Inform 2023 Apr;7:e2200150 [FREE Full text] [doi: 10.1200/CCI.22.00150] [Medline: 37071029]
13. Jensen RE, Snyder CF, Abernethy AP, Basch E, Potosky AL, Roberts AC, et al. Review of electronic patient-reported outcomes systems used in cancer clinical care. J Oncol Pract 2014 Jul;10(4):e215-e222 [FREE Full text] [doi: 10.1200/JOP.2013.001067] [Medline: 24301843]
14. Warrington L, Absolom K, Conner M, Kellar I, Clayton B, Ayres M, et al. Electronic systems for patients to report and manage side effects of cancer treatment: systematic review. J Med Internet Res 2019 Jan 24;21(1):e10875 [FREE Full text] [doi: 10.2196/10875] [Medline: 30679145]
15. Kotronoulas G, Kearney N, Maguire R, Harrow A, Di Domenico D, Croy S, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. J Clin Oncol 2014 May 10;32(14):1480-1501 [FREE Full text] [doi: 10.1200/JCO.2013.53.5948] [Medline: 24711559]

XSL•FO
RenderX

16. Larson JL, Rosen AB, Wilson FA. The effect of telehealth interventions on quality of life of cancer patients: a systematic review and meta-analysis. Telemed J E Health 2018 Jun;24(6):397-405. [doi: 10.1089/tmj.2017.0112] [Medline: 29112484]
17. Xu A, Wang Y, Wu X. Effectiveness of e-health based self-management to improve cancer-related fatigue, self-efficacy and quality of life in cancer patients: Systematic review and meta-analysis. J Adv Nurs 2019 Dec;75(12):3434-3447. [doi: 10.1111/jan.14197] [Medline: 31566769]
18. Kontopantelis E, Reeves D. Metaan: random-effects meta-analysis. Stata J 2010 Sep 01;10(3):395-407 [FREE Full text] [doi: 10.1177/1536867X1001000307]
19. Li Y, Li J, Hu X. The effectiveness of symptom management interventions based on electronic patient-reported outcomes (ePROs) for symptom burden, quality of life, and overall survival among patients with cancer: a meta-analysis of randomized controlled trials. Int J Nurs Stud 2023 Nov;147:104588. [doi: 10.1016/j.ijnurstu.2023.104588] [Medline: 37690275]
20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
21. Aromataris E, Munn Z. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2020.
22. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
23. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons Ltd; 2011.
24. Velikova G, Booth L, Smith AB, Brown PM, Lynch P, Brown JM, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. J Clin Oncol 2004 Feb 15;22(4):714-724. [doi: 10.1200/JCO.2004.06.078] [Medline: 14966096]
25. Pappot H, Baeksted CW, Nissen A, Knoop A, Mitchell SA, Christensen J, et al. Clinical effects of assessing electronic patient-reported outcomes monitoring symptomatic toxicities during breast cancer therapy: a nationwide and population-based study. Breast Cancer 2021 Sep;28(5):1096-1099. [doi: 10.1007/s12282-021-01244-x] [Medline: 33837509]
26. Post DM, Shapiro CL, Cegala DJ, David P, Katz ML, Krok JL, et al. Improving symptom communication through personal digital assistants: the CHAT (Communicating Health Assisted by Technology) project. J Natl Cancer Inst Monogr 2013 Dec;2013(47):153-161 [FREE Full text] [doi: 10.1093/jncimonographs/lgt027] [Medline: 24395985]
27. Absolom K, Warrington L, Hudson E, Hewison J, Morris C, Holch P, et al. Phase III randomized controlled trial of eRAPID: eHealth intervention during chemotherapy. J Clin Oncol 2021 Mar 01;39(7):734-747 [FREE Full text] [doi: 10.1200/JCO.20.02015] [Medline: 33417506]
28. Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol 2016 Feb 20;34(6):557-565 [FREE Full text] [doi: 10.1200/JCO.2015.63.0830] [Medline: 26644527]
29. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
30. Basch E, Schrag D, Henson S, Jansen J, Ginos B, Stover AM, et al. Effect of electronic symptom monitoring on patient-reported outcomes among patients with metastatic cancer: a randomized clinical trial. JAMA 2022 Jun 28;327(24):2413-2422 [FREE Full text] [doi: 10.1001/jama.2022.9265] [Medline: 35661856]
31. Denis F, Lethrosne C, Pourel N, Molinier O, Pointreau Y, Domont J, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients. J Natl Cancer Inst 2017 Sep 01;109(9):436. [doi: 10.1093/jnci/djx029] [Medline: 28423407]
32. Greer JA, Jacobs JM, Pensak N, Nisotel LE, Fishbein JN, MacDonald JJ, et al. Randomized trial of a smartphone mobile app to improve symptoms and adherence to oral therapy for cancer. J Natl Compr Canc Netw 2020 Feb;18(2):133-141. [Medline: 32023526]
33. Kearney N, McCann L, Norrie J, Taylor L, Gray P, McGee-Lennon M, et al. Evaluation of a mobile phone-based, advanced symptom management system (ASyMS) in the management of chemotherapy-related toxicity. Support Care Cancer 2009 Apr;17(4):437-444 [FREE Full text] [doi: 10.1007/s00520-008-0515-0] [Medline: 18953579]
34. Maguire R, McCann L, Kotronoulas G, Kearney N, Ream E, Armes J, et al. Real time remote symptom monitoring during chemotherapy for cancer: European multicentre randomised controlled trial (eSMART). BMJ 2021 Jul 21;374:n1647 [FREE Full text] [doi: 10.1136/bmj.n1647] [Medline: 34289996]
35. Riis CL, Jensen PT, Bechmann T, Möller S, Coulter A, Steffensen KD. Satisfaction with care and adherence to treatment when using patient reported outcomes to individualize follow-up care for women with early breast cancer - a pilot randomized controlled trial. Acta Oncol 2020 Apr;59(4):444-452. [doi: 10.1080/0284186X.2020.1717604] [Medline: 32000559]
36. Riis CL, Stie M, Bechmann T, Jensen PT, Coulter A, Möller S, et al. ePRO-based individual follow-up care for women treated for early breast cancer: impact on service use and workflows. J Cancer Surviv 2021 Aug;15(4):485-496. [doi: 10.1007/s11764-020-00942-3] [Medline: 33415653]

XSL•FO
RenderX

37. Tolstrup LK, Bastholt L, Dieperink KB, Möller S, Zwisler AD, Pappot H. The use of patient-reported outcomes to detect adverse events in metastatic melanoma patients receiving immunotherapy: a randomized controlled pilot trial. J Patient Rep Outcomes 2020 Oct 30;4(1):88 [FREE Full text] [doi: 10.1186/s41687-020-00255-0] [Medline: 33125537]
38. Tolstrup LK, Pappot H, Bastholt L, Möller S, Dieperink KB. Impact of patient-reported outcomes on symptom monitoring during treatment with checkpoint inhibitors: health-related quality of life among melanoma patients in a randomized controlled trial. J Patient Rep Outcomes 2022 Jan 21;6(1):8 [FREE Full text] [doi: 10.1186/s41687-022-00414-5] [Medline: 35061112]
39. Velikova G, Keding A, Harley C, Cocks K, Booth L, Smith AB, et al. Patients report improvements in continuity of care when quality of life assessments are used routinely in oncology practice: secondary outcomes of a randomised controlled trial. Eur J Cancer 2010 Sep;46(13):2381-2388 [FREE Full text] [doi: 10.1016/j.ejca.2010.04.030] [Medline: 20570138]
40. Wheelock AE, Bock MA, Martin EL, Hwang J, Ernest ML, Rugo HS, et al. SIS.NET: a randomized controlled trial evaluating a web-based system for symptom management after treatment of breast cancer. Cancer 2015 Mar 15;121(6):893-899 [FREE Full text] [doi: 10.1002/cncr.29088] [Medline: 25469673]
41. Yang J, Weng L, Chen Z, Cai H, Lin X, Hu Z, et al. Development and testing of a mobile app for pain management among cancer patients discharged from hospital treatment: randomized controlled trial. JMIR Mhealth Uhealth 2019 May 29;7(5):e12542 [FREE Full text] [doi: 10.2196/12542] [Medline: 31144672]
42. Zhang L, Zhang X, Shen L, Zhu D, Ma S, Cong L. Efficiency of electronic health record assessment of patient-reported outcomes after cancer immunotherapy: a randomized clinical trial. JAMA Netw Open 2022 Mar 01;5(3):e224427 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.4427] [Medline: 35357459]
43. Efficace F, Collins GS, Cottone F, Giesinger JM, Sommer K, Anota A, et al. Patient-reported outcomes as independent prognostic factors for survival in oncology: systematic review and meta-analysis. Value Health 2021 Feb;24(2):250-267 [FREE Full text] [doi: 10.1016/j.jval.2020.10.017] [Medline: 33518032]
44. Caminiti C, Maglietta G, Diodati F, Puntoni M, Marcomini B, Lazzarelli S, et al. The effects of patient-reported outcome screening on the survival of people with cancer: a systematic review and meta-analysis. Cancers (Basel) 2022 Nov 07;14(21):5470 [FREE Full text] [doi: 10.3390/cancers14215470] [Medline: 36358888]
45. Tymms K, O'Sullivan C, Smith T, Littlejohn G, Freeman T, Hoffman D, et al. A novel electronic patient-reported outcome delivery system to implement health-related quality of life measures in routine clinical care: an analysis of 5 years of experience. Front Digit Health 2022;4:1074931 [FREE Full text] [doi: 10.3389/fdgth.2022.1074931] [Medline: 36698650]
46. Hernar I, Graue M, Richards D, Strandberg RB, Nilsen RM, Tell GS, et al. Electronic capturing of patient-reported outcome measures on a touchscreen computer in clinical diabetes practice (the DiaPROM trial): a feasibility study. Pilot Feasibility Stud 2019;5:29 [FREE Full text] [doi: 10.1186/s40814-019-0419-4] [Medline: 30820340]
47. Basch E, Barbera L, Kerrigan CL, Velikova G. Implementation of patient-reported outcomes in routine medical care. Am Soc Clin Oncol Educ Book 2018 May 23;38:122-134 [FREE Full text] [doi: 10.1200/EDBK_200383] [Medline: 30231381]

Abbreviations

ED: emergency department
ePRO: electronic patient-reported outcome
FACT-G: Functional Assessment of Cancer Therapy–General
HRQOL: health-related quality of life
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRO: patient-reported outcome
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by A Mavragani; submitted 17.05.23; peer-reviewed by A Major, Y Chu; comments to author 20.12.23; revised version received 07.02.24; accepted 10.04.24; published 31.07.24.

Please cite as:
Perry MB, Taylor S, Khatoon B, Vercell A, Faivre-Finn C, Velikova G, Marsden A, Heal C, Yorke J
Examining the Effectiveness of Electronic Patient-Reported Outcomes in People With Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e49089
URL: https://www.jmir.org/2024/1/e49089
doi:10.2196/49089
PMID:

XSL•FO
RenderX

©Melissa Betty Perry, Sally Taylor, Binish Khatoon, Amy Vercell, Corinne Faivre-Finn, Galina Velikova, Antonia Marsden, Calvin Heal, Janelle Yorke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Training Service Users in the Use of Telehealth: Scoping Review

Emer Galvin[1], BA, MSc; Shane Desselle[2], RPh, PhD; Blánaid Gavin[3], MB BCh BAO, MSc CBT; Fiona McNicholas[3,4,5], MB MCh BAO, MD, Dip Clin Psychother; Shane Cullinan[1], BPharm, PhD; John Hayden[1], BPharm, MPharm, PhD

[1]School of Pharmacy and Biomolecular Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland
[2]Touro University California, California, CA, United States
[3]School of Medicine, University College Dublin, Dublin, Ireland
[4]Children's Health Ireland, Dublin, Ireland
[5]Lucena Child and Adolescent Mental Health Service (CAMHS), Dublin, Ireland

Corresponding Author:
Emer Galvin, BA, MSc
School of Pharmacy and Biomolecular Sciences
Royal College of Surgeons in Ireland
111 St Stephen's Green
Dublin, D02 YN77
Ireland
Phone: 353 1 402 2100
Email: emergalvin20@rcsi.ie

Abstract

Background: The use of telehealth has rapidly increased, yet some populations may be disproportionally excluded from accessing and using this modality of care. Training service users in telehealth may increase accessibility for certain groups. The extent and nature of these training activities have not been explored.

Objective: The objective of this scoping review is to identify and describe activities for training service users in the use of telehealth.

Methods: Five databases (MEDLINE [via PubMed], Embase, CINAHL, PsycINFO, and Web of Science) were searched in June 2023. Studies that described activities to train service users in the use of synchronous telehealth consultations were eligible for inclusion. Studies that focused on health care professional education were excluded. Papers were limited to those published in the English language. The review followed the Joanna Briggs Institute guidelines for scoping reviews and was reported in line with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Titles and abstracts were screened by 1 reviewer (EG). Full texts were screened by 2 reviewers (EG and JH or SC). Data extraction was guided by the research question.

Results: The search identified 8087 unique publications. In total, 13 studies met the inclusion criteria. Telehealth training was commonly described as once-off preparatory phone calls to service users before a telehealth visit, facilitated primarily by student volunteers, and accompanied by written instructions. The training content included guidance on how to download and install software, troubleshoot technical issues, and adjust device settings. Older adults were the most common target population for the training. All but 1 of the studies were conducted during the COVID-19 pandemic. Overall, training was feasible and well-received by service users, and studies mostly reported increased rates of video visits following training. There was limited and mixed evidence that training improved participants' competency with telehealth.

Conclusions: The review mapped the literature on training activities for service users in telehealth. The common features of telehealth training for service users included once-off preparatory phone calls on the technical elements of telehealth, targeted at older adults. Key issues for consideration include the need for co-designed training and improving the broader digital skills of service users. There is a need for further studies to evaluate the outcomes of telehealth training activities in geographically diverse areas.

(J Med Internet Res 2024;26:e57586) doi:10.2196/57586

KEYWORDS

telehealth; video consultations; training; education; older adults; digital divide; digital literacy; review; scoping review; modality of care; training service; user; users; older adult; gerontology; geriatric; geriatrics; caregiver; caregivers; consultation; consultations;

XSL•FO
RenderX

health care professional; health care professionals; PRISMA-ScR; Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; data extraction; phone; phones; telemonitoring

Introduction

The use of telehealth, defined as live audio and video consultations between service users and health care professionals, has increased exponentially since the onset of the COVID-19 pandemic [1]. The presence of physical distancing restrictions and stay-at-home orders necessitated this form of health care delivery [2]. The use of telehealth has prevailed beyond the pandemic, due to its benefits to service users and clinicians alike. These benefits include timely access to care, removal of logistical barriers, and convenience [3]. While these benefits are lauded, many underserved populations, such as those of low income, still face barriers to accessing telehealth [4,5]. One reason for this disparity in access may be because of the presence of the "digital divide."

The digital divide has been recognized as an important social determinant of health [6] and is recognized as having 3 elements [7]. These are (1) disparities in access to technologies, (2) disparities in skills to be able to use these technologies efficiently, and (3) disparities in people's ability to use these technologies to achieve outcomes to improve their lifestyle [7,8]. While device ownership and internet access are increasing worldwide, gaps in digital skills and competencies are still prevalent [9]. In the United Kingdom, 21% of the population lack basic digital skills [10]. These inequalities in digital skills are recognized as contributing to disparities in telehealth use [5,9,11,12].

It is now critical that there are programs to improve the digital literacy of service users so that they have the confidence and skills to use, and benefit from, telehealth [13]. Research on training health care professionals in telehealth has grown, with recent studies identifying the most necessary competencies and approaches for training professionals of various health care disciplines [13,14]. There are now calls for interventions to improve the digital skills of service users who experience barriers to telehealth [12]. Specifically, researchers have explicitly called for training and education on the use of telehealth to increase access for underserved populations [15,16]. This training could have the potential to narrow the widening disparities in telehealth access yet remains an understudied area of research.

This review aimed to examine the extent to which training and education activities for service users in telehealth are reported in the literature, in addition to summarizing the findings of this research. The review also aimed to identify gaps in the literature and determine future research needs.

Methods

Approach

A scoping review was identified as the most relevant method to answer the review question [17], as the research question was exploratory and aimed to map the breadth and heterogeneity of the literature. This review design was chosen to provide an initial idea of the size and nature of the available research and to identify gaps in the existing literature [17,18].

The Joanna Briggs Institute guidelines for scoping reviews [19] were followed and the review was reported in line with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist (Multimedia Appendix 1 [20]). The 5-step scoping review methodologies by Arksey and O'Malley [18] and Levac et al [21] were followed. The five steps were (1) to identify the research question; (2) to identify relevant studies; (3) to select studies; (4) to chart the data; and (5) to collate, summarize, and report results. This scoping review was not registered. The protocol was published on the Open Science Framework website [22].

Step 1: Identifying the Research Question

The scoping review aimed to answer the following question: What is the nature and extent of the literature related to training and education activities for service users in telehealth? The review aimed to provide a comprehensive overview of the breadth and heterogeneity of published research on the provision of training for service users in telehealth. The scoping review also aimed to identify gaps in the literature, limitations of the research, and directions for future research. The specific aims were to (1) summarize the research on training and education activities for service users in telehealth and (2) summarize the content and main features of these activities.

Step 2: Data Source and Search Strategy

The search strategy aimed to locate published studies on the topic of telehealth training for service users. Initial searches were conducted on PubMed and Google Scholar to identify potentially relevant studies. The terms and keywords of these studies were used to create the search strategies. The search strategies were piloted to identify the most appropriate search terms and subsequently were adapted to the parameters of each database. The lead author (EG) conducted the searches in June 2023 of the following databases: MEDLINE (via PubMed), Embase, CINAHL, PsycINFO, and Web of Science. The full search strategies can be seen in Multimedia Appendix 2. Only studies published in English were included. Backward and forward citation tracking was conducted by the lead author (EG).

Step 3: Eligibility Criteria and Screening Process

EndNote X9 (Clarivate) was used to screen and organize the records. The following eligibility criteria were applied when screening the studies.

Participants

Adult populations, including parents and caregivers, were eligible for inclusion. Child populations (younger than 18 years) were excluded, due to an absence of relevant studies found in the initial searches.

XSL•FO
RenderX

Concept

Telehealth was defined as the provision of health care services to service users by health care professionals, through live, synchronous audio and video consultations. Studies that focused on training adult service users on how to use telehealth, including how to set up telehealth, were included. Studies that focused on how to set up telemonitoring devices (eg, heart monitors and spirometers) only were excluded. Studies that focused on training users to use patient portals, unless there was a telehealth consultation component, were excluded. Studies that focused on health care professional education only were excluded.

Context

Studies of service users in any setting were eligible for inclusion.

Types of Sources

Studies of all research designs, including quantitative, qualitative, and mixed methods studies, were eligible for inclusion. Systematic reviews and literature reviews were excluded. Text and opinion papers were eligible for inclusion if they included adequate information on education and training activities. Conference abstracts were excluded, as they did not provide sufficient information required for this scoping review.

Study or Source of Evidence Selection

The lead author (EG) screened the titles and abstracts of all publications identified in the database searches. Two reviewers (EG and JH or SC) screened the full texts of the papers for inclusion. Any disagreements were resolved by discussion. Reasons for exclusion of studies at full-text screening were recorded and reported. The results of the search and the study inclusion process were presented in a PRISMA-ScR flow diagram [19].

Step 4: Data Extraction

A data extraction tool was developed using Microsoft Excel. The extracted data included author, year of publication, aim of study, sample, description of training, outcome measures, and key findings. Papers were excluded at this stage if the reviewers agreed there was insufficient data on the topic. The lead author (EG) conducted data extraction of all included papers.

Step 5: Data Synthesis

As recommended by Arksey and O'Malley [18], the quality of the evidence was not assessed. A descriptive narrative synthesis and numerical summaries were conducted to present the findings.

Results

Study Characteristics

The database searches yielded 13,997 records. After removing duplicates and applying the eligibility criteria, 13 papers were included in the review [23-35] (see Figure 1 for the PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] flow diagram). All 13 studies were published between 2020 and 2023. All studies were conducted in the United States and were published as journal papers. Study designs included uncontrolled preintervention and postintervention studies (n=4, 31%), postintervention studies (n=4, 31%), brief communications (n=4, 31%), and a retrospective cohort study (n=1, 8%). Studies were conducted in academic medical centers (n=5, 39%), geriatrics clinics (n=3, 23%), community settings (n=3, 23%), a specialty care clinic (n=1, 8%), and a federally qualified health center (n=1, 8%). A total of 12 (92%) studies were conducted during the COVID-19 pandemic, in the context of the rapid implementation of telehealth at this time. In total, 9 (69%) of the training programs were evaluated empirically, using mixed methods (n=5, 39%) and quantitative methods (n=4, 31%). The remaining 4 (31%) studies described training activities in "brief communication" style papers detailing their institution's wider move to telehealth during the pandemic. The full characteristics of the included studies can be seen in Multimedia Appendix 3 [23-35].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for scoping reviews.

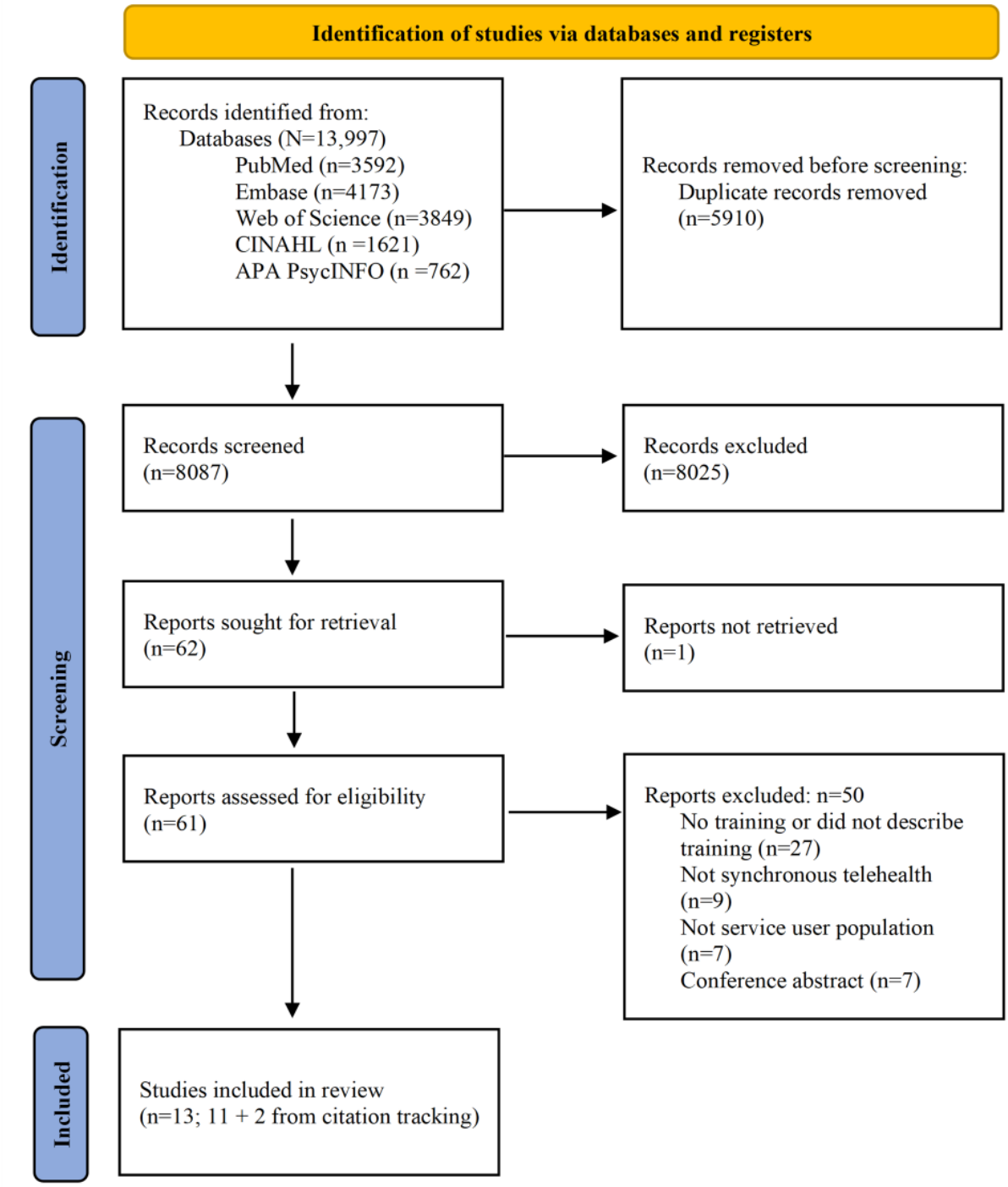

Service User and Trainer Characteristics

Of the 13 studies, the target populations included older adults (n=7, 54%), patients in hospital outpatient settings (age unspecified; n=4, 31%), and adult patients (n=2, 15%). In the 6 studies that reported the demographic information of service users, the majority of participants were female (55%-79%) [23,24,26,29,31,35] and predominantly White (42%-79%) [24,26,29,31,35]. In 1 of the 6 studies, the population was predominantly Black (47%) [23]. A total of 3 (23%) of the 13 studies asked participants about access to technological devices and the internet. Jezewski et al [29] reported that 93% of participants had access to a telehealth-compatible device. Antonio et al [23] reported that 60% of participants had access to a laptop and 99% had access to a cell phone. Hawley et al [27] reported that 64% of participants had access to both the internet and an internet-compatible device with a camera.

The training was delivered primarily by medical or health professions student volunteers (n=7, 54%), followed by hospital staff (n=3, 23%), volunteer trainers (n=1, 8%), graduate students (n=1, 8%), and a member of the research team (n=1, 8%). A total of 4 (31%) studies described the training the trainers undertook, and 5 (39%) studies described that the trainers had materials to assist them such as scripts [24,26] and instructions [23,30,32]. Two training initiatives were facilitated by partnerships with community organizations [29,31].

XSL•FO
RenderX

Training Format, Delivery, and Duration

One-on-one phone calls prior to the telehealth visit were the most commonly used training formats (n=9, 69%). Other training formats included a Microsoft PowerPoint presentation (n=1, 8%), a one-on-one video call (n=1, 8%), a prerecorded video (n=1, 8%), and web-based modules (n=1, 8%). Many studies (n=9, 69%) also provided written instructional documents to participants, primarily on how to download telehealth applications. In addition, 1 (8%) study also included a video including actors that simulated a telehealth consultation, as part of the web-based modules [35]. Phone was the most commonly used delivery modality (n=9, 69%), followed by web-based (including modules, video calls, and prerecorded video; n=3, 23%), and a mix of in-person and paper-based (n=1, 8%). Studies additionally used email (n=5, 39%) and SMS text messaging (n=3, 23%) to communicate with and send materials and web links to participants.

In terms of duration, most of the studies were once-off phone calls (n=8, 62%). The in-person PowerPoint presentation was 20 minutes long but moved to a self-paced paper-based presentation because of COVID-19 restrictions [29]. The web-based modules described by Taylor et al [35] were also self-paced. One training program lasted 1 to 2 hours over 7 sessions, across 2 months in total [31]. Pichan et al [32] described conducting 3 phone calls over the course of 1 week. The prerecorded video described by Spindler et al [34] was less than 2 minutes in duration. A summary of training characteristics is presented in Figure 2.

Figure 2. Summary of training characteristics.

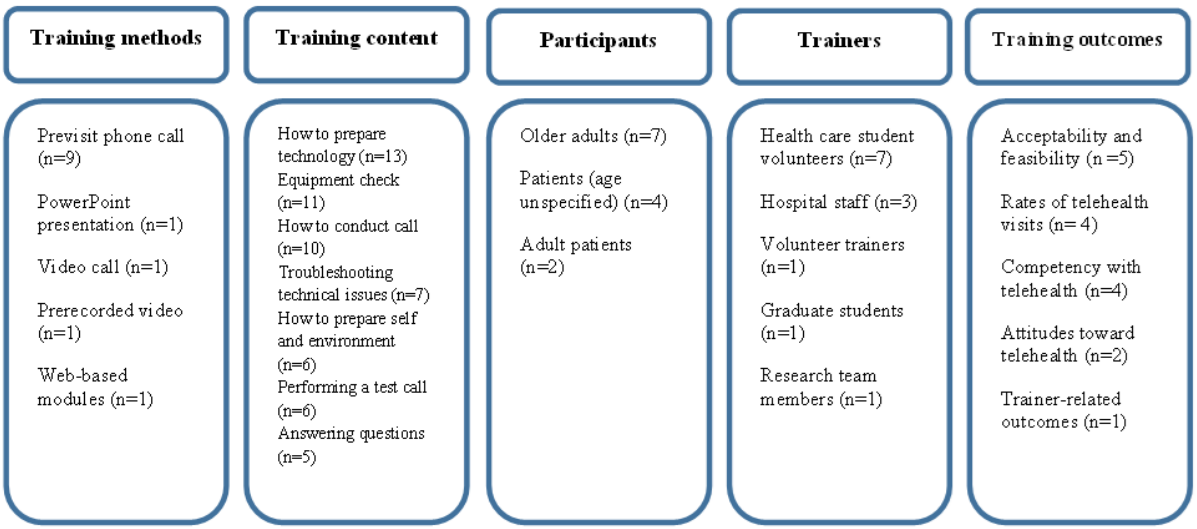

Training Objectives and Content

All studies aimed to train service users on how to use telehealth, primarily to prepare them for upcoming video visits. Additional training-related aims were to reduce the cognitive load demands of telehealth [23], develop health professions students' communication skills [25], and improve the broader digital literacy of service users [31]. The content of the training activities primarily focused on helping service users set up for a video call. All studies provided training on how to prepare the technology including downloading video applications. A majority of 11 (85%) studies described checking with service users that they had the necessary equipment to conduct video calls. A total of 10 (77%) studies described providing guidance on how to conduct a video telehealth call, while 6 (46%) studies described guiding service users on how to prepare themselves or their environment for a video telehealth call. In total, 7 (54%) studies described providing troubleshooting training for technical issues with participants, while 6 (46%) studies described performing a test video call with the service users. A total of 5 (39%) studies described answering service user queries about telehealth.

In total, 4 (31%) studies described explaining to service users what telehealth is, with 2 (15%) studies reported describing to service users how telehealth can be used. A total of 3 (23%) studies provided guidance on how to ensure safety and privacy while using telehealth, with only 1 (8%) study describing the limitations of telehealth [35]. Finally, only 1 (8%) study focused on teaching service users about elements of care during the telehealth visit, including care planning in telehealth, forming a therapeutic relationship, and team-based care in telehealth [35].

Development of Training

In total, 7 (54%) studies reported details on how the training activities were developed. A total of 2 studies described developing the training using the theoretical underpinnings of cognitive load theory [23] and provider telehealth training [35], respectively. A total of 2 studies [29,31] described partnering with community organizations to develop training programs. Hawley et al [27] described categorizing patients into 4 phenotypes based on their interest and capability to complete a home telehealth visit and subsequently creating training to overcome patient-perceived barriers. Pichan et al [32] described that medical students developed the training program, and Gulati et al [25] reported that the geriatrics faculty lead provided guidance on the design of the program.

Reasons for Not Participating or Dropping Out

A total of 4 (31%) studies described reasons why people did not accept training or dropped out of the training program.

XSL•FO
RenderX

Reasons for declining assistance or training included lack of interest in telehealth [23,24], did not feel they needed assistance [24], had already canceled their appointment [24], lack of device [23], and having existing telehealth experience [23]. Reasons for dropping out of training programs included because of health issues [31], schedule conflicts [31], participants not responding [31], and internet problems [31]. Taylor et al [35] reported that lower median household income was significantly associated with lower completion of the training program.

Evaluation of Training Programs

Overview

Studies used a number of methods to examine the outcomes of the training programs. A total of 5 (39%) studies examined qualitative outcomes and 9 (69%) studies examined quantitative outcomes. Studies examined the impact of training on rates of telehealth video visits, telehealth competency and confidence, attitudes toward telehealth, acceptability and feasibility of training, and trainer-related outcomes.

Rates of Telehealth Video Visits

In total, 4 (31%) studies examined the impact of training on rates of telehealth video visits [23,24,26,32]. Gusdorf et al [26] reported that a training call significantly increased the likelihood of a successful video visit. Chu et al [24] reported that the majority of participants who received training (76.5%) were successfully video enabled and that those who declined training had the highest rate of video visit cancellation. Pichan et al [32] reported that rates of video visits increased from before, to after, participating in training. Finally, Antonio et al [23] reported that there was no difference in video visit rates between intervention and nonintervention participants following training and that 40% of intervention participants who were scheduled for video visits ultimately had a phone visit.

Competency and Confidence With Telehealth

Changes in competency and confidence with telehealth were examined in 4 (31%) studies [23,29,31,35]. Taylor et al [35] reported significant improvements in perceived telehealth competency following training. Neumann et al [31] found a significant increase in the mean perceived confidence level for engaging in video visits following training. Antonio et al [23] found no significant differences in self-efficacy or perceived difficulty in using video visits between intervention and nonintervention participants following training. Jezewski et al [29] reported that 36% of participants were familiar with telehealth before training, and after training, 70% of participants understood how to access telehealth. However, 21% of participants reported wanting more information about telehealth after training.

Attitudes Toward Telehealth

A total of 2 (15%) studies examined the impact of training on telehealth attitudes [23,29]. Jezewski et al [29] reported that 39% of participants would use telehealth following training. The authors did not report what this value was before training. In Antonio et al [23], intervention participants were significantly less satisfied with their video visit experience than nonintervention participants following training.

Acceptability and Feasibility of Training

A total of 5 (39%) studies examined the acceptability and feasibility of training [23,27,31,32,35]. Participants in a number of studies appreciated the human, one-on-one elements of training. Antonio et al [23] reported that participants appreciated the communication skills of their trainer and the relationship and rapport they had with the trainer. Similarly, Neumann et al [31] reported that participants valued the patience and reinforcement from the trainers and described having personal relationships with them. This view was mirrored by participants in Pichan et al [32], who appreciated the time the trainers put into helping them.

Certain elements of the training were valued and helpful to participants. Antonio et al [23] reported that participants appreciated the step-by-step guidance offered in the training sessions. Similarly, participants in Antonio et al [23] valued the structure of the training, such as the repetition and self-pacing. Participants in Neumann et al [31] appreciated having access to the devices and training booklets. Having a partnership with a trusted community organization was also appreciated by participants [31]. Pichan et al [32] solicited feedback from providers, who expressed positive views, highlighting that the training program increased their video visits and that patients felt empowered to learn new skills and appreciated the training sessions. Similarly, Antonio et al [23] reported that participants valued being able to access health services and they described having a sense of purpose and control, and not having to rely on family members to help. Hawley et al [27] reported that participants found the training and instructions helpful.

Of the 13 included studies, 4 (31%) studies reported service users' and trainers' challenges relating to the telehealth training programs [23,31,32,35]. The challenges described by learners participating in telehealth training included the accessibility of training [35], concerns about the privacy of the website hosting the training [35], technological security [31], pandemic (isolation) [31], health status and care change [31], and working with older devices [23,32]. Other challenges included adjusting sound and visual settings on devices [23,32], internet access problems [23], switching between applications on small devices [23], people in the background impacting training [23], and patients having to log in to their patient portal to access the telehealth video platform [23]. Taylor et al [35] revised their training modules based on participant feedback. From the clinician's perspective, the challenges included the time-consuming task of going through their patient list and compiling the patient's information for their volunteer trainer [32].

Trainer-Related Outcomes

Gulati et al [25] examined outcomes relating to the trainer who provided the telehealth training calls. They reported that health professions students experienced improvements in their health communication skills.

XSL•FO
RenderX

Discussion

Principal Findings

This scoping review identified 13 studies that described training service users in the use of telehealth. All of the studies were conducted in the United States. The target training groups were primarily older people and patients in academic medical centers. The most commonly reported training method involved a preparatory phone call before a telehealth visit conducted by student volunteers, accompanied by written instructions. The content of training activities was similar across studies, primarily focused on assisting people to download and set up applications and devices. All but 1 of the studies were conducted during the COVID-19 pandemic, with training initiatives conducted to facilitate the rapid move to telehealth during this time.

The studies were primarily quantitative and mixed methods. Training initiatives were evaluated using a variety of methods, including examining video visit rates, conducting pretraining and posttraining surveys, and eliciting qualitative feedback. Some studies reported that telehealth training increased the amount of people participating in video visits. The evidence was limited and mixed about whether telehealth training increases service users' perceived competency with telehealth. The available evidence suggested that training did not improve participants' attitudes toward telehealth. Participants discussed the positive elements and challenges of training programs.

Comparison With Prior Work

To the authors' knowledge, this is the first scoping review to map the literature on training activities for service users in the use of synchronous telehealth. A previous review by Grossman et al [36] explored the impact of interventions to increase patient portal use. They found that individually focused interventions and technical training and assistance had the best evidence for increasing portal use. In our review, individually focused training and technical assistance were the most common elements of telehealth training. However, there was limited evidence on the effectiveness of telehealth training to improve telehealth competency. In addition, some studies reported increased rates of video visits, but it is difficult to ascertain if training alone was responsible for these increases, due to the concurrent increase in the use of video visits during the COVID-19 pandemic.

Much of the previous research on telehealth training has focused on training health care professionals and students on the use of telehealth. Studies have aimed to identify what competencies are needed for health care professionals to be proficient in telehealth. For example, Davies et al [37] conducted a Delphi study to develop a framework for physiotherapists to deliver quality care via telehealth. The framework consisted of 60 capabilities across 7 domains: compliance, patient privacy and confidentiality, patient safety, technology skills, telehealth delivery, assessment and diagnosis, and care planning and management. In this review, technology skills were the primary domain covered in the included studies. The service user training activities focused primarily on the technical elements of setting up and using telehealth. Much fewer studies focused on training in nontechnical elements such as preparing oneself for a visit, getting the most out of a visit, building a therapeutic relationship, and information about the privacy and security of telehealth. Many of the studies were conducted in the context of the COVID-19 pandemic where the priority was moving service users onto telehealth quickly, so it is likely that focusing on downloading and installing the telehealth applications was the priority of institutions.

Implications for Practice and Research

The findings of this review highlight lessons for future training activities for service users in the use of telehealth. In total, 2 studies described partnering with community organizations to provide telehealth training to service users. Partnering with a trusted community organization was appreciated by service users [31] and may be a potential method for facilitating telehealth training. One limitation of the included studies was that there may be a geographical bias toward participants recruited from universities, academic medical centers, and metropolitan towns and cities [32,35]. Partnering with a community organization could increase reach to diverse geographical areas and ensure people are trained in other digital health skills (eg, using patient portals and finding health information on the internet), to improve the broader digital literacy of service users.

As described in 1 of the included studies [35], consulting with service users to develop the training could be a promising method to ensure that telehealth training meets their needs. Specifically, older adults in this study made recommendations to make the training modules more accessible including changes in speech, format, and language use [35]. Considering the accessibility of training for different groups is also important. For example, using closed captioning of videos, or providing word documents that can be read by screen readers, would help to improve the accessibility of training and ensure that people with disabilities are not excluded. Finally, in another study [24], service users declined assistance as they did not have a device available for video visits. To overcome this, Neumann et al [31] gave refurbished iPads to participants. Offering devices to service users or providing the option of telephone visits would ensure that those who do not have access to devices are not excluded from telehealth training or visits.

A number of gaps were identified in the included studies that could be addressed with further research. First, all studies were conducted in the United States. This presents an opportunity for further research to be conducted in more diverse geographical contexts, particularly in low- and middle-income countries where telehealth implementation is not as widespread [38]. Second, the description of training in some of the studies was vague, particularly in the studies that discussed training within their institution's wider transition to telehealth during the pandemic. Future studies should aim to provide more detail about who conducted the training, how it was developed, and what topics were covered. Making the training materials publicly available would be useful for institutions and organizations developing telehealth training programs. In addition, studies did not report on the costs needed to implement the training activities. This is important information for institutions evaluating the feasibility and effectiveness of an intervention,

XSL•FO
RenderX

so future research should report on the cost-effectiveness of training initiatives.

One of the main findings was that there was mixed evidence that training improved service users' perceived competency with telehealth. A total of 2 studies reported significant improvements in perceived telehealth competency and confidence from before, to after, training. However, it is important to note that these studies did not employ a non-training control group. There is a need to conduct further quantitative and qualitative evaluations of training initiatives. Understanding the effectiveness, acceptability, and feasibility of different training initiatives would help to inform best practices for training service users in telehealth. In addition, future research should use validated measures to examine changes in telehealth competency. Furthermore, examining the effectiveness of human-led training compared to self-directed training could be a promising avenue of research, given the anecdotal appreciation of human assistance in the included studies.

Research on the impact of training activities outside the context of the COVID-19 pandemic is also needed. All but 1 of the studies were conducted in response to the need for telehealth implementation during this time, which may have positively impacted participants' interest in partaking in training. In addition, some studies reported increased rates of video visits following training, which may have been confounded by the increased availability of telehealth during this time. Research is needed to investigate if these effects are observed outside the pandemic context. Finally, many of the training activities lacked an educational or theoretically informed basis, so future research should attempt to use theory- and literature-informed training, perhaps consulting the body of research on training health care providers on the use of telehealth.

Strengths and Limitations

A key strength of this review is that it helps to address a gap in the literature of a pertinent research area. The review aimed to describe the existing literature and identify differences and commonalities between training activities. A further strength of the review is that it is reported in line with evidence-based criteria for conducting scoping reviews developed by the Joanna Briggs Institute [19] and the guidelines for scoping reviews described in the PRISMA-ScR checklist [20]. Finally, a comprehensive search of 5 major academic databases was conducted, in addition to backward and forward citation tracking.

Given the nature of scoping reviews, we did not conduct a critical appraisal of the studies included in the review. With the various terms used to describe telehealth, we may have missed some eligible studies. Another limitation of the review is that we did not use other avenues for searching papers such as gray literature searches and contacting key authors in the area. Finally, with the recent, increased use of telehealth, it is likely that further research on this topic will be published in the coming years. There will be a need for an update of this scoping review, particularly to identify studies conducted beyond the context of the pandemic.

Conclusions

This review aimed to map the literature on training activities for service users in telehealth. The common features of training initiatives included a preparatory phone call, volunteer trainers, and the provision of training on how to download and set up telehealth applications. The target populations consisted primarily of older adults. There was limited and mixed evidence that training improved service users' perceived competency with telehealth. Future research should focus on empirically evaluating training activities in geographically diverse settings. The review highlights the need for co-designed training and initiatives to improve the broader digital literacy of service users.

Acknowledgments

EG is funded by a Clement Archer PhD Scholarship from the School of Pharmacy and Biomolecular Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland.

Authors' Contributions

All authors contributed to conceptualization and methodology. EG devised the search strategy, conducted the database searches, performed title and abstract screening, and extracted the data. EG, JH, and SC conducted the full-text screening. EG drafted the manuscript. All authors reviewed and approved the final version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 39 KB - jmir_v26i1e57586_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Search strategy.
[DOCX File , 28 KB - jmir_v26i1e57586_app2.docx]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File , 36 KB - jmir_v26i1e57586_app3.docx]

References

1. Digital health in the European region: the ongoing journey to commitment and transformation. World Health Organization. 2023. URL: https://www.who.int/andorra/publications/m/item/digital-health-in-the-who-european-region-the-ongoing-journey-to-commitment-and-transformation [accessed 2024-07-01]
2. Hong JS, Sheriff R, Smith K, Tomlinson A, Saad F, Smith T, et al. Impact of COVID-19 on telepsychiatry at the service and individual patient level across two UK NHS mental health trusts. Evid Based Ment Health 2021;24(4):161-166 [FREE Full text] [doi: 10.1136/ebmental-2021-300287] [Medline: 34583940]
3. Sugarman DE, Horvitz LE, Greenfield SF, Busch AB. Clinicians' perceptions of rapid scale-up of telehealth services in outpatient mental health treatment. Telemed J E Health 2021 Dec;27(12):1399-1408 [FREE Full text] [doi: 10.1089/tmj.2020.0481] [Medline: 33600272]
4. Roberts ET, Mehrotra A. Assessment of disparities in digital access among Medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020;180(10):1386-1389 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]
5. Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing equity in telemedicine for chronic disease management during the Covid-19 pandemic. NEJM Catalyst 2020;1(3):1-13 [FREE Full text] [doi: 10.1056/CAT.20.0123]
6. Le TV, Galperin H, Traube D. The impact of digital competence on telehealth utilization. Health Policy Technol 2023;12(1):100724. [doi: 10.1016/j.hlpt.2023.100724]
7. Scheerder A, van Deursen A, van Dijk J. Determinants of internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telemat Inform 2017;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]
8. Shaw J, Brewer LC, Veinot T. Recommendations for health equity and virtual care arising from the COVID-19 pandemic: narrative review. JMIR Form Res 2021;5(4):e23233 [FREE Full text] [doi: 10.2196/23233] [Medline: 33739931]
9. Eddison N, Leone E, Healy A, Royse C, Chockalingam N. The potential impact of allied health professional telehealth consultations on health inequities and the burden of treatment. Int J Equity Health 2022;21(1):91 [FREE Full text] [doi: 10.1186/s12939-022-01689-2] [Medline: 35773695]
10. UK Consumer Digital Index 2021. Lloyds Bank. 2021. URL: https://www.lloydsbank.com/assets/media/pdfs/banking_with_us/whats-happening/210513-lloyds-consumer-digital-index-2021-report.pdf [accessed 2024-06-25]
11. Gallegos-Rejas VM, Thomas EE, Kelly JT, Smith AC. A multi-stakeholder approach is needed to reduce the digital divide and encourage equitable access to telehealth. J Telemed Telecare 2023;29(1):73-78. [doi: 10.1177/1357633X221107995] [Medline: 35733379]
12. Cheng J, Arora VM, Kappel N, Vollbrecht H, Meltzer DO, Press V. Assessing disparities in video-telehealth use and eHealth literacy among hospitalized patients: cross-sectional observational study. JMIR Form Res 2023;7:e44501 [FREE Full text] [doi: 10.2196/44501] [Medline: 37171854]
13. Smith K, Torous J, Cipriani A. Teaching telepsychiatry skills: building on the lessons of the COVID-19 pandemic to enhance mental health care in the future. JMIR Ment Health 2022;9(10):e37939 [FREE Full text] [doi: 10.2196/37939] [Medline: 35358948]
14. Crawford A, Sunderji N, López J, Soklaridis S. Defining competencies for the practice of telepsychiatry through an assessment of resident learning needs. BMC Med Educ 2016;16:28 [FREE Full text] [doi: 10.1186/s12909-016-0529-0] [Medline: 26813286]
15. Ng BP, Park C. Accessibility of telehealth services during the COVID-19 pandemic: a cross-sectional survey of Medicare beneficiaries. Prev Chronic Dis 2021;18:E65 [FREE Full text] [doi: 10.5888/pcd18.210056] [Medline: 34197284]
16. Beller HL, Rapp DE, Noona SW, Winkelman AJ, Zillioux JM, Smith RP, et al. Tele-urology during COVID-19: rapid implementation of remote video visits. Urol Pract 2020;7(6):442-447. [doi: 10.1097/UPJ.0000000000000167] [Medline: 37287160]
17. Paré G, Trudel M, Jaana M, Kitsiou S. Synthesizing information systems knowledge: a typology of literature reviews. Inf Manag 2015;52(2):183-199. [doi: 10.1016/j.im.2014.08.008]
18. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
19. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]

XSL•FO
RenderX

20. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. Prisma Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
21. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
22. Galvin E. Training and educating service users in the use telehealth: a scoping review protocol. OSFHOME. 2023. URL: https://osf.io/2mza9/ [accessed 2024-07-03]
23. Antonio MG, Williamson A, Kameswaran V, Beals A, Ankrah E, Goulet S, et al. Targeting patients' cognitive load for telehealth video visits through student-delivered helping sessions at a United States Federally Qualified Health Center: equity-focused, mixed methods pilot intervention study. J Med Internet Res 2023;25:e42586 [FREE Full text] [doi: 10.2196/42586] [Medline: 36525332]
24. Chu JN, Kaplan C, Lee JS, Livaudais-Toman J, Karliner L. Increasing telehealth access to care for older adults during the COVID-19 pandemic at an academic medical center: Video Visits for Elders Project (VVEP). Jt Comm J Qual Patient Saf 2022;48(3):173-179 [FREE Full text] [doi: 10.1016/j.jcjq.2021.11.006] [Medline: 35027304]
25. Gulati A, Lalchandani P, Auchus I, Grandi J, Clelland E, Chen P. A student-led telephone-based clinical learning program for outreach to older adults. J Am Geriatr Soc 2022;70(12):E30-E33. [doi: 10.1111/jgs.18036] [Medline: 36527413]
26. Gusdorf RE, Shah KP, Triana AJ, McCoy AB, Pabla B, Scoville E, et al. A patient education intervention improved rates of successful video visits during rapid implementation of telehealth. J Telemed Telecare 2021;29(8):607-612. [doi: 10.1177/1357633X211008786] [Medline: 33975506]
27. Hawley CE, Genovese N, Owsiany MT, Triantafylidis LK, Moo LR, Linsky AM, et al. Rapid integration of home telehealth visits amidst COVID-19: what do older adults need to succeed? J Am Geriatr Soc 2020;68(11):2431-2439 [FREE Full text] [doi: 10.1111/jgs.16845] [Medline: 32930391]
28. Hoffman PE, London YR, Weerakoon TS, DeLucia NL. Rapidly scaling video visits during COVID-19: the ethos of virtual care at Yale Medicine. Healthcare 2020;8(4):100482 [FREE Full text] [doi: 10.1016/j.hjdsi.2020.100482] [Medline: 33129179]
29. Jezewski E, Miller A, Eusebio M, Potter J. Targeted telehealth education increases interest in using telehealth among a diverse group of low-income older adults. Int J Environ Res Public Health 2022;19(20):13349 [FREE Full text] [doi: 10.3390/ijerph192013349] [Medline: 36293929]
30. Meyer BC, Friedman LS, Payne K, Moore L, Cressler J, Holberg S, et al. Medical undistancing through telemedicine: a model enabling rapid telemedicine deployment in an academic health center during the COVID-19 pandemic. Telemed J E Health 2021;27(6):625-634. [doi: 10.1089/tmj.2020.0327] [Medline: 33030985]
31. Neumann AV, Gonzalez A, Walter LC, Rivera J. Improving older adults' telehealth through a novel community-academic partnership: preliminary data. J Am Geriatr Soc 2023 Dec;71(12):3886-3895. [doi: 10.1111/jgs.18556] [Medline: 37607098]
32. Pichan CM, Anderson CE, Min LC, Blazek MC. Geriatric Education on Telehealth (GET) Access: a medical student volunteer program to increase access to geriatric telehealth services at the onset of COVID-19. J Telemed Telecare 2021;29(10):816-824. [doi: 10.1177/1357633X211023924] [Medline: 34152885]
33. Smith WR, Atala AJ, Terlecki RP, Kelly EE, Matthews CA. Implementation guide for rapid integration of an outpatient telemedicine program during the COVID-19 pandemic. J Am Coll Surg 2020;231(2):216-222.e2 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2020.04.030] [Medline: 32360960]
34. Spindler M, Jacobs D, Yuan K, Tropea T, Teng CW, Perrone C, et al. A department approach to teleneurology. Telemed J E Health 2021;27(9):1078-1084. [doi: 10.1089/tmj.2020.0323] [Medline: 33337290]
35. Taylor S, Souza S, Little L, Odiaga J. Enhancing telehealth competency: development and evaluation of education modules for older adults. OTJR (Thorofare N J) 2023;43(3):478-486 [FREE Full text] [doi: 10.1177/15394492231153115] [Medline: 36757088]
36. Grossman L, Masterson Creber RM, Benda N, Wright D, Vawdrey D, Ancker J. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
37. Davies L, Hinman RS, Russell T, Lawford B, Bennell K, International Videoconferencing Steering Group. An international core capability framework for physiotherapists to deliver quality care via videoconferencing: a Delphi study. J Physiother 2021;67(4):291-297 [FREE Full text] [doi: 10.1016/j.jphys.2021.09.001] [Medline: 34521586]
38. Ong AKS, Kurata YB, Castro SAD, de Leon JPB, Dela Rosa HV, Tomines APJ. Factors influencing the acceptance of telemedicine in the Philippines. Technol Soc 2022;70:102040. [doi: 10.1016/j.techsoc.2022.102040]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.02.24; peer-reviewed by J Mistry, MH Pourasad; comments to author 07.05.24; revised version received 08.05.24; accepted 23.05.24; published 31.07.24.

Please cite as:
Galvin E, Desselle S, Gavin B, McNicholas F, Cullinan S, Hayden J
Training Service Users in the Use of Telehealth: Scoping Review
J Med Internet Res 2024;26:e57586
URL: https://www.jmir.org/2024/1/e57586
doi:10.2196/57586
PMID:

©Emer Galvin, Shane Desselle, Blánaid Gavin, Fiona McNicholas, Shane Cullinan, John Hayden. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Assessing the Effectiveness of eHealth Interventions to Manage Multiple Lifestyle Risk Behaviors Among Older Adults: Systematic Review and Meta-Analysis

Beibei Shi[1], MD; Guangkai Li[1], MD; Shuang Wu[1], MD; Hongli Ge[1], MD; Xianliang Zhang[1], PhD; Si Chen[2], PhD; Yang Pan[1], PhD; Qiang He[1], PhD

[1]School of Physical Education, Shandong University, Jinan, China

[2]School of Nursing and Rehabilitation, Cheeloo College of Medicine, Shandong University, Jinan, China

Corresponding Author:
Qiang He, PhD
School of Physical Education, Shandong University
17922 Jingshi Road, Lixia District
Shandong Province
Jinan, 250061
China
Phone: 86 0531 88395374
Fax: 86 15065092263
Email: hq@sdu.edu.cn

Abstract

Background: Developing adverse lifestyle behaviors increases the risk of a variety of chronic age-related diseases, including cardiovascular disease, obesity, and Alzheimer disease. There is limited evidence regarding the effectiveness of eHealth-based multiple health behavior change (MHBC) interventions to manage lifestyle risk behaviors.

Objective: The purpose of this systematic evaluation was to assess the effectiveness of eHealth MHBC interventions in changing ≥2 major lifestyle risk behaviors in people aged ≥50 years.

Methods: The literature search was conducted in 6 electronic databases—PubMed, Embase, Web of Science, Scopus, Cochrane Library, and SPORTDiscus—from inception to May 1, 2024. Eligible studies were randomized controlled trials of eHealth interventions targeting ≥2 of 6 behaviors of interest: alcohol use, smoking, diet, physical activity (PA), sedentary behavior, and sleep.

Results: A total of 34 articles with 35 studies were included. eHealth-based MHBC interventions significantly increased smoking cessation rates (odds ratio 2.09, 95% CI 1.62-2.70; $P<.001$), fruit intake (standardized mean difference [SMD] 0.18, 95% CI 0.04-0.32; $P=.01$), vegetable intake (SMD 0.17, 95% CI 0.05-0.28; $P=.003$), self-reported total PA (SMD 0.22, 95% CI 0.02-0.43; $P=.03$), and objectively measured moderate to vigorous PA (SMD 0.25, 95% CI 0.09-0.41; $P=.002$); in addition, the interventions decreased fat intake (SMD –0.23, 95% CI –0.33 to –0.13; $P<.001$). No effects were observed for alcohol use, sedentary behavior, or sleep. A sensitivity analysis was conducted to test the robustness of the pooled results. Moreover, the certainty of evidence was evaluated using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework.

Conclusions: eHealth-based MHBC interventions may be a promising strategy to increase PA, improve diet, and reduce smoking among older adults. However, the effect sizes were small. Further high-quality, older adult–oriented research is needed to develop eHealth interventions that can change multiple behaviors.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42023444418; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023444418

(J Med Internet Res 2024;26:e58174) doi:10.2196/58174

KEYWORDS

eHealth; lifestyle risk behaviors; older adults; multiple health behavior change; mobile phone

XSL•FO
RenderX

Introduction

Background

The aging population and the rise in chronic diseases will significantly increase health care expenditures; in the United States, for example, it is estimated that 80% of older adults have 1 chronic disease and 50% have at least 2 [1]. Up to 80% of global cases of heart disease, stroke, and type 2 diabetes, as well as >30% of cancers, could be prevented by reducing tobacco and harmful alcohol use and improving diet and physical activity (PA) [2]. Key indicators of health risk behaviors in older adults include physical inactivity, eating <5 portions of fruits and vegetables per day, obesity, and current tobacco use. The American Heart Association's recently published Life's Essential 8 metrics include healthy sleep as an indicator of cardiovascular health on top of a healthy diet, participation in PA, and avoidance of nicotine [3]. Similarly, many health guidelines now integrate recommendations for PA, sleep, and sedentary behavior (SB) [4,5]. Overall, there is a clear link between these risk factors and poor health outcomes and quality of life [6,7]. Lifestyle behaviors, as important factors in health, provide scientists with valuable and interesting areas of research [8,9].

Engaging in multiple risk behaviors can negatively impact health by increasing the risk of, for example, chronic disease and mortality [10]; moreover, risk behaviors often occur simultaneously. A previous study reported a prevalence rate of approximately 50% for the co-occurrence of unhealthy diet and physical inactivity in adults [11]. Compared to those who do not engage in any of the 4 lifestyle risk behaviors, those who do so face an elevated mortality risk equal to an additional 14 years of aging in both healthy populations [12] and populations with disease [13,14]. This suggests that it may be beneficial to use a holistic intervention approach to collectively change multiple health behaviors rather than individually change a single behavior. Multiple health behavior change (MHBC) interventions have attracted increased attention of late [15,16]. Growing evidence suggests that lifestyle interventions targeting MHBC may have a greater impact on public health than interventions targeting single health behavior change (SHBC) [17,18]. The advantages of MHBC interventions include maximizing health benefits and greater reduction in medical costs [19]; in addition, successfully modifying 1 behavior may increase confidence or motivation to change other health behaviors [16]. Existing systematic reviews have examined the effectiveness of behavioral interventions on multiple health risk behaviors [20].

eHealth refers to the use of information and communication technologies in health and health-related fields to enhance health care services; health surveillance; health literature; and health education, knowledge, and research [21]. There are several advantages over traditional face-to-face interventions [22]. An eHealth intervention can be delivered over long distances; is cost-effective, efficient, and highly accessible; and allows for easy data collection [22,23]. This makes eHealth interventions potentially powerful and scalable tools and enables eHealth-based MHBC interventions to improve ongoing adherence to chronic disease management. However, only a few systematic reviews have specifically examined the effectiveness of eHealth-based MHBC interventions in adult populations; for example, the review by Norman et al [24] focused on dietary behavior change and PA interventions, while Oosterveen et al [25] examined interventions targeting smoking, nutrition, alcohol use, and PA. Overall, previous reviews have included a limited number of health behaviors (<4), focusing primarily on PA and diet.

Objectives

Although studies on eHealth-based MHBC interventions aimed at promoting lifestyle changes in adolescents [26] and adults have been published, to the best of our knowledge, there are no knowledge reviews of such interventions designed for the older population. Older adults have limited access to programs and services that promote healthy lifestyles (eg, gyms) compared to younger adults [27]. eHealth interventions may be the solution to help individuals adopt and maintain a healthy lifestyle. Therefore, it is urgent to investigate whether eHealth can be used to implement evidence-based lifestyle changes. In this review, we targeted studies involving adults aged ≥50 years because this age is often associated with retirement, providing more time and energy to focus on health, and improved quality of life is critical for older adults. A previous systematic review found significant effects of eHealth interventions on PA in older adults; however, lifestyle risk behaviors other than PA were not part of the inclusion criteria, and the findings were inconclusive due to an insufficient number of studies [28]. Engaging in multiple lifestyle risk behaviors increases the risk of chronic disease and all-cause mortality more than the cumulative effect of a single behavior [29,30]. Therefore, we aimed to systematically review the effectiveness of eHealth interventions for risk behavior change in older adults, targeting ≥2 of 6 behaviors of interest: alcohol use, smoking, diet, PA, SB, and sleep.

Methods

Study Design

This systematic review and meta-analysis was registered with PROSPERO (CRD42023444418) and was reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (refer to Multimedia Appendix 1 for the PRISMA checklist) [31]. In addition, this systematic review and meta-analysis was conducted according to the recommendations of the Cochrane Handbook for Systematic Reviews of Interventions [32].

Search Strategy

We conducted a literature search from inception to May 1, 2024, in 6 databases: PubMed, Embase, Cochrane Library, Scopus, Web of Science, and SPORTDiscus. Following the PICOS (Population, Intervention, Comparison, Outcomes, and Study Design) principles and recommendations from the Cochrane Collaboration, we designed the search strategy using Medical Subject Headings terms, text word searches, and Boolean logic (Multimedia Appendix 2). In addition, we incorporated keywords, titles, or abstract terms, including but not limited to

XSL•FO
RenderX

"older adults," "health behavior," "risk reduction behavior," "ehealth," "mobile health," and "telemedicine." The search was restricted to randomized controlled trials (RCTs), and the studies included were limited to those published in English. Furthermore, we manually reviewed the reference lists of the retrieved studies and identified and obtained other relevant literature.

Study Eligibility Criteria

The study eligibility criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Population: mean age: ≥50 years
- Intervention: eHealth is a major component of the intervention
- Comparison: included no intervention, education as usual, or an alternative evidence-based intervention not delivered via eHealth (eg, face-to-face)
- Outcome: targeted ≥2 of the following behaviors: drinking, smoking, diet, physical activity, sedentary behavior, and sleep
- Study design: randomized controlled trials
- Language: English

Exclusion criteria

- Population: mean age: <50 years
- Intervention: eHealth component of the intervention is not predominant (eg, face-to-face interventions complemented by the use of a website)
- Comparison: any intervention delivered via eHealth component
- Outcome: targeted 1 of the following behaviors: drinking, smoking, diet, physical activity, sedentary behavior, and sleep
- Study design: qualitative studies, conference articles, letters, reviews, commentaries, protocols, or pilot studies
- Language: not English

Study Selection and Data Extraction

All titles and abstracts of the identified studies were downloaded and imported into EndNote X9 (Clarivate), and all duplicate articles were removed. To confirm whether the studies met the included criteria, 2 authors (BS and GL) worked independently to screen the titles and abstracts of the studies simultaneously. Next, they reviewed the full text of each paper based on the eligibility criteria. Any disagreements were resolved through discussion or consultation with a third author (QH).

Two authors (BS and GL) extracted data independently from the included studies and entered them into a predesigned data extraction form. The data extracted included the first author, the year of publication, the country where the study took place, study design, intervention frequency and duration, participant population, age, sample size, targeted risk behaviors, measurement tools, intervention types, intervention components, types of control groups, and theoretical basis or behavior change techniques.

Quality Assessment

Two reviewers (BS and GL) independently examined the methodological quality and risk of bias of the included studies using the risk-of-bias tool for randomized trials in the Cochrane Handbook for Systematic Review of Interventions (version 5.1.0). The tool includes 6 domains: selection bias (random sequence generation and allocation concealment), performance bias (blinding of participants and personnel), detection bias (blinding of outcome assessment), attrition bias (incomplete outcome data), reporting bias (selective reporting), and other bias (anything else, ideally prespecified). Each domain has 3 grades: low risk of bias, high risk of bias, and unclear risk of bias. Disagreements were resolved through discussion or consultation with a third author.

In addition, we used the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework to assess the quality of the body of evidence [33].

Data Synthesis and Analysis

For our meta-analysis, the outcomes were prevalence of alcohol use and smoking (*yes* or *no*); intake of fruit and vegetables (mean servings per day or mean portions per day), sugar intake, fat intake, fiber intake (percentage of total energy, grams per 1000 kilocalories, or grams per day), and energy intake (kilojoules per day or kilocalories per day); PA (accelerometer and self-report; minutes per day or minutes per week); sedentary time (minutes per day); and sleep (minutes per day or hours per day). For continuous outcomes, the corresponding variance was calculated using the preintervention and postintervention means and SDs; and for dichotomous outcomes, we extracted the preintervention and postintervention change values and sample sizes. However, if some studies had changes in baseline and postintervention data or if there were significant differences in their baseline data, we used the within-group difference in means and SDs for the intervention and control groups to calculate the effect size. Review Manager (version 5.3; The Cochrane Collaboration) was used to conduct the meta-analysis. We report continuous outcomes using standardized mean differences (SMDs) and dichotomous outcomes using odds ratios (ORs). We present these results using forest plots for each outcome of interest, with the weight (in percentage) indicating the influence of an individual study on the pooled result. The overall effect

XSL•FO
RenderX

difference was considered statistically significant if the 2-tailed P value was <.05. In addition, we assessed the statistical heterogeneity of the included studies using the I^2 statistic and P value. Data were pooled and analyzed with a fixed effects model if $I^2 \leq 50\%$ and $P>.10$; if these values were not met, which meant high heterogeneity among the studies, we used a random effects model to obtain more conservative estimates. We also conducted a sensitivity analysis using the leave-1-study-out method to test the robustness and reliability of the pooled results.

Results

Search Results and Study Selection

Across the 6 electronic databases, a total of 19,073 articles were initially retrieved from the literature search. Of these 19,073 articles, 4407 (23.11%) duplicates were removed. Of the remaining 14,666 articles, 14,379 (98.04%) were excluded after title and abstract screening, leaving 287 (1.96%) full-text articles. Of these 287 articles, 250 (87.1%) were excluded for reasons related to the eligibility criteria, leaving 37 (12.9%) studies for inclusion in the narrative synthesis. Using other methods, an additional 38 records were identified, of which 31 (81.6%) were excluded after title and abstract screening. Of the remaining 7 articles assessed for eligibility, 5 (71%) were excluded, leaving 2 (29%) studies for inclusion. The total number of articles included in the narrative synthesis was 39, of which 5 (13%) articles whose data could not be pooled for meta-analysis were excluded. Ultimately, 34 articles and 35 studies were included in the systematic review. The search and selection processes are shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the article selection process.

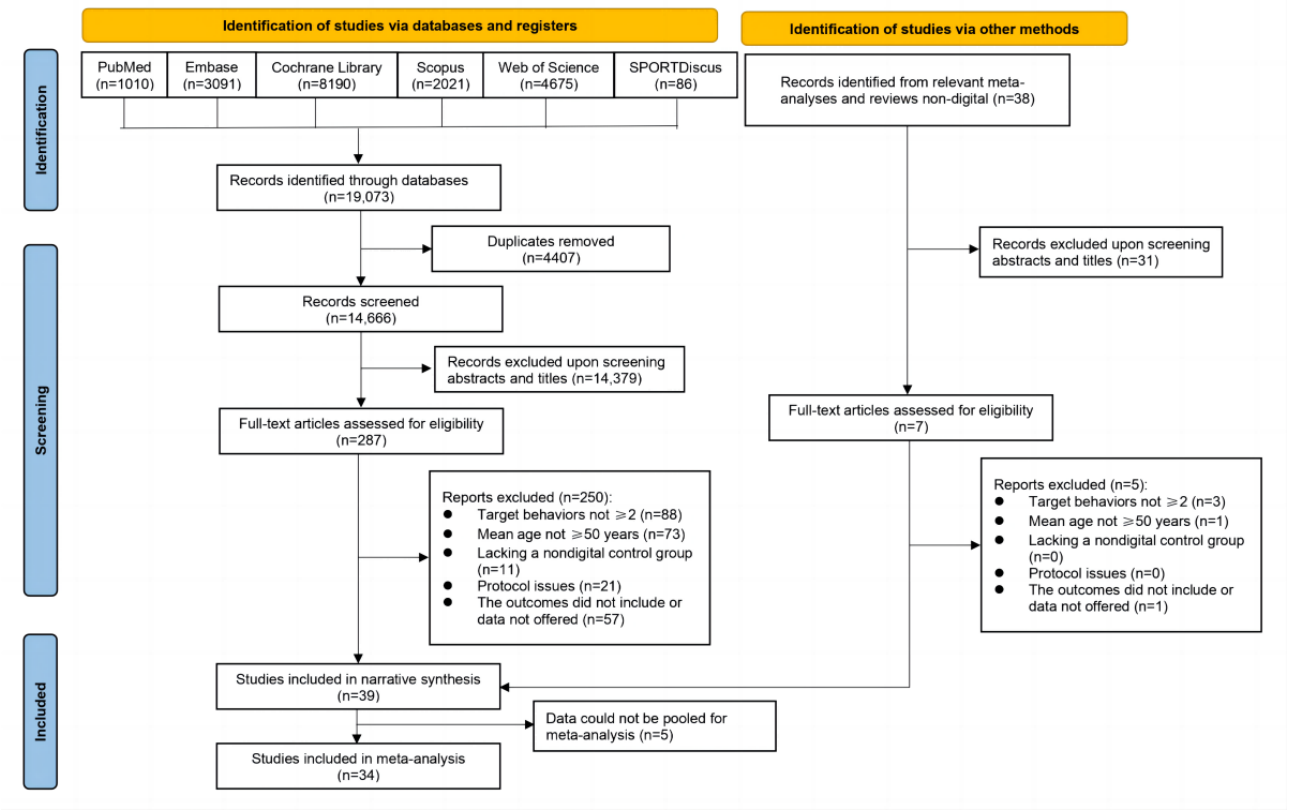

Description of the Included Studies

Study Characteristics

Table 1 presents the study characteristics of the 35 studies from the 34 included articles. All 35 studies are RCTs that were conducted between 2009 and 2024 in 15 countries: the United States (n=10, 29%), Australia (n=5, 14%), the United Kingdom (n=3, 9%), India (n=2, 6%), Canada (n=3, 9%), China (n=2, 6%), Belgium (n=2, 6%), Sweden (n=1, 3%), Germany (n=1, 3%), Italy (n=1, 3%), the Netherlands (n=1, 3%), South Korea (n=1, 3%), Israel (n=1, 3%), Japan (n=1, 3%), and Iran (n=1, 3%).

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Abu-Saad et al [34], 2019; Israel	Patients with type 2 diabetes mellitus	n=50; age: mean 53.0 (SD 7.6) y; 29.58% women	PA[a]; diet	2 arms; 12 mo; personalized information service	Windows software	Questionnaires	The intervention was based on a Windows platform designed as a multifunctional tool that collected data on habitual eating and PA behaviors, personalized by identifying lifestyle behaviors that needed to be improved as targets for education and behavior change, focusing and prioritizing the education and counseling process.	Standard lifestyle advice	Behavior change techniques
Babu et al [35], 2024; India	Patients with fatty liver disease	n=209; age: mean 60.4 (SD 10.0) y; 23.4% women	PA; diet; alcohol use; smoking	2 arms; 6 mo; 29 SMS text messages and 4 videos	Telephone	Self-reporting	Participants in the intervention group were instructed to install the app, and 29 SMS text messages were sent for medication adherence, risk factor control, and lifestyle and behavior change; in addition, 4 videos on various aspects of stroke were sent to participants in the intervention group on scheduled days of the month.	Usual care	NR[b]
Bae et al [36], 2021; South Korea	Patients with coronary heart disease	n=879; age: mean 60.4 (SD 10.5) y; 83.3% women	PA; smoking	2 arms; 6 mo; 4 times/wk	SMS text messages	PA: IPAQ[c]	Messages were based on the TEXT ME[d] trial and the Australian Heart Foundation Healthy Living Guidelines.	Usual care	NR
Bantum et al [37], 2014; United States	Survivors of cancer	n=352; age: mean 50.9 (SD 11.0) y; 82.1% women	PA; diet; sleep	2 arms; 6 mo; 6 wk	Web-based	PA: Godin Leisure Time Exercise Questionnaire; diet: Block Food Frequency Questionnaire; sleep: 5-item validated Women's Health Initiative Insomnia Rating Scale	The intervention was based on online seminars, covering topics such as diet, exercise, and stress management; participants set weekly action plans and receive facilitator feedback on their progress through a *Discussion Center* update; the intervention website includes, apart from *Discussion Center*, sections labeled *My Tools*, *Post Office*, and *Help*, promoting social interaction and personal behavior shaping in addition to providing links to resources.	Waitlist	Behavior change theory
Bloom et al [38], 2020; United States	Patients with chronic obstructive pulmonary disease	n=22; age: mean 70 (SD 7.4) y; 40.9% women	PA; smoking; alcohol use; diet	2 arms; 3 mo; NR	Social networking tool	PA: IPAQ, short-form version; diet: FFQ[e]	The intervention was based on an online social tool that provided individuals with multiple levels of self-management support; the online tool assisted users in mapping personal support networks, understanding preferences, and guiding participation in valuable social activities; the intervention was delivered by trained researchers, and participants had the option of receiving the link via email.	Usual care	NR

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Campbell et al [39], 2009; United States	Patients with colon and rectal cancer	n=735; age: mean 66.5 (SD 10.0) y; 49.4% women	PA: NR; diet	4 arms; 12 mo; 4 times in total	Telephone	PA: 7-Day Physical Activity Recall questionnaire; diet: FFQ	The intervention components included using a client-centered collaborative decision-making approach, giving nonjudgmental feedback, allowing for resistance, and encouraging the participant to make the argument for change; interviewers relied on using open-ended questions and reflections to draw out participants' motives and desires regarding behavior change.	Tailored print communication intervention	Health behavior theories, transtheoretical model (stages of change), and social cognitive theory
Chow et al [40], 2015; Australia	Patients with coronary heart disease	n=710; age: mean 57.6 (SD 9.2) y; 18% women	PA; smoking	2 arms; 6 mo; 4 times/wk	SMS text messages	PA: GPAQ[f]	The SMS text messaging–based prevention program sent participants 4 semipersonalized messages per week for 24 weeks, including advice on diet, PA, and smoking cessation; messages were selected based on an algorithm and participants' baseline characteristics and sent randomly through a management system; participants received brief training to monitor message responses and did not respond interactively.	Usual care	NR
Cicolini et al [41], 2014; Italy	Patients with hypertension	n=198; age: mean 59.0 (SD 14.5) y; 49% women	PA; smoking; alcohol use; diet	2 arms; 6 mo; 1 time/wk	Email and telephone	Questionnaires	The intervention was based on email reminders for 6 months covering advice on diet, exercise, smoking cessation, and more; recommendations were based on current healthy lifestyle guidelines.	Usual care	NR
Djuric et al [42], 2011; United States	Women with breast cancer	n=30; age: mean 52.3 (SD 9.5) y	PA; diet	2 arms; 12 mo; NR	Telephone	PA: Women's Health Initiative questionnaire; diet: dietary recalls	Participants received written educational materials that included recommendations for daily exercise and the Dietary Guidelines for Americans 2005 as well as telephone counseling from a registered dietitian trained in motivational interviewing techniques.	Received written educational materials	Social cognitive theory
Gallagher et al [43], 2022; Australia	Patients with coronary heart disease	n=390; age: mean 61.2 (SD 11.5) y; 17.5% women	PA; smoking	2 arms; 6 mo; self-reported	Mobile app	PA: GPAQ	The intervention was a gamification-based mobile app that encouraged users to engage in a healthy lifestyle and risk factor change through elements such as tracking behaviors, setting short- and long-term challenges, games, and quizzes; the game elements incorporated social cognitive theory strategies to promote sustained behavior change through message prompts, game challenges, and rewards.	Usual care	Social cognitive theory

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Glasgow et al [44], 2012 United States	Patients with type 2 diabetes mellitus with obesity	n=204; age: mean 58.4 (SD 9.2) y; 49.8% women	PA; SB[g]	3 arms; 12 mo; NR	Website-based	PA: CHAMPS[h] questionnaire; SB: NR	Participants were given access to the health website and asked to select goals in 3 areas (medication adherence, PA, and food choices), and they received periodic motivational calls; they recorded their progress and received immediate feedback on their success in meeting their goals over the past 7 days.	Usual care	Behavior change theory; social cognitive theory; social ecological theory
Golshahi et al [45], 2015; Iran	Patients with hypertension	n=822; age: mean 57.35 (SD 8.32) y; 78% women	PA; smoking; diet	4 arms; 5 mo; 8 times in total	SMS text messages	Self-report	Patients were advised to adhere to taking medication daily; increase PA (most days of the week); follow the dietary approach to prevent and manage hypertension (DASH[i] diet), including eating a diet rich in vegetables and reducing dietary sodium to <1500 mg/d; and stop smoking.	Usual care	NR
Grey et al [46], 2019; United Kingdom	Adults with overweight or obesity	n=60; age: mean 50 (SD 8.9) y; 44.1% women	PA; SB; diet	2 arms; 3 mo; NR	Website-based	PA and SB: BodyMedia SenseWear core monitors; diet: 3-day weighed food and fluid records	The intervention was web-based and provided participants with information about PA and healthy eating, as well as a personalized interactive area for participants to set goals and plans related to PA and eating and to monitor their progress toward achieving their goals.	Face-to-face introductory session and NHS[j] online health resources	Behavior change techniques
Harrigan et al [47], 2016; United States	Survivors of breast cancer with a BMI of 25.0 kg/m^2	n=100; age: mean 59.0 (SD 7.5) y; 100% women	PA; diet	3 arms; 6 mo; NR	Telephone	PA: interview-administered PA questionnaire and pedometers; diet: FFQ	This intervention was composed of 11 coaching sessions (each lasting 30 min), represented by a social core curriculum with specific information about nutrition, exercise, and behavior strategies.	Usual care	Social cognitive theory
Hawkes et al [48], 2013; Canada	Patients with colorectal cancer	n=410; age: mean 66.4 (SD 10.6) y; 53.9% women	PA; alcohol use; diet	2 arms, 6 mo; 11 times in total	Telephone	PA: Godin Leisure Time Exercise Questionnaire; diet: FFQ	The intervention was conducted with 11 telephone health coaching sessions over 6 months, and participants were given manuals, motivational postcard prompts, and pedometers; the telephone sessions addressed lifestyle health behaviors and strategies, while the participant handbook included educational information on healthy behaviors.	Usual care	Behavior change techniques
Jane et al [49], 2017; Australia	Adults with overweight and obesity	n=67; age: mean 50.0 (SD 11.2) y; 85% women	PA; alcohol use; diet	3 arms; 3 mo; NR	Social networking tool	Diet: 3-day food records	Participants received a condensed version of the diet, which included detailed information and instructions.	Usual care	NR

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Jennings et al [50], 2014; Australia	Patients with type 2 diabetes mellitus	n=397; age: mean 58.2 (SD 10.3) y; 47.6% women	PA; SB	2 arms; 3 mo; NR	Web - based	PA and SB: IPAQ, long-form version; pedometers	The website program, which used a self-management approach and was developed based on the theory of planned behavior, started a new health module topic each week, and weekly email reminders were distributed to participants; the content changed every week but always included a link to the intervention website.	Waitlist	Theory of planned behavior
Johnston et al [51], 2016; Sweden	Patients with myocardial infarction	n=174; age: mean 57.6 (SD 8.3) y; 18.4% women	PA; smoking	2 arms; 6 mo; every day	Mobile app	PA: PA questionnaires	Participants received a complete interactive patient support tool (web-based application) installed on their smartphones; educational information was sent to participants, including an extended drug adherence e-diary as well as exercise, weight, and smoking modules.	Usual care plus a simplified tool	NR
Kanera et al [52], 2017; Netherlands	Survivors of early cancer diagnoses	n=462; age: mean 55.9 (SD 11.4) y; 79.9% women	PA; diet	2 arms; 6 mo; NR	Web-based	PA: SQUASH[k]; diet: Dutch Standard Questionnaire on Food Consumption	The intervention was a web-based self-management program that collected self-reported lifestyle scores and compared them to guidelines, including advice on what modules (eg, PA and diet modules) were most relevant for them to use; 4 weeks after completing a module, participants were invited to participate in a brief online personalized assessment session; they were given personalized feedback to increase levels of coping self-efficacy to improve behavior maintenance.	Usual care	Theory of planned behavior; self-regulation theory
Krebs et al [53], 2017; United States	Survivors of cancer	n=86; age: mean 59.8 (SD 11.4) y; 82.96% women	PA; diet	2 arms; 3 mo; NR	DVD	PA: Godin Leisure Time Exercise Questionnaire; diet: FFQ	The intervention, which was based on formative evaluation data from a patient telephone survey, was delivered via a DVD; the DVD program included healthy eating and PA recommendations and took approximately 60 minutes to complete.	Usual care plus brief counseling	Social cognitive theory
Lara et al [54], 2016; United States	Retired adults	n=75; age: mean 61 (SD 4) y; 75% women	PA; diet	2 arms; 8 wk; mean 11.4 times in total	Web-based	PA: pedometer; accelerometer; diet: 24-hour recall method	The intervention comprised 5 modules (time, changing work, moving more, being social, and eating well) as well as a diary and a dashboard section to assist with site navigation; the intervention content was personalized to the participant based on information provided during use.	Usual care	Behavior change techniques

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Li et al [55], 2022; United States	Patients with insomnia symptoms	n=21; age: mean 73.3 (SD 6.6) y; 71.4% women	PA; sleep	2 arms; 6 mo; 2 times/mo	Smartphone and smartwatch	PA: Actiwatch 2; PASE[l]; sleep: PSQI[m]	The intervention included (1) mHealth[n] technology learning sessions, (2) personalized PA training, (3) mHealth strategies for PA, (4) financial incentives for completing the prescribed PA, and (5) additional support for mHealth technology.	Program book and a brief 1-time educational session	NR
Liu et al [56], 2018; Canada	Patients with hypertension	n= 129; age: mean 56.9 (SD 9.05) y; 47.7% women	PA; diet	3 arms; 4 mo; voluntary sign-in	Web-based	PA: pedometer; diet: Block Food Frequency Questionnaire	Participants set their own goals or chose interventions to achieve behavioral goals through e-counseling; the user-driven e-consultation group provided weekly information via email and options for areas of lifestyle change through text and video web links that contained information on developing exercise and diet plans, setting behavioral goals, self-monitoring lifestyle behaviors and blood pressure, resolving psychological conflicts, increasing change efficacy, and reviewing social and cognitive behavioral skills for relapse prevention to help participants maintain adherence.	A weekly email newsletter	Transtheoretical model
Morey et al [57], 2009; Canada, United Kingdom, and United States	Survivors of cancer	n=641; age: mean 73.1 (SD 5.1) y; 54.6% women	PA: NR; diet	2 arms; 12 mo; 15 sessions and 8 prompts	Telephone	PA: CHAMPS questionnaire; diet: 24-hour recalls	Participants received a personalized workbook containing advice on lifestyle and weight status; they also received aids such as pedometers, exercise bands, exercise posters, and a food portion schedule guide; each participant was assigned a health counselor to provide support and advice during counseling; in addition, the principal investigator provided support services over the telephone and sent regular progress reports to participants to motivate them to continue to change their behavior.	None	Social cognitive theory; transtheoretical model
Müller et al [58], 2016; Australia	Older Malaysians	n=43; age: mean 63.28 (SD 4.50) y; 74% women	PA; SB	2 arms; 3 mo; 5 times/wk	Text messages	PA and SB: IPAQ, short-form version	Pamphlets were distributed to participants. Participants were introduced specific exercises and provided a printed home-based exercise booklet additional text messages were sent to encourage and remind participants to follow the exercise program.	Exercise booklet	Behavior change techniques

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Pischke et al [59], 2022; Germany	Adults aged ≥60 y	n=204; age: mean 68.7 (SD 5.4) y; 66.2% women	PA; SB	3 arms; 6 mo; NR	Web-based	PA and SB: accelerometer	Intervention materials were provided via a website. participants were asked to provide weekly feedback on PA goal completion; weekly group sessions provided to participants included performing the exercises in groups, going for joint walks, and discussing weekly health education topics.	A print intervention with subjective PA self-monitoring via printed PA pyramid	Self-regulation theory and various behavior change techniques
Poppe et al [60], 2019a; Belgium	Patients with type 2 diabetes mellitus	n=54; age: mean 62.67 (SD 8.4) y; 37% women	PA; SB	3 arms; 5 wk; 1 time/wk	Mobile app	PA and SB: IPAQ; accelerometer	The intervention developed a personal action plan to change selected health behaviors, anticipated potential barriers and found solutions, and enabled users to choose how to monitor their behavior; at the end of the first session, the user's responses were summarized in a printable action plan, and they were provided with optional information on how they could obtain support from a partner, friend, family member, or colleague.	Waitlist	Behavior change theory
Poppe et al [60], 2019b; Belgium	Adults aged ≥50 y	n=63; age: mean 58.68 (SD 7.76) y; 75% women	PA; SB	3 arms; 5 wk; 1 time/wk	Mobile app	PA and SB: IPAQ; accelerometer	The intervention developed a personal action plan to change selected health behaviors, anticipated potential barriers and found solutions, and enabled users to choose how to monitor their behavior; at the end of the first session, the user's responses were summarized in a printable action plan, and they were provided with optional information on how they could obtain support from a partner, friend, family member, or colleague.	Waitlist	Behavior change theory
Prabhakaran et al [61], 2019; India	Patients with hypertension or diabetes mellitus	n=3698; age: mean 55.1 (SD 11.0) y; 44.8% women	smoking; alcohol use	2 arms; 12 mo; NR	Mobile app	Self-reporting	The intervention stored the health records electronically, enabling long-term monitoring and follow-up; it was also equipped to send SMS text message reminders (to take medication and attend follow-up visits) to patients.	Usual care	NR
Sakane et al [62], 2023; Japan	Patients with obesity and hypertension	n=78; age: mean 51.95 (SD 6.5) y; 0% women	PA; diet; smoking	2 arms; 3 mo	Mobile app	Self-reporting	The app provided an initial assessment of the participant, identified a behavioral agenda (exercise, diet, lifestyle habits, and number of steps per day), and enabled self-monitoring.	Usual care	Behavior change techniques

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Swoboda et al [63], 2016; United States	Patients with type 2 diabetes mellitus	n=60; age: mean 56.1 (SD 6.7) y; 68.5% women	PA; SB; diet	3 arms; 4 mo; biweekly telephone calls (7 calls in total)	Telephone	PA and SB: IPAQ, long-form version; diet: FFQ	Self-set goals related to diet and PA were developed based on teleconferences, and action plans were developed to help participants achieve their goals; motivational interviewing and decision-making guidance were used to help participants develop individualized diet and activity goals.	Attention control	NR
Taylor et al [64], 2021; United Kingdom	Patients with inactive exercisers and chronic health conditions	n=450; age: mean 50 (SD, 12) y; 64% women	PA; SB; sleep	2 arms; 4 mo; NR	Web-based	PA and SB: accelerometer and pedometer; sleep: accelerometer	The intervention (e-coachER) offered a range of interactive opportunities to enhance participants' motivation to take up the exercise referral scheme and to maintain a more physically active lifestyle, whether or not they engaged with their local exercise referral scheme; participants were encouraged to make use of the pedometer and the activity record sheets for self-monitoring and goal setting in conjunction with the e-coachER website.	Usual care plus an exercise referral scheme	Self-determination theory
Wang et al [65], 2020; China	Patients with type 2 diabetes mellitus	n=171; age: mean 55.1 (SD 10.8) y; 43% women	PA; smoking; alcohol use	2 arms; 12 mo; 2 times/wk	SMS text messages	Self-reporting	The information design included 5 main areas: health awareness, dietary control, PA, lifestyle habits, and weight control; telephone follow-up was conducted after each stage of the intervention.	Regular education plus general theoretical knowledge messages	Transtheoretical model
Watson et al [66], 2015; United Kingdom	Adults who were obese and inactive with ≥1 CVD[o] risk factors	n=65; age: mean 52.1 (SD 7.4) y; 55% women	PA; diet	2 arms; 12 mo; personalized information service	Web-based	PA: validated RPAQ[p]; diet: 3-month recall	The Imperative Health system generated personalized daily targets (weight loss, PA, and dietary targets) for each participant; automated weekly feedback on the participants' performance, assessed by the self-monitoring devices (weighing scales and accelerometer) and the food diary, was provided.	Usual care	Social support and decisional balance theory

XSL•FO
RenderX

Study, year; country	Population	Sample population; mean age (SD); female population (%)	Targeted risk behaviors	Study design; intervention duration; frequency	Intervention	Measurement tools	Intervention components	Comparison group	Theoretical basis or behavior change techniques
Zheng et al [67], 2019; China	Patients with coronary heart disease and without diabetes mellitus	n=822; age: mean 56.4 (SD 9.5) y; 14.1% women	PA; smoking	2 arms; 6 mo; 6 times/wk	Text messages	PA: IPAQ; short-form version	The intervention was designed as a secondary prevention program that sent participants 1 of each of the following text message types according to a prespecified algorithm: general knowledge of coronary heart disease, blood pressure control, medication adherence, PA, healthy eating, and smoking cessation.	Usual care plus 2 thank-you messages per month	Behavior change theory

[a]PA: physical activity.
[b]NR: not reported.
[c]IPAQ: International Physical Activity Questionnaire.
[d]TEXT ME: Tobacco, Exercise, and Diet Messages.
[e]FFQ: Food Frequency Questionnaire.
[f]GPAQ: Global Physical Activity Questionnaire.
[g]SB: sedentary behavior.
[h]CHAMPS: Community Healthy Activities Model Program for Seniors.
[i]DASH: Dietary Approaches to Stop Hypertension.
[j]NHS: National Health Service.
[k]SQUASH: Short Questionnaire to Assess Health-Enhancing Physical Activity.
[l]PASE: Physical Activity Scale for the Elderly.
[m]PSQI: Pittsburgh Sleep Quality Index.
[n]mHealth: mobile health.
[o]CVD: cardiovascular disease.
[p]RPAQ: Recent Physical Activity Questionnaire.

Characteristics of Participants

A total of 12,931 participants were enrolled in the included studies, with the number of participants ranging from 21 to 879 in individual studies. The mean age of the older adults included ranged between 50.0 (SD 11.2) years [49] and 73.3 (SD 6.6) years [55]. Nearly one-half of the participants (6009/12,931, 46.47%) were female, and 2 (6%) of the 35 studies targeted only women [42,47]. The studies included not only healthy older adults (4/35, 11%) but also people with various chronic conditions, such as cancer (8/35, 23%), type 2 diabetes mellitus (6/35, 17%), hypertension (4/35, 11%), coronary disease (4/35, 11%), and overweight or obesity (3/35, 9%).

Characteristics of eHealth Interventions

The eHealth interventions included in these studies were delivered mainly through 8 intervention methods: SMS text message [36,40,45,58,65,67], telephone [35,39,42,47,48,55,57,63], email [41], website [37,50,52,54,56,59,64,66], mobile app [44,51,60-62], social tool [38,49], DVD [53], and smartwatch [55]; and the most common methods were based on a website (8/35, 23%), SMS text message (6/35, 17%), and mobile app (6/35, 17%). Intervention durations in the included studies ranged from 5 weeks to 21 months, the most common being 6 months (12/35, 34%) and 12 months (8/35, 23%).

Most of the studies (30/35, 86%) targeted 2 to 3 risk factors. Of the 35 studies, 5 (14%) investigated 4 (67%) of the 6 behaviors (PA, smoking cessation, alcohol cessation, and diet) [35,38,41], and only 9% (3/35) of the studies addressed sleep [37,55,64]. The most common combination was PA and diet (8/35, 23%), followed by PA and SB (7/35, 20%) and PA and smoking cessation (5/35, 14%). Of the 35 studies, 25 (71%) used a theoretical basis (n=18, 72%) or behavior change techniques (n=7, 28%). In the studies using theoretical underpinnings, the most common were the social cognitive model (7/25, 28%) and behavior change theory or techniques (12/25, 48%); the studies also used the transtheoretical model, the theory of planned behavior, the self-regulation model, and the self-determination model.

Characteristics of Controls

All participants in the control groups received usual care during the study period. In addition, 6 (17%) of the 35 studies supplied a paper version of the guidance booklet [39,42,55,58,59,64]. Zheng et al [67] and Wang et al [65] sent thank-you text messages or general theoretical knowledge information, Grey et al [46] provided face-to-face introductory sessions and

XSL•FO
RenderX

National Health Service online health resources, and Swoboda et al [63] conducted an attention control intervention.

Outcome Measures

A variety of outcomes were subjectively or objectively assessed before and after the intervention. The measurement tools used for the different outcomes are shown in Multimedia Appendix 3.

The end points of self-reported total PA (TPA), objectively measured TPA, self-reported moderate to vigorous PA (MVPA), objectively measured MVPA, daily steps, regular exercise, and inactive PA were used to describe PA behavior. More than 10 different questionnaires were used to assess participants' self-reported PA levels and SB levels, the most common of which was the International Physical Activity Questionnaire. Objective measurements of TPA and MVPA were obtained from accelerometer measurements, and a pedometer was used to measure daily steps. Regular exercise was defined as accumulating >30 minutes of moderate exercise performed ≥5 days per week. PA was categorized into active and inactive on the basis of established recommendations regarding the total level: ≥600 and <600 metabolic equivalent of task minutes per week, respectively. Eight indicators were selected to describe diet behavior: energy intake as well as intake of fruits, vegetables, fruits and vegetables (together), fats, proteins, sugars, and fiber. Of the 35 studies, 17 (49%) used various versions of the Food Frequency Questionnaire and dietary records or recall from different periods to evaluate the dietary indicators. In addition, the Dutch Standard Questionnaire on Food Consumption was used in the study by Kanera et al [52]. The 3 studies involving sleep used 3 measuring tools, namely the 5-item validated Women's Health Initiative Insomnia Rating Scale, the Pittsburgh Sleep Quality Index, and an accelerometer. The prevalence of alcohol and tobacco use was collected through interviews or questionnaires.

Risk of Bias

The included studies were assessed for risk of bias using the risk-of-bias tool for randomized trials in the Cochrane Handbook for Systematic Review of Interventions (version 5.1.0). A summary of the overall risk-of-bias assessment for all included studies as well as the judgments regarding specific studies can be found in Multimedia Appendix 4 [34-67].

According to our assessment, which was conducted by 2 independent reviewers (BS and GL), 17 (49%) of the 35 publications had a high risk of bias for not blinding participants and study personnel to the intervention assignment; however, blinding in intervention trials for behavior change is challenging. For most of the publications (20/35, 57%), the risk associated with allocation concealment was high (6/35, 17%) or unclear (14/35, 40%). Of the 35 studies, 9 (26%) did not provide sufficient detail to determine whether researchers were blinded to the assessment of outcomes; therefore, the risk was rated as unclear.

Certainty of Evidence

Of all outcomes listed in the summary of findings presented in Multimedia Appendix 5, we deemed the quality of the evidence to be very low to high, as determined by the GRADE framework. We judged the quality of evidence for most of the diet-related outcome indicators to be high. The evidence for outcome indicators related to PA and SB was low, mainly because of the high risk of bias, inconsistent study results, and imprecise measurements. In addition, evidence relating to smoking, alcohol use, and sleep was deemed to be very low or low because of the high risk of bias, small sample sizes, and imprecise measurements.

Meta-Analysis Results

Overview

We have summarized the forest plot results for continuous (Figure 2) and dichotomous (Figure 3) variable outcomes. Overall, eHealth-based MHBC interventions may be able to improve PA, diet, and smoking cessation in older adults, with no effect on alcohol use, SB, or sleep. Forest plots for each of the outcomes are presented in Multimedia Appendix 6 [34-67].

XSL•FO
RenderX

Figure 2. Forest plot integration for continuous outcomes. MVPA: moderate to vigorous physical activity; PA: physical activity; SB: sedentary behavior; SMD: standard mean difference. *P<.05, **P<.01, ***P<.001.

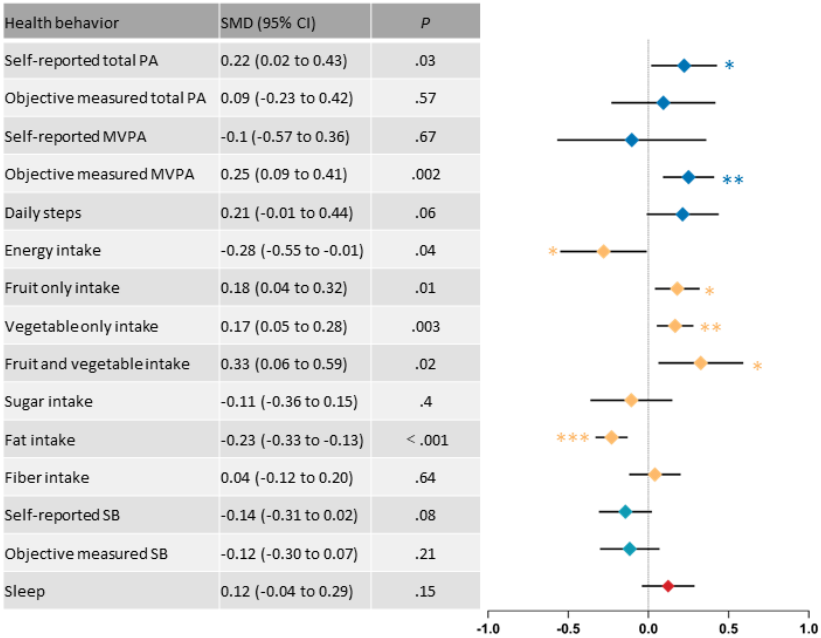

Figure 3. Forest plot integration for dichotomous outcomes. PA: physical activity. ***P<.001.

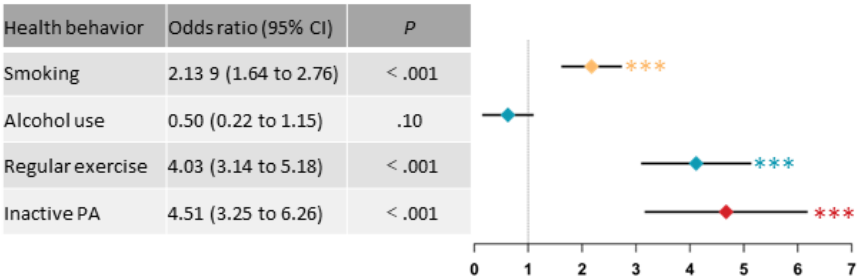

PA Assessment

The effectiveness of eHealth interventions on self-reported TPA levels in older adults was assessed in 17 (49%) of the 35 studies. Compared with the control group, the eHealth interventions significantly increased self-reported TPA (SMD 0.22, 95% CI 0.02-0.43; P=.03), with substantial heterogeneity between the studies (I^2=79%; P<.001; Figure S1A in Multimedia Appendix 6). The sensitivity analysis suggested a similar result (SMD 0.12, 95% CI 0.04-0.20; P=.006), and the heterogeneity dropped significantly (I^2=0%; P=.80) after excluding the study by Cicolini et al [41] (Multimedia Appendix 7 [34-67]). A meta-analysis of the effectiveness of eHealth interventions on objectively measured TPA by a fixed effects model (I^2=19%; P=.79) showed the opposite result (SMD 0.09, 95% CI –0.23 to 0.42; P=.57; Figure S1B in Multimedia Appendix 6). No significant increase was observed in self-reported MVPA after the eHealth intervention (SMD 0.10, 95% CI –0.57 to 0.36; P=.67), with greater heterogeneity (I^2=94%; P<.001; Figure S1C in Multimedia Appendix 6). After excluding the studies by Taylor et al [64] and Pischke et al [59], synthesis using fixed effects models showed a significant increase in self-reported MVPA after the intervention (SMD 0.15, 95% CI 0.03-0.27; P=.02), with moderate heterogeneity (I^2=54%; P=.03). However, the eHealth-based MHBC interventions led to a small but significant increase in objectively measured MVPA (SMD 0.25, 95% CI 0.09-0.41; P=.002), without any heterogeneity (I^2=0%; P=.76; Figure S1D in Multimedia Appendix 6). In addition, a meta-analysis of the 7 studies reporting daily steps showed that the eHealth interventions increased the number of steps taken by older adults compared to the control group (SMD 0.21, 95% CI –0.01 to 0.44; P=.06), corresponding to an average estimate of 597 steps, with borderline significant results and low heterogeneity (I^2=0%; P=.64; Figure S1E in Multimedia Appendix 6). Overall, compared to the controls, the eHealth interventions significantly increased the prevalence of regular exercise (OR 4.17, 95% CI 3.42-5.36; P<.001), with low heterogeneity (I^2=0%; P<.001; Figure S1F in Multimedia Appendix 6); and significantly reduced the prevalence of inactive PA (OR 4.53, 95% CI 3.27-6.28; P<.001), with low heterogeneity (I^2=16%; P=.31; Figure S1G in Multimedia Appendix 6). Sensitivity analysis showed that these results were relatively stable.

XSL•FO
RenderX

SB Assessment

Of the 35 studies, 9 (26%) assessed self-reported SB and 7 (20%) assessed objectively measured SB. A meta-analysis showed no effect of the eHealth intervention on self-reported sedentary time compared to the controls (SMD –0.14, 95% CI –0.31 to 0.02; P=.08; I^2=0%; P=.95; Figure S2A in Multimedia Appendix 6). Similarly, there was no effect of the eHealth intervention on objectively measured sedentary time (SMD –0.12, 95% CI –0.30 to 0.07; P=.21; I^2=0%; P=.91; Figure S2B in Multimedia Appendix 6). Sensitivity analysis showed that the result was stable.

Diet

Mean energy intake (kcal) data from 9 (26%) of the 35 trials were suitable for meta-analysis. Energy intake was significantly lower in the eHealth intervention group (SMD –0.28, 95% CI –0.55 to –0.01; P=.04) compared to the control group, with low heterogeneity (I^2=0%; P=.54; Figure S3A in Multimedia Appendix 6). Of the 35 trials, 8 (23%), 10 (29%), and 7 (20%) assessed intake of fruits (servings), vegetables (servings, portion, and times), and fruits and vegetables (servings), respectively, and the eHealth interventions significantly increased intake of fruits (SMD 0.18, 95% CI 0.04-0.32; P=.01; I^2=19%; P=.28; Figure S3B in Multimedia Appendix 6), intake of vegetables (SMD 0.17, 95% CI 0.05-0.28; P=.003; I^2=12%; P=.34; Figure S3C in Multimedia Appendix 6), and intake of fruits and vegetables (SMD 0.18, 95% CI 0.04-0.32; P=.01; I^2=19%; P=.28; Figure S3D in Multimedia Appendix 6) compared to the controls. Of the 35 trials, 5 (14%) assessed sugar intake, which did not change significantly after the intervention compared to the control group (SMD –0.11, 95% CI –0.36 to 0.15; P=.40; I^2=0%; P=.45; Figure S3E in Multimedia Appendix 6). Of the 35 trials, 9 (26%) assessed fat intake changes, and there was a significant reduction in fat intake after the eHealth intervention (SMD –0.23, 95% CI –0.33 to –0.13; P<.001; I^2=0%; P=.93; Figure S3F in Multimedia Appendix 6). Similarly, a meta-analysis of the 7 trials that assessed fiber intake showed no significant changes (SMD 0.04, 95% CI –0.12 to 0.20; P=.64; I^2=14%; P=.32; Figure S3G in Multimedia Appendix 6). Sensitivity analysis showed that the result was relatively stable.

Smoking and Alcohol Use

A meta-analysis of the 8 studies reporting outcomes on smoking showed that the eHealth interventions significantly reduced the prevalence of smoking (OR 2.09, 95% CI 1.62-2.70; P<.001) but with greater heterogeneity (I^2=85%; P<.001; Figure S4A in Multimedia Appendix 6). After excluding the studies by Gallagher et al [43] and Prabhakaran et al [61], sensitivity analysis showed that the result was stable (OR 3.64, 95% CI 2.68-4.93; P<.001), and the results were homogeneous across the trials (I^2=0%; P=.76). Of the 35 studies, 4 (11%) reported outcomes for alcohol use, and meta-analysis showed no effect of the eHealth intervention on the prevalence of alcohol use (OR 0.75, 95% CI 0.26-2.18; P=.60), with less heterogeneity (I^2=65%; P=.03; Figure S4B in Multimedia Appendix 6). After excluding the study by Babu et al [35], sensitivity analysis showed that the result was stable (OR 0.50, 95% CI 0.22-1.15; P=.10), and the results were homogeneous across the trials (I^2=40%; P=.19).

Sleep

Data from 3 (9%) of the 35 studies were used to assess the effectiveness of eHealth interventions on sleep in older adults. Due to the low level of article heterogeneity (I^2=44%; P=.17) and the small number of included studies, a fixed effects model was used. Meta-analysis showed that the eHealth interventions did not affect sleeping time in older adults (SMD 0.12, 95% CI –0.04 to 0.29; P=.15; Figure S5 in Multimedia Appendix 6). Sensitivity analysis showed that the result was stable.

Discussion

Overview

This systematic review is the first study to systematically examine the effectiveness of eHealth interventions targeting ≥2 of the following behaviors in a population of older adults (aged ≥50 y): PA, diet, smoking, alcohol consumption, sleep, and SB. A total of 35 studies with 12,931 older adults aged ≥50 years met the eligibility criteria for this review. The majority (6009/12,931, 46.47%) of the participants were female, with a mean age ranging from 50.0 (SD 11.2) years to 73.3 (SD 6.6) years. The overall methodological quality was moderate, according to the GRADE framework. Compared to usual care or waitlist, eHealth-based MHBC interventions significantly changed diet, SB, and smoking but had little effect on PA, alcohol consumption, and sleep.

Principal Findings

PA Assessment

Our data suggest that eHealth-based MHBC interventions may improve PA levels and daily steps in older adults. Previous meta-analyses also revealed that eHealth-based MHBC interventions significantly promoted PA among people with noncommunicable diseases and adults compared to the control conditions [22,68]. It is worth noting that the PA referred to in both studies was a synthetic size of multiple effect sizes within a particular scope of outcomes (eg, energy expenditure, steps, and time spent in PA or MVPA). We conducted a separate meta-analysis of PA-related metrics. First, the eHealth-based MHBC intervention significantly improved older adults' self-reported TPA compared with the control group. This is consistent with a previous meta-analysis [28], which found a mean increase in TPA of 90.7 minutes per week according to questionnaires in the eHealth-based MHBC intervention group compared to the control group. Another review involving people aged ≥50 years found that interventions containing smart technology significantly increased self-reported TPA compared to face-to-face interventions (SMD 0.17, 95% CI 0.02-0.32) [69], which is broadly comparable to our effect size. Second, although we also found that the eHealth-based MHBC intervention significantly increased the time spent in MVPA in older adults, this result needs to be interpreted with caution because 4 (80%) of the 5 data points are from the same article. Interestingly, this review found very different results for self-reported and objectively measured TPA or MVPA. Another

XSL•FO
RenderX

review reported similar results, with eHealth interventions significantly increasing objectively measured MVPA but not self-reported MVPA [69]. This may be because the MHBC intervention intervened on a larger number of target behaviors than the SHBC intervention, and the intervention effect was unstable [70,71]. Another reason may be that the effect of eHealth interventions on the PA time measured by ActiGraph devices and that measured by questionnaires was quite different [72]. Third, this review also showed that participants in the eHealth intervention group walked more than those in the control group (mean difference 597 steps/d). This is a conclusion we believe to be quite credible [28,69]. Earlier studies have shown that older people commonly cite inconvenience and a lack of access to a PA program as reasons for avoiding performing PA [73]. Walking is the most commonly targeted PA for older people because it does not require access to specific programs and can be practiced anywhere [74]. Finally, we also meta-analyzed changes in the prevalence of regular exercise as well as inactive PA and found a positive effect of an eHealth-based MHBC intervention. With up to 40% of people in industrialized countries not engaging in any regular PA [75], increasing PA rates by the rate produced by these interventions would have an important impact on health outcomes.

Diet

Our review found that all MHBC interventions involving dietary interventions in older adults were performed together with PA, with a combination of both of them being the most common (8/35, 23%). This combination was also found by previous authors to be the most common combination in adolescent and adult MHBC interventions [26,68]. The results of our study suggest that eHealth-based MHBC interventions can significantly improve the diet of older adults. Specifically, there was an increase in fruit or vegetable intake. This finding aligns with previous research demonstrating that both the eHealth-based SHBC intervention and the eHealth-based MHBC intervention significantly boosted fruit or vegetable intake in the intervention group [26,53,76]. Furthermore, we also found significant reductions in fat intake and energy intake in this population. eHealth interventions similarly significantly reduced dietary fat intake in adults but not in adolescents [26,76]. This may be because the health benefits of reducing fat intake are greater for middle-aged and older adults; therefore, they have greater motivational self-efficacy [77]. eHealth interventions have an impact not only on dietary intake (33%-35%) but also on diet-related cognitive variables (69%-79%) [78]. In addition, we found no evidence that eHealth interventions can change the dietary intake of sugar and fiber in older people. Although other studies have shown that it is possible to reduce children's consumption of sugary beverages, reducing unhealthy eating behaviors is often considered more difficult than starting new healthy behaviors [79,80]. The contents of eHealth interventions that are communicated to participants about behavior change and the feedback they receive can also influence changes in participants' behavior. Further research is needed to examine intervention contents regarding these dietary outcomes to increase the size and sustainability of the effects of MHBC interventions, particularly among older adults.

Smoking and Alcohol Use

The effectiveness of eHealth-based MHBC interventions for stopping drinking is generally weak; however, promising results were seen for smoking cessation. The quality of evidence for alcohol use was deemed to be very low because the review only included 4 RCTs with relatively small sample sizes, which might have affected the pooled effect. In general, the longer the duration of the drinking and smoking interventions, the more effective they are [81]. In this review, we included studies with intervention durations of 3 months as well as 6 months, in addition to the 12-month intervention duration. Due to the limited number of studies, we did not conduct further analyses. Indeed, the evidence for the effectiveness of eHealth-based smoking and alcohol cessation interventions is currently controversial [82,83]. There is also evidence that eHealth interventions may have similar effects in reducing drinking and smoking when compared to face - to - face treatment or no support [84,85]. Moreover, SMS text messaging may be particularly powerful for smoking behaviors such as abstinence and reduction of use, particularly automated reminders and motivating messages that can be sent during times of cravings [86]. Thus, further work is needed to detect specific moderators of interventions such as follow-up length and message frequency.

SB Assessment

The lack of a significant reduction in either self-reported sedentary time (P=.08) or objectively measured sedentary time (P=.06) was an unexpected finding, leaving ample room for improvement. Champion et al [26] found that eHealth-based MHBC interventions led to decreased sedentary time compared to controls (SMD –0.09, 95% CI –0.17 to –0.01). Although in the same direction and a similar effect size to our result, their result was statistically significant. Previous literature suggests that eHealth interventions can improve sedentary time within, but not between, groups of older adults. The review by Yerrakalva et al [87] showed that mobile health app interventions had the potential to reduce SB and increase PA in older adults. The MHBC studies for SB included in this review are all combined with PA, and most of the comments received by participants in the eHealth interventions (via SMS text messaging and telephone) supported their focus on PA and the step goals rather than on SB. In addition, research suggests that given the automaticity of SB, different and more effortful strategies are required to break existing habits compared to forming new habits [88,89]. This hints that future interventions regarding MHBC interventions could use multiple intervention strategies to reduce sedentary time [90].

Sleep

Sleep duration is a commonly used indicator for assessing sleep quality, and suboptimal sleep duration has been associated with increased morbidity and mortality [91], with short sleep durations in particular being detrimental to health [92]. We found that eHealth-based MHBC interventions may not significantly improve sleep duration in older adults. Of the 3 studies on sleep that we included, 2 (67%) found significant improvements in sleep quality. However, the third study did not find that the eHealth intervention did not improve sleep

XSL•FO
RenderX

duration. This may be because, in this study, sleep was used as a secondary outcome, and the intervention only set specific PA goals for the participants without any recommendations for sleep. In contrast to our findings, Howarth et al [93] evaluated the impact of a workplace digital health intervention on health-related outcomes and found that it could improve participants' sleep quality. Inconsistencies in conclusions may be due to possible differences in the effects of eHealth interventions on various indicators of sleep. In addition to objectively measured sleep duration, a range of subjectively measured sleep outcomes (nocturnal sleep quality, sleep duration, sleep efficiency, sleep latency, and sleep medication use) are equally important in older adults with insomnia or insomnia symptoms [94]. PA resulted in significant improvements in overall sleep quality, sleep quality, and sleep latency but not in sleep duration, efficiency, or disturbance [95]. According to the American National Sleep Foundation, nonpharmacological treatment options are the preferred first choice of treatment for sleep problems [96]. Therefore, the doubt whether MHBC interventions can improve sleep problems in older adults is important and needs to be explored in more RCTs.

Strengths and Limitations

The strengths of this study include the focus on a variety of lifestyle behaviors associated with adverse health outcomes and our rigorous methodology.

However, our study has several limitations. First, despite our best efforts to conduct a thorough literature search in a limited number of databases, the diversity of outcome indicator descriptions may still have resulted in the omission of appropriate topics or relevant studies. Second, this review included some participants aged 50 to 60 years because some of the trials were designed to recruit older adults, but they did not specifically exclude people aged <60 years. Future research should explicitly target older adults (aged ≥60 y) for RCTs of eHealth interventions. Third, although all included studies used eHealth interventions, there was still considerable variation in terms of participants (eg, cultural background, age, and health status), intervention characteristics (eg, intervention channel, content, and duration), and outcome measures, and the small number of studies may lead to cautious interpretations of the pooled results. In particular, SHBC interventions differ from MHBC interventions in terms of effectiveness. Finally, because we only searched for articles published in English, the results of the meta-analysis may be affected by language bias.

Implications

This paper highlights important directions for future research. First, there is a lack of qualified interventions to address sleep problems, although many older adults report sleep deprivation or sleep problems [97], and there is a growing recognition that exercise behaviors (such as PA), SB, and sleep are interdependent. Second, there is a general lack of follow-up analyses in intervention designs. Most of the included studies measured outcomes twice (before the intervention and after the intervention). Thus, the long-term and maintenance effects of eHealth interventions on multiple risk behavior changes in older adults have not been validated. As advocated [6,13], the next phase of this study is to explore how and under what conditions these initial changes can be maintained by adding a longer follow-up design. Finally, the intervention components included in the study focused primarily on health behavior education and counseling, with a lack of substantive behavioral coaching; this may be due to the limitations of current eHealth intervention channels. A commonly used intervention paradigm is to select a specific health behavior change theory (eg, behavior change theory, social cognitive theory, or self-regulation theory) as a framework and further promote a theory of choice for effective elements (eg, motivation, planning, and self-regulation). As the results of previous reviews have confirmed the effectiveness of this intervention paradigm, to further improve intervention effectiveness, it may be prudent to use dual-process approaches (ie, focusing on conscious and unconscious processes of behavior change) [54,56] alongside socioecological approaches (ie, involving policy dimensions, environment, and individual factors) [59,64].

In addition, we need to pay extra attention to the synergistic or transfer effects involved in MHBC interventions. Accordingly, lessons, skills, and knowledge learned concerning 1 behavior apply to another, thereby improving multiple behaviors. In addition, success in changing ≥1 lifestyle behaviors may also increase confidence or self-efficacy to improve risky behaviors in individuals who are less motivated to change [22]. In addition to potentially greater efficacy and impact on health, MHBC interventions have greater real-world applicability and provide information about the effective treatment of co-occurring behaviors [23]. This is particularly true among older adults, who are more likely to have co-occurring health conditions than younger adults. Future research should seek to explore this effect in trials of MHBC interventions.

Conclusions

A wide range of eHealth-based MHBC interventions were effective in improving diet, aiding smoking cessation, and increasing daily step counts. However, effect sizes were small, and the overall quality of the evidence was low. Further high-quality research is needed to develop eHealth interventions that are effective in simultaneously modifying multiple risk factors for chronic disease, including substance use and sleep outcomes. It is recommended that eHealth-based MHBC interventions be included in the guidelines to improve the quality of life of older persons and reduce the risk of chronic diseases in later life.

Acknowledgments

The authors would like to express their sincere gratitude to the authors who provided detailed information and data regarding their studies.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BS drafted the manuscript. BS and QH were responsible for the concept and design of the study. BS, GL, and QH screened all abstracts as well as full texts, extracted all data, performed the risk-of-bias evaluation, and conducted the quality assessment and statistical analyses. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e58174_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 23 KB - jmir_v26i1e58174_app2.docx]

Multimedia Appendix 3
The measurement tools used for different outcomes.
[DOCX File , 16 KB - jmir_v26i1e58174_app3.docx]

Multimedia Appendix 4
Risk-of-bias summary.
[DOCX File , 443 KB - jmir_v26i1e58174_app4.docx]

Multimedia Appendix 5
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) assessment results.
[DOCX File , 24 KB - jmir_v26i1e58174_app5.docx]

Multimedia Appendix 6
Forest plots.
[DOCX File , 1746 KB - jmir_v26i1e58174_app6.docx]

Multimedia Appendix 7
Sensitivity analysis.
[DOCX File , 531 KB - jmir_v26i1e58174_app7.docx]

References

1. The state of aging and health in America. Centers for Disease Control and Prevention. 2011. URL: https://www.cdc.gov/aging/data/stateofaging.htm [accessed 2024-04-29]
2. 2008-2013 action plan for the global strategy for the prevention and control of noncommunicable diseases : prevent and control cardiovascular diseases, cancers, chronic respiratory diseases and diabetes. World Health Organization. 2009. URL: https://www.who.int/publications/i/item/9789241597418 [accessed 2024-04-29]
3. Lloyd-Jones DM, Allen NB, Anderson CA, Black T, Brewer LC, Foraker RE, et al. Life's essential 8: updating and enhancing the American heart association's construct of cardiovascular health: a presidential advisory from the American heart association. Circulation 2022 Aug 02;146(5). [doi: 10.1161/cir.0000000000001078]
4. Ross R, Tremblay M. Introduction to the Canadian 24-hour movement guidelines for adults aged 18-64 years and adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2)):v-xi. [doi: 10.1139/apnm-2020-0843] [Medline: 33054330]
5. Draper C, Tomaz S, Biersteker L, Cook CJ, Couper J, de Milander M, et al. The South African 24-hour movement guidelines for birth to 5 years: an integration of physical activity, sitting behavior, screen time, and sleep. J Phys Act Health 2020 Jan 01;17(1):109-119 [FREE Full text] [doi: 10.1123/jpah.2019-0187] [Medline: 31877557]

XSL•FO
RenderX

6. Tremblay MS, Carson V, Chaput JP. Introduction to the Canadian 24-hour movement guidelines for children and youth: an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2016 Jun;41(6 Suppl 3):iii-iiv. [doi: 10.1139/apnm-2016-0203] [Medline: 27306430]
7. Rattray B, Northey J, Pumpa K, Smee DJ, Cherbuin N, Anstey KJ. Objective assessment of physical activity, sedentary behaviour and sleep in older Australians. Medicine and science in sports and exercise. Med Sci Sports Exerc 2015;47(5):880. [doi: 10.1249/01.mss.0000479121.32482.2a]
8. Cunningham C, O' Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta-analyses. Scand J Med Sci Sports 2020 May;30(5):816-827. [doi: 10.1111/sms.13616] [Medline: 32020713]
9. Hu Z, Qin L, Kaminga AC, Xu H. Relationship between multiple lifestyle behaviors and health-related quality of life among elderly individuals with prediabetes in rural communities in China: a STROBE-compliant article. Medicine (Baltimore) 2020 Apr;99(15):e19560 [FREE Full text] [doi: 10.1097/MD.0000000000019560] [Medline: 32282705]
10. Myint PK, Luben RN, Wareham NJ, Bingham SA, Khaw KT. Combined effect of health behaviours and risk of first ever stroke in 20,040 men and women over 11 years' follow-up in Norfolk cohort of European prospective investigation of cancer (EPIC Norfolk): prospective population study. BMJ 2009 Feb 19;338(feb19 2):b349 [FREE Full text] [doi: 10.1136/bmj.b349] [Medline: 19228771]
11. Meader N, King K, Moe-Byrne T, Wright K, Graham H, Petticrew M, et al. A systematic review on the clustering and co-occurrence of multiple risk behaviours. BMC Public Health 2016 Jul 29;16(1):657 [FREE Full text] [doi: 10.1186/s12889-016-3373-6] [Medline: 27473458]
12. Khaw KT, Wareham N, Bingham S, Welch A, Luben R, Day N. Combined impact of health behaviours and mortality in men and women: the EPIC-Norfolk prospective population study. PLoS Med 2008 Jan 08;5(1):e12 [FREE Full text] [doi: 10.1371/journal.pmed.0050012] [Medline: 18184033]
13. Han H, Cao Y, Feng C, Zheng Y, Dhana K, Zhu S, et al. Association of a healthy lifestyle with all-cause and cause-specific mortality among individuals with type 2 diabetes: a prospective study in UK biobank. Diabetes Care 2022 Feb 01;45(2):319-329. [doi: 10.2337/dc21-1512] [Medline: 34857534]
14. Lu Q, Zhang Y, Geng T, Yang K, Guo K, Min X, et al. Association of lifestyle factors and antihypertensive medication use with risk of all-cause and cause-specific mortality among adults with hypertension in China. JAMA Netw Open 2022 Feb 01;5(2):e2146118 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.46118] [Medline: 35103793]
15. Richardson KM, Jospe MR, Saleh AA, Clarke TN, Bedoya AR, Behrens N, et al. Use of biological feedback as a health behavior change technique in adults: scoping review. J Med Internet Res 2023 Sep 25;25:e44359 [FREE Full text] [doi: 10.2196/44359] [Medline: 37747766]
16. Gómez-Gómez I, Barquero-Jiménez C, Johnson E, Conejo-Cerón S, Moreno-Peral P, Bellón J, et al. Effectiveness of multiple health behavior change interventions in reducing symptoms of anxiety in the adult population: a systematic review and meta-analysis of randomized controlled trials. Prev Med 2024 Mar;180:107847. [doi: 10.1016/j.ypmed.2024.107847] [Medline: 38199592]
17. Burgess E, Hassmén P, Welvaert M, Pumpa KL. Behavioural treatment strategies improve adherence to lifestyle intervention programmes in adults with obesity: a systematic review and meta-analysis. Clin Obes 2017 Apr 15;7(2):105-114. [doi: 10.1111/cob.12180] [Medline: 28199047]
18. Wagner KH, Schwingshackl L, Draxler A, Franzke B. Impact of dietary and lifestyle interventions in elderly or people diagnosed with diabetes, metabolic disorders, cardiovascular disease, cancer and micronutrient deficiency on micronuclei frequency - a systematic review and meta-analysis. Mutat Res Rev Mutat Res 2021 Jan;787:108367 [FREE Full text] [doi: 10.1016/j.mrrev.2021.108367] [Medline: 34083034]
19. Amato K, Park E, Nigg CR. Prioritizing multiple health behavior change research topics: expert opinions in behavior change science. Transl Behav Med 2016 Jun;6(2):220-227 [FREE Full text] [doi: 10.1007/s13142-015-0381-5] [Medline: 27356992]
20. Champion KE, Newton NC, Gardner LA, Chapman C, Thornton L, Slade T, et al. Health4Life eHealth intervention to modify multiple lifestyle risk behaviours among adolescent students in Australia: a cluster-randomised controlled trial. Lancet Digit Health 2023 May;5(5):e276-e287 [FREE Full text] [doi: 10.1016/S2589-7500(23)00028-6] [Medline: 37032200]
21. Digital health. World Health Organization. 2021. URL: https://www.who.int/health-topics/digital-health#tab=tab_1 [accessed 2024-04-29]
22. Duan Y, Shang B, Liang W, Du G, Yang M, Rhodes RE. Effects of eHealth-based multiple health behavior change interventions on physical activity, healthy diet, and weight in people with noncommunicable diseases: systematic review and meta-analysis. J Med Internet Res 2021 Feb 22;23(2):e23786 [FREE Full text] [doi: 10.2196/23786] [Medline: 33616534]
23. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017 Sep 15;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
24. Norman GJ, Zabinski MF, Adams MA, Rosenberg DE, Yaroch AL, Atienza AA. A review of eHealth interventions for physical activity and dietary behavior change. Am J Prev Med 2007 Oct;33(4):336-345 [FREE Full text] [doi: 10.1016/j.amepre.2007.05.007] [Medline: 17888860]

XSL•FO
RenderX

25. Oosterveen E, Tzelepis F, Ashton L, Hutchesson MJ. A systematic review of eHealth behavioral interventions targeting smoking, nutrition, alcohol, physical activity and/or obesity for young adults. Prev Med 2017 Jun;99:197-206. [doi: 10.1016/j.ypmed.2017.01.009] [Medline: 28130046]
26. Champion KE, Parmenter B, McGowan C, Spring B, Wafford QE, Gardner LA, et al. Effectiveness of school-based eHealth interventions to prevent multiple lifestyle risk behaviours among adolescents: a systematic review and meta-analysis. Lancet Digit Health 2019 Sep;1(5):e206-e221. [doi: 10.1016/s2589-7500(19)30088-3]
27. Besnier F, Gayda M, Nigam A, Juneau M, Bherer L. Cardiac rehabilitation during quarantine in COVID-19 pandemic: challenges for center-based programs. Arch Phys Med Rehabil 2020 Oct;101(10):1835-1838 [FREE Full text] [doi: 10.1016/j.apmr.2020.06.004] [Medline: 32599060]
28. Kwan RY, Salihu D, Lee PH, Tse M, Cheung DS, Roopsawang I, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020 Apr 21;17(1):7 [FREE Full text] [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
29. Ding D, Rogers K, van der Ploeg H, Stamatakis E, Bauman AE. Traditional and emerging lifestyle risk behaviors and all-cause mortality in middle-aged and older adults: evidence from a large population-based Australian cohort. PLoS Med 2015 Dec 8;12(12):e1001917 [FREE Full text] [doi: 10.1371/journal.pmed.1001917] [Medline: 26645683]
30. Loef M, Walach H. The combined effects of healthy lifestyle behaviors on all cause mortality: a systematic review and meta-analysis. Prev Med 2012 Sep;55(3):163-170. [doi: 10.1016/j.ypmed.2012.06.017] [Medline: 22735042]
31. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
32. Michaelis R, Tang V, Wagner JL, Modi AC, LaFrance Jr WC, Goldstein LH, et al. Cochrane systematic review and meta-analysis of the impact of psychological treatments for people with epilepsy on health-related quality of life. Epilepsia 2018 Feb 03;59(2):315-332 [FREE Full text] [doi: 10.1111/epi.13989] [Medline: 29313968]
33. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
34. Abu-Saad K, Murad H, Barid R, Olmer L, Ziv A, Younis-Zeidan N, et al. Development and efficacy of an electronic, culturally adapted lifestyle counseling tool for improving diabetes-related dietary knowledge: randomized controlled trial among ethnic minority adults with type 2 diabetes mellitus. J Med Internet Res 2019 Oct 16;21(10):e13674 [FREE Full text] [doi: 10.2196/13674] [Medline: 31621640]
35. Babu V, Sylaja PN, Soman B, Varma RP, Ms M, Gl G, et al. A randomized controlled trial of medication adherence and management of risk factors for secondary prevention of stroke (MaMoRS) using a smartphone-based application. Int J Stroke 2024 Apr 22:17474930241245612. [doi: 10.1177/17474930241245612] [Medline: 38533606]
36. Bae JW, Woo SI, Lee J, Park SD, Kwon SW, Choi SH, et al. mHealth interventions for lifestyle and risk factor modification in coronary heart disease: randomized controlled trial. JMIR Mhealth Uhealth 2021 Sep 24;9(9):e29928 [FREE Full text] [doi: 10.2196/29928] [Medline: 34559058]
37. Bantum EO, Albright CL, White KK, Berenberg JL, Layi G, Ritter PL, et al. Surviving and thriving with cancer using a web-based health behavior change intervention: randomized controlled trial. J Med Internet Res 2014 Feb 24;16(2):e54 [FREE Full text] [doi: 10.2196/jmir.3020] [Medline: 24566820]
38. Bloom I, Welch L, Vassilev I, Rogers A, Jameson K, Cooper C, et al. Findings from an exploration of a social network intervention to promote diet quality and health behaviours in older adults with COPD: a feasibility study. Pilot Feasibility Stud 2020 Feb 06;6(1):15 [FREE Full text] [doi: 10.1186/s40814-020-0553-z] [Medline: 32042439]
39. Campbell MK, Carr C, Devellis B, Switzer B, Biddle A, Amamoo MA, et al. A randomized trial of tailoring and motivational interviewing to promote fruit and vegetable consumption for cancer prevention and control. Ann Behav Med 2009 Oct;38(2):71-85 [FREE Full text] [doi: 10.1007/s12160-009-9140-5] [Medline: 20012809]
40. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
41. Cicolini G, Simonetti V, Comparcini D, Celiberti I, Di Nicola M, Capasso LM, et al. Efficacy of a nurse-led email reminder program for cardiovascular prevention risk reduction in hypertensive patients: a randomized controlled trial. Int J Nurs Stud 2014 Jun;51(6):833-843. [doi: 10.1016/j.ijnurstu.2013.10.010] [Medline: 24225325]
42. Djuric Z, Ellsworth JS, Weldon AL, Ren J, Richardson CR, Resnicow K, et al. A diet and exercise intervention during chemotherapy for breast cancer. Open Obes J 2011 Jun 02;3(1):87-97 [FREE Full text] [doi: 10.2174/1876823701103010087] [Medline: 22238561]
43. Gallagher R, Chow CK, Parker H, Neubeck L, Celermajer DS, Redfern J, et al. The effect of a game-based mobile app 'MyHeartMate' to promote lifestyle change in coronary disease patients: a randomized controlled trial. Eur Heart J Digit Health 2023 Jan;4(1):33-42 [FREE Full text] [doi: 10.1093/ehjdh/ztac069] [Medline: 36743873]

XSL•FO
RenderX

44. Glasgow RE, Kurz D, King D, Dickman JM, Faber AJ, Halterman E, et al. Twelve-month outcomes of an internet-based diabetes self-management support program. Patient Educ Couns 2012 Apr;87(1):81-92 [FREE Full text] [doi: 10.1016/j.pec.2011.07.024] [Medline: 21924576]
45. Golshahi J, Ahmadzadeh H, Sadeghi M, Mohammadifard N, Pourmoghaddas A. Effect of self-care education on lifestyle modification, medication adherence and blood pressure in hypertensive adults: randomized controlled clinical trial. Adv Biomed Res 2015;4(1):204 [FREE Full text] [doi: 10.4103/2277-9175.166140] [Medline: 26601092]
46. Grey EB, Thompson D, Gillison FB. Effects of a web-based, evolutionary mismatch-framed intervention targeting physical activity and diet: a randomised controlled trial. Int J Behav Med 2019 Dec 25;26(6):645-657 [FREE Full text] [doi: 10.1007/s12529-019-09821-3] [Medline: 31654276]
47. Harrigan M, Cartmel B, Loftfield E, Sanft T, Chagpar AB, Zhou Y, et al. Randomized trial comparing telephone versus in-person weight loss counseling on body composition and circulating biomarkers in women treated for breast cancer: the lifestyle, exercise, and nutrition (LEAN) study. J Clin Oncol 2016 Mar 01;34(7):669-676 [FREE Full text] [doi: 10.1200/JCO.2015.61.6375] [Medline: 26598750]
48. Hawkes AL, Chambers SK, Pakenham KI, Patrao TA, Baade PD, Lynch BM, et al. Effects of a telephone-delivered multiple health behavior change intervention (CanChange) on health and behavioral outcomes in survivors of colorectal cancer: a randomized controlled trial. J Clin Oncol 2013 Jun 20;31(18):2313-2321. [doi: 10.1200/JCO.2012.45.5873] [Medline: 23690410]
49. Jane M, Hagger M, Foster J, Ho S, Kane R, Pal S. Effects of a weight management program delivered by social media on weight and metabolic syndrome risk factors in overweight and obese adults: a randomised controlled trial. PLoS One 2017;12(6):e0178326 [FREE Full text] [doi: 10.1371/journal.pone.0178326] [Medline: 28575048]
50. Jennings CA, Vandelanotte C, Caperchione CM, Mummery WK. Effectiveness of a web-based physical activity intervention for adults with type 2 diabetes-a randomised controlled trial. Prev Med 2014 Mar;60:33-40. [doi: 10.1016/j.ypmed.2013.12.011] [Medline: 24345601]
51. Johnston N, Bodegard J, Jerström S, Åkesson J, Brorsson H, Alfredsson J, et al. Effects of interactive patient smartphone support app on drug adherence and lifestyle changes in myocardial infarction patients: a randomized study. Am Heart J 2016 Aug;178:85-94 [FREE Full text] [doi: 10.1016/j.ahj.2016.05.005] [Medline: 27502855]
52. Kanera IM, Willems RA, Bolman CA, Mesters I, Verboon P, Lechner L. Long-term effects of a web-based cancer aftercare intervention on moderate physical activity and vegetable consumption among early cancer survivors: a randomized controlled trial. Int J Behav Nutr Phys Act 2017 Feb 10;14(1):19 [FREE Full text] [doi: 10.1186/s12966-017-0474-2] [Medline: 28187725]
53. Krebs P, Shtaynberger J, McCabe M, Iocolano M, Williams K, Shuk E, et al. An eHealth intervention to increase physical activity and healthy eating in older adult cancer survivors: summative evaluation results. JMIR Cancer 2017 Mar 01;3(1):e4 [FREE Full text] [doi: 10.2196/cancer.6435] [Medline: 28410171]
54. Lara J, O'Brien N, Godfrey A, Heaven B, Evans EH, Lloyd S, et al. Pilot randomised controlled trial of a web-based intervention to promote healthy eating, physical activity and meaningful social connections compared with usual care control in people of retirement age recruited from workplaces. PLoS One 2016 Jul 29;11(7):e0159703 [FREE Full text] [doi: 10.1371/journal.pone.0159703] [Medline: 27472560]
55. Li J, Szanton SL, McPhillips MV, Lukkahatai N, Pien GW, Chen KC, et al. An mHealth-facilitated personalized intervention for physical activity and sleep in community-dwelling older adults. J Aging Phys Act 2022 Apr 01;30(2):261-270. [doi: 10.1123/japa.2020-0463] [Medline: 34489366]
56. Liu S, Brooks D, Thomas SG, Eysenbach G, Nolan RP. Effectiveness of user- and expert-driven web-based hypertension programs: an RCT. Am J Prev Med 2018 Apr;54(4):576-583. [doi: 10.1016/j.amepre.2018.01.009] [Medline: 29456025]
57. Morey MC, Snyder DC, Sloane R, Cohen HJ, Peterson B, Hartman TJ, et al. Effects of home-based diet and exercise on functional outcomes among older, overweight long-term cancer survivors: RENEW: a randomized controlled trial. JAMA 2009 May 13;301(18):1883-1891 [FREE Full text] [doi: 10.1001/jama.2009.643] [Medline: 19436015]
58. Müller AM, Khoo S, Morris T. Text messaging for exercise promotion in older adults from an upper-middle-income country: randomized controlled trial. J Med Internet Res 2016 Jan 07;18(1):e5 [FREE Full text] [doi: 10.2196/jmir.5235] [Medline: 26742999]
59. Pischke CR, Voelcker-Rehage C, Ratz T, Peters M, Buck C, Meyer J, et al. Web-based versus print-based physical activity intervention for community-dwelling older adults: crossover randomized trial. JMIR Mhealth Uhealth 2022 Mar 23;10(3):e32212 [FREE Full text] [doi: 10.2196/32212] [Medline: 35319484]
60. Poppe L, de Bourdeaudhuij I, Verloigne M, Shadid S, van Cauwenberg J, Compernolle S, et al. Efficacy of a self-regulation-based electronic and mobile health intervention targeting an active lifestyle in adults having type 2 diabetes and in adults aged 50 years or older: two randomized controlled trials. J Med Internet Res 2019 Aug 02;21(8):e13363 [FREE Full text] [doi: 10.2196/13363] [Medline: 31376274]
61. Prabhakaran D, Jha D, Prieto-Merino D, Roy A, Singh K, Ajay VS, et al. Effectiveness of an mHealth-based electronic decision support system for integrated management of chronic conditions in primary care: the mWellcare cluster-randomized controlled trial. Circulation 2019 Jan 15;139(3):380-391 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.118.038192] [Medline: 30586732]

XSL•FO
RenderX

62. Sakane N, Suganuma A, Domichi M, Sukino S, Abe K, Fujisaki A, et al. The effect of a mHealth app (KENPO-app) for specific health guidance on weight changes in adults with obesity and hypertension: pilot randomized controlled trial. JMIR Mhealth Uhealth 2023 Apr 12;11:e43236 [FREE Full text] [doi: 10.2196/43236] [Medline: 37043287]
63. Swoboda CM, Miller CK, Wills CE. Setting single or multiple goals for diet and physical activity behaviors improves cardiovascular disease risk factors in adults with type 2 diabetes: a pragmatic pilot randomized trial. Diabetes Educ 2016 Aug 18;42(4):429-443. [doi: 10.1177/0145721716650043] [Medline: 27194048]
64. Taylor A, Taylor RS, Ingram W, Dean SG, Jolly K, Mutrie N, et al. Randomised controlled trial of an augmented exercise referral scheme using web-based behavioural support for inactive adults with chronic health conditions: the e-coachER trial. Br J Sports Med 2021 Apr;55(8):444-450 [FREE Full text] [doi: 10.1136/bjsports-2020-103121] [Medline: 33247001]
65. Wang X, Liu D, Du M, Hao R, Zheng H, Yan C. The role of text messaging intervention in inner Mongolia among patients with type 2 diabetes mellitus: a randomized controlled trial. BMC Med Inform Decis Mak 2020 May 14;20(1):90 [FREE Full text] [doi: 10.1186/s12911-020-01129-7] [Medline: 32410608]
66. Watson S, Woodside JV, Ware LJ, Hunter SJ, McGrath A, Cardwell CR, et al. Effect of a web-based behavior change program on weight loss and cardiovascular risk factors in overweight and obese adults at high risk of developing cardiovascular disease: randomized controlled trial. J Med Internet Res 2015 Jul 16;17(7):e177 [FREE Full text] [doi: 10.2196/jmir.3828] [Medline: 26183659]
67. Zheng X, Spatz ES, Bai X, Huo X, Ding Q, Horak P, et al. Effect of text messaging on risk factor management in patients with coronary heart disease: the CHAT randomized clinical trial. Circ Cardiovasc Qual Outcomes 2019 Apr;12(4):e005616. [doi: 10.1161/CIRCOUTCOMES.119.005616] [Medline: 30998400]
68. Tzelepis F, Mitchell A, Wilson L, Byrnes E, Haschek A, Leigh L, et al. The long-term effectiveness of internet-based interventions on multiple health risk behaviors: systematic review and robust variance estimation meta-analysis. J Med Internet Res 2021 Dec 21;23(12):e23513 [FREE Full text] [doi: 10.2196/23513] [Medline: 34931990]
69. D'Amore C, Reid JC, Chan M, Fan S, Huang A, Louie J, et al. Interventions including smart technology compared with face-to-face physical activity interventions in older adults: systematic review and meta-analysis. J Med Internet Res 2022 Oct 31;24(10):e36134 [FREE Full text] [doi: 10.2196/36134] [Medline: 36315229]
70. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? systematic review and meta-analysis. J Med Internet Res 2019 Mar 19;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
71. Schoeppe S, Alley S, van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 07;13(1):127 [FREE Full text] [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
72. Hagstromer M, Ainsworth BE, Oja P, Sjostrom M. Comparison of a subjective and an objective measure of physical activity in a population sample. J Phys Act Health 2010 Jul;7(4):541-550. [doi: 10.1123/jpah.7.4.541] [Medline: 20683097]
73. Nahm ES, Resnick B, Brown C, Zhu S, Magaziner J, Bellantoni M, et al. The effects of an online theory-based bone health program for older adults. J Appl Gerontol 2017 Sep 16;36(9):1117-1144 [FREE Full text] [doi: 10.1177/0733464815617284] [Medline: 26675352]
74. Barnett A, van den Hoek D, Barnett D, Cerin E. Measuring moderate-intensity walking in older adults using the ActiGraph accelerometer. BMC Geriatr 2016 Dec 08;16(1):211 [FREE Full text] [doi: 10.1186/s12877-016-0380-5] [Medline: 27931188]
75. Bauman A, Bull F, Chey T, Craig CL, Ainsworth BE, Sallis JF, et al. The international prevalence study on physical activity: results from 20 countries. Int J Behav Nutr Phys Act 2009 Mar 31;6:21 [FREE Full text] [doi: 10.1186/1479-5868-6-21] [Medline: 19335883]
76. Amireault S, Fong AJ, Sabiston CM. Promoting healthy eating and physical activity behaviors: a systematic review of multiple health behavior change interventions among cancer survivors. Am J Lifestyle Med 2018;12(3):184-199 [FREE Full text] [doi: 10.1177/1559827616661490] [Medline: 30202391]
77. Ochsner S, Scholz U, Hornung R. Testing phase-specific self-efficacy beliefs in the context of dietary behaviour change. Appl Psychol Health Well Being 2013 Mar 17;5(1):99-117. [doi: 10.1111/j.1758-0854.2012.01079.x] [Medline: 23457086]
78. Belogianni K, Baldwin C. Types of interventions targeting dietary, physical activity, and weight-related outcomes among university students: a systematic review of systematic reviews. Adv Nutr 2019 Sep 01;10(5):848-863 [FREE Full text] [doi: 10.1093/advances/nmz027] [Medline: 31181143]
79. MacArthur G, Caldwell DM, Redmore J, Watkins SH, Kipping R, White J, et al. Individual-, family-, and school-level interventions targeting multiple risk behaviours in young people. Cochrane Database Syst Rev 2018 Oct 05;10(10):CD009927 [FREE Full text] [doi: 10.1002/14651858.CD009927.pub2] [Medline: 30288738]
80. Holland RW, Aarts H, Langendam D. Breaking and creating habits on the working floor: a field-experiment on the power of implementation intentions. J Exp Soc Psychol 2006 Nov;42(6):776-783. [doi: 10.1016/j.jesp.2005.11.006]
81. Guo YQ, Chen Y, Dabbs AD, Wu Y. The effectiveness of smartphone app-based interventions for assisting smoking cessation: systematic review and meta-analysis. J Med Internet Res 2023 Apr 20;25:e43242 [FREE Full text] [doi: 10.2196/43242] [Medline: 37079352]

XSL•FO
RenderX

82. Etter JF, Khazaal Y. The Stop-tabac smartphone application for smoking cessation: a randomized controlled trial. Addiction 2022 May 19;117(5):1406-1415 [FREE Full text] [doi: 10.1111/add.15738] [Medline: 34738687]
83. Houston TK, Chen J, Amante DJ, Blok AC, Nagawa CS, Wijesundara JG, et al. Effect of technology-assisted brief abstinence game on long-term smoking cessation in individuals not yet ready to quit: a randomized clinical trial. JAMA Intern Med 2022 Mar 01;182(3):303-312 [FREE Full text] [doi: 10.1001/jamainternmed.2021.7866] [Medline: 35072714]
84. Cadigan JM, Haeny AM, Martens MP, Weaver CC, Takamatsu SK, Arterberry BJ. Personalized drinking feedback: a meta-analysis of in-person versus computer-delivered interventions. J Consult Clin Psychol 2015 Apr;83(2):430-437 [FREE Full text] [doi: 10.1037/a0038394] [Medline: 25486373]
85. Kant R, Yadav P, Bairwa M. Effectiveness of the internet-based versus face-to-face interaction on reduction of tobacco use among adults: a meta-analysis. Cureus 2021 Nov;13(11):e19380 [FREE Full text] [doi: 10.7759/cureus.19380] [Medline: 34925983]
86. Scott-Sheldon LA, Lantini R, Jennings EG, Thind H, Rosen RK, Salmoirago-Blotcher E, et al. Text messaging-based interventions for smoking cessation: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016 May 20;4(2):e49 [FREE Full text] [doi: 10.2196/mhealth.5436] [Medline: 27207211]
87. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
88. Gardner B, Lally P, Wardle J. Making health habitual: the psychology of 'habit-formation' and general practice. Br J Gen Pract 2012 Dec;62(605):664-666 [FREE Full text] [doi: 10.3399/bjgp12X659466] [Medline: 23211256]
89. Matson TE, Anderson ML, Renz AD, Greenwood-Hickman MA, McClure JB, Rosenberg DE. Changes in self-reported health and psychosocial outcomes in older adults enrolled in sedentary behavior intervention study. Am J Health Promot 2019 Sep;33(7):1053-1057. [doi: 10.1177/0890117119841405] [Medline: 30957508]
90. Greaves CJ, Sheppard KE, Abraham C, Hardeman W, Roden M, Evans PH, et al. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. BMC Public Health 2011 Feb 18;11:119 [FREE Full text] [doi: 10.1186/1471-2458-11-119] [Medline: 21333011]
91. Svensson T, Saito E, Svensson AK, Melander O, Orho-Melander M, Mimura M, et al. Association of sleep duration with all- and major-cause mortality among adults in Japan, China, Singapore, and Korea. JAMA Netw Open 2021 Sep 01;4(9):e2122837 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.22837] [Medline: 34477853]
92. Tobaldini E, Fiorelli EM, Solbiati M, Costantino G, Nobili L, Montano N. Short sleep duration and cardiometabolic risk: from pathophysiology to clinical evidence. Nat Rev Cardiol 2019 Apr 8;16(4):213-224. [doi: 10.1038/s41569-018-0109-6] [Medline: 30410106]
93. Howarth A, Quesada J, Silva J, Judycki S, Mills PR. The impact of digital health interventions on health-related outcomes in the workplace: a systematic review. Digit Health 2018;4:2055207618770861 [FREE Full text] [doi: 10.1177/2055207618770861] [Medline: 29942631]
94. Vanderlinden J, Boen F, van Uffelen JG. Effects of physical activity programs on sleep outcomes in older adults: a systematic review. Int J Behav Nutr Phys Act 2020 Feb 05;17(1):11 [FREE Full text] [doi: 10.1186/s12966-020-0913-3] [Medline: 32024532]
95. Lang C, Kalak N, Brand S, Holsboer-Trachsler E, Pühse U, Gerber M. The relationship between physical activity and sleep from mid adolescence to early adulthood. a systematic review of methodological approaches and meta-analysis. Sleep Med Rev 2016 Aug;28:32-45. [doi: 10.1016/j.smrv.2015.07.004] [Medline: 26447947]
96. Safe use of sleep aids. National Sleep Foundation. 2019. URL: https://www.sleepfoundation.org/sleep-aids [accessed 2024-04-29]
97. Rodriguez JC, Dzierzewski JM, Alessi CA. Sleep problems in the elderly. Med Clin North Am 2015 Mar;99(2):431-439 [FREE Full text] [doi: 10.1016/j.mcna.2014.11.013] [Medline: 25700593]

Abbreviations

GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MHBC: multiple health behavior change
MVPA: moderate to vigorous physical activity
OR: odds ratio
PA: physical activity
PICOS: Population, Intervention, Comparison, Outcomes, and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SB: sedentary behavior
SHBC: single health behavior change
SMD: standardized mean difference
TPA: total physical activity

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.03.24; peer-reviewed by A Ramadas, J Mistry; comments to author 29.04.24; revised version received 24.05.24; accepted 01.06.24; published 31.07.24.

Please cite as:
Shi B, Li G, Wu S, Ge H, Zhang X, Chen S, Pan Y, He Q
Assessing the Effectiveness of eHealth Interventions to Manage Multiple Lifestyle Risk Behaviors Among Older Adults: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e58174
URL: https://www.jmir.org/2024/1/e58174
doi:10.2196/58174
PMID:

©Beibei Shi, Guangkai Li, Shuang Wu, Hongli Ge, Xianliang Zhang, Si Chen, Yang Pan, Qiang He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Telehealth-Supported Exercise or Physical Activity Programs for Knee Osteoarthritis: Systematic Review and Meta-Analysis

Xiao-Na Xiang[1,2,3*], PhD; Ze-Zhang Wang[1,2,3*], BS; Jing Hu[1,2,3], DPT; Jiang-Yin Zhang[1,2,3], MS; Ke Li[4,5], PhD; Qi-Xu Chen[4,5], MS; Fa-Shu Xu[6], PhD; Yue-Wen Zhang[4,5], MS; Hong-Chen He[1,2,3], PhD; Cheng-Qi He[1,2,3], PhD; Si-Yi Zhu[1,2,3], PhD

[1]Rehabilitation Medicine Center and Institute of Rehabilitation Medicine, West China Hospital, Sichuan University, Chengdu, China
[2]Key Laboratory of Rehabilitation Medicine in Sichuan Province, West China Hospital, Sichuan University, Chengdu, China
[3]School of Rehabilitation Sciences, West China School of Medicine, Sichuan University, Chengdu, China
[4]Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China
[5]Joint Lab of Data Science and Business Intelligence, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China
[6]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
[*]these authors contributed equally

Corresponding Author:
Si-Yi Zhu, PhD
Rehabilitation Medicine Center and Institute of Rehabilitation Medicine
West China Hospital
Sichuan University
No 37, Guoxue Alley
Wuhou District
Chengdu, 610041
China
Phone: 86 028 85422847
Fax: 86 028 85422847
Email: hxkfzsy@scu.edu.cn

Abstract

Background: The integration of telehealth-supported programs in chronic disease management has become increasingly common. However, its effectiveness for individuals with knee osteoarthritis (KOA) remains unclear.

Objective: This study aimed to assess the effectiveness of telehealth-supported exercise or physical activity programs for individuals with KOA.

Methods: A comprehensive literature search encompassing Embase, MEDLINE, CENTRAL, Web of Science, PubMed, Scopus, PEDro, GreyNet, and medRxiv from inception to September 2023 was conducted to identify randomized controlled trials comparing telehealth-supported exercise or physical activity programs to a control condition for KOA. Data were extracted and qualitatively synthesized across eligible studies, and a meta-analysis was performed to evaluate the effects. The study was reported according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020.

Results: In total, 23 studies met eligibility criteria, with 20 included in the meta-analysis. Results showed that telehealth-supported exercise or physical activity programs reduced pain (g=–0.39; 95% CI –0.67 to –0.11; P<.001), improved physical activity (g=0.13; 95% CI 0.03-0.23; P=.01), and enhanced physical function (g=–0.51; 95% CI –0.98 to –0.05; P=.03). Moreover, significant improvements in quality of life (g=0.25; 95% CI 0.14-0.36; P<.001), self-efficacy for pain (g=0.72; 95% CI 0.53-0.91; P<.001), and global improvement (odds ratio 2.69, 95% CI 1.41-5.15; P<.001) were observed. However, self-efficacy for physical function (g=0.14; 95% CI –0.26 to 0.53; P=.50) showed insignificant improvements. Subgroup analyses based on the World Health Organization classification of digital health (pain: χ^2_2=6.5; P=.04 and physical function: χ^2_2=6.4; P=.04), the type of teletechnology in the intervention group (pain: χ^2_4=4.8; P=.31 and function: χ^2_4=13.0; P=.01), and active or inactive controls (pain: χ^2_1=5.3; P=.02 and physical function: χ^2_1=3.4; P=.07) showed significant subgroup differences.

Conclusions: Telehealth-supported exercise or physical activity programs might reduce knee pain and improve physical activity, physical function, quality of life, self-efficacy, and global improvement in individuals with KOA. Future research should consider

XSL•FO
RenderX

longer implementation durations and assess the feasibility of incorporating wearables and standardized components into large-scale interventions to evaluate the effects.

Trial Registration: PROSPERO CRD42022359658; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=359658

(J Med Internet Res 2024;26:e54876) doi:10.2196/54876

KEYWORDS

telehealth; knee osteoarthritis; physical activities; quality of life; systematic review and meta-analysis; systematic review; meta-analysis; knee; physical activity; exercise; chronic disease; chronic disease management; effectiveness; physical function; self-efficacy

Introduction

Osteoarthritis is a prevalent degenerative joint disease affecting more than 500 million individuals globally [1], with over 260 million individuals experiencing knee osteoarthritis (KOA) alone, resulting in significant health and socioeconomic burdens [2]. The prevalence of KOA among individuals aged 60 years is 26.8% [3], and projections indicate that by the year 2100, KOA will affect approximately 2.37 billion people aged 65 years and older and 866 million individuals aged 80 years and older worldwide [4]. At the early stage of KOA, pain and stiffness are the predominant symptoms; thus, management strategies aim to alleviate pain and increase functional capacities [5]. Physical activity reflects individuals' overall activity levels, while exercise denotes a planned, structured, and repetitive subset of physical activity [6]. The level of daily physical activity correlates positively with the physical function and quality of life of patients with KOA while also controlling pain intensity [7]. Clinical guidelines advocate for physical activity and exercise as the first-line management strategy for KOA [2,8,9]. Exercise and physical activity interventions have been shown to alleviate KOA symptoms and delay functional impairment and eventual joint replacement [10,11]. However, gaps exist in the clinical application of active lifestyle and exercise for KOA, with a tendency to overly rely on medication and surgery [12]. Effective interventions are desperately needed to address decreased function associated with an inactive lifestyle and aging. However, in-person health care services, particularly physical therapy, can be expensive in terms of time and other costs, such as consultation fees and transportation, for individuals with incurable KOA requiring long-term intervention [13]. Furthermore, lacking motivation to maintain an active lifestyle is also a barrier since adherence to home-based exercise decreased by 94.7% at 3 months following discharge [14].

Telehealth is defined as "the delivery and facilitation of health and health-related services including medical care, provider and patient education, health information services, and self-care via telecommunications and digital communication technologies" [15,16]. An accumulating body of evidence suggests that telehealth-supported exercise interventions have been proven as a preferable form of intervention, especially due to the "social distancing" requirement imposed by the COVID-19 pandemic. Hence, the need for advice or interventions via telehealth has soared [17,18]. However, because of its complicated operating system, ambiguous instructions, and need for Wi-Fi or cellular data, digital rehabilitation may not be as beneficial as face-to-face rehabilitation for people with KOA who are typically older.

The modes of remote rehabilitation are diverse, and many design factors, such as reminders, supervision, and communication, are closely related to the ultimate intervention outcomes. The high heterogeneity in the design of existing clinical trials on remote rehabilitation poses challenges in comparing the effectiveness of methods and summarizing experiences. Several reviews have attempted to evaluate the efficacy of telehealth-supported exercise programs in individuals with KOA [19-21], but few have focused on the efficacy of telehealth-supported physical activity programs. Our previous meta-analysis (n=4) [21] indicated that internet-based rehabilitation relieved pain in patients with KOA, but its effect on physical function was unclear due to the limited inclusion of original studies. Regarding the treatment effect of computer- or virtual reality–supported exercise on patients with KOA, another meta-analysis (n=12) [22] found no improvement in physical function, which might result from heterogeneity in exercise programs and experimental design. Additionally, these reviews did not examine how telehealth-supported exercise programs affect physical activity, self-efficacy in coping with symptoms, or global improvement experienced by patients.

To enhance our comprehension of the impact of telehealth-supported exercise or physical activity programs in individuals with KOA, as examined by multiple studies [23-32], we conducted a systematic review and meta-analysis. The objective of this investigation was to assess the effect of telehealth-supported exercise or physical activity programs on pain, physical activity, physical function, self-efficacy, quality of life, and global improvement with a comprehensive bias assessment. Additionally, the study applied the World Health Organization (WHO) classification of digital health [33] to articulate the functionalities of each program and provided an analysis of the minimally important differences (MIDs), which are important considerations in clinical decision-making.

Methods

Selection Process, Search Strategy, and Eligibility Criteria

The review protocol was registered with PROSPERO (CRD42022359658) and reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendations, version 2020 (PRISMA checklist is present in Multimedia Appendix 1) [34]. Multimedia Appendix 2 contained a list of modifications to the study

XSL•FO
RenderX

protocol. For randomized controlled trials (RCTs) published in English-language peer-reviewed journals, the following databases were searched: Embase (via OVID platform), MEDLINE (via OVID platform), CENTRAL (via the Cochrane Library), Web of Science, PubMed, Scopus, and PEDro from inception to September 2023. The specialist registers GreyNet (GreyNet International) and medRxiv (Cold Spring Harbor Laboratory) were searched for gray literature. To identify possibly pertinent studies, the reference lists of studies included in the full-text screening process were manually searched. We developed a search strategy, and the full search strategy is listed in Multimedia Appendix 3. The entire search process was assisted by a librarian from Sichuan University.

A wide definition of a telehealth-supported structured exercise or physical activity programs was established, covering interventions delivery via telephone (voice calls), SMS text messages, mobile app (app-based), internet (web-based), and wearable device (electronics). Studies were included within the Participants, Intervention, Comparison, Outcome, and Study Design (PICOS) framework (Textbox 1). Studies were excluded if the research simultaneously addressed other forms of arthritis or included unclear statistical data. Studies identified from literature research were imported into Review Manager (version 5.4; Nordic Cochrane Centre, Cochrane Collaboration). Titles and abstracts were independently screened by 2 authors (XNX and ZZW) to identify studies for full-text screening. Any disagreements were resolved through discussion under the guidance of a third reviewer (SYZ). The process for full-text screening remained consistent.

Textbox 1. Inclusion criteria according to the Participants, Intervention, Comparison, Outcome, and Study Design framework.

Participants:

- Participants regardless of age with a diagnosis of knee osteoarthritis

Intervention:

- Telehealth-supported structured exercise or physical activity programs delivered by telephone, SMS, mobile app, internet, and wearable device, or applications combined with wearable devices

Comparison:

- Telehealth-supported programs without exercise or physical activity, waiting list, or nontelemedicine interventions (ie, usual care, conventional exercise programs, and patient education)

Outcomes:

- Primary outcomes were pain, physical activity, and physical function and secondary outcomes were quality of life, self-efficacy for pain and function, and overall global improvement

Study design:

- Randomized controlled trial

Data Extraction

Two independent authors (ZZW and JYZ) extracted data (author, year of publication, country, participants, intervention content and duration, forms of monitoring, forms of telehealth, delivery model, the WHO classification of digital health, and results) with a standardized data template. Specifically, the WHO classification of digital health [33] was used to systematically categorize the telehealth interventions and support the synthesis of research and evidence. Accordingly, studies were classified into (1) interventions for clients, (2) interventions for health care providers, and (3) interventions for both clients and health care providers. Disagreements between the 2 reviewers were resolved through consensus, and if necessary, by consultation with a third reviewer (SYZ). For each outcome of interest, means, SDs, and sample sizes were extracted for each comparison. If SDs were missing for continuous data, other statistics (ie, 95% CI; SEs; and t, F, or P values) were used for the calculation of SD via the calculator tool from Review Manager.

Data Synthesis and Analysis

The effect sizes of each study were quantified using standardized mean differences (SMDs), which were computed by dividing the difference in means between the 2 groups by the pooled SD of the measurement [35]. The SMDs for each study were obtained from the changes in outcome measures prior to and after the intervention indicating the intervention's influence on the outcome measures.

Data analysis was performed with Review Manager and R (version 4.2.1; R Foundation for Statistical Computing). Hedges g with a 95% CI was used to analyze continuous variables, irrespective of whether specific outcomes were identified. Heterogeneity was assessed with Cochrane Q statistic (significance level at $P<.10$) and quantified with I^2 (substantial heterogeneity at $I^2>50\%$) [36,37]. The presumed variability across the included studies led to the application of the random-effects model. The results of fixed-effect model results were reported when heterogeneity was absent ($\tau^2=0$). Otherwise, the between-study differences were explained with the results of the random-effects model. Egger regression test, Begg rank correlation test, and funnel plot of the primary outcomes were

XSL•FO
RenderX

used to assess the potential publication bias [38]. If the test for asymmetry was significant, the trim-and-fill method was used to address missing studies and estimate the pooled effect to adjust for possible bias. Influence analysis was used to identify outliers. Sensitivity analyses were performed on primary outcomes to confirm robustness, using the fixed-effect model and implementing the "leave-one-out" method [39], excluding outliers. Methodological assistance was provided by a researcher from the MAGIC China Center or Cochrane China Center at West China Hospital, Sichuan University.

Meta-Analysis

A meta-regression test and subgroup analysis of the primary outcomes (pain, physical activity, and physical function) were conducted to identify factors contributing to heterogeneity. The meta-regression test considered variables that might influence the intervention effect and heterogeneity. Based on the regression results, selected factors were used for the subgroup analysis. Hedges *g* cut-off points of 0.20, 0.50, and 0.80, respectively, represented a small, moderate, and large effect. A *P* value <.05 was deemed statistically significant [35]. Hedges *g* and representative SDs (pooled from the intervention and control groups in trials using the scale) were used to calculate the MIDs for primary outcomes [40], which were then compared to the reported MIDs. The anchor-based estimates were applied when no MID was reported [41].

Quality Assessment

The Cochrane Collaboration's risk of bias tool, specifically Vision 2, was used to assess bias. We assessed biases in the following domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result [42]. Every element was classified as low, some concerns, or high risk. Furthermore, the PEDro scale was used to assess the quality of included studies [43].

Quality of Evidence Assessment

The GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach was applied to evaluate the certainty of evidence for each outcome. The overall certainty of evidence for each outcome was graded as high, moderate, low, or very low. Evidence was downgraded by 1 level for each serious problem identified in the domains of risk of bias, inconsistency (substantial heterogeneity: I^2>50%), indirectness, imprecision (such as small sample size), and publication bias [44].

Results

Study Selection

A total of 14,081 papers were initially identified from databases, with an additional 13 retrieved. After removing duplicates, 4021 records were screened for titles and abstracts. Subsequently, 86 full-text papers were assessed for eligibility. Of these, 23 papers [23-32,45-59] were included in the systematic review (Figure 1). Excluded studies at the full-text screening stage are listed in Multimedia Appendix 4, with reasons for exclusion. Three papers [46,58,59] were excluded from the meta-analysis due to uncalculated SDs, resulting in the inclusion of 20 (87%) papers.

XSL•FO
RenderX

Figure 1. Flow diagram describing the literature review process. PICOS: Participants, Intervention, Comparison, Outcome, and Study Design; RCT: randomized controlled trial.

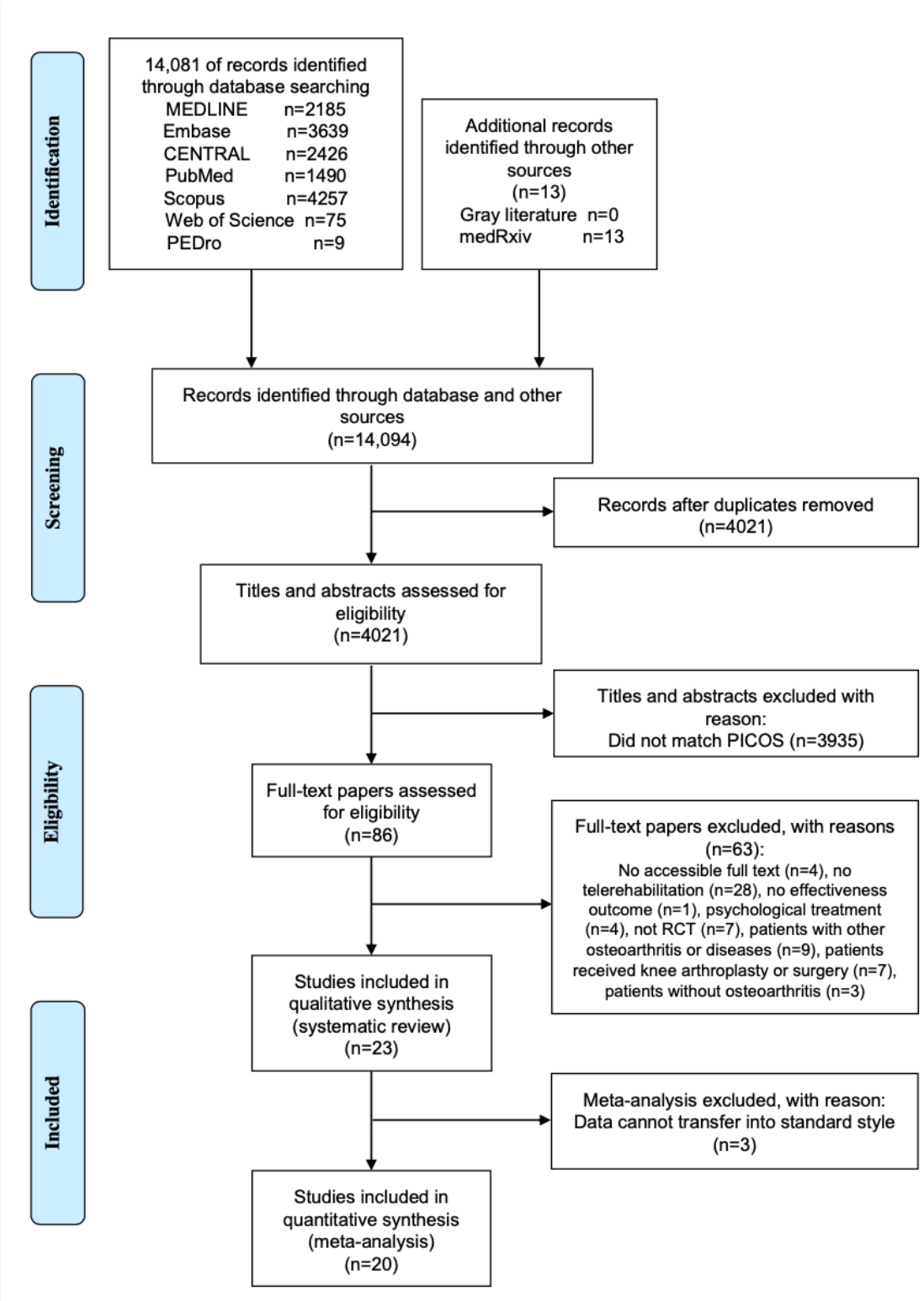

Characteristics of Included Studies

The characteristics of the included studies are presented in Multimedia Appendix 5 [24-32,45-54,56-59]. A total of 23 studies involving 3824 patients, of whom 2294 (60%) were female, were included in the systematic review. One study [26] exclusively examined the effect of telehealth-supported programs on the female population, while other studies recruited participants of both genders. Among these studies, 2 (9%) studies were conducted in Europe, 5 (22%) in North America, 5 (22%) in Asia, 9 (39%) in Australia, and 2 (9%) in Africa. The study period varied between 4 and 96 weeks, with 7 (30%) trials performing the telehealth-based intervention for less than 3 months, 6 (26%) for 3 to 6 months, and 10 (39%) for more than 6 months. A total of 13 (57%) studies provided telehealth-supported exercise programs, 4 (17%) provided

XSL•FO
RenderX

physical activity programs, and 6 (26%) provided treatments in combination. Interventions were delivered through various digital technologies, including mobile app (n=4), telephone (n=4), internet-based platforms (n=4), SMS text messages (n=2), and combinations (n=9). Moreover, studies tested different factors related to program design and delivery. Various forms of monitoring were performed, such as reminder alone (n=4), remote coaching alone (n=3), remote monitoring alone (n=3), combined remote reminder and monitoring (n=2), combined remote coaching and monitoring (n=7), and fewer studies focused on combined remote reminder and coaching (n=1) or combined all (n=1). Concerning interaction during the delivery of telehealth, 8 (35%) studies delivered through virtual contact (non–face-to-face contact), no interacting contact occurred in 5 (22%) studies, 5 (22%) studies used mixed forms, and 1 (4%) study used in-person delivery.

The primary outcomes included pain, physical activity, and physical function. In the 23 included studies, pain was measured using the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain subscale (n=12), Visual Analogue Scale (n=2), Knee Injury and Osteoarthritis Outcome Score (KOOS) pain subscale (n=5), or Numeric Pain Rating Scale (n=10). Physical activity levels were assessed using the Physical Activity Scale for the Elderly (n=6), time spent in daily moderate to vigorous physical activity (n=2), or the International Physical Exercise Questionnaire (n=1). Physical function was evaluated using the WOMAC function subscale (n=15), KOOS function subscale (n=2), Timed Up and Go test (n=6), or Ibadan Knee/Hip Osteoarthritis Outcome Measure (n=1).

The secondary outcomes included quality of life, self-efficacy for pain and function, and global improvement. Self-efficacy was measured using the Arthritis Self-Efficacy Scale, specifically its pain (n=6) and function (n=4) subscales. Quality of life was assessed using either the KOOS Quality of Life subscale (n=4) or the Assessment of Quality of Life (n=7). The global improvement was evaluated based on self-reported scores on a 7-point Likert scale (n=4). The outcome measurements identified in the included studies are detailed in Multimedia Appendix 5.

Risk of Bias

The risk of bias analysis is presented in Multimedia Appendices 6 and 7 [24-32,45-54,56-59]. In total, 19 [24,25,29,30,32,45-59] of the 23 papers followed intention-to-treat analysis, while the remaining 4 [26-28,31] followed per-protocol analysis. Biases across 5 domains were identified and reported in corresponding studies. Concerns regarding the randomization process were noted in 6 (26%) studies [25,27,28,46,57,58]. Given the inherent difficulty in blinding participants in telehealth-based exercise and physical activity programs, concerns arose in the "deviations from intended interventions" domain among 12 (52.17%) studies [24,26-28,30,31,46,51,52,57-59]. All studies have reported strategies for incomplete outcome data. One (4%) study [26] exhibited a high risk of bias in outcome measurement due to insufficient information regarding blinded assessment. Concerns regarding the selection of the reported result were raised in 2 (9%) studies [49,51]. Overall, the risk of bias judgment indicated high risk in 2 (8.70%) studies [26,58] and low risks in 10 (43.48%) studies [29,32,45,47-50,53,54,56]. Meanwhile, the assessment outcomes of the PEDro scale are presented in Multimedia Appendix 7. Each of the 23 papers included in the study received PEDro scores exceeding 5 points. Notably, 12 papers obtained PEDro scores ranging between 6 and 8 points, thus falling within the classification of "good." Furthermore, 11 papers achieved PEDro scores equal to or exceeding 9 points, indicating an "excellent" quality level.

Main Analyses About Effects of Telehealth-Based Exercise or Physical Activity Programs

Overview

Meta-analysis results of the effects of telehealth-based exercise or physical activity programs on primary outcomes are presented in Table 1. The GRADE summary of findings is listed in Multimedia Appendix 8.

XSL•FO
RenderX

Table 1. Meta-analysis results of the primary effects of telehealth-based exercise or physical activity programs.

Primary effects	RCTs[a], n	Hedges *g* (95% CI)	*P* value of Egger regression test	*P* value of Begg rank correlation test	Calculated MID[b]	Reported MID	Quality of evidence (GRADE[c])[d]
Pain	19	–0.39 (–0.67 to –0.11)	.41	.92	1.3	2.0	⊕⊕ [e,f]
Physical activity	9	0.13 (0.03 to 0.23)	.46	.40	9.0	46.0	⊕⊕ [e,g]
Physical function	18	–0.51 (–0.98 to –0.05)	.19	.73	5.3	10.1	⊕⊕ [e,h]

[a]RCT: randomized controlled trial.

[b]MID: minimally important difference.

[c]GRADE: Grading of Recommendations Assessment, Development and Evaluation.

[d]GRADE Working Group grades of evidence: High quality (⊕⊕⊕⊕): Further research is very unlikely to change our confidence in the estimate of effect. Moderate quality (⊕⊕⊕): Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. Low quality (⊕⊕): Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. Very low quality (⊕): We are very uncertain about the estimate.

[e]Downgraded for risk of bias: Participants and personnel were unblended.

[f]Downgraded for inconsistency: Considerable heterogeneity (I^2=83%).

[g]Downgraded for imprecision: Small sample size [54].

[h]Downgraded for inconsistency: Considerable heterogeneity (I^2=87%).

Pain

In 19 studies, a significant difference and a small effect size were observed (n=2512; *g*=–0.39; 95% CI –0.67 to –0.11; *P*<.001; forest plot Figure 2A [24,26-32,45,47-54,56,57]), indicating a favorable impact of the telehealth-based intervention on pain. However, substantial heterogeneity was noted (I^2=83%; τ^2=0.3498; *P*<.001). The calculated MID of pain was 1.3, which was smaller than the reported MID (2.0 units for the WOMAC pain subscale) [60]. Overall, the evidence suggests a low certainty that telehealth-based exercise or physical activity programs lead to a small yet statistically significant reduction in pain, although the clinical significance of this reduction might be limited.

XSL•FO
RenderX

Figure 2. Forest plot of (A) pain, (B) physical activity, and (C) physical function.

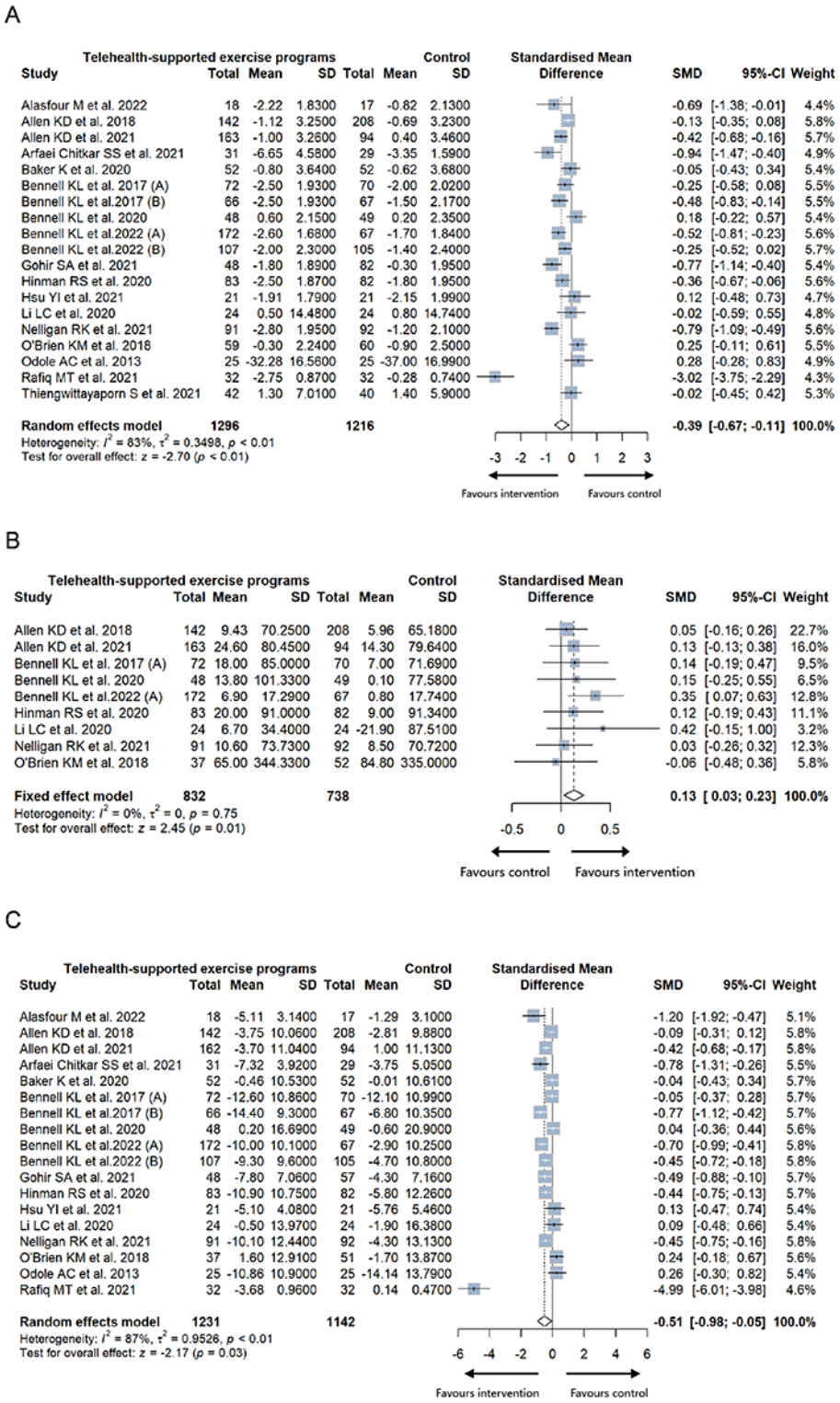

Physical Activity

The result of meta-analysis favored the telehealth-based intervention in promoting physical activity level of the population with KOA (n=1570; g=0.13; 95% CI 0.03-0.23; P=.01; forest plot Figure 2B) with negligible heterogeneity (I^2=0%; τ^2=0; P=.75). The calculated MID for physical activity, determined by Physical Activity Scale for the Elderly, was 9.0, which was less than the reported MID of 46.0 units [61]. Overall, the evidence was of low certainty and suggested that the telehealth-based programs might increase physical activity in a significant and very small way but not in a clinically meaningful manner.

Physical Function

The meta-analysis supported the telehealth-based intervention with a Hedges g effect size of –0.51 (n=2373; 95% CI –0.98 to –0.05; P=.03; forest plot Figure 2C) with considerable

XSL•FO
RenderX

heterogeneity (I^2=87%; τ^2=0.9526; P<.001) for improving physical function. The reported MID (10.1 units for the WOMAC physical function subscale) was higher than the calculated MID of 5.3 [62]. Overall, there was low-certainty evidence to suggest that telehealth-based programs could improve the physical function of the population with KOA to a moderately significant degree, though not reaching a clinically meaningful.

Secondary Outcomes

Compared to populations in control groups, a larger improvement was observed in terms of secondary outcomes within the intervention groups. These included quality of life (n=1301; g=0.25; 95% CI 0.14-0.37; P<.001; heterogeneity: I^2=5%; τ^2=0.0033; P=.39; see "Quality of life" in Multimedia Appendix 9 [24,25,27,29,48-54]), self-efficacy for pain (n=1337; g=0.73; 95% CI 0.52-0.94; P<.001; heterogeneity: I^2=4%; τ^2=0.0056; P=.39; see "Self-efficacy for pain" in Multimedia Appendix 9), and global improvement (n=1042; odds ratio 2.69, 95% CI 1.41-5.15; P<.001; heterogeneity: I^2=79%; τ^2=0.3296; P<.001; see "Global improvement" in Multimedia Appendix 9). However, a nonsignificant trend and moderate heterogeneity were observed for self-efficacy for physical function (n=578; g=0.14; 95% CI –0.26 to 0.53; P=.50; heterogeneity: I^2=52%; τ^2=0.0833; P=.10; see "Self-efficacy for physical function" in Multimedia Appendix 9).

Meta-Regression and Subgroup Analysis

A meta-regression test was conducted for several relative factors that might affect the intervention effect and heterogeneity, including coaching, monitoring, reminders, delivery form, intervention duration, sample size, quality of study, region of study, and other factors; details and results are presented in Multimedia Appendix 10. Notably, the regression results in type of teletechnology, WHO classification, and active or inactive control were significant.

Subgroup analyses based on the WHO classification revealed significant differences in pain (χ^2_2=6.5; P=.04; Figure 3A [24-32,45,47-54,56,57]) and physical function (χ^2_2=6.4; P=.04; Figure 3B). Specifically, within the subgroup of interventions for clients and health care providers, telehealth-based intervention demonstrated significant effects on both pain (g=–0.29; 95% CI –0.49 to –0.09; heterogeneity: I^2=63%; τ^2=0.0646; P<.001) and physical function (g=–0.36; 95% CI –0.63 to –0.08; heterogeneity: I^2=75%; τ^2=0.1293; P<.001).

XSL•FO
RenderX

Figure 3. Subgroup analysis of (A) pain and (B) physical function based on the World Health Organization classification.

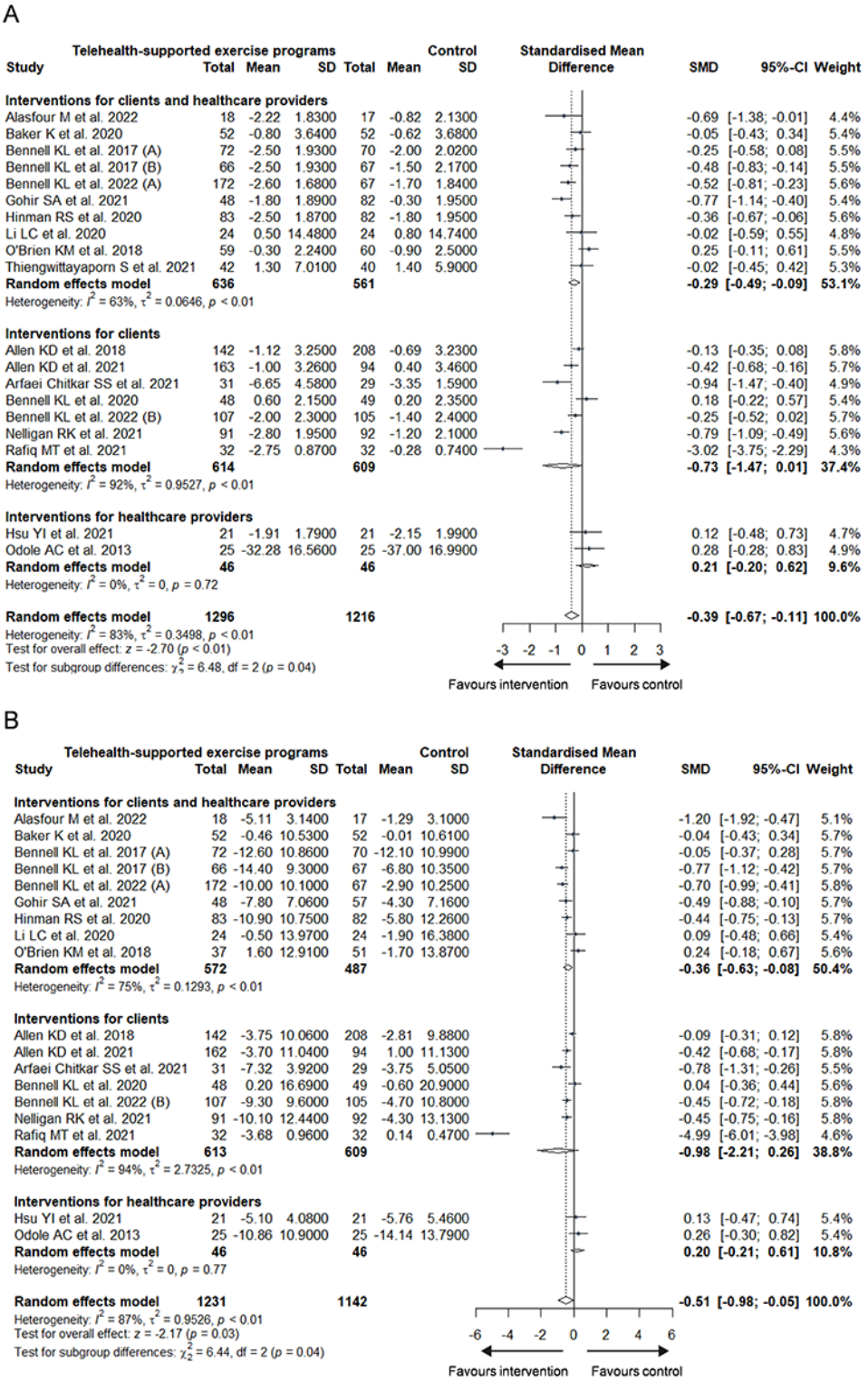

As for the type of digital technologies applied in the programs, the subgroup differences were significant in physical function (χ^2_4=13.0; P=.01; Figure 4B [24-32,45,47-54,56,57]) but not in pain (χ^2_4=4.8; P=.31; Figure 4A). Significant improvements were noticed in physical function across mobile app subgroup (g=–0.73; 95% CI –1.10 to –0.36; heterogeneity: I^2=34%; τ^2=0.0349; P=.22), internet subgroup (g=–0.42; 95% CI –0.80 to –0.04; heterogeneity: I^2=82%; τ^2=0.0920; P<.001), and mixed type of intervention (g=–0.28; 95% CI –0.54 to –0.02; heterogeneity: I^2=68%; τ^2=0.0839; P<.001).

XSL•FO
RenderX

Figure 4. Subgroup analysis of (A) pain and (B) physical function based on the type of digital technology.

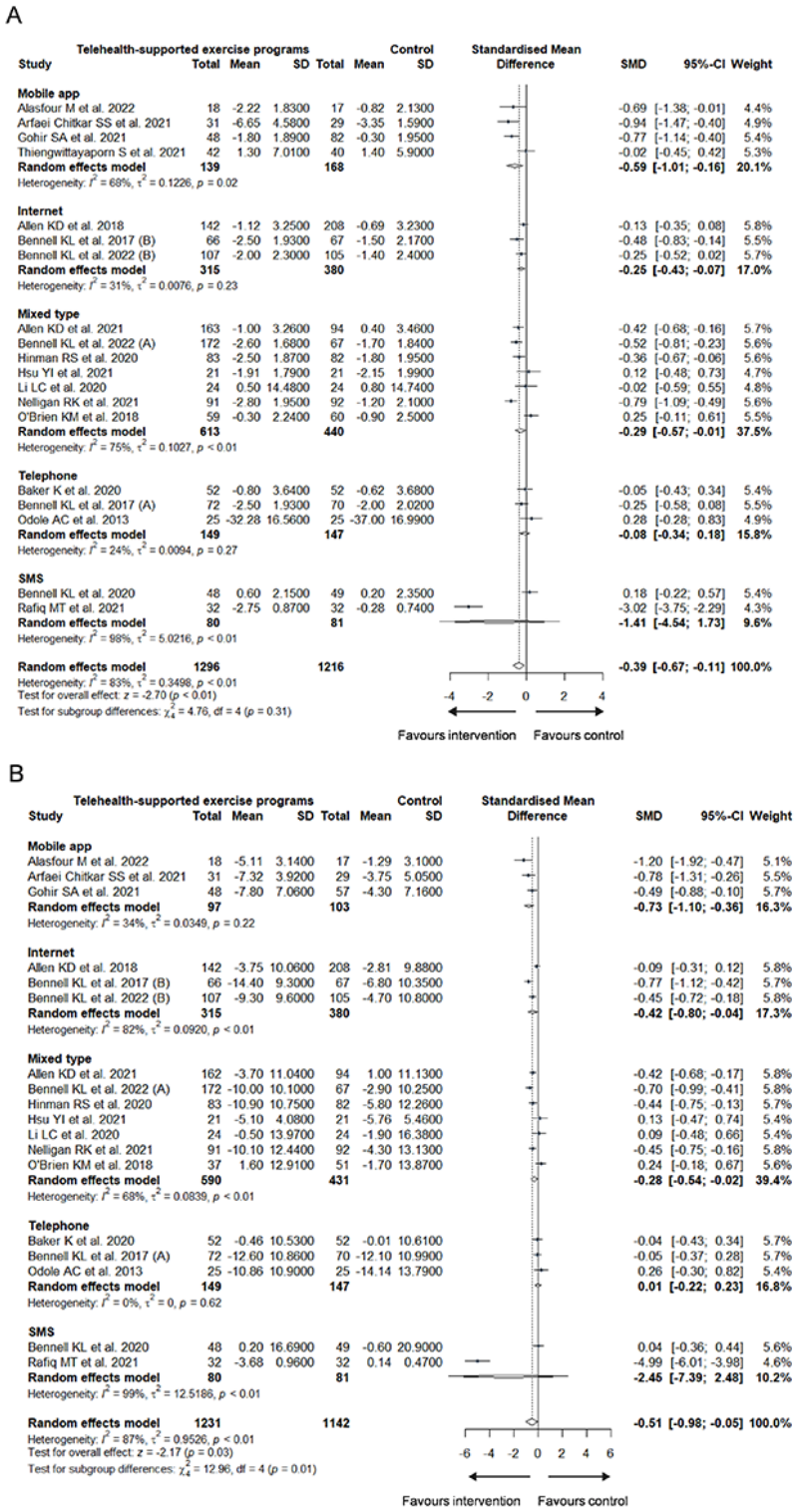

The forms of intervention in the control groups were divided into active controls (ie, exercise, physical therapy, pharmacotherapy, and self-management) or inactive controls (ie, education, usual care, and waitlist), which led to subgroup differences (pain: χ^2_1=5.3; *P*=.02; Figure 5A [24-32,45,47-54,56,57] and physical function: χ^2_1=3.4; *P*=.07; Figure 5B). Compared with inactive control groups, statistically significant pain reduction (*g*=–0.63; 95% CI –1.08 to –0.18; heterogeneity: I^2=87%; τ^2=0.5370; *P*<.001) and function improvement (*g*=–0.79; 95% CI –1.54 to –0.03; heterogeneity: I^2=90%; τ^2=0.9526; *P*<.001) were found in the intervention groups, while the differences between active controlled groups and intervention groups were not significant.

XSL•FO
RenderX

Figure 5. Subgroup analysis of (A) pain and (B) physical function based on the control group.

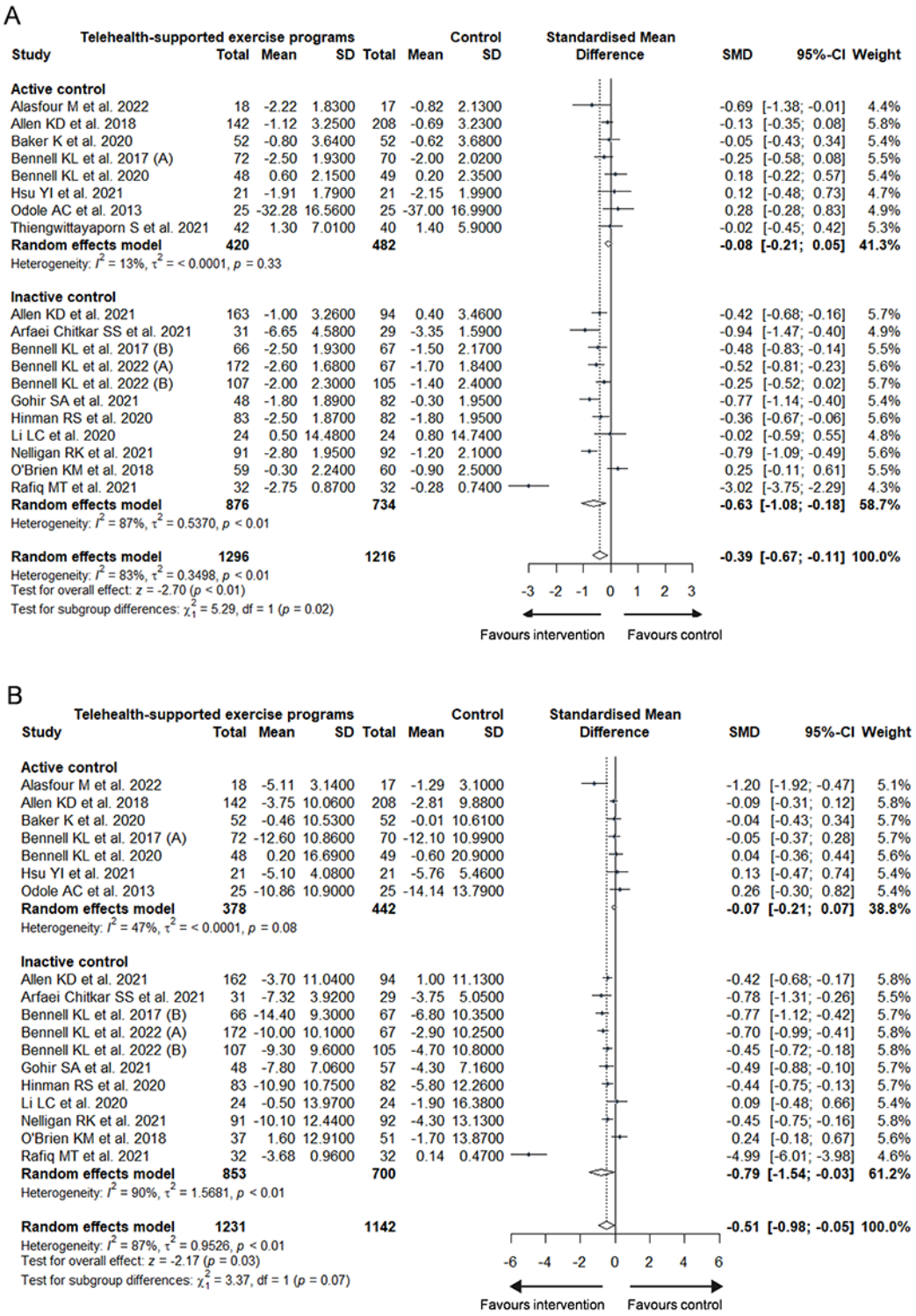

Outliers and Influence Analysis

Multimedia Appendix 11 [24-32,45,47-54,56,57] demonstrates the Baujat plot and influence analyses. The study from Rafiq et al [28] in the pain category was identified as a possible outlier. Meanwhile, studies of Bennell et al [24], Rafiq et al [28], and Allen et al [45] were considered as outliers for physical function.

Sensitivity Analysis

In the fixed-effect model result, a significant difference with a small effect size favoring telehealth-based intervention for pain relief was observed (see "Fix-effects model of pain" in Multimedia Appendix 12 [24,26-32,45,47-54,56,57]). After removing the outlier study [28], the pain relief effect size remained small (g=–0.28; 95% CI –0.44 to –0.11; see "Leave-one-out" analysis of pain" in Multimedia Appendix 12) with reduced heterogeneity (I^2=69%; τ^2=0.2849; P<.001).

XSL•FO
RenderX

Furthermore, even after excluding studies with an overall high risk of bias and the outlier, the effect size for pain relief remained small (g=–0.29; 95% CI –0.47 to –0.11; see "Sensitivity analysis omitted studies with high risk of bias and outliers of pain" in Multimedia Appendix 12) and was presented with heterogeneity (I^2=71%; τ^2=0.0846; P<.001), indicating the robustness of the finding.

For physical function, the positive effect of the telehealth-based intervention was supported by the result of meta-analysis using the fixed-effect model (see "Fix-effects model of physical function" in Multimedia Appendix 12). Subsequent "leave-one-out" analyses confirmed the overall significant effect size for physical function improvement, which remained small (g=–0.30; 95% CI –0.47 to –0.13; see "Leave-one-out" analyses of physical function" in Multimedia Appendix 12), with reduced heterogeneity (I^2=69%; τ^2=0.2895; P<.001) upon removal of outliers [24,28,45]. Even after removing studies with an overall high risk of bias and outliers, the effect size for physical function improvement remained small (g=–0.29; 95% CI –0.46 to –0.11; see "Sensitivity analysis omitted studies with high risk of bias and outliers of physical function" in Multimedia Appendix 12) and heterogeneity (I^2=62%; τ^2=0.0613; P<.001).

Publication Bias

Visual analysis of funnel plots for physical activity did not reveal publication bias. However, asymmetry was noted in studies reporting pain relief and improvement in physical function (Multimedia Appendix 13 [24,26-32,45,47-54,56,57]). Notably, studies pertaining to pain relief and physical function tended to have smaller SEs corresponding with larger sample sizes. More included studies tended to have greater SMD. The study with a small sample size from Rafiq et al [28] might contribute to this asymmetry. However, quantitative tests were unable to identify publication bias for pain relief or physical function. Specifically, neither the Egger regression test nor the Begg rank correlation test found evidence of publication bias (Table 1). No studies were trimmed or imputed in analyzing pain relief and improvement in physical function.

Discussion

Key Results

We discovered low-certainty evidence suggesting that telehealth-supported programs for population with KOA might have a small benefit on pain, a minimal positive impact on physical activity, and a moderate benefit on physical function, even though the changes in these primary outcomes were not clinically meaningful. Furthermore, this review revealed moderate-certainty evidence that telehealth-supported exercise or physical activity programs could lead to a statistically significant improvement in quality of life and self-efficacy for pain, with low-certainty evidence that telehealth-supported interventions enhanced global improvement. Low-certainty evidence, however, demonstrated that telehealth-supported programs had little impact on improving self-efficacy for physical function. In studies involving interventions for patients and health care providers, better results were observed in the group using telehealth-supported exercise or physical activity programs. Moreover, greater effects of the telehealth-supported exercise or physical activity programs were observed when these programs were delivered via mobile apps, the internet, or a combination of both. These effects were particularly notable in studies where the control group received inactive interventions. These findings suggested a definite role of telehealth-supported exercise or physical activity programs in the management of KOA.

Multiple factors influenced the efficacy of telehealth-based interventions, leading to high heterogeneity. We performed influence and subgroup analysis to identify the source of heterogeneity, identifying Rafiq et al [28] as a significant contributor, where notable improvements were observed in outcome measurement in the intervention group. It is possible that extra improvements observed in the study resulted from clients being reminded to complete their exercise sessions [28]. Reminders, monitoring, and coaching are considered important components of telehealth-supported programs [63], our analysis found no significant differences in subgroups according to the presence of these elements. Counseling or coaching could be beneficial for health information communication and helping people to make decisions [64]. While in-person contact with clinicians is often viewed as essential for providing accessible expertise, it may be controversial to use virtual contact skills between remote clients and health care providers, such as consultations or coaching [65], targeted client alerts and reminders [66], and remote client monitoring [67]. To better understand the clinical significance of telehealth-supported exercise or physical activity programs and standardize those formats and components, further studies with rigorous designs should investigate the impact of remote consultations, coaching, reminders, and monitoring.

Our study identified that targeted primary users, digital technology, and comparators were more important than intervention components. Since the WHO classification (version 1.0) [33] is a useful and effective tool for identifying the particulars of telemedicine, we used it to categorize telehealth-based programs. Our findings underscore the importance of patient and health care provider communication in improving pain and physical function outcomes. Additionally, control interventions could potentially be a source of heterogeneity. Patients engaging in telehealth-supported exercise or physical activity programs demonstrated superior outcomes compared to inactive controls and achieved comparable outcomes in pain reduction and physical function improvement to active controls. Consequently, patients might be motivated by the significant advantages of telehealth results that are obtained at a reasonable cost and with ease, especially in circumstances where medical resources may be limited [68].

Comparison With Other Studies

Several reviews have focused on digital health technologies in KOA, each offering unique insights. A narrative review [69] encompassing 91 studies found that digital health interventions were efficacious in enhancing patient education, promoting physical activity, and facilitating exercise interventions for patients with KOA. Similarly, another scoping review [70] concluded that digital health programs were comparably

XSL•FO
RenderX

beneficial to traditional therapy in ameliorating pain, enhancing physical function, and improving quality of life outcomes. However, these reviews did not provide quantified outcomes.

Prior systematic reviews and meta-analyses have synthesized the treatment effect of telehealth-based exercise or physical activity programs. Xiang et al [20] analyzed that 6 RCTs incorporated 8 different telerehabilitation strategies. Yang et al [71] and Chen et al [22] included 9 studies on telehealth-based exercise interventions (as of June 2021) and 12 RCTs on technology-supported exercise programs (as of August 2020), respectively. In total, 11 RCTs about telehealth-supported programs were reviewed to analyze the effect on pain by McHugh et al [19]. In our previous work, which involved 4 studies, we reported a positive impact of internet-based rehabilitation programs on pain of patients with KOA but not on physical function [21]; the meta-analysis by Xiang et al [20] supported this conclusion. However, in this study, we found that both pain and physical function were positively affected by digital exercise or physical activity programs. The differences in conclusions between our previous and current meta-analyses may stem from the increased number of included studies and the diversity of telehealth-supported programs. Moreover, in contrast to the preceding 2 meta-analytical studies [22,71], our research yielded positive results concerning physical function and quality of life. The subgroup analysis on digital technology in this review aligned with the findings from Yang et al [71], suggesting that programs delivered via websites or telephones might yield superior outcomes. The review conducted by McHugh et al [19] specifically focused on the pain symptom and described the distinction between inactive and active comparators without incorporating quantitative analysis. In contrast, this study provided a comprehensive subgroup analysis and demonstrated that the intervention group exhibited superior outcomes in terms of pain and physical functions when compared to the inactive comparator.

In addition to pain and physical function, this study also explored how telehealth exercise or physical activity programs affect the physical activity level of patients with KOA, a topic not extensively discussed in the previous meta-analysis. By focusing specifically on digital exercise or physical activity programs for KOA and their therapeutic effect on physical activity performance, this study expanded upon prior knowledge by including updated evidence and providing a comprehensive summary regarding the effectiveness of telehealth-based exercise or physical activity programs. It was notable that telehealth-based exercise or physical activity interventions led to a slight but significant improvement in physical activity. Furthermore, we used the MIDs to evaluate the variation of pain, physical activity, and physical function rather than solely focusing on statistical significance. This approach holds greater significance in the context of clinical reasoning.

Strengths and Limitations

This study has several key strengths. First, this review included telehealth-supported exercise or physical activity programs delivered through a wide variety of platforms, from SMS text messages and voice calls to internet-based applications or websites, providing a comprehensive understanding of the benefits of these programs. Second, the WHO classification of digital health interventions was introduced to label and describe the functions of each program, leading to new insight into subgroup analysis. Additionally, the study's systematic review focused on the effect of telehealth-supported exercise or physical activity programs on physical activity and physical function with self-efficacy, resulting in new evidence and identifying crucial factors for future research, which could facilitate the development of more effective telehealth-supported exercise or physical activity programs. Nevertheless, this study has some limitations. First, included RCTs were heterogeneous due to the highly variable designs of telehealth programs. The high level of heterogeneity observed may have resulted from the difficulty in applying thorough blinding techniques in RCTs for telehealth. Nonetheless, by using multiple bias analysis, it was possible to partially identify the source of heterogeneity; consequently, the conclusion can be deemed strong and conclusive. Second, the long-term benefits of telehealth-supported exercise or physical activity programs remained unknown due to the limited number of studies that completed long-term assessments beyond 3 months.

Future Research

Despite the established advantages of telehealth-supported exercise or physical activity programs, there are still obstacles to their widespread implementation as a first-line therapy for KOA, including the variation in technologies and management models. The WHO classification of digital health interventions aligns with the current intervention model and potential trend of digital technologies, which includes the nascent Internet of Things technologies, such as robotic-assisted training, wearable activity tracking, or visualized physical therapy [72]. Numerous recent experiments have used mobile apps and the internet, reflecting the growing trend toward using personal devices as telemedicine channels, in response to current interaction and technological advancements. Wearable devices, such as smartwatches or suits, are experiencing significant adoption, as they represent an innovative means of expanding health care services to everyday life situations. The rapid development of this technology, particularly direct-to-consumer wearable tracking technologies [73], holds promise for increasing patient acceptance and standardizing the application of remote technologies. There is a growing interest in various aspects related to remote technologies, such as the efficiency of transmitting information, the quality and consistency of interaction, and the cost or payment of new technologies [74]. These concerns may significantly facilitate the proliferation of telehealth-supported programs.

Conclusions

Low-certainty evidence from this systematic review suggested that patients with KOA might benefit from telehealth-supported exercise or physical activity programs in terms of reduced pain intensity, increased physical activity, and improved physical function, although the improvement may not be clinically meaningful. Additionally, moderate-certainty evidence suggested that telehealth-supported programs targeting KOA led to improvements in quality of life and self-efficacy for pain. The general health condition of the population with KOA was

XSL•FO
RenderX

improved by the telehealth-supported programs, though the level of certainty was low. However, this form of intervention was not effective in improving patients' self-efficacy for physical function. Future research should consider the application of wearable technologies and expand the intervention duration to examine the long-term effect. Most significantly, further research should standardize norms of telehealth-supported exercise or physical activity programs to improve evidence for clinical practice.

Acknowledgments

The authors thank Qing-Yang Shi and She-Yu Li at MAGIC China Center or Cochrane China Center, West China Hospital for the methodology assistance and the librarian team from Sichuan University. This study was supported by the National Natural Science Foundation of China (82272599), Natural Science Foundation of Sichuan Province (2024NSFSC0533), 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYGD23014), Sichuan University "Research Special Project on the Comprehensive Reform of Innovative Educational Practices Enabled by Artificial Intelligence" (2024-44), and National Innovation and Entrepreneurship Training Project for Undergraduates (C2023123256). The funders played no role in the design, conduct, or reporting of this study. We also extend our gratitude to Dr. Shaojun Zhang and his colleagues from Mianzhu People's Hospital (Deyang Clinical Research Center for Rehabilitation Medicine, Science and Technology Bureau of Deyang City, 2023-64) for their invaluable assistance in reviewing the manuscript.

Authors' Contributions

CQH, SYZ, and XNX were responsible for the conception, study design, and search strategy for this review. XNX and ZZW did all database searching and collating of results. XXN, ZZW, JYZ, and SYZ did the paper screening, data extraction, and critical appraisal. All authors contributed to conflict resolution during screening. ZZW, SYZ, and XNX were responsible for data curation. XNX, KL, QXC, YWZ, and FSX contributed to data analysis and data interpretation. XNX, JH, and SYZ drafted the manuscript. All authors contributed to reviewing and editing of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-aAnalyses) 2020 checklist.
[PDF File (Adobe PDF File), 1121 KB - jmir_v26i1e54876_app1.pdf]

Multimedia Appendix 2
Changes to the study protocol.
[DOCX File , 15 KB - jmir_v26i1e54876_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 25 KB - jmir_v26i1e54876_app3.docx]

Multimedia Appendix 4
List of studies excluded at full-text screening stage.
[DOCX File , 92 KB - jmir_v26i1e54876_app4.docx]

Multimedia Appendix 5
Characteristics of included studies.
[DOCX File , 33 KB - jmir_v26i1e54876_app5.docx]

Multimedia Appendix 6
Cochrane risk of bias graphical summary and Cochrane risk of bias summary per study.
[DOCX File , 598 KB - jmir_v26i1e54876_app6.docx]

Multimedia Appendix 7
PEDro scale quality assessment.
[DOCX File , 18 KB - jmir_v26i1e54876_app7.docx]

XSL•FO
RenderX

Multimedia Appendix 8
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) summary of findings.
[DOCX File , 20 KB - jmir_v26i1e54876_app8.docx]

Multimedia Appendix 9
Forest plots of secondary outcomes.
[PNG File , 1703 KB - jmir_v26i1e54876_app9.png]

Multimedia Appendix 10
Result of meta-regression test.
[DOCX File , 16 KB - jmir_v26i1e54876_app10.docx]

Multimedia Appendix 11
Results of outlier and influence analyses for primary outcomes.
[PNG File , 562 KB - jmir_v26i1e54876_app11.png]

Multimedia Appendix 12
Results of sensitivity analyses for primary outcomes.
[PNG File , 1928 KB - jmir_v26i1e54876_app12.png]

Multimedia Appendix 13
Funnel plots of primary outcomes.
[PNG File , 556 KB - jmir_v26i1e54876_app13.png]

References

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019 Apr 27;393(10182):1745-1759. [doi: 10.1016/S0140-6736(19)30417-9] [Medline: 31034380]
2. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American College of Rheumatology/Arthritis Foundation Guideline for the management of osteoarthritis of the hand, hip, and knee. Arthritis Care Res (Hoboken) 2020 Feb;72(2):149-162 [FREE Full text] [doi: 10.1002/acr.24131] [Medline: 31908149]
3. Nguyen UDT, Zhang Y, Zhu Y, Niu J, Zhang B, Felson DT. Increasing prevalence of knee pain and symptomatic knee osteoarthritis: survey and cohort data. Ann Intern Med 2011 Dec 06;155(11):725-732 [FREE Full text] [doi: 10.7326/0003-4819-155-11-201112060-00004] [Medline: 22147711]
4. Vollset SE, Goren E, Yuan C, Cao J, Smith AE, Hsiao T, et al. Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study. Lancet 2020 Oct 17;396(10258):1285-1306 [FREE Full text] [doi: 10.1016/S0140-6736(20)30677-2] [Medline: 32679112]
5. Bijlsma JWJ, Berenbaum F, Lafeber FPJG. Osteoarthritis: an update with relevance for clinical practice. Lancet 2011 Jun 18;377(9783):2115-2126. [doi: 10.1016/S0140-6736(11)60243-2] [Medline: 21684382]
6. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 1985;100(2):126-131 [FREE Full text] [Medline: 3920711]
7. Kraus V, Sprow K, Powell KE, Buchner D, Bloodgood B, Piercy K, et al. Effects of physical activity in knee and hip osteoarthritis: a systematic umbrella review. Med Sci Sports Exerc 2019 Jun;51(6):1324-1339 [FREE Full text] [doi: 10.1249/MSS.0000000000001944] [Medline: 31095089]
8. Biver E, Berenbaum F, Valdes AM, Araujo de Carvalho I, Bindels LB, Brandi ML, et al. Gut microbiota and osteoarthritis management: an expert consensus of the European Society for Clinical and Economic Aspects of Osteoporosis, Osteoarthritis and Musculoskeletal Diseases (ESCEO). Ageing Res Rev 2019 Nov;55:100946 [FREE Full text] [doi: 10.1016/j.arr.2019.100946] [Medline: 31437484]
9. Bannuru RR, Osani MC, Vaysbrot EE, Arden NK, Bennell K, Bierma-Zeinstra SMA, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. Osteoarthritis Cartilage 2019 Nov;27(11):1578-1589 [FREE Full text] [doi: 10.1016/j.joca.2019.06.011] [Medline: 31278997]
10. Beckwée D, Vaes P, Cnudde M, Swinnen E, Bautmans I. Osteoarthritis of the knee: why does exercise work? a qualitative study of the literature. Ageing Res Rev 2013 Jan;12(1):226-236. [doi: 10.1016/j.arr.2012.09.005] [Medline: 23026409]
11. Skou ST, Pedersen BK, Abbott JH, Patterson B, Barton C. Physical activity and exercise therapy benefit more than just symptoms and impairments in people with hip and knee osteoarthritis. J Orthop Sports Phys Ther 2018 Jun;48(6):439-447. [doi: 10.2519/jospt.2018.7877] [Medline: 29669488]
12. Allen KD, Golightly YM, White DK. Gaps in appropriate use of treatment strategies in osteoarthritis. Best Pract Res Clin Rheumatol 2017 Oct;31(5):746-759. [doi: 10.1016/j.berh.2018.05.003] [Medline: 30509418]

XSL•FO
RenderX

13. Petursdottir U, Arnadottir SA, Halldorsdottir S. Facilitators and barriers to exercising among people with osteoarthritis: a phenomenological study. Phys Ther 2010 Jul;90(7):1014-1025. [doi: 10.2522/ptj.20090217] [Medline: 20466741]
14. Nicolson PJA, Hinman RS, Kasza J, Bennell KL. Trajectories of adherence to home-based exercise programs among people with knee osteoarthritis. Osteoarthritis Cartilage 2018 Apr;26(4):513-521 [FREE Full text] [doi: 10.1016/j.joca.2018.01.009] [Medline: 29360592]
15. Tuckson RV, Edmunds M, Hodgkins ML. Telehealth. N Engl J Med 2017 Oct 19;377(16):1585-1592. [doi: 10.1056/NEJMsr1503323] [Medline: 29045204]
16. Catalyst NEJM. What is telehealth? NEJM Catalyst. 2018. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0268 [accessed 2024-07-09]
17. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
18. George MD, Danila MI, Watrous D, Reddy S, Alper J, Xie F, et al. Disruptions in rheumatology care and the rise of telehealth in response to the COVID-19 pandemic in a community practice-based network. Arthritis Care Res (Hoboken) 2021 Aug;73(8):1153-1161 [FREE Full text] [doi: 10.1002/acr.24626] [Medline: 33973389]
19. McHugh C, Kostic A, Katz J, Losina E. Effectiveness of remote exercise programs in reducing pain for patients with knee osteoarthritis: a systematic review of randomized trials. Osteoarthr Cartil Open 2022 Sep;4(3):100264 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100264] [Medline: 36474946]
20. Xiang W, Wang J, Ji B, Li L, Xiang H. Effectiveness of different telerehabilitation strategies on pain and physical function in patients with knee osteoarthritis: systematic review and meta-analysis. J Med Internet Res 2023 Dec 04;25:e40735 [FREE Full text] [doi: 10.2196/40735] [Medline: 37982411]
21. Xie S, Wang Q, Wang L, Wang L, Song K, He C. Effect of internet-based rehabilitation programs on improvement of pain and physical function in patients with knee osteoarthritis: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Jan 05;23(1):e21542 [FREE Full text] [doi: 10.2196/21542] [Medline: 33399542]
22. Chen T, Or CK, Chen J. Effects of technology-supported exercise programs on the knee pain, physical function, and quality of life of individuals with knee osteoarthritis and/or chronic knee pain: a systematic review and meta-analysis of randomized controlled trials. J Am Med Inform Assoc 2021 Feb 15;28(2):414-423 [FREE Full text] [doi: 10.1093/jamia/ocaa282] [Medline: 33236109]
23. Östlind E, Eek F, Stigmar K, Sant'Anna A, Hansson EE. Promoting work ability with a wearable activity tracker in working age individuals with hip and/or knee osteoarthritis: a randomized controlled trial. BMC Musculoskelet Disord 2022 Feb 03;23(1):112 [FREE Full text] [doi: 10.1186/s12891-022-05041-1] [Medline: 35114983]
24. Bennell KL, Lawford BJ, Keating C, Brown C, Kasza J, Mackenzie D, et al. Comparing video-based, telehealth-delivered exercise and weight loss programs with online education on outcomes of knee osteoarthritis: a randomized trial. Ann Intern Med 2022 Feb;175(2):198-209. [doi: 10.7326/M21-2388] [Medline: 34843383]
25. Egerton T, Bennell KL, McManus F, Lamb KE, Hinman RS. Comparative effect of two educational videos on self-efficacy and kinesiophobia in people with knee osteoarthritis: an online randomised controlled trial. Osteoarthritis Cartilage 2022 Oct;30(10):1398-1410 [FREE Full text] [doi: 10.1016/j.joca.2022.05.010] [Medline: 35750241]
26. Alasfour M, Almarwani M. The effect of innovative smartphone application on adherence to a home-based exercise programs for female older adults with knee osteoarthritis in Saudi Arabia: a randomized controlled trial. Disabil Rehabil 2022 Jun;44(11):2420-2427. [doi: 10.1080/09638288.2020.1836268] [Medline: 33103499]
27. Thiengwittayaporn S, Wattanapreechanon P, Sakon P, Peethong A, Ratisoontorn N, Charoenphandhu N, et al. Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial. Arch Orthop Trauma Surg 2023 Feb;143(2):729-738 [FREE Full text] [doi: 10.1007/s00402-021-04149-8] [Medline: 34453570]
28. Rafiq MT, Abdul Hamid MS, Hafiz E. The effect of rehabilitation protocol using mobile health in overweight and obese patients with knee osteoarthritis: a clinical trial. Adv Rheumatol 2021 Oct 24;61(1):63. [doi: 10.1186/s42358-021-00221-4] [Medline: 34689837]
29. Nelligan RK, Hinman RS, Kasza J, Crofts SJC, Bennell KL. Effects of a self-directed web-based strengthening exercise and physical activity program supported by automated text messages for people with knee osteoarthritis: a randomized clinical trial. JAMA Intern Med 2021 Jun 01;181(6):776-785 [FREE Full text] [doi: 10.1001/jamainternmed.2021.0991] [Medline: 33843948]
30. Hsu YI, Chen YC, Lee CL, Chang NJ. Effects of diet control and telemedicine-based resistance exercise intervention on patients with obesity and knee osteoarthritis: a randomized control trial. Int J Environ Res Public Health 2021 Jul 21;18(15):7744 [FREE Full text] [doi: 10.3390/ijerph18157744] [Medline: 34360036]
31. Arfaei Chitkar SS, Mohaddes Hakkak HR, Saadati H, Hosseini SH, Jafari Y, Ganji R. The effect of mobile-app-based instruction on the physical function of female patients with knee osteoarthritis: a parallel randomized controlled trial. BMC Womens Health 2021 Sep 14;21(1):333 [FREE Full text] [doi: 10.1186/s12905-021-01451-w] [Medline: 34521400]

XSL•FO
RenderX

32. Allen KD, Woolson S, Hoenig HM, Bongiorni D, Byrd J, Caves K, et al. Stepped exercise program for patients with knee osteoarthritis: a randomized controlled trial. Ann Intern Med 2021 Mar;174(3):298-307 [FREE Full text] [doi: 10.7326/M20-4447] [Medline: 33370174]
33. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2024-07-09]
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
35. Andrade C. Mean difference, standardized mean difference (SMD), and their use in meta-analysis: as simple as it gets. J Clin Psychiatry 2020 Sep 22;81(5):20f13681 [FREE Full text] [doi: 10.4088/JCP.20f13681] [Medline: 32965803]
36. Hoit G, Whelan DB, Dwyer T, Ajrawat P, Chahal J. Physiotherapy as an initial treatment option for femoroacetabular impingement: a systematic review of the literature and meta-analysis of 5 randomized controlled trials. Am J Sports Med 2020 Jul;48(8):2042-2050. [doi: 10.1177/0363546519882668] [Medline: 31774704]
37. Takeshima N, Sozu T, Tajika A, Ogawa Y, Hayasaka Y, Furukawa TA. Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference? BMC Med Res Methodol 2014 Feb 21;14:30 [FREE Full text] [doi: 10.1186/1471-2288-14-30] [Medline: 24559167]
38. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
39. Patsopoulos NA, Evangelou E, Ioannidis JP. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. Int J Epidemiol 2008 Oct;37(5):1148-1157 [FREE Full text] [doi: 10.1093/ije/dyn065] [Medline: 18424475]
40. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. BMJ 2019 Jan 22;364:k4817 [FREE Full text] [doi: 10.1136/bmj.k4817] [Medline: 30670455]
41. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020 Jun 04;369:m1714 [FREE Full text] [doi: 10.1136/bmj.m1714] [Medline: 32499297]
42. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
43. Cashin AG, McAuley JH. Clinimetrics: Physiotherapy Evidence Database (PEDro) Scale. J Physiother 2020 Jan;66(1):59 [FREE Full text] [doi: 10.1016/j.jphys.2019.08.005] [Medline: 31521549]
44. Ebadi S, Henschke N, Forogh B, Nakhostin Ansari N, van Tulder MW, Babaei-Ghazani A, et al. Therapeutic ultrasound for chronic low back pain. Cochrane Database Syst Rev 2020;7(7):CD009169 [FREE Full text] [doi: 10.1002/14651858.CD009169.pub3] [Medline: 32623724]
45. Allen KD, Arbeeva L, Callahan LF, Golightly YM, Goode AP, Heiderscheit BC, et al. Physical therapy vs internet-based exercise training for patients with knee osteoarthritis: results of a randomized controlled trial. Osteoarthritis Cartilage 2018 Mar;26(3):383-396 [FREE Full text] [doi: 10.1016/j.joca.2017.12.008] [Medline: 29307722]
46. Azma K, RezaSoltani Z, Rezaeimoghaddam F, Dadarkhah A, Mohsenolhosseini S. Efficacy of tele-rehabilitation compared with office-based physical therapy in patients with knee osteoarthritis: a randomized clinical trial. J Telemed Telecare 2018 Sep;24(8):560-565 [FREE Full text] [doi: 10.1177/1357633X17723368] [Medline: 28771070]
47. Baker K, LaValley MP, Brown C, Felson DT, Ledingham A, Keysor JJ. Efficacy of computer-based telephone counseling on long-term adherence to strength training in elderly patients with knee osteoarthritis: a randomized trial. Arthritis Care Res (Hoboken) 2020 Jul;72(7):982-990 [FREE Full text] [doi: 10.1002/acr.23921] [Medline: 31074576]
48. Bennell K, Nelligan RK, Schwartz S, Kasza J, Kimp A, Crofts SJ, et al. Behavior change text messages for home exercise adherence in knee osteoarthritis: randomized trial. J Med Internet Res 2020 Sep 28;22(9):e21749 [FREE Full text] [doi: 10.2196/21749] [Medline: 32985994]
49. Bennell KL, Campbell PK, Egerton T, Metcalf B, Kasza J, Forbes A, et al. Telephone coaching to enhance a home-based physical activity program for knee osteoarthritis: a randomized clinical trial. Arthritis Care Res (Hoboken) 2017 Jan;69(1):84-94. [doi: 10.1002/acr.22915] [Medline: 27111441]
50. Bennell KL, Nelligan R, Dobson F, Rini C, Keefe F, Kasza J, et al. Effectiveness of an internet-delivered exercise and pain-coping skills training intervention for persons with chronic knee pain: a randomized trial. Ann Intern Med 2017 Apr 04;166(7):453-462. [doi: 10.7326/M16-1714] [Medline: 28241215]
51. Bennell KL, Schwartz S, Teo PL, Hawkins S, Mackenzie D, McManus F, et al. Effectiveness of an unsupervised online yoga program on pain and function in people with knee osteoarthritis : a randomized clinical trial. Ann Intern Med 2022 Oct;175(10):1345-1355. [doi: 10.7326/M22-1761] [Medline: 36122378]
52. Gohir SA, Eek F, Kelly A, Abhishek A, Valdes AM. Effectiveness of internet-based exercises aimed at treating knee osteoarthritis: the iBEAT-OA randomized clinical trial. JAMA Netw Open 2021 Feb 01;4(2):e210012 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0012] [Medline: 33620447]

XSL•FO
RenderX

53. Hinman RS, Campbell PK, Lawford BJ, Briggs AM, Gale J, Bills C, et al. Does telephone-delivered exercise advice and support by physiotherapists improve pain and/or function in people with knee osteoarthritis? telecare randomised controlled trial. Br J Sports Med 2020 Jul;54(13):790-797. [doi: 10.1136/bjsports-2019-101183] [Medline: 31748198]
54. Li LC, Feehan LM, Xie H, Lu N, Shaw CD, Gromala D, et al. Effects of a 12-week multifaceted wearable-based program for people with knee osteoarthritis: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 03;8(7):e19116 [FREE Full text] [doi: 10.2196/19116] [Medline: 32618578]
55. Mecklenburg G, Smittenaar P, Erhart-Hledik JC, Perez DA, Hunter S. Effects of a 12-week digital care program for chronic knee pain on pain, mobility, and surgery risk: randomized controlled trial. J Med Internet Res 2018 Apr 25;20(4):e156 [FREE Full text] [doi: 10.2196/jmir.9667] [Medline: 29695370]
56. O'Brien KM, Wiggers J, Williams A, Campbell E, Hodder RK, Wolfenden L, et al. Telephone-based weight loss support for patients with knee osteoarthritis: a pragmatic randomised controlled trial. Osteoarthritis Cartilage 2018 Apr;26(4):485-494 [FREE Full text] [doi: 10.1016/j.joca.2018.01.003] [Medline: 29330101]
57. Odole AC, Ojo OD. A telephone-based physiotherapy intervention for patients with osteoarthritis of the knee. Int J Telerehabil 2013;5(2):11-20 [FREE Full text] [doi: 10.5195/ijt.2013.6125] [Medline: 25945214]
58. Odole AC, Ojo OD. Is telephysiotherapy an option for improved quality of life in patients with osteoarthritis of the knee? Int J Telemed Appl 2014;2014:903816 [FREE Full text] [doi: 10.1155/2014/903816] [Medline: 24778645]
59. Skrepnik N, Spitzer A, Altman R, Hoekstra J, Stewart J, Toselli R. Assessing the impact of a novel smartphone application compared with standard follow-up on mobility of patients with knee osteoarthritis following treatment with hylan G-F 20: a randomized controlled trial. JMIR Mhealth Uhealth 2017;5(5):e64 [FREE Full text] [doi: 10.2196/mhealth.7179] [Medline: 28487266]
60. Messier SP, Mihalko SL, Beavers DP, Nicklas BJ, DeVita P, Carr JJ, et al. Effect of high-intensity strength training on knee pain and knee joint compressive forces among adults with knee osteoarthritis: the START randomized clinical trial. JAMA 2021 Feb 16;325(7):646-657 [FREE Full text] [doi: 10.1001/jama.2021.0411] [Medline: 33591346]
61. Dupont J, Antonio L, Dedeyne L, O'Neill TW, Vanderschueren D, Rastrelli G, et al. Inflammatory markers are associated with quality of life, physical activity, and gait speed but not sarcopenia in aged men (40-79 years). J Cachexia Sarcopenia Muscle 2021 Dec;12(6):1818-1831 [FREE Full text] [doi: 10.1002/jcsm.12785] [Medline: 34523822]
62. Kim MS, Koh IJ, Choi KY, Sung YG, Park DC, Lee HJ, et al. The minimal clinically important difference (MCID) for the WOMAC and factors related to achievement of the MCID after medial opening wedge high tibial osteotomy for knee osteoarthritis. Am J Sports Med 2021 Jul;49(9):2406-2415. [doi: 10.1177/03635465211016853] [Medline: 34115533]
63. Mosnaim GS, Stempel DA, Gonzalez C, Adams B, BenIsrael-Olive N, Gondalia R, et al. The impact of patient self-monitoring via electronic medication monitor and mobile app plus remote clinician feedback on adherence to inhaled corticosteroids: a randomized controlled trial. J Allergy Clin Immunol Pract 2021 Apr;9(4):1586-1594. [doi: 10.1016/j.jaip.2020.10.064] [Medline: 33212239]
64. Gask L. Role of specialists in common chronic diseases. BMJ 2005 Mar 19;330(7492):651-653 [FREE Full text] [doi: 10.1136/bmj.330.7492.651] [Medline: 15774998]
65. Wilhelm S, Weingarden H, Greenberg JL, Hoeppner SS, Snorrason I, Bernstein EE, et al. Efficacy of app-based cognitive behavioral therapy for body dysmorphic disorder with coach support: initial randomized controlled clinical trial. Psychother Psychosom 2022;91(4):277-285 [FREE Full text] [doi: 10.1159/000524628] [Medline: 35588706]
66. Mahmud N, Asch DA, Sung J, Reitz C, Coniglio MS, McDonald C, et al. Effect of text messaging on bowel preparation and appointment attendance for outpatient colonoscopy: a randomized clinical trial. JAMA Netw Open 2021 Jan 04;4(1):e2034553 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.34553] [Medline: 33492374]
67. Mehta SJ, Hume E, Troxel AB, Reitz C, Norton L, Lacko H, et al. Effect of remote monitoring on discharge to home, return to activity, and rehospitalization after hip and knee arthroplasty: a randomized clinical trial. JAMA Netw Open 2020 Dec 01;3(12):e2028328 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.28328] [Medline: 33346847]
68. Duffy S, Lee TH. In-person health care as option B. N Engl J Med 2018 Jan 11;378(2):104-106. [doi: 10.1056/NEJMp1710735] [Medline: 29320653]
69. Shah N, Costello K, Mehta A, Kumar D. Applications of digital health technologies in knee osteoarthritis: narrative review. JMIR Rehabil Assist Technol 2022 Jun 08;9(2):e33489 [FREE Full text] [doi: 10.2196/33489] [Medline: 35675102]
70. Kitagawa T, Hayashi M. mHealth for the self-management of knee osteoarthritis: scoping review. J Med Internet Res 2023 May 08;25:e38798 [FREE Full text] [doi: 10.2196/38798] [Medline: 37155233]
71. Yang Y, Li S, Cai Y, Zhang Q, Ge P, Shang S, et al. Effectiveness of telehealth-based exercise interventions on pain, physical function and quality of life in patients with knee osteoarthritis: a meta-analysis. J Clin Nurs 2023 Jun;32(11-12):2505-2520. [doi: 10.1111/jocn.16388] [Medline: 35872635]
72. Masuki S, Morikawa M, Nose H. Internet of Things (IoT) system and field sensors for exercise intensity measurements. Compr Physiol 2020 Jul 08;10(3):1207-1240. [doi: 10.1002/cphy.c190010] [Medline: 32941686]
73. Fogel AL, Kvedar JC. Reported cases of medical malpractice in direct-to-consumer telemedicine. JAMA 2019 Apr 02;321(13):1309-1310 [FREE Full text] [doi: 10.1001/jama.2019.0395] [Medline: 30938788]
74. Romanick-Schmiedl S, Raghu G. Telemedicine—maintaining quality during times of transition. Nat Rev Dis Primers 2020 Jun 01;6(1):45 [FREE Full text] [doi: 10.1038/s41572-020-0185-x] [Medline: 32483168]

XSL•FO
RenderX

Abbreviations

GRADE: Grading of Recommendations Assessment, Development, and Evaluation
KOA: knee osteoarthritis
KOOS: Knee Injury and Osteoarthritis Outcome Score
MID: minimally important difference
PICOS: Participants, Intervention, Comparison, Outcome, and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference
WHO: World Health Organization
WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

Edited by N Cahill, T Leung; submitted 26.11.23; peer-reviewed by T Kitagawa, N Nakhostin Ansari, J Zhou, C Zhang; comments to author 17.04.24; revised version received 08.05.24; accepted 03.07.24; published 02.08.24.

Please cite as:
Xiang XN, Wang ZZ, Hu J, Zhang JY, Li K, Chen QX, Xu FS, Zhang YW, He HC, He CQ, Zhu SY
Telehealth-Supported Exercise or Physical Activity Programs for Knee Osteoarthritis: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e54876
URL: https://www.jmir.org/2024/1/e54876
doi:10.2196/54876
PMID:

©Xiao-Na Xiang, Ze-Zhang Wang, Jing Hu, Jiang-Yin Zhang, Ke Li, Qi-Xu Chen, Fa-Shu Xu, Yue-Wen Zhang, Hong-Chen He, Cheng-Qi He, Si-Yi Zhu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Implementing AI in Hospitals to Achieve a Learning Health System: Systematic Review of Current Enablers and Barriers

Amir Kamel Rahimi[1,2], BSc, MSc; Oliver Pienaar[3], BSc (Hons), MSc; Moji Ghadimi[3], PhD; Oliver J Canfell[1,2,4,5], PhD; Jason D Pole[1,6,7], PhD; Sally Shrapnel[1,3], PhD; Anton H van der Vegt[1], PhD; Clair Sullivan[1,8], MBBS(Hons), MD

[1]Queensland Digital Health Centre, Faculty of Medicine, The University of Queensland, Brisbane, Australia
[2]Digital Health Cooperative Research Centre, Australian Government, Sydney, Australia
[3]The School of Mathematics and Physics, The University of Queensland, Brisbane, Australia
[4]Business School, The University of Queensland, Brisbane, Australia
[5]Department of Nutritional Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom
[6]Dalla Lana School of Public Health, The University of Toronto, Toronto, ON, Canada
[7]ICES, Toronto, ON, Canada
[8]Metro North Hospital and Health Service, Department of Health, Queensland Government, Brisbane, Australia

Corresponding Author:
Amir Kamel Rahimi, BSc, MSc
Queensland Digital Health Centre
Faculty of Medicine
The University of Queensland
Health Sciences Building
Herston Campus
Brisbane, QLD4006
Australia
Phone: 61 0733465350
Email: amir.kamel@uq.edu.au

Abstract

Background: Efforts are underway to capitalize on the computational power of the data collected in electronic medical records (EMRs) to achieve a learning health system (LHS). Artificial intelligence (AI) in health care has promised to improve clinical outcomes, and many researchers are developing AI algorithms on retrospective data sets. Integrating these algorithms with real-time EMR data is rare. There is a poor understanding of the current enablers and barriers to empower this shift from data set–based use to real-time implementation of AI in health systems. Exploring these factors holds promise for uncovering actionable insights toward the successful integration of AI into clinical workflows.

Objective: The first objective was to conduct a systematic literature review to identify the evidence of enablers and barriers regarding the real-world implementation of AI in hospital settings. The second objective was to map the identified enablers and barriers to a 3-horizon framework to enable the successful digital health transformation of hospitals to achieve an LHS.

Methods: The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were adhered to. PubMed, Scopus, Web of Science, and IEEE Xplore were searched for studies published between January 2010 and January 2022. Articles with case studies and guidelines on the implementation of AI analytics in hospital settings using EMR data were included. We excluded studies conducted in primary and community care settings. Quality assessment of the identified papers was conducted using the Mixed Methods Appraisal Tool and ADAPTE frameworks. We coded evidence from the included studies that related to enablers of and barriers to AI implementation. The findings were mapped to the 3-horizon framework to provide a road map for hospitals to integrate AI analytics.

Results: Of the 1247 studies screened, 26 (2.09%) met the inclusion criteria. In total, 65% (17/26) of the studies implemented AI analytics for enhancing the care of hospitalized patients, whereas the remaining 35% (9/26) provided implementation guidelines. Of the final 26 papers, the quality of 21 (81%) was assessed as poor. A total of 28 enablers was identified; 8 (29%) were new in this study. A total of 18 barriers was identified; 5 (28%) were newly found. Most of these newly identified factors were related

XSL•FO
RenderX

to information and technology. Actionable recommendations for the implementation of AI toward achieving an LHS were provided by mapping the findings to a 3-horizon framework.

Conclusions: Significant issues exist in implementing AI in health care. Shifting from validating data sets to working with live data is challenging. This review incorporated the identified enablers and barriers into a 3-horizon framework, offering actionable recommendations for implementing AI analytics to achieve an LHS. The findings of this study can assist hospitals in steering their strategic planning toward successful adoption of AI.

(J Med Internet Res 2024;26:e49655) doi:10.2196/49655

KEYWORDS

life cycle; medical informatics; decision support system; clinical; electronic health records; artificial intelligence; machine learning; routinely collected health data

Introduction

Background

The growing adoption of electronic medical records (EMRs) in many high-income countries has resulted in improvements in health care delivery through the implementation of clinical decision support systems at the point of care [1]. To meet the ever-accelerating demands for clinical care, various innovative models have been developed to harness the potential of EMR data [2-4]. These new care models aim to enable health care organizations to achieve the quadruple aim of care, which includes enhancing patient experience, advancing providers' experience, improving the health of the population, and reducing health care costs [5].

Artificial intelligence (AI) holds the potential to improve health system outcomes by enhancing clinical decision support systems [6,7]. AI aims to augment human intelligence through complicated and iterative pattern recognition, generally on large data sets that exceed human abilities [8]. While a large body of academic literature has demonstrated the efficacy of AI models in various health domains, most of these models remain as proof of concept and have never been implemented in real-world workflows [9]. This demonstrates the relatively inconsequential endeavors of many AI studies that fail to produce any meaningful impact in the real world. Even with the substantial investments made by the health industry, the implementation of AI analytics in complex clinical practice is still at an early stage [10]. In a limited number of instances, AI has been successfully implemented, largely for nonclinical uses such as service planning or trained on limited static data sets such as chest x-rays or retinal photography [11]. The factors influencing the success or failure of AI implementations in health are poorly investigated [12]. Understanding these barriers and enablers increases the likelihood of successful implementation of AI for the digital transformation of the health system [13,14], ultimately aiding in achieving the quadruple aim of health care [5].

Toward the Digital Transformation of Health Care

A 3-horizon framework has been previously published to help health systems create an iterative pathway for successful digital health transformation (Figure 1 [15]). Horizon 1 aims to optimize the routine collection of patient data during every interaction with the health system. In horizon 2, the data collected during routine care are leveraged in real or near real time to create analytics. Finally, in horizon 3, the insights from data and digital innovations are collated to develop new models of care. A health care system focused on continuous improvement is referred to as a learning health system (LHS) that uses routinely collected data to monitor and enhance health care outcomes consistently [16]. When health care organizations reach the third horizon, they can leverage data in near real time to create ongoing learning iterations and enhance patient care, leading to the establishment of an LHS [17].

Figure 1. The 3-horizon framework for digital health transformation (adapted from Sullivan et al [15] with permission from CSIRO Publishing).

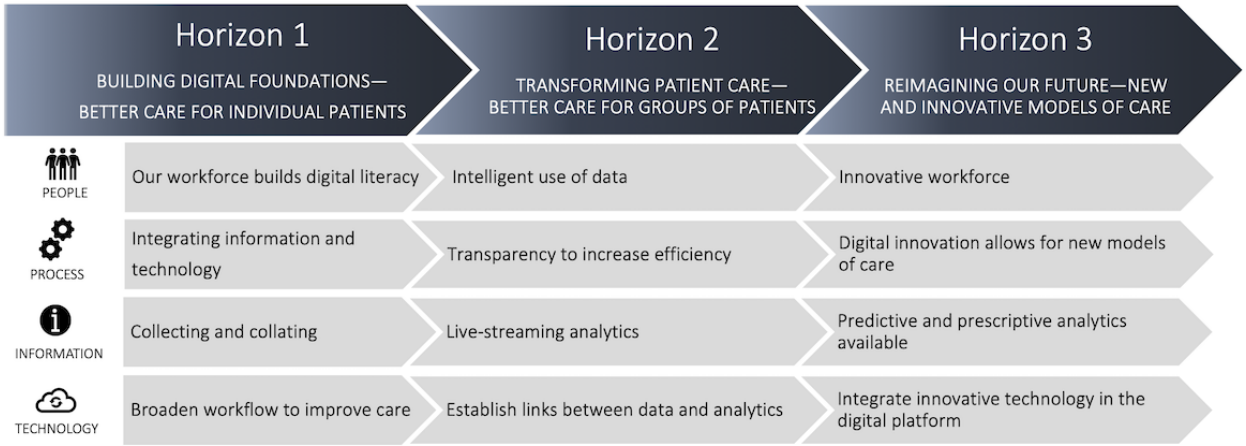

XSL•FO
RenderX

Regarding the 3-horizon model, EMRs are the foundation of horizon 1 (Figure 1). While many health organizations have successfully adopted EMRs into their existing workflows, the transition to horizons 2 and 3 has been challenging for many of these health care facilities [18]. A critical phase in this transition involves moving beyond the capture of EMR data for delivering analytics, including AI, aiming to improve clinical outcomes. There is little published evidence to assist health systems in making this transition [19,20].

Analysis of Prior Work

Before conducting our review, we performed a manual search on Google Scholar using our Medical Subject Heading (MeSH) terms along with the "review" keyword to identify previous review papers that aimed at reviewing studies on the implementation of clinical AI in health care settings. We also included review papers known to our research team. Between 2020 and 2022, we identified 4 reviews that were relevant to the implementation of AI in health care systems [21-24]. Overall, these papers reviewed 189 studies between 2010 and 2022. The characteristics of these reviews, outlined in Table 1, were the year of publication, the targeted care settings, the source of data, the predictive algorithm, and whether the predictive algorithm was implemented.

Table 1. The inclusion criteria for this study and previous work.

Study	Year	Health care setting	Data source	Predictive algorithm	Implementation state
Lee et al [22]	2020	Any	EMR[a]	Any	Implemented
Wolff et al [23]	2021	Any	Any	AI[b] and ML[c]	Implemented
Sharma et al [21]	2022	Any	Any	AI and ML	Implemented
Chomutare et al [24]	2022	Any	Any	AI and ML	Implemented or developed
Our study	2023	Hospitals	EMR	AI and ML	Implemented or guidelines

[a]EMR: electronic medical record.

[b]AI: artificial intelligence.

[c]ML: machine learning.

The prior works identified 20 enablers and 13 barriers to AI implementation in health care across 4 categories: people, process, information, and technology (Multimedia Appendix 1 [21-24]). Overall, the findings derived from these review papers hold significant potential in providing valuable insights for health systems to navigate the path toward digital health transformation. One prevailing shortcoming of these studies is the absence of alignment with evidence-based digital health transformation principles to provide health care organizations with actionable recommendations to enable an LHS [17], therefore limiting their applicability for strategic planning within hospital organizations.

Research Significance and Objectives

Hospitals are intricate hubs within the health care ecosystem, playing a central role in providing comprehensive medical care and acting as crucial pillars supporting the foundations of health care systems worldwide. Understanding the factors influencing the success or failure of AI in hospitals provides valuable insights to optimize the integration of these emerging technologies into hospital facilities. While the previous reviews included all health care settings [21-24], our study only focused on hospital settings. Given the limited instances regarding the implementation of AI in hospital facilities, this study explored the real-world case studies that have practically reported their AI implementation solutions in hospital facilities, aiming to synthesize the evidence of enablers and barriers within their implementation process. In addition to the inclusion of these implementation case studies, we incorporated implementation guidelines as they can potentially assist in the overall understanding of AI implementation in hospitals. This study also focused on aligning the evidence of enablers and barriers within the 3-horizon framework [15], offering a way to establish an empirical infrastructure. As a result, this can enable health care organizations to learn, adapt, and accelerate progress toward an LHS [25].

This review investigated the following research questions (RQs): (1) What enablers and barriers are identified for the successful implementation of AI with EMR data in hospitals? (RQ 1) and (2) How can the identified enablers of and barriers to AI implementation lead to actions that drive the digital transformation of hospitals? (RQ 2).

In addressing these questions, our objectives were to (1) conduct a systematic review of the literature to identify the evidence of enablers of and barriers to the real-world implementation of AI in hospital settings and (2) map the identified enablers and barriers to a 3-horizon framework to enable the successful digital health transformation of hospitals to achieve an LHS.

Methods

Search Strategy

This study followed an extended version of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to outline the review methodology with comprehensive details [26]. PubMed, Scopus, Web of Science, and IEEE Xplore were searched on April 13, 2022. We reviewed prior work to determine potential MeSH keywords relevant to our study [21-24]. A research librarian helped with the definition of the MeSH keywords in PubMed and the translation of that search strategy to all platforms searched. The search strategies were applied across the 4 databases (Multimedia Appendix 2). The MeSH keywords used to search

XSL•FO
RenderX

PubMed were as follows: *product lifecycle management*, *artificial intelligence*, *machine learning*, *deep learning*, *natural language processing*, *neural networks*, *computer*, *deep learning*, *big data*, *hospital*, *inpatient*, *medical*, *clinic*, *deploy*, *integrate*, *monitor*, *post prediction*, *data drift*, and *regulatory*. Using the Boolean operator *OR*, their synonyms were joined to form search phrases. Combining search phrases using the *AND* operator produced the final search string. We incorporated the term "data drift" to the title and abstract, and full-text search as it is a prominent concept for the continuous integration of AI. The term "regulatory" was also added to our search criteria because it is a relevant term for the implementation of AI in health care within the domain of software as a medical device. The reference lists of the included studies were examined to ensure that all relevant papers were included.

Eligibility Criteria

The inclusion criteria were articles published from January 1, 2010, to April 13, 2022, that included case studies and guidelines on the implementation of AI analytic tools in hospital settings using EMR data. Given the scarcity of real-world AI tools in hospital settings, especially the scarcity of published case studies of unsuccessful implementations of clinical AI tools, we specifically included case studies that successfully implemented AI within hospitals to understand lessons learned and provide use cases that other jurisdictions may learn from. On the basis of a review of frameworks for AI implementation in health care practice, we defined the term *implementation* as "an intentional effort designed to change or adapt or uptake interventions into routines" [19]. The term "barrier" was defined as "experiences that impeded, slowed, or made implementations difficult in some way" [20]. In contrast, the term *enablers* was defined as factors, experiences, or processes that facilitated the implementation process. Studies conducted in community or primary care settings were excluded as our main focus was hospital facilities. Studies that did not use AI models were also excluded. We also eliminated non–English-language and conference articles. Studies that focused on regulatory domains and challenges, opportunities, requirements, and recommendations were also excluded as they did not demonstrate real-world AI implementation. The selection of studies was based on the criteria specified in Textbox 1.

Textbox 1. Inclusion criteria for this study.

Inclusion criteria

- Population: adults (aged ≥18 y); inpatients
- Intervention: successfully implemented artificial intelligence (AI) and machine learning (ML) tools using hospital electronic medical record data
- Study design: case studies that implemented AI and ML in the real world; guidelines on the real-world implementation of AI and ML
- Publication date: January 2010 to April 2022
- Language: English

Exclusion criteria

- Population: nonadults (aged <18 y); outpatients
- Intervention: traditional statistical methods; rule-based systems; systems without AI and ML
- Study design: studies without implementation of AI and ML; studies focused on AI and ML development, regulatory-related domains, challenges, opportunities, and recommendations; conference papers; primary care or community settings
- Language: non-English

Screening

For the screening and data extraction procedures, the Covidence (Veritas Health Innovation) systematic review software was used [27]. A 2-stage screening process was performed with the involvement of 2 reviewers (AKR and OP). In the initial stage, the reviewers assessed the relevance of titles and abstracts based on the inclusion criteria. Subsequently, in the second stage, the full texts of the included articles were reviewed by AKR and OP independently. Consensus was reached through discussion between the reviewers whenever necessary.

Data Extraction and Synthesis

AKR and OP conducted the procedure of data extraction. The following study characteristics were extracted from all final included studies: country, clinical setting, study type (case study or guideline), and aim of study. With the adoption of EMR as a prerequisite for AI development, our focus was on extracting evidence of enablers and barriers solely within horizons 2 (implementation) and 3 (creating new models of care). In total, 2 reviewers (AKR and OP) independently extracted evidence regarding enablers and barriers (RQ 1), subsequently reaching consensus through weekly discussions and analysis. The extracted data were disseminated among our research team for review and to gather additional feedback.

To address the second RQ (RQ 2), we mapped the findings from previous reviews along with the found factors in this study across horizons 2 and 3 of the digital transformation framework [15]. Following the data extraction phase, 2 reviewers independently mapped the identified enablers and barriers to 4 categories (people, process, information, and technology). During the mapping of a given enabler or barrier, if it was related to the development of AI analytics, it was mapped to horizon 2 considering its relevance across the 4 domains (people, technology, information, and processes). When an enabler or barrier was associated with the postdevelopment phase focusing on establishing new care models, it was mapped to horizon 3.

XSL•FO
RenderX

Consensus was reached between AKR and OP through a meeting to finalize the mapping phase.

Quality Assessment

For the included use case studies, we used the Mixed Methods Appraisal Tool (MMAT) [28] to conduct a quality assessment. The choice of the MMAT was suitable as the included use case studies exhibited a range of qualitative, quantitative, and mixed methods designs. For evaluating the methodology of guideline studies, we followed the ADAPTE framework [29]. With 9 modules for guideline development, this framework was designed to streamline and enhance the process of creating guidelines within the health domain. The quality assessment was conducted independently by 2 authors (AKR and OP), and any discrepancies were resolved through a meeting.

Results

Study Selection

The search strategy retrieved 1247 papers from PubMed, Scopus, IEEE Xplore, and Web of Science for analysis, and 67 (5.37%) duplicates were identified and eliminated using the EndNote (Clarivate Analytics) citation manager. After screening titles and abstracts, 92.03% (1086/1180) of the studies were removed as the inclusion criteria were not satisfied. A total of 7.97% (94/1180) of the papers remained for full-text review following title and abstract screening. In total, 48% (45/94) of papers were excluded because AI models were not implemented in clinical care. A total of 19% (18/94) of the studies were excluded because they focused on regulatory domains. In total, 9% (8/94) of the studies were excluded due to being the wrong intervention (eg, studies that did not develop AI models). A total of 3% (3/94) of the studies were found to have a clinical population that did not align with our inclusion criteria (eg, hospitalized patients). One study was not in English and was excluded. In addition, 7 studies were discovered by scanning the reference lists of the included articles. In total, 26 studies were included in this review, comprising 9 (35%) guideline studies and 17 (65%) papers with successful implementation examples (Table 2). Figure 2 presents the PRISMA flow diagram outlining the outcomes of this review.

XSL•FO
RenderX

Table 2. Characteristics of the studies included in this review.

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
Wilson et al [30], 2021	United Kingdom	General	Guideline	• To provide advice from health care experts on clinical AI[a] development and implementation	• A team of multidisciplinary experts, including clinicians, software developers, data scientists, and hospital IT staff • Senior, experienced individuals can be particularly useful to overcome implementation barriers • The appointment of a data champion • Staff training in the data science field • Using data scientists or a trusted research environment with appropriate tools can ensure adequate data privacy • A common language with necessary terminologies is suggested within the CST[b] • Clinicians can assist in understanding and resolving the quality and reliability of AI solutions	• HCPs'[c] inexperience with AI • The integration of disparate data sources is one of the barriers to AI solutions in the current workflow
Svedberg et al [31], 2022	Sweden	General	Guideline	• To develop an AI implementation framework in health care • To conduct AI implementation studies to provide direction for further improvement of the framework • To implement the proposed framework in routine care	• The co-design process among clinicians, data scientists, and end users • The national and regional initiatives to facilitate AI implementation into practice • Several major investments facilitated the establishment of the infrastructure design and development of this study • Literature review and the existing theory-driven frameworks and strategies • Technological knowledge and awareness of challenges, including social, cultural, and organizational barriers	• Lack of accessibility of AI implementation science to individuals who could potentially benefit from it

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
Subbaswamy and Saria [32], 2019	United States	General	Guideline	• To explain data shift and overview the types of existing solutions	• Graphical representation can be used to assess the stability of AI models and identify potential performance shifts but requires domain for interpretation • Proactive learning approaches allow models to be stable against anticipated shifts in the future, including the use of stable algorithms that are robust to future shift	• Data set shift is prevalent and problematic in clinical AI settings and needs to be accounted for to prevent performance decay
Pianykh et al [33], 2020	United States	Radiology	Guideline	• To examine the key principles and issues involved in integrating AI with continuous learning in radiology	• Radiologists and clinicians are important to the successful implementation of continuous-learning AI to provide feedback • Continuous learning is a viable method to combat data drift	• Not specified
Leiner et al [34], 2021	The Netherlands	Radiology	Guideline	• To demonstrate the necessity for a vendor-neutral AI implementation infrastructure • To provide a plan for a vendor-neutral AI implementation infrastructure • To discuss prominent issues, including governance, quality control, and ethics	• A team of multidisciplinary experts, including clinicians, data scientists, and IT staff • Platforms are suggested as vendor-neutral infrastructures shared by researchers and clinicians and allow AI systems to receive iterative feedback from clinicians • The accessibility of the AI results at the time of care without requiring physicians to switch workstations or launch specialized software • Consistency between AI implementation methods used within one hospital • The messaging standards, such as HL7[d] • Using the containerization concept to concurrently run multiple instances of AI analytics • Training end users and clinicians for using and interpreting AI results	• Not specified
	Germany	General	Guideline			

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
Gruendner et al [35], 2019				• To implement a secure platform to develop and deploy ML[e] models in health care settings	• The FHIR[f] standard was used to exchange health data between different health care points in a consistent manner • The OMOP-CDM[g] database structure was used as a standard method to organize health care data consistently across various data points. This also enabled the availability of data to researchers and end users. • Containerization allowed for a flexible development environment. It enabled clinicians and ML developers to collaborate and improve performance. • The proposed platform provides scientists with a secure, privacy-preserving, flexible research infrastructure to develop and deploy statistical models within a hospital's IT infrastructure • Using appropriate data privacy techniques can allow for model training using data from multiple hospitals in parallel • Collaboration among the research team	• Predictions can be extremely slow with large input data due to hardware limitations; therefore, the AI may output results not in real time • Generalizable platforms such as KETOS are versatile, but as a result, they are relatively inefficient and may require further customizations and fine-tunings at the local level
Eche et al [36], 2021	United States	Radiology	Guideline	• To provide strategies to tackle overfitting and underspecification of AI models	• Underspecification (the lack of generalizability) can be addressed with the use of artificial or real shifts in test data	• Overfitting and underspecification can negatively impact the generalizability of AI in health care • There is a trade-off between performance and generalizability when addressing underspecification
Allen et al [37], 2021	United States	Radiology	Guideline	• Guideline of evaluation of AI in a radiology setting before implementation in the workflow to assist in purchase decisions and monitoring of the performance afterward		• Model evaluation can be difficult and restricted to larger, informatics-familiar institutions

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
					• Enriched site-specific data can facilitate AI evaluation, allowing that the target population is well-represented before implementation • In the AI evaluation process, capturing the metadata about equipment manufacturers, the protocol used, and demographics in the AI data registry can reveal performance decline and show whether the decline is related to specific machines or manufacturers • QA[h] allows AI to perform according to the implementation requirements	
Verma et al [38], 2021	Canada	General	Guideline	• To provide an approach for developing and implementing AI in health care	• Multidisciplinary team • Safety monitoring • Data quality • User-friendly user interface • Nondisruptive to the current workflow • End-user trust • Continuous evaluation of performance	• Not reported
Wiggins et al [39], 2021	United States	Radiology	Case study	• To develop an AI solution that can generate, consume, and provide outcomes within the clinical radiology process	• Collaboration among developers, radiologists, and AI vendors • Interoperability standards and robust methodologies, such as HL7, FHIR, and SOLE[i] • The use of metadata such as hardware or software specifications • Radiologists should be able to provide feedback on AI results • Raising awareness and providing the required training regarding the potential of AI technologies among clinicians and patients can help increase AI adoption	• Not reported
Wang et al [40], 2021	China	Radiology	Case study			

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
				• To create an AI system that analyzes CT[j] scans automatically to promptly detect COVID-19 pneumonia in hospitals	• Co-designing with clinicians • The AI model was externally validated to assess the generalizability before deployment • Preconfigured model development allowed for very quick deployment • Continuously collected data can lead to better generalizability of AI products and are considered a crucial aspect of epidemic response	• Lack of reception of continuous data for retraining the model may result in data drift and underfitting
Strohm et al [41], 2020	The Netherlands	Radiology	Case study	• To explore barriers to and enablers of AI implementation in radiology	• Collaboration among HCPs in radiology • Financial challenges in the Dutch health care system • The optimism toward AI potential • The existing strategies and initiatives in digital health • The appointment of a data champion	• Inconsistent efficacy of AI output • Lack of robust implementation procedures • Unclear added value of AI applications in routine care • Trust issue of HCPs
Soltan et al [42], 2022	United Kingdom	ED[k] triage	Case study	• To implement an AI application to screen patients with COVID-19 in an ED and perform multicenter external validation	• Conducted multicenter validation across 4 hospitals, including both temporal and geographical validations • Deployment occurred in parallel with the preexisting method, allowing for a direct comparison of performance • The AI only using laboratory tests already routinely done allowed for minimal interruption of regular clinical workflow • Temporal and geographical external validation allowed for the assessment of the generalizability of the AI tool	• Validation only performed in 1 geographical region
Sohn et al [43], 2020	United States	Radiology	Case study	• To develop an infrastructure for the implementation of ML models in routine radiology workflow		• Preexisting pipelines for clinical AI deployment often rely on third-party software, which can be problematic due to complexity, privacy, and maintenance issues

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
					• Collaboration of a multidisciplinary team • The minimum disruption to the current workflow can increase the AI uptake • An open-source pipeline facilitates the integration of additional algorithms • An ML model agnostic to the hospital systems for easier modification and retraining without impacting the existing infrastructure • The use of a QA framework by end users, clinicians, and software testers to identify model errors and submit those errors for model update • Minimum disruption to the existing radiology workflow • QA evaluation • A dedicated server for the AI applications	
Pierce et al [44], 2021	United States	Radiology	Case study	• To implement an AI-enabled mobile x-ray scanner detecting pneumothoraxes in a radiology clinical workflow	• Compatibility of the clinic's system with the vendor along with the vendor's willingness to collaborate • Granting user access privileges according to their specific roles • Staff training • The model received continuous training • Training and education of users in the use of AI can be beneficial • Minimum disruption to the current workflow	• Not reported
Kanakaraj et al [45], 2022	United States	Radiology	Case study	• To develop and demonstrate a clinical image AI validation tool with a convenient user-friendly front end while meeting important security and privacy standards		• Lack of appropriate procedure to capture users' feedback for continuous improvement of AI model

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
					• Use of secure software—PACS[l] image management, HTTPS service, and REDCap[m] database • The AI imaging incubator successfully provided an architecture for executing clinical AI models and displaying results in a clinician-friendly manner while meeting key security and privacy standards (HIPAA[n] compliance)	
Jauk et al [46], 2020	Austria	General	Case study	• To implement ML models to forecast the occurrence of delirium among patients admitted to hospitals	• Clinical staff were involved in the implementation process • Training for nurses and physicians involved is beneficial	• Performance analysis can be complicated for early-warning intervention AI systems • The incidence of delirium was lower than anticipated, impacting the calibration • Sometimes, the algorithm would underperform on patients with fewer previous hospital stays due to reduced EHR[o] data
Davis et al [47], 2019	United States	General	Case study	• To outline a procedure for selecting updating methods to combat clinical prediction model drift	• The procedure effectively recommended updating methods proportional to the need • This procedure can be applied to any type of model • The procedure is conservative compared with others	• The procedure provides no guarantee of clinically appropriate improvement to model performance
Blezek et al [48], 2021	United States	Radiology	Case study	• To outline and demonstrate a system for general AI deployment in radiology and discuss use cases and requirements		• Vended implementation platforms are also imperfect

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
					• The Agile development approach was used to deliver the AI product • Radiology IT support was significantly involved • Computational and storage resources were appropriately configured to properly handle the current and future processing requirements • Radiologists received training on the use of the new system • Custom solutions can fit and function seamlessly in clinical workflows but are susceptible to some issues • Radiologists approved of the ability to conveniently decide the correctness of the results and the system's seamless and intuitive integration into their workflow	
Pantanowitz et al [49], 2020	United States	Pathology	Case study	• To clinically validate an AI algorithm for detecting prostate adenocarcinoma, grade tumors, and detect clinically important features • To deploy the AI algorithm in clinical workflow	• Substantial increase in pathology workload and job complexity makes it a prime candidate for AI uptake • External validation • The use of unseen data sets for performance validation • Small calibration data set was effective for adapting the algorithm to a new environment • Combining target categories into clinically significant groups reduced computational requirements, allowing for real-time analysis	• Discrepancy in labeling data due to discordance among physicians for cancer grading
Fujimori et al [50], 2022	Japan	ED	Case study	• To evaluate the enablers of and barriers of implementing AI in emergency care		• Low performance in workflow • Alert fatigue • The risk of bias on a clinician's decision when using the AI application

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
					• Data explanation and visualization were used to justify the alerts • Robust validations are required to avoid undesired consequences • Alert fatigue was avoided by processing background information and presenting visual data • Training clinicians	
Joshi et al [20], 2022	United States	General	Case study	• To examine the implementation of a sepsis CDS[p] tool with ML models and rule-based approach from the viewpoint of those leading the implementation	• Ease of integration and ability to customize the AI model	• Difficulties with the definition of optimal alerts • Alerts were said to be disruptive to the workflow • Alert fatigue • Concerns about the clinical relevance of the new system • Difficult to explain and understand ML outputs • Trust issue with the output due to misunderstanding the output • High financial cost
Pou-Prom et al [51], 2022	Canada	General	Case study	• To develop an AI application for predicting the risk of clinical deterioration in hospitals	• Multidisciplinary team • Security measures were adopted • Clinical relevance to the targeted cohort • Temporal validation • Conducted a pilot test to understand the model output • User training • Model update to avoid data drift	• Lack of external validation • Lack of generalizability
Baxter et al [52], 2020	United States	General	Case study	• To identify barriers to AI uptake in workflow	• Co-design with end users	• End users' concerns about whether the new solutions are relevant to their workflow • Potential disruption to the routine workflow and unintended consequences • Lack of customization capability
Sandhu et al [53], 2020	United States	ED	Case study	• To examine the variables influencing the implementation of ML applications for predicting sepsis incidence		• Clinicians' trust • Lack of understanding of the output • Alert fatigue • Disruption to the workflow

XSL•FO
RenderX

Study, year	Country	Clinical setting	Study type	Aim of study	Enablers	Barriers
					• Co-design with nurses and clinical staff • Introduced a new job title responsible for the integration • Having the required clinical knowledge about sepsis • Training end users	
Sendak et al [54], 2020	United States	ED	Case study	• To report a deep learning sepsis detection and management system	• Multidisciplinary team • Co-design with clinical staff • Hospital leaders and external research partners • Training staff • Data scientists with the required clinical background • Personnel time for integration of new ML system • Shared infrastructure for development and deployment	• Lack of evidence-based implementation guidelines • Disruption to the workflow • Lack of feedback loop for continuous updating

[a]AI: artificial intelligence.
[b]CST: collaborative science team.
[c]HCP: health care provider.
[d]HL7: Health Level 7.
[e]ML: machine learning.
[f]FHIR: Fast Healthcare Interoperability Resources.
[g]OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model.
[h]QA: quality assurance.
[i]SOLE: Standardized Operational Log of Events.
[j]CT: computerized tomography.
[k]ED: emergency department.
[l]PACS: picture archiving and communication system.
[m]REDCap: Research Electronic Data Capture.
[n]HIPAA: Health Insurance Portability and Accountability Act.
[o]EHR: electronic health record.
[p]CDS: clinical decision support.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for study selection.

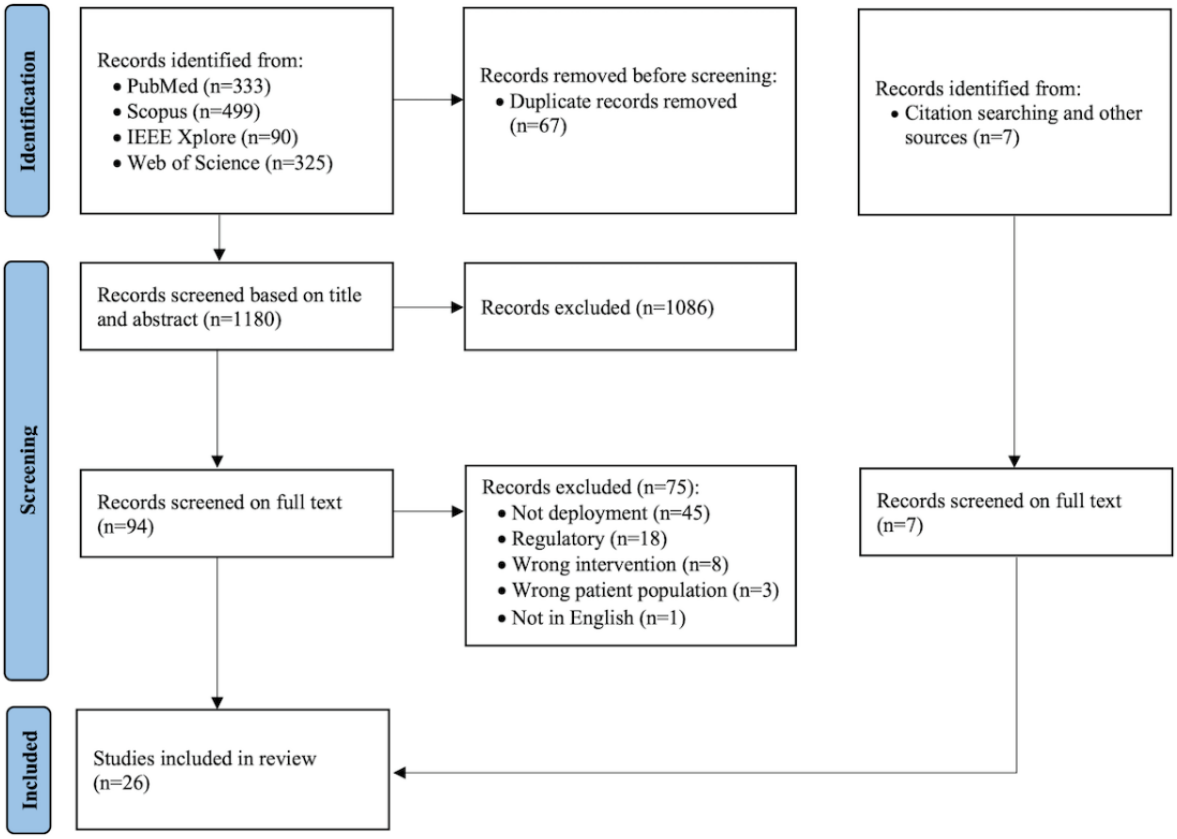

Study Characteristics

Table 2 outlines the characteristics of the included studies in this review. The publication dates of the included studies ranged from 2019 to 2022 [20,30-54]. In total, 65% (17/26) of the studies were case studies on the implementation of AI in hospitals [20,39-54], whereas the remaining 35% (9/26) were implementation guidelines [30-38].

Of the 26 identified studies, 15 (58%) originated from the United States [20,32,33,36,37,39,43-45,47-49,52-54]; 2 (8%) originated from the United Kingdom [30,42]; 2 (8%) originated from the Netherlands [34,41]; and 1 (4%) originated from China [55], Australia [46], Japan [50], Canada [51], Austria [46], Germany [35], and Sweden [31] each.

Radiology was the clinical setting in 46% (12/26) of the studies [33,34,36,37,39-41,43-45,48,49]. A total of 38% (10/26) of the studies were conducted in general inpatient wards [20,30-32,35,38,46,47,51,52], and 15% (4/26) were conducted in emergency departments [42,50,53,54].

Quality Assessment

Regarding the 35% (9/26) of guideline studies, none fully adhered to the ADAPTE framework [29]. Although these included guideline studies had clear scopes and purposes aligned with this review, they all lacked details concerning the assessment of quality, external validation, and aftercare planning procedures. The details of this assessment for all the guideline studies can be found in Multimedia Appendix 3 [20,30-54].

With respect to the 65% (17/26) of case studies, they were classified into 3 groups: quantitative descriptive (12/17, 71%) [39,40,42-49,51,54], qualitative (4/17, 24%) [20,41,52,53], and mixed methods (1/17, 6%) [50]. Overall, 5 of the case studies met the MMAT criteria: all 4 (80%) qualitative studies and the one mixed methods study. The remaining 71% (12/17) of quantitative descriptive studies failed to fully adhere to the MMAT criteria. In all but 17% (2/12) of these quantitative descriptive studies, an appropriate data sampling strategy was not used to represent their target population [40,49]. The statistical analysis of the findings was assessed as appropriate in 58% (7/12) of the quantitative descriptive studies [42,43,46,47,49,51,54]. Overall, our assessment revealed that the quality of 81% (21/26) of the included studies was poor due to insufficient reporting of their methodologies (Multimedia Appendix 3).

RQ Findings

RQ 1A Findings: Enablers of AI Implementation in Hospitals

A total of 28 enablers extracted from both prior work and this study (n=8, 29% were new enablers identified in our study) are presented in Table 3. Most of these newly identified enablers (7/8, 88%) related to the information and technology categories, highlighting the potential opportunities for hospitals regarding data readiness and required technologies for the successful implementation of AI. A total of 54% (15/28) of the enablers were shared findings between the previous reviews and this study.

XSL•FO
RenderX

Table 3. Consolidated view of research question 1A (enablers to artificial intelligence [AI] implementation; N=26)[a].

Horizon and category	Source		Studies, n (%)
	Previous studies	This study	
Horizon 2: creating AI analytics			
People			12 (46)
Enabler 1: multidisciplinary team	• Sharma et al [21]	• Blezek et al [48] • Pierce et al [44] • Pou-Prom et al [51] • Sendak et al [54] • Sohn et al [43] • Strohm et al [41] • Wang et al [40] • Wiggins et al [39] • Gruendner et al [35] • Leiner et al [34] • Verma et al [38] • Wilson et al [30]	
Enabler 2: experienced data scientists	__[b]	• Sendak et al [54] • Strohm et al [41] • Wilson et al [30]	
Process			22 (85)
Enabler 3: co-design with clinicians	• Sharma et al [21]	• Baxter et al [52] • Pierce et al [44] • Sandhu et al [53] • Sendak et al [54] • Sohn et al [43] • Strohm et al [41] • Wang et al [40] • Wiggins et al [39] • Gruendner et al [35] • Pianykh et al [33] • Svedberg et al [31] • Wilson et al [30]	
Enabler 4: robust performance monitoring and evaluation	• Sharma et al [21] • Chomutare et al [24]	• Blezek et al [48] • Fujimori et al [50] • Pou-Prom et al [51] • Sohn et al [43] • Soltan et al [42] • Allen et al [37] • Verma et al [38]	
Enabler 5: seamless integration	• Sharma et al [21] • Lee et al [22]	• Blezek et al [48] • Pierce et al [44] • Sohn et al [43] • Soltan et al [42] • Leiner et al [34] • Verma et al [38]	
Enabler 6: organizational resources	• Sharma et al [21] • Lee et al [22] • Chomutare et al [24]	• Strohm et al [41] • Wiggins et al [39] • Svedberg et al [31] • Wilson et al [30]	
Enabler 7: evidence of clinical and economic AI added value	• Sharma et al [21] • Wolff et al [23] • Chomutare et al [24]	• Joshi et al [20] • Blezek et al [48] • Strohm et al [41]	
Enabler 8: addressing data shift	• Lee et al [22]	• Wang et al [40] • Davis et al [47] • Eche et al [36]	

XSL•FO
RenderX

Horizon and category	Source		Studies, n (%)
	Previous studies	This study	
Enabler 9: improved team communication	• Sharma et al [21]	—	
Information			9 (35)
Enabler 10: data quality	• Chomutare et al [24]	• Pantanowitz et al [49] • Pou-Prom et al [51] • Wang et al [40] • Wiggins et al [39] • Allen et al [37]	
Enabler 11: data security	• Lee et al [22]	• Kanakaraj et al [45] • Pou-Prom et al [51] • Gruendner et al [35]	
Enabler 12: data visualization	—	• Fujimori et al [50] • Subbaswamy and Saria [32]	
Technology			15 (58)
Enabler 13: continuous learning capability	• Sharma et al [21]	• Pierce et al [44] • Pou-Prom et al [51] • Wang et al [40] • Pianykh et al [33] • Subbaswamy and Saria [32]	
Enabler 14: containerization	—	• Pierce et al [44] • Sohn et al [43] • Wang et al [40] • Gruendner et al [35] • Leiner et al [34]	
Enabler 15: interoperability	• Chomutare et al [24]	• Kanakaraj et al [45] • Wiggins et al [39] • Gruendner et al [35] • Leiner et al [34]	
Enabler 16: shared infrastructure	—	• Blezek et al [48] • Sendak et al [54] • Gruendner et al [35] • Leiner et al [34]	
Enabler 17: customization capability	• Sharma et al [21] • Lee et al [22] • Wolff et al [23]	• Joshi et al [20] • Blezek et al [48] • Sohn et al [43]	
Enabler 18: vendor-agnostic infrastructure	—	• Sohn et al [43] • Leiner et al [34]	
Enabler 19: computational and storage resources	—	• Blezek et al [48] • Pantanowitz et al [49]	
Enabler 20: alert considerations	—	• Fujimori et al [50]	
Enabler 21: ease of integration	—	• Joshi et al [20]	
Horizon 3: implementation of new models of care			
People			8 (31)

XSL•FO
RenderX

Horizon and category	Source		Studies, n (%)
	Previous studies	This study	
Enabler 22: skilled end users	• Sharma et al [21] • Lee et al [22] • Chomutare et al [24]	• Blezek et al [48] • Jauk et al [46] • Pierce et al [44] • Sandhu et al [53] • Sendak et al [54] • Gruendner et al [35] • Pianykh et al [33] • Wilson et al [30]	
Enabler 23: hospital leadership	• Chomutare et al [24]	• Sendak et al [54]	
Enabler 24: innovation champions	• Sharma et al [21] • Lee et al [22] • Chomutare et al [24]	—	
Process			9 (35)
Enabler 25: staff training	• Chomutare et al [24]	• Blezek et al [48] • Fujimori et al [50] • Jauk et al [46] • Pierce et al [44] • Pou-Prom et al [51] • Sandhu et al [53] • Sendak et al [54] • Leiner et al [34] • Wilson et al [30]	
Enabler 26: provide incentives when using AI	• Sharma et al [21] • Lee et al [22]	—	
Enabler 27: limiting non-AI solutions	• Wolff et al [23]	—	
Information			1 (4)
Enabler 28: usability	• Chomutare et al [24]	• Verma et al [38]	

[a]Enablers identified in previous reviews and this review were mapped to 4 categories of the 3-horizon framework [15].

[b]Not specified.

Within the scope of the 3-horizon framework [15], most included studies in this paper (22/26, 85%) indicated that the process domain facilitated the development of AI analytics within horizon 2 [20,30,31,33-44,47,48,50-54]. Co-design with clinicians was the most commonly reported enabler in 46% (12/26) of the papers in horizon 2 [30,31,33,35,39-41,43,44,52-54]. The process domain was also highlighted as having a facilitative role in the creation of new care models with AI (horizon 3) in 35% (9/26) of the papers [30,34,44,46,48,50,51,53,54]. Training end users to adopt AI solutions and interpret the insights was reported in all these 9 studies as an enabling factor in horizon 3.

Technological factors were highlighted in 58% (15/26) of the studies as enablers within horizon 2 [20,32-35,39,40,43-45,48-51,54], with the most commonly reported factor being continuous learning capability of AI analytics [32,33,40,44,51] and containerization capability by providing separated development environments [34,35,40,43,44] and applying the interoperability techniques ensuring seamless integration of diverse formats of clinical data from different hardware and software sources [34,35,39,45].

Of all the included studies, 46% (12/26) [30,34,35,38-41,43,44,48,51,54] and 31% (8/26) [30,33,35,44,46,48,53,54] identified people-related enablers across horizons 2 and 3, respectively, with multidisciplinary teams in horizon 2 and trained end users in horizon 3 being the 2 most reported enablers.

Enabling factors related to the information domain were discussed in 35% (9/26) of the included studies in this review [32,35,37,39,40,45,49-51], with data quality being the most reported enabler of the successful implementation of AI in hospitals in >50% of these papers (5/9, 56%) [37,39,40,49,51]. The enablers of the AI adoption in hospitals were reported to include factors such as considerations of data security [35,45,51] and data visualization [32,50] in horizon 2 along with AI usability [38] solutions in horizon 3.

RQ 1B Findings: Barriers to AI Implementation in Hospitals

Overall, a total of 18 barriers to AI implementation in hospitals were extracted from both prior work and this study, with 5 (28%) found to be new in this study (Table 4). Most of these newly identified barriers (4/5, 80%) were related to the

XSL•FO
RenderX

information and technology categories. A total of 50% (9/18) of the identified barriers were found to be shared findings between the previous work and this study. In our analysis, some factors played dual roles, acting as both enablers and barriers. For instance, "Seamless integration" served as an enabler (enabler 5; Table 3), whereas "Disruptive integration" acted as a barrier (barrier 3; Table 4). We reported both enablers and barriers with such reversed meanings to highlight the real-world complexities due to which such factors can exhibit this duality.

Regarding the 3-horizon framework [15], 58% (15/26) of the included studies in this review showed that the process domain hindered the development of AI within horizon 2 [20,31,37,40-43,45-47,50-54]. The lack of sufficient performance assessment within horizon 2 was the most commonly reported barrier in 27% (7/26) of the papers [37,41,42,46,47,50]. The factors related to the process domain were also reported as barriers to the implementation of AI within horizon 3, with 8% (2/26) of the papers reporting alert fatigue as an obstacle to AI adoption for creating new models of care [20,53].

Information-related factors were highlighted in 31% (8/26) of the studies as barriers within horizon 2 [20,35,36,46,51], with the most commonly mentioned one being poor data quality [20,35,36,46,51]. The challenge with data shift was reported as part of the information domain within horizon 3 [32].

Technology-related challenges in horizon 2 were identified in 19% (5/26) of the studies, including issues such as the lack of customization capability and computational limitations of hardware [35,43,48,50,52].

Within horizon 3, a total of 19% (5/26) of the included papers highlighted the barriers related to the people domain [20,30,41,50,53], with lack of trust by clinicians and inexperienced end users in using AI within their routine workflows being 2 barriers reported in these studies.

XSL•FO
RenderX

Table 4. Consolidated view of research question 1B (barriers to artificial intelligence [AI] implementation)[a].

Horizon and category	Source		Studies, n (%)
	Previous studies	This study	
Horizon 2: creating AI analytics			
Process			15 (58)
Barrier 1: insufficient performance assessment	• Chomutare et al [24]	• Fujimori et al [50] • Jauk et al [46] • Soltan et al [42] • Strohm et al [41] • Davis et al [47] • Allen et al [37]	
Barrier 2: lack of standardized guidelines for AI implementation	• Sharma et al [21] • Wolff et al [23] • Chomutare et al [24]	• Pou-Prom et al [51] • Sendak et al [54] • Soltan et al [42] • Strohm et al [41] • Svedberg et al [31]	
Barrier 3: disruptive integration	• Lee et al [22] • Chomutare et al [24]	• Joshi et al [20] • Baxter et al [52] • Sandhu et al [53] • Sendak et al [54]	
Barrier 4: inadequate continuous learning	• Chomutare et al [24]	• Kanakaraj et al [45] • Sendak et al [54] • Wang et al [40]	
Barrier 5: complexity of maintenance	• Wolff et al [23] • Chomutare et al [24]	• Sohn et al [43]	
Barrier 6: lack of clear consensus on alert definitions	—[b]	• Joshi et al [20]	
Barrier 7: insufficient data preprocessing	• Wolff et al [23]	—	
Information			8 (31)
Barrier 8: poor data quality	• Lee et al [22] • Wolff et al [23] • Chomutare et al [24]	• Joshi et al [20] • Jauk et al [46] • Pou-Prom et al [51] • Eche et al [36] • Gruendner et al [35]	
Barrier 9: data heterogeneity	—	• Pantanowitz et al [49] • Wilson et al [30]	
Barrier 10: data privacy	—	• Sohn et al [43]	
Barrier 11: challenges with data availability	• Lee et al [22] • Wolff et al [23] • Chomutare et al [24]	—	
Technology			5 (19)
Barrier 12: lack of customization capability	—	• Baxter et al [52] • Blezek et al [48] • Sohn et al [43]	
Barrier 13: computational limitations of hardware	—	• Fujimori et al [50] • Gruendner et al [35]	
Horizon 3: implementation of new models of care			
People			5 (19)

XSL•FO
RenderX

Horizon and category	Source		Studies, n (%)
	Previous studies	This study	
Barrier 14: inexperienced end users with AI output	• Chomutare et al [24]	• Joshi et al [20] • Sandhu et al [53] • Wilson et al [30]	
Barrier 15: lack of clinician trust	• Lee et al [22] • Chomutare et al [24]	• Fujimori et al [50] • Sandhu et al [53] • Strohm et al [41]	
Process			2 (8)
Barrier 16: alert fatigue	• Lee et al [22] • Chomutare et al [24]	• Joshi et al [20] • Sandhu et al [53]	
Barrier 17: difficulties with understanding AI outputs	• Chomutare et al [24]	—	
Information			1 (4)
Barrier 18: data shift	• Lee et al [22]	• Subbaswamy and Saria [32]	

[a]Barriers identified in previous reviews and this review were mapped to 4 categories of the 3-horizon framework [15].
[b]Not specified.

RQ 2 Findings: Mapping the Findings to the 3-Horizon Framework

The identified enablers and barriers to AI implementation in hospitals (RQ 1) were mapped to the 3-horizon framework [15] across 4 categories: people, process, information, and technology within horizons 2 and 3 (Figure 3 [15]).

In horizon 2, we identified a total of 21 enablers, with most associated with technology (n=9, 43%) and processes (n=7, 33%). Moving to horizon 3, a total of 7 enablers were identified, spanning the categories of people (n=3, 43%), processes (n=3, 43%), and information (n=1, 14%). Regarding barriers, horizon 2 presented a total of 13 barriers, with >50% (n=7, 54%) falling into the process category. In horizon 3, we identified a total of 5 barriers primarily distributed among the people (n=2, 40%), process (n=2, 40%), and information (n=1, 20%) categories.

Figure 3. Mapping the identified enablers and barriers to the 3-horizon framework (adapted from Sullivan et al [15] with permission from CSIRO Publishing). *Enablers described in Table 3; **Barriers described in Table 4. AIML: artificial intelligence machine learning; B: barrier; E: enabler; EMR: electronic medical record.

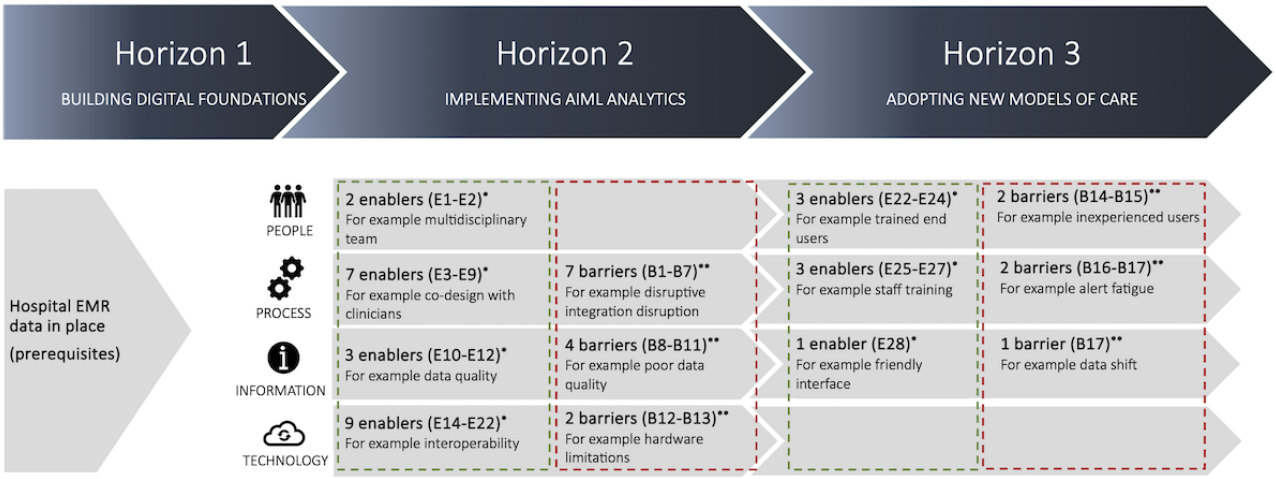

Discussion

Principal Findings

The health care industry needs to adopt new models of care to respond to the ever-growing demand for health services. Over the last decade, the academic community has shown considerable interest in the application of AI to explore new innovative models of care. Despite the numerous papers published each year exploring the potential of AI in various health domains, only a few studies have been implemented into routine workflows. Investigating the factors that lead to the success or failure of AI in health care could potentially provide

XSL•FO
RenderX

actionable insights for the effective implementation of AI in clinical workflows. In this review, we explored the current state of the literature focusing on the implementation of AI in hospitals. Our review of 26 studies revealed several enablers of and barriers to the implementation of AI in digital hospitals. Although our search for studies dated back to 2010, all 26 case studies and guidelines found in our study were published from 2019 onward. This is not surprising considering the significant progress made in AI implementation across many fields in recent years. Given such substantial advancements, implementation science needs to be further developed to accommodate these new AI innovations in health care [19]. This paper can serve as a road map for decision makers, presenting key actionable items to translate AI into hospital settings and leveraging it for potential new models of care.

While this paper extends the findings of previous reviews by examining the factors associated with AI implementation in health care [22-24], a significant aspect found in both previous reviews and our study underscores the significance of process-related factors for creating AI analytics. A large number of papers identified in this study (22/26, 85%) reported process factors as enablers of their AI implementation, aligning with the factors found in all previous reviews (enablers 3-9; Table 3). This commonality indicates the significant opportunity for hospitals to leverage their existing workflows as a strategic approach to enable AI adoption. In the context of developing innovative care models through AI analytics, obstacles associated with people (barriers 14 and 15; Table 4) were identified in 19% (5/26) of the included studies, consistent with findings in 2 previous reviews [22,24]. This highlights the influence of human factors in facilitating the integration of AI in practice.

Apart from the common findings between this and previous reviews, there are several novel aspects to this study. First, it centered specifically on hospitals, the largest and richest source of clinical data. Second, it incorporated AI implementation guidelines from the included studies, allowing for a broader understanding of AI implementation. Third, our review identified new enablers of AI implementation regarding technology and information that can facilitate AI implementation, including quality of data, shared infrastructure for continuous development, and capabilities regarding hardware resources. Fourth, this paper identified new barriers to AI implementation, with most of them being within the domains of process, information, and technology. These barriers included challenges such as data privacy, dealing with heterogeneous data, limitations with the customization of AI analytics, and ambiguity surrounding the design of alert definitions. Finally, the study findings were mapped to a 3-horizon framework encompassing 4 key categories: people, information, process, and technology. This framework offers a clear and practical road map for health care organizations planning to create new AI analytics.

It is important to note that, while our primary focus was on hospital facilities, the findings of this review may exhibit variations across other health care settings. For example, the incorporation of AI in outpatient care may demand different technological infrastructures to enable AI development. Future research can expand upon this study by investigating the evidence of enablers and barriers associated with AI implementation in wider health care settings, including primary care and outpatient care, as we expect that the outcomes of this study may differ in other health care settings. Moreover, the incorporation of studies related to regulatory aspects can be a crucial component for a more comprehensive understanding of the trajectory of AI adoption within health care systems.

Toward AI Implementation in Hospitals

Actionable Recommendations

In this section, we consolidate the findings of this study and prior work within the scope of a 3-horizon framework [15] and provide recommendations for health care organizations that plan to implement AI analytics in hospitals (Textbox 2). These recommendations are not the ultimate solution but rather a flexible action plan to facilitate AI implementation and mitigate potential challenges regarding the digital transformation of hospitals.

XSL•FO
RenderX

Textbox 2. Recommendations for artificial intelligence (AI) adoption in hospitals.

Horizon 1: establishing digital infrastructure

- Implement functional electronic medical record system
- Focus on improving data quality
- Maintain data privacy and security
- Facilitate data availability

Horizon 2: create AI analytics

- Co-design with multidisciplinary team
- Employ experienced data scientists
- Adopt interoperability methods
- Focus on AI usability
- Continuously develop and evaluate AI results
- Enhance data security and privacy
- Improve computational capabilities
- Focus on seamless integration
- Enhance customization capability
- Demonstrate AI added value
- Improve team communication
- Define design standards for AI output
- Focus on vendor-agnostic architecture

Horizon 3: create new models of care

- Restructure the clinical care models using insights from AI analytics
- Provide user training
- Continuously improve quality to produce reliable AI output and minimize data shift and alert fatigue
- Leverage hospital leaders to drive AI adoption
- Appoint innovation managers
- Provide incentive for using AI

Horizon 1: Establishing Digital Infrastructure

Data form the core of AI development to create clinical analytics. Some information barriers emerging in horizon 2, presented in Table 4, may be associated with challenges regarding EMR data, for example, quality of data (barrier 8), data heterogeneity (barrier 9), and data privacy (barrier 10). In the integration of EMR systems within hospital settings, careful attention must be paid to the functionality of the system to enable routine data collection to support the continuous development of AI analytics. Prioritizing the enhancement of data quality through the implementation of rigorous validation processes is a key factor in producing generalizable, reliable, and effective AI outputs. It is also imperative to ensure strict adherence to data privacy protocols during the EMR implementation, safeguarding sensitive patient information and maintaining ethical standards in handling health care data.

Horizon 2: Creating Analytics

Horizon 2 primarily focuses on data extraction and developing AI analytics. The successful implementation of AI in this horizon will be discussed within the following themes.

Form a Diverse Team of Experts

There is evidence suggesting that building a multidisciplinary team consisting of clinicians, nurses, end users, and data scientists can facilitate the successful design and implementation of AI in hospitals (enabler 1; Table 3). Experienced data scientists can potentially increase the success of AI in health care by ensuring accurate, reliable, and fair AI output in addition to identifying biases, handling complex medical data effectively, and optimizing AI algorithms (enabler 2; Table 3).

Enhance the Existing Processes

While horizon 2 revolves around technical aspects of AI implementation, the evidence indicates that involving clinicians, end users, and technical staff in the design and implementation stages is needed for successful integration (enabler 3; Table 3).

XSL•FO
RenderX

The co-design strategy can alleviate challenges such as the lack of consensus on alert definitions (barrier 6; Table 4), leading to usability improvement (enabler 28; Table 3). Enhancing the understanding of AI output through training end users has the potential to alleviate concerns about the usability of AI output, fostering a smoother adoption of AI technologies in hospitals.

The studies recognized that minimizing workflow disruption is key for the successful implementation of AI (enabler 5; Table 3). To minimize workflow disruption and ensure a smooth transition when implementing new AI solutions in hospitals, it is important to engage end users from the early stage of the development process [56], although training and education should be provided to help staff members effectively incorporate the AI solution into their daily routines. For successful implementation of less disruptive technologies such as AI, it is recommended to establish a clear vision and communication by the leadership team (enabler 23; Table 3), have innovation champions (enabler 24; Table 3), and provide incentives (enabler 26; Table 3) to drive long-term adoption and habit formation [57].

Continuous AI development with the use of routinely collected data and clinicians' feedback ensures that AI results accurately reflect the current clinical situations in hospital settings (enabler 13; Table 3). This can support clinicians in making more accurate diagnoses and treatment decisions by leveraging the latest insights derived from AI analytics. While insufficient assessment of AI performance in hospital settings is considered a prominent obstacle to successful implementation (barrier 1; Table 4), continuous development and monitoring helps avoid "data drift," a phenomenon in which AI models lose accuracy over time due to changes in the data or environment [32,47].

Strive for Better Data Quality and Security

The studies indicated that the implementation of AI is hindered by data privacy concerns (barrier 10; Table 4). Hospitals can mitigate the risks associated with data handling and storage by adopting standardized data frameworks and interoperability techniques (enabler 15; Table 3). These measures help minimize vulnerabilities and enhance overall data security.

The quality of data in developing AI analytics refers to the accuracy, completeness, consistency, reliability, and relevance of the data used to implement AI analytics and is considered a crucial enabler for successful AI implementation in hospitals [58]. Hospitals are encouraged to improve their data quality by implementing robust data governance protocols [21,23,31,41,42], adopting standardized data protocols to facilitate interoperability [24,34,35,39,45], and actively validating and verifying the accuracy of the data with clinicians and data scientists [30,41,54].

Strengthen Technological Infrastructures

The use of third-party hardware and software in AI solutions can limit control and raise security and privacy concerns [43]. Open-source software can improve transparency and accountability by allowing experts to identify vulnerabilities, but it can also make it easier for malicious actors to exploit them [35]. To mitigate this risk, hospitals can adopt validated open-source software with appropriate security and privacy measures, such as standardized databases and interoperability protocols [24,34,35,39,45].

Horizon 3: New Models of Care

The objective of horizon 3 is to restructure the clinical care model by harnessing the insights generated from AI analytics. While the main focus of this horizon is on clinicians and processes, fewer practical experiences are available for health organizations to help in shaping the implementation strategy.

Training end users to understand AI output is suggested to enhance the adoption of AI in hospitals (enabler 25; Table 3). Hospital leadership plays a pivotal role in facilitating the adoption of AI by providing strategic guidance, allocating necessary resources, and fostering a supportive environment for the implementation of AI initiatives (enabler 23; Table 3). Hospitals are suggested to appoint innovation managers to actively promote and facilitate the applications of AI, fostering uptake and driving the implementation process in health care (enabler 24; Table 3). Resourcing is the crucial enabler of AI integration, in particular adequate skill sets. Experienced clinicians who can interpret AI results are essential for ensuring that AI systems are used effectively and responsibly in health care organizations (enabler 22; Table 3). As a result, this can redefine the traditional models of care by advocating for evidence-based practices, patient-centered care, collaborative care, and continuous quality improvement to enhance patient outcomes and the overall quality of the care provided by health care organizations.

Limitations

Our search strategy identified 26 studies that met the inclusion criteria. All 26 studies were conducted in high-income countries. As a result, the diversity and applicability of the findings to other health care systems were constrained.

By excluding regulatory frameworks from this review in the rapidly evolving regulatory landscape, we may have limited the important implementation guidelines that ensure patient safety and ethical use of AI provided by health care regulatory bodies.

We conducted a thorough examination of the reference lists in the included studies to ensure the inclusion of all relevant papers. Despite a valid research methodology, this approach may introduce publication bias, a factor to consider when appraising the study's findings.

The methodological reporting of most studies included in this review was assessed as poor, potentially limiting the quality of the findings of this study. While consensus discussions were held after the quality assessment to mitigate potential discrepancies in the final evaluations, it is worth recognizing that this process is subjective and the perspectives of reviewers may evolve over time, resulting in variations when assessed by different individuals.

Although our intention was to identify successful implementations, it is possible that we missed significant enablers or barriers present in failed implementations.

XSL•FO
RenderX

Conclusions

This review incorporated the identified enablers of and barriers to the implementation of AI into a 3-horizon framework to guide future implementations of hospital AI analytics to evolve practice toward an LHS. Successful AI implementation in hospitals requires a shift in conventional resource management to support a new AI implementation and maintenance strategy. Using analytics to enable better decisions in hospitals is critical to enable the ever-increasing need for health care to be met.

Acknowledgments

The authors would like to acknowledge the valuable assistance of Mr Lars Eriksson, a research librarian at the Faculty of Medicine, University of Queensland, Australia, for his expertise and guidance in identifying the search strategy for this study. The authors did not use any generative artificial intelligence tools for this research paper. This study was funded by Digital Health Cooperative Research Centre Limited (DHCRC). DHCRC is funded under the Commonwealth Government's Cooperative Research Centres program. AKR and OJC are supported by DHCRC (DHCRC-0083). The funder played no role in the study design, data collection, analysis and interpretation of the data, or writing of this manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors' Contributions

AKR, CS, SS, JDP, and OJC conceptualized this paper. This research was supervised by our senior researchers CS, SS, JDP, MG, and OJC. The development of the search strategy was conducted by AKR, CS, OJC, and JDP. In total, 2 authors (AKR and OP) conducted the screening process and extracted excerpts that were included in the tables of this paper. AKR, CS, SS, JDP, MG, AV, and OJC reviewed the findings. AKR drafted the manuscript with input from CS, SS, JDP, AV, MG, and OJC. AKR prepared all the figures in this manuscript. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Enablers and barriers identified in previous reviews.
[DOC File , 77 KB - jmir_v26i1e49655_app1.doc]

Multimedia Appendix 2
Search strategies across 4 databases for this review.
[DOC File , 54 KB - jmir_v26i1e49655_app2.doc]

Multimedia Appendix 3
Quality assessment of the included studies.
[XLSX File (Microsoft Excel File), 24 KB - jmir_v26i1e49655_app3.xlsx]

Multimedia Appendix 4
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.
[DOCX File , 33 KB - jmir_v26i1e49655_app4.docx]

References

1. Williams F, Boren SA. The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. Inform Prim Care 2008;16(2):139-145 [FREE Full text] [doi: 10.14236/jhi.v16i2.685] [Medline: 18713530]
2. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 01;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
3. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020 Jun 01;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
4. Barlow J, Bayer S, Curry R. Implementing complex innovations in fluid multi-stakeholder environments: experiences of 'telecare'. Technovation 2006 Mar;26(3):396-406. [doi: 10.1016/j.technovation.2005.06.010]

XSL•FO
RenderX

5. Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf 2015 Oct;24(10):608-610. [doi: 10.1136/bmjqs-2015-004160] [Medline: 26038586]
6. Kamel Rahimi A, Canfell OJ, Chan W, Sly B, Pole JD, Sullivan C, et al. Machine learning models for diabetes management in acute care using electronic medical records: a systematic review. Int J Med Inform 2022 Apr 02;162:104758. [doi: 10.1016/j.ijmedinf.2022.104758] [Medline: 35398812]
7. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
8. Maddox TM, Rumsfeld JS, Payne PR. Questions for artificial intelligence in health care. JAMA 2019 Jan 01;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
9. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
10. Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
11. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]
12. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10:53 [FREE Full text] [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
13. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
14. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. J Med Internet Res 2020 Jul 29;22(7):e18228 [FREE Full text] [doi: 10.2196/18228] [Medline: 32723713]
15. Sullivan C, Staib A, McNeil K, Rosengren D, Johnson I. Queensland Digital Health Clinical Charter: a clinical consensus statement on priorities for digital health in hospitals. Aust Health Rev 2020 Sep;44(5):661-665. [doi: 10.1071/AH19067] [Medline: 31744594]
16. Etheredge LM. A rapid-learning health system. Health Aff (Millwood) 2007;26(2):w107-w118. [doi: 10.1377/hlthaff.26.2.w107] [Medline: 17259191]
17. Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, et al. Scalable collaborative infrastructure for a learning healthcare system (SCILHS): architecture. J Am Med Inform Assoc 2014;21(4):615-620 [FREE Full text] [doi: 10.1136/amiajnl-2014-002727] [Medline: 24821734]
18. Lim HC, Austin JA, van der Vegt AH, Rahimi AK, Canfell OJ, Mifsud J, et al. Toward a learning health care system: a systematic review and evidence-based conceptual framework for implementation of clinical analytics in a digital hospital. Appl Clin Inform 2022 Mar;13(2):339-354 [FREE Full text] [doi: 10.1055/s-0042-1743243] [Medline: 35388447]
19. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
20. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Apr 18;5(2):ooac022 [FREE Full text] [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
21. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022 Oct 05;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
22. Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics (MDPI) 2020 Sep;7(3):25 [FREE Full text] [doi: 10.3390/informatics7030025] [Medline: 33274178]
23. Wolff J, Pauling J, Keck A, Baumbach J. Success factors of artificial intelligence implementation in healthcare. Front Digit Health 2021 Jun 16;3:594971 [FREE Full text] [doi: 10.3389/fdgth.2021.594971] [Medline: 34713083]
24. Chomutare T, Tejedor M, Svenning TO, Marco-Ruiz L, Tayefi M, Lind K, et al. Artificial intelligence implementation in healthcare: a theory-based scoping review of barriers and facilitators. Int J Environ Res Public Health 2022 Dec 06;19(23):16359 [FREE Full text] [doi: 10.3390/ijerph192316359] [Medline: 36498432]
25. Platt JE, Raj M, Wienroth M. An analysis of the learning health system in its first decade in practice: scoping review. J Med Internet Res 2020 Mar 19;22(3):e17026 [FREE Full text] [doi: 10.2196/17026] [Medline: 32191214]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372(1):n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2022-04-13]
28. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract 2018 Jun;24(3):459-467. [doi: 10.1111/jep.12884] [Medline: 29464873]

XSL•FO
RenderX

29. Fervers B, Burgers JS, Voellinger R, Brouwers M, Browman GP, Graham ID, et al. Guideline adaptation: an approach to enhance efficiency in guideline development and improve utilisation. BMJ Qual Saf 2011 Mar;20(3):228-236. [doi: 10.1136/bmjqs.2010.043257] [Medline: 21209134]
30. Wilson A, Saeed H, Pringle C, Eleftheriou I, Bromiley PA, Brass A. Artificial intelligence projects in healthcare: 10 practical tips for success in a clinical environment. BMJ Health Care Inform 2021 Jul;28(1):e100323 [FREE Full text] [doi: 10.1136/bmjhci-2021-100323] [Medline: 34326160]
31. Svedberg P, Reed J, Nilsen P, Barlow J, Macrae C, Nygren J. Toward successful implementation of artificial intelligence in health care practice: protocol for a research program. JMIR Res Protoc 2022 Mar 09;11(3):e34920 [FREE Full text] [doi: 10.2196/34920] [Medline: 35262500]
32. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020 Apr 01;21(2):345-352. [doi: 10.1093/biostatistics/kxz041] [Medline: 31742354]
33. Pianykh OS, Langs G, Dewey M, Enzmann DR, Herold CJ, Schoenberg SO, et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology 2020 Oct;297(1):6-14. [doi: 10.1148/radiol.2020200038] [Medline: 32840473]
34. Leiner T, Bennink E, Mol CP, Kuijf HJ, Veldhuis WB. Bringing AI to the clinic: blueprint for a vendor-neutral AI deployment infrastructure. Insights Imaging 2021 Feb 02;12(1):11 [FREE Full text] [doi: 10.1186/s13244-020-00931-1] [Medline: 33528677]
35. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. KETOS: clinical decision support and machine learning as a service - a training and deployment platform based on Docker, OMOP-CDM, and FHIR Web Services. PLoS One 2019 Oct 03;14(10):e0223010 [FREE Full text] [doi: 10.1371/journal.pone.0223010] [Medline: 31581246]
36. Eche T, Schwartz LH, Mokrane FZ, Dercle L. Toward generalizability in the deployment of artificial intelligence in radiology: role of computation stress testing to overcome underspecification. Radiol Artif Intell 2021 Oct 27;3(6):e210097 [FREE Full text] [doi: 10.1148/ryai.2021210097] [Medline: 34870222]
37. Allen B, Dreyer K, Stibolt RJ, Agarwal S, Coombs L, Treml C, et al. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: try it, buy it, check it. J Am Coll Radiol 2021 Nov;18(11):1489-1496. [doi: 10.1016/j.jacr.2021.08.022] [Medline: 34599876]
38. Verma AA, Murray J, Greiner R, Cohen JP, Shojania KG, Ghassemi M, et al. Implementing machine learning in medicine. CMAJ 2021 Aug 30;193(34):E1351-E1357 [FREE Full text] [doi: 10.1503/cmaj.202434] [Medline: 35213323]
39. Wiggins WF, Magudia K, Schmidt TM, O'Connor SD, Carr CD, Kohli MD, et al. Imaging AI in practice: a demonstration of future workflow using integration standards. Radiol Artif Intell 2021 Oct 27;3(6):e210152 [FREE Full text] [doi: 10.1148/ryai.2021210152] [Medline: 34870224]
40. Wang B, Jin S, Yan Q, Xu H, Luo C, Wei L, et al. AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system. Appl Soft Comput 2021 Jan;98:106897 [FREE Full text] [doi: 10.1016/j.asoc.2020.106897] [Medline: 33199977]
41. Strohm L, Hehakaya C, Ranschaert ER, Boon WP, Moors EH. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct;30(10):5525-5532 [FREE Full text] [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
42. Soltan AA, Yang J, Pattanshetty R, Novak A, Yang Y, Rohanian O, et al. Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. Lancet Digit Health 2022 Apr;4(4):e266-e278 [FREE Full text] [doi: 10.1016/S2589-7500(21)00272-7] [Medline: 35279399]
43. Sohn JH, Chillakuru YR, Lee S, Lee AY, Kelil T, Hess CP, et al. An open-source, vender agnostic hardware and software pipeline for integration of artificial intelligence in radiology workflow. J Digit Imaging 2020 Aug;33(4):1041-1046 [FREE Full text] [doi: 10.1007/s10278-020-00348-8] [Medline: 32468486]
44. Pierce JD, Rosipko B, Youngblood L, Gilkeson RC, Gupta A, Bittencourt LK. Seamless integration of artificial intelligence into the clinical environment: our experience with a novel pneumothorax detection artificial intelligence algorithm. J Am Coll Radiol 2021 Nov;18(11):1497-1505. [doi: 10.1016/j.jacr.2021.08.023] [Medline: 34597622]
45. Kanakaraj P, Ramadass K, Bao S, Basford M, Jones LM, Lee HH, et al. Workflow integration of research AI tools into a hospital radiology rapid prototyping environment. J Digit Imaging 2022 Aug;35(4):1023-1033 [FREE Full text] [doi: 10.1007/s10278-022-00601-2] [Medline: 35266088]
46. Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. J Am Med Inform Assoc 2020 Jul 01;27(9):1383-1392 [FREE Full text] [doi: 10.1093/jamia/ocaa113] [Medline: 32968811]
47. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc 2019 Dec 01;26(12):1448-1457 [FREE Full text] [doi: 10.1093/jamia/ocz127] [Medline: 31397478]
48. Blezek DJ, Olson-Williams L, Missert A, Korfiatis P. AI integration in the clinical workflow. J Digit Imaging 2021 Dec;34(6):1435-1446 [FREE Full text] [doi: 10.1007/s10278-021-00525-3] [Medline: 34686923]
49. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment

XSL•FO
RenderX

study. Lancet Digit Health 2020 Aug;2(8):e407-e416 [FREE Full text] [doi: 10.1016/S2589-7500(20)30159-X] [Medline: 33328045]
50. Fujimori R, Liu K, Soeno S, Naraba H, Ogura K, Hara K, et al. Acceptance, barriers, and facilitators to implementing artificial intelligence-based decision support systems in emergency departments: quantitative and qualitative evaluation. JMIR Form Res 2022 Jun 13;6(6):e36501 [FREE Full text] [doi: 10.2196/36501] [Medline: 35699995]
51. Pou-Prom C, Murray J, Kuzulugil S, Mamdani M, Verma AA. From compute to care: lessons learned from deploying an early warning system into clinical practice. Front Digit Health 2022;4:932123 [FREE Full text] [doi: 10.3389/fdgth.2022.932123] [Medline: 36133802]
52. Baxter SL, Bass JS, Sitapati AM. Barriers to implementing an artificial intelligence model for unplanned readmissions. ACI Open 2020 Jul;4(2):e108-e113 [FREE Full text] [doi: 10.1055/s-0040-1716748] [Medline: 33274314]
53. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421 [FREE Full text] [doi: 10.2196/22421] [Medline: 33211015]
54. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform 2020 Jul 15;8(7):e15182 [FREE Full text] [doi: 10.2196/15182] [Medline: 32673244]
55. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. Yearb Med Inform 2019 Aug;28(1):16-26 [FREE Full text] [doi: 10.1055/s-0039-1677908] [Medline: 31419814]
56. Li JP, Liu H, Ting DS, Jeon S, Chan RV, Kim JE, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res 2021 May;82:100900 [FREE Full text] [doi: 10.1016/j.preteyeres.2020.100900] [Medline: 32898686]
57. Sibbald M, Zwaan L, Yilmaz Y, Lal S. Incorporating artificial intelligence in medical diagnosis: a case for an invisible and (un)disruptive approach. J Eval Clin Pract 2024 Feb;30(1):3-8. [doi: 10.1111/jep.13730] [Medline: 35761764]
58. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: a report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. Health Inf Manag 2021;50(1-2):88-92. [doi: 10.1177/1833358319887743] [Medline: 31805788]

Abbreviations

AI: artificial intelligence
EMR: electronic medical record
LHS: learning health system
MeSH: Medical Subject Heading
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

Edited by T de Azevedo Cardoso, S Ma; submitted 11.08.23; peer-reviewed by M Nair, L Weik; comments to author 13.10.23; revised version received 08.02.24; accepted 22.05.24; published 02.08.24.

Please cite as:
Kamel Rahimi A, Pienaar O, Ghadimi M, Canfell OJ, Pole JD, Shrapnel S, van der Vegt AH, Sullivan C
Implementing AI in Hospitals to Achieve a Learning Health System: Systematic Review of Current Enablers and Barriers
J Med Internet Res 2024;26:e49655
URL: https://www.jmir.org/2024/1/e49655
doi:10.2196/49655
PMID:

©Amir Kamel Rahimi, Oliver Pienaar, Moji Ghadimi, Oliver J Canfell, Jason D Pole, Sally Shrapnel, Anton H van der Vegt, Clair Sullivan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Predictors of Health Care Practitioners' Intention to Use AI-Enabled Clinical Decision Support Systems: Meta-Analysis Based on the Unified Theory of Acceptance and Use of Technology

Julius Dingel[1*], BSc, MSc; Anne-Kathrin Kleine[1*], BSc, MSc, PhD; Julia Cecil[1], BSc, MSc; Anna Leonie Sigl[2]; Eva Lermer[1,2], BSc, MSc, PhD; Susanne Gaube[3], BSc, MSc, PhD

[1]Human-AI-Interaction Group, Center for Leadership and People Management, Ludwig Maximilian University of Munich, Munich, Germany
[2]Department of Liberal Arts and Sciences, Technical University of Applied Sciences Augsburg, Augsburg, Germany
[3]Human Factors in Healthcare, Global Business School for Health, University College London, London, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Anne-Kathrin Kleine, BSc, MSc, PhD
Human-AI-Interaction Group
Center for Leadership and People Management
Ludwig Maximilian University of Munich
Geschwister-Scholl-Platz 1
Munich, 80539
Germany
Phone: 49 8921809775
Email: anne-kathrin.kleine@psy.lmu.de

Abstract

Background: Artificial intelligence–enabled clinical decision support systems (AI-CDSSs) offer potential for improving health care outcomes, but their adoption among health care practitioners remains limited.

Objective: This meta-analysis identified predictors influencing health care practitioners' intention to use AI-CDSSs based on the Unified Theory of Acceptance and Use of Technology (UTAUT). Additional predictors were examined based on existing empirical evidence.

Methods: The literature search using electronic databases, forward searches, conference programs, and personal correspondence yielded 7731 results, of which 17 (0.22%) studies met the inclusion criteria. Random-effects meta-analysis, relative weight analyses, and meta-analytic moderation and mediation analyses were used to examine the relationships between relevant predictor variables and the intention to use AI-CDSSs.

Results: The meta-analysis results supported the application of the UTAUT to the context of the intention to use AI-CDSSs. The results showed that performance expectancy (r=0.66), effort expectancy (r=0.55), social influence (r=0.66), and facilitating conditions (r=0.66) were positively associated with the intention to use AI-CDSSs, in line with the predictions of the UTAUT. The meta-analysis further identified positive attitude (r=0.63), trust (r=0.73), anxiety (r=–0.41), perceived risk (r=–0.21), and innovativeness (r=0.54) as additional relevant predictors. Trust emerged as the most influential predictor overall. The results of the moderation analyses show that the relationship between social influence and use intention becomes weaker with increasing age. In addition, the relationship between effort expectancy and use intention was stronger for diagnostic AI-CDSSs than for devices that combined diagnostic and treatment recommendations. Finally, the relationship between facilitating conditions and use intention was mediated through performance and effort expectancy.

Conclusions: This meta-analysis contributes to the understanding of the predictors of intention to use AI-CDSSs based on an extended UTAUT model. More research is needed to substantiate the identified relationships and explain the observed variations in effect sizes by identifying relevant moderating factors. The research findings bear important implications for the design and implementation of training programs for health care practitioners to ease the adoption of AI-CDSSs into their practice.

(J Med Internet Res 2024;26:e57224) doi:10.2196/57224

XSL•FO
RenderX

KEYWORDS

Unified Theory of Acceptance and Use of Technology; UTAUT; artificial intelligence–enabled clinical decision support systems; AI-CDSSs; meta-analysis; health care practitioners

Introduction

Background

The past decade has witnessed major advancements in the field of health care, particularly through the integration of artificial intelligence (AI). AI may be described as machines that mimic cognitive functions associated with the human mind, such as learning and problem-solving [1]. An area of progress involves the development of AI-enabled clinical decision support systems (AI-CDSSs) [2-4]. AI-CDSSs use machine learning algorithms to process vast amounts of data and provide case-specific advice to health care practitioners to aid clinical decision-making [5-7]. AI-CDSSs use clinical data both from structured (eg, laboratory test results) and unstructured (eg, clinician notes or imaging) sources. The interpretation of text-based data can be performed using natural language processing to transform text into usable data for clinical predictions [8]. In addition, deep learning models, including neural networks, can be used to generate recommendations based on image data, for example, in the detection of pneumonia from chest radiographs [9]. AI-CDSSs may improve the accuracy and efficiency of medical decision-making in several ways.

First, AI-CDSSs may offer structured rationales underpinning clinical decisions that can complement traditional care methods. This structured approach paves the way for clearer understanding, improved communication, and better tracking of the decision-making process in clinical settings [10,11]. Second, AI-CDSSs can integrate data from various sources to provide a comprehensive and personalized recommendation for every patient case [7]. Finally, AI-CDSSs promote the consistency of medical decisions. The use of AI algorithms may ensure that the same set of facts will consistently produce the same recommendations, thus minimizing harmful consequences due to human error [9].

Despite these advantages, the implementation of AI-CDSSs in clinical practice must still overcome numerous barriers. A major challenge in the deployment of AI-CDSSs is the variability in performance. This can occur when the data used to develop the AI models do not adequately represent the population for which the tool is intended. Another issue is when AI-CDSSs are not used as designed, which can be due to a range of factors, including user interface problems, lack of integration into clinical workflows, or insufficient training of health care professionals on how to use the system [7,12-14]. The resulting low performance casts doubt on the value of AI-CDSSs in assisting with clinical decision-making [12,15]. In addition, the lack of understanding of how AI recommendations are derived heightens clinicians' reservations about using these systems [16-18]. There are also challenges related to the alignment of AI-CDSSs with existing workflows that can cause additional workload when new AI systems are incorporated into clinical procedures [7,19-21].

As the development of high-performing AI-CDSSs proceeds, understanding the factors that influence health care practitioners' intention to use these systems becomes increasingly relevant. One of the most comprehensive theories to explain individual technology adoption is the Unified Theory of Acceptance and Use of Technology (UTAUT) [22]. The UTAUT proposes that a person's intention to use a technology is determined by their beliefs and attitudes toward that technology, such as the perception of its performance or the perceived effort it would require to use it. The UTAUT's comprehensive nature and its ability to account for various determinants of technology acceptance make it an appropriate model for examining the predictors of health care practitioners' intention to use AI-CDSSs.

Research to identify predictors of the intention to use AI-CDSSs has accumulated over the past years [4,23-25]. However, the existing literature remains scattered and in need of systematic synthesis. Therefore, the overarching goal of this study was to quantitatively integrate existing studies on the predictors of health care practitioners' intention to use AI-CDSSs. The proposed hypotheses were based on the UTAUT model and existing empirical evidence. With this meta-analysis, we make 4 major contributions to theory and practice. First, we used meta-analytic techniques to estimate the relationship between the predictors of the UTAUT and the intention to use AI-CDSSs, thus providing insights into the applicability of the UTAUT to the context of AI-CDSSs. Second, we identified additional predictors based on the existing literature and examined the relative contribution of the UTAUT and additional predictors in explaining the intention to use AI-CDSSs. With this approach, we contribute to a theoretical refinement and potential extension of the UTAUT model to the context of AI-CDSSs. Third, based on the UTAUT, we examined the role of contextual factors as moderators of the relationships between relevant predictors and use intention, thus shedding light on the conditions that influence the strength of these relationships. Finally, in line with the UTAUT model, this is the first meta-analysis that examines the role of mediators, thus allowing for a better understanding of the complex mechanisms through which use intention may be explained. The study protocol, including all hypotheses and research questions (RQs), has been preregistered through the Open Science Framework [26].

Theory and Hypothesis Development

The UTAUT and the Intention to Use AI-CDSSs

The UTAUT integrates 8 former technology use theories and has become one of the most prominent technology use models [22,27]. The UTAUT has been applied to investigate factors influencing the acceptance and use of technology in different contexts, including health care [28-30]. The primary outcome measure considered in the UTAUT, alongside actual use, is the intention to use a technology [22,31,32]. Intentions are indicators of motivation and reflect the level of determination that individuals have to actually perform a certain behavior [33].

XSL•FO
RenderX

The successful deployment of any technology depends largely on the user's intention to use it [34]. Accordingly, understanding the predictors of the intention to use AI-CDSSs may help overcome individual-level impediments thwarting the adoption of AI-CDSSs in health care.

The UTAUT consists of 4 core predictors of individual use intention: performance expectancy, effort expectancy, social influence, and facilitating conditions [22]. The relationships between these variables and use intention are proposed to be moderated by gender, age, experience, and voluntariness of use [22]. The UTAUT model is shown in Figure 1. All relationships included in the UTAUT were proposed as hypotheses, whereas all additional relationships and moderators that were derived based on empirical findings and other theories were proposed as RQs.

Figure 1. The Unified Theory of Acceptance and Use of Technology model.

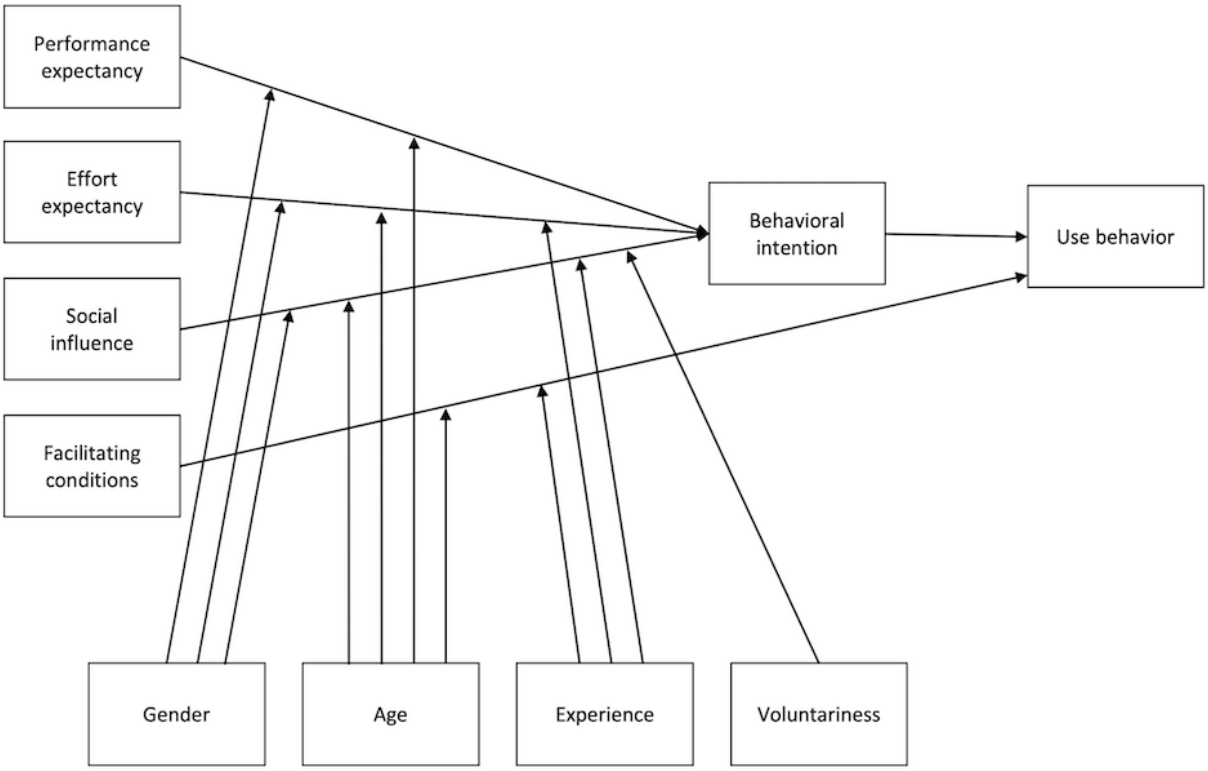

Predictors of the Intention to Use AI-CDSSs Based on the UTAUT

Performance expectancy refers to the extent to which individuals believe that using a technology will improve their job performance. AI-CDSSs have the potential to enhance job performance by aiding clinicians in deriving diagnoses or making treatment decisions [35]. If clinicians perceive their decisions to be improved by using AI-CDSSs, then performance expectancy will be high [36,37]. Hypothesis 1 is that performance expectancy is positively related to the intention to use AI-CDSSs.

Effort expectancy concerns the perceived ease of use of a technology. It is suggested that a system that is perceived to be easy to use is more likely to be accepted than one that is perceived to be complicated to use [22]. If, for example, the perceived effort of using an AI-CDSS in one's existing clinical workflows is perceived to be high, health care practitioners may be less willing to use it [3,21,23]. Hypothesis 2 is that effort expectancy is positively related to the intention to use AI-CDSSs.

Social influence refers to the impact of social factors, such as the expectations and influence of peers, on an individual's intention to use a technology. The positive relationship between social influence and the intention to use AI-CDSSs has consistently been supported in empirical studies [23,37]. For example, it has been found that medical professionals holding the belief that their colleagues, top management, and professional bodies endorse the use of AI-CDSSs in clinical settings are more willing to adopt them [37]. Hypothesis 3 is that social influence is positively related to the intention to use AI-CDSSs.

Facilitating conditions represent the organizational and technical infrastructure necessary for technology adoption [22]. It has been argued that, if users believe that the resources and support are in place to facilitate the use of AI-CDSSs, they are more likely to intend to use them [4,22,38]. Hypothesis 4.1 is that facilitating conditions are positively related to the intention to use AI-CDSSs. In addition, according to the UTAUT, there is

XSL•FO
RenderX

a direct relationship between facilitating conditions and actual technology use [22]. Facilitating conditions refer to the resources and support available to use a technology, including the access to the necessary tools and knowledge. This practical aspect makes the influence of facilitating conditions on use immediate as users are more likely to use technology when they perceive a supportive environment and available resources. Unlike other predictors in the UTAUT, facilitating conditions are proposed as direct antecedents of actual use [22]. Hypothesis 4.2 is that facilitating conditions are positively related to actual use of AI-CDSSs.

Additional Predictors of the Intention to Use AI-CDSSs

The UTAUT has been modified, and additional predictors have been added over time to account for various settings and technologies [31,39-41]. However, a meta - analytic review is limited to the relationships that have been studied in the literature. Following previous research and methodological best practices, we included additional predictors beyond the UTAUT in the meta-analysis that were examined in at least 3 independent samples [42,43]. Following this criterion, we identified attitude, trust, perceived risk, AI anxiety, and personal innovativeness as additional predictors of the intention to use AI-CDSSs.

Individual behavior is driven by intention, which is, in turn, a function of an individual's attitude toward the behavior and subjective norms [31,44]. Indeed, a positive attitude toward AI-CDSSs has been identified as a relevant predictor of the intention to use AI-CDSSs [45-47]. Because the relationship between positive attitude and use intention is not included in the UTAUT, we propose the following RQ (RQ 1) to explore the relationship between positive attitude and the intention to use AI-CDSSs [22]: is there a positive relationship between a positive attitude toward AI-CDSSs and the intention to use AI-CDSSs?

Trust becomes relevant if the outcome of a situation is uncertain or the possibility of undesirable outcomes exists [48]. Trust has been argued to be a particularly relevant predictor of the intention to use AI-CDSSs due to a lack of transparency of how recommendations are derived and the high stakes of erroneous decisions in health care [23,37]. Generally, we may differentiate between initial trust as the judgment of the truster before being exposed to the trustee and knowledge-based trust that may be established after the truster has interacted with the trustee [48]. In the context of AI-CDSSs, some studies refer to initial trust in terms of beliefs in the reliability and safety of AI-CDSSs before the user has been exposed to or actively used the system [3,37,49]. An example item for initial trust is "I believe AI could provide completely accurate diagnosis assistant service" [3]. Another aspect of trust that has been explored in empirical studies is trust in different attributes of the technology, namely, its functionality (being able to do a required task), its helpfulness or benevolence (being able to provide effective help when needed or act in the best interest of the user), and its integrity (operating reliably or consistently without failing) [48,50,51]. An example item for trust in the system's integrity regarding data security is "I trust that recommendations from the AI-powered care pathway are reliable" [51]. Because trust is not included in the UTAUT model, we propose an RQ (RQ 2) to explore whether there is a positive relationship between trust and the intention to use AI-CDSSs [22]: is there a positive relationship between trust and the intention to use AI-CDSSs?

Perceived risk is determined by the unpredictability and perceived intensity of outcomes [52]. In the context of AI-CDSSs, perceived risk refers to the perceived potential negative consequences associated with their use, including performance failure and data insecurity [4]. An example item for perceived risk of a performance failure is "There is a possibility of malfunction and performance failure, so the system might fail to deliver accurate contouring areas and could mislead my work with inaccurate contouring" [4]. Health care professionals may be reluctant to engage with new services fearing that their perceived risk may result in negative user experience or even harm to them or their patients [53]. Different forms of perceived risk have been found to be negatively associated with the intention to use AI-CDSSs [4,37,53,54]. For example, it has been found that performance and legal risk associated with AI-CDSSs are negatively related to the intention to use AI-CDSSs [37]. Because perceived risk is not included in the UTAUT model, we propose the following RQ (RQ 3) to investigate whether perceived risk is negatively associated with the intention to use AI-CDSSs [22]: is there a negative relationship between perceived risk and the intention to use AI-CDSSs?

AI anxiety encompasses general fears and insecurities regarding AI technology. It represents an intuitive, negative affective reaction to AI technologies, for example, based on the fear of making mistakes [55,56]. AI anxiety is often measured using the AI anxiety scale [22]. An example item is "I feel apprehensive about using the system." If health care professionals experience anxiety in using AI-CDSSs, their intention to use them is presumably low. Indeed, AI anxiety has been identified as a negative predictor of the intention to use AI in health care [24]. However, because AI anxiety is not included as a predictor of use intention in the UTAUT, we propose the following RQ (RQ 4) to explore whether AI anxiety is negatively associated with the intention to use AI-CDSSs [22]: is there a negative relationship between AI anxiety and the intention to use AI-CDSSs?

Personal innovativeness describes an individual's readiness to experiment with and embrace a new technology [57]. Those demonstrating a high degree of personal innovativeness have greater capabilities and, therefore, demonstrate greater readiness to use a new technology [58,59]. Indeed, there is empirical evidence for a positive link between personal innovativeness and the intention to use AI-CDSSs [3,36]. RQ 5 is as follows: is there a positive relationship between personal innovativeness and the intention to use AI-CDSSs?

The Relationship Between AI-CDSS Use Intention and Actual Use

The UTAUT proposes that an individual's intention to use a technology is the main predictor of its actual use [22]. However, this relationship has not yet been extensively researched in the context of AI-CDSSs. The limited investigation of actual use may be attributed to the restricted number of AI-CDSSs implemented in clinical practice [60]. Nonetheless, some

XSL•FO
RenderX

evidence indicates that use intention predicts the actual use of AI-CDSSs [4,47]. RQ 6 is as follows: what is the relationship between the intention to use AI-CDSSs and their actual use?

The Relative Contribution of the UTAUT Predictors and Additional Predictors in Explaining AI-CDSS Use Intention

Existing empirical research has explored the extent to which the UTAUT predictors account for variance in technology use intention [61]. For example, performance expectancy has often emerged as the strongest predictor of use intention [62-64]. Other research has found that trust has a stronger effect on the intention to use AI-CDSSs than performance expectancy [37]. As the roles of the UTAUT and additional predictors in explaining the intention to use AI-CDSSs remain unclear, we propose the following RQ (RQ 7): what is the relative contribution of the UTAUT predictors and additional predictors in explaining the intention to use AI-CDSSs?

Moderators of the Relationships Between UTAUT Predictors and the Intention to Use AI-CDSSs

The relationships between UTAUT predictors and use intention are proposed to be moderated by age, gender, user experience with the system, and voluntariness of using the system [22]. First, it has been suggested that younger workers prioritize extrinsic rewards such as improved job performance, thus exhibiting a stronger relationship between performance expectancy and technology use intention [22]. In contrast, it has been suggested that older workers generally face greater software challenges and are more likely to place increased relevance on social influences. Accordingly, they may rely more on effort expectancy and social influence when deciding to use a technology [22]. Hypothesis 5 is that the relationship between (1) performance expectancy and the intention to use AI-CDSSs becomes weaker and the relationships between (2) effort expectancy and (3) social influence and the intention to use AI-CDSSs become stronger with increasing age.

Second, the impact of performance expectancy on use intention is expected to be stronger among men, whereas the relationships between effort expectancy and social influence and use intention would be more pronounced among women [22]. Hypothesis 6 is that the relationship between (1) performance expectancy and the intention to use AI-CDSSs is stronger for men and the relationships between (2) effort expectancy and (3) social influence and the intention to use AI-CDSSs are stronger for women.

Third, according to the UTAUT, limited experience increases the strength of the relationship between effort expectancy and social influence and use intention because individuals with limited experience tend to overestimate the challenges associated with using a new technology and their opinions are more susceptible to social influence [22]. In contrast, as experience increases, facilitating conditions have been proposed to exhibit a greater impact on actual technology use as more experienced users know better in terms of how to take advantage of facilitating conditions when using the system [22]. Hypothesis 7 is that the relationships between (1) effort expectancy and (2) social influence and intention to use AI-CDSSs become weaker with increasing experience and the relationship between (3) facilitating conditions and actual use of AI-CDSSs becomes stronger with increasing experience.

Finally, the UTAUT distinguishes between voluntary (ie, individuals decide themselves whether to use a technology) and mandatory (eg, the use of a technology is mandated by the supervisor) adoption settings [22]. It has been suggested that social influence affects use intention in mandatory situations more because relevant others have the capacity to either incentivize desired actions or penalize noncompliance [22]. Hypothesis 8 is that the relationship between social influence and the intention to use AI-CDSSs is stronger in mandatory adoption settings.

In addition to the UTAUT moderators, we investigated the influence of additional contextual moderators that are studied in the literature, namely, occupation, type of AI-CDSS, and culture. All additional moderators were selected based on a comprehensive preliminary review of the literature. First, health care practitioners may work in different contexts requiring them to complete different tasks. These differences may influence their perceptions, beliefs, and attitudes toward AI-CDSSs [24,53]. For instance, one study found differences in the relationship between social influence and perceived risk and use intention between clinicians (eg, surgery and orthopedics) and nonclinicians (eg, radiologists and pathologists). Specifically, for nonclinicians, social influence positively predicted the intention to use AI-CDSSs, whereas perceived risk did not emerge as a significant predictor. In contrast, among clinicians, the reverse pattern was observed [53]. Second, the type of AI-CDSS likely influences practitioners' use intention. Specifically, health care practitioners may place greater emphasis on the effectiveness and safety of treatment AI-CDSSs compared to diagnostic AI-CDSSs as an erroneous treatment decision is associated with more severe consequences [24]. Finally, cultural differences may influence the intention to use AI-CDSSs in health care [65,66]. For example, one study found perceived ease of use to be a more relevant predictor of the intention to use IT among Taiwanese compared to American physicians [66]. Accordingly, we propose the following RQ (RQ 8): do (1) the practitioner's occupation, (2) the type of AI-CDSS, and (3) the cultural background moderate the relationship between UTAUT predictors and the intention to use AI-CDSSs?

Finally, we investigated the influence of methodological moderators such as publication year and the scale used to measure AI-CDSS use intention. In a meta-analysis based on the UTAUT, it was found that some effect sizes were stronger in more recent studies [61]. Moreover, while most studies use the intention to use scale introduced by Venkatesh et al [22], some studies use self-developed scales to measure use intention [25,36]. RQ 9 is as follows: do (1) publication year and (2) the use intention scale used moderate the relationship between UTAUT predictors and the intention to use AI-CDSSs?

XSL•FO
RenderX

Performance and Effort Expectancy as Mediators of the Relationship Between Facilitating Conditions and the Intention to Use AI-CDSSs

According to the UTAUT, the effect of facilitating conditions on use intention may be explained through performance and effort expectancy [67]. That is, if the required support infrastructure is provided, a person would perceive the system to be both high performing and easy to use, which, in turn, positively influences their intention to use it. Indeed, effort expectancy has been found to fully mediate the relationship between facilitating conditions and use intention [67]. Accordingly, we propose the following RQ (RQ 10) to investigate the mediating role of performance and effort expectancy: is the relationship between facilitating conditions and intention to use AI-CDSSs mediated through performance and effort expectancy?

Overview of the Hypotheses and RQs

Figure 2 shows all hypotheses and RQs. We omitted the relationship between facilitating conditions and actual use of AI-CDSSs (hypothesis 4.2) as well as the moderators experience (hypothesis 7), voluntariness (hypothesis 8), and occupation (RQ 8.1) from the analyses (see the dashed lines in Figure 2) due to the limited number of available independent samples (<3). All deviations from the preregistration are presented in Table S1 in Multimedia Appendix 1 [3,4,22-25,34,36-38,45,47,49,53,54,56,57,68-83].

Figure 2. The proposed research model. The dashed lines represent preregistered hypotheses and research questions (RQs) that could not be investigated due to the limited number of available independent samples (<3). RQ 7 is omitted from the figure as it refers to the relative weight analysis. AI-CDSS: artificial intelligence–enabled clinical decision support system; H: hypothesis; UTAUT: Unified Theory of Acceptance and Use of Technology.

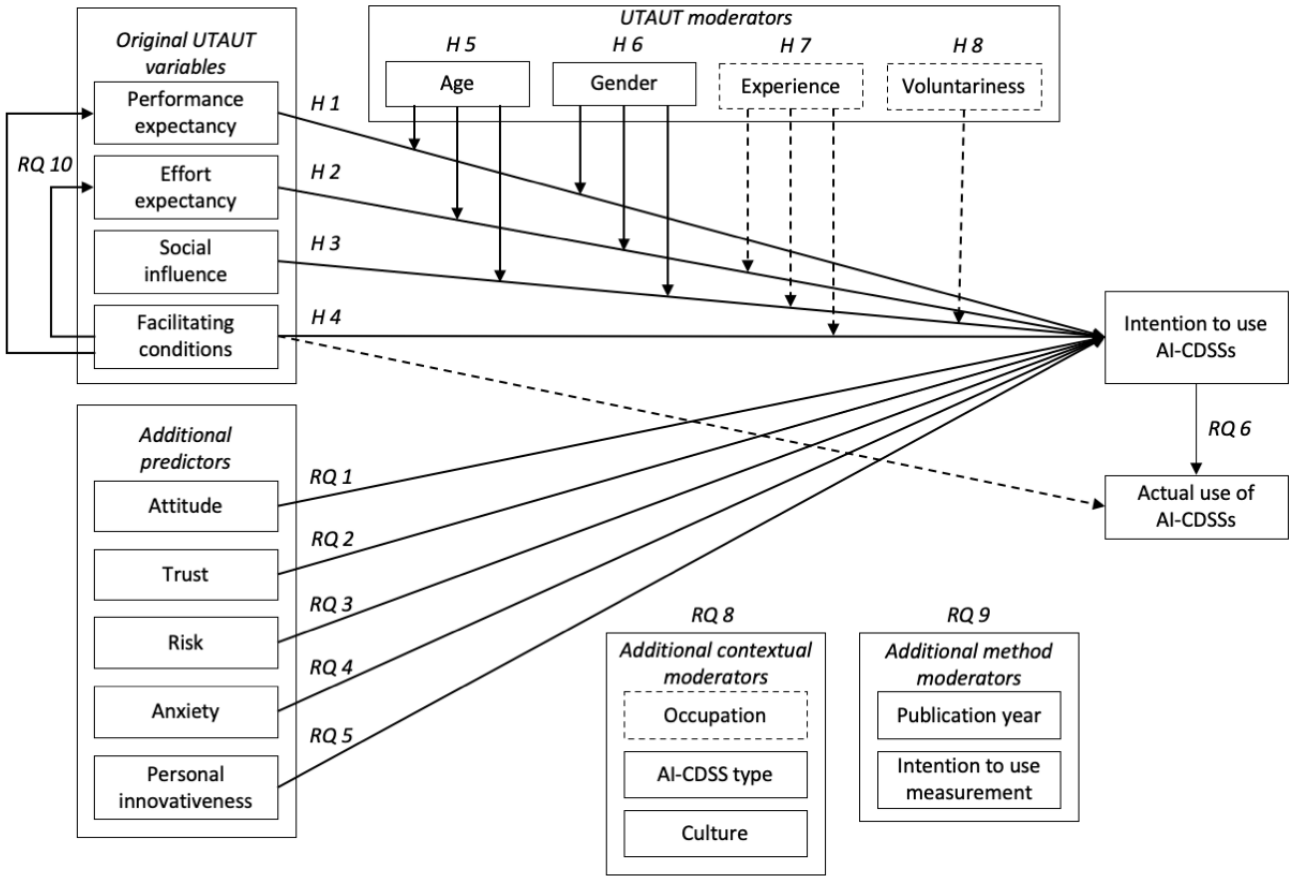

Methods

Inclusion and Exclusion Criteria

To qualify for inclusion, the following criteria had to be met. First, studies had to be published in English. Second, studies had to include AI-CDSSs. The second inclusion criterion was fulfilled if (1) one of the following terms—"artificial intelligence," "AI," "machine learning," "deep learning," or "deep neural networks"—was used to describe the technology [84] and (2) the technology was referred to as a clinical decision support system or it was described as providing recommendations regarding the diagnosis, treatment, or prognosis of health issues [6]. We included studies if AI-CDSSs were mentioned alongside other AI-enabled functionalities [85]. This led to the exclusion of studies that investigated the use intention of other health care technologies, such as telemedicine [86] or the Internet of Medical Things [87]. Notably, one study examined the intention to use explainable and nonexplainable AI-CDSSs in the same sample [38]. Because only one other study examined explainable AI [45], we included only the data for the nonexplainable AI-CDSSs. Third, studies had to include a measure of the intention to use AI-CDSSs as defined in the UTAUT [22], including self-developed scales based on the UTAUT scale. Fourth, studies had to be empirical. This led to the exclusion of nonempirical studies such as reviews or case studies [88]. Fifth, studies had to measure at least one predictor

XSL•FO
RenderX

of the intention to use AI-CDSSs. Sixth, studies had to measure use intention among a sample of health care practitioners or medical students based on the list of health professionals by the World Health Organization [89]. Table S2 in Multimedia Appendix 1 shows a detailed overview of the inclusion criteria per included study.

Search Strategy and Data Extraction

This meta-analysis was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure comprehensive and transparent reporting [90]. We used 5 steps to search for relevant data. First, relevant scientific articles, dissertations, and theses were searched using the electronic databases Embase, MEDLINE, ProQuest, PsycINFO, and Web of Science between October 15, 2022, and January 5, 2023. In total, 2 follow-up searches were conducted on May 2, 2023, and November 7, 2023. The search string was developed based on the participants, intervention, comparators, and outcome framework [91]. The framework was adapted to fit the research purpose, resulting in a 3-tiered search term including the population (health care professionals), technology (AI-CDSSs), and outcome (use intention) of interest. An overview of the search terms is presented in Table S3 in Multimedia Appendix 1. We used the search terms to search titles, abstracts, and keywords. We conducted follow-up searches in Google Scholar using the following search string: (“health care”), AND (“Artificial Intelligence”) AND (“UTAUT”). Second, we conducted forward searching of studies citing the seminal article by Venkatesh et al [22] via Google Scholar and backward searches of review articles [92-96]. Third, abstracts of relevant conference proceedings, including the *Conference on Computer-Supported Cooperative Work and Social Computing*, the *Conference on Human Factors in Computing Systems*, and the *Institute of Electrical and Electronics Engineers*, were searched. Fourth, we sent requests for unpublished articles and data using the mailing list of the German Psychology Association. Finally, authors of articles included in the meta-analysis were contacted and asked for unpublished data sets. No additional unpublished data were obtained.

We reached out to authors when critical information was needed to decide on the inclusion of a study or details essential for the meta-analytic synthesis, such as a correlation table, were missing. From the 24 authors contacted to procure missing information, we successfully obtained 6 data sets. These data sets were used to derive the missing information, for instance, to calculate missing correlations between variables of interest.

Figure S1 in Multimedia Appendix 1 shows the PRISMA diagram with the number of studies identified, included, and excluded, along with reasons for exclusion. The studies from the literature search were assessed following a 3-stage approach. First, titles were screened to identify relevant articles. Second, the abstracts of the remaining articles were reviewed. Third, full article texts were reviewed. As a result of a review of 107 full texts, 17 (15.9%) studies met the inclusion criteria (k=18 independent samples; N=3871).

Following the approach of previous meta-analyses, we only included relationships that were identified in a minimum of 3 separate samples [43,97]. We grouped overlapping variables into construct categories (see Table S4 in Multimedia Appendix 1 for definitions of superordinate constructs and subconstructs. Studies from both the primary and the follow-up literature search were coded by 2 researchers each (AK and SG for the primary search and JD and AK for the follow-up search). Any conflicts in the coding were resolved in weekly consensus meetings. In addition, in line with approaches to ensure accuracy in coding established in previous meta-analyses [98], a random sample of 56% (10/18) of the independent samples was recoded by JC and AS. We included agreement on correlations, reliabilities, and moderator categories into the assessment of interrater agreement. Overall interrater agreement was high (94.7%). Notably, no disagreements were observed regarding correlations. Some mistakes in the coding of reliabilities occurred during recoding due to referencing an incorrect line from the source document. The final code sheet used for the analyses is available on request from the Corresponding Author.

Meta-Analytic Procedures

All analyses were conducted using RStudio (Posit Software, PBC) [99] using the R packages *psychmeta* [100] and *metaSEM* [101].

Bivariate Relationships

To examine the bivariate relationship between the 4 core constructs of the UTAUT (hypotheses 1-4) and the additional predictors (RQs 1-5) with the intention to use AI-CDSSs, a random-effects meta-analysis was conducted [102]. Effect sizes were based on Pearson product-moment correlations. Composites were calculated if multiple measures of the same construct were reported for the same sample [102]. Specifically, a variance-weighted composite (across measures of the same construct) was calculated for each independent sample to combine multiple measures of the same construct into a single effect size per independent sample [102]. Sampling errors were corrected using sample size–weighted correlations. Measurement errors were corrected based on the Cronbach α [102]. In addition to the sample size–weighted correlation (r) and sample size–weighted and reliability-corrected correlation (r_c), the 95% CI and 80% credibility interval (CR) for r_c were reported. Finally, we reported the correlation between observed effects and the influence of the study design artifacts.

Relative Weight Analysis

We conducted relative weight analyses to capture the contribution of the correlated predictors [103]. Specifically, we calculated multivariate meta-analytic regression models based on the pooled correlation matrices to explore the incremental value of the UTAUT predictors and additional predictor variables in explaining the intention to use AI-CDSSs. We used the harmonic mean of the sample size across the correlations considered as the sample size for the estimated regression models [104]. In relative weight analysis, raw relative weights are calculated to reflect the proportion of variance explained in the outcome that is attributed to each of the predictors, whereas rescaled relative weights reflect the percentage of the variance that is explained by each predictor variable [105,106].

XSL•FO
RenderX

Moderation Analyses

Moderator analyses were carried out for constructs that were represented in a minimum of 56% (10/18) of the independent samples to ensure adequate coverage of moderator categories [107]. A total of 5 constructs met this minimum cutoff and were considered for the moderation analyses (ie, performance expectancy, effort expectancy, social influence, trust, and perceived risk). We interpreted categorical moderator effects if each of the levels included ≥3 independent samples. Age was coded as the mean age of study participants, and gender was coded as the percentage of women in the sample. For the type of AI-CDSS, 3 categories were initially identified: diagnostic decision support systems, treatment decision support systems, and systems that combined both diagnostic and treatment decision support. However, the *treatment decision support systems* category had to be excluded from the moderator analysis because of the low number of independent samples focusing on this type of AI-CDSS (2/18, 11%). Culture was operationalized based on the individualism versus collectivism dimension of the country comparison tool by Hofstede [108,109]. A higher score denotes stronger individualism. The publication year was coded chronologically. Finally, the scale used to measure the intention to use AI-CDSSs was coded as a categorical moderator. We differentiated between studies using the scale by Venkatesh et al [22] and studies using self-developed scales. We conducted moderation analyses that were not preregistered as part of exploratory analyses.

Mediation Analysis

To test RQ 6, correlation-based meta-analytic structural equation modeling [110] based on the 2-stage structural equation modeling approach [111,112] was performed. In the first step, the sample size–weighted and reliability-corrected bivariate correlation matrices for each independent sample were pooled together. In 2-stage structural equation modeling, the total sample size is used for the estimation of the meta-analytic structural equation model [112]. In the second step, a path model was fitted to the pooled correlation matrix.

Results

Study Characteristics

The overall mean age of the participants was 36.2 (SD 13.5; range 21-53) years, and 48.7% were female. A total of 41% (7/17) of the studies focused on diagnostic AI-CDSSs, 12% (2/17) focused on treatment AI-CDSSs, 24% (4/17) focused on treatment and diagnostic AI-CDSSs, and 24% (4/17) focused on unspecific AI-CDSSs. In total, 65% (11/17) of the studies were conducted in Asia (6/11, 55% in China), 18% (3/17) were conducted in Europe, 6% (1/17) were conducted in the United States, and 12% (2/17) were conducted worldwide in English-speaking countries.

Meta-Analytic Results

In the following sections, we report sample size–weighted and reliability - corrected correlations (r_c) for the relationships between relevant antecedent variables and AI-CDSS use intention. In line with Cohen [113], we classified our reported effects as weak (r_c=0.1), moderate (r_c=0.3), and strong (r_c=0.5).

Bivariate Relationships

The results of bivariate meta-analytic analyses are shown in Table 1. The UTAUT predictors performance expectancy (*r*c=0.66, 95% CI 0.59-0.73), effort expectancy (*r*c=0.55, 95% CI 0.43-0.67), social influence (*r*c=0.66, 95% CI 0.59-0.72), and facilitating conditions (*r*c=0.66, 95% CI 0.42-0.90) exhibited a strong positive relationship with the intention to use AI-CDSSs. The findings support hypotheses 1 to 3 and 4.1. The relationship between facilitating conditions and actual use was not investigated in a sufficient number of independent samples (k<2). Accordingly, we could not address hypothesis 4.2. Regarding the additional predictors beyond the UTAUT, attitude (*r*c=0.63, 95% CI 0.52-0.73), trust (*r*c=0.73, 95% CI 0.63-0.82), and innovativeness (*r*c=0.54, 95% CI 0.43-0.64) exhibited strong positive relationships, confirming RQs 1, 2, and 5. Perceived risk (*r*c=–0.21, 95% CI –0.35 to –0.08) was weakly negatively related to use intention, supporting RQ 3. Although the estimate for AI anxiety was strong and negative (*r*c=–0.41), the 95% CI included 0 (–0.98 to 0.15). Accordingly, we cannot conclude that AI anxiety is related to use intention, thus not supporting RQ 4. The 80% CRs for effort expectancy (0.27-0.83), facilitating conditions (0.33-0.99), and AI anxiety (–0.81 to –0.01) were wide, suggesting the presence of moderators [101,112]. Finally, the intention to use AI-CDSSs was strongly positively related to the actual use of AI-CDSSs, confirming RQ 6 (3/18, 17% of independent samples; N=478; *r*=0.75; r_c=0.85, SD 0.09, 95% CI 0.63-1.00, 80% CR 0.70-1.00; correlation between observed effects and the influence of the study design artifacts=0.44).

XSL•FO
RenderX

Table 1. Bivariate relationships between predictor variables and artificial intelligence–enabled clinical decision support system use intention (N=18).

Predictor variable	Independent samples, n (%)	Cumulative sample size, N	r [a]	r_c [b] (SD; 95% CI)	80% CR[c]	Correlation between r and statistical artifacts
Performance expectancy	16 (89)	3295	0.59	0.66 (0.13; 0.59 to 0.73)	0.50 to 0.82	0.39
Effort expectancy	15 (83)	3058	0.49	0.55 (0.22; 0.43 to 0.67)	0.27 to 0.83	0.28
Social influence	15 (83)	3058	0.57	0.66 (0.12; 0.59 to 0.72)	0.52 to 0.80	0.46
Facilitating conditions	6 (33)	1048	0.57	0.66 (0.23; 0.42 to 0.90)	0.33 to 0.99	0.25
Attitude	9 (50)	2048	0.51	0.63 (0.14; 0.52 to 0.73)	0.45 to 0.80	0.43
Trust	10 (56)	1840	0.66	0.73 (0.13; 0.63 to 0.82)	0.55 to 0.90	0.35
Perceived risk	10 (56)	2428	–0.19	–0.21 (0.18; –0.35 to –0.08)	–0.45 to 0.02	0.39
Anxiety	3 (17)	391	–0.37	–0.41 (0.23; –0.98 to –0.15)	–0.81 to –0.01	0.38
Innovativeness	5 (28)	843	0.47	0.54 (0.09; 0.43 to 0.64)	0.46 to 0.61	0.81

[a]Sample size–weighted correlation.

[b]Sample size–weighted and reliability-corrected correlation.

[c]CR: credibility interval.

Relative Weight Analysis

It was not possible to explore all 9 predictors in a single relative weight analysis because they were not investigated together in a sufficient number of independent samples (Table S5 in Multimedia Appendix 1). Accordingly, to answer RQ 7, we analyzed 1 model with only the UTAUT predictors (Table 2) and 4 separate extension models consisting of 5 to 6 predictors (Table 3). In the initial model with only the UTAUT predictors, the combined effects of performance expectancy, effort expectancy, social influence, and facilitating conditions explained 50% of the total variance in the intention to use AI-CDSSs. Performance expectancy was the dominant predictor, accounting for 31% of the total variance explained, followed by social influence (28%), facilitating conditions (26%), and effort expectancy (15%). In the extension models, trust emerged as the most influential overall predictor of use intention (between 29% and 35% of the total variance explained). In all 3 models including trust, performance expectancy was the second most influential predictor (between 19% and 24% of the total variance explained). Facilitating conditions (between 20% and 25%) and social influence (between 14% and 21%) consistently explained additional variance in all extension models. In the extension models including trust and perceived risk as well as trust and anxiety, the regression estimate of effort expectancy became negative. Finally, AI anxiety and perceived risk negatively predicted use intention and accounted for 10% (AI anxiety) and 2% (perceived risk) of the total variance explained.

Table 2. Multiple regression models and relative weights for the Unified Theory of Acceptance and Use of Technology predictors[a].

Predictor	B[b] (SE)	t test (df)	P value	Raw RW[c]	RS[d] RW (%)
Performance expectancy	0.31 (0.02)	13.97 (1732)	<.001	0.16	31.19
Effort expectancy	0.08 (0.02)	3.56 (1732)	<.001	0.08	15.2
Social influence	0.27 (0.02)	12.29 (1732)	<.001	0.14	27.91
Facilitating conditions	0.21 (0.02)	9.33 (1732)	<.001	0.13	25.7

[a]$F_{4,1732}$=429.28 (P<.001); R^2=0.498.

[b]Regression estimate.

[c]RW: relative weight.

[d]RS: rescaled.

XSL•FO
RenderX

Table 3. Multiple regression models and relative weights for the Unified Theory of Acceptance and Use of Technology (UTAUT) and additional predictors.

Predictor	B^a (SE)	*t* test (*df*)	*P* value	Raw RW[b]	RS[c] RW (%)
UTAUT extension (attitude and perceived risk; $F_{6,1284}$=222.31; *P*<.001; R^2=0.509)					
Performance expectancy	0.25 (0.03)	9.36 (1284)	<.001	0.12	24
Effort expectancy	0.05 (0.03)	2.04 (1284)	.04	0.06	12.14
Social influence	0.17 (0.03)	6.40 (1284)	<.001	0.10	20.54
Facilitating conditions	0.28 (0.03)	10.82 (1284)	<.001	0.13	25.31
Attitude	0.13 (0.03)	5.02 (1284)	<.001	0.08	15.91
Perceived risk	–0.04 (0.02)	–2.20 (1284)	.03	0.01	2.09
UTAUT extension (trust and innovativeness; $F_{5,1305}$=308.50; *P*<.001; R^2=0.542)					
Performance expectancy	0.22 (0.03)	8.77 (1305)	<.001	0.12	22.72
Effort expectancy	0.05 (0.02)	2.14 (1305)	.03	0.06	11.57
Social influence	0.19 (0.03)	7.62 (1305)	<.001	0.11	20.4
Trust	0.39 (0.03)	15.40 (1305)	<.001	0.19	35.04
Innovativeness	0.04 (0.02)	1.56 (1305)	.12	0.06	10.26
UTAUT extension (trust and perceived risk; $F_{6,1556}$=389.61;*P*<.001; R^2=0.600)					
Performance expectancy	0.18 (0.02)	8.40 (1556)	<.001	0.11	18.76
Effort expectancy	–0.06 (0.02)	–2.65 (1556)	.01	0.05	8.81
Social influence	0.09 (0.02)	3.87 (1556)	<.001	0.09	15.66
Facilitating conditions	0.32 (0.02)	14.99 (1556)	<.001	0.13	22.03
Trust	0.42 (0.02)	19.79 (1556)	<.001	0.20	33
Perceived risk	–0.05 (0.02)	–2.80 (1556)	.01	0.01	1.74
UTAUT extension (trust and anxiety; $F_{6,843}$=241.15; *P*<.001; R^2=0.632)					
Performance expectancy	0.23 (0.03)	8.15 (843)	<.001	0.12	19.25
Effort expectancy	–0.11 (0.03)	–3.92 (843)	<.001	0.05	7.24
Social influence	0.07 (0.03)	2.44 (843)	.02	0.09	14.13
Facilitating conditions	0.31 (0.03)	11.43 (843)	<.001	0.13	20.44
Trust	0.38 (0.03)	13.48 (843)	<.001	0.18	28.68
Anxiety	–0.20 (0.02)	–8.73 (843)	<.001	0.06	10.26

[a]Regression estimate.

[b]RW: relative weight.

[c]RS: rescaled.

Moderation Analyses

Table 4 shows the results of the meta-regression for continuous moderators. Regarding age, older participants showed a weaker relationship between social influence and use intention (*B*=–0.01, 95% CI –0.01 to –0.00), thus *contradicting* hypothesis 5.3, according to which this effect would become stronger with increasing age. The moderation effect is shown in Figure S2 in Multimedia Appendix 1. Age did not moderate any other relationship, thus not confirming hypotheses 5.1 and 5.2. Gender did not moderate any of the relationships, thus not confirming hypotheses 6.1, 6.2, and 6.3. Experience and voluntariness of use were not investigated in a sufficient number of independent samples. Accordingly, we were unable to address hypotheses 7 and 8. Cultural individualism (RQ 8.3) as a contextual moderator that was measured continuously did not influence any of the relationships. Finally, publication year (RQ 9.1) as a methodological moderator that was measured continuously did not influence any of the relationships.

XSL•FO
RenderX

Table 4. Results of the meta-regression (N=18).

Predictor variable and moderator	Independent samples per moderator, n (%)	*B* (SE; 95% CI)	*P* value
Performance expectancy			
Age	4 (22)	<0.01 (<0.01; –0.01 to 0.01)	.56
Gender (percentage women)	16 (89)	<0.01 (<0.01; –0.00 to 0.00)	.88
Individualism	14 (78)	<0.01 (<0.01; –0.00 to 0.00)	.66
Publication year	16 (89)	0.02 (0.03; –0.03 to 0.07)	.42
Effort expectancy			
Age	4 (22)	<0.01 (0.01; –0.02 to 0.02)	.97
Gender (percentage women)	15 (83)	<0.01 (<0.01; –0.01 to 0.00)	.63
Individualism	13 (72)	<0.01 (<0.01; –0.00 to 0.00)	.95
Publication year	15 (83)	0.03 (0.04; –0.06 to 0.09)	.68
Social influence			
Age	4 (22)	–0.01 (<0.01; –0.01 to –0.00)	.03
Gender (percentage women)	15 (83)	<0.01 (<0.01; –0.00 to 0.01)	.09
Individualism	13 (72)	<0.01 (<0.01; –0.00 to 0.00)	.56
Publication year	15 (83)	0.02 (0.02; –0.02 to 0.07)	.21
Trust			
Age	3 (17)	<0.01 (0.01; –0.02 to 0.02)	.90
Gender (percentage women)	10 (56)	<0.01 (<0.01; –0.01 to 0.00)	.88
Individualism	9 (50)	<0.01 (<0.01; –0.00 to 0.00)	.91
Publication year	10 (56)	–0.04 (0.03; –0.10 to 0.03)	.24
Perceived risk			
Gender (percentage women)	10 (56)	<0.01 (<0.01; –0.01 to 0.01)	.64
Individualism	9 (50)	<0.01 (<0.01; –0.00 to 0.01)	.21
Publication year	8 (44)	–0.02 (0.04; –0.11 to 0.06)	.60

The Wald-type pairwise comparisons for each level of categorical moderators are presented in Table 5. We could not investigate RQ 8.1 because information about occupations was not provided in a sufficient number of independent samples. Regarding RQ 8.2, the type of AI-CDSS (diagnostic AI-CDSSs versus diagnostic and treatment AI-CDSSs) did not moderate the relationship between performance expectancy and use intention nor did it moderate the relationship between social influence and use intention. However, the positive relationship between effort expectancy and use intention was stronger for diagnostic AI-CDSSs than for AI-CDSSs that combined diagnostic and treatment recommendations (mean difference=–0.31, 95% CI –0.58 to –0.04). Finally, regarding RQ 9.2, we observed no differences between studies using the scale by Venkatesh et al [22] and those using other measures.

XSL•FO
RenderX

Table 5. Wald-type pairwise comparisons of categorical moderators (N=18)[a].

Predictor variable	Independent samples for moderator level 1, n (%)	Independent samples for moderator level 2, n (%)	*F* test (*df*)	r_{c1}[b]	r_{c2}[c]	Mean difference (95% CI)
AI-CDSS[d] type: diagnostic and treatment AI-CDSSs (level 1) compared to diagnostic AI-CDSSs (level 2)						
Performance expectancy	4 (22)	7 (39)	1.48 (*3*, *3*)	0.62	0.71	–0.09 (–0.31 to 0.13)
Effort expectancy	4 (22)	6 (33)	7.15 (*3*, *3*)	0.41	0.72	–0.31 (–0.58 to –0.04)
Social influence	4 (22)	6 (33)	6.97 (*3*, *3*)	0.59	0.72	–0.14 (–0.28 to <0.01)
Trust	3 (17)	4 (22)	0.09 (*2*, *3*)	0.70	0.72	–0.02 (–0.28 to 0.24)
Use intention scale: other scales (level 1) compared to the scale by Venkatesh et al [22] (level 2)						
Performance expectancy	10 (56)	6 (33)	0.07 (*1*, *10*)	0.66	0.68	–0.02 (–0.18 to 0.15)
Effort expectancy	9 (50)	6 (33)	0.48 (*1*, *9*)	0.59	0.51	0.08 (–0.18 to 0.35)
Social influence	9 (50)	6 (33)	0.82 (*1*, *9*)	0.64	0.70	–0.06 (–0.21 to 0.09)
Trust	5 (28)	5 (28)	0.01 (*1*, *6*)	0.73	0.72	0.01 (–0.19 to 0.20)
Perceived risk	7 (39)	3 (17)	2.41 (*1*, *2*)	–0.16	–0.37	0.21 (–0.10 to 0.53)

[a]Moderator analysis for constructs assessed in at least 10 independent samples.

[b]Sample size–weighted and reliability-corrected correlation for moderator level 1.

[c]Sample size–weighted and reliability-corrected correlation for moderator level 2.

[d]AI-CDSS: artificial intelligence–enabled clinical decision support system.

The Mediating Role of Performance and Effort Expectancy in the Relationship Between Facilitating Conditions and AI-CDSS Use Intention

The role of performance and effort expectancy as mediators of the relationship between facilitating conditions and intention to use AI-CDSSs (RQs 10.1 and 10.2) was analyzed by fitting 2 separate mediation models. The results are shown in Table 6. Performance expectancy and effort expectancy mediated the relationship between facilitating conditions and the intention to use AI-CDSSs (indirect effect for performance expectancy: *B*=0.20, 95% CI 0.12-0.34; indirect effect for effort expectancy: *B*=0.21, 95% CI 0.09-0.37).

XSL•FO
RenderX

Table 6. Mediation models with performance and effort expectancy as mediators.

Path	*B*[a] (95% CI)
Mediator: performance expectancy	
Direct effects	
Facilitating conditions→performance expectancy	0.38 (NA[b] to 0.54)
Performance expectancy→use intention	0.53 (0.36 to 0.70)
Facilitating conditions→use intention	0.29 (–0.01 to 0.57)
Indirect effect	
Facilitating conditions→performance expectancy→use intention	0.20 (0.12 to 0.34)
Mediator: effort expectancy	
Direct effects	
Facilitating conditions→effort expectancy	0.48 (0.35 to 0.62)
Effort expectancy→use intention	0.43 (0.17 to 0.68)
Facilitating conditions→use intention	0.29 (–0.04 to 0.61)
Indirect effect	
Facilitating conditions→effort expectancy→use intention	0.21 (0.09 to 0.37)

[a]Regression estimate.

[b]NA: the lower bound of the CI could not be estimated.

Sensitivity Analysis

To assess the robustness of the meta-analytic findings, we used cumulative meta-analysis. This approach involves conducting a sequence of iterative meta-analyses, with each analysis adding an effect size for a specific relationship. Effect sizes are added in order of decreasing precision, meaning that the initial effect sizes added represent the most accurate population effect size estimates. If less precise studies tend to skew the meta-analytic estimates, this will be observable as a shift in cumulative results when these studies are included [106]. The results of the cumulative meta-analyses are shown in Figure S3 in Multimedia Appendix 1. A total of 5 “drifts” were identified, and all relationships drifted toward stronger effects as less precise studies were added, indicating an overestimation of the true effect. However, meaningful differences were not observed for any of the relationships after half the studies were added compared to after all the studies were added (Table S6 in Multimedia Appendix 1). Accordingly, we conclude that none of the drifts influenced the meta-analytic conclusions.

Discussion

Summary of Findings and Implications for Future Research

The primary goal of the meta-analysis was to gain a better understanding of the predictors of intention to use AI-CDSSs among health care practitioners based on the UTAUT and its extensions. The results of the meta-analysis provide empirical support for the applicability of the UTAUT to the context of AI-CDSSs. As predicted, performance expectancy, effort expectancy, social influence, and facilitating conditions were positively related to the intention to use AI-CDSSs. These findings are largely in line with the findings of UTAUT meta-analyses in other fields [27,32,40,61]. We provide a summary of the main findings for our hypotheses and RQs in Tables 7 and 8.

The results of relative weight analyses showed that all 4 UTAUT predictors together explained 50% of the variance in use intention among health care practitioners, reaffirming the relevance of the UTAUT predictors in the context of AI-CDSSs. Among the UTAUT predictors, performance expectancy emerged as the most relevant, accounting for 31% of the total explained variance, followed by social influence (28%), facilitating conditions (26%), and effort expectancy (15%). In most UTAUT research, performance expectancy is more relevant than effort expectancy, possibly because performance expectancy is inherently connected to the primary motives behind technology use [32,40]. That is, it directly relates to the perceived benefits that users expect to gain from using a technology [32,40]. Effort expectancy refers to the expected ease of using a technology [22]. While important, the ease of use may become a secondary consideration if the technology does not meet the primary performance-related objectives. In other words, users might be willing to overcome a steeper learning curve if they believe the payoff in performance is worthwhile [63]. This could explain why performance expectancy accounts for a higher percentage of the variance in technology acceptance and use intentions compared to effort expectancy. Overall, the findings of this meta-analysis reflect a common finding in technology acceptance research where the anticipated improvement in performance is often found to be a stronger driver of user acceptance than the anticipated effort to learn and use the technology [32,40,61,63].

XSL•FO
RenderX

Table 7. Results of the investigation of the hypotheses (N=18).

Hypothesis	Independent samples, n (%)	Effect size (95% CI)	Result	Main findings
1	16	0.66[a] (0.59 to 0.73)	Supported	Performance expectancy is positively related to the intention to use AI-CDSSs[b].
2	15	0.55[a] (0.43 to 0.67)	Supported	Effort expectancy is positively related to the intention to use AI-CDSSs.
3	15	0.66[a] (0.59 to 0.72)	Supported	Social influence is positively related to the intention to use AI-CDSSs.
4.1	6	0.66 (0.42 to 0.90)	Supported	Facilitating conditions are positively related to the intention to use AI-CDSSs.
4.2	<3	__[c]	—	Not enough independent samples (<3) to examine the relationship between facilitating conditions and actual use of AI-CDSSs
5.1	4	<0.01[d] (–0.01 to 0.01)	Not supported	The relationship between performance expectancy and intention to use AI-CDSSs does not become weaker with increasing age.
5.2	4	<0.01[d] (–0.02 to <0.01)	Not supported	The relationship between effort expectancy and intention to use AI-CDSSs does not become stronger with increasing age.
5.3	4	−0.01[d] (–0.01 to <–0.01)	Not supported	The relationship between social influence and intention to use AI-CDSSs becomes weaker with increasing age.
6.1	16	<0.01[d] (<–0.01 to <0.01)	Not supported	The relationship between performance expectancy and intention to use AI-CDSSs is not stronger for men.
6.2	15	<0.01[d] (–0.01 to <0.01)	Not supported	The relationship between effort expectancy and intention to use AI-CDSSs is not stronger for women.
6.3	15	<0.01[d] (<–0.01 to 0.01)	Not supported	The relationship between social influence and intention to use AI-CDSSs is not stronger for women.
7.1-7.3	<3	—	—	Not enough independent samples (<3) to examine experience as a moderator
8	<3	—	—	Not enough independent samples (<3) to examine mandatory versus voluntary adoption setting as a moderator

[a]Sample size–weighted and reliability-corrected correlation.

[b]AI-CDSS: artificial intelligence–enabled clinical decision support system.

[c]Not applicable.

[d]Regression estimate.

Among the UTAUT predictors, effort expectancy and facilitating conditions had the widest CRs (0.56 and 0.66, respectively), suggesting the presence of moderating influences [102,114]. For example, previous research suggests that radiologists, accustomed to complex machines and heavy workloads, may be willing to invest effort into learning how to use new technology if it reduces their workload, indicating a moderating influence of occupation on the relationship between effort expectancy and use intention [4,37]. In addition, the strength of the relationship between effort expectancy and use intention has been shown to differ between AI-CDSSs for feedback versus decision support [24].

In addition to the core UTAUT variables, we identified attitude, trust, perceived risk, AI anxiety, and personal innovativeness as predictors of the intention to use AI-CDSSs. Although all the included studies (17/17, 100%) reported a negative relationship between AI anxiety and use intention, the CI for AI anxiety included 0. This lack of an observed relationship may be due to the low sample sizes (the total sample size was 391) and the resulting high uncertainty in the true effect. Interestingly, in the relative weight analyses, trust proved to be a more relevant factor than performance expectancy in explaining variance in the intention to use AI-CDSSs. The relevance of trust may be explained by the lack of transparency in how AI recommendations are generated coupled with the high stakes associated with clinical decision-making [115]. Indeed, research has suggested that even highly efficient AI-CDSSs may face resistance in clinical applications if health care practitioners do not trust the system's safety [116-118]. The findings of this meta-analysis align with those of research advocating for the inclusion of trust in the UTAUT model [116].

XSL•FO
RenderX

Table 8. Results of the investigation of the research questions (N=18).

Research question	Independent samples, n (%)	Effect size (95% CI)	Result	Main findings
1	9	0.63[a] (0.52 to 0.73)	Answered	Positive attitude toward AI-CDSSs[b] is positively related to intention to use AI-CDSSs.
2	10	0.73[a] (0.63 to 0.82)	Answered	Trust is positively related to intention to use AI-CDSSs.
3	10	–0.21[a] (–0.35 to –0.08)	Answered	Perceived risk is negatively related to intention to use AI-CDSSs.
4	3	–0.41[a] (–0.98 to –0.15)	Answered	AI[c] anxiety is negatively related to intention to use AI-CDSSs.
5	5	0.54[a] (0.43 to 0.64)	Answered	Personal innovativeness is positively related to intention to use AI-CDSSs.
6	3	0.85[a] (0.63 to 1.00)	Answered	The intention to use AI-CDSSs is positively related to their actual use.
7	__[d]	See Table 3	Partially answered	See Table 3
8.1	<3	—	—	Not enough independent samples (<3) to examine occupation as a moderator
8.2	4 for moderator level 1; 6 for moderator level 2	0.31[e] (–0.58 to –0.04)	Answered	The positive relationship between effort expectancy and use intention was weaker for diagnostic and treatment AI-CDSSs (moderator level 1) than for diagnostic AI-CDSSs (moderator level 2).
8.3	13-14	See Table 4	Answered	Cultural background (individualism) does not moderate the relationship between performance expectancy (14 independent samples), effort expectancy (13 independent samples), and social influence (13 independent samples) and intention to use AI-CDSSs.
9.1	15-16	See Table 4	Answered	Publication year does not moderate the relationship between performance expectancy (16 independent samples), effort expectancy (15 independent samples), and social influence (15 independent samples) and intention to use AI-CDSSs.
9.2	6-10	See Table 5	Answered	There are no differences in the relationships between performance expectancy, effort expectancy, and social influence and intention to use AI-CDSSs between samples that used the scale by Venkatesh et al [22] to measure use intention (moderator level 2) and those that used other scales (moderator level 1).
10.1	16	0.20[f] (0.12 to 0.34)	Answered	Performance expectancy mediates the relationship between facilitating conditions and intention to use AI-CDSSs.
10.2	16	0.21[f] (0.09 to 0.37)	Answered	Effort expectancy mediates the relationship between facilitating conditions and intention to use AI-CDSSs.

[a]Sample size–weighted and reliability-corrected correlation.

[b]AI-CDSS: artificial intelligence–enabled clinical decision support system.

[c]AI: artificial intelligence.

[d]Not applicable.

[e]Mean difference between sample size–weighted and reliability-corrected correlation for moderator levels 1 and 2.

[f]Regression estimate (indirect effect).

Furthermore, this meta-analysis emphasizes the need to consider both drivers and inhibitors of the intention to use AI-CDSSs for a more comprehensive understanding of the adoption process [119]. The relative weight analyses demonstrate that AI anxiety explained approximately 10% of the variance in the intention to use AI-CDSSs after trust (29%), facilitating conditions (20%), performance expectancy (19%), and social influence (14%) and before effort expectancy (7%). The relevance of perceived risk as a predictor of use intention was small (approximately 2% after all other predictors). Risk perception is a cognitive assessment of the potential losses and gains from using AI-CDSSs, which is based on logical evaluation and can be mitigated by providing relevant information [120]. In contrast, AI anxiety is an emotional response that encompasses fears and insecurities about AI technology [121]. Accordingly, AI anxiety is less rational and more difficult to alleviate because it can be deeply rooted in concerns about AI's impact on job security, professional autonomy, and the quality of patient care [24,122,123].

Even for relationships assessed in a substantial number of independent samples, such as performance expectancy, effort

XSL•FO
RenderX

expectancy, attitude, trust, and perceived risk, the CRs were wide (>0.34), suggesting the presence of moderators [102,114]. This observation is supported by the modest amount of variance accounted for by statistical artifacts, indicating that there may be other reasons for substantial variance between individual studies [102,114]. While we considered multiple moderators suggested by the UTAUT and additional contextual and methodological moderators, we only found 2 moderation effects.

First, age moderated the relationship between social influence and use intention, with older health care practitioners exhibiting a weaker relationship between social influence and use intention. This finding does not align with the UTAUT proposing that older individuals place more relevance on the opinion of relevant others when intending to use a new technology [22]. An explanation for this discrepancy may be that practitioners value their professional independence increasingly more with age, thus relying less on the opinion of others regarding technology use as they get older. While differences in professional values and behaviors have been shown to exist between younger and older health care practitioners [124], a systematic examination of the effect of age on the relationship between social influence and the intention to use AI-CDSSs is lacking. It has to be noted that the observed moderation effect was based on only 22% (4/18) of independent samples, underscoring the need to systematically study the influence of age on the relationship between social influence and use intention.

Second, the relationship between effort expectancy and use intention was stronger for diagnostic AI-CDSSs than for devices that combined diagnostic and treatment recommendations. When clinicians assess a tool solely for diagnostic purposes, they may find it easier to anticipate the required effort to use it as the task is less complex and the outcome is more direct. This clear understanding may strengthen the relationship between effort expectancy and use intention. Indeed, the perceived risk associated with smart devices has been found to negatively influence the relationship between effort expectancy and use intention [125]. The multifaceted nature of combined tools may make it more challenging for clinicians to evaluate the effort needed to understand and use them. This uncertainty possibly leads to a weaker relationship between effort expectancy and use intention as clinicians may not be able to adequately assess the effort required, thus not being able to use it as a source of information when it comes to indicating their intention to use it. Future research is needed to investigate the moderating influence of device type.

The results of the mediation analyses indicate that the relationship between facilitating conditions and use intention may be explained through effort expectancy and performance expectancy. This finding aligns with the UTAUT proposing that, when performance and effort expectancy are considered, facilitating conditions lose their importance in predicting use intention [22]. An explanation for the relevance of effort expectancy as a mediator may be that issues related to the support infrastructure, a critical aspect of facilitating conditions, are also conceptually addressed by effort expectancy [22]. That is, if health care organizations establish the appropriate support infrastructure, the effort required to use AI-CDSSs becomes lower [22,54]. Similarly, if a user perceives that the technology is supported by adequate facilitating conditions, they may be more likely to believe in the performance benefits when using the system, explaining the mediating role of performance expectancy.

Practical Implications

Performance expectancy and trust emerged as the 2 most relevant predictors of AI-CDSS use intention, suggesting that measures targeted toward health care practitioners' beliefs in the performance and trustworthiness of AI-CDSSs may be effective in enhancing their intention to use them. However, the consistently positive link between performance expectancy and use intention also suggests that health care institutions need to take measures to deter the perception of low-performing systems as high performing, which could potentially cause more harm than benefit [126]. Health care practitioners require transparent communication regarding the performance and limitations of AI-CDSSs alongside adequate training to ensure their correct use. In addition, regulatory bodies such as the Food and Drug Administration need to ensure that available AI-CDSSs meet certain safety and performance standards [84,127,128]. Adequate policies and oversight in these contexts may ensure a balance between the adoption and safe application of AI-CDSSs in health care decision-making.

Trust in technology is a multifaceted construct including users' perceptions of a system's functionality, helpfulness or benevolence, and integrity [48,49]. Consequently, actions taken to enhance performance expectancy may not be sufficient for building trust in a system [116]. If organizations aim to improve health care practitioners' trust in AI-CDSSs, they need to address the various facets relevant to trust in technology. This includes dealing with ethical issues related to data privacy and the potential misuse of AI-CDSSs as well as addressing the lack of transparency and explainability in AI-generated recommendations [129,130]. For example, trust has been associated with the system's capability to explain its decision-making process, emphasizing the role of explainable AI as a path to building trust in AI-CDSSs [116,131]. In addition, regulatory strategies should be designed to promote and maintain trust in AI-CDSSs along with safe patient outcomes. This might include the use of postmarket surveillance systems to monitor the performance of deployed AI-CDSSs over time, which has been suggested as a method for identifying and mitigating issues of utility and safety in real-world clinical settings [132,133]. Developers can integrate user-centered design principles to tailor AI-CDSSs to the needs and workflows of specific clinical specialties and roles. The early inclusion of user feedback may facilitate the development of user-friendly AI-CDSSs and increase trust in these systems [134,135]. Health care administrators may foster practitioners' trust by providing training programs to increase familiarity with the technology and by designing evaluation metrics that can monitor system performance and user satisfaction [136].

Social influence has been demonstrated to be a relevant predictor of health care practitioners' intentions to use AI-CDSSs, particularly among younger professionals. Institutions aiming to adopt AI-CDSSs can leverage the important role of social influence by establishing a culture that values technological

XSL•FO
RenderX

advancements and by engaging key opinion leaders to advocate and exemplify the use of these systems. In addition, trainings can be structured not only to educate but also to establish a shared understanding and a community of practice that positively reinforces the application of AI-CDSSs [137,138]. By addressing the social aspects of technology acceptance, health care institutions can ensure that their investment in AI is met with a user base that is both competent and willing to integrate these tools into their daily practice.

The importance of facilitating conditions underscores the need for health care organizations to provide a supportive infrastructure that simplifies the integration of AI-CDSSs into existing workflows. For instance, the provision of training programs, allowing health care practitioners to gain firsthand experience, and setting up accessible support teams ready to address system-related issues can considerably boost health care practitioners' intention to use such systems [23,139].

AI anxiety emerged as a barrier to the intention to use AI-CDSSs in the relative weight analysis. Therefore, hospitals and other health care institutions should consider measures to counteract any irrational negative emotional reactions to AI before and during the integration of AI-CDSSs into clinical workflows. A potential method to mitigate AI anxiety involves increasing medical staff involvement in the development process [123] or providing more training opportunities to increase their exposure to AI-enabled devices, thus reducing irrational fears [137].

Limitations and Implications for Future Research

This meta-analysis is not without limitations. First, this study offers insights into the predictors of use intention as the key determining factor of actual use. However, some health care practitioners may express intention to use AI-CDSSs but be hesitant when it comes to their actual implementation. Few studies included in the meta-analysis (3/17, 18%) examined the predictors of actual use, underscoring the need for additional research on predictors of the actual use of AI-CDSSs [4,48].

Second, we were unable to explain the considerable variation in some of the effects based on moderator analyses. We could not evaluate 3 UTAUT moderators—experience with AI-CDSSs, voluntariness of use, and occupation—owing to insufficient samples incorporating these variables. In addition, although all studies including AI anxiety (3/17, 18%) reported negative relationships with use intention, the CI of the meta-analytic estimate included 0 due to the low sample size and the associated high uncertainty in the estimate. More studies on the relationship between AI anxiety and intention to use AI-CDSSs are needed. The large CRs and the low correlations between estimates and statistical artifacts suggest the existence of moderating factors not included in the meta-analysis [102,114]. Future research should explore moderating effects such as differences in the observed relationships among health care practitioners working in different fields or roles to better understand the boundary conditions that affect the relationships between predictors and the intention to use AI-CDSSs.

Third, the 9 relevant predictors could not be examined in a single relative weight analysis. The use of multiple models with subsets of predictors is a pragmatic approach to addressing data sparsity. However, the selected approach hinders definitive conclusions regarding the importance of all considered predictors. Furthermore, innovativeness could not be assessed in the relative weight analysis due to a lack of available samples assessing this predictor. The compromises that had to be made in the relative weight analyses highlight the need for an updated meta-analysis that includes complete predictor sets.

Fourth, the insights derived from the meta-analysis are primarily confined to unspecific AI-CDSSs. Given that AI-CDSS adoption is still limited, only a handful of studies have delved into exploring predictors of the use of specific AI-CDSSs with distinctive features [4,24,38]. The results of these studies show that the attitude toward AI-CDSSs may vary depending on use cases and system features. Future research should examine the adoption of individual systems and variations in effects across different types of AI-CDSSs.

Fifth, the existing body of research on AI-CDSS adoption primarily relies on cross-sectional observational studies, with questionnaires as the main method of data collection. These studies inherently limit the establishment of causal relationships, thus underscoring the need for future research to include longitudinal or experimental designs for more robust evidence of causality. Longitudinal studies may also be used to shed light on the development of use intention and the relevance of relevant predictors over time. For example, it is possible that initial trust plays a crucial role during the implementation phase but becomes less relevant once a system has been successfully implemented.

Sixth, we selected the UTAUT as a general theoretical framework to examine the predictors of intention to use AI-CDSSs. However, there has been some criticism of the UTAUT [64,140]. For example, the UTAUT may not answer questions related to the determinants and processes involved in *value-adding* technology use [64,141,142]. We found support for the prediction that beliefs about the performance and ease of use of AI-CDSSs lead to a higher intention to use these systems. However, based on the UTAUT, whether these beliefs are well founded (ie, whether positive expectations actually lead to beneficial use because the system is indeed high performing and easily implementable) may not be resolved. Another criticism pertains to the UTAUT's narrow viewpoint on individual use. Other models such as the nonadoption, abandonment, scale-up, spread, and sustainability framework [143] adopt a system perspective. This approach enables the examination of predictors at micro (individual technology users), meso (organizational processes and systems), and macro (national policy and wider context) levels, thereby more accurately representing the complex processes involved in technology adoption [143,144]. In addition, the UTAUT focuses on an individual's intention to use a technology and does not fundamentally consider how well the technology fits the task it is being used for. Theories such as the task-technology fit model examine the interconnectedness between task and technological characteristics. The model delves into how features of both the assigned task and the technology at hand shape the task-technology fit, thereby influencing the overall performance and use intentions [145].

XSL•FO
RenderX

Finally, some of the predictors included in this meta-analysis, particularly the additional predictors beyond the UTAUT, may not be adequately represented using standard measurement instruments. For instance, AI anxiety has a multitude of dimensions, such as privacy violation anxiety, bias behavior anxiety, job replacement anxiety, learning anxiety, or ethics violation anxiety [146]. This meta-analysis did not distinguish between these different aspects of AI anxiety as separate predictors of AI-CDSS use intention. Similarly, trust in AI is a multifaceted construct that includes perceptions of the system's benevolence, competence, and integrity [48,49]. Moreover, trust may refer to different aspects of an AI-CDSS, such as trust in the reliability of its predictions when being applied to different contexts, trust in legal protection if harm to patients occurs from using the AI-CDSS, and trust in data privacy [4,47,147]. More research is needed that explores the relevance of different elements of trust (ie, benevolence, competence, and integrity) and elements of the AI-CDSSs that may be trusted to different degrees (eg, reliability, legal and liability issues, and privacy concerns) for the intention to use AI-CDSSs in clinical decision-making [147].

Conclusions

This meta-analysis underscores the relevance of the UTAUT to examine the predictors of intention to use AI-CDSSs in health care. The results indicate that performance expectancy, effort expectancy, social influence, and facilitating conditions are positively related to the intention to use AI-CDSSs among health care practitioners. The analyses further revealed the relevance of the additional predictors attitude, trust, personal innovativeness, AI anxiety, and perceived risk. The results of mediation analyses show that effort expectancy and performance expectancy explain the relationship between facilitating conditions and use intention. Despite identifying age and AI-CDSS type as moderating influences, there is scope for future research to investigate other possible moderators to explain the variability in the observed effects. While the UTAUT model provides a theoretical framework for studying health care practitioners' intention to use AI-CDSSs, it remains relatively silent on the predictors of value-adding use of AI-CDSSs. Future research could investigate the conditions that encourage value-adding use by applying comprehensive frameworks that consider both individual and broader organizational processes (eg, clinic systems and administrative hurdles). Finally, the findings of this meta-analysis provide starting points for the development and integration of AI-CDSSs that are likely to be adopted by health care practitioners as end users.

Acknowledgments

This research was funded by the Volkswagen Foundation (grant 98 525).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials including tables regarding preregistration deviation , inclusion criteria per included study, search terms per database, construct and subconstruct definitions, pooled meta-analytic correlations and number of samples per correlation, results of the cumulative meta-analyses and figures regarding the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart, age as a moderator of the social influence–use intention relationship, and results of the cumulative meta-analyses.

[PDF File (Adobe PDF File), 1195 KB - jmir_v26i1e57224_app1.pdf]

References

1. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. New York, NY: Pearson; 2016.
2. Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit Med 2022 Jul 21;5(1):97 [FREE Full text] [doi: 10.1038/s41746-022-00597-7] [Medline: 35864312]
3. Tran AQ, Nguyen LH, Nguyen HS, Nguyen CT, Vu LG, Zhang M, et al. Determinants of intention to use artificial intelligence-based diagnosis support system among prospective physicians. Front Public Health 2021 Nov 26;9:755644 [FREE Full text] [doi: 10.3389/fpubh.2021.755644] [Medline: 34900904]
4. Zhai H, Yang X, Xue J, Lavender C, Ye T, Li JB, et al. Radiation oncologists' perceptions of adopting an artificial intelligence-assisted contouring technology: model development and questionnaire study. J Med Internet Res 2021 Sep 30;23(9):e27122 [FREE Full text] [doi: 10.2196/27122] [Medline: 34591029]
5. Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Berner ES, editor. Clinical Decision Support Systems: Theory and Practice. Cham, Switzerland: Springer; 2016:1-17.
6. Knop M, Weber S, Mueller M, Niehaves B. Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence-enabled clinical decision support systems: literature review. JMIR Hum Factors 2022 Mar 24;9(1):e28639 [FREE Full text] [doi: 10.2196/28639] [Medline: 35323118]

XSL•FO
RenderX

7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
8. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
9. Ramgopal S, Sanchez-Pinto LN, Horvat CM, Carroll MS, Luo Y, Florin TA. Artificial intelligence-based clinical decision support in pediatrics. Pediatr Res 2023 Jan 29;93(2):334-341 [FREE Full text] [doi: 10.1038/s41390-022-02226-1] [Medline: 35906317]
10. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021 Aug 10;18(16):8435 [FREE Full text] [doi: 10.3390/ijerph18168435] [Medline: 34444182]
11. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Cham, Switzerland: Springer; 2021:795-840.
12. Mucha H, Robert S, Breitschwerdt R, Fellmann M. Usability of clinical decision support systems. Z Arbeitswiss 2022 Sep 02;77(1):92-101. [doi: 10.1007/S41449-022-00324-8]
13. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022 May 18;28(5):924-933 [FREE Full text] [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
14. Zhang R, Zhang Z, Wang D, Liu Z. Editorial: responsible AI in healthcare: opportunities, challenges, and best practices. Front Comput Sci 2023 Aug 15;5:1265902 [FREE Full text] [doi: 10.3389/fcomp.2023.1265902]
15. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open 2021 Mar 01;4(3):e211276 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1276] [Medline: 33704476]
16. Choi DJ, Park JJ, Ali T, Lee S. Artificial intelligence for the diagnosis of heart failure. NPJ Digit Med 2020 Apr 08;3:54 [FREE Full text] [doi: 10.1038/s41746-020-0261-3] [Medline: 32285014]
17. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18:medethics-2020-106820. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
18. Wang L, Chen X, Zhang L, Li L, Huang Y, Sun Y, et al. Artificial intelligence in clinical decision support systems for oncology. Int J Med Sci 2023;20(1):79-86 [FREE Full text] [doi: 10.7150/ijms.77205] [Medline: 36619220]
19. Hummelsberger P, Koch TK, Rauh S, Dorn J, Lermer E, Raue M, et al. Insights on the current state and future outlook of AI in health care: expert interview study. JMIR AI 2023 Oct 31;2:e47353 [FREE Full text] [doi: 10.2196/47353] [Medline: 38875571]
20. Lorenzini G, Arbelaez Ossa L, Shaw DM, Elger BS. Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision making. Bioethics 2023 Jun 25;37(5):424-429. [doi: 10.1111/bioe.13158] [Medline: 36964989]
21. van Cauwenberge D, van Biesen W, Decruyenaere J, Leune T, Sterckx S. "Many roads lead to Rome and the artificial intelligence only shows me one road": an interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Med Ethics 2022 May 06;23(1):50 [FREE Full text] [doi: 10.1186/s12910-022-00787-8] [Medline: 35524301]
22. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
23. Cheng M, Li X, Xu J. Promoting healthcare workers' adoption intention of artificial-intelligence-assisted diagnosis and treatment: the chain mediation of social influence and human-computer trust. Int J Environ Res Public Health 2022 Oct 15;19(20):36293889 [FREE Full text] [doi: 10.3390/ijerph192013311] [Medline: 36293889]
24. Kleine AK, Kokje E, Lermer E, Gaube S. Attitudes toward the adoption of 2 artificial intelligence-enabled mental health tools among prospective psychotherapists: cross-sectional study. JMIR Hum Factors 2023 Jul 12;10:e46859 [FREE Full text] [doi: 10.2196/46859] [Medline: 37436801]
25. Tamori H, Yamashina H, Mukai M, Morii Y, Suzuki T, Ogasawara K. Acceptance of the use of artificial intelligence in medicine among Japan's doctors and the public: a questionnaire survey. JMIR Hum Factors 2022 Mar 16;9(1):e24680 [FREE Full text] [doi: 10.2196/24680] [Medline: 35293878]
26. Predictors of the intention to use AI-enabled clinical decision support systems among healthcare practitioners: a meta-analysis and narrative synthesis. Open Science Framework. URL: https://osf.io/b4j3t [accessed 2024-04-29]
27. Dwivedi YK, Rana NP, Chen H, Williams MD. A meta-analysis of the unified theory of acceptance and use of technology (UTAUT). In: Proceedings of the 2011 International Working Conference on Governance and Sustainability in Information Systems. Managing the Transfer and Diffusion of IT. 2011 Presented at: TDIT '11; September 22-24, 2011; Hamburg, Germany p. 155-170 URL: https://link.springer.com/chapter/10.1007/978-3-642-24148-2_10 [doi: 10.1007/978-3-642-24148-2_10]

XSL•FO
RenderX

28. Fujimori R, Liu K, Soeno S, Naraba H, Ogura K, Hara K, et al. Acceptance, barriers, and facilitators to implementing artificial intelligence-based decision support systems in emergency departments: quantitative and qualitative evaluation. JMIR Form Res 2022 Jun 13;6(6):e36501 [FREE Full text] [doi: 10.2196/36501] [Medline: 35699995]
29. Laka M, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901 [FREE Full text] [doi: 10.3390/ijerph18041901] [Medline: 33669353]
30. Seliaman ME, Albahly MS. The reasons for physicians and pharmacists' acceptance of clinical support systems in Saudi Arabia. Int J Environ Res Public Health 2023 Feb 10;20(4):3132 [FREE Full text] [doi: 10.3390/ijerph20043132] [Medline: 36833832]
31. Dwivedi YK, Rana NP, Jeyaraj A, Clement M, Williams MD. Re-examining the unified theory of acceptance and use of technology (UTAUT): towards a revised theoretical model. Inf Syst Front 2017 Jun 8;21(3):719-734. [doi: 10.1007/s10796-017-9774-y]
32. Khechine H, Lakhal S, Ndjambou P. A meta‐analysis of the UTAUT model: eleven years later. Can J Adm Sci 2016 Jun 06;33(2):138-152. [doi: 10.1002/cjas.1381]
33. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
34. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
35. Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. Int J Inf Manage 2019 Oct;48:63-71. [doi: 10.1016/j.ijinfomgt.2019.01.021]
36. Fan W, Liu J, Zhu S, Pardalos PM. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Ann Oper Res 2018 Mar 19;294(1-2):567-592. [doi: 10.1007/s10479-018-2818-y]
37. Prakash AV, Das S. Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: a mixed-methods study. Inf Manag 2021 Nov;58(7):103524. [doi: 10.1016/j.im.2021.103524]
38. Panigutti C, Beretta A, Giannotti F, Pedreschi D. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 22 to May 5, 2022; New Orleans, LA p. 1-9. [doi: 10.1145/3491102.3502104]
39. Akinnuwesi BA, Uzoka FM, Fashoto SG, Mbunge E, Odumabo A, Amusa OO, et al. A modified UTAUT model for the acceptance and use of digital technology for tackling COVID-19. Sustain Oper Comput 2022;3:118-135. [doi: 10.1016/j.susoc.2021.12.001]
40. Dwivedi YK, Rana NP, Tamilmani K, Raman R. A meta-analysis based modified unified theory of acceptance and use of technology (meta-UTAUT): a review of emerging literature. Curr Opin Psychol 2020 Dec;36:13-18. [doi: 10.1016/j.copsyc.2020.03.008] [Medline: 32339928]
41. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157-178. [doi: 10.2307/41410412]
42. Rudolph CW, Katz IM, Lavigne KN, Zacher H. Job crafting: a meta-analysis of relationships with individual differences, job characteristics, and work outcomes. J Vocat Behav 2017 Oct;102:112-138. [doi: 10.1016/j.jvb.2017.05.008]
43. Kleine AK, Rudolph CW, Zacher H. Thriving at work: a meta‐analysis. J Organ Behavior 2019 Jun 07;40(9-10):973-999. [doi: 10.1002/job.2375]
44. Fishbein M, Ajzen I. Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research. London, UK: Addison-Wesley Publishing Company; 1975.
45. Dalvi-Esfahani M, Mosharaf-Dehkordi M, Leong LW, Ramayah T, Jamal Kanaan-Jebna AM. Exploring the drivers of XAI-enhanced clinical decision support systems adoption: insights from a stimulus-organism-response perspective. Technol Forecast Soc Change 2023 Oct;195:122768. [doi: 10.1016/j.techfore.2023.122768]
46. Kohnke A, Cole ML, Bush R. Incorporating UTAUT predictors for understanding home care patients' and clinician's acceptance of healthcare telemedicine equipment. J Technol Manag Innov 2014 Jul;9(2):29-41. [doi: 10.4067/S0718-27242014000200003]
47. Ritter C. User-based barriers to the adoption of artificial intelligence in healthcare [Thesis]. Capella University. 2019. URL: https://www.proquest.com/openview/34ede932048ba5f4ec92a141243b5920/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2024-04-29]
48. Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: an investigation of its components and measures. ACM Trans Manage Inf Syst 2011 Jul 01;2(2):1-25. [doi: 10.1145/1985347.1985353]
49. McKnight DH. Trust in information technology. In: Davis GB, editor. The Blackwell Encyclopedia of Management, Volume 7, Management Information Systems, 2nd edition. London, UK: Wiley-Blackwell; 2005:329-331.
50. Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electron Mark 2020 Oct 01;31(2):447-464. [doi: 10.1007/S12525-020-00441-4]

XSL•FO
RenderX

51. Cornelissen L, Egher C, van Beek V, Williamson L, Hommes D. The drivers of acceptance of artificial intelligence-powered care pathways among medical professionals: web-based survey study. JMIR Form Res 2022 Jun 21;6(6):e33368 [FREE Full text] [doi: 10.2196/33368] [Medline: 35727614]
52. Bauer RA. Consumer behavior as risk taking. In: Baker MJ, editor. Marketing: Critical Perspectives on Business and Management. New York, NY: Routledge; 2001.
53. Pan J, Ding S, Wu D, Yang S, Yang J. Exploring behavioural intentions toward smart healthcare services among medical practitioners: a technology transfer perspective. Int J Prod Res 2018 Nov 29;57(18):5801-5820. [doi: 10.1080/00207543.2018.1550272]
54. Calisto FM, Nunes N, Nascimento JC. Modeling adoption of intelligent agents in medical imaging. Int J Hum Comput Interact 2022 Dec;168:102922. [doi: 10.1016/j.ijhcs.2022.102922]
55. Johnson DG, Verdicchio M. AI anxiety. J Assoc Inf Sci Technol 2017 Jun 22;68(9):2267-2270. [doi: 10.1002/asi.23867]
56. Kim J, Kadkol S, Solomon I, Yeh H, Soh JY, Nguyen TM, et al. AI anxiety: a comprehensive analysis of psychological factors and interventions. SSRN Journal Preprint posted online on September 15, 2023 [FREE Full text] [doi: 10.2139/ssrn.4573394]
57. Agarwal R, Prasad J. A conceptual and operational definition of personal innovativeness in the domain of information technology. Inf Syst Res 1998 Jun;9(2):204-215. [doi: 10.1287/isre.9.2.204]
58. Ciftci O, Berezina K, Kang M. Effect of personal innovativeness on technology adoption in hospitality and tourism: meta-analysis. In: Proceedings of the 2021 eTourism Conference on Information and Communication Technologies in Tourism. 2021 Presented at: ENTER '21; January 19-22, 2021; Virtual Event p. 162-174 URL: https://link.springer.com/chapter/10.1007/978-3-030-65785-7_14 [doi: 10.1007/978-3-030-65785-7_14]
59. Lu J, Yao JE, Yu CS. Personal innovativeness, social influences and adoption of wireless internet services via mobile technology. J Strateg Inf Syst 2005 Sep;14(3):245-268. [doi: 10.1016/j.jsis.2005.07.003]
60. Smith H, Downer J, Ives J. Clinicians and AI use: where is the professional guidance? J Med Ethics 2024 Jun 21;50(7):437-441 [FREE Full text] [doi: 10.1136/jme-2022-108831] [Medline: 37607805]
61. Blut M, Chong AY, Tsigna Z, Venkatesh V. Meta-analysis of the unified theory of acceptance and use of technology (UTAUT): challenging its validity and charting a research agenda in the Red Ocean. J Assoc Inf Syst 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
62. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
63. Liu L, Miguel Cruz A, Rios Rincon A, Buttar V, Ranson Q, Goertzen D. What factors determine therapists' acceptance of new technologies for rehabilitation – a study using the unified theory of acceptance and use of technology (UTAUT). Disabil Rehabil 2015;37(5):447-455. [doi: 10.3109/09638288.2014.923529] [Medline: 24901351]
64. Shachak A, Kuziemsky C, Petersen C. Beyond TAM and UTAUT: future directions for HIT implementation research. J Biomed Inform 2019 Dec;100:103315 [FREE Full text] [doi: 10.1016/j.jbi.2019.103315] [Medline: 31629923]
65. Huang KY, Choi N, Chengalur-Smith I. Cultural dimensions as moderators of the UTAUT model: a research proposal in a healthcare context. In: Proceedings of the 2010 Conference on Americas Conference on Information Systems. 2010 Presented at: AMCIS '10; August 12-15, 2010; Lima, Peru URL: https://aisel.aisnet.org/amcis2010/188
66. Lin HC. An investigation of the effects of cultural differences on physicians' perceptions of information technology acceptance as they relate to knowledge management systems. Comput Human Behav 2014 Sep;38:368-380. [doi: 10.1016/j.chb.2014.05.001]
67. Venkatesh V. Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. Inf Syst Res 2000 Dec;11(4):342-365. [doi: 10.1287/isre.11.4.342.11872]
68. Eiskjær S, Pedersen CF, Skov ST, Andersen M. Usability and performance expectancy govern spine surgeons' use of a clinical decision support system for shared decision-making on the choice of treatment of common lumbar degenerative disorders. Front Digit Health 2023 Aug 15;5:1225540 [FREE Full text] [doi: 10.3389/fdgth.2023.1225540] [Medline: 37654781]
69. Hsieh P. Determinants of physicians' intention to use AI-assisted diagnosis: an integrated readiness perspective. Comput Human Behav 2023 Oct;147:107868. [doi: 10.1016/j.chb.2023.107868]
70. Wang W, Chen L, Xiong M, Wang Y. Accelerating AI adoption with responsible AI signals and employee engagement mechanisms in health care. Inf Syst Front 2021 Jun 29;25(6):2239-2256. [doi: 10.1007/s10796-021-10154-4]
71. Yang X, Man D, Yun K, Zhang S, Han X. Factors influencing doctors' acceptance of artificial intelligence-enabled clinical decision support systems in tertiary hospitals in China. Research Square Preprint posted online on October 31, 2023.. [doi: 10.21203/rs.3.rs-3493725/v1]
72. Miao R, Wu Q, Wang Z, Zhang X, Song Y, Zhang H, et al. Factors that influence users' adoption intention of mobile health: a structural equation modeling approach. Int J Prod Res 2017 Jun 12;55(19):5801-5815. [doi: 10.1080/00207543.2017.1336681]
73. Spil TAM, Schuring RW. The UTAUT Questionnaire items. In: Spil TAM, Schuring RW, editors. E-Health Systems Diffusion and Use: The Innovation, the User and the Use IT Model. Hershey, PA: IGI Global; 2006:83-98.

XSL•FO
RenderX

74. Venkatesh V, Bala H. Technology Acceptance Model 3 and a research agenda on interventions. Decision Sciences 2008 May 09;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
75. Alshehri MA. Using the UTAUT model to determine factors affecting acceptance and use of e-government services in the Kingdom of Saudi Arabia [Doctoral dissertation]. Griffith University. 2013. URL: https://doi.org/10.25904/1912/1770 [accessed 2024-07-29]
76. Sohn K, Kwon O. Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. Telematics and Informatics 2020 Apr;47:101324. [doi: 10.1016/j.tele.2019.101324]
77. Ye T, Xue J, He M, Gu J, Lin H, Xu B, et al. Psychosocial factors affecting artificial intelligence adoption in health care in China: cross-sectional study. J Med Internet Res 2019 Oct 17;21(10):e14316 [FREE Full text] [doi: 10.2196/14316] [Medline: 31625950]
78. Rahman MM, Lesch MF, Horrey WJ, Strawderman L. Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. Accid Anal Prev 2017 Nov;108:361-373. [doi: 10.1016/j.aap.2017.09.011] [Medline: 28957759]
79. Wu I, Li J, Fu C. The adoption of mobile healthcare by hospital's professionals: an integrative perspective. Decision Support Systems 2011 Jun;51(3):587-596. [doi: 10.1016/j.dss.2011.03.003]
80. Sheth JN, Newman BI, Gross BL. Why we buy what we buy: a theory of consumption values. J Bus Res 1991 Mar;22(2):159-170. [doi: 10.1016/0148-2963(91)90050-8]
81. Gulati S, Sousa S, Lamas D. Modelling trust in human-like technologies. In: Proceedings of the 9th Indian Conference on Human-Computer Interaction. 2018 Dec 16 Presented at: IndiaHCI '18; December 16-18, 2018; Bangalore, India p. 1-10. [doi: 10.1145/3297121.3297124]
82. Brady GM, Truxillo DM, Bauer TN, Jones MP. The development and validation of the Privacy and Data Security Concerns Scale (PDSCS). Int J Selection Assessment 2020 Sep 29;29(1):100-113. [doi: 10.1111/ijsa.12311]
83. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020 Jul 22;20(1):170 [FREE Full text] [doi: 10.1186/s12911-020-01191-1] [Medline: 32698869]
84. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020 Sep 11;3(1):118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
85. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. “Brilliant AI doctor” in rural clinics: challenges in AI-powered clinical decision support system deployment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan p. 1-18 URL: https://tinyurl.com/4ksdcxz3 [doi: 10.1145/3411764.3445432]
86. Pikkemaat M, Thulesius H, Milos Nymberg V. Swedish primary care physicians' intentions to use telemedicine: a survey using a new questionnaire - physician attitudes and intentions to use telemedicine (PAIT). Int J Gen Med 2021;14:3445-3455 [FREE Full text] [doi: 10.2147/IJGM.S319497] [Medline: 34295177]
87. Alomari A, Soh B. Determinants of medical internet of things adoption in healthcare and the role of demographic factors incorporating modified UTAUT. Int J Adv Comput Sci Appl 2023;14(7):17-31. [doi: 10.14569/ijacsa.2023.0140703]
88. Wichmann JL, Willemink MJ, de Cecco CN. Artificial intelligence and machine learning in radiology: current state and considerations for routine clinical implementation. Invest Radiol 2020 Jul 29;55(9):619-627. [doi: 10.1097/rli.0000000000000673]
89. Transforming and scaling up health professionals’ education and training. World Health Organization. 2013. URL: https://www.ncbi.nlm.nih.gov/books/NBK298953/pdf/Bookshelf_NBK298953.pdf [accessed 2024-01-27]
90. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-269 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
91. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995;123(3):A12-A13. [Medline: 7582737]
92. Alhashmi SF, Alshurideh M, Al Kurdi B, Salloum SA. A systematic review of the factors affecting the artificial intelligence implementation in the health care sector. In: Proceedings of the 2020 International Conference on Artificial Intelligence and Computer Vision. 2020 Presented at: AICV '20; July 10-12, 2020; Chongqing, China p. 37-49 URL: https://link.springer.com/chapter/10.1007/978-3-030-44289-7_4 [doi: 10.1007/978-3-030-44289-7_4]
93. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci 2021 Nov 09;11(22):10537. [doi: 10.3390/app112210537]
94. Chen M, Zhang B, Cai Z, Seery S, Gonzalez MJ, Ali NM, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med (Lausanne) 2022 Aug 31;9:990604 [FREE Full text] [doi: 10.3389/fmed.2022.990604] [Medline: 36117979]
95. Chong AY, Blut M, Zheng S. Factors influencing the acceptance of healthcare information technologies: a meta-analysis. Inf Manag 2022 Apr;59(3):103604. [doi: 10.1016/j.im.2022.103604]
96. Santomartino SM, Yi PH. Systematic review of radiologist and medical student attitudes on the role and impact of AI in radiology. Acad Radiol 2022 Nov;29(11):1748-1756. [doi: 10.1016/j.acra.2021.12.032] [Medline: 35105524]

XSL•FO
RenderX

97. Jiang L, Lavaysse LM. Cognitive and affective job insecurity: a meta-analysis and a primary study. J Manag 2018 May 08;44(6):2307-2342. [doi: 10.1177/0149206318773853]
98. Hoffman BJ, Woehr DJ. A quantitative review of the relationship between person–organization fit and behavioral outcomes. J Vocat Behav 2006 Jun;68(3):389-399. [doi: 10.1016/j.jvb.2005.08.003]
99. Allaire JJ. RStudio: integrated development environment for R. RStudio. URL: http://www.posit.co/ [accessed 2024-04-29]
100. Dahlke JA, Wiernik BM. psychmeta: an R package for psychometric meta-analysis. Appl Psychol Meas 2019 Jul 05;43(5):415-416 [FREE Full text] [doi: 10.1177/0146621618795933] [Medline: 31235986]
101. Cheung MW. metaSEM: an R package for meta-analysis using structural equation modeling. Front Psychol 2014 Jan 05;5:1521 [FREE Full text] [doi: 10.3389/fpsyg.2014.01521] [Medline: 25601849]
102. Schmidt FL, Hunter JE. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. 3rd edition. Thousand Oaks, CA: Sage Publications; 2015.
103. Johnson JW. A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behav Res 2000 Jan;35(1):1-19. [doi: 10.1207/s15327906mbr3501_1]
104. Viswesvaran C, Ones DS. Theory testing: combining psychometric meta‐analysis and structural equations modeling. Pers Psychol 2006 Dec 07;48(4):865-885. [doi: 10.1111/j.1744-6570.1995.tb01784.x]
105. LeBreton JM, Hargis MB, Griepentrog B, Oswald FL, Ployhart RE. A multidimensional approach for evaluating variables in organizational research and practice. Pers Psychol 2007 May 17;60(2):475-498. [doi: 10.1111/j.1744-6570.2007.00080.x]
106. Tonidandel S, LeBreton JM. Relative importance analysis: a useful supplement to regression analysis. J Bus Psychol 2011 Jan 7;26(1):1-9. [doi: 10.1007/S10869-010-9204-3]
107. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta-Analysis. 2nd edition. Hoboken, NJ: John Wiley & Sons; 2021.
108. Hofstede G. Culture's Consequences: International Differences in Work-Related Values. Thousand Oaks, CA: SAGE Publications; 1984.
109. Minkov M, Kaasa A. Do dimensions of culture exist objectively? a validation of the revised Minkov-Hofstede model of culture with world values survey items and scores for 102 countries. J Int Manag 2022 Dec;28(4):100971. [doi: 10.1016/j.intman.2022.100971]
110. Jak S. Meta-Analytic Structural Equation Modelling. Cham, Switzerland: Springer; 2015.
111. Cheung MW. Meta-Analysis: A Structural Equation Modeling Approach. Hoboken, NJ: John Wiley & Sons; 2015.
112. Cheung MW, Chan W. Meta-analytic structural equation modeling: a two-stage approach. Psychol Methods 2005 Mar;10(1):40-64. [doi: 10.1037/1082-989x.10.1.40]
113. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. New York, NY: Routledge; 1988.
114. Whitener EM. Confusion of confidence intervals and credibility intervals in meta-analysis. J Appl Psychol 1990;75(3):315-321. [doi: 10.1037//0021-9010.75.3.315]
115. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell 2022 Mar;5:4. [doi: 10.21037/jmai-21-25]
116. Berge GT, Granmo OC, Tveit TO, Munkvold BE, Ruthjersen AL, Sharma J. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. BMC Med Inform Decis Mak 2023 Jan 10;23(1):5 [FREE Full text] [doi: 10.1186/s12911-023-02101-x] [Medline: 36627624]
117. Evans RP, Bryant LD, Russell G, Absolom K. Trust and acceptability of data-driven clinical recommendations in everyday practice: a scoping review. Int J Med Inform 2024 Mar;183:105342 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105342] [Medline: 38266426]
118. Lambert SI, Madi M, Sopka S, Lenes A, Stange H, Buszello CP, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. NPJ Digit Med 2023 Jun 10;6(1):111 [FREE Full text] [doi: 10.1038/s41746-023-00852-5] [Medline: 37301946]
119. Philippi P, Baumeister H, Apolinário-Hagen J, Ebert DD, Hennemann S, Kott L, et al. Acceptance towards digital health interventions - model validation and further development of the unified theory of acceptance and use of technology. Internet Interv 2021 Dec;26:100459 [FREE Full text] [doi: 10.1016/j.invent.2021.100459] [Medline: 34603973]
120. Liu CF, Chen ZC, Kuo SC, Lin TC. Does AI explainability affect physicians' intention to use AI? Int J Med Inform 2022 Dec;168:104884. [doi: 10.1016/j.ijmedinf.2022.104884] [Medline: 36228415]
121. Kaya F, Aydin F, Schepman A, Rodway P, Yetişensoy O, Demir Kaya M. The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. Int J Hum Comput Interact 2022 Dec 07;40(2):497-514. [doi: 10.1080/10447318.2022.2151730]
122. Huo W, Zheng G, Yan J, Sun L, Han L. Interacting with medical artificial intelligence: integrating self-responsibility attribution, human–computer trust, and personality. Comput Human Behav 2022 Jul;132:107253. [doi: 10.1016/j.chb.2022.107253]
123. Huo W, Yuan X, Li X, Luo W, Xie J, Shi B. Increasing acceptance of medical AI: the role of medical staff participation in AI development. Int J Med Inform 2023 Jul;175:105073 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105073] [Medline: 37119693]

XSL•FO
RenderX

124. Foster TL, Bowie P. Professional values and behaviours of younger and older general practitioners in Scotland: a cross-sectional survey. leader 2020 Feb 19;4(2):57-63. [doi: 10.1136/leader-2019-000138]
125. Jena RK. Exploring antecedents of peoples' intentions to use smart services in a smart city environment: an extended UTAUT model. J Inf Syst 2022;36(1):133-149. [doi: 10.2308/isys-2020-050]
126. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021 Feb 19;4(1):31 [FREE Full text] [doi: 10.1038/s41746-021-00385-9] [Medline: 33608629]
127. Kleine AK, Lermer E, Cecil J, Heinrich A, Gaube S. Advancing mental health care with AI-enabled precision psychiatry tools: a patent review. Comput Hum Behav Rep 2023 Dec;12:100322. [doi: 10.1016/j.chbr.2023.100322]
128. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195-e203. [doi: 10.1016/s2589-7500(20)30292-2]
129. Hlávka JP. Security, privacy, and information-sharing aspects of healthcare artificial intelligence. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Cambridge, MA: Academic Press; 2020:235-270.
130. Lukyanenko R, Maass W, Storey VC. Trust in artificial intelligence: from a foundational trust framework to emerging research opportunities. Electron Mark 2022 Nov 28;32(4):1993-2020. [doi: 10.1007/s12525-022-00605-4]
131. Yang W, Wei Y, Wei H, Chen Y, Huang G, Li X, et al. Survey on explainable AI: from approaches, limitations and applications aspects. Hum Cent Intell Syst 2023 Aug 10;3(3):161-188. [doi: 10.1007/s44230-023-00038-y]
132. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
133. Goodman KE, Morgan DJ, Hoffmann DE. Clinical algorithms, antidiscrimination laws, and medical device regulation. JAMA 2023 Jan 24;329(4):285-286. [doi: 10.1001/jama.2022.23870] [Medline: 36602795]
134. Neff M, Schaaf J, Noll R, Holtz S, Schütze D, Köhler SM, et al. Initial user-centred design of an AI-based clinical decision support system for primary care. Stud Health Technol Inform 2024 Jan 25;310:1051-1055. [doi: 10.3233/SHTI231125] [Medline: 38269975]
135. Beltrão G, Paramonova I, Sousa S. User interface design for AI-based clinical decision-support system. Zenodo. URL: https://zenodo.org/records/8369358 [accessed 2024-04-29]
136. Wang W, Wang Y, Chen L, Ma R, Zhang M. Justice at the forefront: cultivating felt accountability towards artificial intelligence among healthcare professionals. Soc Sci Med 2024 Apr;347:116717 [FREE Full text] [doi: 10.1016/j.socscimed.2024.116717] [Medline: 38518481]
137. Kwak Y, Ahn JW, Seo YH. Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions. BMC Nurs 2022 Sep 30;21(1):267 [FREE Full text] [doi: 10.1186/s12912-022-01048-0] [Medline: 36180902]
138. Vallo Hult H, Hansson A, Gellerstedt M. Digitalization and physician learning: individual practice, organizational context, and social norm. J Contin Educ Health Prof 2020;40(4):220-227 [FREE Full text] [doi: 10.1097/CEH.0000000000000303] [Medline: 33284172]
139. Thakkar B, Bharathi V. Medical specialists' perception about adoption of artificial intelligence in the healthcare sector. Cardiometry 2023 Feb 14(25):426-434. [doi: 10.18137/cardiometry.2022.25.426434]
140. Bayaga A, du Plessis A. Ramifications of the unified theory of acceptance and use of technology (UTAUT) among developing countries' higher education staffs. Educ Inf Technol 2023 Sep 19;29:9689-9714. [doi: 10.1007/S10639-023-12194-6]
141. Novak LL, Anders S, Gadd CS, Lorenzi NM. Mediation of adoption and use: a key strategy for mitigating unintended consequences of health IT implementation. J Am Med Inform Assoc 2012;19(6):1043-1049 [FREE Full text] [doi: 10.1136/amiajnl-2011-000575] [Medline: 22634157]
142. Shachak A, Montgomery C, Dow R, Barnsley J, Tu K, Jadad AR, et al. End-user support for primary care electronic medical records: a qualitative case study of users' needs, expectations and realities. Health Syst (Basingstoke) 2013 Nov 01;2(3):198-212 [FREE Full text] [doi: 10.1057/hs.2013.6] [Medline: 26225209]
143. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
144. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Analysing the role of complexity in explaining the fortunes of technology programmes: empirical application of the NASSS framework. BMC Med 2018 May 14;16(1):66 [FREE Full text] [doi: 10.1186/s12916-018-1050-6] [Medline: 29754584]
145. Abdekhoda M, Dehnad A, Zarei J. Factors influencing adoption of e-learning in healthcare: integration of UTAUT and TTF model. BMC Med Inform Decis Mak 2022 Dec 09;22(1):327 [FREE Full text] [doi: 10.1186/s12911-022-02060-9] [Medline: 36494800]
146. Li J, Huang JS. Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. Technol Soc 2020 Nov;63:101410. [doi: 10.1016/j.techsoc.2020.101410]
147. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. Med Law Rev 2023 Nov 27;31(4):501-520 [FREE Full text] [doi: 10.1093/medlaw/fwad013] [Medline: 37218368]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
AI-CDSS: artificial intelligence–enabled clinical decision support system
CR: credibility interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by Y Zhuang; submitted 15.02.24; peer-reviewed by A Hassan, GK Gupta, M Wekenborg, I Ali; comments to author 12.03.24; revised version received 03.05.24; accepted 13.05.24; published 05.08.24.

Please cite as:
Dingel J, Kleine AK, Cecil J, Sigl AL, Lermer E, Gaube S
Predictors of Health Care Practitioners' Intention to Use AI-Enabled Clinical Decision Support Systems: Meta-Analysis Based on the Unified Theory of Acceptance and Use of Technology
J Med Internet Res 2024;26:e57224
URL: https://www.jmir.org/2024/1/e57224
doi:10.2196/57224
PMID:

©Julius Dingel, Anne-Kathrin Kleine, Julia Cecil, Anna Leonie Sigl, Eva Lermer, Susanne Gaube. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Longitudinal Coadaptation of Older Adults With Wearables and Voice-Activated Virtual Assistants: Scoping Review

Kristina Marie Kokorelias[1], PhD; Alisa Grigorovich[2,3], PhD; Maurita T Harris[4], PhD; Umair Rehman[5], PhD; Louise Ritchie[6], PhD; AnneMarie M Levy[7], PhD; Kerstin Denecke[8], PhD; Josephine McMurray[7], PhD

[1]Division of Geriatric Medicine, Department of Medicine, Sinai Health System and University Health Network, Toronto, ON, Canada
[2]KITE Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada
[3]Recreation and Leisure Studies, Brock University, St Catherines, ON, Canada
[4]User Experience Design, Wilfrid Laurier University, Waterloo, ON, Canada
[5]Department of Computer Science, University of Western Ontario, London, ON, Canada
[6]Alzheimer Scotland Centre for Policy and Practice, University of West Scotland, Scotland, United Kingdom
[7]Lazaridis School of Business and Economics/Community Health, Wilfrid Laurier University, Brantford, ON, Canada
[8]Institute for Medical Informatics, Bern University of Applied Sciences, Bern, Switzerland

Corresponding Author:
Josephine McMurray, PhD
Lazaridis School of Business and Economics/Community Health
Wilfrid Laurier University
73 George St
Brantford, ON, N3T3Y3
Canada
Phone: 1 548 889 4492
Email: jmcmurray@wlu.ca

Abstract

Background: The integration of smart technologies, including wearables and voice-activated devices, is increasingly recognized for enhancing the independence and well-being of older adults. However, the long-term dynamics of their use and the coadaptation process with older adults remain poorly understood. This scoping review explores how interactions between older adults and smart technologies evolve over time to improve both user experience and technology utility.

Objective: This review synthesizes existing research on the coadaptation between older adults and smart technologies, focusing on longitudinal changes in use patterns, the effectiveness of technological adaptations, and the implications for future technology development and deployment to improve user experiences.

Methods: Following the Joanna Briggs Institute Reviewer's Manual and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines, this scoping review examined peer-reviewed papers from databases including Ovid MEDLINE, Ovid Embase, PEDro, Ovid PsycINFO, and EBSCO CINAHL from the year 2000 to August 28, 2023, and included forward and backward searches. The search was updated on March 1, 2024. Empirical studies were included if they involved (1) individuals aged 55 years or older living independently and (2) focused on interactions and adaptations between older adults and wearables and voice-activated virtual assistants in interventions for a minimum period of 8 weeks. Data extraction was informed by the selection and optimization with compensation framework and the sex- and gender-based analysis plus theoretical framework and used a directed content analysis approach.

Results: The search yielded 16,143 papers. Following title and abstract screening and a full-text review, 5 papers met the inclusion criteria. Study populations were mostly female participants and aged 73-83 years from the United States and engaged with voice-activated virtual assistants accessed through smart speakers and wearables. Users frequently used simple commands related to music and weather, integrating devices into daily routines. However, communication barriers often led to frustration due to devices' inability to recognize cues or provide personalized responses. The findings suggest that while older adults can integrate smart technologies into their lives, a lack of customization and user-friendly interfaces hinder long-term adoption and satisfaction. The studies highlight the need for technology to be further developed so they can better meet this demographic's evolving needs and call for research addressing small sample sizes and limited diversity.

XSL•FO
RenderX

Conclusions: Our findings highlight a critical need for continued research into the dynamic and reciprocal relationship between smart technologies and older adults over time. Future studies should focus on more diverse populations and extend monitoring periods to provide deeper insights into the coadaptation process. Insights gained from this review are vital for informing the development of more intuitive, user-centric smart technology solutions to better support the aging population in maintaining independence and enhancing their quality of life.

International Registered Report Identifier (IRRID): RR2-10.2196/51129

(J Med Internet Res 2024;26:e57258) doi:10.2196/57258

KEYWORDS

older adults; coadaptation; voice recognition; virtual assistant; wearable; artificial intelligence; smart-assistive technology; scoping review; review methods; review methodology; knowledge synthesis; synthesis; scoping; older adult; gerontechnology; technology; smart technology; smart technologies; smart; geriatrics; older people; geriatric; scoping literature review; protocol; Internet of Things; IoT; aging; PRISMA-ScR; Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; user-centered design; design; user centered; mobile phone

Introduction

Background

Technology has revolutionized our lifestyles from communication methods and social interactions to mobility and self-care. The integration of artificial intelligence (AI) into household devices, such as built-in voice-activated virtual assistants, epitomizes the “Internet of Things” (IoT) concept, where devices are interconnected over the internet allowing for remote monitoring and control [1]. Gigli and Koo [2] described the IoT concept as an “ongoing movement to consolidate all resources globally into a shared infrastructure.” This technological integration is particularly significant for individuals aged 60 years and older, whose numbers are expected to rise significantly in the coming decades [3]. Ambient-assisted living (AAL) technologies, which form a core part of IoT applications tailored to personal care, enhance the quality of life and support the independence of older adults within their living environments [4,5]. By leveraging IoT capabilities, AAL technologies can provide continuous assistance and monitoring, thereby enhancing the safety, independence, and well-being of older adults and supporting the concept of aging in place [4,5].

Aging in place is crucial for enabling older adults to stay socially connected to the places where they live and feel comfortable and competent [6-9], thus potentially reduce the demand for institutional care, alleviating the burden on health care systems [10-14]. Smart technologies facilitate access to the benefits of the internet and video calls for social interaction and access to care and are also equipped with sensors and automation systems [15] that enhance self-management of chronic conditions and adapt living spaces to accommodate the changing needs of older adults [16-20]. For example, such technologies can enable remote or automated control of various features in the home, including lighting, temperature, and use of multimedia, as well as provide automated fall detection, all of which can mitigate potential hazards and accidents in the home [21,22]. These technologies are believed to contribute to high-quality care standards in the care of older adults [23,24].

While the benefits and potential applications of IoT technologies to support older adults aging in place are well known [25,26], particularly in areas such as health management and emergency readiness [27], there is a gap in the scientific literature regarding the long-term interaction (eg, beyond the first few weeks of use) between smart technologies and this demographic [15]. Cross-sectional studies have provided valuable insights into older adults’ intentions to use smart technologies [28] and some initial aspects of coadaptation [29]. However, these studies typically capture a single point in time and fail to reflect the evolving nature of use and adaptation. Consequently, there is a scarcity of research that systematically explores the dynamic, ongoing process of coadaptation where behaviors or actions of older adults change beyond becoming familiar with the technology, and smart technologies using automated, algorithm-driven adaptations adjust to user behaviors and preferences in real time [30] and continuously adjust to each other over extended periods [31].

Ongoing interaction is essential to fully understand the potential and challenges of smart technologies in real-world settings. This need for continuous engagement, a process referred to as coadaptation, has been noted as a limitation in previous reviews [28]. Effective integration of technologies relies not only on their technical capabilities but also on a robust process of coadaptation, the dynamic and reciprocal adjustments between older adults and smart technologies that enhance functionality, usability, and overall user experience [32]. This ongoing interaction ensures that the technologies evolve to meet the specific needs of older adults while the users adapt their behaviors to maximize the benefits of the technology. For example, voice-activated devices may adapt to recognize and understand the speech patterns of their users over time, improving interaction quality [33]. Such interactions ensure that technologies evolve to meet the specific needs of older adults, allowing users to optimize use and the benefits of the technology. However, this is an emerging domain as AI is increasingly incorporated into everyday consumer products. The gap we have identified underscores the need for ongoing research to understand and improve how technologies can better adapt to user preferences over time. This is crucial for improving the independence, user experience, and sustained technology adoption of older adults.

XSL•FO
RenderX

Objectives of the Scoping Review

This scoping review addresses the research gap by examining existing literature on the interaction between smart technologies and older adults over time. It focuses on wearables and voice-activated virtual assistants, exploring how these interactions adapt to enhance user experiences and the benefits derived from these technologies. We chose to focus on wearables and voice-activated devices as they are some of the most common smart technologies used by older adults, they are the key demographic in our study, and they offer convenience and accessibility by often interacting with other smart home technologies [34-36]. This study aims to address several critical questions regarding the nature of coadaptation between smart technologies and older adults, the outcomes of this process, strategies used by older adults to adapt, and the methodological approaches in the existing literature. Specifically, this review seeks to address the following questions:

1. What is the extent and nature of the existing scientific literature exploring the coadaptation between smart technologies and older adults, and how do older adults and technology coevolve over time to enhance older adults' experience with the technology?
2. What have been the outcomes of the coadaption between older adults and smart technology?
   - What specific outcome measures have been used in studies investigating the experiences of older adults in terms of coadapting with technology over time?
3. What specific strategies or approaches have participants used to adapt to the changing circumstances, challenges, or opportunities in their environment, and what process do they follow when making adaptive decisions?
4. What are the key characteristics of older adult participants who have been involved in studies examining the coadaptation between smart technologies and older adults?
5. What methodological strengths, limitations, and recommendations have been documented in the literature regarding the exploration of coadaptation between smart technologies and older adults?
6. What research models and theories contribute to the conceptualization of coadaptation?

By answering these questions, this study aims to support further research on tailoring AI-driven technologies to the specific needs of older adults, developing more user-friendly technologies, easing caregiver burden, and formulating policies promoting independent living.

Methods

Design

To address the diverse range of research questions, we used a scoping review methodology recommended for studies aiming to comprehensively map the literature on complex or heterogeneous topics [37]. This scoping review followed the procedures outlined in the Joanna Briggs Institute manual for conducting scoping reviews [38] and reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Multimedia Appendix 1) [39]. Key deviations from the initially registered protocol [15] include changes in the inclusion criteria related to the duration of technology use and expansion of the search terms and databases to ensure comprehensive literature coverage. These adjustments were necessary to address the rapid evolution of technology and its adoption, ensuring the relevance and comprehensiveness of our review.

Eligibility Criteria

We included peer-reviewed empirical studies, dissertations, and conference proceedings that were published in English since 2000 to include the most recent studies and technologies pertinent to the topic under review [15]. The year 2000 was chosen as a starting point for the literature search to focus on the most recent 2 decades of advancements and their application in supporting older adults, ensuring the review captures relevant and contemporary studies. The turn of the millennium saw significant advancements in computing technology, such as faster processors, increased memory capacity, and improved capabilities [40,41]. Studies had to include individuals aged 55 years or older living independently in community-based settings (not institutional care) and explore the interactions, coadaptation, experiences, and outcomes among older adults and smart technologies, specifically wearables and voice-activated virtual assistants. The coadaptive technology had to be automated, algorithm-driven, and adjusting to user behaviors and preferences in real time to be included in the study—a criterion that we believed accommodated the diverse needs of the older adult users who are more likely to face challenges associated with lower technical proficiency [42] and declining physical and cognitive functions [43]. In line with other reviews on technology and older adults, we modified the age threshold for older adults to 55 years or older to ensure inclusivity and comprehensiveness to capture early experiences of aging and technological interaction, thereby providing a comprehensive view of the aging spectrum relevant to technology use [44].

Our review highlights aspects such as quality of life, well-being, social connectedness, independence, and overall user experience over a minimum average period of 8 (SD 2) weeks or more . Based on our preliminary analysis and expert consultations, a minimum duration of an average of >8 (SD 2) weeks was identified as reasonable for observing meaningful coadaptation between older adults and smart technologies. This duration is considered sufficient to allow for user familiarization, routine integration, and initial feedback cycles essential for coadaptive processes to become apparent and is supported by literature on behavioral habituation and technology adoption, which suggest that several weeks (more than a month) are required for users to adapt to and integrate new technologies into their daily lives [31,45,46]. We excluded studies that focused on other types of technologies or forms of literature.

Three reviewers (KMK, AG, and JM) further refined the a priori eligibility criteria. In line with the iterative nature of conducting scoping reviews [47], the original minimum study duration requirement of 6 months or more [15] was adjusted to an average of >8 weeks to accommodate the limited volume of relevant literature. We had originally selected a 6-month duration based

XSL•FO
RenderX

on our knowledge of at least 1 paper that conducted a year-long study, considering 6 months to be a generous timeframe; this paper is included in our review.

Search Strategy

The primary and senior authors (KMK and JM), in collaboration with an experienced medical information specialist, iteratively developed search strategies with feedback from all authors (Multimedia Appendix 2). The keywords for the review included age-related terms ("aged," "seniors," "older adults," "elderly," "people with disabilities," and "cognitive and physical disabilities"), technological terms ("wearable electronic devices," "voice recognition," "artificial intelligence," "smart technology," "smart assistive technology," "virtual reality," and "ambient intelligence"), and interaction-related terms ("coadaptation," "double-loop learning," "coevolution," "human-computer interface," "human-environment interaction," "interaction design," "personalization," "customization," "tinkering," "crafting," "redesign," "modification," "sensemaking," "hacking," "accommodation," "mutual adaptation," "symbiotic evolution," "reciprocal adjustment," "interactive iteration," "user-technology synergy," "dynamic interface refinement," "adaptive coevolution,"" concurrent learning," and "bidirectional refinement").

Peer review of the Ovid MEDLINE search strategy was conducted using the PRESS (Peer Review of Electronic Search Strategies) checklist by an information specialist from another institution [48]. Adjustments in vocabulary and syntax were then applied across multiple electronic databases (Ovid Embase, PEDro, Ovid PsycINFO, EBSCO CINAHL, IEEE Xplore, Web of Science, the Cochrane Library, Scopus, and Global Index Medicus). The initial search was carried out on August 28, 2023, aimed at identifying English-language publications from 2000 onward, complemented by scanning reference lists, forward and backward searching, and seeking expert recommendations to ensure comprehensive coverage. EndNote (version 9.3.3; Clarivate Analytics) facilitated the download and deduplication of retrieved records [49]. The search was updated by the same information specialist on March 1, 2024, but no additional papers were included.

Study Selection

Deduplicated records were imported into Covidence software (Veritas Health Innovation) [50]. Titles and abstracts were screened by at least 2 reviewers from the full research team followed by a full-text screening. In our scoping review, we applied a hierarchical exclusion process prioritizing the wrong study design (eg, commentaries and reviews), incorrect technology, a duration of less than 8 weeks, and the absence of coadaptation. This meant some papers were excluded for multiple reasons. These exclusions were reported from highest to lowest frequency in our PRISMA-ScR flow diagram. Discrepancies during study selection were resolved through discussions among the research team, ensuring that decisions were made collectively and based on a consensus, which adds to the credibility and reliability of the selection process. One reviewer (KMK) conducted the hand screening consisting of the forward and backward screening and the review of reference lists. All team members reviewed potentially included papers to confirm the selection.

Data Extraction and Analysis

Using Covidence, a data extraction form was iteratively developed and pilot-tested by reviewers (KMK, AG, and JM), facilitating the extraction of study characteristics; participant demographics such as sex and gender; underlying research models, theories, and frameworks; study findings; implications for the technology coadaptation process; and any identified study quality or limitations. Data were extracted by 1 team member (KMK), with uncertainties discussed and resolved among the research team.

Data analysis involved a descriptive review of included papers detailing study characteristics, followed by a content analysis by 3 reviewers (KMK, AG, and JM), with a specific focus on changes in user behavior patterns over time, interaction modifications, and instances where both users and technologies engaged in reciprocal feedback loops resulting in iterative adaptations.

Results

Overview

The database literature searches yielded 16,143 unique records that were screened at the title and abstract phase. A total of 94 papers were screened at the full-text stage, and a total of 5 papers were included. To improve transparency and to provide insights into our rigorous selection process, Multimedia Appendix 3 summarizes the primary reasons for excluding the 89 papers that had an insufficient focus on the interactive and mutual adaptation between older adults and technologies. This table aims to further clarify our stringent criteria for coadaptation, which necessitates observing changes and adjustments over time that significantly impact user experience and technology efficacy.

The finding that only 5 papers met the inclusion criteria highlights a significant gap in research on the long-term coadaptation between older adults and smart technologies, underscoring the need for further investigation in this emerging field. One of these papers was found during the hand search. Figure 1 [51] presents the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the paper selection process.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram.

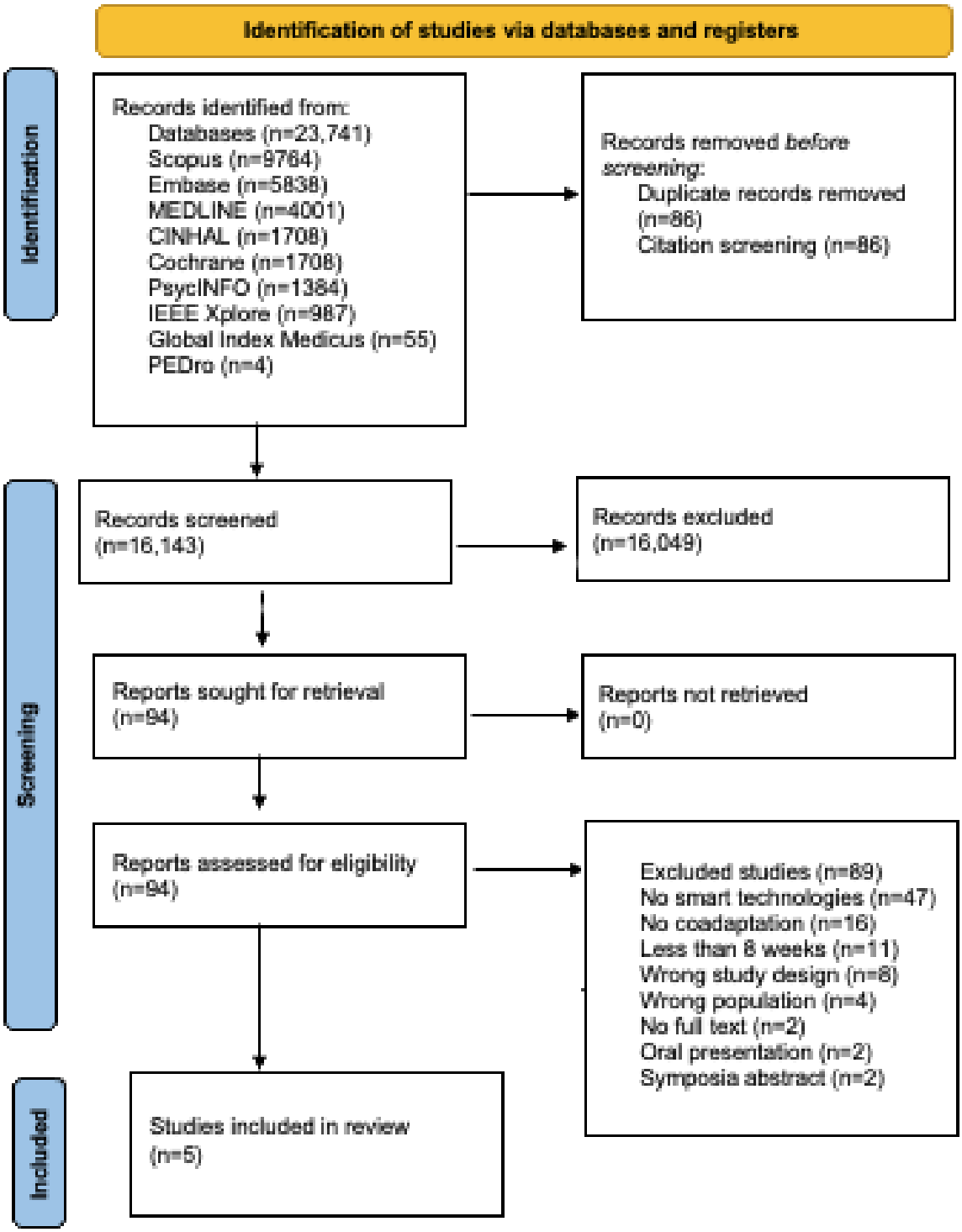

Main Characteristics of the Included Papers

All studies were published between 2021 and 2023. In total, 2 studies were mixed methods [52,53], 2 were qualitative [54,55], and 1 was quasi-experimental [56]. The qualitative data collection methods in these papers included observations [52,55] and interviews [52-54] to understand the experiences, challenges, and interactions of older adults with conversational agents (CAs) and virtual assistants. Three studies kept interaction logs to capture the use of participants [53], including interactions initiated by users through voice commands [52,54]. Another study used a quantitative survey at recruitment and after 2 weeks of use [53].

All but 1 study were conducted in the United States; the remaining study was conducted in the United Kingdom [53]. Three studies involved data collection in people's homes [52,54,55]. Two studies involved data collected at health care

XSL•FO
RenderX

facilities [53,56]. Table 1 presents an overview of the included papers.

XSL•FO
RenderX

Table 1. Characteristics of included peer-reviewed studies.

Author (year), country, and design	Objective	Methods for data collection and number of participants	Length of intervention	Results	Limitations	Recommendations
Cuadra et al (2023) [55], United States, and qualitative	The objective of this study was to explore the experiences and perceptions of older adults who have become users of voice-activated technologies, specifically through the use of a VFAI[a]. The study focuses on understanding how these individuals interact with the VFAI in various contexts, including health data reporting and positive reminiscing, and how their lived experiences influence their use patterns and perceptions of the technology.	• The researchers conducted interviews with 5 participants, reviewed use logs, and provided support to participants to closely observe their interactions with the VFAI over a period of time.	5.9-12.6 weeks (average 8.64 weeks)	The interactions with design probes for health data reporting and positive reminiscing were generally positive, with participants appreciating the value suggested by the probes. However, there were challenges with expectations regarding Alexa's ability to interpret open-ended dialogue and support ambiguity. Participants felt that Alexa did not judge them and provided a safe space for interaction. Overall, participants perceived Alexa as a companion and found the display helpful for voice-first interactions. Challenges included skepticism about health data reporting, hesitance before using the probes, and usability issues such as communication breakdowns. Addressing these challenges could enhance the user experience and increase the value of VFAIs for supporting aging in place.	The study has limitations including nonfunctional design probes, a small sample size of 5 participants in an urban US setting, and the availability of technical support, which may not reflect real-world scenarios for older adults. Additionally, participants were relatively healthy, potentially limiting generalizability, and privacy risks associated with VFAI use require careful consideration in future research.	The study recommends leveraging virtual personal assistants (VFAI) to enhance continuity of care for older adults in their homes. VFAIs can support management continuity by providing consistent health management and reminders for tasks like medication intake and exercise. They can facilitate relational continuity by fostering ongoing therapeutic relationships with users, offering emotional support, and maintaining a nonjudgmental and always available presence. Additionally, VFAIs can promote informational continuity by collecting and using robust information about users' preferences, health history, and context to tailor interactions and support personalized care plans effectively.

XSL•FO
RenderX

Author (year), country, and design	Objective	Methods for data collection and number of participants	Length of intervention	Results	Limitations	Recommendations
Gvozdanovi et al (2022) [53], United Kingdom, and mixed methods	The objective was to assess the feasibility of integrating Vinehealth into brain tumor care, evaluating its capacity to gather real-world and ePRO[b] data and measure subjective improvements in care.	• The study adopted a mixed methodology IDEAL[c] stage 1 design and was conducted at a single tertiary care center. In total, 6 patients were enrolled in the study, of which 4 actively participated by downloading and engaging with the mHealth[d] app throughout the 12-week duration. The primary focus was on collecting real-world and ePRO data through Vinehealth over the study period. Additionally, qualitative feedback was gathered through mixed methodology surveys and semistructured interviews conducted at recruitment and after 2 weeks. This approach allowed for a comprehensive assessment of the feasibility of integrating Vinehealth into brain tumor care as well as evaluating its effectiveness in gathering relevant data and subjective improvements in patient care.	12 weeks	A total of 565 data points were recorded, encompassing various aspects such as symptoms, activity levels, well-being, and medication use. Technical difficulties affected the completion rates of the EORTC[e] QLQ-BN20 and EQ-5D-5L surveys, with rates of 54% and 46%, respectively. However, when ePROs were used, completion rates reached 100%. Participants expressed a desire for more content specific to brain cancer tumors. Despite challenges, all participants endorsed the app and perceived it as enhancing their care.	First, being a proof-of-concept study (IDEAL stage 1), the small sample size precluded the inclusion of a comparator group. Future stages (2 and 3) as per the IDEAL framework would necessitate a larger sample size for formal outcome assessment. While the aim was to onboard all patients before surgery, logistical challenges during the COVID-19 pandemic disrupted this plan. The withdrawal of 2 of 6 participants and the pragmatic recruitment approach highlight the need for a more robust recruitment strategy and a better understanding of withdrawal reasons in future studies. Technical issues and retention rates limited the availability of ePRO data, which could be addressed in subsequent iterations through improved technology and longer follow-up periods with a larger cohort. Additionally, future studies should involve the multidisciplinary team, facilitated by an HCP[f] dashboard for real-time data review and collection of HCP feedback.	__[g]
Kim and Choudhury (2021) [54], United States, and qualitative	This study aims to gather empirical evidence on how the experiences of older adults with a virtual assistant change over time, from novice to experienced users.		16 weeks			

XSL•FO
RenderX

Author (year), country, and design	Objective	Methods for data collection and number of participants	Length of intervention	Results	Limitations	Recommendations
		• This study used a 3-phase interview protocol to investigate older adults' perceptions, challenges, coping strategies, and use patterns with a virtual assistant over a 16-week period. Participants were introduced to a Google Home mini during the first interview, followed by 8 biweekly follow-up interviews. The study also collected device use logs from Google's activity history repository to complement participants' perceptions with actual interaction patterns. • 12 participants		Participants used the virtual assistant primarily for tasks such as playing music, searching for information, making casual conversations, and checking the time and weather. Despite initial challenges, including unfamiliarity with the device and functional errors, participants gradually developed competence and resilience in interacting with the voice-activated virtual assistant. Over time, they appreciated the simplicity and ease of use, convenience of operating without physical interaction, and benefits such as not worrying about making mistakes and building digital companionship.	The study's limitations include potential biases due to convenience sampling from older adult–living communities, the exclusion of older adults with hearing impairments, the lack of comparative studies across different age groups, and the inability to analyze use patterns quantitatively due to data collection through a single Google account.	The recommendations include enriching the conversational capabilities of virtual assistants to enhance user experiences, supporting a learning phase for novice users, and revisiting form factors to align with user expectations and affordances, particularly for older adults who may experience difficulty with new technologies.
Kim (2023) [56], United States, and quantitative	The objective was to assess the impact of integrating information and communication technologies into the health care management of older individuals aged 65 years or older by providing them with the "Health Today" app and a smart speaker for 6 months, measuring changes in depression, self-efficacy, moderate-intensity exercise frequency, relative grip strength, balance, and 5-times-sit-to-stand performance through preintervention and postintervention evaluations.	• 166 participants, aged 65 years or older and enrolled at public health centers, were given access to the "Health Today" app and a smart speaker for 6 months to complete health care tasks. The study divided participants into 2 groups: one group received both the app and speaker, while the other only used the app. Assessments of depression, self-efficacy, frequency of moderate-intensity exercise, grip strength, balance, and 5-times-sit-to-stand performance were conducted before and after the 6-month program.	6 months	Both cohorts demonstrated improved health status and behavioral shifts during the postintervention assessment. Nevertheless, the group given smart speakers, alongside the health management app did not exhibit decreased depression, potentially attributable to the communication and music-listening features.	The study lacked control over attitudes and use patterns toward smart speakers. Follow-up studies should aim to identify and control use patterns for digital devices. Given that prior research predominantly focused on middle-aged individuals, additional studies investigating the effects of mHealth on older adults are warranted. Variables, such as grip strength and balance tests recognized as reliable indicators of frailty, should be taken into account.	—
			3×10 weeks			

XSL•FO
RenderX

Author (year), country, and design	Objective	Methods for data collection and number of participants	Length of intervention	Results	Limitations	Recommendations
Zubatiy et al (2023) [52], United States, and mixed methods	The objective of this research study is to investigate how older adults with MCI[h] and their caregivers use CAs[i] in their homes over time, with a focus on understanding the limitations of existing CAs and identifying requirements for future systems to address these limitations, without externally imposed goals or tasks.	• The study conducted three 10-week deployments of Google Nest Hub devices in the homes of participants, conducting semistructured interviews to understand the impact and use of the devices. User attrition occurred during the deployments, and interaction logs were collected using Google Takeout, allowing for quantitative analysis of user interactions categorized by type and difficulty levels. The goal was to capture perspectives on the usefulness, entertainment value, and frustrations of using the devices in day-to-day life without imposing interaction targets on participants. • 26 older adults and their caregivers		While both care partners and individuals with MCI interacted with the device, there was a consistent increase in use over time, particularly in the third deployment. However, challenges such as multiuser interactions, lack of personalization, and inflexibility in error recovery were identified as persistent themes across all deployments, impacting the overall user experience and effectiveness of the devices.	Limitations include the reliance on a patient population with MCI and their cognitively normal aging spouses, the lack of a true healthy control group, the imperfect metrics used to measure MCI, the limited demographic diversity of the participants, variations in training protocols, and the small sample size, necessitating further research to generalize the findings to broader populations and validate the observed trends.	The study recommendations include the need for future CA systems to prioritize flexibility over efficiency in interactions, encouraging more robust error handling and facilitating user-friendly error recovery mechanisms to minimize frustration and optimize user experience. Additionally, there is a call for enhancing personalization while balancing privacy concerns, suggesting that CA systems should learn and adapt to user preferences over time while respecting user privacy and autonomy. Finally, the study advocates for a shift toward more proactive interactions in CA systems, particularly in multiuser scenarios, to reduce the cognitive burden on users, improve accessibility, and enhance overall usability.

[a]VFAI: voice-first ambient interface.

[b]ePRO: electronic patient-recorded outcome.

[c]IDEAL: Idea, Developments, Exploration, Assessment and Long-term follow-up

[d]mHealth: mobile health.

[e]EORTC: European Organization for Research and Treatment of Cancer.

[f]HCP: health care professional.

[g]Not available.

[h]MCI: mild cognitive impairment.

[i]CA: conversational agent.

Technology Use and Outcomes

Three studies involved older adults using a voice-activated virtual assistant on either a Google or Amazon smart speaker [52,54,55]. Another study used an unnamed smart speaker [56]. One study involved wearables [53]. Two studies involved health apps [53,56,57]. Study intervention periods ranged from 5.9 [55] to 16 weeks [54]. Across all the studies, the use of the technology was consistently maintained [52]. The majority of interactions involved simple user-initiated cues to the technology [52,53], and the mean use frequency averaged 1.8 times daily [54]. The most frequent use of the voice-activated virtual assistant involved playing music, general information searches (eg, weather), casual conversations, and setting reminders or alarms [54]. One study also noted that use patterns revealed use peaks at 9 PM and 7 AM, aligning with bedtime and waking hours [54]. A daytime surge between 1 and 4 PM lacked a defined use pattern, potentially indicating leisure time between meals [54]. Participants tended to value the simplicity and ease of interaction of voice-activated technologies, particularly the fact that they resulted in hands-free convenience [54-56]. In 1 study using wearables, participants appreciated the seamless integration of apps with Google Fit and Apple Health, monitoring vital signs including heart rate and blood pressure, with particular attention to activity levels due to perioperative fatigue in individuals with brain cancer [53].

Reported Characteristics of Study Participants

Sample sizes ranged from 5 [54,55] to 166 participants [56]. All but 1 study reported the average age of participants. One

XSL•FO
RenderX

study reported a range of 45-69 years [53]. The average age of the older adult participants ranged from 73 [55] to 83.8 years [54]. Four (80%) of the 5 studies included mostly female participants [53-56], while another included an equal number of both male and female participants [52]. None of the studies included gender-diverse people (ie, nonbinary).

In regard to health status, one study targeted older adults diagnosed with mild cognitive impairment (MCI; n=13 male and n=13 female) along with their 26 respective care partners (n=10 male, n=16 female; average age 65.7 years) [52]. Caregivers were a spouse or partner or adult child [52]. Diagnosis of MCI had to be validated by a neurologist through a standard set of neuropsychological tests, including the Montreal Cognitive Assessment scores [52]. The focus was on understanding their interactions with CAs in caregiving networks. Another study targeted individuals with brain cancer [53]. In another study with 12 participants, 2 (17%) participants were reported as wearing a hearing aid, and 3 (25%) participants used a wheelchair due to various joint issues [54].

Technology access and comfort were reported in all of the studies. In 1 study, nearly all participants had previously engaged with a CA at some point, although not regularly [52]. Two studies recruited novice users of voice-activated virtual assistants [54,55]. One of these 2 studies indicated that participants had to have no prior experience with a voice-activated virtual assistant [54]. Two studies noted that participants had some familiarity with computers, tablets, and smartphones [54,55]. One study did not comment on computer or smartphone use or experience [52]. In total, 4 (33%) out of 12 participants mentioned having encountered a smart speaker in their children's homes without personal use [54]. Of these, over half of the participants (7/12, 58%) owned a tablet, and all (12/12, 100%) confirmed regular use of computers for tasks such as information searches and email. Participants in 2 studies were observed interacting with multimodal voice-activated virtual assistants in public settings to understand the challenges faced by this demographic [55,58] and their progression from novice to more experienced users [54].

The authors seldom reported on racial, ethnic, cultural, linguistic, educational, or socioeconomic characteristics of the study participants. One study did not reveal specific characteristics about their participants other than identifying their gender, age, and health concerns [54]. Participants in another study were only identified in the limitations section as being a group that reflected "largely White, upper-middle-class households" [52]. None of the studies specified or included information regarding the socioeconomic (income), education, sexual orientation, or religious affiliation of their participants.

Theories, Models, and Frameworks Used During Study Designs

Only 1 study was informed by older adult learning theory [52]. Older adult learning theory provided a framework for interpreting older adults' interactions with technology by emphasizing self-directed learning, relevance to life experiences, and problem-solving orientation [52]. This study posited that enhancing older adults' successful use and acceptance of technology may benefit from diverse self-directed learning approaches, including trial-and-error, errorless learning, repetitive practice, and tailored training materials specifically aimed at this demographic, and involving members of their social support network in the learning process [52]. By contrast, the authors noted that commercial systems for CAs are often marketed as "intuitive" and seldom incorporated these or other types of learning strategies and materials to assist new users [52]. This same study used their prior work to categorize the interactions using an organizational framework by their types, such as asking for weather updates or playing music, and by levels of difficulty [52]. The authors grouped all recorded interactions into 3 broader categories: "out of the box" interactions lacked personalized or account-specific qualities, like inquiring about the weather or setting timers (level 1); personalized interactions, such as setting reminders, managing calendars, or setting alarms (level 2); and interactions requiring access to a third party "Google Action," such as playing games or controlling home automation (level 3) [52].

How Older Adults and Technologies Coadapt or Coevolve

Papers described the distinctions and intricacies involved in tailoring voice-activated virtual assistants' interactions to the needs and capacities of older adults in the future to enhance use and support long-term adoption [54]. Older adult participants formed emotional connections with their voice-activated virtual assistants, seeing them as companions or friends. This relational continuity could potentially increase compliance with health-related tasks [55]. The technologies demonstrated the potential to enable management, relational, and informational continuity of care by providing consistent health management support and retaining knowledge of the users' health histories and preferences, which may enhance users' trust in these technologies [55]. Participants in the included studies often engaged in problem-solving to enhance their use of technology, with some noting that errors were part of their learning process [54,56]. As users began to be more familiar with technology, they exhibited resilience toward errors, attributing some to their own interaction approaches rather than solely blaming device shortcomings [54]. Individuals with brain cancer acclimated to the technology by consistently engaging with the app, inputting health information, and using the educational resources provided, showcasing a level of adaptation that surpasses mere familiarity [53]. This adaptation process was partly automated, guided by the app's algorithms, which customize content and interventions based on the individual's reported conditions and activities [53].

Participants also continually refined their approach to engaging with the voice-activated virtual assistants based on ongoing experiences [54]. For example, some adapted by refining their command phrasing and recognizing the inherent limitations of the technology [54]. Specifically, participants often evaluated the challenges they had with technology to engage in continuous learning and skill development to better navigate technology use [52,54]. In particular, participants developed persistent routines for using the technology [52,54] as potential solutions to challenges they had. For example, playing music before bed and checking time or weather upon waking were prevalent routines among participants [54]. Thus, over time, older adults

XSL•FO
RenderX

integrated technology into their daily lives, especially around bedtime and waking hours [54].

Some participants also adapted their communication style, trialing the use of different tones and intonations with the technology; for some, such adaptation led to more successful interactions, while for others, this resulted in failure [54,55,57]. However, over time, participants began to incorporate additional words, even with mispronunciations or using alternative phrases like "Alexia" and "Alessa," to engage the voice-activated virtual assistant. This also paraphrased commands or repeated them more clearly to enhance the device's understanding [54,55] and adjusted their approach if their initial attempts failed [55]. When observing attempts that were successful by other individuals, participants mimicked these interactions, impacting the virtual assistant's responses [55]. However, this created often interrupted personalized features of the technology being able to understand the primary user [55]. Attending training on the technology by care partners did not significantly inform the use of the technology [52].

Sustained Challenges of Prolonged Technology Use

The most widely identified challenge was the result of human-machine barriers and more specifically the limited capabilities of virtual assistants to mimic human communication patterns. Communication barriers lead to frustration when the technology did not reply to participant's replies, resulting in repeated attempts at communication [52,54]. Specifically, participants' verbal cues aimed at signaling interaction with the technology were sometimes missed by the virtual assistant due to constraints like the need for precise word use [54,55]. Voice-activated virtual assistants with screens failed to synchronize visual prompts with participants' responses to the technology, causing confusion and frustration [52,55]. Participants in one study had access to a technical-support contact person to resolve issues, which is not a realistic scenario for many older adults who may lack such immediate support [55]. The controlled study setting, including the availability of technical support, does not fully reflect the real-world challenges older adults might face when using such technologies independently.

Study Strengths, Limitations, and Recommendations Noted in the Included Studies

None of the study authors explicitly mentioned the strengths of their studies. The most commonly reported limitations were related to the technology under investigation and the demographics of the participant sample. Studies had limited sample sizes and lacked participant diversity. This limitation underscores the need for future research to engage a broader and more representative sample of older adults, considering variables such as geographic location, culture, or socioeconomic status [52,54,55]. Moreover, most participants were relatively healthy; thus, the use of design probes for health data reporting could be challenging for those who are more vulnerable or less technology-savvy, highlighting the need for simpler, more intuitive interfaces [55].

Studies also primarily took place in urban environments, potentially restricting the relevance of the results to similar settings [55]. Additionally, the scope of the technology used in the studies was limited, raising questions about the applicability of findings to other types of voice-activated smart technologies (eg, Google Assistant and Alexa) used in diverse geographical locations and settings [55]. One study noted that subsequent research should aim to include older adults with varying levels of technology expertise and from various demographic backgrounds [55]. For example, Kim and Choudhury [54] noted that the need for future research to focus on addressing voice-activated virtual assistant use among older adults with hearing impairments is crucial to promote inclusivity. This could involve exploring innovative ways to integrate voice-activated functionality discreetly into hearing aids. Exploring how these findings apply to other voice devices is recommended by study authors to broaden the understanding of technology use among older adults [55].

One study noted the challenge of conducting a detailed quantitative analysis of use patterns due to variability in study start dates among participants [54]. Further studies should use a more rigorous recruitment strategy to enhance repeatability and should also endeavor to collect additional information, in a nonintrusive manner, regarding the reasons for participant withdrawal, especially considering that 2 of 6 participants withdrew [53]. This points to the need for methodological approaches that can accommodate and analyze variations in technology use over time.

Discussion

Summary of Findings

With an increasing aging population and the rapid expansion of the IoT, there is growing interest in the role of AAL technologies for supporting aging in place [57-59]. This scoping review aimed to investigate the interaction between these technologies and older adults, with an interest in exploring how both technologies and users adapt to enhance user experiences and generate benefits from these technologies over time. Specifically, this study delved into the coadaptation process between older adults and wearables and voice-activated virtual assistants, a critical yet underexplored area in the scientific literature. Despite the importance of this area for informing technology design and intervention development, we found a gap in long-term interaction research, with only 5 studies, mostly from the United States, addressing this topic. These studies focused on the use of voice-activated virtual assistants in the context of commercially available smart speakers, exploring how older adults engage with and adapt to these technologies over short periods of time. We found that older adults often engage in an iterative learning process, adjusting their interaction styles and approaches with technology over time. Participants tend to develop routines for using the technology and adapted their behaviors, such as words to engage with the technology and to navigate challenges encountered during technology use. However, persistent challenges related to communication barriers between older adults and the technologies were observed, impacting user confidence and hindering personalized interactions by the technology. The included studies also highlighted several limitations to existing research, including

XSL•FO
RenderX

small sample sizes, limited participant diversity, and constraints associated with the technology used. These limitations suggest the need for further research involving larger and more diverse cohorts of older adults, encompassing various demographics and geographic locations. Future studies should also explore the applicability of findings to different virtual assistants and voice-activated devices.

Lack of Diversity

Our review identified that within the scarce literature exploring the coadaptation of older adults using voice-activated virtual assistants and wearables, studies tend to reflect a limited diversity of participants across racial, ethnic, cultural, and socioeconomic backgrounds among the participants. The lack of diversity in the context of technology adaptation has been reported in previous reviews [44], and most smart technologies have been trained using native English accents [60]. This lack of diversity in the design process leads to oversights in accommodating various cultural, socioeconomic, linguistic, and individual differences among older adults [61]. Existing studies have noted that often individuals from racialized groups feel that the voices of virtual assistants do not resonate with their culture in terms of tone and dialect [62]. Similarly, technologies often struggle to identify and respond to accents of individuals for whom English is a second (or third) language [63,64]. Scholars have argued that encouraging the active involvement of diverse older adults throughout the design process spanning from conception to product development and deployment is essential to improving smart technologies to meet the needs of diverse older adults [65]. Future research is encouraged to explore how adults of various linguistic, ethnic, cultural, and socioeconomic backgrounds adapt to voice-activated virtual assistants and CAs, especially those without prior experience with these technologies. This future research can help researchers and technology developers better understand the nuanced requirements of older adults. Moreover, our review noted that participants tended to be relatively healthy, with one study including individuals with MCI [52] and another including those with hearing devices, but no communication difficulties [54]. As such, we lack the opportunity to identify the unique needs of older adults with various forms of disability looking to adopt technology to support their aging in place. Our study therefore highlights the necessity of future research to concentrate on evaluating how older adults with various forms of disability and illness coadapt to technology over time. These studies would shed light on the underlying factors contributing to varying adoption rates among older adults, providing insights into the reasons behind the differing adoption speeds of personal health records. Moreover, evaluations of technology use over time should encompass additional demographic indicators, like educational status (only reported in 1 study), to enhance comprehension of the findings among older adult participants. This inclusion would provide a broader context and facilitate better extrapolation of the study outcomes.

Methods Used for Coadaptation

Our review noted that while papers lacked a precise definition of coevolution or coadaptation, they did note the importance of exploring changes in older adults' use of smart technologies over time and emphasized the need to further tailor commercially available technologies to meet the specific needs and capacities of older adults. This tailoring will be critical to enhance the realization of anticipated benefits of these technologies for supporting aging in place, considering earlier studies highlight the importance of personalized interactions based on individual preferences and habits [66-68]. However, sustained challenges in prolonged technology use were also identified, primarily stemming from communication barriers between humans and machines, mirroring existing cross-sectional studies [69,70]. Moreover, while the design of commercially available technologies often overlooks elements of communication typical of older adults as noted in other studies [71], the reverse is also true with older adults not always understanding the commands and unique features of these technologies, leading to misunderstandings and frustration. Instructions requiring access to additional platforms or technical knowledge added to the confusion, highlighting the need for clearer, more user-friendly interfaces that incorporate multiple types of affordances to help older adults use smart technologies more easily to accomplish desired tasks, thereby leading to more positive user experiences [72]. Comparatively, future longitudinal research could delve deeper into the nuanced dynamics of human-machine communication and interaction over time (eg, longer than 10 weeks), aiming to bridge the gap between older adults and technology. The absence of a consistent definition of coadaptation across studies suggests the importance of developing a unified framework to facilitate clearer communication and comparison of findings in future research on technology adaptation among older adults.

The included studies in this review revealed how older adults use smart technology changes over time, particularly as they become more familiar with their challenges using technology. Future research in this domain could delve deeper into the bidirectional adaptation process, exploring how AI-driven technological systems can become more responsive, intuitive, and personalized for older adults. This exploration might involve investigating adaptive features, machine learning algorithms, or user interface design tailored explicitly to accommodate the evolving needs and challenges encountered by older adults.

Limitations

The inclusion criteria of this scoping review required each selected paper to be published in English from 2000 onward, contain primary data, and include mention of coadaptation between an older adult user and technology over time. We used a comprehensive and methodologically rigorous search strategy, developed with expert consultation and peer-reviewed to ensure accuracy and thoroughness. Despite these efforts, the rapidly evolving nature of technology and the broad conceptual scope of coadaptation might limit the absolute inclusivity of our study selection. In particular, the terminology and application in studies involving CAs, smart speakers, and voice-activated virtual assistants are diverse and continuously developing. While we are confident in the relevance and applicability of the studies included, we acknowledge that the field's dynamic nature may yield additional relevant studies after our review period. This recognition underlines the importance of continuous surveillance and updating of the literature in this domain to capture emerging

XSL•FO
RenderX

insights and innovations that could further elucidate the coadaptation processes between older adults and smart technologies. These limitations highlight areas for methodological refinement in future research to ensure broader applicability and understanding of the coadaptation process.

Conclusions

This scoping review aimed to explore the dynamics of interaction between older adults and 2 prevalent types of smart technologies, wearables, and voice-activated virtual assistants, over time. Through a systematic search across multiple databases using a comprehensive set of keywords, this review endeavored to synthesize existing literature on how older adults engage with these technologies, adapting their use to enhance their daily lives. Despite the rigorous search strategy, the review identified only 5 papers that met the inclusion criteria, revealing a scarcity in the body of literature on this subject.

The findings from the included studies focused on voice-activated virtual assistants and highlighted an iterative learning process undertaken by older adults, who adapt their interaction styles and behaviors to navigate and overcome challenges presented by the technology. Notably, these adaptations reflect a resilience and willingness to integrate technology into their routines, despite encountering persistent communication barriers. These barriers, which often impede personalized interactions and undermine user confidence, highlight critical areas for technological improvement.

Moreover, the review brings to light significant limitations within the existing studies, such as small sample sizes, lack of participant diversity among participants, and a narrow technological focus. These shortcomings signal a pressing need for further research that encompasses broader and more diverse populations of older adults. Such research is essential to developing a deeper understanding of the diverse ways in which older adults from various backgrounds engage with and adapt to smart technologies. By addressing these research gaps, future studies have the potential to significantly contribute to the development of more inclusive, user-friendly, and effective technological solutions tailored to the needs and preferences of the aging population.

In conclusion, this scoping review highlights many unexplored areas for investigation. Expanding the scope of research to include larger, more diverse cohorts and a wider range of technologies will be crucial to advancing the use of smart technology to improve the quality of life and independence of older adults.

Acknowledgments

The authors would like to thank and acknowledge the contributions of Charmaine De Castro, Information Specialist at the Mount Sinai Hospital-Sinai Health System, for providing guidance on the search strategy development. This study was partially funded by the AGE-WELL (Aging Gracefully Across Environments Using Technology to Support Wellness, Engagement and Long Life) network, which was funded by the Government of Canada's Networks of Centres of Excellence Program.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the project idea. KMK, AG, and JM contributed to the study design. KMK wrote the initial manuscript that was first revised by AG and JM. KMK, AG, AML, and JM conducted the preliminary thematic analysis. All authors then shared in the interpretation of the data. All authors were involved in editing and revising the manuscript. All authors approved the final manuscript and are accountable to all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 116 KB - jmir_v26i1e57258_app1.docx]

Multimedia Appendix 2
Search strategies.
[DOCX File , 19 KB - jmir_v26i1e57258_app2.docx]

Multimedia Appendix 3
Summary of papers excluded and rationale.
[DOCX File , 93 KB - jmir_v26i1e57258_app3.docx]

XSL•FO
RenderX

References

1. Rose K, Eldridge S, Chapin L. The internet of things: an overview. Internet Soc (ISOC) 2015;80:1-50 [FREE Full text]
2. Gigli M, Koo S. Internet of things: services and applications categorization. Adv Internet Things 2011;1(2):27-31. [doi: 10.4236/ait.2011.12004]
3. Lutz W, Sanderson W, Scherbov S. The coming acceleration of global population ageing. Nature 2008;451(7179):716-719. [doi: 10.1038/nature06516] [Medline: 18204438]
4. Dohr A, Modre-Opsrian R, Drobics M. The internet of things for ambient assisted living. : IEEE; 2010 Presented at: 2010 Seventh International Conference on Information Technology: New Generations; April 12-14, 2010; Las Vegas, NV, United States. [doi: 10.1109/itng.2010.104]
5. Blackman S, Matlo C, Bobrovitskiy C, Waldoch A, Fang ML. Ambient assisted living technologies for aging well: a scoping review. J Intell Syst 2016;25(1):55-69. [doi: 10.1515/jisys-2014-0136]
6. Callahan JJ. Aging in Place. Delhi: Routledge; 2019.
7. Wong AKC, Wong FKY, Chang K. Effectiveness of a community-based self-care promoting program for community-dwelling older adults: a randomized controlled trial. Age Ageing 2019;48(6):852-858. [doi: 10.1093/ageing/afz095] [Medline: 31437272]
8. Wong KC, Wong FKY, Chang KKP. Health-social partnership intervention programme for community-dwelling older adults: a research protocol for a randomized controlled trial. J Adv Nurs 2015;71(11):2673-2685. [doi: 10.1111/jan.12700] [Medline: 26040946]
9. Wong K, Wong FKY, Yeung WF, Chang K. The effect of complex interventions on supporting self-care among community-dwelling older adults: a systematic review and meta-analysis. Age Ageing 2018;47(2):185-193. [doi: 10.1093/ageing/afx151] [Medline: 28927235]
10. Pastalan LA. Aging in Place: The Role of Housing and Social Supports. Binghamton, NY: Routledge; 2013:142.
11. Friedman C, Caldwell J, Rapp Kennedy A, Rizzolo MC. Aging in place: a national analysis of home- and community-based Medicaid services for older adults. J Disabil Policy Stud 2018;29(4):245-256. [doi: 10.1177/1044207318788889]
12. Marek KD, Popejoy L, Petroski G, Mehr D, Rantz M, Lin WC. Clinical outcomes of aging in place. Nurs Res 2005;54(3):202-211. [doi: 10.1097/00006199-200505000-00008] [Medline: 15897796]
13. Marek KD, Stetzer F, Adams SJ, Popejoy LL, Rantz M. Aging in place versus nursing home care: comparison of costs to Medicare and Medicaid. Res Gerontol Nurs 2012;5(2):123-129. [doi: 10.3928/19404921-20110802-01] [Medline: 21846081]
14. Jutkowitz E, Gitlin LN, Pizzi LT, Lee E, Dennis MP. Cost effectiveness of a home-based intervention that helps functionally vulnerable older adults age in place at home. J Aging Res 2012;2012:680265 [FREE Full text] [doi: 10.1155/2012/680265] [Medline: 21876812]
15. Kokorelias KM, Grigorovich A, Harris MT, Rehman U, Ritchie L, Levy A, et al. Coadaptation between smart technologies and older adults over time: protocol for a scoping review. JMIR Res Protoc 2023;12(1):e51129 [FREE Full text] [doi: 10.2196/51129] [Medline: 37812466]
16. Lee ML, Dey AK. Sensor-based observations of daily living for aging in place. Pers Ubiquit Comput 2014;19(1):27-43. [doi: 10.1007/s00779-014-0810-3]
17. Rantz MJ, Skubic M, Miller SJ, Galambos C, Alexander G, Keller J, et al. Sensor technology to support aging in place. J Am Med Dir Assoc 2013;14(6):386-391 [FREE Full text] [doi: 10.1016/j.jamda.2013.02.018] [Medline: 23562281]
18. Cook D. Sensors in support of aging-in-place: the good, the bad, and the opportunities. In: Mobile Technology for Adaptive Aging: Proceedings of a Workshop. Washington, DC: The National Academies Press; 2020.
19. Tian YJA, Felber NA, Pageau F, Schwab DR, Wangmo T. Benefits and barriers associated with the use of smart home health technologies in the care of older persons: a systematic review. BMC Geriatr 2024;24(1):152 [FREE Full text] [doi: 10.1186/s12877-024-04702-1] [Medline: 38355464]
20. McGarrigle L, Todd C. Promotion of physical activity in older people using mHealth and eHealth technologies: rapid review of reviews. J Med Internet Res 2020;22(12):e22201 [FREE Full text] [doi: 10.2196/22201] [Medline: 33372894]
21. Domb M. Smart home systems based on internet of things. In: Internet of Things (IoT) for Automated and Smart Applications. Rijeka, Croatia: IntechOpen; 2019.
22. Davidovic B, Labus A. A smart home system based on sensor technology. Facta Univ Electron Energ 2016;29(3):451-460. [doi: 10.2298/fuee1603451d]
23. Liu L, Stroulia E, Nikolaidis I, Miguel-Cruz A, Rios Rincon A. Smart homes and home health monitoring technologies for older adults: a systematic review. Int J Med Inform 2016;91:44-59. [doi: 10.1016/j.ijmedinf.2016.04.007] [Medline: 27185508]
24. Chen H, Zhang Y, Wang L. A study on the quality evaluation index system of smart home care for older adults in the community—based on Delphi and AHP. BMC Public Health 2023;23(1):411 [FREE Full text] [doi: 10.1186/s12889-023-15262-1] [Medline: 36859259]
25. Kim KI, Gollamudi SS, Steinhubl S. Digital technology to enable aging in place. Exp Gerontol 2017;88:25-31. [doi: 10.1016/j.exger.2016.11.013] [Medline: 28025126]
26. Mynatt E. Increasing the opportunities for aging in place. 2000 Presented at: Proceedings on the 2000 Conference on Universal Usability; November 16-17, 2000; Arlington, VA, United States. [doi: 10.1145/355460.355475]

XSL•FO
RenderX

27. Choi YK, Thompson HJ, Demiris G. Internet-of-Things smart home technology to support aging-in-place: older adults' perceptions and attitudes. J Gerontol Nurs 2021;47(4):15-21. [doi: 10.3928/00989134-20210310-03] [Medline: 34038251]
28. Tural E, Lu D, Austin Cole D. Safely and actively aging in place: older adults' attitudes and intentions toward smart home technologies. Gerontol Geriatr Med 2021;7:23337214211017340 [FREE Full text] [doi: 10.1177/23337214211017340] [Medline: 34095352]
29. Demiris G, Hensel BK, Skubic M, Rantz M. Senior residents' perceived need of and preferences for "smart home" sensor technologies. Int J Technol Assess Health Care 2008;24(1):120-124. [doi: 10.1017/S0266462307080154] [Medline: 18218177]
30. Fischer G. Adaptive and adaptable systems: differentiating and integrating AI and EUD. : Springer; 2023 Presented at: International Symposium on End User Development; June 6-8, 2023; Cagliari, Italy. [doi: 10.1007/978-3-031-34433-6_1]
31. Mackay W. Responding to cognitive overload : co-adaptation between users and technology. Intellectica 2000;30(1):177-193. [doi: 10.3406/intel.2000.1597]
32. Alahmad R, Robert Jr LP. Capturing the complexity of cognitive computing systems: co-adaptation theory for individuals. 2021 Presented at: Proceedings of the 2021 on Computers and People Research Conference; June 30, 2021; Virtual Event, Germany. [doi: 10.1145/3458026.3462148]
33. Du Y. Voice assistive technology for activities of daily living: developing an alexa telehealth training for adults with cognitive-communication disorders. 2024 Presented at: Proceedings of the CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI, United States p. 1-15. [doi: 10.1145/3613904.3642788]
34. Li J, Ma Q, Chan AH, Man SS. Health monitoring through wearable technologies for older adults: smart wearables acceptance model. Appl Ergon 2019;75:162-169. [doi: 10.1016/j.apergo.2018.10.006] [Medline: 30509522]
35. Malwade S, Abdul SS, Uddin M, Nursetyo AA, Fernandez-Luque L, Zhu X, et al. Mobile and wearable technologies in healthcare for the ageing population. Comput Methods Programs Biomed 2018;161:233-237. [doi: 10.1016/j.cmpb.2018.04.026] [Medline: 29852964]
36. Arnold A, Kolody S, Comeau A, Miguel Cruz A. What does the literature say about the use of personal voice assistants in older adults? a scoping review. Disabil Rehabil Assist Technol 2024;19(1):100-111. [doi: 10.1080/17483107.2022.2065369] [Medline: 35459429]
37. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Meth 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
38. Peters M, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Implement 2021;19(1):3-10. [doi: 10.1097/XEB.0000000000000277] [Medline: 33570328]
39. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
40. Dhillon G, Backhouse J. Technical opinion: information system security management in the new millennium. Commun ACM 2000;43(7):125-128. [doi: 10.1145/341852.341877]
41. Abowd GD, Mynatt ED. Charting past, present, and future research in ubiquitous computing. ACM Trans Comput-Hum Interact 2000;7(1):29-58. [doi: 10.1145/344949.344988]
42. Champagne K, Boot WR. Exploring predictors of mobile device proficiency among older adults. 2017 Presented at: Human-Computer Interaction. Interaction Contexts: 19th International Conference, HCI International 2017, Proceedings, Part II 19; July 9-14, 2017; Vancouver, British Columbia, Canada. [doi: 10.1007/978-3-319-58077-7_13]
43. Gray M, Gills JL, Glenn JM, Vincenzo JL, Walter CS, Madero EN, et al. Cognitive decline negatively impacts physical function. Exp Gerontol 2021;143:111164 [FREE Full text] [doi: 10.1016/j.exger.2020.111164] [Medline: 33232795]
44. Kokorelias KM, Nelson ML, Tang T, Steele Gray C, Ellen M, Plett D, et al. Inclusion of older adults in digital health technologies to support hospital-to-home transitions: secondary analysis of a rapid review and equity-informed recommendations. JMIR Aging 2022;5(2):e35925 [FREE Full text] [doi: 10.2196/35925] [Medline: 35475971]
45. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
46. Lally P, Gardner B. Promoting habit formation. Health Psychol Rev 2013;7(Suppl 1):S137-S158. [doi: 10.1080/17437199.2011.603640]
47. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
48. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
49. Bramer WM, Milic J, Mast F. Reviewing retrieved references for inclusion in systematic reviews using EndNote. J Med Libr Assoc 2017;105(1):84-87 [FREE Full text] [doi: 10.5195/jmla.2017.111] [Medline: 28096751]
50. Babineau J. Product review: Covidence (systematic review software). J Can Health Libr Assoc 2014;35(2):68-71. [doi: 10.5596/c14-016]

XSL•FO
RenderX

51. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
52. Zubatiy T, Mathur N, Heck L, Vickers KL, Rozga A, Mynatt ED. "I don't know how to help with that"—learning from limitations of modern conversational agent systems in caregiving networks. Proc ACM Hum-Comput Interact 2023;7(CSCW2):1-28. [doi: 10.1145/3610170]
53. Gvozdanovic A, Jozsa F, Fersht N, Grover PJ, Kirby G, Kitchen N, et al. Integration of a personalised mobile health (mHealth) application into the care of patients with brain tumours: proof-of-concept study (IDEAL stage 1). BMJ Surg Interv Health Technol 2022;4(1):e000130 [FREE Full text] [doi: 10.1136/bmjsit-2021-000130] [Medline: 36579146]
54. Kim S, Choudhury A. Exploring older adults' perception and use of smart speaker-based voice assistants: a longitudinal study. Comput Hum Behav 2021;124:106914. [doi: 10.1016/j.chb.2021.106914]
55. Cuadra A, Bethune J, Krell R. Designing voice-first ambient interfaces to support aging in place. 2023 Presented at: Proceedings of the 2023 ACM Designing Interactive Systems Conference; July 10-14, 2023; Pittsburgh, PA, United States. [doi: 10.1145/3563657.3596104]
56. Kim D. Can healthcare apps and smart speakers improve the health behavior and depression of older adults? a quasi-experimental study. Front Digit Health 2023;5:1117280 [FREE Full text] [doi: 10.3389/fdgth.2023.1117280] [Medline: 36910571]
57. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Posterior approaches in malleolar fracture: when, why and how. J Biol Regul Homeost Agents 2020;34(3 Suppl 2):89-95. [Medline: 32856446]
58. Cuadra A, Baek H, Estrin D. On inclusion: video analysis of older adult interactions with a multi-modal voice assistant in a public setting. 2022 Presented at: Proceedings of the 2022 International Conference on Information and Communication Technologies and Development; June 27-29, 2022; Seattle, WA, United States. [doi: 10.1145/3572334.3572371]
59. Fadrique LX, Rahman D, Vaillancourt H, Boissonneault P, Donovska T, Morita PP. Overview of policies, guidelines, and standards for active assisted living data exchange: thematic analysis. JMIR Mhealth Uhealth 2020;8(6):e15923 [FREE Full text] [doi: 10.2196/15923] [Medline: 32568090]
60. Hassan MA, Rehmat A, Ghani Khan MU, Yousaf MH. Improvement in automatic speech recognition of South Asian accent using transfer learning of DeepSpeech2. Math Probl Eng 2022:825555. [doi: 10.1155/2022/6825555]
61. Cambre J, Kulkarni C. One voice fits all? social implications and research challenges of designing voices for smart devices. Proc ACM Hum-Comput Interact 2019;3(CSCW):1-19. [doi: 10.1145/3359325]
62. Brewer RN, Harrington C, Heldreth C. Envisioning equitable speech technologies for Black older adults. 2023 Presented at: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; June 12-15, 2023; Chicago, IL, United States. [doi: 10.1145/3593013.3594005]
63. Pal D, Arpnikanondt C, Funilkul S. User experience with smart voice assistants: the accent perspective. 2019 Presented at: 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); July 6-8, 2019; Kanpur, India. [doi: 10.1109/icccnt45670.2019.8944754]
64. Lima L, Furtado V, Furtado E. Empirical analysis of bias in voice-based personal assistants. 2019 Presented at: Companion Proceedings of the 2019 World Wide Web Conference; May 13-17, 2019; San Francisco, CA, United States. [doi: 10.1145/3308560.3317597]
65. Harrington CN, Jelen B, Lazar A, Martin-Hammond A, Pradhan A, Reeder B, et al. Taking stock of the present and future of smart technologies for older adults and caregivers. arXiv. Preprint posted online on March 31, 2021 . [doi: 10.48550/arXiv.2104.00096]
66. Tabassum M, Kosiński T, Frik A, Malkin N, Wijesekera P, Egelman S, et al. Investigating users' preferences and expectations for always-listening voice assistants. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020;3(4):1-23. [doi: 10.1145/3369807]
67. Braun M, Mainz A, Chadowitz R, Pfleging B, Alt F. At your service: designing voice assistant personalities to improve automotive user interfaces. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, United Kingdom. [doi: 10.1145/3290605.3300270]
68. Yefimenko O, Foehr J, Germelmann CC. I'll have what Alexa's having… but only if that's what I'm looking for!—The impact of personalization on recommendation capabilities of smart voice-interaction technology in voice commerce. SMR 2023;7(1):23-38. [doi: 10.5771/2511-8676-2023-1-23]
69. Pedrozo Campos Antunes T, Souza Bulle de Oliveira A, Hudec R, Brusque Crocetta T, Ferreira de Lima Antão JY, de Almeida Barbosa RT, et al. Assistive technology for communication of older adults: a systematic review. Aging Ment Health 2019;23(4):417-427. [doi: 10.1080/13607863.2018.1426718] [Medline: 29451399]
70. Miller DI, Bruce H, Gagnon M, Talbot V, Messier C. Improving older adults' experience with interactive voice response systems. Telemed J E Health 2011;17(6):452-455. [doi: 10.1089/tmj.2010.0204] [Medline: 21631386]
71. Luger E. "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. 2016 Presented at: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, CA, United States. [doi: 10.1145/2858036.2858288]

XSL•FO
RenderX

72. Cho Y, Choi A. Application of affordance factors for user-centered smart homes: a case study approach. Sustainability 2020;12(7):3053. [doi: 10.3390/su12073053]

Abbreviations

AI: artificial intelligence
AAL: ambient-assisted living
CA: conversational agent
IoT: Internet of Things
MCI: mild cognitive impairment
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by G Eysenbach, N Cahill; submitted 09.02.24; peer-reviewed by AKC Wong, SQ Yoong; comments to author 06.03.24; revised version received 14.03.24; accepted 26.06.24; published 07.08.24.

Please cite as:
Kokorelias KM, Grigorovich A, Harris MT, Rehman U, Ritchie L, Levy AM, Denecke K, McMurray J
Longitudinal Coadaptation of Older Adults With Wearables and Voice-Activated Virtual Assistants: Scoping Review
J Med Internet Res 2024;26:e57258
URL: https://www.jmir.org/2024/1/e57258
doi:10.2196/57258
PMID:

©Kristina Marie Kokorelias, Alisa Grigorovich, Maurita T Harris, Umair Rehman, Louise Ritchie, AnneMarie M Levy, Kerstin Denecke, Josephine McMurray. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Enhancing Patient Understanding of Laboratory Test Results: Systematic Review of Presentation Formats and Their Impact on Perception, Decision, Action, and Memory

Frederieke A M van der Mee[1*], MD; Fleur Schaper[2,3*], PhD; Jesse Jansen[1], PhD; Judith A P Bons[3], PhD; Steven J R Meex[3], PhD; Jochen W L Cals[1], MD, Prof Dr

[1]Department of Family Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

[2]Department of Clinical Chemistry, Reinier Medical Diagnostic Center, Delft, Netherlands

[3]Central Diagnostic Laboratory, Maastricht University Medical Center+, Maastricht, Netherlands

[*]these authors contributed equally

Corresponding Author:
Frederieke A M van der Mee, MD
Department of Family Medicine
Care and Public Health Research Institute
Maastricht University
P. Debyeplein 1
Maastricht, 6229 HA
Netherlands
Phone: 31 883887059
Email: frederieke.vandermee@maastrichtuniversity.nl

Abstract

Background: Direct access of patients to their web-based patient portal, including laboratory test results, has become increasingly common. Numeric laboratory results can be challenging to interpret for patients, which may lead to anxiety, confusion, and unnecessary doctor consultations. Laboratory results can be presented in different formats, but there is limited evidence regarding how these presentation formats impact patients' processing of the information.

Objective: This study aims to synthesize the evidence on effective formats for presenting numeric laboratory test results with a focus on outcomes related to patients' information processing, including affective perception, perceived magnitude, cognitive perception, perception of communication, decision, action, and memory.

Methods: The search was conducted in 3 databases (PubMed, Web of Science, and Embase) from inception until May 31, 2023. We included quantitative, qualitative, and mixed methods articles describing or comparing formats for presenting diagnostic laboratory test results to patients. Two reviewers independently extracted and synthesized the characteristics of the articles and presentation formats used. The quality of the included articles was assessed by 2 independent reviewers using the Mixed Methods Appraisal Tool.

Results: A total of 18 studies were included, which were heterogeneous in terms of study design and primary outcomes used. The quality of the articles ranged from poor to excellent. Most studies (n=16, 89%) used mock test results. The most frequently used presentation formats were numerical values with reference ranges (n=12), horizontal line bars with colored blocks (n=12), or a combination of horizontal line bars with numerical values (n=8). All studies examined perception as an outcome, while action and memory were studied in 1 and 3 articles, respectively. In general, participants' satisfaction and usability were the highest when test results were presented using horizontal line bars with colored blocks. Adding reference ranges or personalized information (eg, goal ranges) further increased participants' perception. Additionally, horizontal line bars significantly decreased participants' tendency to search for information or to contact their physician, compared with numerical values with reference ranges.

Conclusions: In this review, we synthesized available evidence on effective presentation formats for laboratory test results. The use of horizontal line bars with reference ranges or personalized goal ranges increased participants' cognitive perception and perception of communication while decreasing participants' tendency to contact their physicians. Action and memory were less frequently studied, so no conclusion could be drawn about a single preferred format regarding these outcomes. Therefore, the use of horizontal line bars with reference ranges or personalized goal ranges is recommended to enhance patients' information

XSL•FO
RenderX

processing of laboratory test results. Further research should focus on real-life settings and diverse presentation formats in combination with outcomes related to patients' information processing.

(J Med Internet Res 2024;26:e53993) doi:10.2196/53993

KEYWORDS

electronic health record; patient access to records; patient portal; laboratory test results; clinical laboratory information systems; health communication; health informatics; patient engagement; patient involvement

Introduction

An increasing number of patients have direct access to their own web-based patient portal. This includes diagnostic test results ordered by their health care professional, such as laboratory test results [1,2]. Providing patients with web-based access to patient portals aims to enhance patient involvement in their health management. Improving patients' knowledge and self-efficacy may enhance disease self-management and interactions with health care providers, and ultimately lead to better health outcomes and increased satisfaction with care [3-6].

However, patient access to web-based patient portals also has potentially negative consequences. For example, misinterpretation or inaccurate knowledge could lead to underestimation of test results and promote a false sense of security [7]. Similarly, gaining insight into medical test results might trigger feelings of insecurity, anxiety, and confusion [8-12]. Previous studies have shown that poor understanding of test results can lead to an increase in telephone calls or doctor consultations, emergency department visits, and even hospitalizations [13-15]. As a result, the overall utility or benefit of providing lab results directly to patients may depend on how these data are presented to and interpreted by the patient [16,17].

Limited health literacy and numeracy skills are significant barriers to the effective use of web-based patient portals and understanding of laboratory test results [18,19]. Although patient understanding can be improved to some extent by avoiding medical jargon and using plain language, overcoming the problem of incomprehension in its entirety remains an ongoing challenge [19-21]. One of the key issues is the numerical presentation of test results, especially for patients with low numeracy skills (ie, those with limited ability to derive meaning from numbers), who have been shown to have difficulties in interpreting basic laboratory test results and identifying results that fall outside the reference range [18]. The lack of supporting information and guidance on interpretation of results contributes to the problem of misinterpretation. This challenge becomes even more pronounced when a larger number of test results are presented [18].

Basic patient portals typically present laboratory test results in a numerical format, often accompanied by a reference range (ie, the range that represents normal values for a particular test) [10,22,23]. An alternative approach to communicating test results is the use of visual displays, such as colors or graphics. These formats require less health literacy and numeracy skills for interpretation and may improve patients' understanding of the results [24-28]. Previous studies have examined a variety of presentation formats for communicating laboratory test results. However, direct comparisons between these studies can be challenging due to the variety of presentation options and clinical contexts. In addition, not all formats may be appropriate for every clinical situation [29].

There is only limited evidence on the effect of specific presentation formats on patients' information processing. As highlighted by Witteman and Zikmund-Fisher [17], laboratory test results often lack meaning for the patients receiving them. Test results represent data, which differs from information and actual knowledge patients commonly encounter in daily life [30,31]. Patients have to complete several steps to go from data perception to usable knowledge. Ancker et al [32] described these steps as well, based on the Wickens model of human information processing [33]. In a sequential order, patients need perception and behavioral intention to achieve actual health behavior. Therefore, it is important that these separate steps of information processing are taken into account when presentation formats are evaluated.

Our systematic review aims to synthesize the existing evidence on effective components of presentation formats for laboratory test results focusing on patients' perception, decision, action, and memory. In this review, we will specifically focus on numeric laboratory test results, and not on results containing only textual or nonnumeric findings.

Methods

This review was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 1) [34]. A protocol for this review was not previously registered.

Search Strategy

The search was conducted in 3 databases (PubMed, Web of Science, and Embase) from inception up to May 31, 2023. In each database, a search was performed, which was developed by the first author (FM) together with an experienced librarian and contained both thesaurus and free text terms. For the search in Embase, a filter was applied to remove preprint records and to exclude MEDLINE citations, since the latter were already covered by the PubMed search. Additionally, 2 authors (FM and FS) performed backward snowballing by screening reference sections of all selected articles to identify relevant publications missed with the search strategy. A fully reproducible search can be found in Multimedia Appendix 2.

Study Selection and Eligibility Criteria

All identified titles and abstracts were downloaded to reference management software (Endnote) and duplicates were removed. Two authors (FM and FS) independently screened for potential

XSL•FO
RenderX

eligible articles using Covidence, a Cochrane's technology platform [35]. First, titles and abstracts were screened against the eligibility criteria. Second, full texts of potentially suitable articles were rescreened using the same criteria. In case of disagreement, consensus was reached by discussion or screening by a third reviewer (JC).

We considered articles fitting for inclusion if they were original research. Studies describing or comparing different ways of presenting diagnostic laboratory test results to patients were included. Only studies examining numeric laboratory test results were included. Furthermore, studies are needed to evaluate the effect of communicating test results on patients' comprehensibility, attitudes, or experiences. Studies conducted in primary care and secondary/tertiary care settings were eligible, as well as studies including healthy volunteers. Studies had to be written in English or Dutch.

Studies were excluded if they (1) were protocols, reviews, systematic reviews, meta-analyses, book chapters, editorials, letters, practice pointers, oral presentations, or poster presentations; (2) were about development, implementation, or adoption of web-based patient portals in general, or about the type of notification of laboratory test results, if they did not consider patients' interpretation of the lab results; (3) focused on web-based access to notes, and not to laboratory test results; (4) did not mention type of presentation format of lab results; (5) focused on the development of web-based lifestyle interventions or web-based applications to collect patient-reported outcomes; (6) focused on the safety or privacy issues of web-based patient portals; (7) were about the effect of communicating test results in web-based patient portals on patients' medication management; (8) tested the effect of test result communication on health care providers; (9) examined communication of other types of diagnostic test results (eg, pharmacogenomics or genomics, radiology, pathology, or microbiology); and (10) examined communication of test results in the context of screening programs.

Data Extraction

Two authors (FM and FS) independently extracted data from the eligible studies into a prepared spreadsheet. The spreadsheet was developed by the multidisciplinary team and piloted by both authors. For each study, the year of publication, country in which the study was performed, study design, number of participants, description of the study population, and the inclusion and exclusion criteria were assessed. Furthermore, information about the presentation of test results in the portal, the type of laboratory tests studied, and whether real or mock data were used, was extracted.

Outcome Measures

Previous research regarding this subject focused on a variety of outcomes related to patients' information processing. As stated above, Ancker et al [32] introduced a taxonomy to categorize different outcome measures when communicating numbers in health care. These categories include sequentially; perception, decision/behavioral intention, action/actual health behavior, and memory. Perception is further divided into 4 subcategories: affective perception, perceived magnitude, cognitive perception, and perception of communication [32,36,37]. An explanation of the categorized outcome measures can be found in Textbox 1 [32]. For this review, the outcome measures of each study were extracted and classified into the categories described.

Textbox 1. Explanation of the outcome measure categories based on patients' information processing by Ancker et al.

Affective perception

- Feelings about the laboratory result communicated.

Perceived magnitude

- Perceived size of risk associated with a test result, captured with measures as "how large or small does this value seem to you?"

Cognitive perception

- Understanding whether a laboratory result is elevated, normal, or below normal. Being able to identify direction of a trend over time.

Perception of communication

- Preference for presentation format of test result.

Decision

- Intention to seek more information or to change behavior after viewing results.

Action

- Change in actual health behavior (eg, search for more information).

Memory

- Recall of a specific test result after viewing (ie, verbatim recall).

XSL•FO
RenderX

Quality Assessment

To assess the quality and risk of bias of all included studies, the Mixed Methods Appraisal Tool (MMAT) was used [38]. The MMAT is designed to concomitantly appraise studies with different designs, such as qualitative, quantitative, and mixed methods studies [39]. Question sets are specific to the study design, notably qualitative studies, quantitative randomized controlled trials, quantitative nonrandomized studies, quantitative descriptive studies, and mixed methods studies. For each suitable study, the appropriate category was chosen and the criteria stated for this specific category were rated as "yes," "no," or "can't tell."

Two authors (FM and FS) discussed both data and quality extraction until a consensus was reached.

Data Synthesis

Due to the heterogeneity of study designs and primary outcomes, meta-analysis was considered inappropriate. Instead, narrative synthesis was used to integrate the findings into descriptive summaries regarding ways of presenting laboratory test results and outcomes of interest.

Results

Overview

The initial search identified 10,537 references. A total of 3490 duplicate records were removed. After applying the exclusion criteria in the primary title and abstract screening, another 6900 records were removed. During full-text screening of the remaining articles (n=146), it appeared that 1 full text was not available. Furthermore, 127 articles were excluded because they did not meet the eligibility criteria. Describing the implementation of web-based patient portals, unrelated to laboratory test results, was the most common exclusion criterion (55/127, 43.3%; Figure 1). A total of 18 studies were found eligible for this systematic review. Cohen κ for interrater reliability was 0.62 for title and abstract screening and 0.80 for full-text screening, indicating respectively a moderate and strong agreement between the 2 reviewers [40].

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

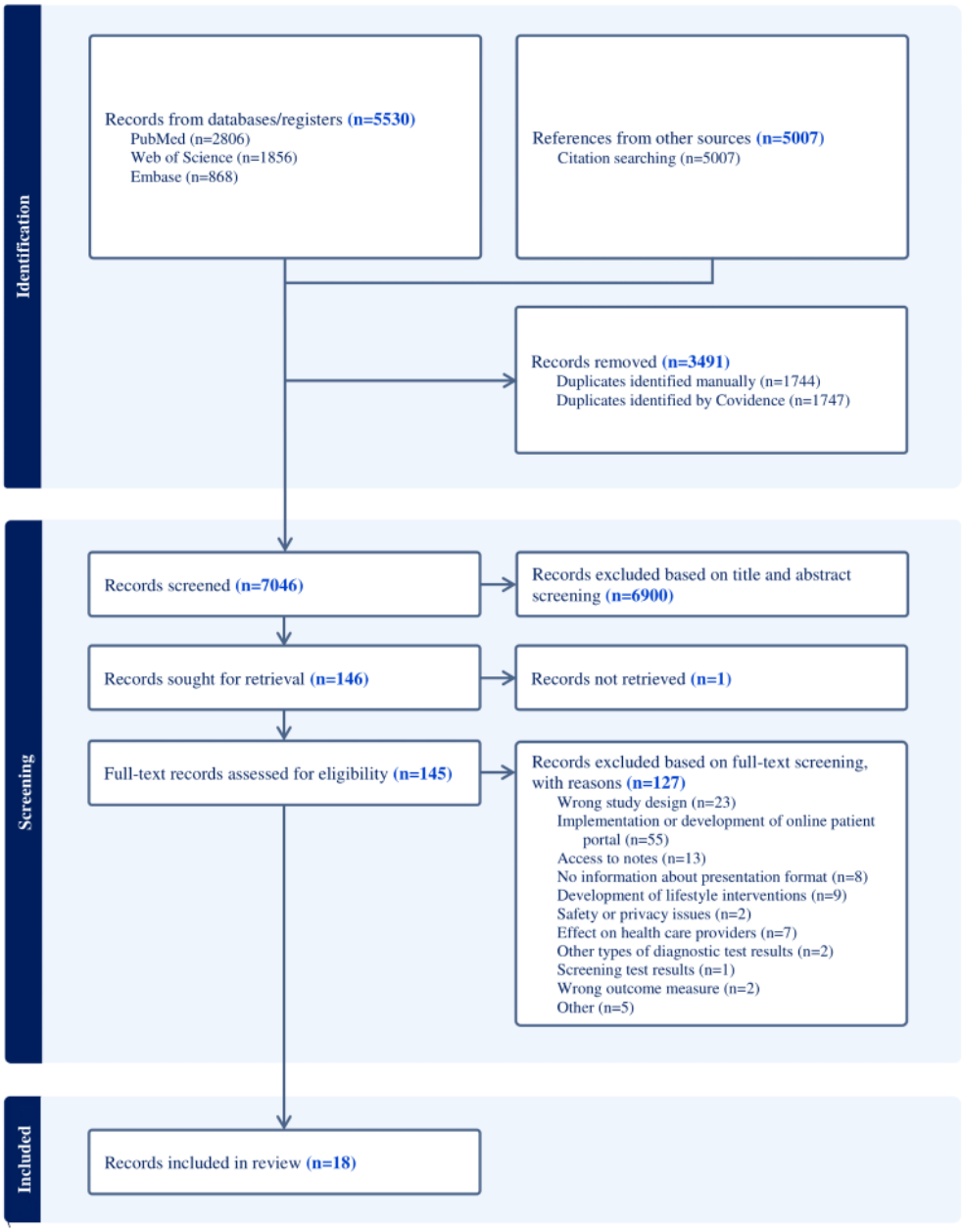

Study Characteristics

A total of 2 qualitative studies, 11 quantitative studies, and 5 mixed methods studies were included (n=18). The included studies were published between 2012 and 2021, and the majority were conducted in the United States (n=13, 72%). The total sample size of the included studies was 12,225 participants, ranging from 8 to 6766 participants. Among the articles reporting the following characteristics, sex was almost equally distributed (6219/13,155, 47.3% female), and participants were predominantly middle-aged (mean 51.1 years) and White (8429/10,865, 77.6% on average). Fourteen (78%) of the 18 studies reported educational level, with 48% (5676/11,813) of the participants reporting a higher education (defined as college-degree or higher). Overall characteristics of the included studies and populations are summarized in Table 1.

XSL•FO
RenderX

Table 1. Study and population characteristics of all included studies (n=18).

Author (year)	Country	Study design	Sample (n)	Population characteristics					Aim of study
				Sample	Sex (% female)	Mean age in years (SD or range)	Race and ethnicity	Education[a]	
Bar-Lev et al (2020) [41]	Israel	Survey	225	Convenience sample	55.9	35 (14)	__[b]	0% low education, 32.6%, middle education, 61.6% high education, and 5.8% other	To examine how different visual displays of personalized medical information affect laypersons' understanding, perceptions, and actions
Brewer et al (2012) [42]	United States	Randomized controlled trial; nonrandomized experimental study	106	Convenience sample	79.2	46 (30-83)	82% White	0% low education, 27% middle education, and 73% high education	To compare the relative usability of tables and horizontal bar graphs for presenting medical test results electronically to consumers
Elder et al (2012) [43]	United States	Qualitative study	12	Convenience sample	67	60 (34-73)	83% White, 8% Black, and 8% Asian	0% low education, 58% middle education, and 42% high education	To understand patients' experiences with, and preferences for, results notification and communication in primary care settings
Fraccaro et al (2018) [7]	United Kingdom	Nonrandomized experimental study	20	Real patients	20	51.8 (10.3)	—	5% low education, 35% middle education, and 60% high education	To investigate if presentations using color improve patients' interpretation of laboratory test results presented through patient portals
Hohenstein et al (2018) [44]	United States	Mixed methods study	301	Volunteers	51	46.0 (16.3, 18-90)	66.8% White, 19.6% Hispanic/ Latino/ Spanish, 12.3% Black/ African American/ Negro, and 4% Asian	0% low education, 38.2% middle education, 48.2% high education, 13.6% unknown	To explore how people interpret medical test results, examined in various interface designs developed to enable self-care and health management
Kelman et al (2016) [45]	United States	Survey	211	Convenience sample	90	52.7 (10.0)	89% White, 4% African American, 6% other and 0.5% preferred not to answer	0.5% low education, 57% middle education, 41% high education, and 1% unknown	To explore ways in which laboratory test results can be communicated in a patient-friendly manner

XSL•FO
RenderX

Author (year)	Country	Study design	Sample (n)	Population characteristics					Aim of study
				Sample	Sex (% female)	Mean age in years (SD or range)	Race and ethnicity	Education[a]	
Morrow et al (2017)[c] [46]	United States	Mixed methods study	36	—	67	77 (65-89)	—	—	A pilot study to finalize development of video-enhanced messages before conducting formal evaluation studies
Morrow et al (2019)[c] [47]	United States	Randomized controlled trial	144	—	71.5	71.9 (60-94)	—	18.8% low education, 13.2% middle education, and 68% high education	To investigate how to support older adult comprehension of and response to patient portal-based numerical information
Nystrom et al (2018) [48]	United States	Mixed methods study	14	Real patients	—	43 (25-73)	—	—	To study patient's ability to generate meaning from each test result and how this meaning would inform their decision-making and subsequent actions
Scherer et al (2018) [49]	United States	Randomized controlled trial	6766	Mixed sample	50.9	49.1 (15.8)	78.2% White, 14.8% African America, and 9.7% other	2% low education, 52.2% middle education, and 45.8% high education	To test the impact of including clinically appropriate goal ranges outside the standard range in the visual displays of laboratory test results
Struikman et al (2020) [50]	The Netherlands	Randomized controlled trial	487	Volunteers	50.3	52.8 (15.4)	—	7.7% low education, 45.8% middle education, 46.4% high education	To discover whether the way of presenting blood test outcomes in an electronic patient portal is associated with patient health engagement and whether this varies across different test outcomes
Talboom-Kamp et al (2020) [51]	The Netherlands	Survey	354	Real patients	—	—	—	—	To investigate attitudes, experiences, and self-efficacy of patients using an online patient portal that communicates laboratory test results

XSL•FO
RenderX

Author (year)	Country	Study design	Sample (n)	Population characteristics					Aim of study
				Sample	Sex (% female)	Mean age in years (SD or range)	Race and ethnicity	Education[a]	
Tao et al (2018) [23]	China	Nonrandomized experimental study	72	Convenience sample	56	Young adult group: 22.3 (2.6); older adult group: 65.8 (3.6)	—	1.4% low education, 33.3% middle education, and 65.3% high education	To examine the effects of 4 graphical formats and age on consumers' comprehension, perceptions, visual attention, and preference of the graphs of the use of self-monitoring test results
Zarcadoolas et al (2013) [52]	United States	Qualitative study	28	Volunteers	64.3	40.0 (12.4, 21-63)	25% Hispanic, 3.6% non-Hispanic White, 67.9% non-Hispanic Black, and 3.6% other	46.4% low education, 53.6% middle education, and 0% high education	To identify vulnerable consumers' response to patient portals, their perceived utility and value, as well as their reactions to specific portal functions
Zhang et al (2020)[c] [15]	United States	Mixed methods study	203	Volunteers	48.3	63.5 between 26-49 years	69.5% White, 4.4% Asian or Pacific islander, 16.7% African American, 5.9% Hispanic or Latino, 2% American Indian, and 1.5% other	0% low education, 19.7% middle education, 79.9% high education, and 0.4% other	To examine the challenges and needs of patients when comprehending laboratory test results
Zhang et al (2021)[c] [53]	United States	Mixed methods study	8	—	50	18-64	—	—	To examine how to help patients understand the connections between their medical context and test results, and the necessary support and actions after receiving these test results
Zikmund-Fisher et al (2017)[d][22]	United States	Survey	1620	Volunteers	52.3	48.9 (15.7)	77.4% White, 13% African American, and 7% other	1.9% low education, 49.9% middle education, and 48.2% high education	To investigate the extent to which different visual displays help people discriminate between test results that do or do not require urgent action

XSL•FO
RenderX

Author (year)	Country	Study design	Sample (n)	Population characteristics					Aim of study
				Sample	Sex (% female)	Mean age in years (SD or range)	Race and ethnicity	Education[a]	
Zikmund-Fisher et al (2018)[d] [54]	United States	Randomized controlled trial	1618	Volunteers	52.1	48.8 (19-89)	77.8% White, 13.2% Black, 13.2% Hispanic, 4% Asian, 0.8% native American, and 4.3% other or multirace	0% low education, 0% middle education, 50.1% high education, and 49.9% unknown	To test the effect of including an additional harm anchor reference point in visual displays of laboratory test results

[a]Low education: primary school. Middle education: secondary, high, or trade school or some college. High education: 4-year, college, associate, university, undergraduate, bachelor's, master's, advanced, professional, or doctorate degree.

[b]Not available.

[c]The following articles are pilot and main studies: Morrow et al (2017) [46] and (2019) [47], as well as Zhang et al (2020) [15] and (2021) [53].

[d]The following articles originate from the same parent study: Zikmund-Fisher et al (2017) [22] and (2018) [54].

The most frequently used laboratory tests were lipid profile (n=10) and hemoglobin A_{1c} (HbA_{1c}) or glucose (n=5). In total, 4 studies used real patients as study population, other studies used healthy volunteers, a convenience sample, or a mixed sample (n=12) or did not define their study population (n=3). Studies used mock test results (ie, hypothetical results; n=16), real results (n=1, with real patients), or both (n=1). The majority of studies used numerical values with reference ranges (n=12) or horizontal line bars with colored blocks (n=12; Table 2). A more detailed overview of the different ways of presenting test results is provided in Multimedia Appendix 3 [7,15,22,23,41-54]. An explanation of the different presentation formats can be found in Figure 2.

XSL•FO
RenderX

Figure 2. Examples of presentation formats used for displaying laboratory test results. The examples are based on a hypothetical HbA_{1c} test result. A combination of different presentation formats is possible. HbA_{1c}: hemoglobin A_{1c}.

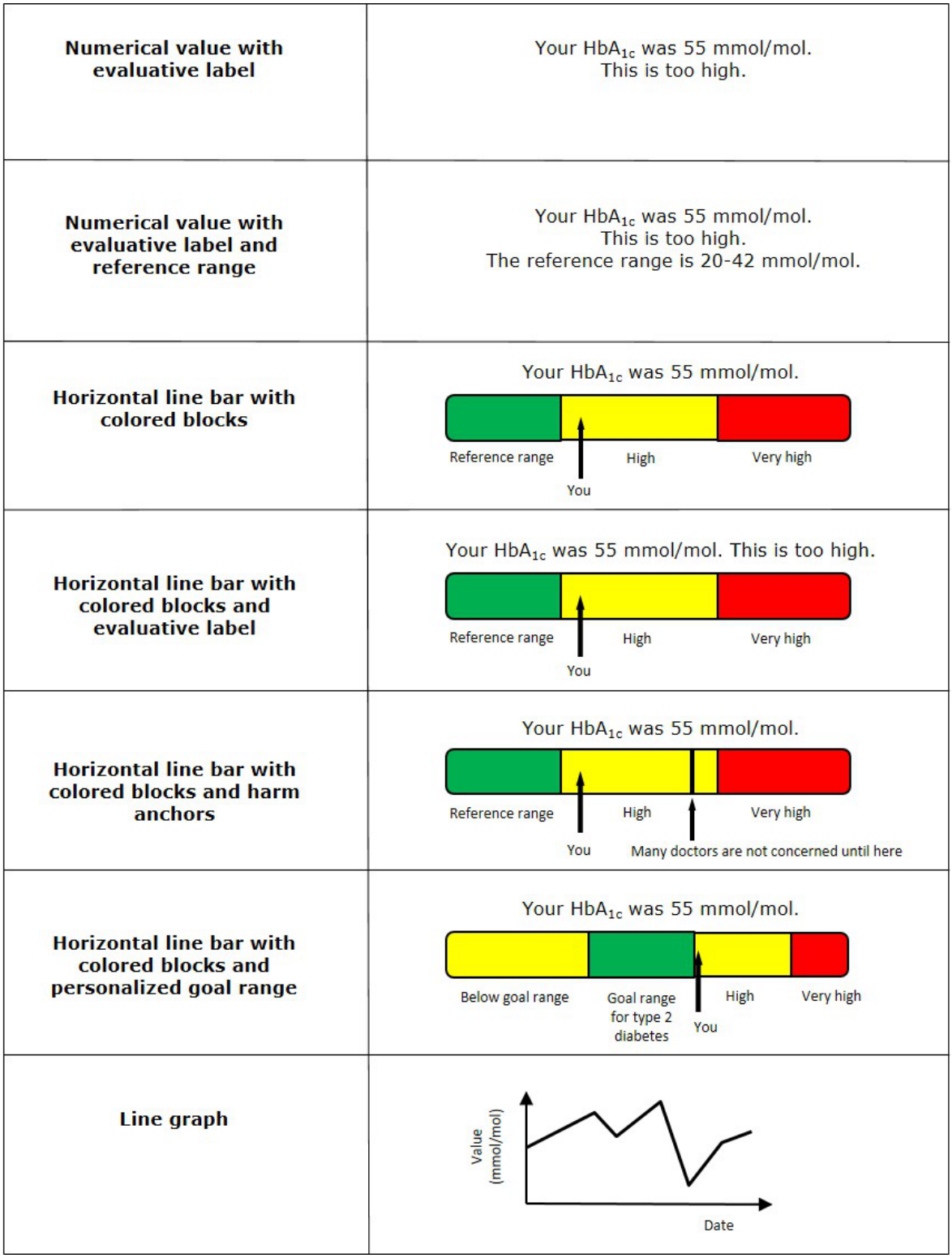

XSL•FO
RenderX

Table 2. Laboratory test characteristics and presentation format used in all included studies (n=18).

Author (year)	Laboratory test information and presentation format						
	Laboratory test	Type of data	Presentation format				
			Numerical	Horizontal line bar	Graph	Video	Text only
Bar-Lev et al (2020) [41]	Hemoglobin, cholesterol, progesterone	Mock	✓		✓		✓
Brewer et al (2012) [42]	Total cholesterol, HDL[a], LDL[b]	Mock	✓	✓			
Elder et al (2012) [43]	Total cholesterol, HDL, LDL	Mock	✓	✓	✓		✓
Fraccaro et al (2018) [7]	Creatinine, eGFR[c], potassium	Mock	✓	✓			
Hohenstein et al (2018) [44]	Vitamin B12, procalcitonin, cholesterol	Mock	✓	✓			
Kelman et al (2016) [45]	Rheumatoid factor	Mock	✓				
Morrow et al (2017) [46]	Total cholesterol, HDL, LDL, TG[d], HbA_{1c}[e]	Mock				✓	
Morrow et al (2019) [47]	Total cholesterol, HDL, LDL, TG, HbA_{1c}	Mock	✓	✓		✓	
Nystrom et al (2018) [48]	Total cholesterol, HDL, LDL, TG	Mock		✓			
Scherer et al (2018) [49]	HbA_{1c}	Mock	✓	✓			
Struikman et al (2020) [50]	Hemoglobin, TSH[f], vitamin D	Mock	✓	✓			
Talboom-Kamp et al (2020) [51]	Type of test differed per patient	Real		✓			
Tao et al (2018) [23]	Glucose (fasting and postprandial)	Mock		✓			
Zarcadoolas et al (2013) [52]	Total cholesterol, HDL, LDL, TG, HbA_{1c}	Mock	✓				
Zhang et al (2020) [15]	Total cholesterol, HDL, LDL, TG	Real and mock	✓				
Zhang et al (2021) [53]	Total cholesterol, HDL, LDL	Mock		✓			
Zikmund-Fisher et al (2017) [22]	Platelet count, ALT[g], creatinine	Mock	✓	✓			
Zikmund-Fisher et al (2018) [54]	Platelet count, ALT, creatinine	Mock		✓			

[a]HDL: high-density lipoprotein.
[b]LDL: low-density lipoprotein.
[c]eGFR: estimated glomerular filtration rate.
[d]TG: triglycerides.
[e]HbA_{1c}: hemoglobin A1c.
[f]TSH: thyroid stimulating hormone.
[g]ALT: alanine aminotransferase.

Quality Assessment

The quality assessment tool (MMAT) includes 5 assessment criteria per study design, each of which is given a score of 20% if present (Multimedia Appendix 4 [7,15,22,23,41-54]). Both qualitative articles (n=2) scored 100%, indicating excellent quality. Quantitative articles (n=11) scored between 0% and 100%, indicating a broad range of quality. These articles lost points mainly for sampling issues (biased sampling strategies and unrepresentative samples), randomization issues (unclear randomization process and incomparable groups at baseline), unclear blinding process, and lack of clarity about the completeness of outcome data and nonresponse bias. Mixed methods articles (n=5) scored between 60% and 100% (low-to-high quality), for the same reasons as described above. In addition, weaknesses in these articles included having an

XSL•FO
RenderX

unclear rationale for using a mixed methods design, unclear presentation format, and failure to adequately interpret the results of the integration of qualitative and quantitative findings.

Outcome Measures

Overview

In all 18 studies, perception was an outcome measure, further categorized into affective perception (n=7), perceived magnitude (n=6), cognitive perception (n=10), and perception of communication (n=14; Table 3 and Textbox 1). Additionally, 10 studies assessed behavioral intention, while memory was considered as an outcome measure in 3 of the included studies.

Table 3. The outcomes assessed in all included studies (n=18).

Author (year)	Perception				Decision	Action	Memory
	Affective perception	Perceived magnitude	Cognitive perception	Perception of communication	Behavioral intention	Health behavior	Verbatim recall
Bar-Lev et al (2020) [41]		✓			✓		
Brewer et al (2012) [42]				✓			✓
Elder et al (2012) [43]			✓	✓			
Fraccaro et al (2018) [7]		✓			✓		✓
Hohenstein et al (2018) [44]	✓		✓	✓			
Kelman et al (2016) [45]			✓	✓	✓		
Morrow et al (2017) [46]	✓		✓	✓			
Morrow et al (2019) [47]	✓	✓		✓	✓		✓
Nystrom et al (2018) [48]				✓	✓		
Scherer et al (2018) [49]	✓		✓		✓		
Struikman et al (2020) [50]	✓		✓		✓		
Talboom-Kamp et al (2020) [51]	✓			✓	✓		
Tao et al (2018) [23]		✓	✓	✓			
Zarcadoolas et al (2013) [52]			✓	✓			
Zhang et al (2020) [15]	✓		✓	✓		✓	
Zhang et al (2021) [53]			✓	✓			
Zikmund-Fisher et al (2017) [22]		✓		✓	✓		
Zikmund-Fisher et al (2018) [54]		✓		✓	✓		

Affective Perception

Several studies explored participants' confidence and concerns while viewing and interpreting laboratory results [15,44,47,49,51]. Talboom-Kamp et al [51] demonstrated that presenting laboratory test results in horizontal line bar format with colored blocks and evaluative labels (ie, textual explanation) enhanced participants confidence in managing their own health, although this effect was not significant. No comparison between different presentation formats and the

XSL•FO
RenderX

influence on confidence was described. These comparisons were also lacking in the other studies.

When results were presented in a horizontal line bar format with colored blocks and a personalized goal range, the negative affect was significantly higher than when results were presented without colored blocks [49].

Scherer et al [49] studied the use of personalized reference values or goal ranges. A type 2 diabetes mellitus scenario was studied, in which standard reference ranges are often not applicable. Replacing standard ranges with goal ranges significantly reduced perceived discouragement compared with situations without goal display, highlighting a positive effect of goal ranges on affective perception [49]. Furthermore, 2 other studies recommended the use of personalized reference ranges in their discussion [44,51].

In 3 studies, whether laboratory test results were within reference ranges seemed to be more important than the presentation format. As results moved further from the reference range, positive emotions decreased and negative emotions increased [15,46,47]. This change in affective perception was not influenced by message format.

Perceived Magnitude

The perceived magnitude of risk of extremely out-of-range results remained unaffected by the presentation formats in all studies. However, for near-normal or slightly out-of-range results participants encountered difficulties in estimating test result severity. Accurate risk perception was lacking, since the severity of these results was inconsistently overestimated or underestimated [7,22,41,47,54]. Zikmund-Fisher et al [54] demonstrated that the incorporation of harm anchors (ie, a threshold line outside the reference range labeled “many doctors are not concerned until here”) significantly enhanced adequate estimations of test result severity for slightly out-of-range results.

Three studies investigated the effect of presentation format on the perceived size of risk [22,23,47]. Morrow et al [47] compared horizontal line bars with both numerical and video-enhanced formats. For both low- and borderline-risk scenarios, the perceived magnitude of risk was significantly higher when horizontal line bars were used, indicating that participants tend to overestimate risk for normal results [47]. Tao et al [23] did not specify whether result normality affected risk perception using different types of horizontal line bars. However, when personalized information was added to the line bar, the risk was perceived as significantly higher. Interestingly, despite this, participants expressed a preference for personalized line bars [23]. Zikmund-Fisher et al [22] compared different types of horizontal line bars with a numerical format. Participants expressed the highest risk perception when near-normal results were presented in a numerical format with a reference range, whereas the perceived risk was lowest when horizontal line bars with gradient colors were used [22].

Cognitive Perception

In all 10 studies assessing this outcome, participants consistently demonstrated the ability to understand or identify out-of-range results. There was consensus among these studies that presenting numbers with a reference range only was insufficient and that tailored information was needed [45,52,53]. A qualitative study revealed that participants preferred the inclusion of evaluative labels [43]. In 2 studies using horizontal line bars as the presentation format, the understanding was significantly increased when color, text, or personalized information (eg, goal range) was added [23,49].

Perception of Communication

The majority of included studies observed a significant association between presentation format, participant satisfaction, and ease of use. In general, satisfaction and ease of use were rated higher when test results were presented using horizontal line bars with colored blocks, as compared with other presentation formats [22,23,42,43,47,51,53]. In one qualitative study, numerical presentation with reference ranges was deemed insufficient, while graphs were considered too complex for easy comprehension [43]. Both quantitative and qualitative studies demonstrated that adding evaluative labels, such as explanations about the meaning and normality of test results, and background information about testing, enhanced understanding and effective use of results. Particularly, the use of lay terms played an important role [15,23,44,45,48,51-53]. Furthermore, 2 studies found a significant positive effect on participant satisfaction when personalized information or goal ranges were incorporated [23,51]. This addition was also recommended by 2 qualitative studies [43,53]. Zikmund-Fisher et al [54] specifically studied different types of horizontal line bars and showed no significant differences in participants’ preferences among the studied formats.

Decision

The behavioral intention was assessed in 10 studies, with varying focuses among them. Some authors examined whether participants would contact their physician [7,22,48,49,54], while others inquired about participants seeking additional web-based information [41,45,48], or making lifestyle changes after reviewing lab results [47,48,51].

Two studies demonstrated that the presentation format did not significantly influence participants’ need to contact their health care provider [7,49]. Conversely, Zikmund-Fisher et al [22,54] demonstrated in 2 studies that participants who viewed near-normal results in a numerical format were significantly more likely to contact their doctor compared with those viewing the same results in one of the horizontal line formats. The use of harm anchors in horizontal line bars substantially reduced the number of participants who would want to contact their physician [22,54].

Participants’ tendency to seek web-based information was significantly influenced by the presentation format, with a significantly higher inclination observed for the numerical format compared with the textual format [41]. Kelman et al [45] and Nystrom et al [48] similarly found that approximately half of the participants would look for additional information after receiving test results in numerical format with reference ranges and evaluative labels, or horizontal line bars with colored blocks,

XSL•FO
RenderX

respectively. However, no comparison was made between presentation formats in these studies [45,48].

Intention to make lifestyle changes after viewing laboratory results was mentioned as an outcome in 3 studies [47,48,51]. Only one of these studies compared several presentation formats but found no significant differences between using a numerical format, horizontal line bars with colored blocks, or video-enhanced format in terms of health-beneficial intentions [47].

Action

There was limited data concerning the actions patients take to comprehend their test results. One mixed methods study used a numerical format with reference ranges as a presentation format [15]. Participants with abnormal test results were significantly more likely to take action compared with those with normal test results. As no comparison between presentation formats was investigated, the effect of format on action remains unstudied.

Memory

Variation in the presentation format of test results, using either a numerical format with reference ranges and evaluative labels, horizontal line bars with colored blocks, video presentation, or grouped presentation, did not significantly impact participant recall [7,42,47]. However, one study found a small but statistically significant effect of test result normality on memory [47].

Struikman et al [50] looked at patient health engagement (PHE), a composite measure comprising affective perception, cognitive perception, and behavioral intention. When test results were presented with explanatory text and visualization, PHE was significantly higher compared with when no explanatory information was provided [50].

Discussion

Principal Findings

Based on reviewing 18 articles assessing various presentation formats of laboratory test results, we can conclude there is not only one optimal presentation format in terms of patients' perception, decision, action, and memory. Nevertheless, the results do indicate that presentation format is important for patients' information processing.

Presentation formats differed between articles, but numerical values with reference ranges or horizontal line bars with colored blocks were most commonly used. All included studies investigated perception as an outcome measure, most frequently perception of communication (n=14). Patients' cognitive perception and perception of communication improved when results were presented using horizontal line bars accompanied with colored blocks and evaluative labels or textual information. Incorporation of reference ranges or personalized goal ranges further enhanced patients' perception levels. Using horizontal line bars with harm anchors significantly reduced the number of participants who would want to contact their physician compared with using a numerical format. Furthermore, using the numerical format significantly increased participants' tendency to search for web-based information, compared with a textual format. Therefore, although no specific format is dissuaded in the included studies, the results suggest that presenting only numbers with reference ranges is suboptimal. Furthermore, adding too many colors and other information to test results could lead to an overload of visual information for some patients, and therefore ultimately decrease the amount of usable knowledge [49]. Action and memory were less frequently studied, respectively in 1 and 3 studies. Action was studied in a descriptive study not comparing different presentation formats, while memory was not significantly impacted by presentation format.

Several studies highlighted that patients' affective perception, action, and memory were not only influenced by presentation format, but also by whether test results were within or outside the reference range. Presentation format appeared to be secondary to test result normality if results were extremely out-of-range. Nevertheless, when results were near-normal, presentation format was more important than result normality regarding effects on patients' information processing.

Overall, the results of this review indicate that presentation format affects patients' information processing, especially in the case of normal or near-normal test results.

Strengths and Limitations

A multidisciplinary team of general practitioners, behavioral scientists, and clinical chemists was involved in this review, which is one of its strengths. Both presentation formats and outcomes used in the included studies were standardized by the authors using a published taxonomy to enable comparison of different studies. As the results of our review are narrative, there is a potential risk of bias when describing them, introduced by the authors. Furthermore, selection bias arising from the heterogeneity of studies represents a notable limitation of this review.

A limitation of the included studies is the use of volunteers or participants recruited via convenience sampling. Only 3 out of 18 studies used real patients, of which one study used real test results. Sixteen studies used mock test results. Displaying mock data is common practice in system evaluation. This method involves less burden and privacy risks for participants, as no personal medical data are collected. Nonetheless, participants lack personal relevance of test results when hypothetical scenarios are used. Therefore, it is possible that most of the included studies did not reflect how participants would respond in real life to their own personal health information. This may limit the generalizability of the findings. However, using personal test results could have negatively affected the comparability between studies, as each participant would have encountered different data.

Among the articles reporting educational level, 48% (5676/11,813) of the participants reported a higher education level, which is higher than in the general population. This may limit the generalizability of the findings to the overall population. Another limitation is the study heterogeneity. Included articles varied widely in methods, presentation formats, and outcome measures used. Comparison of presentation formats

XSL•FO
RenderX

is challenging, especially since laboratory test result communication can have a wide range of possible purposes, from interpreting one single value to identifying important trends on time [24]. Therefore, useful presentation formats may vary per clinical scenario, which presents new challenges for designing a preferred format.

As stated above, patients have to complete several steps to go from data perception to usable knowledge [17,32]. The majority of the included studies studied the first 2 steps of this taxonomy, perception and decision. Only one study examined action as outcome measure, and 3 studies obtained information about memory. Therefore, little is known about the impact of presentation formats on actual health behavior and usable knowledge.

Comparison With Prior Work

An increasing number of patients can directly access their laboratory test results via web; thus, it is becoming more important to make the available data meaningful to laypeople [55]. As highlighted in this review, presentation format affects patients' information processing as described above. In cognitive science, this principle is generally known as information evaluability, in other words using relevant contextual reference information to make it easier to evaluate the meaning of in this case numerical laboratory test results (eg, is this test result good or bad, is it normal or abnormal) [56,57]. The presentation formats for laboratory test results as studied in this review could be considered as different forms of contextual information, or evaluative categories [58]. Prior research has shown that these evaluative categories add both affective and cognitive meaning to numeric test results. This enhances patients' information processing by adding meaning and evaluability to numeric information [58-60]. Furthermore, our findings are in line with recommendations made by Witteman and Zikmund-Fisher [17]. The authors formulated 10 recommendations to communicate laboratory test results via web-based portals in ways that support understanding and actionable knowledge for patients. Our findings align with several of their recommendations, such as the importance of providing a clear takeaway message for each result, establishing thresholds for concern and action whenever feasible, and personalizing the frame of reference by permitting custom reference ranges.

This review explored different strategies to improve patients' interpretation and comprehension of their laboratory test results. The included studies predominantly focused on the effect of graphical presentation only including a subset of the available visualization options. Other formats such as clocks or pie charts been shown in the broader numeracy literature to improve cognitive outcomes and could be the focus of further research in the context of communicating laboratory test results [61]. Graphical presentation formats might mitigate the effects of low numeracy. However, it is important to acknowledge that graphical information may not be automatically useful for individuals with limited graph literacy [62]. Besides numeracy and graph literacy, other factors such as age, educational level, health literacy, and statistical literacy (eg, understanding of concepts of uncertainty and chance) also influence patients' information processing of such graphical results [61-63]. If one of these factors causes patients to not completely understand a specific presentation format, they may consider this format as not suitable. Therefore, some patients may require extra instructions to be able to adequately process and interpret graphical presentation formats [61]. For that reason, the interaction between patients' literacy, numeracy, age, and educational level should be taken into account when performing future work around test result interpretation.

Several initiatives aim to inform and educate patients about laboratory test results while incorporating the insights described above. One example is Lab Tests Online, a website that provides patients with general information about laboratory tests and their meaning [64]. Recently, the usability of ChatGPT (ie, an upcoming tool based on natural language processing) to interpret laboratory test results were examined [65]. ChatGPT appeared to provide somewhat superficial interpretations, which were not always correct, and is therefore not yet usable as a primary information source for patients. However, this may change in the future with the further development of these types of tools. While our review focused on different presentation formats of laboratory test results, interpretative comments provided by laboratory specialists were not studied. Laboratory specialists often add comments to test results to assist general practitioners [66,67]. A pilot study by Verboeket-van de Venne et al [68] demonstrated a positive impact on patient empowerment when patients had access to these patient-specific comments. Therefore, further research should explore the impact of adding interpretative comments to laboratory test results on patients' information processing.

Patients now have web-based access to not only their laboratory test results but also to medical imaging and microbiology results. Given the variations in these types of diagnostic test results, further research is warranted to explore effective components for communicating these other types of test results to patients in their web-based patient portal.

Conclusions

As patients increasingly receive their diagnostic laboratory test results via web-based patient portals, it is becoming more and more important to make test results meaningful to them. Unnecessary confusion or anxiety should be avoided, especially when test results are outside of the reference range. The data from our systematic review suggest that horizontal line bars with colored blocks and reference ranges or personalized goal range increase patients' cognitive perception and perception of communication. Furthermore, this format might reduce patients' concerns and their tendency to contact their physicians. Therefore, to improve patients' understanding of near-normal laboratory test results and prevent anxiety and concerns after viewing these results, implementing horizontal line bars with colored blocks and reference ranges or personalized goal ranges in web-based patient portals would be a prudent choice. Our review highlights the importance of taking end users (ie, patients) into consideration when designing new presentation formats. These results can guide the development and improvement of (new) web-based patient portals. Nevertheless, there is a need for further research that involves more comprehensive data collection and reporting, as well as more

XSL•FO
RenderX

systematic evaluation methods. By using these findings, further research could inform the development of an interpretation support tool for laboratory test results.

Acknowledgments

This study received funding from a partnership between Care Research Netherlands and the Medical Sciences domain of the Dutch Research Council (Dutch abbreviation ZonMW; project 08391052120002). ZonMW had no direct involvement in any facet of the design of the study, analysis or interpretation of the data, or writing of the manuscript.

Authors' Contributions

All authors contributed to the conception of the review. FM and FS conducted the screening and data extraction. FM and FS drafted the manuscript. JJ, JB, SM, and JC gave guidance throughout the whole research process. All authors critically revised and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 23 KB - jmir_v26i1e53993_app1.docx]

Multimedia Appendix 2
Search strategy for PubMed, Web of Science, and Embase up to May 31, 2023.
[DOCX File , 23 KB - jmir_v26i1e53993_app2.docx]

Multimedia Appendix 3
Detailed overview of laboratory test results presentation formats in all included studies (n=18).
[DOCX File , 15 KB - jmir_v26i1e53993_app3.docx]

Multimedia Appendix 4
Detailed overview of quality assessment of all included studies (n=18).
[DOCX File , 17 KB - jmir_v26i1e53993_app4.docx]

References

1. Kaelber DC, Jha AK, Johnston D, Middleton B, Bates DW. A research agenda for personal health records (PHRs). J Am Med Inform Assoc 2008;15(6):729-736 [FREE Full text] [doi: 10.1197/jamia.M2547] [Medline: 18756002]
2. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
3. Ancker JS, Osorio SN, Cheriff A, Cole CL, Silver M, Kaushal R. Patient activation and use of an electronic patient portal. Inform Health Soc Care 2015;40(3):254-266. [doi: 10.3109/17538157.2014.908200] [Medline: 24786648]
4. Lee CI, Langlotz CP, Elmore JG. Implications of direct patient online access to radiology reports through patient web portals. J Am Coll Radiol 2016;13(12 Pt B):1608-1614. [doi: 10.1016/j.jacr.2016.09.007] [Medline: 27888949]
5. Fisher B, Bhavnani V, Winfield M. How patients use access to their full health records: a qualitative study of patients in general practice. J R Soc Med 2009;102(12):539-544 [FREE Full text] [doi: 10.1258/jrsm.2009.090328] [Medline: 19966130]
6. Bhavnani V, Fisher B, Winfield M, Seed P. How patients use access to their electronic GP record--a quantitative study. Fam Pract 2011;28(2):188-194. [doi: 10.1093/fampra/cmq092] [Medline: 21084568]
7. Fraccaro P, Vigo M, Balatsoukas P, van der Veer SN, Hassan L, Williams R, et al. Presentation of laboratory test results in patient portals: influence of interface design on risk interpretation and visual search behaviour. BMC Med Inform Decis Mak 2018;18(1):11 [FREE Full text] [doi: 10.1186/s12911-018-0589-7] [Medline: 29433495]
8. Parker RM, Ratzan SC, Lurie N. Health literacy: a policy challenge for advancing high-quality health care. Health Aff (Millwood) 2003;22(4):147-153. [doi: 10.1377/hlthaff.22.4.147] [Medline: 12889762]
9. Garrido T, Jamieson L, Zhou Y, Wiesenthal A, Liang L. Effect of electronic health records in ambulatory care: retrospective, serial, cross sectional study. BMJ 2005;330(7491):581 [FREE Full text] [doi: 10.1136/bmj.330.7491.581] [Medline: 15760999]
10. Krist AH, Woolf SH. A vision for patient-centered health information systems. JAMA 2011;305(3):300-301 [FREE Full text] [doi: 10.1001/jama.2010.2011] [Medline: 21245186]

XSL•FO
RenderX

11. Pillemer F, Price RA, Paone S, Martich GD, Albert S, Haidari L, et al. Direct release of test results to patients increases patient engagement and utilization of care. PLoS One 2016;11(6):e0154743 [FREE Full text] [doi: 10.1371/journal.pone.0154743] [Medline: 27337092]
12. Sung S, Forman-Hoffman V, Wilson MC, Cram P. Direct reporting of laboratory test results to patients by mail to enhance patient safety. J Gen Intern Med 2006;21(10):1075-1078 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00553.x] [Medline: 16836627]
13. Giardina TD, Baldwin J, Nystrom DT, Sittig DF, Singh H. Patient perceptions of receiving test results via online portals: a mixed-methods study. J Am Med Inform Assoc 2018;25(4):440-446 [FREE Full text] [doi: 10.1093/jamia/ocx140] [Medline: 29240899]
14. Palen TE, Ross C, Powers JD, Xu S. Association of online patient access to clinicians and medical records with use of clinical services. JAMA 2012;308(19):2012-2019. [doi: 10.1001/jama.2012.14126] [Medline: 23168824]
15. Zhang Z, Citardi D, Xing A, Luo X, Lu Y, He Z. Patient challenges and needs in comprehending laboratory test results: mixed methods study. J Med Internet Res 2020;22(12):e18725 [FREE Full text] [doi: 10.2196/18725] [Medline: 33284117]
16. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
17. Witteman HO, Zikmund-Fisher BJ. Communicating laboratory results to patients and families. Clin Chem Lab Med 2019;57(3):359-364 [FREE Full text] [doi: 10.1515/cclm-2018-0634] [Medline: 30407910]
18. Zikmund-Fisher BJ, Exe NL, Witteman HO. Numeracy and literacy independently predict patients' ability to identify out-of-range test results. J Med Internet Res 2014;16(8):e187 [FREE Full text] [doi: 10.2196/jmir.3241] [Medline: 25135688]
19. Keselman A, Smith CA. A classification of errors in lay comprehension of medical documents. J Biomed Inform 2012;45(6):1151-1163 [FREE Full text] [doi: 10.1016/j.jbi.2012.07.012] [Medline: 22925723]
20. McDonald KM, Bryce CL, Graber ML. The patient is in: patient involvement strategies for diagnostic error mitigation. BMJ Qual Saf 2013;22 Suppl 2(Suppl 2):ii33-ii39 [FREE Full text] [doi: 10.1136/bmjqs-2012-001623] [Medline: 23893394]
21. Britto MT, Jimison HB, Munafo JK, Wissman J, Rogers ML, Hersh W. Usability testing finds problems for novice users of pediatric portals. J Am Med Inform Assoc 2009;16(5):660-669 [FREE Full text] [doi: 10.1197/jamia.M3154] [Medline: 19567793]
22. Zikmund-Fisher BJ, Scherer AM, Witteman HO, Solomon JB, Exe NL, Tarini BA, et al. Graphics help patients distinguish between urgent and non-urgent deviations in laboratory test results. J Am Med Inform Assoc 2017;24(3):520-528 [FREE Full text] [doi: 10.1093/jamia/ocw169] [Medline: 28040686]
23. Tao D, Yuan J, Qu X. Presenting self-monitoring test results for consumers: the effects of graphical formats and age. J Am Med Inform Assoc 2018;25(8):1036-1046 [FREE Full text] [doi: 10.1093/jamia/ocy046] [Medline: 29762686]
24. Turchioe MR, Myers A, Isaac S, Baik D, Grossman LV, Ancker JS, et al. A systematic review of patient-facing visualizations of personal health data. Appl Clin Inform 2019;10(4):751-770 [FREE Full text] [doi: 10.1055/s-0039-1697592] [Medline: 31597182]
25. O'Kane M, Freedman D, Zikmund-Fisher BJ. Can patients use test results effectively if they have direct access? BMJ 2015;350:h673. [doi: 10.1136/bmj.h673] [Medline: 25673132]
26. Zikmund-Fisher BJ, Witteman HO, Dickson M, Fuhrel-Forbis A, Kahn VC, Exe NL, et al. Blocks, ovals, or people? icon type affects risk perceptions and recall of pictographs. Med Decis Making 2014;34(4):443-453 [FREE Full text] [doi: 10.1177/0272989X13511706] [Medline: 24246564]
27. Watson ID. Making test results more easily understood by patients. BMJ 2015;350:h1942. [doi: 10.1136/bmj.h1942] [Medline: 25873357]
28. Garcia-Retamero R, Cokely ET. Designing visual aids that promote risk literacy: a systematic review of health research and evidence-based design heuristics. Hum Factors 2017;59(4):582-627. [doi: 10.1177/0018720817690634] [Medline: 28192674]
29. Torsvik T, Lillebo B, Mikkelsen G. Presentation of clinical laboratory results: an experimental comparison of four visualization techniques. J Am Med Inform Assoc 2013;20(2):325-331 [FREE Full text] [doi: 10.1136/amiajnl-2012-001147] [Medline: 23043123]
30. Boisot M, Canals A. Data, information and knowledge: have we got it right? J. Evol. Econ 2004;14:43-67. [doi: 10.1007/s00191-003-0181-9]
31. Chen M, Ebert D, Hagen H, Laramee RS, van Liere R, Ma KL, et al. Data, information, and knowledge in visualization. IEEE Comput Graph Appl 2009;29(1):12-19 [FREE Full text] [doi: 10.1109/mcg.2009.6] [Medline: 19363954]
32. Ancker JS, Benda NC, Sharma MM, Johnson SB, Weiner S, Zikmund‐Fisher BJ. Taxonomies for synthesizing the evidence on communicating numbers in health: goals, format, and structure. Risk Analysis 2022;42(12):2656-2670. [doi: 10.1111/risa.13875]
33. Wickens CD, Helton W, Hollands JG, Parasuraman R, Banbury S. Engineering Psychology & Human Performance. United Kingdom: Routledge; 2021.
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

35. Covidence Systematic Review Software, Veritas Health Innovation, Melbourne, Australia. URL: https://www.covidence.org/ [accessed 2023-08-25]
36. Becker MH. The health belief model and sick role behavior. Health Education Monographs 1974;2(4):324-508. [doi: 10.1177/109019817400200407]
37. Witte K. Putting the fear back into fear appeals: the extended parallel process model. Commun Monogr 1992;59(4):329-349. [doi: 10.1080/03637759209376276]
38. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
39. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot mixed methods appraisal tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
40. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
41. Bar-Lev S, Beimel D. Numbers, graphs and words - do we really understand the lab test results accessible via the patient portals? Isr J Health Policy Res 2020;9(1):58 [FREE Full text] [doi: 10.1186/s13584-020-00415-z] [Medline: 33115536]
42. Brewer NT, Gilkey MB, Lillie SE, Hesse BW, Sheridan SL. Tables or bar graphs? presenting test results in electronic medical records. Med Decis Making 2012;32(4):545-553 [FREE Full text] [doi: 10.1177/0272989X12441395] [Medline: 22472914]
43. Elder NC, Barney K. "But what does it mean for me?" primary care patients' communication preferences for test results notification. Jt Comm J Qual Patient Saf 2012;38(4):168-176. [doi: 10.1016/s1553-7250(12)38022-7] [Medline: 22533129]
44. Hohenstein JC, Baumer EP, Reynolds L, Murnane EL, O'Dell D, Lee S, et al. Supporting accurate interpretation of self-administered medical test results for mobile health: assessment of design, demographics, and health condition. JMIR Hum Factors 2018;5(1):e9 [FREE Full text] [doi: 10.2196/humanfactors.8620] [Medline: 29490894]
45. Kelman A, Robinson CO, Cochin E, Ahluwalia NJ, Braverman J, Chiauzzi E, et al. Communicating laboratory test results for rheumatoid factor: what do patients and physicians want? Patient Prefer Adherence 2016;10:2501-2517 [FREE Full text] [doi: 10.2147/PPA.S104396] [Medline: 28008236]
46. Morrow D, Hasegawa-Johnson M, Huang T, Schuh W, Azevedo RFL, Gu K, et al. A multidisciplinary approach to designing and evaluating electronic medical record portal messages that support patient self-care. J Biomed Inform 2017;69:63-74 [FREE Full text] [doi: 10.1016/j.jbi.2017.03.015] [Medline: 28347856]
47. Morrow D, Azevedo RFL, Garcia-Retamero R, Hasegawa-Johnson M, Huang T, Schuh W, et al. Contextualizing numeric clinical test results for gist comprehension: implications for EHR patient portals. J Exp Psychol Appl 2019;25(1):41-61. [doi: 10.1037/xap0000203] [Medline: 30688498]
48. Nystrom DT, Singh H, Baldwin J, Sittig DF, Giardina TD. Methods for patient-centered interface design of test result display in online portals. EGEMS (Wash DC) 2018;6(1):15 [FREE Full text] [doi: 10.5334/egems.255] [Medline: 30094287]
49. Scherer AM, Witteman HO, Solomon J, Exe NL, Fagerlin A, Zikmund-Fisher BJ. Improving the understanding of test results by substituting (not adding) goal ranges: web-based between-subjects experiment. J Med Internet Res 2018;20(10):e11027 [FREE Full text] [doi: 10.2196/11027] [Medline: 30341053]
50. Struikman B, Bol N, Goedhart A, van Weert JCM, Talboom-Kamp E, van Delft S, et al. Features of a patient portal for blood test results and patient health engagement: web-based pre-post experiment. J Med Internet Res 2020;22(7):e15798 [FREE Full text] [doi: 10.2196/15798] [Medline: 32706704]
51. Talboom-Kamp E, Tossaint-Schoenmakers R, Goedhart A, Versluis A, Kasteleyn M. Patients' attitudes toward an online patient portal for communicating laboratory test results: real-world study using the eHealth impact questionnaire. JMIR Form Res 2020;4(3):e17060 [FREE Full text] [doi: 10.2196/17060] [Medline: 32024632]
52. Zarcadoolas C, Vaughon WL, Czaja SJ, Levy J, Rockoff ML. Consumers' perceptions of patient-accessible electronic medical records. J Med Internet Res 2013;15(8):e168 [FREE Full text] [doi: 10.2196/jmir.2507] [Medline: 23978618]
53. Zhang Z, Kmoth L, Luo X, He Z. User-centered system design for communicating clinical laboratory test results: design and evaluation study. JMIR Hum Factors 2021;8(4):e26017 [FREE Full text] [doi: 10.2196/26017] [Medline: 34842529]
54. Zikmund-Fisher BJ, Scherer AM, Witteman HO, Solomon JB, Exe NL, Fagerlin A. Effect of harm anchors in visual displays of test results on patient perceptions of urgency about near-normal values: experimental study. J Med Internet Res 2018;20(3):e98 [FREE Full text] [doi: 10.2196/jmir.8889] [Medline: 29581088]
55. Zikmund-Fisher BJ. Helping people know whether measurements have good or bad implications: increasing the evaluability of health and science data communications. PIBBS 2019;6(1):29-37. [doi: 10.1177/2372732218813377]
56. Hsee CK. The evaluability hypothesis: an explanation for preference reversals between joint and separate evaluations of alternatives. Organ Behav Hum Decis Process 1996;67(3):247-257. [doi: 10.1006/obhd.1996.0077]
57. Hsee CK, Zhang J. General evaluability theory. Perspect Psychol Sci 2010;5(4):343-355. [doi: 10.1177/1745691610374586] [Medline: 26162182]
58. Peters E, Dieckmann NF, Västfjäll D, Mertz CK, Slovic P, Hibbard JH. Bringing meaning to numbers: the impact of evaluative categories on decisions. J Exp Psychol Appl 2009;15(3):213-227. [doi: 10.1037/a0016978] [Medline: 19751072]

XSL•FO
RenderX

59. Zikmund-Fisher BJ, Fagerlin A, Keeton K, Ubel PA. Does labeling prenatal screening test results as negative or positive affect a woman's responses? Am J Obstet Gynecol 2007;197(5):528 [FREE Full text] [doi: 10.1016/j.ajog.2007.03.076] [Medline: 17880902]
60. Peters E. The functions of affect in the construction of preferences. In: The Construction of Preference. New York: Cambridge University Press; 2006:454-463.
61. Ancker JS, Senathirajah Y, Kukafka R, Starren JB. Design features of graphs in health risk communication: a systematic review. J Am Med Inform Assoc 2006;13(6):608-618 [FREE Full text] [doi: 10.1197/jamia.M2115] [Medline: 16929039]
62. van Weert JCM, Alblas MC, van Dijk L, Jansen J. Preference for and understanding of graphs presenting health risk information. the role of age, health literacy, numeracy and graph literacy. Patient Educ Couns 2021;104(1):109-117 [FREE Full text] [doi: 10.1016/j.pec.2020.06.031] [Medline: 32727670]
63. Ancker JS, Kaufman D. Rethinking health numeracy: a multidisciplinary literature review. J Am Med Inform Assoc 2007;14(6):713-721 [FREE Full text] [doi: 10.1197/jamia.M2464] [Medline: 17712082]
64. Campbell B, Linzer G, Dufour DR. Lab tests online and consumer understanding of laboratory testing. Clin Chim Acta 2014;432:162-165. [doi: 10.1016/j.cca.2013.09.028] [Medline: 24121032]
65. Cadamuro J, Cabitza F, Debeljak Z, de Bruyne S, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. an assessment by the european federation of clinical chemistry and laboratory medicine (EFLM) working group on artificial intelligence (WG-AI). Clin Chem Lab Med 2023;61(7):1158-1166 [FREE Full text] [doi: 10.1515/cclm-2023-0355] [Medline: 37083166]
66. Barlow IM. Are biochemistry interpretative comments helpful? results of a general practitioner and nurse practitioner survey. Ann Clin Biochem 2008;45(Pt 1):88-90. [doi: 10.1258/acb.2007.007134] [Medline: 18275680]
67. Verboeket-van de Venne WPHG, Oosterhuis WP, Keuren JFW, Kleinveld HA. Reflective testing in the Netherlands: usefulness to improve the diagnostic and therapeutic process in general practice. Ann Clin Biochem 2009;46(Pt 4):346-347. [doi: 10.1258/acb.2009.009039] [Medline: 19454541]
68. Verboeket-van de Venne WPHG, Hendriks-Dybicz AM, Oosterhuis WP. Patiënten informeren over laboratoriumuitslagen in het kader van patient empowerment [Inform patients about laboratory test results in the context of patient empowerment]. Ned Tijdschr Klin Chem Labgeneesk [Dutch Journal of Clinical Chemistry and Laboratory Medicine]. 2015. URL: https://www.nvkc.nl/files/ntkc/N60_049408_BW_NVKC_Juli2015_WQ_012.pdf [accessed 2023-09-07]

Abbreviations

HbA1c: hemoglobin A1c
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 26.10.23; peer-reviewed by B Zikmund-Fisher, B Steitz; comments to author 26.02.24; revised version received 07.04.24; accepted 27.05.24; published 12.08.24.

Please cite as:
van der Mee FAM, Schaper F, Jansen J, Bons JAP, Meex SJR, Cals JWL
Enhancing Patient Understanding of Laboratory Test Results: Systematic Review of Presentation Formats and Their Impact on Perception, Decision, Action, and Memory
J Med Internet Res 2024;26:e53993
URL: https://www.jmir.org/2024/1/e53993
doi:10.2196/53993
PMID:

©Frederieke A M van der Mee, Fleur Schaper, Jesse Jansen, Judith A P Bons, Steven J R Meex, Jochen W L Cals. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Leadership for AI Transformation in Health Care Organization: Scoping Review

Abi Sriharan[1,2], MSc, DPhil; Nigar Sekercioglu[2], PhD; Cheryl Mitchell[3], PhD; Senthujan Senkaiahliyan[2], MHSc; Attila Hertelendy[4], PhD; Tracy Porter[5], PhD; Jane Banaszak-Holl[6], PhD

[1]Krembil Centre for Health Management and Leadership, Schulich School of Business, York University, Toronto, ON, Canada
[2]Institute for Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[3]Gustavson School of Business, University of Victoria, Victoria, ON, Canada
[4]College of Business, Florida International University, Florida, FL, United States
[5]Department of Management, Cleveland State University, Cleveland, OH, United States
[6]Department of Health Services Administration, School of Health Professions, University of Alabama Birmingham, Birmingham, OH, United States

Corresponding Author:
Abi Sriharan, MSc, DPhil
Krembil Centre for Health Management and Leadership
Schulich School of Business
York University
MB Room G315
4700 Keele St
Toronto, ON, M3J 1P3
Canada
Phone: 1 3658855898
Email: abisri@yorku.ca

Abstract

Background: The leaders of health care organizations are grappling with rising expenses and surging demands for health services. In response, they are increasingly embracing artificial intelligence (AI) technologies to improve patient care delivery, alleviate operational burdens, and efficiently improve health care safety and quality.

Objective: In this paper, we map the current literature and synthesize insights on the role of leadership in driving AI transformation within health care organizations.

Methods: We conducted a comprehensive search across several databases, including MEDLINE (via Ovid), PsycINFO (via Ovid), CINAHL (via EBSCO), Business Source Premier (via EBSCO), and Canadian Business & Current Affairs (via ProQuest), spanning articles published from 2015 to June 2023 discussing AI transformation within the health care sector. Specifically, we focused on empirical studies with a particular emphasis on leadership. We used an inductive, thematic analysis approach to qualitatively map the evidence. The findings were reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews) guidelines.

Results: A comprehensive review of 2813 unique abstracts led to the retrieval of 97 full-text articles, with 22 included for detailed assessment. Our literature mapping reveals that successful AI integration within healthcare organizations requires leadership engagement across technological, strategic, operational, and organizational domains. Leaders must demonstrate a blend of technical expertise, adaptive strategies, and strong interpersonal skills to navigate the dynamic healthcare landscape shaped by complex regulatory, technological, and organizational factors.

Conclusions: In conclusion, leading AI transformation in healthcare requires a multidimensional approach, with leadership across technological, strategic, operational, and organizational domains. Organizations should implement a comprehensive leadership development strategy, including targeted training and cross-functional collaboration, to equip leaders with the skills needed for AI integration. Additionally, when upskilling or recruiting AI talent, priority should be given to individuals with a strong mix of technical expertise, adaptive capacity, and interpersonal acumen, enabling them to navigate the unique complexities of the healthcare environment.

(J Med Internet Res 2024;26:e54556) doi:10.2196/54556

XSL•FO
RenderX

KEYWORDS

AI implementation; innovation; health care; leadership; AI; artificial intelligence; management; organization; health care organization; strategy

Introduction

Artificial Intelligence in Health Care: Overview

Artificial intelligence (AI) technologies have gained significant momentum in health care, presenting a transformative potential across clinical processes, operational efficiency, decision-making, and workforce optimization [1-3]. The global AI market is projected to shift from US $14.6 billion in 2023 to a formidable estimate of US $102.7 billion by 2028 [4], unveiling a dynamic transformation of unprecedented scale. This investment, coupled with the engagement of nontraditional health care players such as Microsoft, Google, and Amazon and the convergence of technological prowess and health care innovation signaled by generative AI, will place the trajectory of AI in health care in a state of exponential growth [5].

Current investments in health care AI predominantly center on bolstering data capacity, enhancing computational power, and advancing methodological innovations in AI. This includes developing and testing AI models and algorithms tailored for precision medicine, drug discovery, clinical decision-making support, public health surveillance, operational optimization, and process improvement [6,7]. Notably, between August 2022 and July 2023, there were over 150 submissions of drug and biological applications incorporating AI and machine learning components to the US Food and Drug Administration, encompassing a wide array of therapeutic domains and developmental stages [8].

Yet the seamless integration of AI technologies into health care organizational settings presents a multifaceted challenge for health care leaders. This challenge arises from several factors, including the complex nature of AI models, the rapid pace of technological advancement, the imperative of regulatory adherence, ethical concerns surrounding data security and privacy, the risk of perpetuating racial and ethnic biases in data, the necessity of prioritizing human-centric approaches to patient care, and the intricate clinical workflows that must be navigated [9-15]. Furthermore, health care leaders are facing critical and intricate strategic decisions. They must discern which AI solutions merit investment while weighing the merits of in-house development against strategic partnerships with external vendors. Selecting the right vendors and defining the scope of collaboration is pivotal, as is devising a sustainable funding strategy to support both initial development and continuous innovation. Furthermore, they must confront the crucial question of whether to bring in new AI talent or bolster the expertise of their current workforce through upskilling. Each of these decisions will shape the trajectory of health care organizations as they navigate this transformative era. A report by Bain in 2023 revealed that although 75% of surveyed health system executives recognize AI's potential to reshape the health care industry, only 6% have established concrete strategies related to AI [16].

The lack of strategy and strategic failures in AI integration not only have financial consequences for organizations but also erode trust among patients, providers, and organizations [17]. A prominent example is the collaboration between MD Anderson and IBM Watson, aimed at leveraging IBM Watson's cognitive capabilities to combat cancer. This ambitious endeavor, however, incurred a substantial financial toll of over US $62 million for MD Anderson because of setbacks in clinical implementation [18].

Despite a growing body of AI literature, including toolkits such as Canada Health Infoway's "Toolkit for AI Implementers" [19] and guidance from the US Department of Health and Human Services' AI Task Force [20] and the UK National Strategy for AI in Health and Social Care [21], there is still insufficient scholarly attention on how leadership behavior guides AI transformation in health care. Existing reviews focus on AI in medical education [22,23], workforce impact [24], applications in clinical medicine [13,25], barriers to implementation [26,27], and ethical considerations [28,29]. However, no systematic mapping of empirical literature has clarified our understanding of leadership or identified gaps in research. Understanding leadership behavior is crucial for health care organizations considering AI because effective leadership shapes the strategic direction, adoption, and successful implementation of AI technologies.

Research Aim

To address this research gap and to establish a future research agenda this scoping review study aims to address two primary questions: (1) What role does leadership play in AI transformation within health care? and (2) What approaches can health care organizations use to empower their leaders in facilitating AI transformation?

Methods

Research Approach

This review follows scoping review methodology [30] to identify and analyze the current literature and report results following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews; Multimedia Appendix 1) guidelines [31].

Key Definitions

In the context of this study, AI refers to combination of machine learning algorithms, large language models, robotics, and natural language processing systems designed to mimic human cognitive functions, enabling machines to perform tasks autonomously or with minimal human intervention.

AI transformation refers to the systematic changes in clinical, operational, or organizational processes and business models due to the introduction of AI systems to optimize decision-making, automate tasks, improve patient outcomes, and drive organizational change. This involves identifying

XSL•FO
RenderX

opportunities for AI-related innovation, integrating them into processes, and developing strategies to operationalize implementation while ensuring organizational readiness. This is essential for getting health care organizations AI-ready.

Further, in the context of this study, drawing from seminal management and leadership theories, we view leadership as an effective management practice [32]. However, we recognize that leadership roles in health care occur at the clinical, organizational, and systems levels of health systems. At the clinical level, leadership emerges through health care professionals who steer patient care and treatment decisions. At the organizational level, leadership involves middle managers such as unit heads and division leaders guiding health care institutions, administrative units, and personnel toward their goals. At the systems level, leadership encapsulates C-suite leadership responsible for navigating regulatory complexities and organizational and structural silos within complex health systems.

Eligibility Criteria

The following inclusion and exclusion criteria guided our study: (1) focused on AI in health care, (2) contained an evaluation of leadership, (3) were written in English, (4) were published in a peer-reviewed journal, (5) published between January 2015 and June 2023, and (6) used research.

Information Sources and Search Strategy

We adopted comprehensive search strategies for the following electronic databases focused on the health care and business literature: MEDLINE (via Ovid), PsycINFO (via Ovid), CINAHL (via EBSCO), Business Source Premier (via EBSCO), and Canadian Business & Current Affairs (via ProQuest). An academic librarian developed these search strategies with input from the research team. We initially conducted the search in Ovid MEDLINE. We then reviewed our search results using the Peer Review of Electronic Search Strategies tool [33], a checklist for comparing, among other things, the types of errors found in articles and the relative fit of articles to the research question before translating the search strategy into other databases using their command language. Our search was limited to articles published from January 2015 (from the first use of AI-powered chatbots in health care [34] to June 2023. We then ran searches in 4 databases and exported the final search results into the EndNote reference management software (Clarivate), and we removed duplicate articles manually. To capture any papers that may have been missed during the search process, we did forward and reverse citation searches of systematic review articles related to AI [35]. However, we did not find any additional articles that met our criteria. Finally, we imported search results to Covidence (Veritas Health Innovation), a review management software for abstract and title screening, full-text screening, and data charting.

Selection of Sources of Evidence and Data Charting

To minimize selection bias, 2 independent screeners reviewed the titles and abstracts of articles identified via the search against the eligibility criteria using Covidence. We identified articles that met the eligibility criteria for a comprehensive full-text screening. Two independent reviewers then evaluated the full texts against the eligibility criteria using Covidence. In discrepancies between the reviewers, a third reviewer served as the consensus reviewer and used Covidence to resolve conflicts between reviewer 1 and reviewer 2. Following the exclusion of irrelevant articles, we used a predefined data extraction form aligned with our research objectives and guiding questions for systematic data collection. Data extraction categories included data on study characteristics (eg, citations and country); methods (eg, aim, data collection methods, and methodological quality); study context (eg, leadership role, ie, clinical, organizational, or systems); leadership practices (ie, behavior, enablers, and barriers to leadership success); results (ie, main results and author conclusion); and an open-ended reviewer note (ie, capture any relevant information that might aid in the data analysis stage). The data abstraction form was piloted on a random sample of 4 included articles and modified based on feedback from the team. Full data abstraction began only after sufficient agreement had been obtained. Two reviewers independently extracted the data using Covidence, and a third reviewer assessed the data extraction for quality and consensus. Three authors then held a group discussion to resolve any conflicts.

Risk of Bias Assessment

The focus of scoping reviews is to provide a comprehensive overview of the available literature, identifying the extent, range, and nature of research on a particular topic rather than assessing the methodological quality of individual studies [35]. Therefore, we did not perform risk of bias evaluations on the articles included in compliance with the guidelines for scoping reviews.

Data Analysis and Synthesis

Our data analysis was guided by a thematic analysis process [36]. To ensure the accuracy of the emerging themes, we conducted our analysis collaboratively in reviewer pairs [35].

We initially analyzed the extracted data using an open-coding method guided by our research questions. Subsequently, we grouped the codes into categories based on the emerging patterns in the data, which we then synthesized into leadership functional domains, capacities, and context.

In the context of our analysis, functional domains refer to distinct areas of responsibility that a leader must effectively manage a task or a role. Capacity, on the other hand, pertains to the abilities—skills, competencies, or behaviors—that a leader must demonstrate to achieve desired goals. Context refers to the environment, conditions, and situational factors that shape and influence leadership practices and decisions.

Results

Study Selection

As described in Figure 1, the original searches generated 3541 articles published from January 2015 to June 2023. After removing 728 duplicate articles in EndNote, 2813 unique articles were uploaded to Covidence. A total of 2813 relevant studies were then screened using Covidence using the articles' titles and abstracts. We determined that 97 articles met the criteria for a full-text review for eligibility screening. Within these 97

XSL•FO
RenderX

articles, 75 were excluded as they were opinion articles or commentaries without objective data. After conducting the full-text screening, we found that 22 articles met the final inclusion criteria.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

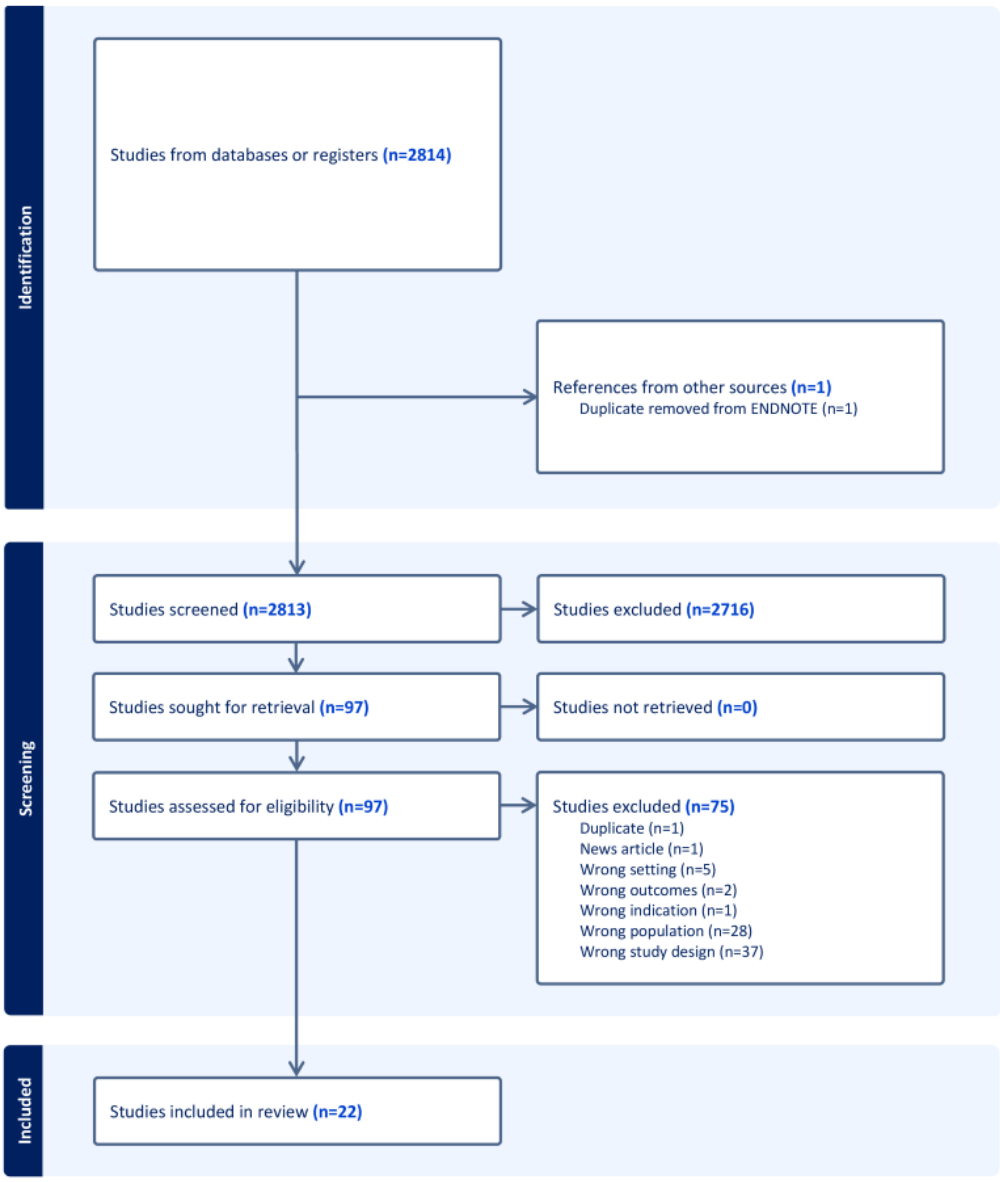

Study Characteristics

Of the 22 studies identified for final inclusion in our review, 12 involved qualitative methods [37-48] such as interviews and case studies, whereas 4 studies involved mixed methods research [49-52] with a qualitative and quantitative strand. There were 3 narrative reports [53-55] based on document synthesis, and 3 studies involved quantitative methods [56-58] such as surveys. These articles focused on clinical, organizational, and systems leadership and came from Canada, China, Finland, Saudi Arabia, Sweden, the Netherlands, the United Kingdom, and the United States. The included papers addressed a broad array of AI applications in health care, including studies focused on improving workflows, quality of care, patient safety, resource optimization, and patient experience. From a clinical domain, researchers focused on primary care, health care systems, radiology, or global health. From a population perspective, the papers covered leadership from the perspective of primary care physicians, radiologists, nurses, nurse managers, public health professionals, global health professionals, health care entrepreneurs, and health care leaders. Table 1 provides a summary of study characteristics.

XSL•FO
RenderX

Table 1. Summary of study characteristics.

Reference	Country	Study context	Leadership level	Theory or framework guiding the research	Study type
Barbour et al [37]	United States	Emergency medicine or medical education	Systems	N/A[a]	Qualitative
Darcel et al [38]	Canada	Primary care	Clinical or systems	Sociotechnological framework	Qualitative
Dicuonzo et al [39]	Canada	Hospital	Organizational or systems	Comprehensive health = technology assessment framework	Qualitative
Dixit et al [53]	Canada	Health care system	Clinical, organizational, or systems	N/A	Narrative report
Ergin et al [58]	Turkey	Nursing	Clinical, organizational, or nursing	N/A	Quantitative
Galsgaard et al [54]	Denmark	Radiology	Clinical	Self-efficacy and professional identity	Narrative report
Ganapathi and Duggal [40]	United Kingdom	Physicians	Clinical	N/A	Qualitative
Gillan [41]	Canada	Radiation medicine and medical imaging technology	Systems or clinical	Normalization Process Theory (NPT)	Qualitative
Hakim et al [49]	Canada	Health care system	Systems or organizational	Health Information and Management Systems Society Adoption Model for Analytics Maturity (AMAM)	Mixed method
Henriksen and Bechmann [42]	Belgium	Technology development	Organizational	Work process and practice-oriented focus	Qualitative
Laukka et al [43]	Finland	Nursing	Organizational, clinical, or nursing	N/A	Qualitative
Li et al [56]	China	Nursing	Organizational, clinical, or nursing	Job Demand-Control-Support (JDCS) model	Quantitative
Morley et al [50]	United Kingdom	Global health	Systems or global health	N/A	Mixed method
Nasseef et al [57]	Saudi Arabia	Health care organization	Systems or public health	Cognitive Fit Theory (CFT)	Quantitative
Olaye and Seixas [44]	United States	Health care startups	Systems or digital health startup	N/A	Qualitative
Petersson et al [45]	Sweden	Health care system	Organizational or systems	N/A	Qualitative
Ronquillo et al [46]	International	Nursing	Systems, clinical, or nursing	N/A	Qualitative
Sawers et al [55]	International	Sustainable development goals—eye health	Systems or global health	N/A	Narrative review
Strohm et al [47]	Netherland	Radiology	Clinical	Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) Framework for new medical technologies in health care organizations.	Qualitative
Upshaw et al [48]	Canada	Primary care	Systems	Sittig and Singh's model for studying Health Information Technology (HIT) in complex adaptive health systems	Qualitative
Willis et al [51]	United Kingdom	Primary care	Clinical	O*NET classification of occupational tasks	Mixed method
Yang et al [52]	China	Hospital	Organizational or systems	Technology-Organization-Environment (TOE) Framework	Mixed method

[a]N/A: not applicable.

XSL•FO
RenderX

Leadership Tasks Essential for AI Transformation in Health Care

We mapped the themes from the included studies across 4 functional domains of leadership task responsibility—technological (AI innovation), strategic (vision and alignment), operational (process and oversight), and organizational (culture and work environment).

The technological functional domain garnered the most significant attention in the literature. The core themes that emerged under the technological domain primarily focused on applying subject matter expertise and AI technical skills to effectively identify AI opportunities, as well as to foster an innovation mindset to develop, tailor, and seamlessly implement AI-driven solutions to address key AI opportunities within health care organizations.

Within the strategic functional domain, the literature underscored the importance of change management and communication as strategic tools for consensus and collaboration related to the AI transformation process. Another core theme that emerged focused on the critical importance of integrating AI solutions into the existing clinical care processes. This strategic alignment is essential for getting support from the staff and ensuring smooth operations of patient care outcomes while embracing the potential of AI solutions. Although the significance of talent strategy related to the recruitment and retention of AI technical expertise within organizations was mentioned, it was not widely seen across the included papers.

Table 2 provides a summary of how the technological and strategic functional domains map across the papers and provides key themes that emerged with the domain area.

Table 2. Leadership competencies related to technical and strategic functional domains.

Reference	Functional domain	Key themes			Functional domain	Key themes		
	Technological	Subject matter expertise	Technical skills	Innovation mindset	Strategic	Change	Communication	Alignment
Barbour et al [37]	✓	✓	✓					
Darcel et al [38]	✓		✓		✓		✓	
Dicuonzo et al [39]	✓		✓		✓		✓	
Dixit et al [53]					✓	✓	✓	
Ergin et al [58]	✓	✓	✓					
Galsgaard et al [54]	✓	✓			✓	✓	✓	
Ganapathi and Duggal [40]	✓	✓	✓		✓			✓
Gillan [41]	✓	✓			✓		✓	
Hakim et al [49]	✓	✓	✓		✓		✓	✓
Henriksen and Bechmann [42]	✓	✓	✓		✓		✓	
Laukka et al [43]	✓	✓						
Li et al [56]	✓			✓	✓	✓		
Morley et al [50]	✓	✓	✓					
Nasseef et al [57]	✓				✓			✓
Olaye and Seixas [44]	✓	✓			✓			✓
Petersson et al [45]	✓		✓		✓		✓	✓
Ronquillo et al [46]	✓	✓	✓		✓	✓		
Sawers et al [55]	✓			✓	✓		✓	
Strohm et al [47]	✓			✓	✓	✓	✓	✓
Upshaw et al [48]	✓	✓	✓					
Willis et al [51]	✓			✓	✓	✓		✓
Yang et al [52]	✓			✓	✓	✓		✓

Emerging evidence in the operational functional domain highlights leaders' need to navigate ethical and risk management issues by establishing robust governance structures prioritizing patient data privacy and security while ethically integrating AI technologies within existing workflows. Additionally, the literature emphasizes that implementing AI in health care will require leaders to ensure new AI solutions comply with existing regulatory and control systems. The literature highlighted that

XSL•FO
RenderX

leaders need to pay attention to process agility through continuous monitoring to ensure AI solutions can adapt to contextual changes.

Finally, the organizational functional domain emerges from the thematic analysis as a pivotal area for AI leadership. The literature emphasizes the importance of stakeholder engagement in building collaboration. Furthermore, it underscores the importance of decision makers' sense-making to enhance their trust in AI opportunities and ensure that AI integration is supported by individuals across the organization. Further, the literature underscored the importance of organizational culture readiness to support physicians and nurses through protected time and incentive pay to engage, innovate, and adopt AI solutions. Table 3 provides a summary of how operational and organizational functional domains map across the papers.

Table 3. Leadership competencies related to operational and organizational functional domains.

Author	Functional domain	Key themes			Functional domain	Key themes		
	Operational	Ethical and risk management	Regulatory compliance	Process agility	Organizational	Stakeholder engagement or collaboration	Trust and sense-making	Organizational culture and readiness
Barbour et al [37]								
Darcel et al [38]	✓	✓	✓		✓	✓		✓
Dicuonzo et al [39]	✓				✓	✓		
Dixit et al [53]	✓	✓				✓		
Ergin et al [58]	✓	✓						
Galsgaard et al [54]					✓	✓		
Ganapathi and Duggal [40]					✓	✓		✓
Gillan [41]	✓	✓		✓	✓	✓	✓	
Hakim et al [49]	✓	✓	✓	✓	✓	✓		
Henriksen and Bechmann [42]					✓	✓	✓	
Laukka et al [43]	✓	✓			✓			
Li et al [56]	✓			✓				
Morley et al [50]	✓	✓	✓		✓	✓		
Nasseef et al [57]	✓	✓				✓		
Olaye and Seixas [44]	✓	✓	✓		✓	✓		
Petersson et al [45]	✓	✓	✓		✓	✓		✓
Ronquillo et al [46]	✓	✓			✓	✓		
Sawers et al [55]	✓	✓	✓		✓	✓		✓
Strohm et al [47]	✓				✓	✓		✓
Upshaw et al [48]	✓	✓						
Willis et al [51]	✓				✓			
Yang et al [52]	✓		✓		✓	✓	✓	

Leadership Skills and Behaviors for Preparing Health Care Organizations for AI Transformation

We categorized the themes related to skills and behaviors into 3 essential capacities that a leader must demonstrate to achieve desired goals—technical capacity, adaptive capacity, and interpersonal capacity. Technical capacity encompasses (1) AI literacy, (2) subject matter knowledge, (3) change leadership skills, and (4) innovation mindset to identify AI innovation opportunities. The interpersonal capacity involves several vital facets such as (1) the ability to foster partnerships among diverse stakeholders, (2) the ability to comprehend diverse stakeholder perspectives and deftly influence adoption, (3) the ability to build trust and collaboration, (4) self-awareness and humility to assemble teams with complementary skills, and (5) the integrity and accountability to embody ethical principles. The adaptive capacity encompasses (1) the foresight and sense-making abilities to discern emerging technologies and

XSL•FO
RenderX

their implications within the health care sphere; (2) the agility to identify and capitalize on transformative opportunities, swiftly adapting and aligning strategies with evolving contexts; and (3) systems thinking to enable an understanding of how elements interconnect and how changes in 1 area can reverberate throughout the entire system.

Contextual Factors Influencing Leadership in AI Transformation

The emerging themes from our review reveal that dynamic environmental and situational factors, including regulatory, technology, and organizational contexts, shape AI transformation within health care organizations. For instance, the regulatory context and frameworks related to health professions and health care organizations play a critical role in how AI can be integrated within the organizations. Similarly, the technology context such as the availability of AI technical talent, the retention of technical expertise, the dynamic nature of AI maturity, and the presence of incentives and technological resources for AI innovation or adoption will significantly influence a leader's ability to effectively drive AI readiness. Finally, the organization context is a critical influence on leaders' capacity for AI adoption and implementation. Organizations that promote and reward innovation and that have transparent communication practices shape leaders' ability to pursue AI opportunities.

Strategies for Empowering Health Care Leaders to Facilitate AI Transformation

For the technological domain, the included papers discussed approaches such as upskilling clinical experts with the necessary AI technical skills and ensuring the presence of specialized experts, such as computer scientists, to enable the subject matter experts to develop, test, and seamlessly integrate AI solutions. Further, the papers discussed collaborative strategies such as clinicians and computer scientists working together to effectively identify AI opportunities and develop, adopt, and implement AI solutions in clinical or operational areas.

For the strategic domain, organizational support was essential in supporting leaders to assess and identify AI opportunities that strategically align with organizational priorities and develop strategies to ensure AI transformation garners support from key stakeholders within the complex regulatory and environmental contexts. The literature also highlighted the competition for AI talent in health care and emphasized the significance of talent retention strategies to preserve the organization's AI technical expertise.

Then, in the operational domain, the emphasis was on establishing governance structures to continuously monitor data quality, patient privacy, and patient care experiences and assess the feasibility and financial implications of AI transformation. These governance structures ensure effective oversight and management of AI initiatives within health care organizations.

Finally, for the organizational domain, the focus was on the pivotal role of organizational culture in AI leadership. Leaders require organizational support to cultivate an environment that fosters innovation and actively incentivizes clinical leaders, such as physicians and nurses, through protected time and incentive pay to innovate and adopt AI solutions. Transparent decision-making processes related to AI solutions are essential cultural elements that build trust in AI systems and promote collaboration among the diverse stakeholders involved in AI transformation within health care organizations.

Discussion

Principal Findings

The purpose of a scoping review is not to draw definitive conclusions but to map the literature, identify emerging patterns, and develop critical propositions. As described in Figure 2, analysis of current literature shows that leading organizations toward AI transformation requires multidimensional leadership. As such, health care organizations need to engage leaders in the technological, strategic, operational, and organizational domains to facilitate AI transformation in their organizations. Further, the reviewed papers suggest that individuals in AI-related leadership roles need to demonstrate (1) technical capacity to understand the technology and innovation opportunities, (2) adaptive capacity to respond to contextual changes, and (3) interpersonal capacity to navigate the human aspects of the AI transformation process effectively. Furthermore, our study illuminates that leaders in the AI-related leadership roles need to navigate regulatory context, the dynamic nature of changing technology context, and organization context.

Figure 2. Multidimensional leadership framework for artificial intelligence (AI) transformation.

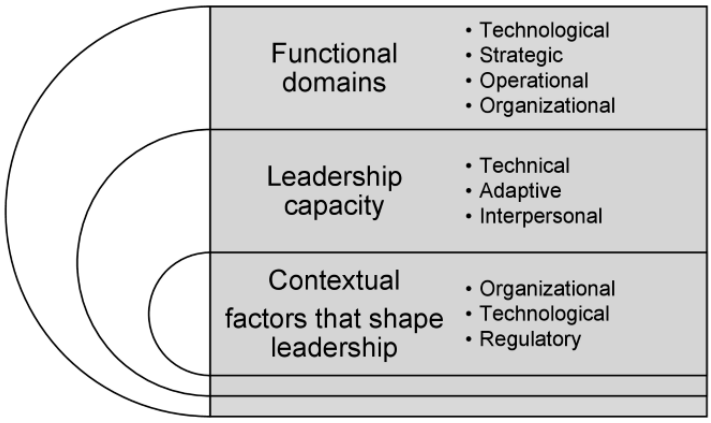

XSL•FO
RenderX

Prior Research

Health care organizations are marked by multifaceted interdependencies among medical facilities, health care providers, patients, administrative units, technology, and the regulatory environment. Therefore, the leadership required for AI transformation—which includes identifying AI opportunities, implementing AI solutions, and achieving full-scale AI adaptation—is not a static role but a continuous and dynamic process. Effective leadership involves the capacity to continuously identify opportunities for AI transformation, influence the thoughts and actions of others, and navigate the complex dynamics of the health care setting and AI technology landscape simultaneously. However, the current literature has not fully articulated this multidimensionality, often focusing on leadership through a linear approach.

Further, multiple situational factors can shape AI transformation. First, the rapid growth of AI technologies introduces an element of uncertainty, making it challenging to anticipate the long-term impact and sustainability of specific AI solutions [6]. Second, AI implementation involves many stakeholders, from technical experts and domain specialists to clinicians, administrators, patients, vendors, and regulatory bodies. Each stakeholder group brings its unique perspectives, priorities, and control systems into the equation, necessitating leaders to navigate competing values, trade-offs, and paradoxes [27]. Third, once alignment is achieved, the integration of AI within an organization triggers a need for a cultural shift, altering work practices and decision-making processes [38,59]. Fourth, the effectiveness of AI solutions hinges on the availability of high-quality data for informed insights and decision-making. When implementing solutions originally developed within different contexts, local organizations must ensure data integrity and the solution's adaptability to the organization's unique context [18]. This challenge is compounded by emerging regulatory frameworks, which add a layer of complexity. Ensuring compliance and the responsible use of AI technologies has become a critical consideration [29,50,60]. Finally, introducing AI may provoke resistance from employees concerned about job displacement or disruptions to established workflows. This problem is further compounded when an organization transitions toward integrating multiple AI systems, as these changes can lead to periods of chaos and confusion [59].

Emerging key opinions and evidence from outside the health domain indicate that leaders must possess an understanding of data quality nuances, assess process risks, and manage AI as a new team member. Additionally, leaders should have a firm grasp of technology, articulate clear business objectives, define precise goals, uphold a long-term vision, prepare their teams for AI transformation, manage data resources effectively, and foster organizational collaboration [3,61-67].

Our findings on the leadership required for AI transformation in health care organizations reinforce this multidimensionality of leadership to effectively navigate the complexities of AI transformation and successfully leverage its potential to drive transformative change. Leaders must operate across different functional domains—technological, strategic, operational, and organizational—while demonstrating technical, adaptive, and interpersonal capacities.

Further, our findings show contingency leadership theories, complexity theory, and transformational leadership theory as relevant theoretical domains for further explaining the different facets of leadership behaviors needed to navigate the multidimensionality of leadership required for AI transformation.

Contingency theories suggest that leadership effectiveness depends on situational factors, which should be considered in future AI implementation studies in the context of AI adaptation and integration within health care organizations [68,69]. Complexity theory provides a framework for examining leadership behaviors in interconnected, dynamic environments where leaders must balance innovation and stability and demonstrate an adaptive approach to challenges, characterized by uncertainty and change [70-73]. Transformational leadership theory emphasizes motivating, empowering, and developing others by fostering trust and collaboration while challenging the status quo to drive organizational change and innovation [74,75]. These theories should be considered in future AI implementation studies within health care organizations.

Future research and training programs related to AI in health care should examine the leadership required for AI transformation through the lens of multidimensionality, providing insights into the interrelatedness of functional domains, leadership capacities, and contextual enablers and barriers, while exploring the key theoretical domains related to contingency, complexity, and transformational leadership to further understand the interpersonal dynamics shaping AI transformation in health care.

Limitations

Some limitations to our scoping review are worth noting. First, given the contextual variability in the included studies and the methodological variations, we could not establish firm correlations about specific leadership domains, capacities, and contextual factors; the effectiveness of leadership approaches; or the moderating effects of contextual factors. Consequently, we have presented only the overarching emergent themes.

Second, our study is limited by the significant variation in conceptual definitions of leadership and leadership competencies found in the current literature, which often lacks more standardized definitions or instruments for measurement. This variation caused conceptual inconsistencies. We addressed the inconsistencies by clearly defining what constitutes a functional domain, capacity, and context before our data analysis to address this. We iteratively coded the data into themes to ensure all relevant aspects were captured.

Third, our search strategy focused on MEDLINE-indexed journals, which may exclude some newer journals indexed in PubMed but not yet in MEDLINE. While this might limit the capture of the very latest advancements in digital health, it does not diminish the robustness of the review. Fourth, we retrieved only articles written in English, which possibly limited the comprehensiveness of our findings. Fifth, we looked at AI as a system and did not look at the relationship between the

XSL•FO
RenderX

implementation of different types of AI tools and leadership behaviors which was beyond the scope of our review. Finally, our analysis used an inductive approach and was not informed by a predetermined theory to aid the mapping of the literature. This may have limited our analysis in capturing different elements of an umbrella theory.

Recommendations for Future Design and Research

Leading organizations toward AI transformation is an adaptive challenge influenced by a myriad of interwoven situational factors that create a dynamic and intricate environment. The body of literature related to AI in health care is rapidly expanding, and the recommendations imparted by this review, alongside the multidimensional leadership framework (Figure 2), stand poised to guide research and practice to empower health care organizations in their AI transformation journey. Future research on AI transformation, which includes innovation identification, implementation, and scaling, can use this framework to understand the role of leadership in driving successful outcomes.

Further, future research must undergo methodological expansion by embracing qualitative and mixed methods approaches to illuminate the intricate temporal aspects of AI transformation and corresponding evolving leadership behaviors.

Conclusions

In summary, emerging evidence shows that multidimensional leadership plays crucial role in AI transformation in health care organization. Leaders must adeptly balance technology opportunities while demonstrating unwavering empathy for stakeholder needs and nimble adaptability to accommodating the ever-changing contextual landscape, which encompasses the regulatory frameworks, the evolution of technology, and the organization's priorities.

Acknowledgments

This work is supported through a grant from the University of Toronto's Connaught Global Challenges. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the study.

Data Availability

The main study data are the data extraction materials and quality ratings of included papers, most of which are included in the study tables. The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in conception and design of the study and approved the protocol. AS and NS were responsible for overseeing the search of databases and literature. AS, NS, and SS were involved in the screening of articles, data extraction and data verification, and analysis of data. All authors were involved in data interpretation, supported in the drafting of the paper, which was led by AS, and all authors supported in revising and formatting of the paper. All authors have provided final approval of the version of the paper submitted for publication, and all authors agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Updated PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist.
[PDF File (Adobe PDF File), 549 KB - jmir_v26i1e54556_app1.pdf]

References

1. Fyfe I. AI predicts pancreatic cancer risk. Nat Rev Gastroenterol Hepatol 2023;20(7):413. [doi: 10.1038/s41575-023-00802-2] [Medline: 37268738]
2. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
3. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
4. AI in healthcare market | size, share, growth | 2022-2027. MarketWatch. 2023. URL: https://www.marketdataforecast.com/market-reports/artificial-intelligence-in-healthcare-market [accessed 2024-03-08]
5. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
6. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Cambridge: Academic Press; 2020:25-60.

XSL•FO
RenderX

7. Kasula BY. AI applications in healthcare: a comprehensive review of advancements and challenges. Int J Manag Educ Sustain Dev 2023;6(6) [FREE Full text]
8. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food & Drug Administration. 2023. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2024-03-08]
9. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
10. Capel T, Brereton M. What is human-centered about human-centered AI? a map of the research landscape. 2023 Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany p. 1-23. [doi: 10.1145/3544548.3580959]
11. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA 2020;323(6):509-510. [doi: 10.1001/jama.2019.21579] [Medline: 31845963]
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
13. Shaheen MY. Applications of Artificial Intelligence (AI) in Healthcare: A Review. Saudia Arabia: ScienceOpen; 2021.
14. Tipton K, Leas BF, Flores E, Jepson C, Aysola J, Cohen J, et al. Impact of healthcare algorithms on racial and ethnic disparities in health and healthcare. AHRQ 2023. [doi: 10.23970/AHRQEPCCER268]
15. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open 2023;6(12):e2345050 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
16. Berger E, Dries M. Beyond hype: getting the most out of generative AI in healthcare today. Bain & Company. 2023. URL: https://www.bain.com/insights/getting-the-most-out-of-generative-ai-in-healthcare/ [accessed 2024-03-08]
17. Khanna NN, Maindarkar MA, Viswanathan V, Fernandes JFE, Paul S, Bhagawati M, et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. Healthcare (Basel) 2022;10(12):2493 [FREE Full text] [doi: 10.3390/healthcare10122493] [Medline: 36554017]
18. Schmidt C. M.D. Anderson breaks with IBM Watson, raising questions about artificial intelligence in oncology. J Natl Cancer Inst 2017;109(5). [doi: 10.1093/jnci/djx113] [Medline: 30053147]
19. Toolkit for AI implementers. Canada Health Infoway. URL: https://www.infoway-inforoute.ca/en/digital-health-initiatives/innovative-technologies/artificial-intelligence/toolkit-for-ai-implementers [accessed 2024-03-08]
20. Liu Q, Huang R, Hsieh J, Zhu H, Tiwari M, Liu G, et al. Landscape analysis of the application of artificial intelligence and machine learning in regulatory submissions for drug development from 2016 to 2021. Clin Pharmacol Ther 2023;113(4):771-774. [doi: 10.1002/cpt.2668] [Medline: 35707940]
21. The national strategy for AI in health and social care. NHS Transformation Directorate. URL: https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/the-national-strategy-for-ai-in-health-and-social-care/ [accessed 2024-02-24]
22. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
23. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019;19(1):460 [FREE Full text] [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
24. Chowdhury S, Dey P, Joel-Edgar S, Bhattacharya S, Rodriguez-Espindola O, Abadie A, et al. Unlocking the value of artificial intelligence in human resource management through AI capability framework. Hum Resour Manage Rev 2023;33(1):100899. [doi: 10.1016/j.hrmr.2022.100899]
25. Shah R, Chircu A. IOT and AI in healthcare: a systematic literature review. Issues Inf Syst 2018;10:48009. [doi: 10.48009/3_iis_2018_33-41]
26. Chomutare T, Tejedor M, Svenning TO, Marco-Ruiz L, Tayefi M, Lind K, et al. Artificial intelligence implementation in healthcare: a theory-based scoping review of barriers and facilitators. Int J Environ Res Public Health 2022;19(23):16359 [FREE Full text] [doi: 10.3390/ijerph192316359] [Medline: 36498432]
27. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
28. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
29. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020;27(3):491-497 [FREE Full text] [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
30. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]

XSL•FO
RenderX

32. Mintzberg H. Managing. California: Berrett-Koehler; 2009.
33. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
34. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020;92(4):807-812. [doi: 10.1016/j.gie.2020.06.040] [Medline: 32565184]
35. Tricco CA, Langlois E, Straus ES. Rapid reviews to strengthen health policy and systems. WHO Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/258698/9789241512763-eng.pdf [accessed 2024-03-08]
36. Boyatzis RE. Transforming Qualitative Information: Thematic Analysis and Code Development. USA: Sage Publications; 1998.
37. Barbour AB, Frush JM, Gatta LA, McManigle WC, Keah NM, Bejarano-Pineda L, et al. Artificial intelligence in health care: insights from an educational forum. J Med Educ Curric Dev 2019;6:2382120519889348 [FREE Full text] [doi: 10.1177/2382120519889348] [Medline: 32064356]
38. Darcel K, Upshaw T, Craig-Neil A, Macklin J, Steele Gray C, Chan TCY, et al. Implementing artificial intelligence in Canadian primary care: barriers and strategies identified through a national deliberative dialogue. PLoS One 2023;18(2):e0281733 [FREE Full text] [doi: 10.1371/journal.pone.0281733] [Medline: 36848339]
39. Dicuonzo G, Donofrio F, Fusco A, Shini M. Healthcare system: moving forward with artificial intelligence. Technovation 2023;120:102510. [doi: 10.1016/j.technovation.2022.102510]
40. Ganapathi S, Duggal S. Exploring the experiences and views of doctors working with artificial intelligence in English healthcare; a qualitative study. PLoS One 2023;18(3):e0282415 [FREE Full text] [doi: 10.1371/journal.pone.0282415] [Medline: 36862694]
41. Gillan C. Healthcare Professional Association Agency in Preparing for Complex Technological Interventions: A Multiple-Case Study of Artificial Intelligence in Radiation Medicine and Medical Imaging (Dissertation). Canada: University of Toronto; 2022.
42. Henriksen A, Bechmann A. Building truths in AI: making predictive algorithms doable in healthcare. Inf Commun Soc 2020;23(6):802-816. [doi: 10.1080/1369118x.2020.1751866]
43. Laukka E, Hammarén M, Kanste O. Nurse leaders' and digital service developers' perceptions of the future role of artificial intelligence in specialized medical care: an interview study. J Nurs Manag 2022;30(8):3838-3846 [FREE Full text] [doi: 10.1111/jonm.13769] [Medline: 35970487]
44. Olaye IM, Seixas AA. The gap between AI and bedside: participatory workshop on the barriers to the integration, translation, and adoption of digital health care and AI startup technology into clinical practice. J Med Internet Res 2023;25:e32962 [FREE Full text] [doi: 10.2196/32962] [Medline: 37129947]
45. Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022;22(1):850 [FREE Full text] [doi: 10.1186/s12913-022-08215-8] [Medline: 35778736]
46. Ronquillo CE, Peltonen LM, Pruinelli L, Chu CH, Bakken S, Beduschi A, et al. Artificial intelligence in nursing: priorities and opportunities from an international invitational think-tank of the nursing and artificial intelligence leadership collaborative. J Adv Nurs 2021;77(9):3707-3717 [FREE Full text] [doi: 10.1111/jan.14855] [Medline: 34003504]
47. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30(10):5525-5532 [FREE Full text] [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
48. Upshaw TL, Craig-Neil A, Macklin J, Gray CS, Chan TCY, Gibson J, et al. Priorities for artificial intelligence applications in primary care: a Canadian deliberative dialogue with patients, providers, and health system leaders. J Am Board Fam Med 2023;36(2):210-220 [FREE Full text] [doi: 10.3122/jabfm.2022.220171R1] [Medline: 36948537]
49. Hakim Z, Ierasts T, Hakim I, D'Penha A, Smith KPD, Caesar MCW. Connecting data to insight: a pan-Canadian study on AI in healthcare. Healthc Q 2020;23(1):13-19. [doi: 10.12927/hcq.2020.26144] [Medline: 32249734]
50. Morley J, Murphy L, Mishra A, Joshi I, Karpathakis K. Governing data and artificial intelligence for health care: developing an international understanding. JMIR Form Res 2022;6(1):e31623 [FREE Full text] [doi: 10.2196/31623] [Medline: 35099403]
51. Willis M, Duckworth P, Coulter A, Meyer ET, Osborne M. Qualitative and quantitative approach to assess of the potential for automating administrative tasks in general practice. BMJ Open 2020;10(6):e032412 [FREE Full text] [doi: 10.1136/bmjopen-2019-032412] [Medline: 32513875]
52. Yang J, Luo B, Zhao C, Zhang H. Artificial intelligence healthcare service resources adoption by medical institutions based on TOE framework. Digit Health 2022;8:20552076221126034 [FREE Full text] [doi: 10.1177/20552076221126034] [Medline: 36211801]
53. Dixit A, Quaglietta J, Gaulton C. Preparing for the future: how organizations can prepare boards, leaders, and risk managers for artificial intelligence. Healthc Manage Forum 2021;34(6):346-352 [FREE Full text] [doi: 10.1177/08404704211037995] [Medline: 34533369]

XSL•FO
RenderX

54. Galsgaard A, Doorschodt T, Holten AL, Müller FC, Ploug Boesen M, Maas M. Artificial intelligence and multidisciplinary team meetings; a communication challenge for radiologists' sense of agency and position as spider in a web? Eur J Radiol 2022;155:110231 [FREE Full text] [doi: 10.1016/j.ejrad.2022.110231] [Medline: 35361507]
55. Sawers N, Bolster N, Bastawrous A. The contribution of artificial intelligence in achieving the sustainable development goals (SDGs): what can eye health can learn from commercial industry and early lessons from the application of machine learning in eye health programmes. Front Public Health 2021;9:752049 [FREE Full text] [doi: 10.3389/fpubh.2021.752049] [Medline: 35004574]
56. Li X, Cheng M, Xu J. Leaders' innovation expectation and nurses' innovation behaviour in conjunction with artificial intelligence: the chain mediation of job control and creative self-efficacy. J Nurs Manag 2022;30(8):3806-3816. [doi: 10.1111/jonm.13749] [Medline: 35899457]
57. Nasseef OA, Baabdullah AM, Alalwan AA, Lal B, Dwivedi YK. Artificial intelligence-based public healthcare systems: G2G knowledge-based exchange to enhance the decision-making process. Gov Inf Q 2022;39(4):101618. [doi: 10.1016/j.giq.2021.101618]
58. Ergin E, Karaarslan D, Şahan S, Çınar Yücel Ş. Artificial intelligence and robot nurses: from nurse managers' perspective: a descriptive cross-sectional study. J Nurs Manag 2022;30(8):3853-3862. [doi: 10.1111/jonm.13646] [Medline: 35474366]
59. Hazarika I. Artificial intelligence: opportunities and implications for the health workforce. Int Health 2020;12(4):241-245 [FREE Full text] [doi: 10.1093/inthealth/ihaa007] [Medline: 32300794]
60. Stanfill MH, Marc DT. Health information management: implications of artificial intelligence on healthcare data and information management. Yearb Med Inform 2019;28(1):56-64 [FREE Full text] [doi: 10.1055/s-0039-1677913] [Medline: 31419816]
61. Davenport T, Foutty J. AI-driven leadership. Sloan Management Review. 2018. URL: https://sloanreview.mit.edu/article/ai-driven-leadership/ [accessed 2024-03-08]
62. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
63. Greenhill RG, Pearson JS, Schmidt RN, Stuart D, Rossettie S. Exploring healthcare leadership competencies for the fourth industrial revolution: a scoping review of the literature. J Health Adm Educ 2021;38(3):695-708.
64. Forum W. Empowering AI leadership: AI C-suite toolkit. World Economic Forum. 2022. URL: https://www3.weforum.org/docs/WEF_Empowering_AI_Leadership_2022.pdf [accessed 2024-03-08]
65. Jones WA. Artificial intelligence and leadership: a few thoughts, a few questions. J Leadersh Stud 2018;12(3):60-62. [doi: 10.1002/jls.21597]
66. Peifer Y, Jeske T, Hille S. Artificial intelligence and its impact on leaders and leadership. Procedia Comput Sci 2022;200:1024-1030. [doi: 10.1016/j.procs.2022.01.301]
67. Smith AM, Green M. Artificial intelligence and the role of leadership. J of Leadership Studies 2018;12(3):85-87. [doi: 10.1002/jls.21605]
68. Hersey P, Blanchard KH, Natemeyer WE. Situational leadership, perception, and the impact of power. Group Organ Manag 2016;4(4):418-428. [doi: 10.1177/105960117900400404]
69. Westbrook D, Kennerley H, Kirk J. Contingency theories of effective leadership. In: The SAGE Handbook of Leadership. London: SAGE; 2011:286-298.
70. Uhl-Bien M, Marion R, McKelvey B. Complexity leadership theory: shifting leadership from the industrial age to the knowledge era. Leadersh Q 2007;18(4):298-318. [doi: 10.1016/j.leaqua.2007.04.002]
71. Uhl-Bien M, Meyer D, Smith J. Complexity leadership in the nursing context. Nurs Adm Q 2020;44(2):109-116. [doi: 10.1097/NAQ.0000000000000407] [Medline: 32134869]
72. Heifetz AR, Grashow A, Linsky M. The Practice of Adaptive Leadership: Tools and Tactics for Changing Your Organization and the World. Cambridge: Harvard Business Press; 2009.
73. Heifetz RA, Laurie DL. The Work of Leadership. New York: Harv Bus Rev; 2001.
74. Bass BM, Riggio RE. Transformational Leadership. New York: Lawrence Erlbaum Associates; 2006.
75. Bass B, Avolio B. Developing transformational leadership: 1992 and beyond. J Eur Ind Train 1990;14(5). [doi: 10.1108/03090599010135122]

Abbreviations

AI: artificial intelligence
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 14.11.23; peer-reviewed by D Chrimes, TAR Sure, S Kommireddy, J Konopik, M Brommeyer; comments to author 20.02.24; revised version received 12.03.24; accepted 15.07.24; published 14.08.24.

Please cite as:
Sriharan A, Sekercioglu N, Mitchell C, Senkaiahliyan S, Hertelendy A, Porter T, Banaszak-Holl J
Leadership for AI Transformation in Health Care Organization: Scoping Review
J Med Internet Res 2024;26:e54556
URL: https://www.jmir.org/2024/1/e54556
doi:10.2196/54556
PMID:39009038

©Abi Sriharan, Nigar Sekercioglu, Cheryl Mitchell, Senthujan Senkaiahliyan, Attila Hertelendy, Tracy Porter, Jane Banaszak-Holl. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Serious Games to Promote Behavior Change in Children With Chronic Diseases: Scoping Review and Development of a Self-Management Learning Framework

Made Ary Sarasmita[1,2*], MClinPharm, PhD; Ya-Han Lee[1,3], MSc; Fan-Ying Chan[1], MSc; Hsiang-Yin Chen[1,3*], PharmD

[1]Department of Clinical Pharmacy, School of Pharmacy, Taipei Medical University, Taipei, Taiwan

[2]Program Study of Pharmacy, Faculty of Mathematics and Science, Udayana University, Badung, Indonesia

[3]Department of Pharmacy, Wan Fang Hospital, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Hsiang-Yin Chen, PharmD
Department of Clinical Pharmacy
School of Pharmacy
Taipei Medical University
Health and Science Building, 7th Fl.
250 Wuxing Street
Taipei, 110
Taiwan
Phone: 886 02 2736 1661 ext 6175
Fax: 886 02 2736 1661
Email: shawn@tmu.edu.tw

Abstract

Background: Digital serious games (SGs) have rapidly become a promising strategy for entertainment-based health education; however, developing SGs for children with chronic diseases remains a challenge.

Objective: In this study, we attempted to provide an updated scope of understanding of the development and evaluation of SG educational tools and develop a framework for SG education development to promote self-management activities and behavior change in children with chronic diseases.

Methods: This study consists of a knowledge base and an analytical base. This study followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. To build the knowledge base, 5 stages of research were developed, including refining the review question (stage 1), searching for studies (stage 2), selecting relevant studies (stage 3), charting the information (stage 4), and collating the results (stage 5). Eligible studies that developed SG prototypes and evaluated SG education for children with chronic diseases were searched for in PubMed, Embase, Google Scholar, and peer-reviewed journals. In the analytical base, the context-mechanism-output approach and game taxonomy were used to analyze relevant behavioral theories and essential game elements. Game taxonomy included social features, presentation, narrative and identity, rewards and punishment, and manipulation and control. A total of 2 researchers selected the domains for the included behavioral theories and game elements. The intended SG framework was finalized by assembling SG fragments. Those SG fragments were appropriately reintegrated to visualize a new SG framework.

Results: This scoping review summarized data from 16 randomized controlled trials that evaluated SG education for children with chronic diseases and 14 studies on SG frameworks. It showed that self-determination theory was the most commonly used behavioral theory (9/30, 30%). Game elements included feedback, visual and audio designs, characters, narratives, rewards, challenges, competitions, goals, levels, rules, and tasks. In total, 3 phases of a digital SG framework are proposed in this review: requirements (phase 1), design and development (phase 2), and evaluation (phase 3). A total of 6 steps are described: exploring SG requirements (step 1), identifying target users (step 2), designing an SG prototype (step 3), building the SG prototype (step 4), evaluating the SG prototype (step 5), and marketing and monitoring the use of the SG prototype (step 6). Safety recommendations to use digital SG-based education for children in the post–COVID-19 era were also made.

Conclusions: This review summarizes the fundamental behavioral theories and game elements of the available literature to establish a new theory-driven step-by-step framework. It can support game designers, clinicians, and educators in designing,

XSL•FO
RenderX

developing, and evaluating digital, SG-based educational tools to increase self-management activities and promote behavior change in children with chronic diseases.

(J Med Internet Res 2024;26:e49692) doi:10.2196/49692

KEYWORDS

children; chronic disease; digital game; patient education; serious game

Introduction

Background

Serious game (SG) educational tools that provide constructive learning with imperative goals for behavior change [1] are increasingly being applied with children with chronic diseases. Training children to self-care for their chronic diseases is highly challenging due to insufficient cognition [2], low attendance [3], complicated treatments [4], and nonadherence to treatments [5]. A properly designed SG educational tool can allow children with chronic diseases to enjoy learning how to overcome real-life challenges [6]. A fully fledged game design provides a safe and controlled environment to experience and practice self-management skills [7]. Holtz et al [8] reported that SG education had positive impacts on self-efficacy, adherence, and knowledge, which drove improvements in behavior and health outcomes over time. Directing the design and application of SGs as an educational intervention to positively support children with chronic diseases could help improve therapeutic outcomes [9].

Behavioral sciences offer insights into how to design effective SG educational tools for children with chronic diseases to achieve the dual goals of internal enjoyment and confidence while promoting their self-care abilities. To understand changes in children's behaviors, basic principles and theories of learning, behavior, and mindset should be identified. The most commonly used theories explaining behavior change include social cognitive theory [10], self-determination theory (SDT) [11], and the mindset theory [12]. Social cognitive theory defines how behavior change can be achieved depending on personal factors, including cognition, capability, self-control, experiences, and expectations, and environmental factors, including emulation, reinforcement, and observation [10]. SDT addresses motivation and influences children to put themselves in situations in which they are exposed to SG education [13]. A growth mindset enhances greater resilience and positivity than a fixed mindset when dealing with challenges and failures [12]. Children with chronic diseases require continuous, specific self-management tasks to achieve levels of their mindset and cognitive development [12]. Incorporating behavioral theories and instructional learning into game mechanics, including practice tasks and challenges, can facilitate the changing process of a growth mindset and enhance motivation [14].

SG educational tools require a sophisticated design to avoid several potential negative consequences for children. The World Health Organization has articulated increased screen media use as a major concern due to the risk of addictive behaviors [15]. Higher gaming behavior is associated with higher levels of social, health, and behavioral problems in children and adolescents [16,17]. Playing SG interventions can promote behavior changes; however, uncontrolled excessive gaming may lead to gaming disorders [18] when games are used above the level of a child's age and mindset [19]. It has also been reported that a large proportion of electronic games may have violent content such as fighting, hitting, destroying, and killing [19], which increases the risk of aggressive behaviors in children. Inappropriate visual designs and game elements may distract from the educational purposes [18].

An SG framework that promotes positive behavior change in children is specifically needed because children's capabilities to respond to emotions and act when encountering difficulties differ from those of adults. Frameworks have been established for developing SG prototypes for adults [20,21]; however, only limited attention has been paid to creating a well-established theoretical SG framework for children. Children are more vulnerable to influences of digital games during their cognitive, social, and emotional development stages [2]. A stepwise SG framework is warranted to guide researchers in designing and evaluating SG educational tools for children with chronic diseases to maximize advantages and avoid unintended effects. This requires pivotal attention by researchers to creatively develop and design appropriate SG educational tools for children that balance the cornerstones of learning and playing.

Objectives

The study purpose was to offer an updated scoping review of SG education focused on delivering self-management activities and promoting behavior change for children with chronic diseases. It provided a scope of understanding of the development and evaluation of SG educational tools and developed a systematic methodological SG-based framework for children with chronic diseases. The intended SG framework was designed according to the methodology of scoping reviews Levac et al [22] and the context-mechanism-output (CMO) approach [23]. The specific aims were to create two bases: (1) to build a knowledge base that covers all the resources required to design and develop an SG framework and (2) to construct an analytical base by integrating behavioral theories and game elements from the knowledge base to design and visualize the intended SG framework through the CMO approach. The findings of this review can benefit researchers developing and evaluating SG-based learning educational tools for children with chronic diseases.

Methods

Study Design

Figure 1 [24-26] describes the process of developing an SG framework for children with chronic diseases using knowledge and analytical bases. This study followed the guidelines of the PRISMA-ScR (Preferred Reporting Items for Systematic

XSL•FO
RenderX

Reviews and Meta-Analyses extension for Scoping Reviews) [27]. For the knowledge base, relevant studies on developing an SG framework and evaluation of SG educational tools for children with chronic diseases were searched and collated throughout 5 stages to cover all the SG resources required. The analytical base used the CMO approach and game taxonomy to build a theory-based foundation for the proposed SG framework. Relevant behavioral theories and essential game elements from relevant studies were appropriately divided into fragments, compared, and assembled to create a visualization of the proposed SG framework. Discussions were conducted throughout the study to resolve any discrepancies among researchers.

Figure 1. Research flowchart. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; SDT: self-determination theory; SG: serious game.

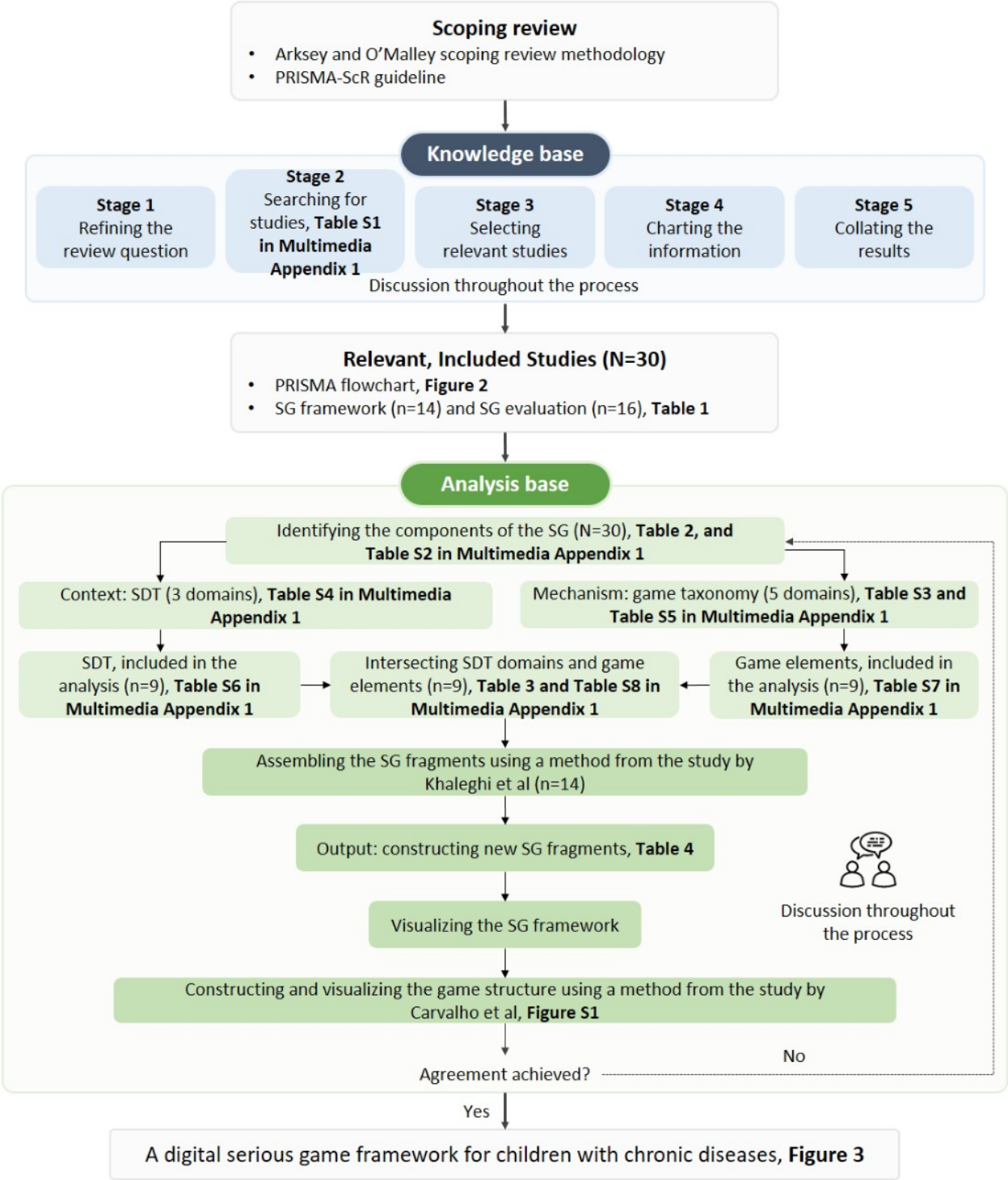

XSL•FO
RenderX

Knowledge Base

There were five stages for building the knowledge base: (1) refining the initial question, (2) identifying relevant studies, (3) selecting relevant studies, (4) charting the information, and (5) collating the results [22].

Refining the Review Question (Stage 1)

SGs are digital games that blend concepts of learning and performing attitudes and are enjoyable to play, with challenging goals [28]. On the basis of the literature, we began exploring the idea of how to develop an SG framework to improve self-management and promote behavior changes in children with chronic diseases. To develop an SG framework, a scientific foundation should be built supported by documented relevant evidence.

Identifying Relevant Studies (Stage 2)

Databases and Search Strategy

Relevant studies that had developed and published established SG frameworks and SG prototypes for children with chronic diseases were searched for using electronic databases, including PubMed, Embase, and Google Scholar. A hand searching method was also used to obtain additional relevant articles in peer-reviewed journals that focused on game research and were indexed in Web of Science, such as *Games for Health Journal* and *JMIR Serious Games*. We searched for articles using keywords obtained from Medical Subject Heading terms, such as "computer game," "video game," and "children," with no restriction on publication year (1980-2024). The panel in Multimedia Appendix 1 shows keyword term variations, and detailed search strategies are described in Table S1 in Multimedia Appendix 1. Reference lists of articles found through the electronic database searching were hand searched to obtain additional relevant information.

Inclusion Criteria

The inclusion criteria were (1) studies involving children aged 5 to 14 years, (2) studies that developed an SG framework for children with chronic diseases, and (3) studies that applied and evaluated the use of SG education for children with chronic diseases. Relevant articles were classified into two groups: (1) studies that focused on developing an SG framework and (2) studies that focused on evaluating SG educational tools for children with chronic diseases. Review or original articles explaining learning theories, behavioral theories, game theories, and game elements or presenting a general or specific SG model or framework that focused on behavior changes in children were included in the "SG framework studies" group. Randomized controlled trials (RCTs) that implemented and evaluated SG-based educational tools for children with chronic diseases were included in the "SG education studies/RCTs" group.

SG-based education is described as the use of SG prototypes or interventions, which are also known as computer games, for educational health and promotion of treatments, health education, patient adherence, therapeutic and side effect monitoring, and patient engagement. In this study, any changes in health-related outcomes in an RCT were descriptively reported. Clinical outcomes referred to any reduction in symptoms of chronic diseases and risks of complications, emergency visits, and hospitalizations. Humanistic outcomes were considered to be any condition that affected physical and social functions [29], including changes in attitudes and behaviors, adherence to treatment and medication, knowledge, quality of life, and patient satisfaction.

Selecting Relevant Studies (Stage 3)

The collected articles were initially imported into reference manager software (EndNote version 20; Clarivate Analysis). After removing duplicates, 2 researchers (MAS and YHL) independently assessed the articles using the inclusion criteria by examining titles and abstracts. Abstracts that met the inclusion criteria were retained for full-text review.

Charting the Information (Stage 4)

Characteristics of SG education for children with chronic diseases that were evaluated in RCTs were charted into a table, including authors' information, conditions, target ages, interventions, comparators, sample sizes, study duration, length of the study, and health-related outcomes. Data were summarized by 2 authors (MAS and YHL). Disagreements were resolved through discussion involving a third reviewer (HYC).

Collating the Results (Stage 5)

The components of SGs were collated from the included studies using the CMO approach [23]. According to the CMO approach, "context" consists of any fundamental principles that enhance the efficacy of SG education to induce behavior changes, including behavioral theories, learning theories, and game theories. "Mechanism" refers to rules of how a game works, the dynamics through which children interact in response to the game, and the game's appearance. It includes game elements to actively engage and motivate target users to perform self-management activities and positive behaviors. "Output" is related to any outcomes, study output, or study prototype.

Analytical Base

Identifying SG Components

Components of SG educational tools comprise behavioral theories, learning theories, game theories (context), and game taxonomy (mechanism). Behavioral theories and game elements were identified after collating all the included studies. Details of the identified SG components based on the CMO approach are presented in Tables S2 and S3 in Multimedia Appendix 1. The most often used behavioral theories were selected for inclusion in the analysis (n=9), as shown in Table S4 in Multimedia Appendix 1. Detailed categories of the game taxonomy are shown in Table S5 in Multimedia Appendix 1. Table S4 in Multimedia Appendix 1 presents SDT domains, consisting of the psychological needs of autonomy, competence, and relatedness to boost motivation. The autonomy domain refers to how users make decisions and boost the sense of control, such as adjusting choices, levels, and difficulties [11]. The sense of control evokes autonomy and fuels users' willingness to continue playing. Competence refers to achieving targeted goals of successful actions, such as challenges, learning tasks, competitions, and rewards [11]. Relatedness expresses how children interact and how their interactions affect others

XSL•FO
RenderX

within the game, such as avatars, feedback, and emotions [11]. Game taxonomy was applied to identify game elements in all studies, including social features, presentation, narrative, identity, rewards and punishments, and manipulation and control [30]. A total of 3 domains of SDT and 5 categories of game taxonomy intersected based on similar characteristics. Those elements were reintegrated to build new, appropriate game elements for children with chronic diseases. These SG components are reviewed and discussed throughout the analysis.

Assembling SG Fragments and Visualizing an SG Framework

SG fragments describe the strategies or systematic procedures for designing, developing, and evaluating SG educational tools for children with chronic diseases. After identifying SG components through SDT and game taxonomy, 2 researchers (MAS and YHL) individually analyzed SG framework studies (n=14) based on their step-by-step procedural techniques (fragments). The collected fragments were appropriately assembled into 5 steps following the method by Khaleghi et al [24]: objective definitions (step 1); users' needs and game element identification (step 2); game concept generation, game mechanics selection, and prototyping (step 3); implementation (step 4); and monitoring (step 5). Fragments of SG development were appropriately reintegrated to build a new SG framework.

Fragments for building up a new SG framework were established and consisted of 3 main phases (requirements, design and development, and evaluation). Each main phase was redesigned to formulate 2 procedural steps and generate output materials. A proposed SG framework was visualized by one researcher (MAS) and then carefully reviewed by 2 other authors (YHL and HYC). Following the method by Carvalho et al [25], a game structure was designed to supplement the SG framework to explain the actions, tools, and achieved goals of learning and gaming (Figure S1 in Multimedia Appendix 1). The process of assembling the SG fragments and then visualizing the SG framework and game structure were discussed among the 3 researchers throughout the study.

Results

Knowledge Base

Overview

We retrieved 1947 articles from PubMed (n=451, 23.16%), Embase (n=131, 6.73%), Google Search (n=512, 26.3%), Google Scholar (n=130, 6.68%), *JMIR Serious Games* (n=272, 13.97%), and *Games for Health Journal* (n=451, 23.16%). After removing duplicates, 738 full-text articles were reviewed. In total, 30 articles were included in the scoping review, consisting of 16 (53%) RCTs that evaluated SG educational tools for children with chronic diseases and 14 (47%) studies on SG frameworks (Figure 2).

Figure 2. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. MeSH: Medical Subject Heading; RCT: randomized controlled trial; SG: serious game.

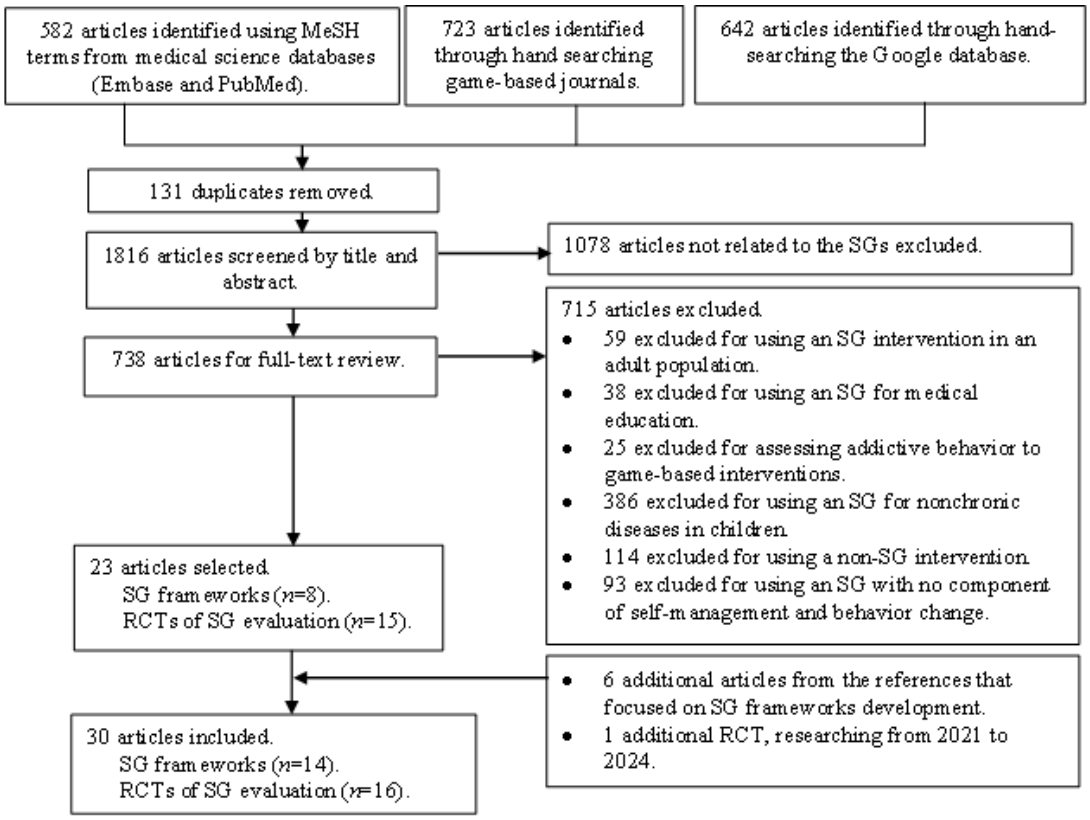

XSL•FO
RenderX

Charting the Information

Table 1 describes the included 16 RCTs that evaluated the use of SG educational tools for children with chronic diseases. SG educational tools that blended the concepts of learning and gaming were developed for children with asthma (7/16, 44%) [31-37], obesity and risk of diabetes or only diabetes (4/16, 25%) [38-41], cancer (2/16, 13%) [42,43], cystic fibrosis (1/16, 6%) [44], cerebral palsy (1/15, 6%) [45], and HIV or AIDS (1/15, 6%) [46]. The number of study participants ranged from 10 to 375; however, 60% (9/15) of the studies had <100 participants, as shown in Table 1. SG educational tools were delivered to children whose ages ranged from 3 to 17 years, with an average duration of play of approximately 40 minutes.

A total of 47% (7/15) of the studies reported improvements in clinical outcomes, including improved energy expenditure, heart rate, and blood pressure [31], or reduced symptoms, such as dyspnea and fatigue [31-33,44], as well as fewer hospitalizations [32] and unscheduled physician visits [38]. Regarding humanistic outcomes, 87% (13/15) of the RCTs evaluated behavioral outcomes. In total, 47% (7/15) of the studies presented improvements in knowledge, and 40% (6/15) reported improvements in behaviors, including asthma self-management [32,33], healthy dietary habits [39], communication with parents [38], disease-related risk communication [46], and lower medication use [34]. A total of 100% (15/15) of the studies evaluated users' acceptance and satisfaction, resulting in 100% (15/15) of the RCTs showing positive acceptance toward SG educational tools and consideration of SG educational tools as enjoyable strategies for learning and practicing self-management tasks. None of the 15 RCTs evaluated economic outcomes.

XSL•FO
RenderX

Table 1. Summary of the included serious game educational tools in randomized controlled trials (N=16).

Study, year	Condition	Age of users (y)	Intervention	Control	Sample size, N	Health outcomes
Rubin et al [33], 1986	Asthma	7-12	Asthma Command	Nonasthma computer game	65	Lower asthma-related acute visits, improved asthma knowledge, and improved asthma management–related behaviors
Bartholomew et al [32], 2000	Asthma	7-17	Watch, Discover, Think, and Act	Usual care	133	Fewer symptoms and hospitalizations, increased asthma knowledge, improved correct decision-making in the game scenario (62%) and engagement in the game (84%), and children felt satisfied (97%)
Yawn et al [37], 2000	Asthma	3-12	Air Academy	Usual health education	87	Improved asthma knowledge, and children and teachers felt satisfied
Huss et al [34], 2003	Asthma	7-12	Wee Willie Wheezie	Written asthma education	101	Children felt that the game could have been more esthetic.
Shames et al [36], 2004	Asthma	5-12	Bronkie's Asthma Adventure	Usual care and video game	119	Increased asthma knowledge, and children had a high interest in the program.
McPherson et al [35], 2005	Asthma	7-14	The Asthma Files	Asthma booklet	101	Lower oral steroid use, improved asthma knowledge, improved internal locus of control, high interest in the program (35/37, 95%), children felt that the game helped them gain asthma knowledge (32/37, 87%), and fewer school absences
Gomes et al [31], 2015	Asthma	5-9	Reflex Ridge from The Kinect Adventures Program	Treadmill session	36	Lower level of FeNO[a], improved asthma control and exercise capacity, higher energy expenditure, higher motivation, and high endurance with the game
Salonini et al [44], 2015	Cystic fibrosis	8-17	The Kinect Adventures Program	Stationary cycle training	30	Less frequent dyspnea and fatigue and high enjoyment of the game
Kato et al [43], 2008	Cancer	13-29	Remission	Noncancer computer game	375	Improved adherence to the use of cancer medications, increased self-efficacy, and greater knowledge
Hamari et al [42], 2019	Cancer	3-16	Nintendo WiiFit	Usual care	36	High acceptability and participation (77%), but the game was not followed as recommended.
Brown et al [38], 1997	Diabetes	8-16	Packy and Marlon	Video games with no health education content	59	Improved self-efficacy, better self-care behavior, increased diabetes-related communication with parents, and fewer unscheduled visits to the physician
Baranowski et al [40], 2011	Obesity and risk of diabetes	10-12	Escape from Diab and Nanoswarm	Knowledge-based internet game	133	Increased habit of eating healthy foods and high enjoyment of the game (80%-90%)
Baranowski et al [39], 2019	Obesity and risk of diabetes	10-12	Escape from Diab and Nanoswarm	Knowledge-based internet game	200	Increased expectations for gameplay
Weiland et al [41], 2022	Obesity and risk of diabetes	9-12	Kids Obesity Prevention or family intervention	Child intervention	23	Increased knowledge gain in children and parents, maintenance of knowledge in parents, and high acceptance of the game
Winskell et al [46], 2018	HIV or AIDS	11-14	Tumaini	Usual care	60	Improved sexual health–related knowledge, greater self-efficacy, and improved intention for risk avoidance strategies and sexual risk communication
Pin and Butler [45], 2019	Cerebral palsy	6-14	Interactive game	Usual care	18	High enjoyment of the game

[a]FeNO: fractional exhaled nitric oxide.

XSL•FO
RenderX

Collating SG Components

Table 2 describes essential components of the SG-based education for developing the proposed SG framework based on the CMO approach (N=30). For context, several studies applied behavioral theories (23/30, 77%) and game theories (22/30, 73%). Regarding mechanisms, several aspects were concerned with embedding social features or feedback (28/30, 93%); presentation or esthetics (30/30, 100%); personalization, including narratives (24/30, 80%), characters (23/30, 77%), and rewards and punishments (26/30, 87%); and manipulation and control, including game genre or rules (23/30, 77%), game goals (28/30, 93%), and challenges (27/30, 90%). Regarding output, 27% (8/30) of the studies generated specific SG frameworks for children with chronic illnesses, including children with diabetes [39,40,47,48], children with cystic fibrosis [49], and children who needed physical rehabilitation [24,50,51]. The most commonly used behavioral theory in the studies was SDT (9/30, 30%). Relevant studies that used SDT as the behavioral theory foundation are identified in Table S6 in Multimedia Appendix 1. Game elements from all the included studies (N=30) are described in Table S7 in Multimedia Appendix 1.

Table 2. Collation of serious game (SG) components based on the context-mechanism-output approach (N=30).

	SG framework studies (n=14), n (%)	Randomized clinical trials (n=16), n (%)	Total studies, n (%)
Context			
Behavioral theories	12 (86)	11 (69)	23 (77)
Game theories	13 (93)	9 (56)	22 (73)
Mechanism			
Social (feedback)	13 (93)	15 (94)	28 (97)
Presentation (esthetic)	14 (100)	15 (94)	29 (100)
Personalization			
Narrative (story and narrative)	13 (93)	11 (69)	24 (83)
Identity (characters and avatars)	12 (86)	11 (69)	23 (79)
Rewards and punishments	13 (93)	13 (81)	26 (90)
Manipulation and control			
Game genre and rules	10 (71)	13 (81)	23 (79)
Game goals	14 (100)	15 (94)	29 (97)
Challenges	12 (86)	15 (94)	27 (93)
Output			
Health outcomes			
Behavioral outcomes	—[a]	14 (88)	14 (47)
Learning outcomes	—	12 (75)	12 (40)
Clinical outcomes	—	11 (69)	11 (38)
Dissemination of the SG framework			
A specific framework for children	8 (57)	—	8 (28)
General framework	6 (43)	—	6 (21)

[a]Not applicable.

Analytical Base

Identifying SG Components

Table 3 summarizes the intersection of the 3 domains of SDT and 5 categories of game taxonomy (n=9). On the basis of our findings, game elements that should be inserted in a proposed SG framework for children with chronic diseases include feedback, such as tailored messages and links to social media (social); visual designs, such as images, videos, animations, and cartoons, and audio designs, such as music and sounds (presentation); avatars, characters, and emotions (identity); storyline (narrative); rewards and progress bar (rewards and punishments); and challenges, choices, competitions, goals, rules, levels, and tasks (manipulation and control). Details of the intersection of SDT and game taxonomy (n=9) are described in Table S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Intersection of the domains of self-determination theory and game taxonomy for the proposed serious game (SG) framework.

Self-determination theory of the proposed SD framework	Game taxonomy				
	Social features	Presentation	Narrative and identity	Rewards and punishments	Manipulation and control
Competence	__[a]	Educative materials [50,51]: learning content and learning instructions	—	Rewards [21,24,39,40,47-51]: points, progress bar, badges, and stars; punishments: NR[b]	Challenges [21,24,39,40,47-51]: competitions [21,24,39,40,48,50,51], levels [21,24,39,40,48,50,51], tasks [21,47,50,51], game rules [24,39,40,50], and goals [24,39,40,47,50]
Autonomy	—	Presentation: visual design [21,24,39,40,47-51]: images, videos, animations, cartoons, and attractive layout; audio design [24,39,40,47-51]: music and sounds	—	—	Choices [24,39,40,47,50,51] and difficulty adjustment [24,39,40,49,50]
Relatedness	Feedback [21,24,39,40,47-51]: tailored messages [39,40,48] and social media [21,49]	—	Narrative [21,24,39,40,47,51]: storyline; identity [21,24,39,40,48-51]: avatars, characters, and emotions	—	Motivation [47]

[a]Not applicable.

[b]NR: not reported.

Assembling SG Fragments and Visualizing the Framework

Table 4 presents the determination of SG fragments from the included studies (14/30, 47%) and then assembles those SG fragments into a proposed SG framework. Each existing study offered different procedural steps for developing an SG prototype, yet the game elements and behavioral theories complemented each other. Only 50% (7/14) of the studies created SG prototypes [24,25,40,47,49,52,53]. Of the 14 SG framework studies, 3 (21%) specifically targeted self-management activities [21,41,47]. On the basis of these findings, we reintegrated those fragments into 3 main phases with 6 step-by-step procedural techniques.

In phase 1 (requirements), there are 2 important steps, including exploring the idea and SG requirements (step 1) using literature reviews and identifying target users' needs (step 2) using iterative discussions or interviews. The outputs of phase 1 are relevant theories, game taxonomy, and children's needs and preferences. In phase 2 (design and development), 2 steps should be considered by designers, including designing the game elements and educative materials (step 3) and building an SG prototype (step 4) using appropriate software programs and hardware equipment. The output of phase 2 is the SG prototype with a game structure. The final phase is phase 3 (evaluation), which is concerned with evaluating the SG prototype using a clinical trial design (step 5) and marketing the SG and monitoring its use (step 6) throughout clinic-based practice. Outputs of phase 3 are health outcome results and the final SG prototype with recommendations for its use.

XSL•FO
RenderX

Table 4. Assembling serious game fragments from the selected studies (N=14).

Study, year	Step 1 (objective definition)	Step 2		Step 3		Step 4 (implementation)	Step 5 (monitoring)
		Users' needs	Game element identification	Game mechanic	Prototyping		
AlMarshedi et al [21], 2016	+[a]	NR[b]	+	+	NR	NR	NR
Baranowski et al [48], 2011	+	+	+	+	+	+	+
Carvalho et al [25], 2015	+	+	+	+	+	+	+
Beristain-Colorado et al [50], 2021	+	NR	+	+	NR	NR	NR
Dörrenbächer et al [51], 2014	+	+	+	+	+	+	+
Epstein et al [23], 2021	+	+	+	+	NR	NR	NR
Hansen [49], 2017	+	+	+	+	+	+	+
Jaccard et al [54], 2021	+	NR	+	+	NR	NR	NR
Khaleghi et al [24], 2021	+	+	+	+	+	NR	NR
Mummah et al [20], 2016	+	NR	+	+	NR	NR	NR
Starks [52], 2014	+	+	+	+	NR	NR	NR
Thompson et al [47], 2010	+	+	+	+	+	+	+
Verschueren et al [53], 2019	+	+	+	+	+	+	+
Wattanasoontorn et al [55], 2013	+	NR	+	+	NR	NR	NR

[a]Present or reported.

[b]NR: not reported.

The Proposed SG Framework

Overview

On the basis of our knowledge base and analytical base, we propose a new SG-based framework that integrates the principles of SDT and game elements into self-management practices, titled Self-Management Interactive Learning and Entertainment for children with chronic diseases, as presented in Figure 3. It consists of 3 main phases, starting from the requirements of SG educational tools (phase 1), design and development (phase 2), and evaluation of the SG educational tools (phase 3). In total, 2 procedural steps are included in each phase, resulting in 6 procedural steps. Each step has input materials as the foundation to support the actions and process and to produce output materials. Output materials in the first phase (phase 1) can be used as the input for the next phase (phase 2), and so forth. Each phase has critical points, revisions, and adjustments that should be considered by any game designer, researcher, or health care provider who would like to create an SG educational tool. Each game should be suitable for target users and their conditions; for example, an SG educational tool for children with asthma should have specific learning materials (asthma action plans and asthma medications), target goals (improved quality of life), and tasks and challenges (asthma self-management activities, breathing technique, and proper asthma medication use) that might differ from those of other diseases. Figure S1 in Multimedia Appendix 1 shows the gaming and learning structure of an SG prototype that blends SDT domains and game elements. The mechanism of how players achieve the target goal by accomplishing challenges should be set in clear game rules. Children will make their first choice by selecting an avatar or character, directly engaging with the game. Learning materials will help children understand their disease management, yet the game instructions will help them simultaneously observe challenges and tasks. After completing the tasks, their performance should be rewarded through points or a performance meter.

XSL•FO
RenderX

Figure 3. The proposed Self-Management Interactive Learning and Entertainment framework.

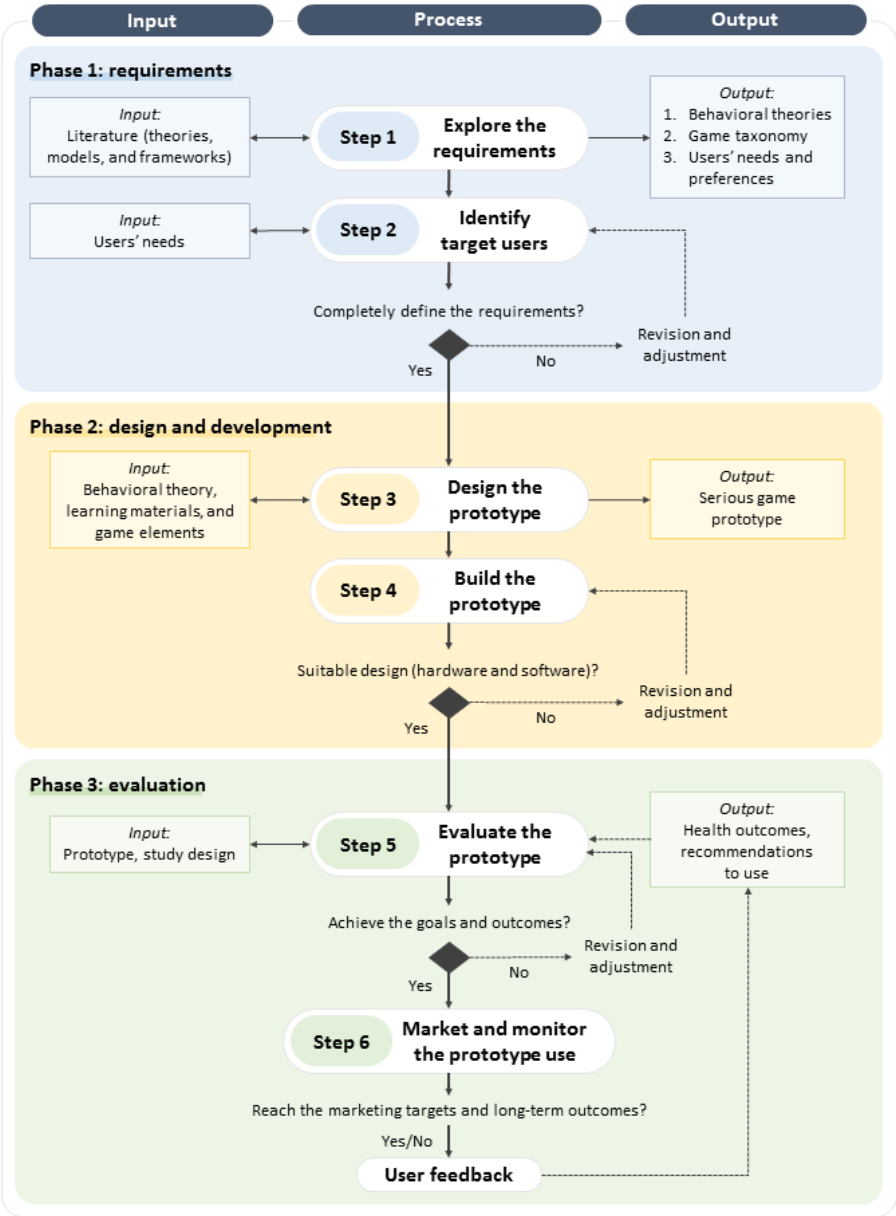

Exploring SG Requirements (Step 1)

A robust theoretical SG-based foundation should be established using literature reviews that gather principles of learning theories, behavioral theories, and game theories. This step is aimed at exploring SG requirements by searching for evidence related to game-based behavior change programs for children using electronic databases, for example, behavioral and game theories. If such evidence is not available, it is recommended to consult established game developers and collect perspectives from target audiences regarding obstacles in their daily lives [21]. According to Bramer et al [56], critical points include how to determine a clear and focused research question, how to choose databases and interfaces to begin, how to use an appropriate search technique, and how to document and translate collected documents. After determining relevant SG literature, game elements and behavioral theories should be translated and adopted for use by children with chronic diseases. Relevant articles can be used as inputs to conduct iterative discussions to identify users' needs.

Identifying Target Users (Step 2)

Step 2 began through iterative discussions with a multidisciplinary, collaborative team. The iterative approach refers to the iterative process of refining, creating, and revising a project until agreement is achieved, and it is commonly used for agile software development [57]. The aim of this step is to collect perspectives on identifying users' profiles and needs, their daily difficulties and barriers related to their chronic conditions, and target outcomes [7]. In this step, critical points emphasize what the players' backgrounds are, what age groups are considered, to which chronic conditions would the SG educational tool be applied, how many users would be involved in the game, and what outcomes need to be achieved [53]. A multidisciplinary team consisting of pediatricians, child psychologists, child educators, game prototype designers, and multimedia experts [54] needs to identify resilient attitudes and

XSL•FO
RenderX

consistent stimuli that suit children's characteristics. Designers should carefully identify users' cultures, beliefs, mindset, and literacy to concisely adopt those preferences into the game's elements [24]. Directly involving children through focus group discussions or in-depth interviews will help the team gamify self-management tasks based on their needs and level of understanding, including medication adherence, physical exercise, and maintenance of healthy dietary habits.

Designing the SG Prototype (Step 3)

A key driver for successful SG education is consolidating a balance between self-management tasks (serious) and game elements (entertainment) [6,58]. This step aims to design the mechanism and user interface of the game itself by consolidating the most appropriate behavioral theories, learning materials, and game elements. Designers begin to create a prototype after establishing selected relevant theories, game elements, and users' needs and outcomes (input). First, designers should elucidate selected, well-established behavioral and learning theories into educational materials and game taxonomy into appropriate game elements for children. Game designers should consider several critical points, including what topics are inserted into the learning materials, which game elements are best suited to achieve the desired outcomes, and how interacting with the game can lead to targeted behaviors [49]. The educational materials should contain disease information, including its pathophysiology, signs and symptoms, treatments and medications, self-management, side effects, the importance of adherence, and daily practices.

It is recommended to insert game elements that offer enjoyment to stimulate children to play, at the same time directly motivating them to learn [51]. Cartoon characters, genres, and stories represent personalization for children [52]. To grow children's mindset, challenges should be designed with competitive levels and rewards provided when a mission is accomplished [12]. A role model with a positive attitude should be inserted into the SG design to encourage children to become masters of practicing self-management activities. Adding these elements facilitates children responding when confronted with conflicts [50] and enhances their sense of resilience. It is important to design an SG prototype that mimics real-world circumstances by setting precise goals and instructing players to perform targeted skills over time [25]. It is also suggested to embed the features of feedback, a progress bar, or trend alerts to evaluate their performance after completing the challenges.

Building the SG Prototype (Step 4)

Step 4 aims to develop an actual prototype based on the selected behavioral theories and game elements. It requires extensive discussions with researchers, multimedia experts, and the game industry to integrate the technology into a game console. A graphic user interface should be built to present the set of game rules. Designers may consider facilitating level adjustments if children fail to win to maintain the developed mindset. Critical points are how the prototype can be built for efficient learning and playing, how to perform such tasks, and how to rapidly respond regarding those performances. Esthetics is an essential aspect to be considered. Game visuals can be appropriately created using 2D or 3D formats [55]. Music and animation can be added to enhance excitement and enjoyment. To effectively promote self-management tasks, virtual reality SGs should be equipped with body movement tools that specifically target childhood chronic diseases that involve physical disabilities [50]. Moreover, privacy should be protected because SG prototypes can be used in multiplayer settings, and the accessibility to enter measured data should be restricted [13].

Evaluating the SG Prototype (Step 5)

Step 5 focuses on evaluating the efficacy of the SG educational tools and unexpected effects after implementation. This step allows researchers to gather feedback from experts and children for further improvements [20,59]. A pilot test, followed by a clinical trial, is recommended, which quantitatively analyzes how the prototype achieves the intended outcomes and qualitatively explores users' experiences. An RCT study design is preferred due to its high quality. The ability to perform a task at an expected level and with minimal adverse events may be set as the intended outcome. It should be carefully determined how long participants will be engaged in the game, how many sessions a child needs to reach the goals, and how long it takes to complete a session. A short duration is associated with unfamiliarity with the tasks, whereas a long duration leads to boredom [60].

Moreover, health outcomes, including clinical, humanistic, and economic outcomes, should be periodically evaluated [61]. Clinical outcomes may include symptom improvement and reduction of morbidity, whereas humanistic outcomes may include knowledge and attitudes, behavior changes, and an improved mindset. As no economic outcomes were available in the studies in this review, researchers are encouraged to evaluate economic outcomes when using the prototype. Possible unexpected impacts of SG interventions on aggressive behaviors should also be evaluated, especially for SG interventions that encompass violent elements [62,63]. Continued discussions with clinicians are still relevant to ensure that the game world setting can be applied to the real world.

Marketing and Monitoring Use of the Prototype (Step 6)

Disseminating and promoting a well-evaluated SG educational tool can enhance access to a broader population that may benefit the most and promptly inform the game industry to invest in such interventions. Commercialization of an SG educational tool for children remains a challenge due to the need for high-end technologies, animated multimedia design, artists, illustrators, and other consoles. Gameplay is rapidly changing due to advances in technology, and it should be developed in line with current modalities to minimize the obsolescence of software and hardware [39]. To ensure market readiness, business experts should be consulted and involved throughout the process. It is recommended for researchers to accelerate partnerships with the gaming industry for sustainable SG maintenance.

Specific practice skills can be designed in a modest simulation. For example, children with type 1 diabetes should be able to use insulin regularly, exercise, maintain a healthy diet, and be aware of the signs of hypoglycemia. Modest instruction will help clinicians in applying SG education for children with chronic diseases in the real world. It is important to underline

XSL•FO
RenderX

that an excellent performance in the game world is not directly associated with mastery in the real world. Practicing self-management skills, such as physical activities and medication use, should regularly be guided by health care professionals. It may be relevant to consult with policy makers and health care associations regarding the establishment of policies and recommendations for appropriate uses of SG educational tools in clinical practice. Postmarketing feedback should continually be collected to improve the SG's quality.

Discussion

Principal Findings

This scoping review proposed a digital framework to design SG educational tools for children with chronic diseases. The SG framework consists of 3 main strategies to guide the planning, design and development, and implementation of SG educational tools to allow children to practice self-management skills for their chronic condition. Major considerations of how each step is conceptualized, including a theory-driven foundation, contents of health education, joyful reinforcement, and use of technology, were discussed. The game elements and game structure should engage children's attention and support them in performing gamified self-management tasks, changing their mindset, and increasing their self-care abilities.

Comparison With Prior Work and Considerations for Using the Proposed Framework

Implementation of SG educational tools for children with chronic diseases has been demonstrated in several previous works [8,9,58], specifically concerning health education [55,64], physical activities [65,66], and self-management [9,67,68]; however, none of them offer a theory-driven framework for behavior change. It has been suggested that researchers articulate a scientific framework for the design of SG educational tools [65]. Although behavioral and self-management interventions can be delivered to children from 5 to 18 years of age [67,68], the health educational content is not applicable to the entire age range. Educational materials for children should be supplemented with communication skills, whereas activities for adolescents should focus on self-monitoring and problem-solving [69]. Multidisciplinary teamwork from conception to marketing is strongly emphasized [64], which was accommodated in this framework throughout the proposed phases.

As game-based interventions are continually growing, researchers are considering developing SG educational tools for children, but questions about how to get started have been raised. Developing an SG educational tool is expensive; therefore, several aspects should be carefully considered before initiating development of SG educational tools, including securing funding and building a collaborative team [69]. Developing SG educational tools for children with chronic diseases differs from entertainment-only video games due to their unique components of behavioral theories and learning materials to boost self-management practices and promote positive behavior changes. For example, children with asthma may need knowledge about preventing asthma triggers and adhering to medication, whereas children with cystic fibrosis may need more physical rehabilitation activities than children with other chronic respiratory diseases. Some of them may need specific, scheduled physical activities, whereas others may need the efforts of encouragement or psychological support and companionship. That is why the game design should be able to address those needs.

Establishing a solid team, which involves experienced game developers or game companies, should be noted. Once members are chosen, clear ground responsibilities and expectations regarding the prototype design should be established. The health care professional team can develop appropriate health learning contents and discuss those materials with the game developer team to analyze and resolve potential problems before programming and prototyping. As there is no reimbursement for SG use as a medical treatment [69], acquiring available funding and resources should be prioritized.

Challenges and Pitfalls of SG Design and Development

Developing appropriate SG educational tools for the specific needs of children with chronic diseases remains a challenge due to the huge investment from ideas to implementation. As the market for SG-based interventions expands across health conditions, there is a trend for SG education to be included as a supportive intervention rather than merely as pure entertainment [23].

On the basis of our heterogeneous results, the procedure through which SG educational tools deliver content might not be the only key contributor to achieve the targeted goals because the intervention should be focused not only on the learning materials but also on the intertwined mechanism of game elements and the elements of behavioral theory. In the context of game-based learning, self-management practices will be correctly performed if users are enjoying themselves, which means having the propensity to engage, blend, and learn. From this perspective, we raised several considerations on the potential of game elements to enhance intrinsic motivation, including how much autonomy (videos, animations or cartoons, choices, and difficulty adjustments of the challenges) must be given to children during play, how can relatedness (narrative or storyline, avatars, characters, and tailored messages) between children and the game be built into SG educational tools, and how can a child's level of competence be defined to challenge them.

Several critical points in each step were pointed out for game designers to avoid failure. First, there can be failure to define suitable educative materials and targeted behaviors for children with specific difficulties. Second, one can fail to generate a dynamic between players and the game while, at the same time, players have to obtain new learning from the SG educational tools. Game levels were revealed to engage players with a positive learning effect; however, this should be in line with the player's skills and cognitive development. A high-challenge game with low-skill, fixed-mindset users may induce frustration; meanwhile, a low-challenge game for users with high skills and a growth mindset may generate feelings of triviality [6]. Given rapid trends in digital technology, SG prototypes should be continually adjusted to prevent them from becoming hackneyed by the time the evaluation trial is finished.

XSL•FO
RenderX

Safety Concerns for Children in the Post–COVID-19 Era

Safety aspects of SG educational tools should be of general concern because these tools are considered a persuasive technology for changing human behaviors. Game-based interventions appear to be most effective in users aged <18 years [23]; nevertheless, children and adolescents vary in their ability to master a mission. Children may feel engaged with customizable avatars, but some of these may contain violent characters [19]. Game designers should ensure that the SG intervention is not dangerous or does not increase risks to children, such as by promoting sedentary or aggressive behaviors [47] or increasing the risk of physical injuries due to practicing skills. Several harmful risks are associated with sleep disorders and internet gaming disorders, such as anxiety, unsuccessful attempts at control, and jeopardizing environments [19,70].

The American Academy of Pediatrics has stated concerns about the influence of digital media on the health and cognitive development of children at the ages of 0 to 5 years, and it has proposed limiting screen use to 1 hour per day for children aged 2 to 5 years [71]. It is also recommended to avoid screen time 1 to 2 hours before bedtime for children and adolescents. In 2020, the American Academy of Ophthalmology recommended the 20-20-20 rule, described as a 20-second break every 20 minutes by looking 20 feet away to prevent and relieve digital eyestrain [72].

The COVID-19 pandemic intensified gaming behaviors among children, especially during school closures, and this garnered the concern of policy makers and health care professionals [15]. Sedentary time in children with chronic diseases might have increased [73] as parents were not well prepared for it due to their attention being focused on social and economic burdens caused by the pandemic. Several SG educational tools were developed during the pandemic to stimulate in-home physical rehabilitation [74,75] and improve anxiety and mood disorders in adolescents [76], and those positive behavioral outcomes should be maintained. Even though the pandemic situation has improved, some parents are continuing to work remotely while simultaneously caring for children, leading to obstacles to maintaining children's learning, especially in households of a low socioeconomic status [77].

Educational, game-based interventions for the post–COVID-19 era should be integrated with appropriate recommendations for their use. Individualized family media use plans are strongly recommended; hence, parental control is central when exposing children to digital media [70]. It is considered important for parents to accompany their children during screen use to foster an effective learning process by understanding the game structure, supporting children in controlling playing times, and monitoring their activities. Instead of giving punishment as a disciplinary matter, by playing together, parents can understand more about SG educational tools and how they can facilitate parent-child interactions. As parents become familiar with their children's games, they will be able to encourage their children to achieve the intended outcomes and avoid addictive behaviors [78].

Limitations

This scoping review has a few limitations. This framework was developed based on a review of the most relevant SG educational tools in several RCTs and SG framework studies instead of a direct participatory approach involving health care professionals and children. When comparing the effects of SG educational tools, most RCTs (9/16, 56%) only captured improvements in humanistic outcomes, such as knowledge [57] and enjoyment. Studies on improving clinical outcomes were limited, and none provided economic outcome evaluations. This is in line with the findings of a previous review that presented a lack of clinical evidence of the implementation of SG educational tools in children with neurodevelopmental disorders [79]. Several studies (5/16, 31%) evaluated changes in knowledge over a relatively short duration on beneficial effects on behaviors. Exploration is still needed as to which game elements can have higher effects on self-management and behavior changes. Moreover, issues of maintenance of intended behaviors after exposure to SG interventions should be carefully addressed. With the limitations of the available literature, this framework should be tested in further studies.

Implications of the Study and Further Research

This framework provides a theory-driven step-by-step approach to help health educators, clinicians, game developers, and policy makers more efficiently develop SG educational tools for children with chronic diseases. Understanding how to integrate the power of SG educational tools offers significant promise for promoting health behavior changes. Only 4% of the top-rated health apps apply the concepts of gamification [80], indicating that the opportunity to develop high-quality SG educational tools for children with chronic diseases is still wide open. Further research should explore the needs for culture-specific SG educational tools and investigate the mediators of behavior change.

Conclusions

A framework of SG-based educational tools promoting self-management activities and behavior changes in children with chronic diseases was developed by incorporating behavioral principles and mechanisms of SGs. It expedites the translation of fundamental behavioral theories and game elements into a scaled-up industrial level in which digital-based game interventions are being created to enhance children's participation and motivation. The effectiveness of SG educational tools in achieving targeted behaviors depends on key designs and elements of how they address problems and mindsets of children with difficulties. Underpinning appropriate behavioral theories, learning materials, game elements, esthetics, and technology should be considered in all phases of research. The design, development, and evaluation of SG educational tools for children with chronic diseases need to be broadly explored with the support of a well-validated game-based framework and the deployment of advanced technologies.

XSL•FO
RenderX

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

MAS contributed to the conceptualization, methodology, software, validation, formal analysis, writing—original draft, and visualization. YHL contributed to formal analysis and writing—review and editing. FYC contributed to visualization and writing—review and editing. HYC contributed to conceptualization, methodology, formal analysis, resources, writing—review and editing, supervision, project administration, and funding acquisition. All authors contributed to data interpretation and manuscript preparation. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 74 KB - jmir_v26i1e49692_app1.docx]

Multimedia Appendix 2
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[PDF File (Adobe PDF File), 80 KB - jmir_v26i1e49692_app2.pdf]

References

1. Ferrante G, Licari A, Marseglia GL, La Grutta S. Digital health interventions in children with asthma. Clin Exp Allergy 2021 Feb;51(2):212-220 [FREE Full text] [doi: 10.1111/cea.13793] [Medline: 33238032]
2. Dall'Oglio I, Gasperini G, Carlin C, Biagioli V, Gawronski O, Spitaletta G, et al. Self-care in pediatric patients with chronic conditions: a systematic review of theoretical models. Int J Environ Res Public Health 2021 Mar 28;18(7):3513 [FREE Full text] [doi: 10.3390/ijerph18073513] [Medline: 33800684]
3. Crimarco A, Mayfield C, Mitchell N, Beets MW, Yin Z, Moore JB. Determinants of attendance at a physical activity focused afterschool program in elementary school children. Int J Exerc Sci 2018 May 1;11(5):137-151 [FREE Full text] [Medline: 29795736]
4. Hilliard ME, Hahn A, Ridge AK, Eakin MN, Riekert KA. User preferences and design recommendations for an mHealth app to promote cystic fibrosis self-management. JMIR Mhealth Uhealth 2014 Oct 24;2(4):e44 [FREE Full text] [doi: 10.2196/mhealth.3599] [Medline: 25344616]
5. Sarasmita MA, Larasanty LP, Kuo LN, Cheng KJ, Chen HY. A computer-based interactive narrative and a serious game for children with asthma: development and content validity analysis. J Med Internet Res 2021 Sep 13;23(9):e28796 [FREE Full text] [doi: 10.2196/28796] [Medline: 34515641]
6. Drummond D, Hadchouel A, Tesnière A. Serious games for health: three steps forwards. Adv Simul (Lond) 2017 Feb 4;2(1):3 [FREE Full text] [doi: 10.1186/s41077-017-0036-3] [Medline: 29450004]
7. Fitzgerald M, Ratcliffe G. Serious games, gamification, and serious mental illness: a scoping review. Psychiatr Serv 2020 Feb 01;71(2):170-183. [doi: 10.1176/appi.ps.201800567] [Medline: 31640521]
8. Holtz BE, Murray K, Park T. Serious games for children with chronic diseases: a systematic review. Games Health J 2018 Oct;7(5):291-301. [doi: 10.1089/g4h.2018.0024] [Medline: 29957082]
9. Charlier N, Zupancic N, Fieuws S, Denhaerynck K, Zaman B, Moons P. Serious games for improving knowledge and self-management in young people with chronic conditions: a systematic review and meta-analysis. J Am Med Inform Assoc 2016 Jan;23(1):230-239 [FREE Full text] [doi: 10.1093/jamia/ocv100] [Medline: 26186934]
10. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004 Apr;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
11. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
12. Dweck CS, Yeager DS. Mindsets: a view from two eras. Perspect Psychol Sci 2019 May 01;14(3):481-496 [FREE Full text] [doi: 10.1177/1745691618804166] [Medline: 30707853]
13. Thompson D. Incorporating behavioral techniques into a serious videogame for children. Games Health J 2017 Apr;6(2):75-86 [FREE Full text] [doi: 10.1089/g4h.2016.0066] [Medline: 28231024]
14. Graafland M, Dankbaar M, Mert A, Lagro J, De Wit-Zuurendonk L, Schuit S, et al. How to systematically assess serious games applied to health care. JMIR Serious Games 2014 Nov 11;2(2):e11 [FREE Full text] [doi: 10.2196/games.3825] [Medline: 25654163]

XSL•FO
RenderX

15. Dong H, Yang F, Lu X, Hao W. Internet addiction and related psychological factors among children and adolescents in China during the coronavirus disease 2019 (COVID-19) epidemic. Front Psychiatry 2020 Sep 2;11:00751 [FREE Full text] [doi: 10.3389/fpsyt.2020.00751] [Medline: 32982806]
16. Li X, Vanderloo LM, Keown-Stoneman CD, Cost KT, Charach A, Maguire JL, et al. Screen use and mental health symptoms in Canadian children and youth during the COVID-19 pandemic. JAMA Netw Open 2021 Dec 01;4(12):e2140875 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.40875] [Medline: 34962557]
17. Werling AM, Walitza S, Grünblatt E, Drechsler R. Media use before, during and after COVID-19 lockdown according to parents in a clinically referred sample in child and adolescent psychiatry: results of an online survey in Switzerland. Compr Psychiatry 2021 Aug;109:152260 [FREE Full text] [doi: 10.1016/j.comppsych.2021.152260] [Medline: 34265597]
18. Zhu S, Zhuang Y, Lee P, Li JC, Wong PW. Leisure and problem gaming behaviors among children and adolescents during school closures caused by COVID-19 in Hong Kong: quantitative cross-sectional survey study. JMIR Serious Games 2021 May 07;9(2):e26808 [FREE Full text] [doi: 10.2196/26808] [Medline: 33960954]
19. Elsayed W. Covid-19 pandemic and its impact on increasing the risks of children's addiction to electronic games from a social work perspective. Heliyon 2021 Dec;7(12):e08503 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e08503] [Medline: 34869925]
20. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
21. AlMarshedi A, Wills G, Ranchhod A. Gamifying self-management of chronic illnesses: a mixed-methods study. JMIR Serious Games 2016 Sep 09;4(2):e14 [FREE Full text] [doi: 10.2196/games.5943] [Medline: 27612632]
22. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
23. Epstein DS, Zemski A, Enticott J, Barton C. Tabletop board game elements and gamification interventions for health behavior change: realist review and proposal of a game design framework. JMIR Serious Games 2021 Mar 31;9(1):e23302 [FREE Full text] [doi: 10.2196/23302] [Medline: 33787502]
24. Khaleghi A, Aghaei Z, Mahdavi MA. A gamification framework for cognitive assessment and cognitive training: qualitative study. JMIR Serious Games 2021 May 18;9(2):e21900 [FREE Full text] [doi: 10.2196/21900] [Medline: 33819164]
25. Carvalho MB, Bellotti F, Berta R, De Gloria A, Sedano CI, Hauge JB, et al. An activity theory-based model for serious games analysis and conceptual design. Comput Educ 2015 Sep;87:166-181 [FREE Full text] [doi: 10.1016/j.compedu.2015.03.023]
26. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Thomas TH, Sivakumar V, Babichenko D, Grieve VL, Klem ML. Mapping behavioral health serious game interventions for adults with chronic illness: scoping review. JMIR Serious Games 2020 Jul 30;8(3):e18687 [FREE Full text] [doi: 10.2196/18687] [Medline: 32729836]
29. Gunter MJ. The role of the ECHO model in outcomes research and clinical practice improvement. Am J Manag Care 1999 Apr;5(4 Suppl):S217-S224 [FREE Full text] [Medline: 10387542]
30. King D, Delfabbro P, Griffiths M. Video game structural characteristics: a new psychological taxonomy. Int J Ment Health Addict 2009 Apr 7;8(1):90-106. [doi: 10.1007/s11469-009-9206-4]
31. Gomes EL, Carvalho CR, Peixoto-Souza FS, Teixeira-Carvalho EF, Mendonça JF, Stirbulov R, et al. Active video game exercise training improves the clinical control of asthma in children: randomized controlled trial. PLoS One 2015 Aug 24;10(8):e0135433 [FREE Full text] [doi: 10.1371/journal.pone.0135433] [Medline: 26301706]
32. Bartholomew LK, Gold RS, Parcel GS, Czyzewski DI, Sockrider MM, Fernandez M, et al. Watch, discover, think, and act: evaluation of computer-assisted instruction to improve asthma self-management in inner-city children. Patient Educ Couns 2000 Feb;39(2-3):269-280. [doi: 10.1016/s0738-3991(99)00046-4] [Medline: 11040726]
33. Rubin DH, Leventhal JM, Sadock RT, Letovsky E, Schottland P, Clemente I, et al. Educational intervention by computer in childhood asthma: a randomized clinical trial testing the use of a new teaching intervention in childhood asthma. Pediatrics 1986 Jan;77(1):1-10. [Medline: 3510016]
34. Huss K, Winkelstein M, Nanda J, Naumann P, Sloand E, Huss R. Computer game for inner-city children does not improve asthma outcomes. J Pediatr Health Care 2003 Mar;17(2):72-78. [doi: 10.1016/S0891-5245(02)88328-6]
35. McPherson AC, Glazebrook C, Forster D, James C, Smyth A. A randomized, controlled trial of an interactive educational computer package for children with asthma. Pediatrics 2006 Apr 01;117(4):1046-1054. [doi: 10.1542/peds.2005-0666] [Medline: 16585298]
36. Shames RS, Sharek P, Mayer M, Robinson TN, Hoyte EG, Gonzalez-Hensley F, et al. Effectiveness of a multicomponent self-management program in at-risk, school-aged children with asthma. Ann Allergy Asthma Immunol 2004 Jun;92(6):611-618. [doi: 10.1016/s1081-1206(10)61426-3]

XSL•FO
RenderX

37. Yawn BP, Algatt-Bergstrom PJ, Yawn RA, Wollan P, Greco M, Gleason M, et al. An in-school CD-ROM asthma education program. J Sch Health 2000 Apr 09;70(4):153-159. [doi: 10.1111/j.1746-1561.2000.tb06462.x] [Medline: 10790839]
38. Brown SJ, Lieberman DA, Germeny BA, Fan YC, Wilson DM, Pasta DJ. Educational video game for juvenile diabetes: results of a controlled trial. Med Inform (Lond) 1997 Jul 12;22(1):77-89. [doi: 10.3109/14639239709089835] [Medline: 9183781]
39. Baranowski T, Baranowski J, Chen TA, Buday R, Beltran A, Dadabhoy H, et al. Videogames that encourage healthy behavior did not alter fasting insulin or other diabetes risks in children: randomized clinical trial. Games Health J 2019 Aug;8(4):257-264 [FREE Full text] [doi: 10.1089/g4h.2018.0097] [Medline: 30964335]
40. Baranowski T, Baranowski J, Thompson D, Buday R, Jago R, Griffith MJ, et al. Video game play, child diet, and physical activity behavior change a randomized clinical trial. Am J Prev Med 2011 Jan;40(1):33-38 [FREE Full text] [doi: 10.1016/j.amepre.2010.09.029] [Medline: 21146765]
41. Weiland A, Reiband N, Schäffeler N, Zurstiege G, Giel KE, Zipfel S, et al. A serious game for the prevention of obesity in school children-impact of parent's involvement: a randomized controlled trial. Life (Basel) 2022 May 24;12(6):779 [FREE Full text] [doi: 10.3390/life12060779] [Medline: 35743810]
42. Hamari L, Järvelä LS, Lähteenmäki PM, Arola M, Axelin A, Vahlberg T, et al. The effect of an active video game intervention on physical activity, motor performance, and fatigue in children with cancer: a randomized controlled trial. BMC Res Notes 2019 Nov 29;12(1):784 [FREE Full text] [doi: 10.1186/s13104-019-4821-z] [Medline: 31783907]
43. Kato PM, Cole SW, Bradlyn AS, Pollock BH. A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. Pediatrics 2008 Aug;122(2):e305-e317. [doi: 10.1542/peds.2007-3134] [Medline: 18676516]
44. Salonini E, Gambazza S, Meneghelli I, Tridello G, Sanguanini M, Cazzarolli C, et al. Active video game playing in children and adolescents with cystic fibrosis: exercise or just fun? Respir Care 2015 Aug 21;60(8):1172-1179 [FREE Full text] [doi: 10.4187/respcare.03576] [Medline: 25899477]
45. Pin TW, Butler PB. The effect of interactive computer play on balance and functional abilities in children with moderate cerebral palsy: a pilot randomized study. Clin Rehabil 2019 Apr;33(4):704-710. [doi: 10.1177/0269215518821714] [Medline: 30599772]
46. Winskell K, Sabben G, Akelo V, Ondeng'e K, Obong'o C, Stephenson R, et al. A smartphone game-based intervention (Tumaini) to prevent HIV among young Africans: pilot randomized controlled trial. JMIR Mhealth Uhealth 2018 Aug 01;6(8):e10482 [FREE Full text] [doi: 10.2196/10482] [Medline: 30068501]
47. Thompson D, Baranowski T, Buday R, Baranowski J, Thompson V, Jago R, et al. Serious video games for health how behavioral science guided the development of a serious video game. Simul Gaming 2010 Aug 01;41(4):587-606 [FREE Full text] [doi: 10.1177/1046878108328087] [Medline: 20711522]
48. Baranowski T, Baranowski J, Thompson D, Buday R. Behavioral science in video games for children's diet and physical activity change: key research needs. J Diabetes Sci Technol 2011 Mar 01;5(2):229-233 [FREE Full text] [doi: 10.1177/193229681100500204] [Medline: 21527086]
49. Hansen OG. What gamification design do users want in a self-management application for chronic diseases? - The case of cystic fibrosis. Norwegian University of Science and Technology. 2017. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2458481 [accessed 2022-04-17]
50. Beristain-Colorado MD, Ambros-Antemate JF, Vargas-Treviño M, Gutiérrez-Gutiérrez J, Moreno-Rodriguez A, Hernández-Cruz PA, et al. Standardizing the development of serious games for physical rehabilitation: conceptual framework proposal. JMIR Serious Games 2021 Jun 24;9(2):e25854 [FREE Full text] [doi: 10.2196/25854] [Medline: 34185003]
51. Dörrenbächer S, Müller PM, Tröger J, Kray J. Dissociable effects of game elements on motivation and cognition in a task-switching training in middle childhood. Front Psychol 2014;5:1275 [FREE Full text] [doi: 10.3389/fpsyg.2014.01275] [Medline: 25431564]
52. Starks K. Cognitive behavioral game design: a unified model for designing serious games. Front Psychol 2014;5:28 [FREE Full text] [doi: 10.3389/fpsyg.2014.00028] [Medline: 24550858]
53. Verschueren S, Buffel C, Vander Stichele G. Developing theory-driven, evidence-based serious games for health: framework based on research community insights. JMIR Serious Games 2019 May 02;7(2):e11565 [FREE Full text] [doi: 10.2196/11565] [Medline: 31045496]
54. Jaccard D, Suppan L, Sanchez E, Huguenin A, Laurent M. The co.LAB generic framework for collaborative design of serious games: development study. JMIR Serious Games 2021 Jul 02;9(3):e28674 [FREE Full text] [doi: 10.2196/28674] [Medline: 34255655]
55. Wattanasoontorn V, Boada I, García R, Sbert M. Serious games for health. Entertain Comput 2013 Dec;4(4):231-247 [FREE Full text] [doi: 10.1016/j.entcom.2013.09.002]
56. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc 2018 Oct 04;106(4):531-541 [FREE Full text] [doi: 10.5195/jmla.2018.283] [Medline: 30271302]
57. Viudes-Carbonell SJ, Gallego-Durán FJ, Llorens-Largo F, Molina-Carmona R. Towards an iterative design for serious games. Sustainability 2021 Mar 17;13(6):3290. [doi: 10.3390/su13063290]

XSL•FO
RenderX

58. Drummond D, Monnier D, Tesnière A, Hadchouel A. A systematic review of serious games in asthma education. Pediatr Allergy Immunol 2017 May 27;28(3):257-265. [doi: 10.1111/pai.12690] [Medline: 27992659]
59. Caserman P, Hoffmann K, Müller P, Schaub M, Straßburg K, Wiemeyer J, et al. Quality criteria for serious games: serious part, game part, and balance. JMIR Serious Games 2020 Jul 24;8(3):e19037 [FREE Full text] [doi: 10.2196/19037] [Medline: 32706669]
60. Buday R, Baranowski T, Thompson D. Fun and games and boredom. Games Health J 2012 Aug;1(4):257-261 [FREE Full text] [doi: 10.1089/g4h.2012.0026] [Medline: 24761316]
61. Kozma CM, Reeder CE, Schulz RM. Economic, clinical, and humanistic outcomes: a planning model for pharmacoeconomic research. Clin Ther 1993;15(6):1121-32; discussion 1120. [Medline: 8111809]
62. DeSmet A, Van Ryckeghem D, Compernolle S, Baranowski T, Thompson D, Crombez G, et al. A meta-analysis of serious digital games for healthy lifestyle promotion. Prev Med 2014 Dec;69:95-107 [FREE Full text] [doi: 10.1016/j.ypmed.2014.08.026] [Medline: 25172024]
63. Espinosa-Curiel IE, Pozas-Bogarin EE, Martínez-Miranda J, Pérez-Espinosa H. Relationship between children's enjoyment, user experience satisfaction, and learning in a serious video game for nutrition education: empirical pilot study. JMIR Serious Games 2020 Sep 17;8(3):e21813 [FREE Full text] [doi: 10.2196/21813] [Medline: 32940609]
64. Sharifzadeh N, Kharrazi H, Nazari E, Tabesh H, Edalati Khodabandeh M, Heidari S, et al. Health education serious games targeting health care providers, patients, and public health users: scoping review. JMIR Serious Games 2020 Mar 05;8(1):e13459 [FREE Full text] [doi: 10.2196/13459] [Medline: 32134391]
65. Tabak M, Dekker-van Weering M, van Dijk H, Vollenbroek-Hutten M. Promoting daily physical activity by means of mobile gaming: a review of the state of the art. Games Health J 2015 Dec;4(6):460-469. [doi: 10.1089/g4h.2015.0010] [Medline: 26397179]
66. Bossen D, Broekema A, Visser B, Brons A, Timmerman A, van Etten-Jamaludin F, et al. Effectiveness of serious games to increase physical activity in children with a chronic disease: systematic review with meta-analysis. J Med Internet Res 2020 Apr 01;22(4):e14549 [FREE Full text] [doi: 10.2196/14549] [Medline: 32234697]
67. Sattoe JN, Bal MI, Roelofs PD, Bal R, Miedema HS, van Staa A. Self-management interventions for young people with chronic conditions: a systematic overview. Patient Educ Couns 2015 Jun;98(6):704-715. [doi: 10.1016/j.pec.2015.03.004] [Medline: 25819373]
68. Brigden A, Parslow RM, Linney C, Higson-Sweeney N, Read R, Loades M, et al. How are behavioural interventions delivered to children (5-11 years old): a systematic mapping review. BMJ Paediatr Open 2019;3(1):e000543 [FREE Full text] [doi: 10.1136/bmjpo-2019-000543] [Medline: 31909219]
69. Baranowski T, Blumberg F, Buday R, DeSmet A, Fiellin LE, Green CS, et al. Games for health for children-current status and needed research. Games Health J 2016 Feb;5(1):1-12 [FREE Full text] [doi: 10.1089/g4h.2015.0026] [Medline: 26262772]
70. Han TS, Cho H, Sung D, Park MH. A systematic review of the impact of COVID-19 on the game addiction of children and adolescents. Front Psychiatry 2022 Aug 18;13:976601 [FREE Full text] [doi: 10.3389/fpsyt.2022.976601] [Medline: 36061296]
71. Council on Communications and Media. Media and young minds. Pediatrics 2016 Nov;138(5):e20162591 [FREE Full text] [doi: 10.1542/peds.2016-2591] [Medline: 27940793]
72. Boyd K. Computers, digital devices, and eye strain. American Academy of Ophthalmology. URL: https://www.aao.org/eye-health/tips-prevention/computer-usage [accessed 2022-04-18]
73. Duan L, Shao X, Wang Y, Huang Y, Miao J, Yang X, et al. An investigation of mental health status of children and adolescents in China during the outbreak of COVID-19. J Affect Disord 2020 Oct 01;275:112-118 [FREE Full text] [doi: 10.1016/j.jad.2020.06.029] [Medline: 32658812]
74. da Silva TD, da Silva PL, Valenzuela EJ, Dias ED, Simcsik AO, de Carvalho MG, et al. Serious game platform as a possibility for home-based telerehabilitation for individuals with cerebral palsy during COVID-19 quarantine - a cross-sectional pilot study. Front Psychol 2021 Feb 2;12:622678 [FREE Full text] [doi: 10.3389/fpsyg.2021.622678] [Medline: 33633648]
75. Demers M, Martinie O, Winstein C, Robert MT. Active video games and low-cost virtual reality: an ideal therapeutic modality for children with physical disabilities during a global pandemic. Front Neurol 2020 Dec 14;11:601898 [FREE Full text] [doi: 10.3389/fneur.2020.601898] [Medline: 33381077]
76. Dietvorst E, Aukes MA, Legerstee JS, Vreeker A, Hrehovcsik MM, Keijsers L, et al. A smartphone serious game for adolescents (Grow It! App): development, feasibility, and acceptance study. JMIR Form Res 2022 Mar 03;6(3):e29832 [FREE Full text] [doi: 10.2196/29832] [Medline: 35238795]
77. Oflu A, Bükülmez A, Elmas EG, Tahta E, Çeleğen M. Comparison of screen time and digital gaming habits of Turkish children before and during the coronavirus disease 2019 pandemic. Turk Arch Pediatr 2021 Jan;56(1):22-26 [FREE Full text] [doi: 10.14744/TurkPediatriArs.2020.41017] [Medline: 34013225]
78. Király O, Potenza MN, Stein DJ, King DL, Hodgins DC, Saunders JB, et al. Preventing problematic internet use during the COVID-19 pandemic: consensus guidance. Compr Psychiatry 2020 Jul;100:152180 [FREE Full text] [doi: 10.1016/j.comppsych.2020.152180] [Medline: 32422427]

XSL•FO
RenderX

79. Kokol P, Vošner HB, Završnik J, Vermeulen J, Shohieb S, Peinemann F. Serious game-based intervention for children with developmental disabilities. Curr Pediatr Rev 2020 Apr 09;16(1):26-32. [doi: 10.2174/1573396315666190808115238] [Medline: 31393252]
80. Edwards EA, Lumsden J, Rivas C, Steed L, Edwards LA, Thiyagarajan A, et al. Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps. BMJ Open 2016 Oct 04;6(10):e012447 [FREE Full text] [doi: 10.1136/bmjopen-2016-012447] [Medline: 27707829]

Abbreviations

CMO: context-mechanism-output
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
SDT: self-determination theory
SG: serious game

Edited by S Ma; submitted 06.06.23; peer-reviewed by R Gorantla, K Spruyt, T Baranowski; comments to author 20.12.23; revised version received 13.04.24; accepted 25.06.24; published 19.08.24.

Please cite as:
Sarasmita MA, Lee YH, Chan FY, Chen HY
Digital Serious Games to Promote Behavior Change in Children With Chronic Diseases: Scoping Review and Development of a Self-Management Learning Framework
J Med Internet Res 2024;26:e49692
URL: https://www.jmir.org/2024/1/e49692
doi:10.2196/49692
PMID:39158952

©Made Ary Sarasmita, Ya-Han Lee, Fan-Ying Chan, Hsiang-Yin Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Prevalence of Health Misinformation on Social Media—Challenges and Mitigation Before, During, and Beyond the COVID-19 Pandemic: Scoping Literature Review

Dhouha Kbaier[1], PhD; Annemarie Kane[2], PhD; Mark McJury[3]; Ian Kenny[1], PhD

[1]School of Computing and Communications, The Open University, Milton Keynes, United Kingdom
[2]Faculty of Arts and Social Sciences, The Open University, Milton Keynes, United Kingdom
[3]School of Physical Sciences, The Open University, Milton Keynes, United Kingdom

Corresponding Author:
Dhouha Kbaier, PhD
School of Computing and Communications
The Open University
Walton Hall
Milton Keynes, MK7 6AA
United Kingdom
Email: dhouha.kbaier@open.ac.uk

Abstract

Background: This scoping review accompanies our research study "The Experience of Health Professionals With Misinformation and Its Impact on Their Job Practice: Qualitative Interview Study." It surveys online health misinformation and is intended to provide an understanding of the communication context in which health professionals must operate.

Objective: Our objective was to illustrate the impact of social media in introducing additional sources of misinformation that impact health practitioners' ability to communicate effectively with their patients. In addition, we considered how the level of knowledge of practitioners mitigated the effect of misinformation and additional stress factors associated with dealing with outbreaks, such as the COVID-19 pandemic, that affect communication with patients.

Methods: This study used a 5-step scoping review methodology following Arksey and O'Malley's methodology to map relevant literature published in English between January 2012 and March 2024, focusing on health misinformation on social media platforms. We defined health misinformation as a false or misleading health-related claim that is not based on valid evidence or scientific knowledge. Electronic searches were performed on PubMed, Scopus, Web of Science, and Google Scholar. We included studies on the extent and impact of health misinformation in social media, mitigation strategies, and health practitioners' experiences of confronting health misinformation. Our independent reviewers identified relevant articles for data extraction.

Results: Our review synthesized findings from 70 sources on online health misinformation. It revealed a consensus regarding the significant problem of health misinformation disseminated on social network platforms. While users seek trustworthy sources of health information, they often lack adequate health and digital literacies, which is exacerbated by social and economic inequalities. Cultural contexts influence the reception of such misinformation, and health practitioners may be vulnerable, too. The effectiveness of online mitigation strategies like user correction and automatic detection are complicated by malicious actors and politicization. The role of health practitioners in this context is a challenging one. Although they are still best placed to combat health misinformation, this review identified stressors that create barriers to their abilities to do this well. Investment in health information management at local and global levels could enhance their capacity for effective communication with patients.

Conclusions: This scoping review underscores the significance of addressing online health misinformation, particularly in the postpandemic era. It highlights the necessity for a collaborative global interdisciplinary effort to ensure equitable access to accurate health information, thereby empowering health practitioners to effectively combat the impact of online health misinformation. Academic research will need to be disseminated into the public domain in a way that is accessible to the public. Without equipping populations with health and digital literacies, the prevalence of online health misinformation will continue to pose a threat to global public health efforts.

(J Med Internet Res 2024;26:e38786) doi:10.2196/38786

XSL•FO
RenderX

KEYWORDS

health misinformation; online health communities; vaccine hesitancy; social media; health professionals; public health; COVID-19; intervention; antivaxxers

Introduction

The global adoption of the internet has made health information more accessible, and the development of digital technology has enabled its rapid dissemination. However, the internet has also made possible the dissemination of false and misleading health misinformation and disinformation, with negative consequences, including the potential to exacerbate health inequalities. Health practitioners have found themselves at the forefront of communicating with patients who have taken on board health misinformation in the context of its proliferation on the web. This paper (associated with the study by Ismail et al [1]) surveyed the current literature concerning online health misinformation to establish the extent and scope of the problem, with special reference to the needs of health practitioners in their efforts to mitigate its impact. Several studies have established useful definitions of the terms *misinformation* and *disinformation* and distinctions between them. Misinformation has been defined as information that is not supported by evidence and contradicts the best-supported evidence available [2,3]. Wang et al [4] made a further distinction between online misinformation and disinformation, in particular on social media platforms. For Wang et al [4], misinformation is information that is not known to be false and is shared without malice. By contrast, disinformation involves the knowing and sharing of false information with the purpose of causing harm. This paper follows the distinctions of Wang et al [4] to use the terms *misinformation* and *disinformation* as appropriate.

It is important to acknowledge at the outset that digital technology in health and social contexts presents both risks and opportunities for equity among different information audiences [5]. However, there has recently been a change in the reception and assessment of the role of the internet, social media in particular, among researchers, even predating the COVID-19 pandemic. In the early days of social media, researchers largely identified prosocial and altruistic uses of social media platforms such as Facebook and Twitter by the public. However, considerable disquiet about the impact of social media and its potential for the spread of “fake news” and the amplification of conspiracy theories has displaced the more positive evaluation that was apparent when social media was in its infancy [6]. In the majority of the current research, there is a view that digital technology, particularly social media, has amplified the problem of health misinformation. The risk most frequently identified, either explicitly or implicitly, is the susceptibility of ordinary users, who may be lacking sophisticated levels of health and digital literacies, to health misinformation. Further risks noted in the literature include disinformation disseminated by organized trolling networks and bots that can be hard to distinguish from human users. The recognition of these risks underpins an emerging policy discourse about the threat of health misinformation, particularly the role of social media in its spread, in which health information and misinformation has become a politicized issue. From one policy perspective, there is an assumption that social media users are vulnerable, even passive, recipients of health misinformation rather than reflective interpreters of the available information. The corollary of this is that correcting misinformation with authoritative knowledge will solve the problem. However, a survey of the literature suggested that neither assumption fully expresses the complexity of how health misinformation is disseminated, received, and used via the internet. This may be because although there is a growing body of evidence on the extent of online health misinformation, there is much less research about what users do with health misinformation, why users consume health misinformation, and why (and which) people believe health misinformation [7-9].

Methods

Overview

In this scoping review, we reviewed the current state of knowledge regarding the prevalence of online misinformation before and during the COVID-19 pandemic and the impact that has on users’ understanding of health information. We considered this context with special reference to patients’ understanding, health practitioners’ practice in response to that, and policy makers’ concerns. The pressures and distractions that health professionals face in attempting to mitigate the impacts of online health misinformation are discussed in relation to patients’ health and digital literacies and the politicization of health information and misinformation.

Information Sources

We conducted a comprehensive literature search to identify relevant studies that explored health misinformation on social media platforms. The search was conducted across multiple electronic databases, including PubMed, Scopus, Web of Science, and Google Scholar.

Search

The search terms included a combination of relevant keywords and phrases, including “health misinformation,” “social media,” “online health communities,” and “COVID-19 pandemic.” The search was not limited by publication date. Detailed search strategies are provided in Multimedia Appendix 1.

Study Selection

Our study selection process followed a scoping approach, where we aimed to identify and include studies that provided insights into the prevalence and challenges of health misinformation on social media platforms. Initially, 2 researchers independently screened titles and abstracts of the identified articles to determine their relevance. Articles that did not meet the inclusion criteria were excluded at this stage.

Inclusion Criteria

Articles were included if they discussed health misinformation on social media, addressed the challenges posed by health

XSL•FO
RenderX

misinformation, or were relevant to the period before, during, and after the COVID-19 pandemic.

Any disagreements between the 2 researchers were resolved through discussion and consensus. Full-text articles were then retrieved for the remaining studies, and a further assessment of eligibility was conducted based on the same inclusion criteria.

Data Extraction

We gathered information on (1) study objectives, (2) research methods, (3) findings, and (4) key themes related to health misinformation. This process was performed independently by 2 researchers, and any discrepancies were resolved through discussion.

Data Synthesis and Analysis

We adopted a scoping review content analysis approach to analyze the data extracted from the selected articles. The analysis process involved identifying key themes and patterns related to health misinformation on social media. The content analysis allowed us to gain a deeper understanding of the challenges posed by health misinformation and the strategies for its mitigation, both before and during the COVID-19 pandemic.

Results

Results of Search

In our article selection process (Figure 1), we initiated our search by identifying a total of 4563 articles from various databases. Following the removal of duplicates, 1295 articles were excluded, leaving us with 3268 unique articles. Subsequently, these articles underwent an initial screening, which involved evaluating their abstracts and titles, resulting in the exclusion of 2635 articles that did not align with our inclusion criteria. Further scrutiny was applied during full-text screening, which was conducted on 633 articles. Among these, 563 articles were found ineligible due to reasons such as not meeting the inclusion criteria (n=378 articles), being classified as literature reviews, editorials, or letters (n=174 articles), or the full texts being inaccessible (n=11 articles). A total of 70 articles were ultimately included in this scoping review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

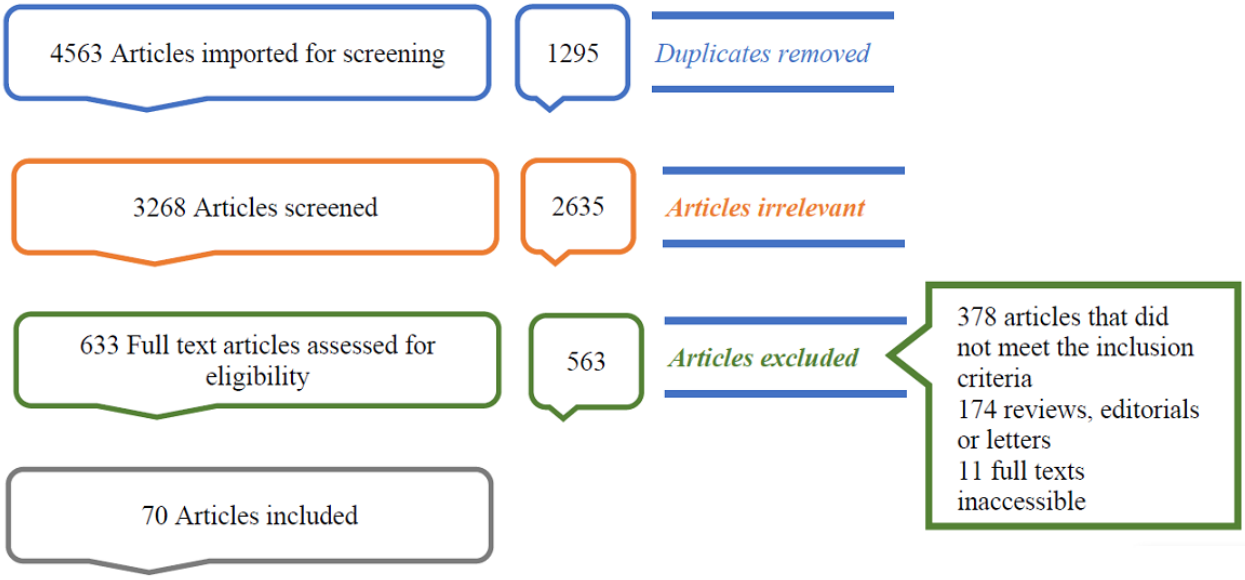

Characteristics of Included Documents (n=70)

The majority (65/70, 93%) of documents were published in the last 10 years and originated predominantly in North America (42/70, 60%), Europe (19/70, 27%), and Asia (11/70, 16%). The funding sources were mainly public (61/70, 87%). The documents were classified as original research papers (38/70, 54%), subjective “opinion” articles (editorials, viewpoints, commentaries, and letters to the journal; 11/70, 16%), and knowledge syntheses or reviews (9/70, 13%) which included systematic reviews (n=6), descriptive reviews (n=2), and 1 integrated theoretic review.

Extent and Impact of Health Misinformation Disseminated Across a Range of Outlets

Overview

This section will review the literature concerning the extent and impact of the problem of health misinformation, including the spread of antivaccination discourse. In a study by Wood et al [10] among health practitioners in North Carolina, 94.2% of the respondents reported encounters with patient health misinformation within the previous year. While the sources of this misinformation were not broken down and identified by Wood et al [10], several other studies linked patient health misinformation to the prevalence of health misinformation on social media sites, identifying the latter as a significant problem [4,11-15]. There is a growing consensus among researchers, health professionals, and policy makers about the need to confront, challenge, and even prevent the online dissemination

XSL•FO
RenderX

of health misinformation [16]. Since the emergence of online social networks, users have increasingly sought and shared health information on social media sites. It is estimated that around 70% of adult internet users search health matters on the web. With the emergence of social media platforms, there has been a rise in "peer-to-peer health care," through which individuals seek and share health information, forming online health communities with others who have similar health concerns [3]. In addition, health organizations and health professionals are increasingly using social media to disseminate and promote health information and advice. The opportunities for sharing and promoting good health information via the internet are evident, and it is important to acknowledge that in online health communities, users share experiences and receive and give different kinds of support, including emotional support, to cope with specific health conditions. However, the medium has also enabled the dissemination of health misinformation, and the prosocial aspects of sharing are also likely to involve the sharing of misinformation, putting the health of users at risk [3].

Misinformation Spreads on Social Media

There is a high degree of consensus among researchers that mainly because of the increasing popularity of social media, the internet has become a space for the dissemination and amplification of "fake news," misleading information, and rumor, including health misinformation and antivaccine conspiracy theories [17]. The COVID-19 pandemic has heightened these concerns, resulting in a proliferation of recent studies and rapid reviews focusing on the online spread of misinformation. Lee et al [18] proposed that the proliferation of health misinformation during the COVID-19 pandemic became a major public health issue. At the earliest signs of the emerging COVID-19 pandemic, the director-general of the World Health Organization, Tedros Adhanom Ghebreyesus, speaking at the February 2020 Munich Security Conference, expressed concern about the risk of an infodemic of health misinformation disseminated via social media, identifying "vaccine hesitancy" as 1 of the top 10 global health threats [19]. Bapaye and Bapaye [20] agreed that the risks of misinformation on social networking sites constitute a global issue, referring specifically to the COVID-19 infodemic.

However, this is not in itself a new problem; longstanding concerns about "fake news" and misinformation in traditional media have been evident since the early decades of the 20th century [21], and the prevalence of misinformation on internet platforms certainly predates the COVID-19 pandemic. Therefore, because the COVID-19 pandemic has only intensified the concern regarding health misinformation, it might be more appropriate to see the pandemic as symptomatic of, and crystallizing, the challenges of countering health misinformation in the digital age, as the development of digital technology and the internet have brought about profound changes in the capacity of both misinformation and disinformation to spread globally and amplify rapidly [4].

Suarez-Lledo and Alvarez-Galvez [16] undertook a review of 69 studies of health misinformation on social media to identify the main health misinformation topics and their frequency on different social media platforms. The studies surveyed used a variety of research methods, including social network analysis (28%), evaluation of content (26%), evaluation of quality (24%), content/text analysis (16%), and sentiment analysis (6%). Suarez-Lledo and Alvarez-Galvez [16] concluded that the incidence of health misinformation was highest on Twitter, in particular, regarding the use of tobacco and other drugs, with some studies citing 87% of such posts containing misinformation. However, health misinformation about vaccines was also prevalent, with around 43% of posts containing misinformation, with the human papillomavirus vaccine being the most affected. This review by Suarez-Lledo and Alvarez-Galvez [16] confirmed many of the findings from earlier surveys. For example, in their survey of 57 articles, Wang et al [4] found that the most frequently discussed topics were regarding vaccination and infectious diseases, including Ebola and the Zika virus. Other topics such as nutrition, cancer, water fluoridation, and smoking were also prevalent. The studies they surveyed had tended to find that a high degree of misinformation on these topics was being shared and liked on social media.

Lee et al [18] conducted a cross-sectional online survey in South Korea to examine the prevalence of COVID-19 misinformation and the impact of exposure to COVID-19 misinformation on beliefs and behaviors. They found that exposure to COVID-19 misinformation was associated with misinformation belief, which then resulted in fewer preventive behaviors. Therefore, they highlighted the potential of misinformation to undermine global efforts in disease control and argued that public health strategies are needed to combat the proliferation of misinformation. Bapaye and Bapaye [20] conducted a cross-sectional online questionnaire survey of 1137 WhatsApp users in India. They noted that most research on the prevalence of misinformation in social media has focused on Twitter and Facebook and on the Global North. Measured by age, researchers found that users aged >65 years were the most vulnerable to accepting the veracity of messages containing health misinformation (K=0.38, 95% CI 0.341-0.419) Respondents aged 19 to 25 years displayed much lower vulnerability (K=0.31, 95% CI 0.301-0.319) than those aged >25 years (P<.05). Measured by occupational category, users employed in nonprofessional occupations had the highest vulnerability (K=0.38, 95% CI 0.356-0.404); this was significantly higher than those of professionals and students (P<.05). Notably, the vulnerability of health professionals was not significantly different from those of other occupation groups (P>.05).

The authors concluded that in a developing country, WhatsApp users aged >65 years and those involved in nonprofessional occupations are the most vulnerable to false information disseminated via WhatsApp. Crucially, they noted that health care workers, who might be expected by laypersons to have expert knowledge, were as likely to be vulnerable to health misinformation as other occupation groups.

Antivaxxer Spread Before, During, and Beyond the COVID-19 Pandemic

Much of the current unease from researchers, understandably, centers on health misinformation about vaccines in the wake of

XSL•FO
RenderX

the COVID-19 pandemic. In particular, there is concern about the growth and spread of so-called antivaxxer misinformation and beliefs. In 2019, the United States had its biggest measles outbreak in 30 years, with most cases involving people who had not been vaccinated. Hotez [22] claimed that much of the reason for the growth of antivaccine beliefs is because of a campaign of misinformation. He argued that social media sites are meeting places for the sharing of antivaccine views. To evade social media platforms' automated moderation tools, which tend to focus on words, several antivaxxer groups, including one with around 250,000 members, began using visual codes, such as the carrot emoji, to hide antivaxxer content.

However, some of the misinformation has gained credibility because it has come from sources that laypersons would expect to be trustworthy. For example, in 1998, the British medical journal *The Lancet* published a paper by Dr Andrew Wakefield claiming a link between the measles, mumps, rubella vaccine and the onset of autism spectrum disorder. Wakefield's paper was later rebutted, and an overwhelming body of evidence now refutes its conclusions [23]. However, despite long being discredited, Wakefield's claims have remained a part of the antivaccine discourse. The persistence of the antivaccination narrative demonstrates the power of such discourses even in the face of evidence to challenge them.

Although strong antivaccine beliefs, and the more ambivalent attitude of vaccine hesitancy, have been around as long as there have been vaccines, until recent decades, they were on the margins. However, evidence supports the claim that they have been gaining momentum in the United States and Europe.

A survey by Skafle et al [24] aimed to synthesize the results from 19 studies in which the effect of social media misinformation on vaccine hesitancy was measured or discussed. The authors noted that the "vast majority" of studies were from industrialized Western countries. Only 1 study contained misinformation about autism as a side effect of COVID-19 vaccines. Nevertheless, the studies implied that information spread on social media had a negative effect on vaccine hesitancy and uptake. The conclusions from Skafle et al [24] were supported by data from online polling agencies. For example, a US YouGov poll from May 2020 found that only 55% of respondents would definitely take a COVID-19 vaccine if one were to become available, whereas 19% of respondents said that they would refuse and 26% were still undecided [25].

While much of the research about online vaccine discourse comes from the United States, there is also evidence that vaccine hesitancy has risen elsewhere. For example, in an Ipsos-MORI survey taken in December 2020, only 40% of respondents in France said they would take a COVID-19 vaccine, a figure symptomatic of a steep and swift decline in vaccine confidence in France [26]. However, interestingly, the same Ipsos-MORI poll indicated a rise in vaccine confidence among respondents in the United States since the earlier YouGov poll, cited earlier, by approximately 10% to 65%, and respondents in the United Kingdom expressed a still higher willingness to take a COVID-19 vaccine at approximately 77%. It is notable that in the United States and United Kingdom, the Ipsos-MORI results came after a period of intermittent lockdowns. The contrast with the results from France is, nevertheless, striking.

Understanding the Challenges Surrounding Health Misinformation

Here, we consider the challenges created by health misinformation on the web: (1) the role played by malicious actors on social media in spreading vaccine disinformation and misinformation and (2) how contextual and cultural issues have different effects on patients' understanding of what is considered genuine, valid, and authentic health information.

Spread of Health Misinformation on Social Media by Malicious Actors

One strand of research presents the issue of health misinformation as a contest between trolls and bots on the one hand and the voices of trustworthy public health agencies on the other [6]. This view was supported by Hotez [22] and Broniatowski et al [11]. The latter investigated the role of bots and trolls as malicious actors mobilizing vaccination discourse on the web. Their study focused specifically on vaccine-related health messaging on Twitter. Comparing the rates of vaccine-related messages, they found that sophisticated bots and Russian trolls tweeted at higher rates than "average users." However, the respective content from bots and trolls differed. Whereas bots communicated antivaccine messages, Russian troll accounts provided a seemingly balanced discussion of both provaccination and antivaccination arguments, implying an equivalence between them. The authors argued that amplifying and normalizing a debate is done with the purpose of sowing discord and may lead to undermining public confidence in scientific consensus about the effectiveness of vaccines. Wang et al [4] acknowledged that it is a challenge to readily distinguish between misinformation and disinformation on the web. They noted that disinformation, such as antivaccine propaganda, can unknowingly be spread by users with genuine concerns [4], as individuals increasingly seek health and healthy lifestyle information via the internet.

Contextual Factors Influencing the Reception of and Responses to Misinformation: Politicization of the Problem of Health Misinformation

The identification of online trolls, bots, and orchestrated networks as major contributors to the spread of health disinformation and misinformation is now part of mainstream political discourse in the United States. On July 16, 2021, a quarrel broke out between the president of the United States, Joe Biden, and Facebook over the spread of health misinformation on the company's social media platforms. Speaking to journalists, Biden blamed social media companies for a rise in the number of deaths from COVID-19 among the unvaccinated in the United States. Referring explicitly to Facebook, the president claimed that by allowing the proliferation of health misinformation on its platforms, the company was "killing people" [27]. Discursive interventions from politicians are never neutral; nevertheless, Biden's claim about the impact of health misinformation on social media is backed up by many of the studies surveyed for this paper. Facebook immediately rebutted Biden's accusation by citing

XSL•FO
RenderX

their rules, introduced in February 2021, which banned posts that make identifiably false claims about vaccines. Furthermore, Facebook challenged Biden's claim by asserting that not only has Facebook provided more authoritative information about COVID-19 and vaccines than any other internet site, reaching 2 billion people with such posts, but also that the platform's vaccine finder tool had been used by more than 3 million Americans.

These figures suggest that although antivaxxer groups find ways to evade detection, their reach may be countered by that of information grounded in current science. A spokesperson for the company said that, far from killing people, "The facts show that Facebook is helping save lives. Period" [27]. The argument between Biden and Facebook may indeed signal more lay awareness of the problem and echo the concerns of the recent academic research about the dissemination of health misinformation by organized bot and troll networks. Framed as it is, in terms of apportioning the blame for the spread of health misinformation, Biden's intervention mirrors much of the academic discourse in the United States on the subject. However, it is also symptomatic of the politicization of health misinformation, arguably accelerated by the COVID-19 pandemic, which may thwart evidence-based decision-making. This point was emphasized strongly by Kyabaggu et al [5]. They framed the problem of pervasive misinformation and disinformation in terms of prime movers and beneficiaries who use it to advance sociopolitical agendas and entrench asymmetrical power, especially in times of uncertainty and threat, such as the COVID-19 pandemic.

Kyabaggu et al [5] identified government failures to adopt evidence-informed decision-making. They noted that such failures have costs that not only are economic but, crucially, result in poorer health outcomes. They cited as an example the United Kingdom government's initial prevaccine herd immunity strategy. The intention of this strategy was to allow SARS-CoV-2 to indiscriminately spread to a critical mass to build up population immunity. The authors noted that this was "a particularly concerning example of evidence framing by a government." Kyabaggu et al [5] argued that public acceptance of health risk messages and adoption of health-protecting behaviors is highly contingent on the degree to which governments engage in evidence-informed decision-making and communicate this basis effectively. The authors cited several instances of government actors failing to recognize misinformation, disseminating inconsistent or inaccurate information, and not using evidence- and information-based decision-making processes. In recent years, the public policy discourse in the United Kingdom has been veering away from evidence- and information-based decision-making, as politicians have denounced "experts" and their "influence" on policy [28,29].

Finally, Gruzd et al [30] reported on the impact of coordinated link-sharing behavior to spread and amplify conspiracy-related misinformation. They found a coalition of Facebook accounts that engaged in coordinated link sharing behavior to promote COVID-19 related misinformation. This coalition included US-based pro-Trump, QAnon, and antivaccination accounts.

Contextual Factors Influencing the Reception of and Responses to Misinformation: Health Literacies and Inequality

While the approach of Broniatowski et al [11], for example, provided a persuasive account of ways in which online health misinformation can be disseminated, there are limitations to this approach, as it did not provide an account of how users respond to the misinformation they encounter. The responses of ordinary users were assumed rather than investigated. Research by Vosoughi et al [31] provided a caveat to the claim that it is bots that accelerate the spread of misinformation. Their work supported that of Broniatowski et al [11] in suggesting that bots spread accurate and false information at the same rate. However, Vosoughi et al [31] also explained that misinformation spreads more rapidly than accurate information because humans, rather than bots, are more likely to spread misinformation [31]. This claim was further supported by Wang [32], who suggested that in democracies, where ideas compete for attention in a marketplace, accurate scientific information, which, for the layperson, may be boring or difficult to understand, is easily crowded out by information that is more easily grasped or sensational. Mokhtari and Mirzaei [12] located this problem specifically in the context of the COVID-19 pandemic. They considered that high mortality from COVID-19, its complexity, and its unknown features resulted in fear, anxiety, and mental pressure among people worldwide. To allay anxiety, people needed health information literacy, defined by the American Library Association as a set of abilities individuals require to recognize when information is needed and to locate, evaluate, and use it effectively [33]. In addition, Wang [32] noted that individuals are differentially vulnerable to health misinformation depending on their level of health literacy and that models need to account for this. Mokhtari and Mirzaei [12] argued that not only information and health literacies but also media literacy are needed. However, studies in the field of health literacy suggest that significant inequalities in health and digital literacies exist.

Researchers have argued that "vastly undervalued and unrecognized" health literacy ought to be considered the best "social vaccine" for preventing COVID-19 in populations [5]. However, inequalities in health literacy persist. Kyabaggu et al [5] defined health literacy as encompassing cognitive and social skills that determine individuals' motivation and ability to access, understand, and use information, including quantitative health risk information, in ways that promote and maintain good health across the life course. They asserted that health literacy is an essential self-management skill and community resource for health, noting that health literacy is positively associated with patients' involvement in clinical decision-making, willingness to express health concerns, and compliance with clinical guidance. However, despite research demonstrating the importance of health literacy, evidence, even from high-income countries, suggested relatively low levels of health literacy.

Kyabaggu et al [5] drew a link between health literacy and digital literacy. They suggested that the latter can be understood as health literacy in digital information and technology spaces. They argued that inequalities in health outcomes are exacerbated

XSL•FO
RenderX

by a widening digital divide. While digital technology in health and social contexts presents both new risks and opportunities for equity in different information audiences, the ways in which power and privilege operated in the COVID-19 misinformation discourse have not been sufficiently examined. Although socially and economically disadvantaged groups were at a greater risk of exposure to COVID-19, their voices and experiences were often marginalized. In addition, inequalities in access to accurate information are not only related to issues of digital access and literacy but are also situational. For example, disadvantaged individuals may have fewer social connections, and low pay may necessitate longer working hours, militating against individuals having the resources of time and energy to seek out accurate health information and enhance their level of health literacy.

The experiences of specific groups may also go unreported. Quraishi [34] addressed the impact of misinformation on South Asian students—a fast-growing group in the United States, but one that often receives little media attention. Quraishi [34] concluded that there is a relationship between the COVID-19 pandemic and students' academic performance and mental health, as well as an increase in the spread of misinformation regarding COVID-19 public safety guidelines.

Older adults can be a vulnerable group in relation to their comparatively poor digital literacy. Zhou et al [35] reported on the accuracy of older adults in judging health information credibility. They found that on average, participants only successfully judged 41.38% of health articles. Attractive headlines increased participant credibility judgments on the content, and of the articles shared with others, 62.5% contained falsehoods.

Contextual Factors Influencing the Reception of and Responses to Misinformation: Cultures and Values

Larson and Broniatowski [19] argued that developing the kinds of literacy advocated by Mokhtari and Mirzaei [12] and Tully et al [2] will not address the deep-seated problems they identified. The work by Kyabaggu et al [5] supported this, and noted that the infodemic crisis is not merely a health and digital literacy issue. Some demographics may be more vulnerable to persuasive communication from broader sociocultural forces. Kyabaggu et al [5] argued that in considering the social determinants of health, attention must be paid not only to digital and health literacies but also to the ways in which these literacies coexist and interact with other influences. Larson and Broniatowski [19] suggested that one of the strongest determinants of vaccine confidence or vaccine hesitancy is the level of trust or distrust in the institutions that produce vaccines. A higher level of trust encourages the willingness to accept a high level of risk for a greater benefit. A lower level of trust militates against the acceptance of even a low level of perceived risk. For Larson and Broniatowski [19], it is not simply the presence of misinformation on social media networks but the social and cultural context of users' reception of that information that influences responses. Health information operates in a complex and contentious social world. Individuals and communities respond to new information in terms of already developed political, cultural, and social values that influence whether they trust or distrust authority. Populations may be characterized by trust or mistrust of scientific institutions and government. Trust has been eroded through the exposure of fraud, research scandals, and misconduct by major multinational pharmaceutical companies, for example. Communities may be predisposed to distrust the government and its agents depending on their own status or identity. According to Goldenberg [36], these contexts can make misinformation and health conspiracy theories compelling.

Strategies to Correct Online Misinformation

We address the additional pressures on health professionals in communicating accurate information to mitigate the effects of misinformation, particularly with regard to the additional requirements imposed as a result of the precautions being taken during the pandemic. One area of disagreement in the literature concerns the usefulness of user correction response.

Research Into User Correction Strategies

There is some disagreement as to whether engagement with misinformation by users spreads and reinforces it or even whether extended debates over health misinformation cause users to doubt the possibility of knowable facts. For example, Broniatowski et al [11] argued that when ordinary users directly confront vaccine-skeptic messages from bots, it only serves to legitimize the "debate." By contrast, Tully et al [2] argued that social media users have a role to play in either spreading or stopping the spread of misinformation across platforms. Their research aimed to uncover what factors influenced users' responses. Tully et al [2] acknowledged that a range of factors can influence the spread or prevention of misinformation, including the behavior of malicious actors such as bots and trolls; the platform's terms of service; and content moderation policies. As already noted, while most users are not creators of misinformation, they may spread and amplify it by liking, sharing, or replying. In opposition to the work of Broniatowski et al [11], Tully et al [2] argued that the content of engagement is particularly important, as their research suggested that multiple corrections by social media users may be required to reduce misperceptions. However, they claimed that most people simply ignore misinformation when they see it on social media.

Tully et al [2] noted the promise in mobilizing users to engage in such correction, given the vast numbers of users on these sites, in comparison with professional fact-checkers and health authorities.

They considered whether the tone of a correction would influence perceptions of the credibility of the message. However, despite some mixed evidence, they concluded that overall, the tone was not a significant factor and that neutral, affirmative, and uncivil corrections were all effective at reducing misperceptions. They found that participants were generally unlikely to reply to the misinformation tweet. However, their content analysis of hypothetical replies suggested that when users did reply, they mainly provided correct information, particularly after seeing other corrections. Tully et al [2] concluded that user corrections offer "untapped potential" in responding to misinformation on social media, but further work is needed to consider how users can be mobilized to provide

XSL•FO
RenderX

corrections, given their overall unwillingness to reply. However, a limitation of the experimental approach of Tully et al [2], acknowledged by the researchers, is that in asking individuals what they would hypothetically do, this may not reflect what they actually do in a real social media setting, especially in relation to an issue they care more strongly about. Although the experiment gauged attitudes, it did not delve into how strongly these attitudes were held. It is also not clear to what degree corrections were effective at reducing misperceptions and how reductions were measured.

By contrast, the results of experimental studies by Ittefaq [37] and Mourali and Drake [38] suggested that correcting misinformation is by no means a straightforward proposition. They noted the previous research on rebuttal, which suggested that properly designed corrections can mitigate the effects of misinformation. However, such studies have tended to compare responses to misinformation followed by correction with responses of a control group that receives no correction or receives an alternative correction. Mourali and Drake [38] argued that this static approach misses the dynamic nature of social media debate. They noted that the correction of misinformation is generally followed up with a rebuke by the original poster, inciting further correction and prolonged back-and-forth debate. Mourali and Drake [38] cited previous studies showing that exposure to conflicting information about health topics, including mammography, nutrition, and the human papillomavirus vaccine, may increase confusion and negative attitudes toward that particular health topic. The researchers found that initial exposure to misinformation had a negative impact on attitudes and intentions toward masking, consistent with previous studies that concluded that exposure to misinformation negatively impacts attitudes and intentions toward behaviors favored by science. Also consistent with previous research, they found that the first correction of the false claim improved attitudes and intentions toward masking. The authors suggested that this effect is partially explained by a decrease in the perceived strength of the argument underlying the false claim. However, this initial improvement diminished on further exposure to false claims and refutation attempts. This finding confirmed their hypothesis that extended exposure to false claims and refutation attempts appears to weaken belief in the possibility of objective knowledge, leading to less positive reactions toward masking as a science-based behavior. They concluded that the level of exposure to contradictory information needs to reach a certain threshold before it affects perceived truth objectivity. However, although people are more likely to share misinformation when its content is consistent with their existing beliefs or when its message is simple, direct, or sensational, correcting misinformation does reduce its likelihood of being shared on social media, an effect that persists even after multiple exposures.

Mourali and Drake [38] noted that each social media platform exhibits particular interaction norms, which may impact how users interpret the conversation. As their study was limited to a single platform, Reddit, and the debate was restricted to 4 exchanges between only 2 protagonists, the researchers acknowledged that these aspects limit the generalizability of the results. They suggested that future research could attempt to replicate their findings on different social media platforms, and to include more than 2 protagonists and more than 4 exchanges. They noted further that although extended debates are common on social media, it is not known how frequently they occur, echoing the comments by Suarez-Lledo and Alvarez-Galvez [16] that the extent of misinformation is not clear.

In contrast to the fairly sanguine view of Tully et al [2] about the potential of users to spread corrective information, Mourali and Drake [38] problematized the position, pointing to the potential for more complex and uncertain outcomes, whereas Larson and Broniatowski [19] argued that although the importance of correcting misinformation, item by item, should not be diminished, only if underlying issues driving misinformation are addressed can, for example, long-term vaccine confidence in populations be sustained. They argue that simply responding to misinformation with factual corrections is not likely to reverse the dissent that has been evident among antivaxxers or to necessarily persuade the more ambivalent vaccine-hesitant individuals. They identified deeper social and cultural issues at play, which have been discussed in this paper in the previous sections.

Research Into Effective Models to Accomplish the Automatic Detection of Health Misinformation in Online Health Communities

Here, we consider examples of research into the automatic detection of health misinformation in online health communities. Zhao et al [3] began from the premise that there is a vast amount of health misinformation, creating a challenge for health communities in identifying misinformation. Rather than relying on users' ability to correct misinformation, they proposed that there is a need for an effective model to achieve automatic detection of health misinformation in online health communities. This view was also put forward by Weinzierl and Harabagiu [39]. Focusing specifically on COVID-19 vaccine misinformation, they argued that automatic detection of misinformation on social media is an essential first step in delivering interventions designed to address vaccine hesitancy.

Zhao et al [3] identified much of the existing analysis as concentrating on the linguistic features of communications only. They wanted to examine the underresearched area of whether integrating user behavioral features with linguistic features, sentiment features, and topic features could effectively distinguish misinformation from accurate information in online health communities. Their study combined the aforementioned features to build a detection model targeting misinformation in online health communities' contexts. The behavioral features targeted were discussion initiation, interaction engagement, influential scope, relational mediation, and informational independence. Descriptions of these behavioral features are reproduced in Table 1.

XSL•FO
RenderX

Table 1. Descriptions of the behavioral features.

Behavioral feature	Measurement	Description
Discussion initiation	The number of threads a user created	To reflect the activity of a user in terms of initiating new discussions
Interaction engagement	The number of replies and the number of replies to a reply a user created	To reflect the activity of a user in terms of interacting with other users
Influential scope	Degree centrality	To reflect the potential communication ability of a user
Relational mediation	Betweenness centrality	To assess the potential of a user for the control of communication in the community
Informational independence	Closeness centrality	To assess the ability of a user to instantly communicate with others without going through many intermediaries

The authors tested their detection model on a data set collected from a real online health community, selecting as their data source Zibizheng Ba, an autism forum on the Baidu Tieba online health community site hosted by the Chinese web service Baidu. Baidu Tieba claims to be one of the largest interest-based discussion platforms in China. Users can generate topic-based discussion forums on the platform, share information, and make friends with other users. Posts on Baidu Tieba are indexed by Baidu, China's most popular search engine, so users can readily find misinformation when searching for health-related information through the search engine. The authors developed a python-based web crawler to collect data from the forum. To train the health misinformation detection model, 5000 records were sampled from the whole data set by stratification according to 3 types of records (ie, thread, reply, and reply to reply) using stratified sampling methods. Therefore, the constituent types of the records (ie, thread, reply, and reply to reply) in the sample data set were consistent with the composition of the whole data set.

The researchers applied the elaboration likelihood model (ELM). The model, originally developed by Petty and Cacioppo [40] to explain attitude change, has been used extensively in advertising to try to influence consumers.

Overall, 4 types of misinformation were identified through their coding analysis, and the model correctly detected about 85% of the health misinformation. Their results also indicated that behavioral features were more informative than linguistic features in detecting misinformation. The authors concluded that their results not only demonstrated the efficacy of behavioral features in health misinformation detection but also offered both methodological and theoretical contributions to misinformation detection by integrating the features of messages as well as the features of message creators. Others have also highlighted the problems posed by misleading visual information [41].

It is worth noting that during the pandemic, the UK National Health Service (NHS) began using Twitter to promote provaccine messaging, which closely follows a combination of the features suggested by Zhao et al [3]. When users searched for the term "vaccine" or related terms, the top post was a message prominently displaying the NHS logo, identifying it as reputable and trustworthy. The tweets contained links to NHS websites providing information about vaccines and COVID-19. The posts differed in linguistic content and visual design. For example, one featured only written text on a white background and stated in bold, "Know the facts." Another featured a large image of a happy minority ethnic family, washing dishes together, with the message that the COVID-19 vaccine decreases household transmission by up to half. The contrasting designs suggest that the message was targeted specifically to users' timelines. It was also apparent that elements of ELM were being applied, combining the features identified by Zhao et al [3] in different ways.

Weinzierl and Harabagiu [39] adopted a different method than Zhao et al [3], reversing the more commonly used classification approach. The authors of each study claimed strong results in identifying health misinformation on social media platforms. However, Nabożny et al [42] argued that the current automatic systems for assessing the credibility of health information are not sufficiently precise to be used without supervision by human medical expert annotators.

Barve and Saini [43] have reported on their use of automated fact-checking using a coded content similarity measure (CSM). In this approach, the CSM showed improved accuracy (91.06%) compared to the accuracy of the Jaccard similarity measure (74.26%). Further, the algorithmic approach outperformed the feature-based method.

Neither Zhao et al [3] nor Weinzierl and Harabagiu [39] recorded what happens when misinformation is detected. Research from Broniatowksi et al [44] suggested that once detected, steps taken by social media platforms such as content removal or deplatforming may not be effective in stemming the spread of misinformation and may even be counterproductive. Social media platforms use a combination of "hard" and "soft" content remedies to reduce the spread of health misinformation. Soft remedies include warning labels attached to content and downranking of some content in web searches, whereas hard remedies include content removal and deplatforming of accounts. Hard remedies are controversial and have given rise to accusations of censorship. For the authors, short-term evidence for the effectiveness of hard remedies is in any case mixed, and long-term evidence is yet to be examined. Their study focused on Facebook and found that while hard remedies did reduce the number of antivaccine posts, they also produced unintended consequences. Provaccine content was removed, and engagement with the remaining antivaccine content repeatedly recovered to prepolicy levels. Worryingly, this content became more misinformative, more politically polarized, and more likely to be seen in users' news feeds. The authors

XSL•FO
RenderX

explain these results as a product of Facebook's architecture, which is designed to promote community formation. Members of communities dedicated to vaccine refusal seek out misinformation. To meet this demand, and to circumvent content moderation efforts, antivaccine content producers post links to external sources of misinformative content, such as Bitchute, Rumble, Gab, and Telegram, in lieu of more mainstream platforms that had implemented similar content removal policies (eg, YouTube and Twitter). Broniatowski et al [44] argued that Facebook's policy reduced the number of posts in antivaccine venues but was not successful in inducing a sustained reduction in engagement with antivaccine content, including misinformation. The authors noted that alternative platforms often host politically extreme right-wing content. Therefore, they argued that Facebook's content removal policies may have the unintended consequence of radicalizing their audiences, and their findings suggested the need to address how social media platform architecture enables community formation and mobilization around misinformative topics when managing the spread of online content.

These studies advocate for the automatic detection of health misinformation. However, work that calls into question the ability of automatic detection to operate without human intervention has also been discussed. In addition, there are questions raised in the literature about what should be done when misinformation is detected and concerns about whether content removal or deplatforming of accounts are the most effective ways to reduce the spread of health misinformation or may even be counterproductive.

The Roles of Health Practitioners

Overview

The discussion so far has highlighted the complex and multifaceted dimensions of the context of online health misinformation in which health practitioners must operate. As noted in our introduction, a study of health practitioners in North Carolina found that nearly 95% had encountered patient health misinformation within the previous year [10]. There is very little research on the amount or effectiveness of training received by health professionals to prepare them for engaging with patients about health misinformation. Wood et al [10] found that most respondents had not received relevant training despite overwhelmingly reporting encountering health misinformation.

Nevertheless, within the literature, there is no shortage of advice from researchers and health professionals addressed to health practitioners on how to approach and correct health misinformation. This advice stems from both original research studies and reviews of best practices featured in peer-reviewed medical and health journals. Such advice centers on the need for health practitioners to understand misinformation and how to address it. Health practitioners are advised of the need to be aware of health myths and urged to dismantle them in providing accurate health guidance [45,46]. Practitioners are further advised that misinformation and pseudoscience are appealing to those seeking certainty because they present information in absolutes, whereas medical science is often ambiguous and contingent. Health practitioners are also encouraged to learn how to message more clearly and to mimic the strategies of misinformation [45]. One study recommends that "practitioners familiarize themselves with the tools of scientific enquiry and consider the pros and cons of various conspiracy evaluation guidelines" [47]. Thompson [48] reports on the activity of health professional influencers and pedagogues in combating misinformation. However, the effectiveness of such social media influencers who are also health professionals remains unclear. At the same time, there is some acknowledgment in this body of literature that misinformation cannot simply be offset with facts, confirming the challenges, discussed earlier, of simply engaging in online refutation. Addressing misinformation also depends on meeting patients' emotional needs [45,49].

In this context, the one-to-one patient-provider relationship in the practice setting is perceived as paramount [45]. As suggested by much of the research, source credibility, or trust, is understood to be the strongest driver of effective correction strategies [50]. It is argued that health care practitioners have the unique opportunity to guide patients toward high-quality, evidence-based medical information [10]. However, it is also noted that practitioners will need patience in their efforts to persuade patients to abandon strongly held self-beliefs, however harmful. Doing so may mean patients relinquishing membership of online communities that have become integral in their lives and even their identities. As noted earlier, belief in misinformation is often persistent in the face of evidence. Success is more likely when individuals are encouraged to reexamine their information sources, alongside new information providing additional context, rather than simply characterizing the individual's beliefs as wrong [51]. Kyabaggu et al [5] commented that good health communication needs to be tailored to the underlying cause of the misinformation problem, and efforts should be made to take on board inequalities within populations to create accurate, low-barrier, targeted health risk messaging. Skafle et al [24] contended that to challenge misconceptions, false claims need to be openly addressed and discussed with both cultural and religious awareness in mind. Guidance for practitioners noted that while responding to patient questions about alternative or unproven therapies may become laborious, a strong bond of trust between health practitioner and patient gives a patient a feeling of being supported and increases their adherence to treatment [52]. Rather than waiting for patients to raise misinformation issues, health care practitioners are advised to anticipate and proactively address potential misinformation and myths with patients. For example, the mortality rate for pediatric cancer has risen during the COVID-19 pandemic because of delayed access to medical care, but misinformation related to COVID-19 may also be a contributing factor [53]. The literature highlights the challenge of navigating the information and misinformation and the need for health practitioners to communicate with their patients more effectively. However, such efforts are not always successful. Some of the factors that may prevent effective communication of good health information have already been raised in this paper. They are revisited and discussed in the next section, along with other stressors for health practitioners.

Stressors for Health Practitioners

Challenges for health practitioners include time pressures and the additional burdens placed on them during the COVID-19

XSL•FO
RenderX

pandemic. These additional pressures add to the issues health practitioners face in trying to mitigate the impact of misinformation. The following is a brief overview of these issues.

On the one hand, administrative burdens placed on practitioners frequently deny them time for dialogue with their patients [52]. On the other, in different contexts, practitioners may be coping with a lack of proper facilities; poor infrastructure for patient care; insufficient or ineffective personal protective equipment; lack of awareness among the general population; poor compliance with preventive methods; and the fear of being infected with the virus, as they too are exposed to misinformation. During the COVID-19 pandemic, health practitioners were considered more vulnerable than other workers to developing psychological problems and other stress-related disorders, as they treated patients confirmed with COVID-19 while also dealing with misinformation [54].

As noted above, practitioners are recommended to invest in developing high levels of patient trust and to proactively correct health misinformation. However, recommendations presuppose that health practitioners necessarily have the resources to do these things well. Some of the materials produced to educate patients are not always reliable or evidence based, resulting ultimately in a loss of trust on the part of patients [52]. In addition, as noted previously, health practitioners themselves are not necessarily immune from accepting health misinformation as credible. Evidence about the level of knowledge and understanding of COVID-19 among practitioners reveals its unevenness. A study of dentists and oral health practitioners' knowledge about COVID-19 suggested that their knowledge was at a relatively high level [55]. By contrast, a study of 310 eye care professionals in Nepal revealed some knowledge but also some acceptance of misinformation. Symptoms of COVID-19 were known to 94% of participants, but only 49% of participants were aware of how the disease is transmitted. More significantly, 41% of participants believed that the consumption of hot drinks helps to destroy the virus, in contradiction to World Health Organization information. The mean overall "knowledge" performance score, as measured by the benchmarks set by the researchers, was 69.65% [56].

A qualitative study to investigate primary health care practitioners' perceptions and understanding of the COVID-19 pandemic was conducted in KwaZulu-Natal, South Africa. The study collected data from 15 participants at 2 different clinics situated in rural KwaZulu-Natal. Participants comprised nurses, physiotherapists, pharmacists, community caregivers, social workers, and clinical associates. Data were collected through individual, in-depth face-to-face interviews using a semistructured interview guide. The participants reported prepandemic and pandemic experiences of fear or denial. There was a perception of poor preparation for the COVID-19 outbreak. The findings also revealed participants' misperceptions regarding the nature of the COVID-19 pandemic. Researchers concluded that respondents' misunderstandings regarding the pandemic were primarily a result of misinformation found on social media [57].

The discussion in this section so far has highlighted the significant potential of health practitioners in mitigating the impact of online health misinformation. However, it has also underlined factors that may militate against health practitioners' ability to do so effectively. Not least of these is the issue of health practitioners' own knowledge, which coexists with other stressors for health practitioners in combating misinformation. The discussion will now consider health information management (HIM) as a tool for supporting health practitioners' knowledge base as one element in a multifaceted strategy for combating misinformation on the web.

HIM as a Mitigation Strategy

We have seen there is a need for health practitioners to be supported with evidence-based knowledge that they can share with patients. Kyabaggu et al [5] argued that the COVID-19 pandemic has demonstrated that in an infectious health crisis, the gathering of accurate and reliable data to assist with the public health response is essential. They highlighted the importance of HIM professionals in supporting contact tracing and syndromic surveillance, as well as in mapping and forecasting health data. They noted that the generation of health information supports the continuum of care and the setting of targets and indicators and aids the planning, monitoring, and evaluation of health programs locally and globally. The health information produced also underpins the development of equitable, efficient, and accessible health care systems, contributing to improving public health initiatives and outcomes. Kyabaggu et al [5] emphasized the importance of an area of HIM, currently in its early stages, that deals with gathering and identifying evidence about the structural inequalities that underlie the disparities in vulnerability to health misinformation discussed in this paper. The collection of rich, high-quality information, including patient-reported experience, outcome measures, and culturally appropriate identity data, can enable health practitioners and public health advisers serving the most disadvantaged and underrepresented communities to use more tools of advocacy for patients.

The authors noted that advances in technology, including artificial intelligence, have the potential to relieve some of the pressures and constraints on health practitioners working on the front line during crises such as the COVID-19 pandemic, allowing more time for one-to-one engagement with patients. Kyabaggu et al [5] advocated for the content expertise of health information managers to serve health practitioners by delivering patient-facing information triaging services; constructing user-friendly knowledge representations, such as data visualizations; and developing information interpretation tools, such as decision aids, plain language summaries, and supplementary explanatory information and metadata. Kyabaggu et al [5] identified the interdisciplinary underpinnings of HIM as essential in contributing to the educational, informational, and decision-making support for addressing current and future infodemic management crises.

XSL•FO
RenderX

Discussion

Summary of Results

Within the literature, there is a consensus that there exists a significant problem of online health misinformation disseminated via the internet on social network platforms, often by online health communities. It is apparent that while users seek trustworthy sources of health information, they are unequally equipped to assess its credibility. This is partly because some groups lack sufficient levels of health and digital literacies, which may be exacerbated by concomitant social and economic inequalities. Reception of, and response to, online health misinformation is also shaped by users' cultural contexts, values, and experiences, which may hinder trust in scientific institutions and governments. Evidence suggests that some demographics are more vulnerable to accepting health misinformation as credible and that health practitioners are unevenly prepared in the context of new global health crises, such as the COVID-19 pandemic. Furthermore, the evidence of disparities in positive and negative attitudes toward vaccination highlights a need to pay specific attention to regional and national settings, even in the current global context. Preexisting levels of local trust in vaccine providers may be a significant factor to consider. While the validity and reliability of YouGov polls are limited, nevertheless, the data from an admittedly narrow range of sources suggests that vaccine confidence may have become more fluctuating and potentially vulnerable to destabilization in the digital era.

While online mitigation strategies such as user correction and automatic detection may have their uses, their effectiveness is contested, and some studies suggest they may even be counterproductive. Our analysis of the available literature indicates that the effectiveness of these strategies varies and needs further evaluation [42,58]. The issue of online health misinformation is further complicated by the operation of malicious actors and politicization of the issue, particularly during the COVID-19 pandemic, militating against the equitable and trusted dissemination of evidence-based knowledge. The role of health practitioners in this context is a challenging one. Research suggests that on the one hand, they are still best placed, at the front line of care, to combat health misinformation with science-based knowledge and advice. On the other hand, the stressors identified in this review create barriers to their abilities to do this well. Constraints of time and lack of supporting infrastructure add to the knowledge deficit noted earlier. Our review underlines the complexity of the environment in which health practitioners operate and calls for greater support and resources to enable effective mitigation of health misinformation [59]. Investment in HIM at local and global levels could address all 3 deficits, creating the potential for health practitioners to enhance their capacity to build trust via knowledgeable one-to-one communication with patients.

Limitations

The limitations of this study are the following: First, the constraints of time and space have necessarily limited the scale and scope of the survey. Second, the study of online health misinformation is a growing field, and inevitably, the nature of the issue means that new evidence is emerging at a rapid rate. In particular, new knowledge and further reflection in the wake of the COVID-19 pandemic will continue to shed new light on the subject. Our study acknowledges these limitations and emphasizes the dynamic nature of the field.

Conclusions

Our survey of the literature on online health misinformation has revealed a complex and multifaceted context in which health practitioners must operate. As the world renormalizes following the pandemic, a collaborative global interdisciplinary effort to provide equitable access to timely, accurate, and complete health information will be needed to support health practitioners in combating the impact of online health misinformation. Academic research will need to be disseminated into the public domain in a way that is accessible to the public to counter misinformation and educate populations concerning how science is carried out. Our conclusions drawn from this review stress the urgency of effective strategies and collaborative efforts to mitigate the prevalence and impact of health misinformation on a global scale. Without strategies for equipping populations with the health and digital literacies required to interpret and use information appropriately, the prevalence of online health misinformation will continue to pose a threat to global public health efforts, disproportionately affecting vulnerable and resource-limited populations. Although social media platforms have a responsibility to correct misinformation, governments will need to engage in evidence-informed decision-making and invest in HIM to support frontline health practitioners in their work, enhance population health literacy, and strengthen evidence-informed decision-making at all levels.

Several issues for further investigation arise from the findings of this review. These include the following:

- The long-term impact of COVID-19 vaccine hesitancy
- Whether the COVID-19 pandemic has intensified or diminished information literacy, and the related question of whether the pandemic will incentivize health information literacy
- The effects of social and cultural differences on the long-term traction of future health misinformation
- Whether social and economic inequalities will become less or more pronounced in the face of a global pandemic
- The comparative effectiveness of strategies to enhance populations' media and digital literacies to facilitate the mitigation of health misinformation and its effects
- The influence of state actors on the propagation of health misinformation on the web
- The extent to which academic research has been disseminated into the public domain in a way that is accessible to the public, and the effectiveness of strategies to do so to counter misinformation and educate populations concerning how science is carried out

XSL•FO
RenderX

Acknowledgments

This research was funded by the School of Computing and Communications at the Open University. It allowed researchers across several faculties to collaborate and build a research team that focused on the experience of health practitioners with misinformation and its impact on their job practice. The authors would also like to thank Tracie Farrell and Nashwa Ismail for their invaluable suggestions and recommendations, as well as their assistance in the article screening process.

Data Availability

The data analyzed in this study are derived from published articles available on Google Scholar. All articles included in the review are cited in the reference list. No additional data or code were collected or generated as part of this study.

Authors' Contributions

The study was conceptualized by DK; funding acquisition was managed by DK; data were curated by DK, AK, MM, and IK; formal analysis was conducted by DK and MM; the investigation was carried out by AK and MM; the methodology was designed by DK and MM; project administration was overseen by DK; resources were provided by DK; supervision was carried out by DK; validation was conducted by DK, AK, MM, and IK; visualization was handled by DK and MM; writing (original draft preparation) was done by DK; and writing (review and editing) was carried out by DK, AK, and MM. All authors reviewed and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed search strategy.
[DOCX File , 28 KB - jmir_v26i1e38786_app1.docx]

References

1. Ismail N, Kbaier D, Farrell T, Kane A. The experience of health professionals with misinformation and its impact on their job practice: qualitative interview study. JMIR Form Res 2022 Nov 02;6(11):e38794 [FREE Full text] [doi: 10.2196/38794] [Medline: 36252133]
2. Tully M, Bode L, Vraga EK. Mobilizing users: does exposure to misinformation and its correction affect users' responses to a health misinformation post? Soc Media Soc 2020 Dec 10;6(4):205630512097837. [doi: 10.1177/2056305120978377]
3. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Inf Process Manage 2021 Jan;58(1):102390. [doi: 10.1016/J.IPM.2020.102390]
4. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
5. Kyabaggu R, Marshall D, Ebuwei P, Ikenyei U. Health literacy, equity, and communication in the COVID-19 era of misinformation: emergence of health information professionals in infodemic management. JMIR Infodemiology 2022;2(1):e35014 [FREE Full text] [doi: 10.2196/35014] [Medline: 35529308]
6. Sutton J. Health communication trolls and bots versus public health agencies' trusted voices. Am J Public Health 2018 Oct;108(10):1281-1282. [doi: 10.2105/AJPH.2018.304661] [Medline: 30207762]
7. Kim S, Capasso A, Ali SH, Headley T, DiClemente RJ, Tozan Y. What predicts people's belief in COVID-19 misinformation? A retrospective study using a nationwide online survey among adults residing in the United States. BMC Public Health 2022 Nov 18;22(1):2114 [FREE Full text] [doi: 10.1186/s12889-022-14431-y] [Medline: 36401186]
8. Morgan JC, Cappella JN. The effect of repetition on the perceived truth of tobacco-related health misinformation among U.S. adults. J Health Commun 2023 Mar 04;28(3):182-189. [doi: 10.1080/10810730.2023.2192013] [Medline: 36938585]
9. Nan X, Wang Y, Thier K. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. Soc Sci Med 2022 Dec;314:115398. [doi: 10.1016/j.socscimed.2022.115398] [Medline: 36327631]
10. Wood JL, Lee GY, Stinnett SS, Southwell BG. A pilot study of medical misinformation perceptions and training among practitioners in North Carolina (USA). Inquiry 2021;58:469580211035742 [FREE Full text] [doi: 10.1177/00469580211035742] [Medline: 34399597]
11. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]

XSL•FO
RenderX

12. Mokhtari H, Mirzaei A. The tsunami of misinformation on COVID-19 challenged the health information literacy of the general public and the readability of educational material: a commentary. Public Health 2020 Oct;187:109-110 [FREE Full text] [doi: 10.1016/j.puhe.2020.08.011] [Medline: 32942170]
13. Schmid P, Altay S, Scherer LD. The psychological impacts and message features of health misinformation - a systematic review of randomized controlled trials. Eur Psychol 2023 Jul;28(3):162-172. [doi: 10.1027/1016-9040/A000494]
14. Southwell BG, Otero Machuca J, Cherry ST, Burnside M, Barrett NJ. Health misinformation exposure and health disparities: observations and opportunities. Annu Rev Public Health 2023 Apr 03;44(1):113-130 [FREE Full text] [doi: 10.1146/annurev-publhealth-071321-031118] [Medline: 36207008]
15. Westberry C, Palmer XL, Potter L. Social media and health misinformation: a literature review. In: Arai K, editor. Proceedings of the Future Technologies Conference (FTC) 2023, Volume 3. FTC 2023. Lecture Notes in Networks and Systems, vol 815. Cham, Switzerland: Springer; 2023.
16. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
17. Borges do Nascimento IJ, Pizarro AB, Almeida J, Azzopardi-Muscat N, Gonçalves MA, Björklund M, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022 Sep 01;100(9):544-561 [FREE Full text] [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
18. Lee JJ, Kang K, Wang MP, Zhao SZ, Wong JYH, O'Connor S, et al. Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. J Med Internet Res 2020 Nov 13;22(11):e22205 [FREE Full text] [doi: 10.2196/22205] [Medline: 33048825]
19. Larson HJ, Broniatowski DA. Why debunking misinformation is not enough to change people's minds about vaccines. Am J Public Health 2021 Jun;111(6):1058-1060. [doi: 10.2105/AJPH.2021.306293] [Medline: 33950733]
20. Bapaye JA, Bapaye HA. Demographic factors influencing the impact of coronavirus-related misinformation on WhatsApp: cross-sectional questionnaire study. JMIR Public Health Surveill 2021 Jan 30;7(1):e19858 [FREE Full text] [doi: 10.2196/19858] [Medline: 33444152]
21. McKernon E. Fake news and the public. Harper's. 1925 Oct. URL: https://harpers.org/archive/1925/10/fake-news-and-the-public/ [accessed 2024-07-30]
22. Hotez P. The physician-scientist: defending vaccines and combating antiscience. J Clin Invest 2019 Apr 29;129(6):2169-2171 [FREE Full text] [doi: 10.1172/JCI129121] [Medline: 31033486]
23. Dyer C. Lancet retracts Wakefield's MMR paper. BMJ 2010 Mar 02;340:c696. [doi: 10.1136/bmj.c696] [Medline: 20124366]
24. Skafle I, Nordahl-Hansen A, Quintana DS, Wynn R, Gabarron E. Misinformation about COVID-19 vaccines on social media: rapid review. J Med Internet Res 2022 Aug 04;24(8):e37367 [FREE Full text] [doi: 10.2196/37367] [Medline: 35816685]
25. New Yahoo News/YouGov coronavirus poll: almost 1 in 5 say they won't get vaccinated. Yahoo News. URL: https://www.yahoo.com/news/new-yahoo-news-you-gov-coronavirus-poll-almost-one-in-five-say-they-wont-get-vaccinated-143852222.html [accessed 2024-08-06]
26. Peretti-Watel P, Verger P, Raude J, Constant A, Gautier A, Jestin C, et al. Dramatic change in public attitudes towards vaccination during the 2009 influenza A(H1N1) pandemic in France. Euro Surveill 2013 Oct 31;18(44):20623 [FREE Full text] [doi: 10.2807/1560-7917.es2013.18.44.20623] [Medline: 24176658]
27. Reuters. ‘They’re killing people’: Biden slams Facebook for Covid disinformation. The Guardian. 2021 Jul 21. URL: https://www.theguardian.com/media/2021/jul/17/theyre-killing-people-biden-slams-facebook-for-covid-misinformation [accessed 2024-07-30]
28. Britain has had enough of experts, says Gove. Financial Times. URL: https://www.ft.com/content/3be49734-29cb-11e6-83e4-abc22d5d108c [accessed 2024-07-30]
29. US news roundup: 19-25 August. Research Professional News. 2022 Aug 25. URL: https://www.researchprofessionalnews.com/rr-news-usa-2022-8-us-news-roundup-19-25-august/ [accessed 2024-07-30]
30. Gruzd A, Mai P, Soares FB. How coordinated link sharing behavior and partisans' narrative framing fan the spread of COVID-19 misinformation and conspiracy theories. Soc Netw Anal Min 2022;12(1):118 [FREE Full text] [doi: 10.1007/s13278-022-00948-y] [Medline: 36004041]
31. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 09;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
32. Wang Y. Systematic review on the social mechanism of health misinformation dissemination in the internet era. Eur J Public Health 2018 Nov;28(suppl_4):cky213.194. [doi: 10.1093/eurpub/cky213.194]
33. Framework for information literacy for higher education. American Library Association. 2016. URL: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/infolit/framework1.pdf [accessed 2024-07-30]
34. Quraishi Z. Addressing mental health, misinformation, and religious tensions among South Asian students across California higher education during the COVID-19 pandemic: A qualitative research study. Heliyon 2023 Jun;9(6):e16396 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e16396] [Medline: 37274702]

XSL•FO
RenderX

35. Zhou J, Xiang H, Xie B. Better safe than sorry: a study on older adults' credibility judgments and spreading of health misinformation. Univers Access Inf Soc 2022 Aug 04;22(3):1-10 [FREE Full text] [doi: 10.1007/s10209-022-00899-3] [Medline: 35966187]
36. Goldenberg MJ. Vaccine Hesitancy: Public Trust, Expertise, and the War on Science. Pittsburgh, PA: University of Pittsburgh Press; 2021.
37. Ittefaq M. "It frustrates me beyond words that I can't fix that": health misinformation correction on Facebook during COVID-19. Health Commun 2023 Nov 12;12:1-11. [doi: 10.1080/10410236.2023.2282279] [Medline: 37953464]
38. Mourali M, Drake C. The challenge of debunking health misinformation in dynamic social media conversations: online randomized study of public masking during COVID-19. J Med Internet Res 2022 Mar 02;24(3):e34831 [FREE Full text] [doi: 10.2196/34831] [Medline: 35156933]
39. Weinzierl MA, Harabagiu SM. Automatic detection of COVID-19 vaccine misinformation with graph link prediction. J Biomed Inform 2021 Dec;124:103955 [FREE Full text] [doi: 10.1016/j.jbi.2021.103955] [Medline: 34800722]
40. Petty RE, Cacioppo JT. Elaboration likelihood model of persuasion. In: Communication and Persuasion. Springer Series in Social Psychology. New York, NY: Springer; 1986.
41. Cowles K, Miller R, Suppok R. When seeing isn't believing: navigating visual health misinformation through library instruction. Med Ref Serv Q 2024;43(1):44-58. [doi: 10.1080/02763869.2024.2290963] [Medline: 38237023]
42. Nabożny A, Balcerzak B, Morzy M, Wierzbicki A, Savov P, Warpechowski K. Improving medical experts' efficiency of misinformation detection: an exploratory study. World Wide Web 2023;26(2):773-798 [FREE Full text] [doi: 10.1007/s11280-022-01084-5] [Medline: 35975112]
43. Barve Y, Saini JR. Detecting and classifying online health misinformation with 'Content Similarity Measure (CSM)' algorithm: an automated fact-checking-based approach. J Supercomput 2023;79(8):9127-9156 [FREE Full text] [doi: 10.1007/s11227-022-05032-y] [Medline: 36644509]
44. Broniatowski D, Gu J, Jamison AM, Simons JR. Facebook's architecture undermines vaccine misinformation removal efforts. arXiv Preprint posted online Feb 4, 2022 [FREE Full text] [doi: 10.48550/arXiv.2202.02172]
45. Russell N. Misinformation during COVID: how should nurse practitioners respond? J Nurse Pract 2021 Jun;17(6):763-764 [FREE Full text] [doi: 10.1016/j.nurpra.2021.03.013] [Medline: 36570561]
46. Wilner T, Holton A. Breast cancer prevention and treatment: misinformation on Pinterest, 2018. Am J Public Health 2020 Oct;110(S3):S300-S304. [doi: 10.2105/AJPH.2020.305812] [Medline: 33001732]
47. MacFarlane D, Hurlstone MJ, Ecker UK. Protecting consumers from fraudulent health claims: A taxonomy of psychological drivers, interventions, barriers, and treatments. Soc Sci Med 2020 Aug;259:112790. [doi: 10.1016/j.socscimed.2020.112790] [Medline: 32067757]
48. Thompson JD. Public health pedagogy and digital misinformation: health professional influencers and the politics of expertise. J Sociol 2022 Sep 28;59(3):646-663. [doi: 10.1177/14407833221128592]
49. Gunter J. Medical misinformation and the internet: a call to arms. Lancet 2019 Jun 08;393(10188):2294-2295. [doi: 10.1016/S0140-6736(19)31206-1] [Medline: 31180024]
50. Sui Y, Zhang B. Determinants of the perceived credibility of rebuttals concerning health misinformation. Int J Environ Res Public Health 2021 Mar 02;18(3):1345 [FREE Full text] [doi: 10.3390/ijerph18031345] [Medline: 33540869]
51. Collier R. Containing health myths in the age of viral misinformation. CMAJ 2018 May 14;190(19):E578 [FREE Full text] [doi: 10.1503/cmaj.180543] [Medline: 29759962]
52. Peterson JS, Swire-Thompson B, Johnson SB. What is the alternative? Responding strategically to cancer misinformation. Future Oncol 2020 Sep;16(25):1883-1888. [doi: 10.2217/fon-2020-0440] [Medline: 32564627]
53. Guidry JP, Miller CA, Ksinan AJ, Rohan JM, Winter MA, Carlyle KE, et al. COVID-19-related misinformation among parents of patients with pediatric cancer. Emerg Infect Dis 2021 Mar;27(2):650-652 [FREE Full text] [doi: 10.3201/eid2702.203285] [Medline: 33496232]
54. Ali S, Khalid A, Zahid E. Is COVID-19 immune to misinformation? A brief overview. Asian Bioeth Rev 2021 Jun;13(2):255-277 [FREE Full text] [doi: 10.1007/s41649-020-00155-x] [Medline: 33777228]
55. Jafari A, Mohammadpour M, Ghanbarzadegan A, Rossi-Fedele G, Bastani P. Oral health practitioners' knowledge, attitude, and awareness about coronavirus: A systematic review and meta-analysis. J Educ Health Promot 2021;10:39 [FREE Full text] [doi: 10.4103/jehp.jehp_939_20] [Medline: 33688548]
56. Sanyam SD, Sah SK, Chaudhary P, Burton MJ, Hoffman JJ. Knowledge and awareness-based survey of COVID-19 within the eye care profession in Nepal: Misinformation is hiding the truth. PLoS One 2021;16(7):e0254761 [FREE Full text] [doi: 10.1371/journal.pone.0254761] [Medline: 34288939]
57. Nxumalo CT, Mchunu GG. A qualitative study to explore primary health care practitioners' perceptions and understanding regarding the COVID-19 pandemic in KwaZulu-Natal, South Africa. Afr J Prim Health Care Fam Med 2021 Nov 26;13(1):e1-e11 [FREE Full text] [doi: 10.4102/phcfm.v13i1.3084] [Medline: 34879694]
58. Abbasi K. Where now in the danse macabre of COVID-19 and misinformation? BMJ 2023 Aug 17;382:1884. [doi: 10.1136/bmj.p1884]

XSL•FO
RenderX

59. Pesko MF, Cummings KM, Douglas CE, Foulds J, Miller T, Rigotti NA, et al. United States public health officials need to correct e-cigarette health misinformation. Addiction 2023 May;118(5):785-788. [doi: 10.1111/add.16097] [Medline: 36507802]

Abbreviations

CSM: content similarity measure
ELM: elaboration likelihood model
HIM: health information management
NHS: National Health Service

Edited by G Eysenbach, T Leung; submitted 15.04.22; peer-reviewed by G Nneji, SF Tsao; comments to author 07.06.22; revised version received 29.09.22; accepted 12.07.24; published 19.08.24.

Please cite as:
Kbaier D, Kane A, McJury M, Kenny I
Prevalence of Health Misinformation on Social Media—Challenges and Mitigation Before, During, and Beyond the COVID-19 Pandemic: Scoping Literature Review
J Med Internet Res 2024;26:e38786
URL: https://www.jmir.org/2024/1/e38786
doi:10.2196/38786
PMID:

©Dhouha Kbaier, Annemarie Kane, Mark McJury, Ian Kenny. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Immersive and Nonimmersive Virtual Reality–Assisted Active Training in Chronic Musculoskeletal Pain: Systematic Review and Meta-Analysis

Hermione Hin Man Lo[1,2], BN, MScN, MPH; Mengting Zhu[2], MSc; Zihui Zou[2], MPH; Cho Lee Wong[1], BN, MSc, PhD; Suzanne Hoi Shan Lo[1], BN, MSc, PhD; Vincent Chi-Ho Chung[2], B Chin Med, BSs Biomed Sci, MSc, PhD; Samuel Yeung-Shan Wong[2], MD, MPH, CCFP; Regina Wing Shan Sit[2], MD, MBBS, DCH, DPD, PDip Community Geriatrics, DipMED

[1]Nethersole School of Nursing, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Regina Wing Shan Sit, MD, MBBS, DCH, DPD, PDip Community Geriatrics, DipMED
Jockey Club School of Public Health and Primary Care
Faculty of Medicine
The Chinese University of Hong Kong
Room 209, Public Health Building, Prince of Wales Hospital
30-32 Ngan Shing Street, Shatin, New Territories
Hong Kong
China (Hong Kong)
Phone: 852 2252 8452
Email: reginasit@cuhk.edu.hk

Abstract

Background: Virtual reality (VR) in different immersive conditions has been increasingly used as a nonpharmacological method for managing chronic musculoskeletal pain.

Objective: We aimed to assess the effectiveness of VR-assisted active training versus conventional exercise or physiotherapy in chronic musculoskeletal pain and to analyze the effects of immersive versus nonimmersive VR on pain outcomes.

Methods: This systematic review of randomized control trials (RCTs) searched PubMed, Scopus, and Web of Science databases from inception to June 9, 2024. RCTs comparing adults with chronic musculoskeletal pain receiving VR-assisted training were included. The primary outcome was pain intensity; secondary outcomes included functional disability and kinesiophobia. Available data were pooled in a meta-analysis. Studies were graded using the Cochrane Risk-of-Bias Tool version 2.

Results: In total, 28 RCTs including 1114 participants with some concerns for a high risk of bias were identified, and 25 RCTs were included in the meta-analysis. In low back pain, short-term outcomes measured post intervention showed that nonimmersive VR is effective in reducing pain (standardized mean difference [SMD] –1.79, 95% CI –2.72 to –0.87; $P<.001$), improving disability (SMD –0.44, 95% CI –0.72 to –0.16; $P=.002$), and kinesiophobia (SMD –2.94, 95% CI –5.20 to –0.68; $P=.01$). Intermediate-term outcomes measured at 6 months also showed that nonimmersive VR is effective in reducing pain (SMD –8.15, 95% CI –15.29 to –1.01; $P=.03$), and kinesiophobia (SMD –4.28, 95% CI –8.12 to –0.44; $P=.03$) compared to conventional active training. For neck pain, immersive VR reduced pain intensity (SMD –0.55, 95% CI –1.02 to –0.08; $P=.02$) but not disability and kinesiophobia in the short term. No statistical significances were detected for knee pain or other pain regions at all time points. In addition, 2 (8%) studies had a high risk of bias.

Conclusions: Both nonimmersive and immersive VR–assisted active training is effective in reducing back and neck pain symptoms. Our study findings suggest that VR is effective in alleviating chronic musculoskeletal pain.

Trial Registration: PROSPERO CRD42022302912; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=302912

(J Med Internet Res 2024;26:e48787) doi:10.2196/48787

XSL•FO
RenderX

KEYWORDS

virtual reality; VR; physical therapy; musculoskeletal; pain; physiotherapy; chronic pain; musculoskeletal pain; low back pain; neck pain; osteoarthritis; knee pain; shoulder pain; disability; kinesiophobia; arthritis; systematic; review methods; review methodology; immersive; simulation; simulations

Introduction

Chronic musculoskeletal pain is a worldwide health problem with varying effects on physical, psychological, and social functioning [1,2]. According to the Global Burden of Disease Study of 2019, chronic musculoskeletal pain, especially low back pain, is the leading cause of disability worldwide, resulted in 149 million years of life lost [3]. The burden is expected to increase with an aging population and longevity [4].

Multimodal care is often needed in the management of chronic pain, with the aim of maintaining physical functioning and psychosocial well-being [5]. Exercise therapy is a well-known nonpharmacological modality in chronic pain management, with positive effects on pain intensity, physical function, sleep, and the quality of life [6]. Its additional benefits on happiness through the release of endorphins, serotonins, dopamine, and other "reward" chemicals have been demonstrated among individuals with chronic pain and depression [7,8]. Recent studies have suggested that technological advancements may increase the attractiveness of these active physical training programs, thus further improving compliance, adherence, and clinical outcomes [9,10].

Virtual reality (VR) is a digital technology that creates "a sense of presence in an computer-generated, three-dimensional, interactive environment with different immersive conditions" through head-mounted devices (HMDs), body-tracking sensors, and direct user input devices [11]. The use of VR promotes physical activity by increasing energy expenditure for fitness [12]. The capability of VR to reduce pain has mostly been attributed to active distraction (visual, auditory, and tactile input through interaction with a VR environment), which is understood to reduce resources available for the perception and elaboration of pain, thus diminishing subjective pain experience [13]. Gaming technology with motivational and affectively rewarding elements, as well as the goal-oriented interaction with a virtual environment, is suggested to have greater pain reduction [14].

An increasing number of trials have evaluated the role of VR-assisted active training in chronic musculoskeletal pain. Yet, within VR applications, an important distinction can be made between immersive and nonimmersive media, which differ in spatial presence [15]. With immersive technology, participants view the full panorama and are essentially inside the created environment. In a nonimmersive environment, virtual content is based on how the device (personal computer, smartphone, or tablet) is moved or rotated, and participants are only external observers. Whether immersive or nonimmersive VR is better for pain management remains unclear.

The aim of this systematic review was to assess the effectiveness of VR-assisted active training versus conventional active controls for musculoskeletal pain in different regions and to analyze the effects of immersive versus nonimmersive VR on validated pain outcomes.

Methods

Study Design

This systematic review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [16]. The protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO) registry (ID CRD42022302912).

Search Strategy

Two independent reviewers (authors HHML and ZZ) independently screened papers from PubMed, Scopus, and the Web of Science. A systematic search was conducted from inception to June 9, 2024. A backward reference search was used to increase the yield of eligible studies. Only English papers with full text available were included.

Our search strategy had 2 components for "chronic primary musculoskeletal pain" and "virtual reality." Keywords for the population were "cervical pain" OR "neck pain" OR "shoulder pain" OR "thoracic pain" OR "back pain" OR "low back pain" OR "joint pain" OR "arthralgia" OR "knee pain" OR "ankle pain" OR "limb pain" OR "osteoarthritis NOT structural" or "degenerative joint." Keywords for the intervention were "virtual reality" OR "augmented reality" OR "mixed reality." Please refer to Multimedia Appendix 1 for detailed search strategies.

Eligibility Criteria

All parallel or cross-over randomized controlled trials (RCTs) that evaluated the effectiveness of VR-assisted active training in chronic musculoskeletal pain were included. Our review included both 2-arm and multiarm trials [17]. Chronic musculoskeletal pain was defined as pain that lasts for more than 3 months persistently or intermittently, including regional pain (joints, limbs, back, or neck), a degenerative joint condition (eg, osteoarthritis), and musculoskeletal complaints that fall under the "chronic primary pain" classification of the *International Classification of Disease, 11th Revision* [18]. We included all VR interventions that create synchronized motion-based interactions with computer-generated objects and provide a sense of "presence" for users. Presence is defined as VR users' feeling of being immersed in a computer-generated environment via HMDs (immersive) or screens (nonimmersive), user input devices, body motion sensors, or commercial video game consoles [19,20]. To compare the effects of VR-assisted active training, we included studies with control groups using conventional exercise therapy or physiotherapy. Detailed inclusion and exclusion criteria are shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Randomized controlled trials (RCTs)
- Aged ≥18 years, with chronic musculoskeletal pain
- Virtual reality (VR)–assisted exercise therapy or physiotherapy
- Active training in comparison groups, including exercise therapy or conventional active physiotherapy

Exclusion criteria

- Cancer-related pain or autoimmune arthritis
- Psychotherapies
- Waitlist controls/daily life routine
- Passive physiotherapy in controls
- VR-assisted controls

Outcome Measures

The primary outcome was pain intensity. To be eligible, studies had to measure pain intensity using the Visual Analog Scale, the Numerical Rating Scale, the McGill Pain Questionnaire, the Chronic Pain Grade Scale, or other validated questionnaires [21,22]. Secondary outcomes were functional disability measured using disease-specific scales, such as the Roland-Morris Disability Questionnaire (RMDQ) [23] or the Oswestry Disability Index (ODI) [24] for chronic low back pain; the Neck Disability Index or the Neck Pain Disability Scale for chronic neck pain [25-27]; the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) for knee pain [28]; the Disabilities of the Arm, Shoulder, and Hand Questionnaire for shoulder pain [29]; and other validated questionnaires. Psychological status was measured through kinesiophobia, an emotional and cognitive factor that leads to maladaptive behaviors [30,31]; studies have shown that kinesiophobia is associated with more pain and disability and a lower quality of life [30]. Kinesiophobia was assessed using the Fear-Avoidance Beliefs Questionnaire [32] and the TAMPA Scale of Kinesiophobia [33,34].

Study Selection and Data Extraction

Studies retrieved from the databases were uploaded to Covidence online systematic review software (Veritas Health Innovation). Two independent reviewers (HHML and ZZ) performed title and abstract screening of the retrieved literature for potentially eligible trials. The full text of selected papers was then retrieved and screened against the inclusion and exclusion criteria. Both reviewers (HHML and ZZ) independently judged the eligibility of the full text retrieved, and disagreements between the 2 reviewers were resolved by a third reviewer (author RWSS).

Data were then extracted from the selected papers by the first reviewer (HHML) and cross-checked by the second reviewer (ZZ). The extraction table headings included the first author, the year of publication, chronic pain subtypes, the sample size analyzed, the intervention group, the control group, the dosage of interventions, the mean age, outcomes (in terms of the mean difference [MD]), and assessment time points in weeks. Two independent reviewers (HHML and ZZ) then extracted MDs and SDs for the following domains: (1) pain intensity, (2) disability, and (3) kinesiophobia.

Risk-of-Bias Assessment

Selected studies were assessed according to the intention to treat using the Cochrane Risk-of-Bias Tool version 2 (RoB 2). RoB 2 covers all types of biases that affect the results of RCTs: (1) bias due to randomization, (2) bias caused by deviations from intended interventions, (3) missing outcome data that results in bias, (4) bias in outcome measurement, and (5) the selection bias of reported results. Two reviewers (HHML and MZ) assessed the risk of bias independently for the selected papers, while a third reviewer (RWSS) acted as the arbitrator. A conclusion for risk-of-bias judgment was made by consensus.

Publication bias was assessed for the meta-analysis with 10 or more RCTs and was determined by constructing a funnel plot with the standard error against the effect size [35].

Quality of Evidence

The Grading of Recommendation Assessment, Development, and Evaluation (GRADE) approach was used to assess the quality of evidence with GRADEpro software [36]. The quality of evidence for pain intensity, disability, and kinesiophobia at different time points was assessed separately. Evidence was downgraded if (1) the risk of bias was apparent (1 study showed a high risk, and 24 studies showed some concerns); (2) inconsistency was demonstrated, with I^2>50%; (3) there was indirectness in participants or comparators (ie, whether participants or comparators aligned and compared with our research question, eg, participants had chronic musculoskeletal pain and comparators used conventional active training/physical therapy); (4) evidence of imprecision (when the effect size was large, ie standardized mean difference [SMD]>0.8 or MD>minimally clinically important difference [MCID; 95% CI], but the total sample size was small) [37]; or (5) there was publication bias (funnel plots were visually inspected when at least 10 trials were included in the meta-analysis). An overall GRADE rating was agreed upon (HHML and MZ) for each included study at 4 levels: very low, low, moderate, and high.

XSL•FO
RenderX

Statistical Analysis

All meta-analyses were conducted using Review Manager (RevMan version 5.4) software (Cochrane). Pairwise meta-analysis was performed using a random effects model according to nonimmersive or immersive VR interventions [38]. Pain intensity, disability, and kinesiophobia were analyzed according to pain regions. Regarding the assessment time points, analysis was conducted in the immediate postintervention period (short term) and at 6 months (intermediate term). In our study, we defined short-term pain outcomes at 12 weeks and intermediate-term pain outcomes at 6 months. These time frames were established based on our clinical experience and the expected response timeline for the pain interventions used. We chose a 12-week period for assessing short-term pain outcomes as it aligns with the typical clinical trajectory observed in patients receiving pain interventions. During this period, patients are likely to experience the initial benefits of the treatment, and early therapeutic effects are most evident. The 24-week mark was selected to represent intermediate-term pain outcomes based on the continued progression of therapeutic effects and sustained pain relief observed in clinical practice. SMDs were used to measure continuous outcomes with more than 1 measuring scale. The SMD was clinically interpreted as Cohen d (SMD 0.2 was considered a small effect, 0.5 was considered a moderate effect and clinically important, and 0.8 was considered a large effect) [39,40]. Weighted mean differences (MDs) with 95% CIs were used to measure continuous outcomes within a domain on similar scales, and relevant minimal clinical important differences were used to assess clinical significance. Heterogeneity between studies was further explored by referring to I^2 statistics for all outcomes: I^2 <25%, low heterogeneity; I^2=25%-50%, moderate heterogeneity; and I^2>50%, high heterogeneity [41].

Results

Eligible Studies

A total of 2753 papers were screened after removing duplicates. After title and abstract screening, 30 (1.1%) papers were selected for full-text retrieval and screening. The remaining reports and 6 papers from the backward reference search were screened by 2 independent reviewers (HHML and MZ), and 8 (22.2%) papers were excluded for the following reasons: not RCTs (n=3, 37.5%), passive physiotherapy in controls (n=1, 12.5%), VR-assisted controls (n=1, 12.5%), waitlist controls (n=1, 12.5%), and duplicate studies (n=2, 25%). A total of 28 (93.3%) of 30 studies were included in the final qualitative synthesis and 25 (83.3%) of the 30 studies were included in the meta-analysis synthesis (Figure 1) [42-65]. Furthermore, 3 (10.7%) studies could not be pooled: 1 (33.3%) single-out nonimmersive VR study on neck pain [66], 1 (33.3%) RCT on different joint pains [60], and 1 (33.3%) study on shoulder pain [59].

XSL•FO
RenderX

Figure 1. PRISMA flowchart for study selection. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT: randomized control trial; VR: virtual reality.

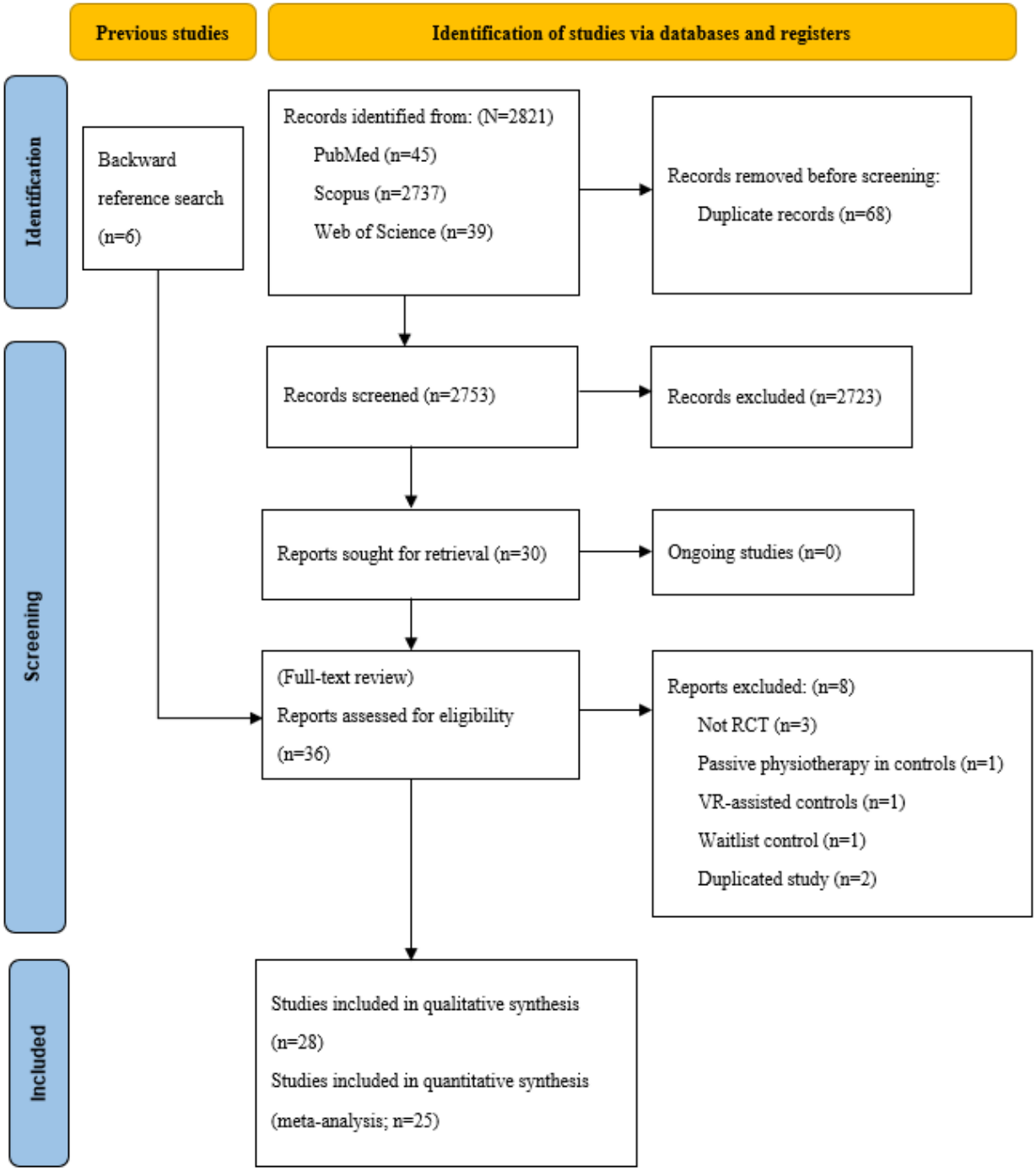

Characteristics of Included Trials

Table S1 in Multimedia Appendix 2 summarizes all the 28 RCTs [42-69] selected. The mean age of the study sample was 42.50 (SD 3.49) years in the VR intervention group and 42.93 (SD 3.73) years in the control group. The sample sizes ranged from 19 (1.7%) to 84 (7.3%), with a total sample size of 1144. The average proportion of female participants was 49.3% (n=564) in all trials, excluding 3 (10.7%) trials [42,58,67] that did not provide participants' sex ratio. In addition, 18 (64.3%) studies [42-47,54-63,66-68] used nonimmersive VR, while 8 studies [48-53,64,65,69] used immersive VR. Each VR session lasted from 10 to 45 minutes (mean 26.54, SD 9.05), except 4 (28.6%) studies [43,48,49,65] that did not report the VR intervention time. The mean total number of sessions was 15.46 (SD 7.18), and the frequency ranged from a single session to 5 sessions/week. Control groups included conventional physical therapy, such as balance and core stabilization training, walking, and audio-guided, isokinetic dynamometer–assisted, proprioceptive, kinematic, and sensorimotor exercises. Furthermore, 4 (14.3%) studies [53,60,64,69] reported no adverse events during the trials. Minor simulator sickness symptoms (eg, dizziness, nausea, headache) were reported in 2 (7.1%) studies [50,51].

XSL•FO
RenderX

Risk-of-Bias Assessment

All included studies were assessed with RoB 2 (Figure 2). Overall, 26 (92.9%) studies [42-44,46-63,66-70] were rated as having "some concerns" of bias, and 2 (7.1%) studies [45,65] were rated as having a "high risk" of bias. In the domain of "bias arising from the randomization process," 15 (53.6%) studies [42-44,46-49,51,57,61,63,65,67-69] had low bias and 13 (46.4%) studies [45,50,52-56,58-60,62,64,66] had some bias. In the domain of "deviations from the intended interventions," all 28 (100%) studies had some bias due to failure to blind participants or delivery of interventions. In the domain of "missing outcome data," 24 (85.7%) studies [42-44,46-51,53-55,57-61,63-66,68,69] had low risk of bias, 3 (10.7%) studies [52,56,62] had some concerns regarding bias, and 1 (3.6%) study [45] had high risk of bias. In the domain of "bias arising from measurement of the outcome," 12 (42.9%) studies [44,46-51,53-55,57-61,63,64,66,67] had low bias and 16 (57.1%) studies [42,43,45,52,53,55,56,62,65,68,69] had some bias. The absence of blinding outcome assessors or data analysts was observed in studies with some bias. In the domain of "selection of reported results," 7 (25%) studies [45-47,51,52,60,64,66] had low risk of bias, 20 (71.4%) studies [42-44,47,49,50,53-59,61-63,67-69] had some concerns regarding bias, and 1 (3.6%) study [65] had high risk of bias. This domain was mainly affected by the absence of prespecified analysis plans in preliminary and pilot studies.

XSL•FO
RenderX

Figure 2. Risk-of-bias assessment of the included studies.

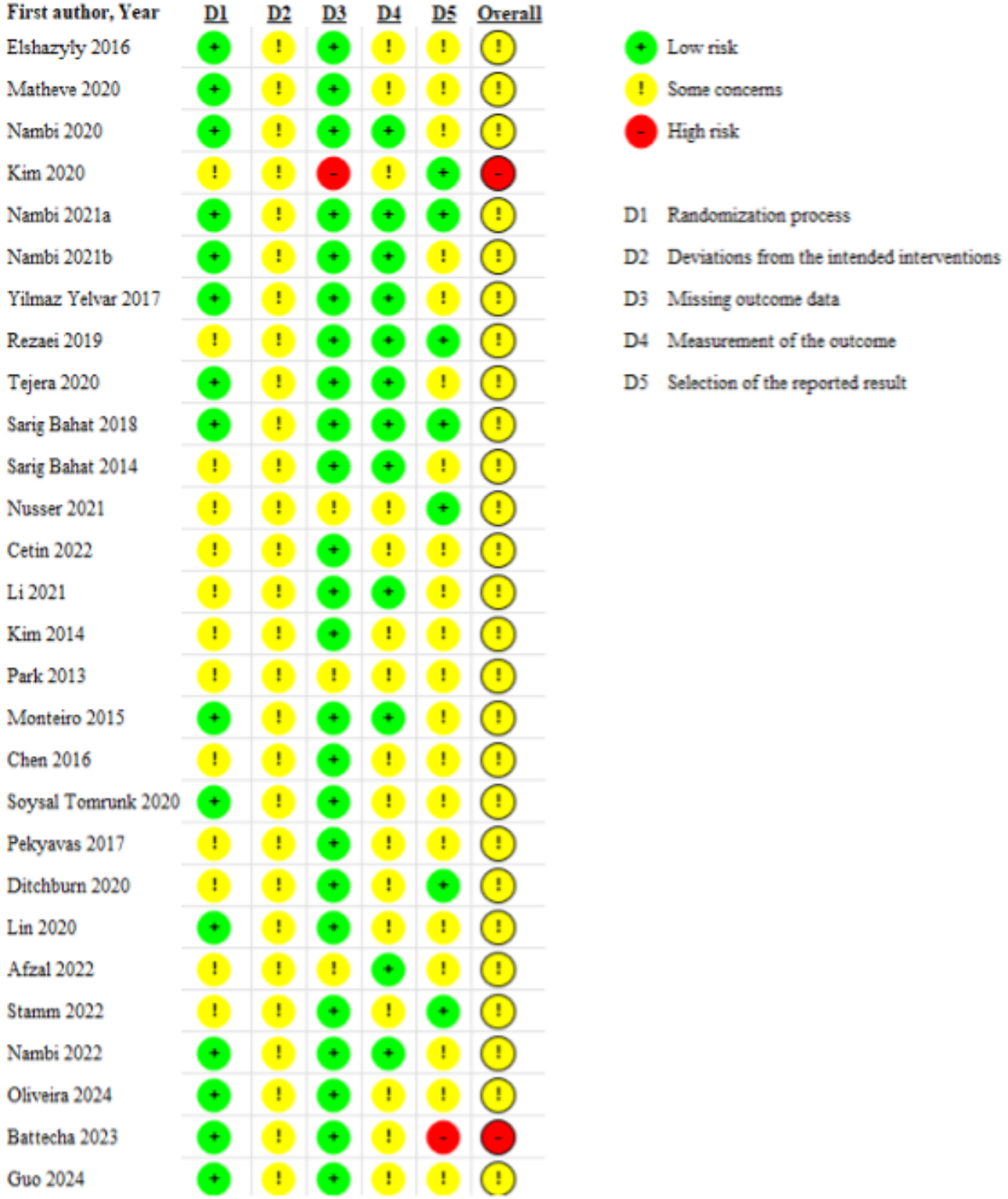

Pain Region: Back

Pain Intensity

In the short term, 15 (53.6%; n=563, 49.2%, participants) of the 28 RCTs [43-48,54-58,62-64,67] were eligible for pooling. Results from 13 (46.4%) RCTs favored the use of nonimmersive VR in reducing pain intensity (SMD –1.79, 95% CI –2.72 to –0.87; P<.001), with high heterogeneity (I^2= 94%) compared to active training. Results of immersive VR from 2 (7.1%) RCTs were statistically insignificant (SMD 0.04, 95% CI –1.10 to 1.19; P=.94), as shown in Figure 3a. Visual inspection of funnel plots (Figure 3b) indicated publication bias in our meta-analysis.

XSL•FO
RenderX

Figure 3. (a) Short-term effects of VR-assisted vs conventional active training in back pain intensity (nonimmersive VR via flat screen; immersive VR via HMDs). (b) Funnel plots for short-term effects of VR-assisted vs conventional active training in back pain intensity. (c) Intermediate effects of nonimmersive VR–assisted vs conventional active training in back pain intensity. HMD: head-mounted device; VR: virtual reality.

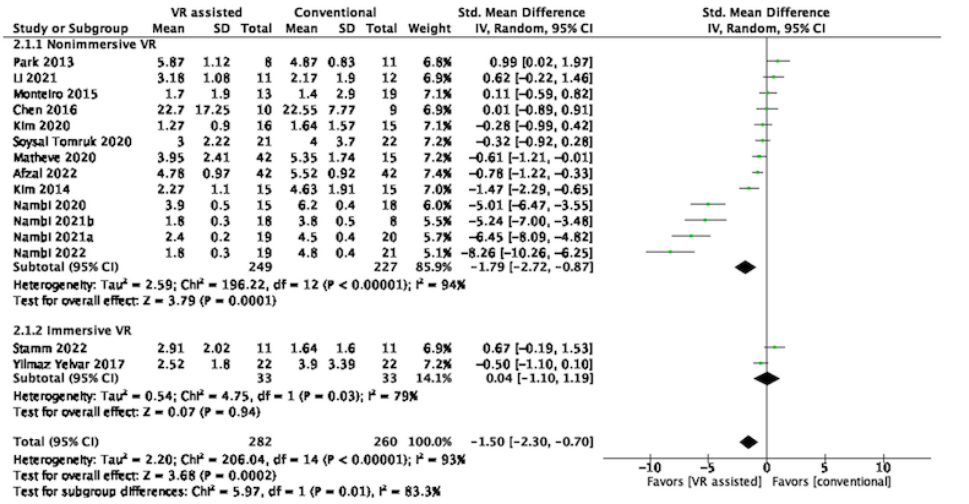

(a)

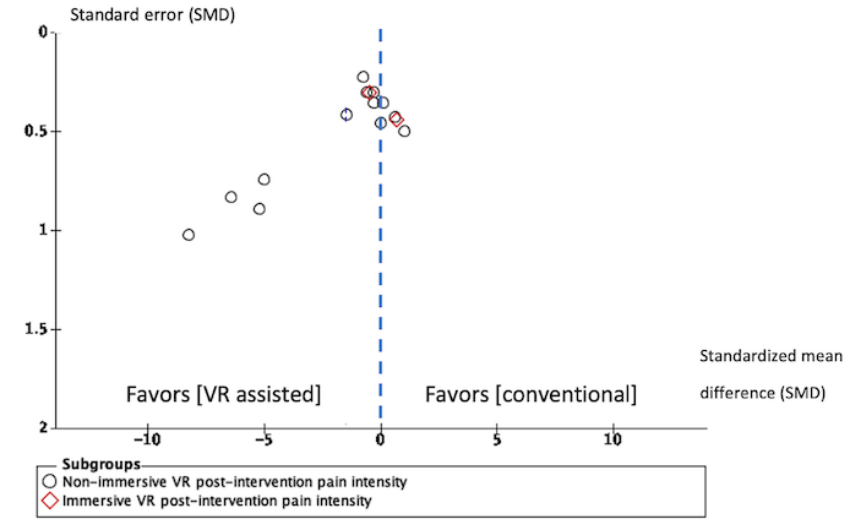

(b)

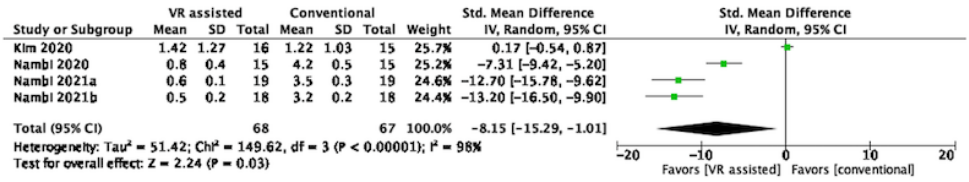

(c)

In the intermediate term, 4 (14.3%; n=135, 11.8%, participants) RCTs [44-47] were eligible for pooling; again, results favored nonimmersive VR in reducing pain intensity (SMD –8.15, 95% CI –15.29 to –1.01; *P*=.03), with high heterogeneity (I^2= 98%), as shown in Figure 3c.

Functional Disability

Of the 28 RCTs, 6 (21.4%; n=229, 20%, participants) [45,48,54,55,58,67] were eligible for pooling. The ODI was extracted over the RMDQ in the pooling because of its higher reliability and relatively lower measurement error [71]. Results favored the use of nonimmersive VR over conventional active training in improving back disability in the short term (SMD –0.44, 95% CI –0.72 to –0.16; *P*=.002) and of low heterogeneity (I^2=6%), as shown in Figure 4.

XSL•FO
RenderX

Figure 4. Short-term effects of nonimmersive VR–assisted vs conventional active training in back functional disability. VR: virtual reality.

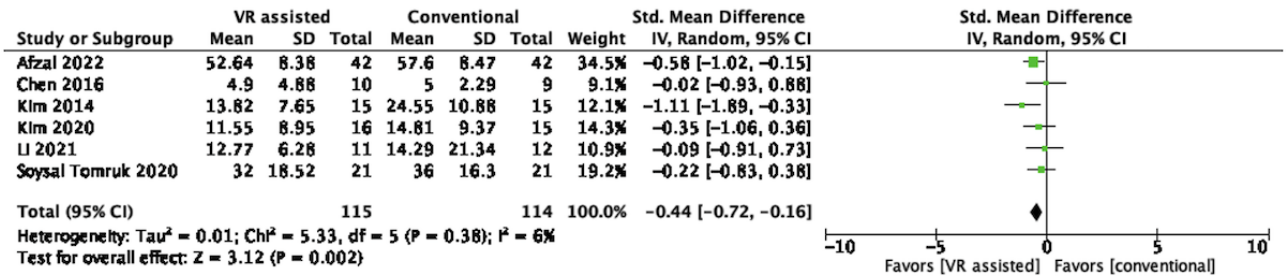

Kinesiophobia

In the short term, pooled results of 5 (17.9%; n=135, 11.8%, participants) of the 28 RCTs [45-47,55] favored nonimmersive VR over conventional active training (SMD –2.94, 95% CI –5.20 to –0.68; *P*=.01), with high heterogeneity (I^2=95%). Pooled results of immersive VR from 2 (7.1%; n=66, 5.8%, participants) RCTs were statistically insignificant (SMD –1.17, 95% CI –2.59 to 0.26; *P*=.11), with high heterogeneity (I^2=85%), as shown in Figure 5a.

Figure 5. (a) Short-term effects of VR-assisted vs conventional active training in back kinesiophobia (nonimmersive VR via flat screen; immersive VR via HMDs). (b) Intermediate effects of nonimmersive VR–assisted vs conventional active training in back kinesiophobia. HMD: head-mounted device; VR: virtual reality.

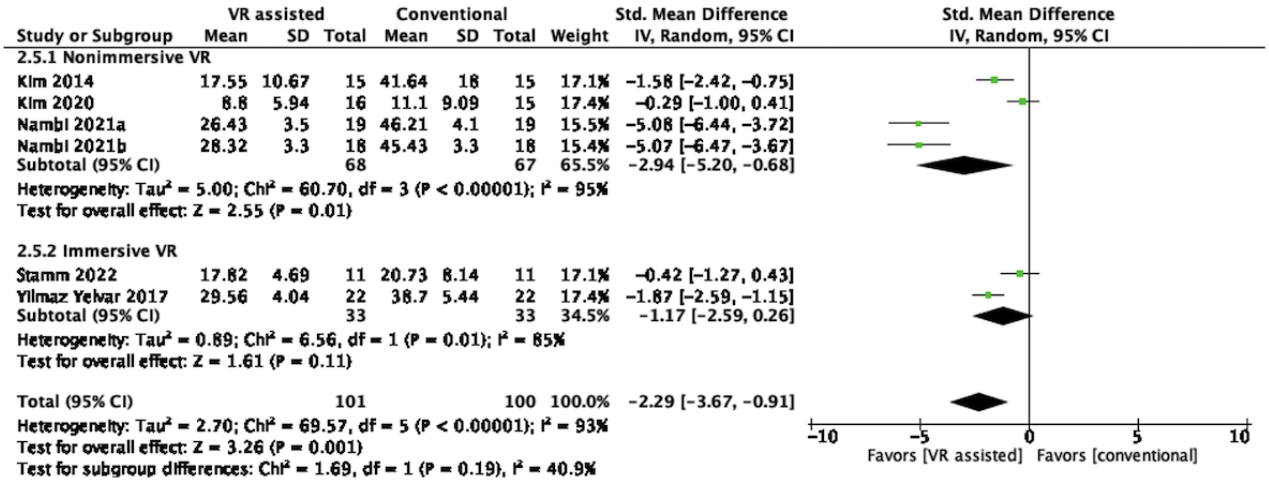

(a)

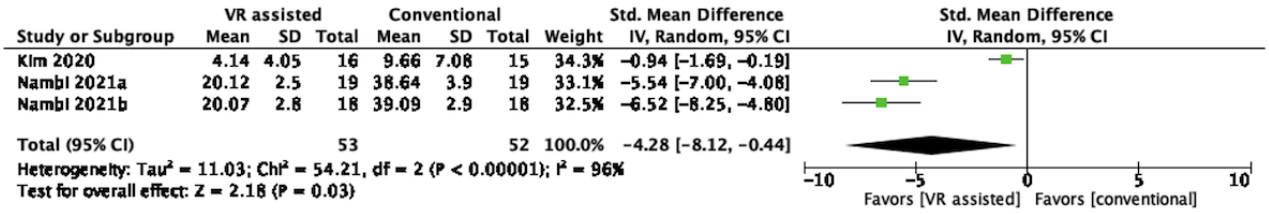

(b)

In the intermediate term, 3 (10.7%; n=105, 9.2%, participants) RCTs [45-47] were pooled. Pooled results favored nonimmersive VR in kinesiophobia (SMD –4.28, 95% CI –8.12 to –0.44; *P*=.03), with high heterogeneity (I^2=96%), as shown in Figure 5b.

Pain Region: Neck

Pain Intensity

Of the 28 RCTs, 7 (25%; n=316, 6.9%, participants) [49-53,65,69] were eligible for pooling for immersive VR. Pooled results favored immersive VR over conventional active training in reducing pain intensity in the short term (SMD –0.55, 95% CI –1.02 to –0.08; *P*=.02), with high heterogeneity (I^2=75%), as shown in Figure 6.

XSL•FO
RenderX

Figure 6. Short-term effects of immersive VR–assisted vs conventional active training in neck pain intensity. VR: virtual reality.

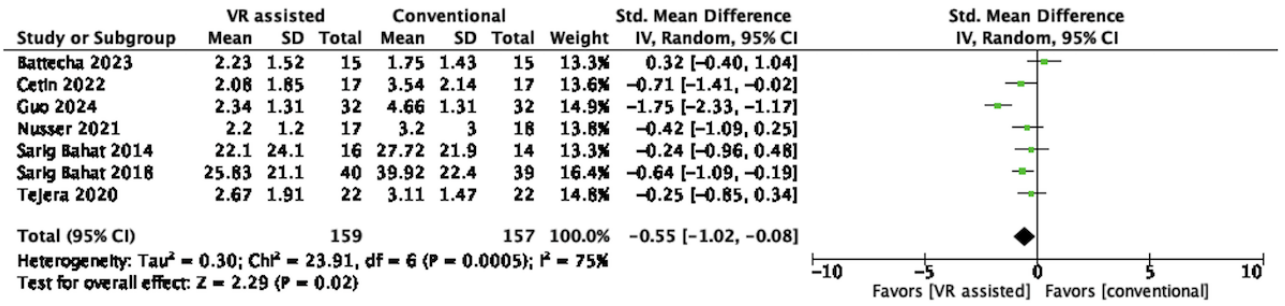

Functional Disability

Of the 28 RCTs, 6 (21.4%; n=282, 24.7%, participants) [49-52,65,69] were eligible for pooling for immersive VR. Pooled results favored immersive VR over conventional active training in reducing neck disability in the short term (MD –2.59, 95% CI –3.51 to –1.67; *P*<.001) and no heterogeneity (I^2=0%), as shown in Figure 7.

Figure 7. Short-term effects of immersive VR–assisted vs conventional active training in neck functional disability. VR: virtual reality.

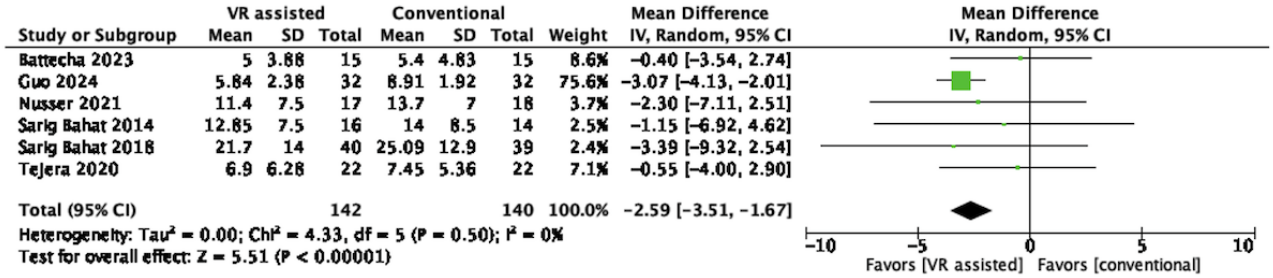

Kinesiophobia

Of the 28 RCTs, 3 (10.7%; n=153, 13.4%, participants) [49-51] were eligible for short-term pooling. Pooled results showed no significant effect of immersive VR no kinesiophobia (SMD –0.09, 95% CI –0.40 to 0.23; *P*=.59) and no heterogeneity (I^2=0%), as shown in Figure 8.

Figure 8. Short-term effects of immersive VR–assisted vs conventional active training in neck kinesiophobia. VR: virtual reality.

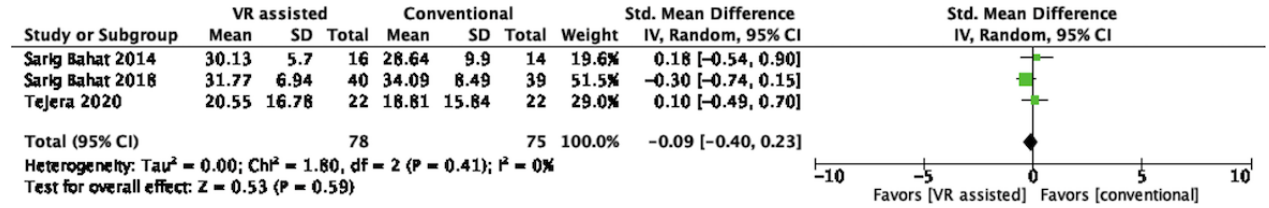

Pain Region: Knee

Pain Intensity

Of the 28 RCTs, 3 (10.7%; n=160, 14%, participants) [42,61,68] on nonimmersive VR were pooled in the short term (SMD –0.74, 95% CI –1.86 to 0.37; *P*=.19), with high heterogeneity (I^2=90%), as shown in Figure 9.

Figure 9. Short-term effects of nonimmersive VR–assisted vs conventional active training in knee pain intensity. VR: virtual reality.

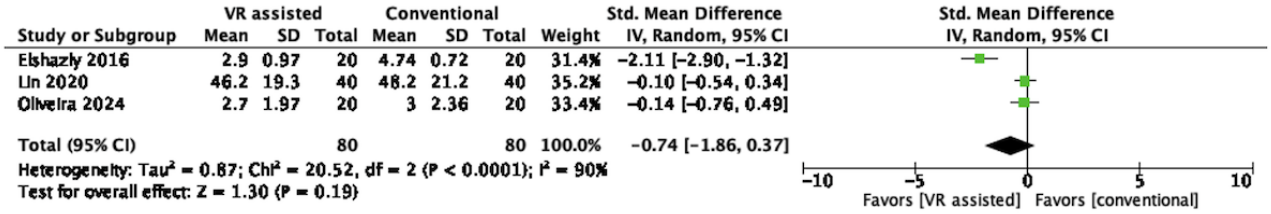

XSL•FO
RenderX

Functional Disability

Of the 28 RCTs, 3 (10.7%; n=160, 14%, participants) [42,61,68] were pooled in the short term. Pooled results (MD –11.36, 95% CI –33.95 to 11.23; *P*=.32) suggested no statistical and clinical significance for nonimmersive VR versus controls, with high heterogeneity (I^2=98%), as shown in Figure 10.

However, due to the small number of studies, pooling was not possible for chronic shoulder, hip, and other joint pain subgroups.

Figure 10. Short-term effects of nonimmersive VR–assisted vs conventional active training in knee functional disability. VR: virtual reality.

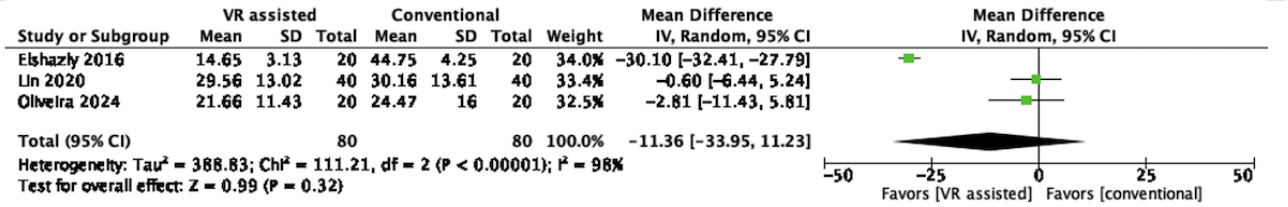

Quality of Evidence With the GRADE Approach

The overall quality of evidence ranged from very low to moderate in both nonimmersive VR (Table 1) and immersive VR (Table 2). In low back pain, the assessment showed very low certainty for both nonimmersive and immersive VR–assisted physical therapy in reducing pain intensity and kinesiophobia in the short and the intermediate term and low certainty for nonimmersive VR in reducing functional disability. For neck pain, we found low-to-moderate certainty for immersive VR–assisted physical therapy in reducing pain, functional disability, and kinesiophobia. For knee pain, very low certainty for nonimmersive VR–assisted physical therapy had no statistical significance for improving pain and functional disability in the short term.

XSL•FO
RenderX

Table 1. Summary of findings: nonimmersive VR[a]–assisted training compared with conventional active training.

Pain region and outcomes	Time point[b]	Illustrative comparative risks	SMD[c]/MD[d] (95% CI), *P* value	Participants (N=1144), n (%)	Studies (N=28), n (%)	Certainty of evidence[e] (GRADE[f])
Back						
Pain intensity	Short term	The nonimmersive VR group showed more improvement in pain intensity compared to the control group.	–1.79 (–2.72 to –0.87), *P*<.01	476 (41.6)	13 (46.4)	Very low[g]
Pain intensity	Intermediate term	The nonimmersive VR group showed more improvement in pain intensity compared to the control group.	–8.15 (–15.29 to –1.01), *P*<.01	135 (11.8)	4 (14.3)	Very low[g]
Functional disability	Short term	The nonimmersive VR group showed more improvement in disability compared to the control group.	–0.44 (–0.72 to –0.16), *P*=.002	229 (20.0)	6 (21.4)	Low[h]
Kinesiophobia	Short term	The nonimmersive VR group showed more improvement in kinesiophobia compared to the control group.	–2.94 (–5.20 to –0.68), *P*=.01	135 (11.8)	4 (14.3)	Very low[g]
Kinesiophobia	Intermediate term	The nonimmersive VR group showed more improvement in kinesiophobia compared to the control group.	–4.28 (–8.12 to –0.44), *P*=.03	105 (9.2)	3 (10.7)	Very low[g]
Knee						
Pain intensity	Short term	There was no statistically significant difference between groups.	–0.74 (–1.86 to 0.37), *P*=.19	160 (14.0)	3 (10.7)	Very low[g]
Functional disability	Short term	There was no statistically significant difference between groups.	–11.36 (–33.95 to 11.23), *P*=.32	160 (14.0)	3 (10.7)	Very low[g]

[a]VR: virtual reality.

[b]Time points: “short term” defined as postintervention; “intermediate term” defined as 6 months.

[c]SMD: standardized mean difference.

[d]MD: mean difference.

[e]Certainty of evidence: high, further research is very unlikely to change our confidence in the estimate of effects; moderate, further research is likely to have an important impact on our confidence in the estimate of effects and may change the estimate; low, further research is very likely to have an important impact on our confidence in the estimate of effects and is likely to change the estimate; very low, any estimate of effect is very uncertain.

[f]GRADE: Grading of Recommendations, Assessment, Development, and Evaluations.

[g]Downgraded by 3 levels as the risk of bias was unclear or high in most included studies (–1), there was inconsistency in results (I^2>50%; –1), there was imprecision due to a large effect size (SMD>0.8 or MD>MCID, 95% CI), but the total sample size was small (–1).

[h]Downgraded by 2 levels as the risk of bias was unclear or high in most included studies (–1), and there was indirectness in comparators (–1).

XSL•FO
RenderX

Table 2. Summary of findings: immersive VR[a]–assisted training compared with conventional active training.

Pain region and outcomes in the short term[b]	Results of meta-analysis	SMD[c]/MD[d] (95% CI), *P* value	Participants (N=1144), n (%)	Studies (N=25), n (%)	Certainty of evidence[e] (GRADE[f])
Back					
Pain intensity	There was no statistically significant difference between groups.	0.04 (–1.10 to 1.19), *P*=.94	66 (5.8)	2 (8.0)	Very low[g]
Kinesiophobia	There was no statistically significant difference between groups.	–1.17 (–2.59 to 0.26), *P*=.11	66 (5.8)	2 (8.0)	Very low[g]
Neck					
Pain intensity	Immersive VR reduced pain more compared to the control group.	–0.55 (–1.02 to –0.08), *P*=.02	316 (27.6)	7 (28.0)	Low[h]
Functional disability	There was no statistically significant difference between groups.	–2.59 (–3.51 to –1.67), *P*<.001	282 (24.7)	6 (24.0)	Low[i]
Kinesiophobia	There was no statistically significant difference between groups.	–0.09 (–0.40 to 0.23), *P*=.59	153 (13.4)	3 (12.0)	Moderate[j]

[a]VR: virtual reality.

[b]"Short term" was defined as postintervention.

[c]SMD: standardized mean difference.

[c]SMD: standardized mean difference.

[d]MD: mean difference.

[e]Certainty of evidence: high, further research is very unlikely to change our confidence in the estimate of effects; moderate, further research is likely to have an important impact on our confidence in the estimate of effects and may change the estimate; low, further research is very likely to have an important impact on our confidence in the estimate of effects and is likely to change the estimate; very low, any estimate of effect is very uncertain.

[f]GRADE: Grading of Recommendations, Assessment, Development, and Evaluations.

[g]Downgraded by 3 levels as the risk of bias was unclear or high in most included studies (–1), there was inconsistency in results (I^2>50%; –1), and there was indirectness in comparators (–1).

[h]Downgraded by 2 levels as the risk of bias was unclear in most included studies (–1) and there was indirectness in comparators (–1).

[i]Downgraded by 2 levels as the risk of bias was unclear in most included studies (–1), there was imprecision due to a large effect size (SMD>0.8 or MD>MCID, 95% CI), but the total sample size was small (–1).

[j]Downgraded by 1 level as the risk of bias was unclear or high in most included studies (–1).

Discussion

Principal Findings

For back pain, very-low-to-low-certainty evidence suggests that nonimmersive VR–assisted training is superior to conventional training in reducing pain, improving disability, and improving kinesiophobia in the short term, and the superior effects on pain and kinesiophobia are sustained in the intermediate term. The effect sizes detected in this study were large for pain intensity in the short and the intermediate term (ie, –1.50 and –8.15, respectively). The effect size for disability of –0.44 was moderate. The effect sizes were also large for kinesiophobia, at –2.29 in the short term and –4.28 in the intermediate term. For neck pain, low-to-moderate-certainty evidence suggests that immersive VR is effective in reducing pain and disability in the short term; the effect size of –0.55 was moderate. The mean difference in disability was –2.59, which was lower than the MCID for a neck disability change of –7.5 [72]. However, no statistically significant effects were detected on kinesiophobia in both short and intermediate terms. For knee pain, only nonimmersive VR was available, and we did not detect any statistically significant difference between nonimmersive VR and control groups in knee pain and function. There are only a few studies on other pain regions, such as the shoulder, hip, and mixed musculoskeletal regions, so pooling was not possible, and the evidence was inconclusive. Our findings suggest that VR-assisted training is superior to conventional active training in managing chronic musculoskeletal pain.

When investigating the effect size of the included studies, especially in nonimmersive VR pooling for chronic back pain, we found that 5 main studies [44,46,47,63,67], which were supervised by physiotherapists in both intervention and control arms, showed large short- and intermediate-term pain reduction effects. Supervision in exercise therapy enhances treatment adherence, achieving the high dosage necessary to demonstrate the VR-assisted effect [73]. Conversely, 5 studies [50,52,53,65,69] on immersive VR–assisted training for neck pain incorporated conservative treatments, such as strengthening and kinematic exercises, to augment their therapeutic effects. These differences in intervention and control designs may have introduced heterogeneity across trials, alongside variations in the participants' mean age, the athletes' training background, the dosage of interventions, and the diversity of different VR hardware and software. For example, the hardware included horse simulators, the ProKin system, the Nintendo Wii system, Kinect Xbox 360, the Biodex Balance system, and high-definition television equipped with motion sensors. Although heterogeneity was high, it was inevitable in VR trials due to the unique features of innovations in digital technology.

XSL•FO
RenderX

Therefore, we suggested that the high heterogeneity might affect the generalizability of results but should not demerit the clinical effects of VR in reducing back and neck pain.

Comparing the clinical effectiveness of immersive and nonimmersive VR is challenging because most studies on chronic low back pain have used nonimmersive VR, while those on chronic neck pain have used immersive VR. This discrepancy may be due to the use of HMDs in detecting cervical kinematics and range of motion during active training for neck pain, which is not required in studies on back pain. Our systematic review revealed that most studies on back pain have used software comprising ready-made recreational VR games or virtual simulated environments. Meanwhile, immersive therapeutic software aimed at creating presence, learning, and habit building has recently emerged for treating chronic low back pain [74]. This development extends the usefulness of immersive VR–assisted interventions by improving pain interference with activity, mood, and stress, which are commonly found in patients with chronic musculoskeletal pain.

Comparison With Other Reviews

Ahern et al [19] conducted a review of VR in patients with neck and back pain, with only 2 trials eligible for quantitative synthesis; they nevertheless reached the same conclusion that VR is effective in reducing back and neck pain intensity. Brea-Gómez et al [75] conducted a comprehensive review of studies on chronic back pain; 14 studies were included in the systematic review and 11 in the meta-analysis. Similar to our results, significant differences were found in favor of VR compared to control interventions in pain intensity and kinesiophobia in the short term, with effect sizes of –1.92 and –8.96, respectively; although they showed a trend favoring VR in reducing disability, only 2 trials were included in pooling, and the results were not statistically significant [75]. We included 6 trials in our pooling, allowing us to detect a more accurate effect size, with smaller CIs and statistical significance. Bordeleau et al [76] also reviewed the use of VR in back pain; 16 trials were included in the meta-analysis, and similar results were found, with VR statistically significantly improving back pain intensity over control interventions. Yet, the authors did not evaluate the role of VR in back pain disability and did not analyze the effects based on the level of immersiveness [76]. Li et al [77] found similar immediate VR effects on back pain but not at 3-6 months, possibly due to high heterogeneity and inconsistency from pooling waitlist controls, violating the assumption of a common effect size [78]. Our findings on reducing pain intensity and disability were also similar to those of the systematic reviews conducted by Guo et al [79] and Brea-Gomez et al [80]. Byra and Czernicki [81] reviewed the effectiveness of VR rehabilitation in knee and hip osteoarthritis with or without arthroplasty; meta-analysis was not performed due to heterogeneous study populations and outcome measurements [81]. Although we found a trend favoring the use of VR in knee pain, the small number of studies limited its power to detect statistical significance [81]. Kantha et al [82] supported the use of VR over conventional physical therapy in reducing pain but not in improving disability; their meta-analysis of 5 included studies also favored the use of nonimmersive VR, which was similar to our results [82]. Yet, their results were drawn from pooling of mixed pain regions.

Strengthens and Limitations

The strengths of this study include a comprehensive review of VR in different musculoskeletal pain regions, not only in the short term but also in the intermediate term. This is the first study to evaluate the degree of immersiveness in VR-assisted active training on validated pain outcomes. We used a rigorous methodology that conformed to best-practice guidelines [16].

There were several limitations. Although we increased the number of VR studies, the total participant sample size was still small, and quantitative syntheses included a small number of studies in most comparisons. For the same reason, we were unable to generate funnel plots to assess publication bias for most outcomes [35]. Finally, since chronic pain is a biopsychosocial condition, the inclusion of only VR-assisted active training and the exclusion of VR-assisted psychotherapy might potentially underestimate the true effect of VR on chronic pain [83].

Future Research and Clinical Implications

Future research needs to focus on the long-term effects of VR-assisted active training on chronic pain management. The joy and pleasure associated with VR interactions are attractive, but the excitement will fade with time. Therefore, it is essential to evaluate participants' adherence to VR interventions [9,84]. Furthermore, trials should be conducted to evaluate other pain, especially knee pain, given that it is the most prevalent condition in the aging population [85]. The mechanism of action of VR in pain regulation will need to be elucidated for the best design of VR apps. Finally, cost-effectiveness should be evaluated to inform resource allocation of VR in clinical practice.

Conclusion

In summary, our study found that nonimmersive VR–assisted active training is superior to conventional active training in reducing pain intensity, functional disability, and kinesiophobia in low back pain in the short and the intermediate term. Immersive VR–assisted active training is effective in reducing the intensity of neck pain. Evidence on knee pain, shoulder pain, and hip pain remains inconclusive due to the small number of studies. Further high-quality VR trials with longer-term follow-up, adequate sample sizes, and cost-effectiveness analysis will inform the role of VR with different immersive levels in chronic musculoskeletal pain management.

Acknowledgments

We thank Dr Alexander Stamenkovic (PhD, Virginia Commonwealth University) and his research team for sharing their findings that supported our manuscript development.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File , 22 KB - jmir_v26i1e48787_app1.docx]

Multimedia Appendix 2
Summary of included studies.
[DOCX File , 36 KB - jmir_v26i1e48787_app2.docx]

Multimedia Appendix 3
PRISMA checklist.
[PDF File (Adobe PDF File), 112 KB - jmir_v26i1e48787_app3.pdf]

References

1. Leveille SG, Ling S, Hochberg MC, Resnick HE, Bandeen-Roche KJ, Won A, et al. Widespread musculoskeletal pain and the progression of disability in older disabled women. Ann Intern Med 2001 Dec 18;135(12):1038-1046. [doi: 10.7326/0003-4819-135-12-200112180-00007] [Medline: 11747382]
2. Scudds RJ, McD Robertson J. Empirical evidence of the association between the presence of musculoskeletal pain and physical disability in community-dwelling senior citizens. Pain 1998 Apr;75(2-3):229-235. [doi: 10.1016/s0304-3959(97)00224-8] [Medline: 9583758]
3. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Dec;396(10267):2006-2017. [doi: 10.1016/s0140-6736(20)32340-0]
4. Blyth FM, Briggs AM, Schneider CH, Hoy DG, March LM. The global burden of musculoskeletal pain—where to from here? Am J Public Health 2019 Jan;109(1):35-40. [doi: 10.2105/ajph.2018.304747]
5. Peterson K, Anderson J, Bourne D, Mackey K, Helfand M. Effectiveness of models used to deliver multimodal care for chronic musculoskeletal pain: a rapid evidence review. J Gen Intern Med 2018 May 9;33(Suppl 1):71-81 [FREE Full text] [doi: 10.1007/s11606-018-4328-7] [Medline: 29633140]
6. Ambrose KR, Golightly YM. Physical exercise as non-pharmacological treatment of chronic pain: why and when. Best Pract Res Clin Rheumatol 2015 Feb;29(1):120-130 [FREE Full text] [doi: 10.1016/j.berh.2015.04.022] [Medline: 26267006]
7. Hoffman MD, Hoffman DR. Does aerobic exercise improve pain perception and mood? A review of the evidence related to healthy and chronic pain subjects. Curr Pain Headache Rep 2007 Apr 13;11(2):93-97. [doi: 10.1007/s11916-007-0004-z] [Medline: 17367586]
8. Dietrich A, McDaniel WF. Endocannabinoids and exercise. Br J Sports Med 2004 Oct 23;38(5):536-541 [FREE Full text] [doi: 10.1136/bjsm.2004.011718] [Medline: 15388533]
9. Palazzo C, Klinger E, Dorner V, Kadri A, Thierry O, Boumenir Y, et al. Barriers to home-based exercise program adherence with chronic low back pain: patient expectations regarding new technologies. Ann Phys Rehabil Med 2016 Apr;59(2):107-113 [FREE Full text] [doi: 10.1016/j.rehab.2016.01.009] [Medline: 27050664]
10. Rousseaux F, Bicego A, Ledoux D, Massion P, Nyssen A, Faymonville M, et al. Hypnosis associated with 3d immersive virtual reality technology in the management of pain: a review of the literature. J Pain Res 2020 May;13:1129-1138. [doi: 10.2147/jpr.s231737]
11. Gerardi M, Cukor J, Difede J, Rizzo A, Rothbaum BO. Virtual reality exposure therapy for post-traumatic stress disorder and other anxiety disorders. Curr Psychiatry Rep 2010 Aug 10;12(4):298-305. [doi: 10.1007/s11920-010-0128-4] [Medline: 20535592]
12. Sween J, Wallington S, Sheppard V, Taylor T, Llanos A, Adams-Campbell LL. The role of exergaming in improving physical activity: a review. J Phys Act Health 2014 May;11(4):864-870 [FREE Full text] [doi: 10.1123/jpah.2011-0425] [Medline: 25078529]
13. Gupta A, Scott K, Dukewich M. Innovative technology using virtual reality in the treatment of pain: does it reduce pain via distraction, or is there more to it? Pain Med 2018 Jan 01;19(1):151-159. [doi: 10.1093/pm/pnx109] [Medline: 29025113]
14. Trost Z, Zielke M, Guck A, Nowlin L, Zakhidov D, France CR, et al. The promise and challenge of virtual gaming technologies for chronic pain: the case of graded exposure for low back pain. Pain Manag 2015 May 14;5(3):197-206. [doi: 10.2217/pmt.15.6] [Medline: 25971643]
15. Ventura S, Brivio E, Riva G, Baños RM. Immersive versus non-immersive experience: exploring the feasibility of memory assessment through 360 technology. Front Psychol 2019 Nov 14;10:2509 [FREE Full text] [doi: 10.3389/fpsyg.2019.02509] [Medline: 31798492]

XSL•FO
RenderX

16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71-n112 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
17. Barthel F, Parmar M, Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design--a reanalysis of 4 trials. Trials 2009 Apr 17;10(1):21 [FREE Full text] [doi: 10.1186/1745-6215-10-21] [Medline: 19374739]
18. Nicholas M, Vlaeyen JW, Rief W, Barke A, Aziz Q, Benoliel R, IASP Taskforce for the Classification of Chronic Pain. The IASP classification of chronic pain for ICD-11: chronic primary pain. Pain 2019 Jan;160(1):28-37. [doi: 10.1097/j.pain.0000000000001390] [Medline: 30586068]
19. Ahern MM, Dean LV, Stoddard CC, Agrawal A, Kim K, Cook CE, et al. The effectiveness of virtual reality in patients with spinal pain: a systematic review and meta‐analysis. Pain Pract 2020 Jul 21;20(6):656-675. [doi: 10.1111/papr.12885] [Medline: 32196892]
20. Chuan A, Zhou JJ, Hou RM, Stevens CJ, Bogdanovych A. Virtual reality for acute and chronic pain management in adult patients: a narrative review. Anaesthesia 2021 May 27;76(5):695-704 [FREE Full text] [doi: 10.1111/anae.15202] [Medline: 32720308]
21. Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain 1986 Oct;27(1):117-126. [doi: 10.1016/0304-3959(86)90228-9] [Medline: 3785962]
22. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res (Hoboken) 2011 Nov 07;63(Suppl 11):S240-S252 [FREE Full text] [doi: 10.1002/acr.20543] [Medline: 22588748]
23. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 1983 Mar;8(2):141-144. [doi: 10.1097/00007632-198303000-00004] [Medline: 6222486]
24. Fairbank J, Couper J, Davies J, O'Brien J. The Oswestry low back pain disability questionnaire. Physiotherapy 1980;66(8):271-273. [doi: 10.1037/t04205-000]
25. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14(7):409-415. [doi: 10.1037/t35122-000]
26. Wheeler AH, Goolkasian P, Baird AC, Darden BV. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine (Phila Pa 1976) 1999 Jul 01;24(13):1290-1294. [doi: 10.1097/00007632-199907010-00004] [Medline: 10404569]
27. Monticone M, Ambrosini E, Vernon H, Brunati R, Rocca B, Foti C, et al. Responsiveness and minimal important changes for the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck pain. Eur Spine J 2015 Dec 7;24(12):2821-2827. [doi: 10.1007/s00586-015-3785-5] [Medline: 25663644]
28. Bellamy N. Pain assessment in osteoarthritis: experience with the WOMAC osteoarthritis index. Semin Arthritis Rheum 1989 May;18(4 Suppl 2):14-17. [doi: 10.1016/0049-0172(89)90010-3] [Medline: 2786253]
29. Hudak PL, Amadio PC, Bombardier C, Beaton D, Cole D, Davis A, et al. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder, and head). Am J Ind Med 1996 Jun;29(6):602-608. [doi: 10.1002/(sici)1097-0274(199606)29:6<602::aid-ajim4>3.0.co;2-l]
30. Luque-Suarez A, Martinez-Calderon J, Falla D. Role of kinesiophobia on pain, disability and quality of life in people suffering from chronic musculoskeletal pain: a systematic review. Br J Sports Med 2019 May 17;53(9):554-559 [FREE Full text] [doi: 10.1136/bjsports-2017-098673] [Medline: 29666064]
31. Van Bogaert W, Coppieters I, Kregel J, Nijs J, De Pauw R, Meeus M, et al. Influence of baseline kinesiophobia levels on treatment outcome in people with chronic spinal pain. Phys Ther 2021 Jun 01;101(6):pzab076. [doi: 10.1093/ptj/pzab076] [Medline: 33611503]
32. Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993;52(2):157-168. [doi: 10.1016/0304-3959(93)90127-b]
33. Miller RP, Kori SH, Todd DD. The Tampa Scale: a measure of kinisophobia. Clin J Pain 1991;7(1):51. [doi: 10.1097/00002508-199103000-00053]
34. French D, France C, Vigneau F, French J, Evans RT. Fear of movement/(re)injury in chronic pain: a psychometric assessment of the original English version of the Tampa scale for kinesiophobia (TSK). Pain 2007 Jan;127(1-2):42-51. [doi: 10.1016/j.pain.2006.07.016] [Medline: 16962238]
35. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002-d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]

XSL•FO
RenderX

36. Marušić MF, Fidahić M, Cepeha CM, Farca LG, Tseke A, Puljak L. Methodological tools and sensitivity analysis for assessing quality or risk of bias used in systematic reviews published in the high-impact anesthesiology journals. BMC Med Res Methodol 2020 May 18;20(1):121 [FREE Full text] [doi: 10.1186/s12874-020-00966-4] [Medline: 32423382]
37. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011 Dec;64(12):1283-1293. [doi: 10.1016/j.jclinepi.2011.01.012] [Medline: 21839614]
38. Nikolakopoulou A, Mavridis D, Salanti G. How to interpret meta-analysis models: fixed effect and random effects meta-analyses. Evid Based Ment Health 2014 May 23;17(2):64. [doi: 10.1136/eb-2014-101794] [Medline: 24778439]
39. Cohen JSPAFTBS. Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. Hillsdale, NJ: L. Erlbaum Associates; 1988.
40. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
41. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
42. Elshazly FAA, Nambi S, Elnegamy TE. Comparative study on virtual reality training (VRT) over sensory motor training (SMT) in unilateral chronic osteoarthritis - a randomized control trial. Int J Med Res Health Sci 2016 Oct;5(8):7-16.
43. Matheve T, Bogaerts K, Timmermans A. Virtual reality distraction induces hypoalgesia in patients with chronic low back pain: a randomized controlled trial. J Neuroeng Rehabil 2020 Apr 22;17(1):55 [FREE Full text] [doi: 10.1186/s12984-020-00688-0] [Medline: 32321516]
44. Nambi G, Abdelbasset WK, Elsayed SH, Alrawaili SM, Abodonya AM, Saleh AK, et al. Comparative effects of isokinetic training and virtual reality training on sports performances in university football players with chronic low back pain-randomized controlled study. Evid Based Complement Alternat Med 2020 Jun 16;2020:2981273-2981210 [FREE Full text] [doi: 10.1155/2020/2981273] [Medline: 32617104]
45. Kim T, Lee J, Oh S, Kim S, Yoon B. Effectiveness of simulated horseback riding for patients with chronic low back pain: a randomized controlled trial. J Sport Rehabil 2020 Feb 01;29(2):179-185. [doi: 10.1123/jsr.2018-0252] [Medline: 30676224]
46. Nambi G, Abdelbasset WK, Alrawaili SM, Alsubaie SF, Abodonya AM, Saleh AK. Virtual reality or isokinetic training; its effect on pain, kinesiophobia and serum stress hormones in chronic low back pain: a randomized controlled trial. Technol Health Care 2021 Jan 15;29(1):155-166. [doi: 10.3233/thc-202301]
47. Nambi G, Abdelbasset W, Alsubaie S, Saleh A, Verma A, Abdelaziz M, et al. Short-term psychological and hormonal effects of virtual reality training on chronic low back pain in soccer players. J Sport Rehabil 2021 Feb 16;30(6):884-893. [doi: 10.1123/jsr.2020-0075] [Medline: 33596538]
48. Yilmaz Yelvar GD, Çırak Y, Dalkılınç M, Parlak Demir Y, Guner Z, Boydak A. Is physiotherapy integrated virtual walking effective on pain, function, and kinesiophobia in patients with non-specific low-back pain? Randomised controlled trial. Eur Spine J 2017 Feb 15;26(2):538-545. [doi: 10.1007/s00586-016-4892-7] [Medline: 27981455]
49. Tejera D, Beltran-Alacreu H, Cano-de-la-Cuerda R, Leon Hernández JV, Martín-Pintado-Zugasti A, Calvo-Lobo C, et al. Effects of virtual reality versus exercise on pain, functional, somatosensory and psychosocial outcomes in patients with non-specific chronic neck pain: a randomized clinical trial. Int J Environ Res Public Health 2020 Aug 16;17(16):5950 [FREE Full text] [doi: 10.3390/ijerph17165950] [Medline: 32824394]
50. Sarig Bahat H, Takasaki H, Chen X, Bet-Or Y, Treleaven J. Cervical kinematic training with and without interactive VR training for chronic neck pain - a randomized clinical trial. Man Ther 2015 Feb;20(1):68-78. [doi: 10.1016/j.math.2014.06.008] [Medline: 25066503]
51. Sarig Bahat H, Croft K, Carter C, Hoddinott A, Sprecher E, Treleaven J. Remote kinematic training for patients with chronic neck pain: a randomised controlled trial. Eur Spine J 2018 Jun 10;27(6):1309-1323. [doi: 10.1007/s00586-017-5323-0] [Medline: 29018956]
52. Nusser M, Knapp S, Kramer M, Krischak G. Effects of virtual reality-based neck-specific sensorimotor training in patients with chronic neck pain: a randomized controlled pilot trial. J Rehabil Med 2021 Feb 10;53(2):jrm00151 [FREE Full text] [doi: 10.2340/16501977-2786] [Medline: 33369684]
53. Cetin H, Kose N, Oge HK. Virtual reality and motor control exercises to treat chronic neck pain: a randomized controlled trial. Musculoskelet Sci Pract 2022 Dec;62:102636. [doi: 10.1016/j.msksp.2022.102636] [Medline: 35952621]
54. Li Z, Yu Q, Luo H, Liang W, Li X, Ge L, et al. The effect of virtual reality training on anticipatory postural adjustments in patients with chronic nonspecific low back pain: a preliminary study. Neural Plast 2021 Jul 27;2021:9975862-9975813 [FREE Full text] [doi: 10.1155/2021/9975862] [Medline: 34367274]
55. Kim S, Min W, Kim J, Lee B. The effects of VR-based Wii Fit yoga on physical function in middle-aged female LBP patients. J Phys Ther Sci 2014 Apr;26(4):549-552 [FREE Full text] [doi: 10.1589/jpts.26.549] [Medline: 24764631]
56. Park J, Lee S, Ko D. The effects of the nintendo Wii exercise program on chronic work-related low back pain in industrial workers. J Phys Ther Sci 2013 Aug;25(8):985-988 [FREE Full text] [doi: 10.1589/jpts.25.985] [Medline: 24259899]
57. Monteiro-Junior R, de Souza C, Lattari E, Rocha N, Mura G, Machado S, et al. Wii-workouts on chronic pain physical capabilities and mood of older women a randomized controlled double blind trial. CNS Neurol Disord Drug Targets 2015 Nov 27;14(9):1157-1164. [doi: 10.2174/1871527315666151111120131] [Medline: 26556092]

XSL•FO
RenderX

58. Chen S, Kim S, Kim K, Lee I, HwangBo G. Effects of horse riding simulator on pain, Oswestry disability index and balance in adults with nonspecific chronic low back pain. J Korean Soc Phys Med 2016 Nov 30;11(4):79-84. [doi: 10.13066/kspm.2016.11.4.79]
59. Pekyavas NO, Ergun N. Comparison of virtual reality exergaming and home exercise programs in patients with subacromial impingement syndrome and scapular dyskinesis: short term effect. Acta Orthop Traumatol Turc 2017 May;51(3):238-242 [FREE Full text] [doi: 10.1016/j.aott.2017.03.008] [Medline: 28446376]
60. Ditchburn J, van Schaik P, Dixon J, MacSween A, Martin D. The effects of exergaming on pain, postural control, technology acceptance and flow experience in older people with chronic musculoskeletal pain: a randomised controlled trial. BMC Sports Sci Med Rehabil 2020 Oct 09;12(1):63 [FREE Full text] [doi: 10.1186/s13102-020-00211-x] [Medline: 33062284]
61. Lin Y, Lee W, Hsieh R. Active video games for knee osteoarthritis improve mobility but not WOMAC score: a randomized controlled trial. Ann Phys Rehabil Med 2020 Nov;63(6):458-465 [FREE Full text] [doi: 10.1016/j.rehab.2019.11.008] [Medline: 31981832]
62. Afzal MW, Ahmad A, Mohseni Bandpei MA, Gillani SA, Hanif A, Sharif Waqas M. Effects of virtual reality exercises and routine physical therapy on pain intensity and functional disability in patients with chronic low back pain. J Pak Med Assoc 2022 Feb 03;72(3):413-417. [doi: 10.47391/jpma.3424]
63. Nambi G, Alghadier M, Kashoo FZ, Aldhafian OR, Nwihadh NA, Saleh AK, et al. Effects of virtual reality exercises versus isokinetic exercises in comparison with conventional exercises on the imaging findings and inflammatory biomarker changes in soccer players with non-specific low back pain: a randomized controlled trial. Int J Environ Res Public Health 2022 Dec 28;20(1):524 [FREE Full text] [doi: 10.3390/ijerph20010524] [Medline: 36612846]
64. Stamm O, Dahms R, Reithinger N, Ruß A, Müller-Werdan U. Virtual reality exergame for supplementing multimodal pain therapy in older adults with chronic back pain: a randomized controlled pilot study. Virtual Real 2022 Feb 11;26(4):1291-1305 [FREE Full text] [doi: 10.1007/s10055-022-00629-3] [Medline: 35194374]
65. Battecha K, Takrouni H, Alzahrani R, Alzahrani R, Alzahrani F, Mohamed HS. Efficacy of virtual reality on neck pain and function in patients with non-specific neck dysfunction: a randomized control trial. Revista iberoamericana de psicología del ejercicio y el deporte 2023;18(5):543-546.
66. Rezaei I, Razeghi M, Ebrahimi S, Kayedi S, Rezaeian Zadeh A. A novel virtual reality technique (Cervigame©) compared to conventional proprioceptive training to treat neck pain: a randomized controlled trial. J Biomed Phys Eng 2018 Nov 18;9(3):355-366. [doi: 10.31661/jbpe.v0i0.556]
67. Soysal Tomruk M, Kara B, Erbayraktar RS. The effect of computer-based training on postural control in patients with chronic low back pain: a randomized controlled trial. J Basic Clin Health Sci 2020;4:329-334. [doi: 10.30621/jbachs.2020.1150]
68. Oliveira LKR, Marques AP, Andrade KFA, Assis JCSD, Brito AL, Souza GS, et al. Virtual reality in improving anticipatory postural adjustments to step initiation in individuals with knee osteoarthritis: a randomized controlled trial. Games Health J 2024 Apr 01;13(2):100-108. [doi: 10.1089/g4h.2023.0154] [Medline: 38436585]
69. Guo Q, Zhang L, Han LL, Gui C, Chen G, Ling C, et al. Effects of virtual reality therapy combined with conventional rehabilitation on pain, kinematic function, and disability in patients with chronic neck pain: randomized controlled trial. JMIR Serious Games 2024 Apr 24;12:e42829 [FREE Full text] [doi: 10.2196/42829] [Medline: 38656775]
70. Stamm O, Dahms R, Müller-Werdan U. Virtual reality in pain therapy: a requirements analysis for older adults with chronic back pain. J Neuroeng Rehabil 2020 Sep 29;17(1):129 [FREE Full text] [doi: 10.1186/s12984-020-00753-8] [Medline: 32993678]
71. Chiarotto A, Maxwell L, Terwee C, Wells G, Tugwell P, Ostelo RW. Roland-Morris Disability Questionnaire and Oswestry Disability Index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? Systematic review and meta-analysis. Phys Ther 2016 Oct;96(10):1620-1637. [doi: 10.2522/ptj.20150420] [Medline: 27081203]
72. Carreon LY, Glassman SD, Campbell MJ, Anderson PA. Neck Disability Index, Short Form-36 Physical Component Summary, and pain scales for neck and arm pain: the minimum clinically important difference and substantial clinical benefit after cervical spine fusion. Spine J 2010 Jun;10(6):469-474. [doi: 10.1016/j.spinee.2010.02.007] [Medline: 20359958]
73. Hayden JA, van Tulder MW, Tomlinson G. Systematic review: strategies for using exercise therapy to improve outcomes in chronic low back pain. Ann Intern Med 2005 May 03;142(9):776-785. [doi: 10.7326/0003-4819-142-9-200505030-00014] [Medline: 15867410]
74. Maddox T, Sparks CY, Oldstone L, Chibbaro M, Sackman J, Judge E, et al. Perspective: the promise of virtual reality as an immersive therapeutic. J Med Ext Real 2024 Jan 01;1(1):13-20. [doi: 10.1089/jmxr.2023.0003]
75. Brea-Gómez B, Torres-Sánchez I, Ortiz-Rubio A, Calvache-Mateo A, Cabrera-Martos I, López-López L, et al. Virtual reality in the treatment of adults with chronic low back pain: a systematic review and meta-analysis of randomized clinical trials. Int J Environ Res Public Health 2021 Nov 11;18(22):11806 [FREE Full text] [doi: 10.3390/ijerph182211806] [Medline: 34831562]
76. Bordeleau M, Stamenkovic A, Tardif P, Thomas J. The use of virtual reality in back pain rehabilitation: a systematic review and meta-analysis. J Pain 2022 Feb;23(2):175-195 [FREE Full text] [doi: 10.1016/j.jpain.2021.08.001] [Medline: 34425250]

XSL•FO
RenderX

77. Li R, Li Y, Kong Y, Li H, Hu D, Fu C, et al. Virtual reality–based training in chronic low back pain: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2024 Feb 26;26:e45406 [FREE Full text] [doi: 10.2196/45406] [Medline: 38407948]
78. Laws KR, Pellegrini L, Reid JE, Drummond LM, Fineberg NA. The inflating impact of waiting-list controls on effect size estimates. Front Psychiatry 2022 Jun 22;13:877089 [FREE Full text] [doi: 10.3389/fpsyt.2022.877089] [Medline: 35815038]
79. Guo Q, Zhang L, Gui C, Chen G, Chen Y, Tan H, et al. Virtual reality intervention for patients with neck pain: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 03;25:e38256 [FREE Full text] [doi: 10.2196/38256] [Medline: 37010891]
80. Brea-Gómez B, Laguna-González A, Pérez-Gisbert L, Valenza MC, Torres-Sánchez I. Virtual reality based rehabilitation in adults with chronic neck pain: a systematic review and meta-analysis of randomized clinical trials. Virtual Real 2024 Mar 27;28(86):1-31. [doi: 10.1007/s10055-024-00979-0]
81. Byra J, Czernicki K. The effectiveness of virtual reality rehabilitation in patients with knee and hip osteoarthritis. J Clin Med 2020 Aug 14;9(8):2639 [FREE Full text] [doi: 10.3390/jcm9082639] [Medline: 32823832]
82. Kantha P, Lin J, Hsu W. The effects of interactive virtual reality in patients with chronic musculoskeletal disorders: a systematic review and meta-analysis. Games Health J 2023 Feb 01;12(1):1-12. [doi: 10.1089/g4h.2022.0088] [Medline: 36706260]
83. Bevers K, Watts L, Kishino N, Gatchel RJ. The biopsychosocial model of the assessment, prevention, and treatment of chronic pain. US Neurol 2016;12(2):98-104. [doi: 10.17925/usn.2016.12.02.98]
84. Yeo SM, Lim JY, Do JG, Lim J, In Lee J, Hwang JH. Effectiveness of interactive augmented reality-based telerehabilitation in patients with adhesive capsulitis: protocol for a multi-center randomized controlled trial. BMC Musculoskelet Disord 2021 Apr 26;22(1):386 [FREE Full text] [doi: 10.1186/s12891-021-04261-1] [Medline: 33902546]
85. Safiri S, Kolahi A, Smith E, Hill C, Bettampadi D, Mansournia MA, et al. Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the Global Burden of Disease Study 2017. Ann Rheum Dis 2020 Jun 12;79(6):819-828. [doi: 10.1136/annrheumdis-2019-216515] [Medline: 32398285]

Abbreviations

GRADE: Grading of Recommendation Assessment, Development, and Evaluation
HMD: head-mounted device
MCID: minimally clinically important difference
MD: mean difference
ODI: Oswestry Disability Index
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RMDQ: Roland-Morris Disability Questionnaire
RoB 2: Cochrane Risk-of-Bias Tool version 2
SMD: standardized mean difference
VR: virtual reality

Edited by T de Azevedo Cardoso; submitted 07.05.23; peer-reviewed by M Gasmi , J Aulenkamp, PG Vajargah; comments to author 10.05.24; revised version received 20.06.24; accepted 04.07.24; published 19.08.24.

Please cite as:
Lo HHM, Zhu M, Zou Z, Wong CL, Lo SHS, Chung VCH, Wong SYS, Sit RWS
Immersive and Nonimmersive Virtual Reality–Assisted Active Training in Chronic Musculoskeletal Pain: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e48787
URL: https://www.jmir.org/2024/1/e48787
doi:10.2196/48787
PMID:

©Hermione Hin Man Lo, Mengting Zhu, Zihui Zou, Cho Lee Wong, Suzanne Hoi Shan Lo, Vincent Chi-Ho Chung, Samuel Yeung-Shan Wong, Regina Wing Shan Sit. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review

Laura Swinckels[1,2,3,4], MSc; Frank C Bennis[5,6,7], PhD; Kirsten A Ziesemer[8], MSc; Janneke F M Scheerman[2,3], PhD; Harmen Bijwaard[3], PhD; Ander de Keijzer[4,9], PhD; Josef Jan Bruers[1,10], PhD

[1]Department of Oral Public Health, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and Vrije Universiteit, Amsterdam, Netherlands

[2]Department Oral Hygiene, Cluster Health, Sports and Welfare, Inholland University of Applied Sciences, Amsterdam, Netherlands

[3]Medical Technology Research Group, Cluster Health, Sport and Welfare, Inholland University of Applied Sciences, Haarlem, Netherlands

[4]Data Driven Smart Society Research Group, Faculty of Engineering, Design & Computing, Inholland University of Applied Sciences, Alkmaar, Netherlands

[5]Quantitative Data Analytics Group, Department of Computer Science, Vrije Universiteit, Amsterdam, Netherlands

[6]Department of Pediatrics, Emma Neuroscience Group, Emma Children's Hospital, Amsterdam UMC, Amsterdam, Netherlands

[7]Amsterdam Reproduction and Development Research Institute, Amsterdam, Netherlands

[8]Medical Library, University Library, Vrije Universiteit, Amsterdam, Netherlands

[9]Applied Responsible Artificial Intelligence, Avans University of Applied Sciences, Breda, Netherlands

[10]Royal Dutch Dental Association (KNMT), Utrecht, Netherlands

Corresponding Author:
Laura Swinckels, MSc
Department of Oral Public Health
Academic Centre for Dentistry Amsterdam (ACTA)
University of Amsterdam and Vrije Universiteit
Gustav Mahlerlaan 3004
Amsterdam, 1081 LA
Netherlands
Phone: 31 205980308
Email: L.Swinckels@acta.nl

Abstract

Background: Electronic health records (EHRs) contain patients' health information over time, including possible early indicators of disease. However, the increasing amount of data hinders clinicians from using them. There is accumulating evidence suggesting that machine learning (ML) and deep learning (DL) can assist clinicians in analyzing these large-scale EHRs, as algorithms thrive on high volumes of data. Although ML has become well developed, studies mainly focus on engineering but lack medical outcomes.

Objective: This study aims for a scoping review of the evidence on how the use of ML on longitudinal EHRs can support the early detection and prevention of disease. The medical insights and clinical benefits that have been generated were investigated by reviewing applications in a variety of diseases.

Methods: This study was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. A literature search was performed in 2022 in collaboration with a medical information specialist in the following databases: PubMed, Embase, Web of Science Core Collection (Clarivate Analytics), and IEEE Xplore Digital Library and computer science bibliography. Studies were eligible when longitudinal EHRs were used that aimed for the early detection of disease via ML in a prevention context. Studies with a technical focus or using imaging or hospital admission data were beyond the scope of this review. Study screening and selection and data extraction were performed independently by 2 researchers.

Results: In total, 20 studies were included, mainly published between 2018 and 2022. They showed that a variety of diseases could be detected or predicted, particularly diabetes; kidney diseases; diseases of the circulatory system; and mental, behavioral, and neurodevelopmental disorders. Demographics, symptoms, procedures, laboratory test results, diagnoses, medications, and

XSL•FO
RenderX

BMI were frequently used EHR data in basic recurrent neural network or long short-term memory techniques. By developing and comparing ML and DL models, medical insights such as a high diagnostic performance, an earlier detection, the most important predictors, and additional health indicators were obtained. A clinical benefit that has been evaluated positively was preliminary screening. If these models are applied in practice, patients might also benefit from personalized health care and prevention, with practical benefits such as workload reduction and policy insights.

Conclusions: Longitudinal EHRs proved to be helpful for support in health care. Current ML models on EHRs can support the detection of diseases in terms of accuracy and offer preliminary screening benefits. Regarding the prevention of diseases, ML and specifically DL models can accurately predict or detect diseases earlier than current clinical diagnoses. Adding personally responsible factors allows targeted prevention interventions. While ML models based on textual EHRs are still in the developmental stage, they have high potential to support clinicians and the health care system and improve patient outcomes.

(J Med Internet Res 2024;26:e48320) doi:10.2196/48320

KEYWORDS

artificial intelligence; big data; detection; electronic health records; machine learning; personalized health care; prediction; prevention

Introduction

Rationale

Digitizing meaningful health information has been proven to contribute to diagnostics. Electronic health records (EHRs) are a digital repository of patient data and contain retrospective, current, and prospective information supporting health care [1]. EHRs contain a wealth of clinical information about early symptoms of a disease and registries of medical treatments [2]. These can be textual or imaging data and include both unstructured clinical notes and structured, coded data. One important aspect of textual EHRs is that they may include risk and preventive factors and early signs before a disease manifests. Especially for patients with multiple visits, many possible indicators are gathered in EHRs, resulting in possible early indications of disease. Therefore, for a good risk assessment, clinicians need the patient's health information, physical examinations, laboratory test results, and history [3] available in EHRs.

In the past 15 years, an explosion in the volume of data registered in EHR systems has occurred [4]. In 2012, the yearly increase in the volume of stored data was up to 150% for hospitals [5]. Not only the number of records continues to increase over time, but EHRs are also quite extensive because of large free texts [6]. Even though the completeness and correctness of EHRs have been found to be at a high level [7], the usability during medical visits lags behind due to this rising volume and variety of EHR data [8]. Consequently, it has even become an experienced usability issue for clinicians to review clinical results and health information from the past [9]. This is quite problematic as some clinicians spend, on average, 32.1% of their time on EHRs reviewing medical care and notes from the past [10]. The increasing EHR workload causes exhaustion and burnout among clinicians [11], negatively affecting the health care quality. This can result in diagnostic errors (missed, delayed, or incorrect diagnoses) because of missed signs [12] registered in the past. In 67.4% of the cases, missing the chief presenting symptoms in EHRs was the reason for missed diagnoses. Overall, meaningful health records have the potential to support risk assessment and early diagnosis, but the increasing amount of data hinder clinicians from using them to their full potential.

It is currently known that supportive tools can simplify complex diagnostic tasks and reduce potential diagnostic errors [13]. There is accumulating evidence suggesting that machine learning (ML) can assist clinicians in analyzing large-scale EHRs as they thrive on high volumes of data. ML is able to fit models specifically adapted to patterns in the data and, compared to traditional statistics, is able to handle multidimensional data [14]. Deep learning (DL) is a subdomain of ML that uses neural networks with multiple (hidden) layers, incorporating complex interactions between variables [15]. Examples of well-developed ML models are based on imaging data for disease detection [16,17] and textual EHRs of hospitalization or intensive care data for predicting disease progression or therapy success [18]. One of the most promising aspects of DL in the context of EHRs containing historical and present clinical data is the ability to incorporate temporality into the model, that is, to base possible risk assessments on hidden patterns over time in clinical parameters. Indeed, DL models have also proved to be more effective by incorporating temporal information (ie, longitudinally processed) rather than cross-sectional information only [19]. Although the techniques of many ML (including DL) models have proved to be effective on EHRs, their focus is often on the engineering of architectures and frameworks [20], but they lack medical outcomes.

Objectives

It is a loss of information if ML developments remain unknown in health care because of the technical perspective of most authors. Especially given that artificial intelligence (AI) is a black box, it is important to clarify the clinical benefits and additional medical insights that can be achieved through these techniques. Therefore, the aim of this review was to perform a scoping review of the evidence on how the use of ML on longitudinal EHRs can support the early detection and prevention of diseases. A preliminary search was conducted, and no current or underway systematic or scoping reviews on the topic were identified. Only 1 review on longitudinal EHRs has been conducted [2], but it focused on methodologies. This study will contribute to what is already known by scoping the substantive medical insights that ML models yield. Given the

XSL•FO
RenderX

aim of this study, the following research questions were addressed:

1. Which diseases have been detected in longitudinal EHRs using ML techniques?
2. What EHR data have been used by ML methods for the early detection and prevention of diseases?
3. What medical insights are generated by developing and using ML models on longitudinal EHRs?
4. What clinical benefits may be reached through the application of ML models on longitudinal EHRs?

Methods

The conduct and reporting of this scoping review adhere to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) statement [21]. A protocol has been registered in the Open Science Framework (DOI: NY2TE).

Eligibility Criteria

Articles were included if they reported on early detection for timely prevention of diseases by using ML on longitudinal EHRs; the full description of eligible participants, concept, context, and types of sources can be found in the protocol. Overall, studies were screened according to several criteria.

Focus

Studies must have a clear focus on health care instead of a technical focus (eg, the article must include disease-specific information and interpretation, preferably executed and written from a health care perspective, and reflect on health or related care outcomes). Studies with a dominant technical focus or an engineering challenge or those using non–real-world data were assumed to be ineligible for this review.

Purpose

ML (including DL) should be aimed at predicting, detecting, or contributing to the risk assessment of diseases. Models aiming for data extraction, clustering, or patient selection for trials did not fit this concept. The purpose also affects the technique used.

Outcome

The prediction target of ML must be (the onset of) a disease or a medical event. By using the *International Classification of Diseases, 11th Revision* [22], we ensured that the primary outcomes were a disease or related medical event (ie, the cause of morbidity or mortality). Thus, studies that predicted disease severity once diagnosed, success of treatment, adverse drug reactions, phenotypes, or events that were not the cause of morbidity or mortality and did not focus on timely prevention were beyond the scope of this research. If the outcome was mortality, these articles were excluded because it is always a consequence of a disease or medical event.

Essential Elements of ML

Studies must incorporate the essential elements of ML, such as training, testing, or validation steps. DL was assumed as a subdomain within ML and, therefore, was included as well.

Data

According to the broadest definition of an EHR [1], data were assumed as EHR data if these contained information supporting continuing, efficient, and quality integrated health care or describing the health status of a patient regardless of the collecting database. Studies must use manually entered EHR data, including textual and numeric values. Both structured (numeric or coded) and unstructured (clinical notes) data were accepted as eligible EHR data. EHRs with solely imaging data (such as x-rays or electrocardiograms) were beyond the scope of this review. EHRs from animals were excluded.

Longitudinal

Studies must use EHRs over time registered at multiple visits (before registering a disease or medical event).

Context

Studies were included if they were conducted in the context of disease prevention. Optimal prevention in health care settings can be reached when participants at risk or signs of a disease are detected as early as possible, and therefore, these studies were eligible in the context of secondary prevention. Secondary prevention emphasizes early disease detection in subclinical forms and seeks to prevent the onset of illness [23]. Studies conducted using data gathered in intensive care settings during a hospital admission or data gathered at the emergency department cannot be viewed in the context of disease prevention because only tertiary preventive measures can be taken to reduce the effects or severity of the established disease as it is too late to influence the onset of disease.

Sample Size

Because ML is data driven (instead of conventional models that are hypothesis driven), only predictions based on >1000 participants in total were considered eligible. This threshold is based on theory (eg, calculations for multivariable predictions of binary outcomes [24]) and practice (eg, the range of sample sizes for disease prediction models on EHRs seen in the literature).

Study Design

Only study designs with clinical, real-world data were considered. If secondary research, such as other reviews, met the aforementioned criteria, the reference list was considered depending on the research question. Conference papers were also considered because of the high quality of evidence in computer science.

Search

After several preliminary searches, 5 bibliographic databases (PubMed, Embase, Web of Science Core Collection [Clarivate Analytics], IEEE Xplore Digital Library, and computer science bibliography) were searched for relevant literature from inception to April 28, 2022. Searches were devised in collaboration with a medical information specialist (KAZ). The following search terms, including synonyms, closely related words, and keywords, were used as index terms or free-text words: “neural network,” “electronic medical record,” and “prediction.” We used only search terms capturing specific ML

XSL•FO
RenderX

techniques that are able to predict or classify. The search strategy was adapted for each included database or information source. The searches contained no methodological search filter or date or language restrictions that would limit results to specific study designs, dates, or languages. We searched computer science bibliography for conference proceedings and hand searched meeting abstracts. Duplicate articles were excluded using the R package *ASYSD* (R Foundation for Statistical Computing), an automated deduplication tool [25], followed by manual deduplication in EndNote (version X20.0.3; Clarivate Analytics) by the medical information specialist (KAZ). The full search strategy used for each database is detailed in Multimedia Appendix 1.

Study Selection

Following the search, all identified citations were collated and uploaded into Rayyan (Rayyan Systems Inc) [26] and EndNote (version X7.8). In total, 2 reviewers (LS and FCB) independently screened all potentially relevant titles and abstracts for eligibility. If necessary, the full-text article was checked against the eligibility criteria. Differences in judgment were resolved through a consensus procedure. The full texts of the selected articles were obtained for further review. As the aim was not to search for "the best available" evidence but to identify and perform a scoping review of all evidence, a critical appraisal was not systematically carried out.

Data Extraction

Data were extracted from the papers included in the scoping review by 2 independent reviewers (LS and FCB) using a data extraction form developed in Microsoft Excel (Microsoft Corp). This form was composed based on full-text findings relevant to the research question and was discussed by the research team. The data extraction sheet captured details about study characteristics, health care discipline, generated medical insights, and clinical benefits for health care and the way EHRs were processed temporally. Multimedia Appendix 2 provides the list and definitions of all data items. This form was piloted using the first 5 articles and was revised and slightly adjusted during the process of extracting data. The extraction of ML techniques was modified to include the extraction of all techniques that were internally compared by appointing the central model and the comparison. Any disagreements between the reviewers were resolved through discussion with additional reviewers. Authors were contacted to request missing or additional data where required.

Synthesis of Results

Extracted data were synthesized into results by frequency counts of concepts and qualitative narratives. Study characteristics, detected diseases, and EHR variables were listed in tabular form. The content of these tables was sorted by disease outcomes according to the *International Classification of Diseases, 11th Revision* disease categories from the World Health Organization. For data concerning medical insights and clinical benefits, a qualitative content analysis was carried out according to the guidance for scoping review knowledge syntheses [27,28]. After each study's key findings were extracted, these were classified into concepts (1-6) and described using a narrative summary. We decided to describe both similarities and exceptions of the generated results and potential impact.

Results

Selection of Evidence

The literature search generated a total of 895 references. After removing duplicates of references that were selected from >1 database, 483 (54%) of the references remained. By screening titles and abstracts, 426 (88.2%) of the articles were excluded. Of the remaining 57 articles, 2 (4%) could not be retrieved because they contained unpublished work. In the second phase, 55 full texts were reviewed for eligibility, and ultimately, 20 (36%) articles were included. Reports were mostly excluded due to wrong data, a technical focus, the absence of a longitudinal aspect, or models based on N<1000. No additional studies were found by checking reference lists. After the final screening, most included articles (18/20, 90%) were found in PubMed. The flowchart of the search and selection process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of study selection. ML: machine learning. DBLP: DataBase systems and Logic Programming.

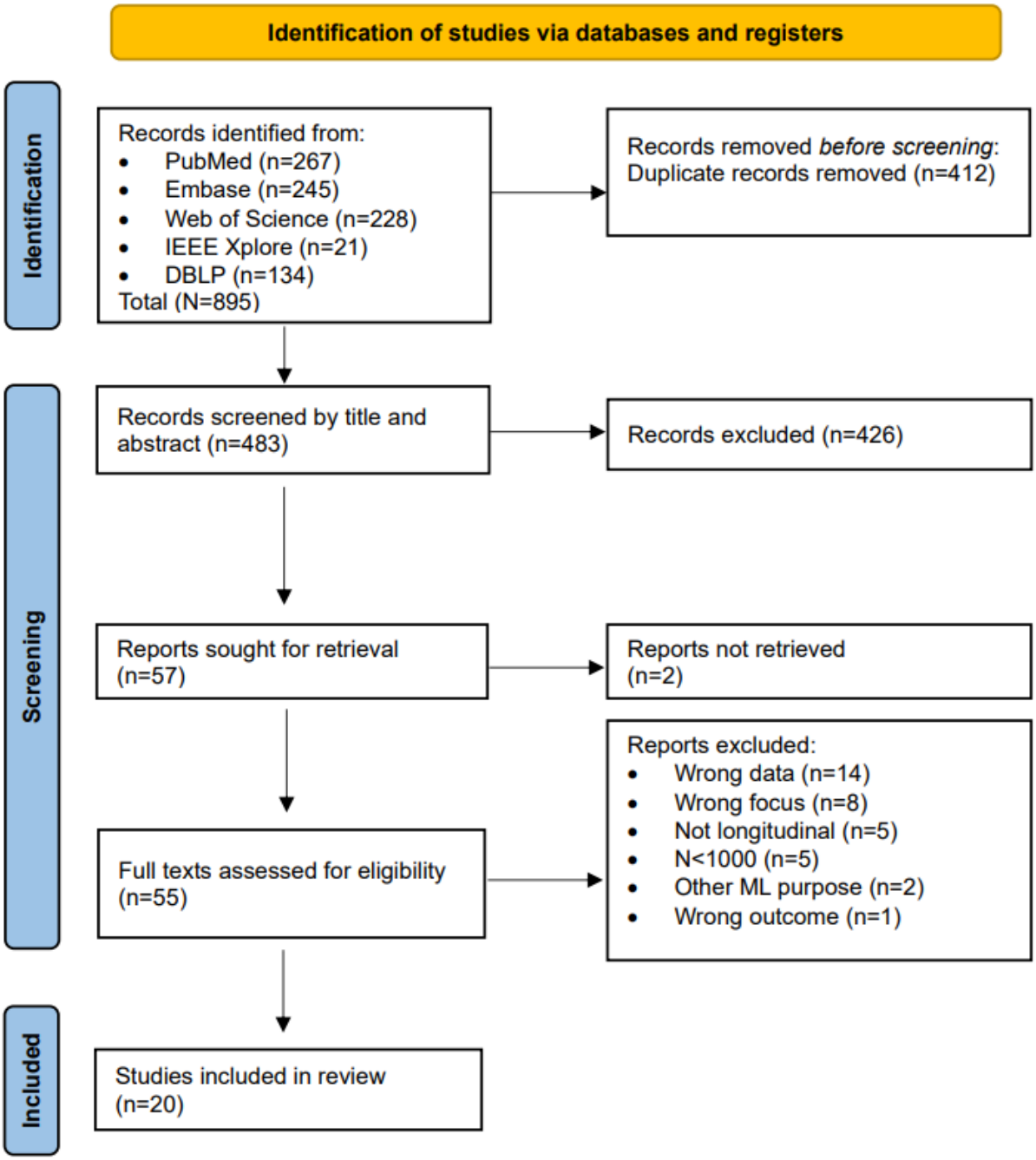

Characteristics of the Included Studies

Of the 20 included articles [29-48], 19 (95%) were published between 2018 and 2022, and 1 (5%) was published in 2016. The aim of these studies to develop an ML or DL model and examine whether it was able to detect the disease of interest in longitudinal EHRs. Detected diseases or related medical events were hepatocellular carcinoma [29], type 2 diabetes or prediabetes mellitus [30,31], mental health conditions [32], dementia [33,36], cognitive impairment [34], psychosis [35], heart failure [37], cardiac dysrhythmia [38], cardiovascular and cerebrovascular events [39], cardiovascular disease [40], knee osteoarthritis [41], kidney function decline [42,43], extreme preterm birth [44], opioid overdose [45], and suicide attempts [46]. One study proposed a health index [47] based on the prediction of 3 important health events, and another study predicted future disease in the next hospital visit [48]. Sample sizes ranged from thousands to millions. In total, 10% (2/20) of the studies used an external validation data set [35,39]. Table 1 shows the included studies and the detected diseases.

XSL•FO
RenderX

Table 1. Overview of the included studies and detected diseases.

Study, year	Disease or medical event	Aim of the study	Sample size, N
Neoplasms			
Ioannou et al [29], 2020	Hepatocellular carcinoma	To examine whether deep learning recurrent neural network models that use raw longitudinal data extracted directly from EHRs[a] outperform conventional regression models in predicting the risk of developing hepatocellular carcinoma	48,151
Endocrine, nutritional, or metabolic diseases (diabetes)			
Alhassan et al [30], 2021	Prediabetes—HbA_{1c}[b] elevation	To identify patients without diabetes that are at a high risk of HbA_{1c} elevation	18,844
Pimentel et al [31], 2018	Type 2 diabetes mellitus	To propose a new prognostic approach for type 2 diabetes mellitus given an EHR and without using the current invasive techniques that are related to the disease	9947
Mental, behavioral, and neurodevelopmental disorders			
Dabek et al [32], 2022	Mental health conditions (anxiety, depression, and adjustment disorder)	To evaluate the utility of machine learning models and longitudinal EHR data to predict the likelihood of developing mental health conditions following the first diagnosis of mild traumatic brain injury	35,451
Ford et al [33], 2019	Dementia	To detect existing dementia before any evidence that the GP[c] had done so, that is, before they had started recording memory loss symptoms or initiating the process of dementia diagnosis	93,120
Fouladvand et al [34], 2019	Mild cognitive impairment	To predict the progression from cognitively unimpaired to mild cognitive impairment and also analyze the potential for patient clustering using routinely collected EHR data	3265
Raket et al [35], 2020	The first episode of psychosis	To develop and validate an innovative risk prediction model (DETECT[d]) to detect individuals at risk of developing a first episode of psychosis through EHRs that contain data from both primary and secondary care	102,030 (training)+43,690 (external validation)
Shao et al [36], 2019	Dementia	To identify cases of undiagnosed dementia by developing and validating a weakly supervised machine learning approach that incorporated the analysis of both structured and unstructured EHR data	11,166
Diseases of the circulatory system			
Choi et al [37], 2016	Heart failure	To explore whether the use of deep learning to model temporal relations among events in EHRs would improve model performance in predicting initial diagnosis of heart failure compared to conventional methods that ignore temporality	32,787
Guo et al [38], 2021	Cardiac dysrhythmia	To predict cardiac dysrhythmias using EHR data for earlier diagnosis and treatment of the condition, thus improving overall cardiovascular outcomes	11,055
Park et al [39], 2019	Cardiovascular and cerebrovascular events	To develop and compare machine learning models predicting high-risk vascular diseases for patients with hypertension so that they can manage their blood pressure based on their risk level	74,535 (training)+59,738 (validation)
Zhao et al [40], 2019	Cardiovascular disease	To apply machine learning and deep learning models to 10-year cardiovascular event prediction by using longitudinal EHRs and genetic data	109,490
Diseases of the musculoskeletal system or connective tissue			
Ningrum et al [41], 2021	Knee osteoarthritis	To develop a deep learning model (Deep-KOA[e]) that can predict the risk of knee osteoarthritis within the next year by using non–image-based electronic medical record data from the previous 3 years	1,201,058
Diseases of the genitourinary system			
Chauhan et al [42], 2020	Rapid kidney function decline	To examine the ability of a prognostic test (KidneyIntelX) that uses machine learning algorithms to predict rapid kidney function decline and kidney outcomes in 2 discrete, high-risk patient populations: type 2 diabetes and APOL1-HR[f]	871 (data set 1); 498 (data set 2)
Inaguma et al [43], 2020	Decline of kidney function (eGFR[g])	To predict the rapid decline in kidney function among patients with chronic kidney disease by using a big hospital database and develop a machine learning–based model	118,584
Conditions originating in the perinatal period			

XSL•FO
RenderX

Study, year	Disease or medical event	Aim of the study	Sample size, N
Gao et al [44], 2019	Extreme preterm birth	To investigate the extent to which deep learning models that consider temporal relations documented in EHRs can predict extreme preterm birth	25,689
External causes of morbidity (self-harm)			
Dong et al [45], 2021	Opioid overdose	To build a deep learning model that can predict patients at high risk of opioid overdose and identify the most relevant features	5,231,614
Walsh et al [46], 2018	Suicide attempts	To evaluate machine learning applied to EHRs as a potential means of accurate large-scale risk detection and screening for suicide attempts in adolescents applicable to any clinical setting with an EHR	1470 (data set 1); 8033 (data set 2); 26,055 (data set 3)
Multi-disease or other			
Hung et al [47], 2020	Health index	To propose a novel health index developed by using deep learning techniques with a large-scale population-based EHR	383,322 (training); 95,746 (testing 1); 102,625 (testing 2)
Wang et al [48], 2020	Multi-disease	To explore how to predict future disease risks in the next hospital visit of a patient when discharged from a hospital	7105 (data set 1); 4170 (data set 2)

[a]EHR: electronic health record.

[b]HbA_{1c}: glycated hemoglobin.

[c]GP: general practitioner.

[d]DETECT: Dynamic Electronic Health Record Detection.

[e]KOA: knee osteoarthritis.

[f]APOL1-HR: apolipoprotein L1 high-risk.

[g]eGFR: estimated glomerular filtration rate.

EHR Data

The EHRs of patients used in the included studies were originally recorded in hospitals or primary care centers. Especially for the detection of mental and behavioral disorders, EHRs were often extracted from military health records [32,36], and for neurodevelopmental and cardiovascular disorders, EHRs were mostly extracted from general practices [33,37]. Most studies (16/20, 80%) used structured EHRs [29-33,35,38-43,45-48], sometimes combined with unstructured data [34,36,37,44], to estimate the risk of a disease or medical event. Demographic information (statically used), symptoms, laboratory (blood) test results, diagnoses, medications, BMI, and clinical notes were commonly used data from EHRs. In addition, the EHR length and hospital admission and visit history were frequently added to the model. Lifestyle data were included for cardiovascular diseases. Clinical and social signs were more frequently used for self-harm and mental, behavioral, and neurodevelopmental disorders. For the prediction of kidney and diabetes outcomes, laboratory test results were frequently extracted. If EHRs were unstructured, natural language processing methods were conducted as a precursor to analyze clinical notes. The central techniques were a basic recurrent neural network (RNN) or long short-term memory (LSTM) [29,31,34,35,39,44,45,49], often compared with logistic regression, support vector machine, or random forest. When techniques were used that could not handle temporal data, a temporal aspect was created in the data. Although not extensively specified, a slope and intercept of variables [31,36]; a mean [30]; minimum, maximum, median, and SD measures [42]; the addition of a time-weight (eg, 0.9 × days from reference point+decay) [43]; different time stamps [42]; or dividing the data into time blocks [33,46] were used. Multimedia Appendix 3 [29-48] provides an overview of the EHR data used and the techniques applied.

Medical Insights

Overview

Disease detection and prevention can be supported by using ML or DL on longitudinal EHRs. First, the development and training of such models on EHRs can generate new medical insights (1-4). Second, when those models are applied (eg, for additional analyses or to “new” data in clinical practice), the following clinical benefits may be achieved (5 and 6). These insights will be summarized in the following sections.

Medical Insight 1: Diagnostic Performance

The use of ML and DL models on EHRs could support the detection of diseases with a high diagnostic accuracy. Performance metrics such as the area under the receiver operating characteristic curve (AUROC), sensitivity (recall), specificity, accuracy, precision, and the area under the precision-recall curve evaluated the detecting ability of the model. The AUROC was by far the most frequently reported metric because it illustrates the diagnostic ability for a binary classification (disease or nondisease) by using the sensitivity versus the specificity. Although it is not our intention to identify the best-performing model, it was observed that the AUROC

XSL•FO
RenderX

of central models varied between 0.73 and 0.97. In 40% (8/20) of the studies, the optimal model had a “good” detection (AUROC between 0.7 and 0.8), 35% (7/20) of the studies succeeded in having a “very good” detection (AUROC between 0.8 and 0.9), and 15% (3/20) of the studies reached an “excellent” detecting performance (AUROC between 0.9 and 1.0) [36,41,46] according to the classification of diagnostic accuracy by Simundic [50]. For the best disease detection, multiple models were compared within the study, or the central model was compared with existing detection tools. The authors of 30% (6/20) of the studies claimed that their model produced a (slightly) higher performance than “conventional” or “traditional” models or ML models in the literature [29,34,37,38,44,45]. In 15% (3/20) of the studies, the central model performed better compared with currently used approaches such as a validated clinical model [42], a surveillance tool on which current health indexes are based [47], and a gold standard in routine clinical practice according to the American College of Cardiology and the American Heart Association [40]. In one study, the prediction scores of the model were validated by experts who agreed 100% through manual record reviewing [36]. The diagnostic accuracy of the included models was not dependent on disease categories but relied on the EHR data given to the model. Many studies (7/20, 35%) mentioned that diseases could be detected more accurately (ie, the predictive performance was increased) when the EHRs were closer to the date of diagnosis [32,33,46] and with an increase in the number of predictors [37,40,43,48]. Overall, the ability of the included models to classify nonhealthy and healthy individuals was close to the registered diagnoses in the EHRs.

Medical Insight 2: Earlier Detection

In 45% (9/20) of the studies, ML and DL models observed all available EHR data to classify patients as a case or control (ie, ML vs human detection) [30,33,34,36,38,39,42,43,45]. However, in the other studies (10/20, 50%), models were able to detect diseases earlier than the moment they were diagnosed by clinicians in EHRs (ie, prediction) [29,31,32,35,37,40,41,44,46-48]. By dividing the participants’ EHRs into 2 pieces, X years were observed (observation period), and based on these data, it was possible to predict the risk of developing a disease or medical event in the future (prediction period). In other words, the prediction was made at an earlier time (x=0) than when it was diagnosed in practice (end of black bars). In some studies (5/20, 25%), it was part of the research to identify what time frame encompasses enough predictive information and, therefore, how much earlier an (accurate) detection was possible [32,33,37,43,46]. For example, Walsh et al [46] used 2 years of EHRs and extended their prediction window more and more to find the earliest moment of an accurate prediction. Raket et al [35] predicted whether a psychosis would occur 1 year before its onset, whereas Zhao et al [40] used 7 years of EHRs to predict the occurrence of cardiovascular events in the following 10 years. Figure 2 [29-48] illustrates the different time frames of longitudinal EHRs and their results according to a possible earlier detection. How much earlier a disease can be detected has a varying clinical meaning and, therefore, needs its own interpretation.

XSL•FO
RenderX

Figure 2. Detection, observation, and prediction periods per disease. A timeline of the electronic health record (EHR) periods that were used. The moment of the prediction (via machine learning) was scaled at x=0. Bars to the left (negative numbers) represent retrospective EHRs from years in the past, and bars to the right (positive values) represent predictions into the future. C: cancer; eGFR: estimated glomerular filtration rate; H: hospitalization; M: mortality.

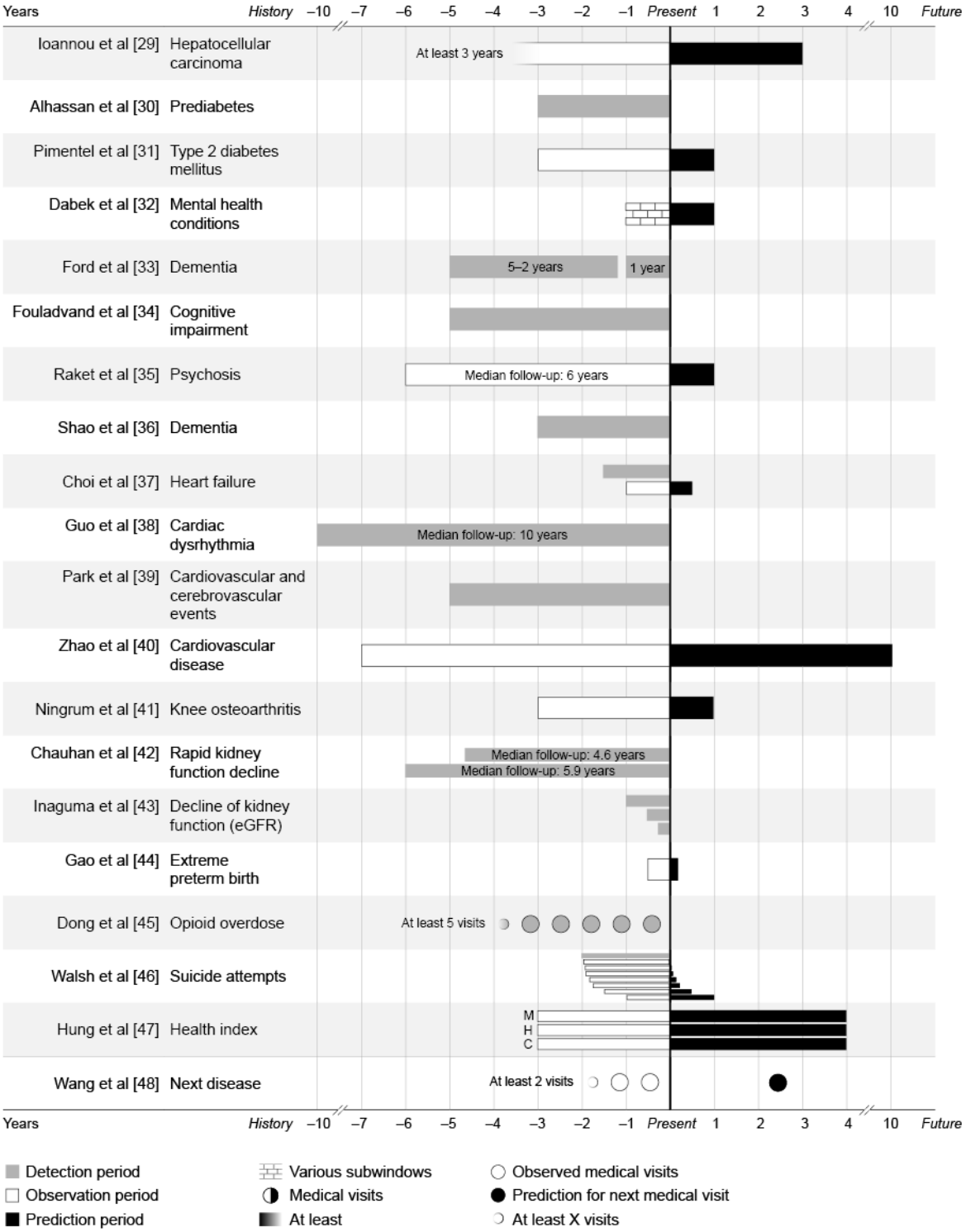

Medical Insight 3: Important Predictors

Another way to support disease detection and prevention was by generating insights into factors, topics, predictors, or indicators contributing to disease prediction [30,31,33,35-41,43-46]. In unstructured clinical notes, relevant topics, related words, and medical concepts were found that contributed to disease detection [36,44]. These words concerned daily living, behavior, and medical history. ML and DL models using structured EHRs generated the most contributing factors and their individual contribution to the outcome [30,31,33,35,37-41,43,45,46]. The most contributing predictors reported among all disease categories were (related to) age, blood pressure, BMI, cholesterol, smoking, and specific

XSL•FO
RenderX

medication. Concerning mental, behavioral, and neurodevelopmental disorders, additional predictors were related to depression, personal difficulties, and personality changes. Some of these identified predictors were new for their discipline (eg, specific medication) [35,41,44] or not yet incorporated into gold standards for clinical diagnostic guidelines (eg, genetic information) [40]. In addition to this, insights into the importance of (known) predictors were generated. For example, Raket et al [35] identified what factors were responsible for the biggest positive and negative change in risk estimation (eg, differential white blood cells) and, therefore, indicated the most effective targets for preventive interventions. Other models found that the contribution of some predictors was not as high as assumed (eg, stress on diabetes) [31]; factors that seemed individually irrelevant turned out to have cumulative important predictive value [35], and the instability of factors, not the factor itself, was a predictor for one disease [40]. The aforementioned factors were identified during model development, but applying such a model to new EHRs would generate responsible factors for that individual.

Medical Insight 4: Other Health Care Indicators

In total, 10% (2/20) of the studies used EHRs not to predict the risk of a disease but to create other health indicators. Hung et al [47] developed a health index based on 3 DL predictions of impactful and costly health indicators (mortality, hospitalization, and cancer). This health index also generated insights into the population's health and was found to be close to the "true risk" and, therefore, a better indicator than baseline models. Another study claimed to forecast what disease an individual would have at the next hospital visit [48]. Their results showed that the developed model generated well-performing results in forecasting medical diagnoses aggregated in 3- and 4-digit International Classification of Diseases, 9th and 10th Revision codes.

Clinical Benefits

Clinical Benefit 5: Preliminary Screening

In 25% (5/20) of the studies, ML models were used to support (preliminary) screening on longitudinal EHRs [29,35,36,42,46]. After developing ML and DL models, risk classes could be generated as a precursor for physical screening. Approximately 90% of the diagnosed cases were concentrated in the highest (10%) risk class. Other studies assessed the utility of ML and DL models by thresholds for the proportion needed to be screened versus the detection possibility [29,42]. For example, to detect 90% of all validated patients with hepatocellular carcinoma, the highest 66% of risk scores (predicted by a DL model) needed to be screened, whereas to detect 80% of all cases, screening from only the highest 51% of risk scores was required [29]. Chauhan et al [42] reasoned the other way around and focused on efficiency. From the 10% highest risk scores for kidney failure, the positive predictive value was 68%. Moreover, the cost benefits for screening options using DL on EHRs were investigated [35]. Disease detection using a DL model was associated with a positive net benefit–to–cost benefit ratio for a single-point risk assessment (1:3) and continuous-time risk assessment (1:16). Reasons for preliminary screening in EHRs were to prioritize those with the highest risk for disciplines with long waiting lists [29,42], before costly or more invasive examinations (eg, image or biomechanical retrieval) [35,41], or to detect cases that might be missed by the current pathway and go undetected [35,36,46].

Clinical Benefit 6: Possible Clinical Benefits

Only 10% (2/20) of the included studies were validated using an external data set, but none of the models have been implemented in clinical practice (yet). Consequently, the benefits for health were not evaluated. However, the authors interpreted their findings and suggested opportunities and possible health care benefits for clinical practice. The authors of 35% (7/20) of the studies mentioned that, if their models were applied in clinical practice, this may improve personalized health care [34-36,42,45-47]. Personalized health care was related to a personalized risk prediction, an individual-level index or output, a tailored care plan, and targeted care and screening. The authors of 60% (12/20) of the studies claimed that prevention could be improved by using their ML and DL models [31-38,42,44,45,47]. Early and timely detection and interventions before disease manifestation were often mentioned. In one case, the use of DL on EHRs could not directly prevent the targeted outcome, but by better preparing health care in an appropriate setting, indirect health outcomes could be prevented [44]. Additional suggestions to improve health care were focused on policies. It was suggested to base health policies on risk classes at a nationwide level [39,42]. Moreover, (predicted) future health conditions may be a better base for health care policies than traditional surveillance models reflecting health conditions from years before [47]. In addition to this, DL support can reduce the clinical workload. Even if the positive predictive value to select a screening population is low, a model with an excellent sensitivity can reduce the clinician's workload by 70% [44]. All studies assumed EHR data to be valuable information to improve health care. The author of one study suggested that even imperfect data can be used as a silver standard to develop risk models [36].

Discussion

Summary of Evidence

The first research question in this study sought to determine which diseases have been detected in longitudinal EHRs using ML techniques. Results showed that a variety of diseases could be detected or predicted, particularly diabetes; kidney diseases; diseases of the circulatory system; and mental, behavioral, and neurodevelopmental disorders [22]. Comparing our findings with those of prior work, only a third of EHR prediction models predict diseases; meanwhile, mortality and hospitalization remain the most prevalent outcomes [51]. Among the studies that have predicted diseases, cancer is the most frequently predicted disease based on EHRs. Another systematic review used clinical notes to identify chronic diseases [52]. It also found diseases of the circulatory system as the most prevalent and explained this by the structure of the data. Not only the structure but also the length of the EHR horizon before diagnosis may explain the diseases that can be detected or predicted. As we determined the scope of diseases that may be prevented, the length of historic data before the diagnosis (in existence of early

XSL•FO
RenderX

signs) reflects the "preventive stage" before the onset of the disease. The literature confirms that the longest EHR time horizon (8-10 years) has been found for diabetes and cardiovascular and kidney diseases [51], which were also prevalent diseases in our scoping review. In the end, the diseases that can be detected rely on available EHR data and, therefore, previous medical visits.

The second research question determined the scope of what EHR data have been used by ML techniques for the early detection and prevention of diseases. This scoping review found that age, sex, BMI, symptoms, procedures, laboratory test results, diagnoses, medications, and clinical notes are frequently used. Diseases that could be detected earlier than when they are currently diagnosed did not use other EHR variables. In addition, the most important predictors found in multiple studies were age, blood pressure, BMI, cholesterol, smoking, and medication. The consistency in the used and most important EHR variables underlines the importance of establishing generalized regulation and standardization of these variables across electronic health software, especially for variables overlapping in various health disciplines [53]. This would also address well-known challenges and limitations with EHR data, which will be discussed later in this section. According to the literature on the use of EHR data, it seems that a larger variable set improves disease prediction [51]. Their systematic review concluded that studies must leverage the full breadth of EHR data by using longitudinal data. In addition, we found that large longitudinal EHR data can successfully be analyzed via RNN and, derived from it, LSTM. These are both neural network architectures that are able to find patterns while incorporating temporality, making them effective for time-series predictions. Other types of neural networks (eg, convolutional neural networks) are well-known for their performance on images [15]. Similar results for techniques were identified in a review on the same topic from a technical perspective [2]. They concluded that RNN (specifically LSTM) was the most prominent technique to capture complex time-varying EHRs. Another review on AI techniques to facilitate earlier diagnoses of cancer also stated that neural networks were the dominant technique applied to EHRs [54]. Our results showed that there was no consistent way to process EHR variables temporally when techniques other than LSTM and RNN were used. Therefore, we can conclude that a basic RNN and LSTM are the most suitable techniques to analyze multivariable, longitudinal EHRs.

The third research question of this review was to determine the scope of medical insights that could be generated. Our results showed that, with the development and training of ML and DL models on EHRs, (1) a high diagnostic accuracy was reached, (2) the most responsible predictors could be identified, (3) diseases could be detected earlier than when they are currently diagnosed, and (4) additional health care indicators were created. The most prominent medical insight was the detection performance of the models. However, how good the performance should be is ambiguous. For example, DL models used to facilitate earlier cancer diagnoses had AUROC values ranging from 0.55 to 0.99 [54], indicating performance from almost random guessing to near-perfect detection. Looking into a more mature domain, the diagnostic accuracy of sepsis predictions ranged from between 0.68 and 0.99 in the intensive care unit to between 0.96 and 0.98 in hospital and between 0.87 and 0.97 in the emergency department [55]. This metric is ideally as high as possible because it induces a high sensitivity (true positives) and specificity (true negatives). For comparison, the diagnostic accuracy of a gut feeling (meta-analysis on cancer diagnosis) had a sensitivity of only 0.40 and a specificity of 0.85 [56]. The diagnostic accuracy of physical examination (for the detection of cirrhosis) had a sensitivity between 0.15 and 0.68 and a specificity between 0.75 and 0.98 [57]. If ML can increase both the sensitivity and specificity of disease detection, nonhealthy persons can be found, and delayed diagnoses can be reduced without overtreating healthy persons misdiagnosed as cases [58]. If the developed model is further evaluated in false-negative and false-positive groups, it may be possible that the model detects even more (true) cases than those registered by clinicians. This is already the case for many DL techniques on imaging data [59]. For now, an even more important finding is the ability of some models to detect disease manifestation earlier than the moment of diagnosis registration in EHRs. These examples of earlier detection are aligned with a study on the onset of diseases [60] that concluded that "slowly progressive diseases are often misperceived as relatively new" (ie, the onset could have been detected earlier). They found that, in 31% of diagnosed cases, the onset of their disease had started >1 year before their diagnosis. When disease predictions are early and accurate enough, it can facilitate disease prevention [23]. Especially with the addition of personally responsible factors and the biggest changers in risk prediction, prevention interventions may be more effective because they are more targeted to the individual. When medical prevention and interventions become based on the unique profile of each individual, personalized health care is delivered [61]. After all, the aforementioned medical insights only show the bright side of ML and DL models.

Our final research question sought the (possible) clinical benefits that could be obtained from using ML on EHRs. We found that preliminary screening was a clinical benefit of applying such models on longitudinal EHRs. Patients were accurately classified into risk classes to prioritize those with the highest risk, and a positive net benefit was found. In addition, the authors of the studies stated that their results (although they were not clinically evaluated) may contribute to a more personalized health care, prevention possibilities, and health care policies and reduce the clinicians' workload. These benefits are perfectly aligned with the near-future vision, strategies, and action foci set by the World Health Organization [62,63]. In particular, the emerging clinical staff shortage makes the future health care system more dependent on technical innovations and the health care system will be forced to be digitally assisted [64]. However, to be adopted in medical practice, ML and DL models require external validation, the absence of bias and drift, and transparency for clinicians. In prior work, benefits have rarely been clinically evaluated either. Even in a more mature health domain regarding ML, the intensive care unit, only 2% of the AI applications are clinically evaluated [65]. In their systematic review, the clinical readiness of AI was explored, but no AI model was found to be integrated into routine clinical practice at the time of writing. The limited amount of publications evaluating the clinical

XSL•FO
RenderX

benefits of the application of ML on EHRs indicates the research gap in the literature. Future studies should explore the follow-up of these AI attempts and the reasons for success or failure in practice.

Up until now, we have only discussed possible beneficial results of using ML and DL on EHRs. However, we cannot ignore the possible risks, obstacles, challenges, or issues. Multiple (systematic) reviews have summarized these well-known issues, challenges, and limitations regarding the application of ML and DL on EHRs [2,51,66,67]. Viewed generally across all studies, practical obstacles influence the scientific and clinical implementation process: ethical considerations, privacy guidelines, legal procedures, equity, and data protection and security [68]. Beyond these obstacles, existing predictions face limitations due to their reliance on the data. First, key issues of using EHRs are irregularity, heterogeneity, sparsity (eg, missing data), temporality, the lack of gold-standard labels, and the volume and quality of data [2,51,66,67]. Second, ML and DL models have limited transparency and interpretability, face domain complexity (vs engineering expertise), may include biases, and often lack external validation. It is not possible to assign specific issues to specific studies; they all suffer more or less from the aforementioned issues. Our point is to become aware of the downside as well. Therefore, all our principal findings must be interpreted with this last discussion point in mind. In our opinion, a consistent, reliable, and valid way of EHR registration will improve the (use of) data and could be the first step toward a data-based health care system. This need for movement and improvement is important not only for research but also for practical convenience for clinicians and, consequently, to succeed in improving health outcomes.

Limitations

A limitation of this scoping review is the time between the search and the publication. As ML and DL have become a popular topic and the amount of research has grown drastically over the last years, new research could have been published between the literature search and the publishing of this scoping review. Consequently, some of our findings may have been overtaken by the progress in research.

Another limitation was the data synthesis regarding the performance outcomes per technique. Due to a wide variety of internal analyses, outcomes were not directly comparable, and therefore, the data extraction and data synthesis were difficult. Some studies just noted the optimal performance value achieved by the central model, while other studies compared a variety of techniques and noted various performance values for different subgroups, different metrics, and different time windows and with the addition of various technical improvements. A few authors discussed their ultimate results and mentioned that their model was better than literature, that is, “traditional” or “conventional,” attempts, which were not always clearly defined. We have attempted to follow the authors’ description to avoid incorrect comparisons. However, some comparisons may have become vague or skewed during data synthesis. Nevertheless, we scoped the optimal AUROC for each study at the meta level.

As we used a broad definition of EHR, we included a greater range of data. This means that the results are not based solely on data directly extracted from clinical record systems but also on data extracted by an intermediate organization, such as insurance companies. Therefore, readers must interpret the results of ML and DL models with this in mind.

Conclusions

Longitudinal EHRs have valuable potential to support the early detection of a variety of diseases. For various diseases, EHR data concerning diagnoses, procedures, vital signs, medication, laboratory tests, BMI, and (early) symptoms have a high predictive value. To analyze multivariable, longitudinal EHRs, a basic RNN and LSTM are the most suitable techniques. For the detection of diseases, using ML (including DL) on EHRs proved to be highly accurate. When the detection occurs at the same moment as the diagnosis of clinicians, it seems not directly relevant for the prevention of diseases. However, the detection of diseases offers the clinical benefits of preliminary screening to prioritize patients from the highest risk class. The prevention of diseases can be supported by ML models that are able to predict or detect diseases earlier than the current clinical practice. The additional information about the most important predictors of the individual and the biggest risk changers allow targeted prevention interventions and, therefore, personalized care. Improved health care policies and workload reduction are frequently cited benefits but have not yet been evaluated in clinical practice. Both ML and DL attempts for disease detection and prevention still remain in the testing and prototyping phase and have a long way to go to be clinically applied.

Acknowledgments

The first author conducted this study as part of her PhD trajectory. Her PhD trajectory was funded by the Centre of Expertise Prevention in Care and Wellbeing from Inholland University of Applied Sciences. JS acknowledges financial support from Regieorgaan SIA RAAK, part of the Netherlands Organisation for Scientific Research (grant HBOPD.2018.05.016). The remaining authors declare no other external sources of funding for this scoping review.

Authors' Contributions

All the authors made substantial contributions to the conception and design, acquisition of data, or analysis and interpretation of data. LS and FCB screened, extracted, analyzed, and interpreted the data. KAZ designed the search strategy and ran, exported, and deduplicated the search results. All authors revised the paper critically and have granted final approval for the version to be published.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File (Adobe PDF File), 96 KB - jmir_v26i1e48320_app1.pdf]

Multimedia Appendix 2
Data extraction instrument.
[PDF File (Adobe PDF File), 36 KB - jmir_v26i1e48320_app2.pdf]

Multimedia Appendix 3
Electronic health record data and applied techniques.
[PDF File (Adobe PDF File), 143 KB - jmir_v26i1e48320_app3.pdf]

Multimedia Appendix 4
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 498 KB - jmir_v26i1e48320_app4.pdf]

References

1. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
2. Xie F, Yuan H, Ning Y, Ong ME, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inform 2022 Mar;126:103980 [FREE Full text] [doi: 10.1016/j.jbi.2021.103980] [Medline: 34974189]
3. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med 2013 Sep;28 Suppl 3(Suppl 3):S660-S665 [FREE Full text] [doi: 10.1007/s11606-013-2455-8] [Medline: 23797912]
4. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
5. Beath C, Becerra-Fernandez I, Ross J, Short J. Finding value in the information explosion. MIT Sloan Manag Rev 2012;53:18-20.
6. de Ruiter HP, Liaschenko J, Angus J. Problems with the electronic health record. Nurs Philos 2016 Jan;17(1):49-58. [doi: 10.1111/nup.12112] [Medline: 26603947]
7. Nilsson G, Ahlfeldt H, Strender LE. Textual content, health problems and diagnostic codes in electronic patient records in general practice. Scand J Prim Health Care 2003 Mar;21(1):33-36. [doi: 10.1080/02813430310000537] [Medline: 12718458]
8. Norgeot B, Glicksberg BS, Trupin L, Lituiev D, Gianfrancesco M, Oskotsky B, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. JAMA Netw Open 2019 Mar 01;2(3):e190606 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0606] [Medline: 30874779]
9. Howe JL, Adams KT, Hettinger AZ, Ratwani RM. Electronic health record usability issues and potential contribution to patient harm. JAMA 2018 Mar 27;319(12):1276-1278 [FREE Full text] [doi: 10.1001/jama.2018.1171] [Medline: 29584833]
10. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
11. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
12. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013 Mar 25;173(6):418-425 [FREE Full text] [doi: 10.1001/jamainternmed.2013.2777] [Medline: 23440149]
13. Farhadian M, Shokouhi P, Torkzaban P. A decision support system based on support vector machine for diagnosis of periodontal disease. BMC Res Notes 2020 Jul 13;13(1):337 [FREE Full text] [doi: 10.1186/s13104-020-05180-5] [Medline: 32660549]
14. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Mark 2021 Apr 08;31(3):685-695. [doi: 10.1007/s12525-021-00475-2]

XSL•FO
RenderX

15. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
16. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505-515 [FREE Full text] [doi: 10.1148/rg.2017160130] [Medline: 28212054]
17. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
18. Syed M, Syed S, Sexton K, Syeda HB, Garza M, Zozus M, et al. Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: systematic review. Informatics (MDPI) 2021 Mar;8(1):16 [FREE Full text] [doi: 10.3390/informatics8010016] [Medline: 33981592]
19. Li Q, Campan A, Ren A, Eid WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. Int J Med Inform 2022 Jul;163:104786 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104786] [Medline: 35512622]
20. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J Biomed Inform 2020 Jan;101:103337 [FREE Full text] [doi: 10.1016/j.jbi.2019.103337] [Medline: 31916973]
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak 2021 Nov 09;21(Suppl 6):206 [FREE Full text] [doi: 10.1186/s12911-021-01534-6] [Medline: 34753471]
23. Kisling LA, Das JM. Prevention strategies. In: StatPearls. Treasure Island, FL: StatPearls Publishing LLC; Aug 01, 2023.
24. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. Stat Med 2019 Mar 30;38(7):1276-1296 [FREE Full text] [doi: 10.1002/sim.7992] [Medline: 30357870]
25. Hair K, Bahor Z, Macleod M, Liao J, Sena E. The Automated systematic search deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. BMC Biol 2023 Sep 07;21(1):189 [FREE Full text] [doi: 10.1186/s12915-023-01686-z] [Medline: 37674179]
26. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
27. Lockwood C, Dos Santos KB, Pap R. Practical guidance for knowledge synthesis: scoping review methods. Asian Nurs Res (Korean Soc Nurs Sci) 2019 Dec;13(5):287-294 [FREE Full text] [doi: 10.1016/j.anr.2019.11.002] [Medline: 31756513]
28. Hendricks L, Eshun-Wilson I, Rohwer A. A mega-aggregation framework synthesis of the barriers and facilitators to linkage, adherence to ART and retention in care among people living with HIV. Syst Rev 2021 Mar 11;10(1):54 [FREE Full text] [doi: 10.1186/s13643-021-01582-z] [Medline: 33568216]
29. Ioannou GN, Tang W, Beste LA, Tincopa MA, Su GL, Van T, et al. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. JAMA Netw Open 2020 Sep 01;3(9):e2015626 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15626] [Medline: 32870314]
30. Alhassan Z, Watson M, Budgen D, Alshammari R, Alessa A, Al Moubayed N. Improving current glycated hemoglobin prediction in adults: use of machine learning algorithms with electronic health records. JMIR Med Inform 2021 May 24;9(5):e25237 [FREE Full text] [doi: 10.2196/25237] [Medline: 34028357]
31. Pimentel A, Carreiro AV, Ribeiro RT, Gamboa H. Screening diabetes mellitus 2 based on electronic health records using temporal features. Health Informatics J 2018 Jun;24(2):194-205 [FREE Full text] [doi: 10.1177/1460458216663023] [Medline: 27566751]
32. Dabek F, Hoover P, Jorgensen-Wagers K, Wu T, Caban JJ. Evaluation of machine learning techniques to predict the likelihood of mental health conditions following a first mTBI. Front Neurol 2021;12:769819 [FREE Full text] [doi: 10.3389/fneur.2021.769819] [Medline: 35185749]
33. Ford E, Rooney P, Oliver S, Hoile R, Hurley P, Banerjee S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. BMC Med Inform Decis Mak 2019 Dec 02;19(1):248 [FREE Full text] [doi: 10.1186/s12911-019-0991-9] [Medline: 31791325]
34. Fouladvand S, Mielke MM, Vassilaki M, Sauver JS, Petersen RC, Sohn S. Deep learning prediction of mild cognitive impairment using electronic health records. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2019 Nov;2019:799-806 [FREE Full text] [doi: 10.1109/bibm47256.2019.8982955] [Medline: 33194303]
35. Raket LL, Jaskolowski J, Kinon BJ, Brasen JC, Jönsson L, Wehnert A, et al. Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. Lancet Digit Health 2020 May;2(5):e229-e239 [FREE Full text] [doi: 10.1016/S2589-7500(20)30024-8] [Medline: 33328055]

XSL•FO
RenderX

36. Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. BMC Med Inform Decis Mak 2019 Jul 09;19(1):128 [FREE Full text] [doi: 10.1186/s12911-019-0846-4] [Medline: 31288818]
37. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017 Mar 01;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]
38. Guo A, Smith S, Khan YM, Langabeer Ii JR, Foraker RE. Application of a time-series deep learning model to predict cardiac dysrhythmias in electronic health records. PLoS One 2021;16(9):e0239007 [FREE Full text] [doi: 10.1371/journal.pone.0239007] [Medline: 34516567]
39. Park J, Kim JW, Ryu B, Heo E, Jung SY, Yoo S. Patient-level prediction of cardio-cerebrovascular events in hypertension using nationwide claims data. J Med Internet Res 2019 Mar 15;21(2):e11757 [FREE Full text] [doi: 10.2196/11757] [Medline: 30767907]
40. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Sci Rep 2019 Jan 24;9(1):717 [FREE Full text] [doi: 10.1038/s41598-018-36745-x] [Medline: 30679510]
41. Ningrum DN, Kung WM, Tzeng IS, Yuan SP, Wu CC, Huang CY, et al. A deep learning model to predict knee osteoarthritis based on nonimage longitudinal medical record. J Multidiscip Healthc 2021;14:2477-2485 [FREE Full text] [doi: 10.2147/JMDH.S325179] [Medline: 34539180]
42. Chauhan K, Nadkarni GN, Fleming F, McCullough J, He CJ, Quackenbush J, et al. Initial validation of a machine learning-derived prognostic test (KidneyIntelX) integrating biomarkers and electronic health record data to predict longitudinal kidney outcomes. Kidney360 2020 Aug 27;1(8):731-739 [FREE Full text] [doi: 10.34067/KID.0002252020] [Medline: 35372952]
43. Inaguma D, Kitagawa A, Yanagiya R, Koseki A, Iwamori T, Kudo M, et al. Increasing tendency of urine protein is a risk factor for rapid eGFR decline in patients with CKD: a machine learning-based prediction model by using a big database. PLoS One 2020;15(9):e0239262 [FREE Full text] [doi: 10.1371/journal.pone.0239262] [Medline: 32941535]
44. Gao C, Osmundson S, Velez Edwards DR, Jackson GP, Malin BA, Chen Y. Deep learning predicts extreme preterm birth from electronic health records. J Biomed Inform 2019 Dec;100:103334 [FREE Full text] [doi: 10.1016/j.jbi.2019.103334] [Medline: 31678588]
45. Dong X, Deng J, Hou W, Rashidian S, Rosenthal RN, Saltz M, et al. Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning. J Biomed Inform 2021 Apr;116:103725 [FREE Full text] [doi: 10.1016/j.jbi.2021.103725] [Medline: 33711546]
46. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J Child Psychol Psychiatry 2018 Dec;59(12):1261-1270. [doi: 10.1111/jcpp.12916] [Medline: 29709069]
47. Hung C, Chen H, Wee LJ, Lin CH, Lee CC. Deriving a novel health index using a large-scale population based electronic health record with deep networks. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:5872-5875. [doi: 10.1109/EMBC44109.2020.9176454] [Medline: 33019310]
48. Wang T, Tian Y, Qiu RG. Long short-term memory recurrent neural networks for multiple diseases risk prediction by leveraging longitudinal medical records. IEEE J Biomed Health Inform 2020 Aug;24(8):2337-2346. [doi: 10.1109/JBHI.2019.2962366] [Medline: 31880573]
49. Wang L, Sha L, Lakin JR, Bynum J, Bates DW, Hong P, et al. Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. JAMA Netw Open 2019 Jul 03;2(7):e196972 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.6972] [Medline: 31298717]
50. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC 2009 Jan;19(4):203-211 [FREE Full text] [Medline: 27683318]
51. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
52. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
53. Vale MD, Perkins DW. Discuss and remember: clinician strategies for integrating social determinants of health in patient records and care. Soc Sci Med 2022 Dec;315:115548. [doi: 10.1016/j.socscimed.2022.115548] [Medline: 36403352]
54. Jones OT, Calanzani N, Saji S, Duffy SW, Emery J, Hamilton W, et al. Artificial intelligence techniques that may be applied to primary care data to facilitate earlier diagnosis of cancer: systematic review. J Med Internet Res 2021 Mar 03;23(3):e23483 [FREE Full text] [doi: 10.2196/23483] [Medline: 33656443]
55. Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020 Mar;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
56. Yao M, Kaneko M, Watson J, Irving G. Gut feeling for the diagnosis of cancer in general practice: a diagnostic accuracy review. BMJ Open 2023 Aug 11;13(8):e068549 [FREE Full text] [doi: 10.1136/bmjopen-2022-068549] [Medline: 37567752]

XSL•FO
RenderX

57. de Bruyn G, Graviss EA. A systematic review of the diagnostic accuracy of physical examination for the detection of cirrhosis. BMC Med Inform Decis Mak 2001;1:6 [FREE Full text] [doi: 10.1186/1472-6947-1-6] [Medline: 11806763]
58. Sørensen J, Hetland ML. Decreases in diagnostic delay are supported by sensitivity analyses. Ann Rheum Dis 2014 Jul;73(7):e45. [doi: 10.1136/annrheumdis-2014-205590] [Medline: 24739326]
59. Killock D. AI outperforms radiologists in mammographic screening. Nat Rev Clin Oncol 2020 Mar;17(3):134. [doi: 10.1038/s41571-020-0329-7] [Medline: 31965085]
60. van Hoorn BT, Wilkens SC, Ring D. Gradual onset diseases: misperception of disease onset. J Hand Surg Am 2017 Dec;42(12):971-7.e1. [doi: 10.1016/j.jhsa.2017.07.021] [Medline: 28899587]
61. McEwen BS, Getz L. Lifetime experiences, the brain and personalized medicine: an integrative perspective. Metabolism 2013 Jan;62 Suppl 1:S20-S26. [doi: 10.1016/j.metabol.2012.08.020] [Medline: 23009787]
62. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-04-29]
63. Health and care workforce in Europe: time to act. World Health Organization. URL: https://www.who.int/europe/publications/i/item/9789289058339 [accessed 2024-04-29]
64. Liu JX, Goryakin Y, Maeda A, Bruckner T, Scheffler R. Global health workforce labor market projections for 2030. Hum Resour Health 2017 Mar 03;15(1):11 [FREE Full text] [doi: 10.1186/s12960-017-0187-2] [Medline: 28159017]
65. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021 Jul;47(7):750-760 [FREE Full text] [doi: 10.1007/s00134-021-06446-7] [Medline: 34089064]
66. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
67. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
68. Cordeiro JV. Digital technologies and data science as health enablers: an outline of appealing promises and compelling ethical, legal, and social challenges. Front Med (Lausanne) 2021;8:647897 [FREE Full text] [doi: 10.3389/fmed.2021.647897] [Medline: 34307394]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
DL: deep learning
EHR: electronic health record
LSTM: long short-term memory
ML: machine learning
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RNN: recurrent neural network

Edited by T de Azevedo Cardoso, S Ma; submitted 19.04.23; peer-reviewed by J Zeng, V Rajan, D Chrimes; comments to author 11.07.23; revised version received 29.09.23; accepted 29.04.24; published 20.08.24.

Please cite as:
Swinckels L, Bennis FC, Ziesemer KA, Scheerman JFM, Bijwaard H, de Keijzer A, Bruers JJ
The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review
J Med Internet Res 2024;26:e48320
URL: https://www.jmir.org/2024/1/e48320
doi:10.2196/48320
PMID:39163096

©Laura Swinckels, Frank C Bennis, Kirsten A Ziesemer, Janneke F M Scheerman, Harmen Bijwaard, Ander de Keijzer, Josef Jan Bruers. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

A Comprehensive Analysis of COVID-19 Misinformation, Public Health Impacts, and Communication Strategies: Scoping Review

Sezer Kisa[1*], PhD; Adnan Kisa[2,3*], PhD

[1]Department of Nursing and Health Promotion, Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway

[2]Department of Health and Exercise, School of Health Sciences, Kristiania University College, Oslo, Norway

[3]Department of International Health and Sustainable Development, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

[*]all authors contributed equally

Corresponding Author:
Sezer Kisa, PhD
Department of Nursing and Health Promotion
Faculty of Health Sciences
Oslo Metropolitan University
St. Olavs plass
Oslo, 0130
Norway
Phone: 47 92501403
Email: sezerkisa@hotmail.com

Abstract

Background: The COVID-19 pandemic was marked by an *infodemic*, characterized by the rapid spread of both accurate and false information, which significantly affected public health. This infodemic led to confusion, mistrust in health authorities, noncompliance with health guidelines, and engagement in risky health behaviors. Understanding the dynamics of misinformation during the pandemic is crucial for developing effective public health communication strategies.

Objective: This comprehensive analysis aimed to examine the complexities of COVID-19 misinformation. Specifically, it sought to identify the sources and themes of misinformation, the target audiences most affected, and the effectiveness of various public health communication strategies in mitigating misinformation.

Methods: This scoping review used the MEDLINE (PubMed), Embase, and Scopus databases to identify relevant studies. An established, methodical framework for scoping reviews was used to review literature published between December 2019 and September 2023. The inclusion criteria focused on peer-reviewed studies published in English that address COVID-19 misinformation and its sources, themes, and target audiences, as well as the effectiveness of public health communication strategies.

Results: The scoping review identified that misinformation significantly impacted mental health, vaccine hesitancy, and health care decision-making. Social media and traditional media were major conduits for spreading misinformation. Key misinformation themes included the origins of the virus, ineffective treatments, and misunderstandings about public health measures. Misinformation sources ranged from social media platforms to traditional media outlets and informal networks. The impact of misinformation was found to vary across different regions and demographic groups, with vulnerable populations being disproportionately affected. Effective strategies to counter misinformation included enhancing health literacy; using digital technology; promoting clear, authoritative communication; and implementing fact-checking mechanisms. In addition, community engagement and targeted health campaigns played a crucial role in addressing misinformation.

Conclusions: The review emphasizes the critical need for accurate and consistent messaging to combat misinformation. Cooperative efforts among policy makers, health professionals, and communication experts are essential for developing effective interventions. Addressing the infodemic is vital for building a well-informed, health-literate society capable of handling misinformation in future global health crises. The study provides valuable insights into the dynamics of misinformation and highlights the importance of robust public health communication strategies. These findings can guide future efforts to mitigate the impact of misinformation during health emergencies.

(J Med Internet Res 2024;26:e56931) doi:10.2196/56931

XSL•FO
RenderX

KEYWORDS

communication strategies; COVID-19; infodemic; misinformation; public health

Introduction

Background

The COVID-19 pandemic, a health crisis of unprecedented scale in the 21st century, was accompanied by an equally significant and dangerous phenomenon—an *infodemic*. The World Health Organization defines an infodemic as the rapid spread and overabundance of information—both accurate and false—that occurs during an epidemic [1]. A tidal wave of misinformation, disinformation, and rumors characterized the infodemic during the COVID-19 pandemic. This led to widespread confusion, mistrust in health authorities, noncompliance with health guidelines, and even risky health behaviors [2-4].

Moreover, the role of political leaders in shaping the narrative around COVID-19 policies significantly influenced these dynamics. In countries such as the United States, Brazil, and Turkey, the intersection of political ideology and crisis management led to increased societal polarization. Leaders in these nations used communication strategies ranging from denying the severity of the pandemic to promoting unproven treatments [5,6]. This complex interplay between leadership communication and public response underscores the critical need for science-based policy communication and the responsible use of social media platforms to combat misinformation and foster societal unity in the face of a global health crisis.

Furthermore, the emergence of the COVID-19 infodemic highlighted the crucial role of social media literacy in combating misinformation. Educating the public on discerning credible information on the web has emerged as a pivotal strategy for mitigating the spread of misinformation and its consequences [7].

Misinformation during public health crises has been a recurring problem. Historical examples from the Ebola outbreak, such as rumors that the virus was a government creation or that certain local practices could cure the disease, highlight how misinformation can hinder public health responses [8]. False beliefs, such as that drinking salt water would cure Ebola or that the disease was spread through the air, led to a mistrust of health workers and avoidance of treatment centers, exacerbating the crisis [9]. In the context of COVID-19, misinformation was particularly pervasive, with false claims about the effectiveness of various nostrums, leading to panic buying and shortages [3,10]. The impact of such misinformation varied across regions [4,11]. These dynamics were often fueled by psychological and social factors, including fear, uncertainty, and the reinforcing nature of social media algorithms, which created echo chambers of false information [12,13]. The wide-ranging consequences affected not only immediate health behaviors but also the trust in, and response to, public health authorities [2,14].

Misinformation during a public health crisis is nothing new. However, the scale and speed at which misinformation spread during the COVID-19 pandemic are unparalleled. This situation was exacerbated by the widespread use of social media and the internet, where rumors can rapidly reach large audiences [9,15]. This spread of misinformation had far-reaching consequences: it undermined public health efforts, promoted harmful practices, contributed to vaccine hesitancy, and possibly prolonged the pandemic [8,10,12,16]. These effects went beyond individual health behaviors; they influenced public health policies and diminished trust in health authorities and the scientific community [12,17,18].

In light of these challenges, the machine learning–enhanced graph analytics (MEGA) framework has emerged as a novel approach to managing infodemics by leveraging the power of machine learning and graph analytics. This framework offers a robust method for detecting spambots and influential spreaders in social media networks, which is crucial for assessing and mitigating the risks associated with infodemics. Such advanced tools are essential for public health officials and policy makers to navigate the complex landscape of misinformation and to develop more effective communication strategies [19]. Furthermore, combating this infodemic necessitates a strategic approach encapsulating the "Four Pillars of Infodemic Management": (1) monitoring information (infoveillance) to track the spread and impact of misinformation; (2) enhancing eHealth literacy and science literacy, empowering individuals to evaluate information critically; (3) refining knowledge quality through processes such as fact checking and peer review, ensuring the reliability of information; and (4) ensuring timely and accurate knowledge translation, minimizing the distortion by political or commercial interests [20]. These measures are essential for mitigating the impact of misinformation and guiding the public and professionals toward quality health information during the pandemic and beyond. The COVID-19 pandemic has highlighted the need for improved public health communication and preparedness strategies, particularly in countering misinformation to prevent similar challenges in future health crises [15,21].

Pertinent Questions

Recognizing the unique challenges posed by the COVID-19 infodemic, this comprehensive scoping review seeks to systematically explore various dimensions of misinformation related to the pandemic. Our investigation is informed by a critical analysis of existing literature, noting a gap in studies that collectively examine the themes, sources, target audiences, impacts, interventions, and effectiveness of public health communication strategies against COVID-19 misinformation. To the best of our knowledge, this is the first review that attempts to bridge this gap by providing a comprehensive and integrated analysis of these key dimensions. While individual aspects of misinformation have been addressed in prior research, there lacks a comprehensive review that integrates these components to offer a holistic understanding necessary for effective countermeasures. Therefore, our review is structured around four pertinent questions, each carefully selected for their significance in advancing our understanding of the COVID-19 infodemic and its counteraction:

XSL•FO
RenderX

1. What is the extent of COVID-19 misinformation? How can it be addressed?
2. What are the primary sources of COVID-19 misinformation?
3. Which target audiences are most affected by COVID-19 misinformation?
4. What public health communication strategies are being used to combat COVID-19 misinformation?

These questions were selected to emphasize critical areas of COVID-19 misinformation that, when addressed, can significantly contribute to bridging technical and knowledge gaps in our response to current and future public health emergencies. By detailing our study's contributions to existing literature, we aim to present distinctive understandings crucial for policy makers, health professionals, and the public in effectively addressing misinformation challenges.

Methods

This scoping review was conducted following the methodology framework defined by Arksey and O'Malley [16] and elaborated upon by Levac et al [17]. This framework, recognized for its systematic approach, involves five stages: (1) defining the research question; (2) identifying relevant studies; (3) selecting appropriate literature; (4) charting the data; and (5) collating, summarizing, and reporting the results.

Databases and Search Strategies

The literature search targeted 3 major databases: MEDLINE (PubMed), Embase, and Scopus. These databases were selected for their comprehensive coverage of medical, health, and social science literature. The search strategy was crafted using a combination of keywords and subject headings related to COVID-19, misinformation, and public health communication. We used ("COVID-19" OR "SARS-CoV-2" OR "Coronavirus") AND ("Misinformation" OR "Disinformation" OR "Fake news" OR "Infodemic") AND ("Public health outcomes" OR "Health impacts") AND ("Communication strategies" OR "Public health communication").

Eligibility Criteria

The inclusion and exclusion criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Article type: peer-reviewed studies
- Language: published in English
- Publication date: published between December 1, 2019, and September 30, 2023
- Focus: addresses COVID-19 misinformation and its sources, themes, and target audiences, as well as the effectiveness of public health communication strategies
- Study design: empirical studies (eg, cross-sectional, observational, randomized controlled trials, qualitative, and mixed methods)

Exclusion criteria

- Article type: non–peer-reviewed articles, opinion pieces, and editorials
- Language: published in languages other than English
- Publication date: published before December 1, 2019, or after September 30, 2023
- Focus: does not address COVID-19 misinformation or its related aspects
- Study design: case studies and anecdotal reports

Study Selection Process

The study selection process involved an initial screening of titles and abstracts to eliminate irrelevant studies, followed by a thorough full-text review of the remaining articles. This critical stage was conducted by the authors, each with expertise in public health communication and health services research, thereby enhancing the thoroughness and reliability of the selection process. In cases of disagreement, the reviewers engaged in discussions until a consensus was reached on the inclusion of each article. In addition, we adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [18] to enhance the thoroughness and transparency of our review (see Multimedia Appendix 1 for the PRISMA-ScR checklist).

Results

Overview

A total of 390 articles were identified from the 3 databases, of which, after removing 134 (34.4%) duplicates, 256 (65.6%) articles remained. Of these 256 articles, 69 (27%) were selected based on abstract searches. Of the 69 full-text articles, 27 (39%) were assessed for eligibility. Of these 27 studies, 21 (78%) were included in the scoping review (Figure 1). This analysis of the 21 studies provides a comprehensive overview of the many impacts of misinformation during the COVID-19 pandemic, including its characteristics, themes, sources, effects, and public health communication strategies.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the search and screening results.

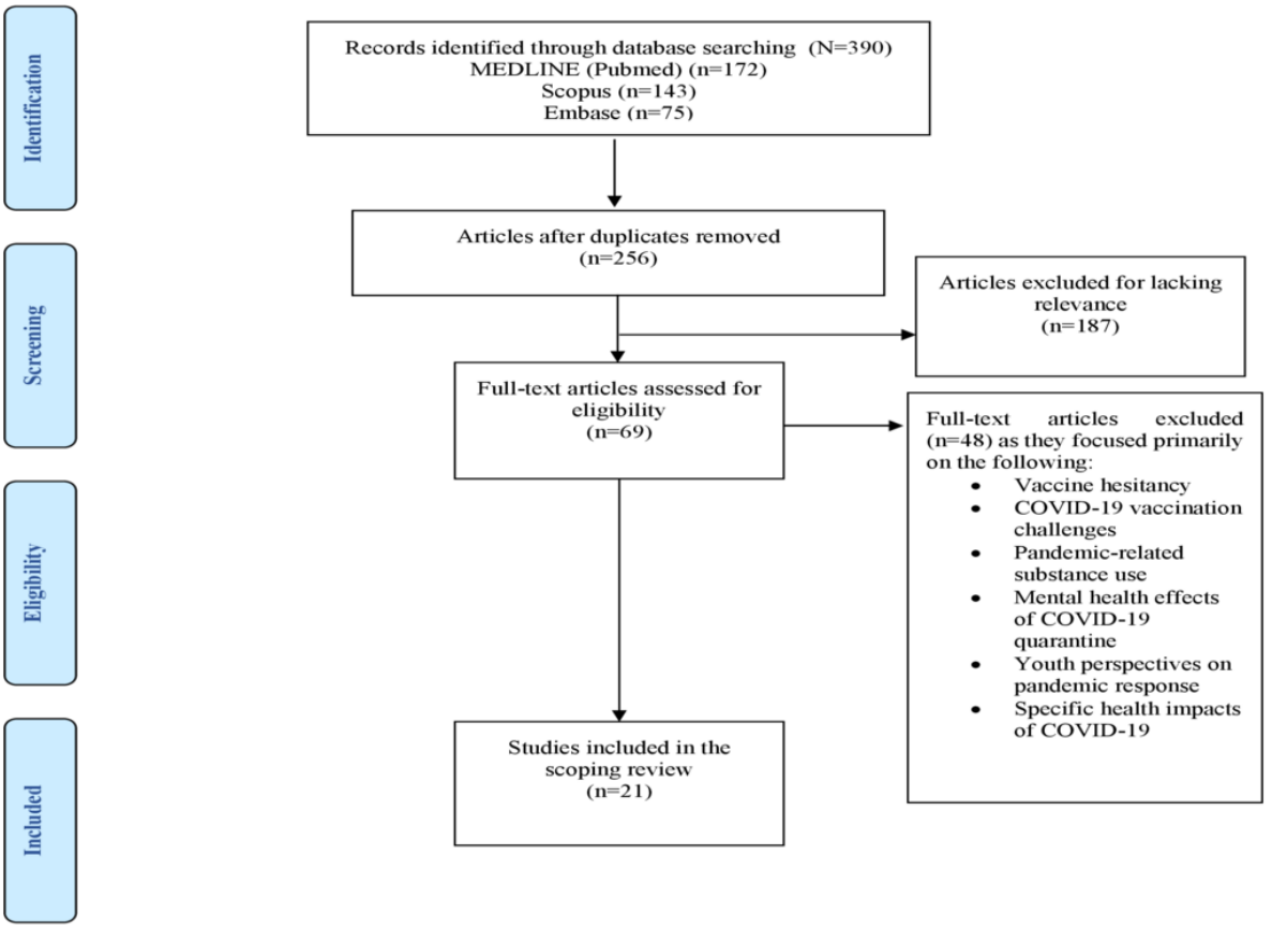

Study Characteristics

The included studies exhibited considerable diversity in terms of their methodologies, geographic focus, and objectives (Table 1). Verma et al [15] conducted a large-scale observational study in the United States, analyzing social media data from >76,000 users of Twitter (subsequently rebranded X) to establish a causal link between misinformation sharing and increased anxiety. By contrast, Loomba et al [11] carried out a randomized controlled trial in both the United Kingdom and the United States to examine the impact of misinformation on COVID-19 vaccination intent across different sociodemographic groups. In the United States, Bokemper et al [22] used randomized trials to assess the efficacy of various public health messages in promoting social distancing. Xue et al [23] used observational methods to explore public attitudes toward COVID-19 vaccines and the role of fact-checking information on social media. These studies collectively used quantitative analysis, web-based surveys, cross-sectional studies, and social network analysis, reflecting the diversity of research approaches. Sample sizes ranged from hundreds to tens of thousands of participants, providing a broad view of the infodemic's impact. Notably, most of the studies (17/21, 81%) were conducted on the web, underlining the infodemic's digital nature. The outcomes assessed various public health aspects, including mental health, communication effectiveness, and behavior change. Kumar et al [13] used social network and topic modeling analyses to gain insights into public perceptions on Reddit, contributing to the methodological diversity within the reviewed literature.

XSL•FO
RenderX

Table 1. General characteristics of the included studies, misinformation themes, sources of misinformation, and target audience.

Study, year; country; method	Aim	Sample, n	Study outcomes	Misinformation themes	Sources of misinformation	Target audience
Datta et al [24], 2020; India; cross-sectional study	Explore COVID-19 information sources among health care professionals	778 adults	High misinformation prevalence, mainly from social media	Incorrect COVID-19 treatments, false diagnoses, virus transmission misconceptions, public health inaccuracies	Social media, family, WhatsApp, television, friends	Health care professionals in India, implications for the general public
Moscadelli et al [3], 2020; Italy; observational study	Measure circulation of fake and verified COVID-19 news	2102 articles shared on social media platforms	Fake news shared 2.35 million times (23.1% of total shares)	Accidental laboratory release, SARS-CoV-2 origin, bioweapon claims, HIV vaccine, vitamin C and D, 5G technology, consuming garlic	Internet articles, social media platforms	General public, especially susceptible individuals
Hou et al [10], 2020; 12 countries; cross-sectional study	Assess public awareness and behavioral responses to COVID-19	Global public responses to COVID-19 from platforms such as Google Trends and Baidu Index	Public response trends, impact of rumors, communication effectiveness	Traditional Chinese medicines, garlic for prevention, antimalarial treatments	Internet rumors, newspapers, political leaders	Global general public; focus on China, the United States, and traditional medicine regions
Agley et al [14], 2021; United States; randomized controlled trial	Examine effects of exposure to infographic on trust in science and COVID-19 misinformation beliefs	1017 adults	Small trust increase, indirect effect on misinformation mediated by trust	5G transmission, laboratory origin, exaggerated deaths, face mask misinformation	Social media, news outlets, other media	American general public
Teovanović et al [25], 2020; Serbia; cross-sectional survey	Explore irrational beliefs and health behaviors during the COVID-19 pandemic	407 participants	Irrational beliefs predict guideline adherence, pseudoscientific practices, and vaccination intentions	Conspiracy theories, pseudoscientific practices, COVID-19 myths, vaccine-related myths	Digital media outlets	Serbian adults, broader audience susceptible to misinformation
Nowak et al [4], 2021; Poland; cross-sectional survey	Assess COVID-19 preventive behaviors, fears, and conspiracy beliefs	1380 adults	Challenges in adherence, influence of misinformation, fear impact	5G technology as pandemic cause, Chinese government conspiracy, pharmaceutical companies' profit motives	Web-based social media, message boards	General public, conspiracy theory believers
Loomba et al [11], 2021; United Kingdom and United States; randomized controlled trial	Measure COVID-19 vaccine misinformation impact on intent	8001 adults	Change in intent due to misinformation	Vaccine importance and safety, 5G links, vaccine trial deaths, pandemic conspiracy theories	Social media, various web-based sources	British and American general public
Scholz et al [21], 2021; Germany; cross-sectional survey	Investigate implementation of quarantine measures	562 adults	Improved understanding and community acceptance of quarantine measures	Quarantine measures, health information	Informal channels, possibly social media	General public, adults not frequently engaging with official channels
Ghaddar et al [2], 2022; Lebanon; cross-sectional study	Explore trust in social media, misinformation, and vaccination intent	2653 adults	Vaccination intent, fake news exposure, trust, and conspiracy beliefs	COVID-19 transmission modes, conspiracy theories, medication effectiveness	WhatsApp, Facebook, television and radio, social media platforms	General population of Lebanon
Kim et al [26], 2022; United States; cross-sectional and observational study	Analyze predictors of belief in COVID-19 misinformation on Facebook	6518 adults	Predictors of belief in misinformation, effects on behaviors, correction strategies	Transmission modes, miracle cures, antivaccine beliefs, political conspiracies	User-generated content on Facebook	General public, especially those exposed to misinformation

XSL•FO
RenderX

Study, year; country; method	Aim	Sample, n	Study outcomes	Misinformation themes	Sources of misinformation	Target audience
Huang et al [27], 2022; China; cross-sectional study	Investigate COVID-19 vaccine hesitancy predictors	4289 adults	Sociodemographic predictors, hesitancy reasons, information sources	Vaccine hesitancy due to negative information	Social media, websites, media outlets	Students, health professionals, workers, general population
Bokemper et al [22], 2022; United States; randomized controlled trial	Test public health message effectiveness on social distancing	3184 adults	Impact on distancing beliefs and intentions	COVID-19 conspiracy theories, severity skepticism, downplaying public health measures	Social media, informal networks, public figures	Individuals valuing personal liberty, government mandate opponents
Kumar et al [13], 2022; United States; quantitative observational study	Analyze perceptions of COVID-19 vaccines on Reddit	266,840 Reddit posts	Vaccine-related events and public attitudes	Vaccine efficacy doubts, conspiracy theories, skepticism regarding science and media	Reddit user posts in subreddits	Reddit users, anti-vaccine subreddit frequent users
Xue et al [23], 2022; United States; observational study	Investigate attitudes toward COVID-19 vaccines on Facebook	12,553 Facebook posts	Public attitude shifts, fact-checking effectiveness	Vaccine efficacy questions, safety views, effectiveness challenges, misinterpretation, emotional manipulation	Politicians, social media, health institutions	General public, vaccine-hesitant people, pro- and antivaccine groups
Mourali and Drake [28], 2022; United States; randomized web-based experiment	Assess social media debates on masking and COVID-19 misinformation	500 adults	Attitude, belief, behavior changes from debates	Masking efficacy, truth objectivity, anti-mask arguments	Reddit thread: user *citizen-health* versus user *Health_Scientist*	General public, online forum users, conspiracy-prone individuals
Verma et al [15], 2022; United States; observational study	Study Twitter (subsequently rebranded X) misinformation impact on anxiety	76,985 Twitter users	Causal link between misinformation sharing and anxiety	Vitamins, gargling, 5G technology, involvement of Bill Gates	Twitter	General Twitter users, vulnerable US women, racial minority individuals
AL-Jalabneh [29], 2023; Jordan; qualitative study	Explore vaccine hesitancy due to misinformation	30 vaccine-hesitant adults	Role of misinformation in increased vaccine hesitancy, safety, and effectiveness concerns	Social media misinformation, conspiracy theories, safety doubts, vaccine distrust	Social media (Facebook and WhatsApp), influencers, foreign health experts	Jordanian citizens, active social media users
Gruzd et al [30], 2023; Canada; observational study	Examine Facebook and YouTube for vaccine misinformation	539 YouTube videos shared on Facebook	Prevalence and nature of vaccine misinformation	Vaccine safety, efficacy, ingredients, conspiracy theories	Facebook groups and pages, YouTube videos	Facebook and YouTube users
Kim et al [12], 2023; United States; cross-sectional survey	Investigate impact of misinformation on trust and compliance	1400 adults	Misinformation linked to lower trust in health experts, guideline compliance	False claims about prevention, treatment, severity	Politicians, media, social networks	American adult population
Kosiyaporn et al [31], 2023; Thailand; mixed methods study	Assess vaccine acceptance factors	193,744 adults	Factors influencing vaccine acceptance: public perceptions, attitudes	Vaccine efficacy, side effects, immunity misconceptions	Lack of trust in media, government, celebrities, social media; trustworthy: health professionals, academics	General population, those with limited access to reliable information
Ugarte and Young [32], 2023; United States; randomized controlled trial	Address misinformation, vaccine hesitancy among essential workers	120 adults	Reduced misinformation beliefs, increased vaccine information requests	COVID-19 vaccine misinformation, natural immunity	Social media (Facebook groups), non–peer-reviewed studies	American essential workers, vaccine-hesitant people

Misinformation Themes and Sources

Misinformation Themes

The results of the studies reported many themes that presented a diverse and interconnected landscape of COVID-19 misinformation. A significant amount of this misinformation related to the virus's origins and transmission, with theories varying from accidental laboratory releases to purported links with 5G technology. These theories often reflected a tendency

XSL•FO
RenderX

to misinterpret scientific data or attribute the pandemic to external and frequently sensational causes (Table 1) [3,4,14].

A significant proportion of misinformation concerned treatments and preventives for COVID-19, where unscientific remedies (accidental or deliberate) and vitamin supplements were touted as effective [10,15,24]. This was coupled with widespread misconceptions and conspiracy theories about COVID-19 vaccines [11,13,27,29,30].

Public health measures such as the effectiveness of masks and social distancing were often mischaracterized or misrepresented, sometimes due to political and economic theories [22,25,28]. Social media played a significant role in amplifying dangerous beliefs and practices [12,29]. The studies demonstrate that misinformation during the pandemic ranged from basic misunderstandings to elaborate conspiracy theories [2,21,23,27,31,32].

Sources of Misinformation

The studies provide a comprehensive analysis of the various sources of COVID-19 misinformation, with a particular focus on social media platforms such as Facebook, WhatsApp, Twitter, Reddit, and YouTube, which were repeatedly identified as primary channels for spreading false information (Table 1) [2-4,11-14,22,24,26,27,29,30,32]. These platforms not only facilitated the spread of misinformation through user-generated content but also through public figures and political leaders, whose remarks often fueled rumors and unsubstantiated claims [10,23,31]. Traditional media sources, including television, newspapers, and radio, also added to the misinformation landscape, either by directly spreading false information or by passing on misleading statements and rumors [2,15]. The influence of informal networks, such as family, friends, and community gatherings, was highlighted, pointing to the significance of word-of-mouth communication in the dissemination of misinformation [21,22,24]. Furthermore, the studies identified specific web-based communities and forums, such as Facebook groups and subreddits, where misinformation was not only shared but also reinforced within echo chambers [13,28,32].

Target Audience of Misinformation

The selected studies revealed a complex landscape of COVID-19 misinformation targeting diverse audiences, with a significant focus on the general public across countries; for instance, Datta et al [24] and Hou et al [10] identified both health care professionals and the broader global population, including those in China, the United States, and countries with traditional medicine practices, as key recipients of misinformation (Table 1). Susceptibility to misinformation was also observed in individuals with low health literacy, depression, or susceptibility to conspiracy theories [3,4,13,25] or vaccine-hesitant individuals and those with a mistrust of vaccines [11,23]. Digital platforms played a significant role in shaping public perceptions, with studies highlighting the impact of misinformation on social media users, online forum participants, and those engaging with user-generated content [12,14,21,22,26,28-30]. Moreover, specific populations such as Serbian adults, American women, racial minority individuals, students, public health professionals, and essential workers were reported as being particularly affected or targeted by misinformation campaigns [15,25,27,31,32].

Impacts of Misinformation on Public Health Outcomes

Identified Negative Impact

The findings presented many negative effects of misinformation on public health (Table 2). One primary consequence was the impact on health care professionals, who faced challenges in discerning accurate information, leading to disruptions in routine decision-making and care practices [24]. The public was also affected, with misdirected responses and increased reliance on unproven remedies, indicating missed opportunities for effective epidemic control [10]. Misinformation significantly disrupted health and risk communication, contributing to social unrest and heightened anxiety [3]. It also directly impacted public health measures, as evidenced by lower intent to accept COVID-19 vaccines [11], reduced adherence to official health guidelines [25], and noncompliance with basic preventive measures such as handwashing [4].

The spread of misinformation resulted in decreased public trust in science [14], undermining the effectiveness of public health messaging [22] and leading to increased vaccine hesitancy [27,29,31,32]. This hesitancy was further exacerbated by the promotion of antivaccine propaganda, posing a barrier to achieving herd immunity [30]. The extent of the impact of misinformation was also evident in the public's mental health, with reports of increased anxiety, suicidal thoughts, and distress [2], as well as in overall public attitudes toward the pandemic [26] and changes in public attitudes toward vaccines, which became increasingly negative over time [23].

XSL•FO
RenderX

Table 2. Impact, strategies, and effectiveness of interventions in addressing misinformation.

Study, year	Identified negative impact	Measured outcomes	Potential contributing factors	Intervention strategies	Intervention method	Platform or channel	Effectiveness metrics	Reported effectiveness
Datta et al [24], 2020	Misinformation prevalence among health care professionals	Survey responses	Social media, infodemics	Information regulation, training for accurate information identification	Training, guidelines	Official websites, media	Misinformation reduction, decision-making	Increased awareness of need for accurate information
Hou et al [10], 2020	Misdirected public responses, reliance on unproven remedies	Search trends, purchasing behaviors	Delayed communication, misinformation amplification	Public awareness enhancement, timely misinformation response	Official guidelines, rumor clarification	Government health commission, World Health Organization	Public behavior correlation with trends	Adoption of protective measures, reduction in rumor-driven behavior
Moscadelli et al [3], 2020	Disruption in health or risk communication	Fake news shares, misinformation proportion	Cognitive biases, digital literacy deficiencies	Health literacy improvement, social media policy reinforcement	Social media analysis, campaigns	Social media	Engagement in fake versus verified news	Increased strategies for discerning false news and stress reduction
Loomba et al [11], 2021	Reduced COVID-19 vaccine acceptance	Vaccine acceptance intent	Misinformation exposure, trust in sources	Misinformation impact assessment	Survey, exposure analysis	Web-based panel, social media	Vaccination intent change	Lower vaccination intent due to misinformation
Nowak et al [4], 2021	Nonadherence to sanitary recommendations, increased fears	Handwashing frequency, disinfectant use	Gender, education, living environment	Accurate information dissemination	Web-based behavior assessment survey	Social media, message boards	Adherence to preventive measures	Increased adherence to sanitary recommendations
Scholz et al [21], 2021	Link between information and quarantine measure reactions	Quarantine concern level, compliance	Demographics, media effectiveness	Continuous media reception for risk communication	Loudspeaker announcements, leaflets	Television, radio, internet	Quarantine measure approval	High acceptance or support limiting disease transmission
Teovanović et al [25], 2020	Reduced adherence to health guidelines	Prevention behavior frequency, vaccination intentions	Belief in conspiracy theories	Countering the negative impacts of irrational beliefs	Factual information and debunking	Social media, digital platforms	Engagement in evidence-based health behavior	Importance of strategies in health behavior modification
Agley et al [14], 2021	Lower trust in science due to misinformation belief	Preventive behavior intentions	Political orientation, demographics	Use of infographic to explain the scientific process	Infographic exposure	Web based (Prolific platform)	Trust in science, misinformation belief	Small increase in trust; indirect misinformation effect
Bokemper et al [22], 2022	Reduced public health messaging effectiveness	Beliefs and social distancing scales	Liberty values endorsement, conspiracy theory belief	Community protection–focused strategies	Random message interventions	Web-based survey platform	Social distancing attitudes, intentions	Improved attitudes and intentions toward distancing
Ghaddar et al [2], 2022	Reduced vaccination intent, increased mental health issues	Vaccination intent, conspiracy belief acceptance	Fake news exposure	Promotion of credible sources, debunking	Public campaigns, educational outreach	Television, radio, official channels	Trust in sources, vaccination intent	Increased trust in information sources
Kim et al [26], 2022	Misdirection in pandemic management	Vaccination intention, mandate compliance	Cultural nonconformity, misinformation spread via social media	Tailored communication to misinformed groups	Categorization, analysis	Social media	Intervention specificity and reach	Strategy specificity and misinformation reduction
Kumar et al [13], 2022	Misinformation increase related to vaccine events	Reddit discussion analysis	Media releases, community dynamics	Countering misinformation, engaging skeptics	Accuracy assessment, evidence-based discussion	Reddit, web-based spaces	Theoretical belief shift, vaccine uptake	Effectiveness proposed based on analysis

XSL•FO
RenderX

Study, year	Identified negative impact	Measured outcomes	Potential contributing factors	Intervention strategies	Intervention method	Platform or channel	Effectiveness metrics	Reported effectiveness
Xue et al [23], 2022	Negative public attitudes toward vaccines	Public attitude change, engagement metrics	Information source impact, emotional response	Use of fact-checking messages	Fact-checking posts, collaboration	Facebook	Public attitude change, engagement metrics	Positive role of third-party fact checkers
Mourali and Drake [28], 2022	Increased confusion, uncertainty, and negative attitudes toward health topics	Attitudes toward masking, truth objectivity, argument strength, source competence, sharing intentions	Extended debates undermining truth objectivity	Debunking misinformation	Web-based randomized study	Reddit-like social media simulation	Masking disposition, truth objectivity, sharing intentions	Correcting misinformation improved masking disposition and reduced sharing but waned with repeated exposure
Verma et al [15], 2022	Increased anxiety, especially among specific demographics	Anxiety levels from Twitter (subsequently rebranded X) data	Prior anxiety, exposure to misinformation	Misinformation exposure limitation, direct interventions	Algorithmic feed adaptation	Social media	Anxiety increase after sharing misinformation	Anxiety increase among misinformation sharers
AL-Jalabneh [29], 2023	Increased vaccine hesitancy	Frequency of misinformation themes	Social media misinformation spread	Media literacy campaigns	Educational campaigns, collaboration	Various media channels	Vaccine attitudes, misinformation reduction	Improved vaccine acceptance and trust
Gruzd et al [30], 2023	Vaccine hesitancy promotion	Proportion of misinformation in content	Social media algorithms	Misinformation removal, evidence-based content promotion	Platform moderation, messaging	Facebook, YouTube	Post or account removals, provaccine content prevalence	Partial success in misinformation reduction
Huang et al [27], 2022	Increased vaccine hesitancy	Vaccine hesitancy scale scores	Infodemic, misinformation impact	Timely health education, authoritative information use	Educational campaigns, messaging	Social media, health care settings	Vaccine hesitancy reduction, willingness to change	Positive impact on vaccination willingness
Kim et al [12], 2023	Lower health guidance compliance	Trust in experts, severity perception	Misinformation exposure, political influences	Improving regulatory efforts to curb the spread of misinformation	Survey research to identify misinformation impact	Web-based survey, media analysis	Trust levels, compliance rates	Improved discernment of false or real news, reduced stress and depression related to the pandemic
Kosiyaporn et al [31], 2023	Vaccine hesitancy due to misinformation	Vaccine acceptance rates, trust levels	Risk perception, discerning true information	Infodemic management, vulnerable group prioritization	Surveys, interviews	Web-based channels, volunteer networks	Vaccine acceptance, misinformation discernment	Increased discernment of true or false information correlated with increased vaccine acceptance
Ugarte and Young [32], 2023	Increased hesitancy and misinformation	Web-based discussion engagement	Web-based misinformation, study limitations	Community peer support	Peer leader educational engagement	Facebook groups	Misinformation on social support posts	Reduction in misinformation posts, social support increase

Measured Outcomes

The studies highlighted the challenges that individuals and communities faced in navigating the pandemic amid a flood of misinformation (Table 2). It was reported that misinformation significantly impacted health care professionals, leading to discomfort, distraction, and difficulty in discerning accurate information. This impact affected decision-making and routine practices [24]. The public's response was manifested by changes in search behaviors and purchasing patterns, reflecting the influence of rumors and celebrity endorsements [10]. It was reported that "fake news" significantly affected the information landscape, skewing the perception of truth versus lies [3]. Hesitancy was reported in intent to receive COVID-19 vaccines across demographic groups [11,27,31]. The misinformation also altered health behaviors, such as handwashing and the use of disinfectants, and influenced preventive behavioral intentions [4,14]. It was also reported that misinformation affected public

XSL•FO
RenderX

adherence to COVID-19 prevention, risk avoidance behaviors, and vaccination intentions [25].

The communication strategies during quarantine, public trust and engagement with authorities, and compliance with quarantine measures were influenced by the level of concern, which was shaped by misinformation [21]. It was reported that misinformation led to changes in social distancing and mask wearing [22]. Social media platforms exhibited a prevalence of antivaccine content and a focus on misinformation in web-based discussions [13,30,32]. The studies also reported that emotional and linguistic features in vaccine-related posts influenced public attitudes toward vaccines, reflecting the impact of different information sources [23]. Anxiety levels were heightened due to exposure to misinformation, especially among specific demographic groups [15]. Some of the studies (2/21, 10%) found that misinformation affected public trust in health experts and government and altered the perceived severity of COVID-19 [12,26].

Potential Contributing Factors

The studies identified a wide array of factors that contributed to the spread of misinformation during the pandemic (Table 2). Key among these were social media and connections with family and friends, which hastened the spread of unregulated information [24]. The issue was further compounded by delayed and nontransparent communication from health authorities, coupled with the absence of early, authoritative responses [10]. Cognitive biases, a lack of digital and health literacy, and the exploitation of social divisions also played significant roles [3]. Factors such as sociodemographic characteristics, trust in information sources, the frequency of social media use, and the nature of misinformation were important [11]. The spread of misinformation was also influenced by gender, education level, and the distinction between urban and rural living [4], as well as age, the effectiveness of media channels, the initial understanding of SARS-CoV-2, and trust in authorities, particularly in relation to quarantine measures [21]. Contributing factors included beliefs in conspiracy theories, cognitive intuition, an overestimation of COVID-19 knowledge, and susceptibility to cognitive biases [25], alongside political orientation and religious commitment [14]. Public behavior was also shaped by concerns about government infringement on personal freedoms [22]. Finally, exposure to fake news and conspiracy stories [2], cultural attitudes toward government mandates, and the spread of misinformation through social media were noted [26].

Public Health Communication Strategies and Their Effectiveness

Intervention Strategies

The studies highlighted the critical role of effective public health communication strategies in addressing COVID-19 misinformation (Table 2). This included a range of approaches such as enhancing health literacy and reinforcing social media policies against fake news [3], along with using fact checking and empathetic communication to debunk misinformation [23]. The importance of timely and accurate information dissemination, particularly through social media, was also noted as a crucial component for authoritative communication [4,10,27].

In addition, several studies advocated for tailored communication approaches. These approaches involve targeting specific misinformed subgroups [26], using infographics to clarify scientific processes [14], and focusing on community protection while reframing reckless behaviors [22]. Essential strategies included training health care professionals to accurately identify credible information, alongside implementing media literacy campaigns and prioritizing groups considered vulnerable in public communication [24,29,31]. Engaging skeptics, particularly vaccine skeptics, through interventions was reported as essential [13,32], with an emphasis on debunking misinformation, promoting credible information sources, and reducing exposure to misinformation [2,15,28,30].

Intervention Methods

The included studies reported various intervention methods to combat misinformation. Key strategies included the use of credible sources [3,24,27], the implementation of targeted campaigns, and the integration of digital technologies such as social media tools and algorithmic analyses (Table 2) [4,10,15]. Educational efforts, ranging from basic loudspeaker announcements to sophisticated web-based educational tools and infographics, were also reported to be effective [2,14,21,29]. The importance of engaging the public through surveys, randomized interventions, and peer discussions was noted [11,22,31,32]. Fact checking, in partnership with third-party organizations and through internal processes, was highlighted as crucial, along with the need for empathetic communication [23]. Finally, some of the studies (2/21, 10%) showed the importance of identifying predictors and using analytical models to refine strategies and better understand public sentiment [26,28].

Platform or Channel for Communication

The studies reported that a diverse array of platforms and channels played a crucial role in effective communication during the COVID-19 pandemic (Table 2). Digital and social media platforms, such as Facebook, Reddit, and YouTube, were extensively used to disseminate facts and counter misinformation, as noted by numerous studies (8/21, 38%) [3,4,13,15,23,26,30,32]. Government websites and official channels, alongside health care settings, were also acknowledged for their value in providing reliable and accurate information [10,24,27,29]. Traditional media forms, including television, radio, and print, were found to be crucial in reaching wide audiences [2,21]. Web-based platforms designed for research and surveys, such as Prolific, played a key role in gauging public perceptions and addressing misinformation [11,14,22,28]. Furthermore, community networks and personal communications were identified as essential, particularly in village health volunteer networks and through engagement with health professionals and academics, demonstrating remarkable effectiveness in local communities and areas with limited digital access [25,31].

XSL•FO
RenderX

Effectiveness Metrics and Reported Effectiveness

In studies on public health communication during the pandemic, effectiveness metrics focused on reducing misinformation and improving health behaviors (Table 2) [13,24,25,27,29,31,32]. Detailed engagement metrics included tracking interactions with verified versus fake news, changes in vaccination intent, and shifts in public attitudes toward vaccines over time [3,11,23]. Unique metrics such as internet search trends correlating with public behavior, adherence to health guidelines, and the impact of misinformation on mental health were also explored [4,10,15,28]. Studies such as that by Gruzd et al [30] analyzed social media for misinformation removal and provaccine content. The reported effectiveness of interventions such as fact checking and clear communication varied across the studies, influencing vaccine attitudes and trust in science to varying degrees [11,14,23]. Some of the studies (8/21, 38%) pointed to increased public support for measures such as quarantine, emphasizing the role of community engagement [21,22], but also noted challenges in maintaining long-term effectiveness and addressing various reactions such as anxiety in response to misinformation [2,13,15,25,26,28]. These studies, often based on computational analyses, existing literature, and theoretical models, highlighted the complex, multifaceted nature of public health communication during the pandemic [3,4,30].

Recommendations, Gaps, and Future Directions

Recommendations for Addressing COVID-19 Misinformation

The included studies recommended a comprehensive approach that included strategic public health communication, educational initiatives, and policy adaptation (Table 3) [2,24]. Key themes included effective information regulation and enhancing discernment skills among health care professionals as well as the general public [2,24], while strategies included considering platform-specific and demographic-focused approaches to combat misinformation [3,31]. Governmental leadership and international coordination were considered crucial [10], and educational strategies were recommended to focus on improving health literacy and researching misinformation inoculation [4,14,25]. Public health messaging and web-based moderation policies were deemed effective [13,22], and technological interventions and comprehensive policy making were recommended [15,30]. Methodological research to understand extended debates and debunking techniques was emphasized [26,28], as well as tailored communication and messaging strategies [11,12,21,27,29] (Table 3).

XSL•FO
RenderX

Table 3. Overview of recommendations, research gaps, and future directions in misinformation management.

Study, year	Recommendation	Specifics of recommendation	Identified gaps	Proposed future research or action
Datta et al [24], 2020	Develop training for information discernment in health care	Focus on skills for identifying and validating medical information in crises	Difficulty discerning authentic versus nonauthentic information; misinformation prevalence on social media	Formulate guidelines for medical information dissemination; enhance crisis communication skills; ethical training in information validation
Moscadelli et al [3], 2020	Strengthen strategies against misinformation in digital media	Enforce policies against fake news; develop demographic-specific communication and health literacy programs	Persistence of fake news; echo chambers on social media; low health literacy and misinformation susceptibility	Conduct research on countering fake news; enhance anti-misinformation measures on social platforms; develop targeted demographic interventions; evaluate health literacy programs
Hou et al [10], 2020	Enhance governmental risk communication and international coordination	Improve transparency and timeliness in risk communication; control misinformation; promote science-backed behaviors	Lack of timely advice for personal protection; inadequate early risk communication; missed opportunities for epidemic control	Assess the impact of government communication on public behavior; study the role of international organizations in outbreak response; develop international partnership strategies
Loomba et al [11], 2021	Adopt targeted communication strategies for vaccine misinformation	Counter misinformation with specific messaging strategies, including altruistic and scientific clarification	Lack of real-world social media research; variable impact of misinformation across demographics	Conduct social media–based studies on vaccine misinformation; establish causal relationships between misinformation types and vaccination intent; tailor public health communication for social media
Scholz et al [21], 2021	Diversify and localize communication strategies for health information	Use various media for rapid communication; address informational needs across demographics; use localized methods in rural settings	Uncertain role of health authorities; evolving media preferences during crises; variable effectiveness in information dissemination	Establish pre-event credibility of health authorities; study media habits in crises; assess long-term behavioral changes after quarantine; evaluate alternative communication methods
Nowak et al [4], 2021	Implement educational initiatives for better public understanding of preventive measures	Focus on accurate information communication and increasing public adherence to preventive measures	Challenges in public adherence to measures; susceptibility to misinformation	Conduct research on communication strategies to increase adherence; focus on demographic-specific interventions; explore psychological factors influencing public responses
Teovanović et al [25], 2020	Develop strategies to mitigate the effects of irrational beliefs and conspiracy theories	Explore and counter distrust in institutions and political cynicism; use factual corrections and debunking techniques	Reliance on self-reported data; lack of cognitive ability control; non-representativeness of sample	Investigate psychological factors affecting health behaviors; create targeted interventions; include observed behaviors in future studies for robust findings
Agley et al [14], 2021	Advance research into strategies for misinformation inoculation	Investigate the efficacy of truthful messaging about scientific processes to combat misinformation	Limited experimental research on misinformation's behavioral effects	Conduct experimental studies testing various methods of communicating scientific processes; focus on misinformation impacts
Bokemper et al [22], 2022	Promote public health messaging to reshape social distancing perceptions and collective responsibility	Reframe social distancing in public messaging; emphasize the importance of collective protection	Uncertainty about which message elements are most effective; observed attitudinal changes not matched by behavioral changes	Dissect effective elements of public health messages; develop strategies to convert attitudes into behaviors; conduct long-term study on message impact

XSL•FO
RenderX

Study, year	Recommendation	Specifics of recommendation	Identified gaps	Proposed future research or action
Kumar et al [13], 2022	Advocate for public health messaging and web-based moderation to address misinformation	Develop tailored communication strategies; engage with committed antivaccine groups; introduce verified-information tags	Challenges in changing beliefs of antivaccine individuals; moderating web-based information	Target interventions at vaccine skeptics; enhance web-based moderation policies; evaluate the effectiveness of these strategies
Kim et al [26], 2022	Focus on methodological research to identify specific misinformation types	Investigate distinct misinformation strains (eg, “vaccine chip” vs “vaccine poison”)	Misalignment between initial misinformation categories and their public health impact; lack of detailed study on antivaccine misinformation	Conduct research on different antivaccine misinformation subtypes; focus on underrepresented communities for comprehensive insights
Huang et al [27], 2022	Strategic communication and interventions for vaccine hesitancy	Target health care providers and the public with educational campaigns	Need for improved information dissemination; lack of health care provider communication training	Research effective communication strategies; create platforms to combat misinformation; design targeted interventions
Xue et al [23], 2022	Comprehensive communication strategies to combat vaccine misinformation	Design posts that will better engage the public; balance negative misinformation with empathetic communication	Underexplored impact of various information sources on vaccine attitudes; emotional responses to health communication not fully understood	Study the influence of information sources on public engagement; investigate emotional appeals in health communication; develop strategies for credible sources to enhance social media influence
Verma et al [15], 2022	Technological and educational interventions for misinformation-related anxiety	Use machine learning and social media data for anxiety detection; use health literacy initiatives	Challenges related to privacy, the First Amendment; limitations in fact-checking resources; unexplored causal relationships	Explore ethically compliant technological interventions; develop efficient resource allocation policies; create inclusive educational programs; conduct extensive studies on psychological and sociodemographic impacts
Mourali and Drake [28], 2022	Extended research on social media debates and debunking techniques	Examine the effectiveness of humor and infographics in debunking; test “prebunking” strategies	Generalizability of findings to other platforms; effectiveness of debunking in extended debates	Quantify occurrence of extended debates; investigate the impact of message elements and sources; examine consequences of engaging with misinformation spreaders
Ghaddar et al [2], 2022	Enhance critical thinking and credibility in public health communication	Promote trusted information sources; evaluate social media content critically	Effectiveness of communication strategies; understanding of belief drivers	Conduct longitudinal studies on public behavior and attitude changes; perform research on social media content engagement
Kim et al [12], 2023	Develop communication strategies to counter misinformation and enhance public trust	Focus on enhancing public trust and compliance with health guidelines	Limited research on misinformation mechanisms	Investigate interventions to mitigate misinformation effects; study impact on public trust and guideline compliance
Gruzd et al [30], 2023	Policy- and platform-based interventions for misinformation management	Strengthen misinformation policies; launch proactive public health campaigns	Inconsistent policy enforcement; persistence of echo chambers	Conduct research on the effectiveness of platform interventions; develop strategies against echo chambers
AL-Jalabneh [29], 2023	Strategic and educational interventions to reduce vaccine hesitancy	Media literacy campaigns; government-media collaboration to improve health literacy	Insufficient health literacy; widespread misinformation on social media	Adopt a collaborative approach to combat misinformation; conduct effectiveness studies of interventions; develop long-term health literacy improvement strategies

XSL•FO
RenderX

Study, year	Recommendation	Specifics of recommendation	Identified gaps	Proposed future research or action
Kosiyaporn et al [31], 2023	Strategic public health communication and infodemic management	Enhance infodemic management; target groups considered vulnerable with specific communication strategies	Lack of large-scale surveys that include noninternet users; limited exploration of misinformation–vaccine acceptance relationship	Monitor misinformation trends; implement fact checking and legal actions; develop communications to debunk myths
Ugarte and Young [32], 2023	Strategy adaptation and research in public health contexts	Apply community peer support and educational engagement to combat misinformation	Small sample size; high engagement skewness; selection bias in Facebook users	Extend intervention duration; increase sample size; focus on factual information dissemination; consider a broader demographic

Identified Gaps in Addressing Misinformation

The studies highlighted several gaps in managing COVID-19 misinformation and public health communication. Challenges included distinguishing authentic information from misinformation, the persistence of fake news, and the presence of echo chambers in social media networks (Table 3) [3,24,30]. Timely, actionable advice for personal protection and effective risk communication during the early stages of the pandemic was lacking [10]. Research limitations included a lack of real-world simulation, leading to challenges in generalizability [11,25,26]. There was insufficient understanding of the role of health authorities as trusted sources, media preference during crises, and the effectiveness of information dissemination in different regions [2,21]. Challenges arising from legal and ethical considerations, resource limitations, disparities in education access, and insufficient exploration of the relationship between misinformation and vaccine acceptance were also noted [15,31,32] (Table 3).

Proposed Future Research and Actions

Future research directions included developing guidelines for medical information dissemination, enhancing crisis communication skills among health care professionals, and creating targeted interventions based on demographics (Table 3) [3,13,24,30]. Evaluating the impact of governmental and international organization communications, conducting research within social media settings, and analyzing the impact of misinformation more accurately were recommended [10,11]. Studying media habits during crises, examining long-term behavioral changes after quarantine, and dissecting the influential aspects of messages were suggested [14,21,25]. Investigating psychological factors, evaluating emotional appeals in health communication, and developing strategies for credible sources to enhance their social media influence were proposed [23,26]. Ethically and legally compliant technological interventions, efficient resource allocation policies, and extensive studies on psychological impacts were recommended [15]. Mourali and Drake [28] proposed quantifying extended debates, studying message elements and sources, and exploring "prebunking." Longitudinal studies, research on user engagement with social media content, and interventions to mitigate misinformation effects were highlighted [2,12,32]. Finally, the studies suggested a holistic approach involving collaboration among companies, governments, and users; continuous monitoring of misinformation trends; regular fact checking; legal actions against sources of misinformation; and specific communications to debunk myths [29,31] (Table 3).

Discussion

Principal Findings

Our study underscores the profound influence of misinformation during the COVID-19 pandemic, particularly in shaping public responses. Misinformation, primarily propagated through social media, led to widespread misconceptions about the severity of COVID-19 infection, triggering public confusion, reluctance to adhere to health guidelines, and increased vaccine hesitancy. This phenomenon significantly impacted vaccine uptake rates. Gallotti et al [33] highlighted the simultaneous emergence of infodemics alongside pandemics, underlining the critical role of both human and automated (bots) accounts in spreading information of questionable quality on platforms such as Twitter. The authors introduced an Infodemic Risk Index to measure the exposure to unreliable news, showing that the early stages of the COVID-19 pandemic saw a significant spread of misinformation, which only subsided in favor of reliable sources as the infection rates increased [33]. This emphasizes the complex challenge of managing infodemics in tandem with biological pandemics, necessitating adaptive public health communication strategies that are responsive to evolving information landscapes. Our findings resonate with historical observations in public health crises, evidenced by studies on the Zika virus outbreak [34], polio vaccination efforts in India and Nigeria [35], and the Middle East respiratory syndrome outbreak [36]. Similar patterns of misinformation were also noted in the H1N1 pandemic and the Ebola outbreak. These instances highlight the critical need for clear, proactive communication strategies to effectively manage misinformation and guide public understanding and responses.

The review also reveals a predominant focus on digital misinformation, underscoring the necessity to comprehend the impact of traditional media and word-of-mouth communication in spreading misinformation. While studies such as that by Basch et al [37] have started to address this gap, there is a clear need for more extensive research, particularly on the long-term effects of misinformation on public health behaviors after a pandemic. This shift toward credible information, as observed by Gallotti et al [33], signals an opportunity for future research to explore capitalizing on changing information consumption patterns in public health messaging. Such observations are crucial for developing effective communication strategies,

XSL•FO
RenderX

highlighting the necessity of integrating infodemic management with pandemic response efforts to mitigate misinformation effects and guide public behavior appropriately. The disparity in the effectiveness of misinformation mitigation strategies points to the need for a nuanced understanding of how misinformation evolves over time. Studies, such as that by Vijaykumar et al [38], highlight the challenges in countering rapidly changing misinformation narratives on digital platforms. Further investigation into the effectiveness of fact checking across different cultures and demographics, as suggested by Chou et al [39], is essential for developing better strategies to combat misinformation in diverse settings.

This review found that various factors, including delayed communication from health authorities, cognitive biases, sociodemographic characteristics, trust in official sources, and political orientation, played a significant role in the spread of misinformation during the pandemic. These findings align with similar observations in other studies. Eysenbach [40] emphasized the importance of trust in government agencies and health care providers in shaping individuals' beliefs and their willingness to share accurate information during public health crises. In addition, Pennycook and Rand [41] highlighted how political beliefs and affiliations can influence people's interpretation of information, thus impacting their acceptance or rejection of official guidance during public health crises. The study by Gallotti et al [33] also highlighted the differentiated roles of verified and unverified users on social media in propagating COVID-19–related information. Their analysis shows that verified users began to point more toward reliable sources over time, hinting at the potential of leveraging social media influencers and verified accounts in directing public attention to factual and scientifically verified information [33].

These insights indicate the critical need for dynamic public health strategies that are adaptable and actionable, aimed at curtailing misinformation through education and technology. It is essential to incorporate digital literacy and clear, audience-specific messaging to effectively counter misinformation, a strategy that has proven successful in health crises beyond the COVID-19 pandemic; for example, during the H1N1 pandemic, targeting specific audience segments with tailored messages significantly improved public understanding and guideline compliance [42]. Likewise, during the Ebola outbreak, proactive and transparent strategies were key in dispelling rumors and building trust in public health authorities [43]. These approaches, based on an understanding of the target audience's concerns and media habits, are consistent with our findings where digital literacy and targeted messaging played a critical role in mitigating COVID-19 misinformation effects. Such strategies are vital not only for immediate crisis response but also for fostering long-term resilience in public health communication, helping to enable the public to distinguish credible information from misinformation, with the ultimate goal of enhancing public health outcomes and trust in health authorities.

In examining the authoritarian responses to the pandemic, particularly in Brazil and Turkey, it is evident that leadership tactics significantly contributed to societal polarization and misinformation. Leaders in these countries used the crisis to suppress dissent and consolidate power, often spreading misinformation and underreporting COVID-19 cases, thereby exacerbating public mistrust and confusion [5]. Similarly, a study of communication strategies across countries with high rates of infection emphasized the variation in political leaders' approaches, where strategies ranged from science-based communications to ideologically influenced messaging [6]. The study highlighted the potential for political leaders to influence public health responses through their communication tactics, further impacting public behavior and trust in health guidelines [6]. In certain situations, the integration of political ideology with public health messaging, as observed in countries such as the United States, Brazil, India, and the United Kingdom, not only perpetuated misinformation but also intensified societal rifts [5,6]. This highlights the paramount role of leadership in navigating public health crises; for instance, in the United States and Brazil, political leaders' approaches to the COVID-19 pandemic—characterized by mixed messaging on mask wearing and social distancing—contributed to public confusion and a politicized response to the pandemic. Similarly, the initial underestimation of the virus's impact in India and the United Kingdom's delayed lockdown response serve as examples of how political decisions can shape public health outcomes and trust in health authorities, emphasizing the profound impact of aligning political views with public health communication [5,6]. In addition, the initial reluctance of the World Health Organization to endorse mask wearing, social distancing, and handwashing, followed by a later reversal of these recommendations, exemplifies the challenges and confusion created by global health leadership during the early stages of the pandemic [44]. Such shifts in guidance contributed to the global spread of misinformation, further complicating public health responses and trust in international health authorities [6]. These approaches, based on an understanding of the target audience's concerns and media habits, are consistent with our findings that digital literacy and targeted messaging played a critical role in mitigating COVID-19 misinformation effects. Such strategies are vital not only for immediate crisis response but also for fostering long-term resilience in public health communication, helping to enable the public to distinguish credible information from misinformation, with the ultimate goal of enhancing public health outcomes and trust in health authorities. Applying the MEGA framework in practical settings could revolutionize public health communication, offering a model for how technology can be harnessed to tackle misinformation more effectively. By processing massive graph data sets and accurately computing infodemic risk scores, MEGA supports the development of targeted communication strategies and interventions. Its approach to preserving crucial feature information through graph neural networks signifies a leap forward in optimizing learning performance, underscoring the framework's utility in crafting evidence-based policies and initiatives to effectively combat misinformation. This emphasizes the importance of integrating advanced technological solutions, such as MEGA, into public health strategies to enhance the precision and effectiveness of infodemic management [19]. The integration of social media literacy into public health strategies is emphasized as essential by Ziapour et al [7], suggesting that a populace equipped with

XSL•FO
RenderX

advanced media literacy skills exhibits greater resilience against misinformation.

Our study reveals the profound impact of the COVID-19 infodemic, which extended beyond public health and eroded trust in health institutions and government authorities. This decline in trust contributed to societal polarization, mirroring the effects seen in the Ebola outbreak, where misinformation led to notable repercussions [45,46]. Further research, similar to that conducted on the Zika outbreak by Basch et al [37], is needed to understand the long-term effects of misinformation on societal cohesion and trust. Addressing this evolving landscape of misinformation requires dynamic and adaptable public health policies. These strategies should integrate insights from various methodologies, using both digital and traditional media for greater reach and impact, drawing lessons from the successful strategies deployed during the H1N1 pandemic, such as those highlighted by Chou et al [39].

Our study advocates for a collaborative approach, uniting governments, the private sector, and the public in a concerted effort to combat misinformation, highlighting the importance of joint action in this global challenge. This approach should include continuous monitoring of misinformation trends, implementing regular fact checking, taking legal action against sources of misinformation, and developing specific communications to debunk myths. Similar findings have been reported in studies addressing misinformation related to the Zika virus [34,47], yellow fever [48], and Ebola [49], emphasizing the importance of a holistic strategy involving all stakeholders [50].

Limitations

The review has several limitations to consider. First, there is a temporal limitation because it included only studies published between December 2019 and September 2023, potentially excluding more recent research that could have offered additional insights. Second, the reliance on specific databases (MEDLINE [PubMed], Embase, and Scopus) as the primary sources for data might have led to the omission of pertinent studies that are not indexed in these databases. Third, the study's sole focus on research articles may have excluded valuable insights from other scholarly works such as conference papers, theses, case studies, and gray literature. Finally, it is important to acknowledge that the study's restriction to English-language publications may have excluded valuable research conducted in other languages. While efforts were made to review the available literature comprehensively, omitting non-English sources could limit the breadth and depth of the findings. Recognizing these limitations, future endeavors should aim to expand the scope of research beyond these constraints, incorporating a more diverse range of sources, languages, and real-world interventions to enrich our understanding of, and response to, misinformation.

Conclusions

The results of this review emphasize the significant and complex challenges posed by misinformation during the COVID-19 pandemic. It shows how misinformation can have a wide impact on public health, societal behaviors, and individual mental well-being. The findings highlight the critical role of effective public health communication strategies in addressing the infodemic. It is essential that these strategies are not only targeted and precise but also adaptable and inclusive, ensuring that they are relevant to diverse demographic and sociocultural contexts.

The review also emphasizes the need for ongoing collaborative research efforts to further explore the nuances of the misinformation spread and its consequences. This requires cooperation among health authorities, policy makers, communication specialists, and technology experts to develop evidence-based approaches and policies to combat misinformation.

Furthermore, the review highlights the importance of refining public health communication strategies to keep up with the ever-changing nature of misinformation, especially in the digital realm. It advocates using advanced technology and data-driven insights to enhance the reach and impact of health communication. By combining scientific rigor, technological innovation, and empathetic communication, these strategies can contribute to building public trust, promoting health literacy, and creating resilient communities capable of recognizing and countering misinformation.

In summary, the lessons learned from the COVID-19 pandemic emphasize the necessity of strengthening public health communication infrastructures. This strengthening is vital for addressing the current misinformation crisis and preparing for future public health emergencies. Implementing these recommendations will play a crucial role in shaping a more informed, aware, and health-literate global community better equipped to confront the challenges posed by misinformation in our increasingly interconnected world. Furthermore, future research directions should explore integrating advanced large language models with frameworks similar to MEGA. This exploration will bolster automated fact checking and infodemic risk management, contributing to more effective strategies in combating misinformation in public health communication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines.
[PDF File (Adobe PDF File), 549 KB - jmir_v26i1e56931_app1.pdf]

XSL•FO
RenderX

References

1. Infodemic. World Health Organization. URL: https://www.who.int/health-topics/infodemic#tab=tab_ [accessed 2020-12-22]
2. Ghaddar A, Khandaqji S, Awad Z, Kansoun R. Conspiracy beliefs and vaccination intent for COVID-19 in an infodemic. PLoS One 2022 Jan 12;17(1):e0261559 [FREE Full text] [doi: 10.1371/journal.pone.0261559] [Medline: 35020721]
3. Moscadelli A, Albora G, Biamonte MA, Giorgetti D, Innocenzio M, Paoli S, et al. Fake news and COVID-19 in Italy: results of a quantitative observational study. Int J Environ Res Public Health 2020 Aug 12;17(16):5850 [FREE Full text] [doi: 10.3390/ijerph17165850] [Medline: 32806772]
4. Nowak BM, Miedziarek C, Pełczyński S, Rzymski P. Misinformation, fears and adherence to preventive measures during the early phase of COVID-19 pandemic: a cross-sectional study in Poland. Int J Environ Res Public Health 2021 Nov 22;18(22):12266 [FREE Full text] [doi: 10.3390/ijerph182212266] [Medline: 34832021]
5. Kisa A. A comparison of the authoritarian strategies used by Brazil and Turkey to tackle the COVID-19 crisis. J Appl Bus Econ 2023 Aug 25;25(4):179-186. [doi: 10.33423/jabe.v25i4.6348]
6. Mohamed Nour M, Kisa A. Political leaders' communication strategies during COVID-19 in highly infected countries: a scoping review. Healthcare (Basel) 2024 Mar 07;12(6):607 [FREE Full text] [doi: 10.3390/healthcare12060607] [Medline: 38540571]
7. Ziapour A, Malekzadeh R, Darabi F, Yıldırım M, Montazeri N, Kianipour N, et al. The role of social media literacy in infodemic management: a systematic review. Front Digit Health 2024 Feb 14;6:1277499 [FREE Full text] [doi: 10.3389/fdgth.2024.1277499] [Medline: 38419808]
8. Smith MJ, Upshur RE. Ebola and learning lessons from moral failures: who cares about ethics? Public Health Ethics 2015 Nov 17;8(3):305-318 [FREE Full text] [doi: 10.1093/phe/phv028] [Medline: 32288786]
9. Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination? BMJ 2014 Oct 14;349(oct14 5):g6178. [doi: 10.1136/bmj.g6178] [Medline: 25315514]
10. Hou Z, Du F, Zhou X, Jiang H, Martin S, Larson H, et al. Cross-country comparison of public awareness, rumors, and behavioral responses to the COVID-19 epidemic: infodemiology study. J Med Internet Res 2020 Aug 03;22(8):e21143 [FREE Full text] [doi: 10.2196/21143] [Medline: 32701460]
11. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar 05;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
12. Kim JW, Lee J, Dai Y. Misinformation and the Paradox of Trust during the COVID-19 pandemic in the U.S.: pathways to Risk perception and compliance behaviors. J Risk Res 2023 Apr 06;26(5):469-484. [doi: 10.1080/13669877.2023.2176910]
13. Kumar N, Corpus I, Hans M, Harle N, Yang N, McDonald C, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. BMC Public Health 2022 Mar 07;22(1):446 [FREE Full text] [doi: 10.1186/s12889-022-12824-7] [Medline: 35255881]
14. Agley J, Xiao Y, Thompson EE, Chen X, Golzarri-Arroyo L. Intervening on trust in science to reduce belief in COVID-19 misinformation and increase COVID-19 preventive behavioral intentions: randomized controlled trial. J Med Internet Res 2021 Oct 14;23(10):e32425 [FREE Full text] [doi: 10.2196/32425] [Medline: 34581678]
15. Verma G, Bhardwaj A, Aledavood T, de Choudhury M, Kumar S. Examining the impact of sharing COVID-19 misinformation online on mental health. Sci Rep 2022 May 16;12(1):8045 [FREE Full text] [doi: 10.1038/s41598-022-11488-y] [Medline: 35577820]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Hang CN, Yu P, Chen S, Tan CW, Chen G. MEGA: machine learning-enhanced graph analytics for infodemic risk management. IEEE J Biomed Health Inform 2023 Dec;27(12):6100-6111. [doi: 10.1109/JBHI.2023.3314632] [Medline: 37713230]
20. Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. J Med Internet Res 2020 Jun 29;22(6):e21820 [FREE Full text] [doi: 10.2196/21820] [Medline: 32589589]
21. Scholz J, Wetzker W, Licht A, Heintzmann R, Scherag A, Weis S, et al. The role of risk communication in public health interventions. An analysis of risk communication for a community quarantine in Germany to curb the SARS-CoV-2 pandemic. PLoS One 2021 Aug 13;16(8):e0256113 [FREE Full text] [doi: 10.1371/journal.pone.0256113] [Medline: 34388211]
22. Bokemper SE, Huber GA, James EK, Gerber AS, Omer SB. Testing persuasive messaging to encourage COVID-19 risk reduction. PLoS One 2022 Mar 23;17(3):e0264782 [FREE Full text] [doi: 10.1371/journal.pone.0264782] [Medline: 35320285]

XSL•FO
RenderX

23. Xue H, Gong X, Stevens H. COVID-19 vaccine fact-checking posts on Facebook: observational study. J Med Internet Res 2022 Jun 21;24(6):e38423 [FREE Full text] [doi: 10.2196/38423] [Medline: 35671409]
24. Datta R, Yadav A, Singh A, Datta K, Bansal A. The infodemics of COVID-19 amongst healthcare professionals in India. Med J Armed Forces India 2020 Jul;76(3):276-283 [FREE Full text] [doi: 10.1016/j.mjafi.2020.05.009] [Medline: 32773929]
25. Teovanović P, Lukić P, Zupan Z, Lazić A, Ninković M, Žeželj I. Irrational beliefs differentially predict adherence to guidelines and pseudoscientific practices during the COVID-19 pandemic. Appl Cogn Psychol 2021 Dec 07;35(2):486-496 [FREE Full text] [doi: 10.1002/acp.3770] [Medline: 33362344]
26. Kim S, Capasso A, Ali SH, Headley T, DiClemente RJ, Tozan Y. What predicts people's belief in COVID-19 misinformation? A retrospective study using a nationwide online survey among adults residing in the United States. BMC Public Health 2022 Nov 18;22(1):2114 [FREE Full text] [doi: 10.1186/s12889-022-14431-y] [Medline: 36401186]
27. Huang Y, Su X, Xiao W, Wang H, Si M, Wang W, et al. COVID-19 vaccine hesitancy among different population groups in China: a national multicenter online survey. BMC Infect Dis 2022 Feb 14;22(1):153 [FREE Full text] [doi: 10.1186/s12879-022-07111-0] [Medline: 35164700]
28. Mourali M, Drake C. The challenge of debunking health misinformation in dynamic social media conversations: online randomized study of public masking during COVID-19. J Med Internet Res 2022 Mar 02;24(3):e34831 [FREE Full text] [doi: 10.2196/34831] [Medline: 35156933]
29. AL-Jalabneh AA. Health misinformation on social media and its impact on COVID-19 vaccine inoculation in Jordan. Commun Soc 2023 Jan 10;36(1):185-200. [doi: 10.15581/003.36.1.185-200]
30. Gruzd A, Abul-Fottouh D, Song MY, Saiphoo A. From Facebook to YouTube: the potential exposure to COVID-19 anti-vaccine videos on social media. Soc Media Soc 2023 Feb 01;9(1):20563051221150403 [FREE Full text] [doi: 10.1177/20563051221150403] [Medline: 36751363]
31. Kosiyaporn H, Netrpukdee C, Pangkariya N, Chandrasiri O, Tangcharoensathien V. The impact of vaccine information and other factors on COVID-19 vaccine acceptance in the Thai population. PLoS One 2023 Mar 7;18(3):e0276238 [FREE Full text] [doi: 10.1371/journal.pone.0276238] [Medline: 36881578]
32. Ugarte DA, Young S. Effects of an online community peer-support intervention on COVID-19 vaccine misinformation among essential workers: mixed-methods analysis. West J Emerg Med 2023 Feb 27;24(2):264-268 [FREE Full text] [doi: 10.5811/westjem.2023.1.57253] [Medline: 36976597]
33. Gallotti R, Valle F, Castaldo N, Sacco P, de Domenico M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nat Hum Behav 2020 Dec 29;4(12):1285-1293. [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
34. Venkatraman A, Mukhija D, Kumar N, Nagpal SJ. Zika virus misinformation on the internet. Travel Med Infect Dis 2016 Jul;14(4):421-422. [doi: 10.1016/j.tmaid.2016.05.018] [Medline: 27267799]
35. Larson HJ, Jarrett C, Eckersberger E, Smith DM, Paterson P. Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007-2012. Vaccine 2014 Apr 17;32(19):2150-2159. [doi: 10.1016/j.vaccine.2014.01.081] [Medline: 24598724]
36. Fung IC, Tse ZT, Chan BS, Fu K. Middle East respiratory syndrome in the Republic of Korea: transparency and communication are key. Western Pac Surveill Response J 2015;6(3):1-2 [FREE Full text] [doi: 10.5365/WPSAR.2015.6.2.011] [Medline: 26668758]
37. Basch CH, Fung IC, Hammond RN, Blankenship EB, Tse ZT, Fu K, et al. Zika virus on YouTube: an analysis of English-language video content by source. J Prev Med Public Health 2017 Mar;50(2):133-140 [FREE Full text] [doi: 10.3961/jpmph.16.107] [Medline: 28372356]
38. Vijaykumar S, Jin Y, Nowak G. Social media and the virality of risk: the risk amplification through media spread (RAMS) model. J Homel Secur Emerg Manag 2015;12(3):653-677 [FREE Full text] [doi: 10.1515/jhsem-2014-0072]
39. Chou WS, Oh A, Klein WM. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
40. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
41. Pennycook G, Rand DG. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc Natl Acad Sci U S A 2019 Feb 12;116(7):2521-2526 [FREE Full text] [doi: 10.1073/pnas.1806781116] [Medline: 30692252]
42. Gray L, MacDonald C, Mackie B, Paton D, Johnston D, Baker MG. Community responses to communication campaigns for influenza A (H1N1): a focus group study. BMC Public Health 2012 Mar 19;12(1):205 [FREE Full text] [doi: 10.1186/1471-2458-12-205] [Medline: 22429559]
43. Sell TK, Hosangadi D, Trotochaud M. Misinformation and the US Ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak. BMC Public Health 2020 May 07;20(1):550 [FREE Full text] [doi: 10.1186/s12889-020-08697-3] [Medline: 32375715]
44. COVID-19 - virtual press conference. World Health Organization. URL: https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-press-conference-full-30mar2020.pdf [accessed 2024-04-29]

XSL•FO
RenderX

45. Das U, Fielding D. Higher local Ebola incidence causes lower child vaccination rates. Sci Rep 2024 Jan 16;14(1):1382 [FREE Full text] [doi: 10.1038/s41598-024-51633-3] [Medline: 38228678]
46. Klofstad CA, Uscinski JE, Connolly JM, West JP. What drives people to believe in Zika conspiracy theories? Palgrave Commun 2019 Apr 02;5(1):36. [doi: 10.1057/S41599-019-0243-8]
47. Dredze M, Broniatowski DA, Hilyard KM. Zika vaccine misconceptions: a social media analysis. Vaccine 2016 Jun 24;34(30):3441-3442 [FREE Full text] [doi: 10.1016/j.vaccine.2016.05.008] [Medline: 27216759]
48. Ortiz-Martínez Y, Jiménez-Arcia LF. Yellow fever outbreaks and Twitter: rumors and misinformation. Am J Infect Control 2017 Jul 01;45(7):816-817. [doi: 10.1016/j.ajic.2017.02.027] [Medline: 28343702]
49. Allgaier J, Svalastog AL. The communication aspects of the Ebola virus disease outbreak in Western Africa--do we need to counter one, two, or many epidemics? Croat Med J 2015 Oct;56(5):496-499 [FREE Full text] [doi: 10.3325/cmj.2015.56.496] [Medline: 26526888]
50. Baseman JG, Revere D, Painter I, Toyoji M, Thiede H, Duchin J. Public health communications and alert fatigue. BMC Health Serv Res 2013 Aug 05;13(1):295 [FREE Full text] [doi: 10.1186/1472-6963-13-295] [Medline: 23915324]

Abbreviations

MEGA: machine learning–enhanced graph analytics
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Mavragani; submitted 30.01.24; peer-reviewed by CN Hang, S Gordon, A Hassan; comments to author 14.03.24; revised version received 02.04.24; accepted 12.06.24; published 21.08.24.

Please cite as:
Kisa S, Kisa A
A Comprehensive Analysis of COVID-19 Misinformation, Public Health Impacts, and Communication Strategies: Scoping Review
J Med Internet Res 2024;26:e56931
URL: https://www.jmir.org/2024/1/e56931
doi:10.2196/56931
PMID:

©Sezer Kisa, Adnan Kisa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Data Quality–Driven Improvement in Health Care: Systematic Literature Review

Anthony Lighterness[1], BBiomedSc, MPhil, MSc; Michael Adcock[1], BSc, MSc; Lauren Abigail Scanlon[1], PhD; Gareth Price[2], PhD

[1]Clinical Outcomes and Data Unit, The Christie NHS Foundation Trust, Manchester, United Kingdom

[2]Radiotherapy Related Research Group, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Anthony Lighterness, BBiomedSc, MPhil, MSc
Clinical Outcomes and Data Unit
The Christie NHS Foundation Trust
550 Wilmslow Road
Manchester, M20 4BX
United Kingdom
Phone: 44 7305054646
Email: anthony.lighterness@nhs.net

Abstract

Background: The promise of real-world evidence and the learning health care system primarily depends on access to high-quality data. Despite widespread awareness of the prevalence and potential impacts of poor data quality (DQ), best practices for its assessment and improvement are unknown.

Objective: This review aims to investigate how existing research studies define, assess, and improve the quality of structured real-world health care data.

Methods: A systematic literature search of studies in the English language was implemented in the Embase and PubMed databases to select studies that specifically aimed to measure and improve the quality of structured real-world data within any clinical setting. The time frame for the analysis was from January 1945 to June 2023. We standardized DQ concepts according to the Data Management Association (DAMA) DQ framework to enable comparison between studies. After screening and filtering by 2 independent authors, we identified 39 relevant articles reporting DQ improvement initiatives.

Results: The studies were characterized by considerable heterogeneity in settings and approaches to DQ assessment and improvement. Affiliated institutions were from 18 different countries and 18 different health domains. DQ assessment methods were largely manual and targeted completeness and 1 other DQ dimension. Use of DQ frameworks was limited to the Weiskopf and Weng (3/6, 50%) or Kahn harmonized model (3/6, 50%). Use of standardized methodologies to design and implement quality improvement was lacking, but mainly included plan-do-study-act (PDSA) or define-measure-analyze-improve-control (DMAIC) cycles. Most studies reported DQ improvements using multiple interventions, which included either DQ reporting and personalized feedback (24/39, 61%), IT-related solutions (21/39, 54%), training (17/39, 44%), improvements in workflows (5/39, 13%), or data cleaning (3/39, 8%). Most studies reported improvements in DQ through a combination of these interventions. Statistical methods were used to determine significance of treatment effect (22/39, 56% times), but only 1 study implemented a randomized controlled study design. Variability in study designs, approaches to delivering interventions, and reporting DQ changes hindered a robust meta-analysis of treatment effects.

Conclusions: There is an urgent need for standardized guidelines in DQ improvement research to enable comparison and effective synthesis of lessons learned. Frameworks such as PDSA learning cycles and the DAMA DQ framework can facilitate this unmet need. In addition, DQ improvement studies can also benefit from prioritizing root cause analysis of DQ issues to ensure the most appropriate intervention is implemented, thereby ensuring long-term, sustainable improvement. Despite the rise in DQ improvement studies in the last decade, significant heterogeneity in methodologies and reporting remains a challenge. Adopting standardized frameworks for DQ assessment, analysis, and improvement can enhance the effectiveness, comparability, and generalizability of DQ improvement initiatives.

(J Med Internet Res 2024;26:e57615) doi:10.2196/57615

XSL•FO
RenderX

KEYWORDS

real-world data; data quality; quality improvement; systematic literature review; PRISMA

Introduction

Background

The landscape of health care, improvement science, and digital technologies increasingly hinges on real-world data (RWD) to improve patient care and outcomes [1,2]. RWD encompasses a vast and dynamic collection of health-related information generated by means of routine clinical care from a diverse range of sources, such as electronic health records (EHRs), electronic medical records (EMRs), hospital information systems (HISs), picture archiving and communication systems (PACSs), national registries, claims data, and wearable devices [2-4]. Despite their long history, the adoption and use of EHRs have become widespread only during the last decade [4,5]. EHRs and EMRs are often used interchangeably in the literature discussing health-related RWD, where some suggest that EMRs are a subset of EHRs [6], but the prominence of EHRs have positioned them as a primary source of RWD due to the comprehensive spectrum of patient information, including genetic testing, treatment modalities, and clinical outcomes [7,8]. To reflect this primary focus, the term *EHR-RWD* will be used throughout this review to denote RWD derived from EHRs.

Real-world evidence (RWE) generated from EHR-RWD holds unprecedented potential to bridge the unmet gaps that exist between controlled clinical trial studies and the complexities of health care delivery in the real world [1,2,7,9]. While randomized control trial studies remain the “gold standard” for assessing the efficacy of new interventions, the essence of RWE lies in its potential to reflect the diversity, heterogeneity, and nuances of patient populations and care settings, thus enabling a more holistic understanding of health outcomes and interventions [7,10]. Data and RWE studies can support the life cycle of drug development, clinical and regulatory decision-making, and health technology assessment [3,8,11]. Moreover, RWE underpins the vision of the learning health care system (LHS), which is a paradigm built upon the cycle of continuous learning to achieve personalized medicine [12]. The transformative potential of RWE, however, hinges on a pivotal caveat—high-quality data.

Despite its potential, EHR-RWD, and by extension RWE, grapple with formidable barriers, and chief among these is data quality (DQ) [1,2,9,10,13,14]. The need for high-quality data was exemplified during the COVID-19 pandemic, when EHR-RWD was critical for research and planning [15]. Regulatory authorities, such as the Food and Drug Administration (FDA), European Medicines Agency (EMA), National Institute for Health and Care Excellence (NICE), and Medicines and Healthcare Products Regulatory Agency (MHRA), recommend the reporting of DQ metrics and dimensions to provide additional context to real-world study outcomes, thus serving as the foundation for trustworthy RWE [3,14,16-19]. These guidelines mostly promote ad hoc DQ assessment and reporting, with the exception that EMA briefly notes the importance of assessing DQ as close as possible to the moment of data capture to help with collection errors [3,16-18]. While ad hoc DQ measurement and reporting can increase transparency and awareness of the limitations of real-world study outcomes, understanding the causes of poor data capture is needed for long-term, sustainable improvement in DQ and, subsequently, the impact of RWE to support decision-making in health care.

In this study, we aimed to evaluate the robustness of studies seeking to improve the quality of structured EHR-derived data. Incorporating quality improvement best practices, we aimed to assess how studies measure DQ, identify which interventions are implemented, and summarize the outcomes to understand which interventions are useful to improve DQ.

DQ Theory

The quality of data describes if the data meet the expectations of a data consumer and, therefore, if they are fit for purpose [20]. This expected behavior can be documented and understood using metadata, that is, additional data that provide meaning and context by describing how the data they are associated with should have been captured, defined, structured, and represented [20]. Metadata, in turn, can inform the design of quantifiable metrics that measure the compliance of data against a set of relevant business rules and constraints during DQ assessment [21]. DQ metrics, which can act as the method by which to measure the respective DQ dimensions, can serve a crucial purpose in understanding areas needing improvement.

Scholars have proposed various multidimensional frameworks for comprehensive DQ assessment. While there may be disagreements on semantic choices and definitions across these frameworks, certain DQ concepts are consistently studied and well represented in the literature, popular frameworks, and DQ profiling software [22]. Despite the challenge, some common concepts can be mapped across frameworks, as exemplified by the 6 core dimensions defined by the Data Management Association (DAMA): completeness, validity, consistency, uniqueness, timeliness, and accuracy (Textbox 1) [20].

Textbox 1 outlines the 6 fundamental DQ dimensions defined by DAMA to enable standardization and comparability of DQ concepts defined, assessed, and improved in the included DQ improvement literature. These dimensions are essential for assessing and improving the quality of RWD.

In an exploration of the DQ literature, we found 10 other reviews that summarized the most frequently represented DQ dimensions across DQ studies, software, and theoretical frameworks [19,21,23-30]. All 10 reviews demonstrated the concept of completeness to be well represented. Of these, 4 reviews agreed that data accuracy and consistency were popular DQ concepts [23,24,26,28]. Timeliness and validity were said to be popular by 3 different reviews, and uniqueness was only raised in the study by Gordon et al [29].

XSL•FO
RenderX

Textbox 1. Core dimensions of data quality as defined by the Data Management Association.

Data quality dimension and description

- Completeness: the presence of the expected data
- Uniqueness: uniqueness of records where duplication is not expected
- Timeliness: a measure of data freshness
- Consistency: a check of consistency between multiple sources of the same data elements
- Validity: the validity of data against data standards or plausible values, ranges, or patterns
- Accuracy: a check of consistency of source data against a reference gold standard

Completeness, also known as missingness, is often reported to be popular among DQ frameworks, tooling, and studies [22-27,29,30]. In general, it refers to the degree to which all required or expected data values or records are present [20]. The most common method to measure completeness consists of counting records with blank, unknown, empty, "NULL," or "NaN" values, though variations may include a measurement of data availability [22,25,26]. Reviews by Weiskopf and Weng [22] and Syed et al [28] also found variations of data completeness assessment involving triangulation of multiple sources to create a gold standard. However, this approach risks assuming the accuracy of available data. Previous work demonstrate that missing data can lower statistical power of research outcomes and lead to biased assumptions with improper use of imputation methods [10,31,32].

Data accuracy is also well studied in DQ literature [22,25,28]. The accuracy dimension measures the extent to which data reflect the truth of events and conform to their actual value [20,22,25,28]. Other terms used to describe accuracy include error, correctness, integrity, trustworthiness, reliability, and validity [22,28]. The most common method to assess accuracy in health care involves the comparison of EHR data to a reference gold standard, which may include paper records, manual data reviews, triangulation of data from multiple sources, or interviews with patients [22]. Measurement of data accuracy can identify issues such as lack of specificity or precision [33]. Previous work found that code precision can be related to staff training or use of multiple EHR systems [33-36]. As Cook et al [37] noted in a review of DQ issues affecting social determinants data, imprecise codified data may affect minority groups disproportionately, which in turn may affect secondary research outcomes.

Various terms and definitions for validity exist in the literature [22,28]. It generally describes the conformance of data to expected value ranges, patterns, formats, general medical knowledge, or data standards as set by local or external authorities [22,26,28,38]. Validity is also termed plausibility, conformance, and integrity and can also be separated into internal and external validity or be incorporated with other data elements such as temporal validity [22,38]. As EHR-derived data contain large volumes of categorical data, such as patient demographic, diagnostic, and treatment-related information, validity constraints are needed to identify areas needing standardization [39,40]. Data standardization has been shown to correlate positively to data sharing capabilities and emergency care [41]. However, improper design of standardized data entry user interfaces, such as the use of excessively long drop-down lists for diagnostic codes, can also increase cognitive demand, lower workflow efficiency, and correlate to clinician burnout [42].

Timeliness refers to several time-related characteristics of data and is, therefore, also termed currency, recency, or freshness [22,38]. For example, time-related data items can measure how closely the recorded information corresponds to the actual event. Factors affecting timely capture of EHR data include workflow inefficiencies, documentation burden, limited access to hardware, and interruptions [43-48]. Batch processing of data long after the event may indicate a lack of timeliness and affects other DQ dimensions such as completeness, accuracy, and validity [28,44,46,48,49].

Consistency, otherwise known as concordance, describes the agreement of similar data elements between multiple sources [20,22]. The existence of multiple data capture systems and RWD sources can give rise to inconsistent data for a given patient, and in the absence of a defined gold standard, the consistency dimension can identify potentially erroneous data [10,50]. Botsis et al [36] identified multiple inconsistencies during a DQ analysis of a cohort of patients with pancreatic cancer stored in the Columbia University Medical Center's EHR data warehouse. These included pancreatitis recorded as chronic in pathology reports but acute in clinical notes and patients with diabetes receiving both codes for type 1 and type 2 in the same EHR source [36]. von Lucadou et al [51] found similar discrepancies when comparing data items between different systems, adding that inconsistencies may be caused by individual documentation habits. Measurement of data consistency highlights potential duplication and redundancy between different EHR sources and within the same EHR system and can thus help improve data capture or data engineering workflows.

The uniqueness dimension identifies where duplication of objects, events, or values are not expected [10,26]. Duplication of patient EHRs frequently occur when disparate data flows that contain overlapping objects are combined [10,38]. Similar to consistency, the uniqueness dimension can identify and resolve redundant and inefficient workflows and processes [52]. This is particularly relevant given that 60% to 90% of clinicians routinely copy and paste data between systems [53]. The "copy and paste" phenomenon is pervasive in health care and is known

XSL•FO
RenderX

to promote inconsistencies, propagate errors, and contribute to documentation burden and clinician fatigue [42,48,53]. As such, the uniqueness dimension is related to consistency and accuracy.

DQ Tooling

DQ measurement involves the process by which defective values are identified and labeled through the application of business rules or automated tooling [21]. The subsequent analysis of DQ results can then be aggregated, analyzed, and summarized, providing key insights for improvement. Tools to support these activities are widely reported and well studied [21,27,29,30,54,55].

While the availability of DQ tools is abundant, the literature reveals a considerable gap in the effective support for DQ improvement efforts, particularly in the realm of health care. Evaluations of DQ profiling software in the studies by Ehrlinger and Wöß [30], Gordon et al [29], and Ozonze et al [21] highlight limitations in the range of DQ metrics offered for assessment, interoperability issues, and complex configuration requirements.

Root cause analysis, a pivotal aspect of DQ management, is also notably lacking in demonstrations within this landscape. Eden et al [48] demonstrated the utility of the Odigos framework in qualitative root cause analysis, which classifies DQ issues that emanate from the material world, such as digital infrastructure or access to hardware; personal world, that is, staff behaviors; and societal world, that is, job roles and social norms. The legal and technical implications associated with data cleaning, as opposed to addressing the root causes of poor data, underscore potential risks to patient safety [56-58]. Consequently, the overarching trend in DQ tool development leans toward prioritizing technical features, leaving a noticeable gap in the demonstration of their utility in the prevention and improvement of poor-quality data capture in real-world health settings.

Quality Improvement

Quality improvement describes the use of systematic continuous approaches to create positive changes in an area of need [59]. Various structured, iterative learning frameworks, such as plan-do-study-act (PDSA), total data quality management (TDQM), define-measure-analyze-improve-control (DMAIC), and the LHS, exist [12,60-63]. Lacking a universally agreed-upon model, each methodology focuses on enhancing different areas, ranging from service evaluation to treatment standards [59].

Developed from the earlier plan-do-check-act cycle by Deming [64], the PDSA cycle enhances the traditional model by prioritizing the “study” stage—a deeper analysis rather than a simple check [65]. This adaptation roots the PDSA cycle firmly in the scientific method, encouraging a disciplined approach to testing and monitoring changes over time [65]. Its flexible and qualitative nature makes it particularly suitable for health care settings where adaptability to complex and variable processes is crucial.

Proposed by Wang [61] in the late 1990s, TDQM adapts traditional total quality management principles specifically to data management, highlighting the importance of data as a key asset or product. In health care, where decision-making increasingly relies on accurate and timely data, TDQM offers a robust framework to ensure the integrity and usability of data. This focus on DQ management is critical as health care systems integrate more digital processes and data-driven decision-making frameworks. TDQM adapts the PDSA planning stage to specifically target the improvement of DQ [61].

A product of Motorola engineers in the 1980s, DMAIC provides a structured, data-driven quality improvement methodology [61]. Unlike the more qualitative PDSA, DMAIC emphasizes quantifiable metrics and statistical analysis to identify and mitigate variations in processes. This makes DMAIC highly suitable for health care areas requiring high levels of measurement precision and control, such as clinical laboratories or any clinical process where outcomes need to meet high standards of care.

While all 3 methodologies share a structured, iterative approach and a reliance on empirical data to drive improvements, they cater to different needs within the health care sector. PDSA’s qualitative and flexible nature is best suited for areas requiring rapid change and adaptability. In contrast, DMAIC’s rigorous, statistical approach fits environments where precision and control are paramount. TDQM’s specific focus on DQ fills a critical niche in ensuring the reliability of health care data systems. What remains unknown is their implementations in the real world. In a review of PDSA cycles aimed at improving treatment standards, Taylor et al [60] demonstrated that <20% of implementations comply with the core features including running multiple iterative learning cycles, the notion of small-scale change, and the use of quantitative data at monthly or more frequent intervals to inform progression of cycles. The stages of each methodology are listed and described in Table 1.

Table 1 presents a comparison of the stages involved in 3 iterative learning frameworks present in health quality improvement and other related literature. The comparison aims to understand common themes in quality improvement methodology and how these can be applied to DQ improvement.

XSL•FO
RenderX

Table 1. Detailed comparison of 3 QI[a] frameworks for iterative learning.

Framework and stage	Description
PDSA[b] [60]	
Plan	Identify a change hypothesis and plan a small test.
Do	Conduct a study plan with the collection of data.
Study	Analyze and interpret the results.
Act	Adapt the change based on feedback and plan the next iteration.
TDQM[c] [61]	
Define	Define target data requirements and DQ[d] dimensions.
Measure	Create metrics to evaluate these dimensions.
Analyze	Investigate root causes for DQ issues.
Improve	Identify key areas for improvement based on DQ root cause analysis.
DMAIC[e] [63]	
Define	Define project scope and objectives.
Measure	Identify and measure baseline service indicators.
Act	Analyze baseline metrics and identify causes of errors.
Improve	Implement changes to reduce or remove root causes of defects.
Control	Put mechanisms in place to ensure sustained improvement.

[a]QI: quality improvement.
[b]PDSA: plan-do-study-act.
[c]TDQM: total data quality management.
[d]DQ: data quality.
[e]DMAIC: define-measure-analyze-improve-control.

Objectives

The rapidly growing body of DQ publications and software tools indicate that this field has gained significant traction, and recent publications illustrate that there is no shortage of DQ concepts, frameworks, and tools [21,24,28-30,54,55,66,67]. While these surveys already provide comprehensive theoretical and functional evaluations on existing DQ concepts and tools for definition and measurement, this represents only the early stages of a bigger picture in DQ improvement and management. Our aim was to evaluate the robustness of studies seeking to use DQ measurement as part of DQ improvement initiatives, focusing on improving the quality of structured EHR-derived data.

Brouwer et al [67], Wiebe et al [68], and Lemma et al [69] have previously published studies on DQ-driven improvement in health care. These studies are compared in Table 2. Wiebe et al [68] included 24 studies aiming to improve EHR documents such as operative reports or discharge summaries. The authors reported that heterogeneity in tools or metrics used to measure the quality of unstructured clinical notes made it difficult to evaluate outcomes. However, 8 included studies used an ad hoc questionnaire and 1 used the validated Physician Documentation Quality Instrument (PDQI-9) tool. Although unstructured notes in health care is a pervasive and ubiquitous source of important patient information, this scope limits the exploration of semiautomated or automated DQ assessment tools or methods [68].

Table 2 aims to compare various literature reviews focused on DQ improvement, identifying current knowledge gaps and evaluating the existing body of research. The goal is to understand current progress and unmet needs.

Brouwer et al [67] evaluated studies published up to 2005 and limited to a general practice setting. With digital health care technology and culture evolving rapidly, a more recent and broader evaluation is needed. Lemma et al [69] focused on low- and middle-income countries, where initiatives generally targeted broader and less-specific DQ improvement compared to high-income countries or technologically advanced institutions. The review expands on these works by evaluating contemporary DQ improvement studies targeting structured her-derived RWD agnostic of health care settings. Our evaluation is guided by quality improvement best practices to understand how studies measure and seek to improve DQ dimensions as defined by the well-recognized DAMA framework [20]. Specifically, we addressed the following three questions:

1. How do quality improvement studies define and measure the quality of data?
2. What interventions are being implemented to improve the quality of RWD?
3. What are the outcomes reported?

XSL•FO
RenderX

Table 2. Comparison of literature reviews evaluating DQ[a] improvement studies.

Comparison	Review paper		
	Brouwer et al [67]	Wiebe et al [68]	Lemma et al [69]
Period covered	Up to 2005	2004-2016	2008-2020
Number of studies	12	24	20
Structured RWD[b]	✓[c]		✓
Completedness	✓	✓	✓
Accuracy	✓	✓	✓
Timeliness		✓	✓
Consistency			✓
Validity			
Uniqueness			
QI[d] framework			

[a]DQ: data quality.
[b]RWD: real-world data.
[c]Did evaluate.
[d]QI: quality improvement.

Methods

Search Strategy and Information Sources

In this review, studies seeking to improve the quality of structured EHR data were examined using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (refer to Multimedia Appendix 1) [70]. The Population - Intervention - Comparison - Outcome - Context framework was used to identify relevant keywords and Medical Subject Headings (MeSH) terms [71,72]. These were combined using Boolean operators to create strategic search queries, which were then used to search the Ovid MEDLINE and PubMed databases for articles published from 1945 to July 2023 (refer to Multimedia Appendix 2 for more information). Additional relevant papers were identified from other publications and manual searches through Google Scholar.

Literature Selection Process

The Ovid MEDLINE and PubMed search results were downloaded as a research information systems file and PubMed text file, respectively. These were then imported into the Mendeley reference manager (Mendeley Ltd) and Rayyan (Rayyan Systems Inc) software for iterative analysis [73]. The Rayyan web app was used to streamline the selection process. Articles were selected based on the following criteria: (1) they describe a DQ assessment or measurement process, (2) they focus on data from an EHR or EMR system, and (3) they involve an intervention aimed at improving DQ over time. The search strategies and article selection process were performed independently by 2 reviewers: AL and MA. Excluded articles were nonempirical studies, improvement studies focusing on quality of care or treatment standards instead of quality of data, studies targeting semistructured or unstructured data or data not captured by an RWD source such as an EHR or EMR system, and studies without an intervention seeking to improve DQ. Table 3 summarizes the key inclusion and exclusion criteria for paper selection. In total, 39 studies were included in the review, as presented by the PRISMA flow diagram for RWD (Figure 1).

Table 3 details the specific criteria used to include or exclude studies in the evaluation of current methods for assessing and improving the quality of structured health RWD. These criteria help to systematically assess the landscape of DQ assessment and improvement strategies.

XSL•FO
RenderX

Table 3. Inclusion and exclusion criteria to evaluate the current landscape of DQ[a] assessment and improvement approaches.

Category	Inclusion criteria	Exclusion criteria
Article type	Empirical, original, or review articles where tools, frameworks, or interventions seek to measure and improve DQ	Nonempirical studies, thesis papers, and non–peer-reviewed publications
Language	Published in English	Articles not published in English
Access	Peer-reviewed and open access articles	Papers that are not free to access
Primary target for quality improvement	Studies that primarily aim to improve the quality of data	Studies that primarily target improvement of treatment standards, standard of care, and clinical workflows without a DQ focus
Study population	Studies targeting structured, tabular data	Studies targeting semistructured or unstructured data
Data source	Data from RWD[b] sources such as EHR[c], EMR[d], PACS[e], or HIS[f]-like systems	Data generated by clinical trial studies
DQ assessment and reporting	Studies that describe a DQ assessment, quantification, or measurement process and implemented an intervention seeking to improve DQ over time	Studies that focus on DQ tool development without demonstration of measurement or improvement of DQ over time
Location or health context	No criteria applied	No criteria applied
Time frame	Studies published since 1945	Studies published before 1945

[a]DQ: data quality.

[b]RWD: real-world data.

[c]EHR: electronic health record.

[d]EMR: electronic medical record.

[e]PACS: picture archiving and communication system.

[f]HIS: hospital information system.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for study selection to evaluate the current literature targeting assessment, analysis, and improvement of the quality of structured real-world data in health care. This PRISMA flowchart illustrates the selection process for reports included in the review, detailing the steps of inclusion and exclusion needed to accurately achieve the intended scope of the study.

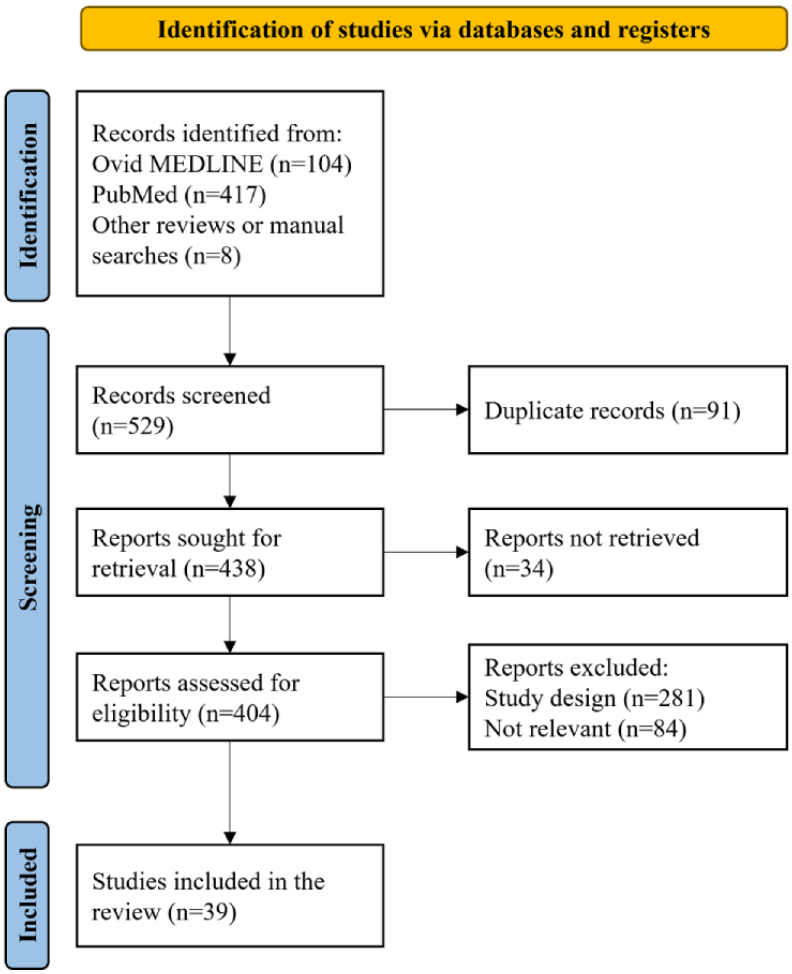

Data Extraction and Synthesis

Following paper selection, we defined a set of data elements essential for addressing the defined research questions. Subsequently, 2 authors (AL and MA) independently extracted and documented this information from each study. The collected data were then cross-checked for notable discrepancies, and any disparities were resolved through consensus. Five key sections of information were extracted: (1) study characteristics (eg, year of publication and health domain), (2) study plans outlined (eg, descriptions of target data and relevant metadata to plan DQ assessment), (3) DQ assessment (eg, methods and dimensions), (4) interventions (eg, which interventions were implemented seeking to improve DQ?), and (5) outcomes (eg, how are results reported?).

Results

Overview of the Study Characteristics

We identified 39 studies describing DQ improvement initiatives in health care that targeted structured RWD sources [74-112]. These are listed in Tables 4 and 5. Most were published between 2009 and 2022 (37/39, 95%), with 2 published in 2002. Studies were affiliated with institutions in the United States (15/39, 38%), followed by Kenya (4/39, 10%), Australia (2/39, 5%), or 3% (1/39) each with 15 other countries. We classified the affiliated institutions into different levels of health care, including primary (11/39, 28%), secondary (10/39, 26%), tertiary (15/39, 38%), and community (3/39, 8%). The primary domains of health care were general practice (4/39, 10%), HIV care (4/39, 10%), intensive care (4/39, 10%), tropical medicine (4/39, 10%), oncology (3/39, 8%), surgery (3/39, 8%), or 3% (1/39) each from 12 other domains.

Table 4 lists the 6 DQ dimensions defined by DAMA, as described, assessed, and/or improved in all included studied. It highlights the framework's role in standardizing and comparing DQ concepts within the reviewed literature.

DQ improvement studies targeted RWD sources that were generated by varying numbers of institutions. Target data were generated by either a single organization (9/39, 23%) or multiple different sites, which ranged from 2 to 10 (11/39, 28%), 11 to 50 (9/39, 23%), or >51 (9/39, 23%) different organizations. A variety of terminology were used to describe the source systems, including EMR (12/39, 31%), national registries or databases (8/39, 21%), EHR (7/39, 18%), HIS (7/39, 18%), clinical information networks (6/39, 15%), PACS (1/39, 3%), or claims data (1/39, 3%).

XSL•FO
RenderX

Table 4. Standardized data quality (DQ) dimensions defined by DAMA[a] across the included studies.

ID	Study, year	Completeness (n=31)	Validity (n=12)	Uniqueness (n=4)	Consistency (n=11)	Timeliness (n=6)	Accuracy (n=12)
1	de Lusignan et al [109], 2002	✓[b]	✓		✓	✓	
2	Wallace et al [108], 2002	✓					✓
3	Nassaralla et al [105], 2009	✓					✓
4	Amoroso et al [112], 2010						
5	Griever et al [107], 2011	✓					✓
6	Ahlbrandt et al [96], 2012						✓
7	Mphatswe et al [110], 2012	✓					✓
8	Rahbar et al [106], 2013	✓	✓		✓		✓
9	Knight et al [101], 2014	✓				✓	
10	Siegel et al [100], 2014	✓					
11	Benard et al [111], 2015	✓					
12	Genet et al [104], 2015						
13	Haskew et al [99], 2015	✓					
14	Smith et al [95], 2015	✓		✓			
15	Soto et al [102], 2015	✓					
16	Taggart et al [98], 2015	✓	✓	✓	✓		
17	Ewing et al [103], 2016						✓
18	Ma et al [77], 2016	✓	✓		✓	✓	
19	Tuti et al [97], 2016	✓					
20	Qin et al [93], 2017	✓					
21	Edgerton et al [92], 2018	✓					
22	Miyoshi et al [89], 2018	✓	✓	✓			
23	Muthee et al [94], 2018	✓			✓		
24	Qualls et al [90], 2018	✓	✓				
25	Daniel et al [88], 2019	✓	✓	✓	✓		
26	Bhattacharya et al [82], 2020	✓			✓	✓	✓
27	Dean et al [75], 2020					✓	✓
28	Koo et al [81], 2020						
29	Moomba et al [85], 2020	✓	✓				
30	Ng et al [78], 2020	✓					✓
31	Njugunaet al [91], 2020		✓				
32	Sinaiko et al [87], 2020	✓					
33	Larrow et al [74], 2020	✓					✓
34	Manesen et al [83], 2021	✓			✓		
35	Olagundoye et al [84], 2021						✓
36	Tizifa et al [86], 2021	✓			✓		
37	Kiogou et al [80], 2022	✓	✓		✓	✓	
38	Pfaff et al [79], 2022	✓	✓		✓		
39	Tuti et al [76], 2022	✓	✓				

[a]DAMA: Data Management Association.

[b]Did assess and target the DQ dimension for improvement.

XSL•FO
RenderX

Table 5. Themes of interventions for DQ[a] improvement in included studies.

ID	Study, year	IT or technical (n=21)	Training (n=17)	DQ report and feedback (n=24)	Workflow (n=5)	Cleaning (n=3)
1	de Lusignan et al [109], 2002			✓[b]		
2	Wallace et al [108], 2002	✓	✓			
3	Nassaralla et al [105], 2009		✓	✓		
4	Amoroso et al [112], 2010	✓	✓	✓		
5	Griever et al [107], 2011		✓		✓	
6	Ahlbrandt et al [96], 2012	✓	✓			
7	Mphatswe et al [110], 2012		✓	✓		
8	Rahbar et al [106], 2013			✓		✓
9	Knight et al [101], 2014	✓	✓	✓		
10	Siegel et al [100], 2014			✓		
11	Benard et al [111], 2015	✓				
12	Genet et al [104], 2015				✓	
13	Haskew et al [99], 2015	✓		✓		
14	Smith et al [95], 2015	✓				
15	Soto et al [102], 2015	✓	✓	✓		
16	Taggart et al [98], 2015			✓		
17	Ewing et al [103], 2016	✓	✓			
18	Ma et al [77], 2016	✓				
19	Tuti et al [97], 2016		✓	✓		
20	Qin et al [93], 2017	✓			✓	
21	Edgerton et al [92], 2018					
22	Miyoshi et al [89], 2018	✓				
23	Muthee et al [94], 2018	✓		✓		
24	Qualls et al [90], 2018			✓		
25	Daniel et al [88], 2019	✓	✓	✓		
26	Bhattacharya et al [82], 2020		✓	✓		
27	Dean et al [75], 2020	✓				
28	Koo et al [81], 2020	✓	✓	✓		
29	Moomba et al [85], 2020				✓	
30	Ng et al [78], 2020	✓	✓	✓		✓
31	Njugunaet al [91], 2020			✓		
32	Sinaiko et al [87], 2020			✓		
33	Larrow et al [74], 2020	✓	✓		✓	
34	Manesen et al [83], 2021	✓		✓		
35	Olagundoye et al [84], 2021	✓	✓	✓		
36	Tizifa et al [86], 2021	✓	✓	✓		
37	Kiogou et al [80], 2022					✓
38	Pfaff et al [79], 2022			✓		
39	Tuti et al [76], 2022			✓		

[a]DQ: data quality.
[b]Did implement.

XSL•FO
RenderX

DQ Assessment Methods

We found various approaches to DQ assessment. The duration of studies ranged from 1 month to 9 years, as did the frequency of DQ assessment. Most studies measured and reported DQ before and after the intervention (38/39, 97%) at varying intervals, including a single before and after comparison (19/39, 49%), yearly (4/39, 10%), quarterly (2/39, 5%), monthly (11/39, 28%), fortnightly (1/39, 3%), weekly (1/39, 3%), or specified data cycles (1/39, 3%). DQ assessment was achieved using manual (15/39, 38%), automated (3/39, 8%), or semiautomatic (13/39, 33%) methods, whereas some methods lacked sufficient description (6/39, 15%) to be classified. Semiautomated methods for DQ assessment mostly involved the scheduling of ad hoc, manually curated programmatic scripts using either R, SAS, or structured query language (SQL) programming languages [76,85,88,90,91,96-99,101,106-108]. Furthermore, 3 studies applied automated methods that used the World Health Organization (WHO) DQ assessment tool and the Open Data Kit [82,91,94] but did not explicitly describe how this was implemented.

The approaches to defining and assessing DQ dimensions varied. We found that 6 studies explicitly referenced 1 of 2 DQ frameworks, including Weiskopf and Weng [22] (3/39, 8%) and Kahn et al [38] (3/39, 8%) [79,80,82,88,90,94]. To enable comparison between studies, the DQ metrics and dimensions reported were extracted and classified according to the DAMA DQ framework. Some DQ concepts lacked sufficient detail to allow classification (eg, studies reporting "error rate," "wrong data," and "percentages of correctly coded" [93,96,112]). These were classified as "unclear" (8/39, 21%). DQ improvement studies assessed the dimensions of completeness (31/39, 79%), accuracy (12/39, 31%), validity (12/39, 31%), consistency (11/39, 28%), timeliness (6/39, 15%), and uniqueness (4/39, 10%). The number of DQ dimensions targeted per study were 1 (13/39, 33%); 2 (14/39, 36%); 3 (2/39, 5%); or 4 (7/39, 18%).

Studies reported inconsistent terminology and definitions for DQ dimensions. For example, although the completeness dimension was generally assessed as the presence or absence of expected data, variations included the proportion of linkage of records between systems [109], use of a gold standard to identify missing patients [92], or overlapping completeness with other dimensions such as validity [76] or accuracy [108]. Validity, also termed conformance or plausibility, was targeted by 12 studies. Of these 39 studies, 6 (15%) used data standards such as WHO International Classification for Disease (ICD) version 9 or 10 [77,85,89,90], SNOMED [77,79,89], Health Language 7 (HL7) [77,88,89], or RxNorm [90]. Others assessed validity by defining business rules that incorporated expected values, formats, or ranges based on local or general medical knowledge [85,89,90,98]. Studies occasionally equated validity with accuracy or correctness [76,89,98].

Of the 12 studies that targeted accuracy, 8 (67%) reported the development and/or use of a gold standard for reference. However, varying definitions for what studies deemed to be a "gold standard" were provided. This included paper charts [93,108]; national data [91,110]; a manually curated data set [84,106]; or manual validation by a trained, expert clinical coder [103]. For example, the "gold standard" in the study by Rahbar et al [106] included 30 patient records that were manually abstracted by a team of experts that included a vascular neurologist clinician before comparing to national stroke registry records. In another study by Ahlbrandt et al [96], the gold standard was described as "the documented and encoded (using OPS [Operationen und Prozedurenschlüssel] Classification, the German modification of ICPM [International Classification of Procedures in Medicine]) surgical procedure," which could be interpreted either as a data standard or patient data in electronic or paper form.

Sometimes DQ dimensions were subsumed by another. For example, data could only be deemed accurate when they were both complete and correct across multiple data elements [105]. In the absence of what is deemed a "gold standard," data consistency was similarly assessed by comparison with paper records [77,83,94], multiple registers [86], or national data [82].

We found that studies' assessment of uniqueness and timeliness were generally consistent with the DAMA definitions. For uniqueness, studies assessed if records were unexpectedly duplicated, for example, in primary keys [79] or patient names [95]. Similarly, timeliness was consistently assessed as the difference in time between point of data capture versus actual timing of events [74,75,77,81,82,104]. In contrast, some of the DQ concepts reported could not be classified according to the DAMA DQ framework, including simplicity, acceptability, flexibility, stability, usefulness [77], and conformance to a specified data model [90]. These were collectively classified as "other" (3/39, 8%).

Interventions for Improvement

Studies varied in their approaches to plan and deliver DQ improvement interventions. In total, 20 studies reported using quantitative or qualitative data analysis before planning an intervention. Qualitative analysis involved assessment of clinical workflow inefficiencies through process mapping techniques or staff surveys [74,108]. In contrast, quantitative analysis involved an assessment of DQ with interpretations of possible root causes [95,96].

To understand the types of interventions studied, we identified 5 common themes, including DQ reporting and feedback (24/39, 62%), IT-related or technical solutions (21/39, 54%), training (17/39, 44%), workflow (5/39, 13%), or data cleaning (3/39, 8%). All studies implemented at least 1 intervention with most implementing multiple interventions (23/39, 59%). DQ reporting and feedback involved assessing DQ and sharing curated results with a specific stakeholder with the aim of encouraging improved data capture behavior. These stakeholders included individual clinical staff or managers [87,98,105] or health care institutions as a whole [82,88,90,101,109,110].

Taggart et al [98] implemented structured DQ reports combined with feedback sessions to improve the quality of EHR data in general practice settings. This approach leveraged regular assessments and direct feedback to practice managers to foster ongoing improvements in data recording practices, illustrating a practical application of DQ feedback mechanisms in a real-world health care setting [98]. In contrast, Sinaiko et al

XSL•FO
RenderX

[87], studied peer comparison feedback emails in a randomized controlled study to assess its effectiveness on improving cancer stage data completeness, underscoring the importance of control groups in validating the impact of DQ interventions.

We found a range of subthemes under the IT-related or technical-based interventions. These improvements involved either introducing a new electronic data capture system, upgrading an existing one [77,81,93,99,102,103,108,111], enhancing front-end user interfaces [74,78,96], or refining back-end data flow processes. Ahlbrandt et al [96] introduced an intervention focusing on improving the graphical user interface of anesthesia information management systems to enhance the validity of the data captured. By shifting from drop-down lists to radio buttons, rearranging the graphical user interface layout, and limiting user options to a set list, they aimed to reduce invalid data entry by making the interface more intuitive and compliant with data standards. This study exemplifies how interface design can directly influence data validity and highlights the impact of front-end modifications [96].

Technology-based interventions often overlapped with training and workflow changes. Ewing et al [103] implemented a browser-assisted clinical coding software along with training, which, in turn, improved efficiencies in clinical workflows. Other studies mainly targeted workflow inefficiencies [74,85,104,107], with Greiver et al [107] introducing a data entry clerk, whereas Moomba et al [85] shifting data entry responsibility from data entry clerks to frontline clinical staff.

To plan, implement, and assess the impact of these DQ improvement interventions, we found that of the 39 studies, only 5 (13%) used a standardized quality improvement framework or iterative learning cycle, such as PDSA (4/39, 10%) [74,81,84,101] or DMAIC (1/39, 3%) [88]. All 4 PDSA studies completed multiple cycles, ranging from 3 to 8. One study reportedly conducted 421 PDSA cycles across 54 different sites [101]. Larrow et al [74] applied the PDSA method to enhance the timeliness of discharge summaries at a pediatric hospital. The study team initiated their quality improvement project by identifying key barriers through qualitative analysis of staff surveys, leading to the strategic redesign of the EHR structured discharge summary template. Notable enhancements included embedded writing tips and standardized drop-down menus for common diagnoses.

Daniel et al [88] applied the DMAIC methodology to define and assess DQ issues. These authors correlated specific DQ dimensions to possible technical issues, for example, data lacking standardization or valid entries may be caused by "errors from data originators, ETL issues or limitations of the EHR data entry tool (inadequate value set constrains, lack of DQ checks)" [88]. By measuring and analyzing these problems in a structured methodology, the team identified key areas that required targeted interventions, such as use of data standards to enforce data validation rules to the data entry system [88].

Table 5 summarizes the various intervention themes implemented to improve the quality of structured RWD in various health care contexts. It provides insights into common strategies in DQ improvement.

Reported Outcomes

To understand and compare the outcomes of DQ improvement initiatives, we identified whether studies reported DQ changes that were better, worse, or showed no change over time. Most studies reported improvements in DQ over time (36/39, 92%). This excludes 3 studies due to results being reported as preliminary [88], potential improvements as opposed to actual [92], or without sufficient detail [79]. Of the 36 studies showing improvement, 9 (23%) also report decreases in DQ [96,100,106,109,111], of which another 4 (10%) also report no changes [82,85,98]. These changes were reported at varying levels of granularity. While most studies reported DQ metrics for specific data items, such as validity of surgical procedure codes [96], others aggregated multiple metrics or dimensions into higher level entities, such as "92% reduction in error rate" [112] and mean monthly accuracy for pediatric early warning scores [75].

We also assessed whether statistical tests were used to demonstrate significance of effect and whether studies compared intervention groups with a control group. When determining significance of treatment effect, 22 studies used at least 1 statistical test or method [74-76,78,81-83,87,93,94,96,98-100,103-107,109,110]. These ranged from chi-square (7/39, 18%) and statistical process control charts (4/39, 10%) to multivariable linear (1/39, 3%) and logistic regression (1/39, 3%). Of these 22 studies, only 1 (5%) study compared the intervention group to a concurrent control group, which reported improvement in completeness of cancer stage data [87].

Discussion

Principal Findings

In this paper, we conducted a systematic literature review to understand the current practices in DQ improvement of structured RWD in a health care context. We found substantial heterogeneity in the approaches to definition, assessment, and interventions across the reviewed literature. The range of definitions for DQ concepts, quality improvement methodologies, and reported outcomes have made synthesis and comparison of the results challenging. In the following sections, we explore these 3 points in greater depth.

DQ Assessment

A key issue in the exploration of DQ is the lack of consensus on theoretical definitions for DQ assessment. Despite the existence of several DQ frameworks, there are no agreed recommendations or guidelines on which frameworks should be used or on how dimensions should be defined, measured, or used to understand real-world issues in data capture, processing, and utility for high-quality RWE generation. This has been demonstrated in a wealth of previous reviews on DQ theory [22,23,25,26,38,48], but to a lesser extent in a quality improvement context. In our review, we found that while some studies did in fact reference theoretical frameworks by Weiskopf and Weng [22] and Kahn et al [38], these account for <15% of all included studies. This indicates a severe lack of uptake of standardized DQ theory in the wider literature and explains the substantial variation and lack of consensus. In turn, the lack of

XSL•FO
RenderX

agreement and consistency makes it difficult to harness the true purpose of DQ assessment, which pertains to its ability to identify issues in real-world processes, behaviors, and resources. While some studies demonstrate qualitative correlations between DQ issues and underlying real-world problems [48], we found that only a small minority of studies implemented a quantitative approach to make a similar connection.

As DQ is a complex, multidimensional construct, each dimension serves to identify context-specific issues in the real-world needing remediation. We found that some authors made this correlation either directly or indirectly; for example, data validity is affected by a lack of standardization of front-end user interfaces on electronic data capture forms [96], timeliness of data indicates possible workflow inefficiencies that delays the point of data capture [74,104], duplication highlights redundant data sources [95], inaccurate data underpin lack of training on medical coding standards [103], and inconsistencies between data sources indicate possible capture of inaccurate data [83,86]. This raises 2 important points: the need to assess DQ beyond completeness or missingness and the importance of standardized frameworks. Without these, crucial error-prone processes in complex clinical pathways may go undiagnosed and continue to generate poor-quality data. This is particularly important given the growing demand for and expectations of real-world health care data, the hype in artificial intelligence, and the growing awareness that maintaining patient records is the leading cause of clinician burnout [42,113].

There was limited reporting of the tools or software used for DQ assessment included in this review. Only 2 DQ assessment tools were reported: the WHO DQ assessment toolkit [82,91,94] and the "Open Data Kit" [86]. Neither of these are explained in sufficient depth to discern how they work or their applicability to other environments. Further investigation into the referenced material also lacked sufficient information. Other tools reported were scheduled programmatic scripts using R, SQL, or SAS software for DQ assessment. Some of these methods are considered "automated" solutions for DQ assessment. This indicates a significant gap between the vast range of DQ software available and the practical implementation of these tools for DQ assessment, causal analysis, and improvement. DQ software must be capable of profiling large volumes of structured data, provide both automated and user specified DQ assessment methods, and facilitate meaningful analysis of possible root causes of poor DQ [21,30]. The limited adoption of existing DQ software might suggest a deficiency in technical proficiency, inadequate documentation clarifying its utility or use cases, or a lack of awareness regarding its availability or relevance.

Quality Improvement Cycles

Another characteristic of the studies included in this review was the limited use of quality improvement frameworks. Only 5 studies [74,81,84,88,101] referenced a quality improvement methodology to plan and implement DQ improvement interventions. This is surprising given the potential benefits these frameworks offer, particularly in fostering systematic, structured, and dynamic approaches to improvement in complex environments.

Quality improvement frameworks, such as PDSA, DMAIC, and TDQM, if implemented robustly, can significantly improve comparability and knowledge sharing between studies, institutions, and organizational teams. This is important given that Siegel et al [100] observed varying improvements across different organizations, stating that systematic and organized quality improvement efforts are needed.

However, the strengths of these frameworks extend beyond iterative learning; they also encourage a deep dive into DQ analysis, helping to unravel the complex relationships between various real-world factors and the root causes of poor DQ. In this way, interventions can be designed in collaboration with the affected stakeholders, that is, frontline clinical staff, to maximize the opportunity for DQ improvement. Knight et al [101] particularly emphasized this point, stating that the quality improvement model, that is, PDSA, "used in this project facilitated the identification and correction of difficulties with the technology of the innovation." Quality improvement frameworks, such as PDSA, TDQM, or DMAIC, can also be adapted to improve the quality of real-world health care data incorporating DQ-driven quantitative analysis alongside real-world issues that can be identified using the Odigos framework [48].

Despite their strengths, the application of these frameworks is not without challenges. One significant constraint is the need for substantial upfront planning and stakeholder engagement, which can be resource intensive. Furthermore, these frameworks require a culture of continuous improvement and openness to change and adoption of data governance practices, which may not be present in all health care settings. This can limit their applicability and effectiveness. In addition, the lack of consistent application and reporting on the use of these frameworks can make it difficult to evaluate their true effectiveness.

Outcomes

We sought to investigate the current approaches to DQ assessment and improvement to synthesize and summarize the lessons learnt from these endeavors. In general, studies reported positive changes in DQ through the implementation of multiple interventions. Lemma et al [69] associated the benefits to DQ when interventions such as training, technical innovation, and DQ feedback were combined. The same authors reported that studies that focused only on single interventions did not generate equally positive DQ changes. In contrast, we found that 17 studies focused on a single intervention showing mostly positive results. For example, Sinaiko et al [87] demonstrated the positive impact of peer comparison emails to completion of cancer stage data when compared to a control group. Similarly, studies that demonstrated a combination of improvements, reductions, and no changes in DQ often implemented multiple interventions.

Studies highlighted the importance of close collaboration with clinical users' needs when implementing digital technologies [101,108,112] and the importance of detailed, personalized feedback on data capture performance provided to data capture end users [87,107,109]. In contrast, Taggart et al [98] found that peer comparison and feedback sessions did not result in better DQ and suggested the need for randomized controlled studies. This raises an important question regarding the need

XSL•FO
RenderX

for control groups in DQ improvement studies. In our review, we found only 1 study that compared an intervention group with a concurrent control group, yet most report successful improvement in DQ.

While some studies highlighted the need for controlled environments, we observed that interventions were chosen for reasons other than baseline quantitative or qualitative analysis or analysis of underlying causes. In other words, only a few studies planned an intervention based on a data-driven approach. When comparing outcomes, most studies compared average baselines before and after the intervention, where only 1 study compared the intervention group to a concurrent nonintervention control group. A simultaneous control group for comparison can facilitate analysis of the cause of effect of treatment interventions along with the implementation of multiple interventions, which occurred in most studies. This, combined with the lack of data-driven approaches when planning or designing interventions, indicates a significant gap in robust, standardized DQ improvement methodologies. Therefore, the reported outcomes should be considered with caution.

The methodological rigor in DQ improvement studies often suffers from a lack of randomized controls and consistent statistical methodologies. As discussed before, only 1 reviewed study included a randomized control group to demonstrate significant improvement [87], and just 22 (56%) out of 39 studies applied diverse statistical tests, such as chi-square and logistic regression. This inconsistency in applying robust analytical techniques can introduce biases, misattributing improvements to interventions rather than actual effects. Moreover, the absence of uniform experimental designs across various health care settings undermines the robustness and generalizability of findings.

Incorporating structured methodologies such as PDSA or DMAIC could significantly enhance the methodological rigor of these studies. These frameworks support systematic implementations and evaluations, facilitating the use of control groups and statistical analysis to reliably isolate intervention effects. By adopting such standardized approaches, future research could more effectively ensure the credibility and applicability of the findings, fostering the development of evidence-based interventions suitable for diverse health care environments.

Future Recommendations

This review highlights the need for standardized and systematic approaches to DQ assessment, analysis, and improvement. This can be addressed in future studies by following quality improvement methodologies, such as the PDSA [114], TDQM [61], or DMAIC [63] iterative learning cycles, and DQ frameworks, such as DAMA [20], Weiskopf and Weng [22], or Kahn et al [38]. Furthermore, understanding the root causes of poor DQ is essential for planning the most appropriate intervention. This intervention should aim to address issues as close to the point of data capture as possible.

The need for standardized DQ assessment is evident. Future research and development should focus on the development and demonstration of DQ tools that are not only grounded in theoretical frameworks, such as those offered by DAMA [20], Weiskopf and Weng [22], or Kahn et al [38], but are also highly accessible and user-friendly. DQ tools should come with comprehensive documentation and practical examples that enable users in making informed decisions about their applicability and relevance in specific health care settings.

As discussed in the Introduction section, DQ tooling currently lack in usability and usefulness [21,29,30]. Ease of use can be overcome by introducing “plug-and-play” functionality that is combined with useful customizable features. This duality can allow users to quickly test and assess the tool’s immediate value and adjust and extend its functionality to fit more complex, specific needs over time. By extension, DQ tools should produce results that are useful for meaningful, in-depth analysis and monitoring of DQ errors.

Current best practices in root cause analysis of poor real-world health care data are unknown but could be facilitated using a framework such as the Odigos framework [48]. Furthermore, understanding causes of poor data can facilitate the design and selection of more relevant interventions needed—an aspect of DQ management that was demonstrated by few articles in this review. Future studies may also wish to compare intervention groups to concurrent control groups and explore 1 intervention at a time instead of multiple. This may help to control for external factors and increase understanding of barriers to high-quality data capture.

Limitations

This review aimed to summarize the lessons learnt from DQ improvement studies. As an abundance of literature already highlights the substantial variation in terminology for DQ concepts, we used the DAMA DQ framework to standardize the heterogeneity in DQ terms and definitions. In doing so, some DQ concepts could not be classified, potentially affecting the frequency counts of DQ dimensions assessed. Another limitation is that we were unable to perform a comprehensive meta-analysis of the methodological constrains and the effect measures of the reported outcomes. We believe the significant scope of this work warrants future research. This was due to considerable variation in the methods for assessment, analysis, and reporting of DQ metrics and changes over time. This posed significant challenges when attempting to objectively elucidate the effect of treatment interventions.

Moreover, while this review captures a rise in DQ improvement studies, with 31 (79%) out of 39 studies published in the last decade, it also includes 8 studies that were published in or before 2013, potentially missing recent advancements in digital health technologies. In addition, the discovery of 6 additional studies from manual searches indicates the likely exclusion of other relevant work. This is particularly the case due to the lack of consistency in DQ terminology and definitions, which made it difficult to capture all possible variations of DQ terms in the search strategy. Despite these challenges, our review included twice as many studies compared to other related reviews, indicating a thorough coverage within the constraints identified. We believe the significant scope of this work warrants future updates to include emerging trends and methodologies in DQ improvement.

XSL•FO
RenderX

Conclusions

The reviewed studies demonstrate that approaches to DQ improvement vary in their methodologies, definitions, and reporting of DQ dimensions. In general, studies implemented multiple interventions and reported positive changes in the quality of structured real-world health care data. In addition to "going paperless" initiatives, studies demonstrated the benefits of engagement with frontline clinical end users, provision of personalized DQ feedback, streamlining clinical workflows, and raising awareness of DQ and data standards aimed at improving DQ in health care settings. Despite this, heterogeneity is a major limitation among DQ literature in general, and we recommend that studies refer to standardized frameworks, such as PDSA cycles for quality improvement and the DAMA DQ framework for assessing DQ dimensions. This would lead to greater consistency and comparison in the reported outcomes.

Acknowledgments

This work was supported by Cancer Research UK RadNet Manchester (C1994/A28701). GP was supported by the NIHR (National Institute for Health and Care Research) Manchester Biomedical Research Centre (NIHR203308). The authors would like to thank Dr Catherine O'Hara, Head of Analytics from the Clinical Outcomes and Data Unit, for her support and guidance in this project, as well as Daniel Livesey, Librarian at The Christie NHS Foundation Trust, for his invaluable training and assistance on defining and implementing the literature search strategy.

Data Availability

All data generated or analyzed during this study are included in this manuscript and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 22 KB - jmir_v26i1e57615_app1.docx]

Multimedia Appendix 2
Search strategy and results.
[DOCX File , 16 KB - jmir_v26i1e57615_app2.docx]

References

1. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis 2023 Mar 23;82(3):306-311 [FREE Full text] [doi: 10.1136/ard-2022-222626] [Medline: 36150748]
2. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 05;22(1):287 [FREE Full text] [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
3. Real-world evidence. U.S. Food & Drug Administration. 2023 May 2. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2023-06-21]
4. Atasoy H, Greenwood BN, McCullough JS. The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization. Annu Rev Public Health 2019 Apr 01;40:487-500. [doi: 10.1146/annurev-publhealth-040218-044206] [Medline: 30566385]
5. Benson T. Why general practitioners use computers and hospital doctors do not--part 1: incentives. BMJ 2002 Nov 09;325(7372):1086-1089 [FREE Full text] [doi: 10.1136/bmj.325.7372.1086] [Medline: 12424171]
6. Anshari M. Redefining electronic health records (EHR) and electronic medical records (EMR) to promote patient empowerment. Int J Inform Dev 2019 Jun 22;8(1):35-39. [doi: 10.14421/ijid.2019.08106]
7. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and impact of real-world clinical data for the practicing clinician. Adv Ther 2018 Nov;35(11):1763-1774 [FREE Full text] [doi: 10.1007/s12325-018-0805-y] [Medline: 30357570]
8. Guinn D, Wilhelm EE, Lieberman G, Khozin S. Assessing function of electronic health records for real-world data generation. BMJ Evid Based Med 2019 Jun;24(3):95-98. [doi: 10.1136/bmjebm-2018-111111] [Medline: 30478146]
9. Kim HS, Lee S, Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. J Korean Med Sci 2018 Jun 26;33(34):e213 [FREE Full text] [doi: 10.3346/jkms.2018.33.e213] [Medline: 30127705]
10. Dang A. Real-world evidence: a primer. Pharmaceut Med 2023 Jan;37(1):25-36 [FREE Full text] [doi: 10.1007/s40290-022-00456-6] [Medline: 36604368]

XSL•FO
RenderX

11. NICE real-world evidence framework corporate document. National Institute for Health and Care Excellence. 2022 Jun 23. URL: https://www.nice.org.uk/corporate/ecd9/chapter/overview [accessed 2024-08-06]
12. Price G, Mackay R, Aznar M, McWilliam A, Johnson-Hart C, van Herk M, et al. Learning healthcare systems and rapid learning in radiation oncology: where are we and where are we going? Radiother Oncol 2021 Nov;164:183-195. [doi: 10.1016/j.radonc.2021.09.030] [Medline: 34619237]
13. Zhang J, Symons J, Agapow P, Teo JT, Paxton CA, Abdi J, et al. Best practices in the real-world data life cycle. PLOS Digit Health 2022 Jan 18;1(1):e0000003 [FREE Full text] [doi: 10.1371/journal.pdig.0000003] [Medline: 36812509]
14. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021 Sep 18;28(10):2251-2257 [FREE Full text] [doi: 10.1093/jamia/ocab132] [Medline: 34313749]
15. Thygesen JH, Tomlinson C, Hollings S, Mizani MA, Handy A, Akbari A, et al. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. Lancet Digit Health 2022 Jul;4(7):e542-e557 [FREE Full text] [doi: 10.1016/S2589-7500(22)00091-7] [Medline: 35690576]
16. NICE real-world evidence framework. National Institute for Health and Care Excellence. 2022 Jun 23. URL: https://www.nice.org.uk/corporate/ecd9/chapter/overview [accessed 2023-06-21]
17. Data quality framework for medicines regulation. European Medicines Agency. URL: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-quality-framework-medicines-regulation [accessed 2024-08-06]
18. MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions. United Kingdom Government. 2021 Dec 16. URL: https://tinyurl.com/2rkjjnfs [accessed 2023-10-09]
19. Castellanos EH, Wittmershaus BK, Chandwani S. Raising the bar for real-world data in oncology: approaches to quality across multiple dimensions. JCO Clin Cancer Inform 2024 Jan;8:e2300046 [FREE Full text] [doi: 10.1200/CCI.23.00046] [Medline: 38241599]
20. Earley S, Henderson D, Sebastian-Coleman L. The DAMA Guide to the Data Management Body of Knowledge (DAMA-DM BOK). Bradley Beach, NJ: Technics Publications, LLC; 2017.
21. Ozonze O, Scott PJ, Hopgood AA. Automating electronic health record data quality assessment. J Med Syst 2023 Feb 13;47(1):23 [FREE Full text] [doi: 10.1007/s10916-022-01892-2] [Medline: 36781551]
22. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
23. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013 Jan;82(1):10-24. [doi: 10.1016/j.ijmedinf.2012.10.001] [Medline: 23122633]
24. Cichy C, Rass S. An overview of data quality frameworks. IEEE Access 2019 Feb 15;7:24634-24648. [doi: 10.1109/access.2019.2899751]
25. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. Int J Environ Res Public Health 2014 May 14;11(5):5170-5207 [FREE Full text] [doi: 10.3390/ijerph110505170] [Medline: 24830450]
26. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc 2020 Dec 09;27(12):1999-2010 [FREE Full text] [doi: 10.1093/jamia/ocaa245] [Medline: 33166397]
27. Mariño J, Kasbohm E, Struckmann S, Kapsner LA, Schmidt CO. R packages for data quality assessments and data monitoring: a software scoping review with recommendations for future developments. Appl Sci 2022 Apr 22;12(9):4238. [doi: 10.3390/app12094238]
28. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615 [FREE Full text] [doi: 10.2196/42615] [Medline: 37000497]
29. Gordon B, Fennessy C, Varma S, Barrett J, McCondochie E, Heritage T, et al. Evaluation of freely available data profiling tools for health data research application: a functional evaluation review. BMJ Open 2022 May 09;12(5):e054186 [FREE Full text] [doi: 10.1136/bmjopen-2021-054186] [Medline: 35534084]
30. Ehrlinger L, Wöß W. A survey of data quality measurement and monitoring tools. Front Big Data 2022 Mar 31;5:850611 [FREE Full text] [doi: 10.3389/fdata.2022.850611] [Medline: 35434611]
31. Parry MG, Sujenthiran A, Cowling TE, Charman S, Nossiter J, Aggarwal A, et al. Imputation of missing prostate cancer stage in English cancer registry data based on clinical assumptions. Cancer Epidemiol 2019 Feb;58:44-51. [doi: 10.1016/j.canep.2018.11.003] [Medline: 30463041]
32. Westerberg M, Beckmann K, Gedeborg R, Irenaeus S, Holmberg L, Garmo H, et al. Choice of imputation method for missing metastatic status affected estimates of metastatic prostate cancer incidence. J Clin Epidemiol 2023 Mar;155:22-30 [FREE Full text] [doi: 10.1016/j.jclinepi.2022.12.008] [Medline: 36538980]
33. Daniels B, Havard A, Myton R, Lee C, Chidwick K. Evaluating the accuracy of data extracted from electronic health records into MedicineInsight, a national Australian general practice database. Int J Popul Data Sci 2022 Jun 29;7(1):1713 [FREE Full text] [doi: 10.23889/ijpds.v7i1.1713] [Medline: 37650032]

XSL•FO
RenderX

34. Palestine AG, Merrill PT, Saleem SM, Jabs DA, Thorne JE. Assessing the precision of ICD-10 codes for uveitis in 2 electronic health record systems. JAMA Ophthalmol 2018 Oct 01;136(10):1186-1190 [FREE Full text] [doi: 10.1001/jamaophthalmol.2018.3001] [Medline: 30054618]
35. Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. AMIA Annu Symp Proc 2018 Apr 16;2017:912-920 [FREE Full text] [Medline: 29854158]
36. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010 Mar 01;2010:1-5 [FREE Full text] [Medline: 21347133]
37. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc 2021 Dec 28;29(1):187-196 [FREE Full text] [doi: 10.1093/jamia/ocab199] [Medline: 34664641]
38. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
39. Bright CJ, Lawton S, Benson S, Bomb M, Dodwell D, Henson KE, et al. Data resource profile: the systemic anti-cancer therapy (SACT) dataset. Int J Epidemiol 2020 Feb 01;49(1):15-1l [FREE Full text] [doi: 10.1093/ije/dyz137] [Medline: 31340008]
40. Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, et al. The secondary use of electronic health records for data mining: data characteristics and challenges. ACM Comput Surv 2022 Jan 18;55(2):1-40. [doi: 10.1145/3490234]
41. Zhang J, Ashrafian H, Delaney B, Darzi A. Impact of primary to secondary care data sharing on care quality in NHS England hospitals. NPJ Digit Med 2023 Aug 14;6(1):144 [FREE Full text] [doi: 10.1038/s41746-023-00891-y] [Medline: 37580595]
42. Diaz-Garelli F, Strowd R, Ahmed T, Lycan TWJ, Daley S, Wells BJ, et al. What oncologists want: identifying challenges and preferences on diagnosis data entry to reduce EHR-induced burden and improve clinical data quality. JCO Clin Cancer Inform 2021 May;5:527-540 [FREE Full text] [doi: 10.1200/CCI.20.00174] [Medline: 33989015]
43. Ullah E, Albrett J, Khan O, Matthews C, Perry I, GholamHosseini H, et al. Workload involved in vital signs-based monitoring & responding to deteriorating patients: a single site experience from a regional New Zealand hospital. Heliyon 2022 Oct 06;8(10):e10955 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10955] [Medline: 36254295]
44. Pérez-Martí M, Casadó-Marín L, Guillén-Villar A. Electronic records with tablets at the point of care in an internal medicine unit: before-after time motion study. JMIR Hum Factors 2022 Feb 10;9(1):e30512 [FREE Full text] [doi: 10.2196/30512] [Medline: 35142624]
45. Byrd JB, Vigen R, Plomondon ME, Rumsfeld JS, Box TL, Fihn SD, et al. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: the VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. Am Heart J 2013 Mar;165(3):434-440. [doi: 10.1016/j.ahj.2012.12.009] [Medline: 23453115]
46. Skyttberg N, Vicente J, Chen R, Blomqvist H, Koch S. How to improve vital sign data quality for use in clinical decision support systems? A qualitative study in nine Swedish emergency departments. BMC Med Inform Decis Mak 2016 Jun 04;16:61 [FREE Full text] [doi: 10.1186/s12911-016-0305-4] [Medline: 27260476]
47. Dall'Ora C, Griffiths P, Hope J, Briggs J, Jeremy J, Gerry S, et al. How long do nursing staff take to measure and record patients' vital signs observations in hospital? A time-and-motion study. Int J Nurs Stud 2021 Jun;118:103921 [FREE Full text] [doi: 10.1016/j.ijnurstu.2021.103921] [Medline: 33812297]
48. Eden R, Syed R, Makasi T, Andrews R, Ter Hofstede A, Wynn M, et al. Revealing the root causes of digital health data quality issues: a qualitative investigation of the Odigos framework. In: Proceedings of the 56th Hawaii International Conference on System Sciences. 2023 Presented at: HICSS 2023; January 3-6, 2023; Maui, HI. [doi: 10.24251/hicss.2023.372]
49. Sockolow PS, Bowles KH, Adelsberger MC, Chittams JL, Liao C. Impact of homecare electronic health record on timeliness of clinical documentation, reimbursement, and patient outcomes. Appl Clin Inform 2014 Apr 30;5(2):445-462 [FREE Full text] [doi: 10.4338/ACI-2013-12-RA-0106] [Medline: 25024760]
50. Lee D, Jiang X, Yu H. Harmonized representation learning on dynamic EHR graphs. J Biomed Inform 2020 Jun;106:103426 [FREE Full text] [doi: 10.1016/j.jbi.2020.103426] [Medline: 32339747]
51. von Lucadou M, Ganslandt T, Prokosch HU, Toddenroth D. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak 2019 Oct 28;19(1):202 [FREE Full text] [doi: 10.1186/s12911-019-0939-0] [Medline: 31660955]
52. Zayas-Cabán T, Haque SN, Kemper N. Identifying opportunities for workflow automation in health care: lessons learned from other industries. Appl Clin Inform 2021 May;12(3):686-697 [FREE Full text] [doi: 10.1055/s-0041-1731744] [Medline: 34320683]
53. Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Systematic review, recommendations, and novel model for health IT collaboration. Appl Clin Inform 2017 Jan 11;8(1):12-34 [FREE Full text] [doi: 10.4338/ACI-2016-09-R-0150] [Medline: 28074211]
54. Quindroit P, Fruchart M, Degoul S, Perichon R, Martignène N, Soula J, et al. Definition of a practical taxonomy for referencing data quality problems in health care databases. Methods Inf Med 2023 May;62(1-02):19-30 [FREE Full text] [doi: 10.1055/a-1976-2371] [Medline: 36356592]

XSL•FO
RenderX

55. Liaw ST, Guo JG, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc 2021 Jul 14;28(7):1591-1599 [FREE Full text] [doi: 10.1093/jamia/ocaa340] [Medline: 33496785]
56. Shi X, Prins C, van Pottelbergh G, Mamouris P, Vaes B, de Moor B. An automated data cleaning method for electronic health records by incorporating clinical knowledge. BMC Med Inform Decis Mak 2021 Sep 17;21(1):267 [FREE Full text] [doi: 10.1186/s12911-021-01630-7] [Medline: 34535146]
57. Phan HT, Borca F, Cable D, Batchelor J, Davies JH, Ennis S. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. Sci Rep 2020 Jun 23;10(1):10164 [FREE Full text] [doi: 10.1038/s41598-020-66925-7] [Medline: 32576940]
58. Dziadkowiec O, Callahan T, Ozkaynak M, Reeder B, Welton J. Using a data quality framework to clean data extracted from the electronic health record: a case study. EGEMS (Wash DC) 2016 Jun 24;4(1):1201 [FREE Full text] [doi: 10.13063/2327-9214.1201] [Medline: 27429992]
59. Backhouse A, Ogunlayi F. Quality improvement into practice. BMJ 2020 Mar 31;368:m865 [FREE Full text] [doi: 10.1136/bmj.m865] [Medline: 32234777]
60. Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed JE. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. BMJ Qual Saf 2014 Apr;23(4):290-298 [FREE Full text] [doi: 10.1136/bmjqs-2013-001862] [Medline: 24025320]
61. Wang RY. A product perspective on total data quality management. Commun ACM 1998 Feb 01;41(2):58-65. [doi: 10.1145/269012.269022]
62. D'Andreamatteo A, Ianni L, Lega F, Sargiacomo M. Lean in healthcare: a comprehensive review. Health Policy 2015 Sep;119(9):1197-1209. [doi: 10.1016/j.healthpol.2015.02.002] [Medline: 25737260]
63. Sokovic M, Pavletic D, Pipan KK. Quality improvement methodologies – PDCA cycle, RADAR matrix, DMAIC and DFSS. J Achiev Mater Manuf Eng 2010;43(1):476-483.
64. Deming WE. Elementary Principles of the Statistical Control of Quality: A Series of Lectures. Tokyo, Japan: Nippon Kagaku Gijutsu Remmei; 1950.
65. Moen R, Norman C. Evolution of the PDCA cycle. In: Proceedings of the 7th ANQ Congress. 2009 Presented at: ANQ 2009; September 15-18, 2009; Tokyo, Japan URL: https://www.anforq.org/activities/congresses/index.html
66. Burns L, Roux NL, Kalesnik-Orszulak R, Christian J, Hukkelhoven M, Rockhold F, et al. Real-world evidence for regulatory decision-making: guidance from around the world. Clin Ther 2022 Mar;44(3):420-437 [FREE Full text] [doi: 10.1016/j.clinthera.2022.01.012] [Medline: 35181179]
67. Brouwer HJ, Bindels PJ, Weert HC. Data quality improvement in general practice. Fam Pract 2006 Oct;23(5):529-536. [doi: 10.1093/fampra/cml040] [Medline: 16868006]
68. Wiebe N, Otero Varela L, Niven DJ, Ronksley PE, Iragorri N, Quan H. Evaluation of interventions to improve inpatient hospital documentation within electronic health records: a systematic review. J Am Med Inform Assoc 2019 Nov 01;26(11):1389-1400 [FREE Full text] [doi: 10.1093/jamia/ocz081] [Medline: 31365092]
69. Lemma S, Janson A, Persson L, Wickremasinghe D, Källestål C. Improving quality and use of routine health information system data in low- and middle-income countries: a scoping review. PLoS One 2020 Oct 8;15(10):e0239683 [FREE Full text] [doi: 10.1371/journal.pone.0239683] [Medline: 33031406]
70. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
71. Carrera-Rivera A, Larrinaga F, Lasa G. Context-awareness for the design of smart-product service systems: literature review. Comput Ind 2022 Nov;142:103730. [doi: 10.1016/j.compind.2022.103730]
72. Yang H, Lee HJ. Research trend visualization by MeSH terms from PubMed. Int J Environ Res Public Health 2018 May 30;15(6):1113 [FREE Full text] [doi: 10.3390/ijerph15061113] [Medline: 29848974]
73. Rayyan homepage. Rayyan. URL: https://www.rayyan.ai/ [accessed 2023-06-30]
74. Larrow A, Chong A, Robison T, Patel A, Kuelbs C, Fisher E, et al. A quality improvement initiative to improve discharge timeliness and documentation. Pediatr Qual Saf 2021 Jul 28;6(4):e440 [FREE Full text] [doi: 10.1097/pq9.0000000000000440] [Medline: 34345753]
75. Dean NP, Cheng JJ, Crumbley I, DuVal J, Maldonado E, Ghebremariam E. Improving accuracy and timeliness of nursing documentation of pediatric early warning scores. Pediatr Qual Saf 2020 Mar 25;5(2):e278 [FREE Full text] [doi: 10.1097/pq9.0000000000000278] [Medline: 32426641]
76. Tuti T, Aluvaala J, Chelangat D, Mbevi G, Wainaina J, Mumelo L, Clinical Information Network Group, et al. Improving in-patient neonatal data quality as a pre-requisite for monitoring and improving quality of care at scale: a multisite retrospective cohort study in Kenya. PLOS Glob Public Health 2022 Oct 20;2(10):e0000673 [FREE Full text] [doi: 10.1371/journal.pgph.0000673] [Medline: 36962543]
77. Ma S, Lawpoolsri S, Soonthornworasiri N, Khamsiriwatchara A, Jandee K, Taweeseneepitch K, et al. Effectiveness of implementation of electronic malaria information system as the national malaria surveillance system in Thailand. JMIR Public Health Surveill 2016 May 06;2(1):e20 [FREE Full text] [doi: 10.2196/publichealth.5347] [Medline: 27227156]

XSL•FO
RenderX

78. Ng C, Payne AS, Patel AK, Thomas-Mohtat R, Maxwell A, Abo A. Improving point-of-care ultrasound documentation and billing accuracy in a pediatric emergency department. Pediatr Qual Saf 2020 Jul 21;5(4):e315 [FREE Full text] [doi: 10.1097/pq9.0000000000000315] [Medline: 32766490]
79. Pfaff E, Girvin A, Gabriel D, Kostka K, Morris M, Palchuk MB, N3C Consortium, et al. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. J Am Med Inform Assoc 2022 Mar 15;29(4):609-618 [FREE Full text] [doi: 10.1093/jamia/ocab217] [Medline: 34590684]
80. Kiogou SD, Chi CL, Zhang R, Ma S, Adam TJ. Clinical data cohort quality improvement: the case of the medication data in the University of Minnesota's clinical data repository. AMIA Jt Summits Transl Sci Proc 2022 May 23;2022:293-302 [FREE Full text] [Medline: 35854717]
81. Koo J, Moyer L, Castello MA, Arain Y. Improving accuracy of handoff by implementing an electronic health record-generated tool: an improvement project in an academic neonatal intensive care unit. Pediatr Qual Saf 2020 Jul 10;5(4):e329 [FREE Full text] [doi: 10.1097/pq9.0000000000000329] [Medline: 32766500]
82. Bhattacharya AA, Allen E, Umar N, Audu A, Felix H, Schellenberg J, et al. Improving the quality of routine maternal and newborn data captured in primary health facilities in Gombe State, Northeastern Nigeria: a before-and-after study. BMJ Open 2020 Dec 02;10(12):e038174 [FREE Full text] [doi: 10.1136/bmjopen-2020-038174] [Medline: 33268402]
83. Manesen R, Mekler KA, Molobi TR, Tyiki AA, Madlavu MJ, Velen K, et al. Data quality assessment of a South African electronic registry for drug-resistant TB, 2015-2016. Public Health Action 2021 Mar 21;11(1):33-39 [FREE Full text] [doi: 10.5588/pha.20.0031] [Medline: 33777719]
84. Olagundoye O, van Boven K, Daramola O, Njoku K, Omosun A. Improving the accuracy of ICD-10 coding of morbidity/mortality data through the introduction of an electronic diagnostic terminology tool at the general hospitals in Lagos, Nigeria. BMJ Open Qual 2021 Mar;10(1):e000938 [FREE Full text] [doi: 10.1136/bmjoq-2020-000938] [Medline: 33674344]
85. Moomba K, Williams A, Savory T, Lumpa M, Chilembo P, Tweya H, et al. Effects of real-time electronic data entry on HIV programme data quality in Lusaka, Zambia. Public Health Action 2020 Mar 21;10(1):47-52 [FREE Full text] [doi: 10.5588/pha.19.0068] [Medline: 32368524]
86. Tizifa TA, Nkhono W, Mtengula S, van Vugt M, Munn Z, Kabaghe AN. Leveraging phone-based mobile technology to improve data quality at health facilities in rural Malawi: a best practice project. Malar J 2021 Apr 27;20(1):203 [FREE Full text] [doi: 10.1186/s12936-021-03742-x] [Medline: 33906650]
87. Sinaiko AD, Barnett ML, Gaye M, Soriano M, Mulvey T, Hochberg E. Association of peer comparison emails with electronic health record documentation of cancer stage by oncologists. JAMA Netw Open 2020 Oct 01;3(10):e2015935 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15935] [Medline: 33021649]
88. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed 2019 Nov;181:104804. [doi: 10.1016/j.cmpb.2018.10.016] [Medline: 30497872]
89. Miyoshi NS, Azevedo-Marques JM, Alves D, Azevedo-Marques PM. An eHealth platform for the support of a Brazilian regional network of mental health care (eHealth-interop): development of an interoperability platform for mental care integration. JMIR Ment Health 2018 Dec 07;5(4):e10129 [FREE Full text] [doi: 10.2196/10129] [Medline: 30530455]
90. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, et al. Evaluating foundational data quality in the national patient-centered clinical research network (PCORnet®). EGEMS (Wash DC) 2018 Apr 13;6(1):3 [FREE Full text] [doi: 10.5334/egems.199] [Medline: 29881761]
91. Njuguna C, Vandi M, Mugagga M, Kanu J, Liyosi E, Chimbaru A, et al. Institutionalized data quality assessments: a critical pathway to improving the accuracy of integrated disease surveillance data in Sierra Leone. BMC Health Serv Res 2020 Aug 07;20(1):724 [FREE Full text] [doi: 10.1186/s12913-020-05591-x] [Medline: 32767983]
92. Edgerton JR, Herbert MA, Hamman BL, Ring WS. Can use of an administrative database improve accuracy of hospital-reported readmission rates? J Thorac Cardiovasc Surg 2018 May;155(5):2043-2047 [FREE Full text] [doi: 10.1016/j.jtcvs.2017.11.071] [Medline: 29329802]
93. Qin Y, Zhou R, Wu Q, Huang X, Chen X, Wang W, et al. The effect of nursing participation in the design of a critical care information system: a case study in a Chinese hospital. BMC Med Inform Decis Mak 2017 Dec 06;17(1):165 [FREE Full text] [doi: 10.1186/s12911-017-0569-3] [Medline: 29212480]
94. Muthee V, Bochner AF, Osterman A, Liku N, Akhwale W, Kwach J, et al. The impact of routine data quality assessments on electronic medical record data quality in Kenya. PLoS One 2018 Apr 18;13(4):e0195362 [FREE Full text] [doi: 10.1371/journal.pone.0195362] [Medline: 29668691]
95. Smith BL, Ramadan M, Corley B, Hablas A, Seifeldein IA, Soliman AS. Measuring the effect of improvement in methodological techniques on data collection in the Gharbiah population-based cancer registry in Egypt: implications for other low- and middle-income countries. Cancer Epidemiol 2015 Dec;39(6):1010-1014 [FREE Full text] [doi: 10.1016/j.canep.2015.11.001] [Medline: 26590335]
96. Ahlbrandt J, Henrich M, Hartmann BA, Bundschuh BB, Schwarz J, Klasen J, et al. Small cause - big effect: improvement in interface design results in improved data quality - a multicenter crossover study. Stud Health Technol Inform 2012;180:393-397. [Medline: 22874219]

XSL•FO
RenderX

97. Tuti T, Bitok M, Malla L, Paton C, Muinga N, Gathara D, et al. Improving documentation of clinical care within a clinical information network: an essential initial step in efforts to understand and improve care in Kenyan hospitals. BMJ Glob Health 2016 May 24;1(1):e000028 [FREE Full text] [doi: 10.1136/bmjgh-2016-000028] [Medline: 27398232]
98. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. Int J Med Inform 2015 Dec;84(12):1094-1098. [doi: 10.1016/j.ijmedinf.2015.09.008] [Medline: 26480872]
99. Haskew J, Rø G, Saito K, Turner K, Odhiambo G, Wamae A, et al. Implementation of a cloud-based electronic medical record for maternal and child health in rural Kenya. Int J Med Inform 2015 May;84(5):349-354 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.01.005] [Medline: 25670229]
100. Siegel EM, Jacobsen PB, Lee JH, Malafa M, Fulp W, Fletcher M, et al. Florida Initiative for Quality Cancer Care: improvements on colorectal cancer quality of care indicators during a 3-year interval. J Am Coll Surg 2014 Jan;218(1):16-25.e1 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2013.09.008] [Medline: 24275073]
101. Knight AW, Szucs C, Dhillon M, Lembke T, Mitchell C. The eCollaborative: using a quality improvement collaborative to implement the National eHealth Record System in Australian primary care practices. Int J Qual Health Care 2014 Aug;26(4):411-417 [FREE Full text] [doi: 10.1093/intqhc/mzu059] [Medline: 24925685]
102. Soto M, Capurro D, Catalán S. Evaluating the data completeness in the electronic health record after the implementation of an outpatient electronic health record. Stud Health Technol Inform 2015;216:885. [Medline: 26262187]
103. Ewing M, Funk GA, Warren AM, Rapier N, Reynolds M, Bennett M, et al. Improving National Trauma Data Bank® coding data reliability for traumatic injury using a prospective systems approach. Health Informatics J 2016 Dec;22(4):1076-1082 [FREE Full text] [doi: 10.1177/1460458215610896] [Medline: 26516133]
104. Genet IC, Firestone KS, Volsko TA. Neonatal respiratory therapist-led rounds can improve staff satisfaction and timeliness of respiratory interventions. Respir Care 2015 Mar;60(3):321-327. [doi: 10.4187/respcare.03489] [Medline: 25389357]
105. Nassaralla CL, Naessens JM, Hunt VL, Bhagra A, Chaudhry R, Hansen MA, et al. Medication reconciliation in ambulatory care: attempts at improvement. Qual Saf Health Care 2009 Oct;18(5):402-407. [doi: 10.1136/qshc.2007.024513] [Medline: 19812105]
106. Rahbar MH, Gonzales NR, Ardjomand-Hessabi M, Tahanan A, Sline MR, Peng H, et al. The University of Texas Houston Stroke Registry (UTHSR): implementation of enhanced data quality assurance procedures improves data quality. BMC Neurol 2013 Jun 15;13:61 [FREE Full text] [doi: 10.1186/1471-2377-13-61] [Medline: 23767957]
107. Greiver CD, Barnsley J, Aliarzadeh B, Krueger P, Moineddin R, Butt DA, et al. Using a data entry clerk to improve data quality in primary care electronic medical records: a pilot study. Inform Prim Care 2011;19(4):241-250 [FREE Full text] [doi: 10.14236/jhi.v19i4.819] [Medline: 22828579]
108. Wallace CJ, Stansfield D, Gibb Ellis KA, Clemmer TP. Implementation of an electronic logbook for intensive care units. Proc AMIA Symp 2002:840-844 [FREE Full text] [Medline: 12463943]
109. de Lusignan S, Stephens PN, Adal N, Majeed A. Does feedback improve the quality of computerized medical records in primary care? J Am Med Inform Assoc 2002;9(4):395-401 [FREE Full text] [doi: 10.1197/jamia.m1023] [Medline: 12087120]
110. Mphatswe W, Mate KS, Bennett B, Ngidi H, Reddy J, Barker PM, et al. Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa. Bull World Health Organ 2012 Mar 01;90(3):176-182 [FREE Full text] [doi: 10.2471/BLT.11.092759] [Medline: 22461712]
111. Benard M, Bouaud J, Marsault C, Boudghene F, Carette MF, Séroussi B. Diagnostic imaging requisition quality when using an electronic medical record: a before-after study. Stud Health Technol Inform 2015;210:364-368. [Medline: 25991167]
112. Amoroso CL, Akimana B, Wise B, Fraser HS. Using electronic medical records for HIV care in rural Rwanda. Stud Health Technol Inform 2010;160(Pt 1):337-341. [Medline: 20841704]
113. Johnson KB, Neuss MJ, Detmer DE. Electronic health records and clinician burnout: a story of three eras. J Am Med Inform Assoc 2021 Apr 23;28(5):967-973 [FREE Full text] [doi: 10.1093/jamia/ocaa274] [Medline: 33367815]
114. Christoff P. Running PDSA cycles. Curr Probl Pediatr Adolesc Health Care 2018 Aug;48(8):198-201. [doi: 10.1016/j.cppeds.2018.08.006] [Medline: 30270135]

Abbreviations

DAMA: Data Management Association
DMAIC: define-measure-analyze-improve-control
DQ: data quality
EHR: electronic health record
EMA: European Medicines Agency
EMR: electronic medical record
FDA: Food and Drug Administration
HIS: hospital information system
HL7: Health Language 7
ICD: International Classification of Diseases
ICPM: International Classification of Procedures in Medicine

XSL•FO
RenderX

LHS: learning health care system
MeSH: Medical Subject Headings
MHRA: Medicines and Healthcare Products Regulatory Agency
NICE: National Institute for Health and Care Excellence
OPS: Operationen und Prozedurenschlüssel
PACS: picture archiving and communication system
PDQI-9: Physician Documentation Quality Instrument
PDQI-9: Physician Documentation Quality Instrument
PDSA: plan-do-study-act
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RWD: real-world data
RWE: real-world evidence
SQL: structured query language
TDQM: total data quality management
WHO: World Health Organization

Edited by A Mavragani; submitted 11.03.24; peer-reviewed by A Hassan, T Khinvasara, A Wani; comments to author 26.04.24; revised version received 10.05.24; accepted 30.05.24; published 22.08.24.

Please cite as:
Lighterness A, Adcock M, Scanlon LA, Price G
Data Quality–Driven Improvement in Health Care: Systematic Literature Review
J Med Internet Res 2024;26:e57615
URL: https://www.jmir.org/2024/1/e57615
doi:10.2196/57615
PMID:

©Anthony Lighterness, Michael Adcock, Lauren Abigail Scanlon, Gareth Price. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Automated Interpretation of Lung Sounds by Deep Learning in Children With Asthma: Scoping Review and Strengths, Weaknesses, Opportunities, and Threats Analysis

Isabelle Ruchonnet-Métrailler[1,2*], MD, PD, PhD; Johan N Siebert[2,3*], MD, PD; Mary-Anne Hartley[4,5], MPH, Prof Dr, MD, PhD; Laurence Lacroix[2,3], MD, PD

[1]Pediatric Pulmonology Unit, Department of Pediatrics, Geneva Children's Hospital, University Hospitals of Geneva, Geneva, Switzerland
[2]Faculty of Medicine, University of Geneva, Geneva, Switzerland
[3]Division of Pediatric Emergency Medicine, Department of Pediatrics, Geneva Children's Hospital, Geneva University Hospitals, Geneva, Switzerland
[4]Intelligent Global Health Research Group, Machine Learning and Optimization Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland
[5]Laboratory of Intelligent Global Health Technologies, Bioinformatics and Data Science, Yale School of Medicine, New Haven, CT, United States
[*]these authors contributed equally

Corresponding Author:
Johan N Siebert, MD, PD
Division of Pediatric Emergency Medicine, Department of Pediatrics
Geneva Children's Hospital
Geneva University Hospitals
47, Avenue de la Roseraie
Geneva, 1205
Switzerland
Phone: 41 (0)795534072
Email: Johan.Siebert@hug.ch

Abstract

Background: The interpretation of lung sounds plays a crucial role in the appropriate diagnosis and management of pediatric asthma. Applying artificial intelligence (AI) to this task has the potential to better standardize assessment and may even improve its predictive potential.

Objective: This study aims to objectively review the literature on AI-assisted lung auscultation for pediatric asthma and provide a balanced assessment of its strengths, weaknesses, opportunities, and threats.

Methods: A scoping review on AI-assisted lung sound analysis in children with asthma was conducted across 4 major scientific databases (PubMed, MEDLINE Ovid, Embase, and Web of Science), supplemented by a gray literature search on Google Scholar, to identify relevant studies published from January 1, 2000, until May 23, 2023. The search strategy incorporated a combination of keywords related to AI, pulmonary auscultation, children, and asthma. The quality of eligible studies was assessed using the ChAMAI (Checklist for the Assessment of Medical Artificial Intelligence).

Results: The search identified 7 relevant studies out of 82 (9%) to be included through an academic literature search, while 11 of 250 (4.4%) studies from the gray literature search were considered but not included in the subsequent review and quality assessment. All had poor to medium ChAMAI scores, mostly due to the absence of external validation. Identified strengths were improved predictive accuracy of AI to allow for prompt and early diagnosis, personalized management strategies, and remote monitoring capabilities. Weaknesses were the heterogeneity between studies and the lack of standardization in data collection and interpretation. Opportunities were the potential of coordinated surveillance, growing data sets, and new ways of collaboratively learning from distributed data. Threats were both generic for the field of medical AI (loss of interpretability) but also specific to the use case, as clinicians might lose the skill of auscultation.

Conclusions: To achieve the opportunities of automated lung auscultation, there is a need to address weaknesses and threats with large-scale coordinated data collection in globally representative populations and leveraging new approaches to collaborative learning.

(J Med Internet Res 2024;26:e53662) doi:10.2196/53662

XSL•FO
RenderX

KEYWORDS

asthma; wheezing disorders; artificial intelligence; deep learning; machine learning; respiratory sounds; auscultation; stethoscope; pediatric; mobile phone

Introduction

Background

Asthma is one of the most common chronic pediatric conditions in the world, affecting almost 10% of children and causing significant morbidity and mortality [1], as well as putting a considerable burden on health care systems [1]. It is the third leading cause of admission among children aged <15 years [2]. In total, 7% of children with asthma (aged <18 years) require hospital admission each year due to exacerbation episodes, 47% of children have sleep disturbances or show limitation in daily activities [3], and 25% of children report increased school absences [4]. Mismanagement of the condition can cause long-lasting pulmonary sequelae with decreased lung function persisting into adulthood [5].

Asthma management is often delegated to patients and caregivers, who estimate treatment requirements according to their subjective understanding of symptom severity. The presence of various inevitable confounding factors may lead to poorly adapted and inequitable treatment. Although a written, symptom-based action plan for patients and caregivers is recommended to improve asthma control and avoid hospitalization, it is not sufficient to significantly reduce the risk for subsequent pediatric emergency department visits [6]. Careful monitoring and management of acute exacerbation episodes and long-term treatment are therefore essential to decrease morbidity and mortality and to improve the quality of life of children with asthma and their families.

Since the invention of the stethoscope by René Laennec in 1816, pulmonary auscultation has become a fundamental part of routine physical examination. However, the interpretation of sounds is ultimately limited by the human capacity for auditory perception and cognitive pattern detection. Recent efforts to digitalize auscultation audio allow the recording of sounds that can then be exploited by various software to assist interpretation [7]. A particular advantage is that recordings capture a higher resolution and wider range of frequencies, which are not usually perceptible to the average (and especially, the aging) human ear (ie, outside 20-20,000 Hz). The digital information obtained from the sounds recorded by electronic auscultation can, in turn, inform computer software, aiming to overcome standard lung auscultation limitations by removing subjectivity, standardizing interpretation through objective, nuanced acoustic pattern recognition, and automating the identification of abnormal auscultations indicating respiratory diseases. Recent advances in signal processing and computational efficiency offer the promise of accessible high-performance analysis for many physiological measurements [7]. In particular, deep learning (commonly referred to as artificial intelligence [AI]) opens up new perspectives to improve diagnostic accuracy and enable the automated monitoring of respiratory diseases by nonphysicians [8,9]. Studies showing this potential have mostly focused on adults [10]. The few studies that focused on pediatric auscultation [11-13] highlighted the high potential impact of standardizing the identification and assessment of obstructive pulmonary sounds, which are known reliable estimators of the severity and treatment response in asthma and wheezing disorders. Promising results have emerged in the automated detection of wheezing disorders and pneumonia [12,13]. Even early iterations in 2008 showed automated recognition of wheezing to be superior to traditional lung auscultation performed by untrained or trained medical staff in outpatient or intensive care units [14]. Beyond improved diagnosis and management, there is the potential to decentralize lung auscultation to home-based monitoring and remote care for earlier detection of asthma exacerbations, potentially reducing morbidity and health care costs [15].

Objective

The purpose of this scoping review is to assess the state of the art of the research conducted in AI applied to digital lung sound analysis in children with asthma and provide a balanced assessment of its strengths, weaknesses, opportunities, and threats (SWOT). To our knowledge, this is the first such review and critical assessment on the topic.

Methods

Review Question

On the basis of the Population targeted, Concept under investigation, Context parameters framework [16], the review question driving the search strategy was “How has AI-driven automated lung sound analysis been used to detect or classify asthma and wheezing disorders in children over the past 23 years (2000-2023), considering its strengths, weaknesses, opportunities, and threats?” (Multimedia Appendix 1).

Information Sources and Search Strategy

The review complies with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [17]. An electronic literature search was performed independently by 2 of the authors with expertise in pediatric emergency medicine and pulmonology (JNS and IR-M). The search was initially focused on academic literature in PubMed, MEDLINE Ovid, Embase, and Web of Science databases, with the date range restricted between January 1, 2000, and May 23. In addition, to broaden the scope, gray literature was sought using Google Scholar with the search limited to the first 10 pages of results (ie, first 250 occurrences) [18,19]. Search terms were chosen to filter out articles on AI-based lung sound analysis for the detection of asthma in humans. The following medical subject headings were explored and connected by Boolean operators with adjustments in syntax adapted to individual databases: artificial intelligence; machine learning; deep learning; neural networks, computer; respiratory sounds; asthma; wheezing disorder. The articles retrieved were further refined for pediatric studies using the following

XSL•FO
RenderX

keywords: “adolescent” OR “children” OR “child, preschool” [20] OR “infant” (Table 1).

Table 1. Search strategy.

Database and filter	Articles, n
PubMed	
Artificial intelligence OR machine learning OR deep learning OR neural networks, computer	262,346
AND respiratory sounds	283
AND (asthma OR wheezing disorder)	147
AND (adolescent OR children OR child, preschool OR infant)	32
MEDLINE Ovid	
(Artificial intelligence or machine learning or deep learning or neural networks, computer).mp	94,359
AND respiratory sounds.mp	132
AND (asthma OR wheezing disorder).mp	25
AND (adolescent OR children OR child, preschool OR infant).mp	16
Embase	
(“artificial intelligence”/exp OR “artificial intelligence” OR (artificial AND (“intelligence”/exp OR intelligence)) OR “machine learning”/exp OR “machine learning” OR (“machine”/exp OR machine) AND (“learning”/exp OR learning)) OR “deepl learning” OR (deepl AND (“learning”/exp OR learning)) OR “neural networks, computer”/exp OR “'neural networks, computer’” OR (neural AND networks, AND (“computer”/exp OR computer))) AND [humans]/lim AND [english]/lim AND [2000-2023]/py	154,919
AND respiratory sounds	222
AND (asthma OR wheezing disorder)	43
AND (adolescent OR children OR child, preschool OR infant)	13
Web of Science	
TS=(Artificial intelligence OR machine learning OR deep learning OR neural networks, computer) and DOP=(200-01-01/2023-05-23) AND LA=(English)	431,019
AND TS=(respiratory sounds)	183
AND TS=(asthma OR wheezing disorder)	35
AND TS=(adolescent OR children OR child, preschool OR infant)	11
Google Scholar (gray literature)	
Artificial intelligence OR machine learning OR deep learning OR neural networks	129,000
AND respiratory sounds	18,800
AND (asthma OR wheezing disorder)	15,300
AND (adolescent OR children OR infant)-cough	8720
Initial occurrences[a]	250

[a]Consistent with recommendations to prioritize the first 200 to 300 results in article titles when searching for gray literature [18], the initial 250 occurrences were retained.

Eligibility Criteria

Identified articles were submitted to inclusion criteria for eligibility: articles should have described automated lung sound analysis in children from birth to 18 years of age with asthma (Textbox 1). Articles should have been written in English. Articles were excluded if the study was conducted on animals. Studies concerning cough rather than lung sounds were excluded. Comments on articles and unpublished work were not considered suitable and were excluded from the review.

XSL•FO
RenderX

Textbox 1. Population, Intervention, Comparator, Outcome, and Study design (PICOS framework) eligibility criteria for the scoping review.

Inclusion criteria

- Population
  - Infants, children, and adolescents aged <18 years
  - Asthma or wheezing disorder
- Intervention
  - Digital auscultation coupled with automated lung sound analysis using artificial intelligence to detect or classify asthma
  - Audio files specifically collected for the purpose of the study or from a publicly available database
- Comparator
  - Asthma lung sounds
- Outcome
  - Predictive measures of performance
- Study design
  - Research articles
- Period
  - January 1, 2000, to May 23, 2023
- Language
  - English

Exclusion criteria

- Population
  - Adult population (aged ≥18 years)
  - Other diseases causing abnormal respiratory sounds
- Intervention
  - No artificial intelligence used
  - Audio files retrieved solely from lung sound databases, which exhibit multiple acquisition flaws inducing systematic biases between predicted labels (eg, disparate diagnoses sourced from various locations, systematically diverse age groups, or using different stethoscopes)
- Comparator
  - Normal or pathological lung sounds other than those associated with asthma
  - Cough or voice sounds
- Outcome
  - No performance metric report
- Study design
  - Abstracts
  - Book chapters
  - Communications
  - Commentaries to articles
  - Conference papers
  - Letters to the editor
  - Study protocols
  - Thesis

XSL•FO
RenderX

- Unpublished work
- Period
  - Outside date range
- Language
  - Other than English

Selection of Sources of Evidence

Academic Literature Search

The articles retrieved from the 4 academic repositories were merged, and duplicates were removed. Any relevant article with a title or abstract including the search terms was independently screened by the same 2 reviewers, who completed the database search process. Disagreements were resolved by consensus with a third reviewer (LL). The reference lists of eligible articles were reviewed to find additional relevant articles and the related papers obtained. This procedure was repeated until the reference lists covered the complete set of articles already retrieved.

Gray Literature Search

Gray literature was screened by one of the reviewers (JNS). Given the inherent content variability and nonpeer-reviewed status of the sources included, gray literature search was excluded from the subsequent quality assessment. This decision was made to maintain consistency in the assessment process and ensure alignment with established quality evaluation criteria applied to peer-reviewed literature. In addition, the reference lists of the identified articles were not reviewed.

Data Charting

Data from the selected articles obtained through academic literature search were extracted by one of the authors (JNS) in the following structured format: publication reference; number of participants and conditions when available, age, number of recordings; study design; number of classes; methodology; recording device; AI model architecture used; findings and performance (metrics) of the algorithm; advantages and limitations of the study; and sharing algorithm code or data as open source. As for the articles from the gray literature, they are provided in the following format: publication reference; article type; and a brief summary of the article.

Critical Appraisal and Data Synthesis

Quality assessment of selected studies from the academic literature search was performed by JNS using the Checklist for the Assessment of Medical Artificial Intelligence (ChAMAI) [21] and then reviewed by LL. An agreement was reached through discussion. The ChAMAI is a 30-item practical assessment-oriented checklist that is recommended to be reported for medical AI studies to encourage reproducible results and evaluation. It aims to distinguish high-quality medical AI studies from the simple application of AI techniques to medical data. The checklist is organized into six dimensions according to the Cross Industry Standard Process for Data Mining methodology [22]: (1) problem understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) validation, and (6) deployment. Each item can be rated as OK (adequately addressed), minor revision (mR; sufficient but requiring mR to improve), or major revision (MR; insufficiently addressed and necessarily requiring MR or justifying rejection of the manuscript). A total of 20 (67%) of the 30 high-priority items can be assigned a value of 2, 1, and 0 depending on their respective rating (OK, mR, or MR). For the remaining 10 (30%) low-priority items, the values are halved (ie, 1, 0.5, and 0). The maximum score is 50 points. Study quality was classified as low (0-19.5), medium (20-34.5), or high (35-50) [23].

SWOT Analysis

A SWOT analysis categorizes factors describing a research subject into 4 distinct groups: strengths, weaknesses, opportunities, and threats. Strengths and weaknesses are internal factors controlled or influenced by the entity itself, such as its resources, capabilities, and organizational structure, while opportunities and threats are external factors beyond the entity's direct control that can significantly impact its activity [24]. Although primarily used in organizational studies [25], we used SWOT as a structured analysis framework to conduct a balanced assessment of the SWOT associated with AI-assisted lung auscultation for pediatric asthma, drawing from the academic literature retrieved in this scoping review.

Results

Selection of Sources of Evidence

The search strategy selected 82 articles from academic database searches, of which 64 (78%) were only screened, and 18 (22%) were assessed for full-text eligibility. In addition, 250 records were retrieved from gray literature search, with 183 (73%) assessed for eligibility, resulting in a total of 201 unique records (Figure 1). Of the 46 articles not included at the screening stage, 9 (20%) pertained to AI-based asthma or wheezing recognition algorithms using cough sounds [26-34]. Of the 18 studies, 11 (61%) identified through the academic literature search were excluded due to ineligible study design (n=4, 36% in adults; n=4, 36% not relying on AI or lung sounds; and n=3, 27% other study types). Among them were 3 quasi-experimental studies [35-37] focused on computerized wheeze detection through respiratory spectrum analysis in children with asthma, not involving the support of AI algorithms. Along with records from the gray literature search, a total of 9% (7/82) and 3.6% (9/250) of studies were respectively included in the scoping review.

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines extension for Scoping Reviews) flowchart depicting included studies from academic databases and gray literature.

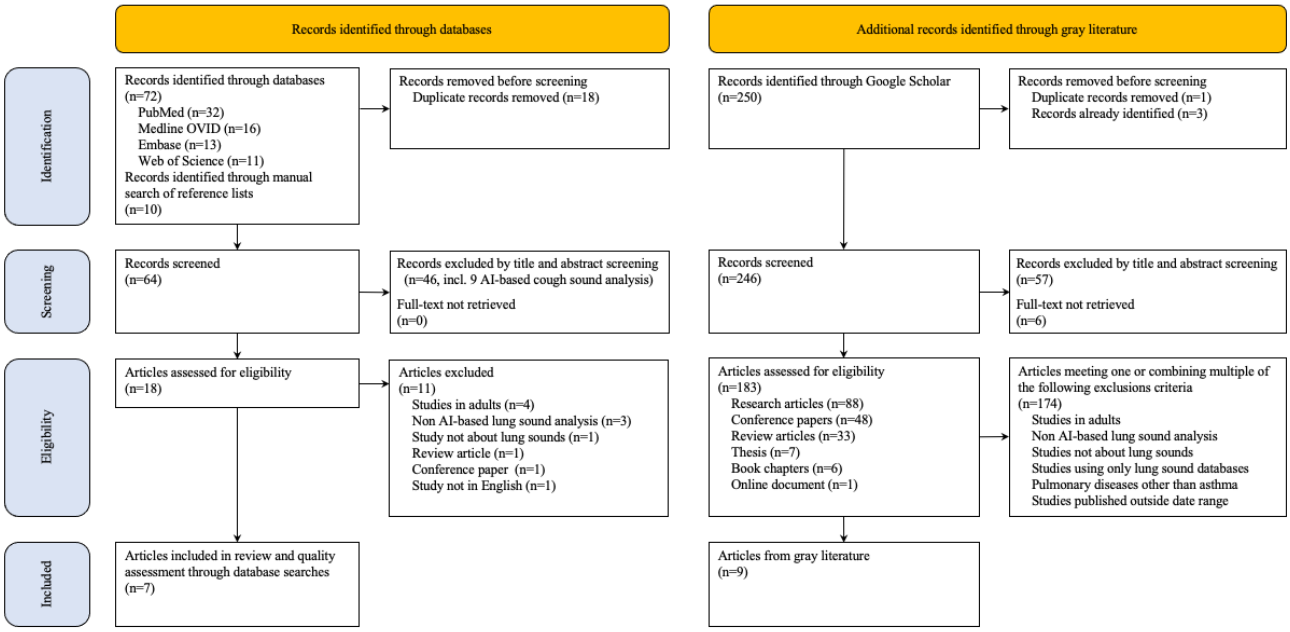

Characteristics of Individual Sources of Evidence

The characteristics of the 7 studies included from the academic literature search are presented in Table 2. They were conducted between 2015, and 2022, across 8 different countries (Croatia, Poland, Australia, China, Singapore, the Republic of Korea, and Russia). With the exception of the study by Mazić et al [38], which was performed in a general hospital, all AI models were implemented in children's hospitals, either in respiratory clinics or inpatient wards. Study participants varied in number, ranging from 16 to 112. The participant count was not specified in 1 study by Gelman et al [39], involving individuals aged 28 days to 18 years. The design of the studies was also variable, with 4 quasi-experimental studies and 3 prospective cohort studies. Lung sounds were collected using a variety of custom-built [38,40] and commercial (StethoMe, Littmann 3200, CliniCloud, Yunting model II, and Jabes) digital stethoscopes (including a mix of various stethoscopes in 1 of the studies) [11,12,41,42], or with built-in or external cell phone microphones [39]. Although most of the studies used patients' lung sounds under realistic conditions in their AI models, the study by Cheng et al [40] also incorporated breath sounds from a commercially available database.

XSL•FO
RenderX

Table 2. Characteristics of individual sources of evidence from academic literature search.

Study	Participants, N[a] (age; recordings, n)	Study design	Classes, n	Methodology	Recording device	AI[b] model architecture	Findings, metrics	Advantages and limitations	Open-source
Mazić et al [38]	16 with obstructive bronchitis or asthma (1-6 y, 45)	Quasi-experimental	2: wheezes and non-wheezes	AI-based binary classifier development using recordings from hospital visits in Croatia under realistic conditions; no lung sound database	Accelerometer	2 cascaded SVM[c] classifiers	A 2-layer parallel-stacked SVM classifier enhances the standard SVM classifier with high reliability of wheezing recognition: accuracy 97.7%	• Advantages: cascade classifier increases accuracy • Limitations: no gold standard. Validation of system not reported. No wheezing in the testing set. A small number of patients. No healthy subjects. No tabular data	No code and no data
Grzywalski et al [11]	50 (1-18 y; 522 total: 124 wheezes, 113 rhonchi, 66 coarse crackles, and 112 fine crackles)	Prospective cohort study	4: wheezes, rhonchi, and fine and coarse crackles	Recordings of Polish inpatients under realistic conditions. No lung sound database. Five pediatricians classified recordings as gold standard. Comparison: 5 pediatricians versus AI in detecting 4 pathological sound types	Digital stethoscope (Littmann3200)	CNNs[d] and StethoMe AI[e]	The AI model slightly outperforms pediatrician auscultations: accuracy 57.7% versus 66.1%, sensibility 78.2% versus 58.1%, specificity 82.2% versus 90.7%, and F_1-score 66.4% versus 61.8%	• Advantages: establishment of a gold standard. Varied pediatric age range • Limitations: human labeling subjectivity. Training and validation set undisclosed. No healthy participants. No tabular data	No code and no data
Kevat et al [12]	25 (median 6.7 y; 192 total: 40 wheezes, 39 wheezes+crackles, 46 cystic fibrosis unspecified sounds, and 67 normal breath sounds)	Quasi-experimental	3: wheezes, crackles, and normal breath sounds	Recordings of Australian inpatients in Australia under realistic conditions. No lung sound database. Classified by 1 pediatric pulmonologist	Digital stethoscopes (Clinicloud and Littman 3200)	CNNs and StethoMe AI	The AI model consistently identified pediatric breath sounds across various devices, while variations based on device type may be present: Clinicloud: sensibility 90% and specificity 97%; Littman: sensibility 80% and specificity 97%	• Advantages: AI blinded to tagged recordings • Limitations: no true gold standard. Human labeling subjectivity. Age range of participants not reported. Training and validation sets undisclosed. Reliance on 2 stethoscopes. No tabular data	No code, No data

XSL•FO
RenderX

Study	Participants, N[a] (age; recordings, n)	Study design	Classes, n	Methodology	Recording device	AI[b] model architecture	Findings, metrics	Advantages and limitations	Open-source
Zhang et al [41]	112, including 5 with asthma, 15 with bronchitis, 15 with bronchiolitis, 1 with bronchiolitis obliterans, 75 with pneumonia, 1 with foreign body aspiration (28 d-18 y; 627 total: 204 wheezes, 159 crackles, and 264 normal breath sounds)	Quasi-experimental	3: wheezes, crackles, and normal breath sounds	Recordings of Chinese inpatients under realistic conditions. No lung sound database. Classified by 2 pediatric pulmonologists as gold standard. Comparison: 6 general pediatricians with over 5 years of experience versus AI in detecting 3 breath sound types	Digital stethoscope (Yunting model II)	SVM	The AI model outperforms general pediatrician auscultations: accuracy 76% versus 64.9%, sensibility 86.4% versus 82.2%, specificity 83% versus 72.1%, and F_1-score 80.9% versus 72.5%	• Advantages: establishment of a gold standard. Training and validation set disclosed. Varied pediatric age range • Limitations: human labeling subjectivity. Few breath sounds of children aged >60 months. Cannot detect crackles and wheezes simultaneously, impacting diagnosis. No tabular data	No code and no data[f]
Cheng et al [40]	73, including 19 with viral wheeze or asthma exacerbation, 17 with well-controlled asthma, 2 with exacerbations of chronic lung disease, 8 with pneumonia, 3 with other conditions, 24 healthy children (5 mo-16 y; 73 total: 18 wheezes, 10 crackles, and 45 normal breath sounds)	Quasi-experimental	3: wheezes, crackles, and normal breath sounds	Recordings of outpatients and inpatients in Singapore under realistic conditions. Use of the RALE lung sound database [43]. Classified by 2 pediatric pulmonologists	Custom-built digital stethoscope with a smartphone	SVM	Suggest a ternary classification model for discriminating among wheezes, crackles, and normal breath sounds: sensibility 91% and specificity 95%	• Advantages: establishment of a gold standard. Varied pediatric age range • Limitations: human labeling subjectivity. Reliance on a custom-built device and a publicly available lung sound database covering all age ranges yet lacking pediatric-specific data. A majority (61%) of breath sounds being normal. No tabular data	No code and no data
Kim et al [42]		Prospective cohort study	2: wheezes versus other breath sounds		Digital stethoscope (JABES GSTechnology)				No code and no data[f]

XSL•FO
RenderX

Study	Participants, N[a] (age; recordings, n)	Study design	Classes, n	Methodology	Recording device	AI[b] model architecture	Findings, metrics	Advantages and limitations	Open-source
	76 children with wheezing and healthy children (1-8 y; total 287: 103 wheezes and 184 unspecified sounds)			Recordings of Korean outpatients under realistic conditions. No lung sound database. Classified by 3 pediatric pulmonologists		3-layer LSTM[g], 4-layer CNN, 4-layer CNN+tabular data, and 34-layer residual network with CBAM[h]+tabular data	Applying a 34-layer residual network with CBAM for audio data and MLP[i] for tabular data exhibits high performance in wheeze sound detection and is suitable for clinical practice: accuracy 91.2%, sensibility 81%, AUC[j] 89.1%, and F_1-score 87.2%	• Advantages: establishment of a gold standard. Combining tabular data improved the AI algorithm performance • Limitations: human labeling subjectivity. Imbalanced data in the training data set. Cannot detect crackles and other sounds simultaneously, impacting diagnosis. 184 unspecified sounds	
Gelman et al [39]	951 patients with asthma and 167 healthy individuals (mo[k]-47 y; 1118)	Prospective cohort study	2: asthma and healthy	Recordings from patients at a polyclinic and children's hospital in Russia. No lung sound database	Built-in or external microphones from a smartphone	NN[l]	Suggest a binary classification model for discriminating between patients with asthma and healthy individuals. Accuracy 89.4%, sensibility 89.3%, and specificity 86%	• Advantage: varied pediatric age range • Limitations: No use of a gold standard. NN trained on mixed adult (12%) and child (88%) lung sounds from oral cavity (4%), tracheal (50%), and chest (46%) recordings. No disclosure of whether patients were inpatients or outpatients. No tabular data	No code and no data

[a]Diagnoses are mentioned when provided by the authors.

[b]AI: artificial intelligence.

[c]SVM: support vector machine.

[d]CNN: convolutional neural network.

[e]StethoMe AI-specific architecture is not explicitly disclosed, but it could potentially involve deep learning, convolutional, or recurrent neural networks.

[f]Data declared available from the authors upon request but not open access.

XSL•FO
RenderX

[g]LSTM: long short-term memory.

[h]CBAM: convolutional block attention module.

[i]MLP: multilayer perceptron.

[j]AUC: area under the curve.

[k]Undisclosed minimum age of children.

[l]NN: neural networks.

Among the identified 7 studies, 4 (57%) [12,39,40,42] developed customized models for their classification tasks, while 3 (43%) [11,38,41] chose a preexisting ready-to-use model (ie, StethoMe AI, and an AI algorithm developed by Tuoxiao Intelligent Technology Company, Shanghai, China). Various model architectures were used for respiratory sound classification, such as support vector machines, neural networks (NNs), and convolutional NNs (Table 2). The algorithms were used for automatic classification of wheezes, crackles, and normal breath sounds. The performance measures varied across studies, with accuracies ranging from 57.7% to 97.7%, sensitivities ranging from 78.2% to 91%, specificities ranging from 82.2% to 97%, and F_1-scores ranging from 66.4% to 87.2% (Table 2). None of the studies shared the algorithm code or data with the community, making reproducibility impossible.

Of the 9 records from gray literature search, the final sample consisted of 6 (67%) conference papers [44-49], 2 (22%) review articles [50,51], and 1 (11%) research article [52] (Multimedia Appendix 2 and Multimedia Appendix 3 [44-53]). Among the total of 201 articles assessed for eligibility across all literature resources, 120 (59.7%) were partially or fully composed of public data set repositories or multimedia resources of breath sounds to train and benchmark their AI-based lung sound algorithms or discuss their relevance in this field. These data sets and their primary features are detailed in Multimedia Appendix 4 [43,53-71]. The International Conference on Biomedical and Health Informatics 2017 Respiratory Sound Database and the Respiratory Acoustic Laboratory Environment represented the 2 most common data sets, used in 61 (50.8%) of 120 studies.

Critical Appraisal and Synthesis of the Results

Table 3 summarizes the quality assessment scores of each dimension of the ChAMAI and the total score in each study. The average score of the included studies was 20.8 (SD 6.7), indicating overall medium quality within the lower range. All studies exhibited scores ranging from 13.5 to 31.5 (low to medium quality). Multimedia Appendix 5 [11,12,21,36-40] shows scores for each of the high- and low-priority items.

Table 3. Quality assessment scores of included studies according to the 6 ChAMAI (Checklist for the Assessment of Medical Artificial Intelligence) dimensions [21].

Study	Problem understanding (total score 10)	Data understanding (total score 6)	Data preparation (total score 8)	Modeling (total score 6)	Validation (total score 12)	Deployment (total score 8)	Total (n=50)
Mazić et al [38]	3.5	0	2	6	4.5	0.5	16.5
Grzywalski et al [11]	5.5	1	0	6	1	1	14.5
Kevat et al [12]	6.5	5	2	4	0.5	1	19
Zhang et al [41]	9	5	1	6	1	2	24
Cheng et al [40]	9	6	2	6	2	1.5	26.5
Kim et al [42]	9	5	3	6	7	1.5	31.5
Gelman et al [39]	6	0	1	5	1	0.5	13.5

SWOT Analysis

A visual representation of the SWOT elements related to AI-assisted lung auscultation for pediatric asthma is presented in Figure 2. It provides a comprehensive understanding of the multifaceted impact of integrating AI into pediatric asthma care, highlighting strength and growth potential, while addressing existing challenges and threats. A detailed breakdown of the SWOT identified in the retrieved academic literature is provided in Multimedia Appendix 6 [11,12,38-42].

XSL•FO
RenderX

Figure 2. Strengths, weaknesses, opportunities, and threats (SWOT) analysis of artificial intelligence (AI)-assisted lung sound analysis for children with asthma.

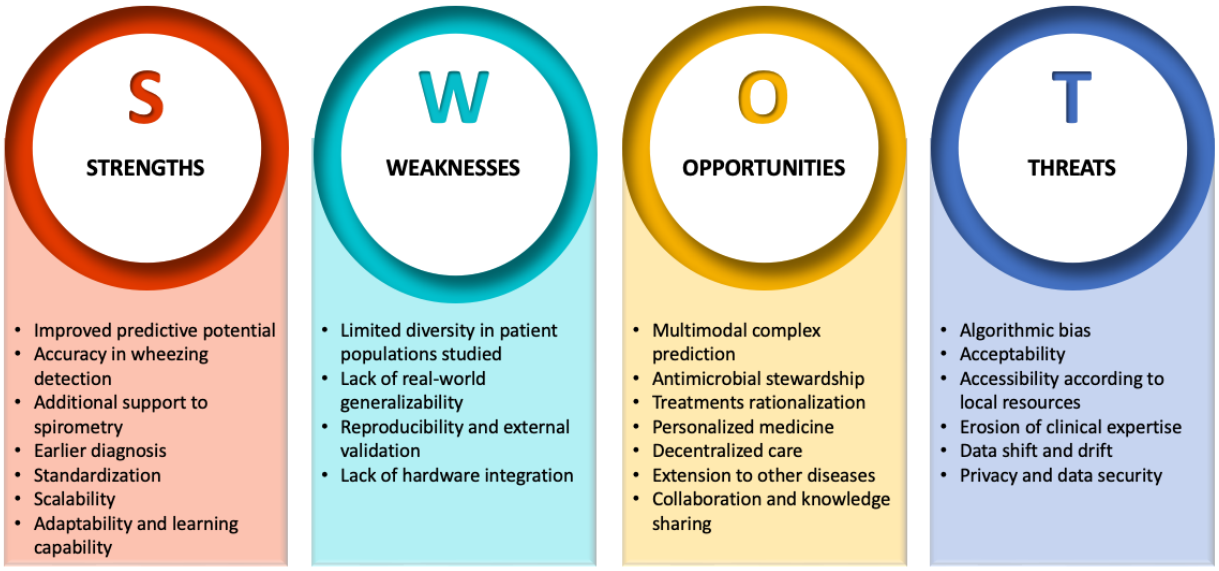

Discussion

Principal Findings

This scoping review provides an overview of the current state of research on AI models for lung sound analysis in pediatric asthma and wheezing disorders. Only 7 (11%) of the 64 studies qualified for review, highlighting the nascent stage of research in this field. Variations among studies can be attributed to heterogeneity in study populations, design, sample sizes, lung sound recording methods, and models used. Notably, although these models showed promising results in distinguishing between wheezing, crackles, and normal breath sounds, none of the studies conducted external validation to assess the reproducibility and generalizability of the model. Furthermore, the lack of open-source code makes reproducibility or external validation impossible.

In terms of quality, significant disparities were observed among the identified studies, which can be characterized as average quality overall. While problem understanding and modeling were satisfying across all studies, with the exception of the study by Mazić et al [38], there were shortcomings in adequately defining participant inclusion and exclusion criteria. Moreover, descriptions related to data preparation, validation, and deployment exhibited areas that could benefit from improvement. Therefore, further research and refinement of AI algorithms are imperative to elevate their performance and enhance their applicability across various pediatric populations and respiratory conditions. These advances will contribute to achieving more accurate and reliable AI-assisted lung sound analysis in pediatric asthma.

Strengths

Improved Predictive Potential

The ability of deep learning to extract predictive patterns from complex medical data has proven utility across several domains of health care [72,73]. As shown in the above scoping review, one of the strengths of AI-based lung sound analysis is its accuracy in the detection of wheezing. These results are in line with the growing evidence found in adult studies on asthma [74].

Additional Support in Assessing Asthma Attack Severity Alongside Spirometry

Spirometry is considered the gold standard for assessing airway obstruction. However, performing it in young patients shows more limitations in their capacity for reproducible and analyzable voluntary breathing maneuvers than that of older individuals [75]. Conducting AI-powered auscultation before and after bronchodilator treatment in young patient population offers the potential to provide valuable additional information beyond the scope of spirometry alone. This could include the detection of subtle changes in breath sounds indicative of airway constriction or responsiveness to bronchodilators, thereby influencing treatment decisions.

Earlier Diagnosis

Among children with persistent, mild-to-moderate asthma, 75% have shown abnormal patterns of lung function growth and increased risk for fixed airflow obstruction and possibly chronic obstructive pulmonary disease in early adulthood [76]. Thus, AI may facilitate timely and precise identification of asthma exacerbations in children, particularly focusing on the detection of the onset of symptoms and worsening of respiratory distress within the home environment [77]. Through frequent monitoring

XSL•FO
RenderX

of breath sounds, AI-powered auscultation can alert caregivers to subtle changes indicative of impending acute asthma exacerbation episodes, thereby allowing for timely proactive intervention before symptoms worsen and better management strategies [78]. This proactive approach not only improves treatment effectiveness but also encourages parents or caregivers to apply necessary preventive measures, such as adjusting medication dosages or avoiding triggers [79]. Consequently, it reinforces asthma management, helping to better control persistent asthma and minimizing further lung damage, thereby enhancing the overall quality of life of affected pediatric patients [80].

Standardization

Interuser variability is one of the most problematic issues in the manual assessment of lung auscultation. Standardizing lung sound analysis through AI may reduce the variability and favor more objective and targeted care in children with asthma.

Scalability

The rising prevalence of asthma [81] requires resources capable of adapting to respiratory epidemics [82,83]. Leveraging open-source and open-access AI algorithms allows for cost-free replication, addressing scalability concerns. Moreover, the convergence of competitively priced digital stethoscopes and AI powered by deep NNs significantly enhances diagnostic accuracy and the scalability of respiratory health care.

Adaptability

While humans rapidly show limited and declining performance when overwhelmed by data [84], AI models feed on large volumes of data, to which they continuously adapt by improving through iterative training and feedback loops. Algorithms can be updated to evolving data (albeit requiring a nonnegligible data pipeline for continuous learning).

Weaknesses

Limited Diversity in Patient Populations Studied

While strengths focus on the potential for improved predictive performance, the limited number of patients and settings make it difficult to assess if applicable to other populations and environments because both model accuracy and performance can be compromised. In our scoping review, most of the studies showed a questionable variety, with small numbers of relatively homogeneous patient cohorts used to train the algorithms.

Real-World Generalizability

Studies collect data in study conditions, which may use specially trained researchers to acquire the data in conditions that do not reflect real-world environments (eg, carefully measured systematic acquisition in quiet rooms on compliant patients). Thus, algorithms might lose their predictive potential under the strain of real-world conditions. Furthermore, as in the study from Cheng et al [40], many studies assessing AI model performances often use publicly available databases (eg, the Respiratory Acoustic Laboratory Environment repository [43], the International Conference on Biomedical and Health Informatics [53], etc) to train their algorithms. However, these databases present fundamental flaws in the data collection process, such as the lack of standardization of the recording environment or the use of different stethoscopes, generating systematic biases in the robustness of their predictive and classification capacity.

Reproducibility and External Validation

As mentioned, no studies performed external validation (testing on a data set representing a new patient population), and none have released their data or models. Thus, making it impossible to assess the validity of the findings.

Lack of Hardware Integration

Accurate collection and standardization of lung audio data require acquisition by digital stethoscopes. To date, few of these devices have built-in AI to assist interpretation.

Opportunities

Multimodal Complex Prediction

The etiology of asthma exacerbations is highly multifactorial, ranging from environmental and genetic factors to the immunological status of the affected child [85]. Predictive patterns are too complex to be assessed without computational assistance. The development of expert-based AI algorithms could help clinicians to identify childhood asthma more quickly, with better reproducibility and affordability [86]. New methods in multimodal decision support [87] allow algorithms to consider diverse data sources at once including imagery, sound, and clinical records. As outlined by Drummond et al [88], the implementation of digital twin systems in health care, illustrated by patient-specific virtual counterparts, integrates multiple data sources to adapt and provide better asthma management in real time. This innovative approach stands to gain from the inclusion of AI-powered lung auscultation as one of its sources in the multimodal approach. The composability of these elements could optimize medical resources to assist children in the daily management of their asthma.

Antimicrobial Stewardship

When faced with diagnostic uncertainty, clinicians tend to err on the side of caution. This behavior is a major issue for antibiotic stewardship and respiratory disease is the leading cause of antibiotic mismanagement. Therefore, an algorithm permitting to enhance the diagnosis of different respiratory conditions could assist in reducing the overuse of antibiotics or steroids [89].

Personalization

As asthma is a chronic condition, it is a good candidate for granular patient-level longitudinal data that could inform on personal evolution. Personalized medicine is a concept based on the following three elements: (1) choosing the right treatment option, (2) adapting it to the right patient, and (3) administering it at the right time [15]. Regarding asthma, the advent of personalized medicine coupled with the digitalization of data should enable specialists to develop a protocol tailored to each patient, with beneficial perspectives in empowerment and disease management [90].

XSL•FO
RenderX

Decentralization

Thanks to advances in cloud computing, telemedicine, and mobile health, patient data can be made available to remote care providers or algorithms. This accessibility can promote equitable access to advanced diagnostic tools and specialized expertise. The opportunity of this is especially obvious when considering resource constraints faced during the peaks of the COVID-19 pandemic. Telemedicine could allow clinicians to safely follow patients at any time regardless of the epidemiological context, thereby maintaining quality, safety, and continuity of care. Digital stethoscopes associated with telemedicine are promising areas of research and development, with many devices already designed and certified to record lung sounds [91]. Digitization and automation played a major role during the COVID-19 pandemic and now offer additional benefits for flexibility and more equitable access to predictive medicine [92,93]. Finally, in low- and middle-income countries, where asthma management and control are frequently insufficient [94], automated interpretation of clinical examinations holds the potential to fill resource gaps, especially in remote settings which have sparse medical expertise or not enough trained staff.

Extension to Other Diseases

While wheezing seems to be a strong signal, automated interpretation of lung sounds has been shown to be promising in various respiratory diseases [9,95]. For instance, a recent multicenter study [13] has shown the performances of a single deep-learning algorithm to detect the audio signatures of several pediatric respiratory conditions, including pneumonia, wheezing respiratory disorders, and bronchiolitis from lung auscultation performed in real clinical conditions and low-resources countries.

Collaboration

By enabling collaboration and knowledge sharing between institutions and the medical community, AI models promote a collaborative environment that favors a collective understanding of respiratory conditions, stimulates joint research efforts, and accelerates the advancement of innovative solutions to improve patient care. New technologies enabling collaborative learning without needing to share data are particularly promising. Like federated learning, where, instead of sharing the data to the source of the model, the model is shared to the source of the data [96].

Threats

Bias

AI-based tools have been evaluated for their inherent risk of bias, such as underdiagnosis of underrepresented populations [97] or algorithmic bias, which challenge the implementation of these new tools. Undetected biases may result from poor-quality or biased selection of the original data. It is essential to ensure representative data collection, large volumes of data to train the algorithms, robust external validation, and close interdisciplinary collaboration to avoid bias. These measures, combined with good user training, the development of transparent and interpretable AI models, the conduct of comprehensive clinical validation studies, and the establishment of clear guidelines, should also address users' concerns about AI-assisted pulmonary auscultation.

Acceptability

The use of AI in medicine is still a new concept to much of the general public and patients (as well as clinicians) may not feel comfortable using or interpreting AI-powered devices. In the case of pediatric asthma, a French study found that the parental perception of AI in parents of patients with asthma was mostly positive. Parents acknowledge the potential of AI to assist physicians in diagnosing and determining the most appropriate treatment, with >50% of them agreeing to adhere to an algorithmic decision system for managing their child's asthma [98]. Furthermore, since the COVID-19 pandemic, parents have shown an increased preference for AI-based management. Interestingly, children seem even more willing to accept AI-based asthma management than their parents [99]. By contrast, the acceptability of AI is contingent on the need for AI systems to be finely tuned and updated regularly. This requirement ensures that AI remains accurate, reliable, and aligned with evolving data and user expectations, which contributes to its continued adoption and trustworthiness.

Accessibility

Seamless integration into existing health care systems, interoperability, and ease of use of AI models require careful planning, adequate resources, and collaboration between technology developers and local health care providers. These considerations are particularly relevant when it comes to equitably addressing the needs of disadvantaged populations or those living in low-income countries where, ironically, asthma prevalence is high [81]. Digital health may offer opportunities to achieve universal health [100]. Nevertheless, these prospects are hindered by a shortage of medical technologies, insufficient digital infrastructure for their support and maintenance, limited financial resources, and low educational levels [101], all of which collectively exacerbate the digital divide.

Clinical Expertise

While the integration of AI into digital lung auscultation analysis offers valuable insights and can enhance diagnostic performances, there is a need to question the adequacy of health care professionals' training in the era of AI-driven auscultation. Relying solely on automated algorithms for this essential task could unwittingly erode the proficiency of clinicians in recognizing subtle nuances and contextual factors useful in making accurate diagnoses. Striking a balance between AI support and maintaining the auscultatory skills of medical practitioners should be a priority in the evolving landscape of digital health.

Data Shift and Drift

As clinical AI systems become more common, it raises questions about the readiness of clinicians to identify issues when these systems fail. One common issue is data shift, where AI underperforms when there is a mismatch between the data used to train the AI model and the data it encounters in the real world [102]. For example, during the COVID-19 pandemic, a widely used sepsis-alerting AI system had to be deactivated because it could not adapt to changing patient demographics. Data drift,

XSL•FO
RenderX

on the other hand, refers to the gradual change in data patterns over time, leading to outdated and less accurate AI models. These threats can result in biased predictions and impair the classification model's performance. To mitigate these threats, AI models need to be constantly monitored and adapted to changes in the associated data to ensure their long-term reliability and effectiveness. Nonetheless, clinicians will need to remain vigilant and report discrepancies between their clinical judgment and AI predictions.

Privacy and Data Security

Finally, 1 of the main threats concerns ethical considerations relating to privacy and data security, given that the use of AI algorithms relies on accessing and processing sensitive patient information. Encryption of data transit is crucial for patient confidentiality and user trust, with access to authorized individuals only [103]. Strikingly, none of the threats and challenges mentioned earlier were addressed, or only to a very limited extent, in the studies identified in our scoping review, as shown by the poor rating of the Deployment dimension of the ChAMAI.

Limitations

This study has limitations. First, we did not conduct a systematic review. Given the novelty of AI in the field of lung auscultation and the rapid evolution of techniques, a scoping review was preferred to a traditional systematic one for a preliminary assessment of a potentially large and diverse body of literature on this broad topic and its application to the pediatric population [104]. Similarly, a formal meta-analysis was not feasible due to the paucity and heterogeneity of available data. Instead, by synthesizing existing evidence in a charting format, this scoping review provides a comprehensive overview of the research landscape, identifies key themes, and highlights areas to inform the decision-making process for further investigation and potentially guide the design of future systematic reviews and meta-analyses. A second limitation relates to the potential omission of significant research related to the use of AI in children with asthma. This has been minimized by an exhaustive exploration of the main medical literature databases, a meticulous analysis of the references within the selected articles, and a double assessment of the identified studies. However, our inclusion criteria may have led to the unintentional exclusion of information. Third, as AI-powered lung sound analysis is an emerging topic and studies take time to be published, studies or gray literature may have been published after the study inclusion date, potentially leading to incomplete results. Finally, using the ChAMAI tool to assess the quality of the selected studies could be criticized. Nevertheless, to the best of our knowledge, this tool is the most comprehensive and advanced tool to date for such an assessment.

Conclusions

This scoping review and the accompanying SWOT analysis provide a comprehensive assessment of the existing literature on the use of AI-assisted lung auscultation in children aged <18 years with asthma or wheezing disorders. The findings highlight the potential benefits of integrating AI algorithms in objective lung sound evaluation, offering improved diagnostic accuracy, personalized management strategies, and remote monitoring capacities. Addressing the identified limitations, promoting standardization, and addressing ethical considerations will be critical for the successful implementation of AI-based systems in pediatric asthma care, ultimately improving the outcome of affected children. Further research and collaborations are necessary to deploy the full potential of AI in revolutionizing the diagnosis and management of pediatric asthma.

Acknowledgments

This study received no funding. The authors attest that there was no use of generative artificial intelligence (AI) technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

All data analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions

JNS and IR-M independently reviewed the articles for inclusion in this scoping review. JNS led the study design. JNS, IR-M, MAH, and LL provided the original conceptualization, wrote the manuscript, and answered reviewers' comments. MAH and IR-M provided expertise and guidance for the work. LL oversaw the conduct of the study. All authors contributed to critical review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Population-Concept-Context (PCC) framework for identifying the main concepts of the scoping review.
[PDF File (Adobe PDF File), 33 KB - jmir_v26i1e53662_app1.pdf]

Multimedia Appendix 2

XSL•FO
RenderX

Distribution of article types according to Google Scholar occurrences (N=183).
[PDF File (Adobe PDF File), 24 KB - jmir_v26i1e53662_app2.pdf]

Multimedia Appendix 3
Characteristics of individual sources of evidence from gray literature search.
[PDF File (Adobe PDF File), 83 KB - jmir_v26i1e53662_app3.pdf]

Multimedia Appendix 4
Public repositories of breath sounds: resources for artificial intelligence–based lung sound algorithm training and benchmarking.
[PDF File (Adobe PDF File), 102 KB - jmir_v26i1e53662_app4.pdf]

Multimedia Appendix 5
Quality assessment of selected studies using the Checklist for the Assessment of Medical Artificial Intelligence.
[PDF File (Adobe PDF File), 180 KB - jmir_v26i1e53662_app5.pdf]

Multimedia Appendix 6
Strengths, weaknesses, opportunities, and threats of retrieved academic literature.
[PDF File (Adobe PDF File), 48 KB - jmir_v26i1e53662_app6.pdf]

Multimedia Appendix 7
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 102 KB - jmir_v26i1e53662_app7.pdf]

References

1. Braman SS. The global burden of asthma. Chest 2006 Jul;130(1 Suppl):4S-12S. [doi: 10.1378/chest.130.1_suppl.4S] [Medline: 16840363]
2. Wu J, Yu Y, Yao X, Zhang Q, Zhou Q, Tang W, et al. Visualizing the knowledge domains and research trends of childhood asthma: a scientometric analysis with CiteSpace. Front Pediatr 2022;10:1019371 [FREE Full text] [doi: 10.3389/fped.2022.1019371] [Medline: 36245730]
3. Ponte EV, Mingotti CF, Mamoni RL, Marchi E, Martinelli JE, de Menezes MB, et al. Hospital admission rate in children and adolescents with mild persistent asthma. Pediatr Pulmonol 2021 Jul;56(7):1889-1895. [doi: 10.1002/ppul.25363] [Medline: 33721424]
4. Sullivan PW, Ghushchyan V, Navaratnam P, Friedman HS, Kavati A, Ortiz B, et al. The national burden of poorly controlled asthma, school absence and parental work loss among school-aged children in the United States. J Asthma 2018 Jun;55(6):659-667. [doi: 10.1080/02770903.2017.1350972] [Medline: 28981368]
5. Mattes J, Gibson PG. The early origins of COPD in severe asthma: the one thing that leads to another or the two things that come together? Thorax 2014 Sep;69(9):789-790. [doi: 10.1136/thoraxjnl-2014-205401] [Medline: 24813191]
6. Kelso JM. Do written asthma action plans improve outcomes? Pediatr Allergy Immunol Pulmonol 2016 Mar 01;29(1):2-5 [FREE Full text] [doi: 10.1089/ped.2016.0634] [Medline: 27158550]
7. Kevat AC, Kalirajah A, Roseby R. Digital stethoscopes compared to standard auscultation for detecting abnormal paediatric breath sounds. Eur J Pediatr 2017 Jul;176(7):989-992. [doi: 10.1007/s00431-017-2929-5] [Medline: 28508991]
8. Palaniappan R, Sundaraj K, Sundaraj S. Artificial intelligence techniques used in respiratory sound analysis--a systematic review. Biomed Tech (Berl) 2014 Feb;59(1):7-18. [doi: 10.1515/bmt-2013-0074] [Medline: 24114889]
9. Sfayyih AH, Sabry AH, Jameel SM, Sulaiman N, Raafat SM, Humaidi AJ, et al. Acoustic-based deep learning architectures for lung disease diagnosis: a comprehensive overview. Diagnostics (Basel) 2023 May 16;13(10):1748 [FREE Full text] [doi: 10.3390/diagnostics13101748] [Medline: 37238233]
10. Kim Y, Hyon Y, Lee S, Woo SD, Ha T, Chung C. The coming era of a new auscultation system for analyzing respiratory sounds. BMC Pulm Med 2022 Mar 31;22(1):119 [FREE Full text] [doi: 10.1186/s12890-022-01896-1] [Medline: 35361176]
11. Grzywalski T, Piecuch M, Szajek M, Bręborowicz A, Hafke-Dys H, Kociński J, et al. Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination. Eur J Pediatr 2019 Jun;178(6):883-890 [FREE Full text] [doi: 10.1007/s00431-019-03363-2] [Medline: 30927097]
12. Kevat A, Kalirajah A, Roseby R. Artificial intelligence accuracy in detecting pathological breath sounds in children using digital stethoscopes. Respir Res 2020 Sep 29;21(1):253 [FREE Full text] [doi: 10.1186/s12931-020-01523-9] [Medline: 32993620]
13. Heitmann J, Glangetas A, Doenz J, Dervaux J, Shama DM, Garcia DH, with the Pneumoscope Study Group. DeepBreath-automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries. NPJ Digit Med 2023 Jun 02;6(1):104 [FREE Full text] [doi: 10.1038/s41746-023-00838-3] [Medline: 37268730]

XSL•FO
RenderX

14. Prodhan P, Dela Rosa RS, Shubina M, Haver KE, Matthews BD, Buck S, et al. Wheeze detection in the pediatric intensive care unit: comparison among physician, nurses, respiratory therapists, and a computerized respiratory sound monitor. Respir Care 2008 Oct;53(10):1304-1309 [FREE Full text] [Medline: 18811991]
15. Louis R, Roche N. Personalised medicine: are we ready? Eur Respir Rev 2017 Sep 30;26(145):170088 [FREE Full text] [doi: 10.1183/16000617.0088-2017] [Medline: 31408006]
16. Peters M, Godfrey C, McInerney P, Munn Z, Tricco A, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. Joanna Briggs Institute Reviewer's Manual. Adelaide, Australia: Joanna Briggs Institute; 2017:1-24.
17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of google scholar in evidence reviews and its applicability to grey literature searching. PLoS One 2015;10(9):e0138237 [FREE Full text] [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]
19. Briscoe S, Nunns M, Shaw L. How do Cochrane authors conduct web searching to identify studies? Findings from a cross-sectional sample of Cochrane Reviews. Health Info Libr J 2020 Dec;37(4):293-318. [doi: 10.1111/hir.12313] [Medline: 32511888]
20. Kastner M, Wilczynski NL, Walker-Dilks C, McKibbon KA, Haynes B. Age-specific search strategies for Medline. J Med Internet Res 2006;8(4):e25 [FREE Full text] [doi: 10.2196/jmir.8.4.e25] [Medline: 17213044]
21. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform 2021 Sep;153:104510 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104510] [Medline: 34108105]
22. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. 2000 Presented at: KDD '20; August 27-31, 2000; Manchester, UK URL: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf
23. Sharan RV, Rahimi-Ardabili H. Detecting acute respiratory diseases in the pediatric population using cough sound features and machine learning: a systematic review. Int J Med Inform 2023 Aug;176:105093 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105093] [Medline: 37224643]
24. Teoli D, Sanvictores T, An J. SWOT Analysis. Treasure Island, FL: StatPearls Publishing; 2023.
25. Pickton DW, Wright S. What's swot in strategic analysis? Strateg Change 1998;7:101-109 [FREE Full text] [doi: 10.1002/(sici)1099-1697(199803/04)7:2<101::aid-jsc332>3.0.co;2-6]
26. Sterling M, Rhee H, Bocko M. Automated cough assessment on a mobile platform. J Med Eng 2014;2014:951621 [FREE Full text] [doi: 10.1155/2014/951621] [Medline: 25506590]
27. Bokov P, Mahut B, Flaud P, Delclaux C. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. Comput Biol Med 2016 Mar 01;70:40-50. [doi: 10.1016/j.compbiomed.2016.01.002] [Medline: 26802543]
28. Porter P, Abeyratne U, Swarnkar V, Tan J, Ng T, Brisbane JM, et al. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. Respir Res 2019 Jun 06;20(1):81 [FREE Full text] [doi: 10.1186/s12931-019-1046-6] [Medline: 31167662]
29. Hee HI, Balamurali B, Karunakaran A, Herremans D, Teoh OH, Lee KP, et al. Development of machine learning for asthmatic and healthy voluntary cough sounds: a proof of concept study. Applied Sciences 2019 Jul 16;9(14):2833. [doi: 10.3390/app9142833]
30. Swarnkar V, Abeyratne U, Tan J, Ng TW, Brisbane JM, Choveaux J, et al. Stratifying asthma severity in children using cough sound analytic technology. J Asthma 2021 Feb;58(2):160-169. [doi: 10.1080/02770903.2019.1684516] [Medline: 31638844]
31. Alqudaihi KS, Aslam N, Khan IU, Almuhaideb AM, Alsunaidi SJ, Ibrahim NM, et al. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. IEEE Access 2021;9:102327-102344 [FREE Full text] [doi: 10.1109/ACCESS.2021.3097559] [Medline: 34786317]
32. Balamurali BT, Hee HI, Kapoor S, Teoh OH, Teng SS, Lee KP, et al. Deep neural network-based respiratory pathology classification using cough sounds. Sensors (Basel) 2021 Aug 18;21(16):5555 [FREE Full text] [doi: 10.3390/s21165555] [Medline: 34450996]
33. Kruizinga MD, Zhuparris A, Dessing E, Krol FJ, Sprij AJ, Doll R, et al. Development and technical validation of a smartphone-based pediatric cough detection algorithm. Pediatr Pulmonol 2022 Mar;57(3):761-767 [FREE Full text] [doi: 10.1002/ppul.25801] [Medline: 34964557]
34. Liao S, Song C, Wang X, Wang Y. A classification framework for identifying bronchitis and pneumonia in children based on a small-scale cough sounds dataset. PLoS One 2022;17(10):e0275479 [FREE Full text] [doi: 10.1371/journal.pone.0275479] [Medline: 36301797]
35. Bentur L, Beck R, Shinawi M, Naveh T, Gavriely N. Wheeze monitoring in children for assessment of nocturnal asthma and response to therapy. Eur Respir J 2003 Apr;21(4):621-626 [FREE Full text] [Medline: 12762346]

XSL•FO
RenderX

36. Bentur L, Beck R, Berkowitz D, Hasanin J, Berger I, Elias N, et al. Adenosine bronchial provocation with computerized wheeze detection in young infants with prolonged cough: correlation with long-term follow-up. Chest 2004 Oct;126(4):1060-1065. [doi: 10.1378/chest.126.4.1060] [Medline: 15486364]
37. Yu C, Tsai TH, Huang SI, Lin CW. Soft stethoscope for detecting asthma wheeze in young children. Sensors (Basel) 2013 Jun 06;13(6):7399-7413 [FREE Full text] [doi: 10.3390/s130607399] [Medline: 23744030]
38. Mazić I, Bonković M, Džaja B. Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds. Biomed Signal Process Control 2015 Aug;21:105-118. [doi: 10.1016/j.bspc.2015.05.002]
39. Gelman A, Furman EG, Kalinina NM, Malinin SV, Furman GB, Sheludko VS, et al. Computer-aided detection of respiratory sounds in bronchial asthma patients based on machine learning method. Sovrem Tekhnologii Med 2022;14(5):45-51 [FREE Full text] [doi: 10.17691/stm2022.14.5.05] [Medline: 37181833]
40. Cheng ZR, Zhang H, Thomas B, Tan YH, Teoh OH, Pugalenthi A. Assessing the accuracy of artificial intelligence enabled acoustic analytic technology on breath sounds in children. J Med Eng Technol 2022 Jan;46(1):78-84. [doi: 10.1080/03091902.2021.1992520] [Medline: 34730469]
41. Zhang J, Wang HS, Zhou HY, Dong B, Zhang L, Zhang F, et al. Real-world verification of artificial intelligence algorithm-assisted auscultation of breath sounds in children. Front Pediatr 2021;9:627337 [FREE Full text] [doi: 10.3389/fped.2021.627337] [Medline: 33834010]
42. Kim BJ, Kim BS, Mun JH, Lim C, Kim K. An accurate deep learning model for wheezing in children using real world data. Sci Rep 2022 Dec 28;12(1):22465 [FREE Full text] [doi: 10.1038/s41598-022-25953-1] [Medline: 36577766]
43. Owens D. Rale lung sounds 3. NursingCenter. 2002. URL: https://www.nursingcenter.com/journalarticle?Article_ID=405020&Journal_ID=54020&Issue_ID=405016 [accessed 2024-04-29]
44. Fernandes T, Rocha BM, Pessoa D, de CP, Paiva RP. Classification of adventitious respiratory sound events: a stratified analysis. In: Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics. 2022 Presented at: BHI '22; September 27-30, 2022; Ioannina, Greece p. 1-5 URL: https://ieeexplore.ieee.org/document/9926841 [doi: 10.1109/bhi56158.2022.9926841]
45. Gouda A, El Shehaby S, Diaa N, Abougabal M. Classification techniques for diagnosing respiratory sounds in infants and children. In: Proceedings of the IEEE 9th Annual Computing and Communication Workshop and Conference. 2019 Presented at: CCWC '19; January 7-9, 2019; Las Vegas, NV p. 354-360 URL: https://ieeexplore.ieee.org/document/8666608 [doi: 10.1109/ccwc.2019.8666608]
46. Liu L, Li W, Jiang C. Time-varying respiratory feature recognition and classification for respiration health and disease screening in children. In: Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics. 2021 Presented at: SMC '21; October 17-20, 2021; Melbourne, Australia p. 1260-1265 URL: https://ieeexplore.ieee.org/document/9659096 [doi: 10.1109/smc52423.2021.9659096]
47. Liu R, Cai S, Zhang K, Hu N. Detection of adventitious respiratory sounds based on convolutional neural network. In: Proceedings of the 2019 International Conference on Intelligent Informatics and Biomedical Sciences. 2019 Presented at: ICIIBMS '19; November 21-24, 2019; Shanghai, China p. 298-303 URL: https://ieeexplore.ieee.org/document/8991459 [doi: 10.1109/iciibms46890.2019.8991459]
48. Milicevic M, Mazic I, Bonkovic M. Classification accuracy comparison of asthmatic wheezing sounds recorded under ideal and real-world conditions. In: Proceedings of the 15th International Conference on Artificial Intelligence, Knowledge Engineering and Databases. 2016 Presented at: AIKED '16; January 29-31, 2016; Venice, Italy p. 101-106 URL: https://www.bib.irb.hr:8443/796936
49. Uwanthika I, Madushanka P, Deraniyagala D. Identification of related technologies associated with asthmatic wheeze detection systems - a review. General Sir John Kotelawala Defence University. 2023. URL: http://ir.kdu.ac.lk/bitstream/handle/345/6578/FOCSS%20Abstract%20Book%2018%20Jan%202023%20Final-63.pdf?sequence=1&isAllowed=y [accessed 2024-04-29]
50. Gurung A, Scrafford CG, Tielsch JM, Levine OS, Checkley W. Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis. Respir Med 2011 Sep;105(9):1396-1403 [FREE Full text] [doi: 10.1016/j.rmed.2011.05.007] [Medline: 21676606]
51. Palaniappan R, Sundaraj K, Ahamed N, Arjunan A, Sundaraj S. Computer-based respiratory sound analysis: a systematic review. IETE Tech Rev 2013;30(3):248. [doi: 10.4103/0256-4602.113524]
52. Kuo HC, Lin BS, Wang YD, Lin BS. Development of automatic wheeze detection algorithm for children with asthma. IEEE Access 2021;9:126882-126890. [doi: 10.1109/access.2021.3111507]
53. Rocha B, Filos D, Mendes L, Vogiatzis I, Perantoni E, Kaimakamis E. A respiratory sound database for the development of automated classification. In: Proceedings of the 2017 Conference on Precision Medicine Powered by pHealth and Connected Health. 2017 Presented at: ICBHI '17; November 18-21, 2017; Thessaloniki, Greece p. 33-37 URL: https://link.springer.com/chapter/10.1007/978-981-10-7419-6_6 [doi: 10.1007/978-981-10-7419-6_6]
54. Fraiwan M, Fraiwan L, Khassawneh B, Ibnian A. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. Data Brief 2021 Apr;35:106913 [FREE Full text] [doi: 10.1016/j.dib.2021.106913] [Medline: 33732827]
55. Home page. 3M Littmann. URL: https://www.littmann.in/3M/en_IN/littmann-stethoscopes-in/education/training/ [accessed 2024-04-29]

XSL•FO
RenderX

56. Pulmonary breath sounds. East Tennessee State University. 2021. URL: http://faculty.etsu.edu/arnall/www/public_html/heartlung/breathsounds/contents.html [accessed 2024-02-18]
57. Nursing starts here. Lippincott NursingCenter. URL: https://www.nursingcenter.com/ncblog/october-2016/breath-sounds-test-yourknowledge [accessed 2024-04-29]
58. One lung sounds library. Thinklabs. 2023. URL: https://www.thinklabs.com/sound-library [accessed 2024-12-18]
59. Breath sound assessment. Medscape. 2023. URL: https://emedicine.medscape.com/article/1894146-overview?form=fpf#a3 [accessed 2024-02-18]
60. Respiratory sounds. Wikipedia. 2023. URL: https://en.wikipedia.org/wiki/Respiratory_sounds [accessed 2024-02-18]
61. Heart and lung sounds reference guide. EasyAuscultation. 2023. URL: https://www.easyauscultation.com/heart-lung-sounds-reference-guide/pulmonic [accessed 2024-02-18]
62. Home page. Reference Database of Respiratory Sounds. 2023. URL: https://www.ers-education.org/e-learning/reference-database-ofrespiratory-sounds/other/ [accessed 2024-02-18]
63. Zhang Q, Zhang J, Yuan J, Huang H, Zhang Y, Zhang B, et al. SPRSound: open-source SJTU paediatric respiratory sound database. IEEE Trans Biomed Circuits Syst 2022 Oct;16(5):867-881. [doi: 10.1109/TBCAS.2022.3204910] [Medline: 36070274]
64. Hsu FS, Huang SR, Huang CW, Huang CJ, Cheng YR, Chen CC, et al. Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database-HF_Lung_V1. PLoS One 2021;16(7):e0254134 [FREE Full text] [doi: 10.1371/journal.pone.0254134] [Medline: 34197556]
65. Respiratory and drug actuation dataset. IEEE DataPort. 2022 Jun. URL: https://ieee-dataport.org/documents/respiratory-and-drug-actuation-dataset [accessed 2024-04-29]
66. Gross V, Hadjileontiadis LJ, Penzel T, Koehler U, Vogelmeier C. Multimedia database "Marburg respiratory sounds (MARS)". In: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2003 Presented at: IEMBS '03; September 17-21, 2003; Cancun, Mexico p. 17-21 URL: https://ieeexplore.ieee.org/document/1279717 [doi: 10.1109/iembs.2003.1279717]
67. Fernando T, Sridharan S, Denman S, Ghaemmaghami H, Fookes C. Robust and interpretable temporal convolution network for event detection in lung sound recordings. IEEE J Biomed Health Inform 2022 Jul;26(7):2898-2908. [doi: 10.1109/JBHI.2022.3144314] [Medline: 35061595]
68. Racineux J. Auscultation pulmonaire. Rueil-Malmaison, France: AstraZeneca respiratoire; 2007.
69. Lehrer S. Understanding Lung Sounds. 3rd edition. Scotts Valley, CA: Createspace Independent Publishing Platform; 2018.
70. Tilkian AG, Conover MB. Understanding Heart Sounds and Murmurs: With an Introduction to Lung Sounds. 4th edition. Philadelphia, PA: W B Saunders Co; 2001.
71. Wilkins RL, Hodgkin JE, Lopez B. Fundamentals of Lung and Heart Sounds. 3rd edition. Maryland Heights, MI: Mosby; 2004.
72. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. Curr Med Sci 2021 Dec;41(6):1105-1115 [FREE Full text] [doi: 10.1007/s11596-021-2474-3] [Medline: 34874486]
73. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
74. Hafke-Dys H, Kuźnar-Kamińska B, Grzywalski T, Maciaszek A, Szarzyński K, Kociński J. Artificial intelligence approach to the monitoring of respiratory sounds in asthmatic patients. Front Physiol 2021;12:745635 [FREE Full text] [doi: 10.3389/fphys.2021.745635] [Medline: 34858203]
75. Jat KR. Spirometry in children. Prim Care Respir J 2013 Jun;22(2):221-229 [FREE Full text] [doi: 10.4104/pcrj.2013.00042] [Medline: 23732636]
76. McGeachie MJ, Yates KP, Zhou X, Guo F, Sternberg AL, Van Natta ML, et al. Patterns of growth and decline in lung function in persistent childhood asthma. N Engl J Med 2016 May 12;374(19):1842-1852 [FREE Full text] [doi: 10.1056/NEJMoa1513737] [Medline: 27168434]
77. van der Kamp MR, Hengeveld VS, Brusse-Keizer MG, Thio BJ, Tabak M. eHealth technologies for monitoring pediatric asthma at home: scoping review. J Med Internet Res 2023 Jul 21;25:e45896 [FREE Full text] [doi: 10.2196/45896] [Medline: 37477966]
78. Martin J, Townshend J, Brodlie M. Diagnosis and management of asthma in children. BMJ Paediatr Open 2022 Apr;6(1):e001277 [FREE Full text] [doi: 10.1136/bmjpo-2021-001277] [Medline: 35648804]
79. Yeung TY, Ewing C, Malanowska A, Zuberbuhler P, Balcom M, Liu J, et al. Home management of childhood asthma exacerbations. Pulm Ther 2018 Dec;4(2):149-157 [FREE Full text] [doi: 10.1007/s41030-018-0061-y] [Medline: 32026392]
80. Lanz MJ, Gilbert I, Szefler SJ, Murphy KR. Can early intervention in pediatric asthma improve long-term outcomes? A question that needs an answer. Pediatr Pulmonol 2019 Mar;54(3):348-357 [FREE Full text] [doi: 10.1002/ppul.24224] [Medline: 30609252]
81. Asher MI, Rutter CE, Bissell K, Chiang CY, El Sony A, Ellwood E, Global Asthma Network Phase I Study Group. Worldwide trends in the burden of asthma symptoms in school-aged children: Global Asthma Network phase I cross-sectional study. Lancet 2021 Oct 30;398(10311):1569-1580 [FREE Full text] [doi: 10.1016/S0140-6736(21)01450-1] [Medline: 34755626]

XSL•FO
RenderX

82. Moscovich DP, Averbuch D, Kerem E, Cohen-Cymberknoh M, Berkun Y, Brooks R, et al. Pediatric respiratory admissions and related viral infections during the COVID-19 pandemic. Pediatr Pulmonol 2023 Jul;58(7):2076-2084. [doi: 10.1002/ppul.26434] [Medline: 37097057]
83. Furlow B. Triple-demic overwhelms paediatric units in US hospitals. Lancet Child Adolesc Health 2023 Feb;7(2):86. [doi: 10.1016/S2352-4642(22)00372-8] [Medline: 36529148]
84. Arnold M, Goldschmitt M, Rigotti T. Dealing with information overload: a comprehensive review. Front Psychol 2023;14:1122200 [FREE Full text] [doi: 10.3389/fpsyg.2023.1122200] [Medline: 37416535]
85. Miller RL, Lawrence J. Understanding root causes of asthma. Perinatal environmental exposures and epigenetic regulation. Ann Am Thorac Soc 2018 Apr;15(Suppl 2):S103-S108 [FREE Full text] [doi: 10.1513/AnnalsATS.201706-514MG] [Medline: 29676631]
86. Seol HY, Wi CI, Ryu E, King KS, Divekar RD, Juhn YJ. A diagnostic codes-based algorithm improves accuracy for identification of childhood asthma in archival data sets. J Asthma 2021 Aug;58(8):1077-1086 [FREE Full text] [doi: 10.1080/02770903.2020.1759624] [Medline: 32315558]
87. Swamy V, Satayeva M, Frej J, Bossy T, Vogels T, Jaggi M. MultiModN- multimodal, multi-task, interpretable modular networks. arXiv. Preprint posted online September 25, 2023 2023 [FREE Full text]
88. Drummond D, Roukema J, Pijnenburg M. Home monitoring in asthma: towards digital twins. Curr Opin Pulm Med 2023 Jul 01;29(4):270-276. [doi: 10.1097/MCP.0000000000000963] [Medline: 37102597]
89. Yu G, Li Z, Li S, Liu J, Sun M, Liu X, et al. The role of artificial intelligence in identifying asthma in pediatric inpatient setting. Ann Transl Med 2020 Nov;8(21):1367 [FREE Full text] [doi: 10.21037/atm-20-2501a] [Medline: 33313112]
90. Chung KF. Personalised medicine in asthma: time for action: number 1 in the series "personalised medicine in respiratory diseases" edited by Renaud Louis and Nicolas Roche. Eur Respir Rev 2017 Sep 30;26(145):170064 [FREE Full text] [doi: 10.1183/16000617.0064-2017] [Medline: 28954768]
91. Lakhe A, Sodhi I, Warrier J, Sinha V. Development of digital stethoscope for telemedicine. J Med Eng Technol 2016;40(1):20-24. [doi: 10.3109/03091902.2015.1116633] [Medline: 26728637]
92. Katzow MW, Steinway C, Jan S. Telemedicine and health disparities during COVID-19. Pediatrics 2020 Aug;146(2):e20201586. [doi: 10.1542/peds.2020-1586] [Medline: 32747592]
93. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020;2020:8830200 [FREE Full text] [doi: 10.1155/2020/8830200] [Medline: 33343657]
94. García-Marcos L, Chiang CY, Asher MI, Marks GB, El Sony A, Masekela R, Global Asthma Network Phase I Study Group. Asthma management and control in children, adolescents, and adults in 25 countries: a Global Asthma Network phase I cross-sectional study. Lancet Glob Health 2023 Mar;11(2):e218-e228 [FREE Full text] [doi: 10.1016/S2214-109X(22)00506-X] [Medline: 36669806]
95. Pramono RX, Bowyer S, Rodriguez-Villegas E. Automatic adventitious respiratory sound analysis: a systematic review. PLoS ONE 2017 May 26;12(5):e0177926. [doi: 10.1371/journal.pone.0177926]
96. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol 2019 Jan 28;10(2):1-19. [doi: 10.1145/3298981]
97. Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021 Dec;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
98. Abdoul C, Cros P, Coutier L, Hadchouel A, Neuraz A, Burgun A, et al. Parents' views on artificial intelligence for the daily management of childhood asthma: a survey. J Allergy Clin Immunol Pract 2021 Apr;9(4):1728-30.e3. [doi: 10.1016/j.jaip.2020.11.048] [Medline: 33290917]
99. Gonsard A, AbouTaam R, Prévost B, Roy C, Hadchouel A, Nathan N, et al. Children’s views on artificial intelligence and digital twins for the daily management of their asthma: a mixed-method study. Eur J Pediatr 2022 Dec 13;182(2):877-888. [doi: 10.1007/s00431-022-04754-8]
100. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-02-18]
101. Clifford GD. E-health in low to middle income countries. J Med Eng Technol 2016;40(7-8):336-341 [FREE Full text] [doi: 10.1080/03091902.2016.1256081] [Medline: 27885920]
102. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
103. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci 2019 Jun;64:277-282. [doi: 10.1016/j.jocn.2019.03.001] [Medline: 30878282]
104. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
ChAMAI: Checklist for the Assessment of Medical Artificial Intelligence
MR: major revision
mR: minor revision
NN: neural network
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SWOT: strengths, weaknesses, opportunities, and threats

Edited by N Cahill; submitted 14.10.23; peer-reviewed by S Lee, Y Li, D Drummond, G Yu; comments to author 02.02.24; revised version received 28.03.24; accepted 10.07.24; published 23.08.24.

Please cite as:
Ruchonnet-Métrailler I, Siebert JN, Hartley MA, Lacroix L
Automated Interpretation of Lung Sounds by Deep Learning in Children With Asthma: Scoping Review and Strengths, Weaknesses, Opportunities, and Threats Analysis
J Med Internet Res 2024;26:e53662
URL: https://www.jmir.org/2024/1/e53662
doi:10.2196/53662
PMID:39178033

©Isabelle Ruchonnet-Métrailler, Johan N Siebert, Mary-Anne Hartley, Laurence Lacroix. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Provenance Information for Biomedical Data and Workflows: Scoping Review

Kerstin Gierend[1], Dipl-Inf(FH); Frank Krüger[2,3], Dr.-Ing.; Sascha Genehr[3], MSc; Francisca Hartmann[1], MSc; Fabian Siegel[1], Dr med; Dagmar Waltemath[4], Dr-Ing; Thomas Ganslandt[5*], Dr med habi.; Atinkut Alamirrew Zeleke[4*], Dr rer medic

[1]Department of Biomedical Informatics, Mannheim Institute for intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Faculty of Engineering, Wismar University of Applied Sciences, Wismar, Germany

[3]Institute of Communications Engineering, University of Rostock, Rostock, Germany

[4]Department of Medical Informatics, University Medicine Greifswald, Greifswald, Germany

[5]Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[*]these authors contributed equally

Corresponding Author:
Kerstin Gierend, Dipl-Inf(FH)
Department of Biomedical Informatics
Mannheim Institute for intelligent Systems in Medicine
Medical Faculty Mannheim, Heidelberg University
Theodor-Kutzer-Ufer 1-3
Mannheim, 68167
Germany
Phone: 49 621383 ext 8087
Email: kerstin.gierend@medma.uni-heidelberg.de

Abstract

Background: The record of the origin and the history of data, known as provenance, holds importance. Provenance information leads to higher interpretability of scientific results and enables reliable collaboration and data sharing. However, the lack of comprehensive evidence on provenance approaches hinders the uptake of good scientific practice in clinical research.

Objective: This scoping review aims to identify approaches and criteria for provenance tracking in the biomedical domain. We reviewed the state-of-the-art frameworks, associated artifacts, and methodologies for provenance tracking.

Methods: This scoping review followed the methodological framework developed by Arksey and O'Malley. We searched the PubMed and Web of Science databases for English-language articles published from 2006 to 2022. Title and abstract screening were carried out by 4 independent reviewers using the Rayyan screening tool. A majority vote was required for consent on the eligibility of papers based on the defined inclusion and exclusion criteria. Full-text reading and screening were performed independently by 2 reviewers, and information was extracted into a pretested template for the 5 research questions. Disagreements were resolved by a domain expert. The study protocol has previously been published.

Results: The search resulted in a total of 764 papers. Of 624 identified, deduplicated papers, 66 (10.6%) studies fulfilled the inclusion criteria. We identified diverse provenance-tracking approaches ranging from practical provenance processing and managing to theoretical frameworks distinguishing diverse concepts and details of data and metadata models, provenance components, and notations. A substantial majority investigated underlying requirements to varying extents and validation intensities but lacked completeness in provenance coverage. Mostly, cited requirements concerned the knowledge about data integrity and reproducibility. Moreover, these revolved around robust data quality assessments, consistent policies for sensitive data protection, improved user interfaces, and automated ontology development. We found that different stakeholder groups benefit from the availability of provenance information. Thereby, we recognized that the term *provenance* is subjected to an evolutionary and technical process with multifaceted meanings and roles. Challenges included organizational and technical issues linked to data annotation, provenance modeling, and performance, amplified by subsequent matters such as enhanced provenance information and quality principles.

XSL•FO
RenderX

Conclusions: As data volumes grow and computing power increases, the challenge of scaling provenance systems to handle data efficiently and assist complex queries intensifies, necessitating automated and scalable solutions. With rising legal and scientific demands, there is an urgent need for greater transparency in implementing provenance systems in research projects, despite the challenges of unresolved granularity and knowledge bottlenecks. We believe that our recommendations enable quality and guide the implementation of auditable and measurable provenance approaches as well as solutions in the daily tasks of biomedical scientists.

International Registered Report Identifier (IRRID): RR2-10.2196/31750

(J Med Internet Res 2024;26:e51297) doi:10.2196/51297

KEYWORDS

provenance; biomedical research; data management; scoping review; health care data; software life cycle

Introduction

Background

Both the use and reuse of electronic medical and patient-related data offers enormous potential for clinical research [1,2]. National programs such as the German Medical Informatics Initiatives support knowledge discovery and data sharing using adequate computational infrastructure and secure processes [3]. In this context, provenance information (Textbox 1) offers access to quality assured, traceable, and credible shared data. These and other advantages of data provenance have been demonstrated, for instance, in the EU Horizon 2020 TRANSFORM project [4] or in the MeDaX Knowledge Graph Prototype [5]. Schröder et al [6] provided an "Electronic Laboratory Notebook" use case in the wet lab to show how provenance supports the understanding and reproducibility of research investigations. At the same time, researchers not considering the origin of data run into the hazard of systematically incomplete or wrong data [7].

Notably, the FAIR (findable, accessible, interoperable, reusable) guiding principles for data stewardship [9] explicitly mention provenance [10,11]. A provenance-oriented approach requires thorough planning, execution, and evaluation of data management processes in the respective application domain [2]. In the scientific context, adherence to criteria such as consistency, interoperability, and confidentiality are generally required across all software tools [1,12,13].

The concept and implementation of provenance are essential for most scientific domains, such as environmental fields (geoprocessing workflows or climate assessments), in nuclear fusion engineering, or material sciences [14,15]. In particular, the biomedical domains demand comprehensive investigation and information about their data management scenarios, including extract, transform, load jobs for data transfer and integration. Reliable data and data pipelines both require provenance data to be embedded in concepts for traceability to understand the relationships between results and source data.

Textbox 1. Provenance terminology.

Definition

- "Provenance" is a description of what happened to a data item [4]. Information models such as the World Wide Web Consortium (W3C) PROV standard formally define provenance of a resource as "a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource" [8].

Objectives

Our work reviews approaches and criteria for provenance tracking in the biomedical domain and discloses current knowledge gaps. This comprises modeling aspects and metadata frameworks for meaningful and usable provenance information during the creation, collection, and processing of scientific biomedical data. The review also covers the examination of quality aspects relating to provenance.

Methods

Overview

We followed the scoping methodological framework developed by Arksey and O'Malley [16] for conducting a scoping review with the following stages: (1) stage 1—identification of the research questions (RQs); (2) stage 2—identification of relevant studies; (3) stage 3—study selection; (4) stage 4—data extraction and charting; and (5) stage 5—collating, summarizing, and reporting the results.

Change From the Original Protocol

The protocol of this scoping review has been published in *JMIR Research Protocols* (international registered report identifier DERR1-10.2196/31750) [17]. In accordance with the original protocol developed for this scoping review, the search period was initially planned to include studies published from January 2006 to March 2021. However, due to the extensive nature of the data extraction and write-up process, it became apparent that additional time was necessary to ensure a comprehensive analysis of the relevant literature. As a result, we extended the search period from 2006 to the end of 2022.

No other changes were made to the original protocol. Thematic analysis methods were applied to analyze the extracted data by organizing themes according to the RQs [18]. In line with the framework developed by Arksey and O'Malley [16], the review

XSL•FO
RenderX

does not attempt to assess the quality of studies or the risk of bias. It also does not assess the generalizability of the results.

Stage 1: Identifying RQs

The main objective of this review was to investigate existing evidence regarding approaches and criteria for provenance tracking and disclosing current knowledge gaps in the biomedical domain. The objective led to the following RQs:

- RQ1: Which potential (methodological) approaches exist for the classification and tracking of provenance criteria and methods in a biomedical or domain-independent context?
- RQ2: How can the potential value of provenance information be harnessed and by whom? How can usability be provided?
- RQ3: What are the challenges and potential problems or bottlenecks for the accomplishment of provenance?
- RQ4: Which guidelines or demands for the consideration of provenance criteria in a biomedical or domain-independent context have to be followed?
- RQ5: How completely can provenance be mapped in the data life cycle or during data management?

Stage 2: Identifying Relevant Studies

Concepts and matching keywords were categorized into 4 groups (Table 1): *target domain* refers to the context of the research topic and includes studies with a biomedical, health care, clinical, or scientific background. In this work, scientific background is limited to domain-independent studies and excludes all other domain-specific studies. *Provenance* concerns the information about the genesis of a given object. *Provenance properties* cover specific requirements tied to the term *provenance*; they also describe selected characteristics in this context. *Objective* includes the purposes or intention of provenance capture. The comprehensive search strategy is recorded in the study protocol [17], and search strings combined with Boolean operators are attached (Multimedia Appendix 1).

Table 1. Concepts and matching keywords (eligibility criteria).

Concepts	Matching keywords (inclusion criteria)
Target domain	biomed*[a], EHR, electronic health record, health care, clinical, scientific[b]
Provenance	provenance, prov, lineage
Provenance properties	interop*, (data NEAR/2 [flow, quality, transformation]), metadata, workflow, semantic, framework, annotat*, ontolog*, management, document*, (model NEAR/2 provenance)
Objective	audit*, decision support, ETL, Extract-Transform-Load, FHIR, record linking, machine learning, reproducib*, transparen*, track*, implement*

[a]The * symbol (wildcard character) replaces or represents one or more characters.

[b]Will be used in a domain-independent context only.

Stage 3: Study Selection

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart depicts the selection process. First, we identified all relevant studies in the PubMed and Web of Science databases based on our search strategy. After deduplication, we launched a transparent screening process by importing all relevant studies into Rayyan [19], a systematic review supporting solution. The studies were then reviewed by at least 2 independent researchers. In the case of vote agreement, the study was either included in the next review phase or excluded from the review. A third independent reviewer was consulted to solve the conflict if no consensus could be reached. The study screening phase started with a title and abstract evaluation for eligibility. Included studies were submitted to a full-text screening, while performing a thorough investigation on the study report. Reviewers voted for inclusion or exclusion considering the inclusion and exclusion criteria. Finally, the residing set of qualified studies was moved into the data extraction pipeline. A description of the study selection is provided in the protocol [17].

Studies were included if they (1) were focused on the biomedical domain or were domain independent, (2) described provenance-tracking approaches, and (3) were written in English. Studies were excluded if they (1) were not specific to the biomedical or general domain, (2) were gray literature, and (3) did not focus on provenance-tracking approaches.

Stage 4: Charting the Data

We followed a collaborative and iterative process to define a charting table for data extraction. Individual reviewers (KG, FK, FH, SG, AAZ, and DW) then scrutinized all studies and extracted central textual occurrences into the data extraction sheet. The variables in the data extraction sheet correspond with the RQs. General characteristics of the studies, approaches for classification and tracking of provenance, and their associated challenges along with the significance and completeness of provenance information in the given context were part of the investigational charting. The reviewers independently charted the data in a structured and consistent way and discussed the results.

Stage 5: Collating, Summarizing, and Reporting the Results

The extracted data were analyzed using summary statistics by calculating the total number and percentages of all studies per category, if applicable. Charts were presented for the distribution of the individual data elements where applicable. The data analysis was partially supported with scripts in Python (version 3.10.0) [20]. Plots were generated with R version 4.0.4 (R Core Team) [21] and version 1.3.0 of the *tidyverse* package [22].

XSL•FO
RenderX

Further analysis was performed using qualitative evaluation. The reporting of the results and outcomes was structured according to the RQs. On the basis of the analysis of the review results, we have developed a road map for a customized provenance framework that considers the life cycle of the software framework (Provenance-Software Framework Life Cycle [SFL]). Implications for future research, practice, and policy makers were outlined. Our reporting adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) reporting guidelines [23].

Results

Literature Search

The search in the PubMed and Web of Science databases resulted in 764 hits and included papers from January 1, 2006, to December 31, 2022. Afterward, 140 duplicates were removed. The remaining 624 papers were subjected to title-abstract screening in an interactive selection process, leaving 118 eligible papers for the full-text review. The full-text papers were further screened to identify papers eligible for the subsequent step of data charting. During this step, additional 52 papers were excluded (see the Stage 4: Charting the Data section). These papers either did not meet the study design context (31/52, 60%) or they lacked the domain concept (15/52, 29%). Four papers reported the same study or contained parts of it, and 2 were not a full paper. A total of 66 articles were included in the data extraction phase (Table 2). The paper selection followed the PRISMA [24] approach (Figure 1), and the PRISMA-ScR checklist is presented in Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. List of included papers.

Number	Title	Study
1	Clinical Text Mining on FHIR	Daumke et al [25], 2019
2	Provenance Solutions for Medical Research in Heterogeneous IT-Infrastructure: An Implementation Roadmap	Parciak et al [26], 2019
3	BioWorkbench: a high-performance framework for managing and analyzing bioinformatics experiments	Mondelli et al [27], 2018
4	Towards structured sharing of raw and derived neuroimaging data across existing resources	Keator et al [28], 2013
5	A unified framework for managing provenance information in translational research	Sahoo et al [13], 2011
6	Towards FAIR protocols and workflows: the OpenPREDICT use case	Celebi et al [29], 2020
7	A Survey on Collecting, Managing, and Analyzing Provenance from Scripts	Pimentel et al [30], 2019
8	Reproducibility Analysis of Scientific Workflows	Bánáti et al [31], 2017
9	Implementing interoperable provenance in biomedical research	Curcin et al [1], 2014
10	Why linked data is not enough for scientists	Bechhofer et al [32], 2013
11	Representing distributed systems using the Open Provenance Model	Groth and Moreau [33], 2011
12	A Semantic Web approach to the provenance challenge	Golbeck and Hendler [34], 2008
13	Applying content management to automated provenance capture	Schuchardt et al [35], 2008
14	The Generalized Data Model for clinical research	Danese et al [36], 2019
15	FHIR Healthcare Directories: Adopting Shared Interfaces to Achieve Interoperable Medical Device Data Integration	Tyndall and Tyndall [37], 2018
16	ProvCaRe: Characterizing Scientific Reproducibility of Biomedical Research Studies using Semantic Provenance Metadata	Sahoo et al [38], 2019
17	Embedding data provenance into the Learning Health System to facilitate reproducible research	Curcin [4], 2017
18	Application of Data Provenance in Healthcare Analytics Software Information Visualisation of User Activities	Xu et al [39], 2018
19	AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance	Huber et al [40], 2020
20	Provenance for distributed biomedical workflow execution	Madougou et al [41], 2012
21	Capturing and Analyzing Provenance from Spark-based Scientific Workflows with SAMbA-RaP	Guedes et al [42], 2020
22	Why-Diff: Exploiting Provenance to Understand Outcome Differences From Non-Identical Reproduced Workflows	Thavasimani et al [43], 2019
23	Deriving scientific workflows from algebraic experiment lines: A practical approach	Marinho et al [44], 2017
24	Access control and view generation for provenance graphs	Danger et al [45], 2015
25	Provenance-based reproducibility in the Semantic Web	Moreau [46], 2011
26	PASSing the Provenance challenge	Holland et al [47], 2008
27	Decentralised provenance for healthcare data	Margheri et al [48], 2020
28	Enhancing Traceability in Clinical Research Data through a Metadata Framework.	Hume et al [12], 2020
29	Scientific Reproducibility in Biomedical Research: Provenance Metadata Ontology for Semantic Annotation of Study Description	Sahoo et al [49], 2016
30	Managing and exploiting routinely collected NHS data for research	Curcin et al [50], 2013
31	The eGenVar data management system—cataloguing and sharing sensitive data and metadata for the life sciences	Razick et al [51], 2014
32	BioQ: tracing experimental origins in public genomic databases using a novel data provenance model	Saccone et al [52], 2012
33	Applications of provenance in performance prediction and data storage optimisation	Woodman et al [53], 2017
34	ProvManager: a provenance management system for scientific workflows	Marinho et al [54], 2012
35	Provenance in collection-oriented scientific workflows	Bowers et al [55], 2008

XSL•FO
RenderX

Number	Title	Study
36	A novel approach to provenance management for privacy preservation	Can and Yilmazer [56], 2020
37	Templates as a method for implementing data provenance in decision support systems	Curcin et al [57], 2017
38	Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv	Khan et al [58], 2019
39	PAV ontology: provenance, authoring and versioning	Ciccarese et al [59], 2013
40	Enabling precision medicine via standard communication of HTS provenance, analysis, and results	Alterovitz et al [60], 2018
41	NeuroProv: Provenance data visualisation for neuroimaging analyses	Arshad et al [61], 2019
42	AVOCADO: Visualization of Workflow-Derived Data Provenance for Reproducible Biomedical Research	Stitz et al [62], 2016
43	Provenance Context Entity (PaCE): Scalable Provenance Tracking for Scientific RDF Data	Sahoo et al [63], 2010
44	A semantic proteomics dashboard (SemPoD) for data management in translational research	Jayapandian et al [2], 2012
45	Providing traceability for neuroimaging analyses	McClatchey et al [64], 2013
46	PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison	Monnin et al [65], 2019
47	Provenance data discovery through Semantic Web resources	Ornelas et al [66], 2018
48	Characterizing workflow-based activity on a production e-infrastructure using provenance data	Madougou et al [67], 2013
49	Storing, reasoning, and querying OPM-compliant scientific workflow provenance using relational databases	Lim et al [68], 2011
50	Blockchain for Healthcare: Securing Patient Data and Enabling Trusted Artificial Intelligence	Jennath et al [69], 2020
51	Bio-Swarm-Pipeline: a light-weight, extensible batch processing system for efficient biomedical data processing	Cheng et al [70], 2009
52	A Comprehensive Query Language for Provenance Information	Jabal and Bertino [71], 2018
53	Provenance trails in the Wings/Pegasus system	Kim et al [72], 2008
54	OPQL: Querying scientific workflow provenance at the graph level	Lim et al [73], 2013
55	Lightweight Distributed Provenance Model for Complex Real-world Environments	Wittner et al [74], 2022
56	A collaborative semantic-based provenance management platform for reproducibility	Samuel and Konig-Ries [75], 2022
57	A novel visualization approach for data provenance	Yazici et al [76], 2022
58	Trellis for efficient data and task management in the VA Million Veteran Program	Ross et al [77], 2021
59	ECO: the Evidence and Conclusion Ontology, an update for 2022	Nadendla et al [78], 2022
60	RepeatFS: a file system providing reproducibility through provenance and automation.	Westbrook et al [79], 2021
61	The BMS-LM ontology for biomedical data reporting throughout the life cycle of a research study: From data model to ontology	Raboudi et al [80], 2022
62	FAIRSCAPE: a Framework for FAIR and Reproducible Biomedical Analytics	Levinson et al [81], 2021
63	FAIRly big: A framework for computationally reproducible processing of large-scale data	Wagner et al [82], 2022
64	FAIR data pipeline: provenance-driven data management for traceable scientific workflows	Mitchell et al [83], 2022
65	Enabling Scientific Reproducibility through FAIR Data Management: An ontology-driven deep learning approach in the NeuroBridge Project	Wang et al [84], 2022
66	The Neuroscience Experiments System (NES)-A Software Tool to Manage Experimental Data and Its Provenance	Ruiz-Olazar et al [85], 2021

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram displaying the paper selection process with the number of studies in the identification and screening phases and all included studies in the scoping review.

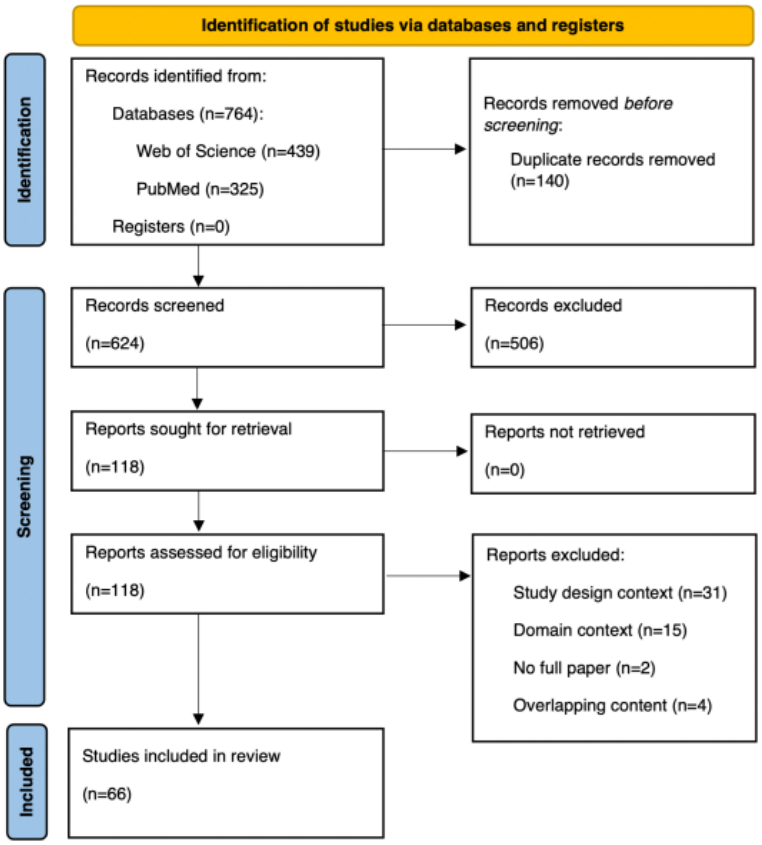

Characteristics of the Included Studies

All papers (n=66) were first published between 2006 and 2022 (Table 2). More than half of the reviewed studies were published within the past 5 years, which we selected for our scoping review. Predominantly, studies originated from the biomedical or health care domain (45/66, 68%), followed by the domain-independent studies (21/66, 32%). Document characteristics of the study corpus are provided in Multimedia Appendix 3.

All studies in this review were screened with regard to the 5 RQs described in the Methods section. The following subsections describe our findings for RQ1 to RQ5.

RQ1: Classification and Tracking of Provenance Criteria in Biomedical Workflows and Data

RQ1.1: Characteristics of Provenance Framework Types

Heterogeneous approaches for classifying and tracking provenance criteria have been reported in the selected literature (n=66). We propose to categorize them by their focus (Multimedia Appendix 4). Most articles (58/66, 88%) focused explicitly on practical provenance management approaches. The remaining theoretical frameworks (8/66, 12%) provided recommendations or reviews. They can be classified into the following subcategories.

Semantics and Models, Ontologies, and Metadata (27/58, 47%)

This comprises provenance-tracking approaches on different granularity, ontology, and model abstraction levels. The semantic Provenance Context Entity approach [63] was developed to track provenance in Resource Description Framework–based semantic web applications. An example of an annotation mechanism was introduced with collection-oriented modeling and design [55]. The Provenance Metadata Model (ProvCaRe S3), supporting scientific reproducibility, was represented with the Web Ontology Language and provenance triples served as a basis for the provenance graph [38]. Later, the NeuroBridge ontology extended ProvCaRe, combined with a deep learning model [84]. Further ontologies include the REPRODUCE-ME ontology, integrated in the CAESAR project [75], or the BioMedical Study–Lifecycle Management core ontology [80]. application programming interfaces for visualization [71] or querying purposes [28] and web services for user access to provenance data [2] were reported.

Scientific Workflows and Workflow Executions (18/58, 31%)

These are mainly Open Provenance Model (OPM)–oriented workflows [86] on different semantic levels, like in the BioWorkbench [27], OpenPREDICT [29], or Web Ontology Language projects. Provenance data were stored in relational databases, like in OPMProv [73] or in graph databases [53,77]. Querying possibilities were offered via a web service or with specific querying languages at the graph level [73].

XSL•FO
RenderX

Privacy Aspects (5/58, 9%)

Decentralized management and General Data Protection Regulation requirements led to the use of blockchain technologies [48] in combination with the PROV model standard. Another scenario incorporated blockchain in a proof-of-concept study [69] to enable an audit trail mechanism for a trusted artificial intelligence model.

Visualization Aspect (4/58, 7%)

The complexity of representing provenance information at different levels of aggregation was examined in the AVOCADO project [62]. The NeuroProv project [61] shows how visualization supports clinicians in information tracking and reproducibility analysis.

General Data Managing Tools (4/58, 7%)

Frameworks provide different modules for data and workflow provenance capture, representation, storage, comparison, and visualization [75] or automatic recomputation of arbitrary data-processing results [82].

RQ1.2: Provenance Model Characteristics

At all, 58 papers reported about provenance model characteristics. The dominant provenance models refer to the PROV [8] specification (25/58, 43%), established by the World Wide Web Consortium (W3C) as the de facto standard for provenance modeling, and the frequently used OPM [86] (17/58, 29%; Multimedia Appendix 5). Other models cite specific solutions (11/58, 19%), are concerned with metadata provision (5/58, 9%), or do not provide any information on the provenance model (8/58, 14%).

OPM is the result of 3 provenance challenges (since 2011 until today). OPM (version 1.1) is exchangeable across systems and supports a process-oriented and dataflow-oriented view. It is based on the notion of the annotated causality graph with nodes as artifacts, processes, and agents. OPM was further developed into a provenance data model. PROV [8] comprises a family of specifications for provenance, designed to promote the publication of provenance information on the web. It offers interoperability across systems and is quite generic.

Figure 2 displays the temporal evolution of the characterized frameworks depending on the applied models. We observed an increased number of papers relating to these provenance management frameworks between 2016 and 2022. At this time, the OPM and W3C PROV standards were extended. The onset of the FAIR principles [29] and the Fast Healthcare Interoperability Resources framework [40] furthermore set new requirements for modeling and implementation projects.

Figure 2. Reported provenance management frameworks per year. The size of the ring corresponds with the number of articles per year that discuss a specific model (color-coded) in the context of the respective framework. OPM: Open Provenance Model; W3C: World Wide Web Consortium.

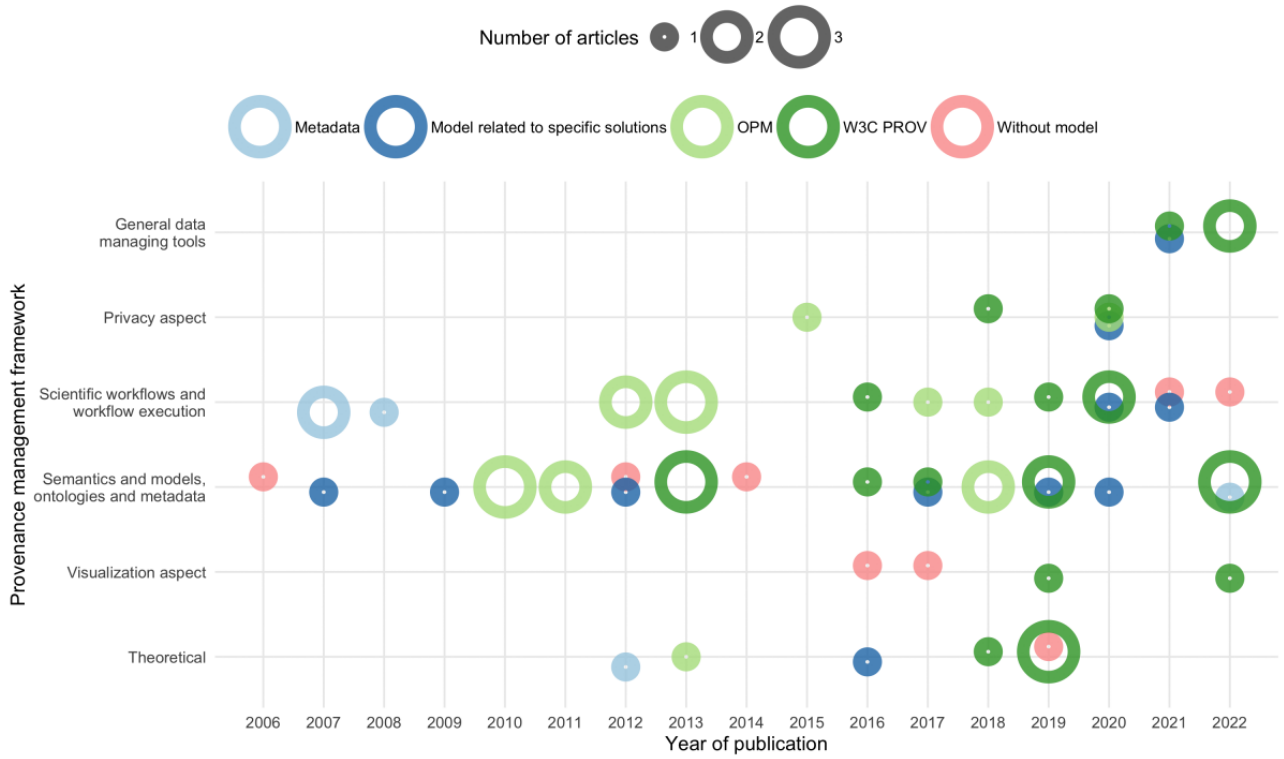

RQ1.3: Validation Status

Most of the studies (n=52) report a successful validation of their provenance solution. The use cases were mostly domain-specific (eg, as part of the AVOCADO [62] project). Other authors chose classical semantic evaluation schemes that demonstrated feasibility by responding to competency questions. Examples are the provenance challenges or proof-of-concept frameworks [13,29,56,69,72].

To pass the provenance challenges, participants needed to solve predefined provenance queries [47,55,73]. Can and Yilmazer [56] evaluated their domain-independent model with an infectious disease use case and implementing the Healthcare

XSL•FO
RenderX

Provenance Information System. Sometimes, more than one approach had been taken for evaluation (eg, use cases and factors, ontology validation, and a user-based study were applied to evaluate the CAESAR approach for scientific experiments [75]). A prototype visualization tool for provenance information used satisfaction surveys to assess the usability of the system [76].

Curcin et al [57] emphasized that the setup of provenance data needs to be modeled and verified separately from the software implementation. Precise validation methods for provenance services focus on usability, performance, scalability, fault tolerance, and functionality [64]. Moreover, they demanded more formal engineering techniques to foster provenance implementation across a broad range of software tools in the biomedical domain and beyond [1,74]. In that sense, formal validation as part of the software engineering process contributes to increased software quality, and formal validation requires testing efforts and testing evidence. However, accurate alignment of testing procedures against predefined requirements in the software life cycle could not be identified.

RQ1.4: Provenance Characteristics

The term *provenance* is subjected to an evolutionary and technical process with multifaceted meanings and roles in the selected papers (n=66). There is agreement that provenance is a piece of history. However, the focus of provenance work ranges from abstract workflow descriptions to summaries of workflow executions to more general knowledge about data sources and result dependencies [2,51,65,72,73]. For example, provenance as semantic metadata was specified in several works between 2007 and 2019. Monnin et al [65] required the encoding of provenance of pharmacogenomics knowledge units. Other works refer to data provenance as knowledge about data sources [45] or as a piece of analytic software [39], as machine-interpretable provenance of data sets, as software and computations, as metadata for all computed results [81], or as the description of the data and its original context, and tracing data history from their creation to their sharing [80].

Sahoo et al [38] stated that the provenance data model together with the PROV Ontology define the minimal categories of provenance metadata terms. Other studies discussed the combined provenance of data and workflows and introduce the terms *prospective*, *retrospective*, and *domain provenance* [1,42,59]. While prospective provenance expresses future abstract workflow information, retrospective provenance gathers past workflow execution and data derivation information. Domain-specific provenance can be defined as an extension to the PROV Ontology. Workflow provenance has repeatedly been mentioned in the context of workflow execution [27,31,34]. Wittner et al [74] introduced the term *provenance backbone*, which covers coarse granularity representation of traceable object artifacts, whereas Mitchell et al [83] expressed provenance in relation to intrinsic and extrinsic metadata. The “FAIRly big framework” [82] demonstrates how records of process provenance are captured and stored in a machine-readable, automatically re-executable way. The 7 Ws (who, what, where, why, when, which, and how) characterize provenance in the study by Ruiz-Olazar et al [85].

RQ1.5: Requirements for Provenance Frameworks

Out of 66 reviewed papers, 44 (67%) papers mentioned ≥1 functional and nonfunctional requirements for the referenced framework type. However, 33% (22/66) of the papers did not identify any specific requirements. For those studies that did, we identified 9 different word fields, matched them (Figure 3), and explained the citations (Multimedia Appendix 6).

Intensive interdisciplinary work on requirements analysis has been undertaken [75,83]. As such, a workshop with scientists from multiple disciplines (biology, computer science, ecology, and chemistry) and an exploratory study [75] contribute to requirements collection in the epidemiological field [83]. Another way to identify requirements is based on exhaustive literature research and interviews with domain specialists [85]. Figure 3 visualizes the reported provenance requirements. The most popular requirements refer to the word fields integrity (16/44, 36%) and reproducibility (13/44, 30%), followed by interoperability, traceability, and performance or scalability related topics (each 9/44, 20%). Others were related to the word fields organizational and security (each 8/44, 18%). Only a few studies reported on trust (5/44, 11%) and usability (3/44, 7%) linked approaches.

XSL•FO
RenderX

Figure 3. Reported provenance requirements or factors by word fields. The line thickness in the first level proportionally reflects the respective characteristics count. The second level displays all occurred requirement classes.

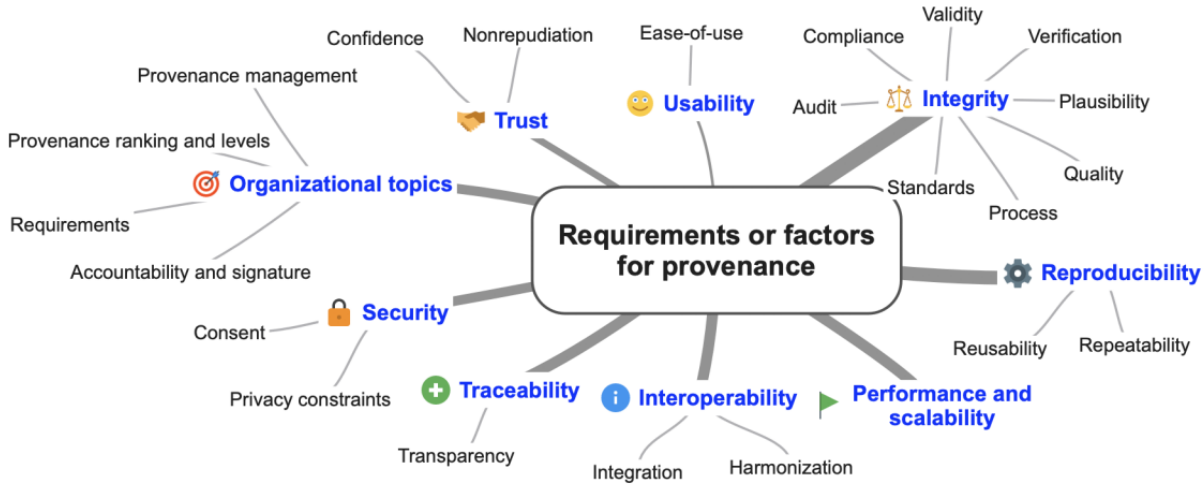

RQ1.6: Domain-Specific Conditions Including Guidelines

Some papers reported domain-specific standards for provenance (n=17). Beyond the W3C standards, we identified the Open Archival Information System Functional Model as a basis for the development of a research object concept [32]. Another example is the International Organization for Standardization (ISO) 15489-1, which defines the term *metadata* [51] or the Technical Committee 276 "Biotechnology" ISO, which standardizes provenance information for the biotechnology domain [74]. The National Institutes of Health guideline "Rigor and Reproducibility" [49] addresses topics impeding the study replicability. The FAIR principles explicitly guided data management practices in studies [77,83]. The FAIRSCAPE framework presented standards and interoperability such as JSON-LD, W3C PROV or CAT, DOI, and EVI's formal model as extension to W3C as evidence chains that support or challenge a result [81].

RQ2: Potential Value of Provenance Information

RQ2.1: Impact of Provenance Information

In our review, a total of 42 papers reported about various impacts (n=99) on different stakeholders (Multimedia Appendix 7). The availability of provenance data impacts the scientific and biomedical communities (Figure 4). With regard to the work of researchers, scientists, academia, investigators, and clinicians (n=64), most papers reported guidance benefits (19/64, 30%) and reproducibility-related effects (12/64, 19%). Considerably fewer papers observed validity (4/64, 6%), managing influence (4/64, 6%), reusability (5/64, 8%) and confidence effects (6/64, 9%). Other studies reported that provenance information impacts the willingness to share knowledge (6/64, 9%), for example, by providing a unified repository for the experimental data for the research group [82,85]. Interestingly, only 13% (8/64) of the studies discussed implications on the quality of research (eg, [46,49,52,72,77,78,84]).

Other involved team or staff members (n=22) such as developers, data managers, or domain experts were also affected by the availability of provenance information. The majority recognizes benefits in validity (5/22, 23%) [26,40,49,60,61] and managing benefits (8/22, 36%) [27,33,41,56,71,76,80,83] followed by guidance benefits (5/22, 23%) [33,53,60,67,83]. In addition, reproducibility (3/22, 14%) [26,60,79] and reusability (1/22, 4%) impacts were mentioned.

Only low impact on patients (n=7) was described, mostly referring to the consent of their data (5/7, 71%) [26,45,48,56,69] to an improved measurable patient outcome and trust in evidence for clinical recommendations (each 1/7, 14%) [57].

Only a few effects on other third parties (n=6) such as data privacy officers, authorities, government, or industry were reported. Related implications concerned mainly the evidence for data validity or sensitive data-processing solutions [26,45,48,56,69,83].

XSL•FO
RenderX

Figure 4. Reported impacts of provenance information. Level 1 presents the stakeholder groups and level 2 presents the impacts on the stakeholders. The line thickness in the second level proportionally reflects the respective counts of the characteristics.

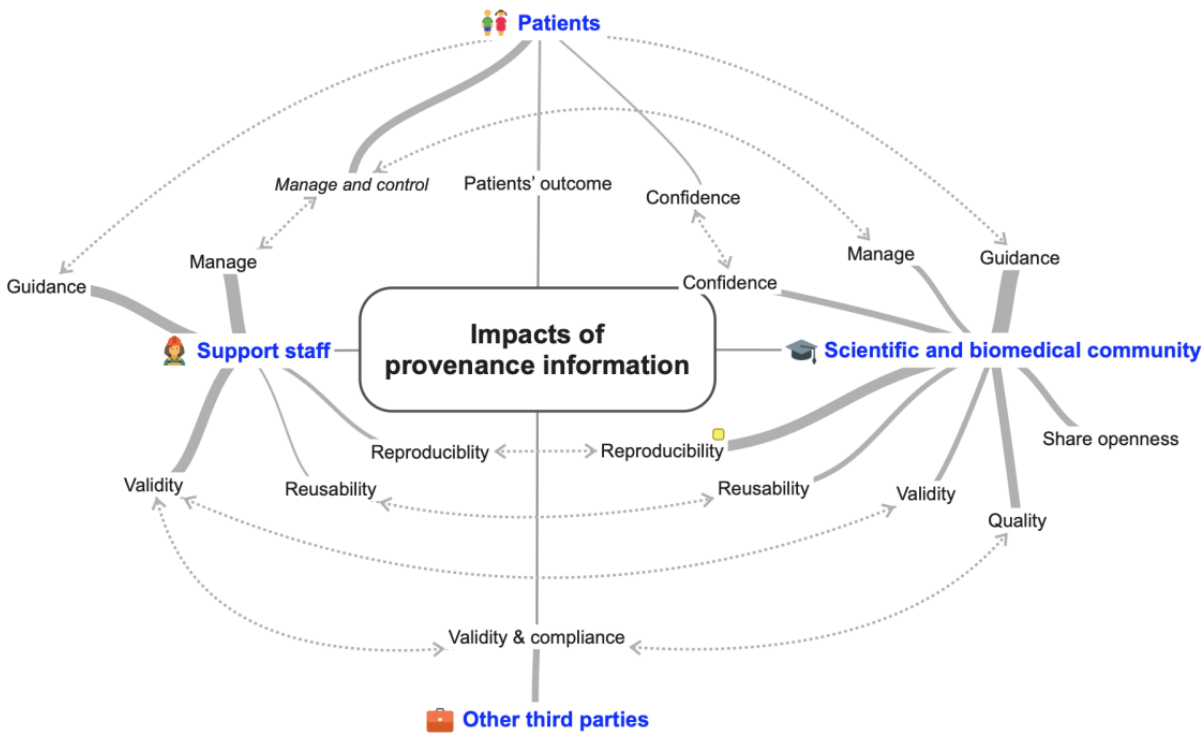

RQ2.2: Data Sources

A large number of papers (n=42) reported studies processed different types of data sources to generate provenance information, for example, neurological data [1,28,84,85], with large-scale data from the brain imaging component of the UK Biobank project [82], electronic health record data [27], study data [62], animals data [80], pathology data [74], omics data [52], biomedical or medical data [64,77,78], computational data [72], time-series analysis of neonatal intensive care unit data [81], and data from hybrid methods [69].

RQ3: Challenges, Problems, and Bottlenecks

Overall, 47 papers reported 74 distinct challenges impeding the implementation of provenance. We categorized these challenges into organizational and technical groups, provided details (Multimedia Appendix 8) and presented a temporal overview of the challenges (Figure 5). In summary, issues are related to data annotation, metadata, and modeling of provenance, as well as performance-related challenges. However, the need for more detailed provenance information; consideration of compliance managing topics (eg, security-related conditions); and adherence to quality and software engineering principles such as exchange, discovery, and interoperability emerged later in the course. Furthermore, usability and scalability questions emerged very early in context with provenance consumption.

More than three-quarters of the reported challenges are technical (64/74, 86%). Thereof, approximately one-fourth is associated with provenance granularity issues (15/64, 23%). Curcin et al [1] pointed out that a granular tracking of relevant human interactions, automated processes, or logging is needed and emphasized the difficulty of choosing a proper level of granularity of provenance and associated with this, the right semantic complexity [4,57]. Beyond that, a balanced trade-off between fast execution and provenance granularity must be found [42]. In fact, a fine-granular provenance level impacts the computing and storage resources [57,58]. Furthermore, managing sensitive data restriction requires the integration of adequate security level granularity into the provenance model [56].

Approximately one-third of the reported challenges (20/64, 31%) either mention the insufficient availability of metadata—which subsequently leads to incomplete provenance models—that claims the terminological heterogeneity in the metadata terms associated with study data sets [84] or does not conform to explicit annotation standards [82]. An improved availability of provenance metadata and FAIR enrichment of the data was demanded [29,38,80]. Furthermore, stakeholders should be involved in the semantic enrichment of provenance data [4,51]. However, during this metadata annotation phase, a lack of semiautomated procedures for ontology selection, semantic modeling, or mapping techniques was reported [2,29,51]. Although the use of existing models is encouraged [59], as it improves semantic interoperability [29], the reuse of vocabularies to represent provenance information remains an extensive task [29]. Cheng et al [70] noted that it was necessary to properly integrate domain-specific demands into the provenance model. Ruiz-Olazar et al [85] claimed that a unified data model for handling metadata is still missing.

One-fifth of the studies (14/64, 22%) reported performance problems during the acquisition of provenance data, such as

XSL•FO
RenderX

workflow overhead [54,73] and scalability [13,85] issues. Increasing data volumes hamper the processing of provenance visualization and stream handling [76]. One proposal with respect to the cost-intensive visualization was to reduce the size of large provenance graphs [39]. Other authors reported challenges with quality [26,29,36,41,52] and usability [54,64,69,70,73,75]. According to the literature, data quality and reuse are lacking due to the deficit in provenance deployment, particularly for observational and administrative studies [26]. Furthermore, the lack of information about experimental origins in genomics data and their related systematic quality control assessment reduce the quality of provenance and the level of creditability [52]. In particular, the low uptake of high-quality semantic models [9] and the unavailability of provenance in general [36] cause information loss and data quality issues. A minor concern is the usability because provenance is recognized to be still in infancy [73]. The challenge of applying more software engineering techniques (4/64, 6%) [4,32,42,67] was reported to facilitate provenance implementation across a broad range of software tools in the biomedical domain and beyond [1].

Significantly fewer organizational challenges (10/74, 13%) [1,4,38,56-58,64,69,73] were reported, partly attributable to a basic unawareness of provenance benefits and less exchange between stakeholders. Khan et al [58] stressed that provenance capture must be established as a standard practice, not as an afterthought. McClatchey et al [64] also recommended working toward gaining the stakeholder's acceptance and confidence in the infrastructure. In the same vein, it is recommended to integrate developers already in the design phase [1]. However, financial challenges were reported due to the necessary investments in provenance-enabled tooling and capabilities [4]. The upcoming relevance of patient-mediated data handling raised new challenges and requirements, especially with respect to policy and governance topics [69] and rigorous validation approaches [74].

Figure 5. Challenges per year of publication. The size of and the numbers in the circles represent the number of articles that reveal a challenge (color-coded). Note that numbers are omitted for single articles per category.

RQ4: Demands

Because of the extensive information obtained from RQ1, we extended the RQs to gain more insights into the provenance tracing and classification requirements identified in RQ1.

Interestingly, most of the 15 papers referred to claims relating to quality aspects.

For example, a more robust assessment of data quality is required [36], and clearer and more consistent policies and policy ontologies are requested to prevent disclosure of sensitive data [56] and more trained staff is required [50], including data managers, software architects, or semantic web specialists. User-friendly interfaces should help scientists in the provenance querying process [54]. Developers should not only recognize technologies but also data model recommendations during the design phase [1]. Performance of provenance reasoning needs to be improved [68] and the further development of ontologies needs to be automated [4,49]. The term "intelligible machines" rather than "intelligent machines" was suggested to better respect the specific aspects of big data technologies in medical research [57]. Integrating the Healthcare Enterprise standards, health care legacy protocols, interoperability, and legacy issues are furthermore mentioned [48], and mappings between entities of various provenance models should be completed [65]. Future integration into a recognized ISO standard similar to BioCompute was proposed [60].

XSL•FO
RenderX

RQ5: Completeness of Provenance Information

The literature predominantly reports on a qualitative evaluation of completeness during the data management processes. However, we found one study describing a data management process dealing with metadata for traceability in clinical studies, which delivered complete provenance in this respect [12]. Curcin [4] proposed to incorporate provenance information in the validation against standards.

One study applied data from 6 clinical research studies and >100 variables to evaluate the coverage of the provenance ontology in the semantic annotation of the study descriptions [49]. Two other documents invoked the need for minimal information elements to ensure sufficient process specification [28] and the existence of rich provenance information for reconstructing and rerunning pipelines [29].

A visualization of provenance data in neuroimaging took a semiqualitative approach for measuring the coverage. They mapped the metrics to use cases for the traceability of results and concluded that there is no absolute measure possible to verify the visualization approach [61]. Arshad et al [61] tested 15% of their workflows for verifiability of results, comparability of workflows, progression of the data for the analysis and origin of results, and evolution to see how data products evolved during an experiment.

Furthermore, Sahoo et al [38] examined the proportion of provenance metadata information across research articles using a qualitative hypothesis method. The method also provides a provenance ranking algorithm for the computation of a reproducibility rank for each article. The outcome of the self-contained DataLad data set presented valid, machine-actionable provenance information for every single result file of the performed data processing [82].

Numerical indication of completeness was not achieved in any of the other papers. However, the papers pointed out the advantages of provenance capture, for example, related to the longevity and accessibility of data after years [40].

Road Map for a Tailor-Made Provenance Framework

On the basis of the insights obtained from the literature review, we developed a road map for the implementation of a tailor-made provenance framework (Provenance-SFL). This approach is based on the SFL for the development, provision, and management of software [87]. The heterogeneous tracking approaches, their artifacts, and varying degrees of fulfillment of the RQs are depicted in Figure 6 and determine our main discussion points.

Figure 6. Road map toward a tailor-made provenance framework (Provenance-Software Framework Lifecycle [SFL]). The road map shows the 4 major processing phases in the inner circle segments: starting with the requirements definition, setup of the design based on the requirements, followed by coding and testing phases related to the given requirements and the implementation after successful testing. The outer and innermost circle present the mapped sections from our research questions approach to the Provenance-SFL. OPM: Open Provenance Model; W3C: World Wide Web Consortium.

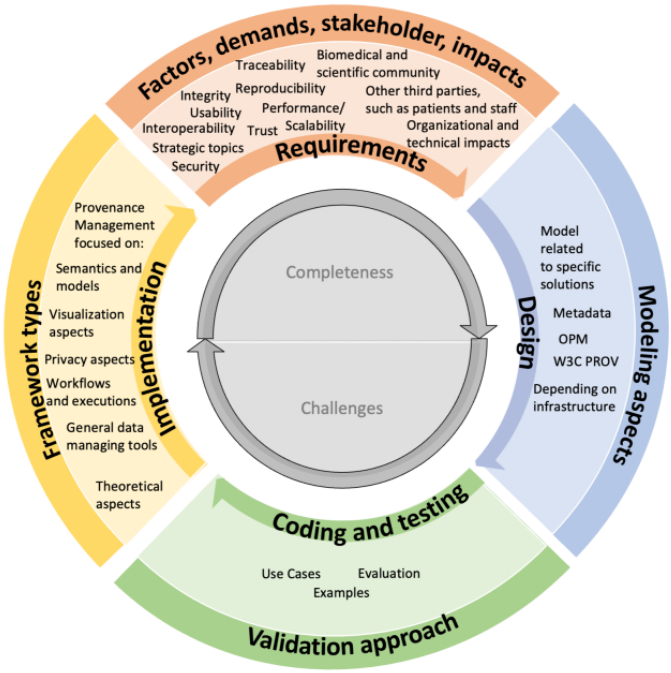

Discussion

Principal Findings

This scoping review presents various approaches and criteria for provenance tracking as a crucial aspect of research, especially in the biomedical domain. Our holistic view leads to an extensive summary of pooled research results, provides possible answers to the 5 RQs and discloses current knowledge gaps.

Following the previously published scoping review protocol and the described deviation led us to include 66 full-text papers from initially 764 papers found in the PubMed and Web of

XSL•FO
RenderX

Science databases. Using a structured and pretested data extraction sheet, contextual, but detailed, results were extracted to answer the 5 RQs defined in the protocol. The results of the scoping review led us to propose a Provenance-SFL road map, which distinguishes between the framework types and model characteristics, the validation status, and the requirement and provenance characteristics (Figure 6).

Most models in this review referred to the W3C PROV and OPM standards. As shown in Figure 2, an increased number of papers relate to the implementation frameworks published between 2016 and 2022. One reason for the increase in implementations might be the substantiation to extend W3C PROV and OPM [58]. Another reason might be the increasing awareness of data management practices. At present, heterogeneous data sources, dynamic infrastructures, data exchange across boundaries, and a lack of standards for quality measures characterize the state of electronic health record data sets [48].

A provenance framework must acknowledge the semantic complexity of the domain and its relevant facets and requirements [58] and the importance of good data management tooling and practices (Figure 2). In addition to requirements analysis, a thorough strategy is necessary to plan the typical data management steps such as collecting, managing, and analyzing data [30]. According to Curcin [4], validation readiness can be achieved by separating modeling and verification of provenance data from the software implementation. We agree that precise requirements analysis, as part of the software life cycle, and the subsequent individual life cycle steps, such as testing and maintenance procedures, support the consequent temporal evolution and hence improve the quality of provenance frameworks and applications.

When incorporated in an inspection, provenance information must be sufficient for a content-related validation against applicable and accepted standards [4]. Therefore, precise validation methods for provenance services regarding usability and performance, scalability, fault tolerance, and functionality are needed [64]. We saw that validation approaches are linked to the evolution of provenance modeling and subsequent implementation attempts. Curcin et al [1] argued that it was necessary to launch more formal software engineering techniques to foster provenance implementation across a broad range of software tools in the biomedical domain and beyond. In that sense, formal validation as part of the software engineering process contributes to increased software and data quality. Formal validation requires testing efforts and testing evidence. Accurate alignment of testing procedures against predefined requirements in the software life cycle could not be identified in the included papers.

Provenance information is of high value for the scientific and biomedical community (eg, researchers); support staff (eg, developers); patients and other third parties (eg, data privacy officer or authority; Figure 4). It is interesting to see that despite the high impact of provenance (Multimedia Appendix 7), only some stakeholders provide sufficient provenance information. Rather, it appears that responsibility for overall provenance management is being shifted to the support staff [88]. We argue that available technology, IT knowledge, and data management skills need to be paired with both domain-specific knowledge and combined with constraints of legal nature or guidance [4,50]. This complexity indeed results in a very time-consuming business. However, automation and metadata collection can support this process [4,6]. As a matter of fact, good provenance information strengthens the credibility of the data and proves that data have not been intentionally or unintentionally changed throughout the data life cycle [89].

We believe that the persisting disagreement on the interpretation of the term *provenance* hampers the uptake of existing frameworks. A unique understanding of the concepts surrounding provenance should be developed, followed by engineering efforts for modeling, implementation, and validation interventions. The ISO 8000-2:2022 standard [90] defines the term *data quality* and clearly recommends defining degrees of requirements. This definition should be considered for use in provenance systems.

With regard to the implementation of provenance systems, we observed that increasing legal and scientific demands require research projects to be implemented more transparently. However, the granularity of provenance could not yet be resolved and so-called knowledge bottlenecks [50,65] persist. It is important to understand that appropriate modeling of provenance information and effective provenance management techniques are required to protect sensitive provenance data.

It furthermore remains unclear how to scale provenance systems for high amounts of data [2,58] (eg, how to store and represent provenance information in an aggregated and efficient manner or how to assist users in sophisticated provenance queries [13]). Without doubt, automated and scalable solutions become impelling due to new challenges arising from the disposal and use of permanently increasing computing power [40]. Growing focus is on the usability of the interface, particularly when provenance systems are implemented in the broad medical community, including patients, physicians, and researchers [73].

The lack of mandatory specifications or guidelines for provenance capture might be the reason why other papers only mention partial completeness. We strongly recommend doing more research on completeness checks as part of provenance tracing. The level of completeness and accuracy of provenance information of core data elements, especially in real-world data, could reveal data integrity issues and thus, affect the overall validity of the study results. Furthermore, reproducibility significantly depends on the accuracy of provenance information. For example, Mondelli et al [27] delivered a tool for better scientific and longitudinal data management, which supports users, reproducibility by provenance, and reproduction through docker containers.

Interestingly, the concept of “quality of provenance” is not clearly defined in any of the papers included for this review. We believe that data quality issues need to be addressed to reach completeness, accuracy, and timeliness of the data and to create trust in it.

Another direction reveals the importance of good and systematic data management practices [51] and the coordination with

XSL•FO
RenderX

relevant stakeholders through the data life cycle. As such, first approaches toward interdisciplinary collaboration started within the MIRAPIE community project [91] and were also recognized in the CAESAR project [75].

In summary, our review collects and structures the challenges during the accomplishment of provenance capture (Figure 5). Challenges that address missing or lacking organizational and technical capabilities were triangulated into more specific subcategories such as organizational (eg, investment and training and administrative) and technical (eg, granularity, performance and modeling and metadata annotation, delimitation reproducibility, and replicability) challenges.

Comparison With Prior Work

In the realm of similar studies, a recent systematic literature review by Sembay et al [92] delved into provenance data management in health information systems. This review is a valuable resource if interested in the technical aspects, encompassing various methods, models, methodologies, and technologies associated with provenance data management in health information systems. It followed the guidelines proposed by Kitchenham and Charters [93] for performing systematic literature reviews in software engineering. Another systematic literature review by Ahmed et al [94] studied the impacts of data provenance in health care and General Data Protection Regulation compliance–based data provenance from a technological perspective. Although our RQ1 shares conceptual and domain-related similarities with the compared studies, there are notable differences in scope, RQs, methodological approaches, and thematic analysis focus.

More than a year after we published our protocol and concluded the thematic analysis of our review findings, a scoping review on biomedical data provenance was published [95]. The authors acknowledged and referenced our protocol. As they had anticipated, our review comprised 5 broad RQs, with the primary overlap occurring in RQ1. However, our scoping review offers more comprehensive results concerning the practical application of provenance and the associated challenges, including aspects such as completeness and validation. In addition, we provide in-depth descriptions of thematic areas and their subtopics, supported by tables, figures, and, most notably, a development of a tailor-made provenance framework road map. These aspects can be considered as the added value of our findings.

Our scoping review shares only few references with the other 2 (in detail: 11 out of 66 references are identical with references used in Johns et al [95] and 2 references are identical with those used in Sembay et al [92]). Therefore, all 4 reviews examine different research results and focus on other aspects. Ahmed et al [94] did not provide the list of the selected 59 articles.

Strengths and Limitations

This work applied a rigorous scoping review methodology using the framework developed by Arksey and O'Malley [16]. All screening stages were carried out by at least 2 independent reviews of 4 members. A previously published protocol [17] guided our review. The fact that the scoping review includes comprehensive results for the 5 related RQs and a road map for a tailor-made Provenance-SFL framework with many additional results as supplements can be considered a strength of this review. The criteria for categorization of provenance impact, requirements, and challenge classes were identified qualitatively in peer review.

One limitation may be because we excluded gray literature from our analysis.

Conclusions

In this paper, we highlighted various approaches and criteria for provenance tracking together with their referenced artifacts, and we developed a road map for a tailor-made Provenance-SFL framework.

Provenance capture benefits all stakeholders involved in data processing, but it is associated with manifold and individual challenges during design, implementation, and the active use scenario phase.

Sophisticated data management planning, documentation, metadata expression, and automation along the sensitive data-processing pipelines need to be scrutinized and implemented throughout the data life cycle and in adherence to the underlying infrastructure condition. With rising data volumes and the legal and scientific demands, there is an urgent need for greater transparency in implementing provenance systems in research projects, despite the challenges of unresolved granularity and knowledge bottlenecks. In addition, the roles and responsibilities of a data stewardship escorting the data should be expressed in this context and intensive training and education measures should be put in place. Guidance and recommendations are requested to provide the systematic measurement of provenance and calls for defining a minimal or gold standard. Governance for good data management and scale-up of good data management capabilities matter in this regard.

The mentioned artifacts, particularly those related to quality aspects, can be seen as transition points resulting from incomplete preliminary work. Therefore, harmonized engineering efforts are now necessary to overcoming the existing hurdles. Awareness of these challenges can facilitate an easier qualified and accurate provenance construction and auditable consumption while enforcing FAIR principles and interoperability standards for data sharing. The effect of provenance for data quality monitoring and the impact of expressive metadata on provenance quality can be considered as open RQs for future work.

XSL•FO
RenderX

Acknowledgments

This research is funded by the German Federal Ministry of Education and Research within the German Medical Informatics Initiative with the grant 01ZZ1801E (Medical Informatics in Research and Care in University Medicine), by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) SFB 1270/2-99150580, and by the National Research Data Infrastructure for Personal Health Data (NFDI4Health) DFG-funded project (Project 442326535). For the publication fee, the authors would like to acknowledge financial support by Heidelberg University.

Data Availability

All data generated or analyzed during this study are included in this published manuscript and its supplementary information files.

Authors' Contributions

KG contributed substantially to the conception, design, screening, data extraction, data charting, graphical representation, data analysis and interpretation, drafting all sections of the manuscript, coordinated reviewing, incorporating the comments from the coauthors, and submitting the paper. FK contributed to the discussion of the concept, screening, data extraction, data charting, and finalization of the manuscript and presented the graphical analysis of the extracted data. SG and FH contributed to data extraction and screening. FS contributed to finalization of manuscript. DW contributed to the discussion of the concept, partial screening, editorial revision, and finalization of the manuscript. AAZ contributed to the discussion of the concept, partial data analysis, screening, data extraction, data charting, editorial revision, and finalization of the manuscript. TG contributed to the discussion of the concept and finalization of the manuscript.

All authors reviewed and approved the submitted version of the manuscript. They agreed both to be personally accountable for the author's contributions and ensured that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strings and queries used for database search.
[DOCX File , 15 KB - jmir_v26i1e51297_app1.docx]

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. Filled-in checklist items for the scoping review.
[DOCX File , 84 KB - jmir_v26i1e51297_app2.docx]

Multimedia Appendix 3
Document characteristics of the study corpus. Document characteristics of the study corpus containing presentation of studies between 2006 and 2022 and allocation to the target domain.
[DOCX File , 76 KB - jmir_v26i1e51297_app3.docx]

Multimedia Appendix 4
Studies and their respective assignment to a framework type. Framework types are displayed by categories and characteristics of provenance. "Framework—practical provenance management" comprises practical efforts for development and implementation of a provenance approach. "Framework—theoretical provenance management" approach includes ideas and generic principles for provenance consideration.
[DOCX File , 83 KB - jmir_v26i1e51297_app4.docx]

Multimedia Appendix 5
Included articles and their related model. Included articles grouped by their related model respectively by using similar approaches. Counts for categories and subcategories are given per model group and approach.
[DOCX File , 16 KB - jmir_v26i1e51297_app5.docx]

Multimedia Appendix 6

XSL•FO
RenderX

Reported requirements or factors for provenance. Studies cited to classification, description, and counting of requirement or factor terms.
[DOCX File , 20 KB - jmir_v26i1e51297_app6.docx]

Multimedia Appendix 7
Reported impacts and stakeholders of provenance information. Included articles and counting of impacts per category; shows the structure and relationship between the individual stakeholders and the reported impacts. The comprehensive meaning of the impact is explained in the column "Description" by the assignment of the individual statements from the mentioned papers.
[DOCX File , 26 KB - jmir_v26i1e51297_app7.docx]

Multimedia Appendix 8
Reported technical and organizational challenges. Included articles and counting per category and subcategory for reported challenges, problems, and bottlenecks during accomplishment of provenance.
[DOCX File , 31 KB - jmir_v26i1e51297_app8.docx]

References

1. Curcin V, Miles S, Danger R, Chen Y, Bache R, Taweel A. Implementing interoperable provenance in biomedical research. Future Gener Comput Syst 2014 May;34:1-16. [doi: 10.1016/j.future.2013.12.001]
2. Jayapandian CP, Zhao M, Ewing RM, Zhang GQ, Sahoo SS. A semantic proteomics dashboard (SemPoD) for data management in translational research. BMC Syst Biol 2012 Dec 17;6(S3):S20. [doi: 10.1186/1752-0509-6-s3-s20]
3. Cuggia M, Combes S. The French health data hub and the German medical informatics initiatives: two national projects to promote data sharing in healthcare. Yearb Med Inform 2019 Aug 16;28(1):195-202 [FREE Full text] [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
4. Curcin V. Embedding data provenance into the learning health system to facilitate reproducible research. Learn Health Syst 2017 Apr 27;1(2):e10019 [FREE Full text] [doi: 10.1002/lrh2.10019] [Medline: 31245557]
5. Wodke JA, Michaelis L, Henkel R. The MeDaX knowledge graph prototype. Stud Health Technol Inform 2023 May 18;302:147-148. [doi: 10.3233/SHTI230089] [Medline: 37203634]
6. Schröder M, Staehlke S, Groth P, Nebe JB, Spors S, Krüger F. Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation. J Biomed Semantics 2022 Jan 31;13(1):4 [FREE Full text] [doi: 10.1186/s13326-021-00257-x] [Medline: 35101121]
7. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the provenance of electronic health record data matters for research: a case example using system mapping. EGEMS (Wash DC) 2014 Apr 16;2(1):1058 [FREE Full text] [doi: 10.13063/2327-9214.1058] [Medline: 25821838]
8. PROV-overview. World Wide Web Consortium. URL: https://www.w3.org/TR/prov-overview/ [accessed 2024-03-29]
9. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
10. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of FAIR (findable, accessible, interoperable, and reusable) data principles in health data stewardship practice: protocol for a scoping review. JMIR Res Protoc 2021 Mar 02;10(2):e22505 [FREE Full text] [doi: 10.2196/22505] [Medline: 33528373]
11. Jauer ML, Deserno TM. Data provenance standards and recommendations for FAIR data. Stud Health Technol Inform 2020 Jun 16;270:1237-1238. [doi: 10.3233/SHTI200380] [Medline: 32570597]
12. Hume S, Sarnikar S, Noteboom C. Enhancing traceability in clinical research data through a metadata framework. Methods Inf Med 2020 May 07;59(2-03):75-85. [doi: 10.1055/s-0040-1714393] [Medline: 32894879]
13. Sahoo SS, Nguyen V, Bodenreider O, Parikh P, Minning T, Sheth AP. A unified framework for managing provenance information in translational research. BMC Bioinformatics 2011 Nov 29;12(1):461 [FREE Full text] [doi: 10.1186/1471-2105-12-461] [Medline: 22126369]
14. Yakutovich AV, Eimre K, Schütt O, Talirz L, Adorf CS, Andersen CW, et al. AiiDAlab – an ecosystem for developing, executing, and sharing scientific workflows. Comput Mater Sci 2021 Feb;188:110165 [FREE Full text] [doi: 10.1016/j.commatsci.2020.110165]
15. Schissel D, Abla G, Flanagan SM, Greenwald M, Lee X, Romosan A, et al. Automated metadata, provenance cataloging and navigable interfaces: ensuring the usefulness of extreme-scale data. Fusion Eng Des 2014 May;89(5):745-749. [doi: 10.1016/j.fusengdes.2014.01.053]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Gierend K, Krüger F, Waltemath D, Fünfgeld M, Ganslandt T, Zeleke AA. Approaches and criteria for provenance in biomedical data sets and workflows: protocol for a scoping review. JMIR Res Protoc 2021 Nov 22;10(11):e31750 [FREE Full text] [doi: 10.2196/31750] [Medline: 34813494]

XSL•FO
RenderX

18. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
19. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
20. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
21. R Core Team. The R project for statistical computing. R Foundation for Statistical Computing. URL: https://www.r-project.org/ [accessed 2024-04-29]
22. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw 2019 Nov;4(43):1686. [doi: 10.21105/joss.01686]
23. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Daumke P, Heitmann KU, Heckmann S, Martínez-Costa C, Schulz S. Clinical text mining on FHIR. Stud Health Technol Inform 2019 Aug 21;264:83-87. [doi: 10.3233/SHTI190188] [Medline: 31437890]
26. Parciak M, Bauer C, Bender T, Lodahl R, Schreiweis B, Tute E, et al. Provenance solutions for medical research in heterogeneous IT-infrastructure: an implementation roadmap. Stud Health Technol Inform 2019 Aug 21;264:298-302. [doi: 10.3233/SHTI190231] [Medline: 31437933]
27. Mondelli ML, Magalhães T, Loss G, Wilde M, Foster I, Mattoso M, et al. BioWorkbench: a high-performance framework for managing and analyzing bioinformatics experiments. PeerJ 2018;6:e5551 [FREE Full text] [doi: 10.7717/peerj.5551] [Medline: 30186700]
28. Keator DB, Helmer K, Steffener J, Turner J, Van Erp T, Gadde S, et al. Towards structured sharing of raw and derived neuroimaging data across existing resources. Neuroimage 2013 Nov 15;82:647-661 [FREE Full text] [doi: 10.1016/j.neuroimage.2013.05.094] [Medline: 23727024]
29. Celebi R, Rebelo Moreira J, Hassan AA, Ayyar S, Ridder L, Kuhn T, et al. Towards FAIR protocols and workflows: the OpenPREDICT use case. PeerJ Comput Sci 2020;6:e281 [FREE Full text] [doi: 10.7717/peerj-cs.281] [Medline: 33816932]
30. Pimentel JF, Freire J, Murta L, Braganholo V. A survey on collecting, managing, and analyzing provenance from scripts. ACM Comput Surv 2019 Jun 18;52(3):1-38. [doi: 10.1145/3311955]
31. Bánáti A, Kacsuk P, Kozlovszky M. Reproducibility analysis of scientific workflows. Acta polytech Hung 2017 May 03;14(2):17. [doi: 10.12700/aph.14.2.2017.2.11]
32. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, et al. Why linked data is not enough for scientists. Future Gener Comput Syst 2013 Feb;29(2):599-611. [doi: 10.1016/j.future.2011.08.004]
33. Groth P, Moreau L. Representing distributed systems using the open provenance model. Future Gener Comput Syst 2011 Jun;27(6):757-765. [doi: 10.1016/j.future.2010.10.001]
34. Golbeck J, Hendler J. A Semantic web approach to the provenance challenge. Concurr Comput 2008;20(5):431-439. [doi: 10.1002/cpe.1238]
35. Schuchardt KL, Gibson T, Stephan E, Chin G. Applying content management to automated provenance capture. Concurr Comput 2008;20(5):541-554. [doi: 10.1002/cpe.1230]
36. Danese MD, Halperin M, Duryea J, Duryea R. The generalized data model for clinical research. BMC Med Inform Decis Mak 2019 Jun 24;19(1):117 [FREE Full text] [doi: 10.1186/s12911-019-0837-5] [Medline: 31234921]
37. Tyndall T, Tyndall A. FHIR healthcare directories: adopting shared interfaces to achieve interoperable medical device data integration. Stud Health Technol Inform 2018;249:181-184. [Medline: 29866978]
38. Sahoo SS, Valdez J, Kim M, Rueschman M, Redline S. ProvCaRe: characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. Int J Med Inform 2019 Jan;121:10-18 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.10.009] [Medline: 30545485]
39. Xu S, Rogers T, Fairweather E, Glenn A, Curran J, Curcin V. Application of data provenance in healthcare analytics software: information visualisation of user activities. AMIA Jt Summits Transl Sci Proc 2018;2017:263-272 [FREE Full text] [Medline: 29888084]
40. Huber SP, Zoupanos S, Uhrin M, Talirz L, Kahle L, Häuselmann R, et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. Sci Data 2020 Sep 08;7(1):300 [FREE Full text] [doi: 10.1038/s41597-020-00638-4] [Medline: 32901044]
41. Madougou S, Santcroos M, Benabdelkader A, van Schaik BD, Shahand S, Korkhov V, et al. Provenance for distributed biomedical workflow execution. Stud Health Technol Inform 2012;175:91-100. [Medline: 22941992]
42. Guedes T, Martins LB, Falci ML, Silva V, Ocaña KA, Mattoso M, et al. Capturing and analyzing provenance from spark-based scientific workflows with SAMbA-RaP. Future Gener Comput Syst 2020 Nov;112:658-669. [doi: 10.1016/j.future.2020.05.031]

XSL•FO
RenderX

43. Thavasimani P, Cala J, Missier P. Why-diff: exploiting provenance to understand outcome differences from non-identical reproduced workflows. IEEE Access 2019;7:34973-34990. [doi: 10.1109/access.2019.2903727]
44. Marinho A, de Oliveira D, Ogasawara E, Silva V, Ocaña K, Murta L, et al. Deriving scientific workflows from algebraic experiment lines: a practical approach. Future Gener Comput Syst 2017 Mar;68:111-127. [doi: 10.1016/j.future.2016.08.016]
45. Danger R, Curcin V, Missier P, Bryans J. Access control and view generation for provenance graphs. Future Gener Comput Syst 2015 Aug;49:8-27. [doi: 10.1016/j.future.2015.01.014]
46. Moreau L. Provenance-based reproducibility in the Semantic Web. J Web Semant 2011 Jul;9(2):202-221. [doi: 10.1016/j.websem.2011.03.001]
47. Holland DA, Seltzer MI, Braun U, Muniswamy‐Reddy KK. PASSing the provenance challenge. Concurr Comput 2008;20(5):531-540. [doi: 10.1002/cpe.1227]
48. Margheri A, Masi M, Miladi A, Sassone V, Rosenzweig J. Decentralised provenance for healthcare data. Int J Med Inform 2020 Sep;141:104197. [doi: 10.1016/j.ijmedinf.2020.104197] [Medline: 32540775]
49. Sahoo SS, Valdez J, Rueschman M. Scientific reproducibility in biomedical research: provenance metadata ontology for semantic annotation of study description. AMIA Annu Symp Proc 2016;2016:1070-1079 [FREE Full text] [Medline: 28269904]
50. Curcin V, Soljak M, Majeed A. Managing and exploiting routinely collected NHS data for research. Inform Prim Care 2013;20(4):225-231 [FREE Full text] [doi: 10.14236/jhi.v20i4.1] [Medline: 23890333]
51. Razick S, Močnik R, Thomas LF, Ryeng E, Drabløs F, Sætrom P. The eGenVar data management system--cataloguing and sharing sensitive data and metadata for the life sciences. Database (Oxford) 2014;2014:bau027 [FREE Full text] [doi: 10.1093/database/bau027] [Medline: 24682735]
52. Saccone SF, Quan J, Jones PL. BioQ: tracing experimental origins in public genomic databases using a novel data provenance model. Bioinformatics 2012 Apr 15;28(8):1189-1191 [FREE Full text] [doi: 10.1093/bioinformatics/bts117] [Medline: 22426342]
53. Woodman S, Hiden H, Watson P. Applications of provenance in performance prediction and data storage optimisation. Future Gener Comput Syst 2017 Oct;75:299-309. [doi: 10.1016/j.future.2017.01.003]
54. Marinho A, Murta L, Werner C, Braganholo V, Cruz SM, Ogasawara E, et al. ProvManager: a provenance management system for scientific workflows. Concurr Comput 2012;24(13):1513-1530. [doi: 10.1002/cpe.1870]
55. Bowers S, McPhillips TM, Ludäscher B. Provenance in collection‐oriented scientific workflows. Concurr Comput 2008;20(5):519-529. [doi: 10.1002/cpe.1226]
56. Can O, Yilmazer D. A novel approach to provenance management for privacy preservation. J Inf Sci 2020;46(2):147-160. [doi: 10.1177/0165551519827882]
57. Curcin V, Fairweather E, Danger R, Corrigan D. Templates as a method for implementing data provenance in decision support systems. J Biomed Inform 2017 Jan;65:1-21 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.022] [Medline: 27856379]
58. Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR. Sharing interoperable workflow provenance: a review of best practices and their practical application in CWLProv. Gigascience 2019 Nov 01;8(11):giz095 [FREE Full text] [doi: 10.1093/gigascience/giz095] [Medline: 31675414]
59. Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJ, Goble C, Clark T. PAV ontology: provenance, authoring and versioning. J Biomed Semantics 2013 Nov 22;4(1):37 [FREE Full text] [doi: 10.1186/2041-1480-4-37] [Medline: 24267948]
60. Alterovitz G, Dean D, Goble C, Crusoe MR, Soiland-Reyes S, Bell A, et al. Enabling precision medicine via standard communication of HTS provenance, analysis, and results. PLoS Biol 2018 Dec 31;16(12):e3000099 [FREE Full text] [doi: 10.1371/journal.pbio.3000099] [Medline: 30596645]
61. Arshad B, Munir K, McClatchey R, Shamdasani J, Khan Z. NeuroProv: Provenance data visualisation for neuroimaging analyses. J Comput Lang 2019 Jun;52:72-87 [FREE Full text] [doi: 10.1016/j.cola.2019.04.004]
62. Stitz H, Luger S, Streit M, Gehlenborg N. AVOCADO: visualization of workflow-derived data provenance for reproducible biomedical research. Comput Graph Forum 2016 Jun 04;35(3):481-490 [FREE Full text] [doi: 10.1111/cgf.12924] [Medline: 29973745]
63. Sahoo SS, Bodenreider O, Hitzler P, Sheth A, Thirunarayan K. Provenance context entity (PaCE): scalable provenance tracking for scientific RDF data. In: Proceedings of the 22nd International Conference on Scientific and Statistical Database Management. 2010 Presented at: SSDBM '10; June 30-July 2, 2010; Heidelberg, Germany p. 461-470 URL: https://link.springer.com/chapter/10.1007/978-3-642-13818-8_32 [doi: 10.1007/978-3-642-13818-8_32]
64. McClatchey R, Branson A, Anjum A, Bloodsworth P, Habib I, Munir K, neuGRID Consortium. Providing traceability for neuroimaging analyses. Int J Med Inform 2013 Sep;82(9):882-894. [doi: 10.1016/j.ijmedinf.2013.05.005] [Medline: 23763909]
65. Monnin P, Legrand J, Husson G, Ringot P, Tchechmedjiev A, Jonquet C, et al. PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. BMC Bioinformatics 2019 Apr 18;20(Suppl 4):139 [FREE Full text] [doi: 10.1186/s12859-019-2693-9] [Medline: 30999867]
66. Ornelas T, Braga R, David JM, Campos F, Castro G. Provenance data discovery through semantic web resources. Concurr Comput 2018;30(6):e4366. [doi: 10.1002/cpe.4366]

XSL•FO
RenderX

67. Madougou S, Shahand S, Santcroos M, van Schaik B, Benabdelkader A, van Kampen A, et al. Characterizing workflow-based activity on a production e-infrastructure using provenance data. Future Gener Comput Syst 2013 Oct;29(8):1931-1942. [doi: 10.1016/j.future.2013.04.019]
68. Lim C, Lu S, Chebotko A, Fotouhi F. Storing, reasoning, and querying OPM-compliant scientific workflow provenance using relational databases. Future Gener Comput Syst 2011 Jun;27(6):781-789. [doi: 10.1016/j.future.2010.10.013]
69. Jennath HS, Anoop VS, Asharaf S. Blockchain for healthcare: securing patient data and enabling trusted artificial intelligence. Int J Interact Multimed Artif Intell 2020;6(3):15. [doi: 10.9781/ijimai.2020.07.002]
70. Cheng X, Pizarro R, Tong Y, Zoltick B, Luo Q, Weinberger DR, et al. Bio-swarm-pipeline: a light-weight, extensible batch processing system for efficient biomedical data processing. Front Neuroinform 2009;3:35 [FREE Full text] [doi: 10.3389/neuro.11.035.2009] [Medline: 19847314]
71. Jabal AA, Bertino E. A comprehensive query language for provenance information. Int J Coop Info Syst 2018 Sep 26;27(03):1850007. [doi: 10.1142/S0218843018500077]
72. Kim J, Deelman E, Gil Y, Mehta G, Ratnakar V. Provenance trails in the Wings/Pegasus system. Concurr Comput 2008;20(5):587-597. [doi: 10.1002/cpe.1228]
73. Lim C, Lu S, Chebotko A, Fotouhi F, Kashlev A. OPQL: querying scientific workflow provenance at the graph level. Data Knowl Eng 2013 Nov;88:37-59. [doi: 10.1016/j.datak.2013.08.008]
74. Wittner R, Mascia C, Gallo M, Frexia F, Müller H, Plass M, et al. Lightweight distributed provenance model for complex real-world environments. Sci Data 2022 Aug 17;9(1):503 [FREE Full text] [doi: 10.1038/s41597-022-01537-6] [Medline: 35977957]
75. Samuel S, König-Ries B. A collaborative semantic-based provenance management platform for reproducibility. PeerJ Comput Sci 2022;8:e921 [FREE Full text] [doi: 10.7717/peerj-cs.921] [Medline: 35494870]
76. Yazici IM, Aktas MS, Yazici IM, Aktas MS. A novel visualization approach for data provenance. Concurr Comput 2022;34(9):4-12. [doi: 10.1002/cpe.6523]
77. Ross PB, Song J, Tsao PS, Pan C. Trellis for efficient data and task management in the VA Million Veteran program. Sci Rep 2021 Dec 01;11(1):23229 [FREE Full text] [doi: 10.1038/s41598-021-02569-5] [Medline: 34853358]
78. Nadendla S, Jackson R, Munro J, Quaglia F, Mészáros B, Olley D, et al. ECO: the evidence and conclusion ontology, an update for 2022. Nucleic Acids Res 2022 Jan 07;50(D1):D1515-D1521 [FREE Full text] [doi: 10.1093/nar/gkab1025] [Medline: 34986598]
79. Westbrook A, Varki E, Thomas WK. RepeatFS: a file system providing reproducibility through provenance and automation. Bioinformatics 2021 Jun 09;37(9):1292-1296 [FREE Full text] [doi: 10.1093/bioinformatics/btaa950] [Medline: 33230554]
80. Raboudi A, Allanic M, Balvay D, Hervé PY, Viel T, Yoganathan T, et al. The BMS-LM ontology for biomedical data reporting throughout the lifecycle of a research study: from data model to ontology. J Biomed Inform 2022 Mar;127:104007 [FREE Full text] [doi: 10.1016/j.jbi.2022.104007] [Medline: 35124236]
81. Levinson MA, Niestroy J, Al Manir S, Fairchild K, Lake DE, Moorman JR, et al. FAIRSCAPE: a framework for FAIR and reproducible biomedical analytics. Neuroinformatics 2021;20(1):187-202 [FREE Full text] [doi: 10.1007/s12021-021-09529-4] [Medline: 34264488]
82. Wagner AS, Waite LK, Wierzba M, Hoffstaedter F, Waite AQ, Poldrack B, et al. FAIRly big: a framework for computationally reproducible processing of large-scale data. Sci Data 2022 Mar 11;9(1):80 [FREE Full text] [doi: 10.1038/s41597-022-01163-2] [Medline: 35277501]
83. Mitchell SN, Lahiff A, Cummings N, Hollocombe J, Boskamp B, Field R, et al. FAIR data pipeline: provenance-driven data management for traceable scientific workflows. Philos Trans A Math Phys Eng Sci 2022 Oct 03;380(2233):20210300 [FREE Full text] [doi: 10.1098/rsta.2021.0300] [Medline: 35965468]
84. Wang X, Wang Y, Ambite JL, Appaji A, Lander H, Moore SM, et al. Enabling scientific reproducibility through FAIR data management: an ontology-driven deep learning approach in the NeuroBridge project. AMIA Annu Symp Proc 2022;2022:1135-1144 [FREE Full text] [Medline: 37128458]
85. Ruiz-Olazar M, Rocha ES, Vargas CD, Braghetto KR. The neuroscience experiments system (NES)-a software tool to manage experimental data and its provenance. Front Neuroinform 2021 Jan 7;15:768615 [FREE Full text] [doi: 10.3389/fninf.2021.768615] [Medline: 35069167]
86. Moreau L, Freire J, Futrelle J, McGrath R, Myers J, Paulson P. The open provenance model: an overview. In: Proceedings of the 2nd International Provenance and Annotation Workshop on Provenance and Annotation of Data and Processes. 2008 Presented at: IPAW '08; June 17-18, 2008; Salt Lake City, UT p. 323-326 URL: https://link.springer.com/chapter/10.1007/978-3-540-89965-5_31 [doi: 10.1007/978-3-540-89965-5_31]
87. Morrison P, Moye D, Pandita R, Williams L. Mapping the field of software life cycle security metrics. Inf Softw Technol 2018 Oct;102:146-159. [doi: 10.1016/j.infsof.2018.05.011]
88. Gierend K, Freiesleben S, Kadioglu D, Siegel F, Ganslandt T, Waltemath D. The status of data management practices across German medical data integration centers: mixed methods study. J Med Internet Res 2023 Nov 08;25:e48809 [FREE Full text] [doi: 10.2196/48809] [Medline: 37938878]
89. Wing JM. The data life cycle. Harv Data Sci Rev 2019 Jun 23(1.1):1-6. [doi: 10.1162/99608f92.e26845b4]

XSL•FO
RenderX

90. ISO 8000-2:2022 data quality. International Organization for Standardization. URL: https://www.iso.org/standard/85032.html [accessed 2022-12-13]
91. Gierend K, Wodke JA, Genehr S, Gött R, Henkel R, Krüger F, et al. TAPP: defining standard provenance information for clinical research data and workflows - obstacles and opportunities. In: Proceedings of the 2023 ACM Web Conference. 2023 Presented at: WWW '23; April 30-May 1, 2023; Austin, TX p. 1551-1554 URL: https://dl.acm.org/doi/10.1145/3543873.3587562 [doi: 10.1145/3543873.3587562]
92. Sembay MJ, de Macedo DD, Júnior LP, Braga RM, Sarasa-Cabezuelo A. Provenance data management in health information systems: a systematic literature review. J Pers Med 2023 Jun 13;13(6):991 [FREE Full text] [doi: 10.3390/jpm13060991] [Medline: 37373980]
93. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. School of Computer Science and Mathematics. 2007. URL: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf [accessed 2024-04-29]
94. Ahmed M, Dar AR, Helfert M, Khan A, Kim J. Data provenance in healthcare: approaches, challenges, and future directions. Sensors (Basel) 2023 Jul 18;23(14):6495 [FREE Full text] [doi: 10.3390/s23146495] [Medline: 37514788]
95. Johns M, Meurers T, Wirth FN, Haber AC, Müller A, Halilovic M, et al. Data provenance in biomedical research: scoping review. J Med Internet Res 2023 Mar 27;25:e42289 [FREE Full text] [doi: 10.2196/42289] [Medline: 36972116]

Abbreviations

FAIR: findable, accessible, interoperable, reusable
ISO: International Organization for Standardization
OPM: Open Provenance Model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
SFL: Software Framework Life Cycle
W3C: World Wide Web Consortium

Edited by A Mavragani; submitted 27.07.23; peer-reviewed by J Cai, K Budde; comments to author 10.05.24; revised version received 30.05.24; accepted 17.06.24; published 23.08.24.

Please cite as:
Gierend K, Krüger F, Genehr S, Hartmann F, Siegel F, Waltemath D, Ganslandt T, Zeleke AA
Provenance Information for Biomedical Data and Workflows: Scoping Review
J Med Internet Res 2024;26:e51297
URL: https://www.jmir.org/2024/1/e51297
doi:10.2196/51297
PMID:

©Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, Atinkut Alamirrew Zeleke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Barriers and Facilitators to the Implementation of Digital Health Services for People With Musculoskeletal Conditions in the Primary Health Care Setting: Systematic Review

Mark Leendert van Tilburg[1], MSc; Ivar Spin[1], MSc; Martijn F Pisters[2,3,4], PhD; J Bart Staal[5,6], PhD; Raymond WJG Ostelo[7,8], PhD; Miriam van der Velde[1,2], MSc; Cindy Veenhof[1,2,3], PhD; Corelien JJ Kloek[1,3], PhD

[1]Innovation of Movement Care Research Group, Research Centre for Healthy and Sustainable Living, HU University of Applied Sciences Utrecht, Utrecht, Netherlands

[2]Department of Rehabilitation, Physiotherapy Science and Sports, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[3]Center for Physical Therapy Research and Innovation in Primary Care, Julius Health Care Centers, Utrecht, Netherlands

[4]Research Group Empowering Healthy Behaviour, Department of Health Innovations and Technology, Fontys University of Applied Sciences, Eindhoven, Netherlands

[5]Musculoskeletal Rehabilitation Research Group, HAN University of Applied Sciences, Radboud University Medical Centre, Nijmegen, Netherlands

[6]Radboud Institute for Health Sciences, IQ Healthcare, Radboud University Medical Center, Nijmegen, Netherlands

[7]Department of Health Sciences, Faculty of Science, VU University, Amsterdam Movement Sciences Research Institute, Amsterdam, Netherlands

[8]Department of Epidemiology and Data Science, Amsterdam University Medical Center, Location Vrije Universiteit, Amsterdam Movement Sciences Research Institute, Amsterdam, Netherlands

Corresponding Author:
Mark Leendert van Tilburg, MSc
Innovation of Movement Care Research Group
Research Centre for Healthy and Sustainable Living
HU University of Applied Sciences Utrecht
Heidelberglaan 7
Utrecht, 3584 CS
Netherlands
Phone: 31 618302750
Email: Mark.vantilburg@hu.nl

Abstract

Background: In recent years, the effectiveness and cost-effectiveness of digital health services for people with musculoskeletal conditions have increasingly been studied and show potential. Despite the potential of digital health services, their use in primary care is lagging. A thorough implementation is needed, including the development of implementation strategies that potentially improve the use of digital health services in primary care. The first step in designing implementation strategies that fit the local context is to gain insight into determinants that influence implementation for patients and health care professionals. Until now, no systematic overview has existed of barriers and facilitators influencing the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting.

Objective: This systematic literature review aims to identify barriers and facilitators to the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting.

Methods: PubMed, Embase, and CINAHL were searched for eligible qualitative and mixed methods studies up to March 2024. Methodological quality of the qualitative component of the included studies was assessed with the Mixed Methods Appraisal Tool. A framework synthesis of barriers and facilitators to implementation was conducted using the Consolidated Framework for Implementation Research (CFIR). All identified CFIR constructs were given a reliability rating (high, medium, or low) to assess the consistency of reporting across each construct.

Results: Overall, 35 studies were included in the qualitative synthesis. Methodological quality was high in 34 studies and medium in 1 study. Barriers (–) of and facilitators (+) to implementation were identified in all 5 CFIR domains: “digital health characteristics” (ie, commercial neutral [+], privacy and safety [–], specificity [+], and good usability [+]), “outer setting” (ie, acceptance by stakeholders [+], lack of health care guidelines [–], and external financial incentives [–]), “inner setting” (ie, change

XSL•FO
RenderX

of treatment routines [+ and –], information incongruence (–), and support from colleagues [+]), "characteristics of the healthcare professionals" (ie, health care professionals' acceptance [+ and –] and job satisfaction [+ and –]), and the "implementation process" (involvement [+] and justification and delegation [–]). All identified constructs and subconstructs of the CFIR had a high reliability rating. Some identified determinants that influence implementation may be facilitators in certain cases, whereas in others, they may be barriers.

Conclusions: Barriers and facilitators were identified across all 5 CFIR domains, suggesting that the implementation process can be complex and requires implementation strategies across all CFIR domains. Stakeholders, including digital health intervention developers, health care professionals, health care organizations, health policy makers, health care funders, and researchers, can consider the identified barriers and facilitators to design tailored implementation strategies after prioritization has been carried out in their local context.

(J Med Internet Res 2024;26:e49868) doi:10.2196/49868

KEYWORDS

eHealth; primary health care; musculoskeletal problems; implementation science; systematic review; mobile phone

Introduction

Background

Approximately 1.71 billion people experience musculoskeletal conditions, which are a major contributor to health care problems worldwide [1]. Worldwide population growth and aging will increase the burden of musculoskeletal conditions on health care in the upcoming decades [2,3]. Therefore, prevention, early detection, and optimal treatment of musculoskeletal conditions, which comprise one of the largest patient groups in primary care, become increasingly important [4-6]. However, patients experience barriers that decrease access to primary care services, such as geographic and transport-related barriers, lack of health insurance, no after-hours access, and a shortage of primary health care professionals [7,8]. A potential solution to optimize prevention and treatment of musculoskeletal conditions, reduce the burden of musculoskeletal conditions on health care, and improve accessibility in primary care is the use of digital health services.

Digital health is an umbrella term encompassing eHealth and mobile health, which are defined as the use of information and communications technology in support of health and health-related fields and the use of mobile wireless technologies for health [9]. Examples are video consultations between a health care professional and patient and the integration of apps within primary care treatment. There are several potential benefits to digital health services, such as improved cost-effectiveness, more information about the health status of the patient, better communication between patients and health care professionals, and more accessibility for patients [10,11]. Previous research supports the effectiveness of digital health services in reducing pain and improving functional disability, catastrophizing, coping ability, and self-efficacy [12,13].

Despite the benefits of digital health services, their use for musculoskeletal conditions in primary care is lagging. Therefore, a thorough implementation is needed, including the development of implementation strategies that potentially improve the use of digital health services for patients with musculoskeletal conditions in primary care. Important stakeholders for designing these implementation strategies are eHealth developers, health care professionals, health care organizations, health policy makers, health care funders, and researchers. The first step to design implementation strategies for local contexts is to perform a determinant analysis in a more specific context to gain insight into determinants that influence implementation from the perspective of patients and health care professionals. Several studies have identified barriers and facilitators to the implementation of digital health services in other settings and populations [14-17]. Some of these barriers for patients or health care professionals in these settings are workflow, resistance to change, costs, reimbursement, intervention design, and digital literacy. However, it remains unclear which barriers and facilitators are applicable for patients with musculoskeletal conditions in the primary health care setting and what the overarching narrative is for this patient population and setting. A generic overview of barriers and facilitators within this more specific context, which is the aim of this systematic review, is useful as a first step for a thorough implementation. A prioritization of these barriers and facilitators for various local contexts, that is, a specific primary care physiotherapy practice, would be the next step to design fitting implementation strategies for the local context [18].

A practical theory-based framework to guide for systematical assessment of barriers and facilitators that influence implementation is the Consolidated Framework for Implementation Research (CFIR) [18]. The CFIR consolidates implementation determinants from a broad array of implementation theories and is composed of 5 domains (intervention [digital health service] characteristics, outer setting, inner setting, characteristics of individuals [health care professionals], and the implementation process), and it provides a systematic way of identifying constructs that have been associated with effective implementation. The use of a framework such as CFIR to structure the overview of barriers and facilitators allows stakeholders undertaking implementation activities to focus on barriers and facilitators that are of most interest to them more easily and design implementation strategies that are specific to their local context [19].

Objectives

No systematic overview of barriers and facilitators influencing the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting exists to support these stakeholders in designing fitting

XSL•FO
RenderX

implementation strategies. Therefore, the aim of this systematic literature review was to identify barriers and facilitators to the implementation of digital health for people with musculoskeletal conditions in the primary health care setting.

Methods

This systematic literature review of qualitative data from qualitative and mixed methods articles is reported following the enhancing transparency in reporting the synthesis of qualitative research statement [20]. Exclusively incorporating qualitative evidence in this systematic review enables a nuanced exploration of the multifaceted factors influencing implementation, providing diverse perspectives and in-depth insights from both patients and health care providers.

Search Strategy

The electronic databases PubMed, Embase, and CINAHL were searched to seek all available studies up to March 2024. The complete search strategy can be found in Multimedia Appendix 1. Inclusion and exclusion criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Domain: adults (aged ≥18 years); musculoskeletal conditions (eg, low back pain, osteoarthritis, and total knee replacement); primary health care setting (eg, general practice and physiotherapy practice)
- Determinant: the health care professional (eg, general practitioners, physiotherapists, and occupational therapists) has provided digital health (eg, synchronous patient-therapist interactions through telephone or video consultations and asynchronous physical exercise training, coaching, and monitoring using web applications, wearables, and platforms) more than once during an intervention
- Outcome: data on barriers or facilitators to the implementation of digital health services that fit into one of the Consolidated Framework for Implementation Research domains; data of patients or health care professionals
- Article type: qualitative and mixed methods designs; full text in English is available

Exclusion criteria

- Determinant: web-based training programs for health care professionals
- Article type: articles with no qualitative component, such as a quantitative survey only

Selection of Studies

The web-based screening tool Rayyan was used for the selection of studies [21]. A total of 3 reviewers (MLvT, IS, and MvdV) conducted the inclusion of eligible articles. Articles were screened independently by 2 reviewers for eligibility based on title and abstract. When an article was potentially eligible for inclusion, a full paper copy of the report was obtained and screened independently by 2 reviewers. Disagreements between the reviewers regarding an article's eligibility were resolved by discussion until consensus was reached. In case of disagreement, a fourth reviewer (CJJK) was consulted. In addition, reference tracking was performed in all included articles. The reasons for exclusion were recorded (Figure 1) [22].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

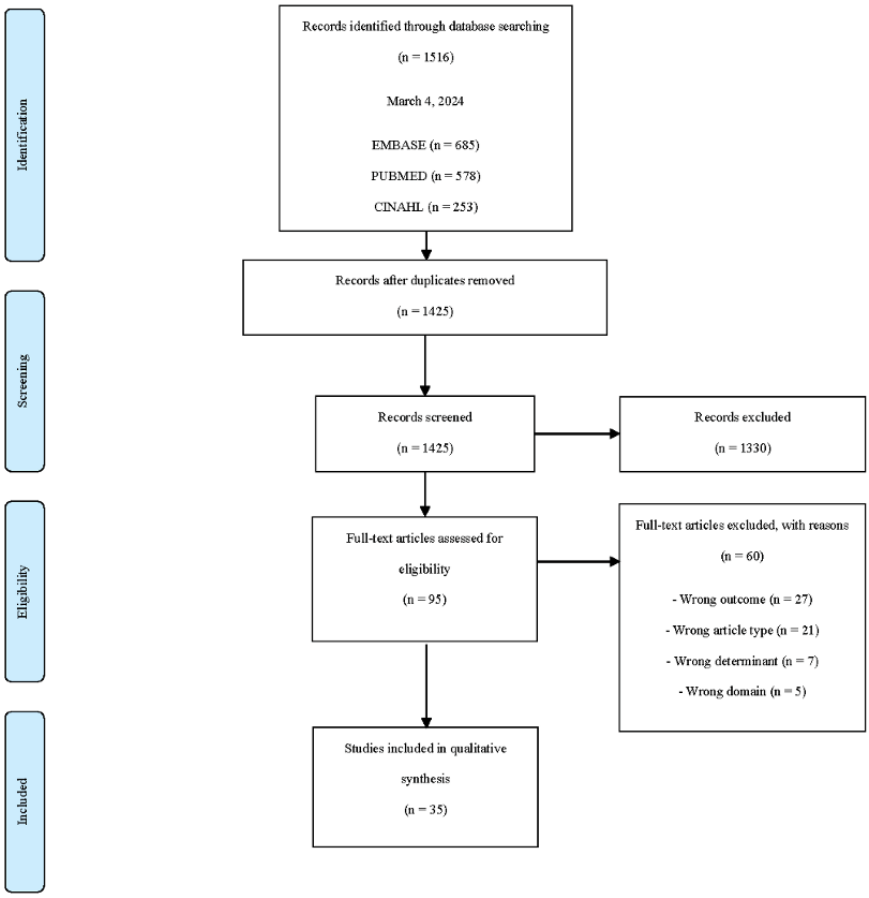

Data Extraction and Management

The reviewers extracted the following data using a standardized extraction form: first author, country, year of publication, aim, design and method of data collection and methods of analysis, sample, description of digital health service, and data on barriers or facilitators to the implementation of digital health services reported in the Results section.

Assessment of Methodological Quality

In total, 3 reviewers (MLvT, IS, and MvdV) independently assessed the methodological quality of the qualitative component of the included articles. The Mixed Methods Appraisal Tool (MMAT) was used to appraise studies for this review [23]. It is a 5-item tool designed to appraise the methodological quality of 5 categories of studies, including qualitative and mixed methods studies. The MMAT has established content validity and has been piloted across the mentioned methodologies [24]. The MMAT encompasses 2 initial screening questions: “Are there clear research questions?” and “Do the collected data allow to address the research questions?” Methodological quality assessment was only performed if “yes” could be answered for both screening questions. A detailed presentation of the individual ratings of each MMAT criterion was provided to inform the quality of the included studies. Overall sum scores were calculated based on the quality of the qualitative component only and presented as number of stars (*), with 0 and 1 star indicating low quality, 2 and 3 stars indicating medium quality, and 4 and 5 stars indicating high quality [25]. These cutoff values were determined by 2 reviewers (MLvT and IS), as the MMAT subscribes, and are arbitrary but useful for transparent data syntheses. Disagreements were resolved in a consensus meeting between the raters. When there was any disagreement, a fourth reviewer (CJJK) could be consulted but was not necessary. As the aim was to describe and synthesize a body of qualitative literature and not determine an effect size, the quality assessment was only included to inform the overall quality of the included articles and to determine the reliability rating.

Data Synthesis

A framework synthesis was performed, with secondary thematic analysis of the results section of the included articles. To synthesize the findings, the CFIR was used, using the 2009 version because the analysis began before the 2022 update [18]. Initially, MLvT and IS used an open coding process to identify barriers and facilitators to implementation and allocated them to the most fitting CFIR construct or subconstruct using a coding manual from the CFIR [26]. During the axial coding process, these open codes were organized into thematic categories representing barriers and facilitators to implementation. MLvT conducted the axial coding, which was reviewed by IS and CJJK on an iterative basis. As the thematic analysis progressed, recurring themes identified across the included studies informed the development of a comprehensive narrative for the generic overview of barriers and facilitators to implementation.

Next, all identified CFIR constructs or subconstructs were given (MLvT, IS, and MvdV) a reliability rating to review the consistency of reporting across each construct and the quality

XSL•FO
RenderX

of the studies that identified them, which was also reported in another systematic review on barriers and facilitators in another context and aims to indicate confidence in the findings [27]. All disagreements were resolved through discussion until consensus was reached. Three levels of reliability were distinguished: (1) high reliability (the construct is consistently supported by >1 study of medium quality and 1 study of high quality or the construct is supported by at least 2 studies of high quality based on the MMAT); (2) medium reliability (the construct is supported by >1 study of medium quality or the construct is identified on the basis of at least 1 high-quality study based on the MMAT); and (3) low reliability (the construct is supported only by studies of low quality or single studies of medium quality based on the MMAT).

Results

Study Selection

The literature search resulted in a total of 1516 articles found in the Embase, PubMed, and CINAHL databases. After removing duplicates, 1425 articles were screened based on title and abstract. This resulted in 95 studies that were screened full text, after which studies were excluded on outcome (n=27, 28%), article type (n=21, 22%), determinant (n=7, 7%), and domain (n=5, 5%). In 2 cases of initial disagreement between reviewers, a fourth reviewer (CJJK) was consulted. Finally, 35 studies were included in the qualitative synthesis [28-62]. No additional studies were found through reference checking. The study selection procedure is presented in Figure 1.

Study Characteristics

Characteristics of the included studies are presented in Table 1. Individual articles are ordered alphabetically within all presented tables. All included articles were published between 2011 and 2024. A total of 10 articles originated from Australia [31,38,40,44,47,48,52,54,57,60]; 9 articles originated from the Netherlands [28-30,35,39,43,45,46,62]; 4 articles originated from Canada [34,37,41,50]; 3 articles originated from the United Kingdom [51,53,59]; 2 articles originated from Brazil [36,58], Sweden [32,55], and France [42,61]; and 1 article originated from Denmark [56]. The digital health services mentioned in the included articles aimed to facilitate synchronous patient-therapist interactions through telephone or video consultations and to support asynchronous physical exercise training, coaching, and monitoring using web applications and platforms. The participants in the included articles primarily consisted of patients, physiotherapists, and general practitioners but also encompassed occupational therapists, dietitians, psychologists, and a pharmacist. Patients presented with a variety of musculoskeletal conditions, including knee and hip osteoarthritis, knee conditions, chronic nonspecific low back pain, Achilles tendinopathy, traumatic hand injury, anterior cruciate ligament reconstruction, shoulder joint replacement, and total knee replacement.

Information about methodological quality of the studies is presented in Table 2. Almost all qualitative components of the included studies, assessed with the MMAT, were of high methodological quality. In total, 31 articles scored 5 stars, and 3 articles scored 4 stars. Qualitative component of 1 article was of medium methodological quality and scored 3 stars.

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Author, year, country	Aims of the study	Methods[a]	Participants[a]	Digital health service
Aily et al [58], 2020; Brazil	• To investigate whether people with knee osteoarthritis would adhere to an exercise therapy program delivered via multiple media • To analyze the effects of intervention on pain and function • To compare acceptability of the telerehabilitation program by middle-aged and older people involved in the study	• Design: mixed methods • Data collection: focus group • Data analysis: inductive thematic analysis	• People with knee osteoarthritis (n=6)	In-person exercise therapy instructions along with a booklet and DVD to take home. Participants also received 6 motivational phone calls throughout the 12-week treatment
Arensman et al [43], 2022; the Netherlands	• To investigate patient perspectives on the acceptability, satisfaction, and performance of an app to support home-based exercise following recommendations from a physiotherapist	• Design: qualitative • Data collection: interviews • Data analysis: framework method	• People with LBPb (n=9) • Women, n=5 (56%) • Age (years), minimum-maximum: 20-71	The Physitrack app that allows physiotherapists to create and share personalized exercise programs with patients. The app allows patients to set reminders to perform their exercises, track their adherence, rate pain scores during the exercises, and send direct messages to their physiotherapists
Barton et al [40], 2022; Australia	• To explore the experiences and attitudes of people receiving physiotherapy telehealth services for musculoskeletal pain conditions during the COVID-19 pandemic	• Design: sequential mixed methods • Data collection: semistructured interviews • Data analysis: inductive thematic analysis	• People with musculoskeletal pain (n=19) • Women, n=11 (60%) • Age (years), mean 53 (SD 17)	Telehealth care from physiotherapists throughout Australia

Bossen et al [62], 2016; the Netherlands	• To develop a blended exercise therapy intervention for people with knee and hip osteoarthritis that matches the values of the users and that can be implemented in the daily routine of physiotherapists
	• To investigate the feasibility through interviews and a pilot study
		• Design: mixed methods
		• Development phase
		  • Data collection: focus group and stakeholder committee
		  • Data analysis: summarizing
		• Pilot study
		  • Data collection: interviews
		  • Data analysis: thematic trend analysis
			• Development phase
			  • Focus group: physiotherapists with extensive experience in the field of osteoarthritis (n=7)
			  • Stakeholder committee: people with knee and hip osteoarthritis, the Royal Dutch Society for Physical Therapy, 2 rehabilitation centers, the Dutch arthritis foundation, an eHealth entrepreneur, and a health insurer (n=7)
			• Pilot study: physiotherapists (n=5) and people with osteoarthritis (n=4)
				E-Exercise is a 12-week intervention, which combines visits with a physiotherapist and a web-based physical activity intervention. Patients receive 4 face-to-face sessions and are supposed to complete 12 web-based assignments. The website has a portal for both patients and physiotherapists and contains text- and video-based information

Button et al [51], 2018; the United Kingdom	• To integrate TRAKc into the physiotherapy outpatient service of 1 National Health Service Health Board and to evaluate patient and physiotherapist use and views of TRAK	• Design: mixed methods • Data collection: interviews • Data analysis: inductive thematic approach	• People with knee conditions (n=16) • Women, n=10 (63%) • Age (years), mean: 39 • Physiotherapists (n=15)	TRAK is a web-based intervention for supporting rehabilitation of knee conditions, with a potential to enhance the quality of treatment components, such as health information provision, rehabilitation monitoring, remote support, and personalized exercise progression

XSL•FO
RenderX

Author, year, country	Aims of the study	Methods[a]	Participants[a]	Digital health service
Martínez de la Cal et al [49], 2021; Spain	• To explore physiotherapists' opinions of the efficacy, benefits, and disadvantages of implementing a web-based telerehabilitation program in the treatment of chronic nonspecific low back pain • To explore the experience of physiotherapists in the management of people with chronic nonspecific LBP	• Design: qualitative • Data collection: interviews • Data analysis: thematic analysis	• Physiotherapists (n=19) • Women, n=8 (42%) • Age (years), mean 39 (SD 8) • Professional experience (years), mean 15 (SD 7)	McKenzie Exercise Therapy and electroanalgesia based on telerehabilitation with the help of 10.1 "Quad Core" tablets
Cottrell et al [52], 2017; Australia	• To evaluate service provider's views on • Current barriers to patients' accessing N/OPSCd and MDSe • The implementation of telerehabilitation within the N/OPSC and MDS	• Design: qualitative • Data collection: semistructured interviews • Data analysis: template analysis	• Service providers (n=26) • Woman, n=16 (61%) • Physiotherapy, n=15 (58%)	Telerehabilitation: delivery of rehabilitation service at a distance using telecommunications technology
Dehainault et al [42], 2024; France	• To explore the content of physical activity for low back pain that general practitioners provide and their opinion about health care smartphone eHealth apps as a support for this advice	• Design: qualitative • Data collection: semistructured interviews • Data analysis: thematic analysis	• General practitioners (n=16) • Women, n=9 (56%) • Age (years), mean (minimum-maximum): 43 (29-64)	Participants were presented with screenshots from the "Mon Coach Dos" and "Activ'Dos" mobile apps, with a standardized explanatory presentation framework. An information sheet about the apps was integrated into the slideshow presenting their description, creator, funding, and data use
Dunphy et al [53], 2017; the United Kingdom	• To evaluate the acceptability of TRAK to people following ACLf reconstruction • To evaluate the acceptability of TRAK to physiotherapists	• Design: qualitative • Data collection: semistructured interviews • Data analysis: pragmatic thematic analysis	• People following ACL reconstruction (n=17) • Woman, n=7 (41%) • Age (years), mean: 30 • Physiotherapists (n=4)	TRAK is a digital intervention developed to support self-management of knee conditions. TRAK provides a platform for individually tailored exercise programs with videos, detailed instructions and progress logs for individual exercises, a health information section, and a contact option that allows a patient to email a physiotherapist for additional support
Egerton et al [54], 2017; Australia	• To identify GPs'g perspectives on potential barriers and facilitators to engagement with a new model to support knee osteoarthritis management with remote delivery options	• Design: qualitative • Data collection: interviews • Data analysis: inductive thematic approach	• General practitioners (n=11) • Woman, n=7 (64%) • Age (years), mean (minimum-maximum): 51 (34-67)	The new model for primary care management of knee osteoarthritis includes a multidisciplinary team of health professionals using remote delivery options (primarily telephone) to provide ongoing "care support." The GP refers the patient to the "care support team" following a brief initial consultation. The "care support team" staff will have skills in health behavior change plus expertise in current best practice for knee osteoarthritis management
Eriksson et al [55], 2011; Sweden	• To describe patients' experiences with home-based physiotherapy via video link after shoulder joint replacement	• Design: qualitative • Data collection: interviews • Data analysis: qualitative content analysis	• People after shoulder joint replacement (n=10) • Age (years), median (minimum-maximum): 70 (53-85) • Woman, n=8 (80%)	A 2-month home-based video physiotherapy program, supervised by an experienced physiotherapist specializing in shoulder problems.

XSL•FO
RenderX

Author, year, country	Aims of the study	Methods[a]	Participants[a]	Digital health service
Ezzat et al [38], 2022; Australia	• To understand patients' perceived acceptability of participating in a telehealth-delivered group-based education and exercise therapy program for knee osteoarthritis	• Design: qualitative • Data collection: semistructured interviews • Data analysis: inductive analysis	• People with knee osteoarthritis (n=19) • Women, n=12 (63%) • Age (years), mean (minimum-maximum): 62 (49-72) • Using telehealth program, n=11 (58%)	GLA:D[h] is a physiotherapist-led 8-week program, which includes 2 group education sessions, followed by 12 supervised, neuromuscular exercise therapy sessions. The program is delivered via telehealth or in person
Ezzat et al [44], 2023; Australia	• To evaluate the implementation of GLA:D via telehealth in Australia using physiotherapist and patient data and applying the RE-AIM QuESTi framework	• Design: convergent mixed methods • Data collection: semistructured interviews • Data analysis: inductive, reflexive thematic analysis	• Physiotherapists (n=23) • Women, n=14 (61%)	GLA:D is a physiotherapist-led 6- to 8-week program, which includes 2 to 3 group education sessions, followed by 12 supervised, neuromuscular exercise therapy sessions. The program is delivered via telehealth or in person
Farzad et al [41], 2023; Canada	• To explore the perspective and experiences of hand therapists from different countries in providing telerehabilitation to understand the barriers and facilitators that the therapists faced during their web-based interventions in hand therapy	• Design: qualitative • Data collection: interviews • Data analysis: qualitative content analysis	• Occupational and physiotherapists (n=14) • Women, n=12 (86%) • Age (years), mean: 44	Web-based hand therapy interventions
Geraghty et al [59], 2020; the United Kingdom	• To explore patients' experiences of using the SupportBack internet intervention, both with and without physiotherapist telephone support	• Design: embedded qualitative • Data collection: interviews • Data analysis: thematic analysis	• People with nonspecific LBP (n=15) • Women, n=10 (67%) • Age (years), mean 60 (SD 15)	SupportBack is a web-based platform to support patients through a self-tailored, 6-week self-management program. Contents include exercises or a walking program with weekly goals, feedback, and advice. Patients also received 3 telephone calls from an musculoskeletal physiotherapist to provide reassurance, address concerns, problem-solve, and encourage continued engagement with the intervention and physical activity goals

Hasani et al [60], 2021; Australia	• To explore the experience of participants and physiotherapists with gym-based exercise interventions for Achilles tendinopathy with weekly telehealth monitoring	• Design: embedded qualitative
		• Data collection: interviews, focus group
		• Data analysis: thematic analysis
			• Interviews
			  • People with Achilles tendinopathy (n=8)
			  • Age (years), minimum-maximum: 38-54
			• Focus group
			  • Physiotherapists (n=7)
			  • Women, n=2 (29%)
			  • Age (years), minimum-maximum: 25-44
			  • Professional experience (years), minimum-maximum: 3-22
				Gym-based exercise program where the participants performed 4 sets of unilateral isotonic standing and seated calf raise exercises in a Smith machine (both sides, one leg at a time) 3 times per week, over 12 weeks
				Physiotherapists supervised 1 session per week via videoconference software (Zoom) that was downloaded to the participants' smartphone

XSL•FO
RenderX

Author, year, country	Aims of the study	Methods[a]	Participants[a]	Digital health service
Hinman et al [48], 2017; Australia	• To explore the experience of patients and physiotherapists using Skype as a service delivery model for physiotherapist-prescribed exercise management of knee osteoarthritis	• Design: embedded qualitative • Data collection: interviews • Data analysis: thematic analysis	• People with knee osteoarthritis (n=12) • Women, n=6 (50%) • Age (years), mean 62 (SD 7)	Participants were provided 7 internet-based Skype-delivered physical therapy sessions for 3 months, with the main purpose being to prescribe an individualized home-based strengthening program to be undertaken 3 times per week
Hjelmager et al [56], 2019; Denmark	• To identify GPs' barriers and facilitators regarding the use of health information technology in the treatment of people with LBP	• Design: qualitative • Data collection: interviews • Data analysis: inductive thematic analysis	• General practitioners (n=8) • Women, n=4 (50%) • Age (years), minimum-maximum: 41-66	Distribution of health-related information via the internet
Kairy et al [50], 2013; Canada	• To explore the perception of people who have undergone a total knee replacement concerning in-home telerehabilitation services	• Design: embedded qualitative • Data collection: interviews • Data analysis: thematic	• People after total knee replacement (n=5) • Women, n=3 (60%) • Age (years), minimum-maximum: 44-72	An in-home telerehabilitation program consisting of twice-a-week physiotherapy sessions for 8 weeks (total 16 sessions) by a videoconferencing system located in the participant's home
Kelly et al [33], 2022; Ireland	• To explore the perceptions of eHealth-mediated supported self-management from the perspective of people with musculoskeletal disorders and physiotherapists	• Design: qualitative • Data collection: semistructured interviews • Data analysis: reflexive thematic analysis	• People with musculoskeletal disorders (n=13) • Women, n=9 (69%) • Age (years), mean (minimum-maximum): 58 (24-77) • musculoskeletal physiotherapists (n=13) • Women, n=6 (46%) • Age (years), mean (minimum-maximum): 35 (26-42)	The use of technological platforms (eg, mobile, computer and tablet) in physiotherapy
Kingston et al [57], 2015; Australia	• To explore the experiences of people receiving medical treatment and rehabilitation for a traumatic hand injury	• Design: qualitative • Data collection: interviews • Data collection: inductive interpretive phenomenological approach	• People receiving medical treatment and rehabilitation for a traumatic hand injury (n=14) • Age (years), minimum-maximum: 24-82	The use of technology, namely, telehealth and the use of the internet
Kloek et al [46], 2020; the Netherlands	• To explore the experiences of physiotherapists and identify determinants that facilitate and hinder the use of the blended intervention e-Exercise	• Design: embedded mixed methods • Data collection: interviews • Data analysis: grounded theory methodology	• Physiotherapists (n=9) • Women, n=3 (33%) • Age (years), minimum-maximum: 24-59	The intervention consists of about 5 physiotherapy sessions in combination with a web-based application (E-Exercise). The web-based application contains a tailored 12-week behavioral graded activity program, videos with strength and mobility exercises, and videos and texts with information about osteoarthritis-related topics.
Lamper et al [39], 2021; the Netherlands	• To explore the feasibility of the eCoach Pain for people with chronic musculoskeletal pain and primary health care professionals	• Design: mixed methods • Data collection: focus group and interviews • Data analysis: thematic analysis	• People with chronic musculoskeletal pain (n=11) • Women, n=8 (73%) • Age (years), mean 60 (IQR 2) • Primary health care professional (n=6) • Women, n=4 (67%)	The eCoach Pain is an electronic coach to facilitate pain rehabilitation. It supports the provision of integrated rehabilitation care with a shared biopsychosocial vision on health. Both patients and primary health care professionals use the eCoach Pain. It comprises a measurement tool for assessing complexity of the pain problem, diaries, pain education sessions, monitoring options, and a chat function

XSL•FO
RenderX

Author, year, country	Aims of the study	Methods[a]	Participants[a]	Digital health service
Lawford et al [47], 2019; Australia	• To qualitatively explore whether physiotherapists' perceptions about telephone-delivered exercise therapy for people with knee osteoarthritis shifted once they had delivered exercise management advice to people with knee osteoarthritis over the telephone	• Design: embedded qualitative • Data collection: pre- and postinterviews • Data analysis: thematic analysis	• Physiotherapists (n=8) • Women, n=4 (50%) • Professional experience (years), mean 14 (SD 8)	The patients received 5 to 10 telephone consultations over a 6-month period. Physiotherapists devised goals and an action plan for each patient that involved both a structured home exercise program and a physical activity plan. Patients also had access to a study website containing video demonstrations of each exercise
Van der Meer et al [29], 2022; the Netherlands	• To assess the needs, facilitators, and barriers of the use of an eHealth application from the perspective of both orofacial physiotherapists and people with TMDj	• Design: qualitative • Data collection: interviews • Data analysis: thematic analysis	• Orofacial physical therapists (n=11) • People with TMD (n=9)	eHealth included in the health care process of people with TMD
Östlind et al [32], 2022; Sweden	• To explore the experiences of using a wearable activity tracker to monitor physical activity and the general perceptions of digital support in osteoarthritis care among people of working age with hip and knee osteoarthritis	• Design: qualitative • Data collection: focus groups • Data analysis: content analysis	• Working individuals with hip or knee osteoarthritis (n=18) • Women, n=13 (72%) • Age (years), mean 58 (SD 6)	A wearable activity tracker (Fitbit Flex 2) in combination with the Fitbit app for 12 weeks. The participants were asked to monitor their activity daily, and they received automatic feedback from the app
Palazzo et al [61], 2016; France	• To assess views of people with chronic LBP concerning barriers to home-based exercise program adherence and to record expectations regarding new technologies	• Design: qualitative • Data collection: semistructured interviews • Data analysis: iterative inductive analysis	• People with chronic LBP (n=29) • Women, n=17 (59%) • Age (years), mean (minimum-maximum): 54 (24-85)	The use of new technologies to decrease the burden of home-based exercise programs in chronic LBP
Passalent et al [37], 2022; Canada	• To understand patient perspectives of the importance of physical activity in the management of axial spondyloarthritis • To describe motivators and barriers associated with adherence to physical activity in people with axial spondyloarthritis • To explore the role of eHealth technology in facilitating physical activity in people with axial spondyloarthritis	• Design: qualitative • Data collection: semistructured interviews • Data analysis: thematic analysis	• People with axial spondyloarthritis (n=12) • Women, n=2 (17%) • Age (years), mean 46 (SD 13)	Technology for encouraging physical activity
Pereira et al [36], 2023; Brazil	• To explore beliefs and expectations of individuals with fibromyalgia about physical exercises delivered through telerehabilitation	• Design: qualitative • Data collection: semistructured interviews • Data analysis: inductive approach	• People with fibromyalgia (n=30) • Women, n=30 (100%) • Age (years), mean 45 (SD 11)	Physical therapy by telerehabilitation
Petrozzi et al [31], 2021; Australia	• To understand the experiences of people with LBP with the Mind Your Back Trial	• Design: qualitative • Data collection: interviews • Data analysis: thematic analysis	• People with LBP (n=25) • Women, n=12 (48%) • Age (years), mean 53 (SD 13)	Physical treatments combined with an internet-delivered psychosocial program called MoodGYM

XSL•FO
RenderX

Author, year, country	Aims of the study	Methods[a]	Participants[a]	Digital health service
Poolman et al [35], 2024; the Netherlands	• To gain insights in how participants experienced the Back2Action intervention	• Design: qualitative • Data collection: semistructured interviews • Data analysis: thematic analysis	• People with nonspecific LBP and neck pain (n=11) • Women, n=6 (55%) • Age (years), median 48 (IQR 25-44)	Back2Action is a newly developed biopsychosocial-blended intervention consisting of in-person physiotherapy sessions blended with psychologically informed digital health. The digital part of the intervention incorporates pain education and behavioral activation
Renard et al [34], 2022; Canada	• To evaluate the acceptability of 2 remote follow - up modalities (telephone and teleconsultation) for patients waiting for public rehabilitation services	• Design: qualitative • Data collection: semistructured interviews • Data analysis: content analysis	• People with nonurgent musculoskeletal conditions (n=10) • Women, n=7 (70%) • Age (years), mean 49 (SD 14)	Teleconsultation follow - ups
Van Tilburg et al [45], 2022; the Netherlands	• To investigate the feasibility of the e-Exercise LBP prototype for patients and physiotherapists	• Design: embedded mixed methods • Data collection: interviews • Data analysis: thematic analysis	• People with LBP (n=7) • Women, n=4 (57%) • Age (years), mean 45 (SD 11) • Physiotherapists (n=7) • Women, n=4 (57%) • Age (years), mean 37 (SD 12)	The patients received a stratified blended intervention, whereby a prognostic stratification tool, a web-based application (e-Exercise), and face-to-face physiotherapy sessions are integrated within physiotherapy treatment to create an optimal combination

Van Tilburg et al [30], 2023; the Netherlands	• To develop physiotherapy-specific matched treatment options as part of a new Stratified Blended Physiotherapy approach for people with neck and shoulder conditions
	• To investigate feasibility of the Stratified Blended Physiotherapy approach for people with neck and shoulder conditions
		• Design: 2-phase mixed methods
		• Phase 1
		  • Data collection: focus groups
		  • Data analysis: thematic analysis
		• Phase 2
		  • Data collection: semstructured interviews
		  • Data analysis: thematic analysis
			• Phase 1
			  • Stakeholders with expertise in the field of eHealth, stratified care and neck and shoulder conditions (n=17)
			• Phase 2
			  • Primary care physiotherapists (n=8)
			  • People with neck and shoulder conditions (n=13)
				A blended physiotherapy treatment (e-Exercise) for people with neck and shoulder conditions in which a smartphone app with personalized information, exercises, and physical activity modules was an integral part of physiotherapy treatment

De Vries et al [28], 2017; the Netherlands	• To explore what patient-, intervention-, and environment-related determinants are determinants of adherence to the web-based component of e-Exercise	• Design: embedded mixed methods • Data collection: interviews • Data analysis: grounded theory methodology	• People with hip and knee osteoarthritis (n=10) • Women, n=7 (70%) • Age (years), mean (minimum-maximum): 60 (51-79)	The web-based component of e-Exercise consists of a 12-week incremental physical activity program based on graded activity, strength and stability exercises, and information on osteoarthritis-related themes. The offline component consists of up to 5 face-to-face physiotherapy sessions

[a]For mixed methods designs, only the data collection, data analysis, and participants from the qualitative component are described.

[b]LBP: low back pain.

[c]TRAK: Taxonomy for the Rehabilitation of Knee Conditions.

[d]N/OPSC: Neurosurgical and Orthopaedic Physiotherapy Screening Clinic.

[e]MDS: multidisciplinary service.

[f]ACL: anterior cruciate ligament.

[g]GP: general practitioner.

[h]GLA:D: Good Life with Osteoarthritis in Denmark.

[i]RE-AIM QuEST: Reach, Effectiveness, Adoption, Implementation, and Maintenance Qualitative Evaluation for Systematic Translation.

[j]TMD: temporomandibular disorder.

XSL•FO
RenderX

Table 2. Methodological quality of the included studies.

Study	Criteria from the Mixed Methods Appraisal Tool: qualitative studies					Total number of stars (based on the qualitative component)
	1.1[a]	1.2[b]	1.3[c]	1.4[d]	1.5[e]	
Aily et al [58], 2020	0[f]	1[g]	1	1	1	***
Arensman et al [43], 2022	1	1	1	1	1	*****
Barton et al [40], 2022	1	1	1	1	1	*****
Bossen et al [62], 2016	1	1	1	0	1	****
Button et al [51], 2018	1	1	1	1	1	*****
Martínez de la Cal et al [49], 2021	1	1	1	1	1	*****
Cottrell et al [52], 2017	1	1	1	1	1	*****
Dehainault et al [42], 2024	1	1	1	1	1	*****
Dunphy et al [53], 2017	1	1	1	1	1	*****
Egerton et al [54]	1	1	1	1	1	*****
Eriksson et al [55], 2011	1	1	1	0	1	****
Ezzat et al [38], 2022	1	1	1	1	1	*****
Ezzat et al [44], 2023	1	1	1	1	1	*****
Farzad et al [41], 2023	1	1	1	1	1	*****
Geraghty et al [59], 2020	1	1	1	1	1	*****
Hasani et al [60], 2021	1	1	1	1	1	*****
Hinman et al [48], 2017	1	1	1	1	1	*****
Hjelmager et al [56], 2019	1	1	1	1	1	*****
Kairy et al [50], 2013	1	1	1	1	1	*****
Kelly et al [33], 2022	1	1	1	1	1	*****
Kingston et al [57], 2015	1	1	1	1	1	*****
Kloek et al [46], 2020	1	1	1	1	1	*****
Lamper et al [39], 2021	1	1	1	1	1	*****
Lawford et al [47], 2019	1	1	1	1	1	*****
van der Meer et al [29], 2022	1	1	1	1	1	*****
Östlind et al [32], 2022	1	1	1	1	1	*****
Palazzo et al [61], 2016	1	1	1	1	1	*****
Passalent et al [37], 2022	1	1	1	1	1	*****
Pereira et al [36], 2023	1	1	1	1	1	*****
Petrozzi et al [31], 2021	1	1	1	1	1	*****
Poolman et al [35], 2024	1	1	1	1	1	*****
Renard et al [34], 2022	1	1	1	1	1	*****
van Tilburg et al [45], 2022	1	1	1	1	1	*****
van Tilburg et al [30], 2023	1	1	1	1	1	*****
De Vries et al [28], 2017	1	1	1	0	1	****

[a]1.1=Is the qualitative approach appropriate to answer the research question?

[b]1.2=Are the qualitative data collection methods adequate to address the research question?

[c]1.3=Are the findings adequately derived from the data?

[d]1.4=Is the interpretation of results sufficiently substantiated by data?

[e]1.5=Is there coherence between qualitative data sources, collection, analysis, and interpretation?

[f]0=no.

XSL•FO
RenderX

[g]1=yes.

Barriers and Facilitators by CFIR

Overview

An overview of CFIR constructs or subconstructs influencing implementation of digital health services for patients with musculoskeletal conditions in the primary health care setting, with the sources and reliability rating, is presented in Table 3. An overview of the data synthesis supported by illustrative quotes, is presented in Table 4.

Table 3. Constructs that influence implementation of digital health.

CFIR[a] domain, construct, and subconstruct	Studies	Reliability
Innovation characteristics		
Innovation source	[42,53,56]	High
Relative advantage	[28-45,47-50,52-57,59-62]	High
Adaptability	[28-34,39,41-46,48,50,52-54,56,57,59,61,62]	High
Complexity	[28-30,37,40,43-45,48,50,51,53,55,56,59,60]	High
Design quality and packaging	[29,33,34,40,45,48,50,53,55,56,59,61,62]	High
Cost	[29,34,40,41,46,47,52,54,55,62]	High
Outer setting		
Patient needs and resources	[28,31,33-36,38,41,42,44-48,50-53,55,57-62]	High
External policy and incentives	[29,31,32,41,44,46,62]	High
Inner setting		
Networks and communications	[39,52,54]	High
Implementation climate		
Tension for change	[29,44,49,52,54,56,57]	High
Compatibility	[29,30,39,44,46-48,51,52,54]	High
Learning climate	[45,46]	High
Readiness for implementation		
Available resources	[28,30,32,33,36,38,40,41,43-45,48,49,51-53,56,60,61]	High
Access to knowledge and information	[45-47,51,52,60]	High
Characteristics of individuals		
Knowledge and beliefs about the intervention	[29,33,42,44-49,51-54,56,60,62]	High
Process		
Engaging		
Opinion leaders	[56]	Medium
Key stakeholders (health care professional)	[50,52,54,60]	High
Executing	[60]	Medium

[a]CFIR: Consolidated Framework for Implementation Research.

XSL•FO
RenderX

Table 4. Overview of the data synthesis supported by illustrative quotes.

CFIR[a] domain, construct, and subconstruct	Barriers (–) and facilitators (+) with illustrative quotes
Digital health service characteristics	
Innovation source	• Commercially neutral (+) • "Some could not see themselves 'offering patients something done by a lab' since it was unlikely the labs were 'doing this for philanthropic reasons.'" [General practitioner] [42] • Link with an institution with a good image (+) • "I think it's really good because I have heard that other hospitals... doesn't have a programme that is as good as yours and physios that look after you." [Patient] [53]

Relative advantage	• Adherence (+)
	  • "In my busy life, the reminders motivated me to take some time to get it done." [Patient] [43]
	• Self-management (+)
	  • "I think it did take away from that expectation of manual therapy. I know when people come into the clinic and they're coming in for a similar issue...because you're in the room with them quite often there is an expectation of manual therapy and being on the phone it just completely takes it out of the equation. You don't have to quite justify why you're not doing the manual therapy quite as much because it's just not an option." [Physiotherapist] [47]
	• Empowerment (+)
	  • "I was at home, I could relax, I could feel okay about what I was doing and I didn't feel intimidated at all." [Patient] [48]
	• Motivation through support (+)
	  • "So it really helped to pick me up and actually having someone talk. Physio phoned up and spoke to me a few times, and that was really, really helpful, because it's really encouraging that, 'No, it's all right keep moving, keep going.'" [Patient] [59]
	• Access to health care (+)
	  • "I think the positive would be that I could do it at home, so I didn't have to incorporate travel time and money for petrol, and trying to get there after work and all that type of stuff." [Patient] [40]
	• Societal awareness (+)
	  • "It is normal to experience back pain and it is often benign, which means that patients don't have to restrict their activities. I sometimes wish that there was a more general understanding of back pain in society. This type of information could easily be shared through an application, I think." [General practitioner] [56]
	• Continuous care chain (+)
	  • "So that when they go away, and they think about it, that they have the opportunity to you know, reengage with the information if they haven't taken it all on board at the time of the consultation." [Physiotherapist] [33]
	• Blended care (+)
	  • "A hybrid model would be awesome for people... maybe the first three weeks in person to really nail technique...then almost last three weeks via Telehealth so that they can learn to exercise in their own home environment." [Physiotherapist] [44]
	• Quality of care (+ and –)
	  • "You don't necessarily need to be putting your hands on [to assess]...that might be 30 s worth...most of the other information we get about that kind of diagnosis and planning is with our eyes, and our ears, and our brains, which we still have over a computer." [Physiotherapist] [44]
	  • "Yeah, some joint mobilities are a little tricky via the computer. Because, again, it is all about knowing and feeling the sensation and the amount of pressure. What amount of distraction and how much is too much."" [Physiotherapist] [41]
	• Patient–health care professional relationship (+ and –)
	  • "You had the time to really investigate what was motivating them or what their main issues were. Whereas I guess if you were more face-to-face and doing more of a traditional role you would be more focused on their range of movement and their strength...it is more about finding out more about them as a person and helping them to remain motivated to continue with the program. I think over the phone facilitated that to a certain degree." [Physiotherapist] [47]
	  • "Humans are social creatures and you sort of lose that when everyone's in their individual rooms online. Yes, you can still see them. Yes, you still engage with them, but it's a different engagement." [Physiotherapist] [44]
	• Privacy and safety (–)
	  • "After reinstalling the app on my phone, I had to look through my old e-mails to find the login code, and it's, of course, strange that if anyone else gets his hands on that email, they can see all my exercises and my private information." [Patient] [43]

XSL•FO
RenderX

CFIR[a] domain, construct, and subconstruct	Barriers (–) and facilitators (+) with illustrative quotes

Adaptability	• Flexibility (+)
	  • "Somehow, you want to prevent it from turning into some kind of assembly line work, and that the therapist no longer thinks about the kind of care that they provide." [Physiotherapist] [30]
	• Specificity (+)
	  • "Basically, I think it is a good app. However, the questions appear too frequent, too standard." [Patient] [39]
	• Suitability (+)
	  • "There is going to be a group both of patients and GPs who just don't want to engage with that type of model. But I think that will be the case no matter what model is designed or developed." [General practitioner] [54]
	• Evolving intervention (+)
	  • "Renewing the exercises, for me it's a good thing, because if you put a little bit of change, that makes it more enjoyable. From the moment you start a new exercise, it will stimulate you." [Patient] [61]

Complexity	• Usability factors (+)
	  • "What you often see in information provision in digital applications is that information is too complicated or too difficult to practically apply." [Physiotherapist] [30]
	• Health care professional management (+)
	  • "I think it's a shame that the physiotherapist did not know how the program worked." [Patient] [28]

Design quality and packaging	• Variety and range of content and functionalities (+)
	  • "Well I suppose the variety. It wasn't just you should be active. There were reasons behind and the self-awareness. I think it's complete." [Patient] [59]
	• Persuasive design (+)
	  • "Options such as ticking off assignments and knowledge that the physiotherapist had insight in the progress were experienced by patients as 'something that serves as a carrot.'" [Patient] [45]
	• Modality (+)
	  • "I would very much like to stress that it should be an app. It's just that it would really help because it is really tricky on the phone. It's hard in the gym I want to look at the examples really quick and remind myself... an app would be better. You can use it offline." [Patient] [53]

Cost	• Reduced number of treatment sessions (+ and –)
	  • "I think it [e-Health] can be very cost-effective for health care, especially for jaw complaints. You can see your orofacial physiotherapist less often because you already have your tools with you. I think it's a very good idea." [Patient] [29]
	  • "I believe this intervention is good for everyone, but especially for the healthcare insurers." [Patient] [46]
	• Patient expenses (+ and –)
	  • "I think the positive would be that I could do it at home, so I didn't have to incorporate travel time and money for petrol, and trying to get there after work and all that type of stuff." [Patient] [40]
	  • "General practitioners generally felt that it should be funded by sources other than patients: 'Ideally it should be...provided for free.'" [Patient] [54]

Outer setting	

XSL•FO
RenderX

CFIR[a] domain, construct, and subconstruct	Barriers (–) and facilitators (+) with illustrative quotes
Patient needs and resources	• Personal traits of patients (–) • "Sufficient Internet skills and self-discipline were described as prerequisite to use the web-based component." [Patient] [28] • Entertaining strategies (+) • "It needs to be fun...like an adventure or detective game. For people like me, it would work." [Patient] [61]
External policy and incentives	• Acceptance by stakeholders (+) • "You sometimes get this kind of pessimism from general practitioners. It's not that they don't want better interventions, it's just that they're sceptical that they will truly become a routine easily accessible part of practice."" [General practitioner] [54] • Health care guidelines (–) • "It would be easier when there would be a national e-Health policy." [Physiotherapist] [46] • Privacy regulations (–) • "We do have big confidentially chunk of potential[lity] issues. We cannot send information over an email without the patient's permission; we cannot send any personalized data over an email." [Physiotherapist] [41] • External financial incentive (–) • "According to physical therapists, this lack of financial incentive was seen as a potential barrier to use the proposed intervention in practice." [Patient] [62]
Inner setting	
Networks and communications	• Communication channels (+) • "I think it comes down to the practicalities to be honest for a lot of these systems whether they succeed or fail, and that's about taking time with the communication that was set up and getting the foundation in place to be effective." [General practitioner] [54] • Personal relationship (+) • "The idea of handing a patient over to an anonymous group of people...I don't see a great attraction." [General practitioner] [54]
Implementation climate	
Tension for change	• Accessibility of health care (+) • "My father is from a small mountain town where there is almost no mobile coverage... and we don't even talk about the internet (laughs)." [Physiotherapist] [49] • Need for trustworthy information (+) • "GPs found that patients would have difficulty in discerning accurate content from inaccurate content." [Patient] [56]
Readiness for implementation	

XSL•FO
RenderX

CFIR[a] domain, construct, and subconstruct	Barriers (–) and facilitators (+) with illustrative quotes
Compatibility	• Change of treatment routines (+ and –) • "Required them to give me a lot more input, you know, describing what's going on a little bit more, it will eliminate, I suppose, some of my normal go-to tactics." [Physiotherapist] [48] • "Once you'd done a couple, it was like—yeah, this is okay, it's going to work. And we learned as we went." [Physiotherapist] [44] • Incompatibility with other initiatives and guidelines (–) and incompatibility with existing payment structures (–) • "There's all these other things that are happening in the background that will influence how general practitioners engage with a programme like this. Thinking about how this will fit into the regular work of a general practitioner will make a big difference, to whether it succeeds or fails." [General practitioner] [54] • Information incongruence (–) • "There's a possibility that...the way that they approach the problem is going to be a little bit different to mine...every now and then it's some seemingly innocent or innocuous comment the patient turns over and then brings it back to you and you have to sort of spend time addressing that." [General practitioner] [54]
Learning climate	• Support from colleagues (+) • "Support from colleagues and the absence of a national e-Health guideline or standard influenced the use of e-Exercise." [Patient] [46] • Professional autonomy (+) • "I had the idea that I was in charge of the treatment." [Physiotherapist] [45]
Knowledge and beliefs about the intervention	
Available resources	• Technology-related issues (–) • "I didn't have earphones so I didn't quite understand this whole process. I think it was the second time that I'd used it. His receptionist was fabulous in coaching me through it and she set it up." [Patient] [40] • Time (+ and –) • "We have more time to focus on therapy, as it is web-based so many small chats with patients are cut and therapy session is focused." [Physiotherapist] [41] • "The physiotherapist thought it was too much [time spent on the app during treatment]. However, I thought, well, you know, if it is necessary, it is necessary." [Patient] [30] • Physical space (+) • "People just sort of popping in or out, or doors opening, and external noise going on, or tradies in the house next-door...that was probably a barrier." [Patient] [44] • Electronic health records • "We need to start looking at developing and rolling out, you know, electronic records...more equipment and more investment...as a nation, we are probably just a little bit behind...particularly in the public system." [Physiotherapist] [33]
Access to knowledge and information	• Health care professionals' training (+) • "I think we [as physiotherapists] got a lot of information prior the trial so for me all the documents that we received actually allowed the process to be very routine and very kind of straight forward and I think obviously once you have done one or two sessions it really starts to become just quite mechanically because you know what you are doing and you know what your expectations are." [Physiotherapist] [60] • Access for patients (+) • "Physiotherapists reported that to improve implementation in the future they would need to improve their own proficiency in using TRAK and allow patients time to explore TRAK before a consultation." [Patient] [51] • Instructions (+) • "To have a bit more resources that you could offer patients... like a video that patients could see and understand what a telehealth session is, whereas I think telehealth has been mentioned in the news a lot and certainly general practitioners use it a lot but they tend to just use it as a phone call, which I think is very, very different to the way physios utilize it." [Physiotherapist] [44]
Characteristics of individuals	

XSL•FO
RenderX

CFIR[a] domain, construct, and subconstruct	Barriers (–) and facilitators (+) with illustrative quotes
Knowledge and beliefs about the intervention	• Health care professional acceptance (+ and –) • "The way you will work and the way you will give information to the patients and counsel people. Changes are coming, I am sure of that." [Physiotherapist] [29] • "It would not suit me at all. I would have preferred to see someone in real life." [General practitioner] [42] • Health care professionals' job satisfaction (+ and –) • "It was easier on my body." [Physiotherapist] [48]
Process	
Engaging	
Opinion leaders	• Peer opinion leaders (+) • "General practitioners who were not familiar with relevant web-based information for low back pain patients expressed that it was not common to actively search for new material to present to their patients. Only if relevant material was presented to them, and preferably by a coworker who could vouch for the material, would they consider recommending it to their patients. Only if relevant material was presented to them, and preferably by a coworker who could vouch for the material, would they consider recommending it to their patients." [Patient] [56]
Executing	
Key stakeholders (health care professional)	• Involvement (+) • "I have used the research concepts to improve the telerehab that I do in the clinic...it was much more vigorous and a bit more standardised [than] what we did so I found it very satisfying and I think I have got more confidence." [Physiotherapist] [60] • Willingness to try (+) • "When asked directly about their level of willingness to try telerehabilitation if introduced into their N/OPSC&MDS facility, almost all participants stated that 'would certainly be willing to give it a go.'" [Patient] [52] • Organizational uncertainties (–) • "The argument will be, with the way that the HHS's [hospital and health service districts] are, who does it and who pays for it?" [Physiotherapist] [52] • Support team (+) • "Participants considered the telerehabilitation technical support team as part of team providing therapy and they all expressed that they felt well supported by the entire team at all times." [Patient] [50] • Justification and delegation (–) • "I thought that they might feel a bit self-conscious being at the gym and chatting away, but most of them wholeheartedly just come and had no issue whatsoever with doing it, so that was good...but I had a couple of the gyms that did not enjoy the patients having their shoes off during the sessions so we had a number of those where either negotiate with the gym or they [patients] had to wear different shoes." [Physiotherapist] [60]

[a]CFIR: Consolidated Framework for Implementation Research.

Domain 1: Digital Health Service Characteristics

Innovation Source (High Reliability)

Commercially neutral digital health services may facilitate implementation according to health care professionals because logos of, for example, pharmaceutical companies could indicate economic instead of public health interests [42,56]. A link with an institution with a good image, such as a specialized hospital, may also be a facilitator to implementation, according to health care professionals, because it promotes trust [53].

Relative Advantage (High Reliability)

When patients or health care professionals experience a relative advantage of digital health services over usual care, this may facilitate implementation. Mentioned relative advantages were promoting *adherence* [28-30,33,43,49,53,61,62], *self-management* [29-31,33,34,36,40,42-44,47,48,53,59], *empowerment* [34,42,44,48,53,55], *motivation through support* [28,31-33,37,38,43,53,59-61], *access to health care* [29,34,36,38,40,41,44,45,47-50,52-54,56,60,62], *creating societal awareness* [56] for specific health problems, and a *continuous care chain* [33,36,39,42,44,55,61]. The integration of digital health and therapy sessions *(blended care)* is described by patients and health care professionals as a facilitator because the digital health service can then be tailored to patient's needs, complementary therapy can be offered, and self-efficacy can be enhanced [28-30,32,35,39-44,48,52,53,56,62]. However, there were also some concerns among health care professionals that *quality of care* may be reduced because, for example, physical examination may not be as thorough compared to usual face-to-face care [30,32-34,36,38-44,52,54,57]. On the contrary, some health care professionals believed that extra time and

XSL•FO
RenderX

encouragement for the patient through a digital health service may result in better treatment outcomes. Advantages and disadvantages related to the *patient–health care professional relationship* were also experienced as both barriers and facilitators to implementation by patients and health care professionals [29-31,33,34,40,41,44,47,50,52,53,55]. Patients reported, for example, that when having health concerns, they prefer face-to-face reassurance over reassurance through a digital health service. Physiotherapists also had some concerns about creating a professional relationship if there are none or less face-to-face sessions. In contrast, they experienced that consulting via telephone forced them to focus on effective conversations, which allowed them to talk at a more personal level with patients. In addition, *privacy and safety* concerns may be barriers to implementation [36,40,42-44,52]. During the COVID-19 pandemic, safely providing health care from home was reported as a facilitator for implementation.

Adaptability (High Reliability)

Both health care professionals and patients agreed that adaptability of digital health services to fit the local context may be an important facilitator to implementation. Digital health services that are *flexible* to tailor to *specific* patient needs and *suitable* for various groups or subgroups of patients facilitate implementation [28-34, 39, 41, 42, 44-46, 48, 50, 52-54, 56, 57, 59, 61, 62]. Another facilitating determinant was an *evolving intervention* [43,56,61]. Use of a digital health service may increase if its content changes and information and features are continuously updated.

Complexity (High Reliability)

Complexity of digital health services that affect implementation is mostly linked to *usability*. Facilitating determinants concerning *usability* may be easy installation; easy to use; simple design and interface; simple navigation; visual support of text; and a not too wordy, manageable content [28-30,37,40,43-45,48,50,51,53,56,59,60]. Barriers concerning *usability* may be functional limitations of digital health services used in health care compared to those available on the commercial market. Another facilitating determinant to implementation was sufficient *health care professional management* for patients, such as updating relevant links and personal plans or the provision of technical aid by health care professionals to reduce complexity [28,53,55].

Design Quality and Packaging (High Reliability)

Experienced excellence in design quality and packaging of digital health services, such as *variety and range of content and functionalities* [29,30,32,35,37,45,53,59,61,62], *persuasive design* [45,53,56], and *modality* [33,34,40,45,53,61], may facilitate implementation according to both patients and health care professionals. Some mentioned functionalities are personal plans, exercise logs with speech notes as an alternative to text input, information modules with educational videos alongside written information, a progress dashboard with milestones, email or chat support, reminder tools, and feedback functions. An app was preferred over a website as modality, in particular, because of offline functionalities of an app.

Cost (High Reliability)

Costs associated with digital health services may be a barrier to implementation. Next to direct costs, a potential *reducednumber of treatment sessions* [29,46,52,55,62] may both be a barrier and facilitator to implementation. Potential loss of income because of substitution of treatment sessions was experienced as a barrier by health care professionals. However, reducing treatment sessions may be a facilitator to some health care professionals because of efficiency, and offering innovative interventions attracts new patients, which is a financial incentive. Some health care professionals mentioned that *patient expenses* for digital health services may be a barrier to implementation [29,34,40,41,47,54]. In addition, digital health services may improve access to care for patients living in remote areas and may save them travel expenses, which was experienced as a facilitator to implementation.

Domain 2: Outer Setting

Patient Needs and Resources (High Reliability)

Needs of patients may influence the participation in digital health. *Personal traits of patients*, such as poor digital literacy [28,33,42,45,46,49,51,52], poor communication skills [34,41,47], higher age [36,41,42,44,45,56], lack of motivation [28,31,35,38,42,44,45,51,53,58,61], maladaptive illness perceptions [36,61], and feeling depressed [61], may be barriers to adherence or participation and therefore to implementation of digital health in primary care. Moreover, *entertaining strategies* for performing exercises, such as exercises in a video game, might improve engagement according to patients, which facilitates implementation [61].

External Policy and Incentives (High Reliability)

Broad acceptance of digital health by patients, health care professionals, and health service funders creates trust for health care professionals that implementation is worthwhile. Therefore, *acceptance by thesestakeholders*, or even the demand by stakeholders such as patients, may be an important facilitator to implementation [29,31,32,44,54]. The absence of *health care guidelines* [44,46], standards, or protocols in using digital health and strict *privacy regulations* [41] may be barriers to implementation. Another barrier to health care professionals may be a lack of *external financial incentive* if the digital health intervention aims to substitute treatment sessions [62].

Domain 3: Inner Setting

Networks and Communications (High Reliability)

Effective, useful, and timely *channels of communication* between health care professionals involved in the use of a digital health intervention may be facilitators to implementation [39,52,54]. An example is the quality and quantity of communication between a general practitioner and a care support team that provided remotely delivered interventions in a multidisciplinary intervention. Another facilitator is some sort of *personal relationship* between health care professionals that are involved in using a digital health service [54].

XSL•FO
RenderX

Implementation Climate—Tension for Change (High Reliability)

Health care professionals and patients agreed that there is a need for change, which was a facilitator to implementation of digital health. Problems that create a tension for change are poor *accessibility to health care* [49,52,57] because of for example medical comorbidities, poor health literacy or inconvenient appointment times, large distance to health care service, high burden of health care on health care professionals, no availability of a (specialized) health care professional, and the *need for trustworthy information* [56].

Implementation Climate—Compatibility (High Reliability)

Integrating digital health services into usual care requires *change of treatment routines*, which may be a barrier to implementation, specifically because of lack of knowledge and practice to adapt routines, lack of confidence, and resistance to change of health care professionals [29,30,39,44,46-48,51,52,54]. Positive experiences with integrating digital health services into usual care may lead to more acceptability and may overcome this barrier. Moreover, *incompatibility with other initiatives and guidelines* may be barriers to implementation [54]. There are many initiatives and guidelines for management of musculoskeletal conditions, and whenever these are incompatible with a digital health service, treatment routines may become complicated and confusing. In addition, *incompatibility with existing payment structures* may lead to inequity of care and was a barrier to implementation according to health care professionals [54]. Health care professionals mentioned that information incongruence could be another barrier to implementation [54]. Safety may be affected when patient advice and information, provided by health care professionals and via digital health services, are incongruent and as a consequence cause the health care professional to spend extra time and effort to deal with conflicting messages.

Implementation Climate—Learning Climate (High Reliability)

The extent to which health care professionals feel as essential, valued, and knowledgeable partners in the implementation process creates a better climate for implementation. Facilitators to implementation of digital health services may be *support from colleagues* and that the *professional autonomy* of health care professionals was maintained [45,46].

Readiness for Implementation—Available Resources (High Reliability)

Available resources, including the availability of suitable infrastructure, may facilitate the implementation of digital health. *Technology-related issues* may be a barrier to implementation [32,33,36,38,40,43,48-51,53,55,60]. Both patients and health care professionals mentioned several technology-related issues, including troubles with initially setting up or operating the technology, insufficient battery life, poor or no internet connection, poor video quality, and audio problems. Moreover, *time* may both be a barrier as well as a facilitator to implementation [28,30,33,41,45,46,51,52,56]. Some health care professionals perceived digital health services as time saving, whereas others perceived it as an additional burden. This issue involves the lack of time to familiarize with, set up, personalize, and use the technology as well as the time investment required from health care professionals to assist patients. In addition, the lack of a quiet *physical space* for health care professionals as well as patients specifically for telerehabilitation may be a barrier to implementation [33,44,52,61]. Moreover, the lack of *electronic health records* may be a barrier to implementation [33].

Readiness for Implementation—Access to Knowledge and Information (High Reliability)

Access of health care professionals and patients to knowledge and information about the use of digital health services may be an important determinant that influences implementation. A *health care professionals' training* before using the digital health intervention may be a facilitator to implementation [41,44-47,52,60]. *Access for patients* to explore the digital health intervention before a consultation and clear *instructions* in the form of a manual, webinar, videos, or face-to-face support were facilitators to implementation [30,32,40,43,44,46,51].

Domain 4: Characteristics of Health Care Professionals (Knowledge and Beliefs About the Intervention: High Reliability)

Health care professionals' acceptance of a digital health intervention may both be a facilitator and barrier [29,33,42,44-47,52,54,60,62]. Resistance to change of health care professionals may be a barrier to implementation, but if health care professionals trust that their efforts to embrace change will be worthwhile, this may facilitate implementation. Most health care professionals are open to digital health services, as long as they have appropriate training and time to familiarize with the intervention and its content. If experiences with a digital health intervention exceeds health care professionals' expectations, this results in intrinsic motivation for the digital health intervention, which promotes implementation. The feeling of maintaining professional autonomy and confidence of health care professionals in being able to deliver the digital health intervention may also facilitate implementation. Concerns about patient information confidentiality, the belief that a digital health intervention will not be as good as face-to-face care, and providing digital health for conditions perceived as low priority may be barriers to implementation related to health care professional acceptance. Moreover, concerns that health care professionals' job satisfaction may diminish may be a barrier to implementation [48,54,60]. However, if digital health services enable more contact with patients, this is experienced as a promotion of health care professionals' satisfaction. Another contribution to satisfaction was that digital health services may lead to less physically demanding care compared to usual care, which all may facilitate implementation.

Domain 5: Process

Engaging—Opinion Leaders (Medium Reliability)

Peer opinion leaders exert influence through their representativeness and credibility. When new digital health services are presented to a health care professional by a coworker who vouches for it (*peer opinion leader*), this may facilitate implementation [56].

XSL•FO
RenderX

Engaging—Key Stakeholders (Health Care Professional; High Reliability)

Involvement of health care professionals in the implementation of digital health services is a facilitating determinant to implementation that promotes confidence in digital health services [60]. Furthermore, the willingness of health care professionals to try digital health services may facilitate implementation [52]. Organizational uncertainties among key stakeholders, such as questions like "Who does it?" and "Who pays for it?" may be barriers to implementation [54]. In addition, setting up a technical *support team* may lead to feelings of support by the health care professional, which may facilitate implementation [50].

Executing (Medium Reliability)

Executing the implementation of digital health services might require some *justification and delegation* to key involved stakeholders, such as gym staff [60]. This may be a barrier to implementation as, for example, content of the digital health intervention (eg, specific gym exercises) may not always be conventional.

Discussion

Principal Findings

In this systematic review, barriers and facilitators to the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting were identified and synthesized according to the CFIR. Barriers and facilitators were identified within all 5 CFIR domains, and almost all constructs or subconstructs of the CFIR with synthesized barriers or facilitators had high reliability. Various stakeholders are involved in the implementation of digital health services for patients with musculoskeletal conditions in the primary care setting. The current determinant analysis provides a generic overview of barriers and facilitators that may be considered by stakeholders, such as digital health intervention developers, health care professionals, health care organizations, health policy makers, health care funders, and researchers, to design fitting implementation strategies [63]. As stakeholders mainly have influence on barriers and facilitators in specific CFIR domains, main results for stakeholders will be presented and discussed accordingly.

Identified barriers and facilitators that may especially be important for developers are from the domain "digital health service characteristics." Facilitators within this domain include the flexibility of digital health services to tailor to specific patient needs, suitability for various subgroups, and high usability. Digital health service developers can consider these facilitators when developing and evaluating their product by using, for example, an eHealth framework, such as the Center for eHealth Research Roadmap [64]. An example of an existing digital health service that uses some of these facilitators is eHealth platform Physitrack, which was experienced by physiotherapists as user friendly, accessible, and helpful in providing personalized care [65,66]. Intervention design with nonoptimal usability was also identified as a barrier to implementation in other contexts, just as costs [14-17]. In this study, financial aspects, such as loss of income for health care providers because of potential substitution or patient expenses, were also shown to be important barriers to implementation for this specific context. Financial strategies to overcome these barriers when implementing digital health services for the context of patients with chronic illnesses living at home, such as changing the (patient) billing systems and fee structures, were suggested in previous research and may be relevant for developers to consider [67].

Identified barriers and facilitators that are especially important to health care professionals are from the domain "digital health service characteristics" and "outer setting." A facilitator within the domain "digital health service characteristics" is the relative advantage of digital health over usual care, such as promoting adherence, self-management, empowerment, and access to health care. Important barriers are the concern that digital health services might negatively affect patient–health care professional relationship and quality of care, experienced additional burden of digital health services, and change of treatment routines. Existing workflow was also shown to be an important barrier in other contexts [16]. To use these facilitators and overcome these barriers, health care professionals might consider using previously developed implementation strategies used in another context, such as conducting educational meetings to train and educate colleague health professionals or conducting cyclical small tests of change [68]. Personal traits of patients, such as digital literacy, maladaptive illness perceptions, poor communication skills, and lack of confidence in the patient's own physical ability, are barriers from the "outer setting." An example of a previously developed tool for physiotherapists is the use of the Checklist Blended Physiotherapy [69]. This clinical decision aid to support the physiotherapist in the decision of whether a digital health service should be an integral part of physiotherapy treatment for an individual patient might be a strategy, which has yet to be evaluated.

Identified barriers and facilitators that are especially important to health policy makers are mostly from the domain "outer setting." The lack of health care guidelines and lack of an external financial incentive were identified as barriers. The World Health Organization developed guideline recommendations on digital health services that can be used to develop guidelines for local contexts [9]. Changing reimbursement policies and clinician incentives are financial strategies that may are recommended to health policy makers [67]. Moreover, broad acceptance of digital health services by patients, health care professionals, and health service funders creates trust for health care professionals that implementation is worthwhile, which may facilitate implementation.

Identified barriers and facilitators that are especially relevant to health care organizations are mostly from the domain "inner setting." Providing access to knowledge and information about the digital health intervention was found to be an important facilitator. In addition, an opinion leader and involvement of health care professionals facilitates implementation. Therefore, it is suggested that health care organizations consider implementation strategies, such as developing and distributing educational material as well as identifying and preparing champions, and inform local opinion leaders to develop

XSL•FO
RenderX

stakeholder interrelationships [68]. Important barriers to overcome are technology-related issues and incompatibility with other initiatives, guidelines, and existing payment structures. Organizational uncertainties, such as questions like "Who does it?" and "Who pays for it?" are barriers to implementation that health care organizations must mainly overcome. To overcome these barriers, health care organizations are suggested to consider new sources of funding, involve executive boards, and try to form or join an innovation network [68].

Researchers can use the generic overview of barriers and facilitators of all domains to prioritize them for a local context, develop implementation strategies, test them, and systematically evaluate implementation outcomes. This is important because determinants are specific to the local context, and local contexts are ever changing [19].

Although several studies have identified barriers and facilitators to the implementation of digital health services in other settings than primary care or complex interventions in the primary care setting, this is the first systematic review of studies identifying and analyzing the facilitators and barriers of digital health services for people with musculoskeletal conditions in the primary health care setting. The results of this study are consistent with findings in other settings or the general health care setting [70]. Although the findings on the level of CFIR domains or subdomains are comparable to other contexts, the nuance in the description of the identified barriers and facilitators are mostly specific to primary care for patients with musculoskeletal conditions.

A strength of this systematic review is that all included articles had a mixed methods or qualitative design, and end-user perspectives of both patients and health care professionals were included, which led to a rich description of barriers and facilitators. However, it is important to note that many of the included studies did not follow a structured implementation process, and it was not possible to discuss whether implementation duration influenced the participants' perspectives. Another strength is the use of the CFIR. Synthesizing according to the CFIR makes our findings easier comparable to other implementation studies and supports the use of common terminology in this field. Despite the careful execution of this study, there are some methodological considerations. The quality of the qualitative component was assessed by presenting stars. Cutoff values were determined by the authors; however, these cutoff values are arbitrary, which may have influenced the interpretation of the quality of included articles. In addition, a reliability rating was used to indicate confidence in the findings. While this approach took consistency and quality of the studies into account, we acknowledge that tools such as GRADE-CERQual were not used, which assesses confidence in findings from a more comprehensive perspective, considering factors such as coherence and adequacy. Incorporating GRADE-CERQual or similar methods in future research could enhance confidence in findings of a qualitative data synthesis [71]. The context of this review was digital health services, the primary care setting, and musculoskeletal conditions. People with musculoskeletal conditions are one of the largest patient groups in the primary health care setting. Although this patient group is very heterogenous, there are some transcendent key recommendations for patients with musculoskeletal conditions in primary health care, which makes the context sufficiently specific to inform relevant stakeholders [72]. Specific types of digital health services researched in the included articles were also very heterogenous. Therefore, it was not possible to specify barriers and facilitators to implementation for different types of digital health services. This should be considered when developing implementation strategies for specific digital health services. This systematic review provides a generic overview, and reliability was presented on the level of subconstructs and not on the level of individual determinants. Therefore, a prioritization of determinants should be carried out for the local context, as a first step in designing implementation strategies [19].

Conclusions

This systematic review provides an extensive description of the barriers and facilitators to the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting. The findings are based on the synthesis of 35 qualitative and mixed methods articles through the CFIR. Barriers and facilitators were identified across all 5 CFIR domains, and nearly all constructs or subconstructs of the CFIR with synthesized barriers or facilitators had high reliability. This suggests that the implementation process can be complex and requires implementation strategies across all CFIR domains. Stakeholders, such as digital health intervention developers, health care professionals, health care organizations, health policy makers, health care funders, and researchers, can consider the identified barriers and facilitators to design tailored implementation strategies after a prioritization has been carried out in their local context.

Acknowledgments

The authors would like to thank the authors of all data used in the review.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Search strategy.
[DOCX File , 17 KB - jmir_v26i1e49868_app1.docx]

Multimedia Appendix 2
Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) checklist.
[DOCX File , 32 KB - jmir_v26i1e49868_app2.docx]

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
2. World report on ageing and health. World Health Organization. 2015. URL: https://www.who.int/publications/i/item/9789241565042#:~:text=The%20World%20report%20on%20ageing,new%20concept%20of%20functional%20ability [accessed 2024-04-29]
3. GBD 2019 DiseasesInjuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
4. Friedberg MW, Hussey PS, Schneider EC. Primary care: a critical review of the evidence on quality and costs of health care. Health Aff (Millwood) 2010 May;29(5):766-772. [doi: 10.1377/hlthaff.2010.0025] [Medline: 20439859]
5. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. Milbank Q 2005 Oct 03;83(3):457-502 [FREE Full text] [doi: 10.1111/j.1468-0009.2005.00409.x] [Medline: 16202000]
6. Shi L. The impact of primary care: a focused review. Scientifica (Cairo) 2012;2012:432892 [FREE Full text] [doi: 10.6064/2012/432892] [Medline: 24278694]
7. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health 2015 Jun;129(6):611-620. [doi: 10.1016/j.puhe.2015.04.001] [Medline: 26025176]
8. Bodenheimer T, Pham HH. Primary care: current problems and proposed solutions. Health Aff (Millwood) 2010 May;29(5):799-805. [doi: 10.1377/hlthaff.2010.0026] [Medline: 20439864]
9. WHO guideline recommendations on digital interventions for health system strengthening. World Health Organization. URL: https://www.ncbi.nlm.nih.gov/books/NBK541902/ [accessed 2024-04-29]
10. Catwell L, Sheikh A. Evaluating eHealth interventions: the need for continuous systemic evaluation. PLoS Med 2009 Aug 18;6(8):e1000126 [FREE Full text] [doi: 10.1371/journal.pmed.1000126] [Medline: 19688038]
11. Elbert NJ, van Os-Medendorp H, van Renselaar W, Ekeland AG, Hakkaart-van Roijen L, Raat H, et al. Effectiveness and cost-effectiveness of ehealth interventions in somatic diseases: a systematic review of systematic reviews and meta-analyses. J Med Internet Res 2014 Apr 16;16(4):e110 [FREE Full text] [doi: 10.2196/jmir.2790] [Medline: 24739471]
12. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020 Jun 05;22(6):e15617 [FREE Full text] [doi: 10.2196/15617] [Medline: 32501277]
13. Kloek CJ, Bossen D, Spreeuwenberg PM, Dekker J, de Bakker DH, Veenhof C. Effectiveness of a blended physical therapist intervention in people with hip osteoarthritis, knee osteoarthritis, or both: a cluster-randomized controlled trial. Phys Ther 2018 Jul 01;98(7):560-570 [FREE Full text] [doi: 10.1093/ptj/pzy045] [Medline: 29788253]
14. Mishuris RG, Stewart M, Fix GM, Marcello T, McInnes DK, Hogan TP, et al. Barriers to patient portal access among veterans receiving home-based primary care: a qualitative study. Health Expect 2015 Dec 12;18(6):2296-2305 [FREE Full text] [doi: 10.1111/hex.12199] [Medline: 24816246]
15. Kruse CS, Atkins JM, Baker TD, Gonzales EN, Paul JL, Brooks M. Factors influencing the adoption of telemedicine for treatment of military veterans with post-traumatic stress disorder. J Rehabil Med 2018 May 08;50(5):385-392 [FREE Full text] [doi: 10.2340/16501977-2302] [Medline: 29700551]
16. Hadjistavropoulos HD, Nugent MM, Dirkse D, Pugh N. Implementation of internet-delivered cognitive behavior therapy within community mental health clinics: a process evaluation using the consolidated framework for implementation research. BMC Psychiatry 2017 Sep 12;17(1):331 [FREE Full text] [doi: 10.1186/s12888-017-1496-7] [Medline: 28899365]
17. Batterham PJ, Sunderland M, Calear AL, Davey CG, Christensen H, Teesson M, et al. Developing a roadmap for the translation of e-mental health services for depression. Aust N Z J Psychiatry 2015 Sep 23;49(9):776-784. [doi: 10.1177/0004867415582054] [Medline: 25907269]
18. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]

XSL•FO
RenderX

19. Nilsen P, Bernhardsson S. Context matters in implementation science: a scoping review of determinant frameworks that describe contextual determinants for implementation outcomes. BMC Health Serv Res 2019 Mar 25;19(1):189 [FREE Full text] [doi: 10.1186/s12913-019-4015-3] [Medline: 30909897]
20. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012 Nov 27;12(1):181 [FREE Full text] [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
22. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009 Jul 21;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
23. Pluye P, Gagnon M, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. Int J Nurs Stud 2009 Apr;46(4):529-546. [doi: 10.1016/j.ijnurstu.2009.01.009] [Medline: 19233357]
24. Souto RQ, Khanassov V, Hong QN, Bush PL, Vedel I, Pluye P. Systematic mixed studies reviews: updating results on the reliability and efficiency of the Mixed Methods Appraisal Tool. Int J Nurs Stud 2015 Jan;52(1):500-501. [doi: 10.1016/j.ijnurstu.2014.08.010] [Medline: 25241931]
25. Hong QN. Reporting the results of the MMAT. Mixed Methods Appraisal Tool. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/ [accessed 2024-04-29]
26. CFIR codebook. Consolidated Framework for Implementation Research. 2014. URL: https://cfirguide.org/wp-content/uploads/2019/08/cfircodebooktemplate10-27-2014.docx [accessed 2024-04-29]
27. Bach-Mortensen AM, Lange BC, Montgomery P. Barriers and facilitators to implementing evidence-based interventions among third sector organisations: a systematic review. Implement Sci 2018 Jul 30;13(1):103 [FREE Full text] [doi: 10.1186/s13012-018-0789-7] [Medline: 30060744]
28. de Vries HJ, Kloek CJ, de Bakker DH, Dekker J, Bossen D, Veenhof C. Determinants of adherence to the online component of a blended intervention for patients with hip and/or knee osteoarthritis: a mixed methods study embedded in the e-Exercise trial. Telemed J E Health 2017 Dec;23(12):1002-1010. [doi: 10.1089/tmj.2016.0264] [Medline: 28525310]
29. van der Meer HA, de Pijper L, van Bruxvoort T, Visscher CM, Nijhuis-van der Sanden MW, Engelbert RH, et al. Using e-Health in the physical therapeutic care process for patients with temporomandibular disorders: a qualitative study on the perspective of physical therapists and patients. Disabil Rehabil 2022 Feb 16;44(4):617-624. [doi: 10.1080/09638288.2020.1775900] [Medline: 32543903]
30. van Tilburg ML, Kloek CJ, Foster NE, Ostelo RW, Veenhof C, Staal JB, et al. Development and feasibility of stratified primary care physiotherapy integrated with eHealth in patients with neck and/or shoulder complaints: results of a mixed methods study. BMC Musculoskelet Disord 2023 Mar 09;24(1):176 [FREE Full text] [doi: 10.1186/s12891-023-06272-6] [Medline: 36890570]
31. Petrozzi MJ, Spencer G, Mackey MG. A process evaluation of the Mind Your Back trial examining psychologically informed physical treatments for chronic low back pain. Chiropr Man Therap 2021 Aug 17;29(1):32 [FREE Full text] [doi: 10.1186/s12998-021-00389-y] [Medline: 34404446]
32. Östlind E, Ekvall Hansson E, Eek F, Stigmar K. Experiences of activity monitoring and perceptions of digital support among working individuals with hip and knee osteoarthritis - a focus group study. BMC Public Health 2022 Aug 30;22(1):1641 [FREE Full text] [doi: 10.1186/s12889-022-14065-0] [Medline: 36042425]
33. Kelly M, Fullen BM, Martin D, Bradley C, McVeigh JG. eHealth interventions to support self-management: perceptions and experiences of people with musculoskeletal disorders and physiotherapists - 'eHealth: It's TIME': a qualitative study. Physiother Theory Pract 2024 May 25;40(5):1011-1021. [doi: 10.1080/09593985.2022.2151334] [Medline: 36426843]
34. Renard M, Gaboury I, Michaud F, Tousignant M. The acceptability of two remote monitoring modalities for patients waiting for services in a physiotherapy outpatient clinic. Musculoskeletal Care 2022 Sep 10;20(3):616-624. [doi: 10.1002/msc.1622] [Medline: 35142425]
35. Poolman EY, Vorstermans L, Donker MH, Bijker L, Coppieters MW, Cuijpers P, et al. How people with persistent pain experience in-person physiotherapy blended with biopsychosocial digital health - a qualitative study on participants' experiences with Back2Action. Internet Interv 2024 Jun;36:100731 [FREE Full text] [doi: 10.1016/j.invent.2024.100731] [Medline: 38465202]
36. Pereira TA, Santos IB, Mota RF, Fukusawa L, Azevedo-Santos IF, DeSantana JM. Beliefs and expectations of patients with fibromyalgia about telerehabilitation during COVID-19 pandemic: a qualitative study. Musculoskelet Sci Pract 2023 Oct;67:102852. [doi: 10.1016/j.msksp.2023.102852] [Medline: 37639980]
37. Passalent L, Cyr A, Jurisica I, Mathur S, Inman RD, Haroon N. Motivators, barriers, and opportunity for e-Health to encourage physical activity in axial Spondyloarthritis: a qualitative descriptive study. Arthritis Care Res (Hoboken) 2022 Jan 20;74(1):50-58. [doi: 10.1002/acr.24788] [Medline: 34928533]

XSL•FO
RenderX

38. Ezzat AM, Bell E, Kemp J, O'Halloran P, Russell T, Wallis J, et al. "Much better than I thought it was going to be": telehealth delivered group-based education and exercise was perceived as acceptable among people with knee osteoarthritis. Osteoarthr Cartil Open 2022 Sep;4(3):100271 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100271] [Medline: 36474949]
39. Lamper C, Huijnen I, de Mooij M, Köke A, Verbunt J, Kroese M. An eCoach-pain for patients with chronic musculoskeletal pain in interdisciplinary primary care: a feasibility study. Int J Environ Res Public Health 2021 Nov 06;18(21):11661 [FREE Full text] [doi: 10.3390/ijerph182111661] [Medline: 34770177]
40. Barton CJ, Ezzat AM, Merolli M, Williams CM, Haines T, Mehta N, et al. "It's second best": a mixed-methods evaluation of the experiences and attitudes of people with musculoskeletal pain towards physiotherapist delivered telehealth during the COVID-19 pandemic. Musculoskelet Sci Pract 2022 Apr;58:102500 [FREE Full text] [doi: 10.1016/j.msksp.2021.102500] [Medline: 35074694]
41. Farzad M, MacDermid J, Ferreira L, Szekeres M, Cuypers S, Shafiee E. A description of the barriers, facilitators, and experiences of hand therapists in providing remote (tele) rehabilitation: an interpretive description approach. J Hand Ther 2023 Oct;36(4):805-816. [doi: 10.1016/j.jht.2023.06.004] [Medline: 37591730]
42. Dehainault M, Gaillard O, Ouattara B, Peurois M, Begue C. Physical activity advice given by French general practitioners for low back pain and the role of digital e-health applications: a qualitative study. BMC Prim Care 2024 Jan 29;25(1):44 [FREE Full text] [doi: 10.1186/s12875-024-02284-w] [Medline: 38287271]
43. Arensman R, Kloek C, Pisters M, Koppenaal T, Ostelo R, Veenhof C. Patient perspectives on using a smartphone app to support home-based exercise during physical therapy treatment: qualitative study. JMIR Hum Factors 2022 Sep 13;9(3):e35316 [FREE Full text] [doi: 10.2196/35316] [Medline: 36098993]
44. M Ezzat A, Kemp JL, J Heerey J, F Pazzinatto M, De Oliveira Silva D, Dundules K, et al. Implementation of the good life with osteoArthritis in Denmark (GLA:D) program via telehealth in Australia: a mixed-methods program evaluation. J Telemed Telecare 2023 Apr 20:1357633X231167620. [doi: 10.1177/1357633X231167620] [Medline: 37082796]
45. van Tilburg ML, Kloek CJ, Staal JB, Bossen D, Veenhof C. Feasibility of a stratified blended physiotherapy intervention for patients with non-specific low back pain: a mixed methods study. Physiother Theory Pract 2022 Feb 20;38(2):286-298. [doi: 10.1080/09593985.2020.1756015] [Medline: 32431201]
46. Kloek CJ, Bossen D, de Vries HJ, de Bakker DH, Veenhof C, Dekker J. Physiotherapists' experiences with a blended osteoarthritis intervention: a mixed methods study. Physiother Theory Pract 2020 May 28;36(5):572-579. [doi: 10.1080/09593985.2018.1489926] [Medline: 29952687]
47. Lawford BJ, Delany C, Bennell KL, Hinman RS. "I was really pleasantly surprised": firsthand experience and shifts in physical therapist perceptions of telephone-delivered exercise therapy for knee osteoarthritis-a qualitative study. Arthritis Care Res (Hoboken) 2019 Apr 28;71(4):545-557. [doi: 10.1002/acr.23618] [Medline: 29885026]
48. Hinman RS, Nelligan RK, Bennell KL, Delany C. "Sounds a bit crazy, but it was almost more personal:" a qualitative study of patient and clinician experiences of physical therapist-prescribed exercise for knee osteoarthritis via skype. Arthritis Care Res (Hoboken) 2017 Dec 02;69(12):1834-1844. [doi: 10.1002/acr.23218] [Medline: 28217864]
49. Martínez de la Cal J, Fernández-Sánchez M, Matarán-Peñarrocha GA, Hurley DA, Castro-Sánchez AM, Lara-Palomo IC. Physical therapists' opinion of e-Health treatment of chronic low back pain. Int J Environ Res Public Health 2021 Feb 16;18(4):1889 [FREE Full text] [doi: 10.3390/ijerph18041889] [Medline: 33669249]
50. Kairy D, Tousignant M, Leclerc N, Côté AM, Levasseur M, Researchers T. The patient's perspective of in-home telerehabilitation physiotherapy services following total knee arthroplasty. Int J Environ Res Public Health 2013 Aug 30;10(9):3998-4011 [FREE Full text] [doi: 10.3390/ijerph10093998] [Medline: 23999548]
51. Button K, Nicholas K, Busse M, Collins M, Spasić I. Integrating self-management support for knee injuries into routine clinical practice: TRAK intervention design and delivery. Musculoskelet Sci Pract 2018 Feb;33:53-60 [FREE Full text] [doi: 10.1016/j.msksp.2017.11.002] [Medline: 29172113]
52. Cottrell MA, Hill AJ, O'Leary SP, Raymer ME, Russell TG. Service provider perceptions of telerehabilitation as an additional service delivery option within an Australian neurosurgical and orthopaedic physiotherapy screening clinic: a qualitative study. Musculoskelet Sci Pract 2017 Dec;32:7-16. [doi: 10.1016/j.msksp.2017.07.008] [Medline: 28787636]
53. Dunphy E, Hamilton FL, Spasić I, Button K. Acceptability of a digital health intervention alongside physiotherapy to support patients following anterior cruciate ligament reconstruction. BMC Musculoskelet Disord 2017 Nov 21;18(1):471 [FREE Full text] [doi: 10.1186/s12891-017-1846-0] [Medline: 29162071]
54. Egerton T, Nelligan R, Setchell J, Atkins L, Bennell KL. General practitioners' perspectives on a proposed new model of service delivery for primary care management of knee osteoarthritis: a qualitative study. BMC Fam Pract 2017 Sep 07;18(1):85 [FREE Full text] [doi: 10.1186/s12875-017-0656-7] [Medline: 28882108]
55. Eriksson L, Lindström B, Ekenberg L. Patients' experiences of telerehabilitation at home after shoulder joint replacement. J Telemed Telecare 2011 Jan 01;17(1):25-30. [doi: 10.1258/jtt.2010.100317] [Medline: 21075802]
56. Hjelmager DM, Vinther LD, Poulsen SH, Petersen LS, Jensen MB, Riis A. Requirements for implementing online information material for patients with low back pain in general practice: an interview study. Scand J Prim Health Care 2019 Mar 31;37(1):60-68 [FREE Full text] [doi: 10.1080/02813432.2019.1569413] [Medline: 30700191]

XSL•FO
RenderX

57. Kingston GA, Judd J, Gray MA. The experience of medical and rehabilitation intervention for traumatic hand injuries in rural and remote North Queensland: a qualitative study. Disabil Rehabil 2015 May 26;37(5):423-429. [doi: 10.3109/09638288.2014.923526] [Medline: 24856789]
58. Aily JB, Barton CJ, Mattiello SM, De Oliveira Silva D, De Noronha M. Telerehabilitation for knee osteoarthritis in Brazil: a feasibility study. Int J Telerehabil 2020 Dec 08;12(2):137-148 [FREE Full text] [doi: 10.5195/ijt.2020.6323] [Medline: 33520101]
59. Geraghty AW, Roberts LC, Stanford R, Hill JC, Yoganantham D, Little P, et al. Exploring patients' experiences of internet-based self-management support for low back pain in primary care. Pain Med 2020 Sep 01;21(9):1806-1817 [FREE Full text] [doi: 10.1093/pm/pnz312] [Medline: 31841156]
60. Hasani F, Malliaras P, Haines T, Munteanu SE, White J, Ridgway J, et al. Telehealth sounds a bit challenging, but it has potential: participant and physiotherapist experiences of gym-based exercise intervention for Achilles tendinopathy monitored via telehealth. BMC Musculoskelet Disord 2021 Feb 04;22(1):138 [FREE Full text] [doi: 10.1186/s12891-020-03907-w] [Medline: 33541314]
61. Palazzo C, Klinger E, Dorner V, Kadri A, Thierry O, Boumenir Y, et al. Barriers to home-based exercise program adherence with chronic low back pain: patient expectations regarding new technologies. Ann Phys Rehabil Med 2016 Apr;59(2):107-113 [FREE Full text] [doi: 10.1016/j.rehab.2016.01.009] [Medline: 27050664]
62. Bossen D, Kloek C, Snippe HW, Dekker J, de Bakker D, Veenhof C. A blended intervention for patients with knee and hip osteoarthritis in the physical therapy practice: development and a pilot study. JMIR Res Protoc 2016 Feb 24;5(1):e32 [FREE Full text] [doi: 10.2196/resprot.5049] [Medline: 26912378]
63. Lau R, Stevenson F, Ong BN, Dziedzic K, Treweek S, Eldridge S, et al. Achieving change in primary care--causes of the evidence to practice gap: systematic reviews of reviews. Implement Sci 2016 Mar 22;11(1):40 [FREE Full text] [doi: 10.1186/s13012-016-0396-4] [Medline: 27001107]
64. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
65. van der Meer HA, Doomen A, Visscher CM, Engelbert RH, Nijhuis-van der Sanden MW, Speksnijder CM. The additional value of e-Health for patients with a temporomandibular disorder: a mixed methods study on the perspectives of orofacial physical therapists and patients. Disabil Rehabil Assist Technol 2024 Feb 12;19(2):433-445. [doi: 10.1080/17483107.2022.2094000] [Medline: 35960692]
66. Be the hero in your client's health journey. Physitrack®. URL: https://www.physitrack.com/ [accessed 2024-04-29]
67. Varsi C, Solberg Nes L, Kristjansdottir OB, Kelders SM, Stenberg U, Zangi HA, et al. Implementation strategies to enhance the implementation of eHealth programs for patients with chronic illnesses: realist systematic review. J Med Internet Res 2019 Sep 27;21(9):e14255 [FREE Full text] [doi: 10.2196/14255] [Medline: 31573934]
68. Waltz TJ, Powell BJ, Matthieu MM, Damschroder LJ, Chinman MJ, Smith JL, et al. Use of concept mapping to characterize relationships among implementation strategies and assess their feasibility and importance: results from the Expert Recommendations for Implementing Change (ERIC) study. Implement Sci 2015 Aug 07;10(1):109 [FREE Full text] [doi: 10.1186/s13012-015-0295-0] [Medline: 26249843]
69. Kloek CJ, Janssen J, Veenhof C. Development of a checklist to assist physiotherapists in determination of patients' suitability for a blended treatment. Telemed J E Health 2020 Aug 01;26(8):1051-1065. [doi: 10.1089/tmj.2019.0143] [Medline: 31804904]
70. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
71. Lewin S, Booth A, Glenton C, Munthe-Kaas H, Rashidian A, Wainwright M, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implement Sci 2018 Jan 25;13(Suppl 1):2 [FREE Full text] [doi: 10.1186/s13012-017-0688-3] [Medline: 29384079]
72. Lin I, Wiles L, Waller R, Goucke R, Nagree Y, Gibberd M, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
MMAT: Mixed Methods Appraisal Tool

XSL•FO
RenderX

Edited by A Mavragani; submitted 13.06.23; peer-reviewed by J Ross, LA Gomes; comments to author 07.10.23; revised version received 01.12.23; accepted 10.04.24; published 27.08.24.

Please cite as:
van Tilburg ML, Spin I, Pisters MF, Staal JB, Ostelo RWJG, van der Velde M, Veenhof C, Kloek CJJ
Barriers and Facilitators to the Implementation of Digital Health Services for People With Musculoskeletal Conditions in the Primary Health Care Setting: Systematic Review
J Med Internet Res 2024;26:e49868
URL: https://www.jmir.org/2024/1/e49868
doi:10.2196/49868
PMID:39190440

©Mark Leendert van Tilburg, Ivar Spin, Martijn F Pisters, J Bart Staal, Raymond WJG Ostelo, Miriam van der Velde, Cindy Veenhof, Corelien JJ Kloek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Framework Development for Reducing Attrition in Digital Dietary Interventions: Systematic Review and Thematic Synthesis

Jian Wang[1,2], MEng; Jinli Mahe[1,2], MPH; Yujia Huo[1,2], LLM; Weiyuan Huang[1,2], MPH; Xinru Liu[1,2], MPH; Yang Zhao[3,4], PhD; Junjie Huang[5], PhD; Feng Shi[6], PhD; Zhihui Li[7], PhD; Dou Jiang[8], BM; Yilong Li[8], BM; Garon Perceval[2], PhD; Lindu Zhao[8], PhD; Lin Zhang[1,2], MD, PhD

[1]The School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[2]Suzhou Industrial Park Monash Research Institute of Science and Technology, Monash University, Suzhou, China
[3]The George Institute for Global Health, University of New South Wales, Sydney, Australia
[4]The George Institute for Global Health, Beijing, China
[5]The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
[6]East China Aviation Personnel Medical Appraisal Center, Shanghai, China
[7]Tsinghua Vanke School of Public Health, Tsinghua University, Shenzhen, China
[8]School of Economics and Management, Southeast University, Nanjing, China

Corresponding Author:
Lin Zhang, MD, PhD
Suzhou Industrial Park Monash Research Institute of Science and Technology
Monash University
Room 805, 8th Floor, Building 2, Sangtian Island Science and Technology Innovation Park
No. 1, Huayun Road
Suzhou
China
Phone: 86 13426423807
Fax: 86 13426423807
Email: lin.zhang2@monash.edu

Abstract

Background: Dietary behaviors significantly influence health outcomes across populations. Unhealthy diets are linked to serious diseases and substantial economic burdens, contributing to approximately 11 million deaths and significant disability-adjusted life years annually. Digital dietary interventions offer accessible solutions to improve dietary behaviors. However, attrition, defined as participant dropout before intervention completion, is a major challenge, with rates as high as 75%-99%. High attrition compromises intervention validity and reliability and exacerbates health disparities, highlighting the need to understand and address its causes.

Objective: This study systematically reviews the literature on attrition in digital dietary interventions to identify the underlying causes, propose potential solutions, and integrate these findings with behavior theory concepts to develop a comprehensive theoretical framework. This framework aims to elucidate the behavioral mechanisms behind attrition and guide the design and implementation of more effective digital dietary interventions, ultimately reducing attrition rates and mitigating health inequalities.

Methods: We conducted a systematic review, meta-analysis, and thematic synthesis. A comprehensive search across 7 electronic databases (PubMed, MEDLINE, Embase, CENTRAL, Web of Science, CINAHL Plus, and Academic Search Complete) was performed for studies published between 2013 and 2023. Eligibility criteria included original research exploring attrition in digital dietary interventions. Data extraction focused on study characteristics, sample demographics, attrition rates, reasons for attrition, and potential solutions. We followed ENTREQ (Enhancing the Transparency in Reporting the Synthesis of Qualitative Research) and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and used RStudio (Posit) for meta-analysis and NVivo for thematic synthesis.

Results: Out of the 442 identified studies, 21 met the inclusion criteria. The meta-analysis showed mean attrition rates of 35% for control groups, 38% for intervention groups, and 40% for observational studies, with high heterogeneity (I^2=94%-99%) indicating diverse influencing factors. Thematic synthesis identified 15 interconnected themes that align with behavior theory

XSL•FO
RenderX

concepts. Based on these themes, the force-resource model was developed to explore the underlying causes of attrition and guide the design and implementation of future interventions from a behavior theory perspective.

Conclusions: High attrition rates are a significant issue in digital dietary interventions. The developed framework conceptualizes attrition through the interaction between the driving force system and the supporting resource system, providing a nuanced understanding of participant attrition, summarized as insufficient motivation and inadequate or poorly matched resources. It underscores the critical necessity for digital dietary interventions to balance motivational components with available resources dynamically. Key recommendations include user-friendly design, behavior-factor activation, literacy training, force-resource matching, social support, personalized adaptation, and dynamic follow-up. Expanding these strategies to a population level can enhance digital health equity. Further empirical validation of the framework is necessary, alongside the development of behavior theory–guided guidelines for digital dietary interventions.

Trial Registration: PROSPERO CRD42024512902; https://tinyurl.com/3rjt2df9

(J Med Internet Res 2024;26:e58735) doi:10.2196/58735

KEYWORDS

thematic synthesis; attrition rate; dropout; behavior change theory; digital dietary intervention; digital health; mHealth; eHealth; mobile apps; email

Introduction

Dietary behavior significantly influences health across various populations [1]. Poor dietary habits are linked to serious diseases and substantial economic burdens [2]. Afshin et al [3] reported that dietary risks were responsible for approximately 11 million deaths and 255 million disability-adjusted life years in 2017. Additionally, unhealthy diets contribute significantly to noncommunicable diseases, which are projected to cost more than US $30 trillion globally in the next decade [4].

Implementing universally accessible dietary interventions is a common approach to improving dietary behaviors [5], and digital technology, known for its efficacy, reach, and affordability, presents promising solutions to the associated challenges [6]. However, attrition—defined as participant dropout before completing an intervention—is prevalent in digital health or eHealth [7-9]. In some formal evaluations of app-based health interventions, attrition rates have reached as high as 75%-99% [7,9]. Many factors contribute to this high attrition rate. For example, digital dietary interventions—dietary programs implemented via digital technology—involve factors such as insufficient motivation [10], lack of interest [11], time constraints [12], inadequate guidance [13], financial constraints [14], limited health care services [15], doubts about efficacy [13], health issues [16], technical problems [17], and overwhelming demands [18]. If attrition remains high, it significantly compromises the validity and reliability of such interventions [9]. Moreover, this influence is not limited to the individual level but also potentially exacerbates health disparities across different social groups—a manifestation of digital health inequity.

From the perspective of behavior theory, interventions aim to achieve behavior change, while attrition represents an interruption in this process. Behavior formation and development involve a multitude of factors, including individual factors (such as attitudes, self-efficacy, skills, and knowledge) and environmental factors (such as health care facilities, social networks, and policies) [19]. These factors are dynamic, arising from both fluctuations in the environment and the internal instability inherent within individuals [19]. When certain essential factors are lacking or insufficient in strength, it becomes impossible to maintain behavior change, leading to attrition [19,20]. Take dietary interventions for type 2 diabetes as an example: this is a long-term process aimed at promoting behavior change, requiring individual belief [21], self-efficacy [22], emotional support from family [23], nutrition advice from dietitians [24], accessible food environments [25,26], and supportive government policies [27]. Inadequate support from these factors can lead to discontinuation and participant attrition. When such attrition arises from common factors, like limited access to professional dietary guidance, it can result in population-level disruptions in behavior and health disparities, exemplifying health inequality [28,29]. Therefore, by viewing attrition as a multifactorial behavior disruption, and using behavior theories to identify the contributing factors and analyze their specific mechanisms, we can provide a novel perspective for understanding and addressing attrition and health disparities.

Nevertheless, research focusing on attrition, particularly investigations into its causes and potential solutions through the lens of behavior theories, remains sparse. This study aims to bridge this gap. Through systematic review and thematic synthesis [30], it comprehensively explores and summarizes the reasons for attrition and potential solutions. These findings are then integrated with concepts from multiple behavior theories to develop a comprehensive theoretical framework. This framework will not only elucidate the behavioral mechanisms behind attrition but also guide future work in designing and implementing more effective digital dietary interventions, thereby reducing attrition rates at the individual level and diminishing health inequalities at the population level.

Methods

Study Design

This study uses a systematic review, meta-analysis, and thematic synthesis to investigate participant attrition in digital dietary interventions. The study protocol is available on International Prospective Register of Systematic Reviews (PROSPERO; CRD42024512902). This review follows the ENTREQ

XSL•FO
RenderX

(Enhancing the Transparency in Reporting the Synthesis of Qualitative Research) guidelines for reporting qualitative syntheses [31] and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards, including the PRISMA checklist in Multimedia Appendix 1.

Search Strategy

A thorough search across 7 electronic databases, including PubMed, MEDLINE, Embase, CENTRAL, Web of Science, CINAHL Plus with Full Text, and Academic Search Complete, was conducted using a predefined set of search terms related to attrition in digital dietary interventions, including synonyms and British spellings, and performed as full-text searches. Examples of search terms used included disengagement rate, churn rate, turnover rate, dropout rate, noncompletion rate, attrition rate, retention rate, adherence rate, compliance rate, follow-up rate, and persistence rate. This aimed to identify relevant English-language studies published from 2013 to 2023. The search period was limited to the years 2013-2023 for 3 reasons. First, 2013 marked a pivotal year in the global mobile internet landscape with the widespread adoption of 4G LTE technology, mobile devices, and applications [32], which laid the foundation for the rapid growth of digital health technologies. Second, this timeframe ensured that the data and findings were current and reflective of the latest trends and methodologies in digital dietary interventions. Third, the volume and quality of research in this field have significantly increased in recent years, providing a robust body of literature for a comprehensive review. The search strategy (see Multimedia Appendix 2) was carefully developed and executed by our experienced research team, ensuring a systematic and thorough review of the literature.

Eligibility Criteria

We focused on original research that either primarily or secondarily explored attrition rates in digital dietary interventions among human populations, encompassing both randomized controlled trials (RCTs) and observational studies. These interventions typically use technologies such as text messaging, social media, web-based platforms, smartphone apps, and personal digital assistants to improve dietary behaviors and support adherence to diet-related therapies, for example, managing chronic diseases and weight control [6,33-35]. Therefore, studies involving nondigital interventions were excluded. Additionally, nonoriginal studies, such as reviews, conference proceedings, commentaries, protocols, and collections, were excluded to concentrate on empirical data. Studies with minimal or unclear relevance to attrition rates were also omitted to ensure reliable data for thematic synthesis.

Study Selection

Zotero 6 (Corporation for Digital Scholarship), a free, open-source research management tool, was used to assist in identifying duplicates and organizing papers. The initial screening of titles and abstracts was performed to eliminate nonoriginal research or studies not pertinent to digital dietary intervention attrition rates. Full-text assessments of potentially eligible studies were then conducted to determine their inclusion based on predefined criteria. This phase excluded studies with marginal or vague relevance to attrition rates. Two independent reviewers (FS and ZL), experienced in systematic reviews and digital health, conducted the selection. Any discrepancies were resolved through discussion or input from a third reviewer (DJ).

Data Extraction

A standardized data abstraction form was developed by the research team to capture specific information from the included studies (see Multimedia Appendix 3). This form comprised 3 parts: the first part focused on study characteristics, including authors, year of publication, study date, targeted dietary behavior, duration, theories or behavioral techniques, study design, and intervention strategy. The second part covered sample characteristics, such as eligibility criteria, sample size, and demographics. The third part collected information relevant to attrition, including the number of participants, number of dropouts, attrition rates, reasons for attrition, and potential solutions. The form was pilot-tested on a sample of studies to ensure clarity and comprehensiveness.

Raw data on intervention strategies, reasons for attrition, and solutions to attrition were abstracted directly from the text of the included studies. This involved line-by-line extraction of relevant excerpts from the abstracts, results, and discussion sections of the included papers, which is the first step in thematic synthesis. In subsequent steps, these excerpts were organized and analyzed to identify descriptive and analytical themes (see “Analysis and Synthesis”).

Data extraction, carried out independently by 2 reviewers (XL and YZ) with backgrounds in public health and behavioral science, also sought consensus in the case of disagreements, facilitated by a third reviewer (JH), if required.

Study Appraisal

Evaluating the quality of included studies is essential prior to thematic synthesis. This process determines each study's contribution to the synthesis process, referred to as its value to the synthesis [36]. In this study, qualitative content (data) on attrition causes and solutions was extracted from the included studies. These were consolidated into descriptive themes and subsequently synthesized into higher-level analytical themes through thematic synthesis. Therefore, assessing the strength of evidence for the descriptive themes is crucial before this process.

Although the included studies, including RCTs and observational studies, are quantitative, the evaluation focuses on the qualitative aspects of the research and, thus, should use qualitative appraisal tools. The method of Walsh and Downe [37] provides a comprehensive, flexible, and practical framework for appraising the quality of qualitative research, and it has been widely applied. Using this framework, we developed a question checklist and an evaluation form (see Multimedia Appendix 4) to appraise the quality of each study. The checklist facilitates the efficient and clear collection of essential evaluation information, while the evaluation form organizes the responses. These responses are used to assess the trustworthiness, transferability, and usefulness of each study based on 34 evaluation criteria, with points accumulated accordingly. The total score, with a perfect score of 34 points

XSL•FO
RenderX

when all criteria are met, serves as the basis for rating the study's value to the synthesis. Due to the small sample size of scores from 21 studies and their nonnormal distribution, we used the more efficient and straightforward 3-quartile method for rating. We calculated the first and second quartiles of the scores, which are 13.0 and 15.67, respectively. Based on these quartiles, we rated the studies as low (<13.0), moderate (13.0 ≤ score < 15.67), and high (≥15.67). One reviewer (YH) conducted the initial evaluation, and a second reviewer (WH) examined the ratings, with both being experienced in designing and evaluating behavior interventions. Discrepancies were resolved through discussion.

Analysis and Synthesis

We conducted a meta-analysis to assess the attrition rates using RStudio (version 4.3.2; Posit) and a random effects model, treating attrition rates as the effect size. Studies from RCTs were categorized into control and intervention groups, while observational studies were classified as a separate group. Sensitivity analyses and publication bias assessments were performed separately for the intervention and observational groups to evaluate the robustness of results and detect potential reporting biases.

Thematic synthesis, involving the systematic extraction and synthesis of qualitative data from multiple studies, can offer deeper and comprehensive insights applicable across various contexts. We used this approach to synthesize descriptions of attrition reasons and solutions with NVivo (version 12; Lumivero). The process entailed: (1) line-by-line coding of data, (2) organizing codes into descriptive themes, and (3) refining themes into overarching analytical themes that elucidate attrition factors and interventions. Two experienced reviewers (JW and YL) in qualitative methods and evidence synthesis independently conducted all steps. After completing each step, they discussed interim results to reach a consensus before proceeding. This approach helped to identify and reduce discrepancies early, ensuring the reliability of the final results. Sensitivity analysis [38] tested the findings' robustness by excluding low-quality study data and reassessing for consistent themes. This analysis confirmed the validity and robustness of our results.

Finally, concepts from multiple behavior theories were used to construct a conceptual framework that better explains the mechanisms of attrition. This approach introduced themes of personal agency from the integrated behavior model [20], habit or impulsive behavior from the reflective-impulsive model [39], cognition and reinforcement from social cognitive theory [40], and diverse resources from the conservation of resources theory [41]. Additionally, the perceived norms theme was expanded to include both subjective or injunctive and descriptive norms, aligning with the integrated behavior model.

Results

Search and Selection Results

Database searches yielded 442 studies. After excluding 55 duplicates and 45 nonretrieved records (due to access restrictions, copyright limitations, or incomplete records), 342 underwent eligibility screening based on predetermined criteria, leading to the exclusion of 321 studies for not meeting inclusion requirements (details in Multimedia Appendix 5). Ultimately, 21 studies were included in the review (Table 1), with the selection process depicted in a PRISMA flow diagram (Figure 1).

XSL•FO
RenderX

Table 1. Summary of extracted study characteristics.

Author	Targeted dietary behavior	Attrition rate (%)	Design types
Brewer et al [11]	Increasing the intake of fruits and vegetables among the participants	__[a]	Randomized controlled trial
Browne et al [42]	Reducing the rate of eating among children with obesity	62.5	Randomized controlled trial
Cheung et al [43]	The intervention targeted dietary behaviors by including text messages that promoted Australian dietary guidelines after pregnancy. This included advice on controlling carbohydrate intake and the use of low carbohydrate vegetables and foods to satiate hunger.	52.5	Randomized controlled trial
Coa and Patrick [10]	Behaviors related to healthy eating	43.0	Observational study
Dawson et al [12]	Improving renal dietary behaviors related to potassium, phosphorus, sodium, and fluid intake, and general healthy eating and lifestyle behaviors.	10.3	Randomized controlled trial
Grutzmacher et al [15]	Nutrition and physical activity	14.3	Observational study
Hawkes et al [44]	Improving diet as one of its main objectives, alongside increasing physical activity and achieving weight loss, to reduce the risk of type 2 diabetes.	63.5	Observational study
Howarth et al [18]	Focusing on resilience, movement, eating, and sleep	—	Observational study
Jiang et al [45]	Targeting optimal nutritional intake using ordinary food and oral nutrition supplements (ONS), tailored to individual needs, preferences, and diet restrictions.	8.3	Randomized controlled trial
Kaul et al [13]	Tracking dietary intake and identifying dietary factors that might influence pain symptoms	38.9	Randomized controlled trial
Linardon et al [14]	Eating disorders, particularly behaviors related to binge eating.	66.0	Randomized controlled trial
Paxton et al [17]	Increasing the intake of fiber, fruits, vegetables, and reducing saturated and trans fats.	35.1	Observational study
Plaete et al [46]	Increasing the intake of fruit and vegetable to promote healthier dietary habits among adults	71.8	Randomized controlled trial
Rom et al [16]	The intervention targeted behaviors associated with binge-eating disorder, focusing on establishing regular eating patterns, self-monitoring of food intake, and addressing thoughts and feelings related to eating and body image.	26.3	Observational study
Schulz et al [8]	Reducing alcohol consumption among adult problem drinkers	52.7	Randomized controlled trial
Silina et al [47]	Increasing physical activity and dietary recommendations for dyslipidemia and weight loss	3.1	Randomized controlled trial
Springer et al [48]	Increasing fruit and vegetable intake	—	Randomized controlled trial
Van der Mispel et al [49]	Increasing fruit and vegetable intake	78.2	Observational study
Whitley et al [50]	Healthy eating and active living behaviors	4.6	Observational study
Young et al [51]	Mediterranean diet	70.3	Observational study
Yuhas et al [52]	Reducing the intake of SSBs[b] by adolescents	11.5	Randomized controlled trial

[a]Not applicable.

[b]SSB: sugar-sweetened beverage.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

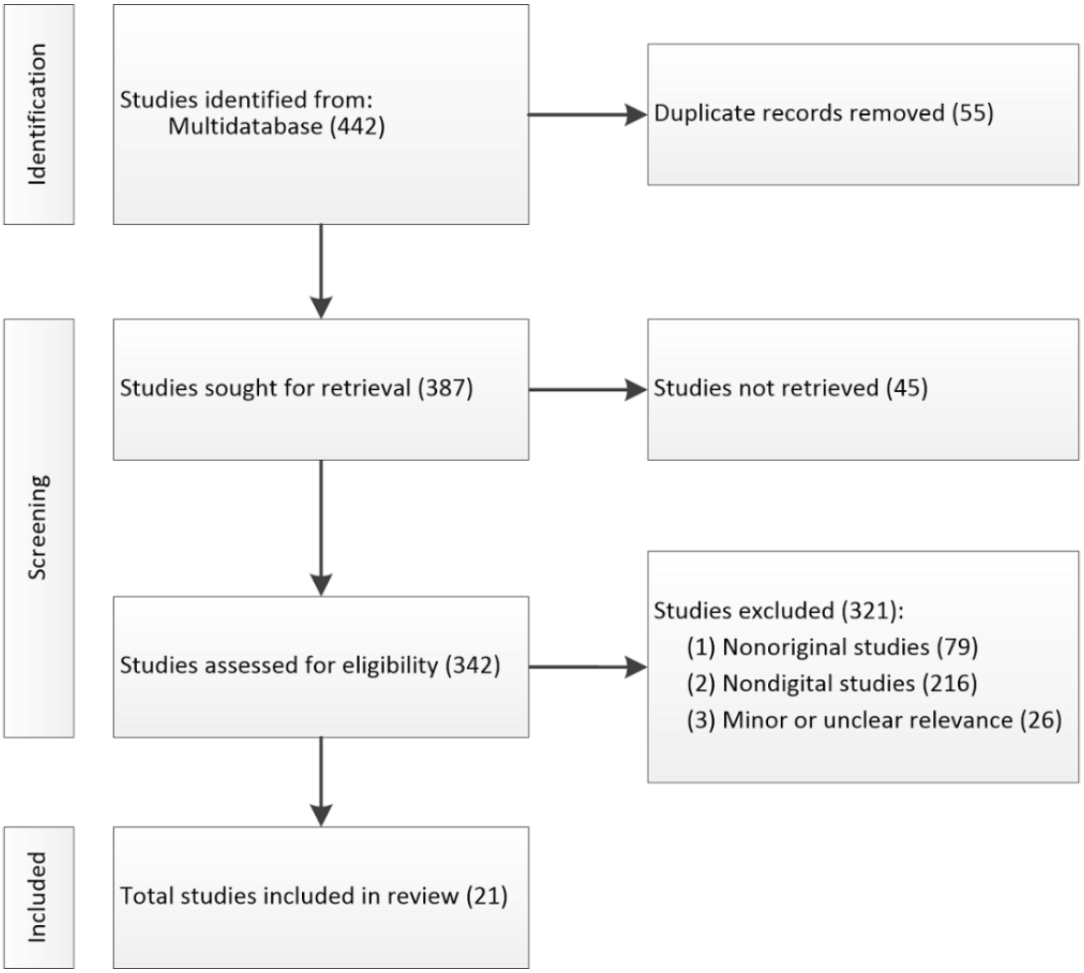

Study Characteristics and Appraisal

The characteristics of the included studies (details in Multimedia Appendix 6 [8,10-18,42-52]) span several countries: global (2 studies, subsequent numbers denote study counts), Latvia (n=1), Australia (n=4), the United States (n=9), Ireland (n=1), Germany (n=1), Belgium (n=2), and China (n=1). In total, 5 studies did not specify gender distribution, while the majority featured predominantly female participants. The interventions used varied digital technologies such as SMS, mobile apps, web-based programs, and email, targeting diverse dietary objectives including carbohydrate intake (n=1), fruit and vegetable intake (n=5), the Mediterranean diet (n=1), oral nutrition supplements intake (n=1), microelement and fluid intake (n=1), diet improvement for type 2 diabetes (n=1), binge eating (n=2), alcohol reduction (n=1), sugar-sweetened beverages intake (n=1), dietary factors related to pain symptoms (n=1), eating rate for obesity (n=1), and general healthy eating habits (n=5). Ethical considerations varied, with 2 studies not reporting on ethics, 1 bypassing review for involving voluntary workplace co-designers, and 18 obtaining clear ethical approvals. Behavior theories or techniques were used in 17 studies. Intervention durations ranged from 28 days to 4 months in 13 studies and 6-12 months in 7 studies, with 1 unspecified. Attrition rates surpassed 20% in 14 studies, peaking at 78.2%.

Among the studies, 12 were RCTs and 9 were observational. Attrition data collection methods varied, with 9 using existing digital systems, 7 detailing only calculation or timing methods, and 5 lacking clarification. Attrition causes and solutions were primarily derived from participant quotes, author interpretations, and trial evidence. Value to the synthesis faced challenges due to missing triangulation, theoretical saturation, representativeness exposition, alternative explanations, conceptual richness, and novel findings. Seven studies was rated high, 8 were rated moderate, and 6 were rated low, with the detailed appraisal process provided in Multimedia Appendix 7 [8,10-18,36,42-52].

Meta-Analysis

Meta-analyses were conducted for control and intervention groups within 12 RCTs, and observational studies were analyzed as a separate group, using a random effects model in RStudio (detailed data in Multimedia Appendix 8 [8,10-17,42-52]). Missing data necessitated the exclusion of 4 studies, resulting in the analysis of 9 RCTs and 8 observational studies. Attrition rate served as the effect size, with results in Figures 2-4. The mean attrition rates were 35% (95% CI 20-52) for control, 38% (95% CI 19-59) for intervention, and 40% (95% CI 21-62) for observational groups. Heterogeneity was assessed using I^2 and τ^2, revealing high heterogeneity with I^2 values of 94%, 97%, and 99%, and τ^2 values of 0.0604, 0.0940, and 0.0922, respectively, all with $P<.001$. The I^2 statistic indicates the percentage of total variation across studies that is due to heterogeneity rather than chance. τ^2 represents the between-study

XSL•FO
RenderX

variance, providing an estimate of the actual variation in effect sizes across the included studies. By using both I^2 and τ^2, we gain a comprehensive understanding of heterogeneity.

Sensitivity analyses, shown in Figures 5-6, used a random-effects model, revealing attrition rates of 34%-45% for intervention groups and 35%-47% for observational groups. The slight variances in 95% CIs suggest that no individual study significantly alters the overall estimate. τ^2 and τ values indicated minimal dispersion and variation, with values ranging from 0.0665 to 0.1074 and 0.2579 to 0.3277 for intervention groups, and 0.0643 to 0.1077 and 0.2536 to 0.3282 for observational groups, respectively. Similarly, I^2 values, clustering around 95%-97% for the intervention group and remaining at 99% for the observational group, reflect substantial heterogeneity yet confirm the limited impact of individual studies on the overall results. These findings indicate that the meta-analysis results are stable and not significantly affected by the inclusion or exclusion of particular studies.

Publication bias was evaluated using Begg's and Egger's tests, alongside funnel plots (Figures 7-8). The intervention group's funnel plot suggested potential bias, though Begg's (*P*=.47) and Egger's (*P*=.20) tests were not statistically significant. The observational group's funnel plot showed more symmetry, implying less bias, supported by nonsignificant Begg's (*P*=.39) and Egger's (*P*=.70) test results. These findings indicate no significant publication bias, suggesting the meta-analysis relatively realistically reflect the actual situation.

Figure 2. Forest plot of attrition rates for control group.

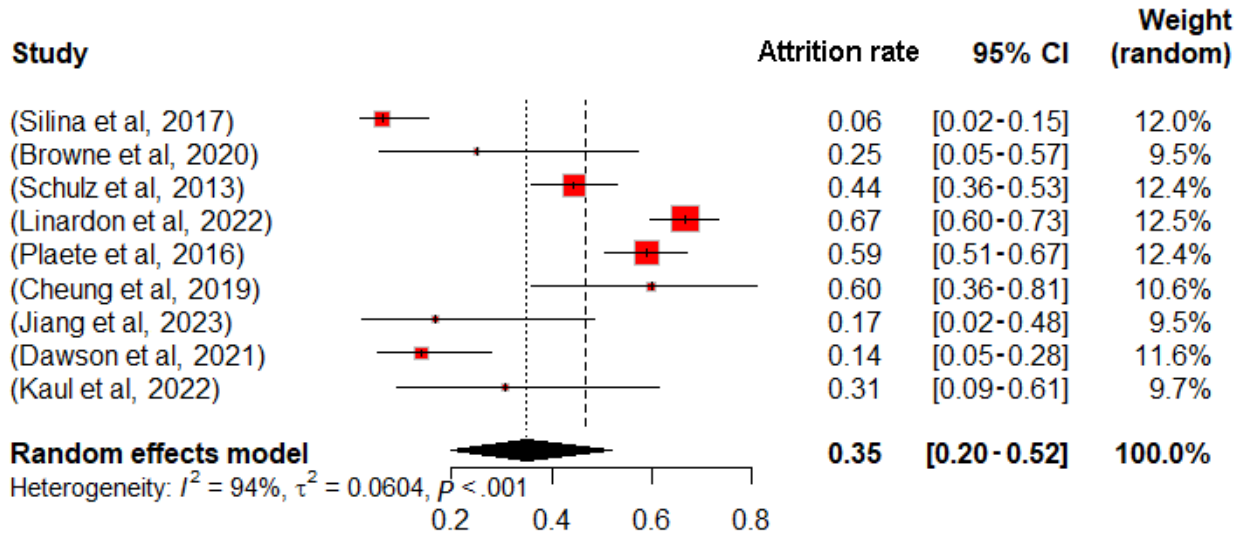

Figure 3. Forest plot of attrition rates for intervention group.

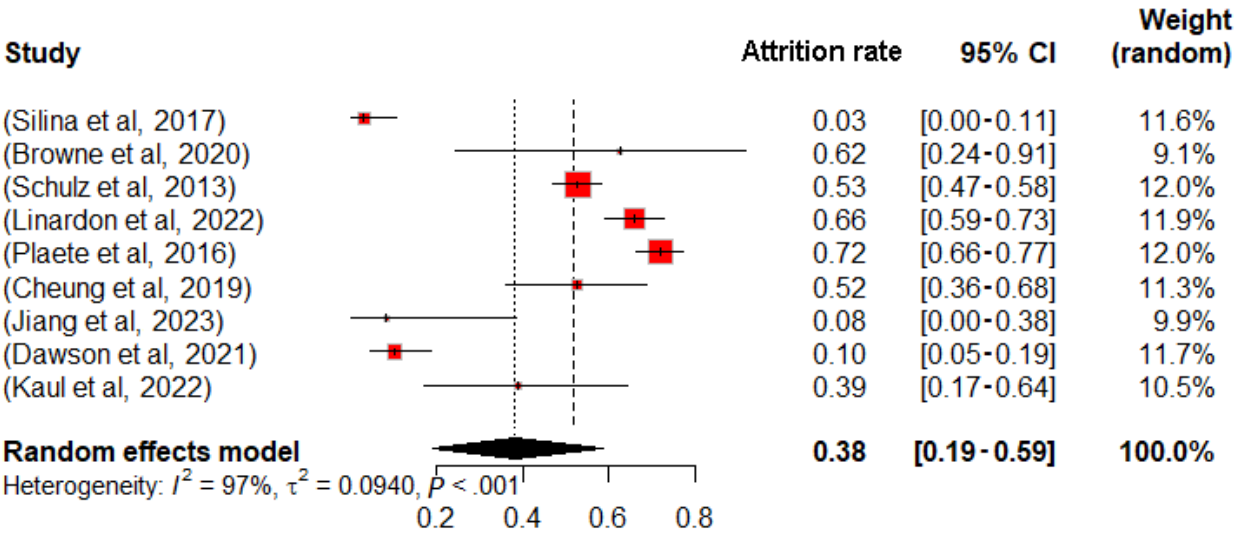

XSL•FO
RenderX

Figure 4. Forest plot of attrition rates for observation group.

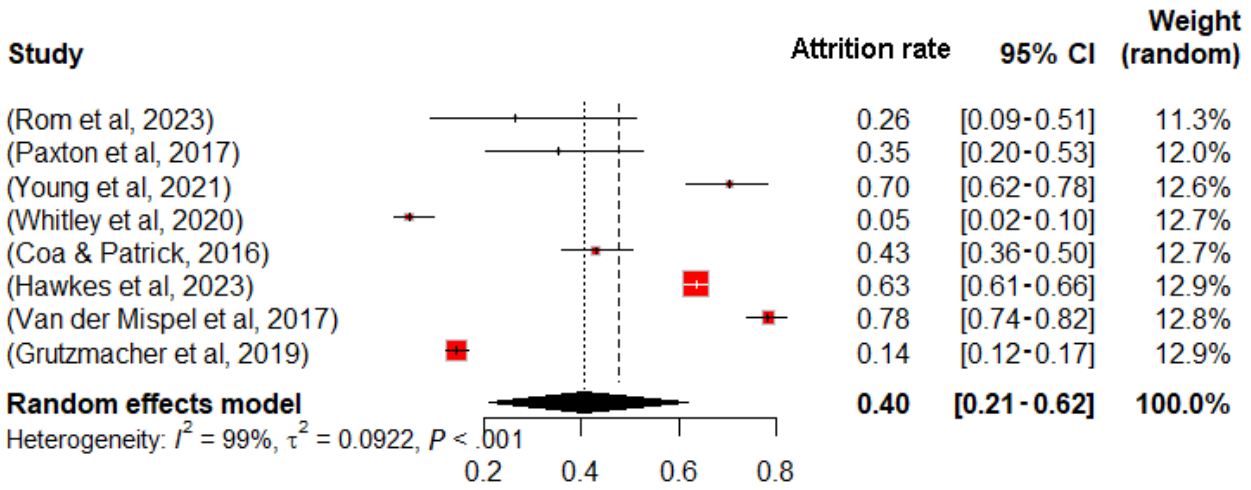

Figure 5. Sensitivity analysis of attrition rates in intervention group.

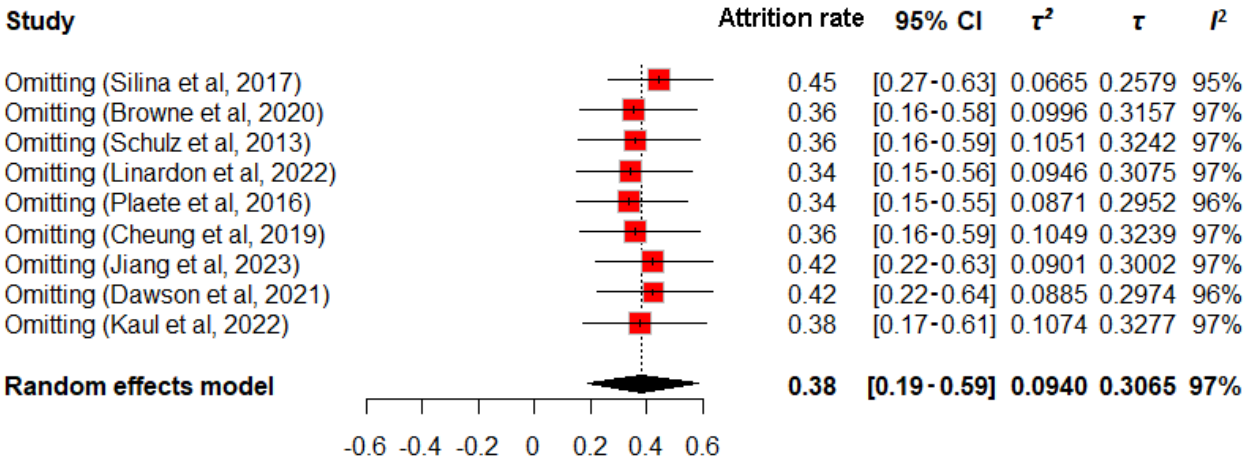

Figure 6. Sensitivity analysis of attrition rates in observational group.

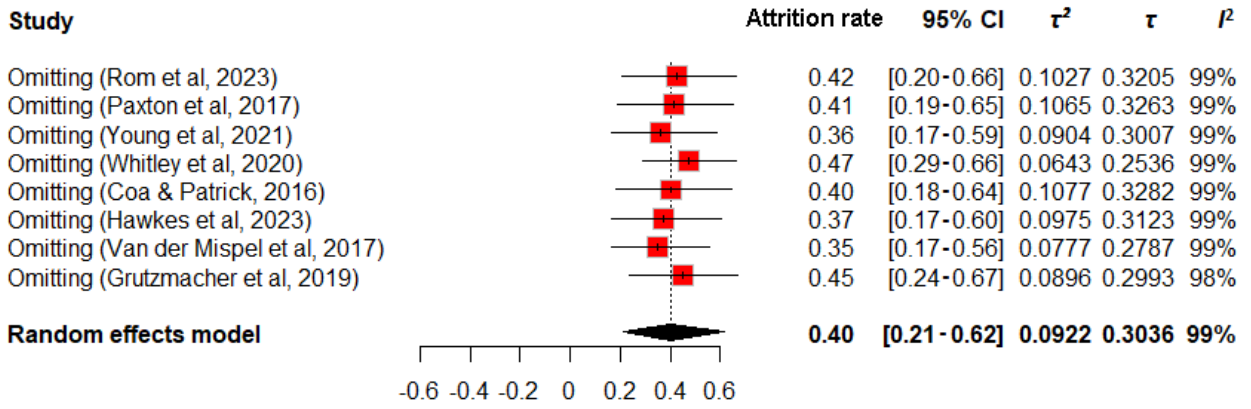

XSL•FO
RenderX

Figure 7. Funnel plot of attrition rate in intervention group.

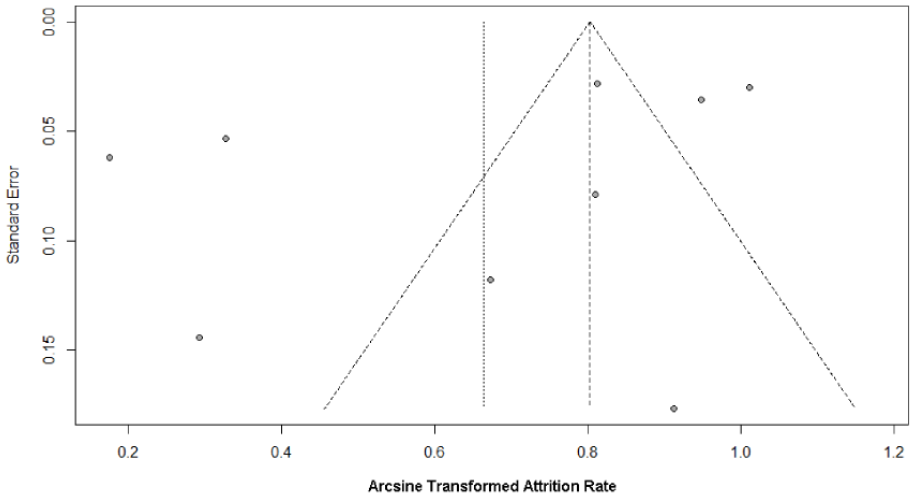

Figure 8. Funnel plot of attrition rates in observational group.

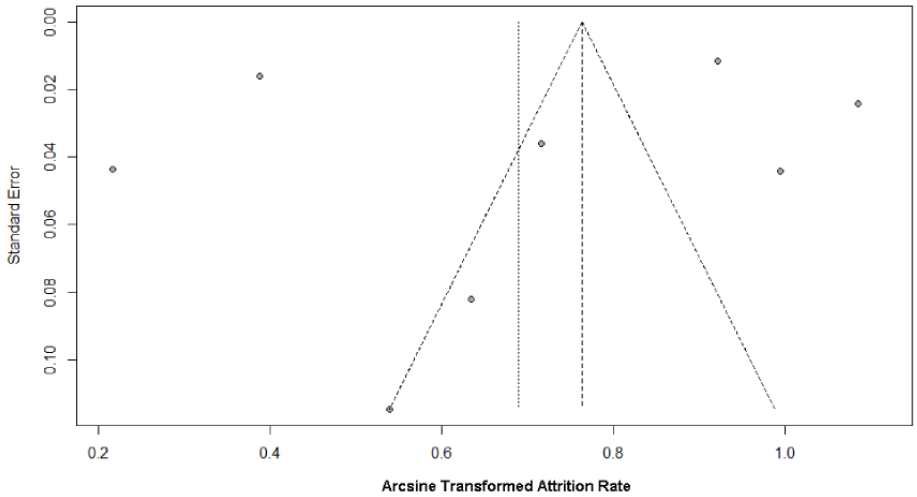

Thematic Synthesis

The thematic synthesis yielded 29 descriptive and 7 analytical themes, along with 19 subthemes, elucidating attrition reasons, as illustrated in Table 2. Furthermore, it identified 20 descriptive themes and 8 analytical themes, with 7 subthemes, addressing potential attrition solutions (details in Table 3). Evidence for each theme was extensively documented in Multimedia Appendix 9 [8,10,13-18,42-52], which also included descriptions not assigned to specific themes due to being nonspecific, unclear, or inefficient. Combining the themes of reasons and corresponding solutions from 2 tables, such as “motivation” with “boost and maintain motivation,” resulted in the formation of 13 merged themes. The finalized themes corresponded with the summary of Eysenbach [9] of factors influencing attrition rates (details in Multimedia Appendix 10).

Subsequently, we drew on concepts from multiple behavior theories, resulting in 15 integrated themes. These themes were conceptually organized to elucidate their interconnections, as shown in Figure 9. This figure illustrates an explanatory framework where participant attrition is influenced by 2 main systems: the driving force system and the supporting resource system. For clarity, this integration is termed the force-resource model. The driving force system encompasses the inherent motivation or tendencies that determine the necessity of behavior execution. In contrast, the resource system provides the essential support required for the feasibility of implementing behaviors. This dichotomy leads to 2 primary causes of participant attrition: firstly, the failure of intervention strategies to generate sufficient motivation or to counteract risky habit or impulsive behaviors; secondly, the presence of inadequate or poorly matched resources. The discussion sections will provide a detailed exploration of these themes and their interplay within this conceptual framework.

XSL•FO
RenderX

Table 2. Themes of attrition reasons from included studies.

Analytical theme and analytical subtheme	Descriptive theme
Motivation	
High motivation	• Higher autonomous motivation • Motivation waning
Attitude	
Negative experiential attitude	• Lake of interest • Strict timeline
Negative instrumental attitude	• Limited usefulness • Doubt regarding efficacy
Positive instrumental attitude	• Goal-connection feeling • Clearer expectation
Subjective or injunctive norm	
Lack of subjective or injunctive norm	• No direct contact
With subjective or injunctive norm	• With patient-provider relationship
Cue	
Distraction cue	• Triggered by stop messages
Reinforcement	
Delayed reinforcement	• Delayed feedback
Positive or immediate reinforcement	• Positive feedback • No reimbursement • Immediate feedback
Resources	
With cognitive load	• Overwhelmed tasks
Low usability	• Technical or usability issues
Limited knowledge or skills	• Limited guidance • Limited technical literacy
Service resource	• Lack of social support • More health services • Few health services
Financial resource	• Financial barriers • Higher income
Time resource	• Time constraint
Personal state	• Health or life issue
Individual differences	
Cultural factor	• Cultural barriers
Education level	• Low education level • High education level

XSL•FO
RenderX

Table 3. Themes of attrition solutions from included studies.

Analytical theme and analytical subtheme	Descriptive theme
Boost and maintain motivation	• Enhance autonomous motivation • Enhance self-affirmation
Improve attitude	
Improve experiential attitude	• Make interventions fun
Improve instrumental attitude	• Educate on intervention
Offer subjective or injunctive norm	• Using health practitioners' referrals
Provide immediate reinforcement	• Set progress markers • Provide immediate information
Provide matching resources	
Decrease cognitive load	• Make interventions easy • Improve usability • Provide guidance
Provide financial resource	• Address financial barriers
Improve personal state	• Improve emotional state
Provide social support	• Using peer encouragement
Personalization strategy	• Use targeted strategies
Personalization strategy	
Based on feedback	• Refine text-messages • Solicit user-feedback
Attend to individual difference	• Screen participants • Understand the impact of participants' characteristics • Address cultural barrier
Dynamic intervention	• Provide tailored follow-up

XSL•FO
RenderX

Figure 9. Force-resource model.

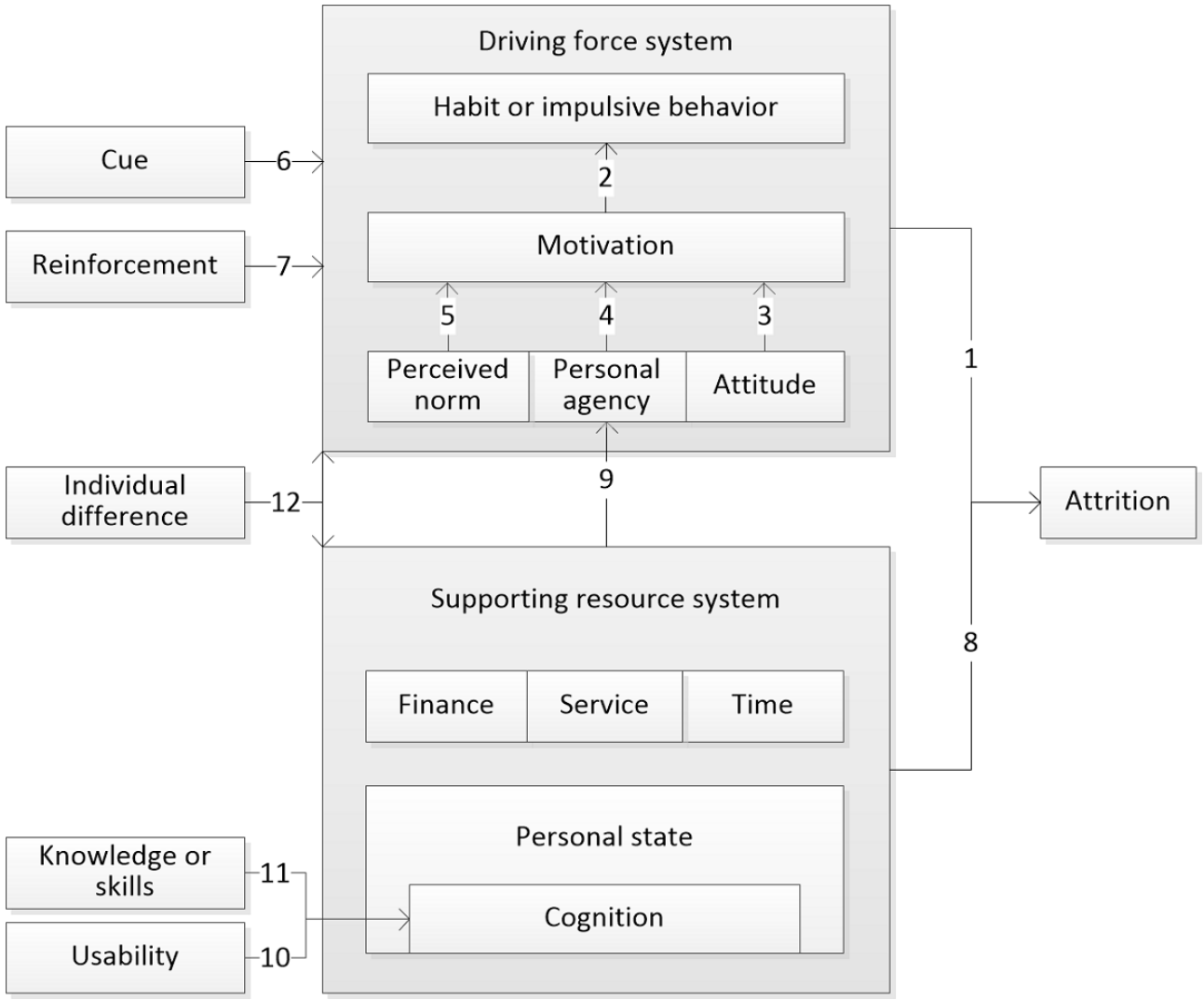

Discussion

Principal Findings

We identified 15 interconnected theoretical themes and integrated behavior theory concepts to construct the force-resource model. As shown in Figure 9, the model comprises 2 subsystems that interact to influence behavior and contribute to attrition. The first subsystem, the driving force system, includes themes of motivation, perceived norms, personal agency, attitude, and habit or impulsive behavior, which collectively guide behavioral directions and trends. Participants are driven by this force system to engage in dietary interventions to improve health. Their behaviors are influenced by attitudes toward the diet-health connection and the efficacy of the intervention, as well as perceived norms and personal agency. This behavioral tendency underpins their initial participation and intention to persist.

The second subsystem is the supporting resources system. The core concept of this system is resource, defined as entities either intrinsically valued or instrumental in achieving valued ends. These resources include physical and psychological states, financial support, time availability, and accessible health services [41]. Adequate resources alongside the force system likely facilitate behavior change, whereas insufficient resources obstruct it. Interestingly, an excess of resources can also contribute to attrition, as it may diminish the perceived value of the intervention. This is attributable to the diverse and competitive nature of motivation; when resources are abundant, previously unattainable desires become attainable, leading participants to pursue more appealing activities, necessitating greater cognitive resources to overcome them. For instance, in food-rich environments, pursuing weight control goals as a self-regulation process demands more cognitive resources than pursuing eating enjoyment [53]. Therefore, ensuring participants have access to appropriate and ample supporting resources is critical to prevent attrition.

In addition to the 2 subsystems, the model includes other key components such as cues, which refer to specific environmental stimuli that trigger actions [54]; reinforcement, which increases the likelihood of a behavior by delivering a rewarding stimulus after the behavior [55]; participants' knowledge and skills; the usability of digital interventions; and individual differences, primarily referring to stable personal attributes [56], including demographic backgrounds, personality traits, and cultural values.

XSL•FO
RenderX

For a more detailed elucidation of the components and mechanisms of the force-resource model, refer to Multimedia Appendix 11 [8,13,15,17,18,20,39,41-43,46,52-78].

Implications for Digital Dietary Interventions

Overview

The force-resource model provides a resource-matching perspective for the design of digital dietary interventions, thus, forming multifaceted behavior intervention strategies tailored to the individual, based on the characteristics and functions of the components, and their interrelationships within the model.

User-Friendly Design

The design of digital interventions should ensure that processes are easy to understand and use, not only by avoiding excessive tasks, such as extensive questionnaires and record-keeping, but also by fully leveraging various technological tools to help simplify these tedious tasks and reduce cognitive load. For example, digital dietary interventions can use image processing and pattern recognition to streamline dietary recording [79], and use artificial intelligence to assess daily dietary quality automatically [80].

Behavior-Factor Activation

Based on behavior theory, behavior change can be achieved by activating corresponding components within the force system. Taking motivation as an example, improving diet is a gradual process that necessitates persistent adherence to achieve long-term health benefits, which presents significant challenges for behavior interventions based on health motivations. We advocate for the integration of health interventions with other activities such as gaming and social interaction that provide immediate feedback, thereby potentially enhancing sustained involvement and reducing attrition. Additionally, applying immediate reinforcement can increase the likelihood or probability of the behavior, while avoiding disruptive cues and highlighting beneficial ones can foster sustained engagement.

Literacy Training

The aim of providing targeted training programs is to enhance 2 types of literacy in individuals. The first is digital literacy, enabling them to use digital health resources more effectively with minimal cognitive load. The second is health literacy, which facilitates the activation of behavior factors such as attitudes and self-efficacy, thereby increasing receptiveness to health interventions. Additionally, there is a complementary relationship between literacy training and user-friendly design: the former helps people better operate digital tools with varying levels of user-friendliness, while interventions with good user-friendly design can reduce the demands on literacy training. Ideally, digital tools that effectively integrate cognitive psychology, behavioral science, and human-computer interaction are designed to be intuitive, engaging, and easy to use without requiring prior training.

Force-Resource Matching

Regarding the resource system, it is vital to ensure the provision or conservation of resources that are compatible with the force system. For instance, the availability of professional dietary counseling significantly influences the success rate of interventions for diabetic patients. Conversely, it is pragmatic to align motivation levels with available resources, recommending the setting of achievable dietary goals accordingly. For example, individuals with limited financial resources and access to medical advice should target a balanced diet as an intervention rather than pursuing antiaging diets or precision nutrition.

Social Support

Rooted in social networks [81], social support offers emotional, instrumental, and informational assistance [82]. Guidance and education fall under the category of informational support, which can alleviate cognitive processing demands while positively influencing attitudes, self-efficacy, and motivation. Emotional support has a wide-ranging impact, as it can enhance subjective well-being and cognitive functioning [83,84], as well as influencing attitudes [57,58], perceived norm [85], self-efficacy [86], and motivation [39,57,83,87,88]. Instrumental support facilitates access to financial and service-related resources, while fostering community and motivation, thus, reducing attrition risk [18,50].

Personalized Adaptation

There are significant variations in the characteristics and attributes of resource and force systems among individuals, arising not only from their personal state and circumstances but also from individual differences. Digital interventions should, therefore, tailor strategies to accommodate these variations, thereby enhancing resonance [89]. For instance, adapting dietary messages to reflect personal emotions and cultural eating habits can maintain engagement across diverse demographics [43]. Personalized adaptation is integral to multiple stages of digital interventions, including design, implementation, evaluation, and optimization, and artificial intelligence holds great potential in addressing this challenge.

Dynamic Follow-Up

The resource and force systems are dynamic, arising from fluctuations in both the environment and individuals' internal states [57,90]. As mentioned earlier, environmental cues can easily trigger distractions and competing motivations, contributing to participant attrition. This implies that static interventions struggle to accommodate such variability. To address this issue, interventions should incorporate real-time adaptability, providing timely and tailored actions [91]. For instance, automated prompts encouraging re-engagement could redirect disengaging users back into the program if metrics indicate disengagement [45]. Additionally, for long-term monitoring of large populations, the rapid data processing capabilities of artificial intelligence can be fully used.

Implications for Digital Health Equity

Participant attrition arises from mismatches between individual resource and force systems. When this phenomenon expands to population level, it essentially creates a form of digital health inequity due to disparities in access to digital health resources. Digital health equity strives for equitable access to and use of resources such as digital health technologies, training programs, digital health care systems, and community support structures,

XSL•FO
RenderX

all designed to improve health outcomes universally [92,93]. In promoting digital health equity, reducing attrition rates is a key strategy [92-96]. This encourages us to broaden the goal of reducing attrition to encompass a larger population and to design solutions at more comprehensive levels based on the multilevel determinants in digital health [92,93]. We have proposed several strategic recommendations based on the force-resource model, as detailed in Multimedia Appendix 12 [92-96].

Complementary Findings

Through a meta-analysis of attrition rates in digital dietary interventions over the past decade, we found that the average attrition rates ranged from 35% to 40%, representing a significant barrier to the efficacy and generalizability of such interventions, irrespective of the study design or the presence of an active intervention component. Notably, the intervention group exhibited a marginally higher attrition rate compared to the control group, with overlapping CIs, suggesting a lack of effectiveness of the investigated interventions in mitigating participant attrition in the included studies. This observation highlights the need for more potent and tailored strategies to promote sustained engagement.

The high degree of heterogeneities underscores the inherent complexity and diversity of factors influencing attrition rates in digital dietary interventions, including variations in study designs, intervention components, participant characteristics, and implementation contexts. This indicates significant room for improvement in standardization of digital dietary interventions. Taking participant characteristics as an example, some may join interventions out of curiosity without a genuine understanding or interest, making them unsuitable for the target group and likely to drop out quickly. These varied and mixed participant characteristics. The “run-in and withdrawal” strategy can mitigate this challenge by including an initial phase where all participants start the intervention [9]. This run-in phase helps identify those less likely to adhere. Participants not fully engaged or committed can then be excluded, leaving a more homogeneous and dedicated group for the remainder of the study. By ensuring a more homogeneous and committed participant group, and standardizing selection criteria, this strategy effectively reduces heterogeneity, thereby leading to more consistent and reliable assessments of intervention effects.

Strengths

Most research on digital interventions prioritizes efficacy, often treating attrition rates as a mere data point rather than a subject of in-depth analysis. This review stands as the first to scrutinize attrition rates within digital systems through the lens of behavior theories, introducing a force-resource model to explore underlying causes and identify possible solutions. It also expands these findings to enhance digital health equity. The insights gained provide a foundational understanding and innovative strategies for developing more effective digital dietary interventions and promoting digital health equity.

Limitations and Future Direction

First, despite performing a thorough search across 7 electronic databases, we acknowledge the limitation in capturing the entire body of literature related to attrition in digital dietary interventions due to the challenge of matching search terms precisely with indexed keywords. To mitigate this, we expanded the search terms and used a comprehensive full-text search strategy. However, there remains the potential for missing studies, as many do not specifically emphasize attrition, leading to its underrepresentation in indexed keywords. This limitation could affect the completeness of our review and underscores the need for more refined search methodologies in future research.

Second, the observed heterogeneities in meta-analysis outcomes highlight a critical issue stemming from significant study variances, emphasizing the need for enhanced standardization and harmonization of protocols and components in digital dietary interventions. Establishing behavior theory–guided guidelines and best practices for designing, implementing, and evaluating these interventions could lead to more consistent and replicable results, thereby improving their generalizability and impact.

Third, this study developed a theoretical framework to clarify attrition mechanisms and guide digital dietary interventions. However, due to the aforementioned limitations and the continuous introduction of new digital tools and strategies in this field, the framework's ability to encompass and explain currently undiscovered and future emerging scenarios remains uncertain. This necessitates further empirical validation and exploration in future research.

Last, reducing attrition rates is a key strategy for promoting digital health equity, encouraging us to expand the findings related to attrition reduction to the population level of digital health equity. However, some unique factors at the population level, such as cultural diversity, social structures, and communication patterns, with significant influence on digital health equity, have not been fully explored in this study. This highlights potential directions for future research.

Conclusions

High attrition rates compromise the effectiveness and sustainability of digital dietary interventions. This review has pioneered the examination of participant attrition in such interventions through the lens of behavior theories, introducing the force-resource model. This framework conceptualizes attrition via the interaction between the driving force system and the supporting resource system, offering a nuanced understanding of participant attrition, summarized as insufficient motivation and inadequate or poorly matched resources. It highlights the critical necessity for digital dietary interventions to dynamically balance motivational components with available resources, ensuring interventions are both compelling and practically feasible for participants. Key recommendations include user-friendly design, behavior-factor activation, literacy training, force-resource matching, social support, personalized adaptation, and dynamic follow-up. These strategies are crucial for promoting sustained engagement and reducing attrition rates. Additionally, by expanding these strategies to a population level, this study contributes to the broader goal of promoting digital health equity.

XSL•FO
RenderX

The comprehensive and reliable validation of this framework requires further empirical investigation, especially with the continuous emergence of new digital tools and strategies in this field. Concurrently, developing behavior theory–guided guidelines for the design, implementation, and evaluation of digital dietary interventions is imperative to enhance their standardization and effectiveness. Furthermore, research should explore the impact of cultural diversity, social structures, and communication patterns on digital health interventions to ensure they are inclusive and effectively address the needs of diverse populations. This comprehensive approach will help overcome barriers to successful intervention outcomes and improve overall public health equity.

Authors' Contributions

JW drafted the manuscript. L Zhang and JW were responsible for the concept and design of the study. FS, ZL, and DJ were involved in the search and study selection. XL, YZ, and JH extracted all data. YH and WH conducted the study appraisal. JW and YL performed the analysis and synthesis. Meta-analysis was conducted by JM and JW. GP conducted a comprehensive English language review and polishing. L Zhao coordinated and managed the organization of the study. All authors participated in developing the review's methodology, contributed to multiple manuscript drafts, and gave their approval for the final version. Authors L Zhao (ldzhao@seu.edu.cn) and L Zhang (lin.zhang2@monash.edu) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) 2020 checklist.
[DOC File , 94 KB - jmir_v26i1e58735_app1.doc]

Multimedia Appendix 2
Search strategies.
[DOC File , 55 KB - jmir_v26i1e58735_app2.doc]

Multimedia Appendix 3
Standardized Data Abstraction Form.
[DOCX File , 18 KB - jmir_v26i1e58735_app3.docx]

Multimedia Appendix 4
Question Checklist and Evaluation Form for Study Quality Appraisal.
[DOC File , 77 KB - jmir_v26i1e58735_app4.doc]

Multimedia Appendix 5
Papers Excluded From Analysis.
[DOC File , 182 KB - jmir_v26i1e58735_app5.doc]

Multimedia Appendix 6
Study Characteristics.
[DOC File , 95 KB - jmir_v26i1e58735_app6.doc]

Multimedia Appendix 7
Study Quality Appraisal.
[DOC File , 306 KB - jmir_v26i1e58735_app7.doc]

Multimedia Appendix 8
Data for Meta-analysis.
[DOC File , 73 KB - jmir_v26i1e58735_app8.doc]

Multimedia Appendix 9
Thematic Synthesis.
[DOC File , 122 KB - jmir_v26i1e58735_app9.doc]

XSL•FO
RenderX

Multimedia Appendix 10
Analytical Themes and Eysenbach's Attrition Factors.
[DOC File , 46 KB - jmir_v26i1e58735_app10.doc]

Multimedia Appendix 11
Force-Resource Model.
[DOC File , 43 KB - jmir_v26i1e58735_app11.doc]

Multimedia Appendix 12
Implications for Digital Health Equity.
[DOC File , 36 KB - jmir_v26i1e58735_app12.doc]

References

1. Brownell KD, Cohen LR. Adherence to dietary regimens. 2: Components of effective interventions. Behav Med 1995;20(4):155-164. [doi: 10.1080/08964289.1995.9933732] [Medline: 7620227]
2. Peeters A, Blake MRC. Socioeconomic inequalities in diet quality: from identifying the problem to implementing solutions. Curr Nutr Rep 2016;5(3):150-159. [doi: 10.1007/s13668-016-0167-5]
3. GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2019;393(10184):1958-1972 [FREE Full text] [doi: 10.1016/S0140-6736(19)30041-8] [Medline: 30954305]
4. Jacka FN, Sacks G, Berk M, Allender S. Food policies for physical and mental health. BMC Psychiatry 2014;14:132 [FREE Full text] [doi: 10.1186/1471-244X-14-132] [Medline: 24884515]
5. Sjöström M, Stockley L. Toward public health nutrition strategies in the European union to implement food based dietary guidelines and to enhance healthier lifestyles. Public Health Nutr 2001;4(2A):307-324. [doi: 10.1017/s1368980001001562] [Medline: 11688436]
6. Chen Y, Perez-Cueto FJA, Giboreau A, Mavridis I, Hartwell H. The promotion of eating behaviour change through digital interventions. Int J Environ Res Public Health 2020;17(20):7488 [FREE Full text] [doi: 10.3390/ijerph17207488] [Medline: 33076239]
7. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
8. Schulz DN, Candel MJ, Kremers SP, Reinwand DA, Jander A, de Vries H. Effects of a web-based tailored intervention to reduce alcohol consumption in adults: randomized controlled trial. J Med Internet Res 2013;15(9):e206 [FREE Full text] [doi: 10.2196/jmir.2568] [Medline: 24045005]
9. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
10. Coa K, Patrick H. Baseline motivation type as a predictor of dropout in a healthy eating text messaging program. JMIR Mhealth Uhealth 2016;4(3):e114 [FREE Full text] [doi: 10.2196/mhealth.5992] [Medline: 27688034]
11. Brewer D, Dickens E, Humphrey A, Stephenson T. Increased fruit and vegetable intake among older adults participating in Kentucky's congregate meal site program. Educ Gerontol 2016;42(11):771-784 [FREE Full text] [doi: 10.1080/03601277.2016.1231511] [Medline: 28642630]
12. Dawson J, Campbell KL, Craig JC, Tong A, Teixeira-Pinto A, Brown MA, et al. A text messaging intervention for dietary behaviors for people receiving maintenance hemodialysis: a feasibility study of KIDNEYTEXT. Am J Kidney Dis 2021;78(1):85-95.e1. [doi: 10.1053/j.ajkd.2020.11.015] [Medline: 33421456]
13. Kaul U, Scher C, Henderson CR, Kim P, Dyhrberg M, Rudin V, et al. A mobile health + health coaching application for the management of chronic non-cancer pain in older adults: results from a pilot randomized controlled study. Front Pain Res (Lausanne) 2022;3:921428 [FREE Full text] [doi: 10.3389/fpain.2022.921428] [Medline: 35959237]
14. Linardon J, Shatte A, Rosato J, Fuller-Tyszkiewicz M. Efficacy of a transdiagnostic cognitive-behavioral intervention for eating disorder psychopathology delivered through a smartphone app: a randomized controlled trial. Psychol Med 2022;52(9):1679-1690. [doi: 10.1017/S0033291720003426] [Medline: 32972467]
15. Grutzmacher SK, Munger AL, Speirs KE, Vafai Y, Hilberg E, Braunscheidel Duru E, et al. Predicting attrition in a text-based nutrition education program: survival analysis of Text2BHealthy. JMIR Mhealth Uhealth 2019;7(1):e9967 [FREE Full text] [doi: 10.2196/mhealth.9967] [Medline: 30664489]
16. Rom S, Miskovic-Wheatley J, Barakat S, Aouad P, Kim M, Fuller-Tyszkiewicz M, et al. The acceptability, feasibility, and preliminary efficacy of a supported online self-help treatment program for binge-eating disorder. Front Psychiatry 2023;14:1229261 [FREE Full text] [doi: 10.3389/fpsyt.2023.1229261] [Medline: 37860164]

XSL•FO
RenderX

17. Paxton RJ, Hajek RA, Newcomb P, Dobhal M, Borra S, Taylor WC, et al. A lifestyle intervention via Email in minority breast cancer survivors: randomized parallel-group feasibility study. JMIR Cancer 2017;3(2):e13 [FREE Full text] [doi: 10.2196/cancer.7495] [Medline: 28935620]
18. Howarth A, Quesada JA, Donnelly T, Mills PR. The development of 'Make One Small Change': an e-health intervention for the workplace developed using the person-based approach. Digit Health 2019;5:2055207619852856 [FREE Full text] [doi: 10.1177/2055207619852856] [Medline: 31210960]
19. Wang J, Shao J, Zhang S, Wang L, Zhang L. Comprehending health behavior change and maintenance: a systematic review and meta-synthesis of behavior theories. Am J Health Educ 2024:1-15. [doi: 10.1080/19325037.2024.2338465]
20. Glanz K, Rimer BK, Viswanath K. Health Behavior and Health Education: Theory Research and Practice. Hoboken, New Jersey, US: John Wiley & Sons; 2008.
21. Sharifirad G, Entezari MH, Kamran A, Azadbakht L. The effectiveness of nutritional education on the knowledge of diabetic patients using the health belief model. J Res Med Sci 2009;14(1):1-6 [FREE Full text] [Medline: 21772854]
22. Strychar I, Elisha B, Schmitz N. Type 2 diabetes self-management: role of diet self-efficacy. Can J Diabetes 2012;36(6):337-344. [doi: 10.1016/j.jcjd.2012.10.005]
23. Albanese AM, Huffman JC, Celano CM, Malloy LM, Wexler DJ, Freedman ME, et al. The role of spousal support for dietary adherence among type 2 diabetes patients: a narrative review. Soc Work Health Care 2019;58(3):304-323 [FREE Full text] [doi: 10.1080/00981389.2018.1563846] [Medline: 30596355]
24. Siopis G, Colagiuri S, Allman-Farinelli M. Effectiveness of dietetic intervention for people with type 2 diabetes: a meta-analysis. Clin Nutr 2021;40(5):3114-3122. [doi: 10.1016/j.clnu.2020.12.009] [Medline: 33413914]
25. Sarkar C, Webster C, Gallacher J. Are exposures to ready-to-eat food environments associated with type 2 diabetes? A cross-sectional study of 347 551 UK biobank adult participants. Lancet Planet Health 2018;2(10):e438-e450 [FREE Full text] [doi: 10.1016/S2542-5196(18)30208-0] [Medline: 30318101]
26. Cradock KA, ÓLaighin G, Finucane FM, McKay R, Quinlan LR, Martin Ginis KA, et al. Diet behavior change techniques in type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 2017;40(12):1800-1810. [doi: 10.2337/dc17-0462] [Medline: 29162585]
27. Ley SH, Hamdy O, Mohan V, Hu FB. Prevention and management of type 2 diabetes: dietary components and nutritional strategies. Lancet 2014;383(9933):1999-2007 [FREE Full text] [doi: 10.1016/S0140-6736(14)60613-9] [Medline: 24910231]
28. McCartney G, Popham F, McMaster R, Cumbers A. Defining health and health inequalities. Public Health 2019;172:22-30 [FREE Full text] [doi: 10.1016/j.puhe.2019.03.023] [Medline: 31154234]
29. Gakidou EE, Murray CJ, Frenk J. Defining and measuring health inequality: an approach based on the distribution of health expectancy. Bull World Health Organ 2000;78(1):42-54 [FREE Full text] [Medline: 10686732]
30. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
31. Tong A, Flemming K, McInnes E, Oliver S, Craig JC. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012;12:181 [FREE Full text] [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
32. Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014-2019 [Internet]. San Jose, CA: Cisco Systems; 2015. URL: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-520862.html [accessed 2024-07-03]
33. Allman-Farinelli M, Gemming L. Technology interventions to manage food intake: where are we now? Curr Diab Rep 2017;17(11):103. [doi: 10.1007/s11892-017-0937-5] [Medline: 28942509]
34. Fakih El Khoury C, Karavetian M, Halfens RJG, Crutzen R, Khoja L, Schols JMGA. The effects of dietary mobile apps on nutritional outcomes in adults with chronic diseases: a systematic review and meta-analysis. J Acad Nutr Diet 2019;119(4):626-651. [doi: 10.1016/j.jand.2018.11.010] [Medline: 30686742]
35. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
36. Greaves C, Poltawski L, Garside R, Briscoe S. Understanding the challenge of weight loss maintenance: a systematic review and synthesis of qualitative research on weight loss maintenance. Health Psychol Rev 2017;11(2):145-163 [FREE Full text] [doi: 10.1080/17437199.2017.1299583] [Medline: 28281891]
37. Walsh D, Downe S. Appraising the quality of qualitative research. Midwifery 2006;22(2):108-119. [doi: 10.1016/j.midw.2005.05.004] [Medline: 16243416]
38. Carroll C, Booth A, Lloyd-Jones M. Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. Qual Health Res 2012;22(10):1425-1434. [doi: 10.1177/1049732312452937] [Medline: 22865107]
39. Strack F, Deutsch R. Reflective and impulsive determinants of social behavior. Pers Soc Psychol Rev 2004;8(3):220-247 [FREE Full text] [doi: 10.1207/s15327957pspr0803_1] [Medline: 15454347]
40. Bandura A. Annals of child development. Six theories of child development. In: Vasta R, editor. Social Cognitive Theory. Stamford, Connecticut, USA: JAI Press; 1989.

XSL•FO
RenderX

41. Hobfoll SE. Conservation of resources. A new attempt at conceptualizing stress. Am Psychol 1989;44(3):513-524. [doi: 10.1037//0003-066x.44.3.513] [Medline: 2648906]
42. Browne S, Kechadi M, O'Donnell S, Dow M, Tully L, Doyle G, et al. Mobile health apps in pediatric obesity treatment: process outcomes from a feasibility study of a multicomponent intervention. JMIR Mhealth Uhealth 2020;8(7):e16925 [FREE Full text] [doi: 10.2196/16925] [Medline: 32673267]
43. Cheung NW, Blumenthal C, Smith BJ, Hogan R, Thiagalingam A, Redfern J, et al. A pilot randomised controlled trial of a text messaging intervention with customisation using linked data from wireless wearable activity monitors to improve risk factors following gestational diabetes. Nutrients 2019;11(3):590 [FREE Full text] [doi: 10.3390/nu11030590] [Medline: 30862052]
44. Hawkes RE, Miles LM, Ainsworth B, Ross J, Meacock R, French DP. Engagement with a nationally-implemented digital behaviour change intervention: usage patterns over the 9-month duration of the national health service digital diabetes prevention programme. Internet Interv 2023;33:100647 [FREE Full text] [doi: 10.1016/j.invent.2023.100647] [Medline: 37502122]
45. Jiang X, Chen J, Yuan X, Lin Y, Chen Y, Li S, et al. Feasibility of an individualized mHealth nutrition (iNutrition) intervention for post-discharged gastric cancer patients following gastrectomy: a randomized controlled pilot trial. Nutrients 2023;15(8):1883 [FREE Full text] [doi: 10.3390/nu15081883] [Medline: 37111102]
46. Plaete J, Crombez G, Van der Mispel C, Verloigne M, Van Stappen V, De Bourdeaudhuij I. Effect of the web-based intervention MyPlan 1.0 on self-reported fruit and vegetable intake in adults who visit general practice: a quasi-experimental trial. J Med Internet Res 2016;18(2):e47 [FREE Full text] [doi: 10.2196/jmir.5252] [Medline: 26929095]
47. Silina V, Tessma MK, Senkane S, Krievina G, Bahs G. Text messaging (SMS) as a tool to facilitate weight loss and prevent metabolic deterioration in clinically healthy overweight and obese subjects: a randomised controlled trial. Scand J Prim Health Care 2017;35(3):262-270 [FREE Full text] [doi: 10.1080/02813432.2017.1358435] [Medline: 28812403]
48. Springer A, Venkatakrishnan A, Mohan S, Nelson L, Silva M, Pirolli P. Leveraging self-affirmation to improve behavior change: a mobile health app experiment. JMIR Mhealth Uhealth 2018;6(7):e157 [FREE Full text] [doi: 10.2196/mhealth.9151] [Medline: 30026179]
49. Van der Mispel C, Poppe L, Crombez G, Verloigne M, De Bourdeaudhuij I. A self-regulation-based eHealth intervention to promote a healthy lifestyle: investigating user and website characteristics related to attrition. J Med Internet Res 2017;19(7):e241 [FREE Full text] [doi: 10.2196/jmir.7277] [Medline: 28698168]
50. Whitley MD, Payán DD, Flórez KR, Williams MV, Wong EC, Branch CA, et al. Feasibility and acceptability of a mobile messaging program within a church-based healthy living intervention for African Americans and Latinos. Health Informatics J 2020;26(2):880-896 [FREE Full text] [doi: 10.1177/1460458219853408] [Medline: 31203706]
51. Young CL, Mohebbi M, Staudacher H, Berk M, Jacka FN, O'Neil A. Int Rev Psychiatry 2021;33(3):266-279. [doi: 10.1080/09540261.2020.1854193] [Medline: 34039236]
52. Yuhas M, Brock DP, Ritterband LM, Chow PI, Porter KJ, Zoellner JM. Retention and engagement of rural caregivers of adolescents in a short message service intervention to reduce sugar-sweetened beverage intake. Digit Health 2023;9:20552076231160324 [FREE Full text] [doi: 10.1177/20552076231160324] [Medline: 36949896]
53. Stroebe W, van Koningsbruggen GMV, Papies EK, Aarts H. Why most dieters fail but some succeed: a goal conflict model of eating behavior. Psychol Rev 2013;120(1):110-138. [doi: 10.1037/a0030849] [Medline: 23230892]
54. Flagel SB, Robinson TE. Neurobiological basis of individual variation in stimulus-reward learning. Curr Opin Behav Sci 2017;13:178-185 [FREE Full text] [doi: 10.1016/j.cobeha.2016.12.004] [Medline: 28670608]
55. Kinyanjui MW, Aloka PJ, Mutisya SK, Ndeke FN, Nyang'ara NM. Classroom instruction reinforcement strategies and factors that influence their implementation in Kenyan primary schools. JESR 2015;5(3):267-278. [doi: 10.5901/jesr.2015.v5n3p267]
56. Sackett PR, Lievens F, Van Iddekinge CH, Kuncel NR. Individual differences and their measurement: a review of 100 years of research. J Appl Psychol 2017;102(3):254-273. [doi: 10.1037/apl0000151] [Medline: 28150988]
57. West R, Brown J. Theory of Addiction. UK: Addiction Press; 2013.
58. Hall PA, Fong GT. Temporal self-regulation theory: a model for individual health behavior. Health Psychol Rev 2007;1(1):6-52. [doi: 10.1080/17437190701492437]
59. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
60. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
61. Shafi AA, Khemka M, Roy Choudhury S. A new approach to motivation: four-drive model. J Hum Behav Soc Environ 2015;26(2):217-226. [doi: 10.1080/10911359.2015.1083505]
62. Louro MJ, Pieters R, Zeelenberg M. Dynamics of multiple-goal pursuit. J Pers Soc Psychol 2007;93(2):174-193. [doi: 10.1037/0022-3514.93.2.174] [Medline: 17645394]
63. Baumeister RF, Bratslavsky E, Muraven M, Tice DM. Ego depletion: is the active self a limited resource? J Pers Soc Psychol 1998;74(5):1252-1265. [doi: 10.1037//0022-3514.74.5.1252] [Medline: 9599441]

XSL•FO
RenderX

64. Bandura A. Self-efficacy. In: Ramachaudran VS, editor. Encyclopedia of Human Behavior. London, United Kingdom: Academic Press; 1994:71-81.
65. Rimal RN. Modeling the relationship between descriptive norms and behaviors: a test and extension of the theory of normative social behavior (TNSB). Health Commun 2008;23(2):103-116. [doi: 10.1080/10410230801967791] [Medline: 18443998]
66. Legros S, Cislaghi B. Mapping the social-norms literature: an overview of reviews. Perspect Psychol Sci 2020;15(1):62-80 [FREE Full text] [doi: 10.1177/1745691619866455] [Medline: 31697614]
67. Rivis A, Sheeran P. Descriptive norms as an additional predictor in the theory of planned behaviour: a meta-analysis. Curr Psychol 2003;22(3):218-233. [doi: 10.1348/014466607X258704] [Medline: 18039428]
68. Ramanathan S, Menon G. Time-varying effects of chronic hedonic goals on impulsive behavior. J Mark Res 2018;43(4):628-641. [doi: 10.1509/jmkr.43.4.628]
69. Pfitzner-Eden F. Why do i feel more confident? Bandura's sources predict preservice teachers' latent changes in teacher self-efficacy. Front Psychol 2016;7:1486 [FREE Full text] [doi: 10.3389/fpsyg.2016.01486] [Medline: 27807422]
70. Parkinson J, David P, Rundle‐Thiele S. Self-efficacy or perceived behavioural control: which influences consumers' physical activity and healthful eating behaviour maintenance? J Consum Behav 2017;16(5):413-423. [doi: 10.1002/cb.1641]
71. Vancouver JB, Thompson CM, Tischner EC, Putka DJ. Two studies examining the negative effect of self-efficacy on performance. J Appl Psychol 2002;87(3):506-516. [doi: 10.1037/0021-9010.87.3.506] [Medline: 12090608]
72. Skinner BF. Science and human behavior. In: The B. F. Skinner Foundation. New York, NY, United States: Free Press; 2014.
73. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977;84(2):191-215. [Medline: 847061]
74. Hofmann W, Friese M, Wiers RW. Impulsive versus reflective influences on health behavior: a theoretical framework and empirical review. Health Psychol Rev 2008;2(2):111-137. [doi: 10.1080/17437190802617668]
75. Bjorklund DF, Harnishfeger KK. The resources construct in cognitive development: diverse sources of evidence and a theory of inefficient inhibition. Dev Rev 1990;10(1):48-71. [doi: 10.1016/0273-2297(90)90004-n]
76. Cabeza R, Nyberg L. Imaging cognition II: an empirical review of 275 PET and fMRI studies. J Cogn Neurosci 2000;12(1):1-47. [doi: 10.1162/08989290051137585] [Medline: 10769304]
77. Baumeister RF, Muraven M, Tice DM. Ego depletion: a resource model of volition, self-regulation, and controlled processing. Soc Cogn 2000;18(2):130-150. [doi: 10.1521/soco.2000.18.2.130]
78. Muraven M, Baumeister RF. Self-regulation and depletion of limited resources: does self-control resemble a muscle? Psychol Bull 2000;126(2):247-259. [doi: 10.1037/0033-2909.126.2.247] [Medline: 10748642]
79. Probst Y, Nguyen DT, Tran MK, Li W. Dietary assessment on a mobile phone using image processing and pattern recognition techniques: algorithm design and system prototyping. Nutrients 2015;7(8):6128-6138 [FREE Full text] [doi: 10.3390/nu7085274] [Medline: 26225994]
80. Shonkoff E, Cara KC, Pei XA, Chung M, Kamath S, Panetta K, et al. The state of the science on artificial intelligence-based dietary assessment methods that use digital images: a scoping review. Current Developments in Nutrition 2022;6:534. [doi: 10.1093/cdn/nzac077.037]
81. Granqvist P, Mikulincer M, Shaver PR. An attachment theory perspective on religionspirituality. In: Vail III K, Routledge C, editors. The Science of Religion, Spirituality, and Existentialism. Cambridge, MA: Academic Press; 2020.
82. Ryan P. Integrated theory of health behavior change: background and intervention development. Clin Nurse Spec 2009;23(3):161-172 [FREE Full text] [doi: 10.1097/NUR.0b013e3181a42373] [Medline: 19395894]
83. Bodenhausen GV. Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In: Mackie DM, Hamilton DL, editors. Affect, Cognition, and Stereotyping. US: Academic Press; 1993:13-37.
84. Halama P. Self-regulation capacity and decision making of Slovak managers in routine situations and in situations with possible negative outcomes. SP 2017;59(2):156-168. [doi: 10.21909/sp.2017.02.737]
85. Bastian B, Pe ML, Kuppens P. Perceived social pressure not to experience negative emotion is linked to selective attention for negative information. Cogn Emot 2017;31(2):261-268. [doi: 10.1080/02699931.2015.1103702] [Medline: 26513588]
86. Stasiewicz PR, Maisto SA. Two-factor avoidance theory: the role of negative affect in the maintenance of substance use and substance use disorder. Behavior Therapy 1993;24(3):337-356. [doi: 10.1016/s0005-7894(05)80210-2]
87. Van DT. Motivational and affective components of workplace learning in some sectors in the Netherlands. In: Chisholm L, Lunardon K, Ostendorf A, Pasqualoni PP, editors. Decoding the Meanings of Learning at Work in Asia and Europe. Innsbruck: Innsbruck University Press; 2012:37-55.
88. Maes S, Gebhardt W. Self-regulation and health behavior. In: Boekaerts M, Pintrich PR, Zeidner M, editors. Handbook of Self-Regulation. London, United Kingdom: Academic Press; 2000:343-368. [doi: 10.1016/b978-012109890-2/50040-8]
89. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]

XSL•FO
RenderX

90. Aunger R, Curtis V. The evo-eco approach to behaviour change. In: Gibson MA, Lawson DW, editors. Applied Evolutionary Anthropology: Advances in the Evolutionary Analysis of Human Behaviour. Midtown Manhattan, New York City: Springer; 2014:271-295. [doi: 10.1007/978-1-4939-0280-4_12]
91. Masthoff J, Grasso F, Ham J. Preface to the special issue on personalization and behavior change. User Model User-Adap Inter 2014;24(5):345-350. [doi: 10.1007/s11257-014-9151-1]
92. Lyles CR, Nguyen OK, Khoong EC, Aguilera A, Sarkar U. Multilevel determinants of digital health equity: a literature synthesis to advance the field. Annu Rev Public Health 2023;44:383-405 [FREE Full text] [doi: 10.1146/annurev-publhealth-071521-023913] [Medline: 36525960]
93. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
94. Smith KP, Christakis NA. Social Networks and Health. Annu Rev Sociol 2008;34(1):405-429. [doi: 10.1146/annurev.soc.34.040507.134601]
95. Ewart CK. Social action theory for a public health psychology. Am Psychol 1991;46(9):931-946. [doi: 10.1037//0003-066x.46.9.931] [Medline: 1958012]
96. Clark MS, Reis HT. Interpersonal processes in close relationships. Annu Rev Psychol 1988;39:609-672. [doi: 10.1146/annurev.ps.39.020188.003141] [Medline: 3278682]

Abbreviations

ENTREQ: Enhancing Transparency in Reporting the Synthesis of Qualitative Research
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial

Edited by G Eysenbach, N Cahill; submitted 23.03.24; peer-reviewed by K Porter; comments to author 24.04.24; revised version received 09.05.24; accepted 26.07.24; published 27.08.24.

Please cite as:
Wang J, Mahe J, Huo Y, Huang W, Liu X, Zhao Y, Huang J, Shi F, Li Z, Jiang D, Li Y, Perceval G, Zhao L, Zhang L
Framework Development for Reducing Attrition in Digital Dietary Interventions: Systematic Review and Thematic Synthesis
J Med Internet Res 2024;26:e58735
URL: https://www.jmir.org/2024/1/e58735
doi:10.2196/58735
PMID:

©Jian Wang, Jinli Mahe, Yujia Huo, Weiyuan Huang, Xinru Liu, Yang Zhao, Junjie Huang, Feng Shi, Zhihui Li, Dou Jiang, Yilong Li, Garon Perceval, Lindu Zhao, Lin Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Approaches to Evaluating Digital Health Technologies: Scoping Review

Anneloek Rauwerdink[1*], MD; Pier Spinazze[2*], MD, MSc, MBA; Harm Gijsbers[3,4,5], MSc; Juul Molendijk[6], MD; Sandra Zwolsman[4], PhD; Marlies P Schijven[3,4,7], MD, MSc, MHsc; Niels H Chavannes[2,8], MD, PhD; Marise J Kasteleyn[2,8], PhD

[1]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, Netherlands

[2]Department of Public Health and Primary Care, Leiden University Medical Centre, Leiden, Netherlands

[3]Department of Surgery, Amsterdam University Medical Center, Amsterdam, Netherlands

[4]Digital Health, Amsterdam Public Health Institute, Amsterdam, Netherlands

[5]Department of Medical Informatics, eHealth Living & Learning Lab, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands

[6]Amsterdam University Medical Center, Amsterdam, Netherlands

[7]Amsterdam Gastroenterology Endocrinology Metabolism, Amsterdam, Netherlands

[8]National eHealth Living Lab, Leiden, Netherlands

[*]these authors contributed equally

Corresponding Author:
Anneloek Rauwerdink, MD
Department of Radiology and Nuclear Medicine
Amsterdam University Medical Center
Meibergdreef 9
Amsterdam, 1105 AZ
Netherlands
Phone: 31 642379975
Email: anneloek@hotmail.com

Abstract

Background: Profound scientific evaluation of novel digital health technologies (DHTs) is key to enhance successful development and implementation. As such, we previously developed the eHealth evaluation cycle. The eHealth evaluation cycle contains 5 consecutive study phases: conceptual, development, feasibility, effectiveness, and implementation.

Objective: The aim of this study is to develop a better understanding of the daily practice of the eHealth evaluation cycle. Therefore, the objectives are to conduct a structured analysis of literature data to analyze the practice of the evaluation study phases and to determine which evaluation approaches are used in which study phase of the eHealth evaluation cycle.

Methods: We conducted a systematic literature search in PubMed including the MeSH term “telemedicine” in combination with a wide variety of evaluation approaches. Original peer-reviewed studies published in the year 2019 (pre-COVID-19 cohort) were included. Nonpatient-focused studies were excluded. Data on the following variables were extracted and systematically analyzed: journal, country, publication date, medical specialty, primary user, functionality, evaluation study phases, and evaluation approach. RStudio software was used to summarize the descriptive data and to perform statistical analyses.

Results: We included 824 studies after 1583 titles and abstracts were screened. The majority of the evaluation studies focused on the effectiveness (impact; 304/824, 36.9%) study phase, whereas uptake (implementation; 70/824, 8.5%) received the least focus. Randomized controlled trials (RCTs; 170/899, 18.9%) were the most commonly used DHT evaluation method. Within the effectiveness (impact) study phase, RCTs were used in one-half of the studies. In the conceptual and planning phases, survey research (27/78, 35%) and interview studies (27/78, 35%) were most frequently used. The United States published the largest amount of DHT evaluation studies (304/824, 36.9%). Psychiatry and mental health (89/840, 10.6%) and cardiology (75/840, 8.9%) had the majority of studies published within the field.

Conclusions: We composed the first comprehensive overview of the actual practice of implementing consecutive DHT evaluation study phases. We found that the study phases of the eHealth evaluation cycle are unequally studied and most attention is paid to the effectiveness study phase. In addition, the majority of the studies used an RCT design. However, in order to successfully

XSL•FO
RenderX

develop and implement novel DHTs, stimulating equal evaluation of the sequential study phases of DHTs and selecting the right evaluation approach that fits the iterative nature of technology might be of the utmost importance.

(J Med Internet Res 2024;26:e50251) doi:10.2196/50251

KEYWORDS

ehealth; digital health; telemedicine; methods; evaluation; literature review; DHI; health technology; health technologies; telehealth; e-health; telemedicine; scoping review; digital health intervention

Introduction

Background

Health care, traditionally a slow adopter of digital innovation, has recently seen an acceleration in the use of digital tools and systems. A transition to more remote patient care and associated services is urgently needed due to an aging global population [1]. In a highly regulated sector in which data privacy and health outcomes are vital, there needs to be a high level of scrutiny and standardized evaluation frameworks in place to limit the inherent risk in rapid innovation and ensure successful outcomes. With an increasing focus on digital health research, it is important to take stock of digital health evaluation methodologies to further guide prospective studies [2].

The World Health Organization (WHO) plays an important role in guiding and accelerating the development of digital health interventions and health innovation globally. In 2018, the WHO developed a *classification scheme of digital health interventions*, aiming to promote a comprehensive and standardized language for health program planners [3]. The scheme organized digital health interventions into the following primary user groups: clients, health care providers, health systems or resource managers, and data services (Multimedia Appendix 1).

Another valuable framework was developed by the National Institute for Health and Care Excellence (NICE) together with the National Health Service in England in 2019 [4,5]. They created the Evidence Standards Framework (ESF) for digital health technologies (DHTs) in order to promote greater consistency when evaluating or commissioning DHTs and thereby enhance the level of scrutiny, which is generally lower than the level of evidence required for drugs or devices. The subjects of NICE's ESF are system services (evidence tier 1); inform, simple monitoring, communicate (evidence tier 2); preventative behavior change, self-manage (evidence tier 3a); and treatment, active monitoring, calculate, and diagnose (evidence tier 3b).

The WHO's digital health classification scheme and the NICE's ESF could work synergistically, providing a much needed standardized and accepted framework for the varied stakeholders involved in digital health to evaluate and improve the development of evidence-based digital health solutions. To provide further granularity to the focus of digital health research, evaluating the sequential study phases—conceptual to implementation—is important [6,7]. Therefore, we previously developed the *eHealth evaluation cycle* (Figure 1) based on existing eHealth evaluation frameworks [8]. The *eHealth evaluation cycle* contains the following 5 consecutive study phases: conceptual and planning; design, development, and usability; pilot (feasibility); effectiveness (impact); and uptake (implementation). In the online version of the *eHealth evaluation cycle*, one can find a synopsis of evaluation approaches (all the methods, study designs, frameworks, and other philosophical approaches) that could be used to evaluate a particular study phase [9]. For example, a "concept mapping study design" can be used to gather information in the conceptual and planning study phase, and an "economic evaluation" can be used in the uptake (implementation) study phase [10]. There are also several types of systematic reviews within the *eHealth evaluation cycle*. For instance, a meta-analysis can be found within the effectiveness (impact) phase, and a narrative review can be found in the conceptual and planning phase when one is, for example, at the start of developing a new DHT and aims to define what is already known.

XSL•FO
RenderX

Figure 1. eHealth evaluation cycle.

Although the number of DHT publications has exponentially grown in recent years, not much is known about the use in daily practice of the consecutive study phases [11]. There are indications that there is too much focus on evaluating *effectiveness* instead of, for example, usability testing [12]. In addition, large-scale implementation of DHT often fails [13-15]. Therefore, in order to potentially improve DHT evaluation and therefore implementation of DHTs, the following research question will be addressed in this study: What is the actual practice of the consecutive evaluation study phases described in the literature?

Aim and Objective

The aim of this study was to develop a better understanding of the daily practice of the *eHealth evaluation cycle*. Therefore, the research objectives were to (1) conduct a structured analysis of data from the literature to analyze the practice of the evaluation study phases and (2) determine which evaluation approaches are used in which study phase of the eHealth evaluation cycle.

Methods

Overall Design

We performed a scoping review subdivided into 2 phases: (1) systematic literature search to find articles concerning the evaluation of DHT, followed by the extraction of data from the selected articles; (2) performing a structured data analysis. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 2).

Search Strategy

The online PubMed database was systematically searched using the MeSH term “telemedicine” in combination with a wide variety of “evaluation approaches” (see Multimedia Appendix 3 for the complete search string). Articles written in English published in the year 2019 were included. The year 2019 was chosen since it is the most recent “pre-COVID-19” year cohort. Herewith, we aimed to avoid possible skewing of results by the temporary surge in COVID-19–related DHTs.

Inclusion and Exclusion Criteria

The following inclusion and exclusion criteria were used. The inclusion criteria consisted of any original, peer-reviewed study evaluating a specific DHT in one or multiple study phases of the eHealth evaluation cycle; all types of literature reviews describing a specific DHT; and tele-education studies when closely related to the outcomes of health care.

The exclusion criteria were papers not written in English; studies not related to DHTs; nonpatient-focused studies (eg, evaluation of technology alone [ie, sensitivity of sensors]); and poster presentations, protocol studies, and opinion papers.

XSL•FO
RenderX

Study Selection

Articles were included based on screening the title and abstract by 2 pairs of researchers. HG and PS reviewed articles with titles starting with the letters A through K, and AR and JM reviewed articles with titles starting with the letters L through Z. Two test series of 20 cases were discussed between the 4 researchers to reach a satisfactory level of consensus before the pairs began screening the titles and abstracts. Rayyan QCRI online software was used to collectively screen articles [16].

Once the inclusion screening was complete, each pair discussed conflicting studies. Disagreements between researchers were resolved by the decision of a third researcher (HG or AR). The 2 sets of included studies were merged and randomly divided into 4 individual sets. Each of the researchers received 1 data set from which to extract the data (described in the Data Extraction section). Most of the data could be extracted from the abstract alone. Therefore, an article's full text was only reviewed in the event that the information in the abstract was insufficient. If the researcher thought the article was not suitable for inclusion in the study after a more in-depth review, it was excluded.

Data Extraction

Data extraction was done using a standardized data extraction form (see example in Multimedia Appendix 4) that was developed by AR and HG using Microsoft Office Excel version 16.52. The form was pilot tested by the group of 4 researchers (AR, HG, JM, and PS) on 10 articles and modified afterwards. The data items that were extracted from the articles are described in Table 1. The variable *functionality* was based upon the NICE ESF for DHTs, and the variable *primary user* was based upon the WHO's digital health classification scheme.

XSL•FO
RenderX

Table 1. Data items extracted from the articles.

Generic variables	
	Journal
	Country[a]
	Publication date
	Medical specialty[b]
Specific variables	
Primary user[c] [3]	Clients
	Health care providers
	Health system or resource managers
	Data services
Functionality[d] [4]	System services
	Inform
	Simple monitoring
	Communicate
	Preventative behavior change
	Self-manage
	Treat
	Active monitoring
	Calculate
	Diagnose
	Multiple functionalities
	Other
	Unclear
Evaluation study phases[e] [8]	Conceptual and planning
	Design, development, and usability
	Pilot (feasibility)
	Effectiveness (impact)
	Multiple phases
Evaluation approach	Free-text field

[a]Country in which the study was conducted, or if this was not clear, the country of the first author.

[b]Medical specialty the digital health technology applies to, with multiple options possible; a standardized list was used.

[c]Categories from the World Health Organization classification scheme of digital health interventions were extracted; multiple categories may be extracted.

[d]Categories from the National Institute for Health and Care Excellence's evidence standards framework were extracted.

[e]Categories from the eHealth evaluation cycles were extracted.

The description of the *evaluation approach* was literally extracted by copying and pasting from the article. If available, multiple evaluation approaches were extracted.

The 4 researchers (AR, HG, JM, and PS) independently extracted the data into the standardized data extraction form. After the data extraction was completed, AR and HG each randomly checked 15 cases from each investigator to evaluate for inconsistencies among the extracted data. When a data item showed an interrater disagreement of more than 10%, a second investigator extracted the data for the specific data item, and a third investigator made the final decision on the inequalities. Considering the extensive workload, this extra step of cross-checking data was performed unblinded. The researchers' group discussion was planned ahead of the second researcher's data extraction to increase the level of consensus. Finally, the completed data extraction forms were merged again, checked by AR, and altered for consistency. The description of each of the extracted evaluation approaches was checked and substituted by a better general description to allow for statistical analysis;

XSL•FO
RenderX

for example, “randomized controlled trial” was changed to “RCT.”

Data Synthesis and Analysis

RStudio vs 2022.07.1 software was used to summarize the descriptive data and to perform statistical analyses. The proportions of the categories for each data item are described using percentages. The variable *evaluation approach* was extracted as a free text field in the data extraction form. It was expected that the data extraction for the variable *evaluation approach* would yield a very wide array of different evaluation approaches. Therefore, to design the clearest visual presentation of the data, only the top 8 (cross-tabulation bar chart) and top 10 (bar chart) most frequently used *evaluation approaches* were used for further analysis. Concerning the variable *medical specialty*, the top 5 (cross-tabulation bar chart) and top 10 (bar chart) were used for further analysis. For the variable *country*, only the top 10 (bar chart) were used for analysis.

To examine relationships between the nominal variables, the following cross-tabulations were executed: *primary user* versus *functionality*, *primary user* versus *evaluation study phase*, *primary user* versus top 8 *evaluation approach*, *functionality* versus *evaluation study phase*, *functionality* versus top 8 *evaluation approach*, *evaluation study phase* versus top 5 *medical specialties*, *evaluation study phase* versus top 8 *evaluation approach*, and top 5 *medical specialty* versus top 8 *evaluation approach*.

Results

After the PubMed database search, 1583 studies were considered relevant according to the inclusion criteria. During the screening of the title and abstract, 716 records were excluded. A further 43 articles were excluded after reading the articles more in depth during the data extraction assessment phase. Finally, 824 studies were included. The PRISMA flowchart summarizes the article selection process (Figure 2).

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

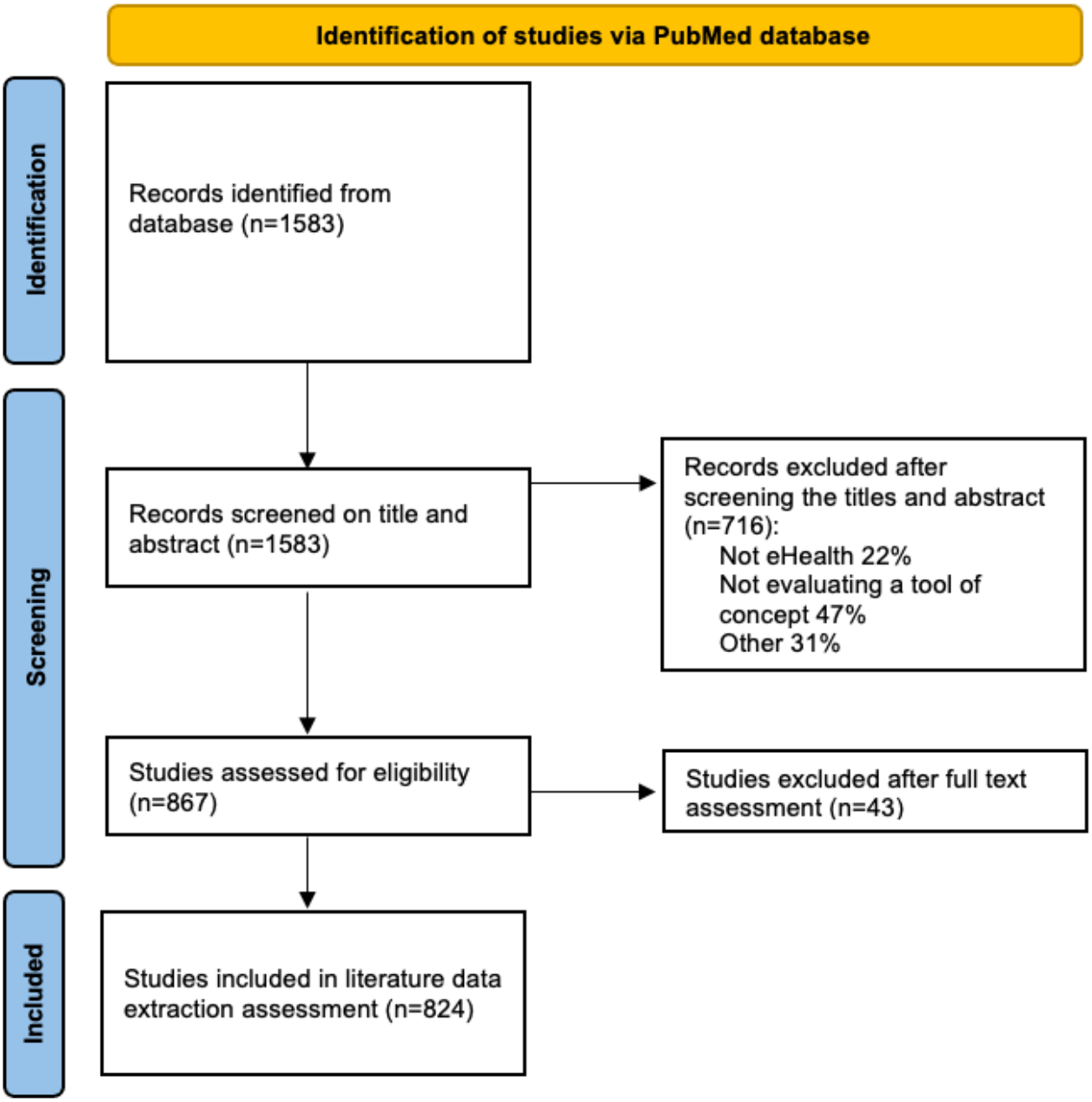

The independent evaluation of the extracted variables from 15 randomly selected records from each of the 4 researchers by AR and HG revealed unacceptable inconsistencies in the extracted data for the variables *functionality* (27/90, 30%) and *evaluation study phases* (18/90, 20%). Following a discussion among the 4 researchers, it was decided that the variables *functionality* and *evaluation study phases* would require a full second review. After the full second review, the variable *functionality* had a mismatch of 29% (250/867), and the variable *evaluation study phases* had a mismatch of 21% (113/867). A third reviewer made the final decision on the discrepancies between the 2 reviewers. All of the researchers participated in the second review, and all made final decisions as a third reviewer.

The top 10 of the general variables *country* and *medical specialty* are shown in Table 2. The complete frequency list of the variables *country*, *medical specialty*, and publishing *journals* from ≥5 studies can be found in Multimedia Appendix 5.

XSL•FO
RenderX

Table 2. Top 10 countries and medical specialties in the 824 publications.

Top 10	Results, n (%)
Country[a]	
United States	307 (36.9)
United Kingdom	63 (7.6)
Australia	55 (6.6)
Canada	49 (5.9)
Netherlands	32 (3.8)
China	31 (3.7)
Germany	29 (3.5)
Spain	20 (2.4)
Italy	15 (1.8)
Brazil	15 (1.8)
Medical specialty[b]	
Psychiatry and mental care	89 (10.6)
Cardiology	75 (8.9)
Neurology	59 (7)
Primary care	50 (6)
Public health	48 (5.7)
Obstetrics, gynecology, and midwifery	44 (5.2)
Pediatrics	43 (5.1)
Endocrinology	38 (4.5)
Applies to all specialties	36 (4.3)
Internal medicine	32 (3.8)

[a]Total countries=73.

[b]Total medical specialities=44.

Table 3 shows the descriptive statics of the specific variables. The majority of DHT-related research was published in the field of psychiatry and mental health (89/840, 10.6%) and cardiology (75/840, 8.9%). Concerning DHT *primary user* categories, clients (491/947, 51.8%) were slightly more frequently targeted than health care providers (425/947, 44.9%). Communication (160/824, 19.4%) and treatment (150/824, 18.2%) were the most frequent categories in the *functionality* variable.

XSL•FO
RenderX

Table 3. Descriptive statics of specific variables of the studies about digital health technologies (DHTs).

Variables	Frequency, n (%)
Primary user (n=947 users)	
Clients	491 (51.8)
Health care providers	425 (44.9)
Health system or resource managers	26 (2.7)
Data services	5 (0.5)
Functionality (n=824 studies)	
Communicate	160 (19.4)
Treat	150 (18.2)
Multiple	103 (12.5)
Diagnose	87 (10.6)
Self-manage	78 (9.5)
Active monitoring	67 (8.1)
Preventative behavior change	61 (7.4)
Inform	49 (5.9)
Simple monitoring	29 (3.5)
System services	26 (3.2)
Calculate	6 (0.7)
Unclear	5 (0.6)
Other	3 (0.4)
Evaluation study phases (n=824 studies)	
Effectiveness (impact)	304 (36.9)
Pilot (feasibility)	232 (28.2)
Design, development, and usability	99 (12)
Conceptual and planning	96 (11.7)
Uptake (implementation)	70 (8.5)
Multiple phases	23 (2.8)
Top 10 evaluation approaches[a] (n=899)	
RCT[b]	170 (18.9)
Survey research	91 (10.1)
Cohort study (prospective)	73 (8.1)
Interview study	58 (6.5)
Cohort study (retrospective)	54 (6)
Mixed methods study design	42 (4.7)
Systematic review	39 (4.3)
Cross-sectional study	36 (4)
Feasibility study	31 (3.4)
Pilot study	25 (2.8)

[a]108 distinct evaluation approaches were listed.

[b]RCT: randomized controlled trial.

The variable *evaluation study phase* showed that nearly 37% (304/824, 36.9%) of the studies were carried out to study the effectiveness of a certain DHT tool. Almost one-third (232/824, 28.2%) of the studies were in the pilot study phase. RCT was the most frequently used (170/899, 18.9%) *evaluation approach.* In total, 108 distinct *evaluation approaches* were encountered

XSL•FO
RenderX

(Multimedia Appendix 6). Although the top 10 consisted of well-known epidemiologic methods, we did encounter novel evaluation methods such as the fit between individual, task, and technology (FITT) framework [17], nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [18], CeHRes Roadmap [19], and systems development life cycle (SDLC) methodology [20]. In addition, several variants of well-known epidemiological methods, such as nonrandomized group comparison study [21] and retrospective record review [22], were identified.

The proportional stacked bar chart in Figure 3 illustrates the cross-tabulation of *evaluation study phase* versus the top 8 *evaluation approaches*. In the effectiveness study phase, RCTs were used in one-half of the studies (124/248, 50%). Prospective and retrospective cohort studies were used in more than one-quarter of the effectiveness studies (68/248, 27.4%). Survey research (27/78, 35%) and interview studies (27/78, 35%) were more commonly used for the conceptual and planning study phase. In the uptake (implementation) study phase, the evaluation approaches used were generally equally divided.

Figure 3. Bar chart evaluation of the study phase versus the top 8 evaluation approaches.

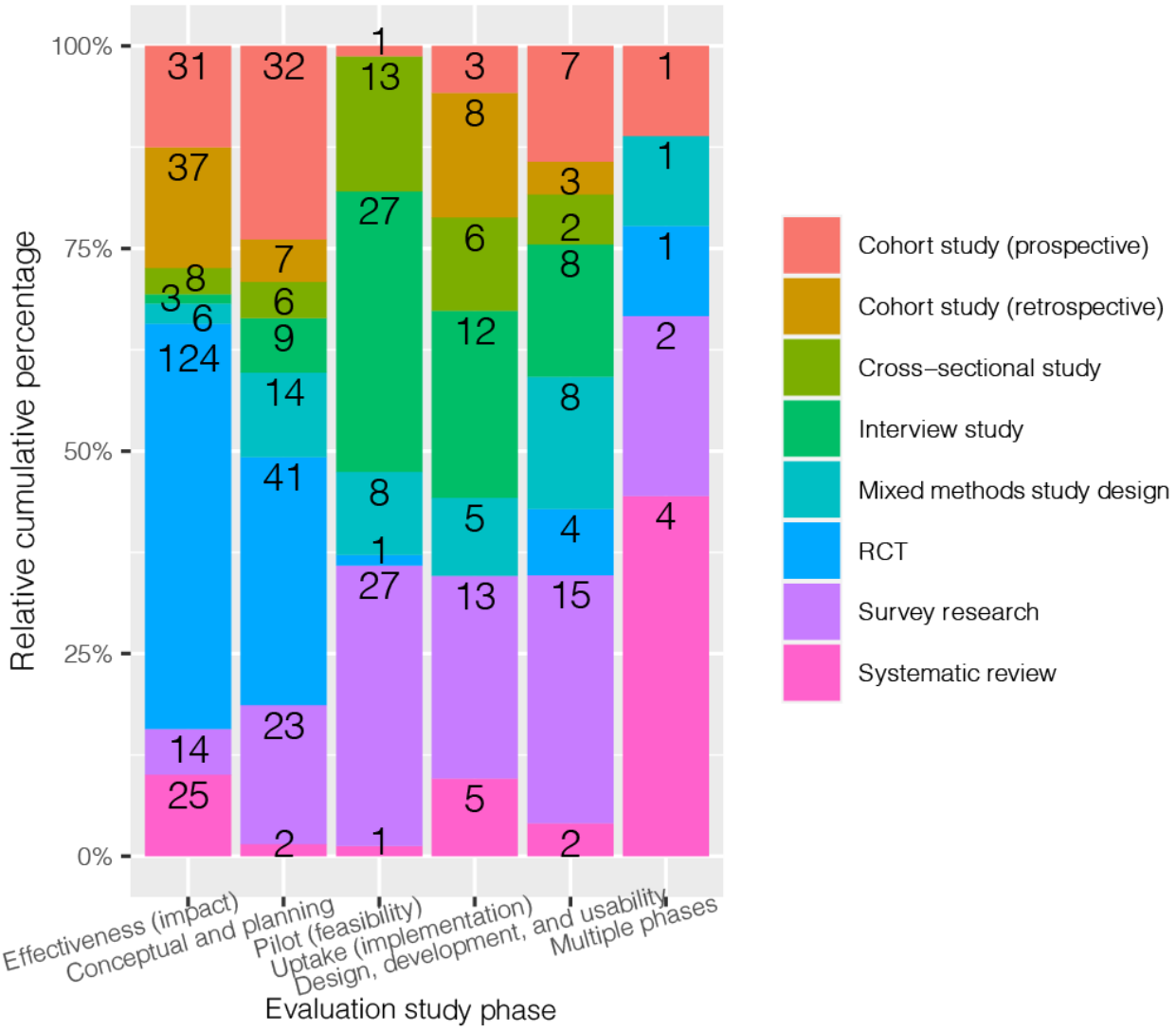

The proportional stacked bar chart of the cross-tabulation of *evaluation study phase* versus the *top 5 medical specialties* illustrates the differences in emphasis of the selected evaluation study phases between the *medical specialties* (Figure 4). Cardiology had the most studies performed in the effectiveness (impact) study phase (40/75, 54%) and fewer in the pilot study phase (14/75, 19%). Within the medical specialty of neurology, the opposite was found; there was more focus on the pilot study phase (26/59, 44%) and less on the effectiveness study phase (19/59, 32%). Primary care had the most evenly divided chart, with the study phase uptake (implementation; 8/50, 16%) appearing to be of greater importance when compared with other specialties.

XSL•FO
RenderX

Figure 4. Bar chart evaluation of study phase versus medical specialty.

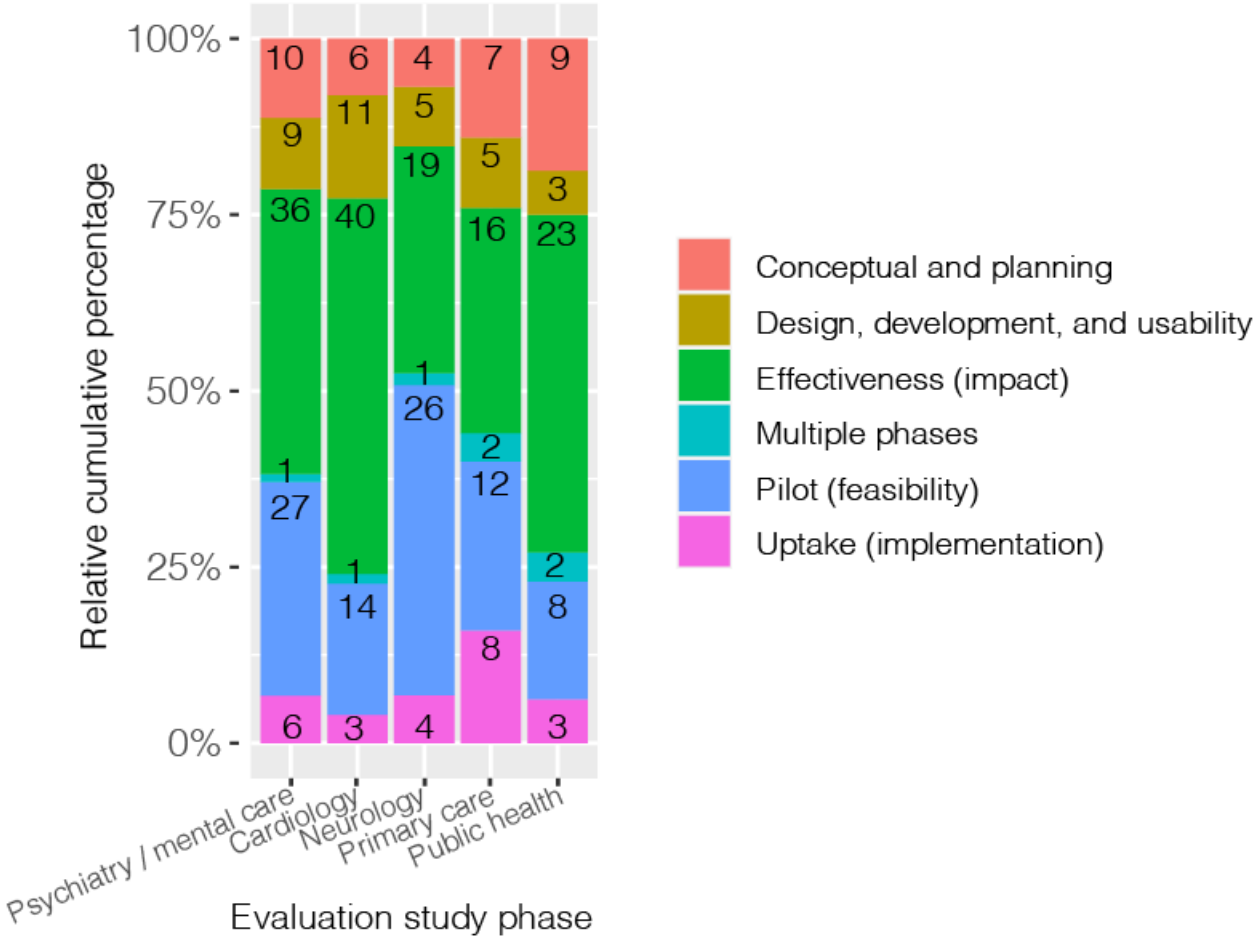

Further, when looking at the proportional stacked bar chart of the cross-tabulation of *medical specialty* versus the *top 8 evaluation approaches*, the chart shows differences concerning medical specialties and the depicted methodology (Figure 5). The RCT represents the biggest share for all medical specialties, with public health as the front-runner (14/30, 47%). Again, primary care had the most equally divided chart when compared with other specialties.

XSL•FO
RenderX

Figure 5. Bar chart evaluation of medical specialty versus the top 8 evaluation approaches.

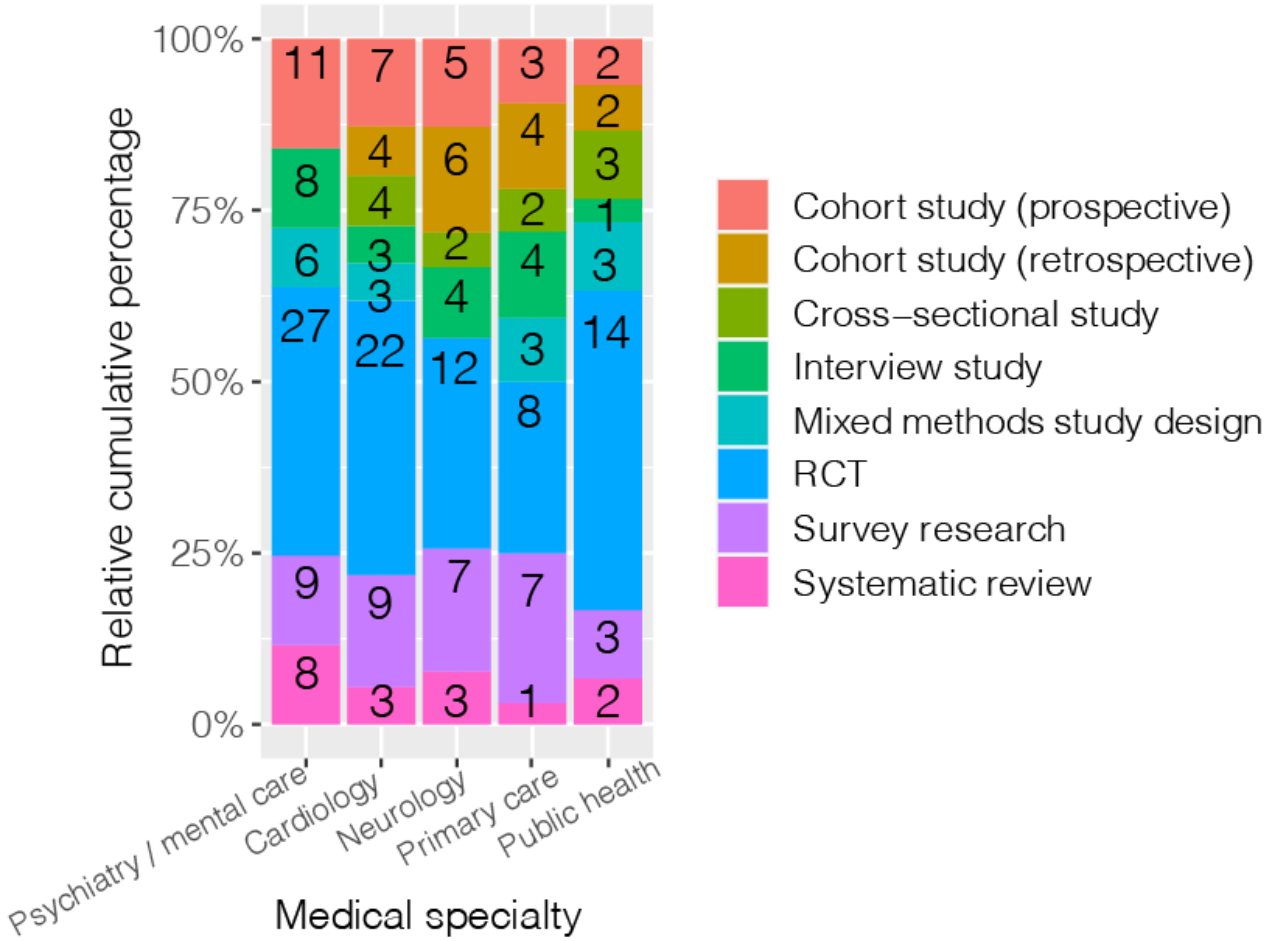

The cross tabulation of *functionality* versus *evaluation study phase* focuses on the breakdown of the study phase in relation to the assumed DHT function (Figure 6). The effectiveness (impact) study phase was most frequently found in the treatment *functionality* category (75/150, 50%). The pilot (feasibility) study phase was most dominant in the diagnose *functionality* category (40/86, 46%). In almost all categories of the *functionality* variable, significantly less attention was paid to the conceptual and planning; design, development, and usability; and uptake (implementation) study phases.

XSL•FO
RenderX

Figure 6. Bar chart evaluation of functionality versus evaluation study phases; the "unclear," calculate," and "other" functionality categories were omitted due to limited results.

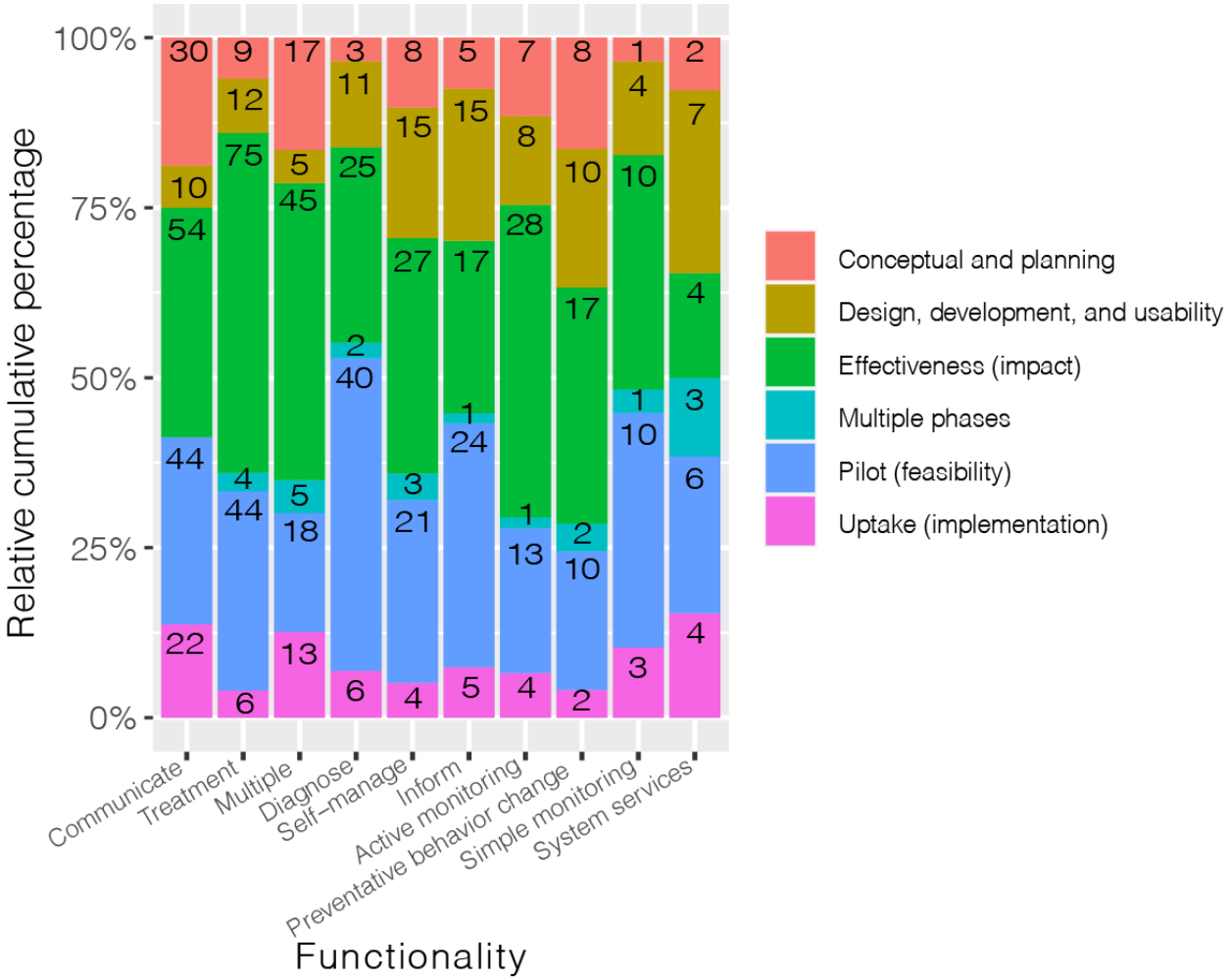

The cross tabulations of *primary user* versus *functionality*, *primary user* versus *evaluation study phase*, *primary user* versus the *top 8 evaluation approaches*, and *functionality* versus the *top 8 evaluation approaches* resulted in an almost equal distribution of the categories. The frequency tables of all the performed cross tabulation analyses, as described in the Methods section, can be found in Multimedia Appendix 7.

Discussion

The aim of this study was to develop a better understanding of the daily practice of the eHealth evaluation cycle. To our knowledge, we composed the first comprehensive overview of the actual practice of the consecutive DHT evaluation study phases. We summarized our main findings in the infographic in Figure 7.

XSL•FO
RenderX

Figure 7. Visual summary of the main findings. RCT: randomized controlled trial.

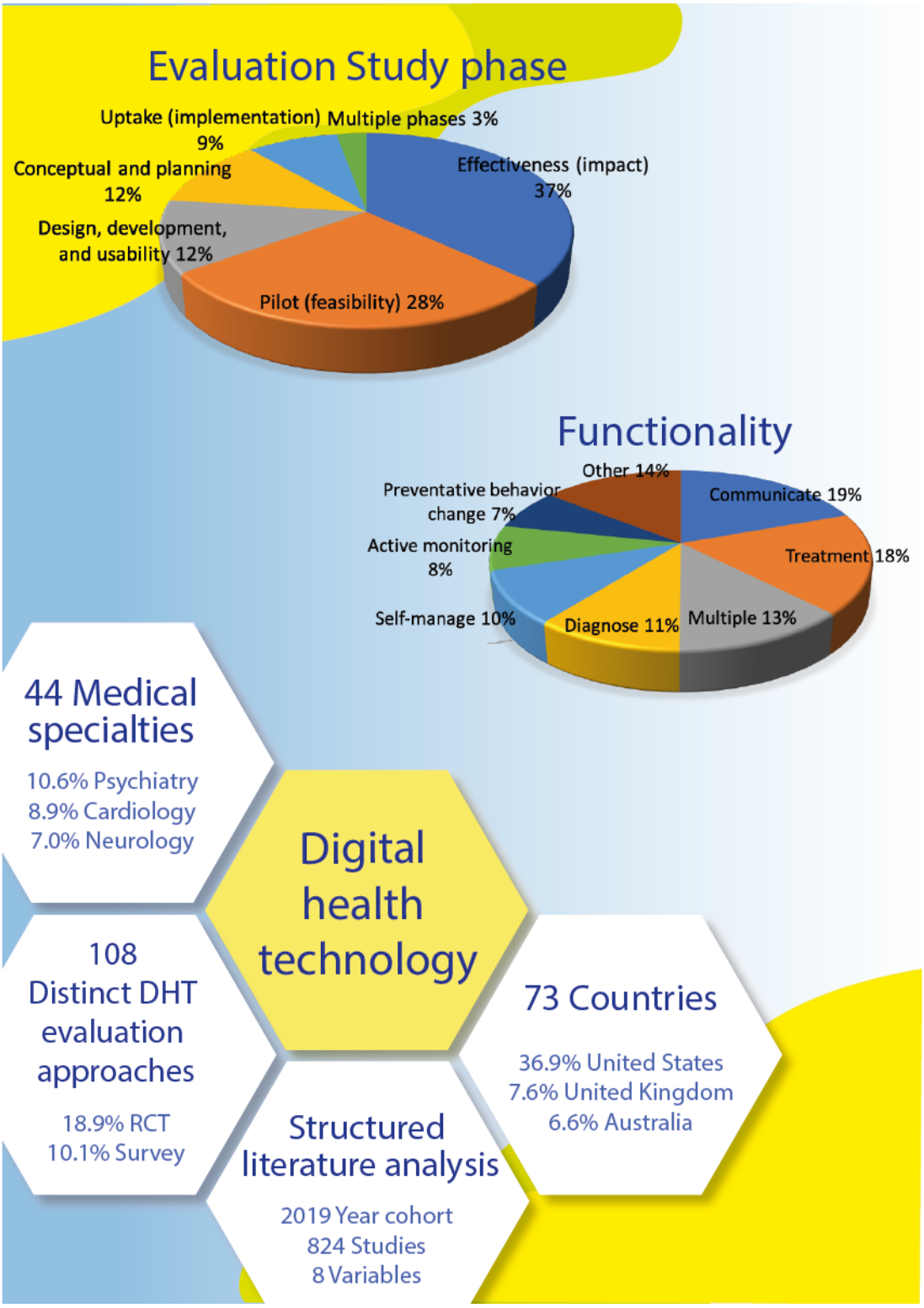

Evaluation Study Phase and Evaluation Approach

Our study highlighted disparities in the attention given to the consecutive evaluation study phases. A predominant focus on the effectiveness and pilot study phases was found. The uptake phase, crucial for successful implementation and scale-up, received the least emphasis. In addition, less focus has been placed on the planning and development study phases. This fundamental mismatch between the context and the technology is the main reason recognized by the WHO that up to 75% of new medical devices fail [23]. A recently published original study from Royle et al [24] also aimed to create awareness about a lack of evaluation of the methodological steps necessary for developing and testing new clinical care pathways. They proposed "the technology clinical trial," describing (1) cocreation of care pathways and addressing information governance, data protection, regulatory, and ethical questions by design; (2) delivery of the technology clinical trial; and (3)

XSL•FO
RenderX

supporting future research and uptake in practice. This 3-step methodology is comparable to the *eHealth evaluation cycle*; however, the relation of the study phase with DHT evaluation approaches is not described.

The RCT emerged as the most frequently used evaluation approach, posing challenges in the context of rapid and iterative DHT development [25]. We did, however, encounter studies using RCT "subdesigns" such as sequential multiple assignment randomized trial (SMART) and microrandomized trials (Multimedia Appendix 5). These designs allow for limited interventional modification within preset boundaries and therewith might bridge the methodological gap. Recently, Hrynyschyn et al [26] conducted a scoping review aimed at providing an overview of existing evaluation methods of DHTs beyond the RCT. They described in detail microrandomization trials, (fractional) factorial RCTs, SMART, and stepped-wedge cluster randomized trials as promising alternatives.

In the conceptual and planning study phase, survey research and interview studies were each used in one-third of the cases. Although these findings seem obvious, as formative research approaches apply well to the earlier study phases of a development cycle, the importance of thorough "background" research in relation to successful implementation cannot be emphasized enough.

Functionality (NICE)

In this study, the *functionality* categories of communication and treatment from the NICE's ESF were the most frequently studied [4,5]. In the *functionality* category of treatment, one-half of the studies focused on the effectiveness (impact) study phase. Concerning the *functionality* category of diagnose, the majority of the studies were found in the pilot (feasibility) study phase. DHTs in the treatment and diagnose categories are both considered "tier 3b" according to the ESF. This implies that designed DHTs must demonstrate effectiveness through a high-quality intervention study (experimental or quasiexperimental design) showing improvements in relevant outcomes. Therefore, it seems to make sense that the majority of the studies focused on the pilot (feasibility) and effectiveness (impact) study phases. Unfortunately, there seems to be too little attention paid to the predevelopment and postdeployment phases, which are also not included in the NICE's ESF.

DHTs are often developed by commercial entities, which then need validation through research to allow for regulatory clearance by governing bodies (eg, the Food and Drug Administration [FDA]) and subsequent adoption and commercialization. Hence, they may only pursue research following the development of a DHT, allowing for speed and rapid prototyping unhindered by slow academic rigor. Often, the development phase involves some form of user testing and prototyping with feedback; however, the value in publishing this is less pronounced and perhaps even avoided to protect intellectual property and avoid competition before commercialization. Despite the time limitations and stakeholder involvement required, perhaps greater attention should be placed on evaluating DHTs focused on treatment in the earlier design and conceptual phases to improve success and uptake at later phases (in terms of costs, usefulness, and adoption, for example).

The CeHRes roadmap is one proposed approach to the development of eHealth interventions. It incorporates both a human-centered design and a business model focus, to create potentially value-adding and sustainable eHealth technologies [27].

Primary User (WHO)

The target end users of studies, as described by the WHO, were primarily focused on clients and health care providers. Health systems or resource managers and data services accounted for less than 4% of studies. This might be due to the limitation of the search to the clinical database PubMed. For example, the IEEE database perhaps would have yielded more relevant results. However, it also might suggest that the majority of technological innovation and implementation is focused on providers and clients (ie, at the point of care rather than at the system and infrastructure levels). This is understandable, considering that health organizations have a level of complexity and fragmented structures that constrain their ability to adopt organization-wide digitization [28,29]. As such, a recent study described how to improve and manage sustainable hybrid (eHealth and face-to-face) health care on an organizational level through the use of the Hybrid Health Care Quality Assessment [30]. The other consideration regarding target end users is the rate of adoption, which is a big contributor to successful implementation and uptake of health technology. Providers remain, to a large extent, the gatekeepers of this, as health care professionals' acceptance is reported as an important requirement for the success of clinical systems [31]. This could explain why the majority of DHTs and their evaluation studies are focused on providers and clients.

Countries and Medical Specialties

The United States published the largest number of DHT evaluation studies. The number of studies per country aligns with the overall ranking of health publications in general per country being the United States, the United Kingdom, Australia, and Canada. However, China, which usually features in third place, featured slightly lower, in sixth place, and the Netherlands placed significantly higher, moving to fifth from 13th place [32]. These discrepancies may be explained by the fact that our study excluded Chinese databases or non-English papers. The numbers for the Netherlands suggest that there is a greater attention to eHealth evaluations in general. According to the Healthcare Information and Management Systems Society (HIMSS), the Netherlands is one of the front-runners in the digitalization of health care [33].

Psychiatry and mental health was the most frequently encountered medical specialty involved in DHT research, followed by cardiology and neurology. A 2018 US-based, weighted survey on the adoption of DHT by physicians in different specialties showed a varied spectrum of adoption rates. Consistent with the data we collected in this scoping review, 27.8% of psychiatrists and 24.1% of cardiologists used DHT for patient interactions [34]. DHTs for mental health have seen increased proliferation, considering inherent benefits such as anonymity, accessibility, and acceptability, possibly explaining why this is the leading specialty in DHT research studies [35].

XSL•FO
RenderX

When relating the medical specialties to the evaluation study phases, the medical specialties of cardiology and neurology showed interestingly opposing results. Of the included cardiology studies, one-half were in the effectiveness phase, and less than one-fifth were in the pilot study phase. As for neurology studies, one-third of the studies were in the effectiveness phase, and almost one-half were in the pilot study phase. As such, perhaps there is less focus on the pilot phase within cardiology, and fewer neurology studies progress to the effectiveness study phase. This could also be an indication of the stage of technological development and application in each specialty and the year of publication we evaluated (2019). For example, technology has been used in cardiology since the first implantable pacemaker in 1958, with studies on topics such as remote monitoring and virtual reality surfacing between 2000 and 2010, and technology has sustained considerable scholarly attention ever since [36,37]. Whereas in neurology, for example, there are very few papers on fall detection using the Internet of Things before 2010, as this is a relatively newer field of study [38].

Limitations

Although our scoping review was thorough, there are some limitations to mention. We only searched in the PubMed database and included only English, peer-reviewed literature, which excludes studies based in foreign contexts (eg, Chinese as well as grey literature), which are used by a large number of commercial institutions involved in the development of DHTs. However, this would have had questionable value regarding methodology and scientific rigor. Only papers published in 2019 were considered in the data extraction step. The decision to include studies from 2019 was made in order to have the most up-to-date data available, without considering the impact of the COVID-19 pandemic on DHT research. Because we only extracted data from studies published in 2019, we might have missed novel evaluation approaches. Further, a possible shift in the evaluation approaches used over time could not be evaluated. It would be interesting to conduct comparisons between time points in a future study (eg, 2015 vs 2019 vs 2023 cohort).

Due to unacceptable discrepancies between the researchers when including data during the data extraction phase for the variables *functionality* and *evaluation study phase*, a second unblinded cross-checking was performed to facilitate concordance, which may have introduced a level of bias. Further, during data extraction, the reviewers standardized the terminology for evaluation approaches, which is open to a level of bias in interpretation. However, this should be minimal considering research uses common terminology in general.

Unfortunately, we were not able to include a variable describing the DHT (eg, “patient portal” or “mHealth” [mobile health]). In the majority of the included studies, the description of the DHT varied too much. Therefore, more in-depth (sub)analyses concerning the type of DHT were not possible. In our opinion, a well-developed, easy-to-use DHT taxonomy would be of great help to classify and evaluate future DHT research. Last, the NICE ESF for DHT was updated in August 2022. However, the majority the subgroups in our study are still used in the NICE framework.

Future

It would be valuable to track which technologies are evaluated in a single study phase or in multiple study phases. However, to do this, each technology would require unique identification that would allow traceability across studies. This would allow visibility on which technologies that begin from a conceptual evaluation phase are eventually implemented and evaluated in subsequent phases for effectiveness and uptake. It would also provide clarity on how many technologies remain in the research stage and fail to see practical clinical implementation.

The results of this study may be used to pinpoint areas of DHT research that require more focus and support the completion of multiple steps of the *eHealth evaluation cycle*. In such, future research could also look at a temporal view of research in relation to medical specialty and evaluation phase, as this could provide further insights as to the stage of digital transformation and where certain fields are lagging behind.

The Excel data file (see the Data Availability section) with the extracted data from the 824 included studies is published online [9,39]. We encourage interested researchers to use the filter options to look at subsets of data for the field of interest. For example, when one is interested in cardiology studies in the pilot phase, one can select “pilot” study phase, and cardiology studies within the pilot study phase from 2019 will be shown. Finally, all encountered evaluation approaches are published in an online freely accessible document (Multimedia Appendix 6). The document will be updated on a regular basis, and we encourage readers of this paper to email additions to the document if any evaluation approach is missing.

Conclusion

Improving the success of the development and implementation of DHTs is crucial to enhancing the transition to an efficient and future-proofed health care system. We developed a better understanding of the practice of the sequential evaluation study phases of the eHealth evaluation cycle and the use of the relative DHT evaluation approaches. In the *eHealth evaluation cycle*, most attention is paid to the pilot (feasibility) and effectiveness (impact) study phases. Whether the evaluation of the earlier study phases and the uptake (implementation) study phase indeed improve successful outcomes is yet to be evaluated. Surveys, interviews, and mixed methods dominated the earlier study phases of DHT research. Although the majority of the evaluation approaches still used an RCT design, the iterative nature of technology may be better suited to more novel assessment approaches. The most often explored DHTs were those focusing on treatment and communication. Interestingly, the specialties of psychiatry and mental health, cardiology, and neurology were more interested in DHT evaluation than others. This offers potential opportunities to focus on unaddressed specialties in the search for DHTs, which can provide novel ways to transform and improve our health care system. Finally, future research might benefit from tracking and sharing which technologies successfully proceed through all stages of the *eHealth evaluation cycle*.

XSL•FO
RenderX

Acknowledgments

The authors declare that no generative AI was used in any portion of the manuscript writing.

Data Availability

The data sets generated during and/or analyzed during this study are available in the online "Evaluation Approaches of Digital Health Technologies" [40].

Authors' Contributions

AR, HG, JM, and PS were involved in screening and data extraction. AR performed the data analyses, which were checked by HG and PS. PS and AR completed the write up of the paper. Review and editing were done by HG, JM, SZ, MPS, MJK, and NHC. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
World Health Organization (WHO) classification scheme.
[DOCX File , 280 KB - jmir_v26i1e50251_app1.docx]

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 84 KB - jmir_v26i1e50251_app2.docx]

Multimedia Appendix 3
PubMed search.
[DOCX File , 15 KB - jmir_v26i1e50251_app3.docx]

Multimedia Appendix 4
Data extraction sheet.
[XLSX File (Microsoft Excel File), 76 KB - jmir_v26i1e50251_app4.xlsx]

Multimedia Appendix 5
Results variables country, medical specialty and journals.
[DOCX File , 32 KB - jmir_v26i1e50251_app5.docx]

Multimedia Appendix 6
Complete list of evaluation approaches.
[DOCX File , 32 KB - jmir_v26i1e50251_app6.docx]

Multimedia Appendix 7
All performed cross tabulations.
[DOCX File , 1863 KB - jmir_v26i1e50251_app7.docx]

References

1. Ageing and health. World Health Organization. 2022 Oct 1. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2022-12-21]
2. Lau F, Kuziemsky C. What is eHealth? In: Lau F, Kuziemsky C, editors. Handbook of eHealth Evaluation: An Evidence-based Approach. Victoria, British Columbia: University of Victoria; Feb 27, 2017.
3. Classification of digital health interventions v1.0. World Health Organization. 2018 Jan 01. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2024-07-14]
4. Evidence Standards Framework for Digital Health Technologies Contents. National Institute for Health and Care Excellence. 2019 Mar. URL: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf [accessed 2022-12-21]

XSL•FO
RenderX

5. Unsworth H, Dillon B, Collinson L, Powell H, Salmon M, Oladapo T, et al. The NICE Evidence Standards Framework for digital health and care technologies - Developing and maintaining an innovative evidence framework with global impact. Digit Health 2021;7:20552076211018617 [FREE Full text] [doi: 10.1177/20552076211018617] [Medline: 34249371]
6. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000 Sep 16;321(7262):694-696 [FREE Full text] [doi: 10.1136/bmj.321.7262.694] [Medline: 10987780]
7. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
8. Bonten TN, Rauwerdink A, Wyatt JC, Kasteleyn MJ, Witkamp L, Riper H, EHealth Evaluation Research Group. Online guide for electronic health evaluation approaches: systematic scoping review and concept mapping study. J Med Internet Res 2020 Aug 12;22(8):e17774 [FREE Full text] [doi: 10.2196/17774] [Medline: 32784173]
9. eHealth methodology guide. Citrienprogramma e-health. URL: https://www.citrienfonds-ehealth.nl/e-health-toolkit/onderzoek/e-health-evaluation-methodology/overview-of-methods/ [accessed 2022-12-21]
10. Trochim W, Kane M. Concept mapping: an introduction to structured conceptualization in health care. Int J Qual Health Care 2005 Jun;17(3):187-191. [doi: 10.1093/intqhc/mzi038] [Medline: 15872026]
11. Safavi K, Cohen A, Ting D, Chaguturu S, Rowe J. Health systems as venture capital investors in digital health: 2011-2019. NPJ Digit Med 2020;3:103 [FREE Full text] [doi: 10.1038/s41746-020-00311-5] [Medline: 32802968]
12. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: A scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
13. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
14. Tomlinson M, Rotheram-Borus MJ, Swartz L, Tsai AC. Scaling up mHealth: where is the evidence? PLoS Med 2013;10(2):e1001382 [FREE Full text] [doi: 10.1371/journal.pmed.1001382] [Medline: 23424286]
15. Aminoff H, Meijer S. Context and complexity in telemedicine evaluation: work domain analysis in a surgical setting. JMIR Perioper Med 2021 Sep 16;4(2):e26580 [FREE Full text] [doi: 10.2196/26580] [Medline: 34528894]
16. Rayyan. URL: https://www.rayyan.ai/ [accessed 2022-12-21]
17. Cho H, Flynn G, Saylor M, Gradilla M, Schnall R. Use of the FITT framework to understand patients' experiences using a real-time medication monitoring pill bottle linked to a mobile-based HIV self-management app: A qualitative study. Int J Med Inform 2019 Nov;131:103949 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.08.009] [Medline: 31561192]
18. Dijkstra A, Heida A, van Rheenen PF. Exploring the challenges of implementing a web-based telemonitoring strategy for teenagers with inflammatory bowel disease: empirical case study. J Med Internet Res 2019 Mar 29;21(3):e11761 [FREE Full text] [doi: 10.2196/11761] [Medline: 30924785]
19. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
20. Kho SES, Lim SG, Hoi WH, Ng PL, Tan L, Kowitlawakul Y. The development of a diabetes application for patients with poorly controlled type 2 diabetes mellitus. Comput Inform Nurs 2019 Feb;37(2):99-106. [doi: 10.1097/CIN.0000000000000485] [Medline: 30407213]
21. Dial HR, Hinshelwood HA, Grasso SM, Hubbard HI, Gorno-Tempini M, Henry ML. Investigating the utility of teletherapy in individuals with primary progressive aphasia. Clin Interv Aging 2019;14:453-471 [FREE Full text] [doi: 10.2147/CIA.S178878] [Medline: 30880927]
22. Hoth AB, Shafer C, Dillon DB, Mayer R, Walton G, Ohl ME. Iowa TelePrEP: a public-health-partnered telehealth model for human immunodeficiency virus preexposure prophylaxis delivery in a rural state. Sex Transm Dis 2019 Aug;46(8):507-512. [doi: 10.1097/OLQ.0000000000001017] [Medline: 31295217]
23. WHO. Medical devices: Managing the Mismatch: An outcome of the Priority Medical Devices project 2010. World Health Organization. URL: https://iris.who.int/bitstream/handle/10665/44407/9789241564045_eng.pdf [accessed 2024-07-14]
24. Royle JK, Hughes A, Stephenson L, Landers D. Technology clinical trials: Turning innovation into patient benefit. Digit Health 2021;7:20552076211012131 [FREE Full text] [doi: 10.1177/20552076211012131] [Medline: 34017607]
25. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-A call for innovative evidence generation approaches. NPJ Digit Med 2020;3:110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
26. Hrynyschyn R, Prediger C, Stock C, Helmer SM. Evaluation methods applied to digital health interventions: what is being used beyond randomised controlled trials?-a scoping review. Int J Environ Res Public Health 2022 Apr 25;19(9):1 [FREE Full text] [doi: 10.3390/ijerph19095221] [Medline: 35564616]
27. Van Velsen L, Wentzel J, Van Gemert-Pijnen JE. Designing eHealth that matters via a multidisciplinary requirements development approach. JMIR Res Protoc 2013 Jun 24;2(1):e21 [FREE Full text] [doi: 10.2196/resprot.2547] [Medline: 23796508]

XSL•FO
RenderX

28. England I, Stewart D, Walker S. Information technology adoption in health care: when organisations and technology collide. Aust Health Rev 2000;23(3):176-185. [doi: 10.1071/ah000176] [Medline: 11186051]
29. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
30. Tossaint-Schoenmakers R, Kasteleyn MJ, Rauwerdink A, Chavannes N, Willems S, Talboom-Kamp EPWA. Development of a quality management model and self-assessment questionnaire for hybrid health care: concept mapping study. JMIR Form Res 2022 Jul 07;6(7):e38683 [FREE Full text] [doi: 10.2196/38683] [Medline: 35797097]
31. Esmaeilzadeh P, Sambasivan M, Kumar N, Nezakhati H. Adoption of technology applications in healthcare: the influence of attitude toward knowledge sharing on technology acceptance in a hospital. U- and E-Service, Science and Technology 2011:17-30 [FREE Full text] [doi: 10.1007/978-3-642-27210-3_3]
32. Scimago. URL: https://www.scimagojr.com/countryrank.php?area=3600&order=itp&ord=desc&year=2019 [accessed 2022-12-21]
33. HIMSS Annual European Digital Health Survey. Healthcare Information and Management Systems Society, Inc. URL: https://www.himss.org/resources/himss-annual-european-digital-health-survey [accessed 2022-12-21]
34. Kane CK, Gillis K. The use Of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
35. Basnet S. The feasibility of eHealth in mental health care. J Addict Res Ther 2014;05(04):1. [doi: 10.4172/2155-6105.1000205]
36. van Hemel NM, van der Wall EE. 8 October 1958, D Day for the implantable pacemaker. Neth Heart J 2008 Oct;16(Suppl 1):S3-S4 [FREE Full text] [Medline: 18958267]
37. Zwack C, Haghani M, Hollings M, Zhang L, Gauci S, Gallagher R, et al. The evolution of digital health technologies in cardiovascular disease research. NPJ Digit Med 2023 Jan 03;6(1):1 [FREE Full text] [doi: 10.1038/s41746-022-00734-2] [Medline: 36596833]
38. Newaz NT, Hanada E. The methods of fall detection: a literature review. Sensors (Basel) 2023 May 30;23(11):1 [FREE Full text] [doi: 10.3390/s23115212] [Medline: 37299939]
39. Projects. NeLL. URL: https://nell.eu/projecten [accessed 2022-12-21]
40. Database Evaluation Approaches of Digital Health Technologies. URL: https://docs.google.com/spreadsheets/d/18TRbIUKE5lqChFsrKli3yo4hhlzQ0oxG6qMRH6cUQIA/edit?usp=sharing [accessed 2024-07-14]

Abbreviations

DHT: digital health technology
ESF: Evidence Standards Framework
FDA: Food and Drug Administration
FITT: fit between individual, task, and technology
HIMSS: Healthcare Information and Management Systems Society
mHealth: mobile health
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
NICE: National Institute for Health and Care Excellence
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SDLC: systems development life cycle
SMART: sequential multiple assignment randomized trial
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 24.06.23; peer-reviewed by M Mirro, A Taylor, B Wolfe-Piggott, H Oh; comments to author 20.10.23; revised version received 22.12.23; accepted 25.06.24; published 28.08.24.

Please cite as:
Rauwerdink A, Spinazze P, Gijsbers H, Molendijk J, Zwolsman S, Schijven MP, Chavannes NH, Kasteleyn MJ
Approaches to Evaluating Digital Health Technologies: Scoping Review
J Med Internet Res 2024;26:e50251
URL: https://www.jmir.org/2024/1/e50251
doi:10.2196/50251
PMID:39196643

XSL•FO
RenderX

©Anneloek Rauwerdink, Pier Spinazze, Harm Gijsbers, Juul Molendijk, Sandra Zwolsman, Marlies P Schijven, Niels H Chavannes, Marise J Kasteleyn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Characteristics, Influence, Prevention, and Control Measures of the Mpox Infodemic: Scoping Review of Infodemiology Studies

XiangYu Yan[1*], PhD; Zhuo Li[2*], MM; Chunxia Cao[1*], PhD; Longxin Huang[3], BMed; Yongjie Li[4], PhD; Xiangbin Meng[5], PhD; Bo Zhang[6], PhD; Maohe Yu[7], DEng; Tian Huang[8], MPH; Jiancheng Chen[9], DEng; Wei Li[10], MM; Linhui Hao[8], BMed; Dongsheng Huang[11], MPH; Bin Yi[12], BMed; Ming Zhang[13], PhD; Shun Zha[8], BMed; Haijun Yang[14], PhD; Jian Yao[15], PhD; Pengjiang Qian[15], PhD; Chun Kai Leung[16,17], PhD; Haojun Fan[1*], PhD; Pei Jiang[18*], MM; Tiejun Shui[8*], MPH

[1]School of Disaster and Emergency Medicine, Tianjin University, Tianjin, China
[2]Department of Epidemiology and Biostatistics, School of Public Health, Jilin University, Changchun, China
[3]Department of Clinical Medicine, North Sichuan Medical College, Nanchong, China
[4]School of Basic Medical Sciences, Peking University, Beijing, China
[5]Department of Cardiology, The First Hospital of Hebei Medical University, Shijiazhuang, China
[6]School of Environmental Science and Engineering, Hainan University, Haikou, China
[7]Tianjin Center for Disease Control and Prevention, Tianjin, China
[8]Yunnan Center for Disease Control and Prevention, Kunming, China
[9]Xiamen Peiyang BCI & Smart Health Innovation Research Institution, Xiamen, China
[10]Yunnan Provincial Hospital of Traditional Chinese Medicine, Kunming, China
[11]Baoshan Prefecture Center for Disease Control and Prevention, Baoshan, China
[12]Lincang Prefecture Center for Disease Control and Prevention, Lincang, China
[13]Department of Radiation Oncology, The Third Affiliated Hospital of Kunming Medical University & Tumor Hospital of Yunnan Province, Kunming, China
[14]Yan'An Hospital of Kunming City, Kunming, China
[15]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[16]Department of Public and International Affairs, City University of Hong Kong, Hong Kong, China
[17]Fairbank Center for Chinese Studies, Harvard University, Cambridge, MA, United States
[18]Department of Public Health, North Sichuan Medical College, Nanchong, China
[*]these authors contributed equally

Corresponding Author:
Tiejun Shui, MPH
Yunnan Center for Disease Control and Prevention
Xianghe Street #1177
Chenggong District
Kunming, 650050
China
Phone: 86 13987165649
Email: shuitiejunynjkedu@163.com

Abstract

Background: The mpox pandemic has caused widespread public concern around the world. The spread of misinformation through the internet and social media could lead to an infodemic that poses challenges to mpox control.

Objective: This review aims to summarize mpox-related infodemiology studies to determine the characteristics, influence, prevention, and control measures of the mpox infodemic and propose prospects for future research.

Methods: The scoping review was conducted based on a structured 5-step methodological framework. A comprehensive search for mpox-related infodemiology studies was performed using PubMed, Web of Science, Embase, and Scopus, with searches completed by April 30, 2024. After study selection and data extraction, the main topics of the mpox infodemic were categorized and summarized in 4 aspects, including a trend analysis of online information search volume, content topics of mpox-related

XSL•FO
RenderX

online posts and comments, emotional and sentiment characteristics of online content, and prevention and control measures for the mpox infodemic.

Results: A total of 1607 articles were retrieved from the databases according to the keywords, and 61 studies were included in the final analysis. After the World Health Organization's declaration of an mpox public health emergency of international concern in July 2022, the number of related studies began growing rapidly. Google was the most widely used search engine platform (9/61, 15%), and Twitter was the most used social media app (32/61, 52%) for researchers. Researchers from 33 countries were concerned about mpox infodemic–related topics. Among them, the top 3 countries for article publication were the United States (27 studies), India (9 studies), and the United Kingdom (7 studies). Studies of online information search trends showed that mpox-related online search volume skyrocketed at the beginning of the mpox outbreak, especially when the World Health Organization provided important declarations. There was a large amount of misinformation with negative sentiment and discriminatory and hostile content against gay, bisexual, and other men who have sex with men. Given the characteristics of the mpox infodemic, the studies provided several positive prevention and control measures, including the timely and active publishing of professional, high-quality, and easy-to-understand information online; strengthening surveillance and early warning for the infodemic based on internet data; and taking measures to protect key populations from the harm of the mpox infodemic.

Conclusions: This comprehensive summary of evidence from previous mpox infodemiology studies is valuable for understanding the characteristics of the mpox infodemic and for formulating prevention and control measures. It is essential for researchers and policy makers to establish prediction and early warning approaches and targeted intervention methods for dealing with the mpox infodemic in the future.

(J Med Internet Res 2024;26:e54874) doi:10.2196/54874

KEYWORDS

mpox; infodemic; infodemiology; information search volume; content topic; digital health

Introduction

Background

Mpox (formerly known as monkeypox) is a zoonotic infectious disease caused by infection with the mpox virus (MPXV) [1]. Clinical manifestations primarily involve fever, skin rash, and lymph node enlargement. The rash is predominantly on the face and limbs, displaying a centrifugal distribution pattern [2]. The main hosts for the MPXV are rodents and primates [3]. Mpox can be transmitted from animal to human and from human to human [4]. Previous studies indicated a degree of cross-protection against the MPXV with the smallpox vaccine, and therefore, the unvaccinated population is generally susceptible [5]. The MPXV belongs to the genus *Orthopoxvirus*, a double-stranded linear DNA virus that is divided into a West African clade and a Congo Basin clade, which is faster spreading and more lethal [6,7]. The first human case of mpox was identified in the Democratic Republic of the Congo in 1970 [8]. Historically, mpox was mainly endemic in West and Central Africa. Reports of mpox outside Africa were rare and primarily involved imported cases or were associated with animals imported from Africa [9]. Since May 2022, numerous mpox cases without travel history to Africa have emerged in multiple countries outside Africa, resulting in international concern as an abrupt public health event [10]. A public health emergency of international concern (PHEIC) was proclaimed for the spreading mpox outbreak on July 23, 2022, by the World Health Organization (WHO) [11]. Most cases involved adults, with most transmission occurring among men who have sex with men (MSM), and the rash is more common in the genital and anal regions [12,13].

In the era of information explosion on the internet, emerging infectious diseases could not only affect human health through human-to-human transmission in the real world but also cause infodemics through the spread of misinformation and rumors on the internet [14]. With the quantity and speed of information rapidly growing, it is challenging for the public to distinguish the true health information and misinformation, which leads to the emergence of an infodemic [15,16]. The lessons we learned from the COVID-19 pandemic revealed the importance of infodemic prevention and control [17]. During the COVID-19 pandemic, incorrect information and rumors distorted the public's understanding of the pandemic, causing more anxiety and panic, affecting people's behavior and adherence to public health indications, which posed a significant challenge to disease control and public health. Incorrect recommendations for treatment methods, prevention, or behaviors could potentially be harmful and worsen the current pandemic [18]. Tedros Adhanom Ghebreyesus, WHO's director-general, has said that "We're not just fighting a pandemic; we're fighting an infodemic" [19].

Similar to the COVID-19 pandemic, the outbreak of mpox in 2022 led to widespread searches and sharing of mpox-related information on search engines and social media platforms [20,21]. Several studies based on social media, such as Twitter, Facebook, and TikTok, have found that expressions of negative sentiments, attitudes toward vaccines, homophobia and racism, and various comments about symptoms and prevention methods are widely discussed on social media [22-25]. This has sounded the alarm of the emergence of a new infodemic.

Infodemiology studies the distribution and determining factors of health information and misinformation on the internet and among populations, which has significant practical applications in the research of infodemic identification and control [26]. It involves analyzing, monitoring, and predicting certain diseases' outbreaks and epidemics and tracking health-related behaviors in populations and other areas based on online data [27]. Infodemiology provides an important perspective and support

XSL•FO
RenderX

for exploring the characteristics of infodemics and can contribute to decision-making in digital health to cope with the ongoing infodemic. At present, there is an urgent need to comprehensively assess the situation and impact of the mpox infodemic to offer recommendations for the prevention and control of mpox through online platforms. However, there is a lack of reviews on this particular topic.

Objective

The aim of this review is to comprehensively summarize mpox-related infodemiology studies in the following aspects and propose prospects for future research: (1) the characteristics of internet search activities and infodemic trends during mpox outbreaks, (2) the views and opinions of social media users and the psychological impact of the mpox infodemic, and (3) researchers' suggestions on prevention and control measures for the mpox infodemic.

Methods

Study Design and Literature Search

The structured 5-step methodological framework by Arksey and O'Malley [28] was used to form this scoping review, including (1) delineating the research query; (2) pinpointing pertinent studies; (3) selecting studies; (4) extracting and charting data; and (5) collating, synthesizing, and presenting the results.

A total of 4 electronic literature databases, including PubMed, Web of Science, Embase, and Scopus were searched comprehensively till April 30, 2024, to avoid missing important studies. The following four parts of keywords were combined as the entry terms when searching the 4 databases: (1) mpox-related keywords: "monkeypox" or "mpox"; (2) infodemic-related keywords: "infodemic" or "infodemiology" or "infosurveillance" or "infoveillance"; (3) website and search engine–related keywords: "internet" or "web" or "website" or "online" or "search engine" or "search volume"; (4) media-related keywords: "media" or "medium" or "social app" or "social application" or "mobile app" or "mobile application" or "software" or "blog" or "blogging" or "forum"; and (5) combinations of #Part 1 AND (#Part 2 OR #Part 3 OR #Part 4).

As these databases collectively included articles from representative preprint repositories, including arXiv, bioRxiv, medRxiv, and Research Square, the searching process could effectively capture cutting-edge studies and articles that had not been peer reviewed.

Study Selection

A study selection procedure was conducted after eliminating duplicates. The inclusion and exclusion criteria of literatures are shown in Textbox 1. A total of 2 reviewers (XY and ZL) screened the literature and documents independently. Disagreements were resolved by discussion with a third reviewer (PJ or L Huang).

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria of study selection.

Inclusion criteria

- Article type
  - Original research
- Language
  - Provide at least a clear abstract or summary in English
- Full text
  - The full text of the study could be found
- Content of the article
  - Focused on mpox
  - Analysis of internet information
  - Focused on online information trends analysis, content analysis, or prevention and control measures for the mpox infodemic

Exclusion criteria

- Article type
  - Nonoriginal research, including review (because the original studies in the review had been included in the study), commentary, and short conference abstract
- Language
  - Both the abstract or summary and main text were not in English
- Full text
  - Only had short abstract or summary, without full text
- Content of the article
  - Not relevant to mpox
  - Not relevant to internet information analysis
  - Only conducted online surveys through the internet, and the content was not relevant to the infodemic

Data Extraction

After a thorough reading of full text, the following information of the selected studies were extracted: (1) article title, (2) journal name, (3) authors and countries, (4) publication date, (5) search engine or social media, and (6) main findings. For articles that mainly focused on the online information search volume or content analysis of social media text, quality assessment was conducted based on the following four criteria: (1) specified the search engine or social media application; (2) clearly reported the content source of the online information, such as Google Trends, Baidu Index, posts, comments, tweets, and videos; (3) clearly reported the period and the amount of data acquired; and (4) reported the statistical analysis method and process. If the abovementioned information could not be fully identified in the article, we considered the article to be incomplete and of low quality, which would be excluded. A total of 2 reviewers (XY and ZL) extracted data and conducted quality assessment, respectively, and reconciled their data records and quality assessment results after this process. A third reviewer (TS) was consulted to make the final decision if there were disagreements.

Data Synthesis

The number of publications per month were calculated, and the chronological trend of publications were described. On the basis of the authors' countries, the geographical distribution of the publications was described using a map. According to the main findings extracted using the abovementioned process, the main focuses of the mpox infodemic were summarized and categorized into 4 aspects, including trends of online information search volume, content topics of mpox-related online posts and comments, emotional and sentiment characteristics of online content, and prevention and control measures for the mpox infodemic. Statistical analyses were performed using SPSS (version 21.0; IBM Corp). Mapping was performed with ArcGIS (version 10.0; ESIR).

Results

General Characteristics of the Included Studies

A total of 1607 articles were retrieved from the databases according to keywords. After the removal of duplicates and study selection based on exclusion criteria, 61 studies were

XSL•FO
RenderX

included in the final analysis (Figure 1). The first study reporting online social media's information of mpox was published in November 2019, which focused on the media reporting of mpox in Nigeria. After that, few studies focused on mpox-related online information or the infodemic until May 2022, when mpox cases emerged in multiple countries outside Africa. With WHO's declaration of an mpox PHEIC, the number of related studies began growing rapidly (Multimedia Appendix 1).

Figure 1. Study selection flowchart of mpox infodemiology studies by April 2024.

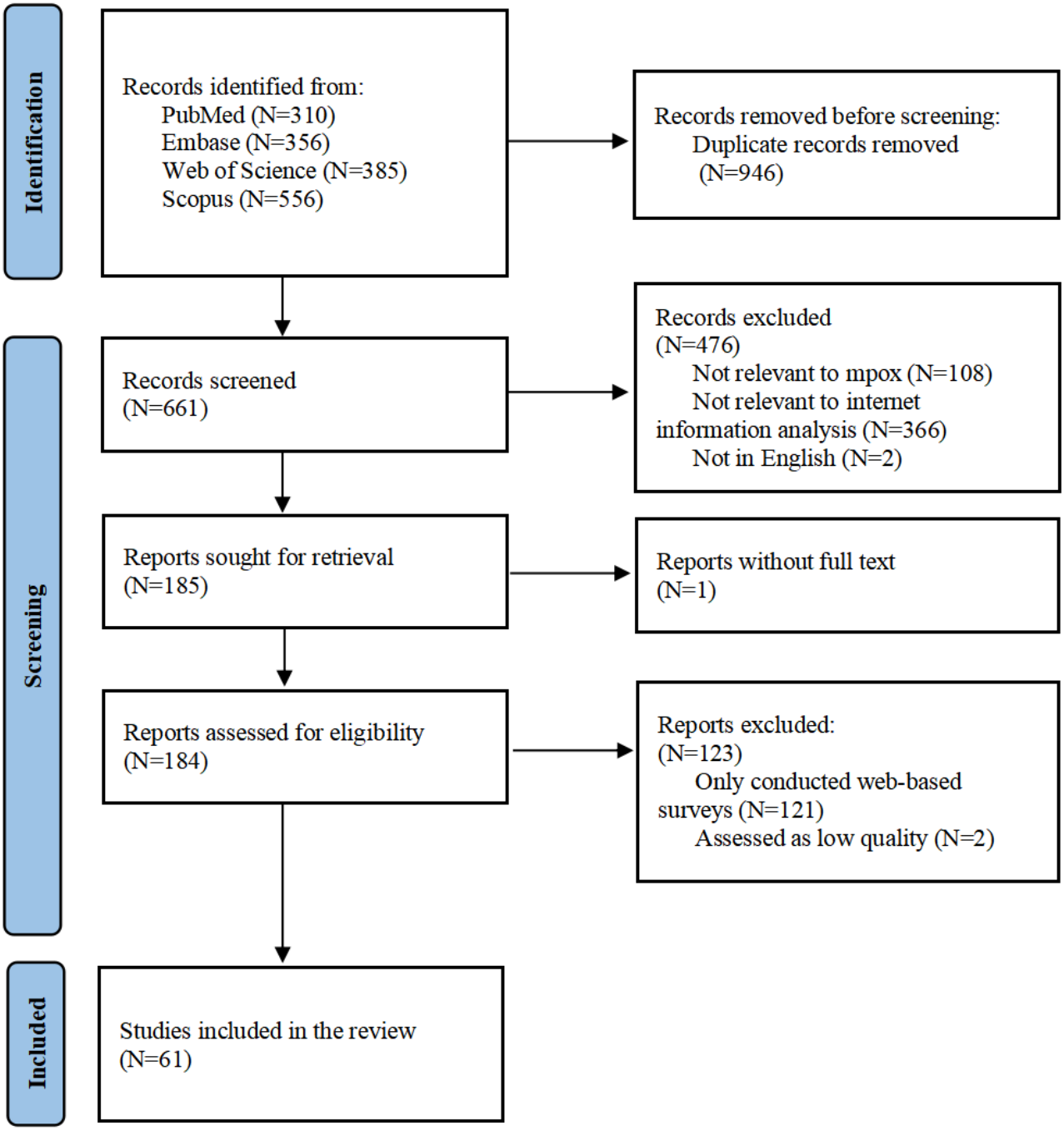

Some popular search engines and social media applications had become the main research platforms and tools for researchers, such as Google, Baidu, TikTok, Twitter, Facebook, and Reddit. Google was the most widely used search engine platform (9/61, 15% studies) for researchers to describe online information search volume. For social media applications, Twitter had become the most used platform for researchers (32/61, 52% studies; Multimedia Appendix 1).

Researchers from 33 countries were concerned about mpox infodemic–related topics. The top 3 countries that published a relatively large number of mpox infodemic articles were the United States (27 studies), India (9 studies), and the United Kingdom (7 studies; Figure 2).

XSL•FO
RenderX

Figure 2. Global geographical distribution of mpox infodemiology studies from November 2019 to April 2024.

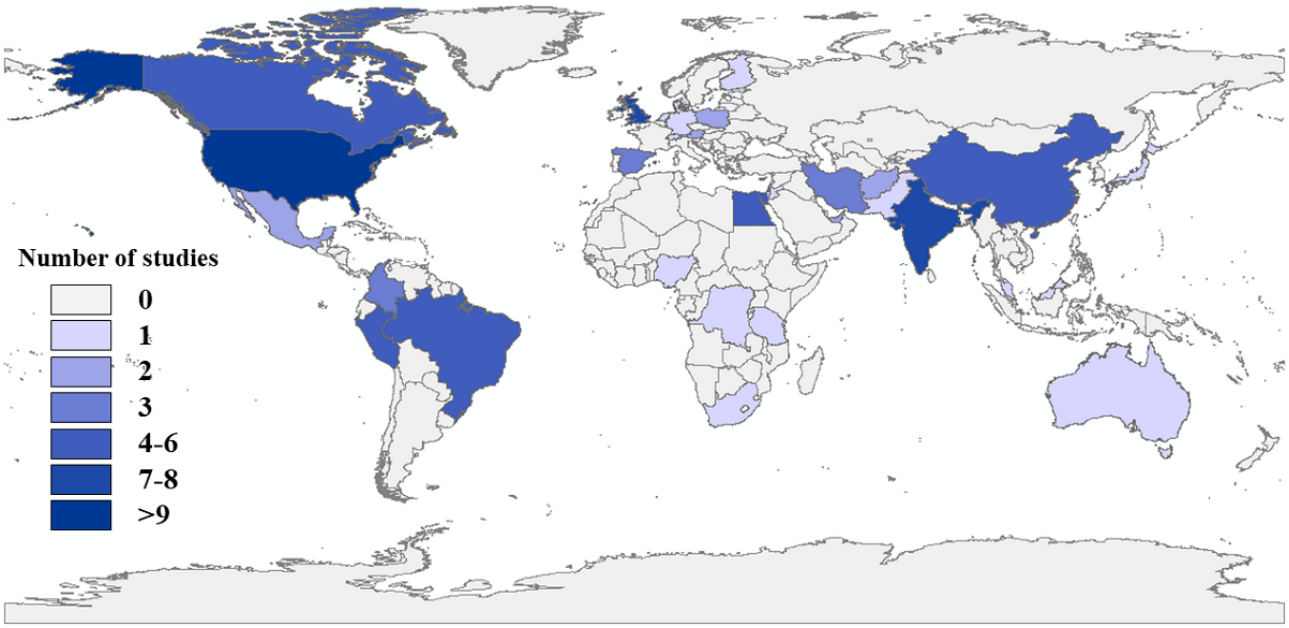

Trends in Online Information Search Volume

Public concern about mpox first showed up in web-based search activities. As an important web search engine, Google provides a useful tool for quantifying web-based search activities named Google Trends [29]. On the basis of Google Trends, Du et al [30] showed that after declaring the outbreak of mpox as a PHEIC, the global web-based search volume of mpox-related keywords increased by 42.85%, and the increase trends occurred among 62 countries or territories, especially for Brazil, the United States, and Indonesia. According to the comparison of Google Trends in mpox endemic countries (eg, Nigeria and Ghana) and nonendemic countries (eg, Spain and Germany) in 2022, Shepherd et al [31] observed a common pattern in most of these countries: online health information–seeking behavior was triggered by the first human mpox case and quickly peaked; however, this increasing trend only lasted for a short time. In addition, compared to the consistently low online search volume for COVID-19 information from mid-May to mid-August 2022, public interest in mpox showed significant fluctuations, with 2 peaks. These peaks occurred when the WHO declared that mpox could be transmitted through sex and when the WHO declared mpox a PHEIC [32]. However, the concerns of COVID-19 still exceeded mpox in most low- and middle-income countries in this period [32]. One study evaluated the search for specific ophthalmic symptoms of mpox in detail (eg, mpox eye), of which Canada, the United States, and the United Kingdom had the top 3 highest search rates [20].

Baidu is a commonly used search engine for Chinese internet users [33]. Similar to the Google Trends tool, the Baidu Index provides the weighted values of keywords search frequencies through the Baidu search engine, which can dynamically reflect the search trend of keywords and public attention [34]. On the basis of the Baidu Index, a study conducted in China showed that the searching trend of mpox in China fluctuated with key events, and the peaks occurred at the following time points: (1) WHO issued the mpox outbreak alert in May 2022; (2) first imported cases of mpox were reported in South Korea and Taiwan, China, in June 2022; (3) WHO declared the mpox outbreak as a PHEIC in July 2022; and (4) Hong Kong and Chongqing, China, reported the first imported cases of mpox [35].

In addition to internet search engines, several studies have revealed the volume of mpox information on other social media. Kato et al [36] investigated media coverage of mpox-related information on 14 media websites in Japan. The volume trends could be divided into 3 periods, including an initial phase between May 19 and May 22, 2022, when the outbreak occurred in the United States and Europe; a judgment period between June 15 and June 16 when the reports on the websites mainly discussed whether mpox should be declared a PHEIC by WHO; and a surge period between June 22 and June 29 when the outbreak began in Asia. Similar to the Google Trends studies, a global infodemiology study based on searches for sexually transmitted infections on YouTube showed a sudden increase began from the WHO's declaration for mpox's spread through sex, and another peak occurred after the WHO's declaration of the PHEIC [21]. In addition, through sampling the text of 105,913,149 tweets, Brian et al [37] detected the number of mpox-related and homophobic tweets from January to October 2022. This study showed a peak of both trends in July and indicated the polarizing trend in mpox-related homophobic and online hate against gay, bisexual, and other MSM (GBMSM) populations, which were catalyzed by the global mpox crisis [37].

Content Topics of Mpox-Related Online Posts and Comments

The contents of the public's concern on online social media were the main focus of the previous infodemiology studies. Public opinions on the 4 popular social media were widely discussed, including TikTok, Twitter, Facebook, and Reddit (Figure 3).

XSL•FO
RenderX

Figure 3. Content topics of mpox-related posts and comments on 4 popular social media from November 2019 to April 2024. LGBTQ+: lesbian, gay, bisexual, transgender, and queer; MSM: men who have sex with men; WHO: World Health Organization.

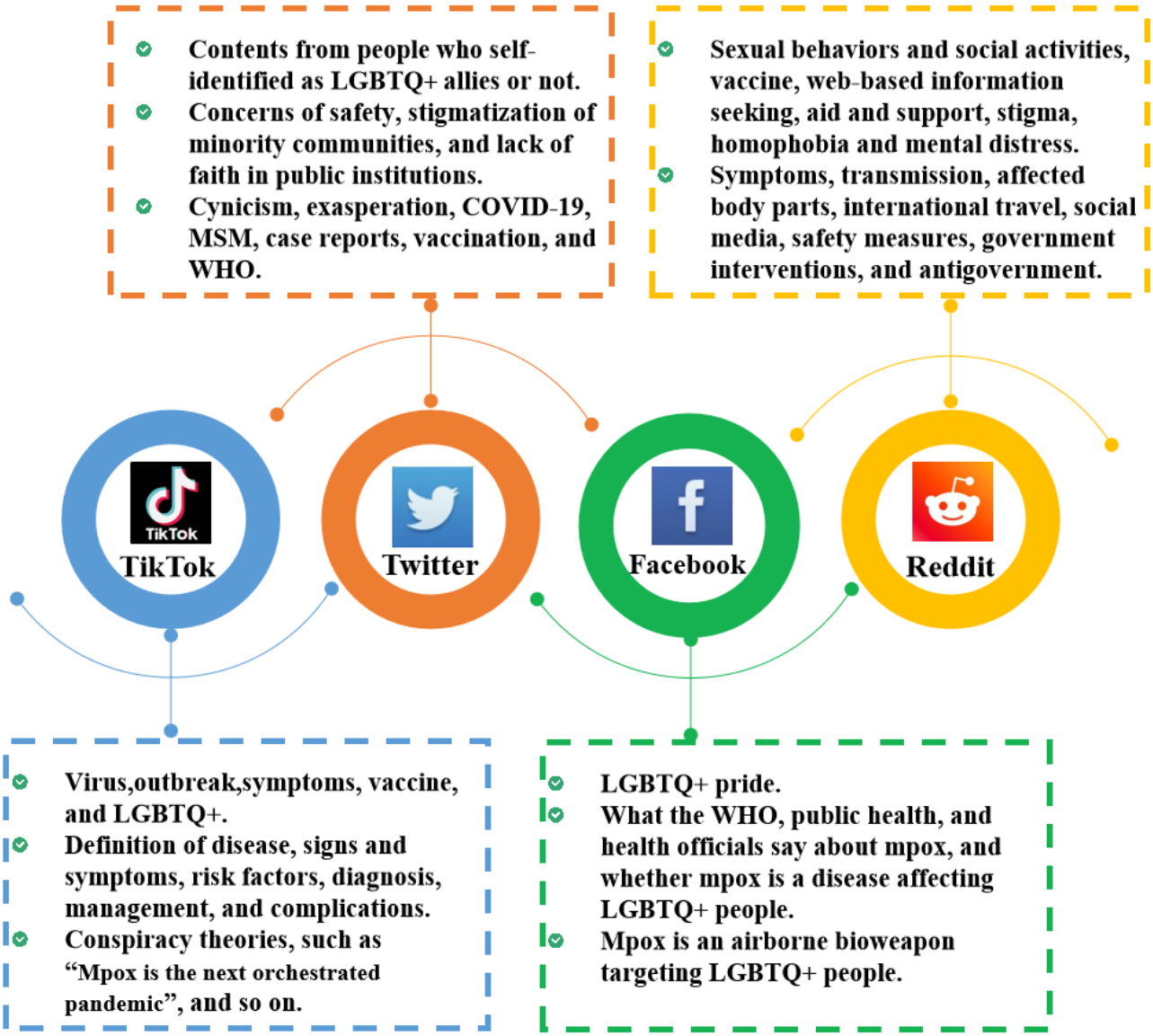

TikTok is a short video mobile platform that allows users to create videos lasting between 15 seconds and 10 minutes [38]. Zenone and Caulfield [39] identified 17.7% (153/864) videos with conspiracy theories or themes related to the mpox outbreak on TikTok, which was classified into 11 categories (eg, "Mpox is the next orchestrated pandemic," "Mpox was introduced to administer vaccines," and "Bill Gates is involved in the mpox outbreak"). By analyzing 100 English videos, Ji-Xu et al [25] found that the video types were mainly educational content (82%) and patient experience (72%); hashtag categories included virus (26%), outbreak (17%), symptoms (15%), vaccine (13%), and lesbian, gay, bisexual, transgender, and queer (LGBTQ+; 12%). For the quality of information, videos uploaded by doctors and science communicators had better information quality and fewer misleading statements compared with other users; however, they had a lower mean number of views and likes [25,40]. In addition, there was inconsistency in the proportion of videos containing misleading statements. Ortiz-Martínez Y and Ortiz-Martínez HM [41] found that only 1% of the most popular videos displayed misinformation, while Ji-Xu et al [25] found that 5 (14%) of the 36 physician videos and 21 (33%) of 64 nonphysician videos had misleading statements. Shi et al [40] studied 85 videos containing public-directed educational information about mpox and found that 33.33% of them were highlighted in clinical practice guidelines (ie, definition of disease, signs and symptoms, risk factors, diagnosis, management, and complications); however, these contents showed incomplete and inaccurate information frequently.

For posts on Twitter, the discussion of mpox-related contents was more extensive. Ortiz-Martínez et al [42] manually checked the top 100 tweets from late May 2022 and found that 52% of tweets contained misinformation and 20% were humorous or nonserious, with the most common misinformation claiming the mpox outbreak was a conspiracy. Edinger et al [43] analyzed 230,163 tweets from May 7, 2022, to July 23, 2022, and identified seven content clusters using the Sentence Bidirectional Encoder Representations from Transformers: (1) cynicism, (2) exasperation, (3) COVID-19, (4) MSM, (5) case reports, (6) vaccination, and (7) WHO. A study conducted by Ng et al [44] focused on topic modeling and manual thematic analysis by using a large data set of >350,000 tweets, which derived 5 topics

XSL•FO
RenderX

into 3 major themes related to the public discourse on the ongoing outbreaks, including concerns of safety, stigmatization of minority communities, and a general lack of faith in public institutions. Farahat et al [45] conducted topic modeling with latent Dirichlet allocation (LDA) using Twitter data collected from May 22, 2022, to August 5, 2022, and identified 12 topics, including the possibility of mpox being pandemic, mpox origin, gay community, spread, preventive measures, declaration of the outbreak, work, the need for mpox vaccinations, confirmed cases, mpox surfacing amid the COVID-19 pandemic, sexually transmitted, and beliefs about mpox. Another study showed that the uncertainty expressions about mpox outbreak turned to be less common among Twitter users in the week after WHO officially declared the mpox PHEIC than the week before WHO's declaration, which indicated that the official public health messaging could correct misperceptions and false narratives [46].

Most discussions on mpox were centered on the gay and other sexual minority groups. Many people have commented that mpox will be spread through their sexual behaviors. Homophobic and discriminatory information could cause stigma among this population. Cooper et al [22] examined 858,581 tweets from May 1, 2022, to July 23, 2022. Topic modeling by LDA identified unique themes among people who self-identified as LGBTQ+ allies or not [22]. Self-identified allies discussed 6 main topics, including informing prevention for mpox infection, geographic reporting of mpox cases, sharing of mpox vaccine information, lessons from mpox and other epidemics, checking hetero-privilege about mpox transmission, and pandemic exasperation about vaccination [22]. As for tweets from non–self-identifying allies or advocates, 9 topics were identified, including growing fear about mpox pandemic, geographic reporting of mpox cases, sexual transmission of mpox, learning from smallpox vaccine for mpox, anger about mpox, concerns about mpox as a global health emergency, global health response and mpox vaccine, declaration of mpox as a global health emergency, and surprise at mpox spread outside Africa [22]. Drawing data from Facebook and Twitter, Movahedi et al [47] used LDA to explore country-level variations in topics related to mpox and two-spirit, lesbian, gay, bisexual, transgender, queer or questioning, intersex, and asexual (2SLGBTQIAP+). They found that a total of 10 different topics were constructed from posts, and 8 were aimed at stigmatizing the 2SLGBTQIAP+ community. For posts on Reddit, Hong [48] used the method of inductive coding approach to analyze 755 Reddit posts that were popular among GBMSM and summarized four important themes: (1) changes in sexual behaviors and social activities; (2) mpox vaccine attitude, uptake, and hesitancy; (3) online information seeking and mutual aid and support; and (4) perceived and experienced stigma as well as homophobia and mental distress. Anoop and Sreelakshmi [49] used LDA and human-in-the-loop qualitative analysis to analyze 289,073 Reddit comments from June 1, 2022, to August 5, 2022. A total of 10 topics were obtained, which included mpox symptoms, mpox transmission, affected body parts, international travel, social media, safety measures, government interventions, mpox vaccination, antigovernment, and homophobia [49].

Emotional and Sentiment Characteristics of Online Content

Emotional and sentiment analysis was also a focus of previous studies. The following methods were used: the National Research Council emotion lexicon [45]; Valence Aware Dictionary and Entiment Reasoner methodology [23]; convolution neural network-long short-term memory network hybrid-based model [50]; certain software packages, such as TextBlob [45], pysentimiento package, and CamemBERT [47,51]; SentiStrength library; and text2emotion library [22].

Although the proportion of different emotional and sentiment types varied in different studies, negative ones accounted for a significant proportion. In a study focusing on mpox tweets in Latin America, which included 55,009 Spanish-language tweets collected between May 31, 2022, and October 31, 2022, the number of negative tweets published by users was much more than positive ones during the 23 weeks analyzed, representing 91.57% [52]. The first public data set of mpox tweets was presented and released by Thakur [23]. This study conducted sentiment analysis of the Twitter posts related to the mpox outbreak from May 7, 2022, to 23 July 2022, revealing a majority neutral sentiment (62.75%) toward mpox, followed by "negative" (20.91%) and "positive" (16.34%) sentiments [23]. Cooper et al [22] analyzed >850,000 English-language tweets from May 1, 2022, to July 23, 2022. They found that 46.5% of tweets were classified as neutral sentiment scores, 39% were negative, and 14.5% were positive. Farahat et al [45] analyzed a collection of >5000 mpox-related tweets from May 22, 2022, to August 5, 2022, and found that 48% of tweets were neutral sentiment, 15% were negative, and 37% were positive. Iparraguirre-Villanueva et al [50] analyzed 84,018 Spanish tweets obtained from Peru in September 2022 and found that 45.42% of people expressed neither positive nor negative opinions, 19.45% expressed negative and fearful feelings, while 35.13% of the opinions regarding the MPXV were positive.

Furthermore, several studies reported that the proportion of posts with positive sentiment has exceeded negative and neutral ones. Anoop and Sreelakshmi [49] analyzed 289,073 Reddit comments from June 1, 2022, to August 5, 2022, and found that nearly 35% of these comments were positive, followed by negative sentiment (34%) and neutral ones (31%). In addition, it seemed that vaccination and immunization-related posts showed more positive sentiment. Rajkhowa et al [51] analyzed 149,133 tweets related to mpox vaccination and immunization from May 1, 2022, to September 23, 2022, and found that the proportion of tweets with positive sentiment was 37.05%, higher than that of negative sentiment (29.89%) and neutral sentiment (33.06%).

A total of 2 articles reported the sentiment difference across different user characteristics. Cooper et al [22] compared sentiment characteristics between posts of self-identified LGBTQ+ allies and nonallies and indicated that most sentiments and emotions values, including happy, surprise, fear, and sadness, except for anger, showed significant differences between the 2 groups. Movahedi et al [47] reported a geographic difference in sentiment intensity among different countries by analyzing 125,424 Twitter and Facebook posts related to mpox

XSL•FO
RenderX

and the 2SLGBTQIAP+ community, of which English posts had higher negative and lower neutral and positive sentiment intensities than Spanish and French posts, and Spanish posts had higher negative and lower positive sentiment intensities than French ones.

In addition, Hong [48] analyzed 755 Reddit posts in several subreddits that were popular among GBMSM and showed obvious anxiety emotion about the outbreak and mpox-related stigma and discrimination among them. These posts also showed their desire and confidence in mpox vaccination in preventing themselves from infection and their concerns about getting vaccinated, such as their eligibility and concerns about vaccine side effects and effectiveness.

Another emotional expression of social media users was pandemic fatigue, which was characterized by people's reduced compliance with public health directives over time [53]. Through analyzing 1,484,042 social media posts from multiple social media platforms between June 23, 2022, and September 21, 2022, White et al [53] reported several emotional themes of pandemic fatigue, including anger, apathy or disinterest toward preventive measures, and overwhelmed with multiple or sustained emergencies.

Prevention and Control Measures for the Mpox Infodemic

Overview

As mentioned in the previous sections, due to the public's lack of understanding of mpox-related knowledge when the mpox outbreak occurred, false news and misinformation filled online platforms and social media, resulting in negative emotions such as fear and panic about mpox and discrimination against mpox key populations such as GBMSM. The mpox infodemic could worsen the actual epidemic situation and conceal the voices of health experts and trustful information [24]. A similar serious infodemic also occurred during the COVID-19 pandemic, especially in the early stage [54]. Therefore, it is essential to learn the lessons of the COVID-19 pandemic and take early prevention and control measures [32]. To prevent the mpox infodemic from becoming worse, previous infodemiology studies on mpox put forward several digital health suggestions, which could be summarized as the following 3 aspects. Detailed suggestions can be seen in Table 1.

XSL•FO
RenderX

Table 1. Main suggestions on mpox prevention and control in infodemiology studies from November 2019 to April 2024.

References	Social media and platforms	Main suggestions on mpox infodemic prevention and control
Zenone and Caulfield [39], 2022	TikTok	• Public health experts may consider greater attention to and investment in monitoring the online environment.
Ortiz-Martínez et al [42], 2022	Twitter	• Offering accurate and reliable information via social media platforms can help combat infodemics, misinformation, and rumors in the current mpox outbreak. • Real-time mpox surveillance on social media.
Ng et al [44], 2022	Twitter	• Authorities should actively counter misinformation, address stigmatizing portrayals, and rebuild the public's faith in institutions. • Primary care practitioners should also be equipped with the tools and knowledge to provide culturally sensitive interventions for LGBTQ+[a] populations. • Public health institutions could also engage respectable key opinion leaders in the LGBTQ+ community to provide the correct public messaging on mpox transmission, signs and symptoms, available help, and prophylaxis.
Martins-Filho [21], 2022	YouTube	• Implementing devices that monitor YouTube content in real time, preventing the spread of false and discriminatory news about mpox.
Martins-Filho et al [32], 2022	Google	• Lessons learned during the COVID-19 pandemic that information spread was driven by the interaction paradigm imposed by specific social media and by the interaction patterns of users engaged with the topic must be put into practice in mpox infodemic control.
Frost and Baldwin [55], 2022	Google	• The need for readability to be taken into account when publishing online resources to ensure simpler information is made available on this topic.
Farahat et al [24], 2022	Multiplatforms	• Raising public awareness and preventing stigma through cooperation with stakeholders. Social media platforms need to provide accurate official news. One health approach should be established and prioritized.
Hong [48], 2023	Reddit	• Understanding gay, bisexual, and men who have sex with men's concerns and prioritizing their needs. • Addressing the historical medical mistrust and promote vaccine education. • To disseminate vaccine information and knowledge through popular opinion leaders. • Using social media as a platform for intervention delivery.
Anoop and Sreelakshmi [49], 2023	Reddit	• The user-generated responses collected and analyzed from social media and news aggregators identify outbreak clusters and hate groups. • To discover public sentiment toward the measures taken by governments and public health organizations so that they can change their strategies and prioritize the action items.

XSL•FO
RenderX

References	Social media and platforms	Main suggestions on mpox infodemic prevention and control
Ji-Xu et al [25], 2023	TikTok	• Develop and validate specific quality assessment tools to evaluate video quality on social media. • Physician creators posting mpox content on TikTok should consider including a discussion of treatments and using hashtags selectively to increase user engagement.
Shi et al [40], 2023	TikTok	• Producing higher quality and richer videos. • Collaboration with influencers can promote health communication messaging, increase disease awareness, and enhance audience engagement.
Ortiz-Martínez Y and Ortiz-Martínez HM [41], 2023	TikTok	• Leveraging the influence of health care professionals and organizations on TikTok.
Basch et al [56], 2023	TikTok	• It is feasible to create engaging videos that mention prevention measures, such as vaccines, disinfection, and safe sex.
Cooper et al [22], 2023	Twitter	• Continued analysis of social media posts could allow public health officials to monitor public opinion on the next outbreak and spread relevant warnings and preventive measures to a wider audience sooner. • To determine groups of patients requiring targeted disease education based on the amount of misinformation disseminated. • Understanding how various demographic groups discuss a disease's symptoms or spread could allow targeted preventions such as vaccines to be mobilized to those populations quicker.
Edinger et al [43], 2023	Twitter	• Social media is monitored in real time to detect the dissemination of information that does not align with public health objectives. • Wide and timely dissemination of important public health information. • Gain insight into potential messaging campaigns that go beyond stigma and stereotypes. • Greater digital surveillance will ensure that information reaches people at a faster rate.
Keum et al [37], 2023	Twitter	• Research examining mpox-based stigma and online homophobia in international contexts is required.
Rajkhowa et al [51], 2023	Twitter	• Policy makers must prioritize disseminating adequate information to physicians and the general population to ensure optimal vaccination coverage. • Various interim guidance on vaccines and immunization for mpox can be adopted. • Strengthening the mental health status among the community. • Creating communication strategies or hosting workshops focused on building trust and effective communication regarding vaccines and immunization can help minimize vaccine-related fears
Movahedi et al [47], 2023	Twitter and Facebook	• Destigmatizing social media through protest, education, and engagement.
Du et al [30], 2023	Google	• The time-lag effect of global online search activity on daily new cases was significant. Online search activity could be used as an early indicator of the outbreak of mpox at the global level and in epidemic countries.

XSL•FO
RenderX

References	Social media and platforms	Main suggestions on mpox infodemic prevention and control
Shepherd et al [31], 2023	Google	• Publication of first mpox case promotes a surge in online health information seeking; the spike in searching behaviors occurs in a relatively short window. Health bodies and governmental organizations must work to ensure that credible and accurate information is searchable and in place as new disease cases or large news publication events are disclosed to the public. • For countries with low internet penetrance, the dissemination of public health–related information (eg, transmission prevention measures) must rely on different traditional infrastructures and media (eg, newspapers and radio).
Zheng et al [35], 2023	Baidu	• After the outbreak of an international epidemic, domestic epidemic information should be disclosed in real time, and core information, such as disease characteristics, transmission routes, and immunization prevention, should be compiled as soon as possible to address issues of public concern. • The content of online health education should be different at different stages of the epidemic. Before the outbreak of imported epidemics, attention should be paid to the health publicity of immunization prevention of the disease, the harm of the disease, and animals to avoid contact with. Later attention should be paid to the publicity and education of early symptoms and transmission routes of the disease. • The health education content should be concise and easy to understand and could be carried out in vivid and interesting ways, such as short videos and cartoons.
Kato et al [36], 2023	14 media websites in Japan	• It is essential to provide easy-to-understand information, enabling the media to disseminate basic information on infectious diseases and messages that lead to coping appraisals.
Dsouza et al [57], 2023	Twitter	• A system is needed to track the dissemination of unregulated media, especially social media. • Each nation needs to strengthen its health system and create comprehensive risk communication strategies and plans to prevent mpox stigma among the LGBTQ+ without jeopardizing their well-being.

[a]LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

Timely and Active Publishing of Professional, High-Quality, and Easy-to-Understand Information Online

Health professionals should be organized to publish authoritative, high-quality mpox information in collaboration with online media influencers in advance of widespread misinformation [25,40,41]. The health guidance information should include information on vaccines, early symptoms, modes of transmission, harm of the disease, diagnosis, and treatment recommendations [22,51,56]. Shepherd et al [31] suggested that it is necessary to grasp the short time window of the searching behaviors spike when the government declared the first local mpox case to publish high-quality health information. The content of online health education should be different at different stages of the epidemic: (1) before the outbreak of imported epidemics, attention should be paid to health publicity of immunization prevention of the disease, the harm of the disease, and animals to avoid contact with and (2) later attention should be paid to the publicity and education of early symptoms and transmission routes of the disease [35]. In addition, the professional information published through online platforms and social media should be easy to understand and have high readability, which can be carried out in vivid and interesting ways, such as short videos and cartoons [35,36,55]. For regions with low internet accessibility, more attention for health information publication needed to be paid to traditional media (eg, newspapers and radio) [31].

XSL•FO
RenderX

Surveillance and Early Warning of the Infodemic Based on Internet Data

Several studies have suggested that it is necessary to consider greater investment and tool development for real-time surveillance of mpox-related content on social media [21,39,42,43,57]. There are 2 advantages to enhance online surveillance of mpox information. On the one hand, real-time grasp of the situation of fake news and public's negative emotions on the internet is conducive to timely intervention and preventing misinformation from occupying the mainstream and forming social panic.

On the other hand, a previous study indicated that there was a significant time lag between the increase of online information search volume and the increase of daily new cases [30]. Du et al [30] has indicated that the higher mpox-related global online search volume, the more daily mpox cases would be observed thereafter, especially after a 7-day and 14-day time lag. Therefore, monitoring online mpox information can provide early indicators for early warning of the mpox epidemic in real world.

Protection for Key High-Risk Mpox Populations, Such as GBMSM

People who were most affected by the infodemic were the mpox high-risk population such as GBMSM. Online platforms and social media were rife with discrimination, stigma, homophobia, and even hatred against this subpopulation [37,49]. Therefore, to reduce the harm caused by cyber violence, it is essential to actively understand their concerns and needs and counter misinformation and stigmatizing portrayals through education and engagement on online platforms [47,48]. At the health system and policy-making level, comprehensive and well-developed risk communication strategies and plans are needed at the beginning of stigmatized diseases' epidemic to prevent stigmatization on the population considered vulnerable without blaming the affected [57]. In addition, it is of great value to turn online platforms and social media into fronts of interventions for high-risk populations. Popular opinion leaders in the communities of these populations can contribute to disseminating vaccine information and correct health knowledge [44,48].

Discussion

Principal Findings

This was an innovative study that reexplored the global mpox pandemic from the perspective of the infodemic and infodemiology. Before the current global pandemic, the mpox infodemic and its influence received little attention, which was due to mpox being locally endemic in West and Central Africa, with few reports outside Africa before 2022 [9]. West and Central African countries lag behind high-income countries in internet infrastructure and social media development, with relatively low population penetration of the internet and mobile devices. The relatively rapid growth trend of mpox infodemic studies after 2022 indicated that the mpox pandemic has indeed aroused widespread concern and discussion among people around the world and has brought anxiety and panic. This further demonstrates that the global infodemic resulting from a new pandemic of emerging infectious diseases cannot be ignored.

The influence of the mpox infodemic was not only reflected in the rapid increase of the relevant online information search volume and the expression of negative emotions on social media but also some conspiracy theories that contributed to hesitation to take the mpox vaccine. Some assertions led to the idea that mpox was a planned pandemic introduced for power, control, and money because vaccine manufacturers and governments could administer or mandate vaccines worldwide [39]. The wide spread of such views online undermined the credibility of government actions and people's confidence in mpox vaccines. In addition, high-risk populations such as GBMSM bore a huge impact of the infodemic. They had experienced stigma and discrimination brought by mpox and their identity as GBMSM, which could exacerbate their vaccine hesitancy and deter them from seeking health care services as early as possible [48]. The abovementioned influences of the mpox infodemic could increase the difficulty and reduce the efficiency of mpox prevention and control work.

Another important finding of this study was that we should attach great importance to the guiding role of information release by authoritative institutions and individuals on online public opinion. For instance, the WHO is the international flagship agency for infectious disease prevention and control. The WHO's declaration of the mpox pandemic as a PHEIC in July 2022 significantly increased the popularity of mpox-related information and discussion on the internet. The studies based on Google Trends and Baidu Index showed that the public's web-based searching behaviors of mpox-related information peaked after WHO's PHEIC declaration [32,35]. Similar phenomena were also shown in the increasing information volume on social media platforms such as YouTube and Twitter [21,37]. In addition, our study showed the increasing trends in the number of mpox infodemic–related articles after July 2022, which indicated that researchers were paying more attention to the mpox infodemic after the WHO's PHEIC declaration. For better prevention and control of the ongoing infodemic, it is vital to make full use of authoritative institutions and opinion leaders to lead the release of correct information on the internet. There were good experiences for us to learn from the control of the COVID-19–related infodemic. During the COVID-19 pandemic, the government's timely correction of misinformation and refutation of rumors could help curb people's panic buying [58]. The study by Wang et al [59] also indicated that the opinion leaders played a key role in leading discussions about COVID-19 vaccines on social media. Therefore, when a public health emergency occurs, prevention and control workers should promptly formulate an infodemic response plan and release professional and high-quality information via authoritative departments and social media opinion leaders.

Perspectives for Future Research

Given the characteristics and limitations of previous explorations of the mpox infodemic, it is essential to involve more digital health approaches to address the new crisis. Researchers can consider the following five aspects to further implement mpox-related infodemiology research and infodemic control.

XSL•FO
RenderX

First, it is essential to establish a comprehensive index system to evaluate the levels of the mpox infodemic based on online information search volume, the proportion of fake news and misinformation, and local mpox case numbers. This system will allow for a quantitative assessment of the risk of the infodemic.

Second, building and optimizing open-source large-scale online information databases can solve the technical bottleneck of data acquisition for researchers and improve research efficiency. Nia et al [60] and Thakur [23] have provided 2 large-scale Twitter data sets for researchers. Large data sets of other social media that are constantly updated are also urgently needed. Importantly, the data acquisition and database building process should comply with internet information and personal privacy security laws, regulations, and ethical requirements, which should also meet the information acquisition stipulations of social media. In addition, to adequately protect privacy and avoid information leakage, the administrators of the open-source databases should pay attention to the removal of personally identifiable information in the internet databases and enter into data security agreements with data users.

Third, it is necessary to carry out a series of comparative studies. It is valuable to compare the infodemic characteristics of countries with different mpox epidemic levels and cultural characteristics. Besides, the comparison of online information search volume and mpox-related content's topics on online platforms and media of different user groups (ie, Blued and Zank used mainly by MSM population and Twitter and TikTok used commonly by general populations) is helpful to formulate targeted online prevention and control strategies.

Fourth, mpox epidemic prediction and early warning models can be constructed based on information on online platforms and social media. This work can provide a new way to improve the timeliness of response for potential outbreaks of mpox.

Fifth, there is a need to develop targeted digital health intervention measures for the mpox infodemic and conduct effectiveness evaluation research for intervention measures. Previous infodemiology studies mainly focused on the searching trends and content of mpox-related information and only put forward suggestions and prospects for prevention and control measures. There is a lack of research to develop and evaluate the effectiveness of specific and feasible infodemic control measures. Therefore, this area of research needs urgent attention from researchers and policy makers.

Limitations

There were 2 limitations of this study. First, only articles written in English were included in this study. Articles published in other languages might be missing. Second, our search of selected databases might have excluded publications from nonindexed sources.

Conclusions

This study comprehensively summarized the characteristics, influence, prevention, and control measures of the global mpox infodemic. The number of related studies began to rise after the WHO's declaration of an mpox PHEIC, which involved researchers from 33 countries around the world. Popular search engines such as Google and social media such as Twitter became the main research platforms and tools for mpox-related infodemiology studies. The peaks of mpox-related online information search volume occurred at the beginning of outbreaks when the first mpox case in a country was reported and when the WHO provided important declarations, such as the PHEIC. However, the peaks always lasted for a short time window. Through analyzing the content topics of posts and comments on social media, a large amount of misinformation with negative sentiment and discriminatory and hostile content against key populations such as GBMSM was found, which posed significant challenges to the control of the mpox epidemic. Given the significant impact of the global mpox infodemic, positive prevention and control measures for the mpox infodemic are needed, including timely and active publishing of professional, high-quality, and easy-to-understand information online, strengthening surveillance and early warning of the infodemic based on internet data, and taking measures to protect key populations from harm from the mpox infodemic.

Acknowledgments

The authors would like to thank the Tianjin University Science and Technology Innovation Leading Talent Cultivation Project for their support (2024XQM-0024).

Data Availability

As this article is a scoping review, most of the data generated during this study are included and listed in this published article. A more detailed data set of the studies included in this review is available from the corresponding author on reasonable request.

Authors' Contributions

XY and TS had full access to all the article data in the study and take responsibility for the integrity of the data. XY and TS were responsible for the concept and design. TH, WL, L Huang, L Hao, and BY were responsible for the literature search and data collection. All authors were responsible for analysis and interpretation of data. XY, ZL, and PJ were responsible for drafting of the manuscript. CC, HF, TS, BZ, YL, and JY were responsible for the critical revision of the manuscript for important intellectual content. XM, MY, JC, DH, MZ, SZ, PQ, and CKL were responsible for administrative, technical, or material support. CC, PJ, and TS were responsible for supervision. All authors read and approved the final version of this manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Number of mpox infodemiology studies from November 2019 to April 2024 and the commonly used search engines or social media. PHEIC: public health emergency of international concern; WHO: World Health Organization.
[PNG File , 250 KB - jmir_v26i1e54874_app1.png]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 66 KB - jmir_v26i1e54874_app2.docx]

References

1. Ladnyj ID, Ziegler P, Kima E. A human infection caused by monkeypox virus in Basankusu territory, Democratic Republic of the Congo. Bull World Health Organ 1972;46(5):593-597 [FREE Full text] [Medline: 4340218]
2. Beer EM, Rao VB. A systematic review of the epidemiology of human monkeypox outbreaks and implications for outbreak strategy. PLoS Negl Trop Dis 2019 Oct 16;13(10):e0007791 [FREE Full text] [doi: 10.1371/journal.pntd.0007791] [Medline: 31618206]
3. Reed KD, Melski JW, Graham MB, Regnery RL, Sotir MJ, Wegner MV, et al. The detection of monkeypox in humans in the Western Hemisphere. N Engl J Med 2004 Jan 22;350(4):342-350. [doi: 10.1056/NEJMoa032299] [Medline: 14736926]
4. Ferdous J, Barek MA, Hossen MS, Bhowmik KK, Islam MS. A review on monkeypox virus outbreak: new challenge for world. Health Sci Rep 2023 Jan 25;6(1):e1007 [FREE Full text] [doi: 10.1002/hsr2.1007] [Medline: 36582623]
5. Petersen E, Kantele A, Koopmans M, Asogun D, Yinka-Ogunleye A, Ihekweazu C, et al. Human monkeypox: epidemiologic and clinical characteristics, diagnosis, and prevention. Infect Dis Clin North Am 2019 Dec;33(4):1027-1043 [FREE Full text] [doi: 10.1016/j.idc.2019.03.001] [Medline: 30981594]
6. Xiang Y, White A. Monkeypox virus emerges from the shadow of its more infamous cousin: family biology matters. Emerg Microbes Infect 2022 Dec 12;11(1):1768-1777 [FREE Full text] [doi: 10.1080/22221751.2022.2095309] [Medline: 35751396]
7. Bunge EM, Hoet B, Chen L, Lienert F, Weidenthaler H, Baer LR, et al. The changing epidemiology of human monkeypox-a potential threat? A systematic review. PLoS Negl Trop Dis 2022 Feb 11;16(2):e0010141 [FREE Full text] [doi: 10.1371/journal.pntd.0010141] [Medline: 35148313]
8. Klingelhöfer D, Braun M, Groneberg DA, Brüggmann D. Global mpox research in the light of the current outbreak: demands, drivers, and obstacles. Emerg Microbes Infect 2023 Dec 15;12(1):2210696 [FREE Full text] [doi: 10.1080/22221751.2023.2210696] [Medline: 37143355]
9. Durski KN, McCollum AM, Nakazawa Y, Petersen BW, Reynolds MG, Briand S, et al. Emergence of monkeypox - west and central Africa, 1970-2017. MMWR Morb Mortal Wkly Rep 2018 Mar 16;67(10):306-310 [FREE Full text] [doi: 10.15585/mmwr.mm6710a5] [Medline: 29543790]
10. Lum FM, Torres-Ruesta A, Tay MZ, Lin RT, Lye DC, Rénia L, et al. Monkeypox: disease epidemiology, host immunity and clinical interventions. Nat Rev Immunol 2022 Oct 05;22(10):597-613 [FREE Full text] [doi: 10.1038/s41577-022-00775-4] [Medline: 36064780]
11. Capobianchi MR, Di Caro A, Piubelli C, Mori A, Bisoffi Z, Castilletti C. Monkeypox 2022 outbreak in non-endemic countries: open questions relevant for public health, nonpharmacological intervention and literature review. Front Cell Infect Microbiol 2022;12:1005955 [FREE Full text] [doi: 10.3389/fcimb.2022.1005955] [Medline: 36204640]
12. Isidro J, Borges V, Pinto M, Sobral D, Santos JD, Nunes A, et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. Nat Med 2022 Aug 24;28(8):1569-1572 [FREE Full text] [doi: 10.1038/s41591-022-01907-y] [Medline: 35750157]
13. Li P, Li J, Ayada I, Avan A, Zheng Q, Peppelenbosch MP, et al. Clinical features, antiviral treatment, and patient outcomes: a systematic review and comparative analysis of the previous and the 2022 mpox outbreaks. J Infect Dis 2023 Aug 16;228(4):391-401 [FREE Full text] [doi: 10.1093/infdis/jiad034] [Medline: 36735342]
14. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb;395(10225):676. [doi: 10.1016/s0140-6736(20)30461-x]
15. Mackey T, Baur C, Eysenbach G. Advancing infodemiology in a digital intensive era. JMIR Infodemiology 2022 Feb 14;2(1):e37115 [FREE Full text] [doi: 10.2196/37115] [Medline: 37113802]
16. Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. J Med Internet Res 2020 Jun 29;22(6):e21820 [FREE Full text] [doi: 10.2196/21820] [Medline: 32589589]
17. Tangcharoensathien V, Calleja N, Nguyen T, Purnat T, D'Agostino M, Garcia-Saiso S, et al. Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. J Med Internet Res 2020 Jun 26;22(6):e19659 [FREE Full text] [doi: 10.2196/19659] [Medline: 32558655]
18. Hua J, Shaw R. Corona virus (COVID-19) "infodemic" and emerging issues through a data lens: the case of China. Int J Environ Res Public Health 2020 Mar 30;17(7):2309 [FREE Full text] [doi: 10.3390/ijerph17072309] [Medline: 32235433]

XSL•FO
RenderX

19. The Lancet Infectious Diseases. The COVID-19 infodemic. Lancet Infect Dis 2020 Aug;20(8):875 [FREE Full text] [doi: 10.1016/S1473-3099(20)30565-X] [Medline: 32687807]
20. Zhu A, Mahajan J, Oydanich MK, Luo MC, Khouri AS. Analysis of google trends and search results for ophthalmic symptoms of monkeypox. Int Ophthalmol 2023 Sep 18;43(9):3149-3155 [FREE Full text] [doi: 10.1007/s10792-023-02714-1] [Medline: 37071346]
21. Martins-Filho PR. Increase in interest in sexually transmitted infections on YouTube during the monkeypox outbreak in 2022: a global infodemiology study. Int J Surg 2022 Nov;107:106970 [FREE Full text] [doi: 10.1016/j.ijsu.2022.106970] [Medline: 36341759]
22. Cooper LN, Radunsky AP, Hanna JJ, Most ZM, Perl TM, Lehmann CU, et al. Analyzing an emerging pandemic on twitter: monkeypox. Open Forum Infect Dis 2023 Apr;10(4):ofad142 [FREE Full text] [doi: 10.1093/ofid/ofad142] [Medline: 37035497]
23. Thakur N. MonkeyPox2022Tweets: a large-scale Twitter dataset on the 2022 monkeypox outbreak, findings from analysis of tweets, and open research questions. Infect Dis Rep 2022 Nov 14;14(6):855-883 [FREE Full text] [doi: 10.3390/idr14060087] [Medline: 36412745]
24. Farahat RA, Head MG, Tharwat S, Alabdallat Y, Essar MY, Abdelazeem B, et al. Infodemic and the fear of monkeypox: call for action. Trop Med Health 2022 Sep 05;50(1):63 [FREE Full text] [doi: 10.1186/s41182-022-00459-8] [Medline: 36058955]
25. Ji-Xu A, Htet KZ, Leslie KS. Monkeypox content on TikTok: cross-sectional analysis. J Med Internet Res 2023 Jan 17;25:e44697 [FREE Full text] [doi: 10.2196/44697] [Medline: 36649057]
26. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
27. Mavragani A. Infodemiology and infoveillance: scoping review. J Med Internet Res 2020 Apr 28;22(4):e16206 [FREE Full text] [doi: 10.2196/16206] [Medline: 32310818]
28. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
29. Google Trends. Google. URL: https://trends.google.com/home.Accessed [accessed 2023-10-20]
30. Du M, Qin C, Yan W, Liu Q, Wang Y, Zhu L, et al. Trends in online search activity and the correlation with daily new cases of monkeypox among 102 countries or territories. Int J Environ Res Public Health 2023 Feb 15;20(4):3395 [FREE Full text] [doi: 10.3390/ijerph20043395] [Medline: 36834089]
31. Shepherd T, Robinson M, Mallen C. Online health information seeking for mpox in endemic and nonendemic countries: Google Trends study. JMIR Form Res 2023 Apr 13;7:e42710 [FREE Full text] [doi: 10.2196/42710] [Medline: 37052999]
32. Martins-Filho PR, de Souza Araújo AA, Quintans-Júnior LJ. Global online public interest in monkeypox compared with COVID-19: Google trends in 2022. J Travel Med 2022 Dec 27;29(8):29. [doi: 10.1093/jtm/taac104] [Medline: 36130219]
33. 2019 China internet users search engine usage research report. China Internet Network Information Center. 2019. URL: https://www.cnnic.net.cn/NMediaFile/old_attach/P020191025506904765613.pdf [accessed 2023-10-20]
34. AI leads the data revolution. Baidu Index. URL: https://index.baidu.com/v2/index.html#/ [accessed 2023-10-20]
35. Zheng J, Yuan Y, Li L. Analysis of online public opinion on global monkeypoxe pidemic in 2022 based on Baidu Index. J Mod Med Health 2023;39:1626-1629 [FREE Full text]
36. Kato M, Yoshimatsu F, Saito T. Trends in media coverage during the monkeypox outbreak: content analysis. J Med Internet Res 2023 Jun 19;25:e45787 [FREE Full text] [doi: 10.2196/45787] [Medline: 37335596]
37. Keum BT, Hong C, Beikzadeh M, Cascalheira CJ, Holloway IW. Mpox stigma, online homophobia, and the mental health of gay, bisexual, and other men who have sex with men. LGBT Health 2023 Jul 01;10(5):408-410 [FREE Full text] [doi: 10.1089/lgbt.2022.0281] [Medline: 36946657]
38. Malik A. TikTok expands max video length to 10 minutes, up from 3 minutes. TechCrunch. URL: https://techcrunch.com/2022/02/28/tiktok-expands-max-video-length-to-10-minutes-up-from-3-minutes/.Accessed [accessed 2023-10-20]
39. Zenone M, Caulfield T. Using data from a short video social media platform to identify emergent monkeypox conspiracy theories. JAMA Netw Open 2022 Oct 03;5(10):e2236993 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.36993] [Medline: 36255729]
40. Shi A, El Haddad J, Cai P, Song S, Wang YJ, Liu Q, et al. Mpox (monkeypox) information on TikTok: analysis of quality and audience engagement. BMJ Glob Health 2023 Mar 14;8(3):e011138 [FREE Full text] [doi: 10.1136/bmjgh-2022-011138] [Medline: 36918216]
41. Ortiz-Martínez Y, Ortiz-Martínez HM. TikTok and its importance in monkeypox public health engagement. J Adolesc Health 2023 Feb;72(2):312 [FREE Full text] [doi: 10.1016/j.jadohealth.2022.10.016] [Medline: 36604012]
42. Ortiz-Martínez Y, Sarmiento J, Bonilla-Aldana DK, Rodríguez-Morales AJ. Monkeypox goes viral: measuring the misinformation outbreak on Twitter. J Infect Dev Ctries 2022 Jul 28;16(7):1218-1220 [FREE Full text] [doi: 10.3855/jidc.16907] [Medline: 35905027]

XSL•FO
RenderX

43. Edinger A, Valdez D, Walsh-Buhi E, Trueblood JS, Lorenzo-Luaces L, Rutter LA, et al. Misinformation and public health messaging in the early stages of the mpox outbreak: mapping the Twitter narrative with deep learning. J Med Internet Res 2023 Jun 06;25:e43841 [FREE Full text] [doi: 10.2196/43841] [Medline: 37163694]
44. Ng QX, Yau CE, Lim YL, Wong LK, Liew TM. Public sentiment on the global outbreak of monkeypox: an unsupervised machine learning analysis of 352,182 twitter posts. Public Health 2022 Dec;213:1-4 [FREE Full text] [doi: 10.1016/j.puhe.2022.09.008] [Medline: 36308872]
45. Farahat RA, Yassin MA, Al-Tawfiq JA, Bejan CA, Abdelazeem B. Public perspectives of monkeypox in Twitter: a social media analysis using machine learning. New Microbes New Infect 2022 Nov;49:101053 [FREE Full text] [doi: 10.1016/j.nmni.2022.101053] [Medline: 36440091]
46. McKinley CJ, Choi JA, Luo Y. Expressing uncertainty and risk about the mpox outbreak: a textual analysis of Twitter messaging. Commun Stud 2024 Feb 09;75(3):283-301. [doi: 10.1080/10510974.2024.2311433]
47. Movahedi Nia Z, Bragazzi N, Asgary A, Orbinski J, Wu J, Kong J. Mpox panic, infodemic, and stigmatization of the two-spirit, lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual community: geospatial analysis, topic modeling, and sentiment analysis of a large, multilingual social media database. J Med Internet Res 2023 May 01;25:e45108 [FREE Full text] [doi: 10.2196/45108] [Medline: 37126377]
48. Hong C. Mpox on Reddit: a thematic analysis of online posts on mpox on a social media platform among key populations. J Urban Health 2023 Dec 14;100(6):1264-1273 [FREE Full text] [doi: 10.1007/s11524-023-00773-4] [Medline: 37580545]
49. Anoop VS, Sreelakshmi S. Public discourse and sentiment during Mpox outbreak: an analysis using natural language processing. Public Health 2023 May;218:114-120. [doi: 10.1016/j.puhe.2023.02.018] [Medline: 37019026]
50. Iparraguirre-Villanueva O, Alvarez-Risco A, Herrera Salazar JL, Beltozar-Clemente S, Zapata-Paulini J, Yáñez JA, et al. The public health contribution of sentiment analysis of monkeypox tweets to detect polarities using the CNN-LSTM Model. Vaccines (Basel) 2023 Jan 31;11(2):312 [FREE Full text] [doi: 10.3390/vaccines11020312] [Medline: 36851190]
51. Rajkhowa P, Dsouza VS, Kharel R, Cauvery K, Mallya BR, Raksha DS, et al. Factors influencing monkeypox vaccination: a cue to policy implementation. J Epidemiol Glob Health 2023 Jun 29;13(2):226-238 [FREE Full text] [doi: 10.1007/s44197-023-00100-9] [Medline: 37119512]
52. Chire-Saire J, Pineda-Briseño A, Oblitas-Cruz J. Sentiment analysis of monkeypox tweets in Latin America. In: Proceedings of the 6th International Conference on Applied Machine Learning and Data Analytics. 2023 Presented at: AMLDA '23; November 9-10, 2023; Lübeck, Germany p. 230-245 URL: https://link.springer.com/chapter/10.1007/978-3-031-55486-5_17 [doi: 10.1007/978-3-031-55486-5_17]
53. White BK, Ishizumi A, Lavery L, Wright A, Foley T, O'Neill R, et al. Expressions of pandemic fatigue on digital platforms: a thematic analysis of sentiment and narratives for infodemic insights. BMC Public Health 2024 Mar 05;24(1):705 [FREE Full text] [doi: 10.1186/s12889-024-17718-4] [Medline: 38443914]
54. Emery N, Dugerdil A, Flahault A. Infodemics in an era of pandemics. Int J Public Health 2022 Aug 17;67:1605209 [FREE Full text] [doi: 10.3389/ijph.2022.1605209] [Medline: 36059583]
55. Frost JC, Baldwin AJ. Readability of online monkeypox patient education materials: improved recognition of health literacy is needed for dissemination of infectious disease information. Infect Dis Health 2023 May;28(2):88-94 [FREE Full text] [doi: 10.1016/j.idh.2022.11.002] [Medline: 36564245]
56. Basch CH, Yalamanchili B, Fera J, Narayanan S. Most liked #monkeypox videos on TikTok: implications for infection control. Am J Infect Control 2024 Jan;52(1):123-124. [doi: 10.1016/j.ajic.2023.06.021] [Medline: 37422067]
57. Dsouza VS, Rajkhowa P, Mallya BR, Raksha D, Mrinalini V, Cauvery K, et al. A sentiment and content analysis of tweets on monkeypox stigma among the LGBTQ+ community: a cue to risk communication plan. Dialogues Health 2023 Dec;2:100095 [FREE Full text] [doi: 10.1016/j.dialog.2022.100095] [Medline: 36573228]
58. Motamed-Jahromi M, Meshkani Z, Mosavi-Negad SM, Momenabadi V, Ahmadzadeh MS. Factors affecting panic buying during COVID-19 outbreak and strategies: a scoping review. Iran J Public Health 2021 Dec 06;50(12):2473-2485 [FREE Full text] [doi: 10.18502/ijph.v50i12.7929] [Medline: 36317022]
59. Wang D, Zhou Y, Ma F. Opinion leaders and structural hole spanners influencing echo chambers in discussions about COVID-19 vaccines on social media in China: network analysis. J Med Internet Res 2022 Nov 18;24(11):e40701 [FREE Full text] [doi: 10.2196/40701] [Medline: 36367965]
60. Nia ZM, Bragazzi NL, Wu J, Kong JD. A Twitter dataset for monkeypox, May 2022. Data Brief 2023 Jun;48:109118 [FREE Full text] [doi: 10.1016/j.dib.2023.109118] [Medline: 37081848]

Abbreviations

2SLGBTQIAP+: two-spirit, lesbian, gay, bisexual, transgender, queer or questioning, intersex, and asexual
GBMSM: gay, bisexual, and other men who have sex with men
LDA: latent Dirichlet allocation
LGBTQ+: lesbian, gay, bisexual, transgender, and queer
MPXV: mpox virus
MSM: men who have sex with men

XSL•FO
RenderX

PHEIC: public health emergency of international concern
WHO: World Health Organization

Edited by A Mavragani; submitted 26.11.23; peer-reviewed by P Velmovitsky, H Peng, F Medina; comments to author 28.02.24; revised version received 30.04.24; accepted 01.07.24; published 30.08.24.

Please cite as:
Yan X, Li Z, Cao C, Huang L, Li Y, Meng X, Zhang B, Yu M, Huang T, Chen J, Li W, Hao L, Huang D, Yi B, Zhang M, Zha S, Yang H, Yao J, Qian P, Leung CK, Fan H, Jiang P, Shui T
Characteristics, Influence, Prevention, and Control Measures of the Mpox Infodemic: Scoping Review of Infodemiology Studies
J Med Internet Res 2024;26:e54874
URL: https://www.jmir.org/2024/1/e54874
doi:10.2196/54874
PMID:

©XiangYu Yan, Zhuo Li, Chunxia Cao, Longxin Huang, Yongjie Li, Xiangbin Meng, Bo Zhang, Maohe Yu, Tian Huang, Jiancheng Chen, Wei Li, Linhui Hao, Dongsheng Huang, Bin Yi, Ming Zhang, Shun Zha, Haijun Yang, Jian Yao, Pengjiang Qian, Chun Kai Leung, Haojun Fan, Pei Jiang, Tiejun Shui. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The SoCAP (Social Communication, Affiliation, and Presence) Taxonomy of Social Features: Scoping Review of Commercially Available eHealth Apps

Ian Kwok[1], PhD; Melanie Freedman[1], MSc; Lisa Kamsickas[2], MA; Emily G Lattie[1], PhD; Dershung Yang[3], PhD; Judith Tedlie Moskowitz[1], MPH, PhD

[1]Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[2]Dana Incorporated, Novi, MI, United States
[3]BrightOutcome Inc, Buffalo Grove, IL, United States

Corresponding Author:
Ian Kwok, PhD
Feinberg School of Medicine
Northwestern University
420 E Superior St
Chicago, IL, 60611
United States
Phone: 1 323 364 1160
Email: iankwok@u.northwestern.edu

Abstract

Background: eHealth interventions have proven to be valuable resources for users with diverse mental and behavioral health concerns. As these technologies continue to proliferate, both academic researchers and commercial app creators are leveraging the use of features that foster a sense of social connection on these digital platforms. Yet, the literature often insufficiently represents the functionality of these key social features, resulting in a lack of understanding of how they are being implemented.

Objective: This study aimed to conduct a methodical review of commercially available eHealth apps to establish the SoCAP (social communication, affiliation, and presence) taxonomy of social features in eHealth apps. Our goal was to examine what types of social features are being used in eHealth apps and how they are implemented.

Methods: A scoping review of commercially available eHealth apps was conducted to develop a taxonomy of social features. First, a shortlist of the 20 highest-rated eHealth apps was derived from One Mind PsyberGuide, a nonprofit organization with trained researchers who rate apps based on their (1) credibility, (2) user experience, and (3) transparency. Next, both mobile- and web-based versions of each app were double-coded by 2 trained raters to derive a list of social features. Subsequently, the social features were organized by category and tested on other apps to ensure their completeness.

Results: Four main categories of social features emerged: (1) communication features (videoconferencing, discussion boards, etc), (2) social presence features (chatbots, reminders, etc), (3) affiliation and identity features (avatars, profiles, etc), and (4) other social integrations (social network and other app integrations). Our review shows that eHealth apps frequently use resource-intensive interactions (eg, videoconferencing with a clinician and phone calls from a facilitator), which may be helpful for participants with high support needs. Furthermore, among commercially available eHealth apps, there is a strong reliance on automated features (eg, avatars, personalized multimedia, and tailored content) that enhance a sense of social presence without requiring a high level of input from a clinician or staff member.

Conclusions: The SoCAP taxonomy includes a comprehensive list of social features and brief descriptions of how these features work. This classification system will provide academic and commercial eHealth app creators with an understanding of the various social features that are commonly implemented, which will allow them to apply these features to enhance their own apps. Future research may include comparing the synergistic effects of various combinations of these social features.

(J Med Internet Res 2024;26:e49714) doi:10.2196/49714

XSL•FO
RenderX

KEYWORDS

eHealth; digital interventions; social features; taxonomy; computer-human interaction; social connection; engagement; eHealth apps; intervention; mental health; behavioral health; mobile app

Introduction

Background

Research on eHealth interventions has demonstrated that technology-driven resources such as apps and web-based platforms can provide effective and easily accessible support to address both mental and behavioral health concerns [1-3]. These technologies have been woven into our daily lives, with the average person spending an estimated 3.5 hours a day on their cell phones [4], which counts toward a daily total of 6.5 hours online [5], as the demand for mobile- and web-based interventions continues to grow. A recent estimate suggests that there may be upwards of 350,000 eHealth apps available for download across app stores [6], with the current number expected to be significantly higher given the growing popularity of such apps in recent years.

However, the potential of these digital health interventions is beset by a lack of engagement among users [3,7-9]. A recent review of 10 self-guided eHealth interventions found that up to 79% of participants did not meet the criteria for minimal use (ie, completing 1 module or assessment or using the intervention at least once), with an average dropout rate of 43% [10]. As a result, improving participant engagement is a vital consideration in the implementation of such technologies.

Several eHealth implementation theories have pointed to the importance of social connection in enhancing adoption and engagement [11]. The normalization process theory framework suggests that the way users interact online can foster the development of new social norms that encourage the utilization of eHealth platforms [12]. Actor-network theory suggests that individual users act on their sense of agency (ie, their ability to choose how to engage virtually) and interact with both non-human (eg, user interfaces, multimedia, and content) and human components (eg, other users and moderators) of a system, which thereby creates meaning and coherence through such ongoing social interactions [13]. Therefore, eHealth technologies with feature sets that cultivate these social processes may enhance participant engagement.

Currently, behavioral or psychological eHealth interventions use various social features to engage their users. The literature on eHealth interventions documents a broad range of features.

Synchronous Communication Features

Real-time communication mediums (ie, voice and video calls, “live” chats, etc) are widely used in eHealth interventions to connect participants with peers, facilitators, and clinicians. An example used videoconferencing to enhance social interaction for socially isolated older adults, by having a trained facilitator meet with participants for 4 sessions a week over the course of 6 months [14]. Videoconferencing interventions may also use a group format, which is structured similarly to in-person support groups. For example, the TeleGAIN program comprised a 12-week videoconferencing group intervention for individuals with aphasia that was led by speech-language pathologists [15].

Asynchronous Communication Features

In contrast, asynchronous communication features, such as discussion boards or forums, allow for interaction among participants even if they are not present online at the same time [16,17]. For example, one study used a discussion board and recruited individuals with a history of mental disorders to moderate and offer peer support. This method encouraged participants to contribute to the discussion boards and to receive support from moderators who had the lived experience of coping with similar challenges [18]. Such features may be particularly helpful for connecting participants who are dispersed across different time zones or who may not have the ability to commit to a fixed meeting schedule.

Social Presence Features

Other types of social features do not require the need for direct communication, while enhancing participants’ sense of *social presence*, defined as the perception that there are other individuals present in a digital environment [19-21]. Additionally, emerging research is being conducted on text- or voice-based conversational agents (ie, chatbots), which can deliver intervention content through language-based user interaction [22-25].

Affiliation and Identity Features

Other features enhance a participant’s ability to represent themselves or affirm their affiliations with others, setting the stage for more meaningful online or offline interactions. For example, the use of *avatars* may enhance a sense of identity and affiliation. In a study of a digital dyadic intervention for family members of veterans with posttraumatic stress disorder, the platform allowed a nominated family member participant to interact with a customized avatar that represented their veteran [26]. Therefore, these virtual representations of identity may help users perceive themselves and others in novel ways, leading to increased engagement.

Social Integrations

Additionally, researchers can leverage the broad user base and rich social feature sets of mature technologies like Instagram, Facebook, and other similar platforms through integration with their own platforms [27].

Taken together, the literature demonstrates how such social features can play a central role in eHealth interventions. The majority of the published research pertains to eHealth resources that have been developed by academic researchers from higher education or medical institutions, which may be distinguished from commercially available apps, which are typically profit-driven enterprises. The findings suggest a tendency toward resource-intensive social features that involve live communication among providers, clinicians, and peers. However, the cost of implementation should be weighed against

XSL•FO
RenderX

the scalability of these interventions [28]. Hence, there may be an opportunity for interventions to leverage social features that enable automation or do not require immediate input from users.

Traditionally, eHealth research has been opaque in providing details about how such features are implemented since the focus of such applied work falls on the population of interest or the outcomes achieved. Therefore, a system of labeling and categorizing these social features will allow academic and commercial eHealth app creators to have a clearer understanding of the various social features that are commonly implemented, enabling them to apply these features in their own work.

This study aimed to establish the SoCAP (social communication, affiliation, and presence) taxonomy of social features by conducting a scoping review of commercially available eHealth apps. A scoping review was suitable for the purpose of the study since this method allows for flexibility in determining inclusion criteria and search methods, which we adapted for identifying and classifying a broad range of social features. First, we conducted a search of commercially available apps, and subsequently categorized the social features to derive our taxonomy. This may help researchers describe their work and draw comparisons between other eHealth platforms.

Methods

Overview

This study may be characterized by a scoping review, which is commonly applied to topics that are complex or have not been extensively reviewed before [29]. Conventionally, this methodology applies to a scope of literature as its focus. However, our method involved determining the variety and categories of social features through a review of commercially available apps. Through this process, we developed our taxonomy, identifying four main categories of social features: (1) communication features (asynchronous and synchronous), (2) social presence features, (3) identity and affiliation features, and (4) other social integrations.

To conduct the scoping review, the 6-step framework methodology of Arksey and O'Mayley [30] was adapted. This included a methodical search, screening, and qualitative synthesis of the findings. This process includes (1) identifying the research question; (2) identifying relevant eHealth and mHealth interventions; (3) selection; (4) charting the data; (5) collating, summarizing, and reporting results; and (6) finally, consultation with experts in the field. No protocol was registered prior to the commencement of the study.

Eligibility and Screening

A shortlist of commercial eHealth and mHealth interventions was selected from One Mind PsyberGuide, a non-profit eHealth app review website maintained by researchers from the University of California, Irvine, and Northwestern University. The site is not affiliated with an industry and has a transparent review process. Each review is conducted by eHealth and mHealth specialists who have demonstrated expertise in the field, as opposed to anonymous reviews on commercial app stores, which may be paid, sponsored, or falsified.

App Rating System

To derive a shortlist of eHealth and mHealth apps, the website's app ratings were referenced, comprising their (1) credibility, (2) user experience, and (3) transparency by trained reviewers from the One Mind Psyberguide. Members of the study team were not involved in the rating process. At the time of writing, 207 apps were listed in the guide, with a subset of 30 apps fully reviewed and scored. In accordance with the guide's rating system, *Credibility* ratings are made on a scale from 1.00 to 5.00 and include components like having a proposed goal, clinical input, research basis, etc. *User Experience* ratings are made on a scale from 1.00 to 5.00 and include components like engagement, functionality, aesthetics, etc. *Transparency* is rated as "acceptable," "questionable," or "unacceptable" and pertains to an app's privacy policies and conforming to data storage and collection policies, with 1.00 representing the lowest rating and 5.00 representing the highest rating.

Based on these ratings, only apps with "acceptable" credibility were selected, and the top 20 apps with the highest aggregate credibility and user experience were shortlisted. Apps that primarily relied on teletherapy were excluded because it would not be possible to see how their social features worked if we met with a counselor or therapist, which was beyond the scope of the review. We also excluded apps that did not have social features. A significant limitation of selecting apps that have been reviewed by One Mind PsyberGuide is drawing from a smaller pool of apps that may not fully represent the scope of apps available. However, given that our team would not be able to review a large number of apps, we felt that this approach would allow us to evaluate apps that had met independent quality benchmarks. The app selection process is charted in Figure 1.

XSL•FO
RenderX

Figure 1. Flow diagram of the app search selection according to PRISMA (preferred reporting items for systematic reviews and meta-analyses) for scoping reviews.

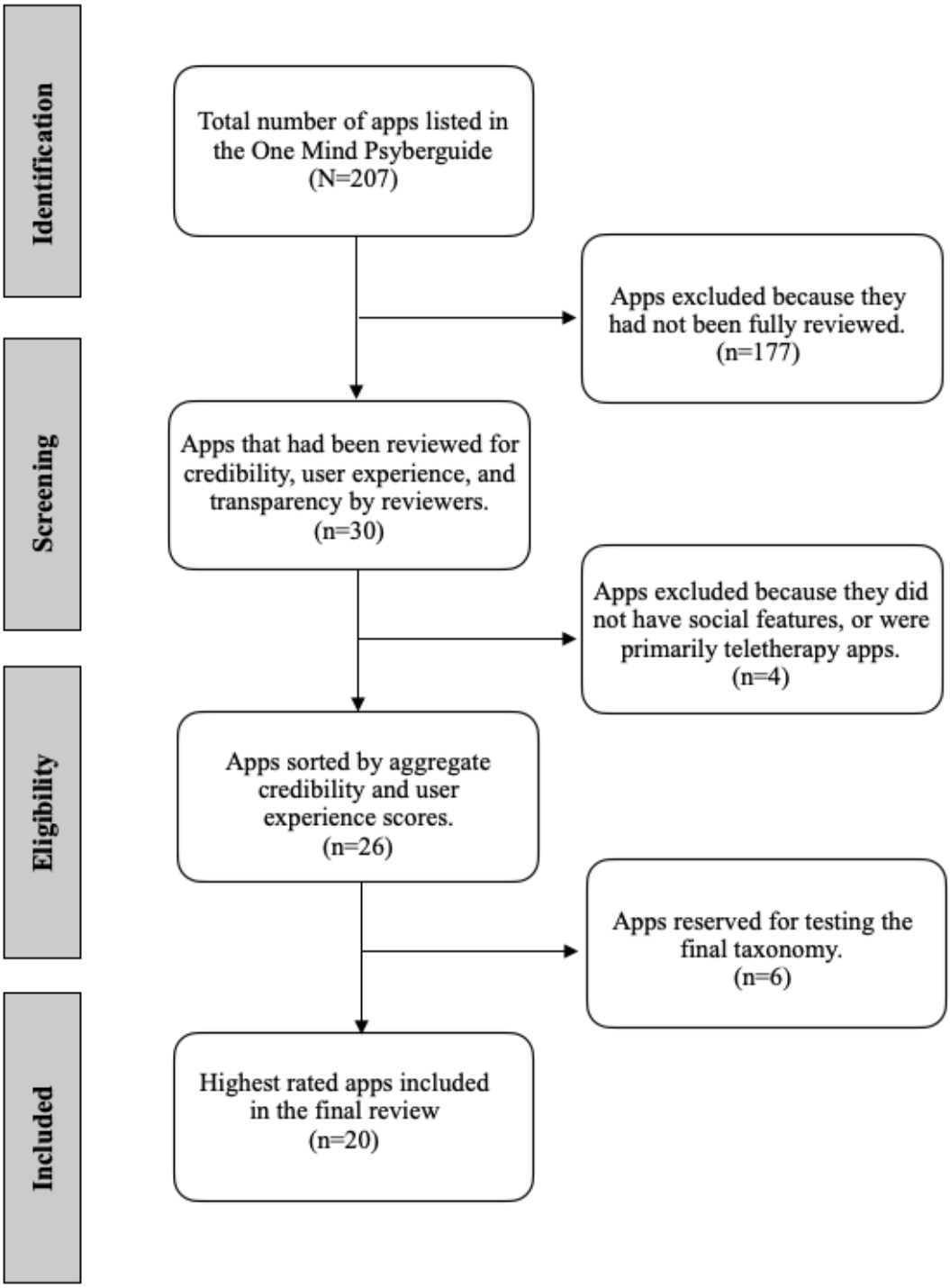

The 20 shortlisted apps, in descending order of aggregate scores, include Headspace [31], Happify [32], Calm [33], SuperBetter [34], CogniFit [35], Sanvello [36], Lumosity [37], MoodMission [38], Wysa [39], Woebot [40], MoodKit [41], GG OCD [42], Virtual Hope Box [43], Sinasprite [44], NOCD [45], MindDoc [46], Replika [47], Simple Habit [48], Shine [49], and 365 Gratitude [50].

Codebook Development

The top 20 apps from One Mind PsyberGuide that met the inclusion criteria were selected. Then, 2 independent coders (IK and MF) used the apps and categorized them according to the types of social features they involved. The remaining 6 apps that met inclusion criteria were reserved to test the SoCAP taxonomy after it was finalized.

Our first step was to create a preliminary codebook by referring to an existing social media feature taxonomy that focused exclusively on peer-based features in eHealth interventions [51]. The taxonomy comprised several categories and their corresponding features. However, the existing list primarily focuses on *peer-to-peer* features (eg, peer SMS, leaderboard,

XSL•FO
RenderX

and peer commenting), which does not fully capture the range of social features in eHealth apps. Hence, the list was expanded to include features that may facilitate interactions with clinicians or study team members. Additionally, the list was expanded to include those that may enhance a sense of social presence (eg, pre-recorded videos and automated messages), even if they did not require direct input from users. Next, we organized the list of social features into a hierarchical structure, with 4 main categories that were informed by the existing literature. The four categories are (1) communication features (asynchronous and synchronous), (2) social presence features, (3) identity and affiliation features, and (4) other social integrations.

Coding Process

The preliminary codebook was initially tested on a set of 5 apps that were randomly selected from the set of 20 apps. The coding process required the coders to note either the presence or absence of features. Both coders first performed independent coding using the preliminary codebook, and then subsequently convened to resolve discrepancies. This was done by reviewing the app together and collaboratively arriving at a decision. Subsequently, the necessary changes were made to the codebook. Next, the actual coding took place, in which both mobile- and web-based versions of each app were double-coded by the 2 coders. A first set of 5 apps was selected, where both coders (1) independently coded the apps, (2) shared their coding results, and (3) resolved any discrepancies in their coding. The double-coding process ensured that each app was extensively tested and reviewed in both mobile app and web-based formats. The next 5 apps were then coded using the same process, which was repeated until all 20 apps were reviewed. The final version of the taxonomy was then tested on 6 apps that had been reviewed by the One Mind PsyberGuide to confirm that no salient categories or features had been overlooked. There were no changes made to the codebook during this process.

As the coders reviewed the apps, specific use-cases that demonstrated innovation were identified. For example, this included (1) *novel* social interaction features, (2) existing features that have been *applied* in new ways for users, (3) features that may uniquely enhance a user's sense of social presence, or (4) features that have not yet been addressed in the existing literature.

Additional Analyses

In addition to noting the absence or presence of the social features, additional descriptive analyses were performed. This included the percentage of apps that addressed (1) psychological concerns (eg, symptoms of depression or anxiety and stress), (2) behavioral and health concerns (eg, exercise, diet, and smoking cessation), and (3) a combination of both psychological and behavioral and health concerns (eg, prenatal mental health). We also calculated the mean number of social features included per app.

Results

Overview

The SoCAP taxonomy consists of four main categories of social features: (1) communication features (asynchronous and synchronous), (2) social presence features, (3) affiliation and identity features, and (4) other social integrations. These categories and their corresponding descriptions are summarized in Table 1, while the specific features and their corresponding descriptions and examples are summarized in Table 2.

Table 1. Main categories of social media features.

Category	Description
Communication features	Features that allow for communication between 2 or more users; comprises (a) *synchronous communication features*, which allow for real-time communication, and (b) *asynchronous communication features*, which allow users to communicate without requiring them to be present at the same time
Social presence features	Features that enhance the perception of others on the platform
Affiliation and identity features	Features that enhance a participant's ability to represent themselves or affirm their affiliations with others
Social integration features	Features that leverage existing social networks or communication platforms

XSL•FO
RenderX

Table 2. The SoCAP (social communication, affiliation, and presence) taxonomy of social features.

Categories and features	Description and examples
1. Communication features	
(a) Synchronous features	
Chatrooms	Real-time discussion groups
Voice calls	Voice-based calls via telephone, cell phone, Google Voice, etc
SMS text messaging	Contacting other users, facilitators, or providers via text
Videoconferencing	Video calls conducted in a group or individual setting
Other	Other features that may be conducted in real-time; eg, live group meditations
(b) Asynchronous features	
Discussion boards	A bulletin board where users can interact by posting messages
Email groups and mailing lists	A collection of users' email addresses, which allows them to send and receive emails from the group
Expert support	Contacting a provider or facilitator through asynchronous means; eg, messaging and requesting support
In-app peer messaging	Contacting other users directly by sending messages
Other	Other features that may be conducted asynchronously; eg, commenting on videos
2. Social presence features	
Chatbot	A feature that simulates human conversation through text or voice interactions
Check-ins and surveys	Includes brief assessments and measures that collect user information
Key quotes	Quotes from other users, thought leaders, etc
Onboarding video and tutorial	Welcome messages or tutorials for users when they first join the platform
Personal multimedia	Includes podcasts, videos, interviews, testimonials, recordings, etc
Reminders and push notifications	Includes prompts, email reminders, notifications, etc
Social data sharing and display	Displaying user activity data, feeds, or polls
Social gamification	Game-like elements involving other users; eg, leaderboards and competitions
Soliciting feedback	Questions or surveys that collect user feedback; eg, “How was this exercise?” and “Rate this lesson”
Tailoring	Content or messages that are adapted based on user input; eg, personalized messaging and tailored information
User counts	Displaying the number of participants who are engaging in online activities; eg, video views and number of users online
Other	Other features that enhance social presence; eg, automated messages on user feeds
3. Affiliation and identity features	
Accessibility panels and options	Includes settings to increase font size, narrate text, or enable voice commands to enhance accessibility
Addressing participants by name	Includes personalized emails, notifications, welcome messages, etc
Characters and avatars	Visual representations of other users, instructors, or characters
Content for underserved groups	May include lessons, skills, activities, and other content for underserved groups; eg, track for anti-Asian hate and meditations for injustice
Expert profiles	Individual profiles for facilitators, providers, and other experts.
Network formation	Features that allow users to form networks; eg, befriending and following
Peer groups and affiliation	Features that allow users to join groups based on their identity, interests, or concerns; eg, a discussion board for LGBTQ[a] individuals and health care workers
Personal avatar and photo upload	Users may be able to upload their own photos, customize avatars, or upload their own photos to present themselves to others
Privacy and sharing options	Providing toggles for users to choose what content they would like to share or withhold from others
Shared and family accounts	Allowing multiple users within one family account
User profiles	A display of personal information about the user that may be public or private

XSL•FO
RenderX

Categories and features	Description and examples
Other	Other affiliation and identity features
4. Social integration features	
Social network integration	Interfacing with social networking sites such as Facebook and Instagram
Other app integration	Interfacing with other apps such as WhatsApp and Viber
Sending invites	Being able to invite others to join the app
Other	Other social integrations and features

[a]LGBTQ: lesbian, gay, bisexual, transgender, and queer.

All (N=20) of the reviewed apps addressed psychological concerns, 85% (n=17) addressed behavioral and health concerns, and 85% (n=17) addressed a combination of both psychological and behavioral and health concerns (PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] flowchart shown in Figure 1). The mean number of social features included per app was 12.2 (SD 5.63, range 2-25).

Communication Features: Asynchronous and Synchronous

Of the features identified in the taxonomy, those that facilitate synchronous (eg, texting and chatrooms) and asynchronous communication (eg, private messaging and discussion boards) provide the highest level of interaction with peers, facilitators, and providers. However, the boundary between "synchronous" and "asynchronous" features can be obscured depending on how these features are used. Therefore, the social features were categorized under one or the other based on their conventional use. An example of an innovative application of a synchronous feature was the "live" meditation feature from Headspace (Figure 2), in which users can follow along to a meditation that was led by a facilitator in real time. While participants are not able to directly interact with other participants, they are able to view how many others are engaging in the live activity, thereby fostering a sense of community among users.

Figure 2. Synchronous features: "Live" meditation feature from Headspace.

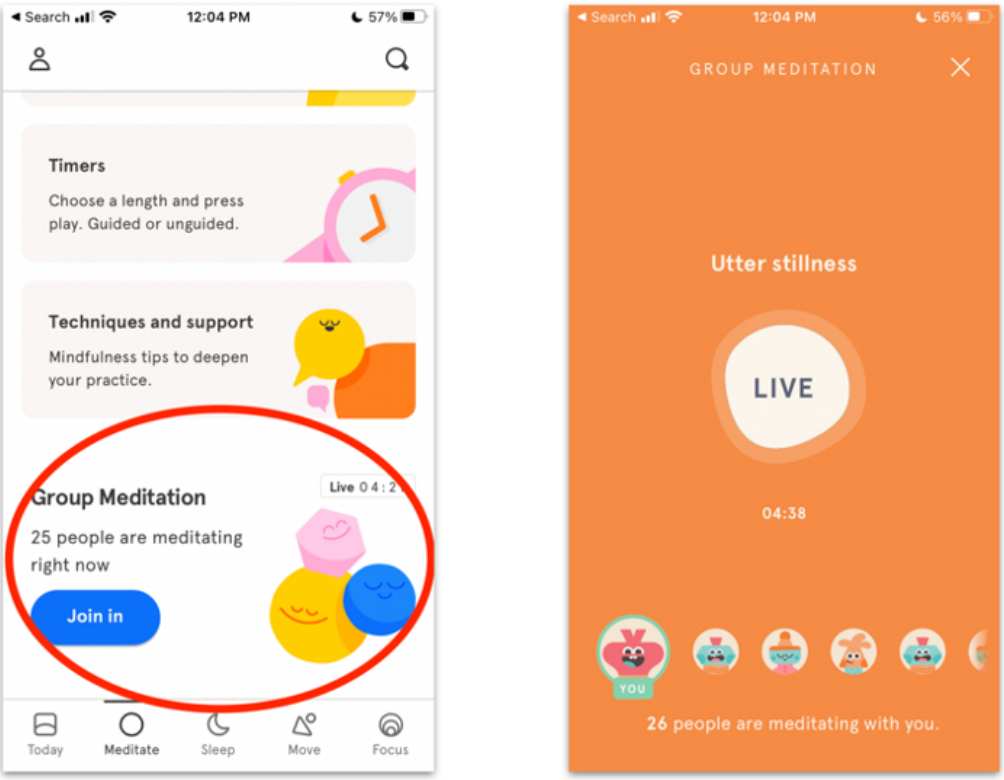

A total of 25% (n=5) of the apps included at least 1 *synchronous* communication feature, with those apps including an average of 1.4 (SD 0.89, range 1-3) of such features. A total of 45% (n=9) included at least 1 *asynchronous* communication feature, with those apps including an average of 2.8 (SD 1.04, range 1-3) of such features. Taken together, 50% (n=10) of all apps did not include either an asynchronous or synchronous communication feature.

XSL•FO
RenderX

The reason for this may be that facilitating and moderating these communications may add an additional layer of complexity to maintaining these products. Each app averaged 0.35 (SD 0.75, range 0-3) synchronous and 0.7 (SD 1.08, range 0-3) asynchronous features; comparatively, our review found a *higher* average number of social presence features (mean 6.85, SD 2.32, range 2-10) and identity representation features per app (mean 4.40, SD 2.68, range 0-8), suggesting a heavy reliance on these automated features that require little or no direct input to enhance a sense of social connection. This may indicate developers' preference for social features that are less resource-intensive.

Social Presence Features

One construct that has been widely studied in online communities is the phenomenon of social presence and its effect on engagement and user satisfaction. This construct was initially adopted for e-learning research but has since been applied to various domains of research involving social interactions in digital spaces.

In our review, social presence features accounted for the most common type of social feature being implemented. 100% (n=20) of the apps included at least 1 social presence feature, with apps including an average of 6.85 (SD 2.32, range 2-10) of such features. These features appeared crucial in enhancing the sense that other users or staff member facilitators or providers were engaged with the app—without requiring a high level of human input. In general, all types of social features, including the previously described asynchronous and synchronous communication features, may enhance a sense of social presence. However, those that facilitate communication require a higher level of input and user interaction. Therefore, this separate category of social presence features was created, which enhances the perception of others without requiring as much input.

Some notable examples of social presence features include the artificial intelligence (AI)–driven chatbot featured in the Woebot app (Figure 3), which relies on conversing with a virtual robot character to access a wide range of psychological and behavioral health intervention content. Another instance where a chatbot is central to the user experience is the replica app, where users can interact with customized avatars.

Another common social presence feature was the application of check-ins and surveys, which were deployed to gauge the participant's mood, interest level, or perceived benefit of certain features or modules. The solicitation of user feedback bolsters the impression that a user's inputs are valued and may enhance a general sense that their preferences are being acknowledged. This enhances the perception that there is an actual person "behind" the app, thereby fostering a sense of social presence. For example, in the posttraumatic stress disorder Coach app (Figure 4), users perform a brief check-in before and after exercises to indicate how they are feeling and to indicate how helpful they found the exercise.

Figure 3. Social presence features: AI-driven chatbots featured in Woebot and Replika.

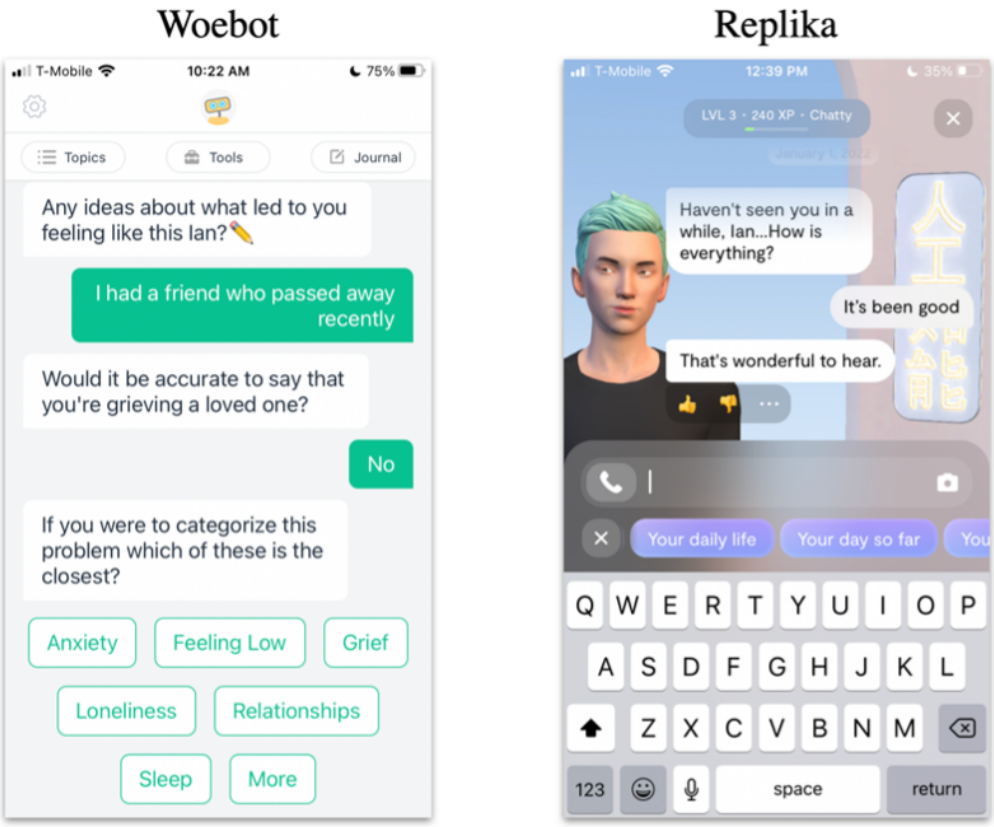

XSL•FO
RenderX

Figure 4. Social presence features: community map and user counts in the Lumosity app.

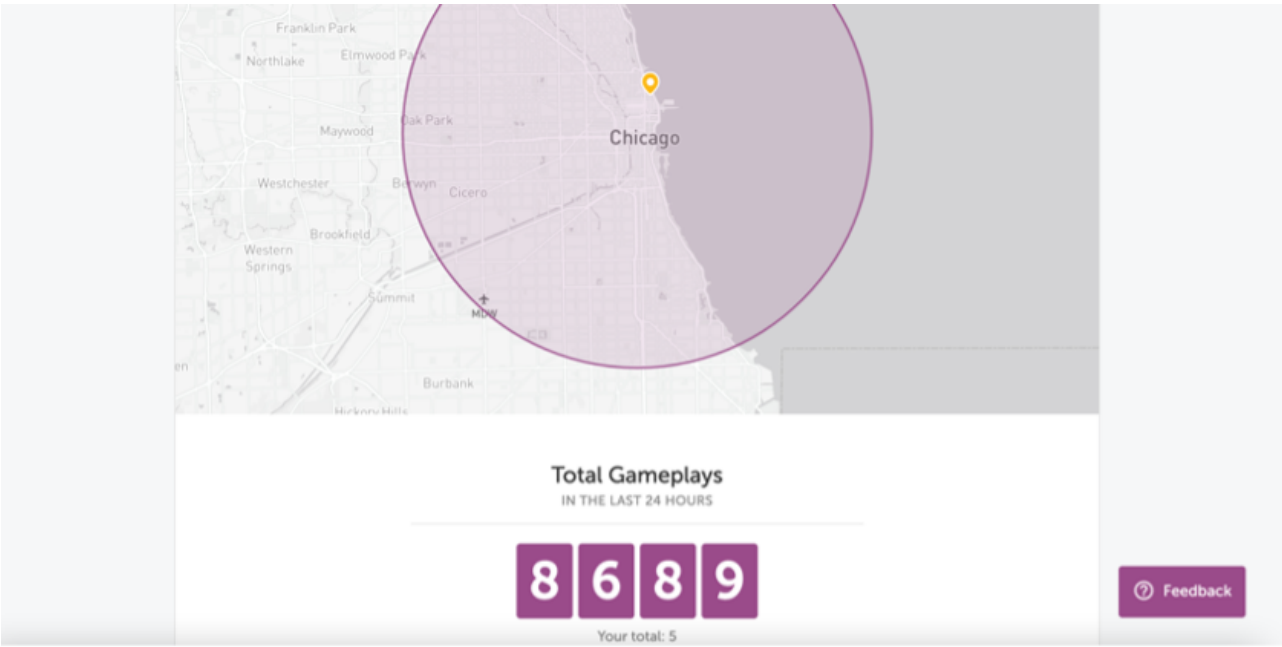

Additional social presence features include user accounts, where apps may display how many people have engaged with certain content or might be on the platform at the same time. An example of this is the community map in the Lumosity app, where participants are shown a map of their location, and are able to view a count of how many people in the geographic region have logged on to the app or have engaged in similar activities as them recently.

Affiliation and Identity Features

Through our review, the category of identity and affiliation features emerged. These are the features that do not require direct communication or interaction with others but provide the foundation necessary for such engagement. Through customization or personalization, these features allow users to present themselves to others in digital spaces. Such features may also reinforce group identities, allowing participants to experience a sense of affiliation and community [52,53]. This may be particularly important in fostering a sense of social connection in the context of eHealth apps, since participants may not have the ability to communicate with each other.

In total, 95% (n=19) of the apps included at least 1 identity and affiliation feature, with apps including an average of (mean 4.40, SD 2.68, range 0-8) of such features. The most common identity and affiliation features were user profiles, avatars, photo upload features, and addressing participants by name. Several apps also provide extensive accessibility options for individuals with different needs. One example is the Headspace app (Figure 5), which uses an EqualWeb toolbar to provide various options to make their website more accessible to users with sensory impairments or disabilities. These inclusive design features improve access, thereby allowing users with unique needs to fully use these eHealth apps. This enables users to participate on their own terms and affirms the presence of users with different support needs.

In the context of fostering affiliation, some features allowed users to form informal networks by following or befriending others, while other features allowed participants to opt into courses or online groups designed for specific populations (eg, expectant mothers and health care workers). Notably, several apps provide content for underserved groups who may be experiencing increased distress due to their racial or other identity. For example, Happify offers a track for Asian American users who may be experiencing racial discrimination (Figure 6). This was developed by a Taiwanese American therapist who teaches users how to manage feelings of fear and anxiety that have arisen from anti-Asian sentiment and also how to experience pride in their unique identities in spite of this adversity. Other examples include the Headspace app, which offers content addressing injustice and privilege, and the Wysa app, which provides support for LGBTQIA (lesbian, gay, bisexual, transgender, queer, intersex, and asexual) individuals. Taken together, such features help foster a sense of social connection by acknowledging the unique needs of these underserved user groups.

XSL•FO
RenderX

Figure 5. Affiliation and identity features: accessibility options in the Headspace app.

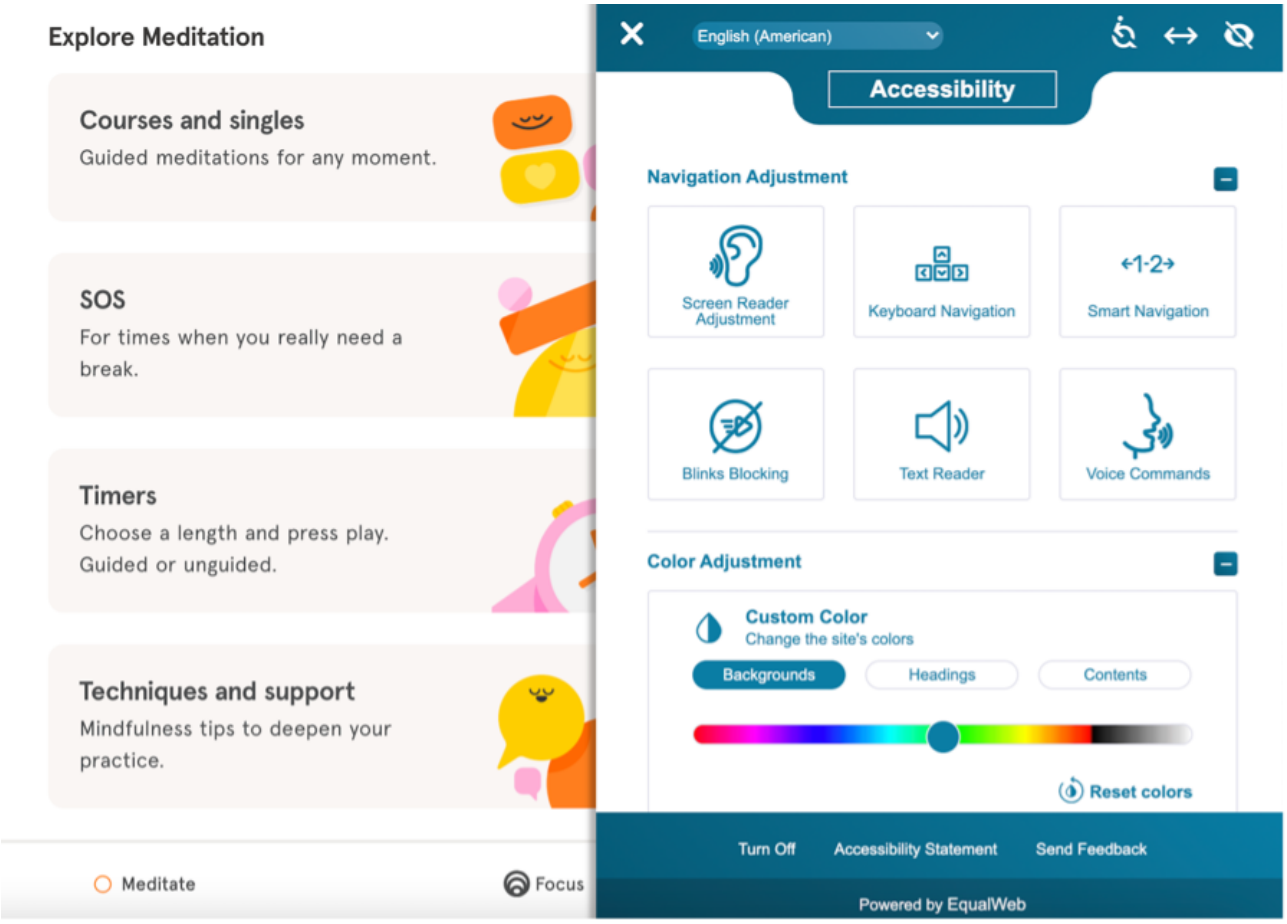

XSL•FO
RenderX

Figure 6. Affiliation and identity features: content for underserved user groups in Happify.

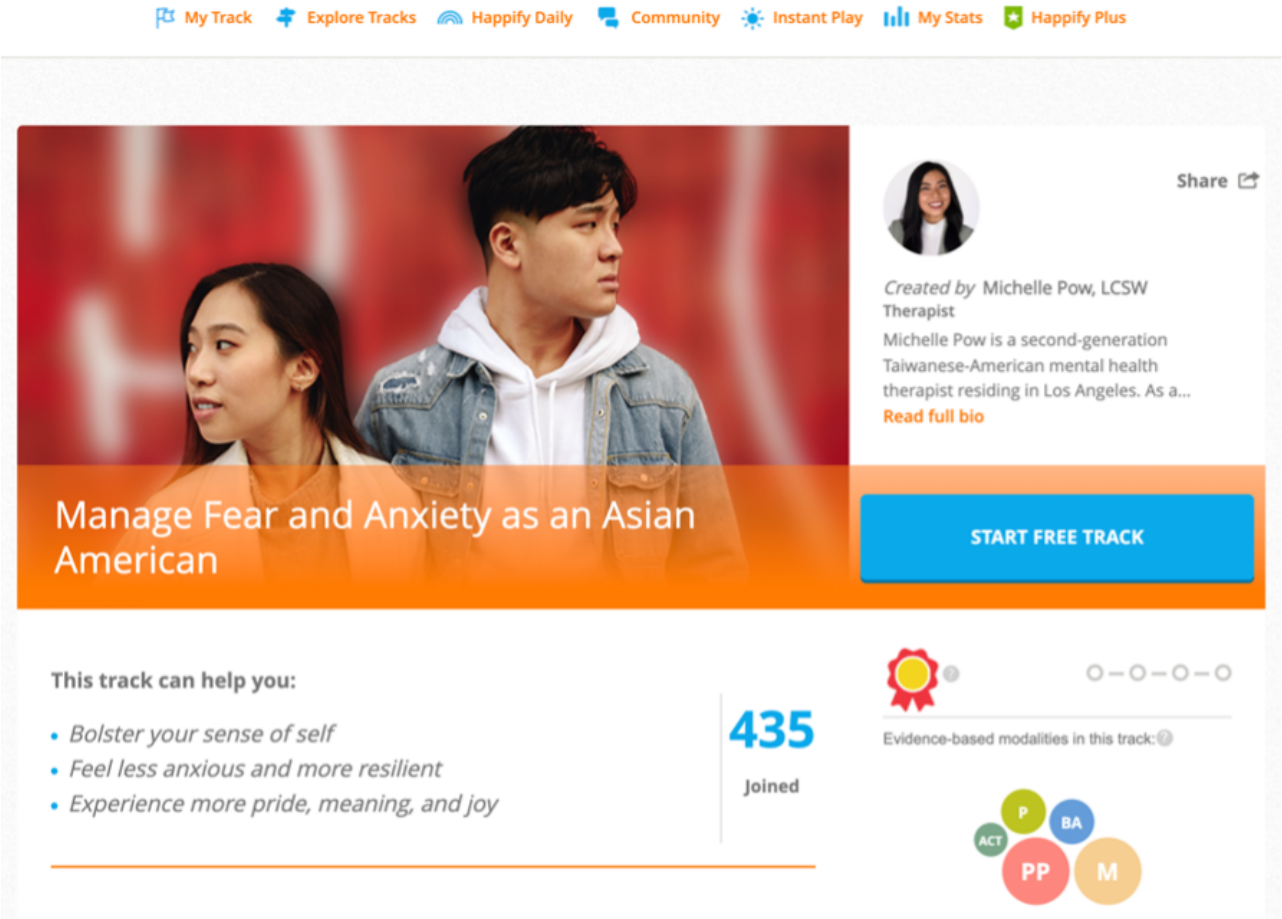

Other Social Integrations and Features

It was found that most eHealth apps have features that leverage existing social networks or communication platforms. For example, the Shine app provides users with membership in a private Facebook group where they can participate in discussions. Another example is the Simple Habit Meditation app, which allows users to share their achievements on Facebook and Instagram. Other related features include allowing users to send email invites to refer others to join the app as members. Such social features maximize the use of existing social networks or communication technologies and may serve as channels to grow their user base. Additionally, such features may potentially enhance the interactivity of the apps, thereby increasing engagement. In total, 85% (n=17) of the apps included at least one form of social integration or other social features, with apps including an average of 2.35 (SD 1.00, range 0-7) of such integrations and features.

Discussion

Principal Findings

This study involved the development of the SoCAP taxonomy of social features through a scoping review of social features in commercial eHealth apps. To derive the taxonomy, a search of commercially available apps was first conducted. Subsequently, we categorized the social features to build the taxonomy. It comprises four categories: (1) communication features (asynchronous and synchronous), (2) social presence features, (3) affiliation and identity features, and (4) other social integrations. It was found that social presence features and identity and affiliation features were more prevalent, encompassing a wider range of functionalities.

In contrast, our literature review found that social features currently used in research-driven eHealth apps tend to leverage synchronous interactions with peers, study staff, or providers; allowing them to meet the needs of high-acuity participants (eg, severe mental illness and patients with medically complex presentations). However, these social features are likely to require significant human resources to maintain since they are often reliant on real-time interactions. Through our initial literature review, it was found that the descriptions of how these features work were often minimal, given that the functionality of these eHealth platforms was typically secondary to the content being delivered.

In terms of social features in commercial eHealth apps, it is likely that there may be additional cost-driven constraints due to an emphasis on profitability. Our review found that commercially available apps lean heavily on automated features, which do not require as much human input as communication-heavy features, which require ongoing curation and moderation. While there are no definitive explanations for this, one hypothesis may be that most commercially available apps are typically profit-driven enterprises, and thus focus on

XSL•FO
RenderX

the most cost-efficient use of human resources to deliver these eHealth products. Additionally, automated features may be more desirable for users in their ability to enhance social connection without others being present on the platform at the same time or to enhance their sense of privacy by not having to engage in direct communication. Therefore, researchers might look to commercial eHealth apps for insight on developing (1) social presence, (2) identity and affiliation, and (3) social integration features that enhance their participants' sense of social connection in a way that is cost-efficient.

Further Research and Implications

The proposed SoCAP taxonomy provides researchers with a framework to evaluate the range of social features currently being implemented in eHealth apps. This enables those who are less familiar with the technical aspects of eHealth intervention design to understand how these features might work and to describe the functionality of the eHealth interventions that they are implementing. The current taxonomy also enables greater transparency about the user-experience design and technical aspects of how eHealth apps may work by establishing common technical terms that both scientists and technologists can use. This fosters a greater understanding of the social technologies that are being applied in both research and industry.

Strengths and Limitations

This study incorporates insight from both academia and industry. This involved combining an overview of social features in the eHealth literature, with a hands-on review of social features in commercially available eHealth apps. Since few of the interventions described in the literature are available for third-party access, the scoping review of commercially available apps allowed us to systematically test and use the apps in order to develop our taxonomy while drawing from the trends and insights gained from the narrative summary of the existing eHealth literature.

However, our scoping review has several limitations. First, it was conducted on a limited set of eHealth apps. This limitation was mitigated by testing the taxonomy on other commercially available apps. However, it remains to be proven how generalizable the taxonomy will be to academic research. Therefore, further research on the taxonomy could include its application to a broader set of eHealth apps in academic research. This may be performed through a conventional systematic literature review or conducted more extensively by contacting researchers directly to gain permission to use and evaluate their interventions. This would allow for further refinement of the taxonomy.

Second, another limitation relates to the nature of the rapidly evolving eHealth landscape; in which new operating systems, data transmission, and mobile technologies are constantly evolving. Therefore, a shortcoming of this screening method is that apps that may have recently been made available might not be included in this list.

This will result in some of the individual social features identified being obsolete over time, with emerging features that should be added to the SoCAP taxonomy. Yet, the broader categories of (1) communication features (asynchronous and synchronous), (2) social presence features, (3) affiliation and identity features, and (4) other social integrations are likely to be flexible enough to include more recent additions to the field. Nonetheless, the taxonomy should continue to be updated to account for these developments and mitigate obsolescence.

Future Directions

Given that social presence and identity and affiliation features were frequently featured in the apps reviewed, we anticipate that there will be continued innovation in these categories. For example, refinements in AI algorithms are allowing chatbots to match users' personalities in order to enhance engagement and satisfaction in these communications [54]. Another potential area of growth would be social features leveraging virtual reality, where users interact with immersive, digital environments and augmented reality technologies that integrate both simulated and real-life experiences. A notable example is Facebook's ambitious efforts to develop virtual reality and augmented reality applications for its Metaverse initiatives. These are likely to result in the development of more immersive social features.

As these emerging social technologies are integrated into eHealth interventions, they will be accompanied by questions on how these innovations lead to measurable outcomes. In this study, we have developed an extensive list of social features. Future research may involve the application of the SoCAP taxonomy to evaluate different feature sets and compare their outcomes, for example, effectiveness and user preference. The growing complexity of eHealth technologies also necessitates an understanding of how an expanding social feature set might burden users—and, on the flip side, establishing a "baseline" amount of social features that could still efficiently enhance social connection. Taken together, the SoCAP taxonomy may provide app developers with a reference for categorizing and evaluating these features.

Conclusions

This study involves the development of SoCAP, a taxonomy of social features in eHealth interventions. Through our scoping review of social features in commercially available eHealth apps, we identified four main categories in the SoCAP taxonomy; these include (1) communication features, (2) social presence features, (3) identity and affiliation features, and (4) other social integrations and features. The taxonomy represents a pioneering initiative to provide researchers with a shared understanding of the scope of social features currently being implemented in eHealth apps. This will allow researchers to apply these features to their own work, thereby fostering further innovation. This study found a heavy reliance on automated features that establish a context for meaningful interactions among users. We anticipate significant innovation in automated features in the future, particularly with the maturation of AI technologies. However, future research may involve further validating the SoCAP taxonomy across a broader range of research-oriented and commercially available apps and identifying which may be linked with enhanced user engagement and outcomes. The continued application and refinement of the SoCAP taxonomy will allow it to evolve as a robust tool for eHealth app development.

XSL•FO
RenderX

Acknowledgments

We would like to thank Dr Stephen Schueller for sharing his knowledge and insight on evaluating eHealth technologies. Generative artificial intelligence was not used in the writing of this manuscript.

Data Availability

The data collected in the coding process may be obtained by contacting the first author.

Authors' Contributions

IK, JTM, EGL, and DY conceived and planned the study. IK and LK conducted the literature search. IK and MF carried out the scoping review. All authors discussed the results and contributed to the final manuscript.

Conflicts of Interest

EGL is currently employed by Lyra Health and Lyra Clinical Associates, receives income from Lyra Health and Lyra Clinical Associates, and has been granted equity in Lyra Health.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 526 KB - jmir_v26i1e49714_app1.pdf]

References

1. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023;53(2):567-579. [doi: 10.1017/S0033291721001999] [Medline: 34047264]
2. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR mHealth uHealth 2020;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]
3. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. Curr Opin Psychol 2020;36:65-70 [FREE Full text] [doi: 10.1016/j.copsyc.2020.04.008] [Medline: 32553848]
4. Wurmser Y. US time spent with mobile 2019. EMARKETER Inc. 2019 May 30. URL: https://www.emarketer.com/content/us-time-spent-with-mobile-2019 [accessed 2024-08-13]
5. Hootsuite. The global state of digital in 2019. 2019. URL: https://www.hootsuite.com/resources/digital-in-2019 [accessed 2024-07-13]
6. Byambasuren O, Beller E, Glasziou P. Current knowledge and adoption of mobile health apps among Australian general practitioners: survey study. JMIR Mhealth Uhealth 2019;7(6):e13199 [FREE Full text] [doi: 10.2196/13199] [Medline: 31199343]
7. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
8. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
9. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-Health: a scoping review. Telemed J E Health 2022;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
10. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
11. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021;23(5):e18500 [FREE Full text] [doi: 10.2196/18500] [Medline: 34057427]
12. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009;43(3):535-554. [doi: 10.1177/0038038509103208]
13. Walsham G. Actor-network theory and IS research: current status and future prospects. In: Information Systems and Qualitative Research. Boston, MA: Springer; 1997:466-480.

XSL•FO
RenderX

14. Dodge HH, Hampstead BM, Potempa K, Struble L, Lindsley J, Rooney WD, et al. P1-639: behavioral randomized controlled trial (RCT) using internet-based social interactions as a tool to enhance cognitive reserve. University of Michigan Library. 2018. URL: https://deepblue.lib.umich.edu/handle/2027.42/152959?show=full [accessed 2024-07-13]
15. Pitt R, Hill AJ, Theodoros D, Russell T. “I definitely think it’s a feasible and worthwhile option”: perspectives of speech-language pathologists providing online aphasia group therapy. Aphasiology 2018;32(9):1031-1053. [doi: 10.1080/02687038.2018.1482403]
16. Petrič G, Atanasova S, Kamin T. Ill literates or illiterates? Investigating the eHealth literacy of users of online health communities. J Med Internet Res 2017;19(10):e331 [FREE Full text] [doi: 10.2196/jmir.7372] [Medline: 28978496]
17. Parikh SV, Huniewicz P. E-health: an overview of the uses of the Internet, social media, apps, and websites for mood disorders. Curr Opin Psychiatry 2015;28(1):13-17. [doi: 10.1097/YCO.0000000000000123] [Medline: 25420193]
18. Alvarez-Jimenez M, Gleeson JF, Bendall S, Penn DL, Yung AR, Ryan RM, et al. Enhancing social functioning in young people at ultra high risk (UHR) for psychosis: a pilot study of a novel strengths and mindfulness-based online social therapy. Schizophr Res 2018;202:369-377. [doi: 10.1016/j.schres.2018.07.022] [Medline: 30031616]
19. Aragon SR. Creating social presence in online environments. New Dir Adult Contin Educ 2003;2003(100):57-68. [doi: 10.1002/ace.119]
20. Lowenthal PR. Social presence. In: Social Computing: Concepts, Methodologies, Tools, and Applications. Hershey, PA: IGI Global; 2010:129-136.
21. Gunawardena CN, Zittle FJ. Social presence as a predictor of satisfaction within a computer‐mediated conferencing environment. Am J Distance Educ 1997;11(3):8-26. [doi: 10.1080/08923649709526970]
22. ter Stal S, Kramer LL, Tabak M, op den Akker H, Hermens H. Design features of embodied conversational agents in eHealth: a literature review. Int J Hum Comput Stud 2020;138:102409. [doi: 10.1016/j.ijhcs.2020.102409]
23. Gabrielli S, Rizzi S, Bassi G, Carbone S, Maimone R, Marchesoni M, et al. Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. JMIR Mhealth Uhealth 2021;9(5):e27965 [FREE Full text] [doi: 10.2196/27965] [Medline: 33950849]
24. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
25. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
26. Interian A, Kline A, Perlick D, Dixon L, Feder A, Weiner MD, et al. Randomized controlled trial of a brief Internet-based intervention for families of veterans with posttraumatic stress disorder. J Rehabil Res Dev 2016;53(5):629-640 [FREE Full text] [doi: 10.1682/JRRD.2014.10.0257] [Medline: 27898154]
27. Jarvis MA, Chipps J, Padmanabhanunni A. “This phone saved my life”: older persons’ experiences and appraisals of an mHealth intervention aimed at addressing loneliness. J Psychol Afr 2019;29(2):159-166. [doi: 10.1080/14330237.2019.1594650]
28. Yang WE, Shah LM, Spaulding EM, Wang J, Xun H, Weng D, et al. The role of a clinician amid the rise of mobile health technology. J Am Med Inform Assoc 2019;26(11):1385-1388 [FREE Full text] [doi: 10.1093/jamia/ocz131] [Medline: 31373364]
29. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
30. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
31. Headspace. 2024. URL: https://www.headspace.com/ [accessed 2024-07-13]
32. Happify. 2024. URL: https://www.happify.com/ [accessed 2024-07-13]
33. Calm. 2024. URL: https://www.calm.com/ [accessed 2024-07-13]
34. SuperBetter. 2024. URL: https://superbetter.com/ [accessed 2024-07-13]
35. CogniFit. 2024. URL: https://www.cognifit.com/ [accessed 2024-07-13]
36. Sanvello. 2024. URL: https://www.sanvello.com/ [accessed 2024-07-13]
37. Lumosity. 2024. URL: https://www.lumosity.com/ [accessed 2024-07-13]
38. MoodMission. 2024. URL: https://moodmission.com/ [accessed 2024-07-13]
39. Wysa. 2024. URL: https://www.wysa.com/ [accessed 2024-07-13]
40. Woebot. 2024. URL: https://woebothealth.com/ [accessed 2024-07-13]
41. MoodKit. 2024. URL: https://www.thriveport.com/products/moodkit/ [accessed 2024-07-13]
42. GGTude. 2024. URL: https://ggtude.com/ [accessed 2024-07-13]
43. Virtual Hope Box. 2024. URL: https://www.militaryonesource.mil/resources/mobile-apps/virtual-hope-box/ [accessed 2024-07-13]
44. Sinasprite. 2024. URL: https://www.litesprite.com/ [accessed 2024-07-13]

XSL•FO
RenderX

45. NOCD. 2024. URL: https://www.treatmyocd.com/ [accessed 2024-07-13]
46. MindDoc. 2024. URL: https://minddoc.com/us/en [accessed 2024-07-13]
47. Replika. 2024. URL: https://replika.com/ [accessed 2024-07-13]
48. Simple Habit. 2024. URL: https://www.simplehabit.com/ [accessed 2024-07-13]
49. Shine. 2024. URL: https://www.theshineapp.com/ [accessed 2024-07-13]
50. 365 Gratitude. 2024. URL: https://365gratitudejournal.com/ [accessed 2024-07-13]
51. Elaheebocus SMRA, Weal M, Morrison L, Yardley L. Peer-based social media features in behavior change interventions: systematic review. J Med Internet Res 2018;20(2):e20 [FREE Full text] [doi: 10.2196/jmir.8342] [Medline: 29472174]
52. Marwick AE. Online identity. In: Hartley J, Burgess J, Bruns A, editors. A Companion to New Media Dynamics. Hoboken, NJ: Wiley-Blackwell; 2013:355-364.
53. Gallos LK, Rybski D, Liljeros F, Havlin S, Makse HA. How people interact in evolving online affiliation networks. Phys Rev X 2012;2(3):031014. [doi: 10.1103/physrevx.2.031014]
54. Shumanov M, Johnson L. Making conversations with chatbots more personalized. Comput Hum Behav 2021;117:106627. [doi: 10.1016/j.chb.2020.106627]

Abbreviations

AI: artificial intelligence
LGBTQIA: lesbian, gay, bisexual, transgender, queer, intersex, and asexual
PRISMA: preferred reporting items for systematic reviews and meta-analyses
SoCAP: social communication, affiliation, and presence

Edited by T Leung, A Mavragani; submitted 21.06.23; peer-reviewed by J Matar, MA Rusandi; comments to author 20.02.24; revised version received 12.03.24; accepted 22.06.24; published 03.09.24.

Please cite as:
Kwok I, Freedman M, Kamsickas L, Lattie EG, Yang D, Moskowitz JT
The SoCAP (Social Communication, Affiliation, and Presence) Taxonomy of Social Features: Scoping Review of Commercially Available eHealth Apps
J Med Internet Res 2024;26:e49714
URL: https://www.jmir.org/2024/1/e49714
doi:10.2196/49714
PMID:

©Ian Kwok, Melanie Freedman, Lisa Kamsickas, Emily G Lattie, Dershung Yang, Judith Tedlie Moskowitz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Remote Patient Monitoring and Machine Learning in Acute Exacerbations of Chronic Obstructive Pulmonary Disease: Dual Systematic Literature Review and Narrative Synthesis

Henry Mark Granger Glyde[1], BSc, PhD; Caitlin Morgan[2], BSc, MRCP(UK), MBChB; Tom M A Wilkinson[3], MBBS, FRCP, MA, PhD; Ian T Nabney[4], BA, MA, PhD; James W Dodd[2], MBBS, FRCP, PhD

[1]EPSRC Centre for Doctoral Training in Digital Health and Care, University of Bristol, Bristol, United Kingdom
[2]Academic Respiratory Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[3]Clinical and Experimental Science, University of Southampton, Southampton, United Kingdom
[4]School of Engineering and Mathematics, University of Bristol, Bristol, United Kingdom

Corresponding Author:
Henry Mark Granger Glyde, BSc, PhD
EPSRC Centre for Doctoral Training in Digital Health and Care
University of Bristol
2nd Floor
Learning and Research Southmead Hospital
Bristol, BS10 5NB
United Kingdom
Phone: 44 7843037078
Email: hglyde18@gmail.com

Abstract

Background: Acute exacerbations of chronic obstructive pulmonary disease (AECOPD) are associated with high mortality, morbidity, and poor quality of life and constitute a substantial burden to patients and health care systems. New approaches to prevent or reduce the severity of AECOPD are urgently needed. Internationally, this has prompted increased interest in the potential of remote patient monitoring (RPM) and digital medicine. RPM refers to the direct transmission of patient-reported outcomes, physiological, and functional data, including heart rate, weight, blood pressure, oxygen saturation, physical activity, and lung function (spirometry), directly to health care professionals through automation, web-based data entry, or phone-based data entry. Machine learning has the potential to enhance RPM in chronic obstructive pulmonary disease by increasing the accuracy and precision of AECOPD prediction systems.

Objective: This study aimed to conduct a dual systematic review. The first review focuses on randomized controlled trials where RPM was used as an intervention to treat or improve AECOPD. The second review examines studies that combined machine learning with RPM to predict AECOPD. We review the evidence and concepts behind RPM and machine learning and discuss the strengths, limitations, and clinical use of available systems. We have generated a list of recommendations needed to deliver patient and health care system benefits.

Methods: A comprehensive search strategy, encompassing the Scopus and Web of Science databases, was used to identify relevant studies. A total of 2 independent reviewers (HMGG and CM) conducted study selection, data extraction, and quality assessment, with discrepancies resolved through consensus. Data synthesis involved evidence assessment using a Critical Appraisal Skills Programme checklist and a narrative synthesis. Reporting followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Results: These narrative syntheses suggest that 57% (16/28) of the randomized controlled trials for RPM interventions fail to achieve the required level of evidence for better outcomes in AECOPD. However, the integration of machine learning into RPM demonstrates promise for increasing the predictive accuracy of AECOPD and, therefore, early intervention.

Conclusions: This review suggests a transition toward the integration of machine learning into RPM for predicting AECOPD. We discuss particular RPM indices that have the potential to improve AECOPD prediction and highlight research gaps concerning patient factors and the maintained adoption of RPM. Furthermore, we emphasize the importance of a more comprehensive examination of patient and health care burdens associated with RPM, along with the development of practical solutions.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52143) doi:10.2196/52143

KEYWORDS

acute exacerbations of COPD; chronic obstructive pulmonary disease; exacerbate; exacerbation; exacerbations; remote patient monitoring; RPM; predict; prediction; predictions; predictive; machine learning; monitoring; remote; COPD; pulmonary; respiratory; lung; lungs; literature review; literature reviews; synthesis; narrative review; narrative reviews; review methods; review methodology

Introduction

Chronic obstructive pulmonary disease (COPD) is a disease defined by airway obstruction, airway inflammation, and in some cases parenchymal destruction (emphysema). COPD accounts for 55% of all chronic respiratory diseases [1] and is characterized by intermittent periods of significantly worsening symptoms known as exacerbations [2]. After a severe exacerbation, the in-hospital mortality is 6.7% [3]. Subsequently, the average mortality rates at 3 and 6 months stand at 18% and 26%, respectively, with a notable 50% mortality rate observed at 3.6 years [3,4]. Exacerbations increase airway and systemic inflammation and disease progression and cause a reduction in quality of life [5-11]. It is estimated that exacerbations in COPD account for 45% of COPD-related costs [12]. Patients who experience frequent acute exacerbations of COPD (AECOPD) have more primary care interactions, increased emergency department (ED) presentations, increased hospitalizations, and increased admissions to the intensive care unit [13]. A recent research priority-setting partnership in COPD found the highest-rated issue by patients or carers to be “identify better ways to prevent exacerbations” [14]. The researchers highlighted the importance of predicting and preventing exacerbations.

There is evidence to suggest that reducing delays in treatment and correct identification of exacerbations can reduce the severity of exacerbations, improve health-related quality of life (HRQoL), and reduce recovery time after an exacerbation. Wilkinson et al [15] found that a longer time to treatment in AECOPD was associated with an increase in the recovery time of exacerbation symptoms [15]. Moreover, they demonstrated that a greater number of correctly identified exacerbations treated by a physician resulted in a better HRQoL, as seen in the lower total St George’s Respiratory Questionnaire scores.

Remote patient monitoring (RPM) is a method of health care delivery that uses wearable devices and sensors to gather patient data outside of traditional health care settings. Using RPM, data that may provide a more detailed picture of the patient’s health become available. The patient, along with a team of health care professionals, can review these data to promptly identify changes in the patient’s health status, enabling early detection of potential exacerbations and facilitating timely intervention. Nevertheless, the current role of RPM in managing AECOPD remains uncertain. While some studies indicate potential benefits, others show no significant effects. Therefore, conducting a systematic review and synthesizing existing evidence becomes crucial to comprehend the current state of the art. This assessment is essential for defining the next steps in the development and testing of technology to address this global health challenge. Notably, leveraging machine learning approaches on the gathered data holds promise in enhancing RPM’s predictive capabilities. Thus, a comprehensive review is imperative to assess the current progress in this field.

In this dual systematic review, we aim to identify various approaches to remote monitoring for exacerbation intervention and prediction in AECOPD. We combine insights from both machine learning and remote clinical monitoring perspectives. The first review focuses on randomized controlled trials (RCTs) using RPM as an intervention to treat or improve AECOPD. The second review investigates studies integrating machine learning with RPM to predict AECOPD. This comprehensive approach enables us to provide a novel understanding of digitally enabled AECOPD interventions. We review the evidence and concepts behind RPM and machine learning; discuss the strengths, limitations, and clinical applications of available systems; and generate recommendations to enhance patient and health care system outcomes.

Methods

We conducted 2 systematic literature searches in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 1) statement [16].

Search Strategy

The searches were conducted between April and May 2023 in 2 electronic databases (Scopus and Web of Science) covering publications since the databases began. For Scopus, this includes records dating back to 1788, and for Web of Science, the database covers literature dating back to the early 1900s. The first search strategy included search strings in 4 main areas: COPD, RPM (telemedicine, telemonitoring, RPM, real-time monitoring, telehealth, mobile health, and digital health), study design (intervention and trial), and outcome (exacerbation frequency, exacerbation duration, ED presentations, hospital admissions, hospital readmissions, primary care interaction, health care costs, quality of life, and days in hospital). The second search strategy also included COPD and RPM but did not include study design, and instead of outcome, the search term was machine learning modeling (machine learning, deep learning, prediction models, and algorithms). The full search strategies for each database are presented in the Multimedia Appendix 2. Articles published in peer-reviewed journals or conference proceedings were considered for review. We did not include abstracts, dissertations, systematic reviews, or case studies.

Study Selection

To be included in the first search, studies were required to (1) specifically examine the use of RPM in COPD; (2) be an RCT; (3) have an exacerbation-related outcome variable, that is,

XSL•FO
RenderX

hospital admissions, exacerbation frequency, and HRQoL; (4) be published between the start date of each electronic database and May 2023; (5) be full freely available articles; and (6) be published in English.

For the second search, fewer studies were available. Therefore, studies were not required to be an RCT, and instead of including an exacerbation-related outcome variable, studies were required to incorporate a form of artificial intelligence modeling, usually machine learning algorithms, for exacerbation prediction.

Papers were excluded from the study for any one of the following reasons: (1) the study was a systematic literature review, (2) the study did not include any one outcome related to either the first search or the second search, (3) the focus of the intervention was behavior change (physical activity, medication adherence, and inhaler technique) or remote rehabilitation (usually pulmonary rehabilitation) rather than remote monitoring, and (4) the main study outcome was cost and did not include patient-related outcomes.

Machine learning studies were intentionally excluded from the first search. This decision was guided by the unique emphasis of each search: the first centered on RPM with clinical monitoring, while the second focused on RPM with an emphasis on machine learning.

Two authors (HMGG and CM) independently assessed the results obtained from the first literature search. Articles were screened in 4 steps: first, duplicates were removed, and then the title, abstract, and keywords were screened. Articles were screened on the inclusion and exclusion criteria outlined above. If authors were unable to determine the suitability during the screening, full-text articles were accessed for inclusion criteria and exclusion criteria. Full-text articles were excluded for not reporting outcomes for patients with COPD or the accuracy of COPD exacerbation prediction separately (in the case of studies with multiple diseases).

Remote monitoring studies were not included in the review if they reported on remote monitoring as an alternative to hospitalization for exacerbation treatment, did not include the specifics of the RPM use, or focused on the diagnosis of AECOPD rather than prediction.

Evidence Assessment and Narrative Synthesis

In the process of narrative synthesis, the initial step involved the identification and documentation of comparator groups. The lead author (HMGG) identified and documented comparator groups, capturing patient numbers, age, sex, and forced expiratory volume in 1 second (FEV1). Subsequently, the lead author recorded the specific RPM indices used in these studies, providing insight into the monitored parameters. These results are detailed in Multimedia Appendix 3. The RPM index, including its percentage occurrence in studies and the average study duration, can be found in Figures 1 and 2. The lead author identified and described the method of detection or prediction of AECOPD. This encompasses either a flagging system or clinical oversight or the application of machine learning approaches. Finally, the outcomes were described, documenting key measures, such as hospitalizations, HRQoL, and performance metrics, like sensitivity and accuracy. The results from the narrative synthesis can be found in Multimedia Appendix 3, which the remaining authors (ITN, JWD, and TMAW) checked.

For the initial search, we use the Critical Appraisal Skills Programme (CASP) RCT checklist. This tool is designed to systematically assess the validity, results, and relevance of RCTs, helping us gauge the quality of the evidence. Questions 9 and 11 were omitted from the evaluation as they are not relevant to this assessment. Studies satisfying 90% (9/10) of the criteria were designated as having the highest level of evidence (strongest evidence). Those meeting 80% (8/10) were categorized as strong evidence, while those meeting 70% (7/10) were classified as moderate evidence. Studies fulfilling 60% (6/10) of the criteria were considered to have limited evidence. For the second search, we used the CASP cohort checklist to assess cohort study quality. Questions 7 and 12 were excluded due to their lack of relevance. As in the first search, studies meeting 90% (9/10), 80% (8/10), 70% (7/10), and 60% (6/10) of the criteria were categorized as strongest, strong, moderate, and limited quality, respectively. The resulting rankings are visualized in Figure 3 [17-44] and described in the *Results* section.

XSL•FO
RenderX

Figure 1. Distribution of remote patient monitoring indices and average study duration for the first search. The figure illustrates the percentage of total studies each remote patient monitoring index appears in, alongside the average duration of these studies. ECG: electrocardiogram; SpO_2: oxygen saturation.

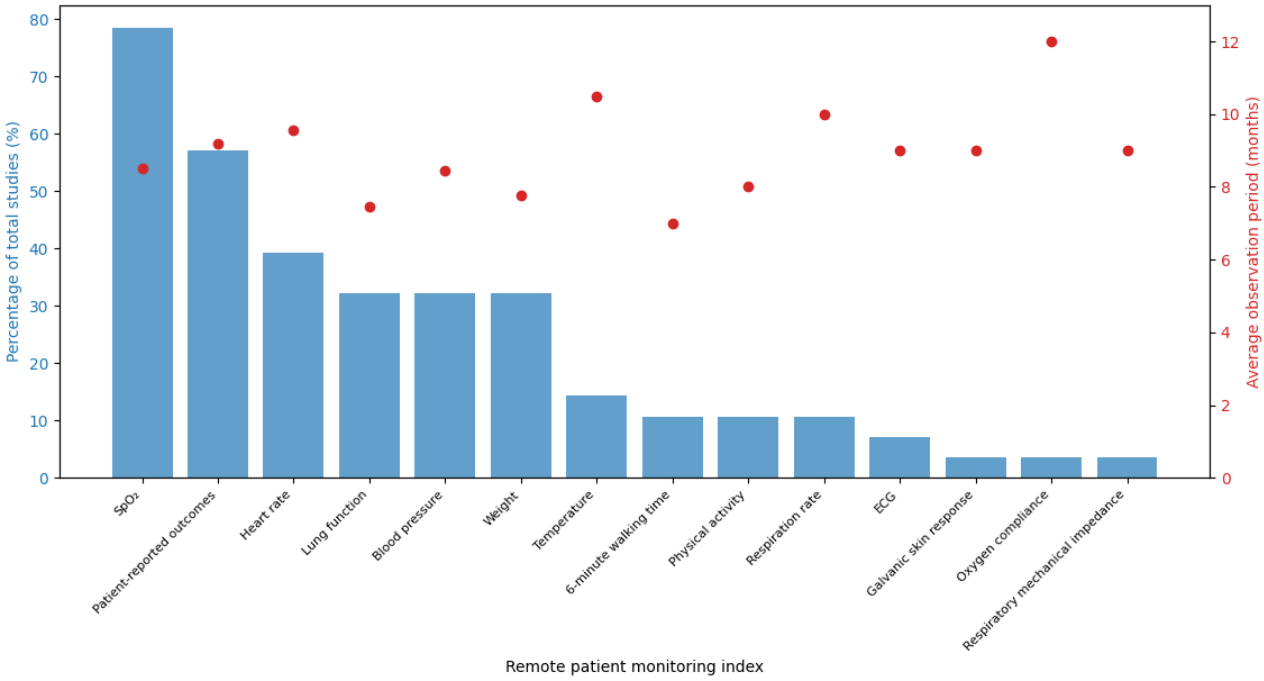

Figure 2. Distribution of remote patient monitoring indices and average study duration for the second search. The figure illustrates the percentage of total studies each remote patient monitoring index appears in, alongside the average duration of these studies. ECG: electrocardiogram; SpO_2: oxygen saturation.

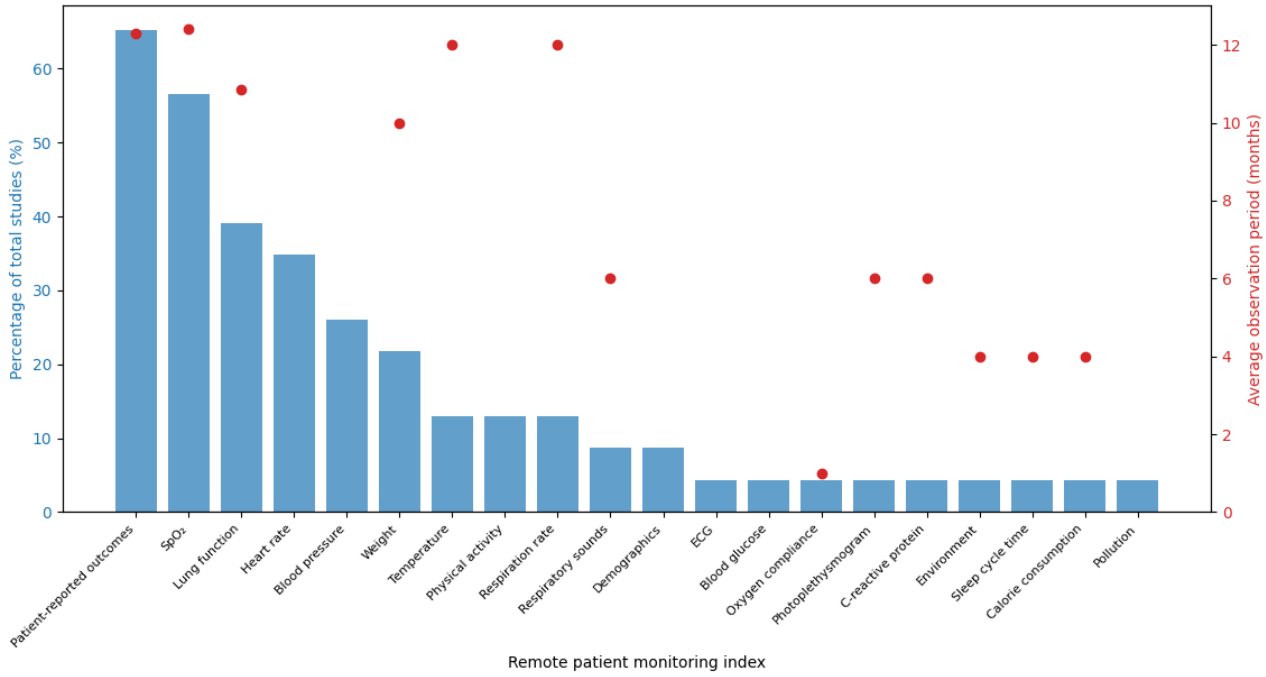

XSL•FO
RenderX

Figure 3. Rankings of RCTs on remote patient monitoring in chronic obstructive pulmonary disease according to the Critical Appraisal Skills Programme RCT checklist. RCT: randomized controlled trial.

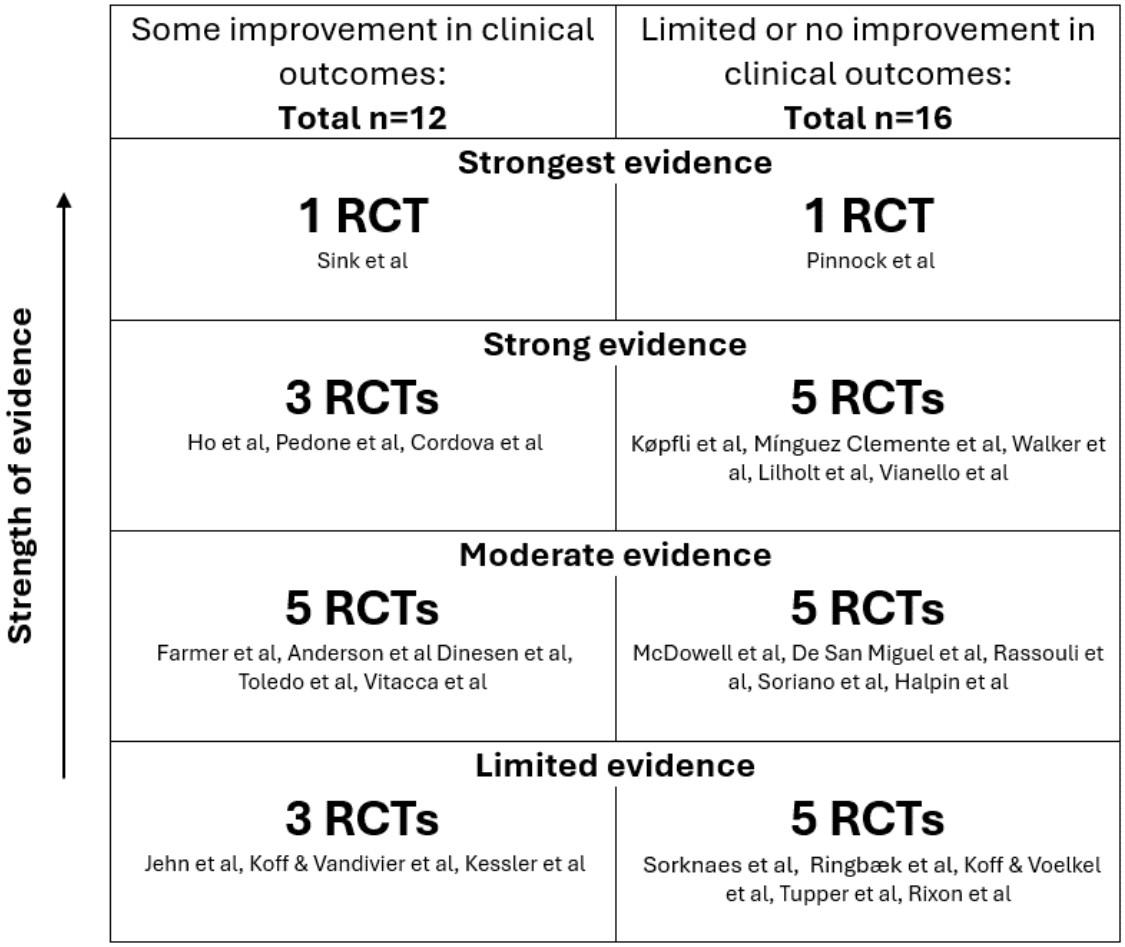

Results

Overview

We screened and analyzed data from April to May 2023. Through the first systematic search, we identified 216 studies, extending from 1998 to 2023. Of these, 28 were included in the review [17-44] (Figure 4).

XSL•FO
RenderX

Figure 4. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the search procedure for randomized controlled trials where remote patient monitoring was used as an intervention to treat or improve acute exacerbations of chronic obstructive pulmonary disease.

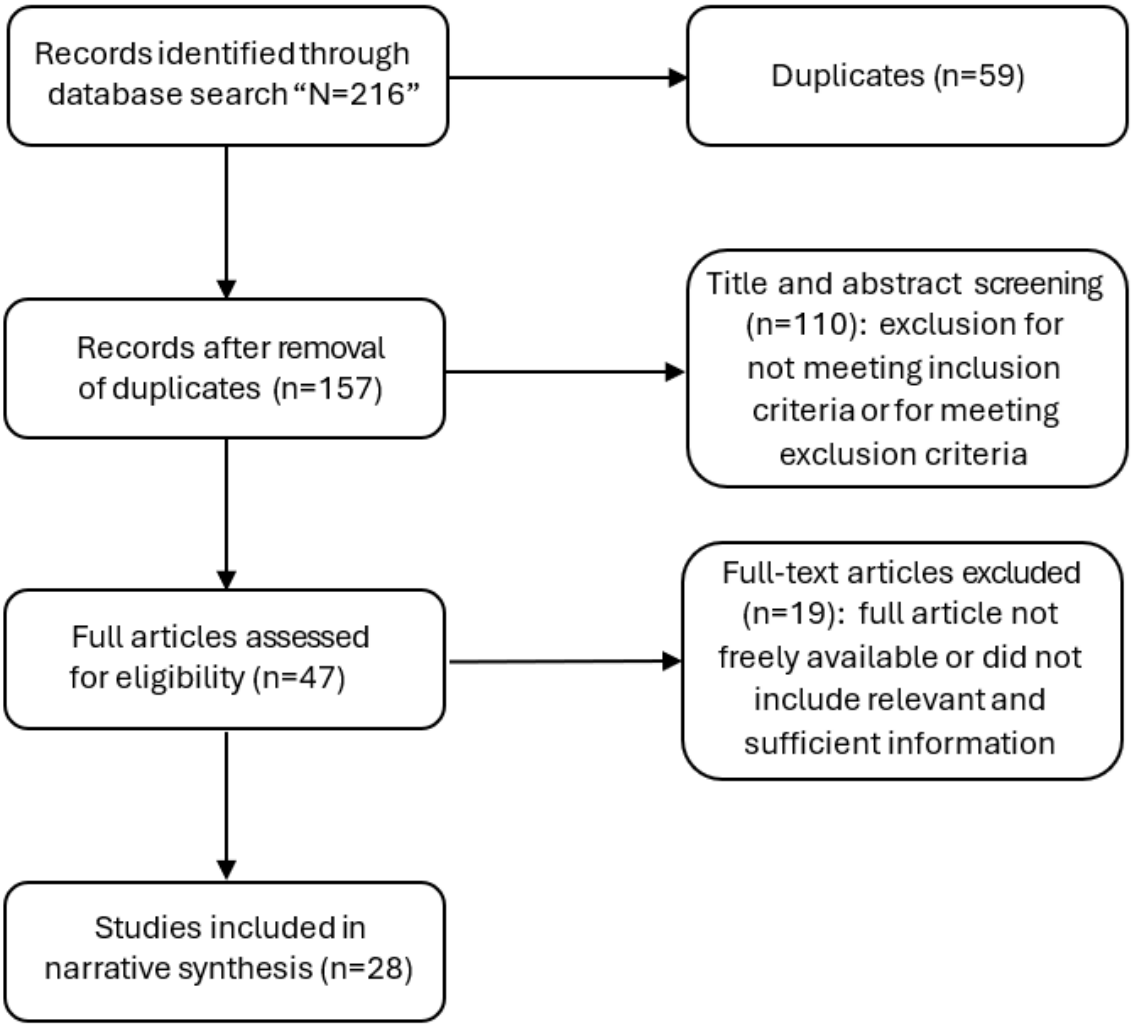

To identify the distribution and duration of studies and RPM indices identified in our first search, we visualized the data using a dual-axis plot (Figure 2). On the x-axis, each RPM index is represented. The left y-axis (blue bar chart) indicates the percentage of studies in which each index appears, while the right y-axis (red points) shows the average duration of these studies. This visualization facilitates a comprehensive comparison between the prevalence and the duration of the studies associated with each RPM index.

We constructed a figure to present the outcomes and rankings derived from the CASP RCT checklist evaluation of the RCTs identified in our first search (Figure 3). The figure is divided into 2 columns, where the left column displays RCTs showing some improvement in clinical outcomes and the right column shows RCTs with limited or no improvement. Within each column, studies are ranked by the strength of evidence, from limited evidence at the bottom to the strongest evidence at the top. Each section indicating evidence strength includes the number of RCTs, with the specific studies cited below this label.

A detailed breakdown of the RCT characteristics, method of intervention delivery, and outcomes are available in Multimedia Appendix 3.

The second systematic literature review search identified 350 articles, extending from 2005 to 2023. Of these, 23 were included in the review [45-68] (Figure 5).

XSL•FO
RenderX

Figure 5. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the search procedure for empirical studies on remote patient monitoring and machine learning to predict acute exacerbations of chronic obstructive pulmonary disease.

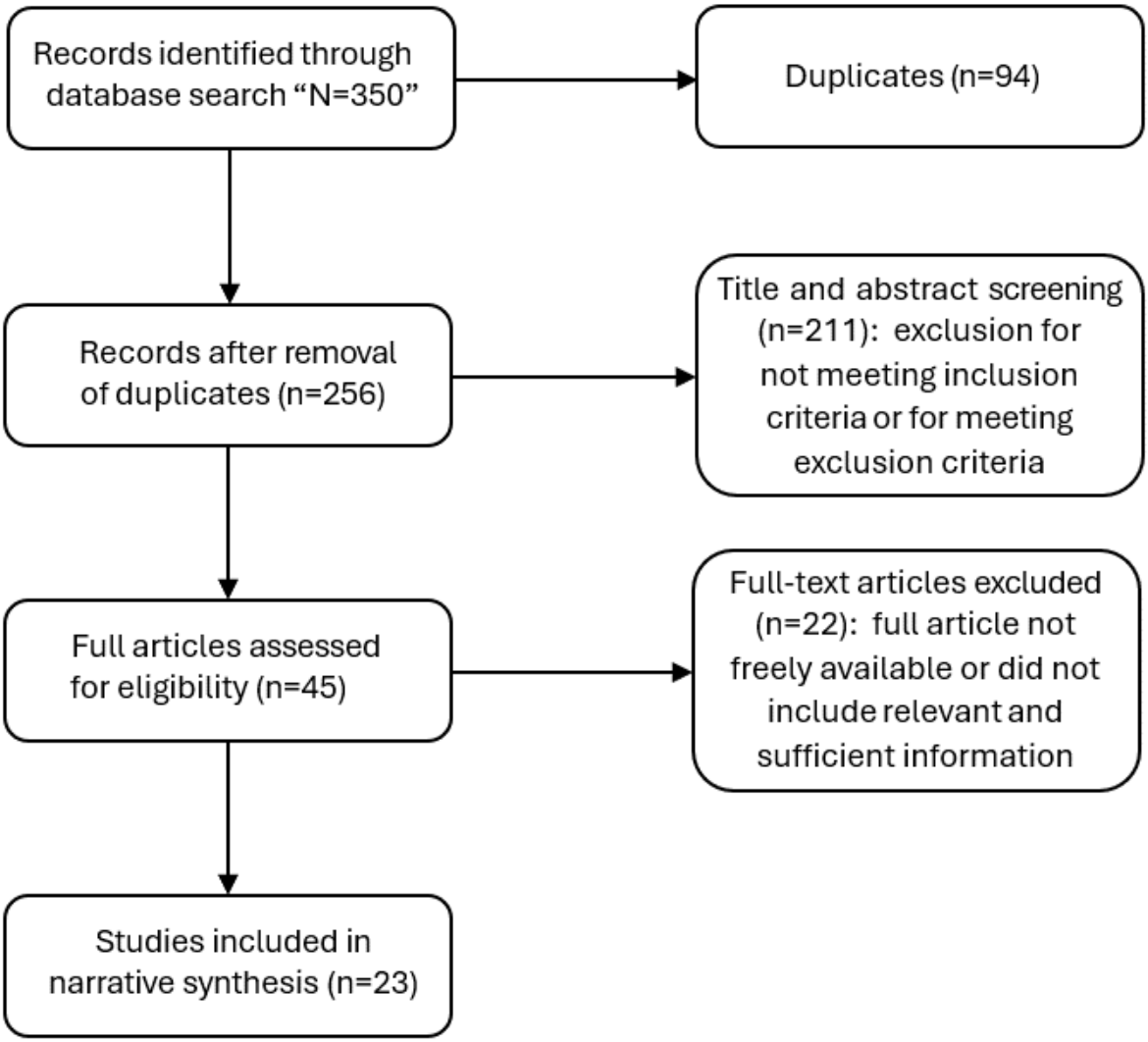

To identify the distribution and duration of studies and RPM indices identified in our second search, we visualized the data using a dual-axis plot (Figure 2). On the x-axis, each RPM index is represented. The left y-axis (blue bar chart) indicates the percentage of studies in which each index appears, whereas the right y-axis (red points) shows the average duration of these studies. This visualization facilitates a comprehensive comparison between the prevalence and the duration of the studies associated with each RPM index.

The studies identified in the second search were ranked according to the CASP cohort checklist. The ranking is categorized by the quality of the study as follows: strongest (studies [52,58-61]), strong (studies [51,54,55,57,64-68]), moderate (studies [48-50,53,56]), and limited (studies [45-47,62,63]). This categorization helps in clearly delineating the quality across the identified studies. A detailed breakdown of the machine learning study characteristics, method of intervention delivery, and outcomes are available in the Multimedia Appendix 3.

RPM for Intervention or Management of AECOPD

Our analysis revealed a tendency among RCTs toward nonimprovement or demonstrated a lack of statistical significance in indicators of improvement in exacerbation management (Figure 3).

Systematic reviews by both Jang et al [69] and Kruse et al [70] identified that RPM often does not show a significant improvement in patient outcomes. This may stem from the diversity of study approaches, as emphasized by Kruse et al [70]: "High variability between the articles and the ways they provided telemonitoring services created conflicting results from the literature review." The varied study designs make comparisons challenging and likely contribute to the diverse reported success levels in the literature. Furthermore, the burden of frequent monitoring may impact the effectiveness of RPM, as the commonly used approaches are often resource-intensive and burdensome for both patients and clinicians, thereby hindering widespread adoption.

XSL•FO
RenderX

RPM and Machine Learning for AECOPD Prediction

Machine learning, a subset of artificial intelligence, centers on leveraging data and algorithms to enable mathematical models for learning without direct instruction. In the context of predicting AECOPD, most studies using machine learning concentrate on assessing accuracy and related performance metrics based on retrospective data. This stands in contrast to efforts using RPM exclusively, which primarily aim at identifying improvements in patient outcomes through RCT. This makes the comparison of these 2 interrelated fields of research challenging, especially with regard to the indication of improving health outcomes for patients. Nevertheless, the narrative synthesis revealed that machine learning in conjunction with RPM facilitates highly accurate AECOPD prediction (Multimedia Appendix 3).

Orchard et al [59] compared machine learning with traditional symptom-counting algorithms in RPM for COPD exacerbation risk [59]. They found that machine learning "outperforms existing predictive algorithms." On a data set from 135 patients over 363 days, basic symptom-counting algorithms had limited accuracy (area under the receiver operating characteristic curve [AUROC] of 0.60 and 0.58). Testing in real-world scenarios reduced the algorithms' performance to no better than random decision-making. Machine learning models achieved the best AUROC of 0.74, exposing traditional algorithms' notable shortcomings, including low accuracy and frequent false positives. This highlights machine learning's potential to significantly enhance RPM accuracy in identifying and predicting COPD exacerbations.

So far, 2 patient-facing studies have been completed using machine learning for AECOPD prediction and intervention. The first tool evaluated was Adaptive Computerized COPD Exacerbation Self-management Support (ACCESS) [61], which demonstrated 97.4% sensitivity, but lower specificity at 65.6%. In an RCT, ACCESS did not significantly impact weeks without exacerbations or hospital admissions [61]. The second tool evaluated was the COPD Predict app that used patient-reported outcomes (PRO), FEV1, and C-reactive protein (CRP), predicting AECOPD with 97.9% sensitivity and 84.0% specificity [65]. For 6 months, COPD Predict reduced hospitalizations by 98%, but lower specificity resulted in 458 false positives, potentially hindering implementation. Using blood CRP levels for diagnosis and prediction may be impractical in real-world scenarios due to cost and access, indicating the need for the development of low-cost at-home biomarker sensors. COPD Predict identified exacerbations earlier than clinician-defined episodes, while ACCESS required major symptoms for 2 consecutive days, potentially delaying timely intervention. Furthermore, the low positive predictive value of access may have affected the trust of the patient-user and reduced the intervention's effectiveness.

Machine Learning Approaches

A total of 5 papers used neural networks for AECOPD prediction [48,50,58,59,68]. Neural networks excel at classification and prediction tasks due to their ability to model complex nonlinear relationships and automatically learn relevant features from data. Furthermore, neural networks can use transfer learning for efficient knowledge reuse, exhibit robustness to noisy data, and generalize well to new, unseen examples.

Nunavath et al [58] used a recurrent neural network (RNN), a subtype of artificial neural networks, for exacerbation prediction [58]. RNNs, commonly used in ordinal or temporal problems, generally demand substantial data for optimal performance. In this study, with a data set of 96 patients and approximately 7300 records collected over 2 years, data augmentation was applied to expand the training data, allowing the application of RNN, potentially resulting in a more potent, robust model for accuracy. van der Heijden et al [46], used Bayesian network algorithms to predict COPD exacerbations, offering advantages in incorporating previous knowledge, handling missing data through probabilistic inference, and providing interpretability by representing dependencies among variables. These models, known for their interpretability, are well-suited for dynamic environments like RPM. In studies by Fernandez-Granero et al [51,56], different machine learning methodologies were used in 2015 and 2017. The 2017 analysis, using a random forest (RF) classifier, demonstrated the benefits of ensemble learning and reduced overfitting compared with the support vector machine (SVM) model in 2015. This analysis also used the Markov chain Monte Carlo method for imputing missing data and feature subset selection, contributing to improved model performance. In addition, logistic regression was used by Shah et al [55] and Kronborg et al [57]. The models created in the study by Shah et al [55] encountered a notable trade-off between sensitivity and specificity (possibly stemming from challenges in data quality), and the model in the study by Kronborg et al [57] achieved an AUROC of 0.74, suggesting logistic regression limitations in capturing complex nonlinear relationships. Researchers should consider diverse approaches, acknowledging their strengths and limitations. Ensemble methods enhance robustness, Bayesian Networks offer interpretability, and techniques such as bootstrapping and data augmentation can improve data quality.

Study Design

The studies identified in systematic searches exhibit heterogeneity in duration, sample size, and outcomes, influencing both RPM interventions and machine learning model development and testing.

Study durations vary from 3 months to 1 year, impacting AECOPD events due to seasonality. Machine learning interventions may face challenges in generalizing beyond specific months. Small studies with 10-30 participants may restrict the capture of exacerbations and hinder model generalizability to larger populations. While recruiting frequent exacerbators can help, it may lead to more false positives in less exacerbation-prone populations. Furthermore, smaller samples will limit the diversity of age, sex, ethnicity, socioeconomic status, disease severity, exacerbation frequency, and time since diagnosis. Diverse participants are crucial to validate intervention effectiveness.

Many studies use machine learning with data from interventional studies or RCTs to develop predictive models. The success of these studies is typically gauged by the accuracy and predictive

XSL•FO
RenderX

capability of the generated models. While this is a crucial initial step to evaluate exacerbation prediction potential, relying solely on machine learning from RCT conditions may restrict applicability to real-world settings and diminish model accuracy. RCTs, conducted in controlled environments with specific criteria, might not fully capture real-world variability and complexity. Despite these difficulties, RCTs offer advantages, particularly in COPD, where a gold-standard diagnostic test for AECOPD is lacking. RCT data often involve rigorous testing, enhancing the certainty of accurately modeling true AECOPD events.

Defining and labeling exacerbations is a complex aspect of predicting AECOPD. It involves determining what qualifies as an exacerbation in the data, which presents significant challenges. Various methods exist for defining exacerbations, including patient-reported symptoms, clinician diagnosis, medication usage, hospitalization, or a combination of these criteria. Each approach has its advantages and drawbacks. Patient-reported symptoms and medication use are categorized as "symptom-based," capturing all exacerbations efficiently but lacking oversight and verification with objective measures. Clinician diagnosis and hospitalization fall into the "event-based" category, with the former possibly being symptom-based if verified remotely. While clinician diagnosis is the gold standard, it can be resource-intensive. An alternative is to use hospitalization, which offers a clinician diagnosis, but may overlook milder events managed at home.

The challenge in labeling exacerbations lies in accurately distinguishing between exacerbation events and periods of stable health during machine learning model training. Failure to do so prevents the identification of the prodromal period and may lead to the misclassification of moments of symptom relief as stable health rather than ongoing exacerbations. To address this, an algorithmic approach is necessary, involving the delineation of data windows. Typically, these windows are created by capturing 14 days of exacerbation-free periods, a duration commonly recognized as the time it takes for an exacerbation to develop after the onset of symptoms. Failure to implement an effective label indicates an AECOPD prediction model may be ineffective in real-world applications.

Patient Factors

Patient engagement and monitoring burden are crucial in RPM. Many studies in this review rely on the daily use of multiple sensors and PRO, yet still show high engagement. However, this may be influenced by the controlled research setting, providing extra attention and support, and motivating participants to engage. These studies often include rigorous monitoring and alerts for missed days. To assess real-world patient engagement, long-term data-gathering studies are often needed.

The burden of daily monitoring is frequently overlooked in interventions and model development studies, despite the requirement for high adherence. Concerns arise about time and effort, impacting daily life. Daily self-monitoring involves tracking symptoms, vital signs, medication adherence, and lifestyle factors, that potentially cause anxiety and stress. Some patients, with comorbidities or limited digital access, may struggle with devices or apps, leading to frustration and nonadherence. Patient and public involvement research is needed to understand user burden, develop strategies, and identify sensor types, minimizing burdens for increased long-term RPM adoption.

A Review of Indices in RPM

The main types of data collected are physiological measures (blood pressure, heart rate [HR] or pulse rate [PR], respiratory rate [RR], weight, oxygen saturation [SpO_2], and temperature), functional measures (lung function and physical activity [PA]), PRO (dyspnea, sputum, sleep quality, depression, anxiety, and HRQoL), self-report (physiological measures, medication-usage, exacerbation history, demographics, and medical history), and meteorological data.

Physiological Measures

In a systematic review, Buekers et al [71] identified 71 papers measuring SpO_2 in patients with COPD before exacerbations, revealing predictive limitations due to scant information on implementation and performance. Milkowska-Dymanowska et al [72] noted a significant SpO_2 decrease before exacerbation, distinguishing it from systolic blood pressure, diastolic blood pressure, or HR. The findings by Shah et al [55] indicated that SpO_2 (AUROC=0.658) outperformed PR (AUROC=0.578) in predicting exacerbations, highlighting SpO_2's superiority, while systolic blood pressure and HR and PR might offer some predictive capability. Mohktar et al [52] emphasized weight as an essential feature for AECOPD prediction. Dinesen et al [21] conducted an RCT that incorporated weight changes for clinical monitoring. Their intervention reduced hospitalization, affirming weight's role in early AECOPD detection. van der Heijden et al [73] established a dependency between exacerbation and body temperature. RCT studies by Ho et al [30] and Pedone et al [24] reported positive patient outcomes using body temperature as an index for exacerbation monitoring. Shah et al [55] found that mean RR increased by 2 in the prodromal period leading to an exacerbation. However, a feasibility study by Chau et al [74], using RR to monitor patients, did not demonstrate improvements in patient outcomes, possibly due to inherent study design limitations. Burton et al [75] observed a mean SpO_2 drop from 93.6% to 92.4% at exacerbation onset, and Shah et al [55] found a decrease from 94% to 93% in the prodromal period, enhancing AECOPD predictive models. Despite these findings, Burton et al [75] identified weak associations between physiological variables and exacerbation episodes, underscoring the need for improved algorithms or additional features for early event detection.

Functional Measures

FEV1 serves as a crucial measure captured by a spirometer for diagnosing and monitoring obstructive lung diseases. Digital spirometry holds the potential to be a potent predictor of AECOPD. Limited data on lung function deterioration shortly before exacerbations exist, but Watz et al [76] through their post hoc analysis observed FEV1 decline 2 weeks before an exacerbation in the WISDOM (Withdrawal of Inhaled Steroids during Optimized Bronchodilator Management) clinical trial, highlighting spirometry's use in identifying exacerbations. Patel

XSL•FO
RenderX

et al [65] demonstrated spirometry's predictive capability, achieving a sensitivity of 97.9% and specificity of 84.0% for AECOPD when combined with CRP and PRO.

PA, although infrequently used, offers a burdenless monitoring approach. Pedone et al [24], in their RCT, incorporating PA as a measure, showed a significant reduction in exacerbation events. Chawla et al [77] found that lower PA in the first week after discharge increased the likelihood of 30-day all-cause readmissions. Wrist-worn wearables measuring PA, prevalent through smartwatches, provide a useful, unobtrusive tool for monitoring for exacerbation prediction. While not a direct alternative to spirometry for lung function, they offer continuous monitoring without user engagement.

PRO Measures

PRO are typically obtained through patient inputs into a digital diary, involving daily responses to yes-or-no or graded questions. Graded inquiries may include assessing chest tightness on a scale of 0 to 5. The rationale for incorporating PRO in RPM is robust. The chronic obstructive pulmonary disease assessment test (CAT), an 8-question validated tool [78], shows a positive correlation with COPD exacerbation risk [79]. A one-unit increase in CAT score signals an 8% higher risk of exacerbation [80]. Similar associations exist for other PRO measures, such as the Modified Medical Research Council dyspnea scale [11]. Most of the RCTs included in this review commonly use PRO as part of RPM. However, many of these studies do not show significant improvements in exacerbation-related outcomes. This might be attributed to the frequent use of nonvalidated, study-specific questionnaires for assessing PRO. To potentially enhance outcomes, a recommended shift would be from using nonvalidated questionnaires to prioritizing validated measures, such as the CAT.

Biological Measures

Exacerbations in COPD correlate with various biomarkers, with CRP being extensively studied [81-84]. Patel et al [65] showcased CRP's efficacy in COPD Predict, emphasizing its high sensitivity and specificity. However, limited CRP use in prediction models stems from challenges in deploying widespread systems with frequent point-of-care testing. Potential solutions include at-home finger-prick blood sampling, mail-in samples, or self-administered point-of-care tests, such as lateral flow tests used during the SARS-CoV-2 pandemic. Inflammatory cytokines in sputum and volatile organic compounds in exhaled breath are also targets for at-home monitoring, contingent on the development of suitable devices [85,86].

Despite the potential benefits, implementing remote biological monitoring encounters challenges in cost and scalability. Substantial expenses for developing and deploying home-based detection equipment, ensuring data accuracy, and managing sensor development, deployment, data transmission, and storage impede practical execution. Achieving scalability for costly digital biological monitoring technology demands a significant investment.

Discussion

Principal Findings

We have undertaken a systematic literature review and narrative synthesis of RPM and machine learning for COPD exacerbations to address the problem of identifying better ways to intervene early or prevent exacerbations to improve outcomes in AECOPD.

The narrative synthesis of available evidence reveals that RCTs using RPM to monitor patients at risk of AECOPD tend to exhibit nonsignificant outcomes. This can be attributed to the heterogeneity in study design and the use of traditional symptom-counting algorithms. In addition, there is limited insight into the accuracy or timeliness of AECOPD detection and intervention. Orchard et al [59] point to limitations relating to the basic algorithms that seem to generate alerts from RPM that may perform no better than chance [60] and highlights the potential of machine learning to significantly enhance the predictive capabilities of RPM.

We chose not to exclusively limit our second search to RCTs due to the lack of available RCTs in the relevant literature. The comparison of these 2 bodies of literature presents a challenge, and instead of considering them as comparators, it is more apt to view them as conjugates. While similarities exist in the methodologies of the 2 bodies of literature, particularly in the RPM indices, there are significant divergences in their outcomes. The first search primarily focuses on patient outcomes, whereas the second centers on the performance of machine learning models. This contrast underscores the importance of a shift in the trajectory of research: RCTs of RPM in AECOPD should introduce machine learning models to identify their efficacy and draw a conclusion to the question if early detection of AECOPD can improve patient outcomes. While early research for machine learning in the prediction of AECOPD is promising, there is much need for further development in the field. Much of the literature evaluates the ability of machine learning models to predict AECOPD yet fails to provide evidence on how these models translate into improving patient outcomes. Notable exceptions include COPD Predict developed by Patel et al [65], which may reduce hospitalization [66], but it requires frequent blood testing and the use of expensive home-based detection equipment. Due to the lack of research in this field, it is too early to determine if exacerbation prediction and intervention can be deployed with machine learning to improve AECOPD outcomes, especially without the incorporation of biological indices.

Some adjustments can be made in future studies that may lead to more robust conclusions. For example, exacerbation definition and labeling are heterogeneous and would benefit from standardization, it is critical to ensure that there is sufficient data capture by increasing the number of participants and keeping the study length to a minimum of 1 year should result in better generalization of predictive models. In addition, integrating data augmentation, resampling, and feature selection techniques can further enhance training data for machine learning. Furthermore, incorporating neural networks into a modeling approach may greatly enhance the predictive power

XSL•FO
RenderX

of AECOPD models. Clinically validated PRO, like that of CAT, should be considered as an alternative to study-specific questionnaires. SpO_2, RR, HR, weight, FEV1, and peak expiratory flow appear frequently in remotely monitored indices and may have the potential for exacerbation prediction if combined with machine learning. In addition, further research could demonstrate the predictive capabilities of physical activity. Patient and public involvement research needs to be conducted with these sensors to explore the feasibility and likelihood of long-term adherence. The incorporation of sensing technologies that monitor in the background and are seemingly burdenless (wrist-worn wearables or smartwatches) should be considered. The use of biological measures is a rarely used method in RPM but research, as seen from the study by Patel et al [65], is needed to develop cost-effective tools for predicting AECOPD.

Limitations

This review has some limitations. The number of studies initially identified in searches is relatively small and limited to those full and freely available in English. This is due to a combination of stringent search terms to identify RCT and machine learning papers and the strict inclusion and exclusion criteria. This may influence our confidence in drawing conclusions regarding the efficacy or accuracy of RPM and machine learning in COPD or, international variations, or those not freely available cannot be reported. However, clearly defined criteria are necessary to minimize bias and subjectivity in the search and enable reproducibility and rigor that add validity to a systematic review. This study did not include a formal risk of bias assessment. However, evaluating the studies using CASP checklists and ranking them should highlight the quality of the studies included in this dual systematic review.

Recommendations

The methodologies for exacerbation intervention using RPM exhibit inconsistency, potentially contributing to the variability in the success of outcomes. The current landscape is hindered by inadequate evidence and the substantial resource burden associated with clinical oversight of remotely monitored indices. Therefore, it becomes apparent that a shift away from conventional intervention methods is warranted, with a compelling case for exploring the integration of machine learning approaches as a more pragmatic and potentially effective alternative.

While the literature emphasizes the accuracy of machine learning in AECOPD prediction, demonstrating the clinical use of these approaches requires evidence from RCT and real-world studies, both of which are currently lacking.

Physiological and functional indices, such as SpO_2, RR, HR, weight, FEV1, and PEF, are commonly found in the literature and have the potential to serve as robust predictors of AECOPD when integrated with machine learning. Exploring the inclusion of PA as a predictive index is justified, and the incorporation of PRO, particularly through clinically validated questionnaires, is strongly recommended.

There is a lack of research into patient factors, there is a need to study the adoption of RPM in the long term and to develop an understanding of the burden of daily or weekly RPM and potential solutions to overcome this burden. Active engagement with patient communities is required to understand their needs, ensuring optimal responses and the best possible outcomes.

Conclusion

RPM has yet to consistently prove a successful intervention in AECOPD, facing significant challenges with accuracy in predicting and identifying exacerbations, resource-intensive processes, and scalability limitations. Although machine learning demonstrates the potential for high accuracy for AECOPD, its clinical use is still to be validated through RCTs. Considering the heterogeneity of COPD, a one-size-fits-all approach may not be suitable. Leveraging machine learning, an individualized approach tailored to each patient's unique data holds the potential to enhance patient outcomes in AECOPD.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e52143_app1.docx]

Multimedia Appendix 2
Search methods and strategy.
[DOCX File , 15 KB - jmir_v26i1e52143_app2.docx]

Multimedia Appendix 3
A detailed breakdown of the study characteristics, method of intervention delivery, and outcomes.
[DOCX File , 42 KB - jmir_v26i1e52143_app3.docx]

References

XSL•FO
RenderX

1. GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Respir Med 2020 Jun;8(6):585-596 [FREE Full text] [doi: 10.1016/S2213-2600(20)30105-3] [Medline: 32526187]
2. Ko FW, Chan KP, Hui DS, Goddard JR, Shaw JG, Reid DW, et al. Acute exacerbation of COPD. Respirology 2016 Oct;21(7):1152-1165 [FREE Full text] [doi: 10.1111/resp.12780] [Medline: 27028990]
3. Suissa S, Dell'Aniello S, Ernst P. Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. Thorax 2012 Nov;67(11):957-963 [FREE Full text] [doi: 10.1136/thoraxjnl-2011-201518] [Medline: 22684094]
4. Hoogendoorn M, Hoogenveen RT, Rutten-van Mölken MP, Vestbo J, Feenstra TL. Case fatality of COPD exacerbations: a meta-analysis and statistical modelling approach. Eur Respir J 2011 Mar;37(3):508-515 [FREE Full text] [doi: 10.1183/09031936.00043710] [Medline: 20595157]
5. Halpin DM, Miravitlles M, Metzdorf N, Celli B. Impact and prevention of severe exacerbations of COPD: a review of the evidence. Int J Chron Obstruct Pulmon Dis 2017;12:2891-2908 [FREE Full text] [doi: 10.2147/COPD.S139470] [Medline: 29062228]
6. Day K, Ostridge K, Conway J, Cellura D, Watson A, Spalluto CM, et al. Interrelationships among small airways dysfunction, neutrophilic inflammation, and exacerbation frequency in COPD. Chest 2021 Apr;159(4):1391-1399. [doi: 10.1016/j.chest.2020.11.018] [Medline: 33245876]
7. Perera WR, Hurst JR, Wilkinson TMA, Sapsford RJ, Müllerova H, Donaldson GC, et al. Inflammatory changes, recovery and recurrence at COPD exacerbation. Eur Respir J 2007 Mar;29(3):527-534 [FREE Full text] [doi: 10.1183/09031936.00092506] [Medline: 17107990]
8. Anzueto A. Impact of exacerbations on COPD. Eur Respir Rev 2010 Jun;19(116):113-118 [FREE Full text] [doi: 10.1183/09059180.00002610] [Medline: 20956179]
9. Kerkhof M, Voorham J, Dorinsky P, Cabrera C, Darken P, Kocks JW, et al. Association between COPD exacerbations and lung function decline during maintenance therapy. Thorax 2020 Sep;75(9):744-753 [FREE Full text] [doi: 10.1136/thoraxjnl-2019-214457] [Medline: 32532852]
10. Hurst JR, Skolnik N, Hansen GJ, Anzueto A, Donaldson GC, Dransfield MT, et al. Understanding the impact of chronic obstructive pulmonary disease exacerbations on patient health and quality of life. Eur J Intern Med 2020 Mar;73:1-6 [FREE Full text] [doi: 10.1016/j.ejim.2019.12.014] [Medline: 31954592]
11. Whittaker H, Rubino A, Müllerová H, Morris T, Varghese P, Xu Y, et al. Frequency and severity of exacerbations of COPD associated with future risk of exacerbations and mortality: a UK routine health care data study. Int J Chron Obstruct Pulmon Dis 2022;17:427-437 [FREE Full text] [doi: 10.2147/COPD.S346591] [Medline: 35264849]
12. de Nigris E, McEwan P, Marshall J, Holmgren U, Foos V. POSB65 economic burden of COPD in the United Kingdom (2021–2040) estimated with the COPD Health Outcome Policy and Intervention (CHOPIN) model. Value in Health 2022 Jan;25(1):S72-S73. [doi: 10.1016/j.jval.2021.11.338]
13. Iheanacho I, Zhang S, King D, Rizzo M, Ismaila AS. Economic burden of chronic obstructive pulmonary disease (COPD): a systematic literature review. Int J Chron Obstruct Pulmon Dis 2020;15:439-460 [FREE Full text] [doi: 10.2147/COPD.S234942] [Medline: 32161455]
14. Alqahtani JS, Aquilina J, Bafadhel M, Bolton CE, Burgoyne T, Holmes S, et al. Research priorities for exacerbations of COPD. Lancet Respir Med 2021 Aug;9(8):824-826. [doi: 10.1016/S2213-2600(21)00227-7] [Medline: 34000234]
15. Wilkinson TMA, Donaldson GC, Hurst JR, Seemungal TAR, Wedzicha JA. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2004 Jun 15;169(12):1298-1303. [doi: 10.1164/rccm.200310-1443OC] [Medline: 14990395]
16. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
17. de Toledo P, Jiménez S, del Pozo F, Roca J, Alonso A, Hernandez C. Telemedicine experience for chronic care in COPD. IEEE Trans Inf Technol Biomed 2006 Jul;10(3):567-573 [FREE Full text] [doi: 10.1109/titb.2005.863877] [Medline: 16871726]
18. Koff PB, Jones RH, Cashman JM, Voelkel NF, Vandivier RW. Proactive integrated care improves quality of life in patients with COPD. Eur Respir J 2009 May;33(5):1031-1038 [FREE Full text] [doi: 10.1183/09031936.00063108] [Medline: 19129289]
19. Vitacca M, Bianchi L, Guerra A, Fracchia C, Spanevello A, Balbi B, et al. Tele-assistance in chronic respiratory failure patients: a randomised clinical trial. Eur Respir J 2009 Feb;33(2):411-418 [FREE Full text] [doi: 10.1183/09031936.00005608] [Medline: 18799512]
20. Halpin DMG, Laing-Morton T, Spedding S, Levy ML, Coyle P, Lewis J, et al. A randomised controlled trial of the effect of automated interactive calling combined with a health risk forecast on frequency and severity of exacerbations of COPD assessed clinically and using EXACT PRO. Prim Care Respir J 2011 Sep;20(3):324-31, 2 p following 331 [FREE Full text] [doi: 10.4104/pcrj.2011.00057] [Medline: 21687919]

XSL•FO
RenderX

21. Dinesen B, Haesum LKE, Soerensen N, Nielsen C, Grann O, Hejlesen O, et al. Using preventive home monitoring to reduce hospital admission rates and reduce costs: a case study of telehealth among chronic obstructive pulmonary disease patients. J Telemed Telecare 2012 Jun;18(4):221-225. [doi: 10.1258/jtt.2012.110704] [Medline: 22653618]
22. De San Miguel K, Smith J, Lewin G. Telehealth remote monitoring for community-dwelling older adults with chronic obstructive pulmonary disease. Telemed J E Health 2013 Sep;19(9):652-657. [doi: 10.1089/tmj.2012.0244] [Medline: 23808885]
23. Jehn M, Donaldson G, Kiran B, Liebers U, Mueller K, Scherer D, et al. Tele-monitoring reduces exacerbation of COPD in the context of climate change--a randomized controlled trial. Environ Health 2013 Nov 21;12:99 [FREE Full text] [doi: 10.1186/1476-069X-12-99] [Medline: 24261700]
24. Pedone C, Chiurco D, Scarlata S, Incalzi RA. Efficacy of multiparametric telemonitoring on respiratory outcomes in elderly people with COPD: a randomized controlled trial. BMC Health Serv Res 2013 Mar 06;13:82 [FREE Full text] [doi: 10.1186/1472-6963-13-82] [Medline: 23497109]
25. Pinnock H, Hanley J, McCloughan L, Todd A, Krishan A, Lewis S, et al. Effectiveness of telemonitoring integrated into existing clinical services on hospital admission for exacerbation of chronic obstructive pulmonary disease: researcher blind, multicentre, randomised controlled trial. BMJ 2013 Oct 17;347:f6070 [FREE Full text] [doi: 10.1136/bmj.f6070] [Medline: 24136634]
26. Sorknaes AD, Bech M, Madsen H, Titlestad IL, Hounsgaard L, Hansen-Nord M, et al. The effect of real-time teleconsultations between hospital-based nurses and patients with severe COPD discharged after an exacerbation. J Telemed Telecare 2013 Dec;19(8):466-474. [doi: 10.1177/1357633X13512067] [Medline: 24227799]
27. McDowell JE, McClean S, FitzGibbon F, Tate S. A randomised clinical trial of the effectiveness of home-based health care with telemonitoring in patients with COPD. J Telemed Telecare 2015 Mar;21(2):80-87. [doi: 10.1177/1357633X14566575] [Medline: 25586812]
28. Ringbæk T, Green A, Laursen LC, Frausing E, Brøndum E, Ulrik CS. Effect of tele health care on exacerbations and hospital admissions in patients with chronic obstructive pulmonary disease: a randomized clinical trial. Int J Chron Obstruct Pulmon Dis 2015;10:1801-1808 [FREE Full text] [doi: 10.2147/COPD.S85596] [Medline: 26366072]
29. Cordova FC, Ciccolella D, Grabianowski C, Gaughan J, Brennan K, Goldstein F, et al. A telemedicine-based intervention reduces the frequency and severity of COPD exacerbation symptoms: a randomized, controlled trial. Telemed J E Health 2016 Feb;22(2):114-122 [FREE Full text] [doi: 10.1089/tmj.2015.0035] [Medline: 26259074]
30. Ho T, Huang C, Chiu H, Ruan S, Tsai Y, Yu C, et al. Effectiveness of telemonitoring in patients with chronic obstructive pulmonary disease in Taiwan-a randomized controlled trial. Sci Rep 2016 Mar 31;6:23797 [FREE Full text] [doi: 10.1038/srep23797] [Medline: 27029815]
31. Vianello A, Fusello M, Gubian L, Rinaldo C, Dario C, Concas A, et al. Home telemonitoring for patients with acute exacerbation of chronic obstructive pulmonary disease: a randomized controlled trial. BMC Pulm Med 2016 Nov 22;16(1):157 [FREE Full text] [doi: 10.1186/s12890-016-0321-2] [Medline: 27876029]
32. Farmer A, Williams V, Velardo C, Shah SA, Yu L, Rutter H, et al. Self-management support using a digital health system compared with usual care for chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2017 May 03;19(5):e144 [FREE Full text] [doi: 10.2196/jmir.7116] [Medline: 28468749]
33. Lilholt PH, Witt Udsen F, Ehlers L, Hejlesen OK. Telehealthcare for patients suffering from chronic obstructive pulmonary disease: effects on health-related quality of life: results from the Danish 'TeleCare North' cluster-randomised trial. BMJ Open 2017 May 09;7(5):e014587 [FREE Full text] [doi: 10.1136/bmjopen-2016-014587] [Medline: 28490555]
34. Rixon L, Hirani SP, Cartwright M, Beynon M, Doll H, Steventon A, et al. A RCT of telehealth for COPD patient's quality of life: the whole system demonstrator evaluation. Clin Respir J 2017 Jul;11(4):459-469. [doi: 10.1111/crj.12359] [Medline: 26260325]
35. Kessler R, Casan-Clara P, Koehler D, Tognella S, Viejo JL, Dal Negro RW, et al. COMET: a multicomponent home-based disease-management programme routine care in severe COPD. Eur Respir J 2018 Jan;51(1):1701612 [FREE Full text] [doi: 10.1183/13993003.01612-2017] [Medline: 29326333]
36. Soriano JB, García-Río F, Vázquez-Espinosa E, Conforto JI, Hernando-Sanz A, López-Yepes L, et al. A multicentre, randomized controlled trial of telehealth for the management of COPD. Respir Med 2018 Nov;144:74-81 [FREE Full text] [doi: 10.1016/j.rmed.2018.10.008] [Medline: 30366588]
37. Tupper OD, Gregersen TL, Ringbaek T, Brøndum E, Frausing E, Green A, et al. Effect of tele-health care on quality of life in patients with severe COPD: a randomized clinical trial. Int J Chron Obstruct Pulmon Dis 2018;13:2657-2662 [FREE Full text] [doi: 10.2147/COPD.S164121] [Medline: 30214183]
38. Walker PP, Pompilio PP, Zanaboni P, Bergmo TS, Prikk K, Malinovschi A, et al. Telemonitoring in Chronic Obstructive Pulmonary Disease (CHROMED). a randomized clinical trial. Am J Respir Crit Care Med 2018 Sep 01;198(5):620-628. [doi: 10.1164/rccm.201712-2404OC] [Medline: 29557669]
39. Mínguez Clemente P, Pascual-Carrasco M, Mata Hernández C, Malo de Molina R, Arvelo LA, Cadavid B, et al. Follow-up with telemedicine in early discharge for COPD exacerbations: randomized clinical trial (TELEMEDCOPD-Trial). COPD 2021 Feb;18(1):62-69. [doi: 10.1080/15412555.2020.1857717] [Medline: 33307857]

XSL•FO
RenderX

40. Sink E, Patel K, Groenendyk J, Peters R, Som A, Kim E, et al. Effectiveness of a novel, automated telephone intervention on time to hospitalisation in patients with COPD: a randomised controlled trial. J Telemed Telecare 2020 Apr;26(3):132-139. [doi: 10.1177/1357633X18800211] [Medline: 30269640]
41. Koff PB, Min S, Freitag TJ, Diaz DLP, James SS, Voelkel NF, et al. Impact of proactive integrated care on chronic obstructive pulmonary disease. Chronic Obstr Pulm Dis 2021 Jan;8(1):100-116 [FREE Full text] [doi: 10.15326/jcopdf.2020.0139] [Medline: 33238087]
42. Rassouli F, Germann A, Baty F, Kohler M, Stolz D, Thurnheer R, et al. Telehealth mitigates COPD disease progression compared to standard of care: a randomized controlled crossover trial. J Intern Med 2021 Mar;289(3):404-410 [FREE Full text] [doi: 10.1111/joim.13230] [Medline: 33428219]
43. Andersen FD, Trolle C, Pedersen AR, Køpfli ML, Børgesen S, Jensen MS, et al. Effect of telemonitoring on readmissions for acute exacerbation of chronic obstructive pulmonary disease: a randomized clinical trial. J Telemed Telecare 2023 Jan 23:1357633X221150279. [doi: 10.1177/1357633X221150279] [Medline: 36683440]
44. Køpfli ML, Børgesen S, Jensen MS, Hyldgaard C, Bell C, Andersen FD. Effect of telemonitoring on quality of life for patients with chronic obstructive pulmonary disease-a randomized controlled trial. Chron Respir Dis 2023;20:14799731231157771 [FREE Full text] [doi: 10.1177/14799731231157771] [Medline: 36775280]
45. Jensen MH, Cichosz SL, Dinesen B, Hejlesen OK. Moving prediction of exacerbation in chronic obstructive pulmonary disease for patients in telecare. J Telemed Telecare 2012 Mar;18(2):99-103. [doi: 10.1258/jtt.2011.110607] [Medline: 22267305]
46. van der Heijden M, Lucas PJ, Lijnse B, Heijdra YF, Schermer TRJ. An autonomous mobile system for the management of COPD. J Biomed Inform 2013 Jun;46(3):458-469 [FREE Full text] [doi: 10.1016/j.jbi.2013.03.003] [Medline: 23500485]
47. Bellos CC, Papadopoulos A, Rosso R, Fotiadis DI. Identification of COPD patients' health status using an intelligent system in the CHRONIOUS wearable platform. IEEE J Biomed Health Inform 2014 May;18(3):731-738. [doi: 10.1109/JBHI.2013.2293172] [Medline: 24808219]
48. Fernández-Granero MA, Sánchez-Morillo D, León-Jiménez A, Crespo LF. Automatic prediction of chronic obstructive pulmonary disease exacerbations through home telemonitoring of symptoms. Biomed Mater Eng 2014;24(6):3825-3832. [doi: 10.3233/BME-141212] [Medline: 25227099]
49. Shah S, Velardo C, Gibson O, Rutter H, Farmer A, Tarassenko L. Personalized alerts for patients with COPD using pulse oximetry and symptom scores. In: Annu Int Conf IEEE Eng Med Biol Soc. 2014 Presented at: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2014 August 26-30; Chicago, IL, USA p. 3164-3167. [doi: 10.1109/EMBC.2014.6944294]
50. Fernandez-Granero MA, Sanchez-Morillo D, Lopez MA, León-Jiménez A. A machine learning approach to prediction of exacerbations of chronic obstructive pulmonary disease. In; 2015 Presented at: International Work-Conference on the Interplay Between Natural and Artificial Computation; 2015 May; Switzerland p. 305-311. [doi: 10.1007/978-3-319-18914-7_32]
51. Fernandez-Granero M, Sanchez-Morillo D, Leon-Jimenez A. Computerised analysis of telemonitored respiratory sounds for predicting acute exacerbations of COPD. Sensors (Basel) 2015 Oct 23;15(10):26978-26996 [FREE Full text] [doi: 10.3390/s151026978] [Medline: 26512667]
52. Mohktar MS, Redmond SJ, Antoniades NC, Rochford PD, Pretto JJ, Basilakis J, et al. Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. Artif Intell Med 2015 Jan;63(1):51-59. [doi: 10.1016/j.artmed.2014.12.003] [Medline: 25704112]
53. Sanchez-Morillo D, Fernandez-Granero MA, Jiménez AL. Detecting COPD exacerbations early using daily telemonitoring of symptoms and k-means clustering: a pilot study. Med Biol Eng Comput 2015 May;53(5):441-451. [doi: 10.1007/s11517-015-1252-4] [Medline: 25725628]
54. Riis HC, Jensen MH, Cichosz SL, Hejlesen OK. Prediction of exacerbation onset in chronic obstructive pulmonary disease patients. J Med Eng Technol 2016;40(1):1-7. [doi: 10.3109/03091902.2015.1105317] [Medline: 26745746]
55. Shah SA, Velardo C, Farmer A, Tarassenko L. Exacerbations in chronic obstructive pulmonary disease: identification and prediction using a digital health system. J Med Internet Res 2017 Mar 07;19(3):e69 [FREE Full text] [doi: 10.2196/jmir.7207] [Medline: 28270380]
56. Fernandez-Granero MA, Sanchez-Morillo D, Leon-Jimenez A. An artificial intelligence approach to early predict symptom-based exacerbations of COPD. Biotechnology & Biotechnological Equipment 2018 Feb 10;32(3):778-784. [doi: 10.1080/13102818.2018.1437568]
57. Kronborg T, Mark L, Cichosz SL, Secher PH, Hejlesen O. Population exacerbation incidence contains predictive information of acute exacerbations in patients with chronic obstructive pulmonary disease in telecare. Int J Med Inform 2018 Mar;111:72-76. [doi: 10.1016/j.ijmedinf.2017.12.026] [Medline: 29425637]
58. Nunavath V, Goodwin M, Fidje JT, Moe CE. Deep neural networks for prediction of exacerbations of patients with chronic obstructive pulmonary disease. 2018 Presented at: Engineering Applications of Neural Networks: 19th International Conference, EANN 2018; 2018 September 3-5; Bristol, UK p. 3-5. [doi: 10.1007/978-3-319-98204-5_18]

XSL•FO
RenderX

59. Orchard P, Agakova A, Pinnock H, Burton CD, Sarran C, Agakov F, et al. Improving prediction of risk of hospital admission in chronic obstructive pulmonary disease: application of machine learning to telemonitoring data. J Med Internet Res 2018 Sep 21;20(9):e263 [FREE Full text] [doi: 10.2196/jmir.9227] [Medline: 30249589]
60. Boer LM, van der Heijden M, van Kuijk NM, Lucas PJ, Vercoulen JH, Assendelft WJ, et al. Validation of ACCESS: an automated tool to support self-management of COPD exacerbations. Int J Chron Obstruct Pulmon Dis 2018;13:3255-3267 [FREE Full text] [doi: 10.2147/COPD.S167272] [Medline: 30349231]
61. Boer L, Bischoff E, van der Heijden M, Lucas P, Akkermans R, Vercoulen J, et al. A smart mobile health tool versus a paper action plan to support self-management of chronic obstructive pulmonary disease exacerbations: randomized controlled trial. JMIR Mhealth Uhealth 2019 Oct 09;7(10):e14408 [FREE Full text] [doi: 10.2196/14408] [Medline: 31599729]
62. Jin Y, Zhang T, Cao Z, Zhao N, Chen C, Wang D, et al. Prediction indicators for acute exacerbations of chronic obstructive pulmonary disease by combining non-linear analyses and machinend Machine. 2018 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018 December 03-06; Madrid, Spain p. 2515-2521. [doi: 10.1109/BIBM.2018.8621430]
63. Iadanza E, Mudura V, Melillo P, Gherardelli M. An automatic system supporting clinical decision for chronic obstructive pulmonary disease. Health Technol 2019 Mar 16;10(2):487-498. [doi: 10.1007/s12553-019-00312-9]
64. Kronborg T, Hangaard S, Cichosz SL, Hejlesen O. A two-layer probabilistic model to predict COPD exacerbations for patients in telehealth. Comput Biol Med 2021 Jan;128:104108. [doi: 10.1016/j.compbiomed.2020.104108] [Medline: 33190010]
65. Patel N, Kinmond K, Jones P, Birks P, Spiteri MA. Validation of COPDPredict™: unique combination of remote monitoring and exacerbation prediction to support preventative management of COPD exacerbations. Int J Chron Obstruct Pulmon Dis 2021;16:1887-1899 [FREE Full text] [doi: 10.2147/COPD.S309372] [Medline: 34188465]
66. Wu C, Li G, Huang C, Cheng Y, Chen C, Chien J, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth 2021 May 06;9(5):e22591 [FREE Full text] [doi: 10.2196/22591] [Medline: 33955840]
67. Chmiel FP, Burns DK, Pickering JB, Blythin A, Wilkinson TM, Boniface MJ. Prediction of chronic obstructive pulmonary disease exacerbation events by using patient self-reported data in a digital health app: statistical evaluation and machine learning approach. JMIR Med Inform 2022 Mar 21;10(3):e26499 [FREE Full text] [doi: 10.2196/26499] [Medline: 35311685]
68. Wu C, Wang S, Su Y, Hsieh T, Chen P, Cheng Y, et al. A precision health service for chronic diseases: development and cohort study using wearable device, machine learning, and deep learning. IEEE J Transl Eng Health Med 2022;10:2700414 [FREE Full text] [doi: 10.1109/JTEHM.2022.3207825] [Medline: 36199984]
69. Jang S, Kim Y, Cho W. A systematic review and meta-analysis of telemonitoring interventions on severe COPD exacerbations. Int J Environ Res Public Health 2021 Jun 23;18(13):6757 [FREE Full text] [doi: 10.3390/ijerph18136757] [Medline: 34201762]
70. Kruse C, Pesek B, Anderson M, Brennan K, Comfort H. Telemonitoring to manage chronic obstructive pulmonary disease: systematic literature review. JMIR Med Inform 2019 Mar 20;7(1):e11496 [FREE Full text] [doi: 10.2196/11496] [Medline: 30892276]
71. Buekers J, De Boever P, Vaes AW, Aerts J, Wouters EFM, Spruit MA, et al. Oxygen saturation measurements in telemonitoring of patients with COPD: a systematic review. Expert Rev Respir Med 2018 Feb;12(2):113-123. [doi: 10.1080/17476348.2018.1417842] [Medline: 29241369]
72. Miłkowska-Dymanowska J, Białas AJ, Obrębski W, Górski P, Piotrowski WJ. A pilot study of daily telemonitoring to predict acute exacerbation in chronic obstructive pulmonary disease. Int J Med Inform 2018 Aug;116:46-51 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.04.013] [Medline: 29887234]
73. van der Heijden M, Velikova M, Lucas PJ. Learning Bayesian networks for clinical time series analysis. J Biomed Inform 2014 Apr;48:94-105 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.007] [Medline: 24361389]
74. Chau JP, Lee DT, Yu DS, Chow AY, Yu W, Chair S, et al. A feasibility study to investigate the acceptability and potential effectiveness of a telecare service for older people with chronic obstructive pulmonary disease. Int J Med Inform 2012 Oct;81(10):674-682. [doi: 10.1016/j.ijmedinf.2012.06.003] [Medline: 22789911]
75. Burton C, Pinnock H, McKinstry B. Changes in telemonitored physiological variables and symptoms prior to exacerbations of chronic obstructive pulmonary disease. J Telemed Telecare 2015 Jan;21(1):29-36. [doi: 10.1177/1357633X14562733] [Medline: 25475218]
76. Watz H, Tetzlaff K, Magnussen H, Mueller A, Rodriguez-Roisin R, Wouters EFM, et al. Spirometric changes during exacerbations of COPD: a post hoc analysis of the WISDOM trial. Respir Res 2018 Dec 13;19(1):251 [FREE Full text] [doi: 10.1186/s12931-018-0944-3] [Medline: 30545350]
77. Chawla H, Bulathsinghala C, Tejada JP, Wakefield D, ZuWallack R. Physical activity as a predictor of thirty-day hospital readmission after a discharge for a clinical exacerbation of chronic obstructive pulmonary disease. Ann Am Thorac Soc 2014 Oct;11(8):1203-1209. [doi: 10.1513/AnnalsATS.201405-198OC] [Medline: 25167366]
78. Jones PW, Harding G, Berry P, Wiklund I, Chen W, Kline Leidy N. Development and first validation of the COPD Assessment Test. Eur Respir J 2009 Sep;34(3):648-654 [FREE Full text] [doi: 10.1183/09031936.00102509] [Medline: 19720809]

XSL•FO
RenderX

79. Min S, Lee Y, Park Y, Yum H, Park M, Yoo J. COPD Assessment Test (CAT) has a positive relationship with exacerbations in COPD patients. Eur Respiratory Soc 2019;54(suppl 63):PA2706. [doi: 10.1183/13993003.congress-2019.PA2706]
80. Rassouli F, Baty F, Stolz D, Albrich WC, Tamm M, Widmer S, et al. Longitudinal change of COPD assessment test (CAT) in a telehealthcare cohort is associated with exacerbation risk. Int J Chron Obstruct Pulmon Dis 2017;12:3103-3109 [FREE Full text] [doi: 10.2147/COPD.S141646] [Medline: 29123387]
81. Lacoma A, Prat C, Andreo F, Domínguez J. Biomarkers in the management of COPD. Eur Respir Rev 2009 Jun;18(112):96-104 [FREE Full text] [doi: 10.1183/09059180.00000609] [Medline: 20956129]
82. Hurst JR, Donaldson GC, Perera WR, Wilkinson TMA, Bilello JA, Hagan GW, et al. Use of plasma biomarkers at exacerbation of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2006 Oct 15;174(8):867-874. [doi: 10.1164/rccm.200604-506OC] [Medline: 16799074]
83. Butler CC, Gillespie D, White P, Bates J, Lowe R, Thomas-Jones E, et al. C-reactive protein testing to guide antibiotic prescribing for COPD exacerbations. N Engl J Med 2019 Jul 11;381(2):111-120. [doi: 10.1056/NEJMoa1803185] [Medline: 31291514]
84. Koutsokera A, Kostikas K, Nicod LP, Fitting J. Pulmonary biomarkers in COPD exacerbations: a systematic review. Respir Res 2013 Oct 21;14(1):111 [FREE Full text] [doi: 10.1186/1465-9921-14-111] [Medline: 24143945]
85. Lawal O, Ahmed WM, Nijsen TME, Goodacre R, Fowler SJ. Exhaled breath analysis: a review of 'breath-taking' methods for off-line analysis. Metabolomics 2017;13(10):110 [FREE Full text] [doi: 10.1007/s11306-017-1241-8] [Medline: 28867989]
86. Kamal F, Kumar S, Edwards MR, Veselkov K, Belluomo I, Kebadze T, et al. Virus-induced volatile organic compounds are detectable in exhaled breath during pulmonary infection. Am J Respir Crit Care Med 2021 Nov 01;204(9):1075-1085 [FREE Full text] [doi: 10.1164/rccm.202103-0660OC] [Medline: 34319857]

Abbreviations

AECOPD: acute exacerbations of chronic obstructive pulmonary disease
AUROC: area under the receiver operating characteristic curve
CASP: Critical Appraisal Skills Programme
CAT: chronic obstructive pulmonary disease assessment test
COPD: chronic obstructive pulmonary disease
CRP: C-reactive protein
ED: emergency department
FEV1: forced expiratory volume in 1 second
HR: heart rate
HRQoL: health-related quality of life
PA: physical activity
PR: pulse rate
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRO: patient-reported outcomes
RCT: randomized controlled trial
RNN: recurrent neural network
RPM: remote patient monitoring
RR: respiratory rate
SpO_2: oxygen saturation
WISDOM: Withdrawal of Inhaled Steroids during Optimized Bronchodilator Management

Edited by T de Azevedo Cardoso; submitted 24.08.23; peer-reviewed by M Crooks, J Kvedar, J Edwards; comments to author 01.12.23; revised version received 29.02.24; accepted 09.07.24; published 09.09.24.

Please cite as:
Glyde HMG, Morgan C, Wilkinson TMA, Nabney IT, Dodd JW
Remote Patient Monitoring and Machine Learning in Acute Exacerbations of Chronic Obstructive Pulmonary Disease: Dual Systematic Literature Review and Narrative Synthesis
J Med Internet Res 2024;26:e52143
URL: https://www.jmir.org/2024/1/e52143
doi:10.2196/52143
PMID:

XSL•FO
RenderX

©Henry Mark Granger Glyde, Caitlin Morgan, Tom M A Wilkinson, Ian T Nabney, James W Dodd. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Detection of Sleep Apnea Using Wearable AI: Systematic Review and Meta-Analysis

Alaa Abd-alrazaq[1], PhD; Hania Aslam[1], MSc; Rawan AlSaad[1], PhD; Mohammed Alsahli[2], PhD; Arfan Ahmed[1], PhD; Rafat Damseh[3], PhD; Sarah Aziz[1], MSc; Javaid Sheikh[1], MD

[1]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Qatar Foundation, Doha, Qatar

[2]Health Informatics Department, College of Health Science, Riyadh, Saudi Electronic university, Riyadh, Saudi Arabia

[3]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

Corresponding Author:
Alaa Abd-alrazaq, PhD
AI Center for Precision Health
Weill Cornell Medicine-Qatar
Qatar Foundation
A31 Luqta street
Education City
Doha
Qatar
Phone: 974 55787845654
Email: aaa4027@qatar-med.cornell.edu

Abstract

Background: Early detection of sleep apnea, the health condition where airflow either ceases or decreases episodically during sleep, is crucial to initiate timely interventions and avoid complications. Wearable artificial intelligence (AI), the integration of AI algorithms into wearable devices to collect and analyze data to offer various functionalities and insights, can efficiently detect sleep apnea due to its convenience, accessibility, affordability, objectivity, and real-time monitoring capabilities, thereby addressing the limitations of traditional approaches such as polysomnography.

Objective: The objective of this systematic review was to examine the effectiveness of wearable AI in detecting sleep apnea, its type, and its severity.

Methods: Our search was conducted in 6 electronic databases. This review included English research articles evaluating wearable AI's performance in identifying sleep apnea, distinguishing its type, and gauging its severity. Two researchers independently conducted study selection, extracted data, and assessed the risk of bias using an adapted Quality Assessment of Studies of Diagnostic Accuracy-Revised tool. We used both narrative and statistical techniques for evidence synthesis.

Results: Among 615 studies, 38 (6.2%) met the eligibility criteria for this review. The pooled mean accuracy, sensitivity, and specificity of wearable AI in detecting apnea events in respiration (apnea and nonapnea events) were 0.893, 0.793, and 0.947, respectively. The pooled mean accuracy of wearable AI in differentiating types of apnea events in respiration (normal, obstructive sleep apnea, central sleep apnea, mixed apnea, and hypopnea) was 0.815. The pooled mean accuracy, sensitivity, and specificity of wearable AI in detecting sleep apnea were 0.869, 0.938, and 0.752, respectively. The pooled mean accuracy of wearable AI in identifying the severity level of sleep apnea (normal, mild, moderate, and severe) and estimating the severity score (Apnea-Hypopnea Index) was 0.651 and 0.877, respectively. Subgroup analyses found different moderators of wearable AI performance for different outcomes, such as the type of algorithm, type of data, type of sleep apnea, and placement of wearable devices.

Conclusions: Wearable AI shows potential in identifying and classifying sleep apnea, but its current performance is suboptimal for routine clinical use. We recommend concurrent use with traditional assessments until improved evidence supports its reliability. Certified commercial wearables are needed for effectively detecting sleep apnea, predicting its occurrence, and delivering proactive interventions. Researchers should conduct further studies on detecting central sleep apnea, prioritize deep learning algorithms, incorporate self-reported and nonwearable data, evaluate performance across different device placements, and provide detailed findings for effective meta-analyses.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58187) doi:10.2196/58187

KEYWORDS

sleep apnea; hypopnea; artificial intelligence; wearable devices; machine learning; systematic review; mobile phone

Introduction

Background

Sleep apnea refers to a health condition where airflow either ceases or decreases episodically during sleep [1]. According to the American Academy of Sleep Medicine, sleep apnea is categorized as a sleep disorder wherein an individual experiences challenges pertaining to breathing when they are asleep [2]. Primarily, there are 3 kinds of sleep apnea. First, obstructive sleep apnea (OSA) is the consequence of issues with the operation of the upper respiratory tract and is considered a chronic breathing disorder associated with sleep [3]. By contrast, a condition where signals required to regulate breathing muscles are not generated or transmitted is referred to as central sleep apnea (CSA). Complex or mixed sleep apnea is a condition that involves a combination of both OSA and CSA [4]. It often begins as OSA and evolves into CSA [4].

According to global estimates, approximately 936 million adults aged between 30 and 69 years experience OSA [5]. A systematic review showed that the global prevalence of OSA is between 9% and 38% [6]. In the United States alone, the number of people struggling with sleep apnea may exceed 30 million, as per the American Medical Association [7]. Furthermore, studies showed that >80% of sleep apnea cases remain undiagnosed [7-10]. If not diagnosed and treated, sleep apnea may result in severe health issues, such as mood disorders [11-13], cardiovascular diseases [14,15], cognitive deterioration [16,17], increased risk of road accidents [18,19], and all-cause mortality [20,21]. Therefore, the timely detection of sleep apnea for prompt initiation of treatment is imperative.

Conventionally, polysomnography is a comprehensive diagnostic test used in the field of sleep medicine to evaluate and monitor various physiological parameters during sleep to help diagnose sleep disorders, such as sleep apnea [22]. Despite being considered the gold standard for diagnosing sleep apnea, it does have some disadvantages and limitations: (1) it is relatively expensive; (2) access to sleep laboratories may be limited, particularly in certain geographic areas; (3) it can be inconvenient for patients, as they must spend a full night in a sleep laboratory with numerous sensors and electrodes attached to their body; (4) the physiological parameters recorded using polysomnography may not fully reflect the individual's typical sleep behavior due to a first-night effect in a sleep laboratory, where sleep patterns are different from those at home due to the novelty of the environment; and (5) it is a subjective process, as analyzing polysomnography data depends on sleep clinicians' experience [22-24]. Hence, there is a dire need to develop and integrate automated technologies and tools that are more efficient and capable of addressing the challenges posed by the current system of diagnosing sleep apnea.

One of the promising solutions that have been used to address the limitations of polysomnography is wearable artificial intelligence (AI), which refers to the integration of AI algorithms and techniques into wearable devices (eg, smartwatches, fitness trackers, and smart glasses) to collect and analyze data (eg, heart rate [HR], respiration rate, and oxygen saturation) to offer various functionalities and insights. Sleep apnea can be efficiently detected with wearable AI due to its convenience, accessibility, affordability, objectivity, and real-time monitoring capabilities. Various types of wearable devices can be used for gathering biomarkers associated with sleep apnea: on-body devices (worn directly on the body or skin), near-body devices (worn close to the body but not touching the skin), in-body devices (implanted within the body), and electronic textiles (clothes with built-in technology). Wearable AI can be used for (1) detecting apnea events in respiration, (2) identifying the type of apnea events in respiration (hypopnea, OSA, CSA, and mixed), (3) detecting patients with sleep apnea, and (4) estimating the severity of sleep apnea.

Research Problem and Aim

In the last decade, numerous investigations have been carried out to evaluate the effectiveness of wearable AI in detecting sleep apnea. Consolidating the results of these studies can contribute to forming more conclusive judgments regarding the effectiveness of wearable AI in detecting sleep apnea. Previous literature reviews attempted to summarize the evidence, but they were constrained by the following limitations. First, most previous reviews were literature reviews rather than systematic reviews [22-28]. Second, many reviews concentrated solely on OSA rather than considering all types of sleep apnea [22,23,25-29]. Third, some reviews focused on a specific type of data, such as HR variability [2,25] and electrocardiography [1,2,25], for sleep apnea detection. Fourth, main databases, such as Embase [1,2,22-29], ACM [1,2,22-29], IEEE [22-25,27-29], and Scopus [1,2,22-25,28], were not incorporated in the searches of previous reviews. Fifth, all prior reviews focused on the performance of various sensors rather than specifically addressing wearable devices [1,2,22-29]. Sixth, one of the reviews focused on non-AI tools for detecting sleep apnea [29]. Seventh, the risk of bias was not taken into account in most of the reviews [1,2,22-28]. Finally, none of these reviews used statistical techniques (eg, meta-analysis) to aggregate findings from previous studies [1,2,22-29]. Hence, this review aimed to bridge all these identified gaps with a focus on examining the performance of wearable AI when it comes to both the detection and prediction of sleep apnea, thereby making it the first of its kind in this field.

Methods

Overview

This review was undertaken and reported in line with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy) guidelines [30]. Multimedia Appendix 1 provides

XSL•FO
RenderX

this review's PRISMA-DTA checklist. Its protocol has been registered with the PROSPERO (CRD42023495554).

Search Strategy

On December 7, 2023, a comprehensive search was performed across the following electronic repositories: MEDLINE (via Ovid), Embase (via Ovid), ACM Digital Library, Scopus, IEEE Xplore, and Google Scholar. MEDLINE and Embase were chosen due to their reputation as authoritative sources for biomedical and health sciences literature. ACM Digital Library and IEEE Xplore were selected for their status as leading repositories for publications in computing, information technology, electrical engineering, and electronics. Scopus was included because of its comprehensive coverage of scientific literature across multiple disciplines, including health sciences, engineering, computer science, and social sciences. Google Scholar was incorporated, as it indexes scholarly literature from diverse sources and serves as a valuable supplementary tool for identifying relevant studies and gray literature. We set an autoalert to run the search query biweekly for 3 months, concluding on March 6, 2024. Because Google Scholar returned a massive number of results, this review assessed only the first 100 results (equivalent to 10 pages). To identify additional relevant studies, we examined the references cited in the studies already included (backward reference list checking) and studies that had cited the included studies (forward reference list checking).

Relevant literature reviews were assessed, and 2 experts holding doctoral degrees in digital health and health informatics were consulted to compile and collate search terms [31]. The final search query combined three categories of search terms: (1) terms related to AI (eg, "artificial intelligence," "machine learning," and "deep learning"), (2) terms associated with wearable devices (eg, "wearable," "smartwatch," and "smartband"), and (3) terms linked to sleep apnea (eg, "sleep apnea" and "sleep aponea"). The Boolean operators "OR" and "AND" were used to combine terms within the same category and across different categories, respectively. The specific search query used for searching each database is detailed in Multimedia Appendix 2 for reference.

Study Eligibility Criteria

This review included studies that used AI algorithms to detect sleep apnea or predict its occurrence by leveraging data derived from wearable devices. The research articles deemed suitable for inclusion in this review were those that concentrated on individuals diagnosed with or suspected of having any type of sleep apnea. No limitations were imposed based on age, gender, or ethnicity. In addition, for inclusion in this review, studies were required to evaluate the performance of AI algorithms in detecting or predicting apnea events in respiration, identifying types of apnea events in respiration, detecting patients with sleep apnea, or estimating the severity of sleep apnea. The studies had to provide the confusion matrix or performance metrics (eg, accuracy, sensitivity, and specificity). Studies using AI solely for detecting sleep quality, sleep stages, or other sleep disorders or forecasting the outcomes of sleep apnea interventions were excluded. This review included studies that gathered data using, at a minimum, on-body wearable devices. Conversely, research papers exclusively relying on the following devices for data collection were not considered: nonwearable devices; handheld devices (eg, mobile phones); near-body wearable devices; in-body wearable devices; wearable devices physically connected to nonwearable devices; and wearable devices necessitating expert oversight, such as those demanding precise electrode placement. This review included only peer-reviewed journal articles, conference papers, and dissertations, without restrictions on study setting, study design, reference standard (ie, ground truth), year of publication, or country of study. However, papers not published in English were excluded from consideration. The decision to exclude studies not written in English was based on practical considerations related to resource constraints and the accessibility of non-English literature. While including studies in languages other than English may enhance the comprehensiveness of the review, it can also pose challenges in terms of language translation, interpretation, and the synthesis of findings. Furthermore, English is widely recognized as the dominant language of scholarly communication in many scientific disciplines, including health care and biomedical research. We have transparently acknowledged its implications for the review's scope and findings in the *Limitations* section. We also excluded studies that fell into the categories of editorials, reviews, protocols, posters, conference abstracts, and research highlights. The decision to exclude these publication types was primarily guided by the need to maintain the focus and rigor of our review process. While editorials, reviews, and research highlights provide valuable insights into and perspectives on a topic, they typically do not present original research findings or empirical data that meet the objectives of our study. Similarly, protocols, posters, and conference abstracts often offer preliminary or incomplete results that may not undergo peer review or provide sufficient detail for a comprehensive analysis. This helps maintain the quality and reliability of the evidence synthesized in our review while minimizing the risk of bias introduced by including non–peer-reviewed or preliminary findings.

Study Selection

The study selection process comprised 3 key steps. Initially, the EndNote (version X9; Clarivate) software was used to eliminate any duplicate papers from the initial pool. Subsequently, 2 reviewers assessed the titles and abstracts of the remaining studies, separately deciding on their inclusion. Finally, the reviewers independently scrutinized the full texts of the remaining articles. Any discrepancies were deliberated upon and resolved through discussion. The level of agreement between the reviewers was substantial, indicated by a κ score of 0.92 for the evaluation of titles and abstracts and 0.95 for the examination of full texts.

Data Extraction

Initially, 5 studies were used to develop and test the data extraction form shown in Multimedia Appendix 3. Independently, 2 reviewers used Excel (Microsoft Corp) to extract metadata from the studies, participants' characteristics, wearable devices' specifications, and AI algorithms' features. In addition to the previously mentioned extracted data, we

XSL•FO
RenderX

collected the highest performance score for each metric, algorithm, and measured outcome. When studies provided raw data or confusion matrices, we calculated all possible performance metrics (eg, accuracy, specificity, and sensitivity). In case of the unavailability of such data, we attempted to obtain them by reaching out to the studies' first and corresponding authors. Any discrepancies between the 2 reviewers were addressed through discussion between them.

Risk of Bias and Applicability Appraisal

To evaluate the quality of the studies included in our review, we adapted the Quality Assessment of Studies of Diagnostic Accuracy-Revised (QUADAS-2) [32] tool to align with our review's specific objectives. This adaptation involved substituting some of the original criteria, which were not applicable to our context, with more relevant criteria from the Prediction Model Risk of Bias Assessment Tool [33]. We modified the QUADAS-2 tool to encompass 4 main domains tailored to our review: "participants," "index test" (focused on AI algorithms), "reference standard" (representing the ground truth), and "analysis." For each domain, we developed 4 targeted questions aligned with our review's objectives. In addition, our evaluation assessed the practical applicability of the results derived from the first 3 domains. To optimize our adapted tool, we initially tested it on 5 studies for fine-tuning purposes. The included studies were independently evaluated by 2 reviewers using the modified QUADAS-2 tool (Multimedia Appendix 4). Any differences in their assessments were discussed and resolved through consensus.

Data Synthesis

We used both narrative and statistical techniques to synthesize the data extracted from the included studies. In our narrative synthesis, we used textual descriptions and tabulated summaries to elucidate the characteristics of the included studies, encompassing study metadata, wearable devices, and AI techniques. As for the statistical approach, a meta-analysis was carried out when at least 2 different studies presented enough data to perform the analysis. We conducted conventional meta-analyses for results associated with the following outcomes, given that they were extracted from different unique studies (ie, independent effect sizes): identification of types of apnea events in respiration, detection of patients with sleep apnea, and estimation of the severity of sleep apnea. Specifically, DerSimonian-Laird random-effects models [34] using the Freeman-Tukey double arcsine transformation [35,36] were performed to pool the extracted results. This method considers variations arising from sampling and accounts for heterogeneity in estimates. The analysis was carried out using the *meta* toolkit within R (version 4.2.2; The R Foundation) [37].

We also performed multilevel meta-analyses for results related to the detection of apnea events in respiration, as certain results originated from the same study (ie, dependent effect sizes) [34,38]. Multilevel meta-analyses were used to address this dependency in effect sizes, thereby minimizing the risk of type I errors. These analyses were carried out using the *metafor* toolkit within R (version 4.2.2) [35].

When applicable, subgroup meta-analyses were conducted to explore how different factors might influence the effectiveness of wearable AI [34,38]. These factors included AI algorithms, the type of algorithm (ie, machine learning [ML] vs deep learning), the number of participants, the type of sleep apnea, the status of the wearable device (ie, commercial vs noncommercial), the placement of the wearable device, data set size, data type, ground truth, and validation method. We considered differences in results between subgroups to be statistically significant if the statistical probability (P value) was <.05.

To assess how consistent the studies were in their findings (heterogeneity), we used 2 statistical tests. The first test is the Cochrane Q statistic, which indicates whether the observed differences in results could be due to chance alone. A P value <.05 indicates significant heterogeneity, meaning the results varied more than expected by chance. The second test is the I^2 statistic, which quantifies the proportion of observed variability due to real differences between studies rather than differences by chance [35,39]. Heterogeneity was considered insignificant when I^2 ranged from 0% to 40%, moderate when I^2 fell within the 30% to 60% range, substantial when I^2 ranged from 50% to 90%, or considerable when I^2 extended from 75% to 100%.

Results

Search Results

As depicted in Figure 1, a total of 615 citations were retrieved when the above-identified databases were searched. Of the retrieved citations, 161 (26.2%) duplicates were removed using EndNote X9, leaving 454 (73.8%) studies. Further, 362 (79.7%) studies were removed after screening the titles and abstracts of these 454 studies. After retrieving and reading the full text of all the remaining 92 (20.3%) studies, it was determined that 57 (62%) of these studies were ineligible for inclusion. The main reasons for exclusion were that they did not use wearable devices (23/92, 25%), did not use AI algorithms (11/92, 12%), did not focus on sleep apnea (6/92, 7%), were irrelevant publication types (16/92, 17%), or were not written in English (1/92, 1%). We identified 3 additional studies relevant to this review through backward reference list checking. In total, 38 studies were included in this review [40-77], and 27 (71%) of them were eligible for meta-analyses [40,41,45-49,52-55,57,58,61-64,66,68,69,71-77].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process. AI: artificial intelligence.

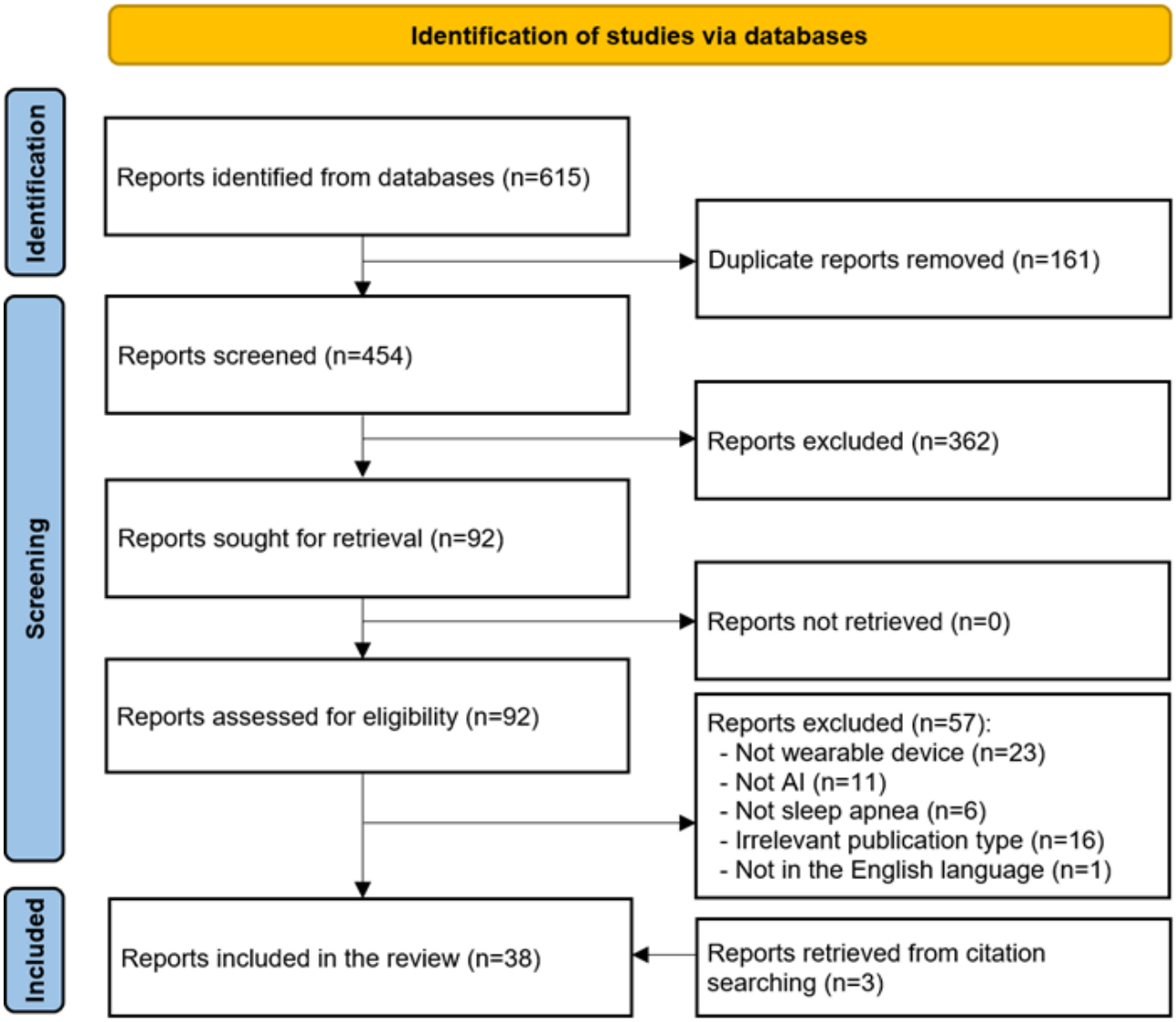

Characteristics of the Included Studies

As displayed in Table 1, the number of studies has varied over the years, with the highest number reached in 2020 (11/38, 29%). While the included studies were conducted in 16 different countries, the studies were predominantly from the United States (9/38, 24%). Most of the studies were journal articles (29/38, 76%), but conference papers also made a substantial contribution (9/38, 24%). The average number of participants across studies was 155.9 (SD 374.9). The number of participants ranged from 4 to 2252. The mean age of participants was identified in 25 (66%) of the 38 included studies and ranged from 25.6 to 61.1 years, with an average of 47.3 (SD 9.3) years. Across 25 studies reporting the proportion of female participants, female participants constituted an average of 37.4% of the total participants, ranging from 12% to 65%. A total of 20 studies reported the BMI, which ranged from 22.1 to 38.7 kg/m^2, with an average of 28.6 (SD 3.81) kg/m^2. About two-thirds (25/38, 66%) of studies did not focus on a specific type of sleep apnea. The characteristics of each included study are listed in Multimedia Appendix 5.

XSL•FO
RenderX

Table 1. Characteristics of the included studies (N=38).

Features	Studies	References
Year of publication, n (%)		
2023	8 (21)	[54,55,60,64,67,71,75,76]
2022	9 (24)	[40,42,46,52,61,63,68,73,77]
2021	4 (11)	[43,53,70,72]
2020	11 (29)	[41,45,47,49,51,57-59,65,66,74]
2019	1 (3)	[48]
2018	3 (8)	[44,50,69]
2014	1 (3)	[62]
2013	1 (3)	[56]
Country of publication, n (%)		
United States	9 (24)	[43,44,46,47,55-57,64,72]
China	8 (21)	[42,63,67,68,70,71,75,76]
South Korea	4 (11)	[45,51,73,77]
Canada	2 (5)	[48,49]
Italy	2 (5)	[40,60]
Norway	2 (5)	[53,54]
Taiwan	2 (5)	[41,69]
Others (<2)	9 (24)	[50,52,58,59,61,62,65,66,74]
Publication type, n (%)		
Conference paper	9 (24)	[43-45,48,50,51,59,62,76]
Journal article	29 (76)	[40-42,46,47,49,52-58,60,61,63-75,77]
Number of participants, mean (SD; range)	155.9 (374.9; 4-2252)	[40-77]
Age (y)		
Value, mean (SD; range)	47.3 (9.3; 25.6-61.1)	[40,41,43,45-49,53,54,57-62,64,66,67,69,71-73,75,77]
Not reported, n (%)	13 (34)	[42,44,50-52,55,56,63,65,68,70,74,76]
Female participants (%)		
Value, mean (SD; range)	37.4 (14.76; 12-65)	[40,41,43,46-49,52-54,57-62,64,66-69,71-73,75]
Not reported, n (%)	13 (34)	[42,44,45,50,51,55,56,63,65,70,74,76,77]
BMI (kg/m^2)		
Value, mean (SD; range)	28.6 (3.813; 22.1-38.7)	[40,41,45,47-49,53,54,57,58,61,62,64,66,68,69,71-73,77]
Not reported, n (%)	18 (47)	[42-44,46,50-52,55,56,59,60,63,65,67,70,74-76]
Sleep apnea type, n (%)		
All	25 (66)	[41,43,45,46,48,49,51-55,57,59-63,66,67,69,70,73-75,77]
Obstructive sleep apnea	12 (32)	[40,42,44,47,56,58,64,65,68,71,72,76]
Central sleep apnea	1 (3)	[50]

Features of Wearable Devices

Commercial wearable devices constituted the majority of wearable devices in the included studies (24/38, 63%; Table 2). The most mentioned wearable device in the included studies was the Belun Ring (3/38, 8%). Wearable devices are placed on various body parts, with the chest (16/38, 42%), wrist (11/38, 29%), and abdomen (9/38, 24%) being the most common locations. Wearable devices were worn for 1 full night (6-8 hours) in 29 studies (76%). The features of wearable devices in each included study are shown in Multimedia Appendix 6.

XSL•FO
RenderX

Table 2. Features of wearable devices (N=38).

Features	Studies, n (%)	References
Status of the wearable device		
Commercial	24 (63)	[40,42-47,51,53,54,57,58,60-66,70,72,73,76,77]
Noncommercial	14 (37)	[41,48-50,52,55,56,59,67-69,71,74,75]
Name of the wearable device		
Belun Ring	3 (8)	[47,64,72]
Patch	2 (5)	[48,49]
T-REX TR100A	2 (5)	[73,77]
Others	17 (45)	[43-46,51,53,54,57,60-62,64-66,71,74,76]
Not reported	14 (37)	[41,42,50,52,55,56,58,59,63,67-70,75]
Placement of the wearable device		
Chest	16 (42)	[41,46,50,52-54,56,57,59-62,65,66,68,69]
Wrist	11 (29)	[40,42-45,51,54,58,63,70,76]
Abdomen	9 (24)	[41,53,54,57,65,69,73,74,77]
Finger	6 (16)	[47,53,64,67,69,72]
Neck	2 (5)	[48,49]
Nose	2 (5)	[53,75]
Face	1 (3)	[55]
Duration of wearing the wearable device		
1 full night	29 (76)	[41-43,45-49,53-56,58,60-69,71-73,75-77]
<1 full night	3 (8)	[50,52,59]
>1 full night	2 (5)	[44,51]
Not reported	3 (8)	[57,70,74]

Features of AI

Classification was the dominant problem-solving approach used in the included studies (38/38, 100%; Table 3). Various AI algorithms were used in the included studies, with convolutional neural networks (CNNs) being the most common (14/38, 37%). Among the 38 included studies, most studies (n=37, 97%) used AI to detect the current sleep apnea, whereas 3 (8%) studies used wearable AI to predict sleep apnea before its occurrence. The mean data set size reported in 28 (74%) studies was 60,554 (SD 133,059), with the range spanning from 12 to 561,480. Most studies (36/38, 95%) used closed-source data, while only 2 (5%) of 38 studies used open-source data. Data were gathered through wearable devices in all studies (38/38, 100%), via self-reported questionnaires in 3 (8%) studies, and using nonwearable devices (eg, smartphones) in 2 (5%) studies. Respiration data (eg, respiratory rate and respiratory efforts; 25/38, 66%) and HR data (eg, HR, HR variability, and interbeat interval; 21/38, 55%) were the most frequently used data for developing the models in the included studies. The number of features reported in 21 (55%) of the 38 studies ranged from 3 to 212, with an average of 44.3 (62.5). Most studies used polysomnography as the ground truth assessment method (26/38, 68%), followed by the wearable device (8/38, 21%) and the context of the experiment (eg, performing different patterns of breathing; 4/38, 11%). In 28 studies that reported the assessor of the ground truth, sleep technicians were the most common assessors (23/38, 61%), followed by sleep physicians (8/38, 21%). American Academy of Sleep Medicine guidelines were followed in 84% (32/38) of studies to assess the ground truth. Train-test split was the most common approach used in the included studies to validate the performance of AI models (20/38, 53%), followed by k-fold cross-validation (17/38, 45%). The included studies used wearable AI to detect apnea events in respiration (24/38, 63%) and patients with sleep apnea (15/38, 40%) and to identify the severity of sleep apnea (21/38, 55%) and types of apnea events in respiration (8/38, 21%). The features of AI in each included study are described in Multimedia Appendix 7.

XSL•FO
RenderX

Table 3. Features of artificial intelligence (N=38).

Features	Studies	References
Problem-solving approaches, n (%)		
Classification	38 (100)	[40-77]
Regression	15 (40)	[45-49,54,56,58,62,64,70,71,76]
AI[a] algorithms, n (%)		
Convolutional neural network	14 (37)	[48,49,53-55,57-60,63,64,68,73,75,77]
Random forest	10 (26)	[40,42,43,46,52-54,67,70,73]
Long short-term memory	9 (24)	[41,44,45,48,49,52-54,66]
Support vector machines	8 (21)	[43,52,53,56,62,67,69,73]
K-nearest neighbors	7 (18)	[42,51,52,54,61,67,70]
Artificial neural network	5 (13)	[47,51,65,72,74]
Multilayer perceptron	5 (13)	[40,50,53,54,73]
Naive Bayes	5 (13)	[41,42,51,52,70]
Decision trees	4 (11)	[42,43,52,70]
AdaBoost	3 (8)	[41,52,61]
Others (<3)	5 (13)	[52,67,71,73,76]
Aim of AI algorithms, n (%)		
Detection	37 (97)	[40-55,57-76]
Prediction	3 (8)	[44,56,65]
Data set size		
Value, mean (SD; range)	60,554 (133,059; 12-561,480)	[40,42-44,46-49,52-55,58-66,68,72-76]
Not reported, n (%)	10 (26)	[41,45,50,51,56,57,67,69-71]
Data sources, n (%)		
Closed source	36 (95)	[40-43,45-64,66-77]
Open source	2 (5)	[44,65]
Data types, n (%)		
Wearable device data	38 (100)	[40-77]
Self-reported data	3 (8)	[44,69,76]
Nonwearable device data	2 (5)	[69,76]
Data input to AI algorithms, n (%)		
Respiration data	25 (66)	[41,43,45,46,48-50,52-54,56-62,65,66,69,73-77]
Heart rate	21 (55)	[40,42,45,47,50-52,56,58,62-64,66-68,70-73,76,77]
Body movement	14 (37)	[40,44,45,47,51,52,58,60,62,64,66,71,72,76]
Oxygen saturation	13 (34)	[41,46,47,53,54,56,60,64,67,69,71,72,76]
Acoustic data	3 (8)	[56,60,76]
Others (<3)	10.5 (4)	[44,55,58,76]
Number of features		
Value, mean (SD; range)	44.33 (62.5; 3-212)	[40-43,45-50,52,56-58,61,64,65,69,70,72,73]
Not reported, n (%)	17 (45)	[44,51,53-55,59,60,62,63,66-68,71,73-77]
Ground truth assessment methods, n (%)		
Polysomnography	26 (68)	[41-43,45-49,55,58,61-73,75-77]
Wearable device	8 (21)	[40,44,51,53,54,56,60,74]
Context	4 (11)	[50,52,57,59]

XSL•FO
RenderX

Features	Studies	References
Guidelines for ground truth assessment, n (%)		
American Academy of Sleep Medicine guidelines	32 (84)	[40-49,51,53-55,58,61-77]
Not reported	6 (16)	[50,52,56,57,59,60]
Assessors of ground truth, n (%)		
Sleep technician	23 (61)	[40-42,44-47,53-56,58,60-62,64-66,69,72,73,76,77]
Sleep physician	8 (21)	[42,57,63,64,68,71,72]
Not reported	10 (26)	[48-52,59,68,70,74,75]
Validation methods, n (%)		
Train-test split	20 (53)	[41,45,47,51,52,55,57-60,63-65,68,71-75,77]
K-fold cross-validation	17 (45)	[42-44,46,48-51,53,54,56,61,63,66,68,70,76]
Leave-one-out cross-validation	5 (13)	[40,54,60,62,69]
Not reported	1 (3)	[67]
Measured outcomes, n (%)		
Apnea events in respiration	24 (63)	[41,43-46,50-59,61,63,66-68,73-75,77]
Sleep apnea severity	21 (55)	[40-43,45-49,53,58,62,64,65,69-73,76,77]
Patients with sleep apnea	15 (40)	[40-44,46-48,53,58,63,64,69,71,72,76]
Type of apnea events	8 (21)	[41,46,53,57,58,60,66,75]

[a]AI: artificial intelligence.

Results of Risk-of-Bias Appraisal

Nearly half of the included studies (17/38, 45%) reported comprehensive details to determine whether an appropriate consecutive or random sample of eligible participants was used. Over half of the studies (22/38, 58%) avoided inappropriate exclusions. A substantial majority, 30 (79%) out of 38 studies, ensured a balanced number of patients across subgroups. In addition, around two-thirds (25/38, 66%) of the studies reported a sufficient sample size. Consequently, a little less than half of the studies (16/38, 42%) were assessed as having a low risk of bias in the "selection of participants" domain, as shown in Figure 2. In terms of matching participants to the predefined requirements in the review question, a low level of concern was identified in nearly 40% (15/38, 40%) of the included studies, as shown in Figure 3.

Figure 2. Results of the assessment of risk of bias in the included studies.

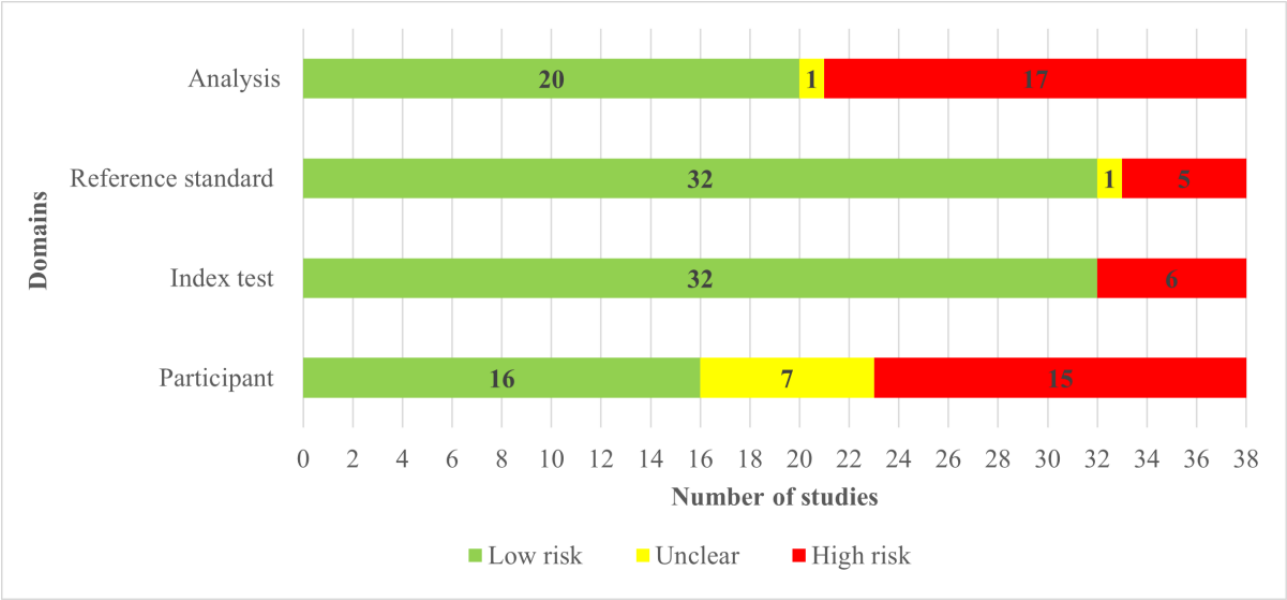

A substantial majority of the included studies comprehensively detailed their AI models, with 34 (89%) out of 38 studies providing thorough descriptions. Almost all, 35 (92%) out of 38 studies, clearly reported the features (predictors) used. Moreover, an overwhelming majority, 36 (95%) out of 38 studies, ensured that these features were sourced without prior

XSL•FO
RenderX

knowledge of the outcome data. Consistency in feature assessment across participants was observed in 35 (92%) out of 38 studies. Consequently, the potential for bias in the "index test" domain was assessed as low in the vast majority of the studies (32/38, 84%), as shown in Figure 2. In addition, 32 (84%) out of 38 studies were found to have minimal concerns regarding the alignment between the model's predictors and the review question's criteria, as illustrated in Figure 3.

Figure 3. Results of the assessment of applicability concerns in the included studies.

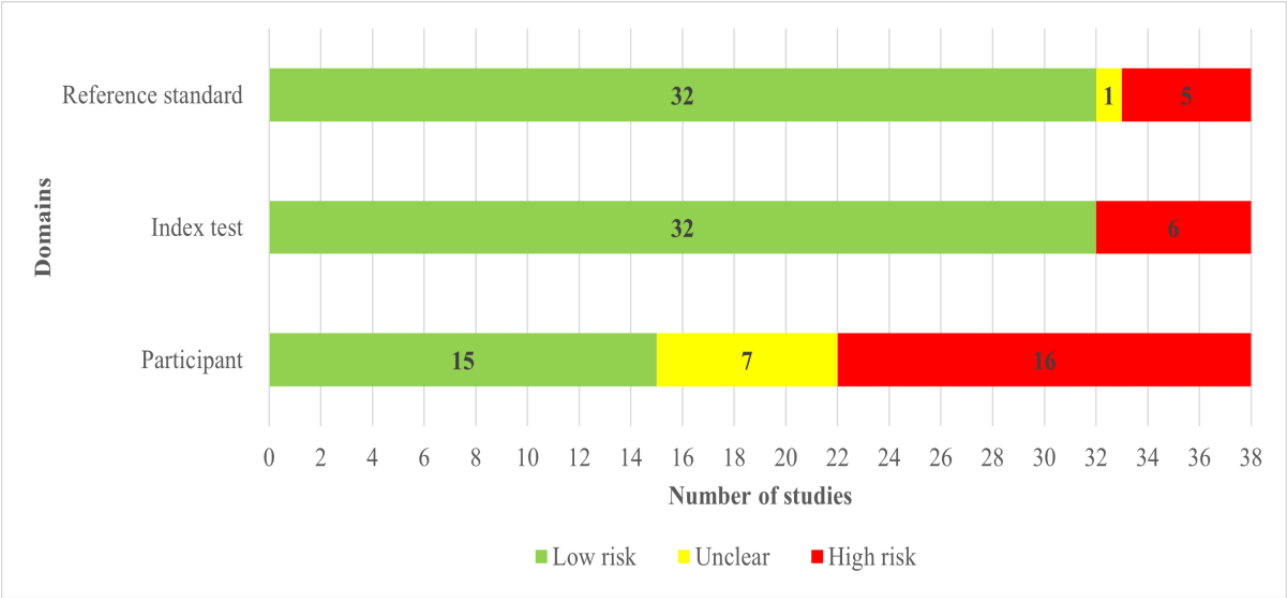

In most of the included studies (32/38, 84%), the outcome of interest, specifically sleep apnea, was consistently assessed using appropriate methodologies. Nearly all studies (37/38, 97%) defined and determined the outcome in a uniform manner for all participants. An overwhelming majority of the studies (36/38, 95%) determined the outcome without prior knowledge of the predictor information. In a substantial portion of the studies (33/38, 87%), the diagnostic test was conducted for an appropriate duration to ensure accurate results. As a result, the potential for bias in the "reference standard" domain was deemed low in the vast majority of the studies (32/38, 84%), as shown in Figure 2. In addition, the same number of studies (32/38, 84%) showed minimal concerns regarding any discrepancies between the outcome's definition, timing, or determination and the review question's criteria, as indicated in Figure 3.

Finally, a significant majority of the studies (34/38, 89%) ensured the inclusion of all enrolled participants in the data analysis. A substantial number of these studies (32/38, 84%) executed proper data preprocessing. Similarly, a high proportion (34/38, 89%) adopted suitable measures to evaluate the performance of their models. Nearly half of the studies (17/38, 45%) demonstrated an appropriate split among training, validation, and test sets. However, the risk of bias in the validation methods used by the remaining studies remained unclear due to insufficient information being provided. Consequently, slightly more than half of the studies (20/38, 53%) were deemed to have a low risk of bias in the "analysis" domain, as indicated in Figure 2. A detailed breakdown of the "risk of bias" and "applicability concerns" for each domain in every study is available in Multimedia Appendix 8.

Results of the Studies

As mentioned earlier, meta-analyses were carried out to pool results related to 4 outcomes: detection of apnea events in respiration, identification of types of apnea events in respiration, detection of patients with sleep apnea, and estimation of the severity of sleep apnea. The following subsections present the results of the meta-analyses for each outcome.

Apnea Events in Respiration

Accuracy

We conducted meta-analyses of 36 estimates of accuracy derived from 2,702,305 respiratory events across 17 (45%) of the 38 studies (Table 4). The pooled mean accuracy of these estimates was 0.893 (95% CI 0.82-0.94). The meta-analyzed evidence exhibited considerable statistical heterogeneity (P<.001; I^2=100%). Further, Table 4 shows that there is a statistically significant difference in the pooled mean accuracy between subgroups in the "algorithms" group (P<.001) and "type of algorithms" group (P=.02), whereas no statistically significant difference (P>.05) was found in the pooled mean accuracy between subgroups in the remaining groups.

XSL•FO
RenderX

Table 4. Pooled mean estimates of accuracy in detecting respiratory events by several factors.

Groups	Studies (N=38), n[a] (%)	Sample size, N	Accuracy (%), range	Pooled mean accuracy (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Algorithm								*<.001* [b]
Convolutional neural network	9 (24)	437,593	0.76-0.97	0.884 (0.84-0.92)	0.0102	8523.6 (<.001)	100	
Recurrent neural networks	6 (16)	665,091	0.73-0.92	0.848 (0.79-0.90)	0.0086	18,227 (<.001)	100	
Random forest	4 (11)	807,225	0.81-0.96	0.867 (0.77-0.94)	0.0154	41,658 (<.001)	100	
Support vector machine	3 (8)	245,745	0.79-0.84	0.807 (0.77-0.84)	0.0014	221.4 (<.01)	99	
K-nearest neighbors	3 (8)	233,647	0.69-0.77	0.736 (0.69-0.78)	0.0019	78.1 (<.01)	97	
AdaBoost	2 (5)	5629	0.71-0.72	0.716 (0.71-0.73)	0.0000	0.99 (.32)	0	
Multilayer perceptron	2 (5)	242,505	0.80-0.81	0.804 (0.79-0.86)	0.0001	11.3 (<.01)	91	
Quadratic discriminant analysis	2 (5)	17,727	0.60-0.73	0.664 (0.53-0.79)	0.0103	218.8 (<.01)	100	
Type of algorithm								*.02*
Machine learning	18 (47)	1,334,180	0.60-0.97	0.831 (0.65-0.92)	0.1926	289,302.4 (<.001)	100	
Deep learning	18 (47)	1,368,125	0.73-1.00	0.899 (0.82-0.94)	0.2659	160,871.2 (<.001)	100	
Sample size, n								.93
<100	24 (63)	326,322	0.60-1.00	0.885 (0.74-0.95)	0.4783	111,302.0 (<.001)	100	
100-200	3 (8)	276,572	0.82-0.93	0.896 (0.82-0.95)	0.0086	2658.6 (<.001)	100	
<200	9 (24)	2,099,411	0.77-0.97	0.907 (0.75-0.97)	0.1238	314,844.0 (<.001)	100	
Type of sleep apnea								>.99
All	34 (89)	2,635,188	0.60-1.00	0.893 (0.80-0.94)	0.3616	457,420.0 (<.001)	96	
Obstructive sleep apnea	2 (5)	67,117	0.88-0.90	0.892 (0.87-0.91)	0.0007	25.4 (<.001)	100	
Status of the WD[c]								.05
Commercial	22 (58)	2,581,505	0.69-0.97	0.844 (0.78-0.89)	0.0705	370,236.6 (<.001)	100	
Noncommercial	14 (37)	120,800	0.60-1.00	0.947 (0.80-0.99)	0.6243	73,404.3 (<.001)	100	
Placement of the WD								.61
Chest	14 (37)	719,055	0.60-0.97	0.845 (0.64-0.94)	0.2173	143,429.2 (<.001)	100	
Abdomen	7 (18)	127,375	0.73-1.00	0.951 (0.13-0.99)	1.4650	84,686.0 (<.001)	100	
Chest and abdomen	3 (8)	261,949	0.76-0.93	0.880 (0.76-0.96)	0.0186	2849.4 (<.001)	100	
Wrist	3 (8)	113,432	0.82-0.88	0.841 (0.80-0.88)	0.0024	790.1 (<.001)	100	

XSL•FO
RenderX

Groups	Studies (N=38), n[a] (%)	Sample size, N	Accuracy (%), range	Pooled mean accuracy (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Data set size, n								.90
<10,000	12 (32)	34,088	0.60-0.97	0.863 (0.39-0.98)	0.4747	1136.8 (<.001)	100	
10,000-50,000	11 (29)	173,171	0.73-1.00	0.907 (0.69-0.97)	0.6016	89,963.2 (<.001)	100	
>50,000	10 (26)	2,218,474	0.73-0.97	0.881 (0.71-0.95)	0.1391	343,840.9 (<.001)	100	
Data type								.30
Respiration data	5 (13)	217,684	0.69-1.00	0.962 (0.53-1.0)	0.9962	56,612.8 (<.001)	100	
HR[d] data	2 (5)	27,735	0.82-0.90	0.864 (0.77-0.94)	0.0078	240.24 (<.001)	100	
Respiration data and HR data	6 (16)	104,439	0.73-0.84	0.813 (0.77-0.85)	0.0116	852.8 (<.001)	100	
Respiration data and SpO_2[e]	10 (26)	2,108,997	0.76-0.97	0.896 (0.75-0.96)	0.1341	319,067.7 (<.001)	100	
Respiration data, HR data, and body movement	12 (32)	238,494	0.60-0.88	0.787 (0.69-0.86)	0.0196	9958.4 (<.001)	100	
Ground truth								.47
Polysomnography	17 (45)	971,896	0.69-0.97	0.877 (0.81-0.92)	0.1433	185,452.6 (<.001)	100	
WD	9 (24)	1,511,429	0.76-1.00	0.949 (0.19-1.0)	1.4613	148,018.6 (<.001)	100	
Experiment context	10 (26)	218,980	0.60-0.93	0.856 (0.45-0.97)	0.2302	13,488.7 (<.001)	100	
Validation method								.31
K-fold cross-validation	12 (32)	2,173,814	0.69-0.97	0.835 (0.64-0.93)	0.1643	347,304 (<.001)	100	
Train-test split	24 (63)	528,491	0.60-1.00	0.911 (0.82-0.96)	0.3667	106,748.9 (<.001)	100	
Overall accuracy	36 (95)	2,702,305	0.60-1.0	0.893 (0.82-0.94)	0.3130	457,567.0 (<.001)	100	__[f]

[a]Many studies were included >1 time in most meta-analyses, given that the studies assessed the performance of >1 algorithm.

[b]Italicized values are statistically significant (*P*<.05).

[c]WD: wearable device.

[d]HR: heart rate.

[e]SpO_2: blood oxygen saturation.

[f]Not applicable.

Sensitivity

As shown in Table 5, meta-analyses were carried out on 22 estimates of sensitivity derived from 872,443 respiratory events across 15 (39%) of the 38 studies. The pooled mean sensitivity of these estimates was 0.793 (95% CI 0.67-0.87). The meta-analyzed evidence has considerable statistical heterogeneity (*P*<.001; I^2=100%). With regard to subgroup analyses, there was no statistically significant difference in the pooled mean sensitivity between subgroups in all groups.

XSL•FO
RenderX

Table 5. Pooled mean estimates of sensitivity in detecting respiratory events by several factors.

Groups	Studies, n[a] (%)	Sample size, N	Sensitivity (%), range	Pooled mean sensitivity (%; 95% CI)	Heterogeneity measures			Test for subgroup difference, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Algorithm								.39
Convolutional neural network	8 (21)	107,274	0.25-0.94	0.752 (0.56-0.90)	0.0862	14,383.9 (<.001)	100	
Recurrent neural networks	6 (16)	279,369	0.68-0.89	0.799 (0.72-0.86)	0.0119	7460.8 (<.001)	100	
Random forest	2 (5)	141,601	0.68-0.80	0.737 (0.61-0.84)	0.0092	1628.9 (<.001)	100	
K-nearest neighbors	2 (5)	114,368	0.30-0.69	0.499 (0.15-0.85)	0.0794	228.8 (<.01)	100	
Type of algorithm								.80
Machine learning	6 (16)	370,337	0.30-0.80	0.682 (0.50-0.80)	0.0696	3872.9 (<.001)	100	
Deep learning	16 (42)	502,106	0.25-0.98	0.819 (0.69-0.90)	0.1397	42,230.7 (<.001)	100	
Sample size, n								.41
<100	10 (26)	41,761	0.25-0.98	0.813 (0.58-0.92)	0.1676	4692.1 (<.001)	100	
100-200	3 (8)	57,022	0.70-0.87	0.801 (0.70-0.89)	0.0113	658.0 (<.001)	100	
<200	9 (24)	773,660	0.44-0.94	0.718 (0.38-0.89)	0.0802	55,208.4 (<.001)	100	
Type of sleep apnea								.99
All	20 (53)	862,060	0.25-0.98	0.791 (0.67-0.87)	0.1006	58,431.0 (<.001)	100	
Obstructive sleep apnea	2 (5)	10,383	0.44-0.93	0.724 (0.18-1.0)	0.1709	2217.2 (<.001)	100	
Status of the WD[b]								.05
Commercial	16 (42)	848,034	0.30-0.94	0.726 (0.61-0.81)	0.0214	58,113.7 (<.001)	100	
Noncommercial	6 (16)	24,409	0.25-0.98	0.830 (0.60-0.97)	0.1051	1790.4 (<.001)	100	
Placement of the WD								.36
Chest	6 (16)	58,175	0.30-0.93	0.745 (0.45-0.89)	0.0079	1621.1 (<.001)	100	
Chest and abdomen	3 (8)	55,783	0.78-0.87	0.826 (0.77-0.87)	0.0032	285.5 (<.001)	99	
Wrist	3 (8)	18,548	0.44-0.70	0.617 (0.44-0.78)	0.0238	1316.3 (<.001)	100	
Data set size, n								.63
<10,000	4 (11)	1657	0.30-0.94	0.796 (0.71-0.97)	0.1338	191.7 (<.001)	99	
10,000-50,000	5 (13)	13,137	0.25-0.98	0.768 (0.48-0.96)	0.1224	2218.8 (<.001)	100	
>50,000	10 (26)	800,627	0.44-0.94	0.718 (0.48-0.86)	0.0523	57,265.8 (<.001)	100	
Data type								.41
Respiration data	5 (13)	35,749	0.30-0.98	0.888 (0.37-0.98)	0.3566	2046.3 (<.001)	100	

XSL•FO
RenderX

Groups	Studies, n[a] (%)	Sample size, N	Sensitivity (%), range	Pooled mean sensitivity (%; 95% CI)	Heterogeneity measures			Test for subgroup difference, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
HR[c] data	2 (5)	7956	0.70-0.93	0.835 (0.56-0.99)	0.0499	600.2 (<.001)	100	
Respiration data and SpO_2[d]	10 (26)	787,542	0.68-0.94	0.814 (0.73-0.87)	0.0876	52,794.7 (<.001)	100	
Respiration data, HR data, and body movement	4 (11)	40,457	0.44-0.80	0.658 (0.50-0.80)	0.0253	1688.2 (<.001)	100	
Ground truth								.29
Polysomnography	11 (29)	95,891	0.25-0.94	0.726 (0.53-0.85)	0.2373	7131.5 (<.001)	100	
WD	9 (24)	742,136	0.69-0.98	0.900 (0.55-0.98)	0.1081	51,466.2 (<.001)	100	
Experiment context	2 (5)	34,416	0.80-0.82	0.813 (0.79-0.83)	0.0002	2.47 (.12)	60	
Validation method								.36
K-fold cross-validation	12 (32)	795,959	0.30-0.94	0.743 (0.59-0.85)	0.0253	54,250.9 (<.001)	100	
Train-test split	10 (26)	76,484	0.25-0.98	0.770 (0.61-0.90)	0.0765	8020.8 (<.001)	100	
Overall sensitivity	22 (58)	872,443	0.25-0.98	0.793 (0.67-0.87)	0.1196	62,433.8 (<.001)	100	__[e]

[a]Many studies were included >1 time in most meta-analyses, given that the studies assessed the performance of >1 algorithm.

[b]WD: wearable device.

[c]HR: heart rate.

[d]SpO_2: blood oxygen saturation.

[e]Not applicable.

Specificity

Meta-analyses were performed to pool 22 estimates of specificity derived from 1,699,503 respiratory events across 15 (39%) of the 38 studies (Table 6). The pooled mean specificity of these estimates was 0.946 (95% CI 0.88-0.98). There was considerable statistical heterogeneity (P<.001; I^2=100%) in the meta-analyzed studies. We also found a statistically significant difference in the pooled mean specificity between subgroups in the "status of wearable device" group (P=.01), while there was no statistically significant difference (P>.05) in the pooled mean specificity between subgroups in the rest of the groups.

XSL•FO
RenderX

Table 6. Pooled mean estimates of specificity in detecting respiratory events by several factors.

Groups	Studies, n[a] (%)	Sample size, N	Specificity (%), range	Pooled mean specificity (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Algorithm								.14
Convolutional neural network	8 (21)	298,315	0.72-0.99	0.932 (0.87-0.98)	0.0236	6463.7 (<.001)	100	
Recurrent neural networks	6 (16)	385,722	0.76-0.95	0.870 (0.81-0.92)	0.0090	11,722.7 (<.001)	100	
Random forest	2 (5)	647,897	0.85-0.98	0.930 (0.74-1.00)	0.0374	28,107.9 (<.001)	100	
K-nearest neighbors	2 (5)	116,039	0.76-0.86	0.812 (0.71-0.90)	0.0079	127.1 (<.01)	99	
Type of algorithm								.10
Machine learning	6 (16)	879,975	0.70-0.98	0.910 (0.38-0.99)	0.5700	259,327.0 (<.001)	100	
Deep learning	16 (42)	819,528	0.72-1.00	0.949 (0.87-0.98)	0.6454	158,624.8 (<.001)	100	
Sample size, n								.94
<100	10 (26)	154,202	0.70-1.00	0.951 (0.79-0.99)	1.0272	123,494.7 (<.001)	100	
100-200	3 (8)	219,550	0.84-0.95	0.922 (0.85-0.97)	0.0112	2860.8 (<.001)	100	
<200	9 (24)	1,325,751	0.85-0.98	0.949 (0.81-0.99)	0.2549	340,888.9 (<.001)	100	
Type of sleep apnea								.97
All	20 (53)	1,642,769	0.70-1.00	0.947 (0.86-0.98)	0.7459	487,691.1 (<.001)	100	
Obstructive sleep apnea	2 (5)	56,734	0.94-0.95	0.943 (0.94-0.95)	0.0100	3.0 (.09)	66	
Status of the WD[b]								*.01* [c]
Commercial	16 (42)	1,629,032	0.70-0.98	0.887 (0.79-0.94)	0.2257	413,387.5 (<.001)	100	
Noncommercial	6 (16)	70,471	0.85-1.00	0.969 (0.92-1.00)	0.0181	2607.4 (<.001)	100	
Placement of the WD								.97
Chest	6 (16)	634,960	0.70-0.98	0.900 (0.65-0.97)	0.3678	157,806.3 (<.001)	100	
Chest and abdomen	3 (8)	206,166	0.72-0.95	0.893 (0.73-0.99)	0.0370	4312.8 (<.001)	100	
Wrist	3 (8)	94,884	0.84-0.94	0.885 (0.82-0.94)	0.0074	2106.2 (<.001)	100	
Data set size, n								.64
<10,000	4 (11)	6511	0.70-0.99	0.923 (0.06-1.00)	0.8131	416.7 (<.001)	100	
10,000-50,000	5 (13)	55,595	0.72-1.00	0.937 (0.81-1.00)	0.0482	9402.4 (<.001)	100	
>50,000	10 (26)	1,417,847	0.76-0.98	0.888 (0.84-0.93)	0.1031	93,530.5 (<.001)	100	
Data type								.37

XSL•FO
RenderX

Groups	Studies, n[a] (%)	Sample size, N	Specificity (%), range	Pooled mean specificity (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Respiration data	5 (13)	181,935	0.70-1.00	0.977 (0.61-1.00)	1.1621	53,359.8 (<.001)	100	
HR[d] data	2 (5)	19,779	0.86-0.95	0.908 (0.80-0.98)	0.0121	217.8 (<.001)	99	
Respiration data and SpO_2[e]	10 (26)	1,321,455	0.72-0.98	0.925 (0.72-0.98)	0.3791	348,529.0 (<.001)	100	
Respiration data, HR data, and body movement	4 (11)	172,117	0.76-0.94	0.864 (0.67-0.95)	0.0117	9381.8 (<.001)	100	
Ground truth								.90
Polysomnography	11 (29)	771,566	0.70-1.00	0.948 (0.86-0.98)	0.5619	202,709.3 (<.001)	100	
WD	9 (24)	769,293	0.72-1.00	0.957 (0.11-1.00)	1.8198	103,183.0 (<.001)	100	
Experiment context	2 (5)	158,644	0.85-0.95	0.908 (0.78-0.98)	0.0158	293.1 (<.001)	100	
Validation method								.10
K-fold cross-validation	12 (32)	1,377,855	0.70-0.98	0.866 (0.61-0.96)	0.3674	401,832.4 (<.001)	100	
Train-test split	10 (26)	321,648	0.84-1.00	0.947 (0.90-0.98)	0.0178	8482.6 (<.001)	100	
Overall specificity	22 (58)	1,699,503	0.70-1.00	0.946 (0.88-0.98)	0.6373	487,706.6 (<.001)	100	—[f]

[a]Many studies were included more than one time in all meta-analyses given that the studies assessed the performance of more than one algorithm.

[b]WD: wearable device.

[c]Italicized values are statistically significant (*P*<.05).

[d]HR: heart rate.

[e]SpO_2: blood oxygen saturation.

[f]Not applicable.

Type of Apnea Events in Respiration

We conducted meta-analyses of 6 estimates of accuracy derived from 637,250 respiratory events across 6 (16%) of the 38 studies (Table 7). The pooled mean accuracy of these estimates was 0.815 (95% CI 0.64-0.94). The meta-analyzed studies exhibited considerable statistical heterogeneity (*P*<.001; I^2=100%). In addition, there was a statistically significant difference in the pooled mean accuracy between subgroups in the "data type" group (*P*=.001), while no statistically significant difference (*P*>.05) was found in the pooled mean accuracy between subgroups in the remaining groups.

XSL•FO
RenderX

Table 7. Pooled mean estimates of accuracy in detecting the type of respiratory events by several factors.

Groups	Studies, n (%)	Sample size, N	Accuracy (%), range	Pooled mean accuracy (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Algorithm								.76
Convolutional neural network	4 (11)	459,163	0.40-0.97	0.829 (0.55-0.99)	0.0973	26,906.1 (<.001)	100	
Long short-term memory	2 (5)	178,087	0.73-0.84	0.788 (0.68-0.88)	0.0073	2299.2 (<.001)	100	
Sample size, n								.27
≤100	3 (8)	309,033	0.74-0.97	0.892 (0.73-0.98)	0.0328	21,243.3 (<.001)	100	
>100	3 (8)	328,217	0.40-0.88	0.724 (0.41-0.95)	0.0824	19,923.0 (<.001)	100	
Status of the WD[a]								.25
Commercial	4 (11)	578,076	0.40-0.93	0.759 (0.51-0.94)	0.0735	41,330.6 (<.001)	100	
Noncommercial	2 (5)	59,174	0.84-0.97	0.909 (0.74-0.99)	0.0249	30.9 (<.01)	97	
Data type								*.001* [b]
Respiration data	2 (5)	189,970	0.93-0.97	0.944 (0.91-0.97)	00022	3.7 (.05)	73	
Respiration data and SpO_2[c]	2 (5)	308,810	0.84-0.88	0.857 (0.81-0.90)	0.0018	697.4 (<.01)	100	
Respiration data, HR[d] data, and body movement	2 (5)	138,470	0.40-0.74	0.574 (0.25-0.87)	0.0590	7875.2 (<.001)	100	
Ground truth								.19
Polysomnography	4 (11)	197,644	0.40-0.97	0.762 (0.49-0.95)	0.0822	12,665.3 (<.001)	100	
Nonpolysomnography	2 (5)	439,606	0.88-0.93	0.905 (0.85-0.95)	0.0039	3353.7 (<.001)	100	
Validation method								.97
K-fold cross-validation	2 (5)	368,849	0.74-0.88	0.812 (0.66-0.93)	0.0164	10,579.9 (<.001)	100	
Train-test split	4 (11)	268,401	0.40-0.97	0.818 (0.54-0.98)	0.0954	28,010.1 (<.001)	100	
Overall accuracy	6 (16)	637,250	0.40-0.97	0.815 (0.64-0.94)	0.0603	41,608.1 (<.001)	100	__[e]

[a]WD: wearable device.

[b]Italicized values are statistically significant (*P*<.05).

[c]SpO_2: blood oxygen saturation.

[d]HR: heart rate.

[e]Not applicable.

Patients With Sleep Apnea

Accuracy

We carried out meta-analyses of 13 estimates of accuracy derived from 2015 participants across 13 (34%) of the 38 studies (Table 8). The pooled mean accuracy of these estimates was 0.869 (95% CI 0.81-0.92). The meta-analyzed estimates showed considerable statistical heterogeneity (*P*<.001; I^2=100%). Further, there was a statistically significant difference in the pooled mean accuracy between subgroups in the "type of sleep apnea" group (*P*=.049). However, no statistically significant difference (*P*>.05) was found in the pooled mean accuracy between subgroups in the remaining groups.

XSL•FO
RenderX

Table 8. Pooled mean estimates of accuracy in detecting sleep apnea by several factors.

Groups	Studies, n (%)	Sample size, N	Accuracy (%), range	Pooled mean accuracy (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Type of algorithm								.20
Machine learning	3 (8)	1141	0.88-0.92	0.896 (0.87-0.92)	0.0003	2.2 (.34)	7	
Deep learning	9 (24)	678	0.71-1.00	0.849 (0.76-0.92)	0.0226	55.4 (<.01)	86	
Sample size, n								.20
≤100	8 (21)	496	0.71-0.96	0.838 (0.77-0.90)	0.0095	23.3 (<.01)	70	
>100	5 (13)	1519	0.75-1.00	0.905 (0.81-0.97)	0.0224	45.5 (<.01)	91	
Type of sleep apnea								*.049* [a]
All	6 (16)	671	0.78-1.00	0.920 (0.84-0.97)	0.0163	24.4 (<.01)	80	
Obstructive sleep apnea	7 (18)	1344	0.71-0.91	0.823 (0.76-0.88)	0.0091	40.5 (<.01)	85	
Status of the WD[b]								.18
Commercial	9 (24)	1632	0.71-0.96	0.841 (0.78-0.89)	0.0098	49.5 (<.01)	84	
Noncommercial	4 (11)	383	0.78-1.00	0.923 (0.81-0.99)	0.0252	22.4 (<.01)	87	
Placement of the WD								.17
Wrist	4 (11)	982	0.74-0.96	0.840 (0.73-0.93)	0.0165	28.4 (<.01)	98	
Finger	3 (8)	212	0.71-0.86	0.805 (0.70-0.89)	0.0070	5.99 (.05)	67	
Chest and abdomen	3 (8)	147	0.86-1.00	0.949 (0.83-1.00)	0.0230	10.6 (<.01)	81	
Data type								.09
Respiration data and SpO_2[c]	4 (11)	556	0.86-1.00	0.938 (0.86-0.99)	0.0139	13.4 (<.01)	78	
Respiration data, HR[d] data, and body movement	4 (11)	408	0.71-0.91	0.840 (0.74-0.92)	0.0111	17.1 (<.01)	82	
Ground truth								.12
Polysomnography	11 (29)	1908	0.71-1.00	0.878 (0.82-0.93)	0.0164	71.1 (<.01)	86	
WD	2 (5)	107	0.74-0.86	0.789 (0.67-0.89)	0.0036	1.6 (.20)	38	
Validation method								.91
K-fold cross-validation	4 (11)	1177	0.78-0.91	0.873 (0.82-0.92)	0.0035	8.2 (.04)	63	
Train-test split	7 (18)	698	0.71-1.00	0.880 (0.78-0.95)	0.0267	59.2 (<.01)	90	
Leave-one-out cross-validation	2 (5)	140	0.74-0.92	0.839 (0.64-0.97)	0.0240	7.7 (<.01)	87	
Overall accuracy	13 (34)	2015	0.71-1.00	0.869 (0.81-0.92)	0.0156	80.3 (<.001)	100	__[e]

[a]Italicized values are statistically significant (*P*<.05).

[b]WD: wearable device.

[c]SpO_2: blood oxygen saturation.

[d]HR: heart rate.

[e]Not applicable.

XSL•FO
RenderX

Sensitivity

As shown in Table 9, meta-analyses were carried out on 13 estimates of sensitivity derived from 1580 participants across 13 (34%) of the 38 studies. The pooled mean sensitivity of these estimates was 0.938 (95% CI 0.89-0.97). The meta-analyzed evidence has considerable statistical heterogeneity (P<.001; I^2=82%). With regards to subgroup analyses, there was no statistically significant difference in the pooled mean sensitivity between subgroups in all groups except for the "placement of wearable device" group (P<.001).

XSL•FO
RenderX

Table 9. Pooled mean estimates of sensitivity in detecting sleep apnea by several factors.

Groups	Studies, n (%)	Sample size, N	Sensitivity (%), range	Pooled mean sensitivity (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Type of algorithm								.78
Machine learning	3 (8)	921	0.89-0.98	0.926 (0.88-0.98)	0.0038	6.9 (.03)	71	
Deep learning	9 (24)	485	0.77-1.00	0.942 (0.87-0.99)	0.0256	59.0 (<.01)	86	
Sample size, n								.50
≤100	8 (21)	363	0.77-1.00	0.953 (0.90-0.99)	0.0154	28.0 (<.01)	75	
>100	5 (13)	1217	0.77-1.00	0.917 (0.83-0.97)	0.0196	33.9 (<.01)	88	
Type of sleep apnea								.06
All	6 (16)	456	0.90-1.00	0.959 (0.93-1.00)	0.0059	13.0 (.02)	62	
Obstructive sleep apnea	7 (18)	1124	0.77-1.00	0.903 (0.83-0.96)	0.0179	37.8 (<.01)	84	
Status of the WD[a]								.06
Commercial	9 (24)	1254	0.77-1.00	0.916 (0.85-0.96)	0.0170	43.2 (<.01)	81	
Noncommercial	4 (11)	326	0.93-1.00	0.974 (0.93-1.00)	0.0048	7.5 (.06)	60	
Placement of the WD								*<.001* [b]
Wrist	4 (11)	833	0.77-0.90	0.837 (0.76-0.91)	0.0073	16.1 (<.01)	81	
Finger	3 (8)	148	0.92-1.00	0.966 (0.90-1.00)	0.0000	5.3 (.07)	62	
Chest and abdomen	3 (8)	130	0.98-1.00	0.997 (0.97-1.00)	0.0083	1.1 (.59)	0	
Data type								.39
Respiration data and SpO_2[c]	4 (11)	378	0.92-1.00	0.980 (0.93-1.00)	0.0080	11.0 (.01)	73	
Respiration data, HR[d] data, and body movement	4 (11)	322	0.92-0.95	0.954 (0.91-0.99)	0.0040	6.2 (.10)	52	
Ground truth								.80
Polysomnography	11 (29)	1495	0.77-1.00	0.941 (0.89-0.97)	0.0126	52.6 (<.01)	81	
WD	2 (5)	85	0.77-1.00	0.917 (0.57-1.00)	0.0773	12.1 (<.01)	92	
Validation method								.89
K-fold cross-validation	4 (11)	941	0.89-1.00	0.944 (0.89-0.98)	0.0063	10.7 (.01)	72	
Train-test split	7 (18)	527	0.77-1.00	0.941 (0.87-0.99)	0.0199	41.9 (<.01)	86	
Leave-one-out cross-validation	2 (5)	112	0.77-0.98	0.896 (0.61-1.00)	0.0542	13.2 (<.01)	92	
Overall sensitivity	13 (34)	1580	0.77-1.00	0.938 (0.89-0.97)	0.0162	67.0 (<.001)	82	__[e]

[a]WD: wearable device.

[b]Italicized values are statistically significant (*P*<.05).

[c]SpO_2: blood oxygen saturation.

[d]HR: heart rate.

[e]Not applicable.

XSL•FO
RenderX

Specificity

Meta-analyses were performed to pool 13 estimates of specificity derived from 436 participants across 13 (34%) of the 38 studies (Table 10). The pooled mean specificity of these estimates was 0.752 (95% CI 0.63-0.86). There was considerable statistical heterogeneity (P<.001; I^2=78%) in the meta-analyzed studies. Our subgroup meta-analyses showed that there was no statistically significant difference in the pooled mean specificity between subgroups in all groups.

XSL•FO
RenderX

Table 10. Pooled mean estimates of specificity in detecting sleep apnea by several factors.

Groups	Studies, n (%)	Sample size, N	Specificity (%), range	Pooled mean specificity (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Type of algorithm								.53
Machine learning	3 (8)	220	0.60-0.88	0.796 (0.63-0.92)	0.0168	7.9 (.02)	75	
Deep learning	9 (24)	194	0.29-1.00	0.735 (0.55-0.89)	0.0575	37.1 (<.01)	78	
Sample size, n								.21
≤100	8 (21)	133	0.29-1.00	0.690 (0.48-0.87)	0.0615	35.8 (<.01)	80	
>100	5 (13)	303	0.72-1.00	0.818 (0.72-0.90)	0.0077	11.6 (.02)	65	
Type of sleep apnea								.62
All	6 (16)	220	0.29-0.89	0.810 (0.55-0.99)	0.0780	34.6 (<.01)	86	
Obstructive sleep apnea	7 (18)	216	0.36-1.00	0.730 (0.64-0.81)	0.0068	12.3 (.06)	51	
Status of the WD[a]								.43
Commercial	9 (24)	378	0.29-1.00	0.784 (0.66-0.89)	0.0295	32.9 (<.01)	76	
Noncommercial	4 (11)	58	0.36-1.00	0.672 (0.39-0.91)	0.0497	11.3 (.01)	73	
Placement of the WD								.71
Wrist	4 (11)	149	0.72-1.00	0.803 (0.62-0.94)	0.0309	10.9 (.01)	73	
Finger	3 (8)	64	0.29-0.88	0.658 (0.29-0.89)	0.0727	11.5 (<.01)	83	
Chest and abdomen	3 (8)	18	0.60-1.00	0.777 (0.47-0.99)	0.0157	2.7 (.26)	26	
Data type								.35
Respiration data and SpO_2[b]	4 (11)	179	0.60-1.00	0.855 (0.66-0.99)	0.0186	5.6 (.13)	47	
Respiration data, HR[c] data, and body movement	4 (11)	86	0.29-0.89	0.700 (0.46-0.90)	0.0411	11.5 (<.01)	74	
Ground truth								.53
Polysomnography	11 (29)	414	0.29-1.00	0.756 (0.62-0.88)	0.0444	52.8 (<.01)	81	
WD	2 (5)	22	0.67-0.75	0.691 (0.46-0.89)	0.0000	0.04 (.85)	0	
Validation method								.48
K-fold cross-validation	4 (11)	236	0.36-0.88	0.713 (0.44-0.93)	0.0547	26.9 (<.01)	89	
Train-test split	7 (18)	172	0.29-1.00	0.793 (0.72-0.86)	0.0456	22.9 (<.01)	74	
Leave-one-out cross-validation	2 (5)	28	0.60-0.67	0.644 (0.45-0.82)	0.0000	0.13 (.72)	0	
Overall specificity	13 (34)	436	0.29-1.00	0.752 (0.63-0.86)	0.0366	54.5 (<.001)	78	—[d]

[a]WD: wearable device.

[b]SpO_2: blood oxygen saturation.

[c]HR: heart rate.

[d]Not applicable.

XSL•FO
RenderX

Severity of Sleep Apnea

Accuracy

We performed meta-analyses of 9 estimates of accuracy derived from 1661 participants across 9 (24%) of the 38 studies (Table 11). The pooled mean accuracy of these estimates was 0.651 (95% CI 0.54-0.75). The meta-analyzed studies exhibited considerable statistical heterogeneity (P<.001; I^2=93%). In addition, there was a statistically significant difference in the pooled mean accuracy between subgroups in "type of sleep apnea" group (P=.03) and "data type" group (P=.01), while no statistically significant difference (P>.05) was found in the pooled mean accuracy between subgroups in the remaining groups.

Table 11. Pooled mean estimates of accuracy in detecting the severity of sleep apnea by several factors.

Groups	Studies, n (%)	Sample size, N	Accuracy (%), range	Pooled mean accuracy (%; 95% CI)	Heterogeneity measures			Test for subgroup difference, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Type of algorithm								.28
Machine learning	3 (8)	1141	0.63-0.80	0.716 (0.60-0.82)	0.0105	40.6 (<.01)	95	
Deep learning	6 (16)	520	0.36-0.89	0.615 (0.46-0.76)	0.0318	47.8 (<.01)	90	
Sample size, n								.28
≤100	4 (11)	274	0.36-0.71	0.584 (0.43-0.74)	0.0219	22.4 (<.01)	87	
>100	5 (13)	1387	0.55-0.89	0.698 (0.56-0.82)	0.0252	72.9 (<.01)	95	
Type of sleep apnea								*.03* [a]
All	4 (11)	591	0.58-0.89	0.757 (0.62-0.87)	0.0196	20.3 (<.01)	85	
Obstructive sleep apnea	5 (13)	1070	0.36-0.67	0.564 (0.46-0.66)	0.0111	23.3 (<.01)	83	
Status of the WD[b]								.08
Commercial	7 (18)	1543	0.36-0.80	0.606 (0.50-0.71)	0.0219	87.9 (<.01)	93	
Noncommercial	2 (5)	118	0.71-0.89	0.809 (0.60-0.95)	0.0181	6.2 (.01)	84	
Placement of the WD								.11
Wrist	3 (8)	922	0.55-0.63	0.596 (0.54-0.65)	0.0011	3.5 (.18)	42	
Finger	3 (8)	212	0.36-0.67	0.542 (0.35-0.72)	0.0235	16.5 (<.01)	88	
Chest and abdomen	2 (5)	118	0.71-0.89	0.809 (0.60-0.95)	0.0219	6.2 (.01)	84	
Data type								*.01*
Respiration data and SpO_2[c]	3 (8)	527	0.71-0.89	0.807 (0.71-0.89)	0.0073	6.2 (.04)	68	
Body movement, HR[d] data, and SpO_2	3 (8)	212	0.36-0.67	0.542 (0.35-0.72)	0.0235	16.5 (<.01)	88	
Validation method								.37
K-fold cross-validation	2 (5)	1079	0.63-0.80	0.719 (0.53-0.87)	0.0195	40.6 (<.01)	98	
Train-test split	6 (16)	520	0.36-0.89	0.615 (0.46-0.76)	0.0318	47.8 (<.01)	90	
Overall accuracy	9 (24)	1661	0.36-0.89	0.651 (0.54-0.75)	0.0243	106.1 (<.001)	93	__[e]

[a]Italicized values are statistically significant (P<.05).

[b]WD: wearable device.

[c]SpO_2: blood oxygen saturation.

[d]HR: heart rate.

[e]Not applicable.

XSL•FO
RenderX

Correlation Coefficient

As shown in Table 12, meta-analyses were carried out on 12 estimates of correlation coefficient (*r*) derived from 1266 participants across 12 (32%) of the 38 studies. The pooled mean *r* of these estimates was 0.877 (95% CI 0.82-0.92). The meta-analyzed evidence has considerable statistical heterogeneity ($P<.001$; $I^2=82\%$). With regard to subgroup analyses, there was a statistically significant difference in the pooled mean r between subgroups in the "placement of wearable device" group ($P<.001$) and the "data type" group ($P<.001$). However, no statistically significant difference ($P>.05$) was found in the pooled mean *r* between subgroups in the remaining groups.

XSL•FO
RenderX

Table 12. Pooled mean estimates of correlation coefficient (r) in detecting the severity of sleep apnea by several factors.

Groups	Studies, n (%)	Sample size, N	Correlation coefficient (%), range	Pooled mean correlation coefficient (%; 95% CI)	Heterogeneity measures			Test for subgroup differences, *P* value
					Tau^2	Q (*P* value)	I^2 (%)	
Type of algorithm								.12
Machine learning	3 (8)	526	0.90 to 0.96	0.922 (0.66 to0.98)	0.0987	25.6 (<.01)	92	
Deep learning	9 (24)	740	0.64 to 0.91	0.856 (0.79 to0.90)	0.0552	53.1 (<.01)	85	
Sample size, n								.79
≤100	9 (24)	541	0.64 to 0.90	0.879 (0.85 to0.90)	0.0001	12.7 (.12)	37	
>100	3 (8)	725	0.71 to 0.96	0.896 (0.12 to0.99)	0.2849	173.0 (<.01)	99	
Type of sleep apnea								.54
All	8 (21)	802	0.64 to 0.96	0.886 (0.81 to0.93)	0.0918	80.2 (<.01)	91	
Obstructive sleep apnea	4 (11)	464	0.71 to 0.90	0.859 (0.68 to0.94)	0.759	34.7 (<.01)	91	
Status of the WD[a]								.35
Commercial	10 (26)	1177	0.64 to 0.96	0.881 (0.82 to 0.92)	0.0977	191.1 (<.01)	95	
Noncommercial	2 (5)	89	0.84 to 0.86	0.856 (0.72 to 0.93)	0.0000	0.1 (.79)	0	
Placement of the WD								*<.001* [b]
Wrist	2 (5)	316	0.71 to 0.91	0.833 (–0.99 to 1.00)	0.1948	20.1 (<.01)	95	
Chest	2 (5)	452	0.87 to 0.96	0.929 (–0.98 to 1.00)	0.1765	16.7 (<.01)	94	
Finger	3 (8)	212	0.89 to 0.90	0.894 (0.88 to0.91)	0.0000	0.1 (.93)	0	
Neck	2 (5)	89	0.84 to 0.86	0.856 (0.72 to0.93)	0.0000	0.1 (.79)	0	
Abdomen	2 (5)	158	0.89 to 0.90	0.894 (0.81 to 0.94)	0.0000	0.1 (.76)	0	
Data type								<.001
Respiration data	2 (5)	89	0.84 to 0.86	0.856 (0.72 to 0.93)	0.0000	0.1 (.79)	0	
Respiration data and SpO_2[c]	2 (5)	438	0.64 to 0.96	0.878 (–1.00 to 1.00)	0.6849	34.5 (<.01)	97	
Respiration data and HR[d] data	2 (5)	158	0.89 to 0.90	0.894 (0.81 to 0.94)	0.0000	0.1 (.76)	0	
Respiration data, HR data, and body movement	3 (8)	369	0.71 to 0.91	0.844 (0.38 to 0.97)	0.1030	24.5 (<.01)	92	
Body movement, HR data, and SpO_2	3 (8)	212	0.89 to 0.90	0.894 (0.88 to 0.91)	0.0000	0.1 (.93)	0	
Validation method								.90
K-fold cross-validation	4 (11)	527	0.64 to 0.96	0.869 (0.49-0.97)	0.2315	59.3 (<.01)	95	
Train-test split	7 (18)	450	0.71 to 0.91	0.877 (0.82-0.92)	0.0486	50.8 (<.01)	88	
Overall accuracy	12 (32)	1266	0.64 to 0.96	0.877 (0.82-0.92)	0.0828	194.5 (<.001)	94	__[e]

[a]WD: wearable device.

[b]Italicized values are statistically significant (*P*<.05).

XSL•FO
RenderX

[c]SpO_2: blood oxygen saturation.

[d]HR: heart rate.

[e]Not applicable.

Discussion

Principal Findings

This systematic review investigated how well wearable AI performs in detecting sleep apnea. Overall, the findings indicate that wearable AI demonstrated a performance that is deemed acceptable, although not optimal, for detecting sleep apnea. Specifically, wearable AI was able to correctly classify apnea events and nonapnea events in 89.3% of respiratory events. This performance was notably higher when using CNN in particular or deep learning algorithms in general. The superiority of CNN architectures can be attributed to their ability to capture the localized dependencies inherent in apnea patterns through convolution kernels. The meta-analyses conducted in this review revealed that wearable AI performed better in detecting nonapnea respiratory events (94.6%) compared to apnea respiratory events (79.3%). This could be linked to the training of AI models using an unrepresentative sample, wherein the number of nonapnea respiratory events (n=1,699,503) was approximately twice as high as the number of apnea respiratory events (n=872,443). This highlights the challenge of applying data balancing techniques for heterogeneous and time-dependent measurements, particularly evident in longitudinal recordings as observed in apnea studies.

Although the sensitivity of wearable AI in detecting apnea events in respiration remained unaffected by any moderating factors, the specificity was influenced by the status of the wearable device, where noncommercial devices exhibited higher specificity than commercial devices. This can be because all studies that used noncommercial wearable devices applied deep learning algorithms, whereas more than one-third (6/16, 38%) of studies that used commercial wearable devices applied ML algorithms (eg, random forest, AdaBoost, and k-nearest neighbors). Introducing scalable AI models, such as deep learning models, into commercial apnea detection applications presents challenges due to their computational expense and resource requirements, thereby complicating market penetration and impacting profit margins. However, recent advancements in tiny ML models and edge AI implementations offer potential solutions to mitigate these challenges. This review also demonstrated that wearable AI was able to correctly differentiate between different types of apnea events (eg, apnea, hypopnea, obstructive apnea, and central apnea) in 81.5% of respiratory events, and this performance was not influenced by any moderators. This can be attributed to the lack of studies (≤4) in all subgroup analyses related to this outcome.

In this review, wearable AI demonstrated 86.9% accuracy in correctly identifying patients with and patients without sleep apnea. This performance was notably higher when the wearable AI was used for detecting sleep apnea in general (92%) rather than OSA in particular (82.3%). This difference may be attributed to the fact that approximately 83% (5/6) of studies focusing on general sleep apnea detection used respiration data to develop the AI models. By contrast, only 29% (2/7) of studies concentrating on OSA detection incorporated respiration data. Given that respiration data are widely acknowledged as the most crucial indicator of sleep apnea, this disparity in use may explain the varying performance levels observed in the review.

Unlike apnea event detection, wearable AI exhibited superior performance in identifying patients with sleep apnea (93.8%) compared to those without sleep apnea (75.2%). This could be associated with the training of AI models using an unrepresentative sample, wherein the number of patients with sleep apnea (1580) was >3 times higher than the number of patients without sleep apnea (436). The specificity of wearable AI in detecting sleep apnea was not affected by any moderator, while its sensitivity was higher when wearable devices were placed on both the chest and abdomen in comparison with other placements (wrist or fingers). This moderation effect could be attributed to the fact that all studies that placed wearable devices on both the chest and abdomen focused on detecting sleep apnea in general, while 6 (86%) out of 7 studies that placed wearable devices in other places focused on detecting OSA in particular. Further, all studies that placed wearable devices on other body parts used commercial wearable devices, whereas only 1 of the studies that placed wearable devices on both the chest and abdomen used commercial wearable devices.

Our meta-analyses also revealed that wearable AI accurately differentiated between various levels of sleep apnea severity (normal, mild, moderate, and severe) in 65.1% of cases. This performance was higher when the wearable AI was used for detecting the severity of sleep apnea in general rather than OSA in particular. This could be linked to the fact that all studies that aimed to detect OSA used commercial devices that were placed on either fingers or wrists, while two-thirds of the studies that focused on sleep apnea in general used noncommercial devices that were placed on both the abdomen and chest. This performance was also higher when the model was developed using both respiration and oxygen saturation data in comparison with using a combination of body movement, HR, and oxygen saturation data. This could be associated with the fact that all studies using the combination of body movement, HR, and oxygen saturation data focused on the detection of OSA using commercial devices placed on fingers, while all studies using both respiration and oxygen saturation data focused on detecting any sleep apnea type using noncommercial devices (in 2, 67% out 3 studies) placed on the abdomen and chest.

Finally, the accuracy of wearable AI in estimating the severity of sleep apnea (ie, the apnea-hypopnea index score) reached 87.7%. This accuracy was higher when the wearable device was placed on the chest and when using both respiration and HR data or a combination of HR, oxygen saturation, and body movement.

Research and Practical Implications

Our analysis revealed that wearable AI shows promise in identifying sleep apnea, distinguishing its type, and gauging its

XSL•FO
RenderX

severity; however, it is not yet ready for widespread use in clinical practices for 3 reasons. First, its current performance falls below the optimal level. Second, only 9 (24%) of the 38 studies were judged to have a low risk of bias in all domains. Third, heterogeneity between studies was considerable in most meta-analyses. Therefore, we cannot propose the use of wearable AI as a replacement for traditional sleep assessments (eg, polysomnography and home sleep apnea testing), but we recommend that wearable AI be used in conjunction with these assessments, taking into account factors such as cost-effectiveness and practical challenges in real-world implementation.

Among all wearable devices used in the included studies, only 1 was specifically designed for diagnosing sleep and monitoring sleep health and obtained clearance from the US Food and Drug Administration. This may be due to a shortage of such wearable devices in the market or a scarcity of studies evaluating them. We urge manufacturers of wearable devices to extend their focus beyond evaluating sleep quality and incorporate AI into their devices for identifying sleep apnea, its various types, and its severity. Further, researchers should pay more attention to such wearable devices in their future studies. The main challenge of conducting such studies is the cost of such wearable devices.

Our meta-analyses indicate that the performance of wearable AI was notably higher when using CNN in particular or deep learning algorithms in general. Therefore, we recommend that manufacturers of wearable devices and researchers prioritize these techniques during the development of devices intended for the detection of sleep apnea. However, obtaining large, high-quality, and standardized data sets for training and validating CNN or deep learning models can be challenging.

Our meta-analysis suggests the need for implementing AI on the edge through specially crafted tiny ML modules with federated learning protocols. Such an approach not only enhances performance metrics but also addresses critical considerations regarding resource efficiency, latency reduction, and privacy preservation inherent in commercial apnea detection systems. However, implementing AI on wearable devices, especially with tiny ML modules, poses challenges related to hardware constraints, such as limited processing power, limited memory, and high energy consumption. Further, ensuring that AI algorithms can run efficiently on resource-constrained devices without compromising performance is a significant challenge. Implementing federated learning protocols for edge devices introduces additional complexities related to communication, synchronization, and security. Designing robust federated learning frameworks that can effectively train AI models across distributed devices while preserving data privacy and security requires careful consideration and expertise.

Most studies included in this review focused on the application of wearable AI for the detection of existing sleep apnea, its type, or its severity, rather than the anticipation of its occurrence. Foreseeing the onset of sleep apnea in the future is as pivotal as, if not more pivotal than, recognizing the current sleep apnea state, as it can pave the way for the development and implementation of proactive interventions. Consequently, we encourage researchers to undertake additional investigations into the capacity of wearable AI to predict future instances of sleep apnea. Such studies collect longitudinal data over an extended period to train and validate predictive models accurately. However, obtaining continuous and comprehensive sleep data from individuals over time can be challenging due to factors such as participant compliance, dropout rates, and the need for long-term monitoring.

In this review, only a single study evaluated the effectiveness of wearable AI in identifying CSA. In addition, only 7 (18%) of the 38 studies investigated the capability of wearable AI to differentiate between different types of sleep apnea. More research is urgently needed to evaluate the performance of wearable AI in these crucial areas. Our study also suggests that more open-source data sets with prepared manual labels for different types of sleep apnea are needed. Collecting large-scale, comprehensive, and well-annotated data from individuals with CSA poses challenges due to the rarity of CSA cases. Furthermore, identifying informative features and physiological signals from wearable devices that can distinguish between different types of sleep apnea is challenging due to the overlap in the clinical presentation and physiological characteristics of different types of sleep apnea, particularly in cases of mixed sleep apnea where both obstructive and central events occur concurrently.

Merely 3 (8%) of the 38 included studies used self-reported data and nonwearable device data alongside wearable device data for the detection of sleep apnea. The inclusion of self-reported data (eg, data regarding demographics, BMI, medical history, family history, and medications) and nonwearable device data (eg, data collected via mobile phones, smart pillows, smart mattresses, voice recorders, and Internet of Things) has the potential to enhance the efficacy of wearable AI in identifying sleep apnea. Hence, manufacturers and researchers are encouraged to take these types of data into consideration, alongside wearable device data, when developing wearable AI for the diagnosis of sleep apnea. However, challenges arise in transferring nonwearable data to wearable devices and the potential impact on the performance of wearable devices in terms of processing speed, memory use, energy consumption, synchronization, and security.

A few studies in this review compared the performance of wearable devices worn on different parts of the body (eg, wrist, abdomen, and chest) and developed wearable AI for not only detecting but also intervening in sleep apnea. This points to a crucial gap in research, which urges further investigation into the different performances of wearable AI with different placements and integrated treatment delivery via wearable AI for sleep apnea management.

Among the 38 studies in our review, 11 (29%) were excluded from the meta-analyses due to insufficient details crucial for their conduct (eg, confusion matrices and the number of apnea and nonapnea cases). They also did not provide multiple performance measures (eg, accuracy, sensitivity, and specificity), which are essential for estimating the necessary information. It is recommended that researchers include these specific details in their reports to facilitate the conduct of meta-analyses by other researchers. However, we acknowledge

XSL•FO
RenderX

that the space constraints imposed by journals and conference proceedings may present a challenge for researchers seeking to include more comprehensive details in their reports.

Limitations

Our review intentionally excluded studies involving (1) nonwearable devices, near-body wearable devices, in-body wearable devices, wearable devices wired to nonwearable devices, and wearable devices requiring an expert for their placement on users; (2) wearable AI in detecting other sleep disorders (eg, insomnia, narcolepsy, and restless legs syndrome); and (3) wearable AI in predicting outcomes of sleep apnea interventions or detecting sleep quality or sleep stages. Therefore, our findings are specifically applicable to wearable AI for sleep apnea detection and may not be generalizable to the excluded devices, disorders, or outcomes. Our findings are based on studies conducted in only 16 countries. Further, while most studies were carried out in hospitals, only 4 (11%) of the 38 studies were conducted in health care centers. Therefore, extrapolating our results to broader populations and clinical settings requires caution. This limitation acknowledges the need for further reviews in these broader areas.

Another limitation of this review is the likelihood of an overestimation or underestimation of the results of our meta-analyses due to 2 reasons. First, some relevant studies could have been overlooked, as our search was confined to English-language publications, and we did not explore other widely used databases, such as CINAHL and Web of Science. Secondly, 11 of the 38 studies in this review were excluded from the meta-analyses, as they did not report details required for meta-analyses.

Conclusions

Our review underscores the potential of wearable AI in identifying sleep apnea, differentiating its type, and gauging its severity. However, wearable AI is not yet ready for integration into routine clinical practices due to its suboptimal performance. Therefore, until further evidence demonstrates an ideal performance, we suggest the concurrent use of wearable AI with traditional sleep apnea assessments (eg, polysomnography and home sleep apnea testing), rather than a complete substitution. Manufacturers need to develop certified commercial wearable devices that can easily detect sleep apnea, predict its occurrence, and deliver proactive interventions. CNN in particular or deep learning algorithms in general should be prioritized during the development of wearable AI for the detection of sleep apnea. Further studies are needed to assess the performance of wearable AI in detecting CSA and distinguishing it from other types of sleep apnea. Researchers should consider incorporating self-reported and nonwearable device data alongside wearable data to enhance the efficacy of wearable AI in detecting sleep apnea. Additional research is required to evaluate the varying performance of wearable devices with different placements. Researchers should also report sufficient details about their findings to enable other researchers to conduct meta-analyses effectively.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

A Abd-alrazaq, A Ahmed, RA, and SA developed the protocol with guidance from and under the supervision of JS. A Abd-alrazaq searched the electronic databases and conducted backward and forward reference list checking. The study selection process and data extraction were carried out by HA and A Abd-alrazaq. A risk-of-bias assessment was conducted by HA and RA. Data synthesis was conducted by A Abd-alrazaq. The *Introduction* and *Methods* sections were written by MA and A Abd-alrazaq. The *Results* section was written by A Abd-alrazaq and RA. The *Discussion* and *Conclusions* sections were written by A Abd-alrazaq and RD. All authors critically revised the manuscript for important intellectual content, approved the manuscript for publication, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy) checklist.
[DOC File , 66 KB - jmir_v26i1e58187_app1.doc]

Multimedia Appendix 2
Search strategy.
[DOCX File , 37 KB - jmir_v26i1e58187_app2.docx]

Multimedia Appendix 3
Data extraction form.
[DOCX File , 26 KB - jmir_v26i1e58187_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Modified version of Quality Assessment of Studies of Diagnostic Accuracy-Revised.
[DOCX File , 27 KB - jmir_v26i1e58187_app4.docx]

Multimedia Appendix 5
Characteristics of each included study.
[DOCX File , 57 KB - jmir_v26i1e58187_app5.docx]

Multimedia Appendix 6
Features of wearable devices.
[DOCX File , 59 KB - jmir_v26i1e58187_app6.docx]

Multimedia Appendix 7
Features of artificial intelligence algorithms.
[DOCX File , 76 KB - jmir_v26i1e58187_app7.docx]

Multimedia Appendix 8
Reviewers' judgments about each "risk of bias" and applicability domain for each included study.
[DOCX File , 72 KB - jmir_v26i1e58187_app8.docx]

References

1. Salari N, Hosseinian-Far A, Mohammadi M, Ghasemi H, Khazaie H, Daneshkhah A, et al. Detection of sleep apnea using machine learning algorithms based on ECG signals: a comprehensive systematic review. Expert Syst Appl 2022 Jan;187:115950. [doi: 10.1016/j.eswa.2021.115950]
2. Singh N, Talwekar RH. "Comparison of machine learning and deep learning classifier to detect sleep apnea using single-channel ECG and HRV: a systematic literature review". J Phys Conf Ser 2022 May 01;2273(1):012015. [doi: 10.1088/1742-6596/2273/1/012015]
3. Ferreira-Santos DA, Amorim P, Silva Martins T, Monteiro-Soares M, Pereira Rodrigues P. Enabling early obstructive sleep apnea diagnosis with machine learning: systematic review. J Med Internet Res 2022 Sep 30;24(9):e39452 [FREE Full text] [doi: 10.2196/39452] [Medline: 36178720]
4. Sateia MJ. International classification of sleep disorders-third edition: highlights and modifications. Chest 2014 Nov;146(5):1387-1394. [doi: 10.1378/chest.14-0970] [Medline: 25367475]
5. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MS, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019 Aug;7(8):687-698. [doi: 10.1016/s2213-2600(19)30198-5]
6. Senaratna CV, Perret JL, Lodge CJ, Lowe AJ, Campbell BE, Matheson MC, et al. Prevalence of obstructive sleep apnea in the general population: a systematic review. Sleep Med Rev 2017 Aug;34:70-81. [doi: 10.1016/j.smrv.2016.07.002] [Medline: 27568340]
7. Berg S. What doctors wish patients knew about sleep apnea. American Medical Association. URL: https://www.ama-assn.org/delivering-care/public-health/what-doctors-wish-patients-knew-about-sleep-apnea [accessed 2024-04-29]
8. Young T, Evans L, Finn L, Palta M. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep 1997 Sep;20(9):705-706. [doi: 10.1093/sleep/20.9.705] [Medline: 9406321]
9. Braley TJ, Dunietz GL, Chervin RD, Lisabeth LD, Skolarus LE, Burke JF. Recognition and diagnosis of obstructive sleep apnea in older Americans. J Am Geriatr Soc 2018 Jul 09;66(7):1296-1302 [FREE Full text] [doi: 10.1111/jgs.15372] [Medline: 29744855]
10. Kapur V, Strohl KP, Redline S, Iber C, O'Connor G, Nieto J. Underdiagnosis of sleep apnea syndrome in U.S. communities. Sleep Breath 2002;06(2):49-54. [doi: 10.1055/s-2002-32318]
11. Lu MK, Tan H, Tsai I, Huang L, Liao X, Lin S. Sleep apnea is associated with an increased risk of mood disorders: a population-based cohort study. Sleep Breath 2017 May 5;21(2):243-253. [doi: 10.1007/s11325-016-1389-x] [Medline: 27495797]
12. Quan SF, Budhiraja R, Batool-Anwar S, Gottlieb D, Eichling P, Patel S, et al. Lack of impact of mild obstructive sleep apnea on sleepiness, mood and quality of life. Southwest J Pulm Crit Care 2014 Jul 25;9(1):44-56 [FREE Full text] [doi: 10.13175/swjpcc082-14] [Medline: 25232509]
13. Lang CJ, Appleton SL, Vakulin A, McEvoy RD, Vincent AD, Wittert GA, et al. Associations of undiagnosed obstructive sleep apnea and excessive daytime sleepiness with depression: an Australian population study. J Clin Sleep Med 2017 Apr 15;13(4):575-582 [FREE Full text] [doi: 10.5664/jcsm.6546] [Medline: 28095971]

XSL•FO
RenderX

14. Marin JM, Agusti A, Villar I, Forner M, Nieto D, Carrizo SJ, et al. Association between treated and untreated obstructive sleep apnea and risk of hypertension. JAMA 2012 May 23;307(20):2169-2176 [FREE Full text] [doi: 10.1001/jama.2012.3418] [Medline: 22618924]
15. Redline S. Obstructive sleep apnea–hypopnea and incident stroke: the sleep heart health study. Am J Respir Crit Care Med 2010 Nov 15;182(10):1332-1333. [doi: 10.1164/ajrccm.182.10.1332b]
16. Marchi NA, Solelhac G, Berger M, Haba-Rubio J, Gosselin N, Vollenweider P, et al. Obstructive sleep apnoea and 5-year cognitive decline in the elderly. Eur Respir J 2023 Apr 16;61(4):2201621 [FREE Full text] [doi: 10.1183/13993003.01621-2022] [Medline: 36796834]
17. Yaffe K, Laffan AM, Harrison SL, Redline S, Spira AP, Ensrud KE, et al. Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women. JAMA 2011 Aug 10;306(6):613-619 [FREE Full text] [doi: 10.1001/jama.2011.1115] [Medline: 21828324]
18. Gottlieb DJ, Ellenbogen JM, Bianchi MT, Czeisler CA. Sleep deficiency and motor vehicle crash risk in the general population: a prospective cohort study. BMC Med 2018 Mar 20;16(1):44 [FREE Full text] [doi: 10.1186/s12916-018-1025-7] [Medline: 29554902]
19. Kales SN, Straubel MG. Obstructive sleep apnea in North American commercial drivers. Ind Health 2014;52(1):13-24 [FREE Full text] [doi: 10.2486/indhealth.2013-0206] [Medline: 24317450]
20. Young T, Finn L, Peppard PE, Szklo-Coxe M, Austin D, Nieto FJ, et al. Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort. Sleep 2008 Aug;31(8):1071-1078 [FREE Full text] [Medline: 18714778]
21. Punjabi NM, Caffo BS, Goodwin JL, Gottlieb DJ, Newman AB, O'Connor GT, et al. Sleep-disordered breathing and mortality: a prospective cohort study. PLoS Med 2009 Aug 18;6(8):e1000132 [FREE Full text] [doi: 10.1371/journal.pmed.1000132] [Medline: 19688045]
22. Abdel-Basset M, Ding W, Abdel-Fatah L. The fusion of internet of intelligent things (IoIT) in remote diagnosis of obstructive sleep apnea: a survey and a new model. Inf Fusion 2020 Sep;61:84-100. [doi: 10.1016/j.inffus.2020.03.010]
23. Aiyer I, Shaik L, Sheta A, Surani S. Review of application of machine learning as a screening tool for diagnosis of obstructive sleep apnea. Medicina (Kaunas) 2022 Nov 01;58(11):1574 [FREE Full text] [doi: 10.3390/medicina58111574] [Medline: 36363530]
24. Ramachandran A, Karuppiah A. A survey on recent advances in machine learning based sleep apnea detection systems. Healthcare (Basel) 2021 Jul 20;9(7):914 [FREE Full text] [doi: 10.3390/healthcare9070914] [Medline: 34356293]
25. Ankitha V, Manimegalai P, Jose PS, Raji P. Literature review on sleep APNEA analysis by machine learning algorithms using ECG signals. J Phys Conf Ser 2021 Jun 01;1937(1):012054. [doi: 10.1088/1742-6596/1937/1/012054]
26. Mendonca F, Mostafa SS, Ravelo-Garcia AG, Morgado-Dias F, Penzel T. A review of obstructive sleep apnea detection approaches. IEEE J Biomed Health Inform 2019 Mar;23(2):825-837. [doi: 10.1109/jbhi.2018.2823265]
27. Pei B, Xia M, Jiang H. Artificial intelligence in screening for obstructive sleep apnoea syndrome (OSAS): a narrative review. J Med Artif Intell 2023 Feb;6:1. [doi: 10.21037/jmai-22-79]
28. Tran NT, Tran HN, Mai AT. A wearable device for at-home obstructive sleep apnea assessment: state-of-the-art and research challenges. Front Neurol 2023 Feb 7;14:1123227 [FREE Full text] [doi: 10.3389/fneur.2023.1123227] [Medline: 36824418]
29. Duarte MA, Pereira-Rodrigues P, Ferreira-Santos D. The role of novel digital clinical tools in the screening or diagnosis of obstructive sleep apnea: systematic review. J Med Internet Res 2023 Jul 26;25:e47735 [FREE Full text] [doi: 10.2196/47735] [Medline: 37494079]
30. McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
31. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Searching for and selecting studies. In: Higgins PT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2019:67.
32. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
33. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51. [doi: 10.7326/m18-1376]
34. Ebert D, Harrer M, Cuijpers P, Furukawa T. Doing Meta-Analysis With R: A Hands-On Guide. Boca Raton, FL: CRC Press; 2012.
35. Borenstein MH, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta‐Analysis. Hoboken, NJ: John Wiley & Sons; 2009.
36. Freeman MF, Tukey JW. Transformations related to the angular and the square root. Ann Math Stat 1950 Dec;21(4):607-611. [doi: 10.1214/aoms/1177729756]
37. Schwarzer G. Meta: an R package for meta-analysis. R News. 2007. URL: https://cran.r-project.org/web/packages/meta/meta.pdf [accessed 2024-04-29]

XSL•FO
RenderX

38. Assink M, Wibbelink CJ. Fitting three-level meta-analytic models in R: a step-by-step tutorial. Quant Method Psychol 2016 Oct 01;12(3):154-174. [doi: 10.20982/tqmp.12.3.p154]
39. Higgins JP, Thompson SG, Deeks JJ, Altman DS. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
40. Benedetti D, Olcese U, Bruno S, Barsotti M, Maestri Tassoni M, Bonanni E, et al. Obstructive sleep apnoea syndrome screening through wrist-worn smartbands: a machine-learning approach. Nat Sci Sleep 2022 May;Volume 14:941-956. [doi: 10.2147/nss.s352335]
41. Chang HC, Wu H, Huang P, Ma H, Lo Y, Huang Y. Portable sleep apnea syndrome screening and event detection using long short-term memory recurrent neural network. Sensors (Basel) 2020 Oct 25;20(21):6067 [FREE Full text] [doi: 10.3390/s20216067] [Medline: 33113849]
42. Chen M, Wu S, Chen T, Wang C, Liu G. Information-based similarity of ordinal pattern sequences as a novel descriptor in obstructive sleep apnea screening based on wearable photoplethysmography bracelets. Biosensors (Basel) 2022 Nov 28;12(12):1089 [FREE Full text] [doi: 10.3390/bios12121089] [Medline: 36551056]
43. Chen X, Xiao Y, Tang Y, Fernandez-Mendoza J, Cao G. ApneaDetector: detecting sleep apnea with smartwatches. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Jun 24;5(2):1-22. [doi: 10.1145/3463514]
44. Fallmann S, Chen L. Detecting chronic diseases from sleep-wake behaviour and clinical features. In: Proceedings of the 5th International Conference on Systems and Informatics. 2018 Presented at: ICSAI '18; November 10-12, 2018; Nanjing, China p. 1076-1084 URL: https://ieeexplore.ieee.org/document/8599388 [doi: 10.1109/icsai.2018.8599388]
45. Fedorin I, Slyusarenko K, Nastenko M. Respiratory events screening using consumer smartwatches. In: Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers. 2020 Presented at: UbiComp/ISWC '20; September 12-17, 2020; Virtual Event p. 25-28 URL: https://dl.acm.org/doi/10.1145/3410530.3414399 [doi: 10.1145/3410530.3414399]
46. Ganglberger W, Bucklin AA, Tesh RA, Da Silva Cardoso M, Sun H, Leone MJ, et al. Sleep apnea and respiratory anomaly detection from a wearable band and oxygen saturation. Sleep Breath 2022 Sep 18;26(3):1033-1044 [FREE Full text] [doi: 10.1007/s11325-021-02465-2] [Medline: 34409545]
47. Gu W, Leung L, Kwok KC, Wu I, Folz RJ, Chiang AA. Belun Ring platform: a novel home sleep apnea testing system for assessment of obstructive sleep apnea. J Clin Sleep Med 2020 Sep 15;16(9):1611-1617 [FREE Full text] [doi: 10.5664/jcsm.8592] [Medline: 32464087]
48. Hafezi M, Montazeri N, Zhu K, Alshaer H, Yadollahi A, Taati B. Sleep apnea severity estimation from respiratory related movements using deep learning. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2019 Presented at: EMBC '19; July 23-27, 2019; Berlin, Germany p. 1601-1604 URL: https://ieeexplore.ieee.org/abstract/document/8857524 [doi: 10.7573/dic.2019-9-4]
49. Hafezi M, Montazeri N, Saha S, Zhu K, Gavrilovic B, Yadollahi A, et al. Sleep apnea severity estimation from tracheal movements using a deep learning model. IEEE Access 2020;8:22641-22649. [doi: 10.1109/access.2020.2969227]
50. Hung PD. Central sleep apnea detection using an accelerometer. In: Proceedings of the 1st International Conference on Control and Computer Vision. 2018 Presented at: ICCCV '18; June 15-18, 2018; Singapore, Singapore p. 106-111 URL: https://dl.acm.org/doi/abs/10.1145/3232651.3232660 [doi: 10.1145/3232651.3232660]
51. Jeon Y, Heo K, Kang SJ. Real-time sleep apnea diagnosis method using wearable device without external sensors. In: Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops. 2020 Presented at: PerCom Workshops '20; March 23-27, 2020; Austin, TX p. 1-5 URL: https://ieeexplore.ieee.org/document/9156119 [doi: 10.1109/percomworkshops48775.2020.9156119]
52. Ji X, Rao Z, Zhang W, Liu C, Wang Z, Zhang S, et al. Airline point-of-care system on seat belt for hybrid physiological signal monitoring. Micromachines (Basel) 2022 Nov 01;13(11):1880 [FREE Full text] [doi: 10.3390/mi13111880] [Medline: 36363901]
53. Kristiansen S, Nikolaidis K, Plagemann T, Goebel V, Traaen GM, Øverland B, et al. Machine learning for sleep apnea detection with unattended sleep monitoring at home. ACM Trans Comput Healthcare 2021 Feb 09;2(2):1-25. [doi: 10.1145/3433987]
54. Kristiansen S, Nikolaidis K, Plagemann T, Goebel V, Traaen GM, Øverland B, et al. A clinical evaluation of a low-cost strain gauge respiration belt and machine learning to detect sleep apnea. Smart Health 2023 Mar;27:100373. [doi: 10.1016/j.smhl.2023.100373]
55. Kwon S, Kim HS, Kwon K, Kim H, Kim YS, Lee SH, et al. At-home wireless sleep monitoring patches for the clinical assessment of sleep quality and sleep apnea. Sci Adv 2023 May 24;9(21):eadg9671 [FREE Full text] [doi: 10.1126/sciadv.adg9671] [Medline: 37224243]
56. Le TQ, Cheng C, Sangasoongsong A, Wongdhamma W, Bukkapatnam ST. Wireless wearable multisensory suite and real-time prediction of obstructive sleep apnea episodes. IEEE J Transl Eng Health Med 2013;1:2700109 [FREE Full text] [doi: 10.1109/JTEHM.2013.2273354] [Medline: 27170854]
57. McClure K, Erdreich B, Bates JH, McGinnis RS, Masquelin A, Wshah S. Classification and detection of breathing patterns with wearable sensors and deep learning. Sensors (Basel) 2020 Nov 13;20(22):6481 [FREE Full text] [doi: 10.3390/s20226481] [Medline: 33202857]

XSL•FO
RenderX

58. Papini GB, Fonseca P, van Gilst MM, Bergmans JW, Vullings R, Overeem S. Wearable monitoring of sleep-disordered breathing: estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography. Sci Rep 2020 Aug 11;10(1):13512 [FREE Full text] [doi: 10.1038/s41598-020-69935-7] [Medline: 32782313]
59. Petrenko A. Breathmonitor: sleep apnea mobile detector. In: Proceedings of the 2nd International Conference on System Analysis & Intelligent Computing. 2020 Presented at: SAIC '20; October 5-9, 2020; Kyiv, Ukraine p. 1-4 URL: https://ieeexplore.ieee.org/document/9239236 [doi: 10.1109/saic51296.2020.9239236]
60. Rossi M, Sala D, Bovio D, Salito C, Alessandrelli G, Lombardi C, et al. SLEEP-SEE-THROUGH: explainable deep learning for sleep event detection and quantification from wearable somnography. IEEE J Biomed Health Inform 2023 Jul;27(7):3129-3140. [doi: 10.1109/jbhi.2023.3267087]
61. Ryser F, Hanassab S, Lambercy O, Werth E, Gassert R. Respiratory analysis during sleep using a chest-worn accelerometer: a machine learning approach. Biomed Signal Process Control 2022 Sep;78:104014. [doi: 10.1016/j.bspc.2022.104014]
62. Selvaraj N, Narasimhan R. Automated prediction of the apnea-hypopnea index using a wireless patch sensor. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:1897-1900. [doi: 10.1109/EMBC.2014.6943981] [Medline: 25570349]
63. Shen Q, Yang X, Zou L, Wei K, Wang C, Liu G. Multitask residual shrinkage convolutional neural network for sleep apnea detection based on wearable bracelet photoplethysmography. IEEE Internet Things J 2022 Dec 15;9(24):25207-25222. [doi: 10.1109/jiot.2022.3195777]
64. Strumpf Z, Gu W, Tsai C, Chen P, Yeh E, Leung L, et al. Belun Ring (Belun Sleep System BLS-100): deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea. Sleep Health 2023 Aug;9(4):430-440 [FREE Full text] [doi: 10.1016/j.sleh.2023.05.001] [Medline: 37380590]
65. Tsouti V, Kanaris A, Tsoutis K, Chatzandroulis S. Development of an automated system for obstructive sleep apnea treatment based on machine learning and breath effort monitoring. Microelectron Eng 2020 Jul;231:111376. [doi: 10.1016/j.mee.2020.111376]
66. van Steenkiste T, Groenendaal W, Dreesen P, Lee S, Klerkx S, de Francisco R, et al. Portable detection of apnea and hypopnea events using bio-impedance of the chest and deep learning. IEEE J Biomed Health Inform 2020 Sep;24(9):2589-2598. [doi: 10.1109/jbhi.2020.2967872]
67. Wang S, Xuan W, Chen D, Gu Y, Liu F, Chen J, et al. Machine learning assisted wearable wireless device for sleep apnea syndrome diagnosis. Biosensors (Basel) 2023 Apr 17;13(4):483 [FREE Full text] [doi: 10.3390/bios13040483] [Medline: 37185558]
68. Wang Z, Peng C, Li B, Penzel T, Liu R, Zhang Y, et al. Single-lead ECG based multiscale neural network for obstructive sleep apnea detection. Internet Things 2022 Nov;20:100613. [doi: 10.1016/j.iot.2022.100613]
69. Wu HT, Wu J, Huang P, Lin T, Wang T, Huang Y, et al. Phenotype-based and self-learning inter-individual sleep apnea screening with a level IV-like monitoring system. Front Physiol 2018 Jul 2;9:723 [FREE Full text] [doi: 10.3389/fphys.2018.00723] [Medline: 30013479]
70. Wu S, Chen M, Wei K, Liu G. Sleep apnea screening based on Photoplethysmography data from wearable bracelets using an information-based similarity approach. Comput Methods Programs Biomed 2021 Nov;211:106442. [doi: 10.1016/j.cmpb.2021.106442] [Medline: 34624633]
71. Xu Y, Ou Q, Cheng Y, Lao M, Pei G. Comparative study of a wearable intelligent sleep monitor and polysomnography monitor for the diagnosis of obstructive sleep apnea. Sleep Breath 2023 Mar 26;27(1):205-212 [FREE Full text] [doi: 10.1007/s11325-022-02599-x] [Medline: 35347656]
72. Yeh E, Wong E, Tsai C, Gu W, Chen P, Leung L, et al. Detection of obstructive sleep apnea using Belun Sleep platform wearable with neural network-based algorithm and its combined use with STOP-Bang questionnaire. PLoS One 2021 Oct 11;16(10):e0258040 [FREE Full text] [doi: 10.1371/journal.pone.0258040] [Medline: 34634070]
73. Yeo M, Byun H, Lee J, Byun J, Rhee HY, Shin W, et al. Respiratory event detection during sleep using electrocardiogram and respiratory related signals: using polysomnogram and patch-type wearable device data. IEEE J Biomed Health Inform 2022 Feb;26(2):550-560. [doi: 10.1109/jbhi.2021.3098312]
74. Yüzer AH, Sümbül H, Nour M, Polat K. A different sleep apnea classification system with neural network based on the acceleration signals. Appl Acoust 2020 Jun;163:107225. [doi: 10.1016/j.apacoust.2020.107225]
75. Zhang H, Fu B, Su K, Yang Z. Long-term sleep respiratory monitoring by dual-channel flexible wearable system and deep learning-aided analysis. IEEE Trans Instrum Meas 2023;72:1-9. [doi: 10.1109/tim.2023.3289535]
76. Zhou G, Zhou W, Zhang Y, Zeng Z, Zhao W. Automatic monitoring of obstructive sleep apnea based on multi-modal signals by phone and smartwatch. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10340237] [Medline: 38083356]
77. Yeo M, Byun H, Lee J, Byun J, Rhee H, Shin W, et al. Robust method for screening sleep apnea with single-lead ECG using deep residual network: evaluation with open database and patch-type wearable device data. IEEE J Biomed Health Inform 2022 Nov;26(11):5428-5438. [doi: 10.1109/jbhi.2022.3203560]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
CSA: central sleep apnea
HR: heart rate
ML: machine learning
OSA: obstructive sleep apnea
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy
QUADAS-2: Quality Assessment of Studies of Diagnostic Accuracy-Revised

Edited by A Mavragani; submitted 08.03.24; peer-reviewed by BS Ibrahim, K Dhou, A Wani; comments to author 26.04.24; revised version received 07.05.24; accepted 23.07.24; published 10.09.24.

Please cite as:
Abd-alrazaq A, Aslam H, AlSaad R, Alsahli M, Ahmed A, Damseh R, Aziz S, Sheikh J
Detection of Sleep Apnea Using Wearable AI: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e58187
URL: https://www.jmir.org/2024/1/e58187
doi:10.2196/58187
PMID:39255014

©Alaa Abd-alrazaq, Hania Aslam, Rawan AlSaad, Mohammed Alsahli, Arfan Ahmed, Rafat Damseh, Sarah Aziz, Javaid Sheikh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Prompt Engineering Paradigms for Medical Applications: Scoping Review

Jamil Zaghir[1,2*], MSc; Marco Naguib[3*], MSc; Mina Bjelogrlic[1,2], PhD; Aurélie Névéol[3], PhD; Xavier Tannier[4], PhD; Christian Lovis[1,2], MPH, MD

[1]Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland

[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[3]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France

[4]Sorbonne Université, INSERM, Université Sorbonne Paris-Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en eSanté, LIMICS, Paris, France

[*]these authors contributed equally

Corresponding Author:
Jamil Zaghir, MSc
Department of Radiology and Medical Informatics
University of Geneva
Chemin des Mines, 9
Geneva, 1202
Switzerland
Phone: 41 022 379 08 18
Email: Jamil.Zaghir@unige.ch

Abstract

Background: Prompt engineering, focusing on crafting effective prompts to large language models (LLMs), has garnered attention for its capabilities at harnessing the potential of LLMs. This is even more crucial in the medical domain due to its specialized terminology and language technicity. Clinical natural language processing applications must navigate complex language and ensure privacy compliance. Prompt engineering offers a novel approach by designing tailored prompts to guide models in exploiting clinically relevant information from complex medical texts. Despite its promise, the efficacy of prompt engineering in the medical domain remains to be fully explored.

Objective: The aim of the study is to review research efforts and technical approaches in prompt engineering for medical applications as well as provide an overview of opportunities and challenges for clinical practice.

Methods: Databases indexing the fields of medicine, computer science, and medical informatics were queried in order to identify relevant published papers. Since prompt engineering is an emerging field, preprint databases were also considered. Multiple data were extracted, such as the prompt paradigm, the involved LLMs, the languages of the study, the domain of the topic, the baselines, and several learning, design, and architecture strategies specific to prompt engineering. We include studies that apply prompt engineering–based methods to the medical domain, published between 2022 and 2024, and covering multiple prompt paradigms such as prompt learning (PL), prompt tuning (PT), and prompt design (PD).

Results: We included 114 recent prompt engineering studies. Among the 3 prompt paradigms, we have observed that PD is the most prevalent (78 papers). In 12 papers, PD, PL, and PT terms were used interchangeably. While ChatGPT is the most commonly used LLM, we have identified 7 studies using this LLM on a sensitive clinical data set. Chain-of-thought, present in 17 studies, emerges as the most frequent PD technique. While PL and PT papers typically provide a baseline for evaluating prompt-based approaches, 61% (48/78) of the PD studies do not report any nonprompt-related baseline. Finally, we individually examine each of the key prompt engineering–specific information reported across papers and find that many studies neglect to explicitly mention them, posing a challenge for advancing prompt engineering research.

Conclusions: In addition to reporting on trends and the scientific landscape of prompt engineering, we provide reporting guidelines for future studies to help advance research in the medical field. We also disclose tables and figures summarizing medical prompt engineering papers available and hope that future contributions will leverage these existing works to better advance the field.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e60501) doi:10.2196/60501

KEYWORDS

prompt engineering; prompt design; prompt learning; prompt tuning; large language models; LLMs; scoping review; clinical natural language processing; natural language processing; NLP; medical texts; medical application; medical applications; clinical practice; privacy; medicine; computer science; medical informatics

Introduction

In recent years, the development of large language models (LLMs) such as GPT-3 has disrupted the field of natural language processing (NLP). LLMs have demonstrated capabilities in processing and generating human-like text, with applications ranging from text generation and translation to question answering and summarization [1]. However, harnessing the full potential of LLMs requires careful consideration of how input prompts are formulated and optimized [2].

Input prompts denote a set of instructions provided to the LLM to execute a task. Prompt engineering, a term coined to describe the strategic design and optimization of prompts for LLMs, has emerged as a crucial aspect of leveraging these models. By crafting prompts that effectively convey tasks or queries, researchers and practitioners can guide LLMs to improve the accuracy and pertinence of responses. The literature defines prompt engineering in various ways: it can be regarded as a prompt structuring process that enhances the efficiency of an LLM to achieve a specific objective [3] or as the mechanism through which LLMs are programmed by prompts [4]. Prompt engineering encompasses a plethora of techniques, often separated into distinct categories such as output customization and prompt improvement [4]. Existing prompt paradigms are presented in more detail in the Methods section.

In the realm of medical NLP, significant advancements have been made, such as the release of LLMs specialized in medical language and the availability of public medical data sets, including in languages other than English [5]. The unique intricacies of medical language, characterized by its terminological precision, context sensitivity, and domain-specific nuances, demand a dedicated focus and exploration of NLP in health care research. Despite these imperatives, to our knowledge, there is currently no systematic review analyzing prompt engineering applied to the medical domain.

The aim of this scoping review is to shed light on prompt engineering, as it is developed and used in the medical field, by systematically analyzing the literature in the field. Specifically, we examine the definitions, methodologies, techniques, and outcomes of prompt engineering across various NLP tasks. Methodological strengths, weaknesses, and limitations of the current wave of experimentation are discussed. Finally, we provide guidelines for comprehensive reporting of prompt engineering–related studies to improve clarity and facilitate further research in the field. We aspire to furnish insights that will inform both researchers and users about the pivotal role of prompt engineering in optimizing the efficacy of LLMs. By gaining a thorough understanding of the current landscape of prompt engineering research, we can pinpoint areas warranting further investigation and development, thereby propelling the field of medical NLP forward.

Methods

Study Design

Our scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines for scoping reviews (available in Multimedia Appendix 1). In this review, we use terminology to denote emerging technical concepts that lack consensus definitions. We propose the following definitions based on previous use in the literature:

- LLM: Object that models language and can be used to generate text by receiving large-scale language modeling pretraining (Luccioni and Rogers [6] define an arbitrary threshold at 1 billion tokens of training data). An LLM can be adapted to downstream tasks through transfer learning approaches such as fine-tuning or prompt-based techniques. Following the study of Thirunavukarasu et al [7] of models for the medical field, we include Bidirectional Encoder Representations From Transformers (BERT)–based and GPT-based models in this definition, although Zhao et al [8] place BERT models in a separate category.
- Fine-tuning: Approach in which the weights of the pretrained LLM are retrained on new samples. The additional data can be labeled and designed to adapt the LLM to a new downstream task.
- Prompt design (PD) [1,2]: Manually building a prompt (named manual prompt or hard prompt), tailored to guide the LLM toward resolving the task by simply predicting the most probable continuity of the prompt. The prompt is usually a set of task-specific instructions, occasionally featuring a few demonstrations of the task.
- Prompt learning (PL) [3]: Manually building a prompt and passing it to an LLM, trained via the masked language modeling (MLM) objective, to predict masked tokens. The prompt often features masked tokens, over which the LLM makes predictions. Those are then projected as predictions for a new downstream task. This approach is also referred to as prompt-based learning.
- Prompt tuning (PT) [9]: Refers to the LLM prompting where part or all the prompt is a trainable vectorial representation (known as continuous prompt or soft prompt) that is optimized with respect to the annotated instances.

Figure 1 illustrates the 4 approaches described above.

XSL•FO
RenderX

Figure 1. Illustration of traditional fine-tuning and the 3 prompt-based paradigms (the fire logo represents trainable parameters, and the flake logo illustrates frozen parameters). LLM: large language model.

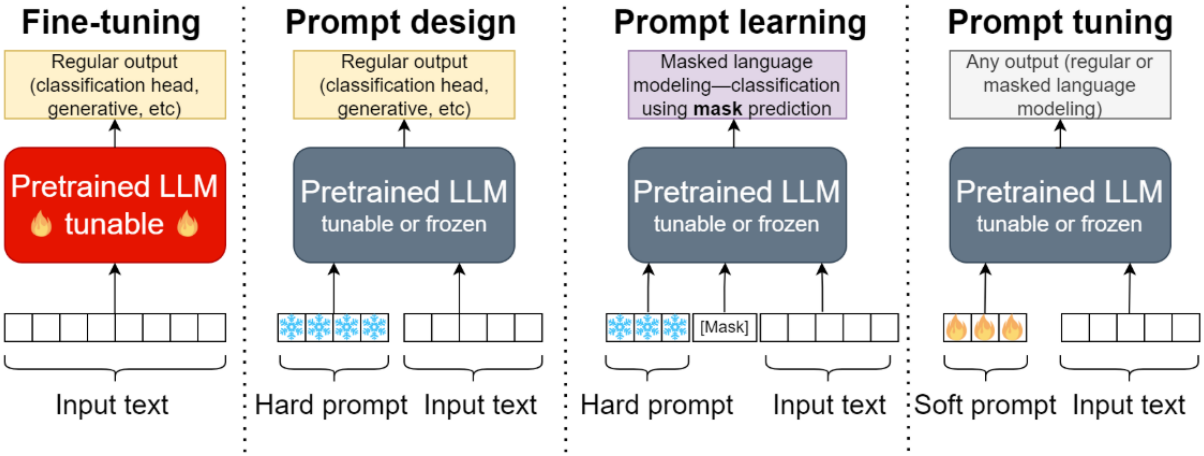

Inclusion and Exclusion Criteria

Studies were included if they met the following criteria: focus on prompt engineering, involvement of at least 1 LLM, relevance to the medical field (biomedical or clinical), pertaining to text-based generation (excluding vision-related prompts), and not focusing on prompting for academic writing purposes. Furthermore, as most of the first studies about prompt engineering emerged in 2022 [2], we added the following constraint: the publication date should be later than 2021.

Screening Process

The initial set of papers retrieved from the searches underwent screening based on titles, abstracts, and keywords. The search strategy is described in Multimedia Appendix 2. Screening was performed by 2 reviewers (JZ and MN), working in a double-blind process. Interannotator agreement was calculated, with conflicts resolved through discussion.

Data Synthesis

We extracted information on prompt paradigms (PD, PL, and PT), involved LLMs, data sets used, studied language, domain (biomedical or clinical), medical subfield (if any), mentioned prompt engineering techniques, computational complexity, baselines, relative performances, and key findings. Additionally, we extracted journal information and noted instances of PD or PL or PT terminology misuse. Details are available in Multimedia Appendix 3. Finally, we compile a list of recommendations based on the positive or negative trends we identify from the selected papers.

Results

Screening Results

The systematic search across sources yielded 398 papers. Following the removal of duplicates, 251 papers underwent screening based on title, abstract, and keywords, leading to the exclusion of 94 studies. During this first screening step, 33 conflicts were identified and resolved among the annotators, resulting in an interannotator agreement of 86.8% (n=218). Subsequently, 157 studies remained, and full-text copies were retrieved and thoroughly screened. This process culminated in the inclusion of a total of 114 papers in this scoping review. The detailed process of study selection is shown in Figure 2. Among the selected papers, 13 are from clinical venues, 33 are from medical informatics sources, 31 are from computer science publications, and 4 are from other sources. Notably, 33 of them are preprints.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the review process.

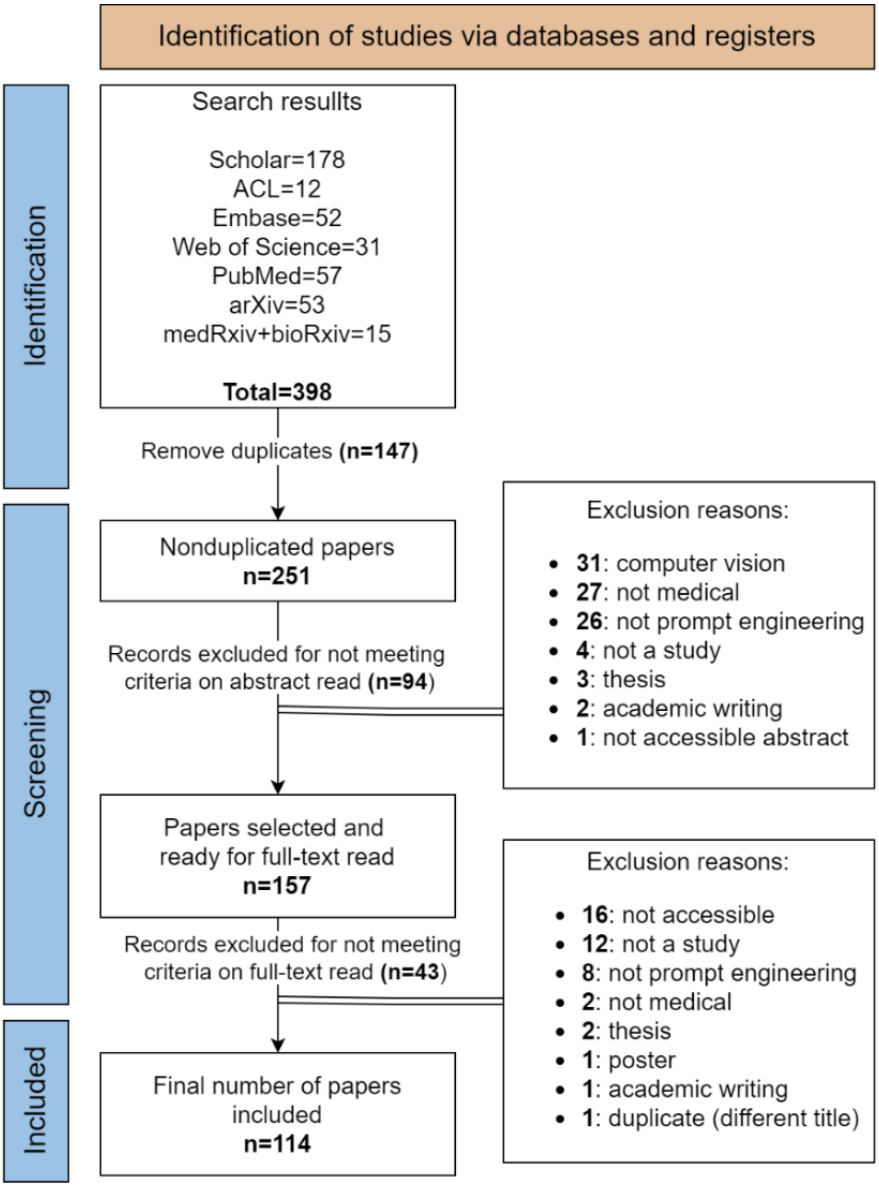

Prompt Paradigms and Medical Subfields

Table 1 depicts the number of papers identified within each prompt paradigm along with their associated medical subfields. Some papers may simultaneously involve several (up to 2 in this review) prompt paradigms. Notably, PD emerged as the predominant category, with a total of 78 papers. These papers spanned across various medical fields, with a greater emphasis on clinical (including specialties) rather than biomedical disciplines. The screening yields 29 PL papers and 19 PT papers, with both paradigms maintaining a balanced distribution between biomedical and clinical domains. However, it is noteworthy that unlike PL and PT, PD encompassed a much broader spectrum of clinical specialties, with a particular interest in psychiatry.

XSL•FO
RenderX

Table 1. Paper distribution by prompt category and medical subfield, with corresponding references.

Prompt paradigm and domain of the topic	References
Prompt design (78)	
Biomedical (17)	[10-26]
Medical licensing examination (12)	[27-38]
Clinical (general) (15)	[39-53]
Psychiatry (10)	[28,54-62]
Oncology (5)	[63-67]
Cardiology (4)	[68-71]
Ophthalmology (3)	[72-74]
Neurology (3)	[69,75,76]
Orthopedics (2)	[77,78]
Clinical trials (2)	[79,80]
Intensive care (2)	[69,81]
Geriatrics (2)	[75,76]
Radiology (2)	[31,82]
Nuclear medicine (1)	[29]
Hepatology (1)	[83]
Endocrinology (1)	[84]
Plastic surgery (1)	[85]
Gastroenterology (1)	[32]
Genetics (1)	[86]
Nursing (1)	[87]
Prompt learning (29)	
Biomedical (13)	[88-100]
Clinical (general) (15)	[41,47,101-113]
Psychiatry (1)	[114]
Prompt tuning (19)	
Biomedical (9)	[16,20,26,90,91,95,98,115,116]
Clinical (general) (6)	[101,105,110,117-119]
Oncology (2)	[120,121]
Psychiatry (1)	[122]
Medical insurance (1)	[123]

Terminology Use

In our review, the consistency of terminology use around prompt engineering was investigated, particularly concerning its 3 paradigms: PD, PL, and PT. Across the papers, we meticulously tracked instances where the terminology was applied differently to the definitions used in the literature and described in the introduction. Notably, PL was used to refer to PD 4 times [12,13,67,86] and PT once [119], while PT was used 5 times to describe PL [88,96,97,99,114] and twice for PD [23,43]. Terminology inconsistencies were identified in only 12 studies. Consequently, while there remains some degree of inconsistency, a significant majority of 102 papers adhered to the definitions identified as commonly used terminology.

Language of Study

Considering the latest developments in NLP research encompassing languages beyond English [124], reporting the language of study is crucial. Several papers do not explicitly state the language of study. In some cases, the language can be inferred from prompt illustrations or examples. In the least informative cases, only the data set of the study is disclosed, indirectly hinting at the language.

Table 2 illustrates the language distribution among the selected papers, noting whether languages are explicitly mentioned, implicitly inferred from prompt illustrations, or simply not stated but implied from the used data set. The language used in 2 papers [60,68] remains unknown.

XSL•FO
RenderX

Table 2. Frequency distribution of papers across various languages. The table also depicts the frequency distribution across venues for papers studying English (N=114).

Language and type of venue	Stated[a], n (%)	Inferred[b], n (%)	Not stated[c], n (%)	Total, n (%)
English				
All	37 (32.5)	48 (42.1)	11 (9.6)	96 (84.2)
Medical informatics	16 (14)	9 (7.9)	2 (1.8)	27 (23.7)
Computer science	8 (7)	18 (15.8)	1 (0.9)	27 (23.7)
Preprint	9 (7.9)	12 (10.5)	5 (4.4)	26 (22.8)
Clinical	1 (0.9)	8 (7)	3 (2.6)	12 (10.5)
Other	3 (2.6)	1 (0.9)	0 (0)	4 (3.5)
Chinese				
All	18 (15.8)	0 (0)	0 (0)	18 (15.8)
French				
All	3 (2.6)	0 (0)	0 (0)	3 (2.6)
Dutch				
All	3 (2.6)	0 (0)	0 (0)	3 (2.6)
Japanese				
All	2 (1.8)	0 (0)	0 (0)	2 (1.8)
Portuguese				
All	2 (1.8)	0 (0)	0 (0)	2 (1.8)
Italian				
All	2 (1.8)	0 (0)	0 (0)	2 (1.8)
Spanish				
All	2 (1.8)	0 (0)	0 (0)	2 (1.8)
Korean				
All	0 (0)	0 (0)	1 (0.9)	1 (0.9)
Basque				
All	1 (0.9)	0 (0)	0 (0)	1 (0.9)
German				
All	1 (0.9)	0 (0)	0 (0)	1 (0.9)
Swedish				
All	1 (0.9)	0 (0)	0 (0)	1 (0.9)
Polish				
All	1 (0.9)	0 (0)	0 (0)	1 (0.9)
Vietnamese				
All	1 (0.9)	0 (0)	0 (0)	1 (0.9)
Unknown				
All	0 (0)	0 (0)	2 (1.8)	2 (1.8)

[a]Stated in the paper.

[b]Inferred from prompt figures and examples.

[c]Inferred from the data set.

Notably, English dominates with 84.2% (n=96) of the selected papers, followed by Chinese at 15.7% (n=18). Then, the other languages are relatively rare, often appearing in studies featuring multiple languages. It is worth mentioning that languages besides English are usually explicitly stated, with the exception of a paper studying Korean [63]. In total, the language had to be inferred from prompt figures and examples in 48 papers, all in English.

XSL•FO
RenderX

Choice of LLMs

Given the diverse array of LLMs available, spanning general or medical, open-source or proprietary, and monolingual or multilingual models, alongside various architectural configurations (encoder, decoder, or both), our study investigates LLM selection across prompt paradigms.

Figure 3 outlines prevalent LLMs categorized by prompt paradigms, though it is not exhaustive and only includes commonly encountered architectures. For example, while encoder-decoder models are absent in PT in Figure 3, there are a few instances where they are used [95,110].

ChatGPT's popularity in PD is unsurprising, given its accessibility. Models from Google, PaLM, and Bard (subsequently rebranded Gemini), all falling under closed models, are also prominent. Among open-source instruct-based LLMs, fewer are used, notably those based on LLaMA-2 with 7 occurrences.

In PL, encoder models, those following the BERT architecture, dominate, covering both general and specialized variants. There are occasional uses of decoder models like GPT-2 in PL-based tasks [103,105]. PT involves all model types, with a preference toward encoders. Further details on the models used are available in Multimedia Appendix 3.

Figure 3. Involved large language models in the prompt engineering studies, covering all prompt paradigms. The number of studies that fit in a node is shown in parentheses. BERT: Bidirectional Encoder Representations From Transformers; RoBERTa: Robustly Optimized BERT Pre-training Approach; T5: Text-to-Text Transfer Transformer.

XSL•FO
RenderX

Topic Domain and NLP Task Trends

Figure 4 [16,20,26,41,47,88-123] illustrates the target tasks used in the PL and PT papers. PL-focused papers predominantly address classification-based tasks such as text classification, named entity recognition, and relation extraction, with text classification being particularly prominent. This aligns with the nature of PL, which centers around an MLM objective. Among other tasks, a study based on text generation [111] makes use of PL to predict masked tokens from partial patient records, aiming to generate synthetic electronic health records. Conversely, PT papers tend to exhibit a slightly broader range of tasks.

Figure 5 [10-87] presents the same analysis for PD-based papers. Unlike PL and PT, a prominent trend observed is that several studies focus on real-world board examinations. Notably, these studies predominantly center around tasks involving answering multiple-choice questions (MCQs). It is worth noting that although MCQs might be cast as a classification task, in practice, it is cast as a generation task using causal LLMs. It is interesting to note that none of the selected PD papers propose the task of entity linking, despite the clear opportunity of leveraging LLMs' in-context learning ability for medical entity linking.

Figure 4. Overview of selected prompt learning and prompt tuning papers, showcasing natural language processing tasks alongside their topic domain (it includes tasks, such as text simplification, where none of the selected papers specifically focused on these tasks). Numbers within square brackets are reference citations.

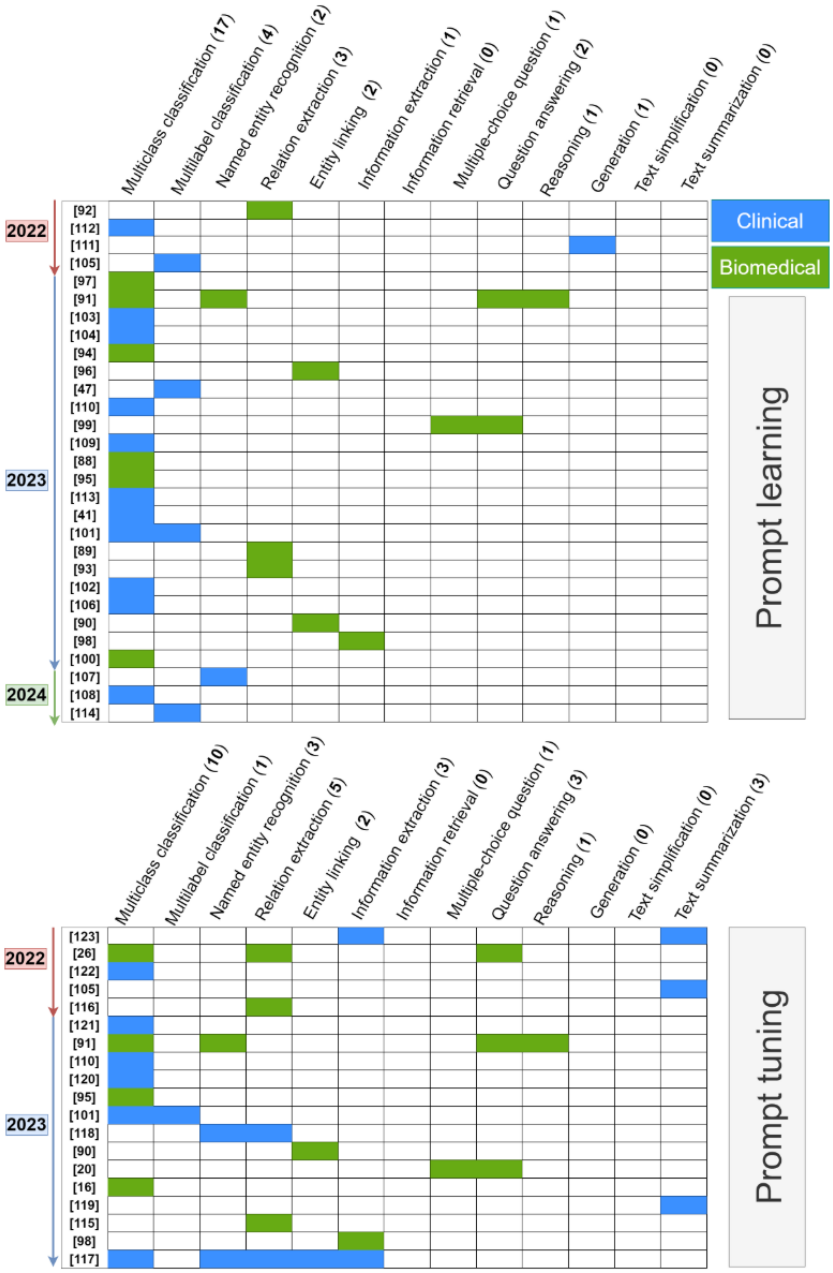

XSL•FO
RenderX

Figure 5. Overview of selected prompt design papers, showcasing natural language processing tasks alongside their topic domain. Numbers within square brackets are reference citations.

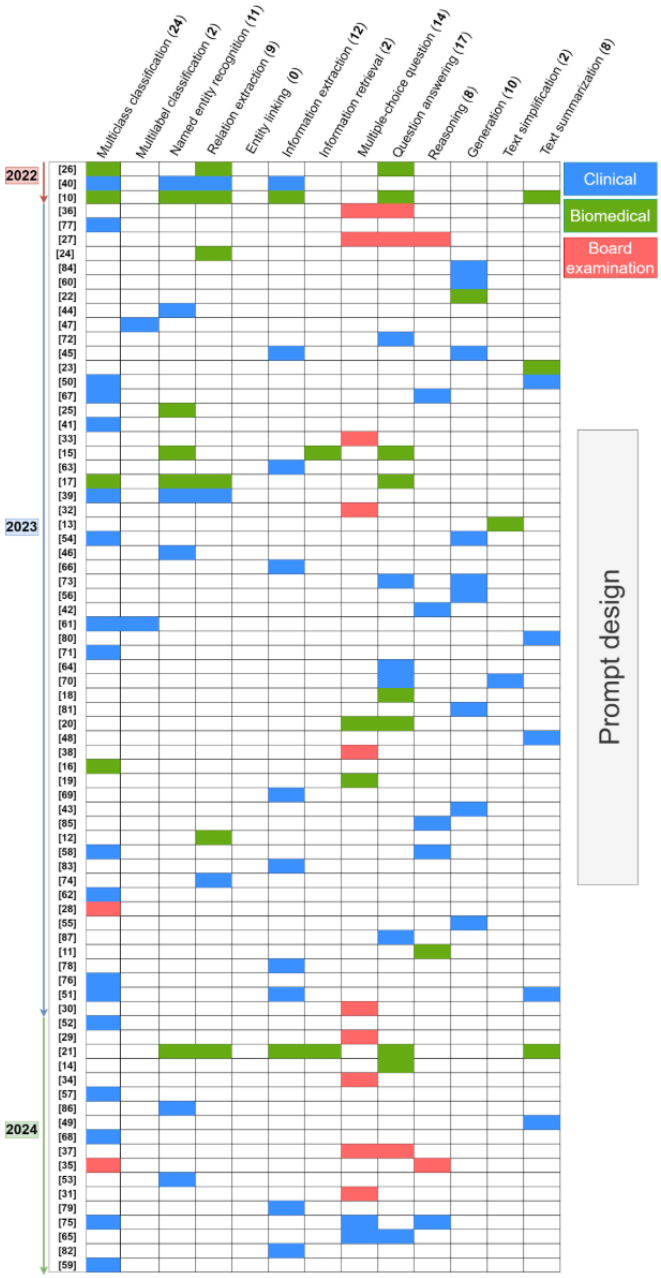

Prompt Engineering Techniques

We extensively investigated the used prompt techniques: among PD papers, 49 studies used zero-shot prompting, 23 used few-shot prompting, and 10 used one-shot prompting. Few shot tends to outperform in MCQs, but its advantage over zero shot is inconsistent in other NLP tasks. We propose a comprehensive summary of the existing techniques in Table 3.

As shown in Table 3, chain-of-thought (CoT) prompting [2] stands as the most common technique, followed by the persona pattern. In medical MCQs, various attempts with CoT can lead to different reasoning pathways and answers. Hence, to improve accuracy, 2 studies [19,20] used self-consistency, a method involving using multiple CoT prompts and selecting the most frequently occurring answer through voting.

Flipped interaction was used for simulation tasks, such as doctor-patient engagement [60] or to provide clinical training to medical students [81]. Emotion enhancement was applied in mental health contexts [58,60], allowing the LLM to produce emotional statements.

More innovative prompt engineering techniques include k-nearest neighbor few-shot prompting [19] and pseudoclassification prompting [78]. The former uses the k-nearest neighbor algorithm to select the k-closest examples in a large annotated data set based on the input before using them in the prompt, and the latter presents to the LLMs all possible labels, asking the model to respond with a binary output for each provided label. Despite its potential, tree-of-thoughts pattern use was limited, with only 1 instance found among the papers [77].

XSL•FO
RenderX

Table 3. Most recurrent prompt techniques found, with the corresponding description, template, and references.

Prompt techniques	Description	Prompt template examples	Count papers	References
Chain-of-thought (CoT)	Asking the large language model (LLM) to provide the reasoning before answering.	• Basic CoT: "<Prompt>. Think step by step." • Another example of CoT: "Solve this math problem. E.g.: You have 3 apples and buy 2 more, how many apples do you have? Solution: Start with 3 apples. Buy 2 more apples. Total apples is 3 + 2 = 5. New problem: You have 5 oranges and give away 2, how many oranges do you have left?"	17	[11,19,20,27,29,32,33,35,39,51,58,67,75,77,82,83,85]
Persona (role-defining)	Assigning the LLM a particular role to accomplish a task related to that role.	• "Act as X (e.g. Act as a Physician, Act as a Psychiatrist, etc)."	10	[32,49,55,56,59-61,82,84,85]
Ensemble prompting	Using multiple independent prompts to answer the same question. The final output is decided by majority vote.	• "Prompt1, Output1, Prompt2, Output2, […], Promptk, Outputk" Final output: Vote	4	[19,20,39,52]
Scene-defining	Simulating a scene related to the addressed task.	• "you are in a hospital, in front of a patient ..."	3	[18,49,61]
Prompt-chaining	Separating a task into multiple subtasks, each resolved with a prompt.	• "Prompt1->Output1, Output1+Prompt2 ->Output2, [...] Outputk-1+Promptk-> Outputk"	3	[37,80,84]
Flipped interaction	Making the LLM take the lead (eg, asking questions) and the user interacting with it passively.	• "I would like you to ask me questions to achieve X. You should ask questions until <condition/goal> is met."	2	[60,81]
Emotion enhancement	Making the LLM more or less expressing human-like emotions.	• "You can have emotional fluctuations during the conversation."	2	[58,60]
Prompt refinement	Using the LLM to refine the prompt such as translating the prompt or rephrasing it.	• "Please translate in English / rephrase this prompt: <P>."	2	[37,48]
Retrieval-augmented generation	Combining an information retrieval component with a generative LLM. Snippets extracted from documents are fed into the system along with the input prompt to generate an enriched output.	• "<List of relevant Snippets> <Input Prompt>"	2	[18,54]
Self-consistency (CoT ensembling)	Ensemble prompting each prompt using CoT. Ideal if a problem has many possible reasoning paths.	• "CoT_Pr1, Output1, CoT_Pr2, Output2, ..., CoT_Prk, Outputk" Final output: Vote	2	[19,20]

Emerging Trends

Figure 6 illustrates a chronological polar pie chart of selected papers and their citation connections, identifying five highly cited papers: (1) Agrawal et al [40] demonstrate GPT-3's clinical task performance, especially in named entity recognition and relation extraction through thorough PD. (2) Kung et al [36] evaluate ChatGPT's (GPT-3.5) ability for the United States Medical Licensing Examination, shortly after the public release of ChatGPT. (3) Singhal et al [20] introduce MultiMedQA and HealthSearchQA benchmarks. The paper also presents instruction PT for domain alignment, a novel paradigm that entails learning a soft prompt prior to the LLM general instruction, which is usually written as a hard prompt. Using this approach on FlanPaLM led to the development of Med-PaLM, improving question answering over FlanPaLM. (4) Nori et al [27] evaluate GPT-4 on the United States Medical Licensing Examination and MultiMedQA, surpassing previous state-of-the-art results, including GPT-3.5 and Med-PaLM. (5) Luo et al [26] release BioGPT, a fine-tuned variant of GPT-2 for biomedical tasks, achieving state-of-the-art results on 6 biomedical NLP tasks with suffix-based PT.

XSL•FO
RenderX

Figure 6. A chronological chart showing the selected papers across the 3 prompt-based paradigms. Papers are classified by different colors according to the venues in which they were published. Different shapes illustrate whether the LLM is fine-tuned, frozen, or both. Solid or striped color indicates whether authors used a nonprompt baseline (including humans) for evaluation. Arrows connecting 2 papers denote direct citations. The nodes in the border of PD, PL, or PT are studies proposing the 2 involved prompt engineering paradigms. Numbers within square brackets are reference citations. LLM: large language model; PD: prompt design; PL: prompt learning; PT: prompt tuning.

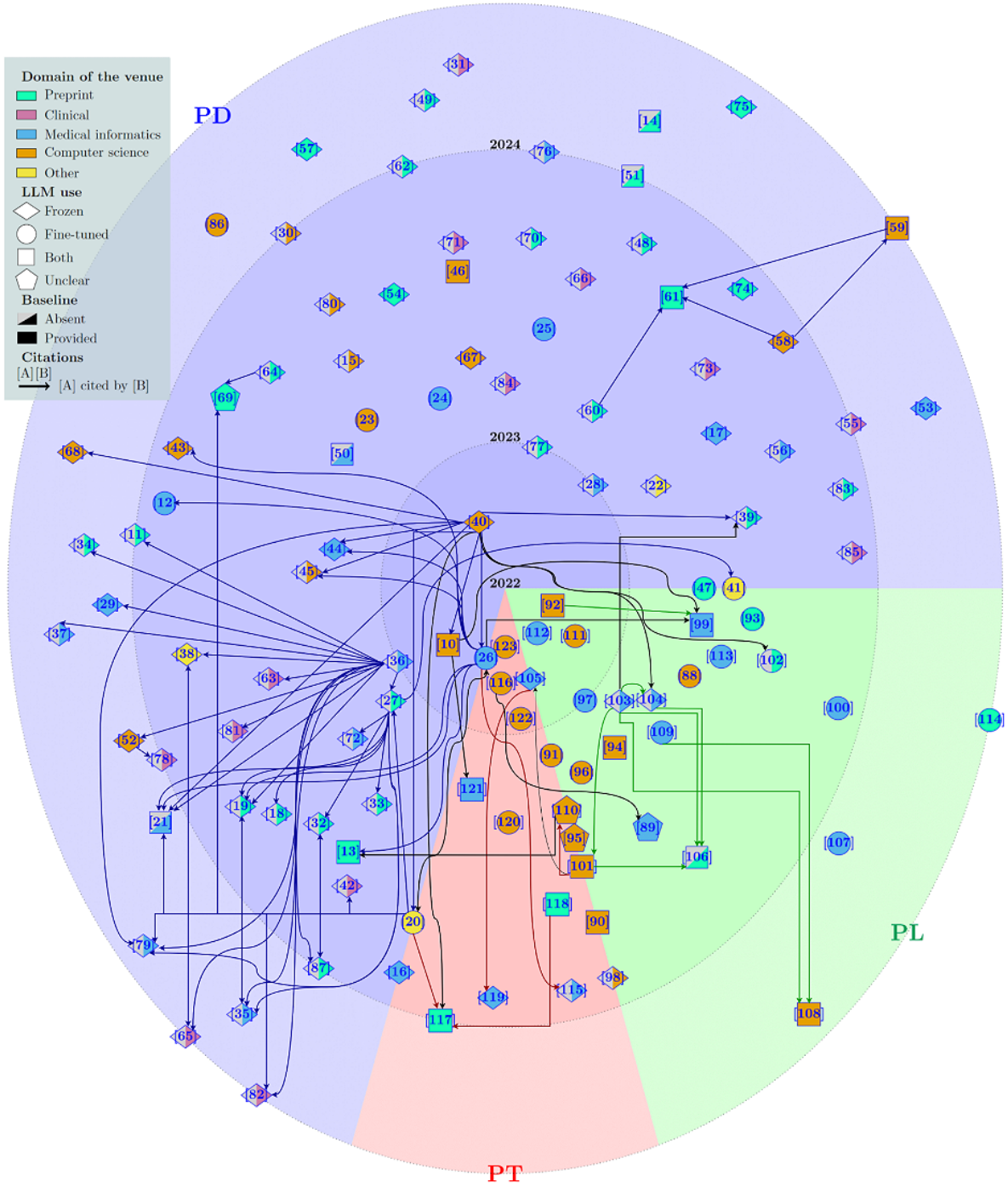

Trends in PD

As shown in Figure 6, the PD paradigm presents multiple trends: all papers disseminated in clinical-based venues, and 27 of 33 (82%) of the encountered preprints adhere to this paradigm. Furthermore, we observed a significant focus on work involving frozen LLMs within the PD domain. This trend is likely due to the frequent use of ChatGPT in 74 instances, as depicted in Figure 3, despite OpenAI offering fine-tuning capabilities for the model. It is worth mentioning that 46 of 78 (59%) PD papers do not include any baseline, including human comparison. This gap will be further explored in a subsequent section.

Trends in PL and PT

Among PL and PT papers, computer science and medical informatics are the most prevalent venues. Although PL has drawn attention to the idea of adapting the MLM objective to downstream tasks without needing to further update the LLM weights, many studies still opt to fine-tune their LLMs, with a nonnegligible amount of them evaluating in few-shot settings

XSL•FO
RenderX

[89,92,93,112]. Unlike PD, PL and PT usually include a baseline, with it often being a traditional fine-tuning version of the evaluated model [92,93,95] to compare it against novel prompt-based paradigms. These studies came to a common conclusion, being that PL is a promising alternative to traditional fine-tuning in few-shot scenarios.

There are 2 ways for conducting PL: one involves filling in the blanks within a text, known as cloze prompts, while the other consists in predicting masked tokens at the end of the sequence, referred to as prefix prompts. A distinct advantage of the latter approach is its compatibility with autoregressive models, as they exclusively predict the appended masks. Among the 29 PL papers, 21 (72%) of them propose cloze prompts, while 15 (52%) use prefix prompting. The involved NLP tasks are well-distributed across these 2 prompt patterns. Another crucial component of PL is the verbalizer. As PL revolves around predicting masked tokens, classification-based tasks require mapping manually selected relevant tokens to each class (manual verbalizer). Alternatively, some studies propose a soft verbalizer, akin to soft prompts, which automatically determines the most relevant token embedding for each label through training. Of the 29 PL papers selected, 16 (55%) studies explicitly mention the use of a manual verbalizer, while 2 explored both verbalizers to assess performance [101,110]. Only 1 exclusively used a soft verbalizer [89]. Another study does not use any verbalizer, as it focuses on generating synthetic data by filling the blanks [111]. Notably, 8 (28%) studies did not report any mention regarding the verbalizer methodology.

Hard prompts, which are related to PD and PL, involve manually crafted prompts. Regarding PT, optimal prompts are attainable through soft prompting (ie, prompts that are trained on a training data set), yet, determining the appropriate soft prompt length remains obscure. In total, 5 of 19 (26%) PT studies tried various soft prompt lengths and reported their corresponding performances [26,105,118,119,122]. While there is no definitive optimal prompt length, a trend emerges: optimal soft prompt length typically exceeds 10 tokens. Surprisingly, 8 (42%) papers omit reporting the soft prompt length. Regarding the placement of soft prompts in relation to the input and the mask, consensus is lacking. A total of 5 (26%) papers prepend the soft prompt at the input's outset, while 4 (21%) append it as a suffix. One paper uses both strategies in a single prompt template [95]. Some innovative methods involve inserting a single soft prompt for each entity that needs to be identified in entity-linking tasks or using token-wise soft prompts, where each token in the textual input is accompanied by a distinct soft prompt. The position of soft prompts remains unreported in 5 (26%) studies. Finally, according to the 6 (32%) studies that used mixed prompts [90,91,95,101,105,110] (a combination of hard and soft prompts), it has consistently been reported that mixed prompts lead to a better performance than hard prompts alone.

Baseline Comparison

Only 62 of the screened papers reported comparisons to established baselines. These include traditional deep learning approaches (eg, fine-tuning approach), classical machine learning algorithms (eg, logistic regression), naive systems (eg, majority class), or human annotation. The remaining papers solely explored prompt-related solutions, without including baseline comparisons. Tables 4-6 traces the presence of a nonprompt baseline among different prompt categories (Table 4), papers sources (Table 5), and NLP tasks addressed (Table 6).

Table 4. Baseline reports among prompt categories (N=114)[a].

Prompt category	No baseline, n (%)	Higher, n (%)	Similar, n (%)	Lower, n (%)	Total, n (%)
Prompt design	48 (42.1)	13 (11.4)	4 (3.5)	13 (11.4)	78 (68.4)
Prompt learning	5 (4.4)	19 (16.7)	3 (2.6)	2 (1.8)	29 (25.4)
Prompt tuning	3 (2.6)	11 (9.6)	2 (1.8)	3 (2.6)	19 (16.7)

[a]Higher or lower indicates that the performance of the proposed prompt-based approach is higher or lower than the baseline.

Table 5. Baseline reports among venues (N=114)[a].

Type of venue	No baseline, n (%)	Higher, n (%)	Similar, n (%)	Lower, n (%)	Total, n (%)
Medical informatics	13 (11.4)	16 (14)	2 (1.8)	2 (1.8)	33 (28.9)
Computer science	7 (6.1)	12 (10.5)	3 (2.6)	9 (7.9)	31 (27.2)
Preprint	21 (18.4)	6 (5.3)	1 (0.9)	5 (4.4)	33 (28.9)
Clinical	13 (11.4)	0 (0)	0 (0)	0 (0)	13 (11.4)
Other	1 (0.9)	2 (1.8)	0 (0)	1 (0.9)	4 (3.5)

[a]Higher or lower indicates that the performance of the proposed prompt-based approach is higher or lower than the baseline.

XSL•FO
RenderX

Table 6. Baseline reports among addressed NLP[a] tasks (N=114)[b].

NLP task	No baseline, n (%)	Higher, n (%)	Similar, n (%)	Lower, n (%)	Total, n (%)
Text classification	13 (11.4)	18 (15.8)	4 (3.5)	11 (9.6)	46 (40.4)
Question answering	13 (11.4)	3 (2.6)	1 (0.9)	2 (1.8)	19 (16.7)
Relation extraction	3 (2.6)	10 (8.8)	0 (0)	3 (2.6)	16 (14)
Information extraction	10 (8.8)	3 (2.6)	0 (0)	2 (1.8)	15 (13.2)
Multiple-choice question	10 (8.8)	3 (2.6)	1 (0.9)	1 (0.9)	15 (13.2)
Named entity recognition	4 (3.5)	5 (4.4)	1 (0.9)	5 (4.4)	15 (13.2)
Text summarization	7 (6.1)	3 (2.6)	0 (0)	1 (0.9)	11 (9.6)
Reasoning	5 (4.4)	3 (2.6)	0 (0)	1 (0.9)	9 (7.9)
Generation	5 (4.4)	2 (1.8)	0 (0)	1 (0.9)	8 (7)
Entity linking	0 (0)	3 (2.6)	0 (0)	0 (0)	3 (2.6)
Coreference resolution	1 (0.9)	1 (0.9)	0 (0)	1 (0.9)	3 (2.6)
Decision support	2 (1.8)	0 (0)	0 (0)	1 (0.9)	3 (2.6)
Conversational	3 (2.6)	0 (0)	0 (0)	0 (0)	3 (2.6)
Text simplification	1 (0.9)	0 (0)	0 (0)	1 (0.9)	2 (1.8)

[a]NLP: natural language processing.

[b]Higher or lower indicates that the performance of the proposed prompt-based approach is higher or lower than the baseline.

Nonprompt-related baselines are often featured in studies focused on PL and PT but not PD. Additionally, PL and PT have a tendency to perform better than their respective reported baselines, PD tends to report less conclusive results. More specifically, among the 22 papers using either PL or PT with an identical fine-tuned model as a baseline, 17 indicate superior performance with the prompt-based approach, 3 observed comparable performance, and 2 studies noted inferior performance.

Significantly, papers from computer science venues tend to include more state-of-the-art baselines than those from medical informatics and clinical venues. Specifically, all 13 papers reviewed from clinical venues did not use any nonprompt baselines. Furthermore, there appears to be no consistent link between the type of NLP tasks and the omission of baselines, indicating that the decision to include baselines is more influenced by the evaluation methodology than by feasibility.

Prompt Optimization

Numerous studies in the literature highlight the few-shot learning capabilities of LLMs, often referred to as "few-shot prompting," wherein they demonstrate proficiency in executing tasks with minimal demonstrations provided, typically through text prompts. However, it is crucial to acknowledge that the annotation cost associated with such frameworks might extend beyond the few annotated demonstrations within the prompt. Many studies claiming to explore few-shot or zero-shot learning through prompt engineering rely on extensive annotated validation data sets to refine PD and formulation. This is, for example, the case in the paper that popularized the term "few-shot learning" [1]. Among the 45 analyzed papers concentrating on few-shot or zero-shot learning, 5 explicitly detail the optimization of prompt formulation using extensive validation data sets. Conversely, 18 of these papers either do not engage in prompt optimization or test various prompts and document all results. Notably, 22 papers present results using only 1 prompt choice, without clarifying whether this choice was made thanks to additional validation data sets.

Discussion

Summary of the Findings

This scoping review aimed to map the current landscape of medical prompt engineering, identifying key themes, gaps, and trends within the existing literature. The primary findings of this study reveal a greater prevalence of PD over PL and PT, with ChatGPT dominating the PD domain. Additionally, many studies omit nonprompt-based baselines, do not specify the language of study, or exhibit a lack of consensus in PL (prefix vs cloze prompt) and PT settings (soft prompt lengths and positions). English is notably dominant as the language of study. These findings suggest that while the field is emerging, there is a pressing need for improved research practices.

Costs, Infrastructure, and LLMs in Clinical Settings

Prompt engineering techniques enable competitive performance in scenarios with limited or no resources as well as in environments with low-cost computing infrastructure. As hospital data and infrastructure are often found in this scenario, these approaches hold great promise in the clinical field. Figure 6 shows the absence of PL- and PT-related works in clinical journals. This trend may stem from the widespread accessibility of ChatGPT, favoring PD-focused investigations. Despite efforts like OpenPrompt [125] to facilitate PL and PT works, the programming barrier likely deters clinical practitioners. Surprisingly, 7 papers use ChatGPT with sensitive clinical data. Despite the recent availability of ChatGPT Enterprise in GPT-4 for secure data handling, it is apparent that most of these studies

XSL•FO
RenderX

have not used this feature since they used GPT-3.5. Limited use of local LLMs, especially LLaMA-based, suggests a need for their increased adoption in future clinical PD studies. The lack of local LLMs may be due to clinicians' limited computational infrastructure.

Prompt Engineering Techniques Effectiveness in Medical Research

In documented prompt engineering techniques, the effectiveness of few-shot prompting compared to zero shot varies by task and scenario. However, CoT shows superior reasoning performance, compelling LLMs to present reasoning pathways and consistently outperforming zero-shot and few-shot methods across PD studies. Its ensemble-based variant, self-consistency, consistently outperforms CoT. Despite the persona pattern's frequent use, there is a lack of ablation studies on its impact on medical task performance, with only 1 paper reporting negligible improvement [61]. Prompt engineering is an emerging field of study that still needs to prove its efficacy. However, almost half of the papers focused only on prompt engineering and failed to report any nonprompt-related baseline performance, despite the availability of such baselines for the addressed NLP tasks. On the whole, the results are far from being systematically in favor of LLM-based methods, greatly attenuating the impression of a technological breakthrough that is generally commented on. Selecting a baseline remains a necessary step toward understanding the actual impact of prompt engineering.

Bender Rule

Regarding the languages, while Table 2 shows the dominance of English in medical literature, many papers studying English fail to explicitly mention the language of study. This oversight is more prevalent in computer science and clinical venues, whereas medical informatics exhibits a more favorable trend, as validated by a chi-square test yielding a P value of .02 (Table S1 in Multimedia Appendix 2). Notably, languages such as Chinese are consistently mentioned across the 18 selected papers. However, the Bender rule, namely "always name the language(s) you are working on," seems to be well respected for languages other than English. This finding has already been documented for NLP research in general [126].

Fine-Tuning Versus Prompt-Based Approaches

While traditional LLM fine-tuning remains a viable method for various NLP tasks, PL and PT are competitive alternatives to fine-tuning, particularly in resource-constrained and low computational scenarios. PL, leveraging predefined prompts to guide model behavior, offers an efficient approach in low-to-no resource environments. Conversely, PT emerges as a viable solution in low computational scenarios, as it requires substantially fewer trainable parameters compared to traditional fine-tuning approaches. Since both prompt-based approaches do not require the LLM to be further trained, they are less prone to catastrophic forgetting [127].

Recommendations for Future Medical Prompt–Based Studies

For future research in prompt engineering, we propose several recommendations aimed at improving research quality, reporting, and reproducibility. From this review, we identified several trends such as the computational advantages or the lack of evaluations on baselines with a lack of ablation studies to evaluate the performance of the prompting strategies. Some studies do not clearly mention the prompt engineering choices they made. For instance, in PL, choices range from using cloze to prefix prompting and from using manual to soft verbalizer. Similarly, PT is characterized by configurations of soft prompts, such as the length and the positions. To clarify these distinctions and enhance methodological transparency and reproducibility in future research, we have developed reporting guidelines available in Textbox 1. Adhering to these reporting guidelines will contribute to advancing prompt engineering methodologies and their practical applications in the medical field.

Textbox 1. Detailed reporting guidelines for future prompt engineering studies.

General reporting recommendations

- For sensitive data, local large language models (LLMs) should be preferred to the ones that use an application programming interface or a web service.
- The language of the study used should be explicitly stated.
- The mention of whether the LLM undergoes fine-tuning should be made explicit.
- The prompt optimization process and results should be documented to ensure transparency, whether it is through different tested manual prompts or through a validation data set.
- The terms "few-shot," "one-shot," and "zero-shot" should not be used in settings where the prompts have been optimized on annotated examples.
- Experiments should include baseline comparisons or at least mention existing results, particularly when data sets originate from previous medical challenges or benchmarks.

Specific to prompt learning and prompt tuning

- Concepts (such as prompt learning and prompt tuning) should be defined and used consistently with the consensus.
- In prompt learning experiments, the verbalizer used (soft and hard) should be explicitly specified, or a clear justification should be provided if the verbalizer is omitted. Additionally, whether the prompt template follows the cloze or the prefix format should be mentioned.
- In prompt tuning experiments, authors should provide details on soft prompt positions, length, and any variations tested, such as incorporating hard or mixed prompts, as part of the ablation study.

XSL•FO
RenderX

Limitations

A limitation was the large number of papers retrieved during the initial search, which was addressed by limiting the search scope to titles, abstracts, and keywords. Furthermore, since some studies may perform prompt engineering techniques without mentioning any of the 4 prompt-related expressions used in the queries, they might be missed by our searches.

Conclusions

Medical prompt engineering is an emerging field with significant potential for enhancing clinical applications, particularly in resource-constrained environments. Despite the promising capabilities demonstrated, there is a pressing need for standardized research practices and comprehensive reporting to ensure methodological transparency and reproducibility. Consistent evaluation against nonprompt-based baselines, prompt optimization documentation, and prompt settings reporting will be crucial for advancing the field. We hope that a better adherence to the recommended guidelines, in Textbox 1, will improve our understanding of prompt engineering and enhance the capabilities of LLMs in health care.

Acknowledgments

JZ is financed by the NCCR Evolving Language, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant #51NF40_180888).

Authors' Contributions

JZ and MN performed the screening and data extraction of the papers and synthesized the findings. AN and XT supervised MN. MB and CL supervised JZ. JZ and MN wrote the manuscript with support from MB, AN, XT, and CL. All authors contributed to the analysis of the results. CL conceived the original idea.

Conflicts of Interest

CL is the editor-in-chief of *JMIR Medical Informatics*. All other authors have no conflict of interest to declare.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 515 KB - jmir_v26i1e60501_app1.pdf]

Multimedia Appendix 2
Search strategy and statistical analysis.
[DOCX File , 20 KB - jmir_v26i1e60501_app2.docx]

Multimedia Appendix 3
Reading notes and details of the reviewed papers.
[XLSX File (Microsoft Excel File), 52 KB - jmir_v26i1e60501_app3.xlsx]

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. 2020 Presented at: Advances in Neural Information Processing Systems; December 6, 2020; Virtual p. 1877-1901 URL: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
2. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. 2022 Presented at: Advances in Neural Information Processing Systems; November 28, 2022; New Orleans p. 22199-22213 URL: https://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html
3. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Jan 16;55(9):1-35. [doi: 10.1145/3560815]
4. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. ArXiv Preprint posted online on February 21, 2023 [FREE Full text] [doi: 10.48550/arXiv.2302.11382]
5. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
6. Luccioni AS, Rogers A. Mind your language (model): fact-checking LLMs and their role in NLP research and practice. ArXiv Preprint posted online on June 1, 2024. [doi: 10.48550/arXiv.2308.07120]
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. ArXiv Preprint posted online on November 24, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.18223]

XSL•FO
RenderX

9. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. 2021 Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; January 10, 2021; Online and Punta Cana, Dominican Republic p. 3045-3059. [doi: 10.18653/v1/2021.emnlp-main.243]
10. Fries J, Weber L, Seelam N, Altay G, Datta D, Garda S, et al. BigBIO: a framework for data-centric biomedical natural language processing. 2022 Presented at: Advances in Neural Information Processing Systems; November 28, 2022; New Orleans p. 25792-25806 URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/a583d2197eafc4afdd41f5b8765555c5-Abstract-Datasets_and_Benchmarks.html
11. Weisenthal SJ. ChatGPT and post-test probability. ArXiv Preprint posted online on July 20, 2024 [FREE Full text] [doi: 10.48550/arXiv.2311.12188]
12. Li L, Ning W. ProBioRE: a framework for biomedical causal relation extraction based on dual-head prompt and prototypical network. 2023 Presented at: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 5, 2023; Istanbul, Turkiye p. 2071-2074 URL: https://tinyurl.com/3n45uwdb
13. Li Z, Belkadi S, Micheletti N, Han L, Shardlow M, Nenadic G. Large language models and control mechanisms improve text readability of biomedical abstracts. ArXiv Preprint posted online on March 16, 2024 [FREE Full text] [doi: 10.48550/arXiv.2309.13202]
14. Li Q, Yang X, Wang H, Liu L, Wang Q, Wang J, et al. From beginner to expert: modeling medical knowledge into general LLMs. ArXiv Preprint posted online on January 7, 2024 [FREE Full text] [doi: 10.48550/arXiv.2312.01040]
15. Ateia S, Kruschwitz U. Is ChatGPT a biomedical expert? 2023 Presented at: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023); September 18-21, 2023; Thessaloniki, Greece p. 73-90 URL: https://ceur-ws.org/Vol-3497/paper-006.pdf
16. Belyaeva A, Cosentino J, Hormozdiari F, Eswaran K, Shetty S, Corrado G, et al. Multimodal LLMs for health grounded in individual-specific data. 2023 Presented at: Machine Learning for Multimodal Healthcare Data; July 29, 2023; Honolulu, Hawaii, United States p. 86-102. [doi: 10.1007/978-3-031-47679-2_7]
17. Chen Q, Sun H, Liu H, Jiang Y, Ran T, Jin X, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics 2023;39(9):btad557 [FREE Full text] [doi: 10.1093/bioinformatics/btad557] [Medline: 37682111]
18. Mollá D. Large language models and prompt engineering for biomedical query focused multi-document summarisation. ArXiv Preprint posted online on November 9, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.05169]
19. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. ArXiv Preprint posted online on November 28, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.16452]
20. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
21. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023;25(1):bbad493 [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
22. Lim S, Schmälzle R. Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Front Commun 2023;8:1129082. [doi: 10.3389/fcomm.2023.1129082]
23. Wu YH, Lin YJ, Kao HY. IKM_Lab at BioLaySumm Task 1: longformer-based prompt tuning for biomedical lay summary generation. 2023 Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; July 13, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.bionlp-1.64]
24. Zhang W, Chen C, Wang J, Liu J, Ruan T. A co-adaptive duality-aware framework for biomedical relation extraction. Bioinformatics 2023;39(5):btad301 [FREE Full text] [doi: 10.1093/bioinformatics/btad301] [Medline: 37220895]
25. Chen P, Wang J, Lin H, Zhao D, Yang Z. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. Bioinformatics 2023;39(8):btad496 [FREE Full text] [doi: 10.1093/bioinformatics/btad496] [Medline: 37549065]
26. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
27. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. ArXiv Preprint posted online on April 12, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.13375]
28. Heinz MV, Bhattacharya S, Trudeau B, Quist R, Song SH, Lee CM, et al. Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. Digit Health 2023;9:20552076231170499 [FREE Full text] [doi: 10.1177/20552076231170499] [Medline: 37101589]
29. Ting Y, Hsieh T, Wang Y, Kuo Y, Chen Y, Chan P, et al. Performance of ChatGPT incorporated chain-of-thought method in bilingual nuclear medicine physician board examinations. Digit Health 2024;10:20552076231224074 [FREE Full text] [doi: 10.1177/20552076231224074] [Medline: 38188855]
30. Casola S, Labruna T, Lavelli A, Magnini B. Testing ChatGPT for stability and reasoning: a case study using Italian medical specialty tests. 2023 Presented at: Proceedings of the 9th Italian Conference on Computational Linguistics; November 30-Decemeber 2, 2023; Venice, Italy URL: https://ceur-ws.org/Vol-3596/paper13.pdf

XSL•FO
RenderX

31. Roemer G, Li A, Mahmood U, Dauer L, Bellamy M. Artificial intelligence model GPT4 narrowly fails simulated radiological protection exam. J Radiol Prot 2024;44(1):013502. [doi: 10.1088/1361-6498/ad1fdf] [Medline: 38232401]
32. Ali S, Shahab O, Al Shabeeb R, Ladak F, Yang JO, Nadkarni G, et al. General purpose large language models match human performance on gastroenterology board exam self-assessments. MedRxi Preprint posted online on September 25, 2023 [FREE Full text] [doi: 10.1101/2023.09.21.23295918]
33. Patel D, Raut G, Zimlichman E, Cheetirala S, Nadkarni G, Glicksberg BS, et al. The limits of prompt engineering in medical problem-solving: a comparative analysis with ChatGPT on calculation based USMLE medical questions. MedRxiv Preprint posted online on August 9, 2023 [FREE Full text] [doi: 10.1101/2023.08.06.23293710]
34. Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, Bard, ChatGPT-3.5, and humans in clinical chemistry multiple-choice questions. MedRxiv Preprint posted online on January 9, 2024 [FREE Full text] [doi: 10.1101/2024.01.08.24300995]
35. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024;7(1):20 [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
36. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
37. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. PLOS Digit Health 2024;3(1):e0000433 [FREE Full text] [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
38. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
39. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. ArXiv Preprint posted online on September 14, 2023 [FREE Full text] [doi: 10.2196/55318]
40. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; May 25, 2022; Abu Dhabi p. 1998-2022. [doi: 10.18653/v1/2022.emnlp-main.130]
41. Dong B, Wang Z, Li Z, Duan Z, Xu J, Pan T, et al. Toward a stable and low-resource PLM-based medical diagnostic system via prompt tuning and MoE structure. Sci Rep 2023;13(1):12595 [FREE Full text] [doi: 10.1038/s41598-023-39543-2] [Medline: 37537202]
42. Gutierrez KLT, Viacrusis PML. Bridging the gap or widening the divide: a call for capacity-building in artificial intelligence for healthcare in the Philippines. JMUST 2023;7(2):1325-1334. [doi: 10.35460/2546-1621.2023-0081]
43. Islam KS, Nipu AS, Madiraju P, Deshpande P. Autocompletion of chief complaints in the electronic health records using large language models. 2023 Presented at: 2023 IEEE International Conference on Big Data (BigData); December 15-18, 2023; Sorrento, Italy p. 4912-4921 URL: https://tinyurl.com/4ajdyddt [doi: 10.1109/bigdata59044.2023.10386778]
44. Meoni S, Ryffel T, De La Clergerie É. Annotate French clinical data using large language model predictions. 2023 Presented at: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); June 26-29, 2023; Houston, TX, United States p. 550-557 URL: https://tinyurl.com/yy2b9fe8 [doi: 10.1109/ichi57859.2023.00099]
45. Meoni S, De la Clergerie E, Ryffel T. Large language models as instructors: a study on multilingual clinical entity extraction. 2023 Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; July 2023; Toronto, Canada p. 178-190 URL: https://aclanthology.org/2023.bionlp-1.15/ [doi: 10.18653/v1/2023.bionlp-1.15]
46. Wang X, Yang Q. LingX at ROCLING 2023 MultiNER-health task: intelligent capture of Chinese medical named entities by LLMs. 2023 Presented at: Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023); October 20-21, 2023; Taipei City, Taiwan URL: https://aclanthology.org/2023.rocling-1.44.pdf
47. Yang Y, Li X, Wang H, Guan Y, Jiang J. Modeling clinical thinking based on knowledge hypergraph attention network and prompt learning for disease prediction. SSRN Preprint posted online on June 30, 2023 [FREE Full text] [doi: 10.2139/ssrn.4496800]
48. Yao Z, Jaafar A, Wang B, Zhu Y, Yang Z, Yu H. Do physicians know how to prompt? The need for automatic prompt optimization help in clinical note generation. ArXiv Preprint posted online on July 5, 2024 [FREE Full text] [doi: 10.48550/arXiv.2311.09684]
49. van Zandvoort D, Wiersema L, Huibers T, van Dulmen S, Brinkkemper S. Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting. ArXiv Preprint posted online on January 19, 2024 [FREE Full text] [doi: 10.5220/0012422600003657]
50. Zhang B, Mishra R, Teodoro D. DS4DH at MEDIQA-Chat 2023: leveraging SVM and GPT-3 prompt engineering for medical dialogue classification and summarization. 2023 Presented at: Proceedings of the 5th Clinical Natural Language Processing Workshop; June 12, 2023; Toronto, Canada p. 536-545. [doi: 10.18653/v1/2023.clinicalnlp-1.57]

XSL•FO
RenderX

51. Zhu W, Wang X, Chen M, Tang B. Overview of the PromptCBLUE Shared Task in CHIP2023. ArXiv Preprint posted online on December 29, 2023 [FREE Full text] [doi: 10.48550/arXiv.2312.17522]
52. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. Expert Syst Appl 2024;235:121186. [doi: 10.1016/j.eswa.2023.121186]
53. Lee Y, Chen C, Chen C, Lee C, Chen P, Wu C, et al. Unlocking the secrets behind advanced artificial intelligence language models in deidentifying Chinese-English mixed clinical text: development and validation study. J Med Internet Res 2024;26:e48443 [FREE Full text] [doi: 10.2196/48443] [Medline: 38271060]
54. Bhaumik R, Srivastava V, Jalali A, Ghosh S, Chandrasekaran R. Mindwatch: a smart cloud-based AI solution for suicide ideation detection leveraging large language models. MedRxiv Preprint posted online on September 26, 2023 [FREE Full text] [doi: 10.1101/2023.09.25.23296062]
55. Heston TF. Safety of large language models in addressing depression. Cureus 2023:15 [FREE Full text] [doi: 10.7759/cureus.50729]
56. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell 2023;6:20 [FREE Full text] [doi: 10.21037/jmai-23-71]
57. Santos WR, Paraboni I. Prompt-based mental health screening from social media text. ArXiv Preprint posted online on May 11, 2024 [FREE Full text] [doi: 10.5753/brasnam.2024.1879]
58. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; January 16, 2023; Singapore p. 6056-6077. [doi: 10.18653/v1/2023.emnlp-main.370]
59. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol 2024;8(1):1-32. [doi: 10.1145/3643540]
60. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. ArXiv Preprint posted online on May 23, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.13614]
61. Qi H, Zhao Q, Li J, Song C, Zhai W, Dan L, et al. Supervised learning and large language model benchmarks on mental health datasets: cognitive distortions and suicidal risks in Chinese social media. ResearchSquare Preprint posted online on November 02, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3523508/v1]
62. Sambath V. Advancements of artificial intelligence in mental health applications?: A comparative analysis of ChatGPT 3.5 and ChatGPT 4. ResearchGate Preprint posted online on December, 2023 [FREE Full text] [doi: 10.13140/RG.2.2.28713.36961]
63. Choi HS, Song JY, Shin KH, Chang JH, Jang B. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023;41(3):209-216 [FREE Full text] [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
64. Lee DT, Vaid A, Menon KM, Freeman R, Matteson DS, Marin MP, et al. Development of a privacy preserving large language model for automated data extraction from thyroid cancer pathology reports. MedRxiv Preprint posted online on November 8, 2023 [FREE Full text] [doi: 10.1101/2023.11.08.23298252]
65. Dennstädt F, Hastings J, Putora PM, Vu E, Fischer GF, Süveg K, et al. Exploring capabilities of large language models such as ChatGPT in radiation oncology. Adv Radiat Oncol 2024;9(3):101400 [FREE Full text] [doi: 10.1016/j.adro.2023.101400] [Medline: 38304112]
66. Zhu S, Gilbert M, Ghanem AI, Siddiqui F, Thind K. Feasibility of using zero-shot learning in transformer-based natural language processing algorithm for key information extraction from head and neck tumor board notes. Int J Radiat Oncol Biol Phys 2023;117(2):e500. [doi: 10.1016/j.ijrobp.2023.06.1743]
67. Zhao X, Zhang M, Ma M, Su C, Wang M, Qiao X, et al. HW-TSC at SemEval-2023 task 7: exploring the natural language inference capabilities of ChatGPT and pre-trained language model for clinical trial. 2023 Presented at: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023); July 10, 2023; Toronto, Canada p. 1603-1608. [doi: 10.18653/v1/2023.semeval-1.221]
68. Nazary F, Deldjoo Y, Di Noia T. ChatGPT-HealthPrompt. Harnessing the power of XAI in prompt-based healthcare decision support using ChatGPT. 2023 Presented at: Artificial Intelligence. ECAI 2023 International Workshops; September 30-October 4, 2023; Kraków, Poland p. 382-397. [doi: 10.1007/978-3-031-50396-2_22]
69. Wang B, Lai J, Cao H, Jin F, Tang M, Yao C, et al. Enhancing real-world data extraction in clinical research: evaluating the impact of the implementation of large language models in hospital setting. ResearchSquare Preprint posted online on November 29, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3644810/v2]
70. Mishra V, Sarraju A, Kalwani NM, Dexter JP. Evaluation of prompts to simplify cardiovascular disease information using a large language model. MedRxiv Preprint posted online on November 9, 2023 [FREE Full text] [doi: 10.1101/2023.11.08.23298225]
71. Feng R, Brennan KA, Azizi Z, Goyal J, Pedron M, Chang HJ, et al. Optimizing ChatGPT to detect VT recurrence from complex medical notes. Circulation 2023;148(Suppl 1):A16401. [doi: 10.1161/circ.148.suppl_1.16401]

XSL•FO
RenderX

72. Chowdhury M, Lim E, Higham A, McKinnon R, Ventoura N, He Y, et al. Can large language models safely address patient questions following cataract surgery? 2023 Presented at: Proceedings of the 5th Clinical Natural Language Processing Workshop; June 10, 2023; Toronto, Canada p. 131-137. [doi: 10.18653/v1/2023.clinicalnlp-1.17]
73. Kleinig O, Gao C, Kovoor JG, Gupta AK, Bacchi S, Chan WO. How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality. Eye (Lond) 2024;38(4):649-653 [FREE Full text] [doi: 10.1038/s41433-023-02772-w] [Medline: 37798360]
74. Arsenyan V, Bughdaryan S, Shaya F, Small K, Shahnazaryan D. Large language models for biomedical knowledge graph construction: information extraction from EMR notes. ArXiv Preprint posted online on December 9, 2023 [FREE Full text] [doi: 10.48550/arXiv.2301.12473]
75. Kwon T, Ong KT, Kang D, Moon S, Lee JR, Hwang D, et al. Large language models are clinical reasoners: reasoning-aware diagnosis framework with prompt-generated rationales. 2024 Presented at: Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence; February 20, 2024; Vancouver p. 18417-18425. [doi: 10.1609/aaai.v38i16.29802]
76. Wang C, Liu S, Li A, Liu J. Text dialogue analysis for primary screening of mild cognitive impairment: development and validation study. J Med Internet Res 2023;25:e51501 [FREE Full text] [doi: 10.2196/51501] [Medline: 38157230]
77. Li J, Wang L, Chen X, Deng X, Wen H, You M, et al. Are you asking GPT-4 medical questions properly?—Prompt engineering in consistency and reliability with evidence-based guidelines for ChatGPT-4: a pilot study. ResearchSquare. Posted online on October 3, 2023 2023:1-20 [FREE Full text] [doi: 10.21203/rs.3.rs-3336823/v1]
78. Zaidat B, Lahoti YS, Yu A, Mohamed KS, Cho SK, Kim JS. Artificially intelligent billing in spine surgery: an analysis of a large language model. Global Spine J 2023:21925682231224753 [FREE Full text] [doi: 10.1177/21925682231224753] [Medline: 38147047]
79. Datta S, Lee K, Paek H, Manion FJ, Ofoegbu N, Du J, et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. J Am Med Inform Assoc 2024;31(2):375-385 [FREE Full text] [doi: 10.1093/jamia/ocad218] [Medline: 37952206]
80. White R, Peng T, Sripitak P, Rosenberg Johansen A, Snyder M. CliniDigest: a case study in large language model based large-scale summarization of clinical trial descriptions. 2023 Presented at: Proceedings of the 2023 ACM Conference on Information Technology for Social Good; September 6-8, 2023; Lisbon, Portugal. [doi: 10.1145/3582515.3609559]
81. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023;9:e49877 [FREE Full text] [doi: 10.2196/49877] [Medline: 37948112]
82. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024;30(2):80-90 [FREE Full text] [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
83. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan SK, et al. From text to tables: a local privacy preserving large language model for structured information retrieval from medical documents. MedRxiv Preprint posted online on December 8, 2023 [FREE Full text] [doi: 10.1101/2023.12.07.23299648]
84. Hamed E, Eid A, Alberry M. Exploring ChatGPT's potential in facilitating adaptation of clinical guidelines: a case study of diabetic ketoacidosis guidelines. Cureus 2023;15(5):e38784 [FREE Full text] [doi: 10.7759/cureus.38784] [Medline: 37303347]
85. Leypold T, Schäfer B, Boos A, Beier JP. Can AI think like a plastic surgeon? Evaluating GPT-4's clinical judgment in reconstructive procedures of the upper extremity. Plast Reconstr Surg Glob Open 2023;11(12):e5471 [FREE Full text] [doi: 10.1097/GOX.0000000000005471] [Medline: 38093728]
86. Yang J, Liu C, Deng W, Wu D, Weng C, Zhou Y, et al. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. Patterns (NY) 2024;5(1):100887 [FREE Full text] [doi: 10.1016/j.patter.2023.100887] [Medline: 38264716]
87. Xiong L, Zeng Q, Deng W, Luo W, Liu R. A novel approach to nursing clinical intelligent decision-making: integration of large language models and local knowledge bases. ResearchSquare Preprint posted online on December 8, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3756467/v1]
88. Zeng Q, Liu Y, He P. A medical question classification approach based on prompt tuning and contrastive learning. 2023 Presented at: The Thirty Fifth International Conference on Software Engineering and Knowledge Engineering (SEKE 2023); July 1-10, 2023; San Francisco, CA, United States p. 632-635. [doi: 10.18293/seke2023-025]
89. Zhao D, Yang Y, Chen P, Meng J, Sun S, Wang J, et al. Biomedical document relation extraction with prompt learning and KNN. J Biomed Inform 2023;145:104459. [doi: 10.1016/j.jbi.2023.104459] [Medline: 37531999]
90. Zhu T, Qin Y, Feng M, Chen Q, Hu B, Xiang Y. BioPRO: context-infused prompt learning for biomedical entity linking. IEEE/ACM Trans Audio Speech Lang Process 2024;32(2023):374-385. [doi: 10.1109/taslp.2023.3331149]
91. Liu C, Zhang S, Li C, Zhao H. CPK-Adapter: infusing medical knowledge into K-adapter with continuous prompt. 2023 Presented at: 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP); April 21-23, 2023; Xi'an, China p. 1017-1023. [doi: 10.1109/icsp58490.2023.10248750]

XSL•FO
RenderX

92. Yeh HS, Lavergne T, Zweigenbaum P. Decorate the examples: a simple method of prompt design for biomedical relation extraction. 2022 Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; August 13, 2024; Marseille, France p. 3780-3787 URL: https://aclanthology.org/2022.lrec-1.403
93. Su Z, Yu X, Chen P. EPTQA: a Chinese medical prompt learning method based on entity pair type question answering. SSRN 2023:24. [doi: 10.2139/ssrn.4563840]
94. Xu H, Zhang J, Wang Z, Zhang S, Bhalerao M, Liu Y, et al. GraphPrompt: graph-based prompt templates for biomedical synonym prediction. 2023 Presented at: Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence; February 7, 2023; Washington, DC, United States p. 10576-10584. [doi: 10.1609/aaai.v37i9.26256]
95. Chen T, Stefanidis A, Jiang Z, Su J. Improving biomedical claim detection using prompt learning approaches. 2023 Presented at: 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML); August 4-6, 2023; Urumqi, China p. 369-376. [doi: 10.1109/prml59573.2023.10348317]
96. Xu Z, Chen Y, Hu B. Improving biomedical entity linking with cross-entity interaction. 2023 Presented at: Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence; February 7, 2023; Washington p. 13869-13877. [doi: 10.1609/aaai.v37i11.26624]
97. Wang Y, Wang Y, Peng Z, Zhang F, Zhou L, Yang F. Medical text classification based on the discriminative pre-training model and prompt-tuning. Digit Health 2023;9:1-14 [FREE Full text] [doi: 10.1177/20552076231193213] [Medline: 37559830]
98. Tian X, Wang P, Mao S. Open-world biomedical knowledge probing and verification. 2023 Presented at: Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG-23); December 8-9, 2023; Tokyo, Japan URL: https://ijckg2023.knowledge-graph.jp/pages/proc/paper_3.pdf
99. Lu K, Potash P, Lin X, Sun Y, Qian Z, Yuan Z, et al. Prompt discriminative language models for domain adaptation. 2023 Presented at: Proceedings of the 5th Clinical Natural Language Processing Workshop; July 14, 2023; Toronto, Canada p. 247-258. [doi: 10.18653/v1/2023.clinicalnlp-1.30]
100. Hu Y, Chen Y, Xu H. Towards more generalizable and accurate sentence classification in medical abstracts with less data. J Healthc Inform Res 2023;7(4):542-556. [doi: 10.1007/s41666-023-00141-6] [Medline: 37927376]
101. Taylor N, Zhang Y, Joyce DW, Gao Z, Kormilitzin A, Nevado-Holgado A. Clinical prompt learning with frozen language models. IEEE Trans Neural Netw Learn Syst 2023:1-11. [doi: 10.1109/TNNLS.2023.3294633] [Medline: 37566498]
102. Landi I, Alleva E, Valentine AA, Lepow LA, Charney AW. Clinical text deduplication practices for efficient pretraining and improved clinical tasks. ArXiv Preprint posted online on September 29, 2023 [FREE Full text] [doi: 10.48550/arXiv.2312.09469]
103. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981 [FREE Full text] [Medline: 37128372]
104. Sivarajkumar S, Wang Y. Evaluation of healthprompt for zero-shot clinical text classification. 2023 Presented at: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); June 26-29, 2023; Houston, TX, United States p. 492-494. [doi: 10.1109/ichi57859.2023.00081]
105. Zhang L, Liu J. Intent-aware prompt learning for medical question summarization. 2022 Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 6-8, 2022; Las Vegas, NV, United States p. 672-679. [doi: 10.1109/bibm55620.2022.9995317]
106. Alleva E, Landi I, Shaw LJ, Böttinger E, Fuchs TJ, Ensari I. Keyword-optimized template insertion for clinical information extraction via prompt-based learning. ArXiv Preprint posted online on October 31, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.20089]
107. Cui Z, Yu K, Yuan Z, Dong X, Luo W. Language inference-based learning for low-resource Chinese clinical named entity recognition using language model. J Biomed Inform 2024;149:104559. [doi: 10.1016/j.jbi.2023.104559] [Medline: 38056702]
108. Ahmed A, Zeng X, Xi R, Hou M, Shah SA. MED-Prompt: a novel prompt engineering framework for medicine prediction on free-text clinical notes. J King Saud Univ Comput Inf Sci 2024;36(2):1-17. [doi: 10.1016/j.jksuci.2024.101933]
109. Lu Y, Liu X, Du Z, Gao Y, Wang G. MedKPL: a heterogeneous knowledge enhanced prompt learning framework for transferable diagnosis. J Biomed Inform 2023;143:104417 [FREE Full text] [doi: 10.1016/j.jbi.2023.104417] [Medline: 37315832]
110. Cui Y, Han L, Nenadic G. MedTem2.0: prompt-based temporal classification of treatment events from discharge summaries. 2023 Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop); July 10-12, 2023; Toronto, Canada p. 160-183. [doi: 10.18653/v1/2023.acl-srw.27]
111. Wang Z, Sun J. PromptEHR: conditional electronic healthcare records generation with prompt learning. 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics; October 11, 2022; Abu Dhabi, United Arab Emirates p. 2873-2855. [doi: 10.18653/v1/2022.emnlp-main.185]
112. Wang S, Tang L, Majety A, Rousseau JF, Shih G, Ding Y, et al. Trustworthy assertion classification through prompting. J Biomed Inform 2022;132:104139 [FREE Full text] [doi: 10.1016/j.jbi.2022.104139] [Medline: 35811026]

XSL•FO
RenderX

113. Yao Z, Tsai J, Liu W, Levy DA, Druhl E, Reisman JI, et al. Automated identification of eviction status from electronic health record notes. J Am Med Inform Assoc 2023;30(8):1429-1437 [FREE Full text] [doi: 10.1093/jamia/ocad081] [Medline: 37203429]
114. Kwon S, Wang X, Liu W, Druhl E, Sung ML, Reisman JI, et al. ODD: a benchmark dataset for the natural language processing based opioid related aberrant behavior detection. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 16, 2024; Mexico City p. 1-22. [doi: 10.18653/v1/2024.naacl-long.244]
115. Su J, Zhang J, Peng P, Wang H. EGDE: a framework for bridging the gap in medical zero-shot relation triplet extraction. 2023 Presented at: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385666]
116. Li Q, Wang Y, You T, Lu Y. BioKnowPrompt: incorporating imprecise knowledge into prompt-tuning verbalizer with biomedical text for relation extraction. Inf Sci 2022;617:346-358. [doi: 10.1016/j.ins.2022.10.063]
117. Peng C, Yang X, Chen A, Yu Z, Smith KE, Costa AB, et al. Generative large language models are all-purpose text analytics engines: text-to-text learning is all your need. J Am Med Inform Assoc 2024 Sep 01;31(9):1892-1903 [FREE Full text] [doi: 10.1093/jamia/ocae078] [Medline: 38630580]
118. Peng C, Yang X, Smith KE, Yu Z, Chen A, Bian J, et al. Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. J Biomed Inform 2024;153:104630 [FREE Full text] [doi: 10.1016/j.jbi.2024.104630] [Medline: 38548007]
119. Duan J, Lu F, Liu J. MVP: optimizing multi-view prompts for medical dialogue summarization. 2023 Presented at: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385916]
120. Rohanian O, Jauncey H, Nouriborji M, Kumar V, Gonalves BP, Kartsonaki C, et al. Using bottleneck adapters to identify cancer in clinical notes under low-resource constraints. 2023 Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; July 23, 2023; Toronto, Canada p. 6239-6278. [doi: 10.18653/v1/2023.bionlp-1.5]
121. Elfrink A, Vagliano I, Abu-Hanna A, Calixto I. Soft-prompt tuning to predict lung cancer using primary care free-text Dutch medical notes. In: Artificial Intelligence in Medicine. Cham: Springer Nature Switzerland; 2023:193-198.
122. Singh Rawat BP, Yu H. Parameter efficient transfer learning for suicide attempt and ideation detection. 2022 Presented at: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI); September 7, 2022; Abu Dhabi, United Arab Emirates p. 108-115. [doi: 10.18653/v1/2022.louhi-1.13]
123. Xu S, Wan X, Hu S, Zhou M, Xu T, Wang H, et al. COSSUM: towards conversation-oriented structured summarization for automatic medical insurance assessment. 2022 Presented at: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; August 14-18, 2022; Washington, DC, United States p. 4248-4256. [doi: 10.1145/3534678.3539116]
124. Shaitarova A, Zaghir J, Lavelli A, Krauthammer M, Rinaldi F. Exploring the latest highlights in medical natural language processing across multiple languages: a survey. Yearb Med Inform 2023;32(1):230-243 [FREE Full text] [doi: 10.1055/s-0043-1768726] [Medline: 38147865]
125. Ding N, Hu S, Zhao W, Chen Y, Liu Z, Zheng HT, et al. OpenPrompt: an open-source framework for prompt-learning. ArXiv Preprint posted online on November 3, 2021 [FREE Full text] [doi: 10.48550/arXiv.2111.01998]
126. Ducel F, Fort K, Lejeune G, Lepage Y. Do we name the languages we study? The #BenderRule in LREC and ACL articles. 2022 Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; June 20-25, 2022; Marseille, France p. 564-573 URL: https://aclanthology.org/2022.lrec-1.60
127. Luo Y, Yang Z, Meng F, Li Y, Zhou J, Zhang Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. ArXiv Preprint posted online on April 2, 2024 [FREE Full text] [doi: 10.48550/arXiv.2308.08747]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
CoT: chain-of-thought
LLM: large language model
MCQ: multiple-choice question
MLM: masked language modeling
NLP: natural language processing
PD: prompt design
PL: prompt learning
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PT: prompt tuning

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 14.05.24; peer-reviewed by B Bhasuran, D Hu, A Jain; comments to author 03.07.24; revised version received 09.07.24; accepted 22.07.24; published 10.09.24.

Please cite as:
Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C
Prompt Engineering Paradigms for Medical Applications: Scoping Review
J Med Internet Res 2024;26:e60501
URL: https://www.jmir.org/2024/1/e60501
doi:10.2196/60501
PMID:39255030

©Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, Christian Lovis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Usability Evaluation Methods Used in Electronic Discharge Summaries: Literature Review

Wubshet Tesfaye[1,2], PhD; Margaret Jordan[1], PhD; Timothy F Chen[1], PhD; Ronald Lynel Castelino[1], PhD; Kamal Sud[3,4], MBBSMD; Racha Dabliz[5], PhD; Parisa Aslani[1], PhD

[1]The University of Sydney School of Pharmacy, Sydney, Australia
[2]The University of Queensland School of Pharmacy, Brisbane, Australia
[3]The University of Sydney School of Medicine, Sydney, Australia
[4]Nepean Kidney Research Centre, Nepean Hospital, Sydney, Australia
[5]Australian Commission on Safety and Quality in Health Care, Sydney, Australia

Corresponding Author:
Parisa Aslani, PhD
The University of Sydney School of Pharmacy
A15, Science Rd, Camperdown NSW 2050
Sydney
Australia
Phone: 61 61 2 9036 6541
Email: parisa.aslani@sydney.edu.au

Abstract

Background: With the widespread adoption of digital health records, including electronic discharge summaries (eDS), it is important to assess their usability in order to understand whether they meet the needs of the end users. While there are established approaches for evaluating the usability of electronic health records, there is a lack of knowledge regarding suitable evaluation methods specifically for eDS.

Objective: This literature review aims to identify the usability evaluation approaches used in eDS.

Methods: We conducted a comprehensive search of PubMed, CINAHL, Web of Science, ACM Digital Library, MEDLINE, and ProQuest databases from their inception until July 2023. The study information was extracted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). We included studies that assessed the usability of eDS, and the systems used to display eDS.

Results: A total of 12 records, including 11 studies and 1 thesis, met the inclusion criteria. The included studies used qualitative, quantitative, or mixed methods approaches and reported the use of various usability evaluation methods. Heuristic evaluation was the most used method to assess the usability of eDS systems (n=7), followed by the think-aloud approach (n=5) and laboratory testing (n=3). These methods were used either individually or in combination with usability questionnaires (n=3) and qualitative semistructured interviews (n=4) for evaluating eDS usability issues. The evaluation processes incorporated usability metrics such as user performance, satisfaction, efficiency, and impact rating.

Conclusions: There are a limited number of studies focusing on usability evaluations of eDS. The identified studies used expert-based and user-centered approaches, which can be used either individually or in combination to identify usability issues. However, further research is needed to determine the most appropriate evaluation method which can assess the fitness for purpose of discharge summaries.

(J Med Internet Res 2024;26:e55247) doi:10.2196/55247

KEYWORDS

electronic discharge summaries; usability testing; heuristic evaluation; heuristics, think-aloud; adoption; digital health; usability; electronic; discharge summary; end users; evaluation; user-centered

XSL•FO
RenderX

Introduction

The adoption of digital health platforms for collecting, sharing, and analyzing health information has shown positive associations with improvements in health care quality, service delivery, and clinical benefits including patient safety [1-4]. Electronic health records (eHRs) have become essential in acute care facilities as they enable the collection, sharing, and analysis of patient-related information, facilitating communication within and across health care settings. However, despite the substantial growth in the digitalization of health information exchange platforms, the complexity of many systems used by health care providers often poses challenges in achieving interoperability across different settings [5,6].

Differences in electronic systems used across different health settings can affect the exchange of relevant patient health and clinical information, especially during transitions of care or clinical handover [7,8]. Evidence indicates that the suboptimal communication between hospitals and external health care providers leads to discrepancies in medication records, duplication of tests, and avoidable delays in service provision, especially affecting vulnerable populations including those with low levels of health care literacy [9,10]. Hence, a coordinated health system with improved health information exchange, usability, and interoperability across health facilities and settings has significant potential to improve postacute care transition and overall patient safety [11].

Hospital discharge is a high-risk event where inaccurate or delayed transfer of clinical information, including medication plans, can significantly risk patient safety and cause medication-related issues [12,13]. Therefore, the clinical handover at the point of hospital discharge is a crucial step in patient care that determines the quality of care and patient safety. The introduction of electronic discharge summaries (eDS) has greatly improved the timely transmission of information to relevant stakeholders, mainly those in the primary care setting [8,14]. eDS, defined as "an end-to-end electronic transfer from the hospital to the community, using a secure messaging system, with the information populated using both pre-populated fields and manual transcription" [15] has seen increased adoption over the past decades. However, to further improve the quality of care and reduce communication delays between health settings, it is crucial that eDS should be user friendly [16]. This will help in minimizing avoidable patient harm incidents caused due to usability issues.

Usability is generally defined as "the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" [17]. In the context of electronic systems, usability refers to whether the system is useful, usable, and satisfying for the intended users to enable completions of intended tasks in certain sequences [18]. Evidence indicates that there are several usability issues identified with eHRs, such as those related to data entry and alerts, interoperability issues, display, automation, and workflow [16]. These usability problems in addition to affecting the implementation of such systems have implications for patient safety such as medication error and use of inappropriate medication doses [16,19]. Evaluation of systems used to prepare eDS provides an opportunity to identify and improve usability issues with existing systems. Usability evaluation involves assessing performance, efficiency, and satisfaction of electronic interfaces and can identify usability issues with eHRs to thereby propose interventions to improve designs of interfaces, their learnability, and service efficiency [20]. While various international organizations have developed and provided guidelines on the content, form, and presentation of eDS [21-27], less is known about the usability of eDS and systems used to display eDS and their potential impact on quality of care.

Evidence from systematic reviews have identified a range of usability evaluation techniques applied broadly to eHRs, which include heuristic evaluation, cognitive walkthrough, think-aloud, user-testing, observation, coupled with use of questionnaires and interviews to assess participants' perspectives and satisfaction [20,28]. However, there is limited evidence on the usability evaluations applied specifically to eDS. Therefore, the aim of this literature review was to identify the usability evaluation techniques that have been used to assess the usability of eDS.

Methods

This literature review is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [29].

Literature Search

We searched PubMed, CINAHL, Web of Science, ACM Digital Library, MEDLINE, and ProQuest databases from their inception until Jul 2023. The main concepts used for developing our search strategy included the following and are tailored for the individual databases. *Concept 1*: "usability evaluation" OR "usability testing" OR "usability test" OR "usability engineering" OR "usability inspection" AND *Concept 2*: ("discharge summar*" OR "discharge communication" OR "continuity of care" OR "transfer of care" OR "clinical handover" OR "electronic discharge" OR "patient discharge").

To capture unpublished and unindexed documents, a gray literature search was conducted using Google Scholar and via a range of governmental and health authorities' websites and guidelines. Reference lists of included studies were also manually searched to identify further eligible studies or government reports which may have been missed during our search. The full search strategy for all databases including gray literature sources is presented in Multimedia Appendix 1.

Study Selection

Search results were screened for eligibility following predefined inclusion and exclusion criteria. The retrieved studies were exported to EndNote and subsequently transferred to Covidence [30]. After removal of duplicates, the remaining documents were screened using title, abstract, and full-text by 2 independent reviewers (WT and MJ), with disagreements resolved via discussion until consensus was reached.

XSL•FO
RenderX

Eligibility Criteria

We included studies that used usability evaluation of eDS or discharge communication or those that evaluated usability issues of eHRs used to prepare an eDS and may also have implication or relevance for eDS. The relevance of eHRs for inclusion was determined based on whether the included studies assessed electronic system interactions without explicitly mentioning eDS (eg, cross-facility health information exchange) or were using an electronic platform that is also known to have an eDS component (eg, My Health Record—an Australian digital platform containing a secure web-based summary of key patient health information, where health care providers can access the system to view and upload information). We also considered studies that focused on electronic health information to patients, with the aim of assessing usability of such information to improve care after discharge. Quantitative, qualitative, and mixed methods studies were all eligible for inclusion.

Studies that evaluated the effectiveness of transfer of care tools or interventions on quality of care or patient outcomes but did not include usability evaluation of eDS were excluded. Studies addressing the use of tools without any usability assessment were also not the focus of this review. Publications in languages other than English were excluded. Finally, we also excluded protocol studies without any preliminary findings.

Operational Definitions

Discharge Summary

A range of information about events during care by a provider or organization, with the goal to provide relevant patient, clinical, and administrative information that enables a continuity of care upon patient's discharge from hospital [21]. While our primary focus is on discharge summaries, we have expanded our scope to include studies addressing usability issues with electronic discharge instructions or information provided to patients or other health care professionals. This was mainly done to understand and address the information needs and preferences of patients during their transition across different types of care.

Electronic Discharge Summary

Refers to a computerized form of discharge summary or instructions typically generated within electronic health records used in tertiary care.

Usability (of eHRs)

Refers to whether the electronic system is useful, usable, and satisfying for the intended users to enable completions of intended tasks in certain sequences [18].

End User

The user of the electronic interfaces, who could be health professionals (eg, physicians, nurses, pharmacists, etc) or consumers (patients or their caregivers).

Data Extraction and Synthesis

We extracted the following information from included studies: study characteristics (authors, publication year, and country), characteristics of end users or participants targeted, study design used (eg, mixed methods, qualitative, etc), usability evaluation method used (eg, questionnaires, interviews, heuristic evaluation, etc), study outcomes reported, and conclusions and limitations. These data were extracted from included studies using a standardized data extraction format that was modified from the Joanna Briggs Institute's manual for evidence synthesis [31], which can be found in Multimedia Appendix 2. Due to the nature of the included studies or heterogeneity of study participants and findings, quantitative analysis or meta-synthesis was not possible; however, we conducted a systematic narrative synthesis of the major study findings and their implications.

Results

Characteristics of Evidence Source

Our search identified a total of 775 records (see PRISMA flowchart in Figure 1). Of these, 34 were relevant for full text review. After removing duplicate and irrelevant records, 12 studies met the eligibility criteria and were included in this review [32-43].

XSL•FO
RenderX

Figure 1. Flow diagram for study selection process.

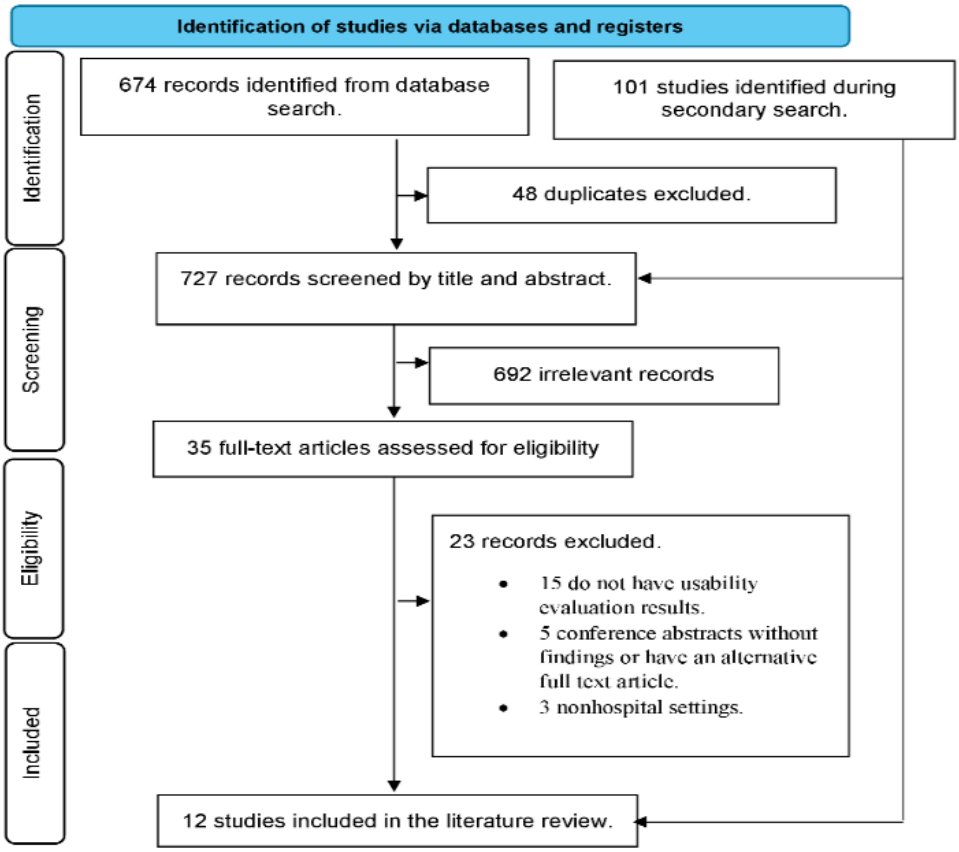

The studies were conducted in the United States (n=5) [32,36,38-40], Australia (n=3) [34,41,42], Germany (n=2) [33,35], and one each from Canada [37] and France [43] and had used qualitative (including document review and semistructured interviews) or mixed methods (n=9) [32,33,35,37-39,41-43], or observational (n=2) [34,36] methods. One document was a thesis containing a study that used experimental and survey methods and presented some findings on usability testing [40]. Table 1 presents the key characteristics and major findings of included studies, while the detailed study findings are summarized in Multimedia Appendix 2.

XSL•FO
RenderX

Table 1. Key study characteristics and major findings.

Study	Country	Study design	Participants	Primary aim	Target of the usability evaluation	Major findings
Barton et al [32]	United States	Qualitative evaluation	Emergency medicine physicians, nurses, geriatrician	To assess a method for integrating diverse expertise such as clinical, patient, care partner, and IT, in the evaluation of patient-facing emergency department after visit summary.	eDS[a]	Identified usability issues related to readability, comprehensibility, and content organization, highlighting the need to integrate experts' perspectives during design.
Busse et al [33]	Germany	Mixed methods (Qualitative evaluation and observational)	Pediatric palliative care health care professionals	To evaluate how potential users from the pediatric palliative care setting perceived an electronic cross-facility system.	Both contents of cross-facility medical records and the system used for presentation	Identified critical need for data transfer automation and suggested improvements in search functions and visualizations.
Doyle et al [34]	Australia	Exploratory mixed methods	Parents of children and physicians	To understand parent and clinician experience of discharge communication and engagement in clinical research.	System used for presenting electronic discharge instructions	High success rates and satisfaction scores were observed for both mobile and desktop interfaces, with most tasks completed successfully.
Kernbeck et al [35]	Germany	Qualitative observational	Pediatric palliative care professionals	To evaluate the acceptance of the medication module from potential users' perspective and to involve them in the development process.	Both contents of cross-facility medical records and the system used for presentation	Identified usability issues related to performance expectancy and learnability, emphasized clarity, and reduced cognitive load.
Naik et al [36]	United States	Observational (user-centered)	People with colorectal cancer	To transform physician-centered discharge warnings into patient-friendly format using health literacy and usability heuristics standards and cognitive interviews.	Both eDS contents and the system used for presentation	Identified inconsistencies in content presentation and readability, highlighted importance of a patient-centered design.
Soto et al [37]	Canada	Mixed methods study	General practitioners, family medicine residents	To improve health information exchange and use of clinical information for decision making.	System used for eDS presentation	Identified usability issues related to drug prescription and medication list visualization.
Tremoulet et al [38]	United States	Qualitative evaluation	Human factors experts, medical professionals	To conduct heuristic evaluation to identify potential usability problems and their level of severity.	Both eDS contents and the system used for presentation	Identified usability issues related to content, comprehensibility, readability, presentation, and organizational aspects of medical documents.
Tremoulet et al [39]	United States	Literature review with mixed methods study	Primary care physicians, nurses, nursing and medical directors, social workers, transition-of-care nurses	To provide insight into how existing acute care eDS support outpatient providers in the coordination of care of older adults.	eDS	Identified usability issues affecting care coordination, emphasized need for standardization of discharge summaries.
Vaigneur [40]	United States	Experimental and survey	Novice readers (caregivers) of discharge instructions	To examine the impact of adjusting readability level of discharge instructions on user comprehension and recall.	eDS	High readability discharge instructions received more attention, better comprehension, and reduced mental demand compared to low readability instructions.

XSL•FO
RenderX

Study	Country	Study design	Participants	Primary aim	Target of the usability evaluation	Major findings
Walsh et al [41]	Australia	Qualitative evaluation	My Health Record users	To identify potential usability issues within My Health Record focusing on eHealth literacy.	Both contents of health information summary and the system used for presentation	Identified usability violations and problems related to language use, website navigation, design elements, and registration processes.
Walsh et al [42]	Australia	Qualitative evaluation	My Health Record users	To identify usability issues with My Health Record through an updated heuristic evaluation.	Both contents of health information summary and the system used for presentation	Identified violations of usability heuristics and highlighted unmet needs for individuals with low eHealth literacy.
Watbled et al [43]	France	Mixed methods	Human factors experts, medical professionals	To apply a combination of methods for longitudinal usability evaluation throughout the system development lifecycle and to identify causes of usability flaws.	System used for presentation of eDS	Identified multiple usability flaws in voice recording systems and emphasized thorough analysis and context-specific evaluations.

[a]eDS: electronic discharge summaries.

Usability Evaluation Methods and Targets

Over half of the included studies [32,36-39,41-43] used a heuristic evaluation method alone or in combination with other methods (Table 2). The method by Nielsen et al or its modified versions [44,45] was the most used heuristic evaluation approach among the included studies [32,36,38,39,41,42]. Watbled et al [43] reported a modified version of the heuristic usability evaluation method known as heuristic walkthroughs.

Table 2. Usability evaluation techniques used.

Author	Heuristic evaluation	Think-aloud	Laboratory testing (*in situ* observation, eye-tracking)	Questionnaire (system usability survey)	Interview	Remote evaluation
Barton et al [32]	✓					
Busse et al [33]		✓			✓	✓
Doyle et al [34]		✓	✓	✓		
Kernbeck et al [35]		✓			✓	
Naik et al [36]	✓		✓			
Soto et al [37]	✓	✓			✓	
Tremoulet et al [38]	✓					
Tremoulet et al [39]	✓			✓	✓	
Vaigneur [40]			✓	✓		
Walsh et al [41]	✓					
Walsh et al [42]	✓					
Watbled et al [43]	✓	✓				

User testing methods such as think-aloud [33,35], eye-tracking and *in situ* observation techniques [40,43] were also used for usability evaluation of eDS systems. A combination of evaluation methods (eg, heuristics with think-aloud technique or use of think-aloud method along with a questionnaire) were also used in certain instances [40,43]. Questionnaires like the system usability survey (SUS) [38-40] and semistructured interviews [33,35,37,39] were also used by multiple studies together with other usability evaluation approaches to assess satisfaction and perception of system users. Remote evaluation (via a Zoom-based videoconference) was successfully applied in 1 study [33].

The usability evaluation studies focused on different participant categories. In the heuristic evaluation, the studies mainly involved experts who assessed usability of interface design, while others focused on either end users or a combination of experts and end users. The targeted end users included clinicians, medical secretaries, nurses, patients, or caregivers, while the experts were human factors experts [36,38,43] and domain

XSL•FO
RenderX

experts (people with knowledge of broader health system and are experienced users of My Health Record) [41,42].

Summary of Major Findings

The included studies identified several usability problems with varying degree of severity in both the eDS as well as the systems used to prepare and display eDS. While some studies focused on the usability of eDS, such as content, comprehensibility, structure, and readability issues [32,36,38-40], other studies evaluated the usability of eDS systems from presentation, design, and ease of use points of view [33,35,41-43]. The studies used different usability metrics such as usefulness, system efficiency, learnability, performance, and satisfaction when evaluating the usability of the targeted systems [33-35,37,38,40]. While the use of heuristic evaluation identified several organizational, layout, and formatting-related usability issues with different systems used to host eDS, the combined approach of using heuristic walkthroughs with user testing proposed by Watbled et al [43] tended to identity more severe problems and also highlighted their potential negative impact. These included issues related to error management, workload, and compatibility. These issues could lead to serious outcomes, such as prolonged deadlines for task completion, mistakes in patient identification, and inadequate error detection by users [43].

Studies that used heuristic evaluation overall identified several content, comprehensibility, readability, and structural usability flaws [32,33,36,38,40]. Visualization and presentation problems (eg, visualization of medication list or diagnosis and clarity and readability of medication documentation) were among the domains identified to have the highest number of usability problems and may have an impact on patient comprehension and safety [32,33,35,37,38]. Further, design readability and layout issues were identified to have an association with longer duration of screen gazes, affecting comprehension of discharge instructions [40]. One study reported that less display fragmentation and data entry requirements can reduce the cognitive load of user, confusion, and usability concerns [35]. Similarly, a study that used the eye-tracking method demonstrated that improving readability and layout was associated with less mental demand [40].

Concerns with language use, interface layout, and lack of audio-visuals were identified as common usability flaws in Australian studies that used usability issues with My Health Record, with implications for people with low electronic health literacy [41,42]. Another Australian study that assessed user satisfaction using the SUS questionnaire highlighted high acceptability of a digital discharge communication tool, with consumers and clinicians reporting high satisfaction scores on the mobile (94%) and desktop (93%) interfaces, respectively [34].

One study, which involved an information technology expert, assessed the likelihood of addressing usability issues for a patient-facing emergency department visit summaries [32]. The study reported that nearly half of the usability issues identified were difficult to address (31/76 issues). These are issues with some information originating from different service vendors or when an eHR vendor was responsible for providing parts of the discharge summaries (eg, headers, content, and order of sections).

Discussion

This review summarizes study findings on usability evaluation approaches used to assess eDS and eDS systems. The limited published evidence revealed the use of heterogenous usability evaluation techniques spanning from one conducted by experts to laboratory and user testing to the use of questionnaires and interviews. Broadly, our findings highlight that the use of heuristics (expert based) and think-aloud (user centered) were the most used methods for evaluation of eDS and eDS systems. Other techniques like eye-tracking, direct observation, questionnaire- and interview-based evaluations were also used in combination with either heuristic or think-aloud approaches.

Heuristic evaluation method, consistent with previous findings on eHRs [28], was used by most of the included studies for evaluation of eDS and eDS systems usability. This technique typically involves the application of a procedure including 3-5 experts to independently apply a set of best practices design (referred as heuristics) to identify usability flaws with system interfaces [44]. The heuristics used in evaluation are either defined *a priori* by experts or are derived from standard guidelines like the ergonomic criteria [46], which has 8 main domains around guidance, workload, explicit control, adaptability, error management, consistency, significance of codes, and compatibility.

Our review indicates that heuristic evaluation method can successfully identify a range of usability issues around readability, comprehensibility, organization, and content of eDS interfaces [38,39]. In addition to identifying usability flaws with user interfaces, heuristic evaluation also enabled assessment of the severity of usability problems. The severity of usability problems is often rated based on the 5-step severity scale developed by Nielsen et al [47], which is a tool widely applied to assess the usability of medical technologies and their impact on patient safety. This severity scale ranges in value from “0” for no usability problem to “4” for usability catastrophe, with mean scores of judgements from multiple evaluators used during heuristic usability evaluation [47].

While heuristic evaluation has the advantage of being more intuitive, efficient, and cheap, with less requirements for advanced planning and involvement of test users [44], it only identifies half of the usability problems that are related to the design of system interfaces [48]. A modified version called heuristic walkthroughs, which also involves the observation of end users, was associated with better detection of usability problems, mainly those characterized as moderate and severe [49]. This has been confirmed by one of the studies that reported heuristic walkthroughs to be effective in the identification of more severe usability problems [43]. Despite the advantages of heuristic evaluation, there are certain limitations associated with this approach. For example, the heterogenous nature of heuristics or guidelines applied in different settings indicates the lack of gold-standard guidelines applicable to every context [50]. Also, because heuristics are broadly defined, they are often interpreted and applied differently by different experts [50]. These

XSL•FO
RenderX

limitations highlighted the need to explore alternative approaches of usability evaluation, preferably those that also consider input from end users.

The think-aloud technique is one of the controlled user testing approaches used by multiple studies in our review to evaluate the usability of eDS [33-35,37,43]. This method requires participants to verbalize their impressions about an interface while using it, enabling data collection from both direct observation and users' self-reported statements [51,52]. This method has the advantage of providing insights into both design and learnability problems associated with systems [50]. Kernebeck et al [35] demonstrated that both effort and performance metrics can be effectively captured using a concurrent think-aloud evaluation approach, and emphasized the critical need to involve actual users from the start of the development process to enable a more transparent evaluation that meets the needs of end users. More importantly, the findings from Watbled et al [43] highlighted that this approach can be successfully integrated with heuristics, offering an advantage of a more holistic assessment of problems from both experts and users.

Eye-tracking is another controlled user testing method used for eDS usability evaluation [40]. In this method, eye-trackers record and analyze information on eye movement, fixation, and screen gaze to assess if the tasks involved are demanding. Questionnaires and semistructured interviews were also used to assess the usability issues with eHRs. Our review identified the use of the SUS [53]—a 10-item Likert tool that provides overall assessment of system usability. The SUS is a nonproprietary self-administered questionnaire with good validity and reliability; however, it is not robust and specific enough in identifying usability issues specific to eHRs.

The usability evaluation techniques used in the included studies, such as the use of heuristics, were not only used to identify issues related to eDS content, such as unnecessary or missing information, poor organization, and inconsistencies in formatting [32,38] but also used to understand the visualizations within eDS systems, including those associated with presenting medication lists and diagnoses [33,35,37,43]. The usability problems identified in the eDS systems had significant consequences, for example, the need for extended deadlines for task completion and errors in patient identification, which ultimately impacted the system's quality and performance [43]. These findings emphasize the importance of improving the speed and quality of systems when designing technologies for use in the context of eDS. It has been proposed that integrating usability testing methods during the development of these systems can potentially reduce adverse health events and outcomes.

In order to provide optimal and safe health services, eDS should provide clinically relevant, accurate, adequate, and clear display of relevant information. The content and quality of discharge summaries have implications for patient outcomes after their discharge from hospital [54]. While technological solutions can significantly improve the content and quality transfer of information, factors such as health literacy and individual patient differences are other important factors to consider during system implementation. This review highlights that this can be achieved through applying rigorous usability assessment techniques that require experts (heuristic evaluation and walkthroughs) approach with a user-based (think-aloud approach) method [50]. However, the limited number of studies assessing the usability of eDS or discharge instructions by patients with different levels of health literacy highlights the need for additional research.

Given the diverse user base of eHRs and discharge summaries in primary care settings, which includes physicians, nurses, pharmacists, and other health professionals, it is crucial to have systems that are easy to navigate, gather and select information, and interpret that information. Therefore, it is important to use a robust usability assessment approach that takes into account the wide range of users, including health professionals and patients, to develop a platform that can be used without significant challenges. Developing systems that can overcome usability issues, such as poor organization and display fragmentation, workflow interference, and cognitive overload, can affect the quality of information required to enable clinical decision making by health professionals and, therefore, continuity of care [55]. With emerging interest around International Patient Summary, which aims to provide a relatively generic means of communication for "unplanned, cross border care" [56], some of the identified usability techniques, especially those applicable to medication and condition summaries, can be used in this broader context.

Although most of the included studies assessed or explored different usability evaluation methods, the usability metrics used were heterogenous in nature. More studies focused on standardized usability metrics like efficiency, effectiveness, and satisfaction, as highlighted in the ISO (International Organization for Standardization) 9241-11 Ergonomics of human-system interaction [17] may shed light into the most effective approach for usability evaluation of eDS. Overall, usability evaluations applied on interfaces should aim to achieve adequate validity, thoroughness, and reliability [57]. In this context, considering the limitations with individual techniques, adopting a multimodal evaluation approach, for example, through combining heuristic evaluation with user testing methods or a questionnaire, may better achieve these objectives. More importantly, there should be an increased focus on developing and implementing usability evaluation techniques that consider factors such as learnability, regular use, error protection, accessibility, and maintainability, as highlighted in the ISO 9241-11 [17]. Another important consideration is the limited geographical locations covered by the included studies, which may limit the applicability of the findings in other settings with different electronic health systems and infrastructures. Lastly, the evidence concerning usability evaluation theories, approaches, and implementation frameworks specific to discharge summaries remains notably scarce. This highlights the need for further research in the area.

Even though we included a range of databases and gray literature sources, it is possible that we may have missed studies indexed in sources not included in this review. Our search strategy was specifically restricted to discharge summaries or instructions, which may have excluded usability evaluation techniques used in the context of EHRs in general. Some of these techniques

XSL•FO
RenderX

identified in previous works focusing on EHRs could also be relevant to eDS [28]. We also acknowledge that despite our systematic and thorough approach, the potential for bias exists due to the reliance on a single reviewer for data extraction and quality appraisal.

Conclusions

We have identified multiple usability evaluation methods that can be used to identify usability concerns applicable to eDS and eDS systems as well as other discharge communication tools. While the evidence in this area is still emerging, especially in terms of standardizing the usability metrics used, published studies indicate the use of a variety of generic methods to effectively assess different aspects of discharge summary contents. These aspects include the presence of necessary information, organization, formatting, as well as the presentation (display and layout) of the systems used to host the eDS.

Heuristic and think-aloud evaluation techniques emerged as the most used methods. They were used either independently or in conjunction with other techniques, such as validated surveys or semistructured interviews. These methods were not only used to identify usability issues with eDS and eDS systems but also revealed severe issues that had implications for the quality and performance of these systems.

Acknowledgments

We would like to thank the Australian Commission on Safety and Quality in Health Care for funding this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 16 KB - jmir_v26i1e55247_app1.docx]

Multimedia Appendix 2
Characteristics of included studies.
[DOCX File , 34 KB - jmir_v26i1e55247_app2.docx]

Multimedia Appendix 3
PRISMA checklist.
[DOC File , 65 KB - jmir_v26i1e55247_app3.doc]

References

1. Manca DP. Do electronic medical records improve quality of care? Yes. Can Fam Physician 2015 Oct;61(10):846-847 [FREE Full text] [Medline: 26472786]
2. Lin YK, Lin M, Chen H. Do Electronic Health Records Affect Quality of Care? Evidence from the HITECH Act. Information Systems Research 2019 Mar;30(1):306-318. [doi: 10.1287/isre.2018.0813]
3. Kern LM, Edwards A, Kaushal R. The patient-centered medical home, electronic health records, and quality of care. Ann Intern Med 2014 Jun 03;160(11):741-749. [doi: 10.7326/M13-1798] [Medline: 24887615]
4. King J, Patel V, Jamoom EW, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health Serv Res 2014 Feb;49(1 Pt 2):392-404 [FREE Full text] [doi: 10.1111/1475-6773.12135] [Medline: 24359580]
5. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. P T 2017;42(9):572-575 [FREE Full text] [Medline: 28890644]
6. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
7. Unnewehr M, Schaaf B, Marev R, Fitch J, Friederichs H. Optimizing the quality of hospital discharge summaries--a systematic review and practical tools. Postgrad Med 2015 Aug;127(6):630-639. [doi: 10.1080/00325481.2015.1054256] [Medline: 26074128]
8. Hesselink G, Schoonhoven L, Barach P, Spijker A, Gademan P, Kalkman C, et al. Improving patient handovers from hospital to primary care: a systematic review. Ann Intern Med 2012 Sep 18;157(6):417-428. [doi: 10.7326/0003-4819-157-6-201209180-00006] [Medline: 22986379]
9. Jones CD, Jones J, Bowles KH, Flynn L, Masoudi FA, Coleman EA, et al. Quality of hospital communication and patient preparation for home health care: results from a statewide survey of home health care nurses and staff. J Am Med Dir Assoc 2019 Apr;20(4):487-491 [FREE Full text] [doi: 10.1016/j.jamda.2019.01.004] [Medline: 30799224]
10. Brody AA, Gibson B, Tresner-Kirsch D, Kramer H, Thraen I, Coarr ME, et al. High prevalence of medication discrepancies between home health referrals and Centers for Medicare and Medicaid Services home health certification and plan of care

XSL•FO
RenderX

and their potential to affect safety of vulnerable elderly adults. J Am Geriatr Soc 2016 Nov;64(11):e166-e170 [FREE Full text] [doi: 10.1111/jgs.14457] [Medline: 27673753]
11. Moore AB, Krupp JE, Dufour AB, Sircar M, Travison TG, Abrams A, et al. Improving transitions to postacute care for elderly patients using a novel video-conferencing program: ECHO-Care transitions. Am J Med 2017 Oct;130(10):1199-1204. [doi: 10.1016/j.amjmed.2017.04.041] [Medline: 28551043]
12. Wilson S, Ruscoe W, Chapman M, Miller R. General practitioner-hospital communications: a review of discharge summaries. J Qual Clin Pract 2001 Dec;21(4):104-108. [doi: 10.1046/j.1440-1762.2001.00430.x] [Medline: 11856404]
13. Perren A, Previsdomini M, Cerutti B, Soldini D, Donghi D, Marone C. Omitted and unjustified medications in the discharge summary. Qual Saf Health Care 2009 Jun;18(3):205-208 [FREE Full text] [doi: 10.1136/qshc.2007.024588] [Medline: 19468003]
14. Reinke CE, Kelz RR, Baillie CA, Norris A, Schmidt S, Wingate N, et al. Timeliness and quality of surgical discharge summaries after the implementation of an electronic format. Am J Surg 2014 Jan;207(1):7-16. [doi: 10.1016/j.amjsurg.2013.04.003] [Medline: 24269034]
15. Safety and Quality Evaluation of Electronic Discharge Summary Systems: Final Report. URL: https://www.safetyandquality.gov.au/publications-and-resources/resource-library/safety-and-quality-evaluation-electronic-discharge-summary-systems-final-report [accessed 2022-10-07]
16. Howe JL, Adams KT, Hettinger AZ, Ratwani RM. Electronic health record usability issues and potential contribution to patient harm. JAMA 2018 Mar 27;319(12):1276-1278 [FREE Full text] [doi: 10.1001/jama.2018.1171] [Medline: 29584833]
17. Bevan N, Carter J, Harker S. ISO 9241-11 revised: what have we learnt about usability since 1998? : Springer; 2015 Presented at: International Conference on Human-Computer Interaction; 2015 Jan 01; UK. [doi: 10.1007/978-3-319-20901-2_13]
18. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. J Biomed Inform 2011 Dec;44(6):1056-1067 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.005] [Medline: 21867774]
19. Ratwani RM, Savage E, Will A, Fong A, Karavite D, Muthu N, et al. Identifying electronic health record usability and safety challenges in pediatric settings. Health Aff (Millwood) 2018 Nov;37(11):1752-1759. [doi: 10.1377/hlthaff.2018.0699] [Medline: 30395517]
20. Ellsworth MA, Dziadzko M, O'Horo JC, Farrell AM, Zhang J, Herasevich V. An appraisal of published usability evaluations of electronic health records via systematic review. J Am Med Inform Assoc 2017 Jan;24(1):218-226 [FREE Full text] [doi: 10.1093/jamia/ocw046] [Medline: 27107451]
21. National Guidelines for On-Screen Presentation of Discharge Summaries. 2017. URL: https://www.safetyandquality.gov.au/publications-and-resources/resource-library/national-guidelines-screen-presentation-discharge-summaries [accessed 2023-10-11]
22. Snow V, Beck D, Budnitz T, Miller DC, Potter J, Wears RL, American College of Physicians, Society of General Internal Medicine, Society of Hospital Medicine, American Geriatrics Society, American College of Emergency Physicians, Society of Academic Emergency Medicine. Transitions of Care Consensus Policy Statement American College of Physicians-Society of General Internal Medicine-Society of Hospital Medicine-American Geriatrics Society-American College of Emergency Physicians-Society of Academic Emergency Medicine. J Gen Intern Med 2009 Aug;24(8):971-976 [FREE Full text] [doi: 10.1007/s11606-009-0969-x] [Medline: 19343456]
23. The SIGN Discharge Document, SIGN. 2012. URL: http://www.sign.ac.uk/guidelines/fulltext/128/index.html [accessed 2022-08-02]
24. National Standard for Patient Discharge Summary Information. 2013. URL: https://www.hiqa.ie/ [accessed 2022-09-29]
25. Standards for the Structure and Content of Health and Care Records. 2018. URL: https://www.rcplondon.ac.uk/projects/outputs/standards-clinical-structure-and-content-patient-records [accessed 2024-08-02]
26. eDischarge summary standard V2.1. 2019. URL: https://theprsb.org/standards/edischargesummary/#:~:text=The%20eDischarge%20Summary%20Standard%20enables,is%20discharged%20from%20hospital%20care [accessed 2024-08-02]
27. UHN Discharge Summary. 2017. URL: https://www.uhnmodules.ca/DischargeSummary/home.html [accessed 2024-08-02]
28. Wronikowska MW, Malycha J, Morgan LJ, Westgate V, Petrinic T, Young JD, et al. Systematic review of applied usability metrics within usability evaluation methods for hospital electronic healthcare record systems: Metrics and Evaluation Methods for eHealth Systems. J Eval Clin Pract 2021 Dec;27(6):1403-1416 [FREE Full text] [doi: 10.1111/jep.13582] [Medline: 33982356]
29. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
30. Covidence. 2020. URL: https://www.covidence.org/reviewers [accessed 2020-06-23]
31. Aromataris E, Munn Z. JBI manual for evidence synthesis. South Australia: JBI; 2020.
32. Barton HJ, Salwei ME, Rutkowski RA, Wust K, Krause S, Hoonakker PL, et al. Evaluating the usability of an emergency department after visit summary: staged heuristic evaluation. JMIR Hum Factors 2023 Mar 09;10:e43729 [FREE Full text] [doi: 10.2196/43729] [Medline: 36892941]

XSL•FO
RenderX

33. Busse TS, Jux C, Kernebeck S, Dreier LA, Meyer D, Zenz D, et al. Participatory design of an electronic cross-facility health record (ECHR) system for pediatric palliative care: a think-aloud study. Children (Basel) 2021 Sep 24;8(10):839 [FREE Full text] [doi: 10.3390/children8100839] [Medline: 34682105]
34. Doyle S, Pavlos R, Carlson SJ, Barton K, Bhuiyan M, Boeing B, et al. Efficacy of digital health tools for a pediatric patient registry: semistructured interviews and interface usability testing with parents and clinicians. JMIR Form Res 2022 Jan 17;6(1):e29889 [FREE Full text] [doi: 10.2196/29889] [Medline: 35037889]
35. Kernebeck S, Jux C, Busse TS, Meyer D, Dreier LA, Zenz D, et al. Participatory design of a medication module in an electronic medical record for paediatric palliative care: a think-aloud approach with nurses and physicians. Children (Basel) 2022 Jan 06;9(1):82 [FREE Full text] [doi: 10.3390/children9010082] [Medline: 35053707]
36. Naik AD, Horstman MJ, Li LT, Paasche-Orlow MK, Campbell B, Mills WL, et al. User-centered design of discharge warnings tool for colorectal surgery patients. J Am Med Inform Assoc 2017 Sep 01;24(5):975-980 [FREE Full text] [doi: 10.1093/jamia/ocx018] [Medline: 28340218]
37. Soto M, Sicotte C, Motulsky A. Using health information exchange: usability and usefulness evaluation. Stud Health Technol Inform 2019 Aug 21;264:1036-1040. [doi: 10.3233/SHTI190382] [Medline: 31438082]
38. Tremoulet P, Krishnan R, Karavite D, Muthu N, Regli SH, Will A, et al. A heuristic evaluation to assess use of after visit summaries for supporting continuity of care. Appl Clin Inform 2018 Jul;9(3):714-724 [FREE Full text] [doi: 10.1055/s-0038-1668093] [Medline: 30208496]
39. Tremoulet PD, Shah PD, Acosta AA, Grant CW, Kurtz JT, Mounas P, et al. Usability of electronic health record-generated discharge summaries: heuristic evaluation. J Med Internet Res 2021 Apr 15;23(4):e25657 [FREE Full text] [doi: 10.2196/25657] [Medline: 33856353]
40. Vaigneur HM. Engineering hospital discharge instructions: An eye-tracking based study. Ann Arbor: Clemson University; 2015:70.
41. Walsh L, Hemsley B, Allan M, Adams N, Balandin S, Georgiou A, et al. The E-health literacy demands of Australia's my health record: a heuristic evaluation of usability. Perspect Health Inf Manag 2017;14(Fall):1f [FREE Full text] [Medline: 29118683]
42. Walsh L, Hemsley B, Allan M, Dahm MR, Balandin S, Georgiou A, et al. Assessing the information quality and usability of within a health literacy framework: What's changed since 2016? Health Inf Manag 2021;50(1-2):13-25 [FREE Full text] [doi: 10.1177/1833358319864734] [Medline: 31370712]
43. Watbled L, Marcilly R, Guerlinger S, Bastien JMC, Beuscart-Zéphir MC, Beuscart R. Combining usability evaluations to highlight the chain that leads from usability flaws to usage problems and then negative outcomes. J Biomed Inform 2018 Feb;78:12-23 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.014] [Medline: 29305953]
44. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: Proceedings of the SIGCHI conference on Human factors in computing systems; 1990 March 01; United States p. 249-256. [doi: 10.1145/97243.97281]
45. Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T. Using usability heuristics to evaluate patient safety of medical devices. J Biomed Inform 2003;36(1-2):23-30 [FREE Full text] [doi: 10.1016/s1532-0464(03)00060-1] [Medline: 14552844]
46. Scapin DL, Bastien JMC. Ergonomic criteria for evaluating the ergonomic quality of interactive systems. Behaviour & Information Technology 1997 Jan;16(4-5):220-231. [doi: 10.1080/014492997119806]
47. Nielsen J. Reliability of severity estimates for usability problems found by heuristic evaluation. 1992 Presented at: Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems; 1992 May 03; United States. [doi: 10.1145/1125021.1125117]
48. Thyvalikakath TP, Monaco V, Thambuganipalle H, Schleyer T. Comparative study of heuristic evaluation and usability testing methods. Stud Health Technol Inform 2009;143:322-327 [FREE Full text] [Medline: 19380955]
49. Sears A. Heuristic walkthroughs: finding the problems without the noise. International Journal of Human-Computer Interaction 1997 Sep;9(3):213-234. [doi: 10.1207/s15327590ijhc0903_2]
50. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009 May;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
51. Bolle S, Romijn G, Smets EMA, Loos EF, Kunneman M, van Weert JCM. Older cancer patients' user experiences with web-based health information tools: a think-aloud study. J Med Internet Res 2016 Jul 25;18(7):e208 [FREE Full text] [doi: 10.2196/jmir.5618] [Medline: 27457709]
52. Fan M, Lin J, Chung C, Truong KN. Concurrent think-aloud verbalizations and usability problems. ACM Trans. Comput.-Hum. Interact 2019 Jul 19;26(5):1-35. [doi: 10.1145/3325281]
53. Brooke J. SUS-a quick and dirty usability scale. Usability evaluation in industry 1996;189:4-7. [doi: 10.1201/9781498710411-35]
54. Lakhaney D, Banker SL. An evaluation of the content of pediatric discharge summaries. Hosp Pediatr 2020;10(11):949-954. [doi: 10.1542/hpeds.2020-0148] [Medline: 33097565]
55. Johnson C, Johnston D, Crowle P. EHR Usability Toolkit: A Background Report on Usability and Electronic Health Records. Rockville, MD: Agency for Healthcare Research and Quality; 2011.
56. Cangioli G. International patient summary implementation guide. The International Patient Summary. 2018. URL: https://international-patient-summary.net/

XSL•FO
RenderX

57. Hartson HR, Andre TS, Williges RC. Criteria For evaluating usability evaluation methods. Int J Hum-Comput Int 2001 Dec;13(4):373-410. [doi: 10.1207/s15327590ijhc1304_03]

Abbreviations

eDS: electronic discharge summaries
eHR: electronic health record
ISO: International Organization for Standardization
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SUS: system usability survey

Edited by G Tsafnat; submitted 07.12.23; peer-reviewed by S Martikainen, A Bamgboje-Ayodele, LA Gomes, L Weinert, F Lau; comments to author 12.04.24; revised version received 17.05.24; accepted 30.06.24; published 12.09.24.

Please cite as:
Tesfaye W, Jordan M, Chen TF, Castelino RL, Sud K, Dabliz R, Aslani P
Usability Evaluation Methods Used in Electronic Discharge Summaries: Literature Review
J Med Internet Res 2024;26:e55247
URL: https://www.jmir.org/2024/1/e55247
doi:10.2196/55247
PMID:

©Wubshet Tesfaye, Margaret Jordan, Timothy F Chen, Ronald Lynel Castelino, Kamal Sud, Racha Dabliz, Parisa Aslani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Artificial Intelligence–Augmented Clinical Decision Support Systems for Pregnancy Care: Systematic Review

Xinnian Lin[1*], PhD; Chen Liang[2,3*], PhD; Jihong Liu[4], SCD; Tianchu Lyu[3], MPH; Nadia Ghumman[3], MBBS, MHA; Berry Campbell[5], MD

[1]School of Education, Fuzhou University of International Studies and Trade, Fuzhou, China
[2]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, United States
[3]Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States
[4]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States
[5]Department of Obstetrics and Gynecology, School of Medicine, University of South Carolina, Columbia, SC, United States
[*]these authors contributed equally

Corresponding Author:
Chen Liang, PhD
Department of Biomedical Informatics and Medical Education
School of Medicine
University of Washington
850 Republican Street
Seattle, WA, 98109
United States
Phone: 1 2065432259
Email: cl0512@uw.edu

Abstract

Background: Despite the emerging application of clinical decision support systems (CDSS) in pregnancy care and the proliferation of artificial intelligence (AI) over the last decade, it remains understudied regarding the role of AI in CDSS specialized for pregnancy care.

Objective: To identify and synthesize AI-augmented CDSS in pregnancy care, CDSS functionality, AI methodologies, and clinical implementation, we reported a systematic review based on empirical studies that examined AI-augmented CDSS in pregnancy care.

Methods: We retrieved studies that examined AI-augmented CDSS in pregnancy care using database queries involved with titles, abstracts, keywords, and MeSH (Medical Subject Headings) terms. Bibliographic records from their inception to 2022 were retrieved from PubMed/MEDLINE (n=206), Embase (n=101), and ACM Digital Library (n=377), followed by eligibility screening and literature review. The eligibility criteria include empirical studies that (1) developed or tested AI methods, (2) developed or tested CDSS or CDSS components, and (3) focused on pregnancy care. Data of studies used for review and appraisal include title, abstract, keywords, MeSH terms, full text, and supplements. Publications with ancillary information or overlapping outcomes were synthesized as one single study. Reviewers independently reviewed and assessed the quality of selected studies.

Results: We identified 30 distinct studies of 684 studies from their inception to 2022. Topics of clinical applications covered AI-augmented CDSS from prenatal, early pregnancy, obstetric care, and postpartum care. Topics of CDSS functions include diagnostic support, clinical prediction, therapeutics recommendation, and knowledge base.

Conclusions: Our review acknowledged recent advances in CDSS studies including early diagnosis of prenatal abnormalities, cost-effective surveillance, prenatal ultrasound support, and ontology development. To recommend future directions, we also noted key gaps from existing studies, including (1) decision support in current childbirth deliveries without using observational data from consequential fetal or maternal outcomes in future pregnancies; (2) scarcity of studies in identifying several high-profile biases from CDSS, including social determinants of health highlighted by the American College of Obstetricians and Gynecologists; and (3) chasm between internally validated CDSS models, external validity, and clinical implementation.

(J Med Internet Res 2024;26:e54737) doi:10.2196/54737

XSL•FO
RenderX

KEYWORDS

artificial intelligence; biomedical ontologies; clinical decision support systems; implementation science; obstetrics; pregnancy; AI; systematic review; CDSS; functionality; methodology; implementation; database query; database queries; bibliography; record; records; eligibility; literature review; prenatal; early pregnancy; obstetric care; postpartum care; pregnancy care; diagnostic support; clinical prediction; knowledge base; therapeutic; therapeutics; recommendation; recommendations; diagnosis; abnormality; abnormalities; cost-effective; surveillance; ultrasound; ontology

Introduction

In the United States, maternal and newborn outcomes (eg, maternal and newborn mortality, preterm birth, low birth weight, congenital abnormalities, and maternal pregnancy complications) are worse than in any other resource-rich countries, where most pregnancy-related mortalities were preventable [1]. Severe maternal mortality and morbidity has led to significant short- or long-term consequences impacting not only pregnant individuals but also their families [2]. Across the health care system in the nation, there remains a limited number of pregnant individuals who have access to evidence-based, comprehensive, and continuous maternity care [1]. Maternity providers often experience insufficient information and limited guidelines to inform clinical decisions, in part because studies conducted in small sample sizes and pregnant and lactating individuals are generally excluded from clinical trials [3].

To address the chasm, the clinical decision support system (CDSS) has broad application in pregnancy care. In the clinical practice guideline provided by the American College of Obstetricians and Gynecologists, the role of CDSS was described along the pathway of patient management and was increasingly exposed within emerging clinical information systems such as electronic health records (EHRs) [4]. Because pregnancy care often is characterized by multi-modal (eg, clinical findings, medical imaging, noninvasive prenatal testing, and genomics), multi-specialty data (eg, obstetrics, maternal-fetal medicine, gynecology, reproductive endocrinology and infertility, and neonatology), and complex episodes (ie, from preconception, conception, prenatal, intrapartum, to post partum), precise and timely clinical decision support requires a very high level of EHR interoperability and clinical validity. Within the context of this study, pregnancy care is defined as the health care for mothers and fetus (or newborn) before, during, and after the pregnancy. In addition to improving patient care management, CDSS also plays a critical role in supporting evidence-based medicine and the pathway toward a learning health system, in which CDS bridges the gap between the increasingly available digital data and much-demanded actionable knowledge for therapeutics and patient care [5,6].

The recent evolution of artificial intelligence (AI) and biomedical informatics has led to new frontiers in clinical and translational medicine. Within the scope of this study, we refer to the definition of AI in health care broadly as the methods and applications that create computer systems capable of activities normally associated with cognitive effort during health care [7]. In the field of CDSS, the application of AI has become increasingly prominent in augmenting knowledge discovery, diagnostics support, risk prediction and alarming, chronic disease management, and patient monitoring, to name a few [8]. In pregnancy care, emerging research has been exploring how AI-augmented CDSS would help improve clinical workflow and patient management. However, with the vast number of clinical guidelines, diverse AI techniques, and different EHR systems and functional modules, the spectrum of capacities and characteristics that such AI-augmented CDSS would further improve pregnancy care remains unclear. Increasing numbers of AI studies in obstetrics and gynecology have been documented [9-11]. However, reviews on how AI-augmented CDSS were specialized in pregnancy care have been missing.

A systematic review is desperately needed to fill several gaps in the literature. First, maternal health decisions are preference-sensitive and have been based on limited evidence-based guidelines. Such characteristics require the CDSS to be designed taking into account both the existing clinical guidelines, carefully selected clinical data, and shared patient consent or preference data. The dynamic of these CDSS designing features appears differently in specific episodes and subspecialty of pregnancy care, which needs to be reviewed systematically yet no existing reviews have achieved this goal. Second, recent AI applications on CDSS for pregnancy care appear to be different from those from one decade ago. Although the concept of AI remains loosely defined, a systematic review is desired to update the state-of-the-art AI methodologies applied to pregnancy-related CDSS, as well as to provide a timely comparative evaluation against those historical AI technologies. Third, existing literature reviews of either AI applications or CDSS in the field of pregnancy care typically focus on the evaluation of various AI methods and the performance of a model (eg, prediction, classification, and information retrieval). No studies have evaluated model performance together with implementation assessment outcomes of the CDSS in real-world settings, which has been a missing perspective of the external validity of CDSS studies.

The objective of this study is to provide a systematic review of empirical studies that examined AI-augmented CDSS in pregnancy care and inform challenges and opportunities with respect to how findings from these emerging studies would improve pregnancy care using the framework of participants, interventions, comparisons, outcomes, and study design as reference. Specifically, we sought to (1) identify specific maternal care domains where AI-augmented CDSS plays a role, (2) characterize the current state of CDSS functions, and (3) identify limitations, challenges, and future opportunities.

Methods

The literature review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [12]. The PRISMA checklist can be found in Multimedia Appendix 1.

XSL•FO
RenderX

Bibliographic Database

We searched 3 electronic bibliographic databases: PubMed/MEDLINE (including MEDLINE and PubMed Central [PMC]), Embase, and ACM Digital Library. PubMed/MEDLINE is a bibliographic database of life sciences and biomedical topics established by the US National Library of Medicine at the National Institutes of Health in 1996. PMC was launched in 2000 as a digital counterpart to the National Library of Medicine's extensive print full-text journal collection of biomedical and life sciences. Some PMC journals are cross-indexed as MEDLINE journals. Embase is a bibliographic database focused on pharmacovigilance. The ACM Digital Library is a comprehensive database of full-text studies and bibliographic literature covering computing and information technology including biomedical informatics and digital health. Licenses for accessing Embase and ACM Digital Library are obtained by the University of South Carolina.

Search Strategy

In the PubMed/MEDLINE database, we searched strings in the field of "text word," which includes "title, abstract, other abstract, Medical Subject Headings (MeSH) terms, MeSH subheadings, publication types, substance names, personal name as subject, corporate author, secondary source, comment or correction notes, and other terms." To supply topics mis-captured or were not precisely captured by "text word," we also searched string in [MeSH Major Topic]. We adjusted the search strategy used for PubMed/MEDLINE in Embase and ACM Digital Library, respectively. We modified the search fields for Embase and ACM Digital Library because MeSH is only used in PubMed/MEDLINE and other fields vary in the three bibliographic databases. For all databases, the time of publications was constrained to be including and before 2022. No language restrictions were applied. The search strings mainly incorporated pregnancy procedures, pregnancy outcomes, CDSS models, CDSS methods, and AI methodologies (see Multimedia Appendix 2 for search strings and criteria). All electronic reference database searches were completed in January 2023. The search strategy was developed by two authors (CL and TL) with consolidated suggestions received from other authors. The search was performed by NG.

Assessment of Eligibility and Biases

One author (NG) removed duplicates when comparing results from each database. In this process, PubMed Identifier, titles, publications, and authors are used to identify unique publications. To perform the eligibility assessment, we used the following inclusion criteria: empirical studies that (1) developed or tested AI methods, (2) developed or tested CDSS or CDSS components, and (3) focused on pregnancy care. Quality and biases of studies were assessed based on several criteria adopted from the Risk of Bias 2 tool [13], which include (1) whether an empirical study, (2) concentration of "pregnancy care," "CDSS," and "AI" in the study, and (3) completeness, clarify, and validity of methods, results, and conclusion as reported in the publications. Full-text manuscripts of potentially relevant studies were reviewed for final inclusion. Following these criteria, two reviewers (TL and NG) independently inspected the candidate publications for inclusion and quality. Discrepancies between the two reviewers were resolved through discussion with a senior reviewer (CL) and then, corrected and finalized. Finally, there were 30 studies selected for review (see Figure 1 for the study selection process).

XSL•FO
RenderX

Figure 1. PRISMA flowchart. AI: artificial intelligence; PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses. Note: Under "Reports excluded," exclusion reasons #1, #2, and #3 are not mutually exclusive.

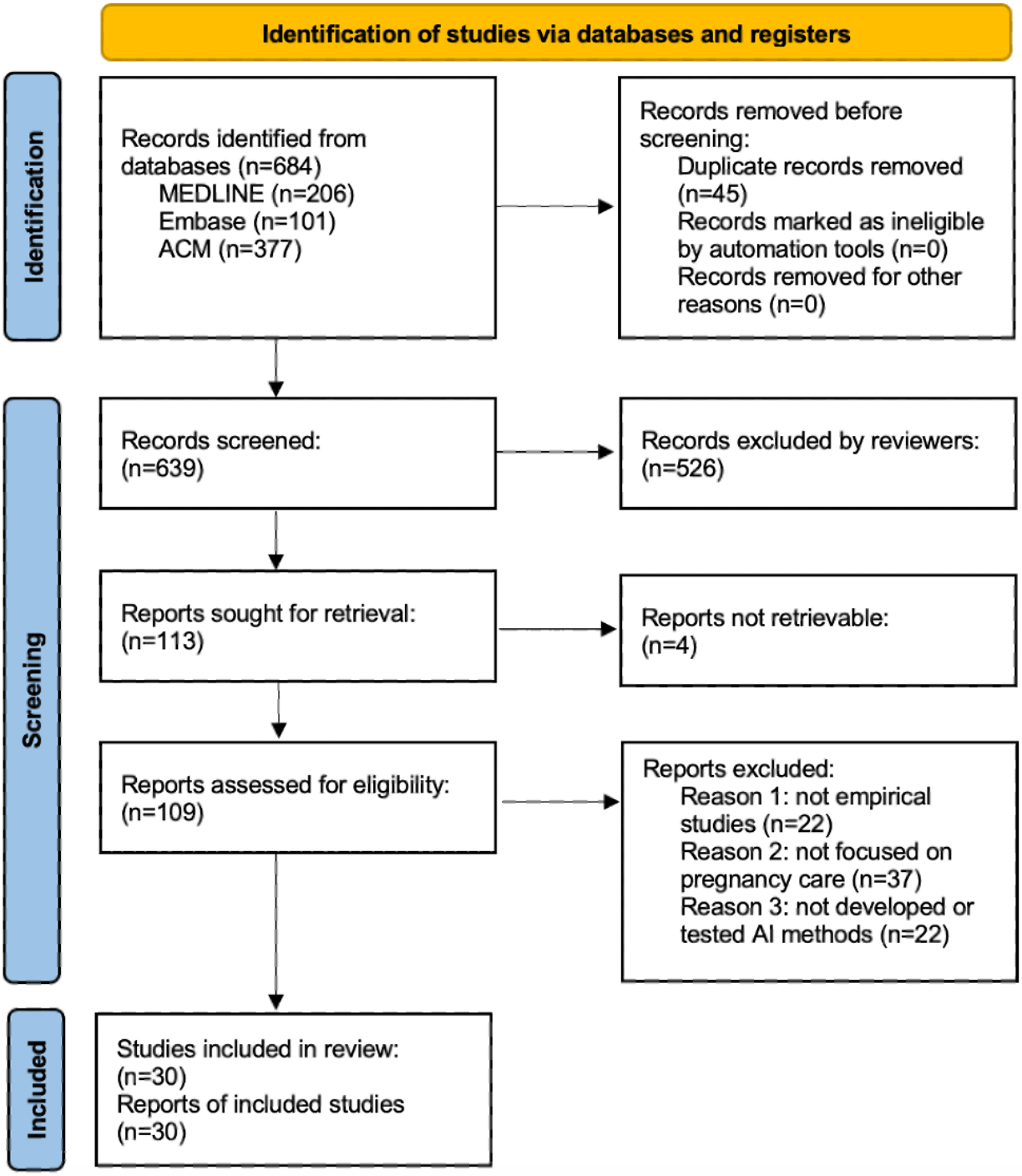

Data Synthesis

When authors reported ancillary information (eg, pilot study), or overlapping outcomes of a study reported from different publications, we grouped such publications as one single unit of publications. Two independent coders (CL and TL) extracted the following study information: authors and year, study objectives, pregnancy care applications, CDSS functionality, data source, study population, AI methods, CDSS performance, validation, and implementation. Pregnancy care applications included three categories: prenatal and early pregnancy care, obstetric care, and postpartum care. In the context of this review, obstetric complications include maternal (eg, perinatal hemorrhage, ectopic pregnancy, eclampsia, and gestational diabetes), fetal (eg, miscarriage, stillbirth, and preterm birth), and neonatal (eg, bradycardia and tachyarrhythmia) adverse events. In the context of this review, obstetric complications include maternal (eg, perinatal hemorrhage, ectopic pregnancy, eclampsia, and gestational diabetes), fetal (eg, miscarriage, stillbirth, and preterm birth), and neonatal (eg, bradycardia and tachyarrhythmia) adverse events. Of note, these three categories are not mutually exclusive. With respect to CDSS functionality, the definition of clinical prediction refers to the prediction of adverse clinical events, outcomes, prognosis, and identification

XSL•FO
RenderX

of at-risk individuals with adverse events. With respect to types of validation, internal validation refers to the process of validating the performance of CDSS models inside the context of the study, by which the context of the study means the model training set and testing set are partitioned from the same data set that has a high degree of homogeneity (eg, from the same clinical site and same patient cohort). Internal validation emphasizes on validity of model accuracy, as well as the sample size. External validation refers to the process of validating the performance of CDSS outside the context of the study, which empathizes with the generalizability of the internally validated CDSS to and across other contexts (eg, clinical sites, patient cohorts, times, and data quality).

Results

Study Selection and Synthesis of Results

We included 206 studies from PubMed/MEDLINE, 101 studies from Embase, and 377 studies from ACM Digital Library. Removal of studies hit by exclusion criteria and duplicates resulted in 30 distinct studies that met the eligibility criteria (Figure 1). We analyzed the 30 studies and characterized findings into structured themes, summarized in Table 1. Over time, the number of relevant studies has increased except for a dip in 2013-2014 (Figure 2).

XSL•FO
RenderX

Table 1. Summary of reviewed studies.

Study	Study objectives[a]	CDSS[b] functions	Data source	Sample	AI[c] methods	Performance	Validation	Implementation
Woolery and Grzymala-Busse (1994) [14]	Expert system for *preterm birth risk assessment*.	Risk prediction	Registry (multiple sites, United States)	18,890 cases	Expert system, machine learning	ACC[d] 53%-88%	External	No
Mongelli et al (1997) [15]	Develop an expert system for the interpretation of *fetal scalp acid-base status*.	Risk prediction	Scalp blood samples (single, England)	2174 samples	Logistic transformations, back-propagation networks, decision tree	N/A[e]	Internal	No
Goodwin et al (2000) [16]	Predict *preterm birth*.	Risk prediction	EHR[f] (single, United States)	19,970 patients	Rule induction, logistic regression, neural network	Customized (AUC[g] 0.75)	Internal	No
Catley et al (2006) [17]	*Obstetrical outcome estimations* in low-risk maternal populations.	Risk prediction	Registry (37 sites, Canada)	48,000 cases	ANN[h]	ROC[i] 0.73	Internal	No
Mueller et al (2006) [18]	Identify *predictors to optimizeextubation decisions for premature infants*.	Risk prediction	EHR (single, United States)	183 infants	ANN, multiple layer regression	AUC >0.9	Internal	Yes
Gorthi et al (2009) [19]	Predict *pregnancy risk* based on patterns from clinical parameters.	Risk prediction	Synthetic cases	200 cases	Decision tree	ACC 82.5	Internal	No
Ocak (2013) [20]	Assess *fetal well-being*.	Risk prediction	Cardiotocogram (single, United States)	1831 samples	SVM[j]	ACC 99.3%	Internal	No
Yılmaz and Kılıkçıer (2013) [21]	Determine the *fetal state* using cardiotocogram data.	Risk prediction	Cardiotocogram (single, United States)	2126 samples	LS-SVM[k]	ACC 91.62%	Internal	No
Spilka et al (2014) [22]	Examine cardiotocogram and support decision-making (*outcomes: diagnostics and risk*).	Diagnostic support	Cardiotocogram (single, United States)	634 samples	Latent class analysis	N/A	Internal	No
Jiménez-Serrano et al (2015) [23]	Detect the *postpartum depression* during 1st week after childbirth. Toward a mobile health app.	Risk prediction	Registry (7 sites, Spain)	1880 women	Logistic regression, naïve bayes, SVM, ANN	ANN (ACC 0.79)	Internal	Conceptual
Ravindran et al (2015) [24]	Assess *fetal well-being*.	Risk prediction	Cardiotocogram (single, United States)	2126 samples	Ensemble: k-NN[l], SVM, Bayesian network, and ELM[m]	ACC 93.61%	External	No

XSL•FO
RenderX

Study	Study objectives[a]	CDSS[b] functions	Data source	Sample	AI[c] methods	Performance	Validation	Implementation
Paydar et al (2017) [25]	Predict *pregnancy outcomes* among systemic lupus erythematosus-affected pregnant women.	Risk prediction	EHR (single, Iran)	149 pregnant women	MLP[n], RBF[o]	MLP (ACC 0.91)	Internal	Conceptual
Dhombres et al (2017) [26]	Develop a knowledge base for ectopic pregnancy.	Knowledge representation	Ultrasound (single, England)	4260 records	Ontology, NLP	Precision 0.83	Internal	No
Maurice et al (2017) [27]	Develop a new knowledge base intelligent system for ultrasound imaging.	Knowledge representation	PubMed (single, United Kingdom)	N/A	Ontology, NLP	F 0.71	Internal	No
Fergus et al (2018) [28]	Classify *cesarean section and vaginal delivery*.	Risk prediction	Registry (single, Czechia)	552 pregnancies	Ensemble: RF, SVM, decision tree, ANN, deferred acceptance	Ensemble (AUC 0.96)	Internal	No
Seitinger et al (2018) [29]	Arden Syntax as medical knowledge representation and processing language in obstetrics.	Knowledge representation	N/A	N/A	Arden syntax	N/A	N/A	No
De Ramón Fernández et al (2019) [30]	Develop a decision support system to make *suggestions for early treatment for ectopic pregnancy*.	Treatment recommendation	EHR (single, Spain)	406 tubal ectopic pregnancies	Multilayer perception, decision rule, SVM, Naïve Bayes	SVM (ACC 0.96)	Internal	No
Wang et al (2019) [31]	Develop a *postpartum depression* prediction model using EHR.	Risk prediction	EHR (single, United States)	179,980 pregnancies	Logistic regression, SVM, decision tree, Naïve Bayes, XGB[p], RF[q]	SVM (AUC 0.79)	Internal	No
Liu et al (2019) [32]	Predict *pregnancies*.	Diagnostic support	Mobile app	65,276 women	Logistic regression, LSTM[r]	AUC 0.67	External	No
Ye et al (2020) [33]	Predict *GDM*[s] and compare their performance with that of logistic regressions.	Risk prediction	EHR (single, China)	22,242 singlet pregnancies	Gradient Boosting Decision Tree, AdaBoost, LightGBM, logistic regression, voting, XGB, decision tree, RF, logistic regression[t]	GBDT[u] (AUC 0.74, 95% CI 0.71-0.76)	Internal	No
Silva et al (2020) [34]	Develop readable and minimal syntax for a web CDSS for antenatal care guidelines.	Knowledge representation	N/A	N/A	Ontology	N/A	No	No

XSL•FO
RenderX

Study	Study objectives[a]	CDSS[b] functions	Data source	Sample	AI[c] methods	Performance	Validation	Implementation
Venkatesh et al (2021) [35]	Predict *the risk of postpartum hemorrhage at labor admission*.	Risk prediction	EHR (Consortium on Safe Labor, United States)	228,438 deliveries	RF, XGB, logistic regression[t], lasso regression[t]	XGB (C statistic0.93; 95% CI 0.92-0.93)	External (multi-site, multi-time)	No
Tissot and Pedebos (2021) [36]	Test embedding strategies in performing risk assessment of miscarriage before or during pregnancy.	Risk prediction	EHR (InfoSaude, Brazil)	4676 pregnancies	Machine learning, ontology embedding	KRAL[v] (F 0.76)	Internal	No
Escobar et al (2021) [37]	Predict *risk of maternal, fetal, and neonatal events*.	Risk prediction	EHR (15 sites, United States)	303,678 deliveries	Gradient boosted, logistic regression[t]	Gradient boosted (AUC 0.786)	External	No
Tao et al (2021) [38]	Construct a hybrid *birth weight* predicting classifier.	Risk prediction	EHR (single, China)	5759 pregnant women	LSTM, CNN[w], RF, SVM, BPNN[x], logistic regression	Hybrid LSTM (MRE[y] 5.65 ± 0.4)	Internal	No
Mooney et al (2021) [39]	Examine RF to predict the *occurrence of hypoxic-ischemic encephalopathy*	Risk prediction	Registry (2 sites, Sweden)	53,000 deliveries	RF	RF (MCC[z] 0.63)	Internal	No
Du et al (2022) [40]	Predict *gestational diabetes mellitus*.	Risk prediction	Registry (single, Ireland)	565 women	XBG, AdaBoost, SVM, RF, logistic regression	SVM (AUC 0.79)	Internal	No
Schmidt et al (2022) [41]	Predict *adverse outcomes in patients with suspected preeclampsia*	Risk prediction	Ultrasound (single, Germany)	1647 patients	Gradient Boosting Decision Tree, RF	GBTree (AUC 0.81)	Internal	No
De Ramón Fernández et al (2022) [42]	Predict *mode of delivery*: cesarean section, eutocia vaginal delivery, instrumental vaginal delivery.	Risk prediction	Registry (single, Spain)	10,565 records	MLP, RF, SVM	ACC >90	Internal	No

XSL•FO
RenderX

Study	Study objectives[a]	CDSS[b] functions	Data source	Sample	AI[c] methods	Performance	Validation	Implementation
Hershey et al (2022) [43]	Predict *spontaneous preterm birth*.	Risk prediction	Surveys, biospecimen (10 centers)	2390 women	SVM	AUC 0.75	Internal	No

[a]The outcomes of a CDSS model are given in italics.
[b]CDSS: clinical decision support system.
[c]AI: artificial intelligence.
[d]ACC: accuracy.
[e]Not applicable.
[f]EHR: electronic health record.
[g]AUC: area under the receiver operating characteristic curve.
[h]ANN: artificial neural network.
[i]ROC: receiver operating characteristic curve.
[j]SVM: support vector machine.
[k]LS-SVM: least-squares support vector machine.
[l]k-NN: k-nearest neighbors.
[m]ELM: extreme learning machine.
[n]MLP: multilayer perceptron neural network.
[o]RBF: radial basis functions neural network.
[p]XGB: XGBoost.
[q]RF: random forest.
[r]LSTM: long-short term memory.
[s]GDM: gestational diabetes.
[t]Benchmark algorithm
[u]GBDT: gradient-boosted decision tree.
[v]KRAL: knowledge representation and artificial learning.
[w]CNN: convolutional neural network.
[x]BPNN: back propagation neural network.
[y]MRE: mean relative error.
[z]MCC: Matthew's correlation coefficient.

XSL•FO
RenderX

Figure 2. Trends in reviewed studies. Top-left: trends in studies by applications in pregnancy care. Top-right: trends in studies by CDSS function. Bottom-left: trends in studies by AI methods. Bottom-right: trends in studies by implementation. AI: artificial intelligence; CDSS: clinical decision support system.

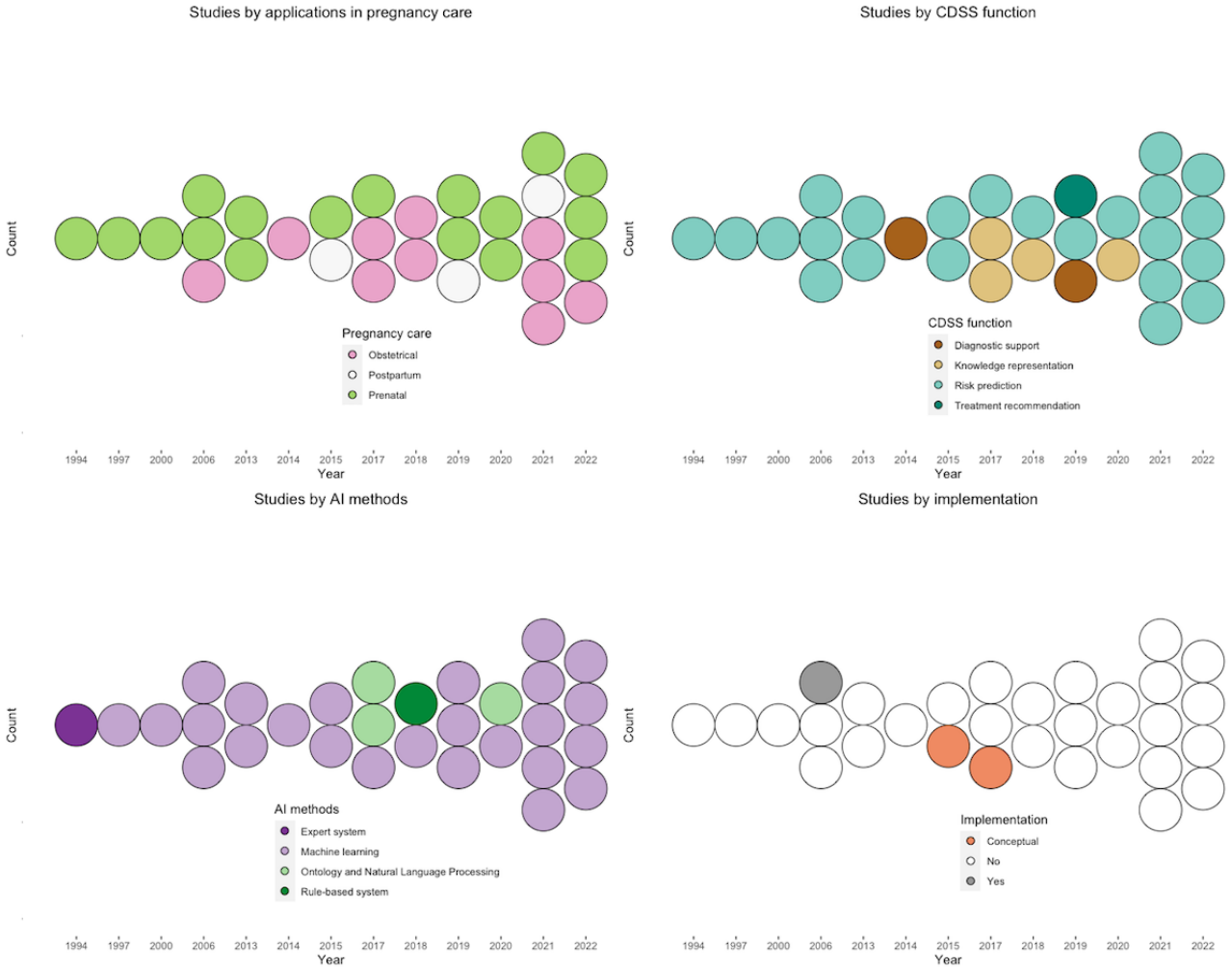

Risk of Bias of Included Studies

Among the 109 studies screened and retrieved, we excluded 22 studies that are not empirically based, 37 studies that pregnancy care is not the primary focus, and 22 studies that do not develop or apply AI methods even though AI-related terms are widely used in the publications. This appraisal process resulted in 30 included studies that have reached a full agreement of quality between the two reviewers (TL and NG) after discussing with the third reviewer (CL). See Figure 1 for the numbers of included and excluded studies from every step. Following the PRISMA guidelines for assessing the risk of bias of included studies, we summarized the assessment outcome in Figure 3.

XSL•FO
RenderX

Figure 3. Traffic light plot for risk-of-bias assessment of included studies.

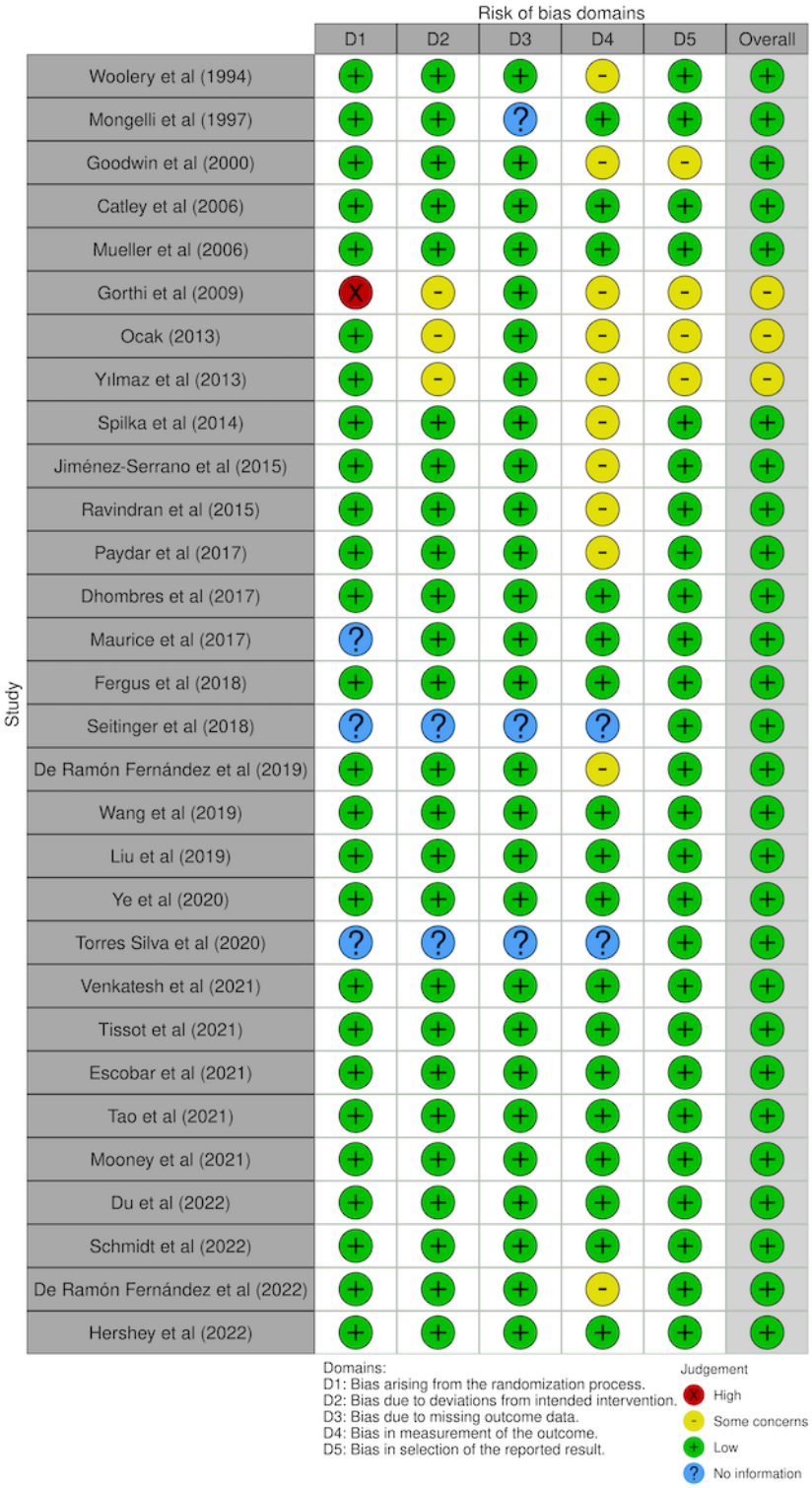

Study Characteristics: Applications in Pregnancy Care

Prenatal and Early Pregnancy Care

Detection of maternal and fetal risk factors and abnormalities during prenatal care is most imperative for timely prevention and intervention (n=17, 57%). Studies using CDSS include the prediction of gestational diabetes mellitus [33,40], miscarriage [25,36], and adverse outcomes resulting from preeclampsia [41] using data from medical history and prenatal care visits. Another use case of CDSS has applied to ectopic pregnancy which is a highly risky condition that often leads to maternal morbidity and mortality [30]. Upon diagnosis, choosing adequate treatment is an important clinical decision process to avoid further complications. Machine learning–based CDSS has been tested to aid providers and patients in better-informed clinical decision-making following ectopic pregnancy [30].

XSL•FO
RenderX

Obstetrical Care

As the number of delivering individuals who experience morbidity and mortality remains high and growing in the United States and many countries [44], an increasing volume of CDSS studies have been focusing on developing predictive models for early detection of adverse events and among at-risk individuals to be used for timely prevention and intervention (n=10, 33%). For example, the identification of individuals with risks of preterm birth can inform advanced medical care planning at the prenatal and perinatal stages [14,17,43]. These studies generally included analyses of risk factors contributing to adverse events, in which machine learning–based studies rely on feature ranking methods for the identification of data highly suggestive of adverse outcomes. Another example is computer-assisted cardiotocography (CTG) trace interpretation to be used before or at labor and delivery to assist decision-making [21,22,28].

Postpartum Care

CDSS has been used for estimating the risk of postpartum hemorrhage at labor and delivery admission (n=3, 30%). Postpartum hemorrhage has been a major source of maternal morbidity and mortality, accounting for nearly one-third of deaths of birthing individuals [45]. Risk estimation used in clinical practice has been based on stratification of risk factors documented in individuals' medical records using parametric statistic models. In recent CDSS studies, researchers have attempted to incorporate nuances beyond known risk factors [35], which may better interpret individual variance, and reduce possible biases from traditional guidelines and theoretical frameworks. Another example is risk assessment and screening of postpartum depression, which is a prevalent postpartum disorder but is often underdiagnosed [23,31].

Study Characteristics: Functionality of CDSSs

Diagnostic Support

Diagnostic support is a classic function known since early stage CDSS (n=2, 7%) [46]. In pregnancy care, this function has been used to assist the interpretation of CTG [21,22,28] because CTG interpretation is known to be challenging due to a great interreviewer variability and accurate interpretation of CTG is important for making proper clinical decisions during prenatal care and labor and delivery (eg, cesarean vs vaginal delivery). Diagnostic support has also been applied for the identification of pregnancy using data collected from mobile devices [32], which may have value for family planning and preventive care.

Clinical Risk Prediction

Clinical prediction as a CDSS function was not prevalent at the inception of CDSS and has started to emerge in pregnancy care recently (n=22, 73%). In pregnancy care, risk prediction tools were broadly used for early detection of adverse maternal and fetal events. Such CDSS applications have been centered around making the prediction of adverse events that may benefit from early detection of abnormalities for timely prevention and intervention, such as eclampsia or preeclampsia [37,41], gestational diabetes [33,40], preterm birth [14,17], miscarriage [25,36], perinatal hemorrhage [35,37,41], hypoxic-ischemic encephalopathy [39], low birth weight [38], and postpartum depression [23,31]. EHR and medical images were often used for training predictive models. A few studies used data from mobile apps [23,32].

Therapeutics Recommendation

Mode of delivery is one of the critical clinical decisions to make during obstetric care (n=2, 7%). Over the last decades, cesarean delivery has been increasing and was found to be associated with increased adverse fetal outcomes, as well as adverse maternal outcomes in subsequential childbirth deliveries [47]. CDSS studies have tested the feasibility of using machine learning to make suggestions out of three delivery modes: cesarean section, eutocic vaginal delivery, and instrumental vaginal delivery [42].

Knowledge Base

CDSS can be categorized as those built on top of a knowledge base and those independent of a knowledge base (n=4, 13%). Several studies reported the design and construction of a knowledge base that can underpin CDSS specialized for pregnancy care. Among these studies, forms of knowledge base include Arden syntax and ontology that have been widely used for formal representation of clinical guidelines and graph-based medical knowledge, as well as XML as a markdown language for web and mobile-based CDSS applications. Specifically, in support of diagnostics and therapeutics recommendations for ectopic pregnancy, ontology has been used for supporting the annotation of medical images (eg, ultrasound images for obstetrics) [26,27]. Arden syntax was used to formalize obstetric clinical guidelines into a knowledge base that supports CDSS functions for obstetrics [29]. XML was used for encoding a knowledge base that underpins mobile app–based CDSS for prenatal care [34].

Study Characteristics: AI Methodologies and Applications

Algorithms

Knowledge-base–independent CDSS typically rely on computational algorithms for learning about decision boundaries, whereas supervised algorithms (eg, classification, prediction, and association rules learning) require human-annotated data as a gold-standard sample whereas unsupervised algorithms (eg, clustering) find decision boundaries without a gold standard. In this review, regression-based algorithms were widely used as benchmark algorithms for clinical prediction tools, diagnostic support, and therapeutants recommendation. Some of the studies used parametric linear statistical models as benchmarks [35]. Among supervised machine learning algorithms, support vector machine, random forest, and gradient boosting algorithms (eg, XGBoost) have been increasingly adopted and have revealed outstanding performance. Simple neural networks (eg, multilayer perceptron and artificial neural networks) have been tested as well especially when the feature space of the model is not overly large in dimensionality and complex [17,18,30]. To incorporate domain-specific medical knowledge and human-curated clinical guidelines into machine learning models, the embedding of ontology was also used in the field of pregnancy care [36]. Additionally, there was the application of deep learning algorithms (eg, convolutional neural network and recurrent

XSL•FO
RenderX

neural network) in the field which has resulted in trained models outperforming other algorithms in comparison [38].

Knowledge-base–dependent CDSS typically rely on rules (eg, if-then and fuzzy logic) or semantic relations (eg, semantic properties defined by ontology). For example, natural language processing (NLP) tasks (eg, named entity recognition and semantic reasoning) were applied in conjunction with the ontology-based knowledge base for the annotation of medical images [26,27]. Studies of rule-based algorithm applications also demonstrated feasibility and robust clinical interpretability [16]. With respect to knowledge base design in these CDSS studies, ontology was commonly used to construct a knowledge base (n=2, 7%) [26,27].

Performance Evaluation

The majority of the studies (n=28, 93%) have tested internal validation to some degree. Validation frameworks used in reviewed studies included hold-out, n-fold cross-validation, and bootstrap or cross-validation. Evaluation metrics used include metrics derived from the theory of information retrieval such as precision, recall, *F* measure, area under the receiver operating characteristics curve [48]; metrics based on probabilistic statistics such as mean squared prediction error, Matthew's correlation coefficient, chi-square, and c-index; metrics based on descriptive statistics such as accuracy; and customized accuracy measures. A few studies used a validation design that allowed for the computing of confidence intervals [33,35], which enhanced the interpretative capability of validation. There were 5 (16%) studies that included external validation [14,24,32,35,37]. These studies generally tested CDSS on separate data sets including those from different clinical sites.

Treatment With Possible Bias

Biases in data sampling, data processing (eg, tackling missing data and data normalization), machine learning model training, validation, and algorithm design could lead to deviated performance and actionable clinical decisions resulting from CDSS. Arguably, because clinical decisions are driven by untrained data, ill-sampled training and validation data are seen to bias the AI-augmented systems, with which CDSS could be one example yet such an issue has not been addressed in the existing studies. Without recognizing or understanding biases in samples, appropriate imputation methods could also omit or amplify biases. Upon review, we did not find comprehensive treatment and discussion for remediating possible biases during the design and development of CDSS.

Study Characteristics: CDSS Implementation

Implementation of CDSS is the final step to incorporate research findings into routine practice. Only a few studies have discussed the conceptual ideas or pilot study design of clinical implementation of reported CDSS (n=3, 10%) [18,23,25]. For implementation studies in a clinical setting, web-based data entry and graphical result presentation were developed for CDSS implementation [18,25]. One study demonstrated the interface of CDSS based on an Android system [23]. We did not see a comprehensive CDSS implementation study design (eg, usability testing) among the reviewed studies.

Discussion

Principal Findings

Over the last decades, we have seen the proliferation of AI applications in clinical and translational medicine. However, AI-augmented CDSS has not been systematically reviewed in the field of obstetrics and gynecology. In this study, we assessed related studies by their health care applications, CDSS functionality, AI methodology, and clinical implementation with the goal of providing the state of the art of studies, as well as summarizing advantages, limitations, and possible future directions. Overall, we have identified 30 related studies published between 1994 and 2022 with an upward trend. There was a notable increase starting in 2021. All studies used data from EHR, registry, and mobile devices, except for those focusing on developing knowledge bases for CDSS. In the field of pregnancy care, functions of existing CDSS include diagnostic support (ie, imaging support), clinical prediction, therapeutics recommendation, and knowledge base. A list of traditional CDSS functions was not seen in the field of pregnancy care, including patient safety (eg, alarms for drug-drug reactions and allergies and computerized provider order entry support), clinical management (eg, point-of-care alters, info button, prompts for vaccination, outreach, and referral), and administrative management (eg, assisted medical coding and documentation) [49-51]. Architectures of CDSS include both knowledge-based dependent and independent, in which ontology remains a primary form for constructing a computerizable knowledge base and was often used jointly with NLP methods (eg, named-entity recognition and semantic reasoning). For CDSS that do not rely on a knowledge base, machine learning algorithms (primarily supervised algorithms including simple neural networks and deep learning) were widely used for learning from empirical data and producing or replicating actionable clinical knowledge. Existing studies confirmed that machine learning, ontology, and NLP have been increasingly applied for modern CDSS in the field of pregnancy care.

Clinical Implication

In the review of the potentials and challenges for adopting existing AI-augmented CDSS for pregnancy care, there are a few aspects. First, well-performed model in individual pregnancy episodes. Existing CDSS have been designed to assist prenatal care, obstetrics, and postpartum care. The majority of the CDSS studies in prenatal care are focused on assisting in the prediction of risks such as miscarriage [36], ectopic pregnancy [26,30], gestational diabetes [33,40], preterm birth [14,16,17], and severe maternal mortality and morbidity events during the prenatal episode [41]. Among predictive models of these CDSS studies, reported predictive performance is generally well. However, we noticed it is often misinterpreted in these studies with respect to how early before an adverse event a CDSS can reliably detect the risk and offer clinical decisions, which is an obvious obstacle before these CDSS models can be used for real-world practice. For obstetrics, CDSS studies have explored ways to assist in making choices of delivery mode [28,42], extubating decisions for preterm infants [18], and diagnostics during birth and delivery [22]. These applications

XSL•FO
RenderX

appear to have the potential to be tested, improved, and adopted in clinical settings. In postpartum care, CDSS applications have performed well in assisting postpartum hemorrhage [52] and depression risk detection [23,31]. When tested and adopted in the real-world scenario, effective data collection could be challenging because of limited patient encounters and access to mobile apps post partum.

Second, model interoperability. Aside from the model performance, interoperability is critically important as it determines the degree clinicians can interpret the model output and make sense of contributing factors and nuances a clinical decision is driven. The knowledge base of CDSS is the most interpretable because any knowledge piece is traceable and can be reasoned in different semantic logics in the knowledge base. Our review has found four studies that used biomedical ontology and semantic web techniques to develop the knowledge base for pregnancy risk, antenatal guidelines, ultrasound imaging, and ectopic pregnancy [26,27,34,36]. With respect to CDSS that function as predictive models, two categories were found among the reviewed studies. Parametric models, including regression models [16,18,23,31-33,37,38,40,52], decision trees [15,19,28,31,33,41], shallow neural networks [17,18,23-25,28,30,42], and expert systems [14,15] can easily reveal decision logic, determinants of a decision, as well as their odds to clinicians. Nonparametric models and deep neural networks [20,21,28,30,38,42,43] have the advantage of taking large and high-dimensional data but have limited capability of making the decision mechanism explicit for clinicians.

In addition to the performance and use-case scenario and clinical interpretability of the CDSS, we noticed that data availability and quality have been a pertinent hurdle for developing CDSS for pregnancy care. The reasons are in part unique for pregnancy care in that during prenatal care data are often generated from either a hospital, outpatient obstetrics group, or outpatient laboratory, where different laboratory technologies and protocols of data entry are the norm. Different initiation time and frequency of prenatal care is another cause of unevenly collected patient data for CDSS training, in which individuals with late initiation and low frequency are generally associated with worse outcomes [53], yet this inequality is underestimated in the experimental phase of CDSS design and testing leading to possible bias when adopting CDSS for real-world practice.

Strengths, Limitations, and Future Directions of Reviewed Studies

Overview

Our review has revealed several strengths of CDSS design in pregnancy care. However, because the application of AI-augmented CDSS in the field remains in its infancy, we also identified a handful of limitations followed by suggestions and future directions.

Toward Robustness in Internal Validation Design

The review of existing studies exhibited several strengths in study design. (1) Existing CDSS experiments were generally based on real-world EHR, registry, and mobile device data, in which several studies have used data from multiple clinical centers [14,17,23,37,39,43]. While the majority of studies have sufficient sample sizes, a few studies used relatively small samples (n=100~300, see Table 1). (2) The majority of the studies (n=19, 73%), excluding knowledge base studies, have tested multiple AI algorithms for comparison with a selected benchmark. (3) Cross-validation or hold-out methods were generally used for internal validation. (4) Majority of studies measured *F* score (including precision and recall), area under the receiver operating characteristics curve, and metrics derived from probabilistic statistics for model performance. A few studies used accuracy alone, which may not be sufficient for a fair performance validation. Overall, reported model performance is acceptable (see Table 1).

Clinical Plausibility

(1) The majority of studies explicitly stated the clinical use scenarios for the reported CDSS, where the capability of early detection of abnormalities, at-risk pregnancies, and risk factors was generally recognized as the core clinical significance of CDSS. Yet, due to the scarcity of prenatal data and difficulties in integrating longitudinal and cross-specialty medical records, early diagnosis and prediction have been limited by data. (2) Existing studies also reflected the challenges of diagnostics and therapeutics recommendations unique to pregnancy care. Such examples include the interpretation of CTG and choosing a delivery mode. However, medical interpretation, such as CTG, historically has shown great variance in interrater reliability, which warrants repeated evaluations for CDSS design. Additionally, there have been controversial discussions pertaining to how the choice of cesarean delivery would adversely affect the fetus and maternal outcomes for subsequential pregnancies [47]. No CDSS studies have considered fetal outcomes or maternal outcomes of subsequential pregnancies for choosing delivery mode in the present pregnancy, limiting the clinical value of this line of applications. With respect to clinical scenarios where AI-augmented CDSS has been applied, we noticed a dearth of research and testing sites in emergency care, which could be interesting to explore in the future.

Possible Biases

The use of CDSS could introduce biases on several occasions. (1) Similar to AI algorithms applied in general clinical practice, bias could be from sampling the data to be used for training and testing CDSS. (2) Racial and ethnic disparities have been well documented in a wide spectrum of maternal mortality and morbidity [54]. For example, using patients sampled from wealthy (or poor) neighborhoods to train a CDSS would introduce bias in their behavior or clinical decision prediction. To mediate, design and clinical implementation of CDSS used for pregnancy care should consider targeted populations and their social determinants of health (SDOH). American College of Obstetricians and Gynecologists has made recommendations for patient screenings to enhance the inclusion of SDOH, avoid stereotyping, acknowledge various forms of racial or ethnic discrimination, and improve clinical decision-making that addresses SDOH [55]. However, existing studies have not included the aforementioned considerations and strategies to reduce biases, which are warranted to be set as future directions.

XSL•FO
RenderX

External Validation and Implementation

Despite brief discussions around conceptual ideas, external validation and implementation were rare in the reviewed CDSS studies. Without external validation, the clinical usability of these CDSS when generalized to different patient cohorts, health care systems, and times remains undertested. Challenges for external validation and implementation include substandard interoperability of CDSS models and clinical information systems where a CDSS could be implemented. Additionally, the implementation of CDSS would require organizational commitment, localized workflow, usability testing, and staff training to be successful [56]. Because of the proliferation of machine learning for CDSS, the generalizability of machine learning models has become a new challenge. Future directions are suggested to address these identified knowledge gaps and challenges.

Limitations of This Study

This review has limitations. Our search strategy is mainly based on keywords and MeSH terms, which may not be comprehensive for capturing CDSS studies that did not explicitly use CDSS-related terminologies. Because the notion of CDSS is often loosely defined, we recognize it as a limitation to eligibility criteria in this review. Despite limitations, our results are timely and pertinent which could be used to guide evidence-based clinical practice and future directions for CDSS studies in the field of pregnancy care. This review study also adheres to the PRISMA guidelines and employed two independent reviewers for study selection and evaluation.

Conclusions

This review summarized state-of-the-art AI-augmented CDSS methods and applications in the field of pregnancy care. This review highlights the proliferation of machine learning-based clinical predictive models and computer-aided diagnostics and therapeutics with acceptable internal validity tested. Recent advances in this line of research include (1) CDSS design targeted for early diagnosis of prenatal abnormalities and early detection of at-risk pregnancies for timely prevention and intervention; (2) challenging medical image interpretation and decision-making that could use the assistance of CDSS; and (3) several knowledge bases needed for specific domains of pregnancy care including, but not limited to, image annotation, adverse events, and clinical guidelines. Future directions are suggested to address possible biases introduced by using AI and CDSS, comprehensive study on external validity and clinical implementation, and continued improvement of clinical plausibility of CDSS.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 623 KB - jmir_v26i1e54737_app1.pdf]

Multimedia Appendix 2
Search strings.
[DOCX File , 14 KB - jmir_v26i1e54737_app2.docx]

References

1. House W. White House blueprint for addressing the maternal health crisis. URL: https://www.whitehouse.gov/wp-content/uploads/2022/06/Maternal-Health-Blueprint.pdf [accessed 2024-08-21]
2. American College of Obstetricians and Gynecologists and the Society for Maternal-Fetal Medicine, Kilpatrick SK, Ecker JL. Severe maternal morbidity: screening and review. Am J Obstet Gynecol 2016;215(3):B17-B22. [doi: 10.1016/j.ajog.2016.07.050] [Medline: 27560600]
3. Clark SL, Saade GA, Tolcher MC, Belfort MA, Rouse DJ, Barton JR, et al. Gestational hypertension and "severe" disease: time for a change. Am J Obstet Gynecol 2023;228(5):547-552. [doi: 10.1016/j.ajog.2022.11.1280] [Medline: 36396107]
4. Clinical practice guidline. American College of Obstetricians and Gynecologists. 2022. URL: https://www.acog.org/clinical/clinical-guidance/clinical-practice-guideline [accessed 2023-02-04]
5. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning learning health system. J Am Med Inform Assoc 2015;22(1):43-50 [FREE Full text] [doi: 10.1136/amiajnl-2014-002977] [Medline: 25342177]
6. Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med 2012;157(3):207-210. [doi: 10.7326/0003-4819-157-3-201208070-00012] [Medline: 22868839]
7. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc 2022;29(4):585-591 [FREE Full text] [doi: 10.1093/jamia/ocac006] [Medline: 35190824]

XSL•FO
RenderX

8. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
9. Abuelezz I, Hassan A, Jaber BA, Sharique M, Abd-Alrazaq A, Househ M, et al. Contribution of artificial intelligence in pregnancy: a scoping review. Stud Health Technol Inform 2022;289:333-336. [doi: 10.3233/SHTI210927] [Medline: 35062160]
10. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of artificial intelligence reported in obstetrics and gynecology journals: systematic review. J Med Internet Res 2022;24(4):e35465 [FREE Full text] [doi: 10.2196/35465] [Medline: 35297766]
11. Davidson L, Boland MR. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. Brief Bioinform 2021;22(5):bbaa369 [FREE Full text] [doi: 10.1093/bib/bbaa369] [Medline: 33406530]
12. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009;62(10):e1-34 [FREE Full text] [doi: 10.1016/j.jclinepi.2009.06.006] [Medline: 19631507]
13. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
14. Woolery LK, Grzymala-Busse J. Machine learning for an expert system to predict preterm birth risk. J Am Med Inform Assoc 1994;1(6):439-446 [FREE Full text] [doi: 10.1136/jamia.1994.95153433] [Medline: 7850569]
15. Mongelli M, Chang A, Sahota D. The development of a hybrid expert system for the interpretation of fetal acid-base status. Int J Med Inform 1997;44(2):135-144. [doi: 10.1016/S1386-5056(96)01244-0] [Medline: 9291005]
16. Goodwin L, Maher S, Ohno-Machado L, Iannacchione MA, Crockett P, Dreiseitl S, et al. Building knowledge in a complex preterm birth problem domain. Proc AMIA Symp 2000:305-309 [FREE Full text] [Medline: 11079894]
17. Catley C, Frize M, Walker CR, Petriu DC. Predicting high-risk preterm birth using artificial neural networks. IEEE Trans Inf Technol Biomed 2006;10(3):540-549. [doi: 10.1109/titb.2006.872069] [Medline: 16871723]
18. Mueller M, Wagner CL, Annibale DJ, Knapp RG, Hulsey TC, Almeida JS. Parameter selection for and implementation of a web-based decision-support tool to predict extubation outcome in premature infants. BMC Med Inform Decis Mak 2006;6:11 [FREE Full text] [doi: 10.1186/1472-6947-6-11] [Medline: 16509967]
19. Gorthi A, Firtion C, Vepa J. Automated risk assessment tool for pregnancy care. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:6222-6225. [doi: 10.1109/IEMBS.2009.5334644] [Medline: 19965084]
20. Ocak H. A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being. J Med Syst 2013;37(2):9913. [doi: 10.1007/s10916-012-9913-4] [Medline: 23321973]
21. Yılmaz E, Kılıkçıer C. Determination of fetal state from cardiotocogram using LS-SVM with particle swarm optimization and binary decision tree. Comput Math Methods Med 2013;2013:487179 [FREE Full text] [doi: 10.1155/2013/487179] [Medline: 24288574]
22. Spilka J, Chudáček V, Janků P, Hruban L, Burša M, Huptych M, et al. Analysis of obstetricians' decision making on CTG recordings. J Biomed Inform 2014;51:72-79 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.010] [Medline: 24747355]
23. Jiménez-Serrano S, Tortajada S, García-Gómez JM. A mobile health application to predict postpartum depression based on machine learning. Telemed J E Health 2015;21(7):567-574. [doi: 10.1089/tmj.2014.0113] [Medline: 25734829]
24. Ravindran S, Jambek AB, Muthusamy H, Neoh SC. A novel clinical decision support system using improved adaptive genetic algorithm for the assessment of fetal well-being. Comput Math Methods Med 2015;2015:283532 [FREE Full text] [doi: 10.1155/2015/283532] [Medline: 25793009]
25. Paydar K, Niakan Kalhori SR, Akbarian M, Sheikhtaheri A. A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus. Int J Med Inform 2017;97:239-246. [doi: 10.1016/j.ijmedinf.2016.10.018] [Medline: 27919382]
26. Dhombres F, Maurice P, Friszer S, Guilbaud L, Lelong N, Khoshnood B, et al. Developing a knowledge base to support the annotation of ultrasound images of ectopic pregnancy. J Biomed Semantics 2017;8(1):1-13 [FREE Full text] [doi: 10.1186/s13326-017-0117-1] [Medline: 28137311]
27. Maurice P, Dhombres F, Blondiaux E, Friszer S, Guilbaud L, Lelong N, et al. Towards ontology-based decision support systems for complex ultrasound diagnosis in obstetrics and gynecology. J Gynecol Obstet Hum Reprod 2017;46(5):423-429. [doi: 10.1016/j.jogoh.2017.03.004] [Medline: 28934086]
28. Fergus P, Selvaraj M, Chalmers C. Machine learning ensemble modelling to classify caesarean section and vaginal delivery types using cardiotocography traces. Comput Biol Med 2018;93:7-16. [doi: 10.1016/j.compbiomed.2017.12.002] [Medline: 29248699]
29. Seitinger A, Rappelsberger A, Leitich H, Binder M, Adlassnig KP. Executable medical guidelines with arden syntax-applications in dermatology and obstetrics. Artif Intell Med 2018;92:71-81. [doi: 10.1016/j.artmed.2016.08.003] [Medline: 27686851]
30. De Ramón Fernández A, Fernández DR, Sánchez MTP. A decision support system for predicting the treatment of ectopic pregnancies. Int J Med Inform 2019;129:198-204. [doi: 10.1016/j.ijmedinf.2019.06.002] [Medline: 31445255]

XSL•FO
RenderX

31. Wang S, Pathak J, Zhang Y. Using electronic health records and machine learning to predict postpartum depression. In: Studies in Health Technology and Informatics. Amsterdam: IOS Press; Aug 21, 2019:888-892.
32. Liu B, Shi S, Wu Y, Thomas D, Symul L, Pierson E, et al. Predicting pregnancy using large-scale data from a women's health tracking mobile application. Proc Int World Wide Web Conf 2019;2019:2999-3005 [FREE Full text] [doi: 10.1145/3308558.3313512] [Medline: 31538145]
33. Ye Y, Xiong Y, Zhou Q, Wu J, Li X, Xiao X. Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. J Diabetes Res 2020;2020:4168340 [FREE Full text] [doi: 10.1155/2020/4168340] [Medline: 32626780]
34. Silva EAT, Uribe S, Smith J, Gomez IFL, Florez-Arango JF. XML data and knowledge-encoding structure for a web-based and mobile antenatal clinical decision support system: development study. JMIR Form Res 2020;4(10):e17512 [FREE Full text] [doi: 10.2196/17512] [Medline: 33064087]
35. Venkatesh KK, Strauss RA, Grotegut CA, Heine RP, Chescheir NC, Stringer JSA, et al. Machine learning and statistical models to predict postpartum hemorrhage. Obstet Gynecol 2020;135(4):935-944 [FREE Full text] [doi: 10.1097/AOG.0000000000003759] [Medline: 32168227]
36. Tissot HC, Pedebos LA. Improving risk assessment of miscarriage during pregnancy with knowledge graph embeddings. J Healthc Inform Res 2021;5(4):359-381 [FREE Full text] [doi: 10.1007/s41666-021-00096-6] [Medline: 35419509]
37. Escobar GJ, Soltesz L, Schuler A, Niki H, Malenica I, Lee C. Prediction of obstetrical and fetal complications using automated electronic health record data. Am J Obstet Gynecol 2021;224(2):137-147.e7. [doi: 10.1016/j.ajog.2020.10.030] [Medline: 33098815]
38. Tao J, Yuan Z, Sun L, Yu K, Zhang Z. Fetal birthweight prediction with measured data by a temporal machine learning method. BMC Med Inform Decis Mak 2021;21(1):26 [FREE Full text] [doi: 10.1186/s12911-021-01388-y] [Medline: 33494752]
39. Mooney C, O'Boyle D, Finder M, Hallberg B, Walsh BH, Henshall DC, et al. Predictive modelling of hypoxic ischaemic encephalopathy risk following perinatal asphyxia. Heliyon 2021;7(7):e07411 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e07411] [Medline: 34278022]
40. Du Y, Rafferty AR, McAuliffe FM, Wei L, Mooney C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Sci Rep 2022;12(1):1170 [FREE Full text] [doi: 10.1038/s41598-022-05112-2] [Medline: 35064173]
41. Schmidt LJ, Rieger O, Neznansky M, Hackelöer M, Dröge LA, Henrich W, et al. A machine-learning-based algorithm improves prediction of preeclampsia-associated adverse outcomes. Am J Obstet Gynecol 2022;227(1):77.e1-77.e30. [doi: 10.1016/j.ajog.2022.01.026] [Medline: 35114187]
42. De Ramón Fernández A, Fernández DR, Sánchez MTP. Prediction of the mode of delivery using artificial intelligence algorithms. Comput Methods Programs Biomed 2022;219:106740 [FREE Full text] [doi: 10.1016/j.cmpb.2022.106740] [Medline: 35338883]
43. Hershey M, Burris HH, Cereceda D, Nataraj C. Predicting the risk of spontaneous premature births using clinical data and machine learning. Inform Med Unlocked 2022;32:101053. [doi: 10.1016/j.imu.2022.101053]
44. Hoyert DL. Maternal mortality rates in the United States, 2021. 2022. URL: https://www.cdc.gov/nchs/data/hestat/maternal-mortality/2021/maternal-mortality-rates-2021.htm [accessed 2023-03-01]
45. Say L, Chou D, Gemmill A, Tunçalp Ö, Moller AB, Daniels J, et al. Global causes of maternal death: a WHO systematic analysis. Lancet Glob Health 2014;2(6):e323-e333 [FREE Full text] [doi: 10.1016/S2214-109X(14)70227-X] [Medline: 25103301]
46. Clancey WJ, Shortliffe EH. Readings in Medical Artificial Intelligence: The First Decade. United States: Addison-Wesley Longman Publishing Co., Inc; 1984.
47. Mengesha MB, Adhanu HH, Weldegeorges DA, Assefa NE, Werid WM, Weldemariam MG, et al. Maternal and fetal outcomes of cesarean delivery and factors associated with its unfavorable management outcomes; in Ayder specialized comprehensive hospital, Mekelle, Tigray, Ethiopia, 2017. BMC Res Notes 2019;12(1):650 [FREE Full text] [doi: 10.1186/s13104-019-4690-5] [Medline: 31590693]
48. Manning CD, Raghavan P, Schutze H. Introduction to Information Retrieval. England: Cambridge University Press; 2008.
49. Menachemi N, Brooks RG. Reviewing the benefits and costs of electronic health records and associated patient safety technologies. J Med Syst 2006;30(3):159-168. [doi: 10.1007/s10916-005-7988-x] [Medline: 16848129]
50. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
51. Kaplan B. Evaluating informatics applications—clinical decision support systems literature review. Int J Med Inform 2001;64(1):15-37. [doi: 10.1016/s1386-5056(01)00183-6] [Medline: 11673100]
52. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med 2021;4(1):147 [FREE Full text] [doi: 10.1038/s41746-021-00518-0] [Medline: 34635760]

XSL•FO
RenderX

53. McDuffie RS, Beck A, Bischoff K, Cross J, Orleans M. Effect of frequency of prenatal care visits on perinatal outcome among low-risk women: a randomized controlled trial. JAMA 1996;275(11):847-851. [doi: 10.1001/jama.1996.03530350029030]
54. Racial and ethnic disparities continue in pregnancy-related deaths. Centers for Disease Control and Prevention. 2019. URL: https://archive.cdc.gov/www_cdc_gov/media/releases/2019/p0905-racial-ethnic-disparities-pregnancy-deaths.html [accessed 2019-09-05]
55. Importance of social determinants of health and cultural awareness in the delivery of reproductive health care. AJOG Committee on Health Care for Underserved Women. URL: https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2018/01/importance-of-social-determinants-of-health-and-cultural-awareness-in-the-delivery-of-reproductive-health-care [accessed 2023-06-12]
56. Campion TR, Waitman LR, May AK, Ozdas A, Lorenzi NM, Gadd CS. Social, organizational, and contextual characteristics of clinical decision support systems for intensive insulin therapy: a literature review and case study. Int J Med Inform 2010;79(1):31-43 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.09.004] [Medline: 19815452]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support systems
CTG: cardiotocography
EHR: electronic health record
MeSH: Medical Subject Headings
NLP: natural language processing
PMC: PubMed Central
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SDOH: social determinants of health

Edited by A Mavragani; submitted 20.11.23; peer-reviewed by S Kommireddy, N Fareed, G Carot-Sans, M Galani; comments to author 06.04.24; revised version received 06.05.24; accepted 24.07.24; published 16.09.24.

Please cite as:
Lin X, Liang C, Liu J, Lyu T, Ghumman N, Campbell B
Artificial Intelligence–Augmented Clinical Decision Support Systems for Pregnancy Care: Systematic Review
J Med Internet Res 2024;26:e54737
URL: https://www.jmir.org/2024/1/e54737
doi:10.2196/54737
PMID:

©Xinnian Lin, Chen Liang, Jihong Liu, Tianchu Lyu, Nadia Ghumman, Berry Campbell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Smartphone-Based Hand Function Assessment: Systematic Review

Yan Fu[1], PhD; Yuxin Zhang[1], MSc; Bing Ye[2,3], MSc; Jessica Babineau[4], MLis; Yan Zhao[5], MSc; Zhengke Gao[1], MSc; Alex Mihailidis[2,3], PhD, PEng

[1]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China

[2]KITE - Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada

[3]Department of Occupational Science and Occupational Therapy, University of Toronto, Toronto, ON, Canada

[4]Library and Information Services, University Health Network, Toronto, ON, Canada

[5]Department of Rehabilitation Medicine, Hubei Province Academy of Traditional Chinese Medicine Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, China

Corresponding Author:
Yuxin Zhang, MSc
School of Mechanical Science and Engineering
Huazhong University of Science and Technology
1037 Luoyu Road, Hongshan District
Wuhan, 430074
China
Phone: 86 17381571416
Email: 953385493@qq.com

Abstract

Background: Hand function assessment heavily relies on specific task scenarios, making it challenging to ensure validity and reliability. In addition, the wide range of assessment tools, limited and expensive data recording, and analysis systems further aggravate the issue. However, smartphones provide a promising opportunity to address these challenges. Thus, the built-in, high-efficiency sensors in smartphones can be used as effective tools for hand function assessment.

Objective: This review aims to evaluate existing studies on hand function evaluation using smartphones.

Methods: An information specialist searched 8 databases on June 8, 2023. The search criteria included two major concepts: (1) smartphone or mobile phone or mHealth and (2) hand function or function assessment. Searches were limited to human studies in the English language and excluded conference proceedings and trial register records. Two reviewers independently screened all studies, with a third reviewer involved in resolving discrepancies. The included studies were rated according to the Mixed Methods Appraisal Tool. One reviewer extracted data on publication, demographics, hand function types, sensors used for hand function assessment, and statistical or machine learning (ML) methods. Accuracy was checked by another reviewer. The data were synthesized and tabulated based on each of the research questions.

Results: In total, 46 studies were included. Overall, 11 types of hand dysfunction–related problems were identified, such as Parkinson disease, wrist injury, stroke, and hand injury, and 6 types of hand dysfunctions were found, namely an abnormal range of motion, tremors, bradykinesia, the decline of fine motor skills, hypokinesia, and nonspecific dysfunction related to hand arthritis. Among all built-in smartphone sensors, the accelerometer was the most used, followed by the smartphone camera. Most studies used statistical methods for data processing, whereas ML algorithms were applied for disease detection, disease severity evaluation, disease prediction, and feature aggregation.

Conclusions: This systematic review highlights the potential of smartphone-based hand function assessment. The review suggests that a smartphone is a promising tool for hand function evaluation. ML is a conducive method to classify levels of hand dysfunction. Future research could (1) explore a gold standard for smartphone-based hand function assessment and (2) take advantage of smartphones' multiple built-in sensors to assess hand function comprehensively, focus on developing ML methods for processing collected smartphone data, and focus on real-time assessment during rehabilitation training. The limitations of the research are 2-fold. First, the nascent nature of smartphone-based hand function assessment led to limited relevant literature, affecting the evidence's completeness and comprehensiveness. This can hinder supporting viewpoints and drawing conclusions.

XSL•FO
RenderX

Second, literature quality varies due to the exploratory nature of the topic, with potential inconsistencies and a lack of high-quality reference studies and meta-analyses.

(J Med Internet Res 2024;26:e51564) doi:10.2196/51564

KEYWORDS

hand function assessment; smartphone-based sensing; rehabilitation; digital health; mobile health; mHealth; mobile phone

Introduction

Background

Hand function assessment is crucial in determining the extent of functional loss in patients and the outcome of surgical and rehabilitative procedures. Subtle changes in hand function could be a good predictor for the early detection of certain neuromuscular degeneration diseases, such as Parkinson disease (PD), which could help take preventive measures to reduce the severity of the illness [1]. However, most current hand function assessments are conducted in a clinical context with the intensive involvement of rehabilitation professionals. Clinical evaluation requires frequent visits and long-duration treatment sessions [2]. Hand function is usually assessed using standard questionnaires, such as the Michigan Hand Outcome Questionnaire and Disability of the Arm, Shoulder, and Hand Index [3]. These measurements are subjective and could result in different assessment results across different test scenarios and medical professionals [4]. Clinical outcomes based on a rating scale are often insensitive to subtle hand function changes and do not support the provision of timely feedback [5]. As such, a hand assessment tool that can overcome the clinical assessment drawbacks of inconvenience, high cost, and imprecision [1,5] and automatically evaluate hand function over time would benefit patients.

Smartphones are equipped with advanced technologies, such as touchscreens, accelerometers, and gyroscopes, which can be used for measuring and evaluating hand function [6]. The application of smartphones in clinical hand dysfunction assessments can exploit built-in sensors (such as accelerometers and gyroscopes) to collect real-time hand movement data with convenience and at low cost [7]. Smartphones can precisely monitor and analyze a patient's hand condition for dysfunction assessment using machine learning (ML) and artificial intelligence algorithms [8]. Moreover, the smartphone-based hand dysfunction assessment can be designed according to clinical criteria to improve the system's reliability and validity [9-11]. Despite recent advances in smartphone-based hand function assessment [12,13], no systematic reviews have been conducted to provide a holistic perspective on how smartphones can be applied to hand function assessment.

Although other technologies, such as wrist-worn or finger-worn sensors, smartwatches, and specialized keyboards, also show potential for automated hand function assessment, they typically focus on simple physiological data collection with limited data processing capabilities and display of basic information [14-16]. However, smartphones offer more extensive data acquisition, accurate data processing, and richer data display options, providing a more comprehensive technological solution [17,18]. Moreover, considering the widespread availability and user-friendly nature of smartphones [19], directing research efforts toward smartphone-centric studies can enhance innovation and application possibilities. This approach not only aligns with the current prevalence of smartphones but also extends a broader scope for future technology transfer and development specific to hand function assessment. Therefore, focusing on smartphone research can lead to more innovation and application possibilities, offering a broader scope for future technology transfer and development. As such, the main goal of this review was to synthesize the present ways in which smartphones are applied in hand function assessment and the extent to which hand function evaluation is achieved using smartphones. It aimed to explore the system development guidelines for the future application of smartphones in hand function assessment.

Research Questions

The research questions (RQs) were as follows: (1) What types of hand dysfunctions are studied, and what assessment inventory tools are used? (2) How are smartphones applied in clinical practice in hand function assessment? (3) What sensors are integrated into smartphones to collect hand function data? (4) What statistics or ML algorithms are used for hand function assessment?

Methods

This systematic review is reported according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1).

Information Sources and Search Strategy

An information specialist (JB) developed and executed a comprehensive search strategy. The following electronic databases were searched: MEDLINE(R) ALL (Ovid), Embase and Embase Classic (Ovid), CENTRAL (Ovid), Scopus, Compendex (Engineering Village), INSPEC (Engineering Village), IEEE Xplore, and ACM Digital Library. The search strategy was first developed in MEDLINE ALL (Ovid) in consultation with the research team. Search terms were also sourced from a previously published review [20]. The search strategy was then adapted into other databases.

Search strategies included the use of text words and subject headings related to two major concepts: (1) smartphone or mobile phone or mHealth and (2) hand function or function assessment. Searches were limited to English-language papers. When possible, searches were also limited to human studies and excluded conference proceedings and trial register records. No date limits were applied. All searches were conducted on June 8, 2023. The complete search strategies for each database are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Study Selection

The studies were imported into Covidence (Veritas Health Innovation) after eliminating duplicates using EndNote (Clarivate). Title and abstract screening and full-text screening were completed by 2 researchers (YZ and YF) independently based on the same inclusion and exclusion criteria. Any disagreement was first discussed and solved by the 2 researchers. Otherwise, a third researcher (BY) was involved to ensure that an agreement was reached.

Neurocognition is evaluated as an independent criterion in clinical hand assessments [21]. Therefore, neurocognitive studies were excluded from this review to focus specifically on aspects related to hand motor control and dysfunction. Although cognitive functions play a significant role in hand motor control, the primary aim of this review was to narrow its scope and focus on the specific factors directly related to the mechanics and dysfunction of the hand, with particular focus on methods and techniques for using smartphones in assessment. Neurocognitive research often involves specialized equipment and methods, for example, neuroimaging techniques such as functional magnetic resonance imaging or electroencephalogram, which may not be practical for assessing hand function in smartphone-related contexts.

After the screening stage, the research quality of selected studies was evaluated using the Mixed Methods Appraisal Tool, a tool designed for the systematic mixed research review evaluation phase [22]. The quality assessment was completed by one researcher and checked by another researcher. A conflict that arose regarding the assessment was discussed between the 2 researchers, and an agreement was reached.

The inclusion and exclusion criteria used for the screening process are presented in Textbox 1.

Textbox 1. The inclusion and exclusion criteria used for the screening process.

Inclusion criteria

- Technology: using smartphone sensors
- Study focus: hand function screening, including hand movement assessment and hand performance measurement
- Clinical assessment: measurement of motor function–related criteria, such as grip strength, posture, and degree of freedom
- Study design: peer-reviewed academic studies
- Language: English

Population: human participants

Exclusion criteria

- Technology: not using a smartphone for hand function assessment
- Study focus: health management and neurocognitive studies
- Clinical assessment: qualitative, non–peer-reviewed, and nonacademic studies
- Study design: systematic reviews, literature reviews, case reports, and letters
- Language: non-English

Population: nonhuman participants

Results

Overview

A total of 8898 records were retrieved from the search. After removing duplicates, 64.31% (5722/8898) of the records were filtered at the title and abstract screening stage. After title and abstract screening, 97.68% (5589/5722) of the records were removed. The remaining 2.32% (133/5722) of the records underwent full-text screening. A total of 46 studies were included after both screening stages and included in the final review. Figure 1 presents the PRISMA [23] flow diagram. Multimedia Appendix 3 [6,9-11,24-58] details the results of the evaluation of included studies based on the Mixed Methods Appraisal Tool. All 46 studies were published after 2012, and 67% (n=31) of them were published between 2017 and 2023.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the screening process for papers included in this study.

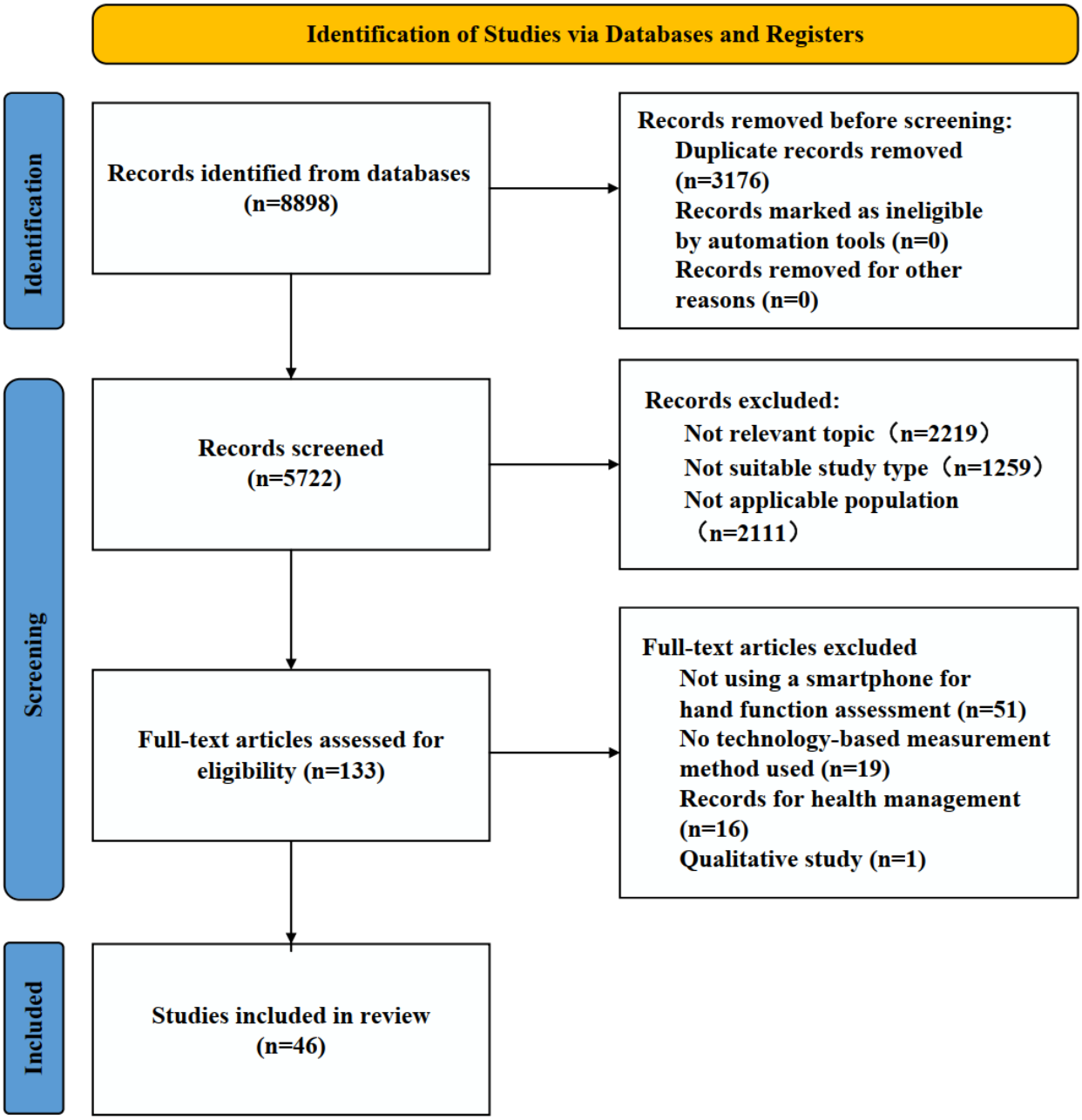

Study Characteristics

Of the 46 studies, 14 (30%) recruited participants with hand dysfunction, 7 (15%) included only healthy participants, and 23 (50%) recruited both types of participants (Table 1). The summarized smartphone specification is shown in Table 2. The age range was 21 to 91 years for patients with hand dysfunction and 17 to 81 years for healthy participants; the sample size varied from 1 to 1815.

XSL•FO
RenderX

Table 1. Characteristics of the studies (n=46).

Characteristic	References
Participants	
Patients only	[9,24-36]
Healthy participants only	[37-41,59,60]
Patients and healthy participants	[6,10,11,29,42-55,61-65]
—[a]	[56,57]
Sex	
Male only	[25,37]
Female only	[58]
Male and female	[6,9-11,26-28,30,31,33-35,39,40,42,43,45,48,49,52,53,55,56,59-65]
—	[24,29,32,36,38,41,44,46,47,50,51,54,57]
Study design	
Quantitative descriptive study	[9,24-29,31-42,45,46,51,56,57,59,60]
Observation study	[43,44,47,48,52,62,64,66]
Nonrandomized study	[6,10,11,25,30,38,43,44,47-50,52-55,58,61-65]
Case-control study	[58]
Study duration	
0-4 minutes	[10,28,29,31,39,58,64]
10 minutes	[59]
1.5 hours	[48]
10 hours	[61]
1-4 weeks	[9,26,42,52]
6-12 weeks	[51,63,66]
—	[11,24,25,27,30,32-38,40,41,43-47,49,50,52-57,60,62,65]
Sample size distribution	
0-32	[10,24-26,31-33,37-39,42,45,52,57,58,60,64]
33-64	[27,28,30,35,40,41,46,47,49,53,55,62,65]
65-95	[11,34,51,61,66]
96-126	[9,44,48,56]
127-189	[43,59]
190-220	—
221-252	[29]
253-598	—
599-629	[63]
630-1851	[36,50,54]

[a]Not applicable.

XSL•FO
RenderX

Table 2. Summary of smartphone specification.

Study, year	Processing power	Operating system	Smartphone type	Sensor sampling rate	Camera resolution
Matera et al [26], 2016	__[a]	Android	Nuans Neo Reloaded and HUAWEI GR5	—	—
Miyake et al [24], 2020	1.2 GHz dual-core processor	—	—	Accelerometer (range +2 to –2 g, 100 Hz)	—
García-Magariño et al [42], 2016	—	Android	Samsung Galaxy Trend Plus	—	—
Bercht et al [25], 2012	—	iOS and Android 4.4.2	iPhone 4S，Samsung Galaxy S4, and Google Nexus 5	—	—
Janarthanan et al [39], 2020	—	Android	LG Optimus G smartphone	—	—
Pan et al [28], 2015	—	iOS	iPhone	Accelerometer (100 Hz)	—
Orozco-Arroyave et al [61], 2020	—	Android	Android smartphone	Accelerometer (100 Hz)	—
Sarwat et al [32], 2021	—	—	—	—	—
Kostikis et al [10], 2015	—	Android	—	Accelerometer and gyroscope (20 Hz)	—
Lee et al [43], 2016	—	Android	Galaxy S3 mini and Android phone	—	—
Lipsmeier et al [6], 2018	—	Android	Tablet	Accelerometer and gyroscope (+66.6 to –10 Hz), magnetometer (+66.6 to –7 Hz), and microphone (44.1 kHz)	—
Sandison et al [45], 2020	—	—	—	—	—
Halic et al [46], 2014	—	iOS	iPhone 5	—	—
Koyama et al [30], 2021	—	—	—	—	—
Chén et al [51], 2020	—	iOS	iPhone 4	—	—
Arroyo-Gallego et al [62], 2017	—	Android 7.0	Huawei P9 Plus	Custom screen keyboard (1.2 GHz)	—
Pratap et al [63], 2020	—	—	Huawei Mate 9 Pro smartphone	—	—
Waddell et al [64], 2021	—	—	—	App touchscreen, accelerometer, and gyroscope (50 Hz)	—
Mousavi et al [56], 2020	—	Android 4.0	—	Mobile accelerometer software (100 Hz)	—
Lee et al [55], 2016	—	—	—	—	—
Hidayat et al [58], 2015	—	—	Huawei P10 Lite	—	—
Wang et al [37], 2016	—	—	—	—	—
Lee et al [38], 2018	—	Android	—	—	—
Iakovakis et al [44], 2019	—	iOS	iPhone XS Max	—	—
Modest et al [47], 2019	—	iOS	iPhone XS Max	—	—
Lendner et al [59], 2019	—	iOS	iPhone	—	—
Tian et al [48], 2019	—	Android	Samsung Galaxy S3 Mini	—	—

XSL•FO
RenderX

Study, year	Processing power	Operating system	Smartphone type	Sensor sampling rate	Camera resolution
Ge et al [27], 2020	—	Android	—	—	20 million pixels
Lee et al [9], 2016	—	Android	LG Optimus S smartphone	—	—
Reed et al [29], 2022	—	Android 5.0	Motorola Moto G II	—	—
Williams et al [31], 2021	—	Android 2.2	HTC Desire smartphone	—	60 frames per second and 1920×1080–pixel resolution
Gu et al [49], 2022	—	Android	Sony Xperia	—	Image resolution: 1980×1080 pixels
Gu et al [60], 2023	—	iOS	iPhone 5 or a newer device	—	Image resolution: 1980×1081 pixels
Prince et al [50], 2018	—	Android	—	—	—
Arora et al [52], 2015	—	—	—	—	—
Kassavetis et al [33], 2015	—	—	Huawei Mate 9 Pro	Smartphone accelerometers (50 Hz)	—
Ienaga et al [41], 2022	—	—	—	—	—
Espinoza et al [34], 2016	—	iOS	iPhone SE	—	—
Chén et al [51], 2020	—	—	—	—	20 million pixels
Surangsrirat et al [36], 2022	—	iOS	iPhone	—	—
Williams et al [53], 2020	—	iOS and Android	iPhone 11 Pro Max	—	60 frames per second, 1920×1080 pixels
Williams et al [11], 2020	—	Android	—	—	—
Prince and de Vos [54], 2018	—	Android	—	Smartphone app, screen, and accelerometer (100 Hz)	—
Santos et al [65], 2017	—	Ios	iPhone 5	—	—
Porkodi et al [40], 2023	—	Android	—	—	2400×1080–pixels and 64 megapixel f/1.89
Akhbardeh et al [57], 2015	—	—	Sony Xperia Z1	—	20.7 mega pixel

[a]Not applicable.

RQ 1: What Types of Hand Dysfunctions Are Studied, and What Clinical Hand Assessment Tools Are Used?

Overview

The hand dysfunctions discussed in the 46 articles were classified as an abnormal hand range of motion (ROM; n=18, 39%), hand tremor (n=15, 33%), hand bradykinesia (n=9, 20%), fine hand use decline (n=9, 20%), hypokinesia (n=4, 9%), and hand arthritis–related hand dysfunction (n=2, 4%). A total of 27 (59%) studies used clinical hand assessment tools (Table 3).

XSL•FO
RenderX

Table 3. Hand dysfunction type.

Hand dysfunction	Reference
Abnormal range of motion	[24-27,32,35,37-41,45-47,49,59,60,65]
Tremor	[6,9,10,28,31,33,36,42,48,51,52,54,56,61,63]
Bradykinesia	[6,9,11,33,36,43,48,53,54]
Decline of fine motor skills	[9,39,44,51,55,61-64]
Hypokinesia	[30,32,34,58]
Hand arthritis–related hand dysfunction	[29,57]

Abnormal Hand ROM

ROM describes how far a joint or muscle can move [67]. The measurement of ROM can indicate joint impairments in patients or the efficacy of rehabilitation programs [67]. Of the 46 studies, 19 (41%) focused on abnormal ROM, 11 (24%) focused on wrist ROM, and 10 (21%) focused on finger ROM. Smartphones were generally placed on the flexor carpi radialis and extensor pollicis longus [25,37,59] to measure wrist ROM and on the distal interphalangeal joint and proximal interphalangeal joint to measure finger ROM [24,25,35,37-39,45,49,60]. In addition, 6 related problems, namely hand injury [24,25,37,38,40,46,65], wrist injury [26,27,46,47,59], stroke [32,37,39,45], after hand surgery [41,60], flexor tendon injury [35], and nerve injury [49], were studied. Most studies (13/19, 68%) showed that the smartphone-based measurement method had the same reliability as the conventional goniometer when evaluating the ROM of healthy people and patients.

Hand Tremor

Hand tremor is a rhythmic, involuntary, and oscillatory (ie, rotating around a central plane) movement involving hand distal joints (eg, fingers and wrist) that is regularly recurrent [68]. All studies, except for 1 study on multiple sclerosis (MS), focused on PD hand tremors. For PD hand tremor assessment, the acceleration, rotational velocity, signal shake number and intensity were collected during daily life activities [6,10,28,36,42,51,56,61]. The number of taps or accuracy of each tap was measured during the finger-tapping activity of the smartphone app [33,48,50,52,63]. Smartphone-based hand dysfunction assessment shows satisfactory repeatability and validity when measured against the Movement Disorder Society of Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [28,33,36,50,52].

Hand Bradykinesia

Hand bradykinesia is characterized by slowness, reduced amplitude of movement, and sequence effect [69]. Hand bradykinesia is observed in patients with PD and patients with MS. PD and MS bradykinesia were detected in touch gestures, including finger tapping [9,11,36,43,53,54] and flick and pinch tactile behaviors [48]. The number of tapping trials and finger positions were examined to assess bradykinesia in hands. Daily activities and finger-to-nose tests were performed when holding the smartphone [6,33]. It was found that smartphones were comparable to conventional methods (such as MDS-UPDRS and Modified Bradykinesia Rating Scale) for assessing hand bradykinesia and may be useful in clinical practice [11,33,36,53].

Fine Hand Use Decline

Fine hand use refers to the use of small hand muscles to create movements, such as the use of a pencil to draw [70]. A total of 4 diseases were mentioned: PD [9,44,51,55,61,62], stroke [39], MS [63], and Huntington disease [64]. This type of hand dysfunction was assessed through smartphone screen interaction, such as playing games and typing activities [39]. Users' hold time, flight time, and pressure sequences during smartphone keystroke typing activity were used to quantify fine motor functions [9,44,51,55,62-64]. Studies show that smartphone has the potential to detect PD symptoms from the users' typing activity, which facilitates the development of digital tools for remote pathological symptom screening [39,44,61].

Hypokinesia

Hypokinesia is a decline in muscle strength that causes the muscle to not contract or move as it used to [71]. Three diseases related to this type of hand dysfunction are stroke [32,58], carpal tunnel syndrome (CTS) [30], and hand arthritis [34]. Patients who had a stroke were asked to perform gestures of grasping and floating [32,58] with a sensor glove worn. Hand information, such as finger position and velocity, were collected from patients with CTS as they played a game [30]. Patients with arthritis participated in power, pinch, and tripod grip tasks to capture grip measures [34]. These new methods show high sensitivity and specificity for disease detection and self-assessment [30,34].

Hand Arthritis–Related Hand Dysfunction

Arthritis is a common condition and is the most frequent cause of disability in American adults [57]. The most common form of arthritis is osteoarthritis, followed by inflammatory arthritis [72]. A method of analyzing hand dysfunction related to hand arthritis involved capturing photographs of each patient's hands. The results indicated that this approach could assist in the primary care, clinical assessment, and management of patients with hand arthritis [29].

Hand assessment tools used in the reviewed studies included clinical scales and instruments (Table 4). Clinical hand assessment tools were used for 2 purposes in 32 (70%) of the 46 studies: task design (n=7, 15% studies) and smartphone assessment outcome validation (n=25, 54% studies). The rest of the studies (14/46, 30%) did not mention the clinical tools. MDS-UPDRS was the most used clinical scale (15/46, 33%),

XSL•FO
RenderX

while a conventional goniometer was the most used instrument (10/46, 22%) [9,24,35,38,40,41,47,49,59,65]. Some studies used the MDS-UPDRS and the alternative finger-tapping test as reference tasks to set up experiment tasks. The effectiveness and reliability of smartphone-based assessment methods were validated by comparing the results with those of the MDS-UPDRS and manual goniometry.

Table 4. Clinical hand assessment tools used.

Clinical scale or instrument	References
For task design	
MDS-UPDRS[a]	[9-11]
CAPSIT-PD[b]	[43]
AFT[c]	[9,50,54,62]
TTT[d]	[9]
For outcome validation	
MDS-UPDRS	[28,31,33,36,38,39,43,50,52,53,62,64]
PDDS[e]	[63]
Neuro-QoL[f]	[63]
UHDRS[g]	[64]
Disease Activity Score-28	[34]
PDQ-8[h]	[36]
MBRS[i]	[11]
Tang criteria	[35]
Conventional goniometer	[9,24,35,38,40,41,47,49,59,65]
Mechanical tappers	[43]
Accelerometer	[61]
Electronic digital caliper	[57]

[a]MDS-UPDRS: Movement Disorder Society of Unified Parkinson's Disease Rating Scale.
[b]CAPSIT-PD: Core Assessment Program for Surgical Interventional Therapies in Parkinson's Disease.
[c]AFT: alternating finger tapping.
[d]TTT: time-tapping test.
[e]PDDS: patient-determined disease step.
[f]Neuro-QoL: quality of life in neurological disorders.
[g]UHDRS: Unified Huntington Disease Rating Scale.
[h]PDQ-8: 8-question Parkinson's Disease Questionnaire.
[i]MBRS: Modified Bradykinesia Rating Scale.

RQ 2: How Are Smartphone-Based Hand Assessment Tools Applied in Clinical Practice?

Smartphone-based hand assessment has been applied in 4 different ways. It has been used for the measurement of function parameters (ie, wrist and finger ROM and hand strength), the early detection of disease-related dysfunction, real-time assessment during rehabilitation, and function assessment and rating (Table 5).

XSL•FO
RenderX

Table 5. Functions of smartphone-based hand assessment tools.

Application setting and task scenario	References
Measurement	
Finger or wrist extension or flexion	[24-27,35,37-41,47,49,59,60]
Finger implement squeeze and finger forward flexor tendon gliding	[25]
A grip force–tracking task	[34,45]
TTT[a]	[9,43]
RAM[b], tremor tracker, and CIT[c]	[9]
Wrist pronation and supination	[65]
(Early) detection	
Daily activity	[42]
Extended and rest versions of MDS-UPDRS[d]	[10,51]
Finger-tapping test	[44,51,52,54,55,62]
Daily motor active tests	[6]
Flick, drag, pinch, and handwriting gestures	[48]
Play a game	[30]
Finger-to-nose test, pronation supination test, and arm-circle exercise	[28]
Photographic capture of the patient's hands	[29,57]
Reaction time test	[52]
Real-time assessment during rehabilitation	
Finger and wrist extension	[26,37,39]
Wrist flexion, wrist extension, finger implement squeeze, and finger forward flexor tendon gliding	[25]
A grip force–tracking task	[45]
Play a game	[39,46]
Grasping, pinching, and waving	[32]
Hand grip and flat	[58]
Function-level rating	
Hanging gestures	[28,31,33]
Finger-to-nose test	[33,61,63]
Photographic capture of the patient's hands	[57]
Grip force–tracking task	[30]
Extended and rest versions of MDS-UPDRS	[29,50,61]
Finger-tapping test	[11,33,36,53,63,64]
Hold the phone	[61]

[a]TTT: time-tapping test.

[b]RAM: rapid alternating movement.

[c]CIT: Cognitive Interference Test.

[d]MDS-UPDRS: Movement Disorder Society of Unified Parkinson's Disease Rating Scale.

Of the 46 studies, 18 (39%) focused on the measurement of hand function parameters such as wrist ROM [26,27,37,40,41,47,59,65], finger ROM [24,25,35,37-39,45,49], hand gesture [49], hand dexterity [9], or hand grip strength [34]. Hand grip strength measurement and hand dexterity measurement were conducted on smartphones and shown to have good constancy with traditional measurement tools [16,23,38].

Furthermore, 15 (33%) out of the 46 papers focused on dysfunction assessment for early disease detection. Dysfunctions, such as hand tremor (10/46, 22%), hand bradykinesia (3/46, 7%), fine hand use decline (5/46, 11%), and hypokinesia (2/46, 4%), were used as biomarkers for certain

XSL•FO
RenderX

diseases, such as PD [6,10,29,30,42,44,48,51,52,54,55,57,61,62], CTS [30], and hand arthritis [57,65]. The detection exhibited high sensitivity and specificity, supporting personalized treatment plan adjustments and enabling early disease diagnosis and optimized management [55].

Among the 46 studies, 14 (30%) concentrated on rating hand dysfunction severity, mostly in PD- or MS-induced hand tremor (8/46, 17%) and bradykinesia (4/46, 9%). The findings demonstrate that smartphones can determine the degree to which the patient is affected by the disease, rating the severity of both the disease and hand dysfunction [45,67,68].

Furthermore, 8 (17%) of the 46 studies explored how smartphones were used for real-time hand function assessment during hand rehabilitation [25,26,32,37,39,45,46,58]. Smartphones provide an interactive interface with guided exercises, therapeutic games, and performance feedback [26,45]. The results of real-time assessment during rehabilitation can help increase patients' motivation and interest, reduce discontinuity in the rehabilitation process, and lower treatment costs [25,26,32,37,39,45,46,58].

RQ 3: How Are Smartphones Used to Assess Hand Function?

The literature showed that smartphones had been used in 4 ways for hand function assessment: data collection (38/46, 83% studies), data display (17/46, 37% studies), data transmission (15/46, 33% studies), and data processing (6/46, 13% studies).

Data Collection

Data were mainly collected via embedded smartphone sensors or smartphone apps [42]. Accelerometers (12/46, 26%) [15,24,26,28,33,36,42,51,54,56,61,64] were the most used built-in smartphone sensors, followed by smartphone cameras (11/46, 24%) [11,27,29,31,35,40,41,49,53,57,60], gyroscopes (5/46, 11%) [6,10,51,59,64], and goniometers (2/46, 4%) [38,47] (Table 6). Some of the smartphone apps (16/46, 35%) [30,33,43,50,52,63,64] were developed to work as a digital tapper to collect the number of trials and position of each tap during the time-tapping test, and AFT task was used to detect hand use, hand tremor, bradykinesia, or ROM. Accelerometers can collect rich information, including angles and the rotational velocity vector of the finger [24,26]. The sampling rate range of accelerometers was 20 to 100 Hz. By using a smartphone's camera, the patient's hand picture can be captured to extract information such as wrist and finger extension and flexion, allowing the measurement of joint ROM or extension [35,41,60]. The camera resolution range was 1920×1080 pixels to 2400×1080 pixels.

XSL•FO
RenderX

Table 6. Built-in sensors involving data collection.

Sensor and measurement	App name	References
Accelerometers		
All angles of DIPj[a], PIPj[b], and MPj[c], including the right and left, active and passive, and extensor and flexor positions	Google LLC and EHMROM	[24]
Still acceleration	HTrembAPP[d]	[42]
The acceleration vector and the rotational velocity vector	DNM[e]	[15]
Accelerometer signal	Roche PD Mobile Application (version; Roche), PD Dr, Apkinson, GEORGE, mPower, and mobile accelerometer software	[28,33,36,51,54,56,61,64]
Orientation, velocity, and motion	HandRehab app	[26]
Smartphone app		
Number, time, velocity, position, consistency, amplitude, and accuracy of each tap	SmT[f], DNM, mPower, Apkinson, elevateMS, ReHand, GEORGE, and HLTapper	[30,33,36,43,50,52,54,55,61-64]
150 test parameters	DNM	[9]
Kinetic tremor and dysmetria in movement	elevateMS	[63]
Pronation, supination, flexion, and extension	DNM and Angulus app	[40,65]
Camera		
Movement and tremor	Did not use an app	[27,31]
Hand video	Did not use an app	[11,31,53]
Hand picture	DNM	[29,35,40,41,49,57,60]
Joints' angles and key point's distance	Did not use an app	[49]
Extension or flexion of the joint	Did not use an app	[35,41,60]
Movement of finger	Did not use an app	[60]
Tapping frequency, amplitude, speed, or rhythm	Did not use an app	[11,53]
Gyroscope		
Gyroscope data in discrete time	DNM	[10]
Gyroscope signal	Roche PD Mobile Application (version 1; Roche) and GEORGE	[51,64,66]
Height, rotation, slope, and acceleration	Gyroscope	[59]
Goniometer		
Finger flexion at MCPj, PIPj, and DIPj and flexion angles of the finger	Goniometer	[38]
Wrist flexion, extension, supination, and pronation ROM[g]	Compass app	[47]
GPS		
Orientation, velocity, and motion	HandRehab app and newly created smartphone apps	[26]
Microphone		
Voice	Roche PD Mobile Application (version 1)	[6]
Pressure sensor		
Pressure-based features	Custom Android app (the name of the app was not mentioned)	[48]
Finger pressure	DNM	[46]
IMU[h]		

XSL•FO
RenderX

Sensor and measurement	App name	References
IMU–based features	Custom Android app (the name of the app was not mentioned)	[48]

[a]DIPj: distal interphalangeal joint.
[b]PIPj: proximal interphalangeal joint.
[c]MPj: metacarpophalangeal joint.
[d]HTrembAPP: Hand Trembling Detector App.
[e]DNM: did not mention.
[f]SmT: smartphone tapper.
[g]ROM: range of motion.
[h]IMU: inertial measurement unit.

Data Display

Data display (17/46, 37%) included the display of raw data (12/46, 26%) [24,26,28,32,34,37,42,45,51,55,58,61], visual instructions (10/46, 22%) [25,26,28,30,37,39,46,55,63,64], and information notification [10,61] (2/46, 4%). Data were frequently displayed in the text form [28,32,34,42,45,51,55,61] and graphic form [24,26,37,58]. Test details, such as date and patient information [26,42,45], were usually displayed. Assessment feedback was also displayed in the form of results or scores [25,45]. The real-time feedback displayed included hand motion data [28,45], virtual 3D representation of finger posture [26], and interactive game interfaces [39].

Data Transmission

Data transmission describes how data are transmitted between smartphones and external devices (Table 7). Due to limited data processing capacity, smartphones generally send data to other resources through Bluetooth, USB dongles, and Wi-Fi for data processing and storage [6,39,43]. Of the 46 studies, 12 (26%) transmitted the data to a cloud server through a unidirectional transfer, meaning data only flowed in 1 direction. Among these 12 studies, 7 (58%) developed a smartphone app to receive the built-in sensor data [6,10,26,28,32,43,61], and the other 5 (42%) designed a smartphone app to receive the training data from external devices (ie, gloves) [25,39,46,58,62]. A total of 3 (%) of the 46 papers reported that smartphones transmitted data with an external device via bidirectional communication [32,45,58], indicating smartphones can send and receive data in both directions. Furthermore, 2 (%) of the 46 papers discussed data privacy and security and referred to Health Insurance Portability and Accountability Act regulations [32].

Table 7. The objects involved in data transmission.

Receiver	References
Remote server	
Computer	[10,56]
Google Drive	[43]
Cloud storage facility	[6]
Cloud computing	[28,47]
Remote server	[32,61,62]
Physician	[6,25,26,46]
External device	
Glove	[32,39,46,58]
HandMATE device	[39,45]

Data Processing

Data processing involves the use of smartphones as terminals to analyze, manipulate, and transform raw data into useful information or machine-readable content [39]. Among the 46 studies, 6 (13%) used a smartphone app to process data [24-26,32,39,42], and 1 (2%) reported the smartphone's processing power [24]. The smartphone processed motion data collected from built-in sensors and external devices. Data collected from built-in sensors, such as ulnar and radius deviations, were converted into ROM and total active motion [24,39,42]. Data from external devices' sensors, such as flex-sensor signals and electromyography, were transformed into flexion and extension angles (in degrees) [26,32]. One of the studies extracted the features from electromyography sensors and then fed them to an ML algorithm for further gesture recognition on smartphone apps [25].

Use of Smartphones for Multiple Functions

A total of 21 (46%) of the 46 studies designed smartphones integrating more than one of the functions mentioned earlier. The most frequent combination was using a smartphone for data transmission and data display [25,26,28,32,39,45,46,58,61] (Table 8). A total of 8 (17%) studies combined ≥3 functions

XSL•FO
RenderX

[24-26,28,32,39,42,61]. For example, in the study by Bercht et al [25], the smartphone was designed to integrate processing capabilities, enabling the real-time reception of game information from the glove's flex sensor and then display of the information on the smartphone screen after local data processing.

Table 8. Use of smartphones for multiple purposes.

Study, year	Data collection	Data processing	Data transmission	Data display
Matera et al [26], 2016	✓	✓	✓	✓
Miyake et al [24], 2020	✓	✓		✓
García-Magariño et al [42], 2016	✓	✓		✓
Bercht et al [25], 2012		✓	✓	✓
Janarthanan et al [39], 2020		✓	✓	✓
Pan et al [28], 2015	✓		✓	✓
Orozco-Arroyave et al [61], 2020	✓		✓	✓
Sarwat et al, 2021 [32]		✓	✓	✓
Kostikis et al [10], 2015	✓		✓	
Lee et al [43], 2016	✓		✓	
Lipsmeier et al [6], 2018	✓		✓	
Sandison et al [45], 2020			✓	✓
Halic et al [46], 2014			✓	✓
Koyama et al [30], 2021	✓			✓
Chén et al [51], 2020	✓			✓
Arroyo-Gallego et al [62], 2017	✓			✓
Pratap et al [63], 2020	✓		✓	
Waddell et al [64], 2021	✓			✓
Mousavi et al [56], 2020	✓			✓
Lee et al [55], 2016	✓		✓	
Hidayat et al [58], 2015	✓			✓

RQ 4: What Statistics or ML Algorithms Are Used for Hand Function Assessment?

Overview

Among the 46 studies, 39 (85%) used statistical methods to process the hand motion data, including parameters such as tapping speed, error, and speed during smartphone screen interaction; 20 (43%) applied ML to analyze the raw data or statistical features; and 17 (37%) used both statistical and ML methods. By contrast, 4 (9%) studies used neither statistics nor ML for data analysis [37,39,47,59].

Statistical Methods

Overall, 21 types of statistical methods were used to process 6 types of hand motion raw data (Table 9). The most used method was summary statistics (23/46, 50%), followed by normalization (7/46, 15%) and Fourier transform (6/46, 13%).

XSL•FO
RenderX

Table 9. Studies classified by statistical methods.

Data processed and statistical method	References
Data collected during the smartphone screen interaction (ie, tapping speed, error, speed, path, pressure, and distance)	
Pythagorean theorem	[43]
Normalization	[33,48,54,61,62,64]
Bootstrap multiple regression	[9]
Summary statistics (range, mean, median, and SD)	[11,30,36,43,50,52,53,55,62]
Akaike information criterion	[9]
Fourier transform	[11,33,53]
Accelerometer values and rotational velocity vector	
ObtainDirection	[42]
ObtainAlpha	[42]
Band-pass filter	[10,64]
Spectral analysis	[10]
Fourier transform	[10,28]
Summary statistics (range, mean, median, and SD)	[34]
Mass univariate	[51]
Feature-wise correlation test	[51]
Regularization	[51]
Butterworth high-pass filter	[33]
EMD[a]	[56]
Smartphone video or picture	
Fourier transform	[31]
Normalization	[57]
Summary statistics (minimum, maximum, mean, median, and SD)	[60]
One-hot encoding categorical and scaling numerical responses	[29]
Savitzky-Golay filter	[11]
Initiating, terminating flexion, extension, and ROM[b]	
RMS[c] error	[27,32,35,40,45]
FSR[d], IMU[e], or pressure sensor signals	
Ōtsu's 11 binarization	[41]
RMS error	[45]
Summary statistics (range, mean, median, and SD)	[32,37]
SMA[f] filtering	[58]
Variables for model prediction (ie, age, sex, and occupation)	
Linear mixed models	[59]
Multiple linear regression	[9]

[a]EMD: empirical mode decomposition.
[b]ROM: range of motion.
[c]RMS: root mean square.
[d]FSR: force sensing resistor.
[e]IMU: inertial measurement unit.
[f]SMA: simple moving average.

XSL•FO
RenderX

ML Methods

In total, 16 types of ML methods were identified (Table 10). They were applied for 4 purposes: disease detection, disease severity evaluation, disease prediction, and feature aggregation. Support vector machines (SVMs) were the most used ML method [10,28,48,49,53,56,62]. The input features of SVMs were preprocessed acceleration signals, such as the sums of squared magnitudes [10] and path- or time-based features [48]. Tian et al [60] reported SVMs as a reliable ML method for early PD detection and multivariate classification with 0.89 sensitivity and 0.88 specificity. Gu et al [49] reported the highest gesture classification accuracy of 1, with a sensitivity of 1 and specificity of 1.

Among the 46 studies, 5 (11%) applied logistic regression for disease severity classification and prediction and hand gesture discrimination [32,51,53,54,62]. The spatiotemporal features from the pixel coordinate data during finger tapping and accelerometer waveforms were the input for this ML method. Logistic regression showed an average accuracy of 88.5% (SD 8.03%; grasp), 83% (SD 10.9%; pinch), and 86.5% (SD 12.57%; wave) [32] and an accuracy of 0.61 and area under the curve (AUC) of 0.59 in PD prediction [53].

Of the 46 studies, 3 (7%) [29,44,54] exploited convolutional neural networks to distinguish patients with PD from healthy controls based on hold time, flight time, and pressure sequences [44]. Convolutional neural networks exploited the finger-tapping rate data for PD severity identification with an AUC of 0.64 and accuracy of 0.62 [54]. They also worked as the base layer for training 2 image preprocessing models and for discriminating PD tremors from other types of tremors with 95% agreement with the accelerometer [29].

Among the 46 studies, 7 (15%) [10,32,48,49,51,53,62] compared the classification performance of different ML algorithms. For example, Kostikis et al [10] applied decision tree (DT), Naive Bayes, C4.5 DT, and a bagged ensemble of DTs for distinguishing patients with PD from healthy participants based on PD hand tremor features. Bagged ensemble of DTs performed better than other classifiers, with an accuracy of 0.90 for the healthy group and 0.82 for the PD group and an AUC of 0.94.

XSL•FO
RenderX

Table 10. Studies classified by MLa algorithms.

ML and feature	Validity and accuracy	References
SVM[b]		
Magα[c], magω[d], sdα[e], and mAmpω[f]	Distinguishing patients with PD[g] from healthy participants: sensitivity=0.56 and specificity=1	[10]
Path-based, time-based, pressure-based, and IMU[h]-based features and additional features for handwriting gestures and pinch gestures	In healthy controls: sensitivity=0.89 and specificity=0.88	[48]
The total, peak and fraction power and average acceleration of the motion data	PD hand resting tremor detection: sensitivity=0.77 and accuracy=0.82	[28]
Angles of the MCPj[i], PIPj[j], DIPj[k], and CMCj[l] of fingers; webspace; etc	Highest gesture classification: accuracy=1, sensitivity=1, and specificity=1	[49]
SFS[m] to select the best feature from the mean, SD, skewness, etc, from accelerometer signals	Tremor activity identified with the highest accuracy of 0.91, specificity of 0.90, and sensitivity of 0.90	[56]
Touchscreen typing features: covariance, skewness, and kurtosis analysis of the timing information	The typing feature aggregated with an AUC[n] of 0.88 (linear-SVM)	[62]
Tapping frequency, amplitude, energy spectral density, and peak-to-peak variability	PD diagnosis predicted with an accuracy of 0.63 and AUC of 0.60 (linear-SVM)	[53]
Tapping frequency, amplitude, energy spectral density, and peak-to-peak variability	PD diagnosis predicted with an accuracy of 0.69 and AUC of 0.68 (SVM-RBF[o])	[53]
Logistic regression		
The mean, RMS[p], SMA[q], and SD for each axis of the accelerometer and gyroscope	Patient performance assessed with average accuracy of 88.5% (SD 8.03%; grasp), 83% (SD 10.9%; pinch), and 86.5% (SD 12.57%; wave)	[32]
Touchscreen typing features: covariance, skewness, and kurtosis analysis of the timing information	The typing feature aggregated with an AUC of 0.87	[62]
Tapping frequency, amplitude, energy spectral density, and peak-to-peak variability	PD diagnosis predicted with an accuracy of 0.61 and AUC of 0.59	[53]
13 spatiotemporal features from the pixel coordinate data about speed, rhythm, accuracy, and fatigue and 28 features from 3 accelerometer waveforms, frequency, and temporal domains	PD severity classified with an AUC of 63.1 (SD 2.11) accuracy of 59.5 (SD 0.96)	[54]
Features selected according to formulas and parameters	Patients with PD distinguished from healthy participants with an accuracy of 0.94, sensitivity of 0.95, and specificity of 0.94 (multivariate logistic regression)	[51]
CNN[r]		
4 statistical features from HT[s], FT[t], and pressure sequences	Classification of patients with PD and healthy controls: in the clinic, mean performance=0.89, sensitivity=0.79, and specificity=0.79; in the wild, mean performance=0.79, sensitivity=0.74, and specificity=0.78	[44]
12 features, such as sex, age, and the duration of symptom	Discriminant PD tremor with 95% agreement with accelerometer	[29]
Raw data of finger tapping	PD severity identified with an AUC of 63.5 (SD 1.56) and accuracy of 62.1 (SD 0.95)	[54]
RF[u]		
Angle of fingers' MCPj, PIPj, DIPj, and CMCj; webspace; etc	Highest gesture classification: accuracy=1, sensitivity=1, and specificity=1	[49]
Mean, SD, and median acceleration	In discriminating participants with PD from controls, sensitivity=0.96 (SD 0.2) and specificity=0.97	[52]
13 spatiotemporal features from the pixel coordinate data about speed, rhythm, accuracy, and fatigue and 28 features from 3 accelerometer waveforms, frequency, and temporal domains	PD severity identified with an AUC of 64.1 (SD 1.08) and accuracy of 60.2 (SD 1.56)	[54]

XSL•FO
RenderX

ML and feature	Validity and accuracy	References
Linear regression		
Magα, magω, sdα, and mAmpω	Patients with PD distinguished from healthy participants with a sensitivity of 0.74 and specificity of 1	[10]
Angle of fingers' MCPj, PIPj, DIPj, and CMCj; webspace; etc	Highest gesture classification: accuracy=1, sensitivity=1, and specificity=1	[49]
AdaBoost		
Magα, magω, sdα, and mAmpω	Patients with PD distinguished from healthy participants with a sensitivity of 0.83 and specificity of 0.85	[34]
Touchscreen typing features: covariance, skewness, and kurtosis analysis of the data timing information	The typing feature aggregated with an AUC of 0.82	[62]
KNN[v]		
Time domain: the signal length, mean value, RMS value, number of vertices, and number of baseline crosses; frequency domain: fundamental frequency, region length, and Fourier variance	Validated with self-defined hand gesture performance classification standards with an accuracy of >95%	[25]
NB[w]		
Magα, magω, sdα, and mAmpω	Patients with PD distinguished from healthy participants with a sensitivity of 0.56% and specificity of 1	[10]
Tapping frequency, amplitude, energy spectral density, and peak-to-peak variability	PD diagnosis predicted with an accuracy of 0.69 and AUC of 0.70	[53]
XGBoost[x]		
Features selected according to formulas and parameters	Patients with PD distinguished from healthy participants with an accuracy of 0.81, a sensitivity of 0.83, and a specificity of 0.9	[51]
The mean, RMS, SMA, and SD for each axis of the accelerometer and gyroscope	Patient performance assessed with average accuracy of 88% (SD 9.88%; grasp), 83.5% (SD 7.74%; pinch), and 82% (SD 14.71%; wave)	[32]
C4.5 DT[y]		
Magα, magω, sdα, and mAmpω	Patients with PD distinguished from healthy participants with a sensitivity of 0.83 and specificity of 0.75	[10]
BagDT[z]		
Magα, magω, sdα, and mAmpω	Patients with PD distinguished from healthy participants with a sensitivity of 0.82 and specificity of 0.90	[10]
DT		
Magα, magω, sdα, and mAmpω	Patients with PD (accuracy rate 82%) distinguished from healthy people (accuracy rate 90%)	[10]
HAR[aa]		
Sustained phonation: MFCC2[ab]; rest tremor: skewness; postural tremor: total power; finger tapping; balance: mean velocity; and gait: turn speed	Unlabeled PD activity test data: PD balance activity test: 99.5%; gait activity test: 96.9%; and distinguishing between resting and gait activities: 98%	[6]
Anomaly detection and an autoencoder		
The position, time, or velocity of the thumb movement	Participants with and participants without CTS[ac] classified with a sensitivity of 0.94, a specificity of 0.67, and an AUC of 0.86	[30]
Elastic net		
Features selected according to formulas and parameters	Patients with PD distinguished from healthy participants with an accuracy of 1, a sensitivity of 0.95, and a specificity of 1	[51]
DNN[ad]		

XSL•FO
RenderX

ML and feature	Validity and accuracy	References
13 spatiotemporal features from the pixel coordinate data, including speed, rhythm, accuracy, and fatigue, and 28 features from 3 accelerometer waveforms, frequency, and temporal domains	PD severity classified with an AUC of 65.7 (SD 1.05) and accuracy of 61.2 (SD 1.07)	[54]

[a]ML: machine learning.
[b]SVM: support vector machine.
[c]magα: the sums of squared magnitudes of the acceleration.
[d]magω: the sums of squared magnitudes of the rotation rate vector.
[e]sdα: the sums of absolute differences in the acceleration vector.
[f]mAmpω: the maximum sums of the 3 axial components of the rotation vector ω calculated by Fourier transform.
[g]PD: Parkinson disease.
[h]IMU: inertial measurement unit.
[i]MCPj: metacarpophalangeal joint.
[j]PIPj: proximal interphalangeal joint.
[k]DIPj: distal interphalangeal joint.
[l]CMCj: carpometacarpal joint.
[m]SFS: feature selection algorithm.
[n]AUC: area under the curve.
[o]RBF: radial basis function.
[p]RMS: root mean square.
[q]SMA: simple moving average.
[r]CNN: convolutional neural network.
[s]HT: hold time.
[t]FT: flight time.
[u]RF: random forest.
[v]KNN: K-nearest neighbor.
[w]NB: naive Bayes.
[x]XGBoost: extreme gradient boosting.
[y]DT: decision tree.
[z]BagDT: bagged ensemble of decision trees.
[aa]HAR: human activity recognition.
[ab]MFCC2: mel-frequency cepstral coefficient2.
[ac]CTS: carpal tunnel syndrome.
[ad]DNN: deep neural network.

Discussion

To the best of our knowledge, this is the first systematic review on the primary design ideas and development of smartphone-based technologies for hand function assessment.

RQ 1: What Types of Hand Dysfunctions Are Studied, and What Assessment Inventory Tools Are Used?

In the literature, smartphones only assessed 6 types of hand dysfunctions, namely abnormal ROM, tremor, bradykinesia, fine motor skill decline, hypokinesia, and hand arthritis–related hand dysfunction. The reason might be that smartphones are limited in capturing the complexity and variety of hand movements to measure all aspects of clinically relevant hand functions [73]. Other types of hand dysfunctions such as decreased grip strength, altered sensation, and impaired coordination are important biomarkers clinically, requiring the future development of smartphones to collect related parameters [74].

ROM is a critical and objective measurement that can reflect various diseases, such as arthritis, trauma, and stroke [75]. Abnormal ROM was the most studied smartphone-based hand function assessment [24-27,32,37-41,45-47,49,51,59,60,65], indicating the advantages of smartphones in obtaining ROM parameters. Therefore, the further development of smartphones to achieve better accuracy and reliability in capturing ROM is warranted. With the advancement of built-in accelerometers and gyroscopes in smartphones, capturing and analyzing hand ROM data have become more accessible [75,76]. Furthermore, smartphones can accurately measure both dynamic ROM and static ROM, providing good potential for long-term monitoring even without the presence of professionals [27].

PD is the most studied disease that causes hand dysfunction. PD can cause multiple hand dysfunctions, such as tremors [6,9,10,42,48], bradykinesia [6,9,43,48], abnormal ROM [37,39,45], and fine hand use decline [44]. It provides evidence that smartphones have the potential to provide a comprehensive assessment platform for multiple hand dysfunctions [9,42-44].

XSL•FO
RenderX

In addition, chronic neurodegenerative diseases, such as PD, exhibit progressive symptoms that require continuous monitoring [7]. However, existing clinical assessment tools, such as MDS-UPDRS, tend to be subjective, time constrained, and time consuming [77]. Smartphone apps could exploit the multiple built-in sensors in smartphones to detect changes indicative of the disease progression or treatment response [78-82], indicating that smartphones can be prosperous tools for managing chronic hand dysfunction in the long run.

Above all, for a reliable clinical application of hand dysfunction assessment, the following should be achieved:

1. Gold standards should be established and validated, specific to the smartphone as an assessment platform.
2. Smartphone assessment should be customizable according to an individual's condition and rehabilitation expectations [83].
3. Smartphone assessment procedures and tasks should adhere to the operational specifications of the clinical assessment criteria [2,84].
4. An individualized rehabilitation plan should be generated from the assessment and evaluated in real-time pace to monitor the individual's rehabilitation progress.

RQ 2: How Are Smartphone-Based Hand Assessment Tools Applied in Clinical Practice?

Real-time assessment during hand rehabilitation is beneficial in clinical practice because it allows the modification of the rehabilitation tasks and goals according to an individual's specific needs and ongoing recovery progress [85]. In our review, studies on real-time smartphone-based assessment were primarily conducted between 2016 and 2022, indicating an emerging trend focusing on real-time hand assessment. A potential technical challenge may lie in identifying the best sensor configuration and feature extraction method for hand function assessment [6,84].

The early detection of a degenerative disease through hand assessment is important because it can help slow down further disease progression [86]. The reviewed literature discussed conditions such as PD and CTS [34,36,37]. Future work could use smartphones for biomarker acquisition to monitor disease-relevant physiological and behavioral symptoms and provide personalized rehabilitation guidance [87-89]. The use of smartphones for biomarker acquisition offers advantages, including portability, accessibility, affordability, noninvasiveness, and continuous monitoring, benefiting both patients and clinicians [90]. However, challenges exist in terms of data quality, reliability, and privacy concerns [91].

RQ 3: How Are Smartphones Used to Assess Hand Function?

Smartphones were mostly used for data collection. With more sensors embedded in smartphones, richer and more dimensional data can be collected for function measurement. For example, the resolution of smartphones' built-in camera is between 1920×1080 and 2400×1080 pixels, which is higher than the commonly used camera resolution in clinical settings, which typically ranges from 1280×720 to 1920×1080 pixels [8]. Compared to smartwatches and ring-shaped sensors, smartphones are more indispensable in people's daily lives, making them an easily available assessment tool and requiring no extra investment like others [92]. While webcams provide high resolution and frame rates, they rely on a stable internet connection and can potentially raise privacy and security concerns [93]. In comparison, smartphones can collect data offline and protect the patient's privacy by encrypting data, anonymizing personal information and storing data locally [40]. This also shows that smartphones, as general-purpose devices, do not require excessive hardware requirements, are available at a low cost, and are easy to access. Smartwatches and wearables usually feature multiple sensors similar to those found in smartphones, allowing for the collection of hand motion and physiological data with real-time feedback. However, their functionality is confined by a fixed position of the body, resulting in the limited scope of data collection [14]. In contrast, smartphones, being portable devices, are not constrained by fixed positions, granting convenience and flexibility for hand dysfunction assessment. Ring-shaped sensors offer high precision and accuracy and provide real-time data. However, their use may be limited due to comfort and portability constraints [16]. Smartphones are equipped with data processing modules, which can analyze and process data in real time, providing better accuracy at the same cost [94]. In terms of user experience, as a more familiar product, smartphones reduce the users' learning cost and provide a more convenient, personalized, and friendly hand dysfunction evaluation experience, which helps improve user participation and satisfaction [19]. However, one of the weaknesses of using a smartphone for data collection may be data errors or biases due to the smartphone user's lack of training, supervision, and quality control [95].

Using smartphones for data processing was the least mentioned in the studies [24-26,32,39,42]. The benefits of smartphone data processing are manifold, including mobility, real-time processing, and interactive nature [96]. This empowers users to access and process data at any time, receive real-time feedback, and seamlessly interact with their smartphones, regardless of location [97]. Despite the advantages, there are also obstacles to overcome, including short battery life, limited storage capacity, and weak processing power [98]. Therefore, most of our reviewed studies focused on the wireless transmission of data to computers or the cloud for subsequent data processing [6,10,39,46]. This approach would allow for efficient data management and processing without consuming the limited storage space available in smartphones (Figure 2) [10].

XSL•FO
RenderX

Figure 2. The primary design ideas for the development of smartphone-based hand function assessment technology. AI: artificial intelligence; FSR: force sensing resistor; IMU: inertial measurement unit; ML: machine learning.

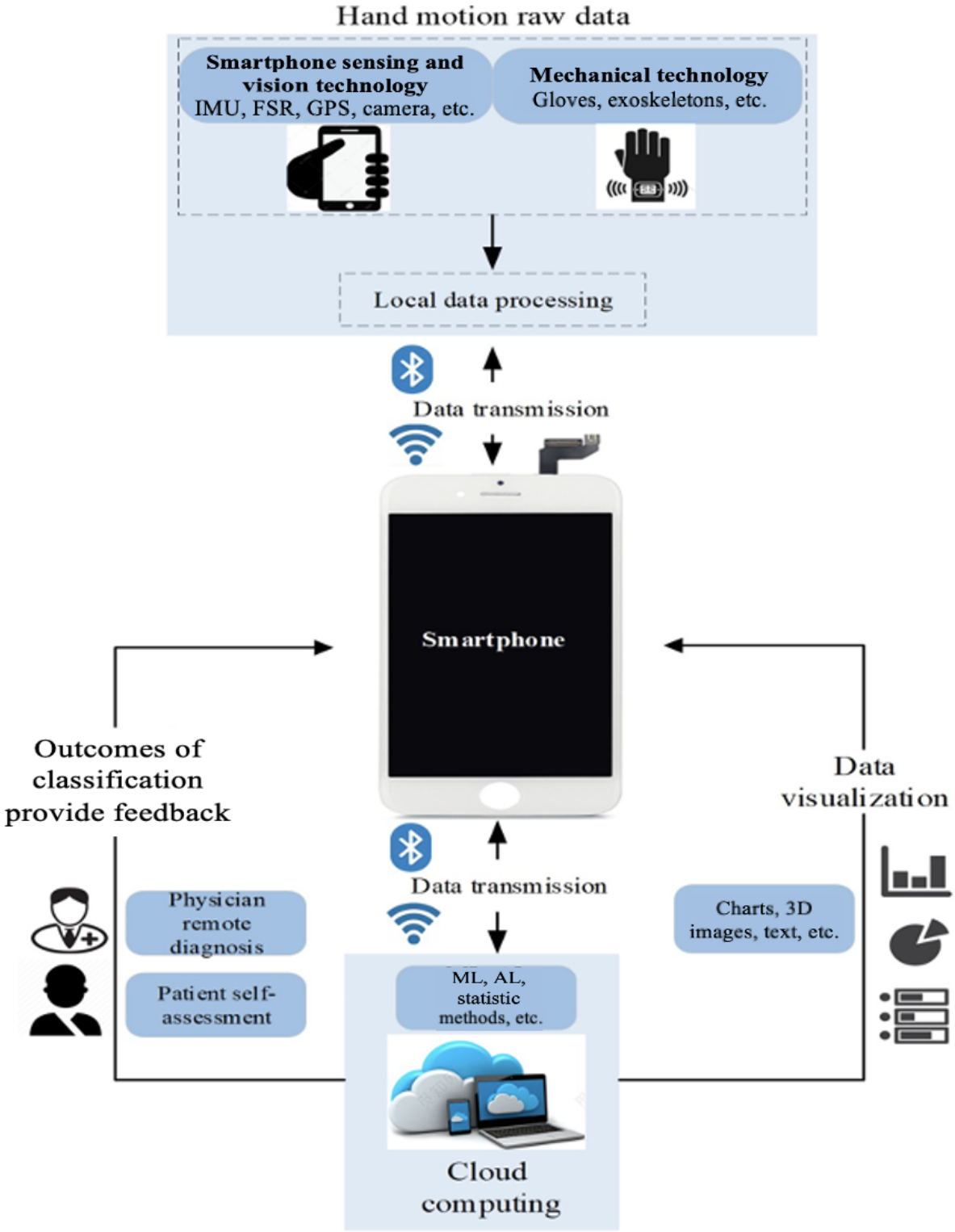

In this review, among the 46 studies, 7 (15%) exclusively involved healthy participants, while 23 (50%) recruited both patients and healthy participants. Consequently, 65% (30/46) of the studies included healthy participants, marking a noteworthy finding. In smartphone-based hand dysfunction assessment, incorporating baseline data from healthy participants is important for several reasons [37-41,59,60]. First, a standard reference range is typically derived from data collected from healthy participants, which could enable a more precise evaluation of a patient's hand dysfunction. By comparing the hand function of patients to that of healthy participants, potential abnormalities can be identified more effectively, assisting in the accurate diagnosis of issues and facilitating the implementation of appropriate treatments [6,10,11,29,42-55,61-65]. Second, during the rehabilitation process, the patient's recovery progress and improvement can

XSL•FO
RenderX

be quantified by comparing against data from health people [45,46]. The effectiveness of the treatment can be more accurately assessed, and rehabilitation protocols could be adjusted for better outcomes. Third, it's necessary to establish a normal reference range from healthy participants, including different ages, sex, and demographic characteristics. A broader set of data is available, ensuring that assessments are not limited to a specific group and can cover a broader population, resulting in a complete and more comprehensive understanding of hand function assessment [99]. In summary, remote assessment platforms have been developed for a wide range of users, including professionals, caregivers, and patients [2,10]. However, certain aspects need to be considered when using smartphones for hand assessment. They are as follows [28,100-102]:

1. Establishing standardized data formats is of utmost importance to ensure compatibility and consistency in data analysis. Inconsistent data formats can pose challenges in data analysis, making it difficult to compare and analyze data obtained from various smartphones.
2. It is necessary to ensure the robustness of smartphone processors or network connections. The effectiveness of the smartphone processor and network can impact the frequency of data updates, which may result in delays when acquiring and displaying real-time data.
3. It is necessary to consider privacy and security. It is important to prioritize data security and privacy by implementing app-appropriate encryption measures during data transmission to mitigate potential ethical and legal issues and ensure compliance with relevant data-protection regulations.

RQ 4: What Statistics or ML Algorithms Are Used for Hand Function Assessment?

Statistical methods (39/46, 85%) were more commonly used than ML methods (20/46, 43%). The most commonly used statistical method was summary statistics such as mean and SD. Summary statistics offer concise insights into data, facilitating comparisons and simplifying analysis [103]. However, they can be subjective, relying on expert experience, and may distort information [104]. In addition, due to the multiple independent variables present in hand function assessment [83], it is important to consider statistical methods that are capable of analyzing a multifactor model, such as multiple linear regression [105].

ML methods have been increasingly used in various health care apps [106]. In the studies in our review, ML methods were mainly used for detecting and classifying patient hand posture, analyzing and classifying behavior patterns (ie, tremor, bradykinesia, and ROM), and identifying disease severity and prediction. Our review found SVMs to be the most commonly used ML algorithm, particularly for disease classification. This may be attributed to the fact that SVMs are capable of effectively addressing multi-dimensional data with small sample sizes while providing a good generalization performance and the ability to work with the primary processing stage data [107]. The main limitation of the SVM algorithm is its inability to handle multiclass classification problems without additional modifications or extensions [108].

Strengths and Limitations of the Study

The strengths of this review are as follows: (1) the relevant database searches were conducted in a comprehensive and reproducible manner; (2) this was the first review that aimed to comprehensively discuss the role of smartphones and their functionalities in hand assessment from a holistic perspective; and (3) this review provides an analytical demonstration of the technical feasibility and advantages of using smartphones for hand function assessment across various domains, including sensor support, clinical practice, and application scenarios. It recommends potential directions for future studies in this field, such as multisensor fusion, gold-standard establishment, real-time assessment, and ML algorithms for data analysis exploration. This review also has some limitations. First, given that smartphone-based hand function assessment is at its nascent stage, the number of relevant studies is limited. This may contribute to a lack of sufficient evidence, completeness, and comprehensiveness in research materials, posing challenges in supporting viewpoints, drawing conclusions, and gaining a comprehensive understanding of the field. Second, this review encompassed only studies in the English language. Third, due to the exploratory and developmental nature of this topic, the literature quality varied, with potential limitations, such as inconsistency and a lack of high-quality reference studies and as well as meta-analyses.

Conclusions and Future Research

This systematic review focused on how smartphones are used for hand function assessment. It covered the evaluation and measurement of hand dysfunction caused by various diseases, different embedded smartphone sensors, and statistical and artificial intelligence methods for hand function assessment. The evidence demonstrated that smartphones could facilitate a convenient, inexpensive, and reliable hand-functional assessment [9,10,44]. Future research could (1) explore how to develop a gold standard for smartphone-based hand function assessment; (2) take advantage of smartphones' integrated systems with multiple sensors to collect patients' data in various dimensions to assess hand function holistically; and (3) develop ML methods that are more suitable for processing data collected by smartphones. On the basis of the growing capabilities of smartphones for data collection and analysis, digital technology holds promise for bringing revolutionary changes to hand function assessment.

Acknowledgments

This work was performed with close collaboration among researchers affiliated with the University of Toronto and Huazhong University of Science and Technology Collaborative Center for Robotics and Eldercare. The authors thank the rehabilitation

XSL•FO
RenderX

professionals at the Hubei Provincial Hospital of Traditional Chinese Medicine. This study was supported by the National Natural Science Foundation of China (71771098).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the conception, design, and methodology of the study and approved the protocol. JB was responsible for overseeing the search of databases and literature. YZ and YF handled the management of database and deduplication of records. YZ, YF, and BY were involved in the screening of citations and data extraction. YZ was responsible for software use, formal analysis, investigation, writing the original draft, reviewing, editing, and visualization. YF and BY were responsible for writing the original draft, supervision, and project administration. YZ, ZG, and AM were responsible for conceptualization, writing, reviewing, and editing. All authors provided support in revising and formatting the manuscript. All authors have provided final approval of the version of the manuscript submitted for publication, and all authors agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOC File , 92 KB - jmir_v26i1e51564_app1.doc]

Multimedia Appendix 2
Search strategy.
[DOCX File , 23 KB - jmir_v26i1e51564_app2.docx]

Multimedia Appendix 3
Mixed Methods Appraisal Tool matrix.
[XLS File (Microsoft Excel File), 185 KB - jmir_v26i1e51564_app3.xls]

References

1. Hathaliya JJ, Modi H, Gupta R, Tanwar S, Sharma P, Sharma R. Parkinson and essential tremor classification to identify the patient’s risk based on tremor severity. Comput Electr Eng 2022 Jul;101:107946 [FREE Full text] [doi: 10.1016/j.compeleceng.2022.107946]
2. Moral-Munoz JA, Zhang W, Cobo MJ, Herrera-Viedma E, Kaber DB. Smartphone-based systems for physical rehabilitation applications: a systematic review. Assist Technol 2021 Jul 04;33(4):223-236 [FREE Full text] [doi: 10.1080/10400435.2019.1611676] [Medline: 31112461]
3. Fowler NK, Nicol AC. Functional and biomechanical assessment of the normal and rheumatoid hand. Clin Biomech (Bristol, Avon) 2001 Oct;16(8):660-666 [FREE Full text] [doi: 10.1016/s0268-0033(01)00057-2] [Medline: 11535347]
4. Fiems CL, Miller SA, Buchanan N, Knowles E, Larson E, Snow R, et al. Does a sway-based mobile application predict future falls in people with Parkinson disease? Arch Phys Med Rehabil 2020 Mar;101(3):472-478 [FREE Full text] [doi: 10.1016/j.apmr.2019.09.013] [Medline: 31669299]
5. Yang K, Xiong WX, Liu FT, Sun YM, Luo S, Ding ZT, et al. Objective and quantitative assessment of motor function in Parkinson's disease-from the perspective of practical applications. Ann Transl Med 2016 Mar;4(5):90 [FREE Full text] [doi: 10.21037/atm.2016.03.09] [Medline: 27047949]
6. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord 2018 Aug 27;33(8):1287-1297 [FREE Full text] [doi: 10.1002/mds.27376] [Medline: 29701258]
7. Zwar N, Harris M, Griffiths R, Roland M, Dennis S, Powell Davies G, et al. A systematic review of chronic disease management. Australian Primary Health Care Research Institute. 2006. URL: https://unsworks.unsw.edu.au/entities/publication/9a36a75a-ba4b-44c0-a5d5-91ace271b0ad [accessed 2024-04-29]
8. Zuo KJ, Guo D, Rao J. Mobile teledermatology: a promising future in clinical practice. J Cutan Med Surg 2013 Nov 01;17(6):387-391. [doi: 10.2310/7750.2013.13030] [Medline: 24138974]
9. Lee W, Evans A, Williams DR. Validation of a smartphone application measuring motor function in Parkinson's disease. J Parkinsons Dis 2016 Apr 02;6(2):371-382. [doi: 10.3233/JPD-150708] [Medline: 27061062]

XSL•FO
RenderX

10. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, Kotsavasiloglou C. A smartphone-based tool for assessing Parkinsonian hand tremor. IEEE J Biomed Health Inform 2015 Nov;19(6):1835-1842. [doi: 10.1109/JBHI.2015.2471093] [Medline: 26302523]
11. Williams S, Zhao Z, Hafeez A, Wong DC, Relton SD, Fang H, et al. The discerning eye of computer vision: can it measure Parkinson's finger tap bradykinesia? J Neurol Sci 2020 Sep 15;416:117003 [FREE Full text] [doi: 10.1016/j.jns.2020.117003] [Medline: 32645513]
12. Gopal A, Hsu WY, Allen DD, Bove R. Remote assessments of hand function in neurological disorders: systematic review. JMIR Rehabil Assist Technol 2022 Mar 09;9(1):e33157 [FREE Full text] [doi: 10.2196/33157] [Medline: 35262502]
13. Mourcou Q, Fleury A, Diot B, Franco C, Vuillerme N. Mobile phone-based joint angle measurement for functional assessment and rehabilitation of proprioception. Biomed Res Int 2015;2015:328142 [FREE Full text] [doi: 10.1155/2015/328142] [Medline: 26583101]
14. González-Cañete FJ, Casilari E. A feasibility study of the use of smartwatches in wearable fall detection systems. Sensors (Basel) 2021 Mar 23;21(6):2254 [FREE Full text] [doi: 10.3390/s21062254] [Medline: 33807104]
15. Kheirkhahan M, Nair S, Davoudi A, Rashidi P, Wanigatunga AA, Corbett DB, et al. A smartwatch-based framework for real-time and online assessment and mobility monitoring. J Biomed Inform 2019 Jan;89:29-40 [FREE Full text] [doi: 10.1016/j.jbi.2018.11.003] [Medline: 30414474]
16. Rovini E, Galperti G, Lorenzon L, Radi L, Fiorini L, Cianchetti M, et al. Design of a novel wearable system for healthcare applications: applying the user-centred design approach to SensHand device. Int J Interact Des Manuf 2023 Dec 14;18(1):591-607. [doi: 10.1007/s12008-023-01676-z]
17. Creagh AP, Simillion C, Scotland A, Lipsmeier F, Bernasconi C, Belachew S, et al. Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape test. Physiol Meas 2020 Jun 19;41(5):054002. [doi: 10.1088/1361-6579/ab8771] [Medline: 32259798]
18. Park YM, Kim CH, Lee SJ, Lee MK. Multifunctional hand-held sensor using electronic components embedded in smartphones for quick PCR screening. Biosens Bioelectron 2019 Sep 15;141:111415. [doi: 10.1016/j.bios.2019.111415] [Medline: 31202189]
19. Talwar Y, Karthikeyan S, Bindra N, Medhi B. Smartphone - a user-friendly device to deliver affordable healthcare - a practical paradigm. J Health Med Inform 2016;7(3):1-7 [FREE Full text] [doi: 10.4172/2157-7420.1000232]
20. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019;2:118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
21. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Movement Disorder Society UPDRS Revision Task Force. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008 Nov 15;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
22. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/efi-180221]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Miyake K, Mori H, Matsuma S, Kimura C, Izumoto M, Nakaoka H, et al. A new method measurement for finger range of motion using a smartphone. J Plast Surg Hand Surg 2020 Apr 24;54(4):207-214 [FREE Full text] [doi: 10.1080/2000656x.2020.1755296]
25. Bercht D, Boisvert T, Lowe J, Stearns K, Ganz A. ARhT: a portable hand therapy system. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:264-267. [doi: 10.1109/EMBC.2012.6345920] [Medline: 23365881]
26. Matera G, Boonyasirikool C, Saggini R, Pozzi A, Pegoli L. The new smartphone application for wrist rehabilitation. J Hand Surg Asian-Pac Vol 2016 Feb 16;21(01):2-7 [FREE Full text] [doi: 10.1142/s2424835516400014]
27. Ge M, Chen J, Zhu ZJ, Shi P, Yin LR, Xia L. Wrist ROM measurements using smartphone photography: reliability and validity. Hand Surg Rehabil 2020 Sep;39(4):261-264 [FREE Full text] [doi: 10.1016/j.hansur.2020.02.004] [Medline: 32171926]
28. Pan D, Dhall R, Lieberman A, Petitti DB. A mobile cloud-based Parkinson's disease assessment system for home-based monitoring. JMIR Mhealth Uhealth 2015 Mar 26;3(1):e29 [FREE Full text] [doi: 10.2196/mhealth.3956] [Medline: 25830687]
29. Reed M, Rampono B, Turner W, Harsanyi A, Lim A, Paramalingam S, et al. A multicentre validation study of a smartphone application to screen hand arthritis. BMC Musculoskelet Disord 2022 May 09;23(1):433 [FREE Full text] [doi: 10.1186/s12891-022-05376-9] [Medline: 35534813]
30. Koyama T, Sato S, Toriumi M, Watanabe T, Nimura A, Okawa A, et al. A screening method using anomaly detection on a smartphone for patients with carpal tunnel syndrome: diagnostic case-control study. JMIR Mhealth Uhealth 2021 Mar 14;9(3):e26320 [FREE Full text] [doi: 10.2196/26320] [Medline: 33714936]

XSL•FO
RenderX

31. Williams S, Fang H, Relton SD, Wong DC, Alam T, Alty JE. Accuracy of smartphone video for contactless measurement of hand tremor frequency. Mov Disord Clin Pract 2021 Jan;8(1):69-75 [FREE Full text] [doi: 10.1002/mdc3.13119] [Medline: 34853806]
32. Sarwat H, Sarwat H, Maged SA, Emara TH, Elbokl AM, Awad MI. Design of a data glove for assessment of hand performance using supervised machine learning. Sensors (Basel) 2021 Oct 20;21(21):6948 [FREE Full text] [doi: 10.3390/s21216948] [Medline: 34770255]
33. Kassavetis P, Saifee TA, Roussos G, Drougkas L, Kojovic M, Rothwell JC, et al. Developing a tool for remote digital assessment of Parkinson's disease. Mov Disord Clin Pract 2015;3(1):59-64 [FREE Full text] [doi: 10.1002/mdc3.12239] [Medline: 30363542]
34. Espinoza F, Le Blay P, Coulon D, Lieu S, Munro J, Jorgensen C, et al. Handgrip strength measured by a dynamometer connected to a smartphone: a new applied health technology solution for the self-assessment of rheumatoid arthritis disease activity. Rheumatology (Oxford) 2016 May;55(5):897-901 [FREE Full text] [doi: 10.1093/rheumatology/kew006] [Medline: 26867731]
35. Chen J, Xian Zhang AI, Jia Qian SI, Jing Wang YU. Measurement of finger joint motion after flexor tendon repair: smartphone photography compared with traditional goniometry. J Hand Surg Eur Vol 2021 Oct;46(8):825-829 [FREE Full text] [doi: 10.1177/1753193421991062] [Medline: 33557680]
36. Surangsrirat D, Sri-Iesaranusorn P, Chaiyaroj A, Vateekul P, Bhidayasiri R. Parkinson's disease severity clustering based on tapping activity on mobile device. Sci Rep 2022 Feb 24;12(1):3142 [FREE Full text] [doi: 10.1038/s41598-022-06572-2] [Medline: 35210451]
37. Wang HP, Guo AW, Bi ZY, Zhou YX, Wang ZG, Lu XY. A novel distributed functional electrical stimulation and assessment system for hand movements using wearable technology. In: Proceedings of the 2016 IEEE Biomedical Circuits and Systems Conference. 2016 Presented at: BioCAS '16; October 17-19, 2016; Shanghai, Chaina p. 74-77 URL: https://ieeexplore.ieee.org/document/7833728 [doi: 10.1109/BioCAS.2016.7833728]
38. Lee H, St Louis K, Fowler JR. Accuracy and reliability of visual inspection and smartphone applications for measuring finger range of motion. Orthopedics 2018 Mar 01;41(2):e217-e221 [FREE Full text] [doi: 10.3928/01477447-20180103-02] [Medline: 29309716]
39. Janarthanan V, Assad-Uz-Zaman MD, Rahman MH, McGonigle E, Wang I. Design and development of a sensored glove for home-based rehabilitation. J Hand Ther 2020;33(2):209-219 [FREE Full text] [doi: 10.1016/j.jht.2020.03.023] [Medline: 32451172]
40. Porkodi J, Karthik V, Mathunny JJ, Ashokkumar D. Reliability and validity of Angulus- smartphone application for measuring wrist flexion and extension. In: Proceedings of the 3rd International conference on Artificial Intelligence and Signal Processing. 2023 Presented at: AISP '23; March 18-20, 2023; Vijaywada, India p. 1-4 URL: https://ieeexplore.ieee.org/document/10135006 [doi: 10.1109/aisp57993.2023.10135006]
41. Ienaga N, Fujita K, Koyama T, Sasaki T, Sugiura Y, Saito H. Development and user evaluation of a smartphone-based system to assess range of motion of wrist joint. J Hand Surg Glob Online 2022;2(6):339-342 [FREE Full text] [doi: 10.1016/j.jhsg.2020.09.004] [Medline: 33083772]
42. García-Magariño I, Medrano C, Plaza I, Oliván B. A smartphone-based system for detecting hand tremors in unconstrained environments. Pers Ubiquit Comput 2016 Sep 8;20(6):959-971 [FREE Full text] [doi: 10.1007/S00779-016-0956-2]
43. Lee CY, Kang SJ, Hong SK, Ma HI, Lee U, Kim YJ. A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease. PLoS One 2016;11(7):e0158852 [FREE Full text] [doi: 10.1371/journal.pone.0158852] [Medline: 27467066]
44. Iakovakis D, Diniz JA, Trivedi D, Chaudhuri RK, Hadjileontiadis LJ, Hadjidimitriou S, et al. Early Parkinson's disease detection via touchscreen typing analysis using convolutional neural networks. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:3535-3538. [doi: 10.1109/EMBC.2019.8857211] [Medline: 31946641]
45. Sandison M, Phan K, Casas R, Nguyen L, Lum M, Pergami-Peries M, et al. HandMATE: wearable robotic hand exoskeleton and integrated android app for at home stroke rehabilitation. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:4867-4872 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9175332] [Medline: 33019080]
46. Halic T, Kockara S, Demirel D, Willey M, Eichelberger K. MoMiReS: mobile mixed reality system for physical and occupational therapies for hand and wrist ailments. In: Proceedings of the 2014 IEEE Innovations in Technology Conference. 2014 Presented at: InnoTek '14; May 16, 2014; Warwick, RI p. 1-6 URL: https://ieeexplore.ieee.org/document/6877376 [doi: 10.1109/innotek.2014.6877376]
47. Modest J, Clair B, DeMasi R, Meulenaere S, Howley A, Aubin M, et al. Self-measured wrist range of motion by wrist-injured and wrist-healthy study participants using a built-in iPhone feature as compared with a universal goniometer. J Hand Ther 2019;32(4):507-514 [FREE Full text] [doi: 10.1016/j.jht.2018.03.004] [Medline: 30017418]
48. Tian F, Fan X, Fan J, Zhu Y, Gao J, Wang D, et al. What can gestures tell?: detecting motor impairment in early Parkinson's from common touch gestural interactions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK p. 1-14 URL: https://dl.acm.org/doi/10.1145/3290605.3300313 [doi: 10.1145/3290605.3300313]

XSL•FO
RenderX

49. Gu F, Fan J, Cai C, Wang Z, Liu X, Yang J, et al. Automatic detection of abnormal hand gestures in patients with radial, ulnar, or median nerve injury using hand pose estimation. Front Neurol 2022;13:1052505 [FREE Full text] [doi: 10.3389/fneur.2022.1052505] [Medline: 36570469]
50. Prince J, Arora S, de Vos M. Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. Physiol Meas 2018 Apr 26;39(4):044005 [FREE Full text] [doi: 10.1088/1361-6579/aab512] [Medline: 29516871]
51. Chén OY, Lipsmeier F, Phan H, Prince J, Taylor KI, Gossens C, et al. Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. IEEE Trans Biomed Eng 2020 Dec;67(12):3491-3500 [FREE Full text] [doi: 10.1109/tbme.2020.2988942]
52. Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. Parkinsonism Relat Disord 2015 Jun;21(6):650-653 [FREE Full text] [doi: 10.1016/j.parkreldis.2015.02.026] [Medline: 25819808]
53. Williams S, Relton SD, Fang H, Alty J, Qahwaji R, Graham CD, et al. Supervised classification of bradykinesia in Parkinson's disease from smartphone videos. Artif Intell Med 2020 Nov;110:101966 [FREE Full text] [doi: 10.1016/j.artmed.2020.101966] [Medline: 33250146]
54. Prince J, de Vos M. A deep learning framework for the remote detection of Parkinson'S disease using smart-phone sensor data. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:3144-3147. [doi: 10.1109/EMBC.2018.8512972] [Medline: 30441061]
55. Lee U, Kang SJ, Choi JH, Kim YJ, Ma HI. Mobile application of finger tapping task assessment for early diagnosis of Parkinson's disease. Electron Lett 2016 Nov;52(24):1976-1978 [FREE Full text] [doi: 10.1049/el.2016.3113]
56. Mousavi SA, Abdulrazzaq MH, Hasan MA, Naghavizadeh M. Diagnosis of hand tremor using a smart phone accelerometer and SVM. In: Proceedings of the 4th International Symposium on Multidisciplinary Studies and Innovative Technologies. 2020 Presented at: ISMSIT '20; October 22-24, 2020; Istanbul, Turkey p. 1-4 URL: https://ieeexplore.ieee.org/document/9254969 [doi: 10.1109/ismsit50672.2020.9254969]
57. Akhbardeh F, Vasefi F, Tavakolian K, Bradley D, Fazel-Rezai R. Toward development of mobile application for hand arthritis screening. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:7075-7078. [doi: 10.1109/EMBC.2015.7320022] [Medline: 26737922]
58. Hidayat AA, Arief Z, Happyanto DC. Mobile application with simple moving average filtering for monitoring finger muscles therapy of post-stroke people. In: Proceedings of the 2015 Conference on International Electronics Symposium. 2015 Presented at: ELECSYM '15; September 29-30, 2015; Surabaya, Indonesia p. 1-6 URL: https://ieeexplore.ieee.org/abstract/document/7380803 [doi: 10.1109/elecsym.2015.7380803]
59. Lendner N, Wells E, Lavi I, Kwok YY, Ho PC, Wollstein R. Utility of the iPhone 4 Gyroscope application in the measurement of wrist motion. Hand (N Y) 2019 May;14(3):352-356 [FREE Full text] [doi: 10.1177/1558944717730604] [Medline: 28918662]
60. Gu F, Fan J, Wang Z, Liu X, Yang J, Zhu Q. Automatic range of motion measurement via smartphone images for telemedicine examination of the hand. Sci Prog 2023;106(1):368504231152740 [FREE Full text] [doi: 10.1177/00368504231152740] [Medline: 36721870]
61. Orozco-Arroyave JR, Vásquez-Correa JC, Klumpp P, Pérez-Toro PA, Escobar-Grisales D, Roth N, et al. Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement. Neurodegener Dis Manag 2020 Jun;10(3):137-157 [FREE Full text] [doi: 10.2217/nmt-2019-0037] [Medline: 32571150]
62. Arroyo-Gallego T, Ledesma-Carbayo MJ, Sanchez-Ferro A, Butterworth I, Mendoza CS, Matarazzo M, et al. Detection of motor impairment in Parkinson's disease via mobile touchscreen typing. IEEE Trans Biomed Eng 2017 Sep;64(9):1994-2002 [FREE Full text] [doi: 10.1109/tbme.2017.2664802]
63. Pratap A, Grant D, Vegesna A, Tummalacherla M, Cohan S, Deshpande C, et al. Evaluating the utility of smartphone-based sensor assessments in persons with multiple sclerosis in the real-world using an app (elevateMS): observational, prospective pilot digital health study. JMIR Mhealth Uhealth 2020 Oct 27;8(10):e22108 [FREE Full text] [doi: 10.2196/22108] [Medline: 33107827]
64. Waddell EM, Dinesh K, Spear K, Elson MJ, Wagner E, Curtis MJ, et al. GEORGE®: a pilot study of a smartphone application for Huntington's disease. J Huntingt Dis 2021 Jun 09;10(2):293-301 [FREE Full text] [doi: 10.3233/jhd-200452]
65. Santos C, Pauchard N, Guilloteau A. Reliability assessment of measuring active wrist pronation and supination range of motion with a smartphone. Hand Surg Rehabil 2017 Oct;36(5):338-345 [FREE Full text] [doi: 10.1016/j.hansur.2017.06.005] [Medline: 28754335]
66. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord 2018 Aug;33(8):1287-1297 [FREE Full text] [doi: 10.1002/mds.27376] [Medline: 29701258]
67. Pratt AL, Ball C. What are we measuring? A critique of range of motion methods currently in use for Dupuytren's disease and recommendations for practice. BMC Musculoskelet Disord 2016 Jan 13;17:20 [FREE Full text] [doi: 10.1186/s12891-016-0884-3] [Medline: 26762197]

XSL•FO
RenderX

68. Lenka A, Jankovic J. Tremor syndromes: an updated review. Front Neurol 2021 Jul 26;12:684835 [FREE Full text] [doi: 10.3389/fneur.2021.684835] [Medline: 34381412]
69. Bologna M, Paparella G, Fasano A, Hallett M, Berardelli A. Evolving concepts on bradykinesia. Brain 2020 Mar 01;143(3):727-750 [FREE Full text] [doi: 10.1093/brain/awz344] [Medline: 31834375]
70. West-Higgins T. Improving reading through fine motor skill development in first grade. Dominican University of California. URL: https://scholar.dominican.edu/masters-theses/343, [accessed 2024-04-29]
71. Sartori R, Romanello V, Sandri M. Mechanisms of muscle atrophy and hypertrophy: implications in health and disease. Nat Commun 2021 Jan 12;12(1):330 [FREE Full text] [doi: 10.1038/s41467-020-20123-1] [Medline: 33436614]
72. Theis KA, Steinweg A, Helmick CG, Courtney-Long E, Bolen JA, Lee R. Which one? What kind? How many? Types, causes, and prevalence of disability among U.S. adults. Disabil Health J 2019 Jul;12(3):411-421. [doi: 10.1016/j.dhjo.2019.03.001] [Medline: 31000498]
73. Mennella C, Alloisio S, Novellino A, Viti F. Characteristics and applications of technology-aided hand functional assessment: a systematic review. Sensors (Basel) 2021 Dec 28;22(1):199 [FREE Full text] [doi: 10.3390/s22010199] [Medline: 35009742]
74. Garcia-Agundez A, Eickhoff C. Towards objective quantification of hand tremors and bradykinesia using contactless sensors: a systematic review. Front Aging Neurosci 2021;13:716102 [FREE Full text] [doi: 10.3389/fnagi.2021.716102] [Medline: 34759810]
75. Keogh JW, Cox A, Anderson S, Liew B, Olsen A, Schram B, et al. Reliability and validity of clinically accessible smartphone applications to measure joint range of motion: a systematic review. PLoS One 2019;14(5):e0215806 [FREE Full text] [doi: 10.1371/journal.pone.0215806] [Medline: 31067247]
76. Theile H, Walsh S, Scougall P, Ryan D, Chopra S. Smartphone goniometer for reliable and convenient measurement of finger range of motion: a comparative study. Australas J Plast Surg 2022 Sep 29;5(2):37-43. [doi: 10.34239/ajops.v5n2.335]
77. Serra-Añó P, Pedrero-Sánchez JF, Inglés M, Aguilar-Rodríguez M, Vargas-Villanueva I, López-Pascual J. Assessment of functional activities in individuals with Parkinson's disease using a simple and reliable smartphone-based procedure. Int J Environ Res Public Health 2020 Jun 09;17(11):4123 [FREE Full text] [doi: 10.3390/ijerph17114123] [Medline: 32527031]
78. Tong HL, Maher C, Parker K, Pham TD, Neves AL, Riordan B, et al. The use of mobile apps and fitness trackers to promote healthy behaviors during COVID-19: a cross-sectional survey. PLOS Digit Health 2022 Aug 18;1(8):e0000087 [FREE Full text] [doi: 10.1371/journal.pdig.0000087] [Medline: 36812578]
79. Ernsting C, Dombrowski SU, Oedekoven M, O Sullivan JL, Kanzler M, Kuhlmey A, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017 Apr 05;19(4):e101 [FREE Full text] [doi: 10.2196/jmir.6838] [Medline: 28381394]
80. Johnson SA, Karas M, Burke KM, Straczkiewicz M, Scheier ZA, Clark AP, et al. Wearable device and smartphone data quantify ALS progression and may provide novel outcome measures. NPJ Digit Med 2023 Mar 06;6(1):34 [FREE Full text] [doi: 10.1038/s41746-023-00778-y] [Medline: 36879025]
81. Weizenbaum EL, Fulford D, Torous J, Pinsky E, Kolachalama VB, Cronin-Golomb A. Smartphone-based neuropsychological assessment in Parkinson's disease: feasibility, validity, and contextually driven variability in cognition. J Int Neuropsychol Soc 2021 May 17;28(4):401-413. [doi: 10.1017/s1355617721000503]
82. Pronovost PJ, Cole MD, Hughes RM. Remote patient monitoring during COVID-19: an unexpected patient safety benefit. JAMA 2022 Mar 22;327(12):1125-1126. [doi: 10.1001/jama.2022.2040] [Medline: 35212725]
83. Duruöz MT. Hand Function: A Practical Guide to Assessment. Cham, Switzerland: Springer; 2014.
84. Krishnan G. Telerehabilitation: an overview. Telehealth Med Today 2021 Nov 30;6(4):1-14 [FREE Full text] [doi: 10.30953/tmt.v6.285]
85. Li C, Cheng L, Yang H, Zou Y, Huang F. An automatic rehabilitation assessment system for hand function based on leap motion and ensemble learning. Cybern Syst 2020 Oct 06;52(1):3-25 [FREE Full text] [doi: 10.1080/01969722.2020.1827798]
86. Govindu A, Palwe S. Early detection of Parkinson's disease using machine learning. Procedia Comput Sci 2023;218:249-261. [doi: 10.1016/j.procs.2023.01.007]
87. Alfalahi H, Khandoker AH, Chowdhury N, Iakovakis D, Dias SB, Chaudhuri KR, et al. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. Sci Rep 2022 May 11;12(1):7690 [FREE Full text] [doi: 10.1038/s41598-022-11865-7] [Medline: 35546606]
88. Goni M, Eickhoff SB, Far MS, Patil KR, Dukart J. Smartphone-based digital biomarkers for Parkinson's disease in a remotely-administered setting. IEEE Access 2022;10:28361-28384. [doi: 10.1109/access.2022.3156659]
89. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/ wearable devices opportunity. NPJ Digit Med 2019 Feb 21;2(1):9 [FREE Full text] [doi: 10.1038/s41746-019-0084-2] [Medline: 31119198]
90. Mohamed T. Digital biomarkers provide a way for doctors and patients to work collaboratively at a distance. URGENT Matter 2023:10 [FREE Full text]
91. Ford E, Milne R, Curlewis K. Ethical issues when using digital biomarkers and artificial intelligence for the early detection of dementia. Wiley Interdiscip Rev Data Min Knowl Discov 2023 Feb 19;13(3):e1492 [FREE Full text] [doi: 10.1002/widm.1492] [Medline: 38439952]
92. Park CS. Examination of smartphone dependence: functionally and existentially dependent behavior on the smartphone. Comput Human Behav 2019 Apr;93:123-128. [doi: 10.1016/j.chb.2018.12.022]

XSL•FO
RenderX

93. Pape M, Geisler BL, Cornelsen L, Bottel L, Te Wildt BT, Dreier M, et al. A short-term manual for webcam-based telemedicine treatment of internet use disorders. Front Psychiatry 2023 Feb 23;14:1053930 [FREE Full text] [doi: 10.3389/fpsyt.2023.1053930] [Medline: 36911137]
94. Carroll A, Heiser G. An analysis of power consumption in a smartphone. In: Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference. 2010 Presented at: USENIXATC '10; June 23-25, 2010; Boston, MA p. 21 URL: https://dl.acm.org/doi/10.5555/1855840.1855861
95. Tomlinson M, Solomon W, Singh Y, Doherty T, Chopra M, Ijumba P, et al. The use of mobile phones as a data collection tool: a report from a household survey in South Africa. BMC Med Inform Decis Mak 2009 Dec 23;9(1):51 [FREE Full text] [doi: 10.1186/1472-6947-9-51] [Medline: 20030813]
96. Boulos M, Wheeler S, Tavares C, Jones R. How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX. BioMed Eng OnLine 2011;10(1):24 [FREE Full text] [doi: 10.1186/1475-925x-10-24]
97. Morikawa C, Kobayashi M, Satoh M, Kuroda Y, Inomata T, Matsuo H, et al. Image and video processing on mobile devices: a survey. Vis Comput 2021;37(12):2931-2949 [FREE Full text] [doi: 10.1007/s00371-021-02200-8] [Medline: 34177023]
98. Trucano M. Using mobile phones in data collection: opportunities, issues and challenges. World Bank. 2014. URL: https://blogs.worldbank.org/en/education/using-mobile-phones-data-collection-opportunities-issues-and-challenges [accessed 2024-04-29]
99. Padilla-Magaña JF, Peña-Pitarch E, Sánchez-Suarez I, Ticó-Falguera N. Quantitative assessment of hand function in healthy subjects and post-stroke patients with the action research arm test. Sensors (Basel) 2022 May 10;22(10):3604 [FREE Full text] [doi: 10.3390/s22103604] [Medline: 35632013]
100. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, Kotsavasiloglou C. Smartphone-based evaluation of parkinsonian hand tremor: quantitative measurements vs clinical assessment scores. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:906-909. [doi: 10.1109/EMBC.2014.6943738] [Medline: 25570106]
101. Harris B, Regan T, Schueler J, Fields SA. Problematic mobile phone and smartphone use scales: a systematic review. Front Psychol 2020 May 5;11:672 [FREE Full text] [doi: 10.3389/fpsyg.2020.00672] [Medline: 32431636]
102. Bull TP, Dewar AR, Malvey DM, Szalma JL. Considerations for the telehealth systems of tomorrow: an analysis of student perceptions of telehealth technologies. JMIR Med Educ 2016 Jul 08;2(2):e11 [FREE Full text] [doi: 10.2196/mededu.5392] [Medline: 27731865]
103. Ozdalga E, Ozdalga A, Ahuja N. The smartphone in medicine: a review of current and potential use among physicians and students. J Med Internet Res 2012 Sep 27;14(5):e128 [FREE Full text] [doi: 10.2196/jmir.1994] [Medline: 23017375]
104. Wang X, Fu Y, Ye B, Babineau J, Ding Y, Mihailidis A. Technology-based compensation assessment and detection of upper extremity activities of stroke survivors: systematic review. J Med Internet Res 2022 Jun 13;24(6):e34307 [FREE Full text] [doi: 10.2196/34307] [Medline: 35699982]
105. Krzywinski M, Altman N. Multiple linear regression. Nat Methods 2015 Dec;12(12):1103-1104 [FREE Full text] [doi: 10.1038/nmeth.3665] [Medline: 26962577]
106. Verma VK, Verma S. Machine learning applications in healthcare sector: an overview. Mater Today Proc 2022;57:2144-2147 [FREE Full text] [doi: 10.1016/j.matpr.2021.12.101]
107. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998 Jul;13(4):18-28 [FREE Full text] [doi: 10.1109/5254.708428]
108. Noble WS. What is a support vector machine? Nat Biotechnol 2006 Dec;24(12):1565-1567 [FREE Full text] [doi: 10.1038/nbt1206-1565] [Medline: 17160063]

Abbreviations

CTS: carpal tunnel syndrome
DT: decision tree
MDS-UPDRS: Movement Disorder Society of Unified Parkinson's Disease Rating Scale
ML: machine learning
MS: multiple sclerosis
PD: Parkinson disease
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROM: range of motion
RQ: research question
SVM: support vector machine

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.08.23; peer-reviewed by S Okita, KY Hsieh; comments to author 11.01.24; revised version received 05.03.24; accepted 24.07.24; published 16.09.24.

Please cite as:
Fu Y, Zhang Y, Ye B, Babineau J, Zhao Y, Gao Z, Mihailidis A
Smartphone-Based Hand Function Assessment: Systematic Review
J Med Internet Res 2024;26:e51564
URL: https://www.jmir.org/2024/1/e51564
doi:10.2196/51564
PMID:

©Yan Fu, Yuxin Zhang, Bing Ye, Jessica Babineau, Yan Zhao, Zhengke Gao, Alex Mihailidis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Exploring Adaptive Virtual Reality Systems Used in Interventions for Children With Autism Spectrum Disorder: Systematic Review

Luna Maddalon[1], MSc; Maria Eleonora Minissi[1], MSc; Thomas Parsons[2,3], PhD; Amaia Hervas[4], MD, PhD; Mariano Alcaniz[1], PhD

[1]Laboratory of Immersive Neurotechnologies, Institute Human-Tech, Universitat Politècnica de València, Valencia, Spain

[2]Grace Center, Edson College, Arizona State University, Tempe, AZ, United States

[3]Computational Neuropsychology and Simulation, Arizona State University, Tempe, AZ, United States

[4]Child and Adolescent Service, University Hospital Mutua Terrassa, Global Institute of Neurodevelopmental Disorders Integrated Care, University of Barcelona, Barcelona, Spain

Corresponding Author:
Luna Maddalon, MSc
Laboratory of Immersive Neurotechnologies
Institute Human-Tech
Universitat Politècnica de València
Camino de Vera s/n
Polytechnic City of Innovation: Access N – Building 8B – 3rd Floor
Valencia, 46022
Spain
Phone: 34 963877518 ext 67018
Email: lmaddal@upv.edu.es

Abstract

Background: Adaptive systems serve to personalize interventions or training based on the user's needs and performance. The adaptation techniques rely on an underlying engine responsible for processing incoming data and generating tailored responses. Adaptive virtual reality (VR) systems have proven to be efficient in data monitoring and manipulation, as well as in their ability to transfer learning outcomes to the real world. In recent years, there has been significant interest in applying these systems to improve deficits associated with autism spectrum disorder (ASD). This is driven by the heterogeneity of symptoms among the population affected, highlighting the need for early customized interventions that target each individual's specific symptom configuration.

Objective: Recognizing these technology-driven therapeutic tools as efficient solutions, this systematic review aims to explore the application of adaptive VR systems in interventions for young individuals with ASD.

Methods: An extensive search was conducted across 3 different databases—PubMed Central, Scopus, and Web of Science—to identify relevant studies from approximately the past decade. Each author independently screened the included studies to assess the risk of bias. Studies satisfying the following inclusion criteria were selected: (1) the experimental tasks were delivered via a VR system, (2) system adaptation was automated, (3) the VR system was designed for intervention or training of ASD symptoms, (4) participants' ages ranged from 6 to 19 years, (5) the sample included at least 1 group with ASD, and (6) the adaptation strategy was thoroughly explained. Relevant information extracted from the studies included the sample size and mean age, the study's objectives, the skill trained, the implemented device, the adaptive strategy used, the engine techniques, and the signal used to adapt the systems.

Results: Overall, a total of 10 articles were included, involving 129 participants, 76% of whom had ASD. The studies included level switching (7/10, 70%), adaptive feedback strategies (9/10, 90%), and weighing the choice between a machine learning (ML) adaptive engine (3/10, 30%) and a non-ML adaptive engine (8/10, 80%). Adaptation signals ranged from explicit behavioral indicators (6/10, 60%), such as task performance, to implicit biosignals, such as motor movements, eye gaze, speech, and peripheral physiological responses (7/10, 70%).

Conclusions: The findings reveal promising trends in the field, suggesting that automated VR systems leveraging real-time progression level switching and verbal feedback driven by non-ML techniques using explicit or, better yet, implicit signal

XSL•FO
RenderX

processing have the potential to enhance interventions for young individuals with ASD. The limitations discussed mainly stem from the fact that no technological or automated tools were used to handle data, potentially introducing bias due to human error.

(J Med Internet Res 2024;26:e57093) doi:10.2196/57093

KEYWORDS

adaptive system; virtual reality; autism spectrum disorder; intervention; training; children; machine learning; biosignal

Introduction

Background

Recent research on assessment, training, and intervention applied to technologies has focused on creating complex adaptive systems [1-5]. According to Almirall et al [6], an adaptive intervention is characterized by a set of clinical decision rules that offer guidance on when and how to adjust the dosage and nature of the treatment, considering specific measures. In this way, adaptivity can be defined as the system's capability to alter its actions in response to the preferences and needs of the user [4,7]. Moreover, adaptive systems have the potential to enhance the individualized training experience and prevent issues such as overtraining, undertraining, cognitive overload, frustration, and boredom [5]. This differs from the nonadaptive approach, where the same settings are used throughout training, or adjustment is based on settings that are unrelated to the participant's performance [8].

Technologies, such as robots, mobile phones, and screen-based systems, provide controlled and engaging environments that facilitate adaptive training and support the development of multiple interaction abilities in a secure and predictable manner [9]. The conventional computer is the predominant hardware choice for adaptive and personalized systems [3,4] because adaptive systems have been built upon existing development tools and infrastructure designed for traditional computers and devices. This approach aims to streamline the development process, ultimately reducing the human effort and time required for implementation. In contrast with traditional computer-based therapies, virtual reality (VR) excels in promoting ecological validity by delivering immersive, lifelike experiences, thus creating a strong sense of presence and facilitating the transfer of learning outcomes to real-world situations [10-13]. VR is intended as a computer-generated simulation of an environment that allows users to interact with and experience an artificial world as if it were real. VR aims to create a sense of presence, where users feel as though they are physically present within the virtual environment, enabling them to explore the simulated space, manipulate objects, and engage with the surroundings in a lifelike manner [12]. Indeed, VR technology can achieve varying degrees of immersion, categorized into 3 levels: nonimmersive, semi-immersive, and fully immersive (depending on the capabilities of the device being used). VR offers various avenues for adaptation, including adjusting the complexity of content, tailoring evaluations, and modifying autonomous virtual agents [14]. Another form of adaptation could involve integrating system input, such as using voice commands and haptic feedback. Moreover, VR not only enables the recording of real-time information but also facilitates the integration of data collected from various devices, each dedicated to monitoring distinct psychophysiological activities [15]. Several studies have yielded significant training findings on implementing adaptive VR interfaces. Among them are interventions related to mental health or neuropsychiatric conditions, such as emotional and affective training [16], treatment of phobias [17], management of pathological stress [18], and therapy for posttraumatic stress disorder [19].

In recent decades, research has focused on using VR in the assessment, treatment, and training of neurodevelopmental disorders such as autism spectrum disorder (ASD) [13,20-25]. Studies increasingly focus on this disease due to the escalating worldwide incidence and the high demand for early interventions [26]. Moreover, extensive research has demonstrated that training with VR can lead to notable enhancements in various domains among the young population with ASD [14,21,23-25]. ASD is a neurodevelopmental condition characterized by impairments in social communication and the presence of restrictive and repetitive behaviors [27]. One distinguishing aspect of ASD is its spectrum nature, indicating significant variability in symptom severity among the individuals affected. This heterogeneity results in clinical phenotypes that differ substantially from one person to another while sharing common underlying features. The heterogeneity in symptom severity observed in ASD requires the implementation of personalized treatment approaches that target each individual's specific symptom configuration. By designing early interventions tailored to each child's characteristics, it becomes possible to provide targeted training and improvements in deficits commonly associated with ASD [14,28]. In this connection, adaptive VR technologies seem to have the potential to pave the path for a new generation of highly efficient technology-driven therapeutic tools for the young population with ASD.

Following this approach, a variety of ASD adaptive technologies have been proposed. Among them, Bian et al [29] presented a VR training system for improving driving skills that autonomously adapted its difficulty levels according to participants' engagement and performance metrics. Another adaptive system was applied to a VR job interview training platform, which dynamically adjusted the conversation according to users' responses and stress levels [30]. Research has demonstrated that using such a solution for rehabilitation yields favorable outcomes in enhancing certain abilities of children with ASD; specifically, the VR system was able to tailor the intervention in accordance with the user's actions [31].

Given the evidence demonstrating the effectiveness of using adaptive VR systems to address deficits associated with ASD and the showcased advantages of early intervention, there arises a need to examine the technical aspects of system adaptation and the complexities of handling user data. To do this, it is

XSL•FO
RenderX

crucial to dissect the components of an adaptive system generally used to train or treat ASD symptoms. The following taxonomy will facilitate the reading of this work.

Adaptive Strategy: Level Switching, Feedback, and Time

A system can be adapted through different strategies, such as level-switching techniques or feedback. Level switching follows a logic of level difficulty, and the choice of switching can be based on progression or regression techniques. The progression technique can be considered the core training principle because it is necessary to increase the difficulty level to continue training certain skills, improve them, and prevent the occurrence of learning effects. Therefore, in this scenario, training is adapted through a gradual increase in difficulty level based on the individual's abilities. Unlike progression, the regression technique allows for a finer, more flexible, and more nuanced approach, adapting the difficulty level according to the arising needs. Rooted in traditional therapy practices, this strategy allows for dynamic adjustments that increase or decrease the difficulty level based on the observed performance of the individual [8]. This would ensure that the individual is able to perform satisfactorily at an easier level before increasing the difficulty level, allowing more time to learn the content [8].

Conversely, feedback consists of a system response transmitted visually, audibly, verbally, or haptically to enhance user understanding. Both strategies can occur in 2 temporal dimensions: real time or deferred time. Real-time adaptation occurs based on real-time measures during training, while deferred adaptation happens before the training or session begins, and it depends on the trainee's prior knowledge, level, preferences, or experience [5]. Underlying these, an engine is responsible for processing the information it receives (refer to the Signals: Explicit and Implicit subsection) to transmit an ad hoc response regarding level switching and feedback. The role of the engine can be performed either by a professional, resulting in person-automatized technologies; or directly by the system itself, resulting in system-automatized technologies.

The Adaptive Engine: Person-Automatized Technologies and System-Automatized Technologies

Within the engine's framework, it is defined as person automatized when a professional is in charge of analyzing the engine input signals and making decisions about the type of adaptation needed according to each case. Often, clinicians adapt the interaction by observing the user without playing any active role in the virtual environment [28]. Alternatively, users may interact with a virtual environment where the main avatars are controlled by real people who adapt the system responses according to the user [32,33]. Although this is a valid choice for adaptation, the limitation lies in the need for technical expertise to effectively use these systems [32]. In person-automatized systems, a professional makes clinical decisions by integrating behavioral and psychophysiological user data to adapt the system. However, this could result in time and bias problems. Conversely, advanced systems that automate clinical integrations through multicomputer interactions stand out as a promising choice [32].

A cost-effective and streamlined approach involves using a system-automatized adaptive engine [5]. In this configuration, the system autonomously processes input signals and directly makes decisions to tailor the intervention. Thereby, together with reducing the signal processing time and adaptation decision delays, there is a lower risk of errors or attention problems for professionals [9,34]. In this context, non–machine learning (ML) techniques can be strategically used to adapt the engine. These techniques rely upon explicitly defined rules or heuristics to guide decision-making and system adaptation. These rules are typically designed by professionals and are based on a set of predetermined conditions and actions. The process involves analyzing input signals and making adaptive decisions based on the application of rule-based methods, handcrafted algorithms, or expert knowledge [35,36]. Engines can also be designed through ML techniques to make adaptive decisions based on learned patterns from data rather than rigid explicit rules. ML models are trained on stored data to make predictions or decisions. These systems then refine their knowledge through data-driven methodologies, automated learning processes, and continuous improvement mechanisms, enabling them to learn and make strategic adaptive decisions autonomously [29,30].

In recent years, the focus has been moving increasingly toward system-automatized technologies that use real-time adaptive techniques [34,37]. Indeed, these adaptive systems are capable of responding in real time to behaviors with accuracy beyond human capabilities [34]. Real-time responses seem to confer advantages in creating effective and realistic interventions [35,38]. To enhance user experience, comfort, or task efficiency, systems may apply ML techniques to recognize patterns in the data, allowing them to make adaptive decisions in real time [29,39]. This is often achieved by using sensors and algorithms to interpret and analyze the data effectively by leveraging the input signals of the system.

Signals: Explicit and Implicit

Concerning what has been reported about adaptive strategies and their features (level switching, feedback, time, and engine), it is now important to explain what kind of information is processed by the engine to make an adaptive decision. A signal refers to user information or data used as input to the engine that guides the adaptive decision-making. The adaptation process occurs through the recording, analyzing, and interpreting of the collected signals to make informed decisions about how the system should adjust its behavior, settings, or functionality. On the basis of the available information, the underlying goal is to improve the system's performance, efficiency, efficacy, and user experience. These signals can be explicit or implicit in nature, and they play a crucial role in enabling the system to adapt and respond effectively to changing conditions. Only through the analysis of the user's signals can an intervention be adapted according to the user's needs. The signal is considered explicit when it corresponds to visible behaviors, such as verbal responses or task performance. The latter typically involves quantifiable metrics of how effectively a user is performing a specific task, with the purpose of informing the system about the quality or efficiency of the task execution [40,41]. By contrast, the implicit signal (or biosignal) refers to data or information inputs that are not explicitly provided by a

XSL•FO
RenderX

user but are inferred or recorded by the system. These implicit biosignals often include nonverbal behaviors such as eye gaze, motor movements, speech patterns, and peripheral physiological responses [30,36,42,43]. Implicit biosignals are valuable because they can uncover insights not directly visible from the user's overt actions.

Moreover, real-time processing of behavioral, biological, and psychological information is used to continuously assess the user's state and adapt the human-machine interface. Adaptive VR systems establish robust communication with users, detecting their current state and adjusting tasks to support specific behavioral goals.

Aim of the Study

In accordance with the aforementioned points, in recent years, the use of adaptive VR systems as a therapeutic tool for individuals with ASD has garnered increasing attention [20,23-25]. Due to the diverse nature of ASD, which manifests differently in individuals and evolves over the lifespan, intervention requirements differ significantly between pediatric and adult populations. Acknowledging the substantial evidence endorsing the advantages of early intervention [44], this systematic review concentrates specifically on the pediatric population affected by ASD. Therefore, considering the relevance of using adaptive systems in ASD to improve the customization of treatment based on young individuals' characteristics and their progress, this systematic review primarily aims to comprehensively investigate and analyze the existing literature on the application of adaptive VR systems in the context of interventions for children with ASD. Specifically, the goal is to discuss those studies that have implemented adaptive VR systems, analyzing the study objectives, the methodologies chosen for the functioning of the adaptive engine, and the types of signals implemented. This review focuses only on system-automatized adaptive engines due to their convenience compared with the person-automatized ones [5,32]. Therefore, the methodology investigated refers to level switching, feedback, time, the autonomous adaptation of system engines, and the signals processed. In addition, this review aims to identify the current body of research on adaptive VR systems used in interventions for children with ASD, highlight methodological considerations, and propose directions for future research to guide the development and refinement of adaptive VR interventions for young individuals on the autism spectrum.

To our knowledge, there are no systematic reviews or meta-analyses on this topic. The only similar work is the study by Alcañiz et al [9], which, however, did not use a systematic approach for the comprehensive screening and review of existing literature. Furthermore, it not only considered those adaptive interventions in VR but followed a broader gaze aimed at any type of technology.

Methods

Approach

The literature search adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [45], ensuring a rigorous search process for this systematic review. The PRISMA checklist is available in Multimedia Appendix 1.

Literature Search Strategy

A comprehensive search was conducted on March 13, 2024, to identify relevant studies in 3 different databases: PubMed Central, Scopus, and Web of Science. These databases were searched using a combination of keywords and Medical Subject Headings terms related to the research question; in addition, the literature was searched for peer-reviewed articles and conference articles in English, with full texts available, published between January 1, 2013, and March 13, 2024. Considering the recent increase of interest in the field and the continual evolution of hardware and software, a literature review that includes sources published in approximately the past decade would provide an up-to-date overview of the state of the art, ensuring the inclusion of the most advanced technologies and the latest methodologies [23,46,47]. The first author used the following Boolean string to search the databases: ("virtual" OR "VR") AND ("adaptive" OR "personalized" OR "customized" OR "individualized" OR "tailored") AND ("training" OR "intervention" OR "therapy" OR "clinical*") AND ("toddler" OR "children" OR "infant" OR "teen*" OR "young adult*") AND ("autistic spectrum disorder" OR "autism" OR "ASD"). Studies that fulfilled the following inclusion criteria were selected: (1) the experimental tasks were presented through a VR system, (2) the adaptation was system automatized, (3) the VR system was designed for intervention or training of ASD symptoms, (4) participants' ages varied between 6 and 19 years, (5) the sample included at least 1 group with ASD, and (6) the adaptive strategy was comprehensively explicative. Accordingly, the following criteria were used to exclude studies from this systematic review: (1) studies not using a VR system for experimental tasks, (2) absence of automated adaptation mechanisms within the VR system, (3) studies whose objective was not specifically aimed at intervention or training of ASD symptoms, (4) studies involving participants outside the age range of 6 to 19 years or absence of specified age, (5) absence of at least 1 group with ASD in the study sample, and (6) studies failing to provide a sufficient or clear explanation of the adaptive strategy used. Applying these criteria aimed to ensure the selection of studies meeting stringent standards for relevance to the review objectives and rigor in methodology.

The selection of studies was divided into 6 stages, as outlined in Textbox 1.

XSL•FO
RenderX

Textbox 1. Study selection stages.

Stages involved in the selection of studies

- Stage 1: the first author conducted the literature search of the databases and manually removed duplicates. The data were recorded in a template to ensure consistency in the data collection process and organized for subsequent analysis. Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendations to avoid bias and reduce errors, the 5 authors operated independently from the second stage onward.
- Stage 2: the titles and abstracts of the retrieved studies were reviewed to identify potentially eligible ones. Reasons for excluding studies were documented.
- Stage 3: a full-text assessment based on predefined inclusion and exclusion criteria was performed. Any discrepancies between the reviewers were resolved through discussion and consensus. Reasons for excluding studies were documented.
- Stage 4: the quality of the selected studies was assessed using the JBI critical appraisal tool [48] to identify and evaluate the risk of bias. Each article was evaluated according to standardized criteria, such as study design, the robustness of the findings, and potential sources of bias.
- Stage 5: relevant data from each included study were systematically collected using a predefined data extraction form. Relevant information categories extracted from the studies were chosen and defined by the 5 authors following the aim of this systematic review.
- Stage 6: the extracted data were combined and analyzed to provide an overall summary of the evidence. The data synthesis involved a narrative synthesis that was tabulated to facilitate the further interpretation of the findings.

No technological or automated tool was used to manage the collected records. Relevant information extracted from the studies included the sample size and mean age, the objectives of the study, the skill trained, the implemented device, the adaptive strategy used, the engine chosen, and the signal used to adapt the systems. The adaptive strategy findings are classified by the mode used to adapt the intervention, such as through level switching (progression or regression) and feedback, the time when adaptation occurs (real time or deferred time), and the engine executing the adaptation (ML based or non-ML based). Feedback can be verbal, auditive, visual, or haptic. The signals reported in the results refer only to those feeding the adaptive engine. Signals recorded for subsequent analyses that did not refer to the adaptive strategy were excluded.

Ethical Considerations

This study was deemed exempt by the Polytechnic University of Valencia.

Results

Review Flow

The study selection process was documented using a PRISMA flow diagram (Figure 1). The search strategy we used yielded 163 records (PubMed Central: n=67, 41.1%; Scopus: n=70, 42.9%; Web of Science: n=26, 16%). From the 163 articles, 26 (16%) duplicates were removed, leaving 137 (84%) unique articles. Of these 137 articles, 113 (82.5%) were excluded based on title and abstract screening. Of the remaining 24 articles, 14 (58%) were excluded after a full-text screening, which was performed to ensure that the studies met the inclusion criteria. Some of the studies (9/14, 64%) initially seemed to meet the inclusion criteria; however, subsequent analysis revealed discrepancies leading to their exclusion; for instance, some of the studies (3/9, 33% 9/14, 64%%) were excluded due to the implementation of person-automatized systems rather than system-automatized processes as required [32,33], while some of the articles (6/9, 67%) were excluded because they lacked an experimental design [49]. Finally, 10 (6.1%) of the initially identified 163 articles satisfied the inclusion criteria and were thus included in the review. The data collected were analyzed following a systematic approach used to identify and synthesize findings.

XSL•FO
RenderX

Figure 1. Flow diagram of the study selection process.

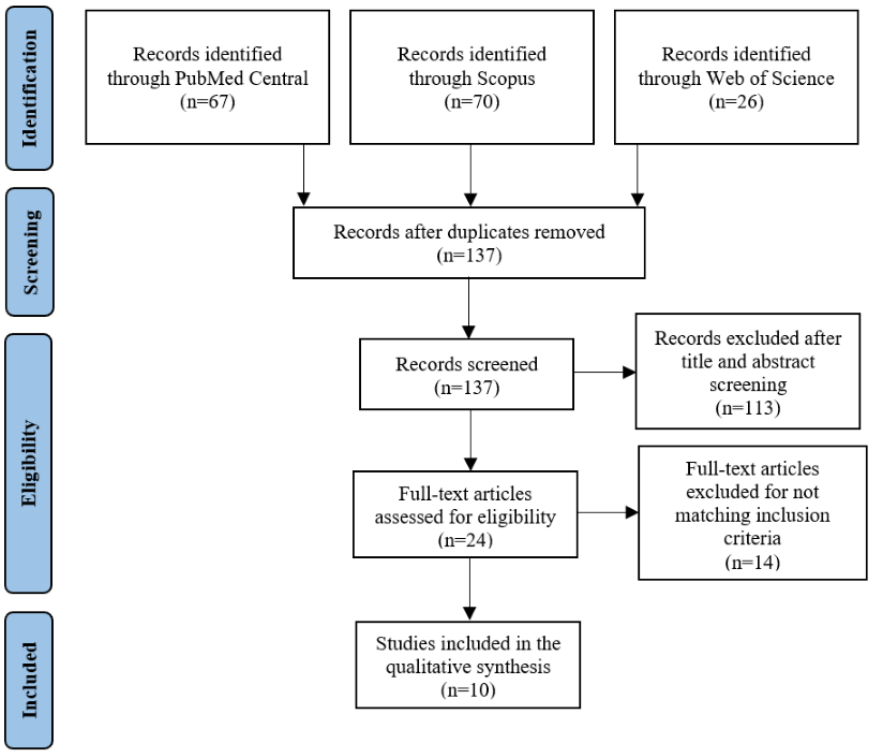

Characteristics of the Selected Studies

The selected studies are presented in Table 1 in alphabetical order. The 10 studies involved 129 participants, of whom 95 (73.6%) were children with ASD. The mean ages of the participants ranged from 8 to 16 years (mean 12.54, SD 2.12 y). With regard to study objectives, of the 10 studies, 6 (60%) were usability studies [36,38,40,42,50,51], 3 (30%) had a further objective aimed at treatment [35,37,41], and 1 (10%) aimed merely at treatment. All included studies designed training through serious games, thus having the purpose of teaching or educating through a game. Most of the studies (7/10, 70%) focused on training social skills [35,36,40,42,43,50,51], while 20% (2/10) concentrated on motor skills [37,38], and 10% (1/10) targeted executive functions [41]. The devices used were desktop computers (8/10, 80%) [35,36,38,40-43,50], a head-mounted display (1/10, 10%) [37], and a tablet computer (1/10, 10%) [51].

Regarding the adaptive strategy, level switching was based either on a progressive strategy (4/10, 40%) [37,41,50,51] or a regressive strategy (3/10, 30%) [36,40,42]. Moreover, the systems provided adaptive feedback through the use of verbal (8/10, 80%) [36,38,40-43,50,51], audio (4/10, 40%) [38,41,50,51], visual (3/10, 30%) [35,38,51], and haptic (1/10, 10%) [38] cues. A real-time adaptive strategy was used in most of the studies (9/10, 90%) [35-38,40-42,50,51], while 20% (2/10) of the studies used a deferred adaptive strategy [41,43]. Finally, regarding the adaptive engine used, either a non-ML–based method was implemented (8/10, 80%) [35,36,38,40-43,50,51], or an ML-based method was applied (3/10, 30%) [37,51]. The number of studies counted does not equal the total number of studies because some of the studies (2/10, 20%) used different adaptive strategies for level switching and feedback, while others (8/10, 80%) used only one of them.

The studies adapted their system through an explicit behavioral signal, such as task performance (6/10, 60%) [36,40-42,50,51]; or an implicit biosignal, such as motor movements (3/10, 30%) [37,38,51], eye gaze (2/10, 20%) [35,42], speech (1/10, 10%) [43], and peripheral physiological responses (1/10, 10%) [36].

XSL•FO
RenderX

Table 1. Characteristics of the selected studies.

Study, year	Sample size (n); age (years), mean (SD)		Study objective	Training focus	Device used	Adaptive strategy				Signal used
	ASD[a]	TD[b]				Level switching	Feedback	Time	Engine	
Bekele et al [35], 2016	6; 15.77 (1.87)	6; 15.20 (1.68)	Treatment, usability	Social skills	Desktop computer	—[c]	Visual	Real time	Non-ML[d] based	Eye gaze
Hocking et al [37], 2022	10; 14.1 (2.6)	—; —	Treatment, usability	Motor skills	HMD[e]	Progression	—	Real time	ML based	Motor movements
Jyoti and Lahiri [50], 2020	20; 8.7 (2.2)	20; 8.7 (2.3)	Usability	Social skills	Desktop computer	Progression	Audio, verbal	Real time	Non-ML based	Task performance
Jyoti and Lahiri [51], 2022	20; 8.61 (2.03)	—; —	Usability	Social skills	Tablet computer	Progression	Audio, visual, verbal	Real time	Level switching: non-ML based; feedback: ML based	Task performance, motor movements
Kuriakose and Lahiri [36], 2016	9; 14.10 (2.6)	—; —	Usability	Social skills	Desktop computer	Regression	Verbal	Real time	Non-ML based	Task performance, peripheral physiological responses
Lahiri et al [42], 2013	8; 16.00 (2.08)	—; —	Usability	Social skills	Desktop computer	Regression	Verbal	Real time	Non-ML based	Task performance, eye gaze
Moon and Ke [43], 2023	4; 12.25 (0.50)	—; —	Treatment	Social skills	Desktop computer	—	Verbal	Deferred time	ML based	Speech patterns
Pradeep Raj and Lahiri [40], 2016	2; 16.25 (3.40)	2; 11.9 (2.7)	Usability	Social skills	Desktop computer	Regression	Verbal	Real time	Non-ML based	Task performance
Vallefuoco et al [41], 2022	10; 11.9 (2.7)	—; —	Treatment, usability	Executive functions	Desktop computer	Progression	Audio, verbal	Level switching: deferred time; feedback: real time	Non-ML based	Task performance
Zhao et al [38], 2018	6; 9.66 (1.36)	6; 10.73 (1.91)	Usability	Motor skills	Desktop computer	—	Audio, visual; verbal; haptic	Real time	Non-ML based	Motor movements

[a]ASD: autism spectrum disorder.

[b]TD: typical development.

[c]Not applicable.

[d]ML: machine learning.

[e]HMD: head-mounted display.

Discussion

Main Findings

The main findings indicate a trend of balanced use among adaptive strategies such as level switching (with slight preferences toward progression techniques) and feedback (with a preference for verbal mode). Furthermore, a widespread use of non-ML techniques was found in engines that used explicit and implicit signals, with a slight tendency toward the latter. Finally, the results showed significant advantages for real-time adaptations.

Overall, the mean age range of the participants in the selected studies (8-16 y) suggests a tendency to prefer a sample composed of adolescents rather than young children. Extended

XSL•FO
RenderX

studies [44,52] highlight the benefits of early intervention (particularly for children younger than adolescents) in achieving significant improvements by targeting the fundamental behavioral and skill deficits associated with ASD. Indeed, younger children tend to make better progress in their treatment, even with lower-intensity programs, than older children [44]. Considering the theoretical foundations and the benefits that early intervention brings to children with ASD, future studies involving adaptive VR systems should consider a younger age sample, as reported in other research fields on ASD treatment, such as interventions using robots [53-55]. In this context, when designing an early intervention, it is crucial to consider pediatric recommendations advocating for a maximum screen time of 1 hour per day for children aged 2 to 5 years because this contributes to establishing a high-quality program [56]. Consequently, the implementation of any technological tools must be carefully crafted and controlled.

Furthermore, concerning study objectives, the findings indicated that most of the studies (9/10, 90%) primarily focused on usability [35-38,40-42,50,51]. This confirms the relatively recent emergence of research in applying adaptive VR systems in ASD interventions [20,23-25]. Indeed, an emerging field must first validate new methods and technologies through usability studies. Moreover, a few of the studies (3/10, 30%) had another clinical goal of validating the designed treatment, which is valuable considering that these are innovative and previously untested adaptive systems [35,37,41]. In addition, the prevalence of training targeting social skills was observed in a majority of the studies (7/10, 70%) [35,36,40,42,43,50,51], aligning with a well-established focus in ASD intervention literature that emphasizes the importance of addressing the social challenges faced by individuals with ASD [57,58].

Another prominent trend observed relates to the device deployed. Notably, among the selected studies, there was a consistent preference for desktop computers and similar devices, such as tablet computers [35,36,38,40-43,50,51], which is consistent with the broader literature on adaptive VR training [5]. The prominent decision to opt for nonimmersive VR devices showcase a practical choice grounded in considerations of accessibility, affordability, and cost-effectiveness, as well as the avoidance of potential discomfort [36,42,50,51,59]. Nevertheless, it is likely that nonimmersive VR systems were chosen because evidence suggests that they could also support individuals with ASD, particularly in addressing comorbidities such as attention-deficit/hyperactivity disorder and anxiety disorders [60,61]. Further research targeting the impact of nonimmersive VR systems on comorbid conditions with ASD would be valuable for elucidating their full potential and applications.

The following subsections are organized as follows: the Adaptive Strategy: Regression, Multimodal, and Real-Time Adaptation subsection delves into findings regarding the adaptive strategy in terms of mode and time; the next subsection, The Adaptive Engine: Non-ML and ML Techniques, delineates the techniques chosen to adapt the engine used; and, finally, the Signals: Embodied Adaptation subsection presents a detailed discussion about the signals involved. Each subsection includes a segment aimed to highlight methodological considerations and guide future research development.

Adaptive Strategy: Regression, Multimodal, and Real-Time Adaptation

The examination of the adaptive systems in the selected studies reveals a critical interplay between level switching and feedback strategies, both essential components for addressing the challenges associated with individualized training. Nevertheless, not all studies implemented both adaptive strategies in their intervention. Considering the level-switching strategy, in studies using progression-based adaptations [37,41,50,51], the difficulty level was automatically increased after sufficient performance, while the same level was repeated after insufficient performance; for example, Jyoti and Lahiri [50] used a performance threshold score of ≥70% as the criterion for increasing the difficulty level. In this application, the performance score was calculated based on the accuracy in recognizing regions of the virtual character's face within the maximum response time allowed [50]. In the study by Hocking et al [37], the performance threshold score was raised to ≥75% to continuously update the challenge. However, performance was quantified by calculating the efficiency, synchrony, and symmetry of the movement performed [37]. Contrastingly, studies that adopted the regression technique [36,40,42] increased the level of difficulty after sufficient performance, while the level remained the same or decreased in the case of semi-insufficient or insufficient performance, respectively. Specifically, Kuriakose and Lahiri [36] designed rules to determine whether the difficulty level should increase (performance score ≥70% and anxiety level <6 units or performance score <70% and anxiety level <6 units), remain the same (performance score ≥70% and anxiety level ≥6 units or performance score <70% and anxiety level <6 units), or decrease (performance score <70% and anxiety level ≥6 units or performance score ≥70% and anxiety level ≥6 units) based on the performance score along with the predicted anxiety level. In the study conducted by Pradeep Raj and Lahiri [40], performance was assessed on cognitive and emotional tasks, with adjustments made solely based on the possibility of increasing the difficulty level when performance was sufficient (≥70%) and decreasing it when performance was not sufficient.

In the training context, feedback plays a pivotal role in guiding and enlightening the system's user regarding their performance, providing responses to both explicit and implicit behaviors. Of the 10 selected studies, 9 (90%) implemented at least 1 feedback mode in their system; the sole exception was the study by Hocking et al [37]. The decision to use one mode over the other seems contingent upon the specific training goals and the nature of the skills being developed. The findings reveal that a majority of the selected studies (8/10, 80%) adopted the verbal feedback mode [36,38,40-43,50,51]. This preference can be attributed to the fact that listening to audio messages or reading written communication allows complex information to be transmitted. Moreover, one-fourth of the studies (4/10, 40%) opted for a combination of feedback modes [38,41,50,51]. In fact, it is well known that multimodal communication is more likely to be received effectively and has a greater informational effect [62]; for example, in the study by Jyoti and Lahiri [51], feedback was provided at the end of each task based on criteria governed by

XSL•FO
RenderX

performance and face orientation. The feedback included verbal, visual (ie, the number of stars), and auditory (ie, clapping hands) components, delivered with positive or negative connotations to reinforce behavior. The exploration of multimodal feedback, as demonstrated by Zhao et al [38], introduces a compelling dimension to the discussion by successfully implementing complex multimodal feedback. The incorporation of 4 different feedback modes—verbal (ie, written message), auditory (ie, crash audio and achievement jingle), visual (ie, reward or warning images), and haptic (ie, friction and spring force)—represents a sophisticated approach to enhancing user engagement [38]. The synergistic effect of multiple sensory modalities, as supported by previous literature [63], suggests that integrating diverse feedback elements can have a cumulative positive impact on the user's learning experience.

This review also delves into the temporal considerations of implementing adaptive strategies. These strategies can be transmitted from the engine to produce an ad hoc adaptive decision either during the session (real time) or between sessions (deferred time). The review's results indicated a preference for adapting the level or the feedback in real time rather than deferred time [35-38,40-42,50,51]. Real-time adaptation seems promising because it aligns with the potential to enhance the training experience by tailoring content in response to the user's performance [5]; for instance, Bekele et al [35] designed a stimuli occlusion paradigm in which the stimuli were revealed gradually based on the eye gaze performance of the participants: the more attentively participants turned their gaze toward the stimulus, the clearer and sharper it became—providing visual feedback. A specific case in the study by Vallefuoco et al [41] involved feedback provided in real time based on the accuracy of participants' actions; here, the difficulty level of the first session was chosen based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, severity levels of ASD symptoms [27], and the difficulty level seemed to progressively increase in line with the game levels and symptom severity. In addition to the duality of real-time (feedback) and deferred (level switching) adaptivity of the 2 strategies, there was further duality inherent in level switching: the initial level was set by the therapist and then automated by the system. Indeed, this study was included in this review because, at first, the system was person automatized and then became system automatized (refer to the next subsection, The Adaptive Engine: Non-ML and ML Techniques), highlighting the advantages of cost-effectiveness and streamlined implementation.

Given these premises, it seems advantageous to introduce at least 1 strategy in real time when designing an adaptive automatized system. Training that uses a scaffolding strategy, which recognizes the need to adjust the level switching up or down as needed (regression technique), has consistently demonstrated better learning outcomes [8]. Similarly, feedback is considered a scaffolding strategy useful for adapting intent and capabilities during the training by promoting critical thinking [64]. Considering the findings, the power of verbal and multimodal feedback [62], and the benefits of in-session administration of the 2 adaptive strategies [5,64], future studies should include progression level switching and real-time verbal feedback in their adaptive interventions. Future research should also carefully consider the interplay between these adaptive strategies and their potential synergistic effects on training efficacy. Practitioners and policy makers involved in designing training interventions should recognize the complementary nature of these strategies and strive to incorporate them synergistically. In conclusion, the comprehensive analysis of adaptive strategies sheds light on their intricate dynamics within individualized training environments. By incorporating these insights into practice frameworks, future researchers can work toward enhancing the effectiveness, accessibility, and scalability of adaptive training interventions across various domains.

The Adaptive Engine: Non-ML and ML Techniques

A system-automatized adaptive engine can independently analyze signals and decide when to apply level switching or feedback to customize the intervention for each individual. In this context, the deployment of an adaptive engine in the reviewed studies confirms the dichotomy in the techniques used for decision-making: non-ML and ML. A significant majority of the reviewed studies (8/10, 80%) lean toward non-ML techniques for adaptive engine design [35,36,38,40-42,50,51]. This approach involves establishing predefined rules, often binary in nature, to determine the appropriateness of the participant's performance. Decisions are made through established cutoffs to determine whether the performance is sufficient in terms of duration, successes, gazed area, accurate movements, and physiological activity levels; for instance, Kuriakose and Lahiri [36] designed an adaptive engine based on the composite effect of physiological index levels (high or low) through the fuzzy logic classification method and the quality of task performance (adequate or inadequate). Similarly, Lahiri et al [42] adapted their VR system through complex fixed rules that blended information from viewing patterns (fixation duration), eye physiology (pupil diameter and blink rate), and task performance. The inflexibility of non-ML techniques is evident in their reliance on a priori knowledge and fixed rules. While effective for binary decisions, the limitation lies in their inability to adapt to more intricate data nuances not easily discernible by human observation. In contrast to non-ML techniques, the primary goal of ML techniques is to predict future observations as accurately as possible and increase efficiency and reproducibility [65,66]. Despite the advantages of adopting ML techniques, a limited number of the selected studies (3/10, 30%) opted for ML techniques to drive adaptive engine decisions [37,43,51]. The use of artificial neural networks by Hocking et al [37] exemplifies the capacity of ML techniques to objectively quantify kinematic features by tracking the biomechanical changes and adapting the challenge level in real time. Specifically, the authors applied a normalized exponential transformation translated into a discrete probability assignment over the potential labels and trained this model until no further enhancements were observed [37]. Similarly, Moon and Ke [43] analyzed participants' speech and used a supervised ML method to categorize the probability values into threshold conditions (from high to low emotional state). Feedback was then tailored based on the individual's threshold condition. The novel integration of ML and non-ML approaches in a single study, as seen in that by Jyoti and Lahiri [51], showcases the potential for a hybrid model. While the task performance level

XSL•FO
RenderX

switching remained non-ML, the use of ML-based feedback signaled a nuanced approach to harnessing the strengths of both techniques [51]. The ML-based feedback relied on image processing using a Haar feature–based cascade frontal face classifier [67] to calculate the percentage of the user's face (gross motor movement) oriented toward the area of interest. However, the method proposed by Viola and Jones [67] is now considered obsolete because it has been replaced with deep learning methods, such as convolutional neural networks [68,69].

Considering the factors discussed, the distinct characteristics of non-ML and ML techniques raise critical considerations for their application in adaptive systems. Non-ML techniques, which are reliant on predetermined rules, demonstrate efficacy in scenarios where binary decision-making aligns with straightforward performance metrics. However, their limited ability to discern subtle patterns in data necessitates caution when confronted with complex and dynamic user interactions that are usually not detectable by humans. Conversely, ML techniques, driven by continuous objective analysis of data sets and pattern recognition, hold promise in offering adaptive systems that improve automatically over time. Considering a growing emphasis on adaptive systems based on ML [5], the ability to replicate outcomes and accurately predict future observations, as highlighted by Orrù et al [65], positions ML as a potent tool to provide more refined and personalized interventions. Future research should carefully weigh the trade-offs between non-ML and ML approaches. This review advocates for increased adoption of ML techniques in future research aiming to enhance ASD interventions.

However, it is notable that among the selected studies that used ML techniques, the samples ranged from 4 to 20 participants [37,43,51]. In ML models, predictions are made by using versatile learning algorithms to find patterns in typically vast data sets; this is why extensive data sets are necessary to yield more accurate algorithm outcomes [65,66]. Acknowledging that larger data sets enhance the generalization performance of ML models [66], the limited sample sizes in these studies raise concerns about the robustness of the ML-based adaptive engines. Future research should prioritize expanding sample sizes to bolster the reliability and effectiveness of ML-driven interventions.

Among other advantages, the adaptive engine, functioning as the autonomous core of the system, exhibits the capability to make real-time decisions. Regardless of the chosen technique (non-ML or ML), an overarching theme in the reviewed studies is the pervasive use of real-time adaptive systems. This aligns with the growing interest in adaptive systems that use ML techniques to recognize patterns in data, enabling real-time responses and an enhanced user experience [34,39]. This is frequently accomplished using sensors and algorithms for effective data interpretation and analysis, harnessing the input signals to optimize the system's performance. It is suggested to implement this approach in future studies to enhance the system's ability to provide meaningful responses based on the gathered dynamic user inputs. Policy makers could support initiatives aimed at advancing sensor technology and algorithm development to facilitate real-time data interpretation and analysis in adaptive systems, which could enhance their effectiveness in clinical settings.

In conclusion, the architectural choices for adaptive engines play a pivotal role in shaping intervention efficacy. The synthesis of real-time adaptability and a preference toward adopting non-ML techniques as well as a consideration of the nuances of sample size in ML-based approaches should guide future endeavors in developing adaptive systems that seamlessly integrate cutting-edge technologies to enhance user experiences in ASD interventions.

Signals: Embodied Adaptation

Given what has been discussed with respect to the adaptive strategies governed by the engine and the techniques used by the engine, it is pertinent to address what drives and prompts these adaptation decisions: signals.

The results of this review showed a similar reliance on explicit and implicit signals, with a slight preference for the latter, in shaping ASD training adaptations. In particular, a third of the studies (3/10, 30%) [40,41,50] exclusively used explicit signals, focusing on behavioral data such as task performance metrics (ie, reaction time and success rates). Another set of studies (3/10, 30%) [36,42,51] adopted a more integrated approach, combining explicit signals (task performance) with implicit biosignals such as gross motor movements, peripheral physiological responses, and eye gaze. Jyoti and Lahiri [50,51] highlighted the importance of adapting training through an implicit measure complementary to performance data. Indeed, in their latest study [51], the authors introduced informative data on the duration of face orientation toward the stimulus presented to infer whether the participant was attending to the task stimulus. Adding implicit biosignals might augment the potential of explicit signals, leading to more valid outcomes. Similarly, Kuriakose and Lahiri [36] adapted their VR system to users' affective states, such as anxiety, through peripheral physiological signals (including galvanic skin response and pulse plethysmogram). Adaptive training based on implicit biosignals has been demonstrated to hold the potential to optimize physiological arousal to maximize training outcomes [5]. Conversely, training programs that fail to adjust to users' affective states may negatively impact performance, leading to either low arousal levels (as in boredom) or high arousal levels (as in anxiety) [29]. To prevent this, it is essential to consider both implicit and explicit biosignals. A combined analysis of users' implicit and explicit biosignals enables refined adaptation decisions based on their needs. During treatment, implicit biosignals can provide valuable insights into the user's inner state, which is often imperceptible to human observation but easily deciphered by computers [36]. Indeed, compared with explicit signals, implicit biosignals emerge as pivotal in providing optimal challenges to users, increasing their engagement and improving their performance [29,30]. The study by Lahiri et al [42] stands out as a pioneering work, in comparison with the other included studies, in anticipating the benefits of adapting training or interventions through implicit biosignals. Indeed, it proposed measuring the engagement level through real-time viewing patterns, subtle eye physiological changes, and performance metrics to adaptively enhance

XSL•FO
RenderX

responses for improved social communication skills. This underscores the clear need to incorporate an implicit biosignal when training or investigating a complex construct.

Recent studies have suggested the integration of advanced technologies (eg, eye gaze monitoring, peripheral response monitoring, motion sensing, and speech input devices) to capture implicit biosignals, thereby steering ASD interventions toward a more intuitive and embodied approach [29,35,40]. As discussed, implicit biosignals unveil hidden meanings and provide insights that extend beyond observable user behaviors. This aligns with the concept of embodied cognition [70], which posits that body and mind are not separate entities but rather maintain a continuous, tight, and ongoing relationship. Embodied cognition theories are based on the concept that every action or reaction is influenced by one's perception of stimuli, typically during goal achievement [70,71]. VR interventions, as embodiments of learning paradigms, represent an illustration of how technology links the mind, body, and environment together and where individuals' perceptions play a crucial role in shaping their bodily experience of the world [72]. The selected studies in this review reflect a growing body of evidence supporting the adaptation of interventions to the unique needs of children with ASD, incorporating devices capable of collecting implicit biosignals of fine [38] or gross [37,51] motor movements, eye gaze [35,42], peripheral physiological responses [36], and speech [43]; for example, Moon and Ke [43] used speech data mining to drive adaptive feedback to identify the emotional states of children.

On the basis of the assumption that embodied cognition theories indicate that cognition involves a continuous interplay of various modules encompassing simulation, environmental, situated action, and bodily state factors [73], implicit biosignals are at the core of the process of cognition because they reflect and explain behaviors that are not directly observable. Thus, future adaptive interventions for children with ASD should go beyond monitoring user performance as a stand-alone datum and adopt an embodied perspective by including implicit aspects, assuming that the arousal component influences the learning experience in training. Cognition rarely proceeds independently of the body; however, it is recognized that many researchers address other forms of cognition assumption [70]. Therefore, the suggestion is to implement the embodied adaptation framework but acknowledge that other adaptations may also be valuable. Practitioners could support initiatives aimed at promoting embodied perspectives in ASD interventions, recognizing their potential to enhance learning experiences and outcomes for individuals with ASD.

Finally, attention should be focused on those included studies (6/10, 60%) that processed implicit biosignals in real time to enhance ASD adaptive interventions [35-38,42,51]. Evidence suggests that VR-automatized systems, which adapt dynamically based on implicit biosignals, are able to foster ongoing communication with users by identifying their present state and aligning activities with specific behavioral objectives [37,38]; for example, sophisticated algorithms have made it possible to quantify and qualify dynamic whole-body motor movement in real time, allowing for the adaptation of motor skills interventions [37]. Even finer motor movements, such as force or precision and control, can be adapted in real time to improve the performance of a kinetic task [38]. In conclusion, the spotlight on studies processing explicit and implicit biosignals in real time underscores their potential to enhance the effectiveness of adaptive interventions for young individuals with ASD. Moreover, as technology continues to advance, the integration of unobtrusive wearable devices that capture implicit biosignals, either independently or in combination with explicit signals, emerges as a promising approach [5]. On the basis of the literature and the reasons discussed, it is recommended to embrace the embodied cognition perspective to comprehensively understand and effectively address the unique needs of children with ASD.

Overall Findings

The literature search showed that there has been a growing emphasis on developing objective adaptive technologies to target the needs of individuals with ASD. This review primarily aimed to shed light on the methodologies of the included studies used for the adaptive engine and the types of signals used. In addition, the review identified methodological insights and offered guidance for improving adaptive VR interventions for individuals on the autism spectrum.

The findings suggested a tendency to focus on adolescent samples, underscoring the need to explore younger age groups in the context of adaptive VR systems. A predominant focus on usability, social skills training, and the use of desktop computers was discussed.

The main findings showed that the use of adaptive strategies varied among studies, with a noteworthy trend toward real-time adaptations. The nuanced application of progression techniques, coupled with the exploration of verbal feedback and real-time adaptations, provides a foundation for future research aimed at optimizing adaptive training systems for therapeutic contexts.

The review highlighted considerations associated with non-ML and ML techniques critical in the engine adaptation. An overarching theme in the reviewed studies is the widespread use of real-time adaptive systems, mostly non-ML, although there has been a growing emphasis over the past 2 years on using ML techniques for refined and enhanced training. Moreover, when choosing an ML engine, a wider sample size should guide the development of future adaptive systems for personalized ASD interventions.

Finally, both explicit and implicit signals were identified as the driving force behind adaptive decisions, with increased attention toward using implicit biosignals for a more comprehensive understanding of user needs. Embracing an embodied adaptation framework is suggested for future ASD interventions, emphasizing real-time processing to optimize adaptation and contribute to improved performance.

By synthesizing and evaluating the existing evidence, this study aims to contribute valuable insights that can inform the design and implementation of more targeted and effective interventions, ultimately enhancing the quality of life for individuals with ASD and their families [74]. In fact, an adaptive system holds the promise of improving outcomes for greater numbers of

XSL•FO
RenderX

children with ASD by guiding clinicians on how to leverage the diversity in treatment responses [6].

Limitations

This review is not without limitations. No technological or automated tools were used to handle the collected data, which could have introduced bias due to human error. However, because the literature search identified only a small number of articles, manual analysis was feasible, and it did not become necessary to use specific automated techniques. The limitations of this review primarily stem from the heterogeneity in the study designs of the selected studies. Future reviews should focus on different inclusion and exclusion criteria with more consistent designs to reduce variation and improve comparability. In addition, the availability of data in the selected studies posed another constraint. Some of the studies (3/10, 30%) lacked sufficient information regarding the adaptation of their systems, making it challenging to conduct specific analyses or fully assess the study's contribution.

Conclusions

In recent years, the field of ASD interventions has witnessed a significant shift toward the development of objective adaptive technologies. This evolving paradigm aims to not only enhance the overall effectiveness of interventions but also customize them to better suit the unique needs of individuals with ASD. A systematic review was conducted to assess the current body of literature concerning the application of adaptive VR systems in ASD interventions.

The findings showed trends that align with theories and similar studies focused on broader themes, such as the significant interplay between level-switching strategies and feedback and a preference for adaptive engines operating in real time [5,64]. However, throughout the paper, we also offered several reflections on how the methodological focus seems to be shifting. Specifically, we observed how implementing multimodal feedback could enhance adaptive interventions in terms of both qualities and opportunities [38,62,63]. We have provided critical insight concerning the growth of adaptation studies implementing ML techniques. Indeed, among the included records, those from the past 2 years have primarily focused on this technique (3/10, 30%) [37,43,51], likely due to its versatile and accurate nature [65,66]. Similarly, the importance of implicit biosignals was rationalized within the theoretical framework of embodied cognition [72,73] for a deeper understanding of unobservable behaviors.

This work can support researchers in designing and testing adaptive VR systems for ASD interventions and help software designers develop more engaging and targeted VR applications, thus improving the effectiveness of digital therapies. In addition, it can contribute to the research of innovative methodologies for assessing and monitoring progress in ASD treatments using advanced VR tools. It can also provide valuable support to clinical professionals in better adapting interventions to patients' states and specific needs. Finally, it can guide practitioners toward a deeper understanding of the potential of VR in ASD treatments, positively influencing decision-making in the formulation of health care and educational policies.

Acknowledgments

The Spanish Ministry of Science, Innovation, and Universities as well as the State Research Agency funded the project ADAPTEA: Biomarker-Driven Adaptive Virtual Reality Stimulation for ASD Interventions (PID2020-116422RB-C21). This study was cofunded by the European Union through the Operational Program of the European Regional Development Fund of the Valencian Community 2014-2020 (IDIFEDER/2018/029 and IDIFEDER/2021/038). The cost of publishing open access was covered by the Polytechnic University of Valencia.

Authors' Contributions

LM performed the literature search. LM, MEM, TP, AH, and MA contributed to conceptualization, conducting the analysis, and revision of the article. The authors critically reviewed the article for essential intellectual content, approved the final publication, and collectively accept responsibility for addressing any questions regarding the work's accuracy and integrity.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 117 KB - jmir_v26i1e57093_app1.pdf]

References

1. Asbee J, Kelly K, McMahan T, Parsons TD. Machine learning classification analysis for an adaptive virtual reality Stroop task. Virtual Real 2023 Jan 11;27(2):1391-1407. [doi: 10.1007/S10055-022-00744-1]
2. Kabudi T, Pappas I, Olsen DH. AI-enabled adaptive learning systems: a systematic mapping of the literature. Comput Educ Artif Intell 2021;2:100017. [doi: 10.1016/j.caeai.2021.100017]

XSL•FO
RenderX

3. Xie H, Chu HC, Hwang GJ, Wang CC. Trends and development in technology-enhanced adaptive/personalized learning: a systematic review of journal publications from 2007 to 2017. Comput Educ 2019 Oct;140:103599. [doi: 10.1016/j.compedu.2019.103599]
4. Vandewaetere M, Desmet P, Clarebout G. The contribution of learner characteristics in the development of computer-based adaptive learning environments. Comput Human Behav 2011 Jan;27(1):118-130. [doi: 10.1016/j.chb.2010.07.038]
5. Zahabi M, Abdul Razak AM. Adaptive virtual reality-based training: a systematic literature review and framework. Virtual Real 2020 Mar 13;24(4):725-752. [doi: 10.1007/s10055-020-00434-w]
6. Almirall D, DiStefano C, Chang Y, Shire S, Kaiser A, Lu X, et al. Longitudinal effects of adaptive interventions with a speech-generating device in minimally verbal children with ASD. J Clin Child Adolesc Psychol 2016;45(4):442-456 [FREE Full text] [doi: 10.1080/15374416.2016.1138407] [Medline: 26954267]
7. Parsons TD, Reinebold JL. Adaptive virtual environments for neuropsychological assessment in serious games. IEEE Trans Consumer Electron 2012 May;58(2):197-204. [doi: 10.1109/tce.2012.6227413]
8. Sampayo-Vargas S, Cope CJ, He Z, Byrne GJ. The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. Comput Educ 2013 Nov;69:452-462. [doi: 10.1016/j.compedu.2013.07.004]
9. Alcañiz M, Maddalon L, Minissi ME, Sirera M, Abad L, Giglioli IA. Intervenciones tecnológicas adaptativas para el trastorno del espectro autista: una revisión bibliográfica. Medicina 2022;82(supl. 1) [FREE Full text]
10. Maddalon L, Minissi ME, Cervera-Torres S, Hervás A, Gómez-García S, Alcañiz M. Multimodal interaction in ASD children: a usability study of a portable hybrid VR system. In: Proceedings of the 17th International Conference on Universal Access in Human-Computer Interaction,Held as Part of the 25th HCI International Conference. 2023 Presented at: UAHCI/HCII '23; July 23-28, 2023; Copenhagen, Denmark p. 614-624 URL: https://dl.acm.org/doi/abs/10.1007/978-3-031-35681-0_40 [doi: 10.1007/978-3-031-35681-0_40]
11. Parsons TD, Gaggioli A, Riva G. Extended reality for the clinical, affective, and social neurosciences. Brain Sci 2020 Nov 30;10(12):1-22 [FREE Full text] [doi: 10.3390/brainsci10120922] [Medline: 33265932]
12. Slater M, Lotto B, Arnold MM, Sanchez-Vives MV. How we experience immersive virtual environments: the concept of presence and its measurement. Anu Psicol 2009;40(2):193 [FREE Full text]
13. Barletta V, Caruso F, Di MT, Piccinno A. Serious games for autism based on immersive virtual reality: a lens on methodological and technological challenges. In: Proceedings of the 12th International Conference on Methodologies and Intelligent Systems for Technology Enhanced Learning. 2023 Presented at: MIS4TEL '22; July 13-15, 2022; L'Aquila, Italy p. 181-195 URL: https://link.springer.com/chapter/10.1007/978-3-031-20617-7_23 [doi: 10.1007/978-3-031-20617-7_23]
14. Maddalon L, Minissi ME, Torres SC, Gómez-García S, Alcañiz M. Virtual humans for ASD intervention: a brief scoping review. Medicina (B Aires) 2023 Mar;83 Suppl 2:48-52 [FREE Full text] [Medline: 36820483]
15. Parsons T, Gaggioli A, Riva G. Virtual reality for research in social neuroscience. Brain Sci 2017 Apr 16;7(4):42 [FREE Full text] [doi: 10.3390/brainsci7040042] [Medline: 28420150]
16. Bermudez I Badia S, Quintero LV, Cameirao MS, Chirico A, Triberti S, Cipresso P, et al. Toward emotionally adaptive virtual reality for mental health applications. IEEE J Biomed Health Inform 2019 Sep;23(5):1877-1887. [doi: 10.1109/JBHI.2018.2878846] [Medline: 30387752]
17. Kritikos J, Alevizopoulos G, Koutsouris D. Personalized virtual reality human-computer interaction for psychiatric and neurological illnesses: a dynamically adaptive virtual reality environment that changes according to real-time feedback from electrophysiological signal responses. Front Hum Neurosci 2021 Feb 12;15:596980 [FREE Full text] [doi: 10.3389/fnhum.2021.596980] [Medline: 33643010]
18. Cosić K, Popović S, Kukolja D, Horvat M, Dropuljić B. Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders. Cyberpsychol Behav Soc Netw 2010 Feb;13(1):73-78. [doi: 10.1089/cyber.2009.0260] [Medline: 20528296]
19. López-Ojeda W, Hurley RA. Extended reality technologies: expanding therapeutic approaches for PTSD. J Neuropsychiatry Clin Neurosci 2022 Jan;34(1):A4-A5. [doi: 10.1176/appi.neuropsych.21100244] [Medline: 35113666]
20. Lorenzo G, Lorenzo-Lledó A, Lledó A, Pérez-Vázquez E. Application of virtual reality in people with ASD from 1996 to 2019. J Enabling Technol 2020 May 25;14(2):99-114. [doi: 10.1108/jet-01-2020-0005]
21. Parsons TD, Carlew AR. Bimodal virtual reality stroop for assessing distractor inhibition in autism spectrum disorders. J Autism Dev Disord 2016 Apr;46(4):1255-1267. [doi: 10.1007/s10803-015-2663-7] [Medline: 26614084]
22. Parsons TD, Riva G, Parsons S, Mantovani F, Newbutt N, Lin L, et al. Virtual reality in pediatric psychology. Pediatrics 2017 Nov 01;140(Supplement 2):S86-S91. [doi: 10.1542/peds.2016-1758i]
23. Mesa-Gresa P, Gil-Gómez H, Lozano-Quilis J, Gil-Gómez J. Effectiveness of virtual reality for children and adolescents with autism spectrum disorder: an evidence-based systematic review. Sensors (Basel) 2018 Aug 01;18(8):2486 [FREE Full text] [doi: 10.3390/s18082486] [Medline: 30071588]
24. Bravou V, Oikonomidou D, Drigas AS. Applications of virtual reality for autism inclusion. A review. Retos 2022;45:779-785 [FREE Full text] [doi: 10.47197/retos.v45i0.92078]
25. Parsons S. Authenticity in virtual reality for assessment and intervention in autism: a conceptual review. Educ Res Rev 2016 Nov;19:138-157. [doi: 10.1016/j.edurev.2016.08.001]

XSL•FO
RenderX

26. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: a systematic review update. Autism Res 2022 May;15(5):778-790 [FREE Full text] [doi: 10.1002/aur.2696] [Medline: 35238171]
27. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). 5th edition. New York, NY: American Psychological Association; 2022.
28. Maskey M, Lowry J, Rodgers J, McConachie H, Parr JR. Reducing specific phobia/fear in young people with autism spectrum disorders (ASDs) through a virtual reality environment intervention. PLoS One 2014;9(7):e100374 [FREE Full text] [doi: 10.1371/journal.pone.0100374] [Medline: 24987957]
29. Bian D, Wade J, Warren Z, Sarkar N. Online engagement detection and task adaptation in a virtual reality based driving simulator for autism intervention. In: Proceedings of the 10th International Conference on Universal Access in Human-Computer Interaction, Users and Context Diversity. 2016 Presented at: UAHCI '16; July 17-22, 2016; Toronto, ON p. 538-547 URL: https://link.springer.com/chapter/10.1007/978-3-319-40238-3_51 [doi: 10.1007/978-3-319-40238-3_51]
30. Adiani D, Itzkovitz A, Bian D, Katz H, Breen M, Hunt S, et al. Career interview readiness in virtual reality (CIRVR): a platform for simulated interview training for autistic individuals and their employers. ACM Trans Access Comput 2022 Mar 04;15(1):1-28. [doi: 10.1145/3505560]
31. Soltiyeva A, Oliveira W, Madina A, Adilkhan S, Urmanov M, Hamari J. My lovely granny's farm: an immersive virtual reality training system for children with autism spectrum disorder. Educ Inf Technol (Dordr) (Forthcoming) 2023 May 20:1-21 [FREE Full text] [doi: 10.1007/s10639-023-11862-x] [Medline: 37361850]
32. Johnson M, Tate AM, Tate K, Laane S, Chang Z, Chapman SB. Charisma™ virtual social training: a digital health platform and protocol. Front Virtual Real 2022 Nov 14;3:4126. [doi: 10.3389/frvir.2022.1004162]
33. Ke F, Im T. Virtual-reality-based social interaction training for children with high-functioning autism. J Educ Res 2013 Nov 02;106(6):441-461. [doi: 10.1080/00220671.2013.832999]
34. Zheng Z, Warren Z, Weitlauf A, Fu Q, Zhao H, Swanson A, et al. Brief report: evaluation of an intelligent learning environment for young children with autism spectrum disorder. J Autism Dev Disord 2016 Nov;46(11):3615-3621 [FREE Full text] [doi: 10.1007/s10803-016-2896-0] [Medline: 27561902]
35. Bekele E, Wade J, Bian D, Fan J, Swanson A, Warren Z, et al. Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with autism spectrum disorders (ASD). In: Proceedings of the 2016 IEEE Conference on Virtual Reality. 2016 Presented at: VR '16; March 19-23, 2016; Greenville, SC p. 121-130 URL: https://ieeexplore.ieee.org/document/7504695 [doi: 10.1109/vr.2016.7504695]
36. Kuriakose S, Lahiri U. Design of a physiology-sensitive VR-based social communication platform for children with autism. IEEE Trans Neural Syst Rehabil Eng 2017 Aug;25(8):1180-1191. [doi: 10.1109/TNSRE.2016.2613879] [Medline: 28114071]
37. Hocking DR, Ardalan A, Abu-Rayya HM, Farhat H, Andoni A, Lenroot R, et al. Feasibility of a virtual reality-based exercise intervention and low-cost motion tracking method for estimation of motor proficiency in youth with autism spectrum disorder. J Neuroeng Rehabil 2022 Jan 07;19(1):1 [FREE Full text] [doi: 10.1186/s12984-021-00978-1] [Medline: 34996473]
38. Zhao H, Zheng Z, Swanson A, Weitlauf A, Warren Z, Sarkar N. Design of a haptic-gripper virtual reality system (Hg) for analyzing fine motor behaviors in children with autism. ACM Trans Access Comput 2018 Nov 16;11(4):1-21. [doi: 10.1145/3231938]
39. Sokolikj Z, Ke F, Chakraborty S, Moon J. Using deep learning to track representational flexibility development of children with autism in a virtual world. In: Proceedings of the 11th International Conference on Information and Education Technology. 2023 Presented at: ICIET '23; March 18-20, 2023; Fujisawa, Japan p. 51-55 URL: https://ieeexplore.ieee.org/document/10111218 [doi: 10.1109/iciet56899.2023.10111218]
40. Pradeep Raj KB, Lahiri U. Design of eyegaze-sensitive virtual reality based social communication platform for individuals with autism. In: Proceedings of the 7th International Conference on Intelligent Systems, Modelling and Simulation. 2016 Presented at: ISMS '16; January 25-27, 2016; Bangkok, Thailand p. 301-306 URL: https://ieeexplore.ieee.org/document/7877229 [doi: 10.1109/isms.2016.58]
41. Vallefuoco E, Bravaccio C, Gison G, Pecchia L, Pepino A. Personalized training via serious game to improve daily living skills in pediatric patients with autism spectrum disorder. IEEE J Biomed Health Inform 2022 Jul;26(7):3312-3322. [doi: 10.1109/JBHI.2022.3155367] [Medline: 35230960]
42. Lahiri U, Bekele E, Dohrmann E, Warren Z, Sarkar N. Design of a virtual reality based adaptive response technology for children with autism. IEEE Trans Neural Syst Rehabil Eng 2013 Jan;21(1):55-64 [FREE Full text] [doi: 10.1109/TNSRE.2012.2218618] [Medline: 23033333]
43. Moon J, Ke F. Effects of adaptive prompts in virtual reality-based social skills training for children with autism. J Autism Dev Disord 2024 Aug 28;54(8):2826-2846. [doi: 10.1007/s10803-023-06021-7] [Medline: 37246166]
44. Fuller EA, Kaiser AP. The effects of early intervention on social communication outcomes for children with autism spectrum disorder: a meta-analysis. J Autism Dev Disord 2020 May;50(5):1683-1700 [FREE Full text] [doi: 10.1007/s10803-019-03927-z] [Medline: 30805766]
45. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]

XSL•FO
RenderX

46. Bölte S, Bartl-Pokorny KD, Jonsson U, Berggren S, Zhang D, Kostrzewa E, et al. How can clinicians detect and treat autism early? Methodological trends of technology use in research. Acta Paediatr 2016 Feb;105(2):137-144 [FREE Full text] [doi: 10.1111/apa.13243] [Medline: 26479859]
47. Cipresso P, Giglioli IA, Raya MA, Riva G. The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. Front Psychol 2018;9:2086 [FREE Full text] [doi: 10.3389/fpsyg.2018.02086] [Medline: 30459681]
48. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2024.
49. Bekele E, Young M, Zheng Z, Zhang L, Swanson A, Johnston R, et al. A step towards adaptive multimodal virtual social interaction platform for children with autism. In: Proceedings of the 7th International Conference on Universal Access in Human-Computer Interaction: User and Context Diversity. 2013 Presented at: UAHCI '13; July 21-26, 2013; Las Vegas, NV p. 464-473 URL: https://link.springer.com/chapter/10.1007/978-3-642-39191-0_51 [doi: 10.1007/978-3-642-39191-0_51]
50. Jyoti V, Lahiri U. Virtual reality based joint attention task platform for children with autism. IEEE Trans Learning Technol 2020 Jan 1;13(1):198-210. [doi: 10.1109/tlt.2019.2912371]
51. Jyoti V, Lahiri U. Portable joint attention skill training platform for children with autism. IEEE Trans Learning Technol 2022 Apr 1;15(2):290-300. [doi: 10.1109/tlt.2022.3169964]
52. Eckes T, Buhlmann U, Holling HD, Möllmann A. Comprehensive ABA-based interventions in the treatment of children with autism spectrum disorder - a meta-analysis. BMC Psychiatry 2023 Mar 02;23(1):133 [FREE Full text] [doi: 10.1186/s12888-022-04412-1] [Medline: 36864429]
53. Ali S, Mehmood F, Dancey D, Ayaz Y, Khan MJ, Naseer N, et al. An adaptive multi-robot therapy for improving joint attention and imitation of ASD children. IEEE Access 2019;7:81808-81825. [doi: 10.1109/access.2019.2923678]
54. Kumazaki H, Warren Z, Swanson A, Yoshikawa Y, Matsumoto Y, Yoshimura Y, et al. Brief report: evaluating the utility of varied technological agents to elicit social attention from children with autism spectrum disorders. J Autism Dev Disord 2019 Apr;49(4):1700-1708 [FREE Full text] [doi: 10.1007/s10803-018-3841-1] [Medline: 30511126]
55. Rudovic O, Lee J, Dai M, Schuller B, Picard RW. Personalized machine learning for robot perception of affect and engagement in autism therapy. Sci Robot 2018 Jun 27;3(19):eaao6760. [doi: 10.1126/scirobotics.aao6760] [Medline: 33141688]
56. McArthur BA, Volkova V, Tomopoulos S, Madigan S. Global prevalence of meeting screen time guidelines among children 5 years and younger: a systematic review and meta-analysis. JAMA Pediatr 2022 Apr 01;176(4):373-383 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.6386] [Medline: 35157028]
57. Grynszpan O, Weiss PL, Perez-Diaz F, Gal E. Innovative technology-based interventions for autism spectrum disorders: a meta-analysis. Autism 2014 May 03;18(4):346-361. [doi: 10.1177/1362361313476767] [Medline: 24092843]
58. Parsons S, Cobb S. State-of-the-art of virtual reality technologies for children on the autism spectrum. Eur J Spec Needs Educ 2011 Aug;26(3):355-366. [doi: 10.1080/08856257.2011.593831]
59. Minissi ME, Gómez-Zaragozá L, Marín-Morales J, Mantovani F, Sirera M, Abad L, et al. The whole-body motor skills of children with autism spectrum disorder taking goal-directed actions in virtual reality. Front Psychol 2023 Apr 6;14:1140731 [FREE Full text] [doi: 10.3389/fpsyg.2023.1140731] [Medline: 37089733]
60. Banneyer KN, Bonin L, Price K, Goodman WK, Storch EA. Cognitive behavioral therapy for childhood anxiety disorders: a review of recent advances. Curr Psychiatry Rep 2018 Jul 28;20(8):65. [doi: 10.1007/s11920-018-0924-9] [Medline: 30056623]
61. Chu L, Shen L, Ma C, Chen J, Tian Y, Zhang C, et al. Effects of a nonwearable digital therapeutic intervention on preschoolers with autism spectrum disorder in China: open-label randomized controlled trial. J Med Internet Res 2023 Aug 24;25:e45836 [FREE Full text] [doi: 10.2196/45836] [Medline: 37616029]
62. Campbell BS, Feldmann A. The power of multimodal feedback. J Curric Teach Learn Leadersh Educ 2017;2(2):1-6 [FREE Full text]
63. Sigrist R, Rauter G, Riener R, Wolf P. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. Psychon Bull Rev 2013 Feb;20(1):21-53. [doi: 10.3758/s13423-012-0333-8] [Medline: 23132605]
64. Munshi A, Biswas G, Baker R, Ocumpaugh J, Hutt S, Paquette L. Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. Compute Assist Learn 2022 Dec 13;39(2):351-368. [doi: 10.1111/jcal.12761]
65. Orrù G, Monaro M, Conversano C, Gemignani A, Sartori G. Machine learning in psychometrics and psychological research. Front Psychol 2019;10:2970 [FREE Full text] [doi: 10.3389/fpsyg.2019.02970] [Medline: 31998200]
66. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci 2017 Nov;12(6):1100-1122 [FREE Full text] [doi: 10.1177/1745691617693393] [Medline: 28841086]
67. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2001 Dec Presented at: CVPR '01; December 8-14, 2001; Kauai, HI p. 2001 URL: https://ieeexplore.ieee.org/document/990517 [doi: 10.1109/cvpr.2001.990517]
68. Ahila Priyadharshini R, Hariharan S, Jagadeeswara R. A CNN-based approach for face recognition under different orientations. In: Proceedings of the 2023 Conference on Computational Intelligence in Pattern Recognition. 2023 Presented

XSL•FO
RenderX

at: CIPR '23; December 13-15, 2023; Virtual Event p. 157-168 URL: https://link.springer.com/chapter/10.1007/978-981-99-3734-9_14 [doi: 10.1007/978-981-99-3734-9_14]
69. Okuno K, Yamashita T, Fukui H, Noridomi S, Arata K, Yamauchi Y, et al. Body posture and face orientation estimation by convolutional network with heterogeneous learning. In: Proceedings of the 2018 International Workshop on Advanced Image Technology. 2018 Presented at: IWAIT '18; January 7-9, 2018; Chiang Mai, Thailand p. 1-4 URL: https://ieeexplore.ieee.org/document/8369677 [doi: 10.1109/iwait.2018.8369677]
70. Barsalou LW. Grounded cognition. Annu Rev Psychol 2008;59:617-645. [doi: 10.1146/annurev.psych.59.103006.093639] [Medline: 17705682]
71. Ghandi M, Blaisdell M, Ismail M. Embodied empathy: using affective computing to incarnate human emotion and cognition in architecture. Int J Archit Comput 2021 Aug 28;19(4):532-552. [doi: 10.1177/14780771211039507]
72. Kopcha TJ, Valentine KD, Ocak C. Editorial: preface to the special issue on embodied cognition and technology for learning. Education Tech Research Dev 2021 Jul 14;69(4):1881-1887. [doi: 10.1007/s11423-021-10023-6]
73. Yuan RQ, Hsieh SW, Chew SW, Chen NS. The effects of gesture-based technology on memory training in adaptive learning environment. In: Proceedings of the 2015 International Conference of Educational Innovation through Technology. 2015 Presented at: EITT '15; October 16-18, 2015; Wuhan, China p. 190-193 URL: https://ieeexplore.ieee.org/document/7446175 [doi: 10.1109/eitt.2015.47]
74. Pv AK, Satheesh A, Abhishek NM, Menon HP, Devasia D. Application of virtual reality (VR) to advance social ability in children with ASD. In: Proceedings of the 2nd International Conference on Automation, Computing and Renewable Systems. 2023 Presented at: ICACRS '23; December 11-13, 2023; Pudukkottai, India p. 1581-1586. [doi: 10.1109/icacrs58579.2023.10404594]

Abbreviations

ASD: autism spectrum disorder
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
VR: virtual reality

Edited by T de Azevedo Cardoso; submitted 05.02.24; peer-reviewed by N Vaughan, E Vallefuoco; comments to author 23.02.24; revised version received 26.03.24; accepted 26.06.24; published 18.09.24.

Please cite as:
Maddalon L, Minissi ME, Parsons T, Hervas A, Alcaniz M
Exploring Adaptive Virtual Reality Systems Used in Interventions for Children With Autism Spectrum Disorder: Systematic Review
J Med Internet Res 2024;26:e57093
URL: https://www.jmir.org/2024/1/e57093
doi:10.2196/57093
PMID:

©Luna Maddalon, Maria Eleonora Minissi, Thomas Parsons, Amaia Hervas, Mariano Alcaniz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Investigating the Use of Telemedicine by Health Care Providers to Diagnose and Manage Patients With Musculoskeletal Disorders: Systematic Review and Meta-Analysis

Raphaël Vincent[1,2,3], PT, MSc; Maxime Charron[2], PT, MSc; Simon Lafrance[1,2], PT, PhD; Audrey-Anne Cormier[2], PT, MSc; Dahlia Kairy[1,3], PT, PhD; François Desmeules[1,2], PT, PhD

[1]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada

[2]Hôpital Maisonneuve-Rosemont Research Center, Université de Montréal Affiliated Research Center, Montréal, QC, Canada

[3]Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal, Institut Universitaire sur la Réadaptation en Déficience Physique de Montréal, Montréal, QC, Canada

Corresponding Author:
François Desmeules, PT, PhD
School of Rehabilitation
Faculty of Medicine
Université de Montréal
7077 Avenue du Parc
Montréal, QC, H3N 1X7
Canada
Phone: 1 514 343 6791
Email: f.desmeules@umontreal.ca

Abstract

Background: Access to care is a major challenge for patients with musculoskeletal disorders (MSKDs). Telemedicine is one of the solutions to improve access to care. However, initial remote diagnosis of MSKDs involves some challenges, such as the impossibility of touching the patient during the physical examination, which makes it more complex to obtain a valid diagnosis. No meta-analysis has been performed to date to synthesize evidence regarding the initial assessment including a physical evaluation using telemedicine to diagnose patients with MSKDs.

Objective: This study aims to appraise the evidence on diagnostic and treatment plan concordance between remote assessment using synchronous or asynchronous forms of telemedicine and usual in-person assessment for the initial evaluation of various MSKDs.

Methods: An electronic search was conducted up to August 2023 using terms related to telemedicine and assessment of MSKDs. Methodological quality of studies was assessed with the Quality Assessment of Diagnostic Accuracy Studies 2 tool. Random-effect model meta-analyses were performed. The Grading of Recommendations, Assessment, Development, and Evaluations framework was used to synthesize the quality and certainty of the evidence.

Results: A total of 23 concordance studies were eligible and included adult participants (N=1493) with various MSKDs. On the basis of high certainty, pooled κ and prevalence-adjusted and bias-adjusted κ for the diagnostic concordance between remote and in-person assessments of MSKDs were 0.80 (95% CI 0.72-0.89; 7 studies, 353 patients) and 0.83 (95% CI 0.76-0.89; 6 studies, 306 patients). On the basis of moderate certainty, pooled Gwet AC1 for treatment plan concordance between remote and in-person assessments of MSKDs was 0.90 (95% CI 0.80-0.99; 2 studies, 142 patients).

Conclusions: The diagnostic concordance for MSKDs is good to very good. Treatment plan concordance is probably good to excellent. Studies evaluating the accuracy to detect red and yellow flags as well as the potential increase in associated health care resources use, such as imaging tests, are needed.

(J Med Internet Res 2024;26:e52964) doi:10.2196/52964

XSL•FO
RenderX

KEYWORDS

telemedicine; musculoskeletal diseases; physical examination; diagnosis; treatment; health care; telecare; meta-analysis; systematic review; telehealth; orthopedic; test; musculoskeletal disorder; MSKD; older adult; older adults; older person; older people; aging; musculoskeletal; mobile phone

Introduction

Background

Almost 1.7 billion people have a musculoskeletal disorder (MSKD) in the world [1]. MSKDs refer to disorders affecting muscles, bones, and joints, such as low back pain, osteoarthritis, or tendinopathy [2]. MSKDs can lead to pain, disability, and poor health-related quality of life, resulting in a significant burden to health care systems as well as to society [2-6]. Evidence shows that a valid diagnosis through a careful initial evaluation and prompt treatment are essential for MSKDs [2,7]. Recommendations for various MSKDs emphasize the importance of screening for signs and symptoms of underlying serious pathologies (red flags) and for psychological factors associated with a poorer prognosis (yellow flags) during the initial evaluation as well as conducting a physical examination that includes measurements of mobility, strength, and use of orthopedic tests [2,8].

Significant health inequalities exist between urban and low-density population areas in several countries, with rural populations having a higher rate of injury and a higher risk of chronic MSKDs [9-11]. These inequalities can partially be explained by difficulties in accessing primary and secondary care [12-14]. The COVID-19 pandemic exacerbated these difficulties, with nearly 1 in 5 individuals unable to access health care during the first year of the COVID-19 pandemic, and patients with MSKDs were more affected by these delays than other patient populations [6,15-17]. Access to care is a major challenge for patients with MSKDs as delays can negatively impact clinical outcomes, such as pain, disability, or quality of life, while also potentially exacerbating psychological distress [18].

Telemedicine has been shown to have many benefits for patients, health care systems, and society [19]. It is an interesting option for optimizing health care access by removing unnecessary hurdles, such as geographical location or for patients with impaired mobility [19-21]. Telemedicine first appeared under asynchronous forms where communication between parties was not happening in real time [22]. Asynchronous forms are still used today for communication between patients and health care providers via messaging systems (emails and instant SMS text messaging) or smartphone apps for follow-up or counselling [23]. Synchronous forms of telemedicine with phone calls and videoconferencing allow real-time interaction between parties [22]. The use of videoconferencing in telemedicine has been a major progress by allowing visual evaluation of movement, edema, or scars, for example, or to observe patient physical performance and function [24]. The use of telemedicine in high-income countries now mainly focuses on patient remote management or follow-ups of patients who have already been assessed in person [25]. Telemedicine is well implanted in medical specialties, such as radiology, dermatology, or psychiatry, but is not widely used in the MSKDs care pathways, particularly for the initial assessment of new patients where a diagnosis is required [25,26]. The use of telemedicine for MSKDs brings many challenges related to the geographical distancing of the patient and the health care provider [27-31]. The loss of physical contact during the physical examination and treatments raises questions among patients and clinicians about the relevance, safety, and effectiveness of remote care [27,30,31]. Patients and health care providers also express concerns about the quality of the therapeutic relationship when consultations are not conducted in person [27,28]. However, telemedicine offers advantages to patients and health care providers, such as better accessibility, greater flexibility, and the possibility to offer interventions that are adapted to a patient's environment as the remote consultation will likely be in a person's home [27].

Previous systematic reviews and meta-analyses have reported that remote patient follow-up management after an in-person initial evaluation is a valid alternative to usual in-person management and leads to similar benefits in clinical outcomes for various MSKDs, such as low back pain, neck pain, shoulder pain, or neurological conditions [32-35]. Previous systematic reviews focusing on synchronous remote initial assessment highlighted that most clinical measures remotely evaluated have a good concurrent validity and an excellent reliability [24,36]. Remote diagnoses were highly concordant with in-person diagnoses, but no meta-analysis was performed on these results, and these systematic reviews only included studies with physiotherapists as evaluators, excluding other health care providers who usually assess patients with MSKDs, such as orthopedic surgeons or primary care physicians [36]. As the diagnosis as well as the treatment plan are key elements of an initial assessment, it is crucial to undertake a comprehensive appraisal, including a meta-analysis of the concordance between remote and in-person assessment for new patients with MSKDs.

Objective

This systematic review with meta-analysis aimed to appraise the available evidence on diagnosis and care concordance after an initial assessment between a remote evaluation and an in-person evaluation for the evaluation of various MSKDs.

Therefore, the research question for this systematic review was as follows: Can a complete, valid, and safe assessment be carried out using asynchronous, synchronous or both forms of telehealth to diagnose various MSKDs?

Methods

Protocol, Registration, and Deviation

The protocol of this systematic review has been registered on PROSPERO (CRD42022335606) [37]. A total of 2 deviations from the protocol have occurred and are as follows: (1) We now only present results for the first specific objective (concordance

XSL•FO
RenderX

of diagnosis and management) in this publication, as there were too much data concerning the second objective (concordance of clinical measures) to be presented in 1 publication (data concerning the second objective will be presented in a second publication); (2) We have now added the calculation of prediction intervals to further assess the heterogeneity of the pooled estimates in addition to the I^2 and τ^2 statistics.

Literature Search

An electronic search was conducted in 4 databases (MEDLINE, Embase, Cochrane Central, and CINAHL) from January 2000 to August 2023 using terms related to telemedicine by any type of health care providers (physicians, orthopedic surgeons, physiotherapists, or other providers), MSKDs, and the assessment or the clinical examination of these conditions. A trained librarian was consulted during the development of our research strategy. We chose to limit our search to the last 2 decades to identify technologies still in use and readily available to clinicians. The full search strategy is available in the Multimedia Appendix 1. Reference lists of identified published studies and previous systematic reviews were checked for any additional studies.

Study Selection

A pair of reviewers (RV and MC or AAC) independently reviewed titles and abstracts to identify studies of interest. Consensus from reviewers was required to include studies. A third reviewer (SL) was available if a consensus was not achieved by the 2 initial reviewers. Textbox 1 shows the inclusion criteria for articles.

Textbox 1. Inclusion criteria for eligible studies.

Inclusion criteria

- Participants with symptoms related to suspected musculoskeletal disorders.
- Participants assessed during a remote evaluation (synchronous or asynchronous) and compared with an in-person evaluation by any health care provider.
- Any type of diagnostic, care or treatment plan concordance outcomes (such as raw agreement, κ, prevalence-adjusted and bias-adjusted κ, or other concordance coefficients).
- Technologies used for the remote assessment had to be accessible for routine clinical use (no experimental setup and no use of sensors to be placed on the participants).
- Full-article or conference abstracts written in French or English.

Data Extraction

Data of included studies were extracted using a predefined standardized form documenting the following: authors' names, year of publication, study design, country, care settings, characteristics of the remote assessment (modality and technology used as well as the presence of a third party to help with patient evaluation), body region or regions affected, number of participants, participant characteristics, health care provider characteristics, and outcome measures. Data extraction was performed by 1 evaluator (RV), and the extracted data were reviewed and verified by a second evaluator (MC or AAC). When data were missing or incomplete, attempts to contact original authors were made to obtain complete data and results.

Methodological Quality Assessment

Methodological quality of included studies was assessed with the valid and reliable Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool [38]. This tool appraises patient selection, index tests, reference standard, flow, and timing. QUADAS-2 also assesses the presence of applicability concerns that would decrease the external validity of the study results. Assessment of methodological quality and applicability concerns was performed by a pair of independent evaluators (RV and MC or SL or AAC or Claudia Cosculluela); the final score was obtained through consensus. In case of disagreement, a third evaluator was available to facilitate consensus (MC or SL). Studies were considered at risk of bias if at least 1 item of the QUADAS-2 was evaluated at high risk of bias.

Data Synthesis

κ, prevalence-adjusted and bias-adjusted κ (PABAK), and Gwet AC1 coefficients were pooled into separate meta-analyses. Weighted means were calculated for raw agreement. Random-effect model meta-analyses were performed using the *metafor* package in RStudio Team (2020, Rstudio: Integrated Development for R. Rstudio) [39,40]. Secondary analyses for different modalities of delivery (asynchronous or synchronous) and for affected body regions were performed. Sensitivity analyses were also performed, including only studies with a low risk of bias (with no item ranked as high risk of bias on the QUADAS-2). Pooled concordance estimates were interpreted as follows: 0.00-0.20=weak, 0.21-0.40=slight, 0.41-0.60=moderate, 0.61-0.80=good, 0.81-0.90=very good, and 0.91-1.00=excellent agreement [41,42]. Coding scripts are available in Multimedia Appendix 2. For all meta-analyses, α levels were set at .05, and 95% CIs were calculated. Statistical heterogeneity was quantified and reported using the I^2 and τ^2 statistics and interpreted according to the Cochrane methodology [43]. Prediction intervals were also calculated using the Comprehensive Meta-Analysis software (version 4.0; H.Biostat) to further assess the heterogeneity of the pooled estimates [44,45]. A narrative synthesis was performed for studies and outcomes not pooled in meta-analyses or for conference abstracts included in this review.

The Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) framework was used for grading the quality and certainty of evidence and for formulating recommendations [46]. Level of evidence was interpreted as

XSL•FO
RenderX

follows: very low certainty: the true effect is probably markedly different from the pooled estimated effect, low certainty: the true effect might be markedly different from the pooled estimated effect, moderate certainty: we believe that the true effect is probably close to the pooled estimated effect, and high certainty: we are very confident that the true effect is similar to the pooled estimated effect.

Results

Overview

After full-text review, 11 studies on in-person remote diagnostic concordance [42,47-56], 7 studies on in-person remote diagnostic and treatment plan concordance [41,57-62], and 3 studies on in-person remote treatment plan concordance [63-65] were included (Figure 1). No studies presented results on triage concordance of surgical candidates. A total of 2 conference abstracts were also included on in-person remote diagnostic concordance [66,67]. Reasons for exclusions are available in Multimedia Appendix 3. Full characteristics of the included studies are presented in Tables 1 and 2 and Table S1 in Multimedia Appendix 4.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for concordance studies selection.

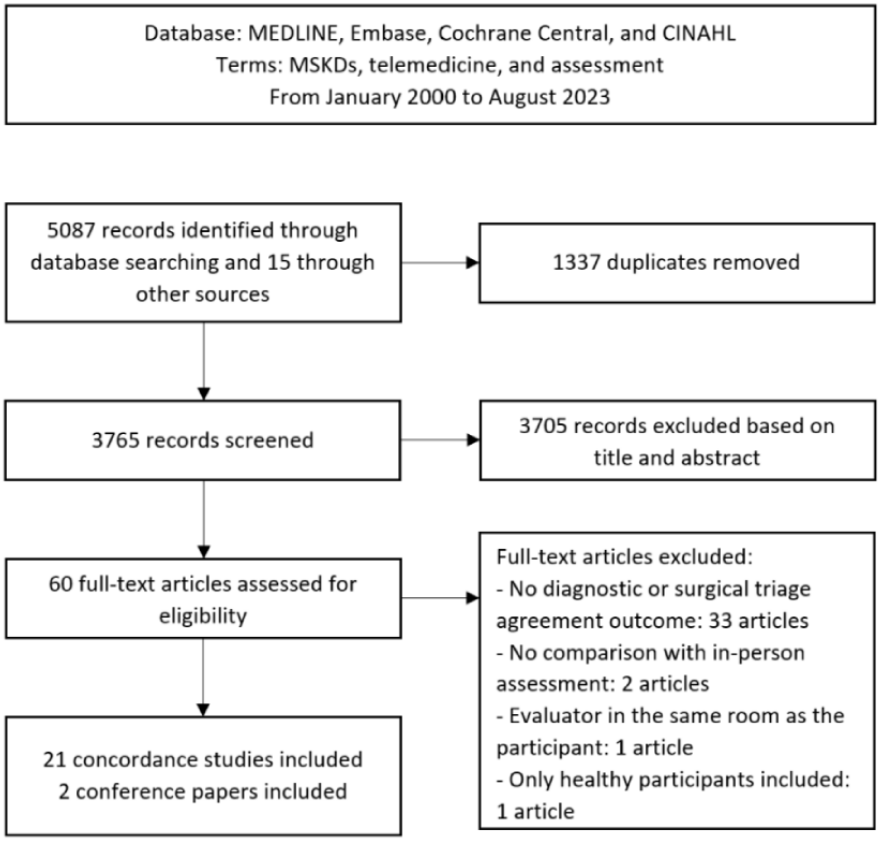

XSL•FO
RenderX

Table 1. Characteristics of the included studies (study settings).

Author and year	Design	Country	Care setting	Modality of assessment	Technology
Abboud et al [57], 2005	Prospective cohort	United States	Orthopedic	Asynchronous	Recorded photographs and data
Bovonratwet et al [58], 2022	Retrospective cohort	United States	Orthopedic	Synchronous	Videoconferencing
Bradley et al [47], 2021	Case control	United States	Orthopedic	Synchronous	Videoconferencing
Demaerschalk et al [48], 2022	Retrospective cohort	United States	Clinic	Synchronous	Videoconferencing
Dias et al [56], 2023	Prospective cohort	Brazil	Orthopedic	Synchronous	Videoconferencing
Exposto et al [42], 2022	Prospective cohort	Denmark	Dentistry	Synchronous	Videoconferencing
Melnick et al [65], 2023	Retrospective cohort	United States	Orthopedic	Synchronous	Videoconferencing
Michaelson et al [62], 2023	Retrospective cohort	United States	Orthopedic	Synchronous	Videoconferencing with the presence of the patient's caregiver
Rabin et al [61], 2022	Prospective cohort	Israel	Orthopedic	Synchronous	Videoconferencing with presence of a research assistant or the patient's caregiver
Tachakra et al [54], 2000	Prospective cohort	United Kingdom	Emergency Department	Synchronous	Videoconferencing with presence of a nurse
Wang et al [55], 2022	Prospective cohort	United States	Orthopedic	Asynchronous	Recorded video of a standardized clinical examination with presence of a research assistant
Cottrell et al [59], 2018	Prospective cohort	Australia	Physiotherapy	Synchronous	Videoconferencing
Lade et al [49], 2012	Prospective cohort	Australia	Physiotherapy	Synchronous	Videoconferencing
Lovo et al [41], 2022	Prospective cohort	Canada	Rural clinic	Synchronous	Videoconferencing with presence of a nurse
Peterson et al [63], 2019	Prospective cohort	United States	Physiotherapy	Synchronous	Videoconferencing
Richardson et al [50], 2017	Prospective cohort	Australia	Physiotherapy	Synchronous	Videoconferencing
Russell et al [52], 2010	Prospective cohort	Australia	Physiotherapy	Synchronous	Videoconferencing
Russell et al [51], 2010	Prospective cohort	Australia	Physiotherapy	Synchronous	Videoconferencing
Steele et al [53], 2012	Prospective cohort	Australia	Physiotherapy	Synchronous	Videoconferencing
Turner et al [60], 2019	Prospective cohort	United Kingdom	Physiotherapy	Synchronous	Telephone
Worboys et al [64], 2018	Prospective cohort	Australia	Occupational therapy	Synchronous	Videoconferencing with presence of an assistant

XSL•FO
RenderX

Table 2. Characteristics of the included studies (patient and evaluator characteristics).

Author and year	Patients				Evaluators		Access to imaging tests to make diagnosis
	Musculoskeletal disorders (MSKDs[a])	Values, n	Age (y), mean (SD)	Female (%)	Profession (number of evaluators)	Training	
Abboud et al [57], 2005	Upper extremity MSKDs	100	43 (NR[b])	50	Orthopedic surgeons (n=3)	NR	Yes
Bovonratwet et al [58], 2022	Spine MSKDs	65	57.5 (14.8)	46	Orthopedic surgeons (n=5)	NR	Yes
Bradley et al [47], 2021	Shoulder MSKDs	50	58.2	53	Orthopedic surgeons (n=9)	NR	No
Demaerschalk et al [48], 2022	MSKDs	300	NR	NR	Physicians, nurses, and physician assistants	NR	NR
Dias et al [56], 2023	Low back pain	122	47.8 (13.1)	66	Orthopedic surgeons (n=17)	NR	NR
Exposto et al [42], 2022	Temporomandibular joint disorders	15	31 (12)	75	Dentists (n=5)	10 hours of training	NR
Melnick et al [65], 2023	Spine MSKDs	152	61.4 (15.4)	43	Orthopedic surgeons (n=7)	NR	NR
Michaelson et al [62], 2023	Shoulder MSKDs	82	52.6 (NR)	NR	Orthopedic surgeons (n=5)	NR	Yes
Rabin et al [61], 2022	Shoulder MSKDs	47	44.6 (22)	36	Orthopedic surgeons (n=2)	4 hours of meeting and training with 5 pilot test participants	Yes
Tachakra et al [54], 2000	Minor trauma injuries	200	NR	31	Emergency physician	NR	Yes
Wang et al [55], 2022	Shoulder MSKDs	32	50.2 (16.2)	47	Orthopedic surgeon (n=1)	NR	Yes
Cottrell et al [59], 2018	Chronic MSKDs	42	52.7 (14.5)	57	Advanced practice physiotherapists (n=6)	Training with 4 pilot test participants	Yes
Lade et al [49], 2012	Elbow MSKDs	10	38 (13)	10	Physiotherapy students (n=3)	Support by an experienced physiotherapist	NR
Lovo et al [41], 2022	Chronic back pain	27	53.7 (18.1)	70	Physiotherapists (n=2) and nurses (n=2)	Interprofessional training before the study	NR
Peterson et al [63], 2019	Low back pain	47	48.6 (15)	70	Physiotherapists (n=2)	NR	NR
Richardson et al [50], 2017	Knee MSKDs	18	23 (7)	55	Physiotherapists (n=3)	Training with 2 pilot test participants	NR
Russell et al [52], 2010	Ankle MSKDs	15	24.5 (10.8)	67	Physiotherapy students (n=3)	Training with 2 pilot test participants	NR
Russell et al [51], 2010	Lower limb MSKDs (nonarticular)	19	26 (13)	74	Physiotherapists (n=3)	Training with 2 pilot test participants	NR
Steele et al [53], 2012	Shoulder MSKDs	22	30.7 (14.2)	27	Physiotherapy students (n=3)	Support by an experienced physiotherapist	NR
Turner et al [60], 2019	MSKDs	55	NR	NR	Physiotherapists (n=22)	NR	NR
Worboys et al [64], 2018	Hand injuries	18	NR	45	Occupational therapists (n=4)	NR	NR

[a]MSKD: musculoskeletal disorder.

XSL•FO
RenderX

[b]NR: not reported.

Countries and Clinical Settings

Studies were conducted in Australia (7/23, 30%), the United States (8/23, 35%), the United Kingdom (4/23, 17%), Denmark (1/23, 4%), Canada (1/23, 4%), Brazil (1/23, 4%), and Israel (1/23, 4%). In terms of care settings, studies were conducted in physiotherapy primary care clinics (8/23, 35%) [49-53,59,60,63], in orthopedic specialized care (8/23, 35%) [47,55-58,61,62,65], and in primary medical care (4/23, 17%) [41,48,66,67]; one study was conducted in an emergency department [54]. One study was concerning temporomandibular disorders in dentistry [42], and 1 study was conducted in an occupational therapy clinic [64]. Synchronous modalities were the most frequent way of contacting patients (21/23, 91%), with 18 (78%) of the 23 studies using a videoconferencing interface [41,42,47-56,58,59,61-63,65] and 3 (13%) using telephone calls [60,66,67]. Support staff was present to help with patient assessment and mobilization during the remote evaluation in 5 (22%) of the 23 studies [41,54,55,61,65]. Asynchronous modalities were used in 2 (8%) of the 23 studies, 1 using photographs of the patients [57] and 1 using video recordings of a self-performed clinical examination under the supervision of an assistant [55].

Participants

A total of 1849 participants were included. Mean age of participants was 49.6 (SD 14.9) years. Female gender accounted for 48% (887/1849) of the participants. Included participants had spinal disorders (703/1849, 38.02%), upper limb disorders (375/1849, 20.3%), lower limb disorders (201/1849, 10.9%), temporomandibular disorders (15/1849, 0.8%), and unspecified various MSKDs (555/1849, 30%). Of the 23 studies, 5 (22%) were conducted in university physiotherapy and dentistry clinics, which included a younger sample of patients and more acute or traumatic pathologies [42,50-53]. Other studies were conducted essentially in usual orthopedic and physiotherapy clinics and included broader samples with various MSKDs clinical representations.

Evaluators and Training

Evaluators were physiotherapists in 35% (8/23) of the studies, physiotherapy students in 13% (3/23) of the studies, or orthopedic surgeons in 35% (8/23) of the studies. Other health care providers involved were nurses, dentists, occupational therapists, and emergency and primary care physicians. In 7 (30%) of the 23 studies, evaluators received specific training on how to perform and adapt their remote assessment [41,42,50-52,59,61]. No information regarding specific training was available for the remaining studies [47,48,54-58,60,62-67]. Of the 23 studies, 3 (13%) involving physiotherapy students allowed them to be advised by a senior physiotherapist if they felt the need [49,52,53]. In 7 (30%) of the 23 studies, evaluators had access to participants' imaging test results [54,55,57-59,61,62] while in 3 (13%) studies they did not [47,66,67]. No information regarding access to imaging results was available for the remaining studies.

Methodological Quality of Included Studies

Of the 21 studies, 3 (14%) had a perfect QUADAS-2 score [55-57]. A total Of the 21 studies, 11 (52%) had at least 1 item of the QUADAS-2 scored as a high a risk of bias [42,47,49,52-54,58,59,62,63,65] and 9 (43%) had at least 1 item of the QUADAS-2 score scored with an unclear risk of bias [41,48,49,51-54,58,61]. The main shortcomings concerned recruitment method or patient selection [50-52,58,59,61,63,64], the use of case-control study design [42,47], and the lack of blinding or validity of the evaluators' diagnosis (diagnoses made by physiotherapy students with inconsistent supervision) [49,52-54,62,65]. Of the 9 (43%) had an applicability concern related to the representativeness of the sample of patients recruited [50-53,58,60,61,63,64]. A study had applicability concerns due to the choice of the index test and of the reference standard for the remote and in-person assessments because the remote and in-person assessments were conducted simultaneously [64]. Details are presented in Table 3.

XSL•FO
RenderX

Table 3. Quality of the included concordance studies based on the Quality Assessment of Diagnostic Accuracy Studies 2 Tool.

Author and year	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Abboud et al [57], 2005	Low	Low	Low	Low	Low	Low	Low
Dias et al [56], 2023	Low	Low	Low	Low	Low	Low	Low
Wang et al [55], 2022	Low	Low	Low	Low	Low	Low	Low
Bovonratwet et al [58], 2022	High[a]	Low	Unclear	Low	High[a]	Low	Low
Bradley et al [47], 2021	High[b]	Low	Low	Low	Low	Low	Low
Cottrell et al [59], 2018	High[c]	Low	Low	Low	Low	Low	Low
Demaerschalk et al [48], 2022	Low	Low	Unclear	Low	Low	Low	Low
Exposto et al [42], 2022	High[b]	Low	Low	Low	Low	Low	Low
Lade et al [49], 2012	Unclear	Low	High[d]	Low	Low	Low	Low
Lovo et al [41], 2022	Unclear	Low	Low	Low	Low	Low	Low
Melnick et al [65], 2023	Low	Low	High[e]	Low	Low	Low	Low
Michaelson et al [62], 2023	Low	Low	High[e]	Unclear	Low	Unclear	Low
Peterson et al [63], 2019	High[f,g]	Low	Low	Low	High[g]	Low	Low
Rabin et al [61], 2022	Low	Unclear	Unclear	Low	High[h]	Low	Low
Richardson et al [50], 2017	Low	Low	Low	Low	High[i]	Low	Low
Russell et al [52], 2010	Unclear	Low	High[d]	Low	High[i]	Low	Low
Russell et al [51], 2010	Unclear	Low	Low	Low	High[i]	Low	Low
Steele et al [53], 2012	Unclear	Low	High[d]	Low	High[i]	Low	Low
Tachakra et al [54], 2022	Unclear	Low	High[j]	Low	Low	Low	Low
Turner et al [60], 2019	Low	Low	Low	Low	Unclear	Low	Low
Worboys et al [64], 2018	Low	Low	Low	Low	High[k]	High[k]	High[k]

[a]Patients who did not have a specific diagnosis and treatment plan during the remote assessment were excluded, which could increase agreement by including only patients who could be easily assessed remotely.

[b]Case-control design was used in the study.

[c]Only a convenience sampling was used for recruitment.

[d]Assessments were conducted by physiotherapy students, which could reduce the validity of the diagnoses made by evaluators.

[e]During the in-person assessment, evaluators had access to the treatment plan proposed after the remote assessment.

[f]Only a convenience sampling was used for recruitment.

[g]Patients with surgical history of the lumbosacral spine, visible deformity, or no reproduction of symptoms with certain orthopedic tests were excluded.

[h]Almost half of the patients (22/47, 47%) were evaluated during a postoperative or nonoperative follow-up consultation, which could increase agreement on diagnosis and treatment plan.

[i]This cohort was composed of younger patients recruited in the university clinic with only acute or subacute injuries, without degenerative pathologies.

[j]The same evaluator conducted the remote and in-person assessments.

[k]Initial assessments were excluded, and a third party was present with the patient to assist with data collection, which could have increased the agreement.

Diagnostic Concordance Between Remote and In-Person Assessments

Of the 23 studies, 17 (74%) reported diagnostic concordance results between in-person and remote assessments, and a study [62] reported diagnostic changes between in-person and remote assessments; the overall raw agreement was 85.9% (1352/1574 patients). Of the 23 studies, diagnostic concordance between in-person and remote assessments pooled estimations were possible for 7 (30%) studies for κ and 6 (26%) for PABAK

XSL•FO
RenderX

estimates. The pooled estimates are presented in Figure 2. The pooled κ was 0.80 (95% CI 0.72-0.89; 7 studies, 353 patients), and the pooled PABAK was 0.83 (95% CI 0.76-0.89; 6 studies, 306 patients), corresponding to very good agreement between both types of assessments. Statistical heterogeneity measured with the I^2 statistic were 65% and 38%, respectively, and considered potentially moderate for these meta-analyses. Prediction intervals for the pooled κ and PABAK were (0.52-0.92) and (0.68-0.91), which were very similar to the calculated CIs for the PABAK but not for the κ, and that difference could be attributed to the κ estimate in the study by Dias et al [56].

Figure 2. Diagnostic concordance (Cohen κs and prevalence-adjusted and bias-adjusted κs [PABAK]) between in-person and remote assessment in patients with musculoskeletal disorders. Meta-analysis with pooled Cohen κs included studies with physiotherapists (6/32, 19%), orthopedic surgeons (23/32, 72%), and physiotherapy students (3/32, 9%) as evaluators. Meta-analysis with pooled PABAKs included studies with physiotherapists (6/30, 33%), orthopedic surgeons (21/30, 50%) and physiotherapy students (3/30, 17%) as evaluators. 95% PI: 95% prediction interval; RE: random-effect model; τ^2/I^2/P: test for heterogeneity [50-52,55-57,61].

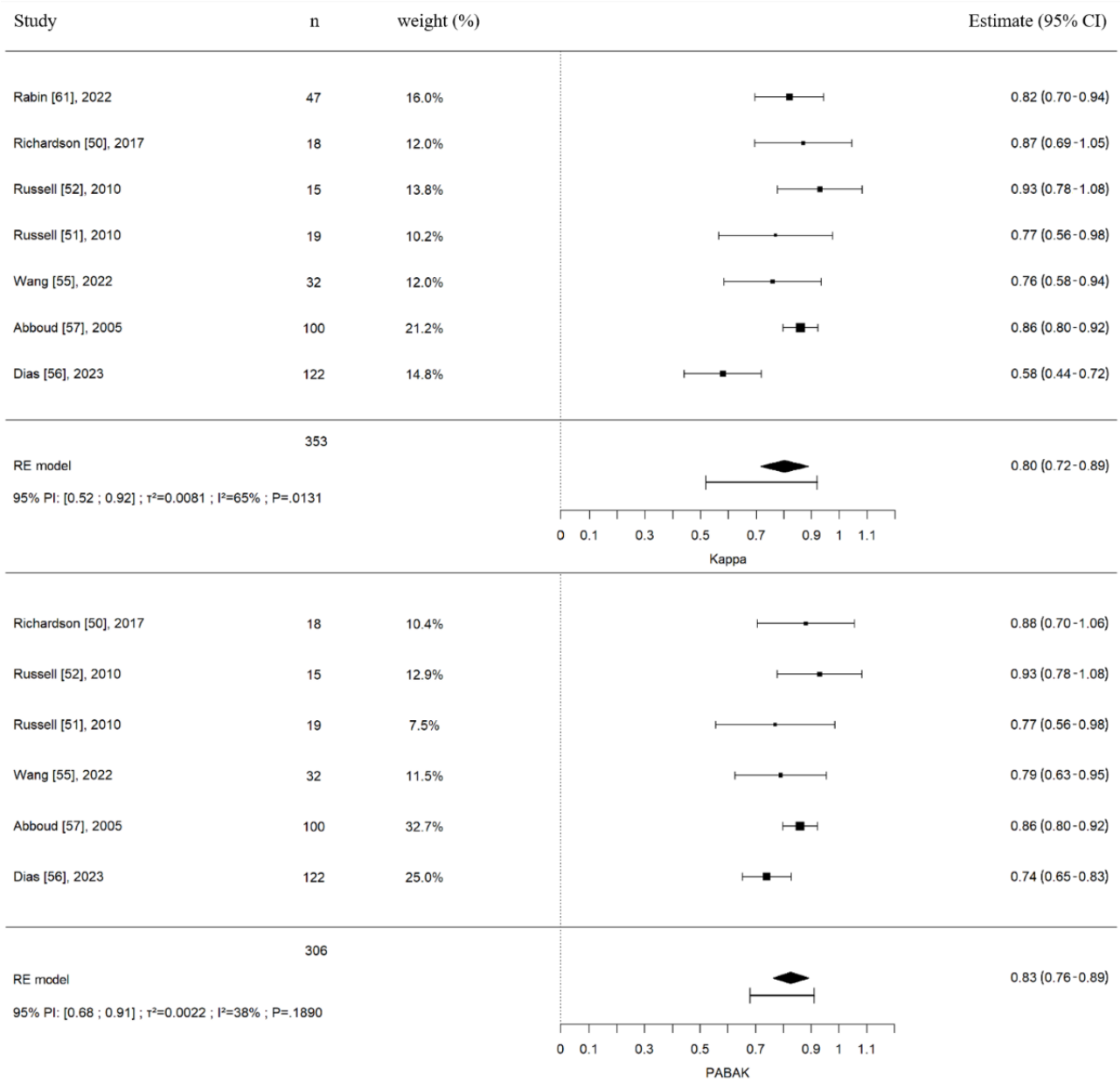

Another study could not be pooled in previous analyses as Fleiss κs were used to evaluate the agreement between in-person and remote assessments [42]. Diagnostic concordance between in-person and remote assessments for the study by Exposto et al [42] for patients with temporomandibular disorders reported Fleiss κs between 0.52 and 1.0, depending on the specific temporomandibular disorders (myalgia of the masseter or the temporalis muscles, temporomandibular arthralgia, or disk displacement with reduction).

Subgroup analyses were performed according to the presence or not of a third party to evaluate the participant during the remote evaluation (Figure S1 in Multimedia Appendix 4), profession of the health care providers (Figures S2 and S3 in Multimedia Appendix 4), body regions (Figures S2 and S4 in Multimedia Appendix 4), and asynchronous or synchronous

XSL•FO
RenderX

evaluation forms of telemedicine used (Figures S5 and S6 in Multimedia Appendix 4). Similar results for diagnostic concordance were reported for all these subgroup analyses. For body regions, subgroup analyses were performed specific to upper and lower limb MSKDs. Analyses excluding high risk-of-bias studies (studies with a QUADAS-2 without items evaluated at high risk of bias) reported similar results for diagnostic concordance (Figure S7 in Multimedia Appendix 4).

Treatment Plan Concordance Between Remote and In-Person Assessments

Of the 23 studies, 8 (35%) reported treatment plan concordance outcomes between in-person and remote assessments and 2 (9%) [62,65] reported treatment plan changes between in-person and remote assessments. Overall raw agreement was 84% (406/483 patients). Treatment plan concordance between in-person and remote assessments pooled estimations were only possible for 2 (9%) of the 23 studies for Gwet AC1 coefficient estimates. The pooled estimates are presented in Figure 3. The pooled Gwet AC1 coefficient was 0.90 (95% CI 0.80-0.99; 2 studies, 142 patients), corresponding to a good to excellent agreement between both types of assessments. Statistical heterogeneity (I^2=50%) was potentially moderate for this meta-analysis. Prediction intervals could not be calculated as this meta-analysis is based on only 2 studies.

Figure 3. Treatment plan concordance (Gwet AC1 coefficients) between in-person and remote assessment in patients with musculoskeletal disorders. Meta-analysis with pooled Gwet AC1 coefficients included studies with advanced practice physiotherapists (6/9, 67%) and orthopedic surgeons (3/9, 33%). Gwet: Gwet AC1 coefficient; RE: random-effect model; τ^2/I^2/P: test for heterogeneity [57,59].

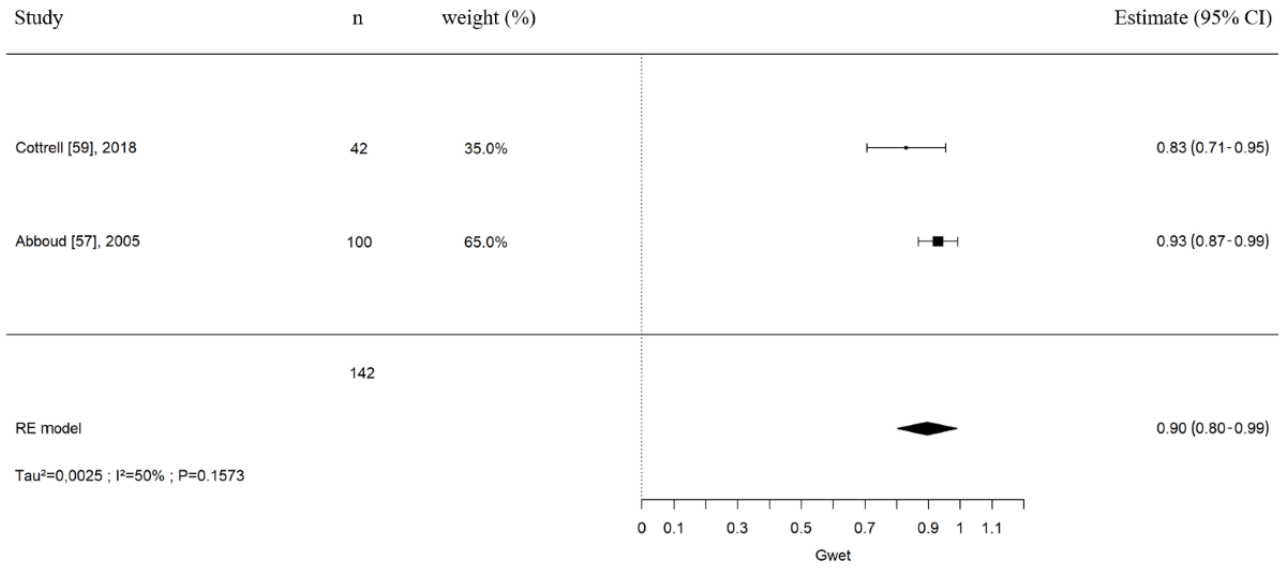

Of the 23 studies, 2 (9%) reporting κ estimates (2 studies, n=147) were not pooled together because of statistical heterogeneity (I^2=94.79%). The first study by Abboud et al [57] compared the treatment plan concordance among 3 orthopedic surgeons for patients with upper limb MSKDs and reported an excellent agreement with a κ of 0.91 (95% CI 0.85-0.97; 100 patients). The other study by Rabin et al [61] compared the treatment plan concordance between 2 orthopedic surgeons for patients with shoulder pathologies. The authors reported only moderate agreement between remote and in-person proposed treatment plans with a κ of 0.43 (95% CI 0.22-0.64; 47 patients). Another study was not pooled in these analyses as the evaluators were physiotherapists and had to choose among 3 physiotherapy treatments for low back pain patients (manipulation, specific exercises, or stabilization exercises). In the study by Peterson et al [63], the treatment plan concordance between in-person and remote assessments evaluated was only moderate with a κ of 0.52 (95% CI 0.32-0.72) and a PABAK of 0.52 (95% CI 0.31-0.72).

GRADE Analysis

Pooled results for diagnostic concordance were of high certainty, highlighting a good to very good concordance between remote and in-person assessments (Figure 4).

XSL•FO
RenderX

Figure 4. Summary of the findings and Grading of Recommendations, Assessment, Development, and Evaluations analysis for diagnostic and treatment plan concordance.

Patient or population: Musculoskeletal disorders
Setting: Primary care, orthopaedic specialized care and physiotherapy
Intervention: Remote patient assessment
Comparison: In-person patient assessment

Outcomes	Pooled concordance [95%CI] Prediction interval (95%PI)	Number of participants (studies)	Certainty of the evidence (GRADE)	Comments
Overall diagnostic concordance kappa	**0.80** [0.72 - 0.89] (0.52 - 0.92)	353[1] (7)	⊕⊕⊕⊕ High	Diagnostic concordance of patients with MSKDs between remote and in-person assessment is good to very good.
Overall diagnostic concordance PABAK	**0.83** [0.76 – 0.89] (0.68 - 0.91)	306[1] (6)		
Treatment plan concordance Gwet	**0.90** [0.80 – 0.99]	142[2] (2)	⊕⊕⊕◯ Moderate[a]	Treatment plan concordance of patients with MSKDs between remote and in-person assessment could be good to excellent. More studies are needed to confirm these results.

1. Diagnostic concordance pooled results included few degenerative pathologies such as osteoarthritis and only low back pain for spinal musculoskeletal disorders. Evaluators included 23 orthopedic surgeons, 6 physiotherapists, and 3 physiotherapy students as assessors for pooled kappa results and 21 orthopedic surgeons, 6 physiotherapists, and 3 physiotherapy students for pooled prevalence-adjusted and bias-adjusted kappa results.

2. Treatment plan concordance results included 3 orthopedic surgeons and 6 advanced practice physiotherapists.

a. Certainty was downgraded for heterogeneity.

GRADE Grading of Recommendations, Assessment, Development, and Evaluations; **GWET** Gwet's AC1 coefficient; **MSKDs** Musculoskeletal disorders; **PABAK** Prevalence-adjusted, bias-adjusted Kappa; **95%CI** 95% Confidence interval; **95%PI** 95% Prediction interval

Pooled results for treatment plan concordance were of moderate certainty, highlighting a good to excellent concordance between remote and in-person assessments (Figure 4). The I^2 was at 50%, and prediction intervals could not be calculated; therefore, certainty of the results was downgraded.

Discussion

Principal Findings

The aim of this systematic review and meta-analysis was to appraise the available evidence on diagnosis and care concordance after an initial assessment between a remote evaluation and an in-person evaluation for the initial evaluation of various MSKDs. On the basis of our results, health care providers can remotely make a diagnosis that is concordant with a usual in-person assessment. These results are robust and among various populations with acute or chronic MSKDs; we conducted several secondary analyses to assess the impact of various factors that may affect the validity and the feasibility of a remote assessment. Presence of assistance with the patient, body region being assessed, technologies used, and different health care providers' profession did not appear to influence the accuracy of the remote assessment, but these findings need to be confirmed due to the limited number of studies in certain analyses. Sensitivity analyses, including only studies without high risk of bias and results from the study by Exposto et al [42], not pooled in the meta-analysis, yielded similar results. The study by Dias et al [56] reported a lower κ than the rest of the pooled studies but a similar raw agreement and PABAK (not calculated in the original paper but extracted and calculated by our team). The distribution of diagnoses in the study by Dias et al [56] shows a very high prevalence of some diagnostic categories and a very low prevalence of others. This imbalance may have biased and lowered the κ statistical value [68,69].

XSL•FO
RenderX

Overall concordance was found to be high in our review, but it may still vary across certain MSKDs, and our secondary analyses with fewer study and participants may not have been able to capture this issue. Moreover, studies on various spinal disorders, such as neck pain, or on degenerative pathologies, such as osteoarthritis, which were only marginally represented in the included studies, but very prevalent, are also needed.

Treatment plans proposed remotely had a good to excellent concordance with the treatment plans proposed in person. It may seem logical that if the remote diagnosis is correct, the treatment plan would be similar. However, there is uncertainty about these results because overall less data were available on the treatment plan in the literature; only 2 studies could be pooled and showed potentially moderate heterogeneity. Raw agreement between remote and in-person treatment plan was 84% (406/483 patients; 9 studies), which can be considered high. Again, these results are promising but need to be confirmed by more studies using statistical estimates to conclude on treatment plan concordance.

However, it should be kept in mind that an initial assessment is not only a diagnosis and a treatment plan. Some other key elements were not or could not have been properly evaluated with a remote assessment in several of the included studies. No data are available on the detection of serious pathologies (red flags) or the evaluation psychosocial risk factors (yellow flags) using a remote assessment [24]. Most red or yellow flags are identified during the subjective examination by elements of the patient's history, specific questions, or questionnaires [8,70]. They may also be identifiable during a remote assessment by using the same screening procedure as during an in-person assessment. However, some tests used to rule out the presence of serious pathologies cannot be performed remotely, such as certain tests of the neurological examination, and patients may be less likely to talk about the psychosocial problems they are experiencing during a remote video or phone encounter. Among the studies we included, none reported having identified serious pathologies. However, given the small sample sizes of the studies and the low prevalence of serious conditions, it is unlikely that several patients with a serious condition were in the included samples [8]. Caution is therefore warranted when concluding that a remote assessment is as safe as traditional in-person assessment and that patients with important psychosocial risk factors are accurately identified. This is the first systematic review with meta-analyses addressing the agreement between remote and in-person diagnoses and treatment plan for MSKDs. Our results are consistent with previous reviews investigating the validity and reliability of different parts of a remote physical examination for MSKDs. Previous reviews have highlighted that most clinical measures assessed by videoconferencing applications, such as pain (visual analog scale) or range of motion assessment in patients with MSKDs, have good concurrent validity and excellent reliability [24,36]. Nevertheless, some measures, such as orthopedic tests, which are usually hands on, have lower concurrent validity and reliability when assessed remotely or simply cannot be performed [47]. Adaptations to the usual in-person physical examination in the musculoskeletal field still need to be developed to achieve validity and reliability comparable to in-person physical examination or to establish that the accuracy of a remote assessment is not as high [24,36,71]. Our results are complementary to other reviews with meta-analyses conducted on remote management of patients with MSKDs that show that remote care after an in-person assessment leads to similar improvements in clinical outcomes as the usual fully in-person approach [32-35]. Guidelines for MSKDs management now emphasize the importance of patient education and self-management as well as physical activity and exercise prescription [2]; all these interventions can be performed remotely [72,73]. In conjunction with our conclusions, the actual body of evidence on remote patient follow-up and clinical measures as well as recommended interventions for MSKDs potentially support the benefits of full remote care pathways for patients with MSKDs.

As stated earlier, more evidence is still needed on the impact of remote assessment for MSKDs. Future studies should also focus on the potential impacts of an initial remote evaluation on health care resources use, such as the use of imaging and other paraclinical investigations, as telemedicine could increase their use. Clinicians and patients express doubts on the validity and safety of a complete remote evaluation because of the impossibility of a hands-on assessment [29,74-76]. This uncertainty induced by the use of telemedicine could increase their use to confirm their diagnosis. There could also be an impact on follow-up visits and referrals to other health care providers and on modification of the relationship between the health care provider and the patient (therapeutic alliance) due to a first video or phone encounter. Telemedicine being an evolution in professional practices, quality standards and proper training of health care providers as well as the integration into the curriculum for trainees must also be anticipated to ensure the quality of remote care and promote the integration of telehealth [28,77,78]. Moreover, digital innovations, such as tools for clinical decision-making aids and improvements in information communication technologies, could compensate for the loss of hands-on assessment and could promote the acceptability of telemedicine by providing additional tools to help clinicians with evaluations and treatments [79-81].

Particular attention should as well be focused on the implementation of telemedicine that does not exclude populations, particularly in rural areas and low-income populations that may be affected by inequalities in access to information and communication technologies, such as access to reliable high-speed internet connections [82]. It is important to develop remote assessment methods using devices available to the public (smartphone and tablet) in contexts similar to clinical reality and not to neglect research of alternatives for populations without access to a high-speed internet connection, such as assessment by phone [60,66,67].

Strengths and Limitations

Strengths of this review include the use of 4 major bibliographical databases, a comprehensive search strategy, the use of the validated QUADAS-2 score to assess methodological quality of included studies, and the use of the Grading of Recommendations, Assessment, Development and Evaluations approach to rate the certainty of the evidence. Prediction

XSL•FO
RenderX

intervals were used to assess the impact of the heterogeneity on the results [45]. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed to ensure a robust review methodology [83]. Furthermore, this study is one of the first meta-analyses to pool concordance outcomes using novel validated statistical approaches [40,84].

However, some limitations should be highlighted. The presence of publication bias could not be assessed because there were <10 studies to conduct a funnel plot [85]. Our results highlighted that the accuracy of remote diagnoses was not influenced by the type of health care provider or the body region involved. However, the studies pooled in the meta-analyses included only orthopedic surgeons or physiotherapists to evaluate mostly adults with peripheral MSKDs. The accuracy of remote diagnoses made by other health care providers or in different clinical settings, such as rheumatology, or for spinal disorders, such as neck pain, could be different. Moreover, the included studies were carried out in experimental settings, which could overestimate the accuracy of the diagnoses compared to real-life conditions when participants are at home and not in a medical environment. Most of the studies (3/6, 50%) included in the meta-analyses used technologies developed specifically for telemedicine, ensuring an optimal quality of exchange between the evaluator and the participant. The limited number of studies included for treatment plans should also be highlighted. Therefore, more methodologically sound studies with larger sample size are needed.

Conclusions

Health care providers can remotely make a diagnosis that is concordant with a usual in-person assessment for various MSKDs. Diagnostic concordance between remote and in-person assessments is good to very good, and future studies are unlikely to modify this conclusion. Type of health care provider, body region, technology used, and the presence of assistance with the patient do not seem to influence the accuracy of remote diagnoses, but more studies are needed to confirm the impact of such factors. Health care providers offer treatment plans that are consistent with those usually proposed in person. Treatment plan concordance between remote and in-person assessments is probably good to excellent, but future studies may modify these conclusions. There are still some considerations, such as the economic impact, modification of therapeutic alliance between health care provider and patient, or detection of serious pathologies and psychosocial risk factors, to investigate to fully appraise the challenges and benefits of the initial remote evaluation for patients with MSKDs. This systematic review with meta-analysis adds support for further development of hybrid or fully remote care pathways for patients with MSKDs to facilitate access to musculoskeletal care.

Acknowledgments

The authors would like to express their gratitude to James M Gregory, Yuri Martins Costa, and Seth Peterson for sharing their data to perform the required analyses and to Claudia Cosculluela for her participation in the risk-of-bias analysis. FD is an Fonds de Recherche du Québéc-Santé FRQ-S Senior Scholar, DK is an Fonds de Recherche du Québéc-Santé Junior-2 Scholar and RV is supported by PhD scholarships from the Faculté de médecine de l'Université de Montréal, the École de réadaptation de l'Université de Montréal, and the Réseau provincial de recherche en adaptation-réadaptation.

Data Availability

All analyses are available in Multimedia Appendix 4. The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 33 KB - jmir_v26i1e52964_app1.docx]

Multimedia Appendix 2
Codes used in RStudio.
[DOCX File , 13 KB - jmir_v26i1e52964_app2.docx]

Multimedia Appendix 3
Manuscripts excluded with reasons.
[DOCX File , 19 KB - jmir_v26i1e52964_app3.docx]

Multimedia Appendix 4
Supplementary materials (characteristics of conference abstracts and subgroup analyses).

XSL•FO
RenderX

[DOCX File , 2694 KB - jmir_v26i1e52964_app4.docx]

Multimedia Appendix 5
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 68 KB - jmir_v26i1e52964_app5.pdf]

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
2. Lin I, Wiles L, Waller R, Goucke R, Nagree Y, Gibberd M, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan 02;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
3. Safiri S, Kolahi AA, Cross M, Hill C, Smith E, Carson-Chahhoud K, et al. Prevalence, deaths, and disability-adjusted life years due to musculoskeletal disorders for 195 countries and territories 1990-2017. Arthritis Rheumatol 2021 Apr;73(4):702-714 [FREE Full text] [doi: 10.1002/art.41571] [Medline: 33150702]
4. Briggs AM, Cross MJ, Hoy DG, Sànchez-Riera L, Blyth FM, Woolf AD, et al. Musculoskeletal health conditions represent a global threat to healthy aging: a report for the 2015 World Health Organization world report on ageing and health. Gerontologist 2016 Apr 18;56 Suppl 2(Suppl 2):S243-S255. [doi: 10.1093/geront/gnw002] [Medline: 26994264]
5. Bevan S. Economic impact of musculoskeletal disorders (MSDs) on work in Europe. Best Pract Res Clin Rheumatol 2015 Jun;29(3):356-373. [doi: 10.1016/j.berh.2015.08.002] [Medline: 26612235]
6. Moir M, Barua B. Waiting your turn: wait times for health care in Canada, 2021 report. Fraser Institute. 2021 Dec 15. URL: https://www.fraserinstitute.org/studies/waiting-your-turn-wait-times-for-health-care-in-canada-2021 [accessed 2023-08-23]
7. Ojha HA, Wyrsta NJ, Davenport TE, Egan WE, Gellhorn AC. Timing of physical therapy initiation for nonsurgical management of musculoskeletal disorders and effects on patient outcomes: a systematic review. J Orthop Sports Phys Ther 2016 Feb;46(2):56-70. [doi: 10.2519/jospt.2016.6138] [Medline: 26755406]
8. Finucane LM, Downie A, Mercer C, Greenhalgh SM, Boissonnault WG, Pool-Goudzwaard AL, et al. International framework for red flags for potential serious spinal pathologies. J Orthop Sports Phys Ther 2020 Jul;50(7):350-372. [doi: 10.2519/jospt.2020.9971] [Medline: 32438853]
9. Grimmer K, Bowman P. Differences between metropolitan and country public hospital allied health services. Aust J Rural Health 1998 Nov;6(4):181-188. [doi: 10.1111/j.1440-1584.1998.tb00310.x] [Medline: 9919074]
10. Bath B, Trask C, McCrosky J, Lawson J. A biopsychosocial profile of adult Canadians with and without chronic back disorders: a population-based analysis of the 2009-2010 Canadian Community Health Surveys. Biomed Res Int 2014;2014:919621 [FREE Full text] [doi: 10.1155/2014/919621] [Medline: 24971357]
11. How healthy are rural Canadians? : an assessment of their health status and health determinants. Canadian Institute for Health Information. 2006. URL: https://secure.cihi.ca/free_products/rural_canadians_2006_report_e.pdf [accessed 2023-08-23]
12. Ono T, Schoenstein M, Buchan J. Geographic imbalances in doctor supply and policy responses. Organisation for Economic Cooperation and Development. 2014 Apr 3. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/geographic-imbalances-in-doctor-supply-and-policy-responses_5jz5sq5ls1wl-en [accessed 2023-08-23]
13. Huguet M. Centralization of care in high volume hospitals and inequalities in access to care. Soc Sci Med 2020 Sep;260:113177 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113177] [Medline: 32712556]
14. Eberhardt MS, Pamuk ER. The importance of place of residence: examining health in rural and nonrural areas. Am J Public Health 2004 Oct;94(10):1682-1686. [doi: 10.2105/ajph.94.10.1682] [Medline: 15451731]
15. Health at a glance: Europe 2022. Organisation for Economic Co-operation and Development. 2022 Dec 5. URL: https://www.oecd.org/en/publications/health-at-a-glance-europe-2022_507433b0-en.html [accessed 2023-08-23]
16. Enquête sur l'accès aux soins de santé et aux produits pharmaceutiques pendant la pandémie (EASSPPP). Statistics Canada. 2021 Jun 2. URL: https://www.statcan.gc.ca/fr/enquete/menages/5346 [accessed 2023-08-23]
17. Pujolar G, Oliver-Anglès A, Vargas I, Vázquez ML. Changes in access to health services during the COVID-19 pandemic: a scoping review. Int J Environ Res Public Health 2022 Feb 03;19(3):1749 [FREE Full text] [doi: 10.3390/ijerph19031749] [Medline: 35162772]
18. Deslauriers S, Déry J, Proulx K, Laliberté M, Desmeules F, Feldman DE, et al. Effects of waiting for outpatient physiotherapy services in persons with musculoskeletal disorders: a systematic review. Disabil Rehabil 2021 Mar 14;43(5):611-620. [doi: 10.1080/09638288.2019.1639222] [Medline: 31304824]
19. Report of the WCPT/INPTRA digital physical therapy practice task force. International Network of Physiotherapy Regulatory Authorities. 2019 May 15. URL: https://www.med.unc.edu/healthsciences/physical/schoolbasedpt/wp-content/uploads/sites/662/2020/04/WCPTINPTRA_DigitalPTPractice.pdf [accessed 2023-08-23]

XSL•FO
RenderX

20. Marcin JP, Shaikh U, Steinhorn RH. Addressing health disparities in rural communities using telehealth. Pediatr Res 2016 Jan 14;79(1-2):169-176 [FREE Full text] [doi: 10.1038/pr.2015.192] [Medline: 26466080]
21. Moffatt JD, Eley DS. The reported benefits of telehealth for rural Australians. Aust Health Rev 2010 Aug 25;34(3):276-281 [FREE Full text] [doi: 10.1071/ah09794]
22. Jagarapu JR, Savani RC. A brief history of telemedicine and the evolution of teleneonatology. Semin Perinatol 2021 Aug;45(5):151416 [FREE Full text] [doi: 10.1016/j.semperi.2021.151416] [Medline: 34006382]
23. Ganjali R, Jajroudi M, Kheirdoust A, Darroudi A, Alnattah A. Telemedicine solutions for clinical care delivery during COVID-19 pandemic: a scoping review. Front Public Health 2022 Jul 22;10:937207 [FREE Full text] [doi: 10.3389/fpubh.2022.937207] [Medline: 35937265]
24. Mani S, Sharma S, Omar B, Paungmali A, Joseph L. Validity and reliability of internet-based physiotherapy assessment for musculoskeletal disorders: a systematic review. J Telemed Telecare 2016 Mar 31;23(3):379-391. [doi: 10.1177/1357633x16642369]
25. WHO Global Observatory for eHealth. Telemedicine: opportunities and developments in Member States: report on the second global survey on eHealth. World Health Organization. 2010. URL: https://iris.who.int/handle/10665/44497 [accessed 2023-08-23]
26. El Aoufy K, Melis MR, Bellando Randone S, Blagojevic J, Bartoli F, Fiori G, et al. The positive side of the coin: Sars-Cov-2 pandemic has taught us how much telemedicine is useful as standard of care procedure in real life. Clin Rheumatol 2022 Feb 05;41(2):573-579 [FREE Full text] [doi: 10.1007/s10067-021-05975-2] [Medline: 34739619]
27. Farzad M, MacDermid J, Ferreira L, Szekeres M, Cuypers S, Shafiee E. A description of the barriers, facilitators, and experiences of hand therapists in providing remote (tele) rehabilitation: an interpretive description approach. J Hand Ther 2023 Oct;36(4):805-816. [doi: 10.1016/j.jht.2023.06.004] [Medline: 37591730]
28. Rethorn ZD, Lee AC, Rethorn TJ. Connecting at the webside: rapid telehealth implementation for musculoskeletal clinicians. J Orthop Sports Phys Ther 2021 Jan;51(1):8-11. [doi: 10.2519/jospt.2021.9902] [Medline: 33076758]
29. Renard M, Gaboury I, Michaud F, Tousignant M. The acceptability of two remote monitoring modalities for patients waiting for services in a physiotherapy outpatient clinic. Musculoskeletal Care 2022 Sep;20(3):616-624 [FREE Full text] [doi: 10.1002/msc.1622] [Medline: 35142425]
30. Malliaras P, Merolli M, Williams CM, Caneiro JP, Haines T, Barton C. 'It's not hands-on therapy, so it's very limited': telehealth use and views among allied health clinicians during the coronavirus pandemic. Musculoskelet Sci Pract 2021 Apr;52:102340 [FREE Full text] [doi: 10.1016/j.msksp.2021.102340] [Medline: 33571900]
31. Rabanifar N, Hoseini MA, Abdi K. Exploring barriers to implementing telerehabilitation from experiences of managers, policymakers, and providers of rehabilitation services in Iran: a qualitative study. Med J Islam Repub Iran 2022 Dec 22;36:157 [FREE Full text] [doi: 10.47176/mjiri.36.157] [Medline: 36660003]
32. Cottrell MA, Galea OA, O'Leary SP, Hill AJ, Russell TG. Real-time telerehabilitation for the treatment of musculoskeletal conditions is effective and comparable to standard practice: a systematic review and meta-analysis. Clin Rehabil 2017 May 02;31(5):625-638. [doi: 10.1177/0269215516645148] [Medline: 27141087]
33. Suso-Martí L, La Touche R, Herranz-Gómez A, Angulo-Díaz-Parreño S, Paris-Alemany A, Cuenca-Martínez F. Effectiveness of telerehabilitation in physical therapist practice: an umbrella and mapping review with meta-meta-analysis. Phys Ther 2021 May 04;101(5):101 [FREE Full text] [doi: 10.1093/ptj/pzab075] [Medline: 33611598]
34. Gava V, Ribeiro LP, Barreto RP, Camargo PR. Effectiveness of physical therapy given by telerehabilitation on pain and disability of individuals with shoulder pain: a systematic review. Clin Rehabil 2022 Jun 01;36(6):715-725. [doi: 10.1177/02692155221083496] [Medline: 35230167]
35. Krzyzaniak N, Cardona M, Peiris R, Michaleff ZA, Greenwood H, Clark J, et al. Telerehabilitation versus face-to-face rehabilitation in the management of musculoskeletal conditions: a systematic review and meta-analysis. Phys Ther Rev 2023 Apr 12;28(2):71-87. [doi: 10.1080/10833196.2023.2195214]
36. Zischke C, Simas V, Hing W, Milne N, Spittle A, Pope R. The utility of physiotherapy assessments delivered by telehealth: a systematic review. J Glob Health 2021 Dec 18;11:04072 [FREE Full text] [doi: 10.7189/jogh.11.04072] [Medline: 34956637]
37. Vincent C, Lafrance S, Charron M, Kairy D, Desmeules F. Remote assessment and management of patients with musculoskeletal disorders: a systematic review with meta-analyses. PROSPERO. 2022. URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=335606 [accessed 2024-09-08]
38. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
39. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
40. Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method 2011;11:145-163. [doi: 10.1007/s10742-011-0077-3]

XSL•FO
RenderX

41. Lovo S, Harrison L, O'Connell ME, Rotter T, Bath B. A physical therapist and nurse practitioner model of care for chronic back pain using telehealth: diagnostic and management concordance. J Telemed Telecare 2022 May 12;30(5):842-850. [doi: 10.1177/1357633x221098904]
42. Exposto FG, Castrillon EE, Exposto CR, Costa DM, Gøkhan MA, Svensson P, et al. Remote physical examination for temporomandibular disorders. Pain 2022 May 01;163(5):936-942. [doi: 10.1097/j.pain.0000000000002455] [Medline: 34433771]
43. Higgins JP, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane handbook for systematic reviews of interventions version 6.4. The Cochrane Collaboration. URL: https://training.cochrane.org/handbook [accessed 2023-08-23]
44. Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
45. IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
46. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
47. Bradley KE, Cook C, Reinke EK, Vinson EN, Mather RC3, Riboh J, et al. Comparison of the accuracy of telehealth examination versus clinical examination in the detection of shoulder pathology. J Shoulder Elbow Surg 2021 May;30(5):1042-1052 [FREE Full text] [doi: 10.1016/j.jse.2020.08.016] [Medline: 32871264]
48. Demaerschalk BM, Pines A, Butterfield R, Haglin JM, Haddad TC, Yiannias J, et al. Assessment of clinician diagnostic concordance with video telemedicine in the integrated multispecialty practice at mayo clinic during the beginning of COVID-19 pandemic from March to June 2020. JAMA Netw Open 2022 Sep 01;5(9):e2229958 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.29958] [Medline: 36053531]
49. Lade H, McKenzie S, Steele L, Russell TG. Validity and reliability of the assessment and diagnosis of musculoskeletal elbow disorders using telerehabilitation. J Telemed Telecare 2012 Oct 19;18(7):413-418. [doi: 10.1258/jtt.2012.120501] [Medline: 23086982]
50. Richardson BR, Truter P, Blumke R, Russell TG. Physiotherapy assessment and diagnosis of musculoskeletal disorders of the knee via telerehabilitation. J Telemed Telecare 2016 Jul 08;23(1):88-95. [doi: 10.1177/1357633x15627237]
51. Russell T, Truter P, Blumke R, Richardson B. The diagnostic accuracy of telerehabilitation for nonarticular lower-limb musculoskeletal disorders. Telemed J E Health 2010 Jun;16(5):585-594. [doi: 10.1089/tmj.2009.0163] [Medline: 20575726]
52. Russell TG, Blumke R, Richardson B, Truter P. Telerehabilitation mediated physiotherapy assessment of ankle disorders. Physiother Res Int 2010 Sep 17;15(3):167-175. [doi: 10.1002/pri.471] [Medline: 20812313]
53. Steele L, Lade H, McKenzie S, Russell TG. Assessment and diagnosis of musculoskeletal shoulder disorders over the internet. Int J Telemed Appl 2012;2012:945745 [FREE Full text] [doi: 10.1155/2012/945745] [Medline: 23193395]
54. Tachakra S, Lynch M, Newson R, Stinson A, Sivakumar A, Hayes J, et al. A comparison of telemedicine with face-to-face consultations for trauma management. J Telemed Telecare 2000 Dec 02;6 Suppl 1(1_suppl):S178-S181. [doi: 10.1258/1357633001934591] [Medline: 10794013]
55. Wang G, Fiedler AK, Warth RJ, Bailey L, Shupe PG, Gregory JM. Reliability and accuracy of telemedicine-based shoulder examinations. J Shoulder Elbow Surg 2022 Aug;31(8):e369-e375. [doi: 10.1016/j.jse.2022.04.005] [Medline: 35562030]
56. Dias JMJ, Mendes AFJ, Pestana de Aguiar E, Silveira LC, Noel Dias MF, Barbosa Raposo NR. Interobserver agreement and satisfaction with the use of telemedicine for evaluating low back pain: a primary, observational, cross-sectional, analytical study. Global Spine J 2023 Aug 16:21925682231194453 (forthcoming) [FREE Full text] [doi: 10.1177/21925682231194453] [Medline: 37585445]
57. Abboud JA, Bozentka DJ, Beredjiklian PK. Telemedicine consultation for patients with upper extremity disorders is reliable. Clin Orthop Relat Res 2005 Jun;435(435):250-257. [doi: 10.1097/01.blo.0000155009.77961.5c] [Medline: 15930947]
58. Bovonratwet P, Song J, Kim YE, Shinn D, Morse KW, Dowdell JE, et al. Telemedicine visits can generate highly accurate diagnoses and surgical plans for spine patients. Spine (Phila Pa 1976) 2022 Sep 01;47(17):1194-1202. [doi: 10.1097/BRS.0000000000004387] [Medline: 35797655]
59. Cottrell MA, O'Leary SP, Swete-Kelly P, Elwell B, Hess S, Litchfield MA, et al. Agreement between telehealth and in-person assessment of patients with chronic musculoskeletal conditions presenting to an advanced-practice physiotherapy screening clinic. Musculoskelet Sci Pract 2018 Dec;38:99-105. [doi: 10.1016/j.msksp.2018.09.014] [Medline: 30366292]
60. Turner D. An exploratory study of physiotherapy telephone assessment. Int J Therapy Rehab 2009 Feb;16(2):97-105. [doi: 10.12968/ijtr.2009.16.2.38897]
61. Rabin A, Dolkart O, Kazum E, Wengier R, Goldstein Y, Maman E, et al. Shoulder assessment by smartphone: a valid alternative for times of social distancing. Arch Orthop Trauma Surg 2022 Jun 13;142(6):979-985 [FREE Full text] [doi: 10.1007/s00402-021-03762-x] [Medline: 33439302]
62. Michaelson EM, Wiesel BB, Siedlarz B, Murthi AM, Sethi PM, Lutton DM, et al. Accuracy of telemedicine for the diagnosis and treatment of patients with shoulder complaints. Curr Orthop Pract 2023;34(3):112-116. [doi: 10.1097/BCO.0000000000001199]

XSL•FO
RenderX

63. Peterson S, Kuntz C, Roush J. Use of a modified treatment-based classification system for subgrouping patients with low back pain: agreement between telerehabilitation and face-to-face assessments. Physiother Theory Pract 2019 Nov 03;35(11):1078-1086. [doi: 10.1080/09593985.2018.1470210] [Medline: 29723124]
64. Worboys T, Brassington M, Ward EC, Cornwell PL. Delivering occupational therapy hand assessment and treatment sessions via telehealth. J Telemed Telecare 2017 Feb 13;24(3):185-192. [doi: 10.1177/1357633x17691861]
65. Melnick K, Porche K, Sriram S, Goutnik M, Cuneo M, Murad G, et al. Evaluation of patients referred to the spine clinic via telemedicine and the impact on diagnosis and surgical decision-making. J Neurosurg Spine 2023 Oct 01;39(4):600-606. [doi: 10.3171/2023.5.SPINE23182] [Medline: 37410606]
66. Higginson R, Hutchinson C. Diagnostic accuracy of knee pathologies by a telephone based Advanced Level Physiotherapy service. Physiotherapy 2017 Dec;103(Supplement 1):e104. [doi: 10.1016/j.physio.2017.11.080]
67. Hutchinson C, Higginson R. The diagnostic accuracy of spinal pathologies referred for by a telephone based advanced level physiotherapy triage service. Physiotherapy 2017 Dec;103(Supplement 1):e99. [doi: 10.1016/j.physio.2017.11.072]
68. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993 May;46(5):423-429 [FREE Full text] [doi: 10.1016/0895-4356(93)90018-v] [Medline: 8501467]
69. Flight L, Julious SA. The disagreeable behaviour of the kappa statistic. Pharm Stat 2015;14(1):74-78 [FREE Full text] [doi: 10.1002/pst.1659] [Medline: 25470361]
70. Stearns ZR, Carvalho ML, Beneciuk JM, Lentz TA. Screening for yellow flags in orthopaedic physical therapy: a clinical framework. J Orthop Sports Phys Ther 2021 Sep;51(9):459-469. [doi: 10.2519/jospt.2021.10570] [Medline: 34465140]
71. Lu AD, Veet CA, Aljundi O, Whitaker E, Smith WB2, Smith JE. A systematic review of physical examination components adapted for telemedicine. Telemed J E Health 2022 Dec 01;28(12):1764-1785. [doi: 10.1089/tmj.2021.0602] [Medline: 35363573]
72. Kennedy CA, Warmington K, Flewelling C, Shupak R, Papachristos A, Jones C, et al. A prospective comparison of telemedicine versus in-person delivery of an interprofessional education program for adults with inflammatory arthritis. J Telemed Telecare 2016 Jul 09;23(2):197-206. [doi: 10.1177/1357633x16635342]
73. van der Kolk NM, de Vries NM, Kessels RP, Joosten H, Zwinderman AH, Post B, et al. Effectiveness of home-based and remotely supervised aerobic exercise in Parkinson's disease: a double-blind, randomised controlled trial. Lancet Neurol 2019 Nov;18(11):998-1008. [doi: 10.1016/S1474-4422(19)30285-6] [Medline: 31521532]
74. Cottrell MA, Hill AJ, O'Leary SP, Raymer ME, Russell TG. Patients are willing to use telehealth for the multidisciplinary management of chronic musculoskeletal conditions: a cross-sectional survey. J Telemed Telecare 2018 Aug;24(7):445-452. [doi: 10.1177/1357633X17706605] [Medline: 28449620]
75. Kairy D, Tousignant M, Leclerc N, Côté AM, Levasseur M, The Telage Researchers. The patient's perspective of in-home telerehabilitation physiotherapy services following total knee arthroplasty. Int J Environ Res Public Health 2013 Aug 30;10(9):3998-4011 [FREE Full text] [doi: 10.3390/ijerph10093998] [Medline: 23999548]
76. Lawford BJ, Bennell KL, Kasza J, Hinman RS. Physical therapists' perceptions of telephone- and internet video-mediated service models for exercise management of people with osteoarthritis. Arthritis Care Res (Hoboken) 2018 Mar 29;70(3):398-408. [doi: 10.1002/acr.23260] [Medline: 28437566]
77. Davies L, Hinman RS, Russell T, Lawford B, Bennell K, International Videoconferencing Steering Group. An international core capability framework for physiotherapists to deliver quality care via videoconferencing: a Delphi study. J Physiother 2021 Oct;67(4):291-297 [FREE Full text] [doi: 10.1016/j.jphys.2021.09.001] [Medline: 34521586]
78. Munce S, Andreoli A, Bayley M, Guo M, Inness EL, Kua A, et al. Clinicians' experiences of implementing a telerehabilitation toolkit during the COVID-19 pandemic: qualitative descriptive study. JMIR Rehabil Assist Technol 2023 Mar 10;10:e44591 [FREE Full text] [doi: 10.2196/44591] [Medline: 36897634]
79. Lowe C, Browne M, Marsh W, Morrissey D. Usability testing of a digital assessment routing tool for musculoskeletal disorders: iterative, convergent mixed methods study. J Med Internet Res 2022 Aug 30;24(8):e38352 [FREE Full text] [doi: 10.2196/38352] [Medline: 36040787]
80. Lervik LC, Vasseljen O, Austad B, Bach K, Bones AF, Granviken F, et al. SupportPrim-a computerized clinical decision support system for stratified care for patients with musculoskeletal pain complaints in general practice: study protocol for a randomized controlled trial. Trials 2023 Apr 11;24(1):267 [FREE Full text] [doi: 10.1186/s13063-023-07272-6] [Medline: 37041631]
81. Létourneau D, Brière S, Maheux MA, Godin C, Warren P, Lauzier G, et al. OpenTera: a framework for telehealth applications. J Open Source Softw 2023 Nov;8(91):5497. [doi: 10.21105/joss.05497]
82. The state of broadband 2020: tackling digital inequalities. International Telecommunication Union and United Nations Educational, Scientific and Cultural Organization. 2020 Sep. URL: https://www.itu.int/dms_pub/itu-s/opb/pol/S-POL-BROADBAND.21-2020-PDF-E.pdf [accessed 2023-08-23]
83. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

84. Lafrance S, Vincent R, Demont A, Charron M, Desmeules F. Advanced practice physiotherapists can diagnose and triage patients with musculoskeletal disorders while providing effective care: a systematic review. J Physiother 2023 Oct;69(4):220-231 [FREE Full text] [doi: 10.1016/j.jphys.2023.08.005] [Medline: 37714771]
85. Page MJ, Higgins JP, Sterne JA. Assessing risk of bias due to missing results in a synthesis. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2019.

Abbreviations

GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
MSKD: musculoskeletal disorder
PABAK: prevalence-adjusted and bias-adjusted κ
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2

Edited by N Cahill; submitted 20.09.23; peer-reviewed by C Zhang, D Singh, R Theiler; comments to author 16.04.24; accepted 24.07.24; published 23.09.24.

Please cite as:
Vincent R, Charron M, Lafrance S, Cormier AA, Kairy D, Desmeules F
Investigating the Use of Telemedicine by Health Care Providers to Diagnose and Manage Patients With Musculoskeletal Disorders: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e52964
URL: https://www.jmir.org/2024/1/e52964
doi:10.2196/52964
PMID:

©Raphaël Vincent, Maxime Charron, Simon Lafrance, Audrey-Anne Cormier, Dahlia Kairy, François Desmeules. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

User Engagement With mHealth Interventions to Promote Treatment Adherence and Self-Management in People With Chronic Health Conditions: Systematic Review

Cyd Eaton[1], PhD; Natalie Vallejo[1]; Xiomara McDonald[1], BS; Jasmine Wu[1]; Rosa Rodríguez[1], BS; Nishanth Muthusamy[1], BA; Nestoras Mathioudakis[1], MHS, MD; Kristin A Riekert[1], PhD

Johns Hopkins School of Medicine, Baltimore, MD, United States

Corresponding Author:
Cyd Eaton, PhD
Johns Hopkins School of Medicine
5200 Eastern Avenue
Baltimore, MD, 21224
United States
Phone: 1 6673066201
Email: ceaton4@jhmi.edu

Abstract

Background: There are numerous mobile health (mHealth) interventions for treatment adherence and self-management; yet, little is known about user engagement or interaction with these technologies.

Objective: This systematic review aimed to answer the following questions: (1) How is user engagement defined and measured in studies of mHealth interventions to promote adherence to prescribed medical or health regimens or self-management among people living with a health condition? (2) To what degree are patients engaging with these mHealth interventions? (3) What is the association between user engagement with mHealth interventions and adherence or self-management outcomes? (4) How often is user engagement a research end point?

Methods: Scientific database (Ovid MEDLINE, Embase, Web of Science, PsycINFO, and CINAHL) search results (2016-2021) were screened for inclusion and exclusion criteria. Data were extracted in a standardized electronic form. No risk-of-bias assessment was conducted because this review aimed to characterize user engagement measurement rather than certainty in primary study results. The results were synthesized descriptively and thematically.

Results: A total of 292 studies were included for data extraction. The median number of participants per study was 77 (IQR 34-164). Most of the mHealth interventions were evaluated in nonrandomized studies (157/292, 53.8%), involved people with diabetes (51/292, 17.5%), targeted medication adherence (98/292, 33.6%), and comprised apps (220/292, 75.3%). The principal findings were as follows: (1) >60 unique terms were used to define user engagement; “use” (102/292, 34.9%) and “engagement” (94/292, 32.2%) were the most common; (2) a total of 11 distinct user engagement measurement approaches were identified; the use of objective user log-in data from an app or web portal (160/292, 54.8%) was the most common; (3) although engagement was inconsistently evaluated, most of the studies (99/195, 50.8%) reported >1 level of engagement due to the use of multiple measurement methods or analyses, decreased engagement across time (76/99, 77%), and results and conclusions suggesting that higher engagement was associated with positive adherence or self-management (60/103, 58.3%); and (4) user engagement was a research end point in only 19.2% (56/292) of the studies.

Conclusions: The results revealed major limitations in the literature reviewed, including significant variability in how user engagement is defined, a tendency to rely on user log-in data over other measurements, and critical gaps in how user engagement is evaluated (infrequently evaluated over time or in relation to adherence or self-management outcomes and rarely considered a research end point). Recommendations are outlined in response to our findings with the goal of improving research rigor in this area.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42022289693; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022289693

(J Med Internet Res 2024;26:e50508) doi:10.2196/50508

XSL•FO
RenderX

KEYWORDS

mobile health; mHealth; digital health; treatment adherence; self-management; user engagement; chronic health conditions; mobile phone

Introduction

Background

As smartphones have become an integral part of modern daily life [1,2], mobile health (mHealth) interventions for treatment adherence and self-management promotion have rapidly developed. These interventions often use existing smartphone features, such as sending SMS text notifications as cues to take prescribed medications [3]; others may integrate external technologies, such as a Bluetooth-enabled glucometer linked to a smartphone app to track blood glucose levels over time with the goal of supporting diabetes management [4]. In general, more frequent engagement, or interaction, with mHealth tools is expected to result in improved treatment adherence or self-management [5]. However, mHealth tools are frequently abandoned by users. Among mobile phone users in the United States, more than half reportedly downloaded an mHealth app but nearly half also stopped using the app due to high data entry burden, low interest, and costs [6].

There is a disconnect between research findings supporting positive correlations between engagement and adherence or self-management and user tendencies to stop using mHealth tools. This discrepancy may reflect an argument put forth by Arigo et al [7] that the mHealth field "lacks a science of engagement." Specifically, there is (1) a lack of consensus in how mHealth user engagement is measured, defined, and reported; (2) no consensus on the optimal level of mHealth user engagement to facilitate meaningful behavior change; and (3) infrequent consideration of user engagement as a research end point. Not treating user engagement as a research end point suggests that this domain is poorly defined and haphazardly evaluated, particularly in terms of how user engagement might evolve over the course of the intervention or relate to behavioral and health outcomes. While the tendency for mHealth use to decline over time suggests that users will not experience maximum benefit from accessing these tools, poor measurement of user engagement has presented a challenge to researchers' abilities to characterize exactly how and why engagement may decrease and how these decreases may affect intervention outcomes. These critical gaps in the science of user engagement significantly limit the utility, effectiveness, uptake, and scalability of mHealth interventions for adherence and self-management promotion.

Recent systematic reviews have examined aspects of user engagement with mHealth interventions for specific diagnoses, including hypertension [8], physical activity [9,10], depressive symptoms [11,12], and mental health conditions [13]. The applicability of these reviews to adherence and self-management mHealth interventions is significantly limited by the small number of studies included and a lack of unified focus on adherence and self-management behaviors. These prior reviews have been further limited by using a very broad definition of *engagement* to include usability, feasibility, user satisfaction, and acceptability [13], limiting the review to studies in which only postassessment retention data were obtained [11] (excludes interventions earlier in the design phase), and focusing on design features associated with user engagement rather than the evaluation of user engagement itself [10]. Another review identified a range of valid and reliable measurement approaches for evaluating user engagement with mHealth interventions for behavior change but used a snowballing method to identify sources rather than a rigorous systematic review of the literature [14]. To encourage the continued use of, and optimal engagement with, mHealth interventions to facilitate adherence and self-management behavior change, it is imperative to comprehensively and systematically evaluate the recent scientific landscape of mHealth user engagement with a clear focus on adherence and self-management behavior.

Objectives

In response to gaps identified in the current scientific literature [7], our registered systematic review aimed to (1) characterize user engagement with interventions promoting mHealth treatment adherence or self-management for adults and youth with health conditions and (2) generate user engagement–focused research recommendations. "User engagement" with mHealth tools can be conceptualized as both behavioral (the extent of use) and experiential (the subjective experience of interacting with the technology) [15]. To enhance the practical application of our review findings, we focused on the behavioral aspects of user engagement to evaluate the degree of use and interaction with the mHealth tool [7] among users with chronic health conditions. Users' behavioral interaction with mHealth tools, features, and associated behavior change components is known as "Little e" engagement. In theory, increased "Little e" engagement is expected to contribute to increased engagement in the desired health behavior, known as "Big E" [16]. Thus, better precision in how behavioral interaction, or "Little e" engagement, is empirically evaluated could help to facilitate greater changes in "Big E" outcomes, thus improving the overall efficacy of mHealth interventions for adherence and self-management. A systematic review approach was selected due to expected heterogeneity in both the measurement of user engagement and adherence and self-management outcomes, which precludes the use of a meta-analysis [17-19]. We specifically aimed to answer the following research questions: (1) How is user engagement defined and measured in studies of mHealth interventions to promote adherence to prescribed medical or health regimens or self-management among people living with a health condition? (2) To what degree are patients engaging with these mHealth interventions? (3) What is the association between user engagement with mHealth interventions and adherence or self-management outcomes? (4) How often is user engagement a research end point?

We also developed the following exploratory question: are there differences in user engagement measurement approaches and

XSL•FO
RenderX

levels between studies that provide monetary compensation and those that do not?

Methods

This systematic review was registered with PROSPERO (CRD42022289693) and prepared in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The review team prepared and followed a standard manual of procedures designed for this systematic review.

Ethical Considerations

Institutional review board approval was not required because the study was not considered human participant research.

Information Sources and Search Strategy

The search strategy (Multimedia Appendix 1) was developed by the first and last authors (CE and KR) in collaboration with an informationist at Johns Hopkins Libraries; implemented in Ovid MEDLINE, Embase, Web of Science, PsycINFO, and CINAHL; and restricted to manuscripts published between 2016 and 2021. All citations returned from the search were imported into our Covidence (Veritas Health Innovation Ltd) [20] database for screening and data extraction.

Eligibility Criteria

The inclusion and exclusion criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Peer-reviewed manuscripts reporting on original qualitative or quantitative investigations published in English between 2016 and 2021
- Participants followed a medical or health regimen or engaged in adherence or self-management activities for a chronic physical or mental health condition
- A mobile health (mHealth) intervention was used by these participants or their caregivers in a home setting, was at least partially automated (could not only include manual 2-way SMS text messaging or video web conferencing), and was accessible on a mobile device (smartphone or tablet device, including internet browser–based programs)
- The primary intervention target was treatment adherence (eg, taking medication, exercising, or following a diet); or self-management of the health condition, often measured by a health outcome associated with treatment adherence behavior (eg, glycated hemoglobin test, viral load, or BMI)
- User engagement (use and interaction) with the mHealth intervention was a measured study outcome, either by objective (eg, app-recorded log-in data) or subjective (eg, user self-report or qualitative interviews) metrics

Exclusion criteria

- Meta-analyses, systematic reviews, published abstracts, dissertations, and published protocols, as well as studies reporting on usability testing or intervention development only

Selection Process

Citations were imported into Covidence; duplicate citations were removed; and title and abstract screening was conducted, followed by a full-text review. At each level of review, 2 separate review team members evaluated each article against the inclusion and exclusion criteria, with discrepancies resolved by CE and N Muthusamy. Any study meeting the inclusion criteria after the full-text review progressed to the data extraction phase (detailed in the next subsection).

Data Extraction

Overview

Data extraction was performed by 2 separate review team members independently of each other using a standard data extraction form developed by the study team, with discrepancies resolved by CE and N Muthusamy. For each study, the reviewers recorded both quantitative and qualitative data relevant to study design (eg, randomized controlled trial and case-control) and methodology (eg, monetary compensation for participation); demographic characteristics of participants; mHealth intervention targets and characteristics; measurement of user engagement, including whether it was a research end point (a key outcome being measured and potentially impacted by participation in the intervention); and the terminology used to describe the behavioral aspects of user engagement (researcher-evaluated user interactions with the mHealth technology).

Study results were summarized as described in the following subsections.

Level of Engagement With the Intervention

This was categorized as “high,” “medium,” “low,” “>1 level reported due to the use of multiple measurement or analytic approaches,” or “not characterized.” Categorizations were assigned based on the language used by the authors to characterize users’ level of engagement with the mHealth intervention (eg, the authors described user engagement with the mHealth intervention as “high”).

Change in Level of Engagement

This was categorized as “increased,” “no change,” “decreased,” “>1 direction reported due to the use of multiple measurement or analytic approaches,” or “not assessed.” For studies that assessed change in engagement over time, categorizations were assigned based on the data presented by the authors (eg, the authors presented data showing that the user engagement measurement decreased over time).

XSL•FO
RenderX

Association With Treatment Adherence or Self-Management Study Outcomes

This was categorized as "higher engagement, positive treatment adherence or self-management outcomes"; "moderate engagement, positive treatment adherence or self-management outcome)"; "lower engagement, positive treatment adherence or self-management outcomes"; "no association"; ">1 association reported due to the use of multiple measurement or analytic approaches"; or "not assessed." For studies that assessed this association, categorizations were assigned based on how the authors reported and framed the study results and conclusions (eg, the authors' reporting and framing of study results and conclusions suggested that higher engagement with the mHealth intervention was associated with positive study outcomes, such as higher treatment adherence or improved self-management outcome or outcomes).

Technology Dosage

This was categorized as "yes, given" (the researchers told participants how often or in what way or ways they should use the mHealth intervention components, such as complete 1 module per week and log medication administration in the app every day) or "no, not given."

Minimum Engagement Research Benchmark

This was categorized as "yes, selected" (the researchers reported in their manuscript that a minimum cutoff for technology engagement was set as an empirical outcome to denote adequate participant engagement; eg, to be adequately engaged, a participant needed to use the Bluetooth-enabled glucometer at least once a day during the study period) or "no, not selected."

No Formal Risk-of-Bias Assessment

We decided not to conduct a formal risk-of-bias assessment, given that the primary aim of this review was to characterize the evaluation and measurement of user engagement rather than certainty in the primary study results. Therefore, it was deemed inappropriate to evaluate the studies using standard risk-of-bias assessment tools.

Synthesis Methods

Statistical analyses were conducted using SPSS software (version 28.0; IBM Corp) [21]. The extracted data were summarized using frequencies and percentages. Methods of measuring user engagement with the mHealth intervention were thematically grouped into discrete measurement categories. Subgroup sensitivity analyses were conducted to examine studies involving pediatric samples (participants aged 0-18 y or aged up to 25 y if the sample was characterized as "pediatric" by the authors) separately from those involving adults only (participants aged >18 y). In these age-based subgroup analyses, of the 292 included studies, 5 (1.7%) were excluded due to including both pediatric and adult participants, and 1 (0.3%) was excluded due to not reporting participant ages. Given that this investigation was designed as a systematic review, no effect measures or meta-regressions were used. No missing summary statistics or data conversions were used.

Results

Search Results and Screening Process

Figure 1 presents the PRISMA flow diagram of the screening process. The initial search returned 3955 citations, from which 70 (1.77%) duplicates were removed. During title and abstract screening, the remaining 3885 studies were screened, and 2736 (70.42%) were excluded. During full-text screening, the remaining 1149 studies were evaluated, and 857 (74.59%) were excluded. The primary reasons for exclusion were as follows: ineligible manuscript type (eg, published abstract; 330/857, 38.5%), user engagement with the mHealth intervention was not measured (206/857, 24%), and ineligible participant population (111/857, 12.9%). The final review included 292 studies [22-313] (refer to Multimedia Appendix 2 for all included studies and characteristics).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the screening process. mHealth: mobile health.

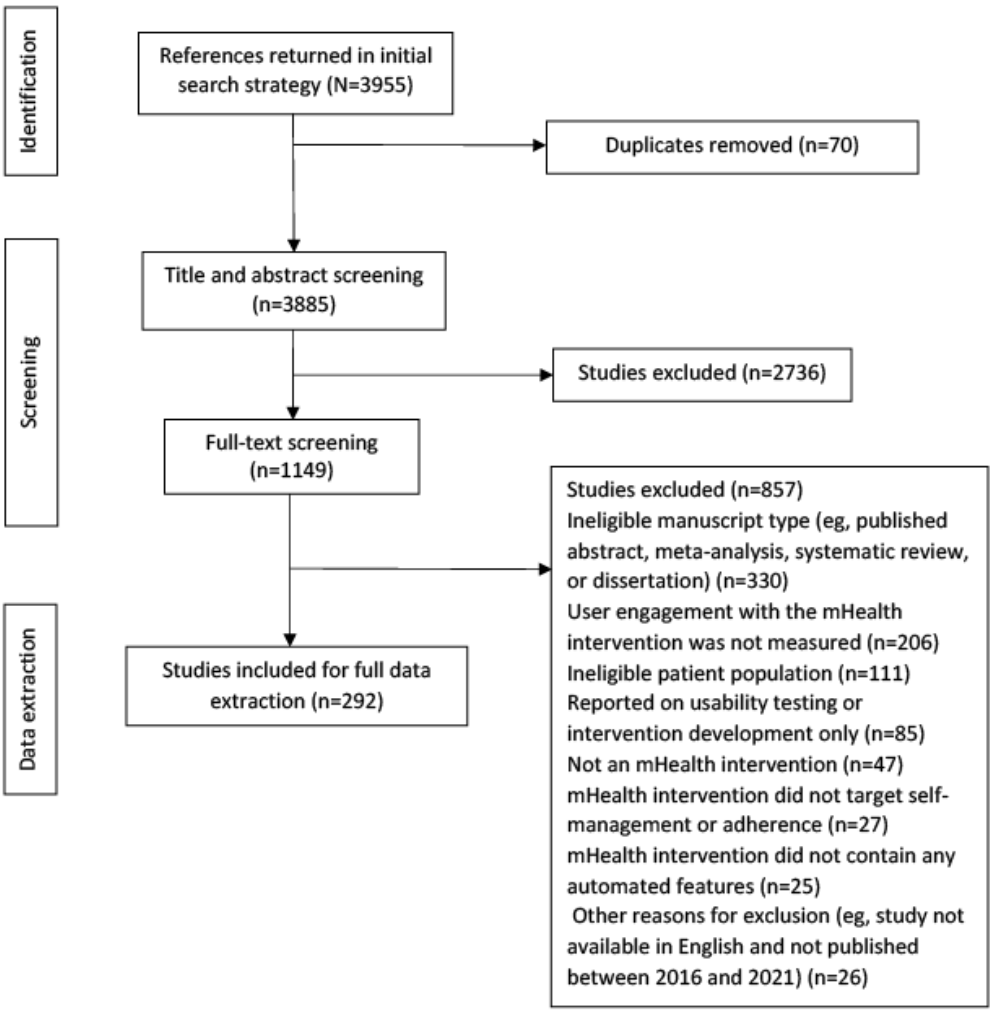

Basic Study and mHealth Intervention Characteristics

Nearly half of the studies (135/292, 46.2%) were conducted in the United States and used a randomized controlled trial design. The median number of participants was 77 (IQR 34-164). Nearly half (135/292, 46.2%) were considered feasibility studies. The median study length was 90 (IQR 60-180) days. Diabetes (51/292, 17.5%) and mental health conditions (35/292, 11.9%) were the most common diagnoses. Study characteristics were similar between adult and pediatric studies, with the exception of health conditions, reflecting expected age-based differences in diagnoses more common among adults than among children (eg, type 2 diabetes and substance use were more commonly studied in adult samples than in pediatric samples).

Table S1 in Multimedia Appendix 3 contains specific details on the mHealth interventions' adherence or self-management targets, intervention components, and intended users. The most frequently targeted adherence or self-management concern was taking medication (98/292, 33.6%), followed by exercise (93/292, 31.8%) and diet (73/292, 25%). Nearly all mHealth interventions were used by the patient (291/292, 99.7%), but some of them included health care providers (58/292, 19.9%) or caregivers (20/292, 6.8%). The majority of mHealth interventions comprised a mobile app (220/292, 75.3%), SMS text messaging or push notifications (74/292, 25.3%), websites or web portals not within a mobile app (48/292, 16.4%), and nonwearable monitoring devices (47/292, 16.1%). Interventions that did not include a mobile app primarily comprised text messaging, a wearable device, a nonwearable device, video web conferencing or telephone calls, or a website or web portal not within a mobile app. Nearly three-quarters of the interventions (216/292, 74%) prompted users to engage with the mHealth intervention. Less than half of the studies (124/292, 42.5%) provided monetary compensation for participation. Only 10.6% (31/292) allowed users to continue using the mHealth intervention after the formal study period. Intervention characteristics were generally similar between adult and pediatric studies, with the exception of intervention target behavior, reflecting expected age-based differences in health concerns more common among adults than among children (eg, exercise, mental health management, and drug or alcohol use or abuse were more commonly targeted in adult samples than in pediatric samples).

How Is User Engagement Defined and Measured?

Definition

Terminology defining user engagement varied widely (Table S2 in Multimedia Appendix 3). There were 33 unique terms

XSL•FO
RenderX

used to define engagement that appeared in at least 2 (0.7%) of the 292 studies. Of the 292 studies, 31 (10.6%) studies each used a unique term that appeared in only that 1 study. “Use” (102/292, 34.9%) and “engagement” (94/292, 32.2%) were the most common terms. Although most terms were synonymous with “use,” “engagement,” or “interaction” with the technology (reflecting our a priori definition of behavioral user engagement and our search strategy), other studies notably used disparate terms, including “acceptability” (26/292, 8.9%), “fidelity” (5/292, 1.7%), “satisfaction” (6/292, 2%), and “perception” (6/292, 2%).

Measurement

Across all studies, 11 distinct user engagement measurement approaches emerged, comprising both objective (n=9, 82%) and subjective (n=2, 18%) methods. User engagement was most frequently evaluated via objective user log-in data from the app or web portal (eg, number of log-ins; 160/292, 54.8%), followed by manually entering data in an app (77/292, 26.4%), qualitative interviews (54/292, 18.5%), and responding to text notifications (49/292, 16.8%). There were “other objective measures” that did not fall into any of the 11 main categories (15/292, 5.1%; eg, notification reading rate or downloading podcasts). These results were similar between adult and pediatric studies (Table 1; Table S3 in Multimedia Appendix 3).

Table 1. How is user engagement measured (n=292)?

Measurement methods	Examples	Studies, n (%)[a]
Objective measures		
User log-in data retrieved from app or website	• Number of log-ins to app or website • Length of time spent in app or website • Frequency of accessing specific features within the app or website	160 (54.8)
Manual user data entry in app-or website-based self-monitoring diaries	• User manually enters data in the app, such as blood glucose level, date and time when medicine was taken, or blood pressure values	77 (26.4)
Response to SMS text messages or push notifications	• User types and sends a response to a SMS text message asking if they took their medicine that day	49 (16.8)
Number or proportion of intervention program modules completed within app or website	• User completes 3 out of 6 possible modules on pain management skills	48 (16.4)
Interacting via chats, phone calls, or social media posts	• Number of times user sends a chat message to care team through app	33 (11.3)
Wearing an electronic monitoring device	• Length of time the user wore a Fitbit device to track daily step count	26 (8.9)
Using a nonwearable electronic monitoring device	• Medication adherence is monitored with an electronic pill bottle that tracks when the bottle is opened and closed to administer medicine	26 (8.9)
Submitting videos via app	• Medication adherence is measured using a mobile app designed to directly observe therapy	5 (1.7)
Other objective measures	• Notification message reading rate or downloading podcasts	13 (4.5)
Subjective measures		
Qualitative interview	• User completes a qualitative interview about their experience using the mHealth[b] app	54 (18.5)
Participant-reported survey	• User self-reports frequency of using the app	29 (9.9)

[a]Percentages do not add up to 100% because studies could fall into >1 category.

[b]mHealth: mobile health.

When examining engagement definitions by measurement approaches, “use” (7/9, 78%) and “engagement” (9/9, 100%) were most commonly used across the nine objective measurement approaches. The exceptions were “wearing an electronic monitoring device” and “submitting videos via app” for which “adherence” (11/26, 42%, 2/5, 40%, respectively) was most commonly used within the measurement approach. Qualitative interviews had the widest range of terminologies used, with “user experience” being the most common (21/54, 39%; Table 2).

XSL•FO
RenderX

Table 2. Associations between user engagement evaluation methods and definitions.

Evaluation methods	Terms[a] used to define user engagement, n (%)[b]
User log-in data retrieved from app or website (n=160)	• Use: 78 (49) • Engagement: 62 (39) • Feasibility: 27 (17) • Adherence: 25 (16)
Manual user data entry in app- or website-based self-monitoring diaries (n=77)	• Use: 31 (40) • Engagement: 24 (31) • Adherence: 19 (25) • Feasibility: 17 (22) • Compliance: 10 (13)
Response to SMS text messages or push notifications (n=49)	• Engagement: 21 (43) • Response: 19 (39)
Number or proportion of intervention program modules completed within app or website (n=48)	• Use: 22 (46) • Engagement: 12 (25) • Adherence: 11 (23)
Interacting via chats, phone calls, or social media posts (n=33)	• Engagement: 14 (42) • Use: 13 (39) • Feasibility: 5 (15) • Compliance: 4 (12)
Wearing an electronic monitoring device (n=26)	• Adherence: 11 (42) • Engagement: 9 (35) • Feasibility: 6 (23)
Using a nonwearable electronic monitoring device (n=26)	• Use: 9 (25) • Adherence: 8 (31) • Engagement: 6 (23)
Submitting videos via app (n=5)	• Adherence: 2 (40) • Use: 2 (40) • Compliance: 1 (20) • Engagement: 1 (20) • Acceptability: 1 (20)
Other objective measures (n=13)	• Engagement: 7 (54) • Feasibility: 2 (15) • Use: 2 (15) • Response: 2 (15)
Qualitative interview (n=54)	• User experience: 21 (39) • Engagement: 18 (33) • Use: 17 (31) • Acceptability: 13 (24) • Feasibility: 8 (15) • Adherence: 8 (15)
Participant-reported survey (n=29)	• Use: 13 (45) • Feasibility: 10 (24) • Engagement: 6 (21) • Acceptability: 6 (21) • Adherence: 5 (17)

[a]We report terms used to describe user engagement in at least 10% of the studies using a given evaluation method; this cutoff was selected to enhance interpretability due to the wide range of terms used to describe user engagement (refer to Table S2 in Multimedia Appendix 3 for details of user engagement definitions).

[b]Percentages within categories do not add up to 100% because studies could fall into >1 category.

The use of user log-in data was the most common measurement method across mHealth intervention components, except for SMS text messaging or push notifications and wearable devices. When SMS text messaging or push notifications was an intervention component, response to SMS text messaging or push notifications was the most common metric (38/74, 51%).

XSL•FO
RenderX

When a wearable device was an intervention component, wearing an electronic monitoring device was the most common metric (26/39, 67%; Table S4 in Multimedia Appendix 3).

To What Degree Are Participants Engaging With These mHealth Interventions?

User Engagement Level

User engagement level was characterized in two-thirds of the reviewed studies (195/292, 66.8%), of which a little more than half (99/195, 50.8%) reported >1 level of engagement due to the use of multiple measurement methods or analyses. Only one-third of the studies (99/292, 33.9%) examined change in engagement over time; when it was examined, engagement tended to decrease (76/99, 77%). These results were similar between adult and pediatric studies (Table 3).

Table 3. Degree of engagement and association with treatment adherence or self-management.

Characteristic	All studies (n=292), n (%)	Adult studies (n=241), n (%)	Pediatric studies (n=45), n (%)
User engagement level			
High	63 (21.6)	55 (22.8)	7 (15.6)
Medium	6 (2.1)	4 (1.7)	2 (4.4)
Low	27 (9.2)	19 (7.9)	6 (13.3)
>1[a] level reported	99 (33.9)	81 (33.6)	18 (40.0)
Not characterized	97 (33.2)	82 (34.0)	12 (26.7)
Change in user engagement level			
Increased	3 (1.0)	3 (0.01)	0 (0)
No change	15 (5.1)	13 (5.4)	1 (2.2)
Decreased	76 (26.0)	61 (25.3)	13 (28.9)
>1 direction reported	5 (1.7)	4 (1.7)	1 (2.2)
Not assessed	193 (66.1)	160 (66.4)	30 (66.7)
Association with adherence or SM[b] outcomes			
Higher engagement, positive adherence or SM outcomes	60 (20.5)	49 (20.3)	8 (17.8)
Moderate engagement, positive adherence or SM outcomes	1 (0.3)	1 (0.4)	0 (0)
Lower engagement, positive adherence or SM outcomes	1 (0.3)	1 (0.4)	0 (0)
No association	18 (6.2)	15 (6.2)	3 (6.7)
>1 association reported	23 (7.9)	18 (7.5)	5 (11.1)
Not assessed	189 (64.7)	157 (65.1)	29 (64.4)

[a]Categories with ">1" finding reflect the use of multiple measurements or analyses, leading to multiple results in different directions (eg, for "Change in user engagement level," engagement is shown to increase and decrease depending on the measurement used).

[b]SM: self-management.

Measurement Approach by User Engagement Level

Compared to studies characterized as having high user engagement, studies with low user engagement tended to measure (≥10% difference) engagement with user log-in data (19/27, 70% vs 36/63, 57%) and module completion (7/27, 26% vs 9/63, 14%). Compared to studies characterized as having low user engagement, studies with high user engagement tended to measure engagement with response to text notifications (14/63, 22% vs 1/27, 4%); interacting via chats, phone calls, or social media posts (13/63, 21% vs 1/27, 4%); wearing an electronic monitoring device (6/63, 10% vs 0); using a nonwearable electronic monitoring device (7/63, 11% vs 0); and qualitative interviews (12/63, 19% vs 2/27, 7%; Table 4).

XSL•FO
RenderX

Table 4. Measurement approach by user engagement level with intervention.

Measurement	Characterization of user engagement level with intervention				
	Low (n=27), n (%)[a]	Medium (n=6), n (%)[a]	High (n=63), n (%)[a]	>1 (n=99), n (%)[a]	Not characterized (n=97), n (%)[a]
User log-in data retrieved from app or website	19 (70)	5 (83)	36 (57)	57 (57)	43 (44)
Manual user data entry in app-or website-based self-monitoring diaries	8 (29)	2 (33)	15 (24)	29 (29)	23 (24)
Response to SMS text messages or push notifications	1 (4)	0 (0)	14 (22)	14 (14)	20 (21)
Number or proportion of intervention program modules completed within app or website	7 (26)	2 (33)	9 (14)	16 (16)	14 (14)
Interacting via chats, phone calls, or social media posts	1 (4)	0 (0)	13 (21)	10 (10)	9 (33)
Wearing an electronic monitoring device	0 (0)	0 (0)	6 (10)	9 (9)	11 (11)
Using a nonwearable electronic monitoring device	0 (0)	0 (0)	7 (11)	11 (11)	8 (8)
Submitting videos via app	2 (7)	0 (0)	0 (0)	2 (20)	1 (20)
Other objective measures	2 (7)	0 (0)	3 (5)	5 (5)	3 (3)
Qualitative interview	2 (7)	1 (17)	12 (19)	16 (16)	23 (24)
Participant-reported survey	1 (4)	2 (33)	4 (6)	10 (10)	12 (12)

[a]Percentages within categories do not add up to 100% because studies could fall into >1 category.

Technology Dosages and Minimum Engagement Research Benchmarks

Technology dosages denote when researchers provided participants with specific recommendations for mHealth intervention use (eg, log in to the app at least 3 times a week). Minimum engagement research benchmarks denoted when researchers set a minimum research cutoff for adequate participant engagement (eg, a participant who responded to ≥75% of SMS text messages during the study period was considered by the researchers to be adequately engaged with the mHealth intervention). A research benchmark could be set without giving a technology dosage and vice versa. Technology dosages were given less than half the time to participants by the researchers across all studies (119/292, 40.8%) and when examined by age group. Whether technology dosages were given or not, researchers characterized engagement level as "low" (14/27, 52% vs 13/27, 48%) and "high" (29/63, 46% vs 34/63, 54%) in relatively equal proportions (<10% difference; Tables 5 and 6).

Table 5. Study characteristics based on technology dosage and minimum engagement research benchmark.

	All studies (n=292), n (%)	Adult studies (n=241), n (%)	Pediatric studies (n=45), n (%)
Technology dosage given to participants			
Yes	119 (40.8)	96 (39.8)	22 (48.9)
No	173 (59.2)	145 (60.2)	23 (51.1)
Minimum engagement research benchmark set			
Yes	81 (27.7)	71 (29.4)	10 (22.2)
No	211 (72.3)	170 (70.5)	35 (77.8)

Table 6. Engagement levels based on technology dosage and minimum engagement research benchmark.

	Low (n=27), n (%)	Medium (n=6), n (%)	High (n=63), n (%)	>1 (n=99), n (%)	Not characterized (n=97), n (%)
Dosage given to participants					
Yes	14 (52)	2 (33)	29 (46)	43 (43)	31 (32)
No	13 (48)	4 (67)	34 (54)	56 (57)	66 (68)
Minimum engagement research benchmark set					
Yes	11 (41)	2 (33)	15 (24)	38 (38)	15 (15)
No	16 (59)	4 (67)	48 (76)	61 (62)	82 (85)

XSL•FO
RenderX

A minimum engagement research benchmark was set as the outcome criterion less than one-third of the time across all studies (81/292, 27.7%) and when examined by age group. When a minimum engagement research benchmark was set, researchers tended to characterize user engagement levels as "low" (11/27, 41%) rather than "high" (15/63, 24%). When no minimum engagement research benchmark was set, researchers tended to characterize user engagement as "high" (48/63, 76%) rather than "low" (16/27, 59%; Table 5).

Among studies that had a minimum engagement research benchmark (81/292, 27.7%), less than half (35/81, 43%) gave a technology dosage to participants. Among the studies that gave a technology dosage to participants (119/292, 40.8%), only 29.4% (35/119) also had a minimum engagement research benchmark.

Of the 292 studies, 35 (12%) gave both a technology dosage and set a minimum engagement research benchmark; these indices matched in 94% (33/35) of the studies. Of these 35 studies, 29 (83%) characterized engagement level, of which the majority (n=13, 45%) reported >1 level of engagement due to the use of multiple measurement methods or analyses.

When a minimum engagement research benchmark was set or a recommended technology dosage was given, engagement tended to be measured with user log-in data (48/81, 59% and 69/119, 58%, respectively), manual data entry (26/81, 32% and 32/119, 26.9%, respectively), or module completion (16/81, 20% and 21/119, 17.6%, respectively; Table 7).

Table 7. User engagement measurement approach when minimum engagement research benchmarks were set or technology dosages were given.

Measurements	User engagement level set by researchers	
	Minimum engagement research benchmark (n=81), n (%)[a]	Technology dosage given (n=119), n (%)[a]
User log-in data retrieved from app or website	48 (59.3)	69 (58.0)
Manual user data entry in app- or website-based self-monitoring diaries	26 (32.1)	32 (26.9)
Response to SMS text messages or push notifications	9 (11.1)	15 (12.6)
Number or proportion of intervention program modules completed within app or website	16 (19.8)	21 (17.6)
Interacting via chats, phone calls, or social media posts	12 (14.8)	14 (11.8)
Wearing an electronic monitoring device	10 (12.3)	18 (15.1)
Using a nonwearable electronic monitoring device	10 (12.3)	14 (11.8)
Submitting videos via app	2 (2.5)	5 (4.2)
Other objective measures	4 (4.9)	7 (5.9)
Qualitative interview	9 (11.1)	16 (13.4)
Participant-reported survey	1 (0.01)	12 (10.1)

[a]Percentages within categories do not add up to 100% because studies could fall into >1 category.

What Is the Association Between User Engagement With mHealth Interventions and Adherence or Self-Management Outcomes?

The association between engagement and treatment adherence or self-management outcomes was only assessed in a little more than one-third of the studies (103/292, 35.3%). Among these 103 studies, 60 (58.3%) tended to report and frame results and conclusions to suggest that higher engagement was associated with positive adherence or self-management outcomes. These results were similar between adult and pediatric studies (Table 3).

How Often Is User Engagement a Research End Point?

User engagement was a research end point in only 19.2% (56/292) of the reviewed studies, with similar results in adult (44/241, 18.3%) and pediatric (11/45, 24%) studies. Of these 56 studies, 30 (54%) used nonrandomized experimental designs, and 27 (48%) were feasibility studies. User engagement was typically defined as "engagement" (24/56, 43%), "adherence" (16/56, 29%), or "use" (16/56, 29%; Table S2 in Multimedia Appendix 3) and was most frequently measured with user log-in data (31/56, 55%) or manual data entry (22/56, 39%; Table S5 in Multimedia Appendix 3).

Exploratory Question: What Is the Association Between Providing Study Participants With Monetary Compensation and User Engagement Level?

Whether study participants were provided monetary compensation for their participation or not (or if compensation was not reported), user engagement measurement methods were used in similar proportions (<10% difference). Similarly, user engagement levels were observed in similar proportions between studies that provided monetary compensation and those that did not (or did not report compensation; Table S6 in Multimedia Appendix 3).

Of the 292 studies, 10 (3.4%) included as an intervention component financial incentives for using the technology or meeting mHealth intervention goals. Of these 10 studies, 7 (70%) reported user engagement levels. Among these 7 studies, in 1 (14%), the user engagement level was characterized as

XSL•FO
RenderX

"high"; in 1 (14%) as "medium"; and in 1 (14%), as "low"; and 4 (57%) reported >1 level due to multiple measurement or analytic approaches.

Discussion

Principal Findings

The principal findings relative to our specific research questions are provided in the following subsections.

How Is User Engagement Defined and Measured in Studies of mHealth Interventions to Promote Adherence to Prescribed Medical or Health Regimens or Self-Management Among People Living With a Health Condition?

Terms used to describe user engagement outcomes were wide ranging, but the most commonly used were "use" (102/292, 34.9%) and "engagement" (94/292, 32.2%). Across all studies reviewed, 11 distinct user engagement measurement approaches were identified, comprising both objective and subjective methods. The most common methods were user log-in data from the app or web portal (160/292, 54.8%), manually entering data in an app (77/292, 26.4%), qualitative interviews (54/292, 18.4%), and responding to SMS text messages or push notifications (49/292, 16.8%).

To What Degree Are Patients Engaging With These mHealth Interventions?

User engagement level was difficult to quantify because it was only characterized in two-thirds of the reviewed studies (195/292, 66.8%), of which a little more than half (99/195, 50.8%) reported >1 level of engagement due to the use of multiple measurement methods or analyses. Only one-third of the studies (99/292, 33.9%) evaluated change in engagement over time, which tended to decrease.

What Is the Association Between User Engagement With mHealth Interventions and Adherence or Self-Management Outcomes?

Only one-third of the studies (103/292, 35.3%) evaluated the association between engagement and treatment adherence or self-management outcomes. When evaluated, the study authors tended to report and frame results and conclusions to suggest that higher engagement was associated with positive adherence or self-management outcomes.

How Often Is User Engagement a Research End Point?

User engagement was rarely considered a research end point (56/292, 19.2%).

Exploratory Question: Are There Differences in User Engagement Measurement Approaches and Levels in Studies That Provide Monetary Compensation Compared to Those That Do Not?

No; whether study participants were provided monetary compensation for their participation or not (or if compensation was not reported), user engagement methods were used in similar proportions, and user engagement levels were observed in similar proportions.

Implications and Future Directions

Despite immense focus in both commercial and research sectors on using mHealth to support chronic illness treatment adherence and self-management [314-316], people do not remain engaged with mHealth interventions in the long term [6]. Our systematic review also found that user engagement tends to decline over time. Consistent with the "Little e, Big E framework" [16], mHealth intervention success hinges on the expectation that people will interact with the technology and thereby experience intended behavior changes that will lead to better overall health. Most importantly, our systematic review revealed critical limitations with how user engagement is defined and evaluated, which significantly impedes our ability to (1) communicate about this topic and (2) draw strong conclusions about how much user engagement is necessary to achieve desired behavior and health changes.

A principal finding of our review was that there is no agreed-upon definition of mHealth user engagement, which is a direct barrier to interdisciplinary communication about this topic. We found >60 terms used to define user engagement, even when limiting inclusion criteria to the behavioral evaluation of this concept [15]. While "use" (102/292, 34.9%) and "engagement" (94/292, 32.2%) were the most common terms, "adherence" (59/292, 20.2%) also appeared frequently. "Adherence" as a term for mHealth user engagement can be problematic because it is closely tied to treatment adherence [317], a key target of many of the mHealth interventions included in this systematic review. "Adherence" can apply to engaging with the mHealth technology as intended, such as logging in to an app [109] or responding to SMS text messages [243] rather than following a prescribed treatment regimen. Engaging with the technology can have a direct connection to treatment adherence (eg, using a Bluetooth-enabled blood pressure cuff reflects adherence to a key part of the hypertension management regimen as well as engaging with the mHealth technology [118,308]). However, such interchangeability and a lack of consistency in terminology will continue to be a barrier to communicating within and outside the field and add to existing challenges with defining and evaluating this domain. The extremely wide range of terms used to define user engagement likely reflects a major critique of mHealth research—that there is a lack of a science of engagement [7]. We encourage standardization in terminology used when the intention is to evaluate users' behavioral engagement with mHealth technology, rather than treatment regimen, with the most common terms found in this systematic review: "use" or "engagement." Integrating standard user engagement language in an internationally adopted clinical terminology system, such as the Systematized Nomenclature of Medicine–Clinical Terms, could help facilitate standardization efforts.

Another key finding was that the use of user log-in data was the default mHealth user engagement outcome, except for wearing a device when the intervention involved a wearable device component and responding to SMS text messages and push notifications when the intervention involved SMS text messaging and push notifications. Reliance on user log-in data over the other 10 major measurement approaches likely reflects that these data can be relatively straightforward to extract from

XSL•FO
RenderX

a web portal. A strength of the use of user log-in data is that it may indicate more effortful and deliberate interactions with the technology (the user likely needs sufficient motivation to open and navigate an app) compared with, for example, quickly responding to a SMS text message or wearing a pedometer on the wrist. User log-in data may be a more accurate reflection of engagement behavior than self-report, which is potentially subject to social desirability and recall bias. The tendency for engagement assessed using user log-in data to be characterized as "low" may reflect the higher user burden of logging in to, and interacting with an, app, in addition to the overwhelming selection of this method for measuring user engagement. Reliance on user log-in data may bias interpretations of user engagement to favor higher quantities of engagement at the expense of higher-quality engagement (eg, infrequent access of a particularly effective app feature that users perceive as interesting, helpful, or motivating may be higher-quality engagement compared with frequent log-ins to the app homepage). Given the ubiquitous selection of user log-in data and the heterogeneity of adherence and self-management outcomes in this systematic review, it is not possible to evaluate which user engagement metric is best. In general, investigators should select the user engagement metric reflecting essential interactions with the mHealth intervention's key technology needed to facilitate behavior and health changes.

We found that the exact degree of user engagement with the interventions reviewed in this study was difficult to estimate and that the association between user engagement and adherence or self-management outcomes were rarely evaluated. Both findings may relate to an important secondary finding that researchers rarely set a priori benchmarks for how much user engagement is considered scientifically adequate (no minimum engagement research benchmark set). Concerningly, when *no* minimum research benchmark was set, researchers tended to characterize user engagement as "high," suggesting overoptimism about study results in the absence of any hypothesized lower boundary for what constituted minimally acceptable mHealth use. Setting cut points for how much user engagement is hypothesized to be adequate would also help to guide how much and in what ways users should interact with the mHealth intervention components to experience meaningful improvements in target outcomes. Matching technology dosages to minimum engagement research benchmarks, when possible, could inform whether engagement recommendations were met and allow researchers to evaluate whether exceeding expectations is associated with even greater improvement in behavior or health outcomes. Of note, studies with technology dosages or minimum engagement research benchmarks seemed to demand more from participants, as evidenced by the more common use of user log-in data and manual data entry to evaluate user engagement. User engagement burden should be considered when setting mHealth engagement expectations.

Setting minimum engagement research benchmarks could also account for variability in user engagement–level characterization depending on the metric selected and help to define what constitutes effective and meaningful engagement. Specifically, engagement assessed using user log-in data tended to be characterized as "low." By contrast, responding to text notifications or interacting via chats, phone calls, or social media posts, which may be more socially rewarding and less effortful to complete, tended to be characterized as "high." A similarly higher pattern of user engagement was observed in studies evaluating the use of wearable (6/63, 10% vs 0) and nonwearable devices (7/63, 11% vs 0), which is often part of the users' regular self-management routines (eg, Bluetooth-enabled pill bottle or glucometer) or passively worn (eg, pedometer) and thus may be less burdensome to use. A higher SMS text message response rate and comparatively lower app log-in rate may both be minimally acceptable and potentially meaningful levels of engagement to effectively facilitate behavior or health change, particularly considering the level of effort and motivation involved in each type of interaction. Given the inconsistencies in how investigators characterize engagement levels, setting minimum engagement research benchmarks could inform whether user engagement is adequate across metrics and associated with key adherence and self-management outcomes. This approach could also improve precision in how engagement levels are characterized (ie, what level of engagement is considered "low" vs "high"?). Such efforts during the study planning phase could reduce the tendency to characterize user engagement at multiple levels due to the use of multiple metrics and analyses. Rethinking engagement levels as "adequate," "acceptable," "effective," or "meaningful" versus "high" or "low" may better inform how much people should realistically use mHealth technology and evaluate whether engagement level is associated with behavior and health changes.

In the mHealth trials reviewed, an important secondary finding was that participants were rarely told how much they should aim to engage with the technology (no technology dosages). This is in contrast to drug trials in which participants are instructed to take a specific dosage of medication on a strict schedule and trials of in-person behavior change interventions (traditional therapy) in which participants are given therapy regimens and expectations for participation (eg, attend 12 therapy sessions and complete assigned homework each week of treatment) [11]. Although an early-phase research goal may be to see how much people interact with the technology in the absence of recommendations, an important future goal is to understand how much engagement is needed to maximize behavior and health changes. To this end, recommending in what ways and how much users should aim to engage with the technology is necessary. Giving technology dosages would also allow for experimental evaluation of different levels of recommended technology use to see what recommendation helps users achieve the most positive behavior and health changes. Furthermore, improvements in health and wellness can take time, and mHealth users may not observe immediate benefits to using mHealth interventions, which could contribute to premature abandonment of the tool. Providing users with clearer guidance on how much and what type of mHealth intervention use is needed to begin effecting positive behavior and health changes could help encourage sustained use.

Another primary finding was that user engagement was rarely considered a research end point. This is a problem because it is generally expected that users will need to interact with the mHealth technology to experience clinical benefit. Thus, user

XSL•FO
RenderX

engagement behavior is a critical aspect of the mHealth intervention itself. Without designating user engagement as a research end point, mHealth intervention trials risk lacking the necessary empirical data and results to help end users understand how to optimally use the technology to see maximal clinical benefit. Lacking these data and results may also be a barrier to informing how these digital tools can be incorporated into regular clinical practice and policy [22]. Thus, in addition to carefully measuring the treatment adherence, self-management behavior, and associated health outcomes directly targeted by the mHealth intervention, equal care should be given to evaluating user engagement, such as how user engagement with the technology may change over time and potentially influence outcomes. Shifting user engagement from afterthought data to an actual research end point would likely improve the quality and rigor of mHealth research and help develop mHealth interventions that motivate users to interact with the technology [318], and, ideally, experience greater clinical benefit.

Our exploratory analysis of studies that provided monetary compensation for study participation versus those that did not (or did not report monetary compensation) showed that engagement-level characterizations were similar between the groups. It is possible that the monetary compensation was provided primarily for completing study procedures (eg, completing study surveys) rather than engaging with the mHealth intervention. In the studies of interventions comprising financial incentives for using the technology or meeting mHealth intervention goals (10/292, 3.4%), no clear pattern emerged for user engagement levels, although this likely reflects the small number of studies incorporating this mHealth component. User engagement evaluation methods were used in similar proportions between the groups, suggesting that monetary compensation was not associated with user engagement measurement selection. An important avenue for future mHealth engagement research is to improve the understanding of how monetary compensation is associated with user engagement levels, particularly when compensation is directed toward mHealth intervention use.

Limitations

This systematic review is among the most comprehensive reviews on this topic, but it has some limitations. First, due to heterogeneity in the study outcomes, we did not conduct a meta-analysis. Thus, this systematic review cannot conclude which user engagement outcome is associated with the most effective interventions. Second, although we found that study authors tended to frame their results and conclusions to suggest that higher user engagement was associated with positive intervention outcomes, this finding should be considered preliminary and hypothesis generating and does not prove that higher engagement leads to better adherence or self-management, particularly given that study authors may have a tendency to report favorable results and frame their findings in a positive light. However, this is an important area to examine in future research to better understand the characteristics of users with higher mHealth engagement as it relates to adherence and self-management outcomes (eg, highly engaged users may represent a patient group with already high adherence and self-management). Third, we did not exclude studies on the basis of their attrition rate; a future direction for improving the science of engagement is to consider how dropout and withdrawal may influence the characterization of user engagement levels. Fourth, a traditional risk-of-bias assessment [319,320] was not conducted due to our interest in user engagement rather than the primary study results targeting adherence and self-management. If user engagement becomes more commonly designated as a research end point, such evaluations of study quality specific to the evaluation of user engagement will be warranted. Fifth, we included a range of medical and mental health conditions but, due to small sample sizes, were unable to conduct subgroup analyses by diagnosis; as the literature on user engagement grows, it may be possible to explore such differences by diagnosis. Sixth, we focused on the behavioral evaluation of user engagement, although this domain likely involves more than behavior alone [15,321], and these other aspects of user engagement may be examined in a future systematic review. Seventh, our interest was in the measurement and evaluation of mHealth user engagement; thus, we did not consider the numerous individual-level factors that could be related to user engagement. Eighth, although study location was not part of the inclusion criteria, nearly half of the reviewed studies (135/292, 46.2%) were conducted in the United States, which may have influenced the types of intervention studies and approaches to evaluating user engagement. Ninth, we conducted our search in specific scientific databases and did not review gray literature or unpublished results; although our search resulted in 292 studies meeting our inclusion criteria, this systematic review may not be generalizable to all mHealth interventions. Finally, it is necessary to select a cutoff date when conducting a systematic review; yet, new studies are constantly being published. An updated review of the mHealth user engagement literature may be warranted in the future. As the science of mHealth user engagement improves, an important next step is to evaluate how factors related to diversity, equity, and inclusion relate to user engagement to promote wider mHealth use and access.

Recommendations for Future Researchers

Our systematic review on mHealth user engagement highlighted critical gaps in mHealth adherence and self-management literature as well as opportunities to improve research in this important area of digital health. mHealth researchers need to prioritize the evaluation of user engagement during the study planning phases. Strengthening mHealth user engagement methodological rigor would likely lead to higher-quality data and more impactful study results to guide practice and policy and ultimately encourage the uptake of interventions promoting mHealth adherence or self-management. Furthermore, improved user engagement evaluation may help researchers to identify effective strategies for supporting meaningful user engagement and sustained interest to help users fully experience the intended behavioral and health benefits of the mHealth intervention [16]. Such research is critical for building the evidence base needed to integrate mHealth interventions into regular clinical care and practice guidelines.

The following recommendations for studies of mHealth user engagement are offered:

XSL•FO
RenderX

- Use consistent terminology. Reflecting the most common trends in the literature reviewed, we recommend referring to the measurement of behavioral mHealth use, interaction, or engagement as “use” or “engagement.” Greater consistency in terminology could help improve rigor and consistency in the measurement itself and facilitate comparisons between studies as well as the evaluation of user engagement in meta-analyses.
- Select mHealth user engagement metrics that reflect interactions with the intervention’s key technology components hypothesized to facilitate behavior and health changes.
- Provide mHealth users with expectations for how much they should aim to interact with the technology (give technology dosages when possible).
- Set minimum engagement research benchmarks to scientifically denote the hypothesized user engagement level for a participant to be considered adequately, acceptably, or meaningfully engaged in the intervention.
- Characterize user engagement levels as adequate or acceptable with minimum engagement research benchmarks to help address issues with engagement being characterized as “high” or “low” depending on the metric selected and consider whether the engagement is adequate, meaningful, or effective across metrics (eg, a higher SMS text message response rate and comparatively lower app log-in rate may both be acceptable and meaningful levels of engagement).
- Designate mHealth user engagement as a research end point to help improve the quality and rigor of the mHealth research and the data collected from these studies. Focusing research efforts on user engagement could lead to mHealth engagement recommendations that could ultimately lead to greater clinical improvement.

Acknowledgments

This systematic review was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (K23DK128573: CE; R01DK125780: N Mathioudakis). The authors thank Valentina Vanos and Jacob White for their contributions to this systematic review.

Data Availability

Select review materials are available as multimedia appendices in this manuscript. The manual of procedures and the data sets generated and analyzed during this study are available from the corresponding author on reasonable request. The authors did not use generative AI to write any portions of this manuscript.

Authors' Contributions

CE was responsible for conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, writing the original draft, and reviewing and editing the manuscript. NV, XM, JW, and RR were responsible for data curation, investigation, writing the original draft, and reviewing and editing the manuscript. N Muthusamy was responsible for data curation, investigation, and reviewing and editing the manuscript. N Mathioudakis and KR were responsible for conceptualization, methodology, and reviewing and editing the manuscript.

Conflicts of Interest

CE serves as an mHealth Advisory Group Member for the Success with Therapies Research Consortium (Cystic Fibrosis Foundation).

Multimedia Appendix 1
Search strategy.
[PDF File (Adobe PDF File), 23 KB - jmir_v26i1e50508_app1.pdf]

Multimedia Appendix 2
Included studies and characteristics.
[XLSX File (Microsoft Excel File), 140 KB - jmir_v26i1e50508_app2.xlsx]

Multimedia Appendix 3
Supplementary tables.
[DOCX File , 38 KB - jmir_v26i1e50508_app3.docx]

Multimedia Appendix 4
PRISMA Checklist for Systematic Reviews.
[PDF File (Adobe PDF File), 119 KB - jmir_v26i1e50508_app4.pdf]

XSL•FO
RenderX

References

1. Teens, social media and technology 2018. Pew Research Center. 2018 May 31. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/#vast-majority-of-teens-have-access-to-a-home-computer-or-smartphone [accessed 2024-02-18]
2. Mobile fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-06-18]
3. Eaton C, Comer M, Pruette C, Psoter K, Riekert K. Text messaging adherence intervention for adolescents and young adults with chronic kidney disease: pilot randomized controlled trial and stakeholder interviews. J Med Internet Res 2020 Aug 14;22(8):e19861 [FREE Full text] [doi: 10.2196/19861] [Medline: 32795983]
4. Osborn CY, Hirsch A, Sears LE, Heyman M, Raymond J, Huddleston B, et al. One drop app with an activity tracker for adults with type 1 diabetes: randomized controlled trial. JMIR Mhealth Uhealth 2020 Sep 17;8(9):e16745 [FREE Full text] [doi: 10.2196/16745] [Medline: 32540842]
5. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020 Jan 13;3(1):4 [FREE Full text] [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
6. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 04;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]
7. Arigo D, Jake-Schoffman DE, Wolin K, Beckjord E, Hekler EB, Pagoto SL. The history and future of digital health in the field of behavioral medicine. J Behav Med 2019 Feb;42(1):67-83 [FREE Full text] [doi: 10.1007/s10865-018-9966-z] [Medline: 30825090]
8. Cao W, Milks MW, Liu X, Gregory ME, Addison D, Zhang P, et al. mHealth interventions for self-management of hypertension: framework and systematic review on engagement, interactivity, and tailoring. JMIR Mhealth Uhealth 2022 Mar 02;10(3):e29415 [FREE Full text] [doi: 10.2196/29415] [Medline: 35234655]
9. Yang Y, Boulton E, Todd C. Measurement of adherence to mHealth physical activity interventions and exploration of the factors that affect the adherence: scoping review and proposed framework. J Med Internet Res 2022 Jun 08;24(6):e30817 [FREE Full text] [doi: 10.2196/30817] [Medline: 35675111]
10. Schwarz A, Winkens LH, de Vet E, Ossendrijver D, Bouwsema K, Simons M. Design features associated with engagement in mobile health physical activity interventions among youth: systematic review of qualitative and quantitative studies. JMIR Mhealth Uhealth 2023 Mar 06;11:e40898 [FREE Full text] [doi: 10.2196/40898] [Medline: 36877551]
11. Lipschitz JM, Van Boxtel R, Torous J, Firth J, Lebovitz JG, Burdick KE, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022 Oct 14;24(10):e39204 [FREE Full text] [doi: 10.2196/39204] [Medline: 36240001]
12. Molloy A, Anderson PL. Engagement with mobile health interventions for depression: a systematic review. Internet Interv 2021 Dec;26:100454 [FREE Full text] [doi: 10.1016/j.invent.2021.100454] [Medline: 34621626]
13. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
14. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
15. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
16. Cole-Lewis H, Ezeanochie N, Turgiss J. Understanding health behavior technology engagement: pathway to measuring digital behavior change interventions. JMIR Form Res 2019 Oct 10;3(4):e14052 [FREE Full text] [doi: 10.2196/14052] [Medline: 31603427]
17. Nieuwlaat R, Wilczynski N, Navarro T, Hobson N, Jeffery R, Keepanasseril A, et al. Interventions for enhancing medication adherence. Cochrane Database Syst Rev 2014 Nov 20;2014(11):CD000011 [FREE Full text] [doi: 10.1002/14651858.CD000011.pub4] [Medline: 25412402]
18. Mellon L, Doyle F, Hickey A, Ward KD, de Freitas DG, McCormick P, et al. Interventions for increasing immunosuppressant medication adherence in solid organ transplant recipients. Cochrane Database Syst Rev 2022 Sep 12;9(9):CD012854 [FREE Full text] [doi: 10.1002/14651858.CD012854.pub2] [Medline: 36094829]
19. Al-Aqeel S, Gershuni O, Al-Sabhan J, Hiligsmann M. Strategies for improving adherence to antiepileptic drug treatment in people with epilepsy. Cochrane Database Syst Rev 2020 Oct 22;10(10):CD008312 [FREE Full text] [doi: 10.1002/14651858.CD008312.pub4] [Medline: 33089492]
20. Covidence systematic review software. Covidence. URL: https://www.covidence.org/ [accessed 2024-04-29]
21. IBM SPSS statistics for Windows. IBM Corp. URL: https://www.ibm.com/products/spss-statistics [accessed 2024-04-29]
22. Forbes G, Newton S, Cantalapiedra Calvete C, Birch J, Dodds J, Steed L, et al. MEMPHIS: a smartphone app using psychological approaches for women with chronic pelvic pain presenting to gynaecology clinics: a randomised feasibility trial. BMJ Open 2020 Mar 12;10(3):e030164 [FREE Full text] [doi: 10.1136/bmjopen-2019-030164] [Medline: 32165549]

XSL•FO
RenderX

23. Pagoto S, Waring ME. A call for a science of engagement: comment on Rus and Cameron. Ann Behav Med 2016 Oct;50(5):690-691. [doi: 10.1007/s12160-016-9839-z] [Medline: 27663577]
24. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
25. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
26. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020 Jul 01;3(7):e2010918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
27. Achtyes ED, Ben-Zeev D, Luo Z, Mayle H, Burke B, Rotondi AJ, et al. Off-hours use of a smartphone intervention to extend support for individuals with schizophrenia spectrum disorders recently discharged from a psychiatric hospital. Schizophr Res 2019 Apr;206:200-208. [doi: 10.1016/j.schres.2018.11.026] [Medline: 30551981]
28. Adams ZW, Sieverdes JC, Brunner-Jackson B, Mueller M, Chandler J, Diaz V, et al. Meditation smartphone application effects on prehypertensive adults' blood pressure: dose-response feasibility trial. Health Psychol 2018 Sep;37(9):850-860 [FREE Full text] [doi: 10.1037/hea0000584] [Medline: 30010353]
29. Agarwal G, Gaber J, Richardson J, Mangin D, Ploeg J, Valaitis R, et al. Pilot randomized controlled trial of a complex intervention for diabetes self-management supported by volunteers, technology, and interprofessional primary health care teams. Pilot Feasibility Stud 2019 Oct 27;5(1):118 [FREE Full text] [doi: 10.1186/s40814-019-0504-8] [Medline: 31673398]
30. Agarwal P, Mukerji G, Desveaux L, Ivers NM, Bhattacharyya O, Hensel JM, et al. Mobile app for improved self-management of type 2 diabetes: multicenter pragmatic randomized controlled trial. JMIR Mhealth Uhealth 2019 Jan 10;7(1):e10321 [FREE Full text] [doi: 10.2196/10321] [Medline: 30632972]
31. Agboola S, Jethwani K, Lopez L, Searl M, O'Keefe S, Kvedar J. Text to move: a randomized controlled trial of a text-messaging program to improve physical activity behaviors in patients with type 2 diabetes mellitus. J Med Internet Res 2016 Nov 18;18(11):e307 [FREE Full text] [doi: 10.2196/jmir.6439] [Medline: 27864165]
32. Aharonovich E, Stohl M, Cannizzaro D, Hasin D. HealthCall delivered via smartphone to reduce co-occurring drug and alcohol use in HIV-infected adults: a randomized pilot trial. J Subst Abuse Treat 2017 Dec;83:15-26 [FREE Full text] [doi: 10.1016/j.jsat.2017.09.013] [Medline: 29129192]
33. Amorim AB, Pappas E, Simic M, Ferreira ML, Jennings M, Tiedemann A, et al. Integrating mobile-health, health coaching, and physical activity to reduce the burden of chronic low back pain trial (IMPACT): a pilot randomised controlled trial. BMC Musculoskelet Disord 2019 Feb 11;20(1):71 [FREE Full text] [doi: 10.1186/s12891-019-2454-y] [Medline: 30744606]
34. Angellotti E, Wong JB, Pierce A, Hescott B, Pittas AG. Combining wireless technology and behavioral economics to engage patients (WiBEEP) with cardiometabolic disease: a pilot study. Pilot Feasibility Stud 2019 Jan 15;5(1):7 [FREE Full text] [doi: 10.1186/s40814-019-0395-8] [Medline: 30675374]
35. Anzaldo-Campos MC, Contreras S, Vargas-Ojeda A, Menchaca-Díaz R, Fortmann A, Philis-Tsimikas A. Dulce wireless Tijuana: a randomized control trial evaluating the impact of project dulce and short-term mobile technology on glycemic control in a family medicine clinic in Northern Mexico. Diabetes Technol Ther 2016 Apr;18(4):240-251 [FREE Full text] [doi: 10.1089/dia.2015.0283] [Medline: 26914371]
36. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016 Dec 20;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
37. Arnold C, Williams A, Thomas N. Engaging With a web-based psychosocial intervention for psychosis: qualitative study of user experiences. JMIR Ment Health 2020 Jun 19;7(6):e16730 [FREE Full text] [doi: 10.2196/16730] [Medline: 32558659]
38. Ashford MT, Olander EK, Rowe H, Fisher JR, Ayers S. Feasibility and acceptability of a web-based treatment with telephone support for postpartum women with anxiety: randomized controlled trial. JMIR Ment Health 2018 Apr 20;5(2):e19 [FREE Full text] [doi: 10.2196/mental.9106] [Medline: 29678804]
39. Athilingam P, Jenkins B, Johansson M, Labrador M. A mobile health intervention to improve self-care in patients with heart failure: pilot randomized control trial. JMIR Cardio 2017 Aug 11;1(2):e3 [FREE Full text] [doi: 10.2196/cardio.7848] [Medline: 31758759]
40. Babbage DR, van Kessel K, Drown J, Thomas S, Sezier A, Thomas P, et al. MS energize: field trial of an app for self-management of fatigue for people with multiple sclerosis. Internet Interv 2019 Dec;18:100291 [FREE Full text] [doi: 10.1016/j.invent.2019.100291] [Medline: 31890637]
41. Bailey JF, Agarwal V, Zheng P, Smuck M, Fredericson M, Kennedy DJ, et al. Digital care for chronic musculoskeletal pain: 10,000 participant longitudinal cohort study. J Med Internet Res 2020 May 11;22(5):e18250 [FREE Full text] [doi: 10.2196/18250] [Medline: 32208358]
42. Batch BC, Spratt SE, Blalock DV, Benditz C, Weiss A, Dolor RJ, et al. General behavioral engagement and changes in clinical and cognitive outcomes of patients with type 2 diabetes using the Time2Focus mobile app for diabetes education: pilot evaluation. J Med Internet Res 2021 Jan 20;23(1):e17537 [FREE Full text] [doi: 10.2196/17537] [Medline: 33470947]

XSL•FO
RenderX

43. Belanger HG, Toyinbo P, Barrett B, King E, Sayer NA. Concussion coach for postconcussive symptoms: a randomized, controlled trial of a smartphone application with Afghanistan and Iraq war Veterans. Clin Neuropsychol 2022 Nov 29;36(8):2093-2119. [doi: 10.1080/13854046.2021.1936188] [Medline: 34184976]
44. Ben-Zeev D, Brian RM, Jonathan G, Razzano L, Pashka N, Carpenter-Song E, et al. Mobile health (mHealth) versus clinic-based group intervention for people with serious mental illness: a randomized controlled trial. Psychiatr Serv 2018 Sep 01;69(9):978-985. [doi: 10.1176/appi.ps.201800063] [Medline: 29793397]
45. Bennell K, Nelligan RK, Schwartz S, Kasza J, Kimp A, Crofts SJ, et al. Behavior change text messages for home exercise adherence in knee osteoarthritis: randomized trial. J Med Internet Res 2020 Sep 28;22(9):e21749 [FREE Full text] [doi: 10.2196/21749] [Medline: 32985994]
46. Bentley CL, Otesile O, Bacigalupo R, Elliott J, Noble H, Hawley MS, et al. Feasibility study of portable technology for weight loss and HbA1c control in type 2 diabetes. BMC Med Inform Decis Mak 2016 Jul 15;16(1):92 [FREE Full text] [doi: 10.1186/s12911-016-0331-2] [Medline: 27418275]
47. Bentley CL, Powell L, Potter S, Parker J, Mountain GA, Bartlett YK, et al. The use of a smartphone app and an activity tracker to promote physical activity in the management of chronic obstructive pulmonary disease: randomized controlled feasibility study. JMIR Mhealth Uhealth 2020 Jun 03;8(6):e16203 [FREE Full text] [doi: 10.2196/16203] [Medline: 32490838]
48. Benzo RP, Kramer KM, Hoult JP, Anderson PM, Begue IM, Seifert SJ. Development and feasibility of a home pulmonary rehabilitation program with health coaching. Respir Care 2018 Feb 24;63(2):131-140 [FREE Full text] [doi: 10.4187/respcare.05690] [Medline: 29066590]
49. Berry D, Blonquist T, Nayak M, Grenon N, Momani T, McCleary N. Self-care support for patients with gastrointestinal cancer: iCancerHealth. Appl Clin Inform 2018 Oct 21;9(4):833-840 [FREE Full text] [doi: 10.1055/s-0038-1675810] [Medline: 30463094]
50. Bhatia A, Kara J, Janmohamed T, Prabhu A, Lebovic G, Katz J, et al. User engagement and clinical impact of the manage my pain app in patients with chronic pain: a real-world, multi-site trial. JMIR Mhealth Uhealth 2021 Mar 04;9(3):e26528 [FREE Full text] [doi: 10.2196/26528] [Medline: 33661130]
51. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR Mhealth Uhealth 2016 Jan 26;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]
52. Blonigen DM, Harris-Olenak B, Kuhn E, Timko C, Humphreys K, Smith JS, et al. Using peers to increase veterans' engagement in a smartphone application for unhealthy alcohol use: a pilot study of acceptability and utility. Psychol Addict Behav 2021 Nov;35(7):829-839 [FREE Full text] [doi: 10.1037/adb0000598] [Medline: 32597665]
53. Boal AL, Abroms LC, Simmens S, Graham AL, Carpenter KM. Combined quitline counseling and text messaging for smoking cessation: a quasi-experimental evaluation. Nicotine Tob Res 2016 May 31;18(5):1046-1053. [doi: 10.1093/ntr/ntv249] [Medline: 26521269]
54. Bobrow K, Farmer AJ, Springer D, Shanyinde M, Yu L, Brennan T, et al. Mobile phone text messages to support treatment adherence in adults with high blood pressure (SMS-Text Adherence Support [StAR]): a single-blind, randomized trial. Circulation 2016 Feb 09;133(6):592-600 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.017530] [Medline: 26769742]
55. Boer L, Bischoff E, van der Heijden M, Lucas P, Akkermans R, Vercoulen J, et al. A smart mobile health tool versus a paper action plan to support self-management of chronic obstructive pulmonary disease exacerbations: randomized controlled trial. JMIR Mhealth Uhealth 2019 Oct 09;7(10):e14408 [FREE Full text] [doi: 10.2196/14408] [Medline: 31599729]
56. Boettcher J, Magnusson K, Marklund A, Berglund E, Blomdahl R, Braun U, et al. Adding a smartphone app to internet-based self-help for social anxiety: a randomized controlled trial. Comput Human Behav 2018 Oct;87(6):98-108. [doi: 10.1016/j.chb.2018.04.052]
57. Boisseau CL, Schwartzman CM, Lawton J, Mancebo MC. App-guided exposure and response prevention for obsessive compulsive disorder: an open pilot trial. Cogn Behav Ther 2017 Nov 05;46(6):447-458. [doi: 10.1080/16506073.2017.1321683] [Medline: 28565937]
58. Bonar EE, Cunningham RM, Sweezea EC, Blow FC, Drislane LE, Walton MA. Piloting a brief intervention plus mobile boosters for drug use among emerging adults receiving emergency department care. Drug Alcohol Depend 2021 Apr 01;221:108625 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108625] [Medline: 33631541]
59. Bonato M, Turrini F, De Zan V, Meloni A, Plebani M, Brambilla E, et al. A mobile application for exercise intervention in people living with HIV. Med Sci Sports Exerc 2020 Feb;52(2):425-433. [doi: 10.1249/MSS.0000000000002125] [Medline: 31415448]
60. Boon M, Calvo-Lerma J, Claes I, Havermans T, Asseiceira I, Bulfamante A, et al. Use of a mobile application for self-management of pancreatic enzyme replacement therapy is associated with improved gastro-intestinal related quality of life in children with cystic fibrosis. J Cyst Fibros 2020 Jul;19(4):562-568 [FREE Full text] [doi: 10.1016/j.jcf.2020.04.001] [Medline: 32335023]

XSL•FO
RenderX

61. Børøsund E, Varsi C, Clark MM, Ehlers SL, Andrykowski MA, Sleveland HR, et al. Pilot testing an app-based stress management intervention for cancer survivors. Transl Behav Med 2020 Aug 07;10(3):770-780 [FREE Full text] [doi: 10.1093/tbm/ibz062] [Medline: 31330023]
62. Bradway M, Pfuhl G, Joakimsen R, Ribu L, Grøttland A, Årsand E. Analysing mHealth usage logs in RCTs: explaining participants' interactions with type 2 diabetes self-management tools. PLoS One 2018 Aug 30;13(8):e0203202 [FREE Full text] [doi: 10.1371/journal.pone.0203202] [Medline: 30161248]
63. Bricker JB, Copeland W, Mull KE, Zeng EY, Watson NL, Akioka KJ, et al. Drug Alcohol Depend 2017 Jan 01;170:37-42 [FREE Full text] [doi: 10.1016/j.drugalcdep.2016.10.029] [Medline: 27870987]
64. Bricker JB, Levin M, Lappalainen R, Mull K, Sullivan B, Santiago-Torres M. Mechanisms of smartphone apps for cigarette smoking cessation: results of a serial mediation model from the iCanQuit randomized trial. JMIR Mhealth Uhealth 2021 Nov 09;9(11):e32847 [FREE Full text] [doi: 10.2196/32847] [Medline: 34751662]
65. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. JAMA Intern Med 2020 Nov 01;180(11):1472-1480 [FREE Full text] [doi: 10.1001/jamainternmed.2020.4055] [Medline: 32955554]
66. Britto MT, Rohan JM, Dodds CM, Byczkowski TL. A randomized trial of user-controlled text messaging to improve asthma outcomes: a pilot study. Clin Pediatr (Phila) 2017 Dec;56(14):1336-1344. [doi: 10.1177/0009922816684857] [Medline: 28056541]
67. Browne S, Kechadi MT, O'Donnell S, Dow M, Tully L, Doyle G, et al. Mobile health apps in pediatric obesity treatment: process outcomes from a feasibility study of a multicomponent intervention. JMIR Mhealth Uhealth 2020 Jul 08;8(7):e16925 [FREE Full text] [doi: 10.2196/16925] [Medline: 32673267]
68. Bucci S, Barrowclough C, Ainsworth J, Machin M, Morris R, Berry K, et al. Actissist: proof-of-concept trial of a theory-driven digital intervention for psychosis. Schizophr Bull 2018 Aug 20;44(5):1070-1080 [FREE Full text] [doi: 10.1093/schbul/sby032] [Medline: 29566206]
69. Bughin F, Bui G, Ayoub B, Blervaque L, Saey D, Avignon A, et al. Impact of a mobile telerehabilitation solution on metabolic health outcomes and rehabilitation adherence in patients with obesity: randomized controlled trial. JMIR Mhealth Uhealth 2021 Dec 06;9(12):e28242 [FREE Full text] [doi: 10.2196/28242] [Medline: 34874887]
70. Buis LR, Roberson DN, Kadri R, Rockey NG, Plegue MA, Danak SU, et al. Understanding the feasibility, acceptability, and efficacy of a clinical pharmacist-led mobile approach (BPTrack) to hypertension management: mixed methods pilot study. J Med Internet Res 2020 Aug 11;22(8):e19882 [FREE Full text] [doi: 10.2196/19882] [Medline: 32780026]
71. Buscemi J, Buitrago D, Iacobelli F, Penedo F, Maciel C, Guitleman J, et al. Feasibility of a smartphone-based pilot intervention for Hispanic breast cancer survivors: a brief report. Transl Behav Med 2019 Jul 16;9(4):638-645 [FREE Full text] [doi: 10.1093/tbm/iby058] [Medline: 29986120]
72. Butryn ML, Martinelli MK, Crane NT, Godfrey K, Roberts SR, Zhang F, et al. Counselor surveillance of digital self-monitoring data: a pilot randomized controlled trial. Obesity (Silver Spring) 2020 Dec 23;28(12):2339-2346 [FREE Full text] [doi: 10.1002/oby.23015] [Medline: 33098278]
73. Canan CE, Waselewski ME, Waldman AL, Reynolds G, Flickinger TE, Cohn WF, et al. Long term impact of PositiveLinks: clinic-deployed mobile technology to improve engagement with HIV care. PLoS One 2020 Jan 6;15(1):e0226870 [FREE Full text] [doi: 10.1371/journal.pone.0226870] [Medline: 31905209]
74. Cartujano-Barrera F, Peña-Vargas CI, Arana-Chicas E, Pérez-Ramos JG, Mattei J, Hurtado-de-Mendoza A, et al. Decídetexto: feasibility and acceptability of a mobile smoking cessation intervention in Puerto Rico. Int J Environ Res Public Health 2021 Feb 03;18(4):1379 [FREE Full text] [doi: 10.3390/ijerph18041379] [Medline: 33546156]
75. Cartujano-Barrera F, Rodríguez-Bolaños R, Arana-Chicas E, Gallegos-Carrillo K, N Flores Y, Pérez-Rubio G, et al. Enhancing nicotine replacement therapy usage and adherence through a mobile intervention: secondary data analysis of a single-arm feasibility study in Mexico. Tob Induc Dis 2020 May 4;18(May):36 [FREE Full text] [doi: 10.18332/tid/120076] [Medline: 32395099]
76. Castensøe-Seidenfaden P, Husted GR, Jensen AK, Hommel E, Olsen B, Pedersen-Bjergaard U, et al. Testing a smartphone app (young with diabetes) to improve self-management of diabetes over 12 months: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jun 26;6(6):e141 [FREE Full text] [doi: 10.2196/mhealth.9487] [Medline: 29945861]
77. Cedars A, Blackmore C. Use of a disease-specific mobile health application in the care of adults with congenital heart disease. Proc (Bayl Univ Med Cent) 2019 Jul 10;32(3):336-339 [FREE Full text] [doi: 10.1080/08998280.2019.1613865] [Medline: 31384182]
78. Chavez K, Palfai TP. Feasibility of a mobile messaging-enhanced brief intervention for high risk heavy drinking MSM: a pre-pilot study. Alcohol Treat Q 2020 Aug 14;38(1):87-105 [FREE Full text] [doi: 10.1080/07347324.2019.1653240] [Medline: 32982032]
79. Chen J, Kaye L, Tuffli M, Barrett MA, Jones-Ford S, Shenouda T, et al. Passive monitoring of short-acting beta-agonist use via digital platform in patients with chronic obstructive pulmonary disease: quality improvement retrospective analysis. JMIR Form Res 2019 Oct 23;3(4):e13286 [FREE Full text] [doi: 10.2196/13286] [Medline: 31647471]

XSL•FO
RenderX

80. Cho H, Flynn G, Saylor M, Gradilla M, Schnall R. Use of the FITT framework to understand patients' experiences using a real-time medication monitoring pill bottle linked to a mobile-based HIV self-management app: a qualitative study. Int J Med Inform 2019 Nov;131:103949 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.08.009] [Medline: 31561192]
81. Cho H, Porras T, Baik D, Beauchemin M, Schnall R. Understanding the predisposing, enabling, and reinforcing factors influencing the use of a mobile-based HIV management app: a real-world usability evaluation. Int J Med Inform 2018 Sep;117:88-95 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.06.007] [Medline: 30032969]
82. Choi SA, Lim K, Baek H, Yoo S, Cho A, Kim H, et al. Impact of mobile health application on data collection and self-management of epilepsy. Epilepsy Behav 2021 Jun;119:107982. [doi: 10.1016/j.yebeh.2021.107982] [Medline: 33946011]
83. Clements MA, Staggs VS. A mobile app for synchronizing glucometer data: impact on adherence and glycemic control among youths with type 1 diabetes in routine care. J Diabetes Sci Technol 2017 May 09;11(3):461-467 [FREE Full text] [doi: 10.1177/1932296817691302] [Medline: 28745097]
84. Coorey G, Peiris D, Scaria A, Mulley J, Neubeck L, Hafiz N, et al. An internet-based intervention for cardiovascular disease management integrated with primary care electronic health records: mixed methods evaluation of implementation fidelity and user engagement. J Med Internet Res 2021 Apr 26;23(4):e25333 [FREE Full text] [doi: 10.2196/25333] [Medline: 33900204]
85. Côté J, Rouleau G, Ramirez-Garcia MP, Auger P, Thomas R, Leblanc J. Effectiveness of a web-based intervention to support medication adherence among people living with HIV: web-based randomized controlled trial. JMIR Public Health Surveill 2020 Apr 20;6(2):e17733 [FREE Full text] [doi: 10.2196/17733] [Medline: 32310145]
86. Coughlin LN, Nahum-Shani I, Philyaw-Kotov ML, Bonar EE, Rabbi M, Klasnja P, et al. Developing an adaptive mobile intervention to address risky substance use among adolescents and emerging adults: usability study. JMIR Mhealth Uhealth 2021 Jan 15;9(1):e24424 [FREE Full text] [doi: 10.2196/24424] [Medline: 33448931]
87. Crafoord M, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
88. Crane D, Garnett C, Michie S, West R, Brown J. A smartphone app to reduce excessive alcohol consumption: identifying the effectiveness of intervention components in a factorial randomised control trial. Sci Rep 2018 Mar 12;8(1):4384 [FREE Full text] [doi: 10.1038/s41598-018-22420-8] [Medline: 29531280]
89. Creary S, Chisolm D, Stanek J, Hankins J, O'Brien SH. A multidimensional electronic hydroxyurea adherence intervention for children with sickle cell disease: single-arm before-after study. JMIR Mhealth Uhealth 2019 Aug 08;7(8):e13452 [FREE Full text] [doi: 10.2196/13452] [Medline: 31397291]
90. Dahne J, Collado A, Lejuez CW, Risco CM, Diaz VA, Coles L, et al. Pilot randomized controlled trial of a Spanish-language behavioral activation mobile app (¡Aptívate!) for the treatment of depressive symptoms among united states Latinx adults with limited English proficiency. J Affect Disord 2019 May 01;250:210-217 [FREE Full text] [doi: 10.1016/j.jad.2019.03.009] [Medline: 30870770]
91. Dahne J, Lejuez C, Diaz VA, Player MS, Kustanowitz J, Felton JW, et al. Pilot randomized trial of a self-help behavioral activation mobile app for utilization in primary care. Behav Ther 2019 Jul;50(4):817-827 [FREE Full text] [doi: 10.1016/j.beth.2018.12.003] [Medline: 31208690]
92. Davies EH, Fieggen K, Wilmshurst J, Anyanwu O, Burman RJ, Komarzynski S. Demonstrating the feasibility of digital health to support pediatric patients in South Africa. Epilepsia Open 2021 Dec 02;6(4):653-662 [FREE Full text] [doi: 10.1002/epi4.12527] [Medline: 34310860]
93. Davis SR, Peters D, Calvo RA, Sawyer SM, Foster JM, Smith LD. A consumer designed smartphone app for young people with asthma: pilot of engagement and acceptability. J Asthma 2021 Feb 06;58(2):253-261. [doi: 10.1080/02770903.2019.1680997] [Medline: 31691618]
94. Davoudi A, Lee NS, Chivers C, Delaney T, Asch EL, Reitz C, et al. Patient interaction phenotypes with an automated remote hypertension monitoring program and their association with blood pressure control: observational study. J Med Internet Res 2020 Dec 03;22(12):e22493 [FREE Full text] [doi: 10.2196/22493] [Medline: 33270032]
95. Dawson J, Campbell KL, Craig JC, Tong A, Teixeira-Pinto A, Brown MA, et al. A text messaging intervention for dietary behaviors for people receiving maintenance hemodialysis: a feasibility study of KIDNEYTEXT. Am J Kidney Dis 2021 Jul;78(1):85-95.e1 [FREE Full text] [doi: 10.1053/j.ajkd.2020.11.015] [Medline: 33421456]
96. de Jong M, van der Meulen-de Jong A, Romberg-Camps M, Degens J, Becx M, Markus T, et al. Development and feasibility study of a telemedicine tool for all patients with IBD: MyIBDcoach. Inflamm Bowel Dis 2017 Apr;23(4):485-493. [doi: 10.1097/MIB.0000000000001034] [Medline: 28267047]
97. de la Vega R, Ritterband L, Palermo TM. Assessing digital health implementation for a pediatric chronic pain intervention: comparing the RE-AIM and BIT frameworks against real-world trial data and recommendations for future studies. J Med Internet Res 2020 Sep 01;22(9):e19898 [FREE Full text] [doi: 10.2196/19898] [Medline: 32870158]
98. DeFulio A, Devoto A, Traxler H, Cosottile D, Fingerhood M, Nuzzo P, et al. Smartphone-based incentives for promoting adherence to antiretroviral therapy: a randomized controlled trial. Prev Med Rep 2021 Mar;21:101318 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101318] [Medline: 33511028]

XSL•FO
RenderX

99. Desteghe L, Kluts K, Vijgen J, Koopman P, Dilling-Boer D, Schurmans J, et al. The health buddies app as a novel tool to improve adherence and knowledge in atrial fibrillation patients: a pilot study. JMIR Mhealth Uhealth 2017 Jul 19;5(7):e98 [FREE Full text] [doi: 10.2196/mhealth.7420] [Medline: 28724512]
100. Desveaux L, Shaw J, Saragosa M, Soobiah C, Marani H, Hensel J, et al. A mobile app to improve self-management of individuals with type 2 diabetes: qualitative realist evaluation. J Med Internet Res 2018 Mar 16;20(3):e81 [FREE Full text] [doi: 10.2196/jmir.8712] [Medline: 29549070]
101. Dietrich JE, Yee DL, Santos XM, Bercaw-Pratt JL, Kurkowski J, Soni H, et al. Assessment of an electronic intervention in young women with heavy menstrual bleeding. J Pediatr Adolesc Gynecol 2017 Apr;30(2):243-246. [doi: 10.1016/j.jpag.2016.10.006] [Medline: 27751907]
102. Dillingham R, Ingersoll K, Flickinger TE, Waldman AL, Grabowski M, Laurence C, et al. PositiveLinks: a mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up. AIDS Patient Care STDS 2018 Jun;32(6):241-250 [FREE Full text] [doi: 10.1089/apc.2017.0303] [Medline: 29851504]
103. Dombrowski SU, McDonald M, van der Pol M, Grindle M, Avenell A, Carroll P, et al. Game of Stones: feasibility randomised controlled trial of how to engage men with obesity in text message and incentive interventions for weight loss. BMJ Open 2020 Feb 25;10(2):e032653 [FREE Full text] [doi: 10.1136/bmjopen-2019-032653] [Medline: 32102807]
104. Downs DS, Savage JS, Rivera DE, Pauley AM, Leonard KS, Hohman EE, et al. Adaptive, behavioral intervention impact on weight gain, physical activity, energy intake, and motivational determinants: results of a feasibility trial in pregnant women with overweight/obesity. J Behav Med 2021 Oct;44(5):605-621 [FREE Full text] [doi: 10.1007/s10865-021-00227-9] [Medline: 33954853]
105. Dunn CG, Turner-McGrievy GM, Wilcox S, Hutto B. Dietary self-monitoring through calorie tracking but not through a digital photography app is associated with significant weight loss: the 2SMART pilot study-a 6-month randomized trial. J Acad Nutr Diet 2019 Sep;119(9):1525-1532 [FREE Full text] [doi: 10.1016/j.jand.2019.03.013] [Medline: 31155474]
106. Dyal N, McAssey K, Agarwal G. Evaluation of a computerized self-management tool for children with type 1 diabetes: a pilot project. Can J Diabetes 2017 Apr;41(2):124-128. [doi: 10.1016/j.jcjd.2016.09.021] [Medline: 28007351]
107. Ellis TD, Cavanaugh JT, DeAngelis T, Hendron K, Thomas CA, Saint-Hilaire M, et al. Comparative effectiveness of mHealth-supported exercise compared with exercise alone for people with Parkinson disease: randomized controlled pilot study. Phys Ther 2019 Feb 01;99(2):203-216. [doi: 10.1093/ptj/pzy131] [Medline: 30715489]
108. Escobar-Viera C, Zhou Z, Morano JP, Lucero R, Lieb S, McIntosh S, et al. The Florida mobile health adherence project for people living with HIV (FL-mAPP): longitudinal assessment of feasibility, acceptability, and clinical outcomes. JMIR Mhealth Uhealth 2020 Jan 08;8(1):e14557 [FREE Full text] [doi: 10.2196/14557] [Medline: 31913127]
109. Fedele DA, Thomas JG, McConville A, McQuaid EL, Voorhees S, Janicke DM, et al. Using mobile health to improve asthma self-management in early adolescence: a pilot randomized controlled trial. J Adolesc Health 2021 Dec;69(6):1032-1040. [doi: 10.1016/j.jadohealth.2021.06.011] [Medline: 34274211]
110. Ferrante JM, Devine KA, Bator A, Rodgers A, Ohman-Strickland PA, Bandera EV, et al. Feasibility and potential efficacy of commercial mHealth/eHealth tools for weight loss in African American breast cancer survivors: pilot randomized controlled trial. Transl Behav Med 2020 Oct 08;10(4):938-948 [FREE Full text] [doi: 10.1093/tbm/iby124] [Medline: 30535101]
111. Fogarty AS, Proudfoot J, Whittle EL, Clarke J, Player MJ, Christensen H, et al. Preliminary evaluation of a brief web and mobile phone intervention for men with depression: men's positive coping strategies and associated depression, resilience, and work and social functioning. JMIR Ment Health 2017 Aug 10;4(3):e33 [FREE Full text] [doi: 10.2196/mental.7769] [Medline: 28798009]
112. Forman EM, Goldstein SP, Crochiere RJ, Butryn ML, Juarascio AS, Zhang F, et al. Randomized controlled trial of OnTrack, a just-in-time adaptive intervention designed to enhance weight loss. Transl Behav Med 2019 Nov 25;9(6):989-1001. [doi: 10.1093/tbm/ibz137] [Medline: 31602471]
113. Garofalo R, Kuhns LM, Hotton A, Johnson A, Muldoon A, Rice D. A randomized controlled trial of personalized text message reminders to promote medication adherence among HIV-positive adolescents and young adults. AIDS Behav 2016 May;20(5):1049-1059 [FREE Full text] [doi: 10.1007/s10461-015-1192-x] [Medline: 26362167]
114. Gazit T, Gutman M, Beatty AL. Assessment of hypertension control among adults participating in a mobile technology blood pressure self-management program. JAMA Netw Open 2021 Oct 01;4(10):e2127008 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.27008] [Medline: 34652447]
115. Gentili C, Zetterqvist V, Rickardsson J, Holmström L, Simons LE, Wicksell RK. ACTsmart: guided smartphone-delivered acceptance and commitment therapy for chronic pain-a pilot trial. Pain Med 2021 Feb 23;22(2):315-328 [FREE Full text] [doi: 10.1093/pm/pnaa360] [Medline: 33200214]
116. Gimbel RW, Rennert LM, Crawford P, Little JR, Truong K, Williams JE, et al. Enhancing patient activation and self-management activities in patients with type 2 diabetes using the US Department of Defense mobile health care environment: feasibility study. J Med Internet Res 2020 May 26;22(5):e17968 [FREE Full text] [doi: 10.2196/17968] [Medline: 32329438]

XSL•FO
RenderX

117. Gire N, Caton N, McKeown M, Mohmed N, Duxbury J, Kelly J, et al. Care co-ordinator in my pocket': a feasibility study of mobile assessment and therapy for psychosis (TechCare). BMJ Open 2021 Nov 16;11(11):e046755 [FREE Full text] [doi: 10.1136/bmjopen-2020-046755] [Medline: 34785541]
118. Godersky ME, Klein JW, Merrill JO, Blalock KL, Saxon AJ, Samet JH, et al. Acceptability and feasibility of a mobile health application for video directly observed therapy of buprenorphine for opioid use disorders in an office-based setting. J Addict Med 2020;14(4):319-325 [FREE Full text] [doi: 10.1097/ADM.0000000000000608] [Medline: 31972762]
119. Godino JG, Golaszewski NM, Norman GJ, Rock CL, Griswold WG, Arredondo E, et al. Text messaging and brief phone calls for weight loss in overweight and obese English- and Spanish-speaking adults: a 1-year, parallel-group, randomized controlled trial. PLoS Med 2019 Sep;16(9):e1002917 [FREE Full text] [doi: 10.1371/journal.pmed.1002917] [Medline: 31553725]
120. Gordon JS, Armin J, D Hingle M, Giacobbi P, Cunningham JK, Johnson T, et al. Development and evaluation of the See Me Smoke-Free multi-behavioral mHealth app for women smokers. Transl Behav Med 2017 Jun;7(2):172-184 [FREE Full text] [doi: 10.1007/s13142-017-0463-7] [Medline: 28155107]
121. Goyal S, Nunn CA, Rotondi M, Couperthwaite AB, Reiser S, Simone A, et al. A mobile app for the self-management of type 1 diabetes among adolescents: a randomized controlled trial. JMIR Mhealth Uhealth 2017 Jun 19;5(6):e82 [FREE Full text] [doi: 10.2196/mhealth.7336] [Medline: 28630037]
122. Greer JA, Jacobs JM, Pensak N, Nisotel LE, Fishbein JN, MacDonald JJ, et al. Randomized trial of a smartphone mobile app to improve symptoms and adherence to oral therapy for cancer. J Natl Compr Canc Netw 2020 Feb;18(2):133-141. [doi: 10.6004/jnccn.2019.7354] [Medline: 32023526]
123. Guarino H, Acosta M, Marsch LA, Xie H, Aponte-Melendez Y. A mixed-methods evaluation of the feasibility, acceptability, and preliminary efficacy of a mobile intervention for methadone maintenance clients. Psychol Addict Behav 2016 Feb;30(1):1-11 [FREE Full text] [doi: 10.1037/adb0000128] [Medline: 26618796]
124. Guo X, Gu X, Jiang J, Li H, Duan R, Zhang Y, et al. A hospital-community-family-based telehealth program for patients with chronic heart failure: single-arm, prospective feasibility study. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13229 [FREE Full text] [doi: 10.2196/13229] [Medline: 31833835]
125. Guthrie NL, Berman MA, Edwards KL, Appelbaum KJ, Dey S, Carpenter J, et al. Achieving rapid blood pressure control with digital therapeutics: retrospective cohort and machine learning study. JMIR Cardio 2019 Mar 12;3(1):e13030 [FREE Full text] [doi: 10.2196/13030] [Medline: 31758792]
126. Haas K, Martin A, Park KT. Text message intervention (TEACH) improves quality of life and patient activation in celiac disease: a randomized clinical trial. J Pediatr 2017 Jun;185:62-7.e2 [FREE Full text] [doi: 10.1016/j.jpeds.2017.02.062] [Medline: 28343658]
127. Hammond AS, Sweeney MM, Chikosi TU, Stitzer ML. Digital delivery of a contingency management intervention for substance use disorder: a feasibility study with DynamiCare Health. J Subst Abuse Treat 2021 Jul;126:108425 [FREE Full text] [doi: 10.1016/j.jsat.2021.108425] [Medline: 34116816]
128. Han A, Min SI, Ahn S, Min S, Hong H, Han N, et al. Mobile medication manager application to improve adherence with immunosuppressive therapy in renal transplant recipients: a randomized controlled trial. PLoS One 2019;14(11):e0224595 [FREE Full text] [doi: 10.1371/journal.pone.0224595] [Medline: 31689320]
129. Hardcastle SJ, Jiménez-Castuera R, Maxwell-Smith C, Bulsara MK, Hince D. Fitbit wear-time and patterns of activity in cancer survivors throughout a physical activity intervention and follow-up: exploratory analysis from a randomised controlled trial. PLoS One 2020 Oct 19;15(10):e0240967 [FREE Full text] [doi: 10.1371/journal.pone.0240967] [Medline: 33075100]
130. Harzand A, Witbrodt B, Davis-Watts ML, Alrohaibani A, Goese D, Wenger NK, et al. Feasibility of a smartphone-enabled cardiac rehabilitation program in male veterans with previous clinical evidence of coronary heart disease. Am J Cardiol 2018 Nov 01;122(9):1471-1476 [FREE Full text] [doi: 10.1016/j.amjcard.2018.07.028] [Medline: 30217377]
131. Heale LD, Dover S, Goh YI, Maksymiuk VA, Wells GD, Feldman BM. A wearable activity tracker intervention for promoting physical activity in adolescents with juvenile idiopathic arthritis: a pilot study. Pediatr Rheumatol Online J 2018 Oct 22;16(1):66 [FREE Full text] [doi: 10.1186/s12969-018-0282-5] [Medline: 30348203]
132. Hébert ET, Ra CK, Alexander AC, Helt A, Moisiuc R, Kendzor DE, et al. A mobile just-in-time adaptive intervention for smoking cessation: pilot randomized controlled trial. J Med Internet Res 2020 Mar 09;22(3):e16907 [FREE Full text] [doi: 10.2196/16907] [Medline: 32149716]
133. Heminger CL, Boal AL, Zumer M, Abroms LC. Text2Quit: an analysis of participant engagement in the mobile smoking cessation program. Am J Drug Alcohol Abuse 2016 Jul 27;42(4):450-458. [doi: 10.3109/00952990.2016.1149591] [Medline: 27120396]
134. Herbec A, Brown J, Shahab L, West R, Raupach T. Pragmatic randomised trial of a smartphone app (NRT2Quit) to improve effectiveness of nicotine replacement therapy in a quit attempt by improving medication adherence: results of a prematurely terminated study. Trials 2019 Sep 02;20(1):547 [FREE Full text] [doi: 10.1186/s13063-019-3645-4] [Medline: 31477166]
135. Herbert LJ, Collier S, Stern A, Monaghan M, Streisand R. A pilot test of the self-management and research technology project: a text message-based diabetes self-management program for adolescents. J Child Health Care 2016 Dec 26;20(4):456-463. [doi: 10.1177/1367493515603829] [Medline: 26396022]

XSL•FO
RenderX

136. Hightow-Weidman L, Muessig KE, Egger JR, Vecchio A, Platt A. Epic Allies: a gamified mobile app to improve engagement in HIV care and antiretroviral adherence among young men who have sex with men. AIDS Behav 2021 Aug;25(8):2599-2617 [FREE Full text] [doi: 10.1007/s10461-021-03222-y] [Medline: 33740213]
137. Hightow-Weidman L, Muessig K, Knudtson K, Srivatsa M, Lawrence E, LeGrand S, et al. A gamified smartphone app to support engagement in care and medication adherence for HIV-positive young men who have sex with men (AllyQuest): development and pilot study. JMIR Public Health Surveill 2018 Apr 30;4(2):e34 [FREE Full text] [doi: 10.2196/publichealth.8923] [Medline: 29712626]
138. Hilliard ME, Cao VT, Eshtehardi SS, Minard CG, Saber R, Thompson D, et al. Type 1 doing well: pilot feasibility and acceptability study of a strengths-based mHealth app for parents of adolescents with type 1 diabetes. Diabetes Technol Ther 2020 Nov;22(11):835-845 [FREE Full text] [doi: 10.1089/dia.2020.0048] [Medline: 32379496]
139. Himelhoch S, Kreyenbuhl J, Palmer-Bacon J, Chu M, Brown C, Potts W. Pilot feasibility study of Heart2HAART: a smartphone application to assist with adherence among substance users living with HIV. AIDS Care 2017 Jul 03;29(7):898-904. [doi: 10.1080/09540121.2016.1259454] [Medline: 28043176]
140. Hoffman V, Söderström L, Samuelsson E. Self-management of stress urinary incontinence via a mobile app: two-year follow-up of a randomized controlled trial. Acta Obstet Gynecol Scand 2017 Oct;96(10):1180-1187 [FREE Full text] [doi: 10.1111/aogs.13192] [Medline: 28718223]
141. Hommel KA, Carmody J, Hershey AD, Holbein C, Kabbouche-Samaha M, Peugh J, et al. Digital therapeutic self-management intervention in adolescents with migraine: feasibility and preliminary efficacy of "migraine manager". Headache 2020 Jun 22;60(6):1103-1110 [FREE Full text] [doi: 10.1111/head.13805] [Medline: 32320052]
142. Hood A, Nwankwo C, Walton A, McTate E, Joffe N, Quinn CT, et al. Mobile health use predicts self-efficacy and self-management in adolescents with sickle cell disease. Transl Behav Med 2021 Oct 23;11(10):1823-1831 [FREE Full text] [doi: 10.1093/tbm/ibab041] [Medline: 33949674]
143. Horner GN, Agboola S, Jethwani K, Tan-McGrory A, Lopez L. Designing patient-centered text messaging interventions for increasing physical activity among participants with type 2 diabetes: qualitative results from the text to move intervention. JMIR Mhealth Uhealth 2017 Apr 24;5(4):e54 [FREE Full text] [doi: 10.2196/mhealth.6666] [Medline: 28438728]
144. Horvath KJ, Lammert S, MacLehose RF, Danh T, Baker JV, Carrico AW. A pilot study of a mobile app to support HIV antiretroviral therapy adherence among men who have sex with men who use stimulants. AIDS Behav 2019 Nov 15;23(11):3184-3198 [FREE Full text] [doi: 10.1007/s10461-019-02597-3] [Medline: 31309348]
145. Huang Z, Tan E, Lum E, Sloot P, Boehm BO, Car J. A smartphone app to improve medication adherence in patients with type 2 diabetes in Asia: feasibility randomized controlled trial. JMIR Mhealth Uhealth 2019 Sep 12;7(9):e14914 [FREE Full text] [doi: 10.2196/14914] [Medline: 31516127]
146. Huberty JL, Green J, Puzia ME, Larkey L, Laird B, Vranceanu A, et al. Testing a mindfulness meditation mobile app for the treatment of sleep-related symptoms in adults with sleep disturbance: a randomized controlled trial. PLoS One 2021 Jan 7;16(1):e0244717 [FREE Full text] [doi: 10.1371/journal.pone.0244717] [Medline: 33411779]
147. Huh U, Tak YJ, Song S, Chung SW, Sung SM, Lee CW, et al. Feedback on physical activity through a wearable device connected to a mobile phone app in patients with metabolic syndrome: pilot study. JMIR Mhealth Uhealth 2019 Jun 18;7(6):e13381 [FREE Full text] [doi: 10.2196/13381] [Medline: 31215513]
148. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152 [FREE Full text] [doi: 10.2196/26152] [Medline: 33872182]
149. Huo X, Krumholz HM, Bai X, Spatz ES, Ding Q, Horak P, et al. Effects of mobile text messaging on glycemic control in patients with coronary heart disease and diabetes mellitus: a randomized clinical trial. Circ Cardiovasc Qual Outcomes 2019 Sep;12(9):e005805. [doi: 10.1161/CIRCOUTCOMES.119.005805] [Medline: 31474119]
150. Hutchesson MJ, Tan CY, Morgan P, Callister R, Collins C. Enhancement of self-monitoring in a web-based weight loss program by extra individualized feedback and reminders: randomized trial. J Med Internet Res 2016 Apr 12;18(4):e82 [FREE Full text] [doi: 10.2196/jmir.4100] [Medline: 27072817]
151. Irizarry T, Allen M, Suffoletto BP, Einhorn J, Burke LE, Kamarck TW, et al. Development and preliminary feasibility of an automated hypertension self-management system. Am J Med 2018 Sep;131(9):1125.e1-1125.e8 [FREE Full text] [doi: 10.1016/j.amjmed.2018.04.038] [Medline: 29806998]
152. Ivanova E, Lindner P, Ly KH, Dahlin M, Vernmark K, Andersson G, et al. Guided and unguided acceptance and commitment therapy for social anxiety disorder and/or panic disorder provided via the internet and a smartphone application: a randomized controlled trial. J Anxiety Disord 2016 Dec;44:27-35. [doi: 10.1016/j.janxdis.2016.09.012] [Medline: 27721123]
153. Jácome C, Almeida R, Pereira AM, Amaral R, Mendes S, Alves-Correia M, et al. Feasibility and acceptability of an asthma app to monitor medication adherence: mixed methods study. JMIR Mhealth Uhealth 2021 May 25;9(5):e26442 [FREE Full text] [doi: 10.2196/26442] [Medline: 34032576]
154. Jaimini U, Thirunarayan K, Kalra M, Venkataraman R, Kadariya D, Sheth A. "How is my child's asthma?" Digital phenotype and actionable insights for pediatric asthma. JMIR Pediatr Parent 2018 Dec 14;1(2):e11988 [FREE Full text] [doi: 10.2196/11988] [Medline: 31008446]

XSL•FO
RenderX

155. Jamison RN, Mei A, Ross EL. Longitudinal trial of a smartphone pain application for chronic pain patients: Predictors of compliance and satisfaction. J Telemed Telecare 2018 Feb;24(2):93-100. [doi: 10.1177/1357633X16679049] [Medline: 27831496]
156. Johansson M, Berman AH, Sinadinovic K, Lindner P, Hermansson U, Andréasson S. Effects of internet-based cognitive behavioral therapy for harmful alcohol use and alcohol dependence as self-help or with therapist guidance: three-armed randomized trial. J Med Internet Res 2021 Nov 24;23(11):e29666 [FREE Full text] [doi: 10.2196/29666] [Medline: 34821563]
157. Jonassaint CR, Kang C, Prussien KV, Yarboi J, Sanger MS, Wilson JD, et al. Feasibility of implementing mobile technology-delivered mental health treatment in routine adult sickle cell disease care. Transl Behav Med 2020 Feb 03;10(1):58-67 [FREE Full text] [doi: 10.1093/tbm/iby107] [Medline: 30508141]
158. Jonathan G, Carpenter-Song EA, Brian RM, Ben-Zeev D. Life with FOCUS: a qualitative evaluation of the impact of a smartphone intervention on people with serious mental illness. Psychiatr Rehabil J 2019 Jun;42(2):182-189 [FREE Full text] [doi: 10.1037/prj0000337] [Medline: 30589278]
159. Juarascio AS, Hunt RA, Lantz Lesser E, Engel SG, Pisetsky EM, Peterson CB, et al. Enhancing integrative cognitive-affective therapy with ecological momentary interventions: a pilot trial. Eur Eat Disord Rev 2021 Jan 26;29(1):152-158. [doi: 10.1002/erv.2800] [Medline: 33104279]
160. Kählke F, Berger T, Schulz A, Baumeister H, Berking M, Auerbach RP, et al. Efficacy of an unguided internet-based self-help intervention for social anxiety disorder in university students: a randomized controlled trial. Int J Methods Psychiatr Res 2019 Jun 27;28(2):e1766 [FREE Full text] [doi: 10.1002/mpr.1766] [Medline: 30687986]
161. Kanera IM, Bolman CA, Willems RA, Mesters I, Lechner L. Lifestyle-related effects of the web-based Kanker Nazorg Wijzer (Cancer Aftercare Guide) intervention for cancer survivors: a randomized controlled trial. J Cancer Surviv 2016 Oct;10(5):883-897 [FREE Full text] [doi: 10.1007/s11764-016-0535-6] [Medline: 26984534]
162. Kanera IM, Willems RA, Bolman CA, Mesters I, Zambon V, Gijsen BC, et al. Use and appreciation of a tailored self-management eHealth intervention for early cancer survivors: process evaluation of a randomized controlled trial. J Med Internet Res 2016 Aug 23;18(8):e229 [FREE Full text] [doi: 10.2196/jmir.5975] [Medline: 27554525]
163. Kanera IM, Willems RA, Bolman CA, Mesters I, Verboon P, Lechner L. Long-term effects of a web-based cancer aftercare intervention on moderate physical activity and vegetable consumption among early cancer survivors: a randomized controlled trial. Int J Behav Nutr Phys Act 2017 Feb 10;14(1):19 [FREE Full text] [doi: 10.1186/s12966-017-0474-2] [Medline: 28187725]
164. Kassavou A, A'Court CE, Chauhan J, Brimocombe JD, Bhattacharya D, Naughton F, et al. Assessing the acceptability of a text messaging service and smartphone app to support patient adherence to medications prescribed for high blood pressure: a pilot study. Pilot Feasibility Stud 2020 Sep 21;6(1):134 [FREE Full text] [doi: 10.1186/s40814-020-00666-2] [Medline: 32974043]
165. Kaushal T, Katz LE, Joseph J, Marowitz M, Morales KH, Atkins D, et al. A text messaging intervention with financial incentive for adolescents with type 1 diabetes. J Diabetes Sci Technol 2022 Jan 30;16(1):120-127 [FREE Full text] [doi: 10.1177/1932296820952786] [Medline: 32864990]
166. Kauw D, Koole MA, Winter MM, Dohmen DA, Tulevski II, Blok S, et al. Advantages of mobile health in the management of adult patients with congenital heart disease. Int J Med Inform 2019 Dec;132:104011 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104011] [Medline: 31654966]
167. Kay MC, Burroughs J, Askew S, Bennett GG, Armstrong S, Steinberg DM. Digital weight loss intervention for parents of children being treated for obesity: a prospective cohort feasibility trial. J Med Internet Res 2018 Dec 20;20(12):e11093 [FREE Full text] [doi: 10.2196/11093] [Medline: 30573449]
168. Kelechi TJ, Prentice MA, Mueller M, Madisetti M, Vertegel A. A lower leg physical activity intervention for individuals with chronic venous leg ulcers: randomized controlled trial. JMIR Mhealth Uhealth 2020 May 15;8(5):e15015 [FREE Full text] [doi: 10.2196/15015] [Medline: 32412419]
169. Kenyon CC, Sundar KG, Gruschow SM, Quarshie WO, Feudtner C, Bryant-Stephens TC, et al. Tailored medication adherence incentives for high-risk children with asthma: a pilot study. J Asthma 2020 Dec;57(12):1372-1378. [doi: 10.1080/02770903.2019.1648503] [Medline: 31389724]
170. Kim H, Faw M, Michaelides A. Mobile but connected: harnessing the power of self-efficacy and group support for weight loss success through mHealth intervention. J Health Commun 2017 May 24;22(5):395-402. [doi: 10.1080/10810730.2017.1296510] [Medline: 28339324]
171. Kim MY, Lee SY, Jo EJ, Lee S, Kang M, Song W, et al. Feasibility of a smartphone application based action plan and monitoring in asthma. Asia Pac Allergy 2016 Jul;6(3):174-180 [FREE Full text] [doi: 10.5415/apallergy.2016.6.3.174] [Medline: 27489790]
172. King-Dowling S, Psihogios AM, Hill-Kayser C, Szalda D, O'Hagan B, Darabos K, et al. Acceptability and feasibility of survivorship care plans and an accompanying mobile health intervention for adolescent and young adult survivors of childhood cancer. Pediatr Blood Cancer 2021 Mar 08;68(3):e28884 [FREE Full text] [doi: 10.1002/pbc.28884] [Medline: 33416214]

XSL•FO
RenderX

173. Klausen SH, Andersen LL, Søndergaard L, Jakobsen JC, Zoffmann V, Dideriksen K, et al. Effects of eHealth physical activity encouragement in adolescents with complex congenital heart disease: the PReVaiL randomized clinical trial. Int J Cardiol 2016 Oct 15;221:1100-1106. [doi: 10.1016/j.ijcard.2016.07.092] [Medline: 27448540]
174. Kliemann N, Croker H, Johnson F, Beeken RJ. Development of the top tips habit-based weight loss app and preliminary indications of its usage, effectiveness, and acceptability: mixed-methods pilot study. JMIR Mhealth Uhealth 2019 May 10;7(5):e12326 [FREE Full text] [doi: 10.2196/12326] [Medline: 31094352]
175. Kooij L, Vos PJ, Dijkstra A, van Harten WH. Effectiveness of a mobile health and self-management app for high-risk patients with chronic obstructive pulmonary disease in daily clinical practice: mixed methods evaluation study. JMIR Mhealth Uhealth 2021 Feb 04;9(2):e21977 [FREE Full text] [doi: 10.2196/21977] [Medline: 33538699]
176. Koot D, Goh PS, Lim RS, Tian Y, Yau TY, Tan NC, et al. A mobile lifestyle management program (GlycoLeap) for people with type 2 diabetes: single-arm feasibility study. JMIR Mhealth Uhealth 2019 May 24;7(5):e12965 [FREE Full text] [doi: 10.2196/12965] [Medline: 31127720]
177. Korus M, Cruchley E, Calic M, Gold A, Anthony SJ, Parekh RS, et al. Assessing the acceptability and efficacy of teens taking charge: transplant-a pilot randomized control trial. Pediatr Transplant 2020 Feb 19;24(1):e13612. [doi: 10.1111/petr.13612] [Medline: 31743564]
178. Kosse RC, Bouvy ML, Belitser SV, de Vries TW, van der Wal PS, Koster ES. Effective engagement of adolescent asthma patients with mobile health-supporting medication adherence. JMIR Mhealth Uhealth 2019 Mar 27;7(3):e12411 [FREE Full text] [doi: 10.2196/12411] [Medline: 30916664]
179. Krishnakumar A, Verma R, Chawla R, Sosale A, Saboo B, Joshi S, et al. Evaluating glycemic control in patients of South Asian origin with type 2 diabetes using a digital therapeutic platform: analysis of real-world data. J Med Internet Res 2021 Mar 25;23(3):e17908 [FREE Full text] [doi: 10.2196/17908] [Medline: 33764306]
180. Kubo A, Aghaee S, Kurtovich EM, Nkemere L, Quesenberry CP, McGinnis MK, et al. mHealth mindfulness intervention for women with moderate-to-moderately-severe antenatal depressive symptoms: a pilot study within an integrated health care system. Mindfulness (N Y) 2021;12(6):1387-1397 [FREE Full text] [doi: 10.1007/s12671-021-01606-8] [Medline: 33723491]
181. Kumar S, Moseson H, Uppal J, Juusola JL. A diabetes mobile app with in-app coaching from a certified diabetes educator reduces A1C for individuals with type 2 diabetes. Diabetes Educ 2018 Jun;44(3):226-236. [doi: 10.1177/0145721718765650] [Medline: 29575982]
182. Lakshminarayana R, Wang D, Burn D, Chaudhuri KR, Galtrey C, Guzman NV, Ben James, et al. Using a smartphone-based self-management platform to support medication adherence and clinical consultation in Parkinson's disease. NPJ Parkinsons Dis 2017;3:2 [FREE Full text] [doi: 10.1038/s41531-016-0003-z] [Medline: 28649602]
183. Lalloo C, Hundert A, Harris L, Pham Q, Campbell F, Chorney J, et al. Capturing daily disease experiences of adolescents with chronic pain: mHealth-mediated symptom tracking. JMIR Mhealth Uhealth 2019 Jan 17;7(1):e11838 [FREE Full text] [doi: 10.2196/11838] [Medline: 30664472]
184. Lambert SD, Grover S, Laizner AM, McCusker J, Belzile E, Moodie EE, et al. Adaptive web-based stress management programs among adults with a cardiovascular disease: a pilot Sequential Multiple Assignment Randomized Trial (SMART). Patient Educ Couns 2022 Jun;105(6):1587-1597. [doi: 10.1016/j.pec.2021.10.020] [Medline: 34753612]
185. Leach CR, Hudson SV, Diefenbach MA, Wiseman KP, Sanders A, Coa K, et al. Cancer health self-efficacy improvement in a randomized controlled trial. Cancer 2022 Feb 01;128(3):597-605 [FREE Full text] [doi: 10.1002/cncr.33947] [Medline: 34668569]
186. Lee H, Uhm KE, Cheong IY, Yoo JS, Chung SH, Park YH, et al. Patient satisfaction with mobile health (mHealth) application for exercise intervention in breast cancer survivors. J Med Syst 2018 Nov 06;42(12):254. [doi: 10.1007/s10916-018-1096-1] [Medline: 30402781]
187. Lee KW, Kim HB, Lee SH, Ha HK. Changes in weight and health-related behavior using smartphone applications in patients with colorectal polyps. J Nutr Educ Behav 2019 May;51(5):539-546. [doi: 10.1016/j.jneb.2019.02.002] [Medline: 30902428]
188. Lee MK, Lee DY, Ahn HY, Park CY. A novel user utility score for diabetes management using tailored mobile coaching: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2021 Feb 24;9(2):e17573 [FREE Full text] [doi: 10.2196/17573] [Medline: 33625363]
189. Lee SE, Park SK, Park YS, Kim KA, Choi HS, Oh SW. Effects of short-term mobile application use on weight reduction for patients with type 2 diabetes. J Obes Metab Syndr 2021 Dec 30;30(4):345-353 [FREE Full text] [doi: 10.7570/jomes21047] [Medline: 34875628]
190. Lefler LL, Rhoads SJ, Harris M, Funderburg AE, Lubin SA, Martel ID, et al. Evaluating the use of mobile health technology in older adults with heart failure: mixed-methods study. JMIR Aging 2018 Dec 04;1(2):e12178 [FREE Full text] [doi: 10.2196/12178] [Medline: 31518257]
191. Leon N, Namadingo H, Cooper S, Bobrow K, Mwantisi C, Nyasulu M, et al. Process evaluation of a brief messaging intervention to improve diabetes treatment adherence in sub-Saharan Africa. BMC Public Health 2021 Aug 21;21(1):1576 [FREE Full text] [doi: 10.1186/s12889-021-11552-8] [Medline: 34418987]

XSL•FO
RenderX

192. Leonard S, Anderson LM, Jonassaint J, Jonassaint C, Shah N. Utilizing a novel mobile health "selfie" application to improve compliance to iron chelation in pediatric patients receiving chronic transfusions. J Pediatr Hematol Oncol 2017 Apr;39(3):223-229. [doi: 10.1097/MPH.0000000000000743] [Medline: 28099398]
193. Levin JB, Sajatovic M, Rahman M, Aebi ME, Tatsuoka C, Depp C, et al. Outcomes of psychoeducation and a text messaging adherence intervention among individuals with hypertension and bipolar disorder. Psychiatr Serv 2019 Jul 01;70(7):608-612 [FREE Full text] [doi: 10.1176/appi.ps.201800482] [Medline: 30991908]
194. Lewis S, Ainsworth J, Sanders C, Stockton-Powdrell C, Machin M, Whelan P, et al. Smartphone-enhanced symptom management in psychosis: open, randomized controlled trial. J Med Internet Res 2020 Aug 13;22(8):e17019 [FREE Full text] [doi: 10.2196/17019] [Medline: 32788150]
195. Loeckx M, Rabinovich RA, Demeyer H, Louvaris Z, Tanner R, Rubio N, et al. Smartphone-based physical activity telecoaching in chronic obstructive pulmonary disease: mixed-methods study on patient experiences and lessons for implementation. JMIR Mhealth Uhealth 2018 Dec 21;6(12):e200 [FREE Full text] [doi: 10.2196/mhealth.9774] [Medline: 30578215]
196. Lopez KE, Salvy SJ, Fink C, Werner J, Wee CP, Hegedus E, et al. Executive functioning, depressive symptoms, and intervention engagement in a sample of adolescents enrolled in a weight management program. Child Obes 2021 Jun 01;17(4):281-290 [FREE Full text] [doi: 10.1089/chi.2020.0334] [Medline: 33826861]
197. Malte CA, Dulin PL, Baer JS, Fortney JC, Danner AN, Lott AM, et al. Usability and acceptability of a mobile app for the self-management of alcohol misuse among Veterans (Step Away): pilot cohort study. JMIR Mhealth Uhealth 2021 Apr 08;9(4):e25927 [FREE Full text] [doi: 10.2196/25927] [Medline: 33830064]
198. Mamykina L, Heitkemper EM, Smaldone AM, Kukafka R, Cole-Lewis H, Davidson PG, et al. Structured scaffolding for reflection and problem solving in diabetes self-management: qualitative study of mobile diabetes detective. J Am Med Inform Assoc 2016 Jan 14;23(1):129-136 [FREE Full text] [doi: 10.1093/jamia/ocv169] [Medline: 26769910]
199. Mantani A, Kato T, Furukawa TA, Horikoshi M, Imai H, Hiroe T, et al. Smartphone cognitive behavioral therapy as an adjunct to pharmacotherapy for refractory depression: randomized controlled trial. J Med Internet Res 2017 Nov 03;19(11):e373 [FREE Full text] [doi: 10.2196/jmir.8602] [Medline: 29101095]
200. Mayberry LS, Berg CA, Greevy RA, Nelson LA, Bergner EM, Wallston KA, et al. Mixed-methods randomized evaluation of FAMS: a mobile phone-delivered intervention to improve family/friend involvement in adults' type 2 diabetes self-care. Ann Behav Med 2021 Mar 16;55(2):165-178 [FREE Full text] [doi: 10.1093/abm/kaaa041] [Medline: 32706852]
201. McClure JB, Anderson ML, Bradley K, An LC, Catz SL. Evaluating an adaptive and interactive mHealth smoking cessation and medication adherence program: a randomized pilot feasibility study. JMIR Mhealth Uhealth 2016 Aug 03;4(3):e94 [FREE Full text] [doi: 10.2196/mhealth.6002] [Medline: 27489247]
202. McGill DE, Volkening LK, Butler DA, Wasserman RM, Anderson BJ, Laffel LM. Text-message responsiveness to blood glucose monitoring reminders is associated with HbA benefit in teenagers with type 1 diabetes. Diabet Med 2019 May 25;36(5):600-605 [FREE Full text] [doi: 10.1111/dme.13929] [Medline: 30734361]
203. McGillicuddy JW, Chandler JL, Sox LR, Taber DJ. Exploratory analysis of the impact of an mHealth medication adherence intervention on tacrolimus trough concentration variability: post hoc results of a randomized controlled trial. Ann Pharmacother 2020 Dec 08;54(12):1185-1193 [FREE Full text] [doi: 10.1177/1060028020931806] [Medline: 32506922]
204. Melilli E, Cestone G, Revuelta I, Meneghini M, Lladó L, Montero N, et al. Adoption of a novel smart mobile-health application technology to track chronic immunosuppression adherence in solid organ transplantation: Results of a prospective, observational, multicentre, pilot study. Clin Transplant 2021 May 16;35(5):e14278. [doi: 10.1111/ctr.14278] [Medline: 33682207]
205. Melissant HC, Verdonck-de Leeuw IM, Lissenberg-Witte BI, Konings IR, Cuijpers P, Van Uden-Kraan CF. 'Oncokompas', a web-based self-management application to support patient activation and optimal supportive care: a feasibility study among breast cancer survivors. Acta Oncol 2018 Jul;57(7):924-934. [doi: 10.1080/0284186X.2018.1438654] [Medline: 29451059]
206. Merwin RM, Moskovich AA, Babyak M, Feinglos M, Honeycutt LK, Mooney J, et al. An open trial of app-assisted acceptance and commitment therapy (iACT) for eating disorders in type 1 diabetes. J Eat Disord 2021 Jan 06;9(1):6 [FREE Full text] [doi: 10.1186/s40337-020-00357-6] [Medline: 33407910]
207. Minen MT, Adhikari S, Padikkala J, Tasneem S, Bagheri A, Goldberg E, et al. Smartphone-delivered progressive muscle relaxation for the treatment of migraine in primary care: a randomized controlled trial. Headache 2020 Nov;60(10):2232-2246 [FREE Full text] [doi: 10.1111/head.14010] [Medline: 33200413]
208. Minen MT, Schaubhut KB, Morio K. Smartphone based behavioral therapy for pain in multiple sclerosis (MS) patients: a feasibility acceptability randomized controlled study for the treatment of comorbid migraine and MS pain. Mult Scler Relat Disord 2020 Nov;46:102489 [FREE Full text] [doi: 10.1016/j.msard.2020.102489] [Medline: 32950893]
209. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res 2019 Jun 08;21(6):e12556 [FREE Full text] [doi: 10.2196/12556] [Medline: 31199319]
210. Mollard E, Michaud K. A mobile app with optical imaging for the self-management of hand rheumatoid arthritis: pilot study. JMIR Mhealth Uhealth 2018 Oct 29;6(10):e12221 [FREE Full text] [doi: 10.2196/12221] [Medline: 30373732]

XSL•FO
RenderX

211. Mollerup A, Harboe G, Johansen JD. User evaluation of patient counselling, combining nurse consultation and eHealth in hand eczema. Contact Dermatitis 2016 Apr 29;74(4):205-216. [doi: 10.1111/cod.12534] [Medline: 26822623]
212. Moore DJ, Pasipanodya EC, Umlauf A, Rooney AS, Gouaux B, Depp CA, et al. Individualized texting for adherence building (iTAB) for methamphetamine users living with HIV: a pilot randomized clinical trial. Drug Alcohol Depend 2018 Aug 01;189:154-160 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.05.013] [Medline: 29958127]
213. Morita PP, Yeung MS, Ferrone M, Taite AK, Madeley C, Stevens Lavigne A, et al. A patient-centered mobile health system that supports asthma self-management (breathe): design, development, and utilization. JMIR Mhealth Uhealth 2019 Jan 28;7(1):e10956 [FREE Full text] [doi: 10.2196/10956] [Medline: 30688654]
214. Moyano D, Morelli D, Santero M, Belizan M, Irazola V, Beratarrechea A. Perceptions and acceptability of text messaging for diabetes care in primary care in Argentina: exploratory study. JMIR Diabetes 2019 Mar 18;4(1):e10350 [FREE Full text] [doi: 10.2196/10350] [Medline: 30882362]
215. Mulvaney SA, Vaala S, Hood KK, Lybarger C, Carroll R, Williams L, et al. Mobile momentary assessment and biobehavioral feedback for adolescents with type 1 diabetes: feasibility and engagement patterns. Diabetes Technol Ther 2018 Jul;20(7):465-474 [FREE Full text] [doi: 10.1089/dia.2018.0064] [Medline: 29882677]
216. Muralidharan S, Ranjani H, Mohan Anjana R, Jena S, Tandon N, Gupta Y, et al. Engagement and weight loss: results from the mobile health and diabetes trial. Diabetes Technol Ther 2019 Sep;21(9):507-513. [doi: 10.1089/dia.2019.0134] [Medline: 31184922]
217. Murphy J, McSharry J, Hynes L, Molloy GJ. A smartphone app to support adherence to inhaled corticosteroids in young adults with asthma: multi-methods feasibility study. JMIR Form Res 2021 Sep 01;5(9):e28784 [FREE Full text] [doi: 10.2196/28784] [Medline: 34468325]
218. Nelson LA, Greevy RA, Spieker A, Wallston KA, Elasy TA, Kripalani S, et al. Effects of a tailored text messaging intervention among diverse adults with type 2 diabetes: evidence from the 15-month REACH randomized controlled trial. Diabetes Care 2021 Jan;44(1):26-34 [FREE Full text] [doi: 10.2337/dc20-0961] [Medline: 33154039]
219. Nelson LA, Mulvaney SA, Gebretsadik T, Ho Y, Johnson KB, Osborn CY. Disparities in the use of a mHealth medication adherence promotion intervention for low-income adults with type 2 diabetes. J Am Med Inform Assoc 2016 Jan;23(1):12-18 [FREE Full text] [doi: 10.1093/jamia/ocv082] [Medline: 26186935]
220. Nelson LA, Spieker A, Greevy R, LeStourgeon LM, Wallston KA, Mayberry LS. User engagement among diverse adults in a 12-month text message-delivered diabetes support intervention: results from a randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17534 [FREE Full text] [doi: 10.2196/17534] [Medline: 32706738]
221. Nichols M, Miller S, Treiber F, Ruggiero K, Dawley E, Teufel Ii R. Patient and parent perspectives on improving pediatric asthma self-management through a mobile health intervention: pilot study. JMIR Form Res 2020 Jul 03;4(7):e15295 [FREE Full text] [doi: 10.2196/15295] [Medline: 32442127]
222. Nilsson A, Sörman K, Klingvall J, Ovelius E, Lundberg J, Hellner C. MyCompass in a Swedish context - lessons learned from the transfer of a self-guided intervention targeting mental health problems. BMC Psychiatry 2019 Jan 31;19(1):51 [FREE Full text] [doi: 10.1186/s12888-019-2039-1] [Medline: 30704424]
223. Oehler C, Scholze K, Reich H, Sander C, Hegerl U. Intervention use and symptom change with unguided internet-based cognitive behavioral therapy for depression during the COVID-19 pandemic: log data analysis of a convenience sample. JMIR Ment Health 2021 Jul 16;8(7):e28321 [FREE Full text] [doi: 10.2196/28321] [Medline: 34115604]
224. Olalla J, García de Lomas JM, Márquez E, González FJ, Del Arco A, De La Torre J, et al. Experience of using an app in HIV patients older than 60 years: pilot program. JMIR Mhealth Uhealth 2019 Mar 06;7(3):e9904 [FREE Full text] [doi: 10.2196/mhealth.9904] [Medline: 30839281]
225. Ong SW, Jassal SV, Miller JA, Porter EC, Cafazzo JA, Seto E, et al. Integrating a smartphone-based self-management system into usual care of advanced CKD. Clin J Am Soc Nephrol 2016 Jun 06;11(6):1054-1062 [FREE Full text] [doi: 10.2215/CJN.10681015] [Medline: 27173169]
226. Osborn CY, van Ginkel JR, Rodbard D, Heyman M, Marrero DG, Huddleston B, et al. One Drop | Mobile: an evaluation of hemoglobin A1c improvement linked to app engagement. JMIR Diabetes 2017 Aug 24;2(2):e21 [FREE Full text] [doi: 10.2196/diabetes.8039] [Medline: 30291059]
227. Oser M, Wallace ML, Solano F, Szigethy EM. Guided digital cognitive behavioral program for anxiety in primary care: propensity-matched controlled trial. JMIR Ment Health 2019 Apr 04;6(4):e11981 [FREE Full text] [doi: 10.2196/11981] [Medline: 30946022]
228. Pagan-Ortiz ME, Goulet P, Kogelman L, Levkoff SE, Weitzman PF. Feasibility of a texting intervention to improve medication adherence among older HIV+ African Americans: A mixed-method pilot study. Gerontol Geriatr Med 2019 Jun 27;5:2333721419855662 [FREE Full text] [doi: 10.1177/2333721419855662] [Medline: 31276016]
229. Pagoto S, Tulu B, Waring ME, Goetz J, Bibeau J, Divito J, et al. Slip Buddy app for weight management: randomized feasibility trial of a dietary lapse tracking app. JMIR Mhealth Uhealth 2021 Apr 01;9(4):e24249 [FREE Full text] [doi: 10.2196/24249] [Medline: 33792547]
230. Palermo TM, de la Vega R, Murray C, Law E, Zhou C. A digital health psychological intervention (WebMAP Mobile) for children and adolescents with chronic pain: results of a hybrid effectiveness-implementation stepped-wedge cluster

XSL•FO
RenderX

randomized trial. Pain 2020 Dec 10;161(12):2763-2774 [FREE Full text] [doi: 10.1097/j.pain.0000000000001994] [Medline: 32658147]
231. Patel ML, Hopkins CM, Brooks TL, Bennett GG. Comparing self-monitoring strategies for weight loss in a smartphone app: randomized controlled trial. JMIR Mhealth Uhealth 2019 Feb 28;7(2):e12209 [FREE Full text] [doi: 10.2196/12209] [Medline: 30816851]
232. Perski O, Watson NL, Mull KE, Bricker JB. Identifying content-based engagement patterns in a smoking cessation website and associations with user characteristics and cessation outcomes: a sequence and cluster analysis. Nicotine Tob Res 2021 Jun 08;23(7):1103-1112 [FREE Full text] [doi: 10.1093/ntr/ntab008] [Medline: 33433609]
233. Phillips S, Kanter J, Mueller M, Gulledge A, Ruggiero K, Johnson M, et al. Feasibility of an mHealth self-management intervention for children and adolescents with sickle cell disease and their families. Transl Behav Med 2021 Apr 07;11(3):724-732 [FREE Full text] [doi: 10.1093/tbm/ibaa132] [Medline: 33410488]
234. Poort H, Ryan A, MacDougall K, Malinowski P, MacDonald A, Markin Z, et al. Feasibility and acceptability of a mobile phone app intervention for coping with cancer as a young adult: pilot trial and thematic analysis. J Med Internet Res 2021 Jun 11;23(6):e25069 [FREE Full text] [doi: 10.2196/25069] [Medline: 34114957]
235. Prasad M, Fine K, Gee A, Nair N, Popp CJ, Cheng B, et al. A smartphone intervention to promote time restricted eating reduces body weight and blood pressure in adults with overweight and obesity: a pilot study. Nutrients 2021 Jun 23;13(7):2148 [FREE Full text] [doi: 10.3390/nu13072148] [Medline: 34201442]
236. Psihogios AM, King-Dowling S, O'Hagan B, Darabos K, Maurer L, Young J, et al. Contextual predictors of engagement in a tailored mHealth intervention for adolescent and young adult cancer survivors. Ann Behav Med 2021 Nov 18;55(12):1220-1230 [FREE Full text] [doi: 10.1093/abm/kaab008] [Medline: 33674863]
237. Rabbi M, Aung MS, Gay G, Reid MC, Choudhury T. Feasibility and acceptability of mobile phone-based auto-personalized physical activity recommendations for chronic pain self-management: pilot study on adults. J Med Internet Res 2018 Oct 26;20(10):e10147 [FREE Full text] [doi: 10.2196/10147] [Medline: 30368433]
238. Ramsey SE, Ames EG, Uber J, Habib S, Clark S, Waldrop D. A preliminary test of an mHealth facilitated health coaching intervention to improve medication adherence among persons living with HIV. AIDS Behav 2021 Nov 12;25(11):3782-3797 [FREE Full text] [doi: 10.1007/s10461-021-03342-5] [Medline: 34117965]
239. Reading M, Baik D, Beauchemin M, Hickey K, Merrill J. Factors influencing sustained engagement with ECG self-monitoring: perspectives from patients and health care providers. Appl Clin Inform 2018 Oct 10;9(4):772-781 [FREE Full text] [doi: 10.1055/s-0038-1672138] [Medline: 30304745]
240. Redfern J, Coorey G, Mulley J, Scaria A, Neubeck L, Hafiz N, et al. A digital health intervention for cardiovascular disease management in primary care (CONNECT) randomized controlled trial. NPJ Digit Med 2020 Sep 10;3(1):117 [FREE Full text] [doi: 10.1038/s41746-020-00325-z] [Medline: 32964140]
241. Reyes AT, Bhatta TR, Muthukumar V, Gangozo WJ. Testing the acceptability and initial efficacy of a smartphone-app mindfulness intervention for college student veterans with PTSD. Arch Psychiatr Nurs 2020 Apr;34(2):58-66. [doi: 10.1016/j.apnu.2020.02.004] [Medline: 32248935]
242. Rico TM, Dos Santos Machado K, Fernandes VP, Madruga SW, Noguez PT, Barcelos CR, et al. Text messaging (SMS) helping cancer care in patients undergoing chemotherapy treatment: a pilot study. J Med Syst 2017 Oct 09;41(11):181. [doi: 10.1007/s10916-017-0831-3] [Medline: 28990135]
243. Rizvi SL, Hughes CD, Thomas MC. The DBT coach mobile application as an adjunct to treatment for suicidal and self-injuring individuals with borderline personality disorder: a preliminary evaluation and challenges to client utilization. Psychol Serv 2016 Nov;13(4):380-388. [doi: 10.1037/ser0000100] [Medline: 27797571]
244. Rodriguez Hermosa JL, Fuster Gomila A, Puente Maestu L, Amado Diago CA, Callejas González FJ, Malo De Molina Ruiz R, et al. Compliance and utility of a smartphone app for the detection of exacerbations in patients with chronic obstructive pulmonary disease: cohort study. JMIR Mhealth Uhealth 2020 Mar 19;8(3):e15699 [FREE Full text] [doi: 10.2196/15699] [Medline: 32191213]
245. Rosenberger EM, DeVito Dabbs AJ, DiMartini AF, Landsittel DP, Pilewski JM, Dew MA. Long-term follow-up of a randomized controlled trial evaluating a mobile health intervention for self-management in lung transplant recipients. Am J Transplant 2017 May;17(5):1286-1293 [FREE Full text] [doi: 10.1111/ajt.14062] [Medline: 27664940]
246. Ross EL, Jamison RN, Nicholls L, Perry BM, Nolen KD. Clinical integration of a smartphone app for patients with chronic pain: retrospective analysis of predictors of benefits and patient engagement between clinic visits. J Med Internet Res 2020 Apr 16;22(4):e16939 [FREE Full text] [doi: 10.2196/16939] [Medline: 32297871]
247. Rudolf I, Pieper K, Nolte H, Junge S, Dopfer C, Sauer-Heilborn A, et al. Assessment of a mobile app by adolescents and young adults with cystic fibrosis: pilot evaluation. JMIR Mhealth Uhealth 2019 Nov 21;7(11):e12442 [FREE Full text] [doi: 10.2196/12442] [Medline: 31750841]
248. Rygh P, Asklund I, Samuelsson E. Real-world effectiveness of app-based treatment for urinary incontinence: a cohort study. BMJ Open 2021 Jan 04;11(1):e040819 [FREE Full text] [doi: 10.1136/bmjopen-2020-040819] [Medline: 33397664]
249. Saberi P, Lisha NE, Erguera XA, Hudes ES, Johnson MO, Ruel T, et al. A mobile health app (WYZ) for engagement in care and antiretroviral therapy adherence among youth and young adults living with HIV: single-arm pilot intervention study. JMIR Form Res 2021 Aug 31;5(8):e26861 [FREE Full text] [doi: 10.2196/26861] [Medline: 34463622]

XSL•FO
RenderX

250. Santiago-Torres M, Mull KE, Sullivan BM, Kwon D, Nollen NL, Zvolensky MJ, et al. Efficacy and utilization of an acceptance and commitment therapy-based smartphone application for smoking cessation among Black adults: secondary analysis of the iCanQuit randomized trial. Addiction 2022 Mar 09;117(3):760-771 [FREE Full text] [doi: 10.1111/add.15721] [Medline: 34890104]
251. Santiago-Torres M, Mull KE, Sullivan BM, Kwon DM, Nez Henderson P, Nelson LA, et al. Efficacy and utilization of smartphone applications for smoking cessation among American Indians and Alaska Natives: results from the iCanQuit trial. Nicotine Tob Res 2022 Mar 01;24(4):544-554 [FREE Full text] [doi: 10.1093/ntr/ntab213] [Medline: 34644389]
252. Schlosser DA, Campellone TR, Truong B, Anguera JA, Vergani S, Vinogradov S, et al. The feasibility, acceptability, and outcomes of PRIME-D: a novel mobile intervention treatment for depression. Depress Anxiety 2017 Jun 18;34(6):546-554 [FREE Full text] [doi: 10.1002/da.22624] [Medline: 28419621]
253. Schnall R, Cho H, Mangone A, Pichon A, Jia H. Mobile health technology for improving symptom management in low income persons living with HIV. AIDS Behav 2018 Oct 3;22(10):3373-3383 [FREE Full text] [doi: 10.1007/s10461-017-2014-0] [Medline: 29299790]
254. Schneider T, Baum L, Amy A, Marisa C. I have most of my asthma under control and I know how my asthma acts: users' perceptions of asthma self-management mobile app tailored for adolescents. Health Informatics J 2020 Mar;26(1):342-353 [FREE Full text] [doi: 10.1177/1460458218824734] [Medline: 30732520]
255. Scott CK, Dennis ML, Johnson KA, Grella CE. A randomized clinical trial of smartphone self-managed recovery support services. J Subst Abuse Treat 2020 Oct;117:108089 [FREE Full text] [doi: 10.1016/j.jsat.2020.108089] [Medline: 32811628]
256. Selter A, Tsangouri C, Ali SB, Freed D, Vatchinsky A, Kizer J, et al. An mHealth app for self-management of chronic lower back pain (Limbr): pilot study. JMIR Mhealth Uhealth 2018 Sep 17;6(9):e179 [FREE Full text] [doi: 10.2196/mhealth.8256] [Medline: 30224333]
257. Seng EK, Prieto P, Boucher G, Vives-Mestres M. Anxiety, incentives, and adherence to self-monitoring on a mobile health platform: a naturalistic longitudinal cohort study in people with headache. Headache 2018 Nov 18;58(10):1541-1555 [FREE Full text] [doi: 10.1111/head.13422] [Medline: 30334248]
258. Serlachius A, Schache K, Kieser A, Arroll B, Petrie K, Dalbeth N. Association between user engagement of a mobile health app for gout and improvements in self-care behaviors: randomized controlled trial. JMIR Mhealth Uhealth 2019 Aug 13;7(8):e15021 [FREE Full text] [doi: 10.2196/15021] [Medline: 31411147]
259. Serper M, Barankay I, Chadha S, Shults J, Jones LS, Olthoff KM, et al. A randomized, controlled, behavioral intervention to promote walking after abdominal organ transplantation: results from the LIFT study. Transpl Int 2020 Jun 12;33(6):632-643 [FREE Full text] [doi: 10.1111/tri.13570] [Medline: 31925833]
260. Shaw RJ, Yang Q, Barnes A, Hatch D, Crowley MJ, Vorderstrasse A, et al. Self-monitoring diabetes with multiple mobile health devices. J Am Med Inform Assoc 2020 May 01;27(5):667-676 [FREE Full text] [doi: 10.1093/jamia/ocaa007] [Medline: 32134447]
261. Shebib R, Bailey JF, Smittenaar P, Perez DA, Mecklenburg G, Hunter S. Randomized controlled trial of a 12-week digital care program in improving low back pain. NPJ Digit Med 2019;2:1 [FREE Full text] [doi: 10.1038/s41746-018-0076-7] [Medline: 31304351]
262. Shingleton RM, Pratt EM, Gorman B, Barlow DH, Palfai TP, Thompson-Brenner H. Motivational text message intervention for eating disorders: a single-case alternating treatment design using ecological momentary assessment. Behav Ther 2016 May;47(3):325-338. [doi: 10.1016/j.beth.2016.01.005] [Medline: 27157027]
263. Siengsukon CF, Silveira Beck Jr E, Drerup M. Feasibility and treatment effect of a web-based cognitive behavioral therapy for insomnia program in individuals with multiple sclerosis: a pilot randomized controlled trial. Int J MS Care 2021;23(3):107-113 [FREE Full text] [doi: 10.7224/1537-2073.2019-122] [Medline: 34177382]
264. Signal V, McLeod M, Stanley J, Stairmand J, Sukumaran N, Thompson D, et al. A mobile- and web-based health intervention program for diabetes and prediabetes self-management (BetaMe/Melon): process evaluation following a randomized controlled trial. J Med Internet Res 2020 Dec 01;22(12):e19150 [FREE Full text] [doi: 10.2196/19150] [Medline: 33258776]
265. Slepian PM, Peng M, Janmohamed T, Kotteeswaran Y, Manoo V, Blades AM, et al. Engagement with manage my pain mobile health application among patients at the transitional pain service. Digit Health 2020 Nov;6(7):2055207620962297 [FREE Full text] [doi: 10.1177/2055207620962297] [Medline: 33117557]
266. Spring B, Pellegrini CA, Pfammatter A, Duncan JM, Pictor A, McFadden HG, et al. Effects of an abbreviated obesity intervention supported by mobile technology: the ENGAGED randomized clinical trial. Obesity (Silver Spring) 2017 Jul;25(7):1191-1198 [FREE Full text] [doi: 10.1002/oby.21842] [Medline: 28494136]
267. St-Jules DE, Woolf K, Goldfarb DS, Pompeii ML, Li H, Wang C, et al. Feasibility and acceptability of mHealth interventions for managing hyperphosphatemia in patients undergoing hemodialysis. J Ren Nutr 2021 Jul;31(4):403-410. [doi: 10.1053/j.jrn.2020.07.009] [Medline: 33160812]
268. Steare T, O'Hanlon P, Eskinazi M, Osborn D, Lloyd-Evans B, Jones R, et al. Smartphone-delivered self-management for first-episode psychosis: the ARIES feasibility randomised controlled trial. BMJ Open 2020 Aug 26;10(8):e034927 [FREE Full text] [doi: 10.1136/bmjopen-2019-034927] [Medline: 32847902]

XSL•FO
RenderX

269. Steinert A, Eicher C, Haesner M, Steinhagen-Thiessen E. Effects of a long-term smartphone-based self-monitoring intervention in patients with lipid metabolism disorders. Assist Technol 2020;32(2):109-116. [doi: 10.1080/10400435.2018.1493710] [Medline: 29944463]
270. Stolz T, Schulz A, Krieger T, Vincent A, Urech A, Moser C, et al. A mobile app for social anxiety disorder: a three-arm randomized controlled trial comparing mobile and PC-based guided self-help interventions. J Consult Clin Psychol 2018 Jun;86(6):493-504. [doi: 10.1037/ccp0000301] [Medline: 29781648]
271. Strauss C, Dunkeld C, Cavanagh K. Is clinician-supported use of a mindfulness smartphone app a feasible treatment for depression? A mixed-methods feasibility study. Internet Interv 2021 Sep;25:100413 [FREE Full text] [doi: 10.1016/j.invent.2021.100413] [Medline: 34401372]
272. Stubbins R, He T, Yu X, Puppala M, Ezeana CF, Chen S, et al. A behavior-modification, clinical-grade mobile application to improve breast cancer survivors' accountability and health outcomes. JCO Clin Cancer Inform 2018 Dec;2:1-11 [FREE Full text] [doi: 10.1200/CCI.18.00054] [Medline: 30652617]
273. Su J, Dugas M, Guo X, Gao G. Influence of personality on mHealth use in patients with diabetes: prospective pilot study. JMIR Mhealth Uhealth 2020 Aug 10;8(8):e17709 [FREE Full text] [doi: 10.2196/17709] [Medline: 32773382]
274. Sundström C, Gajecki M, Johansson M, Blankers M, Sinadinovic K, Stenlund-Gens E, et al. Guided and unguided internet-based treatment for problematic alcohol use - a randomized controlled pilot trial. PLoS One 2016;11(7):e0157817 [FREE Full text] [doi: 10.1371/journal.pone.0157817] [Medline: 27383389]
275. Swendeman D, Sumstine S, Aguilar E, Gorbach P, Comulada W, Gelberg L. Feasibility and acceptability of mobile phone self-monitoring and automated feedback to enhance telephone coaching for people with risky substance use: the QUIT-mobile pilot study. J Addict Med 2021 Apr 01;15(2):120-129 [FREE Full text] [doi: 10.1097/ADM.0000000000000707] [Medline: 32732684]
276. Talboom-Kamp EP, Verdijk NA, Kasteleyn MJ, Harmans LM, Talboom IJ, Numans ME, et al. High level of integration in integrated disease management leads to higher usage in the e-Vita study: self-management of chronic obstructive pulmonary disease with web-based platforms in a parallel cohort design. J Med Internet Res 2017 May 31;19(5):e185 [FREE Full text] [doi: 10.2196/jmir.7037] [Medline: 28566268]
277. Talboom-Kamp EP, Holstege MS, Chavannes NH, Kasteleyn MJ. Effects of use of an eHealth platform e-Vita for COPD patients on disease specific quality of life domains. Respir Res 2019 Jul 10;20(1):146 [FREE Full text] [doi: 10.1186/s12931-019-1110-2] [Medline: 31291945]
278. Talboom-Kamp EP, Verdijk NA, Kasteleyn MJ, Harmans LM, Talboom IJ, Numans ME, et al. Effect of a combined education and eHealth programme on the control of oral anticoagulation patients (PORTALS study): a parallel cohort design in Dutch primary care. BMJ Open 2017 Sep 27;7(9):e017909 [FREE Full text] [doi: 10.1136/bmjopen-2017-017909] [Medline: 28963313]
279. Tighe J, Shand F, Ridani R, Mackinnon A, De La Mata N, Christensen H. Ibobbly mobile health intervention for suicide prevention in Australian Indigenous youth: a pilot randomised controlled trial. BMJ Open 2017 Jan 27;7(1):e013518 [FREE Full text] [doi: 10.1136/bmjopen-2016-013518] [Medline: 28132007]
280. Tincopa MA, Lyden A, Wong J, Jackson EA, Richardson C, Lok AS. Impact of a pilot structured mobile technology based lifestyle intervention for patients with nonalcoholic fatty liver disease. Dig Dis Sci 2022 Feb 03;67(2):481-491 [FREE Full text] [doi: 10.1007/s10620-021-06922-6] [Medline: 33939147]
281. Tombor I, Beard E, Brown J, Shahab L, Michie S, West R. Randomized factorial experiment of components of the SmokeFree Baby smartphone application to aid smoking cessation in pregnancy. Transl Behav Med 2019 Jul 16;9(4):583-593 [FREE Full text] [doi: 10.1093/tbm/iby073] [Medline: 30011020]
282. Tsui JI, Leroux BG, Radick AC, Schramm ZA, Blalock K, Labelle C, et al. Video directly observed therapy for patients receiving office-based buprenorphine - a pilot randomized controlled trial. Drug Alcohol Depend 2021 Oct 01;227:108917 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108917] [Medline: 34399136]
283. Tu YZ, Chang YT, Chiou HY, Lai K. The effects of continuous usage of a diabetes management app on glycemic control in real-world clinical practice: retrospective analysis. J Med Internet Res 2021 Jul 15;23(7):e23227 [FREE Full text] [doi: 10.2196/23227] [Medline: 34264192]
284. Turner-McGrievy GM, Wilcox S, Boutté A, Hutto BE, Singletary C, Muth ER, et al. The Dietary Intervention to Enhance Tracking with Mobile Devices (DIET Mobile) study: a 6-month randomized weight loss trial. Obesity (Silver Spring) 2017 Aug;25(8):1336-1342 [FREE Full text] [doi: 10.1002/oby.21889] [Medline: 28600833]
285. Van Blarigan EL, Chan H, Van Loon K, Kenfield SA, Chan JM, Mitchell E, et al. Self-monitoring and reminder text messages to increase physical activity in colorectal cancer survivors (Smart Pace): a pilot randomized controlled trial. BMC Cancer 2019 Mar 11;19(1):218 [FREE Full text] [doi: 10.1186/s12885-019-5427-5] [Medline: 30866859]
286. Van Blarigan EL, Kenfield SA, Chan JM, Van Loon K, Paciorek A, Zhang L, et al. Feasibility and acceptability of a web-based dietary intervention with text messages for colorectal cancer: a randomized pilot trial. Cancer Epidemiol Biomarkers Prev 2020 Apr;29(4):752-760 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0840] [Medline: 31941707]
287. Van Tiem J, Moeckli J, Suiter N, Fuhrmeister L, Pham K, Dindo L, et al. "A link to the outside:" patient perspectives on a mobile texting program to improve depression self-management. Patient Educ Couns 2021 Sep;104(9):2154-2158. [doi: 10.1016/j.pec.2021.02.018] [Medline: 33627234]

XSL•FO
RenderX

288. Vorrink S, Huisman C, Kort H, Troosters T, Lammers J. Perceptions of patients with chronic obstructive pulmonary disease and their physiotherapists regarding the use of an eHealth intervention. JMIR Hum Factors 2017 Sep 19;4(3):e20 [FREE Full text] [doi: 10.2196/humanfactors.7196] [Medline: 28928110]
289. Wadensten T, Nyström E, Franzén K, Lindam A, Wasteson E, Samuelsson E. A mobile app for self-management of urgency and mixed urinary incontinence in women: randomized controlled trial. J Med Internet Res 2021 Apr 05;23(4):e19439 [FREE Full text] [doi: 10.2196/19439] [Medline: 33818395]
290. Ware P, Dorai M, Ross HJ, Cafazzo JA, Laporte A, Boodoo C, et al. Patient adherence to a mobile phone-based heart failure telemonitoring program: a longitudinal mixed-methods study. JMIR Mhealth Uhealth 2019 Feb 26;7(2):e13259 [FREE Full text] [doi: 10.2196/13259] [Medline: 30806625]
291. Waselewski ME, Flickinger TE, Canan C, Harrington W, Franklin T, Otero KN, et al. A mobile health app to support patients receiving medication-assisted treatment for opioid use disorder: development and feasibility study. JMIR Form Res 2021 Feb 23;5(2):e24561 [FREE Full text] [doi: 10.2196/24561] [Medline: 33620324]
292. Watterson JL, Rodriguez HP, Shortell SM, Aguilera A. Improved diabetes care management through a text-message intervention for low-income patients: mixed-methods pilot study. JMIR Diabetes 2018 Oct 30;3(4):e15 [FREE Full text] [doi: 10.2196/diabetes.8645] [Medline: 30377141]
293. Wei KS, Ibrahim NE, Kumar AA, Jena S, Chew V, Depa M, et al. Habits heart app for patient engagement in heart failure management: pilot feasibility randomized trial. JMIR Mhealth Uhealth 2021 Jan 20;9(1):e19465 [FREE Full text] [doi: 10.2196/19465] [Medline: 33470941]
294. Werner-Seidler A, Wong Q, Johnston L, O'Dea B, Torok M, Christensen H. Pilot evaluation of the Sleep Ninja: a smartphone application for adolescent insomnia symptoms. BMJ Open 2019 May 27;9(5):e026502 [FREE Full text] [doi: 10.1136/bmjopen-2018-026502] [Medline: 31133584]
295. Wright AA, Raman N, Staples P, Schonholz S, Cronin A, Carlson K, et al. The HOPE pilot study: harnessing patient-reported outcomes and biometric data to enhance cancer care. JCO Clin Cancer Inform 2018 Dec;2:1-12 [FREE Full text] [doi: 10.1200/CCI.17.00149] [Medline: 30652585]
296. Xu R, Xing M, Javaherian K, Peters R, Ross W, Bernal-Mizrachi C. Improving HbA with glucose self-monitoring in diabetic patients with EpxDiabetes, a phone call and text message-based telemedicine platform: a randomized controlled trial. Telemed J E Health 2020 Jun 01;26(6):784-793 [FREE Full text] [doi: 10.1089/tmj.2019.0035] [Medline: 31621523]
297. Yanez B, Oswald LB, Baik SH, Buitrago D, Iacobelli F, Perez-Tamayo A, et al. Brief culturally informed smartphone interventions decrease breast cancer symptom burden among Latina breast cancer survivors. Psychooncology 2020 Jan 15;29(1):195-203 [FREE Full text] [doi: 10.1002/pon.5281] [Medline: 31693265]
298. Yang K, Oh D, Noh JM, Yoon HG, Sun J, Kim HK, et al. Feasibility of an interactive health coaching mobile app to prevent malnutrition and muscle loss in esophageal cancer patients receiving neoadjuvant concurrent chemoradiotherapy: prospective pilot study. J Med Internet Res 2021 Aug 27;23(8):e28695. [doi: 10.2196/28695]
299. Yang Q, Hatch D, Crowley MJ, Lewinski AA, Vaughn J, Steinberg D, et al. Digital phenotyping self-monitoring behaviors for individuals with type 2 diabetes mellitus: observational study using latent class growth analysis. JMIR Mhealth Uhealth 2020 Jun 11;8(6):e17730 [FREE Full text] [doi: 10.2196/17730] [Medline: 32525492]
300. Yang Y, Lee EY, Kim HS, Lee SH, Yoon KH, Cho JH. Effect of a mobile phone-based glucose-monitoring and feedback system for type 2 diabetes management in multiple primary care clinic settings: cluster randomized controlled trial. JMIR Mhealth Uhealth 2020 Feb 26;8(2):e16266 [FREE Full text] [doi: 10.2196/16266] [Medline: 32130172]
301. Yen S, Ranney ML, Tezanos KM, Chuong A, Kahler CW, Solomon JB, et al. Skills to enhance positivity in suicidal adolescents: results from an open development trial. Behav Modif 2019 Mar;43(2):202-221 [FREE Full text] [doi: 10.1177/0145445517748559] [Medline: 29258328]
302. Yingling L, Allen NA, Litchman ML, Colicchio V, Gibson BS. An evaluation of digital health tools for diabetes self-management in Hispanic adults: exploratory study. JMIR Diabetes 2019 Jul 16;4(3):e12936 [FREE Full text] [doi: 10.2196/12936] [Medline: 31313657]
303. Yoo HJ, Suh EE. Effects of a smartphone-based self-care health diary for heart transplant recipients: a mixed methods study. Appl Nurs Res 2021 Apr;58:151408. [doi: 10.1016/j.apnr.2021.151408] [Medline: 33745556]
304. You C, Chen Y, Chen CH, Lee C, Kuo P, Huang M, et al. Smartphone-based support system (SoberDiary) coupled with a bluetooth breathalyser for treatment-seeking alcohol-dependent patients. Addict Behav 2017 Feb;65:174-178. [doi: 10.1016/j.addbeh.2016.10.017] [Medline: 27816043]
305. Young MD, Morgan PJ. Effect of a gender-tailored eHealth weight loss program on the depressive symptoms of overweight and obese men: pre-post study. JMIR Ment Health 2018 Jan 09;5(1):e1 [FREE Full text] [doi: 10.2196/mental.8920] [Medline: 29317379]
306. Zaslavsky O, Thompson HJ, McCurry SM, Landis CA, Kitsiou S, Ward TM, et al. Use of a wearable technology and motivational interviews to improve sleep in older adults with osteoarthritis and sleep disturbance: a pilot study. Res Gerontol Nurs 2019 Jul 01;12(4):167-173 [FREE Full text] [doi: 10.3928/19404921-20190319-02] [Medline: 30901479]
307. Zeng EY, Heffner JL, Copeland WK, Mull KE, Bricker JB. Get with the program: adherence to a smartphone app for smoking cessation. Addict Behav 2016 Dec;63:120-124 [FREE Full text] [doi: 10.1016/j.addbeh.2016.07.007] [Medline: 27454354]

XSL•FO
RenderX

308. Zeng Y, Guo Y, Li L, Hong YA, Li Y, Zhu M, et al. Relationship between patient engagement and depressive symptoms among people living with HIV in a mobile health intervention: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e20847 [FREE Full text] [doi: 10.2196/20847] [Medline: 33118956]
309. Zhang L, He X, Shen Y, Yu H, Pan J, Zhu W, et al. Effectiveness of smartphone app-based interactive management on glycemic control in Chinese patients with poorly controlled diabetes: randomized controlled trial. J Med Internet Res 2019 Dec 09;21(12):e15401 [FREE Full text] [doi: 10.2196/15401] [Medline: 31815677]
310. Zhang S, Hamburger E, Kahanda S, Lyttle M, Williams R, Jaser SS. Engagement with a text-messaging intervention improves adherence in adolescents with type 1 diabetes: brief report. Diabetes Technol Ther 2018 May;20(5):386-389 [FREE Full text] [doi: 10.1089/dia.2018.0015] [Medline: 29792749]
311. Zhang Y, Liu C, Luo S, Huang J, Li X, Zhou Z. Effectiveness of Lilly Connected Care Program (LCCP) app-based diabetes education for patients with type 2 diabetes treated with insulin: retrospective real-world study. JMIR Mhealth Uhealth 2020 Mar 06;8(3):e17455 [FREE Full text] [doi: 10.2196/17455] [Medline: 32141838]
312. Zhang Y, Liu C, Luo S, Huang J, Yang Y, Ma X, et al. Effectiveness of the family portal function on the Lilly Connected Care Program (LCCP) for patients with type 2 diabetes: retrospective cohort study with propensity score matching. JMIR Mhealth Uhealth 2021 Feb 05;9(2):e25122 [FREE Full text] [doi: 10.2196/25122] [Medline: 33544081]
313. Zheng X, Spatz ES, Bai X, Huo X, Ding Q, Horak P, et al. Effect of text messaging on risk factor management in patients with coronary heart disease: the CHAT randomized clinical trial. Circ Cardiovasc Qual Outcomes 2019 Apr;12(4):e005616. [doi: 10.1161/CIRCOUTCOMES.119.005616] [Medline: 30998400]
314. Nguyen E, Bugno L, Kandah C, Plevinsky J, Poulopoulos N, Wojtowicz A, et al. Is there a good app for that? Evaluating m-Health apps for strategies that promote pediatric medication adherence. Telemed J E Health 2016 Nov;22(11):929-937. [doi: 10.1089/tmj.2015.0211] [Medline: 27070837]
315. Carmody JK, Denson LA, Hommel KA. Content and usability evaluation of medication adherence mobile applications for use in pediatrics. J Pediatr Psychol 2019 Apr 01;44(3):333-342 [FREE Full text] [doi: 10.1093/jpepsy/jsy086] [Medline: 30358863]
316. Hatem S, Long JC, Best S, Fehlberg Z, Nic Giolla Easpaig B, Braithwaite J. Mobile apps for people with rare diseases: review and quality assessment using mobile app rating scale. J Med Internet Res 2022 Jul 26;24(7):e36691 [FREE Full text] [doi: 10.2196/36691] [Medline: 35881435]
317. Adherence to long-term therapies?: evidence for action. World Health Organization. 2003. URL: https://apps.who.int/iris/handle/10665/42682 [accessed 2024-06-18]
318. Röhricht F, Padmanabhan R, Binfield P, Mavji D, Barlow S. Simple mobile technology health management tool for people with severe mental illness: a randomised controlled feasibility trial. BMC Psychiatry 2021 Jul 16;21(1):357 [FREE Full text] [doi: 10.1186/s12888-021-03359-z] [Medline: 34271902]
319. Gordon K, Dainty KN, Steele Gray C, DeLacy J, Shah A, Resnick M, et al. Experiences of complex patients with telemonitoring in a nurse-led model of care: multimethod feasibility study. JMIR Nurs 2020 Sep 29;3(1):e22118 [FREE Full text] [doi: 10.2196/22118] [Medline: 34406972]
320. Zha P, Qureshi R, Porter S, Chao Y, Pacquiao D, Chase S, et al. Utilizing a mobile health intervention to manage hypertension in an underserved community. West J Nurs Res 2020 Mar;42(3):201-209. [doi: 10.1177/0193945919847937] [Medline: 31057081]
321. McLeod C, Norman R, Litton E, Saville BR, Webb S, Snelling TL. Choosing primary endpoints for clinical trials of health care interventions. Contemp Clin Trials Commun 2019 Dec;16:100486 [FREE Full text] [doi: 10.1016/j.conctc.2019.100486] [Medline: 31799474]

Abbreviations

mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by N Cahill; submitted 18.07.23; peer-reviewed by A Leichtle, W Wei, E Vashishtha, H Veldandi, A Bucher; comments to author 24.01.24; revised version received 27.02.24; accepted 29.07.24; published 24.09.24.

Please cite as:
Eaton C, Vallejo N, McDonald X, Wu J, Rodríguez R, Muthusamy N, Mathioudakis N, Riekert KA
User Engagement With mHealth Interventions to Promote Treatment Adherence and Self-Management in People With Chronic Health Conditions: Systematic Review
J Med Internet Res 2024;26:e50508
URL: https://www.jmir.org/2024/1/e50508
doi:10.2196/50508
PMID:39316431

XSL•FO
RenderX

©Cyd Eaton, Natalie Vallejo, Xiomara McDonald, Jasmine Wu, Rosa Rodríguez, Nishanth Muthusamy, Nestoras Mathioudakis, Kristin A Riekert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Processing of Short-Form Content in Clinical Narratives: Systematic Scoping Review

Amila Kugic[1], MSc; Ingrid Martin[2], MSc; Luise Modersohn[2], MSc; Peter Pallaoro[2], MSc; Markus Kreuzthaler[1], PhD; Stefan Schulz[1], MD; Martin Boeker[2], MD

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

[2]Institute for AI and Informatics in Medicine, School of Medicine and Health, Technical University of Munich, Munich, Germany

Corresponding Author:
Markus Kreuzthaler, PhD
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
Auenbruggerplatz 2/5
Graz, 8036
Austria
Phone: 43 316 385 13591
Email: markus.kreuzthaler@medunigraz.at

Abstract

Background: Clinical narratives are essential components of electronic health records. The adoption of electronic health records has increased documentation time for hospital staff, leading to the use of abbreviations and acronyms more frequently. This brevity can potentially hinder comprehension for both professionals and patients.

Objective: This review aims to provide an overview of the types of short forms found in clinical narratives, as well as the natural language processing (NLP) techniques used for their identification, expansion, and disambiguation.

Methods: In the databases Web of Science, Embase, MEDLINE, EBMR (Evidence-Based Medicine Reviews), and ACL Anthology, publications that met the inclusion criteria were searched according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for a systematic scoping review. Original, peer-reviewed publications focusing on short-form processing in human clinical narratives were included, covering the period from January 2018 to February 2023. Short-form types were extracted, and multidimensional research methodologies were assigned to each target objective (identification, expansion, and disambiguation). NLP study recommendations and study characteristics were systematically assigned occurrence rates for evaluation.

Results: Out of a total of 6639 records, only 19 articles were included in the final analysis. Rule-based approaches were predominantly used for identifying short forms, while string similarity and vector representations were applied for expansion. Embeddings and deep learning approaches were used for disambiguation.

Conclusions: The scope and types of what constitutes a clinical short form were often not explicitly defined by the authors. This lack of definition poses challenges for reproducibility and for determining whether specific methodologies are suitable for different types of short forms. Analysis of a subset of NLP recommendations for assessing quality and reproducibility revealed only partial adherence to these recommendations. Single-character abbreviations were underrepresented in studies on clinical narrative processing, as were investigations in languages other than English. Future research should focus on these 2 areas, and each paper should include descriptions of the types of content analyzed.

(J Med Internet Res 2024;26:e57852) doi:10.2196/57852

KEYWORDS

electronic health records; EHR; clinical narratives; natural language processing; machine learning; deep learning; rule-based approach; short-form expression; disambiguation; word embedding; vector representations; language modeling; human-in-the-loop, feature extraction

XSL•FO
RenderX

Introduction

Background and Significance

Clinical narratives, that is, free text authored by health professionals, are a core component of electronic health records (EHRs) within health care information systems. Today, findings reports, progress notes, surgery reports, and discharge summaries nearly seamlessly document the delivery of health care for every patient. However, with the adoption of EHRs, documentation times appear to increase for hospital staff [1-3], and this added time pressure often results in brevity in both documentation and communication. This explains the frequent use of short-form content (ie, abbreviations and acronyms) in routine documentation. While such jargon is generally well understood within a clinical specialty, professionals from other fields, and especially patients, often have to infer the meaning of these short forms in context. In-depth analyses of clinical narratives have highlighted the semantic and lexical ambiguities introduced by the use of short forms. This highlights the dilemma that, while clinical narratives present key information in a compact manner, overly cryptic formulations, if misunderstood, may severely impact patient safety [4-6].

This situation is further complicated when lexicon lookups for short forms yield multiple possible expansions. For example, the abbreviation "MS" in clinical narratives can stand for "morphine sulfate," "multiple sclerosis," or "mass spectrometry." Determining the correct long form requires understanding the context. The complexity increases further with the ad hoc creation of many short forms and their use being restricted to particular institutions. The fact that the same short form can have entirely different meanings across medical specialties or health care institutions significantly influences the choice of processing methodology [4]. As a result of time pressures, clinicians rarely provide the long form alongside the first occurrence of the short form, as is customary in scientific publications and textbooks.

Natural language processing (NLP) has proposed several solutions to address this issue. The complexity of NLP applications varies across languages, depending not only on their grammatical and morphological characteristics (including those of clinical sublanguages) but also on the availability of lexical resources for each (sub)language. In most cases, low-resource languages have insufficient lexical coverage compared with high-resource languages, particularly English. The limited availability of clinical corpora for research is one reason many studies focus on similar data sets that have been released for research purposes after deidentification and ethics approval, such as Medical Information Mart for Intensive Care (MIMIC)-III [7].

In combination with NLP techniques, researchers have utilized existing methodologies to automatically identify, expand, and disambiguate short forms using data-driven shallow and deep learning (DL) approaches. However, current research does not clearly identify which methodologies are most effective in supporting short-form identification, expansion, and disambiguation. This scoping review, therefore, examines these 3 tasks, which represent distinct yet interconnected methodological aspects of each study, as they are closely linked to determining the correct long form for each short form.

This review covers a 5-year period and focuses on the narrative content of clinical data sets, the processing of short forms using state-of-the-art methodologies, and the short forms themselves. To the best of the authors' knowledge, this is the first systematic scoping review to address these specific aspects of short-form processing in clinical narratives.

Objective

This scoping review was conducted to identify relevant original research papers that apply NLP techniques to process short forms, such as acronyms and other abbreviations, in clinical narratives. All described methodologies need to be evaluated or validated in some form.

The objective of this review is to provide a systematic and structured overview of the literature on (1) short forms in clinical narratives and (2) methods used for their identification, expansion, and disambiguation.

Methods

Study Design

The study design adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [8] for conducting a systematic scoping review (Multimedia Appendix 1).

Eligibility Criteria

Eligible articles were full-text, original, peer-reviewed publications that focused on machine learning (ML) or NLP techniques for the identification, expansion, or disambiguation of short-form content.

Identification focuses on detecting short-form content in clinical narratives, such as recognizing the acronym "RA" using the regular expression [A-Z]{2,}. *Expansion* involves generating possible long forms for short forms, such as expanding "RA" to "rheumatoid arthritis," "right atrium," or "room air." Finally, *disambiguation* pertains to methods that determine the correct expansion for a short form. For example, in the context of "hypertension with RVSP of 46+ RA pressure," the correct expansion is "right atrium" [9].

Papers under review were required to be written in either English or German, as all members of the review team are proficient in both languages. Additionally, the methodologies applied required each paper to focus on human clinical narratives, specifically textual content produced by clinicians in human medicine. This selection is based on the complex wording and structure of these narratives, which often include elements from different languages, such as Latin names for body parts or diseases, or untranslated foreign terms, adding complexity to text processing. Papers that did not focus on short forms were excluded.

Meta-analyses, case reports, collections, abstracts, surveys of patient-reported outcomes, papers lacking performance evaluation and validation, those that did not apply any ML or NLP methodology, and all types of reviews were excluded.

XSL•FO
RenderX

Search Strategies

From January 1, 2018, to February 22, 2023, the literature databases Web of Science, Embase, MEDLINE, EBMR (Evidence-Based Medicine Reviews), and ACL Anthology were searched for relevant papers for this systematic scoping review. Web of Science was accessed through the official Clarivate website [10]. MEDLINE, Embase, and EBMR were searched via Ovid [11], while papers from ACL Anthology were retrieved using a custom Python (Python Foundation) search function, as the integrated ACL search does not support the operators NEAR or ADJ (adjacent). The search field operators assigned in each query, such as "TS" for "topic" or ".mp." for "multiple purposes," enable precise selection of papers from specific fields within the database records, thereby improving the accuracy and relevance of the search results. Examples of these fields are titles, abstracts, keywords, and subject headings. For ACL Anthology, a full-text search was conducted for all articles within the specified timeframe.

The search strategy primarily focused on short-form content and did not include "natural language processing" as a keyword, as this might have been too restrictive for capturing all relevant short-form content processing. Instead, the search terms were selected to cover all types of clinical narratives using MeSH (Medical Subject Headings) terms. Processing of any type of human clinical narrative was relevant to this scoping review. Table 1 presents a comprehensive list of search terms and strategies for each database, and the review process can be found in the *Results* section in a flow diagram.

Table 1. Search queries per database.

Database and query number	Search query
Web of Science and ACL Anthology	
#1	TS=((medic* OR clinic* or allerg* OR androlog* OR anesthesiolog* OR anaesthesiolog* OR bariatric* OR biopharmaceutic* OR cardiolog* OR cardiovascul* OR chiropractic* OR cytopatholog* OR dental OR dentistr* OR dermatolog* OR dietetic* OR emergenc* OR endocrinolog* OR endodontic* OR ethnopharmacolog* OR forensic* OR gastroenterolog* OR genomic* OR geriatric* OR gerontolog* OR geroscienc* OR gynaecolog* OR gynecolog* OR haematolog* OR hematolog* OR immunolog* OR immunopatholog* OR microbiolog* OR midwife* OR nanomedic* OR neonatolog* OR nephrolog* OR neurolog* OR neuropatholog* OR neuropharmacolog* OR neuropsychiatr* OR neuroradiolog* OR neurosurg* OR neurotolog* OR nursing OR nutrigenomic* OR obstetric* OR occupational* OR oncolog* OR ophthalmolog* OR optometr* OR orthodontic* OR orthopedic* OR orthoptic* OR otolaryngolog* OR otolog* OR otorhinolaryngolog* OR paramedic* OR patholog* OR pediatric* OR perinatolog* OR periodontic* OR pharmacogenetic* OR pharmacolog* OR pneumolog* OR pneumonolog* OR podiatr* OR proctolog* OR prosthodontic* OR psychiatr* OR psychopharmacolog* OR pulmonolog* OR radiolog* OR radiology* OR rehabilitati* OR rheumatolog* OR surgery OR surgic* OR telemedic* OR telepatholog* OR teleradiolog* OR telerehabil* OR toxicolog* OR traumatolog* OR urolog* OR venereolog*) NEAR/4 (text* or narrati* or document* or summar* or note* or report*))
#2	TS=(abbrev* OR acronym* OR (short* NEAR/2 form*) OR ((single OR two OR three OR four) NEAR/2 (character or characters)) OR ellips* OR initialism*)
#3	DOP=(2018-01-01/2023-02-22)
#4[a]	#1 AND #2 AND #3
MEDLINE, Embase, and all EBMR[b] reviews	
#1	((medic* OR clinic* or allerg* OR androlog* OR anesthesiolog* OR anaesthesiolog* OR bariatric* OR biopharmaceutic* OR cardiolog* OR cardiovascul* OR chiropractic* OR cytopatholog* OR dental OR dentistr* OR dermatolog* OR dietetic* OR emergenc* OR endocrinolog* OR endodontic* OR ethnopharmacolog* OR forensic* OR gastroenterolog* OR genomic* OR geriatric* OR gerontolog* OR geroscienc* OR gynaecolog* OR gynecolog* OR haematolog* OR hematolog* OR immunolog* OR immunopatholog* OR microbiolog* OR midwife* OR nanomedic* OR neonatolog* OR nephrolog* OR neurolog* OR neuropatholog* OR neuropharmacolog* OR neuropsychiatr* OR neuroradiolog* OR neurosurg* OR neurotolog* OR nursing OR nutrigenomic* OR obstetric* OR occupational* OR oncolog* OR ophthalmolog* OR optometr* OR orthodontic* OR orthopedic* OR orthoptic* OR otolaryngolog* OR otolog* OR otorhinolaryngolog* OR paramedic* OR patholog* OR pediatric* OR perinatolog* OR periodontic* OR pharmacogenetic* OR pharmacolog* OR pneumolog* OR pneumonolog* OR podiatr* OR proctolog* OR prosthodontic* OR psychiatr* OR psychopharmacolog* OR pulmonolog* OR radiolog* OR radiology* OR rehabilitati* OR rheumatolog* OR surgery OR surgic* OR telemedic* OR telepatholog* OR teleradiolog* OR telerehabil* OR toxicolog* OR traumatolog* OR urolog* OR venereolog*) adj4 (text* or narrati* or document* or summar* or note* or report*)).mp.
#2	(abbrev* OR acronym* OR (short* adj2 form*) OR ((single OR two OR three OR four) adj2 (character or characters)) OR ellips* OR initialism*).mp.
#3	#1 AND #2
#4[a]	limit 3 to yr="2018-2023"

[a]Query number #4 is the final applied query, which incorporates all other queries #1, #2, and #3.

[b]EBMR: Evidence-Based Medicine Reviews.

XSL•FO
RenderX

Selection of Studies

For the scoping review, database records were imported into Citavi (Swiss Academic Software GmbH) [12], version 6. During the deduplication stage, these records were reviewed independently by 4 team members (AK, IM, LM, and PP). Each team member was assigned a portion of the records for deduplication: IM, LM, and PP each reviewed 20%, while AK reviewed 40%. After deduplication, AK conducted the merging and final review of all papers. In the screening stage, 4 team members were assigned portions of records to screen titles and abstracts for eligibility. Papers marked for exclusion were moved to a separate directory in Citavi. Each directory was reviewed for validation by at least one supervisor from the review team (MB, MK, or SS). In the eligibility stage, full-text papers were reviewed by 4 team members in pairs (AK and PP/IM and LM). If any team identified reports that matched the exclusion criteria, these were reviewed again by AK and one of the 3 supervisors. Any disagreements were discussed with additional team members until a consensus was reached.

Data Extraction

After the final inclusion decision, all eligible papers were read by all team members to validate the extracted information, which was recorded in a Google Spreadsheet (Alphabet Inc./Xxvi Holdings Inc.). The extracted data included the following information: publication title, authors, year of publication, short task description, setting (where the study was performed), study type, data set description, data set language, type of clinical narrative, scope of clinical narrative, study population, NLP methodologies, experimental setup, benchmarks or ground truth or gold standard, baselines, evaluation and validation, results, performance metrics, limitations, and conceptual description of short forms.

Data Synthesis

Publications were categorized based on (1) the type of short forms being processed and (2) the research methodology. For each type of short form, all processed data sets, data set descriptions, preprocessing methods, and examples provided in the full-text articles were synthesized to determine if any restrictions were placed on the data set before processing. This involved allocating specific restrictions for data set processing, from which short-form types could be derived, or determining if no restrictions were applied. For categorizing research methodologies, the 3 target objectives for short-form processing (identification, expansion, and disambiguation) were extracted from each publication, along with assigning multidimensional research methodologies to each objective. Additional objectives included differentiating between languages studied, noting NLP study recommendations, and identifying applied data sets, inspired by the scoping review by Kersloot et al [13]. Information was extracted from each article, and the characteristics were assigned accordingly. Overall occurrence rates were used for narrative description and summarization.

Results

Overview

Figure 1 illustrates the workflow of this scoping review. A total of 6579 records were identified that matched the search criteria via literature databases. After deduplication, 3878 records proceeded to the screening stage. Reviewing titles and abstracts reduced this number to 81 papers, which were then assessed for eligibility by reading their full texts. An additional manual citation search was conducted using BibliZap [14] based on the included studies, yielding 50 more full-text articles for eligibility screening. A thorough analysis during the eligibility stage resulted in 19 articles being included in the final analysis.

The results were divided into 5 sections, covering (1) types and processing of short forms, (2) applied data sets, (3) languages under investigation, (4) adherence to NLP recommendations, and (5) overall findings.

XSL•FO
RenderX

Figure 1. PRISMA diagram for systematic scoping review. EBMR: Evidence-Based Medicine Reviews; NLP: natural language processing; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; WOS: Web of Science.

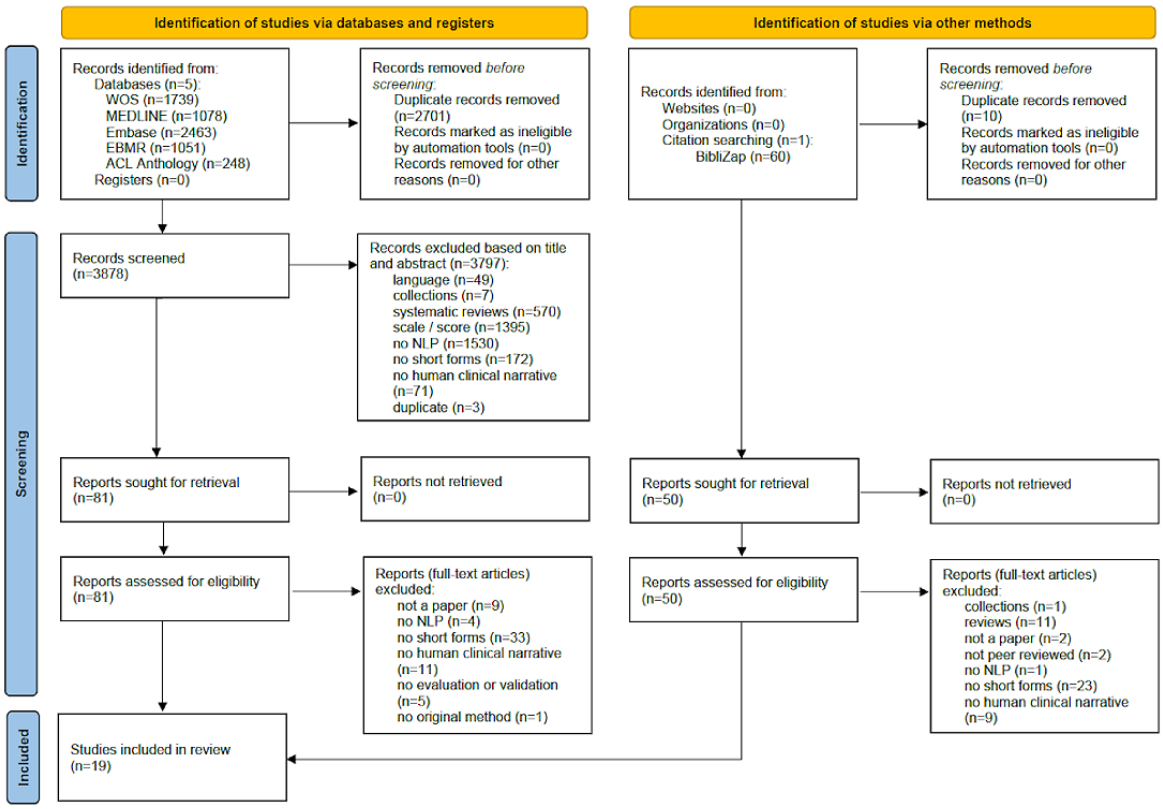

Types of Short Forms

The exact type of short-form content was not explicitly stated in all included articles. It had to be extracted through examples provided, data set evaluations and descriptions, or inference. Unfortunately, a clear typology of short forms does not exist, leading to varying definitions and interpretations of short-form types depending on the data set language used in each article.

In this scoping review, the term "short form" is used as a general category for all types of abbreviations. It includes not only classical acronyms or initialisms, such as "ECG," "5-FU," and "AIDS," but also all types of abbreviations, including single-character abbreviations such as "N" for "nerve" or "neoplasm." Additionally, short forms encompass unfinished words, also known as ad hoc abbreviations, such as "pat" for "patient" or "neonat" for "neonatology." In contrast to acronyms, these short forms are not capitalized and, depending on the language, may end with a dot, for example, "chron." (chronic) or "Dr." (doctor).

Acronyms, such as "RA" for "rheumatoid arthritis," are terms for shortened words or phrases that are written in uppercase letters and are defined based on their pronunciation. However, the inclusion of these acronyms in NLP tasks cannot always be guaranteed, as articles may define short forms in a manner that provides reliable results only for a subset of the data set. For example, only acronyms of 3 characters in length may be processed and evaluated.

Additionally, the occurrence of specialized abbreviations varies depending on the data set language. For example, single-character abbreviations, such as "A" for "arteria" or "aorta," are common in German clinical narratives but less frequent in other languages, such as English. Uppercase letters followed by a period are prevalent in German, such as "M." for "muscle" or "morbus" (disease), and "V." for "vein" or "vulnus" (wound). By contrast, English clinical narratives often use uppercase letters followed by "x," such as "Hx," "Dx," and "Fx" for medical history, diagnosis, and fracture.

Data Set Restrictions

Any restrictions placed on the data set before processing short forms are crucial for evaluating the effectiveness of methodologies for different types of short forms. For example, a restriction might involve analyzing only acronyms within a data set and excluding all other types of short forms.

As many as 7 of the 19 (37%) included papers [15-21] had no restrictions regarding the evaluation of short forms, meaning their investigation of identification, expansion, and disambiguation was not preemptively focused on specific types of short forms. By contrast, the majority (12/19, 63%) of papers [22-33] did impose restrictions on the type of short form, either through length restrictions, such as rule-based filters, or preprocessing guidelines. For example, regarding single-character abbreviations, 1 (5%) article [28] included only a few such abbreviations while excluding others. In 2 (11%) papers [23,26], the inclusion or exclusion of single-character abbreviations was unclear and could not be determined. Table 2 provides a full list of short-form types and the restrictions placed on data sets.

XSL•FO
RenderX

Table 2. Restrictions placed on the data set and selection of types of short forms before being processed.

Paper reference	Restrictions on short forms	Types of short forms selected
Adams et al [22]	Yes	Acronyms
Avdic et al [23]	Yes	Shortened words, abbreviations followed by a period character, physician-specific abbreviations, and common abbreviations
Egorov and Funkner [24]	Yes	Acronyms and abbreviations (at least half of the letters of a word need to be uppercase to be defined as an abbreviation)
Jaber and Martínez [30]	Yes	Acronyms and abbreviations (defined as a short form of a word or phrase)
Jaber and Martínez [25]	Yes	Acronyms and abbreviations (defined as a short form of a word or phrase)
Luo et al [15]	No	N/A[a]
Skreta et al [26]	Yes	Acronyms and abbreviations (without consideration of the casing of the word, all in lowercase)
Grossman et al [16]	No	N/A
Grossman Liu et al [17]	No	N/A
Wang et al [18]	No	N/A
Link et al [27]	Yes	Acronyms: 2 characters in length
Khaleghi et al [28]	Yes	Abbreviations (mostly user-specific word truncation): length less than 3 characters excluded, but hand-selection and inclusion of 2 single-character and 2 two-character abbreviations
Kashyap et al [29]	Yes	Acronyms: more than 2 characters in length
Joopudi et al [19]	No	N/A
Rajkomar et al [20]	No	N/A
Mykowiecka and Marciniak [31]	Yes	Ad hoc abbreviations
Báez et al [21]	No	N/A
Seneviratne et al [32]	Yes	Acronyms
Agrawal et al [33]	Yes	Acronyms

[a]N/A: not applicable.

Processing of Short Forms

Various methods are applied to process clinical narratives across the 3 steps: identification, expansion, and disambiguation of short forms. The first section provides an overview of all methodological classes and their functionalities found in the articles. The second section discusses the methods applied to each of the 3 processing steps.

Overview of Methodologies

The methods described in the selected articles can be roughly categorized into 8 classes: handcrafted rules, classical ML, DL, language modeling, human-in-the-loop, text representations, feature extraction, and software packages. Table 3 presents the classification of methodologies for processing short forms. Predominantly, classical ML techniques are used in combination with handcrafted rules for short-form processing, followed by applications of DL and language modeling. The human-in-the-loop approach is mentioned in only 1 article [18].

XSL•FO
RenderX

Table 3. Methods used for short-form processing.

Paper reference	Handcrafted rules	Classical machine learning	Deep learning	Language modeling	Human-in-the-loop	Text representations	Feature extractions	Software packages
Adams et al [22]	N/A[a]	N/A	✓	✓	N/A	✓	N/A	N/A
Avdic et al [23]	✓	✓	N/A	N/A	N/A	N/A	N/A	N/A
Egorov and Funkner [24]	✓	✓	N/A	N/A	N/A	✓	N/A	N/A
Jaber and Martínez [30]	N/A	✓	N/A	N/A	N/A	✓	✓	N/A
Jaber and Martínez [25]	N/A	N/A	✓	✓	N/A	✓	N/A	N/A
Luo et al [15]	✓	✓	✓	✓	N/A	✓	✓	✓
Skreta et al [26]	N/A	✓	N/A	N/A	N/A	✓	N/A	N/A
Grossman et al [16]	✓	N/A	N/A	N/A	N/A	N/A	✓	✓
Grossman Liu et al [17]	✓	✓	✓	✓	N/A	N/A	✓	✓
Wang et al [18]	N/A	✓	N/A	N/A	✓	N/A	✓	✓
Link et al [27]	✓	✓	N/A	N/A	N/A	✓	N/A	✓
Khaleghi et al [28]	✓	✓	N/A	N/A	N/A	✓	✓	N/A
Kashyap et al [29]	✓	✓	N/A	N/A	N/A	N/A	N/A	N/A
Joopudi et al [19]	✓	✓	✓	N/A	N/A	✓	✓	N/A
Rajkomar et al [20]	✓	✓	✓	✓	N/A	N/A	✓	N/A
Mykowiecka and Marciniak [31]	✓	N/A	✓	N/A	N/A	✓	✓	✓
Báez et al [21]	N/A	✓	✓	N/A	N/A	✓	N/A	✓
Seneviratne et al [32]	✓	N/A	✓	✓	N/A	✓	N/A	N/A
Agrawal et al [33]	✓	N/A	✓	✓	N/A	N/A	N/A	N/A

[a]N/A: not applicable.

Handcrafted rules involve the manual implementation of rule-based approaches, such as regular expressions, for specific tasks during text processing. For example, a regular expression might be used to identify acronyms in clinical texts.

Classical ML methods primarily involve supervised statistical modeling techniques, including support vector machines (SVMs), naïve Bayes classifiers, and decision tree classifiers, as well as unsupervised clustering approaches. These methods are used to classify data points into distinct categories, whether they are seen or new.

Statistical modeling techniques, such as conditional random fields [34], hidden Markov models [35], and logistic regression [36], use statistical functions to learn from the data set and generate predictions based on given dependencies. For instance, logistic regression performs binary classification by applying weights, a sigmoid function, and probabilistic interpretation to train a model for the classification task.

SVMs [37] map data points to a feature space to find the best hyperplane that separates different classes in the data set. Support vectors are the data points closest to this hyperplane. SVMs aim to maximize the margin between support vectors, which enhances the model's performance on unseen data and makes it more reliable and robust to irregularities or outliers in the data set.

Naive Bayes [38] is a probabilistic classifier based on Bayes' theorem, with the "naive" assumption that features are independent of each other. It calculates the probability of existing features to determine the likelihood that unseen features belong to predetermined groups.

Decision tree classifiers [39], including methods such as random forest [40] and gradient boosting [41], use a tree structure with nodes and branches to represent decisions based on features. Internal nodes represent decisions based on feature sets, branches represent possible outcomes of those decisions, and leaf nodes represent the final predictions. This tree structure facilitates traceable and interpretable classifications, allowing the decision path to be traced back to the root node, which aids in the explainability of the model.

DL methods, such as recurrent neural networks, convolutional neural networks (CNNs), sequence-based models, and transfer learning, use artificial neural networks with multiple layers of interconnected nodes. Four key features distinguish DL from classical ML: feature engineering, model complexity, data requirements, and interpretability. In classical ML, features are typically hand-selected or integrated through domain knowledge by engineers, and the trained models have simpler structures. With fewer features, classical ML models need less data and offer more interpretable classifications. By contrast, DL models automatically extract complex features from data sets, reducing the need for manual feature engineering. These models are more

XSL•FO
RenderX

complex due to their multiple layers of interconnected neurons and a larger number of parameters. As a result, DL models require larger data sets and more resources for training compared with classical ML models. Additionally, DL models are often considered "black boxes" and lack interpretability.

Language modeling is closely related to DL. It encompasses techniques that use pretrained language models created from large corpora of text using various NLP algorithms, such as Bidirectional Encoder Representations from Transformers (BERT) [42] and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) [43]. These models use methods such as masked language modeling and bidirectional encoders. Fine-tuning or adapting pretrained language models to specific problem domains with domain-specific data sets can further enhance their performance for NLP tasks, such as classification or prediction. The purpose of language modeling is to capture the statistical properties of language to improve the contextual understanding of texts. A recent advancement in this field is the development of large language models (LLMs), such as generative pretrained transformers (GPTs) [44,45]. A prominent example is ChatGPT [46], which encompasses GPT models developed by OpenAI that generate and synthesize text based on input prompts.

Human-in-the-loop methods [47] integrate human expertise or feedback into the ML process, applicable in both classical ML and DL settings. Examples are active learning, where human input guides the selection of labels or features during training; and model evaluation, where human feedback is used to interpret predictions and improve the model.

Text representations involve transforming textual input before data processing, significantly impacting overall performance depending on the chosen methodology. Examples are bag-of-words, word embeddings, and n-grams.

Bag-of-words represents text as a collection of individual words, disregarding grammar and word order. In this approach, only the presence of words is considered, while word sequences and sentence structure are ignored.

Word embeddings are vector representations of words in a continuous vector space, capturing the similarity between words based on their context and meaning. Each word is represented as a vector with a fixed dimension, and words with similar meanings or contexts are positioned close together in the vector space. This representation preserves semantic information and the context of the input.

N-grams are contiguous sequences of *n* items (or words) from a text. The input text is divided into consecutive sequences of *n* items, which can be words, characters, or similar structures. This representation scheme captures the word order and context within the text.

Feature extraction methods involve extracting or calculating features from text segments. Examples are string similarity metrics, such as those introduced by Levenshtein [48] or Needleman and Wunsch [49], part-of-speech tagging, and sectioning clinical narratives using section header information.

Software packages include implemented techniques and algorithms used by authors either as a baseline or integrated into their own workflow for short-form processing.

Methodologies for Identification, Expansion, and Disambiguation

Table 4 describes the methods for processing short forms, including identification, expansion, or disambiguation. Because of the variability in the objectives of each paper, not all papers addressing expansion also cover disambiguation, and not all papers focusing on disambiguation address identification or expansion of short-form content.

In our review, 14 of the 19 (74%) articles apply methods for identification, 12 (63%) for expansion, and 15 (79%) for disambiguation of short-form content. Of these, only 2 of 14 (14%) identification methods, 2 of 12 (17%) expansion methods, and 15 of 15 (100%) disambiguation methods provide performance metrics.

XSL•FO
RenderX

Table 4. Methodologies for the identification, expansion, and disambiguation of short forms.

Paper reference	Identification	Expansion	Disambiguation
Adams et al [22]	N/A[a]	N/A	Gaussian embedding drawn from both word and metadata prior densities, Bayesian skip-gram model, variational distribution over latent meaning cell, and deep learning (BERT[b], embeddings from language models)
Avdic et al [23]	Rules, lookup in custom lexical resource, baselines, support vector machine, random forest, tree tagger, naive Bayes, and variations of methodologies	N/A	N/A
Egorov and Funkner [24]	Rules and lookup in custom lexical resource	Bag-of-characters vectors, 2 TF-IDF[c] vectorizers	Random forest, logistic regression, extreme gradient boosting, stochastic gradient descent, CatBoost (gradient boosting on decision trees), and support vector classification
Jaber and Martínez [30]	N/A	N/A	Feature extraction, embedding vectors utilized from pretrained models, support vector machine, and naive Bayes skip-gram models
Jaber and Martínez [25]	N/A	N/A	BERT language models, no fine-tuning, context and expansions fed into models as input, token-type IDs, and binary mark IDs
Luo et al [15]	Rules, string similarity (Levenshtein [48]), regular expressions	Lookup and ranking with machine learning	Feature vectors, cosine distances, edit distances, similarity, and plus combinations
Skreta et al [26]	N/A	N/A	Convolutional neural network and max pool over time for local context, IDF[d]-weighted embedding average for global context, and combination of both outputs
Grossman et al [16]	CARD[e] framework [50] (combination of regular expressions and string similarity approaches)	Harmonization of different repositories into 1, cross-mapping synonymous record with filtering, string similarity, and MetaMap [51]	N/A
Grossman Liu et al [17]	CARD framework [50] (combination of regular expressions and string similarity approaches)	Harmonization of different repositories into 1, cross-mapping synonymous record with string similarity, rule-based text feature replacement, and meta inventory potential pair selection	Feed-forward dense neural network, gradient-boosted model, and transformer model
Wang et al [18]	Interactive learning, human annotator input (context + sense, feature + sense)	Dynamic feature representation combined with parameter estimation (logistic regression) = probability predictions	Instance selection is shown to the human annotator if the classifier is unable to select a sense, compared with random sampling, active learning, ReQuery-ReClassify expert method (Wang et al [52]), and feature engineering
Link et al [27]	Rule-based selection of notes with acronyms	Random forest for noisy labels, target sense classification, and word embeddings	Averaging predicted probabilities for final probability selection, prevalence estimation with probability cut-off, and baselines include most-frequent sense and knowledge-based method (Finley et al [53])
Khaleghi et al [28]	Regular expressions (filtering), lemmatization, tokenization, stemming, and string similarity (Levenshtein [48])	String similarity Levenshtein matrix, plus updated matrix, hierarchical agglomerative clustering, empirical selection of cut-off distance, heuristic clustering—sorting the hierarchical agglomerative clustering based on intercluster distances	N/A
Kashyap et al [29]	Regular expressions	Lookup in the PubMed database for text and possible expansions	Logistic regression model

XSL•FO
RenderX

Paper reference	Identification	Expansion	Disambiguation
Joopudi et al [19]	Regular expressions	String matching, feature vectors for each entity (bag-of-words, part-of-speech, clinical note, section, n-gram), assignment of each abbreviation to appropriate cluster, and proportional sampling of sentences from all clusters	Support vector machine, convolutional neural network, and baseline including most-frequent sense
Rajkomar et al [20]	Regular expressions	Needleman and Wunsch [49] global sequence alignment (token level)	Reverse substitution of abbreviations on web data, transfer learning, and chained inference technique to overcome domain shift
Mykowiecka and Marciniak [31]	Rules, part-of-speech tagger (Concraft2 [54]), and morphologic analyzer (Morfeuz2 [55])	Lookup in self-created resources generated through rules, clustering via the Chinese Whispers algorithm [56], and cosine similarity	Word2vec models, bidirectional long-short term memory, and baseline including most-frequent sense
Báez et al [21]	Named entity recognition via Flair framework (bidirectional long-short term memory-conditional random fields architecture) and application of clinical word embeddings	N/A	N/A
Seneviratne et al [32]	Rules	Lookup, replacement of acronyms in sentences with possible expansions to create sentence embeddings	Triplet networks, triplet loss, modeled as a binary classification problem, and baseline including acronym span prediction via pretrained language models (SciBERT and BioBERT)
Agrawal et al [33]	N/A	N/A	Generative pretrained transformer GPT[f]-3 as LLM[g]; LLM prompting for the resolution, postprocessing LLM answer with rules, weak supervision with data set filtering, and fine-tuned PubMedBERT [57] model for evaluation of output

[a]N/A: not applicable.

[b]BERT: bidirectional encoder representations from transformers.

[c]TF-IDF: term frequency-inverse document frequency.

[d]IDF: inverse document frequency.

[e]CARD: Clinical Abbreviation Recognition and Disambiguation.

[f]GPT: generative pretrained transformer.

[g]LLM: large language model.

Identification

The most common NLP approaches for short-form identification (12/14, 86%) include rules and regular expressions, lookups in lexical resources (either custom-created or freely available), and string similarity calculations. Lemmatization, tokenization, and stemming combined with rules were used only by Khaleghi et al [28]. By contrast, Wang et al [18] applied a supervised method, incorporating human annotation input to develop a logistic regression model, rather than using these NLP techniques. Similarly, Mykowiecka and Marciniak [31] used rules in combination with part-of-speech taggers and morphological analyzers to identify ad hoc abbreviations, such as unfinished words, in Polish clinical narratives.

Performance metrics for short-form identification were reported by only 2 studies. Avdic et al [23] conducted the identification and labeling of terms in their Serbian clinical corpus, where 12.9% of the words were abbreviations. By using normalization, stemming, cut-offs, and custom dictionaries, they improved the labeling of Serbian medical terms (including diagnoses, symptoms, medications, etc) to achieve an F_1-score of 0.908.

Báez et al [21] developed the Chilean Waiting List Corpus, an annotated resource comprising deidentified physician-authored referrals from various clinical specialties in Spanish. The annotations include mentions of findings, procedures, diseases, medications, body parts, and abbreviations. Using this annotated data set, they implemented a named entity recognition model with the Flair framework [58], which generated contextual embeddings for each word. Pretrained embeddings, trained on Spanish Wikipedia articles, were compared with embeddings enhanced with clinical data from unannotated parts of the corpus. An abbreviation detection model created with this approach achieved an F_1-score of 0.92 for both the base and enhanced pretrained embedding versions.

Expansion

The expansion of short forms can be achieved either through lookup operations or nonlookup methods. Lookup-based

XSL•FO
RenderX

expansion involves searching for a short form in a list, corpus, sense inventory, or dictionary to retrieve possible expansion candidates. These candidates can then be processed further using other methodologies for disambiguation. Nonlookup expansion methods include data mining techniques, end-to-end encoder-decoder models, text generation workflows, and active learning approaches. Expansion of short forms is accomplished through lookup operations in 8 of 12 (67%) articles [16,17,19,24,27,29,31,32]. These methods involve searching for short forms in lists, corpora, or dictionaries to find possible expansions. By contrast, 3 of 12 (25%) articles used nonlookup methods [18,20,28], which do not rely on predefined lists or dictionaries. Additionally, 1 article (1/12, 8%) by Luo et al [15] presented results from the n2c2 challenge, where participating teams used a mix of lookup operations, such as semantic-type classifiers, vocabulary classifiers, and similarity scores, as well as nonlookup methods, such as edit distance calculations with word embeddings.

Most of these approaches incorporate a variety of methodologies, including feature vector creation and representation (such as a bag of words, part of speech tagging, clinical notes, section information, and n-grams), logistic regression, random forest, word embeddings, retrieve and rank approaches, and clustering. Notably, Rajkomar et al [20] applied Needleman-Wunsch global sequence alignment, a method originally developed for nucleotide and protein sequence alignment, to clinical texts and abbreviation expansion.

Mykowiecka and Marciniak [31] generated expansion candidates either through a data-driven rule-based approach derived from the clinical texts or by applying the Chinese Whispers algorithm [56]. This randomized graph-clustering method clusters occurrences of abbreviations and identifies expansion candidates based on these clusters and cosine similarity.

Grossman et al [16] developed an extensive database of medical short forms, termed the “metathesaurus” by consolidating various repositories. Their processing techniques included lexical normalization, concept identification using MetaMap [59], and cross-mapping of synonymous terms through string similarity. Coverage calculations indicated that their resource achieved very high micro coverage, with 94.3% short-form coverage and 99.6% sense coverage, significantly outperforming the UMLS (Unified Medical Language System) LRABR acronym-abbreviation table, which covered only 74.8% of short forms.

Khaleghi et al [28] utilized similarity matrix calculations using Levenshtein distance [48,60], incorporating rule-based adjustments based on the type of short form processed. The results were then input into k-means partitional clustering, with empirical cut-off distances set and clusters sorted based on intercluster distances. This approach achieved an abbreviation detection accuracy of 90% and a typo detection accuracy of 90.6%.

Disambiguation

The disambiguation approaches are predominantly based on DL methods, which account for 8 of the 15 (53%) studies, including CNNs, feed-forward neural networks, and transformers. Classical ML methods make up 3 of the 15 studies (20%), with SVMs being a notable example. Additionally, embedding representations, such as Gaussian and cosine distance, are used in 6 of the 15 (40%) studies. Logistic regression was used in 2 of the 15 (13%) cases. Other methodologies include Bayesian skip-gram models, variational distributions, gradient boosting, and chained inference combined with transfer learning, among other techniques.

According to performance measures reported by Egorov and Funkner [24], SVM emerged as the best model for short-form expansion, achieving an F_1-score of 0.937. Their study compared SVM with other methods, including random forest, logistic regression, gradient boosting, and stochastic gradient descent classifiers, with the classical ML method outperforming the others in this context.

Adams et al [22] utilized contextualized word representations derived from local context and metadata, combined with predefined inventories of short-form expansions. They used Gaussian embeddings drawn jointly from word and metadata prior densities, and a Bayesian skip-gram model to process surrounding words. This approach resulted in a variational distribution over the latent meaning cell, surpassing the performance of DL strategies. It achieved a weighted mean F_1-score across 5 pretraining runs of 0.69 for MIMIC-III, 0.57 for the CUIMC (Columbia University Irving Medical Center) data set, and 0.51 for the Clinical Abbreviation Sense Inventory (CASI) [9] data set from the University of Minnesota.

Joopudi et al [19] combined string matching, a custom word-sense inventory, and deep-learning methods with feature vectors for SVM and CNN. Among these, CNN with local features outperformed SVM by incorporating clinical narrative metadata and section information, achieving a micro-averaged accuracy of 0.979.

Jaber and Martínez [30] analyzed a subset of 13 acronyms from the CASI data set for acronym disambiguation using SVM and naive Bayes skip-gram models. During the training phase, feature extraction and embedding vectors were utilized alongside pretrained skip-gram models from PubMed Central, Wikipedia, and PubMed abstracts. SVM outperformed naive Bayes, achieving an average accuracy of 0.97 compared with 0.93.

In a follow-up study, Jaber and Martínez [25] utilized a masked language modeling approach with 3 pretrained BERT [42] language models, without fine-tuning for the specific problem domain. They incorporated the context and expansions of each short form as input to the model. This approach achieved an accuracy of 0.991, surpassing the results reported by Adams et al [22] and Joopudi et al [19] on the CASI data set.

Luo et al [15] detailed the methodologies of the top 10 performing teams in the shared task on clinical concept normalization. For the challenging category of single-character abbreviations, 2 teams distinguished themselves by using DL with contextual embeddings. Despite achieving a maximum accuracy of only 0.35, their performance surpassed that of most other teams.

XSL•FO
RenderX

Skreta et al [26] used UMLS term embeddings combined with reverse substitution of terms (replacing expansions with their corresponding abbreviations) in the MIMIC-III data set to generate training examples. By leveraging concept hierarchies from UMLS to augment training sets and searching for related concepts, they adapted the global context in narratives using Euclidean distance to develop an abbreviation disambiguation pipeline with CNNs. By integrating concept hierarchies during pretraining, augmenting with associated medical concepts extracted from the embedding space, and considering the global context of clinical narratives, Skreta et al [26] achieved an accuracy of 0.841 on the CASI data set with their CNN-based abbreviation disambiguation pipeline.

Grossman Liu et al [17] continued the work of Grossman et al [16] by applying DL methodologies to cross-map short forms and develop a metathesaurus of clinical short forms. On clinician-labeled data, their gradient-boosted model, BERT model, and an ensemble approach demonstrated similar performance, with the ensemble achieving an F_1-score of 0.814.

Wang et al [18] utilized an active learning algorithm, incorporating human annotator input to enhance acronym disambiguation. This approach achieved an area under the learning curve score of 0.852 on the CASI data set.

Link et al [27] assessed their semisupervised ensemble ML algorithm (CASEml) on the acronyms “RA,” “MS,” and “MI” using Veterans Affairs EHR data. Their approach, which integrated a visit-level random forest with embeddings-based context representation, achieved accuracy metrics of 0.947 for “RA,” 0.911 for “MS,” and 0.706 for “MI.”

Kashyap et al [29] developed CLASSE GATOR (Clinical Acronym Sense Disambiguator) for acronym disambiguation. The tool, which utilizes full-text research articles from PubMed Central to detect and extract acronym-expansion pairs, trained a logistic regression model. The system achieved an average accuracy of 0.879 in predictive performance.

Rajkomar et al [20] developed a single translation model for the detection, expansion, and disambiguation of clinical acronyms, and evaluated it on multiple data sets. The model achieved accuracies of 0.921 on the CASI data set, 0.957 on the MIMIC-III data set, and 0.965 on the Informatics for Integrating Biology and the Bedside (i2b2) 2014 data set.

Mykowiecka and Marciniak [31] used a bidirectional long short-term memory network architecture for disambiguating ad hoc abbreviations, achieving an F_1-score of 0.726 with 10-fold cross-validation. Further variations in the lists of possible expansion candidates improved the F_1-score to 0.968.

Seneviratne et al [32] adapted embeddings-based approaches for acronym disambiguation by learning sentence embeddings to capture semantic differences. They applied triplet networks and triplet loss methods, drawing inspiration from Siamese Networks [61] for image recognition and triplet neural networks [62] for predicting protein gene ontology. To create the embeddings, acronyms were identified using rules, and based on the CASI data set, abbreviations were replaced with their possible long forms. This approach was modeled as a binary classification problem, where the method determines the correctness of an acronym by comparing the input with the trained embeddings, achieving an F_1-score of 0.87.

Agrawal et al [33] tested the disambiguation of acronyms from the CASI data set using GPT-3, a GPT model developed by OpenAI. This LLM utilizes contextual understanding from the input prompt to generate appropriate responses. In this case, no specific examples for acronym resolution were provided; instead, the model was given only the context in which the acronym appeared, followed by a request to expand the acronym based on that context. This method achieved 0.86 in accuracy and 0.69 in macro F_1-score. With additional data set filtering and fine-tuning of PubMedBERT [57] to distill the GPT-3 model into a smaller, more manageable version, evaluations on MIMIC—restricted for GPT-3 due to data-use agreements—were conducted. The combined GPT-3 and PubMedBERT approach achieved an accuracy of 0.90 for the CASI data set and 0.78 for MIMIC.

Applied Data Sets

For the development and assessment of methodologies, most studies utilize a diverse range of data sets, as detailed in Table 5.

XSL•FO
RenderX

Table 5. Applied data sets for the creation and/or evaluation, and language of clinical narratives being processed, are listed.

Paper reference	Data set for creation or evaluation[a]	Language
Adams et al [22]	*MIMIC-III*[b], University of Minnesota (CASI[c])[d], Columbia University intensive care unit/critical care unit reverse substitution[d,e], MIMIC reverse substitution[e]	English
Avdic et al [23]	*Unstructured EHR*[f]*medical reports* and custom dictionary of medical (eg, diagnoses, medications, Latin terms) and nonmedical terms (stop words, proper nouns)	Serbian
Egorov and Funkner [24]	*Unstructured EHR medical recommendations*, Leo Tolstoy's novel "War and Peace," and an encyclopedic dictionary of medical Russian terms	Russian
Jaber and Martínez [30]	University of Minnesota (CASI)[d]	English
Jaber and Martínez [25]	University of Minnesota (CASI)[d]*and unstructured EHR admission notes/inpatient consult notes/discharge summaries*	English
Luo et al [15]	*i2b2* [g] *2010*	English
Skreta et al [26]	*MIMIC-III*, University of Minnesota (CASI)[d], and *i2b2 2010*	English
Grossman et al [16] Grossman Liu et al [17]	UMLS[h]-LRABR[i], ADAM[i,j], Berman's abbreviations[i], Wikipedia[i], Vanderbilt University inventories from EHR from sign-out and discharge notes[d], Stetson[d], Columbia OBGYN[d], and *MIMIC-III*	English
Wang et al [18]	Medical Subject Headings abbreviations via MEDLINE abstracts[i], University of Minnesota (CASI)[d], and Clinical Abbreviations from Vanderbilt University[d]	English
Link et al [27]	*Unstructured EHR clinical notes*	English
Khaleghi et al [28]	*Unstructured EHR surgical notes*	English
Kashyap et al [29]	*MIMIC-III*, PubMed Central, and University of Minnesota (CASI)[d]	English
Joopudi et al [19]	University of Minnesota (CASI)[d] *and unstructured EHR longitudinal patient records*	English
Rajkomar et al [20]	University of Minnesota (CASI)[d], *MIMIC-III*, synthetic snippets, *i2b2 2014*, Clinical Abbreviations from Vanderbilt University[d], Sign-out note abbreviations[i], Beth Israel Deaconess Medical Center abbreviations[i], and Wikipedia[i]	English
Mykowiecka and Marciniak [31]	*Unstructured EHR clinical notes (interrogation, examination, and recommendations)*	Polish
Báez et al [21]	*Chilean Waiting List Corpus: unstructured EHR clinical notes (referrals)*	Spanish
Seneviratne et al [32]	University of Minnesota (CASI)[d]	English
Agrawal et al [33]	University of Minnesota (CASI)[d] and MIMIC-III reverse substitution[e]	English

[a]Clinical narrative data sets are marked in *italics*.
[b]MIMIC: Medical Information Mart for Intensive Care.
[c]CASI: Clinical Abbreviation Sense Inventory.
[d]A data set that consists of a sense inventory in combination with short forms in context.
[e]Reverse substitution: replacement of long forms with their short forms and labeling it with the original target label.
[f]EHR: electronic health record.
[g]i2b2: Informatics for Integrating Biology and the Bedside.
[h]UMLS: Unified Medical Language System.
[i]Short-form–specific sense inventories.
[j]ADAM: Another Database of Abbreviations in MEDLINE.

Grossman et al [16] and Grossman Liu et al [17] utilized 9 data sets to create a comprehensive deep database of medical abbreviations. Similarly, Rajkomar et al [20] used 8 data sets for clinical abbreviation disambiguation. Additionally, 7 of 19 (37%) studies used unstructured EHRs from their institutions, which are not publicly accessible due to privacy concerns. Available clinical narratives constituted a significant portion of the listed data sets. Notably, 10 of 19 (53%) studies used the CASI [9] from the University of Minnesota, which includes a sense inventory, document-level metadata, and context information for ambiguous clinical abbreviations. The CASI data set incorporates several data sets that are also used independently by other studies reviewed here, including ADAM (Another Database of Abbreviations in MEDLINE) [63]; the UMLS Metathesaurus [64]; and Stedman's Medical Abbreviations, Acronyms and Symbols [65]. Similarly, 6 of the

XSL•FO
RenderX

19 (32%) papers utilized the MIMIC-III [7], which is a deidentified intensive care data set covering over 40,000 patients. This data set includes intensive care notes, tests, orders, billing and code information, demographics, and reports for patients attended to by hospital staff between 2001 and 2012.

Languages Under Investigation

Examining the language distribution of data sets used across the 14 eligible papers, 14 of the 19 (74%) data sets were in English. The remaining data sets were in other languages: Serbian [23], Russian [24], Polish [31], and Spanish [21]. Given that these are considered low-resource languages for clinical NLP, custom dictionaries and sense inventories for short-form content were necessary to aid in expansion and disambiguation. These inventories were then applied to process texts for the predetermined tasks of short-form identification, expansion, and disambiguation.

NLP Recommendations

To assess the quality and reproducibility of the included papers, NLP recommendations across 7 categories were analyzed: source code availability, linking to external data sets, descriptions of internal data sets, application of performance metrics, provision of error analysis, inclusion of confusion matrices, and execution of external validation. These categories were inspired by the NLP recommendations from Kersloot et al [13], established for future studies during a systematic scoping review on NLP algorithms for mapping clinical text fragments onto ontology concepts. The classification of each included paper for short-form processing according to these categories is summarized in Table 6.

Only 3 of 19 (16%) papers met all the recommendations. By contrast, 16 of 19 (84%) only partially fulfilled the criteria for each category. Specifically, of the 19 papers, 10 (53%) did not provide the source code for their methodology, 2 (11%) did not link to the external data sets used, 15 (79%) did not include a confusion matrix for error analysis, 10 (53%) did not conduct an error analysis, and 9 (47%) did not perform external validation.

Table 6. Analysis and classification of included papers according to their fulfillment of natural language processing recommendations for identification, expansion, and disambiguation of short forms.[a]

Paper reference	Overall fulfillment	Source code	External data set linked	Internal data set described	Performance metrics	Error analysis	Confusion matrix	External validation
Adams et al [22]	Partially	Yes	Yes	Partially	Yes	No	No	Yes
Avdic et al [23]	Partially	No	No	Yes	Yes	Yes	No	No
Egorov and Funkner [24]	Partially	No	No	Partially	Yes	No	No	No
Jaber and Martínez [30]	Partially	No	Yes	N/A[b]	Yes	No	No	No
Jaber and Martínez [25]	Partially	No	Yes	Yes	Yes	Yes	No	No
Luo et al [15]	Yes	Yes	Yes	N/A	Yes	Yes	N/A	Yes
Skreta et al [26]	Yes	Yes	Yes	N/A	Yes	Yes	Yes	Yes
Grossman et al [16]	Partially	Yes	Yes	N/A	Yes	No	No	Yes
Grossman Liu et al [17]	Partially	Yes	Yes	N/A	Yes	No	No	Yes
Wang et al [18]	Partially	No	Yes	Yes	Yes	Yes	No	No
Link et al [27]	Yes	Yes	N/A	Yes	Yes	Yes	Yes	Yes
Khaleghi et al [28]	Partially	No	N/A	Yes	Yes	No	Partially	Yes
Kashyap et al [29]	Partially	No	Yes	Yes	Yes	No	No	Yes
Joopudi et al [19]	Partially	No	Yes	Yes	Yes	Yes	No	Yes
Rajkomar et al [20]	Partially	Yes	Yes	Yes	Yes	Yes	No	Yes
Mykowiecka and Marciniak [31]	Partially	No	N/A	Yes	Yes	Yes	No	No
Báez et al [21]	Partially	Yes	Yes	Yes	Yes	No	No	No
Seneviratne et al [32]	Partially	Yes	Yes	N/A	Yes	No	No	No
Agrawal et al [33]	Partially	No	Yes	N/A	Yes	No	No	No

[a]Four possible classifications: yes, no, partially, and N/A.

[b]N/A: not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

Research on abbreviations has not uniformly covered all types, and methodologies have yielded varying results for different subtypes. This disparity is partly due to the limitations of existing short-form repositories, which often focus on specific types of abbreviations. For instance, the CASI from the University of Minnesota includes only acronyms and omits single-character abbreviations. Developing a comprehensive sense inventory for single-character abbreviations is challenging and may require utilizing openly available data sets and databases. The higher ambiguity associated with single-character abbreviations could also explain why they are frequently excluded from studies.

It appears that clinical narratives in different languages have varying distributions of short forms and their subtypes [4]. The included articles indicate that short forms can include the following subtypes: abbreviations, acronyms, abbreviations followed by a period, single-character abbreviations, among others. Our systematic scoping review found that all these subtypes were addressed in at least one of the articles reviewed. However, not all articles clearly specify the types of short-form content they processed. Nevertheless, it is often possible to infer the subtypes covered, based on the methods described or examples from the data sets.

The preemptive exclusion of certain types of short forms from data sets introduces a bias that becomes apparent only when the articles are carefully compared individually.

For identifying and expanding short forms in clinical narratives, the most commonly used approaches were rules, string similarity, and lookups in lexical resources. These methods have a low barrier to entry, being relatively easy to implement and test. By contrast, DL and ML approaches require significantly more resources, including graphics cards, high computing power, and large data sets. Our review indicates a strong preference for DL, particularly for disambiguation tasks, as it heavily relies on contextual detection.

Limitations and Future Research Directions

A limitation of this systematic scoping review was the restriction to the past 5 publication years, which may have excluded important studies published before or after this period. This decision was driven by practical considerations, such as managing the volume of literature within the constraints of available personal resources.

Limitations in effectively processing all types of short forms are primarily due to the data sets and resources used to develop the processing methodologies. Key factors include data accessibility and the resources required for implementing short-form processing techniques. These limitations arise from the need for high-quality data collected from various sites or institutions for each data set language and the establishment of annotation workflows for creating comprehensive short-form sense inventories. These processes would enable the semantic recording of different documentation styles and varied contexts for each short form, thereby enhancing the processing algorithms to better recognize, expand, and disambiguate all types of abbreviations. Given the particularities of clinical language and short forms in each language, bridging the gap in processing between languages requires higher-quality resources. These resources should be generated and made accessible to the research community to improve cross-linguistic short-form processing [66].

Additionally, the articles examined in this scoping review only marginally reflect the impact of LLMs, such as those popularized by ChatGPT. Only 1 article in the review applied LLMs for short-form processing. This limited representation is partly due to the review's timeframe, which concluded just a few months after the release of the GPT-3.5 model. Since then, numerous studies have utilized LLMs for clinical short-form processing tasks. However, these results offer only a snapshot amid the rapidly evolving technological landscape. Currently, LLMs appear to hold significant potential for clinical short-form processing, although their precision is still challenged by issues such as hallucinations, which are difficult to control. Additionally, many of the most effective models are proprietary and cloud-based, which limits their use for processing sensitive data. Therefore, the methodologies analyzed and discussed in this scoping review should not be considered obsolete. We plan to update this review once LLM technology has matured and a sufficient number of new studies—ideally utilizing the same data sets as those reported here—have been published.

Conclusions

Short-form expressions, such as acronyms and other abbreviations, are distinctive elements found in narratives written by clinicians and stored in EHRs. To gain an overview of methods for processing these short-form expressions in clinical texts, we conducted a systematic scoping review of peer-reviewed articles. Our review found that classical ML and DL methodologies demonstrated the best performance for short-form disambiguation tasks, while rule-based and string similarity matching approaches were more commonly used for short-form identification and expansion. The methodologies applied to different short-form types and languages varied, and recommendations for NLP studies were only partially followed. Future research should focus on improving the quality and reproducibility of investigations by providing comprehensive details, including links to used resources and a more detailed description of the short-form content being studied.

Acknowledgments

This work was partially supported by the German Federal Ministry of Education and Research "Bundesministerium für Bildung und Forschung" (BMBF) within the following projects: the Junior Research Group "Clinical Text Analytics: Methods for NLP

XSL•FO
RenderX

on German Texts" (Nachwuchsgruppe "Klinische Textanalytik: Methoden für NLP an deutschen Texten") (DE.xt) and the Data Integration for Future Medicine (DIFUTURE) under grants 01ZZ2009A and 01ZZ2304A, respectively.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Scoping Review) checklist.
[PDF File (Adobe PDF File), 101 KB - jmir_v26i1e57852_app1.pdf]

References

1. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
2. Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: a survey of internal medicine residents and program directors. Arch Intern Med 2010 Feb 22;170(4):377-380. [doi: 10.1001/archinternmed.2009.534] [Medline: 20177042]
3. Christino MA, Matson AP, Fischer SA, Reinert SE, Digiovanni CW, Fadale PD. Paperwork versus patient care: a nationwide survey of residents' perceptions of clinical documentation requirements and patient care. J Grad Med Educ 2013 Dec;5(4):600-604 [FREE Full text] [doi: 10.4300/JGME-D-12-00377.1] [Medline: 24455008]
4. Schwarz CM, Hoffmann M, Smolle C, Eiber M, Stoiser B, Pregartner G, et al. Structure, content, unsafe abbreviations, and completeness of discharge summaries: a retrospective analysis in a university hospital in Austria. J Eval Clin Pract 2021 Dec 09;27(6):1243-1251. [doi: 10.1111/jep.13533] [Medline: 33421263]
5. Awan S, Abid S, Tariq M, Zubairi ABS, Kamal A, Arshad S, et al. Use of medical abbreviations and acronyms: knowledge among medical students and postgraduates. Postgrad Med J 2016 Dec;92(1094):721-725. [doi: 10.1136/postgradmedj-2016-134086] [Medline: 27281817]
6. Chemali M, Hibbert EJ, Sheen A. General practitioner understanding of abbreviations used in hospital discharge letters. Med J Aust 2015 Aug 03;203(3):147, 147e.1-147, 147e.4. [doi: 10.5694/mja15.00224] [Medline: 26224188]
7. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
9. Moon S, Pakhomov S, Melton G. Clinical abbreviation sense inventory. University Digital Conservancy. 2012 Oct 31. URL: http://conservancy.umn.edu/handle/11299/137703 [accessed 2023-09-22]
10. Clarivate - data, insights and analytics for the innovation lifecycle. Clarivate. URL: https://clarivate.com/ [accessed 2024-05-24]
11. Ovid. URL: https://ovidsp.ovid.com/ [accessed 2024-05-24]
12. Swiss Academic Software GmbH. Citavi - Literaturverwaltung und Wissensorganisation. CITAVI. URL: https://www.citavi.com/de [accessed 2024-02-04]
13. Kersloot MG, van Putten FJP, Abu-Hanna A, Cornet R, Arts DL. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semantics 2020 Nov 16;11(1):14-21 [FREE Full text] [doi: 10.1186/s13326-020-00231-z] [Medline: 33198814]
14. BibliZap. URL: https://app.biblizap.org/ [accessed 2024-05-20]
15. Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc 2020 Oct 01;27(10):1529-1537 [FREE Full text] [doi: 10.1093/jamia/ocaa106] [Medline: 32968800]
16. Grossman LV, Mitchell EG, Hripcsak G, Weng C, Vawdrey DK. A method for harmonization of clinical abbreviation and acronym sense inventories. J Biomed Inform 2018 Dec;88:62-69 [FREE Full text] [doi: 10.1016/j.jbi.2018.11.004] [Medline: 30414475]
17. Grossman Liu L, Grossman RH, Mitchell EG, Weng C, Natarajan K, Hripcsak G, et al. A deep database of medical abbreviations and acronyms for natural language processing. Sci Data 2021 Jun 02;8(1):149 [FREE Full text] [doi: 10.1038/s41597-021-00929-4] [Medline: 34078918]
18. Wang Y, Zheng K, Xu H, Mei Q. Interactive medical word sense disambiguation through informed learning. J Am Med Inform Assoc 2018 Jul 01;25(7):800-808 [FREE Full text] [doi: 10.1093/jamia/ocy013] [Medline: 29584896]
19. Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. J Biomed Inform 2018 Oct;86:71-78 [FREE Full text] [doi: 10.1016/j.jbi.2018.07.025] [Medline: 30118854]

XSL•FO
RenderX

20. Rajkomar A, Loreaux E, Liu Y, Kemp J, Li B, Chen M, et al. Deciphering clinical abbreviations with a privacy protecting machine learning system. Nat Commun 2022 Dec 02;13(1):7456-7414 [FREE Full text] [doi: 10.1038/s41467-022-35007-9] [Medline: 36460656]
21. Báez P, Villena F, Rojas M, Durán M, Dunstan J. The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. New York, NY: Association for Computational Linguistics; 2020 Nov 19 Presented at: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics; November 19, 2020; Online p. 291-300. [doi: 10.18653/v1/2020.clinicalnlp-1.32]
22. Adams G, Ketenci M, Bhave S, Perotte A, Elhadad N. Zero-shot clinical acronym expansion via latent meaning cells. Proc Mach Learn Res 2020 Dec;136:12-40 [FREE Full text] [Medline: 34790898]
23. Avdic A, Marovac U, Jankovic D. Automated labeling of terms in medical reports in Serbian. Turk J Elec Eng & Comp Sci 2020 Nov 30;28(6):3285-3303. [doi: 10.3906/elk-2002-9]
24. Egorov M, Funkner A. Automatic extraction and decryption of abbreviations from domain-specific texts. Stud Health Technol Inform 2021 Oct 27;285:281-284. [doi: 10.3233/SHTI210615] [Medline: 34734887]
25. Jaber A, Martínez P. Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques. Methods Inf Med 2022 Jun;61(S 01):e28-e34 [FREE Full text] [doi: 10.1055/s-0042-1742388] [Medline: 35104909]
26. Skreta M, Arbabi A, Wang J, Drysdale E, Kelly J, Singh D, et al. Automatically disambiguating medical acronyms with ontology-aware deep learning. Nat Commun 2021 Sep 07;12(1):5319-5310 [FREE Full text] [doi: 10.1038/s41467-021-25578-4] [Medline: 34493718]
27. Link NB, Huang S, Cai T, Sun J, Dahal K, Costa L, Million Veteran Program. Binary acronym disambiguation in clinical notes from electronic health records with an application in computational phenotyping. Int J Med Inform 2022 Apr 01;162:104753. [doi: 10.1016/j.ijmedinf.2022.104753] [Medline: 35405530]
28. Khaleghi T, Murat A, Arslanturk S, Davies E. Automated surgical term clustering: a text mining approach for unstructured textual surgery descriptions. IEEE J Biomed Health Inform 2020 Jul;24(7):2107-2118. [doi: 10.1109/jbhi.2019.2956973]
29. Kashyap A, Burris H, Callison-Burch C, Boland MR. The CLASSE GATOR (CLinical Acronym SenSE disambiGuATOR): a method for predicting acronym sense from neonatal clinical notes. Int J Med Inform 2020 May;137:104101. [doi: 10.1016/j.ijmedinf.2020.104101] [Medline: 32088556]
30. Jaber A, Martínez P. Disambiguating Clinical Abbreviations using Pre-trained Word Embeddings. Disambiguating Clinical Abbreviations using Pre-trained Word Embeddings: Proc 14th Int Jt Conf Biomed Eng Syst Technol Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications; 2021 Presented at: 14th International Conference on Health Informatics; 11-13 Feb 2021; Online p. 501-508. [doi: 10.5220/0010256105010508]
31. Mykowiecka A, Marciniak M. Experiments with ad hoc ambiguous abbreviation expansion. Stroudsburg, PA: Association for Computational Linguistics; 2019 Presented at: International Workshop on Health Text Mining and Information Analysis (LOUHI 2019); November 3, 2019; Hong Kong p. 44-52. [doi: 10.18653/v1/d19-6207]
32. Seneviratne S, Daskalaki E, Lenskiy A, Suominen H. m-Networks: adapting the triplet networks for acronym disambiguation. Stroudsburg, PA: Association for Computational Linguistics; 2022 Jul 14 Presented at: 4th Clinical Natural Language Processing Workshop; July 14, 2022; Seattle, WA p. 21-29. [doi: 10.18653/v1/2022.clinicalnlp-1.3]
33. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. Stroudsburg, PA: Association for Computational Linguistics; 2022 Dec 07 Presented at: Conference on Empirical Methods in Natural Language Processing; December 7, 2022; Abu Dhabi, United Arab Emirates p. 1998-2022. [doi: 10.18653/v1/2022.emnlp-main.130]
34. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Cambridge, MA: Morgan Kaufmann Publishers Inc; 2001 Jun 28 Presented at: International Conference on Machine Learning; June 28, 2001; San Francisco, CA p. 282-289. [doi: 10.1145/1015330.1015422]
35. Markov AA. An example of statistical investigation of the text concerning the connection of samples in chains. Sci Context 2007 Jan 04;19(4):591-600. [doi: 10.1017/s0269889706001074]
36. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014;8(85):14-2830 [FREE Full text] [doi: 10.3389/fninf.2014.00014] [Medline: 24600388]
37. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
38. Zhang J, Shen D, Zhou G, Su J, Tan C. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. J Biomed Inform 2004 Dec;37(6):411-422 [FREE Full text] [doi: 10.1016/j.jbi.2004.08.005] [Medline: 15542015]
39. Rokach L, Maimon O. Data Mining with Decision Trees: Theroy and Applications. Hackensack, NJ: World Scientific Publishing Co., Inc; 2008.
40. Breiman L. Random forests. Machine Learning 2001 Oct 01;45(1):5-32. [doi: 10.1023/A:1010933404324]
41. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis 2002 Feb;38(4):367-378. [doi: 10.1016/S0167-9473(01)00065-2]
42. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Association for Computational Linguistics. Minneapolis, MN: Association for Computational Linguistics; 2019 Jun 02

XSL•FO
RenderX

Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/2024.naacl-short]
43. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. OpenReview.net. 2019. URL: https://openreview.net/forum?id=r1xMH1BtvB [accessed 2024-07-26]
44. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. Semantic Scholar. 2019. URL: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe [accessed 2024-02-18]
45. Radford A, Narasimhan K. Improving language understanding by generative pre-training. Semantic Scholar. 2018. URL: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035 [accessed 2024-02-18]
46. OpenAI. ChatGPT. URL: https://openai.com/chatgpt [accessed 2024-02-18]
47. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal ?. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 2023;56(4):3005-3054. [doi: 10.1007/s10462-022-10246-w]
48. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl Soviet Union 1966 Feb;10(8):707-710 [FREE Full text]
49. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970 Mar;48(3):443-453. [doi: 10.1016/0022-2836(70)90057-4] [Medline: 5420325]
50. Wu Y, Denny JC, Trent Rosenbloom S, Miller RA, Giuse DA, Wang L, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). J Am Med Inform Assoc 2017 Apr 01;24(e1):e79-e86 [FREE Full text] [doi: 10.1093/jamia/ocw109] [Medline: 27539197]
51. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
52. Wang Y, Zheng K, Xu H, Mei Q. Clinical word sense disambiguation with interactive search and classification. AMIA Annu Symp Proc 2016;2016:2062-2071 [FREE Full text] [Medline: 28269966]
53. Finley GP, Pakhomov SVS, McEwan R, Melton GB. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. AMIA Annu Symp Proc 2016;2016:560-569 [FREE Full text] [Medline: 28269852]
54. Waszczuk J. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. 2012 Dec Presented at: COLING; December 8-15, 2012; Mumbai, India p. 2789-2804 URL: https://aclanthology.org/C12-1170
55. Woliński M. Morfeusz reloaded. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014 May Presented at: The Ninth International Conference on Language Resources and Evaluation (LREC'14); May 26-31, 2014; Reykjavik, Iceland p. 1106-1111 URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/768_Paper.pdf [doi: 10.1007/s10579-005-2692-5]
56. Biemann C. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics; 2006 Jun 05 Presented at: The First Workshop on Graph Based Methods for Natural Language Processing; June 5, 2006; New York City p. 73-80. [doi: 10.3115/1654758.1654774]
57. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
58. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. Stroudsburg, PA: Association for Computational Linguistics; 2018 Aug 20 Presented at: International Conference on Computational Linguistics; August 20, 2018; Santa Fe, NM p. 1638-1649 URL: https://aclanthology.org/C18-1139
59. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
60. Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE Trans Pattern Anal Mach Intell 2007 Jun;29(6):1091-1095. [doi: 10.1109/TPAMI.2007.1078] [Medline: 17431306]
61. Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. Semantic Scholar. 2015 Jul 06. URL: https://www.semanticscholar.org/paper/Siamese-Neural-Networks-for-One-Shot-Image-Koch/f216444d4f2959b4520c61d20003fa30a199670a [accessed 2024-05-21]
62. Zhu Y, Zhang C, Yu D, Zhang Y. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. PLoS Comput Biol 2022 Dec;18(12):e1010793 [FREE Full text] [doi: 10.1371/journal.pcbi.1010793] [Medline: 36548439]
63. Zhou W, Torvik VI, Smalheiser NR. ADAM: another database of abbreviations in MEDLINE. Bioinformatics 2006 Nov 15;22(22):2813-2818. [doi: 10.1093/bioinformatics/btl480] [Medline: 16982707]
64. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
65. Stedman's Medical Abbreviations, Acronyms & Symbols (4th Edition). Baltimore, MD: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

XSL•FO
RenderX

66. Meineke F, Modersohn L, Loeffler M, Boeker M. Announcement of the German Medical Text Corpus Project (GeMTeX). Stud Health Technol Inform 2023 May 18;302:835-836. [doi: 10.3233/SHTI230283] [Medline: 37203512]

Abbreviations

BERT: bidirectional encoder representations from transformers
CARD: Clinical Abbreviation Recognition and Disambiguation
CASI: Clinical Abbreviation Sense Inventory
CLASSE GATOR: Clinical Acronym Sense Disambiguator
CNN: convolutional neural network
CUIMC: Columbia University Irving Medical Center
DL: deep learning
EBMR: Evidence-Based Medicine Reviews
EHR: electronic health record
ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately
GPT: generative pretrained transformer
i2b2: Informatics for Integrating Biology and the Bedside
LLM: large language model
MeSH: Medical Subject Headings
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SVM: support vector machine
UMLS: Unified Medical Language System

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by T Abdullahi, V Leblanc; comments to author 29.03.24; revised version received 24.05.24; accepted 25.07.24; published 26.09.24.

Please cite as:
Kugic A, Martin I, Modersohn L, Pallaoro P, Kreuzthaler M, Schulz S, Boeker M
Processing of Short-Form Content in Clinical Narratives: Systematic Scoping Review
J Med Internet Res 2024;26:e57852
URL: https://www.jmir.org/2024/1/e57852
doi:10.2196/57852
PMID:39325515

©Amila Kugic, Ingrid Martin, Luise Modersohn, Peter Pallaoro, Markus Kreuzthaler, Stefan Schulz, Martin Boeker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Just-in-Time Adaptive Interventions for Behavior Change in Physiological Health Outcomes and the Use Case for Knee Osteoarthritis: Systematic Review

Janis Fiedler[1*], PhD; Matteo Reiner Bergmann[1*], MSc; Stefan Sell[1], PhD; Alexander Woll[1], PhD; Bernd J Stetter[1], PhD

Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany

[*]these authors contributed equally

Corresponding Author:
Janis Fiedler, PhD
Institute of Sports and Sports Science
Karlsruhe Institute of Technology
Engler-Bunte-Ring 15
Karlsruhe, 76131
Germany
Phone: 49 721 608 46978
Email: janis.fiedler@kit.edu

Abstract

Background: The prevalence of knee osteoarthritis (KOA) in the adult population is high and patients profit from individualized therapy approaches. Just-in-time adaptive interventions (JITAIs) are upcoming digital interventions for behavior change.

Objective: This systematic review summarizes the features and effectiveness of existing JITAIs regarding important physiological health outcomes and derives the most promising features for the use case of KOA.

Methods: The electronic databases PubMed, Web of Science, Scopus, and EBSCO were searched using keywords related to JITAIs, physical activity (PA), sedentary behavior (SB), physical function, quality of life, pain, and stiffness. JITAIs for adults that focused on the effectiveness of at least 1 of the selected outcomes were included and synthesized qualitatively. Study quality was assessed with the Quality Assessment Tool Effective Public Health Practice Project.

Results: A total of 45 studies with mainly weak overall quality were included in this review. The studies were mostly focused on PA and SB and no study examined stiffness. The design of JITAIs varied, with a frequency of decision points from a minute to a day, device-based measured and self-reported tailoring variables, intervention options including audible or vibration prompts and tailored feedback, and decision rules from simple if-then conditions based on 1 variable to more complex algorithms including contextual variables.

Conclusions: The use of frequent decision points, device-based measured tailoring variables accompanied by user input, intervention options tailored to user preferences, and simple decision rules showed the most promising results in previous studies. This can be transferred to a JITAI for the use case of KOA by using target variables that include breaks in SB and an optimum of PA considering individual knee load for the health benefits of patients.

(J Med Internet Res 2024;26:e54119) doi:10.2196/54119

KEYWORDS

eHealth; application; applications; digital health; digital technology; digital intervention; digital interventions; physical activity; sedentary behavior; sedentary lifestyle; pain; physical function; quality of life; knee osteoarthritis; osteopathy; bone; arthritis; mobile phone; individualization

XSL•FO
RenderX

Introduction

Background

Osteoarthritis is one of the most common chronic joint conditions and leads to pain, stiffness, and reduced physical function, which in turn diminishes quality of life (QoL) [1,2]. The most commonly affected weight-bearing joints are the knees, followed by the hips [3]. In 2019, about 344 million people worldwide were living with an osteoarthritis severity level that could benefit from effective therapy, and this number will grow over the next few years [4]. Appropriate mechanical stimuli, including physical activity (PA), are joint health protecting factors [5]. Furthermore, an adequate level of PA and reduced sedentary behavior (SB) results in reduced pain and improves physical function and stiffness in patients with knee osteoarthritis (KOA) [6-10]. Thereby, the QoL of people that are affected can also be improved [11]. Treating pain and overcoming functional limitations are essential for preventing a downward spiral of reduced PA and declining QoL [12]. Inadequate PA can lead to excess bodyweight, which increases mechanical joint loading, for example, during locomotion, and accelerates the degenerative process of KOA [13]. Furthermore, individuals with KOA are more prone to cardiovascular disease, as factors, such as reduced PA, are also correlated with elevated cardiovascular disease risk [14]. Consequently, PA and exercise are recommended for preventing such disease [14]. Additionally, PA and exercise have been shown to be beneficial in terms of alleviating KOA symptoms such as pain and stiffness [15]. Previous research has highlighted the importance of new strategies to encourage patients with KOA to participate in greater levels of PA [13,16]. However, KOA is a highly heterogeneous and multifactorial disease that affects individuals differently with variations in disease progression, severity, and response to treatments [13]. Consequently, an individualized therapy approach is suggested to effectively manage the symptoms, slow down the progression of the disease, and promote a healthy lifestyle specifically tailored to each person [17]. While this is also possible for interventions that are delivered in person, mobile health (mHealth) interventions are upcoming methods with the opportunity to deliver such highly individualized interventions at a larger scale. mHealth interventions are defined as medical and public health practices supported by mobile devices, such as mobile phones, patient monitoring devices, PDAs, and other wireless devices [18]. These interventions profit from the fast development of accompanying sensors such as accelerometers and heart rate monitors [19], which allow for ecologically valid real-time interventions. Previous studies show promising results for PA promotion and the reduction of SB if a theoretical foundation and behavior change techniques (BCTs) were used for the intervention [20]. In the context of KOA, reviews point to good usability and the potential for effective interventions, but the research is in an early stage [21,22]. A review including mHealth interventions for KOA points out that most available apps lack scientific studies backing their effectiveness, but those evaluated by RCTs show promising results concerning pain, physical function, and PA [23]. The main features of these apps included exercise prescription and the tracking of symptoms. A special case of mHealth interventions of which patients with KOA could greatly benefit are the highly individualized just-in-time adaptive interventions (JITAIs) or ecological momentary interventions (from here on only the term JITAI is used). These aim to generate real-time or near–real-time feedback by leveraging prior knowledge of individuals and previously collected data. With this compiled, dynamic knowledge, tailored interventions and nudges can be delivered when needed through mobile technologies, aligned with the individual's context and requirements [24]. These nudges can then change proximal outcomes such as step count to achieve the ultimate distal goals of the JITAI, such as better joint health and QoL [25]. An intervention is defined as a JITAI if it corresponds to real-time needs, adapts to input data, and is triggered by the system with the aim to deliver the intervention at a state of vulnerability, opportunity and receptivity of participants [25-28]. In the context of designing JITAIs, decision points represent the specific points in time when the JITAI can be triggered, while decision rules are the criteria that are applied at a decision point to determine if the JITAI should be triggered and which intervention option should be used. Intervention options refer to the possible actions that the JITAI can take at a decision point, while tailoring variables denote the sensor or user input that is used for adaptation [26]. A previous review on JITAIs for PA promotion by Hardeman et al [28] and a meta-analysis for device-based measured health outcomes by Xu and Smit [29] point to the feasibility and preliminary effectivity of JITAIs to promote PA and reduce SB. However, further variables (ie, physical function, stiffness, pain, and QoL) are important to KOA [8-11] that were not considered in previous works. Additionally, previous reviews did not aim to clearly depict decision points, tailoring variables, intervention options, and decision rules from the literature. This, however, is important to build a foundation for future JITAI designs. Finally, JITAIs are an upcoming field, with new studies getting published frequently warranting frequent updates of the literature.

Objective

This review aimed to systematically summarize the literature regarding the features of JITAIs of important physiological health outcomes and proxies (ie, PA, SB, physical function, QoL, pain, and stiffness). The second aim was to extract the most promising features that should be considered by future JITAIs in the treatment of KOA.

Methods

Design

This systematic review was registered on the Open Science Framework [30] on March 7, 2022. It was carried out based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) expanded checklist [31].

Inclusion Criteria

Inclusion and exclusion criteria were formulated according to PICOS (Population, Intervention, Comparison, Outcomes, Study) and are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

1. Population: Participants in the studies must be aged >18 years, regardless of sex and gender. Thus, the only exclusion criterion relates to studies that only included participants aged <18 years of age. Studies were included if <50% of participants were aged >18 years.
2. Intervention: All studies that use a digital physiological treatment that can be defined as a JITAI are included in this review. If studies use other types of interventions in addition to JITAIs, they are still included, and only the JITAI parts were considered. Furthermore, there were no restrictions regarding the duration of the intervention, the number of intervention sessions, or the technical device for the intervention.
3. Comparator: There was no restriction regarding control. We included active control (eg, other physiological treatments or interventions), passive control (no intervention or treatment), and studies with no control group (CG).
4. Outcome: Studies were included whenever they evaluated the effectiveness of a JITAI regarding one of the parameters PA, SB, pain, QoL, physical function, or stiffness as a main outcome. This also included all parameters that are synonyms or components of the previously mentioned parameters.
5. Study design: We included all study types, except reviews or meta-analysis. The only important thing was that the study reported on the effectiveness of the intervention.

Search Strategy

The electronic databases PubMed, Web of Science, Scopus, and EBSCO were used to identify relevant studies. Because search strategies exclusively relying on databases have been shown to be nonexhaustive [32], the reference list of all selected articles was checked for further relevant studies. Initial searches were conducted on October 26, 2022, and the search was rerun on April 9, 2024. We combined different search terms related to the intervention type, the area of mHealth, and the outcome parameters. Further details are available in Multimedia Appendix 1.

Paper Selection

All identified studies were exported to a reference management tool (EndNote [Clarivate] or Citavi [Swiss Academic Software GmbH]). The reference management tools were used to remove all duplicates. Then, all titles were checked for eligibility by MRB and JF independently. Second, the abstracts were screened, and third, the full texts of all remaining papers were screened, and the reasons for exclusion were noted by the 2 authors. Any disagreements at one of the screening steps were resolved by discussion until consensus was found.

Study Quality Assessment

The Quality Assessment Tool Effective Public Health Practice Project was used for the evaluation of the study quality. It is a specifically designed tool to test and support findings in public health research and refers to articles related to a broad number of health-related topics, including chronic diseases [33]. By evaluating 6 different criteria, a global rating was assessed at the end for each study. Every criterion could be rated *weak* (red), *moderate* (yellow), or *strong* (green), which finally influences the global rating. One *weak* rating leads to a *moderate* global rating, and 2 or more *weak* ratings result in a *weak* global rating. Only studies without a *weak* rating can achieve a *strong* global rating [33].

Data Extraction

All important information of the selected studies has been reproduced in tabular form in the Results section. The preparation of the tables and their contents were developed by MRB and JF in consultation with BS. For the most accurate rendering of the study content, reference was made to the study characteristics, including study design, sample size, population, and setting, as well as characteristics of the participants. To provide starting points for future JITAIs, the features and delivery of the JITAIs, including decision points, tailoring variables, intervention options, and decision rules based on the JITAI framework [27], and the duration of the interventions were also extracted from the existing studies. To display effectiveness and theoretical foundation, all data regarding retention, measurement, and the significant within- or between-group differences of the intervention and the use of BCTs stratified by control and intervention group (IG) according to Michie et al [34], and the theoretical foundation of the interventions are displayed in the Results section.

Results

Study Selection

A total of 4340 papers were exported to Endnote (MRB) and Citavi (JF) and screened (Figure 1). Initially, 571 (13.16%) duplicates were excluded by the reference management tools, and another 279 (6.43%) duplicates were removed manually. The remaining 3490 (80.41%) papers were screened for their titles, and 3133 (72.19%) were excluded by the authors. Thus, 357 (8.23%) abstracts were read, and 213 (4.91%) more papers were excluded at this step. The updated search yielded 1215 (28%) additional papers, including 145 (3.34%) duplicates. The 1070 (24.65%) unique papers were screened for title and abstract, and the remaining 88 (2.03%) papers for full texts. The main reasons for exclusion were that it was no intervention study, the intervention did not qualify as a JITAI, the outcome did not fit the inclusion criteria, it was a systematic review or meta-analysis, or all participants were <18 years old. Full texts of the remaining 144 (3.32%) studies were screened, and finally 18 (0.41%) papers were included in the review, supplemented by additional 14 (0.32%) papers from the updated search. During the screening process, 13 more studies were added. Nine studies were adopted from the systematic review by Hardeman et al [28], and 4 additional papers were found during data extraction by checking references of the included studies. Finally, 45 studies are presented and summarized in Table 1 and Multimedia Appendices 2 [30-74] and 3 [35-79].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. JITAI: just-in-time adaptive intervention.

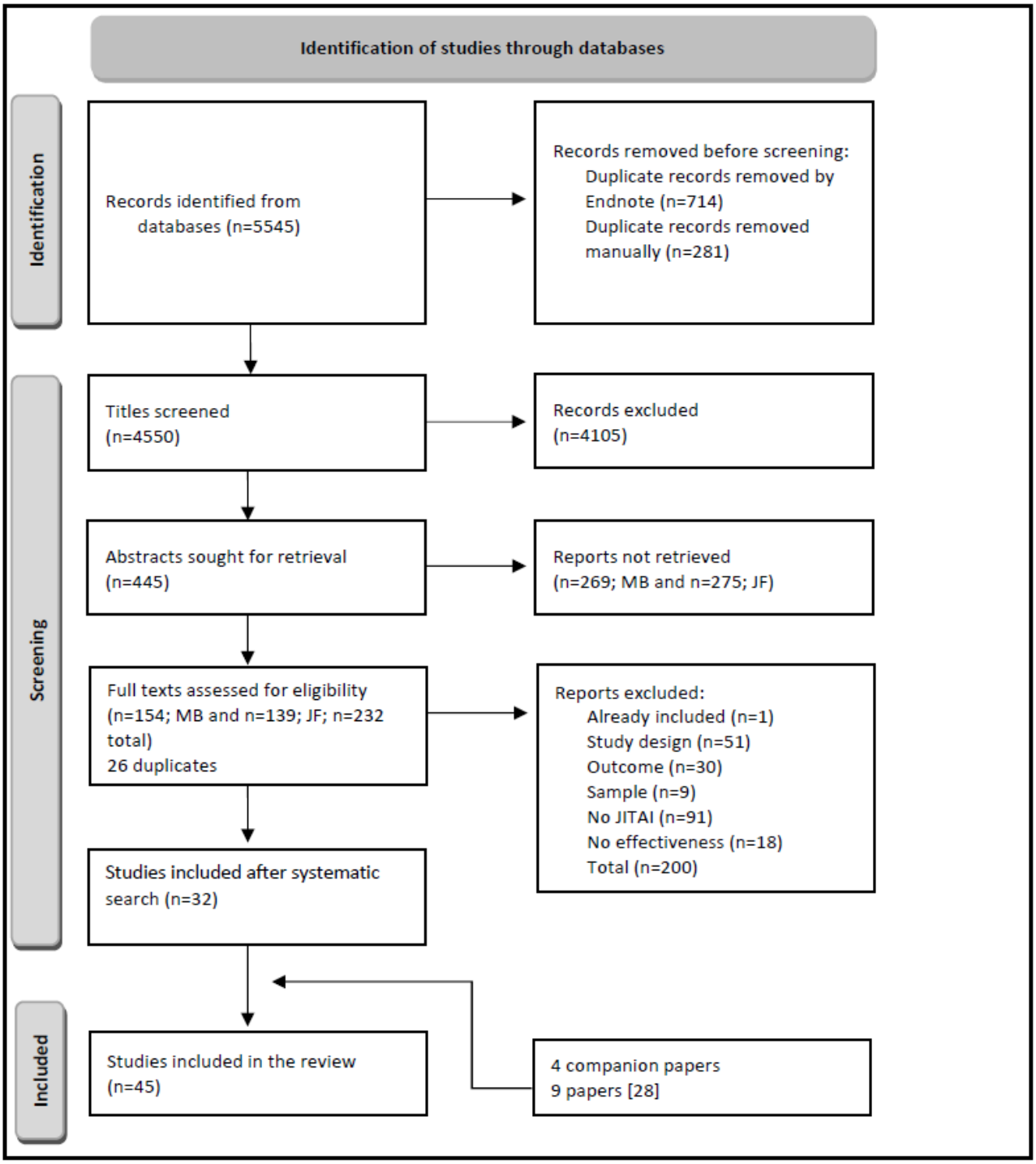

XSL•FO
RenderX

Table 1. Study characteristics.

Author and country of study	Study design	Population and activity profile	Analyzed sample size	Participant characteristics (age, gender, and ethnicity)
Allicock et al [36], United States	• Feasibility study or pilot evaluation • Randomized controlled trial	• Female African American breast cancer survivors. • Healthy enough to engage in moderate to vigorous-intensity PA[a].	• N=22 (IG[b]=13, 59%; and CG[c]=9, 41%)	• Age: mean (SD)=52.2 (9.2) years • No male participants. • 100% (22/22) African American.
Baumann et al [37], Germany	• Multiarm, parallel-group, randomized controlled trial	• Health care professionals (nursing staff and office workers) aged 18 years and older	• N=170 (IG 1=21, 12.4%; IG 2=23, 13.5%; IG 3=7, 4.1%; IG 4=34, 20%; IG 5=16, 9.4%; and CG=69; 40.6%)	• Age: mean (SD)=41.1 (10.9) years. • 39% (66/170) male participants
Bond et al [38], and Thomas and Bond [39], United States	• Feasibility study or pilot evaluation • Randomized within-person study • The participants were randomized each week to one of the 3 different conditions	• Adults who are overweight or obese • No exclusion based on PA level	• N=30	• Age: mean (SD)=47.5 (13.5) years. • 17% (5/30) male participants. • 67% (20/30) White, 13% (4/30) African American, 3% (1/30) American Indian or Alaskan Native, 3% (1/30) Asian, and 13% (4/30) other
Bort-Roig et al [40], Spain	• Randomized controlled trial	• Desk-based employees at-work and away from work • Low activity level outside work • No physical or health problems that limited their ability to stand for bouts of at least 10 minutes	• N=141 (IG=90, 63.8%; and CG=51, 36.2%) • Follow-up: N=64 (IG=42, 66%; and CG=22, 34%)	• Age: mean (SD)=45 (9) years • 18% (25/141) male participants
Brakenridge et al [41,42], Australia	• Cluster-randomized workplace trial • Mixed methods evaluation	• Desk-based office employees • Not reported	• N=153 (IG=66, 43.1%; and CG=87, 56.9%) • Only IG data were used for 2018 results.	• IG age: mean (SD)=37.6 (7.8) years • 47% (31/66) male participants • CG age: mean (SD)=40 (8) years. • 60% (52/87) male participants
Carlozzi et al [43], United States and Wang et al [44], United States	• Carlozzi et al: feasibility study or pilot evaluation • Randomized controlled trial. • Wang et al: pilot evaluation. • 2-arm microrandomized controlled trial.	• Care partners who are aged 18 years and are caring for an individual aged 18 years or older.	• Carlozzi et al N=70 (IG=36; 51%; and CG=34, 49%) • Wang et al N=36 (IG of Carlozzi et al [43])	• IG age: mean (SD)=54.4 (13.1) years • 22% (8/36) male participants • 81% (29/36) Caucasian, 3% (1/36) African American, and 8% (3/36) Asian • CG age: mean (SD)=56.1 (14.5) years • 35% (12/34) male participants. • 97% (33/34) Caucasian and 3% (1/34) African American

XSL•FO
RenderX

Author and country of study	Study design	Population and activity profile	Analyzed sample size	Participant characteristics (age, gender, and ethnicity)
Compernolle et al [45], Belgium	• Single-group pre-post design	• Age >60 years • Able to walk 100 m without difficulties	• N=26	• Age: mean (SD)=64.4 (3.8) years • 50% (13/26) male participants • 100% (26/26) Flemish
Conroy et al [46], United States	• Pilot evaluation • Secondary data analysis using a within-person design	• Insufficiently active young adults aged between 18 and 29 years	• N=58	• Age: mean (SD)=24.6 (3.1) years • 31% (18/58) male participants. • 69% (40/58) White and 31% (18/58) other
Ding et al [47], United States	• Feasibility study or pilot evaluation • Randomized controlled trial	• College students. • Participants who were in the contemplation and preparation stage based on the transtheoretical model.	• N=16 (IG=9, 56%; and CG=7, 44%)	• Age: 18-25 years • 63% (10/16) male participants. • Other characteristics were not reported.
Direito et al [48], New Zealand	• Feasibility study or pilot evaluation • Single-group pre-post design	• Adults who did not meet PA recommendations or who did meet recommendations but intended to decrease sedentary time.	• N=69	• Age: mean (SD)=34.5 (11.8) years • 22% (15/69) male participants • 54% (37/69) New Zealand European, and 46% (32/69) other
Fiedler et al [49], Germany	• Cluster-randomized controlled trial • Secondary data analysis using a within-person design for the IG	• Families (at least 1 parent and 1 child) • Not reported	• N=80	• 59% (47/80) adults • 53% (42/80) male participants
Finkelstein et al [50], United States	• Feasibility study or pilot evaluation • Randomized controlled crossover study	• Sedentary, overweight women • Daily inactivity >3 h	• N=30	• Age: mean (SD)=52 (12) years • No male participants • 47% (14/30) White and 47% (14/30) African American
Freene et al [51], Australia	• Feasibility study or pilot evaluation • Single-group pre-post design	• Adults with a stable coronary heart disease • Able to perform a submaximal walking test	• N=20	• Age: mean (SD)=54 (13) years. • 85% (17/20) male participants • 61% (12/20) Australian
Fundoiano-Hershcovitz et al [52], Israel	• Retrospective cohort study	• Adults with high pain levels and >6 weekly training hours (users with <6 weekly training hours were used for the sensitivity analysis).	• N=981	• Age: mean (SD)=39.2 (29) years • 39% (383/981) male participants
Garland et al [53], United States	• Feasibility study or pilot evaluation • Randomized controlled trial	• Adult patients who are opioid-treated for chronic pain	• N=63	• Age: mean (SD)=53.6 (12.8) years • 59% (37/63) male participants • 92% (58/63) White and 8% (5/63) other
Golbus et al [54], United States	• Randomized clinical trial	• Low to moderate risk patients aged between 18 and 74 years in cardiac rehabilitation	• N=223 (IG=112, 50.2%; and CG=111, 49.8%)	

XSL•FO
RenderX

Author and country of study	Study design	Population and activity profile	Analyzed sample size	Participant characteristics (age, gender, and ethnicity)
				• Age: mean (SD)=59.6 (10.6) years • 69% (154/223) male participants • 84% (187/223) White and 16% (36/223) others
Hermens et al [55], and Tabak [56], the Netherlands	• Feasibility study or pilot evaluation • Quasi-experimental study • Single-group pre-post design	• People living with COPD[d] who had completed a lung rehabilitation program 3 mo before the study	• N=8	• Aged between 49 and 64 years • 60% (5/8) male participants
Hietbrink et al [57], the Netherlands	• Mixed methods longitudinal study	• Adults with type 2 diabetes	• N=20 • N=15 (PA module) • N=5 (nutritional module)	• Age: mean (SD)=68 (8) years. • 70% (14/20) male participants
Hiremath et al [58], United States	• Nonrandomized pilot evaluation • Single-group pre-post design	• Individuals with SCI[e] • Use a manual wheelchair and self-propel them • No active pelvic or thigh wounds, CVD[f], or pregnancy	• N=20	• Age: mean (SD)=39.4 (12.8) years • 80% (16/20) male participants • 12.4 (12.5) years since injury
Ismail and Al Thani [59], Qatar	• Experimental study (between-group design) • Controlled trial	• Adults with a predominantly sedentary job • Able to walk	• N=58 (IG=29; 50%; and CG=29, 50%)	• Aged between 23 and 39 years • 34% (10/29) male participants (IG) and 41% (12/29) male participants (CG)
Klasnja et al [60], United States	• Feasibility study or pilot evaluation • Single-group pre-post design • 2×2 factorial experiment • Within-person microrandomized trial	• Adults who are obese after bariatric surgery	• N=45	• Age: mean (SD)=45 (11.9) years • 16% (7/45) male participants • 80% (36/45) White, 7% (3/45) Black, 2% (1/45) other, and 7% (3/45) mixed
Li et al [61], United States	• Feasibility study or pilot evaluation • Single-group pre-post design	• No prior diagnosis of cognitive impairment or dementia, sedentary lifestyle (self-reported >6 hours of sitting activities per day), poor sleep quality (insomnia severity index ≥8), and no diagnosis of sleep apnea	• N=8	• Age: mean (SD)=74 (5.42) years • 25% (2/8) male participants • 50% (4/8) White and 50% (4/8) Black
Low et al [62], United States	• Usability and feasibility study • Single-group pre-post design	• Adults with abdominal cancer postsurgery • Able to stand and walk unassisted	• N=15	• Age: mean=49.7 years • 20% (3/15) male participants • 87% (13/15) White and 13% (2/15) Black
Low et al [63], United States	• Feasibility study or pilot evaluation • Randomized controlled trial	• Adults scheduled for surgical treatment of metastatic gastrointestinal or peritoneal cancer and able to stand and walk unassisted	• N=26 (IG=13, 50%; and CG=13; 50%)	

XSL•FO
RenderX

Author and country of study	Study design	Population and activity profile	Analyzed sample size	Participant characteristics (age, gender, and ethnicity)
				• Age: mean (SD)=56.2 (10.5) years • 58% (15/26) male participants • 92% (24/26) White, 4% (1/26) Black, and 4% (1/26) more than one
Martin et al [64], United States	• Randomized clinical trial	• Outpatients of an academic cardiac vascular disease prevention center • Self-reported moderate or vigorous leisure-time ≥ 30 min/day at <3 days/week	• N=48 (unblind, no texts=16, 33%; Unblind, texts=16, 33%; and blind=16, 33%).	• Age: mean (SD)=58 (8) years. • 54% (26/48) male participants • 79% (38/48) White
McEntee et al [65], United States	• Randomized trial	• Adults who are inactive and healthy	• N=512 (n=128, 25% per study group)	• Age: mean (SD)=45.5 (9.1) years • 36% (184/512) male participants • 19% (97/512) Hispanic, 6% (31/512) African American, 2% (10/512) Asian, and 84% (430/512) White
Nurmi et al [66], Finland	• Pilot evaluation • N-of-1 randomized controlled trial	• General population >18 years No contraindications to engaging in PA, no use of any PA trackers or PA apps in the previous 6 months, no participation in other trials or behavior change programs in the previous 6 months or during the trial, below 150 MV-PA[g]/week	• N=15	• Age: mean (SD)=42.3 (9.8) years • 27% (4/15) male participants
Pellegrini et al [67], United States	• Feasibility study or pilot evaluation • Quasi-experimental study • Single-group prepost design	• People living with type 2 diabetes • Sedentary occupation or spend ≥75% of the day sitting	• N=9	• Age: mean (SD)=53.1 (10.7) years • 23% (2/9) male participants • 23% (2/9) White and 77% (7/9) Black
Rabbi et al [68], United States	• Feasibility study or pilot evaluation, within-subject • Randomized controlled trial	• Volunteers, including students and professionals • Not reported	• N=17	• Age: mean (SD)=28.3 (6.96) years, 18-49 years • 53% (9/17) male participants
Rabbi et al [69], United States	• Feasibility study or pilot evaluation • Quasi-experimental study • Controlled within-person trial	• Employees of Cornell University • Not reported	• N=16	• Age: 18-29 years: 25% (4/16), 30-39 years: 38% (6/16), 40-49 years: 19% (3/16), and >50 years: 19% (3/16). • 44% (7/16) male participants
Rabbi et al [70], United States	• Feasibility study or pilot evaluation • Single-blinded, controlled within-person trial	• Adults with chronic back pain (>6 months in duration) • Able to walk without mobility aids	• N=10	• Age: 31-60 years • 30% (3/10) male participants

XSL•FO
RenderX

Author and country of study	Study design	Population and activity profile	Analyzed sample size	Participant characteristics (age, gender, and ethnicity)
Radhakrishnan et al [71], United States	• Feasibility study or pilot evaluation • Randomized controlled trial	• Adults with heart failure • Able to walk	• N=38 (IG=19; 50%; and CG=19; 50%)	• Age: 55 years or older • 53% (20/38) male participants • 76% (29/38) White, 16% (6/38) African American, 3% (1/38) Native American, and 5% (2/38) other
Robertson et al [35], United States	• Feasibility study or pilot evaluation • Randomized controlled trial	• Adult cancer survivors • Did not meet the recommended activity levels	• N=78 (IG=39, 50%; and CG=39; 50%)	• Age: 18-34 years: 5% (4/78), 35-49 years: 26% (20/78), 50-64 years: 33% (26/78), 65-74 years: 33% (26/78), and >74 years: 3% (2/78) • 9% (7/78) male participants • 1% (1/78) American Indian or Alaska native, 17% (13/78) Black or African American, 79% (62/78) White, and 1% (1/78) other
Sporrel et al [72], the Netherlands	• Feasibility study or pilot evaluation • Randomized controlled trial design with single-group pre-post evaluation	• Adults between 18 and 55 years • Did not meet WHO[h] PA guidelines and no medical condition that made it unsafe to engage in unsupervised PA	• N=20 (Smart PAUL[i]=11, 55%; and Basic PAUL=9, 45%)	• Age: mean (SD)=30.65 (8.4) years • 15% (3/20) male participants
Stuber et al [73], the Netherlands	• Parallel cluster-randomized controlled trial	• 30 to 80 y old adults, regular shoppers at a participating supermarket (self-reported to purchase >50% weekly groceries at a participating supermarket)	• N=361 (IG=162, 44.9%; and CG=199, 55.1%)	• IG age: mean (SD)=58.9 (11.5) years • 29% (105/361) male participants • CG age: mean (SD)=57.2 (10.2) years • 26% (94/361) male participants
Tabak et al [74,75], the Netherlands	• Randomized controlled pilot trial • Secondary data analysis using a within-person design for the IG	• Patients with COPD • A clinical diagnosis of COPD, no infection or exacerbation in the 4 wk prior to start of the study, and a current or former smoker	• N=30 (IG=14, 47%; and CG=16, 53%)	• IG age: mean (SD)=65.2 (9) years • 57% (8/14) male participants. • CG age: mean (SD)=67.9 (5.7) years • 69% (11/16) male participants
Valle et al [76,77], United States	• Randomized controlled trial	• 18 to 39 years old patients with a cancer diagnosis in the previous 10 years • engaging in <150 min/week of MVPA	• N=280 (IG=140, 50%; and CG=140; 50%)	• Age: mean (SD)=33.4 (4.8) years • 18% (50/280) male participants • 23% (64/280) racial and ethnic minorities
van Dantzig et al [78], the Netherlands	• Feasibility study or pilot evaluation • Study 1: qualitative study • Study 2: randomized controlled trial		• Study 1: N=8 • Study 2: N=86 (IG=40, 47%; and CG=46, 53%)	

XSL•FO
RenderX

Author and country of study	Study design	Population and activity profile	Analyzed sample size	Participant characteristics (age, gender, and ethnicity)
		• Study 1: office workers • Study 2: healthy office workers • Study 1: not reported • Study 2: sedentary job, no known physical handicap or other condition that makes PA (walking) impossible, older than 30 years, not participating in another PA intervention		• Study 1 age: not reported • 50% (4/8) male participants • Study 2 • IG age: mean (SD)=44.5 (7.9); 30-57 years • 58% (23/40) male participants. • CG age: mean (SD)=44.3 (8); 32-63 years • 63% (29/46) male participants
van Dantzig et al [79], the Netherlands	• Substantive evaluation • Randomized controlled trial	• Desk workers with regular (daytime) working hours • Not reported	• N=70 • IG and CG not reported	• Age: 18-65 years • 73% (51/70) male participants (based on N=70 recruited)

[a]PA: physical activity.
[b]IG: intervention group.
[c]CG: control group.
[d]COPD: chronic obstructive pulmonary disease.
[e]SCI: individuals with spinal cord injury.
[f]CVD: cardiovascular disease.
[g]MVPA: moderate to vigorous physical activity.
[h]WHO: World Health Organization.
[i]PAUL: The Playful Active Urban Living app.

Description of the Included Studies

A total of 25 (56%) studies were conducted in the United States, 9 (20%) in the Netherlands, 2 (4%) in Australia, 2 (4%) in Germany, and the others in New Zealand, China, Qatar, Finland, Israel, Spain, and Belgium. Most of them (26/45, 58%) were designed as feasibility study or pilot evaluations, and they were mainly randomized controlled trials (23/45, 51%). In total, 9 of 45 (20%) were conducted in a single-group pre-post design, 3 (7%) as a secondary analysis using a within-person design for the IG, 2 as a controlled within-person trial, 2 (4%) as a mixed-method evaluation, and all other studies were either a retrospective cohort study, a randomized within-person study, or combined different study designs. The minimum number of participants was 8 (3 studies), and the maximum was 981 participants with an average of 99 (SD 175) participants. Two studies focused exclusively on women, and overall, 62% (2434/3960) of all participants were women and 38% (1516/3950) were men. The mean age in 19 studies was >50 years, and they focused mainly on a specific population such as sedentary people, patients with chronic diseases (eg, chronic obstructive pulmonary disease and coronary heart disease), or survivors of cancer. Retention in 3 studies was <50%, 4 studies were below 80%, 21 were ≥80% and <100%, and the remaining 12 studies had full retention. PA was the most common outcome (39 studies), followed by 19 studies that used parameters of SB as an outcome, of which 18 studies considered both PA and SB. In total, 9 studies examined the impact on QoL, 4 used synonyms of physical function, and 3 studies used pain as an outcome. None of the included articles referred to stiffness. Detailed results are displayed in Multimedia Appendix 3. An average of 8 BCTs (SD 7) were included for the IG and 4 (SD 5) for the CG. The number of BCTs in the IG ranged from 2 to 47, compared with 1 to 27 in the CG. The intervention duration ranged from 1 day to a maximum of 12 months, with a typical duration of 3 to 6 weeks. One study had no time restriction but a step count of 166.000 as a cutoff for the intervention duration [35]. Details on BCTs and theoretical foundations can be found in Multimedia Appendix 4 [35-79].

Key Features of JITAIs

Detailed key features of the included JITAIs are displayed in Multimedia Appendix 2.

Most decision points at which a JITAI could potentially be triggered were set to short time intervals, which were either specified to one decision point each minute (n=11) or not clearly specified but indicating short time intervals (n=21). Furthermore, 5 studies set decision points to time frames of 10 minutes up to 2 hours, and 11 were set to multiple times per day up to daily.

Tailoring variables used for JITAI adaptation were mainly device-based measured by accelerometer, smartphone, or smartwatch and concerned PA and SB outcomes supplemented with variables such as time, weather, or location. A total of 10 studies used self-reported tailoring variables through a baseline questionnaire [64], through ecological momentary assessment

XSL•FO
RenderX

(EMA) [36,43,44,57], or through nonspecified methods to assess pain and fatigue [62,63], to enable nonspecified user input [53], or to supplement PA data if the device has not been worn [68-70].

Intervention options at decision points included mainly the delivery of an audible or vibration prompt to participants accompanied by tailored feedback, information about health behavior, and suggestions for behavioral alternatives. One study explicitly explained in the prompt why the prompt occurred [47]. The alternatives based on the decision rules were mainly not to send the prompt for the decision points with short time frames, while those with longer timeframes had different intervention options depending on the tailoring variables.

Decision rules defining opportune moments were mainly based on if-then conditions. This means that an opportune moment was defined if a certain threshold —for example, 20 minutes of SB [67], <100 steps in 1 hour [49], a bout (≥3 min) of moderate to vigorous physical activity (or higher) PA was performed [58], lumbar posture [41,42], time, location, or weather-based thresholds [78] were met. Most decision rules were based on one if-then condition, while some used if-then decision trees. More complex models applied multiarmed bandit algorithms (sequential decision-making algorithm [80]) [68-70], pareto-frontier algorithm (taking user preferences into account [81]) [69], prelearned reinforcement learning models (models that make use of historical data and of data during the trial [82]) [72], k-nearest neighbor classifier (one of the most fundamental and simple classification methods [83]), and real-time prediction and self-learning by a support vector machine (a classical machine learning technique [84]) [55,56]. If the conditions for an opportune moment were met, intervention options were delivered, which included both fixed responses and microrandomized suggestions (eg, Klasnja et al [60]) for behavior change.

Effectiveness

Detailed results for intervention effectiveness are displayed in Multimedia Appendix 3. For RCTs, 5 studies indicated effectiveness compared with a CG for PA [35,37,64,65,68], 3 studies for SB [37,63,79], and 1 study for pain [53]. For pre-post–designed and within-person studies, 8 studies indicated effectiveness in most outcome variables of the study for PA [38,39,46,49,57,61,67,69], 5 for SB [38,39,52,61,67], 1 study for pain [52], and 1 study for QoL [51]. Regarding the opportune-moment identification, decision points, and rules, most interventions that indicated effectiveness had near real-time decision points on a minute-to-minute basis, and simple decision rules depicting opportune moments when a threshold of 20 to 120 minutes of inactivity, SB, or computer activity was reached. Of these, the studies of Rabbi et al [68,69] were the only ones that used more complex algorithms. Effectiveness was also found for longer decision point periods: 2 times per hour [35], 2 to 3 times per day [57,64], and daily [51]. Effective interventions had an intervention duration between 3 weeks and 12 months, where the majority of studies had an intervention length between 3 and 8 weeks. The other included studies that indicated no or limited effectiveness and reported heterogeneous features regarding decision points, decision rules, opportune-moment identifications, and the complexity of algorithms as well.

Study Quality

The quality assessment showed that 4 studies achieved a “strong” global rating (Figure 2 [35-79]). Overall, 15 studies were rated “moderate,” and all the other studies were judged with 2 or more “weak” ratings. The most common weak ratings in 38 and 17 out of 45 were related to the context of blinding and data collection methods, respectively.

XSL•FO
RenderX

Figure 2. Study quality of included studies based on the Quality Assessment Tool Effective Public Health Practice Project. Green indicates a strong rating, yellow a moderate rating, and red a weak rating of the respective category.

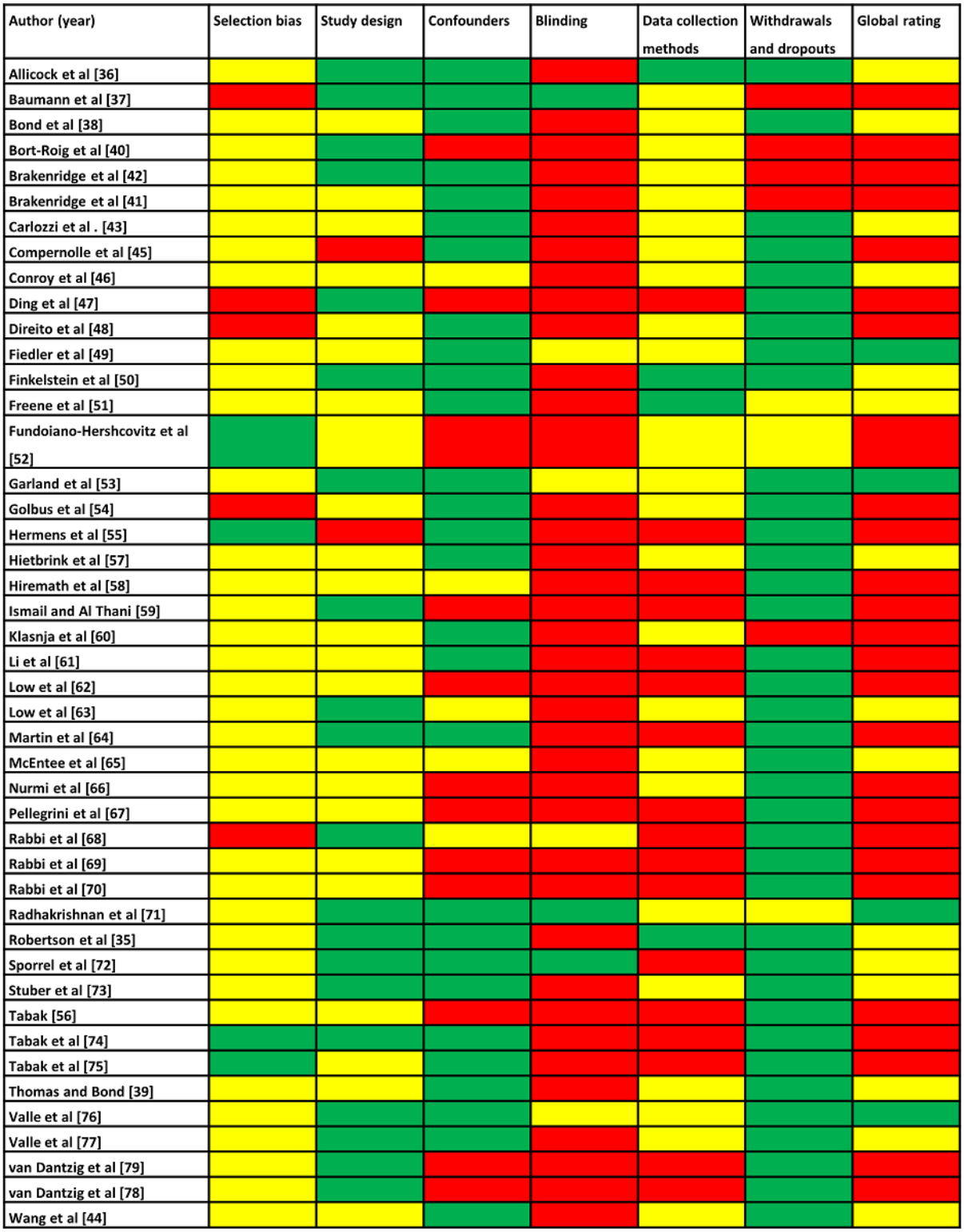

Discussion

Overview

The objective of this study was to provide a systematic overview of the concept of JITAIs and to explore important parameters in KOA management. In total, 45 studies with 38 different JITAIs focusing on the outcomes PA, SB, pain, QoL, physical function, and stiffness were analyzed in this review. The majority of studies investigated PA and SB, while physical function, pain, and QoL were addressed by few studies, and the parameter "stiffness" was not included in any study, indicating a research gap. Reporting and use of the key facets of JITAIs [27], along with the methodology and aspects of study design of the included studies, were heterogeneous. The best evidence for effective JITAIs exists for device-based measurements of PA and SB using frequent decision points and simple algorithms

XSL•FO
RenderX

to deliver prompts after a period of physical inactivity or SB. This can be translated to the use case of KOA and accompanied by a specific knee load management to improve KOA treatments with JITAIs in the future.

JITAIs in the Context of PA, SB, Pain, Stiffness, Physical Function, and QoL

Overall intervention effectiveness is hard to pinpoint and not comparable between studies due to heterogeneous study design, outcome measures, number of statistical comparisons, time windows, sample size, and focus on within- or between-person effects. Stratifying the results by study design, 5 out of 14 RCTs found the intervention to effectively increase PA and 3 studies had a positive effect on SB, and 1 study improved the pain level of participants. Pre-post study designs as well as the within-person studies found an increase in PA in 8 studies, a reduction of SB in 5 studies, a reduction of pain in 1 study, and 1 study positively influenced QoL. No evidence of effectiveness in studies for physical function outcomes was found. To obtain more results about causation under consideration of time-varying effects, it is important to increase the number of microrandomized trials in the future [60] and focus on the state of receptivity of participants. This will also provide more information about what works in which situation and for whom, and it will help to provide valuable information to advance the field of JITAI research.

Although the studies showed a high degree of heterogeneity, some patterns can be observed. For example, decision points with a shorter period (eg, minute-to-minute) and an intervention duration of 3 to 8 weeks have shown good evidence for intervention effectiveness. Here, it is important to add that intervention duration and timing of decision points have to be viewed together to avoid overburdening participants and to keep the content engaging. Otherwise, JITAIs run at a high risk of disengagement [85] that can hinder behavior change. The optimal dose for these parameters has yet to be completely explored. Conroy et al [46] point out that self-monitoring prompts, but not behavioral feedback, show signs of a dose-response association with PA volume in JITAIs. This is an important finding that should be followed up by further studies. The included studies provide limited information regarding the long-term effectivity of JITAIs for behavior change. Of the 15 studies that included a follow-up measurement, 3 studies showed significant benefits of the intervention for QoL [51,55,56] and 2 studies for PA [76,77]. This is not surprising, as JITAIs are an upcoming topic and still focus mainly on the opportune-moment identification, feasibility, usability, and technological challenges of the intervention [27]. In addition, decision rules that represent opportune moments when a threshold of 20 to 120 minutes of inactivity, SB, or computer activity has been reached resulted in behavior change. Here, shorter timeframes might be more useful to break SB, as Thomas and Bond [39] have shown that opportune moments after 30 minutes of being sedentary are more effective than longer decision points of 60 and 120 minutes.

PA or SB was monitored in all significant studies by the smartphone app or by an accelerometer. In addition, prompts were used by each study as an intervention option to interrupt SB, send encouraging messages, give suggestions to increase PA, or send useful and interesting links and information. For example, Freene et al [51] sent prompts in the form of do's to help change behavior and learn a new behavior. The unique feature of this JITAI was that the do's were sent based on a variety of tailoring variables, namely if participants showed a low score in PA, social opportunity, and variety on 3 consecutive days. Although it was not possible to positively influence PA or SB, the parameter QoL could be increased in 4 different domains, both after 6 and after 16 weeks. This result could mean that variety and social opportunity can be more easily influenced by do's than PA. Another reason could be that shorter decision points are more effective for PA while for QoL longer decision points such as days are sufficient.

Importantly, Ding et al [47] highlighted that 43% of the supposed "just-in-time messages" were not perceived as just-in-time by the participants. This is problematic, as sending prompts at moments when participants do not have the opportunity to change their behavior will not lead to the desired effect and might even lead to disengagement with the intervention [26]. The importance of adequate opportune-moment identification has also been highlighted by previous research [25,27] and should be assessed by future JITAI studies. This can be done by a repeated short EMA, which then can also inform and adapt the timing of the intervention and improve adherence and thus behavior change. One example is the study of Rabbi et al [70], who included participants preferences into the opportune-moment identification, which allowed their algorithm to, for example, adapt learned opportune moments that were no longer valid.

Shortcomings of the included studies include the overall weak study quality and the mainly simple intervention design that is not optimized for opportune-moment identification as it seldomly includes the state of receptivity of participants but focuses on the vulnerability. This should be the focus of future studies that consider participants to, for example, not initiate a JITAI trigger when the tailoring variables (eg, calendar or user preferences) indicate that the participant is not receptive [86]. In this context, the role of contextual factors assessed by GPS or EMA [87] are important aspects that are included in a few studies [36,51,54-57,59,60,68-70,72,78,79] and point to the feasibility of such interventions. If interventions including context variables are more effective than those that do not, this cannot be answered by this review as no clear pattern emerges. The additional information provided by these studies is nonetheless very helpful for future interventions and should be explored further, including EMA studies on this topic [88]. The blinding of participants was the main issue for the quality assessment. While it is difficult to blind participants in mHealth research, microrandomization makes this process easier and has a lot of additional benefits, as discussed before. Furthermore, the included studies scored weak to moderate for selection bias. Here, more diverse samples and nonconvenient samples are required in future research to strengthen the evidence base and the generalization of JITAI study results. Additionally, future studies should test different sets of decision points, tailoring variables, decision rules, and intervention options within one

XSL•FO
RenderX

study to improve the current literature on what works for whom and how. This would improve the evidence base for future JITAI interventions in different populations and settings.

Specific Considerations for the Transfer of the Finding to the Use Case of KOA

As a recommendation, the World Health Organization [17] states that a healthy lifestyle should be cultivated and that maintaining a normal body weight plays an important role in preventing osteoarthritis and control the disease progression. Here, PA and SB are behavioral variables that influence pain, physical function, stiffness, and QoL of patients with KOA [6-10]. Consequently, PA and SB can be used to a certain extent as proxy variables to target the distal outcomes pain, physical function, stiffness, and QoL. However, even though appropriate mechanical stimuli are needed for promoting optimal joint function [5], more PA is not necessarily better for patients with KOA [89] and can even support the development of the disease [62]. It is recommended to reduce overuse of joints by refraining from PA that contain start-stop movements, rapid changes in direction, intense jumps, and landings [90].

To design a JITAI for KOA, specific tailoring variables have to be considered. On the basis of our findings, starting points for intervention studies could include device-based measured PA and SB variables in addition to time, location, and pain scores assessed as user input through a mix of event-based and random sampling using EMAs [68] to tailor the intervention to participants needs. From a theoretical perspective, additional JITAIs for stiffness would be very helpful to provide an evidence base for the creation of a JITAI for KOA. A further important tailoring variable that has not been found by this review is the knee load. Here, body-worn measurement technologies such as orthopedic shoe insoles or knee braces with built-in sensors to measure the forces exerted on the knee joint are promising developments for interventions that can adapt the minimum and maximum load limit to each participant and situation [91-94]. Wearable joint monitoring technologies may also help indirectly assess varying levels of joint stiffness.

For example, by quantifying the joint range of motion [95] or assessing relationships between gait biomechanics and fluctuations in KOA symptoms, valuable data can be gained [96]. Additionally, EMA could be used for a targeted assessment of specific symptoms, such as morning stiffness. Pain can also be assessed as user input through a mix of event-based and random sampling using EMAs [68]. This would allow to tailor the JITAI to, for example, a user specific self-reported amount of pain, or knee load under consideration of certain conditions and allow for a highly individualized approach.

A further important intervention method found by this review are educational prompts. These prompts could contain informative facts to educate the user on how to move and exercise appropriately [17] tailored to, for example, their PA, SB, pain, and stiffness. While conventional JITAIs mainly aim to increase PA through classical exercise, JITAIs that focus on the therapy of specific diseases, such as KOA, could use therapeutic exercises and tasks to promote health behavior. For example, it has already been shown that a combination of muscle strengthening and range-of-motion exercises resulted in a significant improvement in osteoarthritis symptoms in the long term [3,97]. Moreover, muscle strengthening was associated with reduced pain and increased function in patients with KOA [98]. These exercises could be encouraged in appropriate moments through the use of a JITAI. In the deployment of digital interventions and JITAIs, it is imperative to account for the technological prerequisites and potential barriers for participants and practitioners. This consideration becomes especially significant when the intended audience includes older patients, exemplified by those with KOA, as underscored in the research conducted by Hinman et al [23].

Designing a JITAI for the Use Case of KOA

Several important decisions need to be made to design a JITAI for KOA. First, the outcome parameters need to be selected. Second, the device for JITAI delivery and assessment of tailoring variables needs to be chosen. Third, tailoring variables, decision points, intervention options, and decision rules need to be defined.

Derived from previous research and the results of this review, the aim of a promising JITAI is to target the proximal outcomes PA and SB in accordance with the recommendations of KOA therapy [17]. This should result in positive developments in the distal outcomes of pain, physical function, stiffness, and QoL. Here, smartphones, accelerometers, or KOA-specific measurement technologies can be used for the intervention and the evaluation of effectiveness. In addition to device-based measured PA and SB, EMAs can be used to gather information through user input about loads of previous activities or provide inferences about parameters such as pain and stiffness. The potential of EMAs for understanding individual behavior is indispensable [99,100] and 2 studies including EMA showed significant pain reduction in participants [52,53]. In addition, the use of sensors for load management would be very useful to provide device-based feedback of knee load and adapt the intervention accordingly to prevent an overloading of the knee joint [93,94]. Further approaches could also include smartphone data about the weather, calendar, or the user's location to improve the opportune-moment identification [27].

The time span in which the decision points are to be made should be of shorter duration, as emerged from the results. Decision points every minute proved to be useful and offer quick adaptation to changing situations. That is especially important to avoid too high knee load in near real time and to break up SB. Prompts to boost overall PA and deliver knowledge could use longer decision points to avoid overburdening the participants [26].

The best evidence regarding intervention options exists for prompts to break prolonged SB or physical inactivity. That is also a relevant parameter for KOA [9] and could be linked with positive reinforcement messages or booster messages, which showed success in the study by Martin et al [64]. Additional intervention options include maintaining an optimal amount of PA while avoiding overloading of the osteoarthritis-affected knee. This could be achieved by continuously tracking and displaying the estimated knee load (eg, using knee-specific feedback technology [94,101]) on the smartphone and prompting participants in near real time if the knee load reaches a critical

XSL•FO
RenderX

individual level. To avoid insufficient PA of participants, daily PA goals could be used. Here, daily prompts with positive reinforcement messages could prompt participants in the morning if they failed to achieve the PA goal of the previous day.

In the next step, the decision rules can be formulated. On the basis of the study by Thomas and Bond [39], interrupting SB after 30 minutes is promising, which is also linked to improved metabolic health [102]. Here, it is important to consider that the prompt really appears in an opportune moment [103] to avoid overburdening participants, especially if such a short time frame is chosen. One option to solve this is to allow users to choose time frames when such triggers occur, enable them to mute the triggers for a certain period, or use more complex decision-making algorithms based on machine learning principles that implement the availability of the user (eg, including calendar). Rules for accompanying positive reinforcement messages and educational information about why PA is useful in the topic of KOA should be formulated to have a maximum number of messages per day. There is also the possibility of including movement recommendations to go for a walk in the nearest park or to explore new places. Decision rules for the optimal amount of PA should only send prompts if participants are closing in on their critical knee load level or if inappropriate kinds of PA such as reoccurring jumps are detected to avoid making participants hesitant about PA. These prompts should always be followed by an explanation for the prompt and recommendations for the future. As these prompts are very hard to define, users should have the option to adapt the prompt in correspondence with a health professional until the prompt is appropriately defined.

Finally, an important consideration for the design of a JITAI in the context of KOA is co-designing of the intervention. Mrklas et al [104] describe a frequently occurring problem within the topic of mHealth apps. Users, in this case people with KOA, are usually not involved in the development of the JITAI-app. That can lead to interventions that are not a good fit for the target population. For better individualization, it is crucial to include the opinions and experiences of those affected. This can help to positively influence the user engagement, which has an impact on the effectiveness of a JITAI, as highlighted by previous research [105-107]. The framework by Choi et al [103] addresses this issue and could be used for a higher user-friendliness and thus a higher engagement in the design of JITAIs.

Limitations and Strengths

Strengths of this review include the detailed search, which has a high number of results within the 4 databases. Furthermore, the chance of finding all relevant articles was increased by examining useful articles for further studies, and thus we were able to include 45 relevant papers. The examination of multiple outcome variables relevant for patients with KOA yields a good fit for a special use case while also providing an overview of JITAIs for health behavior change in general because it was not focused on a specific group of participants.

Limitations of this review include that the terminology and structure of a JITAI have not yet been defined based on a common definition, which hinders screening of eligible articles. A uniform framework for future JITAIs would be of great advantage and would increase comparability [28]. Therefore, this review is oriented and structured based on a JITAI framework [27]. In general, most of the included JITAIs have shown a very high heterogeneity, a low number of participants, and include preliminary studies, which makes it difficult to draw inferences. In addition, either no studies or only very few studies were found that refer to the parameters physical function, stiffness, pain, and also QoL, which limits the explanatory power regarding these parameters. In order to make evidence-based statements, further outcome-specific and high-quality JITAI studies are needed. Finally, body weight and the associated load on the joints are decisive factors in KOA [108], which has not been addressed by this review to maintain a concise focus.

Conclusions

In summary, a variety of JITAIs can be identified that represent interesting and promising results for behavior change in physiological health outcomes. Although a high degree of heterogeneity in the results was found, patterns have been identified that are associated with intervention effectiveness. Using frequent decision points, device-based measurements of tailored variables accompanied by user input, intervention options tailored to user preferences, and simple decision rules showed the most promising results in previous studies. Transferring this knowledge to the use case of KOA can result in a JITAI to break SB, increase PA, and maintain an optimal knee load for the health benefits of patients. Future JITAIs should be oriented toward a uniform structure and terminology, use microrandomized interventions to be able to draw better inferences, integrate new technological aspects such as sensors for the assessment of knee load, and explore optimal dose-response mechanisms and contextual factors that are linked to intervention effectiveness.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors' Contributions

JF and MRB conducted the literature search, article screening, and data extraction. MRB and BJS assessed the study quality. Interpretation of the data and conducting of the first draft of the manuscript were done by JF and MRB with valuable comments from BJS. BJS, SS, and AW revised the manuscript. All authors read and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 22 KB - jmir_v26i1e54119_app1.docx]

Multimedia Appendix 2
Key facets and duration of the just-in-time adaptive interventions.
[DOCX File , 99 KB - jmir_v26i1e54119_app2.docx]

Multimedia Appendix 3
Retention, measurements, and results of the interventions.
[DOCX File , 99 KB - jmir_v26i1e54119_app3.docx]

Multimedia Appendix 4
Theoretical foundation and behavior change techniques.
[DOCX File , 91 KB - jmir_v26i1e54119_app4.docx]

Multimedia Appendix 5
PRISMA checklist.
[PDF File (Adobe PDF File), 100 KB - jmir_v26i1e54119_app5.pdf]

References

1. Allen KD, Thoma LM, Golightly YM. Epidemiology of osteoarthritis. Osteoarthritis Cartilage 2022 Mar;30(2):184-195 [FREE Full text] [doi: 10.1016/j.joca.2021.04.020] [Medline: 34534661]
2. Salaffi F, Carotti M, Stancati A, Grassi W. Health-related quality of life in older adults with symptomatic hip and knee osteoarthritis: a comparison with matched healthy controls. Aging Clin Exp Res 2013 Jul 25;17(4):255-263. [doi: 10.1007/bf03324607]
3. Bennell KL, Hinman RS. A review of the clinical evidence for exercise in osteoarthritis of the hip and knee. J Sci Med Sport 2011 Jan;14(1):4-9. [doi: 10.1016/j.jsams.2010.08.002] [Medline: 20851051]
4. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Dec;396(10267):2006-2017. [doi: 10.1016/s0140-6736(20)32340-0]
5. Hunter DJ, Eckstein F. Exercise and osteoarthritis. J Anat 2009 Mar 21;214(2):197-207 [FREE Full text] [doi: 10.1111/j.1469-7580.2008.01013.x] [Medline: 19207981]
6. Bartels EM, Juhl CB, Christensen R, Hagen KB, Danneskiold-Samsøe B, Dagfinrud H, et al. Aquatic exercise for the treatment of knee and hip osteoarthritis. Cochrane Database Syst Rev 2016 Mar 23;3(3):CD005523 [FREE Full text] [doi: 10.1002/14651858.CD005523.pub3] [Medline: 27007113]
7. Chang WD, Chen S, Lee CL, Lin MY, Lai PT. The effects of Tai Chi Chuan on improving mind-body health for knee osteoarthritis patients: a systematic review and meta-analysis. Evid Based Complement Alternat Med 2016;2016:1813979 [FREE Full text] [doi: 10.1155/2016/1813979] [Medline: 27635148]
8. Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee: a Cochrane systematic review. Br J Sports Med 2015 Dec 24;49(24):1554-1557. [doi: 10.1136/bjsports-2015-095424] [Medline: 26405113]
9. Lee J, Chang RW, Ehrlich-Jones L, Kwoh CK, Nevitt M, Semanik PA, et al. Sedentary behavior and physical function: objective evidence from the Osteoarthritis Initiative. Arthritis Care Res (Hoboken) 2015 Mar 24;67(3):366-373 [FREE Full text] [doi: 10.1002/acr.22432] [Medline: 25155652]
10. Zampogna B, Papalia R, Papalia GF, Campi S, Vasta S, Vorini F, et al. The role of physical activity as conservative treatment for hip and knee osteoarthritis in older people: a systematic review and meta-analysis. J Clin Med 2020 Apr 18;9(4):1167 [FREE Full text] [doi: 10.3390/jcm9041167] [Medline: 32325775]
11. Vitaloni M, Botto-van Bemden A, Sciortino Contreras RM, Scotton D, Bibas M, Quintero M, et al. Global management of patients with knee osteoarthritis begins with quality of life assessment: a systematic review. BMC Musculoskelet Disord 2019 Oct 27;20(1):493 [FREE Full text] [doi: 10.1186/s12891-019-2895-3] [Medline: 31656197]
12. Sinusas K. Osteoarthritis: diagnosis and treatment. Am Fam Physician 2012 Jan 01;85(1):49-56 [FREE Full text] [Medline: 22230308]

XSL•FO
RenderX

13. Mora JC, Przkora R, Cruz-Almeida Y. Knee osteoarthritis: pathophysiology and current treatment modalities. J Pain Res 2018;11:2189-2196 [FREE Full text] [doi: 10.2147/JPR.S154002] [Medline: 30323653]
14. Park D, Park YM, Ko SH, Choi YH, Min DU, Ahn JH, et al. Association between knee osteoarthritis and the risk of cardiovascular disease and the synergistic adverse effects of lack of exercise. Sci Rep 2023 Mar 16;13(1):2777 [FREE Full text] [doi: 10.1038/s41598-023-29581-1] [Medline: 36797339]
15. Li Y, Su Y, Chen S, Zhang Y, Zhang Z, Liu C, et al. The effects of resistance exercise in patients with knee osteoarthritis: a systematic review and meta-analysis. Clin Rehabil 2016 Oct 11;30(10):947-959. [doi: 10.1177/0269215515610039] [Medline: 26471972]
16. Farr JN, Going SB, Lohman TG, Rankin L, Kasle S, Cornett M, et al. Physical activity levels in patients with early knee osteoarthritis measured by accelerometry. Arthritis Rheum 2008 Sep 15;59(9):1229-1236 [FREE Full text] [doi: 10.1002/art.24007] [Medline: 18759320]
17. Osteoarthritis. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/osteoarthritis [accessed 2023-07-26]
18. mHealth: new horizons for health through mobile technologie. World Health Organization. 2011. URL: http://gbv.eblib.com/patron/FullRecord.aspx?p=851142 [accessed 2024-04-29]
19. Reichert M, Giurgiu M, Koch ED, Wieland LM, Lautenbach S, Neubauer AB, et al. Ambulatory assessment for physical activity research: state of the science, best practices and future directions. Psychol Sport Exerc 2020 Sep;50:101742 [FREE Full text] [doi: 10.1016/j.psychsport.2020.101742] [Medline: 32831643]
20. Fiedler J, Eckert T, Wunsch K, Woll A. Key facets to build up eHealth and mHealth interventions to enhance physical activity, sedentary behavior and nutrition in healthy subjects - an umbrella review. BMC Public Health 2020 Oct 23;20(1):1605 [FREE Full text] [doi: 10.1186/s12889-020-09700-7] [Medline: 33097013]
21. Kamilu Sulaiman S, Wong AY, Liangchi Li L, Fordjour Antwi-Afari M, Ou H, Wh Tsang H. The use of mobile health technology in the management of osteoarthritis: a scoping review with scientometric analyses. Int J Med Inform 2023 Feb;170:104937. [doi: 10.1016/j.ijmedinf.2022.104937] [Medline: 36493537]
22. Shah N, Costello K, Mehta A, Kumar D. Applications of digital health technologies in knee osteoarthritis: narrative review. JMIR Rehabil Assist Technol 2022 Jun 08;9(2):e33489 [FREE Full text] [doi: 10.2196/33489] [Medline: 35675102]
23. Hinman RS, Lawford BJ, Nelligan RK, Bennell KL. Virtual tools to enable management of knee osteoarthritis. Curr Treatm Opt Rheumatol 2023 Mar 11:1-21 [FREE Full text] [doi: 10.1007/s40674-023-00202-2] [Medline: 37362068]
24. Spruijt-Metz D, Nilsen W. Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Comput 2014;13(3):13-17 [FREE Full text] [doi: 10.1038/ijo.2013.120] [Medline: 23812094]
25. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219 [FREE Full text] [doi: 10.1037/hea0000306] [Medline: 26651462]
26. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2016 Sep 23;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
27. Wunsch K, Eckert T, Fiedler J, Woll A. Just-in-time adaptive interventions in mobile physical activity interventions - a synthesis of frameworks and future directions. Eur Health Psychol 2022;22(4):842 [FREE Full text]
28. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 03;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
29. Xu Z, Smit E. Using a complexity science approach to evaluate the effectiveness of just-in-time adaptive interventions: a meta-analysis. Digit Health 2023 Jul 02;9:20552076231183543 [FREE Full text] [doi: 10.1177/20552076231183543] [Medline: 37521518]
30. Bergmann M, Fiedler J. Just-in-time adaptive interventions (JITAIs) for knee osteoarthrits: a mixed-method review. Open Science Framework. 2024. URL: https://osf.io/y2hcd/ [accessed 2024-09-02]
31. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
32. Hopewell S, Clarke MJ, Lefebvre C, Scherer RW. Handsearching versus electronic searching to identify reports of randomized trials. Cochrane Database Syst Rev 2007(2):8. [doi: 10.1002/14651858.mr000001]
33. Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. Worldviews Ev Based Nurs 2004 Jun 23;1(3):176-184. [doi: 10.1111/j.1524-475x.2004.04006.x]
34. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug 20;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

XSL•FO
RenderX

35. Robertson MC, Lyons EJ, Liao Y, Baum ML, Basen-Engquist KM. Gamified text messaging contingent on device-measured steps: randomized feasibility study of a physical activity intervention for cancer survivors. JMIR Mhealth Uhealth 2020 Nov 24;8(11):e18364 [FREE Full text] [doi: 10.2196/18364] [Medline: 33231551]
36. Allicock M, Kendzor D, Sedory A, Gabriel KP, Swartz MD, Thomas P, et al. A pilot and feasibility mobile health intervention to support healthy behaviors in African American breast cancer survivors. J Racial Ethn Health Disparities 2021 Mar 8;8(1):157-165. [doi: 10.1007/s40615-020-00767-x] [Medline: 32385847]
37. Baumann H, Heuel L, Bischoff LL, Wollesen B. Efficacy of individualized sensory-based mHealth interventions to improve distress coping in healthcare professionals: a multi-arm parallel-group randomized controlled trial. Sensors (Basel) 2023 Feb 19;23(4):2322 [FREE Full text] [doi: 10.3390/s23042322] [Medline: 36850920]
38. Bond DS, Thomas JG, Raynor HA, Moon J, Sieling J, Trautvetter J, et al. B-MOBILE--a smartphone-based intervention to reduce sedentary time in overweight/obese individuals: a within-subjects experimental trial. PLoS One 2014;9(6):e100821 [FREE Full text] [doi: 10.1371/journal.pone.0100821] [Medline: 24964010]
39. Thomas JG, Bond DS. Behavioral response to a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior in obese adults: implications for JITAI optimization. Health Psychol 2015 Dec;34S:1261-1267 [FREE Full text] [doi: 10.1037/hea0000304] [Medline: 26651467]
40. Bort-Roig J, Chirveches-Pérez E, Giné-Garriga M, Navarro-Blasco L, Bausà-Peris R, Iturrioz-Rosell P, et al. An mHealth workplace-based "sit less, move more" program: impact on employees' sedentary and physical activity patterns at work and away from work. Int J Environ Res Public Health 2020 Nov 28;17(23):8844 [FREE Full text] [doi: 10.3390/ijerph17238844] [Medline: 33260697]
41. Brakenridge CL, Healy GN, Winkler EA, Fjeldsoe BS. Usage, acceptability, and effectiveness of an activity tracker in a randomized trial of a workplace sitting intervention: mixed-methods evaluation. Interact J Med Res 2018 Mar 02;7(1):e5 [FREE Full text] [doi: 10.2196/ijmr.9001] [Medline: 29500158]
42. Brakenridge CL, Fjeldsoe BS, Young DC, Winkler EA, Dunstan DW, Straker LM, et al. Evaluating the effectiveness of organisational-level strategies with or without an activity tracker to reduce office workers' sitting time: a cluster-randomised trial. Int J Behav Nutr Phys Act 2016 Nov 04;13(1):115 [FREE Full text] [doi: 10.1186/s12966-016-0441-3] [Medline: 27814738]
43. Carlozzi NE, Choi SW, Wu Z, Troost JP, Lyden AK, Miner JA, et al. An app-based just-in-time-adaptive self-management intervention for care partners: the CareQOL feasibility pilot study. Rehabil Psychol 2022 Nov;67(4):497-512 [FREE Full text] [doi: 10.1037/rep0000472] [Medline: 36355640]
44. Wang J, Wu Z, Choi SW, Sen S, Yan X, Miner JA, et al. The dosing of mobile-based just-in-time adaptive self-management prompts for caregivers: preliminary findings from a pilot microrandomized study. JMIR Form Res 2023 Sep 14;7:e43099 [FREE Full text] [doi: 10.2196/43099] [Medline: 37707948]
45. Compernolle S, Van Dyck D, Cardon G, Brondeel R. Exploring breaks in sedentary behavior of older adults immediately after receiving personalized haptic feedback: intervention study. JMIR Mhealth Uhealth 2021 May 10;9(5):e26387 [FREE Full text] [doi: 10.2196/26387] [Medline: 33970109]
46. Conroy DE, Wu J, Lee AM, Brunke-Reese D, Lagoa CM. Dose-response relations between the frequency of two types of momentary feedback prompts and daily physical activity. Health Psychol 2023 Mar;42(3):151-160 [FREE Full text] [doi: 10.1037/hea0001271] [Medline: 36862471]
47. Ding X, Xu J, Wang H, Chen G, Thind H, Zhang Y. WalkMore: promoting walking with just-in-time context-aware prompts. In: Proceedings of the 2016 Conference on IEEE Wireless Health. 2016 Presented at: WH '16; October 25-27, 2016; Bethesda, MD p. 1-8 URL: https://ieeexplore.ieee.org/document/7764558 [doi: 10.1109/wh.2016.7764558]
48. Direito A, Tooley M, Hinbarji M, Albatal R, Jiang Y, Whittaker R, et al. Tailored daily activity: an adaptive physical activity smartphone intervention. Telemed J E Health 2020 Apr 01;26(4):426-437. [doi: 10.1089/tmj.2019.0034] [Medline: 31063038]
49. Fiedler J, Seiferth C, Eckert T, Woll A, Wunsch K. A just-in-time adaptive intervention to enhance physical activity in the SMARTFAMILY2.0 trial. Sport Exerc Perform Psychol 2023 Feb;12(1):43-57. [doi: 10.1037/spy0000311]
50. Finkelstein JB, Bedra M, Li X, Wood J, Ouyang P. Mobile app to reduce inactivity in sedentary overweight women. Stud Health Technol Inform 2015;216:89-92. [Medline: 26262016]
51. Freene N, van Berlo S, McManus M, Mair T, Davey R. A behavioral change smartphone app and program (ToDo-CR) to decrease sedentary behavior in cardiac rehabilitation participants: prospective feasibility cohort study. JMIR Form Res 2020 Nov 03;4(11):e17359 [FREE Full text] [doi: 10.2196/17359] [Medline: 33141091]
52. Fundoiano-Hershcovitz Y, Horwitz DL, Tawil C, Cohen O, Goldstein P. The two-stage therapeutic effect of posture biofeedback training on back pain and the associated mechanism: a retrospective cohort study. Front Physiol 2022 Mar 18;13(1):958033 [FREE Full text] [doi: 10.3389/fphys.2022.958033] [Medline: 36589467]
53. Garland EL, Gullapalli BT, Prince KC, Hanley AW, Sanyer M, Tuomenoksa M, et al. Zoom-based mindfulness-oriented recovery enhancement plus just-in-time mindfulness practice triggered by wearable sensors for opioid craving and chronic pain. Mindfulness (N Y) 2023 May 24:1-17 [FREE Full text] [doi: 10.1007/s12671-023-02137-0] [Medline: 37362184]

XSL•FO
RenderX

54. Golbus JR, Gupta K, Stevens R, Jeganathan VS, Luff E, Shi J, et al. A randomized trial of a mobile health intervention to augment cardiac rehabilitation. NPJ Digit Med 2023 Sep 14;6(1):173 [FREE Full text] [doi: 10.1038/s41746-023-00921-9] [Medline: 37709933]
55. Hermens H, op den Akker H, Tabak M, Wijsman J, Vollenbroek M. Personalized coaching systems to support healthy behavior in people with chronic conditions. J Electromyogr Kinesiol 2014 Dec;24(6):815-826. [doi: 10.1016/j.jelekin.2014.10.003] [Medline: 25455254]
56. Tabak M. Telemedicine for patients with COPD: new treatment approaches to improve daily activity behaviour. University of Twente. 2014. URL: https://research.utwente.nl/en/publications/telemedicine-for-patients-with-copd-new-treatment-approaches-to-i-2 [accessed 2024-04-29]
57. Hietbrink EA, Oude Nijeweme-d'Hollosy WO, Middelweerd A, Konijnendijk AA, Schrijver LK, Ten Voorde AS, et al. A digital coach (E-Supporter 1.0) to support physical activity and a healthy diet in people with type 2 diabetes: acceptability and limited efficacy testing. JMIR Form Res 2023 Jul 28;7:e45294 [FREE Full text] [doi: 10.2196/45294] [Medline: 37505804]
58. Hiremath SV, Amiri AM, Thapa-Chhetry B, Snethen G, Schmidt-Read M, Ramos-Lamboy M, et al. Mobile health-based physical activity intervention for individuals with spinal cord injury in the community: a pilot study. PLoS One 2019 Oct 15;14(10):e0223762 [FREE Full text] [doi: 10.1371/journal.pone.0223762] [Medline: 31613909]
59. Ismail T, Al Thani D. Design and evaluation of a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior at work: experimental study. JMIR Form Res 2022 Jan 26;6(1):e34309 [FREE Full text] [doi: 10.2196/34309] [Medline: 35080498]
60. Klasnja P, Rosenberg DE, Zhou J, Anau J, Gupta A, Arterburn DE. A quality-improvement optimization pilot of BariFit, a mobile health intervention to promote physical activity after bariatric surgery. Transl Behav Med 2021 Mar 16;11(2):530-539. [doi: 10.1093/tbm/ibaa040] [Medline: 32421187]
61. Li J, Hodgson N, Lyons MM, Chen K, Yu F, Gooneratne NS. A personalized behavioral intervention implementing mHealth technologies for older adults: a pilot feasibility study. Geriatr Nurs 2020 May;41(3):313-319 [FREE Full text] [doi: 10.1016/j.gerinurse.2019.11.011] [Medline: 31810730]
62. Low CA, Danko M, Durica KC, Kunta AR, Mulukutla R, Ren Y, et al. A real-time mobile intervention to reduce sedentary behavior before and after cancer surgery: usability and feasibility study. JMIR Perioper Med 2020 Mar 23;3(1):e17292 [FREE Full text] [doi: 10.2196/17292] [Medline: 33393915]
63. Low CA, Danko M, Durica KC, Vega J, Li M, Kunta AR, et al. A real-time mobile intervention to reduce sedentary behavior before and after cancer surgery: pilot randomized controlled trial. JMIR Perioper Med 2023 Jan 12;6:e41425 [FREE Full text] [doi: 10.2196/41425] [Medline: 36633893]
64. Martin SS, Feldman DI, Blumenthal RS, Jones SR, Post WS, McKibben RA, et al. mActive: a randomized clinical trial of an automated mHealth intervention for physical activity promotion. J Am Heart Assoc 2015 Nov 09;4(11):e002239 [FREE Full text] [doi: 10.1161/JAHA.115.002239] [Medline: 26553211]
65. McEntee ML, Hurley JC, Phillips CB, Hooker SP, Todd M, Frank LD, et al. The moderating impact of neighborhood walkability on mHealth interventions to increase moderate to vigorous physical activity for insufficiently active adults in a randomized trial. Int J Behav Nutr Phys Act 2023 Aug 15;20(1):97 [FREE Full text] [doi: 10.1186/s12966-023-01494-2] [Medline: 37582736]
66. Nurmi J, Knittle K, Naughton F, Sutton S, Ginchev T, Khattak F, et al. Biofeedback and digitalized motivational interviewing to increase daily physical activity: series of factorial N-of-1 randomized controlled trials piloting the precious app. JMIR Form Res 2023 Nov 23;7:e34232 [FREE Full text] [doi: 10.2196/34232] [Medline: 37995122]
67. Pellegrini CA, Hoffman SA, Daly ER, Murillo M, Iakovlev G, Spring B. Acceptability of smartphone technology to interrupt sedentary time in adults with diabetes. Transl Behav Med 2015 Sep 14;5(3):307-314 [FREE Full text] [doi: 10.1007/s13142-015-0314-3] [Medline: 26327936]
68. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth 2015 May;3(2):e42 [FREE Full text] [doi: 10.2196/mhealth.4160] [Medline: 25977197]
69. Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 707-718 URL: https://dl.acm.org/doi/10.1145/2750858.2805840 [doi: 10.1145/2750858.2805840]
70. Rabbi M, Aung MS, Gay G, Reid MC, Choudhury T. Feasibility and acceptability of mobile phone-based auto-personalized physical activity recommendations for chronic pain self-management: pilot study on adults. J Med Internet Res 2018 Oct 26;20(10):e10147 [FREE Full text] [doi: 10.2196/10147] [Medline: 30368433]
71. Radhakrishnan K, Julien C, Baranowski T, O'Hair M, Lee G, Sagna De Main A, et al. Feasibility of a sensor-controlled digital game for heart failure self-management: randomized controlled trial. JMIR Serious Games 2021 Nov 08;9(4):e29044 [FREE Full text] [doi: 10.2196/29044] [Medline: 34747701]

XSL•FO
RenderX

72. Sporrel K, Wang S, Ettema DD, Nibbeling N, Krose BJ, Deutekom M, et al. Just-in-time prompts for running, walking, and performing strength exercises in the built environment: 4-week randomized feasibility study. JMIR Form Res 2022 Aug 01;6(8):e35268 [FREE Full text] [doi: 10.2196/35268] [Medline: 35916693]
73. Stuber JM, Mackenbach JD, de Bruijn G, Gillebaart M, Hoenink JC, Middel CN, SUPREME NUDGE consortium. Real-world nudging, pricing, and mobile physical activity coaching was insufficient to improve lifestyle behaviours and cardiometabolic health: the Supreme Nudge parallel cluster-randomised controlled supermarket trial. BMC Med 2024 Feb 02;22(1):52 [FREE Full text] [doi: 10.1186/s12916-024-03268-4] [Medline: 38303069]
74. Tabak M, Vollenbroek-Hutten MM, van der Valk PD, van der Palen J, Hermens HJ. A telerehabilitation intervention for patients with chronic obstructive pulmonary disease: a randomized controlled pilot trial. Clin Rehabil 2014 Jun 29;28(6):582-591. [doi: 10.1177/0269215513512495] [Medline: 24293120]
75. Tabak M, op den Akker H, Hermens H. Motivational cues as real-time feedback for changing daily activity behavior of patients with COPD. Patient Educ Couns 2014 Mar;94(3):372-378. [doi: 10.1016/j.pec.2013.10.014] [Medline: 24332934]
76. Valle CG, Diamond MA, Heiling HM, Deal AM, Hales DP, Nezami BT, et al. Effect of an mHealth intervention on physical activity outcomes among young adult cancer survivors: the IMPACT randomized controlled trial. Cancer 2023 Feb 01;129(3):461-472 [FREE Full text] [doi: 10.1002/cncr.34556] [Medline: 36444676]
77. Valle CG, Diamond MA, Heiling HM, Deal AM, Hales DP, Nezami BT, et al. Physical activity maintenance among young adult cancer survivors in an mHealth intervention: twelve-month outcomes from the IMPACT randomized controlled trial. Cancer Med 2023 Aug;12(15):16502-16516 [FREE Full text] [doi: 10.1002/cam4.6238] [Medline: 37317660]
78. van Dantzig S, Bulut M, Krans M, Van Der Lans A, Ruyter B. Enhancing physical activity through context-aware coaching. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: PervasiveHealth '18; May 21-24, 2018; New York, NY p. 187-190 URL: https://dl.acm.org/doi/10.1145/3240925.3240928 [doi: 10.1145/3240925.3240928]
79. van Dantzig S, Geleijnse G, van Halteren AT. Toward a persuasive mobile application to reduce sedentary behavior. Pers Ubiquit Comput 2012 Jul 12;17(6):1237-1246. [doi: 10.1007/s00779-012-0588-0]
80. Tewari A, Murphy SA. From ads to interventions: contextual bandits in mobile health. In: Rehg JM, Murphy SA, Kumar S, editors. Mobile Health: Sensors, Analytic Methods, and Applications. Cham, Switzerland: Springer; 2017:495-517.
81. Roberts JD, Walton RD, Loyer V, Bernus O, Kulkarni K. Open-source software for respiratory rate estimation using single-lead electrocardiograms. Sci Rep 2024 Jan 02;14(1):167 [FREE Full text] [doi: 10.1038/s41598-023-50470-0] [Medline: 38168512]
82. Wang S, Zhang C, Kröse B, van Hoof H. Optimizing adaptive notifications in mobile health interventions systems: reinforcement learning from a data-driven behavioral simulator. J Med Syst 2021 Oct 18;45(12):102 [FREE Full text] [doi: 10.1007/s10916-021-01773-0] [Medline: 34664120]
83. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967 Jan;13(1):21-27. [doi: 10.1109/TIT.1967.1053964]
84. Suthaharan S. Support vector machine. In: Suthaharan S, editor. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning Boston. New York, NY: Springer; 2016:207-235.
85. Park J, Lee U. Understanding disengagement in just-in-time mobile health interventions. Proc ACM Interact Mob Wearable Ubiquitous Technol 2023 Jun 12;7(2):1-27. [doi: 10.1145/3596240]
86. Schueller SM, Tomasino KN, Mohr DC. Integrating human support into behavioral intervention technologies: the efficiency model of support. Clin Psychol Sci Pract 2017 Mar;24(1):27-45. [doi: 10.1037/h0101740]
87. Schneider S, Junghaenel DU, Smyth JM, Fred Wen CK, Stone AA. Just-in-time adaptive ecological momentary assessment (JITA-EMA). Behav Res Methods 2024 Mar 25;56(2):765-783 [FREE Full text] [doi: 10.3758/s13428-023-02083-8] [Medline: 36840916]
88. Giurgiu M, Niermann C, Ebner-Priemer U, Kanning M. Accuracy of sedentary behavior-triggered ecological momentary assessment for collecting contextual information: development and feasibility study. JMIR Mhealth Uhealth 2020 Sep 15;8(9):e17852 [FREE Full text] [doi: 10.2196/17852] [Medline: 32930668]
89. Liu SH, Driban JB, Eaton CB, McAlindon TE, Harrold LR, Lapane KL. Objectively measured physical activity and symptoms change in knee osteoarthritis. Am J Med 2016 May;129(5):497-505.e1 [FREE Full text] [doi: 10.1016/j.amjmed.2015.12.029] [Medline: 26844633]
90. Hootman JM, Dick R, Agel J. Epidemiology of collegiate injuries for 15 sports: summary and recommendations for injury prevention initiatives. J Athl Train 2007;42(2):311-319 [FREE Full text] [Medline: 17710181]
91. Stetter BJ, Ringhof S, Krafft FC, Sell S, Stein T. Estimation of knee joint forces in sport movements using wearable sensors and machine learning. Sensors (Basel) 2019 Aug 25;19(17):3690 [FREE Full text] [doi: 10.3390/s19173690] [Medline: 31450664]
92. Wang L, Jones D, Jones A, Chapman GJ, Siddle HJ, Russell D, et al. A portable insole system to simultaneously measure plantar pressure and shear stress. IEEE Sensors J 2022 May 1;22(9):9104-9113. [doi: 10.1109/jsen.2022.3162713]
93. Faisal AI, Majumder S, Mondal T, Cowan D, Naseh S, Deen MJ. Monitoring methods of human body joints: state-of-the-art and research challenges. Sensors (Basel) 2019 Jun 10;19(11):2629 [FREE Full text] [doi: 10.3390/s19112629] [Medline: 31185629]

XSL•FO
RenderX

94. Stetter BJ, Krafft FC, Ringhof S, Stein T, Sell S. A machine learning and wearable sensor based approach to estimate external knee flexion and adduction moments during various locomotion tasks. Front Bioeng Biotechnol 2020 Jan 24;8:9 [FREE Full text] [doi: 10.3389/fbioe.2020.00009] [Medline: 32039192]
95. Zeni JA, Higginson JS. Dynamic knee joint stiffness in subjects with a progressive increase in severity of knee osteoarthritis. Clin Biomech (Bristol, Avon) 2009 May;24(4):366-371 [FREE Full text] [doi: 10.1016/j.clinbiomech.2009.01.005] [Medline: 19250725]
96. Tan JS, Tippaya S, Binnie T, Davey P, Napier K, Caneiro JP, et al. Predicting knee joint kinematics from wearable sensor data in people with knee osteoarthritis and clinical considerations for future machine learning models. Sensors (Basel) 2022 Jan 07;22(2):446 [FREE Full text] [doi: 10.3390/s22020446] [Medline: 35062408]
97. Thomas KS, Muir KR, Doherty M, Jones AC, O'Reilly SC, Bassey EJ. Home based exercise programme for knee pain and knee osteoarthritis: randomised controlled trial. BMJ 2002 Oct 05;325(7367):752 [FREE Full text] [doi: 10.1136/bmj.325.7367.752] [Medline: 12364304]
98. Bennell KL, Wrigley TV, Hunt MA, Lim BW, Hinman RS. Update on the role of muscle in the genesis and management of knee osteoarthritis. Rheum Dis Clin North Am 2013 Mar;39(1):145-176. [doi: 10.1016/j.rdc.2012.11.003] [Medline: 23312414]
99. Stone AA, Obbarius A, Junghaenel DU, Wen CK, Schneider S. High-resolution, field approaches for assessing pain: ecological momentary assessment. Pain 2021 Jan 21;162(1):4-9 [FREE Full text] [doi: 10.1097/j.pain.0000000000002049] [Medline: 32833794]
100. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr 01;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
101. Teague CN, Heller JA, Nevius BN, Carek AM, Mabrouk S, Garcia-Vicente F, et al. A wearable, multimodal sensing system to monitor knee joint health. IEEE Sensors J 2020 Sep 15;20(18):10323-10334. [doi: 10.1109/jsen.2020.2994552]
102. Healy GN, Dunstan DW, Salmon J, Cerin E, Shaw JE, Zimmet PZ, et al. Breaks in sedentary time: beneficial associations with metabolic risk. Diabetes Care 2008 Apr 05;31(4):661-666. [doi: 10.2337/dc07-2046] [Medline: 18252901]
103. Choi W, Zheng H, Franklin P, Tulu B. mHealth technologies for osteoarthritis self-management and treatment: a systematic review. Health Informatics J 2019 Sep 27;25(3):984-1003 [FREE Full text] [doi: 10.1177/1460458217735676] [Medline: 29090628]
104. Mrklas KJ, Barber T, Campbell-Scherer D, Green LA, Li LC, Marlett N, et al. Co-design in the development of a mobile health app for the management of knee osteoarthritis by patients and physicians: qualitative study. JMIR Mhealth Uhealth 2020 Jul 10;8(7):e17893 [FREE Full text] [doi: 10.2196/17893] [Medline: 32673245]
105. Kelders SM, Kip H, Greeff J. Psychometric evaluation of the TWente Engagement with Ehealth Technologies Scale (TWEETS): evaluation study. J Med Internet Res 2020 Oct 09;22(10):e17757 [FREE Full text] [doi: 10.2196/17757] [Medline: 33021487]
106. Kelders SM, van Zyl LE, Ludden GD. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol 2020 May 27;11:926 [FREE Full text] [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
107. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Mar 19;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
108. Georgiev T, Angelov AK. Modifiable risk factors in knee osteoarthritis: treatment implications. Rheumatol Int 2019 Jul 25;39(7):1145-1157. [doi: 10.1007/s00296-019-04290-z] [Medline: 30911813]

Abbreviations

BCT: behavior change technique
CG: control group
EMA: ecological momentary assessment
IG: intervention group
JITAI: just-in-time adaptive interventions
KOA: knee osteoarthritis
mHealth: mobile health
PA: physical activity
PICOS: Population, Intervention, Comparison, Outcomes, Study
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QoL: quality of life
SB: sedentary behavior

XSL•FO
RenderX

Edited by G Greco; submitted 30.10.23; peer-reviewed by M Robertson, S Mitra, O Malkowski; comments to author 08.04.24; revised version received 13.06.24; accepted 20.07.24; published 27.09.24.

Please cite as:
Fiedler J, Bergmann MR, Sell S, Woll A, Stetter BJ
Just-in-Time Adaptive Interventions for Behavior Change in Physiological Health Outcomes and the Use Case for Knee Osteoarthritis: Systematic Review
J Med Internet Res 2024;26:e54119
URL: https://www.jmir.org/2024/1/e54119
doi:10.2196/54119
PMID:

©Janis Fiedler, Matteo Reiner Bergmann, Stefan Sell, Alexander Woll, Bernd J Stetter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Internet-Based Telehealth Programs in Patients With Hip or Knee Osteoarthritis: Systematic Review and Meta-Analysis

Hao-Nan Wang[1,2*], PT, CSCS, PhD; Pei Luo[3,4*], PT, PhD; Shuyue Liu[5], MSc; Yunyi Liu[6], MSc; Xiao Zhang[5], PhD; Jian Li[1,2], MD

[1]Sports Medicine Center, West China Hospital, Sichuan University, Chengdu, China
[2]Department of Orthopedics and Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China
[3]School of Sport Science, Beijing Sport University, Beijing, China
[4]Key Laboratory of Physical Fitness and Exercise, Ministry of Education, Beijing, China
[5]Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada
[6]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada
[*]these authors contributed equally

Corresponding Author:
Jian Li, MD
Sports Medicine Center
West China Hospital
Sichuan University
No 37, Guoxue Lane, Wuhou District
Chengdu, 610041
China
Phone: 86 18980601388
Fax: 86 28 85582944
Email: lijian_sportsmed@163.com

Abstract

Background: Osteoarthritis (OA) is a chronic musculoskeletal disease that causes pain, functional disability, and an economic burden. Nonpharmacological treatments are at the core of OA management. However, limited access to these services due to uneven regional local availability has been highlighted. Internet-based telehealth (IBTH) programs, providing digital access to abundant health care resources, offer advantages, such as convenience and cost-effectiveness. These characteristics make them promising strategies for the management of patients with OA.

Objective: This study aimed to evaluate the effectiveness of IBTH programs in the management of patients with hip or knee OA.

Methods: We systematically searched 6 electronic databases to identify trials comparing IBTH programs with conventional interventions for hip and knee OA. Studies were selected based on inclusion and exclusion criteria, focusing on outcomes related to function, pain, and self-efficacy. Standardized mean differences (SMDs) with 95% CIs were calculated to compare outcome measures. Heterogeneity was assessed using I^2 and χ^2 tests. The methodological quality of the selected studies and the quality of evidence were also evaluated.

Results: A total of 21 studies with low-to-high risk of bias were included in this meta-analysis. The pooled results showed that IBTH has a superior effect on increasing function (SMD 0.30, 95% CI 0.23-0.37, P<.001), relieving pain (SMD –0.27, 95% CI –0.34 to –0.19, P<.001), and improving self-efficacy for pain (SMD 0.21, 95% CI 0.08-0.34, P<.001) compared to the conventional intervention group. Subgroup analysis revealed that IBTH with exercise can significantly alleviate pain and improve function and self-efficacy, but IBTH with cognitive-behavioral therapy only had the effect of reducing pain.

Conclusions: The meta-analysis provides moderate-quality evidence that IBTH programs have a beneficial effect on improving function, relieving pain, and improving self-efficacy compared to conventional interventions in patients with hip or knee OA. Limited evidence suggests that the inclusion of exercise regimens in IBTH programs is recommended.

Trial Registration: PROSPERO CRD42024541111; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=541111

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55576) doi:10.2196/55576

KEYWORDS

osteoarthritis; knee; hip; telehealth; telemedicine; telerehabilitation; eHealth; exercise; PRISMA

Introduction

Osteoarthritis (OA) is a worldwide chronic musculoskeletal disease. It is characterized by pathologic changes in cartilage, bone, synovium, ligament, muscle, and periarticular fat [1]. OA causes chronic pain, joint stiffness, disability, and a subsequent impact on the quality of life [2]. Worldwide, it is estimated that more than 240 million individuals have activity-limiting, symptomatic OA [3]. Among the major joints, the hip and the knee are most commonly affected by OA [4]. An estimated 24% of the general adult population has OA [5]. Both hip OA and knee OA are the leading causes of disability worldwide, with hip OA ranked as the 11th-highest contributor to disability worldwide and knee OA ranked as the 38th highest in term of years lived with disability [6]. OA-related disability results in substantial direct costs and mortality. On average, individuals with knee OA spend approximately US $15,000 in direct medical costs over their lifetimes [7]. There is currently no cure for OA, and to date, most research has primarily focused on treatments that alleviate pain and prevent functional decline. Nonpharmacological strategies, including patient education, cognitive-behavioral therapy (CBT), weight loss, and exercise, constitute the core management for hip and knee OA before resorting to surgical or pharmacological interventions [8-11]. However, the face-to-face delivery of health care services (eg, exercise and CBT) can be prohibitively expensive in terms of time and other costs for patients with incurable OA who require long-term interventions. However, the COVID-19 pandemic has forced us to change our traditional approach to treatment, so we can strike a balance between physical protection and ongoing care for older adults [12,13]. As the prevalence and treatment costs of hip and knee OA increase, there is a growing recognition of the need to identify effective treatment options that provide timely and equitable access to services regardless of location, accessibility, or public health policies, such as lockdown or quarantine, as in the case of the COVID-19 pandemic [14].

Telehealth is a promising health care delivery method that can help improve access to care. According to the World Health Organization (WHO) definition, telehealth refers to the "delivery of health care services, where patients and providers are separated by distance. Telehealth uses Information and Communication Technologies (ICT) for the exchange of information for the diagnosis and treatment of diseases and injuries, research and evaluation, and for the continuing education of health professionals" [15]. Recently, improvements in technology have dramatically increased the accessibility and quality of digitally delivered care. Despite this, telehealth has yet to be widely adopted due to stringent regulatory laws and a lack of supportive payment structures [16]. During the COVID-19 pandemic, social distancing requirements and lockdown restrictions severely impacted face-to-face health services and interventions. Telehealth has the potential to greatly increase patient access to quality, affordable care, while maintaining physical distance for the safety of both patients and providers [17]. For this reason, telehealth has been adopted and accelerated worldwide as a safe and feasible remote health service [18]. From the patient's perspective, the primary goal of telehealth is to increase access to care and improve the convenience of health care delivery. Telehealth can improve access to health care resources, especially in rural and underdeveloped areas [19]. Telehealth can also significantly reduce the time required to receive medical care, including travel time to health care facilities, time spent in waiting rooms, and time spent actually receiving medical care [20]. Furthermore, telehealth has the potential to contribute to the reduction in health care costs for patients. The implementation of remote monitoring and ongoing management of a patient's health and chronic conditions can reduce hospitalizations and delay joint replacements for patients with OA [21].

Over the past few decades, researchers have been dedicatedly assessing the effectiveness and feasibility of telehealth interventions, and delivery methods have continued to evolve. In the initial stages of research, texting and calling were used as delivery methods for telehealth interventions. However, the outcomes were not consistently satisfactory. For instance, telephone-based weight loss support with weight management and healthy lifestyle services does not reduce knee pain intensity or weight compared to usual care [22]. The advent of the internet in clinical practice as an information-sharing medium has created many opportunities for innovative interventions for people with chronic conditions and their health care providers [23]. Internet-based telehealth (IBTH) programs provide digital access to health resources and care, facilitating the remote transmission of patient demographic and clinical data and images via online platforms or mobile apps [24]. Telehealth access via the internet provides effective and enriched alternative assessments (eg, gait analysis), structured interventions (eg, exercise), and recreational interventions (eg, rehabilitation games) [25]. In addition, IBTH programs that incorporate interactive and continuous self-monitoring, feedback, and information exchange in self-directed or automated modes can provide educational materials and home-based exercises to replace some or all of the need for physiotherapist instruction [26,27]. However, the role of IBTH for people with OA is challenging as they are mostly older and less likely to have access to the internet and smart devices [23].

To date, emerging evidence suggests that IBTH programs have experienced rapid growth; however, evidence of their effectiveness in hip and knee OA remains unclear. Several reviews and meta-analyses have sought to evaluate the effectiveness of telehealth programs for individuals with OA, but the majority of these studies have focused on the efficacy of telehealth-supported exercise interventions [28,29]. In addition, these studies had limitations because more structured interventions, such as CBT, were not included. A previous

XSL•FO
RenderX

meta-analysis showed that internet-based rehabilitation programs may improve pain but not physical function in patients with knee OA [30]. The meta-analysis investigated IBTH programs with multiple structured interventions for knee OA, but the total sample size was too small to allow definite judgments [30]. In addition, further large studies on IBTH for hip and knee OA have been published. Therefore, the purpose of this systematic review and meta-analysis was to evaluate the effectiveness of IBTH programs in function, pain, and self-efficacy in patients with hip or knee OA.

Methods

Search Strategy

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, as outlined in Multimedia Appendix 1 [31]. The priori protocol of this study was reviewed and registered in the International Prospective Register of Systematic Reviews (PROSPERO; CRD42024541111) in May 2024. Two review raters (authors HNW and PL) independently performed a systematic literature search to identify relevant studies. The following databases were searched for papers published between January 1, 2000, and May 3, 2024: PubMed, EMBASE, Web of Science, SPORTDiscus (EBSCO), PEDro, and CINAHL. The search string was divided into 3 sections: the first included synonyms for telehealth, the second consisted of synonyms for the knee or the hip joint, and the third pertained to OA. To ensure that at least 1 search term from each section was included in the results, all synonyms within sections were connected with the “OR” operator, and different sections were connected with the “AND” operator. The search strategies for databases are presented in Multimedia Appendix 2.

Inclusion and Exclusion Criteria

All studies were screened and assessed for eligibility based on the PICOS (Population, Intervention, Comparison, Outcome, and Study) method, in accordance with our specific inclusion and exclusion criteria. Studies were eligible for inclusion if (1) their subjects were diagnosed with OA; (2) they allowed comparisons between telerehabilitation and face-to-face rehabilitation, usual care, waitlisting, and placebo; (3) they contained at least 1 of any objective measures of function, pain, or self-efficacy as the outcome; or (4) they were randomized or quasi-randomized controlled trials (RCTs). Studies were excluded if (1) they were reviews, case reports, conference reports, or observational investigations; (2) participants had undergone a surgical procedure or experienced lower limb trauma; (3) studies or data were duplicated; or (4) studies were published in a language other than English.

Data Selection and Extraction

Research results were imported into EndNote version 20 (Clarivate Analytics) to remove duplications and facilitate selection. Two review raters (PL and HNW) independently screened the titles and abstracts retrieved using the search strategy. The full texts of all studies considered potentially eligible for inclusion were subsequently retrieved and read independently by the 2 review raters, who then made the final selection. Any discrepancies were resolved through discussion and, when necessary, by the involvement of a third independent rater (author XZ). Data were independently extracted from the included studies by the 2 review raters (HNW and PL), including author, year, patient characteristics, intervention characteristics, duration of treatment, outcomes, and time points, and the extracted variables were revised and checked for accuracy by the 2 review raters. In this study, we analyzed the data at endpoints that measured immediately after the IBTH intervention. In the case of incomplete data from a published paper, we contacted the corresponding author for the raw data of trials.

Risk of Bias

The PEDro scale was used to evaluate the quality and bias of the studies in this research. [32]. The PEDro scale, which evaluates study eligibility criteria, random allocation, concealed allocation, baseline comparability, blinded subjects, blinded therapists, blinded assessors, adequate follow-up, intention-to-treat (ITT) analysis, between-group comparisons, and point estimates and variability, was considered valid and reliable [33]. Where available, scores were taken directly from the PEDro database [34]. If scores were unavailable, 2 independent reviewers (HNW and PL) assessed the quality of the papers using the PEDro scale. Any disagreements were resolved by discussion with a third reviewer (XZ).

The revised Cochrane Collaboration’s Risk of Bias version 2 (RoB 2) tool for assessing risk of bias in randomized clinical trials was also used to assess the risk of bias in the included studies by 2 independent authors (HNW and PL). The RoB 2 tool contains algorithms that map responses to signaling questions regarding a proposed risk-of-bias judgment for each outcome assessed in a given study [35]. Therefore, assessment criteria were divided into 5 domains: risk of bias from randomization process, bias due to deviations from intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and risk of bias in selection of the reported result. The risk-of-bias judgment for each of the 5 domains was classified as “low risk of bias,” “some concerns,” or “high risk of bias”. The overall risk of bias on a study level was determined according to the classification of the assessment criteria domains, following guidelines from the RoB 2 tool. Any disagreements were resolved by discussion with a third reviewer (XZ).

Quality of Intervention Reporting

The Template for Intervention Description and Replication (TIDieR) telehealth checklist was used to evaluate the quality of intervention reporting [36]. This checklist was specifically developed for reporting on telehealth interventions evaluated in clinical trials. The checklist consists of 12 items that provide clear and reproducible descriptions of telehealth interventions. The items include a brief name, the rationale (why), materials (what), the procedure (what), the provider (who), the delivery method (how), the location (where), intervention parameters (when and how much), tailoring, modifications, and adherence (how well, both planned and actual):

XSL•FO
RenderX

- Item 1 (brief name): Provide a name or a phrase that describes the intervention.
- Item 2 (why): Describe any rationale, theory, or goal of the elements essential to the intervention.
- Item 3 (what materials): Describe any physical or informational materials used in the intervention, including those provided to participants or used in intervention delivery or in the training of intervention providers. Provide information about where the materials can be accessed.
- Item 4 (what procedures): Describe each of the procedures, activities, or processes used in the intervention, including any enabling or support activities.
- Item 5 (who provided): For each category of intervention provider (eg, psychologist, nursing assistant), describe their expertise, background, and any specific training given.
- Item 6 (how): Describe the modes of delivery (eg, face to face, internet, or telephone) of the intervention and whether it was provided individually or in a group.
- Item 7 (where): Describe the type(s) of location(s) where the intervention occurred, including any necessary infrastructure or relevant features.
- Item 8 (when and how much): Describe the number of times the intervention was delivered and over what period, including the number of sessions; their schedule; and their duration, intensity, or dose.
- Item 9 (tailoring): If the intervention was planned to be personalized, titrated, or adapted, describe what, why, when, and how.
- Item 10 (modifications): If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).
- Item 11 (how well planned): If intervention adherence or fidelity was assessed, describe how and by whom. If any strategies were used to maintain or improve fidelity, describe them.
- Item 12 (how well: actual): If intervention adherence or fidelity was assessed, describe the extent to which the intervention was delivered as planned.

Data Synthesis

This meta-analysis was conducted using R version 4.2.2 (R Foundation for Statistical Computing) with the *meta* and *Robvis* packages. [37]. The standardized mean difference (SMD) with a 95% CI was used to calculate all continuous data. The effect size was categorized as small (<0.20), moderate (0.21-0.79), or large (>0.80) according to Cohen criteria [38,39]. The Higgins statistic (I^2) was calculated to evaluate heterogeneity. When the number of included studies was 5 or fewer, the fixed-effects model was applied [40]. When the number of included studies was more than 5, the random-effects model was applied if substantial heterogeneity (I^2>50% or P<.05) existed; otherwise, the fixed model was used [41]. The significance level was set at P<.05. Sensitivity analysis was performed to evaluate the quality and consistency of results by sequentially omitting each study. In addition, publication bias was assessed using funnel plots and Egger tests [42]. The kappa test was used to evaluate the consistency between raters. In addition to the assessment of the overall effect, subgroup analyses were conducted for the type of IBTH intervention, and the interventions were grouped into 3 clusters: (1) exercise, (2) CBT, and (3) mixed. Moreover, subgroup analyses were conducted for the location of OA, and the interventions were grouped into 3 clusters: (1) hip, (2) knee, and (3) hip/knee.

Quality Assessment

The Grades of Recommendation Assessment, Development, and Evaluation (GRADE) method was used to evaluate the quality of evidence [43]. GRADE assesses the quality of evidence based on the risk of bias, indirectness, inconsistency, imprecision, and the risk of publication bias. The quality of evidence is graded as high, moderate, low, or very low, based on satisfying the following criteria: risk of bias (downgraded if average PEDro scores across studies<7), inconsistency (downgraded if I^2>50 or a single study with N<300), indirectness (downgraded if heterogeneous population or intervention), and imprecision (downgraded if CI>0.25 in either direction or a single study with N<300).

Results

Search Results

The process of paper screening for this study is depicted in Figure 1. A total of 1154 records were identified in the initial search. After removing duplicates, the screening of 886 (76.8%) titles and abstracts retrieved 795 (68.9%) records that did not meet the eligibility criteria. Subsequently, 91 (10.3%) full-text papers were assessed for eligibility. Of these, assessors agreed to include 20 (22%) papers, agreed to exclude 68 (74.7%) papers, and disagreed on whether 3 (3.3%) papers should be included (κ=0.91), indicating good consistency between the 2 raters. Finally, 1 (33.0%) paper was included and 2 (66.7%) were excluded after arbitration by the third rater. This resulted in 21 (44.3%) papers being included in this study.

XSL•FO
RenderX

Figure 1. Flow diagram of the literature search and selection following PRISMA. IBTH: internet-based telehealth; OA: osteoarthritis; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT: randomized controlled trial.

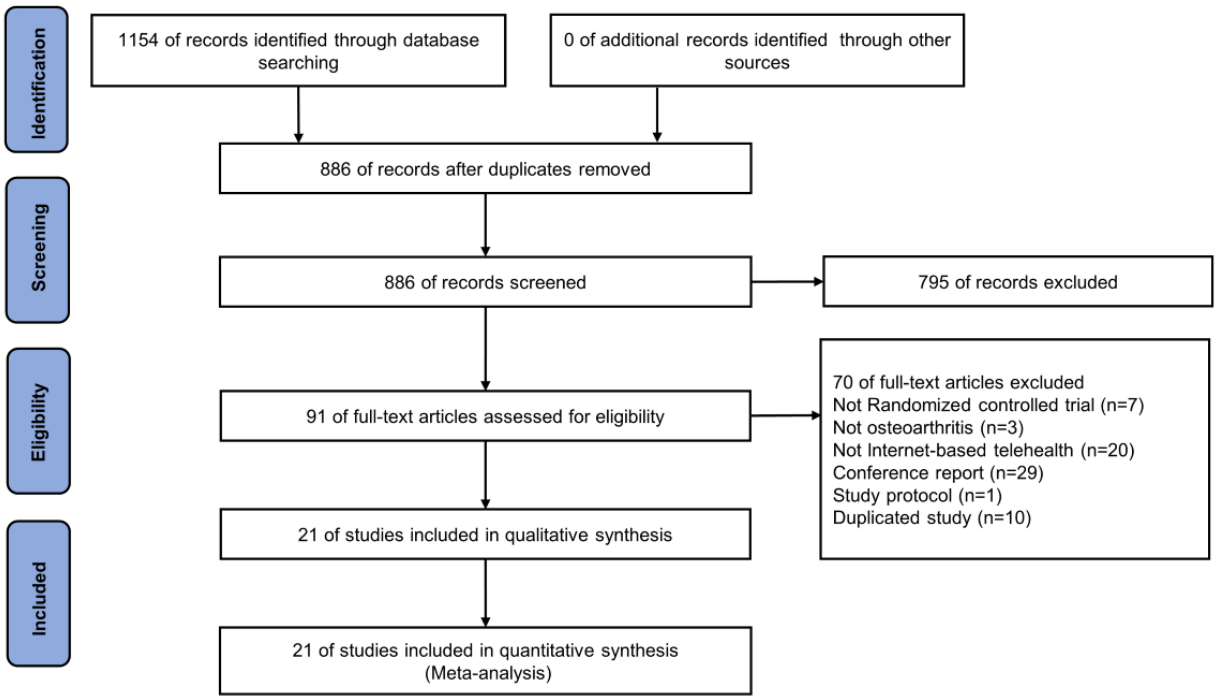

Study Characteristics

A total of 21 papers [27,44-63] with 3186 participants were included to evaluate the effectiveness of IBTH in patients with hip or knee OA. The descriptive characteristics of these included studies are summarized in Table 1. All 21 studies were RCTs published between 2013 and 2024. Of these 21 studies, 6 (28.6%) were conducted in Australia, 4 (19%) in the United States, 4 (19%) in the Netherlands, 2 (9.5%) in Turkey, 1 (4.8%) in the United Kingdom, 1 (4.8%) in Brazil, 1 (4.8%) in Greece, 1 (4.8%) in Saudi Arabia, and 1 (4.8%) in Thailand. In addition, 16 (76.2%) trials included patients with knee OA (n=2149, 67.5%), 1 (4.8%) trial included patients with hip OA (n=144, 4.5%), and 4 (19%) trials included patients with hip and knee OA (n=893, 28%). The age of the patients ranged from 55 to 68.5 years. The intervention duration ranged from 6 weeks to 6 months. Furthermore, 18 (85.7%) studies reported short-term (≤3 months) follow-up, while 3 (14.3%) studies reported medium-term (3-6 months) follow-up.

XSL•FO
RenderX

Table 1. Summary of the characteristics of studies included in the meta-analysis (N=21).

Author, country	Location of OA[a] (sample size, n)	Age (years); female/male patients	Comparison	Intervention	Assessment point	Outcomes
Nelligan et al, Australia [27]	Knee (EG[b]: 103; CG[c]: 103)	EG: 60.3; 66/32 CG: 59.0; 60/33	CG: educational information	EG: web-based home exercise plus SMS	6 months	Function: WOMAC[d] Pain: KOOS[e] Self-efficacy: ASES[f]
Aily et al, Brazil [44]	Knee (EG: 50; CG:50)	EG: 53; 30/20 CG: 55; 30/20	CG: circuit training	EG: internet-based circuit training	14 weeks	Function: WOMAC Pain: VAS[g]
Alasfour and Almarwani, Saudi Arabia [45]	Knee (EG: 20; CG:20)	EG: 54; 20/0 CG: 55; 20/0	CG: exercise	EG: app-based exercise	6 weeks	Function: WOMAC Pain: VAS
Allen et al, United States [46]	Knee (EG: 142; CG1: 140; CG2: 68)	EG: 65.3; 98/44 CG1:65.7; 100/40 CG2: 64.2; 53/9	CG1: physical therapy CG2: waitlisting	EG: internet-based exercise	4 months	Function: WOMAC Pain: WOMAC
Allen et al, United States [47]	Knee (EG: 230; CG: 115)	EG: 59.9; 36/224 CG: 60.2; 17/98	CG: education	EG: internet-based exercise plus telephone counseling	3 months	Function: WOMAC Pain: WOMAC
Bennell et al, Australia [48]	Hip OA (EG: 73; CG: 71)	EG: 61.2; 45/28 CG: 61.3; 37/34	CG: education	EG: pain-coping skill training	8 weeks	Function: WOMAC Pain: WOMAC Self-efficacy: ASES
Bennell et al, Australia [49]	Knee (EG1: 172; EG2: 175; CG: 67)	EG1: 65.4; 93/79 EG2: 64.1; 89/86 CG:65.3; 45/22	CG: education	EG1: telehealth-delivered exercise EG2: telehealth-delivered diet and exercise	6 months	Function: WOMAC Pain: NRS[h]
Bennell et al, Australia [50]	Knee (EG: 107; CG: 105)	EG: 62.8; 70/37 CG:61.8; 78/27	CG: education	EG: online yoga program	12 weeks	Function: WOMAC Pain: NRS Self-efficacy: ASES
Bossen et al, Netherlands [51]	Hip/knee OA (EG: 100; CG: 99)	EG: 61; 60/40 CG: 63; 69/30	CG: waitlisting	EG: web-based physical activity	3 months	Function: KOOS Pain: KOOS Self-efficacy: ASES
Gohir et al, United Kingdom [52]	Knee (EG: 48; CG: 57)	EG:65.2; 34/14 CG: 68.0; 37/20	CG: usual care	EG: digital exercise program	6 weeks	Function: WOMAC Pain: WOMAC
Hunter et al, Australia [53]	Knee (EG: 112; CG: 105)	EG: 63; 71/41 CG: 66; 60/45	CG: usual care	EG: online exercise and CBT[i]	6 months	Function: KOOS ADL[j] Pain: NRS
Kloek et al, Netherlands [54]	Hip/knee (EG: 99; CG: 109)	EG: 63.8; 74/35 CG: 62.3; 67/32	CG: usual care	EG: e-exercise	3 months	Function: KOOS/HOOS[k] ADL Pain: KOOS/HOOS
Moutzouri et al, Greece [55]	Knee (EG: 22; CG: 22)	EG: 65.1; 19/3 CG: 63.5; 17/5	CG: usual care	CG: web-based exercise	6 weeks	Function: KOOS ADL Pain: KOOS
Murphy et al, United States [56]	Knee (EG: 31; CG:15)	EG: 65.1; 19/3 CG: 63.5; 17/5	CG: usual care	CG: online CBT	6 weeks	Function: WOMAC
O'Moore et al, Australia [57]	Knee (EG:49; CG: 20)	EG:63.2; 38/6 CG: 59.68; 17/8	CG: usual care	EG: internet CBT	3 months	Function: WOMAC Pain: WOMAC Self-efficacy: ASES
Pelle et al, Netherlands [58]	Hip/knee (EG: 214; CG: 213)	EG: 62.1; 147/67 CG: 62.1; 159/54	CG: usual care	EG: self-management app with exercise	3 months	Function: KOOS/HOOS ADL Pain: KOOS/HOOS

XSL•FO
RenderX

Author, country	Location of OA[a] (sample size, n)	Age (years); female/male patients	Comparison	Intervention	Assessment point	Outcomes
Rini et al, United States [59]	Knee (EG:58; CG: 55)	EG:68.5; 46/12 CG: 66.7; 45/10	CG: waitlisting	EG: pain-coping skill training	8-10 weeks	Function: AIMS2[l] Pain: AIMS2 Self-efficacy: ASES
Thiengwittayaporn et al, Thailand [60]	Knee (EG:44; CG: 45)	EG: 62.2; 36/6 CG: 63.0; 37/3	CG: conventional education	EG: mobile app–based self-directed exercise guidance	4 weeks	Function: KOOS ADL Pain: KOOS
Tore et al, Turkey [61]	Knee (EG:24; CG: 24)	EG: 55.9; 21/3 CG: 55.8; 22/2	CG: exercise	EG: exercise via video-conferencing	8 weeks	Function: KOOS ADL Pain: KOOS
Tümtürk et al, Turkey [62]	Knee (EG:29; CG: 28)	EG: 53.6; 24/5 CG: 51.5; 16/12	CG: conventional exercise and education	EG: telerehabilitation-based exercise and education	8 weeks	Function: WOMAC Pain: WOMAC
Weber et al, Netherlands [63]	Hip/knee (EG:29; CG: 28)	EG: 53.6; 24/5 CG: 51.5; 16/12	CG: usual care	EG: app-based exercise, physical activity, and education program	12 weeks	Function: KOOS/HOOS ADL Pain: KOOS/HOOS

[a]OA: osteoarthritis.
[b]EG: experimental group.
[c]CG: control group.
[d]WOMAC: Western Ontario and McMaster Universities Osteoarthritis index.
[e]KOOS: Knee Injury and Osteoarthritis Outcome Score.
[f]ASES: Arthritis Self-Efficacy Scale
[g]VAS: Visual Analogue Scale.
[h]NRS: Numeric Rating Scale.
[i]CBT: cognitive-behavioral therapy.
[j]ADL: activities of daily living.
[k]HOOS: Hip Disability and Osteoarthritis Outcome Score.
[l]AIMS2: Arthritis Impact Measurement Scale 2.

The content of the intervention showed significant variation among the pooled studies. As shown in Table 1, a total of 15 (71.4%) studies used interventions that included an internet-based exercise program [27,46,47,49-55,58,60-63], 4 (19%) studies implemented CBT interventions [48,56,57,59], and 1 (4.8%) study used a combination of an exercise program and a CBT intervention [53].

The TIDieR telehealth checklist was assessed for the included studies (Figure 2 and Multimedia Appendix 3). A total of 7(33.3%) studies fulfilled all 12 requirements for the description of a telehealth intervention [46,47,51,52,54,58,60,62]. Only 1 (4.8%) study poorly described the location (item 7) fulfilling the checklist requirements [48]. Around half of the included studies (n=10, 47.6%) [27,44,45,48,50,55-57,59,61] adequately described the tailoring process. All studies reported the actual extent of fidelity for the participants in the intervention. However, 4 (19%) studies [47,53,56,63] inadequately detailed how they planned to maintain or improve fidelity, while 5 (23.8%) studies [47,49,53,56,63] assessed intervention adherence or fidelity. Other items were reported satisfactorily.

XSL•FO
RenderX

Figure 2. Overview of studies reporting intervention information based on the TIDieR-telehealth checklist. TIDieR: Template for Intervention Description and Replication.

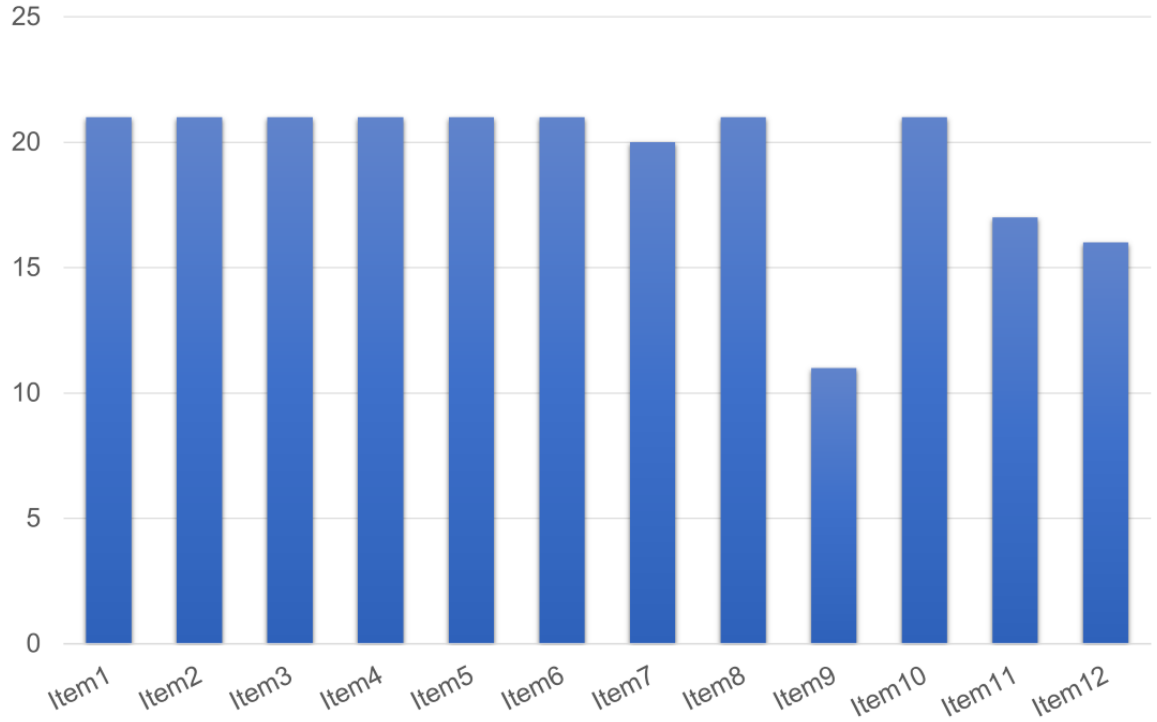

Assessment of Risk of Bias

The risk of bias was assessed for all selected papers on the PEDro scale, and all 21 studies included in the quantitative analysis scored greater than 5, with a maximum score of 10 (Table 2). All included trials used the random allocation method in the study design and comparable outcomes at baseline. A total of 6 (28.6%) trials failed to use allocation concealment [47,54,56,60-62], 2 (9.5%) studies successfully blinded the participants during the trials [48,57], and only 1 (4.8%) study blinded the participants and therapist during the intervention [48], perhaps primarily owing to the characteristics of the interventions. In addition, 11 (52.4%) studies had inadequate blinding of the outcome assessor [27,45,49-52,54,55,58,61,62]. In terms of follow-up, 7 (33.3%) studies did not complete adequate follow-up measurements due to a high attrition rate (>15%) [47,51,52,54,56-58], while 4 (19%) studies did not perform ITT analysis [45,55,56,61]. All studies conducted group comparisons and provided point estimates and variability.

The methodological quality of the eligible studies was also rated using RoB 2 to assess subjective bias (Figure 3). The overall methodological quality of the studies included in the review was mixed. All studies used methods that we judged to have a low risk of bias to randomly assign participants to either the intervention or the control group. This result was due to the selection criteria for RCTs. Thus, RCTs prevented selection bias and were insured against accidental bias. However, 6 (28.6%) trials did not achieve allocation concealment, which contributed to the high risk of bias in the randomization process [47,54,56,60-62]. A potentially important source of bias in this meta-analysis was the methodology of blinding. Only 1 (4.8%) study could successfully blind the participants and therapists [48]. Around half of the studies (n=10, 47.6%) adopted blinding for the outcome assessor [44,46-48,53,56,57,59,60,63]. Hence, a high risk of measurement of outcomes remained possible for outcomes relying on self-report or objective outcomes by outcome assessors who were not blinded to treatment allocation. The majority of the trials (n=17, 81%) used ITT analysis [27,44,46-54,57-60,62,63], which is an analysis method for solving noncompliance and missing outcomes.

XSL•FO
RenderX

Table 2. Methodological classification assessed by the PEDro scale.

Study	Item 2[a]	Item 3[b]	Item 4[c]	Item 5[d]	Item 6[e]	Item 7[f]	Item 8[g]	Item 9[h]	Item 10[i]	Item 11[j]	Total score
Nelligan et al [27]	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	7
Aily et al [44]	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	8
Alasfour and Almarwani [45]	Yes	Yes	No	No	No	No	Yes	No	Yes	Yes	5
Allen et al [46]	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	8
Allen et al [47]	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	7
Bennell et al [48]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	10
Bennell et al [49]	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	7
Bennell et al [50]	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	7
Bossen et al [51]	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	6
Gohir et al [52]	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	7
Hunter et al [53]	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	8
Kloek et al [54]	Yes	No	Yes	No	No	No	No	Yes	Yes	Yes	5
Moutzouri et al [55]	Yes	Yes	Yes	No	No	No	Yes	No	Yes	Yes	6
Murphy et al [56]	Yes	No	Yes	No	No	Yes	No	No	Yes	Yes	5
O'Moore et al [57]	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	7
Pelle et al [58]	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	6
Rini et al [59]	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	8
Thiengwittayaporn et al [60]	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	7
Tore et al [61]	Yes	No	Yes	No	No	No	Yes	No	Yes	Yes	5
Tümtürk et al [62]	Yes	No	Yes	No	No	No	Yes	Yes	Yes	Yes	6
Weber et al [63]	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	8

[a]Item 2: Subjects were randomly allocated to groups.

[b]Item 3: Allocation was concealed.

[c]Item 4: Groups were similar at baseline regarding the most important prognostic indicators.

[d]Item 5: There was blinding of all subjects.

[e]Item 6: There was blinding of all therapists who administered the therapy.

[f]Item 7: There was blinding of all assessors who measured at least 1 key outcome.

[g]Item 8: Measures of at least 1 key outcome were obtained from more than 85% of the subjects initially allocated to groups.

[h]Item 9: All subjects for whom outcome measures were available received the treatment or control condition, as allocated, or where this was not the case, data for at least 1 key outcome was analyzed by the intention to treat (ITT).

[i]Item 10: The results of between-group statistical comparisons were reported for at least 1 key outcome.

[j]Item 11: The study provided both point measures and measures of variability for at least 1 key outcome.

XSL•FO
RenderX

Figure 3. Summary of risk of bias. (A) Risk of bias presented using the Cochrane RoB 2 tool for RCTs and (B) risk of bias as percentages across all included studies. RCT: randomized controlled trial; RoB 2: Risk of Bias version 2.

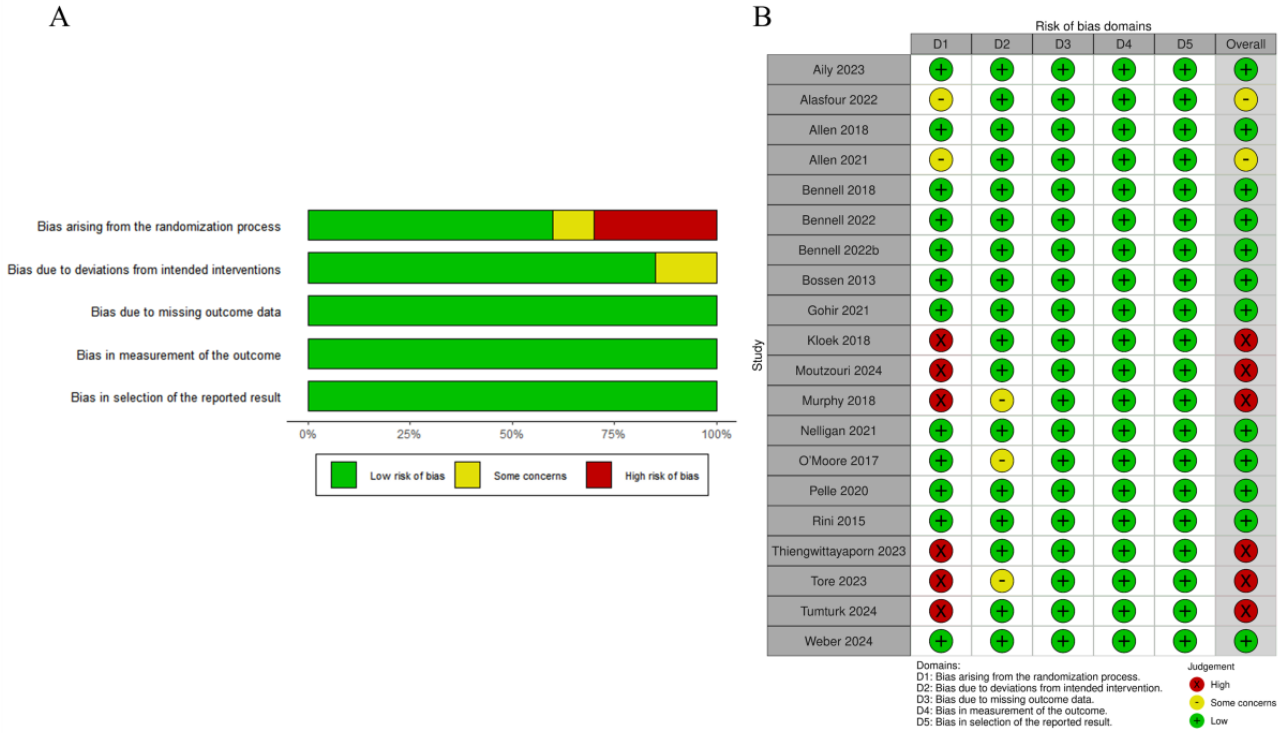

Outcome Measures

Primary Outcome

Function

All 21 studies [27,44-63] evaluated the effects of IBTH compared to control groups. Significant heterogeneity was detected among the included trials (I^2=44%, P=.02). Therefore, a random-effects model was used. Participants with hip or knee OA in the IBTH groups had higher improvement in function (SMD 0.30, 95% CI 0.23-0.37, P<.001), as shown in Figure 4. The GRADE scale established that the certainty of evidence for function comparisons was low (Multimedia Appendix 4). There was no significant subgroup difference in improving function based on the intervention and the period of follow-up. There was no significant difference between IBTH with CBT and the control group concerning improvement of function (SMD 0.02, 95% CI –0.01 to 0.40, I^2=0). In addition, IBTH was not effective in improving function in studies on patients with mixed hip and knee OA (Multimedia Appendix 5).

XSL•FO
RenderX

Figure 4. Forrest plots of comparisons of outcomes between IBTH and control groups. (A) Function, (B) pain, and (C) self-efficacy. IBTH: internet-based telehealth; SMD: standardized mean difference.

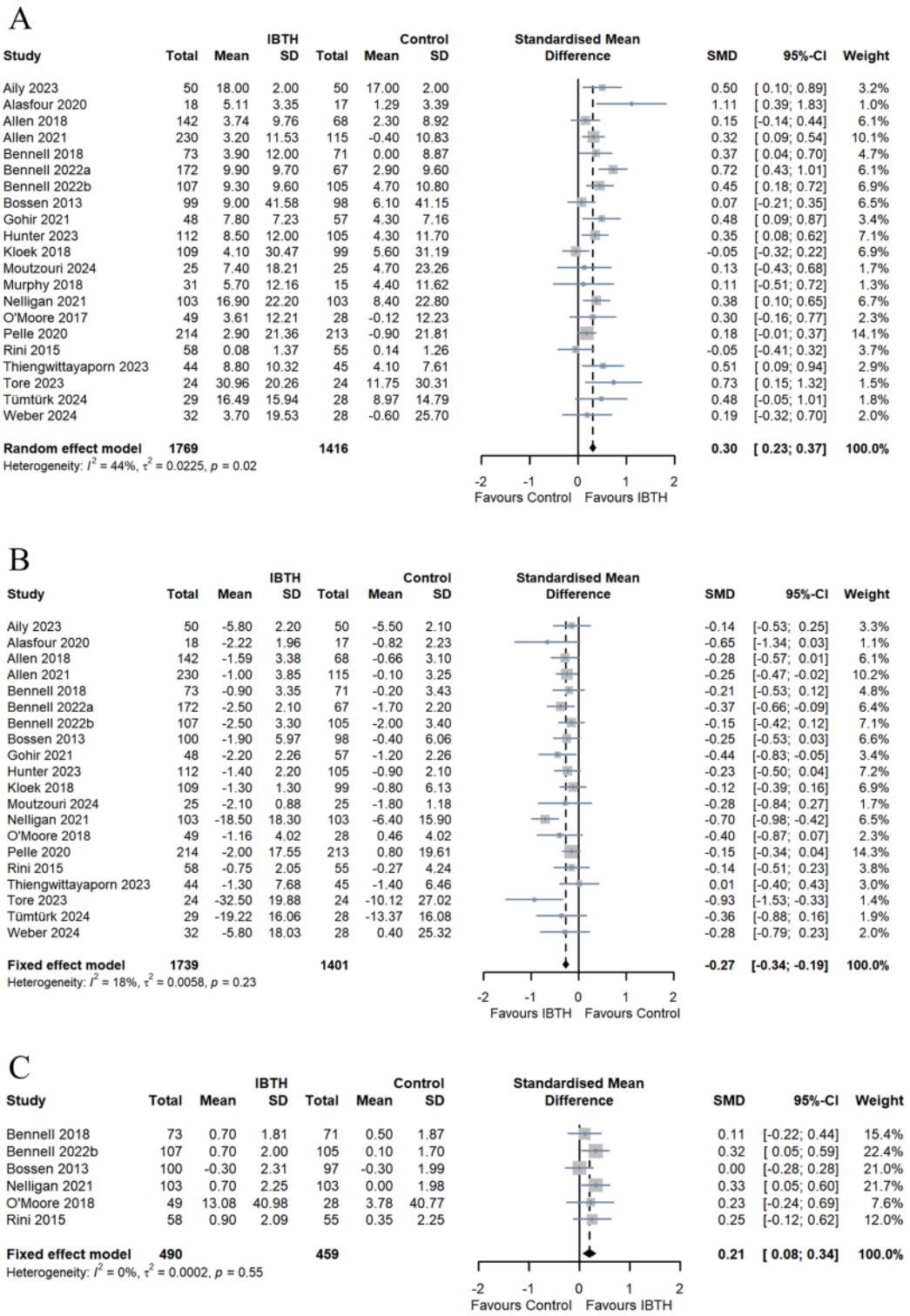

Secondary Outcomes

Pain

A total of 20 (95.2%) studies [27,44-55,57-63] provided data on the pain intensity for IBTH compared to control groups. There was substantial heterogeneity in the pooled studies (I^2=18%, P=.23). Hence, a fixed-effects model was adopted. Participants with hip or knee OA in the IBTH groups had a better effect of pain relief than those in the control groups (SMD –0.27, 95% CI –0.34 to –0.19, P<.001), as shown in Figure 4. The certainty of GRADE evidence for pain was low (Multimedia Appendix 4). Subgroup analyses indicated there was no significant subgroup difference in pain relief based on the intervention, location, and period of follow-up (Multimedia Appendix 5).

XSL•FO
RenderX

Self-Efficacy

Of the 21 studies, 6 (28.6%) [27,48,50,51,57,59] reported outcomes of self-efficacy in both IBTH and control groups. A fixed-effects model was applied in the meta-analysis due to low heterogeneity (I^2=0, P=.55). The results indicated that the IBTH groups showed significant improvement in self-efficacy (SMD 0.21, 95% CI 0.08-0.34, P<.001) compared to the control groups (Figure 4). The certainty of GRADE evidence for self-efficacy was assessed as high (Multimedia Appendix 4). There was no significant subgroup difference in increasing self-efficacy based on the intervention and the period of follow-up. However, IBTH with CBT significantly improved the self-efficacy of patients with hip or knee OA (SMD 0.18, 95% CI –0.03 to 0.40, I^2=0; Multimedia Appendix 5).

Sensitivity Analysis

Sensitivity analysis, involving the sequential omission of individual studies, was conducted to evaluate the impact of each study on the overall outcomes. The removal of any specific study did not alter the significance or substantial heterogeneity of the aggregated results pertaining to function, pain, and self-efficacy. Consequently, the sensitivity analysis affirmed that the findings related to significance and heterogeneity were robust.

Publication Bias

Funnel plots and Egger tests were performed to evaluate publication bias, as illustrated in Multimedia Appendix 6. The funnel plots of the outcomes indicated a potential publication bias in small trials, while Egger tests found that such bias was negligible (P>.05) in this meta-analysis.

Discussion

Principal Findings

In this meta-analysis, we included 21 studies to evaluate the effects of IBTH programs on patients with hip or knee OA. Low-quality evidence was found to suggest that IBTH programs are more effective in relieving pain and improving function compared to the control group. High-quality evidence showed that IBTH programs have greater efficacy in promoting self-efficacy than control interventions. Subgroup analyses indicated that IBTH (either exercise or CBT) is beneficial in reducing pain intensity compared to control interventions. However, IBTH with CBT may improve function and self-efficacy. The role of telehealth in knee OA therapy was consistent with the results of the primary analysis, but for hip or knee/hip OA, results were inconsistent. The results of this systematic review and meta-analysis support the effectiveness of IBTH programs in improving clinical outcomes in patients with hip or knee OA.

The effectiveness of telehealth programs in lower extremity OA has been the subject of investigation in previous studies. Wang et al [64] investigated the effectiveness of telehealth interventions compared with usual care or no intervention in patients with OA who had undergone total knee replacement (TKR) or total hip replacement (THR). They found that telehealth interventions could improve pain and functional mobility in patients who underwent TKR but had no effect on patients undergoing THR. Two other systematic reviews with meta-analysis addressed a similar question, focusing on the effectiveness of telehealth-based exercise [29,65]. The findings showed that telehealth-based exercise is beneficial in reducing pain and improving function in knee OA. Yang et al [65] found that a 3-month telehealth-based exercise intervention using web and smartphone apps was associated with better pain relief and physical function. This finding is consistent with our subgroup analysis for IBTH with exercise for patients with hip or knee OA. Xie et al [30] investigated the effectiveness of internet-based rehabilitation programs specifically in patients with knee OA. The authors concluded that internet-based rehabilitation programs are effective in reducing pain but not improving the function of patients with knee OA. These differences may be explained by differences in inclusion criteria. The most important differences in eligibility criteria were that we included populations with both hip OA and knee OA, while Xie et al [30] focused only on the latter. We also strictly excluded studies published in conference abstracts because of the absence of full text and complete data, whereas they did not. In addition, only 1 study of IBTH for hip OA was included in this study, so more studies on hip OA are needed in the future for more robust results.

Subgroup analyses were performed to assess the impact of disparate intervention methods in IBTH. The results of our subgroup analyses demonstrated that IBTH with exercise can alleviate pain and improve function and self-efficacy. The effects of exercise on symptoms and function have been well studied in hip and knee OA [66,67]. IBTH with exercise may help with OA symptoms due to the reduced excitability of the motor cortex after exercise, which leads to a decrease in motor evoked potentials, thus reducing pain [68]. IBTH integrated with exercise contributes to an increase in muscle strength, a decrease in extension impairments, and an improvement in proprioception, leading to an improvement in the function of patients with OA [69]. In addition, improved self-efficacy results from exercise can also relieve pain and function. Self-efficacy refers to the belief in one's ability to successfully perform a behavior to achieve a desired outcome [70]. A previous study has shown that patients with higher levels of arthritis self-efficacy tend to perceive pain stimuli as less unpleasant and have greater pain tolerance [71]. Somers et al [72] found that patients with OA who possess higher arthritis self-efficacy may be better able to tolerate pain. This increased tolerance may increase their likelihood of engaging in weight management behaviors, such as exercise, that are beneficial for improving function. Our findings show emerging evidence that IBTH with exercise results in greater pain reduction, which is consistent with previous studies. In addition to exercise, our subgroup analyses also included the intervention method of CBT. In this systematic review, the included studies on IBTH with CBT were intended to help patients with OA manage pain by self-regulating thoughts, feelings, and behaviors [73]. These interventions delivered through face-to-face methods are effective in managing OA pain [11,74]. Internet-delivered CBT is effective in multiple pain management outcomes, including disability and depression [57,75]. However, we found that IBTH with CBT is ineffective in improving function. One possible

XSL•FO
RenderX

explanation is that functional impairment is related to joint degeneration and injury. CBT is a psychological intervention beneficial for pain management, but it has little direct effect on improving physical function. Therefore, IBTH with exercise may be more effective in improving function, and the incorporation of exercise protocols into IBTH programs is recommended.

Our study highlights the potential nonpharmacological treatment strategies that can be delivered via the IBTH method. Nonpharmacological treatment strategies, which serve as the core treatment, are among the most common treatments for patients with OA. Compared with conventional face-to-face delivery methods, treatment delivered via IBTH offers advantages such as lower time consumption, convenience, and lower cost [76], potentially improving adherence to the intervention. These strategies include internet automated/self-paced therapeutic exercise programs, web-based programs to promote daily physical activity, internet-based pain-coping skill training, and internet-based CBT programs. In particular, exercise programs in IBTH allow for individualization and progression [46]. Gohir et al [52] established the progression of exercises during the intervention, adjusting them according to complexity, load, and difficulty in relation to each participant's response after completing the exercise.

Although telehealth for patients with OA is now well established, future studies can still focus on areas such as remote assessment, remote monitoring, and the integration of novel technologies into IBTH programs. Remote physiotherapy assessment of musculoskeletal disorders via IBTH has been proven feasible and reliable [77], including assessments of the knee [78] and the hip complex [79]. However, remote assessment is more appropriate for follow-up sessions rather than initial consultations, as physical examinations, blood sampling, and radiographic imaging may be required to make a definitive diagnosis in some cases [80]. Monitoring a patient's health status and interventions is key to a supervised telehealth system. Self-managed telehealth can be facilitated by monitoring health conditions with wearable devices, including measurements such as continuous blood pressure, electrocardiogram, oxygen saturation, and body temperature [81]. For internet-delivered exercise, older patients can use posture and motion monitoring technologies, such as depth cameras and wearable systems, to improve the quality of movement, ensure the effectiveness of exercise, and prevent injury [82-84]. As technology has evolved, the benefits of virtual reality (VR) have been demonstrated across a wide range of conditions [85], making VR attractive to patients with musculoskeletal conditions [86]. In VR interventions, patients interact with the virtual environment to simulate real-life activities, and the rehabilitation process can be gamified [87]. Telehealth via VR can increase patient motivation and improve treatment adherence [88]. As the majority of patients with OA are older adults, the usability of VR-based telehealth should be ensured. As more straightforward platforms have been developed, patients only need a computer or a smartphone to connect to VR via the internet [89]. In addition, some telehealth programs are beginning to use automated programs for dynamic intervention. If telehealth can be combined to enable remote assessment, monitoring, and dynamic adjustments based on patient status, it will provide continuous 24/7 monitoring and intervention, leading to significant cost savings [90].

An improvement in the quality of intervention reporting is warranted. Many of the studies included in this review described intervention details poorly, particularly with regard to the location and intervention parameters (eg, frequency and duration) used in the interventions. These deficiencies may make it infeasible for clinicians to reliably replicate or implement the interventions in future research or practice. For telehealth-based exercise, the American College of Sports Medicine (ACSM) recommends prescribing an exercise program with a variety of parameters [91], including frequency, intensity, time, type, volume, pattern, and progression (FITT-VPP). In addition, the Consensus on Exercise Reporting Template (CERT) has been developed to provide guidance for reporting exercise programs [92]. For IBTH with CBT, specific treatment sessions, frequency, and adherence should be reported, especially for unguided treatment [75]. Hence, researchers should consider using these tools, in addition to the TIDieR telehealth checklist [36,93], when designing and reporting telehealth interventions in the future.

Strength and Limitations

This systematic review selected papers from a wide range of electronic databases, increasing the comprehensiveness of our findings. This study extends previous systematic reviews in comparable populations. Our review included more eligible studies, which allowed us to perform subgroup analyses to determine the influence of telehealth interventions and locations of OA. Therefore, this meta-analysis may provide valuable insights for evidence-based practice among clinicians and therapists.

We performed procedural reviews and an extensive database search. However, it is possible that the search missed some papers due to omission of the gray literature and selective reporting bias, which is a potential limitation. The analyses of different interventions in IBTH programs (eg, exercise, behavior therapy, or mixed) were not performed due to the limited number of included studies. Thus, more high-quality RCTs with larger sample sizes are needed, especially for hip OA. In addition, the outcome measures that were pooled and synthesized were subjective measures. Our ability to pool data from objective measures was limited in this review due to the heterogeneous outcomes identified from the included studies. None of the studies included a long-term intervention, making it impossible to evaluate the effects of different durations of IBTH interventions on patients with OA. Finally, the results were limited to pooled homogeneous outcome measures due to the variety of different measures used to assess pain and function in the included studies.

Conclusion

This meta-analysis provided low-to-high quality evidence that IBTH programs have a beneficial effect on improving function, pain, and self-efficacy compared to conventional interventions in patients with hip or knee OA. Limited evidence supports that

XSL•FO
RenderX

IBTH with exercise may significantly alleviate pain and improve function and self-efficacy. Therefore, it is recommended that exercise protocols be incorporated into IBTH programs. More high-quality, large-scale evidence is needed to further investigate the therapeutic effects of IBTH.

Acknowledgments

We thank Dr Tian-Xue Cui, Faculty of Education, University of Macau, for assisting with statistical analysis. This study was supported by the "1.3.5 Project for Disciplines of Excellence," West China Hospital, Sichuan University (grant number ZYGD21005) and the China Scholarship Council (grant number 202106520013). These funding sources were not involved in the literature review, systematic review, meta-analysis, and writing of the report.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[DOCX File , 35 KB - jmir_v26i1e55576_app1.docx]

Multimedia Appendix 2
Search strategy for the systematic review.
[DOCX File , 40 KB - jmir_v26i1e55576_app2.docx]

Multimedia Appendix 3
TIDieR-telehealth checklist for reporting of interventions in the included studies. TIDieR-telehealth: Template for Intervention Description and Replication.
[DOCX File , 30 KB - jmir_v26i1e55576_app3.docx]

Multimedia Appendix 4
Summary of findings on the efficacy and safety of IBTH compared to control group using the GRADE scale. GRADE: Grades of Recommendation Assessment, Development, and Evaluation; IBTH: internet-based telehealth.
[DOCX File , 18 KB - jmir_v26i1e55576_app4.docx]

Multimedia Appendix 5
Summary of subgroup meta-analyses.
[DOCX File , 20 KB - jmir_v26i1e55576_app5.docx]

Multimedia Appendix 6
Funnel plot and *P* value of Egger tests of the IBTH group versus the control group for (A) function, (B) pain, and (C) self-efficacy. IBTH: internet-based telehealth.
[DOCX File , 146 KB - jmir_v26i1e55576_app6.docx]

References

1. Loeser RF, Goldring SR, Scanzello CR, Goldring MB. Osteoarthritis: a disease of the joint as an organ. Arthritis Rheum 2012 Jun 25;64(6):1697-1707 [FREE Full text] [doi: 10.1002/art.34453] [Medline: 22392533]
2. Katz JN, Arant KR, Loeser RF. Diagnosis and treatment of hip and knee osteoarthritis: a review. JAMA 2021 Feb 09;325(6):568-578 [FREE Full text] [doi: 10.1001/jama.2020.22171] [Medline: 33560326]
3. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015 Aug 22;386(9995):743-800 [FREE Full text] [doi: 10.1016/S0140-6736(15)60692-4] [Medline: 26063472]
4. Crawford DC, Miller LE, Block JE. Conservative management of symptomatic knee osteoarthritis: a flawed strategy? Orthop Rev (Pavia) 2013 Feb 22;5(1):e2 [FREE Full text] [doi: 10.4081/or.2013.e2] [Medline: 23705060]
5. Carmona-Terés V, Lumillo-Gutiérrez I, Jodar-Fernández L, Rodriguez-Blanco T, Moix-Queraltó J, Pujol-Ribera E, et al. Effectiveness and cost-effectiveness of a health coaching intervention to improve the lifestyle of patients with knee osteoarthritis: cluster randomized clinical trial. BMC Musculoskelet Disord 2015 Feb 25;16(1):38 [FREE Full text] [doi: 10.1186/s12891-015-0501-x] [Medline: 25887078]

XSL•FO
RenderX

6. Cross M, Smith E, Hoy D, Nolte S, Ackerman I, Fransen M, et al. The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 2014 Jul 19;73(7):1323-1330. [doi: 10.1136/annrheumdis-2013-204763] [Medline: 24553908]
7. Losina E, Paltiel AD, Weinstein AM, Yelin E, Hunter DJ, Chen SP, et al. Lifetime medical costs of knee osteoarthritis management in the United States: impact of extending indications for total knee arthroplasty. Arthr Care Res (Hoboken) 2015 Feb 27;67(2):203-215 [FREE Full text] [doi: 10.1002/acr.22412] [Medline: 25048053]
8. Bannuru R, Osani M, Vaysbrot E, Arden N, Bennell K, Bierma-Zeinstra S, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. Osteoarthr Cartil 2019 Nov;27(11):1578-1589 [FREE Full text] [doi: 10.1016/j.joca.2019.06.011] [Medline: 31278997]
9. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American College of Rheumatology/Arthritis Foundation guideline for the management of osteoarthritis of the hand, hip, and knee. Arthritis Rheumatol 2020 Feb 06;72(2):220-233 [FREE Full text] [doi: 10.1002/art.41142] [Medline: 31908163]
10. Roos E, Juhl C. Osteoarthritis 2012 year in review: rehabilitation and outcomes. Osteoarthr Cartil 2012 Dec;20(12):1477-1483 [FREE Full text] [doi: 10.1016/j.joca.2012.08.028] [Medline: 22960093]
11. Smith MT, Finan PH, Buenaver LF, Robinson M, Haque U, Quain A, et al. Cognitive-behavioral therapy for insomnia in knee osteoarthritis: a randomized, double-blind, active placebo-controlled clinical trial. Arthritis Rheumatol 2015 May;67(5):1221-1233 [FREE Full text] [doi: 10.1002/art.39048] [Medline: 25623343]
12. Chen Y, Klein SL, Garibaldi BT, Li H, Wu C, Osevala NM, et al. Aging in COVID-19: vulnerability, immunity and intervention. Ageing Res Rev 2021 Jan;65:101205 [FREE Full text] [doi: 10.1016/j.arr.2020.101205] [Medline: 33137510]
13. Reschen ME, Bowen J, Novak A, Giles M, Singh S, Lasserson D, et al. Impact of the COVID-19 pandemic on emergency department attendances and acute medical admissions. BMC Emerg Med 2021 Nov 20;21(1):143 [FREE Full text] [doi: 10.1186/s12873-021-00529-w] [Medline: 34800973]
14. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
15. Telehealth. Analysis of third global survey on eHealth based on the reported data by countries, 2016. World Health Organization. 2016. URL: https://www who int/gho/goe/telehealth/en/ [accessed 2020-05-21]
16. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
17. Papandony M, Chou L, Seneviwickrama M, Cicuttini F, Lasserre K, Teichtahl A, et al. Patients' perceived health service needs for osteoarthritis (OA) care: a scoping systematic review. Osteoarthr Cartil 2017 Jul;25(7):1010-1025 [FREE Full text] [doi: 10.1016/j.joca.2017.02.799] [Medline: 28232144]
18. Duckett S. What should primary care look like after the COVID-19 pandemic? Aust J Prim Health 2020;26(3):207. [doi: 10.1071/py20095]
19. Oh J, Park Y, Jo EC, Kim S. Current status and progress of telemedicine in Korea and other countries. Healthc Inform Res 2015 Oct;21(4):239-243 [FREE Full text] [doi: 10.4258/hir.2015.21.4.239] [Medline: 26618029]
20. Ray KN, Chari AV, Engberg J, Bertolet M, Mehrotra A. Disparities in time spent seeking medical care in the United States. JAMA Intern Med 2015 Dec;175(12):1983-1986 [FREE Full text] [doi: 10.1001/jamainternmed.2015.4468] [Medline: 26437386]
21. Cronström A, Nero H, Lohmander LS, Dahlberg LE. On the waiting list for joint replacement for knee osteoarthritis: are first-line treatment recommendations implemented? Osteoarthr Cartil Open 2020 Jun;2(2):100056 [FREE Full text] [doi: 10.1016/j.ocarto.2020.100056] [Medline: 36474591]
22. O'Brien K, Wiggers J, Williams A, Campbell E, Hodder R, Wolfenden L, et al. Telephone-based weight loss support for patients with knee osteoarthritis: a pragmatic randomised controlled trial. Osteoarthr Cartil 2018 Apr;26(4):485-494 [FREE Full text] [doi: 10.1016/j.joca.2018.01.003] [Medline: 29330101]
23. Wantland DJ, Portillo CJ, Holzemer WL, Slaughter R, McGhee EM. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. J Med Internet Res 2004 Nov 10;6(4):e40 [FREE Full text] [doi: 10.2196/jmir.6.4.e40] [Medline: 15631964]
24. Adewale OS. An internet-based telemedicine system in Nigeria. Int J Inf Manag 2004 Jun;24(3):221-234. [doi: 10.1016/j.ijinfomgt.2003.12.014]
25. Agnew JMR, Hanratty CE, McVeigh JG, Nugent C, Kerr DP. An investigation into the use of mHealth in musculoskeletal physiotherapy: scoping review. JMIR Rehabil Assist Technol 2022 Mar 11;9(1):e33609 [FREE Full text] [doi: 10.2196/33609] [Medline: 35275089]
26. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
27. Nelligan RK, Hinman RS, Kasza J, Crofts SJC, Bennell KL. Effects of a self-directed web-based strengthening exercise and physical activity program supported by automated text messages for people with knee osteoarthritis: a randomized

XSL•FO
RenderX

clinical trial. JAMA Intern Med 2021 Jun 01;181(6):776-785 [FREE Full text] [doi: 10.1001/jamainternmed.2021.0991] [Medline: 33843948]
28. McHugh C, Kostic A, Katz J, Losina E. Effectiveness of remote exercise programs in reducing pain for patients with knee osteoarthritis: a systematic review of randomized trials. Osteoarthr Cartil Open 2022 Sep;4(3):100264 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100264] [Medline: 36474946]
29. Chen T, Or CK, Chen J. Effects of technology-supported exercise programs on the knee pain, physical function, and quality of life of individuals with knee osteoarthritis and/or chronic knee pain: a systematic review and meta-analysis of randomized controlled trials. J Am Med Inform Assoc 2021 Feb 15;28(2):414-423 [FREE Full text] [doi: 10.1093/jamia/ocaa282] [Medline: 33236109]
30. Xie S, Wang Q, Wang L, Wang L, Song K, He C. Effect of internet-based rehabilitation programs on improvement of pain and physical function in patients with knee osteoarthritis: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Jan 05;23(1):e21542 [FREE Full text] [doi: 10.2196/21542] [Medline: 33399542]
31. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009 Jul 21;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
32. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther 2003 Aug;83(8):713-721. [Medline: 12882612]
33. Foley NC, Bhogal SK, Teasell RW, Bureau Y, Speechley MR. Estimates of quality and reliability with the physiotherapy evidence-based database scale to assess the methodology of randomized controlled trials of pharmacological and nonpharmacological interventions. Phys Ther 2006 Jun;86(6):817-824. [Medline: 16737407]
34. Welcome to PEDro, the Physiotherapy Evidence Database. PEDro. URL: https://pedro.org.au/ [accessed 2024-09-20]
35. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
36. Rhon DI, Fritz JM, Kerns RD, McGeary DD, Coleman BC, Farrokhi S, et al. TIDieR-telehealth: precision in reporting of telehealth interventions used in clinical trials - unique considerations for the Template for the Intervention Description and Replication (TIDieR) checklist. BMC Med Res Methodol 2022 Jun 02;22(1):161 [FREE Full text] [doi: 10.1186/s12874-022-01640-7] [Medline: 35655144]
37. Schwarzer G. meta: an R package for meta-analysis. R News 2007;7(3):40-45.
38. Dantas LO, Moreira RDFC, Norde FM, Mendes Silva Serrao PR, Alburquerque-Sendín F, Salvini TF. The effects of cryotherapy on pain and function in individuals with knee osteoarthritis: a systematic review of randomized controlled trials. Clin Rehabil 2019 Aug 08;33(8):1310-1319. [doi: 10.1177/0269215519840406] [Medline: 30957514]
39. Aberson CL. Applied Power Analysis for the Behavioral Sciences. London, UK: Routledge; 2019.
40. Tufanaru C, Munn Z, Stephenson M, Aromataris E. Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness. JBI Evid Implement 2015;13(3):196-207. [doi: 10.1097/xeb.0000000000000065]
41. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
42. Green S, Higgins P, Alderson P, Clarke M, Mulrow D, Oxman D. Chapter 8: assessing risk of bias in included studies. Cochrane Handbook for Systematic Reviews of Interventions. 2011. URL: https://tinyurl.com/3x8arf7x [accessed 2024-09-21]
43. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011 Apr;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
44. Aily JB, de Noronha M, Approbato Selistre LF, Ferrari RJ, White DK, Mattiello SM. Face-to-face and telerehabilitation delivery of circuit training have similar benefits and acceptability in patients with knee osteoarthritis: a randomised trial. J Physiother 2023 Oct;69(4):232-239 [FREE Full text] [doi: 10.1016/j.jphys.2023.08.014] [Medline: 37684147]
45. Alasfour M, Almarwani M. The effect of innovative smartphone application on adherence to a home-based exercise programs for female older adults with knee osteoarthritis in Saudi Arabia: a randomized controlled trial. Disabil Rehabil 2022 Jun 25;44(11):2420-2427. [doi: 10.1080/09638288.2020.1836268] [Medline: 33103499]
46. Allen KD, Arbeeva L, Callahan LF, Golightly YM, Goode AP, Heiderscheit BC, et al. Physical therapy vs internet-based exercise training for patients with knee osteoarthritis: results of a randomized controlled trial. Osteoarthr Cartil 2018 Mar;26(3):383-396 [FREE Full text] [doi: 10.1016/j.joca.2017.12.008] [Medline: 29307722]
47. Allen KD, Woolson S, Hoenig HM, Bongiorni D, Byrd J, Caves K, et al. Stepped exercise program for patients with knee osteoarthritis : a randomized controlled trial. Ann Intern Med 2021 Mar;174(3):298-307. [doi: 10.7326/m20-4447]
48. Bennell KL, Nelligan RK, Rini C, Keefe FJ, Kasza J, French S, et al. Effects of internet-based pain coping skills training before home exercise for individuals with hip osteoarthritis (HOPE trial): a randomised controlled trial. Pain 2018 Sep;159(9):1833-1842. [doi: 10.1097/j.pain.0000000000001281] [Medline: 29794609]
49. Bennell KL, Lawford BJ, Keating C, Brown C, Kasza J, Mackenzie D, et al. Comparing video-based, telehealth-delivered exercise and weight loss programs with online education on outcomes of knee osteoarthritis : a randomized trial. Ann Intern Med 2022 Feb;175(2):198-209. [doi: 10.7326/M21-2388] [Medline: 34843383]

XSL•FO
RenderX

50. Bennell KL, Schwartz S, Teo PL, Hawkins S, Mackenzie D, McManus F, et al. Effectiveness of an unsupervised online yoga program on pain and function in people with knee osteoarthritis : a randomized clinical trial. Ann Intern Med 2022 Oct;175(10):1345-1355. [doi: 10.7326/M22-1761] [Medline: 36122378]
51. Bossen D, Veenhof C, Van Beek KE, Spreeuwenberg PM, Dekker J, De Bakker DH. Effectiveness of a web-based physical activity intervention in patients with knee and/or hip osteoarthritis: randomized controlled trial. J Med Internet Res 2013 Nov 22;15(11):e257 [FREE Full text] [doi: 10.2196/jmir.2662] [Medline: 24269911]
52. Gohir SA, Eek F, Kelly A, Abhishek A, Valdes AM. Effectiveness of internet-based exercises aimed at treating knee osteoarthritis: the iBEAT-OA randomized clinical trial. JAMA Netw Open 2021 Feb 01;4(2):e210012 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0012] [Medline: 33620447]
53. Hunter DJ, Bowden JL, Hinman RS, Egerton T, Briggs AM, Bunker SJ, PARTNER Study Team. Effectiveness of a new service delivery model for management of knee osteoarthritis in primary care: a cluster randomized controlled trial. Arthr Care Res (Hoboken) 2023 Jun;75(6):1320-1332 [FREE Full text] [doi: 10.1002/acr.25037] [Medline: 36205225]
54. Kloek CJJ, Bossen D, Spreeuwenberg PM, Dekker J, de Bakker DH, Veenhof C. Effectiveness of a blended physical therapist intervention in people with hip osteoarthritis, knee osteoarthritis, or both: a cluster-randomized controlled trial. Phys Ther 2018 Jul 01;98(7):560-570 [FREE Full text] [doi: 10.1093/ptj/pzy045] [Medline: 29788253]
55. Moutzouri M, Koumantakis GA, Hurley M, Kladouchou AG, Gioftsos G. Effectiveness of a web-guided self-managed telerehabilitation program enhanced with outdoor physical activity on physical function, physical activity levels and pain in patients with knee osteoarthritis: a randomized controlled trial. J Clin Med 2024 Feb 06;13(4):934 [FREE Full text] [doi: 10.3390/jcm13040934] [Medline: 38398248]
56. Murphy SL, Janevic MR, Lee P, Williams DA. Occupational therapist-delivered cognitive-behavioral therapy for knee osteoarthritis: a randomized pilot study. Am J Occup Ther 2018;72(5):7205205040p1-7205205040p9 [FREE Full text] [doi: 10.5014/ajot.2018.027870] [Medline: 30157016]
57. O'Moore KA, Newby JM, Andrews G, Hunter DJ, Bennell K, Smith J, et al. Internet cognitive-behavioral therapy for depression in older adults with knee osteoarthritis: a randomized controlled trial. Arthr Care Res (Hoboken) 2018 Jan;70(1):61-70. [doi: 10.1002/acr.23257] [Medline: 28426917]
58. Pelle T, Bevers K, van der Palen J, van den Hoogen FHJ, van den Ende CHM. Effect of the Dr. Bart application on healthcare use and clinical outcomes in people with osteoarthritis of the knee and/or hip in the Netherlands; a randomized controlled trial. Osteoarthr Cartil 2020 Apr;28(4):418-427 [FREE Full text] [doi: 10.1016/j.joca.2020.02.831] [Medline: 32119972]
59. Rini C, Porter LS, Somers TJ, McKee DC, DeVellis RF, Smith M, et al. Automated internet-based pain coping skills training to manage osteoarthritis pain: a randomized controlled trial. Pain 2015 May;156(5):837-848 [FREE Full text] [doi: 10.1097/j.pain.0000000000000121] [Medline: 25734997]
60. Thiengwittayaporn S, Wattanapreechanon P, Sakon P, Peethong A, Ratisoontorn N, Charoenphandhu N, et al. Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial. Arch Orthop Trauma Surg 2023 Feb;143(2):729-738 [FREE Full text] [doi: 10.1007/s00402-021-04149-8] [Medline: 34453570]
61. Tore NG, Oskay D, Haznedaroglu S. The quality of physiotherapy and rehabilitation program and the effect of telerehabilitation on patients with knee osteoarthritis. Clin Rheumatol 2023 Mar 24;42(3):903-915 [FREE Full text] [doi: 10.1007/s10067-022-06417-3] [Medline: 36279075]
62. Tümtürk İ, Bakırhan S, Özden F, Gültaç E, Kılınç CY. Effect of telerehabilitation-based exercise and education on pain, function, strength, proprioception, and psychosocial parameters in patients with knee osteoarthritis: a randomized controlled clinical trial. Am J Phys Med Rehabil 2024 Mar 01;103(3):222-232. [doi: 10.1097/PHM.0000000000002335] [Medline: 37678215]
63. Weber F, Kloek C, Stuhrmann S, Blum Y, Grüneberg C, Veenhof C. Usability and preliminary effectiveness of an app-based physical activity and education program for people with hip or knee osteoarthritis - a pilot randomized controlled trial. Arthr Res Ther 2024 Apr 10;26(1):83 [FREE Full text] [doi: 10.1186/s13075-024-03291-z] [Medline: 38600607]
64. Wang X, Hunter DJ, Vesentini G, Pozzobon D, Ferreira ML. Technology-assisted rehabilitation following total knee or hip replacement for people with osteoarthritis: a systematic review and meta-analysis. BMC Musculoskelet Disord 2019 Nov 03;20(1):506 [FREE Full text] [doi: 10.1186/s12891-019-2900-x] [Medline: 31679511]
65. Yang Y, Li S, Cai Y, Zhang Q, Ge P, Shang S, et al. Effectiveness of telehealth-based exercise interventions on pain, physical function and quality of life in patients with knee osteoarthritis: a meta-analysis. J Clin Nurs 2023 Jun 25;32(11-12):2505-2520. [doi: 10.1111/jocn.16388] [Medline: 35872635]
66. Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee: a Cochrane systematic review. Br J Sports Med 2015 Dec 24;49(24):1554-1557. [doi: 10.1136/bjsports-2015-095424] [Medline: 26405113]
67. Fransen M, McConnell S, Hernandez-Molina G, Reichenbach S. Exercise for osteoarthritis of the hip. Cochrane Database Syst Rev 2014 Apr 22;2014(4):CD007912 [FREE Full text] [doi: 10.1002/14651858.CD007912.pub2] [Medline: 24756895]
68. Hoeger Bement MK, Weyer A, Hartley S, Yoon T, Hunter SK. Fatiguing exercise attenuates pain-induced corticomotor excitability. Neurosci Lett 2009 Mar 13;452(2):209-213. [doi: 10.1016/j.neulet.2009.01.038] [Medline: 19383441]

XSL•FO
RenderX

69. Runhaar J, Luijsterburg P, Dekker J, Bierma-Zeinstra SMA. Identifying potential working mechanisms behind the positive effects of exercise therapy on pain and function in osteoarthritis; a systematic review. Osteoarthr Cartil 2015 Jul;23(7):1071-1082 [FREE Full text] [doi: 10.1016/j.joca.2014.12.027] [Medline: 25865391]
70. Schunk DH. Self-efficacy and achievement behaviors. Educ Psychol Rev 1989 Sep;1(3):173-208. [doi: 10.1007/bf01320134]
71. Keefe FJ, Lefebvre JC, Maixner W, Salley AN, Caldwell DS. Self-efficacy for arthritis pain: relationship to perception of thermal laboratory pain stimuli. Arthr Care Res 1997 Jun;10(3):177-184. [doi: 10.1002/art.1790100305] [Medline: 9335629]
72. Somers TJ, Blumenthal JA, Guilak F, Kraus VB, Schmitt DO, Babyak MA, et al. Pain coping skills training and lifestyle behavioral weight management in patients with knee osteoarthritis: a randomized controlled study. Pain 2012 Jun;153(6):1199-1209 [FREE Full text] [doi: 10.1016/j.pain.2012.02.023] [Medline: 22503223]
73. Keefe FJ, Somers TJ, Martire LM. Psychologic interventions and lifestyle modifications for arthritis pain management. Rheum Dis Clin North Am 2008 May;34(2):351-368 [FREE Full text] [doi: 10.1016/j.rdc.2008.03.001] [Medline: 18638681]
74. Wang L, Zhang L, Yang L, Cheng-Qi H. Effectiveness of pain coping skills training on pain, physical function, and psychological outcomes in patients with osteoarthritis: a systemic review and meta-analysis. Clin Rehabil 2021 Mar 26;35(3):342-355. [doi: 10.1177/0269215520968251] [Medline: 33103915]
75. Gandy M, Pang ST, Scott AJ, Heriseanu AI, Bisby MA, Dudeney J, et al. Internet-delivered cognitive and behavioural based interventions for adults with chronic pain: a systematic review and meta-analysis of randomized controlled trials. Pain 2022 Oct 01;163(10):e1041-e1053. [doi: 10.1097/j.pain.0000000000002606] [Medline: 35121696]
76. Tousignant M, Moffet H, Nadeau S, Mérette C, Boissy P, Corriveau H, Belzile, et al. Cost analysis of in-home telerehabilitation for post-knee arthroplasty. J Med Internet Res 2015 Mar 31;17(3):e83 [FREE Full text] [doi: 10.2196/jmir.3844] [Medline: 25840501]
77. Mani S, Sharma S, Omar B, Paungmali A, Joseph L. Validity and reliability of internet-based physiotherapy assessment for musculoskeletal disorders: a systematic review. J Telemed Telecare 2016 Mar 31;23(3):379-391. [doi: 10.1177/1357633x16642369]
78. Richardson BR, Truter P, Blumke R, Russell TG. Physiotherapy assessment and diagnosis of musculoskeletal disorders of the knee via telerehabilitation. J Telemed Telecare 2016 Jul 08;23(1):88-95. [doi: 10.1177/1357633x15627237]
79. Marshall CJ, El-Ansary D, Pranata A, Ganderton C, O'Donnell J, Takla A, et al. Validity and reliability of a novel smartphone tele-assessment solution for quantifying hip range of motion. Sensors (Basel) 2022 Oct 25;22(21):8154 [FREE Full text] [doi: 10.3390/s22218154] [Medline: 36365852]
80. Amin J, Ahmad B, Amin S, Siddiqui A, Alam M. Rehabilitation professional and patient satisfaction with telerehabilitation of musculoskeletal disorders: a systematic review. Biomed Res Int 2022;2022:7366063 [FREE Full text] [doi: 10.1155/2022/7366063] [Medline: 35958819]
81. Ding X, Clifton D, Ji N, Lovell NH, Bonato P, Chen W, et al. Wearable sensing and telehealth technology with potential applications in the coronavirus pandemic. IEEE Rev Biomed Eng 2021;14:48-70. [doi: 10.1109/rbme.2020.2992838]
82. Tey C, An J, Chung W. A novel remote rehabilitation system with the fusion of noninvasive wearable device and motion sensing for pulmonary patients. Comput Math Methods Med 2017;2017:5823740-5823748 [FREE Full text] [doi: 10.1155/2017/5823740] [Medline: 28553367]
83. Bartolo M, Intiso D, Lentino C, Sandrini G, Paolucci S, Zampolini M, Board of the Italian Society of Neurological Rehabilitation (SIRN). Urgent measures for the containment of the coronavirus (Covid-19) epidemic in the neurorehabilitation/rehabilitation departments in the phase of maximum expansion of the epidemic. Front Neurol 2020 Apr 30;11:423 [FREE Full text] [doi: 10.3389/fneur.2020.00423] [Medline: 32425880]
84. Inskip JA, Lauscher HN, Li LC, Dumont GA, Garde A, Ho K, et al. Patient and health care professional perspectives on using telehealth to deliver pulmonary rehabilitation. Chron Respir Dis 2018 Feb 18;15(1):71-80 [FREE Full text] [doi: 10.1177/1479972317709643] [Medline: 28569116]
85. Agostini M, Moja L, Banzi R, Pistotti V, Tonin P, Venneri A, et al. Telerehabilitation and recovery of motor function: a systematic review and meta-analysis. J Telemed Telecare 2015 Feb 22;21(4):202-213. [doi: 10.1177/1357633x15572201]
86. Pastora-Bernal JM, Martín-Valero R, Barón-López FJ, Estebanez-Pérez MJ. Evidence of benefit of telerehabitation after orthopedic surgery: a systematic review. J Med Internet Res 2017 Apr 28;19(4):e142 [FREE Full text] [doi: 10.2196/jmir.6836] [Medline: 28455277]
87. Berton A, Longo UG, Candela V, Fioravanti S, Giannone L, Arcangeli V, et al. Virtual reality, augmented reality, gamification, and telerehabilitation: psychological impact on orthopedic patients' rehabilitation. J Clin Med 2020 Aug 07;9(8):2567 [FREE Full text] [doi: 10.3390/jcm9082567] [Medline: 32784745]
88. Then JW, Shivdas S, Tunku Ahmad Yahaya TS, Ab Razak NI, Choo PT. Gamification in rehabilitation of metacarpal fracture using cost-effective end-user device: a randomized controlled trial. J Hand Ther 2020 Apr;33(2):235-242. [doi: 10.1016/j.jht.2020.03.029] [Medline: 32430167]
89. Pastora-Bernal J, Martín-Valero R, Barón-López FJ, García-Gómez O. Effectiveness of telerehabilitation programme following surgery in shoulder impingement syndrome (SIS): study protocol for a randomized controlled non-inferiority trial. Trials 2017 Feb 23;18(1):82 [FREE Full text] [doi: 10.1186/s13063-017-1822-x] [Medline: 28231815]

XSL•FO
RenderX

90. Markert C, Sasangohar F, Mortazavi BJ, Fields S. The use of telehealth technology to support health coaching for older adults: literature review. JMIR Hum Factors 2021 Jan 29;8(1):e23796 [FREE Full text] [doi: 10.2196/23796] [Medline: 33512322]
91. Garber C, Blissmer B, Deschenes M, Franklin B, Lamonte M, Lee IM, American College of Sports Medicine. American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. Med Sci Sports Exerc 2011 Jul;43(7):1334-1359. [doi: 10.1249/MSS.0b013e318213fefb] [Medline: 21694556]
92. Slade SC, Dionne CE, Underwood M, Buchbinder R. Consensus on Exercise Reporting Template (CERT): explanation and elaboration statement. Br J Sports Med 2016 Dec 05;50(23):1428-1437. [doi: 10.1136/bjsports-2016-096651] [Medline: 27707738]
93. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348(mar07 3):g1687-g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]

Abbreviations

ADL: activities of daily living
AIMS2: Arthritis Impact Measurement Scale 2
ASES: Arthritis Self-Efficacy Scale
CBT: cognitive-behavioral therapy
GRADE: Grades of Recommendation Assessment, Development, and Evaluation
HOOS: Hip Disability and Osteoarthritis Outcome Score
IBTH: internet-based telehealth
ITT: intention to treat
KOOS: Knee Injury and Osteoarthritis Outcome Score
NRS: Numeric Rating Scale
OA: osteoarthritis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB 2: Risk of Bias version 2
SMD: standardized mean difference
THR: total hip replacement
TIDieR: Template for Intervention Description and Replication
TKR: total knee replacement
VAS: Visual Analogue Scale
VR: virtual reality
WOMAC: Western Ontario and McMaster Universities Osteoarthritis index

Edited by A Schwartz, T de Azevedo Cardoso; submitted 16.01.24; peer-reviewed by SH Xie, Y Tsutsumi; comments to author 06.04.24; revised version received 10.08.24; accepted 04.09.24; published 30.09.24.

Please cite as:
Wang HN, Luo P, Liu S, Liu Y, Zhang X, Li J
Effectiveness of Internet-Based Telehealth Programs in Patients With Hip or Knee Osteoarthritis: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e55576
URL: https://www.jmir.org/2024/1/e55576
doi:10.2196/55576
PMID:

©Hao-Nan Wang, Pei Luo, Shuyue Liu, Yunyi Liu, Xiao Zhang, Jian Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review

Hans Eguia[1,2*], MSc, MD; Carlos Luis Sánchez-Bocanegra[2*], PhD; Franco Vinciarelli[1,3*], ITE; Fernando Alvarez-Lopez[4*], PhD; Francesc Saigí-Rubió[2*], PhD

[1]SEMERGEN New Technologies Working Group, Madrid, Spain
[2]Faculty of Health Sciences, Universitat Oberta de Catalunya (UOC), Barcelona, Spain
[3]Emergency Hospital Clemente Álvarez, Rosario (Santa Fe), Argentina
[4]Faculty of Health Sciences, Universidad de Manizales, Manizales, Colombia
[*]all authors contributed equally

Corresponding Author:
Francesc Saigí-Rubió, PhD
Faculty of Health Sciences
Universitat Oberta de Catalunya (UOC)
Rambla del Poblenou, 156
Barcelona, 08018
Spain
Phone: 34 933263622
Email: fsaigi@uoc.edu

Abstract

Background: Ensuring access to accurate and verified information is essential for effective patient treatment and diagnosis. Although health workers rely on the internet for clinical data, there is a need for a more streamlined approach.

Objective: This systematic review aims to assess the current state of artificial intelligence (AI) and natural language processing (NLP) techniques in health care to identify their potential use in electronic health records and automated information searches.

Methods: A search was conducted in the PubMed, Embase, ScienceDirect, Scopus, and Web of Science online databases for articles published between January 2000 and April 2023. The only inclusion criteria were (1) original research articles and studies on the application of AI-based medical clinical decision support using NLP techniques and (2) publications in English. A Critical Appraisal Skills Programme tool was used to assess the quality of the studies.

Results: The search yielded 707 articles, from which 26 studies were included (24 original articles and 2 systematic reviews). Of the evaluated articles, 21 (81%) explained the use of NLP as a source of data collection, 18 (69%) used electronic health records as a data source, and a further 8 (31%) were based on clinical data. Only 5 (19%) of the articles showed the use of combined strategies for NLP to obtain clinical data. In total, 16 (62%) articles presented stand-alone data review algorithms. Other studies (n=9, 35%) showed that the clinical decision support system alternative was also a way of displaying the information obtained for immediate clinical use.

Conclusions: The use of NLP engines can effectively improve clinical decision systems' accuracy, while biphasic tools combining AI algorithms and human criteria may optimize clinical diagnosis and treatment flows.

Trial Registration: PROSPERO CRD42022373386; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=373386

(J Med Internet Res 2024;26:e55315) doi:10.2196/55315

KEYWORDS

artificial intelligence; AI; natural language processing; clinical decision support system; CDSS; health recommender system; clinical information extraction; electronic health record; systematic literature review; patient; treatment; diagnosis; health workers

XSL•FO
RenderX

Introduction

Advancement in medicine continues apace, especially with the emergence of new pathologies such as COVID-19. New treatments are continually being developed to fight not only these diseases but also previous pathologies for which new alternatives are being developed. Consequently, the number of publications in different indexed journals has increased, as shown in search results in various databases such as PubMed. Currently, it is possible to find many articles that mention new treatments or even new diagnostic forms [1,2].

Real and verified information is vital for the treatment and diagnosis of patients and is the cornerstone of medicine. The National Library of Medicine has developed at least 3 major source evaluation systems that provide useful examples for the task at hand: MEDLINE indexing, MedlinePlus indexing, and the Disaster Lit database [3].

Many health workers use the internet to search generally for updated clinical data [4]. However, this method is not the most efficient way to find information, since physicians must determine the type of information they need and then conduct the search themselves in an online medical database. This type of search can not only be time-consuming but also error prone due to not using suitable data. Therefore, automated information recommender systems have been established as a solution that allows medical staff to obtain reliable knowledge very quickly. These types of solutions are known as clinical decision support systems (CDSSs) [5].

CDSSs are composed of multiple platforms that allow the assessment of clinical data and alert clinicians to eventual problems. In addition, decision-making tools can be used to assist clinical staff. For these systems to function properly, they must interact with elements that allow them to obtain updated data for improved development, such as electronic health records (EHRs) [6]. Accordingly, CDSSs are known to focus on 6 specific aspects: data, knowledge, inference, architecture and technology, implementation and integration, and the user [7].

All available technology and tools (eg, artificial intelligence [AI], machine learning, and big data) could be useful for obtaining high-quality, reliable information. Such information could also be obtained by taking a supervised machine learning approach using several natural language processing (NLP) components that are domain independent and related to medical information extraction (text mining) [8]. These resources could include medical sources such as the Unified Medical Language System, different metathesauri, and different medical ontologies.

This study aims to answer the question of whether AI- and NLP-based CDSSs can provide effective results in automated searches that are useful to health care staff. To this end, a systematic review was carried out to assess the current state of these techniques in health care to identify their potential use in EHRs and automated information searches. The results found and conclusions drawn about the research question are subsequently presented.

Methods

Study Design

The protocol for this systematic review was published on November 5, 2022, in PROSPERO (CRD42022373386). This systematic review was performed per the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [9]. A search was conducted in the PubMed, Embase, ScienceDirect, Scopus, and Web of Science online databases for articles published between January 2000 and April 2023 using combinations of the following Medical Subject Headings (MeSH) terms: (((Artificial Intelligence [MeSH Terms]) AND (Natural language processing [MeSH Terms])) AND (Clinical decision support [MeSH Terms])) AND (Electronic health record [MeSH Terms]). The snowballing technique was used to complement the search to find the articles most relevant to the study [10].

Selection Criteria

In total, 2 researchers independently assessed titles and abstracts and analyzed appropriate studies through full-text evaluation. The only inclusion criteria were (1) original research articles and studies on the application of AI-based medical clinical decision support (CDS) using NLP techniques and (2) publications in English. The exclusion criteria were (1) studies describing the use of AI that are not focused on CDS tools; (2) studies related only to NLP; (3) studies related to an algorithm submitted to a challenge; (4) letters to the editor; (5) conference abstracts, books and book reviews; and (6) studies not published in scientific journals (ie, only in science magazines or magazines without a DOI).

Data Extraction and Management

Data were collected as follows: (1) reference, country, and year; (2) objective; (3) study type; (4) research design—intervention; (5) population sample + target (organ); and (6) results and conclusions. Further, 2 researchers independently extracted data. A third investigator resolved discrepancies.

Quality Appraisal of the Studies

The articles were independently assessed by 2 researchers. Disagreements were discussed until a consensus was reached. A Critical Appraisal Skills Programme (CASP) tool for qualitative studies with a 10-item scale (0-10) [11] was used to ensure the quality of the studies, focusing on (1) validity of the study, (2) accuracy of the results, and (3) transferability. A 10-item CASP scale (0-10) was used for systematic reviews, focusing on (1) validity of the study, (2) robustness and relevance of the findings, and (3) applicability and relevance of the results in a local or specific context. Quality appraisal was used to demonstrate the methodological quality of the studies since it would affect the validity of the results and was something that needed to be taken into account when considering the findings of the review.

Ethical Considerations

This study relied on secondary data. No ethics approval or patient consent was therefore required.

XSL•FO
RenderX

Results

Overview

A systematic review was conducted with the aim of assessing the current state of AI and NLP techniques in health care to identify their potential use in EHRs and automated information searches. In the initial search, 707 articles were retrieved. In title and abstract screening, 594 publications were excluded either due to their lack of relevance to the search or duplication. After the initial review, 113 articles were chosen for further examination: 62 from PubMed, 9 from ScienceDirect, 7 from Embase, 18 from Scopus, and 17 from Web of Science. Of the remaining studies, 87 were excluded as they showed examples of data mining and algorithms for a challenge, presented nonscientific stories, or gave ultrashort presentations, among others. Therefore, 26 articles were included in the final analysis. The overview flowchart is shown in Figure 1.

Figure 1. PRISMA 2020 flow diagram for new systematic reviews that include searches of databases and registers only. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

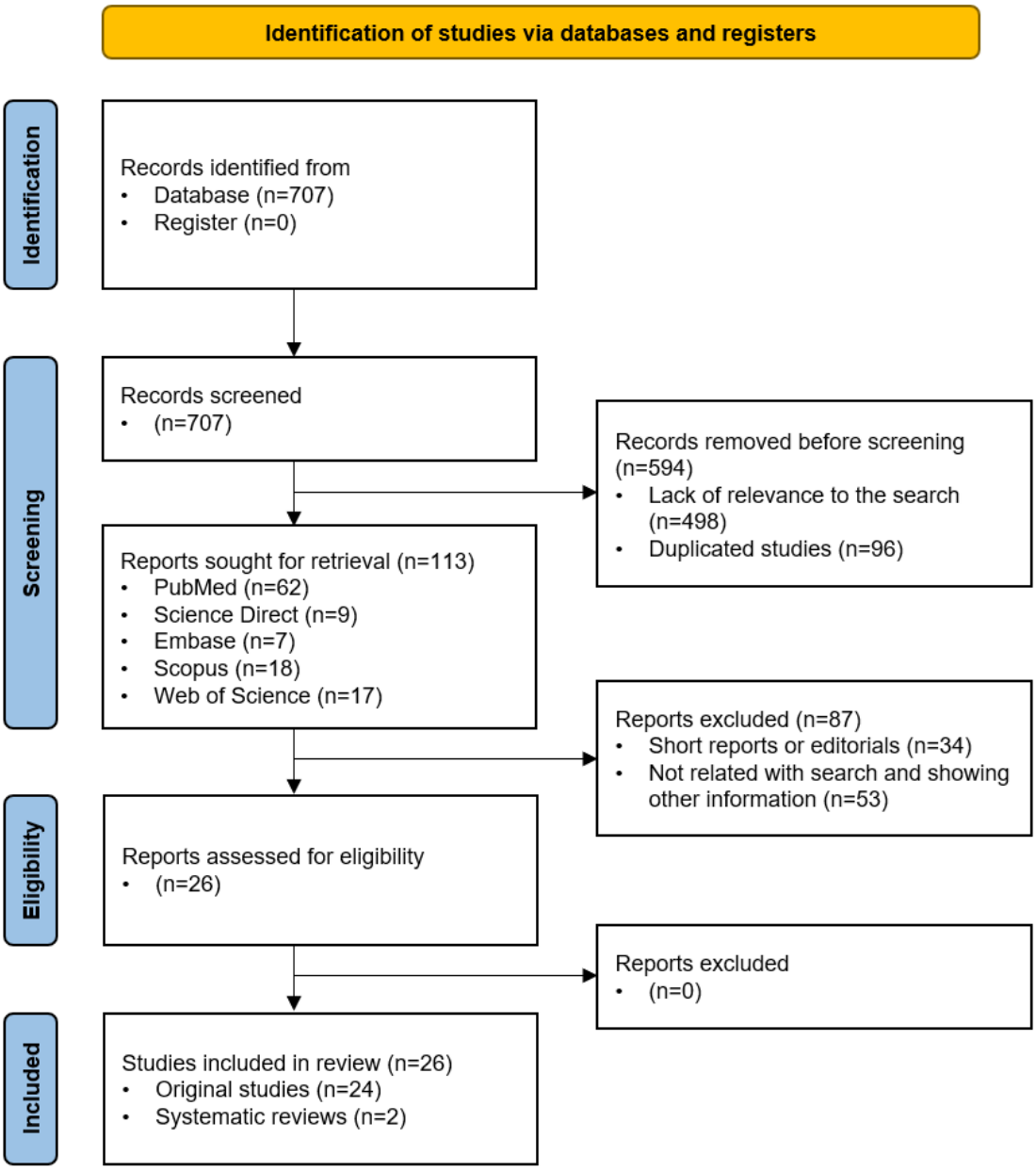

Characteristics of the Included Studies

The characteristics of the included studies are reported in Table 1 and Multimedia Appendix 1. In total, 24 of the evaluated articles were original articles and 2 were systematic reviews [12,13]. It should be noted that 6 of the original articles reviewed were presentations of studies conducted in relation to the 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records [8,14-18].

XSL•FO
RenderX

Table 1. Summary of the studies included in the systematic review (n=26).

Study	Objective	Research design—intervention	Population sample + target (organ)	Results and conclusions
Clark et al [14]	To develop a system for determining the assertion status of clinical reports (extracted from patient records)	• Determining what kind of clinical assertion was made in a clinical record (problem concept-sorted). • 3 assertions: conditional hypothetical and not associated with the patient	• 2010 i2b2[a]/VA[b]: 349 clinical records. • Detection of negation, speculation, and conditionals	• Linguistic information (negation status and temporal attributes) are features that are not always evaluated. They could give a more accurate classification
Patrick et al [15]	To present a method to deal with different extractions of, and classifications in, clinical data	• Developing clinical NLP[c] that includes a proofreading process with validation and correction	• 2010 i2b2/VA: 349 manually annotated notes + 827 raw records: 479 records	• NLP in clinical fields needs to address the issues between model complexity and model accuracy
Roberts and Harabagiu [8]	To develop a framework that can optimally identify medical concepts and adequately classify assertions	• Evaluating and finding the near-optimal subset of features for the 2010 i2b2/VA context task (concept extraction)	• 2010 i2b2/VA: 827 summaries	• Machine-learning–based supervised methods can be improved with feature-selection techniques
Jiang et al [16]	To design and evaluate a machine-learning algorithm to extract clinical information from hospital discharge summaries	• Evaluating the effects of different types of features	• 2010 i2b2/VA: 349 clinical annotated notes	• A system that is useful for concept extraction and assertion
D'Avolio et al [17]	To evaluate if the use of NLP-derived features combined with supervised machine-learning can perform effectively across tasks	• Creating an engine using cTAKES[d] ("noun-phrase bumpers") for text extraction with NLP combined with machine-learning classification	• 2010 i2b2/VA: 349 clinical annotated notes	• NLP clinical extraction is well known. Its adoption with acceptable concept-level extraction performance is also known
Garla et al [19]	To develop a classifier extractor using cTAKES to store documents	• Exploring if the accuracy of rule-based classifiers could be improved with machine-learning algorithms (testing systems after improvement)	• 13,000 abdominal radiology reports, ascites, and liver masses of 395 patients	• cTAKES extensions simplify feature extraction—it enables a better and more efficient exploration of text including the syntactic structure of documents and the negation context of concepts
Wagholikar et al [20]	To develop a CDSS for cervical cancer screening that can interpret free text Papanicolaou reports	• Developing a CDSS[e] that would be able to identify abnormal Papanicolaou reports as well as interpret other variables accurately	• 74 patient cases from cervical cancer screening	• EHR[f] text can be effectively used through NLP to develop CDS[g] tools.
Wagholikar et al [21]	To evaluate a CDSS for cervical cancer screening	• A previous CDSS was created, care providers' recommendations were requested and compared with the CDSS system	• 6053 random patients and 8 reviewers. • Cervical cancer screening	• CDS and human review can facilitate the evaluation of accuracy and would help to identify decision scenarios that may be missed

XSL•FO
RenderX

Study	Objective	Research design—intervention	Population sample + target (organ)	Results and conclusions
Sordo et al [22]	To develop a conceptual schema to represent clinical knowledge for decision support	• Rule interactions based on semantic web technologies (ontologies). Evaluating the feasibility of the modelling strategy by implementing test scenarios	• 3 applications were analyzed: laboratory results; adverse drug reaction monitoring and immunization protocols	• A richer conceptual representation of production rules will facilitate rule authoring consistency in rule implementation and maintenance. • Interactions are not considered a modelling time but they should be
Mehrabi et al [23]	To develop an NLP system to identify (retrospective) patients with pancreatic cysts	• Records were evaluated by the system created. This system double checks negation in the EHR and stratifies the results (low or high risk of cancer)	• Test set of 703 patient records: pancreas and cyst. Training set: 316 control records	• Identification of negation improved the algorithm, increasing recall to 98.12%
Ross et al [12]	To use secondary data and to put data into practice	• Paper-based materials containing similar information. Finding examples of BD[h] use with EHRs	• 84 papers—EHR research (various organs)	• Greater use of EHRs means more clinical data. There is a need to develop new methods to obtain better data from EHRs
Kotfila and Uzuner [18]	To evaluate if NLP techniques can identify phenotypes in unstructured medical notes	• Training support vector machines using individual feature spaces to assess better model performance	• 2 separate i2b2 shared tasks (2008+2014: 790+730 documents)	• A simple feature performed as well as a combination of feature spaces—the addition of training data has weak statistical significance effects
Patterson et al [24]	To determine whether a colonoscopy was performed for screening	• Creating a rule-based model and 3 machine-learning models (automated retrieval console) which also uses cTAKES	• A set of 1000 documents with colonoscopies	• A rule-based classification system appears to be more robust than a machine-learning system alone
Divita et al [25]	To use an NLP algorithm to process large corpora of clinical notes to demonstrate a time decrease in the analyses of a large corpus of clinical information	• Developing an NLP algorithm to detect urinary catheters in hospitalized patients	• 550,000 notes from patients with urinary catheters from a US Veteran Affairs hospital	• Generalizable and widely applicable techniques can aid BD processing and analysis
Mei et al [26]	To evaluate a decision fusion framework for treatment recommendation systems (combining knowledge-driven and data-driven decisions)	• Implementing a fusion engine that obtains inputs from base decision engines	• 3150 records from patients with diabetes in China	• Decision fusion is a promising way toward a more valuable treatment recommendation
Marco-Ruiz et al [27]	To find out if multidisciplinary leverage archetypes and ontologies can model CDSS (better reuse and maintenance)	• An archetype was developed and reviewed by 5 practitioners and 5 information architects. • SNOMED for concept simplification	• 126 cases with respirator and symptoms and signs	• Model archetypes can provide information using ontologies. • EHR standards and terminologies can be used by multidisciplinary teams effectively
Danger et al [28]	To explain the methodology for constructing a clinical prediction rule repository	• Using data mining algorithms for pattern extraction	• Clinical collection for 126,931 patients from primary care with multiple diagnoses	• Important factors to be included in the rules were defined. With the use of clinical evidence rules, a better diagnostic recommendation can be obtained

XSL•FO
RenderX

Study	Objective	Research design—intervention	Population sample + target (organ)	Results and conclusions
Breischneider et al [29]	To introduce a system for the automated processing of clinical reports of patients with mamma carcinoma to extract relevant textual features and derive therapy suggestions	• Use of a rule-based grammar approach for information extraction and deriving therapy suggestions based on the extracted variables and clinical guidelines. The accuracy of predicting therapeutic measures is also reported	• 8766 clinical text reporting on 2096 patients with mamma carcinoma	• Good accuracy of information extraction for different features. The evaluation of therapy suggestions, factors influencing the error between suggested and conducted therapy
Yang et al [30]	To use convolutional neural networks to be able to potentiate extraction without paper construction of rules or knowledge bases	• Proposing a method using neural networks to extract features from EHRs. A tool was used in medical text showing 98.7% accuracy and 96.02% recall	• 18,590 EHRs indicating hypertension, diabetes, COPD[i], gout, arrhythmia, asthma, gastritis, and stomach polyps. After validation: a training set of 7000 EHRs and a test set of 400 EHRs	• The use of neural networks can increase information extraction. This could assist in a feasible and effective diagnosis
Wissel et al [31]	To validate an NLP application using machine-learning to identify patients for epilepsy surgery	• Determining the proportion of potential surgical candidates using EHRs, NLP, and a surgical score	• 1 app and 4211 patients with an epilepsy diagnosis after surgical treatment of epilepsy and controls	• NLP-based app modelled for patient score to identify candidates for epilepsy surgery
Wulff et al [32]	To design a tool to automatically extract important information from medical texts and transform them into standardized data	• NLP consisting of 5 successive tasks was developed. The tool was implemented to extract fragments from EHRs (97% precision and 94% recall)	• 529 medical events from the pediatric cardiology and the intensive care medicine department at Hannover Medical School. Further, 499 events were correctly extracted	• NLP-based solution to extract information from medical text. This will benefit health care by analyzing a large amount of information and presenting the most important data
Kulchak et al [33]	To evaluate if modification of a CDS tool using concepts of human-centered design can improve CDS itself	• Adapting a DDSS[j] (design, NLP with cTAKES, and coordination with EHRs). Creating simulated cases using real but deidentified clinical information (varying complexity)	• 3 simulated genetic cases of increasing complexity were created, 3 clinicians tested the system and another 3 retested it	• Evidence of potential implementation of DDSS using machine-assisted medical record review. The use of mixed technical CDS and human-centered design criteria enhances the real-world clinical workflow
van de Burgt et al [13]	To assess whether data mining can improve the diagnostic and therapeutic processes of CDS	• Combining data mining and CDSSs may improve diagnostic and therapeutic processes, contributing to increased patient safety	• 714 reviewed publications, 39 of which were included	• Barriers: complexity of NLP, EHR incompleteness, validation and performance of the tool, lack of input from an expert team, and the adoption rate among professionals
Suh et al [34]	To evaluate the use of clinical NLP to identify elements relevant to preoperative medical history by analyzing clinical notes	• An NLP pipeline would identify a significant portion of pertinent history captured by a perioperative provider (in an EHR)	• 93 patients with 9765 clinical notes were reviewed	• The use of NLP identified medical conditions relevant to preanesthetic evaluation from unstructured free-text input. This provides CDS or recommends additional testing or evaluation
Park et al [35]				

XSL•FO
RenderX

Study	Objective	Research design—intervention	Population sample + target (organ)	Results and conclusions
	To evaluate the efficacy of an algorithm of 3 levels of search functionality in supporting information retrieval for clinical users from EHR in a simulated clinical environment	• Medical tasks were presented to an evaluation system and a panel of medical raters to assess both the speed and accuracy of task completion	• Medical records for 3 fictional patients, each with 20 documents, were assessed by 60 clinicians of varying specialties and experience levels	• NLP-enhanced search facilitated more accurate task completion than both string search and no search, with similar task speeds between NLP-enhanced search and string search
Afshar et al [36]	To implement a real-time NLP-driven CDS tool for screening opioid misuse in hospitalized adults and assess its effectiveness in providing interventions for substance use disorder treatment	• A 30-month quasi-experimental pre-post study screened inpatients for 24 months of standard care followed by a 6-month tool implementation phase. Cost-effectiveness analysis and user acceptability evaluations were covered	• A total sample size of 12,500 patients (10,000 preintervention and 2500 postintervention)	• The study implemented a real-time NLP-driven CDS tool for screening opioid misuse with a sensitivity of 93% and specificity of 92%. The workflow was reproducible and included a shared pseudocode for cloud service implementation

[a]i2b2: Informatics for Integrating Biology and the Bedside.

[b]VA: US Department of Veterans Affairs.

[c]NLP: natural language processing.

[d]cTAKES: Clinical Text Analysis and Knowledge Extraction System.

[e]CDSS: clinical decision support system.

[f]EHR: electronic health record.

[g]CDS: clinical decision support.

[h]BD: big data.

[i]COPD: chronic obstructive pulmonary disease.

[j]DDSS: diagnostic decision support system.

Of the 26 articles reviewed, 18 (69%) corresponded to authors from the United States [8,12-14,16-25,31,33,34,36]; China [26,30], Germany [29,32], the United Kingdom [28,35] each had 2 (8) articles; and the rest (n=2, 8%) were from Norway [27] and Australia [15].

Of the evaluated articles, 21 (81%) explained the use of NLP as a source for data collection, and 18 (69%) articles used EHRs as a data source; meanwhile, a further 8 (31%) articles were based on clinical data [8,12,14-16,18-21,23-25,29-33,36]. Only 5 (19%) articles showed the use of an NLP tool called Apache cTAKES (Clinical Text Analysis and Knowledge Extraction System) as a set of combined strategies for NLP to obtain clinical data [14,16,17,19,36].

A total of 16 (62%) articles presented stand-alone data review algorithms [20-33,35,36]. Other studies (n=9, 35%) showed that the CDSS alternative was also a way of displaying the information obtained for immediate clinical use [12,20,22,26-28,30,33,36].

Some of the articles focused on specific pathologies such as epilepsy [31]; genomics [33]; pancreatic cysts [23]; radiographic images [19]; diabetes, obesity, hypertension, dyslipidemia, and cardiovascular diseases [18]; colonoscopies [24]; urinary problems or catheters [25]; posttraumatic stress disorder [17]; preanesthetic evaluation [34]; breast cancer [29]; opioid misuse [36]; and cervical cancer assessment [20,21], thus emphasizing the use that can be made of these types of system in almost all medical specialties. In addition, it should be noted that the vast majority of studies reviewed did not consist of just a few cases. This is demonstrated by the fact that 74 patients and their EHRs were reviewed in 1 study [20]; 349 clinical cases were reviewed in several studies [8,14-18]; and more than 126,000 clinical cases were analyzed by information collection systems in another [28].

Quality Appraisal Results

In the CASP checklist for the 24 qualitative studies (Multimedia Appendix 2), all had a clear statement of the aims of the research, an adequate qualitative study design, and clearly defined outcomes. Data collection and analysis were sufficiently rigorous in the 24 studies, all were adequately designed to achieve the research aims, and the results obtained were readily transferrable to other settings. However, only 4 of the evaluated studies [20,21,27,31] indicated interaction with the participants (patients, in this instance), which involved informing them about the study and the use of the data obtained from it. A further 12 studies [18,19,23-25,28-30,32-34,36] also used patient data or health records, but there was no mention of patients being informed. Importantly, while the remaining 7 studies used data for their research, 5 [8,15-17,37] obtained the necessary information from a database for a challenge (i2b2/VA), 1 study generated clinical data specifically for the research [35], and the other 2 [22,26] did not mention the type of data they used or where they obtained it from. Nevertheless, none of this influenced or affected the results of the research. All the studies were analyzed using standard means of content analysis and provided sufficient information on the design to replicate the

XSL•FO
RenderX

study. This was sufficient to demonstrate the credibility of the studies and that the data analysis was sufficiently rigorous.

In the CASP checklist for 2 systematic reviews (Multimedia Appendix 2), both studies had a clear statement of research objectives, an appropriate study design, clearly defined outcomes, and sufficiently rigorous data collection and analysis. Both studies were adequately designed to achieve the research aims, and the results obtained were readily transferable to other settings.

Discussion

Principal Findings

This systematic review produced a synthesis of the current state of AI and NLP techniques in health care to identify their potential use in EHRs and automated information searches. Most of the studies showed good internal validity and decent quality. What stands out from our study is the use of NLP as a source for data collection, and while most of the included studies used EHRs, some were based on clinical data. Only 5 of the articles indicated the use of combined NLP strategies to obtain clinical data. While more than half (16/26, 62%) of the articles presented stand-alone algorithms for data review, others (9/26, 35%) indicated that CDSSs also served to present the information obtained for immediate clinical use.

NLP, as a data mining technique, is considered one of the most appropriate tools to find useful information in the data contained in large databases [12,16]. This is because it is an instrument that enables large amounts of information to be clinically analyzed, showing only the parts with the greatest interest or importance to health professionals [32]. While its use has significantly advanced in extracting concepts from clinical data [17], it faces challenges when dealing with the unstructured format of EHRs, which can impede accurate responses to queries submitted to NLP [15].

To overcome these challenges, various techniques have been proposed. One approach involves the combined use of clinical scores that serve as a guide for obtaining results [31], which could be very useful in improving health systems. Another technique to enhance data collection could be the use of neural networks to increase information extraction (and thus achieve more effective diagnoses) [30]. An alternative option offered by NLP includes a sentiment-based model that goes beyond the traditional collaborative filtering approach. This model uses machine learning algorithms to analyze human language text. The metrics used in sentiment analysis aim to determine whether the overall tone of a text is positive, negative, or neutral [38].

Algorithms are the basis of NLP, which consist of any well-defined computational procedure that takes a value or set of values as input and produces a value or set of values as output [39]. However, despite algorithms being versatile tools used in programming and software development, and predominantly acknowledged for their pivotal role in data mining and AI [40], the process of algorithm development is not always straightforward and can sometimes become complicated. For example, many algorithms have difficulties with “negation,” as it can be interpreted as a positive part of a patient’s clinical history (thus “does not smoke” can be understood as “smoker”). This is a linguistic problem with features that are not always valued, which can lead to inaccurate classification [14,23,41]. It is for this reason that solutions such as NegEx, an algorithm developed in 2001, have been created to try to correct the problem with negation [42]. It should be noted that the use of rules (heuristics) in the search for clinical evidence can generate a better diagnostic recommendation [28], and this is probably because classification systems using rules present more robust machine learning models [24]. The use of tools such as cTAKES is also an alternative, as they are more efficiently and accurately able to scan texts and even the syntactic structure of documents, including negation [19].

The results show the potential of using NLP not only in reviewing clinical notes but also with algorithms that can help find specific information in large volumes of medical information [25]. This may explain its widespread use in epidemiology, public health, and disease surveillance [43,44]. The data obtained could be used to prevent new outbreaks of different diseases worldwide and to identify the main characteristics of pathologies to guide diagnoses even before the disease develops to chronic levels. Health care professionals could benefit from integrating NLP with AI in CDSS to improve medical consultations, streamline tasks such as data analysis, document clinical information in an automated and structured way, and refine treatment strategies and diagnostic processes by automated identification and extraction of key data from medical records [29]. Providing accurate information in real time could improve medical decision-making that better suits each patient’s individual needs, which could translate into better medical outcomes.

When grouping the results by their findings, several conclusions could be drawn, such as that NLP is effective in a CDSS, very accurate, and faster than manual search, especially when accompanied by a human review to facilitate the evaluation of the results and check their accuracy. However, it is necessary to consider the fact that more clinical data from EHRs may complicate its use and that new methods would have to be developed to better obtain large amounts of data. Another striking aspect is that all the reviewed articles focus on the detection of clinical data in EHRs in closed environments. That is, the information obtained was used to account for specific pathologies or diagnostic procedures, and the accuracy was assessed by someone able to understand EHRs. However, none of the articles reviewed referred to the use of external data (medical databases); they all use the data found—using NLP—in the EHRs only. By using external data sources, more appropriate or updated diagnostic aids and treatments could be obtained.

There are also some barriers preventing the development and improvement of NLP systems. One such barrier is the lack of data or incomplete data in EHRs. Another is associated with the lack of use or knowledge of NLP by health professionals. The latter significant issue is the lack of multidisciplinary working practices (health care and computer specialists), which hinders adequate progress concerning NLP algorithms. Establishing a multidisciplinary team involving physicians and information systems professionals would be the most effective approach, as demonstrated in various health care environments

XSL•FO
RenderX

[45-47]. In total, 6 of the reviewed articles described the results of a challenge to find an algorithm that best uses NLP in clinical notes, underscoring the efficacy of such initiatives in catalyzing technological advancement, thereby enhancing the performance of algorithms applicable in AI and big data domains.

Incorporating several medical ontologies to increase the coverage of medical entities may enhance results [48]. A semantic term, representing a single clinical concept, serves as a starting point for ontologies. The combination of these concepts defines a set of properties, allowing interconnections (mapping) between them. This process generates semantic ontologies, characterized by controlled terminology and formal semantic relationships in a particular area of interest using a particular modelling language and terminology [22], such as the terms in EHRs [27]. The incorporation of machine learning techniques into EHRs not only produces better results but also plays a key role in the development of predictive rules. Through the use of ontologies, diagnoses are standardized with a unified vocabulary, facilitating seamless exchange and validation across diverse populations [18,28].

The weight given to ontologies in the studies reviewed varies. While some of them define their use very well [22,27], others only mention ontologies as an important part of information extraction [8,12,15,17-19,23,28,29,33,34] or not at all [13,14,16,21,24-26,30-32,35,36]. Further, 1 study mentions the word "ontology" in the keywords but not in the text. This is surprising, as health ontologies are a fundamental part of clinical data extraction projects, and even more so considering the emergence of new ontologies with almost every new study. The levels of understanding of ontology concepts where the knowledge domains of medicine and computer science intersect could be reviewed as a future line of research.

Regarding the use of ontologies, their inclusion with the use of the semantic web, along with medical NLP, will lead to a better assessment of annotation tasks [49,50]. The use of ontologies is extremely important to overcome the barriers that may arise with the use of NLP. For example, to overcome them, some proposals could be adopted, such as the use of (1) AI assistants (special fusion engines) combining knowledge-based engines and data-based engines; (2) biphasic tools (adding human intervention) with the addition of a human reviewer, which would improve search results and identify potentially lost data; and (3) semantic graphs (sentiment analysis), where ontology-based AI tools would allow relevant information about pathologies in clinical data to be found.

The use of appropriate ontologies in NLP systems would serve to facilitate the real-time extraction of information that could be used for the development of real-time clinical decision tools [51,52]. Ontologies can also be useful for avoiding the ambiguity and inconsistencies found in some health care documents such as EHRs [53]. This is very important because these clinical documents could be converted into more understandable semantic structures by the NLP algorithm, allowing the most important information to be extracted [54]. Thus, a CDSS with incorporated NLP could provide physicians with contextual information, meaning that better clinical decisions could be made to the benefit of patients [55]. Such improvements could take the form of system-generated alerts when alterations in vital sign monitoring or interactions between prescribed drugs are detected [56-58].

Although CDSSs have great potential for use by health care staff to increase adherence to clinical guidelines and to assist in the correct diagnosis, treatment, follow-up, and prevention of various pathologies, with the consequent better maintenance of the population's health [7], some studies suggest that they may disrupt physicians' workflow or alter or be inconsistent with the initial clinical decisions, and may also require technical maintenance with additional costs [59]. Thus, depending on the algorithm and validation, it may present incorrect or low-quality data. Furthermore, as a different program from the EHR one, there may not be adequate interoperability between the two.

With this information, the advantages of using NLP and CDSSs are obvious. However, it is noteworthy that all the studies in our review have a "closed behavior," meaning that only specific information is searched for in the data present in clinical notes, without searching for further information in the large medical databases available. If the latter were to be carried out, it would allow a new line of research to be developed, in which NLP-based algorithms combined with keyword searches in clinical databases such as PubMed could potentially enable better and faster diagnoses to be made, and also updated treatments to be offered, all based on EHR data and in real time.

The ongoing evolution of generative AI, namely large language models (LLMs), represents a type of AI that is capable of generating text through a process of training on large data sets in multiple languages. These models demonstrate the ability to produce "human-like" responses [60]. A well-known example is ChatGPT, whose architecture uses a neural network to process natural language, thus generating responses based on the context of the input text [61]. It is essential to recognize that the synergistic use of these techniques presents a significant opportunity. The integration of tools based on LLM, medical ontologies, and NLP has the potential to offer a substantial positive influence on the health care process [62].

These results support the need to conduct research aimed not only at improving algorithms and generating new knowledge but also at suggesting new research directions for the development of AI tools. This includes the integration of NLP, medical ontologies, and LLM for enhanced search capabilities in EHRs and other external sources. A promising research path could be to develop algorithms whose architecture is based on web systems and contrasted medical databases, supported by AI with NLP, and that gather information about semantic terms from health care ontologies such as those in the National Library of Medicine. Such developments of AI-based tools may have a positive impact on research into their use in certain areas, such as health care [63]. In addition, the development of AI-based skills also enhances the development of further algorithms and research, as evidenced by the publications resulting from the challenges mentioned above. However, it is imperative to acknowledge the potential ethical implications inherent in this field, which require thorough assessment and subsequent integration into clinical practice [64]. While the potential benefits are substantial, it is paramount to rigorously address

XSL•FO
RenderX

ethical considerations and data privacy concerns, emphasizing cybersecurity and privacy requirements to effectively protect patients' sensitive data and ensure their confidentiality [36,65,66].

Limitations

Despite conducting an exhaustive search across 5 databases, which specifically targeted studies on the application of AI-based medical CDS using NLP techniques, a total of 113 studies were initially identified for screening. However, upon thorough review, only 26 studies were deemed to meet the stringent inclusion and exclusion criteria established for this review. Consequently, the representativeness of our findings may be questioned given the number of records primarily identified and the possible paucity of research on this particular study topic. A significant number of articles were excluded from our review due to their failure to establish a clear connection between NLP, AI, medical records, and their integration with CDSS. Despite delving into NLP and AI within the context of medical records, these articles lacked sufficient exploration of their relationship with CDSS [67]. The sources of information were peer-reviewed publications, so relevant information from other sources (eg, gray literature) was omitted. CASP-based quality scores [11] may have reflected incomplete reporting, since the vast majority of studies did not compare their results to those of other studies along similar lines (eg, the 2010 i2b2/VA challenge), or had short lists of references (between 8 and 20) in which nonscientific ones were included [14,22,23,25-28]. Nevertheless, all the articles were very robust in terms of the presentation of their results, which could be extrapolated to different local communities without losing their essence.

Conclusions

The use of NLP engines can effectively obtain results that guide the development of more accurate clinical decision systems. The implementation of decision systems using AI assistants is a potential use of this type of tool. Furthermore, the use of biphasic tools using AI criteria as algorithms combined with human criteria may improve the flow of clinical diagnosis or treatment. Human review can improve the accuracy of the search results as well as identify scenarios that might have been missed. The implementation of a special fusion engine (combining knowledge-driven and data-driven engines) is a promising technique that has shown results in terms of more relevant (or improved) recommendations.

Most CDSSs are designed to recommend text based on keywords. However, this leads to problems regarding the effectiveness of the method using NLP. Some proposals, such as the use of semantic graphs, have been put forward to solve this problem. Some controversy has arisen over the fact that CDSSs endure problems related to a certain coldness in their responses, as well as a paucity of data. A sentiment analysis technique to evaluate user preferences may help to overcome this.

The results found allow us to establish new lines of research for the development of AI tools based on NLP with the use of medical ontologies for information searching in both EHRs and external sources (clinical databases) to obtain better results and extra information that could be used to the benefit of patients.

Acknowledgments

This research received no external funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Main characteristics of the studies included in the systematic review.
[DOCX File , 134 KB - jmir_v26i1e55315_app1.docx]

Multimedia Appendix 2
Quality appraisal—CASP checklist for qualitative studies and for systematic review. CASP: Critical Appraisal Skills Programme.
[DOCX File , 143 KB - jmir_v26i1e55315_app2.docx]

Multimedia Appendix 3
PRISMA 2020 checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[PDF File (Adobe PDF File), 85 KB - jmir_v26i1e55315_app3.pdf]

References

1. Bradley-Ridout G, Nekolaichuk E, Jamieson T, Jones C, Morson N, Chuang R, et al. UpToDate versus DynaMed: a cross-sectional study comparing the speed and accuracy of two point-of-care information tools. J Med Libr Assoc 2021;109(3):382-387 [FREE Full text] [doi: 10.5195/jmla.2021.1176] [Medline: 34629966]
2. Abdill R, Blekhman R. Tracking the popularity and outcomes of all bioRxiv preprints. Elife 2019;8:e45133 [FREE Full text] [doi: 10.7554/eLife.45133] [Medline: 31017570]

XSL•FO
RenderX

3. Kington RS, Arnesen S, Chou WS, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021:10.31478/202107a [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
4. Bocanegra CLS, Ramos JLS, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. BMC Med Inform Decis Mak 2017;17(1):63 [FREE Full text] [doi: 10.1186/s12911-017-0431-7] [Medline: 28506225]
5. Muhiyaddin R, Abd-Alrazaq AA, Househ M, Alam T, Shah Z. The impact of clinical decision support systems (CDSS) on physicians: a scoping review. Stud Health Technol Inform 2020;272:470-473. [doi: 10.3233/SHTI200597] [Medline: 32604704]
6. Marcos M, Maldonado JA, Martínez-Salvador B, Boscá D, Robles M. Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility. J Biomed Inform 2013;46(4):676-689 [FREE Full text] [doi: 10.1016/j.jbi.2013.05.004] [Medline: 23707417]
7. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2016;25(Suppl 1):S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]
8. Roberts K, Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. J Am Med Inform Assoc 2011;18(5):568-573 [FREE Full text] [doi: 10.1136/amiajnl-2011-000152] [Medline: 21724741]
9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021;18(3):e1003583 [FREE Full text] [doi: 10.1371/journal.pmed.1003583] [Medline: 33780438]
10. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005;331(7524):1064-1065 [FREE Full text] [doi: 10.1136/bmj.38636.593461.68] [Medline: 16230312]
11. CASP (qualitative studies) checklist. Critical Appraisal Skills Programme. 2018. URL: https://casp-uk.net/images/checklist/documents/CASP-Qualitative-Studies-Checklist/CASP-Qualitative-Checklist-2018_fillable_form.pdf [accessed 2020-12-22]
12. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform 2014;23(1):97-104 [FREE Full text] [doi: 10.15265/IY-2014-0003] [Medline: 25123728]
13. van de Burgt BWM, Wasylewicz A, Dullemond B, Grouls R, Egberts T, Bouwman A, et al. Combining text mining with clinical decision support in clinical practice: a scoping review. J Am Med Inform Assoc 2023;30(3):588-603 [FREE Full text] [doi: 10.1093/jamia/ocac240] [Medline: 36512578]
14. Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, Yeh A, et al. MITRE system for clinical assertion status classification. J Am Med Inform Assoc 2011;18(5):563-567 [FREE Full text] [doi: 10.1136/amiajnl-2011-000164] [Medline: 21515542]
15. Patrick JD, Nguyen DHM, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Inform Assoc 2011;18(5):574-579 [FREE Full text] [doi: 10.1136/amiajnl-2011-000302] [Medline: 21737844]
16. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-606 [FREE Full text] [doi: 10.1136/amiajnl-2011-000163] [Medline: 21508414]
17. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. J Am Med Inform Assoc 2011;18(5):607-613 [FREE Full text] [doi: 10.1136/amiajnl-2011-000183] [Medline: 21697292]
18. Kotfila C, Uzuner Ö. A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. J Biomed Inform 2015;58 Suppl(Suppl):S92-S102 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.016] [Medline: 26241355]
19. Garla V, Lo Re V, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614-620 [FREE Full text] [doi: 10.1136/amiajnl-2011-000093] [Medline: 21622934]
20. Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012;19(5):833-839 [FREE Full text] [doi: 10.1136/amiajnl-2012-000820] [Medline: 22542812]
21. Wagholikar KB, MacLaughlin KL, Kastner TM, Casey PM, Henry M, Greenes RA, et al. Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. J Am Med Inform Assoc 2013;20(4):749-757 [FREE Full text] [doi: 10.1136/amiajnl-2013-001613] [Medline: 23564631]
22. Sordo M, Rocha BH, Morales AA, Maviglia SM, Oglio EDO, Fairbanks A, et al. Modeling decision support rule interactions in a clinical setting. Stud Health Technol Inform 2013;192:908-912. [Medline: 23920690]
23. Mehrabi S, Schmidt CM, Waters J, Beesley C, Krishnan A, Kesterson J, et al. An efficient pancreatic cyst identification methodology using natural language processing. Stud Health Technol Inform 2013;192:822-826. [Medline: 23920672]

XSL•FO
RenderX

24. Patterson OV, Forbush TB, Saini SD, Moser SE, DuVall SL. Classifying the indication for colonoscopy procedures: a comparison of NLP approaches in a diverse national healthcare system. Stud Health Technol Inform 2015;216:614-618. [Medline: 26262124]
25. Divita G, Carter M, Redd A, Zeng Q, Gupta K, Trautner B, et al. Scaling-up NLP pipelines to process large corpora of clinical notes. Methods Inf Med 2015;54(6):548-552. [doi: 10.3414/ME14-02-0018] [Medline: 26534722]
26. Mei J, Liu H, Li X, Xie G, Yu Y. A decision fusion framework for treatment recommendation systems. Stud Health Technol Inform 2015;216:300-304. [Medline: 26262059]
27. Marco-Ruiz L, Maldonado JA, Karlsen R, Bellika JG. Multidisciplinary modelling of symptoms and signs with archetypes and SNOMED-CT for clinical decision support. Stud Health Technol Inform 2015;210:125-129. [Medline: 25991115]
28. Danger R, Corrigan D, Soler J, Kazienko P, Kajdanowicz T, Majeed A, et al. A methodology for mining clinical data: experiences from TRANSFoRm project. Stud Health Technol Inform 2015;210:85-89. [Medline: 25991107]
29. Breischneider C, Zillner S, Hammon M, Gass P, Sonntag D. Automatic extraction of breast cancer information from clinical reports. 2017 Presented at: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS); June 22-24, 2017; Thessaloniki, Greece p. 213-218. [doi: 10.1109/CBMS.2017.138]
30. Yang Z, Huang Y, Jiang Y, Sun Y, Zhang Y, Luo P. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. Sci Rep 2018;8(1):6329 [FREE Full text] [doi: 10.1038/s41598-018-24389-w] [Medline: 29679019]
31. Wissel BD, Greiner HM, Glauser TA, Holland-Bouley KD, Mangano FT, Santel D, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. Epilepsia 2020;61(1):39-48 [FREE Full text] [doi: 10.1111/epi.16398] [Medline: 31784992]
32. Wulff A, Mast M, Hassler M, Montag S, Marschollek M, Jack T. Designing an openEHR-Based pipeline for extracting and standardizing unstructured clinical data using natural language processing. Methods Inf Med 2020;59(S 02):e64-e78 [FREE Full text] [doi: 10.1055/s-0040-1716403] [Medline: 33058101]
33. Kulchak Rahm A, Walton NA, Feldman LK, Jenkins C, Jenkins T, Person TN, et al. User testing of a diagnostic decision support system with machine-assisted chart review to facilitate clinical genomic diagnosis. BMJ Health Care Inform 2021;28(1):e100331 [FREE Full text] [doi: 10.1136/bmjhci-2021-100331] [Medline: 33962988]
34. Suh HS, Tully JL, Meineke MN, Waterman RS, Gabriel RA. Identification of preanesthetic history elements by a natural language processing engine. Anesth Analg 2022;135(6):1162-1171 [FREE Full text] [doi: 10.1213/ANE.0000000000006152] [Medline: 35841317]
35. Park EH, Watson HI, Mehendale FV, O'Neil AQ, Clinical Evaluators. Evaluating the impact on clinical task efficiency of a natural language processing algorithm for searching medical documents: prospective crossover study. JMIR Med Inform 2022;10(10):e39616 [FREE Full text] [doi: 10.2196/39616] [Medline: 36287591]
36. Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M, et al. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform 2023;11:e44977 [FREE Full text] [doi: 10.2196/44977] [Medline: 37079367]
37. Clark R, Moloney G. Facebook and older adults: fulfilling psychological needs? J Aging Stud 2020;55:100897. [doi: 10.1016/j.jaging.2020.100897] [Medline: 33272457]
38. Osman NA, Mohd Noah SA, Darwich M, Mohd M. Integrating contextual sentiment analysis in collaborative recommender systems. PLoS One 2021;16(3):e0248695 [FREE Full text] [doi: 10.1371/journal.pone.0248695] [Medline: 33750957]
39. Yanofsky NS. Towards a definition of an algorithm. J Log Comput 2010;21(2):253-286. [doi: 10.1093/logcom/exq016]
40. Ding S, Zhao H, Zhang Y, Xu X, Nie R. Extreme learning machine: algorithm, theory and applications. Artif Intell Rev 2013;44:103-115. [doi: 10.1007/s10462-013-9405-z]
41. Guo W, Kraines SB. Semantic content-based recommendations using semantic graphs. Adv Exp Med Biol 2010;680:653-659. [doi: 10.1007/978-1-4419-5913-3_72] [Medline: 20865551]
42. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
43. Kostkova P, Saigí-Rubió F, Eguia H, Borbolla D, Verschuuren M, Hamilton C, et al. Data and digital solutions to support surveillance strategies in the context of the COVID-19 pandemic. Front Digit Health 2021;3:707902 [FREE Full text] [doi: 10.3389/fdgth.2021.707902] [Medline: 34713179]
44. Schoene AM, Basinas I, van Tongeren M, Ananiadou S. A narrative literature review of natural language processing applied to the occupational exposome. Int J Environ Res Public Health 2022;19(14):8544 [FREE Full text] [doi: 10.3390/ijerph19148544] [Medline: 35886395]
45. Martín-Noguerol T, Paulano-Godino F, López-Ortega R, Górriz JM, Riascos R, Luna A. Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. Clin Radiol 2021;76(5):317-324. [doi: 10.1016/j.crad.2020.11.113] [Medline: 33358195]

XSL•FO
RenderX

46. Lavdas I, Glocker B, Rueckert D, Taylor S, Aboagye E, Rockall A. Machine learning in whole-body MRI: experiences and challenges from an applied study using multicentre data. Clin Radiol 2019;74(5):346-356. [doi: 10.1016/j.crad.2019.01.012] [Medline: 30803815]
47. Ponomariov V, Chirila L, Apipie F, Abate R, Rusu M, Wu Z, et al. Artificial intelligence versus doctors' intelligence: a glance on machine learning benefaction in electrocardiography. Discoveries (Craiova) 2017;5(3):e76 [FREE Full text] [doi: 10.15190/d.2017.6] [Medline: 32309594]
48. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc 2011;18(5):594-600 [FREE Full text] [doi: 10.1136/amiajnl-2011-000153] [Medline: 21846787]
49. Madkour M, Benhaddou D, Tao C. Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput Methods Programs Biomed 2016;128:52-68 [FREE Full text] [doi: 10.1016/j.cmpb.2016.02.007] [Medline: 27040831]
50. Ceusters W, Bona J. Ontological foundations for tracking data quality through the internet of things. Stud Health Technol Inform 2016;221:74-78. [Medline: 27071880]
51. Li F, Du J, He Y, Song HY, Madkour M, Rao G, et al. Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events. J Am Med Inform Assoc 2020;27(7):1046-1056 [FREE Full text] [doi: 10.1093/jamia/ocaa058] [Medline: 32626903]
52. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010;17(4):383-388 [FREE Full text] [doi: 10.1136/jamia.2010.004804] [Medline: 20595304]
53. Ceusters W, Smith B. Biomarkers in the ontology for general medical science. Stud Health Technol Inform 2015;210:155-159. [Medline: 25991121]
54. Gaebel J, Kolter T, Arlt F, Denecke K. Extraction of adverse events from clinical documents to support decision making using semantic preprocessing. Stud Health Technol Inform 2015;216:1030. [Medline: 26262330]
55. Garcia-Jimenez A, Moreno-Conde A, Martínez-García A, Marín-León I, Medrano-Ortega F, Parra-Calderón CL. Clinical decision support using a terminology server to improve patient safety. Stud Health Technol Inform 2015;210:150-154. [Medline: 25991120]
56. Jafarpour B, Abidi SR, Ahmad AM, Abidi SSR. INITIATE: an intelligent adaptive alert environment. Stud Health Technol Inform 2015;216:285-289. [Medline: 26262056]
57. Rosier A, Mabo P, Temal L, van Hille P, Dameron O, Deleger L, et al. Remote monitoring of cardiac implantable devices: ontology driven classification of the alerts. Stud Health Technol Inform 2016;221:59-63. [Medline: 27071877]
58. Lardon J, Asfari H, Souvignet J, Trombert-Paviot B, Bousquet C. Improvement of diagnosis coding by analysing EHR and using rule engine: application to the chronic kidney disease. Stud Health Technol Inform 2015;210:120-124. [Medline: 25991114]
59. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
60. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
61. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Sys 2023;2(2):81-83.
62. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol 2023;19(5):598-604. [doi: 10.1016/j.jpurol.2023.05.018] [Medline: 37328321]
63. Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. Lancet Digit Health 2019;1(1):e13-e14 [FREE Full text] [doi: 10.1016/S2589-7500(19)30002-0] [Medline: 33323236]
64. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
65. Bear Don't Walk OJ, Reyes Nieva H, Lee SSJ, Elhadad N. A scoping review of ethics considerations in clinical natural language processing. JAMIA Open 2022;5(2):ooac039 [FREE Full text] [doi: 10.1093/jamiaopen/ooac039] [Medline: 35663112]
66. Fu S, Wang L, Moon S, Zong N, He H, Pejaver V, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: a scoping review. Clin Transl Sci 2023;16(3):398-411 [FREE Full text] [doi: 10.1111/cts.13463] [Medline: 36478394]
67. Gauthier MP, Law JH, Le LW, Li JJ, Zahir S, Nirmalakumar S, et al. Automating access to real-world evidence. JTO Clin Res Rep 2022;3(6):100340 [FREE Full text] [doi: 10.1016/j.jtocrr.2022.100340] [Medline: 35719866]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

CASP: Critical Appraisal Skills Programme
CDS: clinical decision support
CDSS: clinical decision support system
cTAKES: Clinical Text Analysis and Knowledge Extraction System
EHR: electronic health record
LLM: large language model
MeSH: Medical Subject Headings
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 08.12.23; peer-reviewed by D Singh, L Zhu; comments to author 29.02.24; revised version received 20.04.24; accepted 24.07.24; published 30.09.24.

Please cite as:
Eguia H, Sánchez-Bocanegra CL, Vinciarelli F, Alvarez-Lopez F, Saigí-Rubió F
Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review
J Med Internet Res 2024;26:e55315
URL: https://www.jmir.org/2024/1/e55315
doi:10.2196/55315
PMID:

©Hans Eguia, Carlos Luis Sánchez-Bocanegra, Franco Vinciarelli, Fernando Alvarez-Lopez, Francesc Saigí-Rubió. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Patient Safety Incident Reporting and Learning Guidelines Implemented by Health Care Professionals in Specialized Care Units: Scoping Review

Tusiwe Mabel Gqaleni[1], PhD; Sipho Wellington Mkhize[1], PhD; Geldine Chironda[1], PhD

School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa

Corresponding Author:
Tusiwe Mabel Gqaleni, PhD
School of Nursing and Public Health
University of KwaZulu-Natal
238 Mazisi Kunene Road
College of Health Sciences
Durban, 4041
South Africa
Phone: 27 031260400 ext 1559
Fax: 27 0312604270
Email: gqalenit@ukzn.ac.za

Abstract

Background: Implementing Patient Safety Incident Reporting and Learning (PSIRL) guidelines is critical in guiding clinical practice and improving clinical outcomes in specialized care units (SCUs). There is limited research on the evidence of the implemented PSIRL guidelines in SCUs at the global level.

Objective: This review aims to map the evidence of PSIRL guidelines implemented by health care professionals in specialized care units globally.

Methods: A scoping review methodology, according to Joanna Briggs Institute, was adopted. The eligibility criteria were guided by the Population, Concept, and Context (PCC) framework, with the Population including health care professionals, the Concept including PSIRL guidelines, and the Context including specialized units globally. Papers written in English were searched from relevant databases and search engines. The PRISMA-ScR (Preferred Reporting Items for Scoping Reviews and Meta-Analyses extension for Scoping Reviews) checklist for used.

Results: The 13 selected studies were published from 2003 to 2023. Most articles are from the Netherlands and Switzerland (n=3), followed by South Africa (n=2). The nature of implemented PSIRL guidelines was computer-based (n=11) and paper-based incident reporting (n=2). The reporting system was intended for all the health care professionals within the specialized units, focusing on patients, staff members, and families. The outcomes of implemented incident reporting guidelines were positive, as evidenced by improved reporting of incidents, including medication errors (n=8) and decreased rate of incidents and errors (n=4). Furthermore, 1 study showed no change (n=1) in implementing the incident reporting guidelines.

Conclusions: The implementation of reporting of patient safety incidents (PSIs) in specialized units started to be reported around 2002; however, the frequency of yearly publications remains very low. Although some specialized units are still using multifaceted interventions and paper reporting systems in reporting PSIs, the implementation of electronic and computer-based reporting systems is gaining momentum. The effective implementation of an electronic-based reporting system should extend into other units beyond critical care units, as it increases the reporting of PSIs, reducing time to make an informed reporting of PSIs and immediate accessibility to information when needed for analysis. The evidence on the implementation of PSI reporting guidelines in SCUs comes from 5 different continents (Asia, Africa, Australia, Europe, and North America). However, the number identified for certain countries within each continent is very minimal.

(J Med Internet Res 2024;26:e48580) doi:10.2196/48580

XSL•FO
RenderX

KEYWORDS

patient safety incidents; adverse events; harm; near misses; reporting guidelines; implementation guidelines; implementation practices; intervention strategies; critical care units; intensive care units; patient safety; specialized care unit; guidelines; clinical practice; healthcare professional; ICU

Introduction

Background

Reducing the occurrence of patient safety incidents (PSIs) in the health care system has become a global concern. According to the World Health Organization (WHO) [1], the health care system has still demonstrated unacceptably high rates of PSIs and preventable deaths [1]. Patient Safety remains crucial in the improvement of quality patient care and has been defined by the WHO International Classification for Patient Safety as the reduction of the risk of unnecessary and avoidable harm associated with health care to an acceptable minimum [2,3]. Specialized units (critical care and high care units) are no exception, as critically ill patients tend to be more susceptible and exposed to a complex environment, therefore incurring high rates of preventable PSIs and death [4]. The SCU environment is different from the general wards as it is characterized by highly invasive and complicated procedures that make the patients vulnerable and susceptible to PSIs, leading to prolonged lengths of stay in the hospital. Furthermore, these critically ill patients have comorbidities that are life-threatening, leading to various complications that require immediate interventions. Near misses and PSIs require constant surveillance to improve patient safety in acute and critical care units [5]. Near misses are incidents or situations that have the potential to cause harm but did not reach the patient due to timely intervention, whereas PSI is harm caused by medical mismanagement instead of the underlying disease [6]. PSIs contribute to the cost of care, adding to the burden of the patient and, because of malpractice claims, causing mounting and spiraling costs to the health care system and for society at large [7].

Evidence revealed that in high-income countries, it was estimated that 1 in every 10 patients was harmed while receiving hospital care [8]. This harm might have further increased the length of hospitalization, and use of more health care resources, with cost implications. In low and middle-income countries, 134 million PSIs occurred in hospitals due to unsafe care, which resulted in 2.6 million deaths each year (National Academies of Sciences and Medicine) [9]. A study carried out in the Eastern Mediterranean and Africa revealed that almost one-third of patients who experienced PSIs died, and 4 out of 5 of those incidences were preventable [10]. In South Africa, a study conducted in KwaZulu-Natal revealed that PSIs were still high (47%) and were serious in nature, which might have suggested poor implementation of Patient Safety Incident Reporting and Learning (PSIRL) guidelines and a lack of improvement strategies [11,12]. Therefore, mitigation of the occurrence of PSIs remains an important component in rendering quality patient care and improving clinical outcomes.

Implementation guidelines are critical in guiding the clinical practice and improvement of clinical outcomes. Rosa et al [13] and Weiss [14] affirmed that incorporating evidence into critical care practice is recognized as a crucial requirement for the optimal care of critically ill patients. However, implementation of evidence-based practices is often insufficient due to many barriers, resulting in frequent poor adherence to guideline recommendations in critical care settings [13-16]. In response to mitigate the occurrence of PSIs, a global effort was made by the WHO Member States to develop the implementation intervention strategies relevant to their nations to create a safer environment in the health care system [10]. The World Alliance for Patient Safety first drafted the guidelines for a system to report and learn adverse events, which were updated and revised as WHO Guidelines for PSIRL Systems. In South Africa, it was recommended by the National Department of Health that every health establishment was expected to adhere to the PSRIL system as stipulated in this guideline [1]. A patient safety learning system (sometimes called a critical incident reporting system) refers to structured reporting, collation, and analysis of critical incidents [17]. Nevertheless, failure to reduce the PSI occurrence might be related to the poor implementation of PSIRL guidelines, which might have led to negative clinical outcomes, which made it difficult for the policy makers and health care professionals to handle PSIRL guidelines effectively. There is limited research on evidence that looks at the implementation of PSIRL guidelines in specialized units at a global level.

Aim and Questions of the Review

The aim of this review is to map the evidence of PSIRL guidelines implemented by health care professionals in specialized care units (SCUs) globally. The broad question of the review is as follows: what evidence exists on the implementation of PSIRL guidelines by health care professionals? What are the gaps in the implementation of PSIRL guidelines in SCUs?

Methods

Overview

The Joanna Briggs Institute Scoping Review Methods (2020) and scoping reviews described in the 2020 JBI Manual for Evidence Synthesis [18] were used to map the available evidence on PSIRL guidelines implemented by health care professionals in SCUs. A scoping review protocol was developed and registered with the Open Science Framework. We used the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Multimedia Appendix 1) checklist: a checklist and explanations guiding the reporting to ensure that the review conforms to the reporting standards of a scoping review [19].

Eligibility Criteria

The eligibility criteria were guided by the Population, Concept, and Context (PCC) framework, language, timeline, and type of articles, as illustrated in Table 1.

XSL•FO
RenderX

Table 1. PCC[a], language, and timeline to determine the eligibility.

Variable	Inclusion criteria	Exclusion criteria
Population	Heath care workers, health care providers, health care professionals, health personnel, allied health care professionals, nurses, ICU[b] nurses, intensive care nurses, critical care nurses, and medical doctors.	Health care professionals working in nonspecialized units.
Concept	implementation, practices, intervention strategies, patient safety incident reporting guidelines, voluntary patient safety event reporting, risk management, reporting guidelines, patient safety learning systems, critical incident reporting system, adverse events; errors, critical incident, incident reports, and hospital risk reporting.	Studies on the implementation of patient safety reporting guidelines in medical, surgical, theatre, and emergency departments were excluded.
Context	Specialized or specialized care units, ICUs, critical care units, coronary care units, or renal units, or burns units or high care units, worldwide; globally, African continent, European continent, Asian continent, American continent, Australasian continent, WHO[c] regions, and United Nations regions.	Studies written in non-English languages and articles on the implementation of PSIRL[d] in medical, surgical, theatre, and emergency departments were excluded.
Language	Studies written in English were published as of January 2002.	Studies written in non-English languages.
Timeline	From January 2002 to December 2023.	December 2001 and backwards.
Type of studies	Quantitative, qualitative, mixed methods, review articles (systematic, meta-analysis, integrative, and scoping reviews), reports, text and opinion papers, gray literature sources (academic outputs in the form of theses, dissertations, and ongoing research), and professional organizations such as WHO).	Quantitative, qualitative, mixed methods, reviews, reports, text and opinion papers, gray literature sources, and professional organizations with no outcome of interest.

[a]PCC: Population, Concept, and Context.

[b]ICU: intensive care unit.

[c]WHO: World Health Organization.

[d]PSIRL: Patient Safety Incident Reporting and Learning.

Search Terms

The following search terms and electronic databases were used to identify articles for the scoping review (Table 2).

Table 2. Search words.

Criteria	Entries
MeSH[a] terms	Health care professionals, health plan implementation, hospital risk reporting, globally
Search words	(Health care professionals) OR (Heath care workers or health Care Provider or health personnel) OR (Nurses or ICU nurses or intensive care nurses or critical care nurses) OR (medical doctors) AND Implementation or (practices or Intervention strategies) AND (Patient Safety Incident Reporting guidelines) OR (Voluntary Patient Safety Event Reporting or Risk management or Reporting guidelines or Patient safety learning systems or Critical Incident reporting system or Adverse events or errors or critical incident or incident reports or Hospital risk reporting) AND (Specialised or Specialized care units) OR (ICUs; critical care units or Coronary care unit or Renal units or Burns unit or High care units) AND (Worldwide) or (Globally) or (African continent or European continent or Asian continent or American continent or Australasian continent) OR (WHO regions) OR (United Nations regions)

[a]MeSH: Medical Subject Headings.

Search Strategy

This is a secondary analysis where the authors screened electronically published health literature from different databases on patient and safety incident reporting electronic systems within the specialized care units in a hospital setting. The researchers followed a 3-step search strategy to find evidence related to the implementation of PSIRL guidelines implemented by health care professionals in SCUs. The reviewers involved the research librarian in designing and refining the search. First, 2 appropriate databases, namely MEDLINE (PubMed or Ovid) and CINAHL, were searched. Thereafter, an analysis of the text words contained in the title and abstract of retrieved papers and the index terms used to describe the studies were followed. A second search was done using all identified keywords and index terms in all the remaining databases, namely PubMed, EBSCO Host, Web of Science, Scopus, African Journals Online, and Sabinet. A search for gray literature was done to locate unpublished evidence, including academic outputs (theses and dissertations) and ongoing research. The authors search gray literature for completed unpublished academic outputs (theses and dissertations) for further evidence on PSI reporting guidelines implemented by the health care professionals in specialized care units. This was accomplished through searching the ProQuest Dissertation and Theses Global, then search engines Google and Google Scholar. Professional organizations such as WHO were also searched. Finally, the reference list of identified articles was searched for additional sources. A sample of a complete search strategy for at least one major database is included in an appendix to the protocol.

XSL•FO
RenderX

Source of Evidence Selection

The process of source selection was done in 2 stages. The first stage was based on title and abstract examination using the PCC inclusion criteria; thereafter, full-text examination followed as the second stage. All the stages of the review were done by 2 reviewers (TMHG and GC) independently, and any disagreements were solved by consensus. A flowchart of the review process (from the PRISMA-ScR statement) detailing the flow from the search through source selection, duplicates, full-text retrieval, and any additions from the third search, data extraction, and presentation of the evidence was availed. The EndNote (Clarivate) and Ryann (Rayyan Systems, Inc) software were used to manage the search results.

Data Extraction Process, Presentation, and Analysis

Data extraction and verification were done by 2 reviewers (TMHG & GC). A logical and descriptive summary of the results that align with the specific questions was presented in the form of a charting table. The data included the following key information: authors, year of publication, country where the study was done, populations involved, study methods, guidelines or strategies implemented, and outcomes. Analysis of evidence was done through frequency counts of concepts, population, and context. A narrative summary was done to describe the existing evidence on the implementation of PSIRL guidelines by the health care professionals working in SCUs.

Results

Overview

The authors conducted secondary research where the electronically published health literature from different databases on patient and safety incident reporting electronic systems within the SCUs were searched. According to Figure 1, a total of 146 articles were identified from the databases and the search engine Google Scholar. Before the selection process, 6 duplicates were removed to remain with 140 articles for abstract and title screening. After abstract and title screening, 96 articles were excluded based on the following reasons: wrong population (n=7), wrong concept (n=30), and wrong context (n=59). This resulted in the remaining 44 articles for full-text screening. A total of 3 full-text studies could not be retrieved and hence excluded to remain with the 41 full-text studies for second-level screening. Full-text screening yielded the exclusion of 28 articles based on the following reasons: wrong population (n=2), wrong concept (n=13), wrong context (n=9), and unclear outcomes (n=4). As illustrated in Figure 1 below, 13 articles were found to be fitting the inclusion of this scoping review (19-31).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

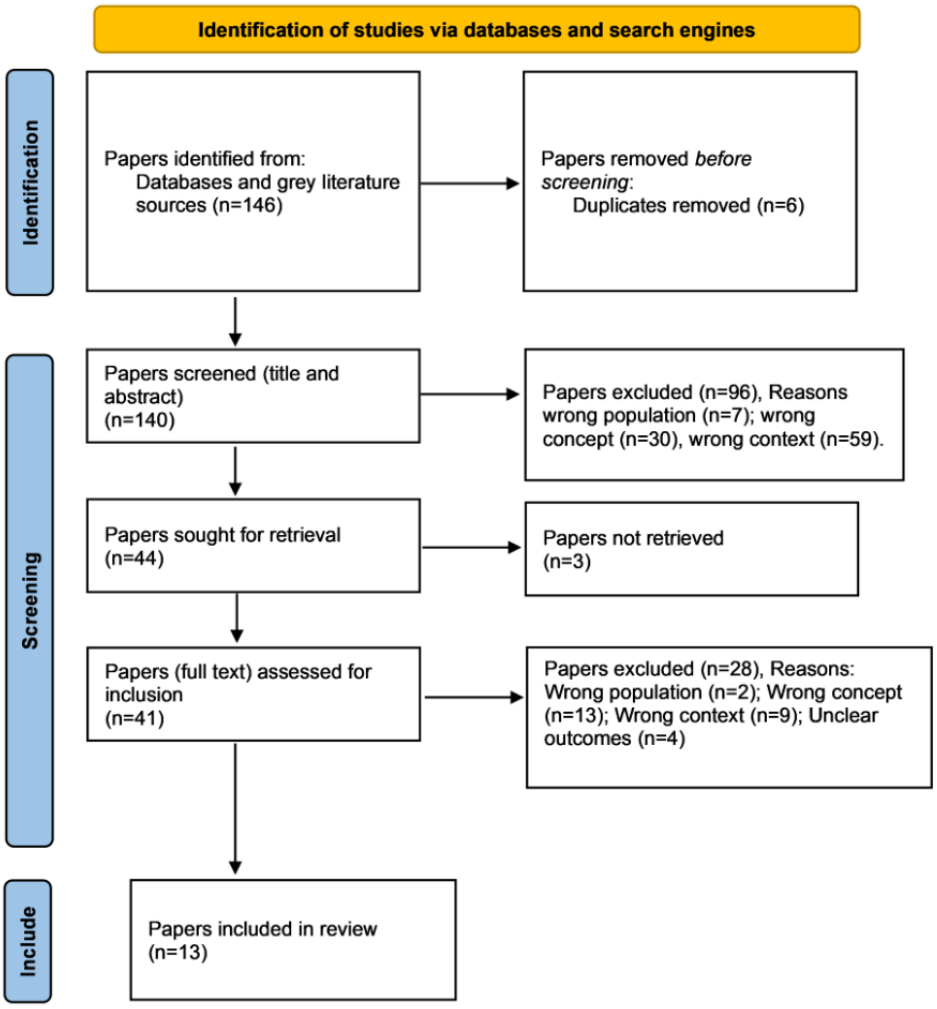

XSL•FO
RenderX

Publication Trends, Distribution, and Characteristics

The articles used quantitative (n=11) and review (n=2) methodologies, and one was not stated. The 13 selected articles (Table S1 in Multimedia Appendix 2) were published in 2003 (n=1), 2007 (n=1), 2009 (n=1), 2010 (n=1); 2011 (n=1), 2012 (n=1), 2013 (n=1), 2014 (n=1), 2017 (n=2), 2018 (n=1), 2020 (n=1), and 2023 (n=1). The articles were identified from 4 different continents, namely Asia (n=2), Africa (n=2), Australia (n=1), North America (n=1), and Europe (n=8). Most articles come from the Netherlands and Switzerland (both with n=3), followed by South Africa (n=2). Other specific countries like Spain, Ireland, Jordan, Australia, North America, and Japan had only 1 article each.

Patient Safety Incident Reporting and Learning Guidelines

The nature of implemented patient safety incident reporting and learning guidelines was computer-based (n=11) and paper-based incident reporting (n=2). The reporting system was for all the health care professionals within the SCUs with a focus on patients, staff members, and families. The outcomes of implemented incident reporting and learning guidelines were positive, as evidenced by improved reporting of incidents, including medication errors (n=9). The strategies implemented in these studies included electronic reporting, a voluntary-nonpunitive reporting system, a medication error checklist, an intensive care unit (ICU) incident registry, and a multifaceted intervention that involved the creation of patient safety peer-leadership role, feedback process, interactive dashboards for patient safety data, and education resources accessible through quick response codes, improved the reporting of PSIs. Furthermore, studies that revealed a decreased rate of incidents and errors (n=2) executed various strategies, including clinical information systems, critical patient transport protocols, and handoff communication processes. In addition, 2 studies did not show any change (n=2) in the implementation of the incident reporting and learning guidelines (Table S1 in Multimedia Appendix 2).

Discussion

Principal Findings

The review aimed to map the evidence of PSIRL guidelines implemented by health care professionals in SCUs globally. The review was specifically focusing more on the SCUs. However, 4 articles might be nonspecialized ICUs, but they were standardized to be used across the hospital, including the specialized units. The evidence on the implementation of PSIRL guidelines in SCUs comes from 5 different continents (Asia, Africa, Australia, Europe, and North America). However, the number of identified countries within each continent is very minimal. Yet globally, there is a higher percentage of PSIs in SCUs [12]. The implementation of reporting of PSIs in SCUs started to be reported around 2002. However, the frequency of yearly publications remains very low. Meanwhile, health care systems in high-income and transitioning countries have implemented the patient safety reporting system, challenges still exist to reach full scale [20]. As safety culture promoters, incident reporting systems (IRS) serve as a starting point of the learning process to prevent the occurrence of the same incident in the future [21,22].

Implementation of the IRS for risk identification and organizational learning is one way of improving patient safety in health care, including specialized settings [23]. The nature of implemented incident reporting guidelines identified in the review were computer-based reporting [24-31], paper-based incident reporting [32-34], and handoff communication [35] The use of electronic and computer-based reporting system is gaining momentum due to its effectiveness in increasing the reporting of PSI events, reducing time to make an informed reporting of PSIs and immediate accessibility to information when needed for analysis [29,36,37]. Adopting the electronic safety program in specialized units like critical care units will enhance the quality of services that are provided for patients as indicated by Muhsein et al [25].

The paper-based reporting system of PSIs has proved to be inefficient as fewer staff members are willing to report the incidents [38]. However, some SCUs are still using multifaceted paper reporting systems in reporting PSIs, errors related to medication, and critical transportation of patients with positive results [32-34]. On the other hand, Fraenkel et al [26] implemented a computerized clinical information system that replaced paper-based charts of patient observations, clinical records, results reporting, and drug prescribing, which resulted in the reduction of the occurrence of PSIs, less documentation, and more time spent on the patient. According to Ramírez et al [39], the implementation of a hospital IRS, including the systematization of the method and analysis of PSIs by the workshop-trained, results in a reduction in the frequency of PSIs.

In some cases, reporting PSIs and handling them at a unit level (SCUs) is not adequate for developing patient safety, hence the need to use multiple methods to strengthen the overall patient safety culture [23,30]. Griffeth et al [30], involved creation of patient safety peer-leadership role, feedback process, interactive dashboards for patient safety data, and education resources accessible through quick response codes, which resulted in patient safety incident reporting increased by 48%. van der Veer et al [29], used the concept of multiple methods where the ICU incident registry was added to the existing registry, and this resulted in double the number of PSIs reported. In addition, the handoff communication process has proved to lower the number of handoff-related incidents and enhance the satisfaction of nurses [35]. While strong evidence of the current review has revealed the positive outcomes of implementing computer-based reporting systems in SCUs such as critical care units, the systematic review carried out by Frey and Schwappach [31] indicated no improvement in critical incident monitoring after the implementation. Similarly, the use of the multifaceted paper reporting strategy highlighted no change for the PSIs related to airway and indwelling lines [34].

Strengths

Scoping reviews ensure high-quality articles are included for data extraction. Publications emanating from different countries worldwide depicted comprehensive information on the implementation of the reporting guidelines and strategies, which

XSL•FO
RenderX

provided further recommendations. Evidence-based information from the results will be used by policy makers to improve patient safety culture in SCUs.

Limitations

The review was limited to SCUs only, therefore, pertinent information from other units and other categories of health care professionals may have been disregarded. In addition, only studies published in English were used; therefore, studies published in other languages whose information may have had a valuable contribution were excluded. There is a paucity of literature on the implementation of PSI reporting and learning guidelines worldwide. Therefore, future research studies need to be conducted, especially in Africa.

Conclusion

The implementation of reporting of PSIs in specialized units started to be reported around 2002; however, the frequency of yearly publications remains very low. Although some specialized units are still using multifaceted interventions and paper reporting systems in reporting PSIs, the implementation of electronic and computer-based reporting systems is gaining momentum. The effective implementation of an electronic-based reporting system should extend into other units beyond critical care units, as it increases the reporting of PSIs, reducing time to make an informed reporting of PSIs and immediate accessibility to information when needed for analysis. The evidence on the implementation of PSI reporting guidelines in SCUs comes from 5 different continents (Asia, Africa, Australia, Europe, and North America). However, the number identified for certain countries within each continent is very minimal.

Acknowledgments

The authors would like to acknowledge the University of KwaZulu-Natal for supporting the publication of this paper. The study was funded by the National Research Foundation Thuthuka funding.

Authors' Contributions

TMG and GC contributed to the conceptualization of the scoping review, data extraction, analysis of results, drafting and editing of the manuscript, and approval of the final version. SWM contributed to the conceptualization and editing of the manuscript and approved the final version of this results paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews extension for Scoping Reviews) checklist.
[DOCX File , 84 KB - jmir_v26i1e48580_app1.docx]

Multimedia Appendix 2
Author, year, country, research design, nature of patient safety incident guidelines, type of health care professional, population recipient, the outcome, and recommendations (n=13).
[DOCX File , 24 KB - jmir_v26i1e48580_app2.docx]

References

1. World Health Organization. National guideline for patient safety incident reporting and learning in the public health sector of South Africa. URL: https://test.idealhealthfacility.org.za/App/Document/Download/53 [accessed 2024-09-05]
2. World Alliance For Patient Safety Drafting Group, Sherman H, Castro G, Fletcher M, World Alliance for Patient Safety, Hatlie M, et al. Towards an international classification for patient safety: the conceptual framework. Int J Qual Health Care 2009;21(1):2-8 [FREE Full text] [doi: 10.1093/intqhc/mzn054] [Medline: 19147595]
3. Runciman W, Hibbert P, Thomson R, Van Der Schaaf T, Sherman H, Lewalle P. Towards an international classification for patient safety: key concepts and terms. Int J Qual Health Care 2009;21(1):18-26 [FREE Full text] [doi: 10.1093/intqhc/mzn057] [Medline: 19147597]
4. Ahmed AH, Thongprayoon C, Schenck LA, Malinchoc M, Konvalinová A, Keegan MT, et al. Adverse in-hospital events are associated with increased in-hospital mortality and length of stay in patients with or at risk of acute respiratory distress syndrome. Mayo Clin Proc 2015;90(3):321-328. [doi: 10.1016/j.mayocp.2014.12.015] [Medline: 25638301]
5. Henneman EA, Gawlinski A, Giuliano KK. Surveillance: a strategy for improving patient safety in acute and critical care units. Critical Care Nurse 2012;32(2):e9-e18 [FREE Full text] [doi: 10.4037/ccn2012166]
6. da Silva MVO, Caregnato RCA. Intensive care unit: safety and monitoring of adverse events. Journal of Nursing UFPE/Revista de Enfermagem UFPE 2019:13 [FREE Full text] [doi: 10.5205/1981-8963.2019.239368]
7. Oyebode F. Clinical errors and medical negligence. Med Princ Pract 2013;22(4):323-333 [FREE Full text] [doi: 10.1159/000346296] [Medline: 23343656]

XSL•FO
RenderX

8. Slawomirski L, Auraaen A, Klazinga NS. The economics of patient safety: strengthening a value-based approach to reducing patient harm at national level.: OECD; 2017. URL: http://www.oecd.org/employment/ [accessed 2024-09-05]
9. Giles S, Fletcher M, Baker M, Thomson R. Incident reporting and analysis. URL: https://tinyurl.com/3mnvuacy [accessed 2024-09-05]
10. World Health Organization (WHO). Patient Safety Tool Kit. Cairo, Egypt: WHO Regional Office for the Eastern Mediterranean; 2015.
11. Gqaleni T, Mkhize S. Healthcare professionals' perception of knowledge and implementation of patient safety incident reporting and learning guidelines in specialised care units, KwaZulu-Natal. South Afr J Crit Care 2023;39(1):25-29 [FREE Full text] [doi: 10.7196/SAJCC.2023.v39i1.559] [Medline: 37521960]
12. Gqaleni TM, Bhengu BR. Analysis of patient safety incident reporting system as an indicator of quality nursing in critical care units in KwaZulu-Natal, South Africa. Health SA 2020;25:1263 [FREE Full text] [doi: 10.4102/hsag.v25i0.1263] [Medline: 32284886]
13. Rosa RG, Teixeira C, Sjoding M. Novel approaches to facilitate the implementation of guidelines in the ICU. J Crit Care 2020;60:1-5. [doi: 10.1016/j.jcrc.2020.07.014] [Medline: 32731099]
14. Weiss CH. Why do we fail to deliver evidence-based practice in critical care medicine? Curr Opin Crit Care 2017;23(5):400-405 [FREE Full text] [doi: 10.1097/MCC.0000000000000436] [Medline: 28858917]
15. Costa DK, White MR, Ginier E, Manojlovich M, Govindan S, Iwashyna TJ, et al. Identifying barriers to delivering the awakening and breathing coordination, delirium, and early exercise/mobility bundle to minimize adverse outcomes for mechanically ventilated patients: a systematic review. Chest 2017;152(2):304-311 [FREE Full text] [doi: 10.1016/j.chest.2017.03.054] [Medline: 28438605]
16. Jeffs L, Hayes C, Smith O, Mamdani M, Nisenbaum R, Bell C, et al. The effect of an organizational network for patient safety on safety event reporting. Eval Health Prof 2014;37(3):366-378. [doi: 10.1177/0163278713491267] [Medline: 23754848]
17. Health Quality Ontario. Patient safety learning systems: a systematic review and qualitative synthesis. Ont Health Technol Assess Ser 2017;17(3):1-23 [FREE Full text] [Medline: 28326148]
18. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
19. Tricco AC, Lillie E, Zarin W, O'Brien K, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
20. Koike D, Ito M, Horiguchi A, Yatsuya H, Ota A. Implementation strategies for the patient safety reporting system using consolidated framework for implementation research: a retrospective mixed-method analysis. BMC Health Serv Res 2022;22(1):409 [FREE Full text] [doi: 10.1186/s12913-022-07822-9] [Medline: 35346182]
21. Torre-Pérez LDL, Granés L, Prat Marín A, Bertran M. A hospital incident reporting system (2016-2019): learning from notifier's perception on incidents' risk, severity and frequency of adverse events. J Healthc Qual Res 2023;38(2):93-104 [FREE Full text] [doi: 10.1016/j.jhqr.2022.08.004] [Medline: 36151046]
22. Harsul W, Irwan AM, Sjattar EL. The relationship between nurse self-efficacy and the culture of patient safety incident reporting in a district general hospital, Indonesia. Clinical Epidemiology and Global Health 2020;8(2):477-481 [FREE Full text] [doi: 10.1016/j.cegh.2019.10.013]
23. Sahlström M, Partanen P, Turunen H. Patient-reported experiences of patient safety incidents need to be utilized more systematically in promoting safe care. Int J Qual Health Care 2018;30(10):778-785. [doi: 10.1093/intqhc/mzy074] [Medline: 29668942]
24. Kabane S. An Evaluation of the Effectiveness of a Hospital Clinical Adverse Event Prevention Programme. Pretoria, South Africa: University of Pretoria; 2013.
25. Muhsein AS, Al-Slehat MA, Ghurra MFB, Al-Shnaaq AAA, Salman ZS, Saleh AM. The impact of implementing electronic safety program on patient safety in intensive care unit, Istishari Hospital. JMENAS 2017;3(2):35-40. [doi: 10.12816/0034678]
26. Fraenkel DJ, Cowie M, Daley P. Quality benefits of an intensive care clinical information system. Crit Care Med 2003;31(1):120-125. [doi: 10.1097/00003246-200301000-00019] [Medline: 12545004]
27. Snijders C, Kollen BJ, van Lingen RA, Fetter WPF, Molendijk H, NEOSAFE Study Group. Which aspects of safety culture predict incident reporting behavior in neonatal intensive care units? A multilevel analysis. Crit Care Med 2009;37(1):61-67. [doi: 10.1097/CCM.0b013e31819300e4] [Medline: 19050606]
28. Brunsveld-Reinders AH, Arbous MS, De Vos R, De Jonge E. Incident and error reporting systems in intensive care: a systematic review of the literature. Int J Qual Health Care 2016;28(1):2-13. [doi: 10.1093/intqhc/mzv100] [Medline: 26660441]
29. van der Veer S, Cornet R, de Jonge E. Design and implementation of an ICU incident registry. Int J Med Inform 2007;76(2-3):103-108. [doi: 10.1016/j.ijmedinf.2006.08.003] [Medline: 17035080]
30. Griffeth EM, Gajic O, Schueler N, Todd A, Ramar K. Multifaceted intervention to improve patient safety incident reporting in intensive care units. Journal of patient safety 2023;19(7):422-428 [FREE Full text] [doi: 10.1097/pts.0000000000001151]

XSL•FO
RenderX

31. Frey B, Schwappach D. Critical incident monitoring in paediatric and adult critical care: from reporting to improved patient outcomes? Curr Opin Crit Care 2010;16(6):649-653. [doi: 10.1097/MCC.0b013e32834044d8] [Medline: 20930624]
32. Truter A, Schellack N, Meyer JC. Identifying medication errors in the neonatal intensive care unit and paediatric wards using a medication error checklist at a tertiary academic hospital in Gauteng, South Africa. SAJCH 2017;11(1):5-10 [FREE Full text] [doi: 10.7196/sajch.2017.v11i1.1101]
33. García PN, Avión RC, Ruiloba ML, Pérez JR, Dobarro AB, García AR. Retrospective study of security in the transfer of critical patients after application of methodology for risk management. Revista Española de Anestesiología y Reanimación (English Edition) 2020;67(3):119-129 [FREE Full text] [doi: 10.1016/j.redare.2019.10.006]
34. Pagnamenta A, Rabito G, Arosio A, Perren A, Malacrida R, Barazzoni F, et al. Adverse event reporting in adult intensive care units and the impact of a multifaceted intervention on drug-related adverse events. Ann Intensive Care 2012;2(1):47 [FREE Full text] [doi: 10.1186/2110-5820-2-47] [Medline: 23174137]
35. Ibrahim MA. Improving nursing handoff process in the cardiovascular intensive care unit.: Royal College of Surgeons in Ireland; 2014. URL: https://repository.rcsi.com/articles/thesis/Improving_Nursing_Handoff_Process_in_the_Cardiovascular_Intensive_Care_Unit/10811981?file=19322018 [accessed 2024-09-05]
36. Kanda H. Development of an online incident-reporting system for management of medical risks at hospital. Yakugaku Zasshi 2011;131(9):1353-1359 [FREE Full text] [doi: 10.1248/yakushi.131.1353] [Medline: 21881311]
37. Giles S, Fletcher M, Baker M, Thomson R. Incident reporting and analysis. Patient Safety: Research into Practice. 2005. URL: https://tinyurl.com/zxjs55v8 [accessed 2024-09-05]
38. Gao X, Yan S, Wu W, Zhang R, Lu Y, Xiao S. Implications from China patient safety incidents reporting system. Therapeutics and clinical risk management 2019;15:259-267 [FREE Full text] [doi: 10.2147/tcrm.s190117]
39. Ramírez E, Martín A, Villán Y, Lorente M, Ojeda J, Moro M, SINOIRES Working Group. Effectiveness and limitations of an incident-reporting system analyzed by local clinical safety leaders in a tertiary hospital: Prospective evaluation through real-time observations of patient safety incidents. Medicine (Baltimore) 2018;97(38):e12509 [FREE Full text] [doi: 10.1097/MD.0000000000012509] [Medline: 30235764]

Abbreviations

ICU: intensive care unit
IRS: incident reporting systems
PCC: Population, Concept, and Context
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PSIRL: Patient Safety Incident Reporting and Learning
PSI: patient safety incident
SCU: specialized care unit
WHO: World Health Organization

Edited by A Mavragani; submitted 02.05.23; peer-reviewed by F Al Dhabbari, V Borza, J Walsh; comments to author 09.04.24; revised version received 31.05.24; accepted 23.07.24; published 04.10.24.

Please cite as:
Gqaleni TM, Mkhize SW, Chironda G
Patient Safety Incident Reporting and Learning Guidelines Implemented by Health Care Professionals in Specialized Care Units: Scoping Review
J Med Internet Res 2024;26:e48580
URL: https://www.jmir.org/2024/1/e48580
doi:10.2196/48580
PMID:

©Tusiwe Mabel Gqaleni, Sipho Wellington Mkhize, Geldine Chironda. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Considering User Experience and Behavioral Approaches in the Design of mHealth Interventions for Atrial Fibrillation: Systematic Review

Sagar Suresh Kumar[1*], MSc; Patricia Connolly[2*], Prof Dr; Anja Maier[1*], Prof Dr

[1]Department of Design, Manufacturing and Engineering Management (DMEM), University of Strathclyde, Glasgow, United Kingdom

[2]Department of Biomedical Engineering, University of Strathclyde, Glasgow, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Sagar Suresh Kumar, MSc
Department of Design, Manufacturing and Engineering Management (DMEM)
University of Strathclyde
James Weir Building
75 Montrose Street
Glasgow, G1 1XJ
United Kingdom
Phone: 44 1415482091
Email: sagar.suresh-kumar@strath.ac.uk

Abstract

Background: Atrial fibrillation (AF) is a leading chronic cardiac disease associated with an increased risk of stroke, cardiac complications, and general mortality. Mobile health (mHealth) interventions, including wearable devices and apps, can aid in the detection, screening, and management of AF to improve patient outcomes. The inclusion of approaches that consider user experiences and behavior in the design of health care interventions can increase the usability of mHealth interventions, and hence, hopefully, yield an increase in positive outcomes in the lives of users.

Objective: This study aims to show how research has considered user experiences and behavioral approaches in designing mHealth interventions for AF detection, screening, and management; the phases of designing complex interventions from the UK Medical Research Council (MRC) were referenced: namely, identification, development, feasibility, evaluation, and implementation.

Methods: Studies published until September 7, 2022, that examined user experiences and behavioral approaches associated with mHealth interventions in the context of AF were extracted from multiple databases. The PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines were used.

Results: A total of 2219 records were extracted, with only 55 records reporting on usability, user experiences, or behavioral approaches more widely for designing mHealth interventions in the context of AF. When mapping the studies onto the phases of the UK MRC's guidance for developing and evaluating complex interventions, the following was found: in the identification phase, there were significant differences between the needs of patients and health care workers. In the development phase, user perspectives guided the iterative development of apps, interfaces, and intervention protocols in 4 studies. Most studies (43/55, 78%) assessed the usability of interventions in the feasibility phase as an outcome, although the data collection tools were not designed together with users and stakeholders. Studies that examined the evaluation and implementation phase entailed reporting on challenges in user participation, acceptance, and workflows that could not be captured by studies in the previous phases. To realize the envisaged human behavior intended through treatment, review results highlight the scant inclusion of behavior change approaches for mHealth interventions across multiple levels of sociotechnical health care systems. While interventions at the level of the individual (micro) and the level of communities (meso) were found in the studies reviewed, no studies were found intervening at societal levels (macro). Studies also failed to consider the temporal variation of user goals and feedback in the design of long-term behavioral interventions.

Conclusions: In this systematic review, we proposed 2 contributions: first, mapping studies to different phases of the MRC framework for developing and evaluating complex interventions, and second, mapping behavioral approaches to different levels of health care systems. Finally, we discuss the wider implications of our results in guiding future mHealth research.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e54405) doi:10.2196/54405

KEYWORDS

atrial fibrillation; wearable devices; lifestyle modification; user experience; design for behavior change; systems thinking; cardiac disease; stroke; heart disease; complication; mobile health; systematic review; usability; mHealth; intervention

Introduction

The Chronic Condition of Atrial Fibrillation and the Goal of Clinical Interventions

Atrial fibrillation (AF) is one of the leading heart diseases in today's world and affects more than 0.5% of the global population. It is caused by irregular pulses in the heart atria, which leads to an increased risk of stroke, heart failure, mortality, and other poor outcomes [1]. The integrated model of multimorbidity and symptom science attributes the progression of chronic conditions, such as AF, to the interplay of various factors across the following 3 domains: risk factors, interventions or conditions or symptoms, and outcomes [2]. AF has both nonmodifiable risk factors, such as age and genetics, as well as several modifiable and lifestyle-based risk factors, such as hypertension, alcohol, and obesity, which account for 30% to 40% of the risk of developing AF [3]. These risk factors can also lead to the co-occurrence of other diseases, such as myocardial infarction, thereby worsening health outcomes [4]. The success of chronic disease interventions lies in their ability to impede or reverse the transition of risk factors to poor patient outcomes through different means [2].

Challenges Faced by Existing AF Clinical Interventions

In the case of AF, early detection and subsequent management through catheter ablation and other interventions can prevent up to 97% of cases from progressing to persistent and permanent forms [5], which are associated with poorer health outcomes, as well as increased long-term costs [6]. However, up to 40% of early cases may be asymptomatic and episodic or paroxysmal, that is, they may not be able to be detected through traditional diagnostic interventions such as routine checkups. Instead, they require continuous monitoring outside of hospitals and other traditional care environments [7]. Moreover, traditional clinical guidelines and their concomitant interventions mostly do not consider the different challenges faced by patients in their respective journeys and daily lives in their target outcomes [8]. The burden of AF goes beyond physical outcomes and also takes a huge toll on the social well-being of patients, by impeding their ability to perform day-to-day activities, affecting their fulfillment of social roles, and burdening family members and caregivers [9]. AF prevalence is expected to increase by more than 60% in the next 30 years. As the condition affects up to 17% of people older than 80 years, the increase in life expectancy, as well as the global population, is expected to pose significant social and economic challenges to the world, especially in middle- and high-income countries [1]. In the United Kingdom, AF care currently accounts for 0.9% to 1.6% of the annual expenditure of the National Health Service and is expected to increase to up to 4.27% over the next 2 decades [10]. The National Health Service has been overwhelmed due to the growing demand for cardiovascular care, with the number of people at the end of 2022 waiting for a heart-related treatment increasing by 280 times compared to before the pandemic [11].

The Potential for Mobile Health in Alleviating Said Challenges

In this regard, mobile health (mHealth) devices consisting of wearable devices and apps can alleviate this inaccessibility of care and workforce constraints by offering alternative and personalized interventions to support care services in the detection and management of chronic conditions [12]. In the detection of AF, different sensing modalities, such as electrocardiography (ECG) [13] and photoplethysmography (PPG) [14], have been integrated within wearable devices, such as smartwatches and smartphones, to facilitate continuous monitoring, which is essential in detecting paroxysmal and asymptomatic cases. Mobile apps can aid patients in improving their lifestyles and reducing risk factors, which in conjunction with regular care can improve recovery, as well as reduce symptomatic burden [15]. For example, a 10% weight reduction in individuals with obesity can result in a 27% reduction in the risk of AF recurrence following catheter ablation [16]. Apps can also assist patients in managing AF by improving medication adherence, increasing their knowledge, improving patient-physician interactions, and so on [17-19].

The Need to Consider User Experiences and Usability of mHealth Interventions

While mHealth technology may have promising potential to improve AF care [20], we need to go beyond its potential clinical benefits and also consider the experiences and perspectives of patients, as well as the implications of introducing such interventions in people's lives. Emotions and feelings experienced by patients as a result of their diagnosed conditions can greatly impact the usability of interventions [21]. Additionally, environmental and individual risk factors can impact an individual's emotional comfort, which is defined as a state of well-being where patients are satisfied with the role of clinical intervention in improving their respective journeys in life [22]. In other chronic disease contexts, patients' emotional comfort with regard to mobile and digital interventions has included various aspects of their daily lives such as diet, physical activity, work, and social occasions [21]. Emotionally comfortable patients are more likely to cooperate with care, thereby increasing the usability of services and interventions [22]. As per International Organization for Standardization (ISO) standard ISO 9241-11:2018—ergonomics and human-system interaction, usability is defined as: "the extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." A user is defined as an individual who interacts with the device of interest [23].

XSL•FO
RenderX

The Lack of Review Papers Considering User Experiences for mHealth Interventions

Currently, there is a lack of systematic reviews that analyze the extent to which patient experiences have been considered in the design of "usable" mHealth interventions for the detection and management of AF. Reviews to date have mostly focused on technical aspects of devices and apps such as accuracy, diagnostic rate, readability, and quality [24-27]. In other contexts, user-centered approaches entailing different co-design methods to examine the needs, perspectives, and overall experiences of patients have been used for the design of more usable health care interventions [28]. For example, in the case of wearable technology to monitor the activity and mobility of patients with dementia, direct observation, questionnaires, and interviews based on a technology adoption model, that is, the Unified Theory of Acceptance and Use of Technology, have yielded positive outcomes. Patients found the ability of such interventions to provide navigation support useful for going out of their homes without the fear of getting lost. They also felt that its benefits had a carryover effect on other aspects of their life including personal autonomy and relationships [29,30]. The discourse in the *JMIR Human Factors* has also focused on reviewing and consolidating the evidence on the usability of mHealth interventions for different chronic conditions [31-33]. Hence, the critical need for a systematic review to ascertain the extent to which user-centered approaches have been embedded in the design of mHealth interventions for AF.

Contextual Interactions Between AF mHealth Interventions and User Experiences of the Wider Health System

When designing mHealth interventions for AF, contextual interactions with the wider health care system [34] need also to be considered as they can further shape experiences and perspectives. This may include the expertise required by those receiving and delivering the intervention [34]; the people involved, for example, patients, carers, and health care professionals; as well as influences through technology, facilities, or regulations [35]. Health care professionals would need to work closely with mHealth devices and apps, and the introduction of new technology impacts clinical workflows, for example, by easing or increasing daily workload [36] and as such impacting human behavior.

Behavior is closely related to user experiences; it can drive as well as be the target for mHealth interventions. Two examples from the studies reviewed are highlighted: in the first case, a user's experience can influence their behavior with regard to technology acceptance and the uptake of interventions. For example, a user's concern about a device and its data causing them unnecessary worry and anxiousness can lead to its rejection [37], and varying levels of adherence to self-monitoring interventions have been reported in the past [38,39]. Preferences regarding the use of mHealth can also vary among different user groups [40]. In the second case, interventions aim for a behavioral or lifestyle modification, where healthier habits are sought to minimize risk factors associated with the condition, and thus, enhance users' experiences in life [41,42]. While studies to date have explored the incorporation of behavior change techniques [43] in wearable devices and mHealth more widely, the focus has not been on AF in particular [44-46]. Furthermore, behavioral interventions can be classified based on the level of the sociotechnical health system they operate within [47]. At the micro level, interventions focus on the individual and their characteristics including knowledge, attitudes, beliefs, or personality traits. Meso- and macro-level approaches are at the level of groups and whole societies, respectively [46]. Long-term behavioral change approaches, which would be necessary given the chronic nature of AF, also need to account for the following: first, the practical limitations of conducting long-term validation studies, and second, as discussed above, the need to address the wider context and complexity of interactions within health care systems [43]. Finally, a certain level of digital literacy and expertise would be required to use such wearable devices and apps [26]. Therefore, AF mHealth interventions can be regarded as complex.

The New Framework for Developing and Evaluating Complex Interventions

Following the new framework commissioned by the National Institute of Health Research and the UK Medical Research Council (MRC) 2021, the design of complex interventions unfolds in the following phases: identification, development, feasibility, evaluation, and implementation [34]. In the identification and development phases, an existing intervention may be adapted for a new context, or it may be developed from scratch. The feasibility testing phase analyses whether the intervention can meet certain intended outcomes. The evaluation phase goes beyond just assessing whether the intervention works and ascertains what other impacts it has and how it interacts in the context in which it is implemented. In the implementation phase, the results of the previous phases are collated to determine steps to facilitate real-world adoption of the intervention in different health care systems. A wide range of other health care interventions ranging from new surgical procedures to redesigning entire health care programs have been successfully realized as complex interventions. For example, the Links Worker program in Glasgow, Scotland, sought to link people in primary care with community resources targeted individual (micro) and groups and communities (meso) levels and assessed multiple measures including bereavement, substance use, employment, and learning difficulties [34].

The Need to Consider User Experiences and Behavioral Approaches in Designing Complex mHealth Interventions

Previous studies entailing mHealth technology for other apps besides AF have failed to address the potential complexity of health system interactions, and thus, have struggled to transition from pilot studies to interventions actively used in health care systems. The reasons attributed range from various issues including behavioral resistance from patients and health care professionals, poor digital literacy surrounding mHealth use, and a lack of interoperability between devices and systems [36]. On the whole, interventions need to be designed to help users reach their respective goals in their journeys, that is, especially social empowerment for patients [9] and workflow integrability

XSL•FO
RenderX

for health care professionals [36]. As such, in this review study, we aimed to ascertain the extent to which studies to date have considered user experiences and related behavioral approaches in different phases of the design of usable complex AF mHealth interventions.

Methods

The systematic review was conducted as per the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [48].

Data Sources and Search Strategy

Overview

The following databases were searched: PubMed, Cochrane, Scopus, Web of Science, and Embase Ovid. The search was not restricted to any particular time period. Only papers in English were considered. The search string used the terms and logic shown in Textbox 1.

Textbox 1. Search strings.

- For PubMed, Cochrane, and Web of Science
  - (usefulness OR usability OR utility OR use OR “user experience” OR “user-experience”) AND (monitoring OR “remote care” OR detection OR management OR treat* OR screening OR assess* OR diagnosis OR “lifestyle modification” OR “lifestyle change” OR “integrated care” OR behavior* OR behaviour* OR “weight loss” OR “risk factor”) AND (wearable OR “emerging technology” OR “health technology” OR “health device” OR “mobile health” OR mHealth OR “handheld ECG” OR smartwatch OR band OR “ECG patch” OR “implantable loop recorder” OR PPG OR smartphone OR phone OR accelerometer OR gyroscope) AND (“atrial fibrillation” OR “atrial arrhythmia”).
- For Scopus
  - (TITLE-ABS-KEY (usefulness OR usability OR utility OR “use” OR “user experience” OR user-experience) AND TITLE-ABS-KEY (monitoring OR “remote care” OR detection OR management OR treat* OR screening OR assess* OR diagnosis OR “lifestyle modification” OR “lifestyle change” OR “integrated care” OR behavior* OR behaviour* OR “weight loss” OR “risk factor”) AND TITLE-ABS-KEY (wearable OR “health technology” OR “emerging technology” OR “health device” OR “mobile health” OR mHealth OR “handheld ECG” OR smartwatch OR band OR “ECG patch” OR “implantable loop recorder” OR ppg OR smartphone OR phone OR accelerometer OR gyroscope) AND TITLE-ABS-KEY (“atrial fibrillation” OR “atrial arrhythmia”)).
- For Embase Ovid
  - usefulness OR usability OR utility OR (user experience) OR user-experience AND (remote care) OR detection OR management OR treat* OR screening OR assess* OR monitoring OR diagnosis OR (lifestyle modification) OR (lifestyle change) OR (integrated care) OR behavior* OR behaviour* OR (weight loss) OR (risk factor) AND (health technology) OR (emerging technology) OR (health device) OR (mobile health) OR mHealth OR (handheld ECG) OR (ECG patch) OR (implantable loop recorder) OR smartwatch OR band OR PPG OR wearable OR smartphone OR phone OR accelerometer OR gyroscope AND (atrial fibrillation) OR (atrial arrhythmia).

Study Selection and Extraction

Studies that were published before September 7, 2022, and met the criteria described in Textbox 2 were selected. The abstracts of the records were reviewed in the first round of literature screening. The aims, methodology, and results of selected papers were assessed after extracting the full-text records.

Textbox 2. Study inclusion and exclusion criteria.

Inclusion criteria

- At least 1 group of patients with atrial fibrillation (AF) received a device, app, or other mobile health (mHealth) intervention
- Studies must consider the usability of the mobile intervention through user experiences, behavior, or some other criteria in their study outcomes in at least 1 phase of the complex intervention design process
- The abstracts of the records were reviewed in the first round of literature screening. The aims, methodology, and results of selected papers were assessed after extracting the full-text records
- Papers published in English

Exclusion criteria

- Did not entail AF or an mHealth intervention
- Did not consider the usability of the intervention in any of the design stages
- Could not access full-text records
- Papers in other languages

XSL•FO
RenderX

Results

Overview

Among the 2219 records identified, with only 55 records met the full inclusion criteria. The results of the systematic literature review are shown in Figure 1.

The interventions in the studies selected based on their application can be classified into 3 types of interventions: detection, screening, and management. The studies focused on the experiences and behavior of users with respect to the focus areas described in Table 1. Most of the studies that focused on the detection of AF using mHealth devices did not specify the type of AF. Biersteker et al [39] proposed an intervention to detect AF postcardiac surgery while Nguyen et al [49] proposed a smartphone-based device to detect pediatric AF and other arrhythmias. A variety of devices, including smartwatches and smartphones, that incorporated ECG, PPG, and other sensors were used to continuously support patients in detecting AF. Some studies aimed at screening, that is, the use of mHealth for the mass diagnosis of AF among large groups of people for different environments, age groups, and risk factors. Orchard et al [50] studied screening in a general practice setting [51-53] and a routine vaccination setting. Macniven et al [54] assessed screening among Aboriginal communities in Australia. Apps for managing AF focused on improving the knowledge of patients, medication adherence, clinical interactions, clinical decision support, and other applications are listed in Table 1. As shown in Table 2, most of the studies (43/55, 78%) considered the user in the feasibility phase through customized or validated usability questionnaires. Some studies have considered multiple phases and types of interventions [34].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram: systematic review procedure and the resulting number of relevant papers.

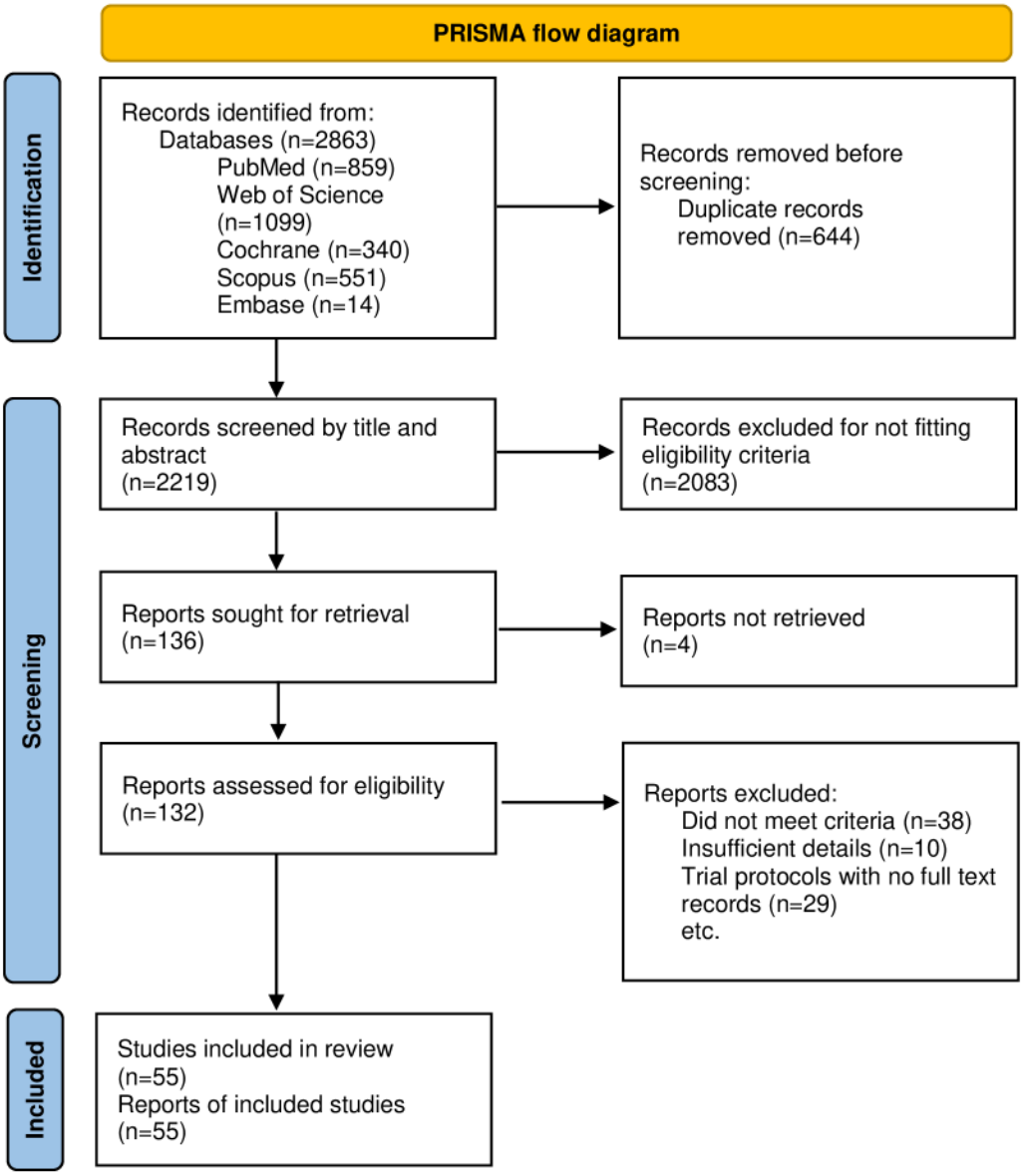

XSL•FO
RenderX

Table 1. The different types of interventions in the selected studies and their focus areas for assessing user experiences and behavioral approaches.

Types of interventions and focus areas	Specific focus areas and examples
Detection	
Device type	• Smartphones [14,38,49,55-59] • Smartwatches and Bands [56,60-63] • Handheld recorders [39,63-67] • Implantable monitors [68,69]
AF[a] type	• Postoperative [58] • Pediatric [49]
Screening	
Environments	• Pharmacy [70] • Aboriginal communities [54,71] • Metropolitan care [52,53] • Rural care [51] • Influenza vaccinations [50]
Patient characteristics	• Age [72-74] • Other risk factors [72]
Management	
Application	• Patient education [13,14,17-19,57,63,64,66,75-81] • Health worker education [82,83] • Clinical decision support [84] • Medication adherence [17,81,85,86] • Therapy adherence [19,87] • Online support group [86] • Sleep apnea management [88]

[a]AF: atrial fibrillation.

Table 2. The number of studies considering user experiences in screening, detection, and management of AF[a] across the phases of complex interventions [34].

	Identification, n	Development, n	Feasibility, n	Evaluation, n	Implementation, n
Screening	2	1	7	1	3
Detection	5	1	23	0	0
Management	4	3	21	0	0

[a]AF: atrial fibrillation.

User Experience and Behavioral Approaches in Phases of Designing Complex Interventions

The various methods used by the selected papers for each phase of the UK MRC framework for the design of interventions [34] are shown in Table 3 and described below.

XSL•FO
RenderX

Table 3. The methods used to understand user experiences and behavioral approaches in the different phases of developing and evaluating complex interventions.

Phase	Methods
Identification	• Survey [72,73,79,87,89,90] • Interviews [65,75,86,87,91]
Development	• Discussions [52,61,87] • Focus groups [61,81]
Feasibility	• Survey [17-19,38,39,49,54-57,59-64,66,67,74,76,77,80,82,84,85,87,88,92-96] • Adherence [14,18,19,38,39,52,55,56,61,64,66,68,69,74,81,84] • Interview [13,18,19,50,52,54,58,62,67,68,71,74] • Observation [39,52,58,67]
Evaluation	• Interviews [52,53] • Observation [52]
Implementation	• Interviews [51,70] • Adherence [51,53,70]

Identification Phase

Studies elicited in the review aimed to understand the attitudes of stakeholders toward mHealth interventions for AF, including the perceived advantages and disadvantages, and case scenarios using surveys and interviews [73,75]. While most studies considered micro level behavior, Waring et al [86] explored the interests of older patients with AF in meso level online community support groups. Most of the surveys and questionnaires were not validated, custom-made, and not designed together with users and stakeholders. Although Ding et al [91] and Nuvvula et al [89] reported that they prepared their survey iteratively by consulting a review panel or "experts," no further elaboration was provided. Reading et al [65] based their interview questions on the Unified Theory of Acceptance and Use of Technology model, where engaged and unengaged users were identified based on device use data. Patients showed an active interest in all interventions barring online support groups, which were more likely to be used by those who were already using web-based tools [86]. Participants with diagnosed AF were more likely to share their data with providers compared to those with undiagnosed AF [89]. Cher et al [75] reported discord between the interests of physicians and patients: physicians wanted a tool that could support their workflow while patients valued tracking and logging their health data. Boriani et al [72] reported that 70% of physician participants in their study were not willing to initiate mass screening using wearables in their current state, and in over half of the cases, the use of wearables was the patients' personal decision. In Ding et al [91], over half of the respondents did not use wearables due to a lack of clinical and other guidelines on the same. In Manninger et al [73], most physicians believed that the costs of wearables should be shared between patients and insurers. Electrophysiologists were more likely to recommend wearable devices compared to general physicians [72,91]. Physicians preferred handheld, single-lead ECG devices over PPG and other devices [72,73]. On the whole, there can be varying and conflicting preferences among users and stakeholders, which need to be considered in identifying suitable mHealth technology in the design of interventions.

Development Phase

Four studies considered a user-centered approach to developing interventions. Dickson et al [61] and Toscos et al [81] proposed focus groups to ascertain user needs and gaps. In Dickson et al [61], selected patients and caregivers along with developers were also invited to a hackathon to improve the usability and interactivity of the smartwatch intervention for detecting AF. Orchard et al [52] developed their initial screening program theory based on previous discussions with doctors, nurses, and researchers. Peleg et al [87] iteratively designed the graphical user interface of their mobile apps by consulting academic and industrial experts and group members, although they did not elaborate on their methods. They used the transtheoretical model to stratify users based on their existing behavior into precontemplation, preparation, and action stages based on their readiness to change with respect to therapy compliance. Patients in the precontemplation stage received encouraging messages once they started using the behavior change app, acknowledging their willingness to get started with therapy. In Toscos et al [81], patients felt that they received inadequate information and support from doctors and health care providers. As a result, Toscos et al [81] developed an app that provided up-to-date information in simple language to patients by both electronic and print means. In Orchard et al [52], health care workers were provided with integrated electronic tools that can automatically identify eligible patients and financial incentives along with the ECG devices to overcome time constraints and the lack of reimbursement policies.

Feasibility Phase

The majority of the studies (43/55, 78%) considered user experiences and behavior of the interventions in the feasibility stage as a part of the progression criteria through questionnaires, interviews, observation, and adherence to interventions. Questionnaires and surveys typically considered ease of use, attractiveness, novelty, and concerns [17,56,85]. Among questionnaires that assessed user satisfaction and the usability of interventions, some studies have used standardized scales, such as the System Usability Scale [61,62] and the Mobile

XSL•FO
RenderX

Application Rating Scale [92], while others have used custom ones with similar Likert scale questions [64]. Adherence to an app or a device was measured as the number of days it was used or the number of signals that were transmitted. Nguyen et al [49] used repeated surveys to track the usability of their intervention across a period. Only one study [93] assessed the impact of an intervention on users' activities of daily living such as moving, sleeping, eating, and using the restroom.

However, some studies asked leading questions in their surveys. For example, in Mcmanus et al [59], some of the survey questions included "The app gives me reassurance" and "This app will improve my general well-being." Qualitative interviews and observation studies were necessary to obtain an in-depth understanding of user perspectives and experiences that surveys and log data with predetermined outcome measures could not do. In the study by Ding et al [62] that assessed the use of a smartwatch for AF monitoring after stroke, patients expressed the need for both continuous monitoring and in the form of a passive system that did not require extensive engagement. In Peleg et al [87], which involved an app to increase patient compliance, interviews confirmed the consistency between participants' and authors' perceptions of "patient engagement." Macniven et al [54] and Ding et al [91] stated that they co-designed interviews and outcome measures together with stakeholders, although they did not provide a detailed description of the same. In Gawalko et al [57] and Nathania et al [18], reminders and positive feedback on the use of heart monitoring devices were used to increase the usability of these interventions. Nathania et al [18] based their surveys and interviews on the Technology Acceptance Model and examined the perceived usefulness, perceived ease of use, and behavioral intent with regard to the self-management app. Participants found the ability of the app to improve AF knowledge, management, and clinical consultations to be useful and felt that having it on their personal devices would improve the ease of use. Most participants were willing to continue using it, although actual use data showed skewed interest for different features, that is, preference for AF-related content and decreased interest for motivational messages and reminders over time. Wong et al [67] used observation studies, where the time taken for physicians to place the electrodes of a handheld ECG device and obtain the signal was considered as a study outcome. Only Desteghe et al [17] considered outcomes at the meso community level in their study on grandparent-grandchildren relationships. A gamification strategy was used to develop a medication reminder mobile app, where the grandparents, who were the patients, and their grandchildren were given rewards on the completion of certain tasks. The grandparents had to take their medicine while the children chose to perform a health-promoting activity such as eating fruits. The rewards were fun activities that both could perform together such as going on a trip.

Evaluation Phase

There was only one study that focused on this stage of mobile AF health interventions. The paper by Orchard et al [52] on screening used the realist evaluation framework that considers the context (C), mechanism (M), and outcome (O) of interventions. Outcomes are effects resulting from the conditions created by mechanisms operating in a context. Contexts included the general practice regulations; mechanisms that lead to success were practice wide engagement and user acceptance; and outcomes related to motivations, behavior change, and barriers to screening. Along with semistructured interviews, they observed health care workers during their practice and noted the different barriers to screening, attitudes to screening, and motivation to screen in rural and metropolitan general practice community settings. Efficient leadership, regular updates, and defined protocols were found to be crucial for the success of screening interventions.

Implementation Phase

Studies proposed semistructured interviews and adherence to intervention implementation in understanding the acceptability, barriers, and facilitators of interventions in different community contexts [51,70]. In Lowres et al [70], which examined the uptake of a public AF screening program in pharmacies, increasing the customer's confidence regarding the ability of the pharmacist ability to use the device was crucial for their participation. The unavailability of Wi-Fi networks and problems with mobile phone reception also challenged customer recruitment. A combination of advertising using flyers and posters and directly approaching customers was necessary to maximize recruitment. Furthermore, pharmacists reported conflicts between their regular work, that is, dispensing drugs and screening especially during busier times. Combining AF screening with other screening services, such as blood pressure and sugar monitoring, resulted in more time-efficient workflows.

Discussion

Principal Findings

While at the outset, it was hoped to ascertain quantitative evidence for the effectiveness of different types of mHealth interventions and their impact on patient's lives, experiences, and patient outcomes; instead, the information presented in the papers reviewed lent itself qualitatively to assess the need and effectiveness of the methods used to elicit user experiences and device usability. The different methods used to study user experiences and behavior with AF mHealth interventions included interviews, surveys, questionnaires, device data, focus groups, hackathons, and observation. Some were custom-made while others were prevalidated, such as the System Usability Scale and Mobile Application Rating Scale, or based on preexisting models, such as the Technology Acceptance Model. In the identification phase, studies showed that there can be considerable differences between the needs of patients and health care workers. In development, user perspectives guided the development of the device and apps, and in providing technical and nontechnical support to ensure the intervention's success. Although several papers considered usability as an outcome in feasibility studies, the interviews, surveys, and other instruments were not designed together with direct users and other stakeholders. Studies using interviews provided a richer understanding of user experiences and perspectives. The evaluation and implementation phase studies entailed unique real-world challenges associated with user participation, acceptance, and workflows that could not be captured by studies in the previous phases.

XSL•FO
RenderX

There was only a single study that specifically mentioned the use of behavioral change approaches in the design of interventions [87]. While the rest of the papers entailed approaches that could be categorized and mapped to different levels of sociotechnical health care systems, the studies did not regard the approaches as behavior-driven or behavioral change approaches in their study objectives. Most of the studies focused on user behavior at the micro level, with none reporting behavior change at the macro level. Furthermore, the long-term applicability of the used methods is unclear. Feedback-based [18,57,87] and goal-based approaches [17,87] have been found to be applicable in the long term in other contexts, although 3 of the selected studies did not entail a variation over time across the respective study durations of 28 weeks [57], 6 weeks [18], and 12 weeks [17]. These approaches require a temporal variation to ensure their long-term applicability, as across time and changing contexts, the same kind of feedback may fail to elicit new behavior, and previous goals may no longer seem beneficial or achievable [43]. In this regard, Peleg et al [87] moved patients between the stages of the transtheoretical model of health behavior change based on their engagement with the app and completion of tasks, even though the authors did not specify the exact duration of their preliminary assessment study.

This systematic review linked studies to different phases of the MRC framework for the design of complex interventions, which was missing in reviews to date. It elucidated the various methods used to understand and influence user experiences and their strengths and drawbacks. It highlighted the lack of emphasis on behavior change and behavior-driven methods, as well as their limited long-term applicability. The results showed that, in most of the studies and especially in the feasibility phase (n=43), the measures for assessing user experiences were based more on authors' own perceptions of an intervention's success and usability and less on users' journeys, goals, and lived experience in health care systems. Going forward, studies aimed at understanding collective behavioral trends at the societal or macro level, which have been unexplored by studies to date, would need to account for political and socioeconomic health system factors such as public approval and government policies [47].

Limitations

This review had some limitations. Only studies in English were considered. The statistical validity of the study outcomes and the existence of biases were not analyzed due to the lack of studies. In addition, some studies were incomplete and only their methodology could be assessed.

Conclusions

This systematic review proposed the novelties of mapping, first, studies of different phases of the MRC framework for the design of complex interventions, and second, behavioral approaches to different levels of health care systems. All in all, most studies to date have not attempted to obtain a deeper understanding of user experiences and related behavior. The challenges associated with introducing mHealth interventions in sociotechnical health care systems, which impact the uptake and usability of mHealth interventions [36,97] and impede the health care system's ability to help users achieve their goals in life [29,30], have not been adequately addressed. The results of this review emphasized the need for a change in the direction of AF mHealth research to focus more on user experiences and behavior. Such a change would be possible through co-designing through life with patient and public involvement and using systems approaches that consider people with complex health needs, for example in the context of multiple long-term conditions. Our analysis of the different works to date with respect to user experiences and behavior serves as a starting point for future research using mHealth interventions to improve the lives and journeys of patients with AF and their wider network of formal and informal carers.

Acknowledgments

This work was partly funded by a Global Research Scholarship from the University of Strathclyde and the Department of Design, Manufacturing, and Engineering Management awarded to the first author. The authors have also been supported by a Systems Engineering Innovation Hubs for Multiple Long-Term Conditions (SEISMIC) development phase grant jointly funded by the UK Research and Innovation (UKRI) Engineering and Physical Science Research Council (EPSRC) and the National Institute for Health and Care Research (NIHR) under grant NIHR157433. The views expressed in this publication are those of the authors and not necessarily those of the UKRI-EPSRC, NIHR, or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 104 KB - jmir_v26i1e54405_app1.pdf]

References

1. Lippi G, Sanchis-Gomar F, Cervellin G. Global epidemiology of atrial fibrillation: an increasing epidemic and public health challenge. Int J Stroke 2021;16(2):217-221. [doi: 10.1177/1747493019897870] [Medline: 31955707]

XSL•FO
RenderX

2. Tripp-Reimer T, Williams JK, Gardner SE, Rakel B, Herr K, McCarthy AM, et al. An integrated model of multimorbidity and symptom science. Nurs Outlook 2020;68(4):430-439 [FREE Full text] [doi: 10.1016/j.outlook.2020.03.003] [Medline: 32482344]
3. Morseth B, Geelhoed B, Linneberg A, Johansson L, Kuulasmaa K, Salomaa V, et al. Age-specific atrial fibrillation incidence, attributable risk factors and risk of stroke and mortality: results from the MORGAM consortium. Open Heart 2021;8(2):e001624 [FREE Full text] [doi: 10.1136/openhrt-2021-001624] [Medline: 34341095]
4. Frederiksen TC, Dahm CC, Preis SR, Lin H, Trinquart L, Benjamin EJ, et al. The bidirectional association between atrial fibrillation and myocardial infarction. Nat Rev Cardiol 2023;20(9):631-644. [doi: 10.1038/s41569-023-00857-3] [Medline: 37069297]
5. Proietti M, Farcomeni A, Goethals P, Scavee C, Vijgen J, Blankoff I, et al. Cost-effectiveness and screening performance of ECG handheld machine in a population screening programme: the Belgian heart rhythm week screening programme. Eur J Prev Cardiol 2019;26(9):964-972. [doi: 10.1177/2047487319839184] [Medline: 30935219]
6. Friedman DJ, Field ME, Rahman M, Goldstein L, Sha Q, Sidharth M, et al. Catheter ablation and healthcare utilization and cost among patients with paroxysmal versus persistent atrial fibrillation. Heart Rhythm O2 2021;2(1):28-36 [FREE Full text] [doi: 10.1016/j.hroo.2020.12.017] [Medline: 34113902]
7. Ballatore A, Matta M, Saglietto A, Desalvo P, Bocchino PP, Gaita F, et al. Subclinical and asymptomatic atrial fibrillation: current evidence and unsolved questions in clinical practice. Medicina (Kaunas) 2019;55(8):497 [FREE Full text] [doi: 10.3390/medicina55080497] [Medline: 31426580]
8. Perry M, Betty SK, Downes N, Andrews N, Mackenzie S, Guideline Committee. Atrial fibrillation: diagnosis and management-summary of NICE guidance. BMJ 2021;373:n1150. [doi: 10.1136/bmj.n1150] [Medline: 34020968]
9. Velleca M, Costa G, Goldstein L, Bishara M, Ming Boo L. A review of the burden of atrial fibrillation: understanding the impact of the new millennium epidemic across Europe. EMJ Cardiol 2019 Oct 17:110-118. [doi: 10.33590/emjcardiol/10310292]
10. Burdett P, Lip GYH. Atrial fibrillation in the UK: predicting costs of an emerging epidemic recognizing and forecasting the cost drivers of atrial fibrillation-related costs. Eur Heart J Qual Care Clin Outcomes 2022;8(2):187-194. [doi: 10.1093/ehjqcco/qcaa093] [Medline: 33346822]
11. Tipping point : why heart care must be prioritised now. British-Heart-Foundation. 2022. URL: https://www.bhf.org.uk/-/media/files/what-we-do/influencing-change/tipping-point-bhf-report.pdf?rev=08914752a2c4dd789746b5e6f4ba617&hash=F7141D4C13A10C6C089824892CA59CDE [accessed 2024-08-15]
12. Tromp J, Jindal D, Redfern J, Bhatt A, Séverin T, Banerjee A, et al. World Heart Federation roadmap for digital health in cardiology. Glob Heart 2022;17(1):61 [FREE Full text] [doi: 10.5334/gh.1141] [Medline: 36051317]
13. Parimbelli E, Sacchi L, Budasu R, Napolitano C, Peleg M, Quaglini S. The role of nurses in e-health: the MobiGuide project experience. Stud Health Technol Inform 2016;225:153-157. [Medline: 27332181]
14. Guo Y, Guo J, Shi X, Yao Y, Sun Y, Xia Y, et al. Mobile health technology-supported atrial fibrillation screening and integrated care: a report from the mAFA-II trial long-term extension cohort. Eur J Intern Med 2020;82:105-111 [FREE Full text] [doi: 10.1016/j.ejim.2020.09.024] [Medline: 33067121]
15. Chung MK, Eckhardt LL, Chen LY, Ahmed HM, Gopinathannair R, Joglar JA, et al. Lifestyle and risk factor modification for reduction of atrial fibrillation: a scientific statement from the American Heart Association. Circulation 2020;141(16):e750-e772 [FREE Full text] [doi: 10.1161/CIR.0000000000000748] [Medline: 32148086]
16. Zhao H, Li X, Yu P, Liu M, Ma J, Wang J, et al. Association between weight loss and outcomes in patients undergoing atrial fibrillation ablation: a systematic review and dose-response meta-analysis. Nutr Metab (Lond) 2023;20(1):5 [FREE Full text] [doi: 10.1186/s12986-023-00724-5] [Medline: 36721216]
17. Desteghe L, Kluts K, Vijgen J, Koopman P, Dilling-Boer D, Schurmans J, et al. The health buddies app as a novel tool to improve adherence and knowledge in atrial fibrillation patients: a pilot study. JMIR Mhealth Uhealth 2017;5(7):e98 [FREE Full text] [doi: 10.2196/mhealth.7420] [Medline: 28724512]
18. Nathania J, Woo BFY, Cher BP, Toh KY, Chia WA, Lim YW, et al. Patient perspectives of the Self-Management and Educational Technology Tool for Atrial Fibrillation (SETAF): a mixed-methods study in Singapore. PLoS One 2022;17(1):e0262033 [FREE Full text] [doi: 10.1371/journal.pone.0262033] [Medline: 35061749]
19. Knaepen L, Delesie M, Theunis R, Vijgen J, Dendale P, Desteghe L, et al. A new smartphone application for integrated transmural care of atrial fibrillation, AF-EduApp: usability and validation study. Digit Health 2021;7:20552076211067105 [FREE Full text] [doi: 10.1177/20552076211067105] [Medline: 34992790]
20. Ruff CT. The promise of mobile health in managing atrial fibrillation. J Am Coll Cardiol 2020;75(13):1535-1537 [FREE Full text] [doi: 10.1016/j.jacc.2020.02.039] [Medline: 32241368]
21. Pal K, Dack C, Ross J, Michie S, May C, Stevenson F, et al. Digital health interventions for adults with type 2 diabetes: qualitative study of patient perspectives on diabetes self-management education and support. J Med Internet Res 2018;20(2):e40 [FREE Full text] [doi: 10.2196/jmir.8439] [Medline: 29463488]
22. Williams AM, Lester L, Bulsara C, Petterson A, Bennett K, Allen E, et al. Patient evaluation of emotional comfort experienced (PEECE): developing and testing a measurement instrument. BMJ Open 2017;7(1):e012999 [FREE Full text] [doi: 10.1136/bmjopen-2016-012999] [Medline: 28122833]

XSL•FO
RenderX

23. ISO 9241-11:2018 ergonomics of human-system interaction part 11: usability: definitions and concepts. Organisation(ISO) IS. 2018. URL: https://www.iso.org/standard/63500.html [accessed 2018-11-01]
24. Gill S, Bunting KV, Sartini C, Cardoso VR, Ghoreishi N, Uh HW, et al. Smartphone detection of atrial fibrillation using photoplethysmography: a systematic review and meta-analysis. Heart 2022;108(20):1600-1607 [FREE Full text] [doi: 10.1136/heartjnl-2021-320417] [Medline: 35277454]
25. Belani S, Wahood W, Hardigan P, Placzek AN, Ely S. Accuracy of detecting atrial fibrillation: a systematic review and meta-analysis of wrist-worn wearable technology. Cureus 2021;13(12):e20362 [FREE Full text] [doi: 10.7759/cureus.20362] [Medline: 35036196]
26. Ayyaswami V, Padmanabhan DL, Crihalmeanu T, Thelmo F, Prabhu AV, Magnani JW. Mobile health applications for atrial fibrillation: a readability and quality assessment. Int J Cardiol 2019;293:288-293 [FREE Full text] [doi: 10.1016/j.ijcard.2019.07.026] [Medline: 31327518]
27. Pearsons A, Hanson CL, Gallagher R, O'Carroll RE, Khonsari S, Hanley J, et al. Atrial fibrillation self-management: a mobile telephone app scoping review and content analysis. Eur J Cardiovasc Nurs 2021;20(4):305-314 [FREE Full text] [doi: 10.1093/eurjcn/zvaa014] [Medline: 33620473]
28. Preece J, Rogers Y, Sharp H. Interaction Design: Beyond Human-Computer Interaction. New York: John Wiley & Sons; 2002.
29. Thorpe JR, Rønn-Andersen KVH, Bień P, Özkil AG, Forchhammer BH, Maier AM. Pervasive assistive technology for people with dementia: a UCD case. Healthc Technol Lett 2016;3(4):297-302 [FREE Full text] [doi: 10.1049/htl.2016.0057] [Medline: 28008366]
30. Thorpe JR, Forchhammer BH, Maier AM. Development of a sensor-based behavioral monitoring solution to support dementia care. JMIR Mhealth Uhealth 2019;7(6):e12013 [FREE Full text] [doi: 10.2196/12013] [Medline: 31199304]
31. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
32. Kane H, Baumgart JG, El-Hage W, Deloyer J, Maes C, Lebas MC, et al. Opportunities and challenges for professionals in psychiatry and mental health care using digital technologies during the COVID-19 pandemic: systematic review. JMIR Hum Factors 2022;9(1):e30359 [FREE Full text] [doi: 10.2196/30359] [Medline: 34736224]
33. Waite M, Martin C, Franklin R, Duce D, Harrison R. Human factors and data logging processes with the use of advanced technology for adults with type 1 diabetes: systematic integrative review. JMIR Hum Factors 2018;5(1):e11 [FREE Full text] [doi: 10.2196/humanfactors.9049] [Medline: 29535079]
34. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of medical research council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
35. Grosios K, Gahan PB, Burbidge J. Overview of healthcare in the UK. EPMA J 2010;1(4):529-534 [FREE Full text] [doi: 10.1007/s13167-010-0050-1] [Medline: 23199107]
36. Porterfield C, Stitt K. Struggling to connect: 2021 state of remote cardiac monitoring. Vector Remote Care. 2021. URL: https://vectorremote.com/white-paper/2021-state-of-remote-cardiac-monitoring-report/ [accessed 2024-08-29]
37. Kumar SS, Andreakou E, Tzachsan M, Maier AM. Mapping the journeys of atrial fibrillation patients and citizens using wearable devices for remote cardiac monitoring. Proc Des Soc 2023;3:2575-2584. [doi: 10.1017/pds.2023.258]
38. Sardana M, Saczynski J, Esa N, Floyd K, Chon K, Chong JW, et al. Performance and usability of a novel smartphone application for atrial fibrillation detection in an ambulatory population referred for cardiac monitoring. J Am Coll Cardiol 2016;67(13):844. [doi: 10.1016/s0735-1097(16)30845-2]
39. Biersteker TE, Boogers MJ, Schalij MJ, de Vries BBLP, Groenwold RHH, van Alem AP, et al. Mobile health vs. standard care after cardiac surgery: results of The box 2.0 study. Europace 2023;25(1):49-58 [FREE Full text] [doi: 10.1093/europace/euac115] [Medline: 35951658]
40. van der Ven J, van den Bemt BJF, van Dijk L, Opdam M, Haegens LL, Vriezekolk JE, et al. Preferences of patients with musculoskeletal disorders regarding the timing and channel of eHealth and factors influencing its use: mixed methods study. JMIR Hum Factors 2023;10:e44885 [FREE Full text] [doi: 10.2196/44885] [Medline: 37756049]
41. Buss VH, Varnfield M, Harris M, Barr M. A mobile app for prevention of cardiovascular disease and type 2 diabetes mellitus: development and usability study. JMIR Hum Factors 2022;9(2):e35065 [FREE Full text] [doi: 10.2196/35065] [Medline: 35536603]
42. Chiang N, Guo M, Amico KR, Atkins L, Lester RT. Interactive two-way mHealth interventions for improving medication adherence: an evaluation using the behaviour change wheel framework. JMIR Mhealth Uhealth 2018;6(4):e87 [FREE Full text] [doi: 10.2196/mhealth.9187] [Medline: 29650504]
43. Cash P, Wrobel A, Maier A, Hansen JP. Understanding long-term behaviour change techniques: a mixed methods study. J Eng Des 2023;34(5-6):383-410. [doi: 10.1080/09544828.2023.2227933]
44. Mercer K, Li M, Giangregorio L, Burns C, Grindrod K. Behavior change techniques present in wearable activity trackers: a critical analysis. JMIR Mhealth Uhealth 2016;4(2):e40 [FREE Full text] [doi: 10.2196/mhealth.4461] [Medline: 27122452]
45. Chia GLC, Anderson A, McLean LA. Behavior change techniques incorporated in fitness trackers: content analysis. JMIR Mhealth Uhealth 2019;7(7):e12768 [FREE Full text] [doi: 10.2196/12768] [Medline: 31339101]

XSL•FO
RenderX

46. Duncan M, Murawski B, Short CE, Rebar AL, Schoeppe S, Alley S, et al. Activity trackers implement different behavior change techniques for activity, sleep, and sedentary behaviors. Interact J Med Res 2017;6(2):e13 [FREE Full text] [doi: 10.2196/ijmr.6685] [Medline: 28807889]
47. Maier A, Cash P. Designing for human behaviour in a systemic world. In: Maier A, Oehmen J, Vermaas PE, editors. Handbook of Engineering Systems Design. Cham: Springer International Publishing; 2022:493-526.
48. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
49. Nguyen HH, van Hare GF, Rudokas M, Bowman T, Silva JNA. SPEAR trial: smartphone pediatric ElectrocARdiogram trial. PLoS One 2015;10(8):e0136256 [FREE Full text] [doi: 10.1371/journal.pone.0136256] [Medline: 26295569]
50. Orchard J, Lowres N, Freedman SB, Ladak L, Lee W, Zwar N, et al. Screening for atrial fibrillation during influenza vaccinations by primary care nurses using a smartphone electrocardiograph (iECG): a feasibility study. Eur J Prev Cardiol 2016;23(2 suppl):13-20. [doi: 10.1177/2047487316670255] [Medline: 27892421]
51. Orchard JJ, Neubeck L, Freedman B, Webster R, Patel A, Gallagher R, et al. Atrial fibrillation screen, management and guideline recommended therapy (AF SMART II) in the rural primary care setting: an implementation study protocol. BMJ Open 2018;8(10):e023130 [FREE Full text] [doi: 10.1136/bmjopen-2018-023130] [Medline: 30385444]
52. Orchard J, Li J, Gallagher R, Freedman B, Lowres N, Neubeck L. Uptake of a primary care atrial fibrillation screening program (AF-SMART): a realist evaluation of implementation in metropolitan and rural general practice. BMC Fam Pract 2019;20(1):170 [FREE Full text] [doi: 10.1186/s12875-019-1058-9] [Medline: 31810441]
53. Orchard J, Neubeck L, Freedman B, Li J, Webster R, Zwar N, et al. eHealth tools to provide structured assistance for atrial fibrillation screening, management, and guideline-recommended therapy in metropolitan general practice: The AF-SMART study. J Am Heart Assoc 2019;8(1):e010959 [FREE Full text] [doi: 10.1161/JAHA.118.010959] [Medline: 30590964]
54. Macniven R, Gwynn J, Fujimoto H, Hamilton S, Thompson SC, Taylor K, et al. Feasibility and acceptability of opportunistic screening to detect atrial fibrillation in aboriginal adults. Aust N Z J Public Health 2019;43(4):313-318 [FREE Full text] [doi: 10.1111/1753-6405.12905] [Medline: 31141280]
55. Beerten SG, Proesmans T, Vaes B. A heart rate monitoring app (FibriCheck) for atrial fibrillation in general practice: pilot usability study. JMIR Form Res 2021;5(4):e24461 [FREE Full text] [doi: 10.2196/24461] [Medline: 33825692]
56. Ding E, Liu D, Soni A, Adaramola O, Han D, Bashar SK, et al. Impressions of older patients with cardiovascular diseases to smart devices for heart rhythm monitoring. 2017 Presented at: 2nd IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017; July 17-19, 2017; Philadelphia, PA. [doi: 10.1109/chase.2017.97]
57. Gawałko M, Duncker D, Manninger M, van der Velden RMJ, Hermans ANL, Verhaert DVM, et al. The european TeleCheck-AF project on remote app-based management of atrial fibrillation during the COVID-19 pandemic: centre and patient experiences. Europace 2021;23(7):1003-1015 [FREE Full text] [doi: 10.1093/europace/euab050] [Medline: 33822029]
58. Lowres N, Freedman SB, Gallagher R, Kirkness A, Marshman D, Orchard J, et al. Identifying postoperative atrial fibrillation in cardiac surgical patients posthospital discharge, using iPhone ECG: a study protocol. BMJ Open 2015;5(1):e006849 [FREE Full text] [doi: 10.1136/bmjopen-2014-006849] [Medline: 25586373]
59. McManus MD, Chong JW, Soni A, Saczynski JS, Esa N, Napolitano C, et al. PULSE-SMART: pulse-based arrhythmia discrimination using a novel smartphone application. J Cardiovasc Electrophysiol 2016;27(1):51-57 [FREE Full text] [doi: 10.1111/jce.12842] [Medline: 26391728]
60. Dhruva SS, Shah ND, Vemulapalli S, Deshmukh A, Beatty AL, Gamble GM, et al. Heart watch study: protocol for a pragmatic randomised controlled trial. BMJ Open 2021;11(12):e054550 [FREE Full text] [doi: 10.1136/bmjopen-2021-054550] [Medline: 35234659]
61. Dickson EL, Ding EY, Saczynski JS, Han D, Moonis M, Fitzgibbons TP, et al. Smartwatch monitoring for atrial fibrillation after stroke—the Pulsewatch study: protocol for a multiphase randomized controlled trial. Cardiovasc Digit Health J 2021;2(4):231-241 [FREE Full text] [doi: 10.1016/j.cvdhj.2021.07.002] [Medline: 35265913]
62. Ding EY, CastañedaAvila M, Tran KV, Mehawej J, Filippaios A, Paul T, et al. Usability of a smartwatch for atrial fibrillation detection in older adults after stroke. Cardiovasc Digit Health J 2022;3(3):126-135 [FREE Full text] [doi: 10.1016/j.cvdhj.2022.03.003] [Medline: 35720675]
63. Turchioe MR, Bakken S, Masterson CRM. Feasibility of using mobile health technology to deliver life's simple 7 messages for atrial fibrillation management. Circulation 2019;140(1) [FREE Full text]
64. Magnani JW, Schlusser CL, Kimani E, Rollman BL, Paasche-Orlow MK, Bickmore TW. The atrial fibrillation health literacy information technology system: pilot assessment. JMIR Cardio 2017;1(2):e7 [FREE Full text] [doi: 10.2196/cardio.8543] [Medline: 29473644]
65. Reading M, Baik D, Beauchemin M, Hickey KT, Merrill JA. Factors influencing sustained engagement with ECG self-monitoring: perspectives from patients and health care providers. Appl Clin Inform 2018;9(4):772-781 [FREE Full text] [doi: 10.1055/s-0038-1672138] [Medline: 30304745]

XSL•FO
RenderX

66. Weng W, Blanchard C, Reed JL, Matheson K, McIntyre C, Gray C, et al. A virtual platform to deliver ambulatory care for patients with atrial fibrillation. Cardiovasc Digit Health J 2020;2(1):63-70 [FREE Full text] [doi: 10.1016/j.cvdhj.2020.11.005] [Medline: 35265891]
67. Wong KC, Thiagalingam A, Kumar S, Marschner S, Kunwar R, Bailey J, et al. User perceptions and experiences of a handheld 12-lead electrocardiographic device in a clinical setting: usability evaluation. JMIR Cardio 2021;5(2):e21186 [FREE Full text] [doi: 10.2196/21186] [Medline: 34435958]
68. Mariani JA, Weerasooriya R, van den Brink O, Mohamed U, Gould PA, Pathak RK, et al. Miniaturized implantable cardiac monitor with a long sensing vector (BIOMONITOR III): insertion procedure assessment, sensing performance, and home monitoring transmission success. J Electrocardiol 2020;60:118-125 [FREE Full text] [doi: 10.1016/j.jelectrocard.2020.04.004] [Medline: 32361086]
69. Tilz RR, Shaik N, Piorkowski C, Hu Y, Connolly AT, Reyes IJ, et al. Real-world adoption of smartphone-based remote monitoring using the confirm Rx™ insertable cardiac monitor. J Innov Card Rhythm Manag 2021;12(8):4613-4620 [FREE Full text] [doi: 10.19102/icrm.2021.120806] [Medline: 34386274]
70. Lowres N, Krass I, Neubeck L, Redfern J, McLachlan AJ, Bennett AA, et al. Atrial fibrillation screening in pharmacies using an iPhone ECG: a qualitative review of implementation. Int J Clin Pharm 2015;37(6):1111-1120. [doi: 10.1007/s11096-015-0169-1] [Medline: 26202627]
71. Gwynne K, Flaskas Y, O'Brien C, Jeffries TL, McCowen D, Finlayson H, et al. Opportunistic screening to detect atrial fibrillation in aboriginal adults in Australia. BMJ Open 2016;6(11):e013576 [FREE Full text] [doi: 10.1136/bmjopen-2016-013576] [Medline: 27852724]
72. Boriani G, Schnabel RB, Healey JS, Lopes RD, Verbiest-van Gurp N, Lobban T, et al. Consumer-led screening for atrial fibrillation using consumer-facing wearables, devices and apps: a survey of health care professionals by AF-SCREEN international collaboration. Eur J Intern Med 2020;82:97-104 [FREE Full text] [doi: 10.1016/j.ejim.2020.09.005] [Medline: 32933842]
73. Manninger M, Kosiuk J, Zweiker D, Njeim M, Antolic B, Kircanski B, et al. Role of wearable rhythm recordings in clinical decision making—the wEHRAbles project. Clin Cardiol 2020;43(9):1032-1039 [FREE Full text] [doi: 10.1002/clc.23404] [Medline: 32700414]
74. Wong KC, Nguyen TN, Marschner S, Turnbull S, Burns MJ, Ne JYA, et al. Patient-led mass screening for atrial fibrillation in the older population using handheld electrocardiographic devices integrated with a clinician-coordinated remote central monitoring system: protocol for a randomized controlled trial and process evaluation. JMIR Res Protoc 2022;11(2):e34778 [FREE Full text] [doi: 10.2196/34778] [Medline: 35103614]
75. Cher BP, Kembhavi G, Toh KY, Audimulam J, Chia WYA, Vrijhoef HJ, et al. Understanding the attitudes of clinicians and patients toward a self-management eHealth tool for atrial fibrillation: qualitative study. JMIR Hum Factors 2020;7(3):e15492 [FREE Full text] [doi: 10.2196/15492] [Medline: 32940611]
76. Desteghe L, Germeys J, Vijgen J, Koopman P, Dilling-Boer D, Schurmans J. The impact of an online directed education platform on the knowledge level of atrial fibrillation patients undergoing cardioversion or pulmonary vein isolation. Europace 2018;20(suppl_1):i24. [doi: 10.1093/europace/euy015.059]
77. Guo Y, Chen Y, Lane DA, Liu L, Wang Y, Lip GYH. Mobile health technology for atrial fibrillation management integrating decision support, education, and patient involvement: mAF app trial. Am J Med 2017;130(12):1388-1396.e6 [FREE Full text] [doi: 10.1016/j.amjmed.2017.07.003] [Medline: 28847546]
78. Peleg M, Shahar Y, Quaglini S, Broens T, Budasu R, Fung N, et al. Assessment of a personalized and distributed patient guidance system. Int J Med Inform 2017;101:108-130. [doi: 10.1016/j.ijmedinf.2017.02.010] [Medline: 28347441]
79. Pezel T, Berthelot E, Gauthier J, Chong-Nguyen C, Iliou MC, Juillière Y, et al. Epidemiological characteristics and therapeutic management of patients with chronic heart failure who use smartphones: potential impact of a dedicated smartphone application (report from the OFICSel study). Arch Cardiovasc Dis 2021;114(1):51-58 [FREE Full text] [doi: 10.1016/j.acvd.2020.05.006] [Medline: 32868257]
80. Rosman L, Armbruster T, Kyazimzade S, Tugaoen Z, Mazzella AJ, Deyo Z, et al. Effect of a virtual self-management intervention for atrial fibrillation during the outbreak of COVID-19. Pacing Clin Electrophysiol 2021;44(3):451-461 [FREE Full text] [doi: 10.1111/pace.14188] [Medline: 33565642]
81. Toscos T, Coupe A, Wagner S, Ahmed R, Roebuck A, Flanagan M, et al. Engaging patients in atrial fibrillation management via digital health technology: the impact of tailored messaging. J Innov Card Rhythm Manag 2020;11(8):4209-4217 [FREE Full text] [doi: 10.19102/icrm.2020.110802] [Medline: 32874747]
82. Ferguson C, Hickman LD, Phillips J, Newton PJ, Inglis SC, Lam L, et al. An mHealth intervention to improve nurses' atrial fibrillation and anticoagulation knowledge and practice: the EVICOAG study. Eur J Cardiovasc Nurs 2019;18(1):7-15. [doi: 10.1177/1474515118793051] [Medline: 30064254]
83. Matthews S, Atkins B, Walton N, Mitchell JA, Jennings G, Buttery AK. Development and use of a cardiac clinical guideline mobile app in Australia: acceptability and multi-methods study. JMIR Form Res 2022;6(8):e35599 [FREE Full text] [doi: 10.2196/35599] [Medline: 35930350]

XSL•FO
RenderX

84. Peleg M, Shahar Y, Quaglini S, Fux A, García-Sáez G, Goldstein A, et al. MobiGuide: a personalized and patient-centric decision-support system and its evaluation in the atrial fibrillation and gestational diabetes domains. User Model User-Adap Inter 2017;27(2):159-213. [doi: 10.1007/s11257-017-9190-5]
85. Desteghe L, Vijgen J, Koopman P, Dilling-Boer D, Schurmans J, Dendale P, et al. Telemonitoring-based feedback improves adherence to non-vitamin K antagonist oral anticoagulants intake in patients with atrial fibrillation. Eur Heart J 2018;39(16):1394-1403. [doi: 10.1093/eurheartj/ehx762] [Medline: 29300888]
86. Waring ME, Hills MT, Lessard DM, Saczynski JS, Libby BA, Holovatska MM, et al. Characteristics associated with Facebook use and interest in digital disease support among older adults with atrial fibrillation: cross-sectional analysis of baseline data from the systematic assessment of geriatric elements in atrial fibrillation (SAGE-AF) cohort. JMIR Cardio 2019;3(2):e15320 [FREE Full text] [doi: 10.2196/15320] [Medline: 31758791]
87. Peleg M, Michalowski W, Wilk S, Parimbelli E, Bonaccio S, O'Sullivan D, et al. Ideating mobile health behavioral support for compliance to therapy for patients with chronic disease: a case study of atrial fibrillation management. J Med Syst 2018;42(11):234. [doi: 10.1007/s10916-018-1077-4] [Medline: 30317403]
88. Verhaert DVM, Betz K, Gawałko M, Hermans ANL, Pluymaekers NAHA, van der Velden RMJ, et al. A virtual sleep apnoea management pathway for the work-up of atrial fibrillation patients in a digital remote infrastructure: VIRTUAL-SAFARI. Europace 2022;24(4):565-575. [doi: 10.1093/europace/euab229] [Medline: 34718525]
89. Nuvvula S, Ding EY, Saleeba C, Shi Q, Wang Z, Kapoor A, et al. NExUS-heart: novel examinations using smart technologies for heart health—data sharing from commercial wearable devices and telehealth engagement in participants with or at risk of atrial fibrillation. Cardiovasc Digit Health J 2021;2(5):256-263 [FREE Full text] [doi: 10.1016/j.cvdhj.2021.08.001] [Medline: 35265917]
90. Manninger M, Zweiker D, Svennberg E, Chatzikyriakou S, Pavlovic N, Zaman JAB, et al. Current perspectives on wearable rhythm recordings for clinical decision-making: the wEHRAbles 2 survey. Europace 2021;23(7):1106-1113. [doi: 10.1093/europace/euab064] [Medline: 33842972]
91. Ding EY, Svennberg E, Wurster C, Duncker D, Manninger M, Lubitz SA, et al. Survey of current perspectives on consumer-available digital health devices for detecting atrial fibrillation. Cardiovasc Digit Health J 2020;1(1):21-29 [FREE Full text] [doi: 10.1016/j.cvdhj.2020.06.002] [Medline: 32924024]
92. Kapoor A, Hayes A, Patel J, Patel H, Andrade A, Mazor K, et al. Usability and perceived usefulness of the AFib 2gether mobile app in a clinical setting: single-arm intervention study. JMIR Cardio 2021;5(2):e27016 [FREE Full text] [doi: 10.2196/27016] [Medline: 34806997]
93. Väliaho ES, Lipponen JA, Kuoppa P, Martikainen TJ, Jäntti H, Rissanen TT, et al. Continuous 24-h photoplethysmogram monitoring enables detection of atrial fibrillation. Front Physiol 2021;12:778775 [FREE Full text] [doi: 10.3389/fphys.2021.778775] [Medline: 35058796]
94. Mathew S, Chambers R. Improving the utility and sustainability of novel health technology to improve clinical outcomes for patients: an East Staffordshire experience of screening for atrial fibrillation with the AliveCor KardiaMobile. BJGP Open 2021;5(2):BJGPO.2020.0169 [FREE Full text] [doi: 10.3399/BJGPO.2020.0169] [Medline: 33622682]
95. Morak J, Kumpusch H, Hayn D, Leitner M, Scherr D, Fruhwald FM, et al. Near field communication-based telemonitoring with integrated ECG recordings. Appl Clin Inform 2011;2(4):481-498 [FREE Full text] [doi: 10.4338/ACI-2010-12-RA-0078] [Medline: 23616890]
96. Santala OE, Halonen J, Martikainen S, Jäntti H, Rissanen TT, Tarvainen MP, et al. Automatic mobile health arrhythmia monitoring for the detection of atrial fibrillation: prospective feasibility, accuracy, and user experience study. JMIR Mhealth Uhealth 2021;9(10):e29933 [FREE Full text] [doi: 10.2196/29933] [Medline: 34677135]
97. Standing C, Standing S, McDermott ML, Gururajan R, Kiani Mavi R. The paradoxes of telehealth: a review of the literature 2000–2015. Syst. Res 2016;35(1):90-101. [doi: 10.1002/sres.2442]

Abbreviations

AF: atrial fibrillation
ECG: electrocardiography
ISO: International Organization for Standardization
mHealth: mobile health
MRC: Medical Research Council
PPG: photoplethysmography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Edited by A Mavragani; submitted 13.11.23; peer-reviewed by M Wright, T Bickmore; comments to author 26.02.24; revised version received 03.06.24; accepted 24.07.24; published 04.10.24.

Please cite as:
Suresh Kumar S, Connolly P, Maier A
Considering User Experience and Behavioral Approaches in the Design of mHealth Interventions for Atrial Fibrillation: Systematic Review
J Med Internet Res 2024;26:e54405
URL: https://www.jmir.org/2024/1/e54405
doi:10.2196/54405
PMID:

©Sagar Suresh Kumar, Patricia Connolly, Anja Maier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Interventions for Managing Medication and Health Care Service Delivery in West Africa: Systematic Review

Emmanuel Oluwatosin Oluokun[1*], PhD; Festus Fatai Adedoyin[1*], PhD; Huseyin Dogan[1*], PhD; Nan Jiang[1*], PhD

Department of Computing and Informatics, Faculty of Science and Technology, Bournemouth University, Poole, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Emmanuel Oluwatosin Oluokun, PhD
Department of Computing and Informatics
Faculty of Science and Technology
Bournemouth University
Fern Barrow, Talbot
Poole, BH12 5BB
United Kingdom
Phone: 44 01202 524111
Email: eoluokun@bournemouth.ac.uk

Abstract

Background: As a result of the recent advancements in technology, the incorporation of digital interventions into the health care system has gained a lot of attention and adoption globally. However, these interventions have not been fully adopted, thereby limiting their impact on health care delivery in West Africa.

Objective: This review primarily aims at evaluating the current digital interventions for medication and health care delivery in West Africa. Its secondary aim is to assess the impacts of digital interventions in managing medication and health care service delivery with the intent of providing vital recommendations that would contribute to an excellent adoption of digital intervention tools in the health care space in West Africa.

Methods: In line with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), a comprehensive search through various databases yielded 529 results. After a rigorous screening, 29 articles that provided information on 3 broad digital health intervention tools were found eligible for this review.

Results: Out of 29 studies, 16 (55%) studies examined phone-based interventions, 9 (31%) studies focused on tele- and e-based interventions, and 4 (14%) studies evaluated digital interventions. These interventions were used for diverse purposes, some of which are monitoring adverse drug reactions, general health, sexual and reproductive health, and training of health care practitioners. The phone-based intervention appears to be the most known and impactful of all the interventions, followed by tele- and e-based, while digital interventions were scarcely used.

Conclusions: Digital interventions have had a considerable level of impact on medication and health care delivery across West Africa. However, the overall impact is limited. Therefore, strategies must be developed to address the challenges limiting the use of digital intervention tools so that these tools can be fully incorporated into the health care space in West Africa.

(J Med Internet Res 2024;26:e44294) doi:10.2196/44294

KEYWORDS

digital interventions; medications delivery; phone-based intervention; tele and e-based intervention; West Africa; management, technology; intervention; medication; tool; smartphone

Introduction

Background

In the last few decades, information technology has witnessed huge transitions [1]. These have facilitated the incorporation of programs and electronic devices using digital technology (digital intervention) for medication delivery and the health care system globally [2,3]. Recently, a group of researchers from the University of Michigan established that mobile health apps permit users to be more knowledgeable about their health, track behavioral changes in health, and obtain digital health assistance through a smartphone or tablet [4]. In the same vein, experts in the health care system have been reported to use mobile devices

XSL•FO
RenderX

to connect with patients for consultation, monitoring, management, and making clinical decisions [5]. Furthermore, the use of mobile devices has been established to advance communication between health care experts and their colleagues [6].

Over the years, several digital intervention tools such as telephone-based interventions, web-based interventions, and mobile apps have been used for supporting smoking cessation, vaccine rates, disease management (type 2 diabetes, hypertension, cancer, etc), and medication delivery [7-13]. Studies have also revealed that digital intervention resulted in health benefits [14]. Hence, these previous studies have therefore provided strong evidence that depicts how digital interventions have been employed in medication delivery and the health care system [15].

There has been a plethora of reviews on the integration of digital interventions into the health care system. Recently, Giravi et al [16] conducted a mini review on adjunct digital therapies for pain management, while Ibrahim et al [17] also conducted a systematic review of studies on digital health for quality health care. The outcome of these reviews revealed that digital interventions have had positive impacts on health. However, most of these reviews have focused on the impacts of integrating digital interventions into the health care system outside Africa, while a few are within Africa. To date, no study has systematically reviewed the digital interventions for medication delivery in West Africa.

In most African countries, health care systems are typically in deplorable situations with severely meager health outcomes [18]. Although, Africa makes up a little above a tenth of the global population, its disease burden stands at 24%, while Sub-Saharan Africa has access to less than 1% of global health expenditure [19]. Sub-Saharan Africa appears to be the worst region in the world, with most countries below the World Health Organization standards for basic health care [20]. Unfortunately, West Africa, being part of sub-Saharan Africa, tops the list of least developed countries in Africa, with a total of 12 countries enlisted [21]. Thus, the need to focus on West Africa. Given these potentials of digital intervention in health care delivery and the peculiarity of West Africa, this study was consequently designed to achieve a systematic review of the digital interventions used for medication and health care delivery in West Africa to assess its impact and provide recommendations that would improve its overall integration into the health care space.

Research Questions

This study aimed to evaluate the current digital interventions for medication and health care delivery in West Africa and also assess their overall impact in the health care space. Consequently, this review paper asks pertinent questions: What digital interventions are currently used for medication and health care delivery in West Africa? How well have these interventions been able to impact the health care system among West Africans and what are the challenges limiting its widespread adoption?

Methods

Overview

This systematic review uses PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [22], and information on search strategy, terminologies, inclusion and exclusion criteria, study selection, and data extraction is given in detail in subsequent sections. However, the review is not registered on the International Prospective Register of Systematic Reviews. This is because, while International Prospective Register of Systematic Reviews accepts systematic review protocols assessing interventions (including qualitative and individual participant data reviews), the nature of intervention for this review is mainly digital interventions in terms of technology and not medical or public health interventions. This is a potential limitation to this review, and future reviews could consider a deeper review and balance of both digital and non-digital (medical and public health) interventions.

Search Strategy

A comprehensive search through databases and various repositories was carried out using the PRISMA model for literature reviews [22]. Google Scholar, CINAHL, Web of Science, HubMed, BioMed, PubMed, Ovid, African Journals Online, and Scopus databases were searched to retrieve relevant information. Given that French and English are the two official languages in West Africa, our search encompassed databases in both languages.

Terminologies

The terminologies used in the search comprised “digital intervention,” “digital health,” “digital technology,” “mHealth,” “eHealth,” “medication delivery,” and “West Africa.” The keywords imputed into the databases included; (“Digital Intervention” AND Nigeria OR “West Africa” AND “Medication management” OR “medication delivery” OR “health care service delivery” OR “healthcare service delivery” AND “Health management” OR “disease management”). The same search approach was followed for all other countries in West Africa, which include Benin, Burkina Faso, Cape Verde, Côte D'Ivoire, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, and Togo with findings from the reviewed studies summarized in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

In line with the recommended guidelines for conducting a systematic review [22], certain conditions were fulfilled to conclude the inclusion and exclusion criteria for the research items used (Table 1).

XSL•FO
RenderX

Table 1. Summary of inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Article type	
• Reviewed journals and articles • Published materials, ie, case reports and case series on digital health intervention and medication delivery or general health care system between 2000 and 2022 • Studies showing significant health improvement and promoted well-being	• Editorials, letters, and protocol papers • Conference abstracts • Studies showing no remarkable improvement in health • Studies without a control group • Publications on the subject matter about countries outside West Africa
Language	
• Studies published in English • Studies published in French	• Studies not published in English • Studies published in French without comprehensive English translations

Data Extraction

Data extraction was independently carried out by 2 reviewers. The following reports were drawn: (1) authors, year, country, and reference; (2) journal; (3) type of health issue and health technology; (4) research approach/method; (5) main findings; and (6) critical appraisal of the strength and limitation of each research item.

Selection of Theme

The digitization of health care systems is referred to by terms such as eHealth, mHealth (mobile health), or digital health [23]. Understanding the significant variability of how these concepts interact with one another and their suitability in the digital health space is crucial given the diversity of terms and definitions used in the field of digital health. As a result, the themes used in this review were designed in line with these terminologies while we critically studied the articles so as to assign the various interventions under these themes, as seen in Figure 1 [24-32].

Figure 1. Core themes in the digitization of the health care system; mHealth, eHealth, digital health, and digital public health [24-32]. mHealth: mobile health.

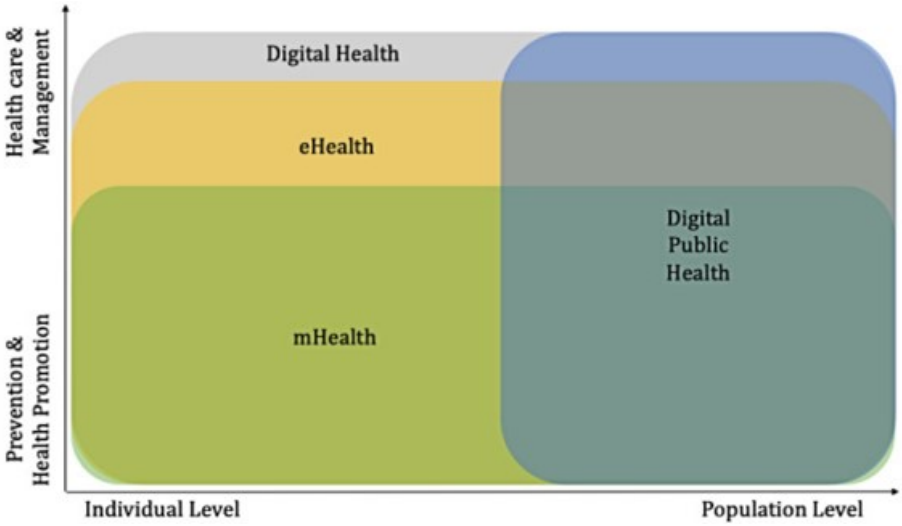

Results

Overview

The papers reviewed examined different digital interventions, including mobile, digital, and electronic technologies used in the health care system. Diverse research approaches were adopted by each of the papers, ranging from qualitative to quantitative, while some used assorted research methods. The terms used in the study include “digital interventions,” “digital technology,” “mobile health,” “eHealth,” and “Telehealth.”

Study Selection and Paper Quality Assessment

After a rigorous screening exercise in line with PRISMA guidelines (Multimedia Appendix 2), the flow diagram shown in Figure 2 presents a stepwise approach to the procedure used in study selection. From the search strategy, 529 records were retrieved from our various sources, out of which 255 duplicates were eliminated. A thorough screening was carried out on 274 records that had potential relevance via a careful examination of the abstracts and titles. Two members of the review team applied the exclusion criteria and found 78 studies eligible. At this point, an independent review of the preliminary result was carried out by 2 other members of the review team, of which only 29 (40%) were included in this review. Meanwhile, 5

XSL•FO
RenderX

articles were obtained by searching the citations of the 29 papers earlier selected. Out of the 5 articles obtained, only one additional article was selected after subjecting them to the selection criteria. Disagreement during screening between 2 reviewers was fixed via deliberation among the entire review team. Although a thorough check of each paper included in the review has been conducted following existing quality checks such as CASP (Critical Appraisal Skills Programme) checklists, other formal checks such as Joanna Briggs Institute's critical appraisal tools can also be beneficial.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection and inclusion process. Flow diagram retrieved from Page et al [22].

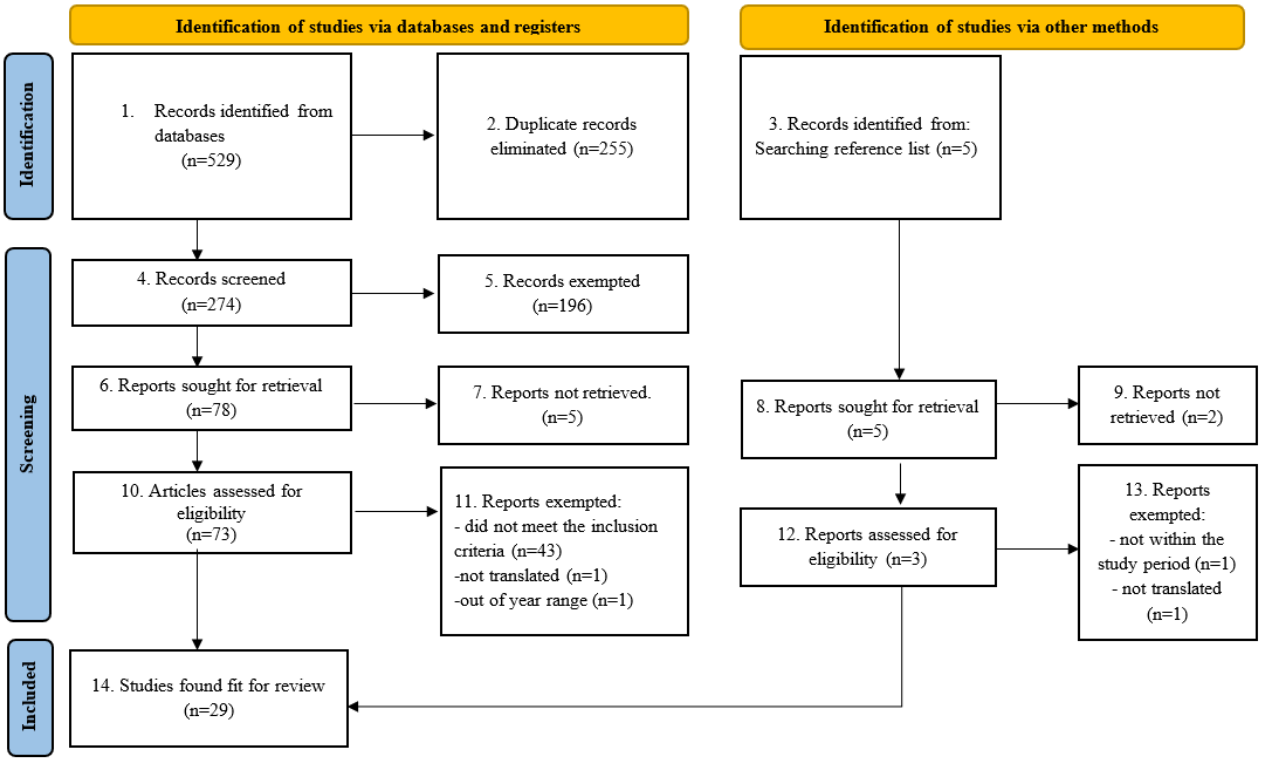

Location of Studies Included

Figure 3 shows that there were 14 papers from Nigeria, 9 from Ghana, 3 from Mali, 1 from Gambia, and 3 from other African countries, while there was no publication from other West African countries.

Figure 3. Papers reviewed per country. Source: Authors' compilation.

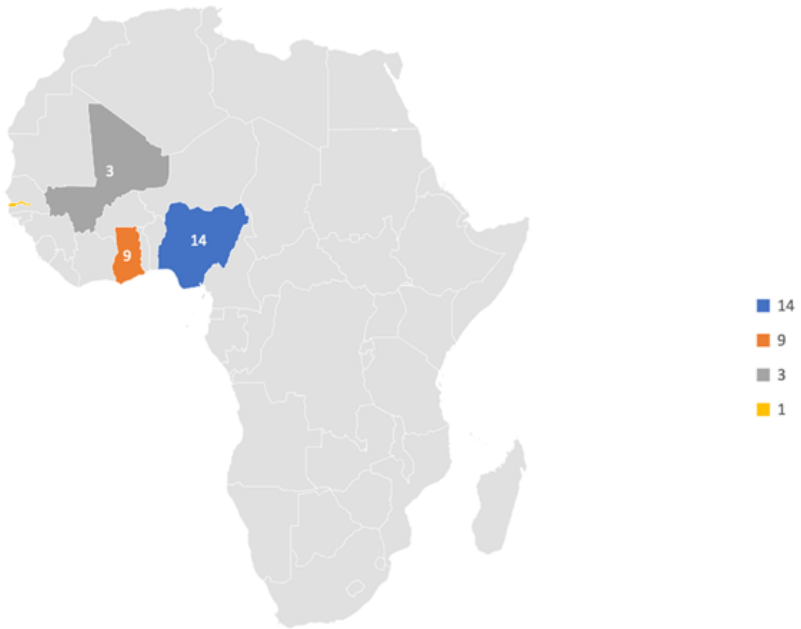

XSL•FO
RenderX

Characteristics of Interventions and Sources of Evidence

This review had 16 articles evaluating phone-based interventions, while 9 articles examined tele- and e-based interventions and 5 focused on digital interventions. These are presented pictorially in Multimedia Appendix 3. We compressed the major findings from the study into 3 themes based on the interventions identified. These themes capture the key digital health interventions used in the health care system in West Africa. Of all the papers reviewed, mobile phone caller tunes, SMS text messages, phone calls, mobile apps, slide projectors, video training, data digitization, amplifier systems, educational digital storytelling, SatCom (satellite communications), and 3G mobile networks are the frequently occurring themes. Nevertheless, we distinguished the definite themes that characterize the core findings of each paper with a tick (Table 2): phone-based interventions, tele- and e-based interventions and digital interventions.

Table 2. Themes coined from the papers under review.

References	Theme 1: Phone-based interventions	Theme 2: Tele- and e-based interventions	Theme 3: Digital interventions
[33]	✓		
[34]	✓		
[35]			✓
[36]	✓		
[37]	✓		
[38]		✓	
[39]			✓
[40]		✓	
[41]		✓	
[42]		✓	
[43]	✓		
[44]	✓		
[45]	✓		
[46]			✓
[47]	✓		
[48]	✓		
[49]	✓		
[50]	✓		
[51]	✓		
[52]	✓		
[53]			✓
[54]	✓		
[55]	✓		
[56]	✓		
[57]			✓
[58]		✓	
[59]		✓	
[60]		✓	
[61]		✓	
[62]		✓	

Theme 1: Phone-Based Interventions

It was found that toll-free telephone lines may be used to improve pharmacovigilance and follow-up on medication response in an environment with a paucity of funds [50]. In the same vein, Appiah et al [51] and Kukula et al [52] also supported that telephone follow-up and mobile caller tunes, respectively, should be used by patients for reporting ADRs. In line with these studies, the National Prescribing Service in Australia has successfully used a helpline to monitor adverse medicine

XSL•FO
RenderX

reactions [63]. In a similar development, Appiah et al [55] supported the development of caller tunes to encourage blood donation, while they [56] opined that those who thought mobile caller tunes could improve medicine adherence were higher among those who approved the application. Universally, the use of mobile phones among health practitioners in refining health care has gained more attention [64]. Andreatta et al [49] suggested that birth attendants may be trained to use mobile phones to report and transfer postpartum hemorrhage data.

Furthermore, the findings of Pop-Eleches et al [44] suggested that reminders in the form of SMS text messages on mobile phones could improve response to antiretroviral therapy in regions with scarce resources. Similarly, Raifman et al [43] opined that receiving mobile SMS text messages can advance the use of antimalarial treatment. The study of Babatunde et al [34] also corroborated that, potentially, mHealth may improve access to health care services. In another study, L'Engle et al [45] engaged mobile phones in improving HIV care and treatment. However, the study was suspended due to issues with the financing source. Similarly, Ishola and Chipps [33] found out that, mHealth could confer psychological flexibility if introduced into the prevention of mother-child HIV transmission. These findings agree with those of Dzansi et al [65], who concluded that mobile phones improved health outcomes in low- and middle-income settings.

Conversely, Stephani et al [48] posited that necessary resources and training are required for the effective adoption of mHealth solutions. He found out that patients at the diabetic center had a good mentality about mobile phones but were unfamiliar with them. The conclusion of Kenny et al [36] also indicates that despite the role of mobile phones in improving the health of sick children, health care providers have identified challenges limiting its widespread adoption. The findings of Gurupur and Wan [66] also concur with these findings.

Mobile technology has not only been effective in general health but also sexual and reproductive health. Based on findings from Rokicki et al [47], text-messaging initiatives have the potential to meaningfully improve reproductive health awareness and abate pregnancy. Similarly, Laing et al [54] reported that communications via mobile phones have made a positive impact on monitoring birth outcomes, and Otu et al [37] also opined that mHealth can transform the delivery of sexual reproductive health.

Theme 2: Tele- and e-Based Interventions

In rural areas, the cost of visiting hospitals often can be high. The COVID-19 pandemic made physical contact dangerous, and lockdowns became inevitable [67]. This has made the preference for telemedicine increase. Based on the outcome of Webb et al [9], he opined that tele-audiology could enhance better access to professional hearing health treatments. In a similar line, the use of telemedicine for equal access to health was evaluated by Bagayoko et al [59]. In the same vein, Bagayoko et al [59] found out that telehealth activities may improve medical diagnostics in cardiology and obstetrics and also the on-site patient monitoring system. It was also reported that health workers who took part in a telepathology program affirmed that the program culminated in increased skills and knowledge about cytopathology [40].

A study carried out by Mbemba et al [62] demonstrated the positive impact of telemedicine on the hiring and retention of medical professionals in Mali's rural locations. This aligns with Hicks et al [58], who found that video training supported by eHealth technology is an acceptable technique for improving clinical knowledge and service delivery to primary health workers in Nigeria.

The knowledge and insights of health care professionals (HCPs) who are at the center of health care delivery about telemedicine are key in determining its future [68]. According to the report of Abodunrin and Akande [42], it was submitted that health care professionals have good knowledge of telemedicine and eHealth. Nevertheless, they alluded to limitations in the application of telemedicine and e-Health in Nigeria to cost and dearth of infrastructure. In the same vein, Monsudi et al [38] reported that the importance of telemedicine in West Africa is known among many HCPs. However, only a few of them were aware of its availability at the hospital. Of all the papers reviewed, only Batta and Iwokwagh [60] opined that the use of news and social media serves as a means of transmitting health information. Nonetheless, its benefit has not been fully harnessed in the health care space. This agrees with the findings of Bekalu et al [69].

Theme 3: Digital Interventions

The report of Tabari [39] maintained the use of low-cost substitutes for printing computerized tomography image scans in developing countries via digital health technology. The outcome of Ebenso et al [53] is that video training and data digitization are promising solutions for promoting changes in maternal-child health care services. Similarly, Akeju et al [57] reported that video training and data digitization improved the level of maternity care delivery and attendance at health facilities. The work of Olu et al [35] reviewed the gains and pains accompanying the implementation of digital health for universal health coverage and proposed a conceptual framework that could make it widely implemented. Interestingly, of all the papers under review, only Ofoegbu et al [46] accessed the impact of education through digital storytelling. According to their study, education through digital storytelling was effective in boosting the risk perception and awareness of HIV among Nigerian adolescents.

Discussion

Overview

This review paper examines the impact of digital interventions on medication delivery in West Africa. It highlights the dire state of health care systems in the region and the potential for digital interventions to address these challenges. The study finds that phone-based interventions are the most used, with positive impacts on drug adherence and data reporting. However, there are limitations in study designs, sample sizes, and statistical methods across the reviewed literature. The paper recommends increased training for health care professionals, public awareness campaigns, and government intervention to address

XSL•FO
RenderX

infrastructure challenges. Despite its thoroughness, the review acknowledges the uneven distribution of literature across West Africa and calls for further research to provide robust information on digital interventions. Overall, the review provides valuable insights into the current state and future potential of digital interventions in West African health care.

Major Findings

We carried out a systematic review of the literature that assessed major digital interventions for medication and health care delivery in West Africa. In this study, mobile devices, tele- and electronic tools, and digital devices were used as interventions. The most frequently used is the phone-based intervention. The use of phone-based interventions has gained access into the health care space more than other interventions. Precisely, 30% of the studies under review focused on tele- and e-based tools as a digital health intervention. It is surprising to see that no study reported the use of news, social media, wearable devices, or videoconferencing as a digital intervention for health care delivery. Overall, 17% of the studies under review used digital tools as interventions. Interestingly, video training and data digitization are gradually gaining acceptance in Africa. Also, remarkably, the report of Ofoegbu et al [46] was striking as it focused on addressing a particular challenge among adolescents using digital storytelling.

The second research question, “how well have these interventions been able to impact the health care system among West Africans and what are the challenges limiting its widespread adoption?” was also answered. We discovered that these digital interventions have impacted the health care system among West Africans in the area of drug adherence and reporting mechanisms, leading to improved data sets. This can therefore serve as a teaching aid to both professionals and people, as a public health strategy and in direct care delivery. These studies point out the fact that digital intervention has had a positive impact and will positively impact the future of medication delivery and the health care system in West Africa. However, the paucity of information on the impacts of digital tools as an intervention in the health care space in West Africa calls for concern.

Critical Appraisal

During the reviewing process, 12 qualitative papers were included and 17 quantitative papers were reviewed. We carried out a critical evaluation with a focus on the papers' strengths and flaws. The evaluation followed the themes that were used in the systematic review. The first theme emphasized phone-based intervention. After reading the work of Adedeji et al [50], it became clear that the novelty of the study was that it was not self-reported. Instead, the authors tracked the purchase of antimalaria medications from 4 patent and proprietary medicine stores and 4 community pharmacies in a particular Nigerian community. The study, however, cannot be extrapolated to all of Nigeria due to the small sample size, time, and financial efficiency of the study. It is interesting to note that the study by Pop-Eleches et al [44] was one of the first of its kind to use a quantitative approach to demonstrate the positive effects of mHealth in the provision of antiretroviral therapy (HIV/AIDS) care. The power of the analysis may be lowered since the study's scope is too wide and the sample size is small. These restrictions mirrored those of Adedeji et al [50].

Following a critique of Ishola and Chipps [33], the study used a randomized control experiment with empirical data to gauge the effectiveness of a mobile health intervention. Scores from the pre- and post-tests were analyzed. The study's conclusions, however, were constrained in that they could not be generalized; hence, it is important to use caution when applying research findings. The test results are constrained by the inadequacies of the measures used. Another study by Kenny et al [36] looked at the difficulties in adopting mHealth from the viewpoint of primary health care workers. Nevertheless, because a specific case study project was chosen, the generalization may be difficult. Furthermore, the hypothesis was not empirically validated, which further restricted the findings' ability to be generalized. In a similar line, the findings of Andreatta et al [49] were deemed strong in that they looked into the significance of cell phones for reporting postpartum hemorrhage data by professional and traditional delivery attendants in rural Africa. The survey data, however, was not quantified using the proper statistical methods for the post-test design. Reporting solely on how the attendant adhered to the established cell phone usage protocol is unreliable. This implies that a significant portion of the findings on digital health interventions in Africa would necessitate a suitable experimental design and precise statistical methods right from the start.

On the other hand, after delving deeply into the conclusions of Kukula et al [52], we realized that the study's strength was a very large sample size and a powerful quantitative analysis that was used to analyze the data gathered. The project could not, however, be completed without financial support from an external institution due to its cost implications. Similar to this, the study by Appiah et al [55] received excellent marks because it tested the hypothesis using structural equation modeling, a powerful statistical technique. However, because the study was carried out in a specific hospital in Ghana, generalizing about rural and urban settings, much alone Sub-Saharan Africa, is inappropriate. Similarly, Rokicki et al [47] randomly assigned their subjects (adolescent girls) into groups using a high sample size and a computer-generated random number. Despite this, all information in the survey was self-reported; participants in trials might have been more tempted to hide their sexual behavior. The survey does not include adolescent girls in the rural sector and is primarily focused on Ghana’s urban areas. Additionally, starting with their initial antenatal appointments, Laing et al [54] followed pregnant women and kept track of any potentially harmful birth outcomes. In order to track stillbirths and neonatal fatalities, the author used a mobile telephone follow-up regardless of whether the expectant woman gave delivery inside or outside of a hospital. However, the outcome of an observational study is somewhat unreliable compared with resolving empirical evidence.

We evaluated other studies and discovered that these had limitations in their outcomes, making them differ from studies with limitations due to study designs, sample sizes, and other statistical methods. Another study [45] was notable for being the first to use a clinical trial to investigate the effects of mobile health on HIV prevention and care among PLHIV in a

XSL•FO
RenderX

large-scale public sector context in Ghana. It was self-reported adherence to antiretroviral therapy, though, and this has notable limitations. This might take the place of laboratory tests or objective measurements from doctors, pharmacists, or lab scientists. Additionally, Raifman et al [43] should be commended for seeking the participants' consent prior to conducting research. This was accomplished using flyers to find volunteers for an intervention and a self-enrolled mobile health service. As a result, those unwilling to engage in the study were not included. Even though they were willing to join, individuals without mobile phones were sadly unable to receive SMS text messages. Similar to this, the findings of Babatunde et al [34] emphasized how rural areas have limited access to mHealth, which hinders their ability to achieve Universal Health Coverage due to issues like inadequate electricity, bad data, limited internet connection, expensive mobile phones, etc. However, the research might be used as more proof rather than a description of the issue and possible solutions, such as mHealth. In a similar line, the findings of Appiah et al from 2021 were convincing about the development of the ATM to enhance medication adherence in Ghana. However, the study does not take into account that moderators can affect the findings. The study setting (diabetes clinic, Kumasi, Ghana) has significant promise for mHealth, and the methods used to analyze the survey results were supported by science, according to Stephani et al [48]. However, only patients who were present in the clinic at the time of the survey were interviewed, and because the responses were self-reported, the internal validity and reliability of the responses could not be determined because the sample selection technique could not accommodate patients outside the clinic. The study by Appiah et al [51] had a strong point in that it examined the significance of mHealth technology for the promotion of reporting ADRs; however, factors that could impact the production and usage of caller tunes to telecommunications were not considered. The findings of Otu et al [37] addressed a specific aspect of health care delivery and were considered and applied to a Nigerian context; however, not much evidence was provided on why, what form, or how the mHealth adoption process can be implemented.

Furthermore, we reviewed holistically the articles that were discussed under theme 2, which focused on tele- and e-based intervention. The results of the study by Mbemba et al [62] were deemed to be strong in that it evaluated the perceived impact of telemedicine on the hiring and retention of medical professionals in rural locations in Mali; however, the study did not include the actual demonstration of telehealth in recruiting and retaining professional workers in a remote area; rather, it only sought the perception of the health care workers in a remote area. In a similar vein, an assessment by Bagayoko et al [59] found that the strength of the study was in the fact that it examined the economic benefit of telehealth in 3 different district hospitals in Mali. To increase the accuracy of the outcome, however, there was hypothesis testing. Frequency and percentage alone are insufficient. In a similar vein, the study by Monsudi et al [38] deserves praise for examining the availability and awareness of telemedicine from the specific employees in a particular region of Nigeria. Although it was hampered by its departure from its intended topic by discussing the number of attendees rather than concentrating solely on addressing the causes for people's ignorance of telemedicine's presence in the hospital. Therefore, the study cannot be extrapolated to the entirety of Nigeria, which conflicts with findings from studies conducted in a different part of the nation. The study conducted by Malami [40] portrayed that the evaluation of the effects of telepathology training on ongoing cytology education was another strength. The study did not, however, carry out the cytology training itself. It only examined the success of earlier telepathology and recommended that cytology training follow suit.

The study by Batta and Iwokwagh [60] was interesting since it showed that the advantages of news and social media have not been sufficiently investigated. The majority of teaching hospitals and specialty clinics use it for publicity. Its shortcomings were that it advocated for Nigeria to adopt the concept of news and social media platforms without outlining the precise steps that may be taken to maximize their usefulness. Similarly, the investigation of the viability and acceptance of eHealth technologies in enhancing service delivery and performance of primary health workers in Nigeria was a strength of the study conducted by Hicks et al [58]. But just 3 Nigerian states were examined in the study. This can be a result of a financial problem. Additionally, there was no mention of reaching the states' rural areas. In contrast, the study by Abodunrin and Akande [42] has the advantage of identifying the potential causes—such as a lack of reliable electricity or internet access or unstable political conditions—that might be impeding Nigeria's adoption of e-Health and telemedicine to the fullest. However, the study was conducted in a specific state in Nigeria. Hence, generalizing is a problem here. The assessment of Bagayoko et al [61] revealed a summary of its strength, which was specific to using Réseau en Afrique Francophone pour la Télémédecine network to improve health care delivery in Mali. The shallowness of the findings and conclusion on the use of Réseau en Afrique Francophone pour la Télémédecine limited the study.

Next, we assessed the pieces that fell within theme 3. The power of the work by Olu et al [35] was demonstrated by the way it presented a range of benefits and drawbacks of Africans adopting digital technology in order to develop a conceptual framework for implementation. The research, however, is very theoretical and general. Using Nigerian college students as a case study, the essay expertly examined the impact of digital storytelling on HIV perception and knowledge [46]. Nevertheless, the sample size is modest when you evaluate Nigeria's entire state from a broad perspective. The study's focus is also restricted to just one particular area of Nigeria. Thus, in a similar discussion of constraints relating to design and analytical techniques, ref. [46] was also included.

On the other hand, the constraints of a few more studies affected the results. For instance, Tabari [39] proposed a low-cost alternative way of producing computerized tomography scan images in the absence of a dedicated camera. This is a really solid argument for the piece. The analysis of the created low-cost computerized topographical image, however, is not sufficiently detailed for replication. As a result, the study is only descriptive and constrained. In a similar vein, Ebenso et al [53] conducted a study in rural Nigeria (3 distinct states) to improve the

XSL•FO
RenderX

provision of maternal and child health care. However, the study's focus is limited to rural areas with few resources. Urban areas might not be aware that digital health interventions are being used. In conclusion, the findings by Akeju et al [57] made a solid case for the distribution of video training and data digitization interventions to 62 health care facilities over a 2-year period (2017-2019) in areas lacking SatCom and 3G mobile networks. Although the report referenced southwest Nigeria, it only included information on one of the 5 Southwestern states' rural health facilities. Therefore, the study does not explore what the government may do to enhance eHealth applications in rural areas. These strengths and flaws mainly focus on experimental designs, statistical tools used, study size, and outcomes. Hence, it is pertinent to consider using the right sample size and statistical tools for a better outcome [70,71].

Overall Integrated Findings and Comparison With Previous Research

Following the critical appraisal, there is also guidance on the overall integrated findings of this review, which informs recommendations that guide evidence-based policy and agenda for future research. According to Stern et al [72] and Lizarondo et al [73], where the core review question uses a mix of qualitative and quantitative papers, authors can use the convergent integrated approach, which allows for transformation of data and for capturing integrated findings from the review. Consequently, our review showed that 53.3% of the studies under review used the phone as an intervention. A previous study [74] conducted among community pharmacists showed that 69.2% of the studies reviewed used the telephone as a digital intervention, while another study [75] showed that 78.9% of the clinical pharmacists preferred the use of the telephone to other digital intervention tools. Both studies agree with the result of our review, as they pitched the telephone as the most regularly used digital health intervention.

Strengths and Limitations

A major strength of this review is the huge and productive efforts put into verifying reasonable parameters. There are also areas of significant input as well as areas of improvement for papers reviewed while we surveyed various digital health interventions. Additionally, the research approach used is another strong point of this review. This opened the researchers to diverse instruments, models, and parameters used to evaluate digital interventions for medication and health care delivery in West Africa.

However, various limitations are worthy of note. First and foremost, this literature review only focuses on a few West African countries. This means that there are many countries that were not included while conducting this review. The focus on a few West African countries limits the generalizability of findings, highlighting the need for broader regional studies. Furthermore, the authors were also constrained by the word count limit, thus, the themes and subthemes could not be adequately explored, while the various columns in the critical review table showing methodology and areas of strengths and weaknesses can be improved. Furthermore, there are very limited studies in West Africa on this subject at the moment. A meta-analysis would be a consideration for future studies, but with more focus on Sub-Saharan Africa to do a deeper and in-depth search on a larger sample/region size. Although a thorough check of each paper included in the review has been conducted following existing quality checks such as CASP checklists, other formal checks such as Joanna Briggs Institute's critical appraisal tools can also be beneficial.

Conclusions

This review paper is the first systematic review of the impacts of digital interventions on medication delivery in West Africa. To ensure we do not miss any relevant articles, we followed the recommendations of the PRISMA statement. This review shows that in West Africa, the availability of literature across West Africa is not even. Thus, more research needs to be done in all regions of West Africa to provide robust information on the digital interventions for medication delivery in West Africa. Overall, we found that phone-based interventions and tele- and e-based interventions were largely more helpful, which is a strong basis to recommend the sustained use of these interventions. Moreover, only a little evidence is available about the solutions to the challenges limiting the use of digital interventions. The critical appraisal of existing studies underscores the need for robust research methodologies and rigorous statistical analysis to ensure the reliability and generalizability of findings. Many studies reviewed exhibited limitations in sample size, study design, and analytical techniques, highlighting the importance of addressing these shortcomings in future research endeavors. Thus, there is a need for comprehensive research to first identify innovative interventions that will be acceptable both in rural and urban areas, proffer solutions to the menace of poor implementation of these interventions, and formulate policies that would encourage a widespread integration of these interventions into the health care system in West Africa.

Policy interventions are crucial to creating an enabling environment for the widespread adoption of digital interventions. Governments and relevant stakeholders should prioritize investments in health care infrastructure and technology while also formulating policies that incentivize the use of digital solutions. Moreover, collaboration between governments, health care institutions, and technology providers is essential to ensure the successful implementation and sustainability of digital health care initiatives in West Africa.

Recommendations

The main recommendation of this study is that the use of digital interventions for medication delivery needs to be significantly improved. Following the papers reviewed, it has been shown clearly that individuals, patients, and health care professionals have adopted some interventions, while some have not been implemented at all. There is, therefore, a need for a holistic implementation of digital intervention tools in the health care space in West Africa.

First, health care professionals need to be trained on the technicalities necessary to integrate digital interventions into health care delivery in West Africa. This is in line with the findings of Feroz et al [76]. Furthermore, people should be sensitized about the use of DI in health care delivery both in

XSL•FO
RenderX

rural and urban areas. In addition, Abodunrin and Akande [42] pinpointed the possible reasons for the poor adoption of digital interventions, such as poor power supply, internet service, the poor state of infrastructure, and so on. Therefore, this review again recommends that the government and other key agencies involved in proffering solutions to these should not hesitate to. Finally, platforms must be established for training HCPs on managing the technicalities involved, sensitizing the users, and formulating policies that would encourage a widespread integration of these interventions into the health care system in West Africa.

Acknowledgments

The authors acknowledge the Department of Computing and Informatics, Bournemouth University, United Kingdom.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. They are made publicly available.

Authors' Contributions

All the authors contributed equally to the writing, review, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of included studies.
[DOCX File , 45 KB - jmir_v26i1e44294_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 36 KB - jmir_v26i1e44294_app2.docx]

Multimedia Appendix 3
Number of papers per intervention. Source: Authors' compilation.
[PNG File , 23 KB - jmir_v26i1e44294_app3.png]

References

1. Valverde S. Major transitions in information technology. Philos Trans R Soc Lond B Biol Sci 2016 Aug 19;371(1701):20150450 [FREE Full text] [doi: 10.1098/rstb.2015.0450] [Medline: 27431527]
2. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
3. Awad A, Trenfield SJ, Pollard TD, Ong JJ, Elbadawi M, McCoubrey LE, et al. Connected healthcare: improving patient care using digital health technologies. Adv Drug Deliv Rev 2021;178:113958. [doi: 10.1016/j.addr.2021.113958] [Medline: 34478781]
4. Lee P, Aikens J, Richardson C, Singer D, Kullgren J, Kirch M, et al. Mobile health app use among older adults. University of Michigan National Poll on Healthy Aging 2020:1-2. [doi: 10.7302/4019]
5. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. Pharm Therap 2014;39(5):356-364 [FREE Full text] [Medline: 24883008]
6. Koivunen M, Niemi A, Hupli M. The use of electronic devices for communication with colleagues and other healthcare professionals - nursing professionals' perspectives. J Adv Nurs 2015;71(3):620-631. [doi: 10.1111/jan.12529] [Medline: 25250739]
7. Pal K, Dack C, Ross J, Michie S, May C, Stevenson F, et al. Digital health interventions for adults with type 2 diabetes: qualitative study of patient perspectives on diabetes self-management education and support. J Med Internet Res 2018;20(2):e40 [FREE Full text] [doi: 10.2196/jmir.8439] [Medline: 29463488]
8. Kassavou A, Houghton V, Edwards S, Brimicombe J, Sutton S. Development and piloting of a highly tailored digital intervention to support adherence to antihypertensive medications as an adjunct to primary care consultations. BMJ Open 2019;9(1):e024121 [FREE Full text] [doi: 10.1136/bmjopen-2018-024121] [Medline: 30613027]

XSL•FO
RenderX

9. Webb J, Peerbux S, Smittenaar P, Siddiqui S, Sherwani Y, Ahmed M, et al. Preliminary outcomes of a digital therapeutic intervention for smoking cessation in adult smokers: randomized controlled trial. JMIR Ment Health 2020;7(10):e22833 [FREE Full text] [doi: 10.2196/22833] [Medline: 33021488]
10. Raijada D, Wac K, Greisen E, Rantanen J, Genina N. Integration of personalized drug delivery systems into digital health. Adv Drug Deliv Rev 2021;176:113857 [FREE Full text] [doi: 10.1016/j.addr.2021.113857] [Medline: 34389172]
11. Hare AJ, Chokshi N, Adusumalli S. Novel digital technologies for blood pressure monitoring and hypertension management. Curr Cardiovasc Risk Rep 2021;15(8):11 [FREE Full text] [doi: 10.1007/s12170-021-00672-w] [Medline: 34127936]
12. Odone A, Gianfredi V, Sorbello S, Capraro M, Frascella B, Vigezzi GP, et al. The use of digital technologies to support vaccination programmes in europe: state of the art and best practices from experts' interviews. Vaccines (Basel) 2021;9(10):1126 [FREE Full text] [doi: 10.3390/vaccines9101126] [Medline: 34696234]
13. Marthick M, McGregor D, Alison J, Cheema B, Dhillon H, Shaw T. Supportive care interventions for people with cancer assisted by digital technology: systematic review. J Med Internet Res 2021;23(10):e24722 [FREE Full text] [doi: 10.2196/24722] [Medline: 34714246]
14. Bevilacqua R, Casaccia S, Cortellessa G, Astell A, Lattanzio F, Corsonello A, et al. Coaching through technology: a systematic review into efficacy and effectiveness for the ageing population. Int J Environ Res Public Health 2020;17(16):5930 [FREE Full text] [doi: 10.3390/ijerph17165930] [Medline: 32824169]
15. Ibeneme S, Ukor N, Ongom M, Dasa T, Muneene D, Okeibunor J. Strengthening capacities among digital health leaders for the development and implementation of national digital health programs in nigeria. BMC Proc 2020;14(Suppl 10):9 [FREE Full text] [doi: 10.1186/s12919-020-00193-1] [Medline: 32714444]
16. Giravi HY, Biskupiak Z, Tyler LS, Bulaj G. Adjunct digital interventions improve opioid-based pain management: impact of virtual reality and mobile applications on patient-centered pharmacy care. Front Digit Health 2022;4:884047 [FREE Full text] [doi: 10.3389/fdgth.2022.884047] [Medline: 35770137]
17. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810 [FREE Full text] [doi: 10.1177/20552076221085810] [Medline: 35340904]
18. Oleribe OO, Momoh J, Uzochukwu BS, Mbofana F, Adebiyi A, Barbera T, et al. Identifying key challenges facing healthcare systems in Africa and potential solutions. Int J Gen Med 2019;12:395-403 [FREE Full text] [doi: 10.2147/IJGM.S223882] [Medline: 31819592]
19. WHO. African Region Health Report. Geneva: United Nations; 2013. URL: https://www.afro.who.int/publications [accessed 2013-05-01]
20. Health care in Africa: IFC report sees demand for investment. International Finance Corporation. 2022. URL: https://www.ifc.org/wps/wcm/connect/news_ext_content/ifc_external_corporate_site/news+and+events/healthafricafeature#:~:text=Health%20care%20in%20Sub%2DSaharan,minimum%20for%20basic%20health%20care [accessed 2022-09-06]
21. Wale-Oshinowo BA, Omobowale AO, Adeyeye MM, Lebura S. Least developed countries in Africa. The Palgrave Encyclopedia of Global Security Studies 2022 Jan 01:1-16 [FREE Full text] [doi: 10.1007/978-3-319-74336-3_346-1]
22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Br Med J 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
23. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022;24(6):e31921 [FREE Full text] [doi: 10.2196/31921] [Medline: 35763320]
24. Nievas SB, García DS, Fernández AA, Bonillo PA, Parrón CT. eHealth: advantages, disadvantages and guiding principles for the future. J Med Internet Res 2019:1-23 [FREE Full text] [doi: 10.2196/preprints.15366]
25. Oh H, Rizo C, Enkin M, Jadad A. What is eHealth (3): a systematic review of published definitions. J Med Internet Res 2005;7(1):e1 [FREE Full text] [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]
26. Shaw T, McGregor D, Brunner M, Keep M, Janssen A, Barnet S. What is eHealth (6)? Development of a conceptual model for eHealth: qualitative study with key informants. J Med Internet Res 2017;19(10):e324 [FREE Full text] [doi: 10.2196/jmir.8106] [Medline: 29066429]
27. Boogerd EA, Arts T, Engelen LJ, van de Belt TH. "What Is eHealth": time for an update? JMIR Res Protoc 2015;4(1):e29 [FREE Full text] [doi: 10.2196/resprot.4065] [Medline: 25768939]
28. Iacono T, Stagg K, Pearce N, Hulme Chambers A. A scoping review of Australian allied health research in ehealth. BMC Health Serv Res 2016;16(1):543 [FREE Full text] [doi: 10.1186/s12913-016-1791-x] [Medline: 27716325]
29. Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan JP, et al. What is eHealth (4): a scoping exercise to map the field. J Med Internet Res 2005;7(1):e9 [FREE Full text] [doi: 10.2196/jmir.7.1.e9] [Medline: 15829481]
30. Davis TL, DiClemente R, Prietula M. Taking mHealth forward: examining the core characteristics. JMIR Mhealth Uhealth 2016;4(3):e97 [FREE Full text] [doi: 10.2196/mhealth.5659] [Medline: 27511612]
31. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]

XSL•FO
RenderX

32. Zeeb H, Pigeot I, Schüz B, Leibniz-Wissenschafts Campus Digital Public Health Bremen. Digital public health-an overview. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020;63(2):137-144. [doi: 10.1007/s00103-019-03078-7] [Medline: 31919531]
33. Ishola AG, Chipps J. The use of mobile phones to deliver acceptance and commitment therapy in the prevention of mother-child HIV transmission in Nigeria. J Telemed Telecare 2015;21(8):423-426. [doi: 10.1177/1357633X15605408] [Medline: 26391511]
34. Babatunde AO, Abdulkareem AA, Akinwande FO, Adebayo AO, Omenogor ET, Adebisi YA, et al. Leveraging mobile health technology towards achieving universal health coverage in Nigeria. Public Health Pract (Oxf) 2021;2:100120 [FREE Full text] [doi: 10.1016/j.puhip.2021.100120] [Medline: 36101601]
35. Olu O, Muneene D, Bataringaya JE, Nahimana MR, Ba H, Turgeon Y, et al. How can digital health technologies contribute to sustainable attainment of universal health coverage in Africa? A perspective. Front Public Health 2019;7:341 [FREE Full text] [doi: 10.3389/fpubh.2019.00341] [Medline: 31803706]
36. Kenny G, O'Connor Y, Eze E, Ndibuagu E, Heavin C. A ground-up approach to mHealth in Nigeria: a study of primary healthcare workers' attitude to mHealth adoption. Proc Comput Sci 2017;121:809-816 [FREE Full text] [doi: 10.1016/j.procs.2017.11.105]
37. Otu A, Ukpeh I, Okuzu O, Yaya S. Leveraging mobile health applications to improve sexual and reproductive health services in Nigeria: implications for practice and policy. Reprod Health 2021;18(1):21 [FREE Full text] [doi: 10.1186/s12978-021-01069-z] [Medline: 33485356]
38. Monsudi KF, Ayanniyi AA, Oguntunde OO. Awareness and practice of telemedicine among staff of the federal medical centre at Birnin Kebbi, Nigeria. J Telemed Telecare 2012;18(7):427-428. [doi: 10.1258/jtt.2012.120606] [Medline: 23104770]
39. Tabari AM. Low-cost printing of computerised tomography (CT) images where there is no dedicated CT camera. J Telemed Telecare 2007;13(6):274-276. [doi: 10.1258/135763307781644861] [Medline: 17785022]
40. Malami SA. Use of telemedicine for cytology training in Africa. J Telemed Telecare 2008;14(2):105-106. [doi: 10.1258/jtt.2007.070803] [Medline: 18348759]
41. Swanepoel DW, Olusanya BO, Mars M. Hearing health-care delivery in sub-Saharan Africa--a role for tele-audiology. J Telemed Telecare 2010;16(2):53-56. [doi: 10.1258/jtt.2009.009003] [Medline: 20008052]
42. Abodunrin OL, Akande TM. Knowledge and perception of e-health and telemedicine among health professionals in lautech teaching hospital, Osogbo, Nigeria. Int J Health Res 2010;2(1):51-58. [doi: 10.4314/ijhr.v2i1.55388]
43. Raifman JRG, Lanthorn HE, Rokicki S, Fink G. The impact of text message reminders on adherence to antimalarial treatment in northern Ghana: a randomized trial. PLoS One 2014;9(10):e109032 [FREE Full text] [doi: 10.1371/journal.pone.0109032] [Medline: 25350546]
44. Pop-Eleches C, Thirumurthy H, Habyarimana JP, Zivin JG, Goldstein MP, de Walque D, et al. Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders. AIDS 2011;25(6):825-834 [FREE Full text] [doi: 10.1097/QAD.0b013e32834380c1] [Medline: 21252632]
45. L'Engle KL, Green K, Succop SM, Laar A, Wambugu S. Scaled-up mobile phone intervention for HIV care and treatment: protocol for a facility randomized controlled trial. JMIR Res Protoc 2015;4(1):e11 [FREE Full text] [doi: 10.2196/resprot.3659] [Medline: 25650838]
46. Ofoegbu TO, Otu MS, Christopher I, Uche A, Nwabuko LO, Ebere I, et al. Impact of an educational digital storytelling intervention on HIV risk perception among Nigerian adolescents. J Int Med Res 2020;48(1):300060519854635 [FREE Full text] [doi: 10.1177/0300060519854635] [Medline: 31208259]
47. Rokicki S, Cohen J, Salomon JA, Fink G. Impact of a text-messaging program on adolescent reproductive health: a cluster-randomized trial in Ghana. Am J Public Health 2017;107(2):298-305. [doi: 10.2105/AJPH.2016.303562] [Medline: 27997236]
48. Stephani V, Opoku D, Otupiri E. Determining the potential of mobilephone-based health interventions in kumasi, Ghana. Ghana Med J 2020;54(2):88-92 [FREE Full text] [doi: 10.4314/gmj.v54i2.6] [Medline: 33536678]
49. Andreatta P, Debpuur D, Danquah A, Perosky J. Using cell phones to collect postpartum hemorrhage outcome data in rural Ghana. Int J Gynaecol Obstet 2011;113(2):148-151 [FREE Full text] [doi: 10.1016/j.ijgo.2010.11.020] [Medline: 21420086]
50. Adedeji AA, Sanusi B, Tella A, Akinsanya M, Ojo O, Akinwunmi MO, et al. Exposure to anti-malarial drugs and monitoring of adverse drug reactions using toll-free mobile phone calls in private retail sector in Sagamu, Nigeria: implications for pharmacovigilance. Malar J 2011;10:230 [FREE Full text] [doi: 10.1186/1475-2875-10-230] [Medline: 21827667]
51. Appiah B, Poudyal A, Burdine JN, Asamoah-Akuoko L, Anum DA, Kretchy IA, et al. Factors that influence the intention to use mobile phone caller tunes for patient reporting of adverse drug reactions: a qualitative study. Ther Adv Drug Saf 2019;10:2042098619871190 [FREE Full text] [doi: 10.1177/2042098619871190] [Medline: 35173950]
52. Kukula VA, Dodoo AAN, Akpakli J, Narh-Bana SA, Clerk C, Adjei A, et al. Feasibility and cost of using mobile phones for capturing drug safety information in peri-urban settlement in Ghana: a prospective cohort study of patients with uncomplicated malaria. Malar J 2015;14:411 [FREE Full text] [doi: 10.1186/s12936-015-0932-8] [Medline: 26481106]
53. Ebenso B, Okusanya B, Okunade K, Akeju D, Ajepe A, Akaba GO, et al. What are the contextual enablers and impacts of using digital technology to extend maternal and child health services to rural areas? Findings of a qualitative study from Nigeria. Front Glob Womens Health 2021;2:670494 [FREE Full text] [doi: 10.3389/fgwh.2021.670494] [Medline: 34816223]

XSL•FO
RenderX

54. Laing S, Remmelzwaal K, Cooper M, N'Dow J. Mobile telephone follow-up to ascertain birth outcomes in The Gambia. Telemed J E Health 2020;26(11):1363-1367 [FREE Full text] [doi: 10.1089/tmj.2019.0119] [Medline: 31976821]
55. Appiah B, Burdine JN, Aftab A, Asamoah-Akuoko L, Anum DA, Kretchy IA, et al. Determinants of intention to use mobile phone caller tunes to promote voluntary blood donation: cross-sectional study. JMIR Mhealth Uhealth 2018;6(5):e117 [FREE Full text] [doi: 10.2196/mhealth.9752] [Medline: 29728343]
56. Appiah B, Kretchy IA, Yoshikawa A, Asamoah-Akuoko L, France CR. Perceptions of a mobile phone-based approach to promote medication adherence: a cross-sectional application of the technology acceptance model. Explor Res Clin Soc Pharm 2021;1:100005 [FREE Full text] [doi: 10.1016/j.rcsop.2021.100005] [Medline: 35479503]
57. Akeju D, Okusanya B, Okunade K, Ajepe A, Allsop MJ, Ebenso B. Sustainability of the effects and impacts of using digital technology to extend maternal health services to rural and hard-to-reach populations: experience from Southwest Nigeria. Front Glob Womens Health 2022;3:696529 [FREE Full text] [doi: 10.3389/fgwh.2022.696529] [Medline: 35211694]
58. Hicks JP, Allsop MJ, Akaba GO, Yalma RM, Dirisu O, Okusanya B, et al. Acceptability and potential effectiveness of eHealth tools for training primary health workers from Nigeria at scale: mixed methods, uncontrolled before-and-after study. JMIR Mhealth Uhealth 2021;9(9):e24182 [FREE Full text] [doi: 10.2196/24182] [Medline: 34528891]
59. Bagayoko CO, Traoré D, Thevoz L, Diabaté S, Pecoul D, Niang M, et al. Medical and economic benefits of telehealth in low- and middle-income countries: results of a study in four district hospitals in mali. BMC Health Serv Res 2014;14(Suppl 1):S9 [FREE Full text] [doi: 10.1186/1472-6963-14-S1-S9] [Medline: 25080312]
60. Batta HE, Iwokwagh NS. Optimising the digital age health-wise: utilisation of new/social media by Nigerian teaching hospitals. Proc Soc Behav Sci 2015;176:175-185. [doi: 10.1016/j.sbspro.2015.01.459]
61. Bagayoko CO, Anne A, Fieschi M, Geissbuhler A. Can ICTs contribute to the efficiency and provide equitable access to the health care system in sub-Saharan Africa? The Mali experience. Yearb Med Inform 2011;6:33-38. [Medline: 21938322]
62. Mbemba GIC, Bagayoko CO, Gagnon MP, Hamelin-Brabant L, Simonyan DA. The influence of a telehealth project on healthcare professional recruitment and retention in remote areas in Mali: a longitudinal study. SAGE Open Med 2016;4:2050312116648047 [FREE Full text] [doi: 10.1177/2050312116648047] [Medline: 27231552]
63. NPS Medicinewise. National Prescribing Service, Medicine wise. URL: https://www.nps.org.au [accessed 2022-09-29]
64. Kalinga A, Munga M, Ngenya A, John W, Kisoka W, Oriyo N, et al. The viability of utilising phone-based text messages in data capture and reporting morbidities due to lymphatic filariasis by community health workers: a qualitative study in Kilwa district, Tanzania. BMC Health Serv Res 2022;22(1):924 [FREE Full text] [doi: 10.1186/s12913-022-08256-z] [Medline: 35854308]
65. Dzansi G, Chipps J, Lartey M. Use of mobile phone among patients with HIV/AIDS in a low-middle income setting: a descriptive exploratory study. Behav Inf Technol 2022;41(4):796-804. [doi: 10.1080/0144929x.2020.1836257]
66. Gurupur VP, Wan TTH. Challenges in implementing mHealth interventions: a technical perspective. mHealth 2017;3:32 [FREE Full text] [doi: 10.21037/mhealth.2017.07.05] [Medline: 28894742]
67. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
68. Ayatollahi H, Sarabi FZP, Langarizadeh M. Clinicians' knowledge and perception of telemedicine technology. Perspect Health Inf Manag 2015;12(Fall):1c [FREE Full text] [Medline: 26604872]
69. Bekalu MA, McCloud RF, Viswanath K. Association of social media use with social well-being, positive mental health, and self-rated health: disentangling routine use from emotional connection to use. Health Educ Behav 2019;46(2_suppl):69-80. [doi: 10.1177/1090198119863768] [Medline: 31742462]
70. Ali Z, Bhaskar SB. Basic statistical tools in research and data analysis. Indian J Anaesth 2016;60(9):662-669 [FREE Full text] [doi: 10.4103/0019-5049.190623] [Medline: 27729694]
71. Andrade C. Sample size and its importance in research. Indian J Psychol Med 2020;42(1):102-103 [FREE Full text] [doi: 10.4103/IJPSYM.IJPSYM_504_19] [Medline: 31997873]
72. Stern C, Lizarondo L, Carrier J, Godfrey C, Rieger K, Salmond S, et al. Methodological guidance for the conduct of mixed methods systematic reviews. JBI Evid Synth 2020;18(10):2108-2118. [doi: 10.11124/JBISRIR-D-19-00169] [Medline: 32813460]
73. Lizarondo L, Stern C, Apostolo J, Carrier J, de Borges K, Godfrey C, et al. Five common pitfalls in mixed methods systematic reviews: lessons learned. J Clin Epidemiol 2022;148:178-183. [doi: 10.1016/j.jclinepi.2022.03.014] [Medline: 35341946]
74. Crilly P, Kayyali R. A systematic review of randomized controlled trials of telehealth and digital technology use by community pharmacists to improve public health. Pharmacy (Basel) 2020;8(3):137 [FREE Full text] [doi: 10.3390/pharmacy8030137] [Medline: 32759850]
75. Park T, Muzumdar J, Kim H. Digital health interventions by clinical pharmacists: a systematic review. Int J Environ Res Public Health 2022;19(1):532 [FREE Full text] [doi: 10.3390/ijerph19010532] [Medline: 35010791]
76. Feroz AS, Khoja A, Saleem S. Equipping community health workers with digital tools for pandemic response in LMICs. Arch Public Health 2021;79(1):1 [FREE Full text] [doi: 10.1186/s13690-020-00513-z] [Medline: 33390163]

XSL•FO
RenderX

Abbreviations

CASP: Critical Appraisal Skills Programme
HCP: health care professional
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SatCom: satellite communications

Edited by A Mavragani; submitted 14.11.22; peer-reviewed by O Ijezie, N Adesina, V Heaslip, F Adeniyi; comments to author 25.07.23; revised version received 06.09.23; accepted 30.05.24; published 09.10.24.

Please cite as:
Oluokun EO, Adedoyin FF, Dogan H, Jiang N
Digital Interventions for Managing Medication and Health Care Service Delivery in West Africa: Systematic Review
J Med Internet Res 2024;26:e44294
URL: https://www.jmir.org/2024/1/e44294
doi:10.2196/44294
PMID:

©Emmanuel Oluwatosin Oluokun, Festus Fatai Adedoyin, Huseyin Dogan, Nan Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Exploring the Need for Medical Futures Studies: Insights From a Scoping Review of Health Care Foresight

Bertalan Meskó[1,2], MD, PhD; Tamás Kristóf[3], PhD; Pranavsingh Dhunnoo[1,4], MD; Nóra Árvai[1,2,5], MSc; Gellért Katonai[1,2,6], MD

[1]The Medical Futurist Institute, Budapest, Hungary
[2]Kálmán Laki Doctoral School of Biomedical and Clinical Sciences, University of Debrecen, Debrecen, Hungary
[3]Institute of Entrepreneurship and Innovation, Corvinus University of Budapest, Budapest, Hungary
[4]Department of Computing, Atlantic Technological University, Letterkenny, Co. Donegal, Ireland
[5]Meducation Hungary Kft, Budapest, Hungary
[6]Department of Family Medicine, Semmelweis University, Budapest, Hungary

Corresponding Author:
Bertalan Meskó, MD, PhD
The Medical Futurist Institute
Povl Bang-Jensen u. 2/B1. 4/1
Budapest, 1118
Hungary
Phone: 36 703807260
Email: berci@medicalfuturist.com

Abstract

Background: The historical development and contemporary instances of futures studies, an interdisciplinary field that focuses on exploring and formulating alternative futures, exemplify the increasing significance of using futures methods in shaping the health care domain. Despite the wide array of these methodologies, there have been limited endeavors to employ them within the medical community thus far.

Objective: We undertook the first scoping review to date about the application of futures methodologies and published foresight projects in health care.

Methods: Through the use of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) method, we identified 59 studies that were subsequently categorized into the following 5 distinct themes: national strategies (n=19), strategic health care foresight (n=15), health care policy and workforce dynamics (n=6), pandemic preparedness and response (n=7), and specialized medical domains (n=12).

Results: Our scoping review revealed that the application of futures methods and foresight has been successfully demonstrated in a wide range of fields, including national strategies, policy formulation, global threat preparedness, and technological advancements. The results of our review indicate that a total of 8 futures methods have already been used in medicine and health care, while there are more than 50 futures methods available. It may underscore the notion that the field is unexploited. Furthermore, the absence of structured methodologies and principles for employing foresight and futures techniques in the health care domain warrants the creation of medical futures studies as a separate scientific subfield within the broad domains of health care, medicine, and life sciences. This subfield would focus on the analysis of emerging technological trends, the evaluation of policy implications, and the proactive anticipation and mitigation of potential challenges.

Conclusions: Futures studies can significantly enhance medical science by addressing a crucial deficiency in the promotion of democratic participation, facilitating interdisciplinary dialogue, and shaping alternative futures. To further contribute to the development of a new research community in medical futures studies, it is recommended to establish a specialized scientific journal. Additionally, appointing dedicated futurists in decision-making and national strategy, and incorporating futures methods into the medical curriculum could be beneficial.

(J Med Internet Res 2024;26:e57148) doi:10.2196/57148

KEYWORDS

foresight; futures studies; health care future; medical futures; technology foresight

XSL•FO
RenderX

Introduction

Analyzing the future of medicine and health care has lately become a major focus for medical journals and the life science community owing to the exceptional technological progress witnessed in health care, digital health's cultural transformation of traditional roles of patients and medical professionals, and the growing importance of preparing for critical changes in the health care ecosystem, which was recently demonstrated by the COVID-19 pandemic.

While "futures studies" has been an established academic discipline for decades, medical professionals and life science researchers have rarely used so-called futures methods in their studies and analyses. However, acknowledging the growing role that a practical domain of futures studies (foresight) plays in shaping the health care landscape and tracking its historical development within this sector can shed light on previous trends and approaches while also establishing a foundation for the emergence and indispensability of medical futures studies as an emerging academic subfield. Textbox 1 describes the related basic terms [1].

Textbox 1. Basic terms.

Futures field: A broader domain that integrates futures studies and foresight to influence decision-making and strategic planning across different sectors.

Futures studies: An academic field focused on exploring and shaping various possible, probable, feasible, and preferred futures through interdisciplinary methods.

Foresight: The strategic practice of exploring and preparing for possible future scenarios to guide present decisions.

Futures methods: Techniques and tools used in futures studies and foresight to systematically explore, create, and test possible and desirable futures to improve decisions.

According to this paper [1], "futures studies" is classified as an academic subject, whereas foresight is considered a practical domain involving more specific strategic planning based on futures knowledge. Both subjects are part of the broader category known as the futures field.

Within this, rooted in social sciences, "futures studies" is an interdisciplinary field with the primary focus to explore and formulate alternative futures [2,3]. Throughout its multiple decades of development, "futures studies" has become a generally accepted academic discipline since the 1960s. Its merit is visible in the international scientific community, having global professional associations, higher education programs, and prestigious scientific journals. Futurists globally adhere to universally accepted theoretical concepts, methodologies, purposes, tenets, ethical principles, and empirical results. Textbox 2 summarizes its short history.

Textbox 2. History of futures studies.

"Futures studies" is an interdisciplinary field that is grounded in the social sciences. Its main objective is to explore and construct alternative futures. The primary goal of futures studies is to analyze and control the causal relationships between events and outcomes in the future using a systematic approach that incorporates feedback loops to foster social and technological innovations [4]. Kristóf and Nováky [3] have provided a comprehensive summary of the developmental trajectory of futures studies. Historically, the term "foresight" has been used to refer to the readiness to address future issues. The concept of "foresight" emerged in the realm of science and technology throughout the 1980s [5]. The term "technology foresight" became widely recognized in the 1990s as a means of creating new policy tools to tackle issues within science, technology, and innovation frameworks [6].

Despite the wide array of available futures methods that could potentially benefit medicine and health care, very few of these attempts have received notable attention from the medical community. Some of these have been performed by consulting companies and agencies such as Deloitte and KPMG [7,8]. Others have been published by institutions such as the World Health Organization [9] and the European Commission [10]. Another prominent example is the so-called Topol Review that outlined recommendations to ensure the United Kingdom's National Health Service is the world leader in using digital technologies to benefit patients [11]. It emphasized the implementation of technologies, such as genomics, digital medicine, artificial intelligence (AI), and robotics, in the future.

Given the recognition of such examples that might be limited in scope and impact, it becomes imperative to further explore the array of futures methodologies [1,12]. This exploration not only highlights the diverse techniques at our disposal but also underscores the untapped potential these methodologies hold in transforming health care planning and policy-making. Our scoping review reveals that the prevailing futures methods employed in the field of medicine and health care encompass trend analysis, Delphi, backcasting, policy analysis, technology assessment, horizon scanning, the Futures Wheel, and scenario analysis (Table 1). A concise overview of these methods is presented below.

XSL•FO
RenderX

Table 1. Comparison of the various foresight methods used in health care.

Foresight method	Description	Key characteristics	Common applications in health care
Trend analysis	Identifying patterns or trends in data over time	Data-driven, quantitative, predictive	Market trends, patient demographics
Delphi	Gathering expert opinions through iterative surveys to reach a consensus	Expert-driven, qualitative, consensus-building	Health care policy, clinical guidelines
Backcasting	Starting with a desired future and working backwards to the present	Goal-oriented, strategic, creative	Long-term health planning, sustainability initiatives
Policy analysis	Evaluating the effectiveness of policies and their potential impacts	Analytical, evaluative, policy-focused	Public health policies, regulatory changes
Technology assessment	Assessing the potential impacts of new technologies	Technology-focused, impact analysis	Adoption of medical technologies, risk assessment
Horizon scanning	Identifying emerging trends, threats, and opportunities	Forward-looking, broad-scope, proactive	Emerging health technologies, risk management
Futures Wheel	Analyzing the consequences of a particular event or statement	Brainstorming exercise	Direct and indirect impacts of technological, policy, or procedural changes
Scenario analysis	Exploring and evaluating possible future scenarios	Exploratory, strategic, scenario-building	Strategic planning, crisis management

Trend analysis refers to the prevailing directions or patterns of change and entails assumed developments in the future with a long-lasting effect and impact on a given field [13]. Trends often emerge from advancements in technology, shifts in patient needs and demographics, evolving health care policies, and new medical research findings. Such trends can be identified through data analysis, expert forecasts, and studying historical patterns in medicine and health care to help shape future strategies and provide insights into potential developments. Trend analysis is often applied to forecasting that is about generating predictions about the outcome of a specific question such as “How many mRNA-based vaccines will have been approved by the US Food and Drug Administration (FDA) by the end of 2030?”

The Delphi method was originally developed as a systematic interactive forecasting method, which relies on a panel of experts [14]. The use of Delphi is prevalent across health sciences research, and it is used to identify priorities, reach consensus on issues of importance, and establish clinical guidelines.

Backcasting is a method that starts with defining one or more preferred futures and then works backwards to identify strategies, policies, projects, or actions that will connect that specified future to the present [15]. Hence, it creates a roadmap toward the imagined future, such as how to effectively provide and organize future health care services.

Policy analysis examines problems calling for policy response and then proceeds to determine and assess possible options for policy action. Studying alternative futures and policy scenarios facilitates the better understanding of how present decisions affect the future [16]. It can indicate what measures improve the effectiveness and sustainability of health systems, or what priorities should be set for health-related research and development.

Technology assessment aims to examine the impacts of new technologies on society, including the capacity to be prepared for what may happen or be needed in the future. It anticipates the possible effects of technology-related evolution by collecting insights into the concerns of various stakeholders [17]. It offers a beneficial approach for health care manufacturers and decision makers to evaluate innovative health technologies prior to their launch. Considering that the successful launch of innovations in health care is substantially time and resource intensive, foresight-based strategic and future-oriented thinking is essential.

Horizon scanning provides strategic foresight through identifying emerging trends, technologies, and potential challenges in health care to anticipate future developments [18]. It involves gathering and analyzing data from a wide range of sources to spot early signs of significant change, also called “weak signals,” and evaluating the implications of these changes for health care systems and policies.

The Futures Wheel is a tool designed to help explore the potential consequences of a particular change or event. This structured brainstorming method, created by futurist Jerome C Glenn in 1971, visually maps out direct and indirect impacts, promoting a deeper understanding of the ripple effects that can arise from a single decision or trend.

Finally, scenario analysis involves creating and exploring multiple plausible future scenarios to understand the potential impacts of various trends, changes, and uncertainties in health care [19]. This could be useful in determining possible risks and opportunities as well as evaluating how resilient the current strategies are to various future circumstances.

Acknowledging the potential of using foresight in medicine and health care and seeing the fragmented attempts at exploiting the even higher potential of other futures methods, we aimed to explore this notion by undertaking the first scoping review to date about the application of futures methodologies and published foresight projects in health care.

XSL•FO
RenderX

Methods

The first author (BM) and second author (TK) of this article performed a scoping review using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) method [20]. The completed PRISMA-ScR checklist is provided in Multimedia Appendix 1. To avoid missing relevant papers, we used a search strategy covering 3 sources. The first included a broad search strategy focusing on futures methods and foresight on PubMed. The second source expanded that search to nonbiomedical journals by using Crossref. For the third source, we used the same search strategy in the 5 most influential (Scopus and Web of Science indexed) journals in the futures field, namely Technological Forecasting and Social Change, Futures, European Journal of Futures Research, Foresight, and Foresight and STI Governance. It is important to note that for now, futures studies and foresight are niche fields with a relatively small community of researchers and a limited number of influential journals, in contrast to larger disciplines. We considered the "most influential" journals to be those that are categorized as Scopus Q1 and are also indexed in Web of Science. The journals must focus specifically on futures studies or foresight. The sequence of mentioning the journals was driven by their latest available Scimago Journal Rank (SJR) index.

We focused on futures methods, studies, and foresight in the medical or health care–related context. As an initial filtering step, we decided to focus on select journal articles only. Articles not published in the English language were also excluded. Using the search query of (medicine OR medical OR health OR healthcare) AND (foresight OR futures studies OR futures methods OR futurology OR futurism), our initial search identified 448 publications in PubMed and 1301 publications in Crossref. Subsequently, to mitigate the risk of overlooking potentially pertinent future-oriented sources, we proceeded to employ an identical search query on the websites of the 5 aforementioned futures-related journals, resulting in 31 more records.

Throughout the subsequent screening process, Crossref hits were ranked in Google Scholar, and the first 100 items were selected for further analysis. The second author (TK) evaluated this list, conforming to the below outlined inclusion and exclusion criteria. Ultimately, 18 articles were selected from Crossref.

For PubMed, 403 records were excluded from the original list, and 45 records were checked for eligibility. Of these, 20 more were excluded. Therefore, 25 records were selected from PubMed.

Subsequently, the first author (BM) conducted a search within the previously mentioned 5 futures-related journals using the same keywords. Four relevant items were shortlisted from Technological Forecasting and Social Change, 4 from Foresight, 3 from the European Journal of Futures Research, 2 from Futures, and 1 from Foresight and STI Governance. To prevent duplication, articles already identified from Crossref were not selected again via journal search. Overall, 16 articles were selected for the final analysis from this search process.

We applied a range of inclusion and exclusion criteria to filter the literature. Articles were included if they met these criteria: (1) focus on foresight in medicine or health care, (2) original studies or reviews, and (3) use of recognized futures methodologies such as trend analysis, Delphi, backcasting, policy analysis, technology assessment, horizon scanning, or scenario analysis.

Articles were excluded if (1) the search terms were mentioned tangentially or in a context not related to futures in the fields of medicine and health care (eg, visual foresight in optometry); (2) the studies were not directly focused on futurist perspectives, even if they mentioned "foresight;" (3) the papers were not primary research articles or reviews, such as editorials, commentaries, book reviews, or news pieces; (4) the papers did not use or address recognized futures methodologies or the methodology was unclear or not well-defined; (5) the papers were from non-peer-reviewed journals as these might not meet the rigorous standards set by more reputable journals; and (6) the papers did not have at least an English abstract.

We ended up including 59 articles in our review. Table 2 describes these articles briefly, while the extended description can be found in Multimedia Appendix 2. The PRISMA-ScR flow diagram illustrating this process is presented in Figure 1.

XSL•FO
RenderX

Table 2. Identified and analyzed articles.

Theme	Title	Author	Journal	Year	Methods/design
Health care policy and workforce dynamics	Management challenges for future digitalization of healthcare services	Gjellebæk et al [21]	Futures	2020	Scenarios, backcasting
Health care policy and workforce dynamics	Drivers, trends and scenarios for the future of health in Europe. Impressions from the FRESHER project	Wepner et al [22]	European Journal of Futures Research	2018	Scenarios, trend analysis
Health care policy and workforce dynamics	Rethinking health workforce planning: Capturing health system social and power interactions through actor analysis	Rees et al [23]	Futures	2018	Document analysis, semistructured interviews
Health care policy and workforce dynamics	Futures-oriented drugs policy research: Events, trends, and speculating on what might become	Rhodes et al [24]	International Journal of Drug Policy	2021	Speculative research/big event, mega trend
Health care policy and workforce dynamics	Looking at the fringes of MedTech innovation: a mapping review of horizon scanning and foresight methods	Gonzalez-Moral et al [25]	BMJ Open	2023	Horizon scanning, meta-analysis
Health care policy and workforce dynamics	An overview of future EU health systems. An insight into governance, primary care, data collection and citizens' participation	Quaglio et al [26]	Journal of Public Health	2018	Workshops, STI[a] policy analysis
National strategies	Integrated foresight for the healthcare sector in Turkey	Kalan et al [27]	International Journal of Foresight and Innovation Policy	2007	Delphi, systems dynamics modeling
National strategies	The knowledge triangle in the healthcare sector — The case of three medical faculties in Norway	Brorstad Borlaug et al [28]	Foresight and STI Governance	2018	STI policy analysis
National strategies	Healthcare 2025: alternative views and underlying values	Rowley [29]	Foresight	2003	Scenarios
National strategies	Foresight in medicine. Lessons from three European Delphi studies	Wild et al [30]	European Journal of Public Health	2000	Delphi, meta-analysis
National strategies	The Dutch Public Health Foresight Study 2018: an example of a comprehensive foresight exercise	Verschuuren et al [31]	European Journal of Public Health	2019	Scenarios
National strategies	Forecasting Japanese health futures with the BFT	Watanabe et al [32]	Futures	1995	Forecasting
National strategies	The Futures Wheel model of the effects of the emerging infectious diseases pandemics on the elderly in Iran	Eidgahi et al [33]	Salmand	2024	Futures Wheel, expert panel, STEEP[b] analysis
National strategies	Possible future scenarios of the general health social security system in Colombia for the year 2033	Aguilar et al [34]	European Journal of Futures Research	2023	Scenarios, in-depth interviews, system dynamics modeling
National strategies	Health-related R&D priorities until 2030: Russian experience	Saygitov [35]	Foresight	2017	Trend analysis, STI policy analysis
National strategies	Perceptions of future hospital management in Finland	Pihlainen et al [36]	Foresight	2019	Delphi, inductive content analysis
National strategies	Using foresight to explore the impacts of flooding in Houston on health, poverty, and equity out to 2050	Daniels et al [37]	Futures	2021	Delphi, cross-impact analysis, interviews
National strategies	Health scenarios and policy making: Lessons from the Netherlands	Schreuder [38]	Futures	1995	Scenarios
National strategies	Trend Impact Analysis (TIA) of community-based futures study for pediatric obesity in Iran	Taghizadeh et al [39]	BMC Pediatrics	2023	Trend impact analysis, Delphi
National strategies	The effect of coronavirus (COVID-19) pandemic on medical sciences education in Iran	Rezaei et al [40]	Journal of Education and Health Promotion	2021	Futures Wheel

XSL•FO
RenderX

Theme	Title	Author	Journal	Year	Methods/design
National strategies	The future of AIDS in Africa: lessons from two scenario projects	Fourie [41]	African Journal of AIDS Research	2009	Scenarios
National strategies	The impact of future world events on Iranians' social health: A qualitative futurology	Damari et al [42]	Iranian Journal of Public Health	2016	Cross-impact analysis, deep interviews
National strategies	Key factors in the future of oral and dental health in Iran using scenario writing approach	Mehrolhassani et al [43]	BMC Oral Health	2024	Scenarios, focus groups, cross-impact analysis
National strategies	Exploring the driving forces and scenario analysis for catastrophic and impoverishing health expenditures in Iran	Hedayati et al [44]	BMC Health Services Research	2024	Scenarios, systematic review
National strategies	Medical technology decisions in The Netherlands: How to solve the dilemma of technology foresight versus market research?	Postma et al [45]	Technological Forecasting and Social Change	2007	Delphi, scenarios, technological forecasting
Pandemic preparedness and response	How to anticipate and control possible future health crises through foresight approaches using the COVID-19 pandemic crisis?	Didier et al [46]	Archives of Community Medicine and Public Health	2023	Scenarios, VUCA[c]
Pandemic preparedness and response	Foresight in the time of COVID-19	Gariboldi et al [47]	The Lancet Regional Health – Western Pacific	2021	Think tank workshops, scenarios, backcasting
Pandemic preparedness and response	Preparedness for emerging infectious diseases: pathways from anticipation to action	Brookes et al [48]	Epidemiology and Infection	2015	Environmental scanning, horizon scanning, multi-criteria decision analysis, simulation modeling
Pandemic preparedness and response	Steering vaccinomics innovations with anticipatory governance and participatory foresight	Ozdemir et al [49]	OMICS: A Journal of Integrative Biology	2011	Anticipatory governance, participatory technology design
Pandemic preparedness and response	The importance of uncertainty in health scenarios: A scoping review on COVID-19 scenarios	Hosseini Golkar et al [50]	Medical Journal of the Islamic Republic of Iran	2023	Scenarios, systematic review
Pandemic preparedness and response	COVID-19 and the politics of hope: A comparative analysis of Greek and Ecuadorian letters from a desired post-pandemic future	Gross et al [51]	Futures	2023	Letters from the future, visioning, backcasting
Pandemic preparedness and response	The value of mass-produced COVID-19 scenarios: A quality evaluation of development processes and scenario content	Crawford et al [52]	Technological Forecasting and Social Change	2022	Scenarios, meta-analysis
Specialized medical domains	The impact of foresight studies on human healthcare in the post-genomic era	Damrongchai et al [53]	International Journal of Foresight and Innovation Policy	2010	Scenarios, Delphi
Specialized medical domains	Delphi in a future scenario study on mental health and mental health care	Bijl [54]	Futures	1992	Delphi
Specialized medical domains	Innovative application of strategic foresight to oncology research	Bishop et al [55]	Foresight	2020	Scenarios, framework foresight
Specialized medical domains	Future of bioprinted tissues and organs: A two-wave global survey	Mota et al [56]	Foresight and STI Governance	2022	Surveys
Specialized medical domains	Leveraging strategic foresight to advance worker safety, health, and well-being	Streit et al [57]	International Journal of Environmental Research and Public Health	2021	Three-horizon foresight (weak signals, seeds of change, trend analysis), scenarios
Specialized medical domains	Four futures for occupational safety and health	Felknor et al [58]	International Journal of Environmental Research and Public Health	2023	Scenario-based framework foresight

XSL•FO
RenderX

Theme	Title	Author	Journal	Year	Methods/design
Specialized medical domains	How will the future of work shape the OSH professional of the future? A workshop summary	Felknor et al [59]	International Journal of Environmental Research and Public Health	2020	Workshops
Specialized medical domains	ASHP Foundation Pharmacy Forecast 2018: Strategic planning advice for pharmacy departments in hospitals and health systems	Vermeulen et al [60]	American Journal of Health-System Pharmacy	2018	Surveys, collective wisdom of "wise crowds"
Specialized medical domains	Personalized medicine beyond genomics: alternative futures in big data-proteomics, environtome and the social proteome	Özdemir et al [61]	Journal of Neural Transmission	2017	Technology foresight analysis
Specialized medical domains	Synergy between competitive intelligence (CI), knowledge management (KM) and technological foresight (TF) as a strategic model of prospecting--the use of biotechnology in the development of drugs against breast cancer	Canongia [62]	Biotechnology Advances	2007	Competitive intelligence, technology foresight
Specialized medical domains	The future of ICT for health and ageing: Unveiling ethical and social issues through horizon scanning foresight	Flick et al [63]	Technological Forecasting and Social Change	2020	Horizon scanning

[a]STI: science, technology, and innovation.

[b]STEEP: social, technological, economic, environmental, and political.

[c]VUCA: volatility, uncertainty, complexity, and ambiguity.

The following data were extracted from each included study: title, authors, journal, year of publication, summary, methods/study design, whether the study is available in open access, and access URL. These data were entered into a Microsoft Excel spreadsheet for analysis. The corresponding author (BM) performed the initial analysis, and the second author (TK) cross-checked it.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flowchart of the study selection process. The flowchart illustrates the study selection process for the scoping review. Initially, 1780 records were identified through database searches. After removing duplicates, title and abstract screening excluded 465 records, leaving 114 records for full-text assessment. Of these, 55 were excluded for not meeting the inclusion criteria, resulting in 59 studies for the final analysis. GS: Google Scholar.

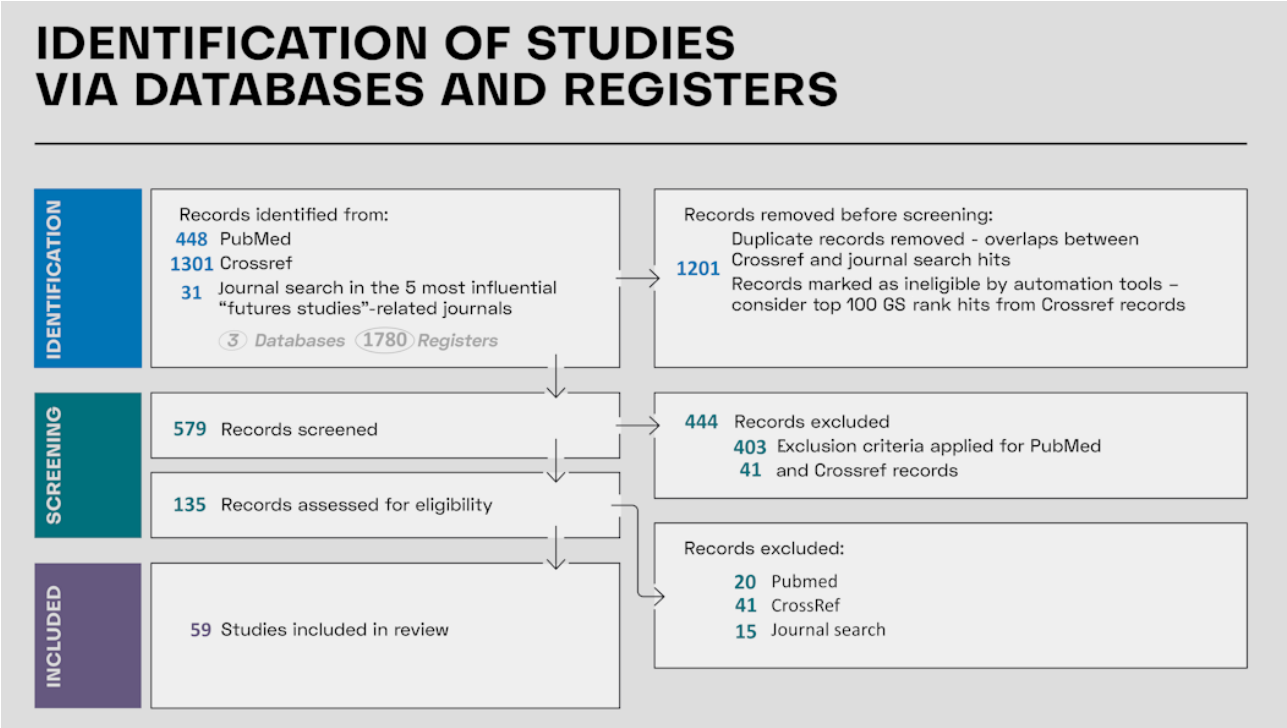

XSL•FO
RenderX

Results

Themes

We identified the following 5 themes and assigned each selected study to one of those themes (Table 2; Multimedia Appendix 2):

1. National strategies (n=19): These studies focus on the unique health-related prospects, challenges, and trends within specific countries. They offer insights tailored to each nation's health care system, demographics, and sociocultural milieu.
2. Strategic health care foresight (n=15): These studies employ futures methodologies to provide high-level analyses of future medical scenarios, trends, and challenges. They are broad in their approach, often looking at overarching factors that may influence the medical field as a whole.
3. Health care policy and workforce dynamics (n=6): This theme encompasses studies that focus on the impact of governmental policies on health care as well as the dynamics of managing health care professionals. They explore strategies and potential reforms that could shape the future of medical governance and human resource management.
4. Pandemic preparedness and response (n=7): These studies address issues related to infectious diseases, with a particular emphasis on vaccination strategies and the challenges posed by the COVID-19 pandemic. They discuss potential future scenarios, challenges, and strategies related to managing and preventing infectious disease outbreaks.
5. Specialized medical domains (n=12): This theme centers on specific areas or disciplines within medicine, such as personalized medicine, oncology, or biotechnology. Papers under this category delve deeply into their respective niche, highlighting future challenges and opportunities.

Theme 1: National Strategies

In this section, we present key findings from a selection of studies that illuminate the integration of futures methodologies in national health care strategies. These studies collectively highlight the growing recognition of foresight as a useful tool in shaping the future of health care policy and practice.

Iran clearly stands out in terms of foresight and futures studies as 6 studies focused on Iranian perspectives. Rezaei et al [40] used the Futures Wheel method to identify the effects of COVID-19 on medical sciences education. Nejatzadehgan Eidgahi et al [33] used the same technique to understand the different effects of emerging infectious disease pandemics on elderly people based on the experiences from the COVID-19 pandemic. Damari et al [42] performed a qualitative study of futurology with cross-impact analysis to find potential domains that may have an impact on Iranians' social health. Mehrolhassani et al [43] conducted a study in 3 stages, including a scenario writing approach, qualitative methods, and exploratory future research, to analyze the future of oral and dental health. Moreover, Hedayati et al [44] conducted a literature review and obtained experts' opinions using the Scenario Wizard software to provide insights into potential future scenarios of health expenditures in the country. Taghizadeh et al [39] conducted a community-based futures study using trend impact analysis on pediatric obesity, accentuating the effectiveness of foresight in policy evaluation, localization, and communication for obesity prevention.

Kalan et al [27] introduced an integrated foresight process within Turkey's health care sector, exemplifying the role of foresight in creating reflexive knowledge, fostering futures literacy, and developing strategic plans. The study provided evidence to promote futures literacy that helped define the causal relationships in the sector to elaborate and accomplish strategic action plans. At the same time, Postma et al [45] proposed a similar approach involving a method merging expert opinion forecasts with market-oriented scenarios in a Dutch case study on cancer imaging techniques, addressing the dilemma of short-term market perspectives versus long-term technology foresight.

Interdisciplinary approaches were highlighted in 2 studies. Verschuuren et al [31] presented the Dutch Public Health Foresight Study 2018, emphasizing a multidisciplinary participatory approach to explore the future of public health and health care in the Netherlands. It combined quantitative scenarios with qualitative thematic studies and participatory stakeholder consultations. The results indicated that well-established decision-making for the future requires an interdisciplinary approach, which is essential in completing complex foresight projects. Similarly, Daniels et al [37] employed foresight to examine the impact of severe flooding on health, poverty, and equity in Houston. They presented plausible scenarios to engage stakeholders and policymakers in shaping an inclusive future for the city's residents. The study concluded that foresight is a critical competence that equips practitioners with the ability to anticipate the implications of decisions.

Two papers focused on specific communities. Saygitov [35] provided insights into Russian efforts to establish research priorities in the health sector and contribute to health-related research and development and innovation policies through the use of trend analysis and expert workshops. UNAIDS and the South African financial services group Metropolitan published a set of scenarios regarding the future impact of HIV/AIDS in Africa and South Africa, respectively [41]. The paper highlighted the many lessons to be gleaned for HIV-related health planning and policy-making in general.

Wild et al [30] examined 3 technology foresight exercises (Delphi studies) focusing on health care innovations, drawing trends from German and British Delphi approaches and a more focused Austrian Delphi. The paper concluded that health policy must not only respond to these innovations but also proactively guide them, emphasizing the need for cooperation and coordination among all stakeholders. Hence, it is an important purpose and responsibility of futures studies, including in medical foresight, to enhance democratic participation, as participating actors have a key role in shaping the future. Pihlainen et al [36] explored Finnish experts' perceptions of hospital management and leadership for the year 2030, using a 3-round Argument Delphi process. The results of the foresight exercise contributed to the development of management and

XSL•FO
RenderX

leadership training and future-oriented management practices in hospitals.

Value-driven scenarios dominated the rest of the papers in this theme. Brorstad Borlaug et al [28] explored the significance of education in the Norwegian health care sector, highlighting the interplay between education, research, and innovation, and its implications for national health care policy. Rowley [29] employed scenarios to initiate public dialogue on health care values in the United States, highlighting the role of values in shaping diverse health care futures and providing a framework for evaluating value-driven components. It is also revealed that futures methods nowadays are able to evaluate values in an objective and future-oriented manner. Aguilar et al [34] formulated future scenarios for the Colombian General Health and Social Security System in 2033, promoting anticipatory governance and participatory policy development.

Another paper shared lessons of the scenario projects the Dutch government's Steering Committee on Futures Health Scenarios organized and carried out to facilitate research and debate on alternative futures in public health [38].

Another example demonstrated the power of forecasting. Japan's Institute of Health Systems Development created Bioforecasting Technology that involves integration of data regarding the population, the natural and social environment, and the health, medical, or social services with projections on the future population's demand for different kinds of health services. They claim that its forecasts have proved to be highly accurate [32].

Theme 2: Strategic Health Care Foresight

Studies that explore the strategic application of futures methodologies in the health care industry are presented in this section, with an emphasis on a range of topics such as ethics, innovation, sustainability, and participatory approaches.

Hemmat et al [64] reviewed and compared research on the future of health IT. The study concluded that while forecasting is suitable for smaller objectives, foresight approaches are better for achieving larger-scale goals. It helps establish reasonable long-term expectations for improvements in health IT. Furthermore, the use of mixed methods in foresight projects can better illuminate the necessary actions to accomplish preferred futures, whereas forecasting was found to be more theoretical than practical.

Mittelstadt et al [65] conducted a systematic meta-analysis of the academic literature to understand the ethical implications of "big data," especially in the biomedical field. It identified 5 primary areas of ethical concern and highlighted 6 additional emerging areas needing attention. This framework aims to guide future ethical assessments and governance of big data practices, especially in the very sensitive context of biomedical research.

With respect to upcoming innovations in health, Butter et al [66] evaluated the degree of alignment between the EU Framework Program (FP7) and the foresight community. Using the Dynamo approach, the study analyzed 140 flagship foresight activities and offered insights for the foresight community to prioritize unaddressed health innovation themes.

Masum et al [67] used compelling cases to highlight the usefulness of 5 futures methods (forecasting, scenario planning, Delphi, technology road mapping, and mass collaboration) in medicine and global health decision-making. It provided evidence that applying futures methods to health can help prepare for the future, including policy development and health initiatives.

Kulkova et al [68] focused on identifying key and atypical factors influencing the development of medicine, particularly in light of the challenges posed by COVID-19. The study, which employed the Delphi method, identified well-being technologies, data-informed personalization, and climate change as the 3 main drivers for the next 1 to 50 years. The results point out the processes that will influence the industry's transformation and suggest that stakeholders take appropriate measures to respond organically to events.

From a multi-stakeholder standpoint, Pereno et al [69] addressed the future of sustainable health care with the goal of defining long-term scenarios and key strategies for transitioning to sustainable health systems. It involved a collaborative foresight process through workshops with stakeholders from the Nordic countries, including health industries, providers, managing authorities, universities, research centers, nongovernmental organizations, and health care networks. The study identified 3 future horizons and drivers for reshaping stakeholder roles, underlining sociotechnical transition toward collaborative, distributed dynamic networks.

In order to envision health care futures in 2040 and beyond, Pau et al [70] investigated a methodology for participatory futures studies involving nonexperts, particularly in contexts with extreme unknowns. Rather than focusing on opinion leaders or scenario writers, the article highlights the role of futurists and foresight practitioners as facilitators.

Lamé et al [71] examined the use of scenario planning in the health care sector, focusing on a project at a large hospital's cancer division. The authors proposed directions for future scenario planning in hospital settings and for research on matching scenario approaches with specific application contexts.

Using science, technology, and knowledge indicators, Antunes et al [72] presented a methodology for technological foresight and scanning with a focus on health care. The approach incorporates indicators involving patents (technology), research and development (science), and competencies (human resources) through international databases and a data-and text-mining tool to create knowledge maps for decision-making support.

Karlsen et al [73] reported on 20 experts envisioning the future of addictions and lifestyles in Europe by 2030 and beyond. The project used a combined approach, blending in-situ idea generation with established futures tools, which culminated in a 2-day electronic scenario workshop. Using backcasting, the experts proposed an optimistic outlook for Europe's future addictions and lifestyles, focusing on collective values, long-term planning, and restitutive solutions to reframe current drug disorder policies.

Five articles focused on a more high-level approach about using futures methods in health care. Bezold [74] described a special

XSL•FO
RenderX

issue of Futures, reflected on the significant growth of health futures, and argued that the future of health futures will be shaped by the ability of futures techniques and those who use them to provide value to organizations and individuals. Pau et al [70] described the importance of long-term thinking in health care through presenting an 18-month design research project that developed a methodology for participatory futures studies with nonexperts. Haghdoost et al [75] contemplated about the emergence of futures methods in health care and concluded that there is a lack of a unique and dominant discourse in health sciences around those methods. Hosseini Golkar et al [76] analyzed the usefulness and importance of scenario planning and health scenario writing for short- and medium-term futures. Finally, Doos et al [77] performed a systematic search to identify the methods used in forecasting studies to predict future health technologies within 2 decades.

Theme 3: Health Care Policy and Workforce Dynamics

This theme's studies explored the dynamics of health care policy and employed futures methodologies to prepare the workforce for the future.

Rhodes et al [24] discussed the need for a more speculative approach in drug policy research, moving beyond predictable, probability-based analyses to embrace possibilities. The authors outlined the use of "Big Event" and "Mega Trend" as tools for speculative intervention, helping to trace and project complex factors influencing drug futures. The aim of this approach was to influence the present about what futures could be.

In order to identify horizon scanning and other approaches used for medical technology foresight in health care decision-making, Garcia Gonzalez-Moral et al [25] conducted a review. It revealed a diverse range of methods, highlighting the need for greater transparency and consistency in methodology reporting. The results shed light on the fact that available methods used in combination can overcome the limitations posed by single methods and no best practice exists in this regard.

Quaglio et al [26] summarized the outcomes from a European Parliament workshop titled "Health systems for the future," where experts and decision-makers discussed enhancing the effectiveness and sustainability of EU health systems. The paper concluded that in order to equip EU health systems to handle upcoming challenges, leadership in policy-making will be essential.

Gjellebæk et al [21] explored middle management strategies for promoting workplace learning in the context of introducing eHealth in health care. The study established the need for a shift toward learning-oriented leadership and adaptive management, focusing on employee involvement and continuous skill development. Scenario planning and backcasting were found to be effective tools for navigating the complexities of changing work practices.

Rees et al [23] demonstrated the importance of actor analysis for identifying strategic positions on health workforce planning, acknowledging the need for accommodating changes in health care roles and environments. The study employed a mixed methods design, combining document analysis, semistructured interviews from 2 health subsectors, and numerical data transformation for actor analysis software.

Wepner et al [22] considered stakeholders from a range of sectors as they investigated alternative policy approaches for noncommunicable diseases in OECD (Organization for Economic Co-operation and Development) countries. The study analyzed trends to develop 4 scenarios of possible futures. The article concluded that out of the box thinking is needed to consider the complexity of future health systems that need to include aspects like equity, literacy, mobility, or urban planning. It makes the case for a systematic and holistic approach to address all drivers and determinants leading to a healthy life and well-being.

Theme 4: Pandemic Preparedness and Response

This theme highlights studies that focused on the application of futures methodologies in pandemic preparedness and response.

The application of futures methodologies in the Western Pacific region to enhance public health emergency response, specifically for the COVID-19 pandemic, was detailed by Gariboldi et al [47]. They used agile think tank sprints, a 6-step foresight process, and backcasting to develop pandemic scenarios over an 18-month horizon, resulting in tailored recommendations for World Health Organization response and country support, and demonstrating foresight's value in public health emergencies. Through the creation of diverse future scenarios that considered not only predictable trends but also areas of uncertainty and divergence, potential shocks to the system, and complex interactions, a more future-proof strategy could be devised.

In their review, Brookes et al [48] suggested integrating tools for identifying, prioritizing, and investigating emerging and re-emerging infectious diseases into a comprehensive framework to improve tactical and strategic planning for emerging risk preparedness. Techniques included environmental scanning, foresight programs, horizon scanning, surveillance, transparent prioritization methods, risk assessment, and simulation modeling. The results suggested that the coordinated application of multiple foresight tools provides a greater overall benefit than individual tools applied in an ad hoc manner.

Since anticipatory governance is becoming increasingly crucial for responsible innovation and scientific practice, particularly in the uncertain and high-stakes field of vaccinomics, Ozdemir et al [49] argued in favor of its application with foresight. This article corroborated one of the core tenets of futures studies, namely that uncertainty is not an accident of the scientific method but is its very substance. Anticipatory governance with participatory foresight offers a mechanism to respond to such inherent sociotechnical uncertainties in the field of vaccinomics by reflective generation of futures literacy.

The evolution of prospective studies toward an integrated, diachronic approach to envisioning the future, especially in uncertain situations like the COVID-19 crisis, was analyzed by Didier et al [46]. It highlighted foresight as a discipline that explores possible futures and informs present choices to shape desirable futures and strategize against unwanted scenarios,

XSL•FO
RenderX

particularly in health-related contexts like pandemics and epidemics. The core message of the article lies in the finding that foresight creates a space for a wide range of dialogues and facilitates thinking ahead about the actions that must be taken now to achieve the world we prefer in the future.

Crawford et al [52] evaluated 213 COVID-19 scenarios developed during the pandemic's first wave, addressing concerns about the value of these mass-produced scenarios. The study concluded that scenario interventions, particularly in crises, should be more proactive and inclusive of stakeholders, decision-makers, and affected communities to ensure effectiveness and relevance.

Gross et al [51] conducted a unique experiment in which they invited people living in countries in Europe, North and South America, Asia, and Africa to imagine what the future could look like, beyond the COVID-19 pandemic, by writing a letter from the future. In their study, they analyzed those letters to reveal a particular relationship between the present and the future: a relationship of hope.

Hosseini Golkar et al [50] examined the phenomenon of uncertainties regarding the COVID-19 pandemic in relation to scenario planning. They made recommendations about improving the accuracy of scenario planning.

Theme 5: Specialized Medical Domains

This theme features studies demonstrating the use of futures methodologies in specialized medical domains, including occupational safety and health (OSH), pharmacy practice, psychiatry, or breast cancer drug development.

Streit et al [57] examined the use of strategic foresight as a tool to address the evolving factors in OSH related to the workforce. The study reviewed work-related future scenarios, presenting a tailored foresight framework for OSH. The article paved the way for the integration of strategic foresight into OSH research and practice to enhance worker safety, health, and well-being. Generating the implications of plausible work-related futures can help OSH prepare for, plan, and influence the future by combating human tendencies to over- and underpredict change, which are critical errors in decision-making.

Felknor et al [58] detailed the US National Institute for Occupational Safety and Health inaugural strategic foresight project aimed at enhancing the OSH response to rapid workplace changes. Using futures studies and strategic management, the project developed 4 alternative future scenarios for OSH through multidisciplinary expert collaboration. This study provided evidence that widely used corporate foresight approaches can be successfully adopted in other fields.

In another article, Felknor et al [59] summarized a workshop focusing on the future of work and its implications for OSH education and training. The workshop featured national experts and thought leaders in OSH, strategic foresight, systems thinking, and industry, addressing challenges and opportunities in both academic and industrial contexts. The main point is that a transdisciplinary perspective should be applied to incorporate multiple disciplines, professions, and technologies into OSH academic training.

The sixth edition of the Pharmacy Forecast report, which emphasizes its role in advancing pharmacy practice leadership, was presented by Vermeulen et al [60]. This edition benefitted from contributions by pharmacists, advisors, panelists, and chapter authors. It also offered insights into emerging trends in pharmacy and health systems, aiming to stimulate strategic planning rather than making precise predictions.

Özdemir et al [61] discussed the integration of big data technologies like proteomics in psychiatry, emphasizing its role as a next-generation biomarker and diagnostic tool. The article proposed a 3-tiered nested governance structure for big data science, involving scientists, ethicists, and scholars in ethics of ethics, to ensure equitable distribution of political power in innovation processes.

With a focus on biotechnology for breast cancer drugs, Canongia's paper [62] presented a strategic model that combines technology foresight, competitive intelligence, and knowledge management to support decision-making in sustainable development and innovation. The findings showed that technological foresight has a common base in the process of gathering strategic information, the implementation of analytical and reflective capacities, and the creation of futures literacy.

Damrongchai et al [53] evaluated the impact of foresight on policy, demonstrating its use through a case study on health care in the postgenomic era. The study highlighted foresight as a crucial tool for understanding the complexity of potential impacts, reconciling different perspectives, and forming a solid foundation for policies, with assessment of the project's impact against various future outcomes.

Horizon scanning was used by Flick et al [63] to explore the future of information and communication technologies (ICT) for health and aging, with a particular focus on the ethical and social issues in this sector. In order to discuss the potential and likely futures of ICT for health and aging, the study analyzed interviews, literature reviews, policy documents, and online public discussions to identify key and weak signals in areas such as future technologies, companies, older people, and environments.

With a focus on personalized medicine, Stelzer et al [78] introduced a novel method that combines bibliometrics and the scenario technique for technology foresight. The method involves an 8-step scenario approach, incorporating emerging technologies and their effects on each scenario, supported by co-citation analysis and a bibliographic coupling network. The findings show that by addressing each other's shortcomings, these 2 approaches together produce a helpful, all-around tool for conducting technology foresight.

Bishop et al [55] applied Framework Foresight, a method from the University of Houston, to explore the future of cancer and cancer research for strategic planning in a cancer research center. It identified 4 scenarios that highlighted strategic issues and offered guidance for not only cancer centers but also medical practice in general. This paper marks the first application of Framework Foresight to a medical topic. As the University of Houston is currently the world leading higher education institution in the field of futures studies, this case study

XSL•FO
RenderX

demonstrates fruitful cooperation among academic futurists, foresight practitioners, and stakeholders from the medical sector.

Experts in tissue engineering from around the world were surveyed in 2 stages by Mota et al [56] on the potential applications of 3D and 4D bioprinting in preclinical research and clinical settings. The study also proposed a practical tool for expert identification and interviewing, which is useful for foresight studies. The results revealed that the influence of technology on the health sector tends to further increase over time. Henceforth, preparing for the future is a necessity for not only those involved in research, clinical practice, or technology management but also those responsible for health care delivery and developing and implementing public health policies.

As the final example of applying futures methods to health care in this theme, Bijl [54] highlighted the benefits of the Delphi method for analyzing issues of futures research on mental health and its potential impact on policy-oriented futures research.

Discussion

Principal Findings

This investigation not only demonstrates the various futures methods available to the medical community but also emphasizes the unexplored possibilities these approaches possess in transforming health care planning and policy formulation.

The studies in the first theme collectively demonstrate how futures methodologies are becoming more widely acknowledged as crucial instruments in shaping national health care strategies and policies, with implications for future health care practices and innovations.

Studies under theme 2 highlight the strategic applications of futures methodologies in health care, ranging from ethics and innovation to sustainability and participatory approaches, enriching the field with diverse perspectives and insights.

Theme 3 addresses the complexities of preparing the health care workforce for future challenges. Studies under this theme offered insights into health system transformation, the training of medical professionals, actor analysis, and policy alternatives for future health care implications.

Studies under theme 4 underscored the significance of using futures methodologies in pandemic preparedness and response, offering insights into their application in public health emergencies, which is a timely subject owing to COVID-19's impact on health care.

The fifth theme demonstrated the wide-ranging applications of futures methodologies in specialized medical domains, from fostering innovation and strategic planning to helping health care professionals make well-informed decisions.

It is possible to conclude from the abovementioned narratives that "futures studies" has been used either sparingly in the health care industry, concentrating only on specialized areas and pandemic preparedness, or infrequently, addressing topics related to national strategies or the health care workforce. Although we managed to identify signs of various futures methods being used to anticipate health care futures, scenarios, and potential challenges, their theoretical and methodological underpinnings are not yet comprehensively developed.

In summary, our scoping review found that the use of futures methods has been successfully demonstrated in a wide range of fields from national strategies and policy-making to preparing for global threats and technological breakthroughs. Overall, 8 futures methods, including trend analysis, Delphi, backcasting, policy analysis, technology assessment, horizon scanning, the Futures Wheel, and scenario analysis, have already been used in medicine and health care, whereas over 50 futures methods exist. It may underscore the notion that the field is unexploited. These findings demonstrate that there are unmet potentials in the field, and therefore, we advocate for a broader interdisciplinary approach to future-oriented studies in medicine and health care.

We argue that these findings create an imperative need for the establishment of "medical futures studies" as a distinctive subdiscipline within the expansive domains of health care, medicine, and life sciences. This pioneering branch is envisioned to be committed to the systematic application of futures methodologies, with the primary objective of proactively anticipating and mitigating potential challenges, deliberating on policy considerations, and discerning emerging technological trends within the health care sector. This inclusive term could encompass foresight, futures methods, and practices, effectively bridging strategic planning with academic exploration.

This is a result of our realization that "futures studies" in this field is currently underutilized, which we consider to be a lost opportunity. We propose that the integration of futures studies has the potential to greatly enhance and advance medical science by filling a crucial gap in accomplishing well-founded anticipation, innovation, futures literacy, scenario formulation, visioning, and strategic planning.

We further elucidate why, among a multitude of fields, health care stands as the prime candidate for the incorporation of futures methods, and conversely, why futures methods are paramount for the evolution of health care (Figure 2).

XSL•FO
RenderX

Figure 2. A summary of the reasons why health care is the prime candidate for the incorporation of futures studies and why the use of futures methods would be paramount for the evolution of health care.

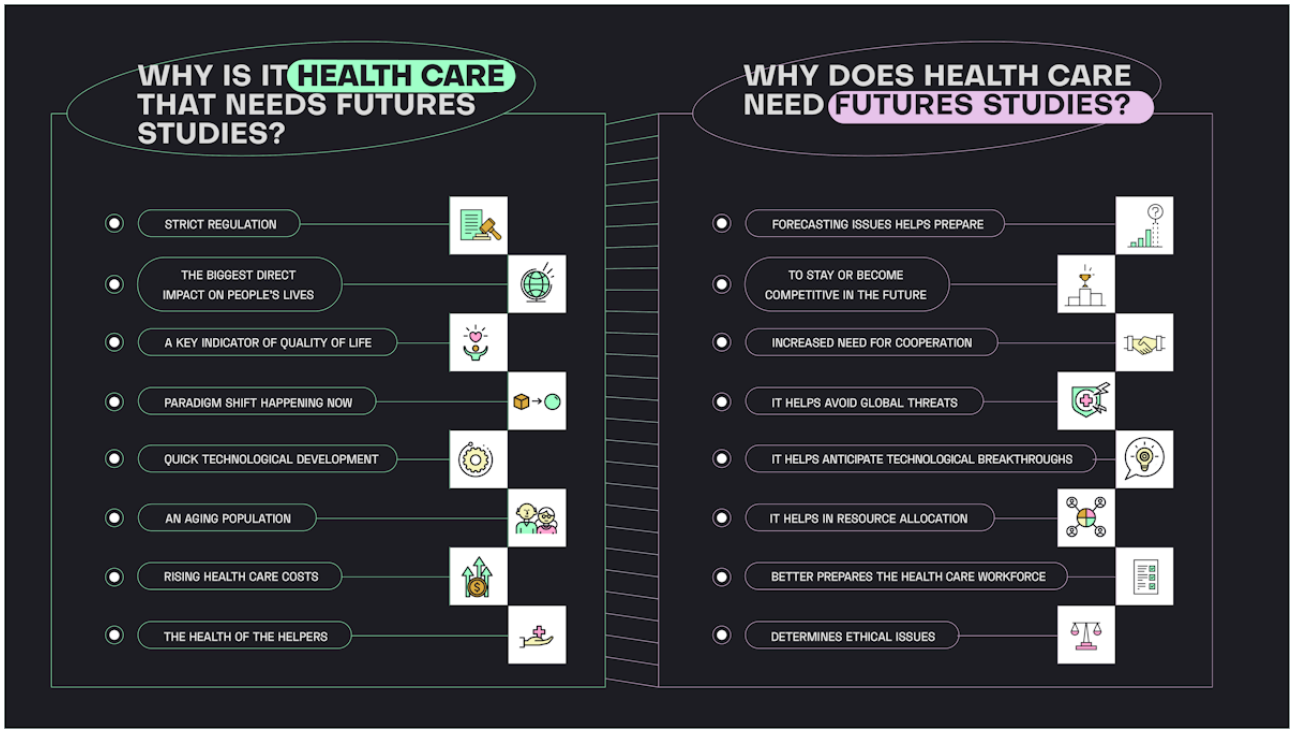

First, we elaborate on why *health care is the prime candidate for the incorporation of futures studies*. Since health care regulations are uniquely stricter than those in other fields, it would be beneficial for regulators to anticipate challenges and benefits of changes in science, technology, and culture. In this manner, innovations could reach the market as a result of prompt and suitable regulations. The US FDA regulating AI in the health care sector is an exemplary story to underscore this [79]. It should be highlighted that one of the most important purposes of futures studies is to develop reflexive futures literacy and create knowledge about the future in order to underpin policy alternatives and decisions in such a way as to shape the future and promote the realization of preferred futures.

Health care has the biggest direct impact on people's lives. From physical, mental, and emotional well-being and longevity, access to quality health care directly affects a person's lifespan [80]. Moreover, good health enables people to work, contribute to the economy, and pursue their aspirations. Besides these observations, effective health care systems monitor and control the spread of infectious diseases. It is naturally supported by futures studies as its general mission is to contribute to the creation of a better, more livable world for human beings within the framework of a forward-looking approach.

Health care is also a key indicator of quality of life. According to Britannica, it is "the degree an individual is healthy, comfortable, and able to take part in or enjoy life events" [81]. Without good health, individuals may find it challenging to work or pursue education. Good health, supported by quality health care, allows individuals to engage in social activities, foster relationships, and participate in community events.

Health care's biggest paradigm shift is happening in this decade [82]. Paradigm shifts have taken place in every industry. In the media, there has been a shift from print to digital platforms. In retail, there has been a shift from stores to e-commerce. In transportation, there has been a shift from gasoline to electric vehicles. Health care's paradigm shift called digital health, during which patients become the point-of-care and the newest member of the medical team, has been taking place during recent years.

Consequently, an incredibly quick technological development is happening in health care that also comes with a cultural revolution. Medicine is highly dependent on technology, and the pace of technological advancement in health care is staggering. Futures methods can help anticipate the impact of emerging technologies, such as AI, gene editing, and telemedicine, on health care delivery, patient outcomes, and the overall health care landscape.

As the global population ages and health care needs become more complex, there is a growing need for innovative approaches to address the demographic challenges associated with elderly people. Research with an eye toward the future can help in envisioning the health care needs of aging populations and developing strategies to meet those needs effectively.

Health care costs are a significant concern worldwide, leading to one of the biggest challenges in the 21st century. Future-oriented research can explore ways to make health care more efficient, cost-effective, and accessible.

The health of helpers is also an important aspect. It is crucial to underscore the documented negative impact on a patient's recovery when health care professionals are overworked or experiencing burnout. By conducting future-oriented research, we can explore innovative approaches to alleviate the burdens on health care professionals, promote their well-being, and ensure they can provide optimal care to patients. Moreover,

XSL•FO
RenderX

addressing health care worker burnout benefits not only patients but also health care practitioners themselves. It can make the health care profession more enjoyable, fulfilling, and effective. This can lead to increased job satisfaction, reduced turnover rates, and a more motivated and engaged health care workforce.

We also describe why *the use of futures methods would be paramount for the evolution of health care*.

Anticipating issues and challenges would be of value for health care to better consider the advantages and disadvantages of technologies. Forecasting has been used for that purpose, among other things, in economics but has not been widely used in health care. As much as traditions are comfortable and give a sense of security, the emergence of evidence-based medicine and the need for scientific rigor marked a significant paradigm shift in the realm of medicine. Information has emerged as an invaluable commodity, the cornerstone of research, and the basis for every treatment decision. However, as important as it is, it is anchored in the present cross-sections. The dynamic nature of health care, which involves societal changes, unprecedented technological advances, and global events, yearns for a forward-looking perspective that futures studies can offer.

Foresight exercises can reveal that the challenges facing health policy involve not only responding to technological innovations by offering a universal promotional environment, but also steering and guiding them in a desirable future direction at an early stage. It could help detect high-level trends for decision makers, such as patients, becoming the point-of-care through a digital health paradigm shift, the art of medicine arriving with the era of AI, or digital health being a cultural transformation [83]. Understanding high-level trends might put decision-makers in a better position to foresee challenges, risks, potentials, and issues that might not yet exist.

Futures studies could help avoid global threats or help prepare in time for global threats. The emergence of new diseases, epidemics, or pandemics poses a significant risk to global public health. Medical futures studies can help anticipate and prepare for these threats, allowing governments and organizations to develop prevention and response strategies that minimize the impact on public health. A Canadian AI technology providing alerts of a possible outbreak in the early days of COVID-19 before national public health authorities did is a prime example of that [84].

In addition, futures studies can help anticipate and prepare for breakthroughs in medical technologies, such as gene editing, personalized medicine, and AI-driven diagnostics. By doing so, the medical community can ensure the readiness of infrastructure, policies, and regulatory environments for seamless adoption of these innovations. Medical associations have been trying to benefit from that [85].

Health care professionals can be better prepared for the quickly changing medical landscape, stay on the cutting edge of knowledge and practice throughout their careers, and identify new critical skills like prompt engineering by incorporating futures studies into medical education [86]. Current medical curricula run the risk of becoming outdated and sticking to less effective and cumbersome practices. The ways students now consume and understand subjects have evolved, with short-format videos, engaging AI or augmented reality projections, virtual reality, and AI-driven patient simulations taking precedence over traditional textbooks and quizzes. The new curriculum should also familiarize students with the basics of digital health and address the growing needs of patients in the face of a decreasing health care workforce. The imminent ubiquity of AI in medical practice means that medical professionals unequipped with these tools risk lagging in their field. Futures studies can provide solutions to these challenges and clarify whether they relate to the evolving dynamic between patients and doctors or the need to equip medical professionals with the digital skills of the future.

Foresight exercises on health technologies can also facilitate envisioning technological developments in order to stay or become competitive in the future. This can be relevant to not only companies trying to contribute to the progress of medicine but also individuals working in and for health care.

Future health challenges are becoming complex and require intersectoral solutions. With an emphasis on broad stakeholder engagement and the integration of varied forms of evidence to undertake collective visioning, futures approaches can strengthen long-term capacities for ongoing participatory evidence-based decision-making.

Moreover, technological innovations emerge from the imagination and skills of individuals acting in an appropriate setting. Foresight exercises identify common grounds of experts from different fields with respect to their understanding of future developments in their profession on a broad base, including policy options. In the future, there will be an increased need for cooperation and coordination among all actors in health care.

Futures studies can also help health care providers and policy-makers to optimize resource allocation by forecasting the demand for health care services and identifying potential bottlenecks or shortages in care [87].

Lastly, applying foresight could aid in determining ethical issues. As advancements in medical technology raise ethical concerns, medical futures studies can help anticipate and address these issues proactively. By engaging in ethical debates early on, the medical community can ensure responsible development and use of new technologies.

An obvious follow-up to this scoping review could be performing systematic reviews of the abovementioned futures methods individually to determine and demonstrate their use cases, benefits, and limitations specifically in medicine and health care.

Limitations

This study has some limitations. We may have missed studies dedicated to foresight and futures methods as such scientific studies are spread across different types of journals ranging from biomedicine to business management. The nomenclature also varies as some studies use foresight, while others might refer to the same methods using other expressions such as futurology, futuristics, or futurism.

XSL•FO
RenderX

As many of the identified studies used different futures methods, it is difficult to compare them to each other. As we only focused on health care and the potential of futures studies to be used in that field, we might have missed other niche areas where futures research could also fill important gaps.

Conclusion

Based on the findings of this study, we conclude that there is an imperative need for the introduction of medical futures studies. In addition to helping policy makers and medical professionals better anticipate emerging technological trends, breakthroughs, challenges, and ethical considerations as medicine advances, a systematic and guideline-based application of futures methods could close significant gaps in health care. We argue that launching a dedicated scientific journal in medical futures studies to foster a new research community, appointing dedicated futurists in decision-making and national strategy, and teaching basic futures methods in the medical curriculum could provide further contribution.

In this study, we conclude that embracing medical futures studies in health care is essential for guiding advancements, informing policy, and shaping education, ultimately leading to a more prepared, innovative, and ethical medical landscape.

Data Availability

All data extracted from the analyzed studies are available in the Multimedia Appendix section.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 497 KB - jmir_v26i1e57148_app1.pdf]

Multimedia Appendix 2
Extended database of the identified and analyzed articles.
[XLSX File (Microsoft Excel File), 30 KB - jmir_v26i1e57148_app2.xlsx]

References

1. Glenn J, Gordon T. Futures Research Methodology — Version 3.0. The Millennium Project. 2009. URL: https://www.millennium-project.org/publications-2/futures-research-methodology-version-3-0/ [accessed 2024-09-24]
2. Kuosa T. Evolution of futures studies. Futures 2011 Apr;43(3):327-336 [FREE Full text] [doi: 10.1016/j.futures.2010.04.001]
3. Kristóf T, Nováky E. The story of futures studies: An interdisciplinary field rooted in social sciences. Social Sciences 2023 Mar 21;12(3):192. [doi: 10.3390/socsci12030192]
4. Motti V. Futures Studies. In: The Palgrave Encyclopedia of the Possible. Cham: Palgrave Macmillan; 2022:1-5.
5. Martin BR. The origins of the concept of ‘foresight’ in science and technology: An insider's perspective. Technological Forecasting and Social Change 2010 Nov;77(9):1438-1447 [FREE Full text] [doi: 10.1016/j.techfore.2010.06.009]
6. Miles I. The development of technology foresight: A review. Technological Forecasting and Social Change 2010 Nov;77(9):1448-1456 [FREE Full text] [doi: 10.1016/j.techfore.2010.07.016]
7. Health Care Foresight: Identifying megatrends. Deloitte. URL: https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/life-sciences-health-care/sg-lshc-healthcare-foresight-megatrends.pdf [accessed 2024-09-24]
8. Healthcare Foresight. KPMG. URL: https://kpmg.com/xx/en/what-we-do/industries/healthcare/healthcare-foresight.html [accessed 2024-09-24]
9. WHO Foresight: Monitoring emerging technologies and building futures-thinking. World Health Organization. URL: https://www.who.int/activities/who-foresight---monitoring-emerging-technologies-and-building-futures-thinking [accessed 2024-09-24]
10. Strategic Foresight Report 2023. Publications Office of the European Union. URL: https://commission.europa.eu/system/files/2023-07/SFR-23-beautified-version_en_0.pdf [accessed 2024-09-24]
11. The Topol Review: Preparing the healthcare workforce to deliver the digital future. The Topol Review, Health Education England, NHS. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2024-09-24]
12. Georghiou L. The Handbook of Technology Foresight. Cheltenham, United Kingdom: Edward Elgar Publishing; 2009.
13. Agami N, Omran A, Saleh M, El-Shishiny H. An enhanced approach for Trend Impact Analysis. Technological Forecasting and Social Change 2008 Nov;75(9):1439-1450 [FREE Full text] [doi: 10.1016/j.techfore.2008.03.006]
14. Zartha Sossa JW, Halal W, Hernandez Zarta R. Delphi method: analysis of rounds, stakeholder and statistical indicators. FS 2019 Sep 10;21(5):525-544. [doi: 10.1108/fs-11-2018-0095]
15. Dreborg K. Essence of backcasting. Futures 1996 Nov;28(9):813-828 [FREE Full text] [doi: 10.1016/s0016-3287(96)00044-4]

XSL•FO
RenderX

16. Wright D, Stahl B, Hatzakis T. Policy scenarios as an instrument for policymakers. Technological Forecasting and Social Change 2020 May;154:119972 [FREE Full text] [doi: 10.1016/j.techfore.2020.119972]
17. Bechtold U, Fuchs D, Gudowsky N. Imagining socio-technical futures – challenges and opportunities for technology assessment. Journal of Responsible Innovation 2017 Aug 31;4(2):85-99. [doi: 10.1080/23299460.2017.1364617]
18. Akartuna E, Johnson S, Thornton A. Enhancing the horizon scanning utility of futures-oriented systematic and scoping reviews. Futures 2024 Apr;158:103340 [FREE Full text] [doi: 10.1016/j.futures.2024.103340]
19. Spaniol M, Rowland N. The scenario planning paradox. Futures 2018 Jan;95:33-43 [FREE Full text] [doi: 10.1016/j.futures.2017.09.006]
20. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
21. Gjellebæk C, Svensson A, Bjørkquist C, Fladeby N, Grundén K. Management challenges for future digitalization of healthcare services. Futures 2020 Dec;124:102636 [FREE Full text] [doi: 10.1016/j.futures.2020.102636]
22. Wepner B, Giesecke S. Drivers, trends and scenarios for the future of health in Europe. Impressions from the FRESHER project. Eur J Futures Res 2018 Jan 10;6(1):2. [doi: 10.1007/s40309-017-0118-4]
23. Rees G, Crampton P, Gauld R, MacDonell S. Rethinking health workforce planning: Capturing health system social and power interactions through actor analysis. Futures 2018 May;99:16-27 [FREE Full text] [doi: 10.1016/j.futures.2018.03.009]
24. Rhodes T, Lancaster K. Futures-oriented drugs policy research: Events, trends, and speculating on what might become. Int J Drug Policy 2021 Aug;94:103332 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103332] [Medline: 34148724]
25. Garcia Gonzalez-Moral S, Beyer FR, Oyewole AO, Richmond C, Wainwright L, Craig D. Looking at the fringes of MedTech innovation: a mapping review of horizon scanning and foresight methods. BMJ Open 2023 Sep 14;13(9):e073730 [FREE Full text] [doi: 10.1136/bmjopen-2023-073730] [Medline: 37709340]
26. Quaglio G, Figueras J, Mantoan D, Dawood A, Karapiperis T, Costongs C, et al. An overview of future EU health systems. An insight into governance, primary care, data collection and citizens' participation. J Public Health (Oxf) 2018 Dec 01;40(4):891-898. [doi: 10.1093/pubmed/fdy054] [Medline: 29590431]
27. Kalan O, Alsan A. Integrated foresight for the healthcare sector in Turkey. IJFIP 2007;3(4):420. [doi: 10.1504/ijfip.2007.016468]
28. Brorstad Borlaug S, Aanstad S. The knowledge triangle in the healthcare sector — The case of three medical faculties in Norway. Foresight and STI Governance 2018 Mar 25;12(1):80-87 [FREE Full text] [doi: 10.17323/2500-2597.2018.1.80.87]
29. Rowley WR. Healthcare 2025: alternative views and underlying values. Foresight 2003;5(5):16-21 [FREE Full text] [doi: 10.1108/14636680310507271]
30. Wild C. Foresight in medicine. Lessons from three European Delphi studies. The European Journal of Public Health 2000 Jun 01;10(2):114-119. [doi: 10.1093/eurpub/10.2.114]
31. Verschuuren M, Hilderink H, Vonk R. The Dutch Public Health Foresight Study 2018: an example of a comprehensive foresight exercise. Eur J Public Health 2020 Feb 01;30(1):30-35. [doi: 10.1093/eurpub/ckz200] [Medline: 31647528]
32. Watanabe Y, Tsubo T. Forecasting Japanese health futures with the BFT. Futures 1995 Nov;27(9-10):959-966 [FREE Full text] [doi: 10.1016/0016-3287(95)00062-3]
33. Nejatzadehgan Eidgahi Z, Dehnavieh R, Borhaninejad V, Alikhani A, Rouhi V. The Future Wheel model of the effects of the emerging infectious diseases pandemics on the elderly in Iran. Salmand 2024 Jan 01;18(04):606-619. [doi: 10.32598/sija.2023.3641.1]
34. Aguilar PLC, Vásquez JEM. Possible future scenarios of the general health social security system in Colombia for the year 2033. Eur J Futures Res 2023 Feb 10;11(1):1. [doi: 10.1186/s40309-022-00213-7]
35. Saygitov RT. Health-related R and D priorities until 2030: Russian experience. FS 2017 Sep 11;19(5):501-510. [doi: 10.1108/fs-06-2016-0022]
36. Pihlainen V, Kivinen T, Lammintakanen J. Perceptions of future hospital management in Finland. JHOM 2019 Aug 08;33(5):530-546. [doi: 10.1108/jhom-02-2018-0045]
37. Daniels K, Grim T, Morgan T. Using foresight to explore the impacts of flooding in Houston on health, poverty, and equity out to 2050. Futures 2021 Aug;131:102754 [FREE Full text] [doi: 10.1016/j.futures.2021.102754]
38. Schreuder R. Health scenarios and policy making. Futures 1995 Nov;27(9-10):953-958 [FREE Full text] [doi: 10.1016/0016-3287(95)00061-5]
39. Taghizadeh S, Alesaeidi S, Jafari-Koshki T, Valizadeh-Otaghsara SM, Poursheikhali A, Tousi AZ, et al. Trend Impact Analysis (TIA) of community-based futures study for pediatric obesity in Iran. BMC Pediatr 2023 Feb 08;23(1):66 [FREE Full text] [doi: 10.1186/s12887-023-03880-y] [Medline: 36750801]
40. Rezaei H, Haghdoost A, Javar HA, Dehnavieh R, Aramesh S, Dehgani N, et al. The effect of coronavirus (COVID-19) pandemic on medical sciences education in Iran. J Educ Health Promot 2021;10:136 [FREE Full text] [doi: 10.4103/jehp.jehp_817_20] [Medline: 34222511]
41. Fourie P. The future of AIDS in Africa: lessons from two scenario projects. Afr J AIDS Res 2007 Aug;6(2):97-107. [doi: 10.2989/16085900709490404] [Medline: 25866059]

XSL•FO
RenderX

42. Damari B, Hajian M, Minaee F, Riazi-Isfahani S. The impact of Future World events on Iranians' social health: A qualitative futurology. Iran J Public Health 2016 Jun;45(6):795-805 [FREE Full text] [Medline: 27648424]
43. Mehrolhassani MH, Mozhdehifard M, Rahimisadegh R. Key factors in the future of oral and dental health in Iran using scenario writing approach. BMC Oral Health 2024 May 14;24(1):563 [FREE Full text] [doi: 10.1186/s12903-024-04354-y] [Medline: 38745163]
44. Hedayati M, Maleki M, Asl IM, Fazaeli AA, Goharinezhad S. Exploring the driving forces and scenario analysis for catastrophic and impoverishing health expenditures in Iran. BMC Health Serv Res 2024 Feb 26;24(1):245 [FREE Full text] [doi: 10.1186/s12913-024-10551-w] [Medline: 38409010]
45. Postma T, Alers J, Terpstra S, Zuurbier A. Medical technology decisions in The Netherlands: How to solve the dilemma of technology foresight versus market research? Technological Forecasting and Social Change 2007 Nov;74(9):1823-1833 [FREE Full text] [doi: 10.1016/j.techfore.2007.05.011]
46. Didier L, Camille S, Franck C, Zeina M. How to anticipate and control possible future health crises through foresight approaches using the COVID-19 pandemic crisis? Arch Community Med Public Health 2023 Jun 23;9(2):034-037. [doi: 10.17352/2455-5479.000199]
47. Gariboldi MI, Lin V, Bland J, Auplish M, Cawthorne A. Foresight in the time of COVID-19. Lancet Reg Health West Pac 2021 Jan;6:100049 [FREE Full text] [doi: 10.1016/j.lanwpc.2020.100049] [Medline: 33521742]
48. Brookes VJ, Hernández-Jover M, Black PF, Ward MP. Preparedness for emerging infectious diseases: pathways from anticipation to action. Epidemiol. Infect 2014 Dec 12;143(10):2043-2058. [doi: 10.1017/s095026881400315x]
49. Ozdemir V, Faraj SA, Knoppers BM. Steering vaccinomics innovations with anticipatory governance and participatory foresight. OMICS 2011 Sep;15(9):637-646. [doi: 10.1089/omi.2011.0087] [Medline: 21848419]
50. Hosseini Golkar M, Mowlaei MM, Behzadi A, Keshavarz-Turk M, Zolfagharnasab A, Hosseini Golkar M. The importance of uncertainty in health scenarios: A scoping review on COVID-19 scenarios. Med J Islam Repub Iran 2023 Oct 20;37:137. [doi: 10.47176/mjiri.37.137] [Medline: 38435831]
51. Gross J, Davids T, Sools A, Saghai Y. Covid-19 and the politics of hope: A comparative analysis of Greek and Ecuadorian letters from a desired post-pandemic future. Futures 2023 May;149:103115 [FREE Full text] [doi: 10.1016/j.futures.2023.103115] [Medline: 36777995]
52. Crawford M, Wright G. The value of mass-produced COVID-19 scenarios: A quality evaluation of development processes and scenario content. Technol Forecast Soc Change 2022 Oct;183:121937 [FREE Full text] [doi: 10.1016/j.techfore.2022.121937] [Medline: 35945976]
53. Damrongchai N, Johnston R, Tegart G. The impact of foresight studies on human healthcare in the post-genomic era. IJFIP 2010;6(1/2/3):151. [doi: 10.1504/ijfip.2010.032671]
54. Bijl R. Delphi in a future scenario study on mental health and mental health care. Futures 1992 Apr;24(3):232-250 [FREE Full text] [doi: 10.1016/0016-3287(92)90033-c]
55. Bishop P, Tamarchak R, Williams C, Radvanyi L. Innovative application of strategic foresight to oncology research. FS 2020 Nov 12;22(5/6):533-550. [doi: 10.1108/fs-03-2020-0028]
56. Mota F, Braga L, Cabral B, Filho C. Future of bioprinted tissues and organs: A two-wave global survey. Foresight and STI Governance 2022;16(1):6-20 [FREE Full text]
57. Streit JMK, Felknor SA, Edwards NT, Howard J. Leveraging strategic foresight to advance worker safety, health, and well-being. Int J Environ Res Public Health 2021 Aug 11;18(16):8477 [FREE Full text] [doi: 10.3390/ijerph18168477] [Medline: 34444224]
58. Felknor S, Streit J, Edwards N, Howard J. Four futures for occupational safety and health. Int J Environ Res Public Health 2023 Feb 28;20(5):4333 [FREE Full text] [doi: 10.3390/ijerph20054333] [Medline: 36901347]
59. Felknor S, Streit J, Chosewood L, McDaniel M, Schulte P, Delclos G. How will the future of work shape the OSH professional of the future? A workshop summary. Int J Environ Res Public Health 2020 Sep 30;17(19):7154 [FREE Full text] [doi: 10.3390/ijerph17197154] [Medline: 33007820]
60. Vermeulen L, Kolesar J, Crismon M, Flynn A, Stevenson J, Almeter P, et al. ASHP Foundation Pharmacy Forecast 2018: Strategic planning advice for pharmacy departments in hospitals and health systems. Am J Health Syst Pharm 2018 Jan 15;75(2):23-54 [FREE Full text] [doi: 10.2146/sp180001] [Medline: 29158305]
61. Özdemir V, Dove ES, Gürsoy U, Şardaş S, Yıldırım A, Yılmaz, et al. Personalized medicine beyond genomics: alternative futures in big data-proteomics, environtome and the social proteome. J Neural Transm (Vienna) 2017 Jan 8;124(1):25-32. [doi: 10.1007/s00702-015-1489-y] [Medline: 26645377]
62. Canongia C. Synergy between Competitive Intelligence (CI), Knowledge Management (KM) and Technological Foresight (TF) as a strategic model of prospecting--the use of biotechnology in the development of drugs against breast cancer. Biotechnol Adv 2007;25(1):57-74 [FREE Full text] [doi: 10.1016/j.biotechadv.2006.10.001] [Medline: 17156964]
63. Flick C, Zamani E, Stahl B, Brem A. The future of ICT for health and ageing: Unveiling ethical and social issues through horizon scanning foresight. Technological Forecasting and Social Change 2020 Jun;155:119995 [FREE Full text] [doi: 10.1016/j.techfore.2020.119995]
64. Hemmat M, Ayatollahi H, Maleki M, Saghafi F. Future research in health information technology: A review. Perspect Health Inf Manag 2017;14(Winter):1b. [Medline: 28566991]

XSL•FO
RenderX

65. Mittelstadt BD, Floridi L. The ethics of big data: Current and foreseeable issues in biomedical contexts. Sci Eng Ethics 2016 Apr;22(2):303-341. [doi: 10.1007/s11948-015-9652-2] [Medline: 26002496]
66. Butter M, Hoogendoorn J. Foresight versus FP7: comparing innovations in healthcare. Foresight 2008;10(6):39-61 [FREE Full text] [doi: 10.1108/14636680810918577]
67. Masum H, Ranck J, Singer PA. Five promising methods for health foresight. Foresight 2010 Feb 23;12(1):54-66. [doi: 10.1108/14636681011020182]
68. Kulkova J, Kulkov I, Rohrbeck R, Lu S, Khwaja A, Karjaluoto H, et al. Medicine of the future: How and who is going to treat us? Futures 2023 Feb;146:103097 [FREE Full text] [doi: 10.1016/j.futures.2023.103097]
69. Pereno A, Eriksson D. A multi-stakeholder perspective on sustainable healthcare: From 2030 onwards. Futures 2020 Sep;122:102605 [FREE Full text] [doi: 10.1016/j.futures.2020.102605] [Medline: 32834076]
70. Pau S, Hall A. New spaces for healthcare futures studies: Connecting existing theory to deeper participatory practice. Futures 2021 Feb;126:102689 [FREE Full text] [doi: 10.1016/j.futures.2020.102689]
71. Lamé G, Jouini O, Stal-Le Cardinal J. Methods and contexts: Challenges of planning with scenarios in a hospital's division. Futures 2019 Jan;105:78-90 [FREE Full text] [doi: 10.1016/j.futures.2018.09.005]
72. Antunes A, Canongia C. Technological foresight and technological scanning for identifying priorities and opportunities: the biotechnology and health sector. Foresight 2006;8(5):31-44 [FREE Full text] [doi: 10.1108/14636680610703072]
73. Karlsen JE, Gual A, Anderson P. Foresighting addiction and lifestyles in Europe 2030+. Eur J Futures Res 2013 Oct 6;1(1):19. [doi: 10.1007/s40309-013-0019-0]
74. Bezold C. The future of health futures. Futures 1995 Nov;27(9-10):921-925 [FREE Full text] [doi: 10.1016/0016-3287(95)00059-3]
75. Haghdoost A, Pourhosseini SS, Emami M, Dehnavieh R, Barfeh T, Mehrolhassani MH. Foresight in health sciences using CLA method. Med J Islam Repub Iran 2017 Dec 30;31(1):84-499 [FREE Full text] [doi: 10.14196/mjiri.31.84] [Medline: 29951385]
76. Hosseini Golkar M, Ghazinoory SS, Saghafi F, Eivazi M, Poursheikhali A, Dehnavieh R. Appropriate types of futures studies scenarios in health. Med J Islam Repub Iran 2017 Dec 30;31(1):82-483 [FREE Full text] [doi: 10.14196/mjiri.31.82] [Medline: 29951383]
77. Doos L, Packer C, Ward D, Simpson S, Stevens A. Past speculations of the future: a review of the methods used for forecasting emerging health technologies. BMJ Open 2016 Mar 10;6(3):e010479 [FREE Full text] [doi: 10.1136/bmjopen-2015-010479] [Medline: 26966060]
78. Stelzer B, Meyer-Brötz F, Schiebel E, Brecht L. Combining the scenario technique with bibliometrics for technology foresight: The case of personalized medicine. Technological Forecasting and Social Change 2015 Sep;98:137-156 [FREE Full text] [doi: 10.1016/j.techfore.2015.06.008]
79. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020 Sep 11;3(1):118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
80. Determinants of health. World Health Organization. 2017. URL: https://www.who.int/news-room/questions-and-answers/item/determinants-of-health [accessed 2024-09-24]
81. Quality of life. Britannica. URL: https://www.britannica.com/topic/quality-of-life [accessed 2024-09-24]
82. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017 Sep;3:38-38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
83. Meskó B. The real era of the art of medicine begins with artificial intelligence. J Med Internet Res 2019 Nov 18;21(11):e16295 [FREE Full text] [doi: 10.2196/16295] [Medline: 31738169]
84. Stieg C. How this Canadian start-up spotted coronavirus before everyone else knew about it. CNBC. URL: https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html [accessed 2024-09-24]
85. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: A report from the National Academy of Medicine. JAMA 2020 Feb 11;323(6):509-510. [doi: 10.1001/jama.2019.21579] [Medline: 31845963]
86. Meskó B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
87. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]

Abbreviations

AI: artificial intelligence
FDA: Food and Drug Administration
ICT: information and communication technologies
OSH: occupational safety and health
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

XSL•FO
RenderX

Edited by G Eysenbach, A Mavragani; submitted 06.02.24; peer-reviewed by H Ali, A Hassan, F Mota; comments to author 29.03.24; revised version received 19.04.24; accepted 23.08.24; published 09.10.24.

Please cite as:
Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G
Exploring the Need for Medical Futures Studies: Insights From a Scoping Review of Health Care Foresight
J Med Internet Res 2024;26:e57148
URL: https://www.jmir.org/2024/1/e57148
doi:10.2196/57148
PMID:

©Bertalan Meskó, Tamás Kristóf, Pranavsingh Dhunnoo, Nóra Árvai, Gellért Katonai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Impact of Skin Pigmentation on Pulse Oximetry Blood Oxygenation and Wearable Pulse Rate Accuracy: Systematic Review and Meta-Analysis

Sanidhya Singh[1], BSE; Miles Romney Bennett[2], MS; Chen Chen[2], PhD; Sooyoon Shin[2], PhD; Hamid Ghanbari[1*], MD, MPH; Benjamin W Nelson[2,3*], PhD

[1]University of Michigan Medical School, Ann Arbor, MI, United States
[2]Verily Life Sciences LLC, South San Francisco, CA, United States
[3]Division of Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Sanidhya Singh, BSE
University of Michigan Medical School
1301 Catherine Street
Ann Arbor, MI, 48109
United States
Phone: 1 7349364000
Email: sansin@med.umich.edu

Abstract

Background: Photoplethysmography (PPG) is a technology routinely used in clinical practice to assess blood oxygenation (SpO_2) and pulse rate (PR). Skin pigmentation may influence accuracy, leading to health outcomes disparities.

Objective: This systematic review and meta-analysis primarily aimed to evaluate the accuracy of PPG-derived SpO_2 and PR by skin pigmentation. Secondarily, we aimed to evaluate statistical biases and the clinical relevance of PPG-derived SpO_2 and PR according to skin pigmentation.

Methods: We identified 23 pulse oximetry studies (n=59,684; 197,353 paired SpO_2-arterial blood observations) and 4 wearable PR studies (n=176; 140,771 paired PPG-electrocardiography observations). We evaluated accuracy according to skin pigmentation group by comparing SpO_2 accuracy root-mean-square values to the regulatory threshold of 3% and PR 95% limits of agreement values to +5 or –5 beats per minute (bpm), according to the standards of the American National Standards Institute, Association for the Advancement of Medical Instrumentation, and the International Electrotechnical Commission. We evaluated biases and clinical relevance using mean bias and 95% CI.

Results: For SpO_2, accuracy root-mean-square values were 3.96%, 4.71%, and 4.15%, and pooled mean biases were 0.70% (95% CI 0.17%-1.22%), 0.27% (95% CI –0.64% to 1.19%), and 1.27% (95% CI 0.58%-1.95%) for light, medium, and dark pigmentation, respectively. For PR, 95% limits of agreement values were from –16.02 to 13.54, from –18.62 to 16.84, and from –33.69 to 32.54, and pooled mean biases were –1.24 (95% CI –5.31 to 2.83) bpm, –0.89 (95% CI –3.70 to 1.93) bpm, and –0.57 (95% CI –9.44 to 8.29) bpm for light, medium, and dark pigmentation, respectively.

Conclusions: SpO_2 and PR measurements may be inaccurate across all skin pigmentation groups, breaching U.S. Food and Drug Administration guidance and industry standard thresholds. Pulse oximeters significantly overestimate SpO_2 for both light and dark skin pigmentation, but this overestimation may not be clinically relevant. PRs obtained from wearables exhibit no statistically or clinically significant bias based on skin pigmentation.

(J Med Internet Res 2024;26:e62769) doi:10.2196/62769

KEYWORDS

photoplethysmography; pulse oximetry; arterial blood gas; skin tone; skin pigmentation; bias; digital technology

XSL•FO
RenderX

Introduction

Photoplethysmography (PPG) technology has been used in medicine since the 1970s to assess pulse rate (PR) and blood oxygenation (SpO_2). The accuracy of PPG-based SpO_2 and PR is critical for medical practice, clinical decision-making, and patient outcomes [1].

Technological advancements have led to rapid expansion of this technology into consumer devices [2]. This has enabled consumers to continuously and unobtrusively track health status, generating information that is becoming commonly used in health care settings and as a source of research end points [3,4].

Researchers have identified several factors that influence the accuracy of PPG-based measurements, including skin temperature, sensor contact pressure, skin thickness, and hydration level [5,6]. Notably, darker skin pigmentation may influence PPG-based SpO_2 and PR readings [7-10]. Patients with darker skin pigmentation are more likely than patients with lighter skin color to have overestimated SpO_2 readings, leading to lower hospital admission; higher occult hypoxemia; and delayed or no access to dexamethasone, therapeutic oxygen, and COVID-19 therapies, resulting in increased hospital readmission, organ dysfunction, and mortality [11-18].

Starting in 2013, a series of U.S. Food and Drug Administration (FDA) guidances, safety communications, and more recently, statements from attorneys general have called to address darker skin pigmentation bias in pulse oximeters [19]; however, these guidances lacked standards for assessing skin tone [20]. Recent pulse oximeter research has rekindled interest in skin pigmentation disparities in PPG sensor accuracy, leading to increased media and regulatory attention [21-24]. In November 2023, more than 24 attorneys general wrote a letter calling on the FDA to take urgent action to address pulse oximeter skin pigmentation disparities [25,26].

More recent research has also highlighted the potential presence of these biases in research-grade and consumer PPG-based wearable devices [4,27-29], leading to calls to address inequity, bias, and discrimination in wearable health technology and clinical practice algorithms [30,31].

Accuracy guidance for pulse oximeters and heart rate or PR devices has been delineated by regulatory bodies, industry, and medical standards. Overall accuracy of pulse oximeters can be assessed with accuracy root-mean-square (A_{rms}), which combines mean bias and precision (SD of bias) into a single metric [32]. FDA guidance has a threshold of A_{rms}≤3% [19] for transmittance devices and a threshold of A_{rms}≤3.5% [19] for ear clip and reflectance devices, while international thresholds are set at A_{rms}≤4% [32]. For heart rate or PR devices, overall accuracy can be assessed with 95% limits of agreement (LoAs). The American National Standards Institute (ANSI), Association for the Advancement of Medical Instrumentation (AAMI), and International Electrotechnical Commission (IEC) have set the recommendation that electrocardiography (ECG) devices have mean bias of +5 or –5 beats per minute (bpm) or mean absolute percentage error≤10%, whichever is greater [33,34]. Devices with accuracy measures breaching these accuracy thresholds can produce questionable results.

It is, therefore, critical to generate evidence to inform the design and calibration of these devices to reduce algorithmic bias and improve accuracy, so that PPG-based technology can generalize to all segments of the population, mitigating racial disparities in health outcomes. There have been systematic reviews on PPG skin pigmentation bias, but these have been limited to either consumer device PR [35] or pulse oximeter SpO_2 [36-38]. Only 1 meta-analysis focused on pulse oximetry for this topic is available [36], and an additional 7 studies have been published recently, providing an additional 50,980 participants and 182,369 paired observations. Therefore, performing a comprehensive meta-analysis to examine PPG accuracy, potential bias, and clinical relevance for SpO_2 and PR by skin pigmentation is timely.

Methods

Search Strategy and Selection Criteria

We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [39] and developed a MEDLINE systematic search. Searches were performed between April 2022 and June 2023, and a final ad hoc search was conducted in June 2023. No additional studies were included after this date.

We included studies that (1) investigated PPG-derived SpO_2 or PR test devices per definition [40]; (2) used arterial blood gas (SaO_2) or ECG as the reference device; and (3) reported mean bias and SD, SE, 95% LoA, or 95% CI by race, ethnicity, or skin tone. Inclusion disagreements were resolved by consensus.

Exclusion criteria included (1) literature reviews, systematic reviews, commentary, and meta-analyses; (2) non-English manuscripts; (3) irretrievable full source texts; (4) studies on remote PPG; and (5) those that lack SaO_2 or ECG as the reference device. We chose to not exclude papers based on measurement hardware or underlying algorithms given that all measurement devices relied on contact-based PPG to measure the same end points of interest—SpO_2 and PR.

Data Analysis

Data Extraction

Data were independently extracted from published manuscripts (BWN and SS). A third and fourth reviewer (MRB and HG) adjudicated any differences between the initial 2 reviewers and resolved any disagreements by checking manuscripts. The first author verified the quality of manuscripts. Standardized data extraction was developed to extract the study characteristics (see the Materials—Data Extraction section in Multimedia Appendix 1). For 1 paper, we used participant-level data from the supplementary materials to calculate mean bias and SE by skin pigmentation category [41]. This resulted in 27 studies in our final analysis (Figure 1 and Table 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram. ECG: electrocardiography; PPG: photoplethysmography; SpO_2: blood oxygenation.

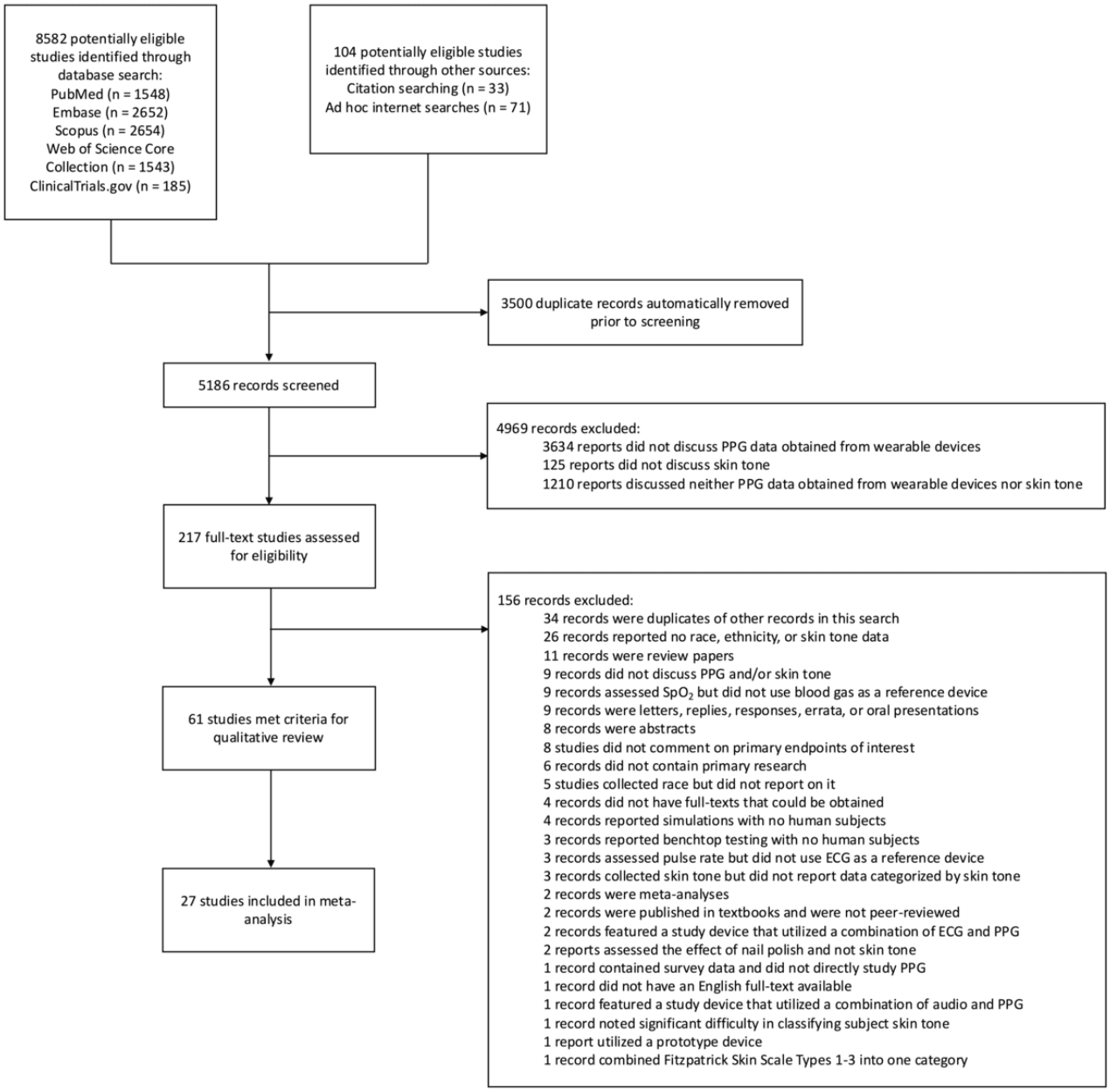

XSL•FO
RenderX

Table 1. Characteristics of included studies.

Device type and study (author, year)	Sample size, n	Test device evaluated	Reference device	SaO_2[a] (%) range	Participant population	Research setting	Skin pigmentation method
Pulse Oximetry							
Abrams et al [42], 2003	200	• Nellcor N-200	Radiometer ABL-520	Not reported	Medical—adult patients with cirrhosis	Inpatient	Race—Black and White
Adler et al [43], 1998	298	• Nellcor D-25	4-wavelength spectrophotometer, or co-oximeter (Radiometer OSM3)	50-99	Medical—adult emergency department patients	Inpatient	Skin tone—Munsell color system categorized into light, medium, or dark
Andrist et al [44], 2022	1061	• Not reported	Not reported	Not reported	Medical—pediatric	Inpatient	Race
Barker and Wilson [45], 2023	75	• Masimo SET[b] pulse oximeters with RD-SET sensors	Radiometer ABL-835 Flex CO-Oximeter	70-100	Mixed—healthy and mild systemic disease	Research—laboratory	Race
Bickler et al [10], 2005	23	• Nellcor N-595 with Nellcor OxiMax A finger probe • Novametrix 513s models (2 types) • Nonin Onyx models (2 types)	Radiometer OSM3	60-100	Healthy—nonsmoking	Research—laboratory	Ethnicity—light (Northern European) and dark (African American)
Bothma et al [46], 1996	100	• Simed S100e • Nihon Koden • Ohmeda 3740	IL482 Co-oximeter System	87.8-99.2	Medical—critically ill adult patients	Inpatient	Skin tone—EEL[c] reflectance spectrophotometer; all participants had dark skin tone
Burnett et al [47], 2022	46,253	• Unspecified Nellcor and Masimo devices	GEMStat Premier 3000	Not reported	Medical—patients receiving anesthesia	Inpatient	Race and ethnicity
Crooks et al [48], 2022	2997	• Not reported	Not reported	Not reported	Medical—COVID-19	Inpatient	Race
Ebmeier et al [49], 2018	394	• Marquette Rac-4A monitors with Masimo sensors and Philips IntelliVue • MP70 monitors with Philips • Adult Reusable SpO_2[d] sensors	Radiometer ABL 800 FLEX SaO_2 analyzer	Not reported	Medical—multiple conditions	Inpatient	Ethnicity
Fawzy et al [17], 2022	1216	• Not reported	ABL825, ABL827, or ABL90 blood gas analyzers	Not reported	Medical—COVID-19	Inpatient	Race and ethnicity

XSL•FO
RenderX

Device type and study (author, year)	Sample size, n	Test device evaluated	Reference device	SaO_2[a] (%) range	Participant population	Research setting	Skin pigmentation method
Feiner et al [8], 2007	36	• Nellcor N-595 (OxiMax A adhesive probe) • Nellcor N-595 (a clip-type probe) • Masimo Radical (clip probe) • Masimo Radical (adhesive disposable probe) • Nonin 9700 (clip-type probe) • Nonin 9700 (disposable adhesive probe)	Radiometer OSM3 multiwavelength oximeter	60-100	Healthy—nonsmoking	Research—laboratory	Ethnicity—light (Caucasian), intermediate (Hispanic, Indian, Filipino, or Vietnamese), and dark (African American) categories
Foglia et al [50], 2017	36	• Nellcor Oximax (Covidien) • Masimo Rainbow SET Radical 7	Siemens Rapidlab 1265	60-92	Medical—infants with cyanotic congenital heart disease and oxygen saturation <90%	Inpatient	Skin tone—Munsell color system
Jubran and Tobin [51], 1990	54	• Nellcor pulse oximeter with disposable or reusable probes • Ohmeda-Biox3700 pulse oximeter with reusable probe	CO-oximetry	Not reported	Medical—critically ill, ventilator-dependent patients	Inpatient	Race—Black and White
McGovern et al [52], 1996	8	• Ohmeda 3700	IL 482 Co-oximeter	Not reported	Medical—adults with stable condition with severe COPD[f]	Research—laboratory	Race—all White
Muñoz et al [53], 2008	846	• Minolta Pulsox-7	IL 682 co-oximeter	Not reported	Medical—adults under assessment for long-term home oxygen therapy	Outpatient	Race—all White
Pilcher et al [54], 2020	400	• Carescape B450 monitor with Nellcor probe • GE Dash 3000 • Masimo Radical 7 • Masimo SET Quartz (unspecified) • Masimo SET Quartz Q400 • Nonin 2120 • Nonin 2140 • Nonin Avant (unspecified) • Nonin Avant 4000 • Nonin Avant 9700 • Nonin Lifesense Medair • Novametrix Model 512 • Ohmeda Biox 3700E with a GE TruSignal or Nellcor probe • Philips Intellivue MP70 with a GE TruSignal Nellcor or Philips probe • Welch Allyn with a Nellcor probe	Radiometer ABL800	72-100	Medical—hospitalized adult patients	Inpatient and outpatient	Skin Tone—Fitzpatrick Scale categorized into light, medium, and dark
	774	• Not reported	Not reported			Inpatient	Race

XSL•FO
RenderX

Device type and study (author, year)	Sample size, n	Test device evaluated	Reference device	SaO_2[a] (%) range	Participant population	Research setting	Skin pigmentation method
Ruppel et al [55], 2023				Not reported	Medical—cardiac catheterization		
Sudat et al [56], 2023	8735	• Not reported	Not reported	Not reported	Medical—multiple conditions	Inpatient	Race
Thrush and Hodges [57], 1994	25	• Critikon Dinamap Plus Model 8700 • Critikon Oxyshuttle Ohmeda 3700 Catalyst Research MiniOx IV	IL482 co-oximeter	80-100	Healthy—non-smoking adults	Research—laboratory	Race—all White
Valbuena et al [12], 2022	372	• Not reported	Not reported, blood gas analysis	Not reported	Medical—adult patients with respiratory failure or COVID-19	Inpatient	Race and ethnicity—White, Black, Hispanic, and Asian
Vesoulis et al [58], 2021	294	• Nellcor SpO_2 module with Neonatal-Adult MAX-N adhesive SpO_2 sensor (Covidien) (used with either Philips IntelliVue MP70 or MX800 monitors)	Radiometer ABL800 Flex	Not reported	Medical—preterm infants at neonatal intensive care unit	Inpatient	Race—White and Black
Wiles et al [59], 2022	194	• Nellcor reusable SpO_2 probes or Mindray disposable SpO_2 probes (GE Healthcare B1x5 M/P monitor)	RAPIDpoint 500 analyser (Siemens Healthcare GmbH)	Not reported	Medical—adult patients with COVID-19 pneumonitis	Inpatient	Race—Asian, Black, White, and other
Zeballos and Weisman [9], 1991	33	• Hewlett-Packard HP-47201A • Ohmeda Biox IIA	IL282 co-oximeter	Not reported	Healthy—non-smoking volunteers	Research—laboratory	Race—all Black
Pulse Rate							

XSL•FO
RenderX

Device type and study (author, year)	Sample size, n	Test device evaluated	Reference device	SaO_2[a] (%) range	Participant population	Research setting	Skin pigmentation method
Bent et al [41], 2020	53	• Empatica E4 • Apple Watch • Fitbit Charge • Garmin Vivosmart 3 • Xiaomi Miband • Biovotion Everion	BittiumFaros 180, Bittium Inc.	N/A[g]	Healthy	Research — laboratory	Skin tone—Fitzpatrick Scale: • 1 (n=7) • 2 (n=8) • 3 (n=10) • 4 (n=9) • 5 (n=9) • 6 (n=10)
Nelson and Allen [60], 2019	1	• Apple Watch 3 • Fitbit Charge 2	Vrije Universiteit Ambulatory Monitoring System	N/A	Healthy	Research — real world	Skin tone—Fitzpatrick Scale: • 1 (n=1)
Sanudo et al [61], 2019	45	• Apple Watch (version not reported)	Polar Chest Strap	N/A	Healthy	Research — laboratory	Skin tone—Fitzpatrick Scale: • 2 (n=15) • 3 (n=15) • 4 (n=15)
Chow and Yang [62], 2020	40	• Garmin Vivosmart HR+ • Xiaomi Mi Band 2	Polar H7 Chest Strap	N/A	Healthy	Research — laboratory	Ethnicity and skin tone—East Asian, Fitzpatrick Scale 3 and 4

[a]SaO_2: arterial blood gas.

[b]SET: Signal Extraction Technology.

[c]EEL: electron energy loss.

[d]SpO_2: blood oxygenation.

[f]COPD: chronic obstructive pulmonary disease.

[g]N/A: not applicable.

Quality Assessment of the Overall Evidence

QUADAS-2 tool was used to evaluate risk of bias and applicability (Figure S1 in Multimedia Appendix 1). Funnel plots evaluated publication bias using the *metafor* package [63] (Figures S2 and S3 in Multimedia Appendix 1).

Statistical Analysis

Skin Tone Categorization

We mapped skin tone, race, and ethnicity into 3 primary skin pigmentation groups of light, medium, and dark, following published methodology [36] (see Table S1 in Multimedia Appendix 1) and examined biases for SpO_2 and PR by each skin pigmentation group (described below).

We elected to use this same skin pigmentation categorization schema as has previously been used, with the goal of expanding upon the analysis of Shi et al [36].

Statistical Analysis

The objective of the study was to assess whether the devices were accurate in estimating SpO_2 and PR when compared with a SaO_2 and ECG reference device, respectively, for each skin pigmentation group. If these measures were found to be inaccurate, biases were quantified and their clinical relevance was assessed. The analytical approach to execute these research objectives was formulated based on methodologies used in prior meta-analyses within the discipline [36].

Evaluation End Points and Criteria

A summary of the evaluation criteria used can be found in Table 2.

XSL•FO
RenderX

Table 2. Evaluation end points and criteria for blood oxygenation (SpO_2) and pulse rate studies.

Category and objective	Evaluation end point	Evaluation criteria
SpO_2		
Evaluate accuracy	A_{rms}[a]	Accurate if A_{rms}≤3%
Statistically significant bias	Mean bias (95% CI)	Statistically significant bias if 95% CI does not contain 0
Clinically relevant bias	Mean bias (95% CI)	Clinically relevant bias if either upper bound of 95% CI≤4% or lower bound of 95% CI>4%
Pulse rate		
Evaluate accuracy	95% LoA[b]	Accurate if 95% LoA is bounded by +5 or –5 bpm[c]
Statistically significant bias	Mean bias (95% CI)	Statistically significant bias if 95% CI does not contain 0
Clinically relevant bias	Mean bias (95% CI)	Clinically relevant bias if either upper bound of 95% CI ≤5 bpm or lower bound of 95% CI >5 bpm

[a]A_{rms}: accuracy root-mean-square.

[b]LoA: limits of agreement.

[c]bpm: beats per minute.

Accuracy

To evaluate the accuracy of SpO_2 , A_{rms} was used to measure accuracy, because of its commonality in the regulatory space [19]. It was calculated as , where precision is the SD of bias [42,43]. A_{rms} is useful in clinical settings because it provides a single metric that accounts for both the systematic error (bias) and the random error (precision) in measurements and it provides a comprehensive assessment of how close the measurements are to the true values. A lower A_{rms} value indicates better accuracy, meaning that the measurement is closer to the true value, whereas a higher A_{rms} value suggests lower accuracy, indicating that the measurement is further away from the true value.

For SpO_2, our study used the stricter A_{rms}>3% threshold to define inaccuracy as per FDA guidance [19]. We used this threshold as 22 (96%) out of 23 of pulse oximeter studies used a transmittance device and only one used an ear clip, which would have set a more liberal threshold of 3.5%. To evaluate accuracy of PR, the 95% LoA of bias was used, and LoA was constructed as mean bias (SD 1.96). For PR, a pooled 95% LoA not bounded by +5 or –5 bpm was considered inaccurate per ANSI/AAMI/IEC standards [33,34].

Bias

To evaluate bias of both SpO_2 and PR, mean bias and its 95% CI were used. For each data point that compares the test with the reference device, bias was constructed as *test device – reference device*. Based on the final results, a measure was considered to have statistically significant bias if the 95% CI of its mean bias did not contain 0, with the estimated mean bias over 0 indicating overestimation and under 0 indicating underestimation.

Clinical Relevance of Estimated Bias

To evaluate the clinical relevance of the estimated bias for SpO_2, if the 95% CI of mean bias was out of the +4% to –4% range it was considered clinically relevant. This threshold is inferred from FDA guidance on *Pulse Oximeters—Premarket Notification Submissions [510(k)s]: Guidance for Industry and Food and Drug Administration Staff* [19]. To evaluate the clinical relevance of the estimated bias for PR, if the 95% CI of mean bias was out of the bound of +5 or –5 bpm, it was considered clinically relevant as per ANSI/AAMI/IEC standards [33,34].

Statistical Meta-Analysis

To obtain pooled results for above listed end points, we collected sample size, paired observations, mean bias, and SD from various studies. When not available, we either transformed relevant parameters, such as 95% LoA and 95% CI, into these statistics or obtained them by analyzing the raw data.

Methods to Pool Measures of Bias

Correlated and hierarchical effect (CHE) models were used to pool mean biases from various studies. Specifically, 3-level hierarchical models were constructed, with Level 1 representing individual data points collected in a study, Level 2 representing potentially multiple comparisons within a research study, and Level 3 representing various studies included in the meta-analysis. Within a study, there may be multiple devices compared using the same participants' data, and these effect sizes are dependent. To account for this dependence, the CHE model used the robust variance estimation (RVE) method, which allowed us to combine data from single-measure design studies with multiple dependent estimates of effect size, even when the dependence is unknown [64]. To execute the CHE model with RVE, we needed to plug in assumed correlation coefficients for these dependent effect sizes. We used correlation coefficient 0.9 following published methods [36], and small (=0.30) and moderate (=0.60) correlation coefficients were also explored in sensitivity analyses. Additionally, one report [48] had substantially larger SE, so we conducted a separate analysis excluding its results as part of the sensitivity analysis. To evaluate the heterogeneity of the studies, we reported the overall I^2 (percentage of variability in the effect sizes that is not caused

XSL•FO
RenderX

by sampling error) and its breakdown of within-study (I^2_{Level2}) and between-study (I^2_{Level3}) portions. We conducted subgroup analyses to see if there were statistically significant differences in bias between skin pigmentation categories (Table S2 from Multimedia Appendix 1). Forest plots were provided to visualize the mean bias (Figures S2 and S3 from Multimedia Appendix 1).

Methods to Pool Measures of Accuracy

When it comes to providing pooled 95% LoA and A_{rms}, no established methodologies exist, so we followed an analytical approach used in prior meta-analyses within the discipline [36]. Specifically, SD of bias across studies were pooled using CHE models similar to those used to pool mean biases. The pooled mean bias and pooled SD were then used to provide pooled estimates of 95% LoA and A_{rms}. Specifically,

- Overall accuracy A_{rms} = [illegible]
- Overall 95% LoA = pooled mean bias ± 1.96 × pooled SD

Other Methods

Descriptive statistics were provided for study characteristics, the device intended use, mean bias based on different skin tones, and reference devices. All analyses were conducted using R (version 4.3.1; R Foundation for Statistical Computing) with statistical models based on R packages *metafor* [63] and *clubSandwich* [65].

Results

Study Selection and Characteristics

Search strategy resulted in 8582 records. We selected 27 studies for full review (pulse oximetry: n=23, PR: n=4; Figure 1, Table 1, and Table 2). A total of 23 pulse oximetry studies involving 59,684 participants with 197,353 paired SpO_2-SaO_2 observations were included. Additionally, 4 PR studies with 176 participants and 140,771 paired PR-ECG observations were analyzed (Table 3).

XSL•FO
RenderX

Table 3. Summary characteristics of included studies[a].

Item and subitem	Pulse oximetry (n=23), n (%)	Pulse rate (n=4), n (%)
Participants (pulse oximetry: n=59,684; pulse rate: n=176)		
Light skin pigmentation	40,416 (67.72)	31 (17.61)
Medium skin pigmentation	9967 (16.70)	129 (73.30)
Dark skin pigmentation	9301 (15.58)	16 (9.09)
Paired observations (pulse oximetry: n=197,353; pulse rate: n=140,771)		
Light skin pigmentation	131,008 (66.38)	43,116 (30.60)
Medium skin pigmentation	32,095 (16.26)	90,733 (64.50)
Dark skin pigmentation	34,250 (17.35)	6922 (4·92)
Device type included in study		
Medical	23 (100)	0 (0)
Nonregulated	0 (0)	4 (100)
Sensor type		
Transmittance	22 (96)	0 (0)
Reflectance	1 (4)	4 (100)
Patient population		
Healthy	4 (17)	3 (75)
Medical	18 (78)	0 (0)
Healthy and medical	1 (4)	0 (0)
Not reported	0 (0)	1 (25)
Research setting		
Medical inpatient	15 (65)	0 (0)
Medical outpatient	1 (4)	0 (0)
Medical combined	1 (5)	0 (0)
Research laboratory	6 (26)	3 (75)
Research in the real world	0 (0)	1 (25)
Skin pigmentation method		
Skin tone	4 (17)	4 (100)
Race	13 (57)	0 (0)
Ethnicity	3 (13)	0 (0)
Race and ethnicity	3 (13)	0 (0)

[a]Medical device defined as devices that received regulatory clearance for either blood oxygenation (SpO_2) or pulse rate (PR).

Descriptive Statistics on Race, Ethnicity, and Skin Tone

Pulse Oximetry

Skin pigmentation was classified by race in 13 (57%) out of 23 studies, ethnicity in 3 (13%) studies, skin tone in 4 (17%) studies, and both race and ethnicity in 3 (13%) studies (Table 3). Of the 59,684 patients with 197,353 paired observations, there were a total of 40,416 (67.72%) patients with light skin pigmentation with 131,008 (66.38%) paired observations, 9967 (16.70%) patients with medium skin pigmentation with 32,095 (16.26%) paired observations, and 9301 (15.58%) patients with dark skin pigmentation with 34,250 (17.35%) paired observations.

PR Results

Skin pigmentation was classified by race or ethnicity in 0 (0%) out of 4 studies and skin tone in 4 (100%) studies (Table 3). Of the 176 patients with 140,771 paired observations, there were a total of 31 (17.61%) participants with light skin pigmentation with 43,116 (30.60%) paired observations, 129 (73.30%) participants with medium skin pigmentation with 90,733 (64.50%) paired observations, and 16 (9.09%) participants with dark skin pigmentation with 6922 (4.92%) paired observations.

XSL•FO
RenderX

PPG Accuracy and Bias by Skin Pigmentation

Pulse Oximetry

The pooled A_{rms} across different skin pigmentation groups was 3.6%, 3.96%, 4.71%, and 4.15% for combined, light, medium, and dark skin pigmentation, respectively (Table 4 and Figures 2-4 [8-11,17,42-51,54-56,58,59]). Of note, studies implementing multiple trial conditions or using multiple study devices were shown multiple times in Figures 2-4 to delineate different devices used within the same study. We observed a pooled mean percent bias of 0.82% (95% CI 0.29%-1.35%) across all skin pigmentation groups using the CHE model. Between-study heterogeneity (I^2_{Level3}) accounted for 14.53% of the total variation, while within-study heterogeneity (I^2_{Level2}) explained 84.02% of total variation. Delineating by skin pigmentation, the pooled mean percent bias from the CHE model was 0.70% (95% CI 0.17%-1.22%) for light skin, 0.27% (95% CI –0.64% to 1.19%) for medium skin, and 1.27% (95% CI 0.58%-1.95%) for dark skin (Tables S3 and S4 in Multimedia Appendix 1). Subgroup analyses found no statistically significant difference between light and medium skin pigmentation (estimate=0.113, SE 0.259; 95% CI –0.459 to 0.686), but there was a statistically significant difference between light and dark skin pigmentation (estimate=0.596, SE 0.240; 95% CI 0.069-1.123), such that pooled bias for those with darker skin was higher as compared to those with lighter skin (Table S2 in Multimedia Appendix 1).

Figure 2. Forest plot showing bias in SpO_2 measurements in patients with light skin pigmentation. Multiple entries from the same study are included, each representing different devices evaluated. Squares denote study weight; center of squares denote observed study effect size; vertical lines denote study CIs; and diamond denotes pooled effect. CHE: correlated and hierarchical effect; SpO_2: blood oxygenation.

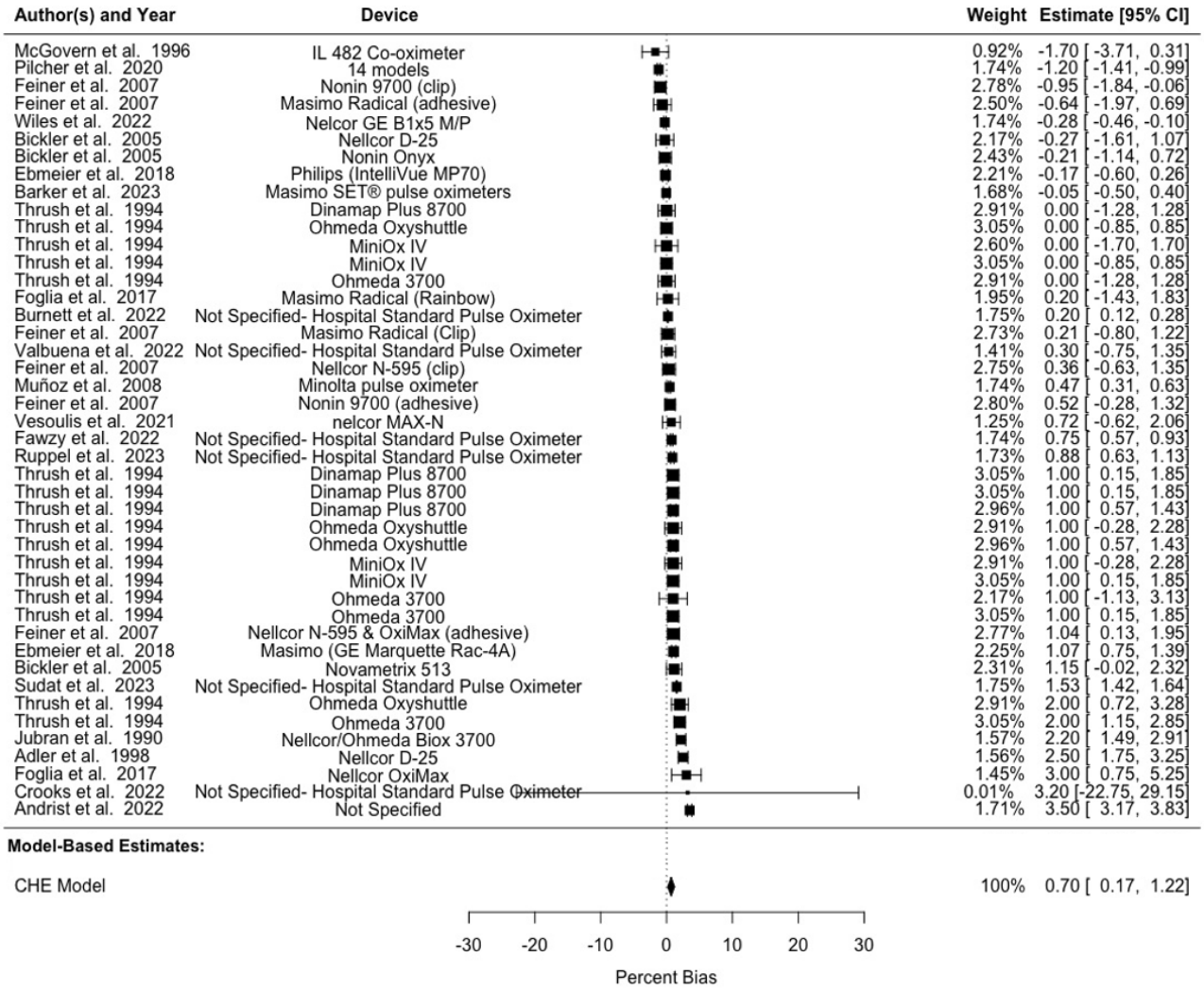

XSL•FO
RenderX

Figure 3. Forest plot showing bias in SpO_2 measurements in subjects with medium skin pigmentation. Multiple entries from the same study are included, each representing different devices evaluated. Note: squares denote study weight, center of squares denote observed study effect size, vertical lines denote study CIs, and diamond denotes pooled effect. CHE: correlated and hierarchical effect; SpO_2: blood oxygenation.

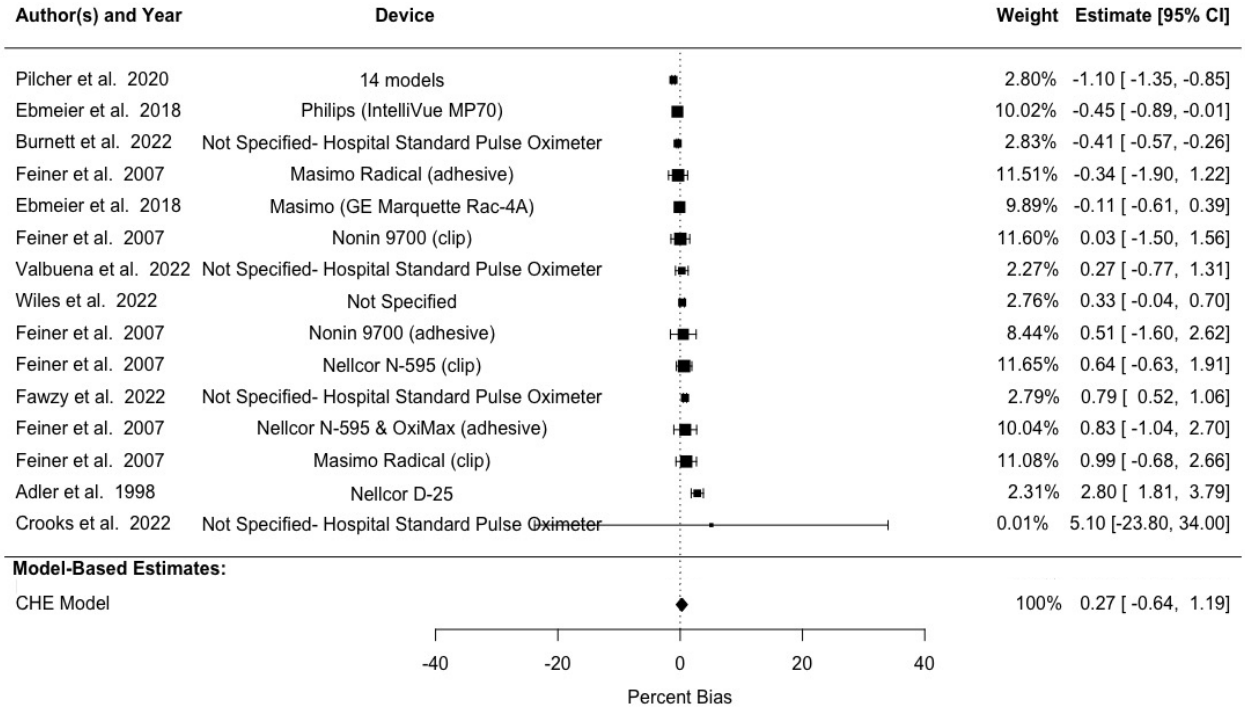

Figure 4. Forest plot showing bias in SpO_2 measurements in patients with dark skin pigmentation. Multiple entries from the same study are included, each representing different devices evaluated. Squares denote study weight; center of squares denote observed study effect size; vertical lines denote study CIs; and diamond denotes pooled effect. CHE: correlated and hierarchical effect; SpO_2: blood oxygenation.

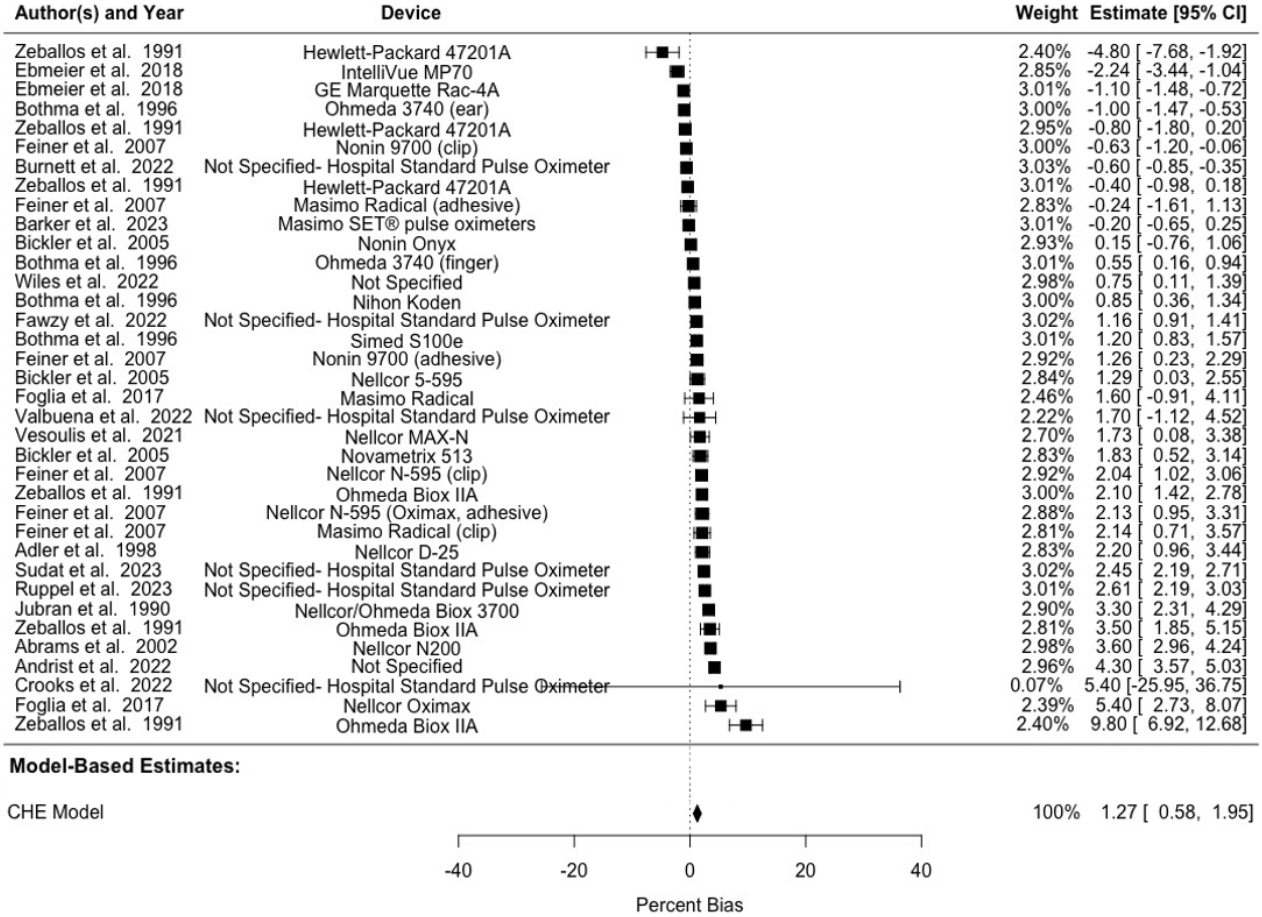

XSL•FO
RenderX

Table 4. Pulse rate and pulse oximetry bias by skin pigmentation[a].

Device type and skin pigment category	Studies (evaluations), n	Sample size (data pairs), n	Unit	Pooled mean bias (95% CI)	Pooled SD (SE)	95% LoA[b]	A_{rms}[c] (%)	Overall I^2 (between- and within-study heterogeneity)
Pulse oximetry								
Light	20 (44)	40,416 (131,008)	Percent	*0.70 (0.17 to 1.22)* [d]	3.90 (1.36)	–6.94 to 8.34	3.96[e]	97.45% (46.44% and 51.01%)
Medium	9 (15)	9,967 (32,095)	Percent	0.27 (–0.64 to 1.19)	4.71 (1.71)	–8.95 to 9.50	4.71[e]	95.31% (81.16% and 14.14%)
Dark	19 (36)	9,301 (34,250)	Percent	*1.27 (0.58 to 1.95)*	3.96 (1.30)	–6.49 to 9.02	4.15[e]	98.46% (0.00% and 98.46%)
Combined	23 (95)	59,684 (197,353)	Percent	*0.82 (0.29 to 1.35)*	3.50 (1.28)	–6.04 to 7.68	3.60[e]	98.55% (14.53% and 84.02%)
Pulse rate								
Light	3 (9)	31 (43,116)	bpm[f]	–1.24 (–5.31 to 2.83)	7.54 (2.13)	–16.02[g] to 13.54[g]	—[h]	10.99% (0.00% and 10.99%)
Medium	3 (9)	129 (90,733)	bpm	–0.89 (–3.70 to 1.93)	9.05 (1.75)	–18.62[g] to 16.84[g]	—	25.01% (0.00% and 25.01%)
Dark	1 (6)	16 (6,922)	bpm	–0.57 (–9.44 to 8.29)	16.89 (1.31)	–33.69[g] to 32·54[g]	—	13.70% (N/A[i] and 13.70%)
Combined	4 (24)	176 (140,771)	bpm	–0.29 (–3.87 to 3.29)	8.64 (1.67)	–17.23[g] to 16.65[g]	—	30.66% (26.76% and 3.90%)

[a]ρ=0.9 was used in correlated and hierarchical effect models to pool both mean bias and SD.
[b]LoA: limits of agreement.
[c]A_{rms}: accuracy root-mean-square
[d]Italicization denotes statistical significance.
[e]Exceeds U.S. Food and Drug Administration guidance for pulse oximetry.
[f]bpm: beats per minute.
[g]Exceeds American National Standards Institute standards for pulse rate.
[h]Not applicable.
[i]N/A: not available.

PR Results

Analysis using the CHE model revealed LoA values of –17.23 to 16.65 bpm and a mean bias of –0.29 (95% CI –3.87 to 3.29) bpm across all studies. Heterogeneity analysis demonstrated that 26.76% of the variation in bias (I^2_{Level3}) stemmed from between-study differences, while 73.34% (I^2_{Level2}) originated from within-study variation. Our analysis revealed 95% LoA of –16.02 to 13.54 bpm, –18.62 to 16.84 bpm, and –33.69 to 32.54 bpm for light, medium, and dark skin pigmentation groups, respectively. Mean biases were –1.24 (95% CI –5.31 to 2.83) bpm, –0.89 (95% CI –3.70 to 1.93) bpm, and –0.57 (95% CI –9.44 to 8.29) bpm for the corresponding groups (Table 4 and Figures 5-7 [41,60-62]). Detailed results are provided in Tables S4 and S5 in Multimedia Appendix 1. Subgroup analyses found no statistically significant difference between light and medium or between light and dark skin pigmentation pooled bias ($P \geq .05$; Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 5. Forest plot showing pulse rate measurement bias in patients with light skin pigmentation. Squares denote study weight; center of squares denotes observed study effect size; vertical lines denote study CIs; and diamond denotes pooled effect. CHE: correlated and hierarchical effect.

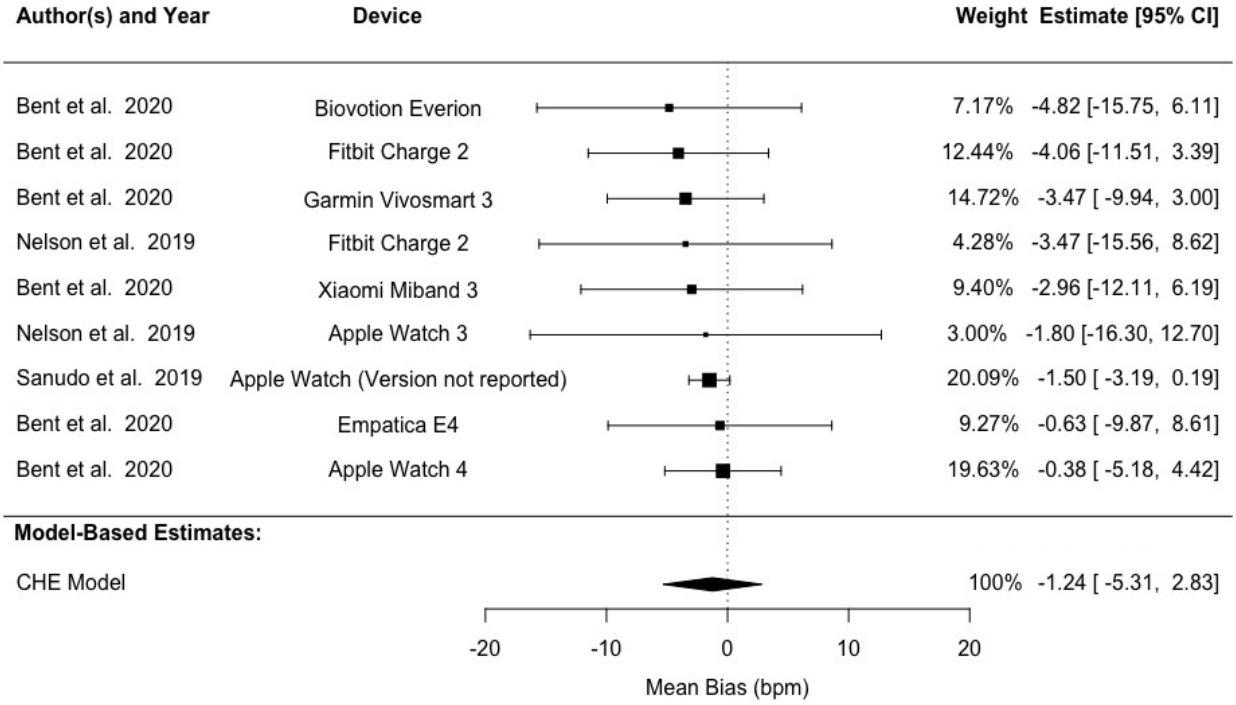

Figure 6. Forest plot showing pulse rate measurement bias in patients with medium skin pigmentation. Squares denote study weight; center of squares denotes observed study effect size; vertical lines denote study CIs; and diamond denotes pooled effect. CHE: correlated and hierarchical effect.

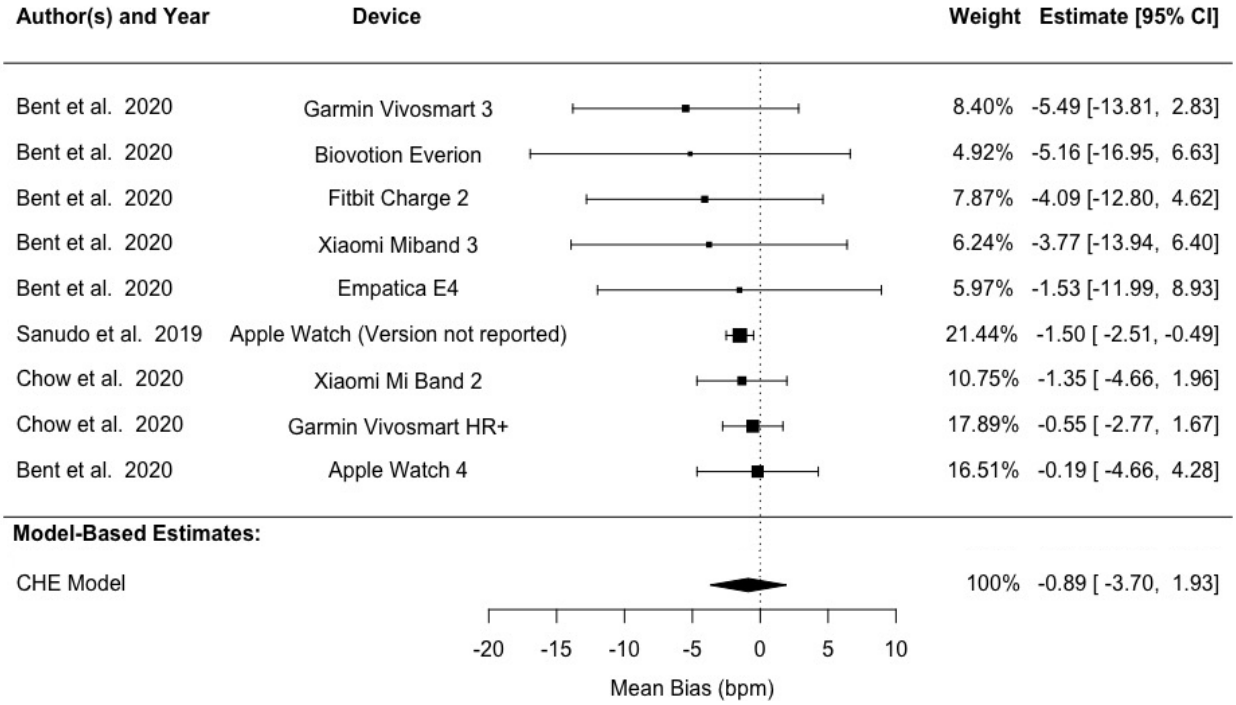

XSL•FO
RenderX

Figure 7. Forest plot showing pulse rate measurement bias in patients with dark skin pigmentation. Squares denote study weight; center of squares denotes observed study effect size; vertical lines denote study CIs; and diamond denotes pooled effect. CHE: correlated and hierarchical effect.

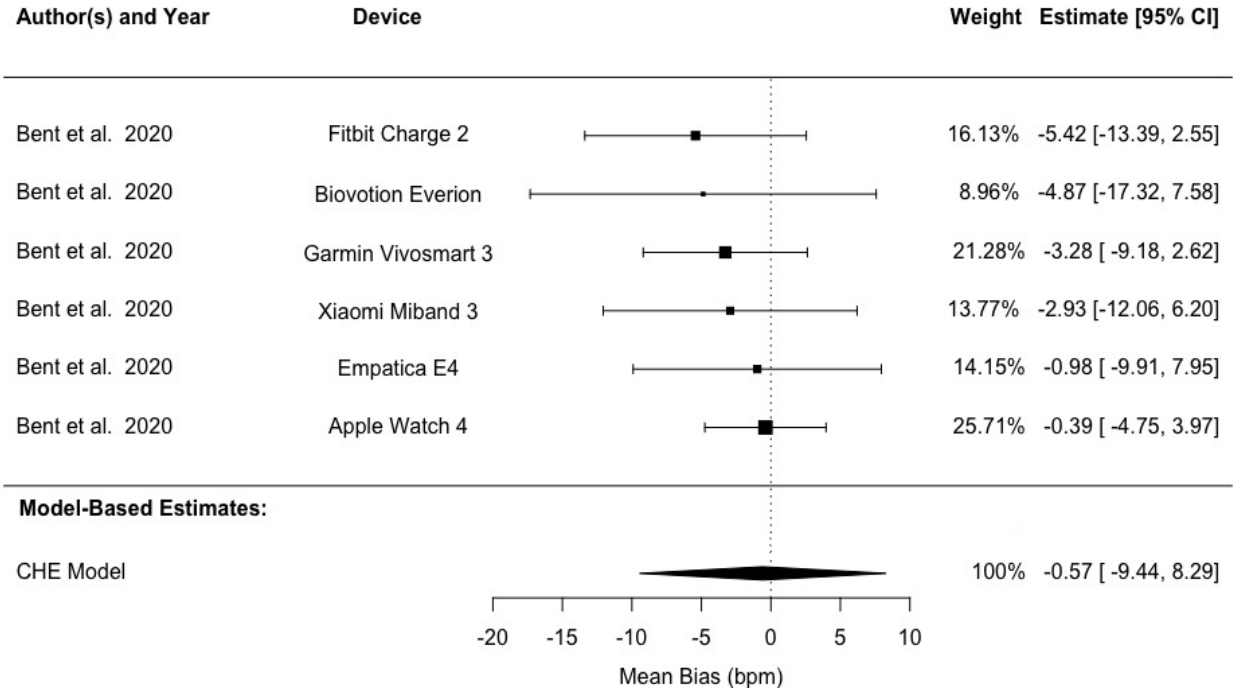

Sensitivity Analyses

Sensitivity analyses were performed for pulse oximetry SpO_2 by removing Crooks et al [48] as an outlier due to substantially large SE (Tables S6 and S7 in Multimedia Appendix 1). The pooled mean percent biases and 95% LoA were essentially unchanged, while the pooled A_{rms} across all skin pigmentation groups was 3.00%, 2.91%, and 3.40% for light, medium, and dark skin pigmentation, respectively. Sensitivity analyses using CHE models with a range of correlation coefficients (including small and moderate values) yielded similar conclusions to those obtained with a high correlation coefficient (ρ=0.9), except for the PR analysis in the light skin pigmentation group. This analysis revealed a statistically significant bias when a low correlation coefficient (ρ=0.3) was used. Details are provided in Table S6 in Multimedia Appendix 1.

Discussion

Principal Findings

The study revealed a paucity of studies properly assessing skin pigmentation, a tendency for uneven distribution across different skin pigmentation groups (overrepresentation of light skin for SpO_2 and a single study with representation of dark skin for PR), and lack of consumer-device reporting despite growing use in clinical setting. Results suggest inaccurate SpO_2 and PR measurements across all skin pigmentation groups as they breach FDA and ANSI/AAMI/IEC standards, respectively, with wearable accuracy varying considerably depending on the model, which may be due to date of model production or algorithm development. Pulse oximeters may also overestimate SpO_2 significantly for light and dark skin pigmentation, but without clinically relevant bias. We did not find statistically significant or clinically relevant bias in wearable PR devices.

Despite not meeting FDA guidance across all groups, pulse oximeter SpO_2 was inaccurate only across medium and dark skin pigmentation groups when compared to the more liberal international thresholds. Additionally, in the sensitivity analyses without the outlier study of Crooks et al [48], all pooled A_{rms} values dropped, resulting in inaccurate pulse oximeter SpO_2 only for dark skin pigmentation and no group exceeding international thresholds.

The results showing pulse oximeters significantly overestimating SpO_2 were expected for dark pigmentation and supported by findings on patient outcomes [11-15,17], but overestimated values for light pigmentation were unexpected. Two possible reasons come to mind. First, less melanin in lighter skin could distort the PPG signal [18]. Second, devices calibrated on individuals with medium skin pigmentation (note that 48% of the US population is categorized as Fitzpatrick Skin Tone Scale III [66]) may lead to inaccurate readings for both lighter and darker skin pigmentation, since both may be suboptimally represented during algorithm training and testing.

Lastly, it should be noted that both SpO_2 and PR studies had lower between-study heterogeneity and higher within-study heterogeneity, indicating that the studies that were included in these analyses were largely consistent with one another and that most of the variation came from within studies potentially due to higher variability across devices used or participants enrolled within each study.

Overall, these findings suggest that when pulse oximetry devices are deployed in their setting of intended use (ie, uncontrolled settings, such real-world medical settings and home environments with diverse patient populations), the performance observed in analytical validation studies may not generalize.

XSL•FO
RenderX

Strengths and Limitations

There were a few limitations, mostly from limitations inherent in prior studies, that should be noted. First, our skin pigmentation categorization approach has strengths but also limitations. It was an effort to overcome the reporting heterogeneity and the tendency in the published literature to conflate data collection on race, ethnicity, and skin tone. This is particularly problematic as skin tone is a physiological concept (determined by the melanin amount in the basal layer of the epidermis), while ethnicity and race are largely social constructs, with high underlying physiological heterogeneity [66]. To reduce heterogeneity, this meta-analysis reclassified race, ethnicity, and skin tone into a universal schema for skin pigmentation based on the system used by Shi et al [36]. This method, however, classifies most White people in the United States as light rather than medium skin pigmentation [66]. Second, there were only 4 prior PR studies that collected participant skin tone that also reported on device accuracy and only 1 that had participants with dark pigmentation. Third, our study used stated FDA and ANSI/AAMI/IEC standards as set thresholds to gauge device performance. It is possible that guidelines and thresholds cited in our study may change in the future, potentially limiting the applicability of the conclusions drawn in this paper. Fourth, our study chose to group papers using a variety of patient populations and testing methodologies with the goal of aggregating the largest pool of data possible on which to draw conclusions across multiple contexts. As described above, low between-study heterogeneity indicates that the studies used in this meta-analysis were largely consistent with one another. This reduces the likelihood of a moderating variable having a significant effect between studies. Instead, it suggests that most of the variation originated within studies, possibly due to higher variability across devices used or participants enrolled within each study. Despite this, we conducted subgroup analyses to examine the potential moderating variables of medical versus healthy populations for pulse oximeters (see Table S8 in Multimedia Appendix 1). We did not conduct these analyses for pulse rate because most studies did not specify whether subjects were healthy or part of medical populations. We did not examine the potential moderating effects of sensor type, transmittance versus reflectance sensors, wavelength of light used by sensors, and location of sensor placement given the core sensor technology studied was PPG across all included studies. The level of SpO_2 was not examined given this study's goal of gathering the largest sample upon which to study the effect of skin tone. Participant activity level was not evaluated as a moderator variable because many included studies did not delineate participant activity level explicitly. Future studies should evaluate the effect of these and other moderating variables on SpO_2 and PR.

Evidence Generation Guidelines for Future Analytical Validation Studies

The 2013 FDA guidance may be insufficient to ensure accuracy in pulse oximeters across all skin pigmentation and settings of intended use [19]. But there are now multiple FDA guidances for digital health tools requiring fit-for-purpose evidence as well as growing concern/guidance on clinical research diversity [4,30,31,67-73]. The FDA currently categorizes consumer devices as low-risk wellness products, exempting them from stringent regulatory oversight. However, as these devices are integrated into clinical decision-making and used as tools in clinical research [3,4], it becomes crucial to understand and communicate their advantages and limitations. Increased reliance on consumer devices increases the demand for accurate devices whose performance features and potential impact on health outcomes are known with transparency. To generate fit-for-purpose evidence applicable to diverse population, here, we propose 5 recommendations based on FDA guidance and literature (Textbox 1) [4,30,31,67-73]:

Textbox 1. Recommendations for future analytical validation studies.

Recommendation 1

It is vitally important for medical pulse oximetry devices as well as nonregulated research and consumer devices to incorporate the V3 framework [74] for sensor verification and analytical validation of derived SpO_2 and PR values for regulatory submission before these metrics can be responsibly deployed in medical, consumer, and research settings.

Recommendation 2

When possible, analytically validate devices in settings of intended use [74], rather than relying on controlled laboratory settings where digits may be warmed prior to testing, and confirm device accuracy in all subgroups (sex, race, skin pigmentation, healthy vs medical populations).

Recommendation 3

Use objective measures of skin pigmentation, rather than relying on race and ethnicity, as this will reduce heterogeneity in studies and allow for a more accurate understanding of how skin pigmentation impacts device performance.

Recommendation 4

Industry should set a priori maximum allowable difference thresholds using FDA and ANSI guidelines, properly power each subgroup, and require that 95% LoA fit within these standards for each subgroup and in the setting of intended use before receiving regulatory approval, production, and deployment.

Recommendation 5

Future studies should report device and firmware versions, as firmware updates may include changes in underlying algorithms influencing accuracy of metric generation, as described previously in the literature [4].

XSL•FO
RenderX

Conclusions

PPG has been applied in clinical practice for decades, and its accuracy in patients with different skin pigmentation has long been in question. Whether this technology contributes to diagnostic biases and by how much is only more pressing for clinicians and patients with the advent of consumer wearable PPG sensors and the growing interest and incorporation of these devices into clinical practice and in clinical research. This systematic review and meta-analysis found that pulse oximeter SpO_2 and wearable PR were inaccurate across all skin pigmentation groups as the resulting accuracy values breached FDA guidance and ANSI/AAMI/IEC standard thresholds, respectively, although pulse oximeter SpO_2 was only found to be inaccurate for dark skin pigmentation in sensitivity analyses. In addition, despite not exceeding clinically relevant bias thresholds, pulse oximeters were found to significantly overestimate SpO_2 for light and dark skin pigmentation. No systematic or clinically relevant bias was found in estimation of PR. The recommendations in this paper can help advise patients, study participants, care providers, device manufacturers, application developers, researchers, and legislators on best practices going forward.

Data Availability

Open code and data are available on Open Science Framework [75] (see the Materials—Open Code and Data section in Multimedia Appendix 1).

Authors' Contributions

SS contributed to study concept and design, data collection, and writing portions of the *Methods* section of the manuscript. HG contributed to study concept and design, data collection, and writing portions of the *Methods* and *Discussion* sections of the manuscript. BWN contributed to study concept and design; data collection; data analysis and interpretation; and writing portions of the *Abstract*, *Introduction*, *Methods*, *Results*, and *Discussion* sections of the manuscript. MRB contributed to data collection and writing portions of the *Methods* section of the manuscript. CC contributed to statistical analysis method determination, code review, and writing portions of the *Methods* section of the manuscript. All authors (SS, MRB, CC, SS, HG, and BWN) reviewed the final manuscript. HG and BWN contributed to this project as shared senior authors.

Conflicts of Interest

BWN, MRB, SS, and CC report past or current employment and/or equity ownership in Verily Life Sciences. HG serves or has served as a consultant for Verily Life Sciences, Boston Scientific, Huxley Medical, and Happitech.

Multimedia Appendix 1
Supplementary information.
[DOCX File , 619 KB - jmir_v26i1e62769_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 89 KB - jmir_v26i1e62769_app2.pdf]

References

1. Kyriacou PA, Allen J. Photoplethysmography: Technology, Signal Analysis and Applications. Cambridge, MA: Academic Press; 2021.
2. Knowles M, Krasniansky A, Nagappan A. Consumer adoption of digital health in 2022: moving at the speed of trust. Rock Health. 2023 Feb 21. URL: https://rockhealth.com/insights/consumer-adoption-of-digital-health-in-2022-moving-at-the-speed-of-trust/ [accessed 2023-03-01]
3. Al-Alusi MA, Khurshid S, Wang X, Venn RA, Pipilas D, Ashburner JM, et al. Trends in consumer wearable devices with cardiac sensors in a primary care cohort. Circ Cardiovasc Qual Outcomes 2022;15(7):e008833 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.121.008833] [Medline: 35758032]
4. Nelson BW, Low CA, Jacobson N, Areán P, Torous J, Allen NB. Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. NPJ Digit Med 2020;3:90 [FREE Full text] [doi: 10.1038/s41746-020-0297-4] [Medline: 32613085]
5. Khan M, Pretty CG, Amies AC, Elliott R, Shaw GM, Chase JG. Investigating the effects of temperature on photoplethysmography. IFAC-PapersOnLine 2015;48(20):360-365. [doi: 10.1016/j.ifacol.2015.10.166]
6. Fine J, Branan KL, Rodriguez AJ, Boonya-Ananta T, Ajmal, Ramella-Roman JC, et al. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. Biosensors (Basel) 2021;11(4):126 [FREE Full text] [doi: 10.3390/bios11040126] [Medline: 33923469]

XSL•FO
RenderX

7. Ries AL, Prewitt LM, Johnson JJ. Skin color and ear oximetry. Chest 1989;96(2):287-290. [doi: 10.1378/chest.96.2.287] [Medline: 2752811]
8. Feiner JR, Severinghaus JW, Bickler PE. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. Anesth Analg 2007;105(6 Suppl):S18-S23. [doi: 10.1213/01.ane.0000285988.35174.d9] [Medline: 18048893]
9. Zeballos RJ, Weisman IM. Reliability of noninvasive oximetry in black subjects during exercise and hypoxia. Am Rev Respir Dis 1991;144(6):1240-1244. [doi: 10.1164/ajrccm/144.6.1240] [Medline: 1741533]
10. Bickler PE, Feiner JR, Severinghaus JW. Effects of skin pigmentation on pulse oximeter accuracy at low saturation. Anesthesiology 2005;102(4):715-719 [FREE Full text] [doi: 10.1097/00000542-200504000-00004] [Medline: 15791098]
11. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2020;383(25):2477-2478 [FREE Full text] [doi: 10.1056/NEJMc2029240] [Medline: 33326721]
12. Valbuena VSM, Seelye S, Sjoding MW, Valley TS, Dickson RP, Gay SE, et al. Racial bias and reproducibility in pulse oximetry among medical and surgical inpatients in general care in the veterans health administration 2013-19: multicenter, retrospective cohort study. BMJ 2022;378:e069775 [FREE Full text] [doi: 10.1136/bmj-2021-069775] [Medline: 35793817]
13. Gottlieb ER, Ziegler J, Morley K, Rush B, Celi LA. Assessment of racial and ethnic differences in oxygen supplementation among patients in the intensive care unit. JAMA Intern Med 2022;182(8):849-858 [FREE Full text] [doi: 10.1001/jamainternmed.2022.2587] [Medline: 35816344]
14. Okunlola OE, Lipnick MS, Batchelder PB, Bernstein M, Feiner JR, Bickler PE. Pulse oximeter performance, racial inequity, and the work ahead. Respir Care 2022;67(2):252-257 [FREE Full text] [doi: 10.4187/respcare.09795] [Medline: 34772785]
15. Wong AI, Charpignon M, Kim H, Josef C, de Hond AAH, Fojas JJ, et al. Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. JAMA Netw Open 2021;4(11):e2131674 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.31674] [Medline: 34730820]
16. Fawzy A, Wu TD, Wang K, Sands KE, Fisher AM, Arnold Egloff SA, et al. Clinical outcomes associated with overestimation of oxygen saturation by pulse oximetry in patients hospitalized with COVID-19. JAMA Netw Open 2023;6(8):e2330856 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30856] [Medline: 37615985]
17. Fawzy A, Wu TD, Wang K, Robinson ML, Farha J, Bradke A, et al. Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19. JAMA Intern Med 2022;182(7):730-738 [FREE Full text] [doi: 10.1001/jamainternmed.2022.1906] [Medline: 35639368]
18. Keller MD, Harrison-Smith B, Patil C, Arefin MS. Skin colour affects the accuracy of medical oxygen sensors. Nature 2022;610(7932):449-451. [doi: 10.1038/d41586-022-03161-1] [Medline: 36261563]
19. Pulse oximeters - premarket notification submissions [510(k)s] guidance for industry and Food and Drug Administration staff. U.S. Food and Drug Administration. 2013 Mar 04. URL: https://www.fda.gov/media/72470/download [accessed 2024-08-24]
20. Review of pulse oximeters and factors that can impact their accuracy. U.S. Food and Drug Administration. 2022. URL: https://www.fda.gov/media/162709/download [accessed 2024-08-24]
21. McFarling UL. FDA panel asks for improvements in pulse oximeters. STAT News. 2022. URL: https://www.statnews.com/2022/11/01/fda-panel-asks-for-improvements-in-pulse-oximeters/ [accessed 2023-01-09]
22. American Psychiatric Association. Wearable Devices as Therapy Tools. Monitor on Psychology. 2021 Sep. URL: https://www.apa.org/monitor/2021/2021-09-monitor.pdf [accessed 2023-01-10]
23. Center for Devices and Radiological Health. Pulse oximeter accuracy and limitations: FDA safety communication. U.S. Food and Drug Administration. URL: https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication [accessed 2023-01-10]
24. November 1, 2022: anesthesiology and respiratory therapy devices panel of the medical devices advisory committee meeting announcement. U.S. Food and Drug Administration. 2023 Apr 4. URL: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-1-2022-anesthesiology-and-respiratory-therapy-devices-panel-medical-devices-advisory [accessed 2023-10-05]
25. McFarling UL. Pulse oximeters' inaccuracies in darker-skinned people require urgent action, AGs tell FDA. STAT News. 2023 Nov 07. URL: https://www.statnews.com/2023/11/07/pulse-oximeters-attorneys-general-urge-fda-action/#:~:text=Pulse%20oximeters?%20overestimation%20of%20oxygen,for%20severe%20Covid%2D19%20infections [accessed 2023-11-07]
26. Attorney General letter to FDA. State of California Office of the Attorney General. 2023 Nov 01. URL: https://oag.ca.gov/system/files/attachments/press-docs/23PR353%20Health%20Equity%20General%20Matter%20Multistate.pdf [accessed 2023-11-01]
27. Zinzuwadia A, Singh JP. Wearable devices-addressing bias and inequity. Lancet Digit Health 2022;4(12):e856-e857 [FREE Full text] [doi: 10.1016/S2589-7500(22)00194-7] [Medline: 36335031]
28. Shachar C, Gerke S. Prevention of bias and discrimination in clinical practice algorithms. JAMA 2023;329(4):283-284. [doi: 10.1001/jama.2022.23867] [Medline: 36602791]
29. Goodman KE, Morgan DJ, Hoffmann DE. Clinical algorithms, antidiscrimination laws, and medical device regulation. JAMA 2023;329(4):285-286. [doi: 10.1001/jama.2022.23870] [Medline: 36602795]

XSL•FO
RenderX

30. Colvonen P, DeYoung P, Bosompra N, Owens R. Limiting racial disparities and bias for wearable devices in health science research. Sleep 2020;43(10):zsaa159 [FREE Full text] [doi: 10.1093/sleep/zsaa159] [Medline: 32893865]
31. State of California Department of Justice. Attorney General Bonta launches inquiry into racial and ethnic bias in healthcare algorithms. State of California Office of the Attorney General. 2022 Aug 31. URL: https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-inquiry-racial-and-ethnic-bias-healthcare [accessed 2023-11-08]
32. Medical electrical equipment particular requirements for basic safety and essential performance of pulse oximeter equipment (British standard). ANSI. URL: https://tinyurl.com/ayfrdpau [accessed 2023-10-23]
33. ANSI/AAMI/IEC. IEC 60601-2-27 particular requirements for the basic safety and essential performance of electro cardiographic monitoring. ITC India. URL: https://www.itcindia.org/iec-60601-2-27-particular-requirements-for-the-basic-safety-and-essential-performance-of-electro-cardiographic-monitoring/#:~:text=What%20is%20IEC%2060601%2D2,processing%2C%20alarms%2C%20and%20displays [accessed 2024-08-24]
34. ANSI/AAMI EC13-2002 cardiac monitors, heart rate meters, and alarms. ANSI. 2002. URL: https://webstore.ansi.org/standards/aami/ansiaamiec132002 [accessed 2023-10-30]
35. Koerber D, Khan S, Shamsheri T, Kirubarajan A, Mehta S. Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: a systematic review. J Racial Ethn Health Disparities 2023;10(6):2676-2684 [FREE Full text] [doi: 10.1007/s40615-022-01446-9] [Medline: 36376641]
36. Shi C, Goodall M, Dumville J, Hill J, Norman G, Hamer O, et al. The accuracy of pulse oximetry in measuring oxygen saturation by levels of skin pigmentation: a systematic review and meta-analysis. BMC Med 2022;20(1):267 [FREE Full text] [doi: 10.1186/s12916-022-02452-8] [Medline: 35971142]
37. Cabanas AM, Fuentes-Guajardo M, Latorre K, León D, Martín-Escudero P. Skin pigmentation influence on pulse oximetry accuracy: a systematic review and bibliometric analysis. Sensors (Basel) 2022;22(9):3402 [FREE Full text] [doi: 10.3390/s22093402] [Medline: 35591092]
38. Al-Halawani R, Charlton PH, Qassem M, Kyriacou PA. A review of the effect of skin pigmentation on pulse oximeter accuracy. Physiol Meas 2023;44(5):05TR01 [FREE Full text] [doi: 10.1088/1361-6579/acd51a] [Medline: 37172609]
39. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
40. Center for Devices and Radiological Health. How to determine if your product is a medical device. U.S. Food and Drug Administration. 2022 Sep 29. URL: https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device [accessed 2023-11-16]
41. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2020;3:18 [FREE Full text] [doi: 10.1038/s41746-020-0226-6] [Medline: 32047863]
42. Abrams G, Sanders MK, Fallon MB. Utility of pulse oximetry in the detection of arterial hypoxemia in liver transplant candidates. Liver Transpl 2002;8(4):391-396 [FREE Full text] [doi: 10.1053/jlts.2002.32252] [Medline: 11965585]
43. Adler JN, Hughes LA, Vivilecchia R, Camargo CA. Effect of skin pigmentation on pulse oximetry accuracy in the emergency department. Acad Emerg Med 1998;5(10):965-970 [FREE Full text] [doi: 10.1111/j.1553-2712.1998.tb02772.x] [Medline: 9862586]
44. Andrist E, Nuppnau M, Barbaro RP, Valley TS, Sjoding MW. Association of race with pulse oximetry accuracy in hospitalized children. JAMA Netw Open 2022;5(3):e224584 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.4584] [Medline: 35357460]
45. Barker SJ, Wilson WC. Racial effects on Masimo pulse oximetry: a laboratory study. J Clin Monit Comput 2023;37(2):567-574 [FREE Full text] [doi: 10.1007/s10877-022-00927-w] [Medline: 36370242]
46. Bothma PA, Joynt GM, Lipman J, Hon H, Mathala B, Scribante J, et al. Accuracy of pulse oximetry in pigmented patients. S Afr Med J 1996;86(5 Suppl):594-596. [Medline: 8914569]
47. Burnett G, Stannard B, Wax D, Lin HM, Pyram-Vincent C, DeMaria S, et al. Self-reported race/ethnicity and intraoperative occult hypoxemia: a retrospective cohort study. Anesthesiology 2022;136(5):688-696 [FREE Full text] [doi: 10.1097/ALN.0000000000004153] [Medline: 35231085]
48. Crooks CJ, West J, Morling JR, Simmonds M, Juurlink I, Briggs S, et al. Pulse oximeter measurements vary across ethnic groups: an observational study in patients with COVID-19. Eur Respir J 2022;59(4):2103246 [FREE Full text] [doi: 10.1183/13993003.03246-2021] [Medline: 35086839]
49. Ebmeier SJ, Barker M, Bacon M, Beasley RC, Bellomo R, Knee Chong C, et al. A two centre observational study of simultaneous pulse oximetry and arterial oxygen saturation recordings in intensive care unit patients. Anaesth Intensive Care 2018;46(3):297-303. [doi: 10.1177/0310057X1804600307] [Medline: 29716488]
50. Foglia EE, Whyte RK, Chaudhary A, Mott A, Chen J, Propert KJ, et al. The effect of skin pigmentation on the accuracy of pulse oximetry in infants with hypoxemia. J Pediatr 2017;182:375-377.e2 [FREE Full text] [doi: 10.1016/j.jpeds.2016.11.043] [Medline: 27939107]
51. Jubran A, Tobin MJ. Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. Chest 1990;97(6):1420-1425. [doi: 10.1378/chest.97.6.1420] [Medline: 2347228]

XSL•FO
RenderX

52. McGovern JP, Sasse SA, Stansbury DW, Causing LA, Light RW. Comparison of oxygen saturation by pulse oximetry and co-oximetry during exercise testing in patients with COPD. Chest 1996;109(5):1151-1155. [doi: 10.1378/chest.109.5.1151] [Medline: 8625659]
53. Muñoz X, Torres F, Sampol G, Rios J, Martí S, Escrich E. Accuracy and reliability of pulse oximetry at different arterial carbon dioxide pressure levels. Eur Respir J 2008;32(4):1053-1059 [FREE Full text] [doi: 10.1183/09031936.00126507] [Medline: 18480106]
54. Pilcher J, Ploen L, McKinstry S, Bardsley G, Chien J, Howard L, et al. A multicentre prospective observational study comparing arterial blood gas values to those obtained by pulse oximeters used in adult patients attending Australian and New Zealand hospitals. BMC Pulm Med 2020;20(1):7 [FREE Full text] [doi: 10.1186/s12890-019-1007-3] [Medline: 31918697]
55. Ruppel H, Makeneni S, Faerber JA, Lane-Fall MB, Foglia EE, O'Byrne ML, et al. Evaluating the accuracy of pulse oximetry in children according to race. JAMA Pediatr 2023;177(5):540-543 [FREE Full text] [doi: 10.1001/jamapediatrics.2023.0071] [Medline: 36939727]
56. Sudat SEK, Wesson P, Rhoads K, Brown S, Aboelata N, Pressman AR, et al. Racial disparities in pulse oximeter device inaccuracy and estimated clinical impact on COVID-19 treatment course. Am J Epidemiol 2023;192(5):703-713 [FREE Full text] [doi: 10.1093/aje/kwac164] [Medline: 36173743]
57. Thrush D, Hodges MR. Accuracy of pulse oximetry during hypoxemia. South Med J 1994;87(4):518-521. [doi: 10.1097/00007611-199404000-00019] [Medline: 8153783]
58. Vesoulis Z, Tims A, Lodhi H, Lalos N, Whitehead H. Racial discrepancy in pulse oximeter accuracy in preterm infants. J Perinatol 2022;42(1):79-85 [FREE Full text] [doi: 10.1038/s41372-021-01230-3] [Medline: 34642469]
59. Wiles MD, El-Nayal A, Elton G, Malaj M, Winterbottom J, Gillies C, et al. The effect of patient ethnicity on the accuracy of peripheral pulse oximetry in patients with COVID-19 pneumonitis: a single-centre, retrospective analysis. Anaesthesia 2022;77(2):143-152 [FREE Full text] [doi: 10.1111/anae.15581] [Medline: 34542168]
60. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019;7(3):e10828 [FREE Full text] [doi: 10.2196/10828] [Medline: 30855232]
61. Sañudo B, de Hoyo M, Muñoz-López A, Perry J, Abt G. Pilot study assessing the influence of skin type on the heart rate measurements obtained by photoplethysmography with the apple watch. J Med Syst 2019;43(7):195. [doi: 10.1007/s10916-019-1325-2] [Medline: 31119387]
62. Chow H, Yang C. Accuracy of optical heart rate sensing technology in wearable fitness trackers for young and older adults: validation and comparison study. JMIR Mhealth Uhealth 2020;8(4):e14707 [FREE Full text] [doi: 10.2196/14707] [Medline: 32343255]
63. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Soft 2010;36(3):1-48.
64. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. Prev Sci 2022;23(3):425-438. [doi: 10.1007/s11121-021-01246-3] [Medline: 33961175]
65. Pustejovsky J. clubSandwich: cluster-robust (sandwich) variance estimators with small-sample corrections. R Project. URL: https://CRAN.R-project.org/package=clubSandwich [accessed 2024-05-29]
66. Keiser E, Linos E, Kanzler M, Lee W, Sainani KL, Tang JY. Reliability and prevalence of digital image skin types in the United States: results from national health and nutrition examination survey 2003-2004. J Am Acad Dermatol 2012;66(1):163-165. [doi: 10.1016/j.jaad.2011.02.044] [Medline: 22177642]
67. Sjoding MW, Iwashyna TJ, Valley TS. Change the framework for pulse oximeter regulation to ensure clinicians can give patients the oxygen they need. Am J Respir Crit Care Med 2023;207(6):661-664 [FREE Full text] [doi: 10.1164/rccm.202209-1773ED] [Medline: 36260769]
68. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Women in Science, Engineering, and Medicine, Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials and Research. In: Bibbins-Domingo K, Helman A, editors. Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups. Washington, DC: National Academies Press; Jul 13, 2022.
69. Lee NT, Resnick P, Barton G. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings. 2019 May 22. URL: https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ [accessed 2023-11-08]
70. Framework for the use of digital health technologies in drug and biological product development. U.S. Food and Drug Administration. 2023 Mar. URL: https://www.fda.gov/media/166396/download?attachment [accessed 2024-08-24]
71. Center for Drug Evaluation and Research. Digital health technologies for remote data acquisition in clinical investigations. U.S. Food and Drug Administration. 2023 Dec. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations [accessed 2023-12-05]
72. Office of the Commissioner. Diversity plans to improve enrollment of participants from underrepresented racial and ethnic populations in clinical trials; draft guidance for industry. U.S. Food and Drug Administration. URL: https://www.fda.gov/

XSL•FO
RenderX

regulatory-information/search-fda-guidance-documents/diversity-plans-improve-enrollment-participants-underrepresented-racial-and-ethnic-populations [accessed 2023-12-05]
73. Center for Drug Evaluation and Research. Enhancing the diversity of clinical trial populations — eligibility criteria, enrollment practices, and trial designs guidance for industry. U.S. Food and Drug Administration. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial [accessed 2023-11-30]
74. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med 2020;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
75. Singh S, Bennett M, Chen C, Ghanbari H, Nelson B. Photoplethysmography pulse oximetry and pulse rate accuracy by skin tone: a meta-analysis. Open Science Framework. URL: http://osf.io/qngmz/ [accessed 2024-09-30]

Abbreviations

AAMI: Association for the Advancement of Medical Instrumentation
ANSI: American National Standards Institute
A_{rms}: accuracy root-mean-square
bpm: beats per minute
CHE: correlated and hierarchical effect
ECG: electrocardiography
FDA: Food and Drug Administration
IEC: International Electrotechnical Commission
LoA: limits of agreement
PPG: photoplethysmography
PR: pulse rate
PRIMSA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RVE: robust variance estimation
SaO_2: arterial blood gas
SpO_2: blood oxygenation

Edited by G Eysenbach, A Mavragani; submitted 04.06.24; peer-reviewed by U Sinha, T Nagamine; comments to author 04.07.24; revised version received 25.07.24; accepted 16.08.24; published 10.10.24.

Please cite as:
Singh S, Bennett MR, Chen C, Shin S, Ghanbari H, Nelson BW
Impact of Skin Pigmentation on Pulse Oximetry Blood Oxygenation and Wearable Pulse Rate Accuracy: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e62769
URL: https://www.jmir.org/2024/1/e62769
doi:10.2196/62769
PMID:

©Sanidhya Singh, Miles Romney Bennett, Chen Chen, Sooyoon Shin, Hamid Ghanbari, Benjamin W Nelson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Spaced Digital Education for Health Professionals: Systematic Review and Meta-Analysis

Laura Martinengo[1,2], MD, PhD; Matthew Song Peng Ng[3], MBBS; Tony De Rong Ng[3], MBBS; Yi-Ian Ang[3], MBBS; Ahmad Ishqi Jabir[1,4], BSc, MSc; Bhone Myint Kyaw[5], MBBS, MSc, PhD; Lorainne Tudor Car[1,6], MD, PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore, Singapore

[2]Centre for Behavioural and Implementation Sciences Interventions, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[4]Future Health Technologies, Singapore-ETH Centre, Campus for Research Excellence And Technological Enterprise, Singapore, Singapore

[5]Nanyang Technological University Singapore, Singapore, Singapore

[6]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Lorainne Tudor Car, MD, PhD
Lee Kong Chian School of Medicine
Nanyang Technological University Singapore
11 Mandalay Rd
Level 18, Clinical Sciences Builiding
Singapore, 308232
Singapore
Phone: 65 69041258
Email: l.tudor.car@imperial.ac.uk

Abstract

Background: Spaced digital education applies digital tools to deliver educational content via multiple, repeated learning sessions separated by prespecified time intervals. Spaced digital education appears to promote acquisition and long-term retention of knowledge, skills, and change in clinical behavior.

Objective: The aim of this review was to assess the effectiveness of spaced digital education in improving pre- and postregistration health care professionals' knowledge, skills, attitudes, satisfaction, and change in clinical behavior.

Methods: This review followed Cochrane's methodology and PRISMA (Preferred Reporting Items of Systematic Reviews and Meta-Analyses) reporting guidelines. We searched MEDLINE, Embase, Web of Science, ERIC, PsycINFO, CINAHL, CENTRAL, and ProQuest Dissertation and Theses databases from January 1990 to February 2023. We included randomized controlled trials (RCTs), cluster RCTs, and quasi-RCTs comparing spaced digital education with nonspaced education, spaced nondigital education, traditional learning, or no intervention for pre- or postregistration health care professionals. Study selection, data extraction, study quality, and certainty of evidence were assessed by 2 independent reviewers. Meta-analyses were conducted using random effect models.

Results: We included 23 studies evaluating spaced online education (n=17, 74%) or spaced digital simulation (n=6, 26%) interventions. Most studies assessed 1 or 2 outcomes, including knowledge (n=15, 65%), skills (n=9, 39%), attitudes (n=8, 35%), clinical behavior change (n=8, 35%), and satisfaction (n=7, 30%). Most studies had an unclear or a high risk of bias (n=19, 83%). Spaced online education was superior to massed online education for postintervention knowledge (n=9, 39%; standardized mean difference [SMD] 0.32, 95% CI 0.13-0.51, I^2=66%, moderate certainty of evidence). Spaced online education (n=3, 13%) was superior to massed online education (n=2, 9%) and no intervention (n=1, 4%; SMD 0.67, 95% CI 0.43-0.91, I^2=5%, moderate certainty of evidence) for postintervention clinical behavior change. Spaced digital simulation was superior to massed simulation for postintervention surgical skills (n=2, 9%; SMD 1.15, 95% CI 0.34-1.96, I^2=74%, low certainty of evidence). Spaced digital education positively impacted confidence and satisfaction with the intervention.

Conclusions: Spaced digital education is effective in improving knowledge, particularly in substantially improving surgical skills and promoting clinical behavior change in pre- and postregistration health care professionals. Our findings support the use

XSL•FO
RenderX

of spaced digital education interventions in undergraduate and postgraduate health profession education. Trial Registration: PROSPERO CRD42021241969;

Trial Registration: PROSPERO CRD42021241969; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=241969

(J Med Internet Res 2024;26:e57760) doi:10.2196/57760

KEYWORDS

digital education; e-learning; spaced education; spaced learning; spaced digital education; spaced simulation; health care professionals; continuous medical education; systematic review; meta-analysis

Introduction

Background

Health care service delivery depends on access to a competent workforce. The World Health Organization (WHO) has long warned of severe global shortages of health care workers, particularly in low- and middle-income countries [1,2]. Long working hours, low wages, workplace stress, and lack of professional development opportunities [3,4] appear to drive the health care workforce turnover [3]. An important strategy to increase recruitment and retention of health care workers is providing high-quality education in an affordable, effective, and sustainable way [5].

Digital education is "the act of teaching and learning by means of digital technologies" [6]. Digital education may promote continuous professional development and improve health care workers' competencies by offering convenient and adaptable learning tools that can be accessed at any place and time [6-10]. A recent series of systematic reviews and meta-analyses on different digital education modalities [6,11-18] showed that digital education is well received and noninferior to traditional education.

Spaced Education

Spaced education is an educational technique that promotes long-term knowledge retention via multiple repeated learning sessions separated by prespecified time intervals [19-21]. Spacing knowledge acquisition is a well-studied and effective learning technique, particularly if it is associated with the active retrieval of information [22]. The beneficial effects of spacing and retrieval appear to be independent of the age of students and the length of the spacing interval, although research seems to suggest that longer spacing intervals are correlated to longer-term knowledge retention [22,23]. Similarly, the effectiveness of spaced education appears to be comparable in online, simulation, and classroom settings [24] or in studies assessing the acquisition of theoretical knowledge and practical skills [22,24]. Spaced education has been increasingly used in undergraduate [25] and postgraduate [24] medical education to improve learners' knowledge [26] or skills, such as resuscitation in emergency medicine [27,28], surgical techniques [29], and image interpretation in radiology [30]. Spaced education has also been shown to promote changes in clinical practice [31,32]. By combining the accessibility and flexibility of digital education with the benefits of spaced learning in knowledge retention, spaced digital education might offer a feasible, scalable, and effective solution to providing high-quality education for health care professionals [33].

A recent systematic review [31] assessed the effectiveness of spaced education in continuous development programs for practicing health care professionals, suggesting an improvement in knowledge, skills, clinical behavior, and confidence. However, the review included a variety of research designs and offered a narrative summary of findings. Given the increasing relevance of spaced education in health profession education, particularly when delivered in digital format, we considered it important to update and expand Phillips et al's [31] systematic review. The aim of this systematic review was to assess the effectiveness of spaced digital education in improving pre- and postregistration health care professionals' knowledge, skills, attitudes, satisfaction, and clinical behavior change.

Methods

Study Design

The review followed Cochrane's methodology [34] and was reported according to the PRISMA (Preferred Reporting Items of Systematic Reviews and Meta-Analyses) checklist [35]. The protocol was registered on the International Prospective Register of Systematic Reviews (PROSPERO; CRD42021241969) on April 14, 2021.

Identification of Studies

The search strategy was developed for a series of systematic reviews on the effectiveness of digital health education interventions for the education and training of students and professionals of health care–related professions [5,12-14,17,18,36-47]. The search strategy was developed in collaboration with librarians and information specialists from Karolinska Instituet and has been regularly updated. We searched MEDLINE, CENTRAL, Embase, Web of Science, ERIC, PsycINFO, CINAHL, and ProQuest Dissertation and Theses databases from January 1990 to February 2023. We complemented the search with forward and backward citation searches in all included studies and 2 relevant reviews [24,31]. Multimedia Appendix 1 presents the MEDLINE search strategy.

Eligibility Criteria

This systematic review included randomized controlled trials (RCTs), cluster randomized controlled trials (cRCTs), and quasi-RCTs reporting spaced digital education interventions for health profession education. Randomized cross-over trials were excluded due to the high risk of contamination and carryover effects in educational interventions.

Study participants included preregistration students and postregistration professionals from health professions, such as medicine, dentistry, nursing and midwifery, medical diagnostic

XSL•FO
RenderX

and treatment technology, physiotherapy and rehabilitation, and pharmacy, as defined by in the Health Field of Education and Training (091) of the International Standard Classification of Education (ISCED-F) [48]. Practitioners of traditional, alternative, and complementary medicine were excluded.

We included studies comparing spaced digital education interventions with nonspaced digital education modalities, spaced nondigital education, traditional face-to-face learning, blended learning, or no intervention. We included 2 spaced digital education modalities in this review. Spaced online education was defined for this study as an educational intervention where information, such as case studies or multiple-choice questions, "is presented and repeated over spaced intervals" [24,49] using a digital device, such as a computer, smartphone, or tablet computer. Spaced digital simulation was defined as studies using high-fidelity simulators, including virtual reality (VR)–based instruments, to provide skills training and assessment that was repeated over spaced intervals. There may be 2 or more repetitions of the learning content that should occur in different days.

Studies were excluded if both intervention and comparison groups described a spaced digital education intervention, if the spacing interval was shorter than 1 day, and if the materials were not repeated over the course of the intervention.

Outcomes Measured

We assessed the following *primary outcomes*, measured using validated or nonvalidated instruments:

- Postintervention knowledge
- Postintervention skills
- Postintervention attitudes or perceptions toward the intervention, as well as the interaction with patients and colleagues
- Learners' satisfaction with the intervention
- Postintervention behavior change in the way learners interact with patients or colleagues or modify their practices

We also included the following *secondary outcomes*: patient-related outcomes, cost and cost-effectiveness of the intervention, and negative effects of implementing a spaced education intervention.

Data Collection and Analysis

Selection of Studies

The title and abstract screening for previous updates of the library was performed by pairs of reviewers working independently and in parallel. However, for the February 2023 update of the search strategy, screening was performed by 1 reviewer (author LM) on ASReview [50,51], an open source screening tool using several active learning and machine learning techniques. Following ASReview guidelines, we defined stopping criteria of 1.04% (90/8634 citations) consecutive "irrelevant" records and screening between 8% and 33% of all citations. The stopping criterion for this review was reached after screening 2850 (33.01%) records. Full-text screening for the entire digital health education library was performed by groups of 2 independent reviewers (authors MSPN, TDRN, and YIA) working in parallel. The data extraction and quality assessment of the included studies were also performed by groups of 2 independent reviewers. Differences between reviewers were resolved by discussion or by consultation with a third reviewer (LM) acting as an arbiter.

Data Extraction and Quality Assessment

Data from the included studies were extracted using a standardized Microsoft Excel form. We extracted information about the study design, participants, characteristics of the digital health education intervention, primary and secondary outcomes, and pre- and postintervention results, as reported by the primary studies.

We assessed the methodological quality of the included studies using Cochrane's Risk of Bias Tool 2 (RoB-2) [52,53], which assesses the risk of bias arising from the randomization process, deviations from the intended interventions, missing outcome data, measurement of the outcome, and selection of the outcome result. For cRCTs, we also assessed the bias arising from the timing of identification or recruitment of participants into clusters. Studies were assessed as having a high risk of bias if at least 1 category was assessed as high risk or multiple domains were assessed as "some concerns." Studies with 1 domain assessed as "some concerns" were categorized as "some concerns" risk of bias. Finally, studies with all domains assessed as low risk were classified as having a low risk of bias. The risk-of-bias figures were created using robvis, an online tool [54].

Grading the Quality of the Evidence

We assessed the quality of evidence and presented it as a summary-of-findings table (Multimedia Appendix 2) for 3 comparisons: spaced online education versus no intervention, spaced online education versus massed online education, and spaced digital simulation versus massed simulation. The evaluation followed the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) criteria; assessed the limitation of the studies (risk of bias), inconsistency, indirectness, and imprecision; and rated the quality of evidence as very low, low, moderate, or high [55]. The assessment was performed using GRADE profiler (GRADEpro) software [55].

Statistical Analysis

Study data were analyzed using Cochrane's Review Manager (RevMan) 5.4.1 [56]. We used the immediate postintervention results in the analyses. To ensure consistent reporting, for outcomes reported as continuous variables, we calculated the standardized mean difference (SMD) and 95% CIs. If studies reported *P* values, we calculated the SD using RevMan's inbuilt calculator [56]. If studies reported the median (IQR), we converted those values to means (SDs) following Luo et al [57] and Wan et al [58]. In studies reporting 2 or more groups, for example, studies with more than 1 intervention arm, we combined the groups before the analysis using the *Cochrane Handbook for Systematic Reviews of Interventions* guidelines [59]. Subsequently, we calculated the SMD and 95% CIs, as per the protocol. Effect size interpretation followed the Hedges' g statistic and considered an intervention to have a small effect if the SMD was between 0.2 and 0.5, a moderate effect if it was between 0.5 and 0.8, and a large effect if it was 0.8 and above.

XSL•FO
RenderX

For studies reporting dichotomous data, we calculated the relative risks (RRs) and associated 95% CIs across studies. For cRCTs, we extracted individual-level data, as reported in the primary studies. If relevant outcome data were missing from the primary studies, we contacted the authors to obtain information. We did not input any missing data. When possible, data were analyzed on an intention-to-treat basis.

We described our findings narratively using Miller's classification of clinical competence [60] to classify the outcomes. When possible, we performed a meta-analysis using a random effects model. We used the I^2 statistic to evaluate heterogeneity, with I^2<25%, 25%-75%, and >75% representing low, moderate, and high heterogeneity, respectively [34].

We attempted to assess publication bias and perform subgroup analyses; however, the small number of studies in each category and the heterogeneity of data precluded these assessments.

Results

Characteristics of Included Studies

The search strategy yielded 85,755 publications, of which 23 (0.03%) studies, reported in 22 papers (0.03%), were included in the review. The study selection process was summarized using a PRISMA flow diagram (Figure 1) [35]. Multimedia Appendix 3 also presents a summary of the included studies, and Multimedia Appendix 4 lists the studies excluded after reviewing the full text.

Figure 1. Study selection flowchart.

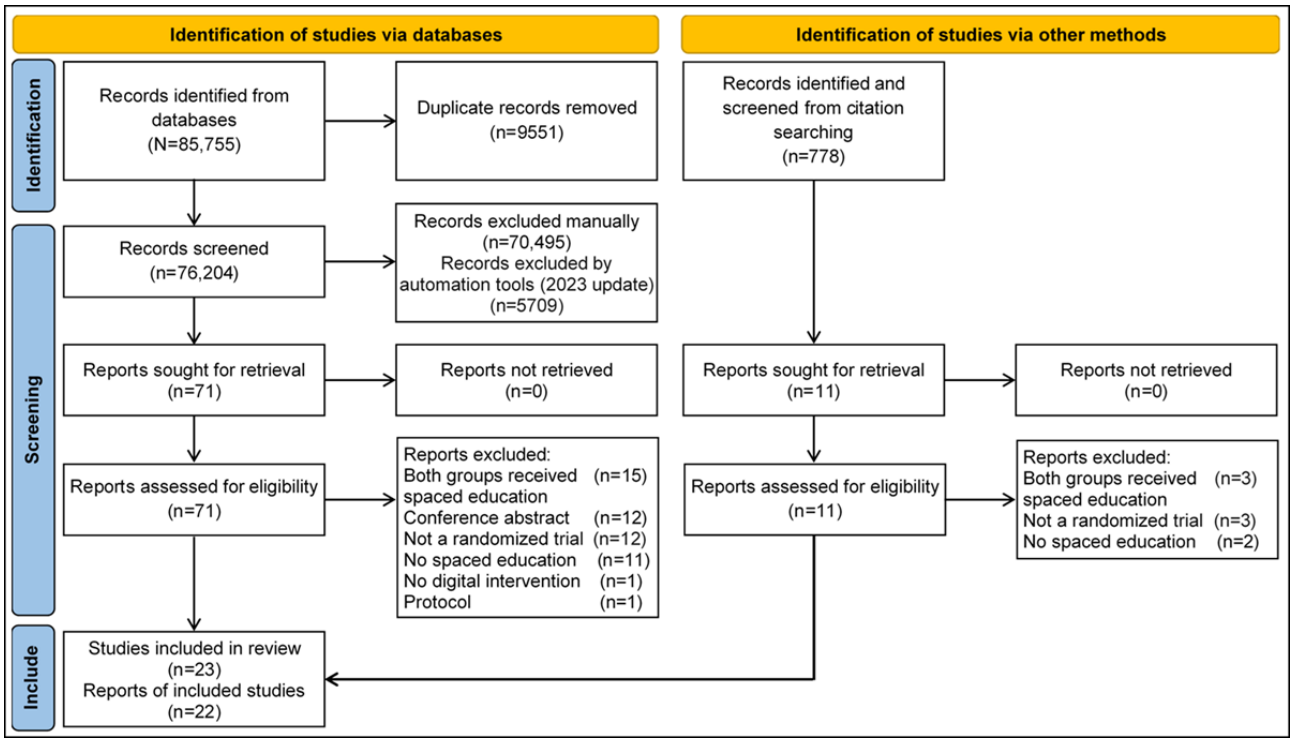

Table 1 presents the characteristics of the included studies. Most studies (n=19, 83%) were RCTs. Participants were randomized into 2 groups, except in 1 (4%) study [61], which randomized participants into 3 groups. Most studies (n=19, 83%) were conducted in high-income countries [49,61-77], 3 (13%) studies were conducted in upper middle–income countries [78-80], and only 1 (4%) study was conducted in a lower middle–income country [81]. The studies included a total of 3371 participants. Study populations comprised postregistration health professionals (n=20, 87%) and preregistration medical [61,76] and nursing [79] students (n=3, 13%). Of the 23 studies, 9 (39%) [61,62,65,66,69,72,74,78,80] reported 1 and 7 (30%) [49,63,67,70,73,76,77] reported 2 primary outcomes. The most common outcome was knowledge (n=15, 65%) [49,62-68,72-75,77,81], followed by skills (n=9, 39%) [61,69,70,72,76,78-81], attitudes (n=8, 35%) [64,68,71-73,75,77,79], behavior change (n=8, 35%) [49,63,67,68,71,75,77], and satisfaction with the intervention (n=7, 31%) [64,70-72,75,76,81].

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Characteristics	Studies (N=23), n (%)
Year of publication	
Before 2019	13 (57)
After 2019	10 (43)
Country	
United States	14 (61)
Türkiye	2 (9)
Australia	1 (4)
Canada	1 (4)
China	1 (4)
France	1 (4)
Nigeria	1 (4)
Singapore	1 (4)
United States and Canada	1 (4)
Study design	
RCT[a]	19 (83)
cRCT[b]	3 (13)
Quasi-RCT	1 (4)
Participants	
Students of medicine	2 (9)
Students of nursing	1 (4)
Medical and surgical specialty interns and residents (trainee physicians)	3 (13)
Medical specialty residents (trainee physicians)	4 (17)
Surgical specialty residents (trainee physicians)	5 (22)
Practicing physicians	2 (9)
Multidisciplinary primary care providers	6 (26)
Clinical domain: medical specialties	
Internal medicine	2 (9)
Oncology	1 (4)
Pediatrics	2 (9)
Primary care	5 (22)
Clinical domain: surgical specialties	
General surgery	2 (9)
Obstetrics and gynecology	2 (9)
Orthopedics	2 (9)
Urology	2 (9)
Clinical domain: medical and surgical specialties	2 (9)
Intervention: spaced online education	
Via email	10 (43)
Using proprietary software (QStream/Moodle)	4 (17)
Via email with gamified elements	1 (4)
Via mobile apps	1 (4)

XSL•FO
RenderX

Characteristics	Studies (N=23), n (%)
Via online games	1 (4)
Intervention: s paced digital simulation training	
High-fidelity simulation	4 (17)
VR[c] simulation	1 (4)
Enhanced with mobile learning	1 (4)
Control: spaced online education	
No intervention	7 (30)
Massed online education via email	5 (22)
Online access to papers	2 (9)
Massed digital education via app with no alerts/reminders	1 (4)
Slideshow-based online program	1 (4)
Standard curriculum	1 (4)
Control: spaced digital simulation training	
Nonspaced high-fidelity simulation	3 (13)
Nonspaced simulation	2 (9)
No intervention	1 (4)
Number of study outcomes	
1	9 (39)
2	7 (30)
3	5 (22)
4	2 (9)
Category of study outcomes	
Knowledge	15 (65)
Skills	9 (39)
Attitudes	8 (35)
Satisfaction	7 (31)
Behavior change	8 (35)
Study quality	
Low risk of bias	11 (48)
Unclear risk of bias	5 (22)
High risk of bias	7 (30)

[a]RCT: randomized controlled trial.

[b]cRCT: cluster randomized controlled trial.

[c]VR: virtual reality.

Characteristics of Interventions

Spaced online education was described in 17 (74%) studies. The interventions consisted of brief learning modules delivered at prespecified intervals using email (n=10, 59%) [49,63,65,66,69-72,77]; proprietary software, such as Qstream (n=3, 18%) [64,68,74] and Moodle (n=1, 6%) [62]; an online game (n=1, 6%) [75]; and a smartphone app (n=1, 6%) [73]. Typically, in 15 (88%) of the 17 studies, the learning modules presented a case study or a small learning unit, followed by 1 or more multiple-choice questions (MCQs) and the correct answer with an explanation to each question. Alternatively, 2 (12%) [69,70] of the 17 studies offered small learning units without a testing component. The information was repeated 2 or more times at predetermined intervals or according to participant responses in adaptive systems (n=10, 59%) [62,63,67,68,71-74,77]. The duration of the interventions varied from 1 week [75] to 16 months [73].

Additionally, 6 studies (26%) described spaced digital simulation interventions delivered using high-fidelity simulators (n=4, 67%) [76,78-80], a VR system (n=1, 17%) [61], or low-fidelity simulators enhanced with weekly SMS text messages (n=1, 17%) [81]. The duration of the interventions varied from 2 training sessions of 4 hours each delivered over

XSL•FO
RenderX

1 week (n=1, 17%) [79] to 4 training sessions delivered over 4 weeks (n=2, 33%) [76,80].

Risk of Bias

The risk of bias (Figure 2) was low in 9 (39%) studies [49,61,64,66,67,69,70,76,78], with some concerns in 4 (17%) studies [62,63,65,80], and high in 7 (30%) studies [71,72,74,75,77,79]. The most common source of bias was missing outcome data. The risk-of-bias assessment for 3 (13%) cRCTs [68,73,81] reflected some concerns in one study, while it was low in 2 studies.

Figure 2. Risk-of-bias assessment for RCTs and cRCTs. cRCT: cluster randomized controlled trial; RCT: randomized controlled trial.

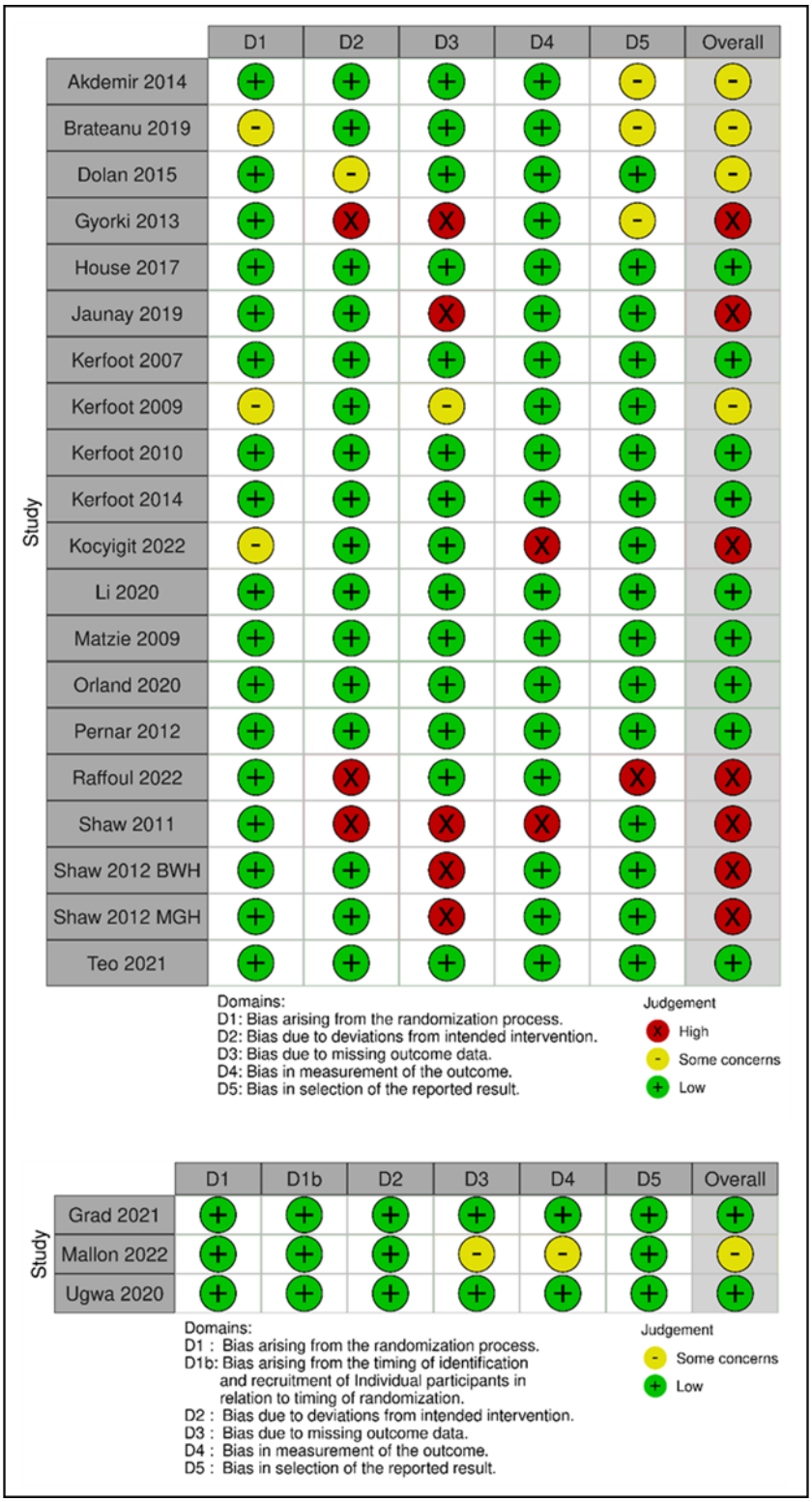

XSL•FO
RenderX

Primary Outcomes

Postintervention Knowledge

Knowledge was reported by 15 (65%) studies (2924/3371, 86.74%, participants), including 14 (93%) studies on spaced online education [49,62-68,72-75,77] and 1 (7%) study on spaced digital simulation [81]. In total, 12 (80%) of 15 studies assessed knowledge using MCQs [49,62-68,72,74,81], of which 7 (47%) studies used validated questionnaires [49,63,65-68,81]. The other studies used a combination of MCQs and short-answer questions [75], the short-answer component of the College of Family Physicians of Canada [73], or true/false questions developed for the study [77]. Figure 3 [63,72] summarizes the pooled estimates for postintervention knowledge outcomes. Three studies presented knowledge retention data [64,65,81].

Figure 3. Forest plot for postintervention knowledge. Note: SD values for Dolan et al and Shaw et al (Brigham and Women's Hospital and Massachusetts General Hospital) were derived from *P* values using RevMan's inbuilt calculator.

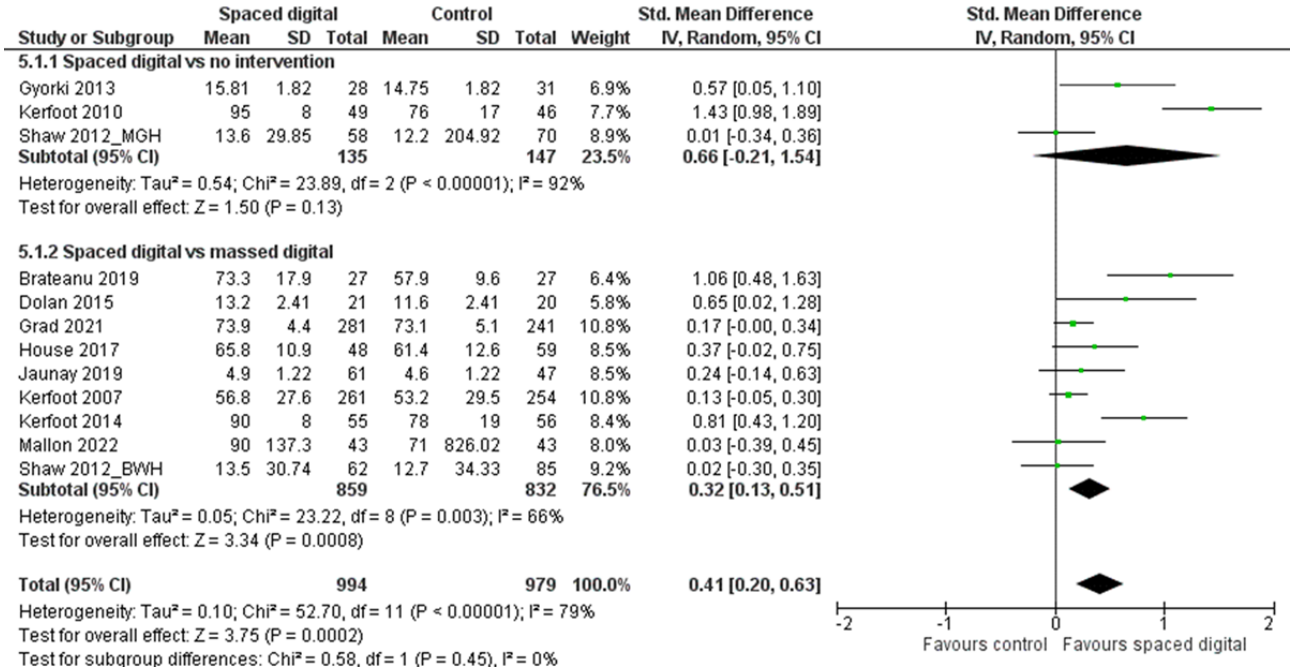

Spaced Online Education Versus No Intervention

Of 4 (27%) studies (499/2924, 17.07%, participants) [49,72,74,77], 3 (75%) studies (n=282, 56.5%, participants) [49,72,74] were included in the meta-analysis. The pooled estimate showed no significant difference in knowledge (SMD 0.66, 95% CI –0.21 to 1.54, I^2=92%) between the intervention and control groups.

Another study (133/499, 26.7%, participants) [77], not included in the meta-analysis, randomized primary care providers (PCPs) to receive 2 months of spaced online education or no intervention after watching an educational video on eating disorders and reported no further improvement in knowledge in the intervention group.

Spaced Online Education Versus Massed Education

Of the 15 studies, 9 (60%) studies (1691/2924, 57.83%, participants) [62-64,66-68,72,73,75] comparing spaced online education with massed digital education were included in the meta-analysis. The pooled estimate showed a small effect size (SMD 0.32, 95% CI 0.13-0.51, I^2=66%) favoring spaced online education.

Spaced Digital Simulation Versus Massed Simulation

Ugwa et al [81] compared low-dose onsite simulation enhanced with mobile mentoring with traditional massed simulation. Knowledge improved in the traditional arm immediately after completing the intervention.

Knowledge Retention

Of the 15 studies, 2 studies (313/2924, 10.7%, participants) [64,65] comparing spaced online education and massed education reported knowledge retention. House et al [64] reported a nonsignificant effect on knowledge retention at 3 months' follow-up (SMD 0.46, 95% CI –0.05 to 0.98), while Kerfoot [65] reported a small effect size 2 years after the completion of a trial on spaced versus massed online education (SMD 0.34, 95% CI 0.02-0.67). The pooled estimate showed a small difference in knowledge retention favoring spaced online education over controls (SMD 0.38, 95% CI 0.10-0.65, I^2=0%). Figure 4 presents the pooled estimates for knowledge retention.

XSL•FO
RenderX

Figure 4. Forest plot for knowledge retention.

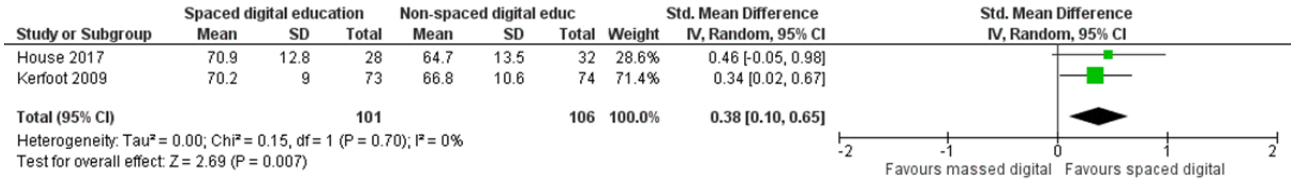

Additionally, Ugwa et al [81] reported that knowledge retention scores at 3 and 12 months' follow-up were similar in the spaced digital simulation and control arms (RR 1.02, 95% CI 0.91-1.13).

Postintervention Skills

A total of 9 (39%) studies (784/3371, 23.26%, participants) reported learners' skills, including 3 (33%) studies on spaced online education [69,70,72] and 6 (67%) studies on spaced digital simulation [61,76,78-81]. Postintervention skills were measured using online surveys distributed among medical students to assess the faculty's [70] or residents' [69] teaching skills, anonymized video recordings [61,72], task completion reports obtained from the simulation instrument [76,78,80], the Objective Structured Clinical Examination (OSCE) [81], and self-reported questionnaires [79].

Spaced Online Education Versus No Intervention

A total of 2 (22%) studies (84/784, 2.4%, participants) [69,70] compared the impact of spaced online education on improving teaching skills. Matzie et al [69] reported that residents receiving the intervention offered more effective feedback to medical students (RR 1.43, 95% CI 1.08-1.90), while Pernar et al [70] reported no difference in faculty's feedback quality between intervention and control groups (RR 0.96, 95% CI 0.72-1.27). The pooled estimate for both studies reflected a nonsignificant effect of spaced online education on improving teaching skills (RR 1.21, 95% CI 0.99-1.47, I^2=74%). Figure 5 presents the forest plot.

Figure 5. Forest plot for postintervention skills (spaced online education vs no intervention).

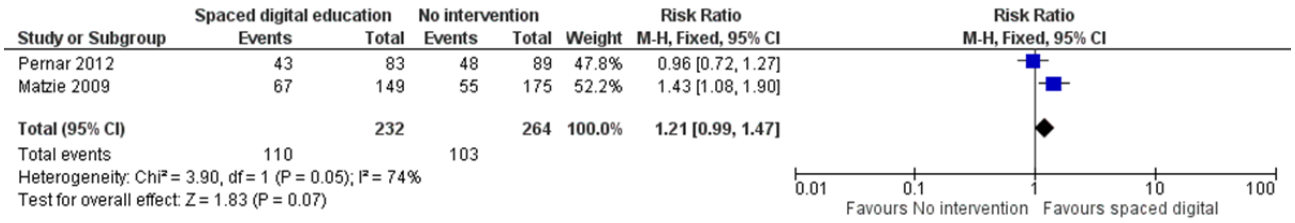

Spaced Online Education Versus Massed Education

Shaw et al [72] assessed interns' compliance (Brigham and Women's Hospital: 195/784, 24.9%, participants) with patient safety procedures while completing a central venous catheterization simulation, reporting a nonsignificant effect (SMD 0.34, 95% CI –0.05 to 0.73).

Spaced Digital Simulation Versus Massed Simulation

A total of 5 (56%) studies (483/784, 61.6%, participants) [61,76,78,79,81] compared spaced digital simulation with massed simulation (Figure 6 [11]). The interventions aimed to improve nursing students' skills to manage an adult patient with chronic lymphocytic leukemia [79], residents' arthroscopic skills [78], medical students' intramedullary tibial nailing [61] and microsurgical suturing skills [76], and health workers' maternal and newborn care skills [81]. In addition, 3 (60%) studies [76,79,81] reported skill improvement as overall scores. The pooled estimate of 2 (67%) studies [76,79] showed that spaced digital simulation is superior to massed simulation (SMD 1.24, 95% CI 0.84-1.64, I^2=74%) in improving learners' skills. However, Ugwa et al [81] reported a nonsignificant effect in learners' skills using a spaced simulation intervention enhanced with mobile mentoring (RR 1.05, 95% CI 0.95-1.16) compared to controls.

XSL•FO
RenderX

Figure 6. Forest plots for learners' skill outcomes. Note: SD values for Li et al were derived from the SE using RevMan's inbuilt calculator.

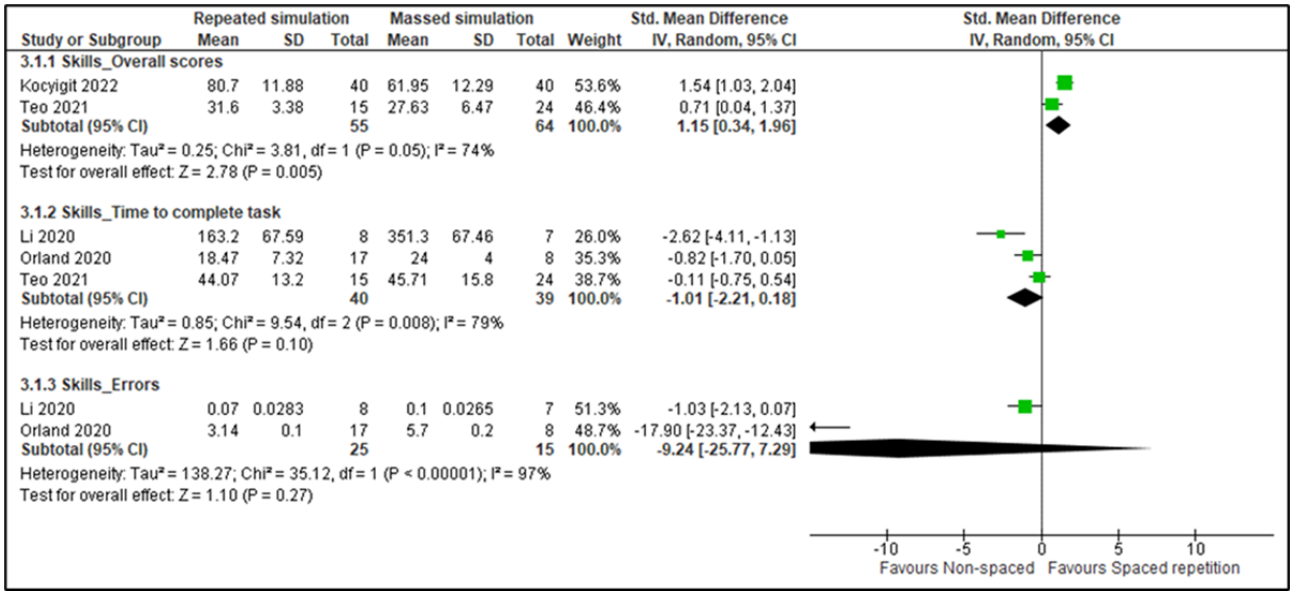

In addition, 3 (60%) studies [61,76,78] reported itemized skill outcomes, including the time to complete a task, economy of movements, the number of errors, and the number of hints requested, with improved skills representing decreased values. Li et al [78] reported a decrease in task completion time and camera path length but no difference in the extent of cartilage injury in orthopedic residents training in arthroscopic skills using spaced high-fidelity simulation. Orland et al [61] reported a shorter completion time, decreased errors and number of requested hints, and a higher rate of completed tasks in medical students using spaced VR simulation with or without technique guidance compared to controls. Teo et al [76] reported a nonsignificant difference in the time to complete the test strip between students receiving spaced digital simulation and controls. The pooled estimates for the time to complete a task (3, 60%, studies) and the number of errors (2, 40%, studies) showed a nonsignificant difference between intervention and control groups.

Spaced Digital Simulation Versus No Intervention

Akdemir et al [80] (22/784, 2.8%, participants) described that a spaced digital simulation intervention to improve gynecology residents' laparoscopic skills resulted in a significant reduction in the time to complete tasks (SMD –1.50, 95% CI –2.46 to –0.53) and the number of movements (SMD –1.96, 95% CI –3.01 to –0.91) and a nonsignificant change in the number of errors (SMD –0.26, 95% CI –1.10 to 0.58).

Postintervention Attitudes

A total of 8 (35%) studies (1539/3371, 45.65%, participants) reported attitudes, including confidence [64,68,71,72], comfort in screening [77], engagement with the intervention (time spent [75] and the number of completed cases [73]), self-efficacy [79], and anxiety [79].

Spaced Online Education Versus No Intervention

Raffoul et al [77] reported that PCPs in the intervention group were more comfortable screening for eating disorders. Shaw et al [71] showed that PCPs receiving spaced online education after a conference reported greater confidence in managing 4 conditions presented in the intervention.

Spaced Online Education Versus Massed Education

The pooled estimate of 3 (38%) studies showed improved confidence in managing clinical cases [64,68] and patient safety protocols [72] (SMD 0.28, 95% CI 0.01-0.55, I^2=43%), as shown in Figure 7 [64,72].

Figure 7. Forest plot for learners' postintervention confidence (spaced online education vs massed education). Note: SD values for Shaw et al (Brigham and Women's Hospital) were derived from *P* values, and SD values for House et al were derived from CIs using RevMan's inbuilt calculator.

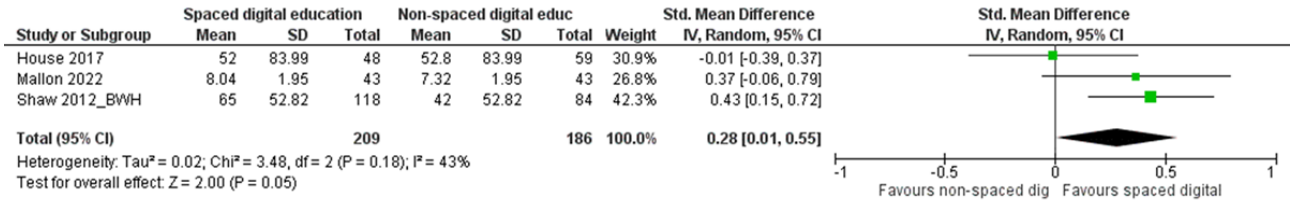

Additionally, Jaunay et al [75] reported that general practitioners (GPs) who received spaced education via an online game spent 45-60 minutes in the intervention compared to 10-20 minutes in the control group. Grad et al [73] compared 2 versions of a study app presenting family medicine case studies and showed that adding reminders that enabled spaced learning resulted in residents completing more case studies, although the difference was nonsignificant.

Spaced Digital Simulation Versus Massed Simulation

Kocyigit and Karagozoglu [79] reported that spaced digital simulation improved nursing students' self-efficacy (SMD 0.57,

XSL•FO
RenderX

95% CI 0.12-1.02), while there was no significant change in state anxiety (SMD 0.25, 95% CI –0.19 to 0.69) between the study groups.

Postintervention Satisfaction

A total of 7 (30%) studies (982/3371, 29.13%, participants) [64,70-72,75,76,81] reported satisfaction with spaced digital education. Satisfaction was assessed using a postintervention survey [75], Likert scale surveys [64,71,72,76], an email-based questionnaire [70], and a satisfaction questionnaire plus focus group discussion [81]. Satisfaction was measured only in the intervention group, except Teo et al [76], who administered the survey to both intervention and control groups and reported similar satisfaction (8.47/10 in spaced digital simulation vs 8.0/10 in massed simulation; *P*=.23). In general, participants considered spaced digital education engaging and enjoyable [72,75] and said that would like to engage in similar interventions [64,72,75] in the future. Participants also reported greater satisfaction in improved patient outcomes [81]. However, 1 (14%) study [70] reported that surgery faculty were not satisfied with an intervention consisting of weekly emails with tips to improve their teaching skills, because the content was vague and generic and it added to the burden of numerous daily emails.

Postintervention Behavior Change

Of the 23 studies, 8 (35%) studies (858/3371, 25.45%, participants) reported a change in clinical behavior, including 7 (88%) studies on spaced online education [49,63,67,68,71,75,77] and 1 (12%) study on spaced digital simulation [79]. The behavior change data were obtained from clinical records [49,63,67,68] and self-reported questionnaires developed by the study authors [71,75,77,79]. Figure 8 [49,63] shows the meta-analyses results.

Figure 8. Forest plot for learners' postintervention change in clinical behavior. Note: SD values for Dolan et al and Kerfoot et al were derived from *P* values using RevMan's inbuilt calculator.

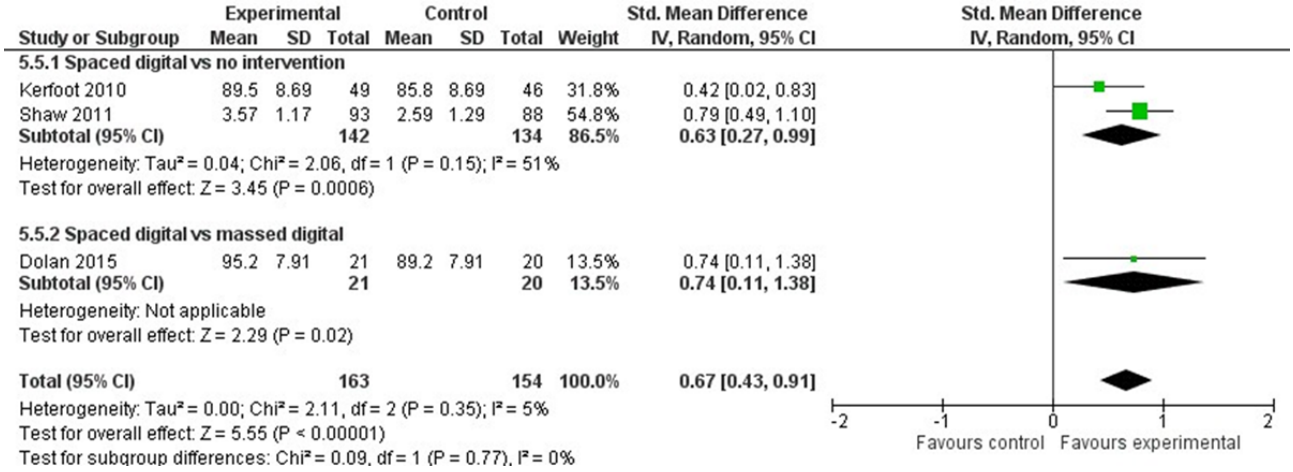

Spaced Online Education Versus No Intervention

The pooled estimate from 2 (25%) studies [49,71] showed that spaced online education promotes a change in clinicians' behavior (SMD 0.63, 95% CI 0.27-0.99, I^2=51%). Kerfoot et al [49] showed that spaced online education improved clinicians' prostate cancer–screening behavior while receiving the intervention (SMD 0.42, 95% CI 0.02-0.83) and during the 72-week follow-up (SMD 0.53, 95% CI 0.12-0.94). Shaw et al [71] reported a change in PCPs' clinical behavior for 4 common disorders (hypertension, diabetes, nonalcoholic steatohepatitis, and HIV) after receiving a spaced online education intervention (SMD 0.36, 95% CI 0.07-0.66 for hypertension to SMD 0.95, 95% CI 0.65-1.26 for HIV). Finally, Raffoul et al [77] reported that PCPs receiving spaced online education were more comfortable screening for eating disorders, although there was no difference between the intervention and control groups in self-reported screening or referral behavior.

Spaced Online Education Versus Massed Education

Dolan et al [63] reported that residents receiving spaced online education appropriately screened more patients for osteoporosis (SMD 0.74, 95% CI 0.11-1.38) and were more likely to treat patients with bisphosphonates (SMD 0.69, 95% CI 0.06-1.32). Jaunay et al [75] reported that GPs receiving an intervention involving a spaced online education game used the new knowledge to improve clinical practice more often (RR 1.45, 95% CI 1.07-1.96), compared to controls. Kerfoot et al [67] demonstrated improved patient outcomes (shorter time to blood pressure target) for PCPs receiving spaced online education (hazard ratio 1.043, 95% CI 1.007-1.081, *P*=.018). Finally, Mallon et al [68] reported that a spaced online constipation management program delivered to pediatric PCPs resulted in lower patient referrals to specialist pediatric gastroenterology clinics but no reduction in constipation-related visits to the emergency department or urgent care clinics.

Spaced Digital Simulation Versus Massed Simulation

Kocyigit and Karagozoglu [79] reported that nursing students' medical error tendency significantly decreased after 2 spaced digital simulation sessions (80.7 vs 61.9 points, *P*=0, with higher values reflecting fewer medical errors).

Discussion

Principal Findings

In this systematic review, the first to include a meta-analysis, we found that spaced digital education is effective in improving knowledge, skills, and confidence and promoting change in

XSL•FO
RenderX

clinical practice for pre- and postregistration health professionals. In particular, marked improvements were demonstrated in physicians' adoption of evidence-based practices after engaging in spaced online learning and in students' and residents' surgical skills after engaging in spaced digital simulation interventions compared to active or inactive controls. The results for knowledge and behavior change, albeit reported in a smaller number of papers, showed that these improvements are maintained over time. Yet, our results should be interpreted with caution, given the substantial heterogeneity of included studies and the small number of studies populating the different spaced digital education categories.

Our systematic review categorized the spaced interventions into 2 distinct categories: spaced online education interventions repeatedly delivered small amounts of information while testing learners' knowledge, while spaced digital simulation interventions provided practical, repeated, and spaced training sessions using high-fidelity simulators to improve learners' skills. The studies included in this review evaluated 3 learning outcomes well aligned with the 4 levels of Miller's classification of clinical competence: knowledge ("knows"), skills ("knows how," "shows how"), and behavior change ("does") [59]. Although measuring learners' attitudes and satisfaction with the digital health education intervention is important and may facilitate learners' engagement with the intervention, knowledge, skills, and behavior change are particularly important to determine the impact of the intervention. Knowledge or skills provide a direct evaluation of the intervention, and selecting one over the other will directly depend on the type of educational intervention content to be delivered. Alternatively, behavior change would more directly reflect the impact of the digital health education intervention on health professionals' daily practice and the potential benefits to patients.

The findings reported in this review support the use of spaced digital education interventions in undergraduate and postgraduate health profession education. However, medical schools and providers of continuous professional development programs have yet to fully embrace this learning modality [82-84]. Simultaneously, students are increasingly turning to commercial learning platforms to extend their learning and to prepare for licensing examinations, such as the USMLE (United States Medical Licensing Examination) program [82,85]. These platforms often follow spaced education and active recall principles in the development of learning materials. About one-third of second-year medical students in the United States routinely use online materials not provided by their medical school to prepare for courses and particularly in preparation for the USMLE [86-88], creating a "parallel, medical student–driven curriculum" [86]. As digital health education platforms become more pervasive, it is critical that institutions providing undergraduate and continuous postgraduate development programs adapt their curriculum to incorporate evidence-based learning approaches to improve students' learning outcomes, while increasing student satisfaction and acceptability of the educational content.

The studies included in this review described interventions with substantially diverse spacing intervals. However, none of the studies provided a rationale for this differing length. Previous studies from the educational research field have shown that the length of the spacing interval is directly related to the knowledge retention period, suggesting that the ideal interval would be around 10% of the test delay [23,89]. Furthermore, expanding the spacing interval appears to further improve retention [89]. Most of these studies have reported solely on knowledge, such as vocabulary recall [23]. Similar research in health profession education is lacking. Therefore, it is imperative that future research evaluate the role of the spacing effect on knowledge retention in health care professionals, as well as the impact of spaced digital education on other learning outcomes, such as skills or clinical behavior change.

Spaced online education interventions have been mostly delivered via email, which may not fully leverage more interactive, engaging features often included in other digital interventions. For example, a recent literature review [90] suggested that the use of game design elements in the development of digital health education interventions may increase user engagement and enhance collaboration. In addition, microlearning, an educational approach that consists of brief, focused lessons, is gaining acceptance as a professional development method in the corporate sector and industry [91], encouraging the development of gamified digital platforms customizable to user needs. However, these platforms do not appear to be widely used in health care settings, despite the potential benefits of including multimedia content, as well as reminders, leaderboards, or other gamified elements, to increase user engagement [92].

Strengths and Limitations

There are several strengths of this systematic review. We followed gold-standard Cochrane guidelines, and we used a comprehensive search strategy developed by expert librarians that was previously used in a series of systematic reviews on the effectiveness of digital health education modalities in health profession education [5,12-14,17,18,36,38,40,42-47]. The search strategy was supplemented by forward and backward reference screening.

However, the review also has limitations. First, our search strategy was developed to retrieve clinical trials on digital health education interventions without including spaced education–specific search terms. However, to compensate for this limitation, we added a backward and forward citation search of all included studies. Second, the included studies showed substantial heterogeneity: they reported diverse outcomes and used diverse measurement instruments and reporting variables that limited further data analyses. Third, the low quality of the included studies, reflected in the majority of unclear- and high-risk-of-bias assessments, added uncertainty to the validity of our results.

Future Research

Several aspects of spaced digital education for health care professionals warrant further research. First, the optimal spacing interval for spaced interventions needs to be defined, accounting for the type of learning outcome targeted by the intervention (eg, knowledge, skills, behavior change, satisfaction), the health care profession, the career stage (pre-registration students or

XSL•FO
RenderX

postregistration professionals), and the intended retention period. Second, the potential impact of newer, more interactive platforms in influencing active engagement of participants, their compliance with the intervention, and the improvement in measured outcomes warrants further investigation. Third, more research is needed to identify the areas of learning where a spaced intervention may be more appropriate. Fourth, studies evaluating possible adverse events associated with interventions are needed as adverse events were not reported in any of the studies included in this review. In addition, cost-effectiveness studies are required to assess the feasibility of large-scale implementation of this learning modality.

Conclusion

This systematic review showed that spaced digital education is effective in improving knowledge, skills, and clinical behavior change in pre- and postregistration health care professionals. The improvement of surgical skills using spaced digital simulation and clinical behavior change after spaced online education interventions are particularly noteworthy. These findings suggest that spaced digital education could be effectively added to undergraduate health profession education and continuous professional education programs for practicing professionals. For the successful implementation of spaced digital education interventions in the health professions, we suggest considering the use of newer, more interactive platforms to deliver the interventions. Future studies should evaluate areas of learning where the intervention will be most effective and assess the ideal length of the intervention and the spacing interval.

Acknowledgments

We gratefully acknowledge contributions by Ms Yasmin Lynda Munro from the LKCMedicine Medical Library for her assistance in updating the digital health education library in February 2022 and February 2023.

Data Availability

The data sets supporting the conclusions of this paper are included within the paper and Multimedia Appendices.

Authors' Contributions

LM and LTC had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. LTC conceptualized and designed the study. LM updated the study library. LM, MSPN, TDRN, YIA, and AIJ selected studies for inclusion, extracted data, and assessed the study quality. LM, BMK, and LTC performed statistical analyses and performed Grading of Recommendations, Assessment, Development and Evaluation (GRADE) assessments. LM, MSPN, TDRN, and YIA drafted the manuscript. All authors reviewed and provided comments for the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

LTC is Associate Editor of *JMIR Medical Education*.

Multimedia Appendix 1
Medline search strategy.
[DOCX File , 24 KB - jmir_v26i1e57760_app1.docx]

Multimedia Appendix 2
Grading of Recommendations, Assessment, Development and Evaluation (GRADE)–based evaluation: summary of findings.
[DOCX File , 30 KB - jmir_v26i1e57760_app2.docx]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File , 49 KB - jmir_v26i1e57760_app3.docx]

Multimedia Appendix 4
Excluded studies.
[DOCX File , 25 KB - jmir_v26i1e57760_app4.docx]

Multimedia Appendix 5
PRISMA 2020 checklist.
[DOCX File , 51 KB - jmir_v26i1e57760_app5.docx]

XSL•FO
RenderX

References

1. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? BMJ Glob Health 2022 Jun;7(6):e009316 [FREE Full text] [doi: 10.1136/bmjgh-2022-009316] [Medline: 35760437]
2. Global strategy on human resources for health: workforce 2030. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511131 [accessed 2023-07-01]
3. De Vries N, Lavreysen O, Boone A, Bouman J, Szemik S, Baranski K, et al. Retaining healthcare workers: a systematic review of strategies for sustaining power in the workplace. Healthcare (Basel) 2023 Jun 29;11(13):1887 [FREE Full text] [doi: 10.3390/healthcare11131887] [Medline: 37444721]
4. Poon YR, Lin YP, Griffiths P, Yong KK, Seah B, Liaw SY. A global overview of healthcare workers' turnover intention amid COVID-19 pandemic: a systematic review with future directions. Hum Resour Health 2022 Sep 24;20(1):70 [FREE Full text] [doi: 10.1186/s12960-022-00764-7] [Medline: 36153534]
5. Martinengo L, Yeo NJY, Markandran KD, Olsson M, Kyaw BM, Car LT. Digital health professions education on chronic wound management: a systematic review. Int J Nurs Stud 2020 Apr;104:103512 [FREE Full text] [doi: 10.1016/j.ijnurstu.2019.103512] [Medline: 32086027]
6. Car J, Carlstedt-Duke J, Tudor Car L, Posadzki P, Whiting P, Zary N, et al. Digital education in health professions: the need for overarching evidence synthesis. J Med Internet Res 2019 Mar 14;21(2):e12913 [FREE Full text] [doi: 10.2196/12913] [Medline: 30762583]
7. Digital education for building health workforce capacity. World Health Organization. 2020. URL: https://apps.who.int/iris/bitstream/handle/10665/331524/9789240000476-eng.pdf?sequence=1&isAllowed=y [accessed 2023-09-07]
8. Carroll C, Booth A, Papaioannou D, Sutton A, Wong R. UK health-care professionals' experience of on-line learning techniques: a systematic review of qualitative data. J Contin Educ Health Prof 2009;29(4):235-241. [doi: 10.1002/chp.20041] [Medline: 19998474]
9. Reeves S, Fletcher S, McLoughlin C, Yim A, Patel KD. Interprofessional online learning for primary healthcare: findings from a scoping review. BMJ Open 2017 Aug 04;7(8):e016872 [FREE Full text] [doi: 10.1136/bmjopen-2017-016872] [Medline: 28780560]
10. Rouleau G, Gagnon M, Côté J, Payne-Gagnon J, Hudson E, Dubois C, et al. Effects of e-learning in a continuing education context on nursing care: systematic review of systematic qualitative, quantitative, and mixed-studies reviews. J Med Internet Res 2019 Oct 02;21(10):e15118 [FREE Full text] [doi: 10.2196/15118] [Medline: 31579016]
11. Vaona A, Banzi R, Kwag KH, Rigon G, Cereda D, Pecoraro V, et al. E-learning for health professionals. Cochrane Database Syst Rev 2018 Jan 21;1(1):CD011736 [FREE Full text] [doi: 10.1002/14651858.CD011736.pub2] [Medline: 29355907]
12. George PP, Zhabenko O, Kyaw BM, Antoniou P, Posadzki P, Saxena N, et al. Online digital education for postregistration training of medical doctors: systematic review by the Digital Health Education Collaboration. J Med Internet Res 2019 Feb 25;21(2):e13269 [FREE Full text] [doi: 10.2196/13269] [Medline: 30801252]
13. Wahabi HA, Esmaeil SA, Bahkali KH, Titi MA, Amer YS, Fayed AA, et al. Medical doctors' offline computer-assisted digital education: systematic review by the Digital Health Education Collaboration. J Med Internet Res 2019 Mar 01;21(3):e12998 [FREE Full text] [doi: 10.2196/12998] [Medline: 30821689]
14. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, et al. Virtual reality for health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959 [FREE Full text] [doi: 10.2196/12959] [Medline: 30668519]
15. Davids J, Manivannan S, Darzi A, Giannarou S, Ashrafian H, Marcus HJ. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. Neurosurg Rev 2021 Aug;44(4):1853-1867 [FREE Full text] [doi: 10.1007/s10143-020-01378-0] [Medline: 32944808]
16. Toale C, Morris M, Kavanagh DO. Training and assessment using the LapSim laparoscopic simulator: a scoping review of validity evidence. Surg Endosc 2023 Mar;37(3):1658-1671 [FREE Full text] [doi: 10.1007/s00464-022-09593-0] [Medline: 36123545]
17. Dunleavy G, Nikolaou CK, Nifakos S, Atun R, Law GCY, Tudor Car L. Mobile digital education for health professions: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Feb 12;21(2):e12937 [FREE Full text] [doi: 10.2196/12937] [Medline: 30747711]
18. Lall P, Rees R, Law GCY, Dunleavy G, Cotič, Car J. Influences on the implementation of mobile learning for medical and nursing education: qualitative systematic review by the Digital Health Education Collaboration. J Med Internet Res 2019 Feb 28;21(2):e12895 [FREE Full text] [doi: 10.2196/12895] [Medline: 30816847]
19. Cepeda NJ, Coburn N, Rohrer D, Wixted JT, Mozer MC, Pashler H. Optimizing distributed practice: theoretical analysis and practical implications. Exp Psychol 2009;56(4):236-246 [FREE Full text] [doi: 10.1027/1618-3169.56.4.236] [Medline: 19439395]
20. Cepeda NJ, Vul E, Rohrer D, Wixted JT, Pashler H. Spacing effects in learning: a temporal ridgeline of optimal retention. Psychol Sci 2008 Nov;19(11):1095-1102. [doi: 10.1111/j.1467-9280.2008.02209.x] [Medline: 19076480]
21. Ebbinghaus H. Memory: A Contribution to Experimental Psychology. New York, NY: Teachers College, Columbia University; 1913.

XSL•FO
RenderX

22. Carpenter SK, Pan SC, Butler AC. The science of effective learning with spacing and retrieval practice. Nat Rev Psychol 2022 Aug 02;1(9):496-511. [doi: 10.1038/s44159-022-00089-1]
23. Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: a review and quantitative synthesis. Psychol Bull 2006 May;132(3):354-380. [doi: 10.1037/0033-2909.132.3.354] [Medline: 16719566]
24. Versteeg M, Hendriks RA, Thomas A, Ommering BWC, Steendijk P. Conceptualising spaced learning in health professions education: a scoping review. Med Educ 2020 Mar;54(3):205-216 [FREE Full text] [doi: 10.1111/medu.14025] [Medline: 31860936]
25. Yeh DD, Park YS. Improving learning efficiency of factual knowledge in medical education. J Surg Educ 2015;72(5):882-889 [FREE Full text] [doi: 10.1016/j.jsurg.2015.03.012] [Medline: 25920573]
26. Kerfoot BP, DeWolf WC, Masser BA, Church PA, Federman DD. Spaced education improves the retention of clinical knowledge by medical students: a randomised controlled trial. Med Educ 2007 Jan;41(1):23-31 [FREE Full text] [doi: 10.1111/j.1365-2929.2006.02644.x] [Medline: 17209889]
27. Patocka C, Cheng A, Sibbald M, Duff JP, Lai A, Lee-Nobbee P, et al. A randomized education trial of spaced versus massed instruction to improve acquisition and retention of paediatric resuscitation skills in emergency medical service (EMS) providers. Resuscitation 2019 Aug;141:73-80 [FREE Full text] [doi: 10.1016/j.resuscitation.2019.06.010] [Medline: 31212041]
28. Lin Y, Cheng A, Grant VJ, Currie GR, Hecker KG. Improving CPR quality with distributed practice and real-time feedback in pediatric healthcare providers - a randomized controlled trial. Resuscitation 2018 Sep;130:6-12. [doi: 10.1016/j.resuscitation.2018.06.025] [Medline: 29944894]
29. Wiseman JJ, Perlmutter JW, Wiseman SM. Learning to learn: spaced education approaches should be adopted to help optimize learning during surgical residency training. Am J Surg 2023 Feb;225(2):439-441 [FREE Full text] [doi: 10.1016/j.amjsurg.2022.09.025] [Medline: 36184327]
30. Choe AI, Woodard S, Thompson BM, Walter V, Fotos JS, Kasales CJ. Spaced education: randomized trial comparing learning efficiency of the adaptive versus nonadaptive spaced education systems among radiology residents. J Am Coll Radiol 2022 Jun;19(6):706-710 [FREE Full text] [doi: 10.1016/j.jacr.2022.03.010] [Medline: 35472369]
31. Phillips JL, Heneka N, Bhattarai P, Fraser C, Shaw T. Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: a systematic review. Med Educ 2019 Sep;53(9):886-902 [FREE Full text] [doi: 10.1111/medu.13895] [Medline: 31144348]
32. Yeung J, Djarv T, Hsieh MJ, Sawyer T, Lockey A, Finn J, Education, Implementation and Team Task Force and Neonatal Life Support Task Force of the International Liaison Committee on Resuscitation (ILCOR). Spaced learning versus massed learning in resuscitation - a systematic review. Resuscitation 2020 Nov;156:61-71 [FREE Full text] [doi: 10.1016/j.resuscitation.2020.08.132] [Medline: 32926969]
33. Hirsch AW, Nagler J. Spaced education through e-learning for ongoing professional development. AEM Educ Train 2020 Oct;4(4):423-427 [FREE Full text] [doi: 10.1002/aet2.10446] [Medline: 33150287]
34. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M. Cochrane Handbook for Systematic Reviews of Interventions version 6.3. Cochrane. 2022. URL: https://training.cochrane.org/handbook [accessed 2023-07-20]
35. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
36. Bajpai S, Semwal M, Bajpai R, Car J, Ho AHY. Health professions' digital education: review of learning theories in randomized controlled trials by the Digital Health Education Collaboration. J Med Internet Res 2019 Mar 12;21(3):e12912 [FREE Full text] [doi: 10.2196/12912] [Medline: 30860483]
37. Brusamento S, Kyaw BM, Whiting P, Li L, Tudor Car L. Digital health professions education in the field of pediatrics: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Sep 25;21(9):e14231 [FREE Full text] [doi: 10.2196/14231] [Medline: 31573906]
38. Divakar U, Nazeha N, Posadzki P, Jarbrink K, Bajpai R, Ho AHY, et al. Digital education of health professionals on the management of domestic violence: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 May 23;21(5):e13868 [FREE Full text] [doi: 10.2196/13868] [Medline: 31124462]
39. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car L, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994 [FREE Full text] [doi: 10.2196/12994] [Medline: 30920375]
40. Huang Z, Semwal M, Lee SY, Tee M, Ong W, Tan WS, et al. Digital health professions education on diabetes management: systematic review by the Digital Health Education Collaboration. J Med Internet Res 2019 Feb 21;21(2):e12997 [FREE Full text] [doi: 10.2196/12997] [Medline: 30789348]
41. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]

XSL•FO
RenderX

42. Kyaw BM, Posadzki P, Dunleavy G, Semwal M, Divakar U, Hervatis V, et al. Offline digital education for medical students: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Mar 25;21(3):e13165 [FREE Full text] [doi: 10.2196/13165] [Medline: 30907731]
43. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967-e12960 [FREE Full text] [doi: 10.2196/12967] [Medline: 3090773]
44. Kyaw BM, Tudor Car L, van Galen LS, van Agtmael MA, Costelloe CE, Ajuebor O, et al. Health professions digital education on antibiotic management: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Sep 12;21(9):e14984 [FREE Full text] [doi: 10.2196/14984] [Medline: 31516125]
45. Posadzki P, Bala MM, Kyaw BM, Semwal M, Divakar U, Koperny M, et al. Offline digital education for postregistration health professions: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Apr 24;21(4):e12968 [FREE Full text] [doi: 10.2196/12968] [Medline: 31017584]
46. Semwal M, Whiting P, Bajpai R, Bajpai S, Kyaw BM, Tudor Car L. Digital education for health professions on smoking cessation management: systematic review by the Digital Health Education Collaboration. J Med Internet Res 2019 Mar 04;21(3):e13000 [FREE Full text] [doi: 10.2196/13000] [Medline: 30829576]
47. Tudor Car L, Kyaw BM, Dunleavy G, Smart NA, Semwal M, Rotgans JI, et al. Digital problem-based learning in health professions: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Feb 28;21(2):e12945 [FREE Full text] [doi: 10.2196/12945] [Medline: 30816846]
48. UNESCO Institute for Statistics. International Standard Classification of Education Fields of Education and Training 2013 (ISCED-F 2013). Montreal, QC: UNESCO Institute for Statistics; 2015.
49. Kerfoot BP, Lawler EV, Sokolovskaya G, Gagnon D, Conlin PR. Durable improvements in prostate cancer screening from online spaced education a randomized controlled trial. Am J Prev Med 2010 Nov;39(5):472-478 [FREE Full text] [doi: 10.1016/j.amepre.2010.07.016] [Medline: 20965387]
50. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021 Feb 01;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
51. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 07;13(7):e072254 [FREE Full text] [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
52. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
53. Higgins J, Savović J, Page M, Elbers R, Sterne J. Chapter 8: assessing risk of bias in a randomized trial. Cochrane Handbook for Systematic Reviews of Interventions version 6.5. 2023. URL: https://training.cochrane.org/handbook/current/chapter-08 [accessed 2024-07-31]
54. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan 26;12(1):55-61 [FREE Full text] [doi: 10.1002/jrsm.1411] [Medline: 32336025]
55. Schünemann H, Higgins J, Vist G, Glasziou P, Akl E, Skoetz N, et al. Chapter 14: completing 'summary of findings' tables and grading the certainty of the evidence. Cochrane Handbook for Systematic Reviews of Interventions version 6.5. 2023. URL: https://training.cochrane.org/handbook/current/chapter-14 [accessed 2024-09-16]
56. Cochrane RevMan. The Cochrane Collaboration. URL: https://training.cochrane.org/online-learning/core-software/revman [accessed 2024-09-16]
57. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 2018 Jun;27(6):1785-1805. [doi: 10.1177/0962280216669183] [Medline: 27683581]
58. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 2014 Dec 19;14:135 [FREE Full text] [doi: 10.1186/1471-2288-14-135] [Medline: 25524443]
59. Higgins J, Li T, Deeks J. Chapter 6: choosing effect measures and computing estimates of effect. Cochrane Handbook for Systematic Reviews of Interventions version 6.5. 2023. URL: https://training.cochrane.org/handbook/current/chapter-06 [accessed 2024-09-16]
60. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
61. Orland MD, Patetta MJ, Wieser M, Kayupov E, Gonzalez MH. Does virtual reality improve procedural completion and accuracy in an intramedullary tibial nail procedure? A randomized control trial. Clin Orthop Relat Res 2020 Sep;478(9):2170-2177 [FREE Full text] [doi: 10.1097/CORR.0000000000001362] [Medline: 32769533]
62. Brateanu A, Strang TM, Garber A, Mani S, Spencer A, Spevak B, et al. Using an adaptive, self-directed web-based learning module to enhance residents' medical knowledge prior to a new clinical rotation. Med Sci Educ 2019 Sep;29(3):779-786 [FREE Full text] [doi: 10.1007/s40670-019-00772-8] [Medline: 34457542]

XSL•FO
RenderX

63. Dolan BM, Yialamas MA, McMahon GT. A randomized educational intervention trial to determine the effect of online education on the quality of resident-delivered care. J Grad Med Educ 2015 Sep;7(3):376-381 [FREE Full text] [doi: 10.4300/JGME-D-14-00571.1] [Medline: 26457142]
64. House H, Monuteaux MC, Nagler J. A randomized educational interventional trial of spaced education during a pediatric rotation. AEM Educ Train 2017 Apr;1(2):151-157 [FREE Full text] [doi: 10.1002/aet2.10025] [Medline: 30051026]
65. Kerfoot BP. Learning benefits of on-line spaced education persist for 2 years. J Urol 2009 Jun;181(6):2671-2673 [FREE Full text] [doi: 10.1016/j.juro.2009.02.024] [Medline: 19375095]
66. Kerfoot BP, Baker HE, Koch MO, Connelly D, Joseph DB, Ritchey ML. Randomized, controlled trial of spaced education to urology residents in the United States and Canada. J Urol 2007 Apr;177(4):1481-1487 [FREE Full text] [doi: 10.1016/j.juro.2006.11.074] [Medline: 17382760]
67. Kerfoot BP, Turchin A, Breydo E, Gagnon D, Conlin PR. An online spaced-education game among clinicians improves their patients' time to blood pressure control: a randomized controlled trial. Circ Cardiovasc Qual Outcomes 2014 May;7(3):468-474 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.113.000814] [Medline: 24847084]
68. Mallon D, Fei L, Farrell M, Anderson JB, Klein M. Randomized controlled trial of interactive spaced education to support constipation management by pediatric primary care providers. J Pediatr Gastroenterol Nutr 2022 May 01;74(5):568-574 [FREE Full text] [doi: 10.1097/MPG.0000000000003405] [Medline: 35149642]
69. Matzie KA, Kerfoot BP, Hafler JP, Breen EM. Spaced education improves the feedback that surgical residents give to medical students: a randomized trial. Am J Surg 2009 Feb;197(2):252-257 [FREE Full text] [doi: 10.1016/j.amjsurg.2008.01.025] [Medline: 18722585]
70. Pernar LIM, Beleniski F, Rosen H, Lipsitz S, Hafler J, Breen E. Spaced education faculty development may not improve faculty teaching performance ratings in a surgery department. J Surg Educ 2012;69(1):52-57 [FREE Full text] [doi: 10.1016/j.jsurg.2011.06.013] [Medline: 22208833]
71. Shaw T, Long A, Chopra S, Kerfoot BP. Impact on clinical behavior of face-to-face continuing medical education blended with online spaced education: a randomized controlled trial. J Contin Educ Health Prof 2011;31(2):103-108 [FREE Full text] [doi: 10.1002/chp.20113] [Medline: 21671276]
72. Shaw TJ, Pernar LI, Peyre SE, Helfrick JF, Vogelgesang KR, Graydon-Baker E, et al. Impact of online education on intern behaviour around joint commission national patient safety goals: a randomised trial. BMJ Qual Saf 2012 Oct;21(10):819-825 [FREE Full text] [doi: 10.1136/bmjqs-2011-000702] [Medline: 22706930]
73. Grad R, Leger D, Kaczorowski J, Schuster T, Adler S, Aman M, et al. Does spaced education improve clinical knowledge among family medicine residents? A cluster randomized controlled trial. Adv Health Sci Educ Theory Pract 2021 Aug;26(3):771-783 [FREE Full text] [doi: 10.1007/s10459-020-10020-z] [Medline: 33389233]
74. Gyorki DE, Shaw T, Nicholson J, Baker C, Pitcher M, Skandarajah A, et al. Improving the impact of didactic resident training with online spaced education. ANZ J Surg 2013 Jun;83(6):477-480 [FREE Full text] [doi: 10.1111/ans.12166] [Medline: 23617607]
75. Jaunay L, Zerr P, Peguin L, Renouard L, Ivanoff A, Picard H, et al. Development and evaluation of a new serious game for continuing medical education of general practitioners (Hygie): double-blinded randomized controlled trial. J Med Internet Res 2019 Nov 20;21(11):e12669 [FREE Full text] [doi: 10.2196/12669] [Medline: 31746775]
76. Teo WZW, Dong X, Yusoff SKBM, Das De S, Chong AKS. Randomized controlled trial comparing the effectiveness of mass and spaced learning in microsurgical procedures using computer aided assessment. Sci Rep 2021 Feb 02;11(1):2810 [FREE Full text] [doi: 10.1038/s41598-021-82419-6] [Medline: 33531563]
77. Raffoul A, Vitagliano JA, Sarda V, Chan C, Chwa C, Ferreira KB, et al. Evaluation of a one-hour asynchronous video training for eating disorder screening and referral in U.S. pediatric primary care: a pilot study. Int J Eat Disord 2022 Sep;55(9):1245-1251 [FREE Full text] [doi: 10.1002/eat.23766] [Medline: 35781822]
78. Li W, Zhang K, Yao S, Xie X, Han W, Xiong W, et al. Simulation-based arthroscopic skills using a spaced retraining schedule reduces short-term task completion time and camera path length. Arthroscopy 2020 Nov;36(11):2866-2872 [FREE Full text] [doi: 10.1016/j.arthro.2020.05.040] [Medline: 32502713]
79. Kocyigit H, Karagozoglu S. Effects of scenario-based high fidelity and repeated simulation methods on the medical error tendency, self-efficacy and state anxiety levels of nursing students. Int J Caring Sci 2022;15(2):1168-1177.
80. Akdemir A, Zeybek B, Ergenoglu AM, Yeniel AO, Sendag F. Effect of spaced training with a box trainer on the acquisition and retention of basic laparoscopic skills. Int J Gynaecol Obstet 2014 Dec;127(3):309-313 [FREE Full text] [doi: 10.1016/j.ijgo.2014.07.015] [Medline: 25176414]
81. Ugwa E, Kabue M, Otolorin E, Yenokyan G, Oniyire A, Orji B, et al. Simulation-based low-dose, high-frequency plus mobile mentoring versus traditional group-based trainings among health workers on day of birth care in Nigeria; a cluster randomized controlled trial. BMC Health Serv Res 2020 Jun 26;20(1):586 [FREE Full text] [doi: 10.1186/s12913-020-05450-9] [Medline: 32590979]
82. Laynor G. Are spaced repetition study tools changing health professions education? J Electron Resour Med Libr 2023 Aug 08;20(2-3):94-97. [doi: 10.1080/15424065.2023.2244364]

XSL•FO
RenderX

83. Spicer JO, Ujunwa KC, Witt LS, Meka J, Gooding HC. Using instructor-developed study resources to increase evidence-based learning strategies among medical students: a mixed-methods study. Med Teach 2023 Dec;45(12):1380-1386 [FREE Full text] [doi: 10.1080/0142159X.2023.2218537] [Medline: 37270765]
84. Wlodarczyk S, Dhaliwal G. The learning sciences meet the learning health system. JAMA Netw Open 2022 Jul 01;5(7):e2223113 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.23113] [Medline: 35881402]
85. Finn E, Ayres F, Goldberg S, Hortsch M. Brave new e-world: medical students' preferences for and usage of electronic learning resources during two different phases of their education. FASEB Bioadv 2022 May;4(5):298-308 [FREE Full text] [doi: 10.1096/fba.2021-00124] [Medline: 35520393]
86. Wu JH, Gruppuso PA, Adashi EY. The self-directed medical student curriculum. JAMA 2021 Nov 23;326(20):2005-2006 [FREE Full text] [doi: 10.1001/jama.2021.16312] [Medline: 34724030]
87. Medical school year two questionnaire 2022 all schools summary report. Association of American Medical Colleges. URL: https://www.aamc.org/data-reports/students-residents/report/year-two-questionnaire-y2q [accessed 2024-01-02]
88. Gamlin R. Build or buy? Curricular design and external basic sciences learning resources. Acad Med 2017 Apr;92(4):430 [FREE Full text] [doi: 10.1097/ACM.0000000000001597] [Medline: 28350600]
89. Carpenter SK, Cepeda NJ, Rohrer D, Kang SHK, Pashler H. Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction. Educ Psychol Rev 2012 Aug 4;24(3):369-378. [doi: 10.1007/s10648-012-9205-z]
90. Krishnamurthy K, Selvaraj N, Gupta P, Cyriac B, Dhurairaj P, Abdullah A, et al. Benefits of gamification in medical education. Clin Anat 2022 Sep;35(6):795-807 [FREE Full text] [doi: 10.1002/ca.23916] [Medline: 35637557]
91. Thillainadesan J, Le Couteur DG, Haq I, Wilkinson TJ. When I say … microlearning. Med Educ 2022 Aug;56(8):791-792 [FREE Full text] [doi: 10.1111/medu.14848] [Medline: 35654438]
92. Ng MSP, Jabir AI, Ng TDR, Ang Y, Chia JL, Tan DNH, et al. Evaluating TESLA-G, a gamified, telegram-delivered, quizzing platform for surgical education in medical students: protocol for a pilot randomised controlled trial. BMJ Open 2023 Jun 28;13(6):e068740 [FREE Full text] [doi: 10.1136/bmjopen-2022-068740] [Medline: 37380211]

Abbreviations

cRCT: cluster randomized controlled trial
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
MCQ: multiple-choice question
PCP: primary care provider
PRISMA: Preferred Reporting Items of Systematic Reviews and Meta-Analyses
RR: relative risk
SMD: standardized mean difference
USMLE: United States Medical Licensing Examination
VR: virtual reality

Edited by N Cahill; submitted 26.02.24; peer-reviewed by A Henri-Bhargava, I Martínez-García, Y Asada; comments to author 21.05.24; revised version received 14.08.24; accepted 26.08.24; published 10.10.24.

Please cite as:
Martinengo L, Ng MSP, Ng TDR, Ang YI, Jabir AI, Kyaw BM, Tudor Car L
Spaced Digital Education for Health Professionals: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e57760
URL: https://www.jmir.org/2024/1/e57760
doi:10.2196/57760
PMID:

©Laura Martinengo, Matthew Song Peng Ng, Tony De Rong Ng, Yi-Ian Ang, Ahmad Ishqi Jabir, Bhone Myint Kyaw, Lorainne Tudor Car. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Technology-Based Interventions in Tobacco Use Treatment Among People Who Identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native: Scoping Review

Emily Hichborn[1*], MA; Avery Turner[1*], MS; Sarah Moore[1], PhD; Phoebe Gauthier[1], MA, MPH; Kathleen Bell[1], MSc; LaTrice Montgomery[2], PhD; Jesse Boggis[1,3], MPH; Chantal Lambert-Harris[1], MA; Elizabeth Saunders[1], PhD; Jesse Dallery[4], PhD; Bethany McLeman[1], BA; Lisa Marsch[1], PhD

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[2]College of Medicine, University of Cincinnati, Cincinnati, OH, United States

[3]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States

[4]College of Liberal Arts and Sciences, University of Florida, Gainesville, FL, United States

[*]these authors contributed equally

Corresponding Author:
Emily Hichborn, MA
Center for Technology and Behavioral Health
Geisel School of Medicine
Dartmouth College
Suite 315
46 Centerra Parkway
Lebanon, NH, 03766
United States
Phone: 1 603 646 7000
Email: Emily.G.Hichborn@KU.Edu

Abstract

Background: Although tobacco use has significantly declined in the general population, traditional tobacco use treatment uptake and success rates remain disproportionately low among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native. Technology-based interventions (TBIs) for tobacco use are promising alternatives to traditional tobacco use treatments.

Objective: This scoping review aims to investigate the extent to which the use of digital TBIs in tobacco use treatment research promotes health equity among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native.

Methods: This scoping review identifies US-based studies (between January 2000 and March 2021) that enlist TBIs for tobacco use treatment and include people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native at ≥50% of the sample when combined; features studies that are also race and ethnicity conscious; and highlights health equity–promoting insights from included studies.

Results: In 85% (22/26) of the studies, the largest proportion of the sample was African American/Black, most participants had low socioeconomic status, and recruitment was most commonly from medical settings. In total, 58% (15/26) of the studies were race and ethnicity conscious, and 67% (10/15) of these studies sought to partner with potential end users. An array of TBIs were represented; however, SMS text messaging was most prevalent. Most TBIs were combined with other evidence-based intervention components (eg, nicotine replacement therapy). Approximately one-third of the studies (8/26, 31%) required participants to have their own device or internet access. The majority were underpowered to detect substantial differences.

Conclusions: The modest number of studies, particularly for persons who identify as Hispanic/Latina/o and American Indian/Alaska Native, demonstrates the limited application of TBIs for tobacco use and that additional research is needed to determine the extent to which TBIs for tobacco use promote health equity among these populations.

XSL•FO
RenderX

International Registered Report Identifier (IRRID): RR2-10.2196/34508

(J Med Internet Res 2024;26:e50748) doi:10.2196/50748

KEYWORDS

health disparities; underrepresented; social determinants of health; tobacco use; technology-based interventions; scoping review; mobile phone

Introduction

Background

Pharmacological and behavioral treatments have demonstrated success in reducing tobacco use in the general population, although the use of these services remains disproportionately low among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native [1]. People who use tobacco and identify as African American/Black and Hispanic/Latina/o are more likely to have made a quit attempt in the past year than White people who smoke [2]; however, their attempts to quit are less likely to be successful due, in part, to their lower use of nicotine replacement therapy (NRT) and other cessation services [2,3]. Hispanic/Latina/o and American Indian/Alaska Native people who smoke are less likely to receive advice to quit from a health professional than White people who smoke [2]. People who identify as American Indian/Alaska Native have the highest smoking prevalence of any racial or ethnic group in the United States [4]. The lack of antitobacco laws on tribal lands and the predatory advertising practices used by tobacco companies to target individuals who identify as American Indian/Alaska Native contribute to the high rates of tobacco use in this group [5]. While effective pharmacological and behavioral treatments for tobacco use exist, higher smoking rates and lower quit success rates among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native indicate that traditional approaches to tobacco use have not yielded optimized and proportional benefits to these populations.

This confluence of factors suggests that novel approaches are needed to address tobacco use among members of racial and ethnic underrepresented groups [6]. Technology-based interventions (TBIs) include the use of technologies, such as computer-based and web-based interventions, SMS text messaging (TM), interactive voice recognition, smartphone apps, and other emerging technologies, to deliver interventions aimed at reducing tobacco use [6,7]. According to the National Telecommunications and Information Administration, 76.9% of Americans who identify as African American/Black, 76.7% of Americans who identify as Hispanic/Latina/o, and 74.6% of Americans who identify as American Indian/Alaska Native reported accessing the internet at least occasionally using any type of device in the past 6 months [8]. Because the National Telecommunications and Information Administration began collecting this information in 1994, internet use has dramatically increased across all racial groups in the United States [8]. Disparities persist in internet access among persons with disabilities, older persons, Americans with low socioeconomic status (SES), and Americans who identify as African American/Black and Hispanic/Latina/o; however, these groups increased their internet use between 2019 and 2021, while use remained stagnant for Americans who identify as White in the same time period [9].

Owing to widespread access to technology in the United States, TBIs are a promising approach to tobacco use treatment and provide an opportunity to promote health equity among marginalized and underserved communities [10]. TBIs increase reach to vulnerable populations by decreasing informational barriers and allowing patients on-demand access to innovative, cost-effective tobacco use treatments [11] without the traditional barriers to accessing treatment, including high costs of treatment and travel to clinics and low rates of treatment fidelity [12]. In addition, some TBIs allow clinicians and researchers greater access to real-time patient data, allowing them to provide more personalized tobacco use treatment [11] and the opportunity to prevent, detect, and resolve health disparities in a timelier manner [10]. The development and implementation of effective TBIs for positive health behavior change should result in decreased use and cost of health services over time [13]. Owing to the high prevalence of internet use among people who identify as racial or ethnic minority individuals and the increased access to tobacco use treatment provided by TBIs, TBIs are a promising approach to address tobacco use among members of underrepresented populations [11].

In addition to providing increased access to treatment, TBIs for tobacco use have demonstrated effectiveness [14]. Two recent Cochrane reviews on the use of mobile phones [15] and internet-based interventions [16] for tobacco use support both the prevalence and promise of TBIs. Whittaker et al [15] analyzed 12 studies that used mobile phone–based tobacco use treatment and found that smokers who were given access to mobile phone–based interventions were 1.7 times more likely to have remained abstinent 6 months after the treatment than smokers who did not have access to the TBI. Taylor et al [16] reviewed 67 randomized controlled trials (RCTs) that used internet interventions for tobacco use. They found that tailored and interactive internet-based tobacco use treatment increased smoking cessation at 6 months after the treatment when compared with nonactive controls, although they cautioned the interpretation of these findings due to high statistical heterogeneity [16]. In addition, they found that internet-based interventions combined with behavioral support were more effective than nonactive controls at increasing cessation rates 6 months after the treatment [16].

While TBIs pose an opportunity to promote health equity among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native, these communities have not historically been included in digital health research [11]. People who identify as African American/Black and American Indian/Alaska Native have traditionally been

XSL•FO
RenderX

exploited and traumatized through participation in research, resulting in a deep-seated mistrust in researchers [17,18]. It is imperative that researchers repair relationships with these individuals, as their input is critical for the development and dissemination of equitable and effective TBIs. Digital technologies that are not designed for, or tested with, samples that include members of underrepresented groups may inadvertently exacerbate existing health inequities [11]. Brewer et al [10] highlight the lack of socioculturally tailored TBIs that include people from diverse racial and ethnic backgrounds and warn of the downstream effects of using TBIs that are "acontextually" developed. Without engagement and input from underrepresented communities, TBIs may largely benefit 1 subgroup of the population while perpetuating or even worsening health-related outcomes for another [10]. For example, an analysis of SmokefreeTXT (developed by the National Cancer Institute), a national text-based tobacco use treatment, found that 12.9% of smokers who completed the program self-reported abstinence from tobacco use at the end of treatment [19]. However, African American/Black users of SmokefreeTXT were less likely than White users to report tobacco use abstinence at all time points during treatment [20]. While this study did not meet the criteria for our scoping review, findings such as these underscore the need to ensure that TBIs for tobacco use are being developed, tested, and disseminated equitably. Therefore, an exploration of the research on TBIs for tobacco use that includes people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native is warranted.

Study Aims

The following research question guided this scoping review: To what extent does the use of digital TBIs in tobacco use treatment research promote health equity among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native?

We operationally defined promoting health equity as the *inclusion* of individuals (≥50% of the sample) who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native in the research of TBIs for tobacco use, as well as the extent to which the research is *race/ethnicity conscious*. To explore whether this subset of research is race and ethnicity conscious, we determined whether the research was explicit about members of underrepresented groups, beyond their inclusion as study participants, in the title, introduction, methods, results, and discussion of the manuscript. We examined peer-reviewed literature published between January 2000 and March 2021 that included individuals who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native in tobacco use treatment research using TBIs. Given a lack of synthesized findings on the use of TBIs to address tobacco use among underrepresented populations, a scoping review is an ideal tool to provide an overview of TBI studies for tobacco use and explore whether TBIs included in the review promote health equity in tobacco use treatment among underrepresented groups. No current or underway scoping reviews or systematic reviews were identified in the preliminary search.

Methods

Original Scoping Review

Studies were initially identified as part of a larger scoping review on TBIs being used in substance use treatment with a majority sample of individuals who identify as African American/Black, Hispanic/Latina/o, American Indian/Alaska Native [21]. This review was conducted using the methodological framework proposed by Arksey and O'Malley for scoping reviews [22], and the results were reported according to the PRISMA-ScR (Preferred Reporting Items For Systematic Reviews And Meta-Analyses Extension for Scoping Reviews) guidelines. Hichborn et al [21] provide additional details on the protocol that covers this tobacco review and the larger scoping review.

Data Sources

Preliminary literature searches were completed in Google Scholar and PubMed, netting dozens of US-based peer-reviewed research studies that met the scoping review criteria. On the basis of the initial search, 5 electronic databases (MEDLINE, Scopus, Cochrane Library, CINAHL, and PsychINFO) were selected to identify tobacco use treatment research using TBIs (Multimedia Appendix 1). The literature searches were conducted by 2 Dartmouth College reference librarians.

Study Selection and Inclusion and Exclusion Criteria

Studies eligible for inclusion were peer-reviewed, qualitative, quantitative, and mixed methods studies; randomized trials; RCTs (efficacy and effectiveness); feasibility and acceptability pilots; formative development studies; secondary analyses (eg, mechanisms or moderators); and assessments. Studies were not included if they were considered protocol papers, papers discussing planned or future work, reviews, commentaries, editorials, opinion pieces, student theses, conference abstracts, book chapters, and guidelines. Eligible studies were US-based due to the context-specific and country-specific issues facing the individuals who are the focus of this review. Eligible studies were also conducted in the English language owing to the language fluencies of the study team. Owing to a clear and emerging increase after 2001 surrounding the number of studies published on the design and development of technology-based behavior change interventions [13], eligible studies were selected from those published between January 2000 and March 2021. In addition, eligible studies included a TBI designed for tobacco use treatment with individuals aged ≥12 years and had at least 50% of participants identifying as African American/Black, Hispanic/Latina/o, or American Indian/Alaska Native when combined [23]. Furthermore, studies solely focused on mental health, pharmacological interventions, cost evaluations, telephone counseling only, primary prevention interventions, and substances other than tobacco were excluded.

Screening and Selection Procedure

Following the initial literature search, all netted studies were uploaded to Endnote X9 (Clarivate Analytics) to remove duplicates [24] and were then transferred to Rayyan (Rayyan Systems, Inc), a web-based tool used to assist researchers in screening, selecting, and labeling studies for systematic reviews

XSL•FO
RenderX

[25]. Three teams consisting of ≥2 independent reviewers blinded to each other's selections first screened all identified studies for inclusion by reviewing the title and abstract (N=6897). Articles meeting the initial criteria were then screened at the full-text level (1159/6897, 16.8%). Through this process, the team identified 21 studies meeting inclusion criteria at both the title and abstract and full-text phases of review. A total of 5 additional studies were identified through background searches, for a total of 26 included studies. Across the selection procedure, all disagreements were addressed through discussion or with additional reviewers.

Data Extraction

Using a standardized extraction form (Multimedia Appendix 2) with prespecified extraction fields (eg, aims, study design, and year of publication), data were extracted from included studies by members of the scoping review research team. In addition, the inclusion of members of the specific racial and ethnic groups, nature of the TBIs, primary outcomes, and verbatim race- and ethnicity-conscious text segments related to the scoping review research questions were extracted. Notably, while scoping reviews do not typically assess the quality of included studies, we have included a summary of study outcomes to inform readers about the consistency of this subset of the literature with the larger body of research in terms of primary and secondary outcomes and to point readers to studies with statistically significant findings, falling far short of a systematic evaluation of the rigor of the analyses. Information gathered from each included study was documented on the extraction form using Excel (Microsoft Corporation). We piloted our data extraction spreadsheet with 2 studies to ensure consistency and allow for iterative revision of the extraction variable definitions before the extraction was completed. Quality assurance checks were completed for 100% (N=26) of the sample by a second team member who reviewed the extractions to ensure accuracy, consistency, and thoroughness, primarily focusing on fields with the potential for greater variability. One author was contacted for additional clarification on his program of research.

Race and Ethnicity Consciousness

All included studies were evaluated for evidence of race and ethnicity consciousness in the development, conduct, analysis, and presentation of research findings. Studies were considered race or ethnicity conscious if the published manuscript included consideration of or was explicit about at least one of the aforementioned underrepresented racial or ethnic groups in one or more manuscript sections (ie, title, introduction, method, results, and discussion; Textbox 1).

Textbox 1. Examples of race- or ethnicity-conscious practices in research.

Title

Signaling the focus on one or more racial or ethnic groups in study titles, for example, "'Every day I think about your messages': assessing text messaging engagement among Latino smokers in a mobile cessation program" [26].

Introduction

Provide epidemiological or other relevant information about one or more racial or ethnic groups in the literature review or enlist a theory that is described as one that may help address health disparities among underrepresented groups. An example of this includes sociological trust theory—a bridge between a broad lens of culturally informed design and attention to trust or distrust.

Methods

Race- and ethnicity-conscious methods may include references to recruitment or retention efforts aimed at racial and ethnic groups, such as cultural tailoring of materials or consideration of matching staff race or ethnicity to that of the sample participants (in telemedicine appointments, and in the animations seen in virtual reality, and in computer games), and assessment for measurement equivalence. Race- or ethnicity-conscious analytic plans may include conducting separate analyses for each race and ethnicity, focusing on within-group differences rather than race comparisons, using stratification methods that balance each race and ethnicity across treatment arms, or in some other way considering race or ethnicity in the plan for analyzing the data.

Results

Present findings in ways that highlight differences and similarities for members of different racial or ethnic groups.

Discussion

Interprets findings for members of racial or ethnic groups by locating results in the context of other development or treatment literature.

Data Synthesis and Analysis

On the basis of the guiding research questions regarding inclusion and race and ethnicity consciousness, tables were populated with data extracted relevant to the research questions (ie, refer to Multimedia Appendices 3 [26-51] and 4 [26-51]). The remaining data extracted served to populate additional tables that helped to describe the studies (Multimedia Appendix 5 [26-51]) as well as the range and nature of the TBIs (Multimedia Appendix 6 [26-51]).

Results

Descriptions of Included Studies

Studies that met eligibility criteria were published between January 2000 and March 2021 and represent a range of research designs: RCTs (12/26, 46%) [27-38]; randomized trials (4/26, 15%) [39-42]; formative and development studies (4/26, 15%) [30,34,36,43], including Mason et al [30] and Orr et al [34], which are also counted among RCTs as publications present findings for both designs; feasibility, acceptability, and effectiveness studies (3/26, 12%) [26,37,44]; secondary data

XSL•FO
RenderX

analyses (3/26, 12%) [45-47]; and qualitative analysis (1/26, 4%) [48]. The TBIs were most often combined with other evidence-based intervention components: NRT (11/26, 42%) [27-29,32-34,36,40-43], pre- or postcessation counseling (9/21, 35%) [27,28,32-34,36,41,43,44], printed informational manual, brochure, or reports (5/26, 19%) [28,29,35,36,40], and contingency management (3/26, 12%) [27,39,43]. About 27% (7/26) were stand-alone interventions [26,30,37,38,49-51]. Figure 1 [52-54] shows an overview of the identification of studies.

Figure 1. Identification of studies. TBI: technology-based intervention. *One study was identified during a review of references after the initial search but was not ultimately included in this review. **Includes studies that focused on substances other than tobacco reviewed separately (see [21]). Tobacco use treatment literature was separated from the substance use treatment literature because until recently, many clinicians and scientists regarded tobacco use disorder and substance use disorder as separate. Substance use clinicians face barriers such as limited training on tobacco use treatment and limited resources for addressing tobacco use in substance use treatment settings, resulting in a focus on attaining abstinence from all substances other than tobacco [23, 24]. Clinicians and scientists have advocated for the inclusion of tobacco use treatment in substance use treatment settings, however, given the distinct way in which tobacco use is treated in systems of care compared to other substance use, we examined the tobacco use studies separately.

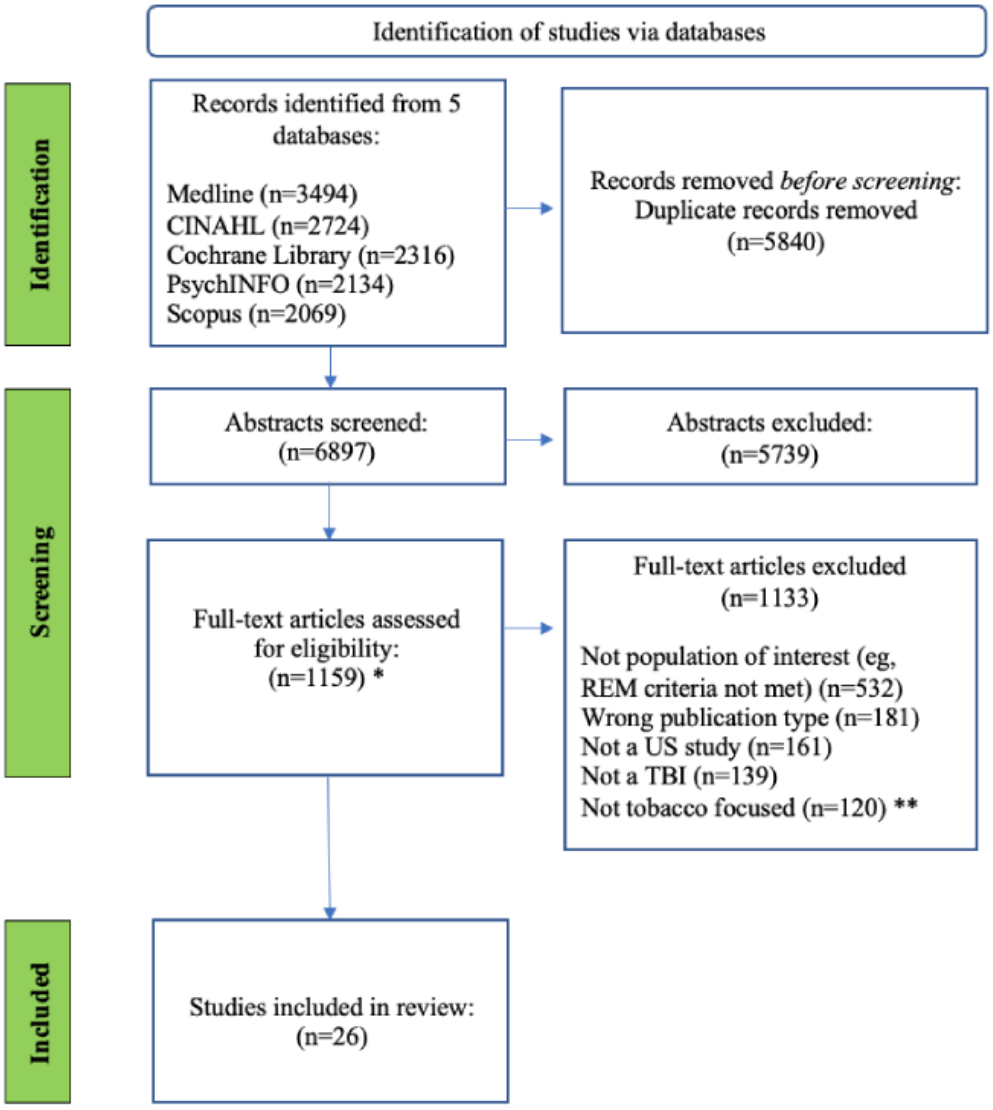

XSL•FO
RenderX

Descriptions of TBIs for Tobacco Use

Multimedia Appendix 6 presents detailed information, including the location of TBI delivery. The range of digital tools supporting the TBIs included apps accessed via the internet (3/26, 12%) [27,43,44], computer-delivered programs (3/26, 12%) [28,29,39], SMS TMs (16/26, 62%) [26,30-37,40,45-49,55], virtual reality (VR; 2/26, 8%) [42,50], DVDs (1/26, 4%) [51], and PDAs (1/26, 4%) [38]. The dosage of exposure and intervention intensity varied widely: apps ranged from 2 times daily for 30 days [44] to 2 times daily for 7 weeks [27,43]; computer-delivered programs ranged from 1 [39] to 8 weekly sessions [28]; SMS TM interventions ranged from 6 to 9 TMs over 5 days [30] to 4 TMs per day over 18 months [41]; VR programs ranged from 7 [50] to 10 [42] weeks of 45 [50] to 60 [42] minute sessions; the lone DVD program exposed participants to two 60-minute videos with 1-month latency between views [51], and the lone PDA-delivered intervention exposed participants to one 20-minute meditation per day for 2 weeks [38]. Of 26 studies, 10 (38%) study teams sought to partner with potential end users, engaging them in the development or adaptation of a TBI [28,34,36,37,39,43,44,48-50] and 12 (46%) study teams did not [27,29,30,32,33,35,38,40-42]. The remaining 4 (15%) studies were secondary analyses by Mason et al [30] and therefore not double counted here. Finally, almost all included studies can be grouped by racial or ethnic cultural focus: 23% (6/26) African American/Black [30,31,45-47,51], 8% (2/26) American Indian/Alaska Native [34,41], 4% (1/26) Hispanic/Latina/o [26] *or* a handful of target populations: 19% (5/26) pregnant and parenting individuals [32,35,37,39,49]; 15% (4/26) people living with comorbidities (ie, HIV [28,36]; schizophrenia, schizoaffective disorder, or psychotic disorder [43]; and posttraumatic stress disorder [27]), 8% (2/26) people experiencing unemployment or economic disadvantage [29,33], and 4% (1/26) adolescents [50]. Of the 26 studies, 5 (19%) [38,40,42,44,48] recruited convenience samples.

Summary of Study Outcomes

The strictest outcome definition in the field of tobacco use cessation research, bioverified smoking cessation (BVSC), was a main outcome measure for 11 (42%) of the 26 studies [26-29,32,33,35,39,41,42,44]. Of 26 studies, 2 (8%) TBIs, including computer-delivered 5As [39] and tailored telemedicine plus NRT with phone counseling [33], demonstrated statistically significant differences relative to respective comparators (P=.02 [39], P=.03 [33]). Bioverified smoking reduction was the primary outcome measure for 1 (4%) study [42]. Smoking rates were significantly lower (P=.045) for the VR skills training group compared with the NRT-only group at the end of treatment. Self-reported smoking cessation (SRSC) was the other primary outcome measure of note across studies [27,29,33,37,38,40,41,43,44,50,51]. TBIs that demonstrated statistically significant differences in SRSC were SMS TM computer-assisted counselling (P=.02) [29]; SmokeFreeTxt (P=.003) [40]; tailored TM with NRT, and phone counseling (P=.03) [33]; VR TBI; The Breathing Room (P=.045) [42]; and Brief mindfulness practice (P<.001) [38]. Self-reported smoking reduction was the primary outcome measure for the SMS TM program of research by Mason et al [30,31]. In both reports, statistically significant differences were demonstrated (P=.04 [30], P=.045 [31]), favoring personalized SMS TM over general health habits SMS TM. It is notable that only 7 (27%) of the 26 studies were sufficiently powered to detect significant findings [28-31,33,35,51]. Multimedia Appendix 5 presents more details, including whether studies were powered to detect differences.

Tobacco TBI Access and Inclusion: Promoting Health Equity

A total of 13 (50%) of the 26 study samples include at least 75% of one underrepresented group targeted in this review [26,27,30-34,36,39,41-47] (articles by Mason et al counted [51] as 1 study). Of 26 studies, 1 (4%) had 100% Hispanic/Latina/o [26], 1 (4%) had 100% African American/Black [51], 2 (8%) had 100% American Indian/Alaska Native [34,41], and 14 (54%) included at least 75% African American/Black participants [27,30-33,36,39,42-47]. In 22 (85%) of the 26 studies, the largest proportion of the sample was African American/Black. Of 26 studies, 9 (35%) studies included Hispanic/Latina/o participants [26,28,29,33,40,42,48-50]; however, only 2 [26,50] included ≥50% of Hispanic/Latina/o participants. Furthermore, nearly all studies other than those involving adolescents (6/26, 23%) and those that did not describe SES variables (4/26, 15%) reported that most of their participants were either unemployed, had annual incomes below the poverty line, did not have educations beyond the high school level, or had public health insurance or no insurance.

The included studies used a variety of recruitment strategies. The most common of which was recruitment from medical settings (11/26, 42%) [28,32,35,37,39-41,43,44,48,49]. Of the 26 studies, 7 (27%) studies used community-based approaches, such as placing flyers in churches or supermarkets; approaching individuals in heavily concentrated public areas; and placing advertisements on the radio, in newspapers, and on Craigslist [26,33,38,41-43,51]. Two (8%) of the 26 studies relied primarily on snowball or respondent-driven sampling [30,31,45-47,51], 2 (8%) other studies recruited participants primarily from schools or universities [34,50], and 1 (4%) study did not describe its recruitment plan [36].

Retention strategies were defined as any effort to collect follow-up data from participants in studies that involve >1 time point [56]. All included studies except for one [49] attempted to collect follow-up data from participants. Compensation, provision of a device or internet access required to use the TBI (platform availability), and technical support provided were highlighted retention strategies because of their relevance for maintaining engagement in the included studies. Most of the (18/26, 69%) included studies compensated participants [27-31,35-39,41,43-47,50,51], and 2 (8%) of the 26 studies additionally provided travel vouchers [28,36]. Of 26 studies, 8 (31%) studies stipulated that a candidate was eligible only if they owned or had access to a device to engage with the TBIs [26,28,32,36,40,44,48,51]. Several (16/26, 62%) studies did not require participants to have access to, own a device, or have internet service (eg, TBI accessible at the study site) [29,42,50]. Technical support was provided to participants in 13 (50%) of the 26 included studies [27,28,30,31,36-38,43-47,50], facilitating continued engagement with the TBI.

XSL•FO
RenderX

Race and Ethnicity Consciousness of Included Studies: Promoting Health Equity

Multimedia Appendix 4 provides further details regarding the 26 studies: 4 (15%) studies signaled the focus on an underrepresented group in study titles [26,34,41,51] (eg, A randomized trial to reduce smoking among American Indians in South Dakota: The walking forward study [41]), 12 (46%) explicitly referenced underrepresented groups in the Introduction section [26,30-34,36,41,45,46,51], 11 (42%) exemplified race- or ethnicity-conscious methods [26,30-32,34,39,41,45,46,50,51], 7 (27%) specifically characterized findings among underrepresented groups in the Results section [26,28,32,36,38,39,51], and finally, 11 (42%) addressed a specific underrepresented group in the Discussion or Conclusion sections [26,30,31,34,36,39,41,45-47,51].

Discussion

Principal Findings

This scoping review identified 26 studies that were published between January 2000 and March 2021, used a TBI designed for tobacco use treatment with individuals aged ≥12 years, and had at least 50% of participants identifying as African American/Black, Hispanic/Latina/o, or American Indian/Alaska Native when combined. Studies included in this scoping review, by definition, promoted health equity by including a majority of members of underrepresented groups in study samples. Slightly more than half (15/26, 58%) were found to be explicit about the inclusion of these members, as well as the potential impact of the TBIs for members of these groups (ie, race or ethnicity conscious). Although predominantly RCTs, only a handful of included studies (7/26, 27%) were powered to detect significant findings. These studies used a variety of outcomes to assess the effectiveness of TBIs. The range and nature of TBIs included in this review were found to be generally representative of TBIs under investigation for use with people experiencing the full range of substance use disorders (SUDs). The review identified a modest number of studies focused on TBIs for tobacco use treatment that include a majority of people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native. In most (22/26, 85%) of the identified studies, participants who identified as African American/Black made up the largest proportion of the sample and most participants reported having low levels of education and income. Participants were recruited from a variety of settings, although most frequently from medical settings, and most (21/26, 81%) of the included studies engaged in efforts to retain participants. Furthermore, multiple study methods' insights designed to promote health equity warrant attention.

Inclusion in Research

Despite the disproportionate impact of tobacco use on people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native [57-60]; the exponential growth in the development of and research on TBIs for behavioral health [13], SUDs [61,62]; and tobacco use specifically [55,63-65] in the United States since 2000, one key finding of this scoping review is that *a modest number of studies met inclusion criteria* (N=26). In 85% (22/26) of the included studies, the largest proportion of participants identified as African American/Black. Only 4 (15%) of the 26 studies included samples with more than half of the participants identifying as Hispanic/Latina/o or American Indian/Alaska Native. *The dearth of research on TBIs for tobacco use treatment among all racial and ethnic groups is problematic: for people who identify as Hispanic/Latina/o or American Indian/Alaska Native, it is profound.* Recruitment of sufficient numbers of racial and ethnic minority participants is imperative to understanding how research affects underrepresented groups [66]. Underrepresentation in research limits the extent to which members of these groups benefit from advances in tobacco use treatment [67].

The prevalence of low SES among participants in the included studies is also notable. Given the extensively supported association between tobacco use and SES factors, such as education and poverty [68-71], the finding that approximately one-third of the studies (8/26, 31%) required participants to own or have access to a device or internet to engage with the TBIs suggests a recommendation for reducing barriers to access among low SES populations in future studies. Despite often-repeated statistics regarding universal access, Americans with lower incomes, who are disproportionately members of underrepresented racial and ethnic groups, have lower levels of technology adoption [72-74]. This pattern holds for smartphones, desktop or laptop computers, home broadband access, and tablet computers. Although not the responsibility of investigators alone, one way that may lower barriers to study enrollment is to make devices and the internet more accessible to participants. Relatedly, removing the requirement for participants to use a specific technological device and instead allowing participants to use a currently owned or preferred device may help promote health equity. However, it is important to highlight that while providing the technology required may reduce some barriers to engagement, other considerations, such as digital literacy, serious mental illness, and organizational policies, may also have a significant impact on real-world populations and should be considered in both future research and clinical applications [75]. Another practice for including individuals who identify as African American/Black, Hispanic/Latina/o, American Indian/Alaska Native is to use community involvement and outreach methods (including respondent-driven sampling, community-based participatory research, and recruiting at locations within the target group's community), which have been shown to be more effective than other methods to reduce study enrollment barriers, thus increasing intervention participation [66]. *The most common recruitment strategy among studies included in this review was recruitment from medical settings,* which may favor insured persons and limit access for members of underrepresented groups due to pervasive mistrust of the medical community [76-80], particularly with respect to research [81]. On the basis of our review of the literature, the inclusion of underrepresented groups recruited from a broad array of settings with a diverse array of methods will help ensure increased representation and generalizability. Suggested alternative settings include community treatment programs, centers, and housing; social media; places of worship; and health fairs through community-based participatory research and respondent-driven sampling [66,82,83].

XSL•FO
RenderX

TBI Findings

The heterogeneity of TBIs for tobacco use included in this review appears to be representative of the larger universe of TBIs for SUDs [7], as does *the high number of TM interventions relative to other platforms* [65]. SMS TM is less technologically complex compared with other digital platforms and being cost-effective, it is also the most popular mobile phone feature nationally among people in SUD treatment, as well as primary care [7,84]. Importantly, 35% of US Latina/o adults and 24% of African American/Black adults rely on smartphones as their primary means of internet access at home [85]. Therefore, mobile health interventions, including SMS TM interventions, may be a first-tier option to address tobacco use among people who identify as Latina/o and possibly also African American/Black if the goal is to lower barriers to access for all [86].

Consistent with the findings of systematic reviews of mobile phone–based [15] and web-based interventions [16] for tobacco use, *the TBIs in this scoping review were most often combined with other evidence-based intervention components*, thus functioning in a supporting role rather than as stand-alone interventions. Of the studies assessing the utility of a TBI (eg, SMS TM) combined with evidence-based treatment, *the combinations of TBI and NRT or TBI, NRT, and cessation counseling* were most commonly associated with statistically significant differences relative to comparators in terms of BVSC or SRSC. Tobacco use pharmacotherapy combined with behavioral intervention is the recommended standard of care [87-89] and is considered a critical component of any quit attempt [90], including for individuals who identify as African American/Black, Hispanic/Latina/o, American Indian/Alaska Native [91,92]. However, pharmacotherapy for tobacco use is underused among members of these groups due to varying attitudes and beliefs about these therapies. For example, people who identify as African American/Black have lower confidence in their effectiveness relative to White people who smoke [93]; people who identify as Hispanic/Latina/o are more likely to distrust or reject medications due to a general avoidance of medications [94] coupled with a belief that willpower alone is sufficient to quit smoking [95]; and people who identify as American Indian/Alaska Native express apprehension about pharmacotherapy side effects and effectiveness [96], as well as distrust related to racism [97]. Two (8%) of the 26 included studies [26,51] mentioned disparities in the use of pharmacotherapy for members of these groups, and 1 (4%) [41] discussed at greater length the negative attitudes about NRT among people who identify as American Indian and the inconsistent availability of NRT on or near American Indian reservations. If TBIs are largely intended to support other evidence-based tobacco use therapies, it is important that *greater attention is given to existing ethnocultural concerns surrounding pharmacotherapies such as NRT* [1]. TBIs that are meant to be combined with pharmacotherapies should address these ethnocultural concerns and, when appropriate, emphasize the safety and efficacy of pharmacotherapies and attempt to dispel myths (ie, people who use NRT lack willpower) that may perpetuate the community's negative perception of them. It is notable, however, that 2 stand-alone TBIs, including computer-delivered 5As [39] and SMS TM intervention by Mason et al [30,31], were found efficacious (BVSC at 10 weeks [39] and self-reported smoking reduction at 6 months [30,31]) compared with controls. These TBIs warrant closer attention, given their potential to impact tobacco use independently.

Race and Ethnicity Consciousness

In her book *Race After Technology* [98], Benjamin poses the question "Can we develop a race-conscious orientation to emerging technology, not only as a mode of critique but as a prerequisite for designing technology differently?" This challenge provided direction for the race and ethnicity consciousness analysis. Our intention was to highlight areas for improvement, as well as the work of researchers whose efforts provide insights into how TBIs for tobacco use have been or may be designed to promote health equity. By reviewing the netted studies that met the *necessary* health equity–promoting condition of *including* underrepresented racial and ethnic group members, we sought to explore the extent to which these studies included consideration of and were explicit about at least one of the aforementioned underrepresented racial or ethnic groups in one or more manuscript sections (ie, title, introduction, method, results, and discussion; Textbox 1). Our finding that slightly more than half of these published manuscripts (15/26, 58%) were race or ethnicity conscious in at least one of the manuscript sections is a promising sign that researchers in this field are beginning to consider the influence of race and ethnicity on TBIs for tobacco use treatment, in addition to their inclusion of members of underrepresented groups.

The predicted growth of both Hispanic/Latina/o and African American/Black population groups, as well as the predicted shrinkage of the non–Hispanic/White population [99], increases the urgency for research that addresses health disparities through best practices for researching diverse groups [100]. Given the limited inclusion of members of underrepresented populations in biomedical research [11], there is a critical learning opportunity when a study does enroll a sample with racial and ethnic heterogeneity. Most studies (15/26, 58%) in this review allocated some manuscript space to acknowledging or being conscious of the fact that race or ethnicity might impact their findings. For example, Burlew et al [101] discuss barriers to methodologically sound health disparities research encountered at each stage of a research project from literature review through responsible dissemination to lay audiences and offer recommendations for addressing these challenges. A qualitative analysis of race-and ethnicity-conscious text segments excerpted from the studies included in this review, as well as a larger scoping review focusing on all other substance use (manuscript in preparation), is planned to explore the use of "best" evidence-based practices when conducting research with diverse groups [66,101,102] and to underscore insights that may help other researchers design studies with outcomes that promote health equity for members of underrepresented groups.

Recent meta-analyses have concluded that *evidence-based behavioral interventions adapted for a specific target group are more effective for that population than generic versions of the same interventions* [103,104]. Culturally appropriate strategies are important when addressing tobacco-related

XSL•FO
RenderX

disparities among members of underrepresented groups, as tobacco use is associated with cultural norms and SES factors, such as education and poverty [105]. In the sample of articles in this review, many of the race- and ethnicity-conscious studies explored the role of cultural traditions, values, norms, and behaviors when providing a rationale for a given use, adaptation, or interpretation of an intervention for members of an underrepresented group. Mason et al [46] offered an exceptional example of exploring culture and context to provide a rationale for an approach to research and intervention. Through the ecological momentary assessment data collected during their RCT exploring the efficacy of their culturally adapted SMS TM intervention for tobacco use among adolescents who identify as African American/Black, the research group further explored the role of culture and context through secondary data analysis with attention to improved understanding of the time-varying effects of urban neighborhoods (tobacco outlet density and evaluations of perceived safety), specifically for adolescent African American/Black people who smoke. The finding that the TBI weakened the effect of tobacco outlet density on smoking for the treatment condition relative to the control highlights the potential that technology, such as ecological momentary assessment, can better interpret nuanced findings and tailor TBIs to align with real-life circumstances of members of underrepresented groups.

Avoiding the trap of "one size fits all" [106], culturally adapted TBIs may enhance the effectiveness of an intervention for people who identify as African American/Black, Hispanic/Latina/o, American Indian/Alaska Native. In their work to tailor an SMS TM intervention for tobacco use in rural American Indian/Alaska Native communities, Orr et al [34] described the results of focus groups with American Indian/Alaska Native tribal college students in Montana. Focus group members strongly disliked "text speak" (eg, "B4" for the word "before") and strongly preferred positive, encouraging messages. These are significant observations that, if undiscovered, could have resulted in extensive resource waste to develop a TBI without impact, or worse, which could have exacerbated health disparities due to perceptions that evidence-based treatments are rarely aligned with American Indian/Alaska Native culture.

Finally, the example provided by Webb Hooper et al [51] regarding the power of exploring culture to interpret the impact of culturally sensitive interventions was instructive. Deep structural adaptations to the TBI were sensitive to the ethnocultural context and focused on topics that included African American/Black smoking statistics, physiological findings, smoking norms, pharmacotherapy concerns, religion and spirituality, family and collectivism, unique stressors, comorbid addiction, environmental influences, targeted marketing, menthol cigarettes, weight concerns, and working against the tobacco industry. Although the study did not find statistically significant differences between the tobacco cessation DVD intervention designed for a general audience and the culturally specific DVD intervention, the researchers did find that exposure to the latter led to greater perceived risks of smoking, in particular the disproportionate risks for African American/Black people compared with White people. Because perceptions of risk are posited to affect behavior change [107], this finding is promising and provides future direction as this was the first study to test the efficacy of the TBI content for its impact on people who identify as African American/Black.

Race- and ethnicity-conscious research makes members of underrepresented groups visible and audible. Tobacco researchers should consider the use of frameworks, such as the Public Health Critical Race Praxis [108], that offer new modes of inquiry and broaden the scope of research priorities to improve health equity. Strategies consistent with those proposed by Public Health Critical Race Praxis include prioritizing the perspectives of members of underrepresented groups through attention to *voice*. Many studies (10/26, 38%) included in this review engaged end users and community members in the assessment of the feasibility, acceptability, usability, and refinement of TBIs, raising the voices of these historically underrepresented groups. In a similar vein, several studies (6/26, 23%) elicited qualitative feedback, allowing for the exploration of different perspectives and a shift of investment away from White normativity [108]. Furthermore, data analytic plans that include separate analyses of racial and ethnic groups or examine race and ethnicity as moderators of outcomes to improve understanding of the social factors that support continued inequity in care also increase the visibility and raise the voices of members of these groups. In several included studies, race is treated as a covariate or controlled, effectively muting the potential impact of race. To promote health equity in research on TBIs for tobacco use treatment, it is necessary to make every effort to ensure that members of underrepresented groups are recruited and included in our samples, that their perspectives on our TBIs are elicited and valued, that whether effect sizes are smaller when compared with those for the overall study sample (and non–Hispanic/White individuals) we consider adaptations to increase effectiveness for these groups, that analyses are planned to enable meaningful interpretation of findings for these groups, and that we plan for dissemination of findings that reaches the communities most likely to benefit.

It is critical to observe that various social determinants of health, including socioeconomic factors, that is, low income and poverty, low educational attainment and literacy, and impediments to accessing health care (eg, lack of transportation, lack of insurance, and lack of trust in health care providers), all have the potential to mediate or moderate the effects of our available evidence-based treatments for tobacco use in the United States. The features of and pathways by which these societal conditions affect health can potentially be altered by evidence-based and informed action. In the case of TBIs for tobacco use in the near term, the low-hanging features and pathways include the provision of tools at no cost to the user; provision of interventions with search interfaces that have high error tolerance regarding spelling, textual information accompanied by audio narration, linear versus hierarchical navigation [74]; and provision of access to interventions without the requirement of transportation (eg, SMS TM instead of computer station at clinic or medical site). In the long term, the tobacco industry continues to profile members of underrepresented communities, especially African American/Black and American Indian/Alaska Native, through its disproportionate promotion of menthol products, the

XSL•FO
RenderX

increased likelihood of tobacco advertisement displays, as well as the decreased likelihood that these communities are protected by smoke-free laws compared with areas with fewer African American/Black residents [109]. Thus, more efficacious, culturally sensitive, and competent educational initiatives focusing on children, as well as ongoing advocacy for policy changes responsive to the tobacco industry's tactics, are simultaneously warranted.

Limitations

There are several limitations to our methodology. First, the search terms used may have limited the findings included in our review, introducing selection bias. Relatedly, including only studies with ≥50% of underrepresented groups likely also precluded the discovery of other relevant, race- or ethnicity-conscious articles [20,64]. Only 7 (27%) of the 26 included studies followed participants for at least 6 months [33,34] (refer to Cochrane reviews), demonstrating the early-stage nature of the included research. Moreover, as many of the studies represent formative research, investigation of the potential differences for participants based on race or ethnicity proxies (unequal economic, educational, and social opportunities) is not to be expected due to small sample sizes. Furthermore, while we recognize that most of the included studies have small sample sizes and are underpowered, we believe that the findings are vital to understanding the state of the research base and emphasize the need for larger trials to remedy these shortcomings. In addition, only English language studies were included in the data set. While we recognize that this approach is common in systematic reviews, we also recognize that it can lead to biased interpretations of findings. In future research, we recommend the expansion of languages and countries included in the review, specifically Spanish and Native US languages, such as Navajo.

Conclusions

Innovation and inequity can go hand in hand [98]. By identifying and comparing publications of TBIs for tobacco use that *include* individuals who identify as African American/Black, Hispanic/Latina/o, American Indian/Alaska Native, we are taking a first step to interrupt the perpetuation of disparities by exploring participant access, equity in outcomes, and the cultural relevance of the research [101]. By exploring the extent to which these studies are explicit about race and ethnicity and the impacts of TBIs on particular people, we further highlight research efforts to promote health equity. By shining a spotlight on this modest body of research and featuring insights related to the promotion of health equity, we hope to encourage other researchers to participate in best practices when engaging diverse groups in research.

Acknowledgments

This scoping review was funded by National Institute on Drug Abuse (grant P30DA029926; principal investigator: L Marsch). The authors would like to thank research and education librarians Heather Blunt and Paige Scudder at the Biomedical Libraries of Dartmouth College for their expertise and support in honing the authors' research questions, developing the search strategy, and running the search through the selected databases, creating the data set used in this scoping review. The authors would also like to acknowledge Dr Kathy Burlew for her knowledge around best practices when researching diverse groups, as well as her personal encouragement with this scoping review.

Data Availability

The data sets generated during this study are available from the corresponding author upon reasonable request. One example of the search terms used in this review is provided in the Multimedia Appendix 1.

Authors' Contributions

SM conceptualized the review. EH, AT, SM, PG, KB, L Montgomery, CL-H, ES, and L Marsch were involved in the methodology of the review. EH, AT, SM, PG, KB, and L Montgomery conducted the investigation. EH, AT, SM, and CLH provided project administration and SM, L Montgomery, and JD supervised the review. EH, AT, SM, PG, KB, and L Montgomery validated the review. EH, AT, and SM wrote the original draft, and EH, AT, SM, PG, KB, L Montgomery, JB, CLH, ES, BM, and L Marsch participated in reviewing and editing the review. L Marsch acquired funding for the study.

Conflicts of Interest

L Marsch is affiliated with Square2 Systems and Pear Therapeutics. These relationships are extensively managed by her academic institution, Dartmouth College. The other authors report no financial relationships with commercial interests.

Multimedia Appendix 1
MEDLINE (Ovid) databases: Ovid MEDLINE and Epub ahead of print, in-process, in-data-review, and other nonindexed citations and daily, 1946 to March 29, 2021.
[DOCX File , 14 KB - jmir_v26i1e50748_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Data extraction of items.
[DOCX File , 15 KB - jmir_v26i1e50748_app2.docx]

Multimedia Appendix 3
Study design, methods, and results of scoping review studies.
[DOCX File , 1099 KB - jmir_v26i1e50748_app3.docx]

Multimedia Appendix 4
Technology-based interventions for tobacco use.
[DOCX File , 4496 KB - jmir_v26i1e50748_app4.docx]

Multimedia Appendix 5
Study technology-based intervention access and inclusion.
[DOCX File , 39 KB - jmir_v26i1e50748_app5.docx]

Multimedia Appendix 6
Race and ethnicity consciousness of included studies.
[DOCX File , 10 KB - jmir_v26i1e50748_app6.docx]

Multimedia Appendix 7
PRISMA-ScR Checklist.
[PDF File (Adobe PDF File), 96 KB - jmir_v26i1e50748_app7.pdf]

References

1. Hooper MW, Payne M, Parkinson KA. Tobacco cessation pharmacotherapy use among racial/ethnic minorities in the United States: considerations for primary care. Fam Med Community Health 2017 Oct 01;5(3):193-203. [doi: 10.15212/fmch.2017.0138]
2. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults - United States, 2000-2015. MMWR Morb Mortal Wkly Rep 2017 Jan 06;65(52):1457-1464 [FREE Full text] [doi: 10.15585/mmwr.mm6552a1] [Medline: 28056007]
3. Trinidad DR, Xie B, Fagan P, Pulvers K, Romero DR, Blanco L, et al. Disparities in the population distribution of African American and non-Hispanic White smokers along the quitting continuum. Health Educ Behav 2015 Dec;42(6):742-751 [FREE Full text] [doi: 10.1177/1090198115577376] [Medline: 25794519]
4. Martell BN, Garrett BE, Caraballo RS. Disparities in adult cigarette smoking - United States, 2002-2005 and 2010-2013. MMWR Morb Mortal Wkly Rep 2016 Aug 05;65(30):753-758 [FREE Full text] [doi: 10.15585/mmwr.mm6530a1] [Medline: 27491017]
5. Carroll DM, Soto C, Baezconde-Garbanati L, Huang LL, Lienemann BA, Meissner HI, et al. Tobacco industry marketing exposure and commercial tobacco product use disparities among American Indians and Alaska natives. Subst Use Misuse 2020;55(2):261-270 [FREE Full text] [doi: 10.1080/10826084.2019.1664589] [Medline: 31544562]
6. Linke SE, Larsen BA, Marquez B, Mendoza-Vasconez A, Marcus BH. Adapting technological interventions to meet the needs of priority populations. Prog Cardiovasc Dis 2016;58(6):630-638 [FREE Full text] [doi: 10.1016/j.pcad.2016.03.001] [Medline: 26957186]
7. Tofighi B, Abrantes A, Stein MD. The role of technology-based interventions for substance use disorders in primary care: a review of the literature. Med Clin North Am 2018 Jul;102(4):715-731 [FREE Full text] [doi: 10.1016/j.mcna.2018.02.011] [Medline: 29933825]
8. NTIA data explorer. National Telecommunications and Information Administration. 2024 Jun 06. URL: https://www.ntia.doc.gov/data/explorer#sel=internetUser&disp=map [accessed 2024-09-20]
9. Goldberg R. New NTIA data show enduring barriers to closing the digital divide, achieving digital equity. National Telecommunications and Information Administration. 2022 May 11. URL: https://www.ntia.gov/blog/2022/new-ntia-data-show-enduring-barriers-closing-digital-divide-achieving-digital-equity [accessed 2024-09-20]
10. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
11. Hamideh D, Nebeker C. The digital health landscape in addiction and substance use research: will digital health exacerbate or mitigate health inequities in vulnerable populations? Curr Addict Rep 2020 Sep;7:317-332. [doi: 10.1007/s40429-020-00325-9]

XSL•FO
RenderX

12. Marsch LA, Dallery J. Advances in the psychosocial treatment of addiction: the role of technology in the delivery of evidence-based psychosocial treatment. Psychiatr Clin North Am 2012 Jun;35(2):481-493 [FREE Full text] [doi: 10.1016/j.psc.2012.03.009] [Medline: 22640767]
13. Taj F, Klein MC, van Halteren A. Digital health behavior change technology: bibliometric and scoping review of two decades of research. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
14. Prutzman YM, Wiseman KP, Grady MA, Budenz A, Grenen EG, Vercammen LK, et al. Using digital technologies to reach tobacco users who want to quit: evidence from the national cancer institute's Smokefree.gov initiative. Am J Prev Med 2021 Mar;60(3 Suppl 2):S172-S184 [FREE Full text] [doi: 10.1016/j.amepre.2020.08.008] [Medline: 33663705]
15. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
16. Taylor GM, Dalili MN, Semwal M, Civljak M, Sheikh A, Car J. Internet-based interventions for smoking cessation. Cochrane Database Syst Rev 2017 Sep 04;9(9):CD007078 [FREE Full text] [doi: 10.1002/14651858.CD007078.pub5] [Medline: 28869775]
17. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897 [FREE Full text] [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
18. Pacheco CM, Daley SM, Brown T, Filippi M, Greiner KA, Daley CM. Moving forward: breaking the cycle of mistrust between American Indians and researchers. Am J Public Health 2013 Dec;103(12):2152-2159 [FREE Full text] [doi: 10.2105/AJPH.2013.301480] [Medline: 24134368]
19. Augustson E, Cole-Lewis H, Sanders A, Schwarz M, Geng Y, Coa K, et al. Analysing user-reported data for enhancement of SmokefreeTXT: a national text message smoking cessation intervention. Tob Control 2017 Nov;26(6):683-689. [doi: 10.1136/tobaccocontrol-2016-052945] [Medline: 27852892]
20. Robinson CD, Wiseman KP, Webb Hooper M, El-Toukhy S, Grenen E, Vercammen L, et al. Engagement and short-term abstinence outcomes among Blacks and Whites in the National Cancer Institute's SmokefreeTXT program. Nicotine Tob Res 2020 Aug 24;22(9):1622-1626 [FREE Full text] [doi: 10.1093/ntr/ntz178] [Medline: 31535690]
21. Hichborn EG, Moore SK, Gauthier PR, Agosti NO, Bell KD, Boggis JS, et al. Technology-based interventions in substance use treatment to promote health equity among people who identify as African American/Black, Hispanic/Latinx, and American Indian/Alaskan native: protocol for a scoping review. JMIR Res Protoc 2022 May 17;11(5):e34508 [FREE Full text] [doi: 10.2196/34508] [Medline: 35579930]
22. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
23. Huey SJJ, Polo AJ. Evidence-based psychosocial treatments for ethnic minority youth. J Clin Child Adolesc Psychol 2008 Jan;37(1):262-301 [FREE Full text] [doi: 10.1080/15374410701820174] [Medline: 18444061]
24. EndNote™. Clarivate. URL: https://clarivate.com/innovation-exchange/solution/endnote/ [accessed 2024-09-20]
25. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
26. Cartujano-Barrera F, Arana-Chicas E, Ramírez-Mantilla M, Perales J, Cox LS, Ellerbeck EF, et al. "Every day I think about your messages": assessing text messaging engagement among Latino smokers in a mobile cessation program. Patient Prefer Adherence 2019;13:1213-1219 [FREE Full text] [doi: 10.2147/PPA.S209547] [Medline: 31413549]
27. Hicks TA, Thomas SP, Wilson SM, Calhoun PS, Kuhn ER, Beckham JC. A preliminary investigation of a relapse prevention mobile application to maintain smoking abstinence among individuals with posttraumatic stress disorder. J Dual Diagn 2017;13(1):15-20 [FREE Full text] [doi: 10.1080/15504263.2016.1267828] [Medline: 27918881]
28. Shuter J, Morales DA, Considine-Dunn SE, An LC, Stanton CA. Feasibility and preliminary efficacy of a web-based smoking cessation intervention for HIV-infected smokers: a randomized controlled trial. J Acquir Immune Defic Syndr 2014 Sep 01;67(1):59-66 [FREE Full text] [doi: 10.1097/QAI.0000000000000226] [Medline: 25118794]
29. Prochaska JJ, Brown-Johnson C, Baiocchi M, Lazaro AS, Chieng A, Stinson S, et al. Treating tobacco dependence to aid re-employment among job-seekers: a randomized controlled trial. Prev Med 2020 Dec;141:106259 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106259] [Medline: 33022318]
30. Mason MJ, Campbell L, Way T, Keyser-Marcus L, Benotsch E, Mennis J, et al. Development and outcomes of a text messaging tobacco cessation intervention with urban adolescents. Subst Abus 2015;36(4):500-506. [doi: 10.1080/08897077.2014.987946] [Medline: 25551337]
31. Mason M, Mennis J, Way T, Zaharakis N, Campbell LF, Benotsch EG, et al. Text message delivered peer network counseling for adolescent smokers: a randomized controlled trial. J Prim Prev 2016 Oct;37(5):403-420. [doi: 10.1007/s10935-016-0439-2] [Medline: 27388626]
32. Forinash AB, Yancey A, Chamness D, Koerner J, Inteso C, Miller C, et al. Smoking cessation following text message intervention in pregnant women. Ann Pharmacother 2018 Nov;52(11):1109-1116. [doi: 10.1177/1060028018780448] [Medline: 29857773]

XSL•FO
RenderX

33. Vidrine DJ, Frank-Pearce SG, Vidrine JI, Tahay PD, Marani SK, Chen S, et al. Efficacy of mobile phone-delivered smoking cessation interventions for socioeconomically disadvantaged individuals: a randomized clinical trial. JAMA Intern Med 2019 Feb 01;179(2):167-174 [FREE Full text] [doi: 10.1001/jamainternmed.2018.5713] [Medline: 30556832]
34. Orr MF, Burduli E, Hirchak KA, Walsh Dotson JA, Young SL, Nelson L, et al. Culturally-tailored text-messaging intervention for smoking cessation in rural American Indian communities: rationale, design, and methods. Contemp Clin Trials Commun 2019 Sep;15:100363 [FREE Full text] [doi: 10.1016/j.conctc.2019.100363] [Medline: 31049463]
35. Pollak KI, Lyna P, Gao X, Noonan D, Bejarano Hernandez S, Subudhi S, et al. Efficacy of a texting program to promote cessation among pregnant smokers: a randomized control trial. Nicotine Tob Res 2020 Jun 12;22(7):1187-1194 [FREE Full text] [doi: 10.1093/ntr/ntz174] [Medline: 31647564]
36. Shuter J, Kim RS, An LC, Abroms LC. Feasibility of a smartphone-based tobacco treatment for HIV-infected smokers. Nicotine Tob Res 2020 Mar 16;22(3):398-407 [FREE Full text] [doi: 10.1093/ntr/nty208] [Medline: 30285151]
37. Tagai EK, Miller SM, Belfiglio A, Xu J, Wen KY, Hernandez E. Persistent barriers to smoking cessation among urban, underserved women: a feasibility study of tailored barriers text messages. Matern Child Health J 2020 Oct;24(10):1308-1317 [FREE Full text] [doi: 10.1007/s10995-020-02963-x] [Medline: 32557133]
38. Ruscio AC, Muench C, Brede E, Waters AJ. Effect of brief mindfulness practice on self-reported affect, craving, and smoking: a pilot randomized controlled trial using ecological momentary assessment. Nicotine Tob Res 2016 Jan;18(1):64-73. [doi: 10.1093/ntr/ntv074] [Medline: 25863520]
39. Ondersma SJ, Svikis DS, Lam PK, Connors-Burge VS, Ledgerwood DM, Hopper JA. A randomized trial of computer-delivered brief intervention and low-intensity contingency management for smoking during pregnancy. Nicotine Tob Res 2012 Mar;14(3):351-360 [FREE Full text] [doi: 10.1093/ntr/ntr221] [Medline: 22157229]
40. Bernstein SL, Rosner J, Toll B. A multicomponent intervention including texting to promote tobacco abstinence in emergency department smokers: a pilot study. Acad Emerg Med 2016 Jul;23(7):803-808 [FREE Full text] [doi: 10.1111/acem.12990] [Medline: 27146116]
41. Dignan MB, Jones K, Burhansstipanov L, Ahamed SI, Krebs LU, Williams D, et al. A randomized trial to reduce smoking among American Indians in South Dakota: the walking forward study. Contemp Clin Trials 2019 Jun;81:28-33 [FREE Full text] [doi: 10.1016/j.cct.2019.04.007] [Medline: 30986536]
42. Bordnick PS, Traylor AC, Carter BL, Graap KM. A feasibility study of virtual reality-based coping skills training for nicotine dependence. Res Soc Work Pract 2012 May;22(3):293-300 [FREE Full text] [doi: 10.1177/1049731511426880] [Medline: 25484549]
43. Wilson SM, Thompson AC, Currence ED, Thomas SP, Dedert EA, Kirby AC, et al. Patient-informed treatment development of behavioral smoking cessation for people with schizophrenia. Behav Ther 2019 Mar;50(2):395-409 [FREE Full text] [doi: 10.1016/j.beth.2018.07.004] [Medline: 30824254]
44. Krishnan N, Elf JL, Chon S, Golub JE. COach2Quit: a pilot randomized controlled trial of a personal carbon monoxide monitor for smoking cessation. Nicotine Tob Res 2019 Oct 26;21(11):1573-1577 [FREE Full text] [doi: 10.1093/ntr/nty182] [Medline: 30169740]
45. Mason M, Mennis J, Way T, Lanza S, Russell M, Zaharakis N. Time-varying effects of a text-based smoking cessation intervention for urban adolescents. Drug Alcohol Depend 2015 Dec 01;157:99-105 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.10.016] [Medline: 26507175]
46. Mason MJ, Mennis J, Zaharakis NM, Way T. The dynamic role of urban neighborhood effects in a text-messaging adolescent smoking intervention. Nicotine Tob Res 2016 May;18(5):1039-1045 [FREE Full text] [doi: 10.1093/ntr/ntv254] [Medline: 26547062]
47. Mason MJ, Coatsworth JD. Adolescents' response to a text message-delivered tobacco use intervention by depressive symptoms and sex. J Child Adolesc Subst Abuse 2020 Jul 17;29(1):1-9. [doi: 10.1080/1067828X.2020.1792021]
48. Grau LE, Pham T, O'Leary T, Weiss J, Toll B, Bernstein SL. Smokers' perspectives on texting for tobacco dependence treatment: a qualitative analysis. Nicotine Tob Res 2017 Mar 01;19(3):307-313. [doi: 10.1093/ntr/ntw184] [Medline: 27613935]
49. Wen KY, Miller SM, Kilby L, Fleisher L, Belton TD, Roy G, et al. Preventing postpartum smoking relapse among inner city women: development of a theory-based and evidence-guided text messaging intervention. JMIR Res Protoc 2014 Apr 03;3(2):e20 [FREE Full text] [doi: 10.2196/resprot.3059] [Medline: 24698804]
50. Woodruff SI, Conway TL, Edwards CC, Elliott SP, Crittenden J. Evaluation of an internet virtual world chat room for adolescent smoking cessation. Addict Behav 2007 Sep;32(9):1769-1786. [doi: 10.1016/j.addbeh.2006.12.008] [Medline: 17250972]
51. Webb Hooper M, Baker EA, Robinson RG. Efficacy of a DVD-based smoking cessation intervention for African Americans. Nicotine Tob Res 2014 Oct;16(10):1327-1335 [FREE Full text] [doi: 10.1093/ntr/ntu079] [Medline: 24838844]
52. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
53. Pagano A, Tajima B, Guydish J. Barriers and facilitators to tobacco cessation in a nationwide sample of addiction treatment programs. J Subst Abuse Treat 2016 Aug;67:22-29 [FREE Full text] [doi: 10.1016/j.jsat.2016.04.004] [Medline: 27296658]

XSL•FO
RenderX

54. Guydish J, Ziedonis D, Tajima B, Seward G, Passalacqua E, Chan M, et al. Addressing tobacco through organizational change (ATTOC) in residential addiction treatment settings. Drug Alcohol Depend 2012 Feb 01;121(1-2):30-37 [FREE Full text] [doi: 10.1016/j.drugalcdep.2011.08.003] [Medline: 21906892]
55. Das S, Tonelli M, Ziedonis D. Update on smoking cessation: e-cigarettes, emerging tobacco products trends, and new technology-based interventions. Curr Psychiatry Rep 2016 May;18(5):51. [doi: 10.1007/s11920-016-0681-6] [Medline: 27040275]
56. Langer SL, Castro FG, Chen AC, Davis KC, Joseph RP, Kim WS, et al. Recruitment and retention of underrepresented and vulnerable populations to research. Public Health Nurs 2021 Nov;38(6):1102-1115 [FREE Full text] [doi: 10.1111/phn.12943] [Medline: 34240459]
57. Kulak JA, Cornelius ME, Fong GT, Giovino GA. Differences in quit attempts and cigarette smoking abstinence between Whites and African Americans in the United States: literature review and results from the international tobacco control US survey. Nicotine Tob Res 2016 Apr;18 Suppl 1(Suppl 1):S79-S87 [FREE Full text] [doi: 10.1093/ntr/ntv228] [Medline: 26980868]
58. Epperson AE, Wallander JL, Elliott MN, Schuster MA. Tobacco use among Latinx adolescents: exploring the immigrant paradox. BMC Pediatr 2018 Nov 30;18(1):379 [FREE Full text] [doi: 10.1186/s12887-018-1355-9] [Medline: 30501626]
59. Lempert LK, Glantz SA. Tobacco industry promotional strategies targeting American Indians/Alaska natives and exploiting tribal sovereignty. Nicotine Tob Res 2019 Jun 21;21(7):940-948 [FREE Full text] [doi: 10.1093/ntr/nty048] [Medline: 29546392]
60. Mickens L, Ameringer K, Brightman M, Leventhal AM. Epidemiology, determinants, and consequences of cigarette smoking in African American women: an integrative review. Addict Behav 2010 May;35(5):383-391 [FREE Full text] [doi: 10.1016/j.addbeh.2009.12.014] [Medline: 20061090]
61. Marsch LA, Carroll KM, Kiluk BD. Technology-based interventions for the treatment and recovery management of substance use disorders: a JSAT special issue. J Subst Abuse Treat 2014 Jan;46(1):1-4 [FREE Full text] [doi: 10.1016/j.jsat.2013.08.010] [Medline: 24041749]
62. Nesvåg S, McKay JR. Feasibility and effects of digital interventions to support people in recovery from substance use disorders: systematic review. J Med Internet Res 2018 Aug 23;20(8):e255 [FREE Full text] [doi: 10.2196/jmir.9873] [Medline: 30139724]
63. Barroso-Hurtado M, Suárez-Castro D, Martínez-Vispo C, Becoña E, López-Durán A. Smoking cessation apps: a systematic review of format, outcomes, and features. Int J Environ Res Public Health 2021 Nov 06;18(21):11664 [FREE Full text] [doi: 10.3390/ijerph182111664] [Medline: 34770178]
64. Webb Hooper M, Carpenter KM, Salmon EE. Web-based tobacco cessation interventions and digital inequality across US racial/ethnic groups. Ethn Dis 2019;29(3):495-504 [FREE Full text] [doi: 10.18865/ed.29.3.495] [Medline: 31367170]
65. Ziedonis D, Das S, Larkin C. Tobacco use disorder and treatment: new challenges and opportunities. Dialogues Clin Neurosci 2017 Sep;19(3):271-280 [FREE Full text] [doi: 10.31887/DCNS.2017.19.3/dziedonis] [Medline: 29302224]
66. Burlew K, McCuistian C, Szapocznik J. Racial/ethnic equity in substance use treatment research: the way forward. Addict Sci Clin Pract 2021 Aug 05;16(1):50 [FREE Full text] [doi: 10.1186/s13722-021-00256-4] [Medline: 34353373]
67. Burlew K, Larios S, Suarez-Morales L, Holmes B, Venner K, Chavez R. Increasing ethnic minority participation in substance abuse clinical trials: lessons learned in the National Institute on Drug Abuse's Clinical Trials Network. Cultur Divers Ethnic Minor Psychol 2011 Oct;17(4):345-356 [FREE Full text] [doi: 10.1037/a0025668] [Medline: 21988575]
68. Casetta B, Videla AJ, Bardach A, Morello P, Soto N, Lee K, et al. Association between cigarette smoking prevalence and income level: a systematic review and meta-analysis. Nicotine Tob Res 2017 Nov 07;19(12):1401-1407. [doi: 10.1093/ntr/ntw266] [Medline: 27679607]
69. Cavelaars AE, Kunst AE, Geurts JJ, Crialesi R, Grötvedt L, Helmert U, et al. Educational differences in smoking: international comparison. BMJ 2000 Apr 22;320(7242):1102-1107 [FREE Full text] [doi: 10.1136/bmj.320.7242.1102] [Medline: 10775217]
70. Maralani V. Understanding the links between education and smoking. Soc Sci Res 2014 Nov;48:20-34. [doi: 10.1016/j.ssresearch.2014.05.007] [Medline: 25131272]
71. U.S. Department of Health and Human Services. The health consequences of smoking—50 years of progress: a report of the surgeon general. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. 2014. URL: https://www.hhs.gov/sites/default/files/consequences-smoking-exec-summary.pdf [accessed 2024-09-20]
72. Mobile fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/?menuItem=d40cde3f-c455-4f0e-9be0-0aefcdaeee00 [accessed 2024-09-20]
73. Vogels EA. Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. 2021 Jun 22. URL: https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/ [accessed 2022-03-29]
74. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 01;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]

XSL•FO
RenderX

75. Kozelka EE, Acquilano SC, Al-Abdulmunem M, Guarino S, Elwyn G, Drake RE, et al. Documenting the digital divide: identifying barriers to digital mental health access among people with serious mental illness in community settings. SSM Ment Health 2023 Dec;4:100241. [doi: 10.1016/j.ssmmh.2023.100241]
76. Bustillo NE, McGinty HL, Dahn JR, Yanez B, Antoni MH, Kava BR, et al. Fatalism, medical mistrust, and pretreatment health-related quality of life in ethnically diverse prostate cancer patients. Psychooncology 2017 Mar;26(3):323-329 [FREE Full text] [doi: 10.1002/pon.4030] [Medline: 26553139]
77. Bracken P, Fernando S, Alsaraf S, Creed M, Double D, Gilberthorpe T, et al. Decolonising the medical curriculum: psychiatry faces particular challenges. Anthropol Med 2021 Dec;28(4):420-428. [doi: 10.1080/13648470.2021.1949892] [Medline: 34282672]
78. Johnson LM, Green HDJ, Koch B, Harding R, Stockman JK, Wagner KD. Correlates of medical mistrust among minority women at risk for HIV and their networks. Health Educ Behav 2021 Dec;48(6):860-872 [FREE Full text] [doi: 10.1177/1090198120986783] [Medline: 33622086]
79. López-Cevallos DF, Flórez KR, Derose KP. Examining the association between religiosity and medical mistrust among churchgoing Latinos in Long Beach, CA. Transl Behav Med 2021 Feb 11;11(1):114-121 [FREE Full text] [doi: 10.1093/tbm/ibz151] [Medline: 31628472]
80. Rusoja EA, Thomas BA. The COVID-19 pandemic, black mistrust, and a path forward. EClinicalMedicine 2021 May;35:100868 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100868] [Medline: 33977258]
81. Jaiswal J. Whose responsibility is it to dismantle medical mistrust? Future directions for researchers and health care providers. Behav Med 2019;45(2):188-196 [FREE Full text] [doi: 10.1080/08964289.2019.1630357] [Medline: 31343959]
82. Subbaraman MS, Laudet AB, Ritter LA, Stunz A, Kaskutas LA. Multi-source recruitment strategies for advancing addiction recovery research beyond treated samples. J Community Psychol 2015 Jun 01;43(5):560-575 [FREE Full text] [doi: 10.1002/jcop.21702] [Medline: 26166909]
83. Jordan A, Babuscio T, Nich C, Carroll KM. A feasibility study providing substance use treatment in the Black church. J Subst Abuse Treat 2021 May;124:108218. [doi: 10.1016/j.jsat.2020.108218] [Medline: 33771290]
84. Rainie L. Internet, broadband, and cell phone statistics. Pew Research Center. 2010 Jan 05. URL: https://www.pewresearch.org/internet/2010/01/05/internet-broadband-and-cell-phone-statistics/ [accessed 2024-09-20]
85. Internet, broadband fact sheet. Pew Research Center. 2024 Jan 31. URL: http://www.pewinternet.org/fact-sheet/internet-broadband/ [accessed 2022-03-30]
86. Gonzalez C, Early J, Gordon-Dseagu V, Mata T, Nieto C. Promoting culturally tailored mHealth: a scoping review of mobile health interventions in Latinx communities. J Immigr Minor Health 2021 Oct;23(5):1065-1077 [FREE Full text] [doi: 10.1007/s10903-021-01209-4] [Medline: 33988789]
87. Fiore MC, Baker TB. Clinical practice. Treating smokers in the health care setting. N Engl J Med 2011 Sep 29;365(13):1222-1231 [FREE Full text] [doi: 10.1056/NEJMcp1101512] [Medline: 21991895]
88. Clinical Practice Guideline Treating Tobacco UseDependence 2008 Update Panel, Liaisons,Staff. A clinical practice guideline for treating tobacco use and dependence: 2008 update. A U.S. public health service report. Am J Prev Med 2008 Aug;35(2):158-176 [FREE Full text] [doi: 10.1016/j.amepre.2008.04.009] [Medline: 18617085]
89. Patnode CD, Henderson JT, Coppola EL, Melnikow J, Durbin S, Thomas RG. Interventions for tobacco cessation in adults, including pregnant persons: updated evidence report and systematic review for the US preventive services task force. JAMA 2021 Jan 19;325(3):280-298. [doi: 10.1001/jama.2020.23541] [Medline: 33464342]
90. Stead LF, Koilpillai P, Fanshawe TR, Lancaster T. Combined pharmacotherapy and behavioural interventions for smoking cessation. Cochrane Database Syst Rev 2016 Mar 24;3(3):CD008286 [FREE Full text] [doi: 10.1002/14651858.CD008286.pub3] [Medline: 27009521]
91. Cox LS, Okuyemi K, Choi WS, Ahluwalia JS. A review of tobacco use treatments in U.S. ethnic minority populations. Am J Health Promot 2011;25(5 Suppl):S11-S30 [FREE Full text] [doi: 10.4278/ajhp.100610-LIT-177] [Medline: 21510783]
92. Robles GI, Singh-Franco D, Ghin HL. A review of the efficacy of smoking-cessation pharmacotherapies in nonwhite populations. Clin Ther 2008 May;30(5):800-812. [doi: 10.1016/j.clinthera.2008.05.010] [Medline: 18555928]
93. Hendricks PS, Westmaas JL, Ta Park VM, Thorne CB, Wood SB, Baker MR, et al. Smoking abstinence-related expectancies among American Indians, African Americans, and women: potential mechanisms of tobacco-related disparities. Psychol Addict Behav 2014 Mar;28(1):193-205 [FREE Full text] [doi: 10.1037/a0031938] [Medline: 23528192]
94. Levinson AH, Borrayo EA, Espinoza P, Flores ET, Pérez-Stable EJ. An exploration of Latino smokers and the use of pharmaceutical aids. Am J Prev Med 2006 Aug;31(2):167-171. [doi: 10.1016/j.amepre.2006.03.022] [Medline: 16829334]
95. Cook JA, Mueser KT. Economic security: an essential component of recovery. Psychiatr Rehabil J 2013 Mar;36(1):1-3. [doi: 10.1037/h0094739] [Medline: 23477642]
96. Burgess D, Fu SS, Joseph AM, Hatsukami DK, Solomon J, van Ryn M. Beliefs and experiences regarding smoking cessation among American Indians. Nicotine Tob Res 2007 Jan;9 Suppl 1:S19-S28. [doi: 10.1080/14622200601083426] [Medline: 17365723]
97. Fu SS, Burgess D, van Ryn M, Hatsukami DK, Solomon J, Joseph AM. Views on smoking cessation methods in ethnic minority communities: a qualitative investigation. Prev Med 2007 Mar;44(3):235-240. [doi: 10.1016/j.ypmed.2006.11.002] [Medline: 17175016]

XSL•FO
RenderX

98. Benjamin R. Race After Technology Abolitionist Tools for the New Jim Code. Hoboken, NJ: Wiley; 2019.
99. Vespa J, Armstrong DM, Medina L. Demographic turning points for the United States: population projections for 2020 to 2060. United States Census Bureau. 2020 Feb. URL: https://www.census.gov/library/publications/2020/demo/p25-1144.html [accessed 2024-09-20]
100. Frey WH. The nation is diversifying even faster than predicted, according to new census data. The Brookings Institution. 2020 Jul 01. URL: https://www.brookings.edu/articles/new-census-data-shows-the-nation-is-diversifying-even-faster-than-predicted/ [accessed 2024-09-20]
101. Burlew AK, Peteet BJ, McCuistian C, Miller-Roenigk BD. Best practices for researching diverse groups. Am J Orthopsychiatry 2019;89(3):354-368. [doi: 10.1037/ort0000350] [Medline: 31070421]
102. Burlew AK, Feaster D, Brecht ML, Hubbard R. Measurement and data analysis in research addressing health disparities in substance abuse. J Subst Abuse Treat 2009 Jan;36(1):25-43 [FREE Full text] [doi: 10.1016/j.jsat.2008.04.003] [Medline: 18550320]
103. Hall GC, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016 Nov;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
104. Smith TB, Rodríguez MD, Bernal G. Culture. J Clin Psychol 2011 Feb;67(2):166-175. [doi: 10.1002/jclp.20757] [Medline: 21105069]
105. Odani S, Armour BS, Graffunder CM, Garrett BE, Agaku IT. Prevalence and disparities in tobacco product use among American Indians/Alaska Natives - United states, 2010-2015. MMWR Morb Mortal Wkly Rep 2017 Dec 22;66(50):1374-1378 [FREE Full text] [doi: 10.15585/mmwr.mm6650a2] [Medline: 29267265]
106. Burlew AK, McCuistian C, Lanaway D, Hatch-Maillette M, Shambley-Ebron D. One size does not fit all: a NIDA CTN inspired model for community engaged cultural adaptation. J Subst Abuse Treat 2020 Mar;112S:28-33 [FREE Full text] [doi: 10.1016/j.jsat.2020.02.006] [Medline: 32220407]
107. Webb MS, Francis J, Hines BC, Quarles FB. Health disparities and culturally specific treatment: perspectives and expectancies of African American smokers. J Clin Psychol 2007 Dec;63(12):1247-1263. [doi: 10.1002/jclp.20437] [Medline: 17972299]
108. Brown CE, Curtis JR, Doll KM. A race-conscious approach toward research on racial inequities in palliative care. J Pain Symptom Manage 2022 May;63(5):e465-e471 [FREE Full text] [doi: 10.1016/j.jpainsymman.2021.11.012] [Medline: 34856335]
109. -. Why tobacco is a racial justice issue. Truth Initiative. 2020 Aug 3. URL: https://truthinitiative.org/research-resources/targeted-communities/why-tobacco-racial-justice-issue [accessed 2023-04-26]

Abbreviations

BVSC: bioverified smoking cessation
NRT: nicotine replacement therapy
PRISMA-ScR: Preferred Reporting Items For Systematic Reviews And Meta-Analyses Extension for Scoping Reviews
RCT: randomized controlled trial
SES: socioeconomic status
SRSC: self-reported smoking cessation
SUD: substance use disorder
TBI: technology-based intervention
TM: text messaging
VR: virtual reality

Edited by T de Azevedo Cardoso; submitted 17.07.23; peer-reviewed by A Thompson, E Kozelka; comments to author 22.01.24; revised version received 17.03.24; accepted 12.08.24; published 10.10.24.

Please cite as:
Hichborn E, Turner A, Moore S, Gauthier P, Bell K, Montgomery L, Boggis J, Lambert-Harris C, Saunders E, Dallery J, McLeman B, Marsch L
Technology-Based Interventions in Tobacco Use Treatment Among People Who Identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native: Scoping Review
J Med Internet Res 2024;26:e50748
URL: https://www.jmir.org/2024/1/e50748
doi:10.2196/50748
PMID:

XSL•FO
RenderX

©Emily Hichborn, Avery Turner, Sarah Moore, Phoebe Gauthier, Kathleen Bell, LaTrice Montgomery, Jesse Boggis, Chantal Lambert-Harris, Elizabeth Saunders, Jesse Dallery, Bethany McLeman, Lisa Marsch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Quality Assessment of Digital Health Apps: Umbrella Review

Maciej Marek Zych[1], BSc; Raymond Bond[1], BSc, PhD; Maurice Mulvenna[1], BSc, MPhil, PhD; Jorge Martinez Carracedo[1], PhD, BSc, MSc; Lu Bai[2], BEng, PhD; Simon Leigh[3], BSc, MSc, PhD

[1]School of Computing, Ulster University, Belfast, United Kingdom

[2]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom

[3]ORCHA, Daresbury, United Kingdom

Corresponding Author:
Maciej Marek Zych, BSc
School of Computing
Ulster University
2-24 York St
Belfast, BT15 1AP
United Kingdom
Phone: 44 7526852505
Email: maciejmarekzych@gmail.com

Abstract

Background: With an increasing number of digital health apps available in app stores, it is important to assess these technologies reliably regarding their quality. This is done to mitigate the risks associated with their use. There are many different guidelines, methods, and metrics available to assess digital health apps with regard to their quality.

Objective: This study aimed to give a holistic summary of the current methods and "condition agnostic" frameworks that are broadly applicable for the quality assessment of all digital health apps.

Methods: A systematic search of literature was conducted on 4 databases: Scopus, PubMed, ACM Digital Library, and IEEE Xplore. We followed the PICOS (Population, Patient, or Problem; Intervention; Comparison; Outcomes; and Study Design) and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodologies when conducting this umbrella review. The search was conducted on January 26, 2024, for review articles published between 2018 and 2023. We identified 4781 candidate papers for inclusion; after title and abstract screening, 39 remained. After full-text analysis, we included 15 review articles in the full review.

Results: Of the 15 review articles, scoping reviews were the most common (n=6, 40%), followed by systematic reviews (n=4, 27%), narrative reviews (n=4, 27%), and a rapid review (n=1, 7%). A total of 4 (27%) review articles proposed assessment criteria for digital health apps. "Data privacy and/or security" was the most mentioned criterion (n=13, 87%) and "Cost" was the least mentioned criterion (n=1, 7%) for the assessment of digital health apps. The Mobile App Rating Scale was the most frequently used framework for quality assessment of digital health apps.

Conclusions: There is a lack of unity or consolidation across identified frameworks, as most do not meet all the identified criteria from the reviewed articles. Safety concerns associated with the use of digital health apps may be mitigated with the use of quality frameworks.

(J Med Internet Res 2024;26:e58616) doi:10.2196/58616

KEYWORDS

mHealth assessment; digital health; quality assessment; health apps quality; assessment criteria; evaluation criteria; health apps criteria; assessment; digital health app; app; umbrella review; risk; mobile phone; frameworks

Introduction

Digital health apps have a potential to make health care more accessible to people of different age groups living with a wide range of health-related conditions. Currently, there are ongoing efforts by the National Health Service (NHS) in the United Kingdom to enhance the use of digital health technologies. These efforts were outlined in the NHS long-term plan [1]. Moreover, these efforts were accelerated by the COVID-19 pandemic [2]. There are approximately 250 new digital health apps available in app stores per day, with a total of approximately 350,000 digital health apps on the market as of

XSL•FO
RenderX

2021 [3]. As digital health apps rise in popularity, so do the risks associated with their use. Digital health apps that are classified as "medical devices" [4] in the United Kingdom are strictly regulated, but digital health apps that are classified as "wellness apps" do not face such regulations.

There are organizations that produce guidelines and frameworks regarding the development and assessment of digital health apps. These include the International Organization for Standardization (ISO): Health and Wellness Apps—Quality and Reliability [5] and the National Institute for Health and Care Excellence (NICE): Evidence Standards Framework [6]. Digital Technology Assessment Criteria (DTAC) [7] is an NHS-developed framework that is being used to provide criteria for the assessment of digital health apps. However, some of the guidelines may be open to interpretation, and there are frameworks that are currently being used to assess digital health apps for their quality.

A scoping review published in 2023 [8] examined the problems and barriers related to the use of digital health apps. It found that "validity," "usability," "technology," "data privacy and security," and "individuality" were addressed in several studies and are partly considered in quality assessment. "Use and adherence," "patient-physician relationship," "knowledge and skills," "implementation," and "costs" of digital health apps were rarely extensively studied. Furthermore, a systematic review published in 2021 [9] found security challenges when developing digital health apps. These and other quality problems surrounding digital health apps may be mitigated by rigorously assessing their quality.

Digital health apps are complex, and the usual methods for assessing medicines (survival, quality of life, and cost), as recommended by NICE, may not be sufficient to monitor the broad spectrum of potential issues arising from their use. Therefore, more specific requirements for assessment of quality are required. However, there is a lack of consensus on how best to achieve this.

In this umbrella review, we define quality as "compliance with best practice standards." The objective of this umbrella review was to give a holistic summary of the current methods and "condition agnostic" frameworks that are broadly applicable for the quality assessment of all digital health apps. Because several review articles have been published regarding the quality assessment of digital health apps, or aspects related to this, we conducted an umbrella review to provide a holistic view of how digital health apps are currently being assessed and where they can improve. This review can be informative to digital health researchers and assessment framework developers. In this umbrella review, we included systematic reviews, scoping reviews, rapid reviews, and narrative reviews.

Methods

Search Strategy and Criteria

For the systematic search of literature, the PICOS (Population, Patient, or Problem; Intervention; Comparison; Outcomes; and Study Design; see Textbox 1) and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodologies have been used. The following search has been conducted using the Scopus, PubMed, ACM Digital Library, and IEEE Xplore databases with the following search objective: find literature, systematic, or scoping review articles for quality assessment of digital health apps. The search for review articles was conducted on January 26, 2024.

Textbox 1. PICOS (Population, Patient, or Problem; Intervention; Comparison; Outcomes; and Study Design) methodology for systematic search of previous or related literature reviews.

Inclusion criteria

- Problem: Review of quality assessment tools for digital health apps, full article, in English, from 2018 to 2023.
- Intervention: Quality assessment frameworks or criteria review applicable to all digital health apps.
- Comparator: Not applicable (N/A)
- Outcome: Review articles regarding quality (or aspects of quality, eg, usability) assessment frameworks or criteria applicable to all digital health apps.
- Study design: Search databases included Scopus, PubMed, ACM Digital Library, and IEEE Xplore. Search query for article title and abstract: ((mhealth OR ehealth OR m-health OR e-health OR "mobile health" OR "electronic health" OR "health app*" OR "medical health app*" OR "digital health app*" OR "digital health product*" OR "digital health intervention*" OR "digital health technolog*" OR "digital health solution*") AND (assurance* OR assessment* OR evaluation* OR audit* OR framework*) AND (review* OR assessment*))

Exclusion criteria

- Problem: Not quality assessment of digital health apps, conference paper, book or book chapter. Not in English.
- Intervention: Not a quality assessment frameworks or criteria review.
- Comparator: N/A
- Outcome: No information on quality assessment frameworks or criteria. Does not focus on digital health apps. Articles targeting specific users (eg, women or adolescence). Focuses on specific feature or category (condition area). Frameworks that focus on user acceptance of technology.
- Study design: N/A

XSL•FO
RenderX

Multimedia Appendix 1 presents the PRIOR checklist and Multimedia Appendix 2 presents the exact search queries that were used for each of the databases.

Study Screen

This study used PICOS methodology to set the inclusion and exclusion criteria (Textbox 1). The review articles included in this study were initially screened by title and abstract. If the review article focused on quality (or its aspect) assessment of digital health apps, it was read in full. If the article met inclusion criteria and none of the exclusion criteria (Textbox 1), it was included in the study. This systematic search followed PRISMA guidelines when screening the articles, where a step-by-step process was set forth. The Rayyan tool has been used to remove article duplicates.

Critical Appraisal

Because 4 (27%) of the 15 review articles were systematic reviews, for those reviews, the Joanna Briggs Institute (JBI) critical appraisal of systematic reviews [10] has been used (see Multimedia Appendix 3). This has been done to assess the quality of the systematic reviews to determine whether they should be included in this umbrella review.

Data Extraction

The characteristics of this study follow the applicable protocols for the umbrella review from the *JBI Manual for Evidence Synthesis* [11]. The following information has been extracted: author-year, objectives, total sample size, number of sources searched, date (year) range of searched and included studies, number of studies included, methods of analysis, and key findings (see Multimedia Appendix 4).

Data Synthesis

In this study, we did not synthesize results statistically. Instead, we narratively synthesized key findings of each of the reviews. This was done due to the presumed heterogeneity of the included reviews.

Results

Overview

Figure 1 shows the PRISMA process of selecting article reviews to be included in the study. The search queries used for each of the databases searched (Scopus, PubMed, ACM Digital Library, and IEEE Xplore) are available in Multimedia Appendix 2. After duplicates were removed, the articles were reviewed by title and abstract, and were included only if they were related to quality assessment of digital health apps and did not meet the exclusion criteria outlined in Textbox 1. Afterward, 39 review articles were read in full and 15 met the inclusion criteria and were included in this umbrella review. This review was not registered and a protocol for systematic reviews was not prepared.

Mendeley (Elsevier) was used to manage all the included review articles (n=15) in the study. The 15 review articles were published between 2018 and 2023. Scoping reviews were the most common (n=6, 40%), followed by systematic reviews (n=4, 27%), narrative reviews (n=4, 27%), and a rapid review (n=1, 7%). Multimedia Appendix 5 depicts which criteria the review articles focused on.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of reviews selected for inclusion in the umbrella review.

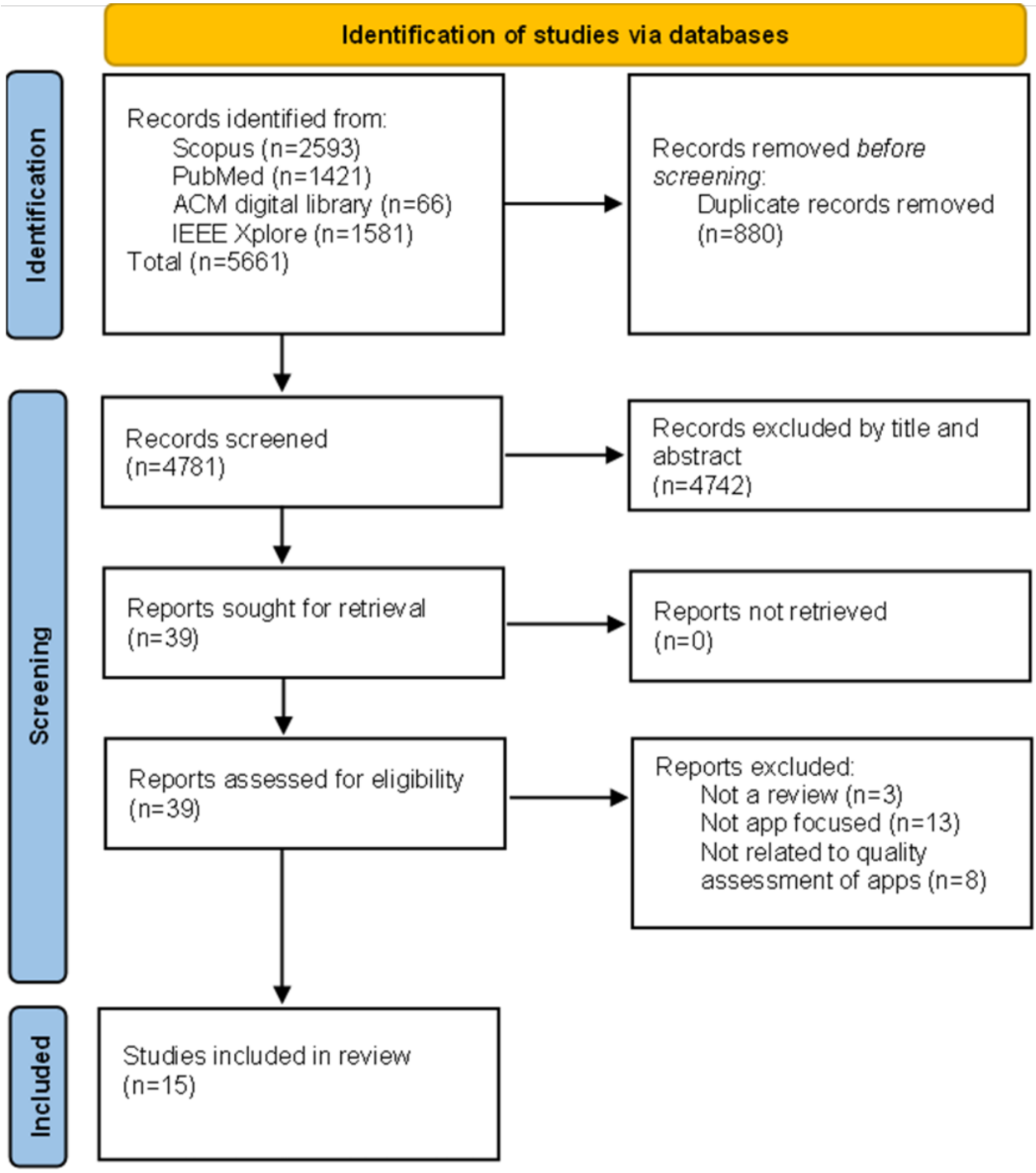

The Assessment Criteria

The objective of this umbrella review was to give a holistic summary of current methods and frameworks used for quality assessment of digital health apps. Table 1 depicts 4 (27%) of the 15 review articles that reported on assessment criteria for digital health apps. Of the 4 review articles, 2 (50%) outlined what was being assessed with frameworks [12,13] and 2 (50%) focused on what should be assessed [14,15]. The remaining 11 review articles included in this umbrella review either focused on a specific criterion (eg, usability), the methods of assessment (eg, Likert scale), or referenced the criteria of a review article already accounted for (Woulfe et al [16] references the Nouri et al [13] criteria). Nouri et al [13] stated that different articles defined assessment criteria differently. For example, usability was mentioned in 14 articles that were reviewed. One article considered usability a subclass of ease of use; other articles placed ease of use under usability.

XSL•FO
RenderX

Table 1. Framework criteria used for the assessment of digital health apps.

Authors, Year	Obtained by[a]	Criteria (assessment domains)
Hensher et al, 2021 [12]	Assessing frameworks	• Clarity of purpose of the app • Developer credibility • Content or information validity • User experience • User engagement or adherence and social support interoperability • Value • Technical features and support • Privacy, security, ethical, or legal • Accessibility
Nouri et al, 2018 [13]	Assessing frameworks	• Design • Information or content • Usability • Functionality • Ethical issues • Privacy and security • User-perceived value
Moshi et al, 2018 [14]	Preset or developed	• Six core criteria: • Current use of the technology • Description and technical characteristics • Effectiveness • Safety • Cost and cost-effectiveness • Organizational aspects • Three optional criteria: • Legal aspects • ethical aspects • social aspects
Lagan et al, 2021 [15]	Preset or developed	• Mobile health Index and Navigation Database (MIND) criteria: • App origin • App functionality • Inputs and outputs • Privacy and security • Evidence and clinical foundation • Features and engagement style • App use • Interoperability and data sharing

[a]"Obtained by" states whether assessment criteria have been identified by assessing assessment frameworks or preset or developed by other means.

Hensher et al [12] and Nouri et al [13] identified criteria by assessing assessment frameworks. Moshi et al [14] and Lagan et al [15] had their own set of criteria and compared assessment frameworks against those criteria. Moshi et al [14] developed a checklist for health technology assessment for mobile medical apps based on previous research [17,18]. Lagan et al [15] used the previously proposed Mobile health Index and Navigation Database (MIND) [19], which was developed with inputs from clinicians, patients, family members, researchers, and policy makers with the aim of providing clinically relevant criteria. Nouri et al [13] stated that there may never be complete digital health apps assessment criteria because these criteria apply to digital health apps that are changing in development continuously.

Data Privacy or Security

Thirteen (87%) of the 15 review articles [12-16,20-27] mention the data privacy or security of digital health apps. Three reviews [13,20,21] found that reviewed articles often considered privacy and security together when evaluating digital health apps. Nurgalieva et al [21] highlighted that although privacy and security overlap, security relates to protection against unauthorized access to data, and privacy is an individual's right to maintain control of their personal data. The review points out that exclusively focusing on security can lead to increased surveillance, which can introduce privacy risks. Furthermore, Nouri et al [13] stated that interpretation of privacy, security, and safety considerations differed when examining different quality assessment tools or methods.

Nurgalieva et al [21] found that methods used to evaluate security are more technical, whereas methods used to evaluate privacy are more user oriented. Some of the review articles [15,21] pointed out that there appears to be greater effort to assess privacy. However, Muro-Culebras et al [22] found lack of assessment regarding developer transparency and policies

XSL•FO
RenderX

regarding user data privacy and security. Lagan et al [15] found that most of the examined assessment frameworks (43/79, 54%) included privacy-related questions. Hensher et al [12] also found privacy and security to be frequently addressed in assessment frameworks.

Grundy [23] discussed how guidelines such as General Data Protection Regulation (GDPR) [28] rely on the users' knowledge and "notice and consent" model. Grundy [23] referenced a study published in 2020 [29] that discusses whether GDPR is fit for purpose, as it assumes that users can know why and how their data are being collected and shared. GDPR also assumes that individual app users can control how their personal data are being processed. Grundy [23] stated that the majority of digital health apps fail to provide assurances around privacy and security. On the other hand, Carmi et al [25] focused on the interpretation of GDPR for mobile health.

Galvin and DeMuro [26] discussed how recent literature has shown that aggregated data that were previously considered deidentified have been shown to be reidentifiable. The review states that data storage and transmission of mobile health data remain a security concern. Moreover, the review points out that because of the lack of privacy policy, or because of the complex language used in such policies, there may be lack of consumer informed consent when using mobile health apps.

Clinical Assurance, Credibility of Information, or Evidence

Eleven (73%) of the 15 review articles [12-16,22-24,26,27,30] included in this umbrella review mentioned clinical assurance, credibility of information of digital health apps, and evidence. Hensher et al [12] stated that in 2 studies included in their review, third-party sponsorships were deemed important due to a possible conflict of interest between developers of an app and sponsors, affecting developers' credibility. Moshi et al [14] found that credibility of information of mobile medical apps was assessed by 15 (33%) of 45 frameworks. Moreover, 27 (60%) of 45 frameworks included questions about mobile medical apps' source of information. None of the frameworks assessed the health impact of mobile medical apps that provide diagnostics or information. The review concluded that none of the included frameworks met all the health technology assessment criteria (see Table 1—Moshi et al [14]) set forth. A total of 3 (7%) of 45 frameworks specifically asked about randomized controlled trials. Lagan et al [15], which expands on the work done by Moshi et al [14], has shown that more frameworks include questions around clinical foundation than in their previous article published in 2019 [31], indicating an increase of interest around clinical foundation assessment.

Muro-Culebras et al [22] stated that many authors use their own personalized questionnaires for assessment, specifically designed to assess the characteristics of their specific digital health apps. The review article stated that because the approach is a personalized questionnaire, there is greater flexibility than in other generic tools. However, lack of validation and reliability (such as inter-rater reliability) raises questions about their suitability for digital health app assessment. Muro-Culebras et al [22] concluded that highly validated tools for the assessment of digital health apps are still a largely unexplored topic in the market. Similarly, Grundy [23] stated that there is a lack of measurement tool validation. Furthermore, users do not appear to have much awareness of the source or validity of health information of a digital health app. Akbar et al [24] found that digital health apps lack domain expert involvement regarding the app content and provide poor evidence base and poor validation.

User Experience, Value, Efficacy or Effectiveness, or Engagement

Eleven (73%) of 15 review articles [12-16,22-24,27,30,32] mentioned user experience, value, efficacy, or engagement. Hensher et al [12] found that aspects of user experience have been frequently assessed by the assessment frameworks. Moreover, the review stated that there is limited evidence on how to evaluate the value domain in the literature. They speculate that this could be due to the current landscape of digital health apps' market being fast and evolving, and the subjectivity of value. The review also stated that studies to demonstrate apps' efficacy and value for money are not undertaken often. Moshi et al [14] stated that of 45 frameworks included in the review, all assessed effectiveness to some extent. Of 45 frameworks, 11 assessed user satisfaction and 30 assessed technical efficacy of mobile medical apps.

Maramba et al [32] focused on methods of usability testing of eHealth applications. Questionnaires, task completion, think-aloud protocol, interviews, heuristic testing, and focus groups were the most frequently used methods of assessment, whereas methods such as eye tracking were rarely used. The review concludes that more investigation needs to be made into assessing usability of eHealth applications. Muro-Culebras et al [22] found that usability among the 8 frameworks assessed in the review was commonly assessed together with engagement, aesthetics, or functionality. Moreover, the review found that 2 (75%) of the 8 frameworks had a user assessor and a professional assessor version. Woulfe et al [16] point out that many digital health apps are not based on any behavior change theory, and in many cases, effectiveness is inadequately assessed. They also address the possibility of using different assessment methodologies in high-, low-, and middle-income countries.

Lagan et al [15] point out that subjective user experience may limit generalizability and standardization of frameworks. This is because assessment would reflect the experience of the assessor. The review further points out that although subjective in nature, information on user friendliness, visual appeal, and interface design may be of great interest to the user and a good predictor of user engagement.

Akbar et al [24] stated that users should be involved in usability testing of digital health apps. Their review indicates that consumers were able to identify many critical issues with digital health apps, such as incorrect information, inappropriate response to their needs, gaps in features, and faults with alarms. Akbar et al [24] suggest that allowing users who are consumers of digital health apps to be involved in usability testing will enable the identification and resolution of usability problems before the apps are made available to the public.

XSL•FO
RenderX

Grundy [23] pointed out that the assessment frameworks mainly focused on content quality and usability, with less attention given to design, security and privacy, functionality, user-perceived value, and ethical issues. Nouri et al [13] stated that usability was treated in different ways by different articles. For example, one article considered usability a subclass of ease of use, and other articles placed ease of use under usability. Azad-Khaneghah et al [27] found that many of the usability and quality rating scales are targeted at professionals. Moreover, the review found that System Usability Scale (SUS) was the most widely used framework or scale (12/40 studies), mainly due to its simplicity. Similarly, Hajesmaeel-Gohari et al [30] found that general questionnaires with fewer questions and higher reliability, such as SUS, have been used more often. Furthermore, the review recommends using frameworks such as the mHealth App Usability Questionnaire [33] that were specifically designed for mobile apps, unlike the SUS.

Safety

Safety (in different contexts) has been mentioned by 9 (60%) of the 15 review articles [12-14,16,20-24]. Safety and risks of digital health apps can arise from different factors. Akbar et al [24] elaborate on different safety concerns and risks associated with the use of digital health apps. The review found 67 safety concerns related to the quality of content, which are grouped into the following 5 categories: incorrect information, incomplete information, variation in content, incorrect output, and inappropriate response to consumer needs. Akbar et al [24] found 13 safety concerns related to software functionality. These are grouped into 5 more categories: gaps in features, lack of validation for user input, delayed processing, response to health dangers, and faulty alarms. Akbar et al [24] further discuss consequence of safety concerns, for example, how one of the digital health apps led to dangerous levels of alcohol consumption among a group of 341 students. Overall, many of the frameworks do not cover the necessary criteria to quality assess digital health apps [14,15]. Moshi et al [14] consider safety as part of the assessment criteria for digital health apps (see Table 1).

Features or Functionality

Six (40%) of the 15 review articles [12-16,27] considered feature or functionality as criteria to be assessed, referred to as technical characteristics by Moshi et al [14]. Furthermore, Nurgalieva et al [21] mention that feature assessment was common when assessing security and privacy. Lagan et al [15] stated that features related to ease of use and visual appeal may be the most important drive of user engagement of mental health apps. Moreover, “subjective questions” around user friendliness, visual appeal and interface design, although difficult to standardize and to assess, may be the greatest predictor of user engagement. Lagan et al [15] and Hensher et al [12] identified digital health apps’ features as part of the assessment criteria (Table 1). Moshi et al [14] and Nouri et al [13] included functionality and technical characteristics, respectively, as part of their assessment criteria for digital health apps. The terms “features,” “functionality,” and “technical characteristics” seem to overlap in the review articles.

Nouri et al [13] stated that 2 (9%) of 23 review articles provided dynamic assessment criteria based on the use cases and features of specific digital health apps, meaning that the assessment criteria were selected for apps based on the use case. They provide an example of the criterion “accuracy of the calculations,” being used only when an app provides at least 1 calculation. Akbar et al [24] stated that many digital health apps had gaps in features that inadequately supported user tasks. They gave an example of “Tele dermatology” apps that did not account for allergies or current medication status.

Cost

Cost as a criterion has been mentioned by 1 (7%) of the 15 review articles [14]. Moshi et al [14] stated that there may be a cost barrier to accessing mobile medical apps because they may contain in-app purchases or require a subscription. They further stated that only 1 (2%) of 45 frameworks assessed cost-effectiveness (in terms of economic assessment) and 11 (24%) of 45 reviewed the cost in terms of price to download or in-app purchases. Nine (60%) of the 15 review articles [12,13,16,20,21,23,24,26,32] mentioned the cost of digital health apps, the cost of data breaches, or equipment cost, but not as a criterion for quality assessment.

Ethical or Legal Issues

Ethical or legal concepts have been mentioned by 10 (67%) of the 15 review articles [12-14,16,20,21,23-26]. Nurgalieva et al [21] and Benjumea et al [20] discuss ethical and legal concepts regarding data and data breaches. Nurgalieva et al [21] discuss how in their review of privacy and security of digital health apps, there was a considerable lack of discussion in the reviewed articles (n=83) regarding privacy or security ethics. Grundy [23] stated that quality assessment frameworks give less attention to ethical issues. The author also stated that legal compliance around content and intellectual property is an aspect of quality regarding commercial apps.

Moshi et al [14] discuss health technology assessment criteria for the assessment of mobile medical apps, including ethical and legal aspects. The review found that 4 (9%) of 45 frameworks discussed legal aspects and 24 (53%) of 45 frameworks discussed ethics. Hensher et al [12] also included ethics and legal aspects as part of their assessment criteria (see Table 1). For example, in the reviews of both Moshi et al [14] and Hensher et al [12], ethical aspects included privacy policies and legal aspects with mention of disclaimers. Nouri et al [13] included ethical issues as part of the assessment criteria (see Table 1). Akbar et al [24] stated that health care professionals may hesitate to promote digital health apps in part due to legal issues.

Assessment Methods and Metrics

Hensher et al [12] from Deakin University conducted a scoping review focusing on the time frame from 2011 to April 2020 using search terms that were synonyms of “health apps,” “evaluation,” and “frameworks” [12]. This review examined 97 evaluation frameworks and studies that included general digital health app evaluation frameworks, such as the Mobile App Rating Scale (MARS), and more domain-specific

XSL•FO
RenderX

frameworks, such as the SUS and Software Usability Measurement Inventory (SUMI).

The scoring and rating techniques varied within the different frameworks, that is, 23% of frameworks used a 5-point scale, 6% a 3-point scale, 3% a 7-point scale, 2% a 4-point scale, and 1% a 10-point scale. A total of 24% frameworks did not elaborate on the scaling system, 20% used a mixed approach, 13% were dichotomous, and 8% did not use a scaling system [12]. Moreover, the frameworks' scoring modalities also varied as 37% did not report a score, 23% used a mean score, 13% used a total sum, 11% used mixed approaches, 9% other, and 6% did not use scoring at all [12]. In Nurgalieva et al [21], it appears that the evaluation of self-declared data from app developers was the most common privacy assessment method.

Hensher et al [12] examined the domain or criteria needed to evaluate digital health apps and found that user experience together with information validity has been the most evaluated criteria. However, this scoping review included frameworks such as SUS and SUMI, and such frameworks that are designed to evaluate usability are not tailored to digital health apps. In their "count," if a framework evaluated an aspect of user experience (UX), it was considered a scale for evaluating UX. Usability is only one aspect of UX.

Lagan et al [15] found that most evaluation frameworks for health apps were concerned with evidence, clinical foundation, and privacy. This study suggests that it is unclear whether engagement has been adequately predicted with the existing frameworks. The study also suggests that a balance between objective and subjective questions is a challenge for evaluation frameworks.

Discussion

Principal Findings

This umbrella review included 15 review articles that were obtained via a systematic search of the literature (see Textbox 1 and Figure 1). The objective of this review was to give a holistic summary of current methods and frameworks used for quality assessment of digital health apps. It appears that frameworks are a common way of assessing digital health apps. Four (27%) of the 15 review articles tried to establish appropriate assessment criteria for digital health apps. Two (50%) of these 4 review articles [12,13] reviewed frameworks for digital health app assessment and derived their criteria. The other 2 (50%) of these 4 review articles [14,15] had predefined criteria based on previous research in the area (see Table 1). In the review articles, there was a lack of discussion regarding the unethical use of dark patterns, defined by UX dictionary [34] as "deceptive design patterns used to mislead users to make them do something they would not do on their own. They are primarily used to generate sales, increase subscriptions, and hit target business numbers." Also, there was no mention of equality, diversity, and/or inclusion.

MARS has been the most frequently used framework for quality assessment according to 2 review articles [12,27]. Across the review articles included in this study, 13 (87%) of 15 included data privacy or security as a criterion. Hence, data privacy or security was the most common criterion. The least mentioned criterion was cost, with 1 (7%) of 15 review articles mentioning it (see Multimedia Appendix 5). Having an overarching assessment framework would reduce the need to apply several separate frameworks in the pursuit of identifying digital health app quality. However, when using frameworks for the quality assessment of digital health apps, it is important to remember the various limitations of frameworks. For example:

1. A framework can contain a lot of questions that sometimes need to be answered in a specific order. Hence, using it may require training.
2. A framework may not "capture" all necessary aspects of a product or system for evaluation.
3. A framework may be better at "capturing" one aspect of a product or system, for example, effectively evaluate ease of use, but lack in clinical assurance.
4. Evaluation is never perfect; frameworks rely on good interrater and intrarater reliability.
5. Frameworks are limited by their domain or condition area, meaning a framework may be suitable to assess all digital health apps (such as Enlight [34]), but not pick up on a condition area that is specific to quality issues, such as insulin intake for diabetes apps.

Speculation can be made that health condition–specific frameworks may provide a more accurate view of quality. This is because specific features may be required by digital health apps targeted at a specific health condition. A framework for a specific health condition, for example, may include questions such as insulin intake for diabetic users of an app. A generic, all-encompassing frameworks may include questions such as "did the app development involve a relevant health expert?" which could only indicate, not verify, whether necessary features are included.

Woulfe et al [16] point out that a framework coined Enlight [34] is a far-reaching and all-encompassing framework for the assessment of digital health apps. However, Enlight includes many questions, more than other generic frameworks used for the assessment of digital health apps. Hence, using it may take more time and curtail its use. However, the use of such frameworks may help in the mitigation of variety of problems with digital health apps [8],9], whereas using inadequate assessment frameworks may lead to overlooking flaws associated with the use of digital health apps. Hence, choosing a framework that adheres to all or most of the criteria mentioned in Table 1, such as DTAC [7] or Enlight [34], should enable a selection of a good quality app, despite not being focused on a specific condition area.

When developing new quality assessment frameworks, it may be helpful for the assessor to assess related criteria together; however, on the flip side, it can create confusion. For example, Nurgalieva et al [20] pointed out that exclusively focusing on security can lead to increased surveillance, which can introduce privacy risks. It can be speculated that if one criterion is being assessed, such as security, then privacy is also being assessed, leading to the omission of questions or areas in the assessment where security and privacy may be in conflict. Hence, caution needs to be taken when merging criteria into one in a framework.

XSL•FO
RenderX

Textbox 2 provides a list of recommendations to improve the development of quality assessment frameworks.

Textbox 2. Recommendations to improve the development of quality assessment frameworks.

1. Standardize definitions of the criteria used for the quality assessment of digital health apps.
2. When choosing any framework, ensure that it has been validated regarding its content and inter-rater reliability.
3. Choosing a framework that adheres to all or most of the criteria mentioned in Table 1 should enable a selection of a good quality app.
4. Ensure that improving compliance with one criterion does not sabotage the other. For example, an increase in security does not sabotage data privacy.
5. Ensure that when frameworks combine multiple criteria into one, for example, data privacy and security, there are no omissions of questions related to the original criteria.
6. More focus on the unethical use of dark patterns in app design.
7. Include criteria related to equality, diversity, and inclusion in order to ensure that digital health apps are widely available to different groups of people.
8. Choosing a framework for a specific health condition may allow for the assessment of specific or necessary features that would not be covered by a generic framework.
9. Ensure that third-party sponsorships do not lead to conflict of interest between developers of an app and sponsors, thus affecting the developers' credibility.
10. Ensure that the language in the privacy policy is easily understandable. Frameworks should point out when privacy policy of an app contains language that is unnecessarily unclear or vague.
11. Specifying the context in which safety concerns are being assessed can reduce confusion, for example, safety related to data privacy or evidence for clinical assurance.

Limitations

This review is based on 15 review articles, most commonly scoping reviews (n=6, 40%), followed by systematic reviews (n=4, 27%), narrative reviews (n=4, 27%), and a rapid review (n=1, 7%). Using different search queries and searching wider publication dates could yield more results. The screening of articles was conducted by 1 coauthor. This review only included article reviews that contained the word "review" in the title. Any review about quality of digital health apps that was condition specific (eg, diabetes) was excluded from this umbrella review.

Conclusions

The majority of frameworks do not meet all the criteria identified from the reviewed articles. Safety concerns associated with the use of digital health apps may be mitigated with the use of quality frameworks. Some criteria for the assessment of digital health apps may conflict with each other. For example, overly focusing on security may lead to privacy concerns. Research indicates that subjective questions, although difficult to standardized, may be the most useful when assessing engagement.

Acknowledgments

The study discussed in this paper has been conducted as part of a PhD Co-operative Awards in Science and Technology (CAST) award, with funding from the Department for the Economy in Northern Ireland and the ORCHA (Organisation for the Review of Care and Health Applications) in the United Kingdom.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting Items for Overviews of Reviews (PRIOR) checklist.
[PDF File (Adobe PDF File), 239 KB - jmir_v26i1e58616_app1.pdf]

Multimedia Appendix 2
Search queries.
[DOCX File , 15 KB - jmir_v26i1e58616_app2.docx]

Multimedia Appendix 3
JBI critical appraisal of systematic reviews.

XSL•FO
RenderX

[DOCX File , 45 KB - jmir_v26i1e58616_app3.docx]

Multimedia Appendix 4
Article characteristics.
[DOCX File , 105 KB - jmir_v26i1e58616_app4.docx]

Multimedia Appendix 5
Criteria addressed by each paper.
[DOCX File , 53 KB - jmir_v26i1e58616_app5.docx]

References

1. The NHS long term plan. URL: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/ [accessed 2021-12-08]
2. The NHS long term plan and COVID-19. The Health Foundation. URL: https://www.health.org.uk/publications/reports/the-nhs-long-term-plan-and-covid-19 [accessed 2021-10-27]
3. Kern J, Skye A, Krupnick M, Pawley S, Pedersen A, Preciado K. Written consent of IQVIA and the IQVIA Institute. Digital Health Trends. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/digital-health-trends-2021 [accessed 2021-06-06]
4. Regulating medical devices in the UK. GOV.UK. URL: https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk [accessed 2023-06-27]
5. ISO/TS 82304-2:2021 Health software Part 2: Health and wellness apps—quality and reliability. ISO. URL: https://www.iso.org/standard/78182.html [accessed 2023-02-15]
6. Evidence standards framework for digital health technologies. NICE. URL: https://www.nice.org.uk/corporate/ecd7 [accessed 2022-06-02]
7. Digital Technology Assessment Criteria (DTAC). NHS. URL: https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/ [accessed 2021-11-06]
8. Giebel GD, Speckemeier C, Abels C, Plescher F, Börchers K, Wasem J, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023;25:e43808 [FREE Full text] [doi: 10.2196/43808] [Medline: 37171838]
9. Aljedaani B, Babar MA. Challenges with developing secure mobile health applications: systematic review. JMIR Mhealth Uhealth 2021;9(6):e15654 [FREE Full text] [doi: 10.2196/15654] [Medline: 34152277]
10. Checklist for systematic reviews and research syntheses. JBI. 2017. URL: https://jbi.global/sites/default/files/2019-05/JBI_Critical_Appraisal-Checklist_for_Systematic_Reviews2017_0.pdf [accessed 2023-12-07]
11. Aromataris EC, Fernandez R, Godfrey C, Holly C, Khalil H, Tungpunkom P. Umbrella reviews (2020). In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, South Australia: JBI; 2024.
12. Hensher M, Cooper P, Dona S, Angeles M, Nguyen D, Heynsbergh N, et al. Scoping review: development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers. J Am Med Inform Assoc 2021;28(6):1318-1329 [FREE Full text] [doi: 10.1093/jamia/ocab041] [Medline: 33787894]
13. Nouri R, Kalhori SRN, Ghazisaeedi M, Marchand G, Yasini M. Criteria for assessing the quality of mHealth apps: a systematic review. J Am Med Inform Assoc 2018;25(8):1089-1098 [FREE Full text] [doi: 10.1093/jamia/ocy050] [Medline: 29788283]
14. Moshi MR, Tooher R, Merlin T. Suitability of current evaluation frameworks for use in the health technology assessment of mobile medical applications: a systematic review. Int J Technol Assess Health Care 2018;34(5):464-475. [doi: 10.1017/S026646231800051X] [Medline: 30201060]
15. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open 2021;11(3):e047001 [FREE Full text] [doi: 10.1136/bmjopen-2020-047001] [Medline: 33741674]
16. Woulfe F, Fadahunsi KP, Smith S, Chirambo GB, Larsson E, Henn P, et al. Identification and evaluation of methodologies to assess the quality of mobile health apps in high-, low-, and middle-income countries: rapid review. JMIR Mhealth Uhealth 2021;9(10):e28384 [FREE Full text] [doi: 10.2196/28384] [Medline: 34636737]
17. Velasco M, Perleth M, Drummond M, Gürtner F, Jørgensen T, Jovell A, et al. Best practice in undertaking and reporting health technology assessments. Working group 4 report. Int J Technol Assess Health Care 2002;18(2):361-422. [doi: 10.1017/s0266462302000284] [Medline: 12053427]
18. Merlin T, Tamblyn D, Ellery B, INAHTA Quality Assurance Group. What's in a name? Developing definitions for common health technology assessment product types of the International Network of Agencies for Health Technology Assessment (INAHTA). Int J Technol Assess Health Care 2014;30(4):430-437. [doi: 10.1017/S0266462314000543] [Medline: 25397872]
19. Mobile health index and navigation database, app evaluation resources. MINDapps. URL: https://mindapps.org/ [accessed 2023-11-09]

XSL•FO
RenderX

20. Benjumea J, Ropero J, Rivera-Romero O, Dorronzoro-Zubiete E, Carrasco A. Privacy assessment in mobile health apps: scoping review. JMIR Mhealth Uhealth 2020;8(7):e18868 [FREE Full text] [doi: 10.2196/18868] [Medline: 32459640]
21. Nurgalieva L, O'Callaghan D, Doherty G. Security and privacy of mHealth applications: a scoping review. IEEE Access 2020;8:104247-104268 [FREE Full text] [doi: 10.1109/access.2020.2999934]
22. Muro-Culebras A, Escriche-Escuder A, Martin-Martin J, Roldán-Jiménez C, De-Torres I, Ruiz-Muñoz M, et al. Tools for evaluating the content, efficacy, and usability of mobile health apps according to the consensus-based standards for the selection of health measurement instruments: systematic review. JMIR Mhealth Uhealth 2021;9(12):e15433 [FREE Full text] [doi: 10.2196/15433] [Medline: 34855618]
23. Grundy Q. A review of the quality and impact of mobile health apps. Annu Rev Public Health 2022;43:117-134 [FREE Full text] [doi: 10.1146/annurev-publhealth-052020-103738] [Medline: 34910582]
24. Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. J Am Med Inform Assoc 2020;27(2):330-340 [FREE Full text] [doi: 10.1093/jamia/ocz175] [Medline: 31599936]
25. Carmi L, Zohar M, Riva GM. The European General Data Protection Regulation (GDPR) in mHealth: theoretical and practical aspects for practitioners' use. Med Sci Law 2023;63(1):61-68. [doi: 10.1177/00258024221118411] [Medline: 35950240]
26. Galvin HK, DeMuro PR. Developments in privacy and data ownership in mobile health technologies, 2016-2019. Yearb Med Inform 2020;29(1):32-43 [FREE Full text] [doi: 10.1055/s-0040-1701987] [Medline: 32823298]
27. Azad-Khaneghah P, Neubauer N, Miguel Cruz A, Liu L. Mobile health app usability and quality rating scales: a systematic review. Disabil Rehabil Assist Technol 2021;16(7):712-721. [doi: 10.1080/17483107.2019.1701103] [Medline: 31910687]
28. General Data Protection Regulation. URL: https://gdpr-info.eu/ [accessed 2022-05-10]
29. Marelli L, Lievevrouw E, van Hoyweghen I. Fit for purpose? The GDPR and the governance of European digital health. Policy Studies 2020;41(5):447-467 [FREE Full text] [doi: 10.1080/01442872.2020.1724929]
30. Hajesmaeel-Gohari S, Khordastan F, Fatehi F, Samzadeh H, Bahaadinbeigy K. The most used questionnaires for evaluating satisfaction, usability, acceptance, and quality outcomes of mobile health. BMC Med Inform Decis Mak 2022;22(1):22 [FREE Full text] [doi: 10.1186/s12911-022-01764-2] [Medline: 35081953]
31. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digit Health 2019;1(2):e52-e54 [FREE Full text] [doi: 10.1016/S2589-7500(19)30013-5] [Medline: 33323229]
32. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
33. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline: 30973342]
34. Baumel A, Faber K, Mathur N, Kane JM, Muench F. Enlight: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based ehealth interventions. J Med Internet Res 2017;19(3):e82 [FREE Full text] [doi: 10.2196/jmir.7270] [Medline: 28325712]

Abbreviations

DTAC: Digital Technology Assessment Criteria
GDPR: General Data Protection Regulation
ISO: International Organization for Standardization
JBI: Joanna Briggs Institute
MARS: Mobile App Rating Scale
MIND: Mobile health Index and Navigation Database
NHS: National Health Service
NICE: National Institute for Health and Care Excellence
PICOS: Population, Patient, or Problem; Intervention; Comparison; Outcomes; and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SUMI: Software Usability Measurement Inventory
SUS: System Usability Scale
UX: user experience
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by YC Foong, FA Causio; comments to author 07.06.24; revised version received 23.07.24; accepted 25.07.24; published 10.10.24.

Please cite as:
Zych MM, Bond R, Mulvenna M, Martinez Carracedo J, Bai L, Leigh S
Quality Assessment of Digital Health Apps: Umbrella Review
J Med Internet Res 2024;26:e58616
URL: https://www.jmir.org/2024/1/e58616
doi:10.2196/58616
PMID:

©Maciej Marek Zych, Raymond Bond, Maurice Mulvenna, Jorge Martinez Carracedo, Lu Bai, Simon Leigh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Co-Designing Remote Patient Monitoring Technologies for Inpatients: Systematic Review

Jennifer Sumner[1,2*], BSc, Msc, PhD; Si Ying Tan[2,3*], PhD; Yuchen Wang[4*], BSc; Camille Hui Sze Keck[2*], BSc; Eunice Wei Xin Lee[5*], MBBS; Emily Hwee Hoon Chew[2*], BSc; Alexander Wenjun Yip[2*], MBBS

[1]Medical Affairs-Research, Alexandra Hospital, Singapore, Singapore

[2]Alexandra Research Centre for Healthcare In the Virtual Environment (ARCHIVE), Alexandra Hospital, Singapore, Singapore

[3]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

[4]School of Computing, National University of Singapore, Singapore, Singapore

[5]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[*]all authors contributed equally

Corresponding Author:
Jennifer Sumner, BSc, Msc, PhD
Medical Affairs-Research
Alexandra Hospital
378 Alexandra Road
Singapore, 159964
Singapore
Phone: 65 6908 2222
Email: jennyssumner@gmail.com

Abstract

Background: The co-design of health technology enables patient-centeredness and can help reduce barriers to technology use.

Objective: The study objectives were to identify what remote patient monitoring (RPM) technology has been co-designed for inpatients and how effective it is, to identify and describe the co-design approaches used to develop RPM technologies and in which contexts they emerge, and to identify and describe barriers and facilitators of the co-design process.

Methods: We conducted a systematic review of co-designed RPM technologies for inpatients or for the immediate postdischarge period and assessed (1) their effectiveness in improving health outcomes, (2) the co-design approaches used, and (3) barriers and facilitators to the co-design process. Eligible records included those involving stakeholders co-designing RPM technology for use in the inpatient setting or during the immediate postdischarge period. Searches were limited to the English language within the last 10 years. We searched MEDLINE, Embase, CINAHL, PsycInfo, and Science Citation Index (Web of Science) in April 2023. We used the Joanna Briggs Institute critical appraisal checklist for quasi-experimental studies and qualitative research. Findings are presented narratively.

Results: We screened 3334 reports, and 17 projects met the eligibility criteria. Interventions were designed for pre- and postsurgical monitoring (n=6), intensive care monitoring (n=2), posttransplant monitoring (n=3), rehabilitation (n=4), acute inpatients (n=1), and postpartum care (n=1). No projects evaluated the efficacy of their co-designed RPM technology. Three pilot studies reported clinical outcomes; their risk of bias was low to moderate. Pilot evaluations (11/17) also focused on nonclinical outcomes such as usability, usefulness, feasibility, and satisfaction. Common co-design approaches included needs assessment or ideation (16/17), prototyping (15/17), and pilot testing (11/17). The most commonly reported challenge to the co-design process was the generalizability of findings, closely followed by time and resource constraints and participant bias. Stakeholders' perceived value was the most frequently reported enabler of co-design. Other enablers included continued stakeholder engagement and methodological factors (ie, the use of flexible mixed method approaches and prototyping).

Conclusions: Co-design methods can help enhance interventions' relevance, usability, and adoption. While included studies measured usability, satisfaction, and acceptability—critical factors for successful implementation and uptake—we could not determine the clinical effectiveness of co-designed RPM technologies. A stronger commitment to clinical evaluation is needed. Studies' use of diverse co-design approaches can foster stakeholder inclusivity, but greater standardization in co-design terminology is needed to improve the quality and consistency of co-design research.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58144) doi:10.2196/58144

KEYWORDS

remote patient monitoring; technology; inpatient; care transition; systematic review; health technology; patient-centeredness; technology use; effectiveness; study design; assessment; pilot testing; health care; technologies; terminology; quality and consistency; telehealth; telemonitoring

Introduction

In recent decades, health care systems have significantly transformed thanks to innovations in wearable devices, telemedicine, and artificial intelligence [1,2]. Adopting technology into health care brings various advantages, including enhancements to patient-centric care, gains in operational efficiency, and diversification of care delivery approaches. Additionally, care providers are increasingly viewing technology as a solution to workforce shortages [3,4]. One area undergoing rapid growth is remote patient monitoring (RPM) [5-7]. In brief, RPM refers to tools that capture health data that can be reviewed remotely [8]. RPM technologies encompass a wide range of tools, including wearable devices, telehealth platforms, mobile apps, implantable devices, and Internet of Things devices.

RPM technologies serve diverse purposes, from treating and observing acute conditions in the inpatient and care transition period to longer-term maintenance of chronic diseases in the outpatient setting. During inpatient care, RPM can enable continuous, real-time monitoring of patients, facilitating early detection of complications, and allowing for timely interventions [5,6,9-12]. As patients transition from inpatient to care in their homes, RPM ensures continuity in monitoring, supporting adherence to discharge instructions, and facilitating prompt responses to declines in health [5,6,9-12]. Furthermore, RPM technologies can assist with earlier discharge or hospital-at-home care models, reducing the length of stay [5,6,9-12]. In the outpatient setting, RPM technologies shift away from acute care to long-term maintenance of chronic diseases, providing oversight and self-management support [5,6,9-12].

While RPM technology holds significant promise, many potentially impactful technologies fail to be adopted or scaled in practice [13]. One prevailing theory suggests that inadequate stakeholder engagement during the technology development phase inhibits the relevance and adoption of the resulting device [13,14]. The omission of stakeholder insights can lead to poor consideration of user needs, leading to the creation of solutions that fail to fulfill their intended purpose effectively. Accordingly, participatory design of technology is advisable to enhance effectiveness and acceptance [15]; one such approach is co-design. Co-design helps to ensure that solutions closely align with user requirements, improving the relevance and usability of the final product and broader adoption [15,16].

Co-design involves a collaborative effort among stakeholders, including technologists, patients, and health care providers, to develop interventions or services [15,17,18]. The process is iterative and entails multiple phases of development, refinement, and evaluation to achieve the final outcome. Built on the premise that users are experts in their experiences, co-design engages users to improve and innovate services or products [19,20]. There are multiple successful examples of co-designed technology in health care that have improved patient satisfaction, enhanced outcomes, and reduced costs [17,21-24].

As investment and the need for RPM technology grows [25], optimizing the technology design becomes crucial to meet stakeholders' needs and ensure uptake. Therefore, our aim was to gain a comprehensive understanding of the role and impact of participatory design methods, specifically co-design, in the RPM domain. As RPM technology in the inpatient and outpatient settings has unique purposes and requirements, we specifically focus on RPM technology during the inpatient and care transition period to home (ie, RPM technology for acute rather than chronic care needs). Our specific objectives were to identify what RPM technology has been co-designed for inpatients and how effective it is, to identify and describe the co-design approaches used to develop RPM technologies and in which contexts they emerge, and to identify and describe barriers and facilitators of the co-design process.

Methods

Study Design

This study is reported and conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [26]. The systematic review protocol was prospectively registered on the PROSPERO database of systematic reviews (registration CRD42024505427). A copy of the PRISMA checklist is included in Multimedia Appendix 1.

Literature Search

We searched MEDLINE, Embase, CINAHL, PsycInfo, and Science Citation Index (Web of Science) in April 2023 using a combination of MeSH terms and keywords around the themes of co-design and RPM technology. The initial strategy was developed in MEDLINE and adapted for use in the other databases. The reference lists of included papers were also searched for further relevant references. A copy of the MEDLINE search strategy is included in Multimedia Appendix 2.

Study Selection

Citations were downloaded and managed in EndNote X9 (Clarivate). Five reviewers (JS, YW, CHSK, EWXL, and EHHC) independently screened titles and abstracts for inclusion according to predefined screening criteria (Textbox 1). When reviewers could not reach a screening decision, the group discussed the paper until a consensus was reached. All reviewers screened the first 50 papers and then met to discuss screening alignment and disagreements; this helped to ensure screening consistency among the reviewers and aided with the refinement of the eligibility criteria. Shortlisted papers were then full-text

XSL•FO
RenderX

screened by 4 reviewers in pairs (YW, CHSK, EWXL, and EHHC). Uncertainties around eligibility were discussed and resolved by a fifth reviewer (JS or SYT). If publications were unclear, efforts were made to contact the authors and obtain further details.

Textbox 1. Participants, intervention, control, outcomes, and study types eligibility criteria.

Participants: Any stakeholders involved in the use or development of remote patient monitoring (RPM) technologies, including health care workers, patients admitted to hospital, technologists, designers, or researchers.

Intervention: Any co-designed RPM technology for use in the inpatient hospital setting or during the immediate postdischarge period (ie, care transition). RPM technology may include but is not limited to wearable devices, telehealth platforms, mobile apps, implantable devices, and Internet of Things devices. We defined co-design as "The participation and equal collaboration between service providers, users, carers and the broader community to co-design health-related technology." To be considered as a "co-design" study, studies had to include the following attributes: (1) involvement of all relevant stakeholders; AND (2) evidence of collaboration between stakeholders beyond only information gathering from consumers; AND (3) evidence that the stakeholders were involved in the development process at more than 1 time point, that is, demonstrates meaningful contribution of stakeholders to project; AND (4) multiple, iterative stages of development, such as needs assessment, ideation, prototyping, pilot testing (ie, usability), and impact evaluation.

Control: For interventional studies, the control group is defined as those not using a co-designed technology. Studies without a control group are also eligible for inclusion if the other criteria are met.

Outcomes: Any papers with clinical or patient-reported health and well-being outcomes. Papers with data on the experience of the co-design process, including facilitators of and barriers to co-design, were also eligible.

Study types: All study types were considered: that is, experimental, observational, quantitative, and qualitative.

Other: Studies were restricted to English language-only papers, and the search was limited to the last 10 years. The date restriction was chosen to identify relevant technologies for the modern-day context.

Data Extraction and Management

Five reviewers were involved in data extraction (JS, YW, CHSK, EWXL, and EHHC). One reviewer extracted the data, and a different reviewer checked the data for accuracy. Any disagreements were discussed and resolved by a third reviewer. Extracted data items included study and population characteristics, intervention details, co-design process information, facilitators and barriers to the co-design process, and outcome measures. In cases where multiple publications were identified for one study, data from the primary study publication were extracted, and the additional publications were scanned for additional information. The extraction sheet was piloted on 3 papers, and refinements were made to the extraction sheet before the remaining papers were extracted.

Quality Assessment

We used the Joanna Briggs Institute critical appraisal checklist to assess the risk of bias for quasi-experimental studies that evaluated a final intervention. The checklist includes 9 questions grouped into 7 domains, which prompt the reviewer to assess the risk of bias [27]. For qualitative studies that evaluated a final intervention, we used the Joanna Briggs Institute critical appraisal for qualitative research [28]. The checklist contains 10 items, which prompt the reviewer to appraise the quality of the study. One reviewer assessed each study (JS), and a second reviewer checked the decisions (YW). Any disagreements were discussed until a consensus was reached.

Data Synthesis

Study outcomes are presented in a narrative synthesis. We descriptively report on the study characteristics, type and purpose of the included interventions, the target population for each intervention, and the technology components used. Each intervention was categorized into 1 of 6 groups based on its primary purpose: educational, monitoring of health, adherence, safety alerts, 2-way communication, and intervention delivery. Categories were agreed upon through discussion. We then plotted the categories (inner circle) into a sunburst plot with associated examples (outer circle). As a meta-analysis was impossible due to a lack of data, we summarize the number of studies reporting efficacy measures, the types of outcomes measured, and the corresponding impact as text. Co-design approaches are described narratively, highlighting the typical stages of co-design undertaken, the methods used, those involved in the process, and the evaluation metrics reported. Finally, we describe the barriers and enablers of the co-design process narratively. In addition, we generated a word cloud to visualize the co-design process barriers and enablers and their frequency of occurrence. To generate the word cloud, the study team first categorized the types of barriers and enablers to co-design through mutual discussion. One reviewer (JS) applied the categories to the data, and a second reviewer then cross-checked the categories for relevance and accuracy (SYT). The number of publications mentioning a particular barrier or enabler was then mapped to each category. A web-based word cloud generator, Flourish, was then used to create the image.

Results

Overview

We identified 3334 reports, including 315 duplicates, which we removed. Independent screening of the remaining 3019 reports led to a further exclusion of 2952 reports. Following the full-text screening, we excluded 39 reports, leaving 28 included (for 17 projects; Figure 1). The most common reason for exclusion was due to the use of non–co-design methodologies to develop the RPM technology. For example, some studies only engaged in needs assessments and did not actively collaborate with stakeholders. In other cases, the design team did not include key stakeholders (ie, direct technology users). A few studies reported the start of a co-design process but had yet to undertake

XSL•FO
RenderX

the work; these we excluded. Finally, RPM technologies not developed for the hospital setting or the immediate postdischarge period (ie, care transition) were excluded. For example, we excluded community-based self-management tools designed for long-term use.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

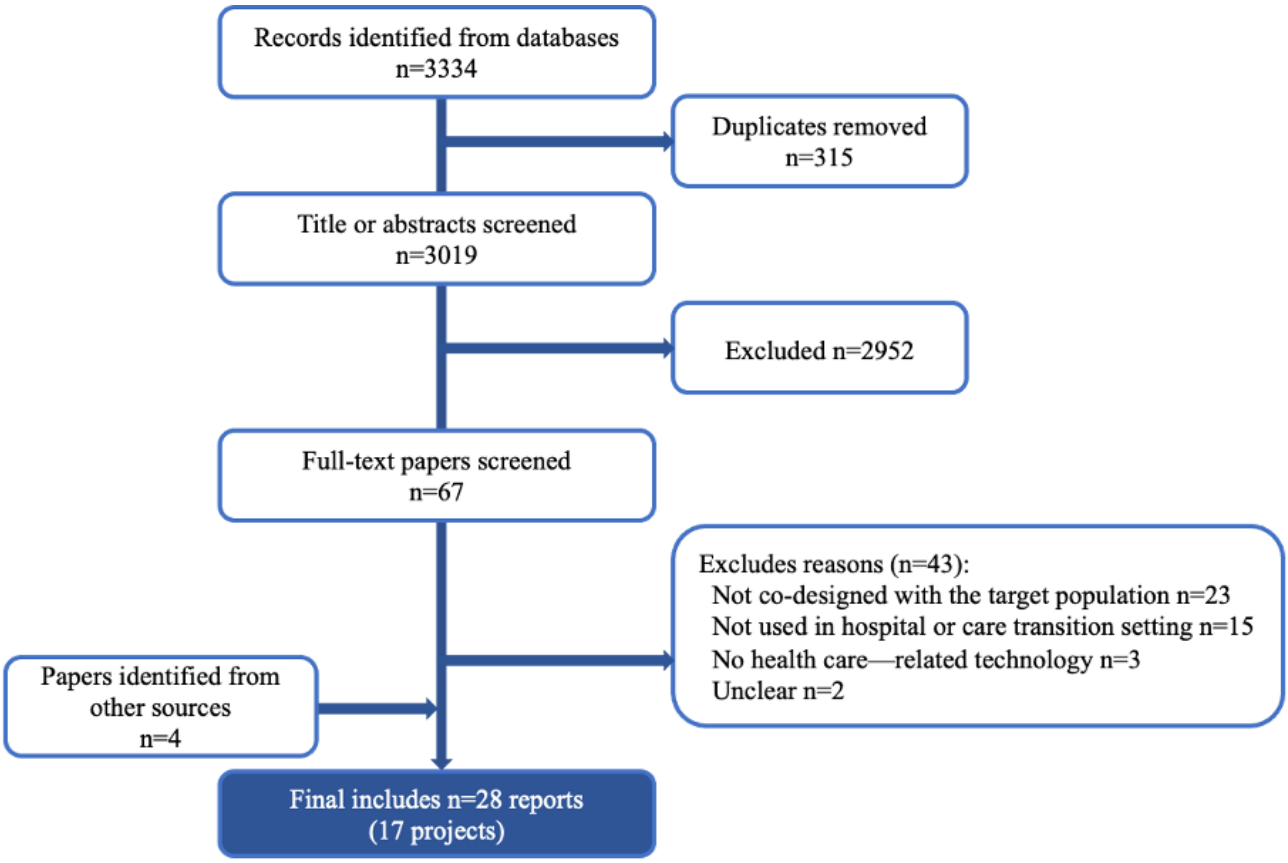

Study Characteristics

Table 1 presents the characteristics of the 17 included projects. Projects were from Europe (n=10) [29-38], the United States (n=4) [39-42], Canada (n=1) [43], New Zealand (n=1) [44], and Sri Lanka (n=1) [45]. Interventions were designed for postsurgical monitoring (n=5) [29-32,39], patients in intensive care (n=2) [33,45], posttransplant monitoring (n=3) [34,35,40], rehabilitation (n=4) [36-38,43], acute inpatients (n=1) [44], presurgical monitoring (n=1) [41], and postpartum care (n=1) [42].

Figure 2 shows the 6 categories of interventions (based on their primary purpose: educational, monitoring of health, adherence, safety alerts, 2-way communication, and intervention delivery) and specific examples from the included reports. Interventions used different technology components including apps (15/17) [30-32,34-45], wearable sensors (3/17) [32,33,36], dashboards (3/17) [32-34], a chatbot (1/17) [30], a video consultation platform (1/17) [34], a smart garment (1/17) [36], and a website (1/17) [29].

XSL•FO
RenderX

Table 1. Included study characteristics.

Author (year), country	Target population and health technology developed	Stages of co-design conducted	Stakeholders involved
Hartup et al (2022, 2023) [29,46], Robinson et al (2023) [47], United Kingdom	Patients undergoing surgery (breast). A web-based intervention (ePainQ) to monitor postoperative pain and facilitate advice provision.	• Needs assessment n=N/R[a] • Pilot n=69	Patients, health care professionals, and academics
Londral et al (2022) [30], Portugal	Patients undergoing surgery (cardiac). A postoperative digital telemonitoring service to monitor postsurgical wounds and improve health literacy.	• Needs assessment n=N/R • Ideation n=N/R • Prototyping n=30 • Pilot n=60	Patients, family caregivers, surgeons, nurses, and technologists
Miller et al (2020) [31], United Kingdom	Patients undergoing surgery (colorectal). An app to support self-care and monitoring post-surgery.	• Needs assessment n=19 • Ideation n>19 • Prototyping n=30	Patients, physicians, nurses, and pharmacy staff
Naeemabadi et al (2019, 2020) [32,48], Denmark	Patients undergoing surgery (knee replacement). A telerehabilitation program solution using sensor technologies to facilitate in-home rehabilitation.	• Needs assessment n=8 • Ideation n=19 • Prototyping n=18 • Pilot 1 n=7; pilot 2 n=4	Patients, health care professionals (physiotherapists, nurses, and orthopedic surgeons), researchers, students, and software developers
Sanger et al (2014, 2016) [39,49], Gunter et al (2016, 2018) [50,51], United States	Patients with surgical wounds. A mobile health wound monitoring app: Mobile Post-Operative Wound Evaluator.	• Needs assessment n=37 • Ideation n=24 • Prototyping n=21 • Pilot 1 n=12; pilot 2 n=40	Patients, providers, health informatics specialists, interaction designers, computer scientists, and patient advisors
Kariyawasam et al (2017) [45], Sri Lanka	Patients in the surgical intensive care unit. An app to support monitoring of hand hygiene compliance.	• Needs assessment n=N/R • Ideation n=N/R • Prototyping n=N/R • Pilot n=N/R	Intensive care unit staff, hospital infection control staff, hospital administrators, and microbiologists
Poncette et al (2022) [33], Germany	Patients in intensive care. Redesign of the user interface for the Vital Sync (version 2.4; Medtronic); a web-based inpatient monitoring platform.	• Needs assessment n=10 • Ideation n=2 • Prototyping n=10	Physicians and medical students
Vaughn et al (2020) [40], United States	Patients undergoing blood and marrow transplant. An app for posttransplant symptom monitoring.	• Prototyping n=32 • Pilot n=36	Patients, a physician researcher, nurse scientists, a nurse practitioner, a pediatric nurse, and child therapists
Duettmann et al (2021) [34], Germany	Patients undergoing a kidney transplant. An app to monitor posttransplant vital signs, well-being, and medication intake and facilitate physician communications and consultations.	• Needs assessment n=N/R • Ideation n=N/R • Prototyping n=N/R • Pilot n=131	Patients and health care professionals
Nielsen et al (2019, 2020) [35,52-54], Denmark	Patients undergoing a kidney transplant. An app to replace in-person follows through tracking of vital sign data, other self-reported health measures, medication use, and a messaging function.	• Needs assessment n=48 • Ideation n=44 • Prototyping n=19 • Pilot 1 n=36	Patients, patient representatives, doctors, nurses, nursing assistants, secretaries, physiotherapists, dieticians, IT designers, innovation consultants, and researchers
Baig et al (2014, 2015, 2020) [44,55,56], New Zealand	Patients in acute care (general medicine). An app that interprets data on blood pressure, heart rate, oxygen saturation, blood glucose, and temperature for patient monitoring in acute care settings.	• Needs assessment N/A[b] • Prototyping n=10 • Pilot 1 n=20; pilot 2 n=30	Nurses, rapid response team members, or emergency response team members
Blair et al (2022) [41], United States	Patients with a congenital heart defect. An app to support home monitoring during the high-risk interstage period (period between surgical procedures) after discharge.	• Needs assessment n=6 (review of web-based blogs) • Ideation n=N/R • Pilot n=11	Caregivers and clinicians
Logsdon et al (2020) [42], United States	Postpartum women. An app to increase a mother's ability to monitor her own health after childbirth.	• Ideation n=5 • Prototyping 1 n=5; prototyping 2 n=22	Postpartum women

XSL•FO
RenderX

Author (year), country	Target population and health technology developed	Stages of co-design conducted	Stakeholders involved
Mortenson et al (2019) [43], Canada	Patients with spinal cord injury (rehabilitation). An app to build self-management skills needed to prevent secondary complications following recent inpatient rehabilitation for spinal cord injury.	• Needs assessment n=N/R • Ideation n=N/R • Prototyping n=N/R	Patients with spinal cord injury and formal and informal caregivers
Perego et al (2022) [36], Italy	Patients who experienced a stroke (rehabilitation). A wearable system for monitoring and evaluating motor rehabilitation activities at home.	• Needs assessment n=N/R • Ideation n=N/R • Prototyping n=10	Patients, physicians and other health care workers from rehabilitation, technologists, and designers
Timmerman et al (2016) [37], Netherlands	Patients with lung cancer (rehabilitation). An app-based rehabilitation program, including physical exercise and symptom or physical activity tracking.	• Needs assessment n=21 • Ideation n=10 • Prototyping n=17	Patients, surgeons, a rehabilitation physician, physiotherapists, and a nurse
An et al (2021) [38], United Kingdom	Patients with chronic obstructive pulmonary disease (rehabilitation). A digital rehabilitation app following hospitalization for chronic obstructive pulmonary disease exacerbation.	• Needs assessment n=26 • Ideation n=26 • Prototyping n=26 • Pilot 1 n=7	Patients, providers, and designers

[a]N/R: not reported.

[b]N/A: not applicable.

Figure 2. Sunburst plot of main technology purpose (inner circle) and associated examples from included reports (outer circle).

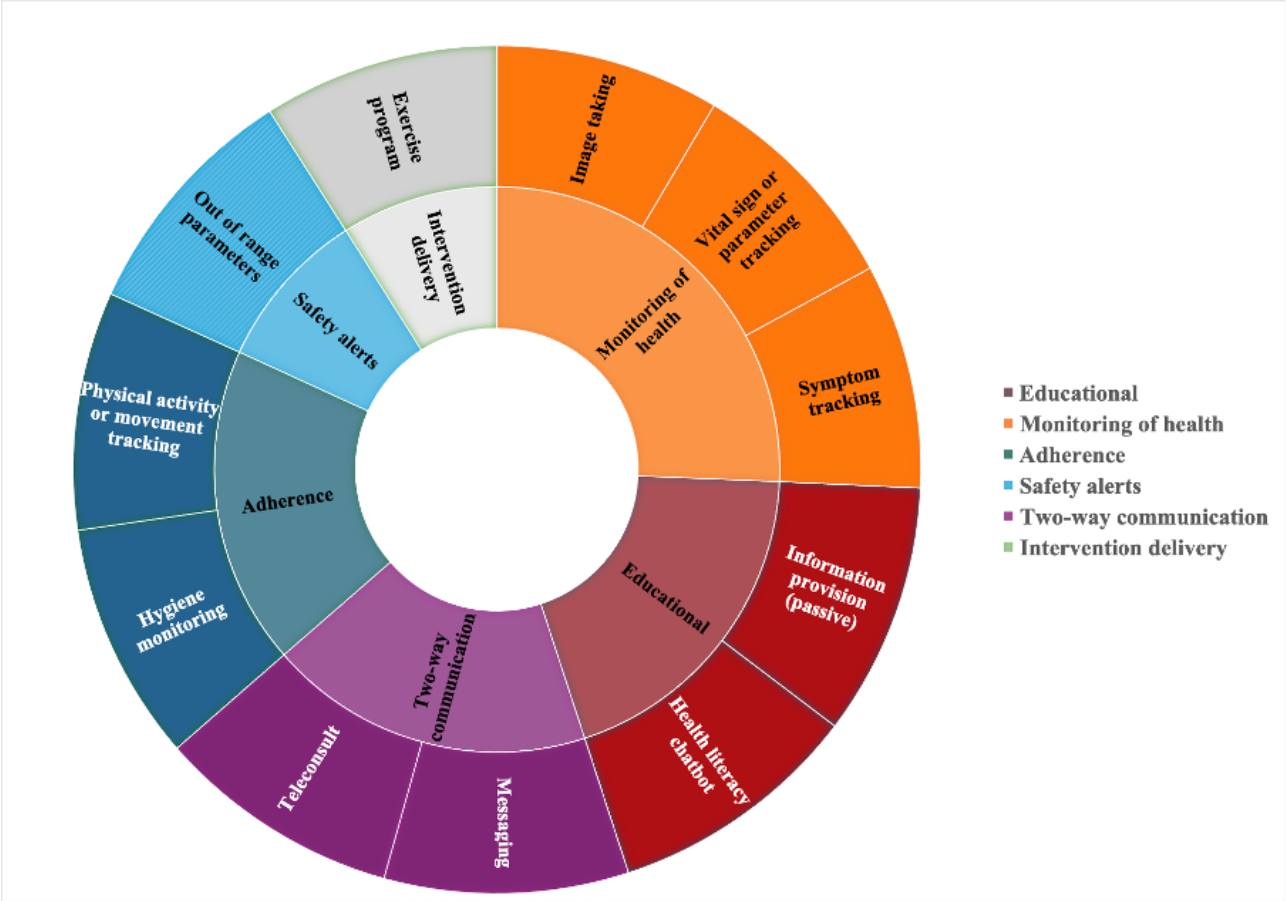

Technology Effectiveness

None of the 17 projects included evaluated the efficacy of their co-designed RPM technology. Three pilot studies [30,45,51] reported on clinical outcomes. Londral et al [30] compared a prospective sample of patients (n=29) using a postoperative digital telemonitoring service to a retrospectively matched control group receiving usual care (n=30). They found that the average critical incident rate (readmission, surgery, or death) was lower in the telemonitored group compared to the control group (n=60; *P*=.01). Kariyawasam et al [45] conducted an uncontrolled pilot (sample size unknown) assessing the effectiveness of a hand hygiene monitoring platform. They found the proportion of participants with positive cultures after an intensive care unit stay, antibiotic use after an intensive care unit stay, and signs and symptoms of infection were 12%, 79%,

XSL•FO
RenderX

and 60%, respectively. Gunter et al [51] conducted an uncontrolled pilot (n=40) evaluating a mobile wound monitoring app. They identified 7 wound complications and 1 false negative during the 2-week evaluation period. Two participants were also readmitted to the hospital but for reasons unrelated to wound infection.

Quality Assessment

For the 3 pilot studies that reported clinical outcomes, the risk of bias was low to moderate across the domains. Londral et al [30] was found to have a moderate risk of bias in 5 domains, Kariyawasam et al [45] in 2 domains, and Gunter et al [51] in 3 domains. All other domains had a low risk of bias. For the remaining quantitative studies (not reporting clinical outcomes), a moderate risk of bias was reported in 1 domain for each study [29,34,44,50,56] and an unclear risk of bias was reported in 5 domains for one study [34] and in 2 domains for another [29]. All other domains had a low risk of bias. For the studies reporting qualitative outcomes, bias was detected in 3 domains for 2 studies [40,48] and in 2 domains for 3 studies [38,41,48]. Unclear bias was also found in 2 domains for one study [38] and in 1 domain for another study [40]. All other domains had a low risk of bias. The quality assessment findings are summarized in Multimedia Appendices 3 and 4.

Co-Design Approaches Used

Overview

Co-design is an iterative process typically involving several stages. We found that not every study conducted every co-design stage (Table 1). In total, 12 of the 17 projects included needs assessment, ideation with stakeholders, and solution prototyping [30-39,43,45]. Two studies did not report on a needs assessment phase [40,42], 1 of which went straight to prototyping [40]. Three studies did not report on ideation activities [29,40,44], and 2 studies developed no prototype and went straight to perform pilot evaluations [29,41].

Health care providers were the most commonly included stakeholders in co-design (16/17) [29-41,43-45]. All but 3 studies involved patients or caregivers; 2 were for vital sign tracking systems [33,44], and 1 was for monitoring staff hygiene [45]. The lack of patient involvement was unsurprising, as patients were not users in these examples. Finally, technologists or designers were involved in 6 of 17 studies [30,32,35,36,38,39]. Other less common stakeholders were health service administrators and patient advocacy groups. No study involved commissioners, service managers, or policy makers.

Needs and Ideation

The number of participants involved in this co-design stage ranged between 1 and 48. Methodological approaches common to both stages were interviews, focus groups, literature reviews, workshops, and questionnaires. Other approaches (only used in needs assessment) included ethnographic observations, analysis of blog data, and audit of existing services. The method only used in ideation was written feedback from stakeholders or domain experts (ie, email correspondence).

Prototyping

Prototyping involved the creation of mock-ups of the proposed intervention which stakeholders fed back on. The number of participants involved in these activities ranged from 10 to 48, and in some cases, multiple iterations of the prototype were developed and refined with stakeholders. Prototypes ranged in sophistication from light fidelity pen and paper sketches (eg, storyboard) to fully functional interventions (eg, fully functional apps). Stakeholder feedback on the prototype was sought through interviews, focus groups, work groups, structured prototype interaction with a task list, unguided prototype interaction, think-aloud feedback, and questionnaires.

Pilot Evaluation

Pilot studies typically assessed nonclinical outcomes such as usability, usefulness, feasibility, and satisfaction (Table 2). Clinical outcomes were reported in 3 projects [30,45,51]. Studies ranged in size from 7 to 131 participants. Participant feedback was gathered through interviews, focus groups, questionnaires, and utility metrics.

XSL•FO
RenderX

Table 2. Pilot study evaluation metrics.

Study	Outcomes	Measurement approach
Baig et al (2014, 2015, 2020) [44,55,56]	Satisfaction, mobility, usability, comfort to operate, and acceptability	• Questionnaire
Blair et al (2022) [41]	Usability and acceptability	• Interviews
Duettmann et al (2021) [34]	Uptake	• Enrollment figures
Hartup et al (2022, 2023) [29,46], Robinson et al (2023) [47]	Acceptability, usability, and usefulness	• Questionnaires (EORTC C30 and BR23[a], EQ-5D, HADS[b], BPI[c], and patient activation) • Interviews
Kariyawasam et al (2017) [45]	App use, behavioral attitudinal changes, user experience, percentage with positive cultures, percentage with antibiotic use after intensive care unit stay, and signs and symptoms of infection	• App usage metrics • Interviews or focus groups • Observations • Clinical data
Londral et al (2022) [30]	User experience, patient adoption, recovery experience, adherence, requirement for technology support, and critical incident rate	• User experience questionnaire • Net promoter score • Questionnaire with open-ended questions • App usage metrics • Clinical data
Naeemabadi et al (2019, 2020) [32,48]	User experience, satisfaction, and usability	• Observations • Interviews • Questionnaire
Nielsen et al (2019, 2020) [35,52-54]	User experience	• Interviews or focus groups
Sanger et al (2014, 2016) [39,49], Gunter et al (2016, 2018) [50,51]	Usability, satisfaction, adherence, satisfaction, burden to workflows, surgical site infection, and hospital readmission	• Systems usability questionnaire • Interviews (staff and patients) • App usage metrics • Clinical data
Vaughn et al (2020) [40]	User experience, feasibility, usability, and acceptability	• Interviews • Questionnaire
An et al (2021) [38]	Usability, usefulness, satisfaction, completion of tasks, and error rate	• Interviews • Completion of tasks and task error rate

[a]EORTC C30 BR23: European Organisation for the Research and Treatment of Cancer Quality of Life questionnaire and Breast Cancer Specific Module.

[b]HADS: Hospital and Anxiety Depression Scale.

[c]BPI: Brief Pain Inventory.

Barriers and Facilitators of the Co-Design Process

In total, 16 projects reported on the barriers and facilitators of the co-design process (Multimedia Appendix 5) [30-45]. The most common barrier to co-design is related to the representativeness of the design group and, correspondingly, the generalizability of the output [30,32,33,35,39,40,43]. Studies noted that co-design groups were often small and may not represent the views of a wider audience. Another commonly reported barrier to co-design was bias [30,37,39,43]. Some mentioned that the methodological approach may have influenced the project direction and biased the output. Others reflected that stakeholder feedback may have been biased due to a predilection for socially desirable responses and the involvement of "early adopters" who may not represent others' views. The time- and resource-consuming nature of co-design was also frequently noted and impacted studies in several ways [30-32,35,43], for instance, the inability to create comprehensive prototypes for participant feedback, difficulties maintaining user engagement with the project, and constraints on the number of participants who could be involved in the co-design process. Other less common barriers included stakeholder conflict and the requirement for good facilitation, stakeholder recruitment challenges, logistical challenges (ie, scheduling), and lack of institutional buy-in, which may impact adoption [31-33,35,39].

Perceived value or buy-in was the most commonly reported enabler of co-design [30,33-35,37]. Buy-in helped maintain stakeholder engagement over time and will ultimately aid with

XSL•FO
RenderX

adoption. Furthermore, sustained stakeholder engagement ensured continuous feedback as concepts developed, the ability to effectively troubleshoot, and the enablement of human-centered design [30,31,33,34,41]. Flexibility in the methods was also noted; this helped to deal with logistical barriers and accommodate stakeholders' needs [31,35]. For instance, stakeholders may have full-time jobs and are unable to participate during "office hours." Other less common enablers included the importance of close communication, the use of prototyping that aided early usability assessments, mutual learning among stakeholders, early stakeholder engagement, participant ownership, and the use of mixed methodologies [30,34-36,39].

Discussion

Principal Findings

In this systematic review of 17 co-designed RPM technologies for the hospital setting or during the immediate postdischarge period, we found no studies evaluating the efficacy of their devices. In the included papers, the co-design development phases spanned needs assessment, ideation, prototyping, and pilot testing (ie, focus on usability). The most commonly reported challenges to the co-design process were generalizability of the output, time and resource constraints, and participant bias. Common enablers included ongoing engagement and stakeholder buy-in. Overall, authors frequently assess usability, satisfaction, and acceptability, facilitating the development of stakeholder-centric solutions. Nevertheless, no reported data supported the clinical effectiveness of the co-designed interventions.

Existing evidence from non–co-designed RPM technologies demonstrates promising results in enhancing patient safety, reducing the length of stay while maintaining adherence [9-12], and proving to be cost-effective [57,58]. However, we found no projects evaluating the clinical impact of co-designed RPM technologies. This evidence gap is consistent with other reviews of co-designed health care interventions in different settings, which consistently highlight a lack of clinical evaluation for co-designed interventions [24,59,60]. While metrics such as usability, satisfaction, and acceptability are frequently measured and are certainly important for successful implementation and adoption [61], the absence of robust clinical evidence remains a critical concern that warrants attention.

Notwithstanding the need for technological innovation to improve health care and address ongoing sustainability challenges [62], health systems must be mindful of introducing new vulnerabilities through the introduction of technology. Populations lacking digital literacy or access to technology, such as older patients, individuals with disabilities, or those with lower socioeconomic status, may struggle to use and adhere to RPM approaches, exacerbating existing health disparities [63,64]. Accordingly, involving these vulnerable populations in the design process becomes all the more crucial. Stakeholders can offer invaluable insights into their unique needs, preferences, and challenges, improving an intervention's usability, effectiveness, and safety [15]. Moreover, stakeholder involvement can help to mitigate barriers to adoption [13]. While our included studies targeted a range of conditions and scenarios, there was no specific focus on vulnerable populations. Future efforts should consider the active engagement of these demographics to ensure technology relevance and usability.

To engage stakeholders effectively, including vulnerable groups, the co-design literature emphasizes flexibility in the approach [65-67], a principle reflected in our included studies. We found a wide variety in terms of the stages of co-design undertaken (ie, needs, ideation, and prototyping), the number of stakeholders involved, and the methods used. On the one hand, the diversity in the co-design process helps accommodate the unique needs and abilities of each stakeholder and aids in developing equitable solutions. For example, working professionals are typically time-constrained, while patients with underlying conditions may struggle to engage in conventional discussions. On the other hand, diversity in co-design processes may also be attributed to inconsistencies in the co-design terminology and definition [60]. Accordingly, these different interpretations of the concept of co-design may influence how researchers execute their study, making it difficult to compare and evaluate the value of co-design outcomes across different projects. In time, the development of the Cocreation Research Standards (CORES) may help with greater standardization [68].

Recommendations

While RPM technologies have several reported benefits, avoiding exacerbation of existing health disparities through the introduction of technology is crucial. Designers may wish to consider frameworks, such as the digital health equity framework [63], to inform their co-design process. Furthermore, those engaging in co-design of technology must be mindful to accommodate mixed stakeholder capabilities. To ensure inclusivity and representativeness, researchers should strive to involve a wide range of individuals and groups. However, this will require careful management of stakeholder interactions. For instance, ensuring a shared understanding of the problem and technological possibilities will help to aid stakeholder engagement. Using a variety of engagement approaches will also give different stakeholders a platform to be involved and prevent unwanted power dynamics that may hinder collaboration. Finally, greater investment in the impact evaluation of co-designed technologies is required to establish the clinical value of co-designed interventions and further justify the resource-intensive process.

Limitations

Due to variability in the co-design terminology, it is feasible that we missed some relevant papers. We attempted to minimize this risk by incorporating a broad range of terms in our search strategy and working with an information specialist. As no unified definition for co-design exists, we developed our own criteria, others may have different interpretations. Thus, we may have excluded relevant papers. We also limited our search to the past 10 years, and we felt this was necessary due to rapid development within the RPM field.

Conclusions

Health care organizations increasingly adopt co-design to improve intervention relevance, usability, and subsequent

XSL•FO
RenderX

adoption. However, we could not conclude the effectiveness of co-designed RPM technologies. A more significant commitment to clinical evaluation is needed. Greater standardization in the co-design terminology is also needed to improve the quality and consistency of co-design research.

Acknowledgments

The authors would like to acknowledge the assistance of Melissa Harden, Wong Suei Nee, and Toh Kim Kee for their guidance on the search strategy. This research is supported by the Singapore Ministry of Health's National Medical Research Council RIE2025 Centre Grant Programme (grant NMRC/CG3/003/2022-AH/MOH-0010130-00).

Authors' Contributions

JS, SYT, and YW were responsible for the conceptualization and methodology of the study. Data curation was carried out by JS, SYT, YW, CHSK, EWXL, and EHHC. JS, SYT, and YW conducted the formal analysis, with JS also handling the visualization. JS prepared the original draft, and the writing was reviewed and edited by JS, SYT, YW, CHSK, EWXL, EHHC, and AWY. AWY oversaw the project administration and secured funding, while supervision was provided by JS.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 80 KB - jmir_v26i1e58144_app1.pdf]

Multimedia Appendix 2
Search strategy.
[PDF File (Adobe PDF File), 41 KB - jmir_v26i1e58144_app2.pdf]

Multimedia Appendix 3
Results of quality assessment for experimental studies.
[PDF File (Adobe PDF File), 79 KB - jmir_v26i1e58144_app3.pdf]

Multimedia Appendix 4
Results of quality assessment for qualitative studies.
[PDF File (Adobe PDF File), 63 KB - jmir_v26i1e58144_app4.pdf]

Multimedia Appendix 5
Word cloud depicting frequency of reported barriers (dark blue) and enablers (light blue) to the co-design process (larger words represent higher frequency of reporting).
[PDF File (Adobe PDF File), 60 KB - jmir_v26i1e58144_app5.pdf]

References

1. Nguyen TN, Piuri V, Qi L, Mumtaz S, Lee WH. Guest editorial innovations in wearable, implantable, mobile, & remote healthcare with IoT & sensor informatics and patient monitoring. IEEE J Biomed Health Inform 2023;27(5):2152-2154. [doi: 10.1109/jbhi.2023.3265411]
2. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health 2023;9:20552076231186520 [FREE Full text] [doi: 10.1177/20552076231186520] [Medline: 37426593]
3. How tech can tackle workloads and staff shortages to optimize healthcare. World Economic Forum. URL: https://www.weforum.org/agenda/2023/01/healthcare-workloads-and-staff-shortages-davos23/ [accessed 2024-09-10]
4. Global strategy on human resources for health: workforce 2030. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789241511131 [accessed 2024-09-18]
5. Condry MW, Quan XI. Remote patient monitoring technologies and markets. IEEE Eng Manag Rev 2023;51(3):59-64. [doi: 10.1109/emr.2023.3285688]
6. Hayes CJ, Dawson L, McCoy H, Hernandez M, Andersen J, Ali MM, et al. Utilization of remote patient monitoring within the United States health care system: a scoping review. Telemed J E Health 2023;29(3):384-394. [doi: 10.1089/tmj.2022.0111] [Medline: 35819861]

XSL•FO
RenderX

7. Shaik T, Tao X, Higgins N, Li L, Gururajan R, Zhou X, et al. Remote patient monitoring using artificial intelligence: current state, applications, and challenges. WIREs Data Min Knowl 2023;13(2):e1485. [doi: 10.1002/widm.1485]
8. Remote patient monitoring. Agency for Healthcare Research and Quality. URL: https://psnet.ahrq.gov/perspective/remote-patient-monitoring [accessed 2024-09-10]
9. Su D, Michaud TL, Estabrooks P, Schwab RJ, Eiland LA, Hansen G, et al. Diabetes management through remote patient monitoring: the importance of patient activation and engagement with the technology. Telemed J E Health 2019;25(10):952-959. [doi: 10.1089/tmj.2018.0205] [Medline: 30372366]
10. Mir O, Ferrua M, Fourcade A, Mathivon D, Duflot-Boukobza A, Dumont S, et al. Digital remote monitoring plus usual care versus usual care in patients treated with oral anticancer agents: the randomized phase 3 CAPRI trial. Nat Med 2022;28(6):1224-1231. [doi: 10.1038/s41591-022-01788-1] [Medline: 35469070]
11. Fudim M, Mirro M, Cheng HM. Audicor remote patient monitoring: FDA breakthrough device and technology for heart failure management. JACC Basic Transl Sci 2022;7(3):313-315 [FREE Full text] [doi: 10.1016/j.jacbts.2022.02.007] [Medline: 35411320]
12. Taylor ML, Thomas EE, Snoswell CL, Smith AC, Caffery LJ. Does remote patient monitoring reduce acute care use? A systematic review. BMJ Open 2021;11(3):e040232 [FREE Full text] [doi: 10.1136/bmjopen-2020-040232] [Medline: 33653740]
13. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
14. Shah MU, Guild PD. Stakeholder engagement strategy of technology firms: a review and applied view of stakeholder theory. Technovation 2022;114:102460. [doi: 10.1016/j.technovation.2022.102460]
15. Steen M, Manschot M, Koning ND. Benefits of co-design in service design projects. Int J Design 2011;5(2):53-60 [FREE Full text]
16. Bate P, Robert G. Experience-based design: from redesigning the system around the patient to co-designing services with the patient. Qual Saf Health Care 2006;15(5):307-310 [FREE Full text] [doi: 10.1136/qshc.2005.016527] [Medline: 17074863]
17. Garmann-Johnsen NF, Helmersen M, Eikebrokk TR. Employee-driven digitalization in healthcare: codesigning services that deliver. Health Policy Technol 2020;9(2):247-254. [doi: 10.1016/j.hlpt.2020.03.001]
18. Nusir M, Rekik M. Systematic review of co-design in digital health for COVID-19 research. Univers Access Inf Soc 2022:1-15 [FREE Full text] [doi: 10.1007/s10209-022-00964-x] [Medline: 36618758]
19. Co-designing for social good: the role of citizens in designing and delivering social services, part one. Centre for Social Impact. URL: https://www.yacwa.org.au/wp-content/uploads/2016/09/An-Introduction-to-Co-Design-by-Ingrid-Burkett.pdf [accessed 2024-09-10]
20. Noorbergen TJ, Adam MTP, Roxburgh M, Teubner T. Co-design in mHealth systems development: insights from a systematic literature review. AIS Trans Hum Comput Interact 2021;13(2):175-205. [doi: 10.17705/1thci.00147]
21. Howard J, Tasker LH, Fisher Z, Tree J. Assessing the use of co-design to produce bespoke assistive technology solutions within a current healthcare service: a service evaluation. Disabil Rehabil Assist Technol 2024;19(1):42-51. [doi: 10.1080/17483107.2022.2060355] [Medline: 35426354]
22. Papoutsi C, Wherton J, Shaw S, Morrison C, Greenhalgh T. Putting the social back into sociotechnical: case studies of co-design in digital health. J Am Med Inform Assoc 2021;28(2):284-293 [FREE Full text] [doi: 10.1093/jamia/ocaa197] [Medline: 33043359]
23. Sanz MF, Acha BV, García MF. Co-design for people-centred care digital solutions: a literature review. Int J Integr Care 2021;21(2):16 [FREE Full text] [doi: 10.5334/ijic.5573] [Medline: 33981193]
24. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
25. Remote patient monitoring devices market size, share & industry analysis, by type (products/devices {multi-parameter monitoring, respiratory monitoring, blood glucose monitoring, cardiac monitoring, and others}, and services), by application (oncology, diabetes, cardiovascular diseases, and others), by end-user (payers, providers, and patients), and regional forecast. Fortune Business Insights. URL: https://www.fortunebusinessinsights.com/remote-patient-monitoring-devices-market-106328 [accessed 2024-09-10]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Barker TH, Habibi N, Aromataris E, Stone JC, Leonardi-Bee J, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evid Synth 2024;22(3):378-388. [doi: 10.11124/JBIES-23-00268] [Medline: 38287725]
28. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015;13(3):179-187. [doi: 10.1097/XEB.0000000000000062] [Medline: 26262565]

XSL•FO
RenderX

29. Hartup S, Ashley L, Briggs M, Velikova G, Johnson M. Developing an effective self-management web-based intervention using co-designed patient and multidisciplinary research: ePainQ. Eur J Cancer 2022;175:S56. [doi: 10.1016/s0959-8049(22)01499-x]
30. Londral A, Azevedo S, Dias P, Ramos C, Santos J, Martins F, et al. Developing and validating high-value patient digital follow-up services: a pilot study in cardiac surgery. BMC Health Serv Res 2022;22(1):680 [FREE Full text] [doi: 10.1186/s12913-022-08073-4] [Medline: 35597936]
31. Miller M, Roxburgh CS, McCann L, Connaghan J, Van-Wyk H, McSorley S, et al. Development of a remote monitoring application to improve care and support patients in the first 30 days following colorectal cancer surgery. Semin Oncol Nurs 2020;36(6):151086 [FREE Full text] [doi: 10.1016/j.soncn.2020.151086] [Medline: 33218885]
32. Naeemabadi M, Dinesen B, Andersen OK, Madsen NK, Simonsen OH, Hansen J. Developing a telerehabilitation programme for postoperative recovery from knee surgery: specifications and requirements. BMJ Health Care Inform 2019;26(1):e000022 [FREE Full text] [doi: 10.1136/bmjhci-2019-000022] [Medline: 31039126]
33. Poncette AS, Mosch LK, Stablo L, Spies C, Schieler M, Weber-Carstens S, et al. A remote patient-monitoring system for intensive care medicine: mixed methods human-centered design and usability evaluation. JMIR Hum Factors 2022;9(1):e30655 [FREE Full text] [doi: 10.2196/30655] [Medline: 35275071]
34. Duettmann W, Naik MG, Schmidt D, Pfefferkorn M, Kurz M, Graf V, et al. Digital home-monitoring of patients after kidney transplantation: the MACCS platform. J Vis Exp 2021(170):e61899. [doi: 10.3791/61899] [Medline: 33900281]
35. Nielsen C, Agerskov H, Bistrup C, Clemensen J. User involvement in the development of a telehealth solution to improve the kidney transplantation process: a participatory design study. Health Informatics J 2020;26(2):1237-1252 [FREE Full text] [doi: 10.1177/1460458219876188] [Medline: 31566460]
36. Perego P, Scagnoli M, Sironi R. Co-design the acceptability of wearables in the healthcare field. In: IoT Technologies for Health Care. Cham: Springer International Publishing; 2022:21-32.
37. Timmerman JG, Tönis TM, Dekker-van Weering MGH, Stuiver MM, Wouters MWJM, van Harten WH, et al. Co-creation of an ICT-supported cancer rehabilitation application for resected lung cancer survivors: design and evaluation. BMC Health Serv Res 2016;16:155 [FREE Full text] [doi: 10.1186/s12913-016-1385-7] [Medline: 27121869]
38. An Q, Kelley MM, Yen PY. Using experience-based co-design to develop mHealth App for digital pulmonary rehabilitation management of patients with chronic obstructive pulmonary disease (COPD). 2021 Presented at: HCI International 2021—Late Breaking Posters. 23rd HCI International Conference, HCII 2021, Proceedings; July 24-29, 2021; Virtual Event p. 125-133. [doi: 10.1007/978-3-030-90179-0_17]
39. Sanger PC, Hartzler A, Lordon RJ, Armstrong CA, Lober WB, Evans HL, et al. A patient-centered system in a provider-centered world: challenges of incorporating post-discharge wound data into practice. J Am Med Inform Assoc 2016;23(3):514-525 [FREE Full text] [doi: 10.1093/jamia/ocv183] [Medline: 26977103]
40. Vaughn J, Shah N, Jonassaint J, Harris N, Docherty S, Shaw R. User-centered app design for acutely ill children and adolescents. J Pediatr Oncol Nurs 2020;37(6):359-367 [FREE Full text] [doi: 10.1177/1043454220938341] [Medline: 32646317]
41. Blair L, Vergales J, Peregoy L, Seegal H, Keim-Malpass J. Acceptability of an interstage home monitoring mobile application for caregivers of children with single ventricle physiology: toward technology-integrated family management. J Spec Pediatr Nurs 2022;27(3):e12372. [doi: 10.1111/jspn.12372] [Medline: 35365917]
42. Logsdon MC, Lauf A, Stikes R, Revels A, Vickers-Smith R. Partnering with new mothers to develop a smart phone app to prevent maternal mortality after hospital discharge: a pilot study. J Adv Nurs 2020;76(1):324-327. [doi: 10.1111/jan.14219] [Medline: 31588581]
43. Mortenson WB, Singh G, MacGillivray M, Sadeghi M, Mills P, Adams J, et al. Development of a self-management app for people with spinal cord injury. J Med Syst 2019;43(6):145. [doi: 10.1007/s10916-019-1273-x] [Medline: 31011881]
44. Baig MM, GholamHosseini H, Connolly MJ, Kashfi G. Real-time vital signs monitoring and interpretation system for early detection of multiple physical signs in older adults. 2014 Presented at: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); June 1-4, 2014; Valencia, Spain p. 355-358. [doi: 10.1109/bhi.2014.6864376]
45. Kariyawasam N, Wong MC, Turner P. User centred design and nosocomials in surgical ICUs: a mobile application for peer monitoring and training in hand hygiene. Stud Health Technol Inform 2017;245:743-747. [Medline: 29295197]
46. Hartup SM, Ashley L, Briggs M, Velikova G, Johnson M. Abstract P6-05-57: developing and feasibility testing a web-based intervention (ePainQ) to support post-operative pain and symptom self-management following surgery for breast cancer. Cancer Res 2023;83(Suppl 5):P6-05-57. [doi: 10.1158/1538-7445.sabcs22-p6-05-57]
47. Robinson OC, Pini S, Flemming K, Campling N, Fallon M, Richards SH, et al. Exploration of pain assessment and management processes in oncology outpatient services with healthcare professionals: a qualitative study. BMJ Open 2023;13(12):e078619 [FREE Full text] [doi: 10.1136/bmjopen-2023-078619] [Medline: 38151273]
48. Naeemabadi M, Søndergaard JH, Klastrup A, Schlünsen AP, Lauritsen REK, Hansen J, et al. Development of an individualized asynchronous sensor-based telerehabilitation program for patients undergoing total knee replacement: participatory design. Health Informatics J 2020;26(4):2492-2511 [FREE Full text] [doi: 10.1177/1460458220909779] [Medline: 32175788]

XSL•FO
RenderX

49. Sanger PC, Hartzler A, Han SM, Armstrong CAL, Stewart MR, Lordon RJ, et al. Patient perspectives on post-discharge surgical site infections: towards a patient-centered mobile health solution. PLoS One 2014;9(12):e114016 [FREE Full text] [doi: 10.1371/journal.pone.0114016] [Medline: 25436912]
50. Gunter R, Fernandes-Taylor S, Mahnke A, Awoyinka L, Schroeder C, Wiseman J, et al. Evaluating patient usability of an image-based mobile health platform for postoperative wound monitoring. JMIR Mhealth Uhealth 2016;4(3):e113 [FREE Full text] [doi: 10.2196/mhealth.6023] [Medline: 27683059]
51. Gunter RL, Fernandes-Taylor S, Rahman S, Awoyinka L, Bennett KM, Weber S, et al. Feasibility of an image-based mobile health protocol for postoperative wound monitoring. J Am Coll Surg 2018;226(3):277-286 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2017.12.013] [Medline: 29366555]
52. Nielsen C, Agerskov H, Bistrup C, Clemensen J. 'The hospital and everyday life are two worlds': patients' and healthcare professionals' experiences and perspectives on collaboration in the kidney transplantation process. Nurs Open 2019;6(4):1491-1500 [FREE Full text] [doi: 10.1002/nop2.349] [Medline: 31660177]
53. Nielsen C, Agerskov H, Bistrup C, Clemensen J. Evaluation of a telehealth solution developed to improve follow-up after kidney Transplantation. J Clin Nurs 2020;29(7-8):1053-1063. [doi: 10.1111/jocn.15178] [Medline: 31889357]
54. Nielsen C, Clemensen J, Bistrup C, Agerskov H. Balancing everyday life-patients' experiences before, during and four months after kidney transplantation. Nurs Open 2019;6(2):443-452 [FREE Full text] [doi: 10.1002/nop2.225] [Medline: 30918695]
55. Baig MM, GholamHosseini H, Ahmad F. A usability and user-experience analysis of VitalsAssist: a mHealth application to monitor vital signs in acute care settings. Annu Int Conf IEEE Eng Med Biol Soc 2020;2020:5132-5135. [doi: 10.1109/EMBC44109.2020.9176467] [Medline: 33019141]
56. Baig MM, Gholamhosseini H, Linden M, Connolly MJ. Review of vital signs monitoring systems—patient's acceptability, issues and challenges. NBE 2015 Jan 22;2(1):2-13. [doi: 10.2174/2213385202999140820115756]
57. Zhang Y, Peña MT, Fletcher LM, Lal L, Swint JM, Reneker JC. Economic evaluation and costs of remote patient monitoring for cardiovascular disease in the United States: a systematic review. Int J Technol Assess Health Care 2023;39(1):e25. [doi: 10.1017/S0266462323000156] [Medline: 37114456]
58. De Guzman KR, Snoswell CL, Taylor ML, Gray LC, Caffery LJ. Economic evaluations of remote patient monitoring for chronic disease: a systematic review. Value Health 2022;25(6):897-913 [FREE Full text] [doi: 10.1016/j.jval.2021.12.001] [Medline: 35667780]
59. Voorberg WH, Bekkers VJJM, Tummers LG. A systematic review of co-creation and co-production: embarking on the social innovation journey. Public Manag Rev 2014;17(9):1333-1357. [doi: 10.1080/14719037.2014.930505]
60. Sumner J, Chong LS, Bundele A, Wei Lim Y. Co-designing technology for aging in place: a systematic review. Gerontologist 2021;61(7):e395-e409 [FREE Full text] [doi: 10.1093/geront/gnaa064] [Medline: 32506136]
61. Klaic M, Kapp S, Hudson P, Chapman W, Denehy L, Story D, et al. Implementability of healthcare interventions: an overview of reviews and development of a conceptual framework. Implement Sci 2022;17(1):10 [FREE Full text] [doi: 10.1186/s13012-021-01171-7] [Medline: 35086538]
62. Dall TM, Gallo PD, Chakrabarti R, West T, Semilla AP, Storm MV. An aging population and growing disease burden will require a large and specialized health care workforce by 2025. Health Aff (Millwood) 2013;32(11):2013-2020. [doi: 10.1377/hlthaff.2013.0714] [Medline: 24191094]
63. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]
64. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]
65. What is co-design. Interaction Design Foundation. URL: https://www.interaction-design.org/literature/topics/codesign [accessed 2024-09-10]
66. Co-Design Toolkit. URL: https://aci.health.nsw.gov.au/projects/co-design [accessed 2024-09-10]
67. Giolitto L. Co-design toolkit. Western Australian Council of Social Service. 2017. URL: https://www.wacoss.org.au/library/co-design-toolkit/ [accessed 2024-09-23]
68. Pearce GHN, Magee P, Christensen F, Darlington E, Bernard S, Olivo M, et al. Protocol for the development of the CO-creation REporting Standards (CORES) for research. Coventry University. 2023. URL: https://www.equator-network.org/wp-content/uploads/2023/04/Protocol-for-the-development-of-CORES.pdf [accessed 2024-09-18]

Abbreviations

CORES: Cocreation Research Standards
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RPM: remote patient monitoring

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.03.24; peer-reviewed by E Baker, N Alvarado; comments to author 14.04.24; revised version received 21.06.24; accepted 25.07.24; published 15.10.24.

Please cite as:
Sumner J, Tan SY, Wang Y, Keck CHS, Xin Lee EW, Chew EHH, Yip AW
Co-Designing Remote Patient Monitoring Technologies for Inpatients: Systematic Review
J Med Internet Res 2024;26:e58144
URL: https://www.jmir.org/2024/1/e58144
doi:10.2196/58144
PMID:39405106

©Jennifer Sumner, Si Ying Tan, Yuchen Wang, Camille Hui Sze Keck, Eunice Wei Xin Lee, Emily Hwee Hoon Chew, Alexander Wenjun Yip. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Exploring the Role of Mobile Apps for Insomnia in Depression: Systematic Review

Yi-Hang Chiu[1,2*], MD; Yen-Fen Lee[3*], PhD; Huang-Li Lin[4,5*], MD; Li-Chen Cheng[3*], PhD

[1]Department of Psychiatry, Wan Fang Hospital, Taipei Medical University, Taipei City, Taiwan
[2]Psychiatric Research Center, Wan Fang Hospital, Taipei Medical University, Taipei City, Taiwan
[3]Department of Information and Finance Management, National Taipei University of Technology, Taipei City, Taiwan
[4]Department of Psychiatry, Chang Gung Memorial Hospital, Linkou, Taoyuan, Taiwan
[5]College of Medicine, Chang Gung University, Taoyuan, Taiwan
[*]all authors contributed equally

Corresponding Author:
Li-Chen Cheng, PhD
Department of Information and Finance Management
National Taipei University of Technology
Number 1, Section 3
Zhongxiao East Road, Da'an District
Taipei City, 10608
Taiwan
Phone: 886 2771 2171 ext 6720
Email: lijen.cheng@gmail.com

Abstract

Background: The COVID-19 pandemic has profoundly affected mental health, leading to an increased prevalence of depression and insomnia. Currently, artificial intelligence (AI) and deep learning have thoroughly transformed health care–related mobile apps, offered more effective mental health support, and alleviated the psychological stress that may have emerged during the pandemic. Early reviews outlined the use of mobile apps for dealing with depression and insomnia separately. However, there is now an urgent need for a systematic evaluation of mobile apps that address both depression and insomnia to reveal new applications and research gaps.

Objective: This study aims to systematically review and evaluate mobile apps targeting depression and insomnia, highlighting their features, effectiveness, and gaps in the current research.

Methods: We systematically searched PubMed, Scopus, and Web of Science for peer-reviewed journal articles published between 2017 and 2023. The inclusion criteria were studies that (1) focused on mobile apps addressing both depression and insomnia, (2) involved young people or adult participants, and (3) provided data on treatment efficacy. Data extraction was independently conducted by 2 reviewers. Title and abstract screening, as well as full-text screening, were completed in duplicate. Data were extracted by a single reviewer and verified by a second reviewer, and risk of bias assessments were completed accordingly.

Results: Of the initial 383 studies we found, 365 were excluded after title, abstract screening, and removal of duplicates. Eventually, 18 full-text articles met our criteria and underwent full-text screening. The analysis revealed that mobile apps related to depression and insomnia were primarily utilized for early detection, assessment, and screening (n=5 studies); counseling and psychological support (n=3 studies); and cognitive behavioral therapy (CBT; n=10 studies). Among the 10 studies related to depression, our findings showed that chatbots demonstrated significant advantages in improving depression symptoms, a promising development in the field. Additionally, 2 studies evaluated the effectiveness of mobile apps as alternative interventions for depression and sleep, further expanding the potential applications of this technology.

Conclusions: The integration of AI and deep learning into mobile apps, particularly chatbots, is a promising avenue for personalized mental health support. Through innovative features, such as early detection, assessment, counseling, and CBT, these apps significantly contribute toward improving sleep quality and addressing depression. The reviewed chatbots leveraged advanced technologies, including natural language processing, machine learning, and generative dialog, to provide intelligent and autonomous

XSL•FO
RenderX

interactions. Compared with traditional face-to-face therapies, their feasibility, acceptability, and potential efficacy highlight their user-friendly, cost-effective, and accessible nature with the aim of enhancing sleep and mental health outcomes.

(J Med Internet Res 2024;26:e51110) doi:10.2196/51110

KEYWORDS

depression; insomnia; chatbots; conversational agents; medical apps; systematic review; technical aspects; PRISMA

Introduction

Background

Insomnia and depression have been significantly correlated with one another. Clinically, the incidence of insomnia is higher in patients with depression than in those without, and the symptoms tend to be more severe [1]. In addition to the fact that depression is often accompanied by symptoms of insomnia, insomnia itself is a risk factor for depression [2,3], and severe insomnia can exacerbate depressive symptoms. Insomnia may also serve as a prodromal symptom of depression, as many patients with depression also report insomnia complaints. According to Gebara et al [4], treating insomnia in patients with both depression and insomnia symptoms can contribute to emotional improvement. The treatments for depression and insomnia are interconnected; improvements in depression often lead to relief from insomnia. However, specifically addressing insomnia can also facilitate the alleviation of depressive symptoms.

In clinical practice, the treatment of insomnia primarily involves cognitive behavioral therapy (CBT) and medication, with most patients responding well to these approaches. Particularly during the COVID-19 pandemic, when patients were unable to visit medical institutions for in-person treatment, telemedicine and mobile device–based medical services became increasingly important [5]. As in-person hospital visits become less feasible, the significance of telemedicine models has increased. While these models have not entirely replaced traditional medical practices, they have proven effective in assisting with the treatment of various conditions, including psychiatric disorders. Examples of such models include internet-based services, mobile devices, and software apps [6-8]. Telemedicine models, including video clinics and technology-assisted mobile health (mHealth) approaches, as well as flexible eHealth solutions, have become pivotal in addressing the escalating demand for mental health support amid the scarcity of conventional health care resources [9]. These approaches aim to tackle the challenges posed by global mental health issues and shortages in traditional health care facilities [10,11].

Recently, artificial intelligence (AI) apps have significantly advanced psychiatric diagnostic assistance, symptom tracking, course prediction, morbidity risk assessment, and mental health education [12]. These apps are increasingly being utilized as therapeutic aids, offering advantages such as high accessibility, convenience, low cost, and sustainability. Notably, innovative AI-assisted therapies, such as chatbots, can provide ongoing follow-up and extended treatment while mitigating the stigma associated with disclosing mental illness, thereby enhancing personal well-being [13-17].

Considering past research, to the best of our knowledge, no review papers have examined mobile apps targeting insomnia in individuals with depression. Therefore, this systematic review aims to comprehensively categorize the thematic features of mobile apps designed for addressing insomnia and depression, shedding light on strategies to improve insomnia among individuals with depression. This study offers a systematic overview of the research field, enhances understanding, and reveals new potential applications and research directions.

Related Work

To gain a deeper understanding of the impact of technological advancements on mobile apps in psychiatry, we organized and evaluated systematic and literature reviews published in the past 5 years. This 5-year limit for the literature search is justified by the rapid advancements in AI and mHealth technology. In recent years, significant advancements in medical technologies have rendered older studies potentially less relevant in the current context [18-21]. This study aimed to capture the evolving trends in mental health–related mobile apps resulting from these technological advancements, providing comprehensive insight into this field.

We assessed recent systematic and literature reviews, revealing a focus on 2 main aspects: insomnia and depression. Regarding insomnia, 8 papers [22-29] primarily addressed the application of CBT principles [22], smartphone app support for sleep self-management [23,24], the usability and principles of smartphone apps for insomnia [25,27], wearable device–based sleep monitoring [26], and the study of CBT within electronic mental health apps [28,29]. Regarding depression, 9 papers [30-38] examined topics such as therapist-supported internet-based CBT [30,31,37], user experience analysis of depression apps [32,34], the adoption of mobile apps in the domains of depression and anxiety [33,35,36], and a review of mobile mental health apps [38].

However, there is a lack of comprehensive research on patients with depression who also experience insomnia symptoms. Although depression and insomnia are common comorbidities in clinical settings, the exploration of these specific conditions has not been thoroughly investigated. Therefore, this study primarily aimed to fill this research gap and provide a more comprehensive understanding of the development and application of mobile apps for patients with both depression and insomnia symptoms.

Methods

Study Identification and Selection

This study adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [39]

XSL•FO
RenderX

for the literature search and article selection (Multimedia Appendix 1). Details regarding methodological protocol amendments and clarifications can be found in Multimedia Appendix 2.

Search Strategy

In the first step, a comprehensive systematic search was conducted across multiple databases, including PubMed, Scopus, and the ISI Web of Science, using keywords such as "mobile applications," "chatbots," "insomnia," "insomnia treatment," and "depression." The complete list of search terms is available in Multimedia Appendix 3. The search criteria were further refined by specifying additional constraints, including publication in peer-reviewed academic journals in English and restricting the time frame to articles published between 2017 and early 2023. This ensured that the selected literature from the most recent 5 years aligned with current technologies and the latest research findings. Systematic reviews typically limit inclusion to studies published within a specified time frame to ensure relevance and capture current evidence. This approach helps maintain the validity of the findings by focusing on recent developments and excluding outdated or less pertinent research [18-21]. A preliminary search yielded 383 articles.

Screening Process

Two experienced researchers meticulously screened the studies retrieved from the databases to ensure that they (1) focused on mobile apps addressing both depression and insomnia, (2) involved young or adult participants, and (3) provided data on treatment efficacy. The screening process was conducted in 2 stages: first, they independently screened the titles and abstracts to identify studies that potentially met the inclusion criteria; second, they reviewed the full texts of the selected studies to determine final inclusion. During the coding process, the researchers read the papers and engaged in discussions to reach a consensus on any inconsistencies in the codes. The risk of bias was assessed using the Cochrane Risk of Bias framework [40].

The screening process was conducted based on various aspects, including research topic characteristics, participant demographics, app features, therapeutic efficacy, operating platforms, technical foundations, feasibility, and acceptability. This approach ensured that the selected studies aligned with the research objectives, thereby enhancing the quality and relevance of the final sample. After reviewing the titles, abstracts, and full texts, 18 articles were included in the final sample. Exclusion criteria encompassed nonacademic journal articles, lack of relevance to the topic, insufficient discussion content, and methodological limitations. This review process ensured the high quality and relevance of the final sample to the research objectives.

Data Extraction and Quality Assessment

Data Extraction

A preliminary coding scheme was developed based on an initial review of the included articles. Subsequently, 2 researchers (YHC and HLL), utilized this scheme to code 10 articles, after which they convened to discuss the practicality of the initial and potential additional coding categories. Throughout the coding process, the 2 experienced researchers read the papers and engaged in discussions to reach a consensus on any inconsistencies in the coding. The final coding scheme comprised 7 dimensions: Research topics (eg, cognitive and emotional support, cognitive behavioral interventions, psychological education, sleep diaries, self-help psychotherapy, conversational agents, examining depression symptoms, and digital behavioral interventions); Participants (eg, young people or adults); App functions (eg, early detection, assessment and screening, counseling and mental support, and CBT); Therapeutic efficacy (eg, the effectiveness of mobile apps as sleep intervention tools, the effectiveness of chatbots in improving sleep quality and insomnia symptoms, and the effectiveness of chatbots in improving depression and its symptoms); Operating platforms (eg, Android, iOS, Android and iOS, multiplatform support, and unreported operating platforms); Technical foundation and feature characteristics (eg, AI chatbots, the ExpiWell app, RASA-based open-source conversational AI); and Feasibility and acceptability (ie, the acceptability and feasibility of chatbots, mobile apps as sleep intervention tools, and mobile apps providing depression intervention and emotional support).

Quality Assessment

The risk of bias was assessed using the Cochrane risk of bias framework [40]. One reviewer (YHC) initially evaluated each study's risk of bias, which was then verified by a second reviewer (HLL). The assessment encompassed sequence generation, allocation concealment, blinding, completeness of outcome data, selective reporting, and other potential biases. Any discrepancies were resolved through consensus between the 2 reviewers.

Results

Study Inclusion

Of the 383 abstracts initially identified, 365 were excluded based on title and abstract screening, as well as duplication; 18 full-text articles met the inclusion criteria. The search process is illustrated using the PRISMA flowchart (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart illustrating the literature search and selection process of studies included in this review.

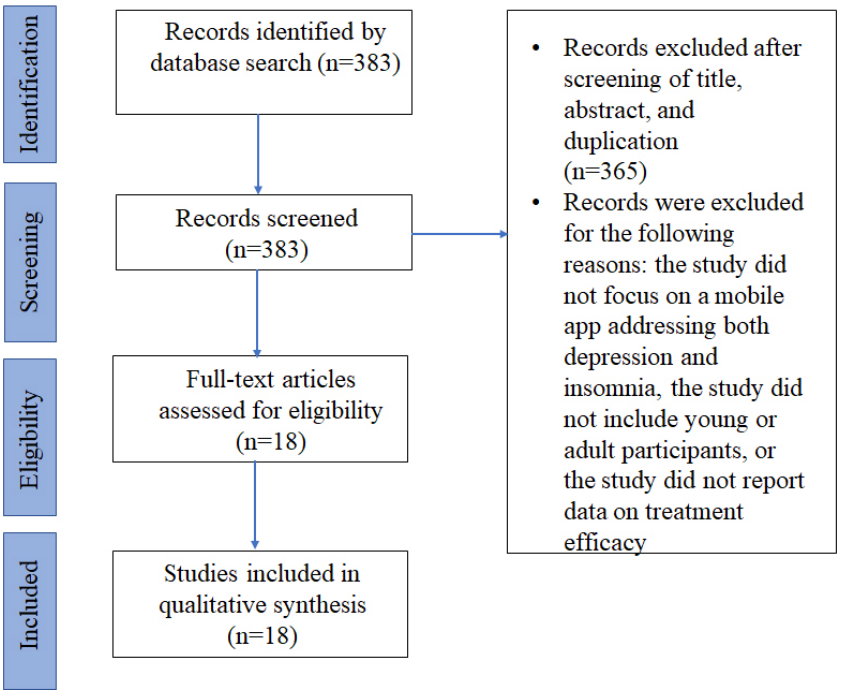

Study Characteristics

Table 1 presents the 18 identified apps, categorizing their primary functions in addressing insomnia and depression. This categorization enhances the understanding of the correlation between insomnia and depression, offering a comprehensive perspective [23,24]. Chatbots and smartphone apps are commonly classified as mobile apps. We explored the shared factors and therapeutic approaches within these mobile apps that contribute to improving sleep and mental health.

XSL•FO
RenderX

Table 1. Study characteristics regarding insomnia and depression chatbots in the review (N=18).

Design and app	Study	Aim	Total sample (age)
Cognitive behavioral interventions (n=10)			
Woebot	[5]	Assessed the feasibility of implementing an app-based cognitive behavioral therapy intervention for adolescents undergoing treatment for depression. This investigation focused on the intervention's acceptability, practicality, and safety.	N=18 (13-17 years)
Sleep Ninja	[41]	Aimed to develop a sleep app that offers personalized sleep information, allows flexible completion of sleep diaries, incorporates game-like features, and has an engaging layout. Participants expressed a preference for a trusted developer, offline accessibility, and free access to the app.	N=21 (12-16 years)
MIND MORE	[42]	Explored the relationships among subjective sleep quality, memory complaints, and depressive symptoms. Evaluated the usability of MIND MORE among older adults (urban Koreans) and analyzed changes in sleep quality during a 1-week intervention. Participants assessed their adherence to the app.	N=41 (≥60 years)
WeChat Mini Program	[43]	Assessed the efficacy of cognitive behavioral therapy in preventing the progression from episodic to persistent insomnia. Supported online self-guided sleep interventions for individuals with persistent insomnia.	N=200 (≥18 years)
CBT-I Coach	[44]	Examined the feasibility, acceptability, and effects on adherence and sleep outcomes of using the CBT-I Coach app for cognitive behavioral therapy for insomnia. Patients who utilized the app as intended reported high levels of acceptability.	N=18 (46-50 years)
Emohaa	[45]	Evaluated the effectiveness of Emohaa, a cognitive behavioral therapy–based conversational agent, in alleviating symptoms of mental distress. The results indicated significant improvements in mental distress, particularly among individuals experiencing depression and insomnia.	N=301 (20-30 years)
XiaoE	[46]	Investigated the impact of XiaoE, a cognitive behavioral therapy–based mental health chatbot, on young adults with depression during the COVID-19 pandemic. The chatbot offered easily accessible and self-directed mental health support.	N=148 (17-34 years)
Kokoro-App	[47]	Accessed via the app, the cognitive behavioral therapy program, connected to a remote server, effectively alleviated depression. It gathered information on participant utilization to enhance the program and mobile health features.	N=164 (40 years)
Sleep Ninja	[48]	Utilizes cognitive behavioral therapy for insomnia in adolescents through a chatbot format that provides personalized information. It offers psychoeducational content, sleep hygiene tips, and sleep-focused cognitive therapy to improve sleep.	N=47(12-16 years)
ExpiWell	[49]	Assessed the feasibility and acceptability of smartphone-based cognitive behavioral therapy interventions for sleep disorders. The study provides a low-intensity, accessible intervention, contributing to the more regular delivery of sleep interventions.	N=14 (22-57 years)
Cognitive and emotional support (n=4)			
Autonomous AI[a] chatbot	[50]	Assists individuals with insomnia by utilizing AI and deep learning to enhance interactions and provide cognitive and emotional support. A friendly chatbot was developed to aid in understanding and treating insomnia.	Not mentioned
CARO	[51]	Proposes CARO, a chatbot designed to provide empathetic conversations and medical advice for individuals with major depression. It senses context, intent, and emotions to facilitate personalized interactions.	Not mentioned
Tess	[52]	Tess is a feasible option for providing emotional support but is not intended to replace trained therapists. Results indicate that AI has the potential to reduce symptoms of depression and anxiety through cognitive behavioral therapy–based conversations.	N=74 (22 years)

XSL•FO
RenderX

Design and app	Study	Aim	Total sample (age)
KANOPEE	[53]	Developed KANOPEE, a smartphone app featuring a virtual agent for autonomous screening and digital behavioral interventions. This study demonstrates the feasibility of using virtual agents and highlights the potential of empathetic human-machine interfaces in eHealth solutions.	N=47 (18-65 years)
Examining symptoms or interventions for depression (n=4)			
Tess	[54]	Assessed the intervention's effects on depression and anxiety symptoms in university students using the chatbot Tess. The pilot trial indicated a decrease in anxiety symptoms among the experimental group.	N=181 (18-33 years)
DEPRA	[55]	An AI-based chatbot designed for the early detection of depression. Integration with social media attracted over 40 participants, and user surveys indicated high levels of satisfaction.	N>40 (18-33 years)
XiaoNan	[56]	Explored the efficacy of chatbot therapy for depression among university students, aiming to provide a convenient, affordable, and interactive self-help psychotherapy approach.	N=83 (19-28 years)
DPs	[57]	Proposed a personalized design for conversational agents aimed at treating depression, incorporating transparency and autonomy as key design principles based on interviews with individuals diagnosed with depression.	N=15 (14-17 years)

[a]AI: artificial intelligence.

Research Topics

Table 1 shows that the mobile apps discussed in the context of insomnia and depression can be broadly categorized into 3 major areas: cognitive behavioral interventions, cognitive and emotional support, and the examination of symptoms or interventions for depression.

A comprehensive research synthesis revealed that studies focusing on "Cognitive Behavioral Interventions" are particularly prominent, totaling 10 [5,41-49]. This underscores researchers' concentrated emphasis on cognitive behavioral interventions in mobile apps addressing insomnia and depression. Additionally, 4 studies [50-53] focused on "Cognitive and Emotional Support," highlighting the importance of mobile apps in providing cognitive and emotional assistance. Furthermore, 4 studies [54-57] were dedicated to "Examining Symptoms or Interventions for Depression," underscoring a sustained interest in apps aimed at the thorough exploration and management of depressive symptoms.

Study Participants

The target audience for these apps included young people (aged 13-25 years; n=12) [5,41,43,45,46,48,49,52,54-57] and adults (aged 18 years and older; n=11) [42-47,52-56] (Table 1).

The primary purpose of mobile apps is to provide personalized support [5,50-52,54,56,57]. In addressing stress, depression, and insomnia, these chatbots offer effective tools and resources to help users improve their mental well-being. They deliver tailored advice, CBT exercises [5,41,43,44,46-50,52,56], and emotional support [45,46,53].

This suggests that the design and development of mental health apps should be tailored to the specific needs and characteristics of different age groups to more effectively address mental health challenges across diverse age ranges. By focusing on both young individuals (13-25 years) and adults (18 years and older), these apps aim to provide targeted, relevant, and personalized mental health support and therapy. This approach contributes to innovation and advancement in the mental health field.

App Functions

The main focus was on early detection, assessment, and screening (n=5) [42,51,54,55,57], counseling and psychological support (n=3) [45,46,53], and CBT (n=11) [5,41,43,44,46-50,52,56]. These apps demonstrate an active exploration of mental health issues, addressing a range of interventions from early detection to personalized psychological support and the provision of CBT.

Primarily adopting CBT-based psychological health chatbots, these apps emphasize feasibility, acceptability, practicality, and safety in their research. They offer diverse solutions that advance the field of mental health technology by prioritizing scientifically grounded approaches. Additionally, they focus on the practical feasibility and acceptability of real-world apps with the potential to achieve comprehensive, profound effects while promoting mental well-being.

Therapeutic Efficacy

Mobile apps targeting both depression and insomnia have been extensively researched for their therapeutic efficacy. Numerous studies support their positive impact on mental health, demonstrating their effectiveness in addressing both conditions simultaneously. Depression and insomnia often co-occur, exacerbating one another in a vicious cycle. Numerous studies [41-44,46-52,55] have shown that AI-powered chatbots in these apps provide personalized support, including sleep guidance, feedback, and adjustment recommendations, which effectively alleviate insomnia symptoms while positively impacting depression. By offering tools and strategies grounded in CBT, these apps enhance users' self-regulation skills, helping them cope with negative emotions and stress, thereby reducing the impact of depression.

XSL•FO
RenderX

Furthermore, the emotional support systems integrated into these apps not only promote mental well-being but also empower patients by enhancing their sense of self-efficacy. This enables them to better manage psychological challenges, leading to significant improvements in their quality of life [58]. With technological advancements, mental health apps are showing increasingly diverse and effective therapeutic outcomes, offering more options and support to individuals dealing with both depression and insomnia. For example, the autonomous AI chatbot leverages AI and deep learning to deliver more interactive cognitive and emotional support therapy, with a focus on understanding and treating insomnia. ExpiWell evaluates the feasibility and acceptability of smartphone-based CBT interventions for sleep disorders. MIND MORE explores the relationship between subjective sleep quality and depressive symptoms, assessing its usability among older age Koreans. Emohaa, a CBT-based app, has made significant progress in alleviating symptoms of psychological distress. Woebot examines the feasibility of implementing app-based CBT interventions for adolescents receiving care for depression.

Mobile App Characteristics

These apps operate across various platforms, including both Android and iOS (n=8) [41-44,48,52-54]. Some are limited to Android only (n=3) [5,46,57], while 1 app supports multiple platforms (n=1) [46]. Additionally, 5 apps did not specify their operating platforms (n=5) [45,47,49-51] (Table 2).

Of these apps, 8 focused on delivering intelligent and personalized mental health support through AI and deep learning techniques. Among the insomnia-related apps, only 2 chatbots, Emohaa and MIND MORE, specifically targeted the connection between depression and insomnia [42,45], with an emphasis on addressing their interrelation. These apps offer effective psychological interventions and guide users in enhancing both sleep quality and mental health. By utilizing AI, deep learning techniques, and personalized processing, they can provide customized advice, techniques, and CBT for addressing depression and insomnia. These chatbots are designed to deliver convenient and accessible support, enabling users to actively engage in and manage their mental health, ultimately leading to positive changes in their daily lives (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of chatbots regarding insomnia and depression in the review (n=16).

Targeted area and app name	Description	Main feature	Operating system	Available for download?
Insomnia				
Autonomous AI chatbot	Based on AI[a] and deep learning, an AI chatbot for people with insomnia that includes natural language processing principles, such as stemming and a bag-of-words. Training data are provided by the app.	The chatbot is designed as a human chat to reduce anxiety through pleasant conversations. The app includes (1) training data, (2) training data "bag of words," (3) feed-forward neural net, and (4) natural language processing.	Not mentioned	No
ExpiWell	The ExpiWell app has a prescheduled "My Sleep Programme" and a user-accessible "Resources" section. The content is tailored to participant responses, using branching logic.	The app consists of 6 weekly core modules and a user-initiated "Resources" section. Content with tailored feedback is provided with access to module summaries.	Not mentioned	No
Sleep Ninja	Sleep Ninja is a CBT[b] for insomnia–based smartphone app targeting various sleep difficulties in adolescents.	The app uses psychoeducation, stimulus control, sleep hygiene education, and sleep-focused cognitive therapy to improve sleep. Its features are based on user input and sleep profiles. Users complete 6 training sessions and finish with a black belt in sleep.	Android and iOS	Yes
WeChat Mini Program	eCBTI[c] will be delivered using the program via the WeChat app. eCBTI is efficient, flexible, and time-saving, and may improve mood and quality of life compared with face-to-face CBTI[d].	The app provides daily personalized support, encompassing behavioral and cognitive elements, along with relaxation audios and sleep hygiene education following CBTI guidelines.	Android and iOS	Yes
CBT-I Coach	The CBT-I Coach app was designed to supplement face-to-face CBT[e] therapy for insomnia and improve adherence to the protocol by completing CBT for insomnia tasks.	The app includes education on sleep processes, positive sleep routines, and environments. It features a sleep diary, tools, guided exercises, reminders for sleep habits, and alarms for prescribed bedtime and wake times.	Android and iOS	Yes
KANOPEE	KANOPEE is a smartphone app for insomnia; it uses a virtual agent for autonomous screening and interventions during COVID-19.	KANOPEE uses C# in Unity 3D with a virtual agent named Louise, whose natural voice and body movements were recorded using motion capture technologies. The interaction scenario adapts to the user's answers through decision trees.	Android and iOS	Yes
Depression				
XiaoNan	XiaoNan is a chatbot powered by open-source conversational AI RASA. Its text content is based on CBT principles approved by professional therapists.	XiaoNan accepts both text and voice messages. Voice messages are transcribed using a natural speech recognition service from the IFLYTEK open platform before being processed by the chatbot's natural language understanding module.	Windows, macOS, Android, and iOS	Yes
Woebot	Woebot uses a fully automated, relational conversational agent to deliver evidence-based therapeutic elements in brief conversations.	This self-guided intervention uses natural language processing and machine learning to incorporate elements of CBT, interpersonal psychotherapy for adolescents, and dialectical behavior therapy.	Android	Yes
XiaoE	XiaoE is a CBT-based chatbot, which interacts automatically with users through text, images, and voice, primarily for depression screening, prevention, and self-help.	Its chatbot dialog system is built on the open-source RASA framework. Utilizes natural language processing and deep learning techniques.	Android	Yes
DPs	The chatbot prototype explores personalized agents with various elements and degrees.	The system allows users to customize their name, gender, avatar, and social roles. It supports personalized modules for behavioral activation and sleep hygiene, with content selected based on user feedback.	Android	Yes

XSL•FO
RenderX

Targeted area and app name	Description	Main feature	Operating system	Available for download?
Tess	Tess is an AI chatbot for depression and anxiety. It supports therapy-based conversations and adjusts interventions based on user feedback.	Tess is an AI chatbot that supports CBT, emotion-focused therapy, and solution-focused brief therapy.	Android and iOS	Yes
DEPRA	The DEPRA chatbot supports early detection of depression.	DEPRA, developed using Dialogflow and Node.js, communicates with participants via Facebook Messenger. Data are stored in an AWS[f] database managed with a MySQL Workbench client.	Android and iOS	Yes
CARO	CARO is an empathetic chatbot app that provides medical advice for major depression.	This study used automated metrics, namely, the Bilingual Evaluation Understudy score and Bidirectional Encoder Representations from Transformers score, and the text is classified into conversations or medical advice.	Not mentioned	No
Kokoro-App	A smartphone CBT app that supports self-monitoring, behavioral activation, and cognitive restructuring skills.	The app has 4 parts: sessions, mind maps, actions, and thoughts. It allows users to enter details of situations, emotions, thoughts, and behaviors.	Not mentioned	No
Insomnia and depression				
Emohaa	A CBT-based conversational agent for reducing mental distress that has demonstrated significant improvements in depression and insomnia.	This conversational agent has 2 platforms: a template-based platform with predefined options and exercises based on CBT principles and a generative dialog platform that allows open-ended conversations and provides emotional support.	Not mentioned	No
MIND MORE	The app analyzed the relationships among sleep quality, memory complaints, depressive symptoms, and app usability and adherence. It is a multimodal CBT for insomnia app with sleep education components.	The main sleep hygiene education program has 3 sessions that include quizzes to evaluate user learning. It helps users understand the mechanisms sustaining insomnia and correct unhelpful sleep-related beliefs, anxiety, arousal, and sleep-disruptive habits. This improves adherence to the app's recommendations.	Android and iOS	Yes

[a]AI: artificial intelligence.

[b]CBT: cognitive behavioral therapy.

[c]eCBTI: electronic cognitive behavioral therapy for insomnia.

[d]CBTI: cognitive behavioral therapy for insomnia.

[e]CBT: cognitive behavioral therapy.

[f]AWS: Amazon Web Services.

The Feasibility and Acceptability of Mobile Apps in Sleep and Depression Interventions

Taylor et al [49] highlighted the feasibility, acceptability, and potential usefulness of smartphone-based CBT interventions delivered through apps aimed at treating sleep problems in individuals with psychosis, demonstrating that CBT is both feasible and acceptable.

In another study, Nicol et al [5] examined the feasibility of providing an app-based CBT intervention for adolescents aged 13-17 years with depression. The study assessed the feasibility, acceptability, practicality, and safety of using a CBT-based chatbot over a 12-week trial. The results indicated that the use of the CBT-based chatbot was feasible, acceptable, and safe for adolescents experiencing moderate depressive symptoms.

As sleep intervention tools, mobile apps not only help improve sleep but also record users' sleep data, provide sleep tracking and reporting, and enhance users' understanding of their sleep habits and progress [59]. Mobile apps have demonstrated the potential to improve sleep disturbances and offer more convenient and personalized support [7,28]. Similarly, these apps have shown strengths in depression interventions by providing emotional support, which can, in turn, help alleviate depression symptoms [60].

The Effectiveness of Mobile Apps as a Depression and Sleep Intervention Tool

Two studies [42,45] evaluated the effectiveness of mobile apps as alternative intervention tools for depression and sleep. Sabour and colleagues [45] confirmed that the demand for mental health support is increasing. Their study included 134 participants, and the results indicated that those using Emohaa experienced significant improvements in symptoms of mental distress. The inclusion of emotional support had a complementary effect on both depression and insomnia. Participants expressed high satisfaction with the Emohaa platform, viewing it as a practical and effective tool. These results offer additional options for individuals seeking emotional support.

XSL•FO
RenderX

Chung et al [42] examined the relationship between sleep quality, depressive symptoms, and the usability of a smartphone-based self-help CBT for an insomnia app called MIND MORE for older adults. They compared participants' sleep quality before and after a 1-week intervention period, along with an assessment of adherence to the app. During the study, 41 participants used the app and completed usability questionnaires both before and after the intervention. The results indicated significant improvements associated with the use of the MIND MORE app.

AI Chatbots' Potential and Applications

These apps showcase the diverse capabilities of AI in chatbots, incorporating technologies such as natural language processing (NLP), ideographic recognition, machine learning, and sentiment analysis [13,61]. These advancements hold the potential for chatbots to be utilized in health care, enabling efficient, equitable, and personalized care. They can assist clinicians in understanding patient conditions and provide decision aids [15,62,63]. Applied functional technologies for AI chatbots encompass a range of components and features that inform future designs and specific apps. Their features are as follows.

An autonomous AI chatbot incorporates various components, including training data, a bag-of-words, a feed-forward neural network, and NLP. These components enable effective understanding and response to user inputs [50].

Emohaa is a generative dialog platform designed to facilitate open-ended conversations and provide emotional support. Users can engage in discussions with a chatbot, allowing them to freely express their thoughts and feelings [45].

XiaoNan is a natural speech-recognition service that utilizes the "IFLYTEK Open Platform" to convert spoken language into text. The chatbot's natural language understanding module processes this text to effectively comprehend user queries [56].

Woebot utilizes both NLP and machine learning techniques, incorporating elements of CBT, interpersonal psychotherapy for adolescents, and dialectical behavior therapy. By personalizing conversations, Woebot aids adolescents in developing emotion regulation skills for their daily lives [5].

XiaoE is a dialog system built on the open-source RASA framework. By utilizing NLP and deep learning techniques, it provides personalized modules for behavioral activation and sleep hygiene, enabling users to improve their daily routines and sleep patterns [46].

Tess provides therapy-based conversations developed by mental health experts, offering individuals professional support and guidance to manage their mental health [52].

DEPRA utilizes Dialogflow and Node.js to communicate with participants via Facebook Messenger. By leveraging Facebook's AI Empathetic Dialogue and a web-scraped Medical Q&A data set, the data are stored in an AWS database managed by a MySQL Workbench client. AWS Educate is used to save all interaction records [55].

CARO utilizes Facebook's AI Empathetic Dialogue and a web-scraped Medical Q&A data set. It uses automated metrics, such as the Bilingual Evaluation Understudy and Bidirectional Encoder Representations from Transformers scores, to assess text, which is classified into conversational or medical advice. Two long short-term memory–based models were trained to generate empathetic responses and provide medical answers [51].

Discussion

Principal Findings

We reviewed articles published between 2017 and 2023 and found that mobile apps related to insomnia and depression primarily focus on cognitive behavioral interventions. Additionally, the review revealed 8 apps that emphasize the use of AI and deep learning technologies to provide intelligent and personalized mental health support. Directions for future research are discussed in the following sections.

AI and Deep Learning for Personalized Mental Health Support With Intelligence

This systematic review investigated mobile apps, categorized the chatbot-themed features of insomnia and depression apps, and explored the correlation between insomnia and depression. It aims to provide comprehensive insights into the common features and therapeutic approaches of chatbot apps, thereby contributing to the improvement of sleep and mental health.

A total of 9 reviewed apps focused on utilizing AI and deep learning technologies to provide intelligent and personalized mental health support [5,45,46,49-51,54-56]. These apps emphasized early detection, assessment, screening, counseling, psychological support, and CBT. CBT-based chatbots record users' sleep data, provide sleep tracking and reporting, and enable users to better understand their sleep habits and progress. They also offer effective psychological interventions and guidance to improve sleep quality and mental health, along with customized advice and techniques. The goal is to make mental health support readily available and convenient for users. By proactively engaging in and managing their mental health, individuals can achieve positive changes and enhance their daily lives.

Specifically, an autonomous AI chatbot integrates various components, including training data, bag-of-words training data, feedforward neural networks, and NLP, to effectively understand and respond to user input [50]. For example, Emohaa is a generative dialog platform that facilitates open-ended conversations and offers emotional support, enabling users to freely express their thoughts and emotions [45]. XiaoNan, part of the "IFLYTEK Open Platform," converts spoken language into text, which is then processed by the chatbot's NLP module to comprehend user queries [56]. Woebot combines NLP and machine learning techniques to integrate CBT, interpersonal psychotherapy for adolescents, and dialectical behavior therapy, helping adolescents develop emotion-regulation skills through customized conversations [5]. XiaoE, built using RASA, utilizes NLP and deep learning to create personalized behavioral activation and sleep hygiene modules aimed at improving users' routines and sleep patterns [46]. Tess offers therapy-based conversations to provide professional mental health support and

XSL•FO
RenderX

guidance [52]. DEPRA uses Dialogflow and Node.js, leveraging Facebook AI Empathetic Dialogue and a web-scraped Medical Q&A data set, which is stored in an AWS database managed by MySQL Workbench. AWS Educate is used to save the interaction records [55]. CARO uses Facebook AI Empathetic Dialogue along with a web-scraped Medical Q&A data set to evaluate text using Bilingual Evaluation Understudy and Bidirectional Encoder Representations from Transformers scores [51]. It provides empathetic responses and medical advice through 2 long short-term memory–based models. The diversity of these chatbots enriches the choices available within apps, catering to the varied needs of different users.

Mobile Apps Offer the Potential to Improve Insomnia in Depression Support

Mobile apps hold significant potential for addressing insomnia and depression. The primary focus has been on early detection, assessment, and screening (n=5) [42,51,54,55,57]; counseling and psychological support (n=3) [45,46,53]; and CBT (n=11) [5,41,43,44,46-50,52,56]. These apps represent a proactive exploration of mental health issues, providing personalized psychological support and CBT from the early detection stage. This comprehensive approach addresses the issues of depression and insomnia by offering users multilayered support. This includes screening and assessment features, real-time counseling, psychological support, and therapy grounded in CBT.

The focal point of these studies has been the diversity of apps within the mental health field, encompassing everything from early intervention to specific therapeutic methods. By leveraging AI and deep learning technologies, these apps can comprehensively understand users' needs and provide tailored solutions. The application of CBT is particularly noteworthy, reflecting the ongoing exploration of therapeutic methods and the pursuit of innovative technological apps.

Overall, these apps have demonstrated significant potential to improve mental health by offering users convenient and personalized support. However, as this field continues to evolve, several directions for future research warrant further exploration. Enhancing the effectiveness of these apps, improving the accuracy of screening and assessment, and more effectively integrating psychological support with traditional therapeutic approaches are all valuable avenues for future investigation.

Limitations

This review investigates chatbot apps within mental health–related mobile apps, focusing on the use of AI and deep learning technologies. During the screening process, researchers examined various aspects, including research topics, participant characteristics, app features, and therapeutic outcomes, resulting in the inclusion of 18 articles. However, this review also has certain limitations. First, subjectivity in the screening process may have been influenced by the researchers' personal judgments. Second, the inclusion of only 18 articles raises questions about whether this selection adequately represents the diversity and comprehensiveness of the entire field, as some relevant studies might have been overlooked. Additionally, it is worth noting that research on the relationship between depression and sleep is relatively scarce, indicating that this area presents a promising opportunity for further investigation. We hope our review serves as a foundation for research in this field and stimulates further academic discussion. While AI and deep learning are evolving rapidly, the long-term effects of their apps remain unclear, warranting additional research. This review offers a preliminary understanding of the use of chatbots in mental health–related mobile apps. However, readers should exercise caution in considering these potential limitations.

This study identified several risks associated with mobile apps targeting insomnia and depression, including the lack of well-defined clinical designs and usage guidelines, insufficient privacy protection, inadequate monitoring to prevent harm from inappropriate user interactions, and challenges in integrating these apps with existing medical models. These issues can result in misinterpretation and potentially ineffective or harmful interventions. In addition, chatbots may lack crisis warning mechanisms and appropriate privacy safeguards [64,65]. Although mobile apps targeting insomnia and depression have shown promise in providing suitable treatment, there is a need for stronger and more robust evidence to comprehensively support their effectiveness and acceptability. Further research is necessary to address these issues and enhance the efficacy of insomnia-related interventions within mobile apps for mental health. In particular, integrating chatbots into existing health care models faces challenges due to potential misconceptions, which could lead to ineffective or harmful interventions. While advancements have been made in the field of mental health, additional high-quality research is necessary to confirm the effectiveness and acceptability of mobile apps in this area. Further studies will help address these challenges and enhance the effectiveness and reliability of mobile apps targeting depression and insomnia symptoms.

Future Prospects

According to our study, agent prompts that integrate human intelligence with deep learning offer intelligent and personalized cognitive and emotional support. This creates numerous opportunities for researchers to combine NLP with human intelligence in developing chatbots specifically designed to address mental health issues. These mental health chatbots can serve as effective sleep intervention tools by providing assistance, recording users' sleep data, and offering sleep tracking and reporting capabilities, thereby enabling individuals with depression to gain a better understanding of their sleep habits and progress. When combined with recommendations from experienced therapists to establish guidelines, these chatbots can be more effective in alleviating the severity of insomnia associated with depression, particularly because long-term interventions have proven to be more effective in improving insomnia severity.

Conclusions

After synthesizing critical information from previous literature on research topics, participants, app functionalities, therapeutic effects, operating platforms, technical foundations, and functional features, as well as feasibility and acceptability, this review offers a systematic summary of the application and effectiveness of mobile apps in the field of mental health. This

XSL•FO
RenderX

will serve as a guiding framework for related studies and future developments. Furthermore, by analyzing app technologies—particularly, the use of AI and deep learning techniques to provide personalized mental health support—mobile apps have shown significant potential in addressing insomnia in individuals with depression. Our research outlines specific directions for future exploration in this domain and contributes to the existing literature.

Currently, AI is unable to fully replace mental health treatment, primarily due to its limitations in replicating the comprehensive skill set of psychiatrists and psychotherapists. Although AI can utilize patient information for diagnosis and prognostic predictions, integrating the diverse layers of biological, psychological, and social data needed to formulate a holistic treatment plan remains a challenge. AI struggles to perceive nuanced emotions, exhibit empathic responses, interpret relational dynamics, and observe subtle behavioral and emotional changes. While AI cannot fully replace existing mental health treatments, integrating AI technology with human therapeutic approaches holds the potential to enhance treatment outcomes, foster innovation, and drive advancements in mental health.

Acknowledgments

This study was supported in part by Wan-Fang Hospital, Taipei Medical University, Taipei, Taiwan, under contract number 112-wf-ntut-04. This study was also supported by the Ministry of Science and Technology of Taiwan (grant MOST 109-2511-H-216-001-MY3 to LCC) and the National Taipei University of Technology and Wan-Fang Hospital Joint Research Program (grant NTUT-WFTMU-112-04). We deeply appreciate the financial support from these institutions, which contributed to the successful completion of this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 32 KB - jmir_v26i1e51110_app1.docx]

Multimedia Appendix 2
Methodological protocol amendments and clarifications log.
[DOCX File , 37 KB - jmir_v26i1e51110_app2.docx]

Multimedia Appendix 3
Indexed and keyword terms.
[DOCX File , 17 KB - jmir_v26i1e51110_app3.docx]

References

1. Choi YH, Yang KI, Yun C, Kim W, Heo K, Chu MK. Impact of insomnia symptoms on the clinical presentation of depressive symptoms: a cross-sectional population study. Front Neurol 2021 Aug 9;12:716097 [FREE Full text] [doi: 10.3389/fneur.2021.716097] [Medline: 34434165]
2. Li L, Wu C, Gan Y, Qu X, Lu Z. Insomnia and the risk of depression: a meta-analysis of prospective cohort studies. BMC Psychiatry 2016 Nov 05;16(1):375-316 [FREE Full text] [doi: 10.1186/s12888-016-1075-3] [Medline: 27816065]
3. Buysse D. J.., Prevalence, course, and comorbidity of insomnia and depression in young adults. Sleep 2008;31(4):473-480. [doi: 10.1093/sleep/31.4.473]
4. Gebara MA, Siripong N, DiNapoli EA, Maree RD, Germain A, Reynolds CF, et al. Effect of insomnia treatments on depression: a systematic review and meta-analysis. Depress Anxiety 2018 May 21;35(8):717-731. [doi: 10.1002/da.22776]
5. Nicol G, Wang R, Graham S, Dodd S, Garbutt J. Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: feasibility and acceptability study. JMIR Form Res 2022 Nov 22;6(11):e40242 [FREE Full text] [doi: 10.2196/40242] [Medline: 36413390]
6. Jain P. Psychiatric chatbot for COVID-19 using machine learning approaches. In: Machine Vision for Industry 4.0. Boca Raton, FL: CRC Press; 2022:169-194.
7. Bhirud N, Tataale S, Randive S, Nahar S. A literature review on chatbots in healthcare domain. Int J Sci Tech Res 2019;8(7):225-231 [FREE Full text]
8. Albalawi U, Mustafa M. Current artificial intelligence (AI) techniques, challenges, and approaches in controlling and fighting COVID-19: a review. Int J Environ Res Public Health 2022 May 12;19(10):5901. [doi: 10.3390/ijerph19105901] [Medline: 35627437]

XSL•FO
RenderX

9. Ahmed A, Aziz S, Khalifa M, Shah U, Hassan A, Abd-Alrazaq A, et al. Thematic analysis on user reviews for depression and anxiety chatbot apps: machine learning approach. JMIR Form Res 2022 Mar 11;6(3):e27654 [FREE Full text] [doi: 10.2196/27654] [Medline: 35275069]
10. Sweeney C, Potts C, Ennis E, Bond R, Mulvenna MD, O'neill S, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthcare 2021 Jul 15;2(3):1-15. [doi: 10.1145/3453175]
11. Gieselmann A, Podleschka C, Rozental A, Pietrowsky R. Communication formats and their impact on patient perception and working mechanisms: a mixed-methods study of chat-based vs. face-to-face psychotherapy for insomnia. Behav Ther 2021 Mar;52(2):430-441. [doi: 10.1016/j.beth.2020.06.001] [Medline: 33622511]
12. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Feb 12;49(09):1426-1448. [doi: 10.1017/s0033291719000151]
13. Suta P, Lan X, Wu B, Mongkolnam P, Chan JH. An overview of machine learning in chatbots. IJMERR 2020;9(4):502-510. [doi: 10.18178/ijmerr.9.4.502-510]
14. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021 Nov 29;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
15. King MR. The future of AI in medicine: a perspective from a chatbot. Ann Biomed Eng 2023 Feb 26;51(2):291-295. [doi: 10.1007/s10439-022-03121-w] [Medline: 36572824]
16. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q 2022 Feb 25;93(1):249-253. [doi: 10.1007/s11126-022-09973-8]
17. Ray A, Bhardwaj A, Malik YK, Singh S, Gupta R. Artificial intelligence and psychiatry: an overview. Asian J Psychiatr 2022 Apr;70:103021 [FREE Full text] [doi: 10.1016/j.ajp.2022.103021] [Medline: 35219978]
18. Gennatas ED, Chen JH. Artificial Intelligence in Medicine: Past, Present, and Future. Amsterdam, The Netherlands: Elsevier; 2021:3-18.
19. Gupta NS, Kumar P. Perspective of artificial intelligence in healthcare data management: a journey towards precision medicine. Computers in Biology and Medicine 2023 Aug;162:107051. [doi: 10.1016/j.compbiomed.2023.107051] [Medline: 37271113]
20. MacMillan F. Conducting a systematic review: a practical guide. In: Handbook of Research Methods in Health Social Sciences. Singapore City, Singapore: Springer; 2019:805-817.
21. Schork NJ. Artificial intelligence and personalized medicine. In: Precision Medicine in Cancer Therapy. Cham, Switzerland: Springer; 2019:265-283.
22. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. J Med Internet Res 2017 Nov 28;19(11):e399 [FREE Full text] [doi: 10.2196/jmir.8598] [Medline: 29187342]
23. Shin JC, Kim J, Grigsby-Toussaint D. Mobile phone interventions for sleep disorders and sleep quality: systematic review. JMIR Mhealth Uhealth 2017 Sep 07;5(9):e131 [FREE Full text] [doi: 10.2196/mhealth.7244] [Medline: 28882808]
24. Choi YK, Demiris G, Lin S, Iribarren SJ, Landis CA, Thompson HJ, et al. Smartphone applications to support sleep self-management: review and evaluation. J Clin Sleep Med 2018 Oct 15;14(10):1783-1790 [FREE Full text] [doi: 10.5664/jcsm.7396] [Medline: 30353814]
25. Yu J, Kuhn E, Miller KE, Taylor K. Smartphone apps for insomnia: examining existing apps' usability and adherence to evidence-based principles for insomnia management. Transl Behav Med 2019 Jan 01;9(1):110-119. [doi: 10.1093/tbm/iby014] [Medline: 30590862]
26. Guillodo E, Lemey C, Simonnet M, Walter M, Baca-García E, Masetti V, HUGOPSY Network, et al. Clinical applications of mobile health wearable-based sleep monitoring: systematic review. JMIR Mhealth Uhealth 2020 Apr 01;8(4):e10733 [FREE Full text] [doi: 10.2196/10733] [Medline: 32234707]
27. Arroyo AC, Zawadzki MJ. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. JMIR Mhealth Uhealth 2022 Apr 04;10(4):e33527 [FREE Full text] [doi: 10.2196/33527] [Medline: 35377327]
28. Denecke K, Schmid N, Nüssli S. Implementation of cognitive behavioral therapy in e-mental health apps: literature review. J Med Internet Res 2022 Mar 10;24(3):e27791 [FREE Full text] [doi: 10.2196/27791] [Medline: 35266875]
29. Deng W, M J J van der Kleij R, Shen H, Wei J, Brakema EA, Guldemond N, et al. eHealth-based psychosocial interventions for adults with insomnia: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Mar 14;25:e39250 [FREE Full text] [doi: 10.2196/39250] [Medline: 36917145]
30. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
31. Lau Y, Htun TP, Wong SN, Tam WSW, Klainin-Yobas P. Therapist-supported internet-based cognitive behavior therapy for stress, anxiety, and depressive symptoms among postpartum women: a systematic review and meta-analysis. J Med Internet Res 2017 Apr 28;19(4):e138 [FREE Full text] [doi: 10.2196/jmir.6712] [Medline: 28455276]

XSL•FO
RenderX

32. Stawarz K, Preist C, Tallon D, Wiles N, Coyle D. User experience of cognitive behavioral therapy apps for depression: an analysis of app functionality and user reviews. J Med Internet Res 2018 Jun 06;20(6):e10120 [FREE Full text] [doi: 10.2196/10120] [Medline: 29875087]
33. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
34. Teles A. Mobile mental health: a review of applications for depression assistance. New York, NY: IEEE; 2019 Presented at: IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); June 5-7, 2019; Cordoba, Spain p. 708-713. [doi: 10.1109/cbms.2019.00143]
35. Wasil AR, Venturo-Conerly KE, Shingleton RM, Weisz JR. A review of popular smartphone apps for depression and anxiety: assessing the inclusion of evidence-based content. Behav Res Ther 2019 Dec;123:103498. [doi: 10.1016/j.brat.2019.103498] [Medline: 31707224]
36. Khademian F, Aslani A, Bastani P. The effects of mobile apps on stress, anxiety, and depression: overview of systematic reviews. Int J Technol Assess Health Care 2020 Dec 14;37(1):e4. [doi: 10.1017/s0266462320002093]
37. McCall HC, Hadjistavropoulos HD, Sundström CRF. Exploring the role of persuasive design in unguided internet-delivered cognitive behavioral therapy for depression and anxiety among adults: systematic review, meta-analysis, and meta-regression. J Med Internet Res 2021 Apr 29;23(4):e26939 [FREE Full text] [doi: 10.2196/26939] [Medline: 33913811]
38. Molloy A, Anderson PL. Engagement with mobile health interventions for depression: a systematic review. Internet Interv 2021 Dec;26:100454 [FREE Full text] [doi: 10.1016/j.invent.2021.100454] [Medline: 34621626]
39. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
40. Higgins J, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
41. Werner-Seidler A, O'Dea B, Shand F, Johnston L, Frayne A, Fogarty AS, et al. A smartphone app for adolescents with sleep disturbance: development of the Sleep Ninja. JMIR Ment Health 2017 Jul 28;4(3):e28 [FREE Full text] [doi: 10.2196/mental.7614] [Medline: 28754651]
42. Chung K, Kim S, Lee E, Park JY. Mobile app use for insomnia self-management in urban community-dwelling older Korean adults: retrospective intervention study. JMIR Mhealth Uhealth 2020 Aug 24;8(8):e17755 [FREE Full text] [doi: 10.2196/17755] [Medline: 32831177]
43. Yang Y, Luo X, Paudel D, Zhang J, Li SX, Zhang B. Effects of e-aid cognitive behavioural therapy for insomnia (eCBTI) to prevent the transition from acute insomnia to chronic insomnia: study protocol for a randomised controlled trial. BMJ Open 2019 Nov 18;9(11):e033457 [FREE Full text] [doi: 10.1136/bmjopen-2019-033457] [Medline: 31740476]
44. Koffel E, Kuhn E, Petsoulis N, Erbes CR, Anders S, Hoffman JE, et al. A randomized controlled pilot study of CBT-I Coach: feasibility, acceptability, and potential impact of a mobile phone application for patients in cognitive behavioral therapy for insomnia. Health Informatics J 2018 Mar 27;24(1):3-13 [FREE Full text] [doi: 10.1177/1460458216656472] [Medline: 27354394]
45. Sabour S, Zhang W, Xiao X, Zhang Y, Zheng Y, Wen J, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023 May 4;5:1133987 [FREE Full text] [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
46. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719 [FREE Full text] [doi: 10.2196/40719] [Medline: 36355633]
47. Furukawa TA, Horikoshi M, Fujita H, Tsujino N, Jinnin R, Kako Y, et al. Cognitive and behavioral skills exercises completed by patients with major depression during smartphone cognitive behavioral therapy: secondary analysis of a randomized controlled trial. JMIR Ment Health 2018 Jan 11;5(1):e4 [FREE Full text] [doi: 10.2196/mental.9092] [Medline: 29326098]
48. Werner-Seidler A, Wong Q, Johnston L, O'Dea B, Torok M, Christensen H. Pilot evaluation of the Sleep Ninja: a smartphone application for adolescent insomnia symptoms. BMJ Open 2019 May 27;9(5):e026502. [doi: 10.1136/bmjopen-2018-026502] [Medline: 31133584]
49. Taylor KM, Bradley J, Cella M. A novel smartphone-based intervention targeting sleep difficulties in individuals experiencing psychosis: a feasibility and acceptability evaluation. Psychol Psychother 2022 Sep 28;95(3):717-737 [FREE Full text] [doi: 10.1111/papt.12395] [Medline: 35481697]
50. Shaikh T. Autonomous AI chatbot therapy for patient with insomnia. New York, NY: IEEE; 2022 Presented at: 2022 IEEE 7th International Conference for Convergence in Technology (I2CT); April 7-9, 2022; Mumbai, Maharashtra, India p. 1-5. [doi: 10.1109/i2ct54291.2022.9825008]
51. Harilal N, Shah R, Sharma S, Bhutani V. CARO: an empathetic health conversational chatbot for people with major depression. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. New York, NY: Association for Computing

XSL•FO
RenderX

Machinery; 2020 Presented at: 7th ACM IKDD CoDS and 25th COMAD; January 5-7, 2020; Hyderabad, Telangana, India p. 349-350. [doi: 10.1145/3371158.3371220]
52. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64 [FREE Full text] [doi: 10.2196/mental.9782] [Medline: 30545815]
53. Philip, Dupuy L, Morin CM, de Sevin E, Bioulac S, Taillard J, et al. Smartphone-based virtual agents to help individuals with sleep concerns during COVID-19 confinement: feasibility study. J Med Internet Res 2020 Dec 18;22(12):e24268 [FREE Full text] [doi: 10.2196/24268] [Medline: 33264099]
54. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021 Aug 12;5(8):e20678. [doi: 10.2196/20678] [Medline: 34092548]
55. Kaywan P. DEPRA: an early depression detection analysis chatbot. Cham, Switzerland: Springer; 2021 Oct 25 Presented at: 10th International Conference: Health Information Science (HIS 2021); October 25–28, 2021; Melbourne, VIC, Australia p. 193-204. [doi: 10.1007/978-3-030-90885-0_18]
56. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022 Mar;27:100495 [FREE Full text] [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
57. Kuhlmeier F, Gnewuch U, Lüttke S, Brakemeier EL, Mädche A. A personalized conversational agent to treat depression in youth and young adults–a transdisciplinary design science research project. Cham, Switzerland: Springer; 2022 Jun 1 Presented at: 17th International Conference on Design Science Research in Information Systems and Technology (DESRIST 2022); June 1-3, 2022; St Petersburg, FL, USA. [doi: 10.1007/978-3-031-06516-3_3]
58. Lim SM, Shiau CWC, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
59. Aarts T, Markopoulos P, Giling L, Vacaretu T. Snoozy: a chatbot-based sleep diary for children aged eight to twelve. 2022 Presented at: IDC '22: Interaction Design and Children; June 27–30, 2022; Braga, Portugal URL: https://tinyurl.com/r4d68mn9 [doi: 10.1145/3501712.3529718]
60. Duvvuri V. Predicting depression symptoms from Discord chat messaging using AI medical chatbots. In: ICMLSC '22: Proceedings of the 2022 6th International Conference on Machine Learning and Soft Computing. New York, NY: ACM; 2022 Presented at: ICMLSC '22: The 6th International Conference on Machine Learning and Soft Computing; January 15-17, 2022; Haikou, China p. 111-119. [doi: 10.1145/3523150.3523168]
61. Adamopoulou E, Moussiades L. An overview of chatbot technology. Cham, Switzerland: Springer; 2020 Presented at: 16th IFIP WG 12.5 International Conference, AIAI 2020; June 5–7, 2020; Neos Marmaras, Greece p. 373-383. [doi: 10.1007/978-3-030-49186-4_31]
62. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
63. Chew HSJ. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022 Apr 13;10(4):e32578. [doi: 10.2196/32578] [Medline: 35416791]
64. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/nejmsr2214184]
65. Moldt J, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec 28;28(1):2182659 [FREE Full text] [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 21.07.23; peer-reviewed by M Sakata, CN Hang; comments to author 08.11.23; revised version received 01.01.24; accepted 22.09.24; published 18.10.24.

Please cite as:
Chiu YH, Lee YF, Lin HL, Cheng LC
Exploring the Role of Mobile Apps for Insomnia in Depression: Systematic Review
J Med Internet Res 2024;26:e51110
URL: https://www.jmir.org/2024/1/e51110
doi:10.2196/51110
PMID:39423009

©Yi-Hang Chiu, Yen-Fen Lee, Huang-Li Lin, Li-Chen Cheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Technology-Supported Physical Activity and Its Potential as a Tool to Promote Young Women's Physical Activity and Physical Literacy: Systematic Review

Kimberley Watson-Mackie[1]; Lauren Arundell[2], PhD; Natalie Lander[2], PhD; Fiona H McKay[3], PhD; Alethea Jerebine[1], PhD; Fotini Venetsanou[4], PhD; Lisa M Barnett[5], PhD

[1]School of Health and Social Development, Faculty of Health, Deakin University, Burwood, Australia
[2]Institute for Physical Activity and Nutrition, Faculty of Health, School of Exercise and Nutrition Sciences, Deakin University, Burwood, Australia
[3]Institute for Health Transformation, Faculty of Health, School of Health and Social Development, Deakin University, Burwood, Australia
[4]School of Physical Education and Sport Science, National and Kapodistrian University of Athens, Athens, Greece
[5]Institute for Physical Activity and Nutrition, Faculty of Health, School of Health and Social Development, Deakin University, Burwood, Australia

Corresponding Author:
Kimberley Watson-Mackie
School of Health and Social Development, Faculty of Health
Deakin University
221 Burwood Highway
Burwood, 3125
Australia
Email: kwatsonm@deakin.edu.au

Abstract

Background: Despite the known benefits of physical activity (PA), rates of engagement in PA remain low globally. Low engagement in PA among young women can impact their health. Technology-supported PA may increase PA and physical literacy (PL; skills that can support PA) among young women.

Objective: This systematic review aims to investigate the (1) associations between technology-supported PA and PA levels, (2) associations between technology-supported PA and PL levels, and (3) types of technology-supported PA that are associated with higher levels of PA engagement among young women aged 13 to 24 years. This age range was chosen as it includes two transitional periods characterized by decreases in PA.

Methods: We searched 6 databases: Applied Science and Technology Source, Education Source, Embase, MEDLINE Complete, Global Health, and SPORTDiscus. Eligible studies were original research published in English between January 1, 2010, and April 24, 2024; focused on young women; and involving either technology-supported PA interventions or research exploring the correlation between technology and PA and PL. The findings of the review were presented descriptively. Study quality was assessed using the JBI Critical Appraisal tools. There were no deviations from the registered protocol.

Results: In total, 23 (0.1%) studies (10,233 participants) from 23,609 records were included: randomized controlled trials (n=9, 39%), nonrandomized or retrospective observational studies (n=9, 39%), and cross-sectional studies (n=5, 22%). Of the 23 studies, 12 (52%) focused on young adults (aged ≥19 y), 9 (39%) involved adolescents (aged <19 y), and 2 (9%) focused on both groups. Nine studies (39%) were theory based. Common types of technology-supported PA were interactive websites or social media platforms (10/23, 43%), wearable fitness trackers (4/23, 17%), and mobile apps (4/23, 17%). PA or PL were predominantly self-reported (18/23, 78%). A total of 53 PA outcomes were measured: 36% (19/53) reported a positive impact on PA from mobile apps (9/15; 60% of analyses), interactive websites or social media platforms (8/27; 30% of analyses) and wearable fitness trackers (2/11; 18% of analyses). The impact on PL was weak (2/7; 29% of analyses). Eight studies (35%) were rated high, 7 (30%) medium, and 8 (35%) low in quality.

Conclusions: There was limited evidence that technology-supported PA improved young women's PA or PL. The evidence was limited by poor study quality and a lack of theoretical frameworks. In addition, little information was provided on the designs of the technology used. Future interventions seeking to improve young women's PA and PL should focus on the development of mobile apps underpinned by behavior change theory and addressing whole domains of PL rather than specific elements. Given

XSL•FO
RenderX

that technology continues to rapidly advance, further studies are needed to demonstrate the impact of technology-supported PA in improving PA and PL among young women.

(J Med Internet Res 2024;26:e52302) doi:10.2196/52302

KEYWORDS

physical literacy; young women; girls; adolescents; technology; mobile; digital; fitness; exercise; physical activity; technology-supported; systematic review; digital health; eHealth; health informatics; mobile phone

Introduction

Background

Physical activity (PA) is considered one of the most effective ways to maintain good health across the lifespan [1,2]. The World Health Organization (WHO) PA guidelines recommend at least 60 minutes of PA per day for those aged 5 to 17 years and 150 to 300 minutes per week for those aged 18 to 64 years [3]. Meeting the WHO PA guidelines helps individuals maintain a healthy weight, reduces the risk of developing noncommunicable diseases such as type 2 diabetes and obesity [1,2,4], and may improve mental well-being and increase academic performance [5,6]. Despite the known benefits, global rates of PA engagement are low, with 37% failing to meet guidelines [7]. Women are significantly less likely to meet PA guidelines than men (32% vs 23%, respectively) [7]. The lowest level of PA is seen in adolescent girls aged 11 to 17 years, with 85% failing to meet guidelines, compared to 78% of adolescent boys [7].

Adolescence is the transitional period from childhood to adulthood, characterized by rapid growth and changing social expectations [8,9]. The period of adolescence is generally considered to be between the ages of 13 and 17 years; therefore, health interventions seeking to improve adolescent health often focus on this age group [3,8,9]. Conversely, the United Nations Children's Fund (UNICEF) defines adolescence as individuals aged 10 to 19 years [10], with recent research suggesting that the period of adolescence should be extended further to include those aged 10 to 24 years [9]. Sawyer et al [9] suggest that a focus on "young people," which would include traditional adolescents (aged 13-17 y) and young adults (aged 18-24 y), would account for varying growth patterns and changes in the timing of social role transitions across different countries. Another consideration is the development of lifelong PA habits. Young people move through 2 transitional periods: starting secondary school and then higher education [1,11,12]. It is during this period that individuals become responsible for their own health and develop their PA beliefs and behaviors, which generally remain consistent across the rest of their lifespan [1,11-13]. Although adolescents and young adults have different experiences, focusing on both age groups targets the transitional periods in which PA engagement decreases. For these reasons, this review is focused on young women aged 13 to 24 years.

Given the known benefits of PA and the importance of developing healthy habits during adolescence, there have been numerous interventions seeking to increase PA engagement during this life stage [14-18]. These interventions are primarily aimed at individuals aged <18 years, with none examining those aged 13 to 24 years specifically [14,16-18]. They also generally focus on both male and female individuals within the target age range [14,16-18]. One review [16] of school-based PA interventions identified a small increase in PA. Another review of 39 reviews of child and adolescent PA interventions reported a small positive effect [14]; however, the review noted that the positive impact of the interventions was small.

As previous PA interventions seeking to improve young women's PA have had limited success, a different approach may be required [2,14,16,17,19-21]. One area of interest is the use of technology-supported PA as a strategy to increase PA [22-26]. Technology has been used as a tool for health promotion since the first mobile fitness apps were released in 2010, and technology-supported PA use increased during the COVID-19–related lockdowns [24,27-30]. Technology-supported PA can be defined as the use of some form of interactive technology or digitally accessed information to promote PA through (1) the demonstration of PA (eg, prerecorded or live-streamed fitness classes), (2) interaction with a device that provides feedback (eg, smartphones and wearable fitness trackers), and (3) interaction with fitness professionals (eg, web-based personal training) or other users of technology-supported PA (eg, through a fitness app or social media platform) [24,27,29,30]. Technology-supported PA is either self-led, with individuals using the technology in their own time (eg, apps and wearable fitness trackers), or facilitated, with sessions conducted by fitness professionals (eg, personal trainers or yoga instructors) in real time, allowing trainers to interact with their clients directly [24,27,29,30].

Recent systematic reviews focusing on the general population have suggested that technology-supported PA use is associated with increased PA and that the most successful types used behavior change techniques; were easy to use; and included gamification, such as offering some competition or challenge [8,18,22,25,31]. These reviews have generally focused only on a single form of technology (eg, apps or wearable fitness trackers). The level of effectiveness of technology-supported PA in increasing PA varies, with some studies reporting an overall significant improvement in PA for intervention groups compared to comparison groups [32,33] and others reporting no change in PA [34]. Reviews that did report improvements in PA, such as the studies by Champion et al [32] (22 publications involving 18,873 participants) and Lee et al [33] (16 interventions), noted that only a few of the included studies had postintervention follow-ups, and when they were included, it seemed that the improvements were not maintained [32,33].

In addition to the potential benefits for young women's PA engagement, technology-supported PA could also improve young women's physical literacy (PL) [35-37]. PL is a holistic approach to health that goes beyond simply engaging in PA;

XSL•FO
RenderX

rather, it is focused on developing the skills, behaviors, and confidence needed to lead an active life [36,38,39]. PL is complex in nature, and it is defined and conceptualized in various ways across the globe [35-37,40]. One of the most comprehensive understandings of PL is provided in the Australian Physical Literacy Framework [38]. This framework groups the elements needed to improve PL into 4 domains: physical (eg, strength and movement skills), psychological (eg, confidence and motivation), social (eg, relationships and collaboration), and cognitive (eg, content knowledge and reasoning) [38]. Emerging research suggests that the development of PA habits is tied to the development of PL [18,35,38,39]. Studies seeking to improve PL have focused primarily on school-age children, with none examining young women specifically [35,41,42]. No review has examined whether technology-supported PA could impact each domain of PL [8,18,22,25].

Objectives

Further research is needed to investigate the effectiveness of technology-supported PA use by young women and to identify the types of technology that may facilitate increased PA engagement and improve PL. The purpose of this systematic review was to investigate, in young women aged 13 to 24 years, the associations between different types of technology-supported PA and (1) PA engagement and (2) PL.

Methods

Overview

The selection of studies, analysis of data, and reporting of study results were conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [43]. The study was registered with PROSPERO in December 2022 (CRD42022382471), and there were no deviations from this protocol.

Search Strategy

We conducted a systematic search of 6 databases: MEDLINE Complete, SPORTDiscus, Global Health, Education Source, Applied Science and Technology Source, and Embase. These databases were selected with advice from the Deakin University librarian due to their alignment with the review objectives. The search focused on articles published between January 1, 2010, and April 24, 2024. The strategy combined synonyms for “young women,” “technology-supported physical activity,” and “physical activity” (refer to Multimedia Appendix 1 for a full list of the terms used in the search), with truncation used to maximize search results.

Eligibility Criteria

Eligible studies included randomized controlled trials (RCTs), nonrandomized interventions or retrospective observational studies investigating the effectiveness of technology-supported PA, as well as longitudinal and cross-sectional studies investigating the potential correlation between technology-supported PA use and PA engagement. Articles that were peer reviewed, contained original research, and published in English after 2010 were considered. The period from 2010 onward was selected because this year marked the release of the first PA mobile apps [44]. Studies that focused on women aged between 13 and 24 years were the primary target. The reference lists of the included articles were also searched for studies that may have been missed in the initial literature search.

Study Selection and Screening

Studies were imported into Covidence (Veritas Health Innovation Ltd) [45] for screening. Qualitative studies, duplicates, articles not available in English, opinion articles, conference abstracts, systematic reviews, and study protocols were excluded. These study types were excluded because they either did not provide original data on the effectiveness of technology-supported PA or, in the case of conference abstracts and non-English articles, did not provide sufficient detail for data analysis. Full-text inclusion and exclusion criteria (Multimedia Appendix 2) were applied by the research team to identify articles for data extraction. Search terms related to PL were not included in the database search because the review focused on interventions that increased PA as the primary outcome of interest; in our analysis, we investigated whether these interventions included the elements of PL. Studies were excluded if they did not provide data on the target population, did not focus on increasing PA, or involved technology targeting specialized populations or those with chronic conditions. However, we included interventions focused on decreasing obesity because preliminary searches indicated that one of the main aims of obesity interventions was increasing PA, while interventions for other chronic conditions did not typically have PA as a key focus. In cases where the sample was the target age and included both male and female participants, studies were excluded if the results for young women were not reported separately.

Data Extraction and Synthesis

Extraction was conducted using Covidence and Microsoft Excel. Extracted data included author, country, intervention setting, study design, theoretical framework, participant characteristics and demographic information, study aims, PA measurement methods, and the data collection tools used. For the intervention studies, information on the duration of the intervention and any follow-up conducted was also extracted. Details of the intervention and any control group were included, as was the method of data analysis. Finally, information was extracted on the effectiveness of the intervention and the effect size of any changes (if these were reported). Title and abstract screening and data collection were conducted by at least 2 authors independently. Disagreements on inclusion were discussed with a third author as needed. Before the full data extraction process, we extracted data from a sample of the studies and compared the findings to ensure consistency with data extraction.

To assess the effectiveness of the interventions and the reported associations between the use of technology-supported PA and PA in the cross-sectional studies, the reviewers divided the studies into 3 categories of technology-supported PA: interactive website or social media platforms, PA-tracking mobile apps, and wearable fitness trackers. Analyses were then conducted based on the effectiveness of each category of technology on PA (improvement or positive association, decline or negative

XSL•FO
RenderX

association, or null effect). Analyses of the effect that the interventions had on the category of PA measured (accelerometry, PA intensity, guideline adherence, energy expenditure, step count, time spent walking, at-home exercise sessions, and increased exercise) were also conducted. The effectiveness of the interventions was assessed using a scale adapted from work conducted by Page et al [46]. If ≤30% of the studies reported a positive effect on PA or PL, the impact was coded as "no likely effect" ("0"). If 31% to 60% of the studies reported improvements, these were coded as "uncertain" ("?") [46]. Finally, in instances where ≥61% of the studies reported changes in the expected direction (either PA improvements or decreases in sedentary behavior), the results were coded as "positive" ("+").

The included studies were assessed according to their alignment to any of the 30 elements of the Australian Physical Literacy Framework [38]. PL is a relatively new concept, and definitions regarding what it is and is not vary depending on the framework used [47]. In addition, many studies explore outcomes that can be considered PL without defining it as PL [48]. We agreed that outcomes related to changes in social support and self-efficacy would not be included as PL outcomes. Although some aspects of social support could relate to elements within the social domain of the Australian Physical Literacy Framework, and self-efficacy could be related to the framework's psychological domain, they are not included as part of this particular framework [38].

Alongside assessing the effectiveness of the interventions on the PA of young women aged 13 to 24 years, the research team divided the sample into 2 subgroups: adolescents aged <19 years and young adults aged ≥19 years. This was to account for the potentially significant physical, physiological, social, and environmental differences between these subgroups as well as the different PA guidelines recommended for each age group to be considered physically active.

Quality Assessment

Two authors independently assessed each publication for quality and risk of bias using the JBI Critical Appraisal tools [43]. Three checklists were used to account for the different study designs used in the included articles: RCT, cross-sectional, and quasi-experimental. To ensure consistency between the quality assessments and the reliability of the quality analysis, we discussed the steps we would take to ensure consistency across all checklists and the different study designs included in the review; for example, we adapted questions in the cross-sectional and quasi-experimental checklists slightly to make them more consistent with those in the RCT checklist. Originally, only the RCT checklist required an exploration of both the validity and reliability of the tools used, while the other checklists only required an investigation of reliability; however, this requirement was added to the cross-sectional and quasi-experimental checklists. Furthermore, the validity and reliability of the data collection methods was considered in reference to the target population of young women. Within the RCT checklist, some questions explored the blinding of both participants and assessors. We decided that participant blinding to treatment assignment would be considered only if the study explicitly stated that blinding occurred or if each group received some form of intervention, the rationale being that blinding would be likely in this scenario because there would be no reason for the researchers to inform participants whether they were in the intervention or control group. When it came to assessor blinding, this was only considered if data collection was conducted in person and not via online self-report. Each study was assessed using the most suitable JBI checklist, due to variation in study designs the research team felt that the quality of retrospective observational studies was better assessed with the cross-sectional checklist rather than the quasi-experimental one. So, while retrospective observational studies would be considered nonrandomized in analysis they were grouped with the cross-sectional studies for quality analysis. If agreement on study quality could not be reached, a third author resolved the discrepancy by undertaking an independent assessment of the publication, and a final decision was made by consensus among the 3 authors.

Results

Description of Studies

The database searches yielded 23,609 records (Applied Science and Technology Source: n=1408, 5.96%; Education Source: n=6292, 26.65%; Embase: n=5432, 23.01%; Global Health: n=2336, 9.89%; MEDLINE Complete: n=5724, 24.24%; and SPORTDiscus: n=2417; 10.24%), of which 4716 (19.98%) were duplicates and removed, and 18,893 (80.02%) were screened by 2 authors based on title and abstract. At this stage, there were 23 disagreements that were resolved via discussion with the 2 authors who screened the papers. After title and abstract screening was completed, 219 articles (n=7, 3.2% involved disagreements that were resolved) were included in the full-text screening. Manual searches of the references lists of these articles were conducted, but no additional relevant articles were found (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

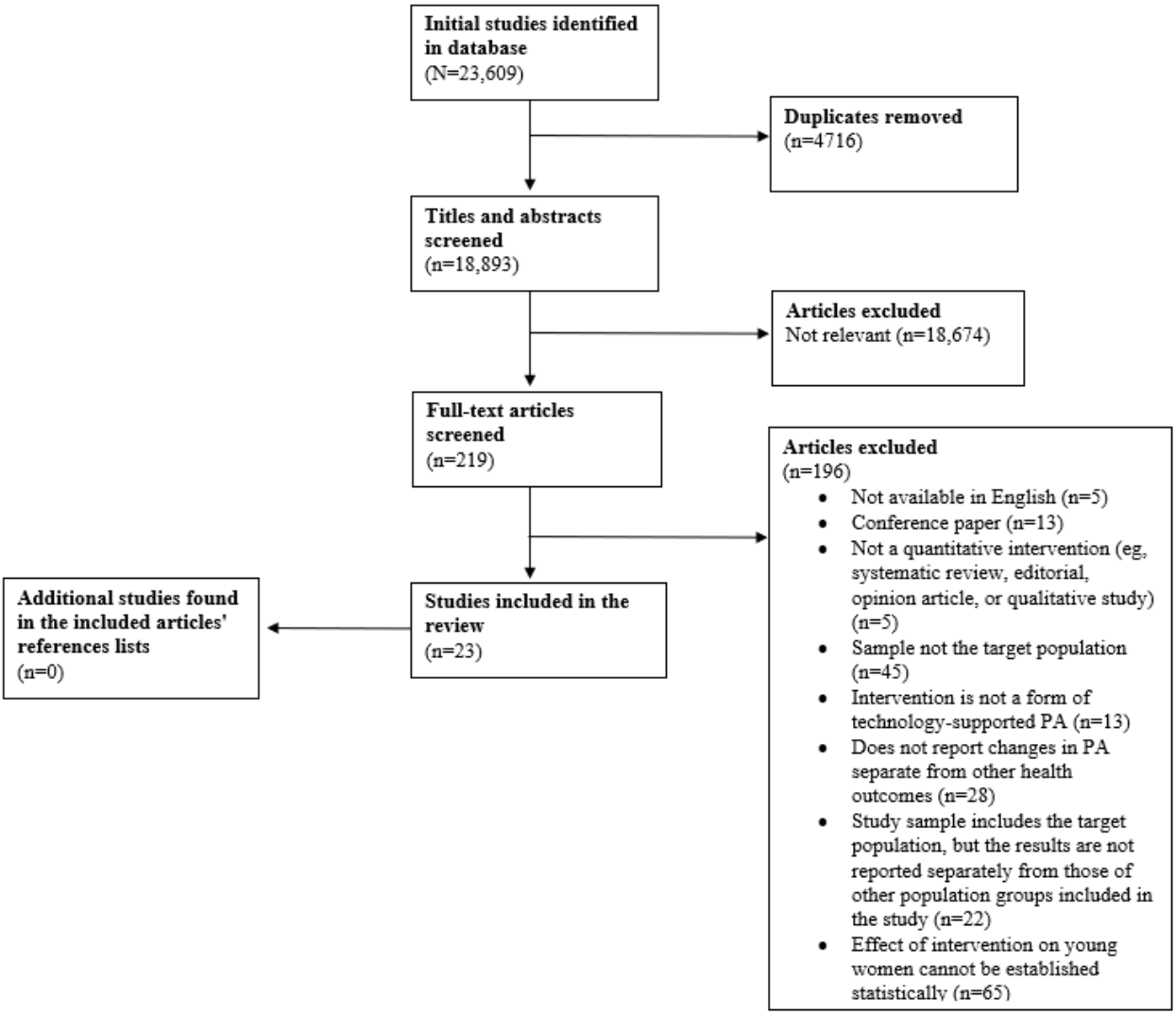

Study Characteristics

A total of 23 studies comprising 10,233 participants were included in the final review. Of these 23 studies, 18 (78%) involved interventions (RCTs: n=9, 50%; nonrandomized trials or retrospective observational studies: n=9, 50%): 12 (67%) included a control group (Multimedia Appendix 3) [40,49-59], and 6 (33%) involved a single sample (Multimedia Appendix 4) [60-65]. The remaining studies (5/23, 22%) were cross-sectional and investigated correlation between technology-supported PA and PA engagement (Multimedia Appendix 5) [66-70]. Of the 23 studies, 5 (22%) [40,60,62,63,69] explored variables that can be classified as elements of PL, although none mentioned PL specifically (Multimedia Appendix 6). Of these 5 studies, 4 (80%) included elements within the psychological domain [40,62,63,69] of the Australian Physical Literacy Framework [38], while 1 (20%) included elements within the framework's physical domain [60].

Sample sizes ranged from 16 [60] to 4128 [68] participants. Of the 23 studies, 12 (52%) included only the target population [40,50-52,54,55,57,60-62,64,66], while the remaining 11 (48%) also included other populations, such as male participants, and individuals outside the target population age range [49,53,56,58,59,63,65,67-70]. Of the 23 studies, 9 (39%) included only adolescents [49,52,56-59,63,64,68], 12 (52%) focused on young adults [40,50,53-55,60-62,66,67,69,70], and 2 (9%) looked at both groups [51,65]; moreover, 6 (26%) sampled participants who were overweight or obese [40,52,54,55,61,62], and 3 (13%) focused on participants who were insufficiently active [51,58,64].

The types of technology-supported PA varied. Of the 18 interventions (RCTs: n=9, 50%; nonrandomized trials or retrospective observational studies: n=9, 50%), 4 (22%) used wearable fitness trackers [54,56,58,63], 10 (56%) used an interactive website or social media platform [40,49-51,53,59-62,64], and 4 (22%) used mobile apps [52,55,57,65].

Of the 4 wearable fitness tracker interventions, 3 (75%) were conducted with adolescents [56,58,63], and 1 (25%) was conducted with young adults [54].

Of the 10 interactive website or social media platform interventions, 3 (30%) targeted adolescents [49,59,64], 6 (60%) focused on young adults [40,50,53,60-62], and 1 (10%) focused on both groups [51].

XSL•FO
RenderX

Of the 4 interventions using mobile apps, 2 (50%) were conducted with adolescents [52,57], 1 (25%) was focused on young adults [55], and 1 (25%) focused on both groups [65].

Of the 18 intervention studies, 8 (44%) used forms of technology-supported PA that had been designed specifically for the study either as a stand-alone intervention or combined with a commercially available form of technology [49-53,58,61,62].

The cross-sectional studies compared various types of technology-supported PA, including mobile apps, and wearable fitness trackers [66-70]. Of the 5 cross-sectional studies, 1 (20%) focused on adolescents [68], and the other 4 (80%) involved young adults [66,67,69,70].

The most common intervention setting was universities [40,50,51,53-55,61,62,66,69,70], followed by high schools [49,52,56-59,63], while 5 (22%) of the 23 studies were conducted with the general population [60,64,65,67,68]. Of the 23 studies, 11 (48%) took place in the United States [51,53,54,59,61,62,64-67,69]; 2 (9%) were conducted in Australia [56,60]; and 1 (4%) study each was conducted in Saudi Arabia [40], the United Arab Emirates [50], Israel [49], China [70], Pakistan [55], Poland [52], Singapore [57], the Netherlands [58], and the United Kingdom [63]. Furthermore, 1 (4%) of the 23 studies compared data between Finland and Ireland [68].

Of the 23 studies, 9 (39%) reported being theory-based [50,53,54,58,61-64,66], although Kattelmann et al [53] and Melton et al [54] did not specify the theory on which their intervention was based. Social cognitive theory was the theory most commonly used (5/9, 56%) [50,61,62,64,66], with McFadden [66] combining it with the transtheoretical model of behavior change. The remaining studies (2/9, 22%) were based on self-determination theory [58,63]. Of the 9 theory-based studies, 3 (33%) focused on adolescents [58,63,64], and 6 (67%) involved young adults [50,53,54,61,62,66].

The most common method of data collection was self-report surveys, with 12 (52%) of the 23 studies using this method [40,49,52,53,58-60,66-70]. Of these 12 studies, 5 (42%) involved adolescents [49,52,58,59,68], and 7 (58%) focused on young adults [40,53,60,66,67,69,70]. Device-based data were collected in 5 (22%) of the 23 studies with an accelerometer [63,64], commercial wearable fitness tracker [54], or step-tracking commercial mobile app [55,65]. Both accelerometer studies were conducted with adolescents [63,64], tracking apps were used with young adults in 2 (40%) [54,55] of the 5 studies, and the remaining study (1/5, 20%) focused on both age groups [65]. Of the 23 studies, 6 (26%) used a combination of subjective and device-based measures [50,51,56,57,61,62]: 2 (33%) with adolescents [56,57] and 3 (50%) with young adults [50,61,62], while 1 (17%) investigated both groups [51].

Levels of PA were recorded as time spent in PA (days or minutes per week) in 8 (89%) of the 9 adolescent studies [49,52,56,58,59,63,64,68], 5 (42%) of the 12 studies with young adults [50,60,66,69,70], and 1 (50%) of the 2 studies that investigated all young women [51]. Both studies conducted on young adults by Joseph et al [61,62] combined days or minutes of weekly PA with interactive website engagement. Of the 4 studies that measured PA as total reported steps, 2 (50%) were conducted with young adults [54,55], 1 (25%) with adolescents [57], and 1 (25%) with both age groups [65]. The final tools of PA measurement—attendance at in-person PA sessions [40], total metabolic equivalents of task (METs) [53], and level of exercise engagement [67]—were all used in interventions focused on young adults.

A little more than half of the studies that included a control and comparison group (7/12, 58%) reported a positive effect on PA. Of these 7 studies, 4 (57%) were RCTs [51,52,54,58], and 3 (43%) were nonrandomized or retrospective observational studies (Multimedia Appendix 3) [40,49,50]. Of the 6 single-sample studies included in the review, 3 (50%) reported a positive outcome [61,64,65], 2 (33%) had no impact [60,62], and 1 (17%) reported a decrease in PA after the intervention (Multimedia Appendix 4) [63]. All cross-sectional studies included in the review (5/23, 22%) [66-70] reported a positive association between technology-supported PA use and PA engagement (Multimedia Appendix 5).

Results of the Quality Assessment

The quality of the 23 studies varied, with 8 (35%) found to be of high quality [51,54,58,61,65,67,68,70], 7 (30%) rated as medium quality [49,52,53,56,59,60,62], and 8 (35%) determined to be low quality [40,50,55,57,63,64,66,69].

All RCTs included in the review (9/23, 39%; Multimedia Appendix 7) reported that true randomization was used to assign participants to the intervention and control groups, that the groups were treated identically, and that the outcome measurements used were the same for both groups [51-59]. However, some of the RCTs (4/9, 44%) did not provide information on participant or assessor blinding; of the 9 RCTs, only 2 (22%) reported that allocation to treatment groups was concealed [55,58], and only 3 (33%) used participant blinding [51,52,54]. No RCT reported whether those delivering the intervention were blind to the treatment group, although the question was not applicable in a little more than half of these RCTs (5/9, 56%) [51-55] because the intervention was self-led. In addition, no RCT stated whether the assessors measuring the outcomes of the intervention were blind to the treatment group. Of the 9 RCTs, 5 (56%) used self-report as the data collection method [51,53,54,57,58]; therefore, the question was not considered applicable. However, the remaining RCTs (4/9, 44%) did use outcome assessors and either reported that the assessors were not blinded [56] or did not report on it at all [52,55,59].

The quasi-experimental studies assessed with the relevant JBI checklist (Multimedia Appendix 7) were found to be of lower quality, but this was in part due to the variation in study designs. As such, questions related to the use of comparison or control groups only related to the studies by Al-Eisa et al [40], Ali et al [50], and Glaser et al [49]. Al-Eisa et al [40] and Ali et al [50] reported using a control group and assessing the intervention and control groups in the same way but did not provide information on the similarities between the 2 groups. Glaser et al [49] reported assessing the intervention and control groups in the same way but also provided information on the

XSL•FO
RenderX

similarities between the groups at baseline, which was why the study was considered medium quality (56%).

All 8 quasi-experimental studies [40,49,50,60-64] outlined the independent and dependent variables under investigation, but none reported why the chosen method of data analysis was used. Only 3 (38%) of the 8 studies collected data at multiple points during the intervention period [60,61], and an intervention follow-up was only conducted by Joseph et al [61] and Larsen et al [64]. Only 1 (12%) of these 8 studies assessed with the JBI quasi-experimental checklist was found to be of high quality [61]. Of the overall 8 studies that were considered low quality, 4 (50%) were from this group [40,50,63,64].

Of the 6 cross-sectional and retrospective observational studies assessed with the relevant JBI checklist (Multimedia Appendix 7), 4 (67%) were considered high quality [65,67,68,70], while 2 (33%) were considered low quality [66,69]. All 6 studies provided details on the study participants and setting, the validity and reliability of the outcome measurements, and the appropriateness of the statistical analysis method used [65-70]. However, only half of the studies (3/6, 50%) clearly defined the participant inclusion criteria and reported whether objective standard criteria had been used for the measurement of the condition under investigation [65,68,70], and only 3 (50%) of the 6 studies identified potential confounding factors and explained how these factors had been addressed [67,68,70]. After conducting the quality assessment, we found that only Ng et al [68] and Wang et al [70] provided enough information to positively answer all checklist questions.

Summary of the Findings

Associations With PA Outcomes

A summary of the study results according to the 3 types of technology-supported PA (interactive website or social media platform, PA-tracking mobile app, and wearable fitness tracker) and the type of PA measures assessed is provided in Table 1.

XSL•FO
RenderX

Table 1. Summary of physical activity (PA) results in the expected direction classified by type of PA measurement.

	Interactive website or social media platform			PA-tracking mobile app			Wearable fitness tracker			Proportion of positive findings
	Positive[a]	Negative[a]	Null[a]	Positive	Negative	Null	Positive	Negative	Null	
Accelerometry										
Accelerometry counts	—[b]	Larsen et al [64], 2018[c]	Joseph et al [62], 2016[c]	—	—	—	—	—	Melton et al [54], 2016[d]; Ridgers et al [56], 2021[d]	0/4
Total	0	1	1	0	0	0	0	0	2	0/4 (0%); positive outcomes code: 0[e]
PA intensity and duration										
Light										
Min/wk	—	—	—	—	—	—	—	—	Slootmaker et al [58], 2010[d]	0/1
Moderate										
D/wk	Ali et al [50], 2021[d]	—	Whittemore et al [59], 2013[d]	—	—	—	—	—	—	1/2
Min/d	—	—	Ali et al [50], 2021[d]	—	—	—	—	—	Ridgers et al [56], 2021[d]	0/2
Min/wk	—	—	Curtis et al [60], 2020[c]; Kattelmann et al [53], 2014[d]; Papalia et al [69], 2018[f]	Wang et al [70], 2019[f]	—	Seah and Koh [57], 2021[d,g]	Slootmaker et al [58], 2010[d]	—	—	2/6
Vigorous										
D/wk	Ali et al [50], 2021[d]	—	Whittemore et al [59], 2013[d]	—	—	—	—	—	—	1/2
Min/d	Ali et al [50], 2021[d]	—	—	—	—	—	—	—	—	1/1
Min/wk	Kattelmann et al [53], 2014[d]; Papalia et al [69], 2018[f]	—	Curtis et al [60], 2020[c]	Wang et al [70], 2019[f]	—	Seah and Koh [57], 2021[g]	—	—	Slootmaker et al [58], 2010[d]	3/6
Moderate to vigorous										
D/wk	Glaser et al [49], 2024[d]	—	—	Dzielska et al [52], 2020[d,h]	—	—	—	—	—	2/2
Min/d	—	—	—	—	—	—	—	Kerner et al [63], 2019[c]	—	0/1

XSL•FO
RenderX

	Interactive website or social media platform			PA-tracking mobile app			Wearable fitness tracker			Proportion of positive findings
	Positive[a]	Negative[a]	Null[a]	Positive	Negative	Null	Positive	Negative	Null	
Min/wk	Larsen et al [64], 2018[c]	—	Joseph et al [61], 2015[c]; Joseph et al [62], 2016[c]	—	—	—	—	—	Slootmaker et al [58], 2010[d]	1/4
Total	7	0	9	3	0	2	1	1	4	11/27 (41%); positive outcomes code: ?[i]
PA guideline adherence										
Meeting American College of Sports Medicine PA guidelines	—	—	—	—	—	—	McFadden [66], 2021[f]	—	—	1/1
Meeting World Health Organization PA guidelines	—	—	—	Ng et al [68], 2021[f]; Wang et al [70], 2019[f]	—	—	Ng et al [68], 2021[f]	—	—	3/3
Total	0	0	0	2	0	0	2	0	0	4/4 (100%); positive outcomes code: +[j]
Energy expenditure										
Min/wk of total metabolic equivalents of task	—	—	Curtis et al [60], 2020[c]; Kattelmann et al [53], 2014[d]	—	—	Seah and Koh [57], 2021[d,g]	—	—	—	0/3
PA (heavy, kcal)	—	—	Cavallo et al [51], 2012[d]	—	—	—	—	—	—	0/1
PA (light, kcal)	—	—	Cavallo et al [51], 2012[d]	—	—	—	—	—	—	0/1
PA (moderate, kcal)	—	—	Cavallo et al [51], 2012[d]	—	—	—	—	—	—	0/1
PA (total, kcal)	—	—	Cavallo et al [51], 2012[d]	—	—	—	—	—	—	0/1
Total	0	0	6	0	0	1	0	0	0	0/7 (0%); positive outcomes code: 0
Step count										
Total step count	—	—	—	Xian et al [65], 2017[c]	Memon et al [55], 2016[d]	Seah and Koh [57], 2021[d,g]	—	Melton et al [54], 2016[d]	—	1/4

XSL•FO
RenderX

	Interactive website or social media platform			PA-tracking mobile app			Wearable fitness tracker			Proportion of positive findings
	Positive[a]	Negative[a]	Null[a]	Positive	Negative	Null	Positive	Negative	Null	
Total	0	0	0	1	1	1	0	1	0	1/4 (25%); positive outcomes code: 0
Walking										
D/wk	—	—	Ali et al [50], 2021[d]	—	—	Wang et al [70], 2019[f]	—	—	—	0/2
Min/d	Ali et al [50], 2021[e]	—	Curtis et al [60], 2020[c]	—	—	—	—	—	—	1/2
Min/wk	—	—	Kattelmann et al [53], 2014[d]	—	—	—	—	—	—	0/1
Total	1	0	3	0	0	1	0	0	0	1/5 (20%); positive outcomes code: 0
At-home exercise sessions										
Number of at-home exercise sessions	Al-Eisa et al [40], 2016[d]	—	—	—	—	—	—	—	—	1/1
Total	1	0	0	0	0	0	0	0	0	1/1 (100%); positive outcomes code: +
Increased exercise										
Increased exercise over the last year	—	—	—	Nagata et al [67], 2021[f]	—	—	—	—	—	1/1
Total	0	0	0	1	0	0	0	0	0	1/1 (100%); positive outcomes code: +

XSL•FO
RenderX

	Interactive website or social media platform			PA-tracking mobile app			Wearable fitness tracker			Proportion of positive findings
	Positive[a]	Negative[a]	Null[a]	Positive	Negative	Null	Positive	Negative	Null	
Overall impact on PA	—	—	— —	—	—		—	—	—	19/53 (36%); positive outcomes code: ?

[a]Impact or association of study in the hypothesized direction.

[b]Not applicable.

[c]Intervention study with a single sample.

[d]Intervention study with comparison groups.

[e]0: no likely effect reported when ≤30% of the studies found changes in the expected direction (adapted from Page et al [46]).

[f]Cross-sectional study; shows correlation between technology and PA.

[g]Only looked at PA engagement on weekends.

[h]Only participants who were overweight or obese.

[i]Uncertain effect reported when 31% to 60% of the studies found changes in the expected direction (adapted from Page et al [46]).

[j]Positive effect reported when 61% to 100% of the studies found changes in the expected direction (adapted from Page et al [46]).

The included studies explored 53 different PA measures, which were grouped into 8 types: accelerometry, PA intensity and duration, PA guideline adherence, energy expenditure, step count, walking, at-home exercise sessions, and increased exercise. Most of the studies explored >1 type.

The most common PA outcome was self-reported intensity (15/23, 65%; 27 different analyses), which included both the type of PA, such as light, moderate, or vigorous, as well as the duration measured in minutes or days per week. Across these 15 studies [49,50,52,53,56-64,69,70], a positive effect or association was reported in 7 (47%; 11/27, 41% analyses); therefore, this was rated as uncertain in terms of effect (“?”).

Energy expenditure was the next most common outcome (4/23, 17%; 7 different analyses), which included METs and the number of calories expended per PA type. Of the 4 studies that included this measure, 3 (75%) measured minutes per week of METs [53,57,60], while Cavallo et al [51] measured PA in terms of kcal. The results of these interventions indicated that technology-supported PA had no likely effect on participant PA [51,53,57,60]; thus, this was rated as “0.”

Walking was measured as self-reported days or minutes per week or minutes per day of PA in 3 (17%) of the 18 intervention studies [50,53,60] and 1 (20%) of the 5 cross-sectional studies [70]. Of these 4 studies, only 1 (25%) reported that the use of technology-supported PA had a positive effect [50], while 3 (75%) reported a null result [53,60,70]; thus, this was rated as having no effect (“0”).

Adherence to PA guidelines (4 different analyses) was used as a measure in 3 (60%) of the 5 cross-sectional studies [66,68,70]. The guidelines used were those provided by the WHO [68,70] and the American College of Sports Medicine [66]. All forms of technology-supported PA measured were associated with greater adherence to PA guidelines [66,68,70]; therefore, this was rated as having a positive effect (“+”).

The overall positive impact of technology-supported PA was uncertain because only 36% (19/53) of the analyses reported a positive effect or changes in the expected direction.

Associations by Type of Technology-Supported PA

When comparing different types of technology-supported PA, data synthesis suggests that interactive websites or social media platforms (8/27, 30% analyses) and wearable fitness trackers (2/11, 18% analyses) had no likely effect on PA. The effect of mobile apps was more promising, but the full impact was uncertain (9/15, 60% analyses). A summary of these findings is reported in Table 2.

XSL•FO
RenderX

Table 2. Summary of physical activity (PA) results in the expected direction classified by type of intervention.

Type of intervention	Improvement or positive association reported in PA	Decline or negative association reported in PA	Null result in PA	Summary of results in the expected direction: analyses, n/N (%)	Code
Interactive website or social media platform	• Intervention study with comparison groups: • Al-Eisa et al [40], 2016[a]; Ali et al [50], 2021[b,c,d]; Kattelmann et al [53], 2014[c]; Glaser et al [49], 2014[e] • Intervention study with a single sample: • Larsen et al [64], 2018[e] • Cross-sectional study: • Papalia et al [69], 2018[c]	• Intervention study with a single sample: • Larsen et al [64], 2018[f]	• Intervention study with comparison groups: • Ali et al [50], 2021[b,d]; Cavallo et al [51], 2012[b,g,h]; Kattelmann et al [53], 2014[b,d,i]; Whittemore et al [59], 2013[b,c] • Intervention study with a single sample: • Curtis et al [60], 2020[b,c,d,i]; Joseph et al [61], 2015[e]; Joseph et al [62], 2016[e,f] • Cross-sectional study: • Papalia et al [69], 2018[b]	8/27 (30)	0[j]
PA-tracking mobile app	• Intervention study with comparison groups: • Dzielska et al [52], 2020[e,k] • Intervention study with a single sample: • Xian et al [65], 2017[l] • Cross-sectional study: • Ng et al [68], 2021[m,n]; Nagata et al [67], 2021[o]; Wang et al [70], 2019[b,c,m]	• Intervention study with comparison groups: • Memon et al [55], 2016[l]	• Intervention study with comparison groups: • Seah and Koh [57], 2021[b,c,i,l] • Cross-sectional study: • Wang et al [70], 2019[d]	9/15 (60)	?[p]
Wearable fitness tracker	• Intervention study with comparison groups: • Slootmaker et al [58], 2010[b] • Cross-sectional study: • McFadden [66], 2021[q]	• Intervention study with comparison groups: • Melton et al [54], 2016[l] • Intervention study with a single sample: • Kerner et al [63], 2019[e]	• Intervention study with comparison groups: • Melton et al [54], 2016[f]; Ridgers et al [56], 2021[b,f]; Slootmaker et al [58], 2010[g,c,e] • Cross-sectional study: • Ng et al [68], 2021[m]	2/11 (18)	0
Overall impact	__[r]	—	—	19/53 (36)	?

[a]Number of home exercise sessions.
[b]Moderate PA.
[c]Vigorous PA.
[d]Walking.
[e]Moderate to vigorous PA.
[f]Accelerometry counts.
[g]Light PA.
[h]Heavy PA.
[i]Metabolic equivalent of task.
[J]0: no likely effect reported when ≤30% of the studies found changes in the expected direction (adapted from Page et al [46]).

XSL•FO
RenderX

[k]Improvement seen only in participants who were overweight or obese.

[l]Total step counts.

[m]Meeting World Health Organization PA guidelines.

[n]Investigated multiple forms of technology-supported PA.

[o]Increased exercise in the previous year.

[p]Uncertain effect reported when 31% to 60% of the studies found changes in the expected direction (adapted from Page et al [46]).

[q]Meeting American College of Sports Medicine PA guidelines.

[r]Not applicable.

Associations With PL Outcomes

The elements of PL were explored in 4 (40%) of the 10 interactive website or social media platform studies [40,60,62,69] and in 1 (25%) of the 4 wearable fitness tracker studies (Multimedia Appendix 6) [63]. Improvements were reported in 2 (40%) [40,69] of these 5 studies for motivation [40] and engagement or enjoyment [69], while the other 3 (60%) [60,62,63] reported that the intervention had no effect. Overall, there was no effect on PL (2/7, 29% analyses; Multimedia Appendix 6).

Differences in the Effectiveness of Technology-Supported PA Between Adolescent and Young Adult Subgroups

Of the 15 studies that reported a positive impact or association between the use of technology-supported PA and PA engagement, 8 (52%) were conducted with young adults [40,50,54,61,66,67,69,70]. By comparison, only 33% (5/15) of the effective interventions focused on adolescents, while both the interventions that looked at both groups reported a positive result [51,65].

Mobile apps were only associated with positive PA outcomes when used by adolescents or across both age groups [65]. Both age groups reported positive PA outcomes when using interactive website or social media platforms (2/5, 40% with adolescents; 3/5, 60% with young adults) [40,49,50,61,64], while the adolescent [58] and young adult [54] subgroup each had 1 wearable fitness tracker intervention that reported a positive outcome. Overall, 56% (5/9) of the adolescent and 67% (8/12) of the young adult interventions reported a positive outcome.

Effectiveness of Technology-Supported PA According to Study Quality and Theoretical Framework

Of the 23 studies included in the review, 9 (39%) reported being underpinned by a theoretical design [50,53,54,58,61-64,66]. Of these 9 studies, 6 (67%) reported a positive outcome or association [50,54,58,61,64,66]. Among the 6 effective theory-based studies, social cognitive theory was the theory most commonly used, with 4 (67%) interventions drawing from this theory [50,61,62,64] and the cross-sectional study by McFadden [66] combining it with the transtheoretical model of behavior change. However, of the 4 effective theory-based studies, only 1 (25%) was considered high quality [61], with the other 3 (75%) considered poor quality [50,64,66]. The remaining effective theory-based studies (2/6, 33%) were considered high quality [54,58]; Slootmaker et al [58] used the self-determination theory, while Melton et al [54] did not provide information on the theory used.

Analysis of the non–theory-based interventions showed that 64% (9/14) had positive outcomes or associations between technology-supported PA use and PA engagement [40,49,51,52,65,67-70]. Of these 9 studies, 4 (44%) were cross-sectional studies [67-70]; therefore, only a positive association between technology-supported PA and PA use could be reported. Of the 9 effective non–theory-based interventions, 5 (56%) were considered high quality [51,65,67,68,70], 2 (22%) were considered medium quality [49,52], and 2 (22%) were rated poor quality [40,69] (Multimedia Appendix 7).

Discussion

Principal Findings

Overview

The primary aim of this review was to investigate the effectiveness of various types of technology-supported PA in increasing young women's PA engagement. The secondary aim was to assess whether any of these interventions explored the elements of PL and whether the interventions led to improvements in PL. There were 3 main types of technology-supported PA investigated in these studies: mobile apps, wearable fitness trackers, and interactive websites or social media platforms. Analysis of the study findings did not indicate that technology-supported PA is an effective method of increasing young women's PA, although, when breaking the findings down by age group, technology-supported PA may have a greater impact on PA engagement for young adults than for adolescents. There was no evidence that technology-supported PA is an effective way of increasing young women's PL.

Effectiveness of Mobile Apps

While the overall impact of technology-supported PA on young women's PA was uncertain, mobile apps may hold some promise, with positive results reported in 2 (50%) of the 4 studies [52,65] and 60% (9/15) of the measured PA outcomes. Currently, >80% of the global population own a mobile device, and these rates continue to increase [71]. Mobile phone use is especially prevalent among adolescents, with data indicating that in some countries, up to 95% of those aged 13 to 19 years have a mobile device, and individuals in this age group report higher levels of daily use than other age groups [72]. This higher rate of ownership and use may make mobile apps aiming to improve levels of PA engagement more effective in those aged <19 years, although only 4 (17%) of the 23 studies in this review focused on mobile apps, highlighting how little research has been conducted on this form of technology-supported PA [52,55,57,65]. Another consideration when investigating the

XSL•FO
RenderX

effectiveness of mobile apps is the use of various forms of "gamification" [22,31-33]. Gamification, such as offering some competition or challenge, has been linked to improved intrinsic motivation and higher levels of app engagement [22,31-33]. Both mobile app studies included in this review that reported improvements in PA used gamification [52,65]. This result is in line with previously conducted studies, with several systematic reviews reporting that gamified apps led to higher levels of PA engagement than apps without elements of gamification [22,31-33]. However, just because mobile apps are popular does not necessarily mean that they are effective [22,34,73-75]. A common limitation of PA mobile apps is a lack of evidence and theory-based design, which has been reported in several systematic reviews [22,34,73-75]. This limitation can be seen in the interventions included in this review because both the mobile app interventions that did not improve PA used commercial fitness apps without providing evidence of reliability or a theoretical framework [55,57].

Effectiveness of Interactive Websites or Social Media platforms

Interventions based on interactive websites or social media platforms were not found to be an effective way of improving young women's PA. Only 30% (8/27) of the measured PA outcomes assessed in interventions using an interactive website or social media platforms reported a positive effect or changes in the expected direction. Social media platforms are reported to be well liked by young women and considered an effective way of improving PA facilitators such as motivation or social support [23,76-79]. Nevertheless, these findings suggest that while social media platforms are popular, it does not necessarily mean that they are effective. This disconnect between engagement and effectiveness has been explored in previous research [76,78]. The study conducted by Duplaga [78] reported that only 33% of young adults following fitness influencers on social media platforms engaged in regular PA, while a survey by Camacho-Miñano et al [76] (37 participants aged 13-17 y) reported that Instagram fitness groups were associated with negative outcomes such as body dissatisfaction [40,60,76,78].

Effectiveness of Wearable Fitness Trackers

The least effective form of technology-supported PA was wearable fitness trackers, with only 18% (2/11) of the measured PA outcomes reporting a positive effect or changes in the expected direction. Only 2 (50%) of the 4 wearable fitness tracker interventions reported a positive result [54,58], and 1 (25%) decreased participant PA [63]; the authors theorized that this decrease may be due to wearable fitness trackers only increasing external motivation rather than autonomous motivation, which promotes long-term behavior change [80]. The limited effectiveness of wearable fitness trackers could be due to the design of the devices used in the interventions. Of the 4 wearable fitness tracker interventions, 3 (75%) used commercial devices [54,56,63], which are often not theory or evidenced based [74]. This may mean that the wearable fitness tracker used might not produce significant improvements in PA, even if the intervention is otherwise well designed. A lack of compliance with the devices and insufficient wear time could also have been a factor. Of the 4 studies, 3 (75%) reported issues with wearable fitness tracker compliance, and previous research has reported that this is a common issue with wearable fitness tracker interventions [81,82]. Another consideration is the age group in which the wearable fitness tracker interventions were conducted. Of the 4 interventions, 3 (75%) focused on adolescents [56,58,63], but a recent study by Shandhi et al [83] reported that young adults were the age group most likely to own and use a wearable fitness tracker; therefore, it is possible that the limited effectiveness of wearable fitness trackers reported in this review is due to the interventions being conducted primarily with adolescents [54,56,58,63,83].

Variation Between Study Designs

The results may not only be attributed to the type of technology-supported PA. The different characteristics of the studies, including study designs, PA outcomes, and measurement tools should be considered; for example, theory-based interventions are reported to result in better outcomes and have more generalizable findings than those that are not [84,85]. This was evident in the review findings because of the 14 studies that were effective, only 6 (43%) had a theoretical framework [50,54,58,61,64,66]. In comparison, only 3 (38%) of the 8 studies that were not effective, reported being theory-based [53,62,63], although it is important to remember that a study being based on a theoretical framework does not automatically make it well designed. This can be seen in the findings of this review; it was noted that only 3 (33%) of the 9 effective theory-based interventions were high quality [54,58,61]. Our findings also noted that studies in the review were more likely to report a positive outcome or association between technology-supported PA use and PA engagement if they were high quality, regardless of the study design [51,54,58,61,65,67,68,70]. This is in line with recent reviews of health research, which note that studies that report higher quality are more likely to report positive outcomes [86,87].

Another consideration is the method of data collection. In PA research, data are primarily collected via self-report or device-based measures (eg, pedometers and accelerometers), and both methods have benefits and limitations [81,82]. Self-report methods are low cost and easy to administer but are prone to recall bias and overreporting of PA [81,82]. While device-based methods reduce the risk of bias, they are limited by cost, reduced generalizability, incorrect use, and the fact that the devices may not be suitable for all types of PA [81]. An example of this can be seen in the study by Larsen et al [64], who reported an increase in self-reported moderate to vigorous PA (from 24.7 min to 79.4 min) but a decrease in device-measured moderate to vigorous PA (from 21.4 to 10.4 min), attributed to changes in the type of PA the sample was engaging in after the intervention that was not effectively measured by accelerometers [64].

Impact on PL

There was little evidence that technology-supported PA could improve young women's PL, with only 29% (2/7) of the PL elements assessed reporting a positive effect or changes in the expected direction. Only 2 (40%) of the 5 included studies reported positive changes in PL [40,69]. The remaining interventions (3/5, 60%) reported no likely effect on PL

XSL•FO
RenderX

[60,62,63]. This is not surprising, given that these 3 interventions also failed to increase PA engagement [60,62,63]. Only 1 (20%) of the 5 PL interventions focused on adolescents [63]. The primary domain explored in these interventions was the psychological domain. This focus on the psychological domain makes sense, given that these elements are known facilitators of PA for young women [31,88-90]. It should be noted that although these studies were found to explore aspects of PL, none of them mentioned the term "physical literacy" specifically [40,60,62,63,69]. However, this was expected because PL is a newer concept and the Australian Physical Literacy Framework was only published in 2019 [38]. Only 2 (40%) of the 5 PL studies were published after this date [60,63].

Strengths and Limitations

To our knowledge, this review is the first synthesis of the impact that various types of technology-supported PA could have on the PA of young women (aged 13-24 y). This is a critical period for health promotion because this age range is when young women's PA begins to decrease, and lifelong PA habits are formed [1,2]. This review also investigated and compared different types of technology-supported PA, while other reviews have focused on a single form of technology such as mobile health or mobile apps [22,33,34]. Focusing on a single form of technology ignores the fast-paced changes in technological advances, especially in a post–COVID-19 world. This is also the first review to investigate the potential impact that technology-supported PA could have on PL. The review was further strengthened by the use of a comprehensive quality assessment. It was also written in accordance with the PRISMA 2020 checklist (Multimedia Appendix 8). Alongside these strengths, there are also some limitations. The first is the inclusion of several single-sample studies because these do not have the same high quality found in RCTs. In addition, several cross-sectional studies were included; these cannot provide a causal link between technology-supported PA use and PA engagement. However, given the limited data available on the use of technology-supported PA by the target population, valuable insights would have been missed if the review had only included RCT studies. The variety of study designs included in this review also made assessing the quality of the studies more complex. Three different quality checklists were used in this review. The risk of this impacting the results was mitigated because we agreed on ways to make the different checklists more consistent. The analysis of the results of the included interventions could also have been hindered by the number of different PA measurements. There were 53 different ways in which PA outcomes were measured; therefore, we sought to address this by combining the different measurements into subgroups. The reported positive impact of mobile apps must be considered with caution due to the small number of studies (4/23, 17%) that focused on this form of technology-supported PA. It is possible that mobile apps may only seem to be more effective than interactive websites or social media platforms and wearable fitness trackers because the limited outcomes measured skewed the results. Another consideration is the poor quality of the included studies and a lack of theory-based design. We cannot be certain that the findings reported in this review are due to the limited effectiveness of technology-supported PA. It may instead be related to the quality of the studies and the types of technology-supported PA used. The findings of this review would have been strengthened if a meta-analysis of the findings could have been conducted; however, the variation between study designs, the measurement of PA outcomes, and study quality did not make this viable.

Future Research

The findings of this review suggest 2 areas for future research. First, more research focused on young women's use of technology-supported PA is needed. Only 12 (52%) of the 23 studies in this review examined young women specifically [40,50-52,54,55,57,60-62,64,66]. Previous research reports that young women experience unique barriers to PA that may not be targeted in interventions involving wider populations [60,89-91]. The second area that requires investigation is the potential of facilitated technology-supported PA. This review was unable to investigate the effectiveness of facilitated technology-supported PA because it was not used in any of the included studies. This finding was not surprising because facilitated technology-supported PA was not common before the COVID-19 pandemic. However, its use increased significantly during the pandemic, and it has continued to remain popular [27,30,92]. Emerging research indicates that facilitated technology-supported PA is most popular with young women, and fitness professionals see it as an effective method of client engagement [27,92]. Another consideration when addressing these areas is how researchers choose to measure PA. Greater consistency in the type of PA measured when investigating technology-supported PA would make comparison between interventions more effective.

Conclusions

This is the first systematic review exploring the use of various types of technology-supported PA. It highlighted that there is no evidence yet for the benefits of technology-supported PA for young women. The review also highlighted how little research has been conducted in this area. Many of the studies included in this review were of poor quality and not grounded in theory; in addition, none investigated facilitated, technology-supported PA. Nevertheless, some of our findings indicate areas of promise for the future. Future interventions could focus on mobile apps because they may be more effective than interactive website or social media platforms and commercial wearable fitness trackers. Furthermore, interventions that combine multiple forms of support, such as mentors or in-person instruction, may be more effective than a single form of technology-supported PA. Adolescents and young adults may experience different barriers and facilitators to PA and the use of technology-supported PA. This must be considered when conducting research on this age group because it may be that different forms of technology are needed for each subgroup of young women. The review findings noted that interventions that were theory-based may be more effective than those that were not. Researchers should consider developing interventions underpinned by behavior change theory, with follow-ups after the intervention to see whether improvements in PA and the use of technology-supported PA have been maintained. In

XSL•FO
RenderX

addition, more research is needed on the impact that technology-supported PA could have on adolescents because the review findings of the impact on this age group are very limited. This will not only improve the health outcomes of young women in the short term but also help them develop the skills and confidence needed to engage in PA across the lifespan. As technology-supported PA continues to improve and become more common, there is a greater need for well-designed evidence-based research exploring the impact that these new types of technology could have on young women's PA and PL.

Acknowledgments

LA is supported by an Australian Research Council Discovery Early Career Researcher Award (DE220100847). KW-M is supported by a PhD candidature from the School of Health and Social Development, Deakin University. NL is supported by an Australian Research Council Discovery Early Career Researcher Award (DE240100452). The authors would like to thank the reviewers for their comments.

Authors' Contributions

KW-M, LMB, NL, FHM, and LA conceived the idea for this review and developed the search strategy. KW-M conducted the literature search and screened all articles at both the title and abstract and full-text stages. LMB, NL, FHM, LA, AJ, and FV assisted with the screening and worked with KW-M to select the articles for inclusion in the review. Preliminary data extraction of these articles was conducted by KW-M and checked by LMB, NL, LA, AJ, and FV. KW-M, LMB, and AJ assessed the included physical activity articles to see whether they included aspects of physical literacy (PL) and worked together to extract the PL data. The quality assessment of all included articles was conducted by KW-M, LMB, FHM, and FV, and the quality assessment of the study elements related to PL were conducted by KW-M, LMB, and AJ. The summary and results tables included in the review were developed by KW-M, LMB, LA, AJ, and FV. KW-M wrote the first and second drafts of the manuscript, and initial feedback was provided by LMB, NL, LA, and FHM. Before submission, all authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File (Adobe PDF File), 188 KB - jmir_v26i1e52302_app1.pdf]

Multimedia Appendix 2
Inclusion and exclusion criteria.
[PDF File (Adobe PDF File), 362 KB - jmir_v26i1e52302_app2.pdf]

Multimedia Appendix 3
Intervention and result details from comparison group studies.
[PDF File (Adobe PDF File), 243 KB - jmir_v26i1e52302_app3.pdf]

Multimedia Appendix 4
Intervention and result details from single-sample studies.
[PDF File (Adobe PDF File), 401 KB - jmir_v26i1e52302_app4.pdf]

Multimedia Appendix 5
Result details from cross-sectional studies.
[PDF File (Adobe PDF File), 622 KB - jmir_v26i1e52302_app5.pdf]

Multimedia Appendix 6
Summary of results in the expected direction by domain and element of physical literacy.
[PDF File (Adobe PDF File), 208 KB - jmir_v26i1e52302_app6.pdf]

Multimedia Appendix 7
Quality assessment for each study design.
[PDF File (Adobe PDF File), 686 KB - jmir_v26i1e52302_app7.pdf]

XSL•FO
RenderX

Multimedia Appendix 8
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 463 KB - jmir_v26i1e52302_app8.pdf]

References

1. Feil K, Allion S, Weyland S, Jekauc D. A systematic review examining the relationship between habit and physical activity behavior in longitudinal studies. Front Psychol 2021;12:626750 [FREE Full text] [doi: 10.3389/fpsyg.2021.626750] [Medline: 33746848]
2. Posadzki P, Pieper D, Bajpai R, Makaruk H, Könsgen N, Neuhaus AL, et al. Exercise/physical activity and health outcomes: an overview of Cochrane systematic reviews. BMC Public Health 2020 Nov 16;20(1):1724 [FREE Full text] [doi: 10.1186/s12889-020-09855-3] [Medline: 33198717]
3. WHO guidelines on physical activity and sedentary behaviour (contract no.: CC BY-NC-SA 3.0 IGO). World Health Organization. 2020. URL: https://iris.who.int/bitstream/handle/10665/336656/9789240015128-eng.pdf?sequence=1 [accessed 2024-04-29]
4. The Lancet Public Health. Time to tackle the physical activity gender gap. Lancet Public Health 2019 Aug;4(8):e360 [FREE Full text] [doi: 10.1016/S2468-2667(19)30135-5] [Medline: 31345750]
5. Andermo S, Hallgren M, Nguyen TT, Jonsson S, Petersen S, Friberg M, et al. School-related physical activity interventions and mental health among children: a systematic review and meta-analysis. Sports Med Open 2020 Jul 16;6(1):25 [FREE Full text] [doi: 10.1186/s40798-020-00254-x] [Medline: 32548792]
6. Biddle SJ, Ciaccioni S, Thomas G, Vergeer I. Physical activity and mental health in children and adolescents: an updated review of reviews and an analysis of causality. Psychol Sport Exerc 2019 May;42:146-155. [doi: 10.1016/j.psychsport.2018.08.011]
7. Global action plan on physical activity 2018-2030: more active people for a healthier world (contract no.: CC BY-NC-SA 3.0). World Health Organization. URL: https://iris.who.int/bitstream/handle/10665/336656/9789240015128-eng.pdf?sequence=1 [accessed 2024-04-29]
8. Rossi L, Behme N, Breuer C. Physical activity of children and adolescents during the COVID-19 pandemic-a scoping review. Int J Environ Res Public Health 2021 Oct 30;18(21):11440 [FREE Full text] [doi: 10.3390/ijerph182111440] [Medline: 34769956]
9. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health 2018 Mar;2(3):223-228. [doi: 10.1016/S2352-4642(18)30022-1] [Medline: 30169257]
10. Adolescent development and participation. UNICEF. 2022. URL: https://www.unicef.org/adolescence [accessed 2024-04-29]
11. Hawlader MD, Mozid NE, Sharmin S, Monju IH, Ahmed SB, Sarker W, et al. The art of forming habits: applying habit theory in changing physical activity behaviour. J Public Health (Berl.) 2022 Nov 02;31(12):2045-2057. [doi: 10.1007/S10389-022-01766-4]
12. Marginson S. Research on international and global higher education: six different perspectives. Oxf Rev Educ 2022 Jul 28;48(4):421-438. [doi: 10.1080/03054985.2022.2087619]
13. Park AH, Zhong S, Yang H, Jeong J, Lee C. Impact of COVID-19 on physical activity: a rapid review. J Glob Health 2022 Apr 30;12:05003 [FREE Full text] [doi: 10.7189/jogh.12.05003] [Medline: 35493780]
14. Messing S, Rütten A, Abu-Omar K, Ungerer-Röhrich U, Goodwin L, Burlacu I, et al. How can physical activity be promoted among children and adolescents? A systematic review of reviews across settings. Front Public Health 2019 Mar 19;7:55 [FREE Full text] [doi: 10.3389/fpubh.2019.00055] [Medline: 30941342]
15. Violant-Holz V, Gallego-Jiménez MG, González-González CS, Muñoz-Violant S, Rodríguez MJ, Sansano-Nadal O, et al. Psychological health and physical activity levels during the COVID-19 pandemic: a systematic review. Int J Environ Res Public Health 2020 Dec 15;17(24):9419 [FREE Full text] [doi: 10.3390/ijerph17249419] [Medline: 33334073]
16. Owen MB, Curry WB, Kerner C, Newson L, Fairclough SJ. The effectiveness of school-based physical activity interventions for adolescent girls: a systematic review and meta-analysis. Prev Med 2017 Dec;105:237-249. [doi: 10.1016/j.ypmed.2017.09.018] [Medline: 28964852]
17. Sims J, Scarborough P, Foster C. The effectiveness of interventions on sustained childhood physical activity: a systematic review and meta-analysis of controlled studies. PLoS One 2015 Jul 20;10(7):e0132935 [FREE Full text] [doi: 10.1371/journal.pone.0132935] [Medline: 26193472]
18. Yomoda K, Kurita S. Influence of social distancing during the COVID-19 pandemic on physical activity in children: a scoping review of the literature. J Exerc Sci Fit 2021 Jul;19(3):195-203 [FREE Full text] [doi: 10.1016/j.jesf.2021.04.002] [Medline: 34135976]
19. Amiri Farahani L, Asadi-Lari M, Mohammadi E, Parvizy S, Haghdoost AA, Taghizadeh Z. Community-based physical activity interventions among women: a systematic review. BMJ Open 2015 Apr 01;5(4):e007210 [FREE Full text] [doi: 10.1136/bmjopen-2014-007210] [Medline: 25833668]
20. Cleland V, Granados A, Crawford D, Winzenberg T, Ball K. Effectiveness of interventions to promote physical activity among socioeconomically disadvantaged women: a systematic review and meta-analysis. Obes Rev 2013 Mar;14(3):197-212. [doi: 10.1111/j.1467-789X.2012.01058.x] [Medline: 23107292]

XSL•FO
RenderX

21. Condello G, Puggina A, Aleksovska K, Buck C, Burns C, Cardon G, DEDIPAC consortium. Behavioral determinants of physical activity across the life course: a "DEterminants of DIet and Physical ACtivity" (DEDIPAC) umbrella systematic literature review. Int J Behav Nutr Phys Act 2017 May 02;14(1):58 [FREE Full text] [doi: 10.1186/s12966-017-0510-2] [Medline: 28464958]
22. Angosto S, García-Fernández J, Valantine I, Grimaldi-Puyana M. The intention to use fitness and physical activity apps: a systematic review. Sustainability 2020 Aug 17;12(16):6641. [doi: 10.3390/SU12166641]
23. Bicen A, Uzunboylu H, Burgul NS. Evaluation of participants' opinions on online physical fitness training. J Sport Psychol 2020;29(4):25-32 [FREE Full text]
24. Füzéki E, Schröder J, Groneberg DA, Banzer W. Online exercise classes during the COVID-19 related lockdown in Germany: use and attitudes. Sustainability 2021 Jul 09;13(14):7677. [doi: 10.3390/SU13147677]
25. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Feb 19;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
26. van Sluijs EM, Ekelund U, Crochemore-Silva I, Guthold R, Ha A, Lubans D, et al. Physical activity behaviours in adolescence: current evidence and opportunities for intervention. Lancet 2021 Jul 31;398(10298):429-442 [FREE Full text] [doi: 10.1016/S0140-6736(21)01259-9] [Medline: 34302767]
27. Bratland-Sanda S, Mathisen TF, Sundgot-Borgen C, Sundgot-Borgen J, Tangen JO. The impact of COVID-19 pandemic lockdown during spring 2020 on personal trainers' working and living conditions. Front Sports Act Living 2020;2:589702 [FREE Full text] [doi: 10.3389/fspor.2020.589702] [Medline: 33345164]
28. Parker K, Uddin R, Ridgers ND, Brown H, Veitch J, Salmon J, et al. The use of digital platforms for adults' and adolescents' physical activity during the COVID-19 pandemic (our life at home): survey study. J Med Internet Res 2021 Feb 01;23(2):e23389 [FREE Full text] [doi: 10.2196/23389] [Medline: 33481759]
29. Dor-Haim H, Katzburg S, Revach P, Levine H, Barak S. The impact of COVID-19 lockdown on physical activity and weight gain among active adult population in Israel: a cross-sectional study. BMC Public Health 2021 Aug 06;21(1):1521 [FREE Full text] [doi: 10.1186/s12889-021-11523-z] [Medline: 34362319]
30. COVID-19 - fitness industry impact report. Fitness Australia. 2020. URL: https://bp-fitnessaustralia-production.s3.amazonaws.com/uploads/uploaded_file/file/472634/COVID-19_Fitness_Industry_Impact_report_F.pdf [accessed 2024-04-29]
31. Knight RL, McNarry MA, Sheeran L, Runacres AW, Thatcher R, Shelley J, et al. Moving forward: understanding correlates of physical activity and sedentary behaviour during COVID-19-an integrative review and socioecological approach. Int J Environ Res Public Health 2021 Oct 17;18(20):10910 [FREE Full text] [doi: 10.3390/ijerph182010910] [Medline: 34682653]
32. Champion KE, Parmenter B, McGowan C, Spring B, Wafford QE, Gardner LA, Health4Life team. Effectiveness of school-based eHealth interventions to prevent multiple lifestyle risk behaviours among adolescents: a systematic review and meta-analysis. Lancet Digit Health 2019 Sep;1(5):e206-e221 [FREE Full text] [doi: 10.1016/S2589-7500(19)30088-3] [Medline: 33323269]
33. Lee AM, Chavez S, Bian J, Thompson LA, Gurka MJ, Williamson VG, et al. Efficacy and effectiveness of mobile health technologies for facilitating physical activity in adolescents: scoping review. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e11847 [FREE Full text] [doi: 10.2196/11847] [Medline: 30747716]
34. Böhm B, Karwiese SD, Böhm H, Oberhoffer R. Effects of mobile health including wearable activity trackers to increase physical activity outcomes among healthy children and adolescents: systematic review. JMIR Mhealth Uhealth 2019 Apr 30;7(4):e8298 [FREE Full text] [doi: 10.2196/mhealth.8298] [Medline: 31038460]
35. Liu Y, Chen S. Physical literacy in children and adolescents: definitions, assessments, and interventions. Eur Phys Educ Rev 2020 Jun 11;27(1):96-112. [doi: 10.1177/1356336X20925502]
36. Longmuir PE, Tremblay MS. Top 10 research questions related to physical literacy. Res Q Exerc Sport 2016;87(1):28-35. [doi: 10.1080/02701367.2016.1124671] [Medline: 26889582]
37. Sum RK, Cheng CF, Wallhead T, Kuo CC, Wang FJ, Choi SM. Perceived physical literacy instrument for adolescents: a further validation of PPLI. J Exerc Sci Fit 2018 May;16(1):26-31 [FREE Full text] [doi: 10.1016/j.jesf.2018.03.002] [Medline: 30662489]
38. Australian physical literacy framework. Sport Australia. 2019. URL: https://www.sportaus.gov.au/__data/assets/pdf_file/0019/710173/35455_Physical-Literacy-Framework_access.pdf [accessed 2024-04-29]
39. Sum RK, Ha AS, Cheng CF, Chung PK, Yiu KT, Kuo CC, et al. Construction and validation of a perceived physical literacy instrument for physical education teachers. PLoS One 2016;11(5):e0155610 [FREE Full text] [doi: 10.1371/journal.pone.0155610] [Medline: 27195664]
40. Al-Eisa E, Al-Rushud A, Alghadir A, Anwer S, Al-Harbi B, Al-Sughaier N, et al. Effect of motivation by "Instagram" on adherence to physical activity among female college students. Biomed Res Int 2016;2016:1546013 [FREE Full text] [doi: 10.1155/2016/1546013] [Medline: 27034927]
41. Carl J, Barratt J, Wanner P, Töpfer C, Cairney J, Pfeifer K. The effectiveness of physical literacy interventions: a systematic review with meta-analysis. Sports Med 2022 Dec;52(12):2965-2999 [FREE Full text] [doi: 10.1007/s40279-022-01738-4] [Medline: 35994237]

XSL•FO
RenderX

42. Filho VC, Pereira WM, Farias BD, Moreira TM, Guerra PH, Queiroz AC, et al. Scoping review on interventions for physical activity and physical literacy components in Brazilian school-aged children and adolescents. Int J Environ Res Public Health 2021 Aug 06;18(16):8349 [FREE Full text] [doi: 10.3390/ijerph18168349] [Medline: 34444097]
43. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
44. Muntaner-Mas A, Martinez-Nicolas A, Lavie CJ, Blair SN, Ross R, Arena R, et al. A systematic review of fitness apps and their potential clinical and sports utility for objective and remote assessment of cardiorespiratory fitness. Sports Med 2019 Apr;49(4):587-600 [FREE Full text] [doi: 10.1007/s40279-019-01084-y] [Medline: 30825094]
45. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2024-04-29]
46. Page ZE, Barrington S, Edwards J, Barnett LM. Do active video games benefit the motor skill development of non-typically developing children and adolescents: a systematic review. J Sci Med Sport 2017 Dec;20(12):1087-1100. [doi: 10.1016/j.jsams.2017.05.001] [Medline: 28600111]
47. Edwards LC, Bryant AS, Keegan RJ, Morgan K, Cooper SM, Jones AM. 'Measuring' physical literacy and related constructs: a systematic review of empirical findings. Sports Med 2018 Mar 15;48(3):659-682 [FREE Full text] [doi: 10.1007/s40279-017-0817-9] [Medline: 29143266]
48. Essiet IA, Lander NJ, Salmon J, Duncan MJ, Eyre EL, Ma J, et al. A systematic review of tools designed for teacher proxy-report of children's physical literacy or constituting elements. Int J Behav Nutr Phys Act 2021 Oct 08;18(1):131 [FREE Full text] [doi: 10.1186/s12966-021-01162-3] [Medline: 34620185]
49. Glaser M, Green G, Barak S, Bord S, Levi S, Jakobovich R, et al. The effects of the Friendship Online intervention program on physical activity, substance abuse, psychosomatic symptoms, and well-being among at-risk youth. J Adolesc 2024 Feb;96(2):251-265. [doi: 10.1002/jad.12272] [Medline: 37985148]
50. Ali HI, Attlee A, Alhebshi S, Elmi F, Al Dhaheri AS, Stojanovska L, et al. Feasibility study of a newly developed technology-mediated lifestyle intervention for overweight and obese young adults. Nutrients 2021 Jul 26;13(8):2547 [FREE Full text] [doi: 10.3390/nu13082547] [Medline: 34444707]
51. Cavallo DN, Tate DF, Ries AV, Brown JD, DeVellis RF, Ammerman AS. A social media-based physical activity intervention: a randomized controlled trial. Am J Prev Med 2012 Nov;43(5):527-532 [FREE Full text] [doi: 10.1016/j.amepre.2012.07.019] [Medline: 23079176]
52. Dzielska A, Mazur J, Nałęcz H, Oblacińska A, Fijałkowska A. Importance of self-efficacy in eating behavior and physical activity change of overweight and non-overweight adolescent girls participating in healthy me: a lifestyle intervention with mobile technology. Nutrients 2020 Jul 17;12(7):2128 [FREE Full text] [doi: 10.3390/nu12072128] [Medline: 32709005]
53. Kattelmann KK, Bredbenner CB, White AA, Greene GW, Hoerr SL, Kidd T, et al. The effects of young adults eating and active for health (YEAH): a theory-based web-delivered intervention. J Nutr Educ Behav 2014;46(6):S27-S41. [doi: 10.1016/j.jneb.2014.08.007] [Medline: 25457733]
54. Melton BF, Buman MP, Vogel RL, Harris BS, Bigham LE. Wearable devices to improve physical activity and sleep. J Black Stud 2016 Jul 27;47(6):610-625. [doi: 10.1177/0021934716653349]
55. Memon AR, Masood T, Awan W, Waqas A. The effectiveness of an incentivized physical activity programme (active student) among female medical students in Pakistan: a randomized controlled trial. J Pak Med Assoc 2018 Oct;68(10):1438-1445. [Medline: 30317338]
56. Ridgers ND, Timperio A, Ball K, Lai SK, Brown H, Macfarlane S, et al. Effect of commercial wearables and digital behaviour change resources on the physical activity of adolescents attending schools in socio-economically disadvantaged areas: the RAW-PA cluster-randomised controlled trial. Int J Behav Nutr Phys Act 2021 Apr 12;18(1):52 [FREE Full text] [doi: 10.1186/s12966-021-01110-1] [Medline: 33845853]
57. Seah ML, Koh KT. The efficacy of using mobile applications in changing adolescent girls' physical activity behaviour during weekends. Eur Phys Educ Rev 2020 Jun 17;27(1):113-131. [doi: 10.1177/1356336x20930741]
58. Slootmaker SM, Chinapaw MJ, Seidell JC, van MW, Schuit AJ. Accelerometers and internet for physical activity promotion in youth' feasibility and effectiveness of a minimal intervention [ISRCTN93896459]. Prev Med 2010 Jul;51(1):31-36. [doi: 10.1016/j.ypmed.2010.03.015] [Medline: 20380847]
59. Whittemore R, Jeon S, Grey M. An internet obesity prevention program for adolescents. J Adolesc Health 2013 May;52(4):439-447 [FREE Full text] [doi: 10.1016/j.jadohealth.2012.07.014] [Medline: 23299003]
60. Curtis RG, Ryan JC, Edney SM, Maher CA. Can Instagram be used to deliver an evidence-based exercise program for young women? A process evaluation. BMC Public Health 2020 Oct 06;20(1):1506 [FREE Full text] [doi: 10.1186/s12889-020-09563-y] [Medline: 33023559]
61. Joseph RP, Dutton GR, Cherrington A, Fontaine K, Baskin M, Casazza K, et al. Feasibility, acceptability, and characteristics associated with adherence and completion of a culturally relevant internet-enhanced physical activity pilot intervention for overweight and obese young adult African American women enrolled in college. BMC Res Notes 2015 Jun 02;8:209 [FREE Full text] [doi: 10.1186/s13104-015-1159-z] [Medline: 26032016]

XSL•FO
RenderX

62. Joseph RP, Pekmezi D, Dutton GR, Cherrington AL, Kim Y, Allison JJ, et al. Results of a culturally adapted internet-enhanced physical activity pilot intervention for overweight and obese young adult African American women. J Transcult Nurs 2016 Mar;27(2):136-146 [FREE Full text] [doi: 10.1177/1043659614539176] [Medline: 24934566]
63. Kerner C, Burrows A, McGrane B. Health wearables in adolescents: implications for body satisfaction, motivation and physical activity. Int J Health Promot Educ 2019 Mar 04;57(4):191-202. [doi: 10.1080/14635240.2019.1581641]
64. Larsen B, Benitez T, Cano M, Dunsiger SS, Marcus BH, Mendoza-Vasconez A, et al. Web-based physical activity intervention for Latina adolescents: feasibility, acceptability, and potential efficacy of the Niñas Saludables study. J Med Internet Res 2018 May 09;20(5):e170 [FREE Full text] [doi: 10.2196/jmir.9206] [Medline: 29743151]
65. Xian Y, Xu H, Xu H, Liang L, Hernandez AF, Wang TY, et al. An initial evaluation of the impact of Pokémon GO on physical activity. J Am Heart Assoc 2017 May 16;6(5):e005341 [FREE Full text] [doi: 10.1161/JAHA.116.005341] [Medline: 28512111]
66. McFadden C. Wearable exercise technology and the impact on college women's physical activity. Quest 2021 Mar 09;73(2):179-191. [doi: 10.1080/00336297.2021.1891553]
67. Nagata JM, Hazzard VM, Ganson KT, Hahn SL, Neumark-Sztainer D, Eisenberg ME. Digital technology use and muscle-building behaviors in young adults. Int J Eat Disord 2022 Feb 14;55(2):207-214 [FREE Full text] [doi: 10.1002/eat.23656] [Medline: 34904744]
68. Ng K, Kokko S, Tammelin T, Kallio J, Belton S, O'Brien W, et al. Clusters of adolescent physical activity tracker patterns and their associations with physical activity behaviors in Finland and Ireland: cross-sectional study. J Med Internet Res 2020 Oct 01;22(9):e18509 [FREE Full text] [doi: 10.2196/18509] [Medline: 32667894]
69. Papalia Z, Wilson O, Bopp M, Duffey M. Technology-based physical activity self-monitoring among college students. Int J Exerc Sci 2018;11(7):1096-1104 [FREE Full text] [Medline: 30338023]
70. Wang T, Ren M, Shen Y, Zhu X, Zhang X, Gao M, et al. The association among social support, self-efficacy, use of mobile apps, and physical activity: structural equation models with mediating effects. JMIR Mhealth Uhealth 2019 Sep 25;7(9):e12606 [FREE Full text] [doi: 10.2196/12606] [Medline: 31573936]
71. Olson JA, Sandra DA, Colucci É, Al Bikaii A, Chmoulevitch D, Nahas J, et al. Smartphone addiction is increasing across the world: a meta-analysis of 24 countries. Comput Human Behav 2022 Apr;129:107138. [doi: 10.1016/j.chb.2021.107138]
72. Fortunato L, Lo Coco G, Teti A, Bonfanti RC, Salerno L. Time spent on mobile apps matters: a latent class analysis of patterns of smartphone use among adolescents. Int J Environ Res Public Health 2023 Jul 25;20(15):6439 [FREE Full text] [doi: 10.3390/ijerph20156439] [Medline: 37568981]
73. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017 Apr 19;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
74. Paganini S, Terhorst Y, Sander LB, Catic S, Balci S, Küchler A, et al. Quality of physical activity apps: systematic search in app stores and content analysis. JMIR Mhealth Uhealth 2021 Jun 09;9(6):e22587 [FREE Full text] [doi: 10.2196/22587] [Medline: 34106073]
75. Panicker RM, Chandrasekaran B. "Wearables on vogue": a scoping review on wearables on physical activity and sedentary behavior during COVID-19 pandemic. Sport Sci Health 2022;18(3):641-657 [FREE Full text] [doi: 10.1007/s11332-021-00885-x] [Medline: 35018193]
76. Camacho-Miñano MJ, MacIsaac S, Rich E. Postfeminist biopedagogies of Instagram: young women learning about bodies, health and fitness. Sport Educ Soc 2019 May 22;24(6):651-664. [doi: 10.1080/13573322.2019.1613975]
77. Drehlich M, Naraine M, Rowe K, Lai SK, Salmon J, Brown H, et al. Using the technology acceptance model to explore adolescents' perspectives on combining technologies for physical activity promotion within an intervention: usability study. J Med Internet Res 2020 Mar 06;22(3):e15552 [FREE Full text] [doi: 10.2196/15552] [Medline: 32141834]
78. Duplaga M. The use of fitness influencers' websites by young adult women: a cross-sectional study. Int J Environ Res Public Health 2020 Oct 01;17(17):32882887 [FREE Full text] [doi: 10.3390/ijerph17176360] [Medline: 32882887]
79. Parker K, Gould L, Nand M, Rawstorn JC, Contardo Ayala AM, Maddison R, et al. Understanding Australian adolescent girls' use of digital technologies for healthy lifestyle purposes: a mixed-methods study. BMC Public Health 2022 Aug 01;22(1):1464 [FREE Full text] [doi: 10.1186/s12889-022-13869-4] [Medline: 35915431]
80. Ntoumanis N, Ng JY, Prestwich A, Quested E, Hancox JE, Thøgersen-Ntoumani C, et al. A meta-analysis of self-determination theory-informed intervention studies in the health domain: effects on motivation, health behavior, physical, and psychological health. Health Psychol Rev 2021 Jun;15(2):214-244. [doi: 10.1080/17437199.2020.1718529] [Medline: 31983293]
81. Pedišić Ž, Bauman A. Accelerometer-based measures in physical activity surveillance: current practices and issues. Br J Sports Med 2015 Feb;49(4):219-223. [doi: 10.1136/bjsports-2013-093407] [Medline: 25370153]
82. Steene-Johannessen J, Anderssen SA, van der Ploeg HP, Hendriksen IJ, Donnelly AE, Brage S, et al. Are self-report measures able to define individuals as physically active or inactive? Med Sci Sports Exerc 2016 Feb;48(2):235-244 [FREE Full text] [doi: 10.1249/MSS.0000000000000760] [Medline: 26322556]

XSL•FO
RenderX

83. Shandhi MM, Singh K, Janson N, Ashar P, Singh G, Lu B, et al. Assessment of ownership of smart devices and the acceptability of digital health data sharing. NPJ Digit Med 2024 Feb 22;7(1):44 [FREE Full text] [doi: 10.1038/s41746-024-01030-x] [Medline: 38388660]
84. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
85. Heath G, Cooke R, Cameron E. A theory-based approach for developing interventions to change patient behaviours: a medication adherence example from paediatric secondary care. Healthcare (Basel) 2015 Dec 04;3(4):1228-1242 [FREE Full text] [doi: 10.3390/healthcare3041228] [Medline: 27417822]
86. Dal Santo T, Rice DB, Amiri LS, Tasleem A, Li K, Boruff JT, et al. Methods and results of studies on reporting guideline adherence are poorly reported: a meta-research study. J Clin Epidemiol 2023 Jul;159:225-234. [doi: 10.1016/j.jclinepi.2023.05.017] [Medline: 37271424]
87. Pirosca S, Shiely F, Clarke M, Treweek S. Tolerating bad health research: the continuing scandal. Trials 2022 Jun 02;23(1):458 [FREE Full text] [doi: 10.1186/s13063-022-06415-5] [Medline: 35655288]
88. Dechrai IM, Mazzoli E, Hanna L, Morgan PJ, Young MD, Grounds JA, et al. Are gender-stereotyped attitudes and beliefs in fathers and daughters associated with girls' perceived motor competence? Phys Educ Sport Pedagogy 2022 Jun 06:1-14. [doi: 10.1080/17408989.2022.2083097]
89. Rosselli M, Ermini E, Tosi B, Boddi M, Stefani L, Toncelli L, et al. Gender differences in barriers to physical activity among adolescents. Nutr Metab Cardiovasc Dis 2020 Aug 28;30(9):1582-1589. [doi: 10.1016/j.numecd.2020.05.005] [Medline: 32605880]
90. Watson A, Eliott J, Mehta K. Perceived barriers and facilitators to participation in physical activity during the school lunch break for girls aged 12–13 years. Eur Phys Educ Rev 2015 Feb 03;21(2):257-271. [doi: 10.1177/1356336X14567545]
91. Depper A, Howe PD. Are we fit yet? English adolescent girls' experiences of health and fitness apps. Health Sociol Rev 2016 Jul 04;26(1):98-112. [doi: 10.1080/14461242.2016.1196599]
92. García-Fernández J, Gálvez-Ruiz P, Bohórquez MR, Grimaldi-Puyana M, Cepeda-Carrión I. The relationship between technological capabilities and organizational impact: direct and indirect routes for employed and self-employed personal fitness trainers. Sustainability 2020 Dec 11;12(24):10383. [doi: 10.3390/SU122410383]

Abbreviations

MET: metabolic equivalent of task
PA: physical activity
PL: physical literacy
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
UNICEF: United Nations Children's Fund
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 30.08.23; peer-reviewed by T Baranowski, W Li, G Hayes; comments to author 04.04.24; revised version received 29.05.24; accepted 27.06.24; published 18.10.24.

Please cite as:
Watson-Mackie K, Arundell L, Lander N, McKay FH, Jerebine A, Venetsanou F, Barnett LM
Technology-Supported Physical Activity and Its Potential as a Tool to Promote Young Women's Physical Activity and Physical Literacy: Systematic Review
J Med Internet Res 2024;26:e52302
URL: https://www.jmir.org/2024/1/e52302
doi:10.2196/52302
PMID:39423006

©Kimberley Watson-Mackie, Lauren Arundell, Natalie Lander, Fiona H McKay, Alethea Jerebine, Fotini Venetsanou, Lisa M Barnett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Designing and Evaluating IT Applications for Informal Caregivers: Scoping Review

Shweta Premanandan[1], PhD; Awais Ahmad[2], MSc; Åsa Cajander[2], PhD; Pär Ågerfalk[1], PhD; Lisette van Gemert-Pijnen[3], PhD

[1]Department of Informatics and Media, Uppsala University, Uppsala, Sweden

[2]Division of Visual Information and Interaction, Department of Information Technology, Uppsala University, Uppsala, Sweden

[3]Department of Psychology, Health, and Technology, Faculty of Behavioral, Management and Social Sciences, University of Twente, Enschede, Netherlands

Corresponding Author:
Shweta Premanandan, PhD
Department of Informatics and Media
Uppsala University
Ekonomikum (Plan 3)
Kyrkogårdsgatan 10
Uppsala, 751 20
Sweden
Phone: 46 707689331
Email: shwetaprem@gmail.com

Abstract

Background: Informal caregivers, often family members or friends, play a crucial role in supporting individuals with chronic illnesses, disabilities, or age-related needs. However, the demands of caregiving can be overwhelming, leading to stress, burnout, and negative impacts on caregivers' well-being. IT applications have emerged as potential solutions to support informal caregivers, but their design and evaluation often lack a comprehensive understanding of caregivers' needs and preferences. By understanding caregivers' perspectives on these issues, this review aimed to inform the development of more effective and user-centered IT solutions that truly support caregivers' needs.

Objective: The purpose of this study was to conduct a scoping review to outline design recommendations for IT applications gathered from informal caregivers. In addition, this study presents evaluations of the use of IT applications by informal caregivers.

Methods: A five-step scoping review methodology was used to map relevant literature in the following manner: (1) research question identification, (2) relevant study identification, (3) selection of pertinent studies for review, (4) data charting from the selected literature, and (5) summarization and reporting of results. A structured search was conducted across the PubMed, Scopus, IEEE Xplore Digital Library, Web of Science, and ACM Digital Library databases. In addition, reference list hand searches and keyword searches in Google Scholar were undertaken. The inclusion criteria comprised research articles (journal and conference) focusing on IT applications tailored for informal caregivers, primarily qualitative studies. Two reviewers independently identified articles for review and extracted the data. Conflicts were resolved through discussion, with a third reviewer consulted if consensus could not be attained. Thematic analysis was used to analyze the data.

Results: A total of 82 articles were selected for the scoping review—44 (54%) related to design and 39 (48%) related to evaluation of IT applications. Six overarching themes emerged related to designing IT applications: information access, support systems, streamlining care-related tasks, fostering informal caregivers' well-being, communication with formal health care professionals, and task support. Furthermore, 3 broad themes emerged related to the evaluation of IT applications for informal caregivers: facilitators of using IT applications, barriers to using IT applications, and suggestions for improving IT applications for informal caregivers.

Conclusions: To our knowledge, this is the first study to map the literature on the design and evaluation of IT applications for informal caregivers. This scoping review outlines current practices and recommendations for designing and evaluating the use of IT applications for informal caregivers. It identifies 6 key design themes and 3 evaluation themes, offering valuable insights for future development in this field. These findings provide a road map for enhancing user-centric IT solutions in informal caregiver support technologies.

International Registered Report Identifier (IRRID): RR2-10.2196/47650

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57393) doi:10.2196/57393

KEYWORDS

burnout; caregiver; design guidelines; design recommendations; evaluation; health care services; implementation; informal caregiver; long-term care; mobile app; facilitators; barriers; usability; work-life balance

Introduction

Background

The demand for informal caregiving is increasing due to reduced hospital and nursing home resources, an aging population, and rising disability rates [1]. Informal caregivers, hereafter referred to as caregivers, are friends or family members who offer care to individuals dealing with illness or disability [1,2]. Informal caregivers provide their relatives with essential care, including practical support and nursing care, saving society considerable expenses [3,4]. The motivations for providing informal care are often rooted in personal and relational factors, such as reciprocity, affection, and family values [5]. However, this role can significantly impact their employment, health, and family life [6]. Caregivers often lack information and support and have unmet needs for respite, financial assistance, and recognition [7,8]. Despite these challenges, informal caregivers are often the only option for care when formal services are not available [9]. However, many informal caregivers do not receive adequate support or assistance [10].

Some informal caregivers report positive experiences from their caregiving journey, including personal accomplishment, strengthened relationships, and personal growth [11,12]. These positive aspects are influenced by factors such as good medical counseling, family and friends' support, and effective cognitive-emotional regulation [11,12]. However, they also face challenges, such as emotional and social aspects, loneliness, and caregiving burdens [11,13,14]. The intensity of caregiving and the presence of other responsibilities can impact their happiness [15]. Moreover, female caregivers in particular often provide more intensive and complex care, leading to poor emotional health [16]. Despite these challenges, informal caregivers can experience personal growth and resilience, which may help them sustain their support over time [14]. However, there is a care gap that means that the need for informal caregivers is rapidly increasing but the pool of potential informal caregivers is shrinking [17]. The existence of this care gap highlights the need to explore IT applications to support caregivers.

IT Applications for Informal Caregivers

IT applications for informal caregivers are designed to provide support, resources, and assistance to individuals caring for family members or friends with chronic illnesses, disabilities, or aging-related issues. These applications can range from mobile apps to web-based platforms and aim to streamline the process of building an informal care team, provide educational information, help remember physicians' appointments, and coordinate care among all involved parties [18]. IT applications are beneficial for informal caregivers in various ways. They can provide access to information and support, reduce social isolation, and improve caregivers' and care recipients' quality of life [18-20]. Various IT applications have been developed to support informal caregivers' caregiving activities. These applications include mobile apps that provide information, resources, and solutions to common problems [19] and web-based platforms that offer knowledge about caring and facilitate the formation of support networks [21]. Smartphone apps have also been designed to deliver personalized health information to caregivers of patients with chronic illnesses [22]. However, deploying these solutions is not without challenges, including technology-related, organizational, socioeconomic, and ethical issues [23]. Information, comprehension, motivation, time available, perceived burden, and perceived informal caregiving competency have been identified as factors influencing technology adoption and use by informal caregivers [24]. Hence, it is recommended that designers involve informal caregivers in the design process, build their digital skills, and raise awareness of the potential benefits of IT solutions [23]. Despite these challenges, the potential of IT applications to support informal caregivers is notable, and further research is needed to explore their design and use on a broader scale [22,25]. There is a need to consolidate existing knowledge on design recommendations for IT applications for caregivers. This would offer designers a valuable starting point when developing new solutions, serving as a comprehensive resource. In addition, it would provide an overview of current evaluations of IT applications for informal caregivers, helping designers understand which features are valued and which are less effective from the caregivers' perspective.

There are some reviews in the field of IT applications for informal caregivers in the recent literature. Hassan [23] identified challenges in deploying information and communications technology solutions, including technological, organizational, socioeconomic, and ethical issues. Sala-González et al [19] found that mobile apps for caregivers often address their needs but the evidence of their effectiveness is limited. Furthermore, Romero-Mas et al [26] highlighted the potential of virtual communities of practice in supporting informal caregivers. Krick et al [27] emphasized the need for high-quality evaluations of digital technologies for caregivers. Hence, a review of current practices in designing and evaluating the use of IT applications for informal caregivers is crucial due to the increasing role of informal caregivers in health care and the potential of IT applications to support them [18]. Table 1 provides an overview of recent review studies in this area. Through our scoping review, we contribute to the existing literature on IT applications for informal caregivers by identifying and summarizing knowledge about design recommendations and evaluation of the use of IT applications to support informal caregivers.

XSL•FO
RenderX

Table 1. Overview of recent review studies.

Study, year	Review study aims	Articles reviewed, N
Hassan [23], 2020	Identify challenges in deploying ICT[a] solutions, including technological, organizational, socioeconomic, and ethical issues	31
Sala-González et al [19], 2021	Evaluate current mobile apps for informal caregivers and assess whether they were developed considering the needs of caregiver users	11
Romero-Mas et al [26], 2020	Highlight the potential of virtual communities of practice in supporting informal caregivers	46
Krick et al [27], 2019	Review digital technologies in both informal and formal care settings focusing on their AEE[b]	715
Martínez-Alcalá et al [28], 2016	Identify the opportunities that ICT offers to health services, specifically for patients with dementia and their families	26
Powell et al [29], 2008	Understand the effectiveness of networked ICT interventions in supporting carers of people with dementia	15
Guessi Margarido et al [22], 2022	Investigate the nature and extent of native smartphone apps for informal caregivers of patients with chronic illnesses	36
Meyer et al [30], 2018	Conducted a systematic review of novel technology as a platform for interventions for caregivers and individuals with severe mental illnesses	11

[a]ICT: information and communications technology.

[b]AEE: acceptance, effectiveness, and efficiency.

Objectives

The purpose of this study was to conduct a scoping review of current practices and recommendations for designing and evaluating the use of IT applications by informal caregivers. Design recommendations for IT applications are best practices and guidelines that designers can use to create user-friendly, efficient, and effective applications. For instance, a design recommendation could involve providing support groups to informal caregivers, ensuring that the application can provide an online forum or space for caregivers to share their experiences and have a sense of community.

We also provide a summary of evaluations of the use of IT applications for informal caregivers. Although many applications for caregivers are designed based on identified needs through interviews and focus groups, the existing literature on this topic has not been compiled comprehensively. The research questions (RQs) that this scoping review addressed were as follows:

1. What are the design recommendations for developing IT applications for informal caregivers? (RQ 1)
2. How is the use of IT applications evaluated by informal caregivers? (RQ 2)

Methods

Overview

The methodological framework by Arksey and O'Malley [31] for scoping reviews was followed for this review. The framework outlines the following five stages: (1) identification of the RQ; (2) identification of relevant studies; (3) selection of relevant studies; (4) charting the data from the selected literature; and (5) collating, summarizing, and reporting the results. Reporting follows the guidelines defined by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Multimedia Appendix 1) [32] checklist. Full details of the methods can be found in our published protocol [18]. There were minor updates from the protocol. The second author conducted the full-text review for a portion of the study. The published protocol originally reported the number of articles from the initial database searches in 2021. However, a subsequent database review was conducted in 2023 along with manual searches. These updated numbers are now reflected in the flowchart.

Selection Criteria

The literature search was carried out with the support of health psychology information specialists from the University of Twente, and we also had consultations with university librarians at Uppsala University. The selection criteria were defined based on the population, intervention, comparator, outcome, and study design (PICOS) approach [33].

Informal adult caregivers, typically family members, partners, or friends aged >18 years, provide unpaid care for individuals with long-term illnesses or disabilities. This review included studies focusing on informal caregivers and excluded those focused on care recipients, formal caregivers, or health care professionals. The interventions considered were IT applications aimed at informal caregivers, ranging from mobile apps to web platforms providing assistance, resources, and support. Studies on applications targeting both care recipients and informal caregivers were included if the caregivers' role was substantial. IT applications designed for care recipients but influenced by informal caregivers' opinions were excluded.

As this review did not compare studies with other treatments, a comparator was not applicable. The included studies focused on designing or evaluating IT applications for caregivers or both. Design-related papers covered caregivers' needs and preferences for web-based applications or interventions, whereas evaluation papers addressed usability, evaluation, assessment, feasibility, and user experience. All study designs were eligible except descriptive or quantitative studies on system adoption

XSL•FO
RenderX

based on theoretical models. Studies on IT applications targeting patients, even if developed with input from caregivers, patients, and health care professionals, were also excluded.

During the screening, the reviewers used the following inclusion and exclusion criteria. The inclusion criteria were (1) conference articles, journal articles, and early-access articles; (2) studies on usability tests of IT applications or platforms; (3) studies describing user needs or requirements for IT applications for informal caregiving; (4) studies that described the evaluation of IT applications focusing on what works or what is preferred, along with user experience; (5) studies describing the development process of IT applications for informal caregiving; and (6) studies on IT applications to train informal caregivers in their caregiving activities. The exclusion criteria were (1) review articles, editorial and opinion papers, news articles, books and book chapters, theses, and clinical case papers; (2) descriptive or quantitative studies assessing the adoption of a system or criteria for adoption based on a theoretical model; (3) studies focused on care recipients and formal caregivers; and (4) studies on IT applications targeted at patients but developed based on interviews with patients and informal caregivers along with health care professionals.

Search Strategy

Electronic database searches were conducted in PubMed, the IEEE Xplore and ACM digital libraries, Scopus, and Web of Science on metadata such as titles, abstracts, and keywords.

The literature search strategy was constructed based on terms related to the following PICOS criteria: (1) informal caregivers (eg, *caregiver*, *spouse*, and *partner*), (2) IT applications (eg, *internet*, *app*, or *eHealth*), and (3) design and evaluation (eg, *design*, *evaluation*, *needs*, *usability*, or *experience*). Librarians were consulted when constructing the search string. Textbox 1 provides an overview of the different groups of keywords. There were no date restrictions. Complete search strategy is available in Multimedia Appendix 2.

Textbox 1. Keywords for the search string.

Group 1: informal caregivers

"Home nursing" OR "informal care*" OR "family care*"

Group 2: IT solutions

"Mobile application*" OR "ICT solution*" OR "ICT" OR eHealth OR "e-coaching system*" OR "*coaching system*" OR "digital solution*" OR "IT solution*" OR "internet-based interventions*" OR telehealth*

Group 3: design or evaluation

Design OR evaluation OR effectiveness OR usability OR requirements OR needs OR perspective OR "user experience*"

The secondary search strategy included conducting forward and backward citation screening for all included studies. In addition, we conducted manual searches in Google Scholar using keywords. The Google Scholar search was conducted using the keyword field in Publish or Perish (version 8.9.4538) [34]. Multiple combinations using keywords from each search group were used for the Google Scholar search on metadata, and 4 combinations yielded the most relevant results. Each of the following search strings was searched separately, and 100 hits were collected from each search: (1) *family carer design OR evaluation "internet intervention,"* (2) *family carer design OR evaluation ICT*, (3) *family carer design OR evaluation "digital intervention,"* and (4) *family carer design OR evaluation application*.

Study Selection

The initial review process (as seen in the protocol [18]) was carried out in November 2021. Search results were first deduplicated [35] and stored using EndNote X9 (Clarivate Analytics) [36]. A total of 428 deduplicated records were imported into Rayyan (Rayyan Systems Inc) [37], where title and abstract screening was carried out by 1 reviewer (SP). Full-text screening of the initial results was conducted independently by 2 reviewers (SP and AA) using the established criteria. Conflicts during the screening process were resolved through discussion, and a third reviewer (ÅC) was consulted as needed. Records retrieved from an updated search for papers published between October 2021 and November 2023 (537 deduplicated records) were screened by 1 reviewer (SP). Secondary searches were conducted by a librarian. For these records, title and abstract screening and full-text review were conducted by 1 reviewer (SP). Abstracts, editorial and opinion papers, news articles, books and book chapters, theses, and clinical case papers were excluded as these sources lacked rigor. Reviews and protocols were not included; however, published results of relevant protocols were sought. In addition, unpublished results of relevant protocols were sought from the authors if published results were not yet available.

Data Extraction

Details from the included records, such as authors, publication year, study purpose, study location, study design, and condition cared for, were extracted by the first reviewer (SP) using a Microsoft Excel spreadsheet (version 2021; Microsoft Corp) and confirmed to be accurate and complete by the second reviewer (AA). Each record was classified as either a design or evaluation study. The articles were imported into the NVivo software (version 12; Lumivero). A table with a summary of the included articles can be found in Multimedia Appendix 3 [8,38-118].

Data Analysis

This study used the NVivo software (version 12) to gather and systematically structure study data. Full-text articles were imported as PDF files into NVivo for data extraction, analysis, and coding. Analysis followed an iterative method involving combining, categorizing, summarizing, and comparing information across studies. Relevant data were identified and

XSL•FO
RenderX

categorized under broader emergent themes related to each RQ. As Braun and Clarke [119] outlined, inductive thematic analysis was applied to analyze texts from the included studies. Identifiable design recommendations and perceptions of evaluations of the use of IT applications by informal caregivers were extracted from analyzing the included articles, forming overarching themes. Any design recommendations and evaluations of use not fitting into established themes were iteratively added as new themes [120,121]. Two reviewers (SP and AA) conducted the formal thematic analysis of the included articles involving coding and identifying themes. Afterward, they met to discuss their findings, and through this discussion, finalized the themes. As their initial themes were largely aligned, the discussion helped solidify the results, ensuring the trustworthiness and validity of the findings. An expert reviewer (ÅC) with extensive experience in this area performed a final review of the codes and themes. The expert also reviewed the list of included articles to ensure that no significant studies were overlooked.

Results

Overview

The database searches yielded 1424 records (Figure 1). After duplicate records were removed (459/1424, 32.23%), titles and abstracts (965/1424, 67.77%) were screened before full texts (203/965, 21%) were retrieved for eligibility screening. A total of 29 of the included reports were identified using secondary search strategies, namely, reference tracking (n=20, 69%) and manual keyword searches in Google Scholar (n=9, 31%). In total, 82 reports were included in the study.

First, results for designing IT applications, which addresses the first RQ, are presented. Of the 82 included articles, this study reviewed 44 (54%) with design recommendations for IT applications supporting informal caregivers. A total of 6 main themes emerged: optimizing information access, support systems for informal caregivers, streamlining care-related tasks for informal caregivers, fostering informal caregivers' well-being, communication with formal health care professionals, and task support. Most of the included studies on design used qualitative methods, including interviews (20/44, 45%) and focus groups (10/44, 23%) with caregivers, whereas some also used mixed methods (5/44, 11%). Some of the included design studies (14/44, 32%) were focused on dementia caregivers, but we also found a mix of informal caregivers for cancer (5/44, 11%), older adults (8/44, 18%), and children with medical complexities (4/44, 9%).

Thereafter, the findings of evaluations of use of IT applications, which addresses the second RQ, are presented. The analysis of 48% (39/82) of the studies revealed 3 main themes: facilitators of using IT applications, barriers to using IT applications, and suggestions to improve IT applications. All studies indicated that IT applications had a positive reception among informal caregivers. Most evaluations (28/39, 72%) were conducted through qualitative interviews, with some using mixed methods (11/39, 28%). Qualitative approaches provided detailed insights into caregivers' needs, whereas some medical studies (4/39, 10%) used randomized controlled trials to assess the impact of IT applications on caregivers and patients. Although many studies (18/39, 46%) focused on mental impairments, there was a research gap concerning caregivers managing physical diseases, highlighting an area for future study.

XSL•FO
RenderX

Figure 1. Study selection flow diagram.

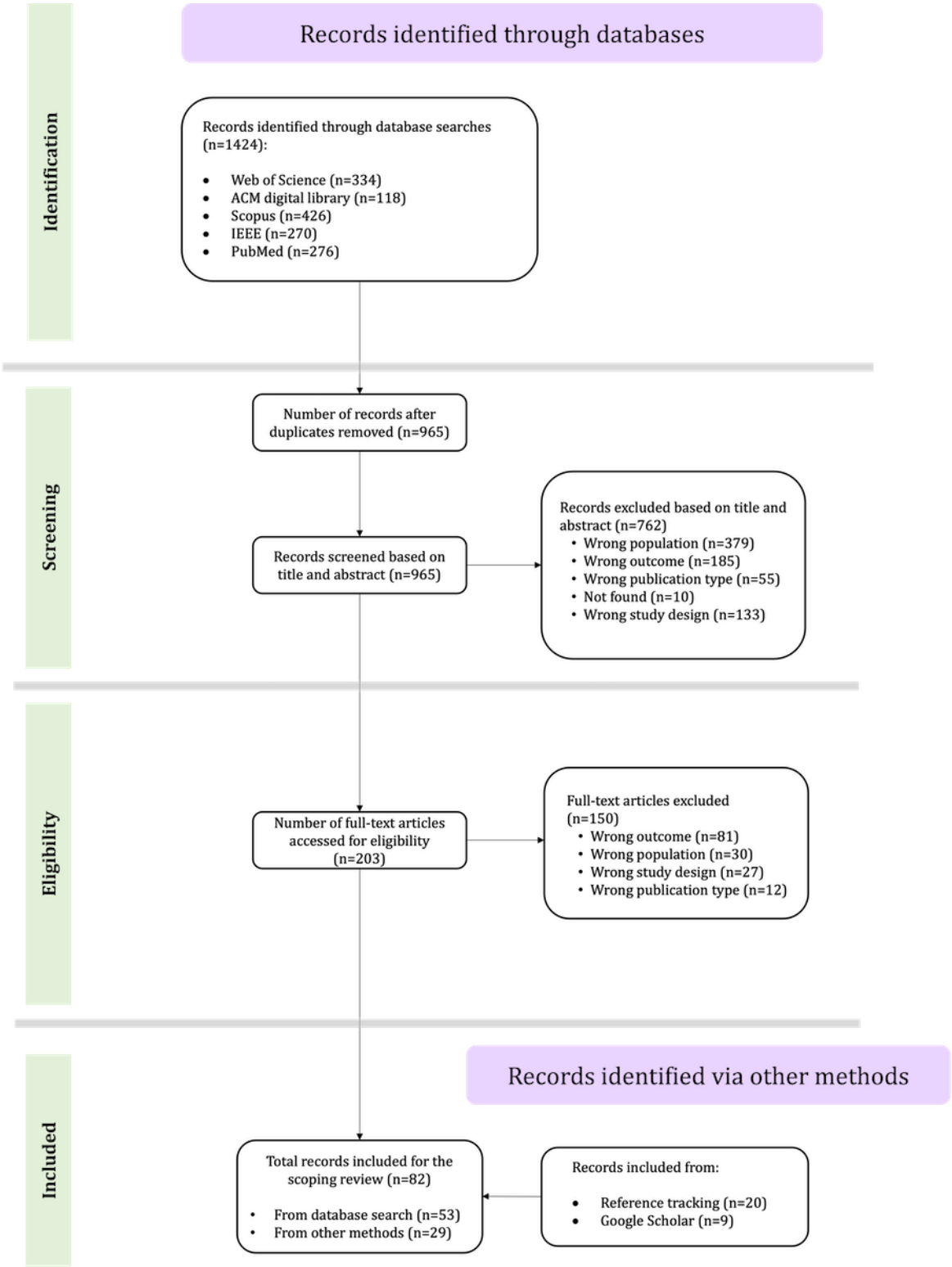

Design Recommendations for IT Applications

Overview

In this section, we present 6 themes and several subthemes of design recommendations for IT applications for informal caregivers. A summary of all the themes and their corresponding subthemes is presented in Table 2. Figure 2 illustrates the number of articles in each theme and subtheme in the form of a bubble chart.

XSL•FO
RenderX

Table 2. Summary of themes and subthemes for the design recommendations for IT applications with references.

Theme and subtheme	References
Optimizing information access and utility	
Access to information and resources	Premanandan et al [8], Rathnayake et al [38], Dickman Portz et al [42], Heynsbergh et al [45], Tixier and Lewkowicz [49], McHugh et al [53], Ahmad et al [57], Premanandan et al [58], Ciuffreda et al [60], Haji Mukhti et al [62], Fan et al [63], Egan et al [64], Köhle et al [68], Masterson-Algar et al [70], Moreno-Cámara et al [71], Sepehri et al [72], Siddiqui et al [74], Williamson et al [76], and Liverpool and Edbrooke-Childs [77]
Personalized information delivery	Nurgalieva et al [43], Hwang et al [54], Moberg et al [56], Premanandan et al [58], Egan et al [64], and Ducharme et al [79]
Content format and accessibility	Ahmad et al [57], Molinari-Ulate et al [59], Masterson-Algar et al [70], and Shreve et al [73]
Clear language	Molinari-Ulate et al [59], Masterson-Algar et al [70], and Xu et al [78]
Support systems	
Facilitating community and peer support	Rathnayake et al [38], Lobão et al [40], Leslie et al [41], Dickman Portz et al [42], Heynsbergh et al [45], Allemann et al [47], Tixier and Lewkowicz [49], Bosch and Kanis [51], Meiland et al [52], McHugh et al [53], Renyi et al [55], Ahmad et al [57], Premanandan et al [58], Molinari-Ulate et al [59], Fan et al [63], Hashemi et al [67], Köhle et al [68], Moreno-Cámara et al [71], Shreve et al [73], Siddiqui et al [74], Vaughan et al [75], and Xu et al [78]
Access to support networks	Nurgalieva et al [43], Heynsbergh et al [45], McHugh et al [53], Renyi et al [55], and Egan et al [64]
Providing reassurance through testimonials	Premanandan et al [8], McHugh et al [53], Ahmad et al [57], Premanandan et al [58], and Giroux et al [65]
Streamlining care-related tasks	Rathnayake et al [38], Macaden et al [39], Lobão et al [40], Chaar et al [46], Meiland et al [52], Premanandan et al [58], Ciuffreda et al [60], Gris et al [61], Haji Mukhti et al [62], Egan et al [64], Gutierrez and Ochoa [66], Sepehri et al [72], Williamson et al [76], Liverpool and Edbrooke-Childs [77], and Xu et al [78]
Fostering caregivers' well-being	
Emotional and psychological support	Lederman et al [44], Heynsbergh et al [45], Chaar et al [46], Bosch and Kanis [51], Ahmad et al [57], Fan et al [63], Hashemi et al [67], Moreno-Cámara et al [71], Sepehri et al [72], Shreve et al [73], Xu et al [78], and Ducharme et al [79]
Physical health and lifestyle management	Rathnayake et al [38], Lederman et al [44], Heynsbergh et al [45], Chaar et al [46], McNaney et al [48], Fan et al [63], Hashemi et al [67], Moreno-Cámara et al [71], Sepehri et al [72], Shreve et al [73], Xu et al [78], and Ducharme et al [79]
Communication with formal health care professionals	
Information exchange and resource access	Heynsbergh et al [45], Allemann et al [47], Schorch et al [50], Premanandan et al [58], Molinari-Ulate et al [59], Fan et al [63], and Giroux et al [65]
Formal support and service coordination	Premanandan et al [58], Moreno-Cámara et al [71], Shreve et al [73], and Ducharme et al [79]
Task support	Premanandan et al [8], Meiland et al [52], Moberg et al [56], Molinari, Ulate et al [59], Fan et al [63], Egan et al [64], Gutierrez and Ochoa [66], and Masterson-Algar et al [70]

XSL•FO
RenderX

Figure 2. Bubble chart illustrating the number of articles in each theme and subtheme for the design recommendations for IT applications.

Theme 1: Optimizing Information Access

Overview

This theme includes design recommendations with regard to providing information access to informal caregivers to help them in their caregiving role. A total of 59% (26/44) of the articles reviewed were classified under this theme, and most (19/44, 43%) referred to information access for informal caregivers. This includes education and training opportunities tailored to informal caregivers' needs, facilitating access to diverse resources and services, delivering personalized information, ensuring enhanced content formats and accessibility, and emphasizing clear language and clarity in communication.

Access to Information and Resources

In the literature reviewed, IT applications provided access to training opportunities to assist informal caregivers in developing their skills for effective caregiving [38,71]. Tailored resources, such as educational materials, classes, and programs were integrated to offer practical support and guidance [63,64]. Some applications also included provisions to facilitate advanced care planning as a crucial element in caregivers' routines [42]. In addition, these applications facilitated access to peer-reviewed research, recognizing the importance of evidence-based insights and strategies for caregivers [63,64]. IT applications for informal caregivers had quick access to relevant links and references [45,53,62,68,70,72,74]. They facilitated the establishment of a common knowledge base among caregivers, fostering a shared

XSL•FO
RenderX

understanding and effective care strategies [49,60]. In addition, these applications offered guidance on accessing formal services such as professional caregivers or housekeeping [63], acknowledging the importance of aiding informal caregivers in navigating available resources. Some IT applications also reported providing access to medical health records [76], which is important for informal caregivers in delivering comprehensive care. A few of these applications also had tailored web-based resources [53,62,72,77] specifically aimed at informal caregivers. These web-based resources may include materials such as information articles, guides, tutorials, or forums that provide information, tips, and assistance relevant to the needs and challenges faced by informal caregivers.

Personalized Information Delivery

IT applications also prioritized delivering tailored information specific to caregivers' unique situations [43,64]. Customizing information for caregivers consistently assisted them in their daily caregiving responsibilities [54]. In addition, these applications featured relevant personalized information addressing caregivers' individual needs and challenges [43,56,64,79]. This tailored approach regarding IT applications emphasizes that caregivers prefer personalized over generic information. For instance, caregivers can benefit from personalized information based on the condition that they provide care for, their relationship with the relative, the stage of caregiving, location, and more. Furthermore, the emphasis on personalized information within these applications is indicative of the diverse and ever-changing nature of caregiving responsibilities. This is because the type of information and support may change based on the progression of the condition for which care is provided.

Content Format and Accessibility

Various articles reviewed (4/44, 9%) discussed IT applications that emphasized delivering content to informal caregivers in varied formats, including video, audio, images, and snippets [59]. In addition, content in short text format was particularly valuable for informal caregivers as they preferred short and easily understandable information [70]. Moreover, the studies stressed the importance of illustrations and appropriate visuals [70] in aiding comprehension and engagement among informal caregivers. The dynamic information needs of informal caregivers require diverse content formats to accommodate their evolving requirements, ensuring accessibility and relevance to their caregiving tasks [73].

Clear Language

Numerous studies (3/44, 7%) stressed the importance of using good, easy, and appropriate language in content targeted at informal caregivers [59,70]. Simplicity and clarity are key, along with the need to avoid jargon that might hinder effective comprehension of information by caregivers [59,70]. Furthermore, the studies also indicated that a few applications focused on debunking illness myths [78], which is important for shaping caregivers' understanding. In addition, caregivers actively sought knowledge and a comprehensive history of the illnesses they provide care for [78].

Theme 2: Support Systems for Informal Caregivers

This theme presents recommendations that pertain to providing support structures for informal caregivers in the form of online forums, peer support, and social support. It also includes the provision of testimonials.

Facilitating Community and Peer Support

The functionalities required by caregivers from IT applications to facilitate community and peer support involved a diverse set of tools aimed at fostering connections and shared experiences among informal caregivers. In the 59% (26/44) of reviewed articles that were classified under this theme, these functionalities prioritized creating a sense of community and shared experiences [57,58], allowing caregivers to relate to and connect with others in similar situations. Peer support is crucial, enabling caregivers to gain insights and emotional reinforcement from their peers [41,42,53,58].

Moreover, the granularity of the information exchanged should also be considered, allowing caregivers to exchange comprehensive experiences and connect with peers [49]. Enabling the sharing of experiences among caregivers becomes vital for building social ties [38,40,47,55,68,74] and solidarity within the caregiving network [51-53]. Feedback from peer groups becomes an invaluable resource, providing diverse perspectives and guidance to caregivers facing similar challenges [59,63,67,71,73,75,78].

Engagement in dedicated caregiver groups [75] and the provision of virtual meeting spaces [57,75] were important functionalities, offering platforms for interactions and support irrespective of geographical limitations. This engagement could be in terms of reading other caregivers' experiences, asking questions, answering questions, venting their frustrations, or just speaking with other caregivers in similar situations. In addition, facilitating face-to-face group engagements [45] and providing access to online forums [57,58] such as Facebook groups satisfies varying preferences, allowing caregivers to seek and extend support.

Access to Support Networks

Improving access to support networks is crucial for the well-being of informal caregivers.

The capability to reach or access a care network is essential as it allows caregivers to connect with relevant support systems [43,53]. Finding peers and gaining access to social contacts of fellow caregivers enable the expansion of one's support network, fostering connections and shared experiences [45].

The inclusion of a Pinboard feature serves as a valuable resource, allowing caregivers to gather and share pertinent information and resources [55]. Access to social media and messaging platforms becomes crucial, providing channels for communication and information exchange among caregivers [64]. In addition, having access to an interaction book [72] offers a centralized space for storing and sharing essential caregiving-related information.

XSL•FO
RenderX

Providing Reassurance Through Testimonials

This recommendation of providing reassurance through testimonials is crucial as it serves as a means to offer confidence and affirmation to caregivers in their caregiving journey. IT applications provide caregivers with reassurance, instilling a sense of trust and confidence [53]. In addition, providing a space for caregivers to incorporate useful testimonials regarding caregiving experiences becomes essential as it can offer insights and real-life experiences to guide fellow caregivers regarding available support services or useful tips [65].

Theme 3: Streamlining Care-Related Tasks for Informal Caregivers

This theme emphasizes design recommendations specifically to assist with caregiving tasks. It includes features for monitoring medication, tracking the care recipient's condition, and streamlining care coordination. A total of 34% (15/44) of the articles reviewed were classified under this theme. In addition, these designs include comprehensive support for daily tasks, offering tailored assistance that caters to the unique needs of caregivers and care recipients.

Some IT applications had a provision for caregivers to record and document caregiver problems [38]. This feature can facilitate effective communication between caregivers and health care professionals, ensuring that the needs and concerns of the caregivers are adequately addressed. Effective illness management is a critical caregiving task, with IT applications offering features that assist caregivers in handling the health needs of their care recipients [38]. This may include tools for tracking symptoms, medication management, and accessing relevant health information. Continuously monitoring and tracking health parameters becomes integral for assessing progress and adjusting interventions as needed, ensuring the best health outcomes for the person under care [39,52,58,60,78]. Many applications also provided support for daily activities and access to services and facilities that can aid caregivers in their responsibilities [38]. This may encompass scheduling assistance, access to community resources, and information about support services available to caregivers and care recipients. Incorporating information and resources related to using care equipment was another useful application feature [38]. This may involve guidance on selecting and using appropriate care equipment and information on where to access such equipment. Some applications also supported caregivers in effectively managing the care of their patients, providing tools for organizing and coordinating care tasks [40]. This may include features for care coordination, communication with health care providers, and managing appointments and treatments. Features related to appointment management, treatment tracking, and medication management were essential in IT applications for informal caregivers [40]. These features can help caregivers stay organized and ensure that care recipients receive medical attention and treatments. Some applications also supported caregivers in learning and adapting to new caregiving routines [46]. This may involve providing educational resources, tips for managing changing care needs, and guidance on adapting to evolving caregiving responsibilities. Incorporating features that provide support in emergencies was important for applications for informal caregivers [52]. This may include emergency contact information, guidance on handling emergencies, and access to emergency resources. Many applications also included provisions for managing various care activities, including household tasks, personal hygiene, mobility support, medical care, and care coordination [61]. These features can help caregivers effectively organize and manage their caregiving responsibilities. Some caregivers were involved in physiotherapy activities for themselves and their care recipients, and hence, those applications included resources and guidance on technical skills related to physiotherapy [62]. This may include instructional materials, exercise demonstrations, and information on physiotherapy techniques. Incorporating evidence-based activity plans can be beneficial for caregivers, and the IT applications also provided access to such resources [64]. This may involve evidence-based caregiving strategies, activity recommendations, and guidelines for promoting the well-being of caregivers and care recipients. IT applications also need to include features that allow caregivers to track the evolution of their roles and access contextual information and services accordingly [66]. This may involve documenting changes in caregiving responsibilities, accessing relevant support based on evolving needs, and staying informed about available resources. A caregiving timeline dashboard can be a valuable feature in applications, providing caregivers with a visual representation of caregiving activities and milestones [72,77]. This feature can help caregivers track and manage their caregiving responsibilities effectively. Incorporating features for tracking medications was essential in applications for informal caregivers [76]. This may involve medication schedules, medication administration reminders, and medication use tracking. IT applications should assist caregivers with comprehensive knowledge on medication use and potential side effects [78]. This may include medication information, dosage guidelines, and information on potential adverse reactions. In addition, incorporating resources and information on symptom management was also [78]. This may involve guidance on recognizing and managing symptoms, accessing relevant health care information, and seeking appropriate support.

Theme 4: Fostering Informal Caregivers' Well-Being

Overview

This theme presents design recommendations that can assist in fostering informal caregivers' well-being through IT applications, and 32% (14/44) of the articles reviewed were classified under this theme. There were 2 subthemes, namely, emotional and psychological support and physical health and lifestyle management. This theme presents various recommendations, which include physical, mental, emotional, and social support to ensure that caregivers receive comprehensive assistance and are better equipped to navigate the challenges of caregiving while ensuring their well-being.

Emotional and Psychological Support

Emotional and psychological support is essential for caregivers, which includes various well-being features. Focusing on assisting in enhancing informal caregivers' well-being is crucial, acknowledging their important role [51,57,67]. It is imperative that caregivers feel empowered and have autonomy to enable

XSL•FO
RenderX

them to navigate challenges independently and feel effective in their roles [44]. Accessing support opportunities and incorporating mindfulness and counseling was another feature provided in applications to help caregivers manage stress and emotional complexities [45]. The stress of caregiving requires strategies that balance the needs of both the caregiver and the care recipient [46,79]. In addition, some IT applications offered features like positive activity interaction, pleasant activity scheduling, gratitude journaling, and a "positive piggy bank," all of which can help promote well-being [46]. Encouraging a positive, solution-oriented approach fosters resilience, enabling adaptive coping mechanisms when faced with challenges [51,57]. Using tools such as a life journal becomes beneficial for reflection and self-awareness, aiding in emotional processing and personal growth [72]. Moreover, accessing emotional support through interactive IT applications ensures easy accessibility and engagement, creating a supportive environment [71,73,78]. Encouraging self-care practices acknowledges the importance of caregivers prioritizing their well-being [71]. In addition, integrating mental health support becomes crucial, recognizing and addressing the psychological aspects of caregiving stress [63].

Physical Health and Lifestyle Management

Physical health and lifestyle management include various components essential for caregivers' and care recipients' well-being. Incorporating entertainment such as music into IT applications becomes a creative way to boost engagement and enjoyment, contributing to caregivers' physical and emotional wellness [38,48,73]. Sharing medical and well-being data becomes crucial for transparency and collaboration between caregivers and care recipients, reducing misunderstandings and distress while aiding informed decision-making [43]. In addition, focusing on physical well-being through exercise, diet, and nutrition becomes fundamental for promoting healthy lifestyles among caregivers [64,76]. Acknowledging the significance of financial aid as part of health management addresses the broader socioeconomic factors influencing well-being [62,72,79]. Recognizing and appreciating efforts becomes a motivating factor, reinforcing positive behaviors and fostering a supportive environment for caregivers to maintain their well-being [8].

Theme 5: Communication With Formal Health Care Professionals

Overview

This theme emphasizes the significance of robust communication channels for informal caregivers with formal health care professionals, beginning with comprehensive information about health care facilities. A total of 23% (10/44) of the articles reviewed were classified under this theme. Equipping caregivers with details about available facilities and services aids informed decision-making regarding patient care. Quick access to relevant contacts and references ensures prompt assistance, whereas encouraging interactions between caregivers and health care professionals facilitates necessary guidance and support.

Information Exchange and Resource Access

Offering information about hospitals in IT applications provides caregivers essential knowledge, enabling them to take quick action during emergencies [45]. IT applications can be instrumental in sharing critical hospital information, such as location, phone number, and services offered. Caregivers were also interested in accessing hospital-specific navigation in the applications [45]. Providing quick references and details of health care contact persons in the applications can ensure quick assistance when required [45]. Facilitating communication with health care professionals through the applications establishes a crucial link between caregivers and expert advice, supporting informed decision-making [47]. IT applications can help coordinate caregiving and healthcare-related tasks for care recipients by streamlining caregiving responsibilities and improving the efficiency of care provision [50]. In addition, guidance in accessing formal services through IT applications bridges the gap between caregivers and available support systems, ensuring comprehensive care [58]. Caregivers can also benefit from timely feedback from health care professionals using the applications, enhancing the quality of care by incorporating expert advice and recommendations [59]. Leveraging referrals to specialized resource persons using IT applications assists in expanding the support network, providing access to specialized knowledge and guidance [65]. Transparent communication through the IT applications about the care recipient's health condition becomes crucial for informed decision-making [63]. In addition, guidance in finding appropriate care facilities and accessing facility information using IT applications can contribute to the overall well-being of care recipients [63].

Formal Support and Service Coordination

First, formal support, including caregiver assistance and respite care [71,73], provides essential relief and specialized aid to caregivers, ensuring uninterrupted and high-quality care while allowing caregivers crucial moments to relax and recuperate. In addition, receiving help at the right time holds immense significance, guaranteeing assistance precisely when it is most needed, thereby maximizing the efficiency of support services [79].

Theme 6: Task Support

A total of 18% (8/44) of the articles reviewed were classified under this theme. The functionalities that were crucial for caregivers in an IT application catering to their needs encompassed a variety of essential features. First, support for memory using calendar functions and reminders becomes crucial, providing organizational assistance to help caregivers manage their schedules and tasks effectively [52]. Integrating supportive dialogues within the system creates a user-friendly and empathetic environment, offering guidance and encouragement when necessary [56]. In addition, providing support for users unfamiliar with IT applications ensures accessibility for individuals with varying levels of technological proficiency [8]. A multi-platform format enhances accessibility, enabling users to use the application across different devices and platforms [59]. Task-based notifications and reminders serve as prompts for caregivers, aiding in task management and

XSL•FO
RenderX

ensuring that important activities are not overlooked [63,64]. Incorporating a search feature facilitates quick information retrieval, enabling users to find specific information efficiently. Furthermore, using persuasive strategies [58] within the application can motivate caregivers to engage more effectively with their caregiving tasks [66]. Integrating read-aloud options contributes to accessibility, allowing users with diverse needs or preferences to access information through auditory means [70].

Evaluation of Use of IT Applications Among Informal Caregivers

Overview

This section provides an overview of informal caregivers' perceptions and evaluations of the use of IT applications. The analysis of the selected studies focused on evaluating IT applications designed to support informal caregivers. In total, 3 major themes emerged: facilitators of using IT applications, barriers to using IT applications, and suggestions to improve IT applications. Figure 3 illustrates the number of articles in each theme and subtheme in the form of a bubble chart. A summary of all the themes and their corresponding subthemes is presented in Table 3.

Figure 3. Bubble chart illustrating the number of articles in each theme and subtheme for the evaluation of the use of IT applications.

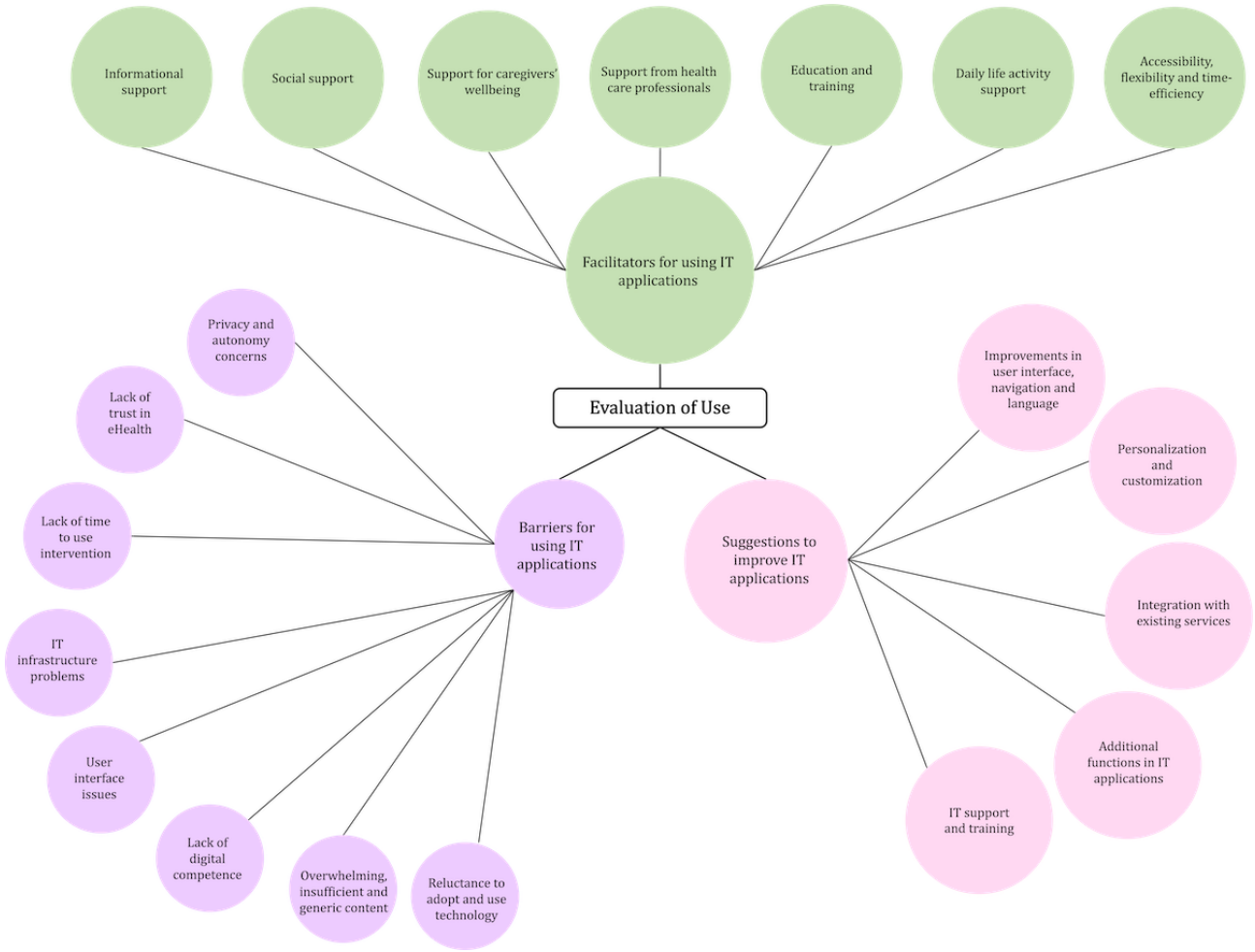

XSL•FO
RenderX

Table 3. Summary of themes and subthemes for the evaluation of the use of IT applications with references.

Theme and subtheme	References
Facilitators of using IT applications	
Informational support	Honary et al [117], Rettinger et al [115], Iribarren et al [113], Crotty et al [112], Tibell et al [108], Brown et al [111], Lam et al [105], Husebø [101], Chiu and Eysenbach [98], Andersson et al [93], and Ploeg et al [114]
Social support	Batchelor et al [116], Ploeg et al [114], Barbabella et al [110], Tibell et al [108], Kwok et al [107], Boutilier et al [106], Hermaszewska and Sin [102], Torp et al [99], Chiu and Eysenbach [98], Blusi et al [96], Lundberg [95], Andersson et al [93], Dam et al [87], Quinn et al [91], Honary et al [117], Blusi et al [97], and Rottenberg and Williams [103]
Support for caregivers' well-being	Quinn et al [91], Andersson et al [93], Torp et al [99], Andersson et al [100], Hermaszewska and Sin [102], Boutilier et al [106], Tibell et al [108], Austrom et al [109], Barbabella et al [110], Crotty et al [112], Ploeg et al [114], Batchelor et al [116], and Sin et al [118]
Support from health care professionals	Lundberg [95], Blusi et al [96], Boutilier et al [106], Austrom et al [109], and Iribarren et al [113]
Education and training	Wilding et al [88], Gomes et al [90], Brouns et al [92], Kajaks et al [94], Lundberg [95], Chiu and Eysenbach [98], Hermaszewska and Sin [102], Egan et al [104], Austrom et al [109], Brown et al [111], Ploeg et al [114], and Rettinger et al [115]
Daily-life activity support	Dam et al [87], Chiu and Eysenbach [98], Boutilier et al [106], Tibell et al [108], Brown et al [111], and Ploeg et al [114]
Accessibility, flexibility, and time efficiency	Lam et al [105], Andersson et al [93], Boutilier et al [106], Austrom et al [109], Brouns et al [92], Blusi et al [96], Kajaks et al [94], and Teles et al [89]
Barriers to using IT applications	
Privacy and autonomy concerns	Dam et al [87], Brouns et al [92], Blusi et al [96], and Honary et al [117]
Lack of trust in eHealth	Teles et al [89], Barbabella et al [110], and Honary et al [117]
Lack of time to use the intervention	Brown et al [111] and Ploeg et al [114]
IT infrastructure problems	Xiao et al [83], Wilding et al [88], Brouns et al [92], Husebø [101], and Lam et al [105]
User interface issues	Costa Stutzel et al [82], Thompson et al [85], Wilding et al [88], Gomes et al [90], and Lam et al [105]
Lack of digital competence	Andersson et al [93] and Husebø [101]
Overwhelming, insufficient, and generic content	Blusi et al [96], Brown et al [111], and Crotty et al [112]
Reluctance to adopt and use technology	Wilding et al [88], Lam et al [105], and Tibell et al [108]
Suggestions to improve IT applications	
Improvements in user interface, navigation, and language	Renati et al [84], Teles et al [89], Gomes et al [90], Andersson et al [93], Husebø [101], Egan et al [104], Boutilier et al [106], Kwok et al [107], Tibell et al [108], Crotty et al [112], Ploeg et al [114], Honary et al [117], and Sin et al [118]
Personalization and customization	Cheng et al [81], Renati et al [84], Teles et al [89], Andersson et al [93], Egan et al [104], Boutilier et al [106], Crotty et al [112], Ploeg et al [114], and Honary et al [117]
Integration with existing services	Brown et al [111] and Batchelor et al [116]
Additional functions in IT applications	Gomes et al [90], Blusi et al [96], Lam et al [105], Boutilier et al [106], Kwok et al [107], Austrom et al [109], and Ploeg et al [114]
IT support and training	Wan et al [80], Xiao et al [83], Blusi et al [96], Husebø [101], Barbabella et al [110], Brown et al [111], and Sin et al [118]

Theme 1: Facilitators of Using IT Applications

Overview

Facilitators of using IT applications emerged as one of the major themes. In this section, we present the facilitators of using IT applications by informal caregivers. A total of 95% (37/39) of the articles reviewed were classified under this theme.

Informational Support

Several studies (11/39, 28%) highlighted the usefulness of the information provided by IT applications in assisting caregivers in patient care [93,98,101,105,108,111-115,117]. The studies showed that the information was delivered through various formats, including videos, PDFs, and web pages, covering topics such as patient care, external resources, and the health care system. Caregivers emphasized the value of information on exercise, diet, and medication management, such as handling

XSL•FO
RenderX

elevated blood sugar with an endocrinologist's help [113]. Dementia caregivers appreciated receiving timely, customized information, which helped them understand and cope with illness-related changes and plan for future treatments [114]. The information modules provided quick, accessible advice when needed [115]. Technology-savvy caregivers expressed the need for increased web-based access to accommodate their schedules [113]. Caregivers also valued information on supportive resources, including local support groups and national charities [117]. These resources offered crucial details on medication management and legal rights. Navigating complex medical insurance was a significant challenge, requiring additional support [113]. Another study highlighted the importance of resource links, such as power of attorney information, to help caregivers manage new responsibilities such as financial and health care decisions [114].

Social Support

In several studies (17/39, 43%), caregivers consistently highlighted the usefulness and significance of the social support received through various forms in different IT applications [87,91,93,95-99,102,103,106-108,110,114,116,117]. Caregivers appreciated web-based platforms' interactive services, which connected them with others facing similar challenges [110]. They recommended improving the web-based community-building process to enhance interaction and create a more engaging social environment. Social interaction with those facing similar caregiving challenges was valued as it allowed them to share experiences and find mutual support. Many caregivers found comfort in personal conversations even when problems could not be solved, knowing that they were not alone [117]. While caregivers liked the discussion forums, they suggested promoting popular posts and adding open-ended boards and features such as small chat rooms or focused caregiver cohorts [107].

Support for Caregivers' Well-Being

Caregivers acknowledged the importance of support to improve their well-being, recognizing that the stress associated with caregiving often led to their compromised health and quality of life [91,93,99,100,102,106,108-110,112,114,116,118]. Caregivers acknowledged that the stress of caregiving contributed to physical and psychological issues, affecting their ability to provide care [112]. IT applications encouraged self-reflection, exploration of feelings, care strategies, and goal setting [106,110]. Several studies (6/39, 15%) highlighted that caregivers viewed IT applications as beneficial for psychological well-being, promoting empowerment, competence, and validation [91,93,99,100,114,116]. Using IT applications may also reinforce a sense of competence among working caregivers, helping them meet caregiving demands and see the positive aspects of their situation [93].

Support From Health Care Professionals

The possibility of communicating and receiving support from health care providers was highly valued in IT applications [95,96,106,109,113]. The studies indicated that caregivers needed IT applications to include professional support from physicians and nurses along with easy web-based appointment scheduling for tailored health information [110]. They often requested improved communication with health care providers to seek reassurance on patient care decisions [113]. Some IT applications offered videoconferencing with other caregivers and health care professionals, which was highly appreciated [95,109]. Caregivers had positive views on group and video conferencing for support and communication, with studies noting improvements in anxiety, depression, and physical health as a result of these interactions with health care professionals [109].

Education and Training

In several IT applications, caregivers found the education and training provided to take care of patients highly useful and helpful for their caregiving responsibilities [88,90,92,94,95,98,102,104,109,111,114,115]. Education and training were provided to caregivers in various forms across different studies. An interesting example involved creating different activities to engage both caregivers and care recipients [115]. Caregivers felt motivated to try out new activities in the applications and adapt them to the personal preferences and possibilities of the relatives they cared for. Participants engaged in activities such as playing ball, watching the news together, decorating the house for Easter, wrapping wool, or looking at photo albums. In addition, most participants expressed satisfaction with the way in which the activities were displayed and described [115].

Daily-Life Activity Support

Caregivers participating in various studies found valuable information and functions that aided them in performing daily-life activities [87,98,106,108,111,114]. For instance, a daily care activity checklist was perceived as useful [106]. One study highlighted the effectiveness of using a timeline to organize caregiving activities, with caregivers appreciating a dedicated care book that provided contact details and practical care insights [87]. In addition, information on coping strategies for patients with dementia was valuable for caregivers in managing daily activities and fostering better interactions [98].

Accessibility, Flexibility, and Time Efficiency

Caregivers found IT applications to be useful as they offered easily accessible, cost-effective, flexible, and time-efficient solutions. These applications gave caregivers quick access to useful features, ensured smooth functionality, and helped save time during their interactions with other caregivers or healthcare professionals [89,92-94,96,105]. Caregivers found that these IT applications were easy to use and flexible and aligned well with their daily-life activities [105]. One major benefit of IT applications was their accessibility at all times, allowing users to access them when needed and at their convenience in their busy schedules [92]. Caregivers were thankful for customer support as it significantly improved their ability to access and use the IT applications with greater ease and efficiency [106,109].

Theme 2: Barriers to Using IT Applications

Overview

The theme of barriers to using IT applications explores the challenges faced by informal caregivers in using IT applications

XSL•FO
RenderX

for their caregiving tasks. These barriers were privacy and autonomy concerns; lack of trust in IT applications; lack of time to use eHealth interventions; IT infrastructure problems; user interface issues; lack of digital competence; overwhelming, insufficient, and generic content; and reluctance to adopt and use technology. Understanding these barriers is crucial for enabling informal caregivers to effectively leverage IT applications to support their caregiving responsibilities and enhance the quality of care provided to their loved ones. A total of 49% (19/39) of the articles reviewed were classified under this theme.

Privacy and Autonomy Concerns

Various studies (4/39, 10%) highlighted caregivers' concerns about privacy and autonomy [87,92,96,117]. Caregivers expressed apprehension regarding the security of their information on web-based platforms and stressed the importance of increasing their privacy levels [87,117]. Patients viewed therapists' access to their data not as a breach of privacy but as a motivating factor. However, health care professionals identified internet connections lacking privacy assurances as a barrier to using a web-based platform, emphasizing the necessity for secure data transport [92].

Lack of Trust in IT Applications

Several studies (3/39, 8%) highlighted that the lack of trust in the credibility and accuracy of the information provided through IT applications negatively impacted caregivers' perceptions of their use [89,110,117]. There was extensive discussion about the crucial need for trust in information sources, emphasizing their reliability and legitimacy [89,117]. It is not just the provision of health-related information that matters—establishing trust and credibility is also crucial [117]. Another study identified factors such as a clear layout, interactive features, owner authority, ease of use, and quality content influencing trust in health programs. Caregivers specifically highlighted the importance of layout, owner authority, and content as elements promoting trust [89].

Lack of Time to Use the Intervention

Caregivers expressed a strong interest in using the system yet faced difficulties in finding the time to engage with it, attributed to the demanding nature of caregiving and additional responsibilities [111,114]. The challenge is particularly pronounced for overwhelmed caregivers, especially those caring for patients undergoing surgeries or coping with serious conditions, who find it daunting to enter information into the system [111].

IT Infrastructure Problems

The challenges in IT infrastructure, such as issues with internet accessibility and hardware, along with technical issues, were a factor hindering the use of IT applications [83,88,92,101,105]. One study highlighted the challenges posed by the lack of consistent internet access, impeding access to health information and engagement in video calls. In addition, hardware issues, particularly those associated with monitoring blood pressure at home using remote devices, were identified as obstacles to effective health monitoring [105]. The absence of reimbursement for IT applications and related infrastructure also posed a barrier for caregivers who could not afford them. Health care professionals cited the lack of reimbursement as a significant obstacle to program implementation due to the associated costs [92].

User Interface Issues

Several studies (5/39, 13%) highlighted the problems with user interfaces that affected caregivers' ease of using the system [82,85,88,90,105]. Caregivers faced several challenges in comprehending the options within the interface, including difficulties understanding specific features such as accessing and using the physical activity function, navigating the notification menu flow, handling alarms, and managing message notifications [82]. They emphasized the need for the interface to provide clear instructions on accessing and using certain functions, underscoring the importance of clarity in using the interface effectively [82,105].

Lack of Digital Competence

Many studies (5/39, 13%) emphasized that a lack of digital competence posed significant obstacles for users. These challenges included difficulties in using and navigating digital tools and technologies, struggles in accessing and using web-based information, limitations in effective communication through digital channels, and potential risks related to digital security [83,88,93,101,105]. Especially among older adults, challenges in using IT applications were observed, primarily attributed to insufficient digital competence [93].

Overwhelming, Insufficient, and Generic Content

The information provided in the IT applications was overwhelming, too generic, or irrelevant for caregivers, which was perceived as a significant barrier to using IT applications [96,111,112]. Caregivers noted that, especially at the time of diagnosis, they were overloaded with information yet paradoxically felt a need for a personalized search for information [112]. They acknowledged the positive impact of the support received but also underscored additional unmet support needs. Specifically, caregivers expressed receiving support not aligned with their specific requirements, encountering limitations in the availability of support, and feeling that the support provided was structured to meet the provider's needs rather than addressing the caregivers' unique needs [96].

Reluctance to Adopt and Use Technology

Reluctance to adopt and use technology was evident in studies in which caregivers expressed hesitation toward incorporating new technological advancements into their daily routines. In addition, there was apprehension about using the chat forum stemming from previous negative experiences [108]. Concerns included the potential for the chat to focus excessively on individual issues or lead to cyberbullying based on caregivers' past encounters with other social media platforms [108].

Theme 3: Suggestions to Improve IT Applications

Overview

This theme explores the proposed improvements to and adaptations of IT applications aimed at addressing the specific needs and challenges faced by informal caregivers. These are

XSL•FO
RenderX

improvements in the user interface, navigation, and language; personalization and customization; integration with existing IT services; additional functions in IT applications; and IT support and training. A total of 56% (22/39) of the articles reviewed were classified under in this theme.

Improvements in User Interface, Navigation, and Language

In several studies (12/39, 31%), enhancements to the user interface, navigation, and language were suggested for system improvement [84,89,90,93,101,104,106,107,112,114,117,118]. This involved maintaining consistent and commonly used menu icons, ensuring uniform language across the system for effective communication with end users [85,118], and implementing plain icons with hover-over features to enhance system responsiveness [104]. These adjustments aimed to improve user experience and make the system more intuitive and accessible. The caregivers also suggested adding visual aids, warnings for potentially distressing content, and features such as frequently viewed buttons for easier navigation [116]. They also proposed rolling discussion topics and emoji reactions to enhance engagement with the forum [116]. Moreover, customization options were recommended for a more personalized user experience [106]. Caregivers recommended incorporating introductory text for created content, including videos and educational materials [108]. This serves to provide context and enhance the understanding of the content, offering users a more informed and engaging experience.

Personalization and Customization

Caregivers expressed a strong desire for personalized and customized IT applications that aligned with their specific needs and use context [81,84,89,93,104,106,112,114,117]. For web-based family support systems to be truly effective, it is essential to introduce them promptly and tailor them to the individual preferences of working caregivers, accommodating each caregiver's unique circumstances [93]. A persistent concern voiced by participants was the importance of maintaining a sense of ownership during caregiving, particularly when they felt that their caregiving role dominated other aspects of their lives, such as being a partner, parent, or sibling [117]. The suggestion was to shift the focus from designing solely for the caregiver as a user to exploring designs that consider caregivers with diverse roles and responsibilities, such as managing a career or caring for young children. The goal was to develop online support that enables caregivers to preserve their identity, emphasizing the broader spectrum of roles they fulfill rather than exclusively focusing on their caregiving responsibilities [117].

Integration With Existing IT Services

Caregivers and patients use various IT services, including health care journals, throughout the caregiving journey. The caregivers expressed a desire for seamless integration between existing services and the proposed ones [111,116]. Specifically, linking CareHeroes (a proposed IT application) with other telehealth technologies, such as monitoring blood pressure, blood sugar, and mobility, was recommended for better usability [111]. This integration was anticipated to substantially amplify the platform's potential impact, creating a more comprehensive and interconnected health care ecosystem [111].

Additional Functions in IT Applications

Several studies (3/39, 8%) suggested additional features for IT applications. Caregivers noted the need for multiple accounts to allow various family members and friends to participate in caregiving through the application, enabling a collaborative approach [90,105,106]. They also requested notifications and reminders for tasks but were concerned that automated reminders might feel "generic" and overwhelming. Customizable, personalized reminders were recommended to improve the user experience [106,107]. Caregivers suggested customizing messages in reminders, allowing for personalized and context-specific notifications [107]. Caregivers also struggled to find information within some applications, prompting the suggestion of a search function to enhance content accessibility and navigation [96,109,114].

IT Support and Training

Caregivers strongly desired easily accessible IT support and training as they encountered various issues while using IT applications [80,83,96,101,110,111,118]. To overcome these issues, caregivers recommended providing education and training on computer and internet use coupled with technical support for working caregivers [93]. The proposal included the addition of a "support" section in the main menu, enabling participants to request technical or emotional support directly from the application [118]. To enhance understanding of IT applications, instructional videos were recommended to educate caregivers about specific challenging tasks [111]. Furthermore, digital skill training was emphasized, particularly for caregivers with less experience with web services [110]. These comprehensive measures aimed to empower caregivers with the necessary support and knowledge to navigate and use IT applications effectively.

Discussion

Principal Findings

In this study, of the 82 selected articles, we examined 44 (54%) that presented design recommendations for IT applications for informal caregivers. Six overarching themes emerged: optimizing information access, support systems for informal caregivers, streamlining care-related tasks for informal caregivers, fostering informal caregivers' well-being, communication with formal health care professionals, and task support. While these themes provide distinct categories, there are instances in which they overlap, indicating interconnections and complementary aspects among these identified areas. These design recommendations can contribute to further design of support applications for informal caregivers. Most of the included studies (44/82, 54%) on design used qualitative methods, including interviews and focus groups with caregivers, whereas some (5/44, 11%) also used mixed methods. Some of the included design studies (14/44, 32%) were focused on dementia caregivers, but we also found a mix of informal caregivers for cancer, older adults, stroke, and children with medical complexities. In this study, we also examined the perceptions and evaluations of IT applications among informal caregivers, analyzing 48% (39/82) of the selected studies in this category, which evaluated IT applications. In total, 3 broad

XSL•FO
RenderX

themes emerged: facilitators of using IT applications, barriers to using IT applications, and suggestions to improve IT applications for informal caregivers. Interestingly, all the studies uniformly indicated a positive reception among informal caregivers regarding the use of IT applications. Most studies (28/39, 72%) used qualitative interviews for the evaluation of IT applications, with some (11/39, 28%) using mixed methods. Qualitative approaches provided detailed insights into caregivers' contexts and preferences. Some studies from the medical field (4/39, 10%) used randomized controlled trials, which offered systematic assessments of the broader medical impacts of IT applications on both caregivers and patients. While many evaluation studies (18/39, 46%) were focused on mental impairments, a notable gap exists in research addressing caregivers dealing with various physical diseases, suggesting a potential area for future exploration.

One major strength of this review is its comprehensive search strategy, spanning diverse fields such as information systems, human-computer interaction, software development, health informatics, and health care. The identified facilitators of using IT applications, barriers to using IT applications, and improvements for IT applications for informal caregivers provide a foundation for further in-depth exploration. In total, 2 authors independently conducted the analysis, subsequently engaging in discussions to enhance the credibility of this review and reduce the risk of potential biases.

Designing IT Applications for Informal Caregivers

Tailored education, training, and accessible resources are important in assisting informal caregivers [68,122]. Providing this in an appropriate format may reduce the considerable burden on informal caregivers. Recent studies emphasize the importance of curated information, ensuring that caregivers access relevant and easily understandable content [123,124]. However, these information needs vary throughout the caregiving journey [125,126]. Integrating evidence-based insights into these applications is an essential aspect of supporting caregivers effectively [63,64].

Community and peer support play significant roles in caregivers' well-being [8,127]. Recent findings emphasize the value of shared experiences and connections among informal caregivers, providing emotional reinforcement and practical guidance [128,129]. The literature also suggests that feeling part of a community can help reduce the social isolation that many caregivers may experience [130]. Testimonials and shared experiences within caregiving communities foster confidence and a sense of solidarity among caregivers [131].

Efforts to streamline caregiving tasks encompass various dimensions crucial for effective support systems [132]. Coordinating appointments, treatments, and caregiving responsibilities has been identified as pivotal for optimizing care provision [133]. This process involves various activities, including communication, monitoring, and information sharing, and is influenced by patient-centeredness [134]. Effective care coordination interventions have been shown to benefit patients with specific conditions, particularly chronic conditions [135]. Informal caregivers play a crucial role in managing the health of older adults, particularly in medication management [136,137]. Monitoring technologies can enable caregivers to provide more care without increasing burden [138]. However, the impact of these changes on caregivers' health and well-being is poorly understood [1].

Recent studies have highlighted the potential of IT applications for supporting caregiver well-being by reducing caregiver burden and stress [139]. Lorca-Cabrera et al [140] and Lam and Lam [141] highlighted the positive effects of web-based interventions and internet use on caregiver well-being, including improved mental health and reduced anxiety and distress. The literature also indicates the importance of IT applications for improving patients' and their caregivers' quality of life [28,142]. These IT interventions are known to reduce caregivers' stress and depression [29].

Our findings suggest that informal caregivers benefit from communication with formal health care professionals. Effective communication between formal health care professionals and informal caregivers is crucial for the well-being of care recipients and the reduction of caregiver burden [143]. However, this communication is often hindered by a lack of clarity on the informational needs of caregivers [144], confusion about available services [145], and a lack of consideration for the views of informal caregivers [146]. The use of technology to mediate this communication has shown potential, particularly in aged care [43]. In the context of dementia care, formal caregivers face challenges in communication due to the impairments of care recipients [147]. Robust communication channels between caregivers and health care entities significantly impact caregiving experiences [147]. Enabling access to health care information, formal support coordination, and guidance from health care professionals is crucial for informed decision-making and ensuring optimal care provision.

Evaluation of IT Applications for Informal Caregivers

Our findings from the studies evaluating the use of IT applications indicate that informal caregivers indeed valued the usefulness of these applications. However, they voiced concerns about the ease of use associated with these applications, in line with previous research [23,148]. A recent scoping review on the evaluation of smartphone apps for informal caregivers of patients with chronic illnesses also indicated that IT applications could play a significant role in informal caregiving, and many of the apps found in this review provided new opportunities for caregivers to access health information [22]. In another systematic review focusing on the implementation of IT applications for informal caregivers of individuals with dementia, it was determined that these interventions have the potential to significantly improve the well-being of caregivers [149]. Furthermore, by empowering caregivers to offer more effective and sustained care to patients, these eHealth interventions hold promise for reducing future burdens on health care services [149]. The importance of caregivers' well-being was evident in our findings as well. Caregivers highly valued functions such as online forums, video meetings, and educational content. They highlighted how these features provided valuable information and support, contributing significantly to the enhancement of their psychological well-being.

XSL•FO
RenderX

Our findings showed that the most identified facilitators of using IT applications for caregivers were related to caregivers' individual needs for personalized information, social support, education, and training to support their caregiving responsibilities. The information, education, and training should be personalized according to the caregivers' situation, their specific caregiving needs, and the care recipient's medical condition. Personalization and customization demand a user-centered and holistic approach to designing IT applications; active and early involvement of users is necessary to create applications that are tailor-made according to their needs [23,150].

It is noteworthy that the facilitators of using IT applications for caregivers, as well as the suggestions for improvement, were in line with the design recommendations gathered in this study. This alignment suggests that caregivers appreciate the implementation of design recommendations in IT applications when they use them. Most of the design recommendations were derived from the potential users of these applications. This underscores the significance of adopting a user-centered approach in which involving users in the requirement-gathering process significantly enhances the usability of these applications [148,151].

Limitations and Future Research

Certain constraints related to this review necessitate consideration. Only 1 reviewer conducted the article identification, screening of titles, and data extraction for this review. Hence, there is a possibility that some relevant articles were not identified. Another limitation is that the review's broad nature limits the in-depth exploration of specific themes. Future research could focus on providing more detailed insights into the identified themes. Future research could also focus on a review that compares the IT application requirements of informal caregivers with those of formal caregivers. This could provide a broader context for researchers. While our review included a few articles that dated back a decade (10/82, 12%), the core design recommendations and evaluations of use by informal caregivers remained relatively stable over time. This is because the fundamental principles guiding caregiver support and design considerations have not drastically changed, and these studies still offered relevant insights into these principles. In addition, while adhering to a stringent scoping review structure, there remains a possibility of overlooking pertinent research, particularly when searching through an extensive array of evidence, including gray literature. The studies often lacked diversity. Many studies primarily included women, potentially excluding experiences of male caregivers. Where reported, studies often lacked representation of various ethnicities and socioeconomic backgrounds, suggesting that the findings might not apply equally to diverse caregiver populations. Due to the self-selecting nature of participants in many reviewed studies, the findings might not represent all family caregivers. This is because individuals who self-select are often more motivated and actively engage in using IT applications. Consequently, caregivers facing different challenges or lacking those specific motivations might have been unintentionally excluded. This limitation raises concerns about the generalizability of the review's conclusions to the broader caregiver population.

Conclusions

This review is the first to explore the design recommendations and evaluation of the use of IT applications for informal caregivers. It provides a summary of design recommendations to begin the design of an IT application for informal caregivers. It also provides a compilation of facilitators of using IT applications, barriers to using IT applications, and suggestions for improving IT applications for informal caregivers based on previous literature. This review was based on 82 articles, of which 44 (54%) were focused on design recommendations and 39 (48%) were focused on evaluations of the use of IT applications for informal caregivers. Six overarching themes emerged related to designing IT applications: information access, support systems, streamlining care-related tasks, fostering informal caregivers' well-being, communication with formal health care professionals, and task support. Furthermore, 3 broad themes emerged related to the evaluation of IT applications for informal caregivers: facilitators of using IT applications, barriers to using IT applications, and suggestions for improving IT applications for informal caregivers. These findings provide a road map for enhancing user-centric IT applications in informal caregiver support technologies.

Acknowledgments

The authors would like to thank Agnes Kotka and Görel Sundström (librarians at Uppsala University) and Carla Gerritsen and Roberto Cruz Martinez (information specialists, Faculty of Behavioral, Management, and Social Sciences, University of Twente) for assisting in the development of the search strategy and the manual search. This research was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant (814072) for the 4-year innovative training network ENTWINE for informal care. This work was also supported by the Swedish Research Council (2019-01231) and the Swedish Cancer Society (20 1014 PjF).

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors' Contributions

SP is the principal author of this manuscript. SP designed the review; identified the research questions; worked with the search strategy; conducted the title and abstract screening, complete full-text screening, and formal analysis; and wrote the manuscript.

XSL•FO
RenderX

AA conducted full-text screening of the initial results, conducted the formal analysis, and wrote parts of the manuscript. PÅ contributed to the study design. ÅC, PÅ, and LvG-P critically revised the manuscript drafts. All authors approved the final draft of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) fillable checklist.
[PDF File (Adobe PDF File), 519 KB - jmir_v26i1e57393_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 14 KB - jmir_v26i1e57393_app2.docx]

Multimedia Appendix 3
Summary table of included articles.
[DOCX File , 112 KB - jmir_v26i1e57393_app3.docx]

References

1. Herrman H, Singh B, Schofield H, Eastwood R, Burgess P, Lewis V, et al. The health and wellbeing of informal caregivers: a review and study program. Aust J Public Health 1993 Sep;17(3):261-266. [doi: 10.1111/j.1753-6405.1993.tb00146.x] [Medline: 8286501]
2. Morrison V, Zarzycki M, Vilchinsky N, Sanderman R, Lamura G, Fisher O, et al. A multinational longitudinal study incorporating intensive methods to examine caregiver experiences in the context of chronic health conditions: protocol of the ENTWINE-iCohort. Int J Environ Res Public Health 2022 Jan 12;19(2):821 [FREE Full text] [doi: 10.3390/ijerph19020821] [Medline: 35055644]
3. Decarlo P, Folkman S. Are informal caregivers important in AIDS care? Posit Outlook 1997;4(2):26-27. [Medline: 12293155]
4. Barber CF. Mental health and informal carers. Br J Ment Health Nurs 2018 Aug 02;7(4):158-162. [doi: 10.12968/bjmh.2018.7.4.158]
5. Zarzycki M, Seddon D, Bei E, Morrison V. Why do they care? A qualitative systematic review and meta-synthesis of personal and relational motivations for providing informal care. Health Psychol Rev 2023 Jun;17(2):344-376. [doi: 10.1080/17437199.2022.2058581] [Medline: 35383541]
6. Bauer JM, Sousa-Poza A. Impacts of informal caregiving on caregiver employment, health, and family. Population Ageing 2015 Feb 14;8(3):113-145. [doi: 10.1007/S12062-015-9116-0]
7. Dixe MD, da Conceição Teixeira LF, Areosa TJ, Frontini RC, de Jesus Almeida Peralta T, Querido AI. Needs and skills of informal caregivers to care for a dependent person: a cross-sectional study. BMC Geriatr 2019 Sep 18;19(1):255 [FREE Full text] [doi: 10.1186/s12877-019-1274-0] [Medline: 31533637]
8. Premanandan S, Ahmad A, Cajander Å, Ågerfalk P, Dolezel M, van Gemert-Pijnen L. Designing a mobile e-Coaching app for immigrant informal caregivers: qualitative study using the persuasive system design model. JMIR Mhealth Uhealth 2023 Nov 09;11:e50038 [FREE Full text] [doi: 10.2196/50038] [Medline: 37943598]
9. Hoffmann F, Rodrigues R. Informal carers: who takes care of them?*. Semantic Scholar. 2010. URL: https://www.semanticscholar.org/paper/Informal-Carers%3A-Who-Takes-Care-of-Them*-Hoffmann-Rodrigues/537f4ec02d9025edb94ecf58726cd7a7e9ea3138 [accessed 2023-12-08]
10. Donelan K, Hill CA, Hoffman C, Scoles K, Feldman PH, Levine C, et al. Challenged to care: informal caregivers in a changing health system. Health Aff (Millwood) 2002 Jul;21(4):222-231. [doi: 10.1377/hlthaff.21.4.222] [Medline: 12117133]
11. Lindeza P, Rodrigues M, Costa J, Guerreiro M, Rosa MM. Impact of dementia on informal care: a systematic review of family caregivers' perceptions. BMJ Support Palliat Care (Forthcoming) 2020 Oct 14. [doi: 10.1136/bmjspcare-2020-002242] [Medline: 33055092]
12. Yu DS, Cheng ST, Wang J. Unravelling positive aspects of caregiving in dementia: an integrative review of research literature. Int J Nurs Stud 2018 Mar;79:1-26. [doi: 10.1016/j.ijnurstu.2017.10.008] [Medline: 29128685]
13. Vasileiou K, Barnett J, Barreto M, Vines J, Atkinson M, Lawson S, et al. Experiences of loneliness associated with being an informal caregiver: a qualitative investigation. Front Psychol 2017;8:585 [FREE Full text] [doi: 10.3389/fpsyg.2017.00585] [Medline: 28469589]
14. Hoang VL, Green T, Bonner A. Informal caregivers' experiences of caring for people receiving dialysis: a mixed-methods systematic review. J Ren Care 2018 Jun;44(2):82-95. [doi: 10.1111/jorc.12235] [Medline: 29357407]

XSL•FO
RenderX

15. van Campen C, de Boer AH, Iedema J. Are informal caregivers less happy than noncaregivers? Happiness and the intensity of caregiving in combination with paid and voluntary work. Scand J Caring Sci 2013 Mar 10;27(1):44-50. [doi: 10.1111/j.1471-6712.2012.00998.x] [Medline: 22571683]
16. Navaie-Waliser M, Feldman PH, Gould DA, Levine C, Kuerbis AN, Donelan K. The experiences and challenges of informal caregivers: common themes and differences among whites, blacks, and hispanics. Gerontologist 2001 Dec;41(6):733-741. [doi: 10.1093/geront/41.6.733] [Medline: 11723341]
17. Roth DL, Fredman L, Haley WE. Informal caregiving and its impact on health: a reappraisal from population-based studies. Gerontologist 2015 Apr;55(2):309-319. [doi: 10.1093/geront/gnu177] [Medline: 26035608]
18. Premanandan S, Ahmad A, Cajander Å, Ågerfalk P, van Gemert-Pijnen L. Designing and evaluating IT applications for informal caregivers: protocol for a scoping review. JMIR Res Protoc 2023 Jul 06;12:e47650 [FREE Full text] [doi: 10.2196/47650] [Medline: 37410533]
19. Sala-González M, Pérez-Jover V, Guilabert M, Mira JJ. Mobile apps for helping informal caregivers: a systematic review. Int J Environ Res Public Health 2021 Feb 10;18(4):1702 [FREE Full text] [doi: 10.3390/ijerph18041702] [Medline: 33578819]
20. Carretero S, Stewart J, Centeno C. Information and communication technologies for informal carers and paid assistants: benefits from micro-, meso-, and macro-levels. Eur J Ageing 2015 Jan 24;12(2):163-173 [FREE Full text] [doi: 10.1007/s10433-015-0333-4] [Medline: 26346568]
21. Torp S, Bing-Jonsson PC, Hanson E. Experiences with using information and communication technology to build a multi-municipal support network for informal carers. Inform Health Soc Care 2013 Sep;38(3):265-279. [doi: 10.3109/17538157.2012.735733] [Medline: 23245624]
22. Guessi Margarido M, Shah A, Seto E. Smartphone applications for informal caregivers of chronically ill patients: a scoping review. NPJ Digit Med 2022 Mar 21;5(1):33 [FREE Full text] [doi: 10.1038/s41746-022-00567-z] [Medline: 35314766]
23. Hassan AY. Challenges and recommendations for the deployment of information and communication technology solutions for informal caregivers: scoping review. JMIR Aging 2020 Jul 29;3(2):e20310 [FREE Full text] [doi: 10.2196/20310] [Medline: 32723720]
24. Woo K, Tark A, Baik D, Dowding D. Informal caregiver decision-making factors associated with technology adoption and use in home healthcare: a systematic scoping review. Home Healthc Now 2019;37(6):328-336. [doi: 10.1097/NHH.0000000000000811] [Medline: 31688467]
25. Lucero RJ, Fehlberg EA, Patel AG, Bjarnardottir RI, Williams R, Lee K, et al. The effects of information and communication technologies on informal caregivers of persons living with dementia: a systematic review. Alzheimers Dement (N Y) 2019;5:1-12 [FREE Full text] [doi: 10.1016/j.trci.2018.11.003] [Medline: 30623020]
26. Romero-Mas M, Gómez-Zúñiga B, Cox AM, Ramon-Aribau A. Designing virtual communities of practice for informal caregivers of Alzheimer's patients: an integrative review. Health Informatics J 2020 Dec;26(4):2976-2991 [FREE Full text] [doi: 10.1177/1460458220950883] [Medline: 32951497]
27. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019 Jun 20;19(1):400 [FREE Full text] [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
28. Martínez-Alcalá CI, Pliego-Pastrana P, Rosales-Lagarde A, Lopez-Noguerola JS, Molina-Trinidad EM. Information and communication technologies in the care of the elderly: systematic review of applications aimed at patients with dementia and caregivers. JMIR Rehabil Assist Technol 2016 May 02;3(1):e6 [FREE Full text] [doi: 10.2196/rehab.5226] [Medline: 28582258]
29. Powell J, Chiu T, Eysenbach G. A systematic review of networked technologies supporting carers of people with dementia. J Telemed Telecare 2008 Mar;14(3):154-156. [doi: 10.1258/jtt.2008.003018] [Medline: 18430288]
30. Meyer TD, Casarez R, Mohite SS, La Rosa N, Iyengar MS. Novel technology as platform for interventions for caregivers and individuals with severe mental health illnesses:a systematic review. J Affect Disord 2018 Jan 15;226:169-177. [doi: 10.1016/j.jad.2017.09.012] [Medline: 28987999]
31. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
32. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
33. Costantino G, Montano N, Casazza G. When should we change our clinical practice based on the results of a clinical study? Searching for evidence: PICOS and PubMed. Intern Emerg Med 2015 Jun 2;10(4):525-527. [doi: 10.1007/s11739-015-1225-5] [Medline: 25833085]
34. Publish or perish. Harzing.com. URL: https://harzing.com/resources/publish-or-perish [accessed 2024-02-19]
35. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc 2016 Jul;104(3):240-243 [FREE Full text] [doi: 10.3163/1536-5050.104.3.014] [Medline: 27366130]
36. The best citation and reference management tool. EndNote. URL: https://endnote.com/ [accessed 2023-11-08]

XSL•FO
RenderX

37. AI powered tool for systematic literature reviews. Rayyan. URL: https://www.rayyan.ai/ [accessed 2023-05-11]
38. Rathnayake S, Moyle W, Jones C, Calleja P. Co-design of an mHealth application for family caregivers of people with dementia to address functional disability care needs. Inform Health Soc Care 2021 Mar 02;46(1):1-17. [doi: 10.1080/17538157.2020.1793347] [Medline: 32706282]
39. Macaden L, Muirhead K, Melchiorre G, Mantle R, Ditta G, Giangreco A. Relationship-centred CogniCare: an academic–digital–dementia care experts interface. Work Older People 2020 Dec 24;25(1):73-83. [doi: 10.1108/wwop-05-2020-0016]
40. Lobão MJ, Guan Y, Curado J, Goncalves M, Melo R, Silva C, et al. Solution to support informal caregivers of patients with dementia. Procedia Comput Sci 2021;181:294-301. [doi: 10.1016/j.procs.2021.01.149]
41. Leslie M, Gray RP, Khayatzadeh-Mahani A. What is 'care quality' and can it be improved by information and communication technology? A typology of family caregivers' perspectives. Scand J Caring Sci 2021 Mar 13;35(1):220-232. [doi: 10.1111/scs.12837] [Medline: 32168399]
42. Dickman Portz J, Ford K, Bekelman DB, Boxer RS, Kutner JS, Czaja S, et al. "We're taking something so human and trying to digitize": provider recommendations for mHealth in palliative care. J Palliat Med 2020 Feb;23(2):240-247 [FREE Full text] [doi: 10.1089/jpm.2019.0216] [Medline: 31526220]
43. Nurgalieva L, Baez M, Adamo G, Casati F, Marchese M. Designing interactive systems to mediate communication between formal and informal caregivers in aged care. IEEE Access 2019;7:171173-171194. [doi: 10.1109/access.2019.2954327]
44. Lederman R, Gleeson J, Wadley G, D'alfonso S, Rice S, Santesteban-Echarri O, et al. Support for carers of young people with mental illness: design and trial of a technology-mediated therapy. ACM Trans Comput Hum Interact 2019 Feb 21;26(1):1-33. [doi: 10.1145/3301421]
45. Heynsbergh N, Heckel L, Botti M, Livingston PM. Development of a smartphone app for informal carers of people with cancer: processes and learnings. JMIR Form Res 2019 Apr 11;3(2):e10990 [FREE Full text] [doi: 10.2196/10990] [Medline: 30973346]
46. Chaar D, Shin J, Mazzoli A, Vue R, Kedroske J, Chappell G, et al. A mobile health app (Roadmap 2.0) for patients undergoing hematopoietic stem cell transplant: qualitative study on family caregivers' perspectives and design considerations. JMIR Mhealth Uhealth 2019 Oct 24;7(10):e15775 [FREE Full text] [doi: 10.2196/15775] [Medline: 31651402]
47. Allemann H, Thylén I, Ågren S, Liljeroos M, Strömberg A. Perceptions of information and communication technology as support for family members of persons with heart failure: qualitative study. J Med Internet Res 2019 Jul 16;21(7):e13521 [FREE Full text] [doi: 10.2196/13521] [Medline: 31313662]
48. McNaney R, Vines J, Mercer J, Mexter L, Welsh D, Young T. DemYouth: co-designing and enacting tools to support young people's engagement with people with dementia. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO p. 1313-1325 URL: https://dl.acm.org/doi/10.1145/3025453.3025558 [doi: 10.1145/3025453.3025558]
49. Tixier M, Lewkowicz M. "Counting on the group": reconciling online and offline social support among older informal caregivers. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 545-558 URL: https://dl.acm.org/doi/10.1145/2858036.2858477 [doi: 10.1145/2858036.2858477]
50. Schorch M, Wan L, Randall D, Wulf V. Designing for those who are overlooked: insider perspectives on care practices and cooperative work of elderly informal caregivers. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. 2016 Presented at: CSCW '16; February 27-March 2, 2016; San Francisco, CA p. 787-799 URL: https://dl.acm.org/doi/10.1145/2818048.2819999 [doi: 10.1145/2818048.2819999]
51. Bosch LB, Kanis M. Design opportunities for supporting informal caregivers. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2016 Presented at: CHI EA '16; May 7-12, 2016; San Jose, CA p. 2790-2797 URL: https://dl.acm.org/doi/10.1145/2851581.2892354 [doi: 10.1145/2851581.2892354]
52. Meiland FJ, Hattink BJ, Overmars-Marx T, de Boer ME, Jedlitschka A, Ebben PW, et al. Participation of end users in the design of assistive technology for people with mild to severe cognitive problems; the European Rosetta project. Int Psychogeriatr 2014 Feb 10;26(5):769-779. [doi: 10.1017/s1041610214000088]
53. McHugh JE, Wherton JP, Prendergast DK, Lawlor BK. Identifying opportunities for supporting caregivers of persons with dementia through information and communication technology. Gerontechnology 2012;10(4):220-230. [doi: 10.4017/gt.2012.10.4.003.00]
54. Hwang AS, Truong KN, Mihailidis A. Using participatory design to determine the needs of informal caregivers for smart home user interfaces. In: Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare. 2012 Presented at: PervasiveHealth '12; May 21-24, 2012; San Diego, CA p. 41-48 URL: https://ieeexplore.ieee.org/document/6240361 [doi: 10.4108/icst.pervasivehealth.2012.248671]
55. Renyi M, Teuteberg F, Kunze C. ICT-Based support for the collaboration of formal and informal caregivers: a user-centered design study. In: Abramowicz W, Paschke A, editors. Business Information Systems. Cham, Switzerland: Springer; 2018:400-411.
56. Moberg C, Grundberg Å, Konradsen H, Kallström AP, Leung AY, Kabir ZN. Meeting own needs and supporting ability to care: family caregivers' and health care professionals' perspectives on professional support provided through a potential

XSL•FO
RenderX

mobile application. Dementia (London) 2022 May 18;21(4):1120-1134. [doi: 10.1177/14713012211065313] [Medline: 35300535]
57. Ahmad A, Premanandan S, Langegård U, Cajander Å, Johansson B, Carlsson M, et al. Positive design framework for carer eSupport: qualitative study to support informal caregivers of patients with head and neck cancer in Sweden. JMIR Cancer 2023 May 30;9:e45748 [FREE Full text] [doi: 10.2196/45748] [Medline: 37252765]
58. Premanandan S, Ahmad A, Cajander Å, Ågerfalk PJ, van Gemert-Pijnen L. Design suggestions for a persuasive e-coaching application: a study on informal caregivers' needs. Digit Health 2023 May 30;9:20552076231177129 [FREE Full text] [doi: 10.1177/20552076231177129] [Medline: 37284014]
59. Molinari-Ulate M, Guirado-Sánchez Y, Platón L, van der Roest HG, Bahillo A, Franco-Martín M. Cultural adaptation of the iSupport online training and support programme for caregivers of people with dementia in Castilla y León, Spain. Dementia (London) 2023 Jul 21;22(5):1010-1026 [FREE Full text] [doi: 10.1177/14713012231165578] [Medline: 36942726]
60. Ciuffreda I, Amabili G, Casaccia S, Benadduci M, Margaritini A, Maranesi E, et al. Design and development of a technological platform based on a sensorized social robot for supporting older adults and caregivers: GUARDIAN ecosystem. Int J of Soc Robotics 2023 Sep 06. [doi: 10.1007/S12369-023-01038-5]
61. Gris F, D'Amen B, Lamura G, Paciaroni L, Socci M, Melchiorre MG. Personalized technological support for informal caregivers of older people with dementia: a co-design approach involving potential end users and healthcare professionals in three focus groups in Italy. Healthcare (Basel) 2023 Sep 28;11(19):2640 [FREE Full text] [doi: 10.3390/healthcare11192640] [Medline: 37830677]
62. Haji Mukhti M, Ibrahim M, Tengku Ismail TA, Nadal IP, Kamalakannan S, Kinra S, et al. Exploring the need for mobile application in stroke management by informal caregivers: a qualitative study. Int J Environ Res Public Health 2022 Oct 10;19(19):12959 [FREE Full text] [doi: 10.3390/ijerph191912959] [Medline: 36232257]
63. Fan Q, DuBose L, Ory MG, Lee S, Hoang MN, Vennatt J, et al. Financial, legal, and functional challenges of providing care for people living with dementia and needs for a digital platform: interview study among family caregivers. JMIR Aging 2023 Sep 05;6:e47577 [FREE Full text] [doi: 10.2196/47577] [Medline: 37526513]
64. Egan KJ, Hodgson W, Dunlop MD, Imperatore G, Kirk A, Maguire R. A novel mobile app ("CareFit") to support informal caregivers to undertake regular physical activity from home during and beyond COVID-19 restrictions: co-design and prototype development study. JMIR Form Res 2021 Oct 01;5(10):e27358 [FREE Full text] [doi: 10.2196/27358] [Medline: 34406969]
65. Giroux D, Tremblay M, Latulippe K, Provencher V, Poulin V, Giguere A, et al. Promoting identification and use of aid resources by caregivers of seniors: co-design of an electronic health tool. JMIR Aging 2019 Aug 22;2(2):e12314 [FREE Full text] [doi: 10.2196/12314] [Medline: 31518284]
66. Gutierrez FJ, Ochoa SF. Making visible the invisible: understanding the nuances of computer-supported cooperative work on informal elderly caregiving in Southern Cone families. Pers Ubiquit Comput 2020 May 01;25(2):437-456. [doi: 10.1007/S00779-020-01404-4]
67. Hashemi M, Irajpour A, Taleghani F. Caregivers needing care: the unmet needs of the family caregivers of end-of-life cancer patients. Support Care Cancer 2018 Mar 27;26(3):759-766. [doi: 10.1007/s00520-017-3886-2] [Medline: 28952034]
68. Köhle N, Drossaert CH, Oosterik S, Schreurs KM, Hagedoorn M, van Uden-Kraan CF, et al. Needs and preferences of partners of cancer patients regarding a web-based psychological intervention: a qualitative study. JMIR Cancer 2015 Dec 29;1(2):e13 [FREE Full text] [doi: 10.2196/cancer.4631] [Medline: 28410157]
69. Martínez-Alcalá CI, Ramirez-Salvador JA, Rosales-Lagarde A, Jiménez-Rodríguez B. An eService platform for the assistance and support of primary caregivers. In: Proceedings of the 12th Iberian Conference on Information Systems and Technologies. 2017 Presented at: CISTI '17; June 21-24, 2017; Lisbon, Portugal p. 1-5 URL: https://ieeexplore.ieee.org/document/7975767 [doi: 10.23919/cisti.2017.7975767]
70. Masterson-Algar P, Egan K, Flynn G, Hughes G, Spector A, Stott J, et al. iSupport for young carers: an adaptation of an e-health intervention for young dementia carers. Int J Environ Res Public Health 2022 Dec 22;20(1):127 [FREE Full text] [doi: 10.3390/ijerph20010127] [Medline: 36612450]
71. Moreno-Cámara S, Palomino-Moral P, Moral-Fernández L, Frías-Osuna A, Parra-Anguita L, Del-Pino-Casado R. Perceived needs of the family caregivers of people with dementia in a mediterranean setting: a qualitative study. Int J Environ Res Public Health 2019 Mar 19;16(6):993 [FREE Full text] [doi: 10.3390/ijerph16060993] [Medline: 30893902]
72. Sepehri K, Holsti L, Niasati S, Chan V, Maclean KE. Beyond the bulging binder: family-centered design of a digital health information management system for caregivers of children living with health complexity. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany p. 1-19 URL: https://dl.acm.org/doi/10.1145/3544548.3581459 [doi: 10.1145/3544548.3581459]
73. Shreve J, Baier R, Epstein-Lubow G, Gardner R. Dementia caregivers' technology preferences: Design insights from qualitative interviews. Gerontechnology 2016;14(2):89-96 [FREE Full text] [doi: 10.4017/gt.2016.14.2.004.00]
74. Siddiqui F, Varghese D, Singh P, Bayyavarapu SB, Lindsay S, Chandrasekara D, et al. Exploring the digital support needs of caregivers of people with serious mental illness. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany p. 1-16 URL: https://dl.acm.org/doi/10.1145/3544548.3580674 [doi: 10.1145/3544548.3580674]

XSL•FO
RenderX

75. Vaughan C, Trail TE, Mahmud A, Dellva S, Tanielian T, Friedman E. Informal caregivers' experiences and perceptions of a web-based peer support network: mixed-methods study. J Med Internet Res 2018 Aug 28;20(8):e257 [FREE Full text] [doi: 10.2196/jmir.9895] [Medline: 30154074]
76. Williamson SS, Gorman PN, Jimison HB. A mobile/web app for long distance caregivers of older adults: functional requirements and design implications from a user centered design process. AMIA Annu Symp Proc 2014;2014:1960-1969 [FREE Full text] [Medline: 25954469]
77. Liverpool S, Edbrooke-Childs J. A caregiver digital intervention to support shared decision making in child and adolescent mental health services: development process and stakeholder involvement analysis. JMIR Form Res 2021 Jun 15;5(6):e24896 [FREE Full text] [doi: 10.2196/24896] [Medline: 34128821]
78. Xu X, Griva K, Koh M, Lum E, Tan WS, Thng S, et al. Creating a smartphone app for caregivers of children with atopic dermatitis with caregivers, health care professionals, and digital health experts: participatory co-design. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e16898 [FREE Full text] [doi: 10.2196/16898] [Medline: 33118949]
79. Ducharme F, Kergoat MJ, Coulombe R, Lévesque L, Antoine P, Pasquier F. Unmet support needs of early-onset dementia family caregivers: a mixed-design study. BMC Nurs 2014 Dec 19;13(1):49 [FREE Full text] [doi: 10.1186/s12912-014-0049-3] [Medline: 25550685]
80. Wan SW, Chong CS, Toh EL, Lim SH, Loi CT, Lew YF, et al. A theory-based, multidisciplinary approach to cocreate a patient-centric digital solution to enhance perioperative health outcomes among colorectal cancer patients and their family caregivers: development and evaluation study. J Med Internet Res 2021 Dec 07;23(12):e31917 [FREE Full text] [doi: 10.2196/31917] [Medline: 34878991]
81. Cheng CF, Werner NE, Doutcheva N, Warner G, Barton HJ, Kelly MM, et al. Codesign and usability testing of a mobile application to support family-delivered enteral tube care. Hosp Pediatr 2020 Aug;10(8):641-650 [FREE Full text] [doi: 10.1542/hpeds.2020-0076] [Medline: 32616602]
82. Costa Stutzel M, Filippo MP, Sztajnberg A, da Costa RM, Brites AD, da Motta LB, et al. Multi-part quality evaluation of a customized mobile application for monitoring elderly patients with functional loss and helping caregivers. BMC Med Inform Decis Mak 2019 Jul 22;19(1):140 [FREE Full text] [doi: 10.1186/s12911-019-0839-3] [Medline: 31331309]
83. Xiao L, Yan X, Emery A. Design and evaluation of web interfaces for informal care providers in senior monitoring. Proc of Assoc for Info 2014 May 08;50(1):1-10. [doi: 10.1002/meet.14505001034]
84. Renati R, Bonfiglio NS, Rollo D. Dealing with loved ones' addiction: development of an app to cope with caregivers' stress. Int J Environ Res Public Health 2022 Nov 30;19(23):15950 [FREE Full text] [doi: 10.3390/ijerph192315950] [Medline: 36498025]
85. Thompson Z, Tamplin J, Vieira Sousa T, Carrasco R, Flynn L, Lamb KE, et al. Content development and validation for a mobile application designed to train family caregivers in the use of music to support care of people living with dementia. Front Med (Lausanne) 2023;10:1185818 [FREE Full text] [doi: 10.3389/fmed.2023.1185818] [Medline: 37250645]
86. Păsărelu CR, Kertesz R, Dobrean A. The development and usability of a mobile app for parents of children with ADHD. Children (Basel) 2023 Jan 14;10(1):164 [FREE Full text] [doi: 10.3390/children10010164] [Medline: 36670714]
87. Dam AE, van Boxtel MP, Rozendaal N, Verhey FR, de Vugt ME. Development and feasibility of Inlife: a pilot study of an online social support intervention for informal caregivers of people with dementia. PLoS One 2017 Sep 8;12(9):e0183386 [FREE Full text] [doi: 10.1371/journal.pone.0183386] [Medline: 28886056]
88. Wilding C, Davis H, Rasekaba T, Hamiduzzaman M, Royals K, Greenhill J, et al. Volunteers' support of carers of rural people living with dementia to use a custom-built application. Int J Environ Res Public Health 2021 Sep 20;18(18):9909 [FREE Full text] [doi: 10.3390/ijerph18189909] [Medline: 34574832]
89. Teles S, Paúl C, Lima P, Chilro R, Ferreira A. User feedback and usability testing of an online training and support program for dementia carers. Internet Interv 2021 Sep;25:100412 [FREE Full text] [doi: 10.1016/j.invent.2021.100412] [Medline: 34401371]
90. Gomes N, Caroço J, Rijo R, Martinho R, Querido A, Peralta T, et al. Evaluation of an e-health platform for informal caregivers and health professionals: the case study of Help2Care. Inform Health Soc Care 2022 Apr 03;47(2):144-158. [doi: 10.1080/17538157.2021.1964509] [Medline: 34404326]
91. Quinn CC, Staub S, Barr E, Gruber-Baldini A. Mobile support for older adults and their caregivers: dyad usability study. JMIR Aging 2019 May 23;2(1):e12276 [FREE Full text] [doi: 10.2196/12276] [Medline: 31518271]
92. Brouns B, Meesters JJ, Wentink MM, de Kloet AJ, Arwert HJ, Vliet Vlieland TP, et al. Why the uptake of eRehabilitation programs in stroke care is so difficult-a focus group study in the Netherlands. Implement Sci 2018 Oct 29;13(1):133 [FREE Full text] [doi: 10.1186/s13012-018-0827-5] [Medline: 30373611]
93. Andersson S, Erlingsson C, Magnusson L, Hanson E. The experiences of working carers of older people regarding access to a web-based family care support network offered by a municipality. Scand J Caring Sci 2017 Sep 01;31(3):487-496. [doi: 10.1111/scs.12361] [Medline: 27481631]
94. Kajaks T, Longfield A, Orozco F, Holyoke P, Dutta T. Pilot quality improvement study of Safeback: a mobile application for estimating low back compression forces for caregivers in the field. Proc Hum Factors Ergon Soc Annu Meet 2016 Dec 20;59(1):610-614. [doi: 10.1177/1541931215591134]

XSL•FO
RenderX

95. Lundberg S. The results from a two-year case study of an information and communication technology support system for family caregivers. Disabil Rehabil Assist Technol 2014 Jul 15;9(4):353-358. [doi: 10.3109/17483107.2013.814170] [Medline: 23855387]
96. Blusi M, Dalin R, Jong M. The benefits of e-health support for older family caregivers in rural areas. J Telemed Telecare 2014 Mar;20(2):63-69. [doi: 10.1177/1357633X13519901] [Medline: 24446251]
97. Blusi M, Asplund K, Jong M. Older family carers in rural areas: experiences from using caregiver support services based on Information and Communication Technology (ICT). Eur J Ageing 2013 Sep 29;10(3):191-199 [FREE Full text] [doi: 10.1007/s10433-013-0260-1] [Medline: 28804294]
98. Chiu TM, Eysenbach G. Theorizing the health service usage behavior of family caregivers: a qualitative study of an internet-based intervention. Int J Med Inform 2011 Nov;80(11):754-764. [doi: 10.1016/j.ijmedinf.2011.08.010] [Medline: 21958550]
99. Torp S, Hanson E, Hauge S, Ulstein I, Magnusson L. A pilot study of how information and communication technology may contribute to health promotion among elderly spousal carers in Norway. Health Soc Care Community 2008 Jan;16(1):75-85. [doi: 10.1111/j.1365-2524.2007.00725.x] [Medline: 18181817]
100. Andersson NB, Hanson E, Magnusson L, Nolan M. Views of family carers and older people of information technology. Br J Nurs 2002 Jun 27;11(12):827-831. [doi: 10.12968/bjon.2002.11.12.10306] [Medline: 12131833]
101. Husebø AM. Stakeholders' perspectives on eHealth support in colorectal cancer survivorship: qualitative interview study. JMIR Cancer 2021 Sep 07;7(3):e28279 [FREE Full text] [doi: 10.2196/28279] [Medline: 34491210]
102. Hermaszewska S, Sin J. End-user perspectives on the development of an online intervention for parents of children on the autism spectrum. Autism 2021 Jul 11;25(5):1234-1245 [FREE Full text] [doi: 10.1177/1362361320984895] [Medline: 33423522]
103. Rottenberg S, Williams A. Web-based delivery of the caregiving essentials course for informal caregivers of older adults in Ontario: mixed methods evaluation study. JMIR Aging 2021 Jun 15;4(2):e25671 [FREE Full text] [doi: 10.2196/25671] [Medline: 34128815]
104. Egan KJ, Hodgson W, Imperatore G, Dunlop MD, Maguire R, Kirk A. Supporting physical activity for informal caregivers during and beyond COVID-19: exploring the feasibility, usability and acceptability of a digital health smartphone application, 'CareFit'. Int J Environ Res Public Health 2022 Sep 30;19(19):12506 [FREE Full text] [doi: 10.3390/ijerph191912506] [Medline: 36231803]
105. Lam SK, Chau JP, Lo SH, Siow EK, Lee VW, Shum EW, et al. User engagement in the development of a home-based virtual multidisciplinary stroke care clinic for stroke survivors and caregivers: a qualitative descriptive study. Disabil Rehabil 2022 Oct;44(20):5983-5989. [doi: 10.1080/09638288.2021.1955305] [Medline: 34297647]
106. Boutilier JJ, Loganathar P, Linden A, Scheer E, Noejovich S, Elliott C, et al. A web-based platform (CareVirtue) to support caregivers of people living with Alzheimer disease and related dementias: mixed methods feasibility study. JMIR Aging 2022 Aug 04;5(3):e36975 [FREE Full text] [doi: 10.2196/36975] [Medline: 35925654]
107. Kwok I, Lattie EG, Yang D, Summers A, Grote V, Cotten P, et al. Acceptability and feasibility of a socially enhanced, self-guided, positive emotion regulation intervention for caregivers of individuals with dementia: pilot intervention study. JMIR Aging 2023 Sep 06;6:e46269 [FREE Full text] [doi: 10.2196/46269] [Medline: 37672311]
108. Tibell LH, Alvariza A, Kreicbergs U, Wallin V, Steineck G, Holm M. Web-based support for spouses of patients with life-threatening illness cared for in specialized home care - A feasibility study. Palliat Support Care 2022 Dec 20:1-9. [doi: 10.1017/S1478951522001602] [Medline: 36537025]
109. Austrom MG, Geros KN, Hemmerlein K, McGuire SM, Gao S, Brown SA, et al. Use of a multiparty web based videoconference support group for family caregivers: innovative practice. Dementia (London) 2015 Sep 25;14(5):682-690 [FREE Full text] [doi: 10.1177/1471301214544338] [Medline: 25062788]
110. Barbabella F, Poli A, Hanson E, Andréasson F, Salzmann B, Döhner H, et al. Usage and usability of a web-based program for family caregivers of older people in three European countries: a mixed-methods evaluation. Comput Inform Nurs 2018 May;36(5):232-241. [doi: 10.1097/CIN.0000000000000422] [Medline: 29505433]
111. Brown EL, Ruggiano N, Page TF, Roberts L, Hristidis V, Whiteman KL, et al. CareHeroes web and Android™ apps for dementia caregivers: a feasibility study. Res Gerontol Nurs 2016;9(4):193-203 [FREE Full text] [doi: 10.3928/19404921-20160229-02] [Medline: 29977440]
112. Crotty BH, Asan O, Holt J, Tyszka J, Erickson J, Stolley M, et al. Qualitative assessment of unmet information management needs of informal cancer caregivers: four themes to inform oncology practice. JCO Clin Cancer Inform 2020 Nov(4):521-528. [doi: 10.1200/cci.19.00159]
113. Iribarren S, Stonbraker S, Suero-Tejeda N, Granja M, Luchsinger JA, Mittelman M, et al. Information, communication, and online tool needs of Hispanic family caregivers of individuals with Alzheimer's disease and related dementias. Inform Health Soc Care 2019;44(2):115-134. [doi: 10.1080/17538157.2018.1433674] [Medline: 29504837]
114. Ploeg J, McAiney C, Duggleby W, Chambers T, Lam A, Peacock S, et al. A web-based intervention to help caregivers of older adults with dementia and multiple chronic conditions: qualitative study. JMIR Aging 2018 Apr 23;1(1):e2 [FREE Full text] [doi: 10.2196/aging.8475] [Medline: 31518230]

XSL•FO
RenderX

115. Rettinger L, Zeuner L, Werner K, Ritschl V, Mosor E, Stamm T, et al. A mixed-methods evaluation of a supporting app for informal caregivers of people with dementia. In: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments. 2020 Presented at: PETRA '20; June 30-July 3, 2020; Corfu, Greece p. 1-8 URL: https://dl.acm.org/doi/10.1145/3389189.3397981 [doi: 10.1145/3389189.3397981]
116. Batchelor R, Gulshan S, Shritharan H, Williams E, Henderson C, Gillard S, et al. Perceived acceptability and experiences of a digital psychoeducation and peer support intervention (COPe-support): interview study with carers supporting individuals with psychosis. J Med Internet Res 2022 Feb 02;24(2):e27781 [FREE Full text] [doi: 10.2196/27781] [Medline: 35107439]
117. Honary M, Fisher NR, McNaney R, Lobban F. A web-based intervention for relatives of people experiencing psychosis or bipolar disorder: design study using a user-centered approach. JMIR Ment Health 2018 Dec 07;5(4):e11473 [FREE Full text] [doi: 10.2196/11473] [Medline: 30530457]
118. Sin J, Woodham LA, Henderson C, Williams E, Sesé Hernández A, Gillard S. Usability evaluation of an eHealth intervention for family carers of individuals affected by psychosis: a mixed-method study. Digit Health 2019 Aug 28;5:2055207619871148 [FREE Full text] [doi: 10.1177/2055207619871148] [Medline: 31489205]
119. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
120. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
121. Vaismoradi M, Jones J, Turunen H, Snelgrove S. Theme development in qualitative content analysis and thematic analysis. J Nurs Educ Pract 2016 Jan 15;6(5):100. [doi: 10.5430/jnep.v6n5p100]
122. Adams E, Boulton M, Watson E. The information needs of partners and family members of cancer patients: a systematic literature review. Patient Educ Couns 2009 Nov;77(2):179-186. [doi: 10.1016/J.PEC.2009.03.027]
123. Bee PE, Barnes P, Luker KA. A systematic review of informal caregivers' needs in providing home-based end-of-life care to people with cancer. J Clin Nurs 2009 May 08;18(10):1379-1393. [doi: 10.1111/j.1365-2702.2008.02405.x] [Medline: 18624779]
124. Ramadurai D, Corder S, Churney T, Graney B, Harshman A, Meadows S, et al. Understanding the informational needs of patients with IPF and their caregivers: 'you get diagnosed, and you ask this question right away, what does this mean?'. BMJ Open Qual 2018 Jan 30;7(1):e000207 [FREE Full text] [doi: 10.1136/bmjoq-2017-000207] [Medline: 29464212]
125. Washington KT, Meadows SE, Elliott SG, Koopman RJ. Information needs of informal caregivers of older adults with chronic health conditions. Patient Educ Couns 2011 Apr;83(1):37-44. [doi: 10.1016/j.pec.2010.04.017] [Medline: 20452165]
126. Lamm K, Kristensson Hallström I, Landgren K. Parents' experiences of living with a child with paediatric feeding disorder: an interview study in Sweden. Scand J Caring Sci 2023 Dec 10;37(4):949-958. [doi: 10.1111/scs.13070] [Medline: 35142385]
127. Darcy J, Brunsden V, Hill R. Exploring online support: informal caregivers' usage of a mental health discussion board. J Cyber Ther Rehabil 2011;4(4):477 [FREE Full text]
128. Khemai C, Meijers JM, Mujezinovic I, Bolt SR, Pieters S, Moser A, et al. Interprofessional collaboration in palliative dementia care through the eyes of informal caregivers. Dementia (London) 2022 Aug 10;21(6):1890-1913 [FREE Full text] [doi: 10.1177/14713012221098259] [Medline: 35535552]
129. Kontrimiene A, Sauseriene J, Blazeviciene A, Raila G, Jaruseviciene L. Qualitative research of informal caregivers' personal experiences caring for older adults with dementia in Lithuania. Int J Ment Health Syst 2021 Jan 20;15(1):12 [FREE Full text] [doi: 10.1186/s13033-020-00428-w] [Medline: 33472676]
130. Newman K, Wang AH, Wang AZ, Hanna D. The role of internet-based digital tools in reducing social isolation and addressing support needs among informal caregivers: a scoping review. BMC Public Health 2019 Nov 09;19(1):1495 [FREE Full text] [doi: 10.1186/s12889-019-7837-3] [Medline: 31706294]
131. Bjerregaard K, Haslam SA, Mewse A, Morton T. The shared experience of caring: a study of care-workers' motivations and identifications at work. Age Soc 2015 Oct 12;37(1):113-138. [doi: 10.1017/s0144686x15000860]
132. Schulz R, Czaja SJ. Family caregiving: a vision for the future. Am J Geriatr Psychiatry 2018 Mar;26(3):358-363. [doi: 10.1016/j.jagp.2017.06.023] [Medline: 28774786]
133. Council on Children with Disabilities and Medical Home Implementation Project Advisory Committee. Patient- and family-centered care coordination: a framework for integrating care for children and youth across multiple systems. Pediatrics 2014 May 28;133(5):e1451-e1460. [doi: 10.1542/peds.2014-0318] [Medline: 24777209]
134. Kianfar S, Carayon P, Hundt AS, Hoonakker P. Care coordination for chronically ill patients: identifying coordination activities and interdependencies. Appl Ergon 2019 Oct;80:9-16 [FREE Full text] [doi: 10.1016/j.apergo.2019.05.002] [Medline: 31280815]
135. Espinel-Flores V, Vargas I, Eguiguren P, Mogollón-Pérez AS, Ferreira de Medeiros Mendes M, López-Vázquez J, et al. Assessing the impact of clinical coordination interventions on the continuity of care for patients with chronic conditions: participatory action research in five Latin American countries. Health Policy Plan 2022 Jan 13;37(1):1-11 [FREE Full text] [doi: 10.1093/heapol/czab130] [Medline: 34718564]
136. Look KA, Stone JA. Medication management activities performed by informal caregivers of older adults. Res Social Adm Pharm 2018 May;14(5):418-426 [FREE Full text] [doi: 10.1016/j.sapharm.2017.05.005] [Medline: 28528023]

XSL•FO
RenderX

137. Gillespie R, Mullan J, Harrison L. Managing medications: the role of informal caregivers of older adults and people living with dementia. A review of the literature. J Clin Nurs 2014 Dec;23(23-24):3296-3308. [doi: 10.1111/jocn.12519] [Medline: 24354583]
138. Alwan M, Kell S, Turner B, Dalal S, Mack D, Felder R. Psychosocial impact of passive health status monitoring on informal caregivers and older adults living in independent senior housing. In: Proceedings of the 2nd International Conference on Information & Communication Technologies. 2006 Presented at: ICICT '06; April 24-28, 2006; Damascus, Syria p. 808-813 URL: https://ieeexplore.ieee.org/document/1684477 [doi: 10.1109/ictta.2006.1684477]
139. Spann A, Vicente J, Abdi S, Hawley M, Spreeuwenberg M, de Witte L. Benefits and barriers of technologies supporting working carers- a scoping review. Health Soc Care Commun 2022 Jan 25;30(1):e1-15 [FREE Full text] [doi: 10.1111/hsc.13421] [Medline: 34036665]
140. Lorca-Cabrera J, Martí-Arques R, Albacar-Riobóo N, Raigal-Aran L, Roldan-Merino J, Ferré-Grau C. Mobile applications for caregivers of individuals with chronic conditions and/or diseases: quantitative content analysis. Int J Med Inform 2021 Jan;145:104310. [doi: 10.1016/j.ijmedinf.2020.104310] [Medline: 33161319]
141. Lam L, Lam M. The use of information technology and mental health among older care-givers in Australia. Aging Ment Health 2009 Jul;13(4):557-562. [doi: 10.1080/13607860802607306] [Medline: 19629780]
142. Demiris G, Afrin LB, Speedie S, Courtney KL, Sondhi M, Vimarlund V, et al. Patient-centered applications: use of information technology to promote disease management and wellness. A white paper by the AMIA knowledge in motion working group. J Am Med Inform Assoc 2008 Jan;15(1):8-13 [FREE Full text] [doi: 10.1197/jamia.M2492] [Medline: 17947617]
143. McPherson KM, Kayes NK, Moloczij N, Cummins C. Improving the interface between informal carers and formal health and social services: a qualitative study. Int J Nurs Stud 2014 Mar;51(3):418-429. [doi: 10.1016/j.ijnurstu.2013.07.006] [Medline: 23928324]
144. Morris SM, Thomas C. The need to know: informal carers and information. Eur J Cancer Care (Engl) 2002 Sep;11(3):183-187. [doi: 10.1046/j.1365-2354.2002.00337.x] [Medline: 12296835]
145. Wiles J. Informal caregivers' experiences of formal support in a changing context. Health Soc Care Community 2003 May;11(3):189-207. [doi: 10.1046/j.1365-2524.2003.00419.x] [Medline: 12823424]
146. Wittenberg Y, Kwekkeboom R, Staaks J, Verhoeff A, de Boer A. Informal caregivers' views on the division of responsibilities between themselves and professionals: a scoping review. Health Soc Care Community 2018 Jul 18;26(4):e460-e473. [doi: 10.1111/hsc.12529] [Medline: 29250848]
147. Kamalraj P, Savundranayagam MY, Orange JB, Kloseck M. Communication in home care: understanding the lived experiences of formal caregivers communicating with persons living with dementia. Int J Older People Nurs 2021 Nov;16(6):e12401. [doi: 10.1111/opn.12401] [Medline: 34337872]
148. Ahmad A, Mozelius P. Human-computer interaction for older adults - a literature review on technology acceptance of eHealth systems. J Engg Res Sci 2022 Apr;1(4):119-126. [doi: 10.55708/js0104014]
149. Christie HL, Bartels SL, Boots LM, Tange HJ, Verhey F, de Vugt ME. A systematic review on the implementation of eHealth interventions for informal caregivers of people with dementia. Internet Interv 2018 Sep;13:51-59 [FREE Full text] [doi: 10.1016/j.invent.2018.07.002] [Medline: 30206519]
150. Peek ST, Wouters EJ, van Hoof J, Luijkx KG, Boeije HR, Vrijhoef HJ. Factors influencing acceptance of technology for aging in place: a systematic review. Int J Med Inform 2014 Apr;83(4):235-248 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.01.004] [Medline: 24529817]
151. Norman DA. Cognitive engineering. In: Norman DA, Draper SW, editors. User Centered System Design; New Perspectives on Human-Computer Interaction. Boston, MA: L. Erlbaum Associates; 1986:61.

Abbreviations

PICOS: population, intervention, comparator, outcome, and study design
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Review
RQ: research question

XSL•FO
RenderX

Edited by N Cahill; submitted 19.02.24; peer-reviewed by G Myreteg, S Brdnik, T Heričko; comments to author 21.06.24; revised version received 17.07.24; accepted 10.09.24; published 23.10.24.

Please cite as:
Premanandan S, Ahmad A, Cajander Å, Ågerfalk P, van Gemert-Pijnen L
Designing and Evaluating IT Applications for Informal Caregivers: Scoping Review
J Med Internet Res 2024;26:e57393
URL: https://www.jmir.org/2024/1/e57393
doi:10.2196/57393
PMID:

©Shweta Premanandan, Awais Ahmad, Åsa Cajander, Pär Ågerfalk, Lisette van Gemert-Pijnen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Acceptability of Digital Mental Health Interventions for Depression and Anxiety: Systematic Review

Carrie K Y Lau[1,2], MSW; Anthony Saad[1], BSc, MPH, MD; Bettina Camara[3], MSc; Dia Rahman, BES; Blanca Bolea-Alamanac[1,4], MSc, MD, PhD

[1]Department of Psychiatry, Women's College Hospital, Toronto, ON, Canada
[2]Factor-Inwentash Faculty of Social Work, University of Toronto, Toronto, ON, Canada
[3]Department of Immunology, Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[4]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Blanca Bolea-Alamanac, MSc, MD, PhD
Department of Psychiatry
Women's College Hospital
76 Grenville St
Toronto, ON, M5S 1B2
Canada
Phone: 1 647 331 2707
Email: Blanca.BoleaAlamanac@wchospital.ca

Abstract

Background: Depression and anxiety disorders are common, and treatment often includes psychological interventions. Digital health interventions, delivered through technologies such as web-based programs and mobile apps, are increasingly used in mental health treatment. Acceptability, the extent to which an intervention is viewed positively, has been identified as contributing to patient adherence and engagement with digital health interventions. Acceptability, therefore, impacts the benefit derived from using digital health interventions in treatment. Understanding the acceptability of digital mental health interventions among patients with depression or anxiety disorders is essential to maximize the effectiveness of their treatment.

Objective: This review investigated the acceptability of technology-based interventions among patients with depression or anxiety disorders.

Methods: A systematic review was performed based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PROSPERO (International Prospective Register of Systematic Reviews) guidelines. We searched PubMed, Web of Science, and Ovid in May 2022. Studies were included if they evaluated digital interventions for the treatment of depression or anxiety disorders and investigated their acceptability among adult patients. Studies were excluded if they targeted only specific populations (eg, those with specific physical health conditions), investigated acceptability in healthy individuals or patients under the age of 18 years, involved no direct interaction between patients and technologies, used technology only as a platform for traditional care (eg, videoconferencing), had patients using technologies only in clinical or laboratory settings, or involved virtual reality technologies. Acceptability outcome data were narratively synthesized by the direction of acceptability using vote counting. Included studies were evaluated using levels of evidence from the Oxford Centre for Evidence-Based Medicine. The risk of bias was assessed using a tool designed for this review and GRADE (Grading of Recommendations, Assessment, Development, and Evaluation).

Results: A total of 143 articles met the inclusion criteria, comprising 67 (47%) articles on interventions for depression, 65 (45%) articles on interventions for anxiety disorders, and 11 (8%) articles on interventions for both. Overall, 90 (63%) were randomized controlled trials, 50 (35%) were other quantitative studies, and 3 (2%) were qualitative studies. Interventions used web-based programs, mobile apps, and computer programs. Cognitive behavioral therapy was the basis of 71% (102/143) of the interventions. Digital mental health interventions were generally acceptable among patients with depression or anxiety disorders, with 88% (126/143) indicating positive acceptability, 8% (11/143) mixed results, and 4% (6/143) insufficient information to categorize the direction of acceptability. The available research evidence was of moderate quality.

XSL•FO
RenderX

Conclusions: Digital mental health interventions seem to be acceptable to patients with depression or anxiety disorders. Consistent use of validated measures for acceptability would enhance the quality of evidence. Careful design of acceptability as an evaluation outcome can further improve the quality of evidence and reduce the risk of bias.

Trial Registration: Open Science Framework Y7MJ4; https://doi.org/10.17605/OSF.IO/SPR8M

(J Med Internet Res 2024;26:e52609) doi:10.2196/52609

KEYWORDS

acceptability; mental health; depression; anxiety; mobile app; internet; mobile phone; mobile health; mHealth; digital health

Introduction

Background

Depression and anxiety disorders are common [1] and have high and increasing societal costs [2,3]. Patients with depression or anxiety disorders can benefit from psychological treatments but face a variety of barriers to accessing traditional mental health services.

Digital health interventions include those delivered with technologies such as web and mobile apps. They have garnered optimism for their potential to lower barriers to accessing mental health services [4,5]. Digital mental health interventions can facilitate access to services by increasing convenience for patients and reducing demands on their time, eliminating geographic limitations, reducing costs, and increasing the capacity of health systems [6].

The use of digital mental health interventions increased during the COVID-19 pandemic [7,8]. Their expanded reach occurred amid public health measures that limited in-person mental health treatment and predictions of a long tail of mental health repercussions [9]. Digital mental health interventions are now commonly available to patients alongside other mental health services, as an adjunct to in-person or synchronous video-based mental health treatment, or as a stopgap for long waitlists [10,11].

However, uncertainty exists around patient adherence to digital mental health interventions. Significant variations in adherence levels, ranging from low to high, have been noted among digital interventions [12]. Issues with nonadherence cast some doubt on whether the potential benefits of digital mental health interventions can be realized. Related issues have been found with user engagement (initiation and continuation of use of a digital health intervention) [13] among digital mental health interventions.

Acceptability, the degree to which individuals view a treatment or potential treatment positively, has been identified as predicting adherence and user engagement [14], and ultimately the intervention's effectiveness [15]. Acceptability among patients is therefore critical to realizing the potential of a digital mental health intervention.

Research into acceptability appears hampered by its varying conceptualizations. Sekhon and colleagues [16] propose that acceptability is the patient's perception of the treatment as appropriate, arising from their cognitive and affective reactions. They also provide a theoretical framework for acceptability with 7 components: affective attitude, burden, perceived effectiveness, ethicality, intervention coherence, opportunity costs, and self-efficacy [16]. Furthermore, Sekhon and colleagues [16] also recommend measuring acceptability at different time points during the intervention.

By contrast, Proctor and colleagues [17] define acceptability as the patient's perception of the intervention as satisfactory or tolerable. Zauszniewski [18] describes acceptability as patients' belief that the intervention is logical and suitable to them. They also describe 3 main components of acceptability, such as the acceptability of an intervention's delivery method, acceptability of the content of an intervention, and acceptability of the provider of the intervention, although they acknowledge that acceptability is often assessed only as a whole [18]. In contrast, the technology acceptance model proposed by Davis [19], regarding attitudes toward technology and its use more generally, presents acceptability as hinging upon users' perceptions of ease in using the technology and the technology's usefulness.

Conflict between the constructs of acceptability and satisfaction has also been identified, with recommendations that acceptability be understood as distinct from satisfaction [16,17]. However, studies continue to operationalize acceptability as a patient's satisfaction with a treatment, indicating continued resonance with the construction of satisfaction as a representation of acceptability, or at least a lack of consensus.

These examples demonstrate the variation with which acceptability has been understood and investigated in evaluations of digital mental health interventions thus far. Despite past recommendations for improvement (eg, Sekhon and colleagues [16]), evaluations have not yet converged on a particular approach to patient acceptability.

Acceptability has implications for user engagement and adherence [14]. Acceptability can impact the benefit that a patient can derive from an intervention and, in turn, compromise or enhance the individual's investment into digital mental health interventions. It is of particular importance to investigate the acceptability of digital interventions for patients with depression and anxiety disorders, given the prevalence of these diagnoses and the frequency with which clinicians are required to identify appropriate psychosocial interventions [1].

Aims and Objectives

This review summarizes the acceptability of digital mental health interventions for depression and anxiety disorders among adult patients. The research question we investigated was the following: "How acceptable to adult patients with depression or anxiety are digital mental health interventions (such as mobile

XSL•FO
RenderX

apps) used in psychiatric care, and what is the evidence for their acceptability?" We accomplished this by identifying evaluation studies of digital mental health interventions where acceptability is one of the outcome measures, reviewing the quality of the available evidence for acceptability, and summarizing the acceptability of digital mental health interventions to patients with depression and anxiety disorders.

Methods

Overview

We started with Sekhon and colleagues' [16] theoretical framework for acceptability, which identified in acceptability the component constructs of affective attitude, burden, perceived effectiveness, ethicality, intervention coherence, opportunity costs, and self-efficacy. We selected this theoretical framework because these component constructs can readily be identified in research studies having acceptability as an outcome measure. Where a study had an outcome measure that assessed one or more of these component constructs, whether or not the same terminology was used, the measure was considered relevant to our study.

In addition, the abovementioned theoretical framework for acceptability recognizes different temporal assessments of acceptability, such as prospective acceptability (before the intervention), concurrent acceptability (during the intervention), and retrospective acceptability (following the intervention) [16]. Each temporal assessment of acceptability is understood to achieve a different purpose and to provide different information [16]. For the purposes of this review, we included studies that assessed acceptability at any of these multiple time points but took note of the time points selected by the researchers for further analysis and discussion.

Sekhon and colleagues [16] suggest that satisfaction and acceptability are constructs that are often confounded but should be differentiated, referencing the fact that satisfaction can only be retrospectively assessed, while acceptability can be assessed at all time points. For this study, satisfaction is considered an aspect of retrospective acceptability. We therefore took into account measures of satisfaction as relevant to our review, despite satisfaction not being a part of Sekhon and colleagues' [16] theoretical framework.

Other constructs not named within Sekhon and colleagues' [16] theoretical framework of acceptability and included in this review are credibility, expectancy, and usability. For this review, studies evaluating credibility, expectancy, and usability were eligible for potential inclusion as long as the measures assess one or more component constructs of acceptability. Studies identified by their authors as collecting feedback, experiences, attitudes, subjective appraisal, and subjective benefit were similarly eligible for inclusion if they provided information on one or more component constructs of acceptability.

Inclusion and Exclusion Criteria

Studies were included if (1) they investigated digital mental health interventions that monitor or treat depression or anxiety disorders; (2) they evaluated internet-based, mobile, or computer-based mental health interventions, and could involve the use of such technologies as mobile apps or web-based treatment modules; (3) their participants were 18 years of age or older; and (4) they collected data on the acceptability of the intervention to their participants.

Excluded were studies that (1) focused on behavioral or lifestyle changes in healthy people; (2) focused on a physical health condition, a neurodevelopmental condition, or a mental health diagnosis other than depression or an anxiety disorder; (3) focused on a specific population such as military personnel, health care workers, students, or a target group representing only one component of a larger population; (4) focused on family members or caregivers; (5) evaluated technologies that act as a conduit to traditional care but add no potential therapeutic value, such as videoconferencing platforms used for remote psychotherapy; (6) evaluated interventions that used virtual reality technology; (7) did not have participants interacting directly with the technology for therapeutic purposes (eg, sensors used by clinicians to collect data without offering therapeutic feedback to participants); (8) had participants <18 years of age; (9) had participants using the technology exclusively in clinical or laboratory settings; or (10) were not completed at the time of searching (eg, study protocols captured in database searches).

Database Review

The review was conducted in 3 locations, PubMed, Web of Science (all databases in their Core Collection), and Ovid (all databases in their Health Sciences, Life Sciences, and Social Sciences categories). Search terms comprising both free text or natural language and Medical Subject Headings (MeSH) were applied in these searches (Multimedia Appendix 1). The searches were performed in May 2022 with no filters used to restrict the time period of the articles, and as a result, the oldest article in our search dated back to 1987. Reverse searches were also conducted by reviewing the references of relevant systematic reviews and meta-analyses, and by reviewing the references of included articles. Only articles with English abstracts were considered, but main texts could be in English, French, German, Italian, Spanish, or Chinese. Searches for gray literature using the same terms identified relevant published dissertations. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PROSPERO (International Prospective Register of Systematic Reviews) [20] guided this review. The review was conducted using conventions for systematic reviews, screening first by title and abstract, followed by reviewing the complete article text. Furthermore, 2 reviewers (CKYL and one of AS, BC, or DR) independently assessed each title and abstract, and each complete text, to determine inclusion or exclusion. Disagreements in these assessments were resolved with further discussion (involving the reviewers and BB-A). The database software Covidence (Veritas Health Innovation Ltd) was used to organize and review the studies [21].

Data Extraction and Synthesis

Data were captured from the included articles by the first author (CKYL) in a table format, checked by the last author (BB-A), and approved by all the authors. These data comprised the following: (1) depression or anxiety disorders needed for study eligibility, (2) details of the intervention (ie, mode of technology

XSL•FO
RenderX

used, program name if one exists, and treatment paradigm), (3) details of the participant sample (ie, sample size, breakdown by group allocation, and gender), (4) details of acceptability-related outcomes (ie, outcome name, instrument used, and time point of data collection), and (5) results of acceptability-related outcomes. From these data, we noted considerable variation in instruments used to measure acceptability and infrequent use of validated instruments. Acceptability outcome data were therefore narratively synthesized by the direction of acceptability and not by its effect size. The narrative synthesis applied vote counting to the tabulated results, was conducted by the first author (CKYL), and was checked by the last author (BB-A).

Levels of Evidence

Included studies were evaluated using the Oxford Centre for Evidence-Based Medicine–levels of evidence [22]. In this system, each study receives a designation based on a grading classification where 1 is the highest level (randomized controlled trials [RCTs] or systematic reviews of RCTs) and 5 is the lowest (mechanism-based reasoning).

Reported evidence for patient acceptability was further evaluated using an assessment tool designed for this review (Multimedia Appendix 2). The purpose of the assessment tool was to evaluate the quality of acceptability measurement and reporting. An assessment tool was developed because, to our knowledge, one did not already exist. Given the frequency (and variation) with which acceptability is reported as an evaluation outcome, the assessment tool allows summarization of our concerns with the current evidence. The tool emphasizes acquiring and providing sufficient information to evaluate, operationalize, and measure acceptability. Sekhon and colleagues' [16] recommendation that acceptability be measured at multiple time points informed item 4 of the tool. The assessment tool produces a score ranging from 0 to 15, where scores from 0 to 5 indicate low quality of evidence for patient acceptability, scores 6-10 indicate moderate quality, and scores 11-15 indicate high quality. Assessments using this tool were performed by one author (CKYL) and reviewed by another (BB-A), with discrepancies identified and consensus achieved through discussion.

Risk of Bias

Included studies were individually assessed on aspects of risk of bias using the assessment tool designed for this review (Multimedia Appendix 2). Specifically, the completeness of outcome data and the appropriateness of measurements (ie, selection of instruments) were determined using this assessment tool. These criteria were selected to assess each included study for their direct relevance to the single implementation outcome (ie, acceptability) that is the focus of this review. Preliminary assessments on these criteria indicated a significant risk of bias across the included studies, in terms of incomplete outcome data and low use of validated instruments.

Given these preliminary results suggesting a significant risk of bias, and this review's focus on a single implementation outcome, risk of bias was assessed using GRADE (Grading of Recommendations, Assessment, Development and Evaluation) [23]. GRADE assesses the quality of evidence for risk of bias across a group of studies in relation to a particular outcome [23]. Across a body of evidence, the risk of bias is assessed as low, unclear, or high [23]. This evaluation was done independently by 2 authors (CKYL and BB-A), with discrepancy resolved with further discussion.

Results

Overview

From the database review process, a total of 2409 articles were identified for screening. Of these articles, 2266 were excluded as they did not meet the inclusion criteria. Included in the review were 143 articles, investigating digital mental health interventions for depression (n=67, 47%), anxiety disorders (n=65, 45%), and a combination of depression and anxiety disorders (n=11, 8%; Figure 1 provides a flow diagram of the review).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for each step of the screening process.

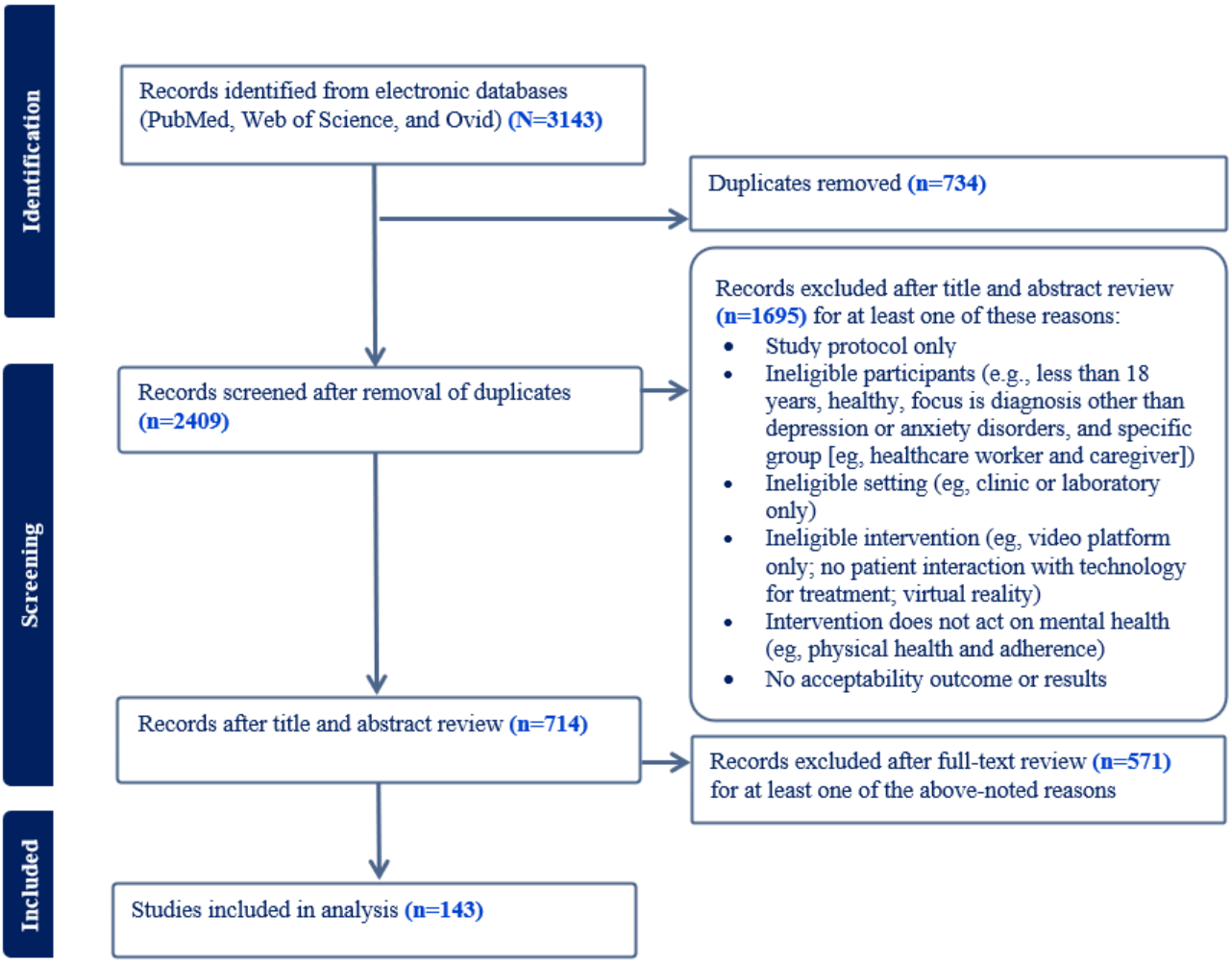

In these articles, researchers most often reported on satisfaction as their acceptability-related construct of choice, with satisfaction investigated in 105 (73%) of the 143 included articles. Out of 143 articles, acceptability or acceptance was reported on in 78 (55%) articles, credibility in 26 (18%) articles, expectancy in 20 (14%) articles, and usability in 19 (13%) articles. There were also 21 (15%) articles using such terminology as feedback, experiences, attitudes, subjective appraisal, and subjective benefit.

For the evidence provided for acceptability and its component constructs, the vast majority, at 137 (96%) out of 143 articles, evaluated acceptability "globally" for each digital mental health intervention (ie, across features or components), and 20 (14%) articles assessed acceptability for specific features or components of each digital mental health intervention.

Referring to the component constructs of acceptability identified in Sekhon and colleagues' [16] theoretical framework, each article investigated a subset of those component constructs. Specifically, out of 143 articles, 121 (85%) investigated affective attitude, 73 (51%) investigated perceived effectiveness, 37 (26%) investigated burden, and 35 (24%) investigated intervention coherence. The remaining component constructs were each investigated in 10% (15/143) or less of the articles (ie, opportunity cost: n=15, 10%; and self-efficacy: n=7, 5%). We did not find any articles that investigated ethicality.

Other constructs that were not explicit components of Sekhon and colleagues' [16] theoretical framework for acceptability but were investigated in the reviewed articles included perceived quality (17/143, 12%) and acceptability of program tempo or length (10/143, 7%).

In terms of temporal assessments of acceptability or its component constructs, each article captured one or more of prospective acceptability (before the intervention began), concurrent acceptability (during the intervention), and retrospective acceptability (following the intervention), also set out in Sekhon and colleagues' [16] theoretical framework. Prospective acceptability appeared in 42 (29%) of the 143 articles and concurrent acceptability in 9 (6%) articles. Retrospective acceptability was separated for this review into posttest (125/143, 87%) and follow-up (15/143, 10%). Of the total 143, only 43 (30%) articles contained acceptability assessments for more than 1 time point.

Assessment of Quality

Using our assessment tool for quality of acceptability (Multimedia Appendix 2), articles investigating digital mental health interventions for depression had scores ranging from 2 (low quality of evidence) to 14 (high quality of evidence), out of a total possible 15. The mean score for the quality of evidence for acceptability was 7.3 (SD 2.0). From the information captured with this assessment tool, out of 143 articles, 74 (53%)

XSL•FO
RenderX

used one or more validated instruments for acceptability or its component constructs, while 66 (47%) did not use any validated instruments for acceptability or its component constructs.

Risk of Bias

The overall risk of bias was assessed as between high and unclear, considering the predominance of studies that lacked blinding (for RCTs), failed to adequately control confounding, had incomplete accounting of patients and outcome events (particularly losses to follow-up), and other limitations (eg, used measures that are not validated).

Interventions for Depression

From the 67 articles addressing digital mental health interventions for depression, 40 (60%) articles involved web-based programs, 21 (31%) involved mobile apps, and 6 (9%) involved computer-based programs. In terms of treatment modality, 36 (54%) articles explored digital mental health interventions that drew from cognitive behavioral therapy (CBT). These apps or programs were often structured as multimedia modules, with varying levels of clinician or technician support and patient feedback.

The 67 articles investigating digital mental health interventions for depression were comprised of 41 (61%) RCTs, 23 (34%) non-RCT quantitative studies, and 3 (5%) qualitative studies. These studies are reviewed in Table S1 [24-65], Table S2 [66-89], and Table S3 [43,90-92], respectively, in Multimedia Appendix 3. The 64 quantitative studies (RCT and non-RCT) indicated overall positive ratings of acceptability among digital mental health interventions for depression. Among these 64 quantitative studies, we categorized 55 (86%) studies as indicating positive acceptability overall, 6 (9%) studies mixed acceptability, and 3 (5%) studies as having insufficient information to categorize the direction of acceptability. We considered results to be “mixed” if a study had more than 1 acceptability outcome measure and at least 1 positive and 1 negative result. These categorizations are presented in the “Summarized acceptability result” column within each table in Multimedia Appendix 3 [24-92].

The 3 qualitative studies (completed secondarily or subsequently to RCTs) had more mixed results. In 2014, Schneider and colleagues [90] reported that open-ended participant responses on satisfaction levels were more negative than positive, though satisfaction or dissatisfaction were reported as arising from matters related and unrelated to the interventions, such as individual time management choices made during the study. Comments specific to the web-based program were more positive than negative. Knowles and colleagues [91] reported in 2015 that they conducted interviews with participants using digital mental health interventions for depression (MoodGYM [e-hub Health] and Beating the Blues [Manage My Health], both web-based programs in CBT). Responses were coded as positive, negative, or ambivalent toward the digital mental health intervention, with the largest group being ambivalent (17/36 participants, 47%), then negative (10/36 participants, 28%), and then positive (9/36 participants, 25%). Also in 2015, Ly and colleagues [92] held semistructured interviews with participants using a mobile app informed by CBT and mindfulness techniques. Participants’ responses were divided, with 5 (42%) of 12 participants indicating an overall positive experience, 4 (33%) indicating a neutral experience, and 3 (25%) indicating a negative experience.

Interventions for Anxiety Disorders

From the 65 articles addressing digital mental health interventions for anxiety, 52 (80%) articles involved web-based programs, 8 (12%) involved mobile apps, 3 (5%) involved computer-based programs, and 2 (3%) involved others. In terms of treatment modality, 58 (89%) articles explored digital mental health interventions that drew from CBT.

The 65 articles investigating digital mental health interventions for anxiety disorders were comprised of 43 (66%) RCTs and 22 (34%) non-RCT quantitative studies. None were qualitative studies. The studies are reviewed in Table S4 [93-135] and Table S5 [136-157] in Multimedia Appendix 3. In general, the results indicated positive ratings of acceptability. We categorized 59 (91%) studies as indicating positive acceptability overall, 3 (5%) studies mixed acceptability, and 3 (5%) studies as having insufficient information to categorize the direction of acceptability. We considered results to be “mixed” if a study had more than 1 acceptability outcome measure and at least 1 positive and 1 negative result. These categorizations are presented in the “Summarized acceptability result” column within each table in Multimedia Appendix 3 [93-157].

Interventions for Depression and Anxiety Disorders Together

Among the 11 articles addressing digital mental health interventions for both depression and anxiety disorders, 8 (73%) articles involved web-based programs, 2 (18%) involved mobile apps, and 1 (9%) involved computer-based programs. In terms of treatment modality, 8 (73%) articles explored digital mental health interventions that drew from CBT.

The 11 articles investigating digital mental health interventions for both depression and anxiety disorders were comprised of 6 (55%) RCTs and 5 (45%) non-RCT quantitative studies (reviewed in Table S6 [158-163] and Table S7 [164-168] in Multimedia Appendix 3). None were qualitative studies. In general, the results indicated positive ratings of acceptability. We categorized all 11 studies as indicating positive acceptability overall, with none that indicated mixed results or insufficient information to categorize the direction of acceptability. These categorizations are presented in the “Summarized acceptability result” column within each table in Multimedia Appendix 3 [158-168].

Discussion

Principal Findings

Digital mental health interventions for depression and anxiety disorders were generally found to be acceptable to patients. Of the 143 included articles, 125 (87%) indicated positive acceptability, 12 (9%) had mixed results, and 6 had insufficient information (4%). These results arose out of evidence generated primarily from RCTs (90/143, 63%) and other quantitative studies (50/143, 35%), and largely investigations of web-based

XSL•FO
RenderX

apps (100/143, 70%) and CBT-based content (102/143, 71%). These findings came primarily from views on a digital mental health intervention as a whole (137/143, 96%) rather than specific features, and centered on patient satisfaction (105/143, 73%) or Sekhon and colleagues' [16] "affective attitude" (121/143, 85%) at posttest (125/143, 87%). We noted no substantial differences in the acceptability noted for digital mental health interventions for depression, anxiety, or depression and anxiety together.

Our findings suggest some receptiveness among those with depression or anxiety disorders to digital mental health interventions. If the overall positive experience of digital mental health interventions translates into patient adherence to and engagement with treatment, digital mental health interventions can improve patient access to care and expand the capacity of health systems. Since some patients appear open to digital mental health interventions, clinicians can explore patients' views on these technologies in the course of treatment planning and respond accordingly. Researchers can further our understanding by investigating patient and intervention characteristics that predict or improve acceptance and deepen our understanding of acceptability beyond satisfaction or affective attitudes. Developers can recognize the potential in continuing to invest in digital mental health interventions and take steps to improve user experiences and strengthen acceptability.

The findings of positive acceptability toward digital mental health interventions must nevertheless be qualified by gaps in the existing research.

First, the assessment tool revealed that studies had a wide range in quality of acceptability-related evidence and were generally only of moderate quality. This finding suggests a need to increase standardization of acceptability assessments. We propose that our assessment tool be used as a checklist when planning evaluation research. The results outlined above indicate that there is an opportunity to use validated measures for acceptability, measure acceptability at different time points, and consider measures that address multiple acceptability components.

The second gap is the continued absence of a consensus definition for acceptability. As an example, Sekhon and colleagues [16] make clear that satisfaction ought to be treated as a construct separate from acceptability, but many studies continue to use satisfaction as an indicator of acceptability, or otherwise consider satisfaction alone when perhaps the construct of interest is acceptability. Similarly, Sekhon and colleagues [16] identified 7 component constructs of acceptability, but studies are generally designed to address only some of these components and vary by the components chosen. These findings suggest that there may not only be variation in how acceptability is investigated, but also variation in how acceptability is understood. It is necessary to establish a consensus on how acceptability should be defined, to increase the consistency and relevance of future acceptability research.

Third, some aspects of acceptability such as affective attitude and perceived effectiveness were more frequently investigated, while other aspects were far less likely to be explored, particularly ethicality, self-efficacy, and opportunity cost. Likely, this relates to some component constructs being considered more important than others, or the feasibility of investigating some component constructs over others. Investigating all components of acceptability is important to better understand how users relate to digital mental health interventions.

Finally, we noted a paucity of research on how acceptability changes over time, with most studies examining only retrospective acceptability. Evaluating acceptability as a variable that is continuous and changeable over time would enhance development processes for digital mental health interventions. It would allow us to identify the modifications that can be made to a technology to strengthen acceptability at different stages, to increase willingness to initiate use of a digital mental health intervention, to continue using it, and to use it as frequently or often as is needed to reap therapeutic benefits.

Limitations

An important limitation of research on acceptability, as identified in the risk of bias assessment, lies with evidence generally being provided by patients who have agreed to use the digital mental health intervention and continued to engage with the study or the digital mental health intervention. Often missing from these findings is the acceptability of digital mental health interventions among individuals who had the opportunity to participate in a study involving a digital mental health intervention but chose not to. Nonparticipation and early termination in digital health have long prompted calls for more research attention on attrition, for example, Eysenbach [169] in 2005, along with related constructs such as adherence and engagement. There remains a need to establish consensus approaches to analyzing attrition, adherence, and engagement [170], and their relationship with acceptability. It is recommended that appropriate methods for such analyses be developed, to be used consistently in future evaluations of digital mental health interventions.

A second limitation comes from potential conflict in the research evidence between the acceptability of the intervention technology and the acceptability of the intervention as a whole. Future research should differentiate acceptability of the technology from the acceptability of other aspects, such as the timing or pacing of the intervention, or the treatment paradigm applied.

A further limitation is that the terminology used to identify acceptability is varied and changeable between studies. Data on acceptability not labelled as such may be difficult to integrate into acceptability research at large.

Finally, literature on this topic has increased exponentially following the COVID-19 pandemic [171]. Any review can only represent a snapshot of the available research at a point in time. At the same time, despite this proliferation of research, diverse populations (in terms of ethnicity, income and education levels, sexual and gender minority status, etc) remain underrepresented [172]. More evaluation studies are needed that ensure their inclusion and provide demographic data as appropriate, to

XSL•FO
RenderX

further our understanding of their acceptability of digital mental health interventions.

Conclusions

There is evidence for the acceptability of digital mental health interventions among patients with depression or anxiety disorders, with most studies in this review indicating positive acceptability. However, the available research evidence for acceptability is, as a whole, of only moderate quality. It is suggested that future intervention studies be planned from the outset to produce higher quality evidence for acceptability. Enhanced research evidence for acceptability will provide better insights into the appropriateness of digital mental health interventions for patients with depression or anxiety disorders.

Acknowledgments

Support was provided by Emily Stowe Society funding at Women's College Hospital Foundation (a nonprofit foundation).

Authors' Contributions

Each author contributed substantially to this work. BB-A conceived of the review and all other authors provided feedback on the design. CKYL, AS, BC, and DR were involved in data collection. All authors participated in data analysis and interpretation. BB-A and CKYL wrote the initial draft of the paper. All authors contributed to revisions of the manuscript. All authors read and approved the final submitted manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOC File , 65 KB - jmir_v26i1e52609_app1.doc]

Multimedia Appendix 2
Acceptability assessment tool.
[DOC File , 61 KB - jmir_v26i1e52609_app2.doc]

Multimedia Appendix 3
Description of acceptability results in the included studies.
[DOCX File , 86 KB - jmir_v26i1e52609_app3.docx]

Multimedia Appendix 4
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 282 KB - jmir_v26i1e52609_app4.pdf]

Multimedia Appendix 5
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist for abstracts.
[PDF File (Adobe PDF File), 267 KB - jmir_v26i1e52609_app5.pdf]

References

1. Report from the Canadian Chronic Disease Surveillance System: Mood and Anxiety Disorders in Canada, 2016. Public Health Agency of Canada. 2016. URL: https://www.canada.ca/content/dam/canada/health-canada/migration/healthy-canadians/publications/diseases-conditions-maladies-affections/mood-anxiety-disorders-2016-troubles-anxieux-humeur/alt/mood-anxiety-disorders-2016-troubles-anxieux-humeur-eng.pdf [accessed 2023-07-08]
2. Greenberg PE, Fournier AA, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics 2021 Jun;39(6):653-665 [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
3. Konnopka A, Leichsenring F, Leibing E, König HH. Cost-of-illness studies and cost-effectiveness analyses in anxiety disorders: a systematic review. J Affect Disord 2009 Apr;114(1-3):14-31. [doi: 10.1016/j.jad.2008.07.014] [Medline: 18768222]
4. Lal S. E-mental health: Promising advancements in policy, research, and practice. Healthc Manage Forum 2019 Mar;32(2):56-62 [FREE Full text] [doi: 10.1177/0840470418818583] [Medline: 30739487]
5. Moroz N, Moroz I, D'Angelo MS. Mental health services in Canada: barriers and cost-effective solutions to increase access. Healthc Manage Forum 2020 Nov;33(6):282-287. [doi: 10.1177/0840470420933911] [Medline: 32613867]

XSL•FO
RenderX

6. Lal S, Adair CE. E-mental health: a rapid review of the literature. Psychiatr Serv 2014 Jan 01;65(1):24-32. [doi: 10.1176/appi.ps.201300009] [Medline: 24081188]
7. Gratzer D, Torous J, Lam RW, Patten SB, Kutcher S, Chan S, et al. Our digital moment: innovations and opportunities in digital mental health care. Can J Psychiatry 2021 Jan;66(1):5-8 [FREE Full text] [doi: 10.1177/0706743720937833] [Medline: 32603188]
8. Sorkin DH, Janio EA, Eikey EV, Schneider M, Davis K, Schueller SM, et al. Rise in use of digital mental health tools and technologies in the United States during the COVID-19 pandemic: survey study. J Med Internet Res 2021 Apr 16;23(4):e26994 [FREE Full text] [doi: 10.2196/26994] [Medline: 33822737]
9. Vadivel R, Shoib S, El Halabi S, El Hayek S, Essam L, Gashi Bytyçi D, et al. Mental health in the post-COVID-19 era: challenges and the way forward. Gen Psychiatr 2021;34(1):e100424 [FREE Full text] [doi: 10.1136/gpsych-2020-100424] [Medline: 33644689]
10. Strudwick G, Sockalingam S, Kassam I, Sequeira L, Bonato S, Youssef A, et al. Digital interventions to support population mental health in Canada during the COVID-19 pandemic: rapid review. JMIR Ment Health 2021 Mar 02;8(3):e26550 [FREE Full text] [doi: 10.2196/26550] [Medline: 33650985]
11. Wellness Together Canada. URL: https://www.wellnesstogether.ca/ [accessed 2023-07-22]
12. Kelders SM, Kok RN, Ossebaard HC, van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
13. Borghouts J, Eikey E, Mark G, de Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
14. Gulliver A, Calear AL, Sunderland M, Kay-Lambkin F, Farrer LM, Batterham PJ. Predictors of acceptability and engagement in a self-guided online program for depression and anxiety. Internet Interv 2021 Sep;25:100400 [FREE Full text] [doi: 10.1016/j.invent.2021.100400] [Medline: 34026569]
15. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
16. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
17. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
18. Zauszniewski JA. Intervention development: assessing critical parameters from the intervention recipient's perspective. Appl Nurs Res 2012 Feb;25(1):31-39. [doi: 10.1016/j.apnr.2010.06.002] [Medline: 20974101]
19. Davis FD. A technology acceptance model for empirically testing new end-user information systems : theory and results. Dissertation Massachusetts Institute of Technology. 1986. URL: http://hdl.handle.net/1721.1/15192 [accessed 2023-07-08]
20. Welcome to the new preferred reporting items for systematic reviews and meta-analyses (PRISMA) website. PRISMA. URL: http://www.prisma-statement.org/ [accessed 2023-07-04]
21. Covidence. URL: https://www.covidence.org/ [accessed 2023-07-04]
22. Oxford Centre for Evidence-Based Medicine: levels of evidence (March 2009). Oxford Centre for Evidence-Based Medicine. URL: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009 [accessed 2023-07-22]
23. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE handbook. GRADEpro. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2024-02-18]
24. Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther 2011;40(4):251-266. [doi: 10.1080/16506073.2011.616531] [Medline: 22060248]
25. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR Mhealth Uhealth 2016 Jan 26;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]
26. Blackwell SE, Browning M, Mathews A, Pictet A, Welch J, Davies J, et al. Positive imagery-based cognitive bias modification as a web-based treatment tool for depressed adults: a randomized controlled trial. Clin Psychol Sci 2015 Jan;3(1):91-111 [FREE Full text] [doi: 10.1177/2167702614560746] [Medline: 25984421]
27. Bücker L, Schnakenberg P, Karyotaki E, Moritz S, Westermann S. Diminishing effects after recurrent use of self-guided internet-based interventions in depression: randomized controlled trial. J Med Internet Res 2019 Oct 02;21(10):e14240 [FREE Full text] [doi: 10.2196/14240] [Medline: 31579014]
28. Buntrock C, Ebert D, Lehr D, Riper H, Smit F, Cuijpers P, et al. Effectiveness of a web-based cognitive behavioural intervention for subthreshold depression: pragmatic randomised controlled trial. Psychother Psychosom 2015;84(6):348-358. [doi: 10.1159/000438673] [Medline: 26398885]

XSL•FO
RenderX

29. Burton C, Szentagotai Tatar A, McKinstry B, Matheson C, Matu S, Moldovan R, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare 2016 Sep;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]
30. de Graaf LE, Huibers MJH, Riper H, Gerhards SAH, Arntz A. Use and acceptability of unsupported online computerized cognitive behavioral therapy for depression and associations with clinical outcome. J Affect Disord 2009 Aug;116(3):227-231. [doi: 10.1016/j.jad.2008.12.009] [Medline: 19167094]
31. Geraedts AS, Kleiboer AM, Wiezer NM, Cuijpers P, van Mechelen W, Anema JR. Feasibility of a worker-directed web-based intervention for employees with depressive symptoms. Internet Interventions 2014 Jul;1(3):132-140. [doi: 10.1016/j.invent.2014.07.001]
32. Gómez Penedo JM, Babl AM, Grosse Holtforth M, Hohagen F, Krieger T, Lutz W, et al. The association of therapeutic alliance with long-term outcome in a guided internet intervention for depression: secondary analysis from a randomized control trial. J Med Internet Res 2020 Mar 24;22(3):e15824 [FREE Full text] [doi: 10.2196/15824] [Medline: 32207689]
33. Høifødt RS, Lillevoll KR, Griffiths KM, Wilsgaard T, Eisemann M, Waterloo K, et al. The clinical effectiveness of web-based cognitive behavioral therapy with face-to-face therapist support for depressed primary care patients: randomized controlled trial. J Med Internet Res 2013 Aug 05;15(8):e153 [FREE Full text] [doi: 10.2196/jmir.2714] [Medline: 23916965]
34. Johansson O, Bjärehed J, Andersson G, Carlbring P, Lundh LG. Effectiveness of guided internet-delivered cognitive behavior therapy for depression in routine psychiatry: a randomized controlled trial. Internet Interv 2019 Sep;17:100247 [FREE Full text] [doi: 10.1016/j.invent.2019.100247] [Medline: 31249791]
35. Kelders SM, Bohlmeijer ET, Pots WTM, van Gemert-Pijnen JEWC. Comparing human and automated support for depression: Fractional factorial randomized controlled trial. Behav Res Ther 2015 Sep;72:72-80. [doi: 10.1016/j.brat.2015.06.014] [Medline: 26196078]
36. Kenter RMF, Cuijpers P, Beekman A, van Straten A. Effectiveness of a web-based guided self-help intervention for outpatients with a depressive disorder: short-term results from a randomized controlled trial. J Med Internet Res 2016 Mar 31;18(3):e80 [FREE Full text] [doi: 10.2196/jmir.4861] [Medline: 27032449]
37. Kok G, Bockting C, Burger H, Smit F, Riper H. Mobile cognitive therapy: adherence and acceptability of an online intervention in remitted recurrently depressed patients. Internet Interventions 2014 Apr;1(2):65-73. [doi: 10.1016/j.invent.2014.05.002]
38. Lappalainen P, Langrial S, Oinas-Kukkonen H, Tolvanen A, Lappalainen R. Web-based acceptance and commitment therapy for depressive symptoms with minimal support: a randomized controlled trial. Behav Modif 2015 Nov;39(6):805-834. [doi: 10.1177/0145445515598142] [Medline: 26253644]
39. Levin W, Campbell DR, McGovern KB, Gau JM, Kosty DB, Seeley JR, et al. A computer-assisted depression intervention in primary care. Psychol Med 2011 Jul;41(7):1373-1383. [doi: 10.1017/S0033291710001935] [Medline: 20961474]
40. Löbner M, Pabst A, Stein J, Dorow M, Matschinger H, Luppa M, et al. Computerized cognitive behavior therapy for patients with mild to moderately severe depression in primary care: a pragmatic cluster randomized controlled trial (@ktiv). J Affect Disord 2018 Oct 01;238:317-326. [doi: 10.1016/j.jad.2018.06.008] [Medline: 29902736]
41. Lüdtke T, Pult LK, Schröder J, Moritz S, Bücker L. A randomized controlled trial on a smartphone self-help application (Be Good to Yourself) to reduce depressive symptoms. Psychiatry Res 2018 Nov;269:753-762. [doi: 10.1016/j.psychres.2018.08.113] [Medline: 30273901]
42. Lukas CA, Berking M. Blending group-based psychoeducation with a smartphone intervention for the reduction of depressive symptoms: results of a randomized controlled pilot study. Pilot Feasibility Stud 2021 Feb 24;7(1):57 [FREE Full text] [doi: 10.1186/s40814-021-00799-y] [Medline: 33627198]
43. Lukas CA, Eskofier B, Berking M. A gamified smartphone-based intervention for depression: randomized controlled pilot trial. JMIR Ment Health 2021 Jul 20;8(7):e16643 [FREE Full text] [doi: 10.2196/16643] [Medline: 34283037]
44. Ly KH, Trüschel A, Jarl L, Magnusson S, Windahl T, Johansson R, et al. Behavioural activation versus mindfulness-based guided self-help treatment administered through a smartphone application: a randomised controlled trial. BMJ Open 2014 Jan 09;4(1):e003440 [FREE Full text] [doi: 10.1136/bmjopen-2013-003440] [Medline: 24413342]
45. Ly KH, Topooco N, Cederlund H, Wallin A, Bergström J, Molander O, et al. Smartphone-supported versus full behavioural activation for depression: a randomised controlled trial. PLoS One 2015;10(5):e0126559 [FREE Full text] [doi: 10.1371/journal.pone.0126559] [Medline: 26010890]
46. Meyer B, Berger T, Caspar F, Beevers CG, Andersson G, Weiss M. Effectiveness of a novel integrative online treatment for depression (Deprexis): randomized controlled trial. J Med Internet Res 2009 May 11;11(2):e15 [FREE Full text] [doi: 10.2196/jmir.1151] [Medline: 19632969]
47. Meyer B, Bierbrodt J, Schröder J, Berger T, Beevers CG, Weiss M, et al. Effects of an internet intervention (Deprexis) on severe depression symptoms: randomized controlled trial. Internet Interventions 2015 Mar;2(1):48-59. [doi: 10.1016/j.invent.2014.12.003]
48. Moritz S, Schilling L, Hauschildt M, Schröder J, Treszl A. A randomized controlled trial of internet-based therapy in depression. Behav Res Ther 2012 Aug;50(7-8):513-521. [doi: 10.1016/j.brat.2012.04.006] [Medline: 22677231]

XSL•FO
RenderX

49. Motter JN, Grinberg A, Lieberman DH, Iqnaibi WB, Sneed JR. Computerized cognitive training in young adults with depressive symptoms: effects on mood, cognition, and everyday functioning. J Affect Disord 2019 Feb 15;245:28-37. [doi: 10.1016/j.jad.2018.10.109] [Medline: 30366235]
50. Oehler C, Görges F, Rogalla M, Rummel-Kluge C, Hegerl U. Efficacy of a guided web-based self-management intervention for depression or dysthymia: randomized controlled trial with a 12-month follow-up using an active control condition. J Med Internet Res 2020 Jul 14;22(7):e15361 [FREE Full text] [doi: 10.2196/15361] [Medline: 32673233]
51. Perini S, Titov N, Andrews G. Clinician-assisted internet-based treatment is effective for depression: randomized controlled trial. Aust N Z J Psychiatry 2009 Jun;43(6):571-578. [doi: 10.1080/00048670902873722] [Medline: 19440890]
52. Pinto MD, Greenblatt AM, Hickman RL, Rice HM, Thomas TL, Clochesy JM. Assessing the critical parameters of eSMART-MH: a promising avatar-based digital therapeutic intervention to reduce depressive symptoms. Perspect Psychiatr Care 2016 Jul;52(3):157-168. [doi: 10.1111/ppc.12112] [Medline: 25800698]
53. Reins JA, Boß L, Lehr D, Berking M, Ebert DD. The more I got, the less I need? Efficacy of Internet-based guided self-help compared to online psychoeducation for major depressive disorder. J Affect Disord 2019 Mar 01;246:695-705. [doi: 10.1016/j.jad.2018.12.065] [Medline: 30611913]
54. Ruwaard J, Schrieken B, Schrijver M, Broeksteeg J, Dekker J, Vermeulen H, et al. Standardized web-based cognitive behavioural therapy of mild to moderate depression: a randomized controlled trial with a long-term follow-up. Cogn Behav Ther 2009;38(4):206-221. [doi: 10.1080/16506070802408086] [Medline: 19221919]
55. Smith J, Newby JM, Burston N, Murphy MJ, Michael S, Mackenzie A, et al. Help from home for depression: a randomised controlled trial comparing internet-delivered cognitive behaviour therapy with bibliotherapy for depression. Internet Interv 2017 Sep;9:25-37 [FREE Full text] [doi: 10.1016/j.invent.2017.05.001] [Medline: 30135834]
56. Stephens VC. Effects on depressive symptoms of a web-based cognitive bias modification-interpretation (CBM-I) program for emotion recognition: a randomised controlled trial. University of Exeter. URL: http://hdl.handle.net/10871/16430 [accessed 2024-09-17]
57. Adams S, Penton-Voak IS, Harmer CJ, Holmes EA, Munafò MR. Effects of emotion recognition training on mood among individuals with high levels of depressive symptoms: study protocol for a randomised controlled trial. Trials 2013 Jun 01;14:161 [FREE Full text] [doi: 10.1186/1745-6215-14-161] [Medline: 23725208]
58. Titov N, Dear BF, Staples LG, Terides MD, Karin E, Sheehan J, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: a randomized controlled trial. J Anxiety Disord 2015 Oct;35:88-102 [FREE Full text] [doi: 10.1016/j.janxdis.2015.08.002] [Medline: 26422822]
59. Titov N, Andrews G, Davies M, McIntyre K, Robinson E, Solley K. Internet treatment for depression: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010 Jun 08;5(6):e10939 [FREE Full text] [doi: 10.1371/journal.pone.0010939] [Medline: 20544030]
60. Tønning ML, Faurholt-Jepsen M, Frost M, Martiny K, Tuxen N, Rosenberg N, et al. The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: The RADMIS randomized controlled trial. J Affect Disord 2021 Mar 01;282:354-363. [doi: 10.1016/j.jad.2020.12.141] [Medline: 33421863]
61. Wahbeh H. Internet mindfulness meditation intervention (IMMI) improves depression symptoms in older adults. Medicines (Basel) 2018 Nov 02;5(4):119 [FREE Full text] [doi: 10.3390/medicines5040119] [Medline: 30400211]
62. Watts S, Mackenzie A, Thomas C, Griskaitis A, Mewton L, Williams A, et al. CBT for depression: a pilot RCT comparing mobile phone vs. computer. BMC Psychiatry 2013 Feb 07;13:49 [FREE Full text] [doi: 10.1186/1471-244X-13-49] [Medline: 23391304]
63. Williams AD, Blackwell SE, Mackenzie A, Holmes EA, Andrews G. Combining imagination and reason in the treatment of depression: a randomized controlled trial of internet-based cognitive-bias modification and internet-CBT for depression. J Consult Clin Psychol 2013 Oct;81(5):793-799 [FREE Full text] [doi: 10.1037/a0033247] [Medline: 23750459]
64. Williams AD, O'Moore K, Blackwell SE, Smith J, Holmes EA, Andrews G. Positive imagery cognitive bias modification (CBM) and internet-based cognitive behavioral therapy (iCBT): a randomized controlled trial. J Affect Disord 2015 Jun 01;178:131-141 [FREE Full text] [doi: 10.1016/j.jad.2015.02.026] [Medline: 25805405]
65. Wong VWH, Ho FYY, Shi NK, Tong JTY, Chung KF, Yeung WF, et al. Smartphone-delivered multicomponent lifestyle medicine intervention for depressive symptoms: a randomized controlled trial. J Consult Clin Psychol 2021 Dec;89(12):970-984. [doi: 10.1037/ccp0000695] [Medline: 35025538]
66. Addington EL, Cheung EO, Bassett SM, Kwok I, Schuette SA, Shiu E, et al. The MARIGOLD study: feasibility and enhancement of an online intervention to improve emotion regulation in people with elevated depressive symptoms. J Affect Disord 2019 Oct 01;257:352-364 [FREE Full text] [doi: 10.1016/j.jad.2019.07.049] [Medline: 31302525]
67. Berman MI, Buckey JC, Hull JG, Linardatos E, Song SL, McLellan RK, et al. Feasibility study of an interactive multimedia electronic problem solving treatment program for depression: a preliminary uncontrolled trial. Behav Ther 2014 May;45(3):358-375 [FREE Full text] [doi: 10.1016/j.beth.2014.02.001] [Medline: 24680231]

XSL•FO
RenderX

68. Burns MN, Begale M, Duffecy J, Gergle D, Karr CJ, Giangrande E, et al. Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res 2011 Aug 12;13(3):e55 [FREE Full text] [doi: 10.2196/jmir.1838] [Medline: 21840837]
69. Callan JA, Dunbar Jacob J, Siegle GJ, Dey A, Thase ME, DeVito Dabbs A, et al. CBT MobileWork©: user-centered development and testing of a mobile mental health application for depression. Cogn Ther Res 2020 Sep 25;45(2):287-302. [doi: 10.1007/s10608-020-10159-4]
70. Caplan S, Sosa Lovera A, Reyna Liberato P. A feasibility study of a mental health mobile app in the Dominican Republic: the untold story. International Journal of Mental Health 2019 Apr 04;47(4):311-345. [doi: 10.1080/00207411.2018.1553486]
71. Cartreine JA, Locke SE, Buckey JC, Sandoval L, Hegel MT. Electronic problem-solving treatment: description and pilot study of an interactive media treatment for depression. JMIR Res Protoc 2012 Sep 25;1(2):e11 [FREE Full text] [doi: 10.2196/resprot.1925] [Medline: 23611902]
72. Coutinho E, Alshukri A, de Berardinis J, Dowrick C. POLYHYMNIA mood – empowering people to cope with depression through music listening. 2021 Presented at: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers; September 24, 2021; Virtual event p. 188-193. [doi: 10.1145/3460418.3479334]
73. Dahne J, Kustanowitz J, Lejuez CW. Development and preliminary feasibility study of a brief behavioral activation mobile application (Behavioral Apptivation) to be used in conjunction with ongoing therapy. Cogn Behav Pract 2018 Feb;25(1):44-56 [FREE Full text] [doi: 10.1016/j.cbpra.2017.05.004] [Medline: 29725198]
74. Dear BF, Zou J, Titov N, Lorian C, Johnston L, Spence J, et al. Internet-delivered cognitive behavioural therapy for depression: a feasibility open trial for older adults. Aust N Z J Psychiatry 2013 Feb;47(2):169-176. [doi: 10.1177/0004867412466154] [Medline: 23152358]
75. Drake G, Csipke E, Wykes T. Assessing your mood online: acceptability and use of Moodscope. Psychol Med 2013 Jul;43(7):1455-1464 [FREE Full text] [doi: 10.1017/S0033291712002280] [Medline: 23149120]
76. Fuller-Tyszkiewicz M, Richardson B, Klein B, Skouteris H, Christensen H, Austin D, et al. A mobile app-based intervention for depression: end-user and expert usability testing study. JMIR Ment Health 2018 Aug 23;5(3):e54 [FREE Full text] [doi: 10.2196/mental.9445] [Medline: 30139722]
77. Gould CE, Carlson C, Ma F, Forman-Hoffman V, Ranta K, Kuhn E. Effects of mobile app-based intervention for depression in middle-aged and older adults: mixed methods feasibility study. JMIR Form Res 2021 Jun 29;5(6):e25808 [FREE Full text] [doi: 10.2196/25808] [Medline: 34185000]
78. Harper Shehadeh MJ, Abi Ramia J, Cuijpers P, El Chammay R, Heim E, Kheir W, et al. Step-by-step, an e-mental health intervention for depression: a mixed methods pilot study from Lebanon. Front Psychiatry 2019;10:986 [FREE Full text] [doi: 10.3389/fpsyt.2019.00986] [Medline: 32116815]
79. Jacmon J, Malouff J, Taylor N. Treatment of major depression: effectiveness of cognitive behavior therapy with an internet course as a central component. E-J Appl Psychol 2010;5(2):1-8.
80. Kooistra LC, Ruwaard J, Wiersma JE, van Oppen P, van der Vaart R, van Gemert-Pijnen JE, et al. Development and initial evaluation of blended cognitive behavioural treatment for major depression in routine specialized mental health care. Internet Interv 2016 May;4:61-71 [FREE Full text] [doi: 10.1016/j.invent.2016.01.003] [Medline: 30135791]
81. Mayer G, Hummel S, Oetjen N, Gronewold N, Bubolz S, Blankenhagel K, et al. User experience and acceptance of patients and healthy adults testing a personalized self-management app for depression: a non-randomized mixed-methods feasibility study. Digit Health 2022;8:20552076221091353 [FREE Full text] [doi: 10.1177/20552076221091353] [Medline: 35425641]
82. Schlosser DA, Campellone TR, Truong B, Anguera JA, Vergani S, Vinogradov S, et al. The feasibility, acceptability, and outcomes of PRIME-D: a novel mobile intervention treatment for depression. Depress Anxiety 2017 Jun;34(6):546-554 [FREE Full text] [doi: 10.1002/da.22624] [Medline: 28419621]
83. Schuster R, Kalthoff I, Walther A, Köhldorfer L, Partinger E, Berger T, et al. Effects, adherence, and therapists' perceptions of web- and mobile-supported group therapy for depression: mixed-methods study. J Med Internet Res 2019 Apr 28;21(5):e11860 [FREE Full text] [doi: 10.2196/11860] [Medline: 31066700]
84. Stiles-Shields C, Montague E, Kwasny MJ, Mohr DC. Behavioral and cognitive intervention strategies delivered via coached apps for depression: Pilot trial. Psychol Serv 2019 May;16(2):233-238 [FREE Full text] [doi: 10.1037/ser0000261] [Medline: 30407055]
85. Strauss C, Dunkeld C, Cavanagh K. Is clinician-supported use of a mindfulness smartphone app a feasible treatment for depression? A mixed-methods feasibility study. Internet Interv 2021 Sep;25:100413 [FREE Full text] [doi: 10.1016/j.invent.2021.100413] [Medline: 34401372]
86. Elliott R, Slatick E, Urman M. Qualitative change process research on psychotherapy: alternative strategies. Psychologische Beitrage 2001;43(3):69-111.
87. Tomasino KN, Lattie EG, Ho J, Palac HL, Kaiser SM, Mohr DC. Harnessing peer support in an online intervention for older adults with depression. Am J Geriatr Psychiatry 2017 Oct;25(10):1109-1119 [FREE Full text] [doi: 10.1016/j.jagp.2017.04.015] [Medline: 28571785]
88. Walsh SM. Development of an online intervention using positive psychology for depression. Queen Mary University of London. UK; 2018. URL: http://qmro.qmul.ac.uk/xmlui/handle/123456789/31871 [accessed 2024-09-17]

XSL•FO
RenderX

89. Yeung A, Wang F, Feng F, Zhang J, Cooper A, Hong L, et al. Outcomes of an online computerized cognitive behavioral treatment program for treating Chinese patients with depression: a pilot study. Asian J Psychiatr 2018 Dec;38:102-107. [doi: 10.1016/j.ajp.2017.11.007] [Medline: 29146042]
90. Schneider J, Sarrami Foroushani P, Grime P, Thornicroft G. Acceptability of online self-help to people with depression: users' views of MoodGYM versus informational websites. J Med Internet Res 2014 Mar 28;16(3):e90 [FREE Full text] [doi: 10.2196/jmir.2871] [Medline: 24681717]
91. Knowles SE, Lovell K, Bower P, Gilbody S, Littlewood E, Lester H. Patient experience of computerised therapy for depression in primary care. BMJ Open 2015 Nov 30;5(11):e008581 [FREE Full text] [doi: 10.1136/bmjopen-2015-008581] [Medline: 26621513]
92. Ly KH, Janni E, Wrede R, Sedem M, Donker T, Carlbring P, et al. Experiences of a guided smartphone-based behavioral activation therapy for depression: a qualitative study. Internet Interventions 2015 Mar;2(1):60-68. [doi: 10.1016/j.invent.2014.12.002]
93. Allen AR, Newby JM, Mackenzie A, Smith J, Boulton M, Loughnan SA, et al. Internet cognitive-behavioural treatment for panic disorder: randomised controlled trial and evidence of effectiveness in primary care. BJPsych Open 2016 Mar;2(2):154-162 [FREE Full text] [doi: 10.1192/bjpo.bp.115.001826] [Medline: 27703768]
94. Bell CJ, Colhoun HC, Carter FA, Frampton CM. Effectiveness of computerised cognitive behaviour therapy for anxiety disorders in secondary care. Aust N Z J Psychiatry 2012 Jul;46(7):630-640. [doi: 10.1177/0004867412437345] [Medline: 22327097]
95. Berger T, Hohl E, Caspar F. Internet-based treatment for social phobia: a randomized controlled trial. J Clin Psychol 2009 Oct;65(10):1021-1035. [doi: 10.1002/jclp.20603] [Medline: 19437505]
96. Berger T, Caspar F, Richardson R, Kneubühler B, Sutter D, Andersson G. Internet-based treatment of social phobia: a randomized controlled trial comparing unguided with two types of guided self-help. Behav Res Ther 2011 Mar;49(3):158-169. [doi: 10.1016/j.brat.2010.12.007] [Medline: 21255767]
97. Berger T, Boettcher J, Caspar F. Internet-based guided self-help for several anxiety disorders: a randomized controlled trial comparing a tailored with a standardized disorder-specific approach. Psychotherapy (Chic) 2014 Jun;51(2):207-219. [doi: 10.1037/a0032527] [Medline: 24041199]
98. Boettcher J, Leek L, Matson L, Holmes EA, Browning M, MacLeod C, et al. Internet-based attention bias modification for social anxiety: a randomised controlled comparison of training towards negative and training towards positive cues. PLoS One 2013;8(9):e71760 [FREE Full text] [doi: 10.1371/journal.pone.0071760] [Medline: 24098630]
99. Boettcher J, Aström V, Påhlsson D, Schenström O, Andersson G, Carlbring P. Internet-based mindfulness treatment for anxiety disorders: a randomized controlled trial. Behav Ther 2014 Mar;45(2):241-253 [FREE Full text] [doi: 10.1016/j.beth.2013.11.003] [Medline: 24491199]
100. Boettcher J, Hasselrot J, Sund E, Andersson G, Carlbring P. Combining attention training with internet-based cognitive-behavioural self-help for social anxiety: a randomised controlled trial. Cogn Behav Ther 2014;43(1):34-48 [FREE Full text] [doi: 10.1080/16506073.2013.809141] [Medline: 23898817]
101. Carlbring P, Westling BE, Ljungstrand P, Ekselius L, Andersson G. Treatment of panic disorder via the internet: a randomized trial of a self-help program. Behavior Therapy 2001;32(4):751-764. [doi: 10.1016/S0005-7894(01)80019-8]
102. Carlbring P, Ekselius L, Andersson G. Treatment of panic disorder via the Internet: a randomized trial of CBT vs. applied relaxation. J Behav Ther Exp Psychiatry 2003 Jun;34(2):129-140. [doi: 10.1016/s0005-7916(03)00026-0] [Medline: 12899896]
103. Carlbring P, Nilsson-Ihrfelt E, Waara J, Kollenstam C, Buhrman M, Kaldo V, et al. Treatment of panic disorder: live therapy vs. self-help via the internet. Behav Res Ther 2005 Oct;43(10):1321-1333. [doi: 10.1016/j.brat.2004.10.002] [Medline: 16086983]
104. Carlbring P, Bohman S, Brunt S, Buhrman M, Westling BE, Ekselius L, et al. Remote treatment of panic disorder: a randomized trial of internet-based cognitive behavior therapy supplemented with telephone calls. Am J Psychiatry 2006 Dec;163(12):2119-2125. [doi: 10.1176/ajp.2006.163.12.2119] [Medline: 17151163]
105. Carter FA, Bell CJ, Colhoun HC. Suitability and acceptability of computerised cognitive behaviour therapy for anxiety disorders in secondary care. Aust N Z J Psychiatry 2013 Feb;47(2):142-152. [doi: 10.1177/0004867412461384] [Medline: 23047956]
106. Dear BF, Staples LG, Terides MD, Karin E, Zou J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2015 Dec;36:63-77 [FREE Full text] [doi: 10.1016/j.janxdis.2015.09.003] [Medline: 26460536]
107. Dear BF, Staples LG, Terides MD, Fogliati VJ, Sheehan J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for Social Anxiety Disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2016 Aug;42:30-44 [FREE Full text] [doi: 10.1016/j.janxdis.2016.05.004] [Medline: 27261562]

XSL•FO
RenderX

108. Ebenfeld L, Lehr D, Ebert DD, Kleine Stegemann S, Riper H, Funk B, et al. Evaluating a hybrid web-based training program for panic disorder and agoraphobia: randomized controlled trial. J Med Internet Res 2021 Mar 04;23(3):e20829 [FREE Full text] [doi: 10.2196/20829] [Medline: 33661121]
109. Fogliati VJ, Dear BF, Staples LG, Terides MD, Sheehan J, Johnston L, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided internet-delivered treatment for panic disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2016 Apr;39:88-102 [FREE Full text] [doi: 10.1016/j.janxdis.2016.03.005] [Medline: 27003376]
110. Furmark T, Carlbring P, Hedman E, Sonnenstein A, Clevberger P, Bohman B, et al. Guided and unguided self-help for social anxiety disorder: randomised controlled trial. Br J Psychiatry 2009 Nov;195(5):440-447. [doi: 10.1192/bjp.bp.108.060996] [Medline: 19880935]
111. Gallego MJ, Emmelkamp PMG, van der Kooij M, Mees H. The effects of a Dutch version of an Internet- based treatment program for fear of public speaking: a controlled study. Int J Clin Health Psychol 2011;11(3):459-472.
112. Hedman E, Andersson G, Ljótsson B, Andersson E, Rück C, Mörtberg E, et al. Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: a randomized controlled non-inferiority trial. PLoS One 2011 Mar 25;6(3):e18001 [FREE Full text] [doi: 10.1371/journal.pone.0018001] [Medline: 21483704]
113. Johnston L, Titov N, Andrews G, Spence J, Dear BF. A RCT of a transdiagnostic internet-delivered treatment for three anxiety disorders: examination of support roles and disorder-specific outcomes. PLoS One 2011;6(11):e28079 [FREE Full text] [doi: 10.1371/journal.pone.0028079] [Medline: 22132216]
114. Johnston L. Development and evaluation of a transdiagnostic internet-delivered cognitive behavioural therapy program for three anxiety disorders [Thesis]. UNSW Sydney. 2012. URL: http://hdl.handle.net/1959.4/52139 [accessed 2024-10-14]
115. Jones SL, Hadjistavropoulos HD, Soucy JN. A randomized controlled trial of guided internet-delivered cognitive behaviour therapy for older adults with generalized anxiety. J Anxiety Disord 2016 Jan;37:1-9. [doi: 10.1016/j.janxdis.2015.10.006] [Medline: 26561733]
116. Kiropoulos LA, Klein B, Austin DW, Gilson K, Pier C, Mitchell J, et al. Is internet-based CBT for panic disorder and agoraphobia as effective as face-to-face CBT? J Anxiety Disord 2008 Dec;22(8):1273-1284. [doi: 10.1016/j.janxdis.2008.01.008] [Medline: 18289829]
117. Klein B, Richards JC, Austin DW. Efficacy of internet therapy for panic disorder. J Behav Ther Exp Psychiatry 2006 Sep;37(3):213-238. [doi: 10.1016/j.jbtep.2005.07.001] [Medline: 16126161]
118. Klein B, Austin D, Pier C, Kiropoulos L, Shandley K, Mitchell J, et al. Internet-based treatment for panic disorder: does frequency of therapist contact make a difference? Cogn Behav Ther 2009;38(2):100-113. [doi: 10.1080/16506070802561132] [Medline: 19306149]
119. Kok RN, van Straten A, Beekman ATF, Cuijpers P. Short-term effectiveness of web-based guided self-help for phobic outpatients: randomized controlled trial. J Med Internet Res 2014 Sep 29;16(9):e226 [FREE Full text] [doi: 10.2196/jmir.3429] [Medline: 25266929]
120. Pham Q, Khatib Y, Stansfeld S, Fox S, Green T. Feasibility and efficacy of an mHealth game for managing anxiety: "Flowy" randomized controlled pilot trial and design evaluation. Games Health J 2016 Feb;5(1):50-67. [doi: 10.1089/g4h.2015.0033] [Medline: 26536488]
121. Robinson E, Titov N, Andrews G, McIntyre K, Schwencke G, Solley K. Internet treatment for generalized anxiety disorder: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010 Jun 03;5(6):e10942 [FREE Full text] [doi: 10.1371/journal.pone.0010942] [Medline: 20532167]
122. Ruwaard J, Broeksteeg J, Schrieken B, Emmelkamp P, Lange A. Web-based therapist-assisted cognitive behavioral treatment of panic symptoms: a randomized controlled trial with a three-year follow-up. J Anxiety Disord 2010 May;24(4):387-396. [doi: 10.1016/j.janxdis.2010.01.010] [Medline: 20227241]
123. Schneider AJ, Mataix-Cols D, Marks IM, Bachofen M. Internet-guided self-help with or without exposure therapy for phobic and panic disorders. Psychother Psychosom 2005;74(3):154-164. [doi: 10.1159/000084000] [Medline: 15832066]
124. Schröder J, Jelinek L, Moritz S. A randomized controlled trial of a transdiagnostic Internet intervention for individuals with panic and phobias - one size fits all. J Behav Ther Exp Psychiatry 2017 Mar;54:17-24. [doi: 10.1016/j.jbtep.2016.05.002] [Medline: 27227651]
125. Stech EP, Chen AZ, Sharrock MJ, Grierson AB, Upton EL, Mahoney AEJ, et al. Internet-delivered exposure therapy versus internet-delivered cognitive behavioral therapy for panic disorder: A pilot randomized controlled trial. J Anxiety Disord 2021 Apr;79:102382. [doi: 10.1016/j.janxdis.2021.102382] [Medline: 33774558]
126. Titov N, Andrews G, Schwencke G, Drobny J, Einstein D. Shyness 1: distance treatment of social phobia over the Internet. Aust N Z J Psychiatry 2008 Jul;42(7):585-594. [doi: 10.1080/00048670802119762] [Medline: 18612862]
127. Titov N, Andrews G, Schwencke G. Shyness 2: treating social phobia online: replication and extension. Aust N Z J Psychiatry 2008 Jul;42(7):595-605. [doi: 10.1080/00048670802119820] [Medline: 18612863]
128. Titov N, Andrews G, Choi I, Schwencke G, Mahoney A. Shyness 3: randomized controlled trial of guided versus unguided Internet-based CBT for social phobia. Aust N Z J Psychiatry 2008 Dec;42(12):1030-1040. [doi: 10.1080/00048670802512107] [Medline: 19016091]

XSL•FO
RenderX

129. Titov N, Andrews G, Johnston L, Schwencke G, Choi I. Shyness programme: longer term benefits, cost-effectiveness, and acceptability. Aust N Z J Psychiatry 2009 Jan;43(1):36-44. [doi: 10.1080/00048670802534424] [Medline: 19085526]
130. Titov N, Andrews G, Choi I, Schwencke G, Johnston L. Randomized controlled trial of web-based treatment of social phobia without clinician guidance. Aust N Z J Psychiatry 2009 Jan 01;43(10):913-919. [doi: 10.1080/00048670903179160]
131. Titov N, Andrews G, Schwencke G, Solley K, Johnston L, Robinson E. An RCT comparing effect of two types of support on severity of symptoms for people completing internet-based cognitive behaviour therapy for social phobia. Aust N Z J Psychiatry 2009 Jan 01;43(10):920-926. [doi: 10.1080/00048670903179228]
132. Titov N, Andrews G, Robinson E, Schwencke G, Johnston L, Solley K, et al. Clinician-assisted internet-based treatment is effective for generalized anxiety disorder: randomized controlled trial. Aust N Z J Psychiatry 2009 Jan 01;43(10):905-912. [doi: 10.1080/00048670903179269]
133. Titov N, Andrews G, Johnston L, Robinson E, Spence J. Transdiagnostic Internet treatment for anxiety disorders: a randomized controlled trial. Behav Res Ther 2010 Sep;48(9):890-899. [doi: 10.1016/j.brat.2010.05.014] [Medline: 20561606]
134. Titov N, Andrews G, Schwencke G, Robinson E, Peters L, Spence J. Randomized controlled trial of Internet cognitive behavioural treatment for social phobia with and without motivational enhancement strategies. Aust N Z J Psychiatry 2010 Oct;44(10):938-945. [doi: 10.3109/00048674.2010.493859] [Medline: 20932208]
135. Tulbure BT, Szentagotai A, David O, tefan S, Månsson KNT, David D, et al. Internet-delivered cognitive-behavioral therapy for social anxiety disorder in Romania: a randomized controlled trial. PLoS One 2015;10(5):e0123997 [FREE Full text] [doi: 10.1371/journal.pone.0123997] [Medline: 25938241]
136. Aydos L, Titov N, Andrews G. Shyness 5: the clinical effectiveness of Internet-based clinician-assisted treatment of social phobia. Australas Psychiatry 2009 Dec;17(6):488-492. [doi: 10.1080/10398560903284943] [Medline: 20001373]
137. Botella C, Gallego M, Garcia-Palacios A, Baños R, Quero S, Alcañiz M. The acceptability of an Internet-based self-help treatment for fear of public speaking. British Journal of Guidance & Counselling 2009 Aug;37(3):297-311. [doi: 10.1080/03069880902957023]
138. Botella C, Guillen V, Banos RM, García-Palacios A, Gallego MJ, Alcaniz M. Telepsychology and self-help: the treatment of fear of public speaking. Cognitive and Behavioral Practice 2007 Feb;14(1):46-57. [doi: 10.1016/j.cbpra.2006.01.007]
139. Botella C, Hofmann SG, Moscovitch DA. A self-applied, internet-based intervention for fear of public speaking. J Clin Psychol 2004 Aug;60(8):821-830. [doi: 10.1002/jclp.20040] [Medline: 15241810]
140. Carlbring P, Furmark T, Steczkó J, Ekselius L, Andersson G. An open study of Internet-based bibliotherapy with minimal therapist contact via email for social phobia. Clinical Psychologist 2007 Feb 03;10(1):30-38. [doi: 10.1080/13284200500378662]
141. Draper M, Rees CS, Nathan PR. Internet-based self-management of generalised anxiety disorder: a preliminary study. Behav. change 2012 Feb 22;25(4):229-244. [doi: 10.1375/bech.25.4.229]
142. Ebenfeld L, Kleine Stegemann S, Lehr D, Ebert DD, Funk B, Riper H, et al. A mobile application for panic disorder and agoraphobia: insights from a multi-methods feasibility study. Internet Interv 2020 Mar;19:100296 [FREE Full text] [doi: 10.1016/j.invent.2019.100296] [Medline: 31890640]
143. Gruber K, Moran PJ, Roth WT, Taylor CB. Computer-assisted cognitive behavioral group therapy for social phobia. Behavior Therapy 2001;32(1):155-165. [doi: 10.1016/S0005-7894(01)80050-2]
144. Kenardy JA, Dow MGT, Johnston DW, Newman MG, Thomson A, Taylor CB. A comparison of delivery methods of cognitive-behavioral therapy for panic disorder: an international multicenter trial. J Consult Clin Psychol 2003 Dec;71(6):1068-1075. [doi: 10.1037/0022-006X.71.6.1068] [Medline: 14622082]
145. Kenwright M, Marks IM, Gega L, Mataix-Cols D. Computer-aided self-help for phobia/panic via internet at home: a pilot study. Br J Psychiatry 2004 May;184:448-449. [doi: 10.1192/bjp.184.5.448] [Medline: 15123511]
146. Klein B, Shandley K, Austin D, Nordin S. A pilot trial of 'Panic Online' as a self-guided treatment for panic disorder. EJ Appl Psychol 2008;4(2):25-30. [doi: 10.7790/ejap.v4i2.136]
147. Krafft J, Ong CW, Davis CH, Petersen JM, Levin ME, Twohig MP. An open trial of group acceptance and commitment therapy with an adjunctive mobile app for generalized anxiety disorder. Cognitive and Behavioral Practice 2022 Nov;29(4):846-859. [doi: 10.1016/j.cbpra.2021.05.008]
148. Lim MH, Rodebaugh TL, Eres R, Long KM, Penn DL, Gleeson JFM. A pilot digital intervention targeting loneliness in youth mental health. Front Psychiatry 2019;10:604 [FREE Full text] [doi: 10.3389/fpsyt.2019.00604] [Medline: 31507469]
149. Loo Gee B, Batterham PJ, Gulliver A, Reynolds J, Griffiths KM. An ecological momentary intervention for people with social anxiety: a descriptive case study. Inform Health Soc Care 2021 Dec 02;46(4):370-398. [doi: 10.1080/17538157.2021.1896525] [Medline: 33779480]
150. Miller CB, Gu J, Henry AL, Davis ML, Espie CA, Stott R, et al. Feasibility and efficacy of a digital CBT intervention for symptoms of generalized anxiety disorder: a randomized multiple-baseline study. J Behav Ther Exp Psychiatry 2021 Mar;70:101609. [doi: 10.1016/j.jbtep.2020.101609] [Medline: 32950939]
151. Miralles I, Granell C, García-Palacios A, Castilla D, González-Pérez A, Casteleyn S, et al. Enhancing in vivo exposure in the treatment of panic disorder and agoraphobia using location-based technologies: a case study. Clinical Case Studies 2019 Dec 20;19(2):145-159. [doi: 10.1177/1534650119892900]

XSL•FO
RenderX

152. Newman MG, Kenardy J, Herman S, Taylor CB. Comparison of palmtop-computer-assisted brief cognitive-behavioral treatment to cognitive-behavioral treatment for panic disorder. J Consult Clin Psychol 1997 Feb;65(1):178-183. [doi: 10.1037//0022-006x.65.1.178] [Medline: 9103747]
153. Oser M, Wallace ML, Solano F, Szigethy EM. Guided digital cognitive behavioral program for anxiety in primary care: propensity-matched controlled trial. JMIR Ment Health 2019 Apr 04;6(4):e11981 [FREE Full text] [doi: 10.2196/11981] [Medline: 30946022]
154. Pier C, Austin DW, Klein B, Mitchell J, Schattner P, Ciechomski L, et al. A controlled trial of internet-based cognitive-behavioural therapy for panic disorder with face-to-face support from a general practitioner or email support from a psychologist. Ment Health Fam Med 2008 Mar;5(1):29-39 [FREE Full text] [Medline: 22477844]
155. Shandley K, Austin DW, Klein B, Pier C, Schattner P, Pierce D, et al. Therapist-assisted, Internet-based treatment for panic disorder: can general practitioners achieve comparable patient outcomes to psychologists? J Med Internet Res 2008 May 19;10(2):e14 [FREE Full text] [doi: 10.2196/jmir.1033] [Medline: 18487138]
156. Tolin DF, McGrath PB, Hale LR, Weiner DN, Gueorguieva R. A multisite benchmarking trial of capnometry guided respiratory intervention for panic disorder in naturalistic treatment settings. Appl Psychophysiol Biofeedback 2017 Mar;42(1):51-58 [FREE Full text] [doi: 10.1007/s10484-017-9354-4] [Medline: 28194546]
157. Williams AD, O'Moore K, Mason E, Andrews G. The effectiveness of internet cognitive behaviour therapy (iCBT) for social anxiety disorder across two routine practice pathways. Internet Interventions 2014 Oct;1(4):225-229. [doi: 10.1016/j.invent.2014.11.001]
158. Batterham PJ, Calear AL, Sunderland M, Kay-Lambkin F, Farrer LM, Christensen H, et al. A brief intervention to increase uptake and adherence of an internet-based program for depression and anxiety (enhancing engagement with psychosocial interventions): randomized controlled trial. J Med Internet Res 2021 Jul 27;23(7):e23029 [FREE Full text] [doi: 10.2196/23029] [Medline: 34313595]
159. Kleiboer A, Donker T, Seekles W, van Straten A, Riper H, Cuijpers P. A randomized controlled trial on the role of support in Internet-based problem solving therapy for depression and anxiety. Behaviour Research and Therapy 2015 Sep;72:63-71. [doi: 10.1016/j.brat.2015.06.013] [Medline: 26188373]
160. Newby JM, Mackenzie A, Williams AD, McIntyre K, Watts S, Wong N, et al. Internet cognitive behavioural therapy for mixed anxiety and depression: a randomized controlled trial and evidence of effectiveness in primary care. Psychol Med 2013 Dec;43(12):2635-2648. [doi: 10.1017/S0033291713000111] [Medline: 23419552]
161. Proudfoot J, Clarke J, Birch M, Whitton AE, Parker G, Manicavasagar V, et al. Impact of a mobile phone and web program on symptom and functional outcomes for people with mild-to-moderate depression, anxiety and stress: a randomised controlled trial. BMC Psychiatry 2013 Nov 18;13:312 [FREE Full text] [doi: 10.1186/1471-244X-13-312] [Medline: 24237617]
162. Titov N, Dear BF, Johnston L, Lorian C, Zou J, Wootton B, et al. Improving adherence and clinical outcomes in self-guided internet treatment for anxiety and depression: randomised controlled trial. PLoS One 2013;8(7):e62873 [FREE Full text] [doi: 10.1371/journal.pone.0062873] [Medline: 23843932]
163. Titov N, Dear BF, Schwencke G, Andrews G, Johnston L, Craske MG, et al. Transdiagnostic internet treatment for anxiety and depression: a randomised controlled trial. Behav Res Ther 2011 Aug;49(8):441-452. [doi: 10.1016/j.brat.2011.03.007] [Medline: 21679925]
164. Adam A, Jain A, Pletnikova A, Bagga R, Vita A, N Richey L, et al. Use of a mobile app to augment psychotherapy in a community psychiatric clinic: feasibility and fidelity trial. JMIR Form Res 2020 Jul 03;4(7):e17722 [FREE Full text] [doi: 10.2196/17722] [Medline: 32618572]
165. Johnston L, Dear BF, Gandy M, Fogliati VJ, Kayrouz R, Sheehan J, et al. Exploring the efficacy and acceptability of internet-delivered cognitive behavioural therapy for young adults with anxiety and depression: an open trial. Aust N Z J Psychiatry 2014 Sep;48(9):819-827. [doi: 10.1177/0004867414527524] [Medline: 24622977]
166. Kladnitski N, Smith J, Allen A, Andrews G, Newby JM. Online mindfulness-enhanced cognitive behavioural therapy for anxiety and depression: outcomes of a pilot trial. Internet Interv 2018 Sep;13:41-50 [FREE Full text] [doi: 10.1016/j.invent.2018.06.003] [Medline: 30206518]
167. Staples LG, Fogliati VJ, Dear BF, Nielssen O, Titov N. Internet-delivered treatment for older adults with anxiety and depression: implementation of the Wellbeing Plus Course in routine clinical care and comparison with research trial outcomes. BJPsych Open 2016 Sep;2(5):307-313 [FREE Full text] [doi: 10.1192/bjpo.bp.116.003400] [Medline: 27703794]
168. Wright JH, Wright AS, Salmon P, Beck AT, Kuykendall J, Goldsmith LJ, et al. Development and initial testing of a multimedia program for computer-assisted cognitive therapy. Am J Psychother 2002;56(1):76-86. [doi: 10.1176/appi.psychotherapy.2002.56.1.76] [Medline: 11977785]
169. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
170. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727 [FREE Full text] [doi: 10.2196/43727] [Medline: 37566447]

XSL•FO
RenderX

171. Lee P, Abernethy A, Shaywitz D, Gundlapalli AV, Weinstein J, Doraiswamy PM, et al. Digital health COVID-19 impact assessment: lessons learned and compelling needs. NAM Perspect 2022;2022 [FREE Full text] [doi: 10.31478/202201c] [Medline: 35402858]
172. Adu-Brimpong J, Pugh J, Darko DA, Shieh L. Examining diversity in digital therapeutics clinical trials: descriptive analysis. J Med Internet Res 2023 Aug 02;25:e37447 [FREE Full text] [doi: 10.2196/37447] [Medline: 37531157]

Abbreviations

CBT: cognitive behavioral therapy
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial

Edited by T de Azevedo Cardoso; submitted 09.09.23; peer-reviewed by A Parks, M Boeker; comments to author 06.02.24; revised version received 27.02.24; accepted 30.04.24; published 28.10.24.

Please cite as:
Lau CKY, Saad A, Camara B, Rahman D, Bolea-Alamanac B
Acceptability of Digital Mental Health Interventions for Depression and Anxiety: Systematic Review
J Med Internet Res 2024;26:e52609
URL: https://www.jmir.org/2024/1/e52609
doi:10.2196/52609
PMID:39466300

©Carrie K Y Lau, Anthony Saad, Bettina Camara, Dia Rahman, Blanca Bolea-Alamanac. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Implications of Big Data Analytics, AI, Machine Learning, and Deep Learning in the Health Care System of Bangladesh: Scoping Review

Md Ashraful Alam[1], MSc; Md Refat Uz Zaman Sajib[2], BDS; Fariya Rahman[1], MPH; Saraban Ether[1], MPH; Molly Hanson[3], MSc; Abu Sayeed[1], MSc; Ema Akter[1], MS; Nowrin Nusrat[1], MS; Tanjeena Tahrin Islam[1], MS; Sahar Raza[1], MPH; K M Tanvir[1], MS; Mohammod Jobayer Chisti[1], PhD; Qazi Sadeq-ur Rahman[1], MSc; Akm Hossain[1], MSc; MA Layek[4], PhD; Asaduz Zaman[5], MSc; Juwel Rana[6,7], MPH; Syed Moshfiqur Rahman[3], PhD; Shams El Arifeen[1], DrPH; Ahmed Ehsanur Rahman[1], PhD; Anisuddin Ahmed[1,3], MS

[1]Maternal and Child Health Division, International Centre for Diarrheal Disease Research, Bangladesh, Dhaka, Bangladesh
[2]Department of Health and Kinesiology, University of Illinois, Champaign and Urbana, IL, United States
[3]Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
[4]Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh
[5]Faculty of Information Technology, Monash University, Melbourne, Australia
[6]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
[7]Research and Innovation Division, South Asian Institute for Social Transformation, Dhaka, Bangladesh

Corresponding Author:
Anisuddin Ahmed, MS
Department of Women's and Children's Health
Uppsala University
Akademiska sjukhuset 751 85
Uppsala, 751 85
Sweden
Phone: 46 73 041 98 48
Email: anisuddin.ahmed@kbh.uu.se

Abstract

Background: The rapid advancement of digital technologies, particularly in big data analytics (BDA), artificial intelligence (AI), machine learning (ML), and deep learning (DL), is reshaping the global health care system, including in Bangladesh. The increased adoption of these technologies in health care delivery within Bangladesh has sparked their integration into health care and public health research, resulting in a noticeable surge in related studies. However, a critical gap exists, as there is a lack of comprehensive evidence regarding the research landscape; regulatory challenges; use cases; and the application and adoption of BDA, AI, ML, and DL in the health care system of Bangladesh. This gap impedes the attainment of optimal results. As Bangladesh is a leading implementer of digital technologies, bridging this gap is urgent for the effective use of these advancing technologies.

Objective: This scoping review aims to collate (1) the existing research in Bangladesh's health care system, using the aforementioned technologies and synthesizing their findings, and (2) the limitations faced by researchers in integrating the aforementioned technologies into health care research.

Methods: MEDLINE (via PubMed), IEEE Xplore, Scopus, and Embase databases were searched to identify published research articles between January 1, 2000, and September 10, 2023, meeting the following inclusion criteria: (1) any study using any of the BDA, AI, ML, and DL technologies and health care and public health datasets for predicting health issues and forecasting any kind of outbreak; (2) studies primarily focusing on health care and public health issues in Bangladesh; and (3) original research articles published in peer-reviewed journals and conference proceedings written in English.

Results: With the initial search, we identified 1653 studies. Following the inclusion and exclusion criteria and full-text review, 4.66% (77/1653) of the articles were finally included in this review. There was a substantial increase in studies over the last 5 years (2017-2023). Among the 77 studies, the majority (n=65, 84%) used ML models. A smaller proportion of studies incorporated AI (4/77, 5%), DL (7/77, 9%), and BDA (1/77, 1%) technologies. Among the reviewed articles, 52% (40/77) relied on primary

XSL•FO
RenderX

data, while the remaining 48% (37/77) used secondary data. The primary research areas of focus were infectious diseases (15/77, 19%), noncommunicable diseases (23/77, 30%), child health (11/77, 14%), and mental health (9/77, 12%).

Conclusions: This scoping review highlights remarkable progress in leveraging BDA, AI, ML, and DL within Bangladesh's health care system. The observed surge in studies over the last 5 years underscores the increasing significance of AI and related technologies in health care research. Notably, most (65/77, 84%) studies focused on ML models, unveiling opportunities for advancements in predictive modeling. This review encapsulates the current state of technological integration and propels us into a promising era for the future of digital Bangladesh.

(J Med Internet Res 2024;26:e54710) doi:10.2196/54710

KEYWORDS

machine learning; deep learning; artificial intelligence; big data analytics; public health; health care; mobile phone; Bangladesh

Introduction

Background

In recent years, the global surge in digital technology has ushered in a data-driven decision-making era across diverse sectors, including the health care and public health sectors [1]. The convergence of state-of-the-art technologies, such as big data analytics (BDA), artificial intelligence (AI), machine learning (ML), and deep learning (DL), is collectively reshaping the health care and epidemiological landscape [2]. This paradigm shift promises to revolutionize the comprehension and management of health challenges, offering the potential for more efficient and effective interventions [2].

The health care system fundamentally revolves around safeguarding and enhancing individual and community health by integrating public health and medical care services, using health promotion, disease and injury prevention, infectious disease surveillance, response, and research. Interdisciplinary research now explores various factors influencing individual and community well-being, ranging from microbial infections to social determinants, behavioral patterns, and various facets of lifestyle [3]. In addition, the widespread adoption of emerging technologies such as telemedicine, clinical decision support systems, electronic health records, personal health records, and mobile health care have generated a vast reservoir of health-related data [4].

Harnessing the latent potential of these extensive data, digital technologies, such as BDA, AI, ML, and DL, can surpass conventional data management systems [5]. By applying AI algorithms, ML techniques, and DL capabilities, BDA offers invaluable support to clinicians, health care providers, and policy makers in intervention designing, planning, and execution; expediting disease detection; predicting health outcomes; and advancing personalized medicine [6]. This ultimately results in both cost-effectiveness and quality outcomes [7].

On a global scale, extensive health data can be categorized into 5 primary domains: biological measurements of participants (eg, genomic or metabolomic datasets), participant context measurements (eg, environmental variables), administrative medical records, participant tracking data from devices equipped with GPS, and electronic data from various sources (eg, social media or search history records) [8]. Using BDA, AI, ML, and DL algorithms to extract valuable insights from these datasets can significantly enhance individual health management and enable swift disease trend detection and outbreak responses [9]. Beyond health care, BDA, AI, and ML can play crucial roles in enabling molecular data analysis, synthesis of genetic and clinical information to craft personalized treatments, and development of predictive models based on omics data to identify genetic disease markers and develop effective novel drugs [10].

In Bangladesh, which is a densely populated South Asian nation with >160 million inhabitants, integrating modern technologies with the health care system holds immense potential. Bangladesh grapples with various health-related challenges, spanning from infectious diseases, noncommunicable diseases (NCDs) and maternal and child health issues to environmental health concerns [11]. To tackle these multifaceted challenges comprehensively, the government of Bangladesh (GoB) has actively adopted advanced technologies to elevate the quality of health care services. Initiatives such as District Health Information System 2, the open medical record system, and the open-source smart register platform have revolutionized health data collection and service delivery [12]. These automated applications are instrumental in generating specific health-related evidence across different tiers of health care services and facilitating need-based resource allocations. Although there are potential challenges regarding proper data governance and legal framework, the immense potential to use these data has been capturing the attention of interdisciplinary researchers and practitioners eager to explore the potential of BDA, AI, ML, and DL technologies to use this wealth of datasets [12].

Objectives

In Bangladesh, instances of using BDA, AI, ML, and DL technologies in health care and public health research are sporadic. However, the dearth of clear evidence has led to an inadequate understanding of the proper application of these technologies and their potential contributions to health care and public health research [12]. This scoping review is motivated by the need to address this gap, systematically examining existing research that explores the use of BDA, AI, ML, and DL technologies to analyze health care data in Bangladesh. The primary objective was to compile and synthesize available research findings on these domains in the country's health care sector. This includes identifying trends, patterns, and methodologies for applying BDA, AI, ML, and DL technologies. The secondary objective was to assess the limitations faced by researchers in integrating these technologies into health care

XSL•FO
RenderX

research, aiming to provide insights into the challenges that may hinder their widespread adoption. Overall, the insights derived from this review aimed to furnish evidence-based strategies for policy makers, health care professionals, and researchers, ultimately contributing to enhancing health outcomes and the overall well-being of the Bangladeshi population. Moreover, this exploratory journey is guided by the goal of uncovering valuable insights that can significantly shape the future of health care, not only within Bangladesh but also on a global scale.

Methods

Overview

We have followed the standard guidelines for scoping reviews. This scoping review used the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Multimedia Appendix 1) guidelines in line with best practices for research paper selection and reporting [13], and its protocol has been registered with the Open Science Framework [14]. We have provided the PRISMA-ScR checklist.

Ethical Considerations

There were no human participants in this review, and ethics approval is not applicable.

Inclusion and Exclusion Criteria

We included all studies that are primarily focused on health care and public health issues in Bangladesh and used any of the cross-cutting technologies, such as BDA, AI, ML, and DL, in analyzing datasets. The studies that used publicly available global or web-based health care and public health datasets in their research but validated their findings in the context of Bangladesh were also included in this review. All journal articles and conference articles published between January 1, 2000, and September 10, 2023, written in English and publicly accessible, were included regardless of research design and sample sizes.

Studies conducted by Bangladeshi researchers but not focusing on the context of Bangladesh, review articles, book chapters, commentaries, gray literature, and articles published in other languages (other than English) were excluded from the study. Manuscripts that were not accessible through institutional credentials or after contacting the authors were also excluded.

A comparative textbox (Textbox 1) presents the inclusion and exclusion criteria.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Article type: journal articles, conference articles
- Focus: studies primarily focused on public health issues in Bangladesh using big data analytics, artificial intelligence, machine learning, or deep learning
- Data source: studies using publicly available global or web-based public health datasets validated in the context of Bangladesh
- Language: English
- Publication date: between January 1, 2000, and September 10, 2023
- Accessibility: full text accessible through institutional access or direct contact with authors

Exclusion criteria

- Article type: review articles, book chapters, commentaries, gray literature
- Focus: studies not focusing on the context of Bangladesh
- Data source: studies using publicly available global or online public health datasets validated in other countries
- Language: papers published in languages other than English
- Publication date: articles published before January 2000 and after the database search date
- Accessibility: manuscripts not accessible through institutional access or direct contact with authors

Data Sources and Literature Search

We conducted a search across 4 electronic databases: MEDLINE (via PubMed), IEEE Xplore, Scopus, and Embase, on September 10, 2023. Our search was refined by applying a date filter, limiting results from January 1, 2000, to the search date. To identify relevant information, we used a set of key search terms related to the population, concept, and context framework (Textbox 2).

XSL•FO
RenderX

Textbox 2. Key terms of the population, concept, and context framework for developing a comprehensive search strategy.

Population

- Public health
- Community health
- Preventive medicine
- Epidemiology
- Health care management
- Public health and safety
- Population health
- Public Health Surveillance
- Public health informatics

Concept

- Big data analytics
- Artificial intelligence
- Machine learning
- Deep learning

Context

- Bangladesh
- Bangladesh health system

For each database, we meticulously crafted a comprehensive search strategy, tailoring it to the specific databases by incorporating both key terms and database-specific index terms (Multimedia Appendix 2).

The main search keywords included terms such as "Artificial Intelligence," "AI," "Machine Intelligence," and "Cognitive Computing" for AI; "Machine Learning," "Transfer Learning," "Predictive Modeling," and "Pattern Recognition" for ML; "Deep Learning," "Neural Network," "Artificial Neural Network," and "Deep Neural Network" for DL; and "Big Data," "Data Mining," "Big Data Analytics," and "Data Science" for BDA. In addition, we included keywords relevant to the health care context, such as "Healthcare System," "Health Services," "Clinical Settings," and "Public Health," and geographic context-specific terms, such as "Bangladesh," "Bangladeshi," and "Bangladesh Health System."

Using these targeted keywords, we aimed to capture a comprehensive set of studies relevant to the use of BDA, AI, ML, and DL technologies within the health care system of Bangladesh.

Screening Process

Following the retrieval of articles from the databases, they were imported into EndNote (version 9; Clarivate), and duplicates were removed. Subsequently, the review team was divided into 2 groups: group 1 and group 2. Each group was assigned a subset of articles from the total pool of retrieved articles (group 1: 826/1653, 49.97%; group 2: 827/1653, 50.03%) to independently screen the titles and abstracts in accordance with the established inclusion and exclusion criteria.

Each reviewer independently screened the titles and abstracts of their assigned articles and cross-checked their decisions within groups. After this initial screening, full-text screening was performed by 2 individual reviewers. Any disagreements or uncertainties that arose during any of the screening stages were promptly resolved through discussion and consensus among the reviewers. In cases of disagreements, a third reviewer stepped in to refute the disagreements, considering inclusion or exclusion criteria. This rigorous screening process ensured that the selected studies aligned with the defined objectives and inclusion or exclusion criteria, enhancing the robustness and reliability of this scoping review.

Data Extraction, Analysis, and Quality Assessment

We developed a data extraction sheet to capture relevant data in line with the review objectives. The extraction variables comprised descriptive information of the included studies, such as title, author, published year, sample size, dataset information, study design, the objective of the study, the technology used, the domain of public health, results, limitations, and practical implementation of the study result. Following data extraction, we analyzed and categorized the data according to the dataset information, the similarity of the disease domains, methods used for analysis, and the performance of algorithms.

The quality of all the included studies was assessed following the Joanna Briggs Institute Critical Appraisal tool [15]. In total, 2 individual reviewers assessed the quality of the studies, and disagreement was solved upon discussion. For overall grading using the Critical Appraisal tool, we followed a published systematic review that used a rule of thumb for overall grading [16]. For case-control studies, if 0 to 3 "yes" answers were

XSL•FO
RenderX

obtained out of 10 screening questions, the study was marked as poor quality. Consecutively, if 4 to 6 “yes” answers were obtained, the study was marked as moderate quality, and if >6 “yes” answers were obtained, the study was ranked as good quality. For cohort studies, among 11 screening questions, 0 to 3 “yes” answers were marked as poor quality, 4 to 7 “yes” answers as moderate quality, and >6 “yes” answers as good quality. For cross-sectional studies, among 8 screening questions, 0 to 3 “yes” answers were marked as poor quality, 4 to 5 “yes” answers as moderate quality, and >6 “yes” answers as good quality study (Multimedia Appendix 3).

Results

Selection of Sources of Evidence

Of the 1653 articles retrieved from the 4 databases, 78 (4.72%) duplicate publications were removed initially. Next, in the title and abstract screening stage, 83.87% (1321/1575) articles were removed as these did not align with the review objectives or were either workshop articles or not original studies (letters to editors and descriptions of keynotes). A total of 43.7% (111/254) of the publications were also excluded due to inadequate information related to research objectives. The remaining 56.3% (143/254) of the studies were taken for full-text review. Moreover, of the 143 studies, 22 (15.4%) were inaccessible and excluded accordingly. Finally, 121 (84.6%) of the 143 articles were reviewed in full text for eligibility. Subsequently, 44 (30.8%) were excluded due to inadequate information considering research objectives, and finally, 77 (53.8%) articles were selected for data extraction. A detailed description of the article selection process is given in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. AI: artificial intelligence; BDA: big data analytics; DL: deep learning; ML: machine learning.

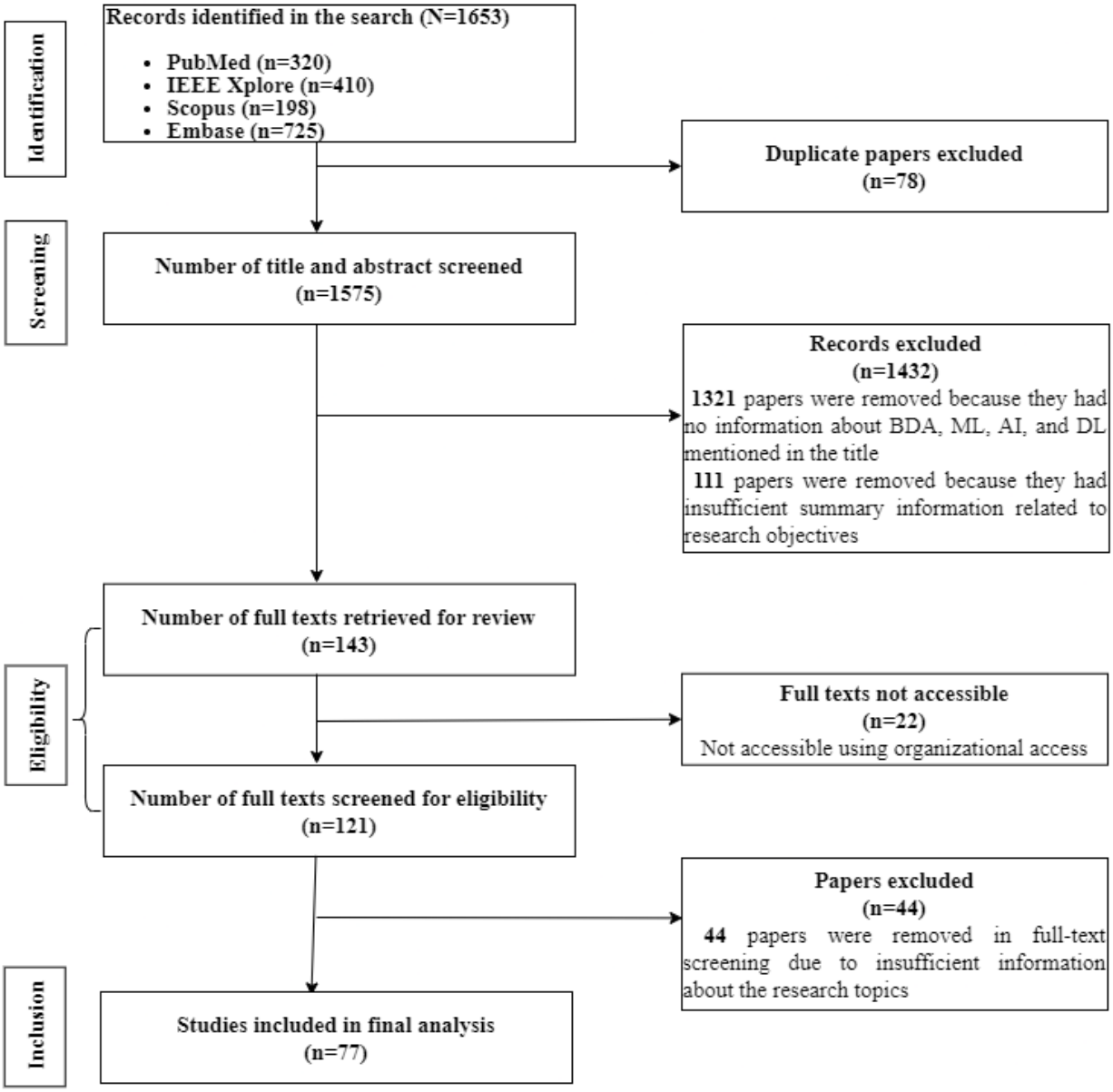

Key Study Characteristics and Thematic Domains

This scoping review identified 77 research articles that examine how BDA, AI, ML, and DL are used in Bangladesh's health care system and public health. The characteristics of the studies were grouped according to the study methodology, data types, and different health domains. Researchers mostly applied ML algorithms in their studies, and in 84% (65/77) of the studies, they used cross-sectional study data. The use of cohort study, case-control study, and retrospective study data was lower among researchers. Table 1 shows the characteristics of the selected studies according to different categories.

XSL•FO
RenderX

Table 1. Characteristics of included studies based on study design, types of data, and health domains (N=77).

Characteristics	Studies, n (%)
Study design	
Cross-sectional study	65 (84)
Cohort study	5 (6)
Case-control study	4 (5)
Retrospective study	3 (4)
Type of data	
Primary	40 (52)
Secondary	37 (48)
Health domain	
Infectious diseases	15 (19)
Noncommunicable diseases	23 (30)
Child health	11 (14)
Mental health	9 (12)
Preventive health	4 (5)
Maternal health	5 (6)
Vector-borne diseases	5 (6)
Others	5 (6)
Technology domain	
Machine learning	65 (84)
Deep learning	7 (9)
Artificial intelligence	4 (5)
Big data analytics	1 (1)

Although the date limiter was set from January 1, 2000, to September 10, 2023, no published articles according to the research objectives were found before 2012. A small number (4/77, 5%) of published research articles were found between 2012 and 2017, focusing on public health issues using BDA, AI, ML, and DL. After 2017, a substantial surge in interest among researchers in using AI, ML, and DL was observed, which mostly focused on NCDs, infectious diseases, and child health. Among the NCD group, predicting type 2 diabetes mellitus, hypertension, and cancer is mostly focused on, and among the infectious diseases, COVID-19 was mostly prevalent throughout the pandemic from 2020 to 2022. In addition, the researchers highlighted using the target technologies for different vector-borne diseases during this timeframe (Figure 2).

XSL•FO
RenderX

Figure 2. Number of articles on different health focuses by year, using big data analytics, artificial intelligence, machine learning, and deep learning. NCD: noncommunicable disease.

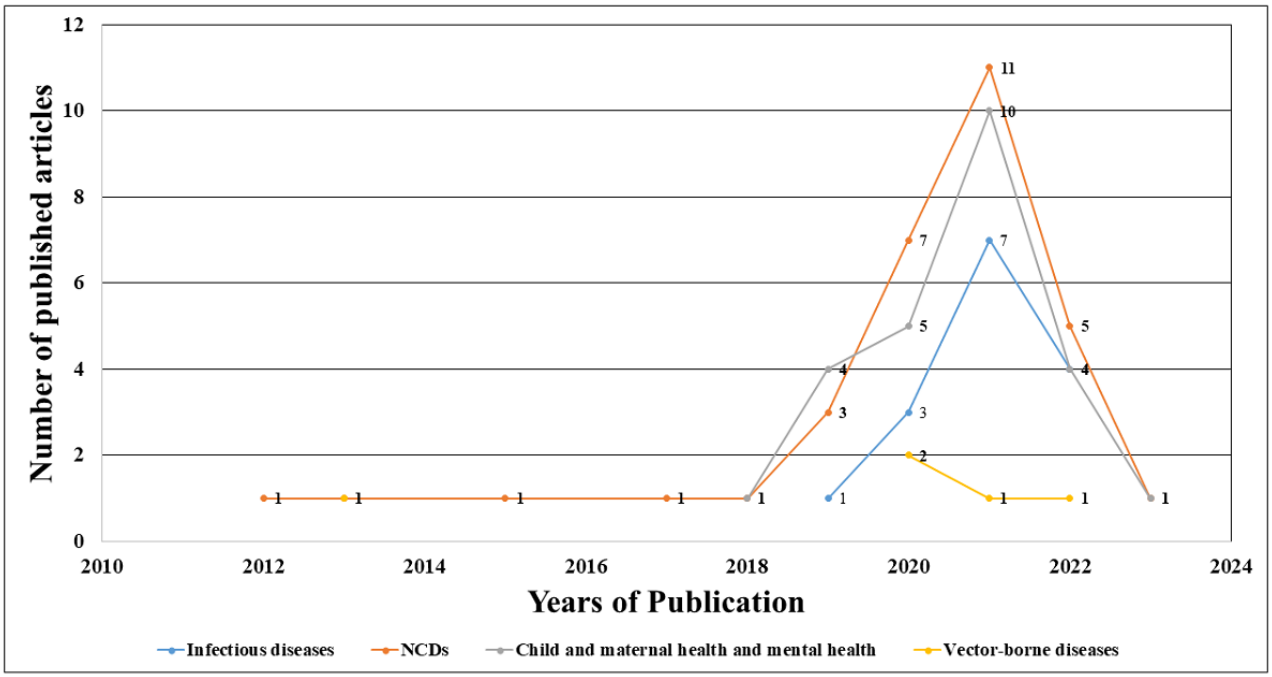

Table 2 summarizes the included articles with respect to their objectives, methods, dataset sizes, and investigated health issues.

XSL•FO
RenderX

Table 2. Summary of the selected studies.

Study	Health focus	Objective of the study	Outcomes
Rahman et al [17], 2021	Infectious diseases (COVID-19)	To categorize the patients with COVID-19 into 3 risk groups (low, moderate, and high) for predicting mortality risk among these patients	LR[a] outperforms other models and its overall accuracy, weighted precision, sensitivity, specificity, and F_1-score are 88%, 88%, 87%, and 90%, respectively.
Satu et al [18], 2021	Infectious diseases (COVID-19)	To develop a short-term forecasting model that predicts the severity of COVID-19	FPM[b] showed better RMSE[c], MAE[d], and R^2 values for predicting COVID-19 cases.
Rafi [19], 2020	Infectious diseases (COVID-19)	To identify COVID-19 cases using CXR[e] and classify them as patients who are normal or patients with COVID-19	Ensemble deep transfer learning model to identify COVID-19 cases with respect to normal cases, with an accuracy of 98.43%.
Sarkar et al [20], 2020	Infectious diseases (COVID-19)	To find the best algorithm that will predict COVID-19 cases, death cases, and recovery cases and the spreading pattern	FPM has a better RMSE value (33,785.55) than other models.
Mohammad Masum et al [21], 2020	Infectious diseases (COVID-19)	To forecast confirmed COVID-19 cases, deaths, and recovery	The LSTM[f] model has an accurate RMSE value than that of SVR[g] and RF[h] on train and test validation sets.
Hassan et al [22], 2021	Infectious diseases (COVID-19)	To create a distributed AI[i] on the edge that is able to deal with big data and to predict the course of the COVID-19 pandemic accurately.	DNN[j] achieved a higher R^2 value of 0.9501 and a better MAPE[k] value of 13.789 in predicting the course of the outbreak.
Chowdhury et al [23], 2021	Infectious diseases (COVID-19)	To predict COVID-19 daily new cases with higher accuracy	The LSTM model performs better than the ANFIS[l] model in predicting COVID-19 cases, with MAPE of 4.51, RMSE of 6.55, and correlation coefficient of 0.75.
Ullah et al [24], 2022	Infectious diseases (COVID-19)	To create AI-based diagnostic tool to detect COVID-19 accurately by examining CXR	At the testing stage, the designed model could interpret CXRs with a precision of 0.98, recall and sensitivity of 0.97, and F_1-score of 67 0.97 for COVID-19. At the validation stage, it showed high sensitivity (90%) and specificity (92%) in detecting COVID-19. The AUC[m] values for COVID-19 and pneumonia were 0.91 and 0.87, respectively.
Leon et al [25], 2021	Infectious diseases (COVID-19)	To predict the number of people infected with COVID-19 and the number of deaths for the upcoming month due to the COVID-19 pandemic.	FPM performs well in predicting the current affected cases with an RMSE value of 318.538, and the ARIMA[n] model performs well in predicting death with an RMSE value of 10.708.
Rahman et al [26], 2019	Infectious diseases (COVID-19)	To predict COVID-19 cases in Bangladesh (daily newly infected cases, daily new fatality, and daily new recovered patients in the next 30 days range)	For daily infected cases forecasting, multiple linear regression and ridge regression perform best and obtain R^2, MSE[o], MAE, and RMSE values of 0.99, 0.004, 0.05, and 0.06, respectively. For daily new fatality cases forecasting, multiple linear regression, ridge regression, and Lasso regression performed best and obtained R^2, MSE, MAE, and RMSE values of 0.99, 0.003, 0.05, and 0.06, respectively. For daily new recovery forecasting, multiple linear regression performed best and obtained R^2, MSE, MAE, and RMSE values of 0.94, 0.05, 0.16, and 0.23, respectively. Overall, multiple linear regression outperforms all other models.
Haq et al [27], 2022	Infectious diseases (COVID-19)	To predict the future direction and spreading pattern of COVID-19	FPM is appropriate for forecasting the COVID-19 pandemic trend in Bangladesh.
Absar et al [28], 2020	Infectious diseases (COVID-19)	To predict the progression of COVID-19 for a period of more than a year under various scenarios in Bangladesh	LSTM showed better RMSE, MAE, and R^2 values for predicting COVID-19 cases.

XSL•FO
RenderX

Study	Health focus	Objective of the study	Outcomes
Karmokar et al [29], 2022	Infectious diseases (COVID-19)	To analyze the effects of meteorological parameters on COVID-19	Cloudy weather has a positive association with COVID-19, temperature has a positive association with number of daily deaths and those who recovered, wind speed has a negative association with COVID-19, air quality has a negative association with daily deaths and recovered, air pressure has a weak negative relation with daily cases and deaths, and rain has a negative relation with daily cases.
Qin et al [30], 2021	Infectious diseases (tuberculosis)	To detect tuberculosis using AI algorithm on CXR images for triaging tuberculosis	Model could identify *Vibrio cholerae* O1 incidence with AUC 93.4%; 95% CI 92.1%-94.7%.
Azman et al [31], 2019	Infectious diseases (cholera)	To identify individuals infected with *V cholerae* O1	All the AI algorithms can be highly accurate and useful triage tools for tuberculosis detection in high-burden regions and outperform human readers with an accuracy of 90.81% (95% CI 90.33-91.29) for qXR, 90.34% (95% CI 89.81-90.87) for CAD4TB, 88.61% (95% CI 88.03-89.20) for Lunit INSIGHT CXR, 84.90% (95% CI 84.27-85.54) for InferRead DR, and 84.89% (95% CI 84.26-85.53) for JF CXR-1. Only qXR (specificity 74.3%; 95% CI 73.3-74.9) and CAD4TB (specificity 72.9%; 95% CI 72.3-73.5) met the TPP at 90% sensitivity.
Pranto et al [32], 2020	NCDs[p] (diabetes)	To predict diabetes among female patients in Bangladesh	RF models showed better performance in predicting the patients with diabetes with an accuracy, F_1-score, and AUC of 78%, 0.84, and 0.83, respectively, in the test set, and KNN[q] showed better accuracy and F_1-score (81.2% and 0.88) and naïve Bayes in AUC (0.84) in validation.
Nishat et al [33], 2021	NCDs (diabetes)	To predict diabetes mellitus	GP[r] emerged as the best-performing algorithm, which is proposed as the most efficient classifier with accuracy, F_1-score, and AUROC[s] of 98.25%, 0.972, and 0.979, respectively.
Islam et al [34], 2020	NCDs (diabetes)	To automatically detect diabetes using ML[t]-based classifiers	Area of living, electricity, wealth index, age, education, working status, smoking status, medicine intake, weight, and BMI are significantly associated with diabetes at a 5% significance level. Furthermore, bagged CART[u] provides the highest accuracy and AUC of 94.3% and 0.600.
Jahan et al [35], 2020	NCDs (diabetes)	To identify patients with a high risk of diabetes using ML	IBK[v] performs well in identifying patients with a high risk of diabetes with 98.73% accuracy.
Emon et al [36], 2021	NCDs (diabetes)	To predict diabetes at an early stage by identifying symptoms and disease related to diabetes	RF performs well in identifying patients with a high risk of diabetes with an accuracy of 98%.
Dutta et al [37], 2022	NCDs (diabetes)	To prepare a diabetes disease classifier dataset so that patients with high risk of diabetes could be identified early	Ensemble-based model (DT[w], RF, XGB[x], and LGBM[y]) proclaims an accuracy of 73.5% and an AUC of 83.2% in predicting diabetes.
Wood [38], 2022	NCDs (diabetes)	To identify the signs and symptoms of type 2 diabetes and predicted type 2 diabetes using ML and statistical method	Support vector classifier predicts early - onset type 2 diabetes with 2.1% errors.
Asaduzzaman et al [39], 2021	NCDs (cancer)	To identify factors that have a greater influence on the severity of cervical cancer and ovarian cancer	AdaBoost performed the best with a classification accuracy of 78.8% in orange and 79% in Sklearn for neurodegenerative disease. In the case of cervical cancer, LR provides the best score of 84.8%, and with Sklearn, the score was 79.3%. On the other hand, SVM[z] shows the best accuracy of 88.3% in orange, and the DT provides 98.6% classification accuracy in Sklearn for ovarian cancer.

XSL•FO
RenderX

Study	Health focus	Objective of the study	Outcomes
Maliha et al [40], 2019	NCDs (cancer)	To predict cancer disease and identified the factors associated with cancer disease	The KNN algorithm achieved a higher accuracy in predicting cancer disease with an accuracy of 98.8%.
Rejaul Islam Royel et al [41], 2021	NCDs (cancer)	To design a tool for early detection of stomach cancer risk level	The study identified 18 significant top risk factors of stomach cancer.
Ahmed et al [42], 2013	NCDs (cancer)	To predict patients with a risk of lung cancer	AprioriTid and DT algorithms can be used to identify the frequent patterns.
Rashid et al [43], 2022	NCDs (diabetes)	To investigate the use of ML approach for predicting cardiac autonomic neuropathy, diabetic peripheral neuropathy, and diabetic retinopathy using only the patient demographic, clinical, and laboratory profiles	RF performs well using diastolic blood pressure, albumin-creatinine ratio, and gender for CAN[aa] testing (98.67%); microalbuminuria, smoking history, and hemoglobin A1c for DPN[ab] testing (67.78%); and hemoglobin A1c, microalbuminuria, and smoking history for RET[ac] testing (84.38%).
Kayyum et al [44], 2020	NCDs (CVD[ad])	To predict the chances of myocardial infarction occurring so that people can take precautions and take measures to prevent it	Bagging can predict patients who are at a high risk of myocardial infarctions with an accuracy of 93.913%, F_1-score of 0.940, and ROC of 0.974.
Tabassum et al [45], 2019	NCDs (CVD)	To predict the cardiac status of a patient or of a person who is unaware of his or her cardiac condition	Neural network showed 82% accuracy.
Chowdhury et al. [46], 2021	NCDs (CVD)	To predict heart disease by analyzing important features from health assessment	SVM has better accuracy in predicting heart disease with an accuracy of 91%.
Fahim et al [47], 2022	NCDs (CVD)	To detect patients with a high risk of CVD at an early stage	XGB performs well in identifying high-risk patients accurately, with an accuracy of 73.72%. Furthermore, the model showed 81.14% accuracy in the dataset, including smoking and alcohol intake.
Islam et al [48], 2021	NCDs (hypertension)	To characterize the risk factors of hypertension among adults in Bangladesh	The combination of SVMRFE-GB[ae] gives the maximum accuracy (66.98%), recall (97.92%), F_1-score (78.99%), and AUC (0.669) compared to others in predicting the risk of hypertension among adults in Bangladesh.
Arefa et al [49], 2019	NCDs (hypertension)	To develop an ML model for a decision support system to predict early hypertension risk among patients and assessed the performance of different ML algorithms	DT, LR, SVM, KNN, and bagged tree showed an accuracy of 100%. Fine Gaussian SVM showed an accuracy of 94.6%. Coarse Gaussian SVM showed an accuracy of 86%.
Islam et al [50], 2022	NCDs (hypertension)	To predict hypertension and its associated factors and compared ML model's performances	XGB, GBM[af], LR, and LDA[ag] models can predict hypertension with a greater accuracy of 90%; DT achieved a precision value of 91%; XGB, GBM, LR, and LDA achieved a recall value of 100%; and XGB, GBM, LR, and LDA scored F_1-score of 95%.
Asadullah et al [51], 2023	NCDs (hypertension)	To predict patients with a high risk of hypertension	Proposed ensemble model showed a higher accuracy, F_1-score, and AUC of 78.17%, 08751, and 0.8634, respectively, in predicting patients with a high risk of hypertension.
Ifraz et al [52], 2021	NCDs (CKD[ah])	To provide a methodology for predicting CKD status using clinical data	The LR model performs better than other models to predict CKD with an accuracy of 97%.
Ehsan et al [53], 2021	NCDs (Parkinson disease)	To screen Parkinson disease using smartphone-based triaxial accelerometer data and cloud-based automated tool	RF classifier produced the optimum result with 88.33% accuracy, 90% sensitivity, and 85% specificity to effectively discriminate between patients with Parkinson disease and healthy individuals.
Akter Hossain et al [54], 2020	NCDs (leukemia disease)	To detect the WBC[ai] components from whole blood images and counted them to know about the possibility of leukemia	Faster-RCNN[aj] model could identify leukemia-positive properties with 90% accuracy.

XSL•FO
RenderX

Study	Health focus	Objective of the study	Outcomes
Shahriar et al [55], 2019	Child health (malnutrition)	To identify the risk factors that had a strong association with malnutrition among children aged 0 to 59 months , predicting malnourished children	ANN[ak] algorithm can predict risk factors better than other algorithm with an accuracy of 86%, 70%, and 67.3%, respectively, for wasting, underweight, and stunting child respectively.
Methun et al [56], 2021	Child health (morbidity)	To determine the factors that regulated the incidence of preventable outbreaks of disease or symptoms among children aged <5 years in Bangladesh	Logistic classifier identified the factors that regulate the incidence of preventable outbreaks of disease or symptoms among children aged <5 years in Bangladesh with an accuracy, F_1-score, and AUC of 0.7, 0.812, and 0.621, respectively.
Borson et al [57], 2020	Child health (low birth weight)	To construct a predictive model for low birth weight prediction by analyzing health and demographic data related to neonatal health conditions in the context of Bangladesh	LR and SVM gained accuracy, precision, and recall scores of 80.30%, 0.743, and 0.803 and 80.29%, 0.803, and 0.803, respectively, in 10-fold cross validation and 81.66%, 0.817, and 0.817 and 81.67%, 0.817, and 0.817, respectively, in training and testing stage.
Rahman et al [58], 2021	Child health (malnutrition)	To predict malnourished children aged <5 years based on their risk factors	RF algorithm can classify stunted growth, wasted growth and underweight children accurately and obtained the highest accuracy of 88.3% for children with stunted growth, 87.7% for children with wasted growth, and 85.7% for children who are underweight.
Talukder and Ahammed [59], 2020	Child health (malnutrition)	To predict malnutrition status among children aged <5 years in Bangladesh	RF accurately identified malnutrition status among children aged <5 years, with an accuracy of 68.51%, a sensitivity of 94.66%, and a specificity of 69.76%.
Khan et al [60], 2021	Child health (malnutrition)	To identify risk factors of stunting among children aged <5 years in Bangladesh	Gradient boosting algorithm predicted stunted children with an accuracy of 67.47%.
Jehan et al [61], 2020	Child health (PTB[al])	To investigate the ability of transcriptomics and proteomics profiling of plasma and metabolomics analysis of urine to identify early biological measurements associated with PTB	PTBs can be predicted using blood and urine samples collected early in the pregnancy, providing opportunities for interventions and the proposed model could be able to identify PTB with an AUROC of 0.83 (95% CI 0.72-0.91).
Islam Pollob et al [62], 2022	Child health (low birth weight)	To determine the risk factors of low birth weight and predicted babies with low birth weight based on ML algorithms	LR-based classifier provided the most accurate classification of babies with low birth weight and has 87.6% accuracy and 0.59 AUC.
Mehedi Hasan et al [63], 2021	Child health (pneumonia)	To classify and predict pneumonia positive and pneumonia negative cases using ML algorithm	DT model can predict pneumonia with high accuracy.
Aftab et al [64], 2021	Child health (GA[am])	To develop and validate programmatically feasible and accurate approaches to estimate newborn GA in low-resource settings	The ML model (10 neonatal characteristics and LMP[an]) estimated GA with a variability of 15.7 days of early ultrasound dating.
Mansur et al [65], 2021	Child health (malnutrition)	To study interactions among various sociodemographic risk factors of childhood stunting in Bangladesh	LR identified malnourished children with an accuracy of 69.4%.
Khudri et al [66], 2023	Maternal health (malnutrition)	To predict the BMI and the risks of malnutrition outcomes for Bangladeshi women of childbearing age from their economic, health, and demographic features	SVM and KNN are the 2 best-performing methods in BMI prediction, with R^2 value of 23.9% and 23.2%, RMSE of 3.372 and 3.387, and MAE of 2.718 and 2.756, respectively.
Islam et al [67], 2022	Maternal health (malnutrition)	To identify potential risk factors of malnourished women and predicted malnourished women using ML algorithm	RF-based classifier provides 81.4% accuracy and 0.837 AUC for underweight prediction and 82.4% accuracy and 0.853 AUC for overweight or obese prediction.
Choudhury et al [68], 2019	Mental health	To identify depression among university students in an early stage using ML	RF algorithm accurately predicted depression successfully, with accuracy and F_1-score of 75% and 60%, respectively.

XSL•FO
RenderX

Study	Health focus	Objective of the study	Outcomes
Haque et al [69], 2021	Mental health	To understand milestone achievement of ASD[ao] using mCARE mobile app and association of milestone achievement with family or child sociodemographic characteristics	For milestone for brushing teeth, LR, KNN, and ANN achieved 95% accuracy; for toilet use, 84% accuracy was achieved with KNN and ANN; and for urinates in toilet or potty and buttons large buttons, ANN achieved an accuracy of 91% and 76%, respectively. For all the parameters, ANN had a higher accuracy (approximately >80% on average).
Tariq et al [70], 2019	Mental health	To classify autism spectrum disorder, neurotypical disorder, and speech and language condition using ML classifier.	The proposed model achieved an accuracy (AUC) of 76% (SD 3%) and sensitivity of 76% (SD 4%) in identifying atypical children from among those with developmental delays. Additionally, the model demonstrated an accuracy (AUC) of 85% (SD 5%) and a sensitivity of 76% (SD 6%) for identifying children with ASD from those predicted to have other developmental delays.
Imtinan Uddin et al [71], 2020	Mental health	To predict early depression risk among technology employees in Bangladesh as well as found some key factors that contributed to depression	AdaBoost DT achieved approximately 98% accuracy in predicting employees with depression.
Khan et al [72], 2018	Mental health	To examine classification algorithm to predict mental disorder	RF produced a better result with an accuracy of 85.1% in predicting mental disorders in real datasets. Furthermore, in an enhanced dataset with synthesized data, it showed 85.9% accuracy.
Rois et al [73], 2021	Mental health	To detect significant risk factors of perceived stress and predicted the prevalence of stress among Bangladeshi university students using ML algorithm	RF model performed better and authentically predicted stress compared with other ML techniques with an accuracy of 0.8972, precision of 0.9241, sensitivity of 0.9250, specificity of 0.8148, AUROC of 0.8715, and k-fold accuracy of 0.8983.
Ahmed et al [74], 2020	Mental health	To detect depression and anxiety using ML	CNN[ap] algorithm has the highest accuracy of 96% for anxiety and 96.8% for depression. In addition, the analysis showed that among Bangladeshi women aged 18 to 35 years, 7.4% experienced profound levels of anxiety and 15.6% experienced chronic depression.
Khan et al [75], 2018	Mental health	To examine the performance of classification algorithms to predict mental disorder	RF has a better performance than the other algorithms.
Moon et al [76], 2022	Mental health	To predict the mental development condition of rural children during the COVID-19 pandemic	RF achieved a higher accuracy of 92.41% in predicting individuals who are mentally hampered
Hu et al [77], 2017	Decision support system	To transform the prediction model into a field-deployable application to predict drug functions to support clinical decisions	GBDT[aq] models could predict drugs with an accuracy of 96.2%.
Baba et al [78], 2015	Decision support system	To recommend high-cost and low-cost diagnostic tests, recommended drugs recommended for patients, and predicted future health-risk levels by assessing patients' health status data	The approach with multiple classifiers successfully reduced the costs of health checkups, a multitask learning method provided accurate recommendations for specific types of drugs, and an active learning method achieved an efficient assignment of health care workers to the follow-up cares of patients.
Al Iqbal [79], 2012	Decision support system	A clinical decision support system to serve the rural medicine center	ACM-FOCL[ar] can identify hepatitis and breast cancer with 90.17% and 84.97 accuracy, respectively, and FOCL[as] can identify lung cancer, diabetes, heart disease, and arrhythmia with 82.67%, 81.60%, 86.88%, and 70.73%, respectively.
Bin Alam et al [80], 2021	Pregnancy and childbirth	To evaluate bagging classifiers in birth mode prediction and identified key factors influencing cesarean section	DT bagging produces the highest accuracy of 0.87 with a 0.60 F_1-score for predicting birth mode.

XSL•FO
RenderX

Study	Health focus	Objective of the study	Outcomes
Kowsher et al [81], 2021	Pregnancy and childbirth	To predict the mode of childbirth that will be appropriate for pregnant women	Quadratic discriminant analysis has the highest accuracy and F_1-score of 0.97992 and 0.979960, respectively, in predicting the mode of childbirth.
Hossain et al [82], 2022	Pregnancy and childbirth	To predict unwanted pregnancies among married women in Bangladesh using ML	Elastic net regression algorithm showed the best results and the most accurate classification for predicting unwanted pregnancy among Bangladeshi women with an accuracy of 77.51%.
Sarma et al [83], 2020	Vector-borne diseases (malaria)	To predict dengue fever at an early stage using ML algorithm	DT algorithm performs well in classifying normal fever and dengue fever, with an accuracy of 79%.
Dey et al [84], 2022	Vector-borne diseases (malaria)	To develop an ML model that can use relevant information about the factors that cause dengue outbreaks within a geographic region	SVR identified the dengue outbreak accurately with a prediction accuracy of 75%.
Islam et al [85], 2021	Vector-borne diseases (malaria)	To predict the probability of dengue fever before taking the pathological test	DT and naïve Bayes achieved a higher accuracy of 100% in predicting dengue fever.
Kassim et al [86], 2021	Vector-borne diseases (malaria)	To detect RBCs[at] in malaria diagnostic smears using dual deep learning	The proposed RBCNet[au] architecture detected RBC with F_1-score, precision, and recall of 97.76%, 97.51%, and 98.07%, respectively.
Vink et al [87], 2013	Vector-borne diseases (malaria)	To detect malaria parasite using vision-based malaria parasite selection	In healthy samples, the system achieved an overall specificity of 99.999978% at the level of (infected) RBCs and a sensitivity of 75% at the cell level, enabling the detection of low parasite densities in a fast and easy-to-use manner.
Masud et al [88], 2020	Preventive health (appendicitis)	To identify risk factors for appendicitis using preoperative symptoms and developed an Android-based app	Apriori algorithm could correctly classify affected patients between medium to very high risk with 99% accuracy.
Sampa et al [89], 2020	Preventive health (uric acid)	To predict the blood uric acid level using ML model based on basic health checkup test results, dietary information, and sociodemographic information	Boosted DT regression can predict patients with a high risk of increased uric acid levels with an RMSE of 0.03.
Ferdowsy et al [90], 2021	Preventive health (obesity)	To predict obesity by identifying risk factors of obesity using ML-based data analysis	The LR algorithm performed better than other algorithms to predict obesity with an accuracy of 97.09%, whereas the gradient boosting algorithm produced the poorest accuracy of 64.08%.
Riajuliislam et al. [91], 2021	Preventive health (thyroid)	To predict hypothyroidism at an early stage by analyzing clinical and demographic data	RFE[av] performed well in predicting features with an accuracy of 99.35%.
Diptu et al [92], 2018	Ophthalmology	To detect glaucoma with the data obtained from OCT[aw] and tonometry test	ANFIS can detect glaucoma from OCT and tonometry tests with an accuracy of 81.25%.

XSL•FO
RenderX

Study	Health focus	Objective of the study	Outcomes
Hassan et al [11], 2021	Health policy	To assess the policy and regulations for medical devices and policy for health data in Bangladesh and compared these with those in the United States and European Union. In addition, identified the gaps in implementing big data and ML in health care	The current regulatory framework of Bangladesh for medical device regulation is inadequate to regulate AI-based medical devices. Furthermore, Bangladesh has neither comprehensive data protection legislation nor specific sectoral laws for health data privacy.

[a]LR: logistic regression.
[b]FPM: Facebook Prophet model.
[c]RMSE: root mean square error.
[d]MAE: mean absolute error.
[e]CXR: chest x-ray.
[f]LSTM: long short-term memory.
[g]SVR: support vector regression.
[h]RF: random forest.
[i]AI: artificial intelligence.
[j]DNN: deep neural network.
[k]MAPE: mean absolute percentage error.
[l]ANFIS: adaptive neuro-fuzzy interference system.
[m]AUC: area under the curve.
[n]ARIMA: autoregressive integrated moving average.
[o]MSE: mean squared error.
[p]NCD: noncommunicable disease.
[q]KNN: k-nearest neighbor.
[r]GP: Gaussian process.
[s]AUROC: area under the receiver operating characteristic curve.
[t]ML: machine learning.
[u]CART: classification and regression tree.
[v]IBK: Instance-Based K-nearest Neighbors.
[w]DT: decision tree.
[x]XGB: extreme gradient boost.
[y]LGBM: light gradient boost Machine.
[z]SVM: support vector machine.
[aa]CAN: Convolutional Attention Network.
[ab]DPN: Dual Path Networks.
[ac]RET: RetinaNet.
[ad]CVD: cardiovascular disease.
[ae]SVMRFE-GB: Support Vector Machine Recursive Feature Elimination with Gradient Boosting.
[af]GBM: gradient boosting machine.
[ag]LDA: linear discriminant analysis.
[ah]CKD: chronic kidney disease.
[ai]WBC: white blood cell.
[aj]RCNN: Region-Based Convolutional Neural Networks.
[ak]ANN: artificial neural network.
[al]PTB: preterm birth.
[am]GA: gestational age.
[an]LMP: last menstrual period.
[ao]ASD: autism spectrum disorder.
[ap]CNN: convolutional neural network.
[aq]GBDT: gradient boosting decision tree.
[ar]ACN-FOCL: Attention Convolutional Network-Focused Clustering.
[as]FOCL: Focused Clustering.
[at]RBC: red blood cell.
[au]RBCNet: Residual Bi-directional Convolutional Network.

XSL•FO
RenderX

[av]REF: Recursive Elimination Feature.

[aw]OCT: optical coherence tomography.

Synthesis of Findings

According to the findings, the studies intended to predict various diseases using various risk factor data (33/77, 43%), classify patients in disease-exposed and unexposed groups (21/77, 27%), predict trends of the outbreak (18/77, 23%), formulate intelligent systems for health service (3/77, 4%), and so on. One study explored the regulatory challenges of using big data technology in the health sector, as well as making recommendations for future policy. Another study tried to detect glaucoma with the data obtained from optical coherence tomography and the tonometry test.

The study's findings indicate that, compared to secondary data, the use of prediction and classification algorithms on primary data has significantly grown (40/77, 52%). In a significant number of studies (37/77, 48%), researchers used data from the Bangladesh Demographic and Health Survey, which is part of the global Demographic and Health Surveys program, designed to collect information on fertility, family planning, and maternal and child health. Furthermore, a substantial portion of studies (12/77, 16%) made use of publicly available data from web-based sources, such as Kaggle, the University of California, Irvine repository, and Worldometer. Researchers validated their prediction algorithms within the context of Bangladesh.

In this review, we observed the adoption of digital interventions, including the use of mobile devices for public health purposes. This review explored articles (4/77, 5%) that leveraged mobile devices for their research, creating intelligent risk prediction tools by analyzing data collected through ML- and DL-based algorithms. The objective was to demonstrate the value of mobile apps in obtaining real-time health data and predicting health status. The primary outcomes of these studies underscored the utility of intelligent mobile apps for self-assessing an individual's health and offering essential therapy recommendations.

On the basis of the data extracted in this review, 44 distinct AI, ML, and DL algorithms commonly used in public health studies were identified. Table 3 presents the algorithms according to different categories and their respective use frequencies.

XSL•FO
RenderX

Table 3. Different categories of available algorithms found in this scoping review (N=77).

Category and subcategory	Algorithm	Frequency, n (%)
Supervised		
Classification	Decision tree	37 (48)
Classification	Random forest	37 (48)
Classification	Logistic regression	35 (45)
Classification	k-nearest neighbor	27 (35)
Classification	Naive Bayes	22 (29)
Classification	Quadratic discriminant analysis	1 (1)
Classification	Bagging classifier	1 (1)
Classification	Bagged classification and regression tree	1 (1)
Classification	CatBoost	1 (1)
Classification or regression	Artificial neural network	11 (14)
Classification or regression	Convolutional neural network	7 (9)
Regression	Gradient boosting	9 (12)
Regression	Linear regression	8 (10)
Regression	Adaptive boosting	7 (9)
Regression	Support vector regression	2 (3)
Regression	Elastic net regression	2 (3)
Regression	Gaussian process	1 (1)
Regression	Ridge regression	1 (1)
Regression	Multiple linear regression	1 (1)
Regression	Least absolute shrinkage and selection operator	4 (5)
Regression	Autoregressive	3 (4)
Time series	Facebook prophet time series model	5 (6)
Time series	Recurrent neural network	2 (3)
Time series	Long short-term memory	2 (3)
Unsupervised		
Clustering	k-means clustering	1 (1)
Dimensionality reduction	Principal component analysis	1 (1)
Time series	Autoregressive integrated moving average	4 (5)
Time series	Apriori algorithm	2 (3)

Among these, the random forest algorithm emerged as a standout performer, delivering the highest accuracy in 21% (16/77) of the studies. The decision tree classifier was also highly effective, with a track record of accurate predictions in 16% (12/77) of the studies. However, for maternal and child health, logistic regression showed substantial promises, while for outbreak trend prediction, the Facebook Prophet model and long short-term memory (LSTM) demonstrated strong performance. In the mental health domain, random forest excelled, and for NCDs, random forest and support vector machine exhibited strong predictive capabilities. In the context of vector-borne diseases, the decision tree algorithm proved effective.

Several other algorithms showcased their utility in various domains, including support vector machine (9/77, 12%), logistic regression (8/77, 10%), Naïve Bayes (4/77, 5%), k-nearest neighbor (3/77, 4%), convolutional neural network (3/77, 4%), gradient boosting (3/77, 4%), LSTM (3/77, 4%), Facebook Prophet time series model (3/77, 4%), artificial neural network (2/77, 3%), XGBoost (2/77, 3%), a priori algorithm (2/77, 3%), and linear discriminant analysis (1/77, 1%). In addition, support vector regression (1/77, 1%), bagging classifier (1/77, 1%), quadratic discriminant analysis (1/77, 1%), multiple linear regression (1/77, 1%), elastic net regression (1/77, 1%), and bagged classification and regression tree (1/77, 1%) displayed their potential in achieving high prediction accuracy in their respective research studies.

This comprehensive analysis highlights the diverse and effective use of these algorithms in public health research, offering valuable insights for researchers and professionals in the field.

XSL•FO
RenderX

Practical Implication of Studies

This review identified a number (9/77, 12%) [17,24,39,42,47,50,54,87,88] of studies where researchers have made web-based and Android-based solutions by applying AI-, ML- and DL-based models after conducting their research. Of the 9 studies, 6 (67%) studies developed apps that mostly focused on NCDs to allow persons to self-assess their own physical condition. The other 3 (33%) studies developed web-based prediction tools for identifying the trends and hot spots of the COVID-19 pandemic, which were found effective when the world suffered a lot from the pandemic situation.

Limitations Mentioned by the Studies

According to the findings, this scoping review observed a significant number of limitations mentioned by the researchers. In approximately 27% (21/77) studies, researchers have expressed their concerns about the datasets they have used. The articles also expressed concerns about using a large amount of data with different dimensions for risk factor identification and prediction of NCDs. Sociodemographic data (8/77, 10%), clinical feature data (15/77, 19%), and lifestyle data (12/77, 16%) were crucial for effective risk factor identification as well as disease prediction, as mentioned in these 21 studies. However, the researchers were able to work on a small sample size with inadequate features, which overfitted their predictions, as reported by the articles.

For outbreak prediction, in 25% (19/77) of the studies, researchers mentioned the necessity to use diversified data from different dimensions with proper tuning of the AI, ML, and DL models on a regular basis. Environmental data, weather data, population movement data, quarantine data, vaccination information, clinical symptoms of patients, as well as crowd-sourced data were required for effective outbreak prediction. However, due to the lack of these features, they expressed concern that their provided prediction model may not work perfectly in all contexts.

Data source is also vital for an effective predictive model, as mentioned in some of the studies (12/77, 16%). Primary-sourced datasets seemed suitable for predictive modeling as they are usually collected in accordance with the research objective, whereas secondary-sourced datasets have their own research objectives, which was not convenient for effective predictive modeling.

Risk factor identification and prediction modeling were revealed as very sensitive for child health–related diseases. Some researchers mentioned using factors from different domains as well as clinical features (10/77, 13%).

Discussion

Principal Findings

This scoping review is an attempt to explore research focusing on Bangladesh's perspective on using BDA, AI, ML, and DL technologies across various health care domains. The 77 studies included in this review were mostly quantitative in nature and focused on 7 distinct health care domains, including infectious diseases, NCDs, maternal health, child health, preventive health, vector-borne diseases, and mental health. While most (65/77, 84%) studies adopted a cross-sectional study design, a few used cohort (5/77, 6%), case-control (4/77, 5%), and retrospective (3/77, 4%) designs. The review revealed 44 distinct AI, ML, and DL algorithms for risk factor identification, classification, prediction, and trend analysis. The most used algorithms were decision tree, random forest, logistic regression, Naïve Bayes, artificial neural network, and k-nearest neighbor. Notably, among the studies, the random forest surpassed others in classification and prediction by accuracy and F_1-score, followed by decision trees. However, in trend analysis, the Facebook Prophet model and LSTM performed better.

These technologies have exhibited effectiveness in classification, prediction, and severity analysis across various health care domains. For instance, time series algorithms, such as Facebook Prophet model and LSTM, can be invaluable in potential outbreak prediction, such as the COVID-19 pandemic. Sarkar et al [20] explored better performance in forecasting and severity analysis of COVID-19 outbreaks in Bangladesh using the Facebook Prophet model. Similarly, these algorithms can be of great use in addressing other epidemics and endemics, such as dengue and diarrhea outbreaks in Bangladesh and other South Asian countries. The use of these algorithms may include but is not limited to hot spot detection, disease pattern recognition, disease severity detection, and resource allocation for effective management. Moreover, Islam et al [48] and Asaduzzaman et al [39] incorporated different ML techniques (decision tree, support vector machine, XGBoost, gradient boosting machine, logistic regression, and linear discriminant analysis) to identify individuals at risk of NCDs, such as diabetes, hypertension, and cancer, as well as explored potential factors influencing the severity, considering the growing burden of NCDs in Bangladesh.

Furthermore, ML-driven classification algorithms can play a pivotal role in categorizing patients based on disease severity, facilitating efficient resource allocation in health care facilities. For example, Rahman et al [17] developed a web application incorporating ML algorithms to identify the severity of COVID-19 among patients in Bangladesh, enabling health care providers to prioritize treatment and intervention strategies effectively. Such categorization technology can reduce human error and overall stress if incorporated into high mobility wards (eg, diarrhea and casualty) and emergency rooms. However, future studies that focus on efficient fine-tuning and validation strategies are needed to ensure higher specificity and prevent false alarms.

Apart from these, several clinical decision support systems have been developed to assist depression diagnosis, considering the recent conversations on mental health, anxiety, and depression [77,79]. Although the models were proven successful in predicting and classifying information to aid public health experts, their real-world validation remains somewhat limited. Few studies (0/77, 13%) tested algorithms in controlled clinical settings with smaller sample sizes [17,32,36,41,49,70,73,76,83,88,89], while others (9/77, 12%) developed AI-, ML-, and DL-based applications without any real-world effectiveness testing [17,24,39,42,47,50,54,87,88].

XSL•FO
RenderX

In Bangladesh, the application of AI, ML, and DL in health care and public health has primarily been confined to risk factors identification, classification of exposed and unexposed disease groups, early danger-sign prediction, trend analysis, and early forecasting of outbreaks on a smaller scale. However, studies from high-income countries have presented the successful implementation of AI, ML, and DL interventions in diverse health care settings as well as mentioned stronger policy evidence [93-96]. For example, scoping reviews mentioned the successful use of BDA for mental health treatment [97], decision support management [98], patient classification, emergency department triage, and hospital resource allocation [99] as well as effective management of virus transmissions such as influenza, Zika virus, AIDS, chikungunya, and Ebola [100]. In Bangladesh, except for 1 policy article on big data, no studies mentioned any call for policies or actions on BDA for the public health domain [11].

While AI and ML have been used to forecast potential outbreaks such as COVID-19, influenza, and dengue in Bangladesh, BDA could be used in such domains for more comprehensive outcomes. Moreover, despite remarkable performances in screening, diagnosis, and prediction of cardiovascular diseases, the use of BDA in clinical settings remains limited [101]. Nonetheless, AI and ML have the potential to improve public health efforts and promote quality health care for all individuals across diverse communities, as mentioned by 1 review article [102].

Research Implications

Our scoping review unveiled that integrating ML algorithms for diagnosis improves accuracy, saving time and effort. Moreover, AI-based decision support systems enhance prescription recommendations, aiding health care professionals in resource-constrained settings such as Bangladesh. Although ML and AI algorithms showed potential regarding outbreak trend prediction and hot spot detection, most studies were conducted on a smaller scale with a smaller sample size and within controlled environments. To gain a better understanding of the effectiveness of AI, ML, and DL algorithm–based prediction models and decision support systems, it is imperative to conduct large-scale studies in diverse environments and real clinical settings. Furthermore, in countries with robust health care systems, ML-based prediction systems tend to be successful in analyzing patient data from electronic health records and generating illness scores for predicting future health care needs [6]. For example, the Cambridge Adjutorium system, developed by a team at the Cambridge Centre for AI in Medicine in the United Kingdom, uses cutting-edge ML techniques to accurately forecast death rates, the necessity for intensive care unit admission, and the need for ventilation in patients with COVID-19 [103]. AI-powered physicians can offer health advice for common medical conditions, reducing hospital visits and staying through digital health coaching. Yet, applying these approaches in lower-income countries such as Bangladesh is limited due to structural, regulatory, and legislative constraints. Further studies and future research are essential to assess the effectiveness of AI-, ML-, and DL-based solutions in genuine clinical and larger population-based settings.

Policy Implications

Technology has become integral to Bangladesh's health care system, with platforms such as District Health Information System 2, electronic management information system, and open medical record system collecting vast amounts of patient data [12]. Despite challenges in structured data acquisition, data security, and processing capacity, research and implementation of AI, ML, and DL initiatives are expanding, promising substantial benefits for resource-limited countries such as Bangladesh. The government of Bangladesh is actively integrating AI, ML, and knowledge-based systems into health care operations as part of its strategic plan [104]. They are supporting technology industries and small teams in developing such solutions. Moreover, the government of Bangladesh is hosting national and international events and offering small grants for AI-driven health care projects, reflecting a strong commitment to incorporating these technologies despite financial limitations.

Training health care workers and ensuring consistent data entry are crucial to enhancing data integration and interoperability. Implementing AI, ML, and DL pilot projects can help optimize resources and gather insights for incremental and broader implementation. Moreover, collaborating with private technology companies, universities, and international organizations through public-private partnerships can codevelop and fund these initiatives, easing the financial burden on the government.

Furthermore, it is essential to develop robust data privacy and security frameworks suited to local needs and establish standards for the ethical use of AI, ML, and DL technologies. Investing in training health care workers and IT professionals, as well as encouraging local universities to include AI and ML in their curricula, will build a skilled workforce. In addition, targeted funding should prioritize high-return investments such as predictive analytics and AI-assisted diagnostic tools supported by thorough cost-benefit analyses.

Despite budget constraints, these strategic initiatives can potentially promote AI- and ML-based solutions in health care settings. By focusing on these actionable steps, Bangladesh can effectively integrate these technologies, ensuring sustainable advancements in digital health and improved overall health outcomes.

Strengths and Limitations

Our study's primary strength lies in the rigorous adherence to the scoping review methodology at every stage. This encompassed thorough screening and quality assessment of the studies included in our analysis. To our knowledge, this is the first initiative to outline the extension and utility of AI, ML, DL, and BDA in health care, along with illustrating the potential scope for interventions, research, and policies around such burgeoning domains in Bangladesh, which is also applicable for similar resource-limited settings. However, further studies are needed to investigate and contrast findings from other settings.

Nonetheless, our study does have other limitations. First, we exclusively considered literature in the English language,

XSL•FO
RenderX

potentially excluding studies in other languages that could have offered valuable insights into the applications of DL, AI, ML, and BDA. However, it is worth noting that previous research has suggested that language restrictions in scoping reviews do not substantially introduce bias [105]. Moreover, we only included MEDLINE, IEEE Xplore, Scopus, and Embase databases to search the relevant articles, whereas we acknowledge that there might be potential useful insights from gray literature, including government of Bangladesh reports and nongovernmental organization publications. However, our included databases thoroughly cover peer-reviewed literature in medicine, engineering, and public health, aligning closely with our scoping review's subject matter and ensuring the reliability and credibility of the evidence synthesized. Furthermore, several studies in our analysis involved small population sizes, and the overall quality of many included studies left room for doubt. Consequently, caution is warranted when interpreting our findings. Another limitation pertained to including studies that relied on secondary datasets and global data sources, primarily due to the scarcity of datasets specific to Bangladesh. We excluded 22 (15.4%) of the 143 articles from our review due to inaccessibility, which may have contributed valuable insights to our review. Finally, some studies did not mention explicit information about their ML models. Because our primary objective did not involve assessing the quality of modeling, these studies were included as part of our analysis.

Conclusions

This scoping review outlines available research on BDA, AI, ML, and DL in the health care system sector of Bangladesh. The increasing demand for these cross-cutting technologies for analyzing and processing health data is steadily expanding. The use of ML for disease prediction and classification can aid health care professionals in enhancing diagnosis and treatments. Advanced algorithmic scoring systems based on multiple health determinants have immense potential to predict diagnosis, readmission, and mortality rates in the near future, potentially saving resources and efforts. Therefore, it is crucial to use BDA, AI, ML, and DL technology in designing future health interventions and policies to maximize the use of health resources and maintain optimum competitiveness. Finally, it is crucial to identify priority areas for the application of these technologies and develop a national strategy for the integration of AI and data science. This strategy should include but not be limited to components such as data monitoring, privacy and security, data availability and validation for research, academic and industry collaboration, e-infrastructure establishment, fostering a supportive legal framework, and training data scientists to ensure a successful transition to a future digital health era.

Acknowledgments

The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors. The first author (MAA) would like to thank Md Al Mamun, manager of library and information services at the International Centre for Diarrhoeal Disease Research, Bangladesh, for their help with the database searches. Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of our research.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors' Contributions

MAA and MRUZS defined the research question of the scoping review, conducted the literature search, and summarized the findings. EA, STE, NN, and TTI helped in the title and abstract screening, full-text screening, and quality assessment. FR, MH, AS, SR, KMT, MJC, QSR, TH, AL, AZ, and JR reviewed the manuscript and gave important feedback. AA, SMR, SEA, and AER supervised the research and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File , 50 KB - jmir_v26i1e54710_app1.docx]

Multimedia Appendix 2
Database search strategy.
[PDF File (Adobe PDF File), 746 KB - jmir_v26i1e54710_app2.pdf]

Multimedia Appendix 3
Quality assessment.

XSL•FO
RenderX

[DOCX File , 28 KB - jmir_v26i1e54710_app3.docx]

References

1. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, Dutch-Covid-19 response team. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med 2020 Sep 16;26(9):1405-1410. [doi: 10.1038/s41591-020-0997-y] [Medline: 32678356]
2. Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, et al. Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manage 2021 Apr;57:101994. [doi: 10.1016/j.ijinfomgt.2019.08.002]
3. Abma T, Banks S, Cook T, Dias S, Madsen W, Springett J, et al. Participatory Research for Health and Social Well-Being. Cham, Switzerland: Springer; 2019.
4. Donelle L, Comer L, Hiebert B, Hall J, Shelley JJ, Smith MJ, et al. Use of digital technologies for public health surveillance during the COVID-19 pandemic: a scoping review. Digit Health 2023 May 17;9:20552076231173220 [FREE Full text] [doi: 10.1177/20552076231173220] [Medline: 37214658]
5. Thayyib P, Mamilla R, Khan M, Fatima H, Asim M, Anwar I, et al. State-of-the-art of artificial intelligence and big data analytics reviews in five different domains: a bibliometric summary. Sustainability 2023 Feb 22;15(5):4026. [doi: 10.3390/su15054026]
6. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
7. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges. Int J Environ Res Public Health 2021 Jan 01;18(1):271 [FREE Full text] [doi: 10.3390/ijerph18010271] [Medline: 33401373]
8. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights 2016;8:1-10 [FREE Full text] [doi: 10.4137/BII.S31559] [Medline: 26843812]
9. Berros N, El Mendili F, Filaly Y, El Bouzekri El Idrissi Y. Enhancing digital health services with big data analytics. Big Data Cogn Comput 2023 Mar 30;7(2):64. [doi: 10.3390/bdcc7020064]
10. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
11. Hassan S, Dhali M, Zaman F, Tanveer M. Big data and predictive analytics in healthcare in Bangladesh: regulatory challenges. Heliyon 2021 Jun;7(6):e07179 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e07179] [Medline: 34141936]
12. Khan MA, Cruz VO, Azad AK. Bangladesh's digital health journey: reflections on a decade of quiet revolution. WHO South East Asia J Public Health 2019 Sep;8(2):71-76. [doi: 10.4103/2224-3151.264849] [Medline: 31441440]
13. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
14. Alam MA. Implication of big data analytics, artificial intelligence, machine learning and deep learning in the public health sector of Bangladesh: a scoping review protocol. Open Science Framework. 2023. URL: https://osf.io/5hgky/metadata/osf [accessed 2024-04-29]
15. Porritt K, Gomersall J, Lockwood C. JBI's systematic reviews: study selection and critical appraisal. Am J Nurs 2014 Jun;114(6):47-52. [doi: 10.1097/01.NAJ.0000450430.97383.64] [Medline: 24869584]
16. Sarker A, Das RA, Ether S, Shariful Islam MA, Saif-Ur-Rahman KA. Non-pharmacological interventions for the prevention of type 2 diabetes in low-income and middle-income countries: a systematic review of randomised controlled trials. BMJ Open 2022 Jun 06;12(6):e062671 [FREE Full text] [doi: 10.1136/bmjopen-2022-062671] [Medline: 35667728]
17. Rahman T, Khandakar A, Hoque ME, Ibtehaz N, Bin Kashem S, Masud R, et al. Development and validation of an early scoring system for prediction of disease severity in COVID-19 using complete blood count parameters. IEEE Access 2021;9:120422-120441. [doi: 10.1109/access.2021.3105321]
18. Satu MS, Howlader KC, Mahmud M, Kaiser MS, Islam SM, Quinn JM, et al. Short-term prediction of COVID-19 cases using machine learning models. Appl Sci 2021 May 08;11(9):4266. [doi: 10.3390/app11094266]
19. Rafi TH. An ensemble deep transfer-learning approach to identify COVID-19 cases from chest X-ray images. In: Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. 2020 Presented at: CIBCB '20; October 27-29, 2020; Via del Mar, Chile p. 1-5 URL: https://ieeexplore.ieee.org/document/9277695 [doi: 10.1109/CIBCB48159.2020.9277695]
20. Sarkar O, Ahamed MF, Chowdhury P. Forecasting and severity analysis of COVID-19 using machine learning approach with advanced data visualization. In: Proceedings of the 23rd International Conference on Computer and Information Technology. 2020 Presented at: ICCIT '20; December 19-21, 2020; Dhaka, Bangladesh p. 1-6 URL: https://ieeexplore.ieee.org/abstract/document/9392704 [doi: 10.1109/iccit51783.2020.9392704]

XSL•FO
RenderX

21. Mohammad Masum AK, Khushbu SA, Keya M, Abujar S, Hossain SA. COVID-19 in Bangladesh: a deeper outlook into the forecast with prediction of upcoming per day cases using time series. Procedia Comput Sci 2020;178:291-300 [FREE Full text] [doi: 10.1016/j.procs.2020.11.031] [Medline: 33520018]
22. Hassan MR, Hassan MM, Altaf M, Yeasar MS, Hossain MI, Fatema K, et al. B5G-enabled distributed artificial intelligence on edges for COVID-19 pandemic outbreak prediction. IEEE Network 2021 May;35(3):48-55. [doi: 10.1109/mnet.011.2000713]
23. Chowdhury AA, Hasan KT, Hoque KK. Analysis and prediction of COVID-19 pandemic in Bangladesh by using ANFIS and LSTM network. Cognit Comput 2021;13(3):761-770 [FREE Full text] [doi: 10.1007/s12559-021-09859-0] [Medline: 33868501]
24. Ullah AN, Rahman MH, Allayear SM, Khan ML, Faysal SM, Chowdhury AB, et al. Helping healthcare providers to differentiate COVID-19 pneumonia by analyzing digital chest X-Rays: role of artificial intelligence in healthcare practice. Int J Biomed 2022 Sep 05;12(3):459-465. [doi: 10.21103/article12(3)_oa21]
25. Leon MI, Iqbal MI, Azim SM, Al Mamun KA. Predicting COVID-19 infections and deaths in Bangladesh using machine learning algorithms. In: Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development. 2021 Presented at: ICICT4SD '21; February 27-28, 2021; Dhaka, Bangladesh p. 70-75 URL: https://ieeexplore.ieee.org/document/9396820 [doi: 10.1109/icict4sd50815.2021.9396820]
26. Rahman M, Ahmed MT, Nur S, Touhidul Islam AZ. The prediction of coronavirus disease 2019 outbreak on Bangladesh perspective using machine learning: a comparative study. Int J Electr Comput Eng 2019;12(4):4276-4287 [FREE Full text] [doi: 10.11591/ijece.v12i4.pp4276-4287]
27. Haq I, Hossain MI, Saleheen AA, Nayan M, Mila MS. Prediction of COVID-19 pandemic in Bangladesh: dual application of susceptible-infective-recovered (SIR) and machine learning approach. Interdiscip Perspect Infect Dis 2022;2022:8570089 [FREE Full text] [doi: 10.1155/2022/8570089] [Medline: 35497651]
28. Absar N, Uddin N, Khandaker MU, Ullah H. The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases. Infect Dis Model 2022 Mar;7(1):170-183 [FREE Full text] [doi: 10.1016/j.idm.2021.12.005] [Medline: 34977438]
29. Karmokar J, Islam MA, Uddin M, Hassan MR, Yousuf MS. An assessment of meteorological parameters effects on COVID-19 pandemic in Bangladesh using machine learning models. Environ Sci Pollut Res Int 2022 Sep 06;29(44):67103-67114 [FREE Full text] [doi: 10.1007/s11356-022-20196-z] [Medline: 35522407]
30. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel AS, Naheyan T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. Lancet Digit Health 2021 Sep;3(9):e543-e554 [FREE Full text] [doi: 10.1016/S2589-7500(21)00116-3] [Medline: 34446265]
31. Azman AS, Lessler J, Luquero FJ, Bhuiyan TR, Khan AI, Chowdhury F, et al. Estimating cholera incidence with cross-sectional serology. Sci Transl Med 2019 Feb 20;11(480):eaau6242 [FREE Full text] [doi: 10.1126/scitranslmed.aau6242] [Medline: 30787170]
32. Pranto B, Mehnaz SM, Mahid EB, Sadman IM, Rahman A, Momen S. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information 2020 Jul 23;11(8):374. [doi: 10.3390/info11080374]
33. Nishat MM, Faisal F, Mahbub MA, Mahbub MH, Islam S. Performance assessment of different machine learning algorithms in predicting diabetes mellitus. Biosci Biotechnol Res Commun 2021 Mar 25;14(1):74-82. [doi: 10.21786/bbrc/14.1/10]
34. Islam MM, Rahman MJ, Chandra Roy D, Maniruzzaman M. Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. Diabetes Metab Syndr 2020;14(3):217-219. [doi: 10.1016/j.dsx.2020.03.004] [Medline: 32193086]
35. Jahan N, Islam A, Al Mamun A. Machine learning with factor scoring to predict diabetes risk level in Bangladesh. Int J Sci Technol Res 2020;9(2):2863-2867 [FREE Full text]
36. Emon MU, Keya MS, Kaiser MS, Islam MA, Tanha T, Zulfiker MS. Primary stage of diabetes prediction using machine learning approaches. In: Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems. 2021 Presented at: ICAIS '21; March 25-27, 2021; Coimbatore, India p. 3643-3647 URL: https://ieeexplore.ieee.org/document/9395968 [doi: 10.1109/icais50930.2021.9395968]
37. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, et al. Early prediction of diabetes using an ensemble of machine learning models. Int J Environ Res Public Health 2022 Sep 28;19(19):12378 [FREE Full text] [doi: 10.3390/ijerph191912378] [Medline: 36231678]
38. Wood DA. Integrated statistical and machine learning analysis provides insight into key influencing symptoms for distinguishing early-onset type 2 diabetes. Chronic Dis Transl Med 2022 Dec 31;8(4):281-295 [FREE Full text] [doi: 10.1002/cdt3.39] [Medline: 36420178]
39. Asaduzzaman S, Ahmed MR, Rehana H, Chakraborty S, Islam MS, Bhuiyan T. Machine learning to reveal an astute risk predictive framework for gynecologic cancer and its impact on women psychology: Bangladeshi perspective. BMC Bioinformatics 2021 Apr 24;22(1):213 [FREE Full text] [doi: 10.1186/s12859-021-04131-6] [Medline: 33894739]
40. Maliha SK, Ema RR, Ghosh SK, Ahmed H, Mollick MR, Islam T. Cancer disease prediction using naive Bayes,K-nearest neighbor and J48 algorithm. In: Proceedings of the 10th International Conference on Computing, Communication and

XSL•FO
RenderX

Networking Technologies. 2019 Presented at: ICCCNT '19; July 6-8, 2019; Kanpur, India p. 1-7 URL: https://ieeexplore.ieee.org/document/8944686 [doi: 10.1109/icccnt45670.2019.8944686]
41. Rejaul Islam Royel M, Ajmanur Jaman M, Al Masud F, Ahmed A, Muyeed A, Ahmed K. Machine learning and data mining methods in early detection of stomach cancer risk. J Appl Sci Eng 2021;24(1):1-8 [FREE Full text] [doi: 10.6180/jase.202102_24(1).0001]
42. Ahmed K, Emran AA, Jesmin T, Mukti RF, Rahman MZ, Ahmed F. Early detection of lung cancer risk using data mining. Asian Pac J Cancer Prev 2013 Jan 31;14(1):595-598 [FREE Full text] [doi: 10.7314/apjcp.2013.14.1.595] [Medline: 23534801]
43. Rashid M, Alkhodari M, Mukit A, Ahmed KI, Mostafa R, Parveen S, et al. Machine learning for screening microvascular complications in type 2 diabetic patients using demographic, clinical, and laboratory profiles. J Clin Med 2022 Mar 09;11(4):903 [FREE Full text] [doi: 10.3390/jcm11040903] [Medline: 35207179]
44. Kayyum S, Miah J, Shadaab A, Minhazu Islam M, Islam M, Nipun SA, et al. Data analysis on myocardial infarction with the help of machine learning algorithms considering distinctive or non-distinctive features. In: Proceedings of the 2020 International Conference on Computer Communication and Informatics. 2020 Presented at: ICCCI '20; January 22-24, 2020; Coimbatore, India p. 179-185 URL: https://ieeexplore.ieee.org/document/9104104 [doi: 10.1109/iccci48352.2020.9104104]
45. Tabassum S, Zaman MI, Ullah MS, Rahaman A, Nahar S, Islam AK. The cardiac disease predictor: IoT and ML driven healthcare system. In: Proceedings of the 4th International Conference on Electrical Information and Communication Technology. 2019 Presented at: EICT '19; December 20-22, 2019; Khulna, Bangladesh p. 1-6 URL: https://ieeexplore.ieee.org/document/9068821 [doi: 10.1109/eict48899.2019.9068821]
46. Chowdhury MN, Ahmed E, Siddik MA, Zaman AU. Heart disease prognosis using machine learning classification techniques. In: Proceedings of the 6th International Conference for Convergence in Technology. 2021 Presented at: I2CT '21; April 2-4, 2021; Mumbai, India p. 1-6 URL: https://ieeexplore.ieee.org/document/9418181 [doi: 10.1109/i2ct51068.2021.9418181]
47. Fahim KE, Yassin H, Amin MH, Dewan PD, Islam A. Detection of cardiovascular disease of patients at an early stage using machine learning algorithms. In: Proceedings of the 2022 International Conference on Healthcare Engineering. 2022 Presented at: ICHE '22; September 23-25, 2022; Johor, Malaysia p. 1-6 URL: https://ieeexplore.ieee.org/document/10179871 [doi: 10.1109/iche55634.2022.10179871]
48. Islam MM, Rahman MJ, Chandra Roy D, Tawabunnahar M, Jahan R, Ahmed NA, et al. Machine learning algorithm for characterizing risks of hypertension, at an early stage in Bangladesh. Diabetes Metab Syndr 2021;15(3):877-884. [doi: 10.1016/j.dsx.2021.03.035] [Medline: 33892404]
49. Arefa I, Alam MS, Siddiquee I, Siddique N. Performance analysis of machine learning algorithms for hypertension decision support system. In: Proceedings of the 2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things. 2019 Presented at: RAAICON '19; November 29-December 1, 2019; Dhaka, Bangladesh p. 15-20 URL: https://ieeexplore.ieee.org/document/9087522 [doi: 10.1109/raaicon48939.2019.8]
50. Islam SM, Talukder A, Awal MA, Siddiqui MM, Ahamad MM, Ahammed B, et al. Machine learning approaches for predicting hypertension and its associated factors using population-level data from three south Asian countries. Front Cardiovasc Med 2022 Mar 31;9:839379 [FREE Full text] [doi: 10.3389/fcvm.2022.839379] [Medline: 35433854]
51. Asadullah M, Hossain MM, Rahaman S, Amin MS, Sumy MS, Parh MY, et al. Evaluation of machine learning techniques for hypertension risk prediction based on medical data in Bangladesh. Indones J Electr Eng Comput Sci 2023 Sep 01;31(3):1794. [doi: 10.11591/ijeecs.v31.i3.pp1794-1802]
52. Ifraz GM, Rashid MH, Tazin T, Bourouis S, Khan MM. Comparative analysis for prediction of kidney disease using intelligent machine learning methods. Comput Math Methods Med 2021 Dec 3;2021:6141470-6141410 [FREE Full text] [doi: 10.1155/2021/6141470] [Medline: 34899968]
53. Ehsan MT, Sajal MS, Mamun KA. An automated cloud-based tool for screening of Parkinson's disease in Bangladesh. In: Proceedings of the 2nd International Conference on Robotics, Electrical and Signal Processing Techniques. 2021 Presented at: ICREST '21; January 5-7, 2021; Dhaka, Bangladesh p. 664-668 URL: https://ieeexplore.ieee.org/document/9331233 [doi: 10.1109/icrest51555.2021.9331233]
54. Akter Hossain M, Islam Sabik M, Muntasir I, Muzahidul Islam AK, Islam S, Ahmed A. Leukemia detection mechanism through microscopic image and ML techniques. In: Proceedings of the 2020 IEEE Region 10 Conference. 2020 Presented at: TENCON '20; November 16-19, 2020; Osaka, Japan p. 61-66 URL: https://ieeexplore.ieee.org/document/9293925/ [doi: 10.1109/tencon50793.2020.9293925]
55. Shahriar MM, Iqubal MS, Mitra S, Das AK. A deep learning approach to predict malnutrition status of 0-59 month's older children in Bangladesh. In: Proceedings of the 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology. 2019 Presented at: IAICT '19; July 1-3, 2019; Bali, Indonesia p. 145-149 URL: https://ieeexplore.ieee.org/document/8784823 [doi: 10.1109/iciaict.2019.8784823]
56. Methun MI, Kabir A, Islam S, Hossain MI, Darda MA. A machine learning logistic classifier approach for identifying the determinants of under-5 child morbidity in Bangladesh. Clin Epidemiol Glob Health 2021 Oct;12:100812. [doi: 10.1016/j.cegh.2021.100812]

XSL•FO
RenderX

57. Borson NS, Kabir MR, Zamal ZZ, Rahman RM. Correlation analysis of demographic factors on low birth weight and prediction modeling using machine learning techniques. In: Proceedings of the 4th World Conference on Smart Trends in Systems, Security and Sustainability. 2020 Presented at: WorldS4 '20; July 27-28, 2020; London, UK p. 169-173 URL: https://ieeexplore.ieee.org/document/9210338 [doi: 10.1109/worlds450073.2020.9210338]
58. Rahman SM, Ahmed NA, Abedin MM, Ahammed B, Ali M, Rahman MJ, et al. Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach. PLoS One 2021;16(6):e0253172 [FREE Full text] [doi: 10.1371/journal.pone.0253172] [Medline: 34138925]
59. Talukder A, Ahammed B. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. Nutrition 2020 Oct;78:110861. [doi: 10.1016/j.nut.2020.110861] [Medline: 32592978]
60. Khan JR, Tomal JH, Raheem E. Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh. Inform Health Soc Care 2021 Dec 02;46(4):425-442. [doi: 10.1080/17538157.2021.1904938] [Medline: 33851897]
61. Jehan F, Sazawal S, Baqui AH, Nisar MI, Dhingra U, Khanam R, Alliance for Maternal and Newborn Health Improvement, the Global Alliance to Prevent Prematurity and Stillbirth, the Prematurity Research Center at Stanford University. Multiomics characterization of preterm birth in low- and middle-income countries. JAMA Netw Open 2020 Dec 01;3(12):e2029655 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.29655] [Medline: 33337494]
62. Islam Pollob SM, Abedin MM, Islam MT, Islam MM, Maniruzzaman M. Predicting risks of low birth weight in Bangladesh with machine learning. PLoS One 2022 May 26;17(5):e0267190 [FREE Full text] [doi: 10.1371/journal.pone.0267190] [Medline: 35617201]
63. Mehedi Hasan M, Omar Faruk M, Biswas Biki B, Riajuliislatn M, Alam K, Farjana Shetu S. Prediction of pneumonia disease of newborn baby based on statistical analysis of maternal condition using machine learning approach. In: Proceedings of the 11th International Conference on Cloud Computing, Data Science & Engineering. 2021 Presented at: Confluence '21; January 28-29, 2021; Noida, India p. 919-924 URL: https://ieeexplore.ieee.org/document/9377169 [doi: 10.1109/confluence51648.2021.9377169]
64. Alliance for Maternal and Newborn Health Improvement (AMANHI) Gestational Age Study Group, Alliance for Maternal and Newborn Health Improvement (AMANHI) GA Study Group. Simplified models to assess newborn gestational age in low-middle income countries: findings from a multicountry, prospective cohort study. BMJ Glob Health 2021 Sep 13;6(9):e005688 [FREE Full text] [doi: 10.1136/bmjgh-2021-005688] [Medline: 34518201]
65. Mansur M, Afiaz A, Hossain MS. Sociodemographic risk factors of under-five stunting in Bangladesh: assessing the role of interactions using a machine learning method. PLoS One 2021;16(8):e0256729 [FREE Full text] [doi: 10.1371/journal.pone.0256729] [Medline: 34464402]
66. Khudri MM, Rhee KK, Hasan MS, Ahsan KZ. Predicting nutritional status for women of childbearing age from their economic, health, and demographic features: a supervised machine learning approach. PLoS One 2023 May 12;18(5):e0277738 [FREE Full text] [doi: 10.1371/journal.pone.0277738] [Medline: 37172042]
67. Islam MM, Rahman MJ, Islam MM, Roy DC, Ahmed NA, Hussain S, et al. Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh. Int J Cogn Comput Eng 2022 Jun;3:46-57. [doi: 10.1016/j.ijcce.2022.02.002]
68. Choudhury AA, Khan MR, Nahim NZ, Tulon SR, Islam S, Chakrabarty A. Predicting depression in Bangladeshi undergraduates using machine learning. In: Proceedings of the 2019 IEEE Conference on Region 10 Symposium. 2019 Presented at: TENSYMP '19; June 7-9, 2019; Kolkata, India p. 789-794 URL: https://ieeexplore.ieee.org/document/8971369 [doi: 10.1109/tensymp46218.2019.8971369]
69. Haque MM, Rabbani M, Dipal DD, Zarif MI, Iqbal A, Schwichtenberg A, et al. Informing developmental milestone achievement for children with autism: machine learning approach. JMIR Med Inform 2021 Jun 08;9(6):e29242 [FREE Full text] [doi: 10.2196/29242] [Medline: 33984830]
70. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, et al. Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: development and validation study. J Med Internet Res 2019 Apr 24;21(4):e13822 [FREE Full text] [doi: 10.2196/13822] [Medline: 31017583]
71. Imtinan Uddin JM, Fatema K, Kumar Dhar P. Depression risk prediction among tech employees in Bangladesh using Adaboosted decision tree. In: Proceedings of the 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering. 2020 Presented at: WIECON-ECE '20; December 26-27, 2020; Bhubaneswar, India p. 147-150 URL: https://ieeexplore.ieee.org/document/9397999 [doi: 10.1109/wiecon-ece52138.2020.9397999]
72. Khan SI, Islam A, Hossen A, Zahangir TI, Latiful Hoque SM. Supporting the treatment of mental diseases using data mining. In: Proceedings of the 2018 International Conference on Innovations in Science, Engineering and Technology. 2018 Presented at: ICISET '18; October 27-28, 2018; Chittagong, Bangladesh p. 339-344 URL: https://ieeexplore.ieee.org/document/8745591 [doi: 10.1109/iciset.2018.8745591]
73. Rois R, Ray M, Rahman A, Roy SK. Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms. J Health Popul Nutr 2021 Nov 27;40(1):50 [FREE Full text] [doi: 10.1186/s41043-021-00276-5] [Medline: 34838133]

XSL•FO
RenderX

74. Ahmed A, Sultana R, Ullas MT, Begom M, Rahi MM, Alam MA. A machine learning approach to detect depression and anxiety using supervised learning. In: Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering. 2020 Presented at: CSDE '20; December 16-18, 2020; Gold Coast, Australia p. 1-6 URL: https://ieeexplore.ieee.org/document/9411642 [doi: 10.1109/csde50874.2020.9411642]
75. Khan SI, Islam A, Zahangir TI, Hoque AM. Knowledge discovery from mental health data. In: Proceeding of the 2018 International Conference on Computer Networks, Big Data and IoT. 2018 Presented at: ICCBI '18; December 19-20, 2018; Madurai, India p. 881-888 URL: https://link.springer.com/chapter/10.1007/978-3-030-24643-3_103 [doi: 10.1007/978-3-030-24643-3_103]
76. Moon NN, Hossain RA, Jahan I, Shakil S, Uddin S, Hassan M, et al. Predicting the mental health of rural Bangladeshi children in coronavirus disease 2019. Int J Electr Comput Eng 2022 Oct 01;12(5):5501. [doi: 10.11591/ijece.v12i5.pp5501-5510]
77. Hu M, Nohara Y, Nakamura M, Nakashima N. Prediction and factor extraction of drug function by analyzing medical records in developing countries. Stud Health Technol Inform 2017;245:403-407. [Medline: 29295125]
78. Baba Y, Kashima H, Nohara Y, Kai E, Ghosh P, Islam R, et al. Predictive approaches for low-cost preventive medicine program in developing countries. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15; August 10-13, 2015; Sydney, Australia p. 1681-1690 URL: https://dl.acm.org/doi/10.1145/2783258.2788587 [doi: 10.1145/2783258.2788587]
79. Al Iqbal R. Hybrid clinical decision support system: an automated diagnostic system for rural Bangladesh. In: Proceedings of the 2012 International Conference on Informatics, Electronics & Vision. 2012 Presented at: ICIEV '12; May 18-19, 2012; Dhaka, Bangladesh p. 76-81 URL: https://ieeexplore.ieee.org/document/6317526 [doi: 10.1109/iciev.2012.6317526]
80. Bin Alam MS, Patwary MJ, Hassan M. Birth mode prediction using bagging ensemble classifier: a case study of Bangladesh. In: Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development. 2021 Presented at: ICICT4SD '21; February 27-28, 2021; Dhaka, Bangladesh p. 95-99 URL: https://ieeexplore.ieee.org/abstract/document/9396909 [doi: 10.1109/icict4sd50815.2021.9396909]
81. Kowsher M, Prottasha NJ, Tahabilder A, Habib K, Abdur-Rakib M, Shameem Alam M. Predicting the appropriate mode of childbirth using machine learning algorithm. Int J Adv Comput Sci Appl 2021;12(5):700-708. [doi: 10.14569/ijacsa.2021.0120582]
82. Hossain MI, Habib MJ, Saleheen AA, Kamruzzaman M, Rahman A, Roy S, et al. Performance evaluation of machine learning algorithm for classification of unintended pregnancy among married women in Bangladesh. J Healthc Eng 2022 May 28;2022:1460908-1460910 [FREE Full text] [doi: 10.1155/2022/1460908] [Medline: 35669979]
83. Sarma D, Hossain S, Mittra T, Bhuiya MA, Saha I, Chakma R. Dengue prediction using machine learning algorithms. In: Proceedings of the 2020 IEEE 8th R10 Humanitarian Technology Conference. 2020 Presented at: R10-HTC '20; December 1-3, 2020; Kuching, Malaysia p. 1-6 URL: https://ieeexplore.ieee.org/document/9357035 [doi: 10.1109/r10-htc49770.2020.9357035]
84. Dey SK, Rahman MM, Howlader A, Siddiqi UR, Uddin KM, Borhan R, et al. Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: a machine learning approach. PLoS One 2022 Jul 20;17(7):e0270933 [FREE Full text] [doi: 10.1371/journal.pone.0270933] [Medline: 35857776]
85. Islam MS, Khushbu SA, Rabby AS, Bhuiyan T. A study on dengue fever in Bangladesh: predicting the probability of dengue infection with external behavior with machine learning. In: Proceedings of the 5th International Conference on Intelligent Computing and Control Systems. 2021 Presented at: ICICCS '21; May 6-8, 2021; Madurai, India p. 1717-1721 URL: https://ieeexplore.ieee.org/document/9432288 [doi: 10.1109/iciccs51141.2021.9432288]
86. Kassim YM, Palaniappan K, Yang F, Poostchi M, Palaniappan N, Maude RJ, et al. Clustering-based dual deep learning architecture for detecting red blood cells in malaria diagnostic smears. IEEE J Biomed Health Inform 2021 May;25(5):1735-1746 [FREE Full text] [doi: 10.1109/JBHI.2020.3034863] [Medline: 33119516]
87. Vink JP, Laubscher M, Vlutters R, Silamut K, Maude RJ, Hasan MU, et al. An automatic vision-based malaria diagnosis system. J Microsc 2013 Jun;250(3):166-178. [doi: 10.1111/jmi.12032] [Medline: 23550616]
88. Masud FA, Royel MR, Sajal MM, Jahan S, Paul BK, Ahmed K. Smart risk prediction tools of appendicitis patients: a machine learning approach. Biointerface Res Appl Chem 2020 Jul 4;11(1):7804-7813. [doi: 10.33263/briac111.78047813]
89. Sampa MB, Hossain MN, Hoque MR, Islam R, Yokota F, Nishikitani M, et al. Blood uric acid prediction with machine learning: model development and performance comparison. JMIR Med Inform 2020 Oct 08;8(10):e18331 [FREE Full text] [doi: 10.2196/18331] [Medline: 33030442]
90. Ferdowsy F, Rahi KS, Jabiullah MI, Habib MT. A machine learning approach for obesity risk prediction. Curr Res Behav Sci 2021 Nov;2:100053. [doi: 10.1016/j.crbeha.2021.100053]
91. Riajuliislam M, Rahim KZ, Mahmud A. Prediction of thyroid disease (hypothyroid) in early stage using feature selection and classification techniques. In: Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Developmen. 2021 Presented at: ICICT4SD '21; February 27-28, 2021; Dhaka, Bangladesh p. 60-64 URL: https://ieeexplore.ieee.org/document/9397052 [doi: 10.1109/icict4sd50815.2021.9397052]
92. Diptu NA, Khan MA, Debnath S, Al Imam A, Rakib AM, Ador KA, et al. Early detection of glaucoma using fuzzy logic in Bangladesh context. In: Proceedings of the 2018 International Conference on Intelligent Systems. 2018 Presented at: IS

XSL•FO
RenderX

'18; September 25-27, 2018; Funchal, Portugal p. 87-93 URL: https://ieeexplore.ieee.org/document/8710490 [doi: 10.1109/is.2018.8710490]
93. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HT, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 03;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]
94. Jimenez G, Spinazze P, Matchar D, Koh Choon Huat G, van der Kleij RM, Chavannes NH, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
95. Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K, Ranmuthu CK, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. Lancet Digit Health 2022 Jun;4(6):e466-e476 [FREE Full text] [doi: 10.1016/S2589-7500(22)00023-1] [Medline: 35623799]
96. Wahl BA, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health 2018;3(4):e000798 [FREE Full text] [doi: 10.1136/bmjgh-2018-000798] [Medline: 30233828]
97. Ahmed A, Agus M, Alzubaidi M, Aziz S, Abd-Alrazaq A, Giannicchi A, et al. Overview of the role of big data in mental health: a scoping review. Comput Methods Programs Biomed Update 2022;2:100076 [FREE Full text] [doi: 10.1016/j.cmpbup.2022.100076]
98. Gaitanou P, Garoufallou E, Balatsoukas P. The effectiveness of big data in health care: a systematic review. In: Proceedings of the 8th Research Conference on Metadata and Semantics Research. 2014 Presented at: MTSR '14; November 27-29, 2014; Karlsruhe, Germany p. 141-153 URL: https://link.springer.com/chapter/10.1007/978-3-319-13674-5_14 [doi: 10.1007/978-3-319-13674-5_14]
99. Brossard PY, Minvielle E, Sicotte C. The path from big data analytics capabilities to value in hospitals: a scoping review. BMC Health Serv Res 2022 Jan 31;22(1):134 [FREE Full text] [doi: 10.1186/s12913-021-07332-0] [Medline: 35101026]
100. Corsi A, de Souza FF, Pagani RN, Kovaleski JL. Big data analytics as a tool for fighting pandemics: a systematic review of literature. J Ambient Intell Humaniz Comput 2021;12(10):9163-9180 [FREE Full text] [doi: 10.1007/s12652-020-02617-4] [Medline: 33144892]
101. Makimoto H, Kohro T. Adopting artificial intelligence in cardiovascular medicine: a scoping review. Hypertens Res 2024 Mar 31;47(3):685-699. [doi: 10.1038/s41440-023-01469-7] [Medline: 37907600]
102. Fisher S, Rosella LC. Priorities for successful use of artificial intelligence by public health organizations: a literature review. BMC Public Health 2022 Nov 22;22(1):2146 [FREE Full text] [doi: 10.1186/s12889-022-14422-z] [Medline: 36419010]
103. van der Schaar M, Alaa AM, Floto A, Gimson A, Scholtes S, Wood A, et al. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. Mach Learn 2021 Dec 09;110(1):1-14 [FREE Full text] [doi: 10.1007/s10994-020-05928-x] [Medline: 33318723]
104. National artificial intelligence policy 2024: draft. Ministry of Posts, Telecommunications and Information Technology. 2024. URL: https://ictd.portal.gov.bd/sites/default/files/files/ictd.portal.gov.bd/page/6c9773a2_7556_4395_bbec_f132b9d819f0/National_AI_Policy_2024_DRAFT.pdf [accessed 2024-04-29]
105. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. Int J Technol Assess Health Care 2012 Apr 26;28(2):138-144. [doi: 10.1017/S0266462312000086] [Medline: 22559755]

Abbreviations

AI: artificial intelligence
BDA: big data analytics
DL: deep learning
LSTM: long short-term memory
ML: machine learning
NCD: noncommunicable disease
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.11.23; peer-reviewed by S Pesälä, J Aarts, E Vashishtha; comments to author 25.04.24; revised version received 14.05.24; accepted 12.09.24; published 28.10.24.

Please cite as:
Alam MA, Sajib MRUZ, Rahman F, Ether S, Hanson M, Sayeed A, Akter E, Nusrat N, Islam TT, Raza S, Tanvir KM, Chisti MJ, Rahman QSU, Hossain A, Layek MA, Zaman A, Rana J, Rahman SM, Arifeen SE, Rahman AE, Ahmed A
Implications of Big Data Analytics, AI, Machine Learning, and Deep Learning in the Health Care System of Bangladesh: Scoping Review
J Med Internet Res 2024;26:e54710
URL: https://www.jmir.org/2024/1/e54710
doi:10.2196/54710
PMID:39466315

©Md Ashraful Alam, Md Refat Uz Zaman Sajib, Fariya Rahman, Saraban Ether, Molly Hanson, Abu Sayeed, Ema Akter, Nowrin Nusrat, Tanjeena Tahrin Islam, Sahar Raza, K M Tanvir, Mohammod Jobayer Chisti, Qazi Sadeq-ur Rahman, Akm Hossain, MA Layek, Asaduz Zaman, Juwel Rana, Syed Moshfiqur Rahman, Shams El Arifeen, Ahmed Ehsanur Rahman, Anisuddin Ahmed. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Experiences of Patients With Cancer Using Electronic Symptom Management Systems: Qualitative Systematic Review and Meta-Synthesis

Siying Zhu[1*], MS; Yan Dong[1*], MS; Yumei Li[2*], BS; Hong Wang[1], MS; Xue Jiang[1], MS; Mingen Guo[1], MS; Tiantian Fan[2], MS; Yalan Song[2], MS; Ying Zhou[1*], PhD; Yuan Han[1*], PhD

[1]School of Nursing, Guangzhou Medical University, Guangzhou, China

[2]Guangzhou Institute of Cancer Research, the Affiliated Cancer Hospital, Guangzhou Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yuan Han, PhD
School of Nursing
Guangzhou Medical University
No. 1, Xinzao
Panyu District
Guangzhou, 511436
China
Phone: 86 020 83079032
Email: hypyx2006@163.com

Abstract

Background: There are numerous symptoms related to cancer and its treatments that can affect the psychosomatic health and quality of life of patients with cancer. The use of electronic symptom management systems (ESMSs) can help patients with cancer monitor and manage their symptoms effectively, improving their health-related outcomes. However, patients' adhesion to ESMSs decreases over time, and little is known about their real experiences with them. Therefore, it is necessary to gain a deep understanding of patients' experiences with ESMSs.

Objective: The purpose of this systematic review was to synthesize qualitative studies on the experiences of patients with cancer using ESMSs.

Methods: A total of 12 electronic databases, including PubMed, Web of Science, Cochrane Library, EBSCOhost, Embase, PsycINFO, ProQuest, Scopus, Wanfang database, CNKI, CBM, and VIP, were searched to collect relevant studies from the earliest available record until January 2, 2024. Qualitative and mixed methods studies published in English or Chinese were included. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement checklist) and the ENTREQ (Enhancing Transparency in Reporting the Synthesis of Qualitative Research) statement were used to improve transparency in reporting the synthesis of the qualitative research. The Critical Appraisal Skills Program (CASP) checklist was used to appraise the methodological quality of the included studies, and a meta-synthesis was conducted to interpret and synthesize the findings.

Results: A total of 21 studies were included in the meta-synthesis. The experiences of patients with cancer using ESMSs were summarized into three major categories: (1) perceptions and attitudes toward ESMSs; (2) the value of ESMSs; and (3) barriers, requirements, and suggestions for ESMSs. Subsequently, 10 subcategories emerged from the 3 major categories. The meta-synthesis revealed that patients with cancer had both positive and negative experiences with ESMSs. In general, patients recognized the value of ESMSs in symptom assessment and management and were willing to use them, but they still encountered barriers and wanted them to be improved.

Conclusions: This systematic review provides implications for developing future ESMSs that improve health-related outcomes for patients with cancer. Future research should focus on strengthening electronic equipment and technical support for ESMSs, improving their functional contents and participation forms, and developing personalized applications tailored to the specific needs and characteristics of patients with cancer.

Trial Registration: PROSPERO CRD42023421730; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=421730

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59061) doi:10.2196/59061

KEYWORDS

electronic symptom management systems; oncology care; access to care; symptom monitoring; self-management; patient-reported outcomes; health-related outcomes; quality of life

Introduction

Background

Globally, cancer incidence and mortality rates have increased rapidly, posing a major public health concern [1]. With the rapid development of cancer treatment methods, survival rates for patients with cancer have increased significantly [2,3]. However, the majority of patients with cancer experience numerous symptoms related to cancer and treatment-related toxicities [4-7], which may result in symptom distress [8,9], financial toxicity [10], prolonged hospitalization, high rates of complications, and even death [11].

Patient-reported outcomes (PROs) are defined as "measurements of any aspect of a patient's health status that come directly from the patient, without any interpretation by a clinician or anyone else" [12]. PROs have been shown to capture symptoms more accurately than physician assessments [13,14]. By using patients' own assessments of the occurrence and severity of symptoms, health care professionals can identify and assess the potential health risks earlier, thereby improving patients' health-related outcomes [15]. In recent years, there has been an explosion of paper versions of PROs, facilitated by a variety of validated instruments [13]. However, in real-world settings, paper scale assessments are prone to data loss and input errors, which make it difficult to guarantee data reliability and add to the burden of data management [16]. Additionally, when patients are not hospitalized, there is a lag in paper-based symptom assessment and management [17]. Given the limitations of paper-based symptom assessments, electronic methods for patients' self-reported symptoms have shown significant promise [18].

Electronic symptom management systems (ESMSs) refer to electronic PRO (ePRO) systems that provide real-time patient assessment and symptom management in oncology practices. Based on patients' responses to symptom assessments, ESMSs can automatically score symptoms and generate warnings. Afterward, health care professionals can receive the data and guide patients, and some ESMSs may also provide evidence-based symptom management recommendations [19-22]. Over the past decades, a growing number of ESMSs have been developed [23-25], and there is equivalence and comparability between electronic and paper-and-pencil symptom assessment measures [26]. Instead of relying on retrospective reporting and delayed manual documentation, ESMSs allow patients to report symptoms via their own electronic devices at home, as well as to potentially document these symptoms automatically in their medical records [27]. Additionally, by providing overviews of symptoms over time, ESMSs can also aid in the early detection and management of symptoms [28] and improve patient-clinician communication as well as patients' quality of life [29-31].

In the past few years, the importance of ESMSs has become increasingly recognized by health care services. However, some studies found that patients' engagement with ESMSs has declined over time [32-35]. Most current systematic reviews of ESMSs focus on intervention effectiveness [17,36] and identifying key mechanisms that improve patients' health-related outcomes [37]. It must be noted that existing evidence fails to fully capture the details of patients' profound experiences, and the underlying reasons for decreased engagement with ESMSs remain unclear. To overcome these deficiencies, systematic reviews of qualitative evidence can facilitate a better understanding of how patients perceive ESMSs. The existing systematic reviews of qualitative studies primarily emphasize patients' experiences with telemedicine [38,39], including not only symptom monitoring and management but also telephone follow-up, digital consultation, virtual simulation, exercise intervention, and so on. However, this comprehensive coverage makes it difficult to probe deeply into patients' specific experiences with ESMSs. Although numerous qualitative studies have been conducted to provide some insight into how patients with cancer perceive ESMSs, systematic reviews of qualitative studies that reflect the specific experiences of patients regarding ESMSs are lacking. Therefore, to better understand patients' experiences with ESMSs, it is necessary to synthesize their experiences systematically.

Objectives

This systematic review aims to synthesize previous qualitative studies on the experiences of patients with cancer using ESMSs, including their perceptions, preferences, barriers, and suggestions. This will contribute to the development of future ESMSs and electronic symptom management interventions tailored to patient needs, increasing the chances of patients reporting their symptoms in a timely manner, thereby resulting in positive outcomes both for patients and health care providers.

Methods

Overview

The protocol was registered in PROSPERO (CRD42023421730) prior to the systematic review. The systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [40] statement checklist and the ENTREQ (Enhancing Transparency in Reporting the Synthesis of Qualitative Research) [41] statement to improve transparency in reporting the synthesis of qualitative research (Multimedia Appendices 1 and 2).

Search Strategy

After assessing the relevant literature reviews, our trained reviewers (authors TF and YS) developed the search strategy and then carefully discussed it with the other reviewers to guarantee a systematic and comprehensive review of the papers.

XSL•FO
RenderX

To develop our search strategy, keywords and derivatives of terms were identified via an initial, limited PubMed search. A combination of MeSH (Medical Subject Headings) terms and free terms were used to ensure all relevant papers were identified. The qualitative literature relevant to the study was searched using systematic electronic databases. The search strategy was developed for PubMed first and then adapted and applied to Web of Science, Cochrane Library, EBSCOhost, Embase, PsycINFO, ProQuest, Scopus, Wanfang database, CNKI, CBM, and VIP. We used the search terms "symptom manage*," "symptom monitor*," "mobile application*," "cell phone," "smartphone," "internet," "telemedicine," "remote consultation," "Computer-Assisted Decision Making," "cancer," "qualitative research," and their MeSH terms. The full search strategy is shown in Multimedia Appendix 3. We searched the electronic databases for eligible studies from database inception to April 25, 2023, and updated it on January 2, 2024. The searches were conducted in English or Chinese. We used EndNote X9 (Clarivate) to upload and store the search results.

Inclusion and Exclusion Criteria

Qualitative studies or qualitative components of mixed methods studies on the experiences of adult patients with cancer using ESMSs were included. Studies in which symptom monitoring and management were telephone-based or using passive monitoring were excluded. Textbox 1 describes the inclusion and exclusion criteria.

Textbox 1. Inclusion criteria and exclusion criteria.

Inclusion criteria

- Patients diagnosed with cancer
- Aged 18 years and older
- Papers reporting on participants' experiences with electronic symptom management systems (ESMSs) used to monitor and/or manage cancer-related symptoms
- Qualitative studies and the qualitative components of mixed methods research

Exclusion criteria

- Gray literature or unpublished peer-reviewed literature
- Published abstracts or conference proceedings
- Literature reviews, systematic reviews, meta-syntheses, etc
- Studies on patients with cancer with incurable illnesses or at the end of their lives
- Primary symptom monitoring and management that was telephone-based (eg, without using any internet-based health tools) or passive monitoring (eg, only using a wearable device)
- Papers published in languages other than English or Chinese

Screening and Data Extraction

Data Screening

All data were imported into Endnote X9 software, and duplicates were removed. Two reviewers (authors SZ and YD) independently screened the papers based on the titles and abstracts. Subsequently, full-text publications that met the inclusion criteria were retrieved and screened by SZ and YD. Disagreements were resolved through discussion with other reviewers (authors YL, HW, XJ, MG, TF, YS, YZ, and YH) to reach a consensus.

Data Extraction

SZ and YD independently extracted the data using a standardized data extraction form (Multimedia Appendix 4). The data extraction focused on identifying specific qualitative results, such as the categories and subcategories related to the phenomenon of interest. For each study, descriptive data included information about the ESMSs, study objectives, methods, analyses, the country/region of study, and the participant demographics. Discrepancies were resolved through joint discussions with the other reviewers.

Quality Appraisal

SZ and YD independently appraised all papers using the Critical Appraisal Skills Program (CASP) [42] criteria. The quality appraisal results were not used as exclusion criteria but helped determine the level of confidence in the findings. We were more concerned with papers that contained depth in data collection and analysis, which could provide valuable insight into participants' experiences with ESMSs.

Meta-Synthesis

The meta-synthesis method was used to synthesize the findings of the included studies. All authors read the included studies to understand the whole research. Three reviewers (authors SZ, YD, and HW) extracted findings that were closely related to our objectives, along with relevant quotations and authors' interpretations. The quotations and interpretations were read and reread by SZ, YD, YL, and HW for coding. The codes were organized into subcategories to form categories. SZ, YD, and YL wrote and continuously refined draft summaries of the categories, and the review team evaluated the appropriateness of the synthesis. Any disagreements were resolved through team discussions.

XSL•FO
RenderX

Results

Summary of the Search Results

Following the search strategy, 1053 papers were found. After the duplicate papers were removed, a total of 644 papers remained. After removing papers with irrelevant titles or abstracts, 62 papers were reviewed in full text. Ultimately, 21 papers were eligible for inclusion in the review. The PRISMA diagram (Figure 1) illustrates the results of the selection process.

Figure 1. Study flow diagram adapted from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. ESMS: electronic symptom management system.

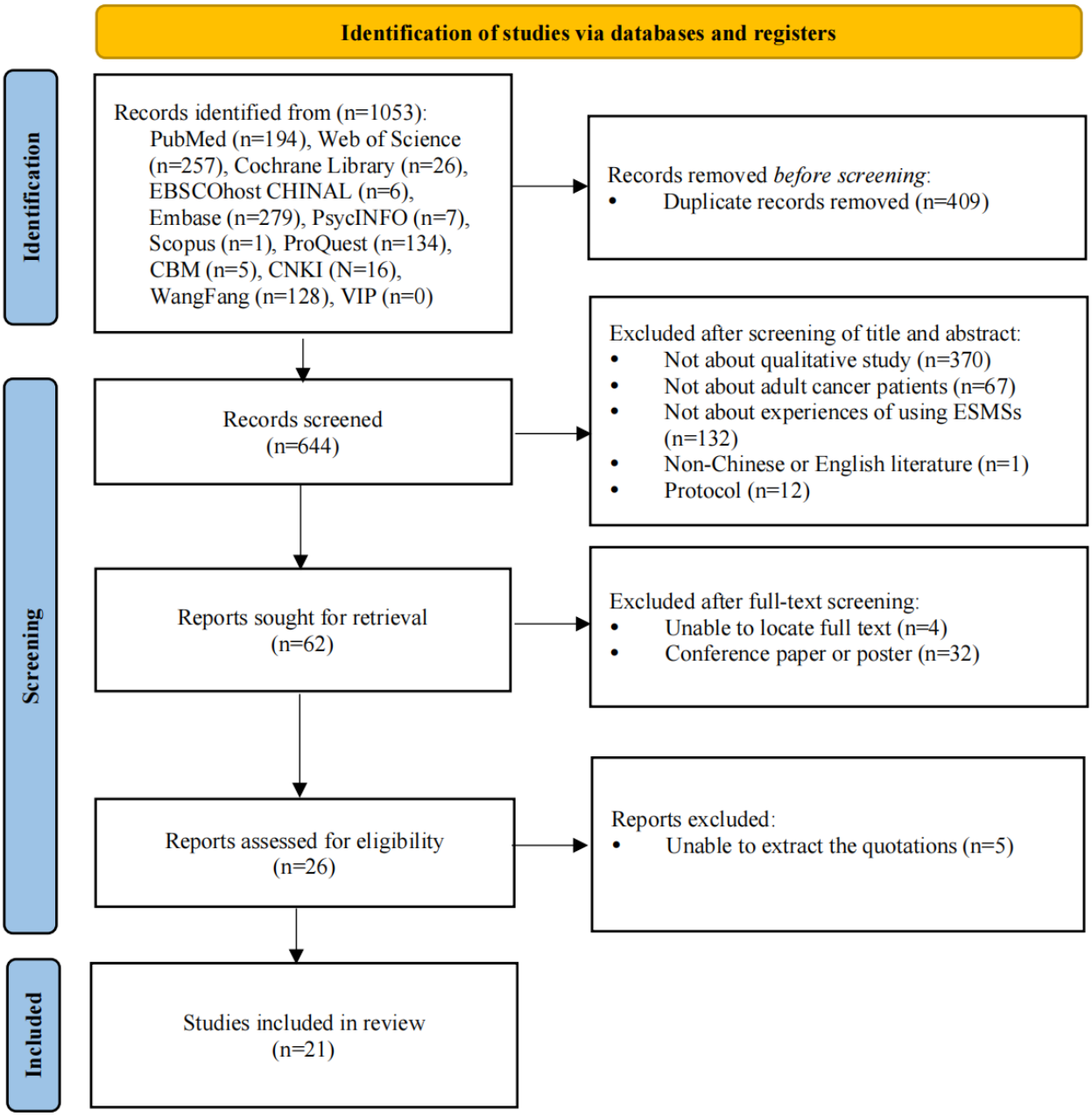

The CASP scores of the papers ranged from 8.5 to 10 (Multimedia Appendix 5), indicating that all 21 papers met an acceptable level of quality and contributed to the meta-synthesis. The CASP results showed that in most included papers (n=20, 95.2%), the relationships between researchers and participants were not elaborated, thereby reducing the quality of the study.

The 21 included papers (Multimedia Appendix 6) were published between 2009 and 2023 and were conducted across diverse geographical settings, including the United Kingdom (n=5, 23.8%) [21,25,43-45], the United States (n=5, 23.8%) [24,46-49], Sweden (n=4, 19%) [23,30,31,50], Australia (n=2, 9.5%) [51,52], China (n=2, 9.5%) [53,54], Ireland (n=1, 4.8%) [20], Canada (n=1, 4.8%) [55], and Norway (n=1, 4.8%) [56]. There were 17 (81 %) mixed methods papers [21,23-25,30,31,43,45-47,49,51-56], while 4 (19%) were purely qualitative [20,44,48,50]. The sample sizes of the qualitative papers ranged from 3 to 131 participants, and 1 (4.8%) paper did not mention sample size [49]. In most (n=20, 95.2%) papers, patients with cancer were middle-aged or older adults. Most (n=15, 71.4%) papers included both male and female participants [20,21,23,25,44-46,48-52,54-56], whereas 2 (9.5%)

XSL•FO
RenderX

papers only reported on males [30,31], 1 (4.8%) only reported on females [24], and 3 (14.3%) did not provide any information about gender at all [43,47,53]. In 5 (23.8%) papers [46,48,49,53,56], the cancer type was not examined, while others focused primarily on gastrointestinal cancer (n=7, 33.3%) [20,43-45,47,51,55], breast cancer (n=4, 19%) [23,24,43,51], prostate cancer (n=3, 14.3%) [23,30,31], lung cancer (n=3, 14.3%) [25,43,54], pancreatic cancer (n=1, 4.8%) [50], chronic myeloid leukemia (n=1, 4.8%) [52], lymphoma (n=1, 4.8%) [55], and malignant pleural mesothelioma (n=1, 4.8%) [21]. A total of 15 (71.4%) papers described treatments given to patients with cancer, including chemotherapy (n=5, 23.8%) [43,47,49,51,55], radiotherapy (n=4, 19%) [24,25,30,31], surgery (n=4, 19%) [44,45,48,50], and oral drugs (n=1, 4.8%) [52]. Moreover, 1 (4.8%) paper included patients receiving neoadjuvant chemotherapy for breast cancer and radiotherapy for prostate cancer [23].

In 8 (38.1%) papers, systematic symptom assessment scales were used in ESMSs, such as the Memorial Symptom Assessment Scale [20,23], Edmonton Symptom Assessment Scale [25], Rotterdam Symptom Checklist [51], Symptom Distress Scale [53], Pain-Intensity Numerical Scale [53], Modified Borg scale [54], Brief Fatigue Inventory [54], Brief Pain Inventory [54], Functional Assessment of Cancer Therapy [20,25,51], and European Organization for Research and Treatment of Cancer [44,45,53]. A total of 5 (23.8%) papers assessed psychological symptoms through ESMSs using the State-Trait Anxiety Inventory [20,25,54], Hospital Anxiety and Depression Scale [51], and Patient Health Questionnaire-Depression Model [53]. In terms of ESMS use time, participants in 15 (71.4%) papers used ESMSs for periods ranging from 10 days to 18 weeks [20,21,23-25,30,31,43-45,47,48,50-52], and most assessed symptoms daily (n=9, 42.9%) [20,21,23,25,30,31,48,50,51]. Additionally, 3 (14.3%) papers were conducted to assess the usability of ESMSs at specific sites [46,55,56], 2 (9.5%) of which required participants to self-report symptoms while waiting in the hospital [53,54], and 1 (4.8%) at the next 2 scheduled chemotherapy appointments [49].

We found that ESMSs varied in the form of symptom management. For the majority of ESMSs mentioned, patients were provided with symptom assessment (n=21, 100%), alerts regarding symptoms were sent to health care professionals (n=12, 57.1%) [21,23-25,30,31,43-45,48,50,55], health care professionals responded to symptoms (n=10, 47.6%) [20,21,23-25,31,43,48,50,55], symptom self-care advices were provided to patients (n=18, 85.7%) [20,21,23,25,30,31,43-48,50-52,54-56], and symptom histories were presented as graphs over time (n=9, 42.9%) [21,23,30,31,44,45,50,51,56]. Furthermore, ESMSs were also capable of allowing patients to create favorites [46], write free text for private health-related information [56], receive daily medication reminders [52], exchange messages with health care professionals [46,56], and share information and experiences with other patients [56].

Results of the Meta-Synthesis

Overview

Participants' experiences with ESMSs were categorized into 3 major categories and 10 subcategories. Multimedia Appendix 7 includes detailed quotations from each of the 3 major categories and 10 subcategories.

Category 1: Perceptions and Attitudes Toward ESMSs

Category Overview

This category demonstrated participants' positive and negative perceptions and attitudes toward ESMSs. The participants reported that ESMSs were easy to use, supported symptom monitoring and management, and became a part of their daily routines, but they also expressed some negative perspectives.

Subcategory 1: Ease of Use

In 13 (61.9%) papers, the participants praised the ease of use of ESMSs [20,21,25,30,43,46-48,50,51,54-56]. In general, participants found ESMSs to be easy to use and navigate [21,25,30,43,46-48]. Participants who were not accustomed to using ESMSs experienced troubles at first, but once they were trained or used them a few times, were able to use them smoothly [50,51,55,56]. A number of participants appreciated how ESMSs asked questions regarding symptoms in lay terms that were easy to understand [20]. Moreover, compared with pen-and-paper formats and face-to-face assessments, the participants found ESMSs more convenient since they could complete the questionnaire at their own pace and correct any incorrect responses themselves [50,54]. Additionally, ESMS also enabled participants to gain a better understanding and mastery of self-management information by providing health information in an accessible format [54].

Subcategory 2: Support Symptom Monitoring and Management

The role of ESMSs in supporting symptom monitoring and management was discussed in 14 (66.7%) papers [20,23,25,31,44-53]. Participants highly appreciated that ESMSs enabled and prompted them to describe symptoms accurately [20,49,50,53]. Some participants believed that ESMSs helped them recall symptoms in general [20,49] and "were good at telling them what's normal or not" [45]. In addition, the participants felt that their symptoms were being tracked by someone [20,46,48,51,52], and the symptoms graphs in ESMSs helped them become aware of and track their symptoms over time [23,44,45,50,51]. Moreover, the participants found ESMSs helpful for reinforcing medical guidance they had received and reminding them of any information they had forgotten [25,44,45,51]. The participants also appreciated that the self-care information provided by ESMSs was specific, appropriate, and achievable, which helped them cope better with their symptoms [20,31,44,45,47,50,51].

Subcategory 3: Symptom Reporting Becomes a Part of Their Daily Routines

As described in 3 (14.3%) papers, the daily requirement of ESMSs to complete the symptom questionnaire became embedded in participants' daily routines [20,30,43]. Although

XSL•FO
RenderX

it took a little time to complete the symptom assessment, this did not negatively impact participants' daily lives, encouraging them to establish a routine.

Subcategory 4: Negative Perspectives

A total of 5 (23.8%) papers reported that participants with negative perspectives regarding ESMSs were less motivated to use them [24,25,30,53,54]. A few participants questioned the clinical relevance of ESMSs [24]. Several participants expressed concern that their symptom reports would be ignored, and if so, they would perceive using ESMSs as a waste of time and worthless [53]. Some participants felt disappointed at not being contacted and doubted the trustworthiness of ESMSs [30]. Some participants considered face-to-face communication to be more reliable and professional than communication through ESMSs [54], and some expressed concern about the additional workload and pressure added to doctors to digest so much information simultaneously [53]. Furthermore, some participants perceived that the self-care information in ESMSs was similar to what their health care providers provided, so they had never or only occasionally read it [25].

Category 2: The Value of ESMSs

Category Overview

As shown in this category, ESMSs were found to be valuable for the participants. The ESMSs helped motivate participants to monitor and manage their symptoms, connect with one another, and communicate effectively, thereby promoting psychological well-being. Despite rarely being mentioned, the participants recognized the potential benefits of ESMSs in alleviating health care burdens.

Subcategory 1: Increasing Motivation for Symptom Monitoring and Management

A total of 9 (42.9%) papers found that ESMSs improved participants' motivation for self-symptom monitoring and management [20,24,30,48-52,54]. ESMSs enhanced the participants' understanding of the causes and effects of symptoms [24,30,49,50,54] and stimulated their consideration of effective methods for symptom prevention and management [20,48,49,51,52]. Moreover, both newly diagnosed and long-term participants reported that using ESMSs improved their adherence by strengthening their sense of accountability for monitoring and managing their symptoms [52].

Subcategory 2: Enhancing Connection and Effective Communication With Others

The benefits of ESMSs in facilitating participants' communication with their health care providers, families, and other patients were reported in 15 (71.4%) papers [20,21,25,30,31,43,44,48-53,55,56]. Participants did not feel that their connection with health care providers was interrupted by using ESMSs, and they were able to maintain a sense of connection after returning home [20,44,50]. Moreover, it was noted that ESMSs' data collection capabilities and alerting mechanisms provided participants with quick access to health care providers [20,25,30,31,43,48,51,55] and could increase their chances of receiving further consultations and treatment [20,21,30,31,44]. Additionally, the ESMSs helped participants prepare for consultations before their visits [49], saving time during medical consultations [53] and facilitating effective communication with health care providers [49,52,53]. The participants also described how ESMS graphs helped them share their feelings and symptoms with their families and friends [30]. Finally, the ESMSs helped facilitate communication and cooperation between patients [56].

Subcategory 3: Gaining Positive Psychological Experiences

A total of 13 (61.9%) papers reported the psychological benefits associated with using ESMSs [20,21,23,25,31,43-45,49-52]. Participants appreciated how ESMSs made them feel reassured, regardless of whether they were experiencing symptoms or not [20,21,23,24,31,43-45,50,51]. The participants also appreciated how ESMSs helped them overcome feelings of uncertainty and concern about symptoms [20,25,45] and reduced anxiety and nervousness [20,44,49]. Interestingly, some participants felt they were being listened to and cared for when using ESMSs [21,52].

Subcategory 4: Potential Benefits for Health Care

As noted in 2 (9.5%) of the reviewed papers [30,43], ESMS use may benefit both patients with cancer and health care professionals. Patients believed that ESMSs would reduce the frequency of their contacting hospitals with symptoms or health-related concerns, thus saving them time spent on phone consultations [43]. Additionally, ESMSs could reduce the burden on the health care system as some patients view them as similar to having medical staff at home [30].

Category 3: Barriers, Requirements and Suggestions for ESMSs

Category Overview

Although most participants reported that ESMSs were easy to operate, several barriers remained. In terms of functionality and content, the participants emphasized the importance of questionnaire items, reminders, alerts, and health information. They also provided suggestions for interface settings based on their needs.

Subcategory 1: Barriers to Using ESMSs

In 6 (28.6%) papers, barriers to using ESMSs were mentioned [24,51-54,56]. Participants who were older or had a lower educational level reported difficulties understanding the questionnaire items and using mobile devices [24,53,54,56]. Participants' health conditions related to cancer treatments are also one of the barriers to using ESMSs, such as being "too tired to follow it through" [51] or worrying about being "too physically or psychologically unwell to participate in the assessment" [53]. Additionally, rural residents, even those who could use mobile devices at home, encountered problems transmitting symptom assessment information to clinical centers due to unstable cellular or internet connectivity [52].

Subcategory 2: Requirements and Suggestions for ESMSs

A total of 12 (57.1%) papers highlighted the requirements and suggestions for ESMSs' symptom assessment, symptom alerts, information push, and interface settings [23,24,30,43-46,48,50,51,53,55]. Participants expressed a desire for improvements in the simplicity and accuracy of the symptom

XSL•FO
RenderX

assessment items [24,30,53], as well as the ability to report additional symptoms [24,43,48]. Moreover, participants with varying disease characteristics had significantly different needs for symptom assessment frequency, ranging from twice a day to once every 6 months [23,24,43,51,53]. Participants also mentioned that they sometimes forgot to log in and record symptoms [51], so some suggested adding an alert to remind them [43]. Furthermore, the option of entering data retrospectively and reediting previously entered data was recommended by participants [51].

While symptom alerts varied from study to study, many participants expressed a desire to decide when or whether to be contacted or to contact health care providers on their own [30,43-45]. Some participants did not consider their symptoms serious enough to warrant consultation [50]. A few even described how they had learned to adjust their responses to avoid being called by nurses [30]. Additionally, some participants did not contact health care providers because their symptoms had already been treated [45] or appointments had already been scheduled [44,45].

Regarding symptom management information push, some participants wanted instant feedback on their symptoms after completing the assessment, as they viewed their symptom results as their personal property [53]. They also hoped that ESMSs would offer better categorized, personalized self-care information [55] and provide different daily messages rather than the same one every day [51]. In addition, having access to health information tailored to their conditions, as well as information on specific topics, was also important to them [51].

In terms of user interface settings, participants reported that they were more likely to use ESMSs with a visually appealing and advanced user interface [46,53,55]. They recommended adding a higher-level menu and a search button to simplify finding information and save time [55]. A larger screen or larger text was considered beneficial for participants with poor eyesight [55]. Additionally, some participants commented that “the screen seemed to be too monotonous” or “clinical” [46,53] and suggested making it more colorful or vivid to attract users’ attention [53,55].

Discussion

Principal Findings

This qualitative systematic review indicated that patients with cancer had both positive and negative experiences with ESMSs. In addition to their ease of use, usefulness, and convenience, patients with cancer benefited from ESMSs physically, psychologically, socially, and economically. However, some patients with cancer were discouraged from using ESMSs due to negative perceptions and barriers. Furthermore, patients with cancer require ESMSs that could be tailored to their specific needs.

This review indicates that ESMSs are useful and acceptable. Consistent with other studies [36,57,58], ESMSs are easy to use, help patients with cancer accurately describe and continuously track their symptoms, provide them with practical self-care information, and do not interfere with their daily lives. Nevertheless, similar to other relevant reviews [59], we found that some patients with cancer had negative experiences with ESMSs, which affected their usage intentions.

Our study indicated that communication with health care providers via mobile phones was not considered more reliable than face-to-face communication. This may be because patients do not want to change their offline medical habits, and they believe using traditional face-to-face forms would help build trust and facilitate better communication [59]. Furthermore, previous studies have shown that the stickiness of ESMSs has decreased over time among patients [32-35], and the reasons behind this decline were explored in our study. Patients with cancer may lose motivation and willingness to continue using ESMSs if they do not see their value, discover that their symptom management information is similar to that provided by their health care providers, or are not provided with timely feedback on their self-reports. According to the Technology Acceptance Model [60], perceived usefulness is one of the key factors affecting users’ attitudes and behavior toward new technologies. In this way, patients with cancer are more likely to use and adhere to ESMSs when they perceive that the ESMSs can aid them with symptom management.

This study further confirms the value of ESMSs in real-world applications. As demonstrated in previous telemedicine studies [38,61], ESMSs can increase patient awareness of and attention to symptoms monitoring and management, facilitate effective communication with health care professionals and other patients, and enhance positive psychological experiences. This study also found long-term potential benefits of ESMSs in reducing health care burdens [30,43]. According to Jibb et al [62], although following up with the health care team after alerts increases intervention costs, it can reduce the rate of emergency department visits, prevent patients from missing the optimal treatment time during diagnosis and treatment queues, and lower overall medical costs. Additionally, Zhang et al [63] found that ESMSs can collect patient symptom information in a timely and continuous manner, thus reducing the cost of manual information collection.

While patients with cancer benefited from the ESMSs, some still encountered barriers to their use. Like other studies regarding technology use in medical settings, patients who were older [59,64], less educated [65,66], and lived in rural areas [67] faced greater challenges with using ESMSs. Older adults tend to have high levels of technology anxiety and even resistance to using technology [68], possibly resulting in a digital divide that hinders their use of ESMSs [69]. Moreover, less educated patients may have difficulty understanding and using the complex functions of ESMSs [38], and patients in rural areas may lack adequate infrastructure and technical support [38]. Due to limited access to network and technological resources, and limited information literacy, mobile medical treatments are challenging for these vulnerable groups [38,70].

This study also provided useful insights into how ESMSs can be further developed and optimized. Patients with cancer have a variety of needs regarding the content, form, and other aspects of ESMSs, and whether these needs are met may affect their adoption and subsequent use [71]. These findings suggest that

XSL•FO
RenderX

ESMSs should be improved in terms of functionality and content design to better meet patients' needs. Consideration should be given to reducing the burden of completing symptom assessments, adjusting symptom alert settings based on whether patients need assistance, and providing more personalized symptom management information. Furthermore, it is imperative to add symptom-filling reminders, modifications, and feedback functions and to create a user-friendly interface.

Strengths and Limitations

To the best of our knowledge, this is the first meta-synthesis focusing on the experiences of patients with cancer who used ESMSs for symptom monitoring and management. This review covered all existing ESMSs regardless of the cancer type, providing a better understanding of how patients with cancer perceive them. Thus, it may provide guidance for future ESMS development and optimization.

Despite the significant insights provided by this review, it faced some limitations. First, this systematic review was limited to articles published in English or Chinese due to the language restrictions of the research group. Although some studies were conducted in non-English speaking areas, they can still result in biased results and be missed. Second, since people with negative attitudes toward ESMSs are likely to be excluded from research studies, this review may not fully reflect all the experiences of patients with cancer. Finally, most of the included studies were from Europe and America and may not be representative of patients with cancer from other regions.

Recommendations for the Future

Based on our meta-synthesis, we have come up with the following recommendations for the future. In the process of using ESMSs, it is important to pay attention to the plight of vulnerable patient groups and provide them with necessary assistance in adapting to this system. Furthermore, future studies should explore what resources patients on different cancer journeys require and prefer rather than simply assuming and prescribing what is deemed beneficial and useful. Additionally, developing ESMSs requires the participation of a multidisciplinary team to bridge the gap between technological innovation and successful service delivery.

Conclusions

This study provides an overview of the qualitative literature on the experiences of patients with cancer using ESMSs. Overall, the review indicated that ESMSs resulted in positive symptom monitoring and management experiences for patients. However, some patients encountered barriers during the use process, so their individual needs and preferences must be carefully considered. Recommendations from this systematic review can be used to improve ESMS development, adoption, and compliance.

Acknowledgments

This study was funded by Guangzhou Education Bureau (Q195003022) and Guangzhou Medical University (Guangyifa 2023 no 16).

Authors' Contributions

SZ, YD, and YL conceptualized and designed the study with input from YZ and YH. TF and YS designed the search strategy. HW, XJ, and MG executed the search strategy. SZ and YD performed article screening, data extraction, and quality appraisal. SZ, YD, YL, and HW performed the meta-synthesis. SY, YD, and YL wrote the first draft of the paper with input from YZ and YH. YZ and YH critically reviewed the manuscript and provided expert input. SZ, YD, and YL revised the manuscript. All authors have read and approved the final version of the manuscript. SZ, YD, and YL contributed equally to this work and should be regarded as co–first authors. YZ and YH contributed equally to this work and should be considered as cocorresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 33 KB - jmir_v26i1e59061_app1.docx]

Multimedia Appendix 2
ENTREQ (Enhancing Transparency in Reporting the Synthesis of Qualitative Research) checklist.
[DOCX File , 24 KB - jmir_v26i1e59061_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 41 KB - jmir_v26i1e59061_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Extracted data form.
[DOCX File , 15 KB - jmir_v26i1e59061_app4.docx]

Multimedia Appendix 5
Critical Appraisal Skills Program (CASP) quality appraisal of the included studies.
[DOCX File , 19 KB - jmir_v26i1e59061_app5.docx]

Multimedia Appendix 6
Sample characteristics of the included studies.
[DOCX File , 28 KB - jmir_v26i1e59061_app6.docx]

Multimedia Appendix 7
Synthesis of the findings.
[DOCX File , 37 KB - jmir_v26i1e59061_app7.docx]

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, et al. Cancer treatment and survivorship statistics, 2022. CA Cancer J Clin 2022 Sep;72(5):409-436 [FREE Full text] [doi: 10.3322/caac.21731] [Medline: 35736631]
3. Zeng H, Chen W, Zheng R, Zhang S, Ji JS, Zou X, et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. Lancet Glob Health 2018 May;6(5):e555-e567 [FREE Full text] [doi: 10.1016/S2214-109X(18)30127-X] [Medline: 29653628]
4. Kalantari E, Kouchaki S, Miaskowski C, Kober K, Barnaghi P. Network analysis to identify symptoms clusters and temporal interconnections in oncology patients. Sci Rep 2022 Oct 12;12(1):17052 [FREE Full text] [doi: 10.1038/s41598-022-21140-4] [Medline: 36224203]
5. So WKW, Law BMH, Ng MSN, He X, Chan DNS, Chan CWH, et al. Symptom clusters experienced by breast cancer patients at various treatment stages: A systematic review. Cancer Med 2021 Apr;10(8):2531-2565 [FREE Full text] [doi: 10.1002/cam4.3794] [Medline: 33749151]
6. Han CJ, Reding K, Cooper BA, Paul SM, Conley YP, Hammer M, et al. Symptom clusters in patients with gastrointestinal cancers using different dimensions of the symptom experience. J Pain Symptom Manage 2019 Aug;58(2):224-234 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.04.035] [Medline: 31077784]
7. Ward Sullivan C, Leutwyler H, Dunn LB, Cooper BA, Paul SM, Conley YP, et al. Differences in symptom clusters identified using symptom occurrence rates versus severity ratings in patients with breast cancer undergoing chemotherapy. Eur J Oncol Nurs 2017 Jun;28:122-132 [FREE Full text] [doi: 10.1016/j.ejon.2017.04.001] [Medline: 28478849]
8. Yeh Y. Symptom distress, stress, and quality of life in the first year of gynaecological cancers: A longitudinal study of women in Taiwan. Eur J Oncol Nurs 2021 Aug;53:101984. [doi: 10.1016/j.ejon.2021.101984] [Medline: 34275743]
9. Wei C, Liang S, Chin C, Lin H, Rosenberg J. Change trajectory of symptom distress, coping strategies, and spiritual wellbeing in colorectal cancer patients undergoing chemotherapy. Healthcare (Basel) 2023 Mar 14;11(6):857 [FREE Full text] [doi: 10.3390/healthcare11060857] [Medline: 36981514]
10. Chan RJ, Gordon LG, Tan CJ, Chan A, Bradford NK, Yates P, et al. Relationships between financial toxicity and symptom burden in cancer survivors: a systematic review. J Pain Symptom Manage 2019 Mar;57(3):646-660.e1 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.12.003] [Medline: 30550833]
11. Pizuorno Machado A, Shatila M, Liu C, Lu Y, Altan M, Glitza Oliva IC, et al. Characteristics, treatment, and outcome of patients with bowel perforation after immune checkpoint inhibitor exposure. J Cancer Res Clin Oncol 2023 Aug;149(9):5989-5998. [doi: 10.1007/s00432-022-04569-y] [Medline: 36611109]
12. U.S. Department of HealthHuman Services FDA Center for Drug EvaluationResearch, U.S. Department of HealthHuman Services FDA Center for Biologics EvaluationResearch, U.S. Department of HealthHuman Services FDA Center for DevicesRadiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006 Oct 11;4:79 [FREE Full text] [doi: 10.1186/1477-7525-4-79] [Medline: 17034633]
13. Gordon BE, Chen RC. Patient-reported outcomes in cancer survivorship. Acta Oncol 2017 Feb;56(2):166-173. [doi: 10.1080/0284186X.2016.1268265] [Medline: 28084867]
14. Falchook AD, Green R, Knowles ME, Amdur RJ, Mendenhall W, Hayes DN, et al. Comparison of patient- and practitioner-reported toxic effects associated with chemoradiotherapy for head and neck cancer. JAMA Otolaryngol Head Neck Surg 2016 Jun 01;142(6):517-523 [FREE Full text] [doi: 10.1001/jamaoto.2016.0656] [Medline: 27149571]

XSL•FO
RenderX

15. Hinami K, Alkhalil A, Chouksey S, Chua J, Trick WE. Clinical significance of physical symptom severity in standardized assessments of patient reported outcomes. Qual Life Res 2016 Sep;25(9):2239-2243. [doi: 10.1007/s11136-016-1261-2] [Medline: 26980417]
16. Aiyegbusi OL. Key methodological considerations for usability testing of electronic patient-reported outcome (ePRO) systems. Qual Life Res 2020 Feb;29(2):325-333 [FREE Full text] [doi: 10.1007/s11136-019-02329-z] [Medline: 31691202]
17. Lai-Kwon J, Cohen JE, Lisy K, Rutherford C, Girgis A, Basch E, et al. The feasibility, acceptability, and effectiveness of electronic patient-reported outcome symptom monitoring for immune checkpoint inhibitor toxicities: a systematic review. JCO Clin Cancer Inform 2023 May;7:e2200185 [FREE Full text] [doi: 10.1200/CCI.22.00185] [Medline: 37220322]
18. Carrasco S, Symes L. Patient use of electronic methods to self-report symptoms: an integrative literature review. Oncol Nurs Forum 2018 May 01;45(3):399-416. [doi: 10.1188/18.ONF.399-416] [Medline: 29683117]
19. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
20. Darley A, Coughlan B, Maguire R, McCann L, Furlong E. A bridge from uncertainty to understanding: The meaning of symptom management digital health technology during cancer treatment. Digit Health 2023;9:20552076231152163 [FREE Full text] [doi: 10.1177/20552076231152163] [Medline: 36714543]
21. Maguire R, Connaghan J, Arber A, Klepacz N, Blyth KG, McPhelim J, et al. Advanced symptom management system for patients with malignant pleural mesothelioma (ASyMSmeso): mixed methods study. J Med Internet Res 2020 Nov 12;22(11):e19180 [FREE Full text] [doi: 10.2196/19180] [Medline: 33180025]
22. Warrington L, Absolom K, Conner M, Kellar I, Clayton B, Ayres M, et al. Electronic systems for patients to report and manage side effects of cancer treatment: systematic review. J Med Internet Res 2019 Jan 24;21(1):e10875 [FREE Full text] [doi: 10.2196/10875] [Medline: 30679145]
23. Crafoord M, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
24. Lapen K, Sabol C, Tin AL, Lynch K, Kassa A, Mabli X, et al. Development and pilot implementation of a remote monitoring system for acute toxicity using electronic patient-reported outcomes for patients undergoing radiation therapy for breast cancer. Int J Radiat Oncol Biol Phys 2021 Nov 15;111(4):979-991 [FREE Full text] [doi: 10.1016/j.ijrobp.2021.07.1692] [Medline: 34314814]
25. Maguire R, Ream E, Richardson A, Connaghan J, Johnston B, Kotronoulas G, et al. Development of a novel remote patient monitoring system: the advanced symptom management system for radiotherapy to improve the symptom experience of patients with lung cancer receiving radiotherapy. Cancer Nurs 2015;38(2):E37-E47. [doi: 10.1097/NCC.0000000000000150] [Medline: 24836956]
26. O'Donohoe P, Reasner DS, Kovacs SM, Byrom B, Eremenco S, Barsdorf AI, et al. Updated recommendations on evidence needed to support measurement comparability among modes of data collection for patient-reported outcome measures: a good practices report of an ISPOR task force. Value Health 2023 May;26(5):623-633 [FREE Full text] [doi: 10.1016/j.jval.2023.01.001] [Medline: 37121630]
27. Basch E. Patient-reported outcomes - harnessing patients' voices to improve clinical care. N Engl J Med 2017 Jan 12;376(2):105-108. [doi: 10.1056/NEJMp1611252] [Medline: 28076708]
28. van Eenbergen MC, van den Hurk C, Mols F, van de Poll-Franse LV. Usability of an online application for reporting the burden of side effects in cancer patients. Support Care Cancer 2019 Sep;27(9):3411-3419. [doi: 10.1007/s00520-019-4639-1] [Medline: 30656400]
29. Suchodolska G, Senkus E. Mobile applications for early breast cancer chemotherapy-related symptoms reporting and management: A scoping review. Cancer Treat Rev 2022 Apr;105:102364 [FREE Full text] [doi: 10.1016/j.ctrv.2022.102364] [Medline: 35231871]
30. Langius-Eklöf A, Christiansen M, Lindström V, Blomberg K, Hälleberg Nyman M, Wengström Y, et al. Adherence to report and patient perception of an interactive app for managing symptoms during radiotherapy for prostate cancer: descriptive study of logged and interview data. JMIR Cancer 2017 Oct 31;3(2):e18 [FREE Full text] [doi: 10.2196/cancer.7599] [Medline: 29089290]
31. Sundberg K, Eklöf AL, Blomberg K, Isaksson A, Wengström Y. Feasibility of an interactive ICT-platform for early assessment and management of patient-reported symptoms during radiotherapy for prostate cancer. Eur J Oncol Nurs 2015 Oct;19(5):523-528. [doi: 10.1016/j.ejon.2015.02.013] [Medline: 25813529]
32. Lehmann J, de Ligt KM, Tipelius S, Giesinger JM, Sztankay M, Voigt S, et al. Adherence to patient-reported symptom monitoring and subsequent clinical interventions for patients with multiple myeloma in outpatient care: longitudinal observational study. J Med Internet Res 2023 Aug 22;25:e46017 [FREE Full text] [doi: 10.2196/46017] [Medline: 37606979]
33. Cherny NI, Parrinello CM, Kwiatkowsky L, Hunnicutt J, Beck T, Schaefer E, et al. Feasibility of large-scale implementation of an electronic patient-reported outcome remote monitoring system for patients on active treatment at a community cancer center. JCO Oncol Pract 2022 Dec;18(12):e1918-e1926 [FREE Full text] [doi: 10.1200/OP.22.00180] [Medline: 36240475]

XSL•FO
RenderX

34. Holch P, Absolom KL, Henry AM, Walker K, Gibson A, Hudson E, et al. Online symptom monitoring during pelvic radiation therapy: randomized pilot trial of the eRAPID intervention. Int J Radiat Oncol Biol Phys 2023 Mar 01;115(3):664-676 [FREE Full text] [doi: 10.1016/j.ijrobp.2022.09.078] [Medline: 36241128]
35. Handa S, Okuyama H, Yamamoto H, Nakamura S, Kato Y. Effectiveness of a smartphone application as a support tool for patients undergoing breast cancer chemotherapy: a randomized controlled trial. Clin Breast Cancer 2020 Jun;20(3):201-208. [doi: 10.1016/j.clbc.2020.01.004] [Medline: 32201165]
36. Li Y, Li J, Hu X. The effectiveness of symptom management interventions based on electronic patient-reported outcomes (ePROs) for symptom burden, quality of life, and overall survival among patients with cancer: A meta-analysis of randomized controlled trials. Int J Nurs Stud 2023 Nov;147:104588. [doi: 10.1016/j.ijnurstu.2023.104588] [Medline: 37690275]
37. Govindaraj R, Agar M, Currow D, Luckett T. Assessing patient-reported outcomes in routine cancer clinical care using electronic administration and telehealth technologies: realist synthesis of potential mechanisms for improving health outcomes. J Med Internet Res 2023 Nov 28;25:e48483 [FREE Full text] [doi: 10.2196/48483] [Medline: 38015606]
38. Ren L, Chen M, Jiang H, Wang Y, Xia L, Dong C. Perceptions of adult patients with cancer towards telemedicine: A qualitative meta-synthesis. Asia Pac J Oncol Nurs 2024 Feb;11(2):100360 [FREE Full text] [doi: 10.1016/j.apjon.2023.100360] [Medline: 38293602]
39. Meneses ADFP, Pimentel FF, da Cruz JPF, Candido Dos Reis FJ. Experiences of women with breast cancer using telehealth: a qualitative systematic review. Clin Breast Cancer 2023 Feb;23(2):101-107. [doi: 10.1016/j.clbc.2022.11.001] [Medline: 36464603]
40. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
41. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012 Nov 27;12:181 [FREE Full text] [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
42. CASP qualitative checklist. Critical Appraisal Skills Programme. URL: https://casp-uk.net/casp-checklists/CASP-checklist-qualitative-2024.pdf [accessed 2024-10-12]
43. McCann L, Maguire R, Miller M, Kearney N. Patients' perceptions and experiences of using a mobile phone-based advanced symptom management system (ASyMS) to monitor and manage chemotherapy related toxicity. Eur J Cancer Care (Engl) 2009 Mar;18(2):156-164. [doi: 10.1111/j.1365-2354.2008.00938.x] [Medline: 19267731]
44. Richards HS, Portal A, Absolom K, Blazeby JM, Velikova G, Avery KNL. Patient experiences of an electronic PRO tailored feedback system for symptom management following upper gastrointestinal cancer surgery. Qual Life Res 2021 Nov;30(11):3229-3239 [FREE Full text] [doi: 10.1007/s11136-020-02539-w] [Medline: 32535864]
45. Richards HS, Blazeby JM, Portal A, Harding R, Reed T, Lander T, et al. A real-time electronic symptom monitoring system for patients after discharge following surgery: a pilot study in cancer-related surgery. BMC Cancer 2020 Jun 10;20(1):543 [FREE Full text] [doi: 10.1186/s12885-020-07027-5] [Medline: 32522163]
46. Lattie EG, Bass M, Garcia SF, Phillips SM, Moreno PI, Flores AM, et al. Optimizing health information technologies for symptom management in cancer patients and survivors: usability evaluation. JMIR Form Res 2020 Sep 21;4(9):e18412 [FREE Full text] [doi: 10.2196/18412] [Medline: 32955450]
47. Gomaa S, Posey J, Bashir B, Basu Mallick A, Vanderklok E, Schnoll M, et al. Feasibility of a text messaging-integrated and chatbot-interfaced self-management program for symptom control in patients with gastrointestinal cancer undergoing chemotherapy: pilot mixed methods study. JMIR Form Res 2023 Nov 10;7:e46128 [FREE Full text] [doi: 10.2196/46128] [Medline: 37948108]
48. McCready TM, Stabile C, Vickers A, Ancker JS, Pusic A, Temple LKF, et al. A remote symptom monitoring tool as part of ambulatory cancer surgery recovery: a qualitative analysis of patient experience. JCO Oncol Pract 2023 Aug;19(8):595-601. [doi: 10.1200/OP.23.00017] [Medline: 37235840]
49. Erickson JM, Ameringer S, Linder L, Macpherson CF, Elswick RK, Luebke JM, et al. Using a heuristic app to improve symptom self-management in adolescents and young adults with cancer. J Adolesc Young Adult Oncol 2019 Apr;8(2):131-141. [doi: 10.1089/jayao.2018.0103] [Medline: 30358473]
50. Gustavell T, Langius-Eklöf A, Wengström Y, Segersvärd R, Sundberg K. Development and feasibility of an interactive smartphone app for early assessment and management of symptoms following pancreaticoduodenectomy. Cancer Nurs 2019;42(3):E1-E10. [doi: 10.1097/NCC.0000000000000584] [Medline: 29596113]
51. Whitehead L, Emery L, Kirk D, Twigg D, Brown D, Dewar J. Evaluation of a remote symptom assessment and management (SAM) system for people receiving adjuvant chemotherapy for breast or colorectal cancer: mixed methods study. JMIR Cancer 2020 Dec 07;6(2):e22825 [FREE Full text] [doi: 10.2196/22825] [Medline: 33284122]
52. Pereira-Salgado A, Westwood JA, Russell L, Ugalde A, Ortlepp B, Seymour JF, et al. Mobile health intervention to increase oral cancer therapy adherence in patients with chronic myeloid leukemia (The REMIND system): clinical feasibility and acceptability assessment. JMIR Mhealth Uhealth 2017 Dec 06;5(12):e184 [FREE Full text] [doi: 10.2196/mhealth.8349] [Medline: 29212628]

XSL•FO
RenderX

53. Chan CWH, Tam W, Cheng KKF, Chui YY, So WKW, Mok T, et al. Piloting electronic self report symptom assessment - Cancer (ESRA-C) in Hong Kong: a mixed method approach. Eur J Oncol Nurs 2011 Sep;15(4):325-334. [doi: 10.1016/j.ejon.2010.10.002] [Medline: 21093371]
54. Tang FWK, Chan CWH, Choy Y, Loong HHF, Chow KM, So WKW. A feasibility study on using tablet personal computers for self-reported symptom assessment in newly diagnosed lung cancer patients. Int J Nurs Pract 2018 Aug;24(4):e12658. [doi: 10.1111/ijn.12658] [Medline: 29642280]
55. Moradian S, Krzyzanowska MK, Maguire R, Morita PP, Kukreti V, Avery J, et al. Usability evaluation of a mobile phone-based system for remote monitoring and management of chemotherapy-related side effects in cancer patients: mixed-methods study. JMIR Cancer 2018 Dec 21;4(2):e10932 [FREE Full text] [doi: 10.2196/10932] [Medline: 30578238]
56. Mirkovic J, Kaufman DR, Ruland CM. Supporting cancer patients in illness management: usability evaluation of a mobile app. JMIR Mhealth Uhealth 2014 Aug 13;2(3):e33 [FREE Full text] [doi: 10.2196/mhealth.3359] [Medline: 25119490]
57. Kim SH, Sung JH, Yoo S, Kim S, Lee K, Oh EG, et al. Effects of digital self-management symptom interventions on symptom outcomes in adult cancer patients: A systematic review and meta-analysis. Eur J Oncol Nurs 2023 Oct;66:102404. [doi: 10.1016/j.ejon.2023.102404] [Medline: 37517339]
58. Cho Y, Zhang H, Harris MR, Gong Y, Smith EL, Jiang Y. Acceptance and use of home-based electronic symptom self-reporting systems in patients with cancer: systematic review. J Med Internet Res 2021 Mar 12;23(3):e24638 [FREE Full text] [doi: 10.2196/24638] [Medline: 33709929]
59. Walker RC, Tong A, Howard K, Palmer SC. Patient expectations and experiences of remote monitoring for chronic diseases: Systematic review and thematic synthesis of qualitative studies. Int J Med Inform 2019 Apr;124:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.01.013] [Medline: 30784430]
60. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263:64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]
61. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019 Jul 10;7(7):e13817 [FREE Full text] [doi: 10.2196/13817] [Medline: 31293246]
62. Jibb LA, Stevens BJ, Nathan PC, Seto E, Cafazzo JA, Johnston DL, et al. Implementation and preliminary effectiveness of a real-time pain management smartphone app for adolescents with cancer: A multicenter pilot clinical study. Pediatr Blood Cancer 2017 Oct;64(10):e26554. [doi: 10.1002/pbc.26554] [Medline: 28423223]
63. Zhang L, Zhang X, Shen L, Zhu D, Ma S, Cong L. Efficiency of electronic health record assessment of patient-reported outcomes after cancer immunotherapy: a randomized clinical trial. JAMA Netw Open 2022 Mar 01;5(3):e224427 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.4427] [Medline: 35357459]
64. Husebø AML. Stakeholders' perspectives on eHealth support in colorectal cancer survivorship: qualitative interview study. JMIR Cancer 2021 Sep 07;7(3):e28279 [FREE Full text] [doi: 10.2196/28279] [Medline: 34491210]
65. Almoajel A, Alshamrani S, Alyabsi M. The relationship between e-health literacy and breast cancer literacy among Saudi women. Front Public Health 2022;10:841102 [FREE Full text] [doi: 10.3389/fpubh.2022.841102] [Medline: 35462821]
66. Zhou J, Wang C. Improving cancer survivors' e-health literacy via online health communities (OHCs): a social support perspective. J Cancer Surviv 2020 Apr;14(2):244-252. [doi: 10.1007/s11764-019-00833-2] [Medline: 31820215]
67. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in Access to and Use of Electronic Personal Health Information Between Rural and Urban Residents in the United States. J Rural Health 2018 Feb;34 Suppl 1(Suppl 1):s30-s38 [FREE Full text] [doi: 10.1111/jrh.12228] [Medline: 28075508]
68. Hamid S, Faith F, Jaafar Z, Abdul Ghani N, Yusop FD. Emerging technology for healthy lifestyle of the middle-age and elderly: a scoping review. Iran J Public Health 2023 Feb;52(2):230-242 [FREE Full text] [doi: 10.18502/ijph.v52i2.11877] [Medline: 37089153]
69. Zhang D, Zhang G, Jiao Y, Wang Y, Wang P. "Digital dividend" or "digital divide": what role does the internet play in the health inequalities among Chinese residents? Int J Environ Res Public Health 2022 Nov 17;19(22):15162 [FREE Full text] [doi: 10.3390/ijerph192215162] [Medline: 36429878]
70. Kyaw TL, Ng N, Theocharaki M, Wennberg P, Sahlen K. Cost-effectiveness of digital tools for behavior change interventions among people with chronic diseases: systematic review. Interact J Med Res 2023 Feb 16;12:e42396 [FREE Full text] [doi: 10.2196/42396] [Medline: 36795470]
71. He X, Zheng X, Ding H. Existing barriers faced by and future design recommendations for direct-to-consumer health care artificial intelligence apps: scoping review. J Med Internet Res 2023 Dec 18;25:e50342 [FREE Full text] [doi: 10.2196/50342] [Medline: 38109173]

Abbreviations

CASP: Critical Appraisal Skills Program
ENTREQ: Enhancing Transparency in Reporting the Synthesis of Qualitative Research
ePRO: electronic patient-report outcome
ESMS: electronic symptom management system
MeSH: Medical Subject Headings

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRO: patient-reported outcome

Edited by N Cahill; submitted 01.04.24; peer-reviewed by J Johnson, J Dean; comments to author 20.05.24; revised version received 02.06.24; accepted 11.09.24; published 28.10.24.

Please cite as:
Zhu S, Dong Y, Li Y, Wang H, Jiang X, Guo M, Fan T, Song Y, Zhou Y, Han Y
Experiences of Patients With Cancer Using Electronic Symptom Management Systems: Qualitative Systematic Review and Meta-Synthesis
J Med Internet Res 2024;26:e59061
URL: https://www.jmir.org/2024/1/e59061
doi:10.2196/59061
PMID:39466301

©Siying Zhu, Yan Dong, Yumei Li, Hong Wang, Xue Jiang, Mingen Guo, Tiantian Fan, Yalan Song, Ying Zhou, Yuan Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Mental Health Interventions for Adolescents in Low- and Middle-Income Countries: Scoping Review

Carolina Wani[1], MSc; Lisa McCann[1], MSc, PhD; Marilyn Lennon[1], BSc, PhD; Caterina Radu[1], MSc

Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom

Corresponding Author:
Carolina Wani, MSc
Department of Computer and Information Sciences
University of Strathclyde
16 Richmond Street
Glasgow, G1 1XH
United Kingdom
Phone: 44 0141 548 3189
Email: carolina.wani@strath.ac.uk

Abstract

Background: Digital mental health interventions (DMHIs) are increasingly recognized as potential solutions for adolescent mental health, particularly in low- and middle-income countries (LMICs). The United Nations' Sustainable Development Goals and universal health coverage are instrumental tools for achieving mental health for all. Within this context, understanding the design, evaluation, as well as the barriers and facilitators impacting adolescent engagement with mental health care through DMHIs is essential.

Objective: This scoping review aims to provide insights into the current landscape of DMHIs for adolescents in LMICs.

Methods: The Joanna Briggs Institute scoping review methodology was used, following the recommendations of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews). Our search strategy incorporated 3 key concepts: population "adolescents," concept "digital mental health interventions," and context "LMICs." We adapted this strategy for various databases, including ACM Digital Library, APA PsycINFO, Cochrane Library, Google Scholar (including gray literature), IEEE Xplore, ProQuest, PubMed (NLM), ScienceDirect, Scopus, and Web of Science. The articles were screened against a specific eligibility criterion from January 2019 to March 2024.

Results: We analyzed 20 papers focusing on DMHIs for various mental health conditions among adolescents, such as depression, well-being, anxiety, stigma, self-harm, and suicide ideation. These interventions were delivered in diverse formats, including group delivery and self-guided interventions, with support from mental health professionals or involving lay professionals. The study designs and evaluation encompassed a range of methodologies, including randomized controlled trials, mixed methods studies, and feasibility studies.

Conclusions: While there have been notable advancements in DMHIs for adolescents in LMICs, the research base remains limited. Significant knowledge gaps persist regarding the long-term clinical benefits, the maturity and readiness of LMIC digital infrastructure, cultural appropriateness, and cost-effectiveness across the heterogeneous LMIC settings. Addressing these gaps necessitates large-scale, co-designed, and culturally sensitive DMHI trials. Future work should address this.

(J Med Internet Res 2024;26:e51376) doi:10.2196/51376

KEYWORDS

adolescents; mental health; low- and middle-income countries; LMICs; digital mental health interventions; DMHIs; cultural appropriateness; implementation; design; evaluation; mobile phone

Introduction

Background

The World Health Organization (WHO) defines mental health as "a state of mental well-being that enables people to cope with the stresses of life, to realize their abilities, to learn well and work well, and to contribute to their communities. Mental health is an integral component of health and well-being and is more than the absence of mental disorder" [1]. The United Nations General Assembly has underscored the key role of universal health coverage (UHC)—where everyone can access the health

XSL•FO
RenderX

services they need without financial hardship—in achieving health for all [1]. Indeed, UHC is central to the health-related Sustainable Development Goals (SDGs), namely target 3.4: "by 2030, reduce by one-third premature mortality from noncommunicable diseases through prevention and treatment and promote mental health and well-being" [1]. While these global agendas are instruments for achieving mental health for all, there is international recognition that adolescent mental health is a distinct and critical public health and human rights issue [2-4].

In 2020, the WHO–UNICEF (United Nations International Children's Emergency Fund)–Lancet Commission called for a renewed focus on the SDGs for advancing child and adolescent health, including mental health [5]. The Commission made a powerful argument that children and adolescents should be at the center of the SDGs to protect their human rights. In addition, this was an ethical and economic investment [5]. Despite this, the *World Mental Health Report* states that approximately 90 WHO member states—less than half of all member states—had a mental health policy or plan specifically for children and adolescents [1]. The report emphasizes the importance of national mental health plans that are fully compliant with human rights instruments, sufficiently resourced, and regularly monitored and evaluated [1].

The onset of mental health disorders typically occurs during adolescence (between the ages of 10-19 years), an important time for developing social and emotional skills [6,7]. It is here where adolescents form coping strategies that enable mental health. Equally, it is a time when young people become vulnerable to risk-taking behaviors, for example, substance abuse. Indeed, suicide is the leading cause of adolescent death [6]. Worldwide, 13% of the world's adolescents (aged 10-19 years) live with a mental disorder [8]; however, widening treatment gaps mean many conditions remain undiagnosed and untreated [6]. Moreover, mental health conditions disproportionately affect adolescents in low- and middle-income countries (LMICs) [1]. LMIC economies are those in which the 2022 gross national income per capita was less than US $13,845 [9]. Approximately 90% of the world's 1.2 billion adolescents reside in LMICs [8]. This population is more vulnerable to human rights violations, and where limited mental health services are available, stigma, discrimination, and social, cultural, and economic challenges are major barriers to treatment access [6,8,10].

The COVID-19 pandemic has worsened the significant burden on this population [8,11,12]. Before the pandemic, researchers explored the feasibility and clinical benefits of digital health technologies for adolescent mental health and care in LMICs [13-15]. Indeed, COVID-19 has driven the adoption of digital health care platforms in LMICs, including digital interventions for health [8,16]. Digital health interventions (DHIs) are increasingly accepted tools for the support of adolescent mental health [17-19].

The WHO defines DHIs as a discrete functionality of digital technology that is applied to achieve health objectives [20]. In this regard, digital mental health interventions (DMHIs) can be understood as a discrete functionality of digital technology that is applied to achieve mental health objectives. DHIs are classified into 4 overarching groups based on the primary user, for instance, clients (service users), health care providers (health service delivery), health system managers (administration and oversight of public health systems), and data services (data collection, management, use, and exchange) [20]. For clients, DHIs/DMHIs can be delivered at an individual or population level; in high and low-income settings; and via several devices, for example, mobile apps, websites, wearables, and smart devices. Such interventions may be self-guided [21] or delivered with lay support, for instance, from teachers [22] or trained psychologists [23]. Furthermore, some DMHIs include theories of behavior change, such as the self-determination theory [24] or the Social Cognitive Theory [23]. Indeed, research suggests that interventions grounded in behavioral theory are more likely to be effective [25,26].

Current Barriers to Existing Mental Health Services in LMICs

Presently, there are significant and systemic barriers that limit adolescent access to mental health services, for example, cost, time, geographic location [27], stigma [28,29], lack of community knowledge and education on mental health [30], and low mental health literacy [31]. Mental health systems in LMICs lack trained mental health specialists, leading to significantly understaffed services [30,32,33]. These shortages are unequal, with urban areas benefiting from more health specialists and resources than rural areas [34,35]. Consequently, this is a potential barrier to treatment because the physical distance may place many services out of reach, thereby reinforcing the rural-urban digital divide [36]. The term digital divide can be understood as "the gap between people who do and do not have access to forms of information and communication technology" [37]. The concept is complex, going beyond the connected versus the unconnected. It is a global phenomenon but even more pronounced in LMIC settings, leading to further social, cultural, and economic inequalities [38,39].

Key Challenges With Adolescent DMHIs in LMICs

Adolescents in LMICs have a distinct disadvantage to digital infrastructure, namely the internet [38]. In 2019, the rural-urban gap in mobile internet use across LMICs was 37%; however, LMICs in sub-Saharan Africa had the widest rural-urban gap, with those living in rural areas being 60% less likely to access the internet than those residing in urban areas [40]. Another report found gaps in literacy and related skills had the greatest impact on mobile internet use [41]. Moreover, the GSM (Global System for Mobile) Communications reports that mobile phone affordability and lack of digital skills were significant barriers in LMICs [42].

There are issues with DMHI implementation, posing a barrier to scalability. Many DMHIs, for example, require additional guidance for adolescent adherence and engagement, such as psychologists, therapists, or lay teacher support [43,44]. This may undermine the scalability of the interventions due to the shortage of health professionals for those roles [44]. Other challenges relate to their feasibility [45], political nature [46], cultural appropriateness [47], design, deployment, evaluation

XSL•FO
RenderX

[48], sustainability [49], cost-effectiveness [50], privacy and data security [51], digital maturity, and readiness [52-54]. Despite this, there is evidence that DMHIs could transform adolescent mental health in low-resource settings [50,55-57].

Rationale and Review Question

We selected the review period 2019-2024 for 3 distinct reasons; first, in 2020, the WHO launched its first guidance on designing DHIs with and for young people, recognizing the significance of youth-centered DHIs and considering their specific needs [58]. Second, that same year, the WHO–UNICEF–Lancet Commission called for a renewed focus on the SDGs for advancing child and adolescent health, including mental health, prioritizing young people in the urgent call to action [5]. Finally, the COVID-19 pandemic has profoundly impacted adolescent mental health worldwide, leading to an increased interest and investment in the delivery of quality, person-centered, remote mental health care [59]. The post–COVID-19 era has provided an opportunity to explore which DMHIs exist for adolescents in LMICs, how they are designed and evaluated, how adolescents are involved in design activities, which theories or models were consulted, and which factors affect adolescent engagement. Examining the current state of adolescent DMHIs in the context of these recent developments allows for a timely and up-to-date insight into gaps in policy, practice, and research. This scoping review aims to collate information on adolescent DMHIs in LIMC settings. The review asks “What is known about DMHIs for adolescents in LMICs, as reported in the literature?”

Subreview Questions

The subreview questions (subRQs) were as follows:

- SubRQ1: How are adolescent DMHIs designed and evaluated within LMICs?
- SubRQ2: What are the reported activities and approaches involving adolescents in designing DMHIs, and what are the perceived outcomes of such involvement?
- SubRQ3: Which frameworks, toolkits, models, or theories were consulted or applied to the DMHI? (including implementation activities related to digital health readiness and preparedness)
- SubRQ4: What factors facilitate or hinder adolescent engagement in DMHIs in LMICs?

Methods

Overview

This scoping review followed the Joanna Briggs Institute Scoping Review Methodology [60] and is reported based on the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [61]. The PRISMA-ScR checklist is shown in Multimedia Appendix 1. Scoping reviews are ideal for disciplines with emerging evidence, for example, DMHIs for adolescents in LMICs, where the limited availability of randomized controlled trials (RCTs) hinders researchers from performing systematic reviews and assessing the quality of evidence [62]. Consistent with the purpose of scoping studies, this review does not seek to assess or appraise the quality and robustness of the evidence, nor does it generalize the findings [61,63,64]. By systematically mapping the available literature, a scoping review can highlight the current state of knowledge, identify research gaps, and potentially reveal patterns or trends that can inform future studies [60]. Previous researchers have used scoping reviews to map available evidence in a similar field [65-67].

Eligibility Criteria

The Joanna Briggs Institute recommends the use of the PCC mnemonic, that is, population, concept, and context, to develop the inclusion criteria [60]. The population consists of adolescents aged 10-19 years, which is consistent with the WHO definition [6]. The concept under study is DMHIs, understood as a discrete functionality of digital technology that is applied to achieve mental health objectives [20]. The context of interest is LMIC economies, defined by the World Bank as nations with a gross national income per capita less than US $13,845 [9], see Textbox 1.

XSL•FO
RenderX

Textbox 1. Eligibility criteria based on the participant, concept, and context framework for Joanna Briggs Institute scoping reviews.

Inclusion criteria

1. Population: Adolescents (World Health Organization [WHO] definition [6]—persons between the ages of 10 and 19 years).
2. Concept: Digital mental health interventions (DMHIs) [20] that are delivered to support the mental health and well-being of adolescents in low- and middle-income countries (LMICs). DMHIs focus on emotional, behavioral, or eating disorders, psychosis, suicide, self-harm, and risk-taking behaviors. Including mental health promotion and prevention interventions and those for early detection and treatment. DMHIs include mobile apps, web applications, smart devices, telehealth, SMS text messaging, and internet-based interventions.
3. Context: Low- and middle-income settings [9].
4. Sources: Published, peer-reviewed, or gray literature of any research study design (eg, randomized controlled trials, systematic reviews, case studies, qualitative, quantitative, or mixed methods, theses or dissertations, white papers, guidelines, conference proceedings, charity reports, or posters).
5. Period: Published between January 2019 and March 2024.
6. Language: Studies published in English.

Exclusion criteria

1. Population: Participants that do not meet the WHO definition of adolescents, that is, are aged younger than 10 years or older than 19 years.
2. Concept: DMHIs that target or only explore the concerns and views of adolescents' parents, caregivers, guardians, teachers, and clinicians. Studies that do not focus on digital mental health or well-being.
3. Context: Not based in low- and middle-income settings.
4. Sources: No full text is available or the paper is not retrievable.
5. Period: Published before January 2019.
6. Language: Not available in English.

Information Sources

The information sources were the following electronic databases: (1) ACM Digital Library, (2) APA PsycINFO, (3) Cochrane Library, (4) Google Scholar (gray literature included, Multimedia Appendix 2), (5) IEEE Xplore, (6) ProQuest, (7) PubMed (NLM), (8) ScienceDirect, (9) Scopus, and (10) Web of Science.

These databases were selected because they collectively align with this review's scope on DHIs and technologies, that is, behavioral science, mental health (psychiatry and psychology), and technology.

Search Strategy

The database identification and search strategy were developed with guidance from a faculty librarian at the University of Strathclyde. The search terms represented the primary concepts of the objectives in the review. These included a range of keywords, free text, and medical subject headings terms and combinations of the Boolean operators. The Joanna Briggs Institute guidelines [60] for scoping reviews recommend a 3-step strategy: (1) an initial search of the database (titles and abstracts) using medical subject heading terms, (2) extending the search query to other databases and adjusting the search strategy for each database, and (3) reviewing the reference list of the selected papers.

Databases were searched from January 2019 to March 2024. This period was selected to capture key developments, for example, the WHO's first digital health guidance for youth [58], the renewed focus on adolescents through the SDGs [5], and the COVID-19 pandemic's exposure of deep and preexisting mental health inequalities among adolescent LMICs [11,12]. Collectively, these highlighted the need for a timely review of adolescent DMHIs in low-resource countries. The search strings and results for each database are presented in Multimedia Appendix 3. The reference lists of all related reviews and included studies were hand-searched for additional papers.

Screening

In total, 1 reviewer (CW) initially searched the databases, focusing on the first 200-300 results per database to manage the workload [68]. The most relevant papers appeared at the top of the search results. CW exported all relevant records to EndNote (Clarivate Analytics) to remove duplicate references. CW then imported the remaining records to Rayyan (Rayyan Systems Inc) for title, abstract, and full-text screening. Rayyan places all imported papers into an "undecided" category, allowing reviewers to independently "exclude," "maybe," or "include" each paper based on the eligibility criteria. The first screening involved only the title and abstract review (CW, LM, and CR). During the second screening, CW retrieved the full text of papers that potentially met the eligibility criteria. Where a study had a published protocol and study outcome paper, only the study outcome paper was included in the review. The relevant papers were identified by CW, LM, and CR and placed in the "include" category in the final screening step. Any discrepancies were discussed and resolved by consensus. All authors agreed that the included papers fully met the eligibility criteria.

Data Charting and Data Items

The authors CW, LM, and ML developed and piloted a data charting form at the protocol stage. This was further refined

XSL•FO
RenderX

iteratively at the review stage, and 2 reviewers (CW and CR) extracted the data (CW extracted the data and CR independently verified it). This approach is appropriate where it is not feasible for reviewers to independently chart the data [60]. Any inconsistencies were discussed and resolved. The data items extracted for each article aligned with the review aims and objectives and included the following:

1. Bibliographic information: lead author, date of publication, and country of lead author.
2. Data items related to subRQ1: DMHI name, description, target mental health disorder, start year of intervention implementation, country of intervention, participant age, sample size, study aims, intervention type, control, evaluation, outcome measures, duration, follow-up, and costs.
3. Data items related to subRQ2: adolescent involvement in DMHI design.
4. Data items related to subRQ3: consulted frameworks, theories, or models.
5. Data items related to subRQ4: factors affecting adolescent engagement with the DMHI (facilitators or barriers).
6. Key findings.

The data charting document can be found in Multimedia Appendix 4.

Collating, Summarizing, Synthesizing, and Reporting the Results

Selected papers were reviewed to identify the similarities and differences between the DMHIs. They were then summarized and categorized into themes related to the review questions.

Results

Overview

Initial searches of 10 databases by CW in March 2024 yielded 2226 articles. Of these, 1291 (57.99%) duplicates were removed in EndNote. The remaining 935 papers were exported to Rayyan for title, abstract, and full-text screening. During the first round of screening, 80 more duplicates were removed, leaving 855 papers. CW identified 25 additional articles by searching the reference lists of excluded reviews and articles selected for inclusion. At title and abstract screening, CW reviewed 100% (880/880) of the records, and 2 reviewers (LM and CR) independently screened 25% (220/880) each, ensuring the records fulfilled the inclusion criteria. LM and CR achieved a 99% alignment score with CW's screened papers. The detailed selection process of the articles is presented in the PRISMA-ScR flow diagram (Figure 1).

Figure 1. The PRISMA-ScR flowchart of the searches of 10 databases. ACM: Association for Computing Machinery; APA: American Psychological Association; IEEE: Institute of Electrical and Electronics Engineers; NLM: National Library of Medicine; LMIC: low- and middle-income country; PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews.

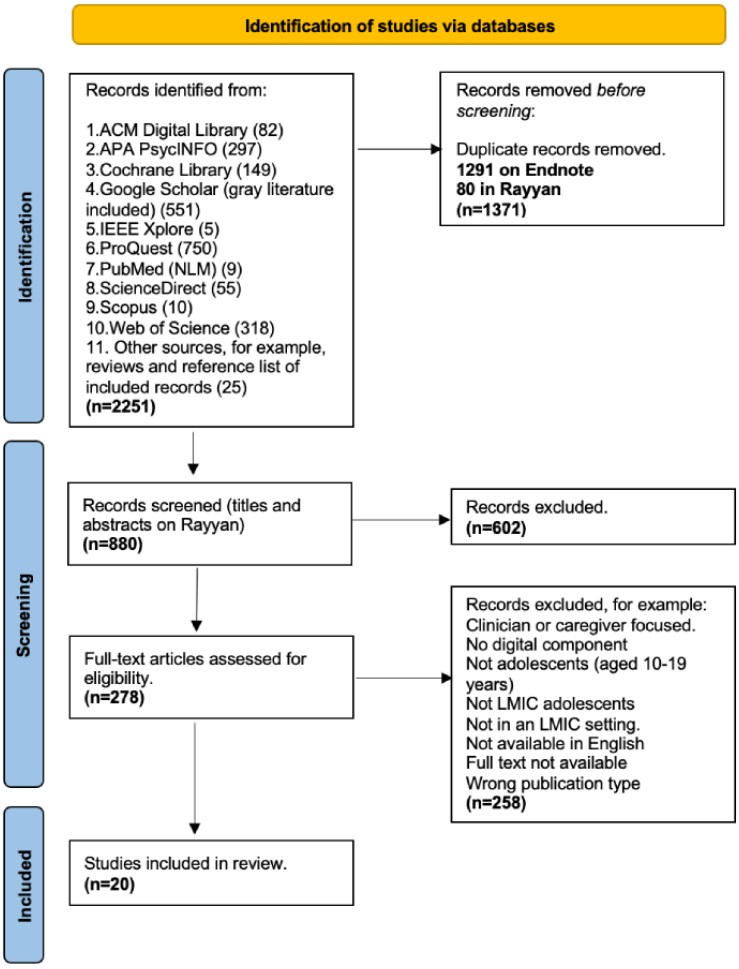

XSL•FO
RenderX

Study Characteristics

Overview

A total of 20 studies were reviewed. The articles were published in English between 2019 and 2024: a total of 20% (4/20) in 2019, a total of 25% (5/20) in 2020, a total of 15% (3/20) in 2021, a total of 20% (4/20) in 2022, a total of 10% (2/20) in 2023, and 10% (2/20) in 2024. Where provided, the start year of the intervention implementation was noted: 2013, a total of 5% (1/20) [24]; 2017, a total of 5% (1/20) [69]; 2018, a total of 5% (1/20) [23]; 2019, a total of 15% (3/20) [21,56,70]; 2020, a total of 15% (3/20) [71-73]; 2021, a total of 15% (3/20) [74-76]; 2022, a total of 5% (1/20) [77]; and 2023, a total of 5% (1/20) [78]. The implementation date or year was unclear in the following references [79-84].

Geographically, the lead authors were from India (5/20, 25%); United States (5/20, 25%); Norway (3/20, 15%); and China, Colombia, Finland, Iran, Kenya, Philippines, and Switzerland (1/20, 5% each).

Concerning study design, RCTs were 25% (5/20) [22,23,56,82,84] of all studies, followed by RCT protocols (3/20, 15%) [74,75,77], qualitative studies (4/20, 20%) [69,76,79,81], mixed methods approaches (5/20, 25%) [71-73,78,80], a quasi-experimental feasibility study (1/20, 5%) [24], a multicycle usability testing approach (1/20, 5%) [70], an intervention design (1/20, 5%) [83], and a pilot study (1/20, 5%) [21]. A total of 15% (3/20) were study protocols, therefore the studies were underway.

Regarding the outcome measures, 70% (14/20) reported on primary and secondary measures, while 30% (6/20) did not. The RCTs used a total of 21 different psychometric tests, scales, and outcome measures (see Multimedia Appendix 4 for the charted data). The most common psychometric measures were the Patient Health Questionnaire-8 and Patient Health Questionnaire-9 for measuring depression and the Generalized Anxiety Disorders-7 for measuring anxiety. In 2 studies, the outcome measures were culturally validated for the target population [23,82].

In total, 100% (20/20) of studies described participants as “adolescents” aged between 10 and 19 years. Overall, the sample size was noted in 90% (18/20) of papers and ranged from 10 to 3960 participants. However, in 2 studies, 10% (2/20) did not report the sample size [76,83].

The duration of the interventions varied widely; for example, 10% (2/20) were single-session interventions [21,75], 5% (1/20) lasted 2 to 4 weeks [76], 10% (2/20) lasted 4 weeks [56,74], 5% (1/20) lasted 5 weeks [71], 5% (1/20) lasted 7 weeks [24], 5% (1/20) lasted 2 months [78], 5% (1/20) lasted 10 weeks [72], 5% (1/20) lasted 11 weeks [70], 5% (1/20) lasted 12 weeks [84], 5% (1/20) was anticipated to last between 2 and 4 months [81], 20% (4/20) lasted 6 months [23,79,80,82], and 5% (1/20) lasted 12 months [77]. The following did not specify the length of the intervention [69,73,83]. Of the 20 papers, 60% (12/20) had a follow-up period ranging from baseline to 12 months, and 40% (8/20) did not report a follow-up period. In 1 study (1/20, 5%), the authors called for a future economic evaluation of the DMHI [77]. Another paper (1/20, 5%) [76], reported on a cost-effective evaluation, noting that the cost of an incremental increase in well-being was US $37, and the cost of reducing emotional and behavioral issues was US $20. Of the other studies, 90% (18/20) did not report on the cost-effectiveness of their DMHI.

DMHIs

A total of 14 different DMHIs were identified, addressing one or more of the following mental health disorders: depression, 65% (13/20); well-being, 30% (6/20); emotional problem-solving, 20% (4/20); anxiety, 30% (6/20); distress, 5% (1/20); suicidal ideation, 10% (2/20); stigma, 10% (2/20); resilience, 5% (1/20); self-harm, 15% (3/20); and general mental health disorders, 10% (2/20). The main combination was depression and anxiety. Most DMHIs (13/20, 65%) were delivered or planned to be delivered in school settings, while others were delivered in clinics (2/20, 10%), hospitals (2/20, 10%), Urban Primary Health Centre in slum clusters (2/20, 10%), and refugee camps (1/20, 5%). Finally, 5% (1/20) did not specify the setting [83] (see Table 1).

XSL•FO
RenderX

Table 1. List of digital mental health interventions, country of implementation, settings, and mode of delivery.

Digital mental health intervention and delivery setting	Description	References
Happy Helping Hand (Attensi Global; Lebanon, Ukraine)—schools and refugee camps	• Digital game for psychosocial support, facilitated by teachers and psychosocial support via iPads, smartphones, or computers (well-being and emotional problem-solving) • Delivery: iPads, smartphones, or computers	[71,72,78]
DepisNet-Thai (Thailand)—schools	• Web program delivered in small groups, supported by teachers (general mental health disorders) • Delivery: web-based	[24]
Intervention via WhatsApp (Meta Platforms, Inc; Kenya)—care clinics	• Intervention delivered via the WhatsApp platform with a trained pediatric HIV adherence and disclosure counselor, group chat (depression) • Delivery: smartphone mobile intervention	[79,80]
SMS text messaging intervention (China)—hospitals	• SMS text messages delivered to reduce deliberate self-harm, individual intervention (self-harm) • Delivery: mobile phones	[81]
DIALOG+S (Unit for Social and Community Psychiatry at Queen Mary University of London and the East London Foundation Trust and Pontificia Universidad Javeriana, Colombia)—schools	• A digital app-supported evidence-based intervention to assist in the interaction between patients with a mental health condition and clinicians, delivered by teachers (mental well-being and resilience) • Delivery: multiple devices such as tablets or cellphones	[82]
PRIDE[a] research program (2016-2022)		
POD[b] Adventures (India)—schools	• Problem-solving game-based intervention (depression and anxiety) • Delivery: smartphone	[69]
PRIDE (2016-2022; India)—schools	• Stepped-care intervention targeting common mental health problems in school-going adolescents, counselor support (general mental health conditions) • Delivery: smartphone and internet-based	[74]
STARS[c] (South Africa)—N/A[d]	• World Health Organization transdiagnostic chatbot for distressed youths (distress) • Delivery: through existing chatbot systems using different technologies (eg, apps, websites, and messaging platforms)	[83]
Shamiri-Digital (Kenya)—schools	• Universal, computerized self-help SSI[e], group-based, lay support (depression, anxiety, and well-being) • Delivery: internet-based	[56,75]
DAD[f] course (Iran)—schools	• Education program, individual delivery, 6 months (8 × 30 min sessions; depression) • Delivery: web-based	[23]
ARTEMIS[g] Intervention (India)—Urban Primary Health Centre in slum clusters	• Combines an antistigma campaign with a digital health intervention to identify and manage mental health conditions (depression, self-harm, suicide risk, or other significant emotional complaints) • Delivery: tablets	[73,77]
Kuamsha app (South Africa and Uganda)—schools	• Delivers behavioral activation to treat depression among adolescents, 6 modules taking 15-20 min to complete, supported by a brief weekly phone call from a peer mentor (depression) • Delivery: smartphone	[70]
Smartteen (India)—hospital	• Computer-assisted CBT[h] intervention—a computer application designed to augment in-person CBT for treatment of depression in adolescents, delivered with therapist support (depression) • Delivery: computer-based	[84]
Self-help game-based mobile app intervention (Philippines)—schools	• Teaches service users skills that help them address mental health (psychosocial intervention), four 30-40-minute sessions that are spanned for 2-4 weeks, group delivery, counselor support (mild or moderate depression and anxiety) • Delivery: school computer laboratories	[76]

XSL•FO
RenderX

Digital mental health intervention and delivery setting	Description	References
3 Computerized SSIs (India)—schools	• Self-guided computerized mental health intervention (depression and anxiety) • Delivery: school computer laboratories	[21]

[a]PRIDE: Premium for Adolescents.

[b]POD: identifying "Problems," generating "Options," and creating a "Do it" plan.

[c]STARS: Sustainable Technology for Adolescents and Youth to Reduce Stress.

[d]N/A: not available.

[e]SSI: single session intervention.

[f]DAD: Dorehye Amozeshie Dokhtaran.

[g]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

[h]CBT: cognitive behavioral therapy.

Of the 20 studies, 30% (6/20) were conducted in India, 20% (4/20) in Kenya, 10% (2/20) in Lebanon, 5% (1/20) in China, 5% (1/20) in Colombia, 5% (1/20) in Iran, 5% (1/20) in the Philippines, 5% (1/20) in South Africa, 5% (1/20) in South Africa and Uganda, 5% (1/20) in Thailand, and 5% (1/20) in Ukraine. There were 4 upper-middle–income economies, 6 lower-middle–income economies, and 1 low-income economy. All met the LMIC definition set by the World Bank (see Tables 2 and 3).

Table 2. Country of digital mental health intervention implementation (N=20).

Study country	Studies, n (%)	References
India	6 (30)	[69,73-75,77,84]
Kenya	4 (20)	[56,75,79,80]
Lebanon	2 (10)	[71,72]
China	1 (5)	[81]
Colombia	1 (5)	[82]
Iran	1 (5)	[23]
Philippines	1 (5)	[76]
South Africa	1 (5)	[83]
South Africa and Uganda	1 (5)	[70]
Thailand	1 (5)	[24]
Ukraine	1 (5)	[78]

Table 3. Range of low- and middle-income countries with adolescent digital mental health interventions.

Economy type[a]	Countries, n	List of countries
Upper-middle–income economies (US $4466 to US $13,845)	4	China, Colombia, South Africa, and Thailand
Lower-middle–income economies (US $1136 to US $4465)	6	India, Iran, Kenya, Lebanon, Philippines, and Ukraine
Low-income economies (US $1135 or less)	1	Uganda

[a]World Bank Classification for the current 2024 fiscal year, low- and middle-income countries are defined as those with a gross national income per capita less than US $13,845.

Adolescent Involvement in DMHI Design

Of the 20 papers, 60% (12/20) involved adolescents in co-designing [85] or participatory approaches at the design stage. For instance, the content of DepisNet-Thai [24] was developed from discussions with adolescents. Elsewhere, co-designing activities informed the development of POD Adventures [69], PRIDE [74], antistigma materials for ARTEMIS (Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums) [73], and the Kuamsha app [70]. Adolescents contributed to intervention designs via group discussions for Smartteen [84] and Shamiri-Digital [21]. Some interventions were existing DMHIs adapted to the target users' cultural contexts, for example, the Happy Helping Hand was initially launched in Norway but was developed in collaboration with adolescent Syrian refugees and Norwegian adolescents. It was later culturally adapted for use with refugee adolescents in Lebanon [71,72]. Furthermore, the Happy Helping Hand

XSL•FO
RenderX

materials were also adapted for Ukrainian adolescents [78]. In an ongoing study (1/20, 5%), an Adolescent Expert Advisory Group will guide researchers on all aspects of the project design, development, and implementation [77].

Consulted Frameworks, Theories, or Models

Of 20 studies, 60% (n=12) engaged with frameworks, toolkits, models, or theories related to behavior change interventions. These included cognitive behavioral models, social learning theory, theories of coping during adolescence, persuasive systems design, and stress-coping theory, among others. Further, 1 (5%) study consulted a digital health framework that included digital maturity and readiness content [69]. Another study referred to guidance for developing and evaluating complex interventions [73]. One paper (5%) reported how their intervention aligned with the national mental health program and the government's national adolescent health program [77], ensuring their DMHI meets the needs and priorities of the nation's adolescents.

Factors Affecting Adolescent Engagement With the DMHI (Facilitators or Barriers)

A total of 90% (18/20) reported on adolescent facilitators and barriers to DMHI engagement. Further, 3 (15%) studies were underway and, therefore, had no findings or results to report at the time of the present review. However, the authors listed factors that facilitated or impeded access to DMHIs for adolescents where possible (see Textbox 2).

XSL•FO
RenderX

Textbox 2. Summary of barriers and facilitators (N=20).

Key barriers

- COVID-19 challenges, for example, exposed adolescents' overlooked social and personal needs.
- Complexity of app implementation, for example, lay counselor training, app development, and mental health service referrals.
- Household and school responsibilities.
- Power outages.
- Affordability, for example, insufficient phone credit, and lack of access to smartphones or mobile phones.
- Limited digital skills, low literacy levels, or both.
- Challenges with user emails reduced acceptance of the program.
- Time constraints on the app delivery by psychosocial staff or teachers.
- Boarding school attendance prevented 1 adolescent participant from participating.
- Duration of intervention.
- Stigma, discrimination, or lack of cultural appropriateness.
- Parent control of device use, for example, parents keeping adolescents' mobile phones for most of the day.
- Lack of privacy and confidentiality.
- Concerns from teachers and parents on the time spent on phones.
- Poor local infrastructure and connectivity.
- Lack of online functionality.
- Lack of digital skills leading to a dependency on teachers or lay staff guiding the intervention.
- Device affordability, for example, phone data running out often and having to wait to top up again.
- Cost of intervention.
- Digital divide evident, for example, lack of digital device ownership and infrastructure.

Key facilitators

- Adapting digital mental health interventions (DMHIs) to the local context, for example, including language, cultural, and religious values.
- Embedding cultural appropriateness by design.
- Access to counselors or lay support.
- Offline access.
- Reading materials in addition to the DMHI.
- A safe space to discuss sensitive and stigmatized issues, for example, HIV-related concerns.
- Value in face-to-face interaction.
- Private school setting meant the availability of computers and internet connection.
- Family support.
- The timing of the intervention, for example, receiving a SMS text message during evenings when self-harm urges are increased.
- Peer groups comprising 10 to 15 adolescents to provide lay counseling and mental health support for adolescents in the community.
- Gamification.
- Tailored content.
- Privacy and confidentiality.
- Free to use, for example, the provision of a smartphone with the WhatsApp app preinstalled.

Discussion

Summary of Evidence

This scoping review aimed to explore what is known about DMHIs for adolescents in LMICs, as reported in the literature. In 2024, the World Bank categorized 134 countries as LMICs [9]. This review located only 11 countries engaged with adolescent DMHIs, illustrating the limited implementation of these solutions across LMICs and reflecting the emergent nature of the field [45]. We analyzed 20 papers to understand how these DMHIs were designed and evaluated, in what capacity

XSL•FO
RenderX

adolescents were involved in the design process, what frameworks or theories were applied, and what factors impacted adolescent engagement with the DMHIs.

Overall, 14 different DMHIs were identified (POD Adventures is part of the PRIDE research program 2016-2022). DMHIs were designed for a limited range of mental health disorders, namely depression and anxiety. Almost all included adolescent input at the design stage, underscoring the importance of person-centered or user-centered approaches in DHIs. The DMHIs were delivered in diverse formats, including group delivery and self-guided methods, with lay staff or mental health professionals, including counselors, psychologists, or psychiatrists. Interventions were administered through smartphones, mobile phones, tablets, computers, and the web, often conducted in schools or clinical settings. Significant structural, psychological, and financial barriers exist in engagement with the DMHIs.

The study designs included RCTs, mixed methods, and qualitative studies. Most studies reported positive outcomes for symptom reduction, feasibility, and acceptability, measured by specific outcome measures. However, most were small scale and not trialed on a large scale over extended periods. While DMHIs were feasible and acceptable among adolescents, there remains a gap in the literature about their long-term cost- and clinical effectiveness.

Design of Adolescent DMHIs

The DMHIs were designed for depression, anxiety, self-harm, well-being, resilience, stigma, stress, and suicide ideation. However, this is not reflective of the broad and complex range of mental health disorders that adolescents develop, including substance misuse and posttraumatic stress disorder [1,6,45]. Concerning intervention settings, most were delivered in schools or hospitals. Indeed, previous studies have recognized schools as critical settings that offer a vital entry point for receiving adolescent mental health services [86,87]. Conversely, out-of-school adolescents or those unlikely to visit those hospitals may miss the opportunity to engage with the DMHIs [88].

While RCTs remain the gold standard for evaluating the effectiveness of DHIs [89,90], they are costly and time-consuming. We identified only 5 RCTs evaluating the effectiveness of adolescent DMHIs [23,56,82,84], with a further 3 RCT protocols underway [74,75,77]. In India, Smartteen [84] was trialed on 21 adolescents with depression against a treatment-as-usual group. The DMHI was feasible, acceptable, and more effective than the treatment-as-usual at reducing symptoms at 12 weeks. Even with reduced therapist time, adolescents adhered to treatment compliance. However, the authors call for more rigorous evaluations at scale. In another study [82], DIALOG+ was adapted for adolescents in educational settings (DIALOG+S), using focus groups with teachers and adolescents. The DMHI was trialed with 70 Colombian adolescents, randomly assigned into DIALOG+S or an active control group (counseling as usual). The intervention was feasible and acceptable and could improve mental health, quality of life, and emotional symptoms. The authors call for larger studies to assess its efficacy. In Iran, the DAD (Dorehye Amozeshie Dokhtaran) [23] was trialed with 128 adolescents, randomly assigned into the DAD or a control group. The intervention showed an improvement in depression symptoms. However, the effects decreased after 12 weeks. The intervention did not affect the outcome expectations or self-efficacy. Elsewhere Shamiri-Digital [75] was trialed with 103 adolescents, randomly assigned into Shamiri-Digital or a study-skills control condition. Shamiri-Digital reduced depressive symptoms compared to the control. However, there were no significant effects on anxiety symptoms, well-being, or happiness. The authors called for replicate trials with extended follow-up periods. In a secondary analysis of this trial, the authors sought to evaluate the costs and cost-effectiveness of Shamiri-Digital through an economic evaluation [91]. Their findings indicate that Shamiri-Digital can be delivered for less than US $4 per student, which is more cost-effective than traditional interventions, for example, 12-16-week cognitive behavioral therapy sessions. However, it is difficult to draw inferences from these studies without critically assessing the quality of the evidence.

A critical design decision came from the inclusion of culturally validated elements. Some studies culturally adapted some components of the intervention [78,82]. Further, 1 study culturally validated an existing psychometric tool for their DMHI, for example, the study [24] adapted the Perceived Stress Scale to make it a reliable and valid instrument in keeping with the Thai culture. Additionally, another study [23] used the Farsi version of the Perceived Social Support Scale-Revised and the Farsi version of the Sherer General Self-Efficacy Questionnaire, reporting that the adapted scales were more appropriate and aligned to Iranian cultures. This aligns with existing literature emphasizing the importance of culturally appropriate DHIs [92,93]. Furthermore, they ensure resonance with the target populations' beliefs, practices, and values leading to improved engagement, treatment adherence, and outcomes [10,44,65,94-96]. In contrast, the Norwegian-designed Happy Helping Hand app was not adapted to the Arab culture for Syrian adolescents, and this was detrimental to the delivery of a DMHI, as themes were deemed religiously and culturally inappropriate, and therefore, certain aspects of the DMHI were skipped [71].

An important design element was the inclusion of lay, peer, or therapist support, with adolescents preferring this option [73,76,79]. The benefits of this approach have been previously reported [55,57,97], where interventions with in-person options were more effective than self-guided or automized interventions [55].

Adolescent Engagement With DMHI Design

The advantages of involving adolescents in participatory and co-design approaches for DMHIs are well documented [67,98-104]. In this review, for example, adolescents were engaged in a series of co-design workshops for exploring existing popular apps, for example, Temple Run (Imangi Studios) and Candy Crush (King); story building, paper prototyping, and discussions about prototype ideas for POD Adventures [69]; the cocreation of an antistigma campaign as part of the main intervention for the ARTEMIS project [77]; and 1 study incorporated feedback from recent high school

XSL•FO
RenderX

graduates into the design of the group-based Shamiri-Digital intervention as an iterative process of the intervention design [56].

Collectively these studies show that participatory approaches are invaluable for generating insights for improving the DMHIs. Nevertheless, there are challenges [105], and authors are calling for the evaluation of co-design processes in diverse contexts and how that impacts technology [102,103] with clear guidance around these processes [98]. These insights will advance understanding of how and why adolescents engage in DHIs [101] and the health care outcomes of such engagements. From the review, reported activities included feasibility, acceptability, and usability studies, workshops, interviews, and focus groups. All were used clearly and meaningfully to advance the DMHI design and evaluation [24,69,70,73,74].

Applied Frameworks and Theories

Incorporating evidence-based theories and techniques that encourage user engagement and behavior change is an essential element of DHI design [25,26,44,106]. Most DMHIs in this review were underpinned by cognitive-behavioral techniques involving behavioral activation and problem-solving features [71,72,76,78,84]. Some used stand-alone theories [23], for example, the social cognitive theory, and others used a combination of behavior change techniques [24], for example, theories of coping during adolescence, the self-determination theory, and the technology acceptance model. However, not all the papers explained how the behavior change theory or technique impacted the target user's mental health. Or, indeed, the evidence on which the selection or combinations of theories were based [106].

Further, incorporating national or international guidance enhanced the intervention design, development, and evaluation. For example, 1 study [69] used the WHO guidelines on monitoring and evaluating DHIs, which addressed digital maturity, readiness, and scalability, whereas another [73] consulted the UK Medical Research Council guidance for developing and evaluating complex interventions focused on process evaluation and contextual factors. Finally, 1 study [77] accessed the WHO's Mental Health Gap Action Programme-Intervention Guide, which considers training requirements and treatment protocols in low-resource settings.

Reported Barriers and Facilitators

Textbox 2 summarizes key barriers and facilitators from the included studies. Access and engagement with DMHIs can be impacted by financial, geographical, psychological, cultural, and structural factors [107-111]. Notable barriers were the affordability of devices, cultural appropriateness, power outages, poor digital and literacy skills, inadequate internet connectivity, and limited access to devices in rural areas. The findings are consistent with other LMICs [38,44,112].

Stigma (self-stigma, eg, individual level, and societal stigma, eg, system level) remains a complex and multifaceted barrier despite the potential for increased privacy, confidentiality, anonymity, and accessibility through DMHIs [113-115]. Multiple factors contribute to persistent stigma, such as societal attitudes, the digital divide, lack of awareness, fear of exposure, discrimination, and cultural sensitivities [113,116,117]. These factors could limit engagement with DMHIs due to the negative social consequences in areas where mental health is a stigmatized issue [118,119].

Further, 1 study described how COVID-19 worsened the already dire human rights and socioeconomic conditions for displaced Syrians living in Lebanon [71], including the loss of income, food, and essential living services, disrupting education and access to learning resources. Another study [82] reported similar findings in Colombia, noting that most adolescents impacted by COVID did not have access to support and treatment. This shortage of mental health professionals and a widening treatment gap, poor education, and community knowledge reinforced the stigma and false beliefs about mental health [82].

Adolescents highlighted some facilitators, such as access to a safe space to discuss stigmatized mental health issues, culturally sensitive DMHIs that embed the local cultural and religious values by design, DMHIs that maintain face-to-face contact while offering therapist-led or lay or peer-led options, access to reading materials in addition to the DMHI, appropriate content that is entertaining, personalized, and has gamified elements, privacy and confidentiality, and DMHIs that are free to use. Previous studies have suggested that addressing the digital divide could mitigate these barriers, providing equitable access to DMHIs, yet care must be taken to ensure that the pursuit of DMHIs does not widen the existing inequalities, further reinforcing the digital divide, undermining the equitable delivery of care [120-122].

Limitations

To the authors' knowledge, 1 scoping review [123] has been published on DMHIs for adolescents in LMICs. Their review [123] was limited to mobile phones only; however, this review has a wider scope, which is more inclusive of the technological infrastructure in LMICs. Furthermore, this review has a more extensive search strategy, searching 10 databases and including gray literature such as dissertations and conference proceedings, and is a more robust evidence synthesis, whereas the review [123] only searched APA PsycINFO, Web of Science, Psychiatry Online, and EBSCOhost. Finally, the eligibility criteria are different, and the previous review is limited by its lack of specificity of publication parameters per publication dates. This review has been purposefully restricted to a 5-year period (2019-2024), as specified earlier in this paper.

There are some limitations to the present review; 1 reviewer performed the database searches, which may have introduced reviewer bias. However, 2 more reviewers independently screened 25% (220/880) of each of the titles and abstracts of potential papers against the eligibility criteria. This step reduces bias and adds credibility and rigor to the review process. In addition, the quality of included studies was not appraised or assessed for risk of bias therefore, there is no assurance of the quality of the evidence in this scoping review. This is consistent with the aim of a scoping review, which is to rapidly map the available literature and not perform a systematic analysis. Finally, the review was limited to papers published in English. This may have significantly reduced the number of eligible studies from LMICs.

XSL•FO
RenderX

Conclusions

The global shortage of access to mental health treatment demonstrates the critical need for effective, low-cost, scalable solutions. The COVID-19 pandemic has highlighted the need for alternative solutions, and leveraging digital technologies through the international frameworks of the UHC and SDGs can alleviate the significant burden of mental health inequalities among adolescents in LMICs.

In this review, adolescent involvement in various stages of the theory-based design and evaluation cycle enhanced the intervention's acceptability, feasibility, and usefulness. While RCTs of DMHIs remain the gold standard in this field, they are sparse. Researchers have called for repeat trials at larger scales in diverse settings to assess their effectiveness, scalability, and feasibility. Moreover, only some of the tools and instruments used to evaluate the DMHIs were culturally validated in LMIC contexts, making it difficult to establish how the DMHI would benefit adolescents beyond the test settings.

Addressing the digital divide and related barriers, such as stigma, will be a critical challenge to ensuring equitable access to universal and affordable mental health care. Understanding the long-term clinical implications and the cost-effectiveness of DMHIs will be another challenge for future research because that is an essential consideration for sustainable DMHIs in resource-constrained settings.

Acknowledgments

We are grateful to Elaine Blair (MSc, Faculty Librarian of Science at the University of Strathclyde) for providing professional guidance on database identification and searches. Funding sources were not involved in the writing of this paper.

Authors' Contributions

CW, LM, and ML conceptualized this study. CW wrote the original draft of this paper. CW extracted the data from the articles, and CR validated the data extraction. Titles and abstracts were screened independently by CW (880/880, 100%), LM (220/880, 25%), and CR (220/880, 25%). CW conducted the full-text screening. CW, LM, and CR resolved conflicts during the screening. ML and LM edited multiple drafts; all authors reviewed and approved the final paper.

Conflicts of Interest

LM and ML are the primary and secondary supervisors, respectively, for CW's PhD at the University of Strathclyde. LM is also the PhD supervisor for CR.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 109 KB - jmir_v26i1e51376_app1.docx]

Multimedia Appendix 2
Google Scholar, custom range search date (2019-2024).
[DOCX File , 18 KB - jmir_v26i1e51376_app2.docx]

Multimedia Appendix 3
The search strings and results for each database.
[DOCX File , 47 KB - jmir_v26i1e51376_app3.docx]

Multimedia Appendix 4
The data charting document.
[DOCX File , 87 KB - jmir_v26i1e51376_app4.docx]

References

1. World mental health report: Transforming mental health for all. World Health Organization. Geneva; 2022. URL: https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf?sequence=1
2. Kieling C, Baker-Henningham H, Belfer M, Conti G, Ertem I, Omigbodun O, et al. Child and adolescent mental health worldwide: evidence for action. Lancet 2011 Oct 22;378(9801):1515-1525. [doi: 10.1016/S0140-6736(11)60827-1] [Medline: 22008427]
3. Helping adolescents thrive toolkit: strategies to promote and protect adolescent mental health and reduce self-harm and other risk behaviours. World Health Organization. 2021. URL: https://iris.who.int/bitstream/handle/10665/341327/9789240025554-eng.pdf?sequence=1 [accessed 2024-03-04]
4. Comprehensive mental health action plan 2013-2030. World Health Organization. 2021. URL: https://iris.who.int/bitstream/handle/10665/345301/9789240031029-eng.pdf?sequence=1&isAllowed=y [accessed 2024-03-04]

XSL•FO
RenderX

5. Clark H, Coll-Seck AM, Banerjee A, Peterson S, Dalglish SL, Ameratunga S, et al. A future for the world's children? A WHO-UNICEF-Lancet Commission. Lancet 2020 Feb 22;395(10224):605-658. [doi: 10.1016/S0140-6736(19)32540-1] [Medline: 32085821]
6. Mental health of adolescents. World Health Organization. 2024 Oct 10. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2024-10-15]
7. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health 2018 Mar;2(3):223-228. [doi: 10.1016/S2352-4642(18)30022-1] [Medline: 30169257]
8. The state of the world's children 2021: on my mind promoting, protecting and caring for children's mental health. United Nations Children's Fund. 2021. URL: https://www.unicef.org/media/108036/file/SOWC-2021-executive-summary.pdf [accessed 2024-03-01]
9. World bank country and lending groups. The World Bank. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2024-03-14]
10. Babatunde GB, van Rensburg AJ, Bhana A, Petersen I. Barriers and Facilitators to Child and Adolescent Mental Health Services in Low-and-Middle-Income Countries: a Scoping Review. Glob Soc Welf 2019 Jun 07;8(1):29-46. [doi: 10.1007/s40609-019-00158-z]
11. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021 Jun;8(6):535-550 [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
12. Sharma M, Idele P, Manzini A. Life in lockdown: child and adolescent mental health and well-being in the time of COVID-19. ERIC. 2021. URL: https://files.eric.ed.gov/fulltext/ED615551.pdf [accessed 2022-10-04]
13. Caplan S, Lovera AS, Liberato PR. A feasibility study of a mental health mobile app in the Dominican Republic: The untold story. Int J Ment Health 2019 Apr 04;47(4):311-345. [doi: 10.1080/00207411.2018.1553486]
14. Kaonga NN, Morgan J. Common themes and emerging trends for the use of technology to support mental health and psychosocial well-being in limited resource settings: a review of the literature. Psychiatry Res 2019 Nov;281:112594. [doi: 10.1016/j.psychres.2019.112594] [Medline: 31605874]
15. Naslund JA, Gonsalves PP, Gruebner O, Pendse SR, Smith SL, Sharma A, et al. Digital innovations for global mental health: opportunities for data science, task sharing, and early intervention. Curr Treat Options Psychiatry 2019 Dec;6(4):337-351 [FREE Full text] [doi: 10.1007/s40501-019-00186-8] [Medline: 32457823]
16. Calvo-Valderrama MG, Marroquín-Rivera A, Burn E, Ospina-Pinillos L, Bird V, Gómez-Restrepo C. Adapting a mental health intervention for adolescents during the COVID-19 pandemic: web-based synchronous focus group study. JMIR Form Res 2021 Nov 03;5(11):e30293 [FREE Full text] [doi: 10.2196/30293] [Medline: 34637395]
17. Mundt AP, Irarrázaval M, Martínez P, Fernández O, Martínez V, Rojas G. Telepsychiatry consultation for primary care treatment of children and adolescents receiving child protective services in Chile: mixed methods feasibility study. JMIR Public Health Surveill 2021 Jul 22;7(7):e25836 [FREE Full text] [doi: 10.2196/25836] [Medline: 34292164]
18. Rodriguez-Villa E, Rozatkar AR, Kumar M, Patel V, Bondre A, Naik SS, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021;8:e30 [FREE Full text] [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
19. Balcombe L, de Leo D. Digital mental health challenges and the horizon ahead for solutions. JMIR Ment Health 2021 Mar 29;8(3):e26811 [FREE Full text] [doi: 10.2196/26811] [Medline: 33779570]
20. Classification of digital health interventions v1.0. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2024-03-24]
21. Wasil AR, Park SJ, Gillespie S, Shingleton R, Shinde S, Natu S, et al. Harnessing single-session interventions to improve adolescent mental health and well-being in India: development, adaptation, and pilot testing of online single-session interventions in Indian secondary schools. Asian J Psychiatr 2020 Apr;50:101980. [doi: 10.1016/j.ajp.2020.101980] [Medline: 32146337]
22. Kacmarek CN, Johnson NE, Osborn TL, Wasanga C, Weisz JR, Yates BT. Costs and cost-effectiveness of Shamiri, a brief, layperson-delivered intervention for Kenyan adolescents: a randomized controlled trial. BMC Health Serv Res 2023 Aug 04;23(1):827 [FREE Full text] [doi: 10.1186/s12913-023-09856-z] [Medline: 37542304]
23. Moeini B, Bashirian S, Soltanian AR, Ghaleiha A, Taheri M. Examining the effectiveness of a web-based intervention for depressive symptoms in female adolescents: applying social cognitive theory. J Res Health Sci 2019 Aug 19;19(3):e00454 [FREE Full text] [Medline: 31586376]
24. Anttila M, Sittichai R, Katajisto J, Välimäki M. Impact of a web program to support the mental wellbeing of high school students: a quasi experimental feasibility study. Int J Environ Res Public Health 2019;16(14):2473 [FREE Full text] [doi: 10.3390/ijerph16142473] [Medline: 31336738]
25. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
26. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change

XSL•FO
RenderX

interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
27. Hamilton J, Siegel D, Carper M. 16—Digital media interventions for adolescent mental health. Handb Adolesc Digital Media Use Ment Health 2022:389 [FREE Full text] [doi: 10.1017/9781108976237.021]
28. Hartog K, Hubbard CD, Krouwer AF, Thornicroft G, Kohrt BA, Jordans MJD. Stigma reduction interventions for children and adolescents in low- and middle-income countries: systematic review of intervention strategies. Soc Sci Med 2020 Feb;246:112749 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112749] [Medline: 31978636]
29. Kanuri N, Arora P, Talluru S, Colaco B, Dutta R, Rawat A, et al. Examining the initial usability, acceptability and feasibility of a digital mental health intervention for college students in India. Int J Psychol 2020 Aug;55(4):657-673 [FREE Full text] [doi: 10.1002/ijop.12640] [Medline: 31867730]
30. Gómez-Restrepo C, Sarmiento-Suárez MJ, Alba-Saavedra M, Bird VJ, Priebe S, van Loggerenberg F. Adapting DIALOG+ in a school setting-a tool to support well-being and resilience in adolescents living in postconflict areas during the COVID-19 pandemic: protocol for a cluster randomized exploratory study. JMIR Res Protoc 2022 Nov 09;11(11):e40286 [FREE Full text] [doi: 10.2196/40286] [Medline: 36350703]
31. Gonsalves PP, Hodgson ES, Bhat B, Sharma R, Jambhale A, Michelson D, et al. App-based guided problem-solving intervention for adolescent mental health: a pilot cohort study in Indian schools. Evid Based Ment Health 2021 Feb;24(1):11-18 [FREE Full text] [doi: 10.1136/ebmental-2020-300194] [Medline: 33208507]
32. Gronholm PC, Makhmud A, Barbui C, Brohan E, Chowdhary N. Qualitative evidence regarding the experience of receiving and providing care for mental health conditions in non-specialist settings in low-income and middle-income countries: a systematic review of reviews. BMJ Ment Health 2023;26(1):e300755 [FREE Full text] [doi: 10.1136/bmjment-2023-300755] [Medline: 37612041]
33. Raj V, Raykar V, Robinson AM, Islam MR. Child and adolescent mental health training programs for non-specialist mental health professionals in low and middle income countries: a scoping review of literature. Community Ment Health J 2022 Jan;58(1):154-165 [FREE Full text] [doi: 10.1007/s10597-021-00805-w] [Medline: 33651235]
34. Devassy SM, Scaria L, Cheguvera N. Task sharing and stepped referral model for community mental health promotion in low- and middle-income countries (LMIC): insights from a feasibility study in India. Pilot Feasibility Stud 2022 Aug 30;8(1):192 [FREE Full text] [doi: 10.1186/s40814-022-01159-0] [Medline: 36042504]
35. Juengsiragulwit D. Opportunities and obstacles in child and adolescent mental health services in low- and middle-income countries: a review of the literature. WHO South East Asia J Public Health 2015;4(2):110-122. [doi: 10.4103/2224-3151.206680] [Medline: 28607309]
36. Naeem F, Husain M, Husain M, Javed A. Digital psychiatry in low- and middle-income countries post-COVID-19: opportunities, challenges, and solutions. Indian J Psychiatry 2020 Sep;62(Suppl 3):S380-S382 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_843_20] [Medline: 33227054]
37. van Dijk JAGM. Digital divide: impact of access. Int Encycl Media Eff 2017:1-11. [doi: 10.1002/9781118783764.wbieme0043]
38. Ferretti A, Vayena E, Blasimme A. Unlock digital health promotion in LMICs to benefit the youth. PLOS Digit Health 2023 Aug;2(8):e0000315 [FREE Full text] [doi: 10.1371/journal.pdig.0000315] [Medline: 37540713]
39. Ragnedda M, Gladkova A. Understanding digital inequalities in the global south. Digital Inequalities Global South 2020:17-30 [FREE Full text] [doi: 10.1007/978-3-030-32706-4_2]
40. Rural-urban gap in mobile internet use in low-and middle-income countries (LMICs) worldwide from 2017 to 2019, by region. Statista. URL: https://www.statista.com/statistics/264067/global-mobile-internet-use-rural-urban-gap-low-middle-income-countries-region/ [accessed 2024-03-06]
41. Most common barriers to mobile internet use in low and middle income countries (LMICs) in 2019, by region. Statista. URL: https://www.statista.com/statistics/265731/share-barriers-mobile-internet-use-low-middle-income-countries-region/ [accessed 2024-03-06]
42. The state of mobile internet connectivity 2022. GSMA. 2022. URL: https://www.gsma.com/r/wp-content/uploads/2022/12/The-State-of-Mobile-Internet-Connectivity-Report-2022.pdf?utm_source=website&utm_medium=download-button&utm_campaign=somic22 [accessed 2024-04-03]
43. Boucher E, Moskowitz JT, Kackloudis GM, Stafford JL, Kwok I, Parks AC. Immediate and long-term effects of an 8-week digital mental health intervention on adults with poorly managed type 2 diabetes: protocol for a randomized controlled trial. JMIR Res Protoc 2020 Aug 04;9(8):e18578 [FREE Full text] [doi: 10.2196/18578] [Medline: 32749998]
44. Borghouts J, Eikey E, Mark G, de Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
45. Faria M, Zin STP, Chestnov R, Novak AM, Lev-Ari S, Snyder M. Mental health for all: the case for investing in digital mental health to improve global outcomes, access, and innovation in low-resource settings. J Clin Med 2023;12(21):6735 [FREE Full text] [doi: 10.3390/jcm12216735] [Medline: 37959201]
46. van Stam G. Conceptualization and practices in digital health: voices from Africa. Afr Health Sci 2022 Mar;22(1):664-672 [FREE Full text] [doi: 10.4314/ahs.v22i1.77] [Medline: 36032450]

XSL•FO
RenderX

47. Chakrabarti S. Digital psychiatry in low-and-middle-income countries: new developments and the way forward. World J Psychiatry 2024 Mar 19;14(3):350-361 [FREE Full text] [doi: 10.5498/wjp.v14.i3.350] [Medline: 38617977]
48. McCool J, Dobson R, Muinga N, Paton C, Pagliari C, Agawal S, et al. Factors influencing the sustainability of digital health interventions in low-resource settings: lessons from five countries. J Glob Health 2020 Dec;10(2):020396 [FREE Full text] [doi: 10.7189/jogh.10.020396] [Medline: 33274059]
49. Greenhalgh T, Abimbola S. The NASSS framework—a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
50. Carter H, Araya R, Anjur K, Deng D, Naslund JA. The emergence of digital mental health in low-income and middle-income countries: a review of recent advances and implications for the treatment and prevention of mental disorders. J Psychiatr Res 2021 Jan;133:223-246 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.12.016] [Medline: 33360867]
51. Holly L, Smith RD, Ndili N, Franz C, Stevens EAG. A review of digital health strategies in 10 countries with young populations: do they serve the health and wellbeing of children and youth in a digital age? Front Digit Health 2022;4:817810 [FREE Full text] [doi: 10.3389/fdgth.2022.817810] [Medline: 35373182]
52. Faulkenberry JG, Luberti A, Craig S. Electronic health records, mobile health, and the challenge of improving global health. Curr Probl Pediatr Adolesc Health Care 2022 Jan;52(1):101111. [doi: 10.1016/j.cppeds.2021.101111] [Medline: 34969611]
53. Kickbusch I, Piselli D, Agrawal A, Balicer R, Banner O, Adelhardt M, Secretariat of the LancetFinancial Times Commission. The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. Lancet 2021 Nov 06;398(10312):1727-1776. [doi: 10.1016/S0140-6736(21)01824-9] [Medline: 34706260]
54. Schueller SM, Torous J. Scaling evidence-based treatments through digital mental health. Am Psychol 2020 Nov;75(8):1093-1104 [FREE Full text] [doi: 10.1037/amp0000654] [Medline: 33252947]
55. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
56. Osborn TL, Rodriguez M, Wasil AR, Venturo-Conerly KE, Gan J, Alemu RG, et al. Single-session digital intervention for adolescent depression, anxiety, and well-being: outcomes of a randomized controlled trial with Kenyan adolescents. J Consult Clin Psychol 2020 Jul;88(7):657-668. [doi: 10.1037/ccp0000505] [Medline: 32391709]
57. Siddiqui S, Gonsalves PP, Naslund JA, Stein DJ, Fineberg NA, Chamberlain SR. 20—Scaling up of mental health services in the digital age: the rise of technology and its application to low- and middle-income countries. Ment Health Digital World 2022:459-479. [doi: 10.1016/b978-0-12-822201-0.00004-6]
58. Youth-centred digital health interventions: a framework for planning, developing and implementing solutions with and for young people. World Health Organization. 2020. URL: https://iris.who.int/bitstream/handle/10665/336223/9789240011717-eng.pdf?sequence=1&isAllowed=y [accessed 2024-03-16]
59. Kim J, Aryee LMD, Bang H, Prajogo S, Choi YK, Hoch JS, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023 Mar 20;10:e43066 [FREE Full text] [doi: 10.2196/43066] [Medline: 36939820]
60. Peters MDJ, McInerney P, Munn Z, Tricco A, Khalil H. Scoping reviews. JBI. URL: https://jbi-global-wiki.refined.site/space/MANUAL/355862497/10.+Scoping+reviews [accessed 2024-03-28]
61. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
62. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
63. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
64. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
65. Kruzan KP, Williams KDA, Meyerhoff J, Yoo DW, O'Dwyer LC, de Choudhury M, et al. Social media-based interventions for adolescent and young adult mental health: a scoping review. Internet Interv 2022 Dec;30:100578 [FREE Full text] [doi: 10.1016/j.invent.2022.100578] [Medline: 36204674]
66. Malloy JA, Partridge SR, Kemper JA, Braakhuis A, Roy R. Co-design of digital health interventions for young adults: protocol for a scoping review. JMIR Res Protoc 2022 Oct 24;11(10):e38635 [FREE Full text] [doi: 10.2196/38635] [Medline: 36279167]
67. Povey J, Raphiphatthana B, Torok M, Nagel T, Shand F, Sweet M, et al. Involvement of Indigenous young people in the design and evaluation of digital mental health interventions: a scoping review protocol. Syst Rev 2021 May 05;10(1):133 [FREE Full text] [doi: 10.1186/s13643-021-01685-7] [Medline: 33952320]
68. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. PLoS One 2015;10(9):e0138237 [FREE Full text] [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]

XSL•FO
RenderX

69. Gonsalves PP, Hodgson ES, Kumar A, Aurora T, Chandak Y, Sharma R, et al. Design and development of the "POD Adventures" smartphone game: a blended problem-solving intervention for adolescent mental health in India. Front Public Health 2019;7:238 [FREE Full text] [doi: 10.3389/fpubh.2019.00238] [Medline: 31508404]
70. Pozuelo JR, Moffett BD, Davis M, Stein A, Cohen H, Craske MG, et al. User-centered design of a gamified mental health app for adolescents in Sub-Saharan Africa: multicycle usability testing study. JMIR Form Res 2023 Nov 30;7:e51423 [FREE Full text] [doi: 10.2196/51423] [Medline: 38032691]
71. Al-Khayat A. Digitalized psychosocial support in education. Exploring the impact of the Happy Helping Hand app for displaced Syrian adolescents in Lebanon. Oslo Metropolitan University. 2021. URL: https://oda.oslomet.no/oda-xmlui/bitstream/handle/11250/2976713/Al-Khayat_flkm2021.pdf?sequence=1&isAllowed=y [accessed 2024-09-19]
72. Raknes S. The happy helping hand used by Syrian displaced adolescents in Lebanon: a pilot study of feasibility, usefulness and impact. White Paper 2020:1-9.
73. Mukherjee A, Yatirajula SK, Kallakuri S, Paslawar S, Lempp H, Raman U, et al. Using formative research to inform a mental health intervention for adolescents living in Indian slums: the ARTEMIS study. Child Adolesc Psychiatry Ment Health 2024 Jan 20;18(1):14 [FREE Full text] [doi: 10.1186/s13034-024-00704-4] [Medline: 38245796]
74. Gonsalves PP, Sharma R, Hodgson E, Bhat B, Jambhale A, Weiss HA, et al. A guided internet-based problem-solving intervention delivered through smartphones for secondary school pupils during the COVID-19 pandemic in India: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2021 Oct 06;10(10):e30339 [FREE Full text] [doi: 10.2196/30339] [Medline: 34586075]
75. Wasil AR, Osborn TL, Weisz JR, DeRubeis RJ. Online single-session interventions for Kenyan adolescents: study protocol for a comparative effectiveness randomised controlled trial. Gen Psychiatr 2021;34(3):e100446 [FREE Full text] [doi: 10.1136/gpsych-2020-100446] [Medline: 34222793]
76. Superable GM. Guided self-help game-based app for adolescents in the Philippines and low-to middle-income countries. 2022. URL: https://www.academia.edu/102203331/Guided_Self_Help_Game_Based_App_for_Adolescents_in_the_Philippines_and_Low_to_Middle_Income_Countries [accessed 2024-09-19]
77. Yatirajula SK, Kallakuri S, Paslawar S, Mukherjee A, Bhattacharya A, Chatterjee S, et al. An intervention to reduce stigma and improve management of depression, risk of suicide/self-harm and other significant emotional or medically unexplained complaints among adolescents living in urban slums: protocol for the ARTEMIS project. Trials 2022 Jul 29;23(1):612 [FREE Full text] [doi: 10.1186/s13063-022-06539-8] [Medline: 35906663]
78. Raknes S, Chorna T. The Helping Hand in Ukraine: feasibility and potential impact. Pilot Feasibility Stud 2024 Jun 29;10(1):96 [FREE Full text] [doi: 10.1186/s40814-024-01520-5] [Medline: 38951948]
79. Chory A, Nyandiko W, Martin R, Aluoch J, Scanlon M, Ashimosi C, et al. HIV-related knowledge, attitudes, behaviors and experiences of Kenyan adolescents living with HIV revealed in WhatsApp group chats. J Int Assoc Provid AIDS Care 2021;20:2325958221999579 [FREE Full text] [doi: 10.1177/2325958221999579] [Medline: 33657911]
80. Chory A, Callen G, Nyandiko W, Njoroge T, Ashimosi C, Aluoch J, et al. A pilot study of a mobile intervention to support mental health and adherence among adolescents living with HIV in Western Kenya. AIDS Behav 2022 Jan;26(1):232-242 [FREE Full text] [doi: 10.1007/s10461-021-03376-9] [Medline: 34292429]
81. Duan S, Wang H, Wilson A, Qiu J, Chen G, He Y, et al. Developing a text messaging intervention to reduce deliberate self-harm in Chinese adolescents: qualitative study. JMIR mHealth uHealth 2020 Jun 11;8(6):e16963 [FREE Full text] [doi: 10.2196/16963] [Medline: 32392173]
82. Gómez-Restrepo C, Sarmiento-Suárez MJ, Alba-Saavedra M, Calvo-Valderrama MG, Rincón-Rodríguez CJ, Bird VJ, et al. Development and implementation of DIALOG+S in the school setting as a tool for promoting adolescent mental well-being and resilience in a post-armed conflict area in Colombia: exploratory cluster randomized controlled trial. JMIR Form Res 2023 Oct 04;7:e46757 [FREE Full text] [doi: 10.2196/46757] [Medline: 37792465]
83. Hall J, Jordan S, van Ommeren M, Au T, Sway RA, Crawford J, et al. Sustainable Technology for Adolescents and youth to Reduce Stress (STARS): a WHO transdiagnostic chatbot for distressed youth. World Psychiatry 2022 Feb;21(1):156-157 [FREE Full text] [doi: 10.1002/wps.20947] [Medline: 35015345]
84. Srivastava P, Mehta M, Sagar R, Ambekar A. Smartteen—a computer assisted cognitive behavior therapy for Indian adolescents with depression—a pilot study. Asian J Psychiatr 2020 Apr;50:101970. [doi: 10.1016/j.ajp.2020.101970] [Medline: 32114331]
85. Sanders EB, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
86. Ali MM, West K, Teich JL, Lynch S, Mutter R, Dubenitz J. Utilization of mental health services in educational setting by adolescents in the United States. J Sch Health 2019 May;89(5):393-401. [doi: 10.1111/josh.12753] [Medline: 30883761]
87. Parikh R, Michelson D, Sapru M, Sahu R, Singh A, Cuijpers P, et al. Priorities and preferences for school-based mental health services in India: a multi-stakeholder study with adolescents, parents, school staff, and mental health providers. Glob Ment Health (Camb) 2019;6:e18 [FREE Full text] [doi: 10.1017/gmh.2019.16] [Medline: 31531228]

XSL•FO
RenderX

88. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
89. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
90. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon MP. Methods to evaluate the effects of internet-based digital health interventions for citizens: systematic review of reviews. J Med Internet Res 2018 Jun 07;20(6):e10202 [FREE Full text] [doi: 10.2196/10202] [Medline: 29880470]
91. Wasil AR, Kacmarek CN, Osborn TL, Palermo EH, DeRubeis RJ, Weisz JR, et al. Economic evaluation of an online single-session intervention for depression in Kenyan adolescents. J Consult Clin Psychol 2021 Aug;89(8):657-667. [doi: 10.1037/ccp0000669] [Medline: 34472893]
92. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
93. Nittas V, Daniore P, Chavez SJ, Wray TB. Challenges in implementing cultural adaptations of digital health interventions. Commun Med (Lond) 2024 Jan 05;4(1):7 [FREE Full text] [doi: 10.1038/s43856-023-00426-2] [Medline: 38182750]
94. Piers R, Williams JM, Sharpe H. Review: Can digital mental health interventions bridge the 'digital divide' for socioeconomically and digitally marginalised youth? A systematic review. Child Adolesc Ment Health 2023 Feb;28(1):90-104. [doi: 10.1111/camh.12620] [Medline: 36510375]
95. Pineda BS, Mejia R, Qin YZ, Martinez J, Delgadillo LG, Muñoz RF. Updated taxonomy of digital mental health interventions: a conceptual framework. mHealth 2023;9:28 [FREE Full text] [doi: 10.21037/mhealth-23-6] [Medline: 37492117]
96. Abuwalla Z, Kadhem Z, Gladstone T, Mikhael E, Bishay A, van Voorhees BW. Proposed model for the cultural adaptation of an internet-based depression prevention intervention (CATCH-IT) for Arab adolescents. Int J Adolesc Med Health 2017;31(1):28719365 [FREE Full text] [doi: 10.1515/ijamh-2016-0147] [Medline: 28719365]
97. Galagali PM, Brooks MJ. Psychological care in low-resource settings for adolescents. Clin Child Psychol Psychiatry 2020 Jul;25(3):698-711 [FREE Full text] [doi: 10.1177/1359104520929741] [Medline: 32567351]
98. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
99. Li S, Achilles M, Spanos S, Habak S, Werner-Seidler A, O'Dea B. A cognitive behavioural therapy smartphone app for adolescent depression and anxiety: co-design of ClearlyMe. tCBT 2022 Mar 16;15:e13 [FREE Full text] [doi: 10.1017/s1754470x22000095]
100. Brown M, Lord E, John A. Adaptation of ACTivate your wellbeing, a digital health and well-being program for young persons: co-design approach. JMIR Form Res 2023 Apr 13;7:e39913 [FREE Full text] [doi: 10.2196/39913] [Medline: 37052994]
101. Liverpool S, Edbrooke-Childs J. A caregiver digital intervention to support shared decision making in child and adolescent mental health services: development process and stakeholder involvement analysis. JMIR Form Res 2021 Jun 15;5(6):e24896 [FREE Full text] [doi: 10.2196/24896] [Medline: 34128821]
102. Ludlow K, Russell JK, Ryan B, Brown RL, Joynt T, Uhlmann LR, et al. Co-designing a digital mental health platform, "Momentum", with young people aged 7-17: a qualitative study. Digit Health 2023;9:20552076231216410 [FREE Full text] [doi: 10.1177/20552076231216410] [Medline: 38033517]
103. Jones RB, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020 Aug;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
104. Bergin AD, Vallejos EP, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. NPJ Digit Med 2020;3:133 [FREE Full text] [doi: 10.1038/s41746-020-00339-7] [Medline: 33083568]
105. Gan DZQ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022;8:20552076221098268 [FREE Full text] [doi: 10.1177/20552076221098268] [Medline: 35677785]
106. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
107. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072 [FREE Full text] [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
108. Poulsen A, Hickie I, Alam M. Overcoming barriers to mHealth co-design in low- and middle-income countries: a research toolkit. Inf Technol Dev 2024:1-20 [FREE Full text] [doi: 10.1080/02681102.2023.2295480]

XSL•FO
RenderX

109. Koly KN, Saba J, Muzaffar R, Modasser RB, Colon-Cabrera D, Warren N. Exploring the potential of delivering mental health care services using digital technologies in Bangladesh: a qualitative analysis. Internet Interv 2022 Sep;29:100544 [FREE Full text] [doi: 10.1016/j.invent.2022.100544] [Medline: 35615404]
110. Karyotaki E, Miguel C, Panagiotopoulou OM, Harrer M, Seward N, Sijbrandij M, et al. Digital interventions for common mental disorders in low- and middle-income countries: a systematic review and meta-analysis. Glob Ment Health (Camb) 2023;10:e68 [FREE Full text] [doi: 10.1017/gmh.2023.50] [Medline: 38024798]
111. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]
112. Ahmed T, Rizvi SJR, Rasheed S, Iqbal M, Bhuiya A, Standing H, et al. Digital health and inequalities in access to health services in Bangladesh: mixed methods study. JMIR mHealth uHealth 2020 Jul 21;8(7):e16473 [FREE Full text] [doi: 10.2196/16473] [Medline: 32706736]
113. Naslund JA, Deng D. Addressing mental health stigma in low-income and middle-income countries: a new frontier for digital mental health. Ethics Med Public Health 2021;19:100719 [FREE Full text] [doi: 10.1016/j.jemep.2021.100719] [Medline: 35083375]
114. Baños R, Herrero R, Vara MD. What is the current and future status of digital mental health interventions? Span J Psychol 2022 Feb 02;25:e5. [doi: 10.1017/SJP.2022.2] [Medline: 35105398]
115. Zhao X, Stadnick NA, Ceballos-Corro E, Castro J, Mallard-Swanson K, Palomares KJ, et al. Facilitators of and barriers to integrating digital mental health into county mental health services: qualitative interview analyses. JMIR Form Res 2023 May 16;7:e45718 [FREE Full text] [doi: 10.2196/45718] [Medline: 37191975]
116. Jardine J, Nadal C, Robinson S, Enrique A, Hanratty M, Doherty G. Between rhetoric and reality: real-world barriers to uptake and early engagement in digital mental health interventions. ACM Trans Comput-Hum Interact 2024 Feb 05;31(2):1-59 [FREE Full text] [doi: 10.1145/3635472]
117. Thornicroft G, Sunkel C, Aliev AA, Baker S, Brohan E, El Chammay R, et al. The Lancet Commission on ending stigma and discrimination in mental health. Lancet 2022 Oct 22;400(10361):1438-1480. [doi: 10.1016/S0140-6736(22)01470-2] [Medline: 36223799]
118. Martínez V, Espinosa-Duque D, Jiménez-Molina Á, Rojas G, Vöhringer PA, Fernández-Arcila M, et al. Feasibility and acceptability of "Cuida tu Ánimo" (take care of your mood): an internet-based program for prevention and early intervention of adolescent depression in Chile and Colombia. Int J Environ Res Public Health 2021;18(18):9628 [FREE Full text] [doi: 10.3390/ijerph18189628] [Medline: 34574553]
119. Maloney CA, Abel WD, McLeod HJ. Jamaican adolescents' receptiveness to digital mental health services: a cross-sectional survey from rural and urban communities. Internet Interv 2020 Sep;21:100325 [FREE Full text] [doi: 10.1016/j.invent.2020.100325] [Medline: 32455121]
120. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719 [FREE Full text] [doi: 10.2196/42719] [Medline: 36853742]
121. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 01;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
122. The state of mobile internet connectivity 2023. GSMA. 2023. URL: https://www.gsma.com/r/wp-content/uploads/2023/10/The-State-of-Mobile-Internet-Connectivity-Report-2023.pdf?utm_source=website&utm_medium=button&utm_campaign=somic23 [accessed 2024-04-03]
123. Madonsela S, Ware LJ, Scott M, Watermeyer J. The development and use of adolescent mobile mental health (m-mhealth) interventions in low- and middle-income countries: a scoping review. S Afr J Psychol 2023 Jul 08;53(4):471-483. [doi: 10.1177/00812463231186260]

Abbreviations

ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums
DHI: digital health intervention
DMHI: digital mental health intervention
GSM: Global System for Mobile
LMIC: low- and middle-income country
PCC: population, concept, and context
POD: identifying "Problems," generating "Options," and creating a "Do it" plan
PRIDE: Premium for Adolescents
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RCT: randomized controlled trial
SDG: Sustainable Development Goal

XSL•FO
RenderX

subRQ: subreview question
UHC: universal health coverage
UNICEF: United Nations International Children's Emergency Fund
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 29.07.23; peer-reviewed by SK Choi, S Lehtimaki, Y Ranjit; comments to author 05.11.23; revised version received 07.05.24; accepted 04.09.24; published 29.10.24.

Please cite as:
Wani C, McCann L, Lennon M, Radu C
Digital Mental Health Interventions for Adolescents in Low- and Middle-Income Countries: Scoping Review
J Med Internet Res 2024;26:e51376
URL: https://www.jmir.org/2024/1/e51376
doi:10.2196/51376
PMID:

©Carolina Wani, Lisa McCann, Marilyn Lennon, Caterina Radu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Provision of Digital Primary Health Care Services: Overview of Reviews

Virgínia Maria Dalfior Fava[1,2*], PhD; Luís Velez Lapão[2,3,4*], PhD

[1]Centro de Estudos Estratégicos Antonio Ivo de Carvalho, Fundação Oswaldo Cruz (Fiocruz), Ministério da Saúde, Rio de Janeiro, Brazil

[2]Intelligent Decision Support Systems Laboratory, Research & Development Unit for Mechanical and Industrial Engineering (UNIDEMI), NOVA School of Science and Technology, Universidade Nova de Lisboa, Caparica, Portugal

[3]Laboratório Associado de Sistemas Inteligentes (LASI), Escola de Engenharia, Universidade do Minho, Guimarães, Portugal

[4]WHO Collaborating Center on Health Workforce Policy and Planning, Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa, Lisboa, Portugal

[*]all authors contributed equally

Corresponding Author:
Virgínia Maria Dalfior Fava, PhD
Centro de Estudos Estratégicos Antonio Ivo de Carvalho
Fundação Oswaldo Cruz (Fiocruz)
Ministério da Saúde
Av. Brasil, 4036, 10º andar, sala 1001
Manguinhos
Rio de Janeiro, 21040-361
Brazil
Phone: 55 21 3882 9133
Email: virginiafava@gmail.com

Abstract

Background: Digital health is a growing field, and many digital interventions have been implemented on a large scale since the COVID-19 pandemic, mainly in primary health care (PHC). The development of digital health interventions and their application in PHC are encouraged by the World Health Organization. The increased number of published scientific papers on this topic has resulted in an overwhelming amount of information, but there is no overview of reviews to summarize this evidence.

Objective: This study aims to provide policy makers, health managers, and researchers with a summary of evidence on digital interventions used in PHC.

Methods: This overview of reviews searched the Web of Science and MEDLINE databases for systematic and scoping reviews on assessments of digital technologies implemented in PHC published from January 2007 to March 2023. Only reviews that addressed digital interventions whose targets were real patients or health care providers (HCPs) were included.

Results: A total of 236 records were identified from the search strategy, of which 42 (17.8%) full-text papers were selected for analysis, and 18 (7.6%) reviews met the eligibility criteria. In total, 61% (11/18) of the reviews focused their analysis on specific digital health interventions (client-to-provider telemedicine, provider-to-provider telemedicine, health worker decision support systems, systems for tracking patients' health status, client participation and self-care platforms, and provision of education and training to health workers), and 39% (7/18) of the reviews focused on specific topics related to PHC (preventive care, chronic disease management, behavioral health disorders, the COVID-19 pandemic, multicomponent PHC interventions, and care coordination). Most studies in the included reviews agreed on barriers to implementation, such as software and apps developed without involving end users, the lack of training of HCPs and patients in digital technology use, and the lack of reimbursement and billing strategies for remote consultations. However, they showed several mixed results related to health service quality and patients' clinical conditions and behavior changes.

Conclusions: Research in digital health applied to PHC is still concentrated in high-income countries, mainly in North America and Europe. The mixed results related to health service quality and patients' clinical conditions or behavior changes may have been caused by deficiencies in the process of implementing digital interventions. It is necessary to examine the entire impact pathway and the causal relationship among implementation, health service quality, and clinical condition outcomes to support the spread of digital health in PHC settings.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53594) doi:10.2196/53594

KEYWORDS

primary health care; digital health; implementation; health service quality; patients' clinical conditions; digital skills; mobile phone

Introduction

Background

Digital health is a growing field combining medical informatics, health care monitoring, and new business models that involves delivering health services and information through the internet and related technologies. The goal is to use information and communications technology to improve health care on a local, regional, and global scale [1]. Although many digital health software and apps have been developed and made available since the 2000s, only since the COVID-19 pandemic did these digital interventions begin to be implemented on a large scale, mainly in primary health care (PHC), highlighting their importance for health care [2,3].

Many studies have demonstrated the contributions of information and communications technologies to enhance health care. Digital health interventions have been effective in improving health behavior, clinical assessment, treatment compliance, and coordination of care [4]. Positive impacts of digital health interventions have been observed regarding improvements in glycated hemoglobin levels in patients diagnosed with type 1 and type 2 diabetes mellitus and prediabetes [5]; greater medication adherence and physical activity, reduced cardiovascular risks, and better detection of arrhythmias in cardiovascular disease care [6]; general health behavior engagement, such as physical activity, dietary behaviors, medication adherence, and sun protection practices [7]; quality of life, psychological outcomes, and screening behaviors in patients with cancer [8]; and mental health [9].

The current relevance of this topic at a global level is demonstrated by the launch of the new Global Initiative on Digital Health by the World Health Organization (WHO) at the Health Ministers' Meeting of the G20 Summit. The Global Initiative on Digital Health is a network of stakeholders to spread digital health, worldwide and within countries, through a comprehensive global ecosystem to promote country capacity and strengthen international cooperation in digital health [10]. The WHO has especially promoted the use of digital health in PHC settings. The Global Conference on Primary Health Care in Astana, Kazakhstan, in October 2018 encouraged the adoption of digital health in PHC settings to improve health care access and quality, patient safety, and care coordination in addition to empowering individuals and communities to take an active role in maintaining their overall health and well-being [11].

Digital technologies can improve PHC in many ways: integrating clinical support tools and referral systems into PHC to help coordinate care and ensure its continuity; supporting and building capacity among the health workforce; supporting patients in home settings through access to personalized information, appointment booking, and tools to manage their chronic conditions; making point-of-care diagnostic testing available for rapid analysis; and identifying risks and reducing harm to guarantee patient safety [12]. In addition, PHC has been critical for three main reasons: (1) it presents features that allow the health system to adapt and respond to a complex and rapidly changing world, (2) it emphasizes promotion and prevention with a people-centered approach that addresses the causes and risk factors of poor health, and (3) it has contributed to universal health coverage and the health-related Sustainable Development Goals. This is possible because PHC has an approach focused on comprehensive care, which includes promotive, protective, preventive, curative, rehabilitative, and palliative care throughout the life course, which addresses the determinants of health in a broad way (social, economic, and environmental factors, as well as individual characteristics and behavior) and seeks to empower individuals, families, and communities to optimize their health [13].

Recent bibliometric studies of scientific production on eHealth or digital health have confirmed the increase in publications on this topic in the last few years, mainly in North America and Europe but also in Australia, China, and India [14-17]. The increased number of published scientific papers on digital health has resulted in an overwhelming amount of information; however, most studies are not actual implementations in the real patient environment [18]. To make this information more accessible and usable, researchers have been summarizing it into review articles. They have also conducted overviews of reviews to address broader questions that go beyond the scope of individual reviews [19]. Policy makers play a crucial role in promoting the implementation of digital technologies in PHC by prioritizing public investment in building the physical infrastructure; deploying information systems and apps; developing a capable health workforce; ensuring a sound legal and regulatory environment; and improving governance, policy, standardization, and interoperability [12].

Objectives

Studies that have conducted overviews of reviews on digital health have addressed many questions focused on technology implementation [20], impacts on health care services [21,22], specific clinical conditions such as cancer [8] and mental health [23], or specific types of digital interventions such as mobile apps [24]. However, there is no overview of reviews to inform policy makers, health managers, and researchers on digital intervention outcomes on PHC. Considering the importance of PHC for global health and the advances in research on the use of digital technologies in PHC settings, this overview of reviews aimed to summarize the evidence on digital interventions applied to PHC to inform policy makers, health managers, and researchers.

XSL•FO
RenderX

Methods

Search Strategy

An overview of reviews was conducted through an electronic literature search of review articles focused on digital technologies applied to PHC, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist [25] (Multimedia Appendix 1). We searched the Web of Science Core Collection and PubMed/MEDLINE databases from the National Library of Medicine. Unpublished and gray literature were not pursued.

The search was conducted in March 2023 and was based on a query with 2 parts (Multimedia Appendix 2): the first one covered topics referring to digital technologies applied to health (ie, *digital health*, *e-health*, *medical informatics*, *health informatics*, and *health IT*), and the second one included topics associated with PHC (ie, *primary care*, *primary healthcare*, and *primary health care*). The review articles had to be published since 2007 and written in English or Portuguese. Duplicated records were excluded.

Eligibility Criteria

The included reviews were required to meet the eligibility criteria established by the population, intervention, comparator, outcome, and study design (PICOS) framework formulated for this study. Textbox 1 outlines the eligibility criteria according to the 5 PICOS components [26].

The digital interventions analyzed by the included reviews were classified according to WHO recommendations [27] into 8 types of interventions, described in Textbox 2.

In total, 2 authors screened the titles and abstracts and excluded studies that did not meet the eligibility criteria. Disagreements were resolved through discussion and consensus between the authors. Full-text versions of the remaining articles were obtained and screened, and papers were also excluded if digital health was not integrated into PHC services or if other settings were included without results presented by level of care. The reasons for exclusion were recorded during the 2 steps (title or abstract and full-text screening).

Textbox 1. Population, intervention, comparator, outcome, and study design (PICOS) framework formulated for this study.

PICOS component and eligibility criteria

- Participants: any patient or health care provider in primary health care
- Intervention: digital interventions implemented in primary health care settings whose targets were real patients or health workers
- Comparisons: no treatment, treatment as usual, or different types of digital technologies
- Outcomes: implementation, health service quality, or patients' clinical conditions or behavior change outcomes
- Study design: systematic or scoping reviews

Textbox 2. Types of digital health interventions selected for this study.

Digital health interventions

- Client-to-provider telemedicine: delivery of health care services in which clients or patients and health workers are separated by distance
- Provider-to-provider telemedicine: delivery of health care services in which ≥2 health workers are separated by distance
- Targeted client communication: transmission of customized health information for different audience segments based on health status or demographic categories
- Health worker decision support: digitized job aids that combine an individual's health information with the health worker's knowledge and clinical protocols to assist health workers in making diagnosis and treatment decisions
- Tracking of patients' health status: digitized records used by health workers to capture and store health information on clients or patients to follow up on their health status and services received
- Provision of education and training to health workers: the management and provision of education and training content in digital form for health professionals, which does not need to be used at the point of care
- Client participation and self-care: digitized records used by patients to capture and monitor their health information and other digital tools for patient communication (client-to-client communication, access by the client to their own medical records, self-monitoring of health or diagnostic data by the client, and active data capture or documentation by the client)
- Laboratory and diagnostics imaging management: digital approaches to manage and exchange laboratory and diagnostic orders and results (transmit diagnostic results to HCPs, transmit and track diagnostic orders, and capture diagnostic results from digital devices)

Data Extraction and Synthesis

The data were managed and stored using Zotero (Corporation for Digital Scholarship) and Excel (Microsoft Corp). Extracted data used to categorize review papers included publication year, authors, title, journal information (title and ISSN), abstract, and language. Additional information was extracted independently by 2 authors from each review paper, including study design, number of studies included, number of databases searched, classification of digital health interventions according to the

XSL•FO
RenderX

WHO recommendations [27], intervention target, outcome category according to the PICOS framework, and a summary of the results.

Risk-of-Bias Assessment

The quality of the evidence in responding to the research questions was assessed independently by 2 authors using the Joanna Briggs Institute Critical Appraisal Checklist for Systematic Reviews tool [28]. The 2 authors discussed the results of the quality appraisal, reaching a consensus in case of any divergence. Question 9 was not assessed because all the reviews summarized qualitative evidence.

Some criteria in the risk-of-bias assessment of scoping reviews were adapted considering the guidelines for preparing this type of study [29]. Question 1 was considered met if the review included clear and explicit statements about the objectives even if it did not explicitly present a research question. Questions 5 and 6 were not evaluated as conducting a critical appraisal of the included sources of evidence is not mandatory.

Data Analysis

A narrative synthesis of the main results from the included reviews was conducted to identify the main types of digital technologies applied to PHC based on the classification of digital health interventions proposed by the WHO [27], which health problems and conditions were addressed, and the results of their implementation from the outcome categories in Textbox 1 (implementation outcomes, health service outcomes, and patients' clinical conditions and behavior change outcomes).

Results

Search Results

A total of 236 records were identified from the search strategy, of which 68 (28.8%) were excluded before analysis because of the publication date or language or for being duplicates. In the title and abstract screening, 75% (126/168) of the records were excluded because of the eligibility criteria. All 42 remaining integral text records were retrieved, and 24 (57%) were excluded after screening. The remaining 18 reviews met the eligibility criteria and were included in the synthesis. Figure 1 shows the PRISMA flow diagram for the total articles assessed.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

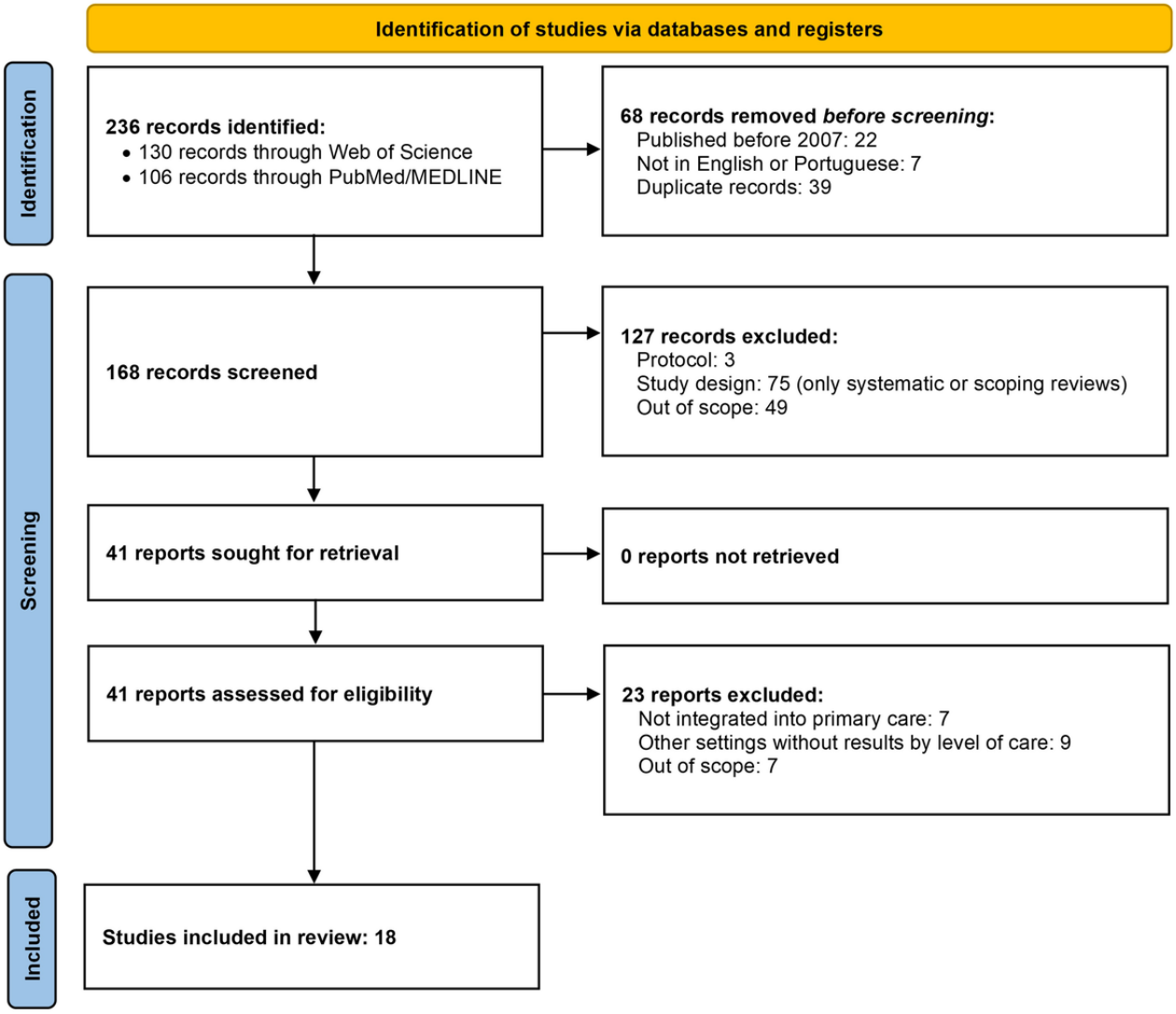

XSL•FO
RenderX

Review Characteristics

Table 1 presents the general characteristics of the reviews, including publication year, study design, number of studies included, number of databases searched, classification of digital health interventions, intervention target, and outcome category. All studies were published in English.

All reviews were published from 2015 onward, with 67% (12/18) published between 2021 and 2023. Of the 18 reviews, 11 (61%) were systematic reviews, and 7 (39%) were scoping reviews. A total of 72% (13/18) of the studies searched for articles in up to 5 databases. The scoping reviews included an average of 35.3 (SD 11.7) reports (outliers excluded: 241), and the systematic reviews included an average of 13.0 reports (outliers excluded: 43 and 48).

Table 2 provides the risk-of-bias assessment details. All systematic reviews except for one (10/11, 91%) presented good methodological quality, meeting at least 8 of the 10 criteria analyzed. Although all systematic reviews conducted critical appraisal to assess the included studies, that analysis was carried out by ≥2 independent reviewers only in 55% (6/11) of them. Similarly, most scoping reviews (5/7, 71%) presented a good methodological approach. The least attended item refers to the appropriate extraction of data, which was not always carried out by 2 independent researchers.

XSL•FO
RenderX

Table 1. General characteristics of the included reviews.

Review	Year	Study design	Databases, n	Included studies, n	Digital health intervention	Outcome categories
Hickson et al [30]	2015	Scoping review	2	24	Client-to-provider telemedicine	Implementation; health services
Beheshti et al [31]	2022	Scoping review	4	43	Client-to-provider telemedicine	Implementation; clinical conditions or behavior changes
Kuo and Dang [32]	2016	Systematic review	1	11	Client-to-provider telemedicine; client participation and self-care	Clinical conditions or behavior changes
Liddy et al [33]	2019	Systematic review	4	43	Provider-to-provider telemedicine	Implementation; health services; clinical conditions or behavior changes
Meunier et al [34]	2023	Systematic review	5	48	Health worker decision support	Implementation
El Asmar et al [35]	2021	Systematic review	6	8	Health worker decision support	Clinical conditions or behavior changes
Agarwal et al [36]	2021	Systematic review	7	8	Health worker decision support	Implementation; health services; clinical conditions or behavior changes
Peyroteo et al [18]	2021	Systematic review	2	18	Tracking of patients' health status	Implementation
Tuan Soh et al [37]	2022	Systematic review	5	13	Client participation and self-care	Implementation
Andrikopoulou et al [38]	2019	Systematic review	7	15	Client participation and self-care	Clinical conditions or behavior changes
Kyaw et al [39]	2019	Systematic review	9	8	Provision of education and training to health workers	Health services; clinical conditions or behavior changes
Willis et al [40]	2022	Scoping review	3	241	Client-to-provider telemedicine; targeted client communication; health worker decision support; tracking of patients' health status; client participation and self-care	Implementation; health services
Xiong et al [41]	2023	Scoping review	3	52	Client-to-provider telemedicine; targeted client communication; health worker decision support; client participation and self-care; provision of education and training to health workers; laboratory and diagnostics imaging management	Health services; clinical conditions or behavior changes
Moon et al [42]	2022	Scoping review	4	20	Client-to-provider telemedicine; targeted client communication; health worker decision support; tracking of patients' health status; client participation and self-care; provision of education and training to health workers	Implementation; clinical conditions or behavior changes
Jonnagaddala et al [43]	2021	Scoping review	4	29	Client-to-provider telemedicine; client participation and self-care; provision of education and training to health workers	Implementation
Silva et al [44]	2022	Scoping review	17	44	Client-to-provider telemedicine; targeted client communication; client participation and self-care; provision of education and training to health workers	Implementation; health services

XSL•FO
RenderX

Review	Year	Study design	Databases, n	Included studies, n	Digital health intervention	Outcome categories
Jimenez et al [45]	2021	Systematic review	2	14	Client-to-provider telemedicine; provider-to-provider telemedicine; targeted client communication; health worker decision support; tracking of patients' health status; client participation and self-care	Implementation; clinical conditions or behavior changes
Maillet et al [46]	2018	Systematic review	5	22	Health worker decision support; tracking of patients' health status; client participation and self-care; laboratory and diagnostics imaging management	Implementation; health services

Table 2. The methodological quality of the reviews based on the Joanna Briggs Institute Critical Appraisal Checklist.

	Q1[a]	Q2[b]	Q3[c]	Q4[d]	Q5[e]	Q6[f]	Q7[g]	Q8[h]	Q9[i]	Q10[j]	Q11[k]
Systematic reviews											
Kuo and Dang [32]	Yes	No	Yes	No	Yes	No	No	Yes	__[l]	Yes	Yes
Liddy et al [33]	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	—	Yes	No
Meunier et al [34]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	—	Yes	Yes
El Asmar et al [35]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes	Yes
Agarwal et al [36]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes	Yes
Peyroteo et al [18]	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	—	Yes	Yes
Tuan Soh et al [37]	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	—	Yes	Yes
Andrikopoulou et al [38]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes	Yes
Kyaw et al [39]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes	Yes
Jimenez et al [45]	Yes	Yes	Yes	No	Yes	No	Yes	Yes	—	Yes	Yes
Maillet et al [46]	Yes	Yes	Yes	Yes	Yes	No	No	Yes	—	Yes	Yes
Scoping reviews											
Hickson et al [30]	Yes	Yes	Yes	No	—	—	No	Yes	—	Yes	Yes
Beheshti et al [31]	Yes	No	Yes	Yes	—	—	Yes	No	—	Yes	Yes
Willis et al [40]	Yes	Yes	Yes	Yes	—	—	Yes	Yes	—	Yes	Yes
Xiong et al [41]	Yes	Yes	Yes	Yes	—	—	Yes	Yes	—	Yes	Yes
Moon et al [42]	Yes	Yes	Yes	Yes	—	—	No	Yes	—	Yes	Yes
Jonnagaddala et al [43]	Yes	Yes	Yes	Yes	—	—	Yes	Yes	—	Yes	Yes
Silva et al [44]	Yes	Yes	Yes	Yes	—	—	Yes	Yes	—	Yes	Yes

[a]Question 1: Is the review question clearly and explicitly stated?
[b]Question 2: Were the inclusion criteria appropriate for the review question?
[c]Question 3: Was the search strategy appropriate?
[d]Question 4: Were the sources and resources used to search for studies adequate?
[e]Question 5: Were the criteria for appraising studies appropriate?
[f]Question 6: Was critical appraisal conducted by ≥2 reviewers independently?
[g]Question 7: Were there methods to minimize errors in data extraction?
[h]Question 8: Were the methods used to combine studies appropriate?
[i]Question 9: Was the likelihood of publication bias assessed?
[j]Question 10: Were recommendations for policy or practice supported by the reported data?
[k]Question 11: Were the specific directives for new research appropriate?
[l]Not applicable.

XSL•FO
RenderX

Types of Digital Health Interventions Applied to PHC

Overview

A total of 11 review papers analyzed 1 specific type of digital health intervention applied to PHC—(1) 3 (27%) reviewed client-to-provider telemedicine interventions, (2) 1 (9%) reviewed provider-to-provider telemedicine interventions, (3) 3 (27%) reviewed health worker decision support interventions, (4) 1 (9%) reviewed interventions tracking patients' health status, (5) 2 (18%) reviewed client participation and self-care interventions, and (6) 1 (9%) reviewed interventions providing education and training to health workers (Textbox 3).

Textbox 3. Review papers according to the type of digital health intervention analyzed.

Client-to-provider telemedicine

- Hickson et al [30]
- Beheshti et al [31]
- Kuo and Dang [32]

Provider-to-provider telemedicine

- Liddy et al [33]

Health worker decision support

- Meunier et al [34]
- El Asmar et al [35]
- Agarwal et al [36]

Tracking patients' health status

- Peyroteo et al [18]

Client participation and self-care interventions

- Tuan Soh et al [37]
- Andrikopoulou et al [38]

Provision of education and training to health workers

- Kyaw et al [39]

Client-to-Provider Telemedicine

A total of 29% (2/7) of the scoping reviews analyzed the literature on the state of client-to-provider telemedicine with a focus on nonurgent PHC practice. One review focused on web-based consultations using mobile devices [30], and the other included all remote health services delivered via real-time communication between the patient and the HCP [31]. Both reviews agreed that this kind of intervention allows for patient follow-up and clinical condition monitoring and aims to provide quick and easy access to meet medical needs.

In addition, a systematic review examined the evidence supporting the use of secure messages associated with other electronic health record (EHR) resources, such as patient portals and personal health records (PHRs; client participation and self-care interventions), focused on diabetes management. Secure messaging was defined as "any electronic communication between a HCP and patient that ensures only those parties can access the communication" [32].

Beheshti et al [31] showed that telehealth has spread in many countries, including papers that analyzed interventions in Europe (half of them, ie, 55.8% from the United Kingdom), United States (34.9%), Asia (2 papers), and Brazil and Zambia (1 paper each). Most studies analyzed in this review focused on investigating the effects of these services on rural and underserved areas.

However, the concentration of studies in Europe and the United States may be associated with the various implementation barriers cited by Hickson et al [30] and Beheshti et al [31]. As obstacles to the implementation of client-to-provider interventions, they cited infrastructure barriers such as lack of security to transmit confidential information, low bandwidth, unsuitable networks, low signal quality, and bad image quality; system barriers such as the lack of integration with electronic medical records, workflow, and other existing systems; legal barriers such as professional regulation when patients and HCPs are in different states or countries; and billing and reimbursement barriers to guarantee the monetary incentives for HCPs to deliver remote health services. These studies also presented patient barriers, such as confidentiality and privacy concerns and the demand for supervisor support, and HCP barriers, such as reluctance toward new approaches to health care delivery, concerns about the absence of patient health literacy, the perception that telehealth technology is unintuitive and inflexible, and the fear of increased workload.

XSL•FO
RenderX

Despite the implementation difficulties, Hickson et al [30] presented studies that showed improvements in health service outcomes. Using remote health services through mobile devices resulted in cost savings for acute care delivery by decreasing in-office spending per member per month and by preventing face-to-face visits in settings that applied web-based consultation reimbursements to a fee-for-service model of patient payment. Beheshti et al [31] presented only 1 study that indicated clinical condition improvements in patients with diabetes with lower blood sugar levels using telemedicine, in addition to other benefits such as better access to health care, less waste of time, and a high degree of satisfaction with health care. In addition, Kuo and Dang [32] reviewed 11 studies, of which 7 suggested a significant improvement in the glycemic control of patients with diabetes (primary outcome) resulting from the use of secure messaging, but the impact on other outcomes such as cholesterol and blood pressure was inconsistent. A total of 10 of the studies included in this review were conducted in North America.

The 2 scoping reviews also specified that there are several communication tools for delivering remote health care but asynchronous technologies (eg, emails and SMS text messages) seemed to be supplementary and did not reduce the frequency of face-to-face visits. The synchronous form is the most dominant approach for delivering telehealth services (eg, telephone calls, videoconferences, and teleconferences). Finally, despite the identified benefits of client-to-provider telemedicine interventions cited by these 2 reviews, such as cost reduction, decreased face-to-face visits, patients' time savings, and health care access and clinical condition improvements, the 2 scoping reviews pointed out that more research is necessary to replicate and expand findings on access to health care and patients' clinical conditions to identify impacts on workflow efficiency and cost-effectiveness depending on billing and reimbursement strategies [30,31].

Provider-to-Provider Telemedicine

A systematic review examined the impact of electronic consultations on the delivery of care, applying the quadruple aim framework that analyzes the effects on population health, the experience of receiving care (patients), the experience of providing care (HCPs), and per-capita costs [33]. This review defined an electronic consultation (e-consult) as an electronic communication tool that allows PHC HCPs to obtain a specialist consultant's expert opinion promptly.

Most studies included in the systematic review were from the United States (19/43, 44%) and Canada (12/43, 28%). In terms of implementation outcomes, e-consults had sustained use and spread, high adoption, and little pushback from HCPs. The range of specialties accessed by PHC HCPs through e-consults expanded beyond dermatology, including multispecialty services that comprise endocrinology, hematology, cardiology, gastroenterology, and neurology.

The results presented by this systematic review showed that most PHC HCPs reported being satisfied with e-consults, they intended to use this technology in the future because their questions were answered, and the specialty support was conclusive without impact on the workload. However, there were still some challenges, such as unclear directions from specialists to PHC HCPs, lack of information or pertinent questions delivered to specialists by PHC HCPs, lack of patient follow-up by the specialists, and a potential increase in workload for specialists.

Regarding health service outcomes, Liddy et al [33] pointed out that e-consults are effective in faster access to specialist advice and substantial avoidance of face-to-face visits and showed that the waiting time for a specialist response was less than that for traditional referrals. Related to patients' clinical condition effects, they indicated a potential for reducing adverse cardiovascular outcomes, but there was a potential for harm due to the lack of patient follow-up.

The authors concluded that there are gaps in the evidence on the implementation of provider-to-provider digital interventions relative to specialists' workload and patients' perceptions. Furthermore, the divergent findings related to impacts on cost-effectiveness and clinical conditions indicate the need for more research in this area.

Health Worker Decision Support

In total, 27% (3/11) of the systematic reviews analyzed health worker decision support interventions [34-36]. All reviews examined aspects related to computerized clinical decision support systems (CDSSs), any software designed to offer patient-specific assessments and recommendations to the HCP based on patient data from EHRs and evidence-based clinical practice guidelines. One review sought to identify and quantify the barriers to and facilitators of the use of CDSSs by PHC HCPs [34]. The second one provided a summary of the effects of CDSS use on the clinical outcomes of adult patients with chronic diseases managed in PHC [35]. The third one focused their analysis on CDSSs accessible via mobile devices by PHC HCPs [36]. Most studies included in these 3 reviews were from North America (United States and Canada), Europe, and Australia, with a few studies from Asia, Africa, and South and Central America.

Meunier et al [34] assessed the implementation of CDSSs using the human, organization, technology, net benefits framework, which comprises the interdependent human, organizational, and technological factors related to health information system adoption in PHC. In the human category, they found that the HCPs' satisfaction and their perceived usefulness of the CDSS recommendations could facilitate the use and adoption of this digital technology. However, when the CDSS offers information overload and recommendations conflicting with HCPs' expertise, it might result in the HCPs' resistance, reluctance, and negative attitude toward CDSS use. Furthermore, providing training in computer skills and software use could be a human facilitator. The main factors related to the organizational category are workflow and teamwork. When the CDSS is integrated into the EHR and the clinical workflow, producing quality reports and reducing workload through the expansion of the skill sets and roles of other health professionals in assisting physicians, it could facilitate the use of this digital technology. Conversely, the workflow disruption caused by CDSS implementation, demanding more teamwork and increasing the workload, might become a barrier. Finally, in the technological category, the ease of use, visual layout, usefulness

XSL•FO
RenderX

of system features and functions, full integration with EHR databases (including data on specialty care), and technical support were pointed out as factors that could influence the use of the CDSS. However, the lack of computers and tablets and the slowness and poor performance of the software were identified as barriers to CDSS use. Agarwal et al [36] sought to verify the HCPs' acceptability of and satisfaction with mobile CDSSs, but they did not find direct evidence of this implementation outcome.

Related to health service outcomes, Agarwal et al [36] reviewed papers that analyzed the HCPs' adherence to recommended practices, guidelines, or protocols and the time between clinical recommendation presentation and appropriate management only for mobile CDSSs. The results were inconclusive for the first outcome because the certainty of evidence was very low, and no evidence was identified for the second outcome.

El Asmar et al [35] and Agarwal et al [36] searched for CDSS use impact on patients' clinical conditions or behavior changes. The quality of evidence of the studies included in both reviews was rated as moderate, low, or very low. El Asmar et al [35] focused on clinical outcomes of the management of chronic diseases and presented data on diabetes mellitus, asthma, angina, hyperlipidemia, and hypertension. Most results indicated no statistically significant differences between CDSS intervention and control participants (usual care or baseline), and it is important to highlight that the study that showed a statistically significant impact on blood pressure between the intervention and control groups revealed extensive training for participants and widespread use by HCPs. Agarwal et al [36] reported similar results for clinical outcomes of chronic disease management—mobile CDSS use made little or no difference in clinical conditions for people with high cardiovascular disease risk, poorly controlled diabetes, or hyperlipidemia. The authors still presented inconclusive results for the impact on maternal and neonatal health outcomes because of the uncertainty of the evidence. However, they pointed out that the use of mobile CDSSs by community workers may increase medication adherence in people with high cardiovascular disease risk. However, the use of this technology by physicians made little or no difference in people with poorly controlled diabetes.

Tracking Patients' Health Status

A systematic review analyzed the literature on remote health monitoring systems for chronic disease management in PHC settings [18]. This type of technology is classified as an intervention tracking patients' health status. The included studies focused on interventions that automatically monitor patient indicators (sensors or wearables) continuously or those in which patients must collect and transmit personal electronic health information in real time or in an asynchronous way. Most studies included were from Europe and the Americas, with a few studies from Asia and Africa.

The studies included in this review indicated that there is a need for more research in this area. Most of the articles presented implementation processes as a proof of concept, pilot study, or clinical trial without evidence of the impact on actual health services or patients' clinical conditions or behavior change outcomes. The authors showed some barriers to the intervention's large-scale implementation, such as the low involvement of health workers, patients, and other stakeholders in the tool design and implementation. The low adherence by HCPs was explained by the lack of technology integration into the existing workflow, limited integration into the information systems already used, and the divergences between the data presented by the remote monitoring systems and the HCPs' expertise in chronic disease management.

Peyroteo et al [18] concluded that it is necessary to fill in the implementation gaps mentioned to ensure the correct and large-scale use of the digital tool. This is the only way to assess the impacts of interventions tracking patients' health status on health services and patients' clinical conditions or behavior change outcomes.

Client Participation and Self-Care

In total, 18% (2/11) of the systematic reviews examined client participation and self-care interventions in PHC. One aimed to identify the benefits of using electronic PHRs based mainly on studies that assessed the patients' and HCPs' perceptions of the implementation of this technology [37]. The other review focused on the PHR features that improve medication adherence in the adult population with chronic diseases [38]. Both reviews defined PHRs as a tool that provides patients with their health care and personal medical information, and most included studies were conducted in North America or Europe.

Regarding implementation outcomes, Tuan Soh et al [37] found evidence indicating users' favorable opinions on using PHR systems. The positive effects pointed out by these authors were on patient empowerment; improvement in self-care management, communication, and the relationship between patients and HCPs; and improvement in quality of care by supporting shared decision-making based on shared data, responsibility for health information that could be updated anytime by the patient, in addition to supporting interventions tracking patients' health status and promoting continuity between the different levels of care. However, the authors indicated that more research is needed to understand structural, organizational, and policy barriers to the implementation of PHR systems and the challenges to ensure data accuracy and reliability.

Andrikopoulou et al [38] focused their review on the analysis of one specific patient behavior change outcome: medication adherence. In total, 13 of the 15 studies included reported positive impacts on medication adherence, and only 2 found no difference. This positive tendency was found despite the heterogeneity in participants' demographics, study location, chronic disease, or medication adherence measurement method. The authors supposed that these results were due to specific PHR features such as reminders, feedback, and alerts on medication management as all studies included at least one of these features. Finally, they highlighted some gaps that demand more research, such as considering the particularities of patients with multimorbidity or polypharmacy when designing PHR resources to guarantee medication adherence.

Provision of Education and Training for Health Workers

A systematic review analyzed the provision of education and training to health workers on antibiotic management in PHC to

XSL•FO
RenderX

summarize evidence on the effectiveness of digital education compared to traditional education for improving clinical practice [39]. For this reason, there were no reports of implementation outcomes, such as human, organizational, or technological barriers and facilitators. All included studies were from Europe and North America, except for one that was conducted in China.

Regarding health services, the authors presented results of specific knowledge outcomes, clinical practice improvements, and physicians' attitudes toward the interventions. They found that digital education (mobile phone or PDA device with internet access) may improve knowledge scores measured using questionnaires compared to traditional education. Some forms of digital education (eg, email with feedback and suggestions and web-based learning resources) showed an impact on reducing antibiotic prescribing or dispensing rates in 4 studies included in the review, and another 2 studies indicated no difference between digital and traditional education when using SMS text messaging or web-based blended education. These results seemed to be more relevant when digital education was combined with web-based training on point-of-care diagnostic tests and enhanced communication skills. Finally, one study reported that almost all physicians were satisfied with the intervention and wanted to continue receiving clinical recommendations about antibiotic prescriptions.

In total, 4 studies included in the review by Kyaw et al [39] assessed patients' clinical conditions measured using reconsultation rate, probability of consultation for a new respiratory tract infection, hospital referral, and postintervention hospital admission rate. In summary, these studies reported little or no difference between digital and traditional education. However, it can be assumed that the real impact of digital education on antibiotic prescription may be on antibiotic resistance.

Health Issues Addressed by Digital Health Interventions in PHC Settings

Overview

A total of 28% (5/18) of the review papers searched for all types of digital interventions applied to six specific health issues in PHC settings: (1) preventive care, (2) chronic disease management, (3) behavioral health disorders, (4) the COVID-19 pandemic, (5) multicomponent PHC interventions, and (6) care coordination (Textbox 4).

Textbox 4. Review papers according to the type of digital health intervention and the topic related to primary health care analyzed.

Preventive care

- Willis et al [40]

Chronic disease management

- Xiong et al [41]

Behavioral health disorders

- Moon et al [42]

COVID-19 pandemic

- Jonnagaddala et al [43]
- Silva et al [44]

Multicomponent primary health care interventions

- Jimenez et al [45]

Care coordination

- Maillet et al [46]

Preventive Care

A scoping review examined the digital interventions used in the US PHC services to enhance and support the delivery of primary, secondary, tertiary, and quaternary preventive care [40]. Most digital interventions identified in this review focused on electronic medical records, with alerts to help health professionals make appropriate clinical decisions (interventions for tracking patients' health status and health worker decision support), or mobile interventions to improve patients' self-care (client participation and self-care interventions). In addition to these types, the authors identified the use of client-to-provider telemedicine and targeted client communication interventions to promote preventive care.

Most studies included in this review showed statistically significant improvements in primary prevention health services and implementation outcomes (adolescent and adult vaccination); secondary prevention health service outcomes (screening rates for colorectal cancer, hepatitis C virus, and osteoporosis but not for colonoscopy); and tertiary prevention health service and clinical condition or behavior change outcomes for managing diabetes, hypertension, asthma, obesity, cardiovascular disease, hyperlipidemia, and mental conditions. There were no studies included on digital interventions in quaternary prevention. Finally, the authors concluded that digital

XSL•FO
RenderX

interventions enhanced PHC by augmenting the comprehensiveness of the care provided but there are still gaps in the evaluations of mobile interventions [40].

Chronic Disease Management

A scoping review summarized the evidence on digital interventions for chronic disease management in PHC in low- and middle-income countries (10 studies from South America, 9 from Asia, and 2 from Africa) [41]. The review identified digital interventions of almost all types: client-to-provider telemedicine, targeted client communication, health worker decision support, client participation and self-care, provision of education and training to health workers, and laboratory and diagnostics imaging management. However, most digital health interventions still used SMS text messaging and smartphone apps for communication, making them easy to use and affordable as they only required cell phones and internet access.

This review indicated consistent positive results for implementation outcomes, such as improvements in accessibility, patient perception of the health service delivery, HCPs' capabilities, and better cost-effectiveness. However, the clinical conditions and behavior change outcomes were highly mixed, with positive, negative, neutral, or nonsignificant results. In addition, the authors found that communication-related digital interventions (client-to-provider telemedicine and targeted client communication interventions) produced more consistent improvements in clinical conditions or behavioral and implementation outcomes but interventions focused on HCPs' capacities (interventions for health worker decision support and provision of education and training to health workers) resulted in substantially mixed clinical condition or behavior change outcomes. These authors defend that research with a larger number of participants followed over a longer time may result in more consistent evidence of the impact of digital interventions [41].

Behavioral Health Disorders

A scoping review was conducted to gather and analyze research on how digital health services can enhance the collaborative care model, which addresses behavioral health disorders in PHC settings [42]. The studies included presented digital health interventions of different types: client-to-provider telemedicine, targeted client communication, health worker decision support, tracking of patients' health status, client participation and self-care, and provision of education and training to health workers.

Regarding implementation outcomes, both patients and HCPs had positive perceptions of these digital interventions, which facilitated the communication between users and freed up HCP time for patients. Despite these positive outcomes, privacy and workload concerns arose when the technology was not fully integrated into the clinical workflow, resulting in low user engagement and high dropout rates. The improvements in clinical conditions or behavior change outcomes presented in the included studies were modest, partly due to methodological factors such as small sample sizes and different control groups [42].

COVID-19 Pandemic

A total of 29% (2/7) of the scoping reviews explored the use of digital interventions in PHC settings during the COVID-19 pandemic. One focused on the Australian PHC response to the pandemic and found interventions on client-to-provider telemedicine, client participation and self-care, and education and training for health workers [43]. The other was a more extensive review including 18 studies conducted in North America, 14 conducted in Europe, 4 conducted in South America, 4 conducted in Asia, and 4 conducted in Oceania; they identified targeted client communication interventions in addition to the digital technologies found in the other review [44]. Both reviews concluded that these digital technologies were implemented to manage COVID-19 as well as other common diseases in PHC settings.

The 2 reviews emphasized that the key facilitators in implementing digital technologies during the COVID-19 pandemic were providing technology training and support for HCPs; offering validated and reliable technologies with a user-centered design that were suitable for the local context; and ensuring user security, acceptability, and satisfaction. However, there were implementation barriers, such as lack of digital facilities in remote and rural areas, safety and privacy concerns, lack of planning for health services and clinical protocols when using digital tools, and the need for new funding arrangements that reimburse for remote services. Although digital technologies during the pandemic had particular importance for chronic condition management, reducing the number of referrals to specialties and hospitalizations and speeding up screening in cases of clinical changes through remote tracking, these tools were not always effective in identifying complex or urgent demands and assisting older adults and children [43,44].

According to Silva et al [44], digital health emerged as the sole substitute for PHC during the COVID-19 pandemic. However, the decrease in the quality of health care could be attributed to the hasty implementation of remote services in not properly equipped facilities and by health professionals who were under immense pressure.

Multicomponent PHC Interventions

A systematic review explored how digital health has enhanced multicomponent PHC interventions, defined as initiatives comprised of several features aimed at enhancing PHC [45]. The review identified 6 types of digital technologies implemented in this context: client-to-provider telemedicine, provider-to-provider telemedicine, targeted client communication, health worker decision support, tracking of patients' health status, and client participation and self-care. Most studies included were from the United States and Europe, except for one from Argentina.

In terms of implementation outcomes, this review showed consistent findings for cost savings, decreased health care expenses, and improved perceptions among HCPs regarding the quality of the workplace and services offered. However, the results were mixed for other outcomes, such as hospital admissions, readmissions, emergency department visits, the

XSL•FO
RenderX

number and cost of drug prescriptions, and patients' perception (increased satisfaction with timely appointments and self-management support but not with PHC performance). Regarding clinical conditions or behavior change outcomes, there were consistent improvements in blood pressure and glycated hemoglobin control and reduction in amputation, smoking, and the number of patients with end-stage renal disease, but this was not the case for outcomes related to cardiovascular diseases [45].

Care Coordination

A systematic review analyzed how digital interventions have supported the laboratory testing process in PHC considering all its phases, from the decision process on the need to carry out a diagnostic test to the interpretation and communication of the test result, contributing to care coordination [46]. The review identified 4 types of digital interventions: health worker decision support, tracking of patients' health status, client participation and self-care, and laboratory and diagnostics imaging management. Most studies included in this review were from the United States, with some conducted in Europe and Peru, and most of them focused on health service outcomes, but some implementation outcomes were related to health professionals' perceptions.

The studies included in this review indicated that digital interventions related to tracking patients' health status and health worker decision support resulted in reducing the amount of effort spent by the staff on collecting information and promoting best care practices but not necessarily reducing the number of diagnostic tests prescribed in the preanalytical phases. The digital laboratory and diagnostics imaging management interventions seemed to impact health care positively in all phases of the testing process, reducing professional workload and the number of errors when the technology was integrated into ambulatory systems and saving time to order prescriptions, reorganize tasks in the intra-analytical phase, and return the test results. Despite this positive impact on health services, in some studies, health professionals indicated that these systems need to be improved in terms of design and reliability, and technical problems and lack of integration between systems were associated with negative effects on the workflow and increased workload. Finally, client participation and self-care interventions to directly communicate the test results to the patient contributed to better collaboration between patients and HCPs and promoted the self-management of chronic diseases [46].

Discussion

Principal Findings

This overview of systematic and scoping reviews on digital health technologies applied to PHC revealed the types of technology and the health issues addressed by these technologies that are currently most in evidence in PHC settings. The included reviews focused on technologies related to interventions on client-to-provider telemedicine [30-32], provider-to-provider telemedicine [33], health worker decision support [34-36], tracking patients' health status [18], client participation and self-care [37,38], and provision of education and training for health workers [39]. There was an expectation that digital health contributes to the improvement of preventive care [40], chronic disease management [41], behavioral health disorder care [42], and delivery of PHC services in pandemic contexts [43,44]. Furthermore, the included reviews indicated the contribution of digital technologies to processes that require coordinating the activities of health workers [45] or different facilities for the provision of health care [46].

Client-to-provider and provider-to-provider telemedicine interventions, previously designed to meet demands from remote or underserved areas, began to be widely used during the COVID-19 pandemic. Social distancing policies spread the implementation of this type of digital intervention as the most viable form of PHC for patients who required continuous monitoring of their clinical conditions. Associated with these technologies, interventions monitoring patients' health status contributed to the management of patients with chronic diseases in the pandemic context [2].

The emphasis on client participation and self-management interventions is a trend that can be observed in several sectors of the economy. This approach allows companies to lower costs while requiring a more active role from consumers. Through the provision of a patient portal, patients can update personal data, consult and schedule appointments and examinations, check examination results, and obtain clinical history information. By doing so, patients take responsibility for their health care and self-manage their clinical conditions. This type of technology is currently used by companies in other sectors, such as airlines that make travelers responsible for purchasing tickets on a web platform [47] and supermarkets that make customers responsible for scanning their merchandise at self-service checkouts [48]. Similarly, the provision of education and training to HCPs through digital technology follows the trend of distance education in general. Learning software has been widespread since the 2000s, when the internet and PCs became more accessible, and is currently an efficient tool for mass education [49,50].

An important topic that emerged from the reviews included in this overview is the need for providing competency training in computer skills and software use for health workers [34,36,41,43-45], in addition to offering guidance to patients [41]. Nonetheless, these reviews did not mention which skills HCPs and patients need to be trained in to use digital technologies. Neve et al [51] argue that HCPs need to learn to work with and interpret the app-derived data and discuss this information with patients, in addition to being trained in communication skills in digital interfaces and workflow management when there are multiple methods of communication between clients and HCPs. Patients also require assistance and guidance in using digital health technologies. Understanding the PHC worker and patient attitudes and needs for ensuring the adequate use of technologies is critical for their effectiveness [2,51,52].

Training capabilities are even more important when discussing the possibility of delegating duties to other professionals, such as community health workers or nurses. Agarwal et al [36] showed positive results in medication adherence in patients with chronic illnesses when assisted by community health agents

XSL•FO
RenderX

supported by computerized CDSS. This discussion is critical, especially in countries with a lack of physicians or underserved areas, and this can be an alternative to improving access to health services in PHC [53].

Some reviews included in this overview (7/18, 39%) also pointed out that adequate use and acceptability across health care workers and patients are influenced by the software or app design, which must be user-friendly and intuitive [18,30,34,38,43,44,46]. Furthermore, technologies must be integrated with systems already used to avoid an increase in the volume of work and rework. Digital health information systems should be designed in line with the functions of PHC. They must also integrate effectively with digital health functions, which include aggregation (patient, diagnostic, and clinician reporting systems and automated collection systems), analysis, and use of information for action [54].

A relevant barrier to the implementation of digital health technologies, mainly those related to client-to-HCP and HCP-to-HCP communication, is the lack of reimbursement and billing strategies for remote consultations [30,31,33]. These new methods of communication between patients and health professionals create new tasks for the latter, which need to be taken into account for payment purposes. These new tasks should also be included in the PHC workflow to avoid excessively increasing the workload. Public policies need to guarantee reimbursement for health care teams for remote care. Research shows that, although this type of policy has been implemented in some countries, results are still mixed in changing clinical practice [2]. Therefore, more research is needed to design reimbursement models that are effective in encouraging the use of this type of digital technology.

The main limitation of this overview of reviews is that most of the included reviews (16/18, 89%) analyzed studies conducted in high-income countries, mainly in North America and Europe. Most of the studies from Europe were conducted in England. The concentration of digital health in high-income countries is a current research concern that was addressed by Xiong et al [41]. They identified 4 main factors that influence digital health service uptake in low- and middle-income countries: governmental regulation and political commitment, the spread of information and communications technology use with interoperability across platforms, user-centered design, and training and incentives for sustainable use of digital tools by health workers and patients. In addition, more research has been conducted to explore this topic, providing insights for developing, implementing, and scaling up digital health initiatives [55,56]; analyzing technologies for improving disease diagnosis and management at the point of care [57]; and mapping digital health interventions for high-risk pregnancies in limited-resource settings [58].

Furthermore, other limitations inherent to the type of methodology used can be highlighted. Overviews of reviews are dependent on the topics covered in published systematic or scoping reviews. Therefore, although this study potentially captured the most relevant types of digital technologies and health issues addressed by these tools in PHC settings, it does not reveal all the possible applications of digital health in PHC. Similarly, overviews of reviews are dependent on the reporting of the included research syntheses. In addition, although the included reviews presented a good methodological quality, this study summarized the results as they were presented by the authors of these reviews. Moreover, relevant studies may have been omitted due to errors in the selection, appraisal, or data extraction processes.

Conclusions

Finally, despite most studies in the included reviews finding consistent results regarding the implementation of digital interventions, they showed several mixed results related to health service quality and patients' clinical conditions or behavior changes [36,41,42,45]. Deficiencies in the implementation of the analyzed technologies may have caused those mixed results. Low adherence to digital intervention protocols by health professionals, lack of acceptability by patients, and difficulties in using the software or app can harm the achievement of results at the level of health services and subsequently in patient-level outcomes. Analyzing the results of a digital health intervention requires examining the entire impact pathway, including the linkages among implementation, health service quality, business models, and clinical condition outcomes, which helps identify bottlenecks throughout the process. The theory of change offers a model to describe the causal assumptions behind the links in the pathway, which explains how change is expected to happen [59]. Therefore, those explanations may contribute to digital health implementation in other settings, which can effectively improve health care quality and patient clinical conditions.

Acknowledgments

This work was supported by the e-Hospital4Future: Building future through an innovated and digital skilled hospital project (EU4Health grant 101101190). The authors acknowledge the Foundation for Science and Technology (Ministry of Science, Technology and Higher Education) for its financial support via projects UIDB/00667/2020 and UIDB/00667/2020 (Research & Development Unit for Mechanical and Industrial Engineering [UNIDEMI]) and Oswaldo Cruz Foundation (Fiocruz) for its institutional support. The authors also acknowledge the contributions of the discussions with IDeaS Laboratory colleagues at UNIDEMI.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e53594_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 15 KB - jmir_v26i1e53594_app2.docx]

References

1. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
2. Pagliari C. Digital health and primary care: past, pandemic and prospects. J Glob Health 2021 Jul 02;11:01005 [FREE Full text] [doi: 10.7189/jogh.11.01005] [Medline: 34221352]
3. Lapão LV, Peyroteo M, Maia M, Seixas J, Gregório J, Mira da Silva M, et al. Implementation of digital monitoring services during the COVID-19 pandemic for patients with chronic diseases: design science approach. J Med Internet Res 2021 Aug 26;23(8):e24181 [FREE Full text] [doi: 10.2196/24181] [Medline: 34313591]
4. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810 [FREE Full text] [doi: 10.1177/20552076221085810] [Medline: 35340904]
5. Stevens S, Gallagher S, Andrews T, Ashall-Payne L, Humphreys L, Leigh S. The effectiveness of digital health technologies for patients with diabetes mellitus: a systematic review. Front Clin Diabetes Healthc 2022;3:936752 [FREE Full text] [doi: 10.3389/fcdhc.2022.936752] [Medline: 36992773]
6. Santo K, Redfern J. Digital health innovations to improve cardiovascular disease care. Curr Atheroscler Rep 2020 Oct 03;22(12):71 [FREE Full text] [doi: 10.1007/s11883-020-00889-x] [Medline: 33009975]
7. Thomas Craig KJ, Morgan LC, Chen CH, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021 May 25;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
8. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327 [FREE Full text] [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
9. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
10. WHO launches a new global initiative on digital health supported by the G20 presidency. World Health Organization. 2023. URL: https://tinyurl.com/5yxecner [accessed 2023-10-17]
11. Declaration of Astana: from Alma-Ata towards universal health coverage and the sustainable development goals. World Health Organization. 2018 Dec 31. URL: https://www.who.int/publications/i/item/WHO-HIS-SDS-2018.61 [accessed 2024-04-29]
12. Digital technologies: shaping the future of primary health care. World Health Organization. 2018. URL: https://tinyurl.com/bdf442pt [accessed 2023-06-02]
13. A vision for primary health care in the 21st century: towards universal health coverage and the sustainable development goals. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/WHO-HIS-SDS-2018.15 [accessed 2023-06-02]
14. Aagja J, Shome S, Chandra A. A bibliometric analysis of digital health and mobile health related global research publications. Hosp Top 2023;101(4):319-325. [doi: 10.1080/00185868.2022.2060155] [Medline: 35380102]
15. Fornazin M, Penteado BE, de Castro LC, de Castro Silva SL. From medical informatics to digital health: a bibliometric analysis of the research field. In: Proceedings of the 27th annual Americas Conference on Information Systems. 2021 Presented at: AMCIS '21; August 9-13, 2021; Virtual Event p. 5-18 URL: https://tinyurl.com/4jvkhvbs
16. Tian H, Chen J. A bibliometric analysis on global eHealth. Digit Health 2022;8:20552076221091352 [FREE Full text] [doi: 10.1177/20552076221091352] [Medline: 35433016]
17. Waqas A, Teoh SH, Lapão LV, Messina LA, Correia JC. Harnessing telemedicine for the provision of health care: bibliometric and scientometric analysis. J Med Internet Res 2020 Oct 02;22(10):e18835 [FREE Full text] [doi: 10.2196/18835] [Medline: 33006571]

XSL•FO
RenderX

18. Peyroteo M, Ferreira IA, Elvas LB, Ferreira JC, Lapão LV. Remote monitoring systems for patients with chronic diseases in primary health care: systematic review. JMIR Mhealth Uhealth 2021 Dec 21;9(12):e28285 [FREE Full text] [doi: 10.2196/28285] [Medline: 34932000]
19. Pollock M, Fernandes RM, Becker LA, Pieper D, Hartling L. Overviews of reviews. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3. Burlington, MA: Wiley-Blackwell; 2022.
20. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
21. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS Med 2011 Jan 18;8(1):e1000387 [FREE Full text] [doi: 10.1371/journal.pmed.1000387] [Medline: 21267058]
22. Rouleau G, Gagnon MP, Côté J, Payne-Gagnon J, Hudson E, Dubois CA. Impact of information and communication technologies on nursing care: results of an overview of systematic reviews. J Med Internet Res 2017 Apr 25;19(4):e122 [FREE Full text] [doi: 10.2196/jmir.6686] [Medline: 28442454]
23. De Witte NA, Joris S, Van Assche E, Van Daele T. Technological and digital interventions for mental health and wellbeing: an overview of systematic reviews. Front Digit Health 2021 Dec 23;3:754337 [FREE Full text] [doi: 10.3389/fdgth.2021.754337] [Medline: 35005695]
24. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. NPJ Digit Med 2018;1:12 [FREE Full text] [doi: 10.1038/s41746-018-0021-9] [Medline: 31304297]
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336-341 [FREE Full text] [doi: 10.1016/j.ijsu.2010.02.007] [Medline: 20171303]
27. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://iris.who.int/handle/10665/311941 [accessed 2023-06-02]
28. Critical appraisal tools. The Joanna Briggs Institute. URL: https://jbi.global/critical-appraisal-tools [accessed 2024-03-06]
29. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
30. Hickson R, Talbert J, Thornbury WC, Perin NR, Goodin AJ. Online medical care: the current state of "eVisits" in acute primary care delivery. Telemed J E Health 2015 Feb;21(2):90-96. [doi: 10.1089/tmj.2014.0022] [Medline: 25474083]
31. Beheshti L, Kalankesh LR, Doshmangir L, Farahbakhsh M. Telehealth in primary health care: a scoping review of the literature. Perspect Health Inf Manag 2022;19(1):1n [FREE Full text] [Medline: 35440933]
32. Kuo A, Dang S. Secure messaging in electronic health records and its impact on diabetes clinical outcomes: a systematic review. Telemed J E Health 2016 Sep;22(9):769-777. [doi: 10.1089/tmj.2015.0207] [Medline: 27027337]
33. Liddy C, Moroz I, Mihan A, Nawar N, Keely E. A systematic review of asynchronous, provider-to-provider, electronic consultation services to improve access to specialty care available worldwide. Telemed J E Health 2019 Mar;25(3):184-198. [doi: 10.1089/tmj.2018.0005] [Medline: 29927711]
34. Meunier P, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023 Jan 23;21(1):57-69 [FREE Full text] [doi: 10.1370/afm.2908] [Medline: 36690490]
35. El Asmar ML, Dharmayat KI, Vallejo-Vaz AJ, Irwin R, Mastellos N. Effect of computerised, knowledge-based, clinical decision support systems on patient-reported and clinical outcomes of patients with chronic disease managed in primary care settings: a systematic review. BMJ Open 2021 Dec 22;11(12):e054659 [FREE Full text] [doi: 10.1136/bmjopen-2021-054659] [Medline: 34937723]
36. Agarwal S, Glenton C, Tamrat T, Henschke N, Maayan N, Fønhus MS, et al. Decision-support tools via mobile devices to improve quality of care in primary healthcare settings. Cochrane Database Syst Rev 2021 Jul 27;7:CD012944 [FREE Full text] [doi: 10.1002/14651858.CD012944.pub2] [Medline: 34314020]
37. Tuan Soh TY, Nik Mohd Rosdy NM, Mohd Yusof MY, Azhar Hilmy SH, Md Sabri BA. Adoption of a digital patient health passport as part of a primary healthcare service delivery: systematic review. J Pers Med 2022 Nov 01;12(11):1814 [FREE Full text] [doi: 10.3390/jpm12111814] [Medline: 36579540]
38. Andrikopoulou E, Scott P, Herrera H, Good A. What are the important design features of personal health records to improve medication adherence for patients with long-term conditions? A systematic literature review. BMJ Open 2019 Sep 26;9(9):e028628 [FREE Full text] [doi: 10.1136/bmjopen-2018-028628] [Medline: 31558449]

XSL•FO
RenderX

39. Kyaw BM, Tudor Car L, van Galen LS, van Agtmael MA, Costelloe CE, Ajuebor O, et al. Health professions digital education on antibiotic management: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Sep 12;21(9):e14984 [FREE Full text] [doi: 10.2196/14984] [Medline: 31516125]
40. Willis VC, Thomas Craig KJ, Jabbarpour Y, Scheufele EL, Arriaga YE, Ajinkya M, et al. Digital health interventions to enhance prevention in primary care: scoping review. JMIR Med Inform 2022 Jan 21;10(1):e33518 [FREE Full text] [doi: 10.2196/33518] [Medline: 35060909]
41. Xiong S, Lu H, Peoples N, Duman EK, Najarro A, Ni Z, et al. Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries. NPJ Digit Med 2023 Feb 01;6(1):12 [FREE Full text] [doi: 10.1038/s41746-023-00764-4] [Medline: 36725977]
42. Moon K, Sobolev M, Kane JM. Digital and mobile health technology in collaborative behavioral health care: scoping review. JMIR Ment Health 2022 Feb 16;9(2):e30810 [FREE Full text] [doi: 10.2196/30810] [Medline: 35171105]
43. Jonnagaddala J, Godinho MA, Liaw ST. From telehealth to virtual primary care in Australia? A Rapid scoping review. Int J Med Inform 2021 Jul;151:104470 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104470] [Medline: 34000481]
44. Silva CR, Lopes RH, de Goes Bay O, Martiniano CS, Fuentealba-Torres M, Arcêncio RA, et al. Digital health opportunities to improve primary health care in the context of COVID-19: scoping review. JMIR Hum Factors 2022 May 31;9(2):e35380 [FREE Full text] [doi: 10.2196/35380] [Medline: 35319466]
45. Jimenez G, Matchar D, Koh CH, van der Kleij R, Chavannes NH, Car J. The role of health technologies in multicomponent primary care interventions: systematic review. J Med Internet Res 2021 Jan 11;23(1):e20195 [FREE Full text] [doi: 10.2196/20195] [Medline: 33427676]
46. Maillet É, Paré G, Currie LM, Raymond L, Ortiz de Guinea A, Trudel M, et al. Laboratory testing in primary care: a systematic review of health IT impacts. Int J Med Inform 2018 Aug;116:52-69. [doi: 10.1016/j.ijmedinf.2018.05.009] [Medline: 29887235]
47. Renny, Guritno S, Siringoringo H. Perceived usefulness, ease of use, and attitude towards online shopping usefulness towards online airlines ticket purchase. Procedia Soc Behav Sci 2013 Jun;81:212-216. [doi: 10.1016/j.sbspro.2013.06.415]
48. Litfin T, Wolfram G. New automated checkout systems. In: Krafft M, Mantrala MK, editors. Retailing in the 21st Century. Berlin, Germany: Springer; 2010:189-203.
49. Bezhovski Z, Poorani S. The evolution of e-learning and new trends. Inf Knowl Manag 2016;6(3):50-57 [FREE Full text]
50. Rana H, Rajiv R, Lal M. E-learning: issues and challenges. Int J Comput Appl 2014 Jul 18;97(5):20-24. [doi: 10.5120/17004-7154]
51. Neve G, Fyfe M, Hayhoe B, Kumar S. Digital health in primary care: risks and recommendations. Br J Gen Pract 2020 Dec;70(701):609-610 [FREE Full text] [doi: 10.3399/bjgp20X713837] [Medline: 33243917]
52. Neves AL, Burgers J. Digital technologies in primary care: implications for patient care and future research. Eur J Gen Pract 2022 Dec;28(1):203-208 [FREE Full text] [doi: 10.1080/13814788.2022.2052041] [Medline: 35815445]
53. Fava VM, Lapão LV. Towards universal health coverage in Brazil: analysis of programs to expand the supply of doctors in primary health care. J Glob Health Rep 2021:5. [doi: 10.29392/001c.19143]
54. Krist A, Phillips R, Leykum L, Olmedo B. Digital health needs for implementing high-quality primary care: recommendations from the National Academies of Sciences, Engineering, and Medicine. J Am Med Inform Assoc 2021 Nov 25;28(12):2738-2742 [FREE Full text] [doi: 10.1093/jamia/ocab190] [Medline: 34537851]
55. Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R, et al. Best practices in scaling digital health in low and middle income countries. Global Health 2018 Nov 03;14(1):103 [FREE Full text] [doi: 10.1186/s12992-018-0424-z] [Medline: 30390686]
56. Kumar M, Mostafa J. Electronic health records for better health in the lower- and middle-income countries. Libr Hi Tech 2020 Mar 10;38(4):751-767. [doi: 10.1108/lht-09-2019-0179]
57. Khubone T, Tlou B, Mashamba-Thompson TP. Electronic health information systems to improve disease diagnosis and management at point-of-care in low and middle income countries: a narrative review. Diagnostics 2020 May 20;10(5):327. [doi: 10.3390/diagnostics10050327]
58. Shahil Feroz A, Afzal N, Seto E. Exploring digital health interventions for pregnant women at high risk for pre-eclampsia and eclampsia in low-income and-middle-income countries: a scoping review. BMJ Open 2022 Feb 08;12(2):e056130 [FREE Full text] [doi: 10.1136/bmjopen-2021-056130] [Medline: 35135777]
59. Mayne J. Useful theory of change models. Can J Program Eval 2015 Sep;30(2):119-142. [doi: 10.3138/cjpe.230]

Abbreviations

CDSS: clinical decision support system
EHR: electronic health record
HCP: health care provider
PHC: primary health care
PHR: personal health record
PICOS: population, intervention, comparator, outcome, and study design

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
WHO: World Health Organization

Edited by A Mavragani; submitted 11.10.23; peer-reviewed by YTD Cheung, C Ndayishimiye; comments to author 26.02.24; revised version received 15.03.24; accepted 24.07.24; published 29.10.24.

Please cite as:
Fava VMD, Lapão LV
Provision of Digital Primary Health Care Services: Overview of Reviews
J Med Internet Res 2024;26:e53594
URL: https://www.jmir.org/2024/1/e53594
doi:10.2196/53594
PMID:

©Virgínia Maria Dalfior Fava, Luís Velez Lapão. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Health Care Professionals' Experience of Using AI: Systematic Review With Narrative Synthesis

Abimbola Ayorinde[1], BSc, MSc, PhD; Daniel Opoku Mensah[1], PhD; Julia Walsh[1], PhD; Iman Ghosh[1], MSc; Siti Aishah Ibrahim[1], MSc; Jeffry Hogg[2], PhD; Niels Peek[3,4], PhD; Frances Griffiths[1], PhD

[1]Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, United Kingdom

[2]AI Digital Health Research and Policy Group, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom

[3]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

[4]The Healthcare Improvement Studies Institute, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

Corresponding Author:
Abimbola Ayorinde, BSc, MSc, PhD
Division of Health Sciences
Warwick Medical School
University of Warwick
Medical School Building
Gibbet Hill Road
Coventry, CV4 7AL
United Kingdom
Phone: 44 2476151098
Email: a.ayorinde.1@warwick.ac.uk

Abstract

Background: There has been a substantial increase in the development of artificial intelligence (AI) tools for clinical decision support. Historically, these were mostly knowledge-based systems, but recent advances include non–knowledge-based systems using some form of machine learning. The ability of health care professionals to trust technology and understand how it benefits patients or improves care delivery is known to be important for their adoption of that technology. For non–knowledge-based AI tools for clinical decision support, these issues are poorly understood.

Objective: The aim of this study is to qualitatively synthesize evidence on the experiences of health care professionals in routinely using non–knowledge-based AI tools to support their clinical decision-making.

Methods: In June 2023, we searched 4 electronic databases, MEDLINE, Embase, CINAHL, and Web of Science, with no language or date limit. We also contacted relevant experts and searched reference lists of the included studies. We included studies of any design that reported the experiences of health care professionals using non–knowledge-based systems for clinical decision support in their work settings. We completed double independent quality assessment for all included studies using the Mixed Methods Appraisal Tool. We used a theoretically informed thematic approach to synthesize the findings.

Results: After screening 7552 titles and 182 full-text articles, we included 25 studies conducted in 9 different countries. Most of the included studies were qualitative (n=13), and the remaining were quantitative (n=9) and mixed methods (n=3). Overall, we identified 7 themes: health care professionals' understanding of AI applications, level of trust and confidence in AI tools, judging the value added by AI, data availability and limitations of AI, time and competing priorities, concern about governance, and collaboration to facilitate the implementation and use of AI. The most frequently occurring are the first 3 themes. For example, many studies reported that health care professionals were concerned about not understanding the AI outputs or the rationale behind them. There were issues with confidence in the accuracy of the AI applications and their recommendations. Some health care professionals believed that AI provided added value and improved decision-making, and some reported that it only served as a confirmation of their clinical judgment, while others did not find it useful at all.

Conclusions: Our review identified several important issues documented in various studies on health care professionals' use of AI tools in real-world health care settings. Opinions of health care professionals regarding the added value of AI tools for supporting clinical decision-making varied widely, and many professionals had concerns about their understanding of and trust in this technology. The findings of this review emphasize the need for concerted efforts to optimize the integration of AI tools in real-world health care settings.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD42022336359; https://tinyurl.com/2yunvkmb

(J Med Internet Res 2024;26:e55766) doi:10.2196/55766

KEYWORDS

artificial intelligence; clinical decision support systems; CDSS; decision-making; quality assessment; clinician experience; health care professionals; health care delivery

Introduction

Artificial intelligence (AI) refers to the ability of machine or computer programs to replicate human intelligence. AI is widely believed to have considerable potential in health care, ranging from the transformation of many aspects of patient care to streamlining administrative processes within and between various organizations involved in health care [1]. Specific potential benefits include earlier detection of disease, improved patient safety, improved estimation of capacity needs, and the facilitation of personalized medicine [1,2]. Globally, there has been great interest in the use of AI in health care settings [3]. For example, the UK government invested £250 million (US $328 million) in a national laboratory to boost AI in the National Health Service in 2019 [2].

AI has various applications in health care, but one of the most well-known applications involves using AI to enhance health care delivery to aid clinical decision-making in a process known as clinical decision support systems (CDSSs) [4]. Historically, CDSSs matched individual patient characteristics to a manually curated, computerized clinical knowledge base [5]. These "knowledge-based" CDSSs comprised a set of "if and then" rules with the system retrieving data from the knowledge base to evaluate the rule and then generating an output or suggested action [4]. More recently, machine learning methods have enabled the development of "non–knowledge-based" systems that use complex representations (patterns, trees, networks, and equations) derived from large datasets to generate outputs or actions [4]. Machine learning models are typically difficult to understand for humans because of their complexity, especially when they have been derived via deep learning, which is the most common method nowadays [4].

The integration of this form of AI tools in health care is notably progressing more slowly than initially anticipated, given the levels of investment, due to several challenges to implementation [6,7]. Some of these challenges extend beyond a specific domain of AI tools and reflect broader issues encountered in the implementation of innovation in practice. Frameworks such as the Consolidated Framework for Implementation Research shed light on these challenges by emphasizing the multifaceted nature of the implementation process [8]. Many challenges in AI implementation are similar to those encountered in the implementation of any new digital technology, and several frameworks have provided comprehensive lens through which these challenges can be examined, for example, the Non-adoption, Abandonment, Scale-up, Spread and Sustainability framework [9] and the meta-analysis–based modified Unified Theory of Acceptance and Use of Technology (meta-UTAUT) [10,11]. However, issues specific to the use of AI technologies, particularly the non–knowledge-based systems have been reported [7]. Health care professionals often struggle with concerns about the reliability, interpretability, and ethical implications of these AI systems [6,7].

This review aimed to qualitatively synthesize the existing evidence on the experiences of health care professionals in using non–knowledge-based CDSSs (referred to as AI tools in the Methods and Results sections) that are deployed in their work setting with a focus on issues specific to non–knowledge-based CDSSs. We believe that focusing on issues specific for non–knowledge-based CDSSs will help with understanding the challenges that remain for the deployment of these systems in health care.

Methods

Search Strategy

This reporting of this systematic review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guideline because the PRISMA-AI (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Artificial Intelligence) guideline was not available at the time of writing [12,13]. We conducted a systematic search of the MEDLINE, Embase, CINAHL, and Web of Science databases for relevant articles. Searches were initially conducted in May 2022, and the searches were updated in June 2023. Overall, 4 key concepts and their possible variations informed the search strategy: AI (such as artificial intelligence, machine learning, and natural language processing); decision support systems (such as decision support system and computer-assisted diagnosis); health care professionals (such as health care professionals, nurse, and physician); and terms relating to experience (such as experience, view, and opinion). Terms within similar categories were combined with OR and then the results from each category were combined with AND (Multimedia Appendix 1). The reference lists of eligible studies were also screened, and experts were contacted for relevant articles that may have been missed. There were no language and date restrictions.

Study Selection

The search results were imported into EndNote (Clarivate Analytics). referencing and bibliography management software to remove duplicates. The titles and abstracts of the retrieved articles were screened by 1 reviewer (AA, JW, or DOM), and a random sample of 20% (1511/7552) of the retrieved articles were screened by a second reviewer. The title and abstract screening was completed on Rayyan software (Rayyan), a web-based application for systematic reviews. The full texts of potentially relevant articles were then retrieved and examined in detail for eligibility by 2 independent reviewers (AA and IG or JW and DOM). Any discrepancies were resolved by

XSL•FO
RenderX

discussion between reviewers and a third reviewer, or the whole project team was consulted when necessary. The most challenging aspect of the study selection was deciding whether the AI described in that particular study is knowledge based or non–knowledge based. When this was unclear, we discussed as a team and deferred to the team member with data science expertise (NP) for final decision.

Inclusion and Exclusion Criteria

Studies were included if (1) they targeted health care or social care professionals; (2) the article described a non–knowledge-based CDSSs deployed for health care; and (3) the study explored health care professionals' experiences of using AI tools in health care or barriers and facilitators to AI tool use in the real world. All study designs were deemed acceptable for inclusion, but nonprimary studies were excluded. However, we screened the references included in the relevant reviews to identify primary studies that may have been missed from our searches. The inclusion and exclusion criteria for study selection are detailed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Eligibility criteria.

Inclusion criteria based on relevant factors

- Population
  - Health care and social care professionals
- Intervention
  - Non–knowledge-based clinical decision support systems (CDSSs) [4] deployed for health care
- Comparator
  - Any (including usual care or no comparator)
- Primary outcomes
  - Health care professionals' experiences of the use of artificial intelligence (AI) tools in health care delivery
  - Benefits and challenges of the use of AI tools in health care
- Secondary outcomes
  - Health care professionals' understanding of AI tools in health care delivery
  - Perceived and actual impact of AI tools on decision-making and the health system (eg, the need for extra clinicians) and clinical workflow and patient pathway
- Types of study
  - Any primary study (empirical study) of any design (qualitative, quantitative, or mixed methods)
  - Studies that focused on issues relating to health care provision, health outcomes, and health service or system configuration were considered for inclusion
- Context
  - Any health care setting from any country

Exclusion criteria based on relevant factors

- Population
  - Nonprofessional caregivers (such as family caregivers)
- Intervention
  - Knowledge-based CDSSs
  - Systems under development or testing
  - AI tools as treatment (eg, robopets)
- Comparator
  - Not applicable
- Primary outcomes
  - We excluded studies which focused on the perception of health care professionals on hypothetical use rather than actual use of AI tools
- Types of study
  - Nonprimary studies (such as literature reviews and opinion papers)
  - Studies exclusively reporting the real-world effectiveness, performance, or diagnostic accuracy were excluded unless accompanied by results relating to health care professionals' experience relating to the outcomes stated earlier
- Context
  - Studies which are not focused on health care settings

XSL•FO
RenderX

Data Extraction and Study Quality Assessment

A data extraction form, developed and piloted by the review team, was used to extract relevant information from each article. The following data were extracted from the eligible studies: author names; publication year; study design; study aims and objectives; geographic location; sample size; characteristics of the health care professionals included (including any indication of how long they have been using AI); description of the AI examined; input data (image or tabular or multimodal); platform (eg, electronic medical record based, standalone app or website, or mobile health embedded); tool development (eg, was it developed in house or by a third party?); study funder; and outcomes.

Data were extracted by one reviewer and checked by a second reviewer. Disagreements were resolved by deliberation between the 2 reviewers and, where necessary, a third reviewer or the project team was consulted.

Two authors independently assessed the quality of all included studies using the Mixed Methods Appraisal Tool [14], which was appropriate due to the variety of study designs included in this review. Disagreements were resolved by discussion between the 2 assessors. A third reviewer was consulted to reach a final decision if the independent assessors were unable to reach a consensus.

Data Analysis

Initially, a theoretically informed thematic synthesis was adopted to analyze the findings of the included studies [15]. The meta-UTAUT was used as a theoretical framework [10,11]. This theoretical framework was selected because it is specific to the acceptance and use of information systems and IT, has been refined through meta-analysis, and is extensively used for understanding adoption of information systems and IT [11]. Two independent reviewers coded the information from the findings of the included studies into descriptive codes and linked these codes to the components of the meta-UTAUT framework (Figure 1). During coding, we remained alert to information that did not fit the framework (Figure 1). We used NVivo software (Lumivero) to facilitate data coding process. Codes and coding were reviewed by the team and modifications discussed. Although the meta-UTAUT allowed us to understand the data, we noticed that the coding was fragmented and the connections between the codes were being missed. Therefore, we decided to describe findings based on the broader themes.

Figure 1. Schematic illustration of the data coding process.

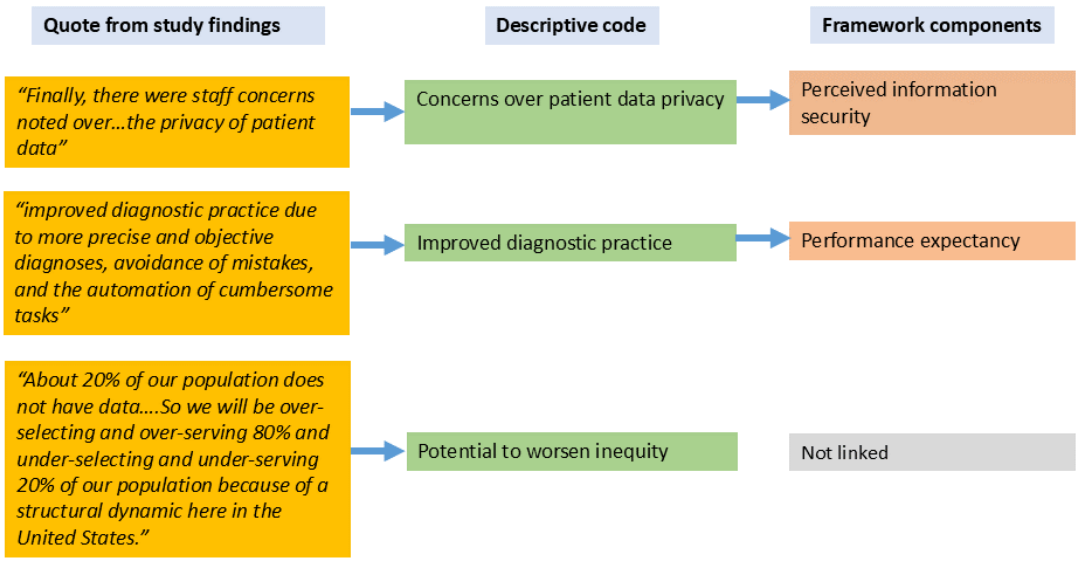

After coding, we made a judgment as to whether codes were specific to AI or not by undertaking a thought experiment supported where necessary by published evidence. Team members (AA and FG) asked themselves, "Could this code be applied to another digital intervention including knowledge-based clinical decision support?" If the answer was yes, we labeled it as "nonspecific." Examples of nonspecific codes include problems with internet connection, because this will cause a problem with any digital system requiring internet connection (eg, video calls); challenges with using digital systems, because it can surface where insufficient training has been given [16]; alert fatigue, because it is a recognized problem for any CDSS [17]; and the importance of ensuring that systems do not adversely affect workflows, because it is a recognized issue for the design of digital interventions [18]. In the Results section, we provide examples of the nonspecific codes. We identified and focused our analysis on the codes specific to AI tools. Through a process of iteration, discussion, and independent verification, we identified 7 overarching themes that encapsulate the data.

Quotes from the findings of the studies were coded into descriptive codes and linked to the components of the meta-UTAUT [10,11]. In cases where no direct link to the meta-UTAUT components existed, those codes were still retained for analysis.

XSL•FO
RenderX

Stakeholder Engagement Workshop

We conducted a web-based stakeholder engagement workshop in August 2022 to obtain feedback on our initial findings and recommendations for future research and practice. We invited patient and public representatives, health care professionals, experts in AI, and policy makers. Potential participants were invited through direct personal contacts. We sent out advertisements to National Institute for Health and Care Research Applied Research Collaboration West Midlands patient and public representatives, Young Person's Mental Health Advisory Group of King's College London, and Cross-Applied Research Collaboration working group on AI. A 2-hour workshop was held over Zoom (Zoom Video Communications, Inc) where we presented our evidence synthesis and invited attendees to provide feedback on our initial findings and recommendations of any additional research that could be conducted in this area as well as any practice implications arising from the existing findings.

Results

Study Selection

A total of 10,743 records were retrieved from the electronic databases. After the removal of duplicate records and screening based on titles and abstract, 172 full-text articles were screened. An additional 10 studies were identified from other sources. Overall, 25 studies describing the experiences of health care professionals from differing disciplines and levels of seniority with various AI met the selection criteria and were included in the final review. The process of study selection is shown in Figure 2. List of excluded studies and the reasons for exclusion are presented in Multimedia Appendix 2.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection.

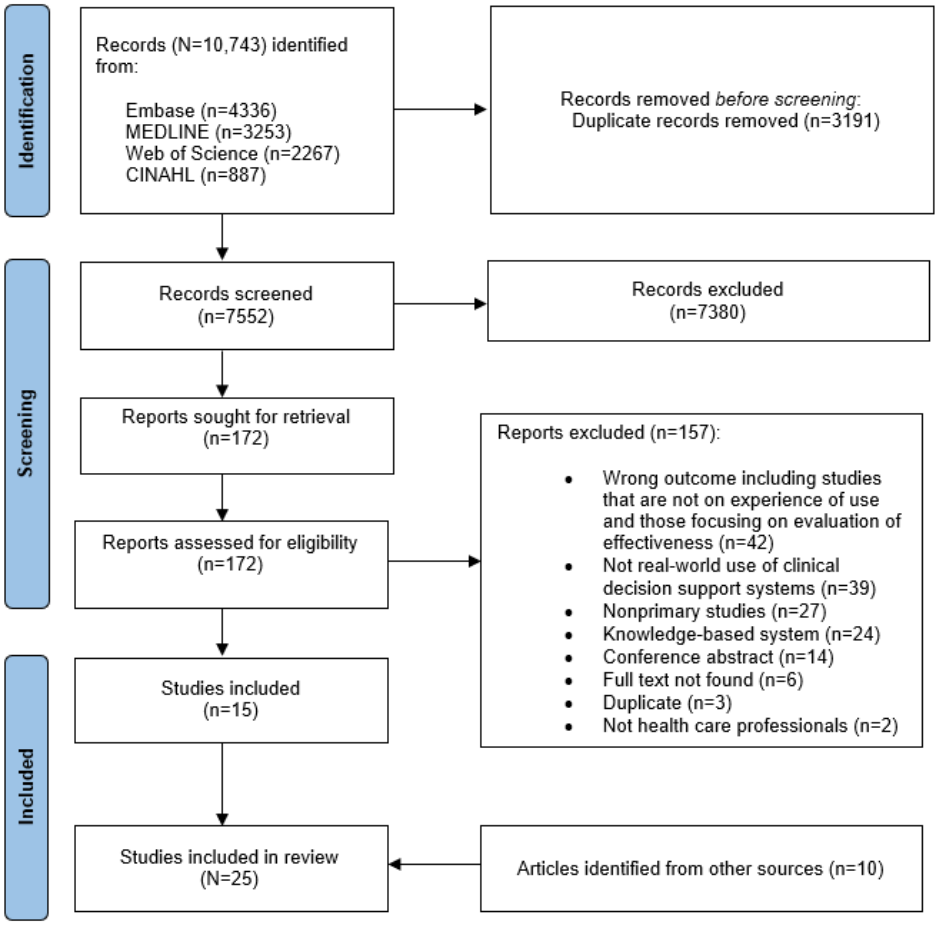

Study Characteristics

The characteristics of the included studies are presented in Multimedia Appendix 2. The included studies (N=25) were published from 1990 to 2022 with 92% (n=23) of the studies published between 2019 and 2023. The studies covered 9 countries, with the majority being reported from the United States (n=12), then the Netherlands (n=2), Canada (n=2), Austria (n=2), and one each from the United Kingdom, Spain, Switzerland, Argentina, Brazil, China, and Thailand. The included studies were primarily qualitative (13/25, 52%), with 36% (9/25) being quantitative and 12% (3/25) being mixed methods studies. Sample sizes ranged from 12 to 724 participants, although 2 studies did not report these data [19,20]. Participants of the included studies comprised practitioners of various disciplines spanning several clinical specialties. The types of AI systems used in the included studies varied widely, covering those used in risk prediction, diagnosis, and treatment (Table 1).

XSL•FO
RenderX

Table 1. Description of the AI[a] systems used in included studies.

Study, year	Study type	Name of AI system	Clinical decision-making task	Setting; country	Target population	Tool development
Beede et al [21], 2020	Qualitative study	No name given	Assessment of diabetic retinopathy	Primary care clinics; Thailand	Patients with diabetes requiring an eye examination	In house
Cruz et al [22], 2019	Quantitative study	Savana system	Notifies physicians in real time about recommendations regarding the health care process to improve adherence rates to clinical pathways	Primary care clinics; Spain	Patients attending primary care	Commercial
Frymoyer et al [23], 2020	Quantitative study	InsightRX-MIPD CDS tool	Determines the optimal dosing regimen of vancomycin	Hospital; United States	Neonates and children for suspected or documented infections with MRSA[b], methicillin-resistant coagulase-negative Staphylococci, and other drug-resistant gram-positive organisms	Commercial
Ginestra et al [24], 2019	Quantitative study	EWS 2.0 alert	Sepsis risk prediction	Hospital; United States	Non–intensive care unit admissions	In house
Gonçalves et al [19], 2020	Qualitative study	Robot Laura	Early identification of sepsis	Hospital; Brazil	Patients in hospital	Commercial
Henkel et al [25], 2022	Quantitative study	AI-Pathway Companion Prostate Cancer VA10B	Diagnosis and management of prostate adenocarcinoma cancer	Hospital; Switzerland	Patients undergoing prostate cancer screening	Commercial
Henry et al [26], 2022	Qualitative study	TREWS[c]	Sepsis detection and treatment management	Hospital; United States	Patients in hospital	Unclear
Jauk et al [27], 2021	Mixed methods study	Unclear	Predicting delirium	Hospital; Austria	Every patient admitted to one of the departments	Unclear
Jauk et al [28], 2022	Quantitative study	Delirium prediction software	Delirium risk prediction	Hospital; Austria	Patients in hospital	In house
Jordan et al [29], 2022	Qualitative study	KATE	Clinical triage	Hospital; United States	Patients at the emergency department	Commercial
Joshi et al [30], 2022	Qualitative study	No name given	Sepsis risk prediction	Hospital; United States	Patients at risk of sepsis	Commercial and in house
Kappen et al [31], 2016	Mixed methods study	No name given	Predicting the risk of PONV[d]	Hospital; Netherlands	All surgical patients	Unclear
Lebovitz et al [32], 2022	Qualitative	No name given	Diagnosis of breast cancer, lung cancer and estimation of bone age	Hospital; United States	Patients requiring diagnostic radiology	Unclear
Marwaha et al [33], 2021	Qualitative study	Face2Gene	Diagnosis of children with rare genetic syndromes	Hospital; Canada	Children with rare genetic syndromes	Unclear
McAdam et al [20], 1990	Quantitative study	Leeds Abdominal Pain System	Diagnosis of acute abdominal pain	Hospital; United Kingdom	Patients presenting with acute abdominal pain	In house
Nehme et al [34], 2023	Quantitative study	GI Genius	Colorectal polyp detection	Cancer center; United States	Patients undergoing elective outpatient colonoscopy	Commercial
Rabinovich et al [35], 2022	Mixed methods study	TRx	Chest x-ray interpretation	Hospital; Argentina	Patients at the emergency department	In house
Romero-Brufau et al [36], 2020	Quantitative study	No name given	Identify patient at risk for poor glycemic control in the next 3 months	Primary care clinics; United States	Patients at risk of poor glycemic control	Commercial

XSL•FO
RenderX

Study, year	Study type	Name of AI system	Clinical decision-making task	Setting; country	Target population	Tool development
Sandhu et al [37], 2020	Qualitative study	Sepsis Watch	Sepsis risk prediction	Hospital; United States	Patients in hospital	In house
Saunders et al [38], 2021	Qualitative study	The mHOMR[e] tool	Mortality risk prediction	Hospital; Canada	Patients with non–cancer serious illnesses	Unclear
Shiang et al [39], 2022	Quantitative study	AI-based decision support system	Detection of pulmonary embolism, intracranial hemorrhage, and acute cervical spine fractures	Hospital; United States	Patients at the radiology department	Commercial
Singer et al [40], 2022	Qualitative study	No name given	Readmission tool: readmission risk prediction	Hospital; United States	All patients admitted to hospital	Unclear
Strohm et al [41], 2020	Qualitative study	BoneXpert	Bone maturity assessments based on x-rays of pediatric patients' hands	Hospital; Netherlands	Pediatric patients	Commercial
Sun and Medaglia [42], 2019	Qualitative study	Watson for oncology	Personalized cancer management	Hospital; China	Patients with cancer	Commercial
Wang et al [43], 2023	Qualitative study	No name given	Predicting risk of PAD[f]	Community or primary care; United States	Patients with PAD who had at least one visit to Duke Health with a PAD-related diagnosis code between January 2015 and March 2016	In house

[a]AI: artificial intelligence.

[b]MRSA: methicillin-resistant *Staphylococcus aureus*.

[c]TREWS: Targeted Real-Time Early Warning System.

[d]PONV: postoperative nausea and vomiting.

[e]mHOMR: modified Hospitalised-patient One-year Mortality Risk.

[f]PAD: peripheral arterial disease.

Quality of Included Studies

The quality of included studies is presented in Multimedia Appendix 2. Overall, the quality of most of the included studies was good. Of the 25 included studies, 12 qualitative studies [19,21,26,29,32,33,37,38,40-43] fulfilled all the relevant quality criteria on Mixed Methods Appraisal Tool while 5 studies fulfilled at least 80% of the quality criteria [20,22,24,27,30]. The remaining 8 studies fulfilled ≤67% of the quality criteria [23,25,28,31,34-36,39], and most of the unmet criteria were assessed as "can't tell" primarily due to the lack of sufficient information.

Excluded Generic Issues

We identified a range of nonspecific issues related to the implementation of digital technologies in health care (see Textbox 2 for examples). These issues have already been extensively addressed in the existing literature [8] and were therefore excluded from analysis in our study.

XSL•FO
RenderX

Textbox 2. Examples of broader issues relating to implementation reported in included studies.

- Issues relating to abandonment [30,41]
- Competing priorities, especially in cases where health care professionals felt the artificial intelligence (AI) was not particularly useful [31,38]
- Limited infrastructure to support the adoption of the technology [19,21]
- Health care professionals' satisfaction with AI applications [19,23,25,30]
- Ease of use [25,27,28,35]
- Integration into the existing systems and accessibility within the existing system [23,27,41]
- Health care professionals appreciated systems that do not affect the dynamics of their workflow [22,27]
- Concerns about alert fatigue or redundancy of alert [30,38]
- Technological challenges relating to health care professionals' proficiency in the use of hardware and software [19]
- Issues relating to information security, including the necessity of secure transmission of data [23], concerns over privacy of patient data [33], and potential threat to national security [42]
- Lack of complete data for patients [43]
- Lack of standards for what and how data are collected [42]
- Issues with internet connections [19,21]
- Challenges with hardware such as recording equipment, smartphones, and tablets [19]

Key Themes

Overview

In this section, we describe the experiences of health care professionals in using AI tools to support clinical decision-making, as reported in the included studies. We focused on the nuances we believe are specific to AI following the process described in the Methods section. We identified seven themes: (1) understanding of AI applications; (2) level of trust and confidence in AI tools; (3) judging the added value of AI; (4) data availability and limitations of AI; (5) time and competing priorities; (6) concern about governance; and (7) collaboration to facilitate the implementation and use of AI tools. We provide detailed exploration of each theme in subsequent sections.

Health Care Professionals' Understanding of AI Applications

In total, 10 studies reported concerns with health care professionals' lack of understanding of AI applications. This includes lack of understanding of the outputs of the AI application and the algorithms or the rationale for the outputs by the AI applications due to the lack of transparency relating to non–knowledge-based AI applications [24,26,27,30,32,37,38,40-42]. For example [42],

> *This lack of transparency is perceived as a major challenge; the AI technology represents as a "black box", and its users have no power to understand its mechanisms, or modify them to tackle potential problems.*

As another example, a study on the use of a machine learning sepsis early warning system reported as follows [37]:

> *Both RRT [rapid response team] nurses and ED [emergency department] physicians said that they lacked the knowledge and understanding required to assess the validity of the machine learning model...Physicians also lacked knowledge about the model and the predictive nature of the model.*

Another study on sepsis risk prediction reported confusion regarding understanding what the alert means [30]. In an attempt to minimize this issue, inclusion of an explanation for the alert firing was introduced in some hospitals. Some felt the explanations were helpful while others found it more confusing [30]:

> *A lot of people get confused...so say you get 25, when the patient's really sick and then the number goes to twenty, does that mean the patients getting better? What do all of the subsequent numbers mean? If it goes up to 30, is the patient getting worse?*
>
> *Concise content with explanations for firing was well-received. A few noted that this was not possible or was more confusing, but most felt inclusion of explanations was helpful.*

Participants in a study on the use of AI tools in radiology reported as follows [32]:

> *Dr V: How does [the AI tool] know that this is a nodule, but this isn't?*
>
> *Dr C: They all look identical to me.*
>
> *Dr K: They expressed frustration in their inability to understand the divergent AI results: What is it telling me to look at? At this tissue? It looks just like the tissue over here, which is perfectly normal...I have no idea what it's thinking.*

However, in a study [26] on an early warning system for sepsis, limited understanding of how the AI system operates was not considered a major barrier to the use of the application.

XSL•FO
RenderX

Level of Trust and Confidence in AI Tools

Findings related to trust in the accuracy and judgments or recommendations of AI tools are mixed [19,21,26-28,30-37,39,41,43]. Some health care professionals believe that the AI tool performs as expected while others do not, and this was reported across various AI tools. For example, in a study [30] on a sepsis risk prediction tool, implementers were unsure whether the AI tools were able to predict sepsis with clinically meaningful specificity when compared with the traditional early warning system using a systemic inflammatory response system, which relies on physiological data (vital sign and laboratory abnormalities). Participants noted feeling disappointed about the predictive potential [30]:

> *The tool...was supposedly predictive, but we discovered...it wasn't predictive...it was really telling providers that they've met the criteria for severe sepsis which...is not really predictive because they've already met it. It wasn't that you were getting it before it happened so even though they were selling it as a predictive model I'm not so convinced it was predictive.*

In another study on the use of AI to support delirium prevention, the following was mentioned [27]:

> *[S]even users (14.9%) did not believe that the application is a useful support for delirium prevention, and seven did not believe that the application can be used to detect delirium at an early stage.*

Referring to the use of an AI tool for lung cancer diagnosis, a study reported the following [32]:

> *Radiologists were deeply committed to providing judgments with maximum certainty, but they expressed difficulty feeling certain given the opacity they experienced when considering divergent AI results: "I just don't know of any radiologist who's not looking closely at the case because they have AI. Because at the end of the day, you're still responsible. How can you trust the machine that much?"* [Dr E]

In a study among health care professionals on the use of BoneXpert, some health care professionals reported struggling to accept outputs from the application [41]:

> *Interestingly, in three hospitals, we found that the referring clinicians did not trust the output of the AI application and redid a manual bone age analysis for every scan. Thus, just like the radiologists, the referring clinicians showed varying levels of acceptance of AI applications.*

Divergence in the opinions of health care professionals and AI recommendations creates uncertainty. However, by using "AI interrogation practices," radiologists are able to use AI in a way that helps them to experience less uncertainty in making final judgments [32]:

> *On the surface, it may seem that using the AI tools (and experiencing opacity) increased the overall uncertainty these radiologists experienced; however, in fact, using the AI tool resulted in radiologists experiencing less uncertainty making their final judgments. They achieved this by using "AI interrogation practices" or practices that human experts enact to relate their own knowledge claims to the AI knowledge claims.*

Other experiences that led to lack of trust in AI tools were the lack of relevant empirical evidence on AI clinical performance and of its impact on health care workflow and quality and the problem for clinicians where AI results do not match their own judgments [32,41]. For example, in a study of the implementation of AI applications in radiology across 7 hospitals in the Netherlands, the authors reported the following [41]:

> *[T]here is a lack of empirical evidence on the effect of AI applications on the radiological workflow, as well as their added value for clinical radiology practice...measuring clinical and organizational benefits of AI on a microlevel is difficult...publications on the validation of the algorithms are based on laboratory rather than clinical settings.*

Societal expectations may also influence confidence in AI tool use. In a study of the use of the Watson tool in China [42], the authors reported that social media and organizations frequently talk about AI and attribute "magic" qualities to AI, which often leads to doctors being disappointed:

> *Hospital managers or doctors report to experiencing frustration when facing the real technology after the societal hype.*

In one study [43], participants suggested that having a track record of the actions of health care professionals based on AI recommendations and the corresponding patient outcomes could facilitate confidence in AI tools.

Judging the Added Value of AI

In some studies, health care professionals reported that AI tools provided added value and improved decision-making [19,20,24,27-30,35,36,39-41]. This included automation of burdensome tasks [41], avoidance of mistakes [32,41], reduction in variability in clinical practice [22], improvement in team communication [24,36], reduction of alert fatigue due to the combination of multiple data points [30], improvement in data collection [20] and team performance [19,27], and provision of more objective and precise diagnosis or prediction [41].

In some studies, some health care professionals found that AI tools served as a reminder or confirmation of their clinical judgments [23,26,27,29,32,35,38,39,42]. For example [27],

> *The prediction helps to corroborate my own estimation when seeing a patient.*

However, not all users of the same tool found it useful [27]:

> *Opinions about the application's usefulness for their own work were mixed: 17 users (36.2%) reported the application to be useful for their work, while 15 users (31.9%) did not find it useful.*

XSL•FO
RenderX

Health care professionals reported some problems relating to AI technologies not addressing issues that are deemed important to them. For example, health care professionals thought postoperative nausea and vomiting (PONV) has low burden and should not be excessively treated [31]. In a study in a US hospital among health care professionals using AI tools for diagnosing breast cancer, lung cancer, and bone age determinations [32], radiologists said AI tools identify abnormalities that they might not see but which were not clinically significant to impact their final judgment:

> *Calcifications can be really little and sometimes hard to see. It [Mammo AI tool] sees those calcifications better than I do. But it also sees all kinds of calcifications that are neither here nor there.* [Dr B]

In a study of peripheral arterial disease (PAD) identification algorithm [43], clinicians emphasized the importance of ensuring that the AI tools focus on supporting clinical decision-making by performing additional analysis that the clinicians do not normally do:

> *I mean, it [the algorithm] has to solve a real problem. Like, I'm not interested in models that in a clinical sense, identify data that I could just identify in the course of my daily work.*

Some studies also reported that health care professionals expressed uncertainty about next steps and that AI tools provided no actionable outputs, therefore not adding value to their clinical decision-making [31,38,43]; adding treatment recommendations would have been more useful [31]. For example, referring to a tool to predict the risks of PONV, a physician said the following [31]:

> *The intuitive use of the predicted PONV [postoperative nausea and vomiting] risk and a stated preference for an actionable recommendation by intervention group interviewees suggested that being presented only with a predicted risk may be difficult to use in a clinical decision. Adding a risk-corresponding treatment recommendation may assist physicians in interpreting the predicted risk for a decision on PONV prophylaxis.* [Theme 4G, quote 43]

Health care professionals also voiced concerns over AI tools not providing actionable information [38]:

> *It would have been nice to have some sort of actionable items, because while the information is good to know, I was never really sure what to do with it. It's like, great, my patient has an elevated one year mortality risk. What can I do about that? What do I do with this information?* [RES02]

From another study [40]:

> *They gave us the risk scores, but we did not really understand what to do with this information, what do these scores mean? It took us a while to figure out what is our threshold for what are we going to consider high risk patients.* [Readmission Risk Tool user]

Furthermore, in a study of an automated mortality prediction model for identification of patients for palliative care, AI helped reduce uncertainties, especially in less-experienced health care professionals [38]. In some situations, AI tools reportedly caused added work for health care professionals [21,29,34,37,41]. For example, in a study on the use of AI for the detection of diabetic retinopathy [21], the AI tool required high quality data and or images, which may be difficult to achieve in some cases. A study participant reported the following:

> *It gives guaranteed results, but it has some limitations. Some images are blurry, and I can still read it, but [the system] can't. P3 shared the same sentiment, "It's good but I think it's not as accurate. If [the eye] is a little obscured, it can't grade it." The system's high standards for image quality is at odds with the consistency and quality of images that the nurses were routinely capturing under the constraints of the clinic, and this mismatch caused frustration and added work.*

In one of the studies on the implementation of AI in diagnostic radiology [41], the clinical benefits or organizational goals for using AI applications were not clearly established before implementation, making it hard to assess after implementation:

> *From an organizational perspective, clinical benefits or organizational goals that might be achieved by using AI applications are not clearly established ex-ante and therefore hard to assess after implementation.*

Health care professionals were concerned that patients may not be willing to have AI tools used for them [42]:

> *They [the patients] have no idea about Watson [a system for designing personalized treatment for cancer patients]. They will think: why do I need a machine to look at [my problem]? I prefer an expert doctor.*

Health care professionals expressed concerns about limiting treatment for palliative care patients based on the risk prediction by AI tools [38]. There were also concerns about clinicians and regulatory agencies becoming overly reliant on AI [26,29]. For example, health care professionals may not actively consider their cases and they may not be able to refine their skills and maintain cultural competence [29].

Data Availability and Limitations of AI

A study reported the users' concern about the potential for AI to worsen inequity as disadvantaged populations are not represented in health datasets [43]:

> *About 20% of our population does not have data.... So we will be over-selecting and over-serving 80% and under-selecting and under-serving 20% of our population because of a structural dynamic here in the United States.*
>
> *And I should also say those even worse, those 20% [who we don't have data on], or a higher proportion of them are older, coming from an economically higher social deprivation index, higher disease burden, nonwhite.*

XSL•FO
RenderX

Concerns about the limits of AI in considering all the aspects of patients' lives were also talked about [31,43]. For example, in the study [31] on an AI tool for PONV, the following was discussed:

> *[T]hey [physicians] felt that a prediction model does not take into account all aspects of a specific patient. The prediction model only predicted the risk for a specific outcome and did not weigh the benefit of treatment against the expected harms and contraindications for a particular patient with specific characteristics and comorbidity.*

In the study of the use of PAD identification algorithm [43], health care professionals reported having access to more data about patient than AI algorithm uses. There were concerns about the AI tool not being able to account for factors such as patient's culture [29] and ground context and therefore limiting the potential impact of the AI [43]:

> *Exercise, nutrition, smoking, and medication adherence are critical components of PAD [peripheral arterial disease] management and clinicians expressed concern about adopting the algorithm to make intervention recommendations without knowing more about these factors...The algorithm's potential impact was limited by an inability to account for barriers patients faced in managing PAD.*

Time and Competing Priorities

Health care professionals reported lack of time to fully utilize AI tools. For example, in a study on the use of the modified Hospitalised-patient One-year Mortality Risk tool, physicians reported lacking time to address the alert because they were often focused on the acute needs of the patient [38]:

> *The inpatient stay often is very compressed. They're in the hospital, they're getting treated, and then they're home. And so, there is not time during the inpatient stay to address these things.*

In the study of the use of AI for PAD identification [43], health care professionals reported limited time to review high-risk patients and having difficulties in choosing which patients to prioritize:

> *The PAD [peripheral arterial disease] ML-driven CDS can be run on all patients in a population, but only a small number of high-risk patients could be reviewed each week. Stakeholders described the difficult trade-offs they had to make when considering how to pick patients to prioritize for Population Rounding.*

Time constraint could be further heightened where AI recommendations differ from the clinical judgment of the health care professionals [32]. However, in a study on the use of AI in diagnostic radiology, radiologists were willing to invest additional time to relate their own knowledge to the AI knowledge claims to build an understanding of the AI results and reconcile divergent views [32]:

> *I know my limitations and I know this [CT AI] is going to help them [nodules] stand out a little better. It's worth the extra time in my mind.*

Concern About Governance

Health care professionals expressed concerns over regulatory and legal uncertainties surrounding AI use [41,42]. For example, uncertainties regarding legal responsibilities for misdiagnosis stemming from AI recommendations were reported in a study in the Netherlands [41], and this may vary for different countries. In another study [42], a health care professionals reported that it is illegal for AI applications to make clinical decisions in China:

> *In China it is illegal for an AI system to make a decision.* [1HP05]

> *In our hospital, we use Watson to assist the Multidisciplinary Team (MDT).... We discussed how to use Watson for a long time.... As the first hospital to use Watson, we find this way [to use Watson together with the MDT].... But this really gives us a heavy burden! Because when we use Watson, we must have at least five doctors to work together with Watson [as required by regulation]. They [the 5 doctors] will sign on the report.* [1HP05]

Collaboration to Facilitate the Implementation and Use of AI

In a study [40] on the use of low bed tool and readmission risk tool, the authors described how health care organizations can enable collaboration between key stakeholders (users, developers, and outside experts) to facilitate the development and implementation of AI. This collaboration allowed users (care management and utilization management team members) to be involved in the identification of initial needs, identification of new users, need for experts, identification of potential sources of inaccuracy, and possible areas of improvement [40].

Having health care professionals serve as champions or leads for AI use and having knowledge exchange platforms supported the use of AI [26,41,43]. For example, the study [41] on the implementation of AI applications in clinical radiology reported that local champions (radiologists with keen interest in AI and good understanding of AI applications) are vital in encouraging implementation within their departments. In another study [43], having a physician lead the integration of AI tool for PAD identification was reported to be beneficial in facilitating the translation to create real-world clinical impact:

> *Dr. XXX obviously was very aware, she was kind of the one that spearheaded [the integration] and wanted to get this going. So I think you know, [targeting clinical need] is her role. I'm sure with her specialty, she probably realized that it was an area that needed more attention.* [Operational stakeholder]

Furthermore, in the study of the use of AI for early warning for sepsis [37], a new role of a sepsis watch nurse was created. The sepsis watch nurses were the primary users of the tool and were extensively trained in the workflow and the tool. They were responsible for communicating the outputs from the AI tool to

XSL•FO
RenderX

physicians to facilitate the translation of the recommendations to patient's bedside. With regard to the skills required to fulfill such roles, the nurses recommended the following [37]:

> *When asked about the skills and knowledge needed to be a good Sepsis Watch nurse, the RRT nurses mentioned good clinical judgment, knowledge of sepsis, and critical care experience...RRT nurses also explained the importance of strong communication skills to confidently speak with attending physicians whom they may not personally know...RRT nurses thought that strong computer skills were not necessary for the role, given the simplicity of the app. RRT nurses also recommended recruiting nurses interested in the role and the need to create buy-in through continuous feedback.*

In another study [26] on the early warning system for sepsis, the deployment team attended staff meetings and also met with individual users, as requested, to explain the system and provide guidance on using the interface. Users were also able to ask questions about the system's behavior using a feedback button [26].

In a study [41] on the implementation of AI tools in diagnostic radiology, it was reported that only 1 of 7 hospitals had a formalized innovation strategy regarding AI, although 3 hospitals had a designated innovation manager. It is worth noting that 4 more hospitals were reported to be developing a structured approach at the time of the study. The lack of structured implementation processes lead to substantial variations in how the application is used in different departments [41]:

> *From a workflow perspective, implementation plans do not specify how the AI application should be integrated into the clinical workflow, which leads to significant variations in the way the application is used in different departments. Furthermore, in all cases, the work done to monitor existing practices or the impact of the implementation of novel technologies on the level of the hospital is currently limited.*

Key Points From Stakeholder Engagement

A total of 18 individuals attended the stakeholder engagement workshop, besides the study team and 1 patient and public engagement officer. This included 9 patient and public representatives, 7 researchers from diverse related fields (including clinical researchers), 1 representative from an AI technology company, and 1 policy maker in data and AI. After being presented with the study findings, the participants agreed that algorithmic transparency and an understanding of how AI makes predictions may be challenging with non–knowledge-based AI. They believed that health care professionals are aware that AI will be used more in the future, but they do not feel prepared for it. The participants highlighted the need for more evidence to showcase good practice and collaborations across specialties or disciplines. They recommended early training of health care professionals to improve their understanding of what AI is and how it works. They acknowledged that some of the training programs may be based on theory rather than wait until AI is fully deployed in all clinical settings. They also recommended the need for AI applications guidelines to include examples of how conflicts between AI recommendations and health care professionals' judgments should be resolved. With regard to suggestions for future research, they suggested that more case studies and mixed methods research would be useful to provide insights into human perspectives of the quantitative research and explore health care professionals' confidence in AI systems. Participants recommended baseline studies on clinicians' knowledge and experiences of AI systems to provide comparison data for future work. There were also suggestions of a need for more information on the perspectives of hospital management and patients on the real-world use of AI systems. Differences in perspectives based on individual characteristics (such as age) could also be explored.

Discussion

Principal Findings

This review describes existing empirical evidence on health care professionals' real-world experiences of using non–knowledge-based CDSSs. While some of the issues emphasized in the included studies can be attributed to issues in implementation of any technology or intervention, we identified many issues that appeared to be particularly challenging for non–knowledge-based CDSSs. We grouped these issues into seven themes: (1) health care professionals' understanding of AI applications; (2) level of trust and confidence in AI tools; (3) judging the added value of AI; (4) data availability and limitations of AI; (5) time and competing priorities; (6) concern about governance; and (7) collaboration to facilitate the implementation and use of AI. The most frequently occurring issues included concerns over lack of understanding of AI outputs and the algorithm or rationale for the AI output. In addition, many studies highlighted the issue of trust and confidence in AI tools where some health care professionals believe that the AI tools function as anticipated while others hold contrary opinions. This lack of trust is further compounded when there are divergent opinions of health care professionals and AI recommendations. Another notable issue relates to the added value of AI. Some health care professionals reported that AI added value whereas some believed it did not add value, especially when the AI technologies did not address issues that were important to the health care professionals or failed to produce any actionable outputs. The challenges are interlinked; lack of understanding of the AI affects the ability to use the output and so trust cannot develop and health care professionals struggle to see the value in the AI tools and remain concerned about clinical responsibility and liability. The regulatory context and how the AI tool was implemented influence their approach to using it.

Comparison With Other Work

Other authors have found that the ability of health care professionals to trust non–knowledge-based CDSSs and understand how they work to benefit patients or improve care delivery are important factors in their adoption from the perspectives of both health care professionals and the public [44-46]. The lack of robust empirical evidence to support the

XSL•FO
RenderX

use of these systems in health care is thought to contribute to these issues [47,48]. One study used the expectancy-value theory [49] and modified extended Unified Theory of Acceptance and Use of Technology [50] to understand how expectancy (effort expectancy and performance expectancy), trust, and perceptions of clinicians related to their intention of using an AI-based CDSS: the Blood Utilisation Calculator [51]. The findings showed that expectancy and perceived risk on using the application have a substantial effect on trust [51]. Establishing a foundation of trust is crucial for the acceptance and effective integration of various AI technologies into health care practices [52].

It has been argued that for machine learning or AI systems to become trusted and accepted in health care, clinicians, researchers, and patients need to feel that the conclusions the systems make and the way that they reach them are interpretable and explainable [53]. This includes the clinician understanding the limitations of the system as a result of limitations of data quality [54] and the complexity of the data [55]. However, studies suggest that whether explainability does add value to AI powered clinical decision support depends on the context it is being used and by whom in addition to technical considerations [56,57]. In contrast, it has been argued that accuracy is sufficient for day-to-day clinical practice, with explainability being a research endeavor [58].

The view that algorithmic bias can exacerbate inequalities is supported by empirical evidence [59]. These biases can be mitigated to some extent [60]. Regulators recommend that the data used in training algorithms be representative of the intended patient population, and an international initiative is underway to develop recommendations for the composition and reporting of datasets used in AI systems for clinical practice [61]. However, as our findings indicate, health care professionals may have data about their patient that is missing from the datasets used in diagnostic CDSSs [4].

The inability of clinicians to process all the information presented to them is not new but potentially worsened by digitally available information, particularly in time-pressured clinical settings [62]. The need for health care organizations to optimize CDSS alerts is not limited to non–knowledge-based systems [63]. The concern expressed about clinical responsibility and liability for non–knowledge-based CDSSs is widely shared [64]. Regulation is in the works, but solutions are not straightforward as there is interaction between AI trustworthiness and transparency and how the AI tool and clinicians work together, which also changes the clinician-patient interaction [64,65]. An extensive systematic review of the barriers and enablers to the implementation of CDSSs for chronic disease identified similar implementation issues as in our review, such as the importance of a champion for the system, the deployment of allocated personnel to use the system, and the need for accessible training [66]. A recent study explored the experiences of various stakeholders, including health care professionals and regulatory bodies in developing and using AI technologies [67]. The study's recommendations for promoting the deployment of AI align with our findings and those identified in the stakeholder engagement workshop. These shared recommendations emphasize the importance of establishing guidelines for AI technology development and adoption; enhancing cocreation; and providing comprehensive education and training for health care professionals, patients, and communities [67].

Future Directions

Many studies have examined the perceptions of health care professionals on the use of AI in the health care settings [46,68]. However, the majority were not focused on AI applications that are fully deployed in real-world settings, and even fewer have focused on non–knowledge-based CDSSs. This review provides a more nuanced understanding of the challenges faced by health care professionals in contexts where non–knowledge-based systems are fully deployed. Many of the challenges identified in our review could be mitigated by collaboration with various stakeholders, including health care professionals, patients, and regulators, during the design and development stage of AI systems [69]. By involving key stakeholders in the process, developers gain invaluable insights into the practical needs and concerns of health care professionals, facilitating the creation of AI tools that are more aligned with the real-world requirements. Regular evaluation and validation of AI tools are essential to generate necessary data that are important for end users [70]. This will provide empirical evidence of the utility of the tools, clinical effectiveness, and generalizability [70]. Furthermore, the provision of education and training is necessary as this would equip health care professionals with the necessary knowledge and skills to enhance their confidence and competence in using AI in their practice [67]. With these issues solved as AI development matures, further research on the experience of health care professionals using AI can focus on identifying unintended consequences of AI use in routine clinical practice.

Limitations

The methods used in this systematic review were established a priori. We used a comprehensive search strategy, searched 4 electronic databases, and contacted experts to ensure we identify relevant studies. Due to limited time and human resources, we were unable to perform independent double screening at the title and abstract screening phase. However, at least 20% of all identified articles were randomly screened by a second reviewer and full-text articles were independently screened by 2 reviewers. Although we dedicated substantial effort to creating a comprehensive search strategy, numerous included studies were identified through recommendations rather than the structured search. This underscores the inherent challenge in rigorously searching databases for this topic, given the diverse language used to describe CDSSs. Moreover, it was difficult to judge sometimes whether the AI tool discussed in a study is knowledge based or non–knowledge based, particularly due to limited descriptions in some studies. Consequently, we recognize the possibility of overlooking some potentially relevant articles. Despite this limitation, our findings align with existing studies, and we believe that this review effectively captures the crucial aspects of health care professionals' experiences in this domain.

XSL•FO
RenderX

Conclusions

This review emphasizes the complex challenges faced by health care professionals using non–knowledge-based CDSSs. Issues related to health care professionals' understanding AI applications, trust in AI tools, and assessing added value are prominent, and the interlinked nature of the challenges identified is evident. As we navigate the evolving landscape of health care technologies, it is important to acknowledge and address these challenges through targeted strategies to improve collaborative development, continuous evaluation or validation, education, and refining the regulatory framework. This will potentially enhance the acceptance and use of AI within health care settings.

Acknowledgments

This project is funded by the Warwick AI National Strategy Fund—Alan Turing Award.

JH is funded by a National Institute for Health and Care Research (NIHR) doctoral fellowship (NIHR301467). AA is affiliated to the NIHR Applied Research Collaboration (ARC) West Midlands (NIHR200165). NP is funded by NIHR ARC Greater Manchester (NIHR200174). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

The authors are grateful to Samantha Johnson, academic support librarian at the University of Warwick, for her invaluable assistance in developing an effective search strategy that greatly enhanced the quality of their work. In addition, the authors extend their appreciation to Niyah Campbell, public involvement and engagement officer, at NIHR ARC West Midlands for his help in coordinating the stakeholder engagement workshop. They acknowledge the contributions from members of the Warwick Medical School Artificial Intelligence Consortium who participated in the stakeholder engagement workshop: Michael Adeyemi, Katherine Barrett, Nurgul Gecer, Lukas Hughes-Noehrer, Karen Keates, Melina Kunar, Sarah Markham, Siobhan O'Connor, Theodora Oikonomidi, Busola Oronti, Kate Ryan and Sagar Jilka. The authors extend their deepest appreciation to all the public representatives, health care professionals, experts in artificial intelligence, and policy makers who willingly shared their insights and perspectives, enriching the depth of their research. They would also like to acknowledge the Cross-ARC Artificial Intelligence Working Group for their collaborative efforts and valuable contributions throughout the course of this project.

Data Availability

All data generated or analyzed during this study are included in this published paper (and its supplementary information files).

Authors' Contributions

AA, FG, and NP were involved in the conceptualization and funding acquisition. AA, DOM, JW, IG, JH, NP, and FG contributed to the design of the review. AA, DOM, JW, IG, and SAI contributed to data curation. AA, DOM, JW, and FG performed formal analysis with input from all authors. AA wrote the original draft of the manuscript, and all authors reviewed the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for all databases.
[DOCX File , 39 KB - jmir_v26i1e55766_app1.docx]

Multimedia Appendix 2
Data extraction, quality assessment, and list of excluded studies.
[XLSX File (Microsoft Excel File), 117 KB - jmir_v26i1e55766_app2.xlsx]

Multimedia Appendix 3
PRISMA checklist.
[PDF File (Adobe PDF File), 938 KB - jmir_v26i1e55766_app3.pdf]

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
2. O'Dowd A. Government pins hopes on £250m AI centre for faster diagnosis and treatment. BMJ 2019 Aug 09;366:l5106. [doi: 10.1136/bmj.l5106] [Medline: 31399443]

XSL•FO
RenderX

3. Zhang J, Whebell S, Gallifant J, Budhdeo S, Mattie H, Lertvittayakumjorn P, et al. An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. Lancet Digit Health 2022 Apr;4(4):e212-e213 [FREE Full text] [doi: 10.1016/S2589-7500(22)00032-2] [Medline: 35337638]
4. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3(1):17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
5. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Mar 14;330(7494):765. [doi: 10.1136/bmj.38398.500764.8f]
6. Hazarika I. Artificial intelligence: opportunities and implications for the health workforce. Int Health 2020 Jul 01;12(4):241-245 [FREE Full text] [doi: 10.1093/inthealth/ihaa007] [Medline: 32300794]
7. Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022 Jul 01;22(1):850 [FREE Full text] [doi: 10.1186/s12913-022-08215-8] [Medline: 35778736]
8. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
9. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
10. Dwivedi YK, Rana NP, Jeyaraj A, Clement M, Williams MD. Re-examining the unified theory of acceptance and use of technology (UTAUT): towards a revised theoretical model. Inf Syst Front 2017 Jun 8;21(3):719-734. [doi: 10.1007/s10796-017-9774-y]
11. Dwivedi YK, Rana NP, Tamilmani K, Raman R. A meta-analysis based modified unified theory of acceptance and use of technology (meta-UTAUT): a review of emerging literature. Curr Opin Psychol 2020 Dec;36:13-18. [doi: 10.1016/j.copsyc.2020.03.008] [Medline: 32339928]
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
13. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med 2023 Jan 16;29(1):14-15. [doi: 10.1038/s41591-022-02139-w] [Medline: 36646804]
14. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/efi-180221]
15. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8(1):45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
16. Upadhyay S, Hu HF. A qualitative analysis of the impact of electronic health records (EHR) on healthcare quality and safety: clinicians' lived experiences. Health Serv Insights 2022 Mar 03;15:11786329211070722 [FREE Full text] [doi: 10.1177/11786329211070722] [Medline: 35273449]
17. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Apr 10;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
18. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
19. Gonçalves LS, Amaro ML, Romero AD, Schamne FK, Fressatto JL, Bezerra CW. Implementation of an artificial intelligence algorithm for sepsis detection. Rev Bras Enferm 2020;73(3):e20180421 [FREE Full text] [doi: 10.1590/0034-7167-2018-0421] [Medline: 32294705]
20. McAdam WA, Brock BM, Armitage T, Davenport P, Chan M, de Dombal FT. Twelve years' experience of computer-aided diagnosis in a district general hospital. Ann R Coll Surg Engl 1990 Mar;72(2):140-146 [FREE Full text] [Medline: 2185682]
21. Beede E, Baylor EE, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI p. 1-12 URL: https://dl.acm.org/doi/proceedings/10.1145/3313831
22. Cruz NP, Canales L, Muñoz JG, Pérez B, Arnott I. Improving adherence to clinical pathways through natural language processing on electronic medical records. Stud Health Technol Inform 2019 Aug 21;264:561-565. [doi: 10.3233/SHTI190285] [Medline: 31437986]

XSL•FO
RenderX

23. Frymoyer A, Schwenk HT, Zorn Y, Bio L, Moss JD, Chasmawala B, et al. Model-informed precision dosing of vancomycin in hospitalized children: implementation and adoption at an academic children's hospital. Front Pharmacol 2020 Apr 29;11:551 [FREE Full text] [doi: 10.3389/fphar.2020.00551] [Medline: 32411000]
24. Ginestra JC, Giannini HM, Schweickert WD, Meadows L, Lynch MJ, Pavan K, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med 2019 Nov;47(11):1477-1484 [FREE Full text] [doi: 10.1097/CCM.0000000000003803] [Medline: 31135500]
25. Henkel M, Horn T, Leboutte F, Trotsenko P, Dugas SG, Sutter SU, et al. Initial experience with AI pathway companion: evaluation of dashboard-enhanced clinical decision making in prostate cancer screening. PLoS One 2022 Jul 20;17(7):e0271183 [FREE Full text] [doi: 10.1371/journal.pone.0271183] [Medline: 35857753]
26. Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit Med 2022 Jul 21;5(1):97 [FREE Full text] [doi: 10.1038/s41746-022-00597-7] [Medline: 35864312]
27. Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. J Med Syst 2021 Mar 01;45(4):48 [FREE Full text] [doi: 10.1007/s10916-021-01727-6] [Medline: 33646459]
28. Jauk S, Veeranki SP, Kramer D, Högler S, Mühlecker D, Eberhartl E, et al. External validation of a machine learning based delirium prediction software in clinical routine. Stud Health Technol Inform 2022 May 16;293:93-100. [doi: 10.3233/SHTI220353] [Medline: 35592966]
29. Jordan M, Hauser J, Cota S, Li H, Wolf L. The impact of cultural embeddedness on the implementation of an artificial intelligence program at triage: a qualitative study. J Transcult Nurs 2022 Oct 08;34(1):32-39. [doi: 10.1177/10436596221129226] [Medline: 36214065]
30. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Jul;5(2):ooac022 [FREE Full text] [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
31. Kappen TH, van Loon K, Kappen MA, van Wolfswinkel L, Vergouwe Y, van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. J Clin Epidemiol 2016 Feb;70:136-145 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.09.008] [Medline: 26399905]
32. Lebovitz S, Lifshitz-Assaf H, Levina N. To engage or not to engage with AI for critical judgments: how professionals deal with opacity when using AI for medical diagnosis. Organ Sci 2022 Jan;33(1):126-148. [doi: 10.1287/orsc.2021.1549]
33. Marwaha A, Chitayat D, Meyn MS, Mendoza-Londono R, Chad L. The point-of-care use of a facial phenotyping tool in the genetics clinic: enhancing diagnosis and education with machine learning. Am J Med Genet A 2021 Apr 08;185(4):1151-1158. [doi: 10.1002/ajmg.a.62092] [Medline: 33554457]
34. Nehme F, Coronel E, Barringer DA, Romero LG, Shafi MA, Ross WA, et al. Performance and attitudes toward real-time computer-aided polyp detection during colonoscopy in a large tertiary referral center in the United States. Gastrointest Endosc 2023 Jul;98(1):100-9.e6. [doi: 10.1016/j.gie.2023.02.016] [Medline: 36801459]
35. Rabinovich D, Mosquera C, Torrens P, Aineseder M, Benitez S. User satisfaction with an AI system for chest X-ray analysis implemented in a hospital's emergency setting. Stud Health Technol Inform 2022 May 25;294:8-12. [doi: 10.3233/SHTI220386] [Medline: 35612006]
36. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. Int J Med Inform 2020 May;137:104072. [doi: 10.1016/j.ijmedinf.2019.104072] [Medline: 32200295]
37. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421. [doi: 10.2196/22421] [Medline: 33211015]
38. Saunders S, Downar J, Subramaniam S, Embuldeniya G, van Walraven C, Wegier P. mHOMR: the acceptability of an automated mortality prediction model for timely identification of patients for palliative care. BMJ Qual Saf 2021 Oct 25;30(10):837-840. [doi: 10.1136/bmjqs-2020-012461] [Medline: 33632758]
39. Shiang T, Garwood E, Debenedectis CM. Artificial intelligence-based decision support system (AI-DSS) implementation in radiology residency: introducing residents to AI in the clinical setting. Clin Imaging 2022 Dec;92:32-37. [doi: 10.1016/j.clinimag.2022.09.003] [Medline: 36183619]
40. Singer SJ, Kellogg KC, Galper AB, Viola D. Enhancing the value to users of machine learning-based clinical decision support tools: a framework for iterative, collaborative development and implementation. Health Care Manage Rev 2022;47(2):E21-E31. [doi: 10.1097/HMR.0000000000000324] [Medline: 34516438]
41. Strohm L, Hehakaya C, Ranschaert ER, Boon WP, Moors EH. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct 26;30(10):5525-5532 [FREE Full text] [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
42. Sun TQ, Medaglia R. Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. Gov Inf Q 2019 Apr;36(2):368-383. [doi: 10.1016/j.giq.2018.09.008]

XSL•FO
RenderX

43. Wang SM, Hogg HD, Sangvai D, Patel MR, Weissler EH, Kellogg KC, et al. Development and integration of machine learning algorithm to identify peripheral arterial disease: multistakeholder qualitative study. JMIR Form Res 2023 Sep 21;7:e43963 [FREE Full text] [doi: 10.2196/43963] [Medline: 37733427]
44. Fan W, Liu J, Zhu S, Pardalos PM. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Ann Oper Res 2018 Mar 19;294(1-2):567-592. [doi: 10.1007/s10479-018-2818-y]
45. Shinners L, Aggar C, Grace S, Smith S. Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: an integrative review. Health Informatics J 2020 Jun 30;26(2):1225-1236 [FREE Full text] [doi: 10.1177/1460458219874641] [Medline: 31566454]
46. Wu C, Xu H, Bai D, Chen X, Gao J, Jiang X. Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. BMJ Open 2023 Jan 04;13(1):e066322 [FREE Full text] [doi: 10.1136/bmjopen-2022-066322] [Medline: 36599634]
47. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJ, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open 2022 Sep 01;5(9):e2233946 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33946] [Medline: 36173632]
48. Jungmann F, Jorg T, Hahn F, Pinto Dos Santos D, Jungmann SM, Düber C, et al. Attitudes toward artificial intelligence among radiologists, IT specialists, and industry. Acad Radiol 2021 Jun;28(6):834-840. [doi: 10.1016/j.acra.2020.04.011] [Medline: 32414637]
49. Wigfield A, Eccles JS. Expectancy–value theory of achievement motivation. Contemp Educ Psychol 2000 Jan;25(1):68-81. [doi: 10.1006/ceps.1999.1015] [Medline: 10620382]
50. Chao CM. Factors determining the behavioral intention to use mobile learning: an application and extension of the UTAUT model. Front Psychol 2019 Jul 16;10:1652 [FREE Full text] [doi: 10.3389/fpsyg.2019.01652] [Medline: 31379679]
51. Choudhury A, Asan O, Medow JE. Effect of risk, expectancy, and trust on clinicians' intent to use an artificial intelligence system -- blood utilization calculator. Appl Ergon 2022 May;101:103708. [doi: 10.1016/j.apergo.2022.103708] [Medline: 35149301]
52. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023 Aug 24;3:1211150 [FREE Full text] [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
53. van der Schaar M, Maxfield N. Making machine learning interpretable: a dialog with clinicians. van der Schaar Lab. 2021. URL: https://www.vanderschaar-lab.com/making-machine-learning-interpretable-a-dialog-with-clinicians/ [accessed 2024-09-26]
54. Miller DD. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. NPJ Digit Med 2019 Jun 28;2(1):62 [FREE Full text] [doi: 10.1038/s41746-019-0138-5] [Medline: 31388566]
55. Chin-Yee B, Upshur R. Three problems with big data and artificial intelligence in medicine. Perspect Biol Med 2019;62(2):237-256. [doi: 10.1353/pbm.2019.0012] [Medline: 31281120]
56. Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, Z-Inspection initiative. To explain or not to explain?: artificial intelligence explainability in clinical decision support systems. PLOS Digit Health 2022 Feb 17;1(2):e0000016 [FREE Full text] [doi: 10.1371/journal.pdig.0000016] [Medline: 36812545]
57. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci 2021 May 31;11(11):5088. [doi: 10.3390/app11115088]
58. Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. Front Genet 2022 Sep 19;13:903600 [FREE Full text] [doi: 10.3389/fgene.2022.903600] [Medline: 36199569]
59. d'Elia A, Gabbay M, Rodgers S, Kierans C, Jones E, Durrani I, et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Fam Med Community Health 2022 Nov 30;10(Suppl 1):e001670 [FREE Full text] [doi: 10.1136/fmch-2022-001670] [Medline: 36450391]
60. Nazer LH, Zatarah R, Waldrip S, Ke JX, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023 Jun 22;2(6):e0000278 [FREE Full text] [doi: 10.1371/journal.pdig.0000278] [Medline: 37347721]
61. Ganapathi S, Palmer J, Alderman JE, Calvert M, Espinoza C, Gath J, et al. Tackling bias in AI health datasets through the STANDING together initiative. Nat Med 2022 Nov 26;28(11):2232-2233. [doi: 10.1038/s41591-022-01987-w] [Medline: 36163296]
62. Sbaffi L, Walton J, Blenkinsopp J, Walton G. Information overload in emergency medicine physicians: a multisite case study exploring the causes, impact, and solutions in four North England National Health Service Trusts. J Med Internet Res 2020 Jul 27;22(7):e19126 [FREE Full text] [doi: 10.2196/19126] [Medline: 32716313]
63. Van Dort BA, Zheng WY, Sundar V, Baysari MT. Optimizing clinical decision support alerts in electronic medical records: a systematic review of reported strategies adopted by hospitals. J Am Med Inform Assoc 2021 Jan 15;28(1):177-183 [FREE Full text] [doi: 10.1093/jamia/ocaa279] [Medline: 33186438]

XSL•FO
RenderX

64. Čartolovni A, Tomičić A, Lazić Mosler E. Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review. Int J Med Inform 2022 May;161:104738. [doi: 10.1016/j.ijmedinf.2022.104738] [Medline: 35299098]
65. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020 Apr 03;47(12):e3 [FREE Full text] [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]
66. Chen W, O'Bryan CM, Gorham G, Howard K, Balasubramanya B, Coffey P, et al. Barriers and enablers to implementing and using clinical decision support systems for chronic diseases: a qualitative systematic review and meta-aggregation. Implement Sci Commun 2022 Jul 28;3(1):81 [FREE Full text] [doi: 10.1186/s43058-022-00326-x] [Medline: 35902894]
67. Nix M, Onisiforou G, Painter A. Understanding healthcare workers' confidence in AI. National Health Service, UK. 2022. URL: https://tinyurl.com/yrhyp5sz [accessed 2024-04-29]
68. Hogg HD, Al-Zubaidy M, Talks J, Denniston AK, Kelly CJ, Malawana J, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
69. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
70. Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, et al. A path for translation of machine learning products into healthcare delivery. EMJ Innov 2020:1-14 [FREE Full text] [doi: 10.33590/emjinnov/19-00172]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
meta-UTAUT: meta-analysis–based Unified Theory of Acceptance and Use of Technology
PAD: peripheral arterial disease
PONV: postoperative nausea and vomiting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-AI: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Artificial Intelligence

Edited by A Mavragani; submitted 22.12.23; peer-reviewed by F Burstein, L Shinners; comments to author 21.03.24; revised version received 10.06.24; accepted 25.07.24; published 30.10.24.

Please cite as:
Ayorinde A, Mensah DO, Walsh J, Ghosh I, Ibrahim SA, Hogg J, Peek N, Griffiths F
Health Care Professionals' Experience of Using AI: Systematic Review With Narrative Synthesis
J Med Internet Res 2024;26:e55766
URL: https://www.jmir.org/2024/1/e55766
doi:10.2196/55766
PMID:

©Abimbola Ayorinde, Daniel Opoku Mensah, Julia Walsh, Iman Ghosh, Siti Aishah Ibrahim, Jeffry Hogg, Niels Peek, Frances Griffiths. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Relation Between Passively Collected GPS Mobility Metrics and Depressive Symptoms: Systematic Review and Meta-Analysis

Yannik Terhorst[1,2,3], BSc, MSc; Johannes Knauer[1], BSc, MSc; Paula Philippi[4], BSc, MSc; Harald Baumeister[1], Prof Dr

[1]Department of Clinical Psychology and Psychotherapy, Institute of Psychology and Education, University Ulm, Ulm, Germany

[2]Department of Psychology, Ludwig Maximilian University of Munich, Munich, Germany

[3]German Center for Mental Health (DZPG), Partner Site Munich-Augsburg, Munich, Germany

[4]Department of Clinical Child and Adolescent Psychology and Psychotherapy, Institute of Psychology, University of Wuppertal, Wuppertal, Germany

Corresponding Author:
Yannik Terhorst, BSc, MSc
Department of Clinical Psychology and Psychotherapy
Institute of Psychology and Education
University Ulm
Lise-Meitner-Str. 16
Ulm, 89081
Germany
Phone: 49 7315032820
Email: yannik.terhorst@uni-ulm.de

Abstract

Background: The objective, unobtrusively collected GPS features (eg, homestay and distance) from everyday devices like smartphones may offer a promising augmentation to current assessment tools for depression. However, to date, there is no systematic and meta-analytical evidence on the associations between GPS features and depression.

Objective: This study aimed to investigate the between-person and within-person correlations between GPS mobility and activity features and depressive symptoms, and to critically review the quality and potential publication bias in the field.

Methods: We searched MEDLINE, PsycINFO, Embase, CENTRAL, ACM, IEEE Xplore, PubMed, and Web of Science to identify eligible articles focusing on the correlations between GPS features and depression from December 6, 2022, to March 24, 2023. Inclusion and exclusion criteria were applied in a 2-stage inclusion process conducted by 2 independent reviewers (YT and JK). To be eligible, studies needed to report correlations between wearable-based GPS variables (eg, total distance) and depression symptoms measured with a validated questionnaire. Studies with underage persons and other mental health disorders were excluded. Between- and within-person correlations were analyzed using random effects models. Study quality was determined by comparing studies against the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) guidelines. Publication bias was investigated using Egger test and funnel plots.

Results: A total of k=19 studies involving N=2930 participants were included in the analysis. The mean age was 38.42 (SD 18.96) years with 59.64% (SD 22.99%) of participants being female. Significant between-person correlations between GPS features and depression were identified: distance (r=–0.25, 95% CI –0.29 to –0.21), normalized entropy (r–0.17, 95% CI –0.29 to –0.04), location variance (r–0.17, 95% CI –0.26 to –0.04), entropy (r=–0.13, 95% CI –0.23 to –0.04), number of clusters (r=–0.11, 95% CI –0.18 to –0.03), and homestay (r=0.10, 95% CI 0.00 to 0.19). Studies reporting within-correlations (k=3) were too heterogeneous to conduct meta-analysis. A deficiency in study quality and research standards was identified: all studies followed exploratory observational designs, but no study referenced or fully adhered to the international guidelines for reporting observational studies (STROBE). A total of 79% (k=15) of the studies were underpowered to detect a small correlation (r=.20). Results showed evidence for potential publication bias.

Conclusions: Our results provide meta-analytical evidence for between-person correlations of GPS mobility and activity features and depression. Hence, depression diagnostics may benefit from adding GPS mobility and activity features as an integral part of future assessment and expert tools. However, confirmatory studies for between-person correlations and further research on within-person correlations are needed. In addition, the methodological quality of the evidence needs to improve.

Trial Registration: OSF Registeries cwder; https://osf.io/cwder

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51875) doi:10.2196/51875

KEYWORDS

smart sensing; digital phenotyping; depression; GPS; global positioning system; meta-analysis; mobile phone; depressive symptoms; smartphone; systematic review; depressive disorders; treatment; mental disorder; mental health; wearable

Introduction

Depressive disorders are one of the most prevalent mental disorders worldwide. The global prevalence rate is estimated to be 4.4% [1,2]. Associated consequences of depression are severe not only for affected individuals (eg, globally, 43 million years lived with disability in 2017) but also for the economy and society [3-5]. Costs are even higher and long-term health consequences are more severe if the disorder is not properly diagnosed or treated [6,7]. To target the global burden caused by depression and to initiate effective treatment, timely diagnosis is a key bottleneck in health care [8-10].

Various diagnostic approaches like structured clinical interviews [11], clinician-rated screening scales [12,13], and self-report screening assessments [14] are available. However, their reliability and accuracy are often hindered by social desirability, recall, or confirmation biases [15-17]. Therefore, augmentations and extensions of existing approaches by objective data sources could potentially make an important contribution to the improvement in the diagnostic process of depression and other mental disorders.

Facing an ever-growing digitalization in daily living, the availability of various sensor data in smartphones, wearable devices, cars, and smart home devices may provide a way toward objective diagnosis in a timely manner and sensor-informed diagnostic support tools for clinical personnel [18-21]. In the context of medical apps, the approach of using sensor data to infer mental health is referred to as smart sensing or digital phenotyping [18,19]. In short, raw data from software (eg, screen status of the smartphone [on or off]) and hardware-based (eg, GPS coordinates) sensors are processed to derive higher-level features (eg, total smartphone usage time, total distance, and circadian rhythm), which are then linked to clinical symptoms or diagnoses [18,19].

Earlier studies highlight the potential of smart sensing [22]: For instance, Opoku Asare et al [23] found an area under the curve (AUC) of the receiver operating curve (ROC) of 94.69% to 99.06% in the prediction of depression status (depressed or not depressed) using a supervised machine learning model on behavioral markers from the smartphone (ie, app usage, screen usage, and network usage features). Furthermore, studies have shown that smart sensing data can improve prediction and increase explained variance over self-report ratings such as ecological momentary assessments [24,25]. In line with these findings, various descriptive reviews underline the potential shown in the present literature [26-33].

While there is a broad array of sensors in the context of depression, in particular, GPS-based mobility and activity features (eg, total distance, places visited, and time spent at home) could be promising candidates for objective and unobtrusive measurement of core symptoms of depression such as reduced activity, fatigue, loss of functioning, and diminished interest [24,27,34,35]. Individual studies find medium to high correlations between depressive symptoms and GPS mobility and activity features (eg, number of location clusters: r=–0.38, circadian movement: r=–0.34, home stay: r=0.22 [34]; eg, [24,35-37]). However, the heterogeneity in the findings and in the methodology between studies is substantial [27,31]. In addition, a mean sample size of 23.1 (SD 27.9) was reported in the systematic review of Cornet and Holden [26], clearly highlighting the issue of underpowered trials in the field. Meta-analytical evidence is highly needed to provide an estimate of the relationship between GPS mobility and activity features and depressive symptoms to determine their applicability for clinical settings and to transfer research from exploratory investigations to confirmatory studies. However, to date, no quantitative meta-analysis has been conducted on the associations between GPS mobility and activity features and depression.

Therefore, this study systematically reviewed and meta-analyzed the current evidence on the associations between GPS mobility and activity features and depression. Besides, core aspects of study quality (eg, adherence to international reporting guidelines, preregistrations, presence of a-priori and post hoc power analysis [31,38,39]) and the potential of publication bias in the field were investigated. Research questions are (1) What is the pooled between-person correlation of GPS mobility and activity features and depression severity? (2) What is the pooled within-person correlation of GPS mobility and activity features and depression severity? (3) To what extent are international reporting guidelines (ie, STROBE [Strengthening the Reporting of Observational studies in Epidemiology] [38]) reflected in the studies? and (4) Are small study effects as an indicator of potential publication bias observable in the literature?

Methods

Study Design

This study is a systematic review and meta-analysis. All procedures have been preregistered in the open science framework under trial registration link. Reporting is conducted in accordance with the updated PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for the reporting of systematic reviews [40] (Multimedia Appendix 1).

Search and Eligibility Criteria

Given the interdisciplinary nature of the field, we searched MEDLINE, PsycINFO, Embase, CENTRAL, ACM, IEEE Xplore, PubMed, and Web of Science. The database-specific search strings can be found in Multimedia Appendix 2. After the removal of duplicates by automatic tools of the databases, all identified articles were screened in a two-stage process: (1) title and abstract screening and (2) full-text screening, both

XSL•FO
RenderX

performed by 2 independent reviewers (YT and JK). If duplicates were encountered in this screening process, they were removed by hand. In addition to the database search, we searched all reference lists of the included studies for further eligible studies. This search and inclusion process was started on December 6, 2022 (database searches from December 06 until June 13), and completed by March 24, 2023 (last reference search). Any disagreements between the reviewers were resolved in discussions.

To be eligible, the following inclusion criteria had to be met: (1) any kind of GPS sensor data collection, (2) the data were collected by a wearable device such as smartphone or smartwatch, (3) an assessment of depressive symptoms was conducted either by self-report or by clinical diagnostic scales, and (4) reported outcomes included correlations between the collected GPS sensor data and depressive symptoms. Studies that (1) comprised participants younger than 18 years of age or (2) included participants with disorders other than depression (eg, bipolar disorder) were excluded.

Measured Variables and Coding

All data were extracted by 2 independent reviewers (JK and YT). The following study characteristics and empirical data points were obtained.

Study Characteristics

To describe the included studies, we extracted the authors' names, publication year, measures of depression (eg, Patient Health Questionnaire-9 [PHQ-9]), GPS-sensor features (ie, mobility and activity features such as total distance), study design, study setting, sensing framework (app name) and sample characteristics (age, gender, population, and country).

Quantitative and Empirical Data

For the statistical analysis, correlation coefficients (both between-person and within-person) and sample size were extracted (see further analysis details below). Between-person correlations were defined as the interindividual associations between 2 variables (eg, do individuals with higher time spent at home tend to show higher depression?). Within-person correlations were defined as intraindividual associations of 2 variables across time (eg, does an individual tend to show higher depression in weeks with more time spent at home?). Corresponding authors of eligible studies with insufficiently reported information for meta-analysis (eg, reporting of *P* values without correlation coefficients) were contacted for the missing information (ie, repeated email containing a study description and extraction template for the needed information). Furthermore, all included studies were compared against international guidelines for reporting observational studies (STROBE [38]). In addition, the presence of a preregistration and a priori or post hoc power analysis were rated as additional criteria (assessment of research standards and small study effects are provided below). The rating was performed independently by 2 researchers (JK and YT). Disagreements were resolved in discussion. Study designs other than exploratory, observational, intensive, longitudinal designs were planned to be compared against corresponding guidelines of the EQUATOR Network but were not present in the included studies.

Statistical Analysis

Meta-Analysis of Correlation

We conducted a random effects meta-analysis of correlations. Pooling was based on inverse variance weighting [41]. Maximum likelihood was used as the estimator. Following the Cochrane handbook for meta-analysis, the minimum required number of studies reporting on the correlations between a feature (eg, home stay) and depression to run meta-analysis was 2 [42]. For all significance tests and range of CI, was set to 5%. Heterogeneity of effect sizes was evaluated by I^2 [43]. For heterogeneity, we defined an I^2 of 25% as the threshold for low, 50% for moderate, and 75% for high heterogeneity in the present meta-analysis. 95% CI and prediction intervals were calculated along the pooled correlation estimates.

Assessment of Research Standards and Small Study Effects

As a general criterion for the adherence to international reporting guidelines, we assessed the reference and adherence to the STROBE guidelines [38] of the included observational studies. The STROBE checklist was rated for each study by 2 independent researchers (YT and JK). Disagreements were resolved in discussion.

Facing the replication crisis in research [39,44], we additionally investigated the percentage of preregistrations and the presence of a priori and post hoc power calculation. Potential publication bias was investigated using (1) funnel plots and (2) Egger test [45,46]. Egger test was only conducted in analyses with at least 10 studies [47]. Funnel plots display the effect sizes reported in studies (x-axis) in respect to the SE (y-axis). The core assumption is that in case of no publication bias, all points in the funnel plot should be distributed equally around an average effect with the high precision studies (low SE) at the top of the figure and closely around the average, and studies with low precision (high SE) should be broadly distributed around the average effect. In contrast to this, a scenario with publication bias (eg, small studies are only published if they show high effects, while large studies with lower effects are published either way due to the number of participants) would lead to an asymmetric funnel plot, where an association between SE and reported effect is visible. Capitalizing on this idea, Eggert test is a regression model investigating whether the SE is significantly influencing the average reported effect. Both asymmetries in funnel plots and significant Eggert tests indicate potential publication bias, and hence, potentially biased meta-analytically results (eg, systematic overestimation of the true effect due to unpublished high SE studies with no effects). For a more in-depth introduction, refer previously published studies (eg, [41]).

Software

All analyses were conducted in R (R Foundation for Statistical Computing). *Meta* was used as the core package in the analysis [48]. For an overview of all loaded packages and versions, see Multimedia Appendix 3. Both the analysis code and the dataset containing the correlations are available under CC-BY 4.0

XSL•FO
RenderX

license at the Open Science Framework (Multimedia Appendix 4).

Results

Study Selection

We identified a total of k=9499 unique records in the systematic literature search. One additional study was screened and included after being forwarded by researchers contacted due to insufficient data reported in an identified study. In the following screening and inclusion process conducted by 2 independent reviewers (YT and JK), k=19 studies were finally included [24,34-37,49-62]. In total, 2 of the studies were eligible for between- and within-person correlation meta-analysis [34,53]. For further details, please refer to the PRIMSA flowchart in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. Records identified from (*) MEDLINE, PsycINFO, Embase, CENTRAL, ACM, IEEE Xplore, PubMed, and Web of Science.

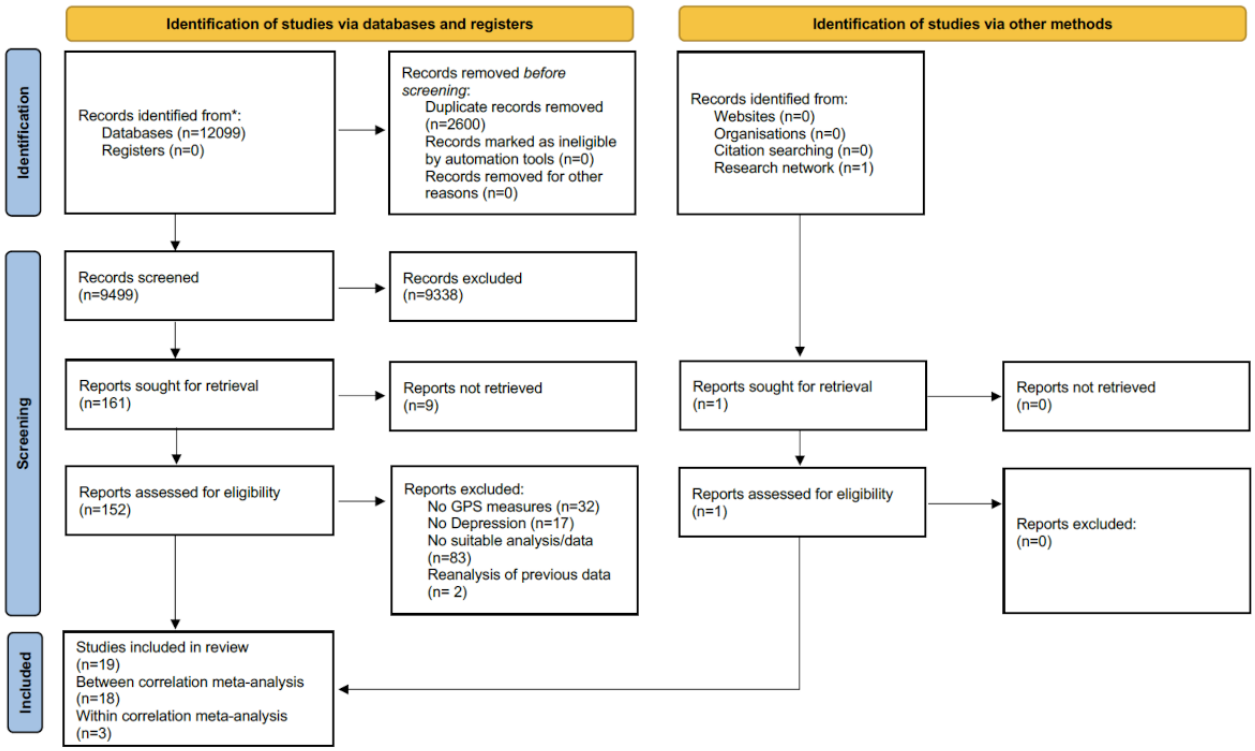

Study Characteristics

Combined, the included studies comprised a total of N=2930 participants. The sample sizes ranged from a minimum of n=18 to a maximum of n=1046 (mean 154.21, SD 235.54; median 72, IQR 83). The mean age across all included studies was 38.42 (SD 18.96) years. The average percentage of female participants was mean 59.64% (SD 22.99%). Of the k=19 included studies, k=8 (42%) studies explicitly targeted students and young adults (eg, recruitment at universities), k=3 (16%) studies recruited from the general public, k=5 (26%) studies used nontailored online and social media recruitment, k=2 (11%) studies targeted older adults, and k=1 (5%) study provided no further information on the recruitment strategies and targeted population. Please refer to Table 1 for further details on the descriptive characteristics of the studies.

A total of k=13 unique sensing frameworks were applied among the included studies. PurpleRobot was the most frequently used framework (k=3/19, 16%). Based on the sensed GPS data, the studies reported n=12 distinct activity and mobility features. The number of unique studies per activity and mobility feature ranged from k=1 to k=14 and a combined sample size of n=69 to n=2287 per feature. A summary of the features can be found in Table 2 for between-person features and Table 3 for within-person features.

XSL•FO
RenderX

Table 1. Main characteristics of included studies.

Reference	Country	N	Depression scale	Age (y), mean (SD)	Software	Period	Target population	Severity
Boukhechba et al [56]	United States	72	DASS-21[a]	19.8 (2.4)	Sensus	2 weeks	Students and young adults	3.5
Canzian and Musolesi [57]	United Kingdom	28	PHQ-8[b]	31 (—[c])	Mood- Traces	Average 71 days	General public	—
Currey and Torous [37]	United States	147	PHQ-9[d]	—	mind-LAMP	4 weeks	Students and young adults	—
DeMasi and Recht [58]	United states	33	BDI[e]	—	—	8 weeks	Students and young adults	12.7
Di Matteo [59]	Canada	71	PHQ-8	30.6 (9.4)	Logger	2 weeks	Nontailored online and social media recruitment	9.1
Farhan et al [62]	United States	79	PHQ-9	—	Life-Rhythm	—	Students and young adults	—
Giannouli et al [60]	Germany	69	GDS[f]	69.5 (4.9)	Ufall & custom app	1 week	Older adults	1.39
Lu et al [50]	United States	103	QIDS[g]	—	Life-Rhythm & Fitbit	—	Students and young adults	—
MacLeod et al [49]	Canada	121	CES-DC[h]	18 (2.76)	Custom app	2 weeks	Students and young adults	32.6
Stamatis et al[i] [61]	United States	1046	PHQ-8	40.9 (—)	LifeSense	16 weeks	Nontailored online and social media recruitment	9.14
Moshe et al [24]	Various European	55	DASS-21	42.8 (11.6)	AWARE	30 days	Nontailored online and social media recruitment	3.8
Nickels et al [51]	United States	379	PHQ-9	—	—	12 weeks	People who are or are not depressed	—
Saeb et al (2015a) [35]	United States	28	PHQ-9	28.9 (10.1)	Purple-Robot	2 weeks	General public	5.6
Saeb et al (2015b) [54]	—	18	PHQ-9	—	Purple-Robot	2 weeks	Nontailored online and social media recruitment	5.8
Saeb et al (2016) [34]	United States	48	PHQ-9	—	Student-Life	10 weeks	Students and young adults	—
Saeb et al (2017) [55]	United States	206	PHQ-9	39.3 (10.3)	Purple-Robot	6 weeks	General public	9.7
Tung et al [52]	Canada	54	GDS	72.6	VALMA	3 days	Older adults	—
Wang et al [36]	United States	83	PHQ-8 and PHQ-4[j]	20.1 (—)	Student-Life	9 weeks	Students and young adults	6.1
Zhang et al [53]	Netherlands, Spain, United Kingdom	290	PHQ-8	—	—	2 weeks	—	—

[a]DASS-21: Depression Anxiety and Stress Scale.

[b]PHQ-8: Patient Health Questionnaire-8.

[c]Not applicable.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]BDI: Beck's Depression Inventory.

[f]GDS: Geriatric Depression Scale.

[g]QIDS: Quick Inventory of Depression Symptomatology.

[h]CES-DC: Center for Epidemiological Studies–Depression.

[i]Stamatis et al [61] provided additional data on depression and GPS mobility metrics obtained from the LifeSense project for this meta-analysis. For further details on the LifeSense project, please refer to [61,63,64].

[j]PHQ-4: Patient Health Questionnaire-4.

XSL•FO
RenderX

Table 2. Between-person features.

GPS feature	Definition	Studies	Frequency	Total N	Value, mean (SD)
Homestay	Percentage of time spent home	Boukhechba et al [56], Currey and Torous [37], Di Matteo [59], Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], Stamatis et al[a] [61], Moshe et al [24], Nickels et al [51], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al 2016 [34], Saeb et al (2017) [55]	14	2216	158.29 (266.4)
Location variance	Variance of latitude and longitude values	Currey and Torous [37], DeMasi and Recht [58], Di Matteo [59], Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], Stamatis et al [61], Moshe et al [24], Nickels et al [51], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34], Zhang et al [53]	14	2287	163.36 (276.1)
Entropy	Distribution of time spent at different location clusters	Currey and Torous [37], Di Matteo [59], Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], Stamatis et al [61], Moshe et al [24], Nickels et al [51], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34], Zhang et al [53]	13	2254	173.38 (284.71)
N clusters	Number of unique location clusters	Currey and Torous [37], Di Matteo [59], Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], Stamatis et al [61], Nickels et al [51], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34], Wang et al [36], Zhang et al [53]	13	2282	175.54 (283.85)
Norm entropy	Entropy normalized by the number of location clusters	Di Matteo [59], Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], MacLeod et al [49], Stamatis et al [61], Moshe et al [24], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34], Zhang et al [53]	12	1849	154.08 (290.46)
Distance	Total distance between coordinates	Di Matteo [59], Lu et al (Android) [50], Lu et al (iOS) [50], MacLeod et al [49], Stamatis et al [61], Moshe et al [24], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34], Tung et al [52], Zhang et al [53]	11	1824	165.82 (301.64)
Speed moving	Speed at GPS data point collection	Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], Stamatis et al [61], Saeb et al (2016) [34]	6	1266	211 (409.36)
Time moving	Time spent in moving states in percentage	Farhan et al (Android) [62], Farhan et al (iOS) [62], Lu et al (Android) [50], Lu et al (iOS) [50], MacLeod et al [49], Zhang et al [53], Di Matteo [59], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34]	10	748	74.8 (81.54)
Circadian movement	Amount of energy in frequency periods (eg, 30 min) based on least-squares spectral analysis	Di Matteo [59], Stamatis et al [61], Saeb et al (2015a) [35], Saeb et al (2015b) [54], Saeb et al (2016) [34]	5	1201	240.2 (450.9)
Life space area	Convex hull of GPS coordinates	Giannouli et al [60]	1	69	69 (—[b])
Maximum action range	Longest (straight-line) distance away from home	Giannouli et al [60]	1	69	69 (—)

[a]Stamatis et al [61] provided additional data on depression and GPS mobility metrics obtained from the LifeSense project for this meta-analysis. For further details on the LifeSense project, please refer to [61,63,64] [b]Not applicable.

[b]Not applicable.

XSL•FO
RenderX

Table 3. Within-person features.

GPS feature	Definition	Studies	Frequency	Total N	Value, mean (SD)
Distance	Total distance between coordinates	Canzian and Musolesi [57], Saeb et al (2016) [34], Zhang et al [53]	3	356	118.67 (148.46)
Entropy	Distribution of time spent at different location clusters	Saeb et al (2016) [34], Zhang et al [53]	2	328	164 (178.19)
Location variance	Variance of latitude and longitude values	Saeb et al (2016) [34], Zhang et al [53]	2	328	164 (178.19)
N clusters	Number of unique location clusters	Saeb et al (2016) [34], Zhang et al [53]	2	328	164 (178.19)
Norm entropy	Entropy normalized by the number of location clusters	Saeb et al (2016) [34], Zhang et al [53]	2	328	164 (178.19)
Circadian movement	Amount of energy in frequency periods (eg, 30 min) based on least-squares spectral analysis.	Saeb et al (2016) [34]	1	38	38 (—[a])
Homestay	Percentage of time spent home	Saeb et al (2016) [34]	1	38	38 (—)
Speed moving	Speed at GPS data point collection	Saeb et al (2016) [34]	1	38	38 (—)
Time moving	Time spent in moving states in percentage	Saeb et al (2016) [34], Zhang et al [53]	2	328	164 (178.19)

[a]Not applicable.

Meta-Analysis: Between Person Correlations

Distance was the mobility and activity feature most strongly associated with depressive symptoms, with a meta-analytically pooled between-person correlation of r=–0.25 (95% CI –0.29 to –0.21), followed by normalized entropy (r=–0.17, 95% CI –0.29 to –0.04), location variance (r=–0.17, 95% CI –0.26 to –0.06), entropy (r=–0.13, 95% CI –0.23 to –0.04), number of clusters (r=–0.11, 95% CI –0.18 to 0.03), and home stay (r=0.10, 95% CI 0.00 to 0.19). In contrast, the features circadian movement, transition time, speed moving and time moving indicate no significant correlations with depression. Table 4 summarizes all of the pooled between-person correlations. Feature-specific forest plots can be found in Multimedia Appendix 5.

Table 4. Pooled between-person correlations.

Feature	N	Correlation	95% CI	Prediction interval	I^2
Circadian movement	1201	–0.36	–0.71 to 0.12	–0.91 to 0.64	86%
Distance	1824	–0.25	–0.29 to –0.21	–0.30 to –0.20	0%
Norm entropy	1849	–0.17	–0.29 to –0.04	–0.44 to 0.13	69%
Location variance	2287	–0.17	–0.26 to –0.06	–0.36 to 0.04	58%
Entropy	2254	–0.13	–0.23 to –0.04	–0.35 to 0.10	57%
N cluster	2282	–0.11	–0.18 to –0.03	–0.25 to 0.04	36%
Homestay	2216	0.10	0.00 to 0.19	–0.09 to 0.27	50%
Speed moving	1266	–0.01	–0.07 to 0.06	–0.09 to 0.07	0%
Time moving	748	–0.05	–0.21 to 0.11	–0.45 to 0.36	68%

Meta-Analysis: Within-Person Correlations

The 3 identified studies reporting on within-person correlations (Table 3) differed widely in the applied methodology of analysis. Saeb et al [34] correlated GPS mobility and activity features with the change in PHQ-9 scores, while Zhang et al [53] used autoregressive models to estimate the correlations over time, and Canzian and Musolesi [57] applied time series analysis to investigate the correlations within each participant across multiple days. Due to this heterogeneity in these studies to derive the reported correlations, we did not perform a meta-analysis for within-person correlations.

Assessment of Research Standards and Small Study Effects

All included studies (k=19) followed an observational study design. Assessment of the international reporting standards for observational studies revealed that not a single study referenced

XSL•FO
RenderX

the international reporting guidelines STROBE. Comparing the k=18 included published studies against the STROBE checklist showed an overall agreement across all items of 73.84% (Stamatis et al [61] excluded from comparison as the provided data originated from additional analyses not covered in that study). Only k=9 (50%) of the studies reported how potential sources of bias were addressed. Regarding missing data, k=6 (33%) reported how missing data was handled and k=3 (17%) reported rates of missingness for the variables of interest. A total of k=7 (39%) of the studies listed reasons for the exclusion of participants at each stage of the study. For all STROBE ratings, please refer to Multimedia Appendix 4.

In addition, neither preregistrations nor study protocols were found for the included studies. A priori power analyses for sample size planning or post hoc for discussion were also not reported in any of the included studies. Power analyses of the included studies indicate that k=16 (84%) studies were sufficiently powered with a power of 80% to detect a correlation of r=.50, k=7 (37%) studies for r=.30, k=4 (21%) studies for r=.20, and k=1 (5%) studies for r=.10.

For all mobility and activity features except for time moving, Egger test showed a significant asymmetry in the funnel plot (Table 5). Funnel plots for features with less than k=10 studies, in addition indicate asymmetry. Please refer to Multimedia Appendix 6 for the funnel plots of all between-person features.

Table 5. Eggert test for funnel plot asymmetry.

Feature	Intercept	95% CI	*P* value
Homestay	1.55	0.56 to 2.54	<.001
Location variance	–1.44	–2.58 to –0.31	.03
Entropy	–2.11	–2.87 to –1.34	<.001
N clusters	–1.29	–2.22 to –0.35	.02
Norm entropy	–2.30	–3.71 to –1.33	<.001
Distance	1.01	0.46 to 1.56	.01
Time moving	–0.42	–2.74 to 2.66	.98

Discussion

Principal Findings

This systematic review with meta-analysis provides pooled correlation estimates for the between-person correlation of GPS-based smart sensing mobility and activity features and depression. We identified robust small to medium between-person correlations for multiple mobility and activity features (ie, distance, normalized entropy, entropy, homestay, number of clusters, and homestay). In contrast to the findings on between-person correlations, this literature for within-person correlations did not allow for meaningful meta-analyses. Furthermore, this study clearly highlights a lack of quality in the literature. While all studies followed an observational study design, not a single study referenced the international reporting guidelines for observational studies (STROBE [38]). Although the reporting in studies addressed STROBE items in most cases (14/19, 74%), deficits were found in addressing sources of bias and missing data. Moreover, we found strong evidence for small study effects potentially indicating publication bias. Most studies were underpowered to detect correlations of the magnitude indicated by the meta-analysis.

Overall, the present analysis highlights the potential of GPS features to infer depression, which is in line with previous published descriptive reviews in the field [26,27,30,32,33]. For instance, with a total sample size of N=1824 and homogenous findings across studies (I^2=0%), distance (r=–0.25) might be a promising objective and unobtrusively collectible GPS feature in expert systems for (assisted) diagnoses, just-in-time interventions and personalization of treatment of depression in future [20,65]. Analogously, normalized entropy, entropy, homestay, number of clusters, and homestay were significant markers (r=0.10-0.17). However, it is important to note that all studies followed an exploratory design. Both the correlations and the clinical feasibility of such smart sensing augmented expert systems need to be proven in confirmatory trials before clinical applications. The identified pooled correlations and their CI offer a strong foundation for a-priori power analysis to guide confirmatory studies in their study and sample planning [66,67].

However, while some of the investigated correlations are statistically significant and pooled estimates can serve as a basis for future study planning from a statistical point, this does not release from the question of what a clinically relevant correlation is. The process involved in obtaining GPS features (eg, collection of GPS coordinates) as well as the features itself (eg, how much time was spent at home or at work) are highly sensitive. Hence, a discussion taking the clinical significance as well as ethical and privacy issues into account is of utmost importance [20,21,65,68-70]. Furthermore, it needs to be discussed to which extent GPS features as a standalone metric can be used in the diagnostic process or rather as an add on to existing procedures and measurements (eg, patient-reported outcome, ecological momentary assessment, medical record data) [20,24,71]. This discussion would strongly benefit from future studies illuminating these questions from patients' and health care providers' (eg, psychotherapists) perspectives.

Another question in the field is, whether GPS features cannot only serve to determine depression from a between-person perspective but also to inform researchers and clinicians about changes within persons (eg, is the increase of time spent at home a reliable indicator for rising depression severity?). This

XSL•FO
RenderX

systematic review showed too sparse evidence and too high heterogeneity in applied analyses to run a meta-analysis on the studies conducted so far. Although this study cannot provide answers regarding the robustness across studies and hence evidence for the significance of within-person correlations, in particular, the included study by Zhang et al [53] on the longitudinal relationship of GPS mobility and activity features and depression found higher within-person correlations than between person correlations, underlining the potential benefit of within-person features in clinical applications and the importance of further studies on within-person correlations.

In line with previous reviews on the methods used in smart sensing studies [26,27,31], this study critically highlights gaps in the adherence to international reporting guidelines, potential publication bias, and high heterogeneity in assessment and analysis methodology. Guidelines like STROBE [38] for observational studies and other study design–specific guidelines offer a strong starting point to increase the study and especially reporting quality in the field. However, research on smart sensing and objective sensor data collection comes with its own challenges like (1) high heterogeneity in hardware (eg, used devices) and GPS sampling (eg, many different manufacturers of sensors and variety in the precision of sensing and data quality), (2) a plethora of data preprocessing decisions (eg, sampling rate, outliner detection or feature selection and definition), (3) missing data handling in large fine-grained datasets, and (4) analyses methodology (eg, how to deal with dependencies resulting from multiple measurements from the same individual [24]) [31]. Preregistration and open-access scripts for data preparation and analyses and extensions to existing international guidelines (eg, standards for feature calculation, reporting guidelines, and missing data handling) are highly needed to move towards a standardized and reliable research field.

The current lack of standardization and heterogeneity between studies in assessment frameworks, samples (eg, age and range), sensor sampling rates, devices, feature calculation, methods of handling missing data or accounting for dependencies in the data structure should also be considered when interpreting these findings. Due to the already low number of studies, we were unable to conduct meaningful sensitivity analysis on more homogenous groups of studies (eg, based on data quality, and devices) or meta-regression models to control for other variables (eg, for systematic mobility and activity differences across age groups). As the number of studies and their quality are likely to increase over time, analyses on homogenous groups and meta-regression analyses might become feasible in future. Nevertheless, we want to point out that the heterogeneity for some features (eg, distance) was very low, underlining the robustness of the pooled correlation estimates for these features. Moreover, the weighted pooling in the meta-analysis gave stronger influence on larger studies with more precise estimations of correlations, counteracting the influence of small studies to some extent [41].

Besides analyses on GPS mobility and activity features in more homogenous study groups, other GPS features should be focused on in the future. For instance, GPS sensor data can be used to derive environmental features such as green space, blue space, air pollution, or even regional data like social deprivation scores (eg, zip code based). First studies using smartphone-based GPS collection to infer depression based on such environmental GPS features are promising, and it needs to be investigated to what extent such features show meta-analytically robust correlations [72-74]. Analogously, meta-analytical evidence for other sensors and features is lacking (eg, smartphone screen usage, app usage, and communication and sociability features) and of high importance to inform researchers and practitioners on which objective markers might be suited for depression diagnosis, early-warning, just-in-time interventions, and other clinical applications [21,30-32,65,69,75]. Furthermore, extensions to other prevalent mental disorders (eg, anxiety) and conditions of leading disease burden (eg, pain) are highly needed to investigate the potential of unobtrusive and objective smart sensing data in health care.

Conclusions

This systematic review and meta-analysis provides evidence for small to medium between-person correlations of GPS mobility and activity features collected through smartphones and depression. In the future, GPS mobility and activity features such as distance may become an important augmentation in the assessment of depression and clinical applications (eg, decision support systems). However, replications in confirmatory studies and improvements in the study quality and standards in the field are needed to draw robust conclusions. Besides, more research on the within-person correlations is necessary to determine the potential of GPS features in not only between-person applications but also from a longitudinal perspective. To fully exploit the potential of smart sensing, further research on associations of depression and other mental disorders with GPS mobility and activity features as well as environmental GPS features (eg, green space) and other sensor modalities (eg, smartphone usage features), and how smart sensing features can be integrated in existing information systems and complex prediction models alongside patient-reported outcomes, clinician ratings, ecological momentary assessments, and medical record data, is of high importance.

Acknowledgments

We would like to thank all the student assistants involved in the project. Generative artificial intelligence tools (eg, ChatGPT) were not used in any portion of the manuscript writing. This study was self-funded by the authors. YT is supported by the initial phase of the German Center for Mental Health (*Deutsches Zentrum für Psychische Gesundheit* [DZPG]; grant 01EE2303A).

XSL•FO
RenderX

Data Availability

The datasets generated during and/or analyzed during this study and analysis codes are available in the Open Science Framework project (Multimedia Appendix 4)

Authors' Contributions

YT contributed to conceptualization, handled methodology, visualization, formal analysis and project administration and wrote original draft. YT, PP, and JK performed data curation. HB handled funding acquisition. YT, PP, and JK performed investigation. YT, PP, JK, and HB managed resources. YT and JK contributed to software. YT and HB performed supervision. YT and JK managed validation. YT, PP, JK, and HB contributed to writing-review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 21 KB - jmir_v26i1e51875_app1.docx]

Multimedia Appendix 2
Search terms.
[DOCX File , 14 KB - jmir_v26i1e51875_app2.docx]

Multimedia Appendix 3
Analysis software.
[DOCX File , 14 KB - jmir_v26i1e51875_app3.docx]

Multimedia Appendix 4
Dataset and code.
[DOCX File , 12 KB - jmir_v26i1e51875_app4.docx]

Multimedia Appendix 5
Forest plots.
[DOCX File , 708 KB - jmir_v26i1e51875_app5.docx]

Multimedia Appendix 6
Funnel plots.
[DOCX File , 150 KB - jmir_v26i1e51875_app6.docx]

References

1. Depression and Other Common Mental Disorders: Global Health Estimates. Geneva, Switzerland: World Health Organization; 2017.
2. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015 Apr;72(4):334-341 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
3. GBD 2017 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
4. Greenberg PE, Fournier AA, Sisitsky T, Pike C, Kessler R. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry 2015;76(2):155-162 [FREE Full text] [doi: 10.4088/JCP.14m09298] [Medline: 25742202]
5. Greenberg PE, Fournier AA, Sisitsky T, Simes M, Berman R, Koenigsberg S, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics 2021;39(6):653-665 [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
6. Trautmann S, Rehm J, Wittchen HU. The economic costs of mental disorders: do our societies react appropriately to the burden of mental disorders? EMBO Rep 2016;17(9):1245-1249 [FREE Full text] [doi: 10.15252/embr.201642951] [Medline: 27491723]

XSL•FO
RenderX

7. Saarni SI, Suvisaari J, Sintonen H, Pirkola S, Koskinen S, Aromaa A, et al. Impact of psychiatric disorders on health-related quality of life: general population survey. Br J Psychiatry 2007;190:326-332. [doi: 10.1192/bjp.bp.106.025106] [Medline: 17401039]
8. Kramer T, Als L, Garralda ME. Challenges to primary care in diagnosing and managing depression in children and young people. BMJ 2015;350:h2512 [FREE Full text] [doi: 10.1136/bmj.h2512] [Medline: 25975867]
9. Wurcel V, Cicchetti A, Garrison L, Kip M, Koffijberg H, Kolbe A, et al. The value of diagnostic information in personalised healthcare: a comprehensive concept to facilitate bringing this technology into healthcare systems. Public Health Genomics 2019;22(1-2):8-15 [FREE Full text] [doi: 10.1159/000501832] [Medline: 31330522]
10. Kroenke K. Depression screening and management in primary care. Fam Pract 2018;35(1):1-3. [doi: 10.1093/fampra/cmx129] [Medline: 29267903]
11. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. Washington, DC: American Psychiatric Association; 2013.
12. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54(5):573-583. [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
13. Fried EI, Flake J, Robinaugh D. Revisiting the theoretical and methodological foundations of depression measurement. Nat Rev Psychol 2022;1(6):358-368 [FREE Full text] [doi: 10.1038/s44159-022-00050-2] [Medline: 38107751]
14. Levis B, Sun Y, He C, Wu Y, Krishnan A, Bhandari P, Depression Screening Data (DEPRESSD) PHQ Collaboration, et al. Accuracy of the PHQ-2 alone and in combination with the PHQ-9 for screening to detect major depression: systematic review and meta-analysis. JAMA 2020;323(22):2290-2300 [FREE Full text] [doi: 10.1001/jama.2020.6504] [Medline: 32515813]
15. Jacobucci R, Grimm KJ. Machine learning and psychological research: the unexplored effect of measurement. Perspect Psychol Sci 2020;15(3):809-816. [doi: 10.1177/1745691620902467] [Medline: 32348703]
16. McNamara ME, Zisser M, Beevers C, Shumake J. Not just "big" data: importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. Behav Res Ther 2022;153:104086. [doi: 10.1016/j.brat.2022.104086] [Medline: 35462242]
17. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]
18. Onnela JP, Rauch S. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016;41(7):1691-1696 [FREE Full text] [doi: 10.1038/npp.2016.7] [Medline: 26818126]
19. Garatva P, Terhorst Y, Messner EM, Karlen W, Pryss R, Baumeister H. Smart sensors for health research and improvement. In: Montag C, Baumeister H, editors. Digital Phenotyping and Mobile Sensing. Berlin, Germany: Springer; 2023:395-411.
20. Terhorst Y, Knauer J, Baumeister H. Smart sensing enhanced diagnostic expert systems. In: Montag C, Baumeister H, editors. Digital Phenotyping and Mobile Sensing. Berlin, Germany: Springer; 2023:413-425.
21. Mohr DC, Zhang M, Schueller S. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
22. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023;6(1):84 [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
23. Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR Mhealth Uhealth 2021;9(7):e26540 [FREE Full text] [doi: 10.2196/26540] [Medline: 34255713]
24. Moshe I, Terhorst Y, Opoku Asare K, Sander L, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247 [FREE Full text] [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
25. Terhorst Y, Messner EM, Opoku Asare K, Montag C, Kannen C, Baumeister H. Which smartphone-based sensing features matter in depression prediction? Results from an observation study. Department of Clinical Psychology and Psychotherapy, Ulm University 2024 [FREE Full text]
26. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
27. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR Mhealth Uhealth 2018;6(8):e165 [FREE Full text] [doi: 10.2196/mhealth.9691] [Medline: 30104184]
28. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. Harv Rev Psychiatry 2020;28(5):296-304. [doi: 10.1097/HRP.0000000000000268] [Medline: 32796192]
29. Dlima SD, Shevade S, Menezes S, Ganju A. Digital phenotyping in health using machine learning approaches: scoping review. JMIR Bioinform Biotechnol 2022;3(1):e39618 [FREE Full text] [doi: 10.2196/39618] [Medline: 38935947]
30. Nouman M, Khoo SY, Mahmud MAP, Kouzani AZ. Recent advances in contactless sensing technologies for mental health monitoring. IEEE Internet Things J 2022;9(1):274-297 [FREE Full text] [doi: 10.1109/jiot.2021.3097801]

XSL•FO
RenderX

31. de Angel V, Lewis S, White K, Oetzmann C, Leightley D, Oprea E, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022;5(1):3 [FREE Full text] [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
32. Zarate D, Stavropoulos V, Ball M, de Sena Collier G, Jacobson N. Correction: exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. BMC Psychiatry 2022;22(1):530 [FREE Full text] [doi: 10.1186/s12888-022-04153-1] [Medline: 35932004]
33. Peterson B, Gonzalez D, Perez-Haddock Y, Frias J, Tourgeman I. PO110 / #755 the utility of passive and portable sensor data for monitoring the symptomatology of depression and anxiety. Neuromodulation: Technology at the Neural Interface 2022;25(7):S247 [FREE Full text] [doi: 10.1016/j.neurom.2022.08.283]
34. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537 [FREE Full text] [doi: 10.7717/peerj.2537] [Medline: 28344895]
35. Saeb S, Zhang M, Karr C, Schueller S, Corden M, Kording K, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
36. Wang R, Wang W, daSilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. 2018 Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; 2018 March 26; New York, NY, United States p. 1-26 URL: https://doi.org/10.1145/3191775 [doi: 10.1145/3191775]
37. Currey D, Torous J. Digital phenotyping correlations in larger mental health samples: analysis and replication. BJPsych Open 2022 Jun 03;8(4):e106 [FREE Full text] [doi: 10.1192/bjo.2022.507] [Medline: 35657687]
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
39. Maxwell SE, Lau MY, Howard GS. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? Am Psychol 2015;70(6):487-498. [doi: 10.1037/a0039400] [Medline: 26348332]
40. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
41. Harrer M, Cuijpers P, Furukawa T, Ebert D. Doing Meta-Analysis With R: A Hands-On Guide. 1st ed. Boca Raton, FL and London: Chapman & Hall/CRC Press; 2021.
42. Cochrane handbook for systematic reviews of interventions. Cochrane Training. URL: https://training.cochrane.org/handbook/current [accessed 2023-07-16]
43. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: i is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
44. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science 2015;349(6251):aac4716. [doi: 10.1126/science.aac4716] [Medline: 26315443]
45. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
46. Mathur MB, VanderWeele T. Sensitivity analysis for publication bias in meta-analyses. J R Stat Soc Ser C Appl Stat 2020;69(5):1091-1119 [FREE Full text] [doi: 10.1111/rssc.12440] [Medline: 33132447]
47. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
48. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019;22(4):153-160 [FREE Full text] [doi: 10.1136/ebmental-2019-300117] [Medline: 31563865]
49. MacLeod L, Suruliraj B, Gall D, Bessenyei K, Hamm S, Romkey I, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR Mhealth Uhealth 2021;9(10):e20638 [FREE Full text] [doi: 10.2196/20638] [Medline: 34698650]
50. Lu J, Shang C, Yue C, Morillo R, Ware S, Kamath J, et al. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. 2018 Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; March 26, 2018; New York, NY p. 1-21 URL: https://doi.org/10.1145/3191753 [doi: 10.1145/3191753]
51. Nickels S, Edwards MD, Poole SF, Winter D, Gronsbell J, Rozenkrants B, et al. Toward a mobile platform for real-world digital measurement of depression: user-centered design, data quality, and behavioral and clinical modeling. JMIR Ment Health 2021;8(8):e27589 [FREE Full text] [doi: 10.2196/27589] [Medline: 34383685]
52. Tung JY, Rose R, Gammada E, Lam I, Roy E, Black S, et al. Measuring life space in older adults with mild-to-moderate alzheimer's disease using mobile phone GPS. Gerontology 2014;60(2):154-162. [doi: 10.1159/000355669] [Medline: 24356464]

XSL•FO
RenderX

53. Zhang Y, Folarin AA, Sun S, Cummins N, Vairavan S, Bendayan R, et al. Longitudinal relationships between depressive symptom severity and phone-measured mobility: dynamic structural equation modeling study. JMIR Ment Health 2022;9(3):e34898 [FREE Full text] [doi: 10.2196/34898] [Medline: 35275087]
54. Saeb S, Zhang M, Kwasny M, Karr C, Kording K, Mohr D. The relationship between clinical, momentary, and sensor-based assessment of depression. 2015 Presented at: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth); May 20-23, 2015; Istanbul, Turkey p. 103-111 URL: https://europepmc.org/abstract/MED/26640739 [doi: 10.4108/icst.pervasivehealth.2015.259034]
55. Saeb S, Lattie EG, Kording KP, Mohr DC. Mobile phone detection of semantic location and its relationship to depression and anxiety. JMIR Mhealth Uhealth 2017;5(8):e112 [FREE Full text] [doi: 10.2196/mhealth.7297] [Medline: 28798010]
56. Boukhechba M, Daros AR, Fua K, Chow PI, Teachman BA, Barnes LE. DemonicSalmon: monitoring mental health and social interactions of college students using smartphones. Smart Health 2018;9-10:192-203 [FREE Full text] [doi: 10.1016/j.smhl.2018.07.005]
57. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. 2015 Presented at: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2015 September 07; Osaka, Japan p. 1293-1304 URL: https://doi.org/10.1145/2750858.2805845
58. DeMasi O, Recht B. A step towards quantifying when an algorithm can and cannot predict an individual's wellbeing. 2017 Presented at: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers; 2017 September 11; Hawaii, Maui p. 763-771 URL: https://doi.org/10.1145/3123024.3125609
59. Di Matteo D. Inference of anxiety and depression from smartphone-collected data. Diss Abstr Int Sect B Sci Eng 2021:1-184 [FREE Full text]
60. Giannouli E, Fillekes MP, Mellone S, Weibel R, Bock O, Zijlstra W. Predictors of real-life mobility in community-dwelling older adults: an exploration based on a comprehensive framework for analyzing mobility. Eur Rev Aging Phys Act 2019;16:19 [FREE Full text] [doi: 10.1186/s11556-019-0225-2] [Medline: 31700551]
61. Stamatis CA, Liu T, Meyerhoff J, Meng Y, Cho YM, Karr CJ, et al. Specific associations of passively sensed smartphone data with future symptoms of avoidance, fear, and physiological distress in social anxiety. Internet Interv 2023;34:100683 [FREE Full text] [doi: 10.1016/j.invent.2023.100683] [Medline: 37867614]
62. Farhan AA, Yue C, Morillo R, Ware S, Lu J, Bi J, et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. 2016 Presented at: 2016 IEEE Wireless Health (WH); October 25-27, 2016; Bethesda, MD p. 1-8 URL: https://nlab.engr.uconn.edu/papers/Farhan-behavior-2016.pdf [doi: 10.1109/WH.2016.7764553]
63. Meyerhoff J, Liu T, Kording KP, Ungar LH, Kaiser SM, Karr CJ, et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. J Med Internet Res 2021;23(9):e22844 [FREE Full text] [doi: 10.2196/22844] [Medline: 34477562]
64. Stamatis CA, Meyerhoff J, Liu T, Sherman G, Wang H, Liu T, et al. Prospective associations of text-message-based sentiment with symptoms of depression, generalized anxiety, and social anxiety. Depress Anxiety 2022;39(12):794-804 [FREE Full text] [doi: 10.1002/da.23286] [Medline: 36281621]
65. Steele R, Hillsgrove T, Khoshavi N, Jaimes L. A survey of cyber-physical system implementations of real-time personalized interventions. J Ambient Intell Human Comput 2021;13(5):2325-2342 [FREE Full text] [doi: 10.1007/s12652-021-03263-0]
66. Schuster R, Kaiser T, Terhorst Y, Messner E, Strohmeier LM, Laireiter AR. Sample size, sample size planning, and the impact of study context: systematic review and recommendations by the example of psychological depression treatment. Psychol Med 2021;51(6):902-908 [FREE Full text] [doi: 10.1017/S003329172100129X] [Medline: 33879275]
67. Barnett I, Torous J, Reeder H, Baker J, Onnela JP. Determining sample size and length of follow-up for smartphone-based digital phenotyping studies. J Am Med Inform Assoc 2020;27(12):1844-1849 [FREE Full text] [doi: 10.1093/jamia/ocaa201] [Medline: 33043370]
68. Montag C, Baumeister H, editors. Digital Phenotyping and Mobile Sensing. 2nd ed. Cham, Switzerland: Springer International Publishing; 2023.
69. Terhorst Y, Weilbacher N, Suda C, Simon L, Messner EM, Sander LB, et al. Acceptance of smart sensing: a barrier to implementation-results from a randomized controlled trial. Front Digit Health 2023;5:1075266 [FREE Full text] [doi: 10.3389/fdgth.2023.1075266] [Medline: 37519894]
70. Nicholas J, Shilton K, Schueller SM, Gray EL, Kwasny MJ, Mohr DC. The role of data type and recipient in individuals' perspectives on sharing passively collected smartphone data for mental health: cross-sectional questionnaire study. JMIR Mhealth Uhealth 2019;7(4):e12578 [FREE Full text] [doi: 10.2196/12578] [Medline: 30950799]
71. Terhorst Y, Sander LB, Ebert DD, Baumeister H. Optimizing the predictive power of depression screenings using machine learning. Digit Health 2023;9:20552076231194939 [FREE Full text] [doi: 10.1177/20552076231194939] [Medline: 37654715]
72. Roberts H, Helbich M. Multiple environmental exposures along daily mobility paths and depressive symptoms: a smartphone-based tracking study. Environ Int 2021;156:106635 [FREE Full text] [doi: 10.1016/j.envint.2021.106635] [Medline: 34030073]

XSL•FO
RenderX

73. Mennis J, Mason M, Ambrus A. Urban greenspace is associated with reduced psychological stress among adolescents: a geographic ecological momentary assessment (GEMA) analysis of activity space. Landsc Urban Plan 2018;174:1-9 [FREE Full text] [doi: 10.1016/j.landurbplan.2018.02.008] [Medline: 29576668]
74. Liu Z, Chen X, Cui H, Ma Y, Gao N, Li X, et al. Green space exposure on depression and anxiety outcomes: a meta-analysis. Environ Res 2023;231(Pt 3):116303. [doi: 10.1016/j.envres.2023.116303] [Medline: 37268208]
75. Rottstädt F, Becker E, Wilz G, Croy I, Baumeister H, Terhorst Y. Enhancing the acceptance of smart sensing in psychotherapy patients: findings from a randomized controlled trial. Front Digit Health 2024;6:1335776 [FREE Full text] [doi: 10.3389/fdgth.2024.1335776] [Medline: 38698889]

Abbreviations

AUC: area under the curve
PHQ: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROC: receiver operating curve
STROBE: Strengthening the Reporting of Observational studies in Epidemiology

Edited by T de Azevedo Cardoso; submitted 11.09.23; peer-reviewed by L Yi, K Aguiar; comments to author 18.06.24; revised version received 02.07.24; accepted 26.07.24; published 01.11.24.

Please cite as:
Terhorst Y, Knauer J, Philippi P, Baumeister H
The Relation Between Passively Collected GPS Mobility Metrics and Depressive Symptoms: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e51875
URL: https://www.jmir.org/2024/1/e51875
doi:10.2196/51875
PMID:

©Yannik Terhorst, Johannes Knauer, Paula Philippi, Harald Baumeister. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis

William J Waldock[1], MBA, MBBCHIR; Joe Zhang[1], MBBS, PhD; Ahmad Guni[1], MBBS; Ahmad Nabeel[2], MBBS; Ara Darzi[1], MBBS; Hutan Ashrafian[2], BSc, MBBS, MBA, PhD

[1]Imperial College London, London, United Kingdom

[2]Institute of Global Health Innovation, Imperial College London, London, United Kingdom

Corresponding Author:
Hutan Ashrafian, BSc, MBBS, MBA, PhD
Institute of Global Health Innovation
Imperial College London
10th Floor, Queen Elizabeth Queen Mother Building, Praed Street
London
United Kingdom
Phone: 44 07799871597
Email: h.ashrafian@imperial.ac.uk

Abstract

Background: Large language models (LLMs) have dominated public interest due to their apparent capability to accurately replicate learned knowledge in narrative text. However, there is a lack of clarity about the accuracy and capability standards of LLMs in health care examinations.

Objective: We conducted a systematic review of LLM accuracy, as tested under health care examination conditions, as compared to known human performance standards.

Methods: We quantified the accuracy of LLMs in responding to health care examination questions and evaluated the consistency and quality of study reporting. The search included all papers up until September 10, 2023, with all LLMs published in English journals that report clear LLM accuracy standards. The exclusion criteria were as follows: the assessment was not a health care exam, there was no LLM, there was no evaluation of comparable success accuracy, and the literature was not original research.The literature search included the following Medical Subject Headings (MeSH) terms used in all possible combinations: “artificial intelligence,” “ChatGPT,” “GPT,” “LLM,” “large language model,” “machine learning,” “neural network,” “Generative Pre-trained Transformer,” “Generative Transformer,” “Generative Language Model,” “Generative Model,” “medical exam,” “healthcare exam,” and “clinical exam.” Sensitivity, accuracy, and precision data were extracted, including relevant CIs.

Results: The search identified 1673 relevant citations. After removing duplicate results, 1268 (75.8%) papers were screened for titles and abstracts, and 32 (2.5%) studies were included for full-text review. Our meta-analysis suggested that LLMs are able to perform with an overall medical examination accuracy of 0.61 (CI 0.58-0.64) and a United States Medical Licensing Examination (USMLE) accuracy of 0.51 (CI 0.46-0.56), while Chat Generative Pretrained Transformer (ChatGPT) can perform with an overall medical examination accuracy of 0.64 (CI 0.6-0.67).

Conclusions: LLMs offer promise to remediate health care demand and staffing challenges by providing accurate and efficient context-specific information to critical decision makers. For policy and deployment decisions about LLMs to advance health care, we proposed a new framework called RUBRICC (Regulatory, Usability, Bias, Reliability [Evidence and Safety], Interoperability, Cost, and Codesign–Patient and Public Involvement and Engagement [PPIE]). This presents a valuable opportunity to direct the clinical commissioning of new LLM capabilities into health services, while respecting patient safety considerations.

Trial Registration: OSF Registries osf.io/xqzkw; https://osf.io/xqzkw

(J Med Internet Res 2024;26:e56532) doi:10.2196/56532

XSL•FO
RenderX

KEYWORDS

large language model; LLM; artificial intelligence; AI; health care exam; narrative medical response; health care examination; clinical commissioning; health services; safety

Introduction

The advent of large language models (LLMs), such as Chat Generative Pretrained Transformer (ChatGPT; OpenAI), has generated extraordinary interest worldwide and transformed the landscape of artificial intelligence (AI). This foremost positioning of transformer models in the public and academic consciousness has been achieved by the remarkable ability of generative artificial intelligence (genAI) models to create new content with human-like semantics and syntax, alongside the capability to accurately replicate learned knowledge in narrative text. Numerous applications in medical research [1], medical education [1], clinical communication or consultation [2], and even diagnosis and risk prediction tasks [2] have been demonstrated to date. There is great positive potential for genAI across all of these pathways and great promise to relieve the increasing pressures and shortage of clinical expertise in health care systems worldwide [2].

The ability of genAI to answer medical examination questions is of particular interest. First, such examinations serve as the gateway for professional qualification. Written examination questions replicate complex clinical scenarios in narrative form and may include the possibility of multiple reasonable differential diagnoses (multiple choice) or require ranking of medically appropriate responses (single-best answer) according to not just clinical knowledge but also contextual decision-making and medical ethics. For decades, this type of examination has been the ultimate test of human clinical judgment and depth of knowledge. The performance of LLMs in this context has far-reaching implications for how medical education is delivered. Second, these expert-developed and expert-validated question-answer pairs are a coherent substitute for real-world training data written in narrative form and may serve to tune genAI models with a clinical consultation, communication, or diagnostic function. This is exemplified by Google's use of medical examination questions to train and test Medical Patient Language Model 2 (Med-PaLM 2) [3]. Finally, these same validated questions are a ready-made benchmark for assessing LLM capabilities in future clinical or medical education–related tasks.

However, the use of LLMs is not without risk. They have a propensity to "hallucinate" false information and produce potentially dangerous inaccuracies [4]. In addition, LLMs are created through a process of pretraining on vast existing text corpuses to enable a general understanding of syntax and semantics. Although models may undergo fine-tuning for particular tasks or domains, this process does not modify the underlying "learned" knowledge but adjusts weights to adapt the model's outputs for a required context. As such, the underlying embedding of our current societal state means that models will also encode societal biases, which will certainly include biases seen in health care provision and outcomes [5]. An understanding of how these problems manifest in real-world tasks is key to developing mitigations and to establish risks and benefits of the use of LLMs in different medical areas.

We conducted a systematic review of LLM accuracy, as tested under health care examination conditions, as compared to known human performance standards. We assessed the reporting quality and risk of bias within existing studies and synthesized a discussion of pitfalls and performance concerns, as reported by study investigators. We discussed how the observed LLM performance impacts medical education and genAI-enabled clinical consultation and recommended a framework for the conduct of future research in this area. In response to this rapidly progressing field, we aimed to establish a baseline performance and quality standard for the current generation of LLMs in narrative medical response tasks.

Methods

Study Design

The systematic review was conducted according to a registered protocol and was reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [6]. The protocol was registered with the Open Science Framework (OSF) [7], under the title "How Accurate are Artificial Intelligence LLMs When Applied to Healthcare Exams and Certificates?", with the secondary research questions "What is the performance of LLM in comparison to required examination standards for humans?" and "What are the primary discovered weaknesses of LLM in addressing narrative health care examination scenarios that may be pertinent to real-world performance in clinical scenarios?"

Eligibility Criteria

The inclusion criteria were all papers up until September 10, 2023, published in English language journals that describe the use of AI solutions in health care examinations and certificates. As reflected in the Medical Subject Headings (MeSH) terms used, the authors screened the manuscripts for "artificial intelligence," which could be described in the following possible ways: "ChatGPT," "GPT," "LLM," "large language model," "machine learning," "neural network," "Generative Pretrained Transformer," "Generative Transformer," "Generative Language Model," or "Generative Model." The exclusion criteria were as follows: the assessment was not a health care examination, there was no LLM, there was no evaluation of comparable success accuracy, and the literature was not original research (ie, commentary, editorials, reviews). We assessed LLMs, first, as applied to health care examinations and, by extension, as applied to clinical problems, including those encountered by individual patients and clinicians, and the likely impact on future medical education. We assessed the outcome of the accuracy of examination response performance and an intervention of the use of LLMs to answer narrative health care examination questions. The additional variable(s)/covariate(s) to consider were the name and country of medical examination; the "pass

XSL•FO
RenderX

mark" and other score boundaries for each medical examination; the average and intervals of human performance for each medical examination that included benchmarks; the identity of LLMs; LLM characteristics, including parameter size; and any fine-tuning for the LLMs prior to testing.

Information Sources

The search included all papers up until September 10, 2023, at which point a preliminary search was conducted and piloting of the study selection process was commenced using MEDLINE/PubMed, CINAHL, ClinicalTrials.gov, Embase, and Google Scholar.

Search Strategy

The literature search included the following MeSH terms used in all possible combinations: "artificial intelligence," "ChatGPT," "GPT," "LLM," "large language model," "machine learning", "neural network", "Generative Pre-trained Transformer", "Generative Transformer," "Generative Language Model," "Generative Model," "medical exam," "healthcare exam," and "clinical exam." Two authors (WJW and AG) independently identified relevant studies, and any discrepancies were resolved by consensus with the help of a third author (HA)

Selection Process

Screening reliability and duplicate removal were maintained by 2 independent screeners reviewing abstracts (WJW and AG), with divergent screener decisions reconciled by a third master screener (HA). Abstracts were downloaded and screened in Covidence software [8] using .rsi and .csv files. Two independent authors (WJW and AG) performed full-text manuscript screening following abstract screening, with discrepancies resolved by consultation with the lead author (HA).

Data Collection, Data Items, and Data Synthesis

Two reviewers (WJW and AG) extracted and synthesized comparative accuracy data from the reviews on Covidence. No automation tools were used. The 2 authors independently extracted data from relevant studies, and any discrepancies were resolved by consensus with the help of a third author (HA) Sensitivity, accuracy, and precision data were extracted, including relevant CIs. The meta-analysis pooling of aggregate data used the random-effects inverse-variance model with DerSimonian-Laird estimate of tau^2. The software used to conduct the meta-analysis was Stata Statistical Software Release 15 (StataCorp).

Risk-of-Bias Assessment and Reporting Bias Assessment

The QUADAS-2 tool [9] was used for the systematic evaluation and assessment of the risk of bias and concerns regarding answer accuracy for clinical examination questions. The evaluation enabled adjudication of the applicability and bias concerns regarding reference standards and training data selection. Two independent authors performed the risk-of-bias assessment, with discrepancies resolved by consultation with the lead author (HA).Results

Results

Study Screening

Based on PRISMA guidelines, the search identified 1673 relevant citations. After removing duplicate results, 1268 (75.8%) papers were screened for titles and abstracts, and 32 (2.5%) [3,10-40] studies were included for full-text review (see Figure 1 and Table S1 in Multimedia Appendix 1).

The LLMs represented in this systematic literature review were Flan-PaLM 2 [3], Generative Pretrained Transformer (GPT)-Neo [10], ChatGPT [11-35], Google Bard [13], Bing Chat [13], PubMedGPT (Stanford University) [36], BioLinkBERT [37] (BERT stands for Bidirectional Encoder Representations from Transformers), PubMedBERT [38], Galactica [39], and DRAGON (Deep Bidirectional Language-Knowledge Graph Pretraining) [40]. All these models are commercial, except BioLinkBERT, GPT-Neo, and DRAGON. The majority of LLMs used in medical examination tasks were pretrained, closed source models, developed and released by commercial organizations, such as ChatGPT. There was no prompt engineering described by the majority of the studies [11,13,15-35] when using ChatGPT, but Kung et al [12] and Gilson et al [14] specifically introduced prompt engineering to mitigate concerns about model "hallucinations" [41]. Stanford University's PubMedGPT 2.7B [36] is an LLM trained on PubMed abstracts and Pile. Flan-PaLM 2 [3], PubMedGPT [36], DRAGON [40], BioLinkBERT [37], Galactica [39], PubMedBERT [38], and GPT-Neo [10] were all evaluated using the same 12,723 United States Medical Licensing Examination (USMLE) open source question dataset [42]. BioLinkBERT [37] is a self-supervised pretraining bidirectional system that leverages graph structures in PubMed. PubMedBERT [38] is a BERT-style model trained on PubMed, while Galactica [39] is a GPT-style model trained on scientific literature that is 44 times the size of PubMedGPT 2.7B [36].

XSL•FO
RenderX

Figure 1. Study selection based on PRISMA guidelines. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

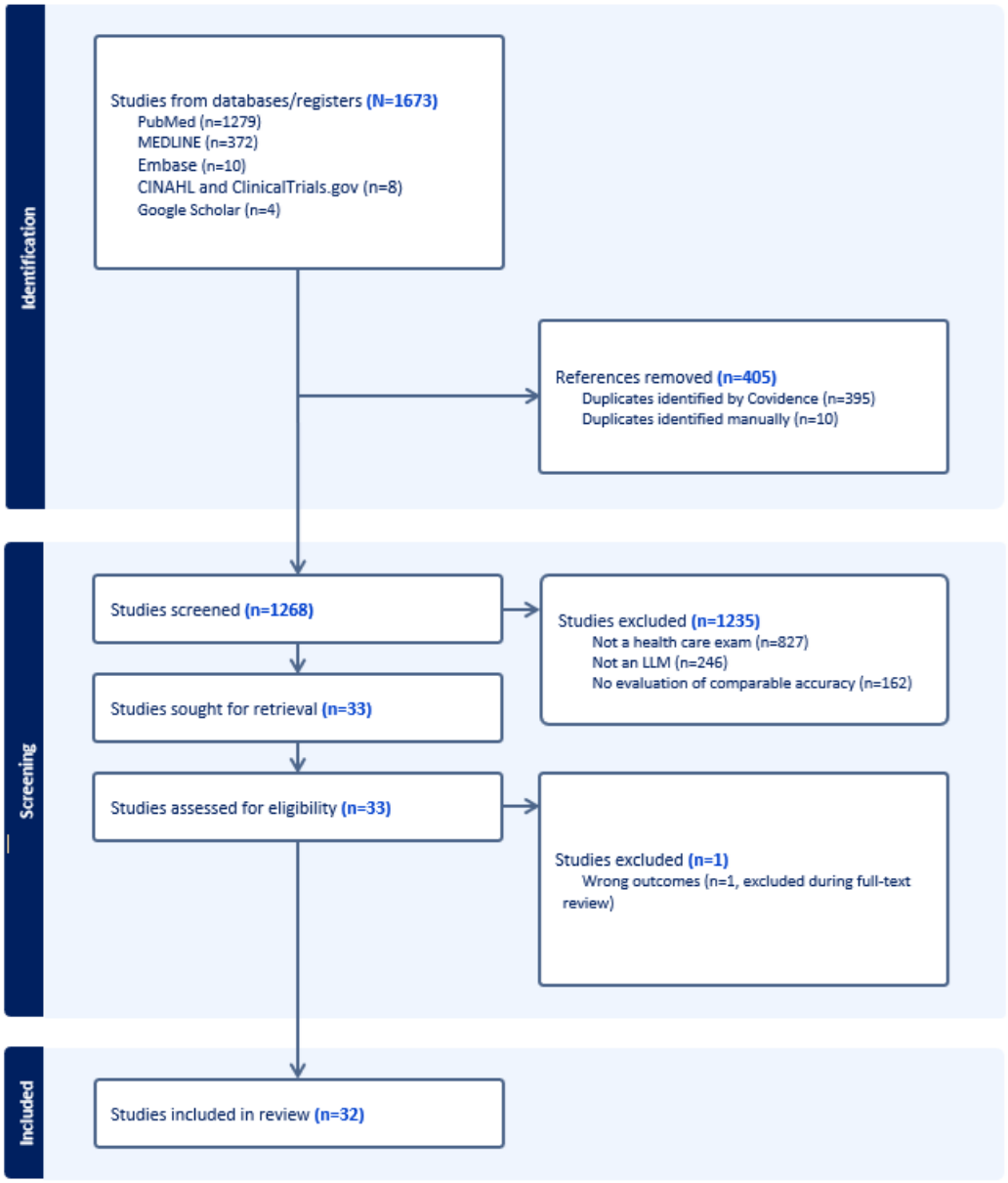

Precision, Sensitivity, and Accuracy

Precision

When assessing the precision of LLMs in all examinations, 2 (6.3%) studies had an overall precision of 0.61 (CI 0.55-0.67) across 189 questions, with a tau^2 heterogeneity of 0.0018 and an I^2 variation attributable to a heterogeneity of 99.6%.

Sensitivity

When assessing the sensitivity of LLMs in all examinations, 2 (6.3%) studies had an overall sensitivity of 1.00 (CI 1.00-1.00) across 189 questions, with a tau^2 heterogeneity of 0.0000 and an I^2 variation attributable to a heterogeneity of 0%.

Accuracy

The overall LLM examination performance, USMLE accuracy, and ChatGPT accuracy were all evaluated by substudy meta-analysis, with question counts moderated for double-counting across multiple substudies. When assessing the accuracy of LLMs in all examinations, 47 substudies had an overall accuracy of 0.61 (CI 0.58-0.64) across 22,347 questions, with a tau^2 heterogeneity of 0.0088 and an I^2 variation attributable to a heterogeneity of 100% (Table 1 and Figure 2).

XSL•FO
RenderX

Table 1. LLM[a] meta-analysis substudies.

Study and substudies	Questions, n	Accuracy
Alessandri Bonetti et al [26]; IRANE (Italian Residency Admission National Exam)	140	0.87
Angel et al [20]		
Bard American Board of Anesthesiology (ABA)	1000	0.46
GPT[b]-3 ABA	1000	0.50
GPT-4 ABA	1000	0.80
Beaulieu-Jones et al [30]		
Data-B	112	0.68
SCORE (Surgical Council on Resident Education)	167	0.71
Bolton et al [36]		
PubMedGPT	12,723	0.50
ChatGPT[c]	1217	0.76
Flores-Cohaila et al [29]; Peruvian National Licensing Medical Examination (PNLME)	180	0.86
Gencer et al [23]; Turkish ChatGPT	105	0.91
Giannos et al [18]; BioMedical Admissions Test (BMAT)	509	0.73
Gilson et al [14]		
ChatGPT A	100	0.44
ChatGPT B	100	0.42
ChatGPT C	87	0.64
ChatGPT D	102	0.58
Gu et al [38]; PubMedBERT[d]	12,723	0.38
Guerra et al [24]; ChatGPT Self-Assessment Neurosurgery (SANS)	643	0.77
Huang et al [21]		
GPT-3 Radiation Oncology in-Training (TXIT)	300	0.62
GPT-4 TXIT	300	0.79
Huang et al [28]; University of Toronto Family Medicine Residency Progress Test (UTFMRPT)	108	0.82
Humar et al [17]; ChatGPT plastic surgery	1129	0.56
Huynh et al [32]; GPT urology	135	0.28
Kufel et al [31]; ChatGPT Polish radiology examination	120	0.52
Kung et al [12]; ChatGPT	376	0.60
Mannam et al [35]; ChatGPT SANS	427	0.67
Morreel et al [16]; ChatGPT Dutch	47	0.50
Oh et al [19]; ChatGPT Korean	280	0.76
Oztermeli et al [22]; GPT-3.5 medical specialty examination (MSE)	1177	0.71
Raimondi et al [13]		
Bard Fellowship of the Royal College of Physicians and Surgeons (Ophthalmology), or FRCOphth, part 1	48	0.63
Bard FRCOphth part 2	43	0.52
Bing Chat FRCOphth part 1	48	0.79
Bing Chat FRCOphth part 2	43	0.83
ChatGPT-3.5 FRCOphth part 1	48	0.55
ChatGPT-3.5 FRCOphth part 2	43	0.50

XSL•FO
RenderX

Study and substudies	Questions, n	Accuracy
LLM chatbot FRCOphth part 1	48	0.66
LLM chatbot FRCOphth part 2	43	0.68
Sharma et al [11]; ChatGPT	119	0.58
Singhal et al [3]; Med-PaLM 2[e]	12,723	0.60
Skalidid et al [34]; ChatGPT cardiology	340	0.59
Strong et al [15]; ChatGPT	28	0.69
Taylor et al [39]; Galactica	12,723	0.44
Venigalla et al [10]; GPT-Neo	12,723	0.33
Wang et al [25]; Chinese National Medical Licensing Examination (CNMLE)	360	0.47
Weng et al [27]; Taiwan Family Medicine Board Exam (TFMBE)	125	0.42
Yasunaga et al [37]; BioLinkBERT	12,723	0.45
Yasunaga et al [40]; DRAGON[f]	12,723	0.48

[a]LLM: large language model.
[b]GPT: Generative Pretrained Transformer.
[c]ChatGPT: Chat Generative Pretrained Transformer.
[d]BERT: Bidirectional Encoder Representations from Transformers.
[e]Med-PaLM 2: Medical Patient Language Model 2.
[f]DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining.

XSL•FO
RenderX

Figure 2. Forest plot of the accuracy of LLM performance on all medical examinations. ABA: American Board of Anesthesiology; BERT: Bidirectional Encoder Representations from Transformers; ChatGPT: Chat Generative Pretrained Transformer; CNMLE: Chinese National Medical Licensing Examination; DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining; FRCOphth: Fellowship of the Royal College of Physicians and Surgeons (Ophthalmology); GPT: Generative Pretrained Transformer; IRANE: Italian Residency Admission National Exam; LLM: large language model; MSE: medical specialty examination; PaLM: Patient Language Model 2; PNLME: Peruvian National Licensing Medical Examination; PRE: Polish radiology examination; SANS: Self-Assessment Neurosurgery; SCORE: Surgical Council on Resident Education; TFMBE: Taiwan Family Medicine Board Exam; TXIT: Radiation Oncology in-Training; UTFMRPT: University of Toronto Family Medicine Residency Progress Test.

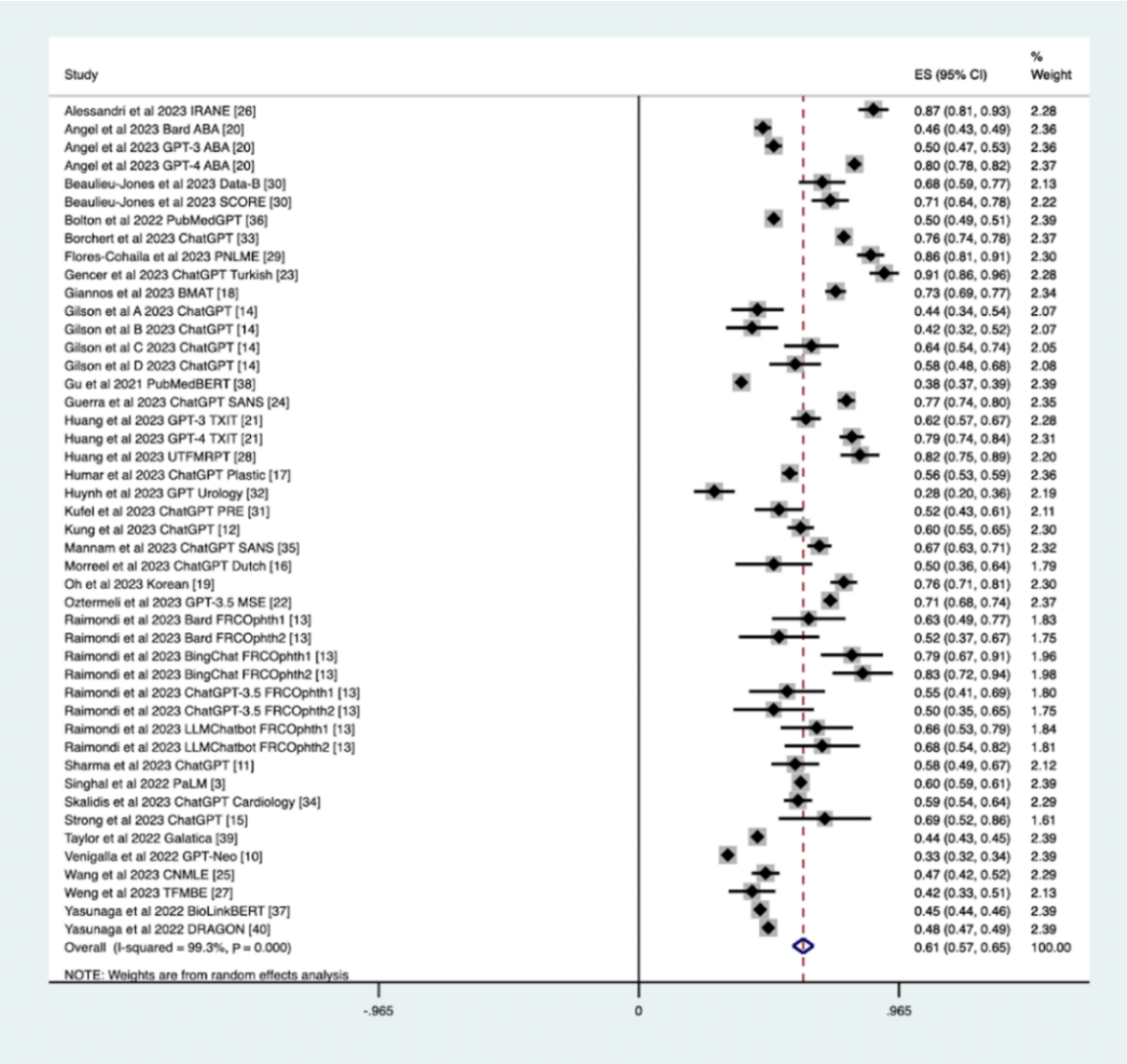

USMLE Accuracy

When assessing the accuracy of LLMs in the USMLE, 14 substudies had an overall accuracy of 0.51 (CI 0.46-0.56) across 13,535 questions, with a tau^2 heterogeneity of 0.0080 and an I^2 variation attributable to a heterogeneity of 100%.

ChatGPT Accuracy

When assessing the accuracy of ChatGPT on medical examinations, 32 substudies had an overall accuracy of 0.64 (CI 0.6-0.67) across 9824 questions, with a tau^2 heterogeneity of 0.0128 and an I^2 variation attributable to a heterogeneity of 100%.

Bias and Narrative Reporting

Among the 32 studies that underwent QUADAS-2 [9,43] risk-of-bias evaluation (Figure 3), only 11 (24.4%) were eligible for meta-analysis. Overall, 10 (31.3%) studies were found to have high bias, 15 (46.9%) studies were found to have some concerns of bias, and 7 (21.9%) studies were found to have low bias. In addition, 3 (9.4%) studies referred to concerns about "hallucinations," but none described the effect nor referred to softer themes, such as empathy. No studies evaluated bias systematically. None of the reviewed literature was systematic reviews, so a TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) adherence to reporting standards analysis [44] was not conducted.

XSL•FO
RenderX

Figure 3. Risk-of-bias. ABA: American Board of Anesthesiology; BERT: Bidirectional Encoder Representations from Transformers; ChatGPT: Chat Generative Pretrained Transformer; CNMLE: Chinese National Medical Licensing Examination; DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining; FRCOphth: Fellowship of the Royal College of Physicians and Surgeons (Ophthalmology); GPT: Generative Pretrained Transformer; IRANE: Italian Residency Admission National Exam; LLM: large language model; MSE: medical specialty examination; PaLM: Patient Language Model 2; PNLME: Peruvian National Licensing Medical Examination; PRE: Polish radiology examination; SANS: Self-Assessment Neurosurgery; SCORE: Surgical Council on Resident Education; TFMBE: Taiwan Family Medicine Board Exam; TFMR: Toronto Family Medicine Residency; TXIT: Radiation Oncology in-Training.

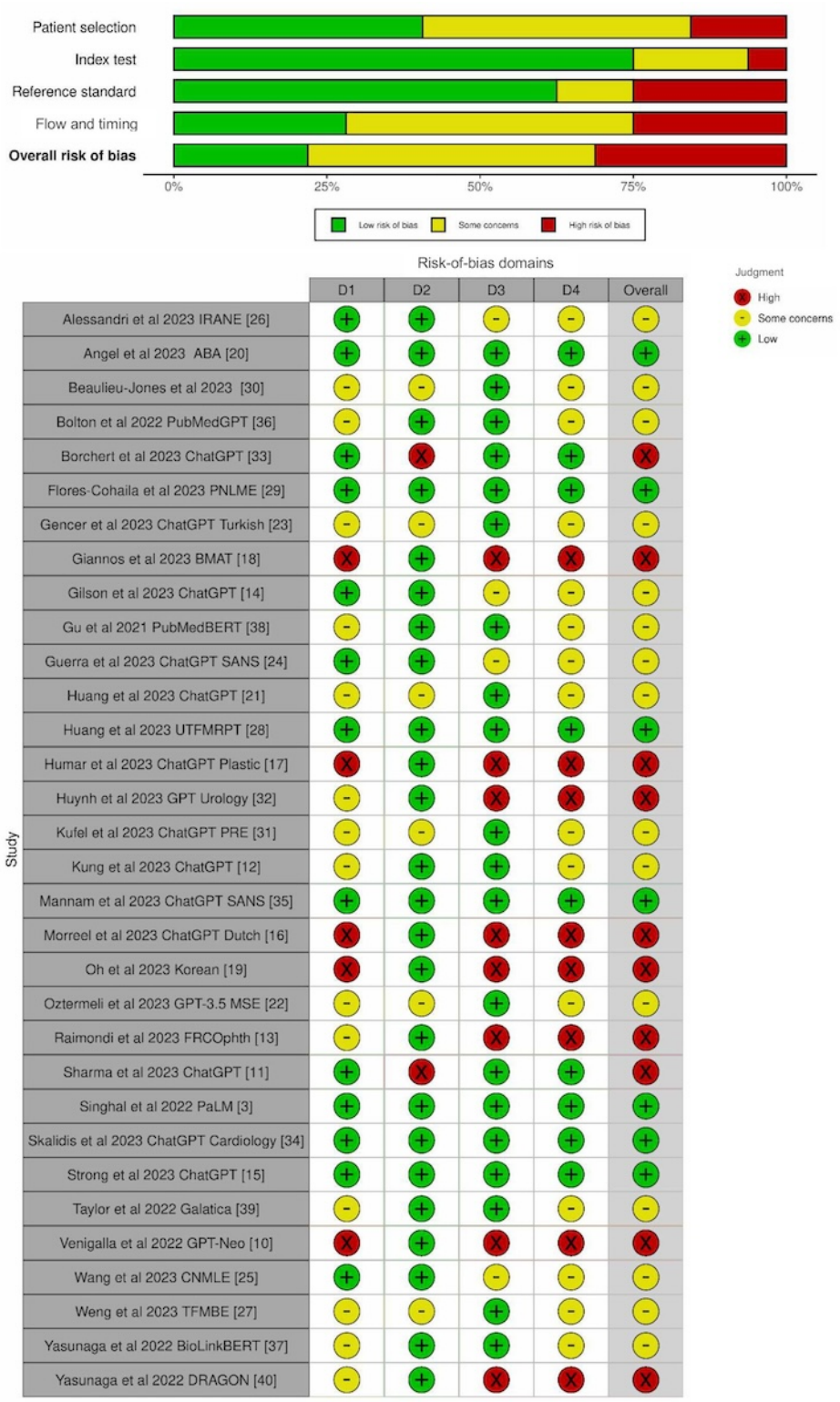

XSL•FO
RenderX

Discussion

Principal Findings

Our meta-analysis suggests that LLMs are able to perform with an overall medical examination accuracy of 0.61 (CI 0.58-0.64) and a USMLE accuracy of 0.51 (CI 0.46-0.56), while ChatGPT can perform with an overall medical examination accuracy of 0.64 (CI 0.6-0.67). We quantified the accuracy of LLMs in responding to health care examination questions and evaluated the consistency and quality of study reporting. The majority of LLMs used in medical examination tasks were pretrained, closed source models, developed and released by commercial organizations, such as ChatGPT. However, we found that minimal research has explored bias, "hallucination," and holistic evaluation of the LLMs themselves. Moreover, neither the risk of bias nor holistic evaluation frameworks exist for LLMs themselves.

There are inherent challenges to integrating LLMs into the education and clinical decision support of human doctors. Use cases for LLMs include grading, detection, prediction, and content generation [45], but the application of these capabilities to the sociocultural elements of medicine are complex. Doctors offer empathetic relationships and formulate clinical reasoning in a more transparent way than current LLMs, raising concerns that the introduction of LLMs will undermine doctor-patient rapport [46] and trust in the ethical compliance of the health care system. LLMs can automate the generation of text content, which offers opportunities to enhance student answer marking and provide responsive learning assistant chat features [45]. However, these features lack transparency, prompting distrust in decision-making [47], and a lack of evidence generation around student engagement [48]. Although these training and infrastructure hurdles must be overcome, there is immense potential for personalized learning experiences with augmented and virtual reality, alongside enhanced curriculum implementation [49].

Medical examinations are not the same as medical practice [50]. The tests that are designed to confirm a human's suitability to practice medicine independently may not be appropriate for an LLM; real-world practice involves greater pathophysiological complexity, diverse holistic care considerations, and important ethical accountability frameworks to ensure empathetic patient-centered health services. Here, we demonstrated LLM capabilities in question-and-answer tasks according to established international benchmarks. Single-best answer questions are designed to simulate clinical decision-making, but there is a lack of relevance of examination questions to real-world tasks [5]. Current models are trained on an unregulated range of both narrow and broad data sets to perform tasks with translational evidence, which currently have unclear significance in clinical practice [5]. LLMs are not yet ready to be a proxy for human education, as questions simplify and isolate scenarios in an imperfect representation of real situations encountered by clinicians. However, the success of LLMs may justify a reconfiguration or even a disruption of medical training. This might involve an initial move toward formative assessments in view of the limitations of summative assessments exposed by the success of LLMs in the USMLE [3]; rather, when offered access to a hitherto untapped wealth of medical information, the role of the doctor may be able to provide judicious medical decisions when presented with intelligent and superintelligent LLM-generated treatment strategies.

Virtual and remote learning opportunities will be enhanced by LLMs [49], but bias, cost, and "hallucination" are the major obstacles to their application in health care. The definition of the threshold for acceptable clinical deployment varies across clinical scenarios and disease states due to the variation in the acceptable tolerance of error. LLMs are developed with parameters that reflect the established sociocultural inequalities in our society and can be perpetuated in LLMs without further intervention. Solutions such as LLM-focused data governance strategies within current and future guidelines and novel approaches, including the use of synthetic data, will likely be needed to ensure those underserved by current data collection pools are not discriminated against in the behavior of the LLMs [51]. With estimates suggesting that US $5 million of graphical processing units (GPUs) [52] are needed at minimum for 1 LLM, their impressive capabilities are unlikely to be ubiquitous across health systems, such as the UK National Health Service (NHS), and may exacerbate inequalities. Finally, there is an inherent danger of "hallucination" with LLMs, undermining the protection of patient data and accurate contributions to live clinical scenarios [53].

Study Limitations

The studies failed to explore the main barriers to LLM implementation in clinical practice, including bias, "hallucinations," usability, cost, and privacy. The extensive variation between studies in the terminology, methodology, outcome measures, and data interpretability could be explained by a lack of consensus on how to conduct and report LLM studies. We have concerns over the reliability of these studies and the small volume of eligible studies for comparison. The lack of consistency in accuracy reporting between studies obstructed evaluation of the relative strengths of each method. There is an inherent challenge in evaluating technology with substantial commercial potential due to producers' understandable reluctance about publishing sensitive details that may enable reproducibility but undermine commercial advantage. Our review concentrated on health care examination LLM performance and so did not account for LLM capability in more generalist evaluations that may still have valuable insights for optimizing health care capabilities.

Future Work

For policy and deployment decisions of LLMs to advance health care, we propose a new framework called *RUBRICC* (Regulatory, Usability, Bias, Reliability [Evidence and Safety], Interoperability, Cost, Codesign–Patient and Public Involvement and Engagement [PPIE]). See Multimedia Appendix 2.

Regulatory

LLMs have unique evaluation requirements. Medicines and Healthcare products Regulatory Agency (MHRA) device standards may categorize some clinical LLMs as type 2b devices [54], although medical knowledge progression (eg, National

XSL•FO
RenderX

Institute for Health and Care Excellence [NICE] guidelines) may require the recall of LLMs due to their capabilities being contained by period updates. Moreover, specific LLM standards for clinical commissioning are yet to be defined. It is important to forecast probable applications of LLMs, such as medical chatbots, clinical documentation, obtaining insurance preauthorization [55], and reviewing research papers [56]. Therefore, the regulatory responsibilities to patient safety and privacy will demand scrutiny on the grounds of LLMs' complexity, hardware, privacy, and real-time adaptation [55]. Developing rigorous and robust regulatory standards will require the commitment and input of key stakeholders, including clinicians, engineers, researchers, ethicists, health policymakers, and patients. Importantly, standards must be regularly adapted and revised to meet the rapidly advancing and evolving nature of LLMs.

Usability

Early adopter contexts will also likely be when the LLM is a clinical decision support tool integrated into various clinical contexts ranging from triage and differential diagnoses to imaging and medication decisions. Different geographies may apply these technologies differently, from the United States' insurance-based federated health landscape, which will likely apply LLMs to local health systems, in contrast to national data connectivity, which offers en masse precision LLM use across specialties, systems, and care tiers, such as in Estonia or the United Kingdom's NHS [57]. Academia will also be impacted, with publication assistance accelerating the role of LLM-coauthored literature [56].

Bias

The systematic review literature deals in terms of bias, which represents the content and function of an AI. The bias discussions in the included papers focused on the following variables: *within-item anchoring bias*, *grounding bias*, *chain-of-thought bias*, and *demographic bias*. By contrast, risk characterizes the contextual impact of an LLM in conversations that inform commissioning of generative medical AI and aligns with current regulatory frameworks for current and future AI tools [56]. Singhal et al [3] evaluated Med-PaLM 2 using the following LLM answer risk framework: *more inaccurate/irrelevant information*, *omits more information*, *more evidence of demographic bias*, *greater extent of harm*, *greater likelihood of harm*. A key consideration is the risk matrix of LLM errors. There are unique requirements for LLM reporting that do not easily map onto established criteria, such as the Standards for Reporting of Diagnostic Accuracy Study (STARD) 2015 checklist [58], and can be incorporated into the upcoming STARD-AI [59]. Associated challenges related to bias include "hallucination" and privacy, threatening the reliability of these LLM services.

Reliability

Evidence

Differences in reference standards and thresholds for diagnostic accuracy make comparison of LLM studies difficult in this nascent field, undermining the pathway to integration into health systems. These problems can only be addressed by specific reporting standards for AI studies [59,60], with design accuracy to address issues of reproducibility, transparency, and efficacy [61]. Further evidence is needed to develop reliable guidelines [62]. We therefore await guidelines that accommodate LLM utility to enable higher-quality and more consistent reporting, which in turn will empower the MHRA and the Food and Drug Administration (FDA) to be able to evaluate LLM risk. Specifically, the development of AI-specific risk-of-bias tools, such as QUADAS-AI, will aid in establishing the risk of bias for evidence synthesis of clinical LLM studies, allowing clinically relevant conclusions to be drawn more confidently [43].

Safety

Multidisciplinary secure data environments (SDEs) [63] must be established with cybersecurity standards to assuage recognized concerns about AI manipulation and displacement of human welfare priorities [64]. There remain established concerns about the regulated integration of LLMs into established clinical workstreams in view of "hallucination" concerns, which will require a quality management system to ensure compliance with best practices to mitigate risk to patients.

Interoperability

Although data flows in the NHS have been mapped [65], there is a growing demand for infrastructural transformation to reduce data inequalities and avoid the digital exclusion of unrepresented and underprivileged groups. A particular challenge includes multimodal data linkages and interoperability with integration of LLM tools in multiple different scenarios across the health service. One must be careful to consider how secondary or primary care data might be used differently to inform population health tools.

Cost

The economic considerations for LLMs can be organized into procurement, data processing, housing and cloud storage, management, and usability costs. Training costs have declined around 80% on models similar to ChatGPT-3 over the past 3 years [62]. The input cost is the number of tokens passed as prompts to the application programming interface (API), and the output cost is dependent on the number of tokens returned [63]. Therefore, for medical free-text record summarization, there is a large input cost dominated by the high quantity of tokens for each prompt. Self-hosted LLMs incur cloud service costs to run the models; it is notable that ChatGPT-4 (32 context length) is priced at US $60 input cost (per million tokens) and US $120 output cost (per million tokens) [66]. Further costs to consider include fine-tuning, which is most effective in improving performance on low-parameter models [67]; the clinical commissioning decisions related to these costs will be linked to the quality-adjusted life years (QALYs) associated with incremental performance improvements.

Codesign-PPIE

Public trust in LLMs can be built through a codesign process, adhering to INVOLVE [68] values, through respect, support, transparency, responsiveness, fairness of opportunity, and accountability. AI raises challenges for the codesign processes due to the disproportionate emphasis on procedures, patients

XSL•FO
RenderX

lacking genuine understanding, and concerns AI may exacerbate inequalities; this is best resolved by a focus on sociotechnical values and design humility to acknowledge to patients what the proposed technology cannot achieve for them [69]. Meanwhile, doctor-patient rapport will likely be enhanced due to LLMs alleviating administrative tasks and helping clinicians answer patient questions [70]. RUBRICC is a nascent area of work that will undergo further development to enable utility and impact in the field.

Conclusion

LLMs offer promise to remediate health care demand and staffing challenges by providing accurate and efficient context-specific information to critical decision makers. However, progress is obstructed by inconsistent reporting and an imbalance of resources between commercial interests and public sector regulators to independently evaluate potential LLM services. The ability of LLMs to pass the USMLE does not mean that the models answer useful questions to practicing clinicians [71]. Although initial results show impressive accuracy in isolated studies, there is an immediate need for a framework, such as RUBRICC, to evaluate this emergent technology and facilitate robust clinical commissioning decisions to benefit patients.

Data Availability

The authors declare that all the data included in this study are available within the paper and multimedia appendices. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Authors' Contributions

WJW was responsible for conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, and writing—original draft; JZ, AN, and AD for conceptualization, supervision, and writing—review and editing; AG for methodology and data curation; and HA for conceptualization, methodology, supervision, and writing—review and editing. No generative AI was used in any portion of the manuscript writing.

Conflicts of Interest

Covidence software was used with funding from the Imperial Healthcare National Health Service (NHS) Trust and Imperial College London. JZ is funded by the Wellcome Trust (grant number 203928/Z/16/Z). AD is chair for the Preemptive Medicine and Health Security Initiative at Flagship Pioneering. HA is chief scientific officer, Preemptive Health and Medicine, Flagship Pioneering.

Multimedia Appendix 1
Summary of studies selected.
[DOCX File , 33 KB - jmir_v26i1e56532_app1.docx]

Multimedia Appendix 2
RUBRICC (Regulatory, Usability, Bias, Reliability [Evidence and Safety], Interoperability, Cost, Codesign–Patient and Public Involvement and Engagement [PPIE]), a framework for LLM clinical policy decisions. AI: artificial intelligence; LLM: large language model.
[PNG File , 216 KB - jmir_v26i1e56532_app2.png]

Multimedia Appendix 3
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 1786 KB - jmir_v26i1e56532_app3.pdf]

References

1. Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthc J 2022 Jul;9(2):190-193 [FREE Full text] [doi: 10.7861/fhj.2022-0013] [Medline: 35928184]
2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887]
3. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. A responsible path to generative AI in healthcare. Google Cloud. 2023 Apr 14. URL: https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model [accessed 2024-10-08]
4. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023 Mar 21;27(1):120 [FREE Full text] [doi: 10.1186/s13054-023-04393-x] [Medline: 36945051]

XSL•FO
RenderX

5. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135 [FREE Full text] [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
6. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097]
7. Systematic review and meta-analysis of the accuracy and capability of artificial intelligence solutions in healthcare exams and certificates. Center for Open Science. 2023 May 26. URL: https://osf.io/xqzkw [accessed 2024-10-08]
8. The world's #1 systematic review tool. Covidence. URL: https://www.covidence.org [accessed 2024-10-08]
9. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
10. Venigalla A, Frankle J, Carbin M. BioMedLM: a domain-specific large language model for biomedical text. Stanford Center for Research on Foundation Models (CRFM) and MosaicML. 2022 Dec. URL: https://www.mosaicml.com/blog/introducing-pubmed-gpt [accessed 2024-10-08]
11. Sharma P, Thapa K, Dhakal P, Upadhaya MD, Adhikari S, Khanal SR. Performance of ChatGPT on USMLE: unlocking the potential of large language models for AI-assisted medical education. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2307.00112]
12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
13. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR, North East Trainee Research in Ophthalmology Network (NETRiON). Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye (Lond) 2023 Dec;37(17):3530-3533 [FREE Full text] [doi: 10.1038/s41433-023-02563-3] [Medline: 37161074]
14. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
15. Strong E, DiGiammarino A, Weng Y, Basaviah P, Hosamani P, Kumar A, et al. Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv Preprint posted online 2023 [FREE Full text] [doi: 10.1101/2023.03.24.23287731] [Medline: 37034742]
16. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. Med Teach 2023 Jun;45(6):665-666. [doi: 10.1080/0142159X.2023.2187684] [Medline: 36905610]
17. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the Plastic Surgery In-Service Exam. Aesthet Surg J 2023 May 04:sjad130. [doi: 10.1093/asj/sjad130] [Medline: 37140001]
18. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 2023 Apr 26;9:e47737 [FREE Full text] [doi: 10.2196/47737] [Medline: 37099373]
19. Oh N, Choi G, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273 [FREE Full text] [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
20. Angel MC, Rinehart JB, Canneson MP, Baldi P. Clinical knowledge and reasoning abilities of AI Large language models in anesthesiology: a comparative study on the ABA exam. medRxiv Preprint posted online 2023 [FREE Full text] [doi: 10.1101/2023.05.10.23289805] [Medline: 37292642]
21. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol 2023;13:1265024 [FREE Full text] [doi: 10.3389/fonc.2023.1265024] [Medline: 37790756]
22. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023 Aug 11;102(32):e34673 [FREE Full text] [doi: 10.1097/MD.0000000000034673] [Medline: 37565917]
23. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci 2023 Oct;366(4):291-295. [doi: 10.1016/j.amjms.2023.08.001] [Medline: 37549788]
24. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023 Nov;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
25. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
26. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How does ChatGPT perform on the Italian Residency Admission National Exam compared to 15,869 medical graduates? Ann Biomed Eng 2024 Apr 25;52(4):745-749. [doi: 10.1007/s10439-023-03318-7] [Medline: 37490183]

XSL•FO
RenderX

27. Weng T, Wang Y, Chang S, Chen T, Hwang S. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 01;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
28. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung F. Assessment of resident and AI chatbot performance on the University of Toronto Family Medicine Residency Progress Test: comparative study. JMIR Med Educ 2023 Sep 19;9:e50514 [FREE Full text] [doi: 10.2196/50514] [Medline: 37725411]
29. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9:e48039 [FREE Full text] [doi: 10.2196/48039] [Medline: 37768724]
30. Beaulieu-Jones BR, Shah S, Berrigan MT, Marwaha JS, Lai S, Brat GA. Evaluating capabilities of large language models: performance of GPT4 on surgical knowledge assessments. medRxiv Preprint posted online 2023 [FREE Full text] [doi: 10.1101/2023.07.16.23292743] [Medline: 37502981]
31. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, et al. Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol 2023;88:e430-e434 [FREE Full text] [doi: 10.5114/pjr.2023.131215] [Medline: 37808173]
32. Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performa poorly on the 2022 Self-assessment Study Program for urology. Urol Pract 2023 Jul;10(4):409-415. [doi: 10.1097/UPJ.0000000000000406] [Medline: 37276372]
33. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the Situational Judgement Test – a professional dilemma-based examination for doctors in the United Kingdom. JMIR Med Educ 2023 Aug 07;9:e48978 [FREE Full text] [doi: 10.2196/48978] [Medline: 37548997]
34. Skalidid I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023 Apr 24;4(3):279-281 Erratum in: Eur Heart J Digit Health. 2023 May 17;4(4):357. doi: 10.1093/ehjdh/ztad034. [doi: 10.1093/ehjdh/ztad029] [Medline: 37265864]
35. Mannam SS, Subtirelu R, Chauhan D, Ahmad HS, Matache IM, Bryan K, et al. Large language model-based neurosurgical evaluation matrix: a novel scoring criteria to assess the efficacy of ChatGPT as an educational tool for neurosurgery board preparation. World Neurosurg 2023 Dec;180:e765-e773 [FREE Full text] [doi: 10.1016/j.wneu.2023.10.043] [Medline: 37839567]
36. Bolton E, Hall D, Yasunaga M, Lee T, Manning C, Liang P. Stanford CRFM introduces PubMedGPT 2.7B. Stanford. 2022. URL: https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b [accessed 2024-10-08]
37. Yasunaga M, Leskovec J, Liang P. LinkBERT: pretraining language models with document links. arXiv Preprint posted online 2022. [doi: 10.18653/v1/2022.acl-long.551]
38. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
39. Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: a large language model for science. arXiv Preprint posted online 2022. [doi: 10.48550/arXiv.2211.09085]
40. Yasunaga M, Bosselut A, Ren H, Zhang X, Manning CD, Liang PS, et al. Deep bidirectional language-knowledge graph pretraining. 2022 Presented at: NeurIPS 2022: 36th Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA p. 37309-37323.
41. Moradi M, Blagec K, Haberl F, Samwald M. GPT-3 models are poor few-shot learners in the biomedical domain. arXiv Preprint posted online 2021. [doi: 10.48550/arXiv.2109.02555]
42. Jin D, Pan E, Oufattole N, Weng W, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci 2021 Jul 12;11(14):6421 [FREE Full text] [doi: 10.3390/app11146421]
43. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021 Oct;27(10):1663-1665 [FREE Full text] [doi: 10.1038/s41591-021-01517-0]
44. Adherence to Tripod. Tripod. URL: https://www.tripod-statement.org/adherence/ [accessed 2024-10-08]
45. Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, et al. Practical and ethical challenges of large language models in education: a systematic literature review. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2303.13379]
46. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 09;22(1):772 [FREE Full text] [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
47. Masoumian Hosseini M, Masoumian Hosseini ST, Qayumi K, Ahmady S, Koohestani HR. The aspects of running artificial intelligence in emergency care; a scoping review. Arch Acad Emerg Med 2023;11(1):e38 [FREE Full text] [doi: 10.22037/aaem.v11i1.1974] [Medline: 37215232]
48. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 07;8(2):e35587 [FREE Full text] [doi: 10.2196/35587] [Medline: 35671077]

XSL•FO
RenderX

49. Mir MM, Mir GM, Raina NT, Mir SM, Mir SM, Miskeen E, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140 [FREE Full text] [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
50. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019 Oct 4;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
51. Synthetic data's role in LLM evolution. Syntheticus. URL: https://tinyurl.com/ye2ud5sr [accessed 2024-10-08]
52. Smith CS. What large models cost you – there is no free AI lunch. Forbes. 2024 Jan 1. URL: https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/ [accessed 2024-10-08]
53. Cath C, Wachter S, Mittelstadt B, Taddeo M, Floridi L. Artificial intelligence and the 'good society': the US, EU, and UK approach. Sci Eng Ethics 2018 Apr 28;24(2):505-528. [doi: 10.1007/s11948-017-9901-7] [Medline: 28353045]
54. Section 5 - classification of general medical devices. Medicines & Healthcare products Regulatory Agency. 2022 Jun 26. URL: https://tinyurl.com/3tyra9t7 [accessed 2024-10-08]
55. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
56. Conroy G. How ChatGPT and other AI tools could disrupt scientific publishing. Nature 2023 Oct 01;622(7982):234-236. [doi: 10.1038/d41586-023-03144-w] [Medline: 37817033]
57. European approach to artificial intelligence. European Commission. URL: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence [accessed 2024-10-08]
58. Bossuyt P, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015 Oct 28;351:h5527 [FREE Full text] [doi: 10.1136/bmj.h5527] [Medline: 26511519]
59. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. Nat Med 2020 Jun 08;26(6):807-808. [doi: 10.1038/s41591-020-0941-1] [Medline: 32514173]
60. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020 Sep 09;370:m3164 [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]
61. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927 [FREE Full text] [doi: 10.1136/bmj.l6927] [Medline: 32198138]
62. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/s1470-2045(19)30149-4]
63. Secure data environment for NHS health and social care data - policy guidelines. Department of Health & Social Care. 2022 Dec 23. URL: https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines [accessed 2024-10-08]
64. AI concerns: manipulating humans, or even replacing them. MIT Sloan. 2023 May 23. URL: https://tinyurl.com/28jemrmy [accessed 2024-10-08]
65. Zhang J, Morley J, Gallifant J, Oddy C, Teo J, Ashrafian H, et al. Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation. Lancet Digital Health 2023 Oct;5(10):e737-e748 [FREE Full text] [doi: 10.1016/s2589-7500(23)00157-7]
66. The economics of large language models. The cost of ChatGPT-like search, training GPT-3, and a general framework for mapping the LLM cost trajectory. Sunyan. 2023 Jan 21. URL: https://sunyan.substack.com/p/the-economics-of-large-language-models [accessed 2024-10-08]
67. The $360K question about large language models economics. TrueFoundry. 2023 Jun 22. URL: https://www.truefoundry.com/blog/economics-of-large-language-models [accessed 2024-10-08]
68. National Institute for Health and Care Research. URL: https://www.nihr.ac.uk/ppi-patient-and-public-involvement-resources-applicants-nihr-research-programmes [accessed 2024-10-08]
69. Donia J, Shaw JA. Co-design and ethical artificial intelligence for health: an agenda for critical research and practice. Big Data Soc 2021 Dec 17;8(2):205395172110652 [FREE Full text] [doi: 10.1177/20539517211065248]
70. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
71. Dash D, Thapa R, Banda JM, Swaminathan A, Cheatham M, Kashyap M, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2304.13714]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

BERT: Bidirectional Encoder Representations from Transformers
ChatGPT: Chat Generative Pretrained Transformer
DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining
genAI: generative artificial intelligence
GPT: Generative Pretrained Transformer
LLM: large language model
Med-PaLM 2: Medical Patient Language Model 2
MeSH: Medical Subject Headings
NHS: National Health Service
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RUBRICC: Regulatory, Usability, Bias, Reliability [Evidence and Safety], Interoperability, Cost, Codesign–Patient and Public Involvement and Engagement (PPIE)
STARD: Standards for Reporting of Diagnostic Accuracy Study
USMLE: United States Medical Licensing Examination

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 18.01.24; peer-reviewed by A Bortkiewicz, N Domingues, Z Yao; comments to author 25.06.24; revised version received 26.06.24; accepted 25.09.24; published 05.11.24.

Please cite as:
Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H
The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e56532
URL: https://www.jmir.org/2024/1/e56532
doi:10.2196/56532
PMID:

©William J Waldock, Joe Zhang, Ahmad Guni, Ahmad Nabeel, Ara Darzi, Hutan Ashrafian. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review

Jie Gong[1], BD; Dandan Gu[1], MD; Suyun Dong[1], BD; Wangqin Shen[2], PhD; Haiou Yan[1], BD; Juan Xie[1], BD

[1]Department of Nursing, Affiliated Hospital of Nantong University, Nursing and Rehabilitation School of Nantong University, Nantong, China
[2]School of Nursing and Rehabilitation, Nantong University, Nantong, China

Corresponding Author:
Juan Xie, BD
Department of Nursing
Affiliated Hospital of Nantong University
Nursing and Rehabilitation School of Nantong University
No. 20 West Temple Road
Nantong, 226001
China
Phone: 86 13912288333
Email: xjjx@ntu.edu.cn

Related Article:

This is a corrected version. See correction statement: http://jmir.org/2024/1/e68783/

Abstract

Background: With the advancement of cervical cancer elimination strategies, promoting human papillomavirus (HPV) vaccination is essential to achieving this goal. The issue of how to structure and develop message content to promote HPV vaccination is a debatable issue.

Objective: The efficacy of gain-loss framing in vaccination contexts is disputed. Our study aimed to elucidate the consequences of message framing on attitudes, intentions, and behavioral tendencies toward HPV vaccination, with the objective of refining message framing strategies and their elements.

Methods: This systematic review adhered strictly to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guideline reporting standards to comprehensively retrieve, extract, and integrate data. We searched databases, including PubMed, Embase, Scopus, and Web of Science, for literature published from database construction to August 15, 2023. Literature screening, data extraction, and quality evaluation were performed by 2 researchers. Intervention studies published in English, conducted with populations with children eligible for HPV vaccination, and involving message framing were included. Attitudes, intentions, and behaviors served as outcome evaluation criteria.

Results: A total of 19 intervention studies were included. Gain-loss framing had no clear effect on vaccination attitudes nor intentions. Loss framing showed a weak advantage at improving HPV vaccination attitudes or intentions, but the evidence was not strong enough to draw definitive conclusions. The impact of gain-loss framing on HPV vaccination behaviors could not be determined due to the limited number of studies and the qualitative nature of the analysis.

Conclusions: Combining gain-loss framing with other message framing approaches may be an effective way to enhance the effect of message framing. More high-quality message framing content and exploring alternative moderator or mediator variables are required to support the conclusion.

Trial Registration: CRD42023451612; https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=451612

(J Med Internet Res 2024;26:e52738) doi:10.2196/52738

KEYWORDS

message framing; gain-loss framing; human papillomavirus; vaccination; attitude; intention; behavior; systematic review; PRISMA

XSL•FO
RenderX

Introduction

Almost 100% of cervical cancers are associated with persistent high-risk human papillomavirus (HPV) infection [1], and HPV 16 and 18, the 2 oncogenic genotypes, are the cause of more than 70% of all cervical cancers around the world [2]. In addition, 90% of warts, 88% of anal cancers, 50% of penile cancers, and 43% of vulvar cancers are attributed to HPV infection [3]. HPV vaccination prevents more than 90% of cancers and precancerous lesions caused by HPV and minimizes the morbidity or mortality associated with HPV-related diseases, while the incidence of adverse events arising from HPV vaccination is relatively modest [4]. The rate of HPV vaccination is reportedly 15% worldwide and is significantly different across nations: 71.5% in the United States, 89.5% in Britain, 8.8% in Singapore, and 2.4%-9.1% in Hong Kong, China [5,6]. Therefore, increasing the rate of HPV vaccination to prevent and control infectious diseases caused by HPV is a global issue that is worth paying attention to.

“Message framing” refers to the distinctive “framing effects” of messages [7]. The essence of “framing” is to select and emphasize messages, and the specific structure and emphasis of messages create different categories of message framing that allow people to view problems through a variety of perspectives and ultimately influence behavioral preferences [8,9]. The most classic type of framing is gain-loss framing, with the gain framing emphasizing the positive outcomes that will result in someone taking action (or negative outcomes that will be avoided) and loss framing emphasizing the negative consequences that will result in not doing something (or positive consequences that will be lost) [10-13]. Message framing plays an essential role in the dissemination of health information and persuading recipients to make behavioral modifications, with the core purpose being to orient them to modify attitudes, intentions, or behaviors toward specific health hot spots [14,15]. According to the theory of reasoned action, attitudes and intentions predict behavior, while attitudes, intentions, and behaviors measure the effectiveness of health messages [16-18].

Controversial effects of gain-loss framing on disease prevention behaviors related to vaccination have been observed. Previous reviews have shown that gain framing was more effective at persuasion for disease prevention scenarios, while loss framing worked better in terms of disease detection [19]. However, Lee and Aaker [20] proposed, based on 6 experiments, that gain framing is more appealing when the message content and settings focus on promoting and facilitating certain behaviors, whereas loss framing is more effective at preventing a phenomenon to occur. However, O'Keefe and Nan [21] performed a meta-analysis to address the influences of message framing on vaccination that revealed no difference between gain and loss framing. Therefore, the aim of our study was to systematically review the effects of interventions based on gain-loss framing for attitudes, intentions, or behaviors related to HPV vaccination and to provide directions and recommendations for designing effective message content.

Methods

Search Strategy

This systematic review is reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [22]. The study protocol was registered in PROSPERO (number CRD42023451612). According to the registered protocol, we initially planned to use RevMan 5.4 software for data synthesis and meta-analysis. However, due to the heterogeneity of the included studies and the limited availability of comparable data, we were unable to perform a quantitative analysis. Therefore, we conducted a qualitative analysis to synthesize the findings. This deviation from the protocol is reported here to ensure transparency.

Computerized retrieval was conducted across 4 databases—PubMed, Embase, Scopus, and Web of Science—with the search period spanning from the inception of each database to August 15, 2023. A combination of MESH terms and keywords was adopted and adjusted to the respective features of the databases. We included 3 essential components in the search strategy and connected them with “AND”: (1) ‘message framing’ or ‘message fram*’ or ‘information framing’ or ‘information framework’ or ‘framing effect*’ or ‘gain fram*’ or ‘loss fram*’ or ‘positive fram*’ or ‘negative fram*’; (2) ‘Human papillomavirus viruses’ or ‘human papillomavirus virus’ or ‘human papillomavirus’ or ‘HPV human papillomavirus’; (3) ‘vaccination’ or vaccin*’ or ‘active immunization’ or ‘mass vaccination’ or ‘vaccination refusal’ or ‘anti-vaccination movement.’ In addition, the references of included studies were searched to obtain supplementary materials.

Study Selection

The recommended age for HPV vaccination is between 9 years and 14 years for both boys and girls, and the decision to administer the vaccination in cases involving minors is typically made by their legal guardians [23,24]. The inclusion criteria consisted of studies that (1) included participants and any of their children eligible for the HPV vaccination, with no specific gender or age restrictions; (2) involved a gain-loss framing of intervention or combined with other message stimulation modalities; (3) compared gain framing with loss framing in groups, compared gain-loss framing against other message framing or against no message framing; (4) measured the effect of gain-loss framing on vaccination and the differences between them with attitudes, intentions, or behaviors as outcome evaluation modalities; (5) were intervention studies (including randomized controlled trials [RCTs] and quasiexperimental studies) published in English. Duplicate publications and publications with missing data were excluded.

We imported the literature into EndNote 21 to remove duplicates, then 2 researchers independently screened the documents and cross-checked them by reading the titles, abstracts, and full texts. Any disagreements were resolved by discussion or consultation with a third researcher.

XSL•FO
RenderX

Data Extraction and Analysis

Data extraction was performed by 2 researchers, and any disagreements that arose were adjudicated by a third researcher. The following content was extracted: (1) basic information of the included studies, such as the first author, year of publication, country, subgroup, theoretical model, and methods or means of message dissemination; (2) baseline characteristics of the population, including the sample size, gender, and age; (3) outcome indicators and main findings. Due to the statistically significant heterogeneity of the publications, quantitative synthesis and a meta-analysis could not be accomplished; therefore, this study focused on the qualitative synthesis. Data extraction was performed using Microsoft Excel 2022.

Quality Assessment

Two researchers independently evaluated the risk of bias for included studies and cross-checked the results, requesting a third researcher to negotiate a solution if they could not reach mutual agreement. The risk of bias was evaluated using the Cochrane Collaboration Risk of Bias (RoB) 2 [25]. For quasiexperimental studies, the selective bias item was rated as high risk because randomization was not conducted [26].

Results

Literature Search

The preliminary screening obtained 4905 relevant studies. After duplicates were removed, the titles and abstracts of 3418 studies were screened, and the full text was screened for 49 of those studies. Finally, 19 studies were included [27-45], 2 of which were quasiexperimental studies that met the inclusion criteria [33,40]. The flowchart of the screening process is shown in Figure 1 [22].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

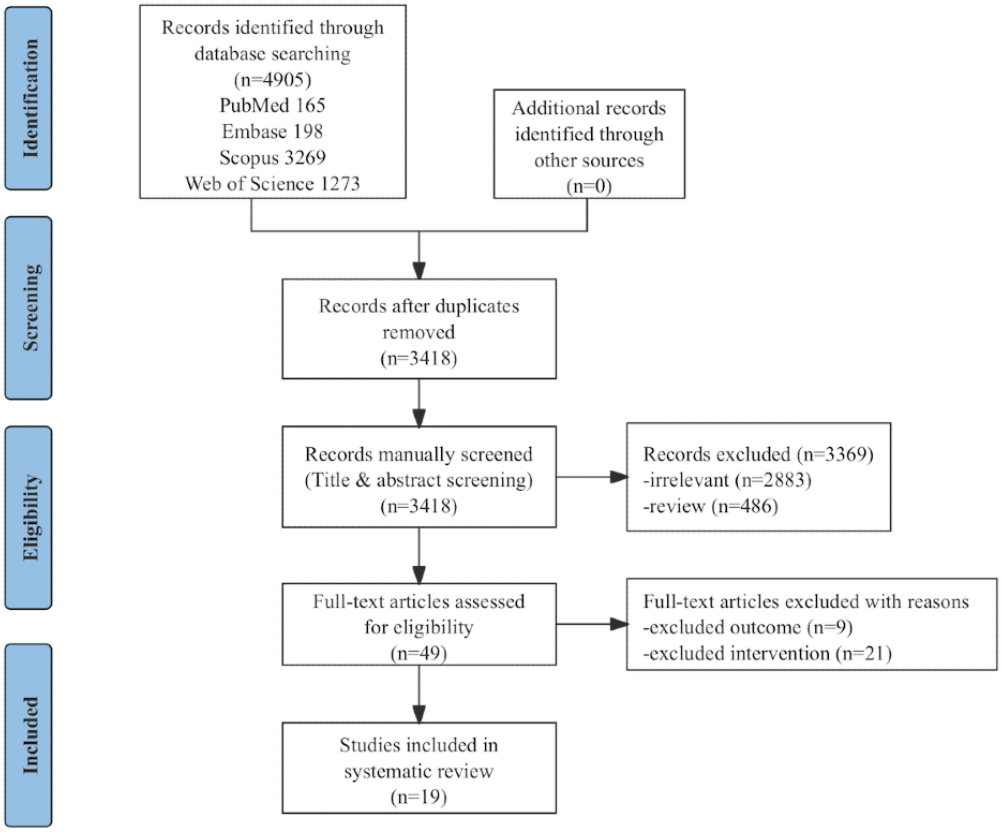

Characteristics of the Included Studies

The included studies were published between 2007 and 2023; 14 publications were from the United States [27-29,31,32,34-37,39,41-44], 2 were from China [38,45], and the other 3 were from Canada [33], Ireland [30], and Singapore [40]. The number of participants ranged from 72 to 739, with 5124 participants in total; the average age ranged from 18 years to 42 years; and 69.1% (3541/5124) of the participants were female.

Participants were categorized into gain-framing or loss-framing groups in 8 studies [27,30-32,35,37,40,44], and the remaining 11 publications added various message stimulation modalities or control groups [28,29,33,34,36,38,39,41-43,45]. Of the 19 studies, 9 delivered messages online in the form of videos or web pages, for example [32-35,39-42,45], and the rest implemented offline interventions via traditional printed materials like booklets and advertisements [27-31,36-38,43,44]. All the studies were based on prospect theory [27-45], and some of the studies applied the health belief model [28,34,36] or theory of planned behavior [30,34]. Participant outcomes were assessed in 13 studies [27-29,32,34-36,38-40,42,43,45], and 6 papers focused on participants' attitudes, intentions, or behaviors toward their children's HPV vaccinations [30,31,33,37,41,44]. Intentions toward HPV vaccination were discussed in 13 studies [27-29,31,32,37-42,44,45], 5 studies described attitudes on HPV

XSL•FO
RenderX

vaccination [30,33,35,36,43], and 1 study evaluated HPV vaccination behaviors directly [34]. Table 1 summarizes the basic information of the included literature.

XSL•FO
RenderX

Table 1. Characteristics of eligible studies about human papillomavirus (HPV) vaccination in the systematic review.

First author, year	Country	Sample, n (M[a] or F[b])	Sample age (years), mean (SD)	Vaccine recipient	Vaccine recipient age (years), mean (SD)	Group	Theory or model	Format (channel)	Results collection time
Gerend, 2007 [27]	US	121 (F)	19.03 (1.09)	Self	__[c]	Gain or loss	PT[d]	Health booklet (offline)	Post intervention
Gerend, 2008 [28]	US	237 (F)	18.6 (1.1)	Self	—	Gain or loss and 1 shot or 6 shots	PT, HBM[e]	Booklet (offline)	Postintervention
Gerend, 2009 [29]	US	126 (M)	19.9 (1.9)	Self	—	Gain or loss and gray or red	PT	Binder (offline)	Pre- and postintervention
Fahy, 2010 [30]	Ireland	72 (F)	41.56 (5.81)	Daughter	8-16[f]	Gain or loss	PT, TPB[g]	Study booklet (offline)	Post intervention
Lechuga, 2011 [31]	US	150 (F)	33.72 (7.95)	Daughter	10.94 (4.01)	Gain or loss	PT	Laminated brochures (offline)	Pre- and postintervention
Nan, 2011 [32]	US	229 (M: 129; F: 100)	20.18 (1.47)	Self	—	Gain or loss	PT	Study website (online)	Postintervention
Gainforth, 2012 [33]	Canada	367 (M, F)	42.55 (4.73)	Daughter or son	11.32 (2.61)	Gain, loss, or mixed	PT, PMT[h], HSM[i]	Messages (online)	Postintervention
Gerend, 2012 [34]	US	739 (F)	21 (1.8)	Self	—	Gain, loss, or no framed information	PT, HBM, TPB	Video (offline + online)	10 months postintervention
Nan, 2012 [35]	US	383 (M: 171, F: 212)	20.05 (1.45)	Self	—	Gain or loss	PT	Web page (online)	Postintervention
Park, 2012 [36]	US	108 (M: 27, F: 81)	20.5 (1.08)	Self	—	Gain or loss and high risk or low risk	PT, HBM	Vaccine advertising (offline)	Postintervention
Nan, 2016 [37]	US	193 (M: 52, F: 141)	36.2 (9.19)	Daughter or son	9-17[f]	Gain or loss	PT	Booklet (offline)	Postintervention
Wen, 2016 [38]	China	156 (M: 56, F: 100)	19.83 (0.91)	Self	—	Gain or loss and present or future	PT, CLT[j], TDT[k]	Text-based message (offline)	Postintervention
Lee, 2017 [39]	US	142 (M: 30, F: 112)	22.44 (1.22)	Self	—	Gain or loss and SNSs[l] or traditional media	PT, HBM	Web page (online)	Postintervention
Kim, 2018 [40]	Singapore	226 (F)	20.39 (1.57)	Self	—	Gain or loss	PT, RFT[m]	Web page (online)	Postintervention
Liu, 2018 [41]	US	431 (F)	30.16 (6.38)	Daughter or son	N/A[n]	Gain, loss, or narrative	PT	Newsletters (online)	Postintervention
Liu, 2018 [42]	US	455 (M, F)	20.42 (3.06)	Self	—	Gain or loss and English or Chinese	PT, CTR[o], CCT[p]	Video (online)	Postintervention
Kim, 2020 [43]	US	347 (M: 132, F: 215)	22.2 (2.62)	Self	—	Gain or loss and future-thinking, past-thinking, or no-thinking	PT, EFM[q], EFT[r]	Health message (offline)	Postintervention
Richards, 2021 [44]	US	184 (M: 50, F: 134)	36.13 (9.07)	Daughter or son	9-17[f]	Gain or loss	PT, PRT[s]	Pamphlet (offline)	Postintervention
Huang, 2023 [45]	China	458 (M)	26.36 (2.88)	Self	—	Gain or loss and self, other, or self-other	PT	Posters (online)	Postintervention

XSL•FO
RenderX

[a]M: male.
[b]F: female.
[c]Not applicable because the vaccine recipient sample is the same as the overall study sample.
[d]PT: prospect theory.
[e]HBM: health belief model.
[f]Range.
[g]TPB: theory of planned behavior.
[h]PMT: protection motivation theory.
[i]HSM: heuristic-systematic model.
[j]CLT: construal level theory.
[k]TDT: temporal discounting theory.
[l]SNSs: social networking sites.
[m]RFT: regulatory focus theory.
[n]N/A: not available.
[o]CTR: cultural theory of risk.
[p]CCT: cultural cognition thesis.
[q]EFM: emotions-as-frame model.
[r]EFT: episodic future thinking.
[s]PRT: psychological reactance theory.

Quality Assessment of the Included Studies

Among the 19 RCTs included, based on the RoB 2 assessment results, 2 studies were judged to have a high risk of bias, while the remaining 17 may have a risk of bias. Although 17 studies used randomization methods, none of them specified the implementation details for the randomization [27-29,31,32,34-39,41-45]. The random sampling methods in 2 studies were not clarified [33,40], while 2 studies reported specific allocation concealment measures [30,32]. Figure 2 summarizes the results of the risk of bias assessments for the included studies.

Figure 2. Risk of bias summary [27-45].

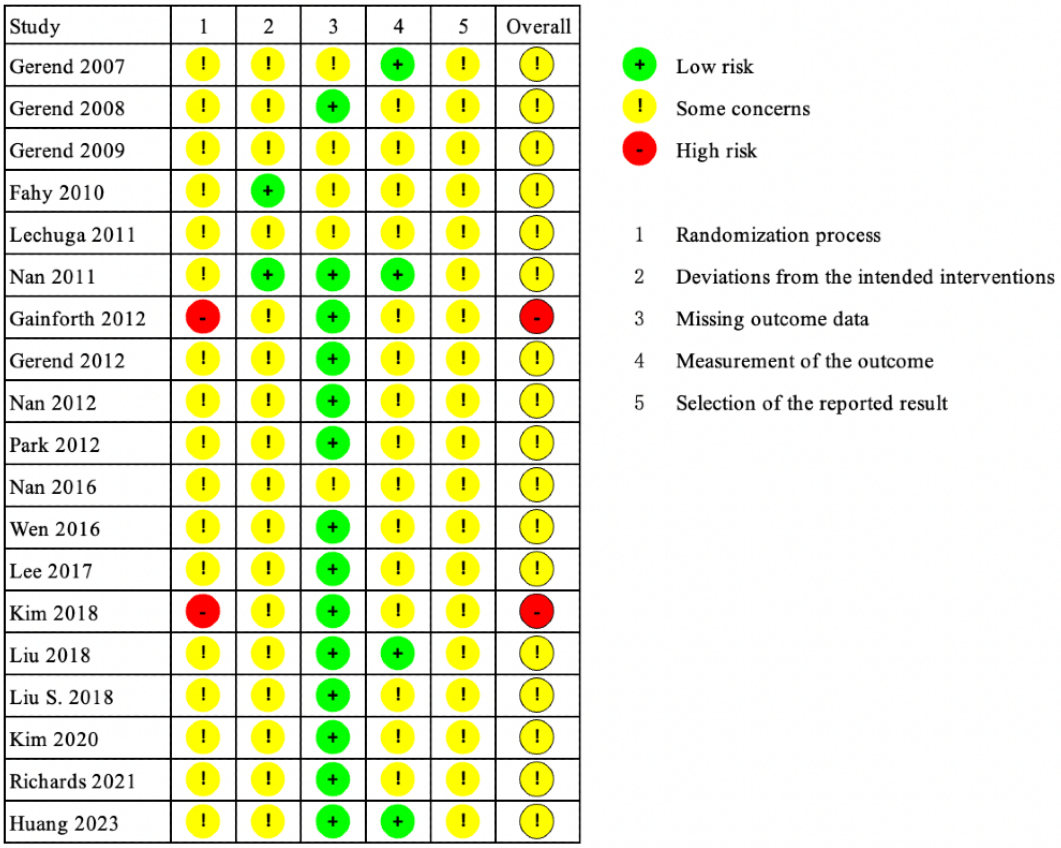

XSL•FO
RenderX

Main Effects of Message Framing

Participant attitudes, intentions, or behaviors toward HPV vaccination were reported in the included literature. Of the 18 studies that evaluated attitudes or intentions toward vaccination, all of them assessed outcomes right after the intervention [27-33,35-45]. A single study measured physical behaviors for HPV vaccination 10 months postintervention [34]. The components of the intervention were related to the benefits of HPV vaccination and the losses of not getting the HPV vaccine. However, the diversity in the measuring instruments and statistical variation across the 19 studies resulted in significant heterogeneity, which prevented quantitative data synthesis. Table 2 summarizes the main findings of the included literature.

Of the 19 studies, 18 reported main or interaction effects of gain-loss framing on vaccination intentions [27-33,35-45]: 8 were focused on the impact of gain-loss framing–based interventions for HPV vaccination intentions [28,29,31,33,36,37,42,43], and 2 of the studies found a significant main effect of gain-loss framing on HPV vaccination intentions, which increased in both the gain-framing and loss-framing groups, as participants' vaccination intentions rose from baseline [28,31]. The gain-loss framing was shown to have no major effect on vaccination intentions in 5 studies [29,33,36,42,43], and mothers' intentions to have their daughters vaccinated were not affected by the gain-loss framing [30]. The differences between gain-framing and loss-framing on HPV vaccination intentions were compared in 12 trials [27-29,31,32,35-39,41,44], of which 8 studies found that the loss framing produced higher vaccination intentions than the gain framing [27-29,32,35-37,39], 3 studies showed no statistically significant difference between the 2 groups after comparison [31,38,41], and 1 RCT revealed loss framing significantly reduces the intention to vaccinate [44].

The influence of message framing on attitudes toward HPV vaccination were explored by 5 studies. [30,33,35,36,43]. Gain-loss framing had no significant impact on attitudes toward HPV vaccination directly [30,33,43]. In contrast, Park [36] found that gain-loss framing had a major effect on HPV vaccination attitudes. Furthermore, a comparison of the effect of 2 message frames on vaccination attitudes revealed that loss framing produced better attitudes about vaccination than gain framing [35].

In 1 study that measured HPV vaccination rates, 6% of participants received one or more doses of an HPV vaccine 10 months after the intervention, and the comparison of the difference in vaccination rates between message frames showed that the rates of HPV vaccination were almost the same with the different message frames [34].

XSL•FO
RenderX

Table 2. Main findings of outcome variables and relevant moderator or mediator variables.

First author, year	Outcome variables	Moderator or mediator variables	Main findings
Gerend, 2007 [27]	Intention: a 6-point scale with 5 items	Sexual behavior, motivational orientation	• The loss-framed message leads to higher vaccination intentions than the gain-framed message for women engaging in risky sexual behaviors or with high avoidance motivation.
Gerend, 2008 [28]	Intention: a 6-point scale with 5 items	Behavioral frequency (1 shot or 6 shots); perceived severity, benefits, barriers, susceptibility, and self-efficacy	• Message framing has a significant main effect on HPV[a] vaccination intentions. • Participants exposed to the loss-framed message had higher vaccination intentions than those exposed to the gain-framed message. • Message framing interacted with behavioral frequency, and the loss-framed message led to greater vaccination intentions than gain-framed message in the case of 1 vaccine shot. • Perceived barriers and severity have no mediating effect; perceived susceptibility and self-efficacy can predict vaccination intentions, while perceived benefits cannot.
Gerend, 2009 [29]	Intention: a 6-point scale with 5 items	Color priming (gray or red)	• Message framing does not have a major impact on vaccination intentions; of the framed messages labeled with red rectangular boxes, those who read the loss-framed message had higher vaccination intentions than those who read the gain-framed message.
Fahy, 2010 [30]	Intention: a 7-point scale with 3 items; attitude: 3 bipolar semantic differential scales	Attitude, normative beliefs;, PBC[b]	• No main effect of message framing on vaccination attitudes. • Mothers' high intentions to have their daughters vaccinated is not influenced by message framing effects. • Attitude, normative beliefs, and PBC explain 69.5% of the variation in intentions.
Lechuga, 2011 [31]	Intention: a 7-point scale with 5 items	Ethnic group: Hispanic, non- Hispanic White, African American	• Vaccination intention is higher with gain framing or loss framing than at baseline. • For the African Americans and Hispanics, the loss-framed message generates higher vaccination intentions; for non-Hispanic Whites, there was no difference between the 2 frames.
Nan, 2011 [32]	Intention: a 7-point scale with 3 items	Motivational orientation	• The loss-framed message leads to higher vaccination intentions than the gain-framed message in avoidance-oriented participants.
Gainforth, 2012 [33]	Attitude: a 7-point scale with 5 semantic differential items; intention: a 7-point scale including 6 items	Sex of the parent and child	• Parents' intentions to vaccinate their children against HPV is not significant. • There was no effect based on gender, message frame, and parents' attitudes.
Gerend, 2012 [34]	Behavior: HPV vaccination rates	Perceived susceptibility, severity, benefits; safety concerns; cost; attitudes; norms; self-efficacy	• HPV vaccination rates are almost the same across different message frames. • Perceived susceptibility, perceived safety, and vaccine cost can predict vaccine uptake, while perceived severity and benefits cannot.
Nan, 2012 [35]	Attitude: a 7-point scale with 3 semantic differential items; intention: a 7-point scale with 3 items	Time orientation	• The loss-framed message leads to greater vaccination attitudes and intentions than the gain-framed message. • Participants with a future-mind had more favorable intentions and attitudes to be vaccinated. • For participants with a present-mind, a loss-framed message leads to greater vaccination attitudes than a gain-framed message, while the 2 frames perform equally in future-minded participants.
Park, 2012 [36]	Attitude: a 7-point scale with 9 semantic differential items; intention: a 7-point scale with 4 items	Perceived risk	• Message framing has a main effect on HPV vaccination attitudes. • Participants exposed to the loss-framed message had higher vaccination intentions. • Participants with high risk perceptions have a strong intention to be vaccinated.

XSL•FO
RenderX

First author, year	Outcome variables	Moderator or mediator variables	Main findings
Nan, 2016 [37]	Intention: 3 scoring items	Perceived susceptibility	• Message framing cannot predict vaccination intentions. • The loss-framed message produced more vaccination intentions than the gain-framed message when perceived susceptibility was low, whereas the gain-framed message produced more vaccination intentions than the loss-framed message when perceived susceptibility was high.
Wen, 2016 [38]	Intention: a scale with 4 items	Temporal distance; prior knowledge	• The interaction effects between message framing and temporal distance were not significant. • There is no difference between loss-framed messages, gain-framed messages, future-framed messages, and present-framed messages in terms of behavioral intention. • Prior knowledge has main effects on behavioral intention.
Lee, 2017 [39]	Intention: a 7-point scale with 4 items	Media channel	• The loss-framed message produced more vaccination intentions than the gain-framed message. • Participants showed a higher level of behavioral intention to get HPV immunization after viewing the loss-framed message post on SNSs[c].
Kim, 2018 [40]	Intention: a 5-point scale with 3 items	Transportation, self-referent, emotions, free vaccine, paid vaccine	• There was an indirect impact of message framing on vaccination intentions through transportation (free vaccine). • Self-referent emotions mediate message framing and vaccination intentions.
Liu, 2018 [41]	Intention: a 7-point scale with 3 items	Time orientation	• The loss-framed message produces high vaccination intention, but the difference is not significant. • The interaction effect between loss-framed messages and time orientation cannot predict vaccination intention.
Liu, 2018 [42]	Intention: a 7-point scale with 3 items	Cultural worldview (English or Chinese)	• Message framing does not have a major impact on vaccination intentions. • An individualistic worldview was positively associated with Chinese participants' willingness to be vaccinated and negatively associated with US participants' willingness to be vaccinated.
Kim, 2020 [43]	Intention: a 7-point scale with 3 items; attitude: a 7-point scale with 5 semantic differential items	Anticipated regret, EFT[d] (future-thinking, past-thinking, or no-thinking)	• The direct effect of message framing on attitudes and intentions toward HPV vaccination is not significant. • Anticipated regret has an indirect effect in message framing and HPV vaccination attitudes and intentions. • EFT and message framing interact on attitudes toward HPV vaccination. • Future thinking produces more favorable attitudes than past thinking in the gain-framed message.
Richards, 2021 [44]	Intention: a 5-point scale with 3 items	Perceived efficacy	• For parents with low perceived efficacy, the loss frame (compared with the gain frame) significantly reduced the intention to vaccinate.
Huang, 2023 [45]	Intention: a 7-point scale with 7 items	Reference point (self, other, or self-other)	• Message framing has no moderating effect between reference points and behavioral intentions.

[a]HPV: human papillomavirus.

[b]PBC: perceived behavioral control.

[c]SNSs: social networking sites.

[d]EFT: episodic future thinking.

Moderator or Mediator Variables of Message Framing

Each of the included studies explored moderating or mediating factors that affect the efficacy of message framing. A study by Fahy and Desmond [30] observed the mediating role between the theory of planned behavior and message framing and discovered that attitudes, normative beliefs, and perceived behavioral control explained 69.5% of the variation in vaccination intentions. The mediating effects of personal perceptions between message framing and intentions or behaviors were examined by 5 studies [28,34,36,37,44]. Gerend et al [28] demonstrated that there was no mediating effect between perceived barriers or perceived severity and message framing; although perceived susceptibility and self-efficacy were able to predict vaccination intentions, perceived benefit did not. Perceived susceptibility, perceived safety, and vaccine cost can predict vaccination rates, whereas perceived severity and benefit do not [34]. Participants with a high risk perception had a strong intention to be vaccinated [36]. Nan et al [37] discovered that loss framing produced more vaccination

XSL•FO
RenderX

intentions than gain framing when perceived susceptibility was low, while when perceived susceptibility was high, gain framing produced more vaccination intentions than loss framing. For parents with low perceived efficacy, loss framing significantly reduced intentions to vaccinate [44]. Kim et al [40] reported that message framing had an indirect effect on vaccination intentions mediated by transportation, self-inferred emotions, and anticipated regret. In addition, there was no impact of gender on vaccination attitudes [33], and prior knowledge had a major effect on behavioral intentions [38].

Nan [35] and Gerend et al [28] discovered that loss framing was associated with higher intentions to vaccinate than gain framing among participants with high avoidance motivation, which also applied to women engaging in risky sexual behaviors [27]. Nan [35] investigated the interaction between message framing and temporal orientation and found a marked interaction between them. The 2 frames worked equivalently in future-minded participants, and participants with a future-minded perspective possessed more intentions and attitudes toward vaccination. For present-minded participants, loss framing resulted in more vaccination attitudes than gain framing. By contrast, another study identified an interaction between loss framing and temporal orientation, which did not predict intentions to vaccinate [41]. Lechuga et al [31] compared the effects of message framing on different racial populations and found that, for African Americans and Hispanics, loss framing produced higher vaccination intentions, whereas for non-Hispanic Whites, there was no difference between the 2 types of message framing. Liu et al [42] identified no remarkable interaction between cultural worldviews and message framing; however, after categorizing by population, a positive interaction was found between individualistic worldviews and vaccination intentions of Chinese recipients and a negative interaction between individualistic worldviews and vaccination intentions of American recipients.

The interaction between message framing and behavioral frequency was explored by Gerend et al [28]. They found that loss framing contributed to higher vaccination intentions than gain framing in cases of 1-shot vaccination; however, there was no difference between the 2 message frames for the 6 doses required. Another study used colored primers to test the effect of message frames labeled red or gray on participants' HPV vaccination intentions. The finding that people who read loss-framed messages with red rectangular labels had higher vaccination intentions than those who read gain-framed messages (with noncolored labels) suggests that the combination of loss framing and a red label may have had a stronger influence on vaccination intentions than gain framing alone [29]. After conducting an intervention with different transmission media, Lee and Cho [39] found that participants expressed higher levels of behavioral intention for HPV vaccination when viewing loss-framed messages posted on social networking media. Kim et al [40] found an interaction between episodic future thinking and message framing, with future-thinking messages producing more favorable attitudes than past-thinking messages in gain-framed messages and no differences between future and past messages in the loss-framed situation on vaccination attitudes. In contrast, Wen and Shen [38] discovered that message framing and temporal distance interacted insignificantly, and there were no significant differences between loss-framed, gain-framed, future-framed, and present-framed messages on vaccination intentions. Huang and Li [45] added information specific to different vaccination targets in message frames and showed that no moderating effect was found between different reference points of message frames and vaccination intentions.

Discussion

Principal Findings

The construction of message framing can effectively affect the public's health attitudes and increase their intentions to practice healthy activities [46]. Therefore, we conducted a systematic review of 19 studies on the persuasive effects of message framing on HPV vaccination and summarized the effects of framing-based message stimulation approaches on participants' attitudes, intentions, or behaviors toward HPV vaccination. Gain-loss framing had no clear effect on vaccination attitudes or intentions. Loss framing showed a weak advantage in improving HPV vaccination attitudes or intentions, but the evidence was not strong enough to draw definitive conclusions. The gain-loss framing could combine various message stimulus contents or modalities to affect HPV vaccination attitudes, intentions, or behaviors distinctively. The impact of gain-loss framing on HPV vaccination behaviors could not be determined due to the limited number of studies and the qualitative nature of the analysis.

Previous reviews have established the advantage of loss framing for improving vaccination attitudes and intentions [47]. However, they did not specifically address the unique context of HPV vaccination. Additionally, we included recent studies published after the last major meta-analysis, ensuring that our review reflects the most up-to-date evidence. Although the direct effect of framing on attitudes and intentions could not be quantitatively assessed due to the limitations of our analysis, our review provides nuanced insights and highlights potential gaps in the existing knowledge. By synthesizing the available qualitative data, we contribute to a more complete understanding of the role of framing in vaccination promotion.

Our study showed that loss framing was more effective than gain framing at improving HPV vaccination attitudes and intentions. However, the distinction in persuasion effects between gain framing and loss framing is not as simple as we assumed. Earlier research on gain-loss framing did not uniformly and clearly present which one was more effective. Rothman and Salovey [48] conducted a literature review of a large number of studies, combing the content of previous studies, and proposed that the effectiveness of message framing must be focused on the specific scenarios of the study design. The direction and strength that specific framing imposes on intentions to adopt healthy behaviors may have diverse or conflicting results across studies depending on the scene setting, researcher manipulation, measurements, and individual traits like educational level and age of the participants. Some studies have suggested no significant difference between gain framing, which emphasizes the benefits of vaccination, and loss framing,

XSL•FO
RenderX

which focuses on the potential risks of not vaccinating, in terms of motivating attitudes or intentions to vaccinate [31,38,41], and others argued that loss framing was more persuasive [27-29,32,35-37,39]. The conflicting findings explained part of our results, although the studies supporting the dominance of loss framing represented more than one-half of the literature, with 11 of the 19 included publications (57.9%) supporting this finding. However, any statistical differences remain uncertain due to the absence of quantitative data integration.

In this systematic review, there was no direct effect of gain-loss framing on vaccination attitudes nor intentions, which supported the results of previous studies [14,49]. All literature included in this systematic review explored moderating or mediating variables between gain-loss framing and vaccination attitudes, intentions, or behaviors. With loss framing, messages combining individualistic worldviews, vaccination frequency, and transmission via social media or labeling message stimulus elements in red could enhance vaccination intentions of participants [28,29,39,42]. Meanwhile, loss framing messages produced higher vaccination intentions for high avoidance motivators, women engaged in risky sexual behaviors, African Americans and Hispanics, present-minded subjects, and low perceived susceptibility participants [27,31,32,35,37]. In addition, attitudes, normative beliefs, perceived behavioral control, perceived susceptibility, self-efficacy, perceived safety, perceived high risk, transportation, self-inferred emotions, and anticipated regret mediated vaccination intentions with gain-loss framing. Earlier opinions have pointed out that a single framing to achieve the expected persuasive effect or to explain altered attitudes and behavioral intentions is not convincing enough. In reality, the architecture of segment information is complex and multifaceted, comprising various message frames that embody diverse design ideas. Conducting research on multiple message frames, therefore, represents a crucial direction for the advancement of message framing theory and practice [50]. Additional high-quality RCTs are needed to verify the accuracy of this result due to the amount of literature that discusses the same variables.

Notably, only 1 study discussed the impact of gain-loss framing on vaccination behavior. Therefore, we were not able to draw conclusions related to the effects of gain-loss framing on vaccination behavior, which provides inspiration for future investigations. Attitudes or intentions are predictors of behavior, and the examination of them is clinically relevant; nevertheless, the ultimate purpose of structuring messages is to instruct the recipients to behave in a particular way [51]. We ought to go beyond the predictors; changing vaccination behaviors and increasing HPV vaccination rates are the objective of implementing interventions.

Limitations

There are several limitations in this systematic review. First, the absence of ongoing studies and only including documents published in English could not guarantee the inclusion of all eligible literature for the systematic review. Second, RoB 2 was selected as the risk of bias evaluation tool, which was influenced by major subjective factors. In addition, assessment tools for outcome indicators and statistical methods varied among the included literature, with a high degree of heterogeneity that diminishes the accuracy of final conclusions. Only 1 study measured behavioral change postintervention, and the other 18 studies assessed outcomes immediately after treatment, without sufficient time for individuals to assimilate and deliberate messages. In addition, the study provided an overview of the influence of attitudes, intentions, or behaviors toward HPV vaccination, but it was unable to clarify the specific causal relations of the 3 dependent variables. Instead, it had to follow former studies that regarded vaccination intentions as the strongest psychological motivation for behavioral occurrences. Due to these limitations, the results should be treated carefully.

Conclusion

This systematic review suggests that loss-framing messages show promise for boosting vaccination attitudes and intentions more effectively than gain-loss framing. The latter did not consistently demonstrate statistically significant advantages. The influence of gain-loss framing on vaccination perspectives varies significantly among diverse ethnic populations, underscoring the importance of cultural considerations in messaging strategies. Integrating gain-loss framing with alternative communication approaches or delivery platforms produces a spectrum of outcomes on vaccine attitudes, intentions, and behaviors, highlighting the potential for tailored interventions. To enhance the efficacy of message-based interventions for vaccination promotion, a broader perspective is imperative. This includes targeting audiences across all age groups, educational backgrounds, socioeconomic strata, and the digital divide while also considering individual media preferences and online health literacy levels. Longitudinal studies are necessary to ascertain the sustained impact of message framing on vaccination behaviors, thereby enriching our understanding of framing effects over time. Future research should prioritize behavioral interventions accompanied by objective outcome measurements, fortifying the evidence base for the strategic application of message framing in public health campaigns.

Acknowledgments

JG acknowledges financial support from the Nantong Health Commission, Jiangsu Province (grant MS2023003).

Authors' Contributions

JG and DG conceived and designed the study. JG and DG collected and analyzed the data. JG, DG, and WS performed the quality assessment and wrote the manuscript. HY and JX revised the manuscript and are responsible for the overall content. All authors reviewed and approved the submission of the final version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 108 KB - jmir_v26i1e52738_app1.pdf]

References

1. Ferrall L, Lin KY, Roden RBS, Hung C, Wu T. Cervical cancer immunotherapy: facts and hopes. Clin Cancer Res 2021 Oct 15;27(18):4953-4973 [FREE Full text] [doi: 10.1158/1078-0432.CCR-20-2833] [Medline: 33888488]
2. Padavu S, Aichpure P, Krishna Kumar B, Kumar A, Ratho R, Sonkusare S, et al. An insight into clinical and laboratory detections for screening and diagnosis of cervical cancer. Expert Rev Mol Diagn 2023 Jan;23(1):29-40. [doi: 10.1080/14737159.2023.2173580] [Medline: 36697264]
3. Pinkiewicz M, Dorobisz K, Zatoński T. Human papillomavirus-associated head and neck cancers. Where are we now? A systematic review. Cancer Manag Res 2022;14:3313-3324 [FREE Full text] [doi: 10.2147/CMAR.S379173] [Medline: 36465708]
4. Cheng L, Wang Y, Du J. Human papillomavirus vaccines: an updated review. Vaccines (Basel) 2020 Jul 16;8(3):1 [FREE Full text] [doi: 10.3390/vaccines8030391] [Medline: 32708759]
5. Esagoff A, Cohen SA, Chang G, Equils O, Van Orman S. Human papillomavirus and Chinese international students in the United States: attitudes, knowledge, vaccination trends, healthcare behaviors, and sexual activity. Hum Vaccin Immunother 2022 Dec 31;18(1):1882283 [FREE Full text] [doi: 10.1080/21645515.2021.1882283] [Medline: 33705223]
6. Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health 2020 Mar;8(2):e191-e203 [FREE Full text] [doi: 10.1016/S2214-109X(19)30482-6] [Medline: 31812369]
7. Buccafusco C, Hemel DJ. Framing vaccine mandates: messenger and message effects. J Law Biosci 2022;9(1):lsac016 [FREE Full text] [doi: 10.1093/jlb/lsac016] [Medline: 35769940]
8. Gursoy D, Ekinci Y, Can AS, Murray JC. Effectiveness of message framing in changing COVID-19 vaccination intentions: Moderating role of travel desire. Tour Manag 2022 Jul;90:104468 [FREE Full text] [doi: 10.1016/j.tourman.2021.104468] [Medline: 34898791]
9. Fridman I, Fagerlin A, Scherr KA, Scherer LD, Huffstetler H, Ubel PA. Gain-loss framing and patients' decisions: a linguistic examination of information framing in physician-patient conversations. J Behav Med 2021 Mar;44(1):38-52. [doi: 10.1007/s10865-020-00171-0] [Medline: 32725580]
10. Litovsky Y, Loewenstein G, Horn S, Olivola CY. Loss aversion, the endowment effect, and gain-loss framing shape preferences for noninstrumental information. Proc Natl Acad Sci U S A 2022 Aug 23;119(34):e2202700119 [FREE Full text] [doi: 10.1073/pnas.2202700119] [Medline: 35972966]
11. Steffen J, Cheng J. The influence of gain-loss framing and its interaction with political ideology on social distancing and mask wearing compliance during the COVID-19 pandemic. Curr Psychol 2023;42(10):8028-8038 [FREE Full text] [doi: 10.1007/s12144-021-02148-x] [Medline: 34341651]
12. Gao R, Guo H, Liu Y, Pang Y, Zhang X, Lian X, et al. The effects of message framing on self-management behavior among people with type 2 diabetes: A randomized controlled trial. Int J Nurs Stud 2023 Jul;142:104491. [doi: 10.1016/j.ijnurstu.2023.104491] [Medline: 37059032]
13. Wang K, Wong EL, Cheung AW, Dong D, Yeoh E. Loss-framing of information and pre-vaccination consultation improve COVID-19 vaccine acceptance: A survey experiment. Front Public Health 2023;11:1063444 [FREE Full text] [doi: 10.3389/fpubh.2023.1063444] [Medline: 36761145]
14. Prakash A, Jeyakumar Nathan R, Kini S, Victor V. Message framing and COVID-19 vaccine acceptance among millennials in South India. PLoS One 2022;17(7):e0269487 [FREE Full text] [doi: 10.1371/journal.pone.0269487] [Medline: 35802760]
15. Lucas T, Thompson HS, Blessman J, Dawadi A, Drolet CE, Hirko KA, et al. Effects of culturally targeted message framing on colorectal cancer screening among African Americans. Health Psychol 2021 May;40(5):305-315 [FREE Full text] [doi: 10.1037/hea0001073] [Medline: 34152784]
16. Akther T, Nur T. A model of factors influencing COVID-19 vaccine acceptance: A synthesis of the theory of reasoned action, conspiracy theory belief, awareness, perceived usefulness, and perceived ease of use. PLoS One 2022 Jan 12;17(1):e0261869 [FREE Full text] [doi: 10.1371/journal.pone.0261869] [Medline: 35020764]
17. Ezati Rad R, Kahnouji K, Mohseni S, Shahabi N, Noruziyan F, Farshidi H, et al. Predicting the COVID-19 vaccine receive intention based on the theory of reasoned action in the south of Iran. BMC Public Health 2022 Mar 04;22(1):229 [FREE Full text] [doi: 10.1186/s12889-022-12517-1] [Medline: 35120486]
18. Orr MG, Thrush R, Plaut DC. The theory of reasoned action as parallel constraint satisfaction: towards a dynamic computational model of health behavior. PLoS One 2013;8(5):e62490 [FREE Full text] [doi: 10.1371/journal.pone.0062490] [Medline: 23671603]

XSL•FO
RenderX

19. Ainiwaer A, Zhang S, Ainiwaer X, Ma F. Effects of message framing on cancer prevention and detection behaviors, intentions, and attitudes: systematic review and meta-analysis. J Med Internet Res 2021 Oct 16;23(9):e27634 [FREE Full text] [doi: 10.2196/27634] [Medline: 34528887]
20. Lee AY, Aaker JL. Bringing the frame into focus: the influence of regulatory fit on processing fluency and persuasion. J Pers Soc Psychol 2004 Mar;86(2):205-218. [doi: 10.1037/0022-3514.86.2.205] [Medline: 14769079]
21. O'Keefe DJ, Nan X. The relative persuasiveness of gain- and loss-framed messages for promoting vaccination: a meta-analytic review. Health Commun 2012;27(8):776-783. [doi: 10.1080/10410236.2011.640974] [Medline: 22292904]
22. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
23. Bai Y, Ip P, Chan K, Ngan H, Yip P. HPV vaccination intentions of female students in Chinese universities: a systematic literature review and meta-analysis. Int J Environ Res Public Health 2022 Aug 17;19(16):1 [FREE Full text] [doi: 10.3390/ijerph191610207] [Medline: 36011838]
24. Markowitz LE, Dunne EF, Saraiya M, Chesson HW, Curtis CR, Gee J, Centers for Disease Control and Prevention (CDC). Human papillomavirus vaccination: recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep 2014 Aug 29;63(RR-05):1-30 [FREE Full text] [Medline: 25167164]
25. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
26. Ryan R, Hill S, Prictor M, McKenzie J. Cochrane Consumers and Communication group Study Quality Guide. Cochrange. Melbourne, Australia: La Trobe University; 2013. URL: https://tinyurl.com/ps469mxf [accessed 2024-10-31]
27. Gerend MA, Shepherd JE. Using message framing to promote acceptance of the human papillomavirus vaccine. Health Psychol 2007 Dec;26(6):745-752. [doi: 10.1037/0278-6133.26.6.745] [Medline: 18020847]
28. Gerend MA, Shepherd JE, Monday KA. Behavioral frequency moderates the effects of message framing on HPV vaccine acceptability. Ann Behav Med 2008 May 12;35(2):221-229. [doi: 10.1007/s12160-008-9024-0] [Medline: 18347893]
29. Gerend MA, Sias T. Message framing and color priming: How subtle threat cues affect persuasion. Journal of Experimental Social Psychology 2009 Jul;45(4):999-1002. [doi: 10.1016/j.jesp.2009.04.002]
30. Fahy A, Desmond DM. Irish mothers' intentions to have daughters receive the HPV vaccine. Ir J Med Sci 2010 Oct 29;179(3):427-430. [doi: 10.1007/s11845-010-0501-7] [Medline: 20585896]
31. Lechuga J, Swain GR, Weinhardt LS. Impact of framing on intentions to vaccinate daughters against HPV: a cross-cultural perspective. Ann Behav Med 2011 Oct 30;42(2):221-226 [FREE Full text] [doi: 10.1007/s12160-011-9273-1] [Medline: 21533624]
32. Nan X. Communicating to young adults about HPV vaccination: consideration of message framing, motivation, and gender. Health Commun 2012 Jan;27(1):10-18. [doi: 10.1080/10410236.2011.567447] [Medline: 22276999]
33. Gainforth HL, Cao W, Latimer-Cheung AE. Message framing and parents' intentions to have their children vaccinated against HPV. Public Health Nurs 2012 Dec;29(6):542-552. [doi: 10.1111/j.1525-1446.2012.01038.x] [Medline: 23078425]
34. Gerend MA, Shepherd JE. Predicting human papillomavirus vaccine uptake in young adult women: comparing the health belief model and theory of planned behavior. Ann Behav Med 2012 Oct 1;44(2):171-180 [FREE Full text] [doi: 10.1007/s12160-012-9366-5] [Medline: 22547155]
35. Nan X. Relative persuasiveness of gain- versus loss-framed human papillomavirus vaccination messages for the present- and future-minded. Human Communication Research 2012;38(1):72-94. [doi: 10.1111/j.1468-2958.2011.01419.x]
36. Park S. The effects of message framing and risk perceptions for HPV vaccine campaigns: focus on the role of regulatory fit. Health Mark Q 2012 Oct;29(4):283-302. [doi: 10.1080/07359683.2012.732847] [Medline: 23210670]
37. Nan X, Madden K, Richards A, Holt C, Wang MQ, Tracy K. Message framing, perceived susceptibility, and intentions to vaccinate children against HPV among African American parents. Health Commun 2016 Jul;31(7):798-805 [FREE Full text] [doi: 10.1080/10410236.2015.1005280] [Medline: 26646190]
38. Wen N, Shen F. Communicating to young Chinese about human papillomavirus vaccination: examining the impact of message framing and temporal distance. Asian Journal of Communication 2016 Mar 23;26(4):387-404. [doi: 10.1080/01292986.2016.1162821]
39. Lee MJ, Cho J. Promoting HPV vaccination online: message design and media choice. Health Promot Pract 2017 Oct 27;18(5):645-653. [doi: 10.1177/1524839916688229] [Medline: 28129691]
40. Kim HK, Lee TK, Kong WY. The interplay between framing and regulatory focus in processing narratives about HPV vaccination in Singapore. Health Commun 2020 Mar 10;35(2):222-232. [doi: 10.1080/10410236.2018.1553022] [Medline: 30526080]
41. Liu S, Yang JZ, Chu H. Now or future? Analyzing the effects of message frame and format in motivating Chinese females to get HPV vaccines for their children. Patient Educ Couns 2019 Jan;102(1):61-67. [doi: 10.1016/j.pec.2018.09.005] [Medline: 30219633]
42. Liu S, Yang JZ, Chu H, Sun S, Li H. Different culture or different mind? perception and acceptance of HPV vaccine in China and in the U.S. J Health Commun 2018 Oct 22;23(12):1008-1016. [doi: 10.1080/10810730.2018.1536729] [Medline: 30346889]

XSL•FO
RenderX

43. Kim HM, Kim E, Murphy S. Testing the effectiveness of message framing and episodic future thinking in promoting HPV vaccination via anticipated regret. Health Commun 2022 May 08;37(5):525-534. [doi: 10.1080/10410236.2020.1855744] [Medline: 33289424]
44. Richards AS, Qin Y, Daily K, Nan X. African American parents' perceived vaccine efficacy moderates the effect of message framing on psychological reactance to HPV vaccine advocacy. J Health Commun 2021 Aug 03;26(8):534-544 [FREE Full text] [doi: 10.1080/10810730.2021.1966688] [Medline: 34427548]
45. Huang G, Li K. Get vaccinated for loved ones: effects of self-other appeal and message framing in promoting HPV vaccination among heterosexual young men. Health Commun 2023 Mar 16;38(2):381-393. [doi: 10.1080/10410236.2021.1953728] [Medline: 34266337]
46. Park J, Kim SH, Kim JG. Effects of message framing and health literacy on intention to perform diabetes self-care: A randomized controlled trial. Diabetes Res Clin Pract 2020 Mar;161:108043. [doi: 10.1016/j.diabres.2020.108043] [Medline: 32006642]
47. Penţa MA, Băban A. Message framing in vaccine communication: a systematic review of published literature. Health Commun 2018 Mar 06;33(3):299-314. [doi: 10.1080/10410236.2016.1266574] [Medline: 28059557]
48. Rothman AJ, Salovey P. Shaping perceptions to motivate healthy behavior: the role of message framing. Psychol Bull 1997 Jan;121(1):3-19. [doi: 10.1037/0033-2909.121.1.3] [Medline: 9000890]
49. Wang J, Gu Y, Xin H, Wang X. Influence of appeal type and message framing on residents' intent to engage in pro-environmental behavior. Int J Environ Res Public Health 2022 Dec 22;19(23):1 [FREE Full text] [doi: 10.3390/ijerph192315431] [Medline: 36497506]
50. Lee AM, Hojjatinia S, Courtney JB, Brunke-Reese D, Hojjatinia S, Lagoa CM, et al. Motivational message framing effects on physical activity dynamics in a digital messaging intervention: secondary analysis. JMIR Form Res 2023 May 21;7:e41414 [FREE Full text] [doi: 10.2196/41414] [Medline: 37083710]
51. Waters EA, Ackermann N, Wheeler CS. Specifying future behavior when assessing risk perceptions: implications for measurement and theory. Med Decis Making 2019 Dec;39(8):986-997 [FREE Full text] [doi: 10.1177/0272989X19879704] [Medline: 31646937]

Abbreviations

HPV: human papillomavirus
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB: risk of bias

Edited by T de Azevedo Cardoso; submitted 13.09.23; peer-reviewed by R Gao, H Guo, SS Shaik, R Mpofu; comments to author 04.06.24; revised version received 26.07.24; accepted 21.10.24; published 07.11.24.

Please cite as:
Gong J, Gu D, Dong S, Shen W, Yan H, Xie J
Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review
J Med Internet Res 2024;26:e52738
URL: https://www.jmir.org/2024/1/e52738
doi:10.2196/52738
PMID:

©Jie Gong, Dandan Gu, Suyun Dong, Wangqin Shen, Haiou Yan, Juan Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Exploration of Features of Mobile Applications for Medication Adherence in Asia: Narrative Review

Tzu Wang[1], PharmD; Yen-Ming Huang[1,2,3*], BPharm, PhD; Hsun-Yu Chan[4*], PhD

[1]School of Pharmacy, College of Medicine, National Taiwan University, Taipei City, Taiwan

[2]Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei City, Taiwan

[3]Department of Pharmacy, National Taiwan University Hospital, Taipei City, Taiwan

[4]Department of Industrial Education, National Taiwan Normal University, Taiepi City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Yen-Ming Huang, BPharm, PhD
Graduate Institute of Clinical Pharmacy
College of Medicine
National Taiwan University
No.33, Linsen S Rd
Zhongzheng Dist
Taipei City, 100025
Taiwan
Phone: 886 33668784
Email: yenming927@ntu.edu.tw

Abstract

Background: Medication is crucial for managing chronic diseases, yet adherence rates are often suboptimal. With advanced integration of IT and mobile internet into health care, mobile apps present a substantial opportunity for improving adherence by incorporating personalized educational, behavioral, and organizational strategies. However, determining the most effective features and functionalities for these apps within the specific health care context in Asia remains a challenge.

Objective: We aimed to review the existing literature, focusing on Asian countries, to identify the optimal features of mobile apps that can effectively enhance medication adherence within the unique context of Asian societies.

Methods: We conducted a narrative review with the SPIDER (sample, phenomenon of interest, design, evaluation, research type) tool. We identified studies on mobile apps for medication adherence from January 2019 to August 2024 on PubMed and Scopus. Key search terms included “Asia,” “chronic disease,” “app,” “application,” “survey,” “experiment,” “questionnaire,” “group,” “medical adherence,” “medication adherence,” “case-control,” “cohort study,” “randomized controlled trial,” “clinical trial,” “observational study,” “qualitative research,” “mixed methods,” and “analysis,” combined using logical operators “OR” and “AND.” The features of mobile apps identified in the studies were evaluated, compared, and summarized based on their disease focuses, developers, target users, features, usability, and use.

Results: The study identified 14 mobile apps designed to enhance medication adherence. Of these, 11 were developed by research teams, while 3 were created by commercial companies or hospitals. All the apps incorporated multiple features to support adherence, with reminders being the most common, present in 11 apps. Patient community forums were the least common, appearing in only 1 app. In total, 6 apps provided lifestyle modification functions, offering dietary and exercise recommendations, generating individualized plans, and monitoring progress. In addition, 6 apps featured health data recording and monitoring functions, with 4 allowing users to export and share records with researchers or health care professionals. Many apps included communication features, with 10 enabling feedback from researchers or health care professionals and 7 offering web-based consultation services. Educational content was available in 8 apps, and 7 used motivation strategies to encourage adherence. Six studies showed that mobile apps improved clinical outcomes, such as blood glucose, lipid, and pressure, while reducing adverse events and boosting physical activities. Twelve studies noted positive humanistic effects, including better medication adherence, quality of life, and user satisfaction.

Conclusions: This review has identified key components integrated into mobile apps to support medication adherence. However, the lack of government and corporate involvement in their development limits the generalizability of any individual app. Beyond basic reminder functions, features such as multiuser support, feedback mechanisms, web-based consultations, motivational tools,

XSL•FO
RenderX

and socialization features hold significant promise for improving medication adherence. Further pragmatic research is necessary to validate the effectiveness of these selected apps in enhancing adherence.

(J Med Internet Res 2024;26:e60787) doi:10.2196/60787

KEYWORDS

Asia; adherence; application; feature; medication; mobile

Introduction

Medication adherence refers to how consistently individuals follow their prescribed medication regimen as advised by health care professionals [1]. While medication is powerful for the management of chronic diseases, the average adherence rate stands at 50% worldwide [2]. A meta-analysis revealed that the prevalence of nonadherence among patients taking antihypertensive medications in Asia is 48% [3]. Suboptimal medication adherence is associated with a range of negative consequences, including reduced treatment outcomes, unfavorable disease prognosis, diminished quality of life, worse complications, and frequent hospitalization [4]. Research suggests that patients decide to take medications based on perceived necessity, effectiveness, and safety, and maintain adherence through accessibility, routine, and memory [5]. Both intentional and unintentional nonadherence may occur, where patients may miss doses accidentally or deliberately alter their regimen without counseling their health care providers [6]. Such behaviors contribute significantly to medication-related hospitalizations and preventable deaths, accounting for US $500 billion (16%) of the entire US health care expenditure annually [7].

Efforts to enhance medication adherence involve strategies across behavioral, educational, and organizational domains [8]. Behavioral interventions focus on adapting patients' surroundings to promote medication adherence, using tools like multidose packs and mapping medication intake to routine [5]. Educational approaches stress the significance of medication in disease management, equipping patients with comprehensive knowledge to empower informed decision-making about their health [9]. Organizational methods streamline medication management through initiatives such as medication synchronization and team-based care, aiming to improve access to medication refill and enhanced adherence [10]. Driven by the advanced integration of IT and mobile internet in health care practice, mobile apps have emerged as a promising avenue for improving adherence by incorporating these strategies and tailoring them to individual needs [11]. Accessible through platforms such as the Apple App Store and Google Play Store, these apps may include reminders, educational resources, and organizational features to support medication adherence. Given the widespread ownership of mobile phones and the tendency of individuals to frequently check their devices, mobile apps offer a convenient and pervasive solution for medication adherence support [12]. Furthermore, they hold substantial potential to enhance patient care and reduce health care costs, particularly considering the significant portion of the population that requires reminders to adhere to their medication regimen [11].

Amidst the onset of the COVID-19 pandemic, digital health interventions such as telehealth and application-based health care emerged in response to the limitations of in-person interventions [13]. The pandemic significantly disrupted the global health system, leading to constrained access to health care services due to surging patient numbers and staff shortages, particularly affecting individuals with chronic diseases who require extensive medical attention [14]. With over half of the global population, approximately 4.3 billion people, now owning smartphones and two-thirds worldwide using the internet, according to the Groupe Speciale Mobile Association's State of Mobile Internet Connectivity Report 2023, the prevalence of digital connectivity has soared [15]. Research indicates that more than half of mobile phone users have downloaded health-related mobile apps, and about 7 in 10 individuals with such apps used them to track progress on health-related goals in 2020 [16]. Consequently, digital health interventions for chronic disease management became imperative and have demonstrated efficacy in various studies [17,18].

In 2018, Ahmed et al [19] discovered 420 adherence apps available on the Apple App Store and the Google Play Store. Their findings suggest that mobile apps could serve as a potential strategy to tackle medication nonadherence. Haase et al [10] reviewed 30 apps focused on medication adherence and selected 5 with the most ideal features based on the Application Score Card. However, these evaluations overlooked the diverse contexts of health care systems, which are pivotal factors influencing patients' medication refill and intake behaviors. Asian and western countries demonstrate notable distinctions in their health care insurance systems and cultural perspectives on medication. For instance, Japan, South Korea, and Taiwan operate universal social health insurance systems that cover all citizens and residents [20]. In contrast, the United States primarily relies on private insurance systems, leading to disparities in access to medication [21]. In East Asian countries (eg, Japan, South Korea, and Taiwan), patients can easily obtain medical services and medications at lower costs compared with patients in the United States, who often encounter high out-of-pocket expenses for medications due to the expensive nature of health care [21,22]. In addition, patients with an Asian cultural background tend to hold more negative perceptions about medication. They often view medications as inherently harmful or addictive substances to be avoided [23,24]. Consequently, patients in Asian countries may face fewer barriers to filling prescriptions and refills but may harbor more doubts regarding medication intake. These cultural and systematic nuances should be considered when designing interventions or recommending adherence apps to ensure their effectiveness across diverse patient populations [25].

XSL•FO
RenderX

Over the past few years, there has been a surge in the development of apps targeting medication adherence improvement. However, determining the most effective features and functionalities for these apps remains uncertain. This study was to review the existing literature, with a particular focus on the Asian countries, to identify the optimal features of mobile apps that can effectively enhance medication adherence within the unique context of Asian societies.

Methods

We used the SPIDER (sample, phenomenon of interest, design, evaluation, research type) framework to craft the research question [26]. So, the research question was "What are the optimal features of mobile apps that are used to improve medication adherence in health care practice in Asian countries?" We followed the principles of the scale for the quality assessment of narrative review articles in conducting this narrative review [27]. Using a qualitative approach, we compared and summarized the features of the selected apps. This involved using an app quality assessment tool, developing a search strategy, applying predefined inclusion criteria for screening apps, comparing and selecting apps, and extracting relevant data.

Search Strategy

We conducted a literature search on 2 major electronic databases, PubMed and Scopus, to identify studies on the use of mobile apps for medication adherence. The SPIDER framework was used to structure the research question and guide the literature search (Multimedia Appendix 1). The key search terms included "Asia," "chronic disease," "app," "application," "survey," "experiment," "questionnaire," "group," "medical adherence," "medication adherence," "case-control," "cohort study," "randomized controlled trial," "clinical trial," "observational study," "qualitative research," "mixed methods," and "analysis," within the timeframe from January 2019 to August 2024. These terms were combined using logical operators "OR" and "AND" and were restricted to searches within titles and abstracts only (Multimedia Appendix 2).

Study Selection

The initial screening of titles and abstracts was conducted by the first author [TW]. We used Endnote to detect and delete duplicates, followed by a manual verification by the corresponding author [YMH] to ensure no duplicates were overlooked. After the initial screening of titles and abstracts, a thorough full-text review of selected articles was conducted by 2 reviewers [TW and YMH], adhering to the inclusion and exclusion criteria outlined in Textbox 1. Any discrepancies between the reviewers were resolved through discussion, and if necessary, a third reviewer [HYC] was consulted to reach a consensus. In addition, hand-searching of grey literature was conducted to identify further relevant papers during the full-text review process.

Textbox 1. Inclusion and exclusion criteria of the selected studies.

Inclusion criteria

- Focusing on Asian adults (18 years or older) with at least 1 chronic disease (defined as conditions lasting 1 year or more, requiring ongoing medical attention or limiting activities of daily living or both).
- Using mobile apps to assist patients in self-management and improve patients' medication adherence.
- Providing descriptions of the features of the apps used in the study.
- Using an experimental design.
- Published in the recent 5 years.

Exclusion criteria

- A review article rather than primary research.
- Publications in languages other than English.
- Insufficient detailed information about the mobile app features used in the intervention.
- Not concerning the link between medication adherence and mobile apps.
- Focusing on diseases for which patients do not require long-term medications.

Data Extraction

Two reviewers [TW and YMH] performed the data extraction process. We jointly decided to extract data from the full-text articles using a pre-established form, which was developed based on literature reviews [10,19,28]. The form encompassed study characteristics, including author, country, and year of publication, as well as details regarding the developer of the mobile app, target population, features, usability, and use of the mobile app. In addition, we documented any additional information about app features not covered in our predefined form. These features of mobile apps across the studies would subsequently be evaluated and compared in our review.

Data Analysis

The findings are depicted in a narrative table (Table 1), evaluating which apps exhibited the most optimal features using a modified mobile app rating scale derived from literature reviews [10,19,28]. A 10-item evaluation sheet was used to determine the number of features captured by each selected app, with a score assigned based on the total number of items covered (Table 1). A higher score indicated that the mobile app possessed

XSL•FO
RenderX

more functions to enhance medication adherence. We summarized and juxtaposed the apps from the selected studies based on their respective disease focus, developers, target users, features, usability, and use. Consequently, we compared and delineated the similarities and disparities in the features of these apps, proposing the ideal features of available apps for enhancing medication adherence.

XSL•FO
RenderX

Table 1. Summary and comparison of the selected mobile apps for improving medication adherence.

Study	Disease focus	Developer type	Target user	Features	Number of app functions
Liu et al [29]	Cardiovascular diseases	Research	Health care professional and patient	• Education • Lifestyle modification (eg, diet, exercise) • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	4
Fan et al [30]	HIV	Research	Health care professional and patient	• Education • Feedback from health care professionals • Motivation • Web-based consultation with health care professionals • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	6
Chen et al [31]	Cardiovascular diseases	Research	Caregiver, health care professional, and patient	• Data record and monitoring • Data export and sharing • Education • Feedback from health care professionals • Lifestyle modification (eg, diet and exercise) • Motivation • Web-based consultation with health care professionals • Patient community forum • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	10
Sunil Kumar et al [32]	Diabetes	Research	Patient	• Data record and monitoring • Feedback from health care professionals • Lifestyle modification (eg, diet and exercise) • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	5
Bozorgi et al [33]	Cardiovascular diseases	Research	Caregiver, health care professional, and patient	• Data record and monitoring • Data export and sharing • Education • Feedback from health care professionals • Lifestyle modification (eg, diet and exercise) • Motivation • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	8
Poorcheraghi et al [34]	Unspecified chronic diseases	Research	Health care professional and patient	• Education • Feedback from health care professionals • Web-based consultation with health care professionals • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	5
Al-Nawayseh et al [35]	Asthma	Research	Patient	• Education • Others	2

XSL•FO
RenderX

Study	Disease focus	Developer type	Target user	Features	Number of app functions
Nurakysh et al [36]	Cardiovascular diseases	Hospital	Caregiver, health care professional, and patient	• Data record and monitoring • Education • Feedback from health care professionals • Lifestyle modification (eg, diet and exercise) • Motivation • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	7
Yang et al [37]	Diabetes	Commercial company	Health care professional and patient	• Data record and monitoring • Data export and sharing • Feedback from health care professionals • Lifestyle modification (eg, diet and exercise) • Motivation • Web-based consultation with health care professionals	6
Chew et al [38]	Unspecified chronic diseases	Research	Caregiver and patient	• Motivation • Web-based consultation with health care professionals • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	4
Park et al [39]	Rheumatoid arthritis	Research	Health care professional and patient	• Education • Web-based consultation with health care professionals • Reminder (eg, medication-taking, follow-up, and daily monitoring)	3
Huang et al [40]	Diabetes	Commercial company	Caregiver and patient	• Data record and monitoring • Data export and sharing • Feedback from health care professionals • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	5
Pang et al [41]	HIV	Research	Health care professional and patient	• Data export and sharing • Feedback from health care professionals • Motivation • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	5
Chen et al [42]	Schizophrenia	Research	Health care professional, patient	• Data export and sharing • Feedback from health care professionals • Motivation • Web-based consultation with health care professionals • Reminder (eg, medication-taking, follow-up, and daily monitoring) • Others	6

XSL•FO
RenderX

Results

Overview

We initially found 247 articles on PubMed and 150 on Scopus. After removing 64 duplicates, we were left with 333 studies. From this pool, 156 irrelevant articles were excluded based on title and abstract screening. Afterwards, 177 articles underwent full-text review for eligibility assessment. Among these, 167 articles were excluded for various reasons, 20 were not primary research, 68 were not conducted in Asia, 31 did not use mobile apps as an intervention or did not provide detailed information about the mobile apps, 24 did not pertain to the association between medication adherence and mobile apps, and 24 were not relevant to chronic diseases. Ultimately, 14 studies [29-42] were included in our review, with an additional 4 results [32,33,35,39] obtained from hand-searching, as illustrated in Figure 1.

Figure 1. Flowchart of the study selection process.

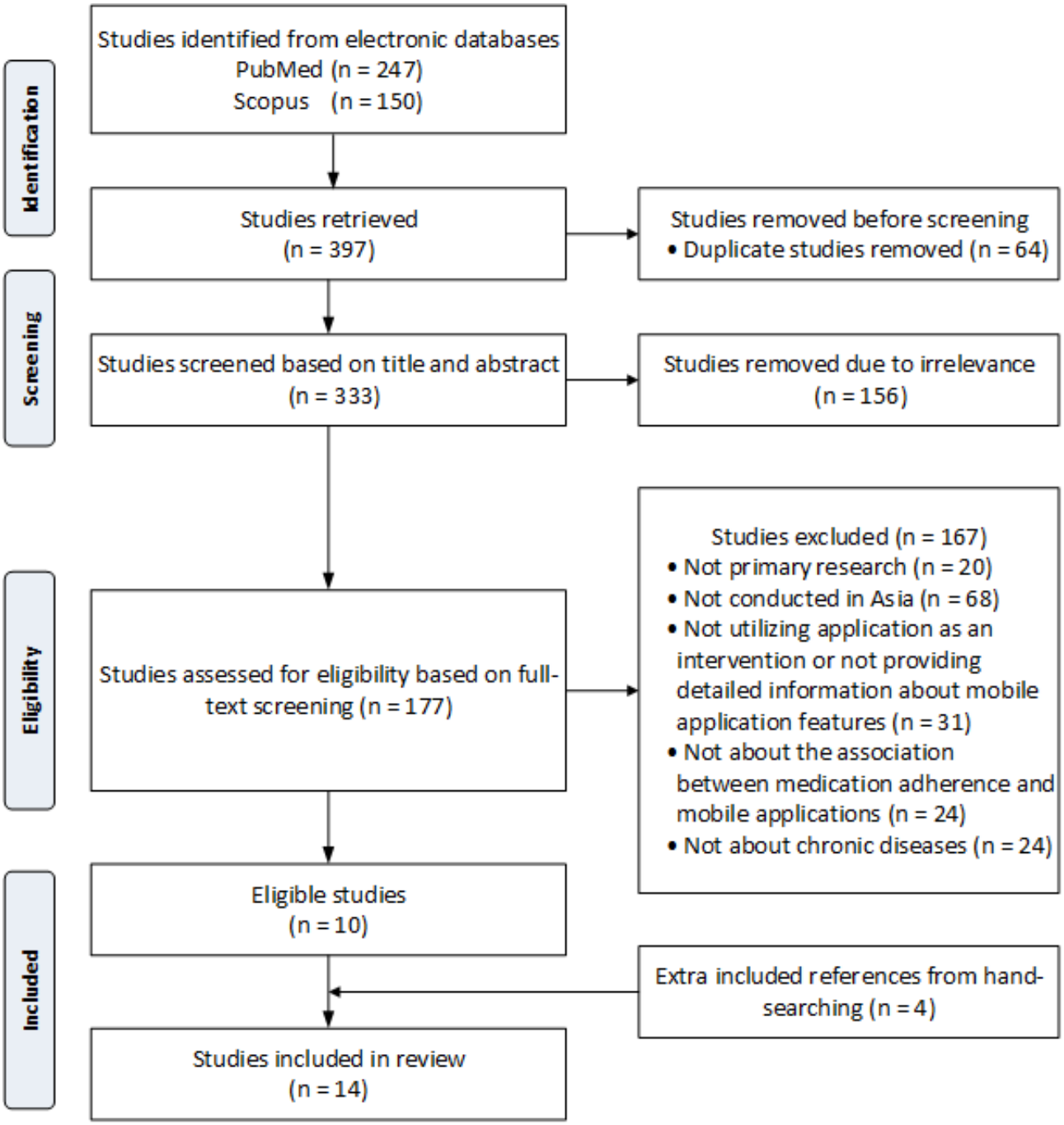

Characteristics of the Included Studies

The majority of the studies (9/14, 64.3%) were conducted in East and Southeast Asia [29-31,37-42], with the remaining studies conducted in India (1/14, 7.1%) [32], Iran (2/14, 14.3%) [33,34], Jordan (1/14, 7.1%) [35], and Kazakhstan (1/14, 7.1%) [36], respectively. All the studies investigated the use of mobile apps to improve medication adherence among patients with chronic diseases, defined as conditions lasting 1 year or more, requiring ongoing medical attention, or restricting activities of

XSL•FO
RenderX

daily living. Among these studies, 4 (28.6%) targeted patients with cardiovascular diseases [29,31,33,36], 3 (21.4%) focused on diabetes mellitus [32,37,40], 2 (14.3%) on HIV management [30,41], 1 (7.1%) on schizophrenia care [42], 1 (7.1%) on asthma treatment [35], 1 (7.1%) on rheumatoid arthritis [39], and 2 (14.3%) included patients with various chronic diseases and polypharmacy [34,38].

Features of the Mobile Apps

Among the mobile apps in these studies, 11 (78.6%) were developed by research teams [29-35,38,39,41,42], while 3 (21.4%) were existing apps developed by commercial companies or hospitals [36,37,40]. The features of the 14 mobile apps are summarized in Table 1, showing significant differences among them. All of the selected apps incorporated multiple features to improve medication adherence. Reminders were the most commonly used feature, while patient community forums were the least integrated feature in the app design.

In total, 12 (85.7%) apps featured a reminder function primarily aimed at reminding patients to take their medications as scheduled [29-36,38-40,42]. Some of these apps also provided reminders for follow-up appointments [30,31,33,35,40], prescription refills [40], or monitoring health status (eg, body weight, blood pressure, or blood sugar) [31,33,35,40]. Out of 14, 6 (42.9%) apps offered lifestyle modification functions by providing dietary and exercise recommendations and generating individualized plans and monitoring [29,31-33,36,37]. Furthermore, 6 (42.9%) apps included health data recording and monitoring functions (eg, body weight, blood pressure, and blood sugar) [29,31-33,36,37], with 4 (28.6%) allowing users to export and share these records with researchers or health care professionals [29,32,33,37]. Two (14.3%) other apps, while unable to record health data, allowed medication intake records to be uploaded and shared with researchers or health care professionals [31,36]. A higher proportion of the apps we reviewed had communication features with researchers or health care professionals. Specifically, 10 (71.4%) apps enabled users to receive feedback from researchers or health care professionals during their usage [30-37,40,42], and 7 (50.0%) apps offered web-based consultation services [30,31,34,37-39,42]. Furthermore, some apps took measures to reinforce patients' willingness to adhere to medications. Eight (57.1%) apps included educational content for patients [29-31,33-36,39], while 7 (50.0%) used motivation strategies to encourage adherence [31,33,35-38,42]. Only 1 app (7.1%) offered a patient community forum, allowing patients to share and discuss their experiences with medication and disease management with peers and caregivers in a safe space [31].

Several unique features were identified across the selected apps in this review (Multimedia Appendix 3). HeartGuardian (Liu et al) supported the management of lipid levels and improved medication adherence by tailoring detailed health plans for patients, recording medication intake, and providing real-time video feedback based on their adherence [29]. Medisafe app offered an easy-to-use reminder feature to help patients take their medications on time while also tracking blood pressure or blood sugar levels. It also assessed patients' health data and medication adherence and allowed caregivers to supervise medication intake [40]. Pang et al [41] created the Mobile Interactive Supervised Therapy (MIST) system to monitor medication adherence among people with HIV, requiring patients to upload videos of their medication intake for researchers to review. Chew et al [38] developed the Med Assist app, which included a 2-way alarm reminder system, allowing users to take medications later if they miss the scheduled time. It could also summarize all medications that patients had to take into a single page and help caregivers supervise medication usage. Furthermore, patients could use the Med Assist app to check medication availability at partnered pharmacies, making prescription refills more convenient. The Med Assist app also featured an adherence scoring system, incorporating gamification where users earned rewards for achieving app objectives [38]. DIAGURU, a mobile app created by Sunil Kumar et al for diabetes management, was capable of recording and monitoring patients' blood sugar levels, insulin usage, and food intake to generate visualized reports and graphs as feedback [32]. In addition, DIAGURU could calculate the caloric value of food intake and provide dietary recommendations based on patients' sugar levels and the caloric consumption [32]. Poorcheraghi et al [34] designed a medication management app targeting older adults, featuring user-friendliness, adjustable font and text sizes, and appropriate background and app item colors. The app recorded the names and images of medications and would audibly announce the medication names and show their pictures when playing the reminders [34]. Asthma mHealth, developed by Al-Nawayseh et al [35], focused on educational functions, providing patients with detailed introduction about asthma and using short videos to demonstrate correct inhaler usage.

Finally, only 5 studies (35.7%) included descriptions of privacy protection measures [30,31,34,39,40]. The most detailed information was provided by Chen et al [31], who implemented user authentication, data encryption for both transmission and storage, and strong password protection for the iCARE app. In addition, the iCARE featured backup and recovery functions and used MongoDB (one of open-source NoSQL databases) for its flexibility and security [31]. The app's development also adhered to international and national data standards [31].

Usability and Use of the Selected Mobile Apps

Although the majority of studies in our review were still in the early stages of app development and pilot testing, 12 (85.7%) reported improvements in clinical or humanistic outcomes [29,31-38,40-42]. As summarized in Table 2, these findings underscore the potential effectiveness of mobile apps in enhancing patient outcomes from clinical and humanistic perspectives.

XSL•FO
RenderX

Table 2. Outcome measures of the selected mobile apps for improving medication adherence.

Study	Clinical outcomes	Humanistic outcomes
Liu et al [29]	• Blood lipid • Medication-associated adverse events	• Medication adherence
Fan et al [30][a]	—[b]	—[b]
Chen et al [31]	—[b]	• Satisfaction • Usefulness
Sunil Kumar et al [32]	—[b]	• Quality of life
Bozorgi et al [33]	• Blood pressure • Physical activity	• Medication adherence • Satisfaction • Usefulness
Poorcheraghi et al [34]	• Medication-associated adverse events	• Medication adherence
Al-Nawayseh et al [35]	• Disease status	• Ease of use • Medication adherence • Satisfaction • Usefulness
Nurakysh et al [36]	—[b]	• Medication adherence
Yang et al [37]	• Blood glucose • Blood pressure	• Medication adherence • Satisfaction
Chew et al [38]	—[b]	• Ease of use • Satisfaction • Usefulness
Park et al [39][a]	—[b]	—[b]
Huang et al [40]	—[b]	• Ease of use • Medication adherence • Satisfaction • Usefulness
Pang et al [41]	• Disease status	• Ease of use • Usefulness
Chen et al [42]	—[b]	• Medication adherence

[a]The study protocol did not include findings on the effectiveness of apps in terms of patient outcomes.

[b]Not applicable.

Six (42.9%) studies highlighted that mobile apps lead to clinical improvement [29,33-35,37,42], including blood glucose, blood pressure, blood lipids, disease status, reduction in medication associated adverse events, and increased physical activities. In addition, 12 (85.7%) studies reported positive humanistic outcomes such as increased medication adherence, improved quality of life, and positive attitude toward the mobile apps, including satisfaction, usability, and ease of use [29,31-38,40-42]. Among these, 8 (57.1%) studies noted better medication adherence in intervention groups using mobile apps compared with control groups [29,33-37,40,42]. Furthermore, 7 (50.0%) studies included in our review assessed users' attitudes toward the mobile apps [31,33,35,37,38,40,41]. Of these, 6 (42.9%) reported positive user attitude and satisfaction with the apps [31,33,35,37,38,40], 6 (42.9%) considered the apps useful and beneficial [31,33,35,38,40,41], and 4 (28.6%) found the apps easy to use [35,38,40,41]. One (7.1%) study suggested that mobile apps could serve as a public health tool to enhance patients' quality of life [32].

Some studies also provided user feedback or detailed how the mobile apps were used. For instance, Chen et al [31] noted that the most frequently visited screens were those related to health recommendations, daily monitoring of health indicators, health behaviors, and medication adherence. Chew et al [38] found that features such as reminder functions, medication summary, and multiuser support were particularly helpful, though users faced challenges when adding new medications and suggested more in-app guidance. Huang et al [40] observed that while the app was useful for older users, they found reminder notifications somewhat annoying. Finally, Pang et al [41] reported that users appreciated features such as medication reminders, a color-coded calendar, and a reward system, though there were privacy

XSL•FO
RenderX

concerns about features that might inadvertently disclose users' HIV status.

Discussion

Principal Findings

This narrative review identified 14 mobile apps used to enhance medication adherence within the Asian context. The findings illustrate the varied features of existing mobile apps used to address medication adherence, aiming for improved health outcomes. Consistent with previous research on enhancing medication adherence, integrating a variety of reminder, behavioral, and educational strategies into intervention content has potential to enhance adherence behavior more efficiently and effectively [9].

Using reminder strategies is a key tactic to help patients adhere to their medication schedules, often implemented through alarms, push notifications, or short message services [19]. To ensure correct medication use, patients need to input or log their medication information into apps, such as the name, dosage, and frequency of their medications. This process requires a level of proficiency in using digital devices. Patients may benefit from user-friendly interfaces or additional assistance from family members or health care professionals to navigate these apps effectively. Furthermore, some apps remind patients of regular follow-up appointments and disease monitoring, providing additional cues to reinforce patient self-efficacy and mitigate forgetfulness, which is one of the most commonly reported reasons for nonadherence to chronic disease medications [4,43]. Therefore, integrating reminder functions for medication-taking with other aspects of disease management may result in improved medication adherence [34,41].

Behavior-focused strategies, such as prompts, social support, and rewards, are effective in encouraging patients to adopt new medication practices [9]. A higher proportion of feedback and web-based consultation features were observed in the mobile apps reviewed in this study. In contrast, many medication adherence apps available on the Apple App Store or Google Play Store lack such functions [10,28]. This divergence may stem from the fact that the apps in our review were mostly developed by research teams for study purposes. This enabled them to integrate more health care professional support for feedback and web-based consultation features. However, most mobile apps available to patients on their personal smartphones are patient-focused and not accessible by pharmacists or physicians [10]. Integrating chatbot functionality to facilitate mutual conversation between patients and health care professionals to address patient concerns requires costly infrastructure, making it more common in research settings than for commercial purposes. Further research is needed to confirm the effectiveness of these features.

In addition, several motivation systems were identified in the apps to motivate patients to adhere to medication regimens. Two apps provided messages about praise and encouragement for patients with good adherence [33,37], while 5 apps used gamification or financial incentives as motivation strategies [31,35,38,41,42]. Patients would receive rewards such as digital coins, scenic photos, or local currency upon confirming their medication adherence. Consistent with previous studies, our review indicated that reward and gamification-based medication adherence features could effectively improve or maintain optimal medication adherence [44,45]. However, the precise correlation and ceiling effect of the incentive value and its impact on medication adherence required further investigation.

Socialization features are valuable for patients as they offer social support from family, friends, peers, and health care professionals [46]. Patient community forums create comfort space for patients with similar backgrounds to share their experiences of tackling medications, enabling immediate consultation with health care professionals or fellow patients in case of any medication-related concerns [47]. Among the apps reviewed, only the iCARE included such a feature. Moving forward, the development of medication adherence apps may consider incorporating such a socialization feature [47].

Apart from medication intake, disease management relies on dietary adjustments and lifestyle modifications [5]. The educational strategies used by the selected apps encompassed various aspects, including disease background, medication introduction, diet guidance, and exercise recommendations. While it is recognized that simply providing knowledge may not always lead to behavior change, offering basic information can assist patients in navigating their health care journey, preventing the consolidation of misinformation. This is particularly beneficial for individuals facing barriers to accessing credible information for medication use and disease management.

A team-based health care approach facilitates collaboration among patients, caregivers, and providers, enhancing health care satisfaction and patient outcomes [30,31]. Implementing the multiple user support function enables caregivers, such as family members, to remotely monitor patients' medication adherence and health status, taking appropriate actions when necessary [37]. However, in our review, only 3 mobile apps included such functionality [31,33,36], grouping patients, caregivers, and health care professionals as target users to enhance medication adherence. Other apps primarily focused on delivering information to patients themselves [32,35] or facilitating communication between patients and their health care providers [29,30,34,37,41,42], overlooking the role of caregivers. Caregivers play a crucial role in supporting patients with chronic diseases and have been shown to enhance patients' medication adherence and treatment outcomes [40]. Participants in the study by Chew et al [37] also expressed high satisfaction with the multiple user support function of the Med Assist, particularly as it is common for working adults to care for older parents in the Asian community, making multiple-user support a convenient and effective method. In summary, the inclusion of multiple user support functionality may be considered an important feature in future mobile app development, provided privacy concerns are addressed.

Privacy and confidentiality concerns can significantly impact the acceptability of mobile apps [41]. Given that these apps often store sensitive data, such as personal identification information, ensuring system security and protecting user

XSL•FO
RenderX

privacy is crucial [39]. Measures such as user authentication, encryption for data transmission and storage, and strong password protection are essential for safeguarding privacy [30,31,34,39,40]. In addition, app features should be carefully designed to prevent unintended disclosure of users' medical conditions. For instance, while the MIST app requires patients to record and report videos of their medication intake, some users worry that this could inadvertently reveal their HIV status, as it might attract unwanted attention from others [41]. Although this feature may improve medication adherence, it also raises concerns about privacy. Currently, there are no unified regulations or guidelines for privacy protection in mobile health apps, making it important for future developments to prioritize privacy safeguards to maintain user trust and protect sensitive information [36].

None of the selected apps were developed by the government, potentially limiting their use in connecting patient health information across institutions. In some Asian countries with single-payer public health insurance, integrating patient data through a single-payer health care system is more feasible than dealing with data fragmentation from a multipayer health care system [48]. The majority of the mobile apps reviewed were developed by research teams, while others were developed by commercial companies or hospitals. None of the selected apps were able to synch to other adherence devices. Consequently, patient health information recorded in these mobile apps may only be accessible to a few health care networks collaborating with the app developers and may not be easily assessed or updated within other health care systems without special requests or permissions. Incorporating patient-generated health data, which includes clinically relevant information captured by patients outside of conventional health care settings, into electronic health records is essential. This integration is critical because leveraging patient-generated health data in clinical practice has the potential to fill health care gaps and enhance personalized treatment [49]. Furthermore, integrating health services into a single reliable and personalized platform could streamline the health care journey, potentially incentivizing patients to adhere to using a single app [40]. While consolidating patient information from various sources into 1 platform can simplify health care processes, developers need to prioritize data security and personal confidentiality. This is essential to prevent users from avoiding or uninstalling an app due to concerns about privacy breaches [12].

Although the mobile apps reviewed in this study come from various Asian countries with differing design features and cultural contexts, we found consistently high user satisfaction and perceived usefulness [31,33,35,37,38,40,41]. Most studies also reported that these apps positively impacted clinical outcomes and medication adherence [29,33-37,40,42]. These results align with previous research conducted in western populations [10,50-52], indicating that mobile app interventions for patients with chronic diseases are both feasible and well-received in the Asian population as well.

Overall, the selected mobile apps consist of a broad spectrum of components to address medication adherence. While these apps offer diverse features for improving medication adherence, it is important to note that the quality and effectiveness of an app cannot be solely determined by the number of features it provides. Without valid studies, it remains unclear which apps are most suitable for use. Recent evidence suggests a shift in the trend toward the use of digital devices as interest in mHealth grows [53]. Therefore, developers should consider users' needs, their proficiency with digital devices, and the resources available to them when designing apps to promote medication adherence.

Limitations of the Research

This study has several limitations. First, the apps selected were obtained from peer-reviewed journals and not from commercial platforms such as the Apple App Store and Google Play Store. Therefore, the included mobile apps may not represent all available options, particularly those not published in academic databases. It is worth noting that paid commercial apps may offer additional features and functionality, and testing such apps could have provided further insights into strategies used to promote medication adherence. Second, we only included articles available in English, potentially overlooking effective apps in other languages. Third, our search was limited to articles published until August 2024, and given the rapid production and release of new apps, we acknowledge that new adherence apps may have been released after our review.

Conclusion

This review has highlighted key components integrated into mobile apps for medication adherence. Nevertheless, the lack of government and corporate engagement during development impedes the generalizability of any single app. In addition, beyond basic reminder functions, features such as multiple user support, feedback mechanisms, web-based consultation, motivational tools, and socialization features show promising potential in improving medication adherence. Further pragmatic research is essential to confirm the efficacy of these selected apps in enhancing medication adherence.

Data Availability

The study materials and detailed analyses are available from the corresponding author upon reasonable request.

Acknowledgments

This research was supported by the National Science and Technology Council (NSTC), Taiwan (NSTC 113-2628-B-002-018-MY3 and NSTC 113-2410-H-003-103-SS2). These funding bodies were not involved in the study design, data collection, statistical analysis, interpretation of findings, or manuscript preparation.

XSL•FO
RenderX

Authors' Contributions

TW contributed to conceptualization, data curation, investigation, formal analysis, and drafting the manuscript. YMH contributed to conceptualization, methodology, investigation, formal analysis, validation, funding acquisition, supervision, as well as drafting and revising the manuscript. HYC contributed to conceptualization, supervision, validation, funding acquisition, and revising the manuscript. All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Using the SPIDER (sample, phenomenon of interest, design, evaluation, research type) framework to create the research question.
[DOCX File , 14 KB - jmir_v26i1e60787_app1.docx]

Multimedia Appendix 2
Searching process on PubMed and Scopus.
[DOCX File , 14 KB - jmir_v26i1e60787_app2.docx]

Multimedia Appendix 3
Unique features of the selected mobile apps for improving medication adherence.
[DOCX File , 18 KB - jmir_v26i1e60787_app3.docx]

References

1. Cramer JA, Roy A, Burrell A, Fairchild CJ, Fuldeore MJ, Ollendorf DA, et al. Medication compliance and persistence: terminology and definitions. Value Health 2008;11(1):44-47 [FREE Full text] [doi: 10.1111/j.1524-4733.2007.00213.x] [Medline: 18237359]
2. De Geest S, Sabaté E. Adherence to long-term therapies: evidence for action. Eur J Cardiovasc Nurs 2003;2(4):323. [doi: 10.1016/S1474-5151(03)00091-4] [Medline: 14667488]
3. Mahmood S, Jalal Z, Hadi MA, Khan TM, Haque MS, Shah KU. Prevalence of non-adherence to antihypertensive medication in Asia: a systematic review and meta-analysis. Int J Clin Pharm 2021;43(3):486-501. [doi: 10.1007/s11096-021-01236-z] [Medline: 33515135]
4. Brown MT, Bussell J, Dutta S, Davis K, Strong S, Mathew S. Medication adherence: truth and consequences. Am J Med Sci 2016;351(4):387-399. [doi: 10.1016/j.amjms.2016.01.010] [Medline: 27079345]
5. Huang YM, Pecanac KE, Shiyanbola OO. "Why am I not taking medications?" barriers and facilitators of diabetes medication adherence across different health literacy levels. Qual Health Res 2020;30(14):2331-2342. [doi: 10.1177/1049732320945296] [Medline: 32723206]
6. Unni EJ, Farris KB. Unintentional non-adherence and belief in medicines in older adults. Patient Educ Couns 2011;83(2):265-268. [doi: 10.1016/j.pec.2010.05.006] [Medline: 20542401]
7. Watanabe JH, McInnis T, Hirsch JD. Cost of prescription drug-related morbidity and mortality. Ann Pharmacother 2018;52(9):829-837 [FREE Full text] [doi: 10.1177/1060028018765159] [Medline: 29577766]
8. Costa E, Giardini A, Savin M, Menditto E, Lehane E, Laosa O, et al. Interventional tools to improve medication adherence: review of literature. Patient Prefer Adherence 2015;9:1303-1314 [FREE Full text] [doi: 10.2147/PPA.S87551] [Medline: 26396502]
9. Conn VS, Ruppar TM. Medication adherence outcomes of 771 intervention trials: systematic review and meta-analysis. Prev Med 2017;99:269-276 [FREE Full text] [doi: 10.1016/j.ypmed.2017.03.008] [Medline: 28315760]
10. Haase J, Farris KB, Dorsch MP. Mobile applications to improve medication adherence. Telemed J E Health 2017;23(2):75-79. [doi: 10.1089/tmj.2015.0227] [Medline: 27248315]
11. Ng R, Carter SR, El-Den S. The impact of mobile applications on medication adherence: a systematic review. Transl Behav Med 2020;10(6):1419-1435. [doi: 10.1093/tbm/ibz125] [Medline: 31384950]
12. Poushter J. Smartphone ownership and internet usage continues to climb in emerging economies. Pew Research Center 2016;22(1):1-44.
13. Rani R, Kumar R, Mishra R, Sharma SK. Digital health: a panacea in COVID-19 crisis. J Family Med Prim Care 2021;10(1):62-65 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_1494_20] [Medline: 34017704]
14. Chudasama YV, Gillies CL, Zaccardi F, Coles B, Davies MJ, Seidu S, et al. Impact of COVID-19 on routine care for chronic diseases: a global survey of views from healthcare professionals. Diabetes Metab Syndr 2020;14(5):965-967 [FREE Full text] [doi: 10.1016/j.dsx.2020.06.042] [Medline: 32604016]
15. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]

XSL•FO
RenderX

16. Johnson C, Richwine C, Patel V. Individuals' access and use of patient portals and smartphone health apps, 2020. ONC Data Brief (Office of the National Coordinator for Health Information Technology) 2021;57:1-14 [FREE Full text]
17. Chong CJ, Bakry MM, Hatah E, Mohd Tahir NA, Mustafa N. Effects of mobile apps intervention on medication adherence and type 2 diabetes mellitus control: a systematic review and meta-analysis. J Telemed Telecare 2023:1357633X231174933. [doi: 10.1177/1357633X231174933] [Medline: 37254516]
18. Banskota S, Healy M, Goldberg EM. 15 smartphone apps for older adults to use while in isolation during the COVID-19 pandemic. West J Emerg Med 2020;21(3):514-525 [FREE Full text] [doi: 10.5811/westjem.2020.4.47372] [Medline: 32302279]
19. Ahmed I, Ahmad NS, Ali S, Ali S, George A, Saleem Danish H, et al. Medication adherence apps: review and content analysis. JMIR Mhealth Uhealth 2018;6(3):e62 [FREE Full text] [doi: 10.2196/mhealth.6432] [Medline: 29549075]
20. Wagstaff A. Health systems in East Asia: what can developing countries learn from Japan and the Asian tigers? Health Econ 2007;16(5):441-456. [doi: 10.1002/hec.1180] [Medline: 17066428]
21. Rice T, Rosenau P, Unruh LY, Barnes AJ. United states: health system review. Health Syst Transit 2020;22(4):1-441 [FREE Full text] [Medline: 33527901]
22. Wu TY, Majeed A, Kuo KN. An overview of the healthcare system in Taiwan. London J Prim Care (Abingdon) 2010;3(2):115-119 [FREE Full text] [doi: 10.1080/17571472.2010.11493315] [Medline: 25949636]
23. Alhomoud F, Dhillon S, Aslanpour Z, Smith F. South Asian and Middle Eastern patients' perspectives on medicine-related problems in the United Kingdom. Int J Clin Pharm 2015;37(4):607-615. [doi: 10.1007/s11096-015-0103-6] [Medline: 25822040]
24. Jin L, Acharya L. Cultural beliefs underlying medication adherence in people of Chinese descent in the United States. Health Commun 2016;31(5):513-521. [doi: 10.1080/10410236.2014.974121] [Medline: 26422467]
25. Horne R, Graupner L, Frost S, Weinman J, Wright SM, Hankins M. Medicine in a multi-cultural society: the effect of cultural background on beliefs about medications. Soc Sci Med 2004;59(6):1307-1313. [doi: 10.1016/j.socscimed.2004.01.009] [Medline: 15210101]
26. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012;22(10):1435-1443. [doi: 10.1177/1049732312452938] [Medline: 22829486]
27. Baethge C, Goldbeck-Wood S, Mertens S. SANRA-a scale for the quality assessment of narrative review articles. Res Integr Peer Rev 2019;4:5 [FREE Full text] [doi: 10.1186/s41073-019-0064-8] [Medline: 30962953]
28. Tao X, Zhang P, Zhang X, Mao L, Peiris D. Features, functions, and quality of mobile applications for type 2 diabetes care in China: systematic search of app stores. Int J Med Inform 2023;180:105273 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105273] [Medline: 37925856]
29. Liu Y, Chen J, Lamb KV, Wu P, Chang P, Cui Y, et al. Smartphone-based self-empowerment app on secondary prevention of patients with cardiovascular disease. Stud Health Technol Inform 2019;264:1712-1713. [doi: 10.3233/SHTI190610] [Medline: 31438306]
30. Fan X, She R, Liu C, Zhong H, Lau JTF, Hao C, et al. Evaluation of smartphone APP-based case-management services among antiretroviral treatment-naïve HIV-positive men who have sex with men: a randomized controlled trial protocol. BMC Public Health 2020;20(1):85 [FREE Full text] [doi: 10.1186/s12889-020-8171-5] [Medline: 31959139]
31. Chen Y, Ji M, Wu Y, Wang Q, Deng Y, Liu Y, et al. An intelligent individualized cardiovascular app for risk elimination (iCARE) for individuals with coronary heart disease: development and usability testing analysis. JMIR Mhealth Uhealth 2021;9(12):e26439 [FREE Full text] [doi: 10.2196/26439] [Medline: 34898449]
32. Sunil Kumar D, Prakash B, Subhash Chandra B, Kadkol PS, Arun V, Thomas JJ. An android smartphone-based randomized intervention improves the quality of life in patients with type 2 diabetes in Mysore, Karnataka, India. Diabetes Metab Syndr 2020;14(5):1327-1332. [doi: 10.1016/j.dsx.2020.07.025] [Medline: 32755831]
33. Bozorgi A, Hosseini H, Eftekhar H, Majdzadeh R, Yoonessi A, Ramezankhani A, et al. The effect of the mobile "blood pressure management application" on hypertension self-management enhancement: a randomized controlled trial. Trials 2021;22(1):413 [FREE Full text] [doi: 10.1186/s13063-021-05270-0] [Medline: 34167566]
34. Poorcheraghi H, Negarandeh R, Pashaeypoor S, Jorian J. Effect of using a mobile drug management application on medication adherence and hospital readmission among elderly patients with polypharmacy: a randomized controlled trial. BMC Health Serv Res 2023;23(1):1192 [FREE Full text] [doi: 10.1186/s12913-023-10177-4] [Medline: 37919712]
35. Al-Nawayseh MK, Al-Iede M, Elayeh E, Hijazeen R, Oweidat KA, Aleidi SM. The impact of using a mobile application to improve asthma patients' adherence to medication in Jordan. Health Informatics J 2021;27(3):14604582211042926 [FREE Full text] [doi: 10.1177/14604582211042926] [Medline: 34587767]
36. Nurakysh S, Kurakbayev K, Kosherbaeva L, Tazhiyeva A, Aimakhanova A, Kulkaeva G, et al. Evaluation of the effectiveness of the mobile application on adherence of patients with arterial hypertension. Acta Inform Med 2022;30(1):18-24 [FREE Full text] [doi: 10.5455/aim.2022.30.18-24] [Medline: 35800910]
37. Yang Y, Lee EY, Kim HS, Lee SH, Yoon KH, Cho JH. Effect of a mobile phone-based glucose-monitoring and feedback system for type 2 diabetes management in multiple primary care clinic settings: cluster randomized controlled trial. JMIR Mhealth Uhealth 2020;8(2):e16266 [FREE Full text] [doi: 10.2196/16266] [Medline: 32130172]

XSL•FO
RenderX

38. Chew S, Lai PSM, Ng CJ. Usability and utility of a mobile app to improve medication adherence among ambulatory care patients in Malaysia: qualitative study. JMIR Mhealth Uhealth 2020;8(1):e15146 [FREE Full text] [doi: 10.2196/15146] [Medline: 32003748]
39. Park JE, Lee JE, Moon BK, Lee H, Park SH, Kim SK, et al. Impact of a pharmaceutical care service for patients with rheumatoid arthritis using a customised mobile device (the PROUD trial): study protocol for a randomised controlled trial. BMJ Open 2022;12(6):e061917 [FREE Full text] [doi: 10.1136/bmjopen-2022-061917] [Medline: 35728909]
40. Huang Z, Tan E, Lum E, Sloot P, Boehm BO, Car J. A smartphone app to improve medication adherence in patients with type 2 diabetes in Asia: feasibility randomized controlled trial. JMIR Mhealth Uhealth 2019;7(9):e14914 [FREE Full text] [doi: 10.2196/14914] [Medline: 31516127]
41. Pang Y, Molton JS, Ooi WT, Paton NI, He HG. Preliminary effects of a mobile interactive supervised therapy intervention on people living with HIV: pilot randomized controlled trial. JMIR Mhealth Uhealth 2020;8(3):e15702 [FREE Full text] [doi: 10.2196/15702] [Medline: 32217500]
42. Chen HH, Hsu HT, Lin PC, Chen CY, Hsieh HF, Ko CH. Efficacy of a smartphone app in enhancing medication adherence and accuracy in individuals with schizophrenia during the COVID-19 pandemic: randomized controlled trial. JMIR Ment Health 2023;10:e50806 [FREE Full text] [doi: 10.2196/50806] [Medline: 38096017]
43. Huang YM, Shiyanbola OO. Investigation of barriers and facilitators to medication adherence in patients with type 2 diabetes across different health literacy levels: an explanatory sequential mixed methods study. Front Pharmacol 2021;12:745749 [FREE Full text] [doi: 10.3389/fphar.2021.745749] [Medline: 34690778]
44. DeFulio A, Silverman K. The use of incentives to reinforce medication adherence. Prev Med 2012;55(Suppl):S86-S94 [FREE Full text] [doi: 10.1016/j.ypmed.2012.04.017] [Medline: 22580095]
45. Tran S, Smith L, El-Den S, Carter S. The use of gamification and incentives in mobile health apps to improve medication adherence: scoping review. JMIR Mhealth Uhealth 2022;10(2):e30671 [FREE Full text] [doi: 10.2196/30671] [Medline: 35188475]
46. Scheurer D, Choudhry N, Swanton KA, Matlin O, Shrank W. Association between different types of social support and medication adherence. Am J Manag Care 2012;18(12):e461-e467 [FREE Full text] [Medline: 23286676]
47. Altuwayrib S, Win KT, Freeman M. Medication adherence support applications for chronic arthritis patients: healthcare providers' perspective in Saudi Arabia. Stud Health Technol Inform 2024;310:1056-1060. [doi: 10.3233/SHTI231126] [Medline: 38269976]
48. Petrou P, Samoutis G, Lionis C. Single-payer or a multipayer health system: a systematic literature review. Public Health 2018;163:141-152. [doi: 10.1016/j.puhe.2018.07.006] [Medline: 30193174]
49. Tiase VL, Hull W, McFarland MM, Sward KA, Del Fiol GD, Staes C, et al. Patient-generated health data and electronic health record integration: a scoping review. JAMIA Open 2020;3(4):619-627 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa052] [Medline: 33758798]
50. Hakami AM, Almutairi B, Alanazi AS, Alzahrani MM. Effect of mobile apps on medication adherence of type 2 diabetes mellitus: a systematic review of recent studies. Cureus 2024;16(1):e51791 [FREE Full text] [doi: 10.7759/cureus.51791] [Medline: 38192533]
51. Puig J, Echeverría P, Lluch T, Herms J, Estany C, Bonjoch A, et al. A specific mobile health application for older HIV-infected patients: usability and patient's satisfaction. Telemed J E Health 2021;27(4):432-440. [doi: 10.1089/tmj.2020.0098] [Medline: 32667858]
52. Morawski K, Ghazinouri R, Krumme A, Lauffenburger JC, Lu Z, Durfee E, et al. Association of a smartphone application with medication adherence and blood pressure control: the MedISAFE-BP randomized clinical trial. JAMA Intern Med 2018;178(6):802-809 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0447] [Medline: 29710289]
53. Fiordelli M, Diviani N, Schulz PJ. Mapping mhealth research: a decade of evolution. J Med Internet Res 2013;15(5):e95. [doi: 10.2196/jmir.2430] [Medline: 23697600]

Abbreviations

MIST: Mobile Interactive Supervised Therapy
SPIDER: sample, phenomenon of interest, design, evaluation, research type

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 21.05.24; peer-reviewed by MJ Wen, Z Ikawati; comments to author 16.07.24; revised version received 20.08.24; accepted 23.09.24; published 08.11.24.

Please cite as:
Wang T, Huang YM, Chan HY
Exploration of Features of Mobile Applications for Medication Adherence in Asia: Narrative Review
J Med Internet Res 2024;26:e60787
URL: https://www.jmir.org/2024/1/e60787
doi:10.2196/60787
PMID:

©Tzu Wang, Yen-Ming Huang, Hsun-Yu Chan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Participation in eHealth Communication Interventions Among Patients Undergoing Hemodialysis: Scoping Review

Anne Deinboll[1], RN, MPK; Cathrine Fredriksen Moe[2], RN, PhD; Mette Spliid Ludvigsen[2,3], RN, PhD

[1]Faculty of Nursing and Health Sciences, Nord University, Mo i Rana, Norway

[2]Faculty of Nursing and Health Sciences, Nord University, Bodø, Norway

[3]Department of Clinical Medicine, Randers Regional Hospital, Aarhus University, Aarhus, Denmark

Corresponding Author:
Anne Deinboll, RN, MPK
Faculty of Nursing and Health Sciences
Nord University
Pb 614
Mo i Rana, 8607
Norway
Phone: 47 75129709
Email: anne.deinboll@nord.no

Abstract

Background: eHealth communication interventions have been shown to offer individuals with chronic kidney disease the opportunity to embrace dialysis therapies with greater confidence, the potential to obtain better clinical outcomes, and an increased quality of life. eHealth is an emerging field that offers diverse, flexible designs and delivery options. However, existing evidence on eHealth communication among patients undergoing hemodialysis is sparse and scattered and lacks systematization.

Objective: This scoping review aims to identify and map the current evidence on patient participation in eHealth communication interventions. We aimed to map the associations between interventions and electronic health records, the participative role of individuals living with chronic kidney disease and undergoing hemodialysis, and the barriers to and facilitators of patient involvement in eHealth communication with health care professionals.

Methods: This study used the Joanna Briggs Institute methodology for conducting a scoping review. Studies eligible for inclusion were those that included adult patients (aged >18 y) undergoing all types of hemodialysis, including prescheduled in-center hemodialysis and conventional home-based hemodialysis. Systematic searches were completed in Ovid MEDLINE, Ovid Embase, EBSCOhost CINAHL with Full Text, Scopus, and ProQuest Dissertations and Theses. Extracted data from the included studies were presented in figures and tables along with descriptions that responded to the research questions. This review was reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines.

Results: In total, 9 peer-reviewed studies were included. The main result was a low participative patient role and a vaguely described link to electronic health records. The key participative facilitators were availability of and access to the intervention; security, trust, and confidence; patient knowledge of their health situation and use of self-care; and patient preparedness for an uncertain future health situation and the ability to relate to family and friends about it. The key participative barriers were lack of availability of and access to information, mistrust and lack of safety, lack of knowledge of health situation and self-care, and relational issues. All barriers and facilitators were related to health literacy.

Conclusions: This scoping review summarizes 4 specific and 3 nonspecific eHealth communication interventions developed and evaluated in various studies involving patients receiving hemodialysis. A knowledge gap exists between low levels of patient participation in eHealth communication and patients' limited access to electronic health records. eHealth communication interventions should implement patient participation and focus on the fact that different modalities of eHealth communication can complement face-to-face communication.

International Registered Report Identifier (IRRID): RR2-10.2196/38615

(J Med Internet Res 2024;26:e51900) doi:10.2196/51900

XSL•FO
RenderX

KEYWORDS

eHealth; electronic health records; hemodialysis; patient participation; renal dialysis; renal insufficiency; chronic; mobile phone

Introduction

Background

Despite substantial improvements in hemodialysis delivery and outcomes over the last decade, patients undergoing hemodialysis continue to have negative experiences [1]. Particularly, patients undergoing hemodialysis still experience poor quality of life (QOL); their symptoms remain prevalent; and their financial burden, morbidity, and mortality remain unacceptably high [2]. Studies have indicated a potential survival advantage with intensive dialysis (ie, an increase in dialysis frequency or duration), a goal that can be achieved with home-based dialysis [3]. Home hemodialysis (HHD) is mostly provided in Western Europe, and increased use of HHD aims to strengthen patients' position and QOL and reduce financial burdens and human resource challenges [2,4]. Patients undergoing hemodialysis depend on effective and real-time communication with renal health care professionals (HCPs) to optimize their QOL owing to the high symptom burden and the complexity of the illness and treatment [5].

The increasing demand for participation in health-related decision-making in both in-center hemodialysis and HHD requires innovative interventions for eHealth communication [6]. Digital home monitoring is in development and enables 2-way communication between patients and HCPs. Many technical challenges have been described in implementing communication interventions and information flows in digital health care systems [7-9]. However, the challenges of the participative role of patients undergoing hemodialysis have rarely been explored.

Approximately 3.9 million individuals living with stage-V chronic kidney disease (CKD) worldwide receive kidney replacement therapy, kidney transplantation, or various forms of dialysis, and the numbers continue to rise [10,11]. In-center hemodialysis is more common than peritoneal dialysis (PD), with approximately 2.3 million individuals undergoing hemodialysis in 2019 [12]. In hemodialysis, blood is pumped out of the body to an artificial kidney machine and returned to the body via tubes connecting the machine to the patient [13], thus involving direct access to the patient's circulatory system. Complications such as severe bleeding, venous thrombosis, infection, or low blood pressure may occur; therefore, patients should be knowledgeable and competent regarding interventions to reduce complications and ensure their safety [14]. Chuasuwan et al [15] reported a lower QOL in patients receiving hemodialysis than in those receiving PD. PD can be performed more easily than hemodialysis at home [15].

Patient participation is increasingly promoted to improve HCPs' responsiveness to patient needs and ensure the legitimacy of decisions affecting patient care [16,17]. Patient participation has individual and collective dimensions. The individual dimension refers to enabling patients to have more influence over their health by increasing their capacity to gain more control over the issues they consider important, including access to their electronic health record (EHR). The collective dimension refers to patient participation in collective activities, in which patients, relatives, representatives, or service users are actively engaged in shaping the development of health care services [18]. The expanding development and dissemination of eHealth interventions is a paradigm shift toward enhanced individuality and patient-centered care [15-17].

Interactions between nurses and patients during hemodialysis sessions involve ongoing dialogues on topics such as hydration status, treatment goals, and practical decisions regarding dialysis procedures [19]. These interactions differ significantly from those in other medical contexts, such as diabetes, owing to frequent and prolonged treatment durations. A critical aspect for patients undergoing hemodialysis involves ongoing consultations regarding health status and eligibility for kidney transplantation, a life-saving procedure for recipients. Inadequate communication about transplantation status between patients and the involved clinicians can compromise patient empowerment and the nurse-patient relationship [20]. Moreover, a proportion of patients receiving hemodialysis do not qualify for kidney transplantation owing to various medical factors [2]. Considerations such as dialysis discontinuation, end-of-life decisions, and subsequent adjustments to supportive treatment and care preferences become particularly crucial [19]. This highlights another significant issue that profoundly impacts this patient group. We hypothesized that individuals undergoing hemodialysis encounter unique issues and interactions owing to the long-term collaboration and communication between patients and HCPs during frequent dialysis sessions whether in a clinical setting or at home. Overall, these factors may necessitate specific eHealth communication interventions tailored to the needs of patients receiving hemodialysis. Therefore, we found it appropriate to focus exclusively on the population undergoing hemodialysis.

Prior Work and Implications for This Study

A previous review reported a low level of eHealth literacy when assessing the availability, acceptability, and use of mobile health in a population with CKD [21]. Another review found a knowledge gap in sociocultural and safety aspects when exploring mobile health use in a population undergoing hemodialysis [22]. However, none of these reviews addressed patient participation, communication, barriers, and facilitators. Therefore, this study explored the participative role of patients undergoing hemodialysis. We used the phrase *eHealth communication* to refer to communication technologies that enable HCP-patient interaction through electronic means [23]. This term encompasses digital and electronic communication interventions [8]. Many eHealth interventions are multimodal, and their definitions may overlap [6]. Distinguishing between *synchronous communication* as video, telephone, and direct messaging functionalities and *asynchronous communication* as email and SMS text messages is common [24]. This study explored the role of patients undergoing hemodialysis in both types of eHealth communication, such as patient portals; telehealth solutions (eg, videoconferencing); and the use of

XSL•FO
RenderX

computers, minicomputers, tablets, networks, or cloud storage for managing and storing medical records.

Furthermore, we explored the potential of 2-way digital communication in patient outcomes. Electronic patient-reported outcome measures (ePROMs) improve patient outcomes [25,26]. Knowledge of patient experience with digital communication related to the use of ePROMs is lacking [27]. In addition, we aimed to investigate the role of patients in digital communication related to ePROMs. The COVID-19 pandemic boosted the development and use of eHealth in nephrology; however, additional knowledge of the barriers to and facilitators of patient participation is required owing to the existing knowledge gap regarding why some patients do not use digital communication. Phelps et al [28] analyzed the use of a UK patient portal from 2009 to 2013 by patients with CKD and found that inactive users had not registered with or logged on to the patient portal. A study of barriers to and facilitators of the use of general eHealth by older adults found that high use of eHealth by older people was not possible [29]. Therefore, we found it necessary to explore the need for digital communication in patients undergoing hemodialysis. A comprehensive overview of the available eHealth communication interventions for patients undergoing hemodialysis and HCPs will aid in identifying the status of patient participation in eHealth communication and the gaps and areas for further development.

Objectives and Research Questions

This scoping review aimed to systemize and map emerging research on eHealth communication interventions for patients undergoing hemodialysis and their participative role in these interventions and to identify barriers to and facilitators of patient participation.

In this review, we were guided by the following research questions (RQs):

1. Which types of eHealth communication interventions for patients undergoing hemodialysis and HCPs can be identified in the literature, and how are they linked to EHRs?
2. Which participative roles of patients undergoing hemodialysis in eHealth communication with HCPs can be identified in the literature?
3. What are the key participative barriers to and facilitators of eHealth communication with HCPs that are encountered by patients undergoing hemodialysis?

This review considered studies that included a population of adult patients (aged >18 y) undergoing hemodialysis. All types of hemodialysis, including prescheduled in-center dialysis and conventional HHD, were included. Studies focusing on patients aged <18 years, patients with stage-I to stage-IV CKD, patients undergoing transplants, and those undergoing PD were excluded. Intervention studies that included patients undergoing hemodialysis and PD were included; however, studies that included only patients undergoing PD were excluded. This review included studies that explored a 3-fold concept. First, the types of eHealth communication interventions included both electronic and digital technology for oral or written communications for example, EHRs (including standardized nursing terminology, electronic patient records or portals, electronic conferencing, and mobile written or oral communication mediated by electronic or digital means). Studies on eHealth interventions with no possibility for communication between patients with CKD and health professionals were excluded (eg, mobile apps for self-efficacy). Second, patient participation in this review referred to the definition by Thompson [30], that is, patient participation requires professionals to engage in 2-way communication. We included studies at all the continuum levels of patient participation, involvement, and similar concepts. Third, barriers to patient participation in eHealth communication included problems, issues, challenges, and obstacles to participation. Facilitators of patient participation included recommendations, interventions or programs, motivation, and experienced results [6]. The barriers and facilitators encountered by patients undergoing hemodialysis were included.

The context for the review was hemodialysis care.

Methods

Design

This review was conducted based on an a priori published protocol [31]. We adopted the Joanna Briggs Institute methodology for scoping reviews to present methodological rigor and transparency [32,33]. This scoping review followed a structured and predefined process that requires rigorous methods to ensure that the results are reliable and meaningful to end users, in line with those of other systematic reviews. We conducted a comprehensive systematic literature search, data extraction, and charting accompanied by a narrative summary of eHealth communication interventions and participative roles of patients undergoing hemodialysis. Scoping reviews do not offer direct recommendations for practice; therefore, no critical appraisal of the methodological quality of the included studies was required. However, the results of scoping reviews are widely used to inform the development of trustworthy clinical guidelines [34]. Thus, this review contributes with an overview of existing research on eHealth communication in hemodialysis and future interventions in this field. This review was reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [35]. PRISMA checklist has been provided in Multimedia Appendix 1.

Eligibility Criteria

The eligibility criteria for this review are outlined in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria. Eligibility criteria for the screening and inclusion process, including article type, study type, and language.

Inclusion criteria

- Article and study type: peer-reviewed primary studies, including qualitative, quantitative, and mixed methods study designs, and PhD theses (PhD theses were included due to the suspected empirical and methodological level in line with peer-reviewed articles; Master's theses, unpublished studies, and conference abstracts were excluded due to lack of peer review)
- Language: English

Exclusion criteria

- Article and study type: conference abstracts, unpublished studies, and Master's theses
- Language: languages other than English

Information Sources

We performed a systematic search using the following electronic databases: MEDLINE (Ovid), Embase (Ovid), CINAHL (EBSCOhost CINAHL with Full Text), Scopus, and ProQuest Dissertations and Theses Open. No restrictions on year range were imposed on the search owing to the novelty of the research field.

Search

A full electronic search strategy for MEDLINE (Textbox 2) was developed and adjusted to the remaining databases. The systematic searches were conducted on May 6, 2022. An identical and updated systematic search was conducted across all databases covering the period from January 1, 2022, to June 4, 2024, to ensure the inclusion of the latest research.

Textbox 2. Search documentation for MEDLINE.

MEDLINE (Ovid) search strategy

- Search dates: May 6, 2022 (190 records retrieved); June 4, 2024 (49 new records retrieved; published since May 6, 2022)
- Participant and context search terms were mutually inclusive. Concept search terms were divided into 2 search blocks.
- No limitations
- Searched: (1) population, context, and Medical Subject Heading (MeSH) terms+text words (TW); (2) concept 1, MeSH terms+TW; (3) concept 2, MeSH terms+TW; (4) 1 AND 2 AND 3

We performed backward citation chasing by reviewing the reference lists of the included records and forward citation chasing of the included records using Google Scholar to ensure a complete selection of studies for the review and that no studies were overlooked [36].

Selection of Sources

All identified records were uploaded to the reference management tool EndNote (version 20; Clarivate Analytics) [37]. Duplicates were removed, and the remaining records were uploaded to the screening software tool Rayyan (Qatar Computing Research Institute) [38]. Initially, the review team conducted a pilot test to ensure consensus on including the first 25 records. Next, titles and abstracts were screened, followed by a full-text screening. We contacted the corresponding authors of 6 studies [39-44] for additional studies or study information to fulfill the study selection process.

The first and second authors discussed the studies during the selection process, and the last author was included in the discussion for agreement on the final included articles.

Data Charting and Items

A revised charting form from the protocol was used to extract data from the included studies. The data were extracted by the first author, and the extracted data were validated against the eligibility criteria by the second author. The study characteristics and data for each RQ were subsequently extracted.

Analysis and Presentation of Results

A scoping review intends to provide an overview rather than synthesize data; nevertheless, a basic analytic framework is required to collate, develop, and present a narrative description of the data answering the RQs [32]. The first 3 phases of the basic qualitative coding framework by Elo and Kyngäs [45] were used to consider the level of patient participation according to the taxonomy by Thompson [30]. In addition, this coding framework was used to identify and clarify barriers and facilitators related to participation in eHealth communication interventions available for patients receiving hemodialysis. Disagreements between the first and second authors were resolved based on consensus through discussion or by including the third author. The reviewer team regularly met during several stages of the review process. The extracted and analyzed data from the included papers were presented in figures and tables along with descriptions that responded to the RQs. The tables were presented according to the distribution of sources of evidence by year of publication, country of origin, hemodialysis context, population, aim, and research methods; type of intervention, association with EHR, type of communication, and level of involvement; facilitators; barriers; and a descriptive summary of the charted results.

XSL•FO
RenderX

Results

Process of Inclusion of Sources

Among the 934 records identified through the systematic search, the titles and abstracts of 510 (54.6%) were screened against the eligibility criteria after removing duplicates (Figure 1 [46]). The full texts of 1.4% (7/510) of the reports [44,47-52] could not be located. In total, 19.6% (100/510) of full-text articles were reviewed. A total of 7 studies on patients' role in communicative eHealth interventions for patients receiving hemodialysis and HCPs were included. The updated search showed an increase in the number of published studies in this area over the 2 years since the first systematic search in 2022. The search led to the inclusion of 2 additional studies. A total of 9 studies were included.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the identification of sources from databases and the screening and inclusion processes [46].

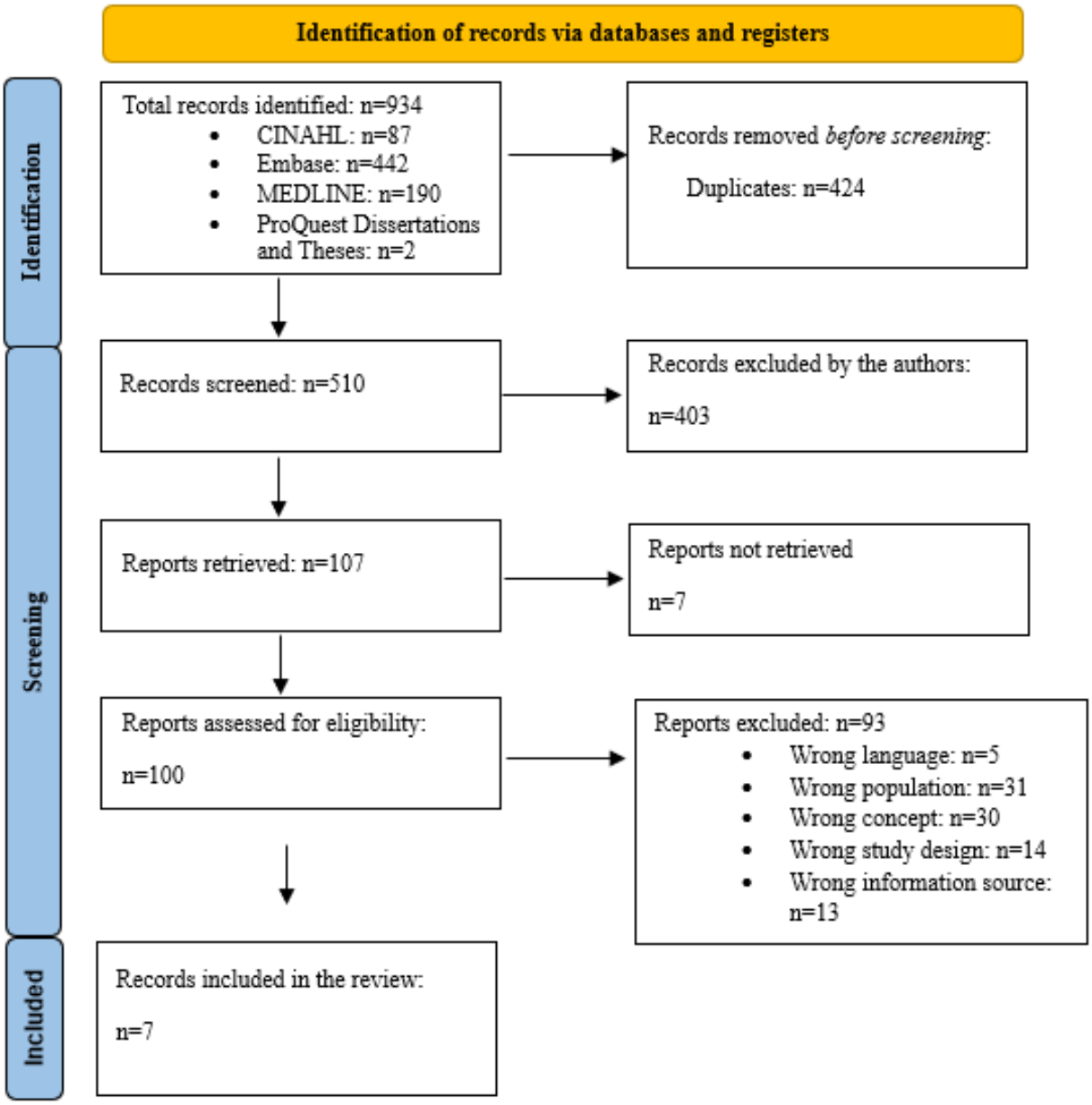

Study Characteristics

The 9 studies [41,44,53-59] (Table 1) were published between 2013 and 2023 and were conducted in North America (United States: n=4, 44%; Canada: n=1, 11%), Australia (n=1, 11%), and Europe (United Kingdom: n=3, 33%). The included sources were all peer-reviewed primary studies, and 11% (1/9) of the studies were published as an editorial letter. In total, 56% (5/9) of the studies included participants from only 1 dialysis unit [53-57], whereas 33% (3/9) of the studies included participants from 2 to 4 dialysis units [41,44,58]. A total of 11% (1/9) of the studies were conducted in the transition from a large dialysis unit to HHD [59]. The studies were all conducted in a context involving hemodialysis, varying from small satellite units to larger central hospital units.

XSL•FO
RenderX

Table 1. Study characteristics. Information about the authors, year of publication, actual hemodialysis (HD) context and participants, aims, study designs, and methods for the 9 studies.

Study and country	HD context	Participants or population (number, age, and sex)	Aim	Study design and method
Chu et al [59], 2023, Australia	Home dialysis (HD and PD[a]) and carers	34 surveys (HD: 8 patients; PD: 19 patients; 7 carers); 21 interviews with patients (HD: 7; PD: 14; 2 of the interviews were together with carers); aged 49-81 years; both sexes	To explore factors influencing the behaviors and perceptions of patients undergoing home dialysis and their carers regarding telehealth-assisted home visits	A concurrent mixed methods design; a combination of surveys and qualitative, semistructured interviews via telephone
Giles et al [58], 2017, United Kingdom	4 renal units including different modalities of renal treatment (2 central hospital units and 2 satellite units)	Phase 1: 57 participants[b] of both sexes aged 34-76 years, including 29 patients (other participants: carers and relatives and HCPs[c]); phase 2: 34 participants[b] aged 18-86 years, including 20 patients (other participants: staff, relatives, and 1 non–English-speaking patient)	(1) To explore the feasibility of patients with renal disease, HCPs, and researchers working in partnership to develop a patient-led quality and safety feedback system within an existing EHR[d] (RPV[e]); (2) to adapt an existing technology (RPV) to accommodate a patient-led quality and safety feedback system and collect a range of qualitative data from patients and staff to establish the main principles and components of the prototype and assess usability	Qualitative design—an evaluation of the first two phases of an eHealth implementation (1) focus group interviews (separate and combined; patients and staff) and (2) individual interviews
Hazara and Bhandari [53], 2016, United Kingdom	1 renal department	69 patients with CKD[b,f] and 16 patients receiving HD; aged 20-83 years; both sexes	To elucidate why some users of the self-management and education website RPV remain inactive	Quantitative design; paper-based questionnaires, including a free-text box
Hudson et al [54], 2020, United Kingdom	1 HD unit (on the southeast coast)	10 adult patients (8 undergoing HD and 2 undergoing transplant)[b]; median ages of 54 years for HD and 63 years for nonusers of RPV; both sexes	To understand how patients with kidney disease use PatientView in their self - care practice	Qualitative design—ethnography; semistructured interviews and observation
Jakubowski et al [55], 2019, United States	A single dialysis clinic	10 adult patients receiving HD (10 initially participated, and 8 patients completed the study); mean age 58.7 years; both sexes	To pilot the feasibility and acceptability of a technology-assisted cognitive behavioral therapy intervention for patients undergoing HD, share design and implementation lessons learned, and provide preliminary results on changes in selected patient-reported symptoms	Mixed methods design—a single-center pilot feasibility study; questionnaire combined with open-ended questions
Kiberd et al [57], 2018, Canada	1 multidisciplinary home dialysis clinic; PD and HD	41 adult patients (HPD[g] and HHD[h])[b] provided consent to join the portal. A total of 27 patients (66%) created a web account and joined the portal. At 6 months, 16 patients had completed follow-up, and at 12 months, 10 patients had completed follow-up for the primary outcome. Mean age 57.1 years; both sexes.		Quantitative design—a single-arm pilot, clinical trial; 4 different questionnaires and in addition the telephone use was compared before and after the patients joined the web portal.

XSL•FO
RenderX

Study and country	HD context	Participants or population (number, age, and sex)	Aim	Study design and method
Ladin et al [41], 2021, United States	4 nephrology sites across the United States[i]	Older patients with stage-4 or stage-5 CKD[b]. Of the 60 interviews, 19 (32%) were with clinicians, 30 (50%) were with patients, and 11 (18%) were with care partners; 16 clinicians (84%) were nephrologists; aged ≥70 years; both sexes	To identify the perceptions of patients, care partners, and nephrologists regarding the patient-centeredness, benefits, and drawbacks of telehealth compared with in-person visits	Qualitative design; semistructured interviews
Lew et al [44], 2023, United States	2 HD units in a district on the southeast coast	94 adult patients receiving HD; mean age 56.5 (SD 13.5) y; both sexes	To evaluate patient experiences with the use of telehealth by their nephrologist for HD during the COVID-19 pandemic	Quantitative design; survey using a Likert scale
Minatodani and Berman [56], 2013, United States	1 outpatient dialysis center; HD	33 participants: 30 patients receiving HD and 3 caregivers; aged 37-87 years; both sexes	(1) To evaluate patients' perceived effectiveness and satisfaction with home telehealth self-monitoring and remote care nurses; (2) to identify perceived facilitators and barriers encountered with remote technology use	Mixed methods design; semistructured individual interviews and an additional survey

[a]PD: peritoneal dialysis.

[b]Mixed population of participants. Patients undergoing hemodialysis were included; however, it was not possible to extract data only from patients receiving hemodialysis.

[c]HCP: health care professional.

[d]EHR: electronic health record.

[e]RPV: Renal Patient View.

[f]CKD: chronic kidney disease.

[g]HPD: home peritoneal dialysis.

[h]HHD: home hemodialysis.

[i]Patients were initially excluded if they were on dialysis, but certain patients initiated hemodialysis during the course of the study and were followed up on until the study ended (a maximum of 18 months).

The studies involved a total of 422 participants, including *others* (eg, HCPs or patients with CKD in varying stages). Overall, 67% (6/9) of the studies included 171 adult patients undergoing hemodialysis [44,53-56,59]; 22% (2/9) of the studies included 79 patients with CKD, including a nonapplicable fraction of patients receiving hemodialysis [41,58]; 11% (1/9) of the studies included 55 patients on dialysis in transition between unit and home dialysis [59]; and 11% (1/9) of the studies included 10 patients on HHD (hemodialysis or PD) [57]. The extractable distribution of patients undergoing hemodialysis was 208 men and 140 women. The number of participants ranged from 8 [55] to 94 [44]. In total, 11% (1/9) of the studies included older patients with CKD aged >70 years [41]. A total of 78% (7/9) of the studies explored the experiences of various populations, including patients with diverse stages of CKD [41,53,54,56-59]. We included these studies because they included patients undergoing hemodialysis, although separating their experiences from those of patients with CKD was not possible. Most of the participants in 33% (3/9) of the studies were undergoing hemodialysis [44,54,56]. Some studies (4/9, 44%) additionally included the experiences of others (caregivers or family) [41,56,58,59] and HCPs [41,58]; nevertheless, the experiences of the patients undergoing hemodialysis were exclusively extracted and separated from the experiences of others. A total of 44% (4/9) of the studies aimed to understand, improve, or provide results on whether patients experienced changes in self-care practices or health-related QOL [41,54-57]. One study elucidated why some users remained inactive [53]. One study aimed to explore factors influencing the behaviors and perceptions of patients on HHD and their carers regarding telehealth-assisted visits and identify the factors influencing their engagement with telehealth [59]. The study was guided by the Behavior Change Wheel framework [60]. One study aimed to evaluate patient experiences with the use of telehealth during the COVID-19 pandemic [44]. One study aimed to evaluate and assess the usability of a patient-led existing technology [58]. The study designs included qualitative (3/9, 33%) [41,54,58], mixed methods (3/9, 33%) [55,56,59], and quantitative (3/9, 33%) [44,53,57] designs. Data collection methods included focus groups and individual oral interviews or questions (6/9, 67%) [41,54-56,58,59], written questionnaires (5/9, 56%) [44,53,56,57,59], and participant observations (1/9, 11%) [54].

XSL•FO
RenderX

Types of eHealth Communication Interventions, Associations With EHRs, and Level of Patient Participation

Overview

A total of 7 eHealth communication interventions were identified. In total, 57% (4/7) were dedicated and named patient portals or eHealth communication platforms: Renal Patient View (RPV), which was investigated in 33% (3/9) of the studies [53,54,58]; RelayHealth [57]; VitelCare Turtle 500 (Vitel Health) [56]; and Vidyo [55]. A total of 43% (3/7) of the interventions were broadly described, unnamed, or general oral video or telehealth consultations [41,44,59] (Table 2).

Table 2. eHealth communication interventions. Information about the eHealth communication interventions and their associations with electronic health records (EHRs), possible communication types, and patient participation level [30].

Study	eHealth communication intervention	Association with EHR	Type of communication	Level of patient participation [30][a]
Chu et al [59], 2023	Telehealth-assisted home visits; telephone or videoconference; telehealth system or platform; —[b]	—	Synchronous	2
Giles et al [58], 2017	An electronic quality and safety feedback system within the web-based patient portal RPV[c]	Link to the EHR for the electronic feedback system; —; yes	Asynchronous	1
Hazara and Bhandari [53], 2016	RPV, a self-management and education website	Yes	Asynchronous	1
Hudson et al [54], 2020	PatientView (formerly termed RPV), a patient portal	Yes	Asynchronous	1
Jakubowski et al [55], 2019	Vidyo videoconferencing platform	—	Synchronous and prearranged	2
Kiberd et al [57], 2018	RelayHealth (McKesson Canada), a web-based eHealth patient portal	Yes	Asynchronous	2
Ladin et al [41], 2021	Telehealth consultations with video and sound	—	Synchronous and prearranged	2
Lew et al [44], 2023	Telehealth consultations, video, and sound via laptop or tablet through a secure video platform	—	Synchronous and spontaneous	2
Minatodani and Berman [56], 2013	VitelCare, a home telehealth monitoring unit with concomitant remote care nurse support	—	Asynchronous, synchronous, and prearranged	2

[a]4=informed decision-making; 3=professional as agent; 2=consultation; 1=information giving; 0=exclusion.

[b]Not applicable.

[c]RPV: Renal Patient View.

RPV Patient Portal

This is a patient portal offered for free, initially piloted and implemented in 2004 and intended for use by patients with renal disease under the care of renal physicians in the United Kingdom. RPV is described as a self-management and education website available anywhere with an internet connection using either an internet browser or a dedicated app. Access to the RPV is strictly controlled through usernames and passwords. Patients can authorize access to their accounts by other HCPs involved in their care (eg, general practitioners or community nurses) [53].

The intervention was renamed PatientView in 2018 and has been implemented and used by several other diagnostic groups. RPV is a common UK eHealth portal used in 90% of all renal units and has been widely implemented since 2008 [54]. We were unable to extract specific information on the functionalities of the patient portal. However, information from the RenalView website indicated the following: the portal provides patients with 3 functions—“Manage” (eg, view blood tests with graphs and trends), “Monitor” their health, and “Message” their health care team [61]. The “Manage” and “Monitor” functions are used by patients, whereas the “Message” function was not clearly described [53,54,58].

RelayHealth

This is a web-based eHealth patient portal that allows patients and HCPs to communicate through a secure, password-protected application. The portal permits the visualization of the messaging histories of the patient and provider. Specialty nurses are trained in portal use. After entering the portal, the patient and home dialysis health care team, which includes nurses, a home dialysis physician, and dieticians, can send messages related to patient care at any time. Messages can be sent between the patient and health care team, including proposed changes to medication, instructions after a clinic visit, upcoming tests, times of new appointments, and questions about care. Messages are electronically stored within the portal, printed, and placed

XSL•FO
RenderX

in the patient's hospital chart to comply with the hospital's legal standards. Both the patients and the health care team are made aware of new messages through email prompts [57].

VitelCare (Vitel Health)

This is a home telehealth monitoring unit that is compliant with Health Insurance Portability and Accountability Act and is broadband enabled [56]. We were unable to extract information on the monitoring unit from the study. The following information about the intervention is from a previous randomized controlled trial: accessories included a blood pressure monitor, scale, pulse oximeter, glucose monitor, video camera, and headset. The patients captured their health data and answered 10 questions on nondialysis days according to an individualized care plan. Videoconferencing between the patient on hemodialysis and "a remote care nurse" was provided using the video camera and headset at arranged times. A trained remote nurse reviewed all incoming data and contacted the patients and caregivers via telephone if the data were outside the normal range or incomplete, as predetermined in the individual plan. The nurses provided additional technical support [62].

Vidyo

This is a secure videoconferencing platform used by behavioral therapists for sessions of live 2-way, technology-assisted, video-recorded cognitive behavioral therapy for patients on hemodialysis. Each session lasts 45 to 60 minutes. Patients use a study laptop with headphones and a microphone and connect in real time with therapists over a secure Wi-Fi hot spot. The platform is used for 8 regularly scheduled sessions per patient during regularly scheduled hemodialysis sessions [55].

Video or Telehealth Consultations in General

The telehealth platforms used by the participants in 4 nephrology clinics in the United States and Australia were not technically described [41,44,59]. In these studies, telehealth or telemedicine was determined following the World Health Organization's definition: "...the provision of healthcare services at a distance... conducted between remote healthcare users seeking health services and healthcare providers (client-to-provider telemedicine)" [44,63]. Telehealth consultations occurred via secure 2-way technology-assisted platforms mediated through mobile phones, PCs, and tablets [21].

Association With EHRs

The association between the eHealth interventions and EHRs was inconsistently described in 2 patient portals: RPV and RelayHealth (Table 2). The descriptions stated that RPV allowed patients to gain secure access to parts of the EHR [53,54,58]. RPV is a written communication platform, and the study investigated the feedback system; however, it lacked a description of its association with the EHR [54,58]. The RPV portal provided access to blood test results from patient records [54]. Patients and HCPs could directly view the EHR. RelayHealth allowed both parties to see the messaging history and provided direct access and linkage to the EHR [57]. We were unable to extract data on the relationship between the telephone and videoconferencing platforms [41,44,55,59] or VitelCare [56] and EHRs.

Level of Patient Participation

The participative role of patients receiving hemodialysis was variously described in the 9 included studies (Table 2). In total, 44% (4/9) of the studies on the 2 patient portals described the possibility for written (email) asynchronous 2-way communication [53,54,57,58]. A total of 44% (4/9) of the studies on videoconferencing platforms described the possibility for oral, (partly) preagreed, and synchronous 2-way communication [41,44,55,59]. One study described a home telehealth monitoring unit with concomitant remote care nurse support—oral and synchronous 2-way communication at prearranged times [56]. VitelCare was vaguely described in the included articles. Therefore, we included additional information about VitelCare from the protocol article by Berman et al [62]—VitelCare also allowed for written, asynchronous communication (eg, email).

In all interventions, patients were involved to varying degrees, depending on the degree of patient involvement, based on our chosen taxonomy [30]. Patients were involved at level 1 ("information giving") in 33% (3/9) of the studies [53,54,58]. This 33% (3/9) of the studies examined the RPV patient portal; however, they examined different populations at different periods and had different aims in examining the experience of patients with CKD regarding RPV use [53,54,58]. Therefore, they highlighted different relevant patient experiences. In total, 56% (5/9) of the studies [41,44,55,57,59] involved patients at level 2 ("consultation"), and 11% (1/9) of the studies [56] involved patients engaged in both asynchronous and synchronous communication, rated as level 2.

Facilitators of and Barriers to Patient Participation in eHealth Communication Interventions

Overview

We were able to identify the facilitators of patient participation in all the included studies (Table 3) and the barriers to patient participation in 67% (6/9) of the studies [41,44,53,56,58,59] (Table 4).

XSL•FO
RenderX

Table 3. Facilitators to patient participation. The 4 categories are availability of and access to information; security, trust, and confidence; patient knowledge of their health situation; and preparedness for the future health situation and be able to talk about it with significant others

Study	4 overarching categories of facilitators			
	Availability of and access to information	Security, trust, and confidence	Patient knowledge of their health situation and use for self-care	Preparedness for the future health situation and be able to talk about it with significant others
Chu et al [59], 2023; part 1: from surveys	• To have the necessary resources AND flexible telehealth options • To have a stable internet connection	• To have support for using telehealth • To attend an appointment when it was arranged using telehealth • To be confident using the technology	• Feeling comfortable talking about health issues via telehealth	__[a]
Chu et al [59], 2023; part 2: from interviews	—	• Reduces travel burden • Telehealth consultations are convenient • Provides flexibility	—	• A belief that telehealth benefits the individual • Telehealth reduces stress for HCPs[b] and positively impacts the health care system • Patients want to do what is expected of them
Giles et al [58], 2017; phase 1	• Easy access to a computer • Availability of free-text option	• To feel confident in reporting • Anonymity • Feedback from reports • Being able to give positive feedback	• Leads to improvements	• Encourage openness between staff and patients
Giles et al [58], 2017; phase 1 continued	—	• Being confident that the feedback is going to the right person • Belief that the feedback system will encourage improvement • Independent voice for patients when giving feedback	—	—
Hazara and Bhandari [53], 2016	—	—	• Check blood test results • The availability of RPV[c], particularly during travel, is a safety concern • The availability despite infrequent use	• Be prepared (have knowledge) to discuss their own situation with their physician
Hudson et al [54], 2020	—	—	• Supporting ways of knowing • Extending existing practice: to be in charge • Translating information into practice	• Engaging family, involving carers, and explaining health problems to carers • Engaging HCPs and achieve an individual acceptance of ownership for health
Jakubowski et al [55], 2019	• 1-click access • Technical logistical support from a person	• Discretion (headphones, screens, and separate rooms)	• "It gives me a chance to get some stuff off my back" • Being able to talk one-on-one with someone was helpful	—

XSL•FO
RenderX

Study	4 overarching categories of facilitators			
	Availability of and access to information	Security, trust, and confidence	Patient knowledge of their health situation and use for self-care	Preparedness for the future health situation and be able to talk about it with significant others
Kiberd et al [57], 2018	• The portal was easy to use	—	—	• Positive impact on access to kidney specialist
Ladin et al [41], 2021	• More convenient, less costly, and more efficient	• Reducing infection risk: telehealth consultations reduce face-to-face contact	—	• Care partner engagement facilitates participation in consultations • The clinician is more dedicated during a web-based session • To be relaxed in one's own home
Lew et al [44], 2023	• Easy access to and availability of the videoconference because renal dietitians facilitated them at the bedside	• Not concerned regarding internet security, privacy, or technical issues • Spending enough time with their physician during web-based visits	• Can address issues and obtain answers to medical questions	—
Minatodani and Berman [56], 2013	—	—	• Being empowered and gaining a sense of control • Self-awareness of health and learned illness self-management skills	• A reminder of how the health situation impacts the caregiver

[a]Not applicable.

[b]HCP: health care professional.

[c]RPV: Renal Patient View.

XSL•FO
RenderX

Table 4. Barriers to patient participation. The 4 categories are unavailability of and challenging access to information, mistrust and unsafety, challenges in knowing their health situation and self-care, and relational challenges.

Study	4 overarching categories of barriers			
	Unavailability of and challenging access to information	Mistrust and unsafety	Patients' challenges in knowing about their health situation and self-care	Relational challenges
Chu et al [59], 2023; part 1: from surveys	• Lack of training or instructions on how to use telehealth • Failing to understand how to use telehealth	—[a]	—	—
Chu et al [59], 2023; part 2: from interviews	• Hearing • Language and body language (eg, lip reading)	• Concerns about the ability to perform routine assessments (eg, leg fluid) • Mistrust in the clinician's ability to do a physical examination	• Entrenched negative attitudes and views regarding telehealth	• Depersonalization • HCPs[b] lacking will or training negatively impact use
Giles et al [58], 2017	• Language • Computer literacy and access to a computer	• Desire for an independent person to investigate the reports • Concerns about repercussions for staff if the patient makes a negative report • Perception that the information will not be dealt with or taken seriously	• A negatively loaded name of the intervention • Patients want to forget about dialysis when at home	• Prefer face-to-face contact for feedback • Fear of deteriorating relationships with HCPs • Lack of support from relatives
Hazara and Bhandari [53], 2016	• Design of website • Log-in problems, loss of log-in credentials, and not knowing who to contact • Access to a computer or the internet	• Personal information may be seen by others	• "Seeing my results online makes me more anxious" • Being too busy to use the website • Patients only use the website to check blood test results	• A feeling that the intervention did not add anything to the patient's existing relationship with the HCP
Hudson et al [54], 2020	—	—	—	—
Jakubowski et al [55], 2019	—	—	—	—
Kiberd et al [57], 2018	—	—	—	—
Ladin et al [41], 2021	• Lack of a private, quiet place • Technical and medical literacy challenges • Not knowing who to contact in the event of technical problems	• A clinician's diagnosis is challenging to trust when both parties do not meet face-to-face	• Measuring blood pressure at home is difficult • Hearing loss and loss of face mimics make web-based communication challenging	• Loss of interpersonal connection and feeling alone
Lew et al [44], 2023	—	• Concerns that the lack of a physical examination hampered the physician's ability to care for them	—	—
Minatodani and Berman [56], 2013	• Technical challenges (eg, outdated software)	—	• Feeling too sick, weak, or tired • Forgetting to send measurements	—

XSL•FO
RenderX

[a]Not applicable.

[b]HCP: health care professional.

Facilitators

The first facilitator (Table 3) was related to availability of and access to the intervention and included easy access to computers and 1-click access, support for log-in problems, and less costly and more efficient use. The second facilitator was related to security and confidence and included being confident that the feedback went to the right person, feeling confident about reporting, discretion (eg, a separate room), and reducing infection risk. The third facilitator addressed patient knowledge of their health situation, including checking blood test results, self-awareness of health, and the availability of the patient portal during travel. This facilitator also included self-care skills, such as translating information into practice, one-to-one talks, and gaining a sense of control. The fourth facilitator addressed preparedness related to changes in health situations and skills in relation to others, family and friends, and HCPs.

Barriers

The first barrier (Table 4) addressed the challenges regarding the availability of and access to the intervention and included physical challenges such as hearing, understanding the language and body language, lack of access to computers or the internet, loss of log-in credentials, and lack of a private place. The second barrier was related to mistrust and unsafety and included worries that personal information may be seen by others, concerns about staff repercussions if the patient made a negative report, the perception that the information would not be dealt with, and worries that a clinician's diagnosis would be challenging to trust when patients and physicians do not physically meet. The third barrier concerned knowledge of the health situation and self-care and included a desire to forget about dialysis at home; anxiety about seeing the measured blood pressure on the web would lead; difficulty measuring blood pressure at home; and feeling too sick, weak, or tired. The fourth barrier addressed the relational challenges for people undergoing hemodialysis concerning other people (eg, HCPs and family members). In addition, it addressed that HCPs lacking will or training in facilitating telehealth consultations negatively impacted their use for the patients. The results indicated preferences for face-to-face contact for feedback, lack of support from relatives, a feeling that the intervention did not add anything to the patient's existing relationship with the HCP, and hearing loss and loss of face mimics causing challenges in web-based communication.

Discussion

Principal Findings

This scoping review aimed to identify and map the available evidence on patient participation in eHealth communication interventions. We identified 4 specific patient portals or conferencing platforms and 3 nonspecified digital audio or video solutions used in eHealth consultations, all used and evaluated through various study designs involving patients undergoing hemodialysis. We found that eHealth communication intervention types were inadequately described, including their link to the EHR. The level of patient participation was predominantly low. Key barriers to participation included unavailability of and limited access to information, mistrust and feelings of insecurity, lack of knowledge about their health situation and self-care, and relational challenges. Conversely, key facilitators of participation were availability of and access to information; security, trust, and confidence; patients' knowledge of their health situation and use for self-care; and preparedness for future health events and explain situation to significant others.

Few Possibilities for Patient-Oriented or Customized Digital Communication and Lack of 2-Way Visibility of Information in EHRs

We only identified 4 specified eHealth communication interventions; nevertheless, the small number of interventions provided different opportunities, both asynchronous and synchronous, for 2-way communication. The lack of description of the interventions' concrete opportunities for 2-way communication obscured the participative role of patients undergoing hemodialysis in digital communication. Only Minatodani and Berman [56] described how the monitoring unit offered both asynchronous and synchronous communication; however, they only vaguely described the functionality without a link to the EHR. No further descriptions were found for the 7 included interventions, and the combined possibility for asynchronous and synchronous communication contributed minimally to modeling future 2-way eHealth communication interventions. Therefore, further research is required in this area that aligns well with the general literature on the subject as a complex eHealth landscape [7].

The patients' low participative role, as evidenced by their hesitation to use or be consistent users of eHealth interventions, may be attributed to a perception of digital interventions as complex, as indicated by our barrier analysis. Regarding barriers, we found that both technical challenges and mistrust may contribute to preventing patients undergoing hemodialysis from using eHealth interventions. This is in line with the experience of patients with CKD who initially had access to but did not persist in using their patient portal [28]. Communication must be patient centered to be therapeutic [64]. Communication can result in a more collaborative relationship between patients and HCPs in which the patients' voice is heard and respected. Our results show that patients undergoing hemodialysis desire a therapeutic and collaborative relationship with their HCPs, and we found no reason to assume that the same desire does not apply to digital communication. Patients may feel disempowered and unheard if patient-centeredness is not adopted in digital communication with HCPs. This can lead to a breakdown in trust and a less collaborative patient-provider relationship.

To overcome the barriers to patient participation at a level higher than that identified in this review [30], we propose that patients undergoing hemodialysis need to feel safer using eHealth communication interventions. Communication

XSL•FO
RenderX

misunderstandings have been documented as a major safety risk for patient harm in renal units [65], and patients require reassurance, education, and ongoing support to view eHealth interventions as an aid to improve communication. Developing a culture of patient safety demands multifaceted efforts; nevertheless, the responsibility to systematically ensure patient safety falls on HCPs and the health organization [66,67]. In addition, patients must find it more meaningful to use eHealth communication interventions and should view the interventions as a modality in addition to or aligned with their existing physical relationship with their HCP. A relational barrier to patient participation in eHealth communication was "...a feeling that the intervention did not add anything to the member's existing relationship with HCP" [53]. Thus, the few interventions that offered asynchronous digital communication, such as email correspondence, offered patients minimal opportunities for higher levels of participation [30]. An example of dissatisfaction with 1-way communication is the desire of patients undergoing hemodialysis for 2-way communication and feedback on their ePROMs, highlighting their desire for communication and relationships with HCPs [25,68]. This desire for 2-way communication has also been documented in general patient populations [69].

One important finding is the lack of detailed descriptions of what patients can view in their EHRs and their experiences with this access. In addition, we were unable to draw any conclusions on how patients perceive the information provided in the EHR and their potential requirements for future interventions. This knowledge gap may also include the challenges that patients face in understanding the medical content in the EHR. We found examples of patients not being able to understand the information in the EHR. The barrier "seeing my result online makes me more anxious" [53] may indicate that the patients may require having their results explained by the HCP and discussing their results with them. The benefits of combining the asynchronous and synchronous functionalities of eHealth communication interventions remain undercommunicated, and the aims and functions of physical and digital communication remain underdiscussed.

Low Levels of Patient Participation

The low levels of patient participation (Table 2) can indicate a passive patient role and limited opportunities to be active, ask questions, demand, or make decisions [30]. However, a low level of patient participation can also indicate that patients do not make decisions about their own situation despite the opportunity for more engagement. The barriers and facilitators in this study do not indicate an explicit formula for eHealth communication interventions to increase engagement by all patients. Higher levels of patient involvement (Table 2), as described by Thompson [30], imply 2-way communication. The highest levels include dialogue, shared decisions, and transmission of power and knowledge from the HCP to the patient [30,31]. Our results show that eHealth use may lead to patients' fear of increased distance from HCPs [41]. The low level of patient participation may be related to the power dynamic between patients and professionals [30,31].

Patients' fear of increased distance from HCPs can further lead to a negative power balance, consistent with the results of a scoping review of good practices for dialysis education, treatment, and eHealth [70]. Further conceptualization and development of interventions may continue without real patient participation if patient experience is not accounted for in the development of evaluation instruments. The patients' wish to maintain interactions and relationships with HCPs is in line with the importance of the nurse-patient interaction described by Berman et al [62]. We can assume that feeling safe and secure is a prerequisite for transition from physical to digital nurse-patient interaction. As described previously, the level of patient involvement and the balance of power between patients and HCPs may be negatively affected if eHealth interventions do not achieve their intended purpose: increased autonomy, a stronger patient voice, and more patient power [7]. Knowledge is an undeniable prerequisite for power [71].

Barriers and Facilitators Related to Health Literacy

The EHR is one of the central sources of patient knowledge about their situation, and the lack of access to EHRs limits patient knowledge about their medical situation. Knowledge of one's health is connected to the concept of health literacy, defined as "...the ability of an individual to obtain and translate knowledge and information in order to maintain and improve health in a way that is appropriate to the individual and system contexts" [72]. This definition includes three broad themes that capture the essence of the health literacy concept: (1) knowledge of health, health care, and health systems; (2) processing and using information in various formats in relation to health and health care; and (3) the ability to maintain health through self-management and working in partnership with health providers [72]. We aimed to demonstrate how our findings on barriers and facilitators may provide insights into the multifaceted approaches required to increase patient participation and health literacy in the hemodialysis context [73]. The themes of the barriers and facilitators relate to different competencies in our understanding of the health literacy concept [74].

First, our findings on the barriers to and facilitators of "availability and access to information" and "trusting the security of the interventions" relate to the first concept of health literacy, that is, "to be able to receive knowledge." The results of the RPV patient portal are in line with those of a study of nonusers of the same portal demonstrating that practical barriers (eg, access to the intervention and lack of assistance with the first log-in) are crucial barriers to persistent use [28]. Facilitators of the sustained use of patient portals include initial support and establishment of access to web-based records. The interventions must be aimed to facilitate a user interface arranged for and personalized to the individual user, with content generalized to the population [28]. Patients undergoing hemodialysis may have limited access to technology, such as smartphones or computers, causing difficulty in implementing digital communication tools. Considering that older adults account for most of the population undergoing hemodialysis worldwide [75] and older persons generally have greater challenges using eHealth, it is reasonable to assume that patients undergoing hemodialysis may require more support than other patient groups [76]. In addition, patients

XSL•FO
RenderX

undergoing hemodialysis may have physical (eg, poorer vision and intention tremors) and cognitive (eg, mood and memory problems) limitations, causing difficulty in using certain types of technology [24,77].

HCPs and patients undergoing hemodialysis meet frequently, and HCPs are in a unique position to individualize support and motivation and appear safe and confident to increase both the technical and safe use of eHealth communication interventions among patients. However, HCPs must be awarded credit by their employers for their resources and time to perform this important part of their daily workflow, thereby facilitating alternative approaches to physical meetings. Access to EHRs by both patients and HCPs could contribute to shared knowledge and improved relationships. The low level of participation among patients undergoing hemodialysis in this study correlates with the self-reported level of health literacy among the population in a Norwegian study, in which older persons and persons with chronic diseases had a lower level of health literacy [78]. We can assume that this result is comparable with those in other countries. Therefore, efforts to provide the population undergoing hemodialysis with digital possibilities to acquire knowledge about their own health situation may demand more resources than those for other populations (eg, younger populations).

Second, we found that the barriers to and facilitators of "security and trust" are related partly to the first concept of health literacy and partly to the second concept, that is, "processing and using the information." Discretion is an explicit barrier to patient participation. Discretion is a challenge when discussing sensitive information in an hemodialysis unit where several patients concurrently undergo hemodialysis. The possibility of digitally exchanging information and knowledge that can be read and processed in private, combined with digital or physical consultations, may be a fruitful solution. A special challenge when promoting digital communication for patients undergoing hemodialysis is their close relationship with their HCPs, and digital communication must offer value that is not obtained through physical meetings. Both patients and HCPs must find worthy interventions. Moreover, the sharing of data about health situations, treatment, results, and challenges can facilitate a higher level of knowledge, which is related to a higher level of patient participation [30] and health literacy [72]. In addition, a higher level of knowledge can facilitate shared understanding through further face-to-face or digital discussions.

Third, we found that the barriers to and facilitators of "preparedness and relation to others" related to the third concept of health literacy: "ability to maintain health through self-management and working in partnerships with health providers." An explicit barrier was that the intervention did not add anything positive to the existing relationship with the HCP [53,58]. The relationships between patients undergoing hemodialysis and HCPs are long term, and patients have a dependent relationship with the HCP. This correlates with a review describing that older people with multimorbidity regard a lack of self-efficacy, knowledge, support, functionality, and information provision regarding the benefits of eHealth as the main barriers to using eHealth [29]. Our review confirmed that important facilitators for improving patient participation include active engagement of target end users in the design and delivery of eHealth programs, support for overcoming privacy concerns, and enhancement of self-efficacy in the use of technology. The integration of eHealth applications and programs across health services to accommodate the patient group of older adults with multimorbidity may be a key facilitator [29].

Patient involvement was challenged for various reasons, including technical issues, safety, self-care, and relational factors, as identified and presented in the overarching categories of barriers and facilitators (Tables 3 and 4). This means that challenges that hinder increased patient involvement require action at different levels. Similarly, a multifaceted approach is required to address the challenges within all 3 broad themes and levels of health literacy. This is in line with the understanding that health literacy encompasses different levels and skills and can be developed through education, particularly if it is well conceptualized and context specific [79]. We support an encompassing approach that includes overcoming barriers and strengthening facilitators to support greater autonomy and personal empowerment in the hemodialysis context.

Limitations

This scoping review solely identified and mapped the current evidence regarding patient participation in eHealth communication interventions. The findings do not offer recommendations for practice or research but rather highlight research gaps and inspire future studies. According to the Joanna Briggs Institute scoping review guidelines [33], assessing the methodological quality of the included studies was not applicable. However, this review's primary strength is its comprehensive systematic searches across multiple databases. Despite identifying only 9 studies on eHealth interventions, we are confident that these studies represent the relevant published sources on patient participation in eHealth communication interventions. Adherence to the standardized a priori protocol [31] and the PRISMA-ScR reporting guidelines [46] aided in ensuring consistency and transparency in the review. One limitation may be that the studies' descriptions of patient involvement were sparse, and we may have misjudged the degree of patient participation. To counteract this, we relied on supplementary sources for more detailed descriptions of the interventions. The predominance of qualitative designs may preclude obtaining valid and honest answers, particularly during oral evaluations. This is particularly true for the frail population of older adults undergoing hemodialysis, who may feel dependent on their relationships with HCPs [55]. All the studies reported facilitators (9/9, 100%) but did not mention barriers (3/9, 33%). However, participants in qualitative research projects may find it challenging to provide objective and explicitly negative evaluations orally or face-to-face.

Implications for Future Research and Practice

This review highlighted the current understanding and gaps in knowledge regarding patient participation in eHealth communication interventions for patients undergoing hemodialysis. We emphasize the need for diverse research methodologies, including qualitative, quantitative, and mixed methods, to explore patient experiences and evaluate the use of these interventions. Research should inform the design,

XSL•FO
RenderX

innovation, and implementation of eHealth projects, ensuring compliance with ethical considerations, to uncover both positive and negative feedback. The needs of patients may vary across different stages of CKD and types of renal replacement therapy, presenting challenges when diverse needs are assessed within the same group. Stakeholder perspectives, including patients, HCPs, and caregivers, should be integrated into cocreational processes to develop accessible and safe eHealth communication models. Research-based interventions should systematically address barriers and enhance patient participation across various health literacy themes. In addition, the health care system must facilitate eHealth communication as part of HCPs' daily workflow supplemented by face-to-face interactions. Finally, additional research is required to fulfill the aim of increased HHD.

Conclusions

Patients undergoing hemodialysis require 2-way communication and relationships with HCPs. Low patient participation is associated with low levels of health literacy in other older populations with chronic illnesses. Working at different levels to overcome barriers to patient participation can facilitate higher levels of patient participation and health literacy. HCPs should be the preferred facilitators of eHealth communication with patients undergoing hemodialysis to improve feelings of safety and security. Facilitating shared knowledge, patient participation, and health literacy is vital to improving the visibility of EHR data for patients.

Acknowledgments

The authors gratefully acknowledge senior librarian Malin Elisabeth Norman from Nord University, Norway, for her invaluable expertise in literature search and associate professor Jorunn Bjerkan, also from Nord University, Norway, for her profound knowledge in the eHealth field and her significant contributions to this study. This study had no external funding.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

AD, CFM, and MSL contributed to conceptualization and methodology. AD performed the data curation and writing of the original draft preparation. AD and MSL contributed to the visualization and investigation. CFM and MSL contributed with supervision. AD and CFM performed the formal analysis, investigation, resources, and validation. AD, CFM, and MSL contributed with writing—review and editing. AD, CFM, and MSL agree to be accountable for all aspects of the work and finally approve the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 121 KB - jmir_v26i1e51900_app1.pdf]

References

1. Gadaen RJ, Kooman JP, Cornelis T, van der Sande FM, Winkens BJ, Broers NJ. The effects of chronic dialysis on physical status, quality of life, and arterial stiffness: a longitudinal study in prevalent dialysis patients. Nephron 2021;145(1):44-54 [FREE Full text] [doi: 10.1159/000510624] [Medline: 33108785]
2. Bello AK, Okpechi IG, Osman MA, Cho Y, Htay H, Jha V, et al. Epidemiology of haemodialysis outcomes. Nat Rev Nephrol 2022 Jun;18(6):378-395 [FREE Full text] [doi: 10.1038/s41581-022-00542-7] [Medline: 35194215]
3. Miller AJ, Perl J, Tennankore KK. Survival comparisons of intensive vs. conventional hemodialysis: pitfalls and lessons. Hemodial Int 2018 Jan;22(1):9-22. [doi: 10.1111/hdi.12559] [Medline: 28425578]
4. Li PK, Cheung WL, Lui SL, Blagg C, Cass A, Hooi LS, et al. Increasing home based dialysis therapies to tackle dialysis burden around the world: a position statement on dialysis economics from the 2nd Congress of the International Society for Hemodialysis. Nephrology (Carlton) 2011 Jan;16(1):53-56. [doi: 10.1111/j.1440-1797.2010.01418.x] [Medline: 21175978]
5. Roberti J, Cummings A, Myall M, Harvey J, Lippiett K, Hunt K, et al. Work of being an adult patient with chronic kidney disease: a systematic review of qualitative studies. BMJ Open 2018 Sep 04;8(9):e023507 [FREE Full text] [doi: 10.1136/bmjopen-2018-023507] [Medline: 30181188]
6. Wang CS, Ku E. eHealth in kidney care. Nat Rev Nephrol 2020 Jul;16(7):368-370 [FREE Full text] [doi: 10.1038/s41581-020-0271-z] [Medline: 32238923]

XSL•FO
RenderX

7. Horgan D, Hackett J, Westphalen CB, Kalra D, Richer E, Romao M, et al. Digitalisation and COVID-19: the perfect storm. Biomed Hub 2020 Sep 17;5(3):1341-1363 [FREE Full text] [doi: 10.1159/000511232] [Medline: 33564668]
8. Choudhuri A, Chatterjee JM, Garg S. Internet of things in healthcare: a brief overview. In: Internet of Things in Biomedical Engineering. Cambridge, MA: Academic Press; 2019.
9. Samir A, Johansen HD. Technical viewpoint of challenges, opportunities, and future directions of policy change and information-flow in digital healthcare systems. In: Proceedings of the Fourteenth International Conference on eHealth, Telemedicine, and Social Medicine. 2022 Presented at: eTELEMED 2022; June 26-30, 2022; Porto, Portugal.
10. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2020 Feb 29;395(10225):709-733. [doi: 10.1016/S0140-6736(20)30045-3] [Medline: 32061315]
11. Huijben JA, Kramer A, Kerschbaum J, de Meester J, Collart F, Arévalo OL, et al. Increasing numbers and improved overall survival of patients on kidney replacement therapy over the last decade in Europe: an ERA Registry study. Nephrol Dial Transplant 2023 Mar 31;38(4):1027-1040. [doi: 10.1093/ndt/gfac165] [Medline: 35974693]
12. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
13. Engels N, de Graav G, van der Nat P, van den Dorpel M, Bos WJ, Stiggelbout AM. Shared decision-making in advanced kidney disease: a scoping review protocol. BMJ Open 2020 Feb 27;10(2):e034142 [FREE Full text] [doi: 10.1136/bmjopen-2019-034142] [Medline: 32111615]
14. Himmelfarb J, Ikizler TA. Hemodialysis. N Engl J Med 2010 Nov 04;363(19):1833-1845. [doi: 10.1056/NEJMra0902710] [Medline: 21047227]
15. Chuasuwan A, Pooripussarakul S, Thakkinstian A, Ingsathit A, Pattanaprateep O. Comparisons of quality of life between patients underwent peritoneal dialysis and hemodialysis: a systematic review and meta-analysis. Health Qual Life Outcomes 2020 Jun 18;18(1):191 [FREE Full text] [doi: 10.1186/s12955-020-01449-2] [Medline: 32552800]
16. Moreira T. The Transformation of Contemporary Health Care: The Market, the Laboratory, and the Forum. Milton Park, UK: Routledge; 2013.
17. Digital hjemmeoppfølging. Helsedirektoratet. URL: https://www.helsedirektoratet.no/faglige-rad/digital-hjemmeoppfolging/digital-hjemmeoppfolging [accessed 2022-11-15]
18. Castro EM, Van Regenmortel T, Vanhaecht K, Sermeus W, Van Hecke A. Patient empowerment, patient participation and patient-centeredness in hospital care: a concept analysis based on a literature review. Patient Educ Couns 2016 Dec;99(12):1923-1939. [doi: 10.1016/j.pec.2016.07.026] [Medline: 27450481]
19. Camedda C, Bici G, Magi CE, Guzzon A, Longobucco Y. The therapeutic nurse-patient relationship in hemodialysis: a pilot mixed-method study on the perceived quality of nurses' attitudes and caring behaviors. Nurs Rep 2023 Jul 20;13(3):990-1003 [FREE Full text] [doi: 10.3390/nursrep13030087] [Medline: 37489409]
20. Lazenby S, Edwards A, Samuriwo R, Riley S, Murray MA, Carson-Stevens A. End-of-life care decisions for haemodialysis patients - 'we only tend to have that discussion with them when they start deteriorating'. Health Expect 2017 Apr;20(2):260-273 [FREE Full text] [doi: 10.1111/hex.12454] [Medline: 26968338]
21. Schrauben SJ, Appel L, Rivera E, Lora CM, Lash JP, Chen J, et al. Mobile health (mHealth) technology: assessment of availability, acceptability, and use in CKD. Am J Kidney Dis 2021 Jun;77(6):941-50.e1 [FREE Full text] [doi: 10.1053/j.ajkd.2020.10.013] [Medline: 33309860]
22. Yang Y, Chen H, Qazi H, Morita PP. Intervention and evaluation of mobile health technologies in management of patients undergoing chronic dialysis: scoping review. JMIR Mhealth Uhealth 2020 Apr 03;8(4):e15549 [FREE Full text] [doi: 10.2196/15549] [Medline: 32242823]
23. Kluge EH. Framework considerations. In: The Electronic Health Record: Ethical Considerations. Cambridge, MA: Academic Press; Jan 2020.
24. Buritica JP. Working together when apart. In: The Holloway Guide to Remote Work. London, UK: Holloway, Incorporated; 2020.
25. Flythe JE, Tugman MJ, Narendra JH, Dorough A, Hilbert J, Assimon MM, et al. Feasibility of tablet-based patient-reported symptom data collection among hemodialysis patients. Kidney Int Rep 2020 Jul;5(7):1026-1039 [FREE Full text] [doi: 10.1016/j.ekir.2020.04.021] [Medline: 32647760]
26. Van Der Veer SN, Ercia A, Caskey FJ, Farrington K, Jury F, Rees M, et al. Developing an intervention to implement electronic patient-reported outcomes in renal services in the UK. Stud Health Technol Inform 2020 Jun 16;270:936-940. [doi: 10.3233/SHTI200299] [Medline: 32570519]
27. Lew SQ, Wallace EL, Srivatana V, Warady BA, Watnick S, Hood J, et al. Telehealth for home dialysis in COVID-19 and beyond: a perspective from the American Society of Nephrology COVID-19 home dialysis subcommittee. Am J Kidney Dis 2021 Jan;77(1):142-148 [FREE Full text] [doi: 10.1053/j.ajkd.2020.09.005] [Medline: 33002530]
28. Phelps RG, Taylor J, Simpson K, Samuel J, Turner AN. Patients' continuing use of an online health record: a quantitative evaluation of 14,000 patient years of access data. J Med Internet Res 2014 Oct 24;16(10):e241 [FREE Full text] [doi: 10.2196/jmir.3371] [Medline: 25344621]

XSL•FO
RenderX

29. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
30. Thompson AG. The meaning of patient involvement and participation in health care consultations: a taxonomy. Soc Sci Med 2007 Mar;64(6):1297-1310. [doi: 10.1016/j.socscimed.2006.11.002] [Medline: 17174016]
31. Deinboll A, Moe CF, Ludvigsen MS. Interventions, participative role, barriers, and facilitators for involvement in eHealth communication for people undergoing hemodialysis: protocol for a scoping review. JMIR Res Protoc 2022 Jul 29;11(7):e38615 [FREE Full text] [doi: 10.2196/38615] [Medline: 35904862]
32. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
33. Pollock D, Peters MD, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 01;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
34. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
35. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
36. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: a tool for transparent and efficient forward and backward citation chasing in systematic searching. Res Synth Methods 2022 Jul 07;13(4):533-545. [doi: 10.1002/jrsm.1563] [Medline: 35472127]
37. EndNote X9. Clarivate Analytics. 2013. URL: https://support.clarivate.com/Endnote/s/article/EndNote-X9-Updates?language=en_US [accessed 2024-10-16]
38. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
39. Coats H, Shive N, Doorenbos AZ, Schmiege SJ. Integration of person-centered narratives into the electronic health record: study protocol. Nurs Res 2020;69(6):483-489 [FREE Full text] [doi: 10.1097/NNR.0000000000000463] [Medline: 32740306]
40. Dawson J, Campbell KL, Craig JC, Tong A, Teixeira-Pinto A, Brown MA, et al. A text messaging intervention for dietary behaviors for people receiving maintenance hemodialysis: a feasibility study of KIDNEYTEXT. Am J Kidney Dis 2021 Jul;78(1):85-95.e1. [doi: 10.1053/j.ajkd.2020.11.015] [Medline: 33421456]
41. Ladin K, Porteny T, Perugini JM, Gonzales KM, Aufort KE, Levine SK, et al. Perceptions of telehealth vs in-person visits among older adults with advanced kidney disease, care partners, and clinicians. JAMA Netw Open 2021 Dec 01;4(12):e2137193 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.37193] [Medline: 34870680]
42. Duncanson E, Bennett PN, Viecelli A, Dansie K, Handke W, Tong A, et al. Feasibility and acceptability of e-PROMs data capture and feedback among patients receiving haemodialysis in the Symptom monitoring WIth Feedback Trial (SWIFT) pilot: protocol for a qualitative study in Australia. BMJ Open 2020 Nov 06;10(11):e039014 [FREE Full text] [doi: 10.1136/bmjopen-2020-039014] [Medline: 33158824]
43. Schick-Makaroff K, Molzahn AE. Evaluation of real-time use of electronic patient-reported outcome data by nurses with patients in home dialysis clinics. BMC Health Serv Res 2017 Jun 26;17(1):439 [FREE Full text] [doi: 10.1186/s12913-017-2377-y] [Medline: 28651528]
44. Lew SQ, Kaur G, Sikka N, Erickson KF. In-center hemodialysis unit patient experience with telehealth. Hemodial Int 2023 Apr;27(2):193-196 [FREE Full text] [doi: 10.1111/hdi.13065] [Medline: 36744583]
45. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
46. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
47. Fouda T, Yasin F, Aly S, Singh PR, Mohammed ME, Elesnawi M, et al. MO851: impact of COVID-19 on dialysis patients: reported experience in Qatar. Nephrol Dial Transplant 2022;37(Supplement_3):gfac083.033. [doi: 10.1093/ndt/gfac083.033]
48. Raduege TJ, Thomson Reuters Accelus. Healthcare reform: delivery reform. Issue Brief Health Policy Track Serv 2015 Dec 28:1-82. [Medline: 27116776]
49. Porteny T, Koch-Weser S, Rifkin DE, Isakova T, Gordon EJ, Wong JB, et al. Decision-making during uncertain times: a qualitative study of kidney patients, care partners, and nephrologists during the COVID-19 pandemic: PO0057. J Am Soc Nephrol 2021;32(10S):73. [doi: 10.1681/ASN.20213210S173a]
50. Ong SW, Jassal V, Seto E, So S, Uddin A, Shier A, et al. Mobile self-management system for CKD patients: conceptual framework and design principles. Am J Kidney Dis 2014 May;63(5):B85. [doi: 10.1053/j.ajkd.2014.01.285]
51. Oni O, Goel A, Savin V, Gandy K, Witten E. Design of pilot study to improve quality-of-life in chronic diseases using custom web-based educational and monitoring platform. J Investig Med 2015;63:670-671.

XSL•FO
RenderX

52. Patient experience. Bringing care closer. Health Serv J 2009 Oct 29;Suppl:9. [Medline: 20128061]
53. Hazara AM, Bhandari S. Barriers to patient participation in a self-management and education website Renal PatientView: a questionnaire-based study of inactive users. Int J Med Inform 2016 Mar;87:10-14. [doi: 10.1016/j.ijmedinf.2015.12.004] [Medline: 26806707]
54. Hudson C, Darking M, Cox J. Understanding the value of Patientview for enabling self-care practice in chronic kidney disease. J Ren Care 2020 Mar 16;46(1):13-24. [doi: 10.1111/jorc.12300] [Medline: 31621197]
55. Jakubowski KP, Jhamb M, Yabes J, Gujral S, Oberlin LE, Bender FH, et al. Technology-assisted cognitive-behavioral therapy intervention for end-stage renal disease. Transl Behav Med 2020 Aug 07;10(3):657-663 [FREE Full text] [doi: 10.1093/tbm/ibz077] [Medline: 31131853]
56. Minatodani DE, Berman SJ. Home telehealth in high-risk dialysis patients: a 3-year study. Telemed e-Health 2013 Jul;19(7):520-522. [doi: 10.1089/tmj.2012.0196]
57. Kiberd J, Khan U, Stockman C, Radhakrishnan A, Phillips M, Kiberd BA, et al. Effectiveness of a web-based eHealth portal for delivery of care to home dialysis patients: a single-arm pilot study. Can J Kidney Health Dis 2018 Sep 07;5:2054358118794415 [FREE Full text] [doi: 10.1177/2054358118794415] [Medline: 30210802]
58. Giles SJ, Reynolds C, Heyhoe J, Armitage G. Developing a patient-led electronic feedback system for quality and safety within Renal PatientView. J Ren Care 2017 Mar;43(1):37-49. [doi: 10.1111/jorc.12186] [Medline: 27990782]
59. Chu G, Silva C, Adams K, Chacko B, Attia J, Nathan N, et al. Exploring the factors affecting home dialysis patients' participation in telehealth-assisted home visits: a mixed-methods study. J Ren Care 2024 Jun 12;50(2):128-137. [doi: 10.1111/jorc.12475] [Medline: 37434485]
60. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
61. PatientView homepage. PatientView. URL: https://www.patient-view.com/ [accessed 2023-07-11]
62. Berman SJ, Wada C, Minatodani D, Halliday T, Miyamoto R, Lindo J, et al. Home-based preventative care in high-risk dialysis patients: a pilot study. Telemed J E Health 2011 May;17(4):283-287. [doi: 10.1089/tmj.2010.0169] [Medline: 21480788]
63. What you need to know about digital health systems. World Health Organization. 2019 Feb 5. URL: https://www.who.int/news/item/05-02-2019-what-you-need-to-know-about-digital-health-systems [accessed 2021-03-16]
64. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021 Sep 03;20(1):158 [FREE Full text] [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
65. Garrick R, Kliger A, Stefanchik B. Patient and facility safety in hemodialysis: opportunities and strategies to develop a culture of safety. Clin J Am Soc Nephrol 2012 Apr;7(4):680-688 [FREE Full text] [doi: 10.2215/CJN.06530711] [Medline: 22282480]
66. Lov om helsepersonell m.v. (helsepersonelloven). Lovdata. URL: https://lovdata.no/dokument/NL/lov/1999-07-02-64 [accessed 2020-06-24]
67. Lov om spesialisthelsetjenesten m.m. (spesialisthelsetjenesteloven). Lovdata. URL: https://lovdata.no/lov/1999-07-02-61 [accessed 2023-07-11]
68. Knowles SE, Ercia A, Caskey F, Rees M, Farrington K, Van der Veer SN. Participatory co-design and normalisation process theory with staff and patients to implement digital ways of working into routine care: the example of electronic patient-reported outcomes in UK renal services. BMC Health Serv Res 2021 Jul 18;21(1):706 [FREE Full text] [doi: 10.1186/s12913-021-06702-y] [Medline: 34273978]
69. Aiyegbusi OL. Key methodological considerations for usability testing of electronic patient-reported outcome (ePRO) systems. Qual Life Res 2020 Feb;29(2):325-333 [FREE Full text] [doi: 10.1007/s11136-019-02329-z] [Medline: 31691202]
70. van Eck van der Sluijs A, Vonk S, van Jaarsveld BC, Bonenkamp AA, Abrahams AC. Good practices for dialysis education, treatment, and eHealth: a scoping review. PLoS One 2021;16(8):e0255734 [FREE Full text] [doi: 10.1371/journal.pone.0255734] [Medline: 34379654]
71. Hattaway M. A Companion To English Renaissance Literature and Culture. Chichester, UK: Wiley-Blackwell; 2002.
72. Liu C, Wang D, Liu C, Jiang J, Wang X, Chen H, et al. What is the meaning of health literacy? A systematic review and qualitative synthesis. Fam Med Community Health 2020 May 14;8(2):e000351 [FREE Full text] [doi: 10.1136/fmch-2020-000351] [Medline: 32414834]
73. O Riordan J, Kane PM, Noble H, Smyth A. Advance care planning and health literacy in older dialysis patients: qualitative interview study. BMJ Support Palliat Care 2021 Nov 15:bmjspcare-2021-003273. [doi: 10.1136/bmjspcare-2021-003273] [Medline: 34782344]
74. Mackert M, Champlin SE, Holton A, Muñoz II, Damásio MJ. eHealth and health literacy: a research methodology review. J Comput Mediat Commun 2014 Apr 12;19(3):516-528. [doi: 10.1111/jcc4.12044]
75. Ladin K, Neckermann I, D'Arcangelo N, Koch-Weser S, Wong JB, Gordon EJ, et al. Advance care planning in older adults with CKD: patient, care partner, and clinician perspectives. JASN 2021 Jun;32(6):1527-1535. [doi: 10.1681/asn.2020091298]
76. Rockmann R, Gewald H. Elderly people in eHealth: who are they? Procedia Comput Sci 2015;63:505-510. [doi: 10.1016/j.procs.2015.08.376]

XSL•FO
RenderX

77. Wu HH, Nixon AC, Dhaygude AP, Jayanti A, Mitra S. Is home hemodialysis a practical option for older people? Hemodial Int 2021 Oct;25(4):416-423. [doi: 10.1111/hdi.12949] [Medline: 34133069]
78. Le C, Finbråten HS, Pettersen KS, Joranger P, Guttersrud Ø. Befolkningens helsekompetanse, del I. Helsedirektoratet. 2019. URL: https://tinyurl.com/2stvpfxt [accessed 2024-10-31]
79. Nutbeam D. Defining and measuring health literacy: what can we learn from literacy studies? Int J Public Health 2009 Jul 30;54(5):303-305. [doi: 10.1007/s00038-009-0050-x] [Medline: 19641847]

Abbreviations

CKD: chronic kidney disease
EHR: electronic health record
ePROM: electronic patient-reported outcome measure
HCP: health care professional
HHD: home hemodialysis
PD: peritoneal dialysis
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QOL: quality of life
RPV: Renal Patient View
RQ: research question

Edited by A Mavragani; submitted 16.08.23; peer-reviewed by N Koizumi, G Martins, A Poliani; comments to author 09.05.24; revised version received 12.08.24; accepted 25.09.24; published 11.11.24.

Please cite as:
Deinboll A, Moe CF, Ludvigsen MS
Participation in eHealth Communication Interventions Among Patients Undergoing Hemodialysis: Scoping Review
J Med Internet Res 2024;26:e51900
URL: https://www.jmir.org/2024/1/e51900
doi:10.2196/51900
PMID:

©Anne Deinboll, Cathrine Fredriksen Moe, Mette Spliid Ludvigsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effects of Virtual Reality–Based Reminiscence Therapies for Older Adults With Cognitive Impairment: Systematic Review

Qian Mao[1]; Zhen Zhao[1]; Lisha Yu[1,2], PhD; Yang Zhao[3], PhD; Hailiang Wang[1], PhD

[1]School of Design, The Hong Kong Polytechnic University, Hong Kong, China

[2]Division of Artificial Intelligence, Lingnan University, Hong Kong, China

[3]School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, China

Corresponding Author:
Hailiang Wang, PhD
School of Design
The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom
Hong Kong, 999077
China
Phone: 852 27665471
Email: hailiang.wang@polyu.edu.hk

Abstract

Background: Reminiscence therapy (RT) is a commonly used nonpharmaceutical treatment for cognitive impairment. Artifacts or conversations are used in RT to recall individuals' memories and past experiences. Virtual reality (VR) has increasingly been used as an assistive technology during RT. However, the effects of VR-based RT (VR-RT) methods remain unclear, and insights into the related benefits and challenges are urgently needed.

Objective: The study aims to systematically review the effects of VR-RTs for older adults with cognitive impairment.

Methods: Seven databases (MEDLINE, Academic Search Premier, CINAHL, Web of Science, PubMed, the Cochrane Central Register of Controlled Trials, and ScienceDirect) were searched to identify relevant articles published from inception to August 10, 2023. Peer-reviewed publications that assessed the effect of VR-RTs (ie, using virtual clues to evoke participants' memories or past experiences) on cognitive-related outcomes were included. Two independent researchers conducted the literature search, review, and data extraction processes. A narrative synthesis approach was used to analyze the extracted data.

Results: Of the 537 identified articles, 22 were ultimately included in the data analysis. The results revealed that VR-RTs could maintain cognitive status (4/4, 100%) and reduce anxiety (2/2, 100%) in older adults with cognitive impairment. Nevertheless, one study found a cognitive improvement after VR-RTs, whereas cognitive degradation was observed at a 3- to 6-month follow-up measure. Around 88% (7/8) of the included studies indicated that VR-RTs improved memory; however, the evidence regarding the beneficial effects of VR-RTs was limited in improving quality of life (1/4, 25%) and reducing apathy (0/2, 0%) and depression (1/3, 33%). The results indicated that VR-RTs are safe, engaging, acceptable, and satisfying for older adults with cognitive impairment. In VR scenarios, personalized stimulus materials related to the users' youth experiences were more effective for treating cognitive impairment than other stimulus materials.

Conclusions: The results of this systematic review demonstrate the potential benefits of VR-RT for older adults with cognitive impairment, especially in improving emotion and memory and maintaining cognitive status. VR-RT is also safe and enjoyable for older adults. However, due to the trial heterogeneity of included studies, we can only provide qualitative results instead of performing meta-analysis to quantify the effect size of VR-RTs. Thus, more randomized controlled trials are required to examine the designs and effects of VR-RTs for groups of older adults with specific needs.

(J Med Internet Res 2024;26:e53348) doi:10.2196/53348

KEYWORDS

virtual reality; reminiscence therapy; cognitive impairment; older adults; mobile phone

XSL•FO
RenderX

Introduction

Cognitive Impairment

The World Health Organization predicts that by 2050, the proportion of older adults worldwide will double and thus account for approximately a quarter of the global population [1]. Many older adults experience cognitive impairment, such as losing memory or other cognitive abilities (eg, language, visual, and spatial perception), as they age [2]. Approximately 5% to 10% of older adults diagnosed with mild cognitive impairment (MCI) develop dementia each year [3], a condition that negatively affects their neuropsychiatric health, as exemplified by the manifestation of anxiety, depression, social isolation, and sleep disorders [4]. Dementia also severely affects the independence of older adults and their ability to function in daily life, thus reducing their quality of life [5]. Dementia further places substantial socioeconomic and care burdens on the family members of affected individuals, health care systems, and even societies. The total global cost associated with dementia, which affects 55.2 million people worldwide, is now estimated to be US $800 billion per year and is predicted to increase to US $2 trillion by 2030 [6]. Approximately 50% of these costs can be attributed to informal care, who were found to deliver an average of 5 hours of care and supervision per day [7]. Therefore, it is becoming increasingly important to address cognitive impairment and defer its progression through timely diagnosis and the provision of effective interventions at an early stage.

Traditional interventions for cognitive impairment include both pharmacological and nonpharmacological therapies. Pharmacological therapies (eg, atypical antipsychotic medications) are major interventions for older adults with cognitive impairment and aim to alleviate symptoms rather than the underlying disease processes [8]. Research efforts have been expanded to nonpharmacological therapies, which can slow functional decline, improve apathy, reduce disruptive behaviors, and slow the progression from cognitive impairment to dementia [9]. Among these, reminiscence therapy (RT) involves the organized discussion and recall of past life experiences guided by props such as videos, pictures, locations, and objects [10]. This approach may improve memory and depressive symptoms in older adults with cognitive impairment [11]. For example, some studies have shown the beneficial effects of personal reminiscing namely, mental stimulation, social interaction, and emotional connections, which can improve older adults' mood, self-esteem, and overall quality of life [12,13]. Older adults can easily perform such RT interventions, and they can be administered by therapists or researchers in various settings [14]. However, traditional methods of RT have some drawbacks that have limited its widespread application. For example, RT requires face-to-face interaction between the therapist and participant during treatment sessions, which may be limited by the available physical space [15]. Another potential limitation is the availability of the interactive resources (eg, personalized props and scenes) needed to offer a highly immersive environment that can evoke emotional responses during the interactions [16]. Thus, some researchers are attempting to support RT interventions by incorporating virtual reality (VR) technologies.

VR-Based RTs

VR refers to an immersive and interactive computer-generated environment that simulates a real or imagined experience [17]. A VR user wears a head-mounted display that provides auditory, visual, and tactile feedback to enhance their sense of presence in the virtual world [18]. In the context of cognitive impairment, VR technology has emerged as a promising intervention and support tool that can enhance cognitive abilities and promote the engagement of older adults [19]. Recent work has underlined the benefits of VR technology in interventions for cognitive impairment. For example, Dockx et al [20] indicated that VR technology's immersive and interactive nature enhanced users' engagement, enjoyment, and therapy adherence. Tarnanas et al [21] emphasized that VR technology offers a safe and controlled environment for practicing cognitive skills as it eliminates the risks associated with real-world activities. VR technology also gives therapists much control over the therapy experience. Liu et al [22] showed that VR technology has surpassed the traditional paper-and-pencil assessment paradigm and can capture users' behavioral data, thus allowing the accurate assessment and identification of people with cognitive impairment. Through controlled experiments, Liao et al [23] found that VR technology significantly improved executive function in older adults with cognitive impairment (P=.03). In summary, the controlled environment, tailored training, immediate feedback, and engagement provided through VR interventions contribute to the cognitive-related benefits among older adults.

Given these effects of VR technology in the context of cognitive impairment interventions, many researchers have combined VR technology with RT, and their studies have provided new evidence for the effectiveness of VR-based RTs (VR-RTs). VR-RTs can improve memory recall by providing cues and prompts (eg, movies, videos, images, or 3D-rendered scenes) that trigger specific memories in the user. Its immersive nature can enhance the retrieval of detailed and vivid memories, thus stimulating cognitive processes and strengthening neural connections [24]. For example, Xu and Wang [25] used a VR system to restore life scenes in China in the 1970s. They demonstrated that this VR-based RT (VR-RT) intervention was highly effective and acceptable among older adults with cognitive impairment. In addition, Saredakis et al [26] found that VR-RT could reduce symptoms of isolation and apathy among older adults with cognitive impairment. Hence, VR-RTs have great potential to alleviate memory loss and promote communication, cognition, and positive mood.

Prior reviews have focused predominantly on either VR technology or RT as interventions for cognitive impairment and synthesized the existing VR technology landscape or the impacts of RT interventions on individuals with cognitive impairment [22,27,28]. To the best of our knowledge, only the study by Reisinho et al [29] has reviewed the applications, benefits, and limitations of VR-RTs for the cognitive rehabilitation of people with cognitive impairment or dementia. However, that review paid little attention to the VR-RT stimulus materials and construction methods, synthesis of the effectiveness of VR-RTs for older adults with cognitive impairment, or user satisfaction. Attention to such aspects is required to optimize the design and

XSL•FO
RenderX

implementation of VR-based cognitive rehabilitation programs, enhance users' engagement, and improve users' cognitive functioning.

We conducted a systematic review to explore and synthesize the evidence of the effects and applicability of VR-RTs among older adults with cognitive impairment. Specifically, this review summarizes the materials, construction methods, and intervention designs of existing VR-RTs. Then, we review and analyze the effectiveness of VR-RTs in improving the quality of life and cognitive-related functions of older adults with cognitive impairment. Finally, we discuss the user satisfaction and side effects of VR-RTs. The findings of this review have led us to provide valuable recommendations for researchers, designers, and caregivers who use VR-RTs to care for older adults with cognitive impairment.

Methods

Search Strategy

Seven electronic databases namely, MEDLINE, Academic Search Premier, CINAHL, Web of Science, PubMed, the Cochrane Central Register of Controlled Trials, and ScienceDirect, were searched for relevant articles up to August 10, 2023.

We used the following search terms and search strategy for potential studies related to VR-based RTs among older adults with cognitive impairments: (virtual reality OR virtual OR VR OR virtual simulation) AND (reminiscence OR narrative OR life review OR life story book) AND (dementia OR cognitive impairment OR MCI OR Alzheimer's disease OR memory disorder OR cognitive decline OR cognitive dysfunction OR memory impairment).

Study Selection

Studies were screened on the basis of the inclusion and exclusion criteria, as shown in Textbox 1. A 3-step screening process was implemented to ensure the comprehensiveness and validity of our systematic review. First, 2 authors (QM and ZZ) independently identified relevant articles during the initial review by screening the titles and abstracts. Second, the full texts of the relevant articles were reviewed to determine their eligibility according to the predefined inclusion and exclusion criteria. Third, the reference lists of the included articles were manually examined to identify any missed studies, and the Google Scholar database was manually searched using search terms and search strategy to identify additional potentially relevant studies. The authors independently assessed the articles in all review steps. Any discrepancies regarding study inclusion were discussed with a third author (HW) until a final decision was reached.

Textbox 1. The inclusion and exclusion criteria for screening studies related to virtual reality–based reminiscence therapies (RT) among older adults with cognitive impairments.

Inclusion criteria

- Article type: used RT, a psychosocial intervention using prompts to evoke participants' memories or discussing participants' past events and experiences, and virtual reality (VR) technology; evaluated the effects of VR-based RTs on individuals with cognitive impairment; published in a peer-reviewed journal or conference proceedings.
- Language: written in English.

Exclusion criteria

- Article type: did not use RT and VR for cognitive impairment intervention; did not evaluate the effects of VR-RTs; books, review articles, and commentary letters, as well as unpublished papers.
- Language: papers not written in English.

Data Extraction

Textbox 2 presents the types of information and data extracted independently by 2 authors (QM and ZZ). Specifically, the beneficial effects of VR-RTs on the quality of life and cognitive-related functions of people with cognitive impairment were summarized. "Quality of life" refers to the overall mental and physical health in daily life. Cognitive-related functions refer to neuropsychiatric issues due to cognitive impairment, such as anxiety and declines in memory and cognitive status. The analysis of user satisfaction was based on side effects and user acceptability, immersion, and engagement. "Side effects" refer to various forms of physical discomfort triggered by VR-RTs, such as eye strain, vertigo, or nausea. "User acceptability" mainly refers to users' attitudes toward the interventions and reveals the level of acceptance of VR-RTs. Immersion indicates the degree of immersiveness of the VR environment, and engagement indicates the level of individuals' active and meaningful participation in an intervention.

XSL•FO
RenderX

Textbox 2. The extracted information on study design, participants, interventions, and outcomes.

Study

- First author
- Year of publication
- Study region
- Study aim
- Study design

Participants

- Sample size
- Average age
- Sex distribution
- Health status

Virtual reality (VR) technology

- Device
- Locations of device
- System type
- Construction approach

VR-based reminiscence therapy

- Contents of the VR-based reminiscence therapy
- Intervention frequency, duration per session, and intervention duration

Outcome measure

- Quality of life
- Cognitive-related functions
- User satisfaction

Data Synthesis

We qualitatively synthesized the included articles to analyze the effects and applicability of VR-RTs among older adults with cognitive impairment. Specifically, we identified the studies and participants' characteristics. We then summarized and analyzed the VR technologies and intervention designs of the VR-RTs according to the extracted data.

We classify, herein, VR technologies into nonimmersive, semi-immersive, or immersive VR technologies [30]. Specifically, nonimmersive VR technologies generally use small platform screens, such as tablets, computers, and televisions, to produce a VR environment with low levels of immersion and interaction. Semi-immersive VR technologies, ranked between nonimmersive and immersive VR, allow users to engage in a partially virtual environment with certain degrees of immersion and interactivity using large-screen projectors and panoramic monitors. In immersive VR, the virtual environment completely replaces the physical world through platforms that provide full views and sensory feedback, such as head-mounted displays and cave automatic virtual environment.

The effectiveness of VR-RTs in improving quality of life and cognitive-related functions was further analyzed according to the differences in the values of the outcomes from before to after the intervention. On the basis of the above-mentioned synthesis, we critically discuss the benefits and challenges of applying VR technologies during RT for older adults with cognitive impairment.

Risk of Bias Assessment

The Cochrane Risk of Bias in Non-Randomized Studies of Interventions tool was used independently by 2 authors (QM and ZZ) to assess the risk of bias in the included studies. The assessment indices of the tool consist of 7 domains, including confounding, selection of participants in the study, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of the reported result. Each domain was classified into low risk, moderate risk, serious risk, or critical risk on the basis of the responses to the items, and an overall assessment was calculated based on the 7 domains [31].

XSL•FO
RenderX

Results

Study Identification

We identified 537 articles in our initial search, of which, 195 were excluded as duplicates. Two authors (QM and ZZ) screened the titles and abstracts of the remaining 342 articles to identify potentially relevant studies. We reviewed the full texts of 46 potentially relevant articles, and 17 articles met our inclusion criteria. We further identified 5 potentially eligible articles by manually searching the reference lists of the included articles and the Google Scholar database for missed studies. Finally, 22 articles were included in the systematic review. Figure 1 presents the literature search and review process.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the literature search and review process for potential articles on virtual reality-based reminiscence therapies among older adults with cognitive impairments.

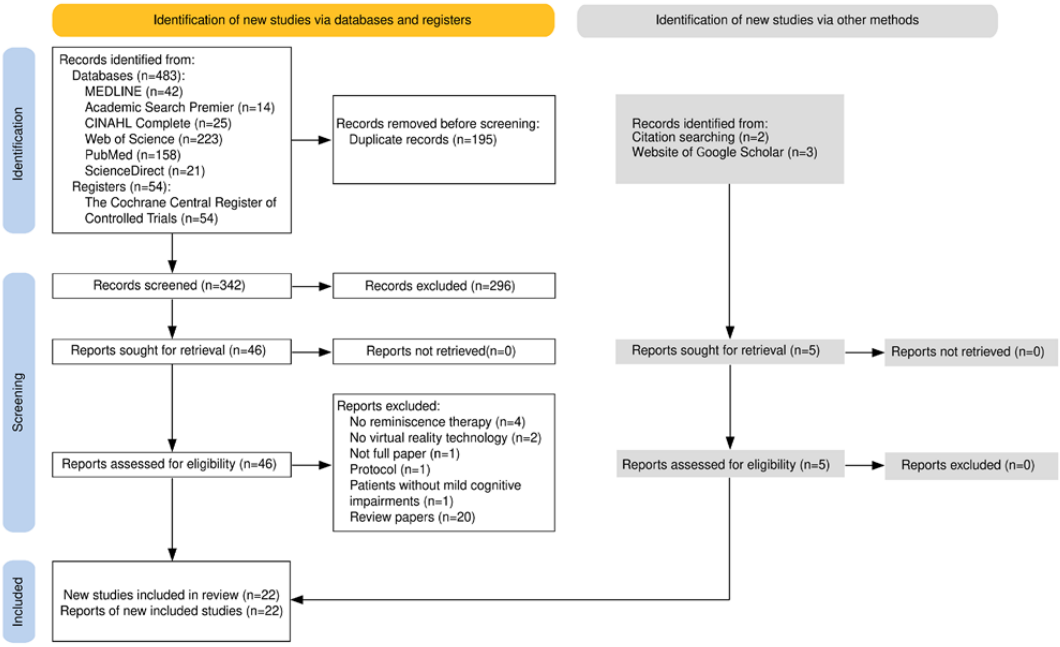

Study Characteristics

The 22 included studies comprised 1 (5%) case study [32], 1 (5%) randomized controlled trial (RCT) study [24], 7 (32%) pilot studies [33-39], 6 (27%) feasibility studies [25,26,40-43], and 7 (32%) prototype studies [18,44-49]. Most of the included studies were conducted in Australia (3/22, 14%) [26,33,35], France (3/22, 14%) [41,42,46], Japan (3/22, 14%) [24,32,39], or Taiwan (3/22, 14%) [37,44,45]. Half of the studies were published in the year 2020 [25,26,35,36,43,47] or 2021 [24,32,33,39,40].

The sample sizes in the included studies ranged from 2 to 63 participants. The mean age of the participants ranged from 63.40 to 89.80 years. Female accounted for 38% (5/13) to 89% (39/44) of the participants in the included studies. More information on the 22 included articles can be found in Multimedia Appendix 1. The detailed information on the PRISMA checklist can be found in Multimedia Appendix 2.

VR Technology

Typically, in the included studies on nonimmersive VR, a computer was used to present scenes [24,34,38,43,45,47,49]. The participants used a Kinect (Microsoft Corporation) [45], mouse and keyboard [34,38,43,47,49], or tactile screen [24] to interact with the VR environment. In semi-immersive VR, BARCO iSpace (Barco NV) or a large projector was used to show the VR environment [41,42,46,48], and users used an ART wireless finger-tracking [41,46], mouse [42], or a Kinect [48] to interact with the environment. Immersive VR was the most popular technology, as it was used in about 60% (13/22) of the included studies [18,25,26,32,33,35-37,39,40,44,47,49]. Oculus (Meta Platforms) (8/22, 36%) and VIVE (HTC Corporation) headsets (3/22, 14%) were used most frequently to present immersive VR scenes. More detailed information about the VR systems is shown in Table 1.

XSL•FO
RenderX

Table 1. Degree of immersion, devices and construction approach for the virtual reality (VR) technology in the included studies.

Degree of immersion, study, and year	Device	Construction approach
Nonimmersive VR		
Man et al [34], 2011	Computer	Geometry-based VR
Lancioni et al [38], 2015	Computer	Image-based VR
Hou et al [45], 2017	Computer and Kinect	Image-based VR
Klein et al [49], 2018	Display	Image-based VR
Tominari et al [24], 2021	Tablet	Image-based VR
Sun et al [47], 2020	Monitors, television, or projectors	Image-based VR
Tabafunda et al [43], 2020	Webcam and display	Image-based and geometry-based VR
Semi-immersive VR		
Siriaraya and Ang [48], 2014	Kinect sensor and projector	Geometry-based VR
Chapoulie et al [46], 2014	BARCO iSpace, glasses with advanced real-time tracking system, and ART wireless finger-tracking devices	Image-based VR
Benoit et al [41], 2015	BARCO iSpace, glasses with advanced real-time tracking system, and ART wireless finger-tracking devices	Image-based VR
Manera et al [42], 2016	A Barco OverView OLSF-721 full HD 3D stereoscopic LED[a] video wall, and Volfoni Edge 1.2 active 3D LCD[b] shutter glasses	Image-based VR
Immersive VR		
Tsao et al [44], 2019	VR case headset	Geometry-based VR
Xu and Wang [25], 2020	HTC VIVE headset	Geometry-based VR
Yu and Choi [18], 2018	VIVE and Oculus headsets	Image-based VR
Klein et al [49], 2018	Headset VR without detailed information	Image-based VR
Webber et al [35], 2020	Oculus Go headset	Image-based VR
Saredakis et al [26], 2020	Oculus Go headset	Image-based VR
Sun et al [47], 2020	Headset VR without detailed information	Image-based VR
Coelho et al [36], 2020	Samsung Gear VR with a smartphone and the Oculus Rift	Image-based VR
Yahara et al [32], 2021	Oculus Go headset	Image-based VR
Saredakis et al [33], 2021	Oculus Go headset	Image-based VR
Huang and Yang [37], 2022	VIVE headset	Image-based VR
Niki et al [39], 2021	Oculus Go headset	Image-based and geometry-based VR
Afifi et al [40], 2021	Oculus Go headset	Image-based and geometry-based VR

[a]LED: light-emitting diode.

[b]LCD: liquid-crystal display.

Next, we classified the approaches used in the included studies to construct VR environments into geometry-based VR and image-based VR [44]. Specifically, 7 (32%) studies used geometry-based VR, which allows users to experience simulated objects and scenes in a virtual environment created using specialized 3D model-building software (eg, 3DS Max, Unity, and AutoCAD). Meanwhile, 18 (82%) studies used images or videos from the real world to establish image-based VR through image processing, image-based rendering, or video editing. More information about the VR technologies used in the 22 included studies can be found in Multimedia Appendix 1.

Characteristics of VR-RT Interventions

The VR-RT stimulus materials were classified according to the content of the VR scenes as virtual travel for addresses, old locations, old items, and personally relevant pictures or videos. In the included studies, virtual travel for addresses usually allowed participants to view places of interest on Google Maps; old locations were used to indicate historical architectural and infrastructural features such as parks, homes, streets, or buildings; old items were products previously used by the participants, including foods, clothing, artifacts, magazines,

XSL•FO
RenderX

posters, music, and other similar items that were once part of their daily lives; and personally relevant pictures or videos were photos and videos of the participants or their friends or family members. We further divided the stimulus materials into personalized and general stimulus materials. The personalized stimulus materials were tailored for the participants on the basis of their past experiences while the general stimulus materials were generic resources applicable to a broad population.

Four (18%) studies used virtual travel for addresses as stimulus materials [26,32,33,35]. Although these studies allowed the participants to view preferred addresses on Google Maps, no correlation to a specific time period (eg, the 1960s) was identified in the studies. Therefore, we considered virtual travel for addresses to be general stimulus materials. Eight (36%) and 6 (27%) studies used old locations [36,37,39-41,43,44,46] and old items [24,25,34,45,48,49], respectively, to evoke the participants' memories. Most of the stimulus materials in these categories pertained to the period between the 1940s and the 1980s and specifically to the participants' periods of childhood and adolescence, as shown in Table 2. Notably, 2 (9%) studies used personalized old locations based on the participants' experiences [36,40], whereas the other studies (12/22, 55%) involving old items and old locations used general stimulus materials. Four (18%) studies used pictures or videos of the participants and their family members in the past as stimulus materials [18,38,40,42].

Table 2. The stimulus materials and intervention characteristics of virtual reality-based reminiscence therapies.

Stimulus materials and type	Content	Intervention duration, mean (SD)
Virtual travel for addresses		
General	Google Maps street view [26,32,33,35].	23.75 (14.93) min for 3.50 (3.11) sessions, 2.00 (1.41) wks
Personalized	__[a]	__[a]
Old locations		
General	Historic houses between the 1950s and 1980s [37,39,44]; familiar and unfamiliar locations in the past [41,43,46].	9.44 (6.15) min for 6.75 (11.50) sessions, 3.75 (5.50) wks
Personalized	Favorite locations or destinations from the past [40]; childhood houses and workplace locations as well as leisure and religious venues [36].	20.00 (14.14) min for 2.50 (2.12) sessions, 1.50 (0.71) wks
Old items		
General	Items in a home setting [25,48] and a convenience store [34]; food, clothing, live, transportation, and festival items [45]; material and cultural artifacts corresponding to childhood [24]; old items from between the 1940s and 1970s [49].	24.00 (14.07) min for 4.20 (4.44) sessions, 3.00 (3.08) wks
Personalized	__[a]	__[a]
Personally relevant pictures or videos		
General	__[a]	__[a]
Personalized	Family photos and videos [40]; personal old photos [18,38,42].	13.33 (14.43) min for 49.50 (68.59) sessions, 2.50 (2.60) wks

[a]Data not available.

We further summarized and calculated the mean intervention durations of the analyzed VR-RTs. After excluding 4 (18%) studies that did not report the intervention duration [18,43,44,48], the results showed that studies using general old items had the longest mean single-session duration (mean 24.00, SD 14.07 min), while studies using general old locations (6/22, 27%) had the shortest mean single-session duration (mean 9.44, SD 6.15 min). In addition, interventions using personally relevant pictures or videos had the most sessions according to the reported designs (mean 49.50, SD 68.59 sessions). However, except for 1 (5%) study that reassessed participants 3 to 6 months after the intervention [24], the remaining studies (21/22, 96%) conducted no follow-up measure.

Furthermore, 5 (23%) included studies compared the effectiveness of VR-RTs with traditional paper-based [24,25,33,41] and talk-based RTs [25,39]. In paper-based RTs, the stimulus materials are presented on paper; in talk-based RTs, participants' memories are recalled through conversation. Three (14%) included studies compared the effectiveness of nonimmersive VR-RTs with nontreatment [33,41,46], semi-immersive VR-RT [41,46], or immersive VR-RT [33], and 2 (9%) of these studies also compared the effectiveness of using familiar versus unfamiliar stimulus materials in VR-RTs [41,46].

Outcome Measures

Life Quality

As shown in Table 3, 4 (18%) studies [24,33,34,36] estimated the effects of VR-RTs on the quality of life of older adults via statistical tests, such as analysis of variance, the 2-tailed *t* test, and the Mann-Whitney *U* test. The participants' quality of life was assessed in various domains, including psychometric status,

XSL•FO
RenderX

ability to live independently, and physical health. One (25%) of these studies indicated that VR-RT using old items had significant beneficial effects on older adults' psychometric status (P<.01) [24], while the remaining studies (3/22, 14%) [33,34,36] found no substantial differences in quality of life from before to after the intervention. Moreover, paper-based RT [24,34] or nontreatment [33] were not found to have substantial beneficial effects on quality of life.

Table 3. Effects of virtual reality–based reminiscence therapies (VR-RTs) on quality of life.

Outcome and measurement tools for older adults' quality of life	Stimulus materials	VR-RTs	NVR-RTs[a]
Lawton Instrumental Activities of Daily Living scale	Old items [34]	No effect	No effect
Quality of life in Alzheimer disease	Virtual travel for addresses [33]	No effect	No effect
Three-item loneliness scale	Virtual travel for addresses [33]	No effect	No effect
Revised Philadelphia Geriatric Center Morale Scale	Old items [24]	Positive effect[b] (P<.01)	No effect
EUROHIS-QOL-8[c] scores	Old locations [36]	No effect	None[d]

[a]NVR-RT: non–virtual reality–based reminiscence therapy.

[b]A significant difference between virtual reality–based reminiscence therapies and non–virtual reality–based reminiscence therapies.

[c]EUROHIS-QOL-8: European Health Interview Survey-Quality of Life 8-item index.

[d]No measurement for non–virtual reality–based reminiscence therapy.

Cognitive-Related Functions

As shown in Table 4, 13 (59%) studies evaluated the effects of VR-RTs on cognitive-related functions, including emotion, memory performance, and cognitive status [24-26,32-39,41,46]. Specifically, 5 (39%) of these studies assessed emotions (eg, anxiety, apathy, and depression) [24,32,33,37,39], 8 (62%) studies estimated memory performance [24-26,34,35,38,41,46], and 4 (31%) studies evaluated cognitive status [24,33,36,37]. As 1 (8%) study performed a qualitative data analysis [35], we reviewed and analyzed qualitative and quantitative results. Regarding quantitative data, 2 (17%) studies measured and reported the changes in outcome values from before to after the intervention [32,38], and the remaining 10 (83%) studies used statistical tests, such as the paired t test, Mann-Whitney U tests, and Wilcoxon signed rank test, to compare the outcome changes.

XSL•FO
RenderX

Table 4. Effects of virtual reality–based reminiscence therapies (VR-RTs) on cognitive-related functions.

Outcome and measurement tools	Stimulus materials	VR-RTs	NVR-RTs[a]
Anxiety			
State-Trait Anxiety Inventory	Old locations [39]	Positive effect	None[b]
State-Trait Anxiety inventory	Virtual travel for addresses [32]	Positive effect	None[b]
Apathy			
Apathy Evaluation Scale	Virtual travel for addresses [32]	Unclear	None[b]
Apathy Evaluation Scale	Virtual travel for addresses [33]	No effect	No effect
Depression			
Geriatric Depression Scale	Virtual travel for addresses [33]	No effect	No effect
Center for Epidemiological Studies Depression scale	Old locations [37]	Positive effect	None[b]
Multidimensional Observation Scale for Elderly Subjects	Old items [24]	No effect	No effect
Memory			
Verbal engagement	Personally relevant pictures or videos [38]	Positive effect	None[b]
Multifactorial Memory Questionnaire	Old items [34]	Positive effect	Positive effect
Fuld Object-Memory Evaluation	Old items [34]	Positive effect	Positive effect
Performance in Autobiographical Fluency Task	Old locations [41]	Positive effect	No effect
Past reconstruction	Virtual travel for addresses [35]	Positive effect	None[b]
TEMPau scale	Old items [25]	Positive effect	Positive effect
Word fluency test	Old items [24]	No effect	No effect
Verbal fluency test	Virtual travel for addresses [26]	Positive effect	None[b]
Verbal fluency test	Old locations [46]	Positive effect	No effect
Cognitive status			
Clinical Dementia Rating	Old locations [37]	No effect	None[b]
Mini-Mental State Examination	Old locations [37]	No effect	None[b]
Mini-Mental State Examination	Old items [24]	Positive effect	Positive effect
Cognitive Abilities Screening Instrument	Old locations [37]	No effect	None[b]
Trail-making test parts A and B	Old items [24]	No effect	No effect
Neuropsychiatric Inventory	Old locations [36]	No effect	None[b]
Addenbrooke Cognitive Examination III	Virtual travel for addresses [33]	No effect	No effect

[a]NVR-RT: non–virtual reality–based reminiscence therapy.

[b]No measurement for non–virtual reality–based reminiscence therapy.

The effectiveness of VR-RTs for mitigating anxiety was verified in 2 studies that used old locations as stimulus materials [32,39]. However, inconsistent results were reported in terms of the effectiveness of VR-RTs in reducing apathy and depression. Three included studies assessed depression [24,33,37], and only 1 study found a positive effect of VR-RTs when using virtual travel for addresses [37]. The effect of VR-RTs on apathy was unclear due to a pair of opposing outcomes in 1 (50%) study [32].

Among the included studies that addressed memory, 7 (88%) out of 8 reported that VR-RTs improved participants' memories [25,26,34,35,38,41,46]. All related (2/2, 100%) studies found that familiar old locations were substantially more effective than unfamiliar environments in VR-RTs [41,46]. In contrast,

XSL•FO
RenderX

no substantial differences in effectiveness were observed between nonimmersive and semi-immersive VR-RTs. VR-RTs were found to have similar positive effects on memory as paper-based RT [25,34] and talk-based RT [25].

Only 1 (25%) included study revealed that VR-RTs could significantly improve the participants' cognitive status ($P<.01$), assessed by the Mini-Mental State Examination after the intervention. At the same time, no substantial positive effect was found during follow-up measures [24]. The remaining 3 (75%) studies showed no substantial beneficial effect of VR-RTs on cognitive enhancement [33,36,37]. Similarly, a paper-based RT did not benefit cognitive status [33]. However, none of the included studies showed substantial declines in cognitive status after VR-RT use, suggesting that VR-RTs could be used to maintain the cognitive status of individuals with cognitive impairment.

User Satisfaction

A total of 14 (64%) of the 22 studies assessed user satisfaction with the VR-RTs [25,26,32,33,35,36,39-43,46,48,49], of which, 11 (79%) studies estimated the side effects [25,26,32,33,35,36,39-42,46], 7 (50%) studies evaluated user acceptability [26,32,35,40,42,43,49], 4 (28.6%) studies examined the immersiveness of VR [25,40,41,46], and 8 (57%) studies examined engagement with the intervention [25,26,33,36,40,41,46,48]. The results of user satisfaction are summarized in Table 5. One (7%) study used the changes in outcome values from before to after the intervention to estimate the side effects of VR-RTs, as well as user acceptability and engagement [32]; 5 (36%) studies assessed user satisfaction via qualitative data analysis [35,36,43,48,49]; and 9 (64%) studies evaluated user satisfaction via statistical analysis [25,26,33,36,39-42,46].

Table 5. User satisfaction with virtual reality–based reminiscence therapies (VR-RTs) and non–VR-based RTs (NVR-RTs).

Outcome and measurement tools	Stimulus materials	VR-RTs	NVR-RTs
Side effects			
Simulator Sickness Questionnaire	Virtual travel for addresses [33]	No effect	None[a]
Simulator Sickness Questionnaire	Virtual travel for addresses [26]; old locations [36]	Positive effect	None[a]
Self-report numerical rating scale	Old locations [40]; old items [25]; personally relevant pictures or videos [40,42]	No effect	No effect
Self-report numerical rating scale	Old locations [41,46]	No effect	None[a]
Self-report numerical rating scale	Virtual travel for addresses [32]; old locations [39]	No effect	None[a]
Interview	Virtual travel for addresses [35]	No effect	None[a]
User acceptability			
Self-report numerical rating scale	Virtual travel for addresses [26,32]	Positive effect	None[a]
Self-report numerical rating scale	Personally relevant pictures or videos [42]	Positive effect[b] (P=.04)	Positive effect
Self-report numerical rating scale	Old locations [40]; personally relevant pictures or videos [40]	Positive effect	Positive effect
Interview	Old locations [43]	Positive effect	None[a]
Interview	Virtual travel for addresses [35]	NE[c]	None[a]
Observation	Old items [49]	Positive effect	None[a]
Immersion			
Self-report numerical rating scale	Old locations [41,46]; old items [25]	Positive effect	None[a]
Self-report numerical rating scale	Old locations [40]; personally relevant pictures or videos [40]	Positive effect	Positive effect
Engagement			
Human and automated coding	Personally relevant pictures or videos [40]; old locations [40]	Positive effect[b] (P=.004)	Positive effect
Self-report numerical rating scale	Virtual travel for addresses [26]; old locations [41,46]	Positive effect	None[a]
Self-report numerical rating scale	Virtual travel for addresses [33]; old items [25]	Positive effect	Positive effect
Observation	Old locations [36]; old items [48]	Positive effect	None[a]

[a]No measurement for non–virtual reality–based reminiscence therapies.

[b]A significant difference between virtual reality–based reminiscence therapies and non–virtual reality–based reminiscence therapies.

[c]NE: negative effect.

XSL•FO
RenderX

The results of 9 (64%) studies suggested that the VR-RTs were highly safe, with negligible side effects [25,32,33,35,39-42,46]. In contrast, 2 (14%) studies found that VR-RTs using virtual travel for addresses or old locations triggered moderate or high side effects in participants [26,36]. Specifically, discomfort from wearable devices, vertigo, a sensation of falling, eye strain, nausea, fullness in the head, burping, and blurred vision were reported [26,36]. No substantial differences in side effects were identified between VR-RTs and paper-based RT [25,42] or talk-based RT [25,40].

In 6 (86%) studies, participants with cognitive impairment reported that the VR-RTs were enjoyable, attractive, easy to use, acceptable, and highly satisfying [26,32,40,42,43,49]. However, a few participants in 1 (14%) study claimed that the VR-RTs had little value and preferred to avoid virtual travel for addresses due to sadness [35]. In addition, the users reported significantly higher satisfaction levels with VR-RTs using personally relevant pictures or videos than with a paper-based RT (P=.04) [42]. The VR-RTs were also highly immersive and engaging [25,26,33,36,40,41,46,48], with some studies reporting higher engagement levels than those recorded for talk-based RT [40].

Risk of Bias

Of the 22 included trials, 14 (64%) articles had a low risk of bias, 1 (5%) had a moderate risk of bias, 2 (9%) had a serious risk of bias, and 5 (23%) had a critical risk of bias (Figure 2 [18,24-26,32-49]).

XSL•FO
RenderX

Figure 2. Assessment results of the risk of bias.

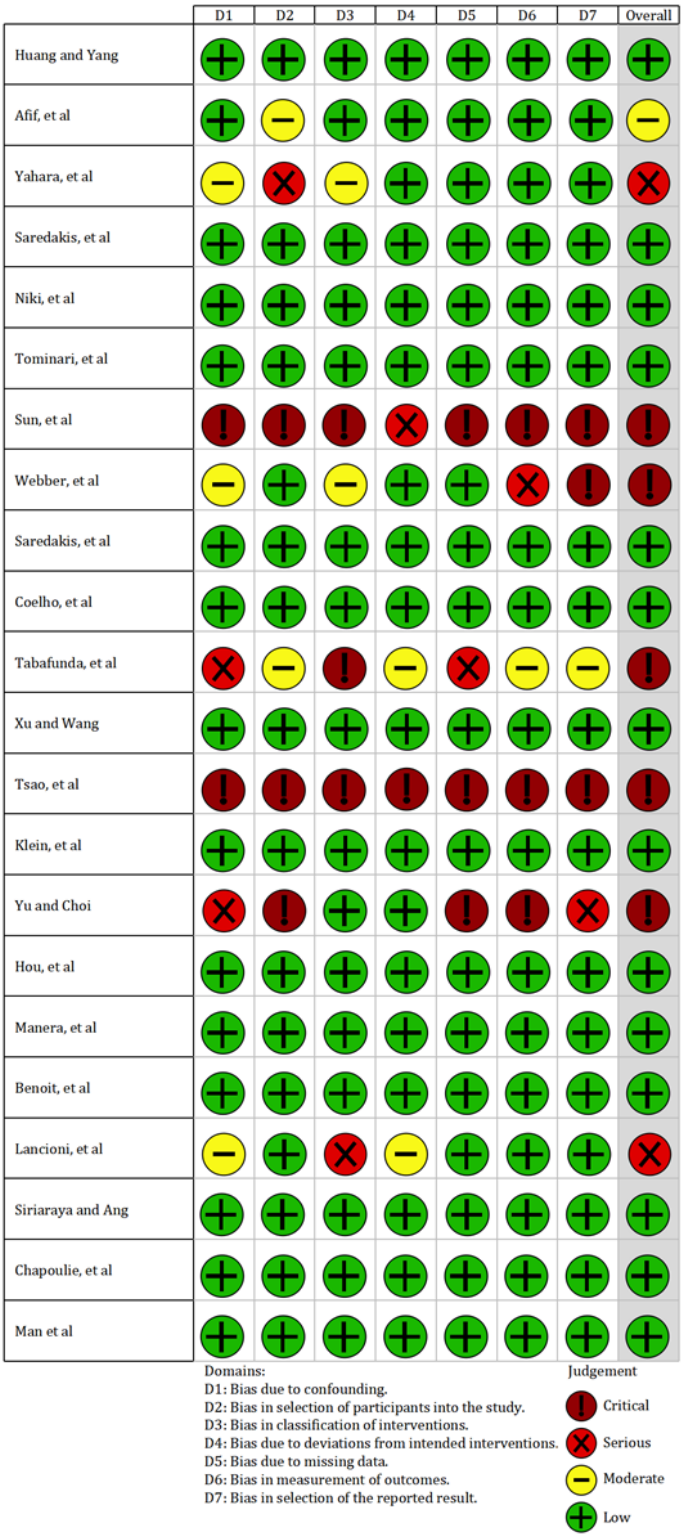

Discussion

Principal Findings

In this systematic review, 22 studies were synthesized comprehensively to evaluate the effects of VR-RTs on older adults with cognitive impairment. The results revealed that VR-RTs could effectively improve emotion and memory but had limited potential to improve quality of life and cognitive status. Concerning user satisfaction, although 2 (9%) studies reported some side effects and 1 (5%) study reported that VR-RTs could trigger negative emotions, most participants reported negligible side effects, indicating that the VR-RTs were safe and enjoyable, as indicated by high levels of immersion and engagement. We further discuss the specific impacts and implications of VR-RTs below.

XSL•FO
RenderX

Effect of VR-Based RTs on Quality of Life

For older adults, RT can enhance life satisfaction and self-confidence by transferring previously gained social skills and enjoyable experiences to real life [50,51]. RTs also allow individuals to reassess and resolve past conflicts, allowing people to gain meaning and a sense of belonging in daily life [52]. However, in contrast with previous studies [51,53,54], we found that most of the studies included in our review showed no substantial positive effect of VR-RTs or traditional RTs on older adults' quality of life, except for 1 (5%) RCT in which VR-RTs effectively increased participants' Philadelphia Geriatric Center Morale Scale scores [24]. The inconsistent effects of VR-RTs and RTs on the participants' quality of life may be attributable to differences in the intervention modalities and settings [55]. For example, one previous study reported that RT was substantially effective only if delivered in 6 or 7 sessions [53]. In addition, most included studies assessed VR-RTs in non-RCTs, possibly leading to result bias. We recommend that additional RCTs with longer intervention durations be conducted to investigate the effect of VR-RTs on quality of life.

Effect of VR-Based RTs on Cognitive-Related Outcomes

We analyzed and synthesized the effects of VR-RTs on older adults' emotions, memory, and cognitive status. According to our results, VR-RTs effectively improved anxiety and memory, consistent with the findings of previous studies on VR-based or RT interventions [54,56]. VR technologies using diverse sensory stimuli considerably enhanced the participants' motivation and interest in the training and elicited pleasant feelings [57]. Therefore, VR technology undeniably played a critical role in RT interventions intended to decrease anxiety and improve emotions. Furthermore, most of the included studies [25,26,41,46] verified the effectiveness of VR-RTs for enhancing autobiographical memory, which pertains to a person's recollection of specific and personal events from their past experiences [58]. This finding is consistent with those of a previous study, suggesting that exposure to an immersive environment containing numerous concrete cues from a person's youth can improve their autobiographical recollection [59]. RT uses memory triggers to recall the long-term memories of people with cognitive impairment [60]. Accordingly, RTs combined with immersive environments were shown to evoke substantial improvements in autobiographical memory. In particular, VR-RTs using familiar old locations had stronger positive effects on autobiographical memory than VR-RTs using unfamiliar environments [41,46]. In addition, 1 (5%) included study revealed that VR-RTs substantially enhanced the temporal orientation of older adults with MCI [24]. In another study, VR-RTs considerably improved participants' objective memory performance, especially their recall of episodic memories immediately and after a delay [34]. Therefore, the potential applications of VR-RTs for memory improvement in older adults with cognitive impairment are enormous.

However, most of the included studies reported that VR-RTs had no substantial effects on depression and apathy. The potential reasons for this lack of effect might be the intervention designs and participants. For example, group RT, which allows several patients to be treated together, was verified to yield greater positive effects and to be more cost-effective than RT with only 1 patient [61]. In addition, RT is not considered an appropriate means of reducing depression in people with severe dementia due to the strong relationship between depression and cognitive impairment [51]. Therefore, we suggest that the VR-RTs may have been somewhat effective for older adults with MCI, given that beneficial interactions in VR-RTs can allow participants to revisit past problems and achieve goals [62]. Although previous studies found that RT had a substantial positive effect on apathy [54,60], the impacts of VR-RTs on apathy and the underlying mechanisms remain unclear due to the limited number of studies. In future studies, it will be crucial to investigate the effects of VR-RTs on depression and apathy in older adults with specific cognitive impairments.

Furthermore, only 1 (25%) study [24] reported that VR-RTs significantly ($P<.01$) improved participants' cognitive status among the included 4 (100%) studies [24,25,33,41]. Previous studies have reported inconsistent effects of RT on cognitive status. Some studies identified a significant moderate or small effect size of RT on cognitive status [51,55]. In contrast, other studies reported no statistically significant effects of RT [54,63]. This discrepancy may be attributable to the various sensitivities and specificities of the assessment tools used to measure cognitive status [59]. Differences in the quality and methodology of studies may also have affected the consistency of the results [64]. Although no significant effectiveness of VR-RTs was reported in improving the participants' cognitive status, the interventions were beneficial in maintaining cognitive status and preventing functional degradation [10], consistent with the findings of previous studies [54,63]. People with cognitive impairment, especially dementia, experience progressive brain shrinkage, which exacerbates cognitive impairment [63]. It is crucial to conduct interventions to preserve cognitive status in this situation. Individual-based or care home-based interventions were verified to improve cognitive status more than general interventions [55]. Thus, we recommend the use of personalized VR-RTs as a nonpharmacological treatment to maintain the cognitive function of people with cognitive impairment.

User Satisfaction With VR-Based RTs

Although VR-based interventions are commonly considered interesting and enjoyable, users may experience side effects due to perceptual discrepancies between the real and virtual environments [65]. Thus, the occurrence and severity of side effects must be considered when assessing user satisfaction with VR-RTs. Two (14%) included studies that used image-based VR reported short-duration side effects such as discomfort caused by wearable devices, eye strain, vertigo, nausea, spatial disorientation, and mild negative emotions [26,36]. Inappropriate placement and inadequate camera resolution were suggested as potential factors that may induce physiological discomfort during VR-RTs [26,36]. Participants also may experience spatial disorientation and confusion due to external verbal interference from reminiscence therapists [36]. Despite these potential side effects of VR-RTs, participants commonly expressed positive attitudes regarding the acceptability of VR-RTs. They were satisfied with the VR-RTs and considered them to be potentially beneficial.

XSL•FO
RenderX

Meanwhile, people preferred immersive VR-RTs to traditional RT or nonimmersive VR-RTs (eg, paper-based and flat screen–based RTs) [33,42]. The finding that immersive VR with multisensory stimulation can enhance biofeedback and presence in the virtual environment [66] prompted the implementation of VR-RTs. However, 1 (5%) included study found that the participants tended to avoid VR-RTs because of negative memories [35]. Hence, research scholars must adequately consider older adults' experiences and avoid potentially adverse RT contexts [58].

In the included studies, the VR-RTs were commonly considered highly immersive and engaging. A high level of immersion gives older adults a sense of being physically together with their family members, especially during VR-RTs that allow older adults and their family members to interact [67]. Older adults who engaged more in the VR-RTs achieved a greater social presence, which prompted them to maintain secure attachments to their families [40]. Notably, older adults with MCI were found to engage more deeply in the VR-RTs than those with dementia, whereas the opposite pattern was observed for immersion [40]. The severity of cognitive impairment may limit physical mobility and interaction [68].

Practical Implications

The included studies used a wide variety of VR technologies and RT content. Given the limited evidence regarding the effects of specific VR-RT programs, we provide preliminary insights into the design of the VR-RT interventions for older adults with cognitive impairment, including the immersion levels and construction approaches of VR technology, the stimulus materials, and the duration of the VR-RT intervention. We also discuss several challenges related to the implementation of VR-RTs.

VR Technology

The immersion level can influence the effectiveness of VR-RTs and user satisfaction by affecting users' sense of presence and engagement [69]. Although immersive VR provides high levels of immersion and engagement, it may cause older adults with severe dementia to experience an increased cognitive load [70]. This may occur because the allocation of brain and sensory resources increases with the degree of immersion in VR [71]. Among the studies included in our review, 2 (14%) studies using immersive VR-RTs reported that participants experienced several side effects [26,36]. In contrast, studies that used semi-immersive VR-RTs [41,42,46] or nonimmersive VR-RTs [72] reported no side effects. However, the optimal immersion level for effective VR-RTs remains uncertain due to the limited number of relevant RCTs. Given the potential side effects, we recommend increased caution when applying immersive VR to older adults.

Image-based VR was considered more acceptable by users compared with geometry-based VR, as it generated more positive emotions [39]. However, the included studies suggested that image-based VR is more likely to induce side effects than geometry-based VR [26,36]. Image-based VR uses highly realistic scenes [72], and the delay between the image presentation and the user's motions might contribute to increased side effects [73]. Thus, we recommend that future studies on the design of image-based VR focus on reducing its side effects.

VR-RT Interventions

The included studies used 4 types of stimulus materials associated with the earlier periods of the participants' lives, consistent with the finding that individuals with Alzheimer disease exhibited better recall of memories from their youth than of recent memories [25]. Familiar stimulus materials were verified as superior to unfamiliar ones for improving the participants' emotions and memory recall [33,41,43,46]. Thus, for VR-RTs, personalized stimulus materials related to the individual's personal history are recommended [49]. In addition, although most of the included studies used visual materials as stimuli, aural materials are also suitable for VR-RTs as they confer benefits such as enjoyment and information encoding [34,43,74]. Therefore, personalized visual and aural stimulus materials associated with the participants' younger days tend to be more effective than general stimulus materials in interventions for cognitive impairment; such materials may enhance cognitive abilities, emotional well-being, and quality of life.

The duration of a VR-RT intervention can also influence its effectiveness. A session duration of around 40 minutes and a total intervention duration of 8 to 12 weeks were considered the most beneficial for participants [64,65]. Although 72% of the users in 1 study left the VR environment after 36 minutes due to cybersickness [72], most included studies restricted the participants' VR exposure time to <30 minutes. Moreover, most studies have only preliminarily evaluated VR-RTs through feasibility studies; thus, the optimal duration of a VR-RT has not been defined. More RCT studies are warranted to determine an appropriate duration of VR-RTs that elicits more positive effects with fewer side effects.

Despite the potential benefits of VR-RTs for older adults with cognitive impairment, the included studies also reported diverse concerns. For example, participants with cognitive impairment in one study expressed concerns regarding vision difficulties in the VR environment [43]. Another study suggested the importance of passive interactivity between patients and caregivers [43], while a third study suggested that an external disturbance could disrupt a participant's attention [36]. VR-RTs may also induce negative emotions in participants due to the recall of unpleasant memories [35]. More attention should be given to the above-mentioned problems and the careful design of interventions to promote the implementation of VR-RTs for older adults with cognitive impairment.

Limitations

This systematic review has several limitations. First, although the findings suggest that VR-RTs are potentially beneficial in improving emotions and memory functions, the methodological restrictions of the included studies prevented us from obtaining statistical evidence of these effects. A meta-analysis is recommended to determine the effectiveness of VR-RTs in terms of specific functional improvements in older adults with cognitive impairment. Second, as few of the included studies were RCTs comparing the effects of VR-RTs and traditional

XSL•FO
RenderX

RTs on cognitive-related outcomes, it remains unclear whether VR-RTs differ significantly from traditional RTs regarding effectiveness. We recommend that future studies compare VR-RTs with traditional RTs through RCTs with large sample sizes and rigorous methodologies. The results of such studies may deepen our knowledge of the potential benefits and limitations of VR-RTs for older adults with cognitive impairment. Third, the intervention design and the participants' demographic characteristics, including age, emotional status, and the severity of cognitive impairment, may influence the effectiveness of a VR-RT. However, we have limited insight into these factors' potential influences and specific mechanisms. Future studies should also diversify the participant groups and VR-RT contents when designing treatment protocols.

Conclusions

We performed this systematic review to provide insights into the effects of VR-RTs among older adults with cognitive impairment. The results reveal that VR-RTs have negligible side effects and high levels of engagement and thus can achieve high levels of user satisfaction. VR-RTs were shown to effectively improve memory functions and emotions and maintain the cognitive status of older adults with cognitive impairment. However, there is limited evidence to support the potential benefits of VR-RTs in terms of improving older adults' quality of life. Moreover, personalized stimulus materials of VR-RTs were found to outperform general stimulus materials for enhancing or maintaining cognitive-related functions. In particular, multisensory stimulus materials related to experiences from their days of youth can considerably enhance the benefits of VR-RTs among older adults. However, some side effects of VR technology use, such as cybersickness and vision difficulties, have raised concerns. In summary, VR-RT applications appear to be a promising option for the nonpharmacological treatment of cognitive impairment.

Acknowledgments

This work was supported by the Shenzhen–Hong Kong–Macao Science and Technology Program Fund (grant SGDX20210823103403028), the Health and Medical Research Fund, the Health Bureau, Hong Kong (grant 20211201), and the Hong Kong Polytechnic University grants (grants P0050656, P0036146, and P0045769).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the 22 papers included in the systematic review.
[DOCX File , 31 KB - jmir_v26i1e53348_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 27 KB - jmir_v26i1e53348_app2.docx]

References

1. Ageing and health: government/agency report. World Health Organization (WHO). URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2022-11-07]
2. Benedict RH, Amato MP, DeLuca J, Geurts JJ. Cognitive impairment in multiple sclerosis: clinical management, MRI, and therapeutic avenues. Lancet Neurol 2020 Oct;19(10):860-871 [FREE Full text] [doi: 10.1016/S1474-4422(20)30277-5] [Medline: 32949546]
3. Sanford AM. Mild cognitive impairment. Clin Geriatr Med 2017 Aug;33(3):325-337. [doi: 10.1016/j.cger.2017.02.005] [Medline: 28689566]
4. Perini G, Cotta Ramusino M, Sinforiani E, Bernini S, Petrachi R, Costa A. Cognitive impairment in depression: recent advances and novel treatments. Neuropsychiatr Dis Treat 2019;15:1249-1258 [FREE Full text] [doi: 10.2147/NDT.S199746] [Medline: 31190831]
5. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020 Aug 08;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
6. Chan KY, Adeloye D, Asante KP, Calia C, Campbell H, Danso SO, Global Dementia Prevention Program (GloDePP). Tackling dementia globally: the Global Dementia Prevention Program (GloDePP) collaboration. J Glob Health 2019 Dec;9(2):020103 [FREE Full text] [doi: 10.7189/jogh.09.020103] [Medline: 31893025]

XSL•FO
RenderX

7. Wimo A, Seeher K, Cataldi R, Cyhlarova E, Dielemann JL, Frisell O, et al. The worldwide costs of dementia in 2019. Alzheimers Dement 2023 Jul 08;19(7):2865-2873 [FREE Full text] [doi: 10.1002/alz.12901] [Medline: 36617519]
8. Abraha I, Rimland JM, Trotta FM, Dell'Aquila G, Cruz-Jentoft A, Petrovic M, et al. Systematic review of systematic reviews of non-pharmacological interventions to treat behavioural disturbances in older patients with dementia. The SENATOR-OnTop series. BMJ Open 2017 Mar 16;7(3):e012759 [FREE Full text] [doi: 10.1136/bmjopen-2016-012759] [Medline: 28302633]
9. Cai Y, Li L, Xu C, Wang Z. The effectiveness of non-pharmacological interventions on apathy in patients with dementia: a systematic review of systematic reviews. Worldviews Evid Based Nurs 2020 Aug 07;17(4):311-318. [doi: 10.1111/wvn.12459] [Medline: 32767834]
10. Lazar A, Thompson H, Demiris G. A systematic review of the use of technology for reminiscence therapy. Health Educ Behav 2014 Oct 01;41(1 Suppl):51S-61S [FREE Full text] [doi: 10.1177/1090198114537067] [Medline: 25274711]
11. Woods B, O'Philbin L, Farrell EM, Spector AE, Orrell M. Reminiscence therapy for dementia. Cochrane Database Syst Rev 2018 Mar 01;3:CD001120 [FREE Full text] [doi: 10.1002/14651858.CD001120.pub3] [Medline: 29493789]
12. Duru Aşiret G, Kapucu S. The effect of reminiscence therapy on cognition, depression, and activities of daily living for patients with Alzheimer disease. J Geriatr Psychiatry Neurol 2016 Jan 06;29(1):31-37. [doi: 10.1177/0891988715598233] [Medline: 26251112]
13. Van Bogaert P, Van Grinsven R, Tolson D, Wouters K, Engelborghs S, Van der Mussele S. Effects of SolCos model-based individual reminiscence on older adults with mild to moderate dementia due to Alzheimer disease: a pilot study. J Am Med Dir Assoc 2013 Jul;14(7):528.e9-528.13. [doi: 10.1016/j.jamda.2013.01.020] [Medline: 23583001]
14. Domingues NS, Verreault P, Hudon C. Reducing burden for caregivers of older adults with mild cognitive impairment: a systematic review. Am J Alzheimers Dis Other Demen 2018 Nov 24;33(7):401-414 [FREE Full text] [doi: 10.1177/1533317518788151] [Medline: 30041535]
15. Khan A, Bleth A, Bakpayev M, Imtiaz N. Reminiscence therapy in the treatment of depression in the elderly: current perspectives. J Ageing Longev 2022 Feb 24;2(1):34-48. [doi: 10.3390/jal2010004]
16. Nebot À, Domènech S, Albino-Pires N, Mugica F, Benali A, Porta X, et al. LONG-REMI: an AI-based technological application to promote healthy mental longevity grounded in reminiscence therapy. Int J Environ Res Public Health 2022 May 15;19(10):5997 [FREE Full text] [doi: 10.3390/ijerph19105997] [Medline: 35627535]
17. Healy D, Flynn A, Conlan O, McSharry J, Walsh J. Older adults' experiences and perceptions of immersive virtual reality: systematic review and thematic synthesis. JMIR Serious Games 2022 Dec 06;10(4):e35802 [FREE Full text] [doi: 10.2196/35802] [Medline: 36472894]
18. Yu S, Choi K. Virtual reality environment based cognitive rehabilitation system. Adv Sci Lett 2018 Mar 01;24(3):1970-1973. [doi: 10.1166/asl.2018.11820]
19. Wu J, Ma Y, Ren Z. Rehabilitative effects of virtual reality technology for mild cognitive impairment: a systematic review with meta-analysis. Front Psychol 2020 Sep 25;11:1811 [FREE Full text] [doi: 10.3389/fpsyg.2020.01811] [Medline: 33101098]
20. Dockx K, Bekkers EM, Van den Bergh V, Ginis P, Rochester L, Hausdorff JM, et al. Virtual reality for rehabilitation in Parkinson's disease. Cochrane Database Syst Rev 2016 Dec 21;12:CD010760 [FREE Full text] [doi: 10.1002/14651858.CD010760.pub2] [Medline: 28000926]
21. Tarnanas I, Tsolakis A, Tsolaki M. Assessing virtual reality environments as cognitive stimulation method for patients with MCI. In: Brooks AL, Brahnam S, Jain LC, editors. Technologies of Inclusive Well-Being: Serious Games, Alternative Realities, and Play Therapy. Berlin, Germany: Springer; 2014:39-74.
22. Liu Q, Song H, Yan M, Ding Y, Wang Y, Chen L, et al. Virtual reality technology in the detection of mild cognitive impairment: a systematic review and meta-analysis. Ageing Res Rev 2023 Jun;87:101889. [doi: 10.1016/j.arr.2023.101889] [Medline: 36806377]
23. Liao YY, Chen IH, Lin YJ, Chen Y, Hsu WC. Effects of virtual reality-based physical and cognitive training on executive function and dual-task gait performance in older adults with mild cognitive impairment: a randomized control trial. Front Aging Neurosci 2019 Jul 16;11:162 [FREE Full text] [doi: 10.3389/fnagi.2019.00162] [Medline: 31379553]
24. Tominari M, Uozumi R, Becker C, Kinoshita A. Reminiscence therapy using virtual reality technology affects cognitive function and subjective well-being in older adults with dementia. Cogent Psychol 2021 Sep 01;8(1):1968991. [doi: 10.1080/23311908.2021.1968991]
25. Xu J, Wang B. Efficacy of VR-based reminiscence therapy in improving autobiographical memory for Chinese patients with AD. In: Proceedings of the AHFE 2020 Virtual Conference on Ergonomics in Design. 2020 Presented at: AHFE '20; July 16-20, 2020; New York, NY p. 339-349 URL: https://link.springer.com/chapter/10.1007/978-3-030-51038-1_47
26. Saredakis D, Keage HA, Corlis M, Loetscher T. Using virtual reality to improve apathy in residential aged care: mixed methods study. J Med Internet Res 2020 Jun 26;22(6):e17632 [FREE Full text] [doi: 10.2196/17632] [Medline: 32469314]
27. Irazoki E, Contreras-Somoza LM, Toribio-Guzmán JM, Jenaro-Río C, van der Roest H, Franco-Martín MA. Technologies for cognitive training and cognitive rehabilitation for people with mild cognitive impairment and dementia. a systematic review. Front Psychol 2020 Apr 9;11:648 [FREE Full text] [doi: 10.3389/fpsyg.2020.00648] [Medline: 32373018]

XSL•FO
RenderX

28. Zhao Y, Yu J, Liu J, Chen Z, Xiang W, Zhu S, et al. Cognitive rehabilitation via head-mounted virtual reality technology in patients with Alzheimer's disease: a systematic review. Digit Med 2022;8:5 [FREE Full text] [doi: 10.4103/digm.digm_42_21]
29. Reisinho P, Raposo R, Zagalo N. A systematic literature review of virtual reality-based reminiscence therapy for people with cognitive impairment or dementia. In: Proceedings of the 2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies. 2022 Presented at: IMET '22; October 4-7, 2022; Limassol, Cyprus p. 1-8 URL: https://ieeexplore.ieee.org/document/9930000
30. Hao J, Xie H, Harp K, Chen Z, Siu KC. Effects of virtual reality intervention on neural plasticity in stroke rehabilitation: a systematic review. Arch Phys Med Rehabil 2022 Mar;103(3):523-541. [doi: 10.1016/j.apmr.2021.06.024] [Medline: 34352269]
31. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
32. Yahara M, Niki K, Ueno K, Okamoto M, Okuda T, Tanaka H, et al. Remote reminiscence using immersive virtual reality may be efficacious for reducing anxiety in patients with mild cognitive impairment even in COVID-19 pandemic: a case report. Biol Pharm Bull 2021;44(7):1019-1023 [FREE Full text] [doi: 10.1248/bpb.b21-00052] [Medline: 34193684]
33. Saredakis D, Keage HA, Corlis M, Ghezzi ES, Loffler H, Loetscher T. The effect of reminiscence therapy using virtual reality on apathy in residential aged care: multisite nonrandomized controlled trial. J Med Internet Res 2021 Sep 20;23(9):e29210 [FREE Full text] [doi: 10.2196/29210] [Medline: 34542418]
34. Man DW, Chung JC, Lee GY. Evaluation of a virtual reality-based memory training programme for Hong Kong Chinese older adults with questionable dementia: a pilot study. Int J Geriatr Psychiatry 2012 May 17;27(5):513-520. [doi: 10.1002/gps.2746] [Medline: 21681818]
35. Webber S, Baker S, Waycott J. Virtual visits: reminiscence in residential aged care with digital mapping technologies. Australas J Ageing 2021 Sep 04;40(3):293-300. [doi: 10.1111/ajag.12902] [Medline: 33538405]
36. Coelho T, Marques C, Moreira D, Soares M, Portugal P, Marques A, et al. Promoting reminiscences with virtual reality headsets: a pilot study with people with dementia. Int J Environ Res Public Health 2020 Dec 12;17(24):9301 [FREE Full text] [doi: 10.3390/ijerph17249301] [Medline: 33322679]
37. Huang LC, Yang YH. The long-term effects of immersive virtual reality reminiscence in people with dementia: longitudinal observational study. JMIR Serious Games 2022 Jul 25;10(3):e36720 [FREE Full text] [doi: 10.2196/36720] [Medline: 35877169]
38. Lancioni GE, Singh NN, O'Reilly MF, Sigafoos J, D'Amico F, Ferlisi G, et al. Patients with moderate Alzheimer's disease engage in verbal reminiscence with the support of a computer-aided program: a pilot study. Front Aging Neurosci 2015 Jun 03;7:109 [FREE Full text] [doi: 10.3389/fnagi.2015.00109] [Medline: 26089797]
39. Niki K, Yahara M, Inagaki M, Takahashi N, Watanabe A, Okuda T, et al. Immersive virtual reality reminiscence reduces anxiety in the oldest-old without causing serious side effects: a single-center, pilot, and randomized crossover study. Front Hum Neurosci 2020;14:598161 [FREE Full text] [doi: 10.3389/fnhum.2020.598161] [Medline: 33536887]
40. Afifi T, Collins NL, Rand K, Fujiwara K, Mazur A, Otmar C, et al. Testing the feasibility of virtual reality with older adults with cognitive impairments and their family members who live at a distance. Innov Aging 2021;5(2):igab014 [FREE Full text] [doi: 10.1093/geroni/igab014] [Medline: 34632105]
41. Benoit M, Guerchouche R, Petit PD, Chapoulie E, Manera V, Chaurasia G, et al. Is it possible to use highly realistic virtual reality in the elderly? A feasibility study with image-based rendering. Neuropsychiatr Dis Treat 2015;11:557-563 [FREE Full text] [doi: 10.2147/NDT.S73179] [Medline: 25834437]
42. Manera V, Chapoulie E, Bourgeois J, Guerchouche R, David R, Ondrej J, et al. A feasibility study with image-based rendered virtual reality in patients with mild cognitive impairment and dementia. PLoS One 2016 Mar 18;11(3):e0151487 [FREE Full text] [doi: 10.1371/journal.pone.0151487] [Medline: 26990298]
43. Tabafunda A, Matthews S, Akhter R, Uribe-Quevedo A, Sun W, Horsburgh S, et al. Development of a non-immersive VR reminiscence therapy experience for patients with dementia. In: Proceedings of the 22nd International Conference on Late Breaking Posters. 2020 Presented at: HCI '20; July 19-24, 2020; Copenhagen, Denmark p. 509-517 URL: https://link.springer.com/chapter/10.1007/978-3-030-60703-6_66 [doi: 10.1007/978-3-030-60703-6_66]
44. Tsao YC, Shu CC, Lan TS. Development of a reminiscence therapy system for the elderly using the integration of virtual reality and augmented reality. Sustainability 2019 Sep 02;11(17):4792. [doi: 10.3390/su11174792]
45. Hou CJ, Huang MW, Zhou JY, Hsu PC, Zeng JH, Chen YT. The application of individual virtual nostalgic game design to the evaluation of cognitive function. Annu Int Conf IEEE Eng Med Biol Soc 2017 Jul;2017:2586-2589. [doi: 10.1109/EMBC.2017.8037386] [Medline: 29060428]
46. Chapoulie E, Guerchouche R, Petit PD, Chaurasia G, Robert P, Drettakis G. Reminiscence therapy using image-based rendering in VR. In: Proceedings of the 2014 IEEE Conference on Virtual Reality (VR). 2014 Presented at: VR '14; March 29-April 02, 2014; Minneapolis, MN p. 45-50 URL: https://ieeexplore.ieee.org/document/6802049

XSL•FO
RenderX

47. Sun W, Horsburgh S, Quevedo A, Liscano R, Tabafunda A, Akhter R, et al. Advancing reminiscence therapy through virtual reality application to promote social connectedness of persons with dementia. Gerontechnology 2020 Oct 03;19(s):1. [doi: 10.4017/gt.2020.19.s.70041]
48. Siriaraya P, Ang CS. Recreating living experiences from past memories through virtual worlds for people with dementia. In: Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON p. 3977-3986 URL: https://dl.acm.org/doi/10.1145/2556288.2557035
49. Klein P, Uhlig M, Will H. The touch and feel of the past—using haptic and VR artefacts to enrich reminiscence therapy for people with dementia. Technologies 2018 Nov 13;6(4):104. [doi: 10.3390/technologies6040104]
50. Lök N, Bademli K, Selçuk-Tosun A. The effect of reminiscence therapy on cognitive functions, depression, and quality of life in Alzheimer patients: randomized controlled trial. Int J Geriatr Psychiatry 2019 Jan;34(1):47-53. [doi: 10.1002/gps.4980] [Medline: 30246408]
51. Saragih ID, Tonapa SI, Yao CT, Saragih IS, Lee BO. Effects of reminiscence therapy in people with dementia: a systematic review and meta-analysis. J Psychiatr Ment Health Nurs 2022 Dec 17;29(6):883-903. [doi: 10.1111/jpm.12830] [Medline: 35348260]
52. Siverová J, Bužgová R. The effect of reminiscence therapy on quality of life, attitudes to ageing, and depressive symptoms in institutionalized elderly adults with cognitive impairment: a quasi-experimental study. Int J Ment Health Nurs 2018 Oct;27(5):1430-1439. [doi: 10.1111/inm.12442] [Medline: 29427397]
53. Park K, Lee S, Yang J, Song T, Hong GR. A systematic review and meta-analysis on the effect of reminiscence therapy for people with dementia. Int Psychogeriatr 2019 Nov 04;31(11):1581-1597. [doi: 10.1017/S1041610218002168] [Medline: 30712519]
54. Thomas JM, Sezgin D. Effectiveness of reminiscence therapy in reducing agitation and depression and improving quality of life and cognition in long-term care residents with dementia: a systematic review and meta-analysis. Geriatr Nurs 2021;42(6):1497-1506. [doi: 10.1016/j.gerinurse.2021.10.014] [Medline: 34735996]
55. O' Philbin L, Woods B, Farrell EM, Spector AE, Orrell M. Reminiscence therapy for dementia: an abridged Cochrane systematic review of the evidence from randomized controlled trials. Expert Rev Neurother 2018 Sep 28;18(9):715-727. [doi: 10.1080/14737175.2018.1509709] [Medline: 30092689]
56. Istvandity L. Combining music and reminiscence therapy interventions for well-being in elderly populations: a systematic review. Complement Ther Clin Pract 2017 Aug;28:18-25. [doi: 10.1016/j.ctcp.2017.03.003] [Medline: 28779927]
57. Domenicucci R, Ferrandes F, Sarlo M, Borella E, Belacchi C. Efficacy of ICT-based interventions in improving psychological outcomes among older adults with MCI and dementia: a systematic review and meta-analysis. Ageing Res Rev 2022 Dec;82:101781. [doi: 10.1016/j.arr.2022.101781] [Medline: 36343879]
58. Subramaniam P, Woods B, Whitaker C. Life review and life story books for people with mild to moderate dementia: a randomised controlled trial. Aging Ment Health 2014 Sep 24;18(3):363-375 [FREE Full text] [doi: 10.1080/13607863.2013.837144] [Medline: 24063317]
59. Kirk M, Rasmussen KW, Overgaard SB, Berntsen D. Five weeks of immersive reminiscence therapy improves autobiographical memory in Alzheimer's disease. Memory 2019 Apr 08;27(4):441-454. [doi: 10.1080/09658211.2018.1515960] [Medline: 30198380]
60. İnel Manav A, Simsek N. The effect of reminiscence therapy with internet-based videos on cognitive status and apathy of older people with mild dementia. J Geriatr Psychiatry Neurol 2019 Mar 06;32(2):104-113. [doi: 10.1177/0891988718819864] [Medline: 30612511]
61. Liu Z, Yang F, Lou Y, Zhou W, Tong F. The effectiveness of reminiscence therapy on alleviating depressive symptoms in older adults: a systematic review. Front Psychol 2021 Aug 17;12:709853 [FREE Full text] [doi: 10.3389/fpsyg.2021.709853] [Medline: 34484066]
62. Morgan S, Woods RT. Life review with people with dementia in care homes: a preliminary randomized controlled trial. Nonpharmacol Ther Dement 2010;1(1):43-60 [FREE Full text]
63. Huang HC, Chen YT, Chen PY, Huey-Lan Hu S, Liu F, Kuo YL, et al. Reminiscence therapy improves cognitive functions and reduces depressive symptoms in elderly people with dementia: a meta-analysis of randomized controlled trials. J Am Med Dir Assoc 2015 Dec;16(12):1087-1094. [doi: 10.1016/j.jamda.2015.07.010] [Medline: 26341034]
64. Panic N, Leoncini E, de Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. PLoS One 2013 Dec 26;8(12):e83138 [FREE Full text] [doi: 10.1371/journal.pone.0083138] [Medline: 24386151]
65. Baniasadi T, Ayyoubzadeh SM, Mohammadzadeh N. Challenges and practical considerations in applying virtual reality in medical education and treatment. Oman Med J 2020 May 19;35(3):e125 [FREE Full text] [doi: 10.5001/omj.2020.43] [Medline: 32489677]
66. Campo-Prieto P, Cancela JM, Rodríguez-Fuentes G. Immersive virtual reality as physical therapy in older adults: present or future (systematic review). Virtual Reality 2021 Jan 03;25(3):801-817. [doi: 10.1007/s10055-020-00495-x]

XSL•FO
RenderX

67. Roberts AR, De Schutter B, Franks K, Radina ME. Older adults' experiences with audiovisual virtual reality: perceived usefulness and other factors influencing technology acceptance. Clin Gerontol 2019 Mar 05;42(1):27-33. [doi: 10.1080/07317115.2018.1442380] [Medline: 29505343]
68. Tolea MI, Morris JC, Galvin JE. Trajectory of mobility decline by type of dementia. Alzheimer Dis Assoc Disord 2016;30(1):60-66 [FREE Full text] [doi: 10.1097/WAD.0000000000000091] [Medline: 25886717]
69. Rose T, Nam CS, Chen KB. Immersion of virtual reality for rehabilitation - review. Appl Ergon 2018 May;69:153-161. [doi: 10.1016/j.apergo.2018.01.009] [Medline: 29477323]
70. Makransky G, Terkildsen TS, Mayer RE. Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learn Instr 2019 Apr;60:225-236. [doi: 10.1016/j.learninstruc.2017.12.007]
71. Slobounov SM, Ray W, Johnson B, Slobounov E, Newell KM. Modulation of cortical activity in 2D versus 3D virtual reality environments: an EEG study. Int J Psychophysiol 2015 Mar;95(3):254-260. [doi: 10.1016/j.ijpsycho.2014.11.003] [Medline: 25448267]
72. Martirosov S, Bureš M, Zítka T. Cyber sickness in low-immersive, semi-immersive, and fully immersive virtual reality. Virtual Real 2022 May 19;26(1):15-32 [FREE Full text] [doi: 10.1007/s10055-021-00507-4] [Medline: 34025203]
73. Bowman DA, McMahan RP. Virtual reality: how much immersion is enough? Computer 2007 Jul;40(7):36-43. [doi: 10.1109/mc.2007.257]
74. Cuevas PE, Davidson PM, Mejilla JL, Rodney TW. Reminiscence therapy for older adults with Alzheimer's disease: a literature review. Int J Ment Health Nurs 2020 Jun 26;29(3):364-371. [doi: 10.1111/inm.12692] [Medline: 31984570]

Abbreviations

MCI: mild cognitive impairment
RCT: randomized controlled trial
RT: reminiscence therapy
VR: virtual reality
VR-RT: virtual reality–based reminiscence therapy

Edited by A Mavragani; submitted 04.10.23; peer-reviewed by E Vashishtha, D Liu; comments to author 28.02.24; revised version received 24.04.24; accepted 02.10.24; published 12.11.24.

Please cite as:
Mao Q, Zhao Z, Yu L, Zhao Y, Wang H
The Effects of Virtual Reality–Based Reminiscence Therapies for Older Adults With Cognitive Impairment: Systematic Review
J Med Internet Res 2024;26:e53348
URL: https://www.jmir.org/2024/1/e53348
doi:10.2196/53348
PMID:

©Qian Mao, Zhen Zhao, Lisha Yu, Yang Zhao, Hailiang Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Telehealth Use and Legal Considerations in Drug Health Services During Pandemics: Systematic Scoping Review

Meryem Jefferies[1,2], LLB, PhD; Robert Graham[1], MD; Marguerite Tracy[1,3], MD, PhD; Scott Read[2,4,5], PhD; Mohammed Eslam[2], MD, PhD; Mark W Douglas[2,6], MD, PhD; Jacob George[2], MD, Prof Dr, PhD

[1]Drug Health, Western Sydney Local Health District, North Parramatta, NSW, Australia

[2]Storr Liver Centre, The Westmead Institute for Medical Research, The University of Sydney at Westmead Hospital, Westmead, NSW, Australia

[3]General Practice Clinical School, Faculty of Medicine and Health, The University of Sydney, Sydney NSW, Australia

[4]Blacktown/Mt Druitt Clinical School and Research Centre, School of Medicine, Western Sydney University, Blacktown, NSW, Australia

[5]Blacktown Hospital, Western Sydney Local Health District, Blacktown, NSW, Australia

[6]Centre for Infectious Diseases and Microbiology, Sydney Infectious Diseases Institute, The University of Sydney at Westmead Hospital, Westmead, Australia

Corresponding Author:
Meryem Jefferies, LLB, PhD
Drug Health, Western Sydney Local Health District
Building 83, Cumberland Hospital 5 Fleet Street
North Parramatta, NSW
Australia
Phone: 61 61 478 168518
Fax: 61 61 2 8838 2025
Email: Meryem.Jefferies@health.nsw.gov.au

Abstract

Background: The COVID-19 pandemic impacted patients with substance use disorder (SUD) more than the general population and resulted in substantially increased emergency department admissions. Routine care of patients attending drug health services during the pandemic transitioned, with telehealth being important in delivering appropriate care. However, telehealth introduces unique risks such as privacy, confidentiality, and data safety. Providing health care through telehealth may fail if the legal impacts are not fully identified and acted on by health professionals. It also poses unintended risks for patients and can result in ineffectiveness, damages, medical negligence, and detracts from the best intentions of governments and health professionals. Understanding the legal framework ensures that medical professionals operate health care through telehealth within the law. Providing health care successfully through telehealth depends on the balance between innovation and legal compliance. By considering these aspects, clinicians and practitioners can provide effective and safe telehealth services during pandemics or any other natural disaster.

Objective: We aimed to explore the legal impact of autonomy consent, confidentiality, privacy, data security, professional indemnity, and liability when delivering telehealth to patients with SUD. The scoping review also aimed to provide legal, ethical, and clinical considerations to minimize legal risks with using telehealth in drug health service outpatient settings.

Methods: We performed a scoping review to provide an overview of existing research, statutes, and case laws for the incorporation of clinical, ethical, and legal considerations into telehealth use. Six databases for medical and 6 databases for legal publications were searched, as well as Australian national and selected international regulatory standards. Medical articles published up to June 2022 were included in this review. Our search yielded 1436 publications, 614 abstracts were reviewed, and 80 published studies met the inclusion criteria from 614 legal and medical search results. Current regulations related to technology use in drug health services, relevant cases, and international regulatory standards are discussed.

Results: In total, 43 legal documents including 15 statutes, 4 case laws, and 37 medical publications were reviewed. The themes arising from the literature were consent and autonomy (20/80, 25%), confidentiality (8/80, 10%), privacy (8/80, 10%), data security (7/80, 9%), and professional indemnity issues (3/80, 4%) in telehealth use. Further, 24 studies identified legal issues associated with telehealth use in patients with SUD.

Conclusions: Our review identified potential legal issues associated with telehealth use in patients with SUD. Several legal and medical research articles provide frameworks, codes of conduct, or suggestions for clinicians to consider, but there was little

XSL•FO
RenderX

discussion or evidence of how legal considerations are being applied when providing telehealth consultations at drug health services. Clinicians should be aware of the medicolegal implications when providing health care via telehealth at drug health services.

(J Med Internet Res 2024;26:e46394) doi:10.2196/46394

KEYWORDS

telehealth use; emergency department admission; substance use disorder; drug addiction; consent; privacy; professional indemnity; confidentiality; data security

Introduction

A recent study showed a substantial increase in emergency department presentations for patients with substance use disorder (SUD) during the COVID-19 pandemic [1]. This population required innovative solutions [2] to provide access to timely health care [1] including the use of telehealth. New technologies however raise new legal, security, and ethical considerations [3]. Due to innovations developing faster [3], implementing an appropriate technology that ensures patient safety while delivering the best possible care to patients with SUDs is important [2]. The time frame for creating appropriate clinical frameworks within the current regulatory structure however is not adequately addressed [4].

There is currently no published study analyzing the use of telehealth and the related issues that arise in the clinical context of drug health. This review outlines using telehealth to deliver routine clinical care to patients attending drug health services in emergencies, identifies specific issues that arise from the use of such platforms with current laws, and describes the elements to be considered by clinicians when using telehealth at drug health services.

The Australian Government responded to the COVID-19 pandemic [5] by legislating the COVID-19 Legislation Amendment (Emergency Measures) Act 2020 [6] to regulate the public health emergency [7]. Likewise, New South Wales (NSW) Health implemented interim actions for surveillance, infection control, laboratory testing, and contact management for COVID-19 [8]. In the context of a pandemic, people attending drug health services have an increased risk of transmission of infectious agents compared to patients with other health conditions. People requiring drug health services are often required to attend clinics daily, increasing the risk of transmission of respiratory viruses. Several studies on the impact of the COVID-19 pandemic on people with SUD have reported higher rates of depression, anxiety, irritability, and posttraumatic stress [9]. To meet the challenges, digital health services were increasingly used in drug health services, including *My Health Record* [10] and telehealth [11].

Using telehealth will continue to provide health care in drug health services. However, there remains a need for a legally regulated model of care for the use of telehealth to deliver best practices for patients with SUD. In this paper, we addressed potential legal issues about autonomy and consent, confidentiality, privacy, data security, professional indemnity, and liability when using telehealth in drug health services.

Methods

Study Design

As this review involved legal and medical considerations, a systematic scoping review is the most suitable methodology. We conducted the scoping review by following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. This checklist is provided in Multimedia Appendix 1 [12,13].

Review Context

This review context is of what is known from the legal and medical literature about the use of telehealth, current regulations related to telehealth use in drug health services, relevant cases, and international regulatory models.

Inclusion Criteria

Inclusion criteria were established before performing the search as shown in Table 1. We performed the search up to June 2022 and used relevance, article type, publication sources, and subject matter for including and excluding publications. As relevance, those related to telehealth use in patients with SUD, autonomy and consent, confidentiality, privacy, data security, and professional indemnity and liability were included. As article type, original studies, reviews and editorials, viewpoints, guidelines, letters to editors and commentaries, and relevant legislation and case laws were included. As publication sources, published papers in a peer-reviewed platform or an institutional report were included. As subject matter, those related to regulations and legal considerations when using telehealth in drug health services, autonomy and consent, confidentiality, privacy, data security, and professional indemnity and liability were included. We excluded any publication that was not related to telehealth use in drug health services. Legal articles that were not published in full or accessible, preprints, or unavailable full sources were also excluded.

XSL•FO
RenderX

Table 1. Inclusion and exclusion criteria.

Type of criteria	Included	Excluded
Relevance	Telehealth, SUD[a], autonomy and consent, confidentiality, privacy, data security, professional indemnity, and liability	Those irrelevant to telehealth use or did not contain a telehealth component in patients with SUD
Article type	Original studies, reviews, editorials, viewpoints, guidelines, letters to editors, commentaries, relevant legislation, and case laws	Preprints
Publication sources	Published in a peer-reviewed platform or institutional report	Unavailable full texts or full names of sources
Subject matter	Regulations and legal considerations, telehealth, drug health service, autonomy and consent, confidentiality, privacy, data security, professional indemnity, and liability	Those related to telehealth modelling, and studies that used technology only for a better understanding of disease dynamics with no help from health professionals

[a]SUD: substance use disorder.

Search Strategy

This review was performed with two search strategies—a medical and a legal search.

Medical Search

Research on telehealth use in patients with SUDs was performed using the key databases MEDLINE, Embase, and CINAHL. Databases including the Science Citation Index were expanded by using database-specific controlled vocabulary (where available) and general free-text terms. Other relevant websites were explored including those of the Global Health Library, WHO, United States Food and Drug Administration, European Medicines Agency, and Australian Government Department of Health. Other search engines such as Google Scholar, ResearchGate, and Science Direct were explored using the search terms. Keywords searched include telehealth use, emergency department admission, SUD, drug addiction, consenting, privacy, confidentiality, and data safety. Credibility and quality of sources were evaluated by reviewing the reputation and authority of the sources. The medical search was performed on titles and abstracts. All publications stating the role of telehealth use in drug health services during the COVID-19 emergency, autonomy and consent, and data security were included in the analysis. Duplicate publications, review articles, opinion articles, and letters were excluded. Those publications which did not provide the principal data or were articles with incomplete data were excluded.

Legal Search

The legal search was undertaken using Australian Case Law databases including Westlaw AU, Lexis Advance Pacific, AustLII, CCH Intelliconnect, and CaseLaw NSW. The relevant legislation in Australia, the United States, the United Kingdom, Canada, and New Zealand were reviewed. The key terms including autonomy consent, confidentiality, privacy, data security, professional indemnity, liability, Privacy Act, and legal basis for telehealth were used.

The search was conducted on titles, abstracts, and a summary of case laws. An elementary search on June 1, 2021, identified a range of evidence on the role of technology in drug health services during the COVID-19 pandemic. A combination of keywords: COVID-19, emergency department admission, telehealth in health care, telehealth in drug health clinic, eHealth, e-consenting, telemedicine, SUD, autonomy, and consent, data security, privacy, and confidentiality were used. Relevant publications and legal authorities were selected for further review. A review of authoritative judgments regarding consent and patient record privacy was included.

Article Selection and Data Extraction

All studies stating the role of telehealth use in drug health services during COVID-19, autonomy and consent, and data security were included in the analysis. Duplicate publications and papers not providing the principal data or articles with incomplete data were excluded. Case authorities reviewed are of superior courts and apply to the current context of drug health in Australia.

We retrieved 300 records from our medical search and 314 records from our legal search. Then 496 records were screened for the relevance of title and abstract, excluding 260 of the records. Additionally, 156 were further excluded as they were not directly relevant to the context of this study. Finally, 80 studies (43 legal and 37 medical studies) were included. The bibliography of all selected articles was reviewed for relevance.

Results

Overview

A summary of medical and legal studies used in this manuscript is outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1). Medical issues arising from the published literature included those at the health system level—cost-effectiveness, access, early intervention [9] to prevent use of emergency services [1]; local level practice issues—cost of technology [2], ensuring data safety and privacy [14]; clinician level—conducting a thorough clinical assessment via telehealth, medicolegal indemnity cover [15]; and at the patient level—increased access to care, potential for misdiagnosis, data security and confidentiality [14], inequitable access to technology and connectivity [16].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the review process.

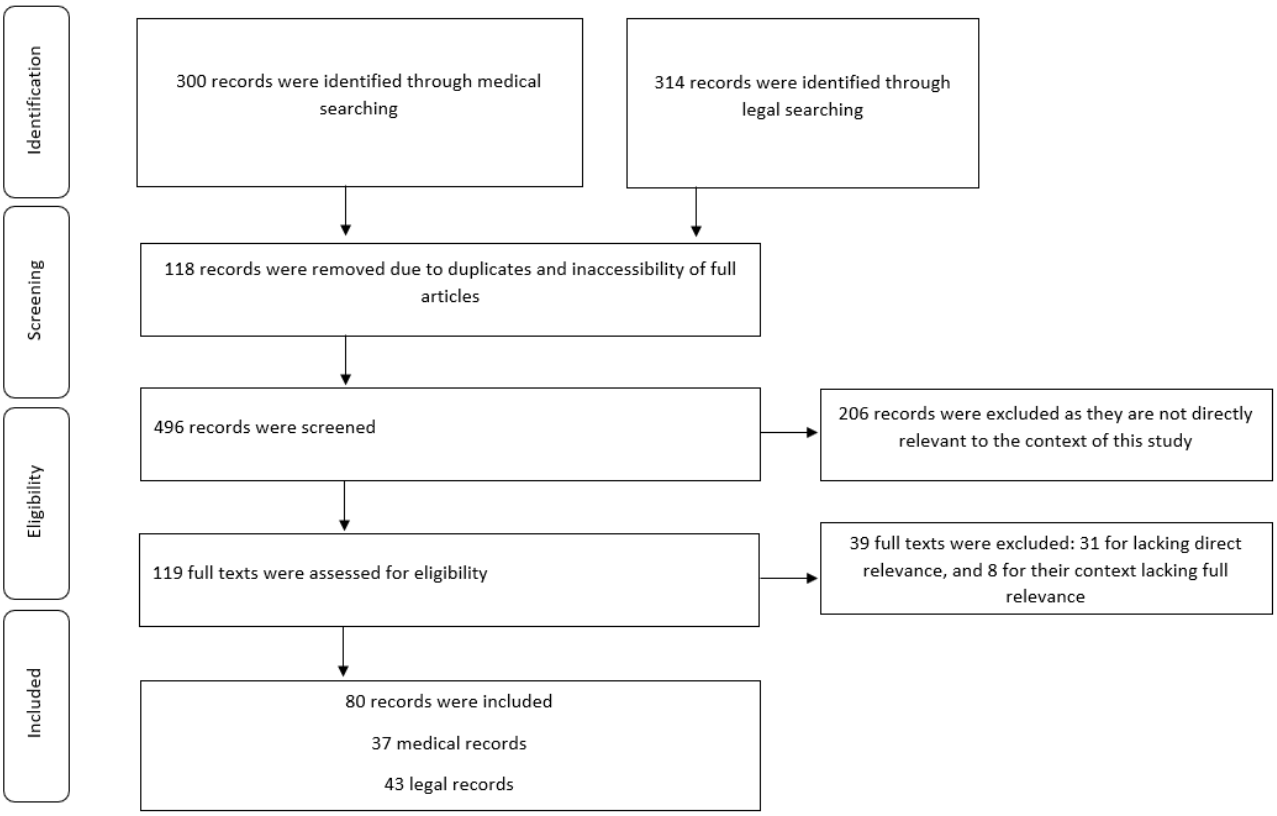

From the records assessed, five domains should be considered by health professionals when providing health care in the drug health context; these include autonomy and consent, confidentiality, privacy, data security, and professional indemnity and liability.

The reviewed literature offered several positive aspects stemming from the use of technology platforms to enhance health care delivery in drug health. These include cost-effectiveness, the ability to provide health care to remote and regional areas, the potential to reduce infection transmission, and a reduction in emergency department admissions.

While several positive aspects of the use of telehealth in drug health care were identified, the literature also identified numerous issues: patient privacy, data privacy [17], the risk of cybercrime [18], data security, lower efficacy for clinical examination, health professional licensure, reimbursement, credentialing, malpractice [19], and high risk of misdiagnosis due to the inability to conduct a complete physical examination [20]. In one report, 63% of total misdiagnoses in health care were the result of failing to perform a physical examination [21]. There are numerous local reports on security breaches that result in a patient's medical information being accessed by unauthorized third parties [22,23] A further issue is legal uncertainty regarding professional indemnity for clinicians during remote consultations [24].

Legal Considerations in the Use of Telehealth in Drug Health Clinical Settings

Autonomy and Consent

This review identified several legal issues regarding obtaining consent from patients with SUD through telehealth. Philosophers and ethicists [21] have proposed arguments to suggest that people with SUD may be unable to provide informed consent to participate in treatment programs [25]. Their reasoning included the inability to say "no" to the offer of heroin [26]. Other research however suggests that addiction does not significantly impair autonomy. In NSW, patient consent to treatment is mandatory [27,28] except for involuntary treatment in cases of severe substance dependence under the Drug and Alcohol Treatment Act 2007 (NSW) [29]. The legislation provides that patient consent to treatment is not required under 3 limited circumstances. First, when an emergency arises and the patient lacks capacity and treatment wishes are unknown [30], when capacity is deemed by law to not exist—such as when patients are legally "incompetent" due to mental health or disability (in such instances legislation provides for alternate means of obtaining consent), and where the law overrides a patient's right to provide consent [31].

Where treatment is given without consent, or it exceeds the terms of the consent, the treatment will constitute a tortious battery and may also be the subject of complaint and disciplinary sanction. This is the case even if the treatment is successful or therapeutic. Where treatment is given and the consent is premised on incorrect or inadequate information, a patient has a cause of action in negligence if they endure damage or loss.

XSL•FO
RenderX

In most treatment contexts, obtaining informed consent is a necessary precondition of treatment. Without consent, a medical practitioner may face disciplinary allegations of professional malpractice or be sued for battery or negligence if detriment or harm is endured. This is despite the fact the treatment may be given in a manner that benefits the patient.

Obtaining patient consent via telehealth in a drug health setting can be problematic. Currently, there is no legal authoritative case regarding breach of consent due to telehealth use. However, the following cases are reflective of patient consent-related issues which may arise in drug health clinics.

The accepted legal test at common law for patient consent and autonomy is established in Rogers vs Whittaker [32,33] and the test has also been adopted in Canada and the United Kingdom [34]. This legal principle potentially applies to telehealth use if a health care professional fails to obtain informed consent and an informed decision from the patient regarding the risks of particular procedures and if consultation is not performed in person. The case is considered a precedent in many countries including the United Kingdom, the United States, New Zealand, and Canada [35]. Courts have no difficulty in assessing the presence and efficacy of consent simply because information and treatment are being provided by using technology.

There are challenges to providing health care via telehealth which present new variables and challenges that are not present in face-to-face consultations. The risk of misreading nonverbal cues in telehealth, the inability of the patient to access relevant information (due to their own limitations or the lack of availability of this information), the digital literacy of the patient and the effect of influence by other persons who may be unknowingly present when telehealth is conducted are all variables that have the potential to impact on the exchange of information, and therefore on the practitioner's ability to determine informed consent.

In Malette vs Shulman [36], the issue of lack of consent and autonomy was considered. The plaintiff (Malette) was seriously injured in a car crash and was rushed to a hospital where she was found to carry a card that identified she was of the Jehovah's Witness faith, and expressly requested that no blood transfusions be given under any circumstances. However, Dr Shulman authorized a life-saving blood transfusion despite knowing that the procedure was contrary to the patient's consent. The patient sued. The Supreme Court of Ontario concluded that the doctor had violated the patient's rights over her body by acting against her intentions without consent [37] and this constituted a battery. Battery is a legal concept that occurs when a person touches or applies force to another person's body in an unauthorized and offensive way. In Malette vs Shulman, the blood transfusions without Georgette's consent raised the question of whether it established a battery. The rules of self-determination and individual autonomy play an essential role. Georgette's right to reject treatment, even if it might have the purpose of lifesaving, constitutes her autonomy. Dr Shulman's action against Georgette's consent in favor of saving her life constituted a legal and ethical tension. The trial judge decided in favor of Georgette, highlighting that the Jehovah's Witness card validly limited Dr Shulman's right to treat her. Dr Shulman's actions constituted a battery as there were no grounds to ignore Georgette's obvious rejection of blood transfusions. Malette vs Shulman provides the nuanced balance between respecting patient autonomy and ensuring medical help. It highlights the legal obligation to obtain informed consent, even in lifesaving or any other emergencies. The case can be considered as a precedent when issues of consent and patient autonomy arise in telehealth use. The case is recognized in multiple countries including Australia, the United States, the United Kingdom, and New Zealand in addition to Canada [35].

The relevant legislation in NSW regarding consent to treatment are the Children and Young Persons (Care and Protection) Act 1998 [29,38,39], Guardianship Act 1987 [40,41], Guardianship Regulation 2016 [42,43], and Minor (Property and Contracts) Act 1970 [42,44], and Drug and Alcohol Treatment Act 2007 (NSW). Obtaining and ensuring informed consent is an essential domain for telehealth consultations of patients for drug health and providing health care without consent can lead to risk of malpractice and medical negligence [45]. Consent must be provided freely [46] and the use of telehealth should acknowledge and account for the limitations and risks of telehealth services when compared to face-to-face treatment. Obtaining consent through telehealth also faces legal and regulatory barriers such as large variations in rules, regulations, and guidelines for practice. Such variation contributes to the confusion of providers engaged in the practice of telehealth [47]. The Medical Board of Australia provides guidelines on technology-based patient consultations and their medicolegal implications [48].

Confidentiality

Patient confidentiality is at risk of being breached when using telehealth. Information related to health is always sensitive and drug health clients always prefer that their medical diagnoses and treatment be handled with a high level of confidentiality [49]. Moreover, health care providers need to inform the patient that they have a right to access information about themselves, to approve the information to be used by other health care providers, and in some cases for deidentified information to be used in clinical research [49].

Skene [15] defined confidentiality from three viewpoints. The first is an ethical perspective, as maintaining confidentiality about a medical record respects the autonomy of the patient [45]. The second is the medical perspective. If a patient knows that what they discuss with their doctor is confidential, they are more likely to be honest about what they disclose, which is beneficial for their treatment [45]. The third is a public policy perspective, that is, it is in society's interest that patients are provided with the best possible treatment through telehealth [15].

Medical records of patients with substance use dependency and mental health issues have strong protection in the United States. In Jaffee vs Redmond [33], communications between psychotherapists and their patients were found to be privileged under the Federal Rules of Evidence [50]. This judgment extended the privilege of patient records to psychotherapists, with the court noting that confidentiality played an important part in effective patient treatment. The case law is authoritative

XSL•FO
RenderX

in other states because of the superiority of the court making the judgment [50]. The court concluded that the privilege is necessary because "the mental wellbeing of our citizenry, no less than its physical health, is a public good of transcendent importance" and "the mere possibility of disclosure may harm establishment of the confidential relationship necessary for successful treatment" [50].

In summary, health professionals must consider confidentiality when providing health care through telehealth by taking reasonable steps such as ensuring the use of telehealth software that prevents screen recording [48]. However, confidentiality can be still breached such as by recording the consultation using another device [48]. Jaffee vs Redmond is one of the key cases for legal and ethical consideration when using telehealth regarding patient confidentiality and is recognized around the world [51].

Privacy

Patient privacy can be compromised in the following ways: first, patients may need to access the internet from public locations such as libraries due to limited connectivity. This situation can compromise privacy, as patients may be using shared devices or networks, potentially exposing sensitive health information. Living in a remote area without an internet connection is not a privacy issue, however, lack of internet connection and the act of using the internet from public places can impact privacy [48]. Second, health professionals might have less ability to perform a controlled examination through telehealth. This can lead to a lack of control over the collection, use, and sharing of private information. Unusually, people may be around the consultant and the patient might not want to overhear what was said or sometimes people are not comfortable discussing health issues from within their home due to a lack of privacy, and an unsafe environment to discuss concerns [52]. Patients for drug health are a vulnerable cohort and when their medical information is disclosed against their will, they frequently feel ashamed, helpless, and stigmatized [53]. Successful treatment may become difficult in such circumstances and the relationship of trust with the doctor may be permanently harmed. Actual or threatened disclosure can be particularly traumatic for the drug health of patients as they need supportive, reliable, and trusting relationships with health care providers.

The following cases are also relevant in Australia. In R vs Osolin [54], the importance of patient-centric privacy was emphasized by the court and the case is recognized by multiple jurisdictions around the world [55]. In this case, disclosure of sexual assault in medical records that were then produced under subpoena was considered by the court to amount to "double victimization" of the patient, with significant adverse consequences for ongoing treatment [56]. Undesired disclosure of patient records impairs treatment and can impact the procedural fairness of court proceedings that use clinical records as evidence [57]. Use of patient records as evidence can result in harm to the person's psychiatric treatment.

Health professionals must consider privacy when providing health care through telehealth by taking reasonable steps to protect personal information from misuse and loss, and from unauthorized access, modification, or disclosure. The clinicians must advise on and obtain patient consent as to how their personal details will be collected, stored, and used [14].

Data Security

Data security has become a far more serious issue stemming from technology use in drug health clinics as greater volumes of data are generated and used in novel ways. Therefore, drug health information requires special protection against threats such as hacking and malware. Malicious or mistaken breaches of sensitive patient data can result in serious consequences such as reputational damage, emotional distress, stigma, and public health concerns [58]. Data security issues can also arise during telehealth consultations of people with SUD by being hacked when using video visits, using unprotected devices, public Wi-Fi, when there is low health digital literacy, or a lack of familiarity with online platforms [59]. Breach of data security during telehealth will have the same legal consequences as breaching data security of the patient's medical record.

Recently, hacking by state actors has gained prominence. This is exemplified in the recent class action that is being brought against the private health insurer Medibank, where thousands of patient's health records were published by ransomware hackers [58]. Frequently, hackers are external to health care and are linked to political or profit motives [58,60]. For example, the Australian National University lost the data of 200,000 patients covering a 19-year period due to a cyberattack believed to have been perpetrated by a sophisticated state actor [16]. In Australia, where those responsible for such attacks can be identified and brought before the courts, a range of criminal offenses exist under the Criminal Code Act 1995 (the Commonwealth; Cth) [61] and the Cybercrime Act 2001 (Cth) [45,62]. Australia participates in international approaches to support data security such as becoming a signatory to Europe's Convention on Cybercrime [45,63]. However, it must be acknowledged that bringing perpetrators to justice is not always possible, and this does not undo the damage to patient confidentiality or privacy.

Thus, when providing health care through telehealth, patients need to be informed of privacy choices and data security measures the practitioner is adopting. Clinicians must avoid disclosure of sensitive personal information. In determining policy on data sharing, the interests of the individual need to be balanced against the risks [64].

Professional Indemnity and Liability

The legal ramifications regarding the provision of health care through telehealth consultation are largely untested. Courts will continue to consider what is reasonable in the provision of health care on a peer standard test. Practitioners who do not adopt appropriate technology or take reasonable and necessary precautions to secure patient information face potential legal action [65]. Clinicians, all users, software developers, and algorithm developers are all potentially liable. Hence, software and algorithm developers cannot be the only ones responsible for issues regarding technology [45].

Thus, health professionals should be aware that they have all the medicolegal risks of face-to-face consultations, and potentially an increased risk because of the limitations of

XSL•FO
RenderX

telehealth in obtaining informed consent and the absence of physical examination. Special attention should be given to prevent errors and omissions, negligent credentialing, breaches of privacy, and interruptions of service due to equipment or technology failures. Clinicians need to be aware of what exactly their liability insurance policy covers, especially when providing telehealth services in other jurisdictions [47].

Regulation of Telehealth Use in Australia

In Australia, technology-based patient consultation, including those for the drug health of patients, is regulated by the Health Practitioner Regulation National Law Act 2009 (NSW) [66], and the principles are defined in Good Medical Practice: a Code of Conduct for Doctors in Australia [67]. Privacy of communication is covered under the Telecommunications (Interception and Access) Act 1979 (Cth) [68], while confidentiality is regulated under The Privacy Act 1988 (Cth) [69]. Medical information handling for patients of drug health services is regulated under the Health Records and Information Privacy Act 2002 (NSW) [10]. Civil liability is determined by the principles of the Civil Liability Act 2002 (NSW) [70]. Finally, data security is regulated under the Criminal Code Act 1995 (Cth) [61], proposed by the Cybercrime Act 2001 (Cth) [62].

In other jurisdictions, for example in New Zealand, the confidentiality and privacy of patients with substance use issues have even stronger protections than in Australia. The New Zealand Evidence Act [71,72] states that strong privilege operates in criminal proceedings to protect communications between patients and health professionals when patients are being treated at drug health clinics for substance dependency under section 59 of the New Zealand Evidence Act. The rationale is that "the broader aim of securing due compliance with the law is more likely to be managed through medical treatment instead of through prosecution." This particularly applies to drug addiction, where legal sanctions have minimum effect [72]. It is argued that New Zealand's legislation better accommodates the needs of patients with drug addiction and mental health issues than its Australian counterpart for several reasons. First, it protects clinical records made when treating conditions that may lead to criminal behavior. Second, the legislation provides that the public interest in protecting vulnerable patients with drug addiction and mental issues is paramount. Finally, it seeks to protect the particular treatment relationship in question and ensures that potential patients are not prevented from seeking help. The New Zealand Evidence Act [71,73] was strengthened when the new Evidence Act was introduced in 2006 [74].

Discussion

Principal Findings

The findings from our review indicate that health care providers are required to consider 5 domains when using telehealth. The key precedent case laws are applicable when a breach of any of these 5 domains occurs. We found that telehealth use in patients with SUD and the arising legal considerations will remain an area of interest for medical and legal journals. This is not an unexpected outcome as telehealth can offer many benefits such as improving access to health care, decreasing medical costs, and increasing patient satisfaction [75].

Most published studies in the databases are in the English language. This is not surprising as English is the dominant language of international scientific publications; however, not all countries use it for medical research communications. Further, there is still a gap between the views of experts and researchers from other non–English-speaking countries [76]. It is worth noting that most published studies focused on using telehealth during the pandemic. The arguments and points about the legal considerations will likely continue and will identify future research areas.

Both medical and legal searches showed that telehealth use-related issues concern the 5 domains that were assessed with precedent case laws and legislation. A comparative legal analysis examining legal systems across different jurisdictions identifies similarities, differences, and best practices. The case laws related to autonomy, consent, confidentiality, and privacy are recognized in multiple countries including Australia, United States, United Kingdom, Canada, and New Zealand. This is related to sharing a common law tradition and similar legal systems. However, multiple other factors affect the recognition of foreign judgment including bilateral or multilateral treaties between countries, accessibility of legal sources in the foreign country, compliance with procedural requirements, public interest, and regulatory considerations. Such factors impact the enforceability and recognition of foreign judgments. In conclusion, recognizing foreign judgments is relevant because it provides cross-jurisdictional insights to support legal decision-making [77].

Potentially Liable Parties From Telehealth Use

Our findings show that the following 3 parties can be legally liable for adverse clinical outcomes that arise due to the use of telehealth: the first is medical professionals and hospitals [45]. This liability depends on the nature of the breach such as whether the medical professional acted as part of good clinical practice as widely accepted by professional peer opinion [34]. The second is the software companies that provide the technology platforms [45]. This option focuses on whether the technology used provided acceptable specifications and protections. The third is that both parties can be liable [45]. This option is not blaming only one party as each party has a responsibility for the outcome.

Regulation of Telehealth

Our study highlighted the regulatory perspective [45] as technology involves complex systems and has multiple medical, technology, and software specialties involved. Hence, it is difficult to regulate according to current law and policy. The regulation of medical devices is currently performed by the Therapeutic Goods Administration (TGA) [78]. The TGA states that Australian regulatory frameworks cannot be used to identify potential risk factors related to software such as in a medical device. Thus, the regulatory structure for the use of telehealth in drug health clinical settings is not established. The essential elements that must be considered in establishing the regulation of technology include (1) accessing the service, (2) controlling

XSL•FO
RenderX

the operation of the service, and (3) identification and management of risks such as external hacking [79]. When new technologies are created there is often some initial concern from users [79], followed by assurances from service providers. When technological failure occurs, the law plays an important role in delivering redress, and through regulation preventing undesirable outcomes. Keeping regulatory frameworks up to date for every new technology is an ongoing journey [45].

Regulation of telehealth use and providing a quality standard from the TGA is difficult as there is no validated control procedure [45]. A collaborative approach comprising all relevant stakeholders including medical professionals, other specialists such as technology providers and insurance companies, is most likely to be successful in regulating telehealth use.

Our review suggested that a critical aspect is to avoid the occurrence of any harm and to ensure that the patient is fully informed of any potential risks or limitations of the proposed clinical service by telehealth [45]. Hence, health professionals conducting a consultation using telehealth in drug health situations should be aware that they may be legally liable in the event of misdiagnosis or treatment without consent that results in harm to the patient [80].

Strengths and Limitations

This study has several strengths including the search of medical and legal basis principles, and the use of case law and statutes as primary legal sources. There are however several limitations to this study. First, this study included the English literature, and this limited comparison to studies published outside this domain and conducted in other countries. Second, this study included publications up to June 2022 as it sought to capture information before and during the COVID-19 pandemic.

Conclusion

Our scoping review demonstrates that the implementation of telehealth in drug health settings is important to connect these patients to the required health care. Delivering health care through telehealth can be safe and effective if consent and autonomy, confidentiality, privacy, data security, and professional indemnity and liability are considered.

Acknowledgments

This study received no specific fund from any funding agency in the public, commercial, or not-for-profit sectors. JG is supported by the Robert W. Storr Bequest to the Sydney Medical Foundation, University of Sydney; a National Health and Medical Research Council of Australia (NHMRC) Program Grant (APP1053206), Project, Ideas and Investigator grants (APP2001692, APP1107178, APP1108422, and APP1196492), and a Cancer Institute NSW grant (2021/ATRG2028). MWD is supported by National Health and Medical Research Council Ideas Grants (APP2002565 and APP20021586) from the Australian Centre for HIV and Hepatitis Virology Research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed Search Strategy.
[DOCX File , 71 KB - jmir_v26i1e46394_app1.docx]

Multimedia Appendix 2
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File , 27 KB - jmir_v26i1e46394_app2.docx]

References

1. Jefferies M, Rashid H, Graham R, Read S, Banik GR, Lam T, et al. COVID-19 impact on Australian patients with substance use disorders: emergency department admissions in Western Sydney before vaccine roll out. Vaccines (Basel) 2022;10(6):889. [doi: 10.3390/vaccines10060889] [Medline: 35746497]
2. Butenko A, Larouche P. Regulation for innovativeness or regulation of innovation? Law, Innovation Technol 2015 Jul;7(1):52-82. [doi: 10.1080/17579961.2015.1052643]
3. White B, McDonald FJ, Willmott L. Health law in Australia. Australia: Lawbook Company; 2014.
4. Blais LE, Wagner WE. Emerging science, adaptive regulation, and the problem of rulemaking ruts. Tex L Rev 2007;86:1701. [doi: 10.1093/obo/9780199363445-0009]
5. Organization WH. Protocol for assessment of potential risk factors for 2019-novel coronavirus (COVID-19) infection among health care workers in a health care setting. Geneva: World Health Organization; 2020.
6. COVID-19 Legislation Amendment (Emergency Measures) Act 2020 (NSW). URL: https://legislation.nsw.gov.au/view/pdf/asmade/act-2020-1 [accessed 2024-10-11]
7. McHattan N. 'COVID-19'update: government response to COVID-19 pandemic. Australian Restructuring Insolvency & Turnaround Association Journal 2020;32(2):6.

XSL•FO
RenderX

8. Australia C. Coronavirus Disease 2019 (COVID-19): CDNA National Guidelines for Public Health Units. Canberra: DoH; 2020.
9. Yazdi K, Fuchs-Leitner I, Rosenleitner J, Gerstgrasser NW. Impact of the COVID-19 pandemic on patients with alcohol use disorder and associated risk factors for relapse. Front Psychiatry 2020;11:620612. [doi: 10.3389/fpsyt.2020.620612] [Medline: 33391060]
10. Health Records and Information Privacy Act 2002 (NSW). URL: https://legislation.nsw.gov.au/view/html/inforce/current/act-2002-071 [accessed 2024-10-11]
11. Kaplan B, Litewka S. Ethical challenges of telemedicine and telehealth. Camb Q Healthc Ethics 2008;17(4):401-416. [doi: 10.1017/S0963180108080535] [Medline: 18724880]
12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
13. Waddington H, White H, Snilstveit B, Hombrados JG, Vojtkova M, Davies P, et al. How to do a good systematic review of effects in international development: a tool kit. J Dev Eff 2012;4(3):359-387. [doi: 10.1080/19439342.2012.711765]
14. Shachar C, Engel J, Elwyn G. Implications for telehealth in a postpandemic future: regulatory and privacy issues. JAMA 2020;323(23):2375-2376. [doi: 10.1001/jama.2020.7943] [Medline: 32421170]
15. Skene L. Law and Medical Practice: Rights, Duties, Claims and Defences. Australia: Lexis Nexis; 1998.
16. Smith M, Urbas G. Technology Law: Australian and International Perspectives. Canberra: Cambridge University Press; 2021.
17. Bokolo AJ. Exploring the adoption of telemedicine and virtual software for care of outpatients during and after COVID-19 pandemic. Ir J Med Sci 2021;190(1):1-10 [FREE Full text] [doi: 10.1007/s11845-020-02299-z] [Medline: 32642981]
18. O'Hara VM, Johnston SV, Browne NT. The paediatric weight management office visit via telemedicine: pre- to post-COVID-19 pandemic. Pediatr Obes 2020;15(8):e12694 [FREE Full text] [doi: 10.1111/ijpo.12694] [Medline: 32627434]
19. Fisk M, Livingstone A, Pit SW. Telehealth in the context of COVID-19: changing perspectives in Australia, the United Kingdom, and the United States. J Med Internet Res 2020;22(6):e19264 [FREE Full text] [doi: 10.2196/19264] [Medline: 32463377]
20. Ngan OMY, Sim JH. Evolution of bioethics education in the medical programme: a tale of two medical schools. Int J Ethics Educ 2020;6(1):37-50. [doi: 10.1007/s40889-020-00112-0]
21. Albers MW, Gilmore GC, Kaye J, Murphy C, Wingfield A, Bennett DA, et al. At the interface of sensory and motor dysfunctions and Alzheimer's disease. Alzheimers Dement 2015;11(1):70-98 [FREE Full text] [doi: 10.1016/j.jalz.2014.04.514] [Medline: 25022540]
22. Bassan S. Data privacy considerations for telehealth consumers amid COVID-19. J Law Biosci 2020;7(1):lsaa075 [FREE Full text] [doi: 10.1093/jlb/lsaa075] [Medline: 33269090]
23. Hood C. Telehealth cybersecurity. In: A Practical Guide to Emergency Telehealth. New York: Oxford University Press; 2021:81-92.
24. Goudkamp T, Goudkamp J. An outline of the Civil Liability Act 2002 [NSW]. Law Soc J: Off J Law Soc N S W 2002;40(7):46-49.
25. Charland LC. Cynthia's dilemma: consenting to heroin prescription. Am J Bioeth 2002;2(2):37-47. [doi: 10.1162/152651602317533686] [Medline: 12189075]
26. Pickard H. The puzzle of addiction. In: The Philosophy and Science of Addiction. UK: Routledge; 2018:9-22.
27. Levy N. Addiction, autonomy, and informed consent: on and off the garden path. J Med Philos 2016;41(1):56-73 [FREE Full text] [doi: 10.1093/jmp/jhv033] [Medline: 26668167]
28. Dietz HP, Caudwell Hall J, Weeg N. Antenatal and intrapartum consent: implications of the NSW consent manual 2020. Aust N Z J Obstet Gynaecol 2021;61(5):802-805. [doi: 10.1111/ajo.13397] [Medline: 34097302]
29. Drug and Alcohol Treatment Act 2007 (NSW). URL: https://legislation.nsw.gov.au/view/whole/html/inforce/current/act-2007-007 [accessed 2024-10-11]
30. Lander F, Gray D, Wilkes E. The alcohol mandatory treatment act: evidence, ethics and the law. Med J Aust 2015;203(1):47-49. [doi: 10.5694/mja15.00173] [Medline: 26126570]
31. Dore GM, Batey RG, Smyth DJ. Involuntary treatment of drug and alcohol dependence in New South Wales: an old act and a new direction. Med J Aust 2013;198(11):583-585. [doi: 10.5694/mja12.10387] [Medline: 23919692]
32. Mulheron R. The defence of therapeutic privilege in Australia. J Law Med 2003;11(2):201-213. [Medline: 14655584]
33. Jaffee v. Redmond, 518 U.S. 1. 1996. URL: https://supreme.justia.com/cases/federal/us/518/1/ [accessed 2024-10-11]
34. Goldring J. The Civil Liability Act 2002 (NSW). Judicial Rev 2003;6(3):273-312.
35. Del Villar K, White B, Willmott L, Downie J, Stackpoole C. Does a refusal of treatment need to be informed? towards a resolution of the current confusion in English, Canadian and Australian common law. McGill J Law Health 2023:82. [doi: 10.2139/ssrn.4576144]
36. Malette v Shulman, in 1990 OJ 450. 1990. URL: https://www.globalhealthrights.org/wp-content/uploads/2016/05/MalettevShulman.pdf [accessed 1990-03-30]

XSL•FO
RenderX

37. Siebrasse N, Shulman MV. The requirement of consent in medical emergencies. McGill LJ 1988;34:1080. [doi: 10.1201/b15014-90]
38. Wales NS. Children and Young Persons (Care and Protection) Act 1998. NSW Gov Inf Serv 1998.
39. Children and Young Persons (Care and Protection) Act 1998 (NSW). 1998. URL: https://legislation.nsw.gov.au/view/html/inforce/current/act-1998-157 [accessed 2024-10-11]
40. Townsend R, Luck M. Protective jurisdiction, patient autonomy and paramedics: the challenges of applying the NSW Mental Health Act. Australas J Paramedicine 2009;7(4):1-11. [doi: 10.33151/ajp.7.4.185]
41. Guardianship Act 1987 (NSW). 1987. URL: https://legislation.nsw.gov.au/view/html/inforce/current/act-1987-257 [accessed 2024-10-11]
42. NSW Health. Consent to medical and healthcare treatment manual. NSW Health Consent to Medical and Healthcare Treatment Manual. 2020 Feb 01. URL: https://www.health.nsw.gov.au/policies/manuals/Publications/consent-manual.pdf [accessed 2024-10-17]
43. Guardianship Regulation 2016 (NSW). 2016. URL: https://legislation.nsw.gov.au/view/html/inforce/current/sl-2016-0556 [accessed 2024-10-11]
44. Minor (Property and Contracts) Act 1970 (NSW). 1970. URL: https://legislation.nsw.gov.au/view/whole/html/inforce/current/act-1970-060 [accessed 2024-10-11]
45. Thomas EE, Haydon HM, Mehrotra A, Caffery LJ, Snoswell CL, Banbury A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2022;28(4):301-308. [doi: 10.1177/1357633X20960638] [Medline: 32985380]
46. De CP. Comparative Healthcare Law. London: Routledge; 2013.
47. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021;17(2):218-221 [FREE Full text] [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
48. Fisher K, Davey AR, Magin P. Telehealth for Australian general practice: the present and the future. Aust J Gen Pract 2022;51(8):626-629 [FREE Full text] [doi: 10.31128/AJGP-11-21-6229] [Medline: 35908757]
49. Conlon D, Raeburn T, Wand T. Decision-making processes of a nurse working in mental health, regarding disclosure of confidential personal health information of a patient assessed as posing a risk. Collegian 2021;28(3):261-267. [doi: 10.1016/j.colegn.2020.08.010]
50. Remley TP, Herlihy B, Herlihy SB. The US supreme court decision in Jaffee v. Redmond: implications for counselors. J Couns Dev 2011;75(3):213-218. [doi: 10.1002/j.1556-6676.1997.tb02335.x]
51. Rozel J, Soliman L. Tool, constraint, liability, context. Emerg Psychiatry 2023:401. [doi: 10.1093/med/9780197624005.003.0020]
52. Imlach F, McKinlay E, Middleton L, Kennedy J, Pledger M, Russell L, et al. Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. BMC Fam Pract 2020;21(1):269 [FREE Full text] [doi: 10.1186/s12875-020-01336-1] [Medline: 33308161]
53. Bizzarri JV, Rucci P, Sbrana A, Gonnelli C, Massei GJ, Ravani L, et al. Reasons for substance use and vulnerability factors in patients with substance use disorder and anxiety or mood disorders. Addict Behav 2007;32(2):384-391. [doi: 10.1016/j.addbeh.2006.04.005] [Medline: 16797139]
54. R v Osolin. 1993. URL: https://decisions.scc-csc.ca/scc-csc/scc-csc/en/item/1092/index.do [accessed 2024-10-11]
55. Furlong A, Koggel CM, Levin C. The questionable contribution of psychotherapeutic and psychoanalytic records to the truth-seeking process. In: Confidential Relationships: Psychoanalytic, Ethical, and Legal Contexts. Leiden, The Netherlands: Brill; 2022:13-29.
56. Busby K. Responding to defense demands for clients' records in sexual violence cases: some guidance for record keepers, in confidential relationships. In: Confidential relationships: Psychoanalytic, ethical, and legal contexts. Leiden, Netherlands: Brill; 2003:207-228.
57. L'Heureux-Dube C. The charter of rights and the administration of criminal justice in Canada-where have we been and where shall we go. Ohio St J Crim L 2005;3:473 [FREE Full text]
58. Schwarz K. Secret's out: the truth about Australian privacy law. LSJ: Law Society Journal 2020(69):38-41.
59. Houser SH, Flite CA, Foster SL. Privacy and security risk factors related to telehealth services—a systematic review. Perspect Health Inf Manag 2023;20(1):1f [FREE Full text] [Medline: 37215337]
60. Hanson F, O'Connor S, Walker M, Courtois L. Hacking democracies. 2019. URL: https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2019-05/Hacking%20democracies.pdf [accessed 2024-10-11]
61. Criminal Code Act 1995 (Cth). 2022. URL: https://www.legislation.gov.au/C2004A04868/2022-11-10/text [accessed 2022-11-22]
62. Cybercrime Act 2001 (Cth). URL: https://rm.coe.int/cybercrime-act-2001/16808e70b4 [accessed 2004-09-06]
63. Council of Europe Convention on Cybercrime 2001. URL: https://rm.coe.int/1680081561 [accessed 2024-10-11]
64. Blandford A, Wesson J, Amalberti R, AlHazme R, Allwihan R. Opportunities and challenges for telehealth within, and beyond, a pandemic. Lancet Glob Health 2020;8(11):e1364-e1365 [FREE Full text] [doi: 10.1016/S2214-109X(20)30362-4] [Medline: 32791119]
65. Moodie S. Intoxication and minors-what role for the law. Newcastle L 2008;11:123.

XSL•FO
RenderX

66. Health Practitioner Regulation National Law Act 2009 (NSW). 2009. URL: https://legislation.nsw.gov.au/view/html/inforce/current/act-2009-86a [accessed 2024-10-11]
67. Fitzgerald J, Burrell C, Bull PW. Health practitioner regulation national law act 2009: attitudes and compliance of the chiropractic profession one year on. Chiropractic J Aust 2012;42(2):51-59.
68. Telecommunications (Interception and Access) Act 1979 (Cth). URL: https://www.homeaffairs.gov.au/nat-security/files/telecommunications-interception-access-act-1979-annual-report-16-17.pdf [accessed 2024-10-11]
69. Privacy Act 1988 (Cth). URL: https://www.legislation.gov.au/C2004A03712/2019-08-13/text [accessed 2019-08-13]
70. Civil Liability Act 2002 (NSW). URL: https://legislation.nsw.gov.au/view/html/inforce/current/act-2002-022 [accessed 2024-10-11]
71. Ministry of Justice New Zealand. Evidence Act 2006 of New Zealand. Evidence Act 2006 New Zealand. 2023 Nov 28. URL: https://www.legislation.govt.nz/act/public/2006/0069/latest/DLM393463.html [accessed 2024-10-17]
72. Howard K, McCann C, Dudley M. What is communication assistance? Describing a new and emerging profession in the New Zealand youth justice system. Psychiatr Psychol Law 2020;27(2):300-314. [doi: 10.1080/13218719.2020.1719378] [Medline: 32944128]
73. Fibbins H, Czosnek L, Stanton R, Davison K, Lederman O, Morell R, et al. Self-reported physical activity levels of the 2017 Royal Australian and New Zealand College of Psychiatrists (RANZCP) conference delegates and their exercise referral practices. J Ment Health 2020;29(5):565-572. [doi: 10.1080/09638237.2018.1521935] [Medline: 30322334]
74. Thomson G, Wilson N. Screening for ministerial appointments? Systems failure in Peter Dunne's appointment as a New Zealand revenue and associate health minister. N Z Med J 2005;118(1226):U1765. [Medline: 16311619]
75. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020;22(12):e24087. [doi: 10.2196/24087] [Medline: 33147166]
76. Powell K. Publishing: foreign tongues. Nature 2012;487(7405):129-131. [doi: 10.1038/nj7405-129a]
77. Douglas M, Keyes M, McKibbin S, Mortensen R. The HCCH judgments convention in Australian law. Federal Law Review 2019;47(3):420-443. [doi: 10.1177/0067205x19856503]
78. Weekes L, Ramzan I. Prescription of compounded ophthalmic medications—a pharmacy perspective. Ramzan, CEOptom-20-09-RV. R. 2020 Mar 08. URL: https://pubmed.ncbi.nlm.nih.gov/32253769/ [accessed 2024-10-17]
79. Smith M, Mann M, URBAS G. Biometrics, Crime and Security. London, United Kingdom: Routledge; 2018.
80. Swanson B. Telehealth and the Law. HIC 99: Proc 1999:315-320.

Abbreviations

Cth: the Commonwealth
NSW: New South Wales
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SUD: substance use disorder
TGA: Therapeutic Goods Administration

Edited by A Mavragani; submitted 10.02.23; peer-reviewed by T Aslanidis, R Davis, WB Lee, M Nolan; comments to author 31.10.23; revised version received 22.01.24; accepted 18.09.24; published 12.11.24.

Please cite as:
Jefferies M, Graham R, Tracy M, Read S, Eslam M, Douglas MW, George J
Telehealth Use and Legal Considerations in Drug Health Services During Pandemics: Systematic Scoping Review
J Med Internet Res 2024;26:e46394
URL: https://www.jmir.org/2024/1/e46394
doi:10.2196/46394
PMID:

©Meryem Jefferies, Robert Graham, Marguerite Tracy, Scott Read, Mohammed Eslam, Mark W Douglas, Jacob George. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning Methods to Personalize Persuasive Strategies in mHealth Interventions That Promote Physical Activity: Scoping Review and Categorization Overview

Annette Brons[1,2,3], MSc; Shihan Wang[2], PhD; Bart Visser[3], PhD; Ben Kröse[1,4], PhD; Sander Bakkes[2], PhD; Remco Veltkamp[2], PhD

[1]Digital Life Center, Amsterdam University of Applied Sciences, Amsterdam, Netherlands
[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
[3]Centre of Expertise Urban Vitality, Amsterdam University of Applied Sciences, Amsterdam, Netherlands
[4]Department of Computer Science, University of Amsterdam, Amsterdam, Netherlands

Corresponding Author:
Remco Veltkamp, PhD
Department of Information and Computing Sciences
Utrecht University
Princetonplein 5
Utrecht, 3584 CC
Netherlands
Phone: 31 621156976
Email: R.C.Veltkamp@uu.nl

Abstract

Background: Although physical activity (PA) has positive effects on health and well-being, physical inactivity is a worldwide problem. Mobile health interventions have been shown to be effective in promoting PA. Personalizing persuasive strategies improves intervention success and can be conducted using machine learning (ML). For PA, several studies have addressed personalized persuasive strategies without ML, whereas others have included personalization using ML without focusing on persuasive strategies. An overview of studies discussing ML to personalize persuasive strategies in PA-promoting interventions and corresponding categorizations could be helpful for such interventions to be designed in the future but is still missing.

Objective: First, we aimed to provide an overview of implemented ML techniques to personalize persuasive strategies in mobile health interventions promoting PA. Moreover, we aimed to present a categorization overview as a starting point for applying ML techniques in this field.

Methods: A scoping review was conducted based on the framework by Arksey and O'Malley and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) criteria. Scopus, Web of Science, and PubMed were searched for studies that included ML to personalize persuasive strategies in interventions promoting PA. Papers were screened using the ASReview software. From the included papers, categorized by the research project they belonged to, we extracted data regarding general study information, target group, PA intervention, implemented technology, and study details. On the basis of the analysis of these data, a categorization overview was given.

Results: In total, 40 papers belonging to 27 different projects were included. These papers could be categorized in 4 groups based on their dimension of personalization. Then, for each dimension, 1 or 2 persuasive strategy categories were found together with a type of ML. The overview resulted in a categorization consisting of 3 levels: dimension of personalization, persuasive strategy, and type of ML. When personalizing the timing of the messages, most projects implemented reinforcement learning to personalize the timing of reminders and supervised learning (SL) to personalize the timing of feedback, monitoring, and goal-setting messages. Regarding the content of the messages, most projects implemented SL to personalize PA suggestions and feedback or educational messages. For personalizing PA suggestions, SL can be implemented either alone or combined with a recommender system. Finally, reinforcement learning was mostly used to personalize the type of feedback messages.

Conclusions: The overview of all implemented persuasive strategies and their corresponding ML methods is insightful for this interdisciplinary field. Moreover, it led to a categorization overview that provides insights into the design and development of personalized persuasive strategies to promote PA. In future papers, the categorization overview might be expanded with additional layers to specify ML methods or additional dimensions of personalization and persuasive strategies.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47774) doi:10.2196/47774

KEYWORDS

artificial intelligence; exercise; mobile app; adaptive; tailoring; supervised learning; reinforcement learning; recommender system

Introduction

Background and Related Work

Regular physical activity (PA) decreases the risk of several diseases, improves quality of life and well-being, and improves mental health. Despite these positive effects, physical inactivity is a serious and still growing worldwide problem. Worldwide, 28% of adults and 81% of adolescents are insufficiently physically active. They do not meet the World Health Organization recommendations of at least 150 minutes per week of moderately intense PA for adults and 60 minutes of moderate PA per day for adolescents [1].

Mobile health (mHealth) interventions have been shown to be effective in promoting PA [2-5]. Intervention success can be improved by implementing effective persuasive strategies, also called behavior change techniques, such as goal setting, monitoring, rewards, reminders, education, activity suggestions, and feedback [6-8]. Moreover, personalized mHealth interventions are more effective in promoting PA than nonpersonalized interventions [9-12]. In mHealth, personalization means adapting intervention strategies to individual characteristics such as gender, disease, coping strategy, current location, current PA level, or performed PA [13,14]. The combination of success factors for personalization and for persuasive strategies can take place on several dimensions, such as the type of feedback, content of the messages, timing of the messages, rewards, and personal settings [9,11,15].

To personalize persuasive health-promoting interventions, machine learning (ML) has become increasingly popular in recent years [16,17]. ML is an important subfield of artificial intelligence, with numerous applications in domains such as robotics, natural language processing, and computer vision. The key methods of ML include supervised learning (SL), unsupervised learning (UL), and reinforcement learning (RL). Several kinds of ML techniques have been used in adaptive interventions that promote PA. SL, in which the prediction is based on labeled data, can be used for classification to predict a category and for regression to predict a quantity. SL for classification, for instance, was used in a chatbot-based digital coach to improve PA levels [18] and a mobile app with personalized feedback to promote PA in cardiovascular disease rehabilitation [19]. SL was used for regression to predict low blood sugar levels in patients with diabetes based on their previous blood glucose levels and performed PA [20]. On the other hand, UL analyzes unlabeled data and can be implemented to cluster data or build a recommender system (RS). An RS provides suggestions that fit the user's needs or preferences. Clustering, for example, was applied in a PA advisor system to increase PA [21], whereas an RS was developed to recommend adequate educational content to users to improve their knowledge regarding PA and a healthy diet [22]. Finally, there is RL, which is an ML method based on rewarding desired actions. RL is often used to personalize the timing of messages, also called a just-in-time adaptive intervention [23]. To choose the appropriate RL algorithm, problems are categorized as Markov decision processes (MDPs) or as multi-armed bandit (MAB) and contextual bandit problems. MDPs can model decision-making problems when the results are partly random and partly controlled by the user. MDP-based RL, for example, was used in an adaptive intervention to personalize motivational coaching messages to promote healthy behavior such as PA [24]. MAB is used when a series of choices has to be made without knowing personal preferences and the consequences of a choice. MAB, for instance, was the basis of an adaptive personalized messaging smartphone app to improve PA levels [25].

There are many studies [5,26-32] that show an overview of the effectiveness of mHealth interventions on PA, each focusing on different target groups and outcomes. Several recent reviews [9,11,32-36] have zoomed in on personalizing persuasive interventions, but none of them have focused on personalization with the use of ML. On the other hand, some reviews [16,37,38] have addressed the use of ML to personalize PA-promoting interventions more generally but have not studied the combination of ML and personalizing persuasive strategies. To start with, Chaudhari et al [37] reviewed personalized PA interventions using ML and other methods but focused on personalizing the timing of PA-promoting messages instead of other persuasive strategies such as content of messages or level of difficulty. In addition, Oyebode et al [38] and Goh et al [16] studied ML methods to personalize interventions for health and well-being—including PA—in which persuasive strategies were implemented but did not focus either on the ML methods used to personalize these persuasive strategies.

Objectives and Contribution

To the best of our knowledge, no review has addressed personalizing persuasive strategies using ML in the field of promoting PA yet. Therefore, insights into the characteristics of studies that did implement ML to personalize persuasive PA-promoting strategies and an indication of missing information is already a contribution to this field. Moreover, considering the interdisciplinary nature of this field, it would be helpful for interventions to be designed and developed in the future to have some guidance in choosing the best-fitting ML technique considering the persuasive strategies. Designers and researchers in the field of persuasive strategies lack such an insightful overview. Therefore, the first objective of our study was to provide an overview of implemented persuasive strategies and their corresponding implemented ML techniques in this field. We conducted a scoping review to reach this goal as this is a useful approach to analyze knowledge gaps and identify key characteristics of studies [39]. Moreover, we aimed to categorize these findings and build a categorization overview as a starting point for applying ML techniques to personalize persuasive strategies in PA-promoting interventions.

XSL•FO
RenderX

Methods

Overview

A scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) criteria [40]. The PRISMA-ScR checklist is provided in Multimedia Appendix 1. The scoping review was conducted using the framework by Arksey and O'Malley [41], later extended by Levac et al [42], which consists of the following five stages: (1) identifying the research question; (2) identifying relevant studies; (3) selecting studies; (4) charting the data; and (5) collating, summarizing, and reporting the results. In addition to these steps, we provide a categorization overview based on the results of the scoping review.

Step 1: Identifying the Research Question

Corresponding to the 2 objectives, our research questions were defined as follows: (1) What ML techniques have been used to personalize persuasive strategies in mHealth interventions that promote PA? (2) How to guide future designers, developers, and researchers in choosing ML techniques to personalize persuasive strategies in PA-promoting interventions?

Step 2: Identifying Relevant Studies

We searched for studies published until February 24, 2023, in 3 databases: Scopus, Web of Science, and PubMed. These databases were searched because they cover relevant studies in the digital health domain. To find relevant papers for answering the research questions, all studies should match the following semantics: *digital persuasive interventionspromoting PA* that are *personalized* to participants' needs through *machine learning* techniques. We searched for these concepts in the title, abstract, and keywords. The exact query that was run on all databases can be found in Multimedia Appendix 2. In addition, the reference lists of included papers and relevant review papers were searched by hand. These papers were categorized as *manual* in the identification section of the PRISMA-ScR flow diagram.

Step 3: Selecting the Studies

After the initial search, duplicates were removed. The remaining articles were screened by 2 reviewers (AB and SW) in 2 phases. First, titles and abstracts were screened using the ASReview software [43,44]. Studies meeting at least one of the exclusion criteria (Textbox 1) were removed. ASReview uses ML to efficiently screen large amounts of titles and abstracts [45]. Moreover, the studies are only reviewed based on content rather than irrelevant metadata such as author and journal names as only titles and abstracts are shown [46]. Both reviewers started with the same set of papers and had to choose at least 2 papers to label before screening to initially train the algorithm. Moreover, the algorithm improves itself based on the labels of the reviewer. Thus, the order of papers shown for screening differs between the reviewers as it depends on both the reviewers' choice of initial papers to label and the labels given during the screening process. On the basis of several relevant and irrelevant studies labeled by the reviewers, a model is trained to sort the remaining studies by order of relevance. When labeling all studies, the model constantly improves itself. Thus, both reviewers did manually screen all titles and abstracts, but because of the order of relevance, it was more efficient than the usual method in a random order.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Articles: peer-reviewed articles, conference proceedings, or book chapters on qualitative, quantitative, and mixed method studies or protocols
- Language: English
- Application: digital intervention, either stand-alone or hybrid with a health care provider, with at least one personalized persuasive strategy
- Machine learning (ML) technique: supervised ML, unsupervised ML, or reinforcement learning
- Health behavior promoted by the intervention: physical activity (including walking, running, cycling, performing exercises, and performing sports activities) either alone or as part of several health behaviors

Exclusion criteria

- Articles: conference abstracts, reviews, letters, editorials, comments, non–peer-reviewed articles or book chapters, white papers, or articles with no full text available
- Language: all other languages
- Application: nonadaptive intervention or persuasive strategy not specified
- ML technique: rule-based if-then systems, predefined formulas, manual personalization, ML method not specified, computer vision, natural language processing, or robotics
- Health behavior promoted by the intervention: health behaviors other than physical activity

Next, the full texts of all the remaining studies were reviewed. Studies meeting all the inclusion criteria (Textbox 1) were included. When both reviewers did not agree, a third reviewer (SB) was asked for the final decision. Both reviewing rounds were conducted blinded, meaning that both reviewers had no access to each other's findings before completing the full task. A dataset consisting of the screened papers together with the reason for exclusion can be found here [47]. Afterward, all the included papers were categorized based on the projects they belonged to. As project names, we used the name of the

XSL•FO
RenderX

intervention or distinctive key terms included in the study. For each project that included several papers, the paper with the most detailed information regarding the implemented persuasive strategies and ML method was labeled as the main paper. From the Target Group subheading in the Results section onward, we only use this main paper as a reference to keep the report clear.

Step 4: Charting Data

Data extraction was performed by one reviewer (AB) and verified by a second reviewer (SW). The extracted data were recorded in a Microsoft Excel form that was tested by 2 reviewers (AB and SW) beforehand. We extracted data regarding general information of the paper, target group, PA intervention, and the implemented technology. A detailed overview of the extracted information is shown in Textbox 2.

Textbox 2. Extracted data from the full texts.

General

- Title
- Authors
- Year of publication
- Type of paper (journal article, conference proceeding, or book chapter)
- Objective
- Main results

Intervention target group

- Adults or children
- Healthy or with disabilities

Intervention

- Description of the intervention
- Persuasive strategy
- Physical activity (PA) application field
- Underlying theories
- Stand-alone or hybrid with human interaction

Technology

- Machine learning (ML) category and method
- ML technology
- Gamification or not
- Types of data as input for algorithm
- Features used for adaptation
- Dimension of personalization
- Platform

Study

- Study type
- Real users or simulation
- Outcomes regarding PA or ML technique

Step 5: Collating, Summarizing, and Reporting the Results

We presented an overview of the literature search using the PRISMA-ScR flow diagram. Moreover, we analyzed and described the extracted data from the included papers. We focused on the developed intervention, target group, implemented ML method, dimension of personalization, underlying persuasive strategy, type of study, and main results. Furthermore, we summarize the underlying psychological theories and some technical details of the developed interventions.

Building the Categorization Overview

On the basis of the resulting overview of implemented ML methods to personalize persuasive strategies in PA-promoting

XSL•FO
RenderX

interventions, we provided a categorization overview. This consists of 3 levels, starting with the dimension of personalization. The persuasive strategies are shown per dimension together with the ML method that was implemented in most of the projects. In addition, we show the projects that support the highlighted ML method for that specific combination of persuasive strategy and dimension of personalization.

The ML methods shown in the categorization overview are based on the number of projects that implemented that ML method for a specific persuasive strategy within a dimension of personalization. As the outcome measures of all the included projects differed widely and not all projects had published results yet, we did not consider the results of the included projects in the categorization overview.

Results

Overview of Literature Search

The search query yielded 400 papers in total. These studies were scanned for duplicates, which resulted in a total of 75.3% (301/400) of unique papers. The abstracts of the unique papers were assessed against the exclusion criteria, resulting in a shortlist of 44.9% (135/301) of the papers. Main reasons for excluding abstracts were no promotion of PA, no personalized intervention, no use of ML to personalize the intervention, and review papers. Reviewing the full texts of these 135 shortlisted papers resulted in 40 (29.6%) included papers, being 13.3% (40/301) of the screened unique papers and belonging to 27 projects. Most excluded papers (53/95, 56%) discussed interventions that were personalized without using ML. A lot of the other papers were removed because the intervention did not promote PA (23/95, 24%) or there was no intervention yet (9/95, 9%). The PRISMA-ScR flow diagram with the aforementioned numbers is shown in Figure 1.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram. ML: machine learning; PA: physical activity.

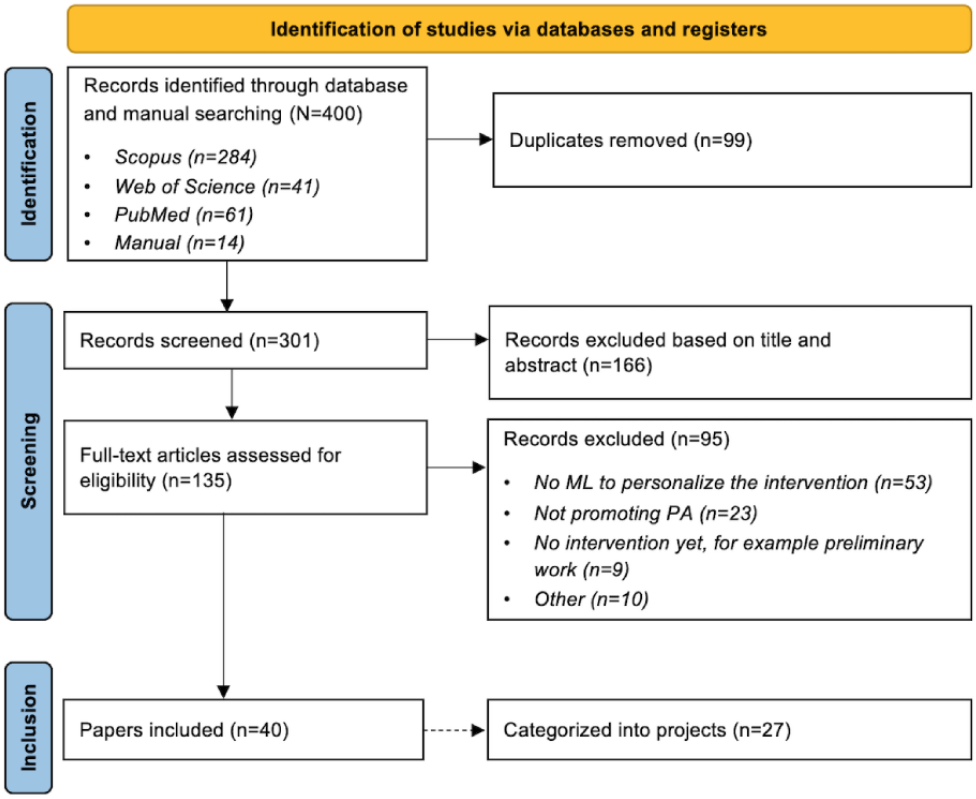

Overview of Included Projects

In total, 40 papers that used ML to personalize persuasive strategies in PA-promoting interventions were included. These papers belonged to 27 projects. All projects and their corresponding papers are shown in Table 1 categorized by the dimension of personalization. Project Power2DM was the only intervention that belonged to 2 dimensions of personalization and, therefore, is included twice to give a complete overview of each dimension.

XSL•FO
RenderX

Table 1. Overview of all the included papers categorized by the dimension of personalization. In the Relevant papers column, the main article is mentioned first.

Project name	Relevant papers	Intervention target group	Persuasive strategy	ML[a] category and method	ML technique	Features used for adaptation	Study type	Study outcome
Dimension of personalization: timing of messages								
Heartsteps	Tomkins et al [48] and Liao et al [49]	Adults with hypertension	Reminders	RL[b]: contextual bandit	IntelligentPooling	Performed PA[c], location, weather, day of the week, and current time	Simulation (to improve algorithm) and pilot	All participants except 1 improved their step count with the intervention; 26% improvement in step count in simulations; IntelligentPooling algorithm performed better than Thompson sampling algorithm
Power2DM[d]	Gönül et al [24] and Glachs et al [50]	Adults with diabetes	Reminders (timing), motivational messages, and feedback (type of feedback)	RL: MDP[e] and contextual bandit	Customized version of eligibility traces	Time, location, PA status, phone screen status, emotional status, and number of reminders and motivational messages that have been sent	Simulation	Number of required messages to achieve the goal was lower for RL MDP than for general RL
PAUL	Wang et al [51,52] and Sporrel et al [53]	Adults	Reminders	RL: MDP	Policy gradient RL (REINFORCE algorithm)	Performed PA, current time, and calendar	Feasibility study	83.3% of reminders provoked PA within 50 min; 66.7% of PAs were performed within 5 hours after the reminder
Ally	Mishra et al [18], Kramer et al [54], and Künzler et al [55]	Adults	Goal setting and monitoring	SL[f]: classification	Logistic regression	Date, time, battery level of phone, device interaction, and PA	Simulation and pilot	Higher receptivity in ML group than in the controls (simulation); participants who were more receptive were more likely to achieve their goals (pilot)
U4Fit	Pilloni et al [56]	Adults	Feedback and monitoring	SL: classification	RF[g]	Covered distance, workout duration, rest time, average speed, burned calories, and time elapsed since previous workout	Simulation	In 70% of cases, the ML intervention correctly predicted whether a sportsman would stop exercising and need feedback or motivational messages
Dimension of personalization: content of messages								
Quantified Self	Erdeniz et al [57]	Adults	PA suggestion	SL: classification; UL[h]: recommender system	K-nearest neighbor; collaborative filtering	Age, gender, location, chronic diseases, oxygen saturation, heart rate, and performed PA	Pilot	≥75% of PA recommendations were accurate

XSL•FO
RenderX

Project name	Relevant papers	Intervention target group	Persuasive strategy	ML[a] category and method	ML technique	Features used for adaptation	Study type	Study outcome
WellHealth	Lo et al [58]	Adults with chronic neck or back pain	PA suggestion	SL: classification	Neural network	Present condition, history of condition, past investigation, social history, drug history, and 24-hour PA pattern	Observational study	Increased time spent on therapeutic exercises
MyBehavior	Rabbi et al [59,60]	Both healthy adults and adults with obesity	PA suggestion	RL: multi-armed bandit; UL: clustering	EXP3[i]; BIRCH[j] online clustering algorithm	Current location, performed PA, and PA behavior pattern	RCT[k]	Significantly (P=.05) better improvement in PA in ML-personalized intervention compared to nonpersonalized intervention
Mining Minds	Banos et al [61]	Adults	PA suggestion	SL: classification; UL: recommender system	SVM[l] and Gaussian mixture modeling; content-based filtering	Accelerometer data and feedback from users given on recommendations to improve the system	Pilot test	Average response time after recommendation was between 3 and 40 min
PA Advisor	Li et al [21]	Adults	PA suggestion	SL: classification; UL: clustering	DT[m], SVM, and RF; hierarchical clustering with Gower distance	Location, PA, and general physiological characteristics	Design	No results yet
Pro-Fit	Dharia et al [62]	Adults	PA suggestion	SL: classification; UL: recommender system	Gradient boosting and DT; collaborative filtering	Accelerometer data, feedback from users on given recommendations, and user profile	Simulation (to improve algorithm) and user study	Most participants felt that the ML-personalized intervention was successful in motivating them to increase PA levels
WalkPal	Sansrimahachai [63]	Older adults	PA suggestion	SL: classification	Neural network	Heart rate, location, and walking statistics	Simulation	Reasonable prediction accuracy of exercise min with a mean average error of –17 to +17 weekly exercise min
Blockchain	Jamil et al [64]	Both healthy adults and adults with obesity	PA suggestion	SL: classification	SVM	BMI, performed PA, and user activity	Simulation	SVM was the best-performing ML method with an accuracy of 92%
PA Recommendation	Zhao et al [65]	Adults	PA suggestion	SL: classification; UL: recommender system	SVM; collaborative filtering	Age, purpose of exercise, vegetarian or not, wake-up time, video game preference, workout with trainer or not, gender, BMI, favorite type of musing during exercise, and favorite sports brand	Simulation and pilot	Prediction accuracy of ≥70% in simulation

XSL•FO
RenderX

Project name	Relevant papers	Intervention target group	Persuasive strategy	ML[a] category and method	ML technique	Features used for adaptation	Study type	Study outcome
Wireless Health	Suh et al [66]	Adults	PA suggestion, feedback, and social comparison	SL: classification (personalized activity); UL: recommender system (feedback and social comparison)	DT; collaborative filtering	Heart rate, accelerometer data, pressure, and gyroscope data	Pilot	Exercises were performed more accurately, and participants reported increase in motivation; increase of 25.54% in calories burned compared to non-personalized training protocol
selfBACK	Mork and Bach [67] and Sandal et al [68]	Adults with chronic low back pain	PA suggestion and education	SL: classification	Case-based reasoning	Demographics, medication, quality of life score, sleep, mood, stress, pain, and PA	Design	No results yet
CoCare	Cerón-Ríos et al [22]	Both healthy adults and adults with obesity	Education	SL: classification; UL: recommender system	DT; collaborative filtering	Current location, indoor location, friends, date, and daily schedule	Feasibility study	90% of recommendations were accurate
drBart	Pelle et al [69]	Adults with osteoarthritis	Goal setting	UL: recommender system	Contextual multi-armed bandit approach	Gender, age, height, weight, symptoms, quality of sleep, and maximum walking distance	Design	No results yet
HNGW	Dijkhuis et al [70]	Adults	Feedback	SL: classification	RF	Steps	Pilot test	RF was the best-performing ML method with an accuracy of 93% and an F_1-score of 0.9
Healthy Behavior Messages	Kadri et al [71]	Adults	Feedback	SL: classification	DT	Accelerometer data	Simulation	DT performed better than the BiLSTM[n] classifier with an F_1-score of 62%
MedFit	Prabhu et al [19]	Adults	Feedback	SL: classification	SVM	Accelerometer data	Design	No results yet
BeWell+	Lane et al [72]	Adults	Feedback	SL: classification	Naïve Bayes	PA	Design	Intervention can operate successfully on consumer smartphones, and users understand the personalized feedback and respond by taking steps
Dimension of personalization: type of messages								

XSL•FO
RenderX

Project name	Relevant papers	Intervention target group	Persuasive strategy	ML[a] category and method	ML technique	Features used for adaptation	Study type	Study outcome
Diabetes Messages	Hochberg et al [73] and Yom-Tov et al [74]	Adults with type 2 diabetes	Feedback	RL: contextual bandit	Linear regression with Bolzmann sampling	Demographics, PA, and reaction to messages	Pilot	Significant difference (*P*=.04)between personalized and nonpersonalized version: increase in PA in the ML-personalized intervention and decrease in PA in the nonpersonalized intervention
Personalization Paradox	Zhu et al [75]	Adults	Feedback, monitoring, and social comparison	RL: multi-armed bandit	__[o]	Performed PA and self-reported motivation	User study	No significant difference in step count between ML-personalized intervention and nonpersonalized intervention; significant difference (*P*=.004) in motivation between ML-personalized intervention and nonpersonalized intervention
Diamante	Aguilera et al [25] and Figueroa et al [76-78]	Adults	Feedback	RL: multi-armed bandit	Thompson sampling	Not discussed	Design	No results yet
Power2DM[d]	Gönül et al [24] and Glachs et al [50]	Adults with diabetes	Reminders (timing), motivational messages, and feedback (type of message)	RL: MDP and contextual bandit	Customized version of eligibility traces	Time, location, PA status, phone screen status, emotional status, and number of interventions sent for planned activity	Simulation	RL MDP performed better than standard RL; number of required messages to achieve the goal was lower for RL MDP than for general RL
Dimension of personalization: level of difficulty of PA								
Maze VR	Huber et al [79]	Adults	PA that fit the capabilities	RL: MDP	Deep RL	ECG[p] and Borg scale	Proof of concept	Prototype was able to adapt both the cognitive and physical difficulty of the game for participants who were unsatisfied with a very easy level; the algorithm helped keep participants in a state of flow

XSL•FO
RenderX

Project name	Relevant papers	Intervention target group	Persuasive strategy	ML[a] category and method	ML technique	Features used for adaptation	Study type	Study outcome
Pathologys	Aguilar et al [80]	Adults	PA that fit the capabilities	SL: classification	DT and neural network	Heart rate and performance	Pilot	Neural network had greater impact on progress and motivation than DT

[a]ML: machine learning.
[b]RL: reinforcement learning.
[c]PA: physical activity.
[d]Indicates that the project was included in the table twice because it belonged to 2 dimensions of personalization.
[e]MDP: Markov decision process.
[f]SL: supervised learning.
[g]RF: random forest.
[h]UL: unsupervised learning.
[i]EXP3: Exponential-weight algorithm for Exploration and Exploitation.
[j]BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies.
[k]RCT: randomized controlled trial.
[l]SVM: support vector machine.
[m]DT: decision tree.
[n]BiLSTM: bidirectional long short-term memory.
[o]Not applicable.
[p]ECG: electrocardiogram.

Projects

Project Heartsteps [48,49] developed a digital coach to improve the PA levels of patients with hypertension by providing PA reminders. Project PAUL [51-53] developed a digital coach that provided PA reminders as well in addition to several other nonpersonalized persuasive strategies. The digital coaches of projects Diabetes Messages [73,74], Diamante [25,76-78], BeWell+ [72], MedFit [19], Healthy Behavior Messages [71], and HNGW [70] all provided personalized feedback to increase PA levels in general, all having their own target group. The digital coach of project Ally [18,54,55] provided personalized goal-setting and monitoring messages. Project U4Fit [56] did not design a digital coach itself but developed an algorithm that predicted whether a sportsman would stop exercising so that their human coach could contact them.

In total, 41% (11/27) of the projects suggested personalized PAs. A total of 55% (6/11) of them focused on a PA plan or schedule, whereas 45% (5/11) of them included specific exercises or activities. The Quantified Self project [57], Mining Minds [61], PA Advisor [21], and PA Recommendation [65] all suggested personalized activities to promote PA in general, the latter focusing on PA recommendations for exergames. Projects WalkPal [63], Pro-Fit [62], and Wireless Health [58] all focused on training. They included a walking exercise plan [63], fitness schedule [62], and interval training schedule [66]. Wireless Health also provided personalized music recommendations and social comparison to motivate users to achieve the proposed training schedule. Both MyBehavior [59,60] and Blockchain [64] combined the suggested PAs with a dietary plan to increase PA and decrease calorie intake. In WellHealth [63] and selfBACK [67,68], the activity schedule was part of a self-management plan for people with chronic low back pain. The first focused on adherence to exercise therapy, whereas the aim of the second project was to prevent chronic pain.

The drBart project [69] focused on self-management as well by providing personal goals. This project aimed to optimize nonsurgical care. Project CoCare [22] suggested personalized multimedia content about PA and a healthy diet. Project Personalization Paradox [75] attempted to increase PA motivation by providing social comparison. Both Maze VR [79] and Pathologys [80] developed exergames with dynamic difficulty adjustment to promote PA. Finally, project Power2DM [24,50] personalized the timing and frequency of motivational coaching messages to promote healthy behavior, including PA.

Target Group

All 27 projects designed adaptive PA-promoting interventions for adults. In total, 26% (7/27) of them were designed for adults with a disease, such as cardiovascular disease [19,48], diabetes [24,74], osteoarthritis [69], and chronic neck or back pain [58,67]. A total of 4% (1/27) of the projects were designed for older adults specifically [63]. In total, 59% (16/27) of the interventions were designed for adults in general [18,21,25,51,56,57,61,62,65,66,70-72,75,79,80], and 11% (3/27) of the interventions had both healthy adults and adults with obesity as their target groups [22,60,64].

Persuasive Strategies

In total, 4 different dimensions of personalization were found in all projects. Except for 4% (1/27) of the projects [24], all projects included only 1 dimension of personalization. Most of the projects (17/27, 63%) personalized the content of the

XSL•FO
RenderX

messages [19,21,22,57,58,60-67,69-72]. A total of 19% (5/27) of the projects focused on the timing of the messages [18,24,48,51,56], whereas 15% (4/27) of the projects personalized the type of feedback in the messages [24,25,74,75]. For instance, this regarded positive or negative feedback and feedback in which someone's results were compared to someone else's or to their own previous results. In addition, 7% (2/27) of the projects personalized the level of difficulty of PAs [79,80], meaning that the difficulty of PAs was aligned with the capabilities of the participants.

In total, 9 different personalized persuasive strategies were incorporated into the developed interventions: reminders, goal setting, monitoring, motivational messages, feedback, social comparison, PA suggestions, education, and PA that fits capabilities. Most projects that personalized the timing of messages (3/5, 60%) included reminders [24,48,51]. Moreover, goal setting [18], monitoring [18,56], and feedback [56] were applied in this group. Most projects that personalized the content of messages (11/17, 65%) used PA suggestions as a persuasive strategy [21,57,58,60-67]. Of those projects, 18% (2/11) also personalized feedback and used social comparison [66] and education [67]. Feedback was personalized in another 15% (4/27) of the projects [19,70-72], and education was the persuasive strategy in another project [22]. The remaining project focused on personalized goal setting [69]. All projects that personalized the type of messages (4/27, 15%) included feedback as a persuasive strategy [24,25,74,75]. In addition, they included motivational messages [24], monitoring, and social comparison [68]. The 7% (2/27) of the projects that personalized the difficulty of PA used PA suggestions that fit the participants' capabilities as a persuasive strategy [79,80]. This means that, for instance, easier PAs were suggested when previous activities appeared to be too hard for the participants.

ML Methods

The ML methods that were implemented in the PA-promoting interventions can be divided into 3 categories: RL, SL, and UL. A variety of methods was used within these categories. To start with, 30% (8/27) of the projects used RL. These projects regarded personalized timing of reminders [24,48,49,51], personalized feedback types [24,25,74,75], personalized PA suggestions [60], and personalized difficulty of PAs [79]. Most (6/8, 86%) used RL based on contextual bandit or MAB problems [24,25,48,60,74,75]. MDP was the basis for the remaining projects that implemented RL algorithms to personalize their persuasive strategies [24,51,79]. One of the projects combined both strategies [24]. Except for project Personalization Paradox, all projects reported the specific RL technique that was used to personalize the persuasive strategies. All projects implemented other specific RL techniques: customized version of eligibility traces [24], policy gradient RL [51], deep RL [79], IntelligentPooling [48], Exponential-weight algorithm for Exploration and Exploitation (EXP3) [60], Thompson sampling [25], and linear regression with Bolzmann sampling [74].

SL was implemented in 56% (15/27) of the projects, and they all used SL for classification. These projects included personalized timing of monitoring and feedback messages [18,56]; personalized level of difficulty of PA [80]; and personalizing the content of the messages, including personalized PA suggestions [57,58,61-67], education [22,67], and feedback [19,67,70-72]. A variety of techniques was applied in the group that used SL: logistic regression [18], random forest (RF) [56,70], neural networks (NNs) [58,63,80], k-nearest neighbor [57], decision tree (DT) [22,62,66,71,80], naïve Bayes [72], case-based reasoning [67], and support vector machine [19,61,64,65]. A total of 19% (5/27) of the projects, which applied SL, combined this with UL for RS [22,57,61,62,66]. Except for 1, all of them personalized the content of the messages. Most (4/5, 80%) used collaborative filtering [22,57,62,66], whereas one project used content-based filtering [61]. In addition, 7% (2/27) of the projects combined SL with UL for clustering using hierarchical clustering with Gower distance [21] and Balanced Iterative Reducing and Clustering Using Hierarchies online clustering [60]. Only 4% (1/27) of the projects implemented UL for RS without SL [69]. They used a contextual MAB approach to personalize goal setting.

Study Outcomes

A total of 22% (6/27) of the projects had PA-related outcomes as the main study results. In total, 67% (4/6) of these projects compared the PA-related outcome between an ML-personalized intervention and a nonpersonalized intervention [60,66,74,75]. In 75% (3/4) of these projects, the ML-personalized version performed better in terms of performed PA than the nonpersonalized version [60,66,74]. This difference was reported to be significant with respectively P=.05 [60] and P=.04 [74] in 67% (2/3) of the projects. No significant difference in PA levels was found in the remaining project, although a significant difference (P=.004) in motivation was found in favor of the ML-personalized intervention [75]. In addition, 15% (4/27) of the projects found positive effects regarding PA when using the ML-personalized intervention without comparison to a nonpersonalized intervention. Improved step count [48], increased time spent on exercises [58], and successful motivation of participants to increase PA levels [62] were reported. Finally, 4% (1/27) of the projects reported that personalizing the difficulty of PAs in an exergame helped keep participants in a state of flow [79].

In total, 33% (9/27) of the projects reported outcomes regarding their developed algorithm as the main study results. A total of 56% (5/9) of these projects focused on the accuracy of their classifications. Accurate PA recommendations with 75% [57], 70% [65], and 92% [64] accuracy were reported, as well as an accuracy of 90% for predicting personalized educational content [22] and 70% for predicting the need for motivational feedback [56]. Moreover, an accuracy of 93% for predicting the probability of achieving a daily step goal [55] and an F_1-score of 62% for predicting the performed PAs [71] were found. In addition, 11% (1/9) of the projects reported a mean average error of –17 to +17 minutes per week on the prediction of performed exercise [63]. A total of 22% (2/9) of the projects reported the response time after a message. The average time of performing PA after receiving a recommendation was between 3 and 40 minutes in one of the projects [61], whereas 66.7% of the PAs were performed within 5 hours after the

XSL•FO
RenderX

reminder in another project [51]. The latter also reported that 83.3% of the reminders provoked PA.

Only 19% (5/27) of the projects compared specific ML techniques, all comparing techniques from the same type of ML. One study reported that the number of messages needed to achieve a goal was lower for the MDP RL algorithm than for general RL algorithms [24]. In total, 40% (2/5) of the projects compared DT with another classifier. DT was preferred over bidirectional long short-term memory (BiLSTM) in one of those projects [71], whereas NNs achieved better results in the other project [80]. DT, logistic regression, support vector machine, and k-nearest neighbor were compared in 40% (2/5) of the projects. In the first project, DT performed best [64]. In the second project, adaptive boosting, NNs, stochastic gradient descent, and RF were added to the comparison, and RF performed best [70].

Of the 21 projects that reported results, 16 (76%) achieved this with real users [18,22,48,51,57,58,60-62,65,66,70,74,75,79,80], whereas the remaining 5 (24%) did so in simulations [24,56,63,64,71]. A total of 11% (3/27) of the projects did describe the design and a protocol for evaluation, but no results had been published yet [25,67,69]. In total, 7% (2/27) of the projects described the design, but no evaluation plan was found [19,21]. One project did evaluate the intervention, but no outcome regarding PA or the developed algorithm was discussed [72].

Other Findings

Only 30% (8/27) of the projects [19,24,51,57,60,67,69,75] reported the underlying theories that supported the choice of their implemented persuasive strategies. Most mentioned behavior change techniques as the basis for their intervention. Although not all studies reported specific behavior change techniques, the Fogg Behavior Model [81]; the Capability, Opportunity, and Motivation–Behavior (COM-B) model [82]; and self-determination theory [83,84] were mentioned several times. In addition, the social cognitive theory [85], the theory of social comparison by Buunk [86], and the theory of planned behavior [87] were included.

Most interventions (16/27, 59%) were designed as mobile phone apps. Only 11% (3/27) of the projects developed a web application. Except for 4% (1/27) of the projects, no human interaction, for instance, with a health care professional, was implemented. The project that did choose a hybrid version instead of a stand-alone digital intervention enabled contact with a human sports coach [56]. A total of 3 of the interventions included gamification elements, of which 2 (67%) developed an exergame with personalized difficulty of PA in the game. Regarding the input features for the ML algorithms, most were based on sensor data such as performed PA, heart rate, and location. In addition, phone log information, questionnaire results, and personal characteristics were used as input features.

Categorization Overview

In the included projects, we searched for differences and similarities in several characteristics: the dimension of personalization, implemented persuasive strategies, intervention target group, ML category and method, ML technique, and study outcome. Clear categories regarding the dimension of personalization and overlapping approaches regarding persuasive strategies and ML methods were found, whereas no categories could be established for the target group, study type, and study outcome. Regarding the target group, not enough differences were found to split the projects into categories. On the other hand, regarding the study outcome and ML technique, too many differences were found to propose categories at this level.

Thus, our categorization overview, shown in Figure 2 [17,18,20-23,45,48,51,56-58,61-68,71,72,75], consists of 3 levels: dimension of personalization, persuasive strategy, and ML method. For the first level, 4 dimensions of personalization were found in the included projects: the timing, content, and type of messages and the level of difficulty of PA. As only 7% (2/27) of the projects personalized the level of difficulty of PA [79,80] and both implemented another type of ML, there was not enough support to include this dimension of personalization in our framework. Therefore, the dimension of personalization level consists of timing, content, and type of messages.

For the timing of messages, the persuasive strategies were divided into 2 groups: the timing of reminders and the timing of messages regarding feedback, monitoring, and goal setting. RL was implemented in the 3 projects in this category that personalized the timing of reminders [24,48,51]. SL for classification was used in both projects [18,56] that personalized the timing of messages with feedback, monitoring, or goal-setting information. The difference in methods between the timing of reminders and that of feedback, monitoring, and goal-setting information might be explained by the idea that reminders often need to be sent at a specific time or time frame related to a lot of activities, whereas feedback, monitoring, and goal-setting messages are often sent at a time frame only related to the performed PA. The latter kind of messages should be sent, for instance, during, directly after, or 15 minutes after performing PA, which can be translated to 3 categories and, therefore, is suitable for solving using SL for classification. On the other hand, reminders for PA should be sent, for example, after 2 hours of inactivity on Tuesday mornings, 10 minutes after taking the bus on Thursday evenings, and also at 8:05 AM on Saturdays. RL is suitable to solve such specific timing problems. Moreover, with RL, the algorithm can be improved over time with additional information, such as the time between the reminder and the actual performed PA. This can be done without training the whole dataset again, which would have been the case when using SL.

XSL•FO
RenderX

Figure 2. Proposed categorization overview as a starting point for choosing machine learning methods when implementing adaptive persuasive strategies in mobile health interventions that promote physical activity [18,19,21,22,24,25,48,51,56-58,61-67,70-73,75]. RL: reinforcement learning; RS: recommender system; SL: supervised learning; UL: unsupervised learning.

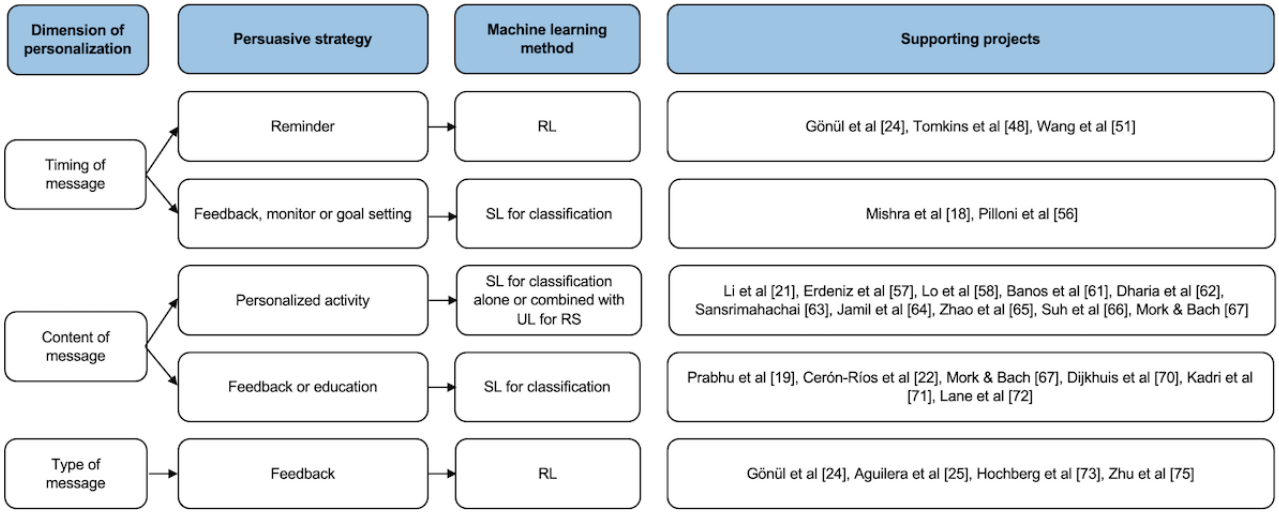

Regarding personalizing the content of the messages, 2 groups of persuasive strategies were found. For personalizing both PA suggestions and feedback or educational messages, SL for classification was used in most projects (15/17, 88%). To personalize PA suggestions, either SL or SL combined with UL for RS was implemented most often. This category is supported by 82% (9/11) of the projects in this group, which performed SL [57,58,61-67], of which 44% (4/9) combined SL for classification with UL for RS [57,61,62,65]. SL for classification can be used to predict which activities are the best fit to the participants' needs and characteristics, such as age, current PA level, or disease. When there is, for instance, a large number of PAs that can be suggested, SL can be combined with RS to improve the recommendations over time based on feedback from the user regarding the quality of the recommendation. The predictions that are performed using SL can then be used as a starting point. Regarding personalizing the content of feedback or educational messages, most projects in this category (6/8, 75%) only implemented SL [19,22,67,70-72]. For example, this can be used to predict which educational topics or feedback messages regarding the performed PA are suitable. The number of options for these types of messages is often much smaller than the number of possible PA suggestions. As RSs are particularly useful in selecting personalized items from a large dataset, improvement of the results using an RS is less suitable for personalizing feedback messages or educational topics than for personalizing PA suggestions. For this reason, and because only 17% (1/6) of the projects in this group combined SL with RS [22], we did not include this combination in our categorization.

Regarding the personalization of the type of messages, only feedback was found as a persuasive strategy. To personalize the type of feedback messages, all 4 projects implemented RL [24,25,74,75]. This, for instance, can be used to predict whether a participant's motivation is best improved using positive, neutral, or negative feedback. When RL is implemented, the algorithm can be improved over time without retraining or relabeling.

Discussion

Principal Findings

The objective of our study was to provide an overview of ML methods used to personalize persuasive strategies in PA-promoting interventions and present a categorization overview based on these findings as a starting point for implementing ML methods in this field. We analyzed the details of the developed PA interventions, the implemented personalized persuasive strategies, the ML methods used, and the results of 40 studies belonging to 27 projects. These papers could be categorized based on the dimensions of personalization, which resulted in 4 groups. For each dimension, 1 or 2 persuasive strategy categories were found and linked to ML methods.

On the basis of these included papers and dimensions found, we provided a categorization overview, now consisting of three layers: (1) the dimension of personalization, (2) the personalized persuasive strategy, and (3) the ML method. To personalize the timing of reminders, RL was mostly implemented, whereas SL for classification was implemented most often for personalizing the timing of messages regarding feedback, monitoring, and goal setting. For personalizing the content of PA suggestions and feedback or educational messages, most projects (15/17, 88%) implemented SL for classification as well. Regarding personalized PA suggestions, this was implemented either alone or combined with UL for RS. Finally, RL was often implemented to personalize the type of feedback messages. The categorization overview can be a starting point in using ML to personalize persuasive strategies in PA-promoting mHealth interventions.

Strengths, Limitations, and Opportunities

A strength of our study is the systematic analysis of multiple mHealth interventions using personalized persuasive strategies with ML to promote PA. We analyzed the persuasive strategies and ML methods of 27 projects on several levels. Inherent to conducting a scoping review, we did not consider the quality

XSL•FO
RenderX

of the studies, which could have led to biased results. Moreover, we initially intended to only include projects that had published results. Although all these projects showed positive results on their own outcome measures, we were not able to compare the results because the outcome measures differed widely. Therefore, we decided to also include projects that had not published results yet. This additional information was in line with the projects that had published results and, therefore, resulted in a stronger support for the presented categorization overview. However, it is a limitation that this categorization overview could only be based on the number of interventions in which ML methods were implemented and that effectiveness and performance differences were not considered. When more projects publish their results in the future, this could lead to enough comparable information regarding study outcomes so that recommendations based on the effectiveness of particular ML methods can be made. Moreover, challenges faced, such as data availability, interpretability, privacy concerns, and performance differences between data collection and labeling approaches, could then be discussed. Other aspects that might be elaborated more on in the future when more information is published are the underlying theories that support the choice of implemented persuasive strategies and the use of human contact to personalize persuasive strategies in health care settings. It is remarkable that, for now, only 30% (8/27) of the projects described such underlying theories and only 4% (1/27) of the projects implemented the possibility to have contact with a human being.

The need for more studies publishing performance or effectiveness to compare study outcomes was also discussed in the studies of Goh et al [16] and Chaudhari et al [37]. Goh et al [16] performed a scoping review on ML techniques to personalize interventions for health and well-being, including PA, and mentioned that more studies should examine interventions in a more mature, developmental stage to appraise the impacts of such interventions more prudently and confidently. Chaudhari et al [37] only studied personalized timing of PA-promoting messages and also included personalization without ML. However, they also discussed that more studies evaluating the effectiveness of the interventions are required to learn which aspects of personalization are promising.

Although we have carefully formulated our inclusion criteria and constructed our search strategy, such restrictions always result in excluding interesting papers and information. For future research, it might be interesting to explore how ML techniques have been used to personalize persuasive strategies specifically. Moreover, the search could be extended to more specific databases regarding health care and behavior change. To widen the perspective, it might also be compelling to compare ML methods and persuasive strategies in PA-promoting interventions with those in interventions that focus on other domains of health promotion.

Because we used ASReview, we could efficiently screen a lot of papers, and we were able to focus on the content rather than on irrelevant metadata. A strength of our study is the dataset with all labels and reasons for exclusion. This improves the reproducibility of our study, and this information is useful for future meta-reviews. However, a limitation is that the trained ASReview models could not be extracted from the ASReview program.

The proposed categorization overview now has 3 levels: the dimension of personalization, persuasive strategy, and type of ML. As the objective of our study was to provide a categorization overview as a starting point for implementing ML to personalize persuasive strategies in PA-promoting interventions, we chose to present the dimension of personalization as the first level. Adding a level on top of this to guide in choosing personalized persuasive strategies is outside the scope of our study. A lot of work has already been published in this field, for instance, studies that address the implementation of effective persuasive strategies [6-8] and the effect of personalized interventions [9-12]. At first, we wanted to add a fourth level to our categorization overview to further specify the ML methods by recommending specific ML techniques. However, not enough overlapping techniques were found in the included projects to support this information for all dimensions of personalization. The only ML technique that had enough support was collaborative filtering as UL method for an RS to personalize the content of PA suggestions. The number of choices per level was defined by the analyzed projects. We found 3 different dimensions of personalization with enough support and, for each dimension, 1 or 2 persuasive strategies with several supporting projects. Remarkably, these 3 dimensions concerned messages. This might indicate that messages are useful for personalization. A possible explanation might be that contact with the participant is the basis of a lot of persuasive strategies. In an mHealth intervention, such contact is often replaced by digital messages.

The current categorization overview might already be helpful for future designers, developers, and researchers that plan to implement ML for personalizing persuasive strategies in PA-promoting interventions. However, the categorization overview is expected to be extended when the results of more interventions are published. For instance, more choices per level with new dimensions of personalization, such as personalizing the level of difficulty of PA, or persuasive strategies could be added. Moreover, a fourth level might then be added to specify ML methods by categorizing ML techniques. It would be interesting to study whether specific ML techniques can be linked to features such as the kind of data used for personalization, the target group, the specific intervention goal, or the underlying psychological theories. In addition, it would be interesting to explore whether large language models will be used more often to personalize the content of messages now that generative artificial intelligence has become available to the public and has recently shown promising results.

Oyebode et al [38] conducted a review on ML techniques in personalized systems for health and well-being. Although they did not focus on persuasive strategies, they did discuss PA-promoting interventions specifically. Our results are in line with theirs, although we included 40 papers belonging to 27 projects, whereas they included 8 papers regarding PA. In addition, they gave some recommendations. These regarded general implementation challenges, such as data quality– and infrastructure-related issues. When designing new personalized

XSL•FO
RenderX

persuasive PA-promoting interventions, it might be helpful to combine our categorization overview with their recommendations. This leads us to another important strength of our study: the provided categorization overview. We did not only show an insightful overview of ML methods used but also categorized this information as a starting point for the design and development of ML-personalized PA interventions.

Conclusions

Our categorization overview framework links 3 dimensions of personalization and several persuasive strategies to implemented ML methods. Regarding the timing of messages, RL was implemented most often to personalize the timing of reminders, and SL was implemented to personalize the timing of feedback, monitoring, and goal-setting messages. When personalizing the content of messages, most projects implemented either SL or both SL and RS for PA suggestions and SL for feedback or educational messages. When personalizing the type of feedback messages, most projects implemented RL algorithms. The provided categorization overview can be used as a starting point in the design and development of personalized persuasive strategies to promote PA. When more of such mHealth intervention results are published in the future, the categorization overview might be expanded with specific ML techniques or with additional dimensions of personalization and persuasive strategies, such as personalizing the level of difficulty of PA. Moreover, when more results regarding performance or effectiveness are published, recommendations might be given on which ML methods to implement.

Acknowledgments

This work was supported by the Institute for Preventive Health and the artificial intelligence program of the EWUU alliance of Eindhoven University of Technology, Wageningen University & Research, Utrecht University, and University Medical Center Utrecht.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 91 KB - jmir_v26i1e47774_app1.pdf]

Multimedia Appendix 2
Search query.
[PDF File (Adobe PDF File), 177 KB - jmir_v26i1e47774_app2.pdf]

References

1. WHO factsheet physical activity. World Health Organization (WHO). URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2022-10-24]
2. Müller AM, Alley S, Schoeppe S, Vandelanotte C. The effectiveness of e- and mHealth interventions to promote physical activity & healthy diets in developing countries: a systematic review. Int J Behav Nutr Phys Act 2016 Oct 10;13(1):109 [FREE Full text] [doi: 10.1186/s12966-016-0434-2] [Medline: 27724911]
3. Gal R, May AM, van Overmeeren EJ, Simons M, Monninkhof EM. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. Sports Med Open 2018 Sep 03;4(1):42 [FREE Full text] [doi: 10.1186/s40798-018-0157-9] [Medline: 30178072]
4. Direito A, Carraça E, Rawstorn J, Whittaker R, Maddison R. mHealth technologies to influence physical activity and sedentary behaviors: behavior change techniques, systematic review and meta-analysis of randomized controlled trials. Ann Behav Med 2017 Apr 18;51(2):226-239. [doi: 10.1007/s12160-016-9846-0] [Medline: 27757789]
5. Fiedler J, Eckert T, Wunsch K, Woll A. Key facets to build up eHealth and mHealth interventions to enhance physical activity, sedentary behavior and nutrition in healthy subjects - an umbrella review. BMC Public Health 2020 Oct 23;20(1):1605 [FREE Full text] [doi: 10.1186/s12889-020-09700-7] [Medline: 33097013]
6. Eckerstorfer LV, Tanzer NK, Vogrincic-Haselbacher C, Kedia G, Brohmer H, Dinslaken I, et al. Key elements of mHealth interventions to successfully increase physical activity: meta-regression. JMIR Mhealth Uhealth 2018 Nov 12;6(11):e10076 [FREE Full text] [doi: 10.2196/10076] [Medline: 30425028]
7. Domin A, Spruijt-Metz D, Theisen D, Ouzzahra Y, Vögele C. Smartphone-based interventions for physical activity promotion: scoping review of the evidence over the last 10 years. JMIR Mhealth Uhealth 2021 Jul 21;9(7):e24308 [FREE Full text] [doi: 10.2196/24308] [Medline: 34287209]
8. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009;28(6):690-701. [doi: 10.1037/A0016136]

XSL•FO
RenderX

9. Davis A, Sweigart R, Ellis R. A systematic review of tailored mHealth interventions for physical activity promotion among adults. Transl Behav Med 2020 Oct 12;10(5):1221-1232. [doi: 10.1093/tbm/ibz190] [Medline: 33044542]
10. Baretta D, Perski O, Steca P. Exploring users' experiences of the uptake and adoption of physical activity apps: longitudinal qualitative study. JMIR Mhealth Uhealth 2019 Feb 08;7(2):e11636 [FREE Full text] [doi: 10.2196/11636] [Medline: 30735143]
11. Sporrel K, Nibbeling N, Wang S, Ettema D, Simons M. Unraveling mobile health exercise interventions for adults: scoping review on the implementations and designs of persuasive strategies. JMIR Mhealth Uhealth 2021 Jan 18;9(1):e16282 [FREE Full text] [doi: 10.2196/16282] [Medline: 33459598]
12. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010;51(3-4):214-221 [FREE Full text] [doi: 10.1016/j.ypmed.2010.06.004] [Medline: 20558196]
13. Kreuter MW, Skinner CS. Tailoring: what's in a name? Health Educ Res 2000 Feb;15(1):1-4 [FREE Full text] [Medline: 10788196]
14. Beck C, McSweeney JC, Richards KC, Roberson PK, Tsai P, Souder E. Challenges in tailored intervention research. Nurs Outlook 2010;58(2):104-110 [FREE Full text] [doi: 10.1016/j.outlook.2009.10.004] [Medline: 20362779]
15. Gosetto L, Ehrler F, Falquet G. Personalization dimensions for mHealth to improve behavior change: a scoping review. Stud Health Technol Inform 2020 Nov 23;275:77-81. [doi: 10.3233/SHTI200698] [Medline: 33227744]
16. Goh YS, Ow Yong JQ, Chee BQ, Kuek JH, Ho CS. Machine learning in health promotion and behavioral change: scoping review. J Med Internet Res 2022 Jun 02;24(6):e35831 [FREE Full text] [doi: 10.2196/35831] [Medline: 35653177]
17. Craig KJ, Morgan LC, Chen C, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021 May 25;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
18. Mishra V, Künzler F, Kramer JN, Fleisch E, Kowatsch T, Kotz D. Detecting receptivity for mHealth interventions in the natural environment. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Jun;5(2):74 [FREE Full text] [doi: 10.1145/3463492] [Medline: 34926979]
19. Prabhu G, Kuklyte J, Gualano L, Venkataraman K, Ahmadi A, Duff O, et al. Design and development of the MedFit app: a mobile application for cardiovascular disease rehabilitation. In: Proceedings of the 7th International Conference on Wireless Mobile Communication and Healthcare. 2017 Presented at: MobiHealth '17; November 14-15, 2017; Vienna, Austria p. 20-28 URL: https://link.springer.com/chapter/10.1007/978-3-319-98551-0_3 [doi: 10.1007/978-3-319-98551-0_3]
20. Elhadd T, Mall R, Bashir M, Palotti J, Fernandez-Luque L, Farooq F, for PROFAST-Ramadan Study Group. Artificial Intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during Ramadan (the PROFAST - IT ramadan study). Diabetes Res Clin Pract 2020 Nov;169:108388 [FREE Full text] [doi: 10.1016/j.diabres.2020.108388] [Medline: 32858096]
21. Li Z, Das S, Codella J, Hao T, Lin K, Maduri C, et al. An adaptive, data-driven personalized advisor for increasing physical activity. IEEE J Biomed Health Inform 2019 May;23(3):999-1010. [doi: 10.1109/JBHI.2018.2879805] [Medline: 30418890]
22. Cerón-Rios G, López DM, Blobel B. Architecture and user-context models of CoCare: a context-aware mobile recommender system for health promotion. Stud Health Technol Inform 2017;237:140-147. [Medline: 28479557]
23. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 03;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
24. Gönül S, Namlı T, Coşar A, Toroslu İ. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. Artif Intell Med 2021 May;115:102062. [doi: 10.1016/j.artmed.2021.102062] [Medline: 34001322]
25. Aguilera A, Figueroa CA, Hernandez-Ramos R, Sarkar U, Cemballi A, Gomez-Pathak L, et al. mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study. BMJ Open 2020 Aug 20;10(8):e034723 [FREE Full text] [doi: 10.1136/bmjopen-2019-034723] [Medline: 32819981]
26. Carter DD, Robinson K, Forbes J, Hayes S. Experiences of mobile health in promoting physical activity: a qualitative systematic review and meta-ethnography. PLoS One 2018;13(12):e0208759 [FREE Full text] [doi: 10.1371/journal.pone.0208759] [Medline: 30557396]
27. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699 [FREE Full text] [doi: 10.2196/26699] [Medline: 33811021]
28. Lee AM, Chavez S, Bian J, Thompson LA, Gurka MJ, Williamson VG, et al. Efficacy and effectiveness of mobile health technologies for facilitating physical activity in adolescents: scoping review. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e11847 [FREE Full text] [doi: 10.2196/11847] [Medline: 30747716]
29. Böhm B, Karwiese SD, Böhm H, Oberhoffer R. Effects of mobile health including wearable activity trackers to increase physical activity outcomes among healthy children and adolescents: systematic review. JMIR Mhealth Uhealth 2019 Apr 30;7(4):e8298 [FREE Full text] [doi: 10.2196/mhealth.8298] [Medline: 31038460]
30. McGarrigle L, Todd C. Promotion of physical activity in older people using mHealth and eHealth technologies: rapid review of reviews. J Med Internet Res 2020 Dec 29;22(12):e22201 [FREE Full text] [doi: 10.2196/22201] [Medline: 33372894]

XSL•FO
RenderX

31. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
32. Monteiro-Guerra F, Rivera-Romero O, Fernandez-Luque L, Caulfield B. Personalization in real-time physical activity coaching using mobile applications: a scoping review. IEEE J Biomed Health Inform 2020 Jun;24(6):1738-1751. [doi: 10.1109/jbhi.2019.2947243]
33. Ma JK, Floegel TA, Li LC, Leese J, De Vera MA, Beauchamp MR, et al. Tailored physical activity behavior change interventions: challenges and opportunities. Transl Behav Med 2021 Dec 14;11(12):2174-2181 [FREE Full text] [doi: 10.1093/tbm/ibab106] [Medline: 34424344]
34. Ghanvatkar S, Kankanhalli A, Rajan V. User models for personalized physical activity interventions: scoping review. JMIR Mhealth Uhealth 2019 Jan 16;7(1):e11098 [FREE Full text] [doi: 10.2196/11098] [Medline: 30664474]
35. Aldenaini N, Orji R, Sampalli S. How effective is personalization in persuasive interventions for reducing sedentary behavior and promoting physical activity: a systematic review. In: Proceedings of the 15th International Conference on Persuasive Technology. 2020 Presented at: Persuasive '20; January 28-29, 2020; Aalborg, Denmark p. 15 URL: https://tinyurl.com/yc66wbju [doi: 10.3389/fcomp.2020.00019]
36. Daryabeygi-Khotbehsara R, Shariful Islam SM, Dunstan D, McVicar J, Abdelrazek M, Maddison R. Smartphone-based interventions to reduce sedentary behavior and promote physical activity using integrated dynamic models: systematic review. J Med Internet Res 2021 Sep 13;23(9):e26315 [FREE Full text] [doi: 10.2196/26315] [Medline: 34515637]
37. Chaudhari S, Ghanvatkar S, Kankanhalli A. Personalization of intervention timing for physical activity: scoping review. JMIR Mhealth Uhealth 2022 Feb 28;10(2):e31327 [FREE Full text] [doi: 10.2196/31327] [Medline: 35225811]
38. Oyebode O, Fowles J, Steeves D, Orji R. Machine learning techniques in adaptive and personalized systems for health and wellness. Int J Hum Comput Interact 2022 Jul 27;39(9):1938-1962. [doi: 10.1080/10447318.2022.2089085]
39. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
40. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
41. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
42. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
43. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021 Feb 01;3(2):125-133. [doi: 10.1038/S42256-020-00287-7]
44. van de Schoot R, de Bruin J. Researcher-in-the-loop for systematic reviewing of text databases. Zenodo. 2020. URL: https://zenodo.org/records/4013207 [accessed 2024-04-29]
45. Ferdinands G, Schram R, de Bruin J, Bagheri A, Oberski D, Tummers L, et al. Active learning for screening prioritization in systematic reviews: a simulation study. OSF Preprint posted online August 10, 2020 [FREE Full text] [doi: 10.31219/osf.io/w6qbg]
46. van de Schoot R, de Bruin J, Schram R. ASReview: active learning for systematic reviews (v0.13.2). Zenodo. URL: https://zenodo.org/records/4158671 [accessed 2024-04-29]
47. sw1989 / ML-methods-for-mHealth. GitHub. URL: https://github.com/sw1989/ML-methods-for-mHealth/tree/main [accessed 2024-04-29]
48. Tomkins S, Liao P, Klasnja P, Murphy S. IntelligentPooling: practical thompson sampling for mHealth. Mach Learn 2021 Sep 21;110(9):2685-2727 [FREE Full text] [doi: 10.1007/s10994-021-05995-8] [Medline: 34621105]
49. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Mar 18;4(1):1-22 [FREE Full text] [doi: 10.1145/3381007] [Medline: 34527853]
50. Glachs D, Namli T, Strohmeier F, Rodríguez Suárez G, Sluis M, Delgado-Lista J, et al. A predictive model-based decision support system for diabetes patient empowerment. Stud Health Technol Inform 2021 May 27;281:963-968. [doi: 10.3233/SHTI210321] [Medline: 34042816]
51. Wang S, Sporrel K, van Hoof H, Simons M, de Boer RD, Ettema D, et al. Reinforcement learning to send reminders at right moments in smartphone exercise application: a feasibility study. Int J Environ Res Public Health 2021 Jun 04;18(11):23 [FREE Full text] [doi: 10.3390/ijerph18116059] [Medline: 34199880]
52. Wang S, Zhang C, Kröse B, van Hoof H. Optimizing adaptive notifications in mobile health interventions systems: reinforcement learning from a data-driven behavioral simulator. J Med Syst 2021 Oct 18;45(12):102 [FREE Full text] [doi: 10.1007/s10916-021-01773-0] [Medline: 34664120]

XSL•FO
RenderX

53. Sporrel K, De Boer RD, Wang S, Nibbeling N, Simons M, Deutekom M, et al. The design and development of a personalized leisure time physical activity application based on behavior change theories, end-user perceptions, and principles from empirical data mining. Front Public Health 2020;8:528472 [FREE Full text] [doi: 10.3389/fpubh.2020.528472] [Medline: 33604321]
54. Kramer JN, Künzler F, Mishra V, Presset B, Kotz D, Smith S, et al. Investigating intervention components and exploring states of receptivity for a smartphone app to promote physical activity: protocol of a microrandomized trial. JMIR Res Protoc 2019 Jan 31;8(1):e11540 [FREE Full text] [doi: 10.2196/11540] [Medline: 30702430]
55. Künzler F, Mishra V, Kramer JN, Kotz D, Fleisch E, Kowatsch T. Exploring the state-of-receptivity for mHealth interventions. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Dec 11;3(4):1-27 [FREE Full text] [doi: 10.1145/3369805] [Medline: 36159209]
56. Pilloni P, Piras L, Boratto L, Carta S, Fenu G, Mulas F. Recommendation in persuasive eHealth systems: an effective strategy to spot users' losing motivation to exercise. In: Proceedings of the 11th ACM Conference on Recommender Systems. 2017 Presented at: RecSys '17; August 27-31, 2017; Como, Italy p. 6-9 URL: https://iris.unica.it/bitstream/11584/243730/1/healthRecSys17_paper_5.pdf
57. Erdeniz SP, Menychtas A, Maglogiannis I, Felfernig A, Tran TN. Recommender systems for IoT enabled quantified-self applications. Evol Syst 2019 Oct 30;11(2):291-304. [doi: 10.1007/S12530-019-09302-8]
58. Lo WL, Lei D, Li L, Huang DF, Tong KF. The perceived benefits of an artificial intelligence-embedded mobile app implementing evidence-based guidelines for the self-management of chronic neck and back pain: observational study. JMIR Mhealth Uhealth 2018 Nov 26;6(11):e198 [FREE Full text] [doi: 10.2196/mhealth.8127] [Medline: 30478019]
59. Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 707-718 URL: https://dl.acm.org/doi/10.1145/2750858.2805840 [doi: 10.1145/2750858.2805840]
60. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth 2015;3(2):e42 [FREE Full text] [doi: 10.2196/mhealth.4160] [Medline: 25977197]
61. Banos O, Bilal Amin M, Ali Khan W, Afzal M, Hussain M, Kang BH, et al. The Mining Minds digital health and wellness framework. Biomed Eng Online 2016 Jul 15;15 Suppl 1:76 [FREE Full text] [doi: 10.1186/s12938-016-0179-9] [Medline: 27454608]
62. Dharia S, Eirinaki M, Jain V, Patel J, Varlamis I, Vora J, et al. Social recommendations for personalized fitness assistance. Pers Ubiquit Comput 2017 Jun 29;22(2):245-257. [doi: 10.1007/S00779-017-1039-8]
63. Sansrimahachai W. Personalized walking exercise support system for elderly based on machine learning. In: Proceedings of the 17th International Joint Conference on Computer Science and Software Engineering. 2020 Presented at: JCSSE '20; November 4-6, 2020; Bangkok, Thailand p. 6-11 URL: https://ieeexplore.ieee.org/document/9268327 [doi: 10.1109/jcsse49651.2020.9268327]
64. Jamil F, Kahng HK, Kim S, Kim DH. Towards secure fitness framework based on IoT-enabled blockchain network integrated with machine learning algorithms. Sensors (Basel) 2021 Feb 26;21(5):1-31 [FREE Full text] [doi: 10.3390/s21051640] [Medline: 33652773]
65. Zhao Z, Arya A, Orji R, Chan G. Physical activity recommendation for exergame player modeling using machine learning approach. In: Proceedings of the 8th International Conference on Serious Games and Applications for Health. 2020 Presented at: SeGAH '20; August 12-14, 2020; Vancouver, BC p. 1-9 URL: https://ieeexplore.ieee.org/document/9201820 [doi: 10.1109/segah49190.2020.9201820]
66. Suh MK, Nahapetian A, Woodbridge J, Rofouei M, Sarrafzadeh M. Machine learning-based adaptive wireless interval training guidance system. Mobile Netw Appl 2011 Jun 30;17(2):163-177. [doi: 10.1007/S11036-011-0331-5]
67. Mork PJ, Bach K, selfBACK Consortium. A decision support system to enhance self-management of low back pain: protocol for the selfBACK project. JMIR Res Protoc 2018 Dec 20;7(7):e167 [FREE Full text] [doi: 10.2196/resprot.9379] [Medline: 30030208]
68. Sandal LF, Stochkendahl MJ, Svendsen MJ, Wood K, Øverås CK, Nordstoga AL, et al. An app-delivered self-management program for people with low back pain: protocol for the selfBACK randomized controlled trial. JMIR Res Protoc 2019 Dec 03;8(12):e14720 [FREE Full text] [doi: 10.2196/14720] [Medline: 31793897]
69. Pelle T, Bevers K, van der Palen J, van den Hoogen FH, van den Ende CH. Development and evaluation of a tailored e-self-management intervention (dr. Bart app) for knee and/or hip osteoarthritis: study protocol. BMC Musculoskelet Disord 2019 Aug 31;20(1):398 [FREE Full text] [doi: 10.1186/s12891-019-2768-9] [Medline: 31472687]
70. Dijkhuis TB, Blaauw FJ, van Ittersum MW, Velthuijsen H, Aiello M. Personalized physical activity coaching: a machine learning approach. Sensors (Basel) 2018 Feb 19;18(2):1 [FREE Full text] [doi: 10.3390/s18020623] [Medline: 29463052]
71. Kadri N, Ellouze A, Ksantini M. Recommendation system for human physical activities using smartphones. In: Proceedings of the 2nd International Conference on Computer and Information Sciences. 2020 Presented at: ICCIS '20; October 13-15, 2020; Sakaka, Saudi Arabia p. 1-4 URL: https://ieeexplore.ieee.org/document/9257671 [doi: 10.1109/iccis49240.2020.9257671]

XSL•FO
RenderX

72. Lane ND, Lin M, Mohammod M, Yang X, Lu H, Cardone G, et al. BeWell: sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Netw Appl 2014 Jan 9;19(3):345-359. [doi: 10.1007/S11036-013-0484-5]
73. Hochberg I, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Yom-Tov E. Encouraging physical activity in patients with diabetes through automatic personalized feedback via reinforcement learning improves glycemic control. Diabetes Care 2016 Apr;39(4):e59-e60. [doi: 10.2337/dc15-2340] [Medline: 26822328]
74. Yom-Tov E, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Hochberg I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. J Med Internet Res 2017 Oct 10;19(10):e338 [FREE Full text] [doi: 10.2196/jmir.7994] [Medline: 29017988]
75. Zhu J, Dallal DH, Gray RC, Villareale J, Ontañón S, Forman EM, et al. Personalization paradox in behavior change apps: lessons from a social comparison-based personalized app for physical activity. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-21. [doi: 10.1145/3449190]
76. Figueroa CA, Luo TC, Jacobo A, Munoz A, Manuel M, Chan D, et al. Conversational physical activity coaches for Spanish and English speaking women: a user design study. Front Digit Health 2021 Oct 8;3:747153 [FREE Full text] [doi: 10.3389/fdgth.2021.747153] [Medline: 34713207]
77. Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal J, et al. Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. Ann Behav Med 2022 Mar 11;56(2):212-218. [doi: 10.1093/abm/kaab028] [Medline: 33871015]
78. Figueroa CA, Aguilera A, Chakraborty B, Modiri A, Aggarwal J, Deliu N, et al. Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. J Am Med Inform Assoc 2021 Jun 12;28(6):1225-1234 [FREE Full text] [doi: 10.1093/jamia/ocab001] [Medline: 33657217]
79. Huber T, Mertes S, Rangelova S, Flutura S, André E. Dynamic difficulty adjustment in virtual reality exergames through experience-driven procedural content generation. In: Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence. 2021 Presented at: SSCI '21; December 5-7, 2021; Orlando, FL p. 1-8 URL: https://ieeexplore.ieee.org/document/9660086 [doi: 10.1109/SSCI50451.2021.9660086]
80. Aguilar JD, Guzman DE, Rengifo CF, Chalapud LM, Guzman JD. Proposal of a game with dynamic difficulty adjustment from physiological signals in the context of an exergame. In: Proceedings of the 40th Central America and Panama Convention. 2022 Presented at: CONCAPAN '22; November 9-12,2021; Panama City, Panama p. 1-6 URL: https://ieeexplore.ieee.org/document/9997775 [doi: 10.1109/CONCAPAN48024.2022.9997775]
81. Fogg BJ. Persuasive Technology: Using Computers to Change What We Think and Do. San Francisco, CA: Morgan Kaufmann Publishers; 2010.
82. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
83. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]
84. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 2000;55(1):68-78. [doi: 10.1037/0003-066X.55.1.68]
85. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001 Feb;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
86. Buunk BP. Comparison direction and comparison dimension among disabled individuals: toward a refined conceptualization of social comparison under stress. Pers Soc Psychol Bull 1995 Apr 01;21(4):316-330. [doi: 10.1177/0146167295214002]
87. Ajzen I. From intentions to actions: a theory of planned behavior. In: Khul J, Beckmann J, editors. Action Control: From Cognition to Behavior. Berlin, Germany: Springer; 1985:11-39.

Abbreviations

BiLSTM: bidirectional long short-term memory
COM-B: Capability, Opportunity, and Motivation–Behavior
DT: decision tree
EXP3: Exponential-weight algorithm for Exploration and Exploitation
MAB: multi-armed bandit
MDP: Markov decision process
mHealth: mobile health
ML: machine learning
NN: neural network
PA: physical activity
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RF: random forest
RL: reinforcement learning

XSL•FO
RenderX

RS: recommender system
SL: supervised learning
UL: unsupervised learning

Edited by A Mavragani; submitted 05.04.23; peer-reviewed by P Gorp, M Simons, M Swartz, R Safa; comments to author 17.09.23; revised version received 07.01.24; accepted 23.07.24; published 15.11.24.

Please cite as:
Brons A, Wang S, Visser B, Kröse B, Bakkes S, Veltkamp R
Machine Learning Methods to Personalize Persuasive Strategies in mHealth Interventions That Promote Physical Activity: Scoping Review and Categorization Overview
J Med Internet Res 2024;26:e47774
URL: https://www.jmir.org/2024/1/e47774
doi:10.2196/47774
PMID:

©Annette Brons, Shihan Wang, Bart Visser, Ben Kröse, Sander Bakkes, Remco Veltkamp. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review

Phawinpon Chotwanvirat[1,2], PhD; Aree Prachansuwan[3], PhD; Pimnapanut Sridonpai[3], MSc; Wantanee Kriengsinyos[3], PhD

[1]Theptarin Diabetes, Thyroid, and Endocrine Center, Vimut-Theptarin Hospital, Bangkok, Thailand

[2]Diabetes and Metabolic Care Center, Taksin Hospital, Medical Service Department, Bangkok Metropolitan Administration, Bangkok, Thailand

[3]Human Nutrition Unit, Food and Nutrition Academic and Research Cluster, Institute of Nutrition, Mahidol University, Nakhon Pathom, Thailand

Corresponding Author:
Wantanee Kriengsinyos, PhD
Human Nutrition Unit, Food and Nutrition Academic and Research Cluster, Institute of Nutrition
Mahidol University
999 Phutthamonthon 4 Rd., Salaya
Nakhon Pathom, 73170
Thailand
Phone: 66 2 800 2380
Fax: 66 2 441 9344
Email: wantanee.krieng@mahidol.ac.th

Abstract

Background: To accurately capture an individual's food intake, dietitians are often required to ask clients about their food frequencies and portions, and they have to rely on the client's memory, which can be burdensome. While taking food photos alongside food records can alleviate user burden and reduce errors in self-reporting, this method still requires trained staff to translate food photos into dietary intake data. Image-assisted dietary assessment (IADA) is an innovative approach that uses computer algorithms to mimic human performance in estimating dietary information from food images. This field has seen continuous improvement through advancements in computer science, particularly in artificial intelligence (AI). However, the technical nature of this field can make it challenging for those without a technical background to understand it completely.

Objective: This review aims to fill the gap by providing a current overview of AI's integration into dietary assessment using food images. The content is organized chronologically and presented in an accessible manner for those unfamiliar with AI terminology. In addition, we discuss the systems' strengths and weaknesses and propose enhancements to improve IADA's accuracy and adoption in the nutrition community.

Methods: This scoping review used PubMed and Google Scholar databases to identify relevant studies. The review focused on computational techniques used in IADA, specifically AI models, devices, and sensors, or digital methods for food recognition and food volume estimation published between 2008 and 2021.

Results: A total of 522 articles were initially identified. On the basis of a rigorous selection process, 84 (16.1%) articles were ultimately included in this review. The selected articles reveal that early systems, developed before 2015, relied on handcrafted machine learning algorithms to manage traditional sequential processes, such as segmentation, food identification, portion estimation, and nutrient calculations. Since 2015, these handcrafted algorithms have been largely replaced by deep learning algorithms for handling the same tasks. More recently, the traditional sequential process has been superseded by advanced algorithms, including multitask convolutional neural networks and generative adversarial networks. Most of the systems were validated for macronutrient and energy estimation, while only a few were capable of estimating micronutrients, such as sodium. Notably, significant advancements have been made in the field of IADA, with efforts focused on replicating humanlike performance.

Conclusions: This review highlights the progress made by IADA, particularly in the areas of food identification and portion estimation. Advancements in AI techniques have shown great potential to improve the accuracy and efficiency of this field. However, it is crucial to involve dietitians and nutritionists in the development of these systems to ensure they meet the requirements and trust of professionals in the field.

(J Med Internet Res 2024;26:e51432) doi:10.2196/51432

XSL•FO
RenderX

KEYWORDS

image-assisted dietary assessment; artificial intelligence; dietary assessment; mobile phone; food intake; image recognition; portion size

Introduction

Background

Dietary assessment is a technique for determining an individual's intake, eating patterns, and food quality choices, as well as the nutritional values of consumed food. However, this technique's procedures are costly, laborious, and time-consuming and rely on specially trained personnel (such as dietitians and nutritionists) to produce reliable results. Consequently, a strong need exists for novel methods having improved measurement capabilities that are accurate, convenient, less burdensome, and cost-effective [1]. Instead of relying solely on client self-report, taking food photos before eating has been incorporated into traditional methods, such as a 3-day food record with food images, to reduce missing food records, incorrect food identification, and errors in portion size estimation. However, this technique still requires well-trained staff to translate food image information into reliable nutritional values and does not solve labor-intensive and time-consuming issues.

The application of computer algorithms to translate food images into representative nutritional values has gained interest in both the nutrition and computer science communities. This combination has resulted in a new field called image-assisted dietary assessment (IADA), and various systems have been developed to address these limitations, ranging from simple estimation equations in early systems to more complex artificial intelligence (AI) models in recent years. By applying IADA alongside the increasing use of smartphones and devices with built-in digital cameras, real-time analysis of dietary intake data from food images has become possible with accurate results, reduced labor, and greater convenience, thus gaining attention among nutrition professionals. However, the technical nature of this field can make it difficult to understand for those without a background in computer science or engineering, leading to the low involvement of nutrition professionals in its development. This gap is the rationale for us to conduct this review.

Objectives

The objective of this review is to bridge that knowledge gap by providing an up-to-date overview of the gradual enhancement of AI integration in dietary assessment based on food images. The information is presented in chronological order and in a manner that is understandable and accessible to those who may not be familiar with the technical jargon and complexity of AI terminologies. In addition, the advantages and limitations of these systems are discussed. Finally, we proposed auxiliary systems to enhance the accuracy of IADA and its potential adoption within the nutrition community.

Methods

Overview

To conduct this scoping review, we followed the methodology suggested by Arksey and O'Malley [2] and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [3].

Search Strategy

We searched 2 web-based databases, PubMed and Google Scholar, between February 2023 and March 2023, using the following terms: ((“food image”[Title/Abstract]) AND (classification[Title/Abstract] OR recognition[Title/Abstract] OR (“computer vision”[Title/Abstract]))) and “artificial intelligence,” “dietary assessment,” “computer vision,” “food image” recognition, “portion size,” segmentation, and classification, respectively.

Eligibility Criteria

This review included studies that focused on AI techniques used for IADA, specifically AI models, systems, or digital methods for food recognition and food volume estimation. For mobile apps or systems, we considered only articles that explain algorithms beyond mobile apps, prototype testing, or conducting clinical research. Studies that used noncomputational techniques, such as using food images as a tool for training human portion estimation, are excluded. Eligible articles were published in peer-reviewed journals or conference papers and written in English.

Selection Process

We used Zotero (Corporation for Digital Scholarship) reference management software to collect search results using the add multiple results function. All automatic data retrieval functions were disabled to prevent data retrieval from exceeding Google Scholar's traffic limitation. Zotero's built-in duplicate merger was used to identify duplicated records, and unduplicated records were exported to Excel online (Microsoft Corp). In Excel, all authors independently screened article types, titles, and abstracts. The screening process removed all nonrelated titles or abstracts, review and editorial articles, non-English articles, or conference abstracts without full text. For thesis articles, the corresponding published articles were identified using keywords from the title, first author, or corresponding author whenever possible. Each article required 2 independent reviewers' approval. In cases of conflict, a full-text review was necessary to resolve disagreements. After the initial screening process, the full texts of articles were obtained to assess eligibility. All full-text articles, whether they were excluded or not, and review articles were thoroughly read to identify interesting or related articles. These were classified as articles from other sources.

XSL•FO
RenderX

Data Extraction

A data extraction table was constructed, including the system name, classification algorithm, portion size estimation algorithm, accuracy of classification or portion estimated results, and the system's noticeable advantages and drawbacks. Data were extracted from full texts.

Results

Literature Findings

We retrieved 44 (8.4%) items from PubMed, while Google Scholar provided 478 (91.6%) results from the search terms, giving a total of 522 items retrieved. In total, 122 (23.4%) duplicate items were removed using Zotero's built-in duplicate merger. The remaining 400 (76.6%) deduplicated items were screened based on their titles and abstracts, resulting in 104 (19.9%) records for full-text review. After the full-text review process, 72 (13.8%) articles were included in this study. In addition, we manually identified and included 12 (2.3%) additional articles from other sources. An overview of the literature identification method and results is shown in Figure 1, and the PRISMA-ScR checklist is available in Multimedia Appendix 1.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flowchart of the structured literature search, screening, and selection methodology.

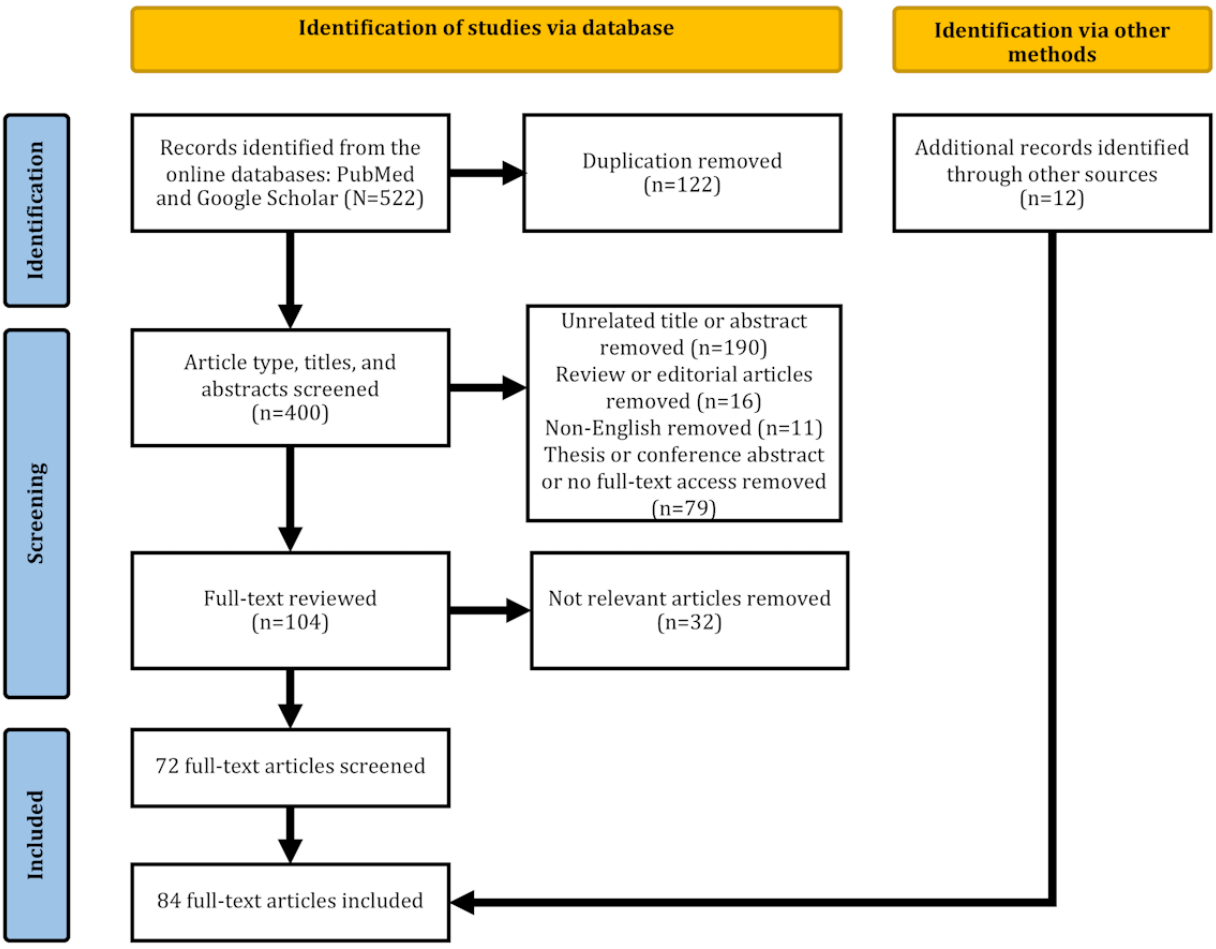

Traditional Dietary Assessment Methods

When measuring individual food intake, dietary assessment methods are typically divided into 2 sequential processes: methods to obtain dietary intake and methods to estimate the nutritional values of food. Principally, obtaining an individual's intake can be done by recording all consumed foods, beverages, herbs, or supplements with their portion sizes on a day-to-day basis or within a specific time frame (eg, a week) based on variation in the nutrients of interest. These methods were developed early on and can be performed manually. Due to their simplicity, some methods are frequently used in nutrition professionals' practices.

The 24-hour dietary recall (24HR) method is the simplest way to measure dietary intake, but accurately obtaining dietary intake information can be very challenging. The participant or their caregiver are asked by a trained interviewer to recall the participant's food intake within the last 24 hours. This method relies heavily on the client's memory and estimation of food portion size [4]. Unintentional misreporting of food intake is common, as clients often forget some foods. Underreporting of portion size is common because clients are not familiar with

XSL•FO
RenderX

estimating food portion sizes [5,6]. In participants who are overweight or obese, intentional underreporting is also common [7]. Although this method is the simplest for determining dietary intake, it takes approximately 1 hour to complete each interview. Moreover, a single 24HR result does not satisfactorily define an individual's usual intake due to day-to-day variations in eating habits.

Estimated food records (EFRs) are more reliable but time-consuming. Clients are asked to record all food and beverage intake during eating times for a specified period. Details of food are needed along with the portion sizes estimated by the client and rounded to household units (eg, half cup of soymilk with ground sesame and 4 tablespoons of kidney beans without syrup). To improve accuracy, training in estimating portion size using standard food models is required. The EFR places a burden on the clients, as they need to record all eating times. Moreover, some clients temporarily change their intake habits during recording to minimize this burden, while others may intentionally not report certain foods to cover up certain eating habits. Food portion size estimation errors are sometimes found, but taking food photographs before and after eating can lower these errors [8-12].

A standardized weighing scale can be used to avoid errors caused by human estimation of portion sizes. This technique is known as weighed food records and is considered the gold standard for determining personal intake. However, it is impractical to weigh all eaten food in the long term because it becomes a burden for the client to measure the weight of food eaten throughout the day [4]. This technique also only eliminates portion size estimation errors, while other issues with EFRs may still persist.

After retrieving dietary intake information from sources, such as 24HR, EFR, or weighed food records, the next step is to estimate the representative nutritional value of the food using a food composition table. If the recorded foods match the food items and their description in an available food composition table, the nutritional values can be obtained by multiplying the consumed food weight directly. However, if the food items are not found, the food needs to be analyzed and broken down into its components. The nutritional values of each component can then be obtained from the food composition table (or its nutrition label) and multiplied by the actual weight of each consumed component. When the portion size is recorded instead of its actual weight, the estimated weight can be obtained using standardized portion sizes from the food composition table. Nutrient analysis software can easily accomplish this task.

IADA Methods

Overview

Digital devices are often used for dietary assessment. The first well-documented attempt to develop such a digital device was called Wellnavi by Wang et al [8]. Although the device yielded accurate results, its usability was limited by the technologies of the time, including short battery life, poor image quality, a bulky body, and a less sensitive touch screen [10].

Several attempts have been made to use generic devices, such as Palm (Palm Inc) PDAs [13], compact digital cameras [14], and smartphones [15], instead of inventing a specific food recording device. In using these devices, users reported a decrease in the burden of completing food recording when compared with traditional methods [16,17]. However, these devices still rely heavily on dietitians or nutritionists to analyze the nutritional values of food items.

Recent advancements in mobile phone technologies, including high-performance processors and high-quality digital cameras, have created the opportunity to invent a food image analysis system on smartphones. While the exact origins of applying AI for IADA research are uncertain, one well-documented attempt to develop a simple system on smartphones was that of DiaWear [18]. The system implemented an artificial neural network, which is a subset of deep learning, a recently advanced technique in the field of AI. Despite achieving an accuracy rate above 75%, which was considered incredible at that time, the system's usefulness was limited because it could identify only 4 types of foods—hamburgers, fries, chicken nuggets, and apple pie. In addition, the system could not determine the portion size of the taken food image; thus, it gave a nutritional value based on a constant portion size directly.

In this paper, the architecture of IADA is divided into multistage architectures, which were prevalent in the early stages of IADA development, and end-to-end architecture, which has emerged more recently with advancements in AI techniques and food image datasets. The multistage architectures, as implied by their name, include 4 individual processes: segmentation, food identification, portion estimation, and nutrient calculations using a food composition table. This sequential process is consistent across all early-stage IADA systems [19-23]. These subprocesses are trained independently because they require specific input variables, and optimization can only be done for each step individually, not for the entire process. By contrast, the end-to-end approach, which replaces a multistep pipeline with a single model, can be fine-tuned as a whole process, making it more advanced and increasingly the focus of researchers today.

Nowadays, multistage architectures are becoming obsolete and are often referred to as traditional IADA. They played a significant role in the IADA timeline before the emergence of the end-to-end approach. Therefore, we delve into the multistage architectures, particularly focusing on food identification and portion estimation algorithms in their subsections, and provide details about the end-to-end approach in the Going Beyond the Traditional Approach With Deep Learning section. For better comparison, Figure 2 illustrates traditional dietary assessment methods and the substitution processes of IADA, along with some notable systems that indicate combining certain processes of the multistage architecture into a single model through deep learning [18,23-31].

XSL•FO
RenderX

Figure 2. Comparison of traditional dietary assessment processes and the image-assisted dietary assessment (IADA) substitution processes for the same tasks, including systems that integrate multistage architecture into a single model using deep learning. Systems referenced include DiaWear from Shroff et al [18], GoCARB from Anthimopoulos et al [23], FIVR from Puri et al [26], Im2Calories from Myers et al [27], Diabetes60 from Christ et al [28], Multitask CNN from Ege and Yanai [29], Fang et al [30], and technologies-assisted dietary assessment (TADA) from Zhu et al [24, 25,31]. 24HR: 24-hour dietary recall; CNN: convolutional neural network; EFR: estimated food record; GAN: generative adversarial network; ResNet50: residual network; SVM: support vector machine; VCG: visual geometry group; WFR: weighed food record.

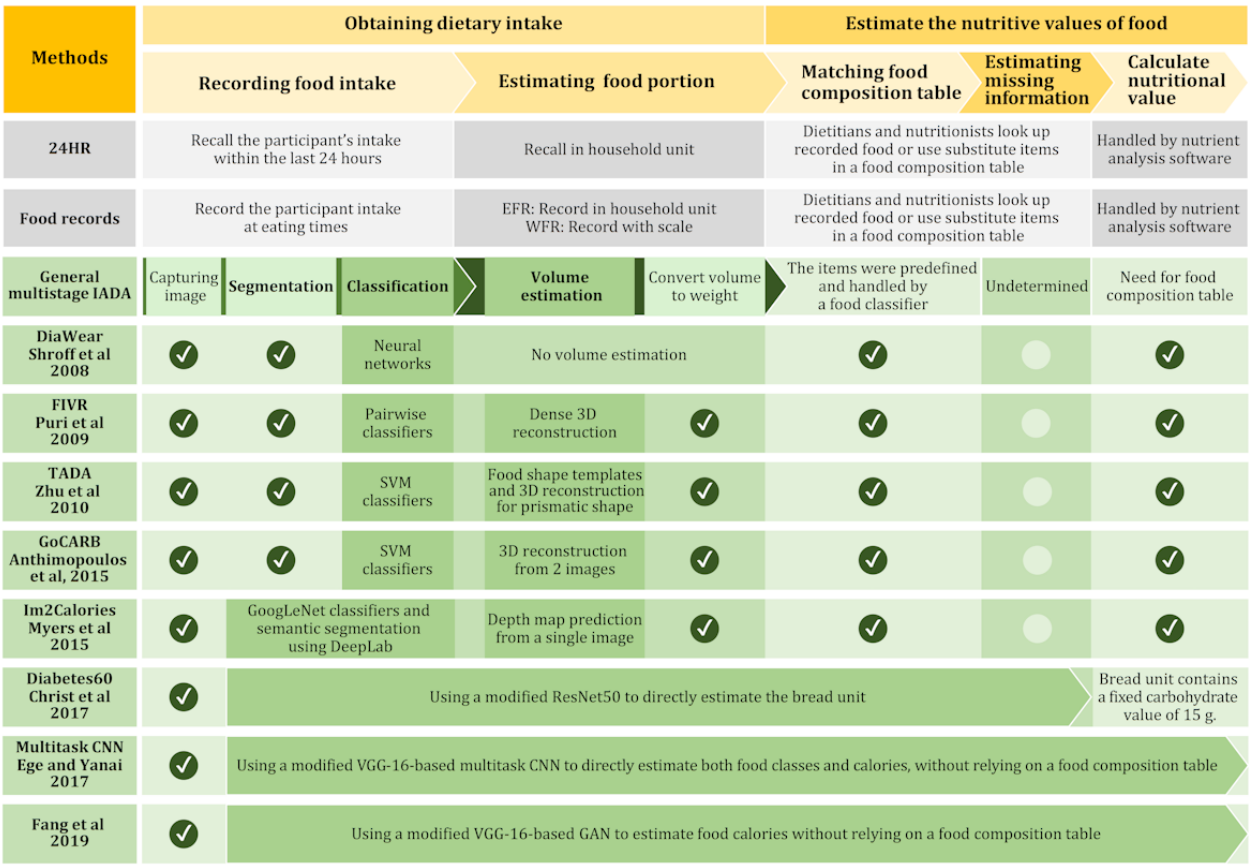

Food Identification System

Image recognition systems are one of the milestones in the computer vision field. The goal is to detect and locate an interesting object in an image. Several researchers have applied this technique to food identification tasks that formerly relied on humans only. The early stages in the development of food identification systems were from 2009 to 2015. Most of the existing systems were powered by machine learning algorithms that required human-designed input information, or technical terms called features. Hence, all machine learning-based algorithms are classified as handcrafted algorithms.

The era of handcrafted algorithms began in 2009 with the release of the Pittsburgh Fast-Food Image Dataset [19], marking a significant historical landmark in promoting research into food identification algorithms. This dataset consisted of 4545 fast-food images, including 606 stereo image pairs of 101 different food items. In addition, researchers provided baseline detection accuracy results of 11% and 24% using only the image color histogram together with the support vector machines (SVMs)-based classifier and the bag-of-scale-invariant feature transform classifier, respectively. Although these classifiers were commonly used during that time, the results were not considered sufficient and demonstrated much room for improvement. Since then, various techniques have been proposed to improve the accuracy of food classification from images. In later studies, the same team used pairwise statistics to detect ingredient relations in food images, achieving an accuracy range of 19% to 28% on the Pittsburgh Fast-Food Image Dataset [20]. Taichi and Keiji [21], from the University of Electro-Communications (UEC) team, used multiple kernel learning, which integrates different image features such as color, texture, and scale-invariant feature transform. This method achieved 61% accuracy on a new dataset of 50 food images and 37.5% accuracy on real-world images captured using a mobile phone [21]. In 2011, Bosch et al [22] from the Technology Assisted Dietary Assessment (TADA) team achieved an accuracy of 86.1% for 39 food classes by using an SVM classifier. This approach incorporated 6 features derived from color and texture [22]. These results suggest that including a larger number of features in the algorithms could potentially improve detection accuracy.

After active research, the accuracy of handcrafted algorithms reached a saturation point for improvement during the 2014 period. The optimized bag-of-features model was applied to food image recognition by Anthimopoulos et al [23]. It achieved an accuracy level of up to 77.8% for 11 classes of food on a food image dataset containing nearly 5000 images for the type 1 diabetes project called GoCARB. Pouladzadeh et al [32]

XSL•FO
RenderX

achieved a 90.41% accuracy for 15 food classes using an SVM classifier with 4 image features: color, texture, size, and shape. Kawano and Yanai [33] (UEC) attained a 50.1% accuracy for a new dataset comprising 256 food classes, using a one-vs-rest classifier with a Fisher vector and a derived feature from a color histogram named RootHoG [33]. While handcrafted algorithms yielded high-accuracy results for their specific test datasets with fewer food classes, they struggled to effectively handle larger class sets and real-world images. This difficulty arose due to factors, such as challenging lighting conditions, image noise, distorted food shapes, variations in food colors, and the presence of multiple items within the same image. Handcrafted algorithms may reach a limitation in their ability to improve further.

In contrast, the novel approach called deep learning, which can automatically extract features from input data, appears to be more suitable for complex tasks such as food identification. The convolutional neural network (CNN), considered to be one of the approaches in deep learning, was developed for handling image analysis in 1998 [34]. CNN reads a group of squared pixels of an input image, referred to as a receptive field, and then applies a mathematical function to the read data. The operation is performed repeatedly from the top-left corner until reaching the bottom-right corner of an input image. This operation is done in a similar manner to matrix multiplication or dot product in linear algebra. CNN and deep learning were applied to the food identification task in 2014 by the UEC team [35]. This system achieved an accuracy of 72.3% on a dataset containing 100 classes of real-world Japanese food images, named UEC FOOD-100, surpassing their previous handcrafted system in 2012, which achieved 55.8% on the same dataset [36]. This marked the beginning of the era of applying deep learning techniques for food identification. Later that year, the UEC team also released an international food image dataset called UEC FOOD-256 that contained 256 food classes to facilitate further research [37]. Simultaneously, the FOOD-101 dataset was made available, comprising nearly 101,000 images of 101 different food items [38]. They also presented baseline classification results from the random forest–based algorithm, one of the handcrafted algorithms, and compared it with CNN. They found that CNN achieved an accuracy of 56.4%, while random forest–based algorithm achieved 50.76% accuracy in this dataset. These food image datasets have become the favored benchmark for subsequent food identification systems.

Another important technique is transfer learning, which is well-known for training many deep learning algorithms, including CNNs. It involves 2 stages: pretraining and fine-tuning. Initially, the model is trained with a large and diverse image dataset, and then it is further trained with a smaller, more specific dataset to enhance detection accuracy. This approach is similar to how humans are educated, where broad knowledge is learned in school followed by deeper knowledge in university. The UEC team applied this training approach to the food identification task in 2015 and successfully achieved an accuracy of 78.77% on the UEC FOOD-100 dataset [39]. It has been reported that pretraining on large-scale datasets for both food and nonfood images could improve the classification system's accuracy beyond 80% [40-45], which is considered to surpass all handcrafted algorithms and be sufficient for real-world applications.

Currently, numerous state-of-the-art object detectors or classifier models, including the pretrain and fine-tune training paradigm, have been developed and are available, such as AlexNet (AlexNet is an object detection model that won the ImageNet Challenge in 2012; it is named after its inventors, Alex Krizhevsky) [46], region-based CNN (R-CNN; an object detection model that significantly improved object detection performance by combining region proposals with CNNs) [47], residual network (ResNet; a deep learning model that won the ImageNet Challenge in 2015, known for its innovative use of residual learning to train very deep networks) [48], You Only Look Once (YOLO; it is an object detection model that introduced a novel approach by framing object detection as a single regression problem, predicting bounding boxes and class probabilities directly from full images in one step evaluation) [49], Visual Geometry Group (VGG) [50], and Inception (this is an object detection model that won the ImageNet Challenge in 2014, recognized for its use of a novel architecture that efficiently leverages computing resources inside the network) [51]. These object detectors have been designed to automatically extract features from input images and learn distinct characteristics of each class during the training process. Deep learning-based object detection models have shown great promise in image recognition tasks, especially in complex tasks such as food identification. These models and their derivatives are commonly found in many of the food identification systems developed later. The use of these state-of-the-art models presents an exciting opportunity for nutrition researchers who may not have a background in computer engineering or data science. They can now create high-performance food identification systems for specific tasks by curating a food image dataset and training the model accordingly. With the various algorithms available, it is crucial to carefully consider their unique characteristics to select the most suitable one for a given application. The notable food identification systems are listed in Table 1.

XSL•FO
RenderX

Table 1. Overview of notable food identification systems, classifier algorithms, selected features, number of classes, name of food dataset (if specified or noted as their own dataset if absent), and accuracy results[a].

Study, year	Projects or team	Classifier	Feature	Class (dataset)	Accuracy results percentages
Shroff et al [18], 2008	• DiaWear	• Neural network	• Color, size, shape, and texture	• 4	• ~75
Chen et al [19], 2009	• PFID[b]	• SVM[c]	• Color • BoSIFT[d]	• 61 (PFID)	• ~11 • ~24
Taichi and Keiji [21], 2009	• UEC[e]	• MKL[f]	• Color, texture, and SIFT[g]	• 50	• 61.34
Hoashi et al [52], 2010	• UEC	• MKL	• BoF[h], Gabor[i], color, HOG[j], and texture	• 85	• 62.53
Yang et al [20], 2010	• PFID	• SVM	• Pairwise local features	• 61 (PFID)	• 78.00
Zhu et al [31], 2010	• TADA[k]	• SVM with Gaussian radial basis kernel	• Color and texture	• 19	• 97.20
Kong and Tan [53], 2011	• DietCam	• Multiclass SVM	• Nearest neighbor Gaussian region detector, and SIFT	• 61 (PFID)	• 84.00
Bosch et al [22], 2011	• TADA	• SVM	• Color, entropy, Gabor, Tamura[l], SIFT, Haar wavelet[m], steerable[n], and DAISY[o]	• 39	• 86.10
Matsuda et al [36], 2012	• UEC	• MKL-SVM	• HOG, SIFT, Gabor, color, and texture	• 100 (UEC-Food100)	• 55.80
Anthimopoulos et al [23], 2014	• GoCARB	• SVM	• HSV[p]-SIFT, optimized BoF, and color moment invariant	• 11	• 78.00
He et al [54], 2014	• TADA	• k-nearest neighbors	• DCD[q], SIFT, MDSIFT[r], and SCD[s]	• 42	• 65.4
Pouladzadeh et al [32], 2014	__[t]	• SVM	• Color, texture, size, and shape	• 15	• 90.41
Kawano and Yanai [35], 2014	• UEC	• Pretrained CNN[u]	—	• 100 (UEC-Food100)	• 72.3
Yanai and Kawano [39], 2015	• UEC	• Deep CNN	—	• 100 (UEC-Food-100)	• 78.77
Christodoulidis et al [40], 2015	• GoCARB	• Patch-wise CNN	—	• 7	• 84.90
Myers et al [27], 2015	• Google	• GoogLeNet	—	• 101	• 79.00

XSL•FO
RenderX

Study, year	Projects or team	Classifier	Feature	Class (dataset)	Accuracy results percentages
Liu et al [41], 2016	—	• DeepFood	—	• Food-101 • UEC-256	• 77.40 • 54.70
Singla et al [42], 2016	—	• GoogLeNet	—	• 11	• 83.60
Hassannejad et al [43], 2016	—	• InceptionV3[v]	—	• 101 (Food-101) • 100 (UEC-Food100) • 256 (UEC-Food256)	• 88.28 • 81.45 • 76.17
Ciocca et al [44], 2017	—	• VGG[w]	—	• 73 (UNIMINB2016)	• 78.30
Mezgec and Koroušić Seljak [45], 2017	—	• NutriNet (Modified AlexNet[x])	—	• 73 (UNIMINB2016)	• 86.72
Pandey et al [55], 2017	—	• Ensemble net	—	• 101 (Food-101)	• 72.10
Martinel et al [56], 2018	—	• WISeR[y]	—	• 101 (Food-101) • 100 (UEC-Food100) • 256 (UEC-Food256)	• 88.72 • 79.76 • 86.71
Jiang et al [57], 2020	—	• MSMVFA[z]	—	• 101 (Food-101) • 172 (VireoFood-172) • 208 (ChineseFoodNet)	• ~90.47 • 90.61 • 81.94
Lu et al [58], 2020	• GoCARB	• Modified InceptionV3	—	• 298 Generic food • Subgroups • Fine-grained • (MADiMA[aa])	• 65.80 • 61.50 • 57.10
Wu et al [59], 2021	—	• Modified AlexNet	—	• 22 styles of Bento sets	• 96.30

[a]Note that convolutional neural network–based classifiers do not require the number of features to be shown as they extract features autonomously.

[b]PFID: Pittsburgh Fast-Food Image Dataset.

[c]SVM: support vector machine.

[d]BoSIFT: bag-of-scale-invariant feature transform.

[e]UEC: University of Electro-Communications.

[f]MKL: multiple kernel learning. This is a machine-learning technique that combines multiple kernels or similarity functions, to improve the performance and flexibility of kernel-based models such as support vector machines.

[g]SIFT: scale-invariant feature transform.

[h]BoF: bag-of-features.

[i]Gabor is a texture feature extraction invented by Dennis Gabor.

[j]HOG: histogram of orientated gradients—a feature descriptor based on color.

[k]TADA: Technology Assisted Dietary Assessment.

[l]Tamura is a 6-texture feature extraction invented by Hideyuki Tamura.

[m]Haar wavelet is a mathematical analysis for wavelet sequence named after Alfréd Haar.

[n]Steerable filter is an image filter introduced by Freeman and Adelson.

[o]DAISY is a local image descriptor introduced by E Tola et al [60], but they did not describe a true acronym of DAISY.

[p]HSV is the name of a red-green-blue color model based on hue, saturation, and value.

[q]DCD: dominant color descriptor.

[r]MDSIFT: multiscale dense scale-invariant feature transform.

[s]SCD: scalable color descriptor.

[t]Not available.

[u]CNN: convolutional neural network.

[v]Inception is an object detection model that won the ImageNet Challenge in 2014, recognized for its use of a novel architecture that efficiently leverages computing resources inside the network.

[w]VGG: visual geometry group—an object detection model named after a research group from the University of Oxford.

XSL•FO
RenderX

[x]AlexNet is an object detection model that won the ImageNet Large-Scale Visual Recognition Challenge (also known as the ImageNet challenge) in 2012; it is named after its inventors, Alex Krizhevsky.

[y]WISeR: wide-slice residual.

[z]MSMVFA: multi-scale multi-view feature aggregation.

[aa]MADiMA: Multimedia Assisted Dietary Management.

Food Portion Size Estimation System

Overview

Food portion size estimation is a challenging task for researchers as it requires more accurate information on the amount of food, ingredients, or cooking methods that cannot be obtained from only a captured image without additional input, which makes it harder to create a food image dataset with portion size annotation. Furthermore, quantifying an object's size from a single 2D image is faced with common image perspective distortion problems [61,62], as shown in Figure 3. First, the size of the object in the image can change due to the distance between the object (food) and the capturing device (smartphone or camera). The size of the white rice in Figure 3A is smaller compared with Figure 3B because the white rice in Figure 3B is closer to the camera. Second, the angle at which the photo is taken also alters the perceived object size. For example, flattened objects such as rice, that are spread out on a 23-cm (9-inch) circular plate appear in their full size in a bird's-eye shot (90°), in Figure 3C, but they appear smaller when taken from approximately 30° from the tabletop as in Figure 3D. Thirdly, there is a loss of depth in a bird's-eye view in Figures 3E and 3F, making it difficult to compare between food B and food C. The weights of foods A, B, C, and D are 48, 49, 62, and 149 grams, respectively. We use these images for teaching image-based portion estimation for dietetics students.

While pretrain and fine-tune training for CNNs is a silver bullet for food image identification, currently there is no equivalent solution for portion estimation. Many researchers are actively finding ways to calibrate the object size within an image to mediate such an error, and several approaches have been discussed here. Basically, portion estimation can be broadly classified, based on complexity, into four progressive categories: (1) pixel density, (2) geometric modeling, (3) 3D reconstruction, and (4) depth camera. Table 2 provides an overview of notable systems for volume estimation.

Figure 3. There are common image perspective distortion problems. Firstly, position distortion: the size of the white rice in (A) is smaller compared to (B) because the white rice in (B) is closer to the camera. Secondly, angle distortion: the white rice in (C) is fully visible at 90 degrees, while it appears smaller when taken from 30 degrees, as in (D). Thirdly, there is a loss of depth information in the bird's-eye view in (E) and (F), making it difficult to compare food B and food C.

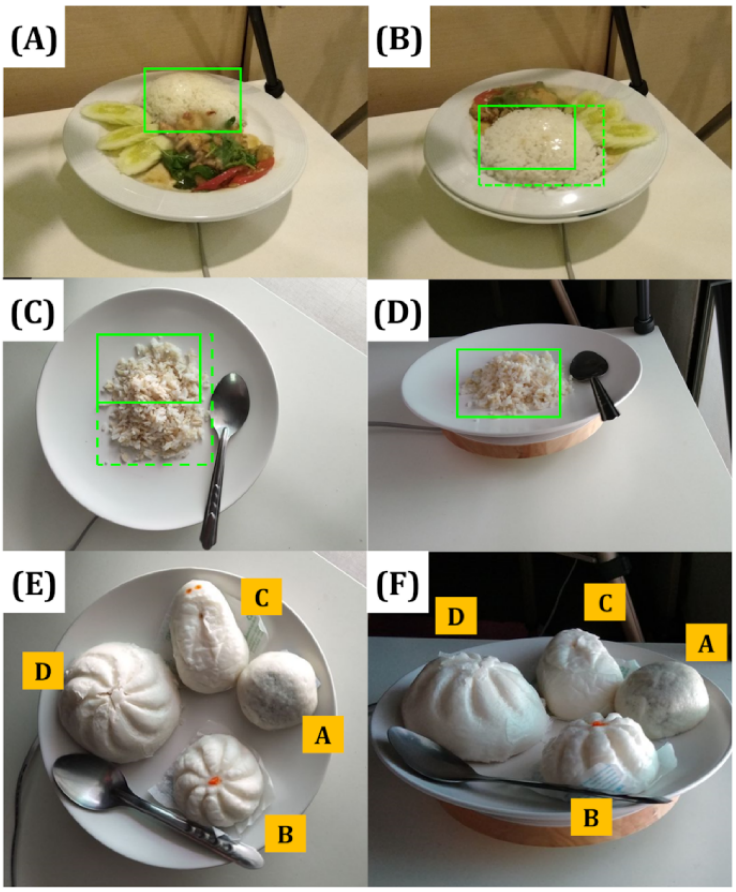

XSL•FO
RenderX

Table 2. A comprehensive overview of notable publications for 4 volume estimation approaches, arranged chronologically.

Approach and study, year	Projects or team	Reference object	Item	Reported error
Pixel density approach				
Martin et al [13], 2009	—[a]	• Physical card	• N/A[b]	• N/A
Jia et al [63], 2012	• University of Pittsburgh	• Circular plate • Circular LED light	• —	• <27.60 • <54.10
Pouladzadeh et al [32], 2014	—	• User's thumb	• 5	• <10
Okamoto and Yanai [64], 2016	• UEC[c]	• Wallet	• 3	• Mean calorie error • Beef rice bowl –242 (SD 55.1) • Croquette –47.08 (SD 52.5) • Salad 4.86 (SD 11.9)
Akpa et al [65], 2017	—	• Chopstick	• 15	• <6.65
Liang and Li [66], 2017	—	• 1-yuan coin	• 19 fruits	• 15 items <20%
Yanai et al [67], 2019 and Ege et al [67], 2019	• UEC	• Rice grain size	• 3	• <10%
Geometric modeling approach				
Zhu et al [24], 2010 and Zhu et al [25], 2008	• TADA[d]	• Checkerboard	• 7	• Spherical 5.65% • Prismatic 28.85%
Chae et al [69], 2011	• TADA	• Checkerboard	• 26	• Cylinders 11.1% • Flattop solid 11.7%
Chen et al [70], 2013	• University of Pittsburgh	• Circular plate	• 17	• 3.69%
Jia et al [71], 2014	• University of Pittsburgh	• Circular plate • Other container	• 100	• <30% from 85/100 of test items
Tanno et al [72], 2018	• UEC	• Apple ARKit	• 3	• Mean calorie error • Beef rice bowl –67.14 (SD 18.8) • Croquette–127.0 (SD 9.0) • Salad –0.95 (SD 0.16)
Yang et al [73], 2019	• University of Pittsburgh	• Augmented reality	• 15	• Large objects 16.65% • Small objects 47.60%
Smith et al [74], 2022	—	• Checkerboard	• 26	• Single food items 32.4%-56.1% • Multiple food items 23.7%-32.6%
3D reconstruction approach				
Puri et al [26], 2009	—	• 3 images • Checkerboard	• 26	• 2%-9.5%
Kong and tan [75], 2012	—	• 3 images • Checkerboard	• 7	• Volume estimation error 20%
Rahman et al [76], 2012	• TADA	• 2 images • Checkerboard	• 6	• 7.70%

XSL•FO
RenderX

Approach and study, year	Projects or team	Reference object	Item	Reported error
Chang et al [77], 2013	• TADA	• Using food silhouettes to reconstruct a 3D object	• 4	• 10%
Anthimopoulos et al [78], 2015	• GoCARB	• 2 images physical card Physical card	N/A	• Volume estimation error 9.4%
Dehais et al [79], 2017	• GoCARB	• 2 images • Physical card	• 45 dishes • 14 meals	• 8.2%-9.8%
Gao et al [80], 2018	—	• SLAMe-based with Rubik cube	• 3	• 11.69%-19.20% for static measurement • 16.32%-27.9% for continuous measurement

Ando et al [81], 2019	• UEC	• Multiple cameras on iPhone X for depth estimation	• 3	• Calorie estimation error
				  • Sweet and sour pork <1%
				  • Fried chicken <1%
				  • Croquette <15%

Lu et al [58], 2020	• GoCARB	• 2 images • Physical card and gravity information	• 234 items from MADiMA[f]	• MARE[g] 19%, while their earlier system, GoCarb (2017), achieved 22.6% on the same task [79].
Depth camera approach				
Shang et al [82], 2011	—	• Specific food recording device	—	• No performance report
Chen et al [83], 2012	—	• Depth camera	—	• No performance report
Fang et al [84], 2016	• TADA	• Camera from this study [85]	• 10	• Depth method overestimates volume than geometric model
Alfonsi et al [86], 2020	—	• iPhone and Android devices	• 200	• Carbohydrate estimation error <10 g
Herzig et al [87], 2020		• iPhone X	• 128	• Relative error of weight estimation 14.0%

[a]Not available.
[b]N/A: not applicable.
[c]UEC: University of Electro-Communications.
[d]TADA: Technology Assisted Dietary Assessment.
[e]SLAM: simultaneous localization and mapping.
[f]MADiMA: Multimedia Assisted Dietary Management.
[g]MARE: mean absolute relative error.

Revisiting the Classic Pixel Density Approach

Pixel density is the simplest approach for providing good and effective estimation. After a food image is segmented, the number of pixels in each segmented section is determined. Mathematical equations or other transformations are then used to calculate the portion size of each section that is presented in the image.

However, this approach suffers from image distortion problems, and several approaches have been implemented to combat this drawback. The simplest method is the use of a physical reference object or fiducial marker for calibrating the size of objects in an image. When the real size of the reference object is known, the real size of an object can be determined relative to the reference object. This method was chosen for food volume estimation during its early development stage [13,88,89]. Various physical objects have been used as reference objects in the literature, including a special patterned card [13,89], a known-size circular plate [63] or bowl [90], chopsticks [65], a 1-yuan coin [66], a wallet [64], a user's thumb [40,91], or even rice grain size [67].

Geometric Modeling Approach

Assuming that the food has a cylindrical shape, such as compressed steamed rice (Figure 4A), its volume can be calculated using the conventional formula $2\pi r^2 \times h$. The radius

XSL•FO
RenderX

r and height h can be determined by counting the pixels in the image. While this approach is effective for geometric shapes, it is less reliable for irregular shapes that lack a specific equation. The demonstration of this approach is shown in Figure 4B, where the user selects a predefined shape and then manually fits (or registers) the geometric model with the image.

The TADA team reported the use of several predefined shapes of foods, including cylindrical, flattop solid, spherical, and prismatic models [24,25,68,69]. Prismatic models were specifically used to estimate portion sizes of irregularly shaped foods. This approach allowed a more accurate estimation of portion sizes by considering the unique characteristics of each food item. The research team at the University of Pittsburgh proposed a similar technique known as wireframe modeling. This technique involves creating a skeletal representation of an object using lines and curves to define its structure to accurately capture the shape and dimensions of food items [70,71]. However, this approach is also affected by common image distortion problems. Initially, a physical reference object was used for calibration.

Geometric modeling shares a fundamental principle with augmented reality (AR), a technology that transforms 2D environmental images into 3D coordinates in a computer system. As AR has become more widely available on smartphones, many researchers have explored the feasibility of using AR as a calibration method instead of using physical reference objects [72,73]. AR-based object length measurement is demonstrated in Figure 5.

Figure 4. This figure demonstrates the various approaches to estimating food volume. (A) A cylindrical shape of 75 grams of brown rice taken from a 60° angle. (B) Geometric modeling with a predefined cylindrical shape, where the user needs to adjust each point manually to fit the object. (C) A predicted depth map from state-of-the-art dense prediction transformation. (D) A 3D reconstructed object using depth information from (C). These images have been adjusted in size for visual comparison purposes.

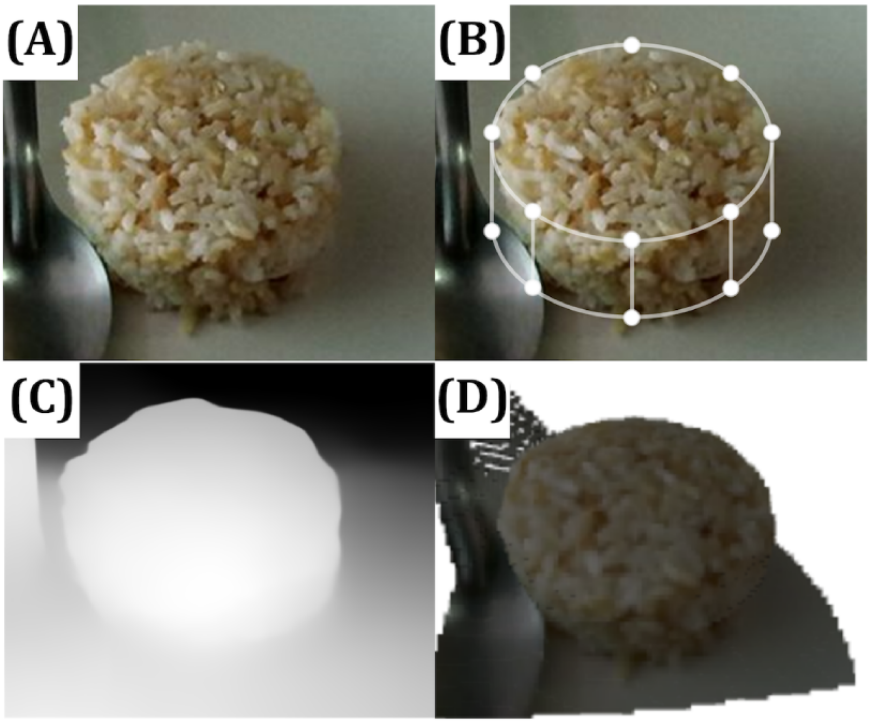

Figure 5. Measuring the size of the same banana can be done using different techniques, as shown in the figure. (A) A standard ruler is used as a ground truth measurement, (B) Samsung augmented reality Zone app, and (C) Apple iPhone Measure app. These apps use the gyroscope or accelerometer sensors in the mobile phone to accurately track the movement of the phone as the measurement line is drawn.

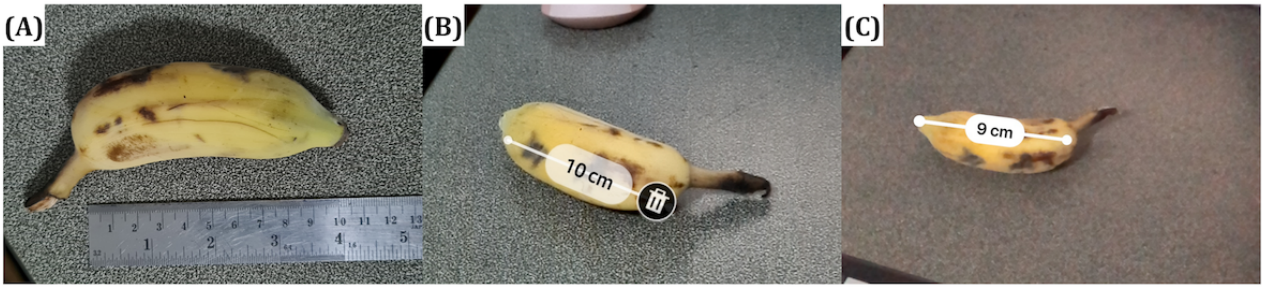

XSL•FO
RenderX

3D Reconstruction

This technique involves using ≥2 images taken from different angles to create virtual 3D objects in 3D coordinates in a computer system. It shares the same principle as both AR and geometric modeling, where reconstructed objects are represented similarly to prismatic models in geometric modeling. Furthermore, this technique allows for the inclusion of shapes beyond traditional geometric shapes.

While several researchers have explored the use of 3D reconstruction [26,75,76], 1 notable example is the GoCARB system [78]. This system requires 2 images taken from different angles to construct a 3D model of the food, achieving an accuracy within 20 grams for carbohydrate content estimation. This level of accuracy is comparable to estimates made by dietitians when the food is completely visible on a single dish with an elliptical plate and flat base [92].

Figures 4C and 4D demonstrate a similar 3D reconstruction approach but implemented using state-of-the-art dense prediction transformation models to predict depth maps from a single image (Figure 4A), followed by the reconstruction of the 3D object using the predicted depth map.

Depth Camera Approach

This method operates on the same principle as geometric modeling and 3D reconstruction, but it requires a special time-of-flight (ToF) sensor (also known as a depth camera) to measure an object's size in 3D coordinates in a computer system. Initially, the application of depth cameras in food volume estimation was limited, primarily due to their high cost and limited availability [82]. However, with the introduction of consumer-grade depth cameras, such as Kinect (Microsoft Corp), Intel RealSense, and smartphones equipped with depth sensors, their accessibility increased, leading to wider use in food volume estimation applications [81,83,84,86,87].

Nevertheless, the availability of depth sensors remains a significant challenge in implementing this system. Currently, only a limited number of mobile phone models are equipped with such sensors. In addition, some manufacturers integrate the sensor with the front camera for authentication purposes, such as Apple's FaceID, making it impractical for capturing object photos. Moreover, certain mobile device manufacturers have omitted the ToF sensor in their recent models [93], further reducing the availability of depth sensors and posing implementation challenges for the depth camera approach.

An example of depth information captured by the Intel Realsense d435i depth camera displayed in RGB (red-green-blue; color model based on additive color primaries) with depth (RGB with depth; RGBD) format is shown in Figure 6B. Rendered objects from a captured polygon file are demonstrated as freely rotatable 3D objects in Figures 6C and 6D, with a regular RGB image shown for comparison in Figure 6A.

Figure 6. (A) A typical red-green-blue image showing 3 Burmese grapes, each weighing approximately 20 grams. (B) A red-green-blue image with depth captured by Intel RealSense d435i from a bird's-eye view. (C) and (D) 3D reconstructed objects from the polygon file, illustrating the height of each fruit from different angles.

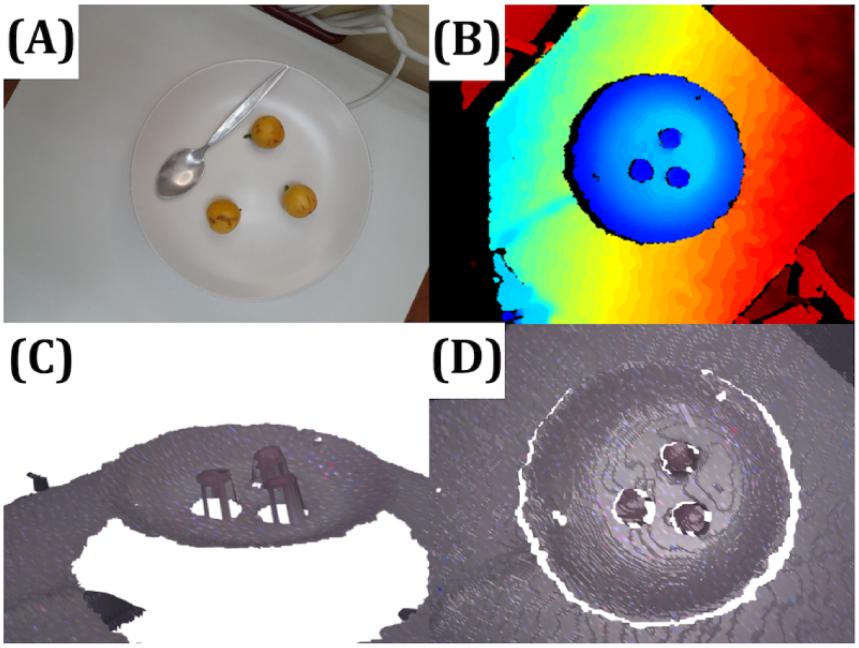

Going Beyond the Traditional Approach With Deep Learning

Advancements in deep learning are opening more possibilities to improve the IADA system by merging some steps (or even all steps) of the multistep pipeline into a single model, which can be fine-tuned as a whole process. Due to the rise in IADA research with the emergence of advanced algorithms, we can only highlight a few reports that demonstrate the gradual enhancements in IADA in this paper.

In 2015, Myers et al [27] from Google proposed the Im2Calories system, using deep learning for all stages of IADA. The classifiers are based on the GoogLeNet architecture, and the

XSL•FO
RenderX

classification results are used to improve the semantic segmentation handled by the DeepLab network. For volume estimation, a new CNN architecture, trained with an RGBD dataset, estimates the depth map from a single RBG image and then converts the depth map to volume in the final step. Although the absolute error for some test foods could exceed 300 ml, the overall volume estimation results were deemed acceptable. The system still requires a food composition database to determine the nutritional values of the food in the final step.

The idea of using deep learning to estimate food volume is gaining popularity, and several systems are transitioning to using deep learning algorithms to estimate food volume without the need for an actual ToF sensor. In 2017, carbohydrate counting algorithms named Diabetes60 were proposed by Christ et al [28]. The system reported food-specific portions called "bread units," which are defined to contain 12 to 15 grams of carbohydrates. This definition closely resembles the "carb unit" widely used in the diabetes field or the "exchange unit" in dietetic practice. The system was based on ResNet50 and trained using an RGBD image dataset that contained human-annotated bread unit information. It achieved a root mean square error of 1.53 (approximately 18.4-23 g of carbohydrate), while humans could achieve a root mean square error of 0.89 (approximately 10.7-13.4 g of carbohydrate) when compared with the ground truth. The modified ResNet was also used for fruit volume estimation, achieving an error of 2.04% to 14.3% for 5 types of fruit and 1 fruit model [94]. Furthermore, Jiang et al [95] introduced a system to classify liquid levels in bottles into 4 categories: 25%, 50%, 75%, and 100%. Using their own designed CNN architecture, they achieved a 92.4% classification accuracy when the system was trained with 3 methods of data augmentation. Furthermore, the system could achieve 100% classification accuracy when the bottle images had labels removed.

One challenge in converting a single 2D image into a 3D object is the difficulty in capturing the back side of an object in single-view images due to factors such as view angle or occlusion. Therefore, the food volume may be underestimated. Point2Volume was introduced in 2020 by Lo et al [96] to address the limitations. The system builds upon 2 of their previous works: a deep learning view synthesis [97] and a point completion network [98]. When a single-depth image is captured, a Mark region-based CNN—a combination of object detection and instance segmentation network—performs classification and segmentation, obtaining only partial point clouds due to occlusion. It then reconstructs the complete shapes and finally estimates the food volumes. This system demonstrated a volume estimation error of 7.7% for synthetic foods and 15.3% for real foods.

While the estimation of exact food volume has improved recently, dietitians and nutritionists often use a different approach. They compare unknown food amounts with known reference volumes, such as a thumb, matchbox, tennis ball, deck of cards, or a series of known portion-size images. Yang et al [99] introduced a system that mimics this mental estimation approach in 2021. The system classifies the unknown portion object to match the system's set of reference volumes and then fine-tunes the predicted volume using the selected set. The system achieved a mean relative volumetric error of around 11.6% to 20.1% for their own real food image dataset. Interestingly, they noted that even when the system chose the wrong set of reference volumes—due to top-1 accuracy being <50% in most cases—the mean relative volumetric error still remained acceptable, implying that fewer reference volume sets might be sufficient.

Another crucial question is how many food classes should be included in the system to achieve usability in day-to-day situations. The goFood system [58], successor to the previous carbohydrate estimation system GoCARB, takes a different approach to expand the coverage beyond their included food classes. Using a modified Inception V3 architecture to classify food into a 3-level hierarchical structure: 18 types of generic food (eg, meat, bread, and dairy), 40 types of subgroups (eg, white bread and red meat), and 319 types of specific foods. This strategy mirrors the concept of a food exchange list, allowing the handling of a large number of foods without the need for an extensive number of fine-grained classifications. This lowers the number of unidentified food objects and results in achieving at least a 3% higher accuracy for food identification than the single-level Inception V3 classifier. Their newer 3D reconstruction algorithm, incorporating gravity data from the smartphone's inertial measurement unit (eg, accelerometer or gyroscope), achieved a mean absolute relative error of 19%, surpassing the algorithm in GoCARB, which had 22.6% error.

Furthermore, CNN and deep learning could potentially estimate nutrients directly without relying on food composition tables, enabling an end-to-end approach for IADA. The originality of this method is unclear, but to the best of our knowledge, the first well-documented system was introduced by Miyazaki et al [100] in 2011. This system extracts 4 features from food images and estimates calories from these features instead of relying on food identification, portion estimation, and food composition tables as in multistage IADA. The system achieved a relative error of approximately 20% for 35% of items and 40% for 79% of items, which is relatively high. This idea inspired subsequent works by Ege and Yanai [29] from UEC in 2017. They applied a multitask CNN, a technique where a model is trained to perform multiple tasks simultaneously, using visual geometry group-16 for feature extraction and a calorie-annotated image dataset for training. The CNN system achieved an estimation error of 20% for 50% of items and 40% for 80% of items in their Japanese food image dataset. However, the system assumed that each food image contained only 1 food item; this limitation was addressed in their later works [101,102]. Multitask CNNs can be fine-tuned for the entire algorithm rather than for each stage as in a multistage architecture. This gives them the potential to surpass multistage architectures, similar to how deep learning and CNNs have outperformed handcrafted food identification algorithms. Therefore, they have gained significant attention from researchers [103-107].

Not only multitask CNNs but also generative adversarial networks, which are the backbone of image generation AI, such as Dall-E (OpenAI), can be used to learn the energy distribution map and estimate food energy directly from a single RGB image. Fang et al [30] from the TADA team applied this approach and

XSL•FO
RenderX

achieved a mean energy estimation error of 209 kcal. Their subsequent work, which included adding food localization networks, improved accuracy by approximately 3.6% [108]. While most system predictions focus on food portions (volume or weight), calories, or macronutrients such as carbohydrates, in 2019, Situju et al [109] used a multitask CNN to predict the salt content of 14 types of food. This was achieved by training the multitask CNN with a dataset annotated for both calories and salt. The relative estimation error was 31.2% (89.6 kcal) for calories and 36.1% (0.74 g) for salt. These works provide evidence that advanced deep learning techniques yield promising results and offer room for improvement in IADA, garnering increasing attention from researchers today.

Advancements and Challenges From the Dietitian's Perspective

Overview

According to recently published information, both image classification and volume estimation techniques are comparable in accuracy to those of untrained humans or even trained professionals in some situations [92,110]. Some limitations exist, however, in relying on traditional methods, which indicates that another auxiliary system might be necessary to improve the overall accuracy and usefulness of a future developed system.

Using Recipe-Specific Nutritional Values

Currently, most existing systems rely on standard food composition tables to calculate the representative nutritional values of foods. While the United States Department of Agriculture National Nutrient Database is considered comprehensive, in practical dietetics, it is important to use recipe-specific nutritional values when available. For example, differentiating between a Subway sandwich (Subway IP LLC) and a Starbucks sandwich (Starbucks Corporation) using a food identification system may be feasible with a large image dataset of these specific sandwiches. However, it could be more straightforward to use location data to determine the brand of the sandwich.

Furthermore, when a food product has a nutrition facts label, it is essential to obtain the representative values directly from the label instead of relying solely on food composition tables. This can be accomplished either through a system equipped with optical character recognition or by accessing a vast nutrition facts label database, such as Open Food Facts [111]. By incorporating these recipe-specific and label-based nutritional values, the accuracy and relevance of food nutrient assessment systems can be significantly improved.

Challenges With Density Determination

The conversion of volume to weight in volume estimation approaches relies on food-specific density values, which can pose technical difficulties [112]. Furthermore, food-specific density is not provided in all food compositions; therefore, it must be obtained through calculation. Most food composition tables provide nutrient content per 100 grams of edible food, as it is derived from direct chemical analysis procedures. By contrast, food portion sizes are often measured in household units, such as teaspoons, tablespoons, or measuring cups.

The portion-specific weight must be divided by the standard volume of the household unit to calculate density. For example, according to the Thai food composition table, cooked mung bean sprouts weigh 78 and 34 grams for 1 serving (240 mL) and 1/3 serving (80 mL), respectively. This results in food-specific densities of 0.325 and 0.425 g/mL. However, relying on a single representative density value may not be appropriate, as it can contribute to overall system errors beyond just volume estimation. To address this challenge, a calibration curve-like method should be used instead of relying on a single density value. The accuracy and reliability of volume estimation systems can be improved, thus ensuring more precise and consistent results.

Guessing Missing Information

When assessing food intake, dietitians and nutritionists often encounter situations where certain food items are not readily available in food composition tables or nutrition databases. In such cases, a comprehensive analysis of the food needs to be conducted, breaking it down into its individual components. Using plain fried rice with egg as an example, the 2 cups of fried rice should be divided into at least 2 components: steamed white rice and chicken egg, which are visible in the image. However, additional components, such as seasonings and cooking oil, must be estimated. Seasonings, such as salt, soy sauce, and sugar, are typically added to enhance flavor, while cooking oil is often used to prevent food from sticking to the pan and to aid in the cooking process. Furthermore, the amount of seasoning and cooking oil may vary based on the personal experience or preference of the nutritionist who analyzes the food. Consequently, in nutrition research, it is recommended to have at least 2 or 3 analysts to reduce individual bias [113]. Using algorithms, which are based on standardized criteria, the variation caused by personal experience and subjectivity can be reduced.

Explainable System and Trust Issues

Using AI in health care has attracted close attention from health care communities worldwide, raising concerns about how to trust unexplained systems [114-116]. This concern is also shared by nutrition professionals. The black-box nature of deep learning algorithms makes it difficult for users to identify incorrect outputs.

When dietitians and nutritionists review a participant's food photo and the estimated calorie intake is lower than expected, it could be due to underreporting or misreporting by the participant, selection of an inappropriate food item, forgetting to include certain amounts of oil in recipe analysis, or underestimating portion sizes. Dietitians and nutritionists can easily identify these errors. However, if the system only provides calorie outputs without additional information, it fails to establish trust with the users. Consequently, involving nutrition professionals in the development and evaluation of these systems is crucial to build trust and ensure that the technology meets their requirements.

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we investigated the AI techniques used for IADA and analyzed the available literature to identify the principal findings in this field. Our scoping review encompassed 522 articles, and after careful evaluation, we included 84 (16.1%) articles for analysis, spanning from 2008 to 2021. After 2015, the increase in the number of published articles in this field can be attributed to various factors, including the growing availability of large datasets, advancements in AI development frameworks, and improved accessibility of hardware resources for AI-related tasks.

The principal findings were categorized into 2 main areas: food identification and food volume estimation. The chronological presentation of the articles allowed a better understanding of the algorithms' complexity and the improvements achieved in accuracy. The transition from handcrafted food identification algorithms to deep learning-based algorithms occurred within a relatively short span of 5 years. This shift demonstrated the transformative power of deep learning in enhancing the accuracy and efficiency of food identification in image-based dietary assessment. Regarding food volume estimation, 4 different approaches were identified. However, all of these approaches share the common goal of translating 2D object views into 3D representations within a computer system and then converting these to weight to estimate representative nutritional values from a food composition table. While these approaches each have their strengths and limitations, the use of depth cameras is straightforward for measuring volume with fewer assumptions and might result in the lowest error rates compared with other methods. Nonetheless, the limited availability of depth cameras in some smartphones poses a significant challenge for implementing this approach. However, recent advancements in deep learning techniques offer promising alternatives to overcome the need for specific hardware to estimate volume and even directly estimate nutritional values without using a food composition table.

Comparison With Prior Work

During our search for relevant studies, we encountered several review articles published before ours. Gemming et al [117] organized notable studies from the early stages of IADA development. Doulah et al [118] primarily focused on computational methods for determining energy intake, including IADA techniques and wearable devices aimed at replacing traditional dietary assessment methods. Lo et al [119] provided detailed explanations of techniques for both food recognition and volume estimation used in IADA studies. The survey from Subhi et al [120] and the systematic review from Dalakleidi et al [121] offer comprehensive comparisons of IADA systems, organized based on the subtasks of multistage architecture. Tay et al [122] provided an exclusive report on computational food volume estimation. While these review articles provide extensive information, they may be difficult to comprehend for nontechnical individuals, such as dietitians and nutritionists. This review is tailored to serve as a starting point for those who may not be familiar with the technical terminology and complexity associated with this field, presenting information in clear chronological order for easy following and comparison.

Strengths and Limitations

While technology has advanced rapidly over the past 2 decades, it is important to acknowledge that some of the studies included in our review may have become outdated in terms of algorithm complexity, measurement techniques, and the accuracy of predicted results. Nonetheless, the findings from these earlier studies remain crucial from a dietitian's perspective and provide valuable insights for future research and solution development. Although our search strategy was comprehensive and systematic, it is important to acknowledge that there may be studies that we were unable to identify or include in this study. Despite this limitation, our analysis provides a comprehensive overview of the principal findings in the field of IADA, shedding light on the potential and challenges of incorporating AI techniques into this domain.

Conclusions

The application of AI has demonstrated promising results in enhancing the accuracy and efficiency of IADA. Advanced technologies, such as deep learning, CNNs, multitask CNNs, and generative adversarial networks, have significantly improved digitization of dietary intake. However, despite their potential, there are still challenges to overcome when implementing these technologies in real-world settings. To achieve broader coverage and increased reliability, integrating various inputs, such as food barcodes, direct label readers through optical character recognition, and location-specific recipes, could enhance the capabilities of IADA systems.

Additional research and development efforts are needed to address persistent issues, such as the limited availability of depth cameras, interassessor variation, missing information, and density estimation. While AI-based approaches offer valuable insights into dietary intake, it is essential to recognize that they were not designed to capture long-term usual intake entirely, which could be determined by aggregating self-reported and objective measures of dietary intake.

Furthermore, combining usual intake with additional aspects of health, such as physical activity, sleep patterns, and body composition, is required for a comprehensive understanding of the relationship between lifestyle, health, and disease. By overcoming these challenges, AI-based approaches have the potential to revolutionize dietary assessment and contribute to a better understanding of an individual's intake, eating patterns, and overall nutritional health.

XSL•FO
RenderX

Acknowledgments

This work was funded by the Program Management Unit for Human Resources & Institutional Development, Research, and Innovation agency under contract B04G640044. The authors would like to thank the Institute of Nutrition, Mahidol University, for the support and use of their facilities. The authors gratefully thank Sabri Bromage for his valuable suggestions and to George Attig for editing the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

PC wrote the manuscript and provided data for tables and figures. PC and WK conceived and designed the conceptual framework. PC, AP, and PS discussed implications, limitations, and potential future directions. All authors reviewed, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist. [PDF File (Adobe PDF File), 101 KB - jmir_v26i1e51432_app1.pdf]

References

1. Thompson FE, Subar AF, Loria CM, Reedy JL, Baranowski T. Need for technological innovation in dietary assessment. J Am Diet Assoc 2010 Jan;110(1):48-51 [FREE Full text] [doi: 10.1016/j.jada.2009.10.008] [Medline: 20102826]
2. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
3. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
4. Gibson RS. Principles of Nutritional Assessment. Oxford, UK: Oxford University Press; 2005.
5. Zegman MA. Errors in food recording and calorie estimation: clinical and theoretical implications for obesity. Addict Behav 1984;9(4):347-350. [doi: 10.1016/0306-4603(84)90033-9] [Medline: 6532141]
6. Livingstone MB, Robson PJ, Wallace JM. Issues in dietary intake assessment of children and adolescents. Br J Nutr 2004 Oct;92 Suppl 2:S213-S222. [doi: 10.1079/bjn20041169] [Medline: 15522159]
7. Goris AH, Westerterp-Plantenga MS, Westerterp KR. Undereating and underrecording of habitual food intake in obese men: selective underreporting of fat intake. Am J Clin Nutr 2000 Jan;71(1):130-134. [doi: 10.1093/ajcn/71.1.130] [Medline: 10617957]
8. Wang DH, Kogashiwa M, Ohta S, Kira S. Validity and reliability of a dietary assessment method: the application of a digital camera with a mobile phone card attachment. J Nutr Sci Vitaminol (Tokyo) 2002 Dec;48(6):498-504. [doi: 10.3177/jnsv.48.498] [Medline: 12775117]
9. Nicklas TA, O'Neil CE, Stuff J, Goodell LS, Liu Y, Martin CK. Validity and feasibility of a digital diet estimation method for use with preschool children: a pilot study. J Nutr Educ Behav 2012;44(6):618-623 [FREE Full text] [doi: 10.1016/j.jneb.2011.12.001] [Medline: 22727939]
10. Wang DH, Kogashiwa M, Kira S. Development of a new instrument for evaluating individuals' dietary intakes. J Am Diet Assoc 2006 Oct;106(10):1588-1593. [doi: 10.1016/j.jada.2006.07.004] [Medline: 17000191]
11. Gregory R, Walwyn L, Bloor S, Amin S. A feasibility study of the use of photographic food diaries in the management of obesity. Pract Diabetes Int 2006 Apr 10;23(2):66-68. [doi: 10.1002/pdi.899]
12. Williamson DA, Allen HR, Martin PD, Alfonso AJ, Gerald B, Hunt A. Comparison of digital photography to weighed and visual estimation of portion sizes. J Am Diet Assoc 2003 Sep;103(9):1139-1145. [doi: 10.1016/s0002-8223(03)00974-x] [Medline: 12963941]
13. Martin CK, Han H, Coulon SM, Allen HR, Champagne CM, Anton SD. A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method. Br J Nutr 2009 Feb;101(3):446-456 [FREE Full text] [doi: 10.1017/S0007114508027438] [Medline: 18616837]
14. Dahl Lassen A, Poulsen S, Ernst L, Kaae Andersen K, Biltoft-Jensen A, Tetens I. Evaluation of a digital method to assess evening meal intake in a free-living adult population. Food Nutr Res 2010 Nov 12;54 [FREE Full text] [doi: 10.3402/fnr.v54i0.5311] [Medline: 21085516]

XSL•FO
RenderX

15. Rollo ME, Ash S, Lyons-Wall P, Russell A. Trial of a mobile phone method for recording dietary intake in adults with type 2 diabetes: evaluation and implications for future applications. J Telemed Telecare 2011;17(6):318-323. [doi: 10.1258/jtt.2011.100906] [Medline: 21844173]
16. Daugherty BL, Schap TE, Ettienne-Gittens R, Zhu FM, Bosch M, Delp EJ, et al. Novel technologies for assessing dietary intake: evaluating the usability of a mobile telephone food record among adults and adolescents. J Med Internet Res 2012 Apr 13;14(2):e58 [FREE Full text] [doi: 10.2196/jmir.1967] [Medline: 22504018]
17. Six BL, Schap TE, Zhu FM, Mariappan A, Bosch M, Delp EJ, et al. Evidence-based development of a mobile telephone food record. J Am Diet Assoc 2010 Jan;110(1):74-79 [FREE Full text] [doi: 10.1016/j.jada.2009.10.010] [Medline: 20102830]
18. Shroff G, Smailagic A, Siewiorek DP. Wearable context-aware food recognition for calorie monitoring. In: Proceedings of the 12th IEEE International Symposium on Wearable Computers. 2008 Presented at: ISWC 2008; September 28-October 1, 2008; Pittsburgh, PA. [doi: 10.1109/iswc.2008.4911602]
19. Chen M, Dhingra K, Wu W, Yang L, Sukthankar R, Yang J. PFID: Pittsburgh fast-food image dataset. In: Proceedings of the 16th IEEE International Conference on Image Processing. 2009 Presented at: ICIP 2009; November 7-10, 2009; Cairo, Egypt. [doi: 10.1109/icip.2009.5413511]
20. Yang S, Chen M, Pomerleau D, Sukthankar R. Food recognition using statistics of pairwise local features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010 Presented at: CVPR 2010; June 13-18, 2010; San Francisco, CA. [doi: 10.1109/cvpr.2010.5539907]
21. Taichi J, Keiji Y. A food image recognition system with Multiple Kernel Learning. In: Proceedings of the 16th IEEE International Conference on Image Processing. 2009 Presented at: ICIP 2009; November 7-10, 2009; Cairo, Egypt. [doi: 10.1109/icip.2009.5413400]
22. Bosch M, Zhu F, Khanna N, Boushey CJ, Delp EJ. Combining global and local features for food identification in dietary assessment. Proc Int Conf Image Proc 2011 Sep;2011:1789-1792 [FREE Full text] [doi: 10.1109/ICIP.2011.6115809] [Medline: 25110454]
23. Anthimopoulos MM, Gianola L, Scarnato L, Diem P, Mougiakakou SG. A food recognition system for diabetic patients based on an optimized bag-of-features model. IEEE J Biomed Health Inform 2014 Jul;18(4):1261-1271. [doi: 10.1109/JBHI.2014.2308928] [Medline: 25014934]
24. Zhu F, Bosch M, Woo I, Kim S, Boushey CJ, Ebert DS, et al. The use of mobile devices in aiding dietary assessment and evaluation. IEEE J Sel Top Signal Process 2010 Aug;4(4):756-766 [FREE Full text] [doi: 10.1109/JSTSP.2010.2051471] [Medline: 20862266]
25. Zhu F, Mariappan A, Boushey CJ, Kerr D, Lutes KD, Ebert DS, et al. Technology-assisted dietary assessment. Proc SPIE Int Soc Opt Eng 2008 Mar 20;6814:681411 [FREE Full text] [doi: 10.1117/12.778616] [Medline: 22128303]
26. Puri M, Zhu Z, Yu Q, Divakaran A, Sawhney H. Recognition and volume estimation of food intake using a mobile device. In: Proceedings of the Workshop on Applications of Computer Vision. 2009 Presented at: WACV 2009; December 7-8, 2009; Snowbird, UT. [doi: 10.1109/wacv.2009.5403087]
27. Myers A, Johnston N, Rathod V, Korattikara A, Gorban A, Silberman N, et al. Im2Calories: towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision. 2015 Presented at: ICCV 2015; December 7-13, 2015; Santiago, Chile. [doi: 10.1109/iccv.2015.146]
28. Christ PF, Schlecht S, Ettlinger F, Grün F, Heinle C, Tatavatry S, et al. Diabetes60 — inferring bread units from food images using fully convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017 Presented at: ICCVW 2017; October 22-29, 2017; Venice, Italy. [doi: 10.1109/iccvw.2017.180]
29. Ege T, Yanai K. Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. 2017 Presented at: Thematic Workshops '17; October 23-27, 2017; Mountain View,CA. [doi: 10.1145/3126686.3126742]
30. Fang S, Shao Z, Kerr DA, Boushey CJ, Zhu F. An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: protocol and methodology. Nutrients 2019 Apr 18;11(4):877 [FREE Full text] [doi: 10.3390/nu11040877] [Medline: 31003547]
31. Zhu F, Bosch M, Boushey CJ, Delp EJ. An image analysis system for dietary assessment and evaluation. Proc Int Conf Image Proc 2010:1853-1856 [FREE Full text] [doi: 10.1109/ICIP.2010.5650848] [Medline: 22025261]
32. Pouladzadeh P, Shirmohammadi S, Al-Maghrabi R. Measuring calorie and nutrition from food image. IEEE Trans Instrum Meas 2014 Aug;63(8):1947-1956. [doi: 10.1109/TIM.2014.2303533]
33. Kawano Y, Yanai K. FoodCam-256: a large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights. In: Proceedings of the 22nd ACM International Conference on Multimedia. 2014 Presented at: MM '14; November 3-7, 2014; Orlando, FL. [doi: 10.1145/2647868.2654869]
34. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998 Nov;86(11):2278-2324. [doi: 10.1109/5.726791]
35. Kawano Y, Yanai K. Food image recognition with deep convolutional features. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. 2014 Presented at: UbiComp '14 Adjunct; September 13-17, 2014; Seattle, WA. [doi: 10.1145/2638728.2641339]

XSL•FO
RenderX

36. Matsuda Y, Hoashi H, Yanai K. Recognition of multiple-food images by detecting candidate regions. In: Proceedings of the IEEE International Conference on Multimedia and Expo. 2012 Presented at: ICME 2012; July 9-13, 2012; Melbourne, Australia. [doi: 10.1109/icme.2012.157]
37. Kawano Y, Yanai K. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Proceedings of the 13th European Conference on Computer Vision. 2014 Presented at: ECCV 2014; September 6-12, 2014; Zurich, Switzerland. [doi: 10.1007/978-3-319-16199-0_1]
38. Bossard L, Guillaumin M, Gool LV. Food-101 – mining discriminative components with random forests. In: Proceedings of the 13th European Conference on Computer Vision. 2014 Presented at: ECCV 2014; September 6-12, 2014; Zurich, Switzerland. [doi: 10.1007/978-3-319-10599-4_29]
39. Yanai K, Kawano Y. Food image recognition using deep convolutional network with pre-training and fine-tuning. In: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. 2015 Presented at: ICMEW 2015; June 29-July 3, 2015; Turin, Italy. [doi: 10.1109/icmew.2015.7169816]
40. Christodoulidis S, Anthimopoulos M, Mougiakakou S. Food recognition for dietary assessment using deep convolutional neural networks. In: Proceedings of the International Conference on Image Analysis and Processing. 2015 Presented at: ICIAP 2015; September 7-8, 2015; Genoa, Italy. [doi: 10.1007/978-3-319-23222-5_56]
41. Liu C, Cao Y, Luo Y, Chen G, Vokkarane V, Ma Y. DeepFood: deep learning-based food image recognition for computer-aided dietary assessment. In: Proceedings of the 14th International Conference on Smart Homes and Health Telematics. 2016 Presented at: ICOST 2016; May 25-27, 2016; Wuhan, China. [doi: 10.1007/978-3-319-39601-9_4]
42. Singla A, Yuan L, Ebrahimi T. Food/non-food image classification and food categorization using pre-trained GoogLeNet model. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management. 2016 Presented at: MADiMa '16; October 16, 2016; Amsterdam, The Netherlands. [doi: 10.1145/2986035.2986039]
43. Hassannejad H, Matrella G, Ciampolini P, De Munari I, Mordonini M, Cagnoni S. Food image recognition using very deep convolutional networks. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management. 2016 Presented at: MADiMa '16; October 16, 2016; Amsterdam, The Netherlands. [doi: 10.1145/2986035.2986042]
44. Ciocca G, Napoletano P, Schettini R. Food recognition: a new dataset, experiments, and results. IEEE J Biomed Health Inform 2017 May;21(3):588-598. [doi: 10.1109/JBHI.2016.2636441] [Medline: 28114043]
45. Mezgec S, Koroušić Seljak B. NutriNet: a deep learning food and drink image recognition system for dietary assessment. Nutrients 2017 Jun 27;9(7):657 [FREE Full text] [doi: 10.3390/nu9070657] [Medline: 28653995]
46. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
47. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014 Presented at: CVPR 2014; June 23-28, 2014; Columbus, OH. [doi: 10.1109/cvpr.2014.81]
48. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: CVPR 2016; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
49. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: CVPR 2016; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.91]
50. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. Preprint posted online on September 4, 2014 2024. [doi: 10.48550/arXiv.1409.1556]
51. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015 Presented at: CVPR 2015; June 7-12, 2015; Boston, MA. [doi: 10.1109/cvpr.2015.7298594]
52. Hoashi H, Joutou T, Yanai K. Image recognition of 85 food categories by feature fusion. In: Proceedings of the IEEE International Symposium on Multimedia. 2010 Presented at: ISM 2010; December 13-15, 2010; Taichung, Taiwan. [doi: 10.1109/ism.2010.51]
53. Kong F, Tan J. DietCam: regular shape food recognition with a camera phone. In: Proceedings of the International Conference on Body Sensor Networks. 2011 Presented at: BSN 2011; May 23-25, 2011; Dallas, TX. [doi: 10.1109/bsn.2011.19]
54. He Y, Xu C, Khanna N, Boushey CJ, Delp EJ. Analysis of food images: features and classification. Proc Int Conf Image Proc 2014 Oct;2014:2744-2748 [FREE Full text] [doi: 10.1109/ICIP.2014.7025555] [Medline: 28572748]
55. Pandey P, Deepthi A, Mandal B, Puhan NB. FoodNet: recognizing foods using ensemble of deep networks. IEEE Signal Process Lett 2017 Dec;24(12):1758-1762. [doi: 10.1109/lsp.2017.2758862]
56. Martinel N, Foresti GL, Micheloni C. Wide-slice residual networks for food recognition. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. 2018 Presented at: WACV 2018; March 12-15, 2018; Lake Tahoe, NV. [doi: 10.1109/wacv.2018.00068]
57. Jiang S, Min W, Liu L, Luo Z. Multi-scale multi-view deep feature aggregation for food recognition. IEEE Trans Image Process 2020;29:265-276. [doi: 10.1109/TIP.2019.2929447] [Medline: 31369375]

XSL•FO
RenderX

58. Lu Y, Stathopoulou T, Vasiloglou MF, Pinault LF, Kiley C, Spanakis EK, et al. goFOOD: an artificial intelligence system for dietary assessment. Sensors (Basel) 2020 Jul 31;20(15):4283 [FREE Full text] [doi: 10.3390/s20154283] [Medline: 32752007]
59. Wu MY, Lee JH, Hsueh CY. A framework of visual checkout system using convolutional neural networks for Bento buffet. Sensors (Basel) 2021 Apr 08;21(8):2627 [FREE Full text] [doi: 10.3390/s21082627] [Medline: 33918027]
60. Tola E, Lepetit V, Fua P. DAISY: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans Pattern Anal Mach Intell 2010 May;32(5):815-830. [doi: 10.1109/TPAMI.2009.77] [Medline: 20299707]
61. Galer M. Photography: Foundations for Art & Design: The Creative Photography. Waltham, MA: Focal Press; 2004.
62. Katz M. Introduction to Geometrical Optics. Singapore, Singapore: World Scientific; 2002.
63. Jia W, Yue Y, Fernstrom JD, Yao N, Sclabassi RJ, Fernstrom MH, et al. Imaged based estimation of food volume using circular referents in dietary assessment. J Food Eng 2012 Mar;109(1):76-86 [FREE Full text] [doi: 10.1016/j.jfoodeng.2011.09.031] [Medline: 22523440]
64. Okamoto K, Yanai K. An automatic calorie estimation system of food images on a smartphone. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management. 2016 Presented at: MADiMa '16; October 16, 2016; Amsterdam, The Netherlands. [doi: 10.1145/2986035.2986040]
65. Akpa EA, Suwa H, Arakawa Y, Yasumoto K. Smartphone-based food weight and calorie estimation method for effective food journaling. SICE J Control Meas Syst Integr 2021 Jan 18;10(5):360-369. [doi: 10.9746/jcmsi.10.360]
66. Liang Y, Li J. Deep learning-based food calorie estimation method in dietary assessment. arXiv. Preprint posted online on June 10, 2017 2024 [FREE Full text]
67. Ege T, Shimoda W, Yanai K. A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In: Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management. 2019 Presented at: MADiMa '19; October 21, 2019; Nice, France. [doi: 10.1145/3347448.3357162]
68. Woo I, Otsmo K, Kim S, Ebert DS, Delp EJ, Boushey CJ. Automatic portion estimation and visual refinement in mobile dietary assessment. Proc SPIE Int Soc Opt Eng 2010 Jan 01;7533:75330O [FREE Full text] [doi: 10.1117/12.849051] [Medline: 22242198]
69. Chae J, Woo I, Kim S, Maciejewski R, Zhu F, Delp EJ, et al. Volume estimation using food specific shape templates in mobile image-based dietary assessment. Proc SPIE Int Soc Opt Eng 2011 Feb 07;7873:78730K [FREE Full text] [doi: 10.1117/12.876669] [Medline: 22025936]
70. Chen HC, Jia W, Yue Y, Li Z, Sun YN, Fernstrom JD, et al. Model-based measurement of food portion size for image-based dietary assessment using 3D/2D registration. Meas Sci Technol 2013 Oct;24(10):10.1088/0957-0233/24/10/105701 [FREE Full text] [doi: 10.1088/0957-0233/24/10/105701] [Medline: 24223474]
71. Jia W, Chen HC, Yue Y, Li Z, Fernstrom J, Bai Y, et al. Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. Public Health Nutr 2014 Aug;17(8):1671-1681 [FREE Full text] [doi: 10.1017/S1368980013003236] [Medline: 24476848]
72. Tanno R, Ege T, Yanai K. AR DeepCalorieCam V2: food calorie estimation with CNN and AR-based actual size estimation. In: Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. 2018 Presented at: VRST '18; November 28-December 1, 2018; Tokyo Japan. [doi: 10.1145/3281505.3281580]
73. Yang Y, Jia W, Bucher T, Zhang H, Sun M. Image-based food portion size estimation using a smartphone without a fiducial marker. Public Health Nutr 2019 May;22(7):1180-1192 [FREE Full text] [doi: 10.1017/S136898001800054X] [Medline: 29623867]
74. Smith SP, Adam MT, Manning G, Burrows T, Collins C, Rollo ME. Food volume estimation by integrating 3D image projection and manual wire mesh transformations. IEEE Access 2022 May 02;10:48367-48378. [doi: 10.1109/ACCESS.2022.3171584]
75. Kong F, Tan J. DietCam: automatic dietary assessment with mobile camera phones. Pervasive Mob Comput 2012 Feb;8(1):147-163. [doi: 10.1016/j.pmcj.2011.07.003]
76. Rahman MH, Li Q, Pickering M, Frater M, Kerr D, Bouchey C, et al. Food volume estimation in a mobile phone based dietary assessment system. In: Proceedings of the Eighth International Conference on Signal Image Technology and Internet Based Systems. 2012 Presented at: SITIS 2012; November 25-29, 2012; Sorrento, Italy. [doi: 10.1109/sitis.2012.146]
77. Chang X, Ye H, Albert P, Edward D, Nitin K, Carol B. Image-based food volume estimation. CEA13 (2013) 2013 Oct;2013:75-80 [FREE Full text] [doi: 10.1145/2506023.2506037] [Medline: 28573255]
78. Anthimopoulos M, Dehais J, Shevchik S, Ransford BH, Duke D, Diem P, et al. Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones. J Diabetes Sci Technol 2015 May;9(3):507-515 [FREE Full text] [doi: 10.1177/1932296815580159] [Medline: 25883163]
79. Dehais J, Anthimopoulos M, Shevchik S, Mougiakakou S. Two-view 3D reconstruction for food volume estimation. IEEE Trans Multimedia 2017 May;19(5):1090-1099. [doi: 10.1109/tmm.2016.2642792]
80. Gao A, Lo FP, Lo B. Food volume estimation for quantifying dietary intake with a wearable camera. In: Proceedings of the IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks. 2018 Presented at: BSN 2018; March 4-7, 2018; Las Vegas, NV. [doi: 10.1109/bsn.2018.8329671]

XSL•FO
RenderX

81. Ando Y, Ege T, Cho J, Yanai K. DepthCalorieCam: a mobile application for volume-based food calorie estimation using depth cameras. In: Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management. 2019 Presented at: MADiMa '19; October 21, 2019; Nice, France. [doi: 10.1145/3347448.3357172]
82. Shang J, Duong M, Pepin E, Zhang X, Sandara-Rajan K, Mamishev A, et al. A mobile structured light system for food volume estimation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. 2011 Presented at: ICCVW 2011; November 6-13, 2011; Barcelona, Spain. [doi: 10.1109/iccvw.2011.6130229]
83. Chen MY, Yang YH, Ho CJ, Wang SH, Liu SM, Chang E, et al. Automatic Chinese food identification and quantity estimation. In: Proceedings of the SIGGRAPH Asia 2012 Technical Briefs. 2012 Presented at: SA '12; November 28-December 1, 2012; Singapore, Singapore. [doi: 10.1145/2407746.2407775]
84. Fang S, Zhu F, Jiang C, Zhang S, Boushey CJ, Delp EJ. A comparison of food portion size estimation using geometric models and depth images. In: Proceedings of the IEEE International Conference on Image Processing. 2016 Presented at: ICIP 2016; September 25-28, 2016; Phoenix, AZ. [doi: 10.1109/icip.2016.7532312]
85. Zhang S. Flexible 3D shape measurement using projector defocusing: extended measurement range. Opt Lett 2010 Apr 01;35(7):934-936 [FREE Full text] [doi: 10.1364/OL.35.000934] [Medline: 20364174]
86. Alfonsi JE, Choi EE, Arshad T, Sammott SA, Pais V, Nguyen C, et al. Carbohydrate counting app using image recognition for youth with type 1 diabetes: pilot randomized control trial. JMIR Mhealth Uhealth 2020 Oct 28;8(10):e22074 [FREE Full text] [doi: 10.2196/22074] [Medline: 33112249]
87. Herzig D, Nakas CT, Stalder J, Kosinski C, Laesser C, Dehais J, et al. Volumetric food quantification using computer vision on a depth-sensing smartphone: preclinical study. JMIR Mhealth Uhealth 2020 Mar 25;8(3):e15294 [FREE Full text] [doi: 10.2196/15294] [Medline: 32209531]
88. Zhang W, Yu Q, Siddiquie B, Divakaran A, Sawhney H. "Snap-n-eat": food recognition and nutrition estimation on a smartphone. J Diabetes Sci Technol 2015 May;9(3):525-533 [FREE Full text] [doi: 10.1177/1932296815582222] [Medline: 25901024]
89. He Y, Xu C, Khanna N, Boushey CJ, Delp EJ. Food image analysis: segmentation, identification and weight estimation. In: Proceedings of the IEEE International Conference on Multimedia and Expo. 2013 Presented at: ICME 2013; July 15-19, 2013; San Jose, CA. [doi: 10.1109/icme.2013.6607548]
90. Yue Y, Jia W, Sun M. Measurement of food volume based on single 2-D image without conventional camera calibration. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:2166-2169 [FREE Full text] [doi: 10.1109/EMBC.2012.6346390] [Medline: 23366351]
91. Pouladzadeh P, Villalobos G, Almaghrabi R, Shirmohammadi S. A novel SVM based food recognition method for calorie measurement applications. In: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops. 2012 Presented at: ICMEW 2012; July 9-13, 2012; Melbourne, Australia. [doi: 10.1109/icmew.2012.92]
92. Vasiloglou MF, Mougiakakou S, Aubry E, Bokelmann A, Fricker R, Gomes F, et al. A comparative study on carbohydrate estimation: GoCARB vs. dietitians. Nutrients 2018 Jun 07;10(6):741 [FREE Full text] [doi: 10.3390/nu10060741] [Medline: 29880772]
93. Samsung Galaxy A24, A34, and A54 to launch without depth sensing cameras. GSMArena. 2022 Jul 21. URL: https://www.gsmarena.com/samsung_is_going_to_remove_useless_depth_sensing_cameras_from_its_a_series_devices_next_year-news-55135.php [accessed 2023-04-09]
94. Li H, Han T. DeepVol: deep fruit volume estimation. In: Proceedings of the 27th International Conference on Artificial Neural Networks. 2018 Presented at: ICANN 2018; October 4-7, 2018; Rhodes, Greece. [doi: 10.1007/978-3-030-01424-7_33]
95. Jiang Y, Schenck E, Kranz S, Banerjee S, Banerjee NK. CNN-based non-contact detection of food level in bottles from RGB images. In: Proceedings of the 25th International Conference on MultiMedia Modeling. 2019 Presented at: MMM 2019; January 8-11, 2019; Thessaloniki, Greece. [doi: 10.1007/978-3-030-05710-7_17]
96. Lo FP, Sun Y, Qiu J, Lo BP. Point2Volume: a vision-based dietary assessment approach using view synthesis. IEEE Trans Ind Inform 2020 Jan;16(1):577-586. [doi: 10.1109/TII.2019.2942831]
97. Lo FP, Sun Y, Qiu J, Lo B. Food volume estimation based on deep learning view synthesis from a single depth map. Nutrients 2018 Dec 18;10(12):2005 [FREE Full text] [doi: 10.3390/nu10122005] [Medline: 30567362]
98. Lo FP, Sun Y, Qiu J, Lo B. A novel vision-based approach for dietary assessment using deep learning view synthesis. In: Proceedings of the IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks. 2019 Presented at: BSN 2019; May 19-22, 2019; Chicago, IL. [doi: 10.1109/bsn.2019.8771089]
99. Yang Z, Yu H, Cao S, Xu Q, Yuan D, Zhang H, et al. Human-mimetic estimation of food volume from a single-view RGB image using an AI system. Electronics (Basel) 2021 Jul 28;10(13):1556 [FREE Full text] [doi: 10.3390/electronics10131556] [Medline: 34552763]
100. Miyazaki T, de Silva GC, Aizawa K. Image-based calorie content estimation for dietary assessment. In: Proceedings of the IEEE International Symposium on Multimedia. 2011 Presented at: ISM 2011; December 5-7, 2011; Dana Point, CA. [doi: 10.1109/ism.2011.66]

XSL•FO
RenderX

101. Ege T, Yanai K. Multi-task learning of dish detection and calorie estimation. In: Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. 2018 Presented at: CEA/MADiMa '18; July 15, 2018; Stockholm, Sweden. [doi: 10.1145/3230519.3230594]
102. Ege T, Yanai K. Simultaneous estimation of dish locations and calories with multi-task learning. IEICE Trans Inf Syst 2019;E102.D(7):1240-1246. [doi: 10.1587/transinf.2018cep0004]
103. Lu Y, Allegra D, Anthimopoulos M, Stanco F, Farinella GM, Mougiakakou S. A multi-task learning approach for meal assessment. In: Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. 2018 Presented at: CEA/MADiMa '18; July 15, 2018; Stockholm, Sweden. [doi: 10.1145/3230519.3230593]
104. He J, Shao Z, Wright J, Kerr D, Boushey C, Zhu F. Multi-task image-based dietary assessment for food recognition and portion size estimation. In: Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval. 2020 Presented at: MIPR 2020; August 6-8, 2020; Shenzhen, China. [doi: 10.1109/mipr49039.2020.00018]
105. Thames Q, Karpur A, Norris W, Xia F, Panait L, Weyand T, et al. Nutrition5k: towards automatic nutritional understanding of generic food. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021 Presented at: CVPR 2021; June 20-25, 2021; Nashville, TN. [doi: 10.1109/cvpr46437.2021.00879]
106. Ruede R, Heusser V, Frank L, Roitberg A, Haurilet M, Stiefelhagen R. Multi-task learning for calorie prediction on a novel large-scale recipe dataset enriched with nutritional information. In: Proceedings of the 25th International Conference on Pattern Recognition. 2020 Presented at: ICPR 2020; January 10-15, 2021; Virtual Event. [doi: 10.1109/icpr48806.2021.9412839]
107. Lu Y, Stathopoulou T, Mougiakakou S. Partially supervised multi-task network for single-view dietary assessment. In: Proceedings of the 25th International Conference on Pattern Recognition. 2020 Presented at: ICPR 2020; July 10-15, 2021; Virtual Event. [doi: 10.1109/icpr48806.2021.9412339]
108. He J, Mao R, Shao Z, Wright JL, Kerr DA, Boushey CJ, et al. An end-to-end food image analysis system. Electron Imaging 2021 Jan;33:285-1-285-7. [doi: 10.2352/issn.2470-1173.2021.8.imawm-285]
109. Situju SF, Takimoto H, Sato S, Yamauchi H, Kanagawa A, Lawi A. Food constituent estimation for lifestyle disease prevention by multi-task CNN. Appl Artif Intell 2019 Apr 23;33(8):732-746. [doi: 10.1080/08839514.2019.1602318]
110. Boushey CJ, Spoden M, Delp EJ, Zhu F, Bosch M, Ahmad Z, et al. Reported energy intake accuracy compared to doubly labeled water and usability of the mobile food record among community dwelling adults. Nutrients 2017 Mar 22;9(3):312 [FREE Full text] [doi: 10.3390/nu9030312] [Medline: 28327502]
111. Open Food Facts. URL: https://world.openfoodfacts.org [accessed 2024-06-13]
112. Partridge EK, Neuhouser ML, Breymeyer K, Schenk JM. Comparison of nutrient estimates based on food volume versus weight: implications for dietary assessment methods. Nutrients 2018 Jul 27;10(8):973 [FREE Full text] [doi: 10.3390/nu10080973] [Medline: 30060455]
113. Conway R, Robertson C, Dennis B, Stamler J, Elliott P, INTERMAP Research Group. Standardised coding of diet records: experiences from INTERMAP UK. Br J Nutr 2004 May;91(5):765-771 [FREE Full text] [doi: 10.1079/BJN20041095] [Medline: 15152639]
114. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC. [doi: 10.1145/3173574.3174156]
115. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
116. Chang EM, Saigal CS, Raldow AC. Explaining health state utility assessment. JAMA 2020 Mar 17;323(11):1085-1086. [doi: 10.1001/jama.2020.0656] [Medline: 32091541]
117. Gemming L, Utter J, Ni Mhurchu C. Image-assisted dietary assessment: a systematic review of the evidence. J Acad Nutr Diet 2015 Jan;115(1):64-77. [doi: 10.1016/j.jand.2014.09.015] [Medline: 25441955]
118. Doulah A, McCrory MA, Higgins JA, Sazonov E. A systematic review of technology-driven methodologies for estimation of energy intake. IEEE Access 2019;7:49653-49668 [FREE Full text] [doi: 10.1109/access.2019.2910308] [Medline: 32489752]
119. Lo FP, Sun Y, Qiu J, Lo B. Image-based food classification and volume estimation for dietary assessment: a review. IEEE J Biomed Health Inform 2020 Jul;24(7):1926-1939. [doi: 10.1109/jbhi.2020.2987943]
120. Subhi MA, Ali SH, Mohammed MA. Vision-based approaches for automatic food recognition and dietary assessment: a survey. IEEE Access 2019 Mar 13;7:35370-35381. [doi: 10.1109/access.2019.2904519]
121. Dalakleidi KV, Papadelli M, Kapolos I, Papadimitriou K. Applying image-based food-recognition systems on dietary assessment: a systematic review. Adv Nutr 2022 Dec 22;13(6):2590-2619 [FREE Full text] [doi: 10.1093/advances/nmac078] [Medline: 35803496]
122. Tay W, Kaur B, Quek R, Lim J, Henry CJ. Current developments in digital quantitative volume estimation for the optimisation of dietary assessment. Nutrients 2020 Apr 22;12(4):1167 [FREE Full text] [doi: 10.3390/nu12041167] [Medline: 32331262]

XSL•FO
RenderX

Abbreviations

24HR: 24-hour dietary recall
AI: artificial intelligence
AR: augmented reality
CNN: convolutional neural network
EFR: estimated food record
IADA: image-assisted dietary assessment
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
ResNet: residual network
RGB: red-green-blue (color model based on additive color primaries)
RGBD: red-green-blue with depth
SVM: support vector machine
TADA: Technology Assisted Dietary Assessment
ToF: time-of-flight
UEC: University of Electro-Communications

Edited by A Coristine; submitted 31.07.23; peer-reviewed by W Jia, K DiFilippo; comments to author 15.02.24; revised version received 13.06.24; accepted 24.09.24; published 15.11.24.

Please cite as:
Chotwanvirat P, Prachansuwan A, Sridonpai P, Kriengsinyos W
Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review
J Med Internet Res 2024;26:e51432
URL: https://www.jmir.org/2024/1/e51432
doi:10.2196/51432
PMID:

©Phawinpon Chotwanvirat, Aree Prachansuwan, Pimnapanut Sridonpai, Wantanee Kriengsinyos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Examining the Role of Large Language Models in Orthopedics: Systematic Review

Cheng Zhang[1,2,3*], MS; Shanshan Liu[1,2,3*], MD; Xingyu Zhou[4], BS; Siyu Zhou[1,2,3], MD; Yinglun Tian[1,2,3], MD; Shenglin Wang[1,2,3], MD; Nanfang Xu[1,2,3*], MD, PhD; Weishi Li[1,2,3*], MD

[1]Department of Orthopaedics, Peking University Third Hospital, Beijing, China

[2]Engineering Research Center of Bone and Joint Precision Medicine, Ministry of Education, Beijing, China

[3]Beijing Key Laboratory of Spinal Disease Research, Beijing, China

[4]Peking University Health Science Center, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Weishi Li, MD
Department of Orthopaedics
Peking University Third Hospital
49 North Garden Road
Beijing, 100191
China
Phone: 86 01082267360
Email: puh3liweishi@163.com

Abstract

Background: Large language models (LLMs) can understand natural language and generate corresponding text, images, and even videos based on prompts, which holds great potential in medical scenarios. Orthopedics is a significant branch of medicine, and orthopedic diseases contribute to a significant socioeconomic burden, which could be alleviated by the application of LLMs. Several pioneers in orthopedics have conducted research on LLMs across various subspecialties to explore their performance in addressing different issues. However, there are currently few reviews and summaries of these studies, and a systematic summary of existing research is absent.

Objective: The objective of this review was to comprehensively summarize research findings on the application of LLMs in the field of orthopedics and explore the potential opportunities and challenges.

Methods: PubMed, Embase, and Cochrane Library databases were searched from January 1, 2014, to February 22, 2024, with the language limited to English. The terms, which included variants of "large language model," "generative artificial intelligence," "ChatGPT," and "orthopaedics," were divided into 2 categories: *large language model* and *orthopedics*. After completing the search, the study selection process was conducted according to the inclusion and exclusion criteria. The quality of the included studies was assessed using the revised Cochrane risk-of-bias tool for randomized trials and CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidance. Data extraction and synthesis were conducted after the quality assessment.

Results: A total of 68 studies were selected. The application of LLMs in orthopedics involved the fields of clinical practice, education, research, and management. Of these 68 studies, 47 (69%) focused on clinical practice, 12 (18%) addressed orthopedic education, 8 (12%) were related to scientific research, and 1 (1%) pertained to the field of management. Of the 68 studies, only 8 (12%) recruited patients, and only 1 (1%) was a high-quality randomized controlled trial. ChatGPT was the most commonly mentioned LLM tool. There was considerable heterogeneity in the definition, measurement, and evaluation of the LLMs' performance across the different studies. For diagnostic tasks alone, the accuracy ranged from 55% to 93%. When performing disease classification tasks, ChatGPT with GPT-4's accuracy ranged from 2% to 100%. With regard to answering questions in orthopedic examinations, the scores ranged from 45% to 73.6% due to differences in models and test selections.

Conclusions: LLMs cannot replace orthopedic professionals in the short term. However, using LLMs as copilots could be a potential approach to effectively enhance work efficiency at present. More high-quality clinical trials are needed in the future, aiming to identify optimal applications of LLMs and advance orthopedics toward higher efficiency and precision.

(J Med Internet Res 2024;26:e59607) doi:10.2196/59607

XSL•FO
RenderX

KEYWORDS

large language model; LLM; orthopedics; generative pretrained transformer; GPT; ChatGPT; digital health; clinical practice; artificial intelligence; AI; generative AI; Bard

Introduction

Background

Large language models (LLMs) typically refer to pretrained language models (PLMs) that have a large number of parameters and are trained on massive amounts of data. In recent years, this area has emerged as one of the most prominent areas of research in artificial intelligence (AI) innovation [1,2]. What makes LLMs different from smaller-scale PLMs is their remarkable emergent abilities to solve complex tasks. Studies have found that LLMs, such as generative pretrained transformer (GPT)-3 with approximately 175 billion parameters, exhibit a significant leap in natural language processing (NLP) capabilities compared to PLMs with fewer parameters, such as GPT-2 with approximately 1.5 billion parameters [2,3]. Generative AI applications developed based on LLMs not only possess the ability to understand natural language but can also generate corresponding text, images, and even videos based on input sources. This human-machine interaction mode holds great potential in medical scenarios.

LLMs have undergone significant advancements in recent years; currently, the most prevalent web-based LLM service is ChatGPT (OpenAI). Launched in November 2022, ChatGPT is a chatbot application developed based on GPT-3.5 or GPT-4 after fine-tuning, and it can quickly respond to questions posed by users. In addition to ChatGPT, applications include Bard (upgraded to Gemini in December 2023) based on Language Model for Dialogue Applications (Google LLC); Med-PaLM 2 (Google LLC); ERNIE Bot (Baidu); and MOSS (Fudan University). GPT-4 can approach or achieve human-level performance in cognitive tasks across various fields, including medical domains [4]. When answering the 2022 United States Medical Licensing Examination questions, without further training or reinforcement, ChatGPT reached or approached a passing level in all 3 examinations [5]. However, answering examination questions does not directly reflect the performance of LLMs in clinical applications. The value and safety of a chatbot that is already in use are still not fully understood, making clinical research both essential and imperative. Published narrative reviews and editorials have explored the medical applications of LLM technology from 3 perspectives: clinical practice, education, and research [1,6-8]. These publications also provide a preliminary assessment of the value and safety of LLMs, offering guidance for exploring their use in specialized medical fields.

Orthopedics is a significant branch of medicine, typically encompassing disciplines such as trauma, spine surgery, joint surgery, sports medicine, hand surgery, and bone oncology. Orthopedic diseases have a broad impact on populations and pose a major global health threat. Low back pain, a common symptom in orthopedics or spine surgery, has been identified as the leading cause of global productivity loss, as measured in years, according to a large-scale epidemiological study covering 195 countries and regions; in 126 countries, low back pain ranks first among the causes of years lived with disability [9]. In traditional health care systems, the annual medical expenditure for low back pain in the United States is estimated to exceed US $100 billion, contributing to a significant socioeconomic burden [10]. Similarly, osteoarthritis is also a critical global health issue. The global prevalence of knee osteoarthritis in adults aged >40 years is 23%, with approximately 61% of adults aged >45 years showing radiographic evidence of knee osteoarthritis [11]. Therefore, applying LLMs in orthopedics holds the potential to alleviate the current heavy socioeconomic burden.

It is worth noting that several pioneers in orthopedics have conducted studies on LLMs across various subspecialties to explore their performance in addressing different issues. However, there are currently few reviews and summaries of these studies. The published reviews primarily focus on introducing and popularizing the basic concepts of LLMs in orthopedics [12,13], or they offer forward-looking perspectives by categorizing LLM applications in clinical practice, education, and research [14]. A systematic summary of existing research is absent. To the best of our knowledge, this review is the first to systematically summarize existing research findings. In contrast to prior works, we place greater emphasis on the quantitative evaluation methods and results of these studies because we believe that these methods and outcomes can help orthopedic and computer science researchers better understand the current state of LLM research and the performance of LLMs. Regarding application categorization, we consider tasks involving NLP in management as another important application area for LLMs in orthopedics. Therefore, this review adds a category for orthopedic management applications to the existing classification framework.

Objectives

The objective of this review was to comprehensively summarize the research findings on the application of LLMs in orthopedics and outline the advantages, limitations, and methodological evaluations, while also exploring the potential opportunities and challenges emerging in this era, for facilitating interdisciplinary collaboration and advancement among researchers in computer science and orthopedics. The ultimate goal is to contribute to improved efficiency and quality of orthopedic care as well as a reduction in medical costs and the associated socioeconomic burden.

Methods

Search Strategy

The protocol for this systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (checklist can be found in the Multimedia Appendix 1) [15]. PubMed, Embase, and Cochrane Library databases were searched, with the language limited to

XSL•FO
RenderX

English. The time frame was set from January 1, 2014, to February 22, 2024. Search terms were divided into 2 categories, with the first category including LLM-related terms and the second containing words related to orthopedics and its subspecialties (Textbox 1). Terms within each category were connected using "OR," while terms within different categories were connected using "AND." The full search strategy can be found in Multimedia Appendix 2.

Textbox 1. Categories and terms applied in the search queries.

Category 1

- "large language model," "LLM," "generative artificial intelligence," "generative AI," "ChatGPT," and "Generative Pre-Trained Transformer"

Category 2

- "orthopedics," "bone," "musculoskeletal," "injury," "wound," "trauma," "articular," "joint," "sports medicine," "hand surgery," "spine," "spinal, "cervical vertebrae," "thoracic vertebrae," "lumbar vertebrae," "sacrum," "coccyx," "spinal canal," "vertebral body," and "intervertebral disc"

Study Selection

The records were downloaded from the databases and imported into EndNote (version 21.2; Clarivate) for article management. The study selection process was conducted independently by 2 investigators (CZ and SL). The inclusion and exclusion criteria are listed in Textbox 2. The results were cross-checked, and discrepancies were resolved through discussion, with the final determination made by a third investigator (YT).

Textbox 2. Inclusion and exclusion criteria.

Inclusion criteria

- Article type
  - Original research
- Language
  - Articles written in English
- Content
  - Studies that use at least 1 large language model (LLM)
  - Studies that are relevant to the field of orthopedics

Exclusion criteria

- Article type
  - Reviews, editorials, letters, and study protocols
- Language
  - Articles written in a language other than English
- Content
  - Studies that do not involve LLMs
  - Studies that use LLMs for tasks such as code generation, debugging, or text generation without any performance evaluation of the model

Quality Assessment of Studies

Quality assessment was conducted by 2 investigators (CZ and SL) independently. First, the study designs were identified. Studies that involved only posing questions to LLMs, did not recruit participants, and did not report a study design were classified as surveys. Given the diverse nature of the survey types included in the review, quality assessments were conducted only for studies that recruited participants. The revised Cochrane risk-of-bias tool for randomized trials [16] was used to assess randomized controlled trials (RCTs), and the CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidance [17] was used to evaluate prospective or retrospective observational studies. The revised Cochrane risk-of-bias tool (version of August 22, 2019) is designed for assessing RCTs and contains 5 domains: bias arising from the randomization process, bias due to deviations from the intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and bias in selection of the reported result. The CONSORT-AI guidance is a new reporting guideline specifically designed for clinical trials that assess interventions with an AI component. The quality assessment domains under this guidance include a statement of the AI algorithm used, details of how the AI intervention fits

XSL•FO
RenderX

within the clinical pathway, inclusion and exclusion criteria for input data, a description of the approaches used to handle unavailable input data, a description of the input data acquisition process for the AI intervention, specifications of human-AI interaction in the collection of input data, the output of the AI algorithm, and explanations of how the AI intervention's outputs contribute to health behavior changes. The results were cross-checked, and discrepancies were resolved through discussion, with the final determination made by another investigator (YT).

Data Extraction and Synthesis

The studies were categorized into 4 groups based on their application areas: clinical practice, education, research, and management. Data extraction and synthesis were conducted by 2 investigators (CZ and SL) independently. In addition to general characteristics, the composition of extracted data varied depending on the specific category. Details of the data extraction strategy can be found in Multimedia Appendix 3. In cases where there were inconsistencies in the process, a third investigator (XZ) participated in the discussion and made the final decision. For studies with high heterogeneity, we did not synthesize the parameters for model performance evaluation and instead focused on providing a descriptive analysis of the data. Microsoft Excel 2019 was used for data collection, analysis, and visualization.

Results

Characteristics of Included Studies

A total of 829 studies were identified; after removing duplicates and screening, 68 (8.2%) studies were selected in the literature review. The inclusion process is shown in Figure 1.

Figure 1. Flowchart of literature screening based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. LLM: large language model.

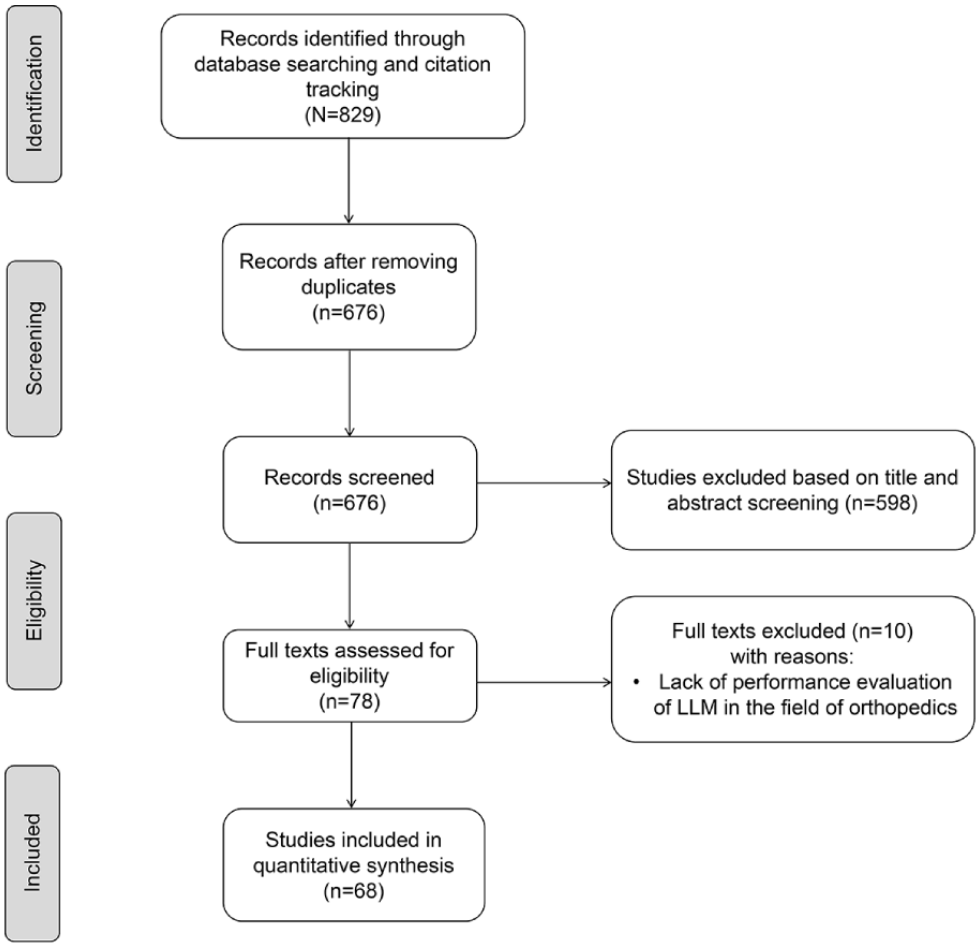

The application of LLMs in orthopedics involves the fields of clinical practice, education, research, and management. Of the 68 included studies, 47 (69%) focused on clinical practice (Table 1) [18-64], 12 (18%) addressed orthopedic education (Table 2) [65-76], 8 (12%) were related to scientific research (Table 3) [77-84], and 1 (1%) pertained to the field of management (Table 3) [85]. Of the 68 studies, 55 (81%) were classified as surveys; furthermore, only 8 (12%) recruited patients, only 1 (1%) was a high-quality study (RCT), and only 1 (1%) was a prospective study. Since June 2023, research on the application of LLMs in orthopedics has increased month by month (Figure 2).

XSL•FO
RenderX

Table 1. Characteristics of the included studies focused on clinical practice.

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Agharia et al [18], 2024	Survey	Formulate clinical decisions	GPT-3.5; GPT-4; Bard	Proportion of most popular response: 68% (GPT-4); 40.2% (GPT-3.5); 45.4% (Bard)	__[b]	Subjective
Anastasio et al [19], 2023	Survey	Generate answers to clinical questions	GPT-3.5	Ratio of responses in different quality grades: bottom-tier rating 4.5%; middle-tier rating 27.3%; top-tier rating 68.2%	—	Subjective
Baker et al [20], 2024	RCT[c]	Assist with writing patient histories	GPT-4	Mean time: 69.8 (SD 26.2) s; mean word count: 135.8 (SD 40.3); mean PDQI-9[d] score: 35.6 (SD 3.1); mean overall rating: 3.8 (SD 0.6); ratio of erroneous documents: 36%	11	Subjective
Christy et al [21], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Ratio of appropriate responses in total responses: 78%; intraclass correlation coefficient: 0.12	—	Subjective
Coraci et al [22], 2023	Cross-sectional study	Create questionnaire for assessment	GPT-3.5	Correlation: acceptable correlation with ODI[e] and QBPDS[f]; no statistical correlation with RMDQ[g] or NRS[h]	20	Subjective
Crook et al [23], 2023	Survey	Generate answers to clinical questions	GPT-3	DISCERN score: 58; JAMA[i] benchmark score: 0/4; FRE[j] score: 34; FKGL[k] score: 15	—	Subjective
Daher et al [24], 2023	Prospective study	Diagnose and manage patients	GPT-3	Accuracy of diagnosis: 93%; accuracy of management: 83%	29	Objective
Decker et al [25], 2023	Cross-sectional study	Generate informed consent documentation	GPT-3.5	Mean readability, accuracy, and completeness scores (surgeons vs LLMs): readability= 15.7 vs 12.9; risks=1.7 vs 1.7; benefits=1.4 vs 2.3; alternatives=1.4 vs 2.7; overall impression=1.9 vs 2.3; composite: 1.6 vs 2.2	—	Subjective
Draschl et al [26], 2023	Survey	Generate answers to clinical questions	GPT-3.5	5-point Likert scores, mean: completeness=3.80 (SD 0.63); misleading=4.04 (SD 0.67); errors=4.14 (SD 0.58); up-to-dateness=3.90 (SD 0.45); suitability for patients=3.69 (SD 0.64); suitability for surgeons=3.63 (SD 0.95)	—	Subjective
Dubin et al [27], 2023	Survey	Generate answers to clinical questions	GPT-3	25% of the questions were similar when performing a Google web search and a search of ChatGPT for all search terms; 75% of the questions were answered by government websites; 55% of the answers were different between Google web search and ChatGPT in terms of numerical questions	—	Subjective
Duey et al [28], 2023	Survey	Generate answers to clinical questions	GPT-3.5; GPT-4	Accuracy: 33% (GPT-3.5); 92% (GPT-4)	—	Subjective
Fabijan et al [29], 2023	Cross-sectional study	Classify cases of single-curve scoliosis	GPT-4; Microsoft Bing with GPT[l]; Scholar AI Premium	GPT-4 and Scholar AI Premium excelled in classifying single-curve scoliosis with perfect sensitivity (100%) and specificity (100%)	56	Objective
Fahy et al [30], 2024	Survey	Generate answers to clinical questions	GPT-3.5; GPT-4	GPT-3.5 vs GPT-4 mean DISCERN score: 55.4 vs 62.09; mean reading grade level score: 18.08 vs 17.90	—	Subjective

XSL•FO
RenderX

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Gianola et al [31], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Internal consistency: 49%; accuracy: 33%	—	Subjective
Hurley et al [32], 2024	Survey	Generate answers to clinical questions	ChatGPT	DISCERN score: 60; JAMA benchmark score: 0; FRE score: 26.2; FKGL score: considered to be that of a college graduate	—	Subjective
Johns et al [33], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Satisfaction rate: 60%	—	Subjective
Johns et al [34], 2024	Survey	Generate answers to clinical questions	GPT-3.5	DISCERN score: 41; FKGL score: 13.4; satisfaction rate: 40%	—	Subjective
Kaarre et al [35], 2023	Survey	Generate answers to clinical questions	GPT-4	Average correctness of responses for patients and physicians: 1.69 and 1.66, respectively (on a scale ranging from 0=incorrect, 1=partially correct, and 2=correct)	—	Subjective
Kasthuri et al [36], 2024	Survey	Generate answers to clinical questions	Microsoft Bing with GPT-4	Mean completeness score: 2.03; mean accuracy score: 4.49	—	Subjective
Kienzle et al [37], 2024	Survey	Generate answers to clinical questions	GPT-4	Mean DISCERN score in overall quality: 3.675	—	Subjective
Kirchner et al [38], 2023	Survey	Rewrite patient education materials	GPT-3.5	Mean FKGL score in patient education materials related to herniated lumbar disk, scoliosis, stenosis, TKA[m], and THA[n]: before rewrite=9.5, 12.6, 10.9, 12.0, and 6.3, respectively; after rewrite=5.0, 5.6, 6.9, 11.6, and 6.1, respectively	—	Subjective
Kuroiwa et al [39], 2023	Survey	Generate answers to clinical questions	GPT-3.5	Ratios of correct answers: 25/25, 1/25, 24/25, 16/25, and 17/25 for carpal tunnel syndrome, cervical myelopathy, lumbar spinal stenosis, knee osteoarthritis, and hip osteoarthritis, respectively	—	Objective
Li et al [40], 2023	Survey	Generate answers to clinical questions	GPT-4	Mean accuracy score (out of 5): 4.3; mean completeness score (out of 3): 2.8	—	Subjective
Li et al [41], 2024	Survey	Generate answers to clinical questions	GPT-3.5	1 response was excellent, requiring no clarification; 4 responses were satisfactory, requiring minimal clarification; 3 responses were satisfactory, requiring moderate clarification; 2 responses were unsatisfactory	—	Subjective
Lower et al [42], 2023	Survey	Deliver safe and coherent medical advice	GPT-4	Mean Likert scale score: 3.2	—	Subjective
Magruder et al [43], 2024	Survey	Generate answers to clinical questions	ChatGPT	Answer grades (from 1 to 5), mean: relevance=4.43 (SD 0.77); clarity=4.22 (SD 0.86); accuracy=4.10 (SD 0.90); evidence based=3.92 (SD 1.01); completeness=3.91 (SD 0.88); consistency=3.54 (SD 1.10)	—	Subjective

XSL•FO
RenderX

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Mika et al [44], 2023	Survey	Generate answers to clinical questions	GPT-3.5	2 responses were excellent, requiring no clarification; 4 responses were satisfactory, requiring minimal clarification; 3 responses were satisfactory, requiring moderate clarification; 1 response was unsatisfactory	—	Subjective
Pagano et al [45], 2023	Retrospective observational study	Formulate diagnosis and potential treatment suggestions	GPT-4	Diagnostic accuracy: 100% for the total cases; concordance in therapeutic recommendations: 83% for the total cases	100	Objective
Mejia et al [46], 2024	Survey	Generate answers to clinical questions	GPT-3.5; GPT-4	Accuracy: 52% (GPT-3.5); 59% (GPT-4); overconclusiveness: 48% (GPT-3.5); 45% (GPT-4)	—	Subjective
Russe et al [47], 2023	Retrospective observational study	Provide accurate fracture classification based on radiology reports	FraCChat; GPT-3.5-Turbo; GPT-4	Accuracy: GPT 3.5=3%; GPT 4=2%; FraCChat 3.5=48%; FraCChat 4=71%	—	Objective
Schonfeld et al [48], 2024	Retrospective cohort study	Predict outcome of adult spinal deformities	Gatortron	AUC[o] scores: 0.565 (pulmonary complication); 0.559 (neurological complication); 0.557 (sepsis); 0.508 (delirium); F_1-scores: 0.545 (pulmonary complication); 0.250 (neurological complication); 0.383 (sepsis); 0.156 (delirium)	209	Objective
Seth et al [49], 2023	Survey	Generate answers to clinical questions	ChatGPT	Mean Likert scale score: 3.1	—	Subjective
Shrestha et al [50], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Accuracy: 44%-65% for different guideline variations	—	Subjective
Sosa et al [51], 2024	Survey	Generate answers to clinical questions	GPT-4; Bard; Bing AI	Ratios of appropriate answers to questions related to bone physiology: 83.3% (GPT-4); 23.3% (Bing AI); 16.7% (Bard)	—	Subjective
Stroop et al [52], 2023	Survey	Generate answers to clinical questions	ChatGPT	Ratio of medically complete correct answers: 52%; ratio of medically complete and comprehensive answers: 55%	—	Subjective
Suthar et al [53], 2023	Retrospective observational study	Generate diagnosis	GPT-4	Accuracy rate in spine cases: 55%	—	Objective
Taylor et al [54], 2024	Survey	Generate answers to clinical questions	ChatGPT	Ratio of surgeons who reported that the questions had been appropriately answered: 91%	—	Subjective
Temel et al [55], 2024	Survey	Generate answers to clinical questions	GPT-4	Ensuring Quality Information for Patients score: mean 43.02 (SD 6.37); FRE score: mean 26.24 (SD 13.81); FKGL score: mean 14.84 (SD 1.79)	—	Subjective
Tharakan et al [56], 2024	Survey	Generate answers to clinical questions	GPT-3	Answers provided by ChatGPT cited more academic references than those provided by a Google search (80% vs 50%)	—	Subjective
Truhn et al [57], 2023	Retrospective observational study	Prioritize treatment recommendations	GPT-4	The overall quality of the treatment recommendations was rated as good or better	20	Subjective

XSL•FO
RenderX

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Warren et al [58], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Answers to fact, policy, and value questions (mean scores): DISCERN=51, 53, and 55, respectively; JAMA benchmark=0, 0, and 0, respectively; FRE=48.3, 42.0, and 38.4, respectively; FKGL=10.3, 10.9, and 11.6, respectively	—	Subjective
Wilhelm et al [59], 2023	Survey	Generate treatment recommendations	Claude-instant-v1.0; GPT 3.5-Turbo; Command-xlarge-nightly; Bloomz	Mean DISCERN quality scores: 3.4 (Claude-instant-v1.0); 2.8 (GPT 3.5-Turbo); 2.2 (Command-xlarge-nightly); 1.1 (Bloomz)	—	Subjective
Wright et al [60], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Mean accuracy score: 4.26; mean comprehensiveness score: 3.79	—	Subjective
Yang et al [61], 2024	Retrospective observational study	Generate diagnosis	GPT-3.5	Accuracy: 0.87; sensitivity: 0.99; specificity: 0.73	1366	Objective
Yang et al [62], 2024	Survey	Generate answers to clinical questions	ChatGPT; Bard	Concordance with the AAOS[p] Clinical Practice Guidelines: 80% (ChatGPT); 60% (Bard)	—	Subjective
Yapar et al [63], 2024	Survey	Generate answers to clinical questions	GPT-4	Accuracy: 79.8%; applicability: 75.2%; comprehensiveness: 70.6%; communication clarity: 75.6%	—	Subjective
Zhou et al [64], 2024	Case study	Generate answers to clinical questions related to the case	GPT-3.5	No statistical results	—	Subjective

[a]LLM: large language model.
[b]Not applicable.
[c]RCT: randomized controlled trial.
[d]PDQI-9: Physician Documentation Quality Instrument-9.
[e]ODI: Oswestry Disability Index.
[f]QBPDS: Quebec Back Pain Disability Scale.
[g]RMDQ: Roland-Morris Disability Questionnaire.
[h]NRS: numerical rating scale.
[i]JAMA: Journal of the American Medical Association.
[j]FRE: Flesch reading ease.
[k]FKGL: Flesch-Kincaid grade level.
[l]GPT: generative pretrained transformer.
[m]TKA: total knee arthroplasty.
[n]THA: total hip arthroplasty.
[o]AUC: area under the curve.
[p]AAOS: American Academy of Orthopaedic Surgeons.

XSL•FO
RenderX

Table 2. Characteristics of the included studies focused on orthopedic education.

Study, year	Study design	Task	LLM[a] tools	Source	Questions, n	Scores or accuracy (%)
Cuthbert and Simpson [65], 2023	Survey	Examination	GPT-3.5	UKITE[b]	134	35.8
Ghanem et al [66], 2023	Survey	Examination	GPT-4	OITE[c]	201	61.2
Han et al [67], 2024	Survey	Examination	GPT-3.5	ASSH[d]	1583	36.2
Hofmann et al [68], 2023	Survey	Examination	GPT-3.5; GPT-4	OITE	410 (GPT-3.5); 396 (GPT-4)	GPT-3.5: 46.3; GPT-4: 63.4
Jain et al [69], 2024	Survey	Examination	GPT-3.5	OITE	360	52.8
Kung et al [70], 2023	Survey	Examination	GPT-3.5; GPT-4	OITE	360	GPT-3.5: 54.3; GPT-4: 73.6
Lum [71], 2023	Survey	Examination	GPT-3.5	OITE	207	47
Massey et al [72], 2023	Survey	Examination	GPT-3.5; GPT-4	ResStudy Orthopaedic Examination Question Bank	180	GPT-3.5: 29.4; GPT-4: 47.2
Ozdag et al [73], 2023	Survey	Examination	GPT-3.5	OITE	102	45
Rizzo et al [74], 2024	Survey	Examination	GPT-3.5-Turbo; GPT-4	OITE	2022: 207; 2021: 213; 2020: 215	2022: GPT-4=67.63; GPT 3.5-Turbo=50.24; 2021: GPT-4=58.69; GPT 3.5-Turbo=47.42; 2020: GPT-4=59.53; GPT 3.5-Turbo=46.51
Saad et al [75], 2023	Survey	Examination	GPT-4	Mock FRCS Orth[e] Part A	240	67.5
Traoré et al [76], 2023	Survey	Examination	GPT-3.5	EBHS[f] diploma examination	18	0

[a]LLM: large language model.

[b]UKITE: United Kingdom and Ireland In-Training Examination.

[c]OITE: Orthopaedic Surgery In-Training Examination.

[d]ASSH: American Society for Surgery of the Hand.

[e]FRCS Orth: Orthopaedic Fellow of the Royal College of Surgeons.

[f]EBHS: European Board of Hand Surgery.

XSL•FO
RenderX

Table 3. Characteristics of the included studies focused on orthopedic research and management.

Study, year	Study design	Task	LLM[a] tools	Input	Key findings
Gill et al [77], 2024	Survey	Improve readability	GPT-3.5	IRB[b]-approved orthopedic surgery research consent forms	ChatGPT can significantly improve the readability of orthopedic clinical research consent forms; 63.2% of the post-ChatGPT consent forms had at least 1 error
Hakam et al [78], 2024	Survey	AI[c]-Generated scientific literature	GPT-3.4; You.com	Five abstracts about meniscal injuries	The AI-generated texts could not be successfully identified
Kacena et al [79], 2024	Survey	Write scientific review articles	GPT-4	Prompts	AI reduced the time for writing but had significant inaccuracies
Lawrence et al [80], 2024	Survey	Generate abstract	GPT-3	A standard set of input commands	Interrater reliability for abstract quality scores was moderate
Lotz et al [81], 2023	Survey	Assist new research hypothesis exploration	Toolkit based on GPT-3.5	Prior studies	LLMs may be useful for analyzing and distinguishing publications, as well as determining the degree to which the literature supports or contradicts emergent hypotheses
Methnani et al [82], 2023	Survey	Calculate sample size	GPT-3.5	All necessary data, such as mean, percentage SD, normal deviations, and study design	In 1 (25%) of the 4 trials, the sample size was correctly calculated
Nazzal et al [83], 2024	Survey	Write a review article	GPT-4	Prompts	The AI-only paper was the most inaccurate, with inappropriate reference use, and the AI-assisted paper had the greatest incidence of plagiarism
Sanii et al [84], 2023	Survey	Perform an orthopedic surgery literature review	GPT-3; Perplexity	Standard prompts	The current iteration of ChatGPT cannot perform a reliable literature review, and Perplexity is only able to perform a limited review of the medical literature
Zaidat et al [85], 2023	Retrospective cohort study	Predict CPT[d] codes	GPT-4	Surgical operative notes	The AUROC[e] score was 0.87, and the AUPRC[f] score was 0.67

[a]LLM: large language model.

[b]IRB: institutional review board.

[c]AI: artificial intelligence.

[d]CPT: current procedural terminology.

[e]AUROC: area under the receiver operating characteristic curve.

[f]AUPRC: area under the precision-recall curve.

Figure 2. Trends in the number of publications.

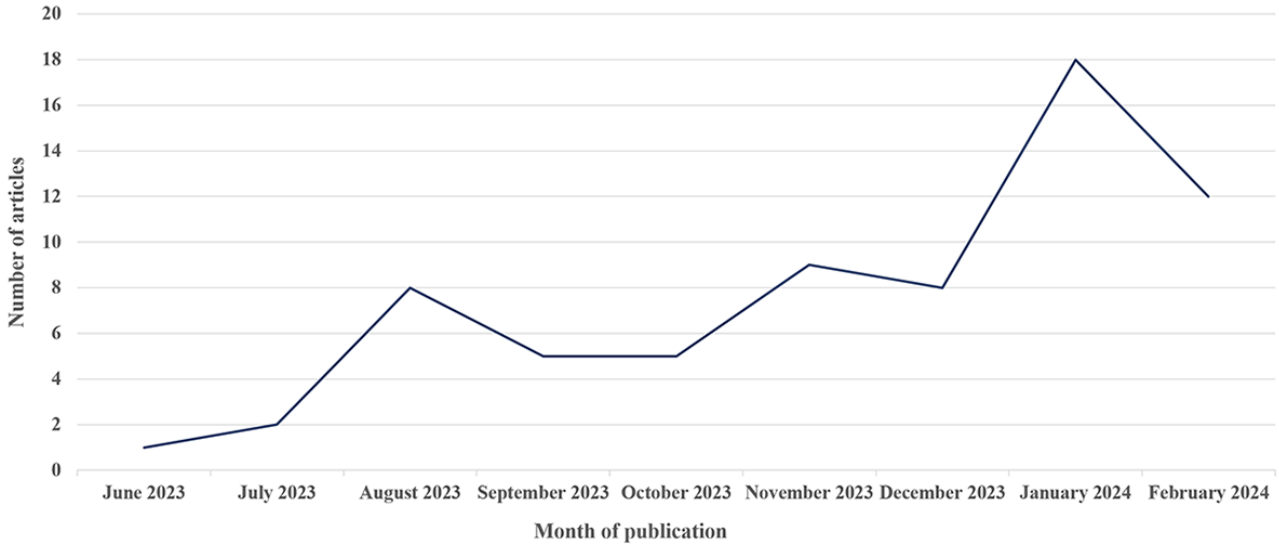

XSL•FO
RenderX

Quality Assessment of Studies

We conducted quality assessments for the 8 studies that recruited participants (Multimedia Appendices 4 and 5). The RCT study was evaluated using the revised Cochrane risk-of-bias tool for randomized trials, and it was found to have a low risk of bias in all 5 domains—bias arising from the randomization process, bias due to deviations from the intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and bias in selection of the reported result—indicating high study quality. The remaining studies (7/8, 88%; observational studies) were evaluated using the CONSORT-AI guidance and were found to be of good quality.

Distribution of LLM Tools

Among all the LLM tools applied, ChatGPT was the most commonly mentioned. Other LLM tools included Bard, Microsoft Bing, Scholar AI, Perplexity, Gatortron, Claude, Command-xlarge-nightly, and Bloomz, as well as software developed by researchers based on commonly used LLM kernels. Currently, there are 2 main versions of ChatGPT available: GPT-3 (including GPT-3.5 and GPT-3.5-Turbo) and GPT-4. Most of the studies (48/68, 71%) specified the version of the tool used. The majority of the studies (25/48, 52%) only used GPT-3 or 3.5, likely due to the publication lag because these studies were conducted before the release of GPT-4. Given that GPT-4 outperforms GPT-3 in most tasks, future research should primarily use GPT-4.

Model Performance Evaluation

As shown in Tables 1-3, there is considerable heterogeneity in the definition, measurement, and evaluation of LLM performance across the included studies. Currently, there is no unified research paradigm for the application of LLMs in medicine. Therefore, this review focused on different model performance evaluation metrics according to various application categories. For clinical applications of LLMs, we were particularly concerned with the accuracy of model reasoning and the readability of the generated text; unfortunately, the majority of the studies (39/47, 83%) relied on subjective assessments of the model's performance. In studies with objective evaluations, the heterogeneity in the subtasks performed by the LLMs (including diagnosis, classification, clinical case analysis, and case text generation) prevented us from pooling the data. For diagnostic tasks alone, the accuracy ranged from 55% to 93% [24,53]. When performing disease classification tasks, GPT-4's accuracy ranged from 2% to 100% [29,47] (Table 1). In studies on readability, the most commonly used metrics are the Flesch Reading Ease and the Flesch-Kincaid Grade Level (FKGL) scores. The FKGL metric correlates reading difficulty with years of education, providing a straightforward reflection of the readability of generated materials. In these studies, the generated texts had FKGL scores ranging from a minimum of 5.0 [38], indicating primary school reading difficulty, to the maximum required years of education [32], showing significant variability. This variability is likely due to differences in the research questions, methodologies, prompts, and evaluators. In the educational applications (eg, answering questions from examination papers), the most frequently used test source (7/12, 58%) was the Orthopaedic Surgery In-Training Examination (OITE). The test scores are widely recognized as performance evaluation metrics for the models, with final scores ranging from 45% to 73.6% due to differences in models and test selections (Table 2). For applications of LLMs in research and management, the flexible and varied nature of the tasks led to substantial differences in performance measurement and evaluation. Therefore, we collected the model inputs and major findings for a descriptive presentation (Table 3).

Discussion

Overview

Despite the relatively short time since their introduction and the absence of rigorous and comprehensive performance evaluation in highly specialized fields such as orthopedics, it is an undeniable fact that LLMs have already been made accessible to the public. Given the increasing acceptance and widespread adoption of LLMs, it is imperative for orthopedic surgeons to possess a comprehensive understanding of their operational mechanisms and limitations. Users should also delineate the safe application boundaries while harnessing the benefits offered by LLMs, all while mitigating potential risks in their daily clinical practice. This section presents a comprehensive overview of application examples and model performance across diverse fields, while providing strategic approaches to address LLMs based on our findings. In addition, in this section, we critically evaluate research methodologies and offer potential recommendations for future investigations.

Application of LLMs in the Field of Orthopedic Education

LLMs can not only provide answers but also offer explanations and even engage in further discussions on a given topic, demonstrating potential value in orthopedic education. Several studies have evaluated the performance of ChatGPT in answering questions related to orthopedics and further discussed its value in the field of orthopedic education. The source of questions includes the OITE [66,68-71,73,74], the ResStudy Orthopaedic Examination Question Bank [72], the Fellowship of the Royal College of Surgeons (Trauma and Orthopaedic Surgery) examination [65,75], and hand surgery examinations in the United States and Europe [67,76]. Accuracy (scores) and whether the answers meet the standard are important evaluation criteria for LLM performance. Another educational indicator is the correctness and reasonableness of answer explanations. Studies evaluating OITE questions usually convert accuracy into postgraduate year (PGY) levels for evaluation. Due to differences in software applications and question selection, different studies have reported varying performances of ChatGPT. ChatGPT with GPT-4 performed at an average level ranging from PGY-2 to PGY-5 [66,68,70,74], while ChatGPT with GPT-3.5 performed slightly better than PGY-1 or below the average level of PGY-1 [68-71,73,74]. For correct answers, ChatGPT can provide explanations and reasoning processes consistent with those of examiners, which helps students understand the questions and general orthopedic principles [66,69]. However, ChatGPT failed to pass the Fellowship of the Royal College of Surgeons (Trauma and Orthopaedic

XSL•FO
RenderX

Surgery) examination and hand surgery examinations in the United States and Europe [65,67,75,76]. In addition, as a language model, ChatGPT cannot analyze medical images correctly [70], limiting its role in orthopedic imaging education.

Although LLMs currently cannot fully replace orthopedic instructors, they can still serve as a valuable supplementary tool for learning. Integrating their responses with authoritative resources for verification and using appropriate prompts can optimize their capacity to offer logical explanations and foster critical thinking.

Application of LLMs in Clinical Practice

Medical Consultation and Physician-Patient Communication

One challenge faced by orthopedic physicians is that, unlike in the case of other clinical interventions, LLMs have already been integrated as medical consultation tools in the diagnosis and treatment process of numerous diseases without sufficient clinical evidence and regulatory review from authorities such as the US Food and Drug Administration. LLMs can be considered an alternative approach for patients who have sustained injuries or experience discomfort before seeking guidance from primary care physicians or specialists. When confronted with medical issues, individuals who rely heavily on the internet for problem-solving in their personal and professional lives often exhibit a tendency to seek treatment decisions on the web [30]. Compared to traditional search engines or Wikipedia, LLMs could potentially become a significant source of medical consultation information, especially in cases of nonacute diseases such as lower back pain or joint pain. Meanwhile, many physicians also hope that LLMs can help alleviate their burden of simple medical consultations and repetitive paperwork related to physician-patient communication (such as preoperative consent forms), which is considered 1 of the important factors contributing to physician burnout [86]. Although LLMs can provide concise, clarified, or simplified responses related to the given topic and deliver high-quality and empathetic answers [66,87], given their imperfect performance in addressing questions related to orthopedics [65-76], caution should be exercised regarding their reliability in orthopedic consultation scenarios.

Studies have evaluated the performance of LLMs in answering questions related to hand surgery [23], spinal cord injuries [55], joint and sports medicine [19,27,35,41,56,58], and preoperative physician-patient communication for lumbar disk herniation [52] and hip replacement surgery [44]. In these studies, the evaluation criteria of interest typically encompass the model's answer accuracy, readability, completeness, and information sources. Evaluation methods often encompass scale assessments or subjective ratings conducted by researchers. The DISCERN score is commonly used to evaluate answer quality [19,23,58,88], while FKGL and Flesch Reading Ease scores are commonly used to measure readability [23,55,58]. The accuracy of LLMs' responses is closely correlated with the specific topic. Questions in the field of joint and sports medicine often receive high-quality responses, while there are serious issues with the quality of answers regarding spinal cord injuries. There are also significant differences in the evaluation of the readability or comprehensibility of LLMs' answers, with some researchers considering them to be easily understood [44,52], while studies using Flesch-related scales suggest that LLMs' answers require a reading level of at least 10 years of education or even university level for full comprehension [23,55,58]. The underlying factors contributing to this phenomenon can be attributed to variations in question topics, prompts, and evaluation methodologies used for answer assessment. Consequently, orthopedic surgeons should exercise caution when interpreting the findings of these studies.

Although LLMs can offer more scholarly health information in comparison to search engines [27,56], they still cannot replace orthopedic physicians in medical consultation and physician-patient communication. Using LLMs as a guiding tool and maintaining communication with physicians during further diagnosis and treatment decisions may be a safer and more effective strategy.

Clinical Workflow

The performance of LLMs in orthopedic examinations suggests that they cannot handle complex tasks independently, but they hold potential to serve as valuable assistants for orthopedic physicians. One possible application is using LLMs to automate simple, repetitive tasks such as writing medical records for common orthopedic diseases [20]. In the context of complex disease management tasks, LLMs can possess a more extensive and specialized knowledge base than less experienced newly graduated physicians and assist them in various aspects of disease management. Some researchers have tested the performance of LLMs using specific clinical questions or guidelines [21,26,28,50], while others have directly inputted clinical case data into the model, allowing it to summarize and provide corresponding diagnostic or treatment decisions autonomously [18,24,45,57,64]. Currently, there is no further research on introducing LLMs into orthopedic operations, likely because of the limited availability of intelligent terminals and digital scenarios that may combine operative procedures with LLMs. Potential docking scenarios for the LLM model could include intelligent surgical applications such as mixed reality operating rooms [89,90] and autonomous laminectomy robots [91,92].

In the context of clinical practice, apart from the fundamental requirement of accurate response, time consumption and work efficiency also serve as crucial reference indicators for evaluating LLMs' performance. Despite variations in the assessment of model accuracy across the included studies, potentially attributed to differences in research objectives, prompt design, evaluation criteria, and assessment tools, no study has presented evidence indicating that LLMs can independently perform clinical work. Therefore, the current models still require rigorous supervision during their use. An RCT study evaluating the performance of ChatGPT in assisting with orthopedic clinical documentation found that there was no significant efficiency advantage in using ChatGPT: the time taken to complete medical history writing was not superior to voice input, and instances of fabricated content were observed within the ChatGPT-generated medical histories [20].

XSL•FO
RenderX

Although LLMs currently have limitations, they remain valuable tools for orthopedic surgeons in their daily practice. It is important to approach cautiously the responses provided by LLMs and seek additional evidence and explanations from the model used when faced with unclear answers. By incorporating evidence-based medicine tools, we can ultimately achieve superior clinical diagnoses and treatment plans, thereby elevating the quality of care delivered by physicians.

Application of LLMs in the Field of Research

Research is generally considered a creative endeavor, and introducing LLMs into the field of research may offer more flexibility. Currently, there are limited attempts to use LLMs in orthopedic research. A study found that lowering the reading threshold of professional texts through LLMs can assist in improving the readability of informed consent forms for orthopedic clinical research, but the forms did not meet the recommended sixth-grade reading level set by the American Medical Association [77]. In addition, the literature summarization and generation capabilities of LLMs can contribute to independent or assisted writing of literature reviews in the orthopedic field [79,83]. On the other side of the coin, concerns about integrity arise when scholars find that the model's output can be deceptively realistic. The abstracts generated by LLMs for studies on *meniscal injuries* and *joint replacement* were indistinguishable from those written by human researchers [78,80]. However, web-based LLMs do not perform well in tasks such as literature review or sample size estimation in sports medicine research [82,84]. Possible reasons may include the potential limitations of LLMs in meeting logical reasoning requirements and the inappropriate use of prompts. For more complex tasks, an optimization approach could involve developing task-specific toolkits based on the fundamental architecture of LLMs. The feasibility of this approach has been validated in interdisciplinary research on the management of back pain [81].

Application of LLMs in Management

Trained NLP models can convert natural language into structured data and have demonstrated superior performance in tasks involving the current procedural terminology for identifying spinal surgery records [93]. However, ChatGPT, with its larger parameters, performs weaker than NLP models in the task of identifying spinal surgery current procedural terminology codes [85]. One possible reason is that traditional NLP models have been trained on more targeted datasets, whereas researchers cannot fine-tune the backend model of ChatGPT using these data. Despite the current model's performance limitations hindering its further application in this field, the potential advancements in "fine-tuning" techniques may enable LLMs to assume a more influential role in orthopedic management in the future.

Current Advantages and Limitations of LLMs in Orthopedic Applications

Overview

In contrast to conventional pretrained machine learning models, LLMs exhibit the advantage of versatility by accurately addressing problems across various domains without necessitating additional training on specific samples. In the field of orthopedics, another advantage of LLMs is their user-friendly and convenient nature. Users do not need to go through the long process of waiting and referral from general practitioners to specialists. By simply accessing apps equipped with LLMs, users can inquire about diverse subspecialties in orthopedics at any time and from anywhere, receiving answers promptly at a minimal cost or even free of charge. This service surpasses the capabilities of current health care systems and is unlikely to be replicated in the foreseeable future.

However, as mentioned previously, these advantages are based on unverified answers and unpredictable risks. The answers provided by LLMs for questions related to orthopedics are less robust than those for everyday common knowledge and have significant limitations in terms of accuracy, readability, reliability, and timeliness, as detailed in the following subsections.

Accuracy

Almost all studies (66/68, 97%) found errors in LLMs' responses, with more noticeable inaccuracies in specialized areas such as hip and knee joints and hand surgery [62,67,76]. Some answers even contradicted fundamental orthopedic knowledge [52]. Therefore, some researchers argue that current expectations for guidance provided by AI platforms should be tempered by both physicians and patients [62]. Possible reasons include the limited availability of publicly accessible orthopedic data for training, especially for specialized diseases, as well as privacy concerns that restrict public access to a large amount of data. In the future, besides waiting for more powerful next-generation LLMs, using existing LLMs to learn orthopedic cases and fine-tuning them may be a potential solution to improve accuracy.

Readability

Some of the included studies (3/68, 4%) suggest that the content generated by LLMs is not satisfactory in terms of readability for the general population [23,55,58]. The potential reasons for the lack of readability may include not only the limited training data but also the quality of the trained data. By incorporating more popular science content and common clinical responses, it may be possible to address the issue of readability through fine-tuning the model.

Reliability

Different ways of asking the same question may yield completely different answers [21]. This instability, particularly in response to specific prompts, not only affects users' experience and trust but also greatly interferes with researchers' homogenized evaluations. It is imperative to establish standardized questioning processes and prompt criteria.

Timeliness

Training LLMs from scratch is both costly and time consuming, leading to significant retraining expenses. However, unlike everyday common knowledge, orthopedics is constantly evolving with new diagnostic and treatment approaches as well as surgical techniques. Therefore, outdated information becomes

XSL•FO
RenderX

an important risk factor leading to inaccurate answers, necessitating caution in this context.

Methodological Limitations of the Selected Studies

Although there are 47 studies related to clinical issues, only 8 (17%) recruited patients [20,22,24,29,45,48,57,61]. Many of the studies (39/47, 83%) only focus on investigation and evaluation, lacking rigorous methods for clinical research, such as RCTs. Furthermore, there is a lack of research end points directly linked to patient outcomes, such as cure rates or improvements in quality of life, making it difficult to find direct evidence of prognosis. Most of the studies (46/68, 68%) rely on subjective methodologies, such as expert ratings, for model evaluation and lack objective criteria and approaches for assessment, leading to unreliable research results. Furthermore, the absence of standardized questioning paradigms has led to instability in LLM responses, posing challenges for reproducibility and limiting the reliability and clinical significance of the study findings.

Limitations of This Review

This systematic review has several limitations. First, only English-language articles were included, which may have led to the exclusion of relevant studies published in other languages. Second, due to significant heterogeneity in study designs, model tasks, and evaluation parameters among the included studies, we did not perform a comprehensive synthesis of most of the data, nor did we conduct a meta-analysis. Third, our search was restricted to commonly used medical research databases such as PubMed, Embase, and Cochrane Library, potentially overlooking relevant studies from other sources, including conference papers and gray literature. Finally, given the limited availability of rigorous clinical studies, we included a considerable number of subjective surveys. Although our objective was to provide a broad overview of LLM-related information, this may have introduced bias into the findings. These limitations are expected to be addressed as more standardized, high-quality clinical studies become available in future research.

Conclusions

Due to the current limitations of LLMs, they cannot replace orthopedic professionals in the short term. However, using LLMs as copilots could be a potential approach to effectively enhance work efficiency at present. In addition, developing task-specific downstream tools based on LLMs is also a potential solution to improve model performance for further use.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82272577). The funding organization had no involvement in the design of the study, data collection, data analysis, data interpretation, writing of the report, or the decision to publish. Its role was strictly limited to providing financial support. We used the generative AI tool ChatGPT with GPT-4 [94] for language refinement purposes. The AI tool was only used to perform grammar checks and assist with enhancing the clarity and fluency of the writing. All intellectual contributions, including research design, data collection, analysis, and interpretation, were made by the authors. The final manuscript has been reviewed and approved by the authors, who take full responsibility for its content.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions

NX, WL, CZ, and SL made contributions to conception and design. XZ, SZ, and YT made contributions to the acquisition, analysis, and interpretation of data. CZ, SL, and XZ drafted the manuscript, and NX, WL, SZ, YT, and SW revised it critically for important intellectual content. All authors approved the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 83 KB - jmir_v26i1e59607_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 25 KB - jmir_v26i1e59607_app2.docx]

Multimedia Appendix 3
Data extraction strategy.
[DOCX File , 22 KB - jmir_v26i1e59607_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Revised Cochrane risk-of-bias tool for randomized trials template for assessment completion.
[PDF File (Adobe PDF File), 232 KB - jmir_v26i1e59607_app4.pdf]

Multimedia Appendix 5
Quality assessment of studies of large language models based on CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidance.
[DOCX File , 53 KB - jmir_v26i1e59607_app5.docx]

References

1. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv Preprint posted online on March 31, 2023 [FREE Full text]
3. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv Preprint posted online on June 15, 2022 [FREE Full text]
4. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv Preprint posted online on March 15, 2023 [FREE Full text]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 05;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
7. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls. Ann Intern Med 2024 Feb;177(2):210-220. [doi: 10.7326/m23-2772]
8. Tang YD, Dong ED, Gao W. LLMs in medicine: the need for advanced evaluation systems for disruptive technologies. Innovation (Camb) 2024 May 06;5(3):100622 [FREE Full text] [doi: 10.1016/j.xinn.2024.100622] [Medline: 38699776]
9. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
10. Knezevic NN, Candido KD, Vlaeyen JW, Van Zundert J, Cohen SP. Low back pain. Lancet 2021 Jul 03;398(10294):78-92. [doi: 10.1016/S0140-6736(21)00733-9] [Medline: 34115979]
11. Duong V, Oo WM, Ding C, Culvenor AG, Hunter DJ. Evaluation and treatment of knee pain: a review. JAMA 2023 Oct 24;330(16):1568-1580. [doi: 10.1001/jama.2023.19675] [Medline: 37874571]
12. Yao JJ, Aggarwal M, Lopez RD, Namdari S. Current concepts review: large language models in orthopaedics: definitions, uses, and limitations. J Bone Joint Surg Am 2024 Jun 19. [doi: 10.2106/JBJS.23.01417] [Medline: 38896652]
13. Fayed AM, Mansur NS, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. J Exp Orthop 2023 Jul 26;10(1):74 [FREE Full text] [doi: 10.1186/s40634-023-00642-8] [Medline: 37493985]
14. Merrell LA, Fisher ND, Egol KA. Large language models in orthopaedic trauma: a cutting-edge technology to enhance the field. J Bone Joint Surg Am 2023 Sep 06;105(17):1383-1387. [doi: 10.2106/JBJS.23.00395] [Medline: 37402227]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
17. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ 2020 Sep 09;370:m3164 [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]
18. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: an analysis of ChatGPT 3.5, ChatGPT 4, and Bard. J Orthop 2024 Apr;50:1-7 [FREE Full text] [doi: 10.1016/j.jor.2023.11.063] [Medline: 38148925]
19. Anastasio AT, Mills FB4, Karavan MPJ, Adams SBJ. Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery. Foot Ankle Orthop 2023 Oct 22;8(4):24730114231209919 [FREE Full text] [doi: 10.1177/24730114231209919] [Medline: 38027458]

XSL•FO
RenderX

20. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. J Am Acad Orthop Surg 2024 Feb 01;32(3):123-129. [doi: 10.5435/JAAOS-D-23-00474] [Medline: 37976385]
21. Christy M, Morris MT, Goldfarb CA, Dy CJ. Appropriateness and reliability of an online artificial intelligence platform's responses to common questions regarding distal radius fractures. J Hand Surg Am 2024 Feb;49(2):91-98. [doi: 10.1016/j.jhsa.2023.10.019] [Medline: 38069953]
22. Coraci D, Maccarone MC, Regazzo G, Accordi G, Papathanasiou JV, Masiero S. ChatGPT in the development of medical questionnaires. The example of the low back pain. Eur J Transl Myol 2023 Dec 15;33(4):12114 [FREE Full text] [doi: 10.4081/ejtm.2023.12114] [Medline: 38112605]
23. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS. Evaluation of online artificial intelligence-generated information on common hand procedures. J Hand Surg Am 2023 Nov;48(11):1122-1127. [doi: 10.1016/j.jhsa.2023.08.003] [Medline: 37690015]
24. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int 2023 Nov;7(6):2534-2541 [FREE Full text] [doi: 10.1016/j.jseint.2023.07.018] [Medline: 37969495]
25. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 02;6(10):e2336997 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
26. Draschl A, Hauer G, Fischerauer SF, Kogler A, Leitner L, Andreou D, et al. Are ChatGPT's free-text responses on periprosthetic joint infections of the hip and knee reliable and useful? J Clin Med 2023 Oct 20;12(20):6655 [FREE Full text] [doi: 10.3390/jcm12206655] [Medline: 37892793]
27. Dubin JA, Bains SS, Chen Z, Hameed D, Nace J, Mont MA, et al. Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. J Arthroplasty 2023 Jul;38(7):1195-1202. [doi: 10.1016/j.arth.2023.04.007] [Medline: 37040823]
28. Duey AH, Nietsch KS, Zaidat B, Ren R, Ndjonko LC, Shrestha N, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J 2023 Nov;23(11):1684-1691. [doi: 10.1016/j.spinee.2023.07.015] [Medline: 37499880]
29. Fabijan A, Polis B, Fabijan R, Zakrzewski K, Nowosławska E, Zawadzka-Fabijan A. Artificial intelligence in scoliosis classification: an investigation of language-based models. J Pers Med 2023 Dec 09;13(12):1695 [FREE Full text] [doi: 10.3390/jpm13121695] [Medline: 38138922]
30. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament injury. J Pers Med 2024 Jan 18;14(1):104 [FREE Full text] [doi: 10.3390/jpm14010104] [Medline: 38248805]
31. Gianola S, Bargeri S, Castellini G, Cook C, Palese A, Pillastrini P, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. J Orthop Sports Phys Ther 2024 Mar;54(3):222-228. [doi: 10.2519/jospt.2024.12151] [Medline: 38284363]
32. Hurley ET, Crook BS, Lorentz SG, Danilkowicz RM, Lau BC, Taylor DC, et al. Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. Arthroscopy 2024 Mar;40(3):726-31.e6. [doi: 10.1016/j.arthro.2023.07.048] [Medline: 37567487]
33. Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. Arthrosc Sports Med Rehabil 2024 Apr;6(2):100893 [FREE Full text] [doi: 10.1016/j.asmr.2024.100893] [Medline: 38375341]
34. Johns WL, Martinazzi BJ, Miltenberg B, Nam HH, Hammoud S. ChatGPT provides unsatisfactory responses to frequently asked questions regarding anterior cruciate ligament reconstruction. Arthroscopy 2024 Jul;40(7):2067-79.e1. [doi: 10.1016/j.arthro.2024.01.017] [Medline: 38311261]
35. Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2023 Nov;31(11):5190-5198 [FREE Full text] [doi: 10.1007/s00167-023-07529-2] [Medline: 37553552]
36. Kasthuri VS, Glueck J, Pham H, Daher M, Balmaceno-Criss M, McDonald CL, et al. Assessing the accuracy and reliability of AI-generated responses to patient questions regarding spine surgery. J Bone Joint Surg Am 2024 Jun 19;106(12):1136-1142. [doi: 10.2106/JBJS.23.00914] [Medline: 38335266]
37. Kienzle A, Niemann M, Meller S, Gwinner C. ChatGPT may offer an adequate substitute for informed consent to patients prior to total knee arthroplasty-yet caution is needed. J Pers Med 2024 Jan 05;14(1):69 [FREE Full text] [doi: 10.3390/jpm14010069] [Medline: 38248771]
38. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? Clin Orthop Relat Res 2023 Nov 01;481(11):2260-2267. [doi: 10.1097/CORR.0000000000002668] [Medline: 37116006]
39. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023 Sep 15;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]

XSL•FO
RenderX

40. Li J, Gao X, Dou T, Gao Y, Zhu W. Assessing the performance of GPT-4 in the filed of osteoarthritis and orthopaedic case consultation. medRxiv Preprint posted online on August 09, 2023 [FREE Full text] [doi: 10.1101/2023.08.06.23293735]
41. Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. Arthroscopy 2024 Jul;40(7):2058-2066. [doi: 10.1016/j.arthro.2023.12.009] [Medline: 38171421]
42. Lower K, Seth I, Lim B, Seth N. ChatGPT-4: transforming medical education and addressing clinical exposure challenges in the post-pandemic era. Indian J Orthop 2023 Sep;57(9):1527-1544. [doi: 10.1007/s43465-023-00967-7] [Medline: 37609022]
43. Magruder ML, Rodriguez AN, Wong JCJ, Erez O, Piuzzi NS, Scuderi GR, et al. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. J Arthroplasty 2024 Aug;39(8):2022-2027. [doi: 10.1016/j.arth.2024.02.023] [Medline: 38364879]
44. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am 2023 Oct 04;105(19):1519-1526. [doi: 10.2106/JBJS.23.00209] [Medline: 37459402]
45. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol 2023 Nov 28;24(1):61 [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
46. Mejia MR, Arroyave JS, Saturno M, Ndjonko LC, Zaidat B, Rajjoub R, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American Spine Society guideline comparison. Neurospine 2024 Mar;21(1):149-158 [FREE Full text] [doi: 10.14245/ns.2347052.526] [Medline: 38291746]
47. Russe MF, Fink A, Ngo H, Tran H, Bamberg F, Reisert M, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep 2023 Aug 30;13(1):14215 [FREE Full text] [doi: 10.1038/s41598-023-41512-8] [Medline: 37648742]
48. Schonfeld E, Pant A, Shah A, Sadeghzadeh S, Pangal D, Rodrigues A, et al. Evaluating computer vision, large language, and genome-wide association models in a limited sized patient cohort for pre-operative risk stratification in adult spinal deformity surgery. J Clin Med 2024 Jan 23;13(3):656 [FREE Full text] [doi: 10.3390/jcm13030656] [Medline: 38337352]
49. Seth I, Xie Y, Rodwell A, Gracias D, Bulloch G, Hunter-Smith DJ, et al. Exploring the role of a large language model on carpal tunnel syndrome management: an observation study of ChatGPT. J Hand Surg Am 2023 Oct;48(10):1025-1033. [doi: 10.1016/j.jhsa.2023.07.003] [Medline: 37530687]
50. Shrestha N, Shen Z, Zaidat B, Duey AH, Tang JE, Ahmed W, et al. Performance of ChatGPT on NASS clinical guidelines for the diagnosis and treatment of low back pain: a comparison study. Spine (Phila Pa 1976) 2024 May 01;49(9):640-651. [doi: 10.1097/BRS.0000000000004915] [Medline: 38213186]
51. Sosa BR, Cung M, Suhardi VJ, Morse K, Thomson A, Yang HS, et al. Capacity for large language model chatbots to aid in orthopedic management, research, and patient queries. J Orthop Res 2024 Jun 21;42(6):1276-1282. [doi: 10.1002/jor.25782] [Medline: 38245845]
52. Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C, et al. Large language models: are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur Spine J 2023 Oct 11 (forthcoming). [doi: 10.1007/s00586-023-07975-z] [Medline: 37821602]
53. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG. Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) "case of the month". Cureus 2023 Aug;15(8):e43958 [FREE Full text] [doi: 10.7759/cureus.43958] [Medline: 37746411]
54. Taylor WL4, Cheng R, Weinblatt AI, Bergstein V, Long WJ. An artificial intelligence chatbot is an accurate and useful online patient resource prior to total knee arthroplasty. J Arthroplasty 2024 Aug;39(8S1):S358-S362. [doi: 10.1016/j.arth.2024.02.005] [Medline: 38350517]
55. Temel MH, Erden Y, Bağcıer F. Information quality and readability: ChatGPT's responses to the most common questions about spinal cord injury. World Neurosurg 2024 Jan;181:e1138-e1144. [doi: 10.1016/j.wneu.2023.11.062] [Medline: 38000671]
56. Tharakan S, Klein B, Bartlett L, Atlas A, Parada SA, Cohn RM. Do ChatGPT and Google differ in answers to commonly asked patient questions regarding total shoulder and total elbow arthroplasty? J Shoulder Elbow Surg 2024 Aug;33(8):e429-e437. [doi: 10.1016/j.jse.2023.11.014] [Medline: 38182023]
57. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023 Nov 17;13(1):20159 [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
58. Warren EJ, Hurley ET, Park CN, Crook BS, Lorentz S, Levin JM, et al. Evaluation of information from artificial intelligence on rotator cuff repair surgery. JSES Int 2024 Jan;8(1):53-57 [FREE Full text] [doi: 10.1016/j.jseint.2023.09.009] [Medline: 38312282]
59. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 2023 Oct 30;25:e49324 [FREE Full text] [doi: 10.2196/49324] [Medline: 37902826]

XSL•FO
RenderX

60. Wright BM, Bodnar MS, Moore AD, Maseda MC, Kucharik MP, Diaz CC, et al. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients? Bone Jt Open 2024 Feb 15;5(2):139-146 [FREE Full text] [doi: 10.1302/2633-1462.52.BJO-2023-0113.R1] [Medline: 38354748]
61. Yang F, Yan D, Wang Z. Large-scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications. J Bone Oncol 2024 Feb;44:100525 [FREE Full text] [doi: 10.1016/j.jbo.2024.100525] [Medline: 38314324]
62. Yang J, Ardavanis KS, Slack KE, Fernando ND, Della Valle CJ, Hernandez NM. Chat generative pretrained transformer (ChatGPT) and bard: artificial intelligence does not yet provide clinically supported answers for hip and knee osteoarthritis. J Arthroplasty 2024 May;39(5):1184-1190. [doi: 10.1016/j.arth.2024.01.029] [Medline: 38237878]
63. Yapar D, Demir Avcı Y, Tokur Sonuvar E, Eğerci Ö, Yapar A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. Jt Dis Relat Surg 2024 Jan 01;35(1):169-176 [FREE Full text] [doi: 10.52312/jdrs.2023.1402] [Medline: 38108178]
64. Zhou Y, Moon C, Szatkowski J, Moore D, Stevens J. Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. Eur J Orthop Surg Traumatol 2024 Feb;34(2):927-955 [FREE Full text] [doi: 10.1007/s00590-023-03742-4] [Medline: 37776392]
65. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass section 1 of the Fellowship of the Royal College of Surgeons (trauma and orthopaedics) examination? Postgrad Med J 2023 Sep 21;99(1176):1110-1114. [doi: 10.1093/postmj/qgad053] [Medline: 37410674]
66. Ghanem D, Covarrubias O, Raad M, LaPorte D, Shafiq B. ChatGPT performs at the level of a third-year orthopaedic surgery resident on the orthopaedic in-training examination. JBJS Open Access 2023;8(4):e23.00103 [FREE Full text] [doi: 10.2106/JBJS.OA.23.00103] [Medline: 38638869]
67. Han Y, Choudhry HS, Simon ME, Katt BM. ChatGPT's performance on the hand surgery self-assessment exam: a critical analysis. J Hand Surg Glob Online 2024 Mar;6(2):200-205 [FREE Full text] [doi: 10.1016/j.jhsg.2023.11.014] [Medline: 38903839]
68. Hofmann HL, Guerra GA, Le JL, Wong AM, Hofmann GH, Mayfield CK, et al. The rapid development of artificial intelligence: GPT-4's performance on orthopedic surgery board questions. Orthopedics 2024;47(2):e85-e89. [doi: 10.3928/01477447-20230922-05] [Medline: 37757748]
69. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res 2024 Jan 03;19(1):27 [FREE Full text] [doi: 10.1186/s13018-023-04467-0] [Medline: 38167093]
70. Kung J, Marshall C, Gauthier C, Gonzalez T, Jackson JB3. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8(3):e23.00056 [FREE Full text] [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
71. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res 2023 Aug 01;481(8):1623-1630. [doi: 10.1097/CORR.0000000000002704] [Medline: 37220190]
72. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023 Dec 01;31(23):1173-1179 [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
73. Ozdag Y, Hayes DS, Makar GS, Manzar S, Foster BK, Shultz MJ, et al. Comparison of artificial intelligence to resident performance on upper-extremity orthopaedic in-training examination questions. J Hand Surg Glob Online 2024 Mar;6(2):164-168 [FREE Full text] [doi: 10.1016/j.jhsg.2023.10.013] [Medline: 38903829]
74. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. J Orthop 2024 Apr;50:70-75. [doi: 10.1016/j.jor.2023.11.056] [Medline: 38173829]
75. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon 2023 Oct;21(5):263-266. [doi: 10.1016/j.surge.2023.07.001] [Medline: 37517980]
76. Traoré SY, Goetsch T, Muller B, Dabbagh A, Liverneaux PA. Is ChatGPT able to pass the first part of the European Board of Hand Surgery diploma examination? Hand Surg Rehabil 2023 Sep;42(4):362-364. [doi: 10.1016/j.hansur.2023.06.005] [Medline: 37353199]
77. Gill B, Bonamer J, Kuechly H, Gupta R, Emmert S, Kurkowski S, et al. ChatGPT is a promising tool to increase readability of orthopedic research consents. J Orthop Trauma Rehab 2024 Jan 22. [doi: 10.1177/22104917231208212]
78. Hakam HT, Prill R, Korte L, Lovreković B, Ostojić M, Ramadanov N, et al. Human-written vs AI-generated texts in orthopedic academic literature: comparative qualitative analysis. JMIR Form Res 2024 Feb 16;8:e52164 [FREE Full text] [doi: 10.2196/52164] [Medline: 38363631]
79. Kacena MA, Plotkin LI, Fehrenbacher JC. The use of artificial intelligence in writing scientific review articles. Curr Osteoporos Rep 2024 Feb;22(1):115-121 [FREE Full text] [doi: 10.1007/s11914-023-00852-0] [Medline: 38227177]

XSL•FO
RenderX

80. Lawrence KW, Habibi AA, Ward SA, Lajam CM, Schwarzkopf R, Rozell JC. Human versus artificial intelligence-generated arthroplasty literature: a single-blinded analysis of perceived communication, quality, and authorship source. Int J Med Robot 2024 Feb 13;20(1):e2621. [doi: 10.1002/rcs.2621] [Medline: 38348740]
81. Lotz JC, Ropella G, Anderson P, Yang Q, Hedderich MA, Bailey J, et al. An exploration of knowledge-organizing technologies to advance transdisciplinary back pain research. JOR Spine 2023 Dec;6(4):e1300 [FREE Full text] [doi: 10.1002/jsp2.1300] [Medline: 38156063]
82. Methnani J, Latiri I, Dergaa I, Chamari K, Ben Saad H. ChatGPT for sample-size calculation in sports medicine and exercise sciences: a cautionary note. Int J Sports Physiol Perform 2023 Oct 01;18(10):1219-1223. [doi: 10.1123/ijspp.2023-0109] [Medline: 37536678]
83. Nazzal MK, Morris AJ, Parker RS, White FA, Natoli RM, Fehrenbacher JC, et al. Using AI to write a review article examining the role of the nervous system on skeletal homeostasis and fracture healing. Curr Osteoporos Rep 2024 Feb 13;22(1):217-221 [FREE Full text] [doi: 10.1007/s11914-023-00854-y] [Medline: 38217755]
84. Sanii RY, Kasto JK, Wines WB, Mahylis JM, Muh SJ. Utility of artificial intelligence in orthopedic surgery literature review: a comparative pilot study. Orthopedics 2024;47(3):e125-e130 [FREE Full text] [doi: 10.3928/01477447-20231220-02] [Medline: 38147494]
85. Zaidat B, Lahoti YS, Yu A, Mohamed KS, Cho SK, Kim JS. Artificially intelligent billing in spine surgery: an analysis of a large language model. Global Spine J 2023 Dec 26:21925682231224753 [FREE Full text] [doi: 10.1177/21925682231224753] [Medline: 38147047]
86. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020 Jul 15;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
87. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
88. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
89. Dennler C, Bauer DE, Scheibler AG, Spirig J, Götschi T, Fürnstahl P, et al. Augmented reality in the operating room: a clinical feasibility study. BMC Musculoskelet Disord 2021 May 18;22(1):451 [FREE Full text] [doi: 10.1186/s12891-021-04339-w] [Medline: 34006234]
90. Verhey JT, Haglin JM, Verhey EM, Hartigan DE. Virtual, augmented, and mixed reality applications in orthopedic surgery. Int J Med Robot 2020 Apr;16(2):e2067. [doi: 10.1002/rcs.2067] [Medline: 31867864]
91. Li Z, Jiang S, Song X, Liu S, Wang C, Hu L, et al. Collaborative spinal robot system for laminectomy: a preliminary study. Neurosurg Focus 2022 Jan;52(1):E11. [doi: 10.3171/2021.10.FOCUS21499] [Medline: 34973664]
92. Li Z, Wang C, Song X, Liu S, Zhang Y, Jiang S, et al. Accuracy evaluation of a novel spinal robotic system for autonomous laminectomy in thoracic and lumbar vertebrae: a cadaveric study. J Bone Joint Surg Am 2023 Jun 21;105(12):943-950. [doi: 10.2106/JBJS.22.01320] [Medline: 36943914]
93. Kim JS, Vivas A, Arvind V, Lombardi J, Reidler J, Zuckerman SL, et al. Can natural language processing and artificial intelligence automate the generation of billing codes from operative note dictations? Global Spine J 2023 Sep;13(7):1946-1955 [FREE Full text] [doi: 10.1177/21925682211062831] [Medline: 35225694]
94. ChatGPT. URL: https://chatgpt.com [accessed 2024-04-29]

Abbreviations

AI: artificial intelligence
CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence
FKGL: Flesch-Kincaid grade level
GPT: generative pretrained transformer
LLM: large language model
NLP: natural language processing
OITE: Orthopaedic Surgery In-Training Examination
PGY: postgraduate year
PLM: pretrained language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by G Eysenbach, A Mavragani; submitted 17.04.24; peer-reviewed by J Wu, X Liu, DG Chang; comments to author 20.06.24; revised version received 01.08.24; accepted 11.09.24; published 15.11.24.

Please cite as:
Zhang C, Liu S, Zhou X, Zhou S, Tian Y, Wang S, Xu N, Li W
Examining the Role of Large Language Models in Orthopedics: Systematic Review
J Med Internet Res 2024;26:e59607
URL: https://www.jmir.org/2024/1/e59607
doi:10.2196/59607
PMID:

©Cheng Zhang, Shanshan Liu, Xingyu Zhou, Siyu Zhou, Yinglun Tian, Shenglin Wang, Nanfang Xu, Weishi Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Human Factors, Human-Centered Design, and Usability of Sensor-Based Digital Health Technologies: Scoping Review

Animesh Tandon[1,2,3], MD, MS; Bryan Cobb[4], PhD; Jacob Centra[5], DPT; Elena Izmailova[6], PhD; Nikolay V Manyakov[7], PhD; Samantha McClenahan[5], PhD; Smit Patel[5], PharmD; Emre Sezgin[8], PhD; Srinivasan Vairavan[9], PhD; Bernard Vrijens[10], PhD; Jessie P Bakker[5,11,12], BS, MS, PhD; Digital Health Measurement Collaborative Community (DATAcc) hosted by DiMe[13]

[1]Division of Cardiology and Cardiovascular Medicine, Department of Heart, Vascular, and Thoracic, Children's Institute, Cleveland Clinic Children's, Cleveland, OH, United States

[2]Cleveland Clinic Children's Center for Artificial Intelligence, Department of Heart, Vascular, and Thoracic, Children's Institute, Cleveland Clinic Children's, Cleveland, OH, United States

[3]Department of Pediatrics, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH, United States

[4]Healthcare Innovations Delivery, Neurology, Medical Affairs, Genentech, San Francisco, CA, United States

[5]Digital Medicine Society, Boston, MA, United States

[6]Koneksa Health, New York, NY, United States

[7]Data Science and Digital Health, Johnson & Johnson Innovative Medicine, Beerse, Belgium

[8]The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

[9]Johnson & Johnson Innovative Medicine, New Jersey, NJ, United States

[10]AARDEX Group, Liège, Belgium

[11]Division of Sleep and Circadian Disorders, Mass General Brigham, Boston, MA, United States

[12]Division of Sleep Medicine, Harvard Medical School, Boston, MA, United States

[13]See Acknowledgements

Corresponding Author:
Jessie P Bakker, BS, MS, PhD
Digital Medicine Society
90 Canal Street
4th Floor
Boston, MA, 02114
United States
Phone: 1 6175833777
Email: jessie.bakker@dimesociety.org

Abstract

Background: Increasing adoption of sensor-based digital health technologies (sDHTs) in recent years has cast light on the many challenges in implementing these tools into clinical trials and patient care at scale across diverse patient populations; however, the methodological approaches taken toward sDHT usability evaluation have varied markedly.

Objective: This review aims to explore the current landscape of studies reporting data related to sDHT human factors, human-centered design, and usability, to inform our concurrent work on developing an evaluation framework for sDHT usability.

Methods: We conducted a scoping review of studies published between 2013 and 2023 and indexed in PubMed, in which data related to sDHT human factors, human-centered design, and usability were reported. Following a systematic screening process, we extracted the study design, participant sample, the sDHT or sDHTs used, the methods of data capture, and the types of usability-related data captured.

Results: Our literature search returned 442 papers, of which 85 papers were found to be eligible and 83 papers were available for data extraction and not under embargo. In total, 164 sDHTs were evaluated; 141 (86%) sDHTs were wearable tools while the remaining 23 (14%) sDHTs were ambient tools. The majority of studies (55/83, 66%) reported summative evaluations of final-design sDHTs. Almost all studies (82/83, 99%) captured data from targeted end users, but only 18 (22%) out of 83 studies captured data from additional users such as care partners or clinicians. User satisfaction and ease of use were evaluated for 83% (136/164) and 91% (150/164) of sDHTs, respectively; however, learnability, efficiency, and memorability were reported for only

XSL•FO
RenderX

11 (7%), 4 (2%), and 2 (1%) out of 164 sDHTs, respectively. A total of 14 (9%) out of 164 sDHTs were evaluated according to the extent to which users were able to understand the clinical data or other information presented to them (understandability) or the actions or tasks they should complete in response (actionability). Notable gaps in reporting included the absence of a sample size rationale (reported for 21/83, 25% of all studies and 17/55, 31% of summative studies) and incomplete sociodemographic descriptive data (complete age, sex/gender, and race/ethnicity reported for 14/83, 17% of studies).

Conclusions: Based on our findings, we suggest four actionable recommendations for future studies that will help to advance the implementation of sDHTs: (1) consider an in-depth assessment of technology usability beyond user satisfaction and ease of use, (2) expand recruitment to include important user groups such as clinicians and care partners, (3) report the rationale for key study design considerations including the sample size, and (4) provide rich descriptive statistics regarding the study sample to allow a complete understanding of generalizability to other patient populations and contexts of use.

(J Med Internet Res 2024;26:e57628) doi:10.2196/57628

KEYWORDS

digital health; remote; decentralized; sensors; connected care; usability; ergonomics; human-centered design; user experience; systematic scoping review; human factors; screening; clinicians; wearable; mobile phone

Introduction

Sensor-based digital health technologies (sDHTs), defined as connected digital medicine products that process data captured by mobile sensors using algorithms to generate measures of behavioral and/or physiological function [1], have been increasingly adopted in both research and health care in recent years [2,3]. sDHTs include products designed to capture data passively (such as continuous glucose monitors and wearables for monitoring sleep) or during active tasks (such as mobile spirometry or smartphone-based cognitive assessments) from wearable, implantable, ingestible, or ambient tools. Implementation of sDHTs requires interactions across the hardware containing the sensor or sensors, the software that is used to convert sensor data to health-related measures, and the users (who could be consumers, patients, clinicians, and more) who interact at one or more stages of data capture. Given this complexity and the increasing use of sDHTs, defining and understanding best practices for human factors, human-centered design, and usability (defined in Textbox 1) of sDHTs is a critical need. Although regulatory guidance focused on the usability of medical devices is well established, sDHTs require unique consideration because (1) sDHTs used in clinical research studies for data capture may or may not be regulated medical devices [4], (2) research participants likely have different motivations and needs related to their use of the technology, (3) sDHTs are often used over much longer time periods in research compared with health care settings, and (4) digital measures captured in large studies may be analyzed with limited human oversight or clinical interpretation.

The methodological approaches taken toward sDHT usability evaluation have varied substantially [5,6], casting light on the many challenges in implementing these tools into clinical trials and patient care at scale across diverse patient populations [7,8]. For example, some studies have adopted questionnaires developed for products and systems other than sDHTs [9], while others have described the approach to participatory design alongside qualitative data capture [10]. Inadequate attention to human-centered design and usability testing approaches can hinder the evaluation of health care interventions, contribute to insufficient adoption, perpetuate health disparities, increase costs, and potentially introduce safety risks [11-14]. Thus, integrating human factors considerations in the design, development, and evaluation of sDHTs is critical to improving their likelihood of being adopted and properly utilized in a way that is safe, effective, inclusive, and optimizes the user experience.

While several systematic reviews have focused on understanding and quantifying the usability of digital health products for specific applications [15-19], their focus has primarily been on study outcomes rather than evaluating methodological approaches. Recognizing the urgency of addressing sDHT usability-related challenges, a precompetitive collaboration within the Digital Health Measurement Collaborative Community (DATAcc) hosted by the Digital Medicine Society (DiMe) undertook a scoping review to highlight studies that have performed a usability-related evaluation for sDHTs, outline the dimensions of usability data that were assessed, and highlight the methods of usability evaluation. Our objective was to explore the current landscape and identify gaps, which will inform the development and dissemination of recommendations and an evidence-driven evaluation framework of sDHTs as being fit for purpose from a usability perspective.

XSL•FO
RenderX

Textbox 1. Definitions.

Human factors

- The application of knowledge about human behavior, abilities, limitations, and other characteristics of users to the design and development of a sensor-based digital health technology (sDHT) to optimize usability within a defined intended use or context of use. This definition incorporates terminology and concepts from the US Food and Drug Administration (FDA) [20], the UK Medicines and Health Care Products Regulatory Agency (MHRA) [21], and the National Medical Products Administration (NMPA) of China (translated) [22].

Human-centered design

- An approach to interactive systems that aims to make systems usable and useful by focusing on the users, their needs and requirements, and by applying human factors and usability knowledge and techniques, as defined in the International Organization for Standardization (ISO) 9241-210:2019 standard [23].

Usability

- The extent to which an sDHT can be used to achieve specified goals with ease, efficiency, and user satisfaction within a defined intended use or context of use. This definition incorporates terminology and concepts from the FDA [20], the MHRA [21], the NMPA (translated) [22], and ISO 9241-210:2019 [23].

Methods

Overview

We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for scoping reviews (Multimedia Appendix 1) [24]. As a scoping review, this work did not meet the criteria for registration on PROSPERO [25]. The protocol is available from the corresponding author.

Literature Search

We completed our literature search in PubMed using search terms designed in 6 layers as follows (terms within each layer were separated by the Boolean operator “OR”, while the layers themselves were separated using “AND” or “NOT”): (1) Medical Subject Heading (MeSH; [26]) term for human participants; (2) MeSH terms related to sDHTs, such as wearable electronic devices and digital technology; (3) keywords related to sDHTs such as wear* (asterisk indicates truncation), remote, and connected; (4) keywords related to human-centered design, usability, human factors, and ergonomics; (5) exclusion of out-of-scope publication types such as editorials and case reports; and (6) published between January 1, 2013, and May 30, 2023. The complete search string is provided in Table S1 in Multimedia Appendix 2.

To avoid potentially overlooking novel or emerging technologies, the search terms did not include descriptions of specific sensor types (such as accelerometer), form factors (such as watch), methodology (such as actigraphy), wear location (such as wrist), or technology make or model.

Study Selection

We systematically screened publications identified in the literature search based on the PICO (patients/participants; intervention; comparator; outcomes) eligibility criteria outlined in Table 1, designed to identify studies describing the incorporation of knowledge about human behavior, abilities, limitations, and other characteristics of users to the design and development process; human-centered design; and ease of use, efficiency, or user satisfaction of sDHTs. Studies reporting sDHT adherence (eg, average wear time) or measurement success metrics (eg, percentage of in-range measurements obtained) were considered out of scope unless they reported one of the aforementioned concepts.

Two independent investigators (JC and JPB) began by screening a random selection of 20% of publications; disagreements were resolved by consensus, and clarifications were made to the wording of the eligibility criteria to reduce ambiguity. The same two investigators then reviewed another random selection of 20% of publications; it was determined a priori that if the reviewers were in agreement for ≥90% of these publications, the remaining 60% would be reviewed by a single investigator (JC) as described elsewhere [27,28].

XSL•FO
RenderX

Table 1. Study selection eligibility criteria.

PICO[a] framework[b]	Eligibility criteria
Patient or participant	• Exclude studies that do not report data collected from human participants
Intervention	• Exclude studies that do not assess a specific sDHT[c], defined according to the definition of BioMeT in the V3 framework[d]: • Connected • Interpreted as a digital method of data transfer from the sDHT to the location of data analysis, either wired or wireless • Mobile • Interpreted as the tool being capable of collecting data in the out-of-clinic setting, although the study may have deployed the tool in clinic • Sensor-based • Interpreted as the tool containing at least one sensor sampling a physical construct such as acceleration, light, or temperature • Used for purposes of measurement, diagnosis, and/or treatment of a behavioral or physiological function
Comparator	• N/A[e]
Outcome or outcomes	• Exclude studies that do not report data on human factors, human-centered design, or usability (see Textbox 1 for definitions)

[a]PICO: patients/participants; intervention; comparator; outcomes.

[b]The PICO framework is described by Eriksen and Frandsen [29].

[c]sDHT: sensor-based digital health technology.

[d]Throughout this review, we refer to "sensor-based digital health technology" (sDHT); however, this was operationalized according to the definition of "biometric monitoring technology" (BioMeT) as described in Goldsack et al [1].

[e]Not applicable.

Data Extraction and Analysis

Data extraction fields included study design and sample characteristics; the type, maturity, make or model, form factor, and wear location (if applicable) of each sDHT evaluated along with the health concept or concepts generated by each sDHT; the methodological approaches; and the types of usability-related data reported in each study. Most fields for data capture were categorical, with categories created in advance to minimize error. Extraction from each publication was undertaken by one of three investigators with adjudication by an independent investigator as needed.

Categories of usability-related data are described in Table 2, and compiled based on the literature including the International Organization for Standardization (ISO) 9241-210:2019 standard [23] and Nielsen's [30] usability attributes, as well as the studies identified in this review; that is, data not clearly fitting into an existing category were extracted and categorized post hoc. We acknowledge that there are various models for capturing data describing usability and related topics [31]; however, there is no single standard that has been widely adopted.

Consistent with the goal of a scoping review, all data were analyzed descriptively.

XSL•FO
RenderX

Table 2. Categories of usability-related data extracted from eligible papers.

Category	Definition[a,b]
User satisfaction	The extent to which a user finds the sDHT[c] to be pleasant to use, which may reflect trust, comfort, aesthetics, engagement, desirability, emotional response, and other considerations. Always captured through self-report.
Ease of use	The ease with which a user is able to perform user tasks. Can be captured through self-report (such as the mental demand or effort required to complete a task) or objective measures (such as the number of actions, number of attempts, or time required to complete a task).
Efficiency	The ease with which a user is able to perform user tasks after having learned how to use the sDHT. Captured according to the definition of ease of use above.
Learnability[d]	The ease with which a user is able to perform user tasks during their first encounter with the sDHT. Captured according to the definition of ease of use above.
Memorability	The ease with which a user is able to perform user tasks after a period of nonuse, assessed in a test-retest paradigm. Captured according to the definition of ease of use above.
Usefulness[e]	The extent to which a user finds the sDHT, or its specific features or functions, to be valuable, productive, or helpful. Always captured through self-report.
Use errors[f]	An action or lack of action that may result in a use-related hazard (a potential source of harm), as well as error recovery defined as the ability of a user to make a correction following a use error in order to complete a task. Can be captured through self-report or objective assessments.
Technical performance or malfunctions	Technical performance, such as page load times, or the number, type, and severity of errors associated with sDHT malfunction. Can be captured through self-report or objective assessments.
Readability	The reading skills a user must possess to understand information presented to them through the sDHT itself, or through written materials such as instructions for use, cautions, warnings, or contraindications; [32]. Always captured through objective assessments, and typically reported as a reading grade.
Understandability or actionability	The extent to which users of diverse backgrounds, languages, and varying levels of health literacy understand (1) the clinical data or other information, such as instructions, cautions, warnings, and contraindications, presented to them; and (2) the actions or tasks they should complete in response, such as an sDHT-derived blood glucose measurement requiring an adjustment to medication [32]. Always captured through objective assessments.

[a]Note that in the definitions, "self-report" includes data captured through surveys, interviews, and focus groups, while "objective" includes data captured through observation (direct or video) or through the sDHT itself (or any related software) such as timestamps, app crash reports, and page load times.

[b]Comfort and trust were extracted separately for the purposes of this review.

[c]sDHT: sensor-based digital health technology.

[d]Learnability refers to the operation of the sDHT rather than a practice effect associated with a research study outcome or endpoint.

[e]We have adopted the term usefulness instead of utility, to avoid confusion with clinical utility, which refers to the extent to which implementing a medical product leads to improved health outcomes or provides useful information about diagnosis, treatment, management, or prevention of disease [33].

[f]Study outcomes that come after "use errors" are not typically considered usability data, but are related concepts often captured during usability evaluations.

Results

Literature Search and Study Selection

The PubMed search conducted on June 1, 2023, yielded 442 results, including one published only as an abstract. After applying the eligibility criteria described in Table 1, a further 356 publications were excluded. As such, 85 studies were determined to be eligible; however, 2 studies were under embargo, leaving 83 studies for data extraction (Figure 1). A complete list of all included studies is provided in Table S2 in Multimedia Appendix 2 [9,10,34-114].

As described above, two investigators reached a consensus on 20% (n=88) of the 442 publications, before any further publications were screened. The same investigators then screened a further 88 publications independently, which resulted in 100% agreement of eligibility. Per protocol, a single investigator screened the remaining 266 papers.

XSL•FO
RenderX

Figure 1. PRISMA flowchart. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; sDHT: sensor-based digital health technology.

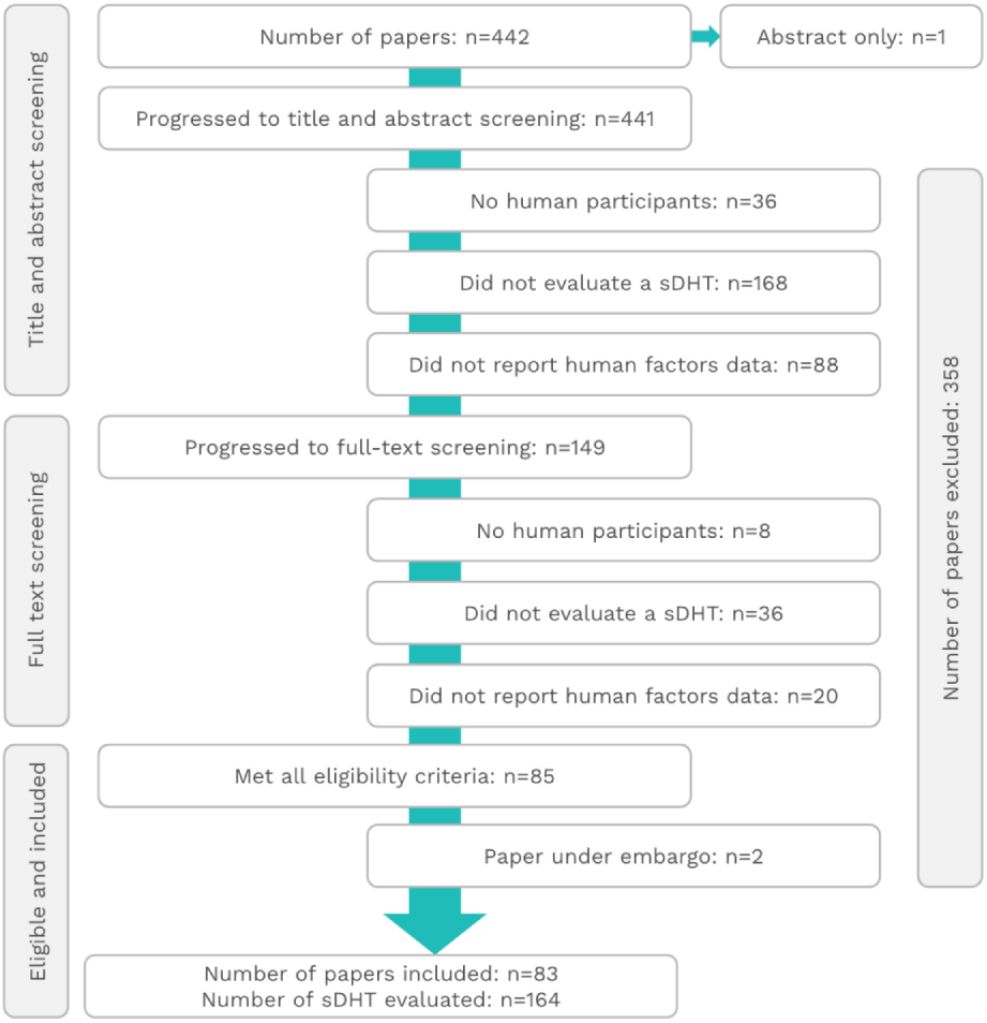

Study Design Considerations

The majority of studies (55/83, 66%; Table 3) reported summative evaluations of products that were marketed or production-equivalent (ie, sample products of final design assembled in a way that differs from—but is equivalent to—the manufacturing processes used for the marketed product [115]). The remaining 28 (34%) out of 83 studies reported formative evaluations of prototype products; we did not identify any reports focused solely on sDHT design. Most studies (53/83, 64%) were conducted partially or completely off-site. Study sample sizes spanned a wide range (range 1-623; median 27, IQR 13-60); however, only 21 (25%) of the full set of 83 studies, and 17 (33%) of the 55 summative studies, reported a rationale for the sample size (with or without a power calculation).

XSL•FO
RenderX

Table 3. Study design and sample characteristics across therapeutic areas.

	Therapeutic area of sDHT[a] end users (number of studies in parenthesis)									
	Aging (n=19)	Cardiovascular (n=9)	Endocrine (n=3)	Neurology (n=13)	Oncology (n=3)	Respiratory (n=6)	Surgery (n=5)	Healthy (n=15)	Other[b] (n=10)	Total (n=83)
Study design, n (%)										
Observational	17 (89)	9 (100)	3 (100)	13 (100)	3 (100)	5 (83)	4 (80)	14 (93)	10 (100)	78 (94)
Interventional	2 (11)	0 (0)	0 (0)	0 (0)	0 (0)	1 (17)	1 (20)	1 (7)	0 (0)	5 (6)
Study focus[c], n (%)										
Summative; sample size rationale	3 (16)	3 (33)	0 (0)	4 (31)	0 (0)	1 (17)	2 (40)	1 (7)	3 (30)	17 (20)
Summative; no sample size rationale	6 (32)	2 (22)	2 (67)	7 (54)	3 (100)	4 (67)	1 (20)	9 (60)	4 (40)	38 (46)
Formative; sample size rationale	2 (11)	1 (11)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (10)	4 (5)
Formative; no sample size rationale	8 (42)	3 (33)	1 (33)	2 (15)	0 (0)	1 (17)	2 (40)	5 (33)	2 (20)	24 (29)
Setting, n (%)										
Remote	11 (58)	3 (33)	1 (33)	5 (38)	2 (67)	6 (100)	1 (20)	7 (47)	6 (60)	42 (51)
On-site	4 (21)	6 (67)	1 (33)	5 (38)	0 (0)	0 (0)	3 (60)	8 (53)	3 (30)	30 (36)
Both remote and on-site	4 (21)	0 (0)	1 (33)	3 (23)	1 (33)	0 (0)	1 (20)	0 (0)	1 (10)	11 (13)
Duration of sDHT data collection, n (%)										
≤1 day	5 (26)	3 (33)	1 (33)	2 (15)	0 (0)	0 (0)	2 (40)	7 (47)	2 (20)	22 (27)
>1 to ≤7 days	2 (11)	3 (33)	0 (0)	4 (31)	0 (0)	2 (33)	2 (40)	2 (13)	2 (20)	17 (20)
>7 to ≤30 days	6 (32)	2 (22)	1 (33)	3 (23)	1 (33)	0 (0)	1 (20)	2 (13)	0 (0)	16 (19)
>31 to ≤90 days	3 (16)	0 (0)	0 (0)	2 (15)	1 (33)	4 (67)	0 (0)	3 (20)	1 (10)	14 (17)
>90 to ≤180 days	1 (5)	1 (11)	0 (0)	0 (0)	1 (33)	0 (0)	0 (0)	0 (0)	3 (30)	6 (7)
>180 days	1 (5)	0 (0)	0 (0)	2 (15)	0 (0)	0 (0)	0 (0)	0 (0)	1 (10)	4 (5)
Not reported	1 (5)	0 (0)	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	1 (10)	4 (5)
Study sample										
Sample size, median (IQR)	30 (13.5-52.5)	24 (10-41)	35 (20-189)	40 (22-70)	30 (22-31.5)	14.5 (9.25-19.75)	29 (15-60)	25 (13-105)	21 (12-81.25)	27 (13-60)
Sample size, range	8-125	5-156	5-343	5-623	14- 33	1-314	10-77	1-243	3-407	1-623
Users, n (%)										
End users[d]	19 (100)	9 (100)	3 (100)	13 (100)	3 (100)	6 (100)	5 (100)	15 (100)	9 (90)	82 (99)
Care partner users[d]	0 (0)	1 (11)	1 (33)	2 (15)	0 (0)	0 (0)	0 (0)	3 (20)	1 (10)	8 (10)
Clinician users[d]	2 (11)	3 (33)	1 (33)	2 (15)	0 (0)	1 (17)	1 (20)	1 (7)	1 (10)	12 (14)
Experts[d]	1 (5)	0 (0)	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	0 (0)	3 (4)
Age, n (%)										
Adults only	19 (100)	7 (78)	1 (33)	10 (77)	2 (67)	3 (50)	5 (100)	11 (73)	7 (70)	65 (78)
Children only	0 (0)	0 (0)	1 (33)	1 (8)	0 (0)	2 (33)	0 (0)	3 (20)	3 (30)	10 (12)
Both adults and children	0 (0)	1 (11)	0 (0)	2 (15)	1 (33)	1 (17)	0 (0)	1 (7)	0 (0)	6 (7)
Not reported	0 (0)	1 (11)	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)
Sex/gender, n (%)										
Male or men only	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	2 (20)	3 (4)

XSL•FO
RenderX

	Therapeutic area of sDHT[a] end users (number of studies in parenthesis)									
	Aging (n=19)	Cardio-vascular (n=9)	En-docrine (n=3)	Neurolo-gy (n=13)	Oncolo-gy (n=3)	Respira-tory (n=6)	Surgery (n=5)	Healthy (n=15)	Other[b] (n=10)	Total (n=83)
Female or women only	1 (5)	0 (0)	0 (0)	0 (0)	1 (33)	1 (17)	0 (0)	0 (0)	2 (20)	5 (6)
Both or all sexes/genders	18 (95)	7 (78)	2 (67)	13 (100)	2 (67)	5 (83)	5 (100)	13 (87)	6 (60)	71 (86)
Not reported	0 (0)	2 (22)	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	0 (0)	4 (5)
Race/ethnicity, n (%)										
Race/ethnicity reported	2 (11)	1 (11)	0 (0)	4 (31)	0 (0)	2 (33)	0 (0)	3 (20)	2 (20)	14 (17)
Race/ethnicity not reported	17 (89)	8 (89)	3 (100)	9 (69)	3 (100)	4 (67)	5 (100)	12 (80)	8 (80)	69 (83)
Number of sDHTs assessed										
Range	1-7	1-3	1-1	1-5	1-6	1-5	1-2	1-7	1-11	1-11

[a]sDHT: sensor-based digital health technology.

[b]"Other" therapeutic area category contains studies with enrollment eligibility focused on anaphylaxis, muscular dystrophy, hemophilia, nocturnal enuresis, blood and marrow transplant, overweight or obesity, pregnancy, and nonspecific hospitalized or chronic illness. One study recruited clinicians only (no end users of the sDHT) and is included in this category.

[c]Studies reporting formative and summative evaluations are categorized as summative.

[d]Categories are not mutually exclusive.

Sample Characteristics

As shown in Table 3, the largest target populations were focused on aging and healthy participants (15 and 19 studies, respectively; 34/83, 41% of all studies). Among the various diseases studied, neurology and cardiovascular were the most common therapeutic areas (13 and 9 studies, respectively; 22/44, 50% of studies assessing nonhealthy individuals). Table S3 in Multimedia Appendix 2 contains a list of conditions falling into each therapeutic area.

Almost all studies (82/83, 99%) captured data from targeted end users; the remaining study captured data only from clinician users [63]. Several studies captured data from multiple user groups; in total, 8 and 12 studies gathered data from care partner users and clinician users, respectively. Three studies involved experts (not considered to be sDHT users); two of these described a formal heuristic evaluation [79,99] while the other described involving experts in design, biomedical engineering, computer science, and mobile health system production in the sDHTs design and formative testing process [103]. Finally, we noted substantial missing participant demographic data; age, sex/gender, and race/ethnicity were not reported in 2, 4, and 69 studies, respectively.

sDHTs Assessed in Eligible Studies

Across the 83 studies included in our review, a total of 164 different sDHTs were assessed (141 wearable and 23 ambient tools; Table 4), ranging from 1 to 11 sDHTs within a single study. Ingestible and implantable sDHTs were in scope, but none were identified in our literature search. A wide range of form factors (22 distinct categories) and wear locations (14 anatomical locations presented in 5 categories) were identified. Digital clinical measures of vital signs (n=76 sDHTs), physical activity (n=61 sDHTs), and mobility (n=35) were most prevalent. Table S4 in Multimedia Appendix 2 contains more comprehensive information regarding wear locations and health concepts captured by sDHTs.

Most sDHTs (126/164, 77%) required only passive interaction by users, meaning that data were captured without user input other than basic tasks such as charging or changing batteries. The remaining 38 (23%) sDHTs required active engagement at specific times such as completion of physical therapy [58], exercise [97], or blood glucose tests [79].

XSL•FO
RenderX

Table 4. sDHT[a] descriptive information across therapeutic areas.

	Therapeutic area of sDHT users (number of sDHTs in parenthesis)									
	Aging (n=27)	Cardiovascular (n=12)	Endocrine (n=3)	Neurology (n=31)	Oncology (n=8)	Respiratory (n=15)	Surgery (n=7)	Healthy (n=35)	Other[b] (n=26)	Total (n=164)
sDHT type, n (%)										
Wearable	26 (96)	9 (75)	1 (33)	30 (97)	8 (100)	11 (73)	7 (100)	33 (94)	16 (62)	141 (86)
Ambient	1 (4)	3 (25)	2 (67)	1 (3)	0 (0)	4 (27)	0 (0)	2 (6)	10 (38)	23 (14)
sDHT maturity, n (%)										
Prototype	9 (33)	5 (42)	1 (33)	5 (16)	0 (0)	0 (0)	2 (29)	5 (14)	11 (42)	38 (23)
Final or marketed	18 (67)	6 (50)	2 (67)	26 (84)	8 (100)	15 (100)	5 (71)	27 (77)	12 (46)	119 (73)
Not reported	0 (0)	1 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (9)	3 (12)	7 (4)
Form factor, n (%)										
Adhesive patch	2 (7)	0 (0)	1 (33)	5 (16)	0 (0)	0 (0)	1 (14)	2 (6)	1 (4)	12 (7)
Balance board	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	1 (<1)
Camera, video, or still	0 (0)	1 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (8)	3 (2)
Clip	4 (15)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	1 (14)	1 (3)	0 (0)	7 (4)
Clothing or shoes	5 (19)	3 (25)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (11)	5 (19)	17 (10)
Contact lens	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (<1)
Cuff or wrap	0 (0)	1 (8)	0 (0)	1 (3)	0 (0)	0 (0)	1 (14)	2 (6)	0 (0)	5 (3)
Electrode or electrodes	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	2 (6)	0 (0)	3 (2)
Exercise equipment	0 (0)	0 (0)	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (8)	3 (2)
Glasses	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	2 (<1)
Gloves	2 (7)	1 (8)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (2)
Glucometer	0 (0)	0 (0)	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (<1)
Handheld thermometer	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	0 (0)	0 (0)	0 (0)	1 (<1)
Mattress pad	0 (0)	1 (8)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (<1)
Medication package	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	0 (0)	0 (0)	1 (4)	2 (<1)
Phone or tablet	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (3)	4 (15)	5 (3)
Probe	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	2 (<1)
Ring	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	0 (0)	0 (0)	0 (0)	1 (<1)
Spirometer	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (13)	0 (0)	0 (0)	0 (0)	2 (<1)
Contactless unit	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	2 (<1)
Strap	12 (44)	4 (33)	0 (0)	20 (65)	8 (100)	9 (60)	4 (57)	21 (60)	9 (35)	87 (53)
Weight scale	0 (0)	1 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (<1)
Wear location, n (%)										
Arms or wrists or hands	11 (41)	5 (42)	0	19 (61)	8 (100)	8 (53)	3 (43)	19 (54)	9 (35)	82 (50)
Head or face	1 (4)	0	0	4 (13)	0	0	0	3 (9)	0	8 (5)
Legs or ankles or feet	2 (7)	2 (17)	0 (0)	1 (3)	0 (0)	0 (0)	2 (29)	3 (9)	0 (0)	10 (6)
Neck or torso or hips	10 (37)	2 (17)	1 (33)	3 (10)	0 (0)	2 (13)	2 (29)	6 (17)	6 (23)	32 (20)

XSL•FO
RenderX

	Therapeutic area of sDHT users (number of sDHTs in parenthesis)									
	Aging (n=27)	Cardiovascular (n=12)	Endocrine (n=3)	Neurology (n=31)	Oncology (n=8)	Respiratory (n=15)	Surgery (n=7)	Healthy (n=35)	Other[b] (n=26)	Total (n=164)
Multiple locations[c]	2 (7)	0 (0)	0 (0)	3 (10)	0 (0)	0 (0)	0 (0)	2 (6)	1 (4)	8 (5)
N/A[d]	1 (4)	3 (25)	2 (67)	1 (3)	0 (0)	5 (33)	0 (0)	2 (6)	10 (38)	24 (15)
Interaction type[e], n (%)										
Passive	24 (89)	7 (58)	1 (33)	26 (84)	8 (100)	12 (80)	3 (43)	30 (86)	15 (58)	126 (77)
Active	3 (11)	5 (42)	2 (67)	5 (16)	0 (0)	3 (20)	4 (57)	5 (14)	11 (42)	38 (23)
Health concepts[f], n (%)										
Activities of daily living	6 (22)	2 (17)	0 (0)	2 (6)	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	11 (7)
Physical activity	16 (59)	4 (33)	1 (33)	6 (19)	9 (113)	9 (6)	1 (14)	15 (43)	0 (0)	61 (37)
Adherence	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (7)	0 (0)	0 (0)	0 (0)	1 (<1)
Electrical activity	0 (0)	0 (0)	0 (0)	15 (48)	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	16 (<1)
Mobility	5 (19)	4 (33)	0 (0)	11 (35)	0 (0)	0 (0)	2 (29)	13 (37)	0 (0)	35 (21)
Sleep	6 (22)	1 (8)	0 (0)	0 (0)	1 (13)	1 (7)	0 (0)	5 (14)	0 (0)	14 (9)
Vital signs	9 (33)	5 (42)	2 (67)	18 (58)	0 (0)	5 (33)	14 (2)	23 (66)	0 (0)	76 (46)
Other[g]	4 (15)	2 (17)	0 (0)	4 (13)	0 (0)	2 (13)	0 (0)	0 (0)	0 (0)	12 (7)

[a]sDHT: sensor-based digital health technology.

[b]"Other" therapeutic area category contains studies with enrollment eligibility focused on anaphylaxis, muscular dystrophy, hemophilia, nocturnal enuresis, blood and marrow transplant, overweight or obesity, pregnancy, and nonspecific hospitalized or chronic illness. One study recruited clinicians only (no end users of the sDHT) and is included in this category.

[c]Refers to multisensor sDHTs worn on different parts of the body, or sDHTs that can be positioned in one of many locations.

[d]Wear location is not applicable to ambient sDHTs. Wear locations are presented in greater detail in Table S4 in Multimedia Appendix 2.

[e]Passive: sDHT data are collected over long time periods without user input other than aspects such as charging or changing batteries (such as actigraphy); includes tools for which the absence of data is meaningful (such as smart packaging for adherence monitoring). Active: sDHT data collection requires user engagement at defined timepoints. Categories described previously [28].

[f]Health concepts are not mutually exclusive; a single sDHT can capture data in multiple categories. Heath concepts are presented in greater detail in Table S4 in Multimedia Appendix 2.

[g]"Other" health concept category includes bladder volume, body habitus, cardiac output, fall detection, gaze or visual movement, intraocular pressure, lung or airway function, and tremor detection.

Methodological Approaches

As described in Table 5, most sDHTs (139/164, 85%) were evaluated in the actual environment in which they were intended to be used, while 25 sDHTs were assessed in a simulated environment only. The vast majority were evaluated during actual use (148/164, 90%) rather than through "look and feel" approaches. Of particular interest, a variety of methods were used to evaluate usability and related concepts, including interviews (49 sDHTs), focus groups (29 sDHTs), direct or video observation (35 sDHTs), think-aloud (15 sDHTs), and heuristic analysis (2 sDHTs). Surveys were the most prevalent method for capturing usability data; 86 sDHTs were evaluated using referenced surveys while 81 sDHTs were evaluated using surveys developed in house by study investigators. Data for 4 sDHTs were captured using the sDHT itself; for example, instances of connectivity loss or data capture drops were recorded as "use errors" or "technical performance or product errors" [42,100]. An illustrative example of a use error that may be addressed through design modification is the report of users turning an sDHT on and off repeatedly as it was not clear whether the product was operating correctly [53]. Additional examples of product errors, distinct from use errors, included instances of system crash [51] and software malfunctions requiring computer program patches [54].

XSL•FO
RenderX

Table 5. Methodological approaches to usability data collection.

	sDHT[a] type (number of sDHTs in parenthesis)		
	Ambient (n=23)	Wearable (n=141)	Total (n=164)
Data collection environment, n (%)			
Actual environment	23 (100)	107 (76)	130 (79)
Simulated environment	0 (0)	25 (18)	25 (15)
Both actual and simulated	0 (0)	9 (6)	9 (5)
Interactions with sDHT, n (%)			
Look and feel	1 (4)	15 (11)	16 (10)
Actual use	22 (96)	126 (89)	148 (90)
Usability evaluation methods[b]			
Interviews	5 (22)	44 (31)	49 (30)
Focus groups	10 (43)	19 (13)	29 (18)
Surveys—referenced	14 (61)	72 (51)	86 (52)
Surveys—in house	7 (30)	74 (52)	81 (49)
Think-aloud	1 (4)	14 (10)	15 (9)
Observation (direct or video)	1 (4)	34 (24)	35 (21)
Measured by the sDHT	0 (0)	5 (4)	5 (3)
Heuristic analysis	1 (4)	1 (<1)	2 (<1)
Type or types of usability data reported, n (%)			
Mixed methods	14 (61)	58 (41)	72 (44)
Quantitative only	6 (26)	60 (43)	66 (40)
Qualitative only	3 (13)	23 (16)	26 (16)
Categories of usability and related data reported, n (%)			
User satisfaction	19 (83)	117 (83)	136 (83)
Comfort	5 (22)	107 (76)	112 (68)
Ease of use; self-report	23 (100)	122 (87)	145 (88)
Ease of use; objectively captured	1 (4)	4 (3)	5 (3)
Learnability	1 (4)	10 (7)	11 (7)
Efficiency	0 (0)	4 (3)	4 (2)
Memorability	0 (0)	2 (<1)	2 (<1)
Usefulness	16 (70)	96 (68)	112 (68)
Use errors	6 (26)	26 (18)	32 (20)
User trust	12 (52)	53 (38)	65 (40)
Readability	0 (0)	0 (0)	0 (0)
Understandability or actionability	1 (4)	13 (9)	14 (9)
Technical performance or product errors	19 (83)	79 (56)	98 (60)
Adherence to sDHT reported, n (%)			
Objectively measured by the sDHT	6 (26)	44 (31)	50 (30)
Self or care partner report	1 (4)	14 (10)	15 (9)
Both objective and self or care partner	0 (0)	4 (3)	4 (2)
Reported but method not described	1 (4)	7 (5)	8 (5)
Adherence not reported	15 (65)	72 (51)	87 (53)

XSL•FO
RenderX

[a]sDHT: sensor-based digital health technology.

[b]Categories are not mutually exclusive.

Categories of Usability-Related Data Reported

User satisfaction was captured for the majority of sDHTs (136/164, 83%), often as a measure of acceptability or user attitudes. Although overall ease of use was also commonly reported, captured through either self-report or objective methods (n=145 and n=5, respectively), the related concepts of learnability, efficiency, and memorability were reported for only 11, 4, and 2 sDHTs, respectively. Technical performance and product errors associated with malfunction were captured for 98 sDHTs, while use errors were captured for only 32 sDHTs. Finally, although none of the studies in our review reported the readability of information presented to the user, 14 sDHTs were evaluated according to the extent to which users were able to understand the data or information presented to them (understandability) or the actions or tasks they should complete in response (actionability).

Finally, adherence (such as wear or use time) was reported for 77 sDHTs. Of these, 50 sDHTs captured adherence data objectively, adherence to 19 sDHTs was assessed through self-report or care partner report, and the method was not described for 8 sDHTs.

The complexity of the relationships in our dataset comparing usability evaluation methods with sDHT form factor, and comparing usability evaluation methods with the categories of usability-related data reported, are depicted in Figures 2 and 3, respectively. For example, the width of each chord in Figure 2 is proportional to the number of sDHTs of the relevant form factor that were assessed using the linked method, demonstrating that surveys (both referenced and in house) were the most common evaluation methods while heuristic analysis was the least common. Similarly, Figure 3 demonstrates that overall satisfaction and self-reported ease of use were captured frequently, in contrast to data related to objective ease of use, efficiency, learnability, and memorability. Both figures contain a large number of linked chords, indicating that specific usability evaluation methods were adopted across diverse sDHT form factors and outcome measures.

Tables S5-S7 in Multimedia Appendix 2 present the data shown in Tables 3-5 for the subset of 55 studies reporting the results of summative evaluations, while Tables S8-S10 in Multimedia Appendix 2 present these data for the subset of 28 studies reporting formative evaluations.

Figure 2. Chord diagram depicting the relationship between sDHT form factors and usability evaluation methods. sDHT: sensor-based digital health technology.

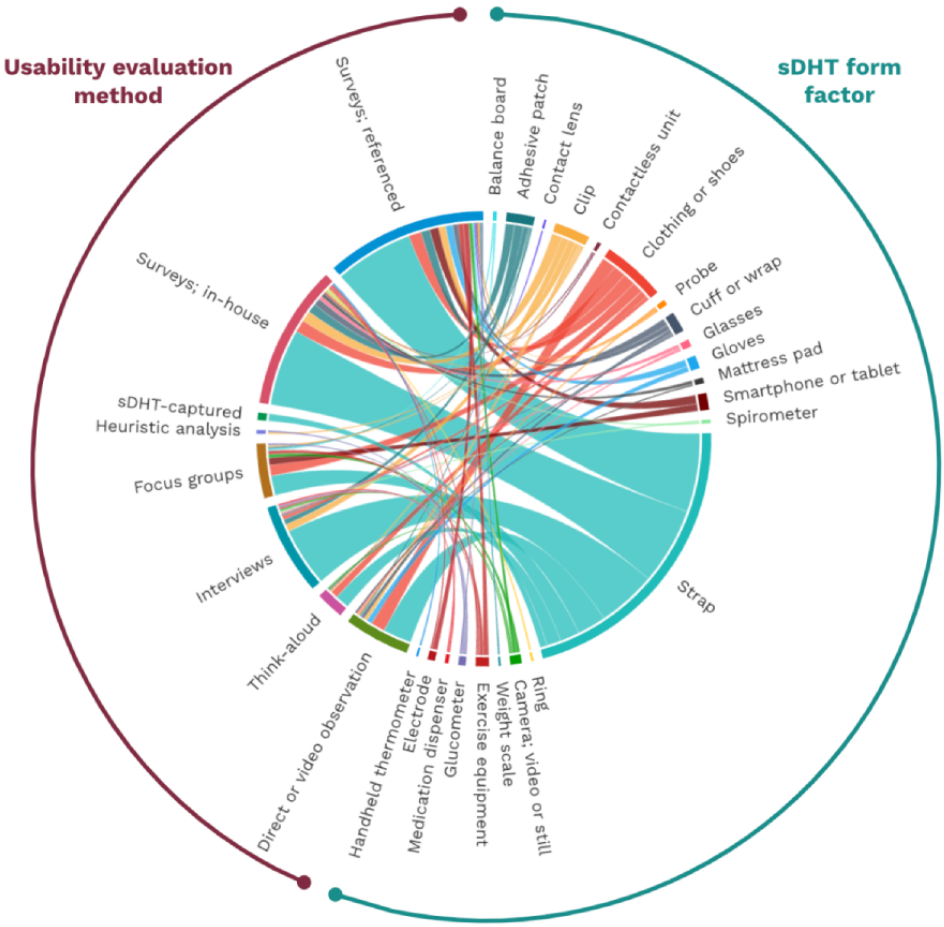

XSL•FO
RenderX

Figure 3. Chord diagram depicting the relationship between usability evaluation methods and categories of usability-related data reported. sDHT: sensor-based digital health technology.

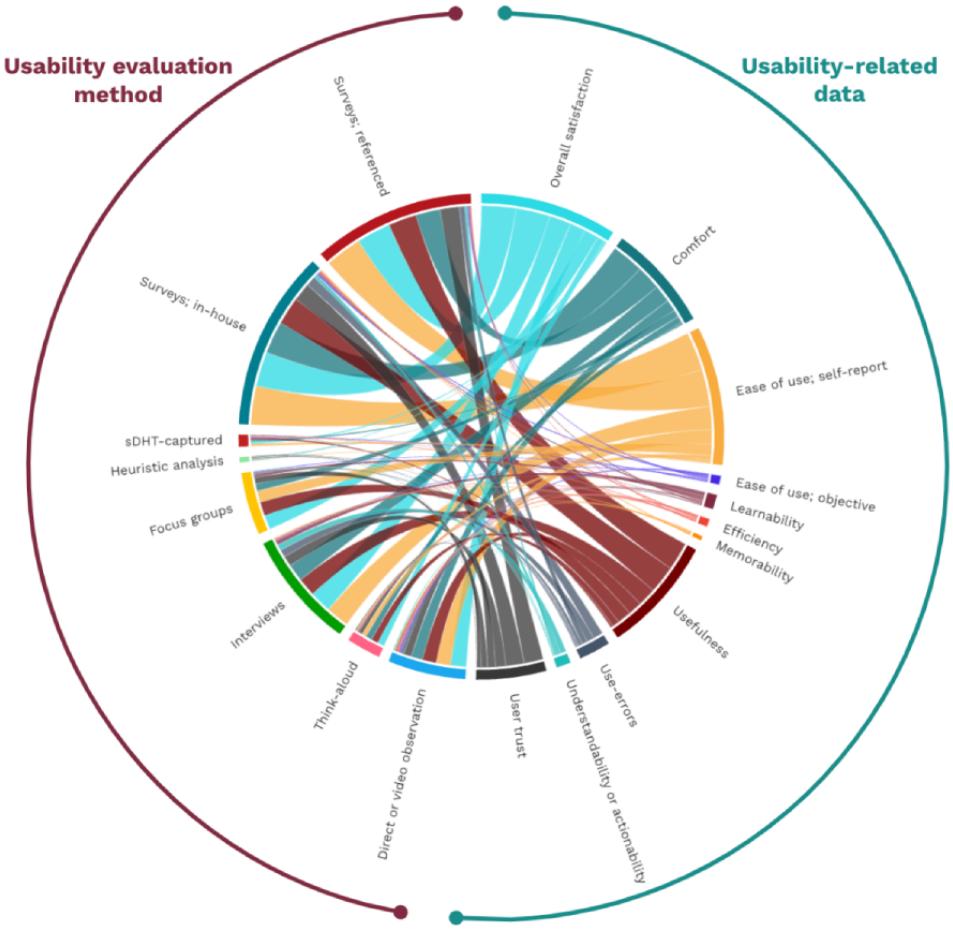

Discussion

Principal Findings

This paper represents the first scoping review reporting the methodological approaches adopted during usability-related studies specifically focused on sDHTs. We identified 83 formative and summative studies published over the decade from 2013 to 2023 that evaluated human factors, human-centered design, or usability for 164 ambient and wearable tools. Most studies (67/83, 81%) recruited nonhealthy individuals, thereby providing informative data regarding sDHT usability across many diseases in addition to other aspects of health such as aging and pregnancy. Most sDHTs were evaluated in the intended use environment, with multiple facets of usability-related data captured via a range of mixed method approaches including heuristic analysis, surveys, observation, think-aloud, focus groups, interviews, and use errors or technical performance errors captured by the sDHT itself such as instances of connectivity loss.

This review highlights 4 notable gaps that warrant attention as the field advances. First, the breadth and scope of usability and related data were fairly simplistic, relying largely on surveys capturing user satisfaction and ease of use (each captured for >80% of sDHTs) with limited reporting of sDHT use errors, learnability, efficiency, or memorability. The extent to which users understood the health- and behavior-related data or other information presented to them (understandability) and the actions or tasks they should complete in response (actionability) was assessed for only 9% (14/164) of sDHTs. Understandability and actionability are particularly important for sDHTs, given that they are often used by patients or participants in out-of-clinic settings without clinical supervision. For the use of sDHTs in clinical care settings, it is imperative that users understand whether and how to react to clinical data [116], and thus the lack of focus on understandability and actionability is concerning and could be due to the early-stage nature of sDHTs in clinical practice. In the context of clinical research, however, sharing sDHT data with participants in real time has the potential to introduce bias and affect user behavior, thereby posing a risk of yielding inaccurate results [4]. Additional dimensions related to understandability and actionability, such as understanding optimal ways of implementing remote examinations, also warrant further investigation.

Second, only 22% (18/83) of studies considered users other than end users (patients or participants), such as care partners and clinicians, who play crucial roles in sDHT implementation and therefore the quality of data captured [117]. Especially in populations where care partners play a key role in sDHT implementation (eg, children, older people, those with language barriers, and those with disabilities), understanding usability from the care partner perspective is vital. Although existing usability data may be available for some sDHTs that are regulated as medical devices, research participants likely have

XSL•FO
RenderX

needs and motivations for using the sDHT that differ from patients using the product as part of usual care. Similarly, the needs of investigator users are likely different from the needs of clinician users, requiring further evaluation.

Third, we found that only 31% (17/55) of summative studies (referred to by the US Food and Drug Administration as "human factors validation studies"; [20]) provided a rationale for the sample size, with or without a power calculation. An understanding of key study design considerations, including sample size, is important for evaluating the robustness of study conclusions.

Finally, as has been noted previously [28,118], we observed a deficiency in reporting basic sample demographics, with studies typically providing information on age and sex/gender but neglecting to include details on the race and ethnicity of participants. Inadequate reporting of descriptive data, including sociodemographics, precludes a complete understanding of generalizability, potentially leading to the need to repeat studies while contributing to disparities and biases in clinical research [119].

As described above, while there are several existing systematic reviews describing the usability of digital health products for specific applications [15-19], few have focused specifically on evaluating methodological approaches. In addition, most prior systematic reviews with similar objectives have focused on digital health technologies that are not sensor-based, such as electronic medical records systems [120] and mobile clinical decision support tools [121], that are not used for remote data capture. In 2023, Maqbool and Herold [5] published a systematic review of usability evaluations describing a broad suite of over 1000 digital health tools consisting mostly of mobile health apps and including a subset of 20 products approximately aligned to our definition of sDHT, including fitness or activity trackers, digital sphygmomanometers, and wearable fall risk assessment systems. Compared to this study, Maqbool and Herold [5] found relatively increased rates of clinician and care partner participation, and reporting of learnability, efficiency, and memorability. Such differences emphasize substantial variability in usability study methodology across subcategories of digital health technologies, as well as differences in definitions and terminology of the concepts reported, underscoring the need for a common evaluation framework.

Strengths and Limitations

Strengths of this review include the robust approach taken to testing our search terms including a careful assessment against a list of target papers identified a priori to ensure that we were capturing appropriate literature. This process was intended to not only ensure the inclusivity of relevant literature but also the reliability of our findings to help provide a foundation for subsequent reviews and meta-analyses. In-depth data extraction across many domains allowed for a thorough comparative analysis of the identified studies. The decision to focus on studies published within the last decade (2013-2023) was also carefully considered, as it encompasses the recent surge in studies reporting sDHT implementation. While sDHTs have a lengthy history prior to 2013, this temporal scope ensures that our findings reflect contemporary developments and trends, offering insights into the current state of sDHT implementation.

A number of limitations are acknowledged. First, we limited our search to the peer-reviewed literature. We acknowledge that many usability studies undertaken by technology manufacturers may be published in the gray literature; however, our ultimate goal is to use the findings of our review to guide the development of a framework representing best practices, and therefore, the peer review process was used as an indicator of methodological rigor and reporting quality. Second, terminology in the field of digital medicine is still evolving and investigators use many different terms to describe sDHTs; by incorporating 25 descriptive keywords in Layer C of our search terms (Table S1 in Multimedia Appendix 2), we found it necessary to rely on MeSH terms developed by the National Library of Medicine [26] as a means of limiting our literature search to a feasible number of publications. As a consequence, we were limited to conducting our search in PubMed as this is the clinical research database for the National Library of Medicine. While MeSH terms are widely accepted and systematically applied, their specificity may have excluded relevant studies using different terminology potentially resulting in unintentional omissions. In addition, the decision to search within one database may have resulted in missed publications. Our hope is that as the field matures, terminology will become harmonized and sDHT-specific indexing will support the identification of studies adopting these technologies. Third, our decision to exclude descriptions of specific sensors, form factors, methodologies, wear location, and technology make or model may have excluded publications that used these types of keywords in the absence of other descriptors and MeSH terms. This approach was taken to reduce the possibility of overlooking novel or emerging technologies in favor of established digital products such as actigraphy tools. Finally, only 40% (176/442) of publications were screened for eligibility by multiple investigators. This approach to study identification, which has been described and adopted previously [27,28], allowed us to screen a greater number of papers which was necessary given the lack of systematic indexing. The high agreement levels between investigators suggest that our quality-control approach maintained a robust screening process, despite part of the work being conducted by a single investigator.

Conclusions and Future Directions

Based on our findings, we suggest 4 actionable recommendations that will help to advance the implementation of sensor-based digital measurement tools in both clinical and research settings. First, we encourage investigators to adopt in-depth assessment and reporting of usability data beyond user satisfaction and ease of use. In particular, it is valuable to understand use errors alongside technical errors, and it is critical to evaluate the extent to which users understand the clinical data and information presented to them and the appropriate tasks to undertake in response, if applicable. Second, it is essential to embrace the diversity of users in all respects, including diversity of stakeholders within the human-centered design process; evaluation of usability across multiple user groups including care partners and clinicians; and ensuring that the participating users are generalizable to the intended use

XSL•FO
RenderX

population in terms of sociodemographics, social determinants of health, and other characteristics. Third, rigorous study design is key. Usability is a heterogeneous concept, and it is often beneficial to evaluate usability alongside other objectives such as analytical or clinical validation; thus, we do not advocate a particular study design or set of study outcome measures. We do, however, believe that careful consideration of usability evaluation criteria, study sample sizes, and predetermined thresholds of success is critical for making go or no-go decisions as to whether a particular sDHT is sufficiently usable for implementation in a particular context of use. Finally, we recommend adhering to reporting and publication checklists such as Annex B in ISO 9241-11:2018 [122] and EVIDENCE [123], the latter of which describes optimal reporting requirements of studies evaluating several aspects of sDHT quality including usability assessments. Ensuring consistency in reporting will enable meaningful comparisons between studies, facilitate better assessments of findings, and enhance the accurate interpretation of results and limitations across studies.

Our long-term goal is to develop and disseminate an evidence-driven framework for evaluating sDHTs as being fit for purpose from a usability perspective, informed in part by the findings of this review. By developing such a framework, we endeavor to contribute to the ongoing discourse surrounding sDHTs, ultimately paving the way for the development of safe and effective tools that lead to a more inclusive and patient-centric health care ecosystem poised to improve clinical trials and clinical practice.

Acknowledgments

The authors wish to acknowledge Bethanie McCrary and Danielle Stefko for assistance with project management, Katerina Djambazova for assistance with data extraction, and Jennifer Goldsack for providing feedback on the manuscript. Chord diagrams were created with flourish.studio online software. This work was undertaken within the Digital Health Measurement Collaborative Community (DATAcc), hosted by the Digital Medicine Society (DiMe).

Authors' Contributions

All authors contributed to the study design, data interpretation, and manuscript preparation. The literature search, literature screening, and data extraction were undertaken by JC, SM, and JPB.

Conflicts of Interest

AT is a consultant for Synergen Technology Labs, LLC; Siemens Healthineers; and Gabi SmartCare. BC is an employee of Genentech, a member of the Roche Group and Roche Pharmaceuticals, and owns company stock. EI is an employee of Koneksa Health and may own company stock. NM and SV are employees of Johnson & Johnson Innovative Medicine and hold company stocks or stock options. JPB reports financial interests (consulting income, shares, or stock) in Philips, Signifier Medical Technologies, Koneksa Health, and Apnimed. ES serves on the editorial board as an Associate Editor of *JMIR Publications*.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 162 KB - jmir_v26i1e57628_app1.pdf]

Multimedia Appendix 2
Supplementary tables.
[DOCX File , 140 KB - jmir_v26i1e57628_app2.docx]

References

1. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digital Med 2020;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
2. Library of digital endpoints. Digital Medicine Society (DiMe). 2021. URL: https://dimesociety.org/get-involved/library-of-digital-endpoints/ [accessed 2024-01-23]
3. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digital Med 2022;5(1):13 [FREE Full text] [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
4. Center for Drug Evaluation, Research. Digital health technologies for remote data acquisition in clinical investigations. US Food and Drug Administration. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations [accessed 2024-01-23]
5. Maqbool B, Herold S. Potential effectiveness and efficiency issues in usability evaluation within digital health: a systematic literature review. J Syst Software 2024;208:111881 [FREE Full text] [doi: 10.1016/j.jss.2023.111881]

XSL•FO
RenderX

6. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digital Med 2019;2:38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
7. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023;25:e42719 [FREE Full text] [doi: 10.2196/42719] [Medline: 36853742]
8. Bent B, Sim I, Dunn JP. Digital medicine community perspectives and challenges: survey study. JMIR Mhealth Uhealth 2021;9(2):e24570 [FREE Full text] [doi: 10.2196/24570] [Medline: 33533721]
9. Evans J, Papadopoulos A, Silvers CT, Charness N, Boot WR, Schlachta-Fairchild L, et al. Remote health monitoring for older adults and those with heart failure: adherence and system usability. Telemed e-Health 2016;22(6):480-488 [FREE Full text] [doi: 10.1089/tmj.2015.0140] [Medline: 26540369]
10. Ummels D, Braun S, Stevens A, Beekman E, Beurskens A. Measure it super simple (MISS) activity tracker: (re)design of a user-friendly interface and evaluation of experiences in daily life. Disabil Rehabil Assist Technol 2022;17(7):767-777. [doi: 10.1080/17483107.2020.1815089] [Medline: 32970493]
11. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. J Med Internet Res 2023;25:e43251 [FREE Full text] [doi: 10.2196/43251] [Medline: 36961506]
12. Levander XA, VanDerSchaaf H, Barragán VG, Choxi H, Hoffman A, Morgan E, et al. The role of human-centered design in healthcare innovation: a digital health equity case study. J Gen Intern Med 2024;39(4):690-695. [doi: 10.1007/s11606-023-08500-0] [Medline: 37973709]
13. Benishek LE, Kachalia A, Biddison LD. Improving clinician well-being and patient safety through human-centered design. JAMA 2023;329(14):1149-1150. [doi: 10.1001/jama.2023.2157] [Medline: 36821124]
14. Tase A, Vadhwana B, Buckle P, Hanna GB. Usability challenges in the use of medical devices in the home environment: a systematic review of literature. Appl Ergon 2022;103:103769 [FREE Full text] [doi: 10.1016/j.apergo.2022.103769] [Medline: 35430424]
15. Ye B, Chu CH, Bayat S, Babineau J, How T, Mihailidis A. Researched apps used in dementia care for people living with dementia and their informal caregivers: systematic review on app features, security, and usability. J Med Internet Res 2023;25:e46188 [FREE Full text] [doi: 10.2196/46188] [Medline: 37824187]
16. Siette J, Dodds L, Sharifi F, Nguyen A, Baysari M, Seaman K, et al. Usability and acceptability of clinical dashboards in aged care: systematic review. JMIR Aging 2023;6:e42274 [FREE Full text] [doi: 10.2196/42274] [Medline: 37335599]
17. Kraaijkamp JJM, van Dam van Isselt EF, Persoon A, Versluis A, Chavannes NH, Achterberg WP. eHealth in geriatric rehabilitation: systematic review of effectiveness, feasibility, and usability. J Med Internet Res 2021;23(8):e24015 [FREE Full text] [doi: 10.2196/24015] [Medline: 34420918]
18. Keogh A, Argent R, Anderson A, Caulfield B, Johnston W. Assessing the usability of wearable devices to measure gait and physical activity in chronic conditions: a systematic review. J Neuroeng Rehabil 2021;18(1):138 [FREE Full text] [doi: 10.1186/s12984-021-00931-2] [Medline: 34526053]
19. Butler S, Sculley D, Santos DS, Fellas A, Gironès X, Singh-Grewal D, et al. Usability of eHealth and mobile health interventions by young people living with juvenile idiopathic arthritis: systematic review. JMIR Pediatr Parent 2020;3(2):e15833 [FREE Full text] [doi: 10.2196/15833] [Medline: 33258786]
20. Center for Devices, Radiological Health. Applying human factorsusability engineering to medical devices. US Food and Drug Administration. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices [accessed 2024-01-23]
21. Department of Health and Social Care. Guidance on applying human factors and usability engineering to medical devices including drug-device combination products in Great Britain. Medicines and Healthcare products Regulatory Agency. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/970563/Human-Factors_Medical-Devices_v2.0.pdf [accessed 2024-01-30]
22. Guiding principles for technical review of human factors design of medical devices [DiMe, Trans.]. National Medical Products Association. URL: https://datacc.dimesociety.org/wp-content/uploads/2023/09/NMPA-Human-Factors-Guidance-English-Translation-FINAL.pdf [accessed 2024-01-30]
23. ISO. ISO 9241-210. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2024-01-23]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. PROSPERO. URL: https://www.crd.york.ac.uk/prospero/ [accessed 2024-01-23]
26. Medical subject headings—home page. U.S. National Library of Medicine. 2020. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2024-01-30]
27. Nussbaumer-Streit B, Sommer I, Hamel C, Devane D, Noel-Storr A, Puljak L, et al. Rapid reviews methods series: guidance on team considerations, study selection, data extraction and risk of bias assessment. BMJ Evidence-Based Med 2023;28(6):418-423 [FREE Full text] [doi: 10.1136/bmjebm-2022-112185] [Medline: 37076266]
28. Olaye IM, Belovsky MP, Bataille L, Cheng R, Ciger A, Fortuna KL, et al. Recommendations for defining and reporting adherence measured by biometric monitoring technologies: systematic review. J Med Internet Res 2022;24(4):e33537 [FREE Full text] [doi: 10.2196/33537] [Medline: 35436221]

XSL•FO
RenderX

29. Eriksen MB, Frandsen TF. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. J Med Libr Assoc 2018;106(4):420-431 [FREE Full text] [doi: 10.5195/jmla.2018.345] [Medline: 30271283]
30. Nielsen J. Usability Engineering. Somerville, Massachusetts: Boston Academic Press; 1993.
31. Shachak A, Kuziemsky C, Petersen C. Beyond TAM and UTAUT: future directions for HIT implementation research. J Biomed Inform 2019;100:103315 [FREE Full text] [doi: 10.1016/j.jbi.2019.103315] [Medline: 31629923]
32. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
33. Bossuyt PMM, Reitsma JB, Linnet K, Moons KGM. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 2012;58(12):1636-1643. [doi: 10.1373/clinchem.2012.182576] [Medline: 22730450]
34. Weenk M, Bredie SJ, Koeneman M, Hesselink G, van Goor H, van de Belt TH. Continuous monitoring of vital signs in the general ward using wearable devices: randomized controlled trial. J Med Internet Res 2020;22(6):e15471 [FREE Full text] [doi: 10.2196/15471] [Medline: 32519972]
35. Bentley CL, Powell L, Potter S, Parker J, Mountain GA, Bartlett YK, et al. The use of a smartphone app and an activity tracker to promote physical activity in the management of chronic obstructive pulmonary disease: randomized controlled feasibility study. JMIR Mhealth Uhealth 2020;8(6):e16203 [FREE Full text] [doi: 10.2196/16203] [Medline: 32490838]
36. Liverani M, Ir P, Wiseman V, Perel P. User experiences and perceptions of health wearables: an exploratory study in Cambodia. Glob Health Res Policy 2021;6(1):33 [FREE Full text] [doi: 10.1186/s41256-021-00221-3] [Medline: 34556184]
37. Nasseri M, Nurse E, Glasstetter M, Böttcher S, Gregg NM, Nandakumar AL, et al. Signal quality and patient experience with wearable devices for epilepsy management. Epilepsia 2020;61 Suppl 1:S25-S35. [doi: 10.1111/epi.16527] [Medline: 32497269]
38. Sala-Cunill A, Luengo O, Curran A, Moreno N, Labrador-Horrillo M, Guilarte M, et al. Digital technology for anaphylaxis management impact on patient behaviour: a randomized clinical trial. Allergy 2021;76(5):1507-1516. [doi: 10.1111/all.14626] [Medline: 33043475]
39. Vaughn J, Gollarahalli S, Shaw RJ, Docherty S, Yang Q, Malhotra C, et al. Mobile health technology for pediatric symptom monitoring: a feasibility study. Nurs Res 2020;69(2):142-148 [FREE Full text] [doi: 10.1097/NNR.0000000000000403] [Medline: 31972852]
40. Ollenschläger M, Kluge F, Müller-Schulz M, Püllen R, Möller C, Klucken J, et al. Wearable gait analysis systems: ready to be used by medical practitioners in geriatric wards? Eur Geriatr Med 2022;13(4):817-824 [FREE Full text] [doi: 10.1007/s41999-022-00629-1] [Medline: 35243600]
41. Matcham F, Carr E, White KM, Leightley D, Lamers F, Siddi S, et al. Predictors of engagement with remote sensing technologies for symptom measurement in major depressive disorder. J Affect Disord 2022;310:106-115. [doi: 10.1016/j.jad.2022.05.005] [Medline: 35525507]
42. Bruno E, Biondi A, Thorpe S, Richardson MP, RADAR-CNS Consortium. Patients self-mastery of wearable devices for seizure detection: a direct user-experience. Seizure 2020;81:236-240 [FREE Full text] [doi: 10.1016/j.seizure.2020.08.023] [Medline: 32889512]
43. Figueiredo J, Carvalho SP, Vilas-Boas JP, Gonçalves LM, Moreno JC, Santos CP. Wearable inertial sensor system towards daily human kinematic gait analysis: benchmarking analysis to MVN BIOMECH. Sensors (Basel) 2020;20(8):2185 [FREE Full text] [doi: 10.3390/s20082185] [Medline: 32290636]
44. Lunsford-Avery JR, Keller C, Kollins SH, Krystal AD, Jackson L, Engelhard MM. Feasibility and acceptability of wearable sleep electroencephalogram device use in adolescents: observational study. JMIR Mhealth Uhealth 2020;8(10):e20590 [FREE Full text] [doi: 10.2196/20590] [Medline: 33001035]
45. Standoli CE, Guarneri MR, Perego P, Mazzola M, Mazzola A, Andreoni G. A smart wearable sensor system for counter-fighting overweight in teenagers. Sensors (Basel) 2016;16(8):1220 [FREE Full text] [doi: 10.3390/s16081220] [Medline: 27517929]
46. Marando CM, Mansouri K, Kahook MY, Seibold LK. Tolerability and functionality of a wireless 24-hour ocular telemetry sensor in African American glaucoma patients. J Glaucoma 2019;28(2):119-124. [doi: 10.1097/IJG.0000000000001141] [Medline: 30688840]
47. Pavic M, Klaas V, Theile G, Kraft J, Tröster G, Guckenberger M. Feasibility and usability aspects of continuous remote monitoring of health status in palliative cancer patients using wearables. Oncology 2020;98(6):386-395 [FREE Full text] [doi: 10.1159/000501433] [Medline: 31336377]
48. Yang K, Meadmore K, Freeman C, Grabham N, Hughes AM, Wei Y, et al. Development of user-friendly wearable electronic textiles for healthcare applications. Sensors (Basel) 2018;18(8):2410 [FREE Full text] [doi: 10.3390/s18082410] [Medline: 30044382]
49. McGillion MH, Dvirnik N, Yang S, Belley-Côté E, Lamy A, Whitlock R, et al. Continuous noninvasive remote automated blood pressure monitoring with novel wearable technology: a preliminary validation study. JMIR Mhealth Uhealth 2022;10(2):e24916 [FREE Full text] [doi: 10.2196/24916] [Medline: 34876396]

XSL•FO
RenderX

50. Carrasco JJ, Pérez-Alenda S, Casaña J, Soria-Olivas E, Bonanad S, Querol F. Physical activity monitoring and acceptance of a commercial activity tracker in adult patients with haemophilia. Int J Environ Res Public Health 2019;16(20):3851 [FREE Full text] [doi: 10.3390/ijerph16203851] [Medline: 31614706]
51. Argent R, Slevin P, Bevilacqua A, Neligan M, Daly A, Caulfield B. Wearable sensor-based exercise biofeedback for orthopaedic rehabilitation: a mixed methods user evaluation of a prototype system. Sensors (Basel) 2019;19(2):432 [FREE Full text] [doi: 10.3390/s19020432] [Medline: 30669657]
52. Tsung-Yin O, Chih-Young H, Che-Wei L. A mixed-methods study of users' journey mapping experience and acceptance of telehealthcare technology in Taiwan. Telemed e-Health 2019;25(11):1057-1070. [doi: 10.1089/tmj.2018.0155] [Medline: 30694730]
53. Simblett SK, Biondi A, Bruno E, Ballard D, Stoneman A, Lees S, et al. Patients' experience of wearing multimodal sensor devices intended to detect epileptic seizures: a qualitative analysis. Epilepsy Behav 2020;102:106717. [doi: 10.1016/j.yebeh.2019.106717] [Medline: 31785481]
54. Chevallier T, Buzancais G, Occean B, Rataboul P, Boisson C, Simon N, et al. Feasibility of remote digital monitoring using wireless Bluetooth monitors, the Smart Angel™ app and an original web platform for patients following outpatient surgery: a prospective observational pilot study. BMC Anesthesiol 2020;20(1):259 [FREE Full text] [doi: 10.1186/s12871-020-01178-5] [Medline: 33032541]
55. Virbel-Fleischman C, Rétory Y, Hardy S, Huiban C, Corvol JC, Grabli D. Body-worn sensors for Parkinson's disease: a qualitative approach with patients and healthcare professionals. PLoS One 2022;17(5):e0265438 [FREE Full text] [doi: 10.1371/journal.pone.0265438] [Medline: 35511812]
56. Meritam P, Ryvlin P, Beniczky S. User-based evaluation of applicability and usability of a wearable accelerometer device for detecting bilateral tonic-clonic seizures: a field study. Epilepsia 2018;59 Suppl 1:48-52. [doi: 10.1111/epi.14051] [Medline: 29873828]
57. Esbjörnsson M, Ullberg T. Safety and usability of wearable accelerometers for stroke detection the STROKE ALARM PRO 1 study. J Stroke Cerebrovasc Dis 2022;31(11):106762 [FREE Full text] [doi: 10.1016/j.jstrokecerebrovasdis.2022.106762] [Medline: 36115106]
58. Hua A, Johnson N, Quinton J, Chaudhary P, Buchner D, Hernandez ME. Design of a low-cost, wearable device for kinematic analysis in physical therapy settings. Methods Inf Med 2020;59(1):41-47. [doi: 10.1055/s-0040-1710380] [Medline: 32535880]
59. Domingos C, Costa P, Santos NC, Pêgo JM. Usability, acceptability, and satisfaction of a wearable activity tracker in older adults: observational study in a real-life context in Northern Portugal. J Med Internet Res 2022;24(1):e26652 [FREE Full text] [doi: 10.2196/26652] [Medline: 35080503]
60. Wing D, Godino JG, Baker FC, Yang R, Chevance G, Thompson WK, et al. Recommendations for identifying valid wear for consumer-level wrist-worn activity trackers and acceptability of extended device deployment in children. Sensors (Basel) 2022;22(23):9189 [FREE Full text] [doi: 10.3390/s22239189] [Medline: 36501894]
61. Cortell-Tormo JM, Garcia-Jaen M, Ruiz-Fernandez D, Fuster-Lloret V. Lumbatex: a wearable monitoring system based on inertial sensors to measure and control the lumbar spine motion. IEEE Trans Neural Syst Rehabil Eng 2019;27(8):1644-1653. [doi: 10.1109/TNSRE.2019.2927083] [Medline: 31283484]
62. Keogh A, Dorn JF, Walsh L, Calvo F, Caulfield B. Comparing the usability and acceptability of wearable sensors among older Irish adults in a real-world context: observational study. JMIR Mhealth Uhealth 2020;8(4):e15704 [FREE Full text] [doi: 10.2196/15704] [Medline: 32310149]
63. Hamilton C, Lovarini M, van den Berg M, McCluskey A, Hassett L. Usability of affordable feedback-based technologies to improve mobility and physical activity in rehabilitation: a mixed methods study. Disabil Rehabil 2022;44(15):4029-4038. [doi: 10.1080/09638288.2021.1884904] [Medline: 33645384]
64. O'Brien J, Mason A, Cassarino M, Chan J, Setti A. Older women's experiences of a community-led walking programme using activity trackers. Int J Environ Res Public Health 2021;18(18):9818 [FREE Full text] [doi: 10.3390/ijerph18189818] [Medline: 34574741]
65. Jiménez-Fernández S, de Toledo P, del Pozo F. Usability and interoperability in wireless sensor networks for patient telemonitoring in chronic disease management. IEEE Trans Biomed Eng 2013;60(12):3331-3339. [doi: 10.1109/TBME.2013.2280967] [Medline: 24021636]
66. Areia C, Young L, Vollam S, Ede J, Santos M, Tarassenko L, et al. Wearability testing of ambulatory vital sign monitoring devices: prospective observational cohort study. JMIR Mhealth Uhealth 2020;8(12):e20214 [FREE Full text] [doi: 10.2196/20214] [Medline: 33325827]
67. Bruno E, Biondi A, Böttcher S, Lees S, Schulze-Bonhage A, Richardson MP, RADAR-CNS Consortium. Day and night comfort and stability on the body of four wearable devices for seizure detection: a direct user-experience. Epilepsy Behav 2020;112:107478. [doi: 10.1016/j.yebeh.2020.107478] [Medline: 33181896]
68. Uomoto JM, Skopp N, Jenkins-Guarnieri M, Reini J, Thomas D, Adams RJ, et al. Assessing the clinical utility of a wearable device for physiological monitoring of heart rate variability in military service members with traumatic brain injury. Telemed e-Health 2022;28(10):1496-1504. [doi: 10.1089/tmj.2021.0627] [Medline: 35231193]

XSL•FO
RenderX

69. Mitsopoulos K, Fiska V, Tagaras K, Papias A, Antoniou P, Nizamis K, et al. NeuroSuitUp: system architecture and validation of a motor rehabilitation wearable robotics and serious game platform. Sensors (Basel) 2023;23(6):3281 [FREE Full text] [doi: 10.3390/s23063281] [Medline: 36991992]
70. Grym K, Niela-Vilén H, Ekholm E, Hamari L, Azimi I, Rahmani A, et al. Feasibility of smart wristbands for continuous monitoring during pregnancy and one month after birth. BMC Pregnancy Childbirth 2019;19(1):34 [FREE Full text] [doi: 10.1186/s12884-019-2187-9] [Medline: 30654747]
71. Pradhan S, Kelly VE. Quantifying physical activity in early Parkinson disease using a commercial activity monitor. Parkinsonism Relat Disord 2019;66:171-175 [FREE Full text] [doi: 10.1016/j.parkreldis.2019.08.001] [Medline: 31420310]
72. Lin WY, Ke HL, Chou WC, Chang PC, Tsai TH, Lee MY. Realization and technology acceptance test of a wearable cardiac health monitoring and early warning system with multi-channel MCGs and ECG. Sensors (Basel) 2018;18(10):3538 [FREE Full text] [doi: 10.3390/s18103538] [Medline: 30347695]
73. Yurkiewicz IR, Simon P, Liedtke M, Dahl G, Dunn T. Effect of Fitbit and iPad wearable technology in health-related quality of life in adolescent and young adult cancer patients. J Adolesc Young Adult Oncol 2018;7(5):579-583. [doi: 10.1089/jayao.2018.0022] [Medline: 29924668]
74. Mackintosh KA, Chappel SE, Salmon J, Timperio A, Ball K, Brown H, et al. Parental perspectives of a wearable activity tracker for children younger than 13 years: acceptability and usability study. JMIR Mhealth Uhealth 2019;7(11):e13858 [FREE Full text] [doi: 10.2196/13858] [Medline: 31682585]
75. Hendriks MMS, Vos-van der Hulst M, Keijsers NLW. Feasibility of a sensor-based technological platform in assessing gait and sleep of in-hospital stroke and incomplete spinal cord injury (iSCI) patients. Sensors (Basel) 2020;20(10):2748 [FREE Full text] [doi: 10.3390/s20102748] [Medline: 32408490]
76. Lin BS, Wong AM, Tseng KC. Community-based ECG monitoring system for patients with cardiovascular diseases. J Med Syst 2016;40(4):80. [doi: 10.1007/s10916-016-0442-4] [Medline: 26802010]
77. Radder B, Prange-Lasonder GB, Kottink AIR, Holmberg J, Sletta K, van Dijk M, et al. Home rehabilitation supported by a wearable soft-robotic device for improving hand function in older adults: a pilot randomized controlled trial. PLoS One 2019;14(8):e0220544 [FREE Full text] [doi: 10.1371/journal.pone.0220544] [Medline: 31386685]
78. Bell KM, Onyeukwu C, McClincy MP, Allen M, Bechard L, Mukherjee A, et al. Verification of a portable motion tracking system for remote management of physical rehabilitation of the knee. Sensors (Basel) 2019;19(5):1021 [FREE Full text] [doi: 10.3390/s19051021] [Medline: 30823373]
79. Pillalamarri SS, Huyett LM, Abdel-Malek A. Novel Bluetooth-enabled tubeless insulin pump: a user experience design approach for a connected digital diabetes management platform. J Diabetes Sci Technol 2018;12(6):1132-1142 [FREE Full text] [doi: 10.1177/1932296818804802] [Medline: 30304951]
80. Burgdorf A, Güthe I, Jovanović M, Kutafina E, Kohlschein C, Bitsch J, et al. The mobile sleep lab app: an open-source framework for mobile sleep assessment based on consumer-grade wearable devices. Comput Biol Med 2018;103:8-16. [doi: 10.1016/j.compbiomed.2018.09.025] [Medline: 30316065]
81. Del-Valle-Soto C, Valdivia LJ, López-Pimentel JC, Visconti P. Comparison of collaborative and cooperative schemes in sensor networks for non-invasive monitoring of people at home. Int J Environ Res Public Health 2023;20(7):5268 [FREE Full text] [doi: 10.3390/ijerph20075268] [Medline: 37047884]
82. Hosseini A, Buonocore CM, Hashemzadeh S, Hojaiji H, Kalantarian H, Sideris C, et al. Feasibility of a secure wireless sensing smartwatch application for the self-management of pediatric asthma. Sensors (Basel) 2017;17(8):1780 [FREE Full text] [doi: 10.3390/s17081780] [Medline: 28771168]
83. Lin WY, Chou WC, Tsai TH, Lin CC, Lee MY. Development of a wearable instrumented vest for posture monitoring and system usability verification based on the technology acceptance model. Sensors (Basel) 2016;16(12):2172 [FREE Full text] [doi: 10.3390/s16122172] [Medline: 27999324]
84. Esfahani MIM, Nussbaum MA. Preferred placement and usability of a smart textile system vs. inertial measurement units for activity monitoring. Sensors (Basel) 2018;18(8):2501 [FREE Full text] [doi: 10.3390/s18082501] [Medline: 30071635]
85. Stubberud A, Tronvik E, Olsen A, Gravdahl G, Linde M. Biofeedback treatment app for pediatric migraine: development and usability study. Headache 2020;60(5):889-901. [doi: 10.1111/head.13772] [Medline: 32052871]
86. Chung J, Brakey HR, Reeder B, Myers O, Demiris G. Community-dwelling older adults' acceptance of smartwatches for health and location tracking. Int J Older People Nurs 2023;18(1):e12490 [FREE Full text] [doi: 10.1111/opn.12490] [Medline: 35818900]
87. Shelley J, Fairclough SJ, Knowles ZR, Southern KW, McCormack P, Dawson EA, et al. A formative study exploring perceptions of physical activity and physical activity monitoring among children and young people with cystic fibrosis and health care professionals. BMC Pediatr 2018;18(1):335 [FREE Full text] [doi: 10.1186/s12887-018-1301-x] [Medline: 30352564]
88. Auerswald T, Meyer J, von Holdt K, Voelcker-Rehage C. Application of activity trackers among nursing home residents—a pilot and feasibility study on physical activity behavior, usage behavior, acceptance, usability and motivational impact. Int J Environ Res Public Health 2020;17(18):6683 [FREE Full text] [doi: 10.3390/ijerph17186683] [Medline: 32937840]
89. Schuurmans MM, Muszynski M, Li X, Marcinkevičs R, Zimmerli L, Lopez DM, et al. Multimodal remote home monitoring of lung transplant recipients during COVID-19 vaccinations: usability pilot study of the COVIDA desk incorporating

XSL•FO
RenderX

wearable devices. Medicina (Kaunas) 2023;59(3):617 [FREE Full text] [doi: 10.3390/medicina59030617] [Medline: 36984618]
90. Cesareo A, Nido SA, Biffi E, Gandossini S, D'Angelo MG, Aliverti A. A wearable device for breathing frequency monitoring: a pilot study on patients with muscular dystrophy. Sensors (Basel) 2020;20(18):5346 [FREE Full text] [doi: 10.3390/s20185346] [Medline: 32961986]
91. Albani G, Ferraris C, Nerino R, Chimienti A, Pettiti G, Parisi F, et al. An integrated multi-sensor approach for the remote monitoring of Parkinson's disease. Sensors (Basel) 2019;19(21):4764 [FREE Full text] [doi: 10.3390/s19214764] [Medline: 31684020]
92. Demiris G, Chaudhuri S, Thompson HJ. Older adults' experience with a novel fall detection device. Telemed e-Health 2016;22(9):726-732 [FREE Full text] [doi: 10.1089/tmj.2015.0218] [Medline: 26959299]
93. Kaiserman K, Buckingham BA, Prakasam G, Gunville F, Slover RH, Wang Y, et al. Acceptability and utility of the mySentry remote glucose monitoring system. J Diabetes Sci Technol 2013;7(2):356-361 [FREE Full text] [doi: 10.1177/193229681300700211] [Medline: 23566993]
94. Drehlich M, Naraine M, Rowe K, Lai SK, Salmon J, Brown H, et al. Using the technology acceptance model to explore adolescents' perspectives on combining technologies for physical activity promotion within an intervention: usability study. J Med Internet Res 2020;22(3):e15552 [FREE Full text] [doi: 10.2196/15552] [Medline: 32141834]
95. Nguyen NH, Hadgraft NT, Moore MM, Rosenberg DE, Lynch C, Reeves MM, et al. A qualitative evaluation of breast cancer survivors' acceptance of and preferences for consumer wearable technology activity trackers. Support Care Cancer 2017;25(11):3375-3384. [doi: 10.1007/s00520-017-3756-y] [Medline: 28540402]
96. Arends J, Thijs RD, Gutter T, Ungureanu C, Cluitmans P, Van Dijk J, et al. Multimodal nocturnal seizure detection in a residential care setting: a long-term prospective trial. Neurology 2018;91(21):e2010-e2019 [FREE Full text] [doi: 10.1212/WNL.0000000000006545] [Medline: 30355702]
97. Finkelstein J, Cisse P, Jeong IC. Feasibility of interactive resistance chair in older adults with diabetes. Stud Health Technol Inform 2015;213:61-64. [Medline: 26152953]
98. Thilo FJS, Hahn S, Halfens RJG, Schols JMGA. Usability of a wearable fall detection prototype from the perspective of older people—a real field testing approach. J Clin Nurs 2019;28(1-2):310-320. [doi: 10.1111/jocn.14599] [Medline: 29964344]
99. Preusse KC, Mitzner TL, Fausset CB, Rogers WA. Older adults' acceptance of activity trackers. J Appl Gerontol 2017;36(2):127-155 [FREE Full text] [doi: 10.1177/0733464815624151] [Medline: 26753803]
100. Hofbauer LM, Rodriguez FS. How is the usability of commercial activity monitors perceived by older adults and by researchers? A cross-sectional evaluation of community-living individuals. BMJ Open 2022;12(11):e063135 [FREE Full text] [doi: 10.1136/bmjopen-2022-063135] [Medline: 36323474]
101. Cajamarca G, Rodríguez I, Herskovic V, Campos M, Riofrío JC. StraightenUp+: monitoring of posture during daily activities for older persons using wearable sensors. Sensors (Basel) 2018;18(10):3409 [FREE Full text] [doi: 10.3390/s18103409] [Medline: 30314352]
102. Caswell N, Kuru K, Ansell D, Jones MJ, Watkinson BJ, Leather P, et al. Patient engagement in medical device design: refining the essential attributes of a wearable, pre-void, ultrasound alarm for nocturnal enuresis. Pharmaceut Med 2020;34(1):39-48. [doi: 10.1007/s40290-019-00324-w] [Medline: 31970684]
103. Imbesi S, Corzani M. Multisensory cues for gait rehabilitation with smart glasses: methodology, design, and results of a preliminary pilot. Sensors (Basel) 2023;23(2):874 [FREE Full text] [doi: 10.3390/s23020874] [Medline: 36679671]
104. Shore L, Power V, Hartigan B, Schülein S, Graf E, de Eyto A, et al. Exoscore: a design tool to evaluate factors associated with technology acceptance of soft lower limb exosuits by older adults. Hum Factors 2020;62(3):391-410. [doi: 10.1177/0018720819868122] [Medline: 31419179]
105. Hart P, Bierwirth R, Fulk G, Sazonov E. The design and evaluation of an activity monitoring user interface for people with stroke. 2014 Presented at: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 26-30, 2014; Chicago, IL p. 5908-5911. [doi: 10.1109/EMBC.2014.6944973]
106. Rupp MA, Michaelis JR, McConnell DS, Smither JA. The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact. Appl Ergon 2018;70:77-87. [doi: 10.1016/j.apergo.2018.02.005] [Medline: 29866329]
107. Drabarek D, Anh NT, Nhung NV, Hoa NB, Fox GJ, Bernays S. Implementation of medication event reminder monitors among patients diagnosed with drug susceptible tuberculosis in rural Viet Nam: a qualitative study. PLoS One 2019;14(7):e0219891 [FREE Full text] [doi: 10.1371/journal.pone.0219891] [Medline: 31329610]
108. Vandelanotte C, Duncan MJ, Maher CA, Schoeppe S, Rebar AL, Power DA, et al. The effectiveness of a web-based computer-tailored physical activity intervention using Fitbit activity trackers: randomized trial. J Med Internet Res 2018;20(12):e11321 [FREE Full text] [doi: 10.2196/11321] [Medline: 30563808]
109. Matsunaga K, Ogasawara T, Kodate J, Mukaino M, Saitoh E. On-site evaluation of rehabilitation patients monitoring system using distributed wireless gateways. 2019 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany p. 3195-3198. [doi: 10.1109/EMBC.2019.8856963]

XSL•FO
RenderX

110. Van Velthoven MH, Oke J, Kardos A. ChroniSense national early warning score study: comparison study of a wearable wrist device to measure vital signs in patients who are hospitalized. J Med Internet Res 2023;25:e40226 [FREE Full text] [doi: 10.2196/40226] [Medline: 36745491]
111. de'Sperati C, Dalmasso V, Moretti M, Høeg ER, Baud-Bovy G, Cozzi R, et al. Enhancing visual exploration through augmented gaze: high acceptance of immersive virtual biking by oldest olds. Int J Environ Res Public Health 2023;20(3):1671 [FREE Full text] [doi: 10.3390/ijerph20031671] [Medline: 36767037]
112. Radder B, Prange-Lasonder G, Kottink AIR, Melendez-Calderon A, Buurke JH, Rietman JS. Feasibility of a wearable soft-robotic glove to support impaired hand function in stroke patients. J Rehabil Med 2018;50(7):598-606 [FREE Full text] [doi: 10.2340/16501977-2357] [Medline: 30003268]
113. McNamara RJ, Tsai LLY, Wootton SL, Ng LWC, Dale MT, McKeough ZJ, et al. Measurement of daily physical activity using the SenseWear armband: compliance, comfort, adverse side effects and usability. Chron Respir Dis 2016;13(2):144-154 [FREE Full text] [doi: 10.1177/1479972316631138] [Medline: 26879695]
114. Radder B, Prange-Lasonder GB, Kottink AIR, Holmberg J, Sletta K, Van Dijk M, et al. The effect of a wearable soft-robotic glove on motor function and functional performance of older adults. Assist Technol 2020;32(1):9-15. [doi: 10.1080/10400435.2018.1453888] [Medline: 29601251]
115. ISO 13485:2016. 2020. URL: https://www.iso.org/standard/59752.html [accessed 2024-02-22]
116. Cox CL. Patient understanding: how should it be defined and assessed in clinical practice? J Eval Clin Pract 2023;29(7):1127-1134. [doi: 10.1111/jep.13882] [Medline: 37338517]
117. Thomas EE, Taylor ML, Banbury A, Snoswell CL, Haydon HM, Rejas VMG, et al. Factors influencing the effectiveness of remote patient monitoring interventions: a realist review. BMJ Open 2021;11(8):e051844 [FREE Full text] [doi: 10.1136/bmjopen-2021-051844] [Medline: 34433611]
118. Sharma Y, Djambazova K, Marquez C, Lyden K, Goldsack J, Bakker J. A systematic review assessing the state of analytical validation for connected, mobile, sensor-based digital health technologies. medRxiv. Preprint posted online on May 23, 2023 2023:1-21. [doi: 10.1101/2023.05.22.23290371]
119. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Women, Bibbins-Domingo K, Helman A. Why diverse representation in clinical research matters and the current state of representation within the clinical research ecosystem. 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK584396/ [accessed 2024-01-23]
120. Wronikowska MW, Malycha J, Morgan LJ, Westgate V, Petrinic T, Young JD, et al. Systematic review of applied usability metrics within usability evaluation methods for hospital electronic healthcare record systems: metrics and evaluation methods for eHealth systems. J Eval Clin Pract 2021;27(6):1403-1416 [FREE Full text] [doi: 10.1111/jep.13582] [Medline: 33982356]
121. Wohlgemut JM, Pisirir E, Kyrimi E, Stoner RS, Marsh W, Perkins ZB, et al. Methods used to evaluate usability of mobile clinical decision support systems for healthcare emergencies: a systematic review and qualitative synthesis. JAMIA Open 2023;6(3):ooad051 [FREE Full text] [doi: 10.1093/jamiaopen/ooad051] [Medline: 37449057]
122. ISO 9241-11:2018. 2023. URL: https://www.iso.org/standard/63500.html [accessed 2024-01-23]
123. Manta C, Mahadevan N, Bakker J, Irmak SO, Izmailova E, Park S, et al. EVIDENCE publication checklist for studies evaluating connected sensor technologies: explanation and elaboration. Digital Biomarkers 2021;5(2):127-147 [FREE Full text] [doi: 10.1159/000515835] [Medline: 34179682]

Abbreviations

MeSH: Medical Subject Heading
PICO: patients/participants; intervention; comparator; outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
sDHT: sensor-based digital health technology

Edited by T de Azevedo Cardoso; submitted 22.02.24; peer-reviewed by S Yu, M Nissen; comments to author 11.04.24; revised version received 28.05.24; accepted 11.09.24; published 15.11.24.

Please cite as:
Tandon A, Cobb B, Centra J, Izmailova E, Manyakov NV, McClenahan S, Patel S, Sezgin E, Vairavan S, Vrijens B, Bakker JP, Digital Health Measurement Collaborative Community (DATAcc) hosted by DiMe
Human Factors, Human-Centered Design, and Usability of Sensor-Based Digital Health Technologies: Scoping Review
J Med Internet Res 2024;26:e57628
URL: https://www.jmir.org/2024/1/e57628
doi:10.2196/57628
PMID:

XSL•FO
RenderX

©Animesh Tandon, Bryan Cobb, Jacob Centra, Elena Izmailova, Nikolay V Manyakov, Samantha McClenahan, Smit Patel, Emre Sezgin, Srinivasan Vairavan, Bernard Vrijens, Jessie P Bakker, Digital Health Measurement Collaborative Community (DATAcc) hosted by DiMe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Use of Digital Tools in Arbovirus Surveillance: Scoping Review

Carolina Lopes Melo[1*], MSc; Larissa Rangel Mageste[1], MSc; Lusiele Guaraldo[1], PhD; Daniela Polessa Paula[2], PhD; Mayumi Duarte Wakimoto[1*], MD, PhD

[1]Instituto Nacional de Infectologia Evandro Chagas, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil
[2]Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brazil
[*]these authors contributed equally

Corresponding Author:
Mayumi Duarte Wakimoto, MD, PhD
Instituto Nacional de Infectologia Evandro Chagas
Fundação Oswaldo Cruz
Avenida Brasil, 4365
Rio de Janeiro, 21040-900
Brazil
Phone: 55 2138659110
Email: mayumi.wakimoto@ini.fiocruz.br

Abstract

Background: The development of technology and information systems has led to important changes in public health surveillance.

Objective: This scoping review aimed to assess the available evidence and gather information about the use of digital tools for arbovirus (dengue virus [DENV], zika virus [ZIKV], and chikungunya virus [CHIKV]) surveillance.

Methods: The databases used were MEDLINE, SCIELO, LILACS, SCOPUS, Web of Science, and EMBASE. The inclusion criterion was defined as studies that described the use of digital tools in arbovirus surveillance. The exclusion criteria were defined as follows: letters, editorials, reviews, case reports, series of cases, descriptive epidemiological studies, laboratory and vaccine studies, economic evaluation studies, and studies that did not clearly describe the use of digital tools in surveillance. Results were evaluated in the following steps: monitoring of outbreaks or epidemics, tracking of cases, identification of rumors, decision-making by health agencies, communication (cases and bulletins), and dissemination of information to society).

Results: Of the 2227 studies retrieved based on screening by title, abstract, and full-text reading, 68 (3%) studies were included. The most frequent digital tools used in arbovirus surveillance were apps (n=24, 35%) and Twitter, currently called X (n=22, 32%). These were mostly used to support the traditional surveillance system, strengthening aspects such as information timeliness, acceptability, flexibility, monitoring of outbreaks or epidemics, detection and tracking of cases, and simplicity. The use of apps to disseminate information to society (P=.02), communicate (cases and bulletins; P=.01), and simplicity (P=.03) and the use of Twitter to identify rumors (P=.008) were statistically relevant in evaluating scores. This scoping review had some limitations related to the choice of DENV, ZIKV, and CHIKV as arboviruses, due to their clinical and epidemiological importance.

Conclusions: In the contemporary scenario, it is no longer possible to ignore the use of web data or social media as a complementary strategy to health surveillance. However, it is important that efforts be combined to develop new methods that can ensure the quality of information and the adoption of systematic measures to maintain the integrity and reliability of digital tools' data, considering ethical aspects.

(J Med Internet Res 2024;26:e57476) doi:10.2196/57476

KEYWORDS

arbovirus infections; dengue; zika virus; chikungunya fever; public health surveillance; digital tool; technology

Introduction

Arboviruses have become relevant public health problems in tropical and subtropical areas due to either socioeconomic or environmental factors, involving inadequate occupation of urban space, poor sanitary conditions, and deforestation [1].

Dengue, caused by the dengue virus (DENV), is 1 of the most important neglected tropical diseases transmitted by mosquitoes in humans. Since its onset in Southeast Asia in the 1950s, the disease has rapidly spread throughout tropical regions and currently remains a health concern worldwide [2].

XSL•FO
RenderX

Chikungunya, caused by the chikungunya virus (CHIKV), was first described in 1952 [3] and has been responsible for outbreaks and epidemics in Asia and Africa. Between 2005 and 2007, 266,000 cases were reported in the Reunions Islands, affecting almost 34% of the island's population [4]. A major outbreak occurred in 2013 with the emergence of an Asian lineage of the virus affecting the Caribbean Saint Martin Island, from where the virus spread to more than 50 countries of the South American continent, leading to over 1 million infections [5]. It is important to highlight the chikungunya disease burden, which includes chronicity, severe infections, increased hospitalization risks, and associated mortality [6].

Zika was first described in 1947 in nonhuman primates, and infections in humans were sporadic and mild. In 2013-2014, an outbreak occurred in French Polynesia, and severe neurological manifestations were reported [7]. In 2015, cases of congenital microcephaly among pregnant women infected with zika virus (ZIKV) were reported in Brazil. Those findings raised concerns about the infection during pregnancy, and a body of evidence showed the association of ZIKV infection with fetal death, growth restriction, and a series of abnormalities in the fetal central nervous system, as well as microcephaly [8,9].

The infection scenario caused by arboviruses has pointed to relevant threats to public health in recent years. The cocirculation of the 3 arboviruses (DENV, CHIKV, ZIKV) has imposed major challenges in surveillance and increased the demand for health service support in affected areas [10]. Among the arboviruses, we chose DENV, CHIKV, and ZIKV as they have great medical importance and similar clinical manifestations and as transmission occurs through the same vector, the mosquito *Aedes aegypti*. Dengue fever has a high incidence in several countries. Chikungunya presents high morbidity, considering the course of the disease in the acute phase, which can lead to chronic symptoms. Zika in pregnant women has an important impact on congenital neurological manifestations, in addition to being an emerging disease.

The continuous and systematic collection, analysis, and interpretation of health-related data are part of the scope of public health surveillance. Monitoring of outbreaks and epidemics, tracking of cases, evaluation of interventions, evaluation of rumors, communication (cases and bulletins), and decision-making by health agencies are essential steps for a surveillance system. Therefore, the effectiveness of a surveillance system is directly related to its ability to control diseases [11,12]. However, the rapid development of data science, including big data and artificial intelligence (AI), and the growth of accessible and heterogeneous health-related data are definitely changing the field of health surveillance [13]. The use of technology in health care has increased in recent years. Digital tools, such as apps, digital forms, online chats, video calls, telemedicine, social media, and games, have been used for data collection, case tracking, and disease risk classification [14-16]. Furthermore, the use of big data composing hybrid systems, through the combination of structured and unstructured data, is a promising strategy to collect electronic health records in real time, potentially impacting infectious disease surveillance [17].

It is noteworthy that public engagement in the digital technological universe is becoming increasingly important worldwide. Therefore, the use of digital tools for arbovirus surveillance seems relevant due to the impact of the diseases and the need for timely and effective control strategies [18].

This review aimed to assess the available evidence and gather information about the use of digital tools for arbovirus (DENV, ZIKV, and CHIKV) surveillance.

Methods

Study Design

We conducted a scoping review to systematically map the use of digital tools in arbovirus surveillance [19], defined as technological resources/electronic devices capable of establishing communication between individuals through data sharing [20,21]. Our protocol was elaborated using the Joanna Briggs Institute PRIMSA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines for scoping reviews (Table S1 in Multimedia Appendix 1) [19,22]. The protocol was registered with the Open Science Framework.

Search

Data were retrieved from the following bibliographic databases: MEDLINE, EMBASE, LILACS, SCIELO, Web of Science, and SCOPUS. The research question was based on the population, context, concept (PCC) approach [19]: population, digital tools; context, arbovirus surveillance; and concept, use of digital tools to perform surveillance. The following Medical Subject Headings (MeSH) descriptors were combined: "arbovirus infections," "dengue," "zika virus," "chikungunya fever," "public health surveillance," "epidemiological monitoring," "technology," "audiovisual aids," "social media," "big data," "mobile apps," and "social networking." The search was performed in April 2023 (Table S2 in Multimedia Appendix 2).

Selection, Reading, and Data Extraction

The selection was independently performed by 2 authors (CLM and LR), and disagreements were resolved by a third author (MDW). The inclusion criterion was defined as studies that described the use of digital tools in arbovirus surveillance.

First, titles and abstracts returned by the search were read, and the following were excluded: letters, editorials, reviews, case reports, series of cases, descriptive epidemiological studies, laboratory and vaccine studies, economic evaluation studies, and studies that did not clearly describe the use of digital tools in surveillance. Next, potentially eligible studies were read in full, and the same inclusion and exclusion criteria were applied. The online software Rayyan was used for the selection process [23].

Data from included studies were collected using a standardized data extraction tool designed for this study. The form included the following sections: identification of the study (authors, journal, year of publication, language), studies characteristics (period, study population, place of study), digital tools used, frequency of data collection, objective of the method used for

XSL•FO
RenderX

the surveillance of arboviruses, and practical applicability. Data charting was implemented using EpiData 3.1 software.

The form was tested initially with 5 papers and subsequently subjected to minor adjustments, such as including new data record fields or changes in format to improve information recording. Two reviewers independently collected data from each included paper. Any disagreements were resolved through discussion between the 2 reviewers or by a third reviewer.

Data Synthesis and Analysis

A descriptive analysis of the methods and results of using digital tools was carried out based on the attributes for evaluating surveillance systems proposed by the Centers for Disease Control and Prevention (CDC) and on some essential surveillance activities. We used the following attributes: (1) sensitivity (proportion and cases of the disease detected by the surveillance system, ability to detect outbreaks, ability to monitor changes in the number of cases over time), (2) opportunity (speed between the steps of the surveillance system), (3) simplicity (structure and ease of operationalization of the surveillance system), (4) acceptability (willingness of people or organizations to participate in the system), (5) flexibility (ability to adapt to changing information needs and operating conditions with minimal need for time, personnel, and resources), (6) specificity (capacity of the system to exclude "noncases" of the disease), and (7) positive predictive value (PPV: proportion of reported cases that actually have the event under surveillance) [24]. Furthermore, the criteria for evaluating the contributions of digital tools to arbovirus surveillance were applied based on the *Epidemiological Surveillance Guide*, and the International Health Regulations (IHR) [25,26]. The following surveillance activities were used: monitoring of outbreaks or epidemics, tracking of cases, decision-making by health agencies, identification of rumors, communication (cases and bulletins), and dissemination of information to society.

If the attributes or activities were not clearly mentioned in the study, each independent observer (CLM and LR) imputed the presence or absence according to the CDC definition or according to the performance of the surveillance system. A third observer (MDW) resolved conflicts. A dichotomous variable was created to evaluate the presence or absence of the contribution of digital tools to arbovirus surveillance. An arbitrary value of 2 was assigned when the attributes sensitivity, opportunity, simplicity, acceptability, flexibility, specificity, and PPV were present and 1 if they were absent.

The sum of the scores of the 7 attributes was calculated for each of the selected studies, and this sum variable was categorized as follows: 7 or 8, unsatisfactory; 9 or 10, moderately satisfactory; 11 or 12, satisfactory; and 13 or 14, very satisfactory. The sum of the scores of the 6 health surveillance activities (monitoring of outbreaks or epidemics, tracking of cases, decision-making by health agencies, identification of rumors, communication [cases and bulletins], and dissemination of information to society) was performed for each of the selected studies, and this sum variable was categorized as follows: 6, unsatisfactory; 7 or 8, moderately satisfactory; 9 or 10, satisfactory; and 11 or 12, very satisfactory.

Next, the percentage of the use of each digital tool in the reviewed studies was calculated. The frequency with which the 7 health surveillance attributes and 6 activities were present in the included studies was also calculated according to the use of digital tools (Multimedia Appendices 3 and 4).

The mean (SD) and *P* value of the sum attributed to the score of each axis were calculated: 7 surveillance system attributes and 6 health surveillance activities according to the digital tools that were used with the greatest value absolute in the included studies (apps, Twitter, Google Trends, and big data), with the aim of identifying whether the groups had a statistically significant difference. Subsequently, we compared the means using ANOVA for the aforementioned digital tools according to the 2 axes to verify the difference between them.

We used R software version 4.3.1 (*summarytools* package for descriptive analysis, *dplyr* package for working with dataframes, and *MASS* and *car* native R packages for ANOVA; R Foundation for Statistical Computing) to perform data analysis.

Results

Study Characteristics

The search strategy retrieved 2227 studies; after removing duplicates and applying the inclusion and exclusion criteria, we included 68 (3%) studies in the review (Figure S1 in Multimedia Appendix 1).

The characteristics of the studies included are outlined in Multimedia Appendix 5. Of the 68 studies, 50% (n=34) were performed in Asia, 16% (n=11) in North America, and 9% (n=6) in Europe or Central America. In addition, 32% (n=22) of the studies were performed in South America, with 17 (77%) of these in Brazil. Dengue was addressed in 50 (74%) studies, followed by zika (n=21, 31%) and chikungunya (n=9, 13%); 6 (9%) studies addressed the 3 arboviruses simultaneously.

The digital tools most studied were apps (n=24, 35%); Twitter, currently called X (n=22, 32%); Google Trends (n=7, 10%); and big data (n=6, 9%). Social media (Twitter, Facebook, Instagram, YouTube, Flirck, Sina Weibo, and blogs) was used in 37% (n=25) of the studies.

The time range between health data collection and dissemination of information was defined by the studies as real time (n=48, 71%), near real time (n=2, 3%), yearly (n=1, 2%), weekly (n=7, 10%), daily (n=8, 12%), other (n=1, 2%), and not clearly described (n=9, 13%), as shown in Multimedia Appendix 5.

Data from digital tools were compared with official data in 16 (24%) studies: [27-42]. Of these, 12 (75%) studies [27,29,31,33-40,42] showed statistically significant correlations ($P<.05$) or strong correlations between official and unofficial data coming from online trends, social media, or big data, of which 6 (50%) addressed the use of Twitter data.

In addition, 21 (31%) studies [28,29,32,34,35,43-58] developed predictions, forecasts, detection of reemergent events, and early warning models, while 7 (10%) studies [44,59-64] presented results regarding the use of digital tools to design a participatory syndromic surveillance system, and 32 (47%) studies evaluated

XSL•FO
RenderX

the use of digital tools for several surveillance activities: prevention and control of arboviruses [40,51,59,65-82], content analysis and rumors [83-91], and mitigation of the lack of epidemiological data in surveillance systems [92,93].

Most studies presented an enhancement in opportunity (n=63, 93%), flexibility (n= 57, 84%), sensitivity (n=63, 93%), and simplicity (n=46, 68%) indicators. In addition, 18 (26%) studies [28,32,39,41,43,45,47,50,52,56,58-60,65,70,79,83,92] presented an enhancement in the ability of the system to exclude "noncases" of the disease (specificity) and the proportion of reported cases that actually had the event under surveillance (PPV).

Furthermore, 63 (93%) studies addressed the monitoring of outbreaks or epidemics, and 60 (88%) studies addressed case tracking. More than half of the studies improved decision-making by health agencies (n=43, 63%) and communication (cases and bulletins; n=40, 59%). Dissemination of information to society and identification of rumors were addressed in 29 (43%) studies.

The use of apps in surveillance enhanced the following indicators: opportunity (n=22, 92%), sensitivity (n=21, 88%), simplicity (n=21, 88%), acceptability (n=19, 79%), and flexibility (n=19, 79%). Furthermore, 23 (34%) studies showed that the use of apps enhanced the monitoring of outbreaks or epidemics (n=22, 96%), the tracking of cases (n=21, 92%), decision-making by health agencies (n=17, 74%), and the availability of information for society (n=14, 61%) [38,44-48,50,57,60-62,64-66,71-74,76,77,79-81,94]. Figure 1 illustrates health surveillance system attributes and contributions to health surveillance activities according to use of apps and Twitter in the 68 studies.

Of the 24 (35%) studies that mentioned the use of apps, 11 (46%) were evaluated as very satisfactory and 9 (38%) as satisfactory according to surveillance activity enhancement. In addition, 7 (29%) studies were very satisfactory and 14 (58%) were satisfactory when surveillance indicators were assessed. No study that mentioned the use of apps was evaluated as unsatisfactory.

Among the studies that addressed the use of Twitter (n=22, 32%) [27,29,30,32,35-37,41,42,49,51,63,75,82-84,86,88-92], 19 (86%) described an increase in speed between the steps of the surveillance system (opportunity) and the ability to adapt to changing information needs, while 18 (82%) described operating conditions with minimal need for time, personnel, and resources (flexibility). More than 12 (55%) studies that addressed the use of Twitter described an enhancement in system sensitivity, and 12 (55%) described the ease of operation and the willingness of people or organizations to participate in the system (acceptability). Less than 7 (32%) studies pointed out an enhancement to system specificity and the PPV.

Furthermore, 18 (82%) studies that addressed the use of Twitter described enhancement in monitoring of outbreaks or epidemics. Of these, 11 (61%) studies described an increase in communication (cases and bulletins) and decision-making by health agencies. In addition, 15 (68%) addressed the identification of rumors, while dissemination of information to society was mentioned in 6 (27%) of the studies (Figure 1).

Of the 22 studies that addressed the use of Twitter, 7 (32%) were evaluated as very satisfactory and 9 (41%) were satisfactory regarding surveillance activities. In addition, 6 (27%) studies were very satisfactory, 11 (50%) were satisfactory, 4 (18%) were moderately satisfactory, and 1 (5%) was unsatisfactory according to the analysis of surveillance system indicators.

Studies that reported the use of Google Trends (n=7, 10%) enhanced the following indicators: simplicity, sensitivity, flexibility, and opportunity. Furthermore, 6 (86%) studies enhanced the monitoring of outbreaks or epidemics and tracking of cases.

Studies that reported the use of big data (n=6, 9%) [28,30,37,43,51,93] assessed sensitivity and flexibility of the surveillance system, as well as opportunity (speed between the surveillance system steps). All of them contributed to the monitoring of outbreaks or epidemics. Additionally, 3 (50%) studies [30,43,93] addressed decision-making by health agencies.

A total of 27 (40%) studies [27,33,39-41,47,48,50,53,56,57,61-64,67,71-73,75,79,82,85,87,91,94] were rated as very satisfactory regarding surveillance activities and addressed the use of the following digital tools: apps, Twitter, game platforms, web-based digital tools, Facebook, Google Trends, Google News, Wikipedia, and Sina Weibo. Of these, 19 (70%) investigated dengue, 7 (26%) investigated zika, and 1 (4%) addressed diarrheal syndrome, respiratory syndrome, arboviral syndrome, chikungunya, zika, and influenza.

In addition, 15 (22%) studies [27,29,36,45,50,52,59-61,65,72,79,84,90,92] were classified as very satisfactory regarding surveillance system indicators and reported the use of the following digital tools: apps, Twitter, Google Trends, Google News, blogs, Wikipedia, telepidemiological surveillance, eletronic bracelets, Google Maps, drones, and the GPS.

In comparing the means of the sums of health surveillance system activities, there was a difference between apps, Twitter, Google Trends, and big data. The activities communication (cases and bulletins; P=.01), dissemination of information to society (P=.02), and identification of rumors (P=.008) showed a statistically significant difference (Table 1).

In comparing the means of the sums of health surveillance system attributes, there was a difference between apps, Twitter, Google Trends, and big data (Table 2).

However, the attribute simplicity (P=.03) showed a statistically significant difference.

XSL•FO
RenderX

Figure 1. Health surveillance activities and surveillance system indicators according to the use of apps and Twitter.

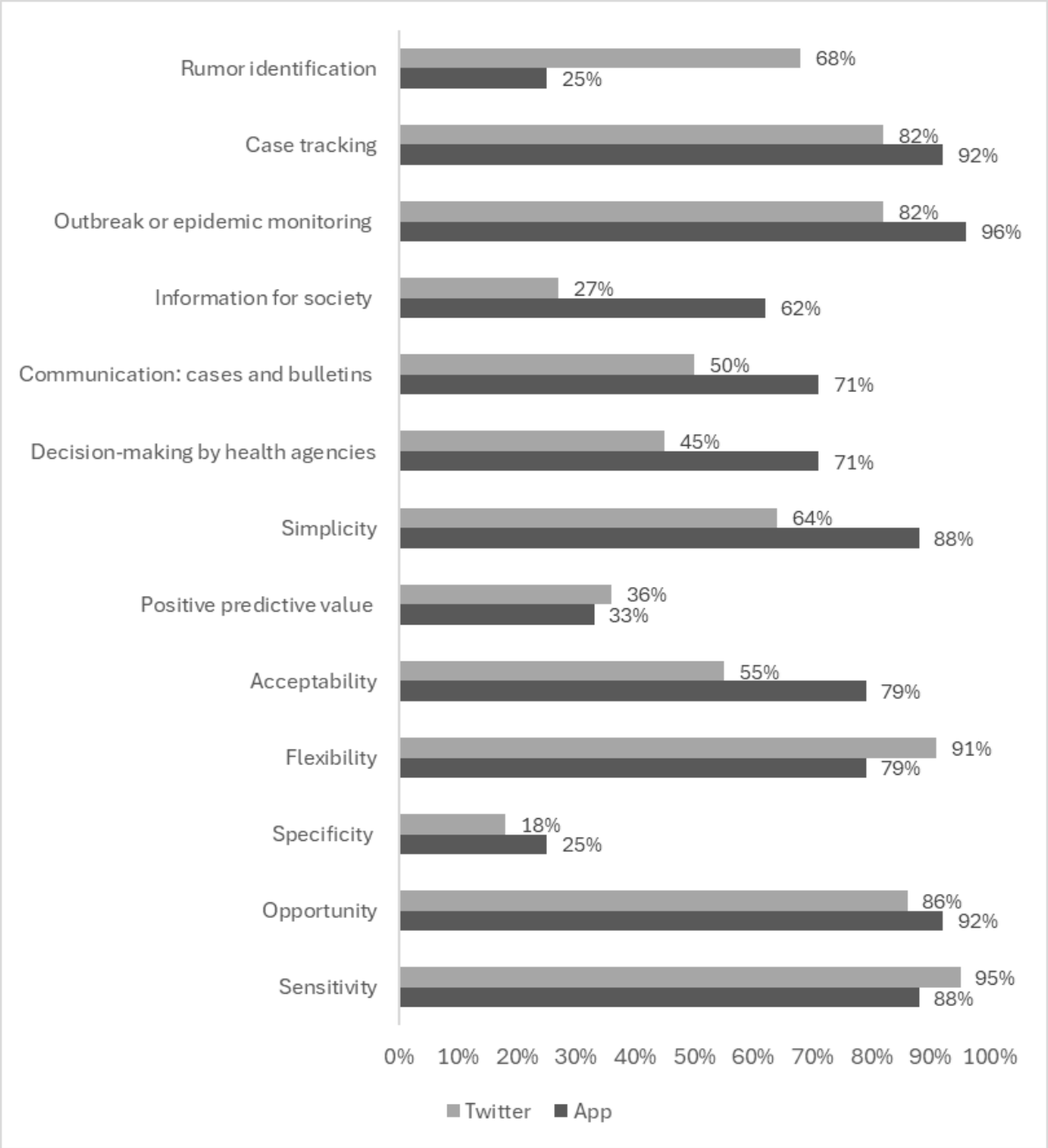

Table 1. ANOVA of health surveillance activity scores.

Health surveillance system activity	Apps (n=24), mean (SD)	Big data (n=6), mean (SD)	Google Trends (n=7), mean (SD)	Twitter (n=22), mean (SD)	*P* value[a]
Decision-making by health agencies	1.7 (0.5)	1.5 (0.5)	1.4 (0.5)	1.5 (0.5)	.31
Communication: cases and bulletins	1.7 (0.5)	1.0 (0.0)	1.7 (0.5)	1.5 (0.5)	.01
Dissemination of information to society	1.6 (0.5)	1.2 (0.4)	1.1 (0.4)	1.3 (0.4)	.02
Outbreak or epidemic monitoring	2.0 (0.2)	2.0 (0.0)	1.9 (0.4)	1.8 (0.4)	.36
Rumor identification	1.2 (0.4)	1.2 (0.4)	1.3 (0.5)	1.7 (0.5)	.008
Case tracking	1.9 (0.3)	2.0 (0.0)	1.9 (0.4)	1.8 (0.4)	0.6

[a]*P* values (ANOVA) of health surveillance system activities considering the use of digital tools (apps, big data, Google Trends, and Twitter).

XSL•FO
RenderX

Table 2. ANOVA of health surveillance system attribute scores.

Health surveillance system attribute	Apps (n=24), mean (SD)	Big data (n=6), mean (SD)	Google Trends (n=7), mean (SD)	Twitter (n=22), mean (SD)	*P* value[a]
Sensitivity	1.9 (0.3)	2.0 (0.0)	2.0 (0.0)	1.9 (0.2)	.94
Specificity	1.8 (0.4)	1.7 (0.5)	1.2 (0.5)	1.2 (0.4)	.64
Opportunity	1.9 (0.3)	2.0 (0.0)	2.0 (0.0)	1.8 (0.4)	.88
Flexibility	1.8 (0.4)	1.7 (0.5)	2.0 (0.0)	1.9 (0.3)	.82
Acceptability	1.8 (0.4)	1.3 (0.5)	1.2 (0.4)	1.5 (0.5)	.12
PPV[b]	1.3 (0.5)	1.3 (0.5)	1.0 (0.0)	1.4 (0.5)	.67
Simplicity	1.8 (0.4)	1.0 (0.0)	1.8 (0.4)	1.5 (0.5)	.03

[a]*P* value (ANOVA) of health surveillance system attributes considering the use of digital tools (apps, big data, Google Trends, and Twitter).

[b]PPV: positive predictive value.

Discussion

Principal Findings

This scoping review demonstrated different approaches for the use of digital tools to prevent and control arboviruses. The use of apps and Twitter in surveillance revealed the best results.

Due to the extensive health crisis caused by COVID-19, health agencies around the world have concentrated efforts to adopt strategies aimed at providing reliable information for the population, detecting symptoms, providing first care in suspected cases, and supporting the detection of new cases and laboratory diagnosis using apps [95]. Therefore, the use of apps has proved to be beneficial, not only for surveillance, but also as a valuable aid in dealing with public health emergencies. Initiatives in Brazil, such as the COVID-19 observatory, Infogripe, and Infodengue, are used to monitor and to inform society about health problems online. In our review, 88% of the studies used a digital tool for tracking cases or monitoring arbovirus outbreaks, and most of them used social media data or apps, which is in line with the measures adopted in the response to the COVID-19 pandemic. Furthermore, our review showed a higher score for apps, indicating statistical relevance in the use of apps to disseminate information to society (P=.02), communication (cases and bulletins; P=.01), and ease of operationalization of the surveillance system (P=.03), with the highest means in the evaluation of scores.

Of the 24 studies that mentioned the use of apps, 96% pointed to monitoring outbreaks or epidemics and 92% mentioned tracking cases. In agreement, a systematic review conducted by Quinn et al [96] evaluated studies of web-based apps, indicator-based surveillance, and the response to communicable disease outbreaks. Their review highlighted the use of apps to improve the early detection of disease outbreaks and disease notification, as well as the active participation of users; however, they indicated a low PPV [96]. As we also observed in our review, the use of apps in arbovirus surveillance contributes either to the opportunity and ease of operationalization of the surveillance system and case detection or to a reduction in costs. Furthermore, in our review, 62% of the studies underlined the availability of information (data) in real time, which implies the triggering of timely actions with an impact on prevention and control measures. It is even more important to have real-time data in an epidemic or pandemic situation, when information and data sources need to be available in a timely manner for the implementation of infection control measures and minimization of risk factors associated with the health of the population [97].

The correlation analysis between unofficial data from social media and official data of arbovirus surveillance may be helpful to assess the potential use of nonofficial data. Twitter data on influenza was monitored for a year in the United States. Data were collected and processed based on geographic information science (GIS) and data mining and then compared with official data from national, regional, and local reports of disease outbreaks. This correlation revealed strong statistical relevance between the data sources [98]. Samaras et al [99] reported the feasibility of building an early detection and forecasting system for influenza epidemics with data from Twitter and Google search engines. This process took place in real time for 23 weeks, and the data collected from the digital tools were compared with official data. The results pointed to a high correlation with Google data and to the usefulness of Twitter data. In our review, 12 studies [27,29,31,33-40,42] showed a statistically significant correlation (P<.05) or strong correlation between official data of arbovirus surveillance and unofficial data from online trends, social media, or big data. Of these, more than half addressed the use of Twitter data. However, data collection and processing are crucial steps that require investment in appropriate techniques. One study showed the correlation was positive and statistically significant but with several limitations [41].

Our data showed good results regarding the use of Twitter according to either surveillance activities or surveillance system indicators. Twitter was used to propose a framework to explore online data sources to mitigate the lack of epidemiological data, assess digital behaviors and complex interaction between new data streams induced by the chikungunya outbreak, identify public health problems during a dengue epidemic, identify rumors, track and monitor cases, and support arbovirus case prediction and early warning models.

The phenomenon of misinformation and fake news became notorious during the COVID-19 pandemic, where the use of social media intensified. The sharing of fake news is a social

XSL•FO
RenderX

problem that threatens public health. Jain [100] proposed an entropy approach to identify and monitor rumors related to COVID-19 based on shared tweets. In our review, the use of Twitter to identify rumors was statistically relevant in evaluating scores (*P*=.008), presenting a higher mean compared to other digital tools.

Moreover, our data showed 21 studies [28,29,32,34,35,43-58] that developed early warning models and enhanced the prediction, forecast, and detection of remergent events. The authors reported the use of apps, big data, Twitter, Google Trends, AI, the internet of things (IoT), and statistical models with algorithm adjustment to predict arbovirus cases. One study [29] presented the possibility of predicting dengue cases up to 8 weeks in advance using data from Twitter, Google Trends, and Wikipedia. In addition, the authors addressed the use of apps with geospatial and meteorological information capable of detecting and predicting possible breeding sites of the mosquito vector, predicting dengue outbreaks, generating detailed reports, and providing users with health education about dengue. Applications of fog computing were also discussed, as well as a model capable of merging large volumes of data through big data to generate early warnings. The use of Google Trends was also reported to predict the COVID-19 outbreak in India 2-3 weeks before routine surveillance [101]. Furthermore, our review showed that the use of Google Trends in arbovirus surveillance can facilitate operationalization of the surveillance system (*P*=.03), with the highest means in the evaluation of scores.

AI algorithms play a key role in rapidly predicting, detecting, classifying, sorting, and diagnosing an infection. An AI-based system is capable of accurately predicting changes in human behavior, contributing to the detection of and response to epidemic risks [102,103]. In this scoping review, one study [55] used AI and IoT as an approach to collect data and predict future situations and support preparedness and response.

Considering that we live in the big data era, and that society is increasingly connected, the use of data available on the web has been growing in several areas, although there is criticism regarding their use. The use of open, nonstructured data to obtain health outcomes requires not only the storage and processing of large volumes of data but also methodological concerns [104,105]. The balance regarding the quantity versus the quality of data remains a challenge. Nevertheless, statistical methods are being rapidly developed to meet public health demands based on the analysis of large volumes of data. The combination of data is a resource that expands the analysis capacity of a system, and this area of epidemiology is increasingly leading this scenario [105,106]. A strategy for expanding the use of social media in the surveillance area would be the "data science–based approach," encompassing multidisciplinary teams, and "app of techniques," with machine learning algorithms and natural language processing (NLP). In our review, the use of big data was addressed in 6 studies [28,30,37,43,51,93] that combined data from social media for tracking cases, monitoring outbreaks or epidemics, disseminating information to society, and identifying children with incomplete immunization. Moreover, the use of big data contributed to decision-making by health agencies and to prevention and control measures concerning emerging and reemerging infectious diseases.

The use of unofficial data from the internet and social media in surveillance has some limitations, such as information overload as well as uncertain quality and validity of data for surveillance purposes. Therefore, more evidence is needed regarding the efficacy and assessment of integrated systems. The phenomenon of information overload could be mitigated by investing in automated technology for monitoring health-related internet-based data so that these strategies could be adopted within the health surveillance system [105,107]. Leal et al [104] argued that in Brazil, there is an immediate lack of technological incorporation to reduce information time and improve the means used in the surveillance routine, which harms the "information for action" issue, the hallmark of public health surveillance.

Although searches by web sources are in continuous growth, it is important to emphasize that digital inclusion is limited worldwide [108]. The mitigation of biases related to the representativeness digital vehicles data is a complex process. Data about a certain disease may be underrepresented in Twitter due to the lack of digital coverage in different locations [106]. Thus, an important point to consider is the inequality of internet access to the population, which can limit the implementation of digital health surveillance strategies [109].

The validation of data through statistical techniques and other approaches is desirable to increase the reliability of the data and their use in the decision-making process for action [12]. It is important to emphasize the relevance of regulation of the use of data from digital tools to ensure the protection of the participant, especially the ethical aspects involved, even if they are available in the social networks [110].

Investment in innovation, technology, and digital tools in the routine of health surveillance is essential, especially as we are experiencing exponential technological advances and an increase in public health demands [111]. However, there is a major methodological challenge in validating information collected from unofficial sources. Additionally, it is necessary to review the regulations to support alternative and complementary surveillance systems, as described.

Limitations

Our scoping review has some limitations related to the choice of the arboviruses DENV, ZIKV, and CHIKV, due to their clinical and epidemiological importance. Furthermore, there are some limitations regarding the assessment of accessibility and digital inclusion of the populations studied. However, these issues were not found in the included studies, despite the search including 6 databases and gray literature, without language and period restrictions.

Conclusion

Our review outlined the use of several digital tools for arbovirus surveillance, with emphasis on the use of apps and Twitter in surveillance. These tools can contribute to surveillance in a complementary way and strengthen the following aspects: dissemination of information to society, rumor identification, information opportunity, acceptability of users to participate in

XSL•FO
RenderX

the system, capacity to adapt to new epidemiological situations, monitoring of outbreaks or epidemics, case detection and tracking, operationalization of the system, and reduction in costs.

In the contemporary scenario, it is no longer possible to ignore the use of web data or social media as a complementary strategy to health surveillance. However, it is important that efforts be combined to develop new methods that can ensure the quality of information and the adoption of systematic measures to maintain the integrity and reliability of digital tools' data, considering ethical aspects.

Acknowledgments

We thank the Professional Master Program in Clinical Research at the Evandro Chagas National Institute of Infectious Diseases, where the first author developed the study to obtain their master's degree.

Data Availability

The data sets used during the study are available from the corresponding author upon request.

Authors' Contributions

CLM and MDW conceived the study. CLM, LRMN, and MDW performed the review steps. CLM and DPP conceived the analysis. MDW and LG supervised the analysis. CLM, LG, and MDW were involved in manuscript preparation. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist and PRISMA flow diagram.
[DOCX File , 96 KB - jmir_v26i1e57476_app1.docx]

Multimedia Appendix 2
Search strategy.
[PDF File (Adobe PDF File), 287 KB - jmir_v26i1e57476_app2.pdf]

Multimedia Appendix 3
Health surveillance activities according to the use of digital tools.
[PDF File (Adobe PDF File), 225 KB - jmir_v26i1e57476_app3.pdf]

Multimedia Appendix 4
Health surveillance system attributes according to the use of digital tools.
[PDF File (Adobe PDF File), 226 KB - jmir_v26i1e57476_app4.pdf]

Multimedia Appendix 5
General characteristics of studies on the use of digital tools for surveillance.
[XLSX File (Microsoft Excel File), 29 KB - jmir_v26i1e57476_app5.xlsx]

References

1. Lorenz C, Azevedo TS, Virginio F, Aguiar BS, Chiaravalloti-Neto F, Suesdek L. Impact of environmental factors on neglected emerging arboviral diseases. PLoS Negl Trop Dis 2017 Sep 27;11(9):e0005959 [FREE Full text] [doi: 10.1371/journal.pntd.0005959] [Medline: 28953892]
2. Guzman MG, Gubler DJ, Izquierdo A, Martinez E, Halstead SB. Dengue infection. Nat Rev Dis Primers 2016 Aug 18;2(1):16055. [doi: 10.1038/nrdp.2016.55] [Medline: 27534439]
3. Robinson MC. An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952-53. I. Clinical features. Trans R Soc Trop Med Hyg 1955 Jan;49(1):28-32. [doi: 10.1016/0035-9203(55)90080-8] [Medline: 14373834]
4. Schuffenecker I, Iteman I, Michault A, Murri S, Frangeul L, Vaney M, et al. Genome microevolution of chikungunya viruses causing the Indian Ocean outbreak. PLoS Med 2006 Jul 23;3(7):e263 [FREE Full text] [doi: 10.1371/journal.pmed.0030263] [Medline: 16700631]

XSL•FO
RenderX

5. Bartholomeeusen K, Daniel M, LaBeaud DA, Gasque P, Peeling RW, Stephenson KE, et al. Chikungunya fever. Nat Rev Dis Primers 2023 Apr 06;9(1):17 [FREE Full text] [doi: 10.1038/s41572-023-00429-2] [Medline: 37024497]
6. Costa LB, Barreto FKDA, Barreto MCA, Santos THPD, Andrade MDMOD, Farias LABG, et al. Epidemiology and economic burden of chikungunya: a systematic literature review. Trop Med Infect Dis 2023 May 31;8(6):301 [FREE Full text] [doi: 10.3390/tropicalmed8060301] [Medline: 37368719]
7. Musso D, Gubler DJ. Zika virus. Clin Microbiol Rev 2016 Jul;29(3):487-524. [doi: 10.1128/cmr.00072-15]
8. Brasil P, Pereira JP, Moreira ME, Ribeiro Nogueira RM, Damasceno L, Wakimoto M, et al. Zika virus infection in pregnant women in Rio de Janeiro. N Engl J Med 2016 Dec 15;375(24):2321-2334. [doi: 10.1056/nejmoa1602412]
9. de Araújo TVB, Rodrigues LC, de Alencar Ximenes RA, de Barros Miranda-Filho D, Montarroyos UR, de Melo APL, et al. Association between zika virus infection and microcephaly in Brazil, January to May, 2016: preliminary report of a case-control study. Lancet Infect Dis 2016 Dec;16(12):1356-1363. [doi: 10.1016/s1473-3099(16)30318-8]
10. BIREME. Pan American Health Organization. 2022. URL: https://www.paho.org/en/bireme [accessed 2023-09-03]
11. Wilder-Smith A, Gubler DJ, Weaver SC, Monath TP, Heymann DL, Scott TW. Epidemic arboviral diseases: priorities for research and public health. Lancet Infect Dis 2017 Mar;17(3):e101-e106. [doi: 10.1016/s1473-3099(16)30518-7]
12. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. J Epidemiol Community Health 2020 Jun 24;74(7):612-616 [FREE Full text] [doi: 10.1136/jech-2018-211654] [Medline: 32332114]
13. Surveillance in emergencies. World Health Organization. 2022. URL: https://www.who.int/emergencies/surveillance [accessed 2023-09-03]
14. Celuppi I, Lima GDS, Rossi E, Wazlawick R, Dalmarco E. Uma análise sobre o desenvolvimento de tecnologias digitais em saúde para o enfrentamento da COVID-19 no Brasil e no mundo. Cad Saúde Pública. 2021 Mar 12. URL: https://doi.org/10.1590/0102-311X00243220 [accessed 2023-09-05]
15. Novoa C, Netto A. Fundamentos em gestão e informática em saúde Internet. UNIFESP - Departamento de Informática em Saúde. 2019. URL: https://repositorio.unifesp.br/handle/11600/51788 [accessed 2023-09-05]
16. Zamora A, Galán-Rodas E, Ramírez E, Rodríguez-Morales AJ, Mayta-Tristán P. Videojuego pueblo pitanga en la lucha contra el dengue en Costa Rica. Rev Peru Med Exp Salud Publica 2015 Jun 19;32(2):397. [doi: 10.17843/rpmesp.2015.322.1640]
17. Simonsen L, Gog JR, Olson D, Viboud C. Infectious disease surveillance in the big data era: towards faster and locally relevant systems. J Infect Dis 2016 Dec 01;214(suppl_4):S380-S385 [FREE Full text] [doi: 10.1093/infdis/jiw376] [Medline: 28830112]
18. Siddiqui TR, Ghazal S, Bibi S, Ahmed W, Sajjad SF. Use of the Health Belief Model for the assessment of public knowledge and household preventive practices in Karachi, Pakistan, a dengue-endemic city. PLoS Negl Trop Dis 2016 Nov 10;10(11):e0005129 [FREE Full text] [doi: 10.1371/journal.pntd.0005129] [Medline: 27832074]
19. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
20. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol 2012 Jul;8(7):e1002616 [FREE Full text] [doi: 10.1371/journal.pcbi.1002616] [Medline: 22844241]
21. Salathé M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy 2018 Jan 04;14(1):1 [FREE Full text] [doi: 10.1186/s40504-017-0065-7] [Medline: 29302758]
22. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
23. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
24. Updated guidelines for evaluating public health surveillance systems. Centers for Disease Control and Prevention. 2001. URL: https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm [accessed 2023-09-03]
25. Guia de Vigilância Epidemiológica - 7 edição. Biblioteca Virtual em Saúde Ministério da Saúde. 2009. URL: https://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_epidemiologica_7ed [accessed 2023-12-16]
26. Regulamento Sanitário Internacional (2005). 1. ed. Organização Mundial de Saúde. 2009. URL: https://www.gov.br/anvisa/pt-br/assuntos/paf/regulamento-sanitario-internacional/arquivos/7181json-file-1 [accessed 2023-12-16]
27. Antunes M, Silva C, Guimaraes C, Rabaco M. Social media monitoring: the dengue e-monitor. TransInformação 2014 Jan 1;26:9-18.
28. Lee C, Yang H, Lin S. Incorporating big data and social sensors in a novel early warning system of dengue outbreaks. 2015 Presented at: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Paris, France, , pp. 1428?33; 2015; Paris, France. [doi: 10.1145/2808797.2808883]
29. Marques-Toledo CDA, Degener CM, Vinhal L, Coelho G, Meira W, Codeço CT, et al. Dengue prediction by the web: tweets are a useful tool for estimating and forecasting dengue at country and city level. PLoS Negl Trop Dis 2017 Jul 18;11(7):e0005729 [FREE Full text] [doi: 10.1371/journal.pntd.0005729] [Medline: 28719659]

XSL•FO
RenderX

30. Carlos MA, Nogueira M, Machado RJ. Analysis of dengue outbreaks using big data analytics and social networks. 2017 Presented at: 4th International Conference on Systems and Informatics (ICSAI); November 1, 2017; Hangzhou, China p. 1592-1597. [doi: 10.1109/icsai.2017.8248538]
31. Husnayain A, Fuad A, Lazuardi L. Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Glob Health Action 2019;12(1):1552652 [FREE Full text] [doi: 10.1080/16549716.2018.1552652] [Medline: 31154985]
32. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS Negl Trop Dis 2017 Jan;11(1):e0005295 [FREE Full text] [doi: 10.1371/journal.pntd.0005295] [Medline: 28085877]
33. Adebayo G, Neumark Y, Gesser-Edelsburg A, Abu Ahmad W, Levine H. Zika pandemic online trends, incidence and health risk communication: a time trend study. BMJ Glob Health 2017;2(3):e000296 [FREE Full text] [doi: 10.1136/bmjgh-2017-000296] [Medline: 29082006]
34. Teng Y, Bi D, Xie G, Jin Y, Huang Y, Lin B, et al. Dynamic forecasting of zika epidemics using Google Trends. PLoS One 2017 Jan 6;12(1):e0165085 [FREE Full text] [doi: 10.1371/journal.pone.0165085] [Medline: 28060809]
35. Masri S, Jia J, Li C, Zhou G, Lee M, Yan G, et al. Use of Twitter data to improve zika virus surveillance in the United States during the 2016 epidemic. BMC Public Health 2019 Jun 14;19(1):761 [FREE Full text] [doi: 10.1186/s12889-019-7103-8] [Medline: 31200692]
36. Mahroum N, Adawi M, Sharif K, Waknin R, Mahagna H, Bisharat B, et al. Public reaction to chikungunya outbreaks in Italy-insights from an extensive novel data streams-based structural equation modeling analysis. PLoS One 2018;13(5):e0197337 [FREE Full text] [doi: 10.1371/journal.pone.0197337] [Medline: 29795578]
37. Klein GH, Guidi Neto P, Tezza R. Big data e mídias sociais: monitoramento das redes como ferramenta de gestão. Saude Soc 2017 Mar;26(1):208-217. [doi: 10.1590/S0104-12902017164943]
38. Rodriguez-Valero N, Luengo Oroz M, Cuadrado Sanchez D, Vladimirov A, Espriu M, Vera I, et al. Mobile based surveillance platform for detecting zika virus among Spanish delegates attending the Rio de Janeiro Olympic Games. PLoS One 2018 Aug 22;13(8):e0201943 [FREE Full text] [doi: 10.1371/journal.pone.0201943] [Medline: 30133492]
39. Monnaka VU, de Oliveira CAC. Google Trends correlation and sensitivity for outbreaks of dengue and yellow fever in the state of São Paulo. Einstein (São Paulo) 2021 Aug;19:eAO5969. [doi: 10.31744/einstein_journal/2021ao5969]
40. Provenzano S, Gianfredi V, Santangelo OE. Insight the data: Wikipedia's researches and real cases of arboviruses in Italy. Public Health 2021 Mar;192:21-29. [doi: 10.1016/j.puhe.2020.12.010] [Medline: 33607517]
41. Coberly JS, Fink CR, Elbert Y, Yoon IK, Velasco JM, Tomayao AD, et al. Tweeting fever: can twitter be used to monitor the incidence of dengue-like illness in the Philippines? Johns Hopkins University Applied Physics Laboratory. 2014. URL: https://secwww.jhuapl.edu/techdigest/content/techdigest/pdf/V32-N04/32-04-Coberly.pdf [accessed 2024-11-02]
42. Gomide J, Veloso A, Meira W, Almeida V, Benevenuto F, Ferraz F. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. 2011 Presented at: WebSci '11: 3rd International Web Science Conference; June 15-17, 2011; Koblenz, Germany URL: https://doi.org/10.1145/2527031.2527049 [doi: 10.1145/2527031.2527049]
43. Othman MK, Danuri MSNM. Proposed conceptual framework of Dengue Active Surveillance System (DASS) in Malaysia. 2016 Presented at: 2016 International Conference on Information and Communication Technology (ICICTM); 2016; Kuala Lumpur, Malasya p. 90-96. [doi: 10.1109/icictm.2016.7890783]
44. Leal-Neto OB, Dimech GS, Libel M, de Souza WV, Cesse E, Smolinski M, et al. Saúde na Copa: the world's first application of participatory surveillance for a mass gathering at FIFA World Cup 2014, Brazil. JMIR Public Health Surveill 2017 May 04;3(2):e26 [FREE Full text] [doi: 10.2196/publichealth.7313] [Medline: 28473308]
45. Kassim M, Ali N, Idris A, Shahbudin S, Rahman R. Dengue attack analysis system on mobile application. 2018 Presented at: 2018 IEEE 8th International Conference on System Engineering and Technology (ICSET); October 15-16, 2018; Bandung, Indonesia p. 151-156. [doi: 10.1109/icsengt.2018.8606397]
46. Ibrahim N, Keong T. Development of aedes-entomological predictive analytical dashboard application. 2018 Presented at: 2018 Seventh ICT International Student Project Conference (ICT-ISPC); 2018; Nakhonpathom, Thailand p. 1-5. [doi: 10.1109/ict-ispc.2018.8523990]
47. Asat A, Mahat A, Hassan R, Mohamed SAA. Development of dengue detection and prevention system (Deng-E) based upon open data in Malaysia. 2107 Presented at: 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI); 2017; Langkawi, Malaysia p. 1-6. [doi: 10.1109/iceei.2017.8312426]
48. Babu AN, Niehaus E, Shah S, Unnithan C, Ramkumar PS, Shah J, et al. Smartphone geospatial apps for dengue control, prevention, prediction, and education: MOSapp, DISapp, and the mosquito perception index (MPI). Environ Monit Assess 2019 Jun 28;191(Suppl 2):393. [doi: 10.1007/s10661-019-7425-0] [Medline: 31254076]
49. Marques-Toledo CA, Bendati MM, Codeço CT, Teixeira MM. Probability of dengue transmission and propagation in a non-endemic temperate area: conceptual model and decision risk levels for early alert, prevention and control. Parasit Vectors 2019 Jan 16;12(1):38 [FREE Full text] [doi: 10.1186/s13071-018-3280-z] [Medline: 30651125]
50. Somboonsak P. Development innovation to predict dengue affected area and alert people with smartphones. Int J Online Biomed Eng IJOE 2020 Feb 12;16(02):62-79 [FREE Full text] [doi: 10.3991/ijoe.v16i02.12425]

XSL•FO
RenderX

51. Rocklöv J, Tozan Y, Ramadona A, Sewe MO, Sudre B, Garrido J, et al. Using big data to monitor the introduction and spread of chikungunya, Europe, 2017. Emerg Infect Dis 2019 Jun;25(6):1041-1049 [FREE Full text] [doi: 10.3201/eid2506.180138] [Medline: 31107221]
52. Polanco González C, Islas Vazquez I, Castañón González JA, Buhse T, Arias-Estrada M. Electronic devices that identify individuals with fever in crowded places: a prototype. Micromachines (Basel) 2017 Jun 24;8(7):202 [FREE Full text] [doi: 10.3390/mi8070202] [Medline: 30400392]
53. Guo P, Zhang Q, Chen Y, Xiao J, He J, Zhang Y, et al. An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data. Sci Total Environ 2019 Jan 10;647:752-762 [FREE Full text] [doi: 10.1016/j.scitotenv.2018.08.044] [Medline: 30092532]
54. Massaro E, Kondor D, Ratti C. Assessing the interplay between human mobility and mosquito borne diseases in urban environments. Sci Rep 2019 Nov 15;9(1):16911 [FREE Full text] [doi: 10.1038/s41598-019-53127-z] [Medline: 31729435]
55. Rahman MS, Safa NT, Sultana S, Salam S, Karamehic-Muratovic A, Overgaard HJ. Role of artificial intelligence-internet of things (AI-IoT) based emerging technologies in the public health response to infectious diseases in Bangladesh. Parasite Epidemiol Control 2022 Aug;18:e00266 [FREE Full text] [doi: 10.1016/j.parepi.2022.e00266] [Medline: 35975103]
56. Parikh N, Daughton A, Rosenberger W, Aberle D, Chitanvis M, Altherr F, et al. Improving detection of disease re-emergence using a web-based tool (RED Alert): design and case analysis study. JMIR Public Health Surveill 2021 Jan 07;7(1):e24132 [FREE Full text] [doi: 10.2196/24132] [Medline: 33316766]
57. Herbuela VRDM, Karita T, Carvajal TM, Ho HT, Lorena JMO, Regalado RA, et al. Early detection of dengue fever outbreaks using a surveillance app (Mozzify): cross-sectional mixed methods usability study. JMIR Public Health Surveill 2021 Mar 01;7(3):e19034 [FREE Full text] [doi: 10.2196/19034] [Medline: 33646128]
58. Cardenas R, Hussain-Alkhateeb L, Benitez-Valladares D, Sánchez-Tejeda G, Kroeger A. The Early Warning and Response System (EWARS-TDR) for dengue outbreaks: can it also be applied to chikungunya and zika outbreak warning? BMC Infect Dis 2022 Mar 07;22(1):235 [FREE Full text] [doi: 10.1186/s12879-022-07197-6] [Medline: 35255839]
59. Galván P, Cane V, Samudio M, Cabello A, Cabral M, Basogain X, et al. [Implementation of a community tele-epidemiological surveillance system using information and communication technologies in Paraguay]. Rev Panam Salud Publica 2014 Jun 1;35(5/6):353-358 [FREE Full text]
60. Olson D, Lamb M, Lopez MR, Colborn K, Paniagua-Avila A, Zacarias A, et al. Performance of a mobile phone app-based participatory syndromic surveillance system for acute febrile illness and acute gastroenteritis in rural Guatemala. J Med Internet Res 2017 Nov 09;19(11):e368 [FREE Full text] [doi: 10.2196/jmir.8041] [Medline: 29122738]
61. Leal-Neto OB, Cruz O, Albuquerque J, Nacarato de Sousa M, Smolinski M, Pessoa Cesse E, et al. Participatory surveillance based on crowdsourcing during the Rio 2016 Olympic Games using the Guardians of Health platform: descriptive study. JMIR Public Health Surveill 2020 Apr 07;6(2):e16119 [FREE Full text] [doi: 10.2196/16119] [Medline: 32254042]
62. Ruiz-Burga E, Bruijning-Verhagen P, Palmer P, Sandcroft A, Fernandes G, de Hoog M, ZIKAction Consortium. Detection of potential arbovirus infections and pregnancy complications in pregnant women in Jamaica using a smartphone app (ZIKApp): pilot evaluation study. JMIR Form Res 2022 Jul 27;6(7):e34423 [FREE Full text] [doi: 10.2196/34423] [Medline: 35896029]
63. Espina K, Estuar MRJE. Infodemiology for syndromic surveillance of dengue and typhoid fever in the Philippines. Procedia Comput Sci 2017;121:554-561 [FREE Full text] [doi: 10.1016/j.procs.2017.11.073]
64. Lwin M, Vijaykumar S, Fernando O, Cheong S, Rathnayake V, Lim G, et al. A 21st century approach to tackling dengue: crowdsourced surveillance, predictive mapping and tailored communication. Acta Trop 2014 Feb;130:100-107 [FREE Full text] [doi: 10.1016/j.actatropica.2013.09.021] [Medline: 24161879]
65. Lwin MO, Vijaykumar S, Rathnayake VS, Lim G, Panchapakesan C, Foo S, et al. A social media mHealth solution to address the needs of dengue prevention and management in Sri Lanka. J Med Internet Res 2016 Jul 01;18(7):e149 [FREE Full text] [doi: 10.2196/jmir.4657] [Medline: 27369296]
66. Lwin MO, Jayasundar K, Sheldenkar A, Wijayamuni R, Wimalaratne P, Ernst KC, et al. Lessons from the implementation of Mo-Buzz, a mobile pandemic surveillance system for dengue. JMIR Public Health Surveill 2017 Oct 02;3(4):e65 [FREE Full text] [doi: 10.2196/publichealth.7376] [Medline: 28970191]
67. Lima T, Barbosa B, Niquini C, Araujo C, Lana R. Playing against dengue design and development of a serious game to help tackling dengue. 2017 Presented at: IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH); April 2-4 2017; Perth, WA URL: http://www.segah.org/2017/docs/Papers/Poster%20Session/P-PS-08-40.pdf [doi: 10.1109/segah.2017.7939294]
68. Lima TFM. A game-based platform to tackle a public health problem. 2018 Presented at: 2018 Annual Symposium on Computer-Human Interaction in Play Companion (CHI PLAY '18); October 28-31, 2018; Melbourne, VIC p. 17-25 URL: https://doi.org/10.1145/3270316.3270605 [doi: 10.1145/3270316.3270605]
69. Sreeram S, Shanmugam L. Autonomous robotic system based environmental assessment and dengue hot-spot identification. 2018 Presented at: IEEE International Conference on Environment and Electrical Engineering and IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe); 2018; Palermo, Italy p. 1-6. [doi: 10.1109/eeeic.2018.8493849]
70. Singh S, Bansal A, Sandhu R, Sidhu J. Fog computing and IoT based healthcare support service for dengue fever. IJPCC 2018 Jun 04;14(2):197-207 [FREE Full text] [doi: 10.1108/ijpcc-d-18-00012]

XSL•FO
RenderX

71. Lwin M, Sheldenkar A, Panchapakesan C, Ng J, Lau J, Jayasundar K, et al. Epihack Sri Lanka: development of a mobile surveillance tool for dengue fever. BMC Med Inform Decis Mak 2019 Jun 13;19(1):111 [FREE Full text] [doi: 10.1186/s12911-019-0829-5] [Medline: 31196073]
72. Minhas K, Tabassam M, Rasheed R, Abbas A, Khattak H, Khan S. A framework for dengue surveillance and data collection in Pakistan. 2019 Presented at: IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC); July 15-19, 2019; Milwaukee, WI p. 275-280. [doi: 10.1109/compsac.2019.10219]
73. Lwin M, Ng J, Jayasundar K, Kensinger A, Tan S. Visual design for a mobile pandemic map system for public health. AI Soc 2020 Jan 18;36(4):1349-1360 [FREE Full text] [doi: 10.1007/s00146-020-00939-7]
74. Idriani E, Rahmaniati MM, Kurniawan R. Dengue surveillance information system: an Android-based early warning system for the outbreak of Dengue in Padang, Indonesia. Ind J Public Health Res Dev 2019;10(5):1386. [doi: 10.5958/0976-5506.2019.01124.0]
75. Souza RC, Assunção RM, Oliveira DM, Neill DB, Meira W. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. Spat Spatiotemporal Epidemiol 2019 Jun;29:163-175. [doi: 10.1016/j.sste.2018.11.005] [Medline: 31128626]
76. Berbudi A, Sujatmiko B, Wiraswati HL, Rohmawaty E. The potential use of open data kit application for the mosquito larvae monitoring program to control dengue vector in Indonesia. Biosci Biotech Res Comm 2020 Mar 30;13(1):79-83 [FREE Full text] [doi: 10.21786/bbrc/13.1/14]
77. Herbuela VRDM, Karita T, Francisco ME, Watanabe K. An integrated mHealth app for dengue reporting and mapping, health communication, and behavior modification: development and assessment of Mozzify. JMIR Form Res 2020 Jan 08;4(1):e16424 [FREE Full text] [doi: 10.2196/16424] [Medline: 31913128]
78. Sareen S, Sood SK, Gupta SK. Secure internet of things-based cloud framework to control zika virus outbreak. Int J Technol Assess Health Care 2017 Apr 24;33(1):11-18. [doi: 10.1017/s0266462317000113]
79. Beltran J, Boscor A, dos SW, Massoni T, Kostkova P. Zika: a new system to empower health workers and local communities to improve surveillance protocols by e-learning and to forecast zika virus in real time in Brazil. 2018 Presented at: DH '18: 2018 International Conference on Digital Health; April 23-26, 2018; Lyon, France p. 90 URL: https://doi.org/10.1145/3194658.3194683 [doi: 10.1145/3194658.3194683]
80. Ocampo CB, Mina NJ, Echavarria MI, Acuña M, Caballero A, Navarro A, et al. VECTOS: an integrated system for monitoring risk factors associated with urban arbovirus transmission. Glob Health Sci Pract 2019 Mar 29;7(1):128-137. [doi: 10.9745/ghsp-d-18-00300]
81. Rodríguez S, Sanz A, Llano G, Navarro A, Parra-Lara L, Krystosik A, et al. Acceptability and usability of a mobile application for management and surveillance of vector-borne diseases in Colombia: an implementation study. PLoS One 2020;15(5):e0233269 [FREE Full text] [doi: 10.1371/journal.pone.0233269] [Medline: 32469894]
82. Amin S, Uddin M, alSaeed D, Khan A, Adnan M. Early detection of seasonal outbreaks from Twitter data using machine learning approaches. Complexity 2021 Mar 15;2021(3):1-12 [FREE Full text] [doi: 10.1155/2021/5520366]
83. Gong J, Li S, Lee J. Space, time, and disease on social media: a case study of dengue fever in China. Geomatica 2018 Dec 01;72(4):112-126 [FREE Full text] [doi: 10.1139/geomat-2018-0016]
84. Shahid F, Ony SH, Albi TR, Chellappan S, Vashistha A, Islam ABMAA. Learning from tweets: opportunities and challenges to inform policy making during dengue epidemic. Proc ACM Hum-Comput Interact 2020 May 29;4(CSCW1):1-27. [doi: 10.1145/3392875]
85. Vijaykumar S, Meurzec RW, Jayasundar K, Pagliari C, Fernandopulle Y. What's buzzing on your feed? Health authorities' use of Facebook to combat Zika in Singapore. J Am Med Inform Assoc 2017 Nov 01;24(6):1155-1159 [FREE Full text] [doi: 10.1093/jamia/ocx028] [Medline: 28449047]
86. Miller M, Banerjee T, Muppalla R, Romine W, Sheth A. What are people tweeting about zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. JMIR Public Health Surveill 2017 Jun 19;3(2):e38 [FREE Full text] [doi: 10.2196/publichealth.7157] [Medline: 28630032]
87. Lwin M, Lu J, Sheldenkar A, Schulz P. Strategic uses of Facebook in zika outbreak communication: implications for the crisis and emergency risk communication model. Int J Environ Res Public Health 2018 Sep 10;15(9):1974 [FREE Full text] [doi: 10.3390/ijerph15091974] [Medline: 30201929]
88. Chan MPS, Winneg K, Hawkins L, Farhadloo M, Jamieson KH, Albarracín D. Legacy and social media respectively influence risk perceptions and protective behaviors during emerging health threats: a multi-wave analysis of communications on zika virus cases. Soc Sci Med 2018 Sep;212:50-59 [FREE Full text] [doi: 10.1016/j.socscimed.2018.07.007] [Medline: 30005224]
89. Vijaykumar S, Nowak G, Himelboim I, Jin Y. Virtual zika transmission after the first U.S. case: who said what and how it spread on Twitter. Am J Infect Control 2018 May;46(5):549-557. [doi: 10.1016/j.ajic.2017.10.015] [Medline: 29306490]
90. Sousa L, de Mello R, Cedrim D, Garcia A, Missier P, Uchôa A, et al. VazaDengue: an information system for preventing and combating mosquito-borne diseases with social networks. Inf Syst 2018 Jun;75:26-42 [FREE Full text] [doi: 10.1016/j.is.2018.02.003]
91. Abouzahra M, Tan J. Twitter vs. zika—the role of social media in epidemic outbreaks surveillance. Health Policy Technol 2021 Mar;10(1):174-181 [FREE Full text] [doi: 10.1016/j.hlpt.2020.10.014]

XSL•FO
RenderX

92. Albinati J, Meira JW, Pappa G, Teixeira M, Marques-Toledo C. Enhancement of epidemiological models for dengue fever based on Twitter data. 2017 Presented at: DH 2017: International Conference on Digital Health; July 2-5, 2017; London UK p. 109-118. [doi: 10.1145/3079452.3079464]
93. Zhou X, Lee E, Wang X, Lin L, Xuan Z, Wu D, et al. Infectious diseases prevention and control using an integrated health big data system in China. BMC Infect Dis 2022 Apr 06;22(1):344 [FREE Full text] [doi: 10.1186/s12879-022-07316-3] [Medline: 35387590]
94. Divi N, Smolinski M. EpiHacks, a process for technologists and health experts to cocreate optimal solutions for disease prevention and control: user-centered design approach. J Med Internet Res 2021 Dec 15;23(12):e34286 [FREE Full text] [doi: 10.2196/34286] [Medline: 34807832]
95. Souza J, Venturini J. Tecnologias e Covid-19 no Brasil: vigilância e desigualdade social na periferia do capitalismo. Heinrich Böll Stiftung, Rio de Janeiro Office. 2020 Jun 30. URL: https://br.boell.org/pt-br/2020/06/04/tecnologias-e-covid-19-no-brasil-vigilancia-e-desigualdade-social-na-periferia-do [accessed 2023-09-03]
96. Quinn E, Hsiao KH, Maitland-Scott I, Gomez M, Baysari MT, Najjar Z, et al. Web-based apps for responding to acute infectious disease outbreaks in the community: systematic review. JMIR Public Health Surveill 2021 Apr 21;7(4):e24330 [FREE Full text] [doi: 10.2196/24330] [Medline: 33881406]
97. Leal-Neto OB, Santos F, Lee J, Albuquerque J, Souza W. Prioritizing COVID-19 tests based on participatory surveillance and spatial scanning. Int J Med Inform 2020 Nov;143:104263 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104263] [Medline: 32877853]
98. Allen C, Tsou M, Aslam A, Nagel A, Gawron J. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. PLoS One 2016;11(7):e0157734 [FREE Full text] [doi: 10.1371/journal.pone.0157734] [Medline: 27455108]
99. Samaras L, García-Barriocanal E, Sicilia M. Comparing social media and Google to detect and predict severe epidemics. Sci Rep 2020 Mar 16;10(1):4747 [FREE Full text] [doi: 10.1038/s41598-020-61686-9] [Medline: 32179780]
100. Jain L. An entropy-based method to control COVID-19 rumors in online social networks using opinion leaders. Technol Soc 2022 Aug;70:102048 [FREE Full text] [doi: 10.1016/j.techsoc.2022.102048] [Medline: 35765463]
101. Venkatesh U, Gandhi PA. Prediction of COVID-19 outbreaks using Google Trends in India: a retrospective analysis. Healthc Inform Res 2020 Jul;26(3):175-184 [FREE Full text] [doi: 10.4258/hir.2020.26.3.175] [Medline: 32819035]
102. Dong J, Wu H, Zhou D, Li K, Zhang Y, Ji H, et al. Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China. J Med Syst 2021 Jul 24;45(9):84 [FREE Full text] [doi: 10.1007/s10916-021-01757-0] [Medline: 34302549]
103. Comito C, Pizzuti C. Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: a focused review. Artif Intell Med 2022 Jun;128:102286 [FREE Full text] [doi: 10.1016/j.artmed.2022.102286] [Medline: 35534142]
104. Leal-Neto OB, Albuquerque J, Souza W, Cesse E, Cruz O. Inovações disruptivas e as transformações da saúde pública na era digital. Cad Saúde Pública 2017 Nov 21;33(11):e00005717 [FREE Full text] [doi: 10.1590/0102-311x00005717]
105. Filho C, Porto AD. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol Serv Saúde 2015 Jun;24(2):325-332 [FREE Full text] [doi: 10.5123/s1679-49742015000200015]
106. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41:101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
107. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q 2014 Mar 06;92(1):7-33 [FREE Full text] [doi: 10.1111/1468-0009.12038] [Medline: 24597553]
108. Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (infoveillance): systematic review. J Med Internet Res 2020 Mar 13;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
109. Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros - TIC Domicílios 2018. CETIC. 2019 Oct 28. URL: https://cetic.br/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nos-domicilios-brasileiros-tic-domicilios-2018 [accessed 2023-09-06]
110. Leal-Neto OB, Dimech G, Libel M, Oliveira W, Ferreira J. Digital disease detection and participatory surveillance: overview and perspectives for Brazil. Rev Saúde Pública 2016;50:17 [FREE Full text] [doi: 10.1590/s1518-8787.2016050006201]
111. Dixon B, Grannis S, McAndrews C, Broyles A, Mikels-Carrasco W, Wiensch A, et al. Leveraging data visualization and a statewide health information exchange to support COVID-19 surveillance and response: application of public health informatics. J Am Med Inform Assoc 2021 Jul 14;28(7):1363-1373 [FREE Full text] [doi: 10.1093/jamia/ocab004] [Medline: 33480419]

Abbreviations

AI: artificial Intelligence
CDC: Centers for Disease Control and Prevention

XSL•FO
RenderX

CHIKV: chikungunya virus
DENV: dengue virus
IoT: internet of things
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ZIKV: zika virus

Edited by A Schwartz; submitted 17.02.24; peer-reviewed by Y Yan, J Mistry, S Pesälä; comments to author 30.05.24; revised version received 10.07.24; accepted 15.10.24; published 18.11.24.

Please cite as:
Melo CL, Mageste LR, Guaraldo L, Paula DP, Wakimoto MD
Use of Digital Tools in Arbovirus Surveillance: Scoping Review
J Med Internet Res 2024;26:e57476
URL: https://www.jmir.org/2024/1/e57476
doi:10.2196/57476
PMID:

©Carolina Lopes Melo, Larissa Rangel Mageste, Lusiele Guaraldo, Daniela Polessa Paula, Mayumi Duarte Wakimoto. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI-Based Noninvasive Blood Glucose Monitoring: Scoping Review

Pin Zhong Chan[1], BSN; Eric Jin[2], MBBS; Miia Jansson[3], BSN, PhD; Han Shi Jocelyn Chew[4], BSN, PhD

[1]Department of Nursing, Ng Teng Fong General Hospital, Singapore, Singapore
[2]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[3]Research Unit of Health Sciences and Technology, University of Oulu, Oulu, Finland
[4]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Han Shi Jocelyn Chew, BSN, PhD
Alice Lee Centre for Nursing Studies
Yong Loo Lin School of Medicine
National University of Singapore
10 Medical Drive
Singapore, 117597
Singapore
Phone: 65 65168687
Email: jocelyn.chew.hs@nus.edu.sg

Abstract

Background: Current blood glucose monitoring (BGM) methods are often invasive and require repetitive pricking of a finger to obtain blood samples, predisposing individuals to pain, discomfort, and infection. Noninvasive blood glucose monitoring (NIBGM) is ideal for minimizing discomfort, reducing the risk of infection, and increasing convenience.

Objective: This review aimed to map the use cases of artificial intelligence (AI) in NIBGM.

Methods: A systematic scoping review was conducted according to the Arksey O'Malley five-step framework. Eight electronic databases (CINAHL, Embase, PubMed, Web of Science, Scopus, The Cochrane-Central Library, ACM Digital Library, and IEEE Xplore) were searched from inception until February 8, 2023. Study selection was conducted by 2 independent reviewers, descriptive analysis was conducted, and findings were presented narratively. Study characteristics (author, country, type of publication, study design, population characteristics, mean age, types of noninvasive techniques used, and application, as well as characteristics of the BGM systems) were extracted independently and cross-checked by 2 investigators. Methodological quality appraisal was conducted using the Checklist for assessment of medical AI.

Results: A total of 33 papers were included, representing studies from Asia, the United States, Europe, the Middle East, and Africa published between 2005 and 2023. Most studies used optical techniques (n=19, 58%) to estimate blood glucose levels (n=27, 82%). Others used electrochemical sensors (n=4), imaging (n=2), mixed techniques (n=2), and tissue impedance (n=1). Accuracy ranged from 35.56% to 94.23% and Clarke error grid (A+B) ranged from 86.91% to 100%. The most popular machine learning algorithm used was random forest (n=10) and the most popular deep learning model was the artificial neural network (n=6). The mean overall checklist for assessment of medical AI score on the included papers was 33.5 (SD 3.09), suggesting an average of medium quality. The studies reviewed demonstrate that some AI techniques can accurately predict glucose levels from noninvasive sources while enhancing comfort and ease of use for patients. However, the overall range of accuracy was wide due to the heterogeneity of models and input data.

Conclusions: Efforts are needed to standardize and regulate the use of AI technologies in BGM, as well as develop consensus guidelines and protocols to ensure the quality and safety of AI-assisted monitoring systems. The use of AI for NIBGM is a promising area of research that has the potential to revolutionize diabetes management.

(J Med Internet Res 2024;26:e58892) doi:10.2196/58892

KEYWORDS

artificial intelligence; blood glucose; diabetes; noninvasive; self-monitoring; machine learning; scoping review; monitoring; management; health informatics; deep learning; accuracy; heterogeneity; mobile phone

XSL•FO
RenderX

Introduction

According to the International Diabetes Federation, around 537 million adults aged 20-79 years were diagnosed with diabetes in 2021, a number that has been projected to increase to 783 million in 2045 [1]. Chronic diabetes mellitus (DM) leads to many severe complications, including stroke, blindness, ulcers, kidney failure, and vascular damage [2]. DM management places a massive burden on health care expenditure, which has more than quadrupled to at least US $966 billion over the last 15 years [1]. The most common and possibly life-threatening complication of DM is hypoglycemia [3], where common symptoms include autonomic (anxiety, tremors, palpitations, and diaphoresis) and neuroglycopenic (blurred vision, dizziness, headache, and loss of consciousness) manifestations [4]. Therefore, individuals with DM are often advised to monitor their blood glucose levels regularly to detect and manage abnormalities [4]. However, current blood glucose monitoring (BGM) methods are often invasive and require repetitive pricking of a finger to obtain blood samples, predisposing individuals to pain, discomfort, and infection [5]. The threshold for the onset of hypoglycemia also differs among patients (ie, typically higher in patients with uncontrolled diabetes), indicating the need for personalized BGM strategies [4].

Besides invasive BGM techniques, minimally invasive and noninvasive techniques have been developed. The most common minimally invasive method adopts the glucose-oxidase principle where a wire-based sensor is inserted in the subcutaneous layer of the skin [6]. It involves a calibration process that measures the current signal from the interstitial fluid rather than from the blood [7]. However, frequent calibration is required to maintain sensor accuracy by using traditional invasive fingerpick samples as a reference [8]. Recent flash glucose monitoring uses factory calibration which does not require calibration by the user but this method requires frequent replacement of the needle electrode every 1-2 weeks [9].

Noninvasive blood glucose monitoring (NIBGM) is ideal for minimizing discomfort, reducing the risk of infection, and increasing convenience. The latest advancements include near-infrared spectroscopy (NIRS), photoplethysmography (PPG), Raman spectroscopy (RS), photoacoustic signals, and biosensors, like saliva and tears [10]. As noninvasive methods do not directly detect blood glucose levels from blood samples, artificial intelligence (AI) could be used to estimate and predict blood glucose levels based on specific features selected. The use of AI could also facilitate the personalized BGM to inform treatment options, including insulin initiation and titration [11,12]. Although AI algorithms have been widely used in various health care settings including decision support systems and warning systems for hypoglycemia in patients with T1DM, little is known regarding the applicability of noninvasive methods [13,14].

Several studies have explored noninvasive methods for measuring and monitoring blood glucose levels in patients with and those without DM [15-17]. However, these reviews did not cover the use of machine learning (ML) systems often embedded in these devices, nor did they perform a comprehensive analysis of the accuracy of the devices. Similarly, some reviews have focused on the use of AI approaches for diabetes diagnosis and management using optical sensors [17] and breath analysis [18]. While these reviews present a comprehensive analysis of the available and used ML models, they often only cover one method of data collection, such as optical sensors. Two reviews focused on heart rate variability analysis [19,20] while another review focused on both electrocardiography (ECG) and PPG signals [21]. Furthermore, another review focused on detailing an overview of ML and AI techniques in the field of DM detection and self-management but not on NIBGM [22]. Therefore, a comprehensive review of the existing literature is needed to understand the current status of the use of AI in NIBGM. Given the novelty of using AI in NIBGM systems, evidence on the accuracy and effectiveness of such technologies is limited. Thus, we conducted a scoping review to rapidly map the key concepts and evidence regarding the use of AI for continuous NIBGM. Our findings would scope the available evidence on this topic, and identify the existing research gaps to inform the value and direction of conducting a full systematic review [23].

Methods

Overview

This scoping review was conducted using Arksey and O'Malley's five-step framework and reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta - Analyses extension for Scoping Reviews) guidelines (Multimedia Appendix 1) [24].

Step 1: Identifying the Research Question

Our primary research question was as follows: (1) what are the use cases of AI-assisted noninvasive BGM systems? Our secondary research question was as follows: (2) what are the AI models developed for noninvasive BGM?

Step 2: Identifying Relevant Studies

A comprehensive literature search was conducted on 08 February 2023 across 8 major databases, namely CINAHL, Embase, PubMed, Web of Science, Scopus, The Cochrane-Central Library, ACM Digital Library, and IEEE Xplore. The following key terms were used in the search: "glucose monitoring"; "monitoring glucose"; "artificial intelligence"; "computer heuristics"; "fuzzy logic"; "knowledge bases"; "machine learning"; "natural language processing"; "neural networks"; and "sentiment analysis" (Multimedia Appendix 2). Content experts were consulted and previous reviews on similar topics were hand searched for additional relevant studies.

Step 3: Study Selection and Methodological Quality Assessment

After the search was completed, duplicate studies were identified and removed. The remaining papers had their abstracts screened in a double-blinded, independent manner by 2 investigators (PZC and HSJC) according to the following inclusion criteria: prospective and retrospective primary studies that described the use of AI for continuous noninvasive BGM among human

XSL•FO
RenderX

participants and written primarily in English. Papers were excluded if they: (1) were reviews or gray literature, (2) did not involve AI in the continuous NIBGM, (3) involved nonhuman participants, or (4) were primarily non–English-language papers. Disputes were resolved via discussion and consensus by the 2 investigators. Both investigators (PZC and HSJC) then combed the relevant journals, bibliography, and conference submissions to identify more relevant papers, and all papers then underwent a full-text sieve based on the inclusion and exclusion criteria.

The methodological quality appraisal of each study was conducted independently by 2 reviewers (PZC and EJ) using the Checklist for assessment of medical AI (ChAMAI) [25]. Reviewers rated each included study on 30 items representing 6 domains namely problem understanding, data understanding, data preparation, modeling, validation, and deployment. Each item (bolded to represent having a high priority) received a rating of OK (adequately addressed), mR (sufficient but improvable), or MR (inadequately addressed, corresponding to a score of 2, 1, and 0, respectively. Items on a low-priority (not bolded) received half the scores and the maximum total score is 50 [26]. Overall scores indicate the study quality to be low (0-19.5), medium (20-34.5), or high (35-50). Discrepancies were resolved by a third reviewer (HSJC).

Step 4: Data Charting

Data were extracted independently and cross-checked by 2 investigators (PZC and HSJC). Any disputes were resolved via discussion and consensus. Study characteristics extracted included the author names, country, type of publication, study design, population characteristics, mean age, types of noninvasive techniques used, and application, as well as characteristics of the BGM systems (use of AI, AI type, AI features, types of data imputed, technology used, dataset, validation, proportion of training and testing dataset, and metrics used).

Results

Step 5: Collating, Summarizing, and Reporting the Results

Our initial search yielded 1270 studies. After removing duplicated citations, screening through 848 titles and abstracts, and 65 full-text papers, 33 papers were included in this scoping review (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA-ScR flowchart. PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta - Analyses extension for Scoping Reviews.

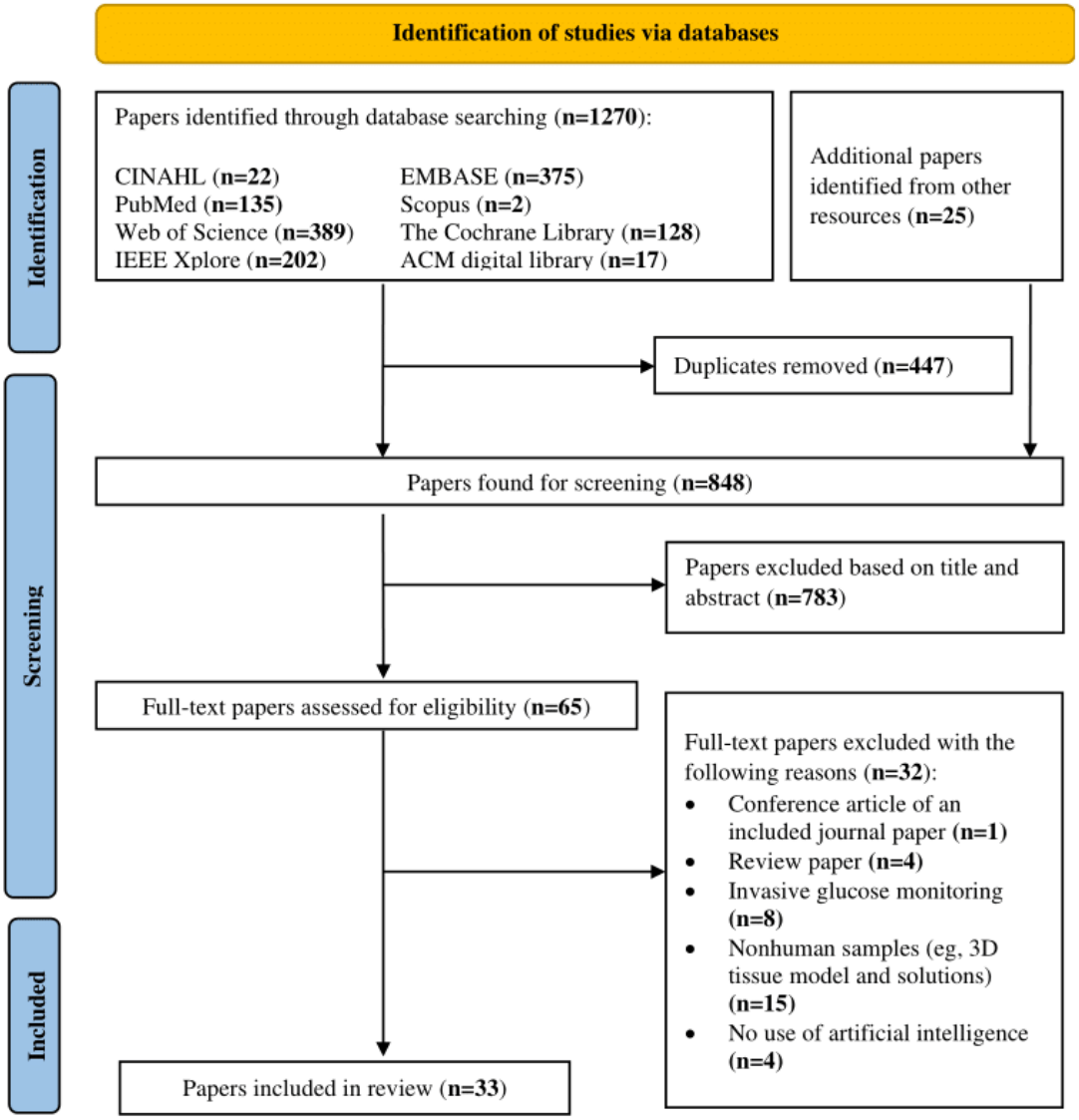

Study Characteristics

Most of the included studies were conducted in Asia (n=21, 64%) [10,27-46], were peer-reviewed journal papers (n=24, 73%) [10,30-41,47-57], were prospective cohort studies (n=20, 61%) [30-33,35-43,47,49,51-53,57,58], and used optical (eg, near-infrared and PPG) techniques (n=19, 58%) [10,27,29,31,33,35-37,39,40,42,44,45,47,51,53,56-58] to estimate blood glucose levels (n=27, 82%) [27,29-31,33-38,40-43,45-48,50-53,55-58] (Table 1). Most of the studies did not report the population characteristics [10,27,29,33,35-39,41,42,44-46,48,54,56-58], mean age [10,27-36,39,41-43,45,46,48,50-53,55,57], and sex [27-29,33,35,36,38-41,43-48,50-52,54,55,57,58].

XSL•FO
RenderX

Table 1. Summary of study characteristics.

Study characteristics	Values (N=33), n (%)
Country	
Algeria [48]	1 (3)
Bangladesh [27]	1 (3)
China [33,35,39-41]	5 (15)
China (Hong Kong) [30]	1 (3)
India [10,31,32,34,43-45]	7 (21)
Indonesia [42]	1 (3)
Israel [37]	1 (3)
Malaysia [46]	1 (3)
Mexico [47,58]	2 (6)
Netherlands [52]	1 (3)
South Korea [36,38]	2 (6)
Spain [56]	1 (3)
Sri Lanka [28,29]	2 (6)
United States [49-51,53-55,57]	7 (21)
Type of publication	
Journal paper [10,30-41,47-57]	24 (73)
Conference papers [27-29,42-46,58]	9 (27)
Study design	
Prospective cohort [30-33,35-43,47,49,51-53,57,58]	20 (61)
Retrospective cohort [10,28,44,48,50,54,56]	7 (21)
Others [27,29,34,45,46,55]	6 (18)
Population characteristics	
Type 1 DM[a] [50,52]	2 (6)
Type 2 DM [28]	1 (3)
Healthy [47,51]	2 (6)
Mixture [30-32,34,40,43,49,53,55]	9 (27)
NR[b] [10,27,29,33,35-39,41,42,44-46,48,54,56-58]	19 (58)
Age (years)	
21-40 [38,44,47,54,56,58]	6 (18)
40-65 [37,40,49]	3 (9)
NR [10,27-36,39,41-43,45,46,48,50-53,55,57]	24 (73)
Sex (male)	
0 [10]	1 (3)
<50 [34,49]	2 (6)
>50 [30-32,37,42,53,56]	7 (21)
NR [27-29,33,35,36,38-41,43-48,50-52,54,55,57,58]	23 (69.70)
Noninvasive techniques	
Optical (NIR[c], PPG[d], Raman) [10,27,29,31,33,35-37,39,40,42,44,45,47,51,53,56-58]	19 (58)
Impedance [55]	1 (3)
Biosensor (breath, saliva, tears) [30,32,34,52]	4 (12)

XSL•FO
RenderX

Study characteristics	Values (N=33), n (%)
Imaging (ECG[e], UWB[f]) [46,48]	2 (6)
Mixture [38,50]	2 (6)
NR [28,41,43,49,54]	5 (15)
Applications	
Predict DM [10,44]	2 (6)
Monitoring by physician [43,53]	2 (6)
Estimate BG[g] levels [10,27,29-42,45-48,50-52,54-58]	27 (82)
Estimate HbA_{1c}[h] levels [49]	1 (3)
Predict future BG levels [28]	1 (3)

[a]DM: diabetes mellitus.
[b]NR: not reported.
[c]NIR: near-infrared.
[d]PPG: photoplethysmography.
[e]ECG: electrocardiography.
[f]UWB: ultrawideband.
[g]BG: blood glucose.
[h]HbA_{1c}: hemoglobin A_{1c}.

Use Cases of AI-Assisted NIBGM

The majority of the use cases were to estimate blood glucose levels (n=29, 88%), [10,27,29-31,33-38,40-48,50-53,55-58], 3 (9%) were to detect DM [10,32,54], 1 (3%) was to estimate suitable insulin doses [44], and another was to predict future blood glucose level (Table 1 and Multimedia Appendix 3 [10,27-58]). Only one study used AI-assisted NIBGM to estimate hemoglobin A_{1c} (HbA_{1c}) levels and blood glucose variability among adults with prediabetes [49], which is noteworthy as glucose variability refers to oscillations in blood glucose levels throughout the day and could suggest the severity of diabetic complications.

BGM Technology

A summary of the technology used for NIBGM is shown in Multimedia Appendices 4 and 5 [10,27-58] and Figure 2. Of the 33 studies, 19 studies experimented with devices to estimate blood glucose levels using optical methods including PPG (n=7, 21%) [27,31,35,40,44,45,56], NIRS (n=7, 21%) [10,29,36,39,42,53,57], RS (n=1, 3%) [51], absorption spectroscopy (n=1, 3%) [33], noninvasive optical analysis of visible light capture by specialized cameras (n=1, 3%) [47], color image sensor (n=1, 3%) [37], and laser beam and light diode resistor (n=1, 3%) [58], 4 studies used devices that detected biological substances, including biosensor for tear glucose [37] and breath analysis [48] (two were not reported [30,52]), 3 (6%) studies used imaging techniques, including ECG [48] and ultrawideband (UWB) [46], 2 studies used mixed methods, including Optical+Electromagnetic+Thermal techniques [36] and Impedance+Multi-Wavelength NIR Spectroscopy [38], and 1 study used a device that measured tissue impedance (Multimedia Appendix 4 [10,27-58]). NIR spectroscopy detects the intensity of the reflected near-infrared light by glucose molecules in the blood to estimate blood glucose levels [35]. PPG operates on the same principles as that of the pulse oximeter, by calculating blood glucose levels based on the light intensity detected on a receiver and sent out by a transmitter [12]. RS works by comparing the Raman light emitted from a scattering medium (tissue) for transcutaneous determination of compositions of molecules, such as glucose, in the tissue-blood matrix [15,51]. RS can noninvasively monitor variations in glucose present at low concentrations in the blood-tissue matrix of the skin due to its distinct characteristic spectral features.

XSL•FO
RenderX

Figure 2. Blood glucose monitoring technology.

NIBGMs based on biosensors use breath, saliva, or tear samples to derive blood glucose concentrations based on their components, such as sodium, potassium, and calcium ions. ECGs measure the electrical activities of the heart and present them as a PQRST wave, with various abnormalities of the waves seemingly correlated with hyper- and hypoglycemia [49]. UWB imaging estimates blood glucose change via changes in the blood dielectric properties [46]. Two devices had mixed methods of detection. The device by Song et al [38] used both impedance and NIR to estimate blood glucose levels, while the device by Abubeker and Baskar [44] integrated ultrasonic, electromagnetic, and thermal data from the patient. Finally, the device by Malinin et al [55] measured the impedance of tissues via bracelet-type electrodes to detect blood glucose levels by monitoring the transfer functions of a tissue segment in the electromagnetic field.

A total of 19 (58%) devices used light-related signals as the input data (PPG signals [27,31,35,36,40,45,56], UWB imaging [46], NIR signals [10,29,39,53,57], Raman Spectra [51], nonvisible light signals [33,37,58], visible light signals [47], and LED light [42]), 5 (15%) devices collected biological samples (tears [52,59], breath [30,32], and saliva [34]), 4 (12%) devices used images or videos (video of finger [27,40], facial video [35], and image of finger or ear [57]), 4 (12%) devices collected vitals (eg, oxygen saturation, heart rate, and skin temperature [41,44,49,50]), 1 (3%) device collected impedance data [55], 1 (3%) device used ECG data [48], 1 (3%) device used a combination of types of inputs (intensity modulated photocurrent spectroscopy and multiwavelength NIRS) [38], 2 (6%) devices extracted various attributes from the patient's lifestyle and background, with one using medication intake, food intake, daily activities, and measured blood glucose levels as the input [28], and the other using pregnancy, BMI, insulin

XSL•FO
RenderX

level, age, blood pressure, skin thickness, glucose, and diabetes pedigree function [54].

AI Models Developed for NIBGM

A summary of the characteristics of ML is shown in Multimedia Appendix 5 [10,27-58]. The accuracy of NIBGM estimating blood glucose ranged from 35.56% [38] to 94.23% [38], mean absolute error (MAE) ranged from 0.248 [54] to 11.8 [15], R^2 ranged from 0.11 [44] to 0.91 [37], and Clarke error grid (CEG; A+B) ranged from 86.91% [50] to 100% [12,15,23,25,36,37,39,41]. Both MAE and R^2 were used to evaluate regression models, with lower MAE scores meaning a more accurate model and higher R^2 scores meaning a model that can cover a greater variety of data points. CEG was developed to measure the efficacy of BGM systems, and it consists of a grid divided into five zones. Zone A represents values that are clinically accurate and safe, while zones B, C, D, and E represent progressively more significant clinical errors. Typically, only data points within zones A and B are accepted by clinicians. A total of 8 (55%) devices achieved a CEG (A+B) of 100%, all of which included supervised learning models.

Various ML and deep learning (DL) algorithms were used. Nine [25,28,31,36,37,40,42,46,51] devices used only DL models (often some kind of neural network [NN]), while 8 [15,23,24,27,38,39,50,58] devices included only ML models. The rest used a mix of models. Among the DL models or NNs, 6 devices [2,31,40,42,49,56] used artificial neural networks (ANN), and three devices [12,25,54] used deep NNs. Among the ML models, 10 devices [12,14,22,27,34,38,41,44,53,54] used random forest (RF), 8 devices [12,34,41,43,44,49,54,56] used linear regression, 5 devices [14,26,29,34,49] used support vector machines (SVM) and 5 devices [22,41,44,53,54] used support vector regression. Datasets were split according to ratios ranging from 70:30 to the traditional 80:20 for training and testing according to different studies.

The most popular ML algorithm used was RF. Incidentally, RF is widely recognized as one of the most effective machine learning algorithms for classification tasks [22]. Increasing the number of trees in the forest improves prediction accuracy, allowing for tailored models based on specific characteristics. One study which used the use of RF had an accuracy of 94.2% [45], while another study that examined the use of RF to predict HbA_{1c} achieved a low mean average percent error of 4.87% [49].

Another popular algorithm used for data classification is SVM. SVM uses nonlinear mapping to transform DM training data into a higher dimension and seeks the optimal linear separating hyperplane [22]. It aims to create distinct margins between different classes, improving the training and testing speed. In a study based on salivary electrochemical signals, SVM outperformed other models in estimating blood glucose levels with 85% accuracy, 84% precision, and 85% sensitivity [34]. SVM had the best performance in another study which used PPG signals with an accuracy of 81.7% [44].

NNs are a popular DL model extensively used for the detection and diagnosis of DM. This was evident in a study that used CNN to estimate blood glucose levels using breath signals [32]. Performance was promising with a low mean square error of 0.14 and area under the curves as 0.97, 0.96, and 0.96 for T1DM, T2DM, and healthy, respectively [32]. ANN performed best when using input from the Pima Indian diabetes dataset, achieving an overall accuracy of 88.6% [54].

Study Quality

The mean overall ChAMAI score on the included papers was 33.5 (SD 3.09), suggesting an average of medium quality (Table 2). Most of the studies were of medium quality ranging between 30 and 41, while 10 studies were of high quality with a score equal to or more than 35. The proportion of “OK,” “mR,” and “MR” in high-priority items range from 20% to 80%, 0% to 6.7%, and 6.7% to 80%, respectively (Figure 3). The proportion of “OK,” “mR,” and “MR” in low-priority items range from 10% to 50%, 0% to 20%, and 50% to 90%, respectively (Figure 4). The interrater agreement in using ChAMAI indicated moderate agreement (Cohen κ=0.49).

XSL•FO
RenderX

Table 2. Study quality rated based on ChAMAI[a].

Author (Year)	Problem understanding (10)	Data understanding (6)	Data preparation (8)	Modeling (6)	Validation (12)	Deployment (8)	Overall (50)
Abubeker and Baskar (2022) [44]	7	4	5	6	7	4	33
Agrawal et al (2022) [10]	8	4	4	6	8	4	34
Alarcón-Paredes et al (2019) [47]	7	3	4	6	10	3	33
Ali et al (2016) [46]	7	4	5	6	9	3	34
Arbi et al (2023) [48]	7	3	5	6	7	4	32
Balasooriya and Nanayakkara (2020) [28]	7	3	4	5	7	3	29
Bent et al (2021) [49]	10	4	4	5	10	4	37
Bogue-Jimenez et al (2022) [50]	7	3	5	6	10	3	34
Enejder et al (2005) [51]	9	4	5	5	8	2	33
Francisco-García et al (2019) [58]	7	4	5	6	8	3	33
Geelhoed-Duijvestijn et al (2021) [52]	10	5	4	5	6	3	33
Guo et al (2012) [30]	7	3	5	6	6	3	30
Habbu et al (2019) [31]	8	4	5	5	9	3	34
Jain et al (2020) [53]	8	4	5	5	10	3	35
Khanam and Foo (2021) [54]	7	3	7	5	8	3	33
Krishnan et al (2020) [45]	5	3	4	5	6	2	25
Lekha and Suchetha (2018) [32]	7	4	5	6	9	3	34
Liu et al (2019) [33]	9	5	4	6	9	4	37
Malik et al (2016) [34]	10	4	5	6	9	3	37
Malinin (2012) [55]	6	3	5	6	9	2	31
Manurung et al (2019) [42]	7	5	5	6	10	3	36
Monte-Moreno (2011) [56]	9	6	7	6	10	3	41
Nanayakkara et al (2018) [29]	9	4	4	5	9	3	34
Nie et al (2023) [35]	9	5	5	6	10	3	38
Rachim and Chung (2019) [36]	6	3	4	6	9	3	31
Rajeshwaran et al (2022) [43]	6	3	5	6	8	3	31
Segman (2018) [37]	9	5	4	6	6	3	33
Song et al (2015) [38]	7	3	4	5	8	3	30

XSL•FO
RenderX

Author (Year)	Problem understanding (10)	Data understanding (6)	Data preparation (8)	Modeling (6)	Validation (12)	Deployment (8)	Overall (50)
Sumaiya et al (2020) [27]	5	3	5	6	8	3	30
Valero et al (2022) [57]	9	4	4	6	8	4	35
Yu et al (2021) [39]	9	4	5	6	8	3	35
Zhang et al (2020) [40]	9	6	5	6	9	3	38
Zhu et al (2021) [41]	6	4	4	6	8	3	31

[a]ChAMAI: Checklist for assessment of medical artificial intelligence.

Figure 3. Proportion of OK=adequately addressed, mR=sufficient but improvable, MR= inadequately addressed ratings on each high priority items.

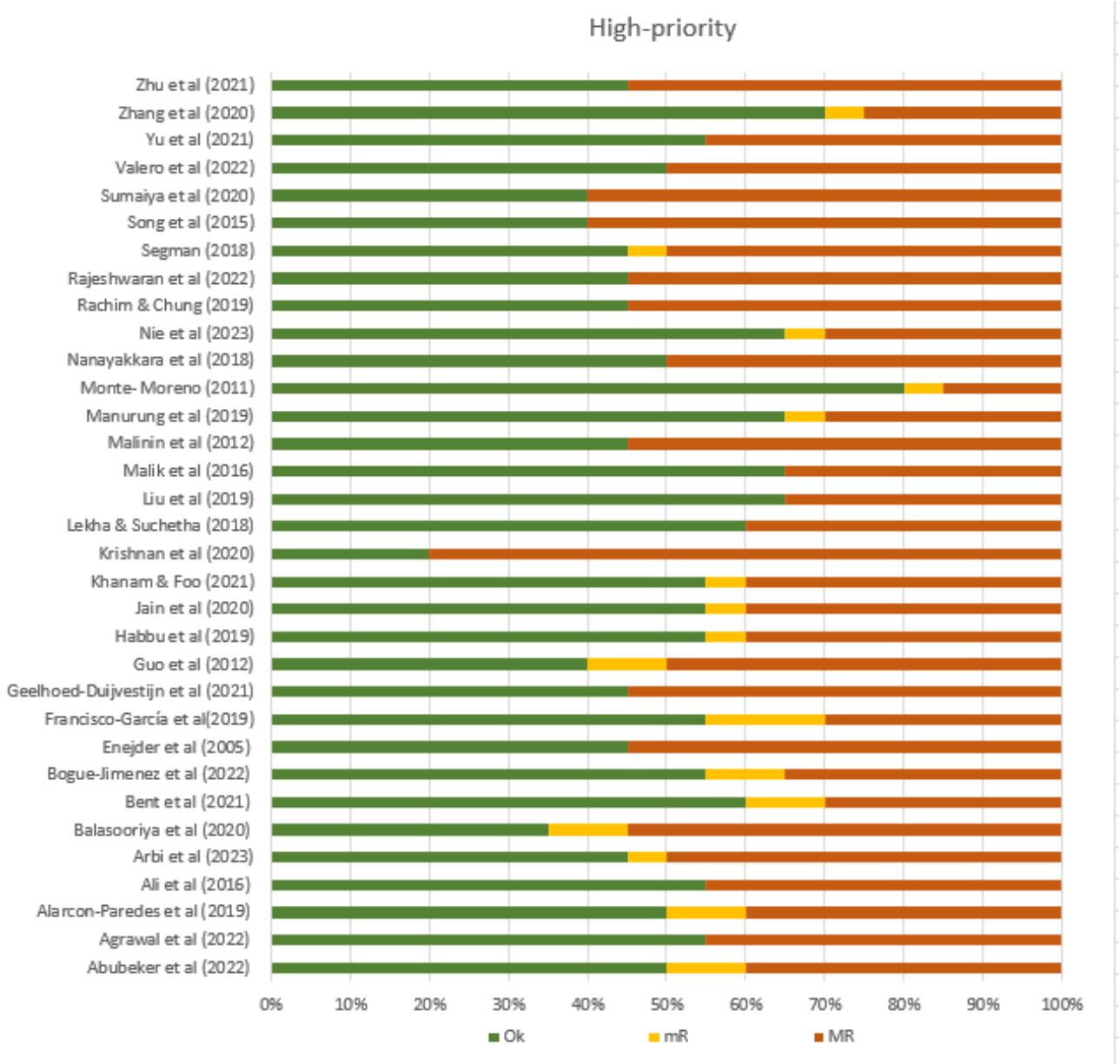

XSL•FO
RenderX

Figure 4. Proportion of OK=adequately addressed, mR=sufficient but improvable MR= inadequately addressed ratings on each low priority items.

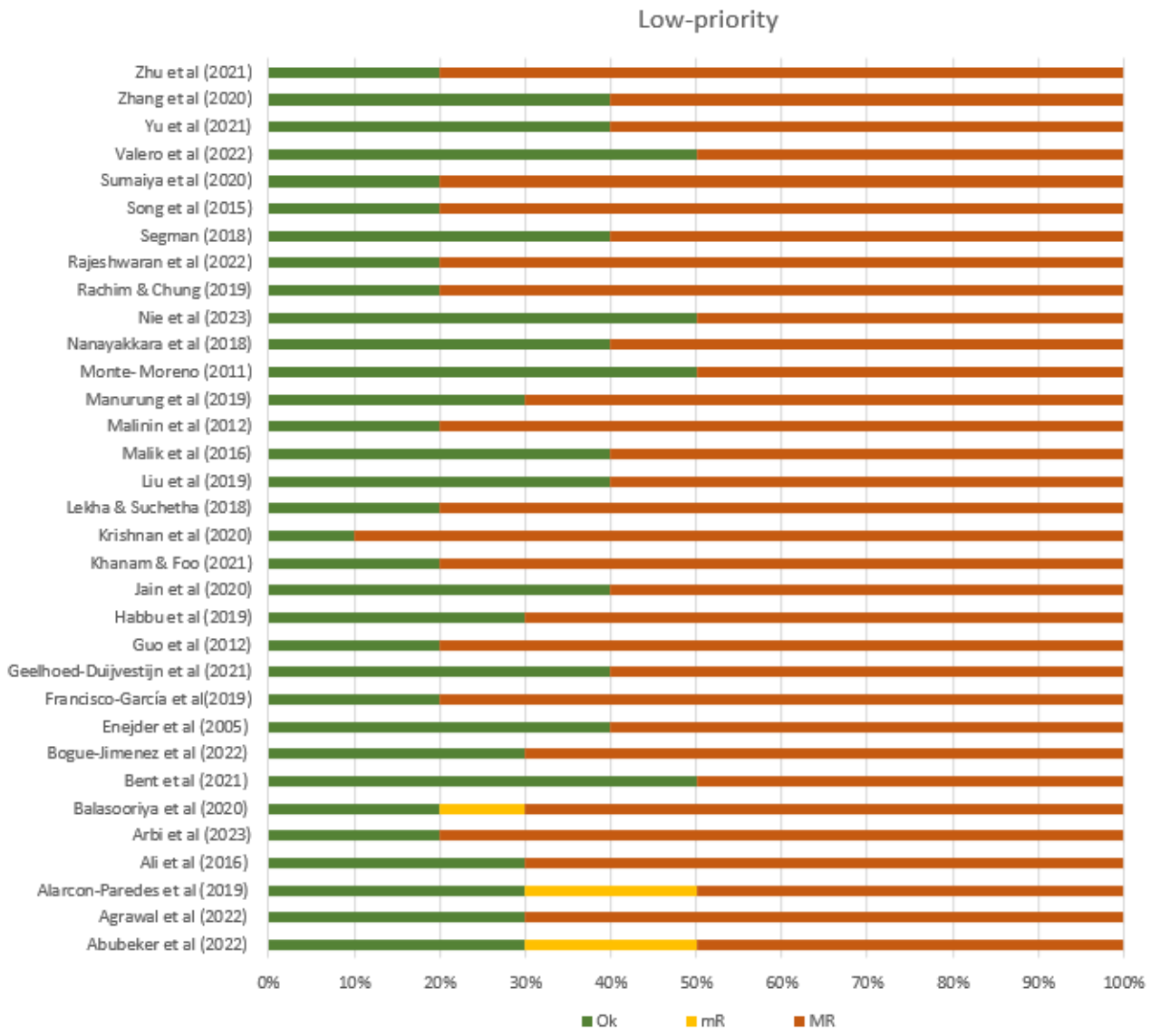

Discussion

Principal Findings

Findings from this scoping review revealed the applications of AI-assisted NIBGM systems, available technology developed, and types of AI algorithms from the 33 included studies published between 2005 and 2023. Most studies (n=20, 60%) originated from just 3 countries mainly China, India, and the United States.

The bulk of the evidence comes from Asian studies, potentially due to the alarming increase in the prevalence of DM in Asia compared to their European counterparts [60]. There was an even mix of studies from low-, middle- and high-income countries but it is unclear whether AI technologies can be made affordable and accessible to individuals in low- and middle-income countries.

More research can be done to determine the cost and accessibility of AI-assisted glucose monitoring systems and their barriers to widespread adoption. A significant number of studies were reported in conference proceedings, which reflect the emerging evidence regarding AI in NIBGM. Perhaps more research relating to diagnostic accuracy can be conducted to increase the strength of evidence for the adoption of such technology over current traditional glucose monitoring systems.

The majority of studies that develop ML algorithms to predict DM used the Pima Indian diabetes dataset which comprises 8 parameters. These criteria include the number of pregnancies, BMI, plasma glucose concentration, blood pressure, skinfold thickness, diabetes pedigree function, and an outcome variable of class 0 or 1 (where 0 denotes patient without diabetes and 1 denotes patient with diabetes) [61,62]. Other features include waveform characteristics from optical signals, such as shape and amplitude, to estimate blood glucose levels [33,36,40]. AI advances in the field of blood glucose estimation research in the context of NIBGM have the potential to improve the quality of life for patients with DM and minimize invasiveness.

Application

AI was mainly used in NIBGM to estimate real-time blood glucose levels using optical, biosensing, imaging, and tissue

XSL•FO
RenderX

impedance measurement technology instead of current widely used methods such as blood tests or finger pricks [63]. AI was also used to predict future blood glucose levels (up to 30 minutes later) [22] and detect DM [10,44], suggesting the potential of AI-assisted NIBGM for continuous BGM and diagnostic purposes.

Technology Used

Two broad classifications for NIBGM emerged namely sample- and non–sample-based methods of detection. Sample-based include studies like Malik et al [34] which use salivary electrochemical signals to train ML models. Concentrations of sodium, potassium, and calcium ions were measured and correlated with blood glucose levels [34]. Other sample-based techniques include the use of breath signals to detect acetone to estimate blood glucose levels [30,32]. A major challenge for the development of NIBGM systems which rely on bodily fluid is that the concentration glucose level is miniscule [64]. Hence, there is a need to enhance sensitivity and remove other interference in such sensors [65].

Out of the non–sample-based noninvasive techniques developed to predict blood glucose levels, PPG, akin to the technology of pulse oximetry, appears the most among the studies followed by other optical techniques such as NIRS and RS. The results were not surprising as the use of optical methods for measuring glucose levels is presently one of the best approaches in noninvasive glucose estimation research [47]. For example, Monte-Moreno [56] used a PPG-based sensor to measure changes in blood volume changes and developed an ML algorithm to estimate blood glucose levels [14]. While traditional PPG requires skin contact, typically using a finger over a smartphone camera, to detect blood volume changes, advanced remote PPG allows the detection of subtle skin color changes to estimate blood volume changes [27,35,40]. These technologies have to be validated against the conventional BGM methods in a larger clinical population to establish their usefulness and efficiency [66].

Such technology may be useful for self-monitoring since it has a low barrier of entry and only requires a smartphone. Such setup may also be useful in clinical settings for monitoring or diagnostic purposes and reduces the need for retraining since staff are familiar with similar setups in the hospitals. On the other hand, others have commented that the use of PPG is often corrupted by measurement artifacts from movements, restricting one's movement during continuous glucose monitoring [36,67].

ML Models

This review maps out the various DL and traditional ML algorithms used by the studies. Previous studies have adopted ML for risk stratification and identification of patients with DM [68]. Several ML processes, such as SVM, regression trees, k-nearest neighbor, ANN, naïve Bayes, and RF, have been used in transforming diabetes care [69].

The main uses of ML processes include feature selection and classification. ML methods require the extraction of features from signals. However, extracting fiducial points from real-life signals can be highly challenging [21]. Not only is it difficult to develop a feature extraction algorithm that can handle diverse waveform types but there is also a need to assess the quality of the computed features as the feature extraction algorithm is unable to effectively operate if the input signal is corrupted [21].

The emergence of DL has facilitated the analysis of large volumes of data without the need for explicit feature extraction. However, DL approaches experience limited interpretability, which can be problematic in a clinical setting where understanding why and how a pathology was detected is crucial for validating the diagnosis [21].

Future Research

In different studies, researchers have used various ML algorithms to construct classification models using derived feature vectors to evaluate the performance of different algorithms on the datasets used. Conversely, some researchers have opted to use a single ML method for their classification model. However, it is important to note that no single ML algorithm is universally optimal for all types of input data [70]. Therefore, it is beneficial to test multiple ML algorithms and determine which one produces the best outcomes for a given task. Comparisons among different AI models can help identify the strengths and limitations of each approach, guiding further improvements in accuracy and performance.

Given the heterogeneity of AI models and input data applied in each study, it is beyond the scope of this review to ascertain the best NIBGM system based on performance metrics alone. Furthermore, the lack of standardized reporting and analysis of results, leads to heterogeneity that hampers the comparison of findings across studies. Perhaps a diagnostic accuracy review may be more suited to address the question of which system is best suited to be adopted in various settings. As with all AI studies, efforts should be made to standardize and regulate the use of AI technologies in diabetes care. Consensus guidelines and protocols should be developed to ensure the quality and safety of AI-assisted monitoring systems [71].

Another potential area of research in the field of NIBGM is the use of digital twin (DT) techniques. DT serves as a digital representation that mirrors the state of a physical entity or system by capturing real-time data through sensors and reflecting it in digital devices [72]. DT offers a powerful solution for real-time monitoring, accurate diagnosis, and effective treatment [72]. However, the main challenges include data acquisition, data privacy, and security concerns [73]. Further advancement in Big Data is required to develop holistic and accurate DTs.

Strengths and Limitations

To the best of our knowledge, this is the first study to report the current state-of-the-art in AI-assisted NIBGM, which informs the direction of future systematic reviews and interventional research. To enhance the rigor of the study, we adhered to the PRISMA-ScR guidelines and had 2 independent reviewers in the paper selection process.

This study had several limitations. First, as this review limits in vivo methods of verification (human participants), certain relevant evidence could have been precluded, such as studies that use in vitro methods for verification of their AI models

XSL•FO
RenderX

such as skin models or varied concentrations of glucose solutions. Second, a simple keyword search strategy and only papers written in English were retrieved, possibly limiting the scope of our findings. However, we conducted a hand search of previous systematic reviews to identify relevant papers.

Conclusions

The use of AI for NIBGM is a promising area of research that has the potential to revolutionize diabetes management. The studies reviewed demonstrate that some AI techniques can accurately predict glucose levels from noninvasive sources while enhancing comfort and ease of use for patients. However, the overall range of accuracy is wide due to the heterogeneity of models and input data. As such, we propose that there is a need for further efforts to standardize and regulate the use of AI technologies in diabetes care, as well as develop consensus guidelines and protocols to ensure the quality and safety of AI-assisted monitoring systems.

Acknowledgments

No funding was received for conducting this study.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

PZC wrote the manuscript with support from EJ. PZC and EJ conducted the methodological appraisal. PZC and HSJC extracted the data. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta - Analyses extension for Scoping Reviews) checklist.
[DOCX File , 17 KB - jmir_v26i1e58892_app1.docx]

Multimedia Appendix 2
Specific search strategy for each database.
[DOCX File , 15 KB - jmir_v26i1e58892_app2.docx]

Multimedia Appendix 3
Characteristics of blood glucose monitoring systems.
[DOCX File , 25 KB - jmir_v26i1e58892_app3.docx]

Multimedia Appendix 4
Characteristics of noninvasive blood glucose monitoring systems.
[DOCX File , 26 KB - jmir_v26i1e58892_app4.docx]

Multimedia Appendix 5
Characteristics of Machine Learning.
[DOCX File , 25 KB - jmir_v26i1e58892_app5.docx]

References

1. IDF Diabetes Atlas 2021. Belgium: International Diabetes Federation; 2021.
2. Duckworth W, Abraira C, Moritz T. Glucose control and vascular complications in veterans with type 2 diabetes. J Vasc Surg 2009 Apr;49(4):1084. [doi: 10.1016/j.jvs.2009.02.026]
3. Lee A, Juraschek SP, Windham BG, Lee CJ, Sharrett AR, Coresh J, et al. Severe hypoglycemia and risk of falls in type 2 diabetes: the atherosclerosis risk in communities (ARIC) study. Diabetes Care 2020 Sep;43(9):2060-2065 [FREE Full text] [doi: 10.2337/dc20-0316] [Medline: 32611607]
4. Nakhleh A, Shehadeh N. Hypoglycemia in diabetes: an update on pathophysiology, treatment, and prevention. World J Diabetes 2021 Dec 15;12(12):2036-2049 [FREE Full text] [doi: 10.4239/wjd.v12.i12.2036] [Medline: 35047118]
5. Rodbard D. Continuous glucose monitoring: a review of recent studies demonstrating improved glycemic outcomes. Diabetes Technol Ther 2017 Jun;19(S3):S25-S37 [FREE Full text] [doi: 10.1089/dia.2017.0035] [Medline: 28585879]

XSL•FO
RenderX

6. Lane JE, Shivers JP, Zisser H. Continuous glucose monitors: current status and future developments. Curr Opin Endocrinol Diabetes Obes 2013 Apr;20(2):106-111. [doi: 10.1097/MED.0b013e32835edb9d] [Medline: 23422244]
7. McGarraugh G. The chemistry of commercial continuous glucose monitors. Diabetes Technol Ther 2009 Jun;11 Suppl 1(S1):S17-S24. [doi: 10.1089/dia.2008.0133] [Medline: 19469674]
8. Rossetti P, Bondia J, Vehí J, Fanelli CG. Estimating plasma glucose from interstitial glucose: the issue of calibration algorithms in commercial continuous glucose monitoring devices. Sensors (Basel) 2010;10(12):10936-10952 [FREE Full text] [doi: 10.3390/s101210936] [Medline: 22163505]
9. Bailey T, Bode BW, Christiansen MP, Klaff LJ, Alva S. The performance and usability of a factory-calibrated flash glucose monitoring system. Diabetes Technol Ther 2015 Nov;17(11):787-794 [FREE Full text] [doi: 10.1089/dia.2014.0378] [Medline: 26171659]
10. Agrawal H, Jain P, Joshi AM. Machine learning models for non-invasive glucose measurement: towards diabetes management in smart healthcare. Health Technol (Berl) 2022;12(5):955-970 [FREE Full text] [doi: 10.1007/s12553-022-00690-7] [Medline: 35996737]
11. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med 2020 Sep;26(9):1380-1384. [doi: 10.1038/s41591-020-1045-7] [Medline: 32908282]
12. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
13. Woldaregay AZ, Årsand E, Botsis T, Albers D, Mamykina L, Hartvigsen G. Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes. J Med Internet Res 2019 May 01;21(5):e11030 [FREE Full text] [doi: 10.2196/11030] [Medline: 31042157]
14. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018 May 30;20(5):e10775 [FREE Full text] [doi: 10.2196/10775] [Medline: 29848472]
15. Bolla AS, Priefer R. Blood glucose monitoring—an overview of current and future non-invasive devices. Diabetes Metab Syndr 2020;14(5):739-751. [doi: 10.1016/j.dsx.2020.05.016] [Medline: 32497964]
16. Alsunaidi B, Althobaiti M, Tamal M, Albaker W, Al-Naib I. A review of non-invasive optical systems for continuous blood glucose monitoring. Sensors (Basel) 2021 Oct 14;21(20):6820 [FREE Full text] [doi: 10.3390/s21206820] [Medline: 34696033]
17. Shokrekhodaei M, Cistola DP, Roberts RC, Quinones S. Non-invasive glucose monitoring using optical sensor and machine learning techniques for diabetes applications. IEEE Access 2021;9:73029-73045 [FREE Full text] [doi: 10.1109/access.2021.3079182] [Medline: 34336539]
18. Puntambekar V. MedTech internship diaries 2018. Natl Med J India 2020;33(1):59. [doi: 10.4103/0970-258x.308259]
19. Gusev M, Poposka L, Spasevski G, Kostoska M, Koteska B, Simjanoska M, et al. Noninvasive glucose measurement using machine learning and neural network methods and correlation with heart rate variability. J Sens 2020 Jan 06;2020:1-13. [doi: 10.1155/2020/9628281]
20. Swapna G, Soman K, Vinayakumar R. Deep Learning Techniques for Biomedical and Health Informatics. New York: Springer; 2019.
21. Zanelli S, Ammi M, Hallab M, El Yacoubi MA. Diabetes detection and management through photoplethysmographic and electrocardiographic signals analysis: a systematic review. Sensors (Basel) 2022 Jun 29;22(13):4890 [FREE Full text] [doi: 10.3390/s22134890] [Medline: 35808386]
22. Chaki J, Ganesh ST, Cidham S, Theertan SA. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review. J King Saud Univ Comp Inf Sci 2022;34(6):3204-3225. [doi: 10.1016/j.jksuci.2020.06.013]
23. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
24. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Meth 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
25. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform 2021 Sep;153:104510 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104510] [Medline: 34108105]
26. Zhou Y, Ge Y, Shi X, Wu K, Chen W, Ding Y, et al. Machine learning predictive models for acute pancreatitis: a systematic review. Int J Med Inform 2022 Jan;157:104641 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104641] [Medline: 34785488]
27. Sumaiya J, Hasan MR, Hossain E. Noninvasive blood glucose measurement using live video by smartphone. 2020 Presented at: 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC); December 01, 2020; Kuching, Malaysia. [doi: 10.1109/r10-htc49770.2020.9357018]
28. Balasooriya K, Nanayakkara ND. Predicting short-term changing blood glucose level of diabetes patients using noninvasive data. 2020 Presented at: 2020 IEEE Region 10 Conference (TENCON); November 16, 2020; Osaka, Japan. [doi: 10.1109/tencon50793.2020.9293823]

XSL•FO
RenderX

29. Nanayakkara ND, Munasingha SC, Ruwanpathirana GP. Non-invasive blood glucose monitoring using a hybrid technique. 2018 Presented at: Moratuwa Engineering Research Conference (MERCon)/4th International Multidisciplinary Engineering Research Conference; June 01, 2018; Katubedda, Sri Lanka. [doi: 10.1109/mercon.2018.8421885]
30. Guo D, Zhang D, Zhang L, Lu G. Non-invasive blood glucose monitoring for diabetics by means of breath signal analysis. Sens Actuators B 2012;173:106-113. [doi: 10.1016/j.snb.2012.06.025]
31. Habbu S, Dale M, Ghongade R. Estimation of blood glucose by non-invasive method using photoplethysmography. Sādhanā 2019 May 6;44(6):135. [doi: 10.1007/s12046-019-1118-9]
32. Lekha S, Suchetha M. Real-time non-invasive detection and classification of diabetes using modified convolution neural network. IEEE J Biomed Health Inform 2018;22(5):1630-1636. [doi: 10.1109/jbhi.2017.2757510]
33. Liu WJ, Huang A, Wang P, Chu C. PbFG: physique-based fuzzy granular modeling for non-invasive blood glucose monitoring. Inf Sci 2019;497:56-76. [doi: 10.1016/j.ins.2019.05.013]
34. Malik S, Khadgawat R, Anand S, Gupta S. Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. Springerplus 2016;5(1):701 [FREE Full text] [doi: 10.1186/s40064-016-2339-6] [Medline: 27350930]
35. Nie Z, Rong M, Li K. Blood glucose prediction based on imagingphotoplethysmography in combination with machine learning. Biomed Signal Process Control 2023;79:104179. [doi: 10.1016/j.bspc.2022.104179]
36. Rachim VP, Chung W. Wearable-band type visible-near infrared optical biosensor for non-invasive blood glucose monitoring. Sens Actuators B 2019;286:173-180. [doi: 10.1016/j.snb.2019.01.121]
37. Segman YJ. Device and method for noninvasive glucose assessment. J Diabetes Sci Technol 2018 Nov;12(6):1159-1168 [FREE Full text] [doi: 10.1177/1932296818763457] [Medline: 29575926]
38. Song K, Ha U, Park S, Bae J, Yoo H. An impedance and multi-wavelength near-infrared spectroscopy IC for non-invasive blood glucose estimation. IEEE J Solid-State Circuits 2015;50(4):1025-1037. [doi: 10.1109/jssc.2014.2384037]
39. Yu Y, Huang J, Zhu J, Liang S. An accurate noninvasive blood glucose measurement system using portable near-infrared spectrometer and transfer learning framework. IEEE Sens J 2021;21(3):3506-3519. [doi: 10.1109/jsen.2020.3025826]
40. Zhang G, Mei Z, Zhang Y, Ma X, Lo B, Chen D, et al. A noninvasive blood glucose monitoring system based on smartphone PPG signal processing and machine learning. IEEE Trans Ind Inf 2020;16(11):7209-7218. [doi: 10.1109/tii.2020.2975222]
41. Zhu J, Zhou Y, Huang J, Zhou A, Chen Z. Noninvasive blood glucose concentration measurement based on conservation of energy metabolism and machine learning. Sensors (Basel) 2021 Oct 21;21(21):6989 [FREE Full text] [doi: 10.3390/s21216989] [Medline: 34770294]
42. Manurung BE, Munggaran HR, Ramadhan GF, Koesoema AP. Non-invasive blood glucose monitoring using near-infrared spectroscopy based on internet of things using machine learning. 2019 Presented at: 7th IEEE Region10 Humanitarian Technology Conference (IEEE R10 HTC); November 12, 2019; Depok, West Java, Indonesia. [doi: 10.1109/r10-htc47129.2019.9042479]
43. Rajeshwaran K, Thirunavukkarasu T, Pooja S, Rajeshkumar S. Machine learning based non-invasive glucose observation for diabetes. 2022 Presented at: 2022 Smart Technologies, Communication and Robotics (STCR); December 11, 2022; Sathyamangalam, India. [doi: 10.1109/stcr55312.2022.10009539]
44. Abubeker KM, Baskar S. A machine learning strategy for internet-of-things-enabled diabetic prediction to mitigate pneumonia risk. 2022 Presented at: 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); October 12, 2022; Noida, India. [doi: 10.1109/icrito56286.2022.9964842]
45. Krishnan SH, Vinupritha P, Kathirvelu D. Non-invasive glucose monitoring using machine learning. 2020 Presented at: 2020 International Conference on Communication and Signal Processing (ICCSP); July 28, 2020; Chennai, India. [doi: 10.1109/iccsp48568.2020.9182434]
46. Ali MS, Shoumy NJ, Khatun S, Kamarudin LM. Non-invasive blood glucose measurement performance analysis through UWB imaging. 2016 Presented at: 2016 3rd International Conference on Electronic Design (ICED); August 11, 2016; Phuket, Thailand. [doi: 10.1109/iced.2016.7804698]
47. Alarcón-Paredes A, Francisco-García V, Guzmán-Guzmán I, Cantillo-Negrete J, Cuevas-Valencia R, Alonso-Silverio G. An IoT-based non-invasive glucose level monitoring system using raspberry pi. Appl Sci 2019 Jul 28;9(15):3046. [doi: 10.3390/app9153046]
48. Arbi KF, Soulimane S, Saffih F, Bechar MA, Azzoug O. Blood glucose estimation based on ECG signal. Phys Eng Sci Med 2023 Mar;46(1):255-264. [doi: 10.1007/s13246-022-01214-3] [Medline: 36595189]
49. Bent B, Cho PJ, Wittmann A, Thacker C, Muppidi S, Snyder M, et al. Non-invasive wearables for remote monitoring of HbA1c and glucose variability: proof of concept. BMJ Open Diabetes Res Care 2021 Jun;9(1):e002027 [FREE Full text] [doi: 10.1136/bmjdrc-2020-002027] [Medline: 36170350]
50. Bogue-Jimenez B, Huang X, Powell D, Doblas A. Selection of noninvasive features in wrist-based wearable sensors to predict blood glucose concentrations using machine learning algorithms. Sensors (Basel) 2022 May 06;22(9):3534 [FREE Full text] [doi: 10.3390/s22093534] [Medline: 35591223]
51. Enejder AMK, Scecina TG, Oh J, Hunter M, Shih W, Sasic S, et al. Raman spectroscopy for noninvasive glucose measurements. J Biomed Opt 2005;10(3):031114 [FREE Full text] [doi: 10.1117/1.1920212] [Medline: 16229639]

XSL•FO
RenderX

52. Geelhoed-Duijvestijn P, Vegelyte D, Kownacka A, Anton N, Joosse M, Wilson C. Performance of the prototype NovioSense noninvasive biosensor for tear glucose in type 1 diabetes. J Diabetes Sci Technol 2021 Nov;15(6):1320-1325 [FREE Full text] [doi: 10.1177/1932296820964844] [Medline: 33095035]
53. Jain P, Joshi AM, Mohanty SP. iGLU: an intelligent device for accurate noninvasive blood glucose-level monitoring in smart healthcare. IEEE Consumer Electron Mag 2020;9(1):35-42. [doi: 10.1109/mce.2019.2940855]
54. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express 2021;7(4):432-439. [doi: 10.1016/j.icte.2021.02.004]
55. Malinin L. Development of a non-invasive blood glucose monitor based on impedance measurements. IJBET 2012;8(1):60. [doi: 10.1504/ijbet.2012.045358]
56. Monte-Moreno E. Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. Artif Intell Med 2011 Oct;53(2):127-138. [doi: 10.1016/j.artmed.2011.05.001] [Medline: 21696930]
57. Valero M, Pola P, Falaiye O, Ingram KH, Zhao L, Shahriar H, et al. Development of a noninvasive blood glucose monitoring system prototype: pilot study. JMIR Form Res 2022 Aug 26;6(8):e38664 [FREE Full text] [doi: 10.2196/38664] [Medline: 36018623]
58. Francisco-García V, Guzmán-Guzmán IP, Salgado-Rivera R, Alarcón-Paredes A. Non-invasive glucose level estimation: a comparison of regression models using the MFCC as feature extractor. In: Pattern Recognition. Cham: Springer International Publishing; 2019:206-215.
59. Ramasahayam S, Haindavi KS, Chowdhury SR. Noninvasive estimation of blood glucose concentration using near infrared optodes. In: Sensing Technology: Current Status and Future Trends IV. Hyderabad, India: Springer International Publishing; 2015:67-82.
60. Rhee EJ. Diabetes in Asians. Endocrinol Metab 2015;30(3):263-269. [doi: 10.3803/enm.2015.30.3.263]
61. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comp Sci 2018;132:1578-1585. [doi: 10.1016/j.procs.2018.05.122]
62. Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, et al. A model for early prediction of diabetes. Inf Med Unlocked 2019;16:100204. [doi: 10.1016/j.imu.2019.100204]
63. Chew HSJ, Lim SL, Kim G, Kayambu G, So BYJ, Shabbir A, et al. Essential elements of weight loss apps for a multi-ethnic population with high BMI: a qualitative study with practical recommendations. Transl Behav Med 2023 Apr 03;13(3):140-148. [doi: 10.1093/tbm/ibac090] [Medline: 36689306]
64. Alhaddad AY, Aly H, Gad H, Al-Ali A, Sadasivuni KK, Cabibihan J, et al. Sense and learn: recent advances in wearable sensing and machine learning for blood glucose monitoring and trend-detection. Front Bioeng Biotechnol 2022;10:876672 [FREE Full text] [doi: 10.3389/fbioe.2022.876672] [Medline: 35646863]
65. Yao H, Shum AJ, Cowan M, Lähdesmäki I, Parviz BA. A contact lens with embedded sensor for monitoring tear glucose level. Biosens Bioelectron 2011 Mar 15;26(7):3290-3296 [FREE Full text] [doi: 10.1016/j.bios.2010.12.042] [Medline: 21257302]
66. Jansson MM, Kögler M, Hörkkö S, Ala-Kokko T, Rieppo L. Vibrational spectroscopy and its future applications in microbiology. Appl Spectrosc Rev 2021;58(2):132-158. [doi: 10.1080/05704928.2021.1942894]
67. Tamura T, Maeda Y, Sekine M, Yoshida M. Wearable photoplethysmographic sensors—past and present. Electronics 2014;3(2):282-302. [doi: 10.3390/electronics3020282]
68. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, Suri HS, Abedin MM, El-Baz A, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst 2018 Apr 10;42(5):92 [FREE Full text] [doi: 10.1007/s10916-018-0940-7] [Medline: 29637403]
69. Dankwa-Mullan I, Rivo M, Sepulveda M, Park Y, Snowdon J, Rhee K. Transforming diabetes care through artificial intelligence: the future is here. Popul Health Manag 2019 Jun;22(3):229-242 [FREE Full text] [doi: 10.1089/pop.2018.0129] [Medline: 30256722]
70. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014 Presented at: International Conference on Learning Representations (ICLR 2015); May 07-09, 2015; San Diego, CA p. 1-14 URL: https://www.semanticscholar.org/paper/Very-Deep-Convolutional-Networks-for-Large-Scale-Simonyan-Zisserman/eb42cf88027de515750f230b23b1a057dc782108
71. Varsi C, Nes LS, Kristjansdottir OB, Kelders SM, Stenberg U, Zangi HA, et al. Implementation strategies to enhance the implementation of eHealth programs for patients with chronic illnesses: realist systematic review. J Med Internet Res 2019 Sep 27;21(9):e14255 [FREE Full text] [doi: 10.2196/14255] [Medline: 31573934]
72. Sun T, He X, Song X, Shu L, Li Z. The digital twin in medicine: a key to the future of healthcare? Front Med (Lausanne) 2022;9:907066 [FREE Full text] [doi: 10.3389/fmed.2022.907066] [Medline: 35911407]
73. Bruynseels K. When nature goes digital: routes for responsible innovation. J Responsible Innovation 2020;7(3):342-360. [doi: 10.1080/23299460.2020.1771144]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
BGM: blood glucose monitoring
CEG: Clarke error grid
ChAMAI: checklist for assessment of medical AI
DL: deep learning
DM: diabetes mellitus
DT: digital twin
ECG: electrocardiography
HbA1c: hemoglobin A1c
MAE: mean absolute error
ML: machine learning
NIBGM: noninvasive blood glucose monitoring
NIRS: near-infrared spectroscopy
NN: neural network
PPG: photoplethysmography
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta - Analyses extension for Scoping Reviews
RF: random forest
RS: Raman spectroscopy
SVM: support vector machines
UWB: ultrawideband

Edited by A Coristine; submitted 01.04.24; peer-reviewed by S Saeedi, AN Ali; comments to author 21.06.24; revised version received 24.06.24; accepted 08.10.24; published 19.11.24.

Please cite as:
Chan PZ, Jin E, Jansson M, Chew HSJ
AI-Based Noninvasive Blood Glucose Monitoring: Scoping Review
J Med Internet Res 2024;26:e58892
URL: https://www.jmir.org/2024/1/e58892
doi:10.2196/58892
PMID:

©Pin Zhong Chan, Eric Jin, Miia Jansson, Han Shi Jocelyn Chew. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Digital Health Interventions in Promoting Physical Activity Among College Students: Systematic Review and Meta-Analysis

Siyuan Bi[1], MA; Junfeng Yuan[1], MA; Yanling Wang[1], MA; Wenxin Zhang[1], MA; Luqin Zhang[1], MA; Yongjuan Zhang[1], MA; Rui Zhu[1], MA; Lin Luo[1], PhD

School of Physical Education, Guizhou Normal University, Guiyang, China

Corresponding Author:
Lin Luo, PhD
School of Physical Education
Guizhou Normal University
Huaxi District, University Town, Siya Road
Guiyang, 550025
China
Phone: 86 86751983
Email: 460022831@gznu.edu.cn

Abstract

Background: Recent studies offer conflicting conclusions about the effectiveness of digital health interventions in changing physical activity behaviors. In addition, research focusing on digital health interventions for college students remains relatively scarce.

Objective: This study aims to examine the impact of digital health interventions on physical activity behaviors among college students, using objective measures as outcome indicators.

Methods: In accordance with the 2020 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a comprehensive literature search was conducted across several databases, including MEDLINE (PubMed), Web of Science, Cochrane Library, and EBSCO (CINAHL Plus with full text), to identify relevant intervention studies published up to June 6, 2023. The inclusion criteria specified studies that examined the quantitative relationships between digital health interventions and physical activity among adults aged 18 years to 29 years, focusing on light physical activity (LPA), moderate to vigorous physical activity (MVPA), sedentary time (ST), or steps. Non–randomized controlled trials were excluded. The quality of the studies was assessed using the Cochrane Risk of Bias tool. Results were synthesized both narratively and quantitatively, where applicable. When sufficient homogeneity was found among studies, a random-effects model was used for meta-analysis to account for variability.

Results: In total, 8 studies, encompassing 569 participants, were included in the analysis. The primary outcomes measured were LPA, MVPA, ST, and steps. Among these studies, 3 reported on LPA, 5 on MVPA, 5 on ST, and 3 on steps. The meta-analysis revealed a significant increase in steps for the intervention group compared with the control group (standardized mean difference [SMD] 0.64, 95% CI 0.37-0.92; $P<.001$). However, no significant differences were observed between the intervention and control groups regarding LPA (SMD –0.08, 95% CI –0.32 to 0.16; $P=.51$), MVPA (SMD 0.02, 95% CI –0.19 to 0.22; $P=.88$), and ST (SMD 0.03, 95% CI –0.18 to 0.24; $P=.78$).

Conclusions: Digital health interventions are effective in increasing steps among college students; however, their effects on LPA, MVPA, and sedentary behavior are limited.

Trial Registration: PROSPERO CRD42024533180; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=533180

(J Med Internet Res 2024;26:e51714) doi:10.2196/51714

KEYWORDS

digital health intervention; college students; physical activity behavior; steps; light intensity physical activity; moderate to vigorous physical activity; sedentary behavior; knowledge synthesis; systematic review; meta-analysis; physical activity; eHealth; digital health; mobile phone

XSL•FO
RenderX

Introduction

Background

The college years represent a critical phase [1] that significantly influences the levels of physical activity and the formation of healthy habits [2-4]. However, research indicates that physical activity levels among college students are generally low and rarely meet the recommended health guidelines [5]. Physical activity and sedentary behavior are interrelated behaviors that affect health and well-being and are risk factors for noncommunicable diseases. The increasing number of sedentary adults negatively impacts the prevalence of noncommunicable diseases and the overall health status of the global population [6-9]. Physical activity is a crucial protective factor that can prevent and manage major noncommunicable diseases such as cardiovascular diseases, stroke, diabetes, breast cancer, and colon cancer. Furthermore, physical activity helps prevent other significant risk factors for noncommunicable diseases, such as hypertension, overweight, and obesity, while also improving mental health, delaying the onset of dementia, and enhancing the quality of life and well-being. Based on this, the World Health Organization has developed global recommendations for promoting health through physical activity, suggesting that adults should engage in at least 150 minutes of moderate-intensity physical activity or 75 minutes of vigorous-intensity physical activity per week. However, currently, a quarter of adults worldwide fail to meet the global physical activity recommendations as set by the World Health Organization [10]. Thus, efforts to increase physical activity and reduce sedentary behavior have been identified as international public health priorities [11-13].

The lack of physical activity negatively affects the health of college students, potentially leading to psychological health issues, academic performance decline, and limited social life [14-16]. To improve college students' physical activity behavior, digital health interventions have been recognized as a promising strategy [17,18]. Digital health is a health strategy that uses digital, mobile, and wireless technologies to support the achievement of health goals, including mobile health (mHealth) and eHealth. These technologies encompass mobile phones, tablets, web-based interventions, SMS, social media, fitness mobile apps, and wearable devices [19]. Digital health interventions may involve the use of mobile applications, social media platforms, online fitness programs, and other digital tools to provide personalized health advice, goal setting, and monitoring [20-23]. A characteristic feature of digital health interventions is their ease of deployment and the ability to target specific population groups [19,24]. The Global Action Plan on Physical Activity by the World Health Organization emphasizes the use of digital health interventions to promote physical activity behavior [25]. The advantages of digital health interventions lie in their reliability, real-time nature, and personalized customization, which can provide effective support for physical activity among college students [26-33]. Considering that college students are characterized as digital natives [34-36], scholars have conducted numerous digital health intervention studies targeting college students' health behaviors. However, despite the extensive research on digital health interventions, the overall effectiveness of these interventions on college students' physical activity behavior remains unclear.

Some studies have found that personalized intelligent SMS interventions have a positive impact on physical activity health outcomes [37,38]. Other research supports the effectiveness of smartphone applications in increasing physical activity [39-44]. However, there are also research findings that suggest digital health interventions may not effectively promote changes in physical activity behavior [45-47]. In addition, some studies have found that internet-based interventions have positive effects on physical activity, but with small effect sizes and heterogeneity [48,49]. Furthermore, other studies suggest that digital health interventions may improve specific health behaviors and outcomes, but with limited effectiveness [50,51]. There are also studies that indicate a small positive correlation between digital health engagement and physical activity, but the heterogeneity in the design, settings, intervention components, and outcomes of the studies may affect the validity of the conclusions [52-54]. In summary, while digital health interventions have the potential to improve college students' physical activity behavior, further research is needed to validate their overall effectiveness.

This study focused on young adults aged 18-29 years, specifically college students in the early adulthood stage [1]. This demographic was chosen due to its unique developmental phase, characterized by a transition from adolescence to adulthood. Individuals in this group face numerous challenges, including academic pressure, changing living environments, and shifting social roles, all of which can contribute to decreased physical activity levels and an increased risk of obesity and related chronic diseases [2-4].

Digital health interventions, as a convenient, low-cost, and easily disseminated health promotion tool, have shown potential in promoting physical activity in recent years [19,24]. However, existing research presents mixed findings regarding the effectiveness of such interventions on college students' physical activity. While some studies report positive impacts on physical activity indicators [37,38], others suggest limited or even insignificant effects [45-47].

To address this ambiguity, a meta-analysis was conducted to synthesize and analyze existing research findings. This rigorous approach aimed to provide more comprehensive and reliable evidence for evaluating the impact of digital health interventions on college students' physical activity behavior. The findings contributed to filling the knowledge gaps in the current literature and offered empirical evidence to support the application of digital health interventions in promoting physical activity among this population.

Objectives

The aim of this study is to assess the overall impact of digital health interventions on college students' physical activity behavior through a meta-analysis. By synthesizing and summarizing existing research findings, the study seeks to obtain an accurate evaluation of the overall effectiveness of digital health interventions in the college student population. In this context, the following research questions are proposed: Do

XSL•FO
RenderX

digital health interventions have a positive impact on college students' physical activity behavior? Can digital health interventions significantly improve the level of physical activity among college students? Are there differences in the effectiveness of digital health interventions across different modes of physical activity?

By addressing these research questions, the study aims to develop a comprehensive understanding of the impact of digital health interventions on college students' physical activity behavior. The findings are intended to provide scientific evidence to support the development of relevant policies and intervention measures.

Methods

Overview

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [55] and was registered with PROSPERO (International Prospective Register of Systematic Reviews) under the identifier CRD42024533180. A rigorous and comprehensive literature search was executed using computer-assisted systematic search strategies. Initially, as specified in our registered protocol (CRD42024533180), the intention was to limit the search to the MEDLINE (PubMed) and Web of Science databases. However, to ensure a more comprehensive inclusion of relevant literature, the search strategy was subsequently expanded to incorporate additional databases, namely the Cochrane Library and EBSCO (CINAHL Plus with full text). Apart from this modification in the search strategy, the study design remained consistent with the original PROSPERO protocol.

Search Strategy

This study aimed to assess the impact of digital health interventions on changes in physical activity behavior among college students. To ensure a comprehensive search, we searched MEDLINE (PubMed) and Web of Science, Cochrane library, and EBSCO (CINAHL Plus with full text) databases for relevant articles published up to June 6, 2023. The detailed search strategy, including specific keywords and search terms, is available in Multimedia Appendix 1.

Selection Criteria

The selection criteria are shown in Textbox 1.

Textbox 1. List of inclusion and exclusion criteria.

Inclusion criteria

- Population: the study population consists of college students aged 18-29 years. Participants with intellectual disabilities, significant cognitive impairments, or severe mobility impairments were excluded from this study.
- Interventions: the included interventions in the study are digital health–based interventions that have an impact on changes in physical activity behavior. These interventions include mHealth and eHealth, such as mobile phones, tablets (such as iPad), web-based interventions (such as web-based chats and teleconferences), SMS, social media, lifestyle or fitness smartphone apps, and wearable devices.
- Control group: the included control groups in the study consist of groups that do not receive any intervention, receive traditional interventions, or receive minimal interventions (such as providing only physical activity goals). If there are multiple intervention groups in the study, the control group's data from and the data from the intervention group with the most digital health intervention features are extracted for analysis.
- Outcomes: this meta-analysis includes studies that measure physical activity or sedentary behavior objectively using devices (such as mobile phone data, accelerometers, pedometers). The mean and SD of the following indicators are extracted: light physical activity and moderate to vigorous physical activity, sedentary time, and step count.
- Study design: this meta-analysis only includes randomized controlled trials, where the experimental group uses technology-based interventions while the control group receives either traditional interventions (primarily oral or written health education materials) or no intervention.

Exclusion criteria

- Review article.
- Letter to the editor.
- Editorial.
- Articles in non-English languages (while acknowledging that excluding non-English articles may introduce language bias and limit the comprehensiveness of the findings, this review focuses solely on research published in English due to resource limitations, specifically regarding translation, time, and funding).
- Other study designs (retrospective studies, quasi-experimental designs, non–randomized controlled trials).
- Articles based on studies focused on research subjects other than college students (aged 18-29 years).
- Articles that did not collect or report data on sex.
- Articles without objective measurement of physical activity (eg, relying solely on self-report).
- Articles exclusively reporting on the feasibility or acceptability of technology-based interventions, or lacking findings pertaining to physical activity behavior outcomes, were excluded.
- Articles with incomplete outcome reporting, hindering reliable synthesis, were excluded from this analysis.

XSL•FO
RenderX

Study Selection

To guarantee the accuracy and reliability of included studies, a rigorous screening process was implemented. Following the guidelines for meta-analysis, 2 independent reviewers conducted a blinded, duplicate screening of all retrieved articles to determine their eligibility for inclusion [56]. The screening process involved 2 stages. Initially, titles and abstracts were screened to exclude studies that clearly did not meet the inclusion criteria. Subsequently, full texts of the remaining articles were reviewed to confirm their eligibility. Throughout the screening process, discrepancies between reviewers were resolved through discussion and consensus.

Data Extraction

The researchers used standardized forms specifically developed for this review to independently extract data from each included study, ensuring comprehensive acquisition of trial data. The extracted data encompassed the study country, study participants (population and sample size), study design (description of digital intervention measures, intervention and control group characteristics, and intervention duration), outcome measurements (measurement tools and timing), and key study outcomes (mean and SD of light physical activity [LPA], moderate to vigorous physical activity [MVPA], sedentary time [ST], and steps, or baseline and follow-up step count statistics). Average changes were computed when necessary, and in cases where SDs were not reported, they were calculated based on SE or 95% CI [57,58]. If the research papers did not adequately report the results, the researchers contacted the authors to obtain additional data. In instances where discrepancies in data extraction arose, the author team engaged in discussions and reexamined the original studies until a consensus was reached.

Risk of Bias (Quality) Assessment

The researchers used the Cochrane Collaboration bias tool, which is integrated within the Review Manager software (version 5.4, The Nordic Cochrane Centre), and supplemented it with the CONSORT (Consolidated Standards of Reporting Trials) checklist to assess the quality of the included studies [59]. This tool comprises domains for selection bias, performance bias, detection bias, attrition bias, and reporting bias. Each study was independently assessed by 2 reviewers. In addition, publication bias was assessed using funnel plot analysis.

Data Analysis

The statistical analysis for this study was performed using Review Manager and Stata (version 16; StataCorp LLC). Considering the challenges posed by various definitions of changes in physical activity behavior, this study decided to include only literature that solely used LPA, MVPA, ST, and steps as pre- and postintervention measurement indicators. Studies that did not provide these measurement data were excluded to ensure interpretability and meaningfulness of the meta-analysis. In addition, studies that did not provide sufficient information for calculating the mean and SD, or used statistical analyses (such as correlation coefficients, multivariate, and adjusted regression coefficients) inconsistent with the relevant meta-analysis, were also excluded. Furthermore, in cases where multiple intervention groups within a study exhibited similar characteristics of digital health interventions, the intervention group with the most digital health intervention features was selected for the meta-analysis.

The primary outcome measures for this meta-analysis were the mean and SD of LPA, MVPA, ST, and steps. For each intervention and control group in each study, the baseline and follow-up mean, SD, and participant count of physical activity were entered into the Review Manager software. Using these data, standardized mean differences (SMDs) between intervention and control group changes for each study were calculated as the measure of effect size. The SMD was chosen because although the included studies used the same units of measurement, the scales or measurement tools used across studies might have varied, making direct comparisons of weighted mean differences difficult. SMD, through standardization, can eliminate the influence of different scales, reflecting the differences in intervention effects more accurately and enhancing the interpretability of the results. If data were available from 3 or more studies, a meta-analysis was conducted, and an overall effect size was computed.

Due to the diversity in study populations and intervention designs, a random effects model approach was used to account for heterogeneity among studies. The I^2 statistic was used to assess the presence of heterogeneity, which describes the percentage of total variation in effect estimates due to heterogeneity [60,61]. The I^2 statistic was chosen as the preferred measure of variance since it is stable with small sample sizes. Heterogeneity was considered significant if the P value from the chi-square test was less than .10 and I^2 was greater than 50% [60,62,63].

The effect size of SMDs was interpreted using Cohen suggestion [64], where 0.2 represents a small effect size, 0.5 represents a medium effect, and 0.8 represents a large effect, with values below 0.2 considered trivial. In addition, sensitivity analyses for LPA, MVPA, ST, and steps were performed using Stata software (version 16.0) to further assess the stability of the study results.

Results

Study Selection

Initially, a total of 14,915 articles were identified for evaluation across 4 databases, such as 1743 from PubMed, 6403 from Web of Science, 4583 from the Cochrane Library, and 1933 from EBSCO (CINAHL Plus with full text). In addition, 253 articles were identified through manual searching of reference lists from key articles. After excluding 4028 duplicate records from multiple databases and search channels, we proceeded to exclude 10,722 articles that were irrelevant to the scope of this study based on their titles and abstracts. From the remaining 165 articles, an additional 157 were deemed ineligible for inclusion in this meta-analysis, as they did not meet the predetermined inclusion criteria. Ultimately, a total of 8 relevant research articles were identified and included in this meta-analysis. The detailed flowchart of the literature search process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the searching and screening processes.

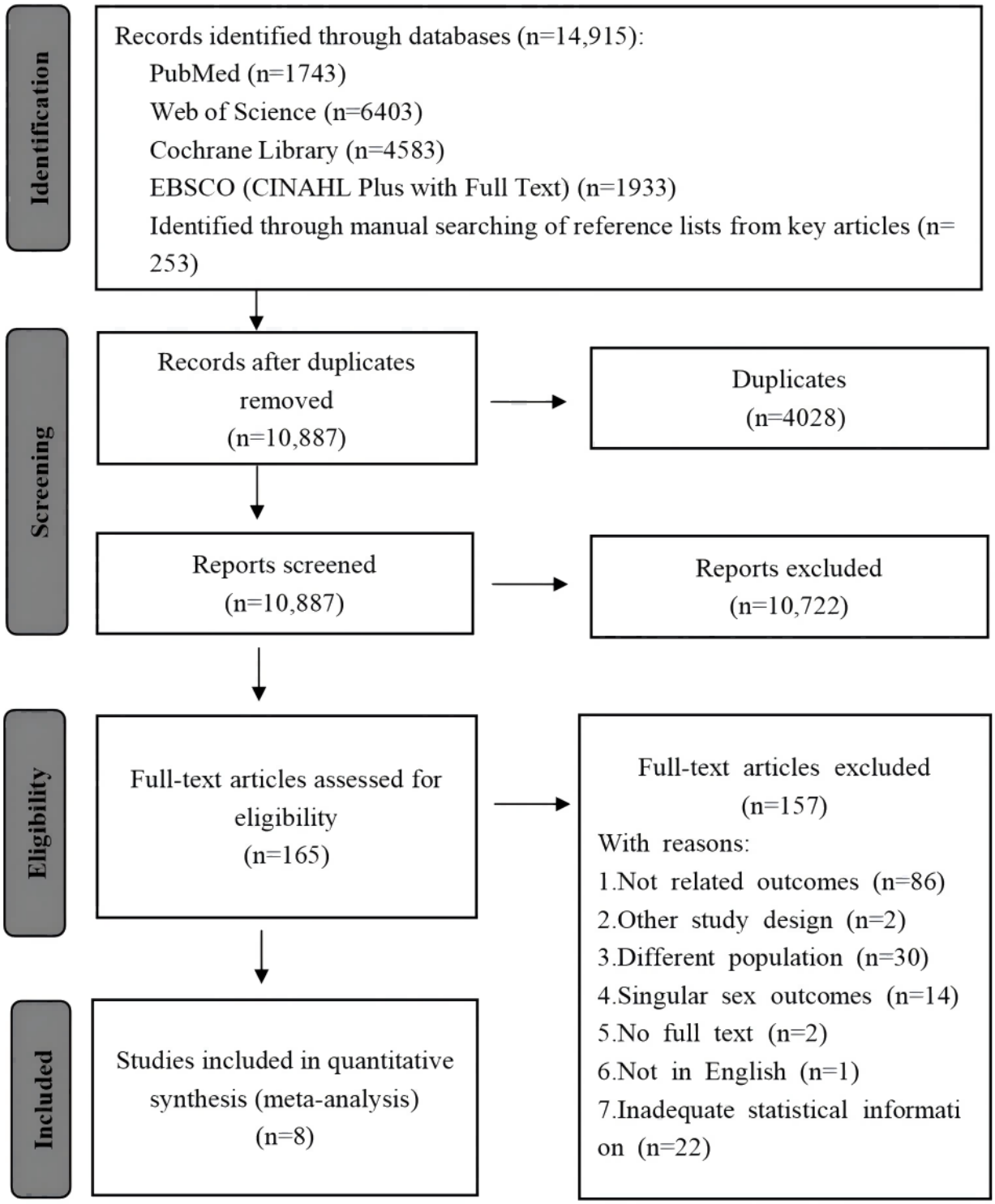

Study Characteristics

Overview

A total of 8 studies were included in this meta-analysis, all of which were randomized controlled trials published in English between 2014 and 2023. The main characteristics of these 8 studies are provided in Multimedia Appendix 2 (PRISMA checklist provided in Multimedia Appendix 3). It was reported that four studies [45,46,50,65] were not registered on ClinicalTrials.gov, while 4 studies [47,51,66,67] were registered on the website. Among these 8 studies, 4 were conducted in Australia [45], Spain [66], China [46], and Germany [65], respectively, while the remaining 4 studies were conducted in the United States [47,50,51,67].

Participants

The sample sizes of these 8 studies ranged from 34 to 187 participants, with a total of 569 participants. All included studies

XSL•FO
RenderX

targeted individuals between the ages of 18 and 30 years, specifically university students. Among these, except for 1 study [50] which did not report or calculate the SD of participant age, the remaining 7 studies [45-47,51,65-67] reported or calculated both the mean and SD of participant age.

The 8 studies exhibited varying gender distributions among their participants. In a study by Hebden et al [45], women comprised 80% (n=41) of the 51 participants. Similarly, a study by Kim et al [47] included 187 participants with 62% (n=116) identifying as female. The study by Miragall et al [66] comprised 52 participants, of whom 86% (n=45) were women. A study by Pope at al [51], with 38 participants, had 74% (n=28) female representation. The study by Al-Nawaiseh et al [67] included 114 participants, with women constituting 81% (n=92) of the sample. The study by Lau et al [46] had a more balanced gender distribution, with 61% (n=34) of its 56 participants being female. The study by Pope and Gao [50], involving 44 participants, had 73% (n=32) female participants. Finally, the study by Kellner et al [65], with the smallest sample size of 34, had the highest female representation at 85% (n=29).

Interventions

All primary interventions in the studies aimed to increase physical activity, although one study [45] aimed to evaluate the impact on body weight, and another study [50] aimed to assess the impact on health outcomes. The methods used in the intervention group included social media [46], smart text messaging [65], mobile digital applications [45,67], smart wearable technology devices (such as smartwatches and pedometers) [47,66], or a combination of social media and smartphone apps [50], or a combination of social media and wearable technology devices [51]. The control group either did not receive any intervention [46,65] or only received traditional/standard interventions [45,47,50,51,66,67]. The duration of the interventions varied from 3 weeks [66] to 4 weeks [46], 5 weeks [65], 10 weeks [50], 12 weeks [45,51,67], and 15 weeks [47].

Outcomes

All included studies reported multiple objective measures of physical activity. In total, 3 studies (n=276) reported LPA, 5 studies (n=369) reported MVPA, 5 studies (n=354) reported ST, and 3 studies (n=215) reported steps.

For measuring LPA, MVPA, and ST, the 8 studies used the following tools: ActiGraph GT3X+ accelerometer [68], ActiGraph accelerometer (GT1M model) [69], SenseWear Pro Armband mobile device [70], ActiGraph Actitrainer (ActiGraph LLC) [71], ActiGraph Link accelerometer [72], and ActivPal accelerometer (Pal Technologies) [73]. The tools used to measure steps included the Pacer pedometer, Accupedo pedometer, and Fitbit One pedometer.

Risk of Bias in Included Studies

The methodological rigor of the included studies was evaluated using the Cochrane Risk of Bias tool, which assesses 6 domains, namely selection bias, performance bias, detection bias, attrition bias, reporting bias, and other biases. Each domain was assessed as having a low risk, high risk, or unclear risk of bias. The overall quality of each study was then classified as high, moderate, or low based on the assessment results (Figure 2 [45-47,50,51,65-67]).

XSL•FO
RenderX

Figure 2. Risk of bias summary. These are authors' judgments of each methodological quality item for each included study. Plus signs (+) indicate high methodological quality (low risk of bias); minus signs (-) indicate low methodological quality (high risk of bias); question marks (?) indicate unclear methodological quality (reported information about what happened in the study was insufficient).

The included studies generally demonstrated moderate methodological quality, meeting average quality assessment standards. Notably, all but one study [46] explicitly described a robust random sequence generation process for participant allocation, minimizing selection bias. Furthermore, all studies reported baseline characteristics and used valid, objective measures for assessing physical activity outcomes, reducing the risk of detection bias.

However, the nature of the interventions presented challenges in achieving blinding of participants and personnel, potentially introducing performance bias. Specifically, while intention-to-treat analysis was used in all but 4 studies

XSL•FO
RenderX

[46,65-67], mitigating attrition bias, the lack of blinding may have influenced participant behavior and data interpretation.

Regarding reporting bias, all studies, with the exception of Lau et al [46], provided sufficient statistical data for analysis. However, only 5 studies [46,47,50,51,67] reported between-group differences in changes over time, potentially limiting the comprehensiveness of the findings.

Meta-Analysis

Figure 3 displays the funnel plots for these measures, assessing publication bias and heterogeneity. The plots for LPA and steps indicate low publication bias and heterogeneity, with data points symmetrically distributed around the funnel's center, suggesting consistent effect sizes. In contrast, the MVPA plot shows higher heterogeneity and potential publication bias, with data points more dispersed and deviating from the center line. The ST plot, while having some points at the funnel's edges, overall indicates slight publication bias and low heterogeneity, with generally consistent effect sizes. These findings highlight the need for researchers to consider heterogeneity and potential publication bias when interpreting results for different physical activity types and sedentary time (ST).

Figure 3. Funnel plot of light physical activity (LPA), moderate to vigorous physical activity (MVPA), sedentary time (ST), and steps.

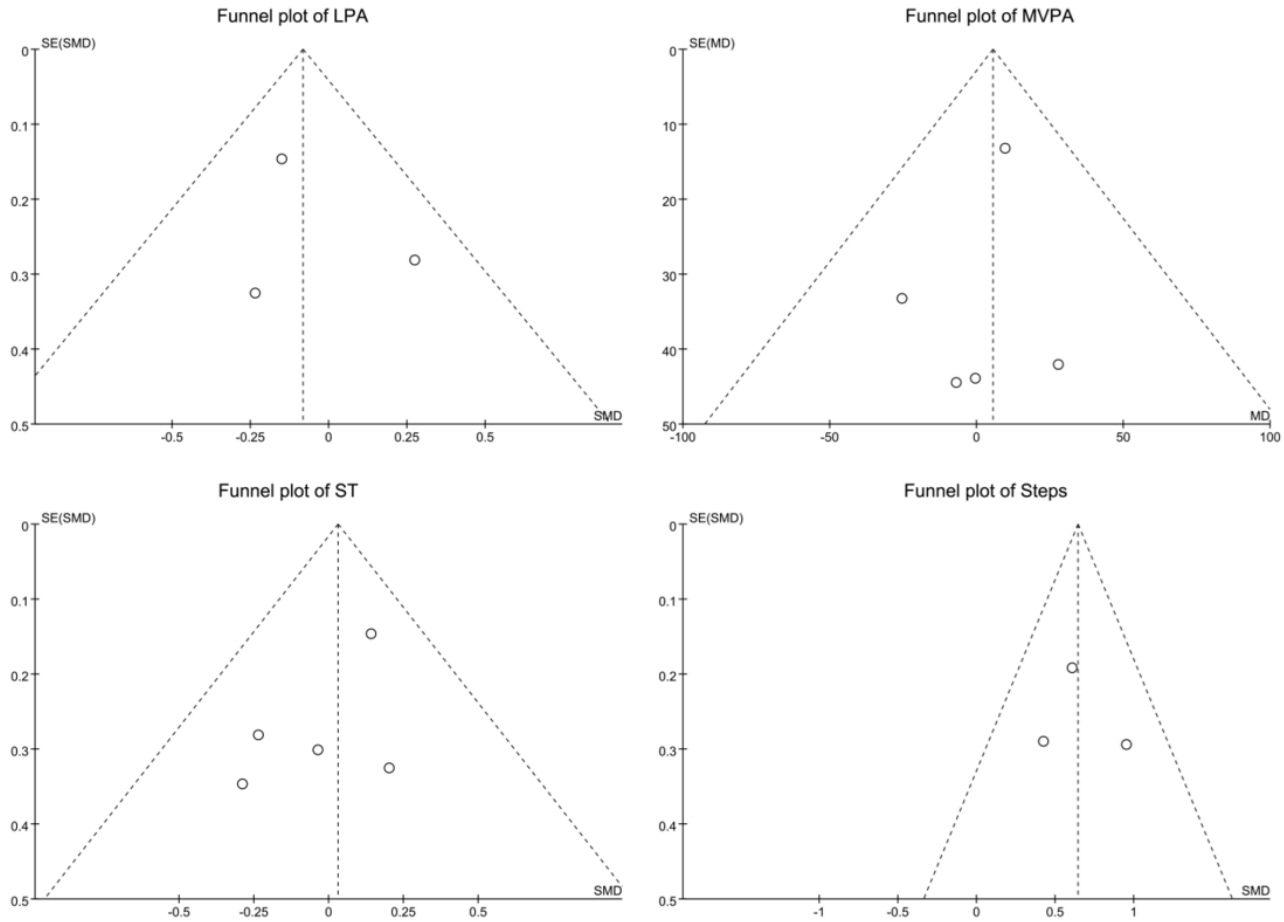

The results of the meta-analysis showed the forest plot (Figure 4 [45-47,50,51,65-67]) for the 4 outcome measures: LPA, MVPA, ST, and steps.

XSL•FO
RenderX

Figure 4. Forest plot of light physical activity (LPA), moderate to vigorous physical activity (MVPA), sedentary time (ST), and steps per week [45-47,50,51,65-67].

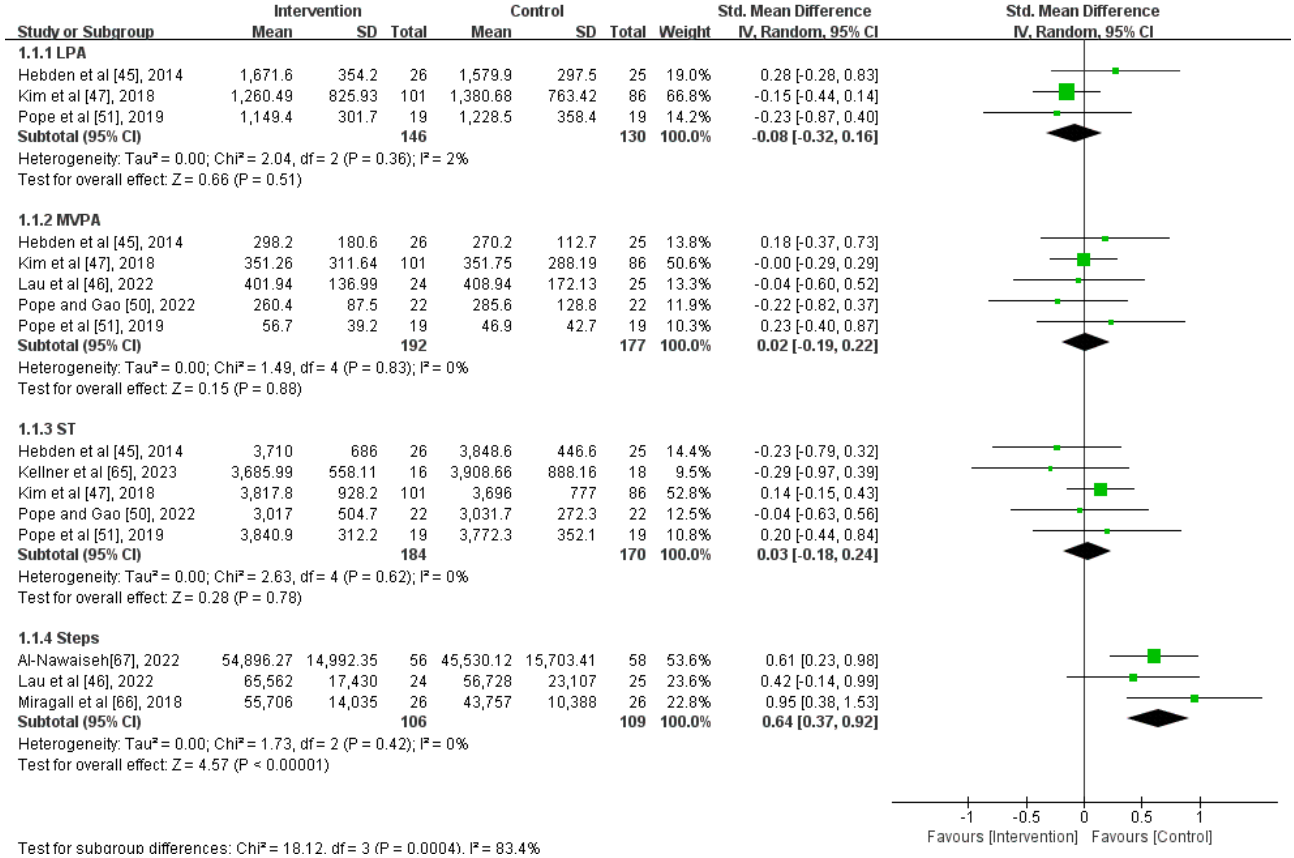

Regarding LPA, 3 studies involving 276 participants reported changes in weekly LPA. The results revealed a small SMD between the intervention and control groups, slightly favoring the control group (SMD –0.08, 95% CI –0.32 to 0.16; Figure 4 [45-47,50,51,65-67]). However, this difference was not statistically significant (*P*=.51). In addition, the results showed very low heterogeneity (I^2=2%).

For MVPA, 5 studies involving 369 participants reported changes in weekly MVPA. The results indicated a small SMD between the intervention and control groups, slightly favoring the intervention group (SMD 0.02, 95% CI –0.19 to 0.22; Figure 4 [45-47,50,51,65-67]). However, this difference was not statistically significant (*P*=.88). Furthermore, the results showed no heterogeneity (I^2=0%).

Regarding ST, 5 studies involving 354 participants reported changes in weekly ST. The results showed a small SMD between the intervention and control groups, slightly favoring the intervention group (SMD 0.03, 95% CI –0.18 to 0.24; Figure 4 [45-47,50,51,65-67]). However, this difference was not statistically significant (*P*=.78). In addition, the results showed no heterogeneity (I^2=0%).

For steps, 3 studies involving 215 participants reported changes in weekly step counts. The results revealed a large SMD between the intervention and control groups, significantly favoring the intervention group (SMD 0.64, 95% CI 0.37-0.92; Figure 4 [45-47,50,51,65-67]), with a statistically significant difference (*P*<.001). Furthermore, the results showed no heterogeneity (I^2=0%).

Sensitivity Analysis

The research findings of the 4 outcome measures, namely LPA, MVPA, ST, and steps, indicate that the point estimates of the combined effect sizes obtained after excluding a specific study fall within the 95% CI of the overall combined effect size. This suggests that the exclusion of a particular study does not significantly alter the results, indicating a considerable stability in the evaluation of the 4 outcome measures. The sensitivity analysis charts corresponding to each outcome measure are presented in figures.

The sensitivity analysis chart for LPA is shown in Figure 5 [45,47,51].

XSL•FO
RenderX

Figure 5. Sensitivity analysis chart of light physical activity (LPA) [45,47,51].

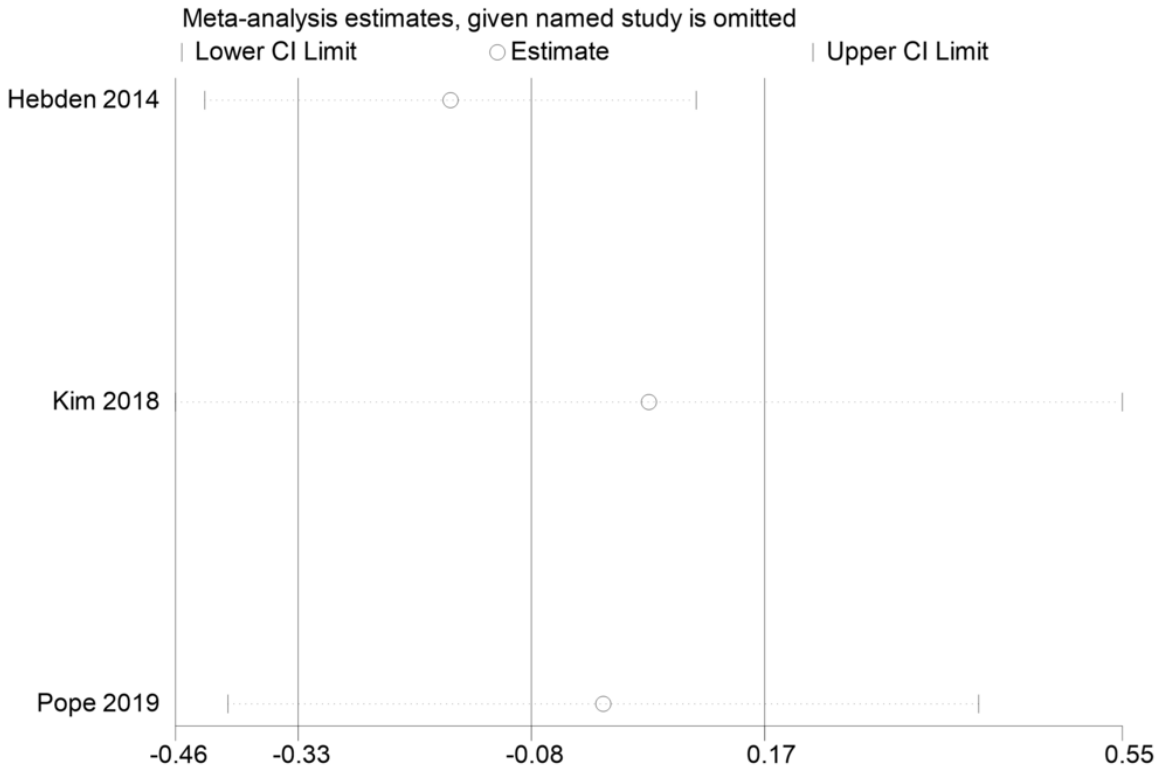

The vertical line labeled "Estimate" represents the combined results of all included studies, while the lines on either side denote the upper and lower confidence intervals of these results. A circle corresponding to any particular study indicates the combined results of the remaining studies after excluding that specific study. If, upon exclusion of a study, the circle remains close to the "Estimate" line and within the original confidence interval, it suggests that the exclusion of this study does not significantly alter the combined results. This consistency implies that the meta-analysis results are robust.

The sensitivity analysis charts for MVPA, ST, and steps are shown in Figures 6 [45-47,50,51], 7 [45,47,50,51,65], and 8 [46,66,67], respectively.

XSL•FO
RenderX

Figure 6. Sensitivity analysis chart of moderate to vigorous physical activity (MVPA) [45-47,50,51].

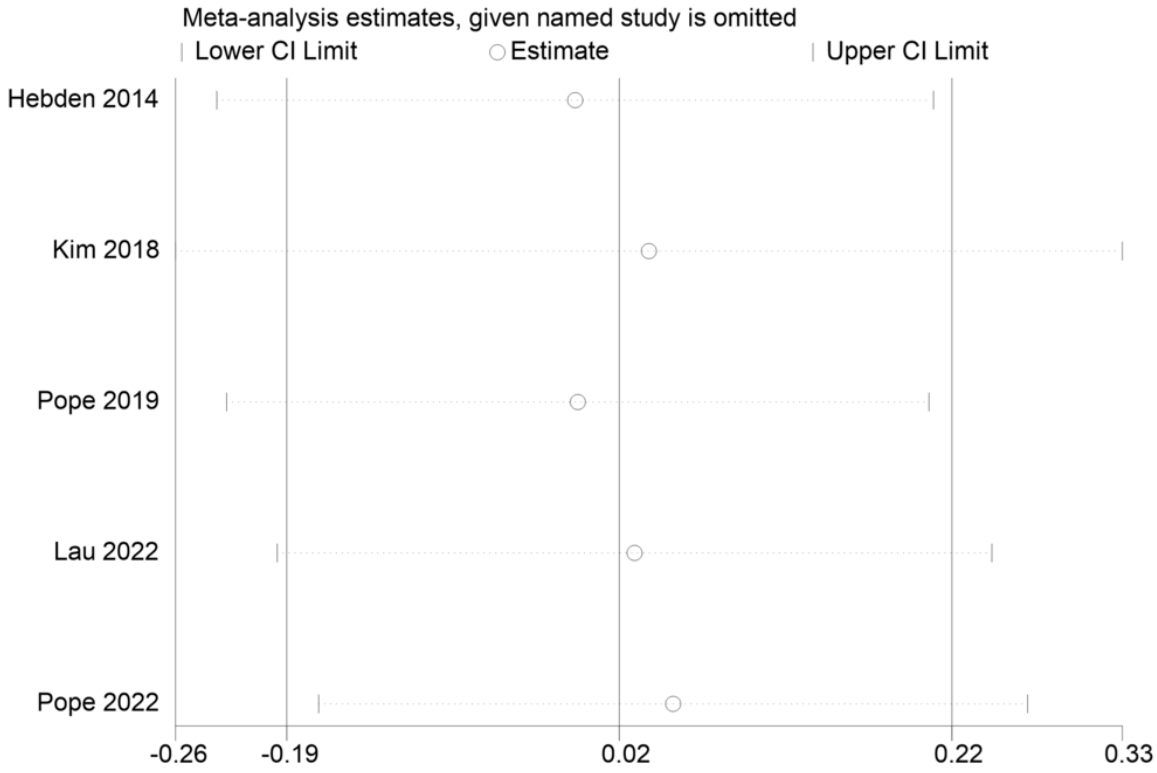

Figure 7. Sensitivity analysis chart of sedentary time (ST) [45,47,50,51,65].

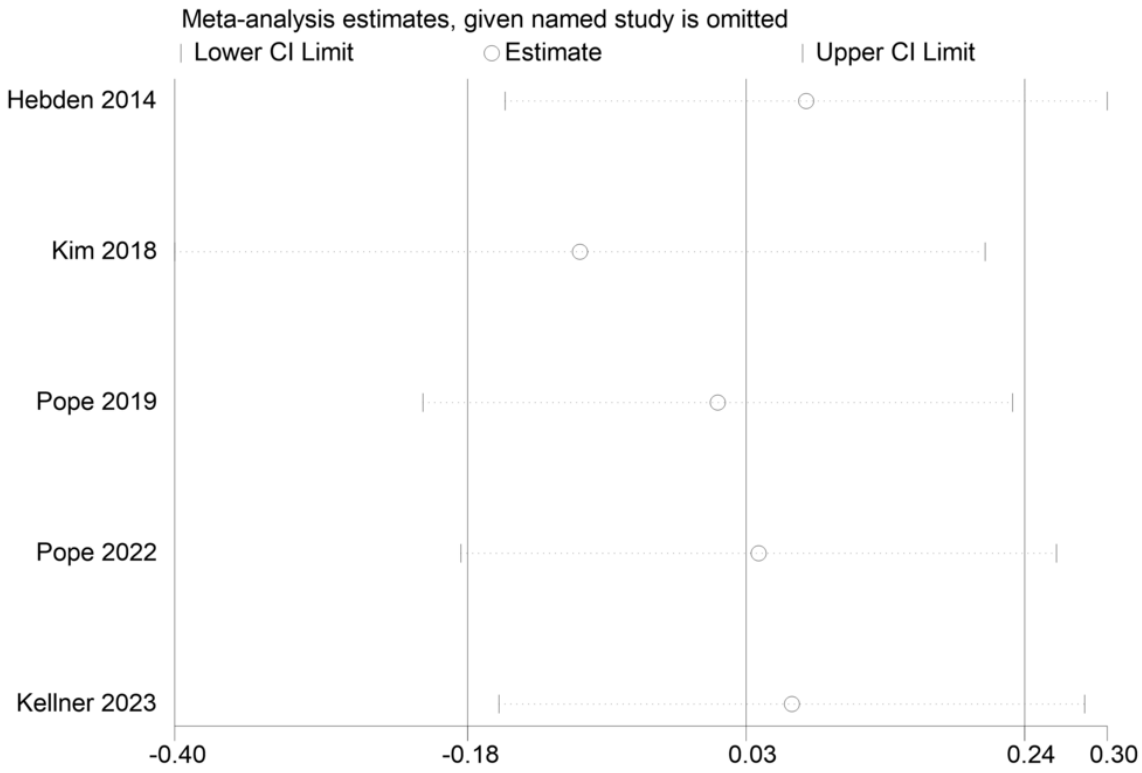

XSL•FO
RenderX

Figure 8. Sensitivity analysis chart of steps [46,66,67].

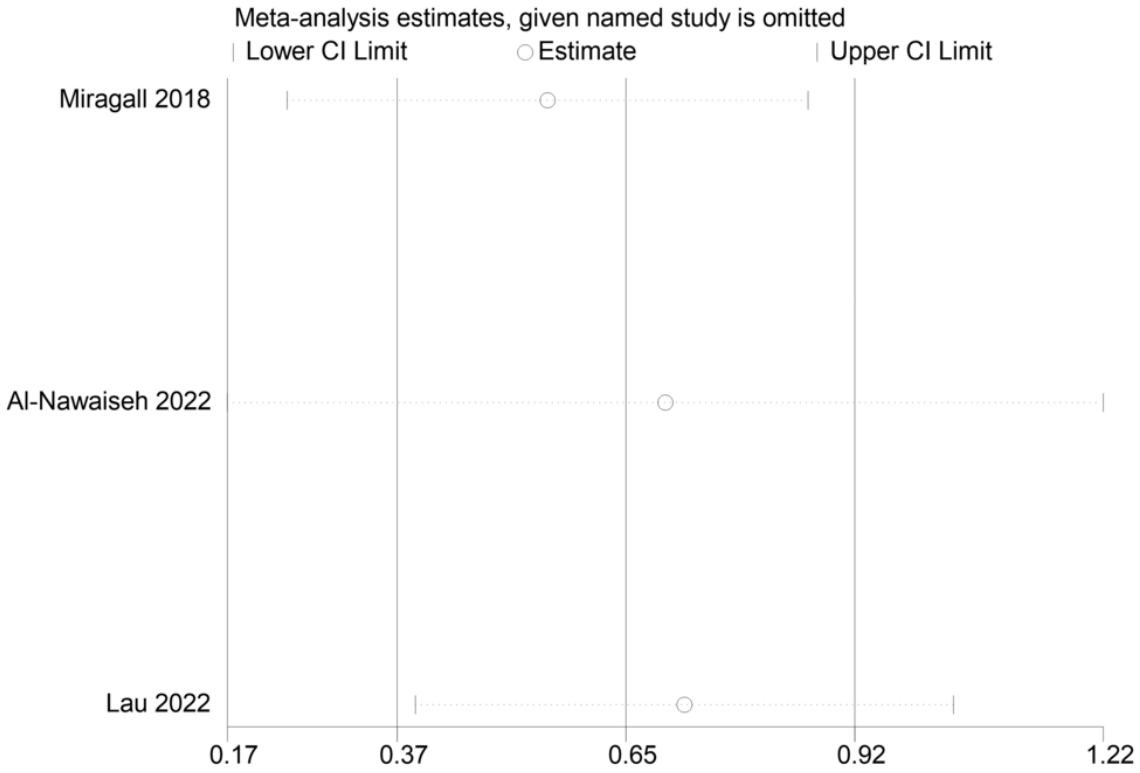

Discussion

Principal Findings

The results of the included analysis demonstrate that digital health interventions are a stable and effective strategy for significantly improving the step count of college students, thereby enhancing their physical health. This conclusion is supported by 3 randomized controlled trials included in the analysis [46,66,67]. With the widespread use of health trackers and mobile devices, daily step count has become a convenient and reliable indicator for measuring individual physical activity levels [74,75]. Research indicates that increasing daily step count has significant benefits for both physical and mental health [76]. First, increasing step count enhances cardiopulmonary function and oxygen uptake, thereby improving cardiovascular function [75,77,78]. Second, walking is a low-intensity aerobic exercise that can help lower blood pressure, control blood sugar levels, and prevent cardiovascular diseases. In addition, walking aids in calorie expenditure, weight control, and reduces the risk of obesity. Furthermore, walking helps release physical and psychological stress, improving emotional and mental well-being [79-81]. Further studies have demonstrated a correlation between daily step count and all-cause mortality. Individuals who walk 8000 to 12,000 steps per day have a lower all-cause mortality rate compared with those who walk only 4000 steps per day. This association has been validated across different age groups, genders, and ethnicities [82,83]. Therefore, focusing on step count and increasing it through intervention measures is of significant importance. Digital health interventions provide a convenient and practical approach to help college students increase their step count. By using digital health technologies such as the internet, smartphone apps, text messaging services, and wearable devices, individuals can monitor their step count, set goals, receive reminders and feedback, and more. This intervention not only alerts individuals to their walking behavior but also provides personalized advice and support to encourage increased step count.

However, the analysis results of this study indicate that digital health interventions may not be an effective method for significantly influencing the levels of low-, moderate-, and high-intensity physical activity and ST among college students. Although previous research has demonstrated that increasing the duration of low, moderate, and high-intensity physical activity can significantly reduce the negative impact of ST and improve risk factors associated with cardiovascular diseases, such as hypertension, hyperglycemia, hyperlipidemia, and mortality [84-87]. However, in this study, no significant impact of digital health interventions on the levels of low, moderate, and high-intensity physical activity and ST among college students was observed. It is worth noting that in this study, the impact of digital health interventions on step count was significant, but not significant for other modes of physical activity. This may be due to the fact that walking is a low-intensity aerobic exercise that is easily accepted and maintained by college students. In contrast, changing high-intensity physical activity and sedentary behaviors may require more personalized support and intervention measures.

Specifically, step count differs from LPA measured by accelerometers, and thus, the significance of step count does not represent the significance of LPA. Loprinzi and Cardinal

XSL•FO
RenderX

[88] found that while an increase in step count is significantly associated with health, the correlation between step count and total LPA is low, indicating that the significance of step count is not equivalent to LPA. Tudor-Locke et al [89] also noted that step count primarily measures walking, whereas LPA encompasses a broader range of low-intensity activities.

In previous studies, the impact on LPA and step count yielded different results. Hebden [45] found that LPA levels significantly increased in both the intervention and control groups, with no significant difference between them. Kim et al [47] reported a significant decrease in LPA time in the intervention group, with no significant change in the control group, yet no significant difference between the groups. Pope et al [51] observed a decrease in LPA time in the intervention group and no significant change in the control group, with no significant difference between the groups. Regarding step count, Miragall et al [66] found a significant increase in daily steps in the intervention group, with no significant change in the control group, and a significant difference between the groups. Lau et al [46] found no significant difference in objectively measured step count between the intervention and control groups. Al-Nawaiseh et al [67] observed a significant increase in weekly steps from baseline to follow-up in the intervention group, with no significant change in the control group, and a significant difference between the groups.

Therefore, to effectively promote low-, moderate-, and high-intensity physical activity among college students and improve their sedentary behaviors, it is necessary to explore more effective methods and continue this investigation in future research. This may involve providing more personalized intervention measures, such as customized exercise plans, individual goal setting, and regular feedback. In addition, combining social support and motivational mechanisms, such as team activities and competitions, may further encourage college students to engage in more high-intensity physical activity and reduce sedentary behaviors.

In light of the above, digital health interventions have been identified as an effective strategy to enhance step count and improve the physical health status of college students. By monitoring step count, setting goals, and providing personalized advice and support, digital health interventions can assist college students in increasing their daily step count. However, further research is needed to investigate the effectiveness of interventions targeting other modes of physical activity and to delve into the mechanisms underlying digital health interventions. With the continuous advancement of digital health technologies, future studies should explore more precise and personalized intervention approaches and conduct in-depth investigations into the mechanisms of digital health interventions to provide a more comprehensive and profound understanding.

Strengths and Limitations

The included studies in this meta-analysis used various intervention measures, including eHealth and mHealth technologies such as the internet, smartphone apps, SMSs, and wearable devices. All studies adhered to the standards of randomized controlled trials and used digital devices for objective measurement of physical activity outcome indicators. The research findings presented detailed results before, during, and even after follow-up, demonstrating high quality and stability. However, a limitation of this study is the relatively small sample size, which calls for further investigation.

The methodological rigor of the included studies was evaluated using the Cochrane Risk of Bias tool. The results showed that the studies generally had moderate methodological quality, meeting average quality assessment standards. Notably, except for 1 study (Lau et al [46]), all studies explicitly described a robust random sequence generation process for participant allocation, minimizing selection bias. In addition, all studies reported baseline characteristics and used valid, objective measures to assess physical activity outcomes, reducing detection bias.

However, the nature of the interventions posed challenges in achieving blinding of participants and personnel, potentially introducing performance bias. Specifically, although most studies [46,65-67] used intention-to-treat analysis to mitigate attrition bias, the lack of blinding could have influenced participant behavior and data interpretation. Regarding reporting bias, except for Lau et al [46], all studies provided sufficient statistical data for analysis. However, only 5 studies [46,47,50,51,67] reported between-group differences in changes over time, potentially limiting the comprehensiveness of the findings.

Subjective measurement of physical activity and studies involving noncollege student populations were excluded in this analysis, reducing the subjectivity and instability associated with self-report data and enhancing the reliability and consistency of the results. The accuracy of measuring LPA, MVPA, and ST may be influenced by the specific devices and wear-time protocols used in each study. This variability could result in observed inconsistencies. Sensitivity analysis further validated the stability and effectiveness of the meta-analysis results. These rigorous inclusion and exclusion criteria minimized heterogeneity and increased the credibility of the results. However, due to the limited number of included studies, subgroup analysis was not conducted, preventing an understanding of the differences in effects among different populations, intervention measures, and time points.

The key strength of this analysis lies in its focus on the impact of digital health interventions on college students, as well as the inclusion of 4 indicators of physical activity behavior. This may be the first meta-analysis on the effects of digital health interventions specifically targeting college students. However, it is important to note that this study only included English articles, which may have influenced the research sample and compromised the stability of the results. In addition, the included studies mainly originated from Western countries, lacking research on college student populations in non-Western contexts, thus potentially limiting the generalizability of the results. Further research is needed to explore the impact of digital health interventions on physical activity among college students in the future.

XSL•FO
RenderX

Conclusions

According to the research findings, digital health interventions have shown significant effects on step counts among college students. However, the impact on LPA, moderate to high-intensity physical activity, and sedentary behavior may not be evident. It is important to note that these conclusions are based on a limited number of studies, and further research is needed to validate these findings. Specifically, more research is needed to investigate the impact of digital health interventions on increasing step counts among college students. Therefore, future studies should further explore the effects of digital health interventions on changes in physical activity behavior among college students, particularly in promoting increases in step counts. This will contribute to a better understanding of the effectiveness of digital health interventions and provide scientific evidence for the development of effective intervention strategies.

Acknowledgments

The funding for this study was provided by the National Natural Science Foundation of China (project 72364006).

Authors' Contributions

SB executed the search strategies. SB and YW conducted the initial screening and full-text review. SB and JY conducted data extraction, while ZWX and ZLQ resolved the conflict. SB conducted narrative analysis and meta-analyses, which were later confirmed by ZYJ and ZR. LI supervised the entire process of the project. All authors contributed to the manuscript drafting.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Searching Strategy.
[PDF File (Adobe PDF File), 107 KB - jmir_v26i1e51714_app1.pdf]

Multimedia Appendix 2
Study Characteristics.
[PDF File (Adobe PDF File), 2661 KB - jmir_v26i1e51714_app2.pdf]

Multimedia Appendix 3
PRISMA checklist.
[DOCX File , 28 KB - jmir_v26i1e51714_app3.docx]

References

1. Arnett JJ. Emerging adulthood. A theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [Medline: 10842426]
2. Yahia N, Wang D, Rapley M, Dey R. Assessment of weight status, dietary habits and beliefs, physical activity, and nutritional knowledge among university students. Perspect Public Health 2016;136(4):231-244. [doi: 10.1177/1757913915609945] [Medline: 26475773]
3. Aceijas C, Waldhäusl S, Lambert N, Cassar S, Bello-Corassa R. Determinants of health-related lifestyles among university students. Perspect Public Health 2017;137(4):227-236. [doi: 10.1177/1757913916666875] [Medline: 27597797]
4. Park MJ, Scott JT, Adams SH, Brindis CD, Irwin CE. Adolescent and young adult health in the United States in the past decade: little improvement and young adults remain worse off than adolescents. J Adolesc Health 2014;55(1):3-16. [doi: 10.1016/j.jadohealth.2014.04.003] [Medline: 24815958]
5. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA 2018;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
6. Abu-Moghli FA, Khalaf IA, Barghoti FF. The influence of a health education programme on healthy lifestyles and practices among university students. Int J Nurs Pract 2010;16(1):35-42. [doi: 10.1111/j.1440-172X.2009.01801.x] [Medline: 20158546]
7. Wisloff U, Lavie CJ. Taking physical activity, exercise, and fitness to a higher level. Prog Cardiovasc Dis 2017;60(1):1-2. [doi: 10.1016/j.pcad.2017.06.002] [Medline: 28793972]
8. Lavie CJ, Arena R, Swift DL, Johannsen NM, Sui X, Lee DC, et al. Exercise and the cardiovascular system: clinical science and cardiovascular outcomes. Circ Res 2015;117(2):207-219 [FREE Full text] [doi: 10.1161/CIRCRESAHA.117.305205] [Medline: 26139859]
9. Lavie CJ, Ozemek C, Carbone S, Katzmarzyk PT, Blair SN. Sedentary behavior, exercise, and cardiovascular health. Circ Res 2019;124(5):799-815. [doi: 10.1161/CIRCRESAHA.118.312669] [Medline: 30817262]

XSL•FO
RenderX

10. Global Strategy on Diet, Physical Activity and Health : A Framework to Monitor and Evaluate Implementation. Geneva, Switzerland: World Health Organization; 2004.
11. More Physical Activity (RUN). Global action plan on physical activity 2018–2030: more active people for a healthier world.: World Health Organization; 2018. URL: https://www.who.int/publications/i/item/9789241514187 [accessed 2023-09-19]
12. Vuori I. World health organization and physical activity. Prog prev med 2018;3(1):e0012 [FREE Full text]
13. Physical activity.: World Health Organization; 2022. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2021-09-19]
14. Donnelly JE, Hillman CH, Castelli D, Etnier JL, Lee S, Tomporowski P, et al. Physical activity, fitness, cognitive function, and academic achievement in children: a systematic review. Med Sci Sports Exerc 2016;48(6):1197-1222 [FREE Full text] [doi: 10.1249/MSS.0000000000000901] [Medline: 27182986]
15. Stubbs B, Vancampfort D, Hallgren M, Firth J, Veronese N, Solmi M, et al. EPA guidance on physical activity as a treatment for severe mental illness: a meta-review of the evidence and position statement from the European psychiatric association (EPA), supported by the international organization of physical therapists in mental health (IOPTMH). Eur Psychiatry 2018;54:124-144. [doi: 10.1016/j.eurpsy.2018.07.004] [Medline: 30257806]
16. Ferreira Silva RM, Mendonça CR, Azevedo VD, Raoof Memon A, Noll PRES, Noll M. Barriers to high school and university students' physical activity: a systematic review. PLoS One 2022;17(4):e0265913 [FREE Full text] [doi: 10.1371/journal.pone.0265913] [Medline: 35377905]
17. Iribarren SJ, Cato K, Falzon L, Stone PW. What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. PLoS One 2017;12(2):e0170581 [FREE Full text] [doi: 10.1371/journal.pone.0170581] [Medline: 28152012]
18. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
19. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. Geneva, Switzerland: World Health Organization; 2016.
20. Smith AJ, Skow Á, Bodurtha J, Kinra S. Health information technology in screening and treatment of child obesity: a systematic review. Pediatrics 2013;131(3):e894-e902. [doi: 10.1542/peds.2012-2011] [Medline: 23382447]
21. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018;6(3):e65 [FREE Full text] [doi: 10.2196/mhealth.9131] [Medline: 29549071]
22. Tong HL, Coiera E, Laranjo L. Using a mobile social networking app to promote physical activity: a qualitative study of users' perspectives. J Med Internet Res 2018;20(12):e11439 [FREE Full text] [doi: 10.2196/11439] [Medline: 30578201]
23. Müssener U, Thomas K, Bendtsen P, Bendtsen M. High school students' preferences and design recommendations for a mobile phone-based intervention to improve psychological well-being: mixed methods study. JMIR Pediatr Parent 2020;3(2):e17044 [FREE Full text] [doi: 10.2196/17044] [Medline: 32673268]
24. Recommendations on Digital Interventions for Health System Strengthening. Geneva, Switzerland: World Health Organization; 2019.
25. WHO Global Observatory for eHealth. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. Geneva, Switzerland: World Health Organization; 2011.
26. Berman AH, Gajecki M, Sinadinovic K, Andersson C. Mobile interventions targeting risky drinking among university students: a review. Curr Addict Rep 2016;3:166-174 [FREE Full text] [doi: 10.1007/s40429-016-0099-6] [Medline: 27226948]
27. Dute DJ, Bemelmans WJE, Breda J. Using mobile apps to promote a healthy lifestyle among adolescents and students: a review of the theoretical basis and lessons learned. JMIR Mhealth Uhealth 2016;4(2):e39 [FREE Full text] [doi: 10.2196/mhealth.3559] [Medline: 27150850]
28. Prosser T, Gee KA, Jones F. A meta-analysis of effectiveness of E-interventions to reduce alcohol consumption in college and university students. J Am Coll Health 2018;66(4):292-301. [doi: 10.1080/07448481.2018.1440579] [Medline: 29452058]
29. Åsberg K, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, Löf M, et al. Multiple lifestyle behaviour mHealth intervention targeting Swedish college and university students: protocol for the buddy randomised factorial trial. BMJ Open 2021;11(12):e051044. [doi: 10.1136/bmjopen-2021-051044]
30. Farrer L, Gulliver A, Chan JK, Batterham PJ, Reynolds J, Calear A, et al. Technology-based interventions for mental health in tertiary students: systematic review. J Med Internet Res 2013;15(5):e101 [FREE Full text] [doi: 10.2196/jmir.2639] [Medline: 23711740]
31. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014;16(5):e130 [FREE Full text] [doi: 10.2196/jmir.3142] [Medline: 24836465]
32. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]

XSL•FO
RenderX

33. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
34. Prensky M. Digital natives, digital immigrants part 1. Horizon J Hum Soc Sci Res 2001;9(5):1-6. [doi: 10.1108/10748120110424816]
35. Prensky M. Digital natives, digital immigrants part 2: do they really think differently? On the Horizon 2001;9(6):1-6. [doi: 10.1108/10748120110424843]
36. Coyne SM, Padilla-Walker LM, Howard E. Emerging in a digital world: a decade review of media use, effects, and gratifications in emerging adulthood. Emerging Adulthood 2013;1(2):125-137. [doi: 10.1177/2167696813479782]
37. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013;97:41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
38. O'Brien LM. Integrating text-messaging and web-based interventions to improve physical activity and nutrition among college students. Boston, MA: Boston University Libraries; 2014.
39. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? systematic review and meta-analysis. J Med Internet Res 2019;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
40. Stephens J, Allen J. Mobile phone interventions to increase physical activity and reduce weight: a systematic review. J Cardiovasc Nurs 2013;28(4):320-329 [FREE Full text] [doi: 10.1097/JCN.0b013e318250a3e7] [Medline: 22635061]
41. Müller AM, Alley S, Schoeppe S, Vandelanotte C. Int J Behav Nutr Phys Act 2016;13(1):109 [FREE Full text] [doi: 10.1186/s12966-016-0434-2] [Medline: 27724911]
42. Gal R, May AM, van Overmeeren EJ, Simons M, Monninkhof EM. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. Sports Med Open 2018;4(1):42 [FREE Full text] [doi: 10.1186/s40798-018-0157-9] [Medline: 30178072]
43. Kim HN, Seo K. Smartphone-based health program for improving physical activity and tackling obesity for young adults: a systematic review and meta-analysis. Int J Environ Res Public Health 2019;17(1):15 [FREE Full text] [doi: 10.3390/ijerph17010015] [Medline: 31861359]
44. Wibowo RA, Kelly P, Baker G. The effect of smartphone application interventions on physical activity level among university/college students: a systematic review protocol. Physical Therapy Reviews 2020;25(2):135-142. [doi: 10.1080/10833196.2020.1756125]
45. Hebden L, Cook A, van der Ploeg HP, King L, Bauman A, Allman-Farinelli M. A mobile health intervention for weight management among young adults: a pilot randomised controlled trial. J Hum Nutr Diet 2014;27(4):322-332. [doi: 10.1111/jhn.12155] [Medline: 23992038]
46. Lau PWC, Wang JJ, Ransdell LL, Shi L. The effectiveness of facebook as a social network intervention to increase physical activity in Chinese young adults. Front Public Health 2022;10:912327 [FREE Full text] [doi: 10.3389/fpubh.2022.912327] [Medline: 35937270]
47. Kim Y, Lumpkin A, Lochbaum M, Stegemeier S, Kitten K. Promoting physical activity using a wearable activity tracker in college students: a cluster randomized controlled trial. J Sports Sci 2018;36(16):1889-1896. [doi: 10.1080/02640414.2018.1423886] [Medline: 29318916]
48. Davies CA, Spence JC, Vandelanotte C, Caperchione CM, Mummery W. Meta-analysis of internet-delivered interventions to increase physical activity levels. Int J Behav Nutr Phys Act 2012;9:52 [FREE Full text] [doi: 10.1186/1479-5868-9-52] [Medline: 22546283]
49. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
50. Pope ZC, Gao Z. Feasibility of smartphone application- and social media-based intervention on college students' health outcomes: a pilot randomized trial. J Am Coll Health 2022;70(1):89-98. [doi: 10.1080/07448481.2020.1726925] [Medline: 32150514]
51. Pope ZC, Barr-Anderson DJ, Lewis BA, Pereira MA, Gao Z. Use of wearable technology and social media to improve physical activity and dietary behaviors among college students: a 12-week randomized pilot study. Int J Environ Res Public Health 2019;16(19):3579 [FREE Full text] [doi: 10.3390/ijerph16193579] [Medline: 31557812]
52. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
53. Strain T, Wijndaele K, Dempsey PC, Sharp SJ, Pearce M, Jeon J, et al. Wearable-device-measured physical activity and future health risk. Nat Med 2020;26(9):1385-1391 [FREE Full text] [doi: 10.1038/s41591-020-1012-3] [Medline: 32807930]
54. Goodyear VA, Wood G, Skinner B, Thompson JL. The effect of social media interventions on physical activity and dietary behaviours in young people and adults: a systematic review. Int J Behav Nutr Phys Act 2021;18(1):72 [FREE Full text] [doi: 10.1186/s12966-021-01138-3] [Medline: 34090469]

XSL•FO
RenderX

55. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
56. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
57. Higgins JPT, Deeks JJ. Selecting studies and collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. Hoboken, New Jersey: John Wiley & Sons, Ltd; 2008:151-185.
58. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins J, et al. Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
59. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
60. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
61. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
62. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
63. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med 1997;127(9):820-826. [doi: 10.7326/0003-4819-127-9-199711010-00008] [Medline: 9382404]
64. Cohen J. Statistical Power Analysis for the Behavioral Sciences 2nd ed. New York: Routledge; 1988.
65. Kellner M, Dold C, Lohkamp M. Objectively assessing the effect of a messenger-based intervention to reduce sedentary behavior in university students: a pilot study. J Prev (2022) 2023;44(5):521-534 [FREE Full text] [doi: 10.1007/s10935-023-00735-1] [Medline: 37171555]
66. Miragall M, Domínguez-Rodríguez A, Navarro J, Cebolla A, Baños RM. Increasing physical activity through an internet-based motivational intervention supported by pedometers in a sample of sedentary students: a randomised controlled trial. Psychol Health 2018;33(4):465-482. [doi: 10.1080/08870446.2017.1368511] [Medline: 28880576]
67. Al-Nawaiseh HK, McIntosh WA, McKyer LJ. An-m-health intervention using smartphone app to improve physical activity in college students: a randomized controlled trial. Int J Environ Res Public Health 2022;19(12):7228. [doi: 10.3390/ijerph19127228] [Medline: 35742477]
68. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport 2011;14(5):411-416. [doi: 10.1016/j.jsams.2011.04.003] [Medline: 21616714]
69. Kozey SL, Staudenmayer JW, Troiano RP, Freedson PS. Comparison of the ActiGraph 7164 and the ActiGraph GT1M during self-paced locomotion. Med Sci Sports Exerc 2010;42(5):971-976 [FREE Full text] [doi: 10.1249/MSS.0b013e3181c29e90] [Medline: 19997000]
70. Jakicic JM, Marcus M, Gallagher KI, Randall C, Thomas E, Goss FL, et al. Evaluation of the SenseWear pro armband to assess energy expenditure during exercise. Med Sci Sports Exerc 2004;36(5):897-904. [doi: 10.1249/01.mss.0000126805.32659.43] [Medline: 15126727]
71. Lee KY, Macfarlane DJ, Cerin E. Comparison of three models of actigraph accelerometers during free living and controlled laboratory conditions. Eur J Sport Sci 2013;13(3):332-339 [FREE Full text] [doi: 10.1080/17461391.2011.643925] [Medline: 23679150]
72. Migueles JH, Cadenas-Sanchez C, Ekelund U, Delisle Nyström C, Mora-Gonzalez J, Löf M, et al. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations. Sports Med 2017;47(9):1821-1845 [FREE Full text] [doi: 10.1007/s40279-017-0716-0] [Medline: 28303543]
73. O'Brien MW, Wu Y, Petterson JL, Bray NW, Kimmerly DS. Validity of the ActivPAL monitor to distinguish postures: a systematic review. Gait Posture 2022;94:107-113. [doi: 10.1016/j.gaitpost.2022.03.002] [Medline: 35276456]
74. Paluch AE, Bajpai S, Bassett DR, Carnethon MR, Ekelund U, Evenson KR, Steps for Health Collaborative. Daily steps and all-cause mortality: a meta-analysis of 15 international cohorts. Lancet Public Health 2022;7(3):e219-e228 [FREE Full text] [doi: 10.1016/S2468-2667(21)00302-9] [Medline: 35247352]
75. Bravata DM, Smith-Spangler C, Sundaram V, Gienger AL, Lin N, Lewis R, et al. Using pedometers to increase physical activity and improve health: a systematic review. JAMA 2007;298(19):2296-2304. [doi: 10.1001/jama.298.19.2296] [Medline: 18029834]
76. Warburton DER, Nicol CW, Bredin SSD. Health benefits of physical activity: the evidence. CMAJ 2006;174(6):801-809 [FREE Full text] [doi: 10.1503/cmaj.051351] [Medline: 16534088]
77. Tudor-Locke C, Bassett DR. How many steps/day are enough? Preliminary pedometer indices for public health. Sports Med 2004;34(1):1-8. [doi: 10.2165/00007256-200434010-00001] [Medline: 14715035]
78. Murtagh EM, Murphy MH, Boone-Heinonen J. Walking: the first steps in cardiovascular disease prevention. Curr Opin Cardiol 2010;25(5):490-496 [FREE Full text] [doi: 10.1097/HCO.0b013e32833ce972] [Medline: 20625280]

XSL•FO
RenderX

79. Moore CC, Evenson KR, Shiroma EJ, Howard AG, Cuthbertson CC, Buring JE, et al. Abstract 069: sporadic step accumulation and all-cause mortality in the women's health study: do all steps count? Circulation 2021;143(Suppl_1):A069.
80. Warburton DER, Bredin SSD. Reflections on physical activity and health: what should we recommend? Can J Cardiol 2016;32(4):495-504. [doi: 10.1016/j.cjca.2016.01.024] [Medline: 26995692]
81. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol 2017;32(5):541-556. [doi: 10.1097/HCO.0000000000000437] [Medline: 28708630]
82. Number of steps per day more important than step intensity.: National Institutes of Health URL: https://www.nih.gov/news-events/nih-research-matters/number-steps-day-more-important-step-intensity [accessed 2023-09-19]
83. Total daily steps, not step intensity, offer more benefits.: Harvard Health Publishing; 2020. URL: https://www.health.harvard.edu/staying-healthy/total-daily-steps-not-step-intensity-offer-more-benefits [accessed 2023-09-19]
84. Buman MP, Hekler EB, Haskell WL, Pruitt L, Conway TL, Cain KL, et al. Objective light-intensity physical activity associations with rated health in older adults. Am J Epidemiol 2010;172(10):1155-1165 [FREE Full text] [doi: 10.1093/aje/kwq249] [Medline: 20843864]
85. Chastin SFM, Palarea-Albaladejo J, Dontje ML, Skelton DA. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. PLoS One 2015;10(10):e0139984 [FREE Full text] [doi: 10.1371/journal.pone.0139984] [Medline: 26461112]
86. Chastin SFM, De Craemer M, De Cocker K, Powell L, Van Cauwenberg J, Dall P, et al. How does light-intensity physical activity associate with adult cardiometabolic health and mortality? Systematic review with meta-analysis of experimental and observational studies. Br J Sports Med 2019;53(6):370-376 [FREE Full text] [doi: 10.1136/bjsports-2017-097563] [Medline: 29695511]
87. Farrahi V, Rostami M, Dumuid D, Chastin SFM, Niemelä M, Korpelainen R, et al. Joint profiles of sedentary time and physical activity in adults and their associations with cardiometabolic health. Med Sci Sports Exerc 2022;54(12):2118-2128 [FREE Full text] [doi: 10.1249/MSS.0000000000003008] [Medline: 35881930]
88. Loprinzi PD, Cardinal BJ. Measuring children's physical activity and sedentary behaviors. J exerc sci fit 2011;9(1):15-23. [doi: 10.1016/s1728-869x(11)60002-6]
89. Tudor-Locke C, Craig CL, Brown WJ, Clemes SA, De Cocker K, Giles-Corti B, et al. How many steps/day are enough? For adults. Int J Behav Nutr Phys Act 2011;8:79 [FREE Full text] [doi: 10.1186/1479-5868-8-79] [Medline: 21798015]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
eHealth: electronic health
LPA: light intensity physical activity
mHealth: mobile health
MVPA: moderate to vigorous physical activity
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
SMD: standardized mean difference
ST: sedentary time

Edited by T de Azevedo Cardoso; submitted 09.08.23; peer-reviewed by J Thomas, H Liu; comments to author 06.04.24; revised version received 03.06.24; accepted 27.08.24; published 20.11.24.

Please cite as:
Bi S, Yuan J, Wang Y, Zhang W, Zhang L, Zhang Y, Zhu R, Luo L
Effectiveness of Digital Health Interventions in Promoting Physical Activity Among College Students: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e51714
URL: https://www.jmir.org/2024/1/e51714
doi:10.2196/51714
PMID:

©Siyuan Bi, Junfeng Yuan, Yanling Wang, Wenxin Zhang, Luqin Zhang, Yongjuan Zhang, Rui Zhu, Lin Luo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Cybersecurity Interventions in Health Care Organizations in Low- and Middle-Income Countries: Scoping Review

Kaede Hasegawa[1], MSc; Niki O'Brien[1], MSc; Mabel Prendergast[1], MBBS; Chris Agape Ajah[1], MBBS, MSc; Ana Luisa Neves[2,3], MBChB, MSc, PhD; Saira Ghafur[1], MBChB, MRCP, MSc

[1]Institute of Global Health Innovation, Imperial College London, London, United Kingdom

[2]Department of Primary Care and Public Health, Imperial College London, London, United Kingdom

[3]RISE-Health, Faculty of Medicine of the University of Porto, Porto, Portugal

Corresponding Author:
Niki O'Brien, MSc
Institute of Global Health Innovation
Imperial College London
10th Floor, St Mary's Hospital
Praed Street
London, W2 1NY
United Kingdom
Phone: 44 020 7594 1419
Email: n.obrien@imperial.ac.uk

Abstract

Background: Health care organizations globally have seen a significant increase in the frequency of cyberattacks in recent years. Cyberattacks cause massive disruptions to health service delivery and directly impact patient safety through disruption and treatment delays. Given the increasing number of cyberattacks in low- and middle-income countries (LMICs), there is a need to explore the interventions put in place to plan for cyberattacks and develop cyber resilience.

Objective: This study aimed to describe cybersecurity interventions, defined as any intervention to improve cybersecurity in a health care organization, including but not limited to organizational strategy(ies); policy(ies); protocol(s), incident plan(s), or assessment process(es); framework(s) or guidelines; and emergency planning, implemented in LMICs to date and to evaluate their impact on the likelihood and impact of attacks. The secondary objective was to describe the main barriers and facilitators for the implementation of such interventions, where reported.

Methods: A systematic search of the literature published between January 2017 and July 2024 was performed on Ovid Medline, Embase, Global Health, and Scopus using a combination of controlled terms and free text. A search of the gray literature within the same time parameters was undertaken on the websites of relevant stakeholder organizations to identify possible additional studies that met the inclusion criteria. Findings from included papers were mapped against the dimensions of the Essentials of Cybersecurity in Health Care Organizations (ECHO) framework and presented as a narrative synthesis.

Results: We included 20 studies in this review. The sample size of the majority of studies (13/20, 65%) was 1 facility to 5 facilities, and the studies were conducted in 14 countries. Studies were categorized into the thematic dimensions of the ECHO framework, including context; governance; organizational strategy; risk management; awareness, education, and training; and technical capabilities. Few studies (6/20, 30%) discussed cybersecurity intervention(s) as the primary focus of the paper; therefore, information on intervention(s) implemented had to be deduced. There was no attempt to report on the impact and outcomes in all papers except one. Facilitators and barriers identified were grouped and presented across national or regional, organizational, and individual staff levels.

Conclusions: This scoping review's findings highlight the limited body of research published on cybersecurity interventions implemented in health care organizations in LMICs and large heterogeneity across existing studies in interventions, research objectives, methods, and outcome measures used. Although complex and challenging, future research should specifically focus on the evaluation of cybersecurity interventions and their impact in order to build a robust evidence base to inform evidence-based policy and practice.

(J Med Internet Res 2024;26:e47311) doi:10.2196/47311

XSL•FO
RenderX

KEYWORDS

computer security; internet security; network security; digital health; digital health technology; cybersecurity; health data; global health; security; data science; LMIC; low income; low resource; scoping review; review methodology; implementation; barrier; facilitator

Introduction

Background

Health systems globally are incorporating technology into every aspect of the delivery of care [1]. The health sector is also one of the most targeted and profitable sectors for cyberattacks [2], defined as the malicious attempt to gain unauthorized access to online systems or computers [3]. The United States National Institute of Standards and Technology (NIST) expands upon this definition, describing a cyberattack as "any kind of malicious activity that attempts to collect, disrupt, deny, degrade, or destroy information system resources or the information itself" [4]. Medical records are more valuable than credit card details on the dark web in some countries, due to the inclusion of both patient identifiers and financial information [5]. Health care organizations globally have seen a significant increase in the frequency of cyberattacks in recent years as technology plays more of a central role in the delivery of care [6,7].

In 2017, the WannaCry malware attack targeted computers running with an unsupported Microsoft Windows operating system [8]. The impact on the UK health system was far-reaching, and the incident remains the largest cyberattack to affect the National Health Service (NHS), with 34 hospital trusts and 603 primary care and other NHS organizations directly affected [8]. In 2021, Ireland's health service faced a ransomware attack after an employee in the Irish Health Service Executive opened a spreadsheet email attachment compromised with malware [9]. In 2023, Pennsylvania Lehigh Valley Health Network was also the target of a ransomware attack in which the clinical images of individuals receiving cancer treatment were released [10]. These attacks serve only as examples of the many cyberattacks the health sector has faced. Health care cybersecurity is underpinned by staff understanding of the significance of threats, as most security breaches occur through human error.

Cyberattacks not only cause massive disruption to health service delivery but also directly impact patient safety. They disrupt and delay treatment, as well as threaten the safe use of personal data, which can lead to mistrust by patients and the public toward health care providers [11]. In extreme cases, cyberattacks may lead to patient morbidity and mortality. For example, analysis of the impacts of the UK NHS WannaCry attack showed a 6% decrease in total admissions per infected hospital per day during WannaCry, with 4% fewer emergency admissions and 9% fewer elective admissions [12]. Research has shown that delays to hospital inpatient admissions in excess of 5 hours from time of arrival at the emergency department are associated with an increase in all-cause 30-day mortality. Furthermore, for every 82 admitted patients whose time to inpatient bed transfer is delayed beyond 6 hours to 8 hours from arrival time, there is 1 additional death [13]. In a separate study, delayed elective surgery patients also reported a loss of working days for themselves and family members; increased disappointment, frustration, and stress; concern for continued symptoms; and deteriorating conditions [14]. Such downstream impacts outline the potential far-reaching impacts cybersecurity attacks and the disruption caused have on patients.

Health care cybersecurity is increasingly relevant to health systems globally, including in low- and middle-income countries (LMICs). In LMICs, the implementation of digital technologies in health care service provision is rapidly accelerating [15] and becoming an increasing part of the health agenda [7]. Health care systems in LMICs are progressively using digital innovations, including health care informatics systems, electronic health care records, and wearables [11].

Health systems in LMICs have also seen an increase in cyber threats [3]. East Asia Saraburi Hospital in Thailand was the target of a ransomware attack in 2020, followed by another cyberattack on the Public Health Ministry. Together, these resulted in the potential theft of 16 million patient records [16]. The incidents led the Thai National Cyber Security Agency to announce 3 subordinate laws under the Cybersecurity Act to develop stronger cybersecurity. Among these laws was a mandate to provide cybersecurity training and conduct risk assessments for those dealing with critical information [16]. Similarly, Life Healthcare, a private provider in South Africa, was attacked in 2020, disrupting the admissions systems, business processing systems, and email servers, though patient care was not impacted due to a swift switch to backup systems [17].

Despite the increase in cyberattacks in the health sector in LMICs, there is limited evidence on the number and nature of cyberattacks as well as interventions put in place to plan for cyberattacks and develop cyber resilience in these settings. Cybersecurity interventions are defined as any intervention to improve cybersecurity in a health care organization. These include but are not limited to organizational strategy(ies); policy(ies) (eg, policy for installation of appropriate malware and virus protection software, access management policy); protocol(s), incident plan(s), or process(es) (eg, processes around identity management, threat detection); framework(s) or guidelines; education (eg, courses, training); and emergency planning. Given increasing cybersecurity threats, it is essential to develop a knowledge base on interventions found to be effective at improving cyber preparedness in the health sector. A first step in achieving this is to understand which interventions have been developed and what their impact has been,

Objectives

The primary objectives of this scoping review were to (1) describe cybersecurity interventions implemented in LMICs to date and (2) evaluate their impact on the likelihood and impact of attacks. As a secondary objective, we aimed to identify the main barriers and facilitators for the implementation of such interventions, where reported.

XSL•FO
RenderX

Methods

This scoping review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines [18].

Search Strategy

A systematic search of the literature published between January 2017 and July 2024 (current) was performed on Ovid Medline, Embase, Global Health, and Scopus using a combination of controlled terms and free text, depending on the database functionality. The detailed search strategy is provided in Multimedia Appendix 1. Searches were conducted in 3 health-specific (Ovid Medline, Embase, and Global Health) databases, and as such, no health care keywords were used in the search. Given the multidisciplinary topic area and the importance of capturing literature from information and communication technology (ICT) and computer science disciplines, Scopus was also searched with specific health care keywords ("health," "healthcare," "medical," "hospital," and "clinic"). We selected Scopus because the IEEE Xplore Library, a leading source of cybersecurity-focused research, is indexed by the database.

A search of the gray literature within the same time parameters was undertaken to identify possible additional studies that met the inclusion criteria. Searches were conducted on the websites of relevant stakeholder organizations (ie, the World Bank and the World Health Organization [WHO] including WHO regional offices and the WHO International Clinical Trials Registry Platform Search Portal), Ponemon Institute, Cybersecurity and Infrastructure Security Agency, CyberPeace Institute, Gartner, and Verizon) and conference proceedings from related conferences (ie, Healthcare Information and Management Systems Society, Global Forum on Cyber Expertise).

Selection Criteria

We included any health care–focused study undertaken in (1) an LMIC country that (2) described 1 or multiple interventions related to cyberattacks or cybersecurity (Textbox 1). Only studies published in English were included. Feasibility studies, commentaries, and editorial papers were excluded.

Textbox 1. Inclusion and exclusion criteria.

Article type

- Inclusion criterion: original research study, including systematic reviews
- Exclusion criteria: feasibility studies, commentaries, and editorial papers

Language

- Inclusion criterion: English
- Exclusion criterion: any other language

LMICs are defined by the World Bank based on gross national income per capita [19]; however, it has been noted that the categorization should only be used when relevant to the research study [20]. Health systems in LMICs are often challenged by resourcing constraints at a national level, which filter down to local systems and providers, and, as such, likely face common resourcing challenges for developing and implementing cybersecurity interventions. Given these shared challenges, focusing the research on LMICs as a general group was agreed by the authors. Notably, our use of the term serves as a starting point to explore cybersecurity interventions rather than an end point, as further nuanced research and evidence generation will be required.

Cybersecurity interventions were defined as any intervention to improve cybersecurity in a health care organization, including but not limited to organizational strategy(ies); policy(ies); protocol(s), incident plan(s), or assessment process(es); framework(s) or guidelines; education (eg, courses, training); and emergency planning.

Title and abstract screening followed by full-text screening were performed by 3 independent researchers (NO, KH, CAA) based on the inclusion and exclusion criteria described. Conflicts were resolved by consensus at each stage of the screening process. Intercoder agreement was measured by calculating the Cohen kappa at each screening phase. The quality of included publications was not assessed.

Data Extraction and Analysis

Studies that met the inclusion criteria were extracted by the first independent researcher (KH) using a standardized Microsoft Excel spreadsheet and reviewed by 2 other independent researchers (NO, ALN) to ensure data quality and consistency. Data extracted included author names, year of publication, study design, setting, population, outcome measures, and main findings.

Analysis Framework

Identified cybersecurity interventions were mapped against the Essentials of Cybersecurity in Health Care Organizations (ECHO) framework, a guide for policymakers and health and care organizations to strengthen their cybersecurity infrastructure, which positions cybersecurity interventions across 6 dimensions including context, the wider conditions within which organizations' cyber planning operates; governance, which includes the policies and protocols to reduce the threat of cyberattacks; organizational strategy, involving policies, planning, and the allocation of responsibility at the organizational level; risk management, which involves identifying, assessing, and mitigating cyber risks; awareness, education, and training, which are actions to ensure that all

XSL•FO
RenderX

stakeholders within the organization have at least core knowledge on cybersecurity; and technical capabilities, or technical requirements needed to safeguard cybersecurity [21,22]. The ECHO framework was developed in 2020 as a health sector–specific framework for ICT professionals and nonexpert stakeholders. The was recently included as a recommended resource in the World Bank's knowledge notes series, entitled "Implementation Know-how Briefs to Support Countries to Prioritize, Connect and Scale for a Digital-in-Health Future (Cybersecurity in health brief)" [23].

A narrative synthesis of the findings was conducted. Relevant findings and outcome measures were grouped into subcategories and organized based on the 6 dimensions of the ECHO framework. Due to heterogeneity of the populations, interventions, comparators, and outcomes across the included studies, a meta-analysis was not performed, and no attempt was made to compare cybersecurity interventions described.

Results

Studies Included in the Analysis

The database search identified a total of 3134 publications. After removal of 1294 duplicates and 38 publications marked as ineligible by automation tools (eg, Covidence automatic deduplication function), 1802 papers underwent screening. Of these, 164 studies underwent full-text screening, and 20 eventually were finally included in the review (Figure 1). Cohen kappa values were 0.29 in the abstract screening and 0.64 in the full-text review.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. LMIC: low- and middle-income country.

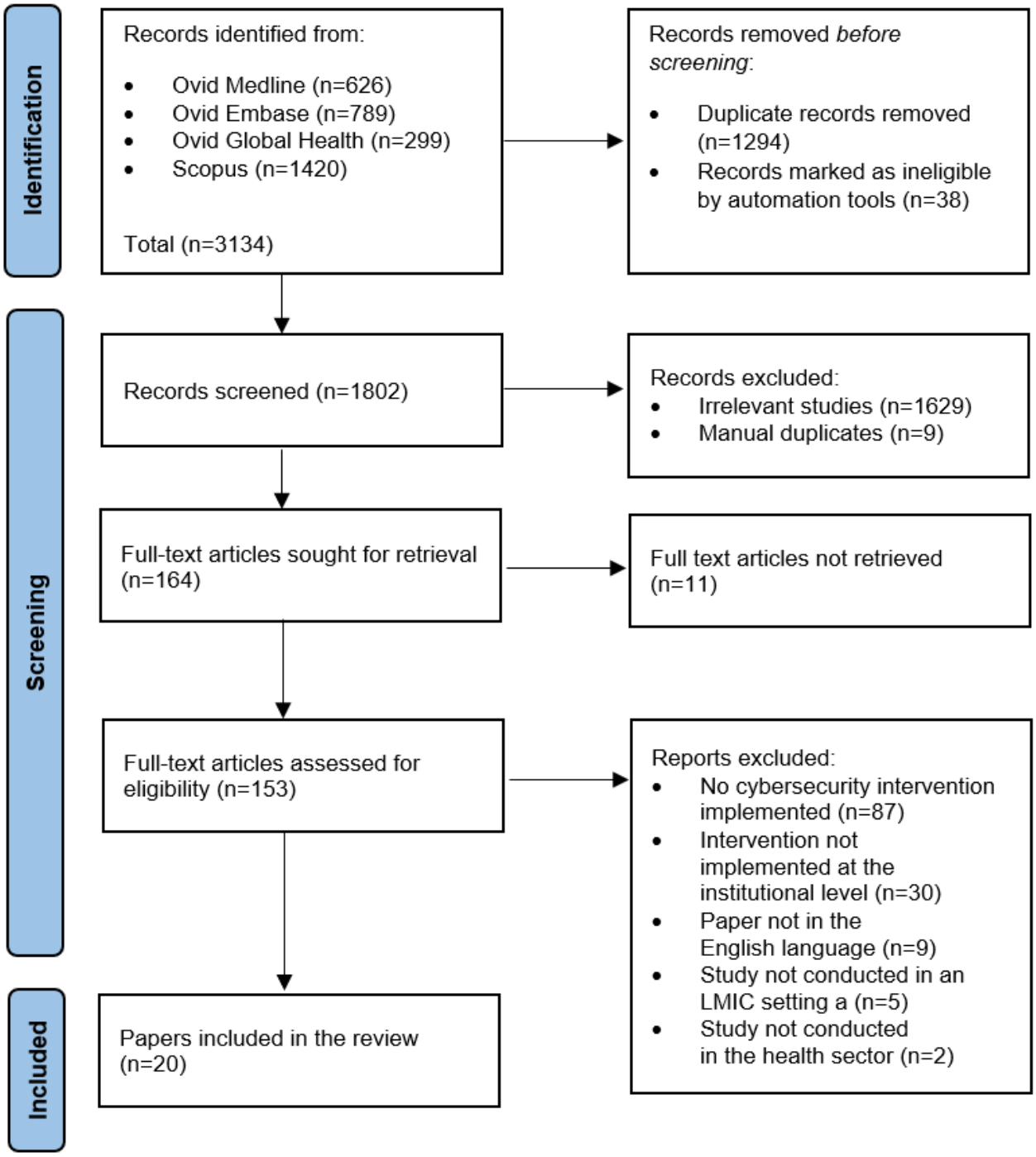

Description of Included Studies

Table 1 presents the country and economic classification status of each of the papers. Notably 1 research study was conducted in both Libya and Yemen, so this was counted as 2 studies for this analysis, resulting in 21 total studies. The countries most represented in the studies included were Peru (3/21, 14%) and South Africa (3/21, 14%), followed by China, Iran, and Indonesia (2/21, 10% each). All other countries included (Brazil, Ecuador, Ghana, Indonesia, Jordan, Libya, Nepal, Turkey, Vietnam, Yemen) were represented in only 1 study each (1/21, 5%). Two-thirds (14/21, 67%) of the studies were conducted in upper middle-income countries, and 6 (6/21, 29%) were conducted in lower-middle income countries, with 1 (5%) conducted in a low-income country setting. One study was conducted in 2 countries (Libya and Yemen). The sample size (ie, number of institutions involved in the study) ranged from 1 facility to 312 facilities: 7 (7/20, 35%) studies were based in a single facility, 6 studies (6/20, 30%) were conducted in 2 facilities to 5 facilities, and 7 (7/20, 35%) studies were conducted in 6 or more facilities.

XSL•FO
RenderX

Table 1. Description of studies by country, World Bank economic classification, and number of facilities.

Characteristic	Studies, n (%)
Country	
Peru	3 (14)
South Africa	3 (14)
China	2 (10)
Indonesia	2 (10)
Iran	2 (10)
Brazil	1 (5)
Ecuador	1 (5)
Ghana	1 (5)
Jordan	1 (5)
Libya	1 (5)
Nepal	1 (5)
Turkey	1 (5)
Vietnam	1 (5)
Yemen	1 (5)
Economic classification[a]	
UMIC[b]	14 (67)
LMIC[c]	6 (29)
LIC[d]	1 (5)
Number of facilities	
1	7 (35)
2-5	6 (30)
≥6	7 (35)

[a]Country status based on the World Bank economic classification [19].

[b]UMIC: upper middle-income country.

[c]LMIC: low- and middle-income country.

[d]LIC: low-income country.

Description of Cybersecurity Interventions Implemented in LMICs

Of the 20 studies, only 6 (30%) discussed the intervention(s) as the primary focus of the paper. Information on intervention(s) implemented had to be deduced from the remaining papers through analysis of responses to staff or hospital-level surveys (12/20, 60%) and case studies (1/20, 5%). The gray literature search undertaken with the same parameters did not find any additional papers to include. Detailed information of the included studies, including a description of cybersecurity intervention(s), is presented in Table 2. Technical capabilities (14/55, 25%) and awareness, education, and training (11/55, 20%) were the 2 most common interventions mentioned in the studies, followed by governance (9/55, 16%), organizational strategy (9/55, 16%), risk management (7/55, 13%), and context (5/55, 9%). Table 3 presents the description of studies by ECHO framework dimension.

XSL•FO
RenderX

Table 2. Detailed information (authors, year, country, institution, intervention) of the included studies.

Authors	Year	Country(s)	Institution name	Description of intervention
Ali Alferjanya et al [24]	2022	Libya, Yemen	Not stated; 5 private hospitals in Libya and the 5 biggest private hospitals in Yemen	No direct discussion of intervention; some of the organizational staff survey reported attendance of ICT[a] security training. Similarly, some staff reported the use of "different passwords across multiple web portals, systems, or applications" suggesting organizational technology-based cybersecurity measures, such as access control.
Ayatollahi and Shagerdi [25]	2017	Iran	Not stated; 27 hospitals located in 1 city in northwest Iran	Organizational staff reported on information security safeguards, including technical, administrative, and physical safeguards. The most common technical safeguards included access control and user authentication and detective control tests.
Chuma and Ngoepe [26]	2022	South Africa	Not stated; 1 public hospital in South Africa	Direct discussion of interventions; an organizational staff survey found the adoption of relevant legislation and the ISO27799 standard. The hospital established policies such as an administrative policy, records management policy, and medicine policy but had no formal security policy. Technical safeguards implemented at the hospital included password and username, data encryption, firewall and antivirus, and risk management activities such as maintaining a security audit log.
Daraghmeh and Brown [27]	2021	Jordan	Not stated; 39 private hospitals in Jordan	Organizational staff reported on various cybersecurity interventions. Some hospitals reported complying with relevant regulations and standards; having business continuity, disaster recovery, and risk management plans in place; and measuring the maturity of cyber measures. Some hospitals also reported requiring their staff to undergo training and various technological measures, including patching, firewalls, and authentication.
Dönmez et al [28]	2020	Turkey	Not stated; 15 public health institutions in Kocaeli, Marmara region	No direct discussion of intervention; organizational staff survey found attendance at ICT security training. Similarly, some staff reported "access to the information security policy documents at the hospital," suggesting organizational planning and strategy; "Personal information is stored via information technologies securely"; and password management systems, suggesting technical controls in some institutions.
Hou et al [29]	2018	China	Not stated; 1 public hospital in Shaanxi province, China	No direct discussion of intervention; an in-depth case study presented the hospital's information security management measures, including regulatory and standards adherence, risk management activities and procedures, staff information security awareness training, technical controls, and organizational security strategies.
Jara et al [30]	2021	Peru	Not stated; 1 private hospital in Lima, Peru	Implementation of a model of cybersecurity capability measurement, calculated by the level of compliance with controls
Kandabongee Yeng et al [31]	2023	Ghana	Not stated; 2 hospitals in Ghana	No direct discussion of intervention; staff "shared their knowledge of security measures related to password use, access control, vulnerability reporting and logging of users' access" as part of the research methodology, suggesting these interventions were in place at the 2 hospitals. Additionally, ICT and administration staff shared their knowledge on security governance, virus control, data backup, and training. Hospital A also reported having a draft security policy.
Khac Hai et al [32]	2017	Vietnam	Not stated; 312 HIV outpatient clinics in Vietnam	No direct discussion of intervention; some of the organizational staff survey reported attendance at ICT security training.
Maeko and Van Der Haar [33]	2018	South Africa	Charlotte Maxeke Academic Hospital, Johannesburg, South Africa	No direct discussion of intervention; some staff reported adhering "to the control procedures of locking or logging off their computers," suggesting organizational technology-based cybersecurity measures, such as access control, and some reported attendance at ICT security training.
Meng et al [34]	2018	China	1 anonymous organization in China implemented the approach	Implementation of a trust-based approach based on Bayesian inference to identify malicious devices in a health care environment
Nistrina and Bin Bon [35]	2019	Indonesia	Soreang Hospital, Bandung, Indonesia	Undertaking of a maturity assessment based on the COBIT[b] 5 capability levels

XSL•FO
RenderX

Authors	Year	Country(s)	Institution name	Description of intervention
Quimiz-Moreira et al [36]	2022	Ecuador	Verdi Cevallos Balda Hospital, IESS[c] Hospital, and Rodríguez Zambrano Hospital, Ecuador	No direct discussion of intervention; some of the hospitals reported on the use of 14 cybersecurity intervention areas including information security policies, roles and responsibilities, organization of information security, safety linked to human resources, asset management, access control, management of patient data (eg, pseudonymization), physical security of ICT equipment, operations security (eg, back-up, event logging), communications security (eg, network security management), acquisition, development and maintenance of information systems, supplier and third-party policies, incident management planning, business continuity, and compliance with security standards, policies, laws, and regulations.
Rijo et al [37]	2020	Brazil	Not stated; 11 hospitals in Brazil	No direct discussion of intervention; some of the hospitals surveyed reported on the use of 12 aspects of cybersecurity: IT security (eg, installing security patches), interoperability, hardware infrastructure, network infrastructure, business process management, use of standards, use of ISO[d] standards and certifications, research and development activities, IT team profile, alignment with strategy, decision-making and procurement, IT investment.
Serrano Rojas et al [38]	2022	Peru	Not stated; 1 health center in Lima, Peru	Implementation of a cybersecurity maturity model for health data privacy and protection
Sanchez Rubio et al [39]	2022	Peru	Not stated; 2 clinics in Lima, Peru	Implementation of a model to identify the level of maturity of the health clinics for preventing data leak scenarios
Sarbaz et al [40]	2019	Iran	Not stated; 12 academic hospitals in Mashhad, northeastern Iran	No direct discussion of intervention; some of the organizational staff survey reported the provision of information security training for staff and use of access control methods within their hospitals. Similarly, some staff reported "My organization runs security controls (such as firewall, antivirus, encryption, etc.) to protect sensitive information" and "My organization has used security procedures to protect important information from being stolen by malware (such as decoder, trojans, and spyware)," suggesting organizational technology-based cybersecurity measures and organizational planning.
Kencana Sari et al [41]	2023	Indonesia	National Cardiovascular Center Harapan Kita (RSPJNHK), a specialized public hospital	No direct discussion of intervention; implementation of interventions was mapped against ISO 27799:2016, and staff surveys were undertaken. The listed interventions included HIS[e] implementation policy, including information security management, set up of a System Maintenance and Security Sub-unit in the IT Department, establishment of procedures for incident reporting, ICT security training, access controls, establishment of a safe area for the data center room, and various technological measures, including antivirus, firewalls, and authentication.
Singh and Singh [42]	2022	South Africa	Not stated; 2 hospitals in northern Kwa-Zulu Natal, South Africa	No direct discussion of intervention; some of the organizational staff survey reported attendance at ICT security training. Most reported this training taking place at the hospital, though others reported training at school or a university or college.
Upadhyaya et al [43]	2018	Nepal	Tertiary hospital for Children, Eye, ENT[f] and Rehabilitation Services and associated community hospital in Chapagaun, Nepal	Implementation of a centralized, permissioned blockchain-based, secured health care data system using hyperledger alongside information security measures including training and adoption of quality and safety standards

[a]ICT: information and communication technology.

[b]COBIT: Control Objectives for Information and Related Technologies.

[c]IESS: Hospital de Especialidades Carlos Andrade Marín.

[d]ISO: International Organization for Standardization.

[e]HIS: health information system.

[f]ENT: ear, nose, throat.

XSL•FO
RenderX

Table 3. Mapping of the 20 included studies against the dimensions of the Essentials of Cybersecurity in Health Care Organizations (ECHO) framework.

ECHO dimension	Studies including the dimension, n (%)[a]	Papers
D1: Context	5 (9)	Daraghmeh and Brown [27], Jara et al [30], Nistrina and Bin Bon [35], Serrano Rojas et al [38], Sanchez Rubio et al [39]
D2: Governance	9 (16)	Chuma and Ngoepe [26], Daraghmeh and Brown [27], Hou et al [29], Kandabongee Yeng et al [31], Quimiz-Moreira et al [36], Rijo et al [37], Sarbaz et al [40], Kencana Sari et al [41], Upadhyaya et al [43]
D3: Organizational strategy	9 (16)	Chuma and Ngoepe [26], Daraghmeh and Brown [27], Dönmez et al [28], Hou et al [29], Kandabongee Yeng et al [31], Quimiz-Moreira et al [36], Rijo et al [37], Sarbaz et al [40], Kencana Sari et al [41]
D4: Risk management	7 (13)	Ayatollahi and Shagerdi [25], Chuma and Ngoepe [26], Daraghmeh and Brown [27], Hou et al [29], Quimiz-Moreira et al [36], Rijo et al [37], Sarbaz et al [40]
D5: Awareness, education, and training	11 (20)	Ali Alferjanya et al [24], Daraghmeh and Brown [27], Hou et al [29], Kandabongee Yeng et al [31], Khac Hai et al [32], Maeko and Van Der Haar [33], Quimiz-Moreira et al [36], Sarbaz et al [40], Kencana Sari et al [41], Singh and Singh [42], Upadhyaya et al [43]
D6: Technical capabilities	14 (25)	Ali Alferjanya et al [24], Ayatollahi and Shagerdi [25], Chuma and Ngoepe [26], Daraghmeh and Brown [27], Dönmez et al [28], Hou et al [29], Kandabongee Yeng et al [31], Maeko and Van Der Haar [33], Meng et al [34], Quimiz-Moreira et al [36], Rijo et al [37], Sarbaz et al [40], Kencana Sari et al [41], Upadhyaya et al [43]

[a]Some studies mentioned multiple interventions that corresponded to different dimensions of the ECHO framework, resulting in an N of 55.

Impact on Interventions of the Likelihood and Impact of Attacks, Clinical Outcomes, or the Quality and Safety of Care

Of the 6 studies that described the implementation of interventions, 5 did not report on outcomes, either related to impact on frequency or scale of cyberattacks, clinical outcomes, or the quality and safety of care; therefore, the impact remains unknown [26,30,34,35,38]. One study, however [43], reported on the impact of a blockchain-based, secured health care data system using hyperledger alongside information security measures (the intervention) on patient satisfaction and noted that the health care organization "did not undergo repetitive work like history taking, investigations etc.," suggesting that administrative waste was decreased, a key component of efficiency as described by the Institute of Medicine Six Domains of Health Quality Care [44]. There was no attempt to report on the impact and outcomes following the interventions in the remaining 14 papers from which interventions had to be deduced.

Main Barriers and Facilitators for the Implementation of Cybersecurity Interventions

The main facilitators and barriers identified are presented in Table 4. Facilitators to implementation were directly reported in only 1 study [43]. Upadhyaya et al [43] noted success factors (facilitators) in the implementation of a blockchain-based, secured health care data system, which included ease of use, improved security and reliability, and training to develop knowledge on its use to enhance diagnosis and treatment.

Table 4. Main facilitators and barriers to implementation identified from the 20 included studies.

Level	Facilitators	Barriers
National or regional	• Clear national legislation and policies (supported in organizations by active monitoring, evaluation, and learning)	• Lack of national or regional policies and guidance for organizations • Threat of natural disasters
Organizational	• Staff engagement and training • Management support	• Lack of knowledge among staff • Lack of ICT[a] experts • Exclusion of top-level managers • Cost of implementation • Political or high-level influence
Individual	• Perception of the importance of cybersecurity • Intervention's ease of use • Security and reliability of intervention	• Disregard of cultural differences • Lack of time

[a]ICT: information and communication technology.

Facilitators to implementation of cybersecurity interventions could be deduced from the discussion in 11 studies but were not directly reported [24,26,28-34,41,42]. At an organizational level, one of the main facilitators found in 4 studies was that national legislation and policies enabled health care organizations to create their own information security

XSL•FO
RenderX

management guidelines and subsequent interventions [26,29,33,34]. However, national legislation and policy was most effective when it was supported by active monitoring and penalization, for example with hospital inspections [26,29]. Jara et al [30] suggested that routine collection of data on information security by organizations enabled better cybersecurity practices due to a continuous process of evaluation and improvement. In addition, the most important factors for positive staff engagement with cybersecurity software are reportedly the appearance features, screen interface, and volume of information on the screen [42]. At an individual level, one of the main facilitators identified by 4 studies was that a perception of the importance of cybersecurity positively influenced cybersecurity intervention adherence [28,31-33]. This perception was influenced by different factors; for instance, Dönmez et al [28] suggested that health care workers recognized the need for improved cybersecurity, while Maeko and Van Der Haar [33] suggested that health care workers recognized the importance of safeguarding their patients. Studies recognized that a facilitator for successful cybersecurity training was when employees created their own cybersecurity goals and action plans [24,29]. Ali Alferjanya et al [24] expanded on this point by suggesting that this was facilitated by the inclusion of staff supervisors to aid staff with creating their goals.

Barriers to implementation were directly reported in 2 studies [25,41]. Maeko and Van Der Haar [33] noted barriers associated with awareness, education, and training and technical capability–focused interventions. Cost was perceived by 49% of respondents as a major barrier to training being offered at the hospital, and implementation barriers associated with multimodal access control systems included cost; political influence, as high-level decision-making can result in employee resistance; and existing cyber hygiene practices. Kencana Sari et al [41] noted that barriers associated with health care worker adoption of security behaviors included perceived downsides (eg, multifactor authentication taking more time) and workload constraints. Barriers to implementation of cybersecurity interventions could be deduced from the discussion in 9 studies [24-26,29,32-34,40,42]. One of the main barriers to implementing interventions was the lack of knowledge among all levels of health care staff [24,32-34,40,42]. Hou et al [29] noted that a barrier to effective training is that national policies do not outline how information security awareness training should be carried out. Meng et al [34] also mentioned that a lack of IT experts within health care organizations hindered cybersecurity practice. With regards to the successful implementation of cybersecurity training, disregard for cultural differences, a lack of time among health care staff, and exclusion of top-level managers were identified as major barriers [24,29,33]. The threat of natural and man-made disasters such as floods and fires, specifically electrical fires that impact electric power transmission, were found to be additional barriers [25,26]. Finally, Chuma and Ngoepe [26] identified a lack of resources such as power and system failures, poor network, and outdated systems as key challenges.

Discussion

Main Findings

The primary outcome of this scoping review was the description of cybersecurity interventions implemented in the LMIC health care organizations described in the 20 studies. Our results indicate that there is limited research focusing directly on identifying and evaluating cybersecurity interventions in our research context. Only 6 of the 20 studies focused directly on cybersecurity interventions [26,30,34,35,38,43]. The remaining 14 studies included an indirect discussion by using surveys or questionnaires to investigate the state of cybersecurity in health care organizations [24,25,27-29,31-33,36,37,39-42].

From the included studies, which directly investigated cybersecurity interventions [26,30,34,35,38,43], 1 study proposed a way of calculating devices' trust values and identifying trusted devices by means of a Bayesian inference approach, and another implemented a blockchain-based, secured health care system to transfer data. Five studies focused on ECHO Dimension 1, specifically investigating cybersecurity capability and maturity, implementing models and frameworks to assess security capabilities [27,30,35,38,39], with 1 study using an existing framework called Control Objectives for Information and Related Technologies (COBIT) 5 [35].The studies that indirectly discussed cybersecurity interventions through surveys or questionnaires most frequently described cybersecurity staff training (ECHO Dimension 5) as an intervention, followed by access control, including use of a password and user authentication (ECHO Dimension 6). Only 1 study reported on impact and outcomes of the interventions, specifically on patient satisfaction.

Despite the high frequency of staff training interventions being described, one of the main barriers mentioned was a lack of cybersecurity knowledge among health care staff [24,32-34,40,42]. This may indicate that existing ICT training is either sporadic or not effective in educating its staff members on cybersecurity and their roles and responsibilities. In this context, it is important to note that evidence is contradictory in what concerns the effectiveness of ICT training [44-46]. One of the main facilitators of training engagement was when staff were aware of the value of training and its importance to their patients [28,32,33]. This could suggest that the problem lies within the health care culture, rather than the training itself. We recommend that interventions should hold a larger focus on changing the culture around the importance of cybersecurity. Given that cybersecurity training is a nontechnical intervention that most staff will be exposed to, it is also likely that this was overreported in staff surveys.

National and regional policy and legislation play an important role in creating a setting for cybersecurity interventions to be implemented [26,29,33,34]. The facilitators of cybersecurity interventions mentioned in the studies often suggested that national legislation on cybersecurity gives health care organizations a direction for the development and implementation of their own policies and interventions. Hou et al [29] specifically mentioned that one of the barriers to effective cybersecurity staff training is that national policy does not

XSL•FO
RenderX

outline how information security awareness training should be carried out. It is recommended that health care organizations in LMICs collaborate with government stakeholders not only in the development of national and regional policy but also in the implementation of cybersecurity interventions. For instance, in the United Kingdom, the NHS England cybersecurity team collaborates with the UK National Cyber Security Centre on interventions to outline what effective digital practice is [47]. LMICs can similarly build on their national policies and collaborate with governments to develop effective cybersecurity interventions.

There was a spread of World Bank economic classifications in the country settings across the included studies. However, upper middle-income countries were most represented, as it is likely that they generally have greater access to resources and technology, as compared with LMIC contexts. Furthermore, 5 of the 20 studies were conducted in private hospitals [24,27,30,37,39], with Daraghmeh and Brown [27] performing a study of 39 private hospitals. These settings are not fully representative of the low-resource contexts in which many cybersecurity interventions must be implemented in public, non-for-profit, and faith-based facilities. More studies need to be conducted in low-resource settings that directly measure the impact of cybersecurity interventions.

Comparison With the Previous Literature

Findings of this scoping review are consistent with previous evidence suggesting that health care organizations globally are scaling up their responses to cyberattacks, particularly in the post-COVID-19 pandemic context [48,49]. However, it is notable that the limited evidence focused on investigation within LMIC health systems makes an in-depth comparison with previous literature in LMICs challenging. In the global context, including high-income country (HIC) health systems, He et al [48] also found that interventions used in the health sector during COVID-19 included increasing security awareness, enabling business continuity, and applying technical controls.

As Table 4 indicates, there were a range of facilitators and barriers to the implementation of cybersecurity interventions in LMIC settings reported or deduced from the evidence base. These findings are consistent with research undertaken in HICs. Coventry et al [50] found that barriers to secure behavior in health care among staff at 3 health care sites in Ireland, Italy, and Greece were a result of a lack of policies and reinforcement of secure behavior, poor awareness of consequences, and security as a barrier to productivity or patient care. Branley-Bell et al [51] also noted time pressures and fatigue as barriers to secure behavior in the health care context. Financial barriers to implementation of cybersecurity interventions identified in this study were also echoed in findings from interviews with cybersecurity experts working in hospitals in Canada and the United States [52]. Notably, a lack of comparable research on facilitators suggests more research is required across HIC and LMIC contexts.

Given the nascent literature on the impact of interventions on likelihood and impact of attacks, clinical outcomes, or the quality and safety of care in the wider literature, analysis of the 1 study outlining an impact on patient satisfaction and efficiency [43] remains limited. It is hoped that the evidence gap will be addressed given the increasing incidences of cyberattacks in health care and the growing use of digital technologies in health service delivery, making cybersecurity an essential element of patient safety in the health sector context.

Strengths and Limitations

This scoping review has several strengths [53]. The review represents the first known attempt to gather information systematically on cybersecurity interventions in health care organizations in LMICs. The broad nature of the scoping review methodology enabled the inclusion of a range of studies with varied aims and using varied methods. The search terms also contributed to the inclusion of 17 papers, enabled discussion on the topic of interest, and identified facilitators and barriers to the implementation of cybersecurity initiatives.

This review has several limitations. First, the scoping review was not registered in a database (eg, PROSPERO) because few results were expected to have detailed information to enable the application of synthesis methods to determine outcomes and effects. It is possible that some papers were missed in the search. However, we sought to mitigate this limitation by developing broad search terms, screening multiple databases both inside and outside medicine and health care, involving a team in the screening process, checking references in the included papers, and undertaking a search of the gray literature. Although care was taken to ensure that technical cybersecurity publications were captured by searching Scopus, some papers may have been missed as the Scopus catalog provides only 89% coverage of the ACM Digital Library, a common database of cybersecurity results. There was no assessment of bias in the included studies. Given our inclusion and exclusion criteria, it is likely that non-English studies published on this topic were missed. This limitation is particularly pertinent as a large number of LMICs are non-English speaking.

Implications for Research and Policy

This scoping review is the first of its kind and identifies a critical lack of investigation on cybersecurity interventions implemented in health care organizations. One of the main gaps in the literature is the direct evaluation of interventions, including the assessment of outcomes and impact, and studies focusing primarily on this as their main focus. As such, there is a need for robust evaluation of impact and outcomes. It is also important that further research is conducted across public, private, not-for-profit, and faith-based health providers in LMICs with a focus on primary data collection as opposed to further literature reviews.

A greater focus on interventions beyond cybersecurity training must be investigated. Expanding primary research methodologies beyond staff surveys and questionnaires also offers the opportunity to capture more information on technical interventions related to cybersecurity. Additionally, further research should evaluate the culture change around the importance of cybersecurity and implications for impact and outcomes. We believe greater research is required, using learnings from the field of implementation science and seeking

XSL•FO
RenderX

to build a robust and generalizable evidence base to inform implementation practice [54].

This review also identified the important role of national and regional policy and legislation as an enabler of cybersecurity intervention implementation. Clear national legislation and regulation, with accompanying guidance on cybersecurity, provides health care organizations with a mandate and direction to develop and implement organizational policies and interventions accordingly. It is also recommended that health care organizations in LMICs collaborate with government stakeholders in the implementation of cybersecurity interventions where possible, to enable government stakeholders to understand the practice-level challenges toward improved policy and guidance long term.

Conclusion

This scoping review presents a comprehensive description of cybersecurity interventions implemented in LMICs. The small number of studies identified highlights the limited body of research published in this topic area and shows large heterogeneity in interventions, research objectives, methods, and outcome measures used. Consistent with wider literature on the impacts of cybersecurity interventions, the impact of cybersecurity interventions on the likelihood and impact of cyberattacks in health care organizations remains unclear, making a reliable analysis of evidence difficult. Nonetheless, it is important to continue research to explore the impact, although complex and difficult to assess, to enable targeted cybersecurity initiatives to be used for the benefit of patients, providers, and health systems.

Current evidence points to clear national legislation and policies, supported in organizations by active monitoring, evaluation, and learning, and provider-level staff engagement and training as facilitators of cybersecurity interventions. Staff perception of the importance of cybersecurity, intervention ease of use, and security and reliability of interventions facilitate successful implementation. However, a lack of policy and guidance for organizations, a lack of knowledge among staff and health ICT experts, and the costs of implementation challenge greater implementation of cybersecurity initiatives in health care organizations. Future research should directly evaluate cybersecurity interventions and expand methodologies to build a robust and generalizable evidence base to inform policy and practice.

Acknowledgments

ALN is funded by the National Institute for Health and Care (NIHR) Applied Research Collaboration North West London (ARC NWL) and NIHR NWL Patient Safety Research Collaboration (NWL PSRC), with infrastructure support from the NIHR Imperial Biomedical Research Centre. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

KH, NO, and SG conceptualized the manuscript. KH developed search terms, with input from NO, and ran the database searches. KH, NO, and CAA undertook the systematic review screening, and ALN completed the quality assessment. KH, NO, and MP wrote the original draft with support from ALN and SG. All authors equally contributed to the reviewing and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms.
[DOCX File , 32 KB - jmir_v26i1e47311_app1.docx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 105 KB - jmir_v26i1e47311_app2.pdf]

References

1. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. Clin Chem Lab Med 2019 Feb 25;57(3):328-335 [FREE Full text] [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
2. Coventry L, Branley D. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas 2018 Jul;113:48-52 [FREE Full text] [doi: 10.1016/j.maturitas.2018.04.008] [Medline: 29903648]

XSL•FO
RenderX

3. Cyber Threats 2020: A Year in Retrospect. PWC. URL: https://www.pwc.co.uk/cyber-security/pdf/pwc-cyber-threats-2020-a-year-in-retrospect.pdf [accessed 2024-09-28]
4. Attack. National Institute of Standards and Technology (NIST) Information Technology Laboratory Computer Security Resource Center. URL: https://csrc.nist.gov/glossary/term/attack [accessed 2024-09-28]
5. Martin G, Martin P, Hankin C, Darzi A, Kinross J. Cybersecurity and healthcare: how safe are we? BMJ 2017 Jul 06;358:j3179 [FREE Full text] [doi: 10.1136/bmj.j3179] [Medline: 28684400]
6. Neprash HT, McGlave CC, Cross DA, Virnig BA, Puskarich MA, Huling JD, et al. Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016-2021. JAMA Health Forum 2022 Dec 02;3(12):e224873 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.4873] [Medline: 36580326]
7. Javaid M, Haleem A, Singh RP, Suman R. Towards insighting cybersecurity for healthcare domains: A comprehensive review of recent practices and trends. Cyber Security and Applications 2023 Dec;1:100016. [doi: 10.1016/j.csa.2023.100016]
8. Investigation: WannaCry cyber attack and the NHS. National Audit Office. 2018. URL: https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf [accessed 2024-09-28]
9. Corera G. Irish health cyber-attack could have been even worse, report says. BBC. 2021 Dec 10. URL: https://www.bbc.co.uk/news/technology-59612917 [accessed 2024-09-28]
10. Devi S. Cyber-attacks on health-care systems. The Lancet Oncology 2023 Apr;24(4):e148. [doi: 10.1016/s1470-2045(23)00119-5]
11. O'Brien N, Ghafur S, Durkin M. Cybersecurity in health is an urgent patient safety concern: We can learn from existing patient safety improvement strategies to address it. Journal of Patient Safety and Risk Management 2021 Mar 03;26(1):5-10 [FREE Full text] [doi: 10.1177/2516043520975926]
12. Ghafur S, Kristensen S, Honeyford K, Martin G, Darzi A, Aylin P. A retrospective impact analysis of the WannaCry cyberattack on the NHS. NPJ Digit Med 2019;2:98 [FREE Full text] [doi: 10.1038/s41746-019-0161-6] [Medline: 31602404]
13. Jones S, Moulton C, Swift S, Molyneux P, Black S, Mason N, et al. Association between delays to patient admission from the emergency department and all-cause 30-day mortality. Emerg Med J 2022 Mar 18;39(3):168-173 [FREE Full text] [doi: 10.1136/emermed-2021-211572] [Medline: 35042695]
14. Herrod PJJ, Adiamah A, Boyd-Carson H, Daliya P, El-Sharkawy AM, Sarmah PB, WES-Pi Study Group on behalf of the East Midlands Surgical Academic Network (EMSAN), WES-Pi Study Group. Winter cancellations of elective surgical procedures in the UK: a questionnaire survey of patients on the economic and psychological impact. BMJ Open 2019 Sep 13;9(9):e028753 [FREE Full text] [doi: 10.1136/bmjopen-2018-028753] [Medline: 31519672]
15. Holeman I, Cookson TP, Pagliari C. Digital technology for health sector governance in low and middle income countries: a scoping review. J Glob Health 2016 Dec;6(2):020408 [FREE Full text] [doi: 10.7189/jogh.06.020408] [Medline: 27648255]
16. Leesa-Nguansuk S. Warning over cyberattack threats. Bangkok Post. 2021 Sep 08. URL: https://www.bangkokpost.com/business/2178051/warning-over-cyberattack-threats [accessed 2024-09-28]
17. South Africa's Life Healthcare hit by cyber attack. Reuters. 2020 Jun 09. URL: https://www.reuters.com/article/us-life-healthcare-cyber-idUSKBN23G0MY [accessed 2024-09-28]
18. Moher D, Liberati A, Tetzlaff J, Altman DG, the PRISMA Group. Reprint—Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Physical Therapy 2009;89(9):873-880 [FREE Full text] [doi: 10.1093/ptj/89.9.873]
19. World Bank Country and Lending Groups. The World Bank. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.Accessed [accessed 2024-09-28]
20. Lencucha R, Neupane S. The use, misuse and overuse of the 'low-income and middle-income countries' category. BMJ Glob Health 2022 Jun 07;7(6):e009067 [FREE Full text] [doi: 10.1136/bmjgh-2022-009067] [Medline: 35672116]
21. O'Brien N, Grass E, Martin G, Durkin M, Darzi A, Ghafur S. Developing a globally applicable cybersecurity framework for healthcare: a Delphi consensus study. BMJ Innov 2020 Nov 23;7(1):199-207. [doi: 10.1136/bmjinnov-2020-000572]
22. O'Brien N, Martin G, Grass E, Durkin M, Ghafur S. Safeguarding our Healthcare Systems: A Global Framework for Cybersecurity. World Innovation Summit for Health. Doha, Qatar: World Innovation Summit for Health; 2020. URL: https://2020.wish.org.qa/app/uploads/2020/09/WISH-2020_Forum-Reports_Cyber-Security-and-Healthcare-Systems_ENG.pdf [accessed 2024-09-28]
23. Implementation Know-how Briefs to Support Countries to Prioritize, Connect and Scale for a Digital-in-Health Future. The World Bank. URL: https://openknowledge.worldbank.org/handle/10986/40215 [accessed 2024-09-28]
24. Ali Alferjanya MAO, Musaed Al-mwald MN, Alias RB. The effect of cyber security knowledge on employees' personal growth: An empirical study in private hospitals in libya and yemen. Health Education and Health Promotion 2022;10(2):369-375 [FREE Full text]
25. Ayatollahi H, Shagerdi G. Information security risk assessment in hospitals. Open Med Inform J 2017;11:37-43 [FREE Full text] [doi: 10.2174/1874431101711010037] [Medline: 29204226]
26. Chuma K, Ngoepe M. Security of electronic personal health information in a public hospital in South Africa. Information Security Journal: A Global Perspective 2021 Mar 01;31(2):179-195 [FREE Full text] [doi: 10.1080/19393555.2021.1893410]
27. Daraghmeh R, Brown R. A big data maturity model for electronic health records in hospitals. 2021 Presented at: International Conference on Information Technology (ICIT); July 14-15, 2021; Amman, Jordan. [doi: 10.1109/ICIT52682.2021.9491781]

XSL•FO
RenderX

28. Dönmez E, Kitapçı N, Kitapçı OC, Yay M, Aksu PK, Köksal L, et al. Readiness for health information technology is associated to information security in healthcare institutions. Acta Inform Med 2020 Dec;28(4):265-271 [FREE Full text] [doi: 10.5455/aim.2020.28.265-271] [Medline: 33627928]
29. Hou Y, Gao P, Nicholson B. Understanding organisational responses to regulative pressures in information security management: The case of a Chinese hospital. Technological Forecasting and Social Change 2018 Jan;126:64-75. [doi: 10.1016/j.techfore.2017.03.023]
30. Jara HLS, Navarro HBP, Armas-Aguirre J. Cybersecurity and Privacy Capabilities Model for Data Management Against Cyber-Attacks in the Health Sector. In: Iano Y, Saotome O, Kemper G, Mendes de Seixas AC, Gomes de Oliveira G, editors. Proceedings of the 6th Brazilian Technology Symposium (BTSym'20). BTSym 2020. Smart Innovation, Systems and Technologies, vol 233. Cham, Switzerland: Springer; 2021:359-367.
31. Kandabongee Yeng P, Yang BM, Stolt Pedersen M. Assessing cyber-security compliance level in paperless hospitals: An ethnographic approach. 2023 Presented at: 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS); November 29-December 1, 2022; Milan, Italy URL: https://ieeexplore.ieee.org/document/10061936 [doi: 10.1109/iotsms58070.2022.10061936]
32. Khac Hai N, Lawpoolsri S, Jittamala P, Thi Thu Huong P, Kaewkungwal J. Practices in security and confidentiality of HIV/AIDS patients' information: A national survey among staff at HIV outpatient clinics in Vietnam. PLoS One 2017;12(11):e0188160 [FREE Full text] [doi: 10.1371/journal.pone.0188160] [Medline: 29136017]
33. Maeko ME, Van Der Haar DT. A framework for user awareness and acceptance of smart card and fingerprint-based access control to medical information systems in South Africa. 2018 Presented at: IST-Africa Week Conference (IST-Africa); May 9-11, 2018; Gaborone, Botswana URL: https://ieeexplore.ieee.org/document/8417340
34. Meng W, Choo K, Furnell S, Vasilakos A, Probst C. Towards bayesian-based trust management for insider attacks in healthcare software-defined networks. T-NSM 2018:761-773 [FREE Full text] [doi: 10.1109/tnsm.2018.2815280]
35. Nistrina K, Bin Bon HAT. Information security for hospital information system using COBIT 5 framework. 2019 Presented at: International Conference on Industrial Engineering and Operations Management; March 5-7, 2019; Bangkok, Thailand URL: http://ieomsociety.org/ieom2019/papers/767.pdf
36. Quimiz-Moreira M, Zambrano-Romero W, Moreira-Zambrano C, Mendoza-Zambrano M, Cedeño-Palma E. Cybersecurity mechanisms for information security in patients of public hospitals in Ecuador. In: Botto-Tobar MS, Gómez O, Rosero Miranda R, Díaz Cadena A, Montes León S, Luna-Encalada W, editors. Trends in Artificial Intelligence and Computer Engineering. ICAETT 2021. Lecture Notes in Networks and Systems, vol 407. Cham, Switzerland: Springer; 2022:211-224.
37. Rijo R, Martinho R, Aparecida Oliveira A, Alves D, Silveira Nogueira Reis Z, Santos-Pereira C, et al. Profiling IT security and interoperability in brazilian health organisations from a business perspective. International Journal of E-Health and Medical Communications (IJEHMC) 2020;11(2):96-114 [FREE Full text] [doi: 10.4018/IJEHMC.2020040106]
38. Serrano Rojas AJ, Paniura Valencia EF, Armas-Aguirre J, Madrid Molina JM. Cybersecurity maturity model for the protection and privacy of personal health data. 2022 Presented at: IEEE 2nd International Conference on Advanced Learning Technologies on Education & Research (ICALTER); November 16-19, 2022; Lima, Peru. [doi: 10.1109/ICALTER57193.2022.9964729]
39. Sanchez Rubio CJ, Villacorta GG, Choque JO, Armas-Aguirre J. Personal health data: A security capabilities model to prevent data leakage in big data environments. 2022 Presented at: 17th Iberian Conference on Information Systems and Technologies (CISTI); June 22-25, 2022; Madrid, Spain. [doi: 10.23919/CISTI54924.2022.9820432]
40. Sarbaz M, Manouchehri Monazah F, Banaye Yazdipour A, Kimiafar K. Views of health information management staff on non-technical security management factors, Mashhad, Iran. Stud Health Technol Inform 2019;258:65-69. [Medline: 30942716]
41. Kencana Sari P, Wuri Handayani P, Nizar Hidayanto A, Wiranda Busro P. How information security management systems influence the healthcare professionals' security behavior in a public hospital in Indonesia. IJIKM 2023;18:583-607 [FREE Full text] [doi: 10.28945/5185]
42. Singh I, Singh Y. Cyber-security knowledge and practice of nurses in private hospitals in northern Durban, Kwazulu-Natal. Journal of Theoretical and Applied Information Technology 2022;100(1):246-247 [FREE Full text]
43. Upadhyaya P, Upadhyay SK, Subedi B, Subedi B, Gaire A. Revolutionizing healthcare systems of a developing country using blockchain. 2019 Presented at: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC); December 13-15, 2018; Madurai, India URL: https://ieeexplore.ieee.org/document/8782417 [doi: 10.1109/ICCIC.2018.8782417]
44. Alhuwail D, Al-Jafar E, Abdulsalam Y, AlDuaij S. Information security awareness and behaviors of health care professionals at public health care facilities. Appl Clin Inform 2021 Aug 29;12(4):924-932 [FREE Full text] [doi: 10.1055/s-0041-1735527] [Medline: 34587638]
45. Nifakos S, Chandramouli K, Nikolaou C, Papachristou P, Koch S, Panaousis E, et al. Influence of human factors on cyber security within healthcare organisations: a systematic review. Sensors (Basel) 2021 Jul 28;21(15):A [FREE Full text] [doi: 10.3390/s21155119] [Medline: 34372354]

XSL•FO
RenderX

46. Arain MA, Tarraf R, Ahmad A. Assessing staff awareness and effectiveness of educational training on IT security and privacy in a large healthcare organization. J Multidiscip Healthc 2019;12:73-81 [FREE Full text] [doi: 10.2147/JMDH.S183275] [Medline: 30666123]
47. What Good Looks Like framework. NHS England. 2021 Oct 4. URL: https://transform.england.nhs.uk/digitise-connect-transform/what-good-looks-like/what-good-looks-like-publication/ [accessed 2024-09-28]
48. He Y, Aliyu A, Evans M, Luo C. Health care cybersecurity challenges and solutions under the climate of COVID-19: scoping review. J Med Internet Res 2021 Apr 20;23(4):e21747 [FREE Full text] [doi: 10.2196/21747] [Medline: 33764885]
49. Feeley A, Lee M, Crowley M, Feeley I, Roopnarinesingh R, Geraghty S, et al. Under viral attack: An orthopaedic response to challenges faced by regional referral centres during a national cyber-attack. Surgeon 2022 Oct;20(5):334-338. [doi: 10.1016/j.surge.2021.09.007] [Medline: 34782238]
50. Coventry L, Branley-Bell D, Sillence E, Magalini S, Mari P, Magkanaraki A, et al. Cyber-Risk in Healthcare: Exploring Facilitators and Barriers to Secure Behaviour. 2020 Presented at: HCI for Cybersecurity, Privacy and Trust: Second International Conference, HCI-CPT 2020, Held as Part of the 22nd HCI International Conference; July 19-24, 2020; Copenhagen, Denmark. [doi: 10.1007/978-3-030-50309-3_8]
51. Branley-Bell D, Coventry L, Sillence E. Promoting cybersecurity culture change in healthcare. Northumbria University Research Portal. 2021. URL: https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/54632089/PETRA_Branley_Bell_et_al_2021_.pdf [accessed 2024-09-28]
52. Wilner A, Luce H, Ouellet E, Williams O, Costa N. From public health to cyber hygiene: Cybersecurity and Canada's healthcare sector. International Journal 2022 Feb 21;76(4):522-543 [FREE Full text] [doi: 10.1177/00207020211067946]
53. Pham M, Rajić A, Greig J, Sargeant J, Papadopoulos A, McEwen S. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Res Synth Methods 2014 Dec;5(4):371-385 [FREE Full text] [doi: 10.1002/jrsm.1123] [Medline: 26052958]
54. Klaic M, Kapp S, Hudson P, Chapman W, Denehy L, Story D, et al. Implementability of healthcare interventions: an overview of reviews and development of a conceptual framework. Implement Sci 2022 Jan 27;17(1):10 [FREE Full text] [doi: 10.1186/s13012-021-01171-7] [Medline: 35086538]

Abbreviations

COBIT: Control Objectives for Information and Related Technologies
ECHO: Essentials of Cybersecurity in Health Care Organizations
HIC: high-income country
ICT: information and communication technology
LMIC: low- and middle-income country
NHS: National Health Service
NIST: National Institute of Standards and Technology
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
WHO: World Health Organization

Edited by T Leung, A Mavragani; submitted 15.03.23; peer-reviewed by Z Zandesh, J Dykstra, M Hölbl; comments to author 07.09.23; revised version received 01.11.23; accepted 07.09.24; published 20.11.24.

Please cite as:
Hasegawa K, O'Brien N, Prendergast M, Ajah CA, Neves AL, Ghafur S
Cybersecurity Interventions in Health Care Organizations in Low- and Middle-Income Countries: Scoping Review
J Med Internet Res 2024;26:e47311
URL: https://www.jmir.org/2024/1/e47311
doi:10.2196/47311
PMID:

©Kaede Hasegawa, Niki O'Brien, Mabel Prendergast, Chris Agape Ajah, Ana Luisa Neves, Saira Ghafur. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Uses and Experiences of Synchronous Communication Technology for Home-Dwelling Older Adults in a Home Care Services Context: Qualitative Systematic Review

Martin Vinther Bavngaard[1], MA; Anne Lund[1], PhD; Björg Thordardottir[1,2], PhD; Erik Børve Rasmussen[3], PhD

[1]Department of Rehabilitation Science and Health Technology, Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway
[2]Faculty of Occupational Therapy, School of Health, Business and Natural Sciences, University of Akureyri, Akureyri, Iceland
[3]Department of Social Work, Child Welfare and Social Policy, Faculty of Social Sciences, Oslo Metropolitan University, Oslo, Norway

Corresponding Author:
Martin Vinther Bavngaard, MA
Department of Rehabilitation Science and Health Technology
Faculty of Health Sciences
Oslo Metropolitan University
Postboks 4, St. Olavs plass
Oslo, 0130
Norway
Phone: 47 67235000
Email: martinvi@oslomet.no

Abstract

Background: European health care systems regard information and communication technology as a necessity in supporting future health care provision by community home care services to home-dwelling older adults. Communication technology enabling synchronous communication between 2 or more human actors at a distance constitutes a significant component of this ambition, but few reviews have synthesized research relating to this particular type of technology. As evaluations of information and communication technology in health care services favor measurements of effectiveness over the experiences and dynamics of putting these technologies into use, the nuances involved in technology implementation processes are often omitted.

Objective: This review aims to systematically identify and synthesize qualitative findings on the uses and experiences of synchronous communication technology for home-dwelling older adults in a home care services context.

Methods: The review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist for reporting. We conducted a cross-disciplinary search in 5 databases for papers published between 2011 and 2023 that yielded 4210 citations. A total of 13 studies were included after 4 screening phases and a subsequent appraisal of methodological quality guided by the Critical Appraisal Skills Programme tool. From these, prespecified data were extracted and incorporated in a 3-stage thematic synthesis producing 4 analytical themes.

Results: The first theme presented the multiple trajectories that older users' technology acceptance could take, namely straightforward, gradual, partial, and resistance laden, notwithstanding outright rejection. It also emphasized both instrumental and emotional efforts by the older adults' relatives in facilitating acceptance. Moving beyond acceptance, the second theme foregrounded the different types of work involved in attempts to integrate the technology by older users, their relatives, and health care providers. Theme 3 highlighted how the older users' physical and cognitive conditions formed a contextual backdrop challenging this integration work, together with challenges related to spatial context. Finally, consequences derived from taking the technology into use could be of a both enabling and complicating nature as integration reconfigured the way users related to themselves and each other.

Conclusions: The acceptance and integration of synchronous communication technology for older adults involves multiple user groups in work tending to the technology, to the users themselves, and to each other through intergroup negotiations. This review's original contribution consists of its attention to the dynamics across different user groups in deriving consequences from using the technology in question, in addition to its assertion that such consequences may be both intentional and unintentional. We argue that our findings may be used to provide nuance to policies addressing—and practices taking place in—contexts that involve similar user technology constellations to the ones explored in this paper.

Trial Registration: PROSPERO CRD42023414243; https://tinyurl.com/wrha6j3f

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59285) doi:10.2196/59285

KEYWORDS

systematic review; qualitative; thematic synthesis; communication technology; relatives; home care services; aging in place; home-dwelling; older adult; aging; gerontology; European; effectiveness; information; technology; health care provider; cross-disciplinary; telehealth; telemonitoring

Introduction

Background

In coming years, the health care needs of increasingly aging populations are projected to supersede the available resources of health care systems in Western societies [1]. Measures have been taken to meet this challenge, with health and social policies across the United Nations prioritizing older adults' ability to "age in place" [2]—a scenario wished for by most older adults [3]. In several countries, especially established welfare states, the home care services sector has emerged as a political focal point, being first in line to adapt their service provision to the increasing number of home-dwelling older adults [4,5]. Concurrently, Western health policy discourses addressing aging in place construe innovative digital solutions for the care of older adults as enablers per default [6]. Such technologies have been associated with a triple-win narrative: benefitting older adults, the national economy, and society overall [7]. As reflected in the 2020 UN Plan of Action [8], which addresses health and care services globally, emphases on bolstering community home care services coexist with the increased and persistent use of digital solutions. Indeed, research comparing the needs of European home care systems reveals a shared ambition across countries to integrate technology in the provision of home care services [9]. Moreover, existing research has foregrounded the potential held by technology for digital patient service communication in heightening the efficiency of services and enabling health care provision to older adults living in rural areas [10].

However, assessing the impacts of using information and communications technology (ICT) for aging in place requires a wide array of research methodologies and data [11]. Evaluating ICT using only parameters of effectiveness and efficiency runs the risk of presenting a polished but distorted picture, overlooking the dynamics and contextual situatedness of human practices involving technology [12,13]. Even randomized controlled trial designs, the purported "gold standard" of evidence-based health research, have been criticized for being unable to "tell us all we need to know where the intervention is in a complex, dynamic context such as homecare for older people" [14]. As contemporary health and care policies continually favor increased technologization, qualitative knowledge developed from those receiving, using, and experiencing such technology firsthand is of utmost importance [15]. However, this recognition of complex dynamics must also extend to technology. From an interaction design perspective, different technologies *afford* different interactions that influence the dynamics of the user technology interplay [16]. A glance toward the philosophical underpinnings of media theory serves as a further reminder to foreground the medium, that is, technology—through which human interactions transpire [17]. An expanding corpus of qualitative research investigating the integration of ICT by older adults attests to this: while personal alarms may conflict with older adults' self-identification as independent individuals [18], digital medicine dispensers may cause frustration by imposing rigid structures for medicine taking or by simply not going with the decor [19]. The need to remain attentive to technology types is underlined by Lindberg et al [20], whose systematic review reveals that the tendency to use varying terminologies to describe ICT in home care research is widespread. Indeed, attempts at systematically reviewing existing research on ICT for older adults in the context of home care have conceptualized ICT with varying degrees of specificity, from using umbrella terms, such as "eHealth" [21], "telemedicine" [22], "telecare" [12], "long-distance caregiving technology" [23], and "digital assistive technology" [24], to making more homogenous technology classifications, such as "active ICT" [25] and "apps" [26]. In light of this terminological multiplicity, this review pursues specificity by centering on synchronous communication technology (SCT).

Review Aim

This systematic review aimed to identify, review, and synthesize results from qualitative studies reporting on the uses and experiences of SCT for home-dwelling older adults in a home care services context. The review operates with the following research question: How is SCT for home-dwelling older adults in a home care services context used and experienced? We also used the following subquestions:

1. What are the types and purposes of SCT investigated in community home care services for home-dwelling older adults?
2. How is the SCT integrated by the users?
3. How is the SCT experienced by the users?

Methods

Search Strategy

A protocol detailing the study aim, inclusion criteria, method of extraction, and analysis was registered in PROSPERO before conducting the search [27]. Following preliminary searches used to select and assess the appropriateness of search terms, a systematic search was conducted by the main reviewer (MVB) on March 10, 2023, in the following databases: Web of Science, MEDLINE, PsycINFO, CINAHL, and ACM Digital Library. The search queries used the Boolean phrase search mode and consisted of 4 keyword search strings as well as database-specific index terms (see Table 1 for a detailed overview).

XSL•FO
RenderX

Table 1. Search query used in this review[a] (n=4210).

Database	Results, n (%)
Web of Science	704 (16.72)
MEDLINE	934 (22.19)
PsycINFO	865 (20.55)
CINAHL	568 (13.49)
ACM Digital Library	1139 (27.05)

[a]The following constitutes the 4 core search strings separated by semicolons: (ICT OR "information and communication* technolog*" OR telecare OR ehealth OR "welfare technolog*" OR "warm technolog*" OR "assistive technolog*" OR "care technolog*" OR "social technolog*" OR "communication technolog*" OR gerontechnolog* OR "digital technolog*" OR "digital health" OR virtual OR video OR tablet OR web* OR smartphone* OR "smart device" OR phone* OR mobile* OR computer OR app OR apps OR "mobile application" OR robot* OR platform); (old* OR "old* people" OR "old* adults" OR elder* OR aging OR aging OR senior* OR "later life" OR aged); ("home care" OR homecare OR "elder care" OR "eldercare" OR "home nurs*" OR "domestic care" OR "home based care" OR "home health care" OR "home healthcare" OR "homebased care" OR "community care" OR "community health care" OR "community healthcare"); (qualitative OR experienc* OR perception* OR perceiv* OR feel* OR attitude* OR adopt* OR thematic OR theme* OR ethnograph* OR interview* OR observation* OR phenomenolog* OR "case stud*" OR "focus group" OR narrative* OR "grounded theory").

Inclusion Criteria

Because of its frequent use in structuring the research questions of systematic reviews, the qualitative adaptation of the Participants, Interest, and Context model [28] was deemed a suitable tool to trace the review's area of interest.

Participants

This review considered studies involving users of SCT for older adults aged ≥65 years and who are living in their own home. Because of the nature of our phenomenon of interest (see the subsequent section), "users" may refer to participants other than the older adults themselves, such as informal caregivers or health care professionals. No inclusion criteria specified whether the older adults should be living alone or with other people (eg, relatives or spouse) nor if they should be living with or without a chronic condition.

Phenomenon of Interest

The review's phenomenon of interest was uses and experiences with SCT. Drawing on the typology of communication systems in health care by Coiera [29], we define SCT as *a discrete device that enables digital mediation of synchronous communication between 2 or more human actors*. This definition did not limit the technology to a specific purpose (eg, SCT for enhancing social connectedness), allowing us to include a variety of potential purposes while establishing a definitional coherence between the types of technology. As evident in Table 1, the delineation was not reflected in the search queries but instead functioned as a criterion against which potentially relevant studies were manually screened.

Context

Only studies investigating the uses and experiences of SCT in home care services were considered for inclusion. Other contexts, for example, hospitals and specialist health care, were excluded. Moreover, this review focused exclusively on original empirical research that was either fully qualitative or with a qualitative component. No other limitations were imposed regarding the study design. Only peer-reviewed publications written in English and published between January 1, 2011, and March 10, 2023, were considered for inclusion. By delimiting this particular period, the review intends to capture the current practices regarding technology in home care services—a setting that has seen an acceleration in the development of technological solutions during the last decade [30]. As such, synthesizing studies published before 2011 may result in lower transferability to current health care systems [31].

Data Processing

Screening and Inclusion

Using the aforementioned search strategy yielded 4210 citations across the 5 databases. Following the search, all identified citations were imported into the reference management software EndNote 20 [32] where 663 (15.74%) duplicates were removed. The remaining 3547 (84.25%) citations were uploaded to the web-based literature review tool Rayyan—Intelligent Systematic Review developed by Ouzzani et al [33] to facilitate the subsequent title and abstract screening process. This process was undertaken in 2 stages. First, 2344 (N=3547, 66.08%) citations that plainly deviated from the inclusion criteria were excluded by the main reviewer (MVB). Second, 3 reviewers (MVB, AL, and EBR) independently screened the title and abstract of the remaining 1203 (N=3547, 33.91%) citations using a screening tool while being blinded to each other's decisions. Pilot-testing the screening tool revealed that titles and abstracts rarely specified the technology investigated by the study, which made it difficult to accurately judge the technology's relevance to our area of interest. As a result, a screening tool operating with broader inclusion criteria regarding technology was developed in accordance with the guidelines of Polanin et al [34]. Following the blinded screening, all cases of conflict (259/1203, 21.52%) were resolved through group discussion while using the screening tool.

A total of 250 (20.78%) out of 1203 citations met the inclusion criteria and were retrieved in full-text form. The full-text assessment comprised 2 stages. Because of the impossibility of applying our inclusion criterion pertaining to SCT in the title and abstract screening process, the first stage involved assessing the full-text specifically against this criterion. This assessment was performed by the main reviewer (MVB) and resulted in the

XSL•FO
RenderX

exclusion of 193 (N=250, 77.2%) citations. In the second stage, the eligibility criteria were applied to the remaining studies (Textbox 1). Each study was read in full, independently by each member of the review team (MVB, AL, BT, and EBR) and assessed for inclusion. After reaching a consensus through subsequent discussion, a corpus of 18 (N=57, 31%) studies was deemed eligible for inclusion. The screening process was guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram [35], modified to reflect our stages of assessment, and is discussed in the Results section.

Textbox 1. Screening tool used for the full-text assessment.

Does the study...

- Investigate the use of "a discrete device that enables digital mediation of synchronous communication between two or more human actors?"
- Relate itself explicitly to older adults?
- Situate itself in a naturalistic home care context?
- Describe its methods as having a qualitative component?

Is the study...

- An original primary study (ie, not proceeding or book chapter)?
- Peer reviewed?
- In English?
- Published on or after January 1, 2011?

Appraisal of Methodological Quality

Because of the divergent traditions within qualitative research, an awareness of the ways in which studies construct knowledge was necessary [36]. Three reviewers (AL, BT, and EBR) each appraised one-third of the included studies, while MVB appraised all 18 studies using the expanded version of the Critical Appraisal Skills Programme tool [37] by Long et al [38]. As the parameters used to judge methodological quality may vary because of a multiplicity of factors not reflecting actual scientific rigor (eg, reporting style or journal preferences), the appraisal did not rank studies based on their quality nor was a risk of study bias assessment conducted. Consensus regarding all exclusions was reached through discussion. Multimedia Appendix 1 [39-51] presents an overview of the team's assessments.

Data Extraction

Data were extracted by the main reviewer (MVB) and cross-checked by coreviewers (AL, BT, and EBR) using a data extraction tool specifically designed for this review. This tool follows an inclusive approach to data extraction [52] by extracting both the "abstract" and "findings" sections as well as the study's bibliographical and methodological details for conveying its general characteristics in a tabular format.

Data Synthesis

To ensure transparency, the review was guided by the PRISMA 2020 checklist (Multimedia Appendix 2) for reporting [35]. However, abiding by the qualitative nature of the study aims means casting scientific studies as data themselves. Such studies not only have been developed within a specific social, cultural, and historical context but also present empirical data that have been selected, interpreted, and curated by the respective authors [53]. Therefore, attentiveness toward the constructed nature of "findings" in studies was crucial in establishing the point of departure for—and trajectory of—the synthesis. To preserve an awareness of the context-sensitive nature of each study, its full-text was imported into the qualitative data analysis software, NVivo 12 (Lumivero). However, only the findings were subjected to line-by-line coding.

The findings of the included studies were synthesized using thematic synthesis as described by Thomas and Harden [52]. By anchoring the analysis thematically, the method allows for the synthesization of data developed across different qualitative traditions. Derived from meta-ethnography, thematic synthesis involves a 3-stage process with each stage following an ascending order of analytic abstraction, from (1) coding the extracted data line-by-line to (2) the generation of descriptive and (3) analytic themes. Initial coding was performed independently by MVB and EBR, resulting in 45 and 129 codes, respectively. Although coding many of the same tendencies, MVB's 45 codes were more developed and thus closer to descriptive themes while EBR's 129 codes reflected each isolated instance of analytical interest. Incidentally, this allowed for the development of descriptive themes from EBR's codes while cross-checking these with MVB's codes—an inductive-deductive pendulation. Where codes did not fit with any descriptive theme, either a new theme was created or the code was deleted. Through this process, 25 descriptive themes were developed. MVB and EBR then proceeded to group the descriptive themes within themes of higher analytical abstraction. In this process, and informed by the account of theme development by Braun and Clarke [54], descriptive themes were dissolved and their codes were distributed into other themes where they proved to be a better fit for the story conveyed by the analytical themes. During this distribution, MVB and EBR checked each code's consistency with its new placement by revisiting its original context. The total amount of descriptive themes consequently decreased to 21, making up 4 analytical themes. Provided with a mapping of analytical and

XSL•FO
RenderX

descriptive themes, AL and BT read through all included studies to assess MVB's and EBR's synthesization. A subsequent discussion meeting was held in which the thematization and theme labels were discussed, refined, and agreed upon.

Ethical Considerations

As the data corpus consisted exclusively of research studies published in peer-reviewed scientific journals, and thus did not entail data collection involving human participants, ethics approval was not sought as part of conducting this review. The ethical integrity of the included studies was assessed as part of the methodological appraisal.

Results

Characteristics of the Included Studies

The screening process, illustrated using the PRISMA flow diagram in Figure 1, produced 18 studies eligible for inclusion. However, during the methodological appraisal, 2 (N=18, 11%) studies were deemed to be of insufficient quality for the review and 3 (16%) studies did not comply with the inclusion criteria. Hence, they were excluded (Multimedia Appendix 3). Following their removal, the final dataset consisted of 13 studies whose characteristics are outlined in Table 2. The included studies were undertaken in Australia, Canada, Denmark, Finland, Germany, Norway, Spain, Sweden, and the United States and involved a total of 270 participants. Older adults were key participants in all studies. In addition, the studies included informal caregivers; formal carers (nurses, occupational therapists, etc); assessors; interpreters; managers; and palliative care consultation workers. In total, 10 (N=13, 76%) studies collected data from different user groups, although 2 (15%) studies did not disclose the number of participants in each user group. Furthermore, 1 (7%) study specified the different user groups but omitted the number of participants from each. The following technologies were investigated by the studies: alarm pendant (5/13, 38%), tablet with videoconferencing software (3/13, 23%), laptop with videoconferencing software (2/13, 15%), picturephone (1/13, 7%), telephone (1/13, 7%), and video phone (1/13, 7%). The duration of use ranged from participants having used the technology once or twice to ≥12 months. The studies' designs were either qualitative (11/13, 84%) or mixed methods (2/13, 15%). However, the quality assessment revealed that details on the theoretical and methodological underpinnings were generally underreported (Multimedia Appendix 1). The data collection methods included qualitative interviews (12/13, 92%), observations (including video recordings; 6/13, 46%), document analysis (2/13, 15%), and focus groups (2/13, 15%). Seven (N=13, 53%) studies used more than one qualitative data collection method. For analyzing the data, studies used thematic analysis (5/13, 38%), qualitative content analysis (3/13, 23%), hermeneutic analysis (2/13, 15%), discourse analysis (1/13, 7%), or conversation analysis (1/13, 7%) or did not report a framework for data analysis (1/13, 7%). Seven (53%) of the 13 studies reported using an explicit theoretical framework to guide the data analysis.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart illustrating the process of identifying and screening studies for inclusion. QA: quality appraisal.

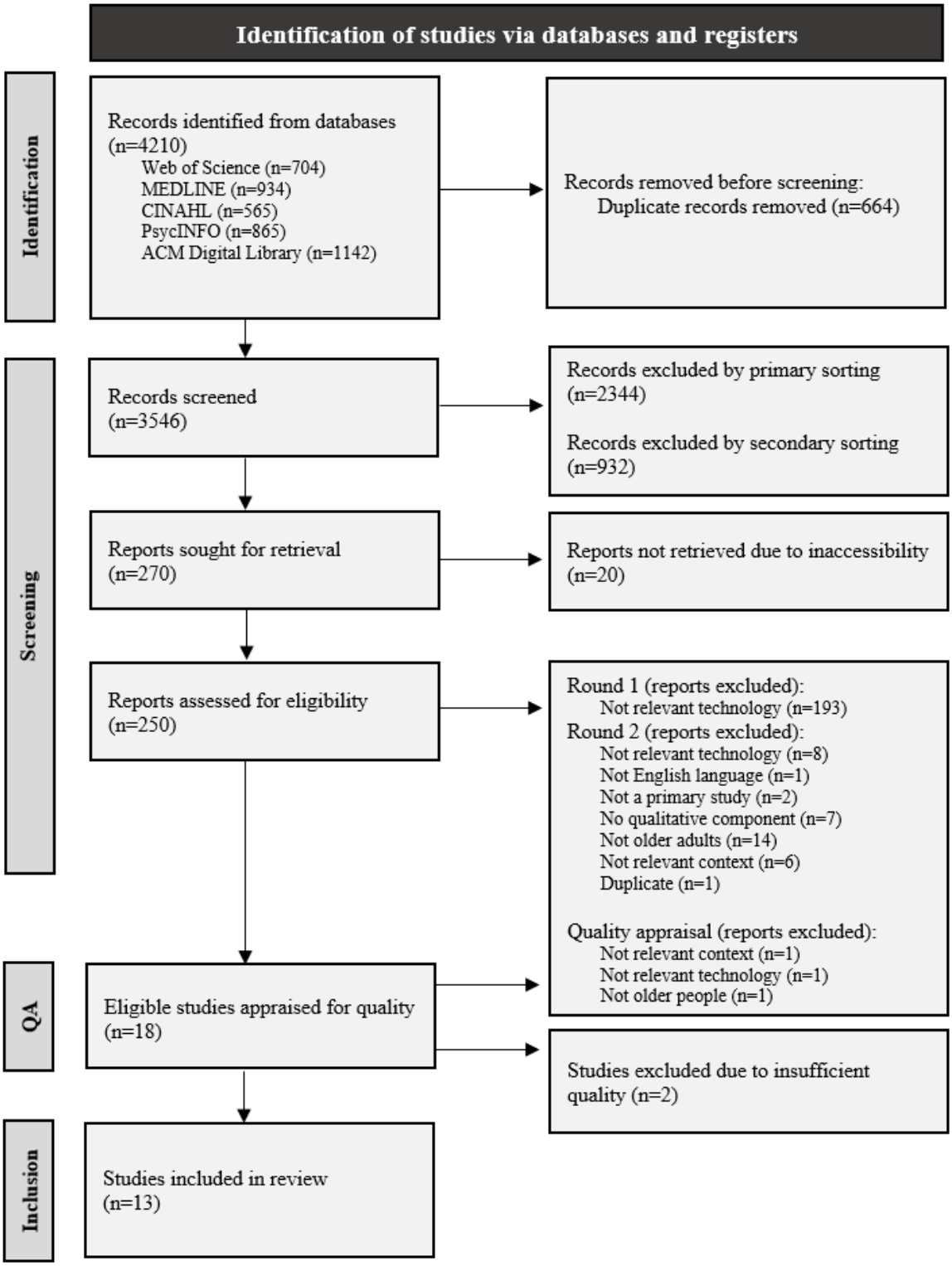

XSL•FO
RenderX

Table 2. General characteristics of included studies.

Study	Country of study	Aim of study	Context of study	Research design	Participants	Type of technology (discrete device)	Type of technology (device-in-network)	Duration of use	Data types	Data analysis methods	Theoretical framework
Correa and Domènech [39], 2013	Catalonia, Spain	Examine the role of technology in the familial care networks of older adults	Home-based telecare	Qualitative	57 (older adults, informal carers, and formal carers)	Alarm pendant	A monitoring system that involves a domestic terminal and pendant through which the user can communicate with a call center manned by teleoperators.	At least 12 months	Interviews and focus groups	Thematic analysis	Actor-Network Theory
Criado and Domènech [40], 2015	Spain	Explore the impact of home telecare services on the care relations for older adults	Home-based telecare	Qualitative	Not reported (older adults, relatives, and technicians)	Social alarm pendant	An alarm hub and social alarm pendant through which the user can communicate with a call center manned by teleoperators.	Not reported	Interviews and observations	Thematic analysis	Science and technology studies
Gilbert et al [41], 2021	Australia	Explore stakeholders' perspectives on using video remote interpreting for home-based cognitive assessments	Home-based cognitive assessment	Qualitative	25 (6 older adults, 9 assessors, and 10 interpreters)	Tablet with videoconferencing software	An iPad or Microsoft Surface Pro tablet placed in the homes of users, connecting them to interpreters via one of 2 videoconferencing platforms set up by agency support staff.	At least once	Interviews	Not reported	__[a]
Ilomäki and Ruusuvuori [42], 2022	Finland	Analyze how video-mediated, nurse-led guiding of medicine taking shapes client autonomy	Home-based telecare service pilot	Qualitative	7 (4 older adults and 3 nurses)	Tablet with videoconferencing software	A tablet with videoconferencing software placed in the home of the user connecting them to home care nurses stationed in a shared office.	At least once	Observations (video recordings)	Conversation analysis	Empirical ethics
Karlsen et al [43], 2017	Norway	Explore how older adults and their family caregivers experience the use of telecare services	Home-based telecare	Qualitative	24 (18 older adults and 7 relatives)	Personal alarm (and other)	A wrist- or neck-worn alarm facilitating 2-way communication, primarily used to connect users to home care services.	At least 5 months	Interviews	Hermeneutic analysis	—

XSL•FO
RenderX

Study	Country of study	Aim of study	Context of study	Research design	Participants	Type of technology (discrete device)	Type of technology (device-in-network)	Duration of use	Data types	Data analysis methods	Theoretical framework
Lee et al [44], 2021	United States	Examine the impact of a telephone reassurance program on older adults	Telephone-based social support	Mixed-methods	17 (older adults)	Telephone	Older adults are assigned older volunteers who initiate calls via a landline or mobile telephone.	12 weeks	Interviews	Thematic analysis	Transactional Model of Stress and Coping
Outila and Kiuru [45], 2021	Finland	Analyze older adults' use of the picturephone in a home care setting	Home-based telecare	Qualitative	8 (older adults)	Picturephone	A tablet placed in the home of the user, connecting them to family, friends, and health care services as well as providing audiovisual content.	At least once	Interviews and observations	Foucauldian discourse analysis	Actor-Network Theory and Foucauldian concepts
Paul et al [46], 2019	Canada	Understand the experience of using mobile web-based videoconferencing for home-based palliative care of older adults	Home-based palliative telehealth	Mixed-methods	30 (8 older adults, 9 relatives, 5 RPC-CT[b] members, and 8 nurses)	Laptop with videoconferencing software	A laptop with an external webcam and speakerphone placed in the home of the user with assistance from a nurse, connecting them to a palliative care physician consultant.	Not reported	Interviews and focus groups	Content analysis	—
Schmidt et al [47], 2019	Germany	Understand how the video-mediated monitoring process of care and case management is experienced by older adults	Case and care management	Qualitative	20 (older adults)	Tablet with videoconferencing software	A tablet with videoconferencing software placed in the home of the user, connecting them to nurses, social workers, and peers with similar characteristics.	9 months	Interviews	Interpretative-hermeneutic analysis	—
Similä et al [48], 2014	Finland	Explore the experienced value of a video-mediated service by health care professionals and older users	Tele-delivered rehabilitation and occupational therapy	Qualitative	8 (5 older adults and 3 occupational therapists)	Laptop with videoconferencing software or broadcasting device	A laptop with videoconferencing software and an external webcam and speakers or Vidyo HD-50 videoconferencing attachment with touch computers for input.	10 months	Interviews, observations, and documents	Content analysis	—

XSL•FO
RenderX

Study	Country of study	Aim of study	Context of study	Research design	Participants	Type of technology (discrete device)	Type of technology (device-in-network)	Duration of use	Data types	Data analysis methods	Theoretical framework
Stokke [49], 2017	Norway	Understand the micro-context of the coproduction of care involving technologies used in care practices	Home-based telecare	Qualitative	24 (11 older adults, 4 relatives, 6 care workers, and 3 key workers or managers)	Social alarm pendant	A unit placed in the home of the user, encompassing a wrist- or neck-worn device connecting the user to a call center responder.	At least 12 months	Interviews and observations	Thematic analysis	Script theory and Domestication theory
Stokke [50], 2017	Norway	Explore how actors using a social alarm relate to, perceive, and articulate their expectations of the technology in everyday life	Home-based telecare	Qualitative	30 (11 older adults, 9 relatives, 7 care workers, and 3 key workers or managers)	Social alarm pendant	A unit placed in the home of the user, encompassing a wrist- or neck-worn device connecting the user to a call center responder.	At least 12 months	Interviews, observations, and documents	Thematic analysis	Script theory and Domestication theory
Åkerlind et al [51], 2018	Sweden	Describe how older adults with e–home care and their relatives understand and experience safety in everyday life	Home-based telecare	Qualitative	20 (12 older adults and 8 relatives)	Video phone (and other)	A video screen–based device placed in the home of the user, connecting to the computer of e–home care staff or relatives and offering video calls.	6 months	Interviews	Content analysis	—

[a]Not applicable.

[b]RPC-CT: rural palliative care consultation team.

Synthesized Findings: Analytical Themes

Overview of Themes

Four analytical themes were developed, forming a synthesized narrative depicting the multifaceted process of accepting and integrating the technology along with factors moderating this process and the various consequences of use. Through this narrative, the synthesis addresses the uses and experiences of SCT for home-dwelling older adults. The four themes are interconnected and sequentially ordered as follows: (1) *crossing the doorstep: trajectories toward technology integration*; (2) *making it fit: types of integration work*; (3) *complicated by context: physical, cognitive, and spatial challenges*; and (4) *putting it to use: (un)intended consequences of using the technology*. For the finalized thematic hierarchy, see Textbox 2.

XSL•FO
RenderX

Textbox 2. Depiction of the synthesis' thematic hierarchy, including themes and subthemes.

Crossing the doorstep: trajectories toward technology integration

- Antecedent perceptions of the technology
- Acceptance
- Temporal acceptance
- Partial acceptance
- Resistance

Making it fit: types of integration work

- Older user's work
- Caregiver work
- Service work
- Shifting roles
- Cultural and technological role expectations

Complicated by context: physical, cognitive, and spatial challenges

- Physical issues
- Significance of spatial context
- Cognitive issues

Putting it to use: (un)intended consequences of using the technology

- Technology as enabling interaction
- Technology as complicating interaction
- Technology as structuring interaction
- Social connectedness
- Creating new situations with uncertain norms
- Safety and security
- Worries
- Relief

Theme 1: Crossing the Doorstep—Trajectories Toward Technology Integration

A recurring topic across the included studies was the initial acceptance or rejection of the technology by the end users, primarily the home-dwelling older adults. As indicated by our sequential ordering of the 4 themes, acceptance is understood here as a precursor to the work involved with integrating the technology into the lives of its users. The present theme, then, centers around the phase of accepting the technology, thereby beginning the process of turning oneself from a nonuser to a user by integrating or "taming" it [55]. To expand upon the dynamics of the theme, we offer the metaphor of the user standing opposite the technology in a doorway.

While all studies reported on cases where the technology was welcomed across the doorstep more or less straightforwardly by older users, the synthesis also showed a variety of other responses. One such response was *gradual* acceptance—the technology was admitted, but with initial suspicion. Illustrating this, Stokke [49] sums up the journey of one older user initially rejecting the technology but later welcoming it: "He needed time to get used to it before accepting it." Full acceptance was achieved over time, conditional on the older user being familiarized with, and learning how to use, the technology [45,48-50].

Another response landed acceptance at a less stable form, namely one of *partiality*. In such cases, the user left the door ajar—not fully welcoming the technology but not rejecting it either. Users did so for different reasons, one being the sense that use involved compromise. Some older users of social alarms felt that use entailed a compromise with their self-perception by accepting their fragility [39,50]. Another reason was that the technology did not live up to their expectations, expressed as users' discontent with the current functions of their picturephone [45]. Similarly, cost also spurred hesitation among some older users [45]. In cases where health care professionals constituted end users, partial acceptance arose from views that video communication technology compromised quality service provision when compared with face-to-face interaction [41,44-46,48].

XSL•FO
RenderX

A final type of response involved attempts by older users to *reject* the technology. Reasons for rejection were dissatisfaction with the technology's qualities, requirements of constant tinkering to ensure functionality, convictions that engaging with the technology was not worth the effort, and unwillingness to pay money for the technology as part of health care services [45,48]. Some attempts at keeping the technology out of the household were successful, as shown in a case description by Criado and Domènech [40]. Here, a potential older user refused to allow a telecare installer to put up the telecare device that her daughter had suggested, instead escorting the installer out of her home. However, not all acts of resistance resulted in outright rejection. Some older users chose simply to ignore the technology in question [45] or to resist through intentional nonuse, as disclosed by one older user of a social alarm pendant: "In the beginning, I must say...I thought it was disgusting to wear...I did not like it. I took it off at night and put it on the bedside table" [50]. Although the technology had slipped past the doorstep, it was treated with derision as an unwelcome guest.

Across the various types of responses, relatives and informal caregivers were commonly reported as being *driving forces of acceptance* [40,43,45,50]. Whether gradual, partial, or more straightforward, acceptance was thus often found to be heavily influenced by efforts from relatives. These efforts were both instrumental, for example, by suggesting and acquiring the technology or directing the older user on how to use it, and emotional, for example, by convincing the older adult to accept the technology (eg, by framing the introduction of technology into their lives as an inevitability) [40,45]. However, the involvement of relatives could also create tensions between them and the older users. For instance, when relatives were interested in the technology but the older users were not [40,43]. One study [39] poignantly illustrates such a scenario through excerpts of a focus group interview with telecare personnel:

> *[W]hen a son or daughter requests it, which happens a lot.... The person who has to want it is them, because if you want it we're going to set it up and they're going to say, "Yes, yes, yes," so we shut up, so that we leave them alone, and when we leave they're going to take it off or leave it on top of the nightstand.*

As indicated by the excerpt, initial acceptance was sometimes provided by relatives rather than the older adults themselves because of concerns regarding their safety [39,40,43,50].

Finally, another driving force was older adults' existing perceptions, either of the technologies in question or of new technology altogether. Continuing the doorway metaphor, such perceptions are perhaps like judging the merits of who is at the door through the peephole. In some cases, existing perceptions were a driving force of rejection. For instance, the peephole glance led to skepticism of the picturephone even before it was taken into use [45]. Similarly, expecting to fail or becoming confused deterred some older adults from initially accepting videoconferencing [48]. However, we stress that the studies also found the opposite, namely older adults who viewed the prospect of new technology with curiosity [45] or as a helpful addition for managing future health-related needs [43,45]. Schmidt et al [47] noted that *all* older users had positive expectations for the videoconferencing technology.

Theme 2: Making It Fit—Types of Integration Work

Having accepted the technology, although sometimes only partially, users began the process of integrating the technology into their respective contexts. This theme illustrates how attempts at integration induce change and foregrounds the types of work performed by different users to manage these changes related to the process of integration. The theme also encompasses roles and cultural expectations, as each manifest through the integration work done by users.

Principally, *older users* engaged in work aimed at making the technology fit into their everyday life. For some, this meant not only moving furniture to accommodate the technology [45] but also learning how to use the technology, for instance, by reorganizing existing schedules or appropriating their technology use to align with their individual interests [40,45]. Several older users were enrolled in work tending to the technology by ensuring that power and internet cables were plugged in [48] and that the technology was working properly [39]. In assessing whether her home terminal was functional, one relative recounted this latter type of work in which an older user of a social alarm accepted her role as guarantor [39]:

> *[S]he says: "I got a call to check on it. I got a call from the alarm center to check to see whether..." Well, you know, if the device is working. They tell her what to press and then that's all. So she tells us about it if we are not at home.*

The work of integrating the technology also invoked *relatives*, often in the role of informal caregivers, to support the use of the technology by the older user on an instrumental level [39]. Such work included instructing the older user to remember to use the technology and continually suggesting use purposes [43]. Some studies found that relatives' work was crucial for sustained use of the technology [39,43,45], with Outila and Kiuru [45] highlighting the breakdown of integration caused by a relative unwilling to support its use:

> *Interviewer: Well, have you talked to anyone about the possibility of contacting your loved ones via picturephone? Have you had any discussions about that?*
>
> *Helena: No, I haven't, not yet. There is, in that paper, [instructions] saying that you can. But my son, he is living his own life, so he doesn't really care about my things.*

Most of the studies described work performed by *service providers* in integrating the technology. First, work was necessary to adapt to the technology-induced changes in service provision. Health care professionals had to negotiate what constituted the appropriate use of videoconferencing technology when assessing older users [46], how to distribute knowledge between teams when conducting video-based cognitive assessments [41], and how to transform content into digital formats [48]. Second, health care services operating the technology needed to allot time and resources to sustain normal operation. Such work included providing adequate technical

XSL•FO
RenderX

support for videoconferencing, allocating more time for video-based health care service provision vis-á-vis usual formats, instructing end users on how to use the technology, and conducting routine checks of social alarms' functionality [39,40,46,48,50].

Furthermore, some integration work enrolled multiple user groups: relatives navigated areas of responsibilities and acted as mediating links between user and home care services [40,43] and services worked to negotiate appropriate contexts for using technology for themselves, older users, and their relatives [50]. During these processes, their work intertwined. Indeed, 4 (80%) of the 5 studies investigating social alarm pendants reported shifts in user roles and role relations among older users, relatives, and health care services. Existing relationships between older users and relatives were destabilized by changes brought about by the technology (or its process of installation), sometimes offering opportunities to renegotiate forms of relatedness [39,40]. However, reports also indicate that precisely because of technology-induced changes to existing roles and role relations, disputes over care responsibilities may arise. Relatives involved in work to ensure the social alarms' functionality may acquire technological expertise surpassing that of health care professionals [43]. Similarly, ensuring the optimal function of the social alarms in some cases meant diverging from care constellations preferred by the older user or meant that relatives otherwise caring for the older user did not fit the role prescribed by social alarms [40].

This notion of roles points to a more general finding synthesized across several studies—that when the technology was integrated into the lives of its users, norms and expectations surfaced, pointing to underlying culturally contingent ideal role figurations between older users, relatives, and health care services [40,43,45,51]. The issue of "role fit" [40] presented earlier indicates how services operating the social alarm inscribed expectations into the device regarding who could and could not fit the role of carer-through-telecare—and how this clashed with cultural expectations of family members as carers. A daughter's comments on watching over her mother using a videophone constitute another example [51]. She explains that she views checking up digitally as a "part of everyday modern life" due to emerging possibilities for monitoring using technological devices. These findings foreground that roles involved in technology integration are not just instantiated based solely on functionality. Rather, roles and role figurations draw on preexisting belief systems grounded in cultural understandings, of both care and technologies.

Theme 3: Complicated by Context—Physical, Cognitive, and Spatial Challenges

As illustrated by the second theme, integrating the technology entail multiple practices. These practices were sometimes complicated by the presence of challenges pertaining to material and health-related circumstances. Only 3 (N=13, 23%) included studies provided a complete overview of the participants' health conditions [43,44,48], although several studies reported that health-related conditions of the older users influenced their ability to integrate the technology. Specifically, the presence of *physical or cognitive issues* complicated or hindered integration work altogether. Three (N=13, 23%) studies [41,44,45] recounted users who viewed the requirements for using the technology as insurmountable due to their deteriorating physical health. One older picturephone user described how her disposition changed due to tiredness despite initially accepting the technology: "I was too tired. I was in such a bad condition that I had to let go. And now, for at least three months, I haven't turned it on at all" [45].

Illustrating practical challenges involving physical health, Gilbert et al [41] note how remote interpreting using videoconferencing software sometimes posed requirements of hearing and vision that the older users could not fulfill. Intertwined with this inability to use the technology was also a decrease in interest to use, although reports omit the specifics of this disinterest. Lee et al [44] highlight a situation in which an older user lost interest during the process of integration as the technology seemed less relevant when viewed from her position of increased vulnerability. Similarly, one older user's son felt that the technology was not a suitable solution for his mother because of her poor physical condition [51]. Thus, physical health issues were reported to influence both ability and interest to use.

However, most of the reported challenges were related to the older users' cognitive capacity. Several studies noted how older users lacked the required cognitive capacity to use the technology. For instance, using videoconferencing software demanded concentration and effort beyond that of face-to-face communication to follow the conversation [51] or required too steep of a learning curve. Some users reported that integrating the technology and becoming a "fluent user" [45] was challenged by users' own perceptions of their capacity to learn at their age. In addition, using the technology could require skills of comprehension not attainable by some older users. According to an occupational therapist, video-mediated therapy requires users to comprehend the norms of video communication, which was not always the case [48]. Similarly, users with cognitive issues sometimes forgot who was on the screen [46]. Stokke [50] vividly depicts how a social alarm, designed to be easy to operate by older adults, nonetheless relies on its user's ability to understand its function when in need:

> *Marie was the one wearing the alarm. She was physically vigorous, even though she had had a stroke some years back and suffered from dementia. Marie did not understand how she should call for help when her husband got ill, and Peter was too heavy for Marie to manage.... When he experienced falls in blood pressure, he passed out for a while. Sometimes she pressed the pendant, sometimes not. She got confused when these things happened, but then he woke up and said, "Press the button!"*

Paralleling the findings of Stokke [50], some studies included comments made by informal caregivers and health care professionals in which the technology in question was deemed inappropriate to use with or by older users with dementia [41,49]. In these ways and others, the older users' deteriorating health came in the way of performing the necessary integration work.

XSL•FO
RenderX

Beyond health, issues related to spatial context were also reported to influence integration work. First, such influences related to the spatial arrangement of the technology, where suboptimal lighting conditions [41,46], background noise, and disturbances by other people present [41,48] were noted by health care professionals as consequences of poor placement of the technology in the home of the older user. By the same token, Ilomäki and Ruusuvuori [42] illustrate that guiding medicine taking may be disrupted by too far of a distance between medicine and videoconferencing equipment. Second, several studies detailed how the physical presence or absence of human actors affected the integration of the technology [41,43,46,47]. For example, the presence of relatives during the installation and maintenance of older users' social alarms was emphasized by Criado and Domènech [40], while Outila and Kiuru [45] stated that a lack of nearby relatives acting as technology aides hindered the use of the picturephone.

Theme 4: Putting It to Use—(Un)intended Consequences of Using the Technology

While the 2 preceding themes synthesized findings concerning (challenges to) work involved in integrating the technology into the lives of its users, this theme focuses on the reported consequences that arise from having taken the technology into use across user groups. Depending on their process of integration, users derived a wealth of different experiences from interacting with technology for different purposes.

First, older users expressed how using SCT gave rise to feelings of *social connectedness* with peers or relatives despite being isolated either at home or in the hospital [45,47,48,51] and decreased feelings of loneliness [44,45]. Even routine checks performed by health care services made a social alarm user feel less alone, as commented by one relative [39].

Second, feelings of increased *safety and security* were reported by many users. Older users felt that using the technology allowed a sense of relatives or services to be "distant but present," which provided them with safety while enabling them to live at home with increased independence [39,43,49,51]. By the same token, relatives connecting to the older user via technology also felt safer [39,43,49,51]. While older users felt both connected and safer, users in health care services felt that video communication enabled effective health care provision at a distance owing to visual access [41,45,50] and made encounters with older users more succinct by sidestepping informal chatter [41] and allowing physicians to interact directly with patients [46]. Only one experience of increased safety was reported by health care services. This was concerned with the ability of videoconferencing technology to mitigate the need for service provision in unfamiliar settings and in the presence of strangers [41].

Related to, but separate from, safety and security were reports that using the technology brought users *relief*. Older users said they experienced relief as the technology reduced personal health care costs by replacing physical visits [45] and as they did not need to involve worried relatives in case of an emergency [39]. Relatives felt that being present despite distance relieved them of some care responsibilities [50], allowing them enough peace of mind to go on vacation or simply sleep at night [39,51]. Using the technology was experienced by services as saving time and costs by circumventing travel [41,46,48] and by ensuring fragile older patients easy access to help if needed [45,50].

Contrasting the preceding points, some users experienced the technology as inducing *worries*. Primarily experienced by older users, worries were related to the use of the technology, revolving around technical malfunctions [39,41,51] and internet security [46]. Some relatives worried that the technology could extend older users' life at home beyond what they deemed appropriate in terms of safety [43,51]. One relative and older user aired their concern that using technology may lessen human contact in future health care service provision for older adults [45,51].

Far from stable cause-and-effect end points, the aforementioned experiences are consequences resting upon the dynamics of users' integration work. We now turn to these dynamics, denoting the ways in which the use of technology produced new situations, sometimes creating uncertain norms challenging the users' actions, habits, and beliefs. Nowhere is this seen clearer than in older users' negotiation of what constitutes appropriate and inappropriate use [39,43-45,49,50]. Exemplifying this, one older user recounts her choice not to use her social alarm pendant after a fall despite it being intended for emergencies [39]:

> *[T]he chair fell down over that way and I fell the other way.... "Ok, everything hurts but no, I haven't broken anything because the fall was soft. No, so get up," and of course I wasn't going to call because you're not going to stir things up just for something like this.*

This excerpt illustrates how acquiring and integrating the technology introduces questions of what it means to be in need. Similarly, one older user sometimes felt the need to test his alarm pendant despite it being for emergency use only and navigated this question of legitimate use by using his humoristic nature and well-established relationship with the health care unit [49]. In several studies, it thus became clear that the established figuration surrounding the technology left open a degree of interpretation for its users to negotiate. Correspondingly, variations in implementation and ambivalences regarding the end purpose of the technology arose [43,45,50].

However, the work of integrating the technology also influenced interactions between users by imposing different interactional premises. Numerous studies reported on how such premises complicated interactions for the respective users, although primarily from the perspective of health care professionals. The imposed format of synchronous audiovisual communication gave rise to interactional dysfluencies: turn-taking became more difficult [41,42,48], conversations were halted by audio and video delays or blackouts [41,46,47], and empathetic or gestural displays became restricted [41]. In addition, some health care professionals stated that discussing sensitive topics was unsuitable using the format [41,46]. Interaction could also be complicated by the spatial separation of interlocutors, as it meant that health care professionals relinquished control of the other user's side. They were unable to gain a sense of the older user's

XSL•FO
RenderX

settings, including who else was present during the conversation [41,46], and found it difficult to direct the older user's attention back to the conversation at hand if they went off-topic [41,48]. Furthermore, the capacity to connect instantly to older users meant not warming up and losing interprofessional coordination in advance of the encounter, thereby complicating interaction [41].

Finally, establishing social alarm "roles" restructured existing figurations and therefore interactions between older users, their relatives, and the involved health care services in ways that created tension [39]. The following excerpt illustrates this restructuring [50]:

> *[T]he care-worker was usually helping other patients when they received an alarm. Due to confidentiality, they had to leave the room to talk and were therefore interrupted in their work. There are many field notes describing how the person receiving help commented on this, as he/she had to wait for the care-worker to return and continue the work.*

Here, the prioritization of social alarms by the home care service meant that all other care tasks became subject to potential disruption in case an alarm went off. Thus, the technology created more fragile situations in which distant actors were able to interfere with the situation at hand.

Discussion

Principal Findings

By identifying, assessing, and synthesizing findings across 13 qualitative studies, this review provides an understanding of the nuances of how SCT for home-dwelling older adults in a home care services context is used and experienced. Using the method proposed by Thomas and Harden [52] for thematic synthesis enabled an *upcycling* [56] of the findings presented by the included primary studies through analytical abstraction, thereby generating a conceptually innovative sequential understanding of technology use. While other reviews share our overall area of interest, they approach this from different perspectives. These either focus exclusively on the experiences of older adults [12,21,57,58] or include other actors but only when the older adults live with dementia [26,31,59]. Few review studies focalizing on SCT exist; only the scoping review by Chelongar and Ajami [25] operates with a similar definition but includes both qualitative and quantitative studies and omits synthesization. Our review's innovative contribution is its attention to both the uses and experiences of SCT as contingent on users other than the older user, as well as the technology's imposition of new premises for their interactions and relatedness.

Although the first theme of the synthesis indicates that acceptance, even if partial, takes as its premise the willingness of the older adult to transform from nonuser to user, one central aspect of our results relates to the enrollment of other users in facilitating acceptance by older users. Relatives and—to a lesser extent—service providers often work to acquire the technology and to persuade the older user to use it. Using the term "warm experts," Olsson and Viscovi [60] emphasize the role of grown children in appointing, acquiring, and supporting the use of technological devices by their parents, drawing on their familiarity with their qualities and needs. This review echoes such findings but also accentuates the sometimes conflicting values and roles between different users—particularly relatives and older adults. These contest the unequivocal *warmth* of relatives' engagements. When relatives pursue acceptance despite their parents' unenthusiasm or act as driving forces in initially acquiring the technology, the notion of older users' acceptance may be construed as a door to be breached rather than to be let through. Although several included studies attest to this inference, a few reported cases juxtaposed the driving forces; for example, a daughter's compliance with her father's wish to age at home using technology despite her conviction that a nursing home might be a better option. The emphasis on the older user's acceptance may rest partly on the review's focus on SCT, as older users must inhabit a role in 1 of 2 possible ends of the medium. However, the prevalent preoccupation with older adults' technology acceptance in research [61-63] lends credence to acceptance being a crucial stage for technology integration. Therefore, it is notable that acceptance among relatives and service providers is a matter largely omitted by the included studies. Given their significance as proxies for older adults' acceptance of SCT, emphasis on these user groups may be productive when investigating SCT for home-dwelling older adults.

The synthesis ventures beyond acceptance, with the second theme bringing different users' continuous integration work to the fore and highlighting the dynamics of roles, expectations, and the technology's fit into existing relations. In this regard, it engages with numerous studies operationalizing the concept of *domestication* to investigate the sociotechnical processes of integration mediating technological "effects" [64-67]. In the context of eldercare, Saborowski and Kollak [68] argue that, rather than being of a technical nature, the question of whether telecare technology is "good" hinges on its domestication into the existing care network. Such understandings of technology integration run parallel to calls for more nuanced approaches to the evaluation of technologies in health care [69-71]. Far from a cry from specialized fields within social science, a recent review investigating knowledge gaps in existing research on the implementation of welfare technology reached a similar conclusion—that research should address longitudinal processes; involve multiple users; and include their attitudes, beliefs, and actual uses [72].

By highlighting the weight given to the physical and cognitive capacities of older users by the included studies, the third theme simultaneously reveals that the competencies of other users and their context—be it organizational or familial—received less attention. This finding aligns with the assertion of Peine and Neven [73] that the development of technology intended for use by older people disproportionally casts their disabilities and needs as the primary issue to solve. In contrast, the study by Nordtug et al [74] on video consultations (a technology comparable to that investigated by 6 of the 13, 46% studies) revealed that uncertainties among general practitioners significantly influenced their integration and use. In this regard, the context surrounding technology use should be considered. All the included studies investigated digitally delivered

XSL•FO
RenderX

home-based care—a cultural context with certain conventions potentially framing users' interaction with the technology [16]. The included SCT can be argued to mirror such a context, as it established a sender-recipient relationship paralleling traditional care pathways. The only exception is the picturephone [45]. Interestingly, this is also the only case in which older users complained of device costs, suggesting that the older users inhabited a consumer role in acquiring the picturephone and—by extension—the selection of that particular care pathway.

The significance of context notwithstanding, our findings point to the general consequences of using the technology across the included studies. Older users benefitted from being able to contact relatives and peers, thereby feeling more socially connected and safer despite isolation. Although only Lee et al [44] explicitly sought to investigate the impacts of technologically mediated interactions on loneliness, a total of 6 (N=13, 46%) studies reported increased social connectedness among older users. Such a discrepancy points to the agency of users in defining the meaning of their engagements with SCT irrespective of its originally intended purposes. Another enablement experienced by older users and, principally, service providers was that of avoiding travel. For service providers, this equated to saving resources. The realization of this particular capacity of SCT is in line with the framing of technology implementation in health care in European countries as a cost-effective measure [8].

Limitations

Qualitative systematic reviews often seek heterogeneity to maximize nuance across included studies [52,75]. In contrast, our review intentionally delineated technologies of interest as those capable of mediating synchronous communication between human actors. On the basis of the assumption that the characteristics of a technology matter more than its status, we sought technological homogeneity to achieve better comparability. This strategy may indeed have resulted in the loss of valuable variations in use cases and experiences. During the database query, our interest in a particular technological type influenced our choice of search terms; we abstained from relying only on umbrella terms (eg, telecare and eHealth) because of their breadth and included specific technologies (eg, tablets and apps). Beyond revealing what technologies we were interested in, this may have skewed the search toward technology types well-known to us. Nonetheless, this delineation proved productive, revealing a potential for future reviews to address other technologies (eg, social robots or electronic medical record systems), while remaining attentive to their specificities.

Choosing the community home care setting as the review's area of inquiry gives precedence to countries operating with this organizational context of home care, thereby compromising transferability to differently organized systems of home care provision. On a more general note, we advocate for a principle of selective and informed transferability when attempting to apply the findings of this qualitative meta-synthesis—that by affording transparency regarding the review process, we provide readers with the means to make well-informed decisions about the applicability of our results to their own context [75]. Thus, we do not assume our results are transferable to all contexts—not only due to the aforementioned technological and organizational specificities but also due to the nature of qualitative research itself [53]. Similarly, several aspects of our approach compromise the reproducibility of this review. Framing subjectivity as a precondition for interpretative work rather than as a risk factor to be eliminated, diverging perspectives emerging between reviewers during the processes of screening, appraising, and synthesizing, were viewed positively—as a crossroad involving productive negotiations between alternative interpretations. Discussion meetings were held throughout these processes to align perspectives and achieve transparency but without the conviction that standardized procedures will or must produce the same outcome across all reviewers [56].

Conclusions

Synthesizing findings across 13 studies, our review aimed to enrich knowledge on the uses and experiences of SCT for home-dwelling older adults in a home care context. Four analytical themes composed a sequential, multifaceted process illustrating first the gradual, partial, and resistance-laden dimensions of accepting technology by older users, and the role of relatives as mediators therein. Succeeding acceptance, the integration of the technology implies the introduction of change, demanding that older users, relatives, and health care providers address these by engaging in work and to tend to the technology's functioning, to its sustained use, and to each other. Contextual issues complicate these practices, a notable concentration of which concern the qualities of older users specifically. Feelings of social connectedness, safety and security, relief, and worries are not just consequences of use but also consequences of profound reconfigurations of practices and relations across different users. To our knowledge, this is the first review to synthesize scientific reports involving this particular constellation of technologies, users, and contexts. On account of this specificity, we argue that our findings may be transferable to policies aimed at—and practices within—similar domains. At the least, it invites a more nuanced understanding of the implications of introducing SCT into the lives of its users.

Acknowledgments

The authors wish to thank Malene Wøhlk Gundersen, head of section at the Division for Research, Development and University Library, OsloMet, for her assistance in constructing the search query used in this review. This review is part of the first author's PhD research and is embedded within a larger research and innovation project that centers on welfare technology implementation in Oslo, which has been funded by the Research Council of Norway (grant 331810).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Decision summary based on the expanded Critical Appraisal Skills Programme tool.
[XLSX File (Microsoft Excel File), 15 KB - jmir_v26i1e59285_app1.xlsx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 68 KB - jmir_v26i1e59285_app2.pdf]

Multimedia Appendix 3
Studies excluded because of insufficient methodological quality.
[DOCX File , 14 KB - jmir_v26i1e59285_app3.docx]

References

1. Ageing and health. World Health Organization. 2024 Oct 1. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2024-10-26]
2. Ageing policy in Europe, North America and Central Asia in 2017-2022. United Nations Economic Commission for Europe. 2022. URL: https://unece.org/sites/default/files/2022-08/Synthesis-report_0.pdf [accessed 2024-09-25]
3. Global health and aging. National Institute on Ageing, National Institutes of Health, U.S. Department of Health and Human Services. URL: https://www.nia.nih.gov/sites/default/files/2017-06/global_health_aging.pdf [accessed 2024-04-06]
4. Communication from the commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions REPowerEU plan. European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A230%3AFIN [accessed 2024-09-26]
5. NOU 2023: 4 tid for handling — personellet i en bærekraftig helse- og omsorgstjeneste. Regjeringen.no. URL: https://www.regjeringen.no/no/dokumenter/nou-2023-4/id2961552/ [accessed 2024-10-26]
6. Lindeman DA, Kim KK, Gladstone C, Apesoa-Varano EC. Technology and caregiving: emerging interventions and directions for research. Gerontologist 2020 Feb 14;60(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1093/geront/gnz178] [Medline: 32057082]
7. Neven L, Peine A. From triple win to triple sin: how a problematic future discourse is shaping the way people age with technology. Societies 2017 Sep 15;7(3):26 [FREE Full text] [doi: 10.3390/soc7030026]
8. UN decade of healthy ageing: plan of action 2021-2030. World Health Organization. URL: https://cdn.who.int/media/docs/default-source/decade-of-healthy-ageing/decade-proposal-final-apr2020-en.pdf?sfvrsn=b4b75ebc_28 [accessed 2024-09-25]
9. Hammink C, Mohammadi M. Uncovering homecare needs and challenges across Europe. Interreg ACE Project. 2024 Apr 18. URL: https://www.interregnorthsea.eu/sites/default/files/2024-04/Uncovering%20Homecare%20needs%20and%20challenges%20across%20Europe_April%202024_0.pdf [accessed 2024-09-25]
10. Rush KL, Singh S, Seaton CL, Burton L, Li E, Jones C, et al. Telehealth use for enhancing the health of rural older adults: a systematic mixed studies review. Gerontologist 2022 Nov 30;62(10):e564-e577. [doi: 10.1093/geront/gnab141] [Medline: 34661675]
11. Connelly K, ur Rehman Laghari K, Mokhtari M, Falk TH. Approaches to understanding the impact of technologies for aging in place: a mini-review. Gerontology 2014;60(3):282-288 [FREE Full text] [doi: 10.1159/000355644] [Medline: 24457288]
12. Karlsen C, Ludvigsen MS, Moe CE, Haraldstad K, Thygesen E. Experiences of community-dwelling older adults with the use of telecare in home care services: a qualitative systematic review. JBI Database System Rev Implement Rep 2017 Dec;15(12):2913-2980. [doi: 10.11124/JBISRIR-2017-003345] [Medline: 29219874]
13. Lynch J, Hughes G, Papoutsi C, Wherton J, A'Court C. "It's no good but at least I've always got it round my neck": a postphenomenological analysis of reassurance in assistive technology use by older people. Soc Sci Med 2022 Jan;292:114553 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114553] [Medline: 34799181]
14. Mort M, Roberts C, Pols J, Domenech M, Moser I, EFORTT investigators. Ethical implications of home telecare for older people: a framework derived from a multisited participative study. Health Expect 2015 Jun;18(3):438-449 [FREE Full text] [doi: 10.1111/hex.12109] [Medline: 23914810]
15. Stokke R, Hellesø R, Sogstad M. Hvorfor er det så vanskelig å integrere velferdsteknologii omsorgstjenesten? Tidsskrift Omsorgsforskning 2019 Dec 20;5(3):7-20. [doi: 10.18261/issn.2387-5984-2019-03-02]
16. Norman D. The Design of Everyday Things. New York, NY: Basic Books; 2013.
17. McLuhan M. Understanding Media: The Extensions of Man. Cambridge, MA: MIT Press; 1994.

XSL•FO
RenderX

18. Peek ST, Luijkx KG, Rijnaard MD, Nieboer ME, van der Voort CS, Aarts S, et al. Older adults' reasons for using technology while aging in place. Gerontology 2016;62(2):226-237 [FREE Full text] [doi: 10.1159/000430949] [Medline: 26044243]
19. Airola E, Rasi P. Domestication of a robotic medication-dispensing service among older people in Finnish Lapland. Hum Technol 2020;16(2):117-138. [doi: 10.17011/ht/urn.202008245639]
20. Lindberg B, Nilsson C, Zotterman D, Söderberg S, Skär L. Using information and communication technology in home care for communication between patients, family members, and healthcare professionals: a systematic review. Int J Telemed Appl 2013;2013:461829 [FREE Full text] [doi: 10.1155/2013/461829] [Medline: 23690763]
21. Lindberg T, Sandström B, Andersson EK, Borg C, Hjelm M, Nilsson L, et al. Older persons' experience of eHealth services in home health care: a meta-ethnography eHealth services in home health care. Health Informatics J 2021;27(4):14604582211059370 [FREE Full text] [doi: 10.1177/14604582211059370] [Medline: 36047741]
22. Şahin E, Yavuz Veizi BG, Naharci MI. Telemedicine interventions for older adults: a systematic review. J Telemed Telecare 2024 Feb;30(2):305-319 [FREE Full text] [doi: 10.1177/1357633X211058340] [Medline: 34825609]
23. Wang J, Fu Y, Lou V, Tan SY, Chui E. A systematic review of factors influencing attitudes towards and intention to use the long-distance caregiving technologies for older adults. Int J Med Inform 2021 Sep;153:104536. [doi: 10.1016/j.ijmedinf.2021.104536] [Medline: 34325206]
24. Holthe T, Halvorsrud L, Lund A. Digital assistive technology to support everyday living in community-dwelling older adults with mild cognitive impairment and dementia. Clin Interv Aging 2022 Apr 18;17:519-544 [FREE Full text] [doi: 10.2147/CIA.S357860] [Medline: 35464157]
25. Chelongar K, Ajami S. Using active information and communication technology for elderly homecare services: a scoping review. Home Health Care Serv Q 2021;40(1):93-104. [doi: 10.1080/01621424.2020.1826381] [Medline: 32990180]
26. Martínez-Alcalá CI, Pliego-Pastrana P, Rosales-Lagarde A, Lopez-Noguerola JS, Molina-Trinidad EM. Information and communication technologies in the care of the elderly: systematic review of applications aimed at patients with dementia and caregivers. JMIR Rehabil Assist Technol 2016 May 02;3(1):e6 [FREE Full text] [doi: 10.2196/rehab.5226] [Medline: 28582258]
27. Bavngaard M, Lund A, Thordardottir B, Rasmussen E. The uses and experiences of communication technology for home-dwelling older adults in a homecare services context: a qualitative systematic review. PROSPERO. 2023. URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=414243 [accessed 2024-10-26]
28. Butler A, Hall H, Copnell B. A guide to writing a qualitative systematic review protocol to enhance evidence-based practice in nursing and health care. Worldviews Evid Based Nurs 2016 Jun;13(3):241-249. [doi: 10.1111/wvn.12134] [Medline: 26790142]
29. Coiera E. Communication systems in healthcare. Clin Biochem Rev 2006 May;27(2):89-98 [FREE Full text] [Medline: 17077879]
30. Moser I. Velferdsteknologi. En Ressursbok. Oslo, Norway: Cappelen Damm; 2019.
31. Boyle LD, Husebo BS, Vislapuu M. Promotors and barriers to the implementation and adoption of assistive technology and telecare for people with dementia and their caregivers: a systematic review of the literature. BMC Health Serv Res 2022 Dec 23;22(1):1573 [FREE Full text] [doi: 10.1186/s12913-022-08968-2] [Medline: 36550456]
32. EndNote. Clarivate. URL: https://clarivate.com/innovation-exchange/solution/endnote/ [accessed 2024-10-26]
33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
34. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large‐evidence systematic reviews and meta‐analyses. Res Synth Methods 2019 Sep;10(3):330-342. [doi: 10.1002/jrsm.1354]
35. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
36. Atkins S, Lewin S, Smith H, Engel M, Fretheim A, Volmink J. Conducting a meta-ethnography of qualitative literature: lessons learnt. BMC Med Res Methodol 2008 Apr 16;8:21 [FREE Full text] [doi: 10.1186/1471-2288-8-21] [Medline: 18416812]
37. CASP checklists. Critical Appraisal Skills Programme. URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2024-04-05]
38. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Res Methods Med Health Sci 2020 Aug 06;1(1):31-42. [doi: 10.1177/2632084320947559]
39. Correa G, Domènech M. Care networking: a study of technical mediations in a home telecare service. Int J Environ Res Public Health 2013 Jul 22;10(7):3072-3088 [FREE Full text] [doi: 10.3390/ijerph10073072] [Medline: 23880730]
40. Criado TS, Domenech M. Older people in a connected autonomy? Promises and challenges in the technologization of care. Span J Sociol Res 2015;152:105-120. [doi: 10.5477/cis/reis.152.105]
41. Gilbert AS, Croy S, Hwang K, LoGiudice D, Haralambous B. Video remote interpreting for home-based cognitive assessments: stakeholders' perspectives. Interpreting 2021 Aug 11;24(1):84-110. [doi: 10.1075/intp.00065.gil]

XSL•FO
RenderX

42. Ilomäki S, Ruusuvuori J. Preserving client autonomy when guiding medicine taking in telehomecare: a conversation analytic case study. Nurs Ethics 2022 May 04;29(3):719-732 [FREE Full text] [doi: 10.1177/09697330211051004] [Medline: 35119321]
43. Karlsen C, Moe CE, Haraldstad K, Thygesen E. Caring by telecare? A hermeneutic study of experiences among older adults and their family caregivers. J Clin Nurs 2019 Apr;28(7-8):1300-1313. [doi: 10.1111/jocn.14744] [Medline: 30552788]
44. Lee K, Fields N, Cassidy J, Kusek V, Feinhals G, Calhoun M. Caring callers: the impact of the telephone reassurance program on homebound older adults during COVID-19. Home Health Care Serv Q 2021 Nov 07;40(4):247-261. [doi: 10.1080/01621424.2021.1997861] [Medline: 34747329]
45. Outila M, Kiuru H. “Picturephone in my home”: actor-network theory and Foucauldian Discourse Analysis on Northern Finnish older adults starting to use a video conferencing service. J Technol Hum Serv 2021 Jan 15;39(2):163-192. [doi: 10.1080/15228835.2020.1869670]
46. Read Paul L, Salmon C, Sinnarajah A, Spice R. Web-based videoconferencing for rural palliative care consultation with elderly patients at home. Support Care Cancer 2019 Sep 7;27(9):3321-3330. [doi: 10.1007/s00520-018-4580-8] [Medline: 30613908]
47. Schmidt S, Behrens J, Lautenschlaeger C, Gaertner B, Luderer C. Experiences with combined personal-online case management and the self-reliance of older people with multimorbidity living alone in private households: results of an interpretative-hermeneutical analysis. Scand J Caring Sci 2019 Dec 17;33(4):931-939. [doi: 10.1111/scs.12691] [Medline: 30994204]
48. Similä H, Harjumaa M, Isomursu M, Ervasti M, Moilanen H. Video communication in remote rehabilitation and occupational therapy groups. Phys Occup Ther Geriatr 2014 May 23;32(2):97-111. [doi: 10.3109/02703181.2014.919050]
49. Stokke R. "Maybe we should talk about it anyway": a qualitative study of understanding expectations and use of an established technology innovation in caring practices. BMC Health Serv Res 2017 Sep 15;17(1):657 [FREE Full text] [doi: 10.1186/s12913-017-2587-3] [Medline: 28915809]
50. Stokke R. Older people negotiating independence and safety in everyday life using technology: qualitative study. J Med Internet Res 2018 Oct 19;20(10):e10054 [FREE Full text] [doi: 10.2196/10054] [Medline: 30341049]
51. Åkerlind C, Martin L, Gustafsson C. eHomecare and safety: the experiences of older patients and their relatives. Geriatr Nurs 2018 Mar;39(2):178-185. [doi: 10.1016/j.gerinurse.2017.08.004] [Medline: 28988833]
52. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
53. Sandelowski M, Barroso J. Handbook for Synthesizing Qualitative Research. Cham, Switzerland: Springer Publishing Company; 2006.
54. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications; 2021.
55. Pols J, Willems D. Innovation and evaluation: taming and unleashing telecare technology. Sociol Health Illn 2011 Mar 17;33(3):484-498 [FREE Full text] [doi: 10.1111/j.1467-9566.2010.01293.x] [Medline: 21241338]
56. Malterud K. Qualitative Metasynthesis: A Research Method for Medicine and Health Sciences. Milton Park, UK: Taylor & Francis; 2019.
57. Silva RC, Zechner M, Yaylagul-Korkmaz N, Efthymiou A, Jøranson N. Experienced barriers in the use of ICT for social interaction in older adults ageing in place: a qualitative systematic review protocol. PROSPERO. 2022. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022370044 [accessed 2024-04-05]
58. Hovland H, Thygesen E, Haraldstad K, Karlsen C. Older adults’ experiences with digital technology to maintain social contact: a qualitative meta-synthesis protocol. PROSPERO. 2022. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022322554 [accessed 2024-04-05]
59. Thordardottir B, Malmgren Fänge A, Lethin C, Rodriguez Gatta D, Chiatti C. Acceptance and use of innovative assistive technologies among people with cognitive impairment and their caregivers: a systematic review. Biomed Res Int 2019;2019:9196729 [FREE Full text] [doi: 10.1155/2019/9196729] [Medline: 30956989]
60. Olsson T, Viscovi D. Warm experts for elderly users: who are they and what do they do? Human Technol 2018 Nov 30;14(3):324-342. [doi: 10.17011/ht/urn.201811224836]
61. Peng L, Man SS, Chan AH, Ng JY. Personal, social and regulatory factors associated with telecare acceptance by Hong Kong older adults: an indication of governmental role in facilitating telecare adoption. Int J Hum Comput Interact 2022 Apr 20;39(5):1059-1071. [doi: 10.1080/10447318.2022.2050539]
62. Martín-García AV, Redolat R, Pinazo-Hernandis S. Factors influencing intention to technological use in older adults. The TAM model application. Res Aging 2022 Dec 28;44(7-8):573-588. [doi: 10.1177/01640275211063797] [Medline: 34962846]
63. Cimperman M, Makovec Brenčič M, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an extended UTAUT model. Int J Med Inform 2016 Jun;90:22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
64. Johannessen LE, Nordtug M, Haldar M. Multi-site domestication: taming technologies across multiple institutional settings. Inf Commun Soc 2023 Sep 11:1-17. [doi: 10.1080/1369118x.2023.2255644]
65. Hämäläinen A, Hirvonen H. Electronic health records reshaping the socio-technical practices in long-term care of older persons. Technol Soc 2020 Aug;62:101316. [doi: 10.1016/j.techsoc.2020.101316]

XSL•FO
RenderX

66. Søraa RA, Nyvoll P, Tøndel G, Fosch-Villaronga E, Serrano JA. The social dimension of domesticating technology: interactions between older adults, caregivers, and robots in the home. Technol Forecast Soc Change 2021 Jun;167:120678. [doi: 10.1016/j.techfore.2021.120678]
67. Jalil S, Hardy D, Myers T, Atkinson I. But it doesn't go with the décor: domesticating a telemedicine diabetes intervention in the home. In: Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design. 2014 Presented at: OzCHI '14; December 2-5, 2014; Sydney, Australia. [doi: 10.1145/2686612.2686653]
68. Saborowski M, Kollak I. "How do you care for technology?" – Care professionals' experiences with assistive technology in care of the elderly. Technol Forecast Soc Change 2015 Apr;93:133-140. [doi: 10.1016/j.techfore.2014.05.006]
69. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
70. Oudshoorn N. Telecare Technologies and the Transformation of Healthcare. London, UK: Palgrave Macmillan; 2011.
71. Pols J. Care at a Distance: On the Closeness of Technology. Amsterdam, The Netherlands: Amsterdam University Press; 2012.
72. Borg J, Gustafsson C, Landerdahl Stridsberg S, Zander V. Implementation of welfare technology: a state-of-the-art review of knowledge gaps and research needs. Disabil Rehabil Assist Technol 2023 Feb 14;18(2):227-239. [doi: 10.1080/17483107.2022.2120104] [Medline: 36103349]
73. Peine A, Neven L. Social-structural lag revisited. Gerontechnology 2011;10(3):129-139. [doi: 10.4017/gt.2011.10.3.002.00]
74. Nordtug M, Assing Hvidt E, Lüchau EC, Grønning A. General practitioners' experiences of professional uncertainties emerging from the introduction of video consultations in general practice: qualitative study. JMIR Form Res 2022 Jun 14;6(6):e36289 [FREE Full text] [doi: 10.2196/36289] [Medline: 35653607]
75. Saini M, Shlonsky A. Systematic Synthesis of Qualitative Research. New York, NY: Oxford University Press; 2012.

Abbreviations

ICT: information and communications technology
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SCT: synchronous communication technology

Edited by A Coristine; submitted 08.04.24; peer-reviewed by D Patil, H Pilabré; comments to author 04.06.24; revised version received 04.07.24; accepted 30.09.24; published 22.11.24.

Please cite as:
Bavngaard MV, Lund A, Thordardottir B, Rasmussen EB
The Uses and Experiences of Synchronous Communication Technology for Home-Dwelling Older Adults in a Home Care Services Context: Qualitative Systematic Review
J Med Internet Res 2024;26:e59285
URL: https://www.jmir.org/2024/1/e59285
doi:10.2196/59285
PMID:

©Martin Vinther Bavngaard, Anne Lund, Björg Thordardottir, Erik Børve Rasmussen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Elements Influencing User Engagement in Social Media Posts on Lifestyle Risk Factors: Systematic Review

Yan Yee Yip[1,2], MPharm, MSc; Mohd Makmor-Bakry[1], PhD; Wei Wen Chong[1], PhD

[1]Centre for Quality Management of Medicines, Faculty of Pharmacy, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

[2]Centre for Clinical Epidemiology, Institute for Clinical Research, National Institutes of Health, Shah Alam, Selangor Darul Ehsan, Malaysia

Corresponding Author:
Wei Wen Chong, PhD
Centre for Quality Management of Medicines
Faculty of Pharmacy
Universiti Kebangsaan Malaysia
Jalan Raja Muda Abdul Aziz
Kuala Lumpur, 50300
Malaysia
Phone: 60 39289711
Email: weiwen@ukm.edu.my

Abstract

Background: The high prevalence of noncommunicable diseases and the growing importance of social media have prompted health care professionals (HCPs) to use social media to deliver health information aimed at reducing lifestyle risk factors. Previous studies have acknowledged that the identification of elements that influence user engagement metrics could help HCPs in creating engaging posts toward effective health promotion on social media. Nevertheless, few studies have attempted to comprehensively identify a list of elements in social media posts that could influence user engagement metrics.

Objective: This systematic review aimed to identify elements influencing user engagement metrics in social media posts by HCPs aimed to reduce lifestyle risk factors.

Methods: Relevant studies in English, published between January 2006 and June 2023 were identified from MEDLINE or OVID, Scopus, Web of Science, and CINAHL databases. Included studies were those that examined social media posts by HCPs aimed at reducing the 4 key lifestyle risk factors. Additionally, the studies also outlined elements in social media posts that influenced user engagement metrics. The titles, abstracts, and full papers were screened and reviewed for eligibility. Following data extraction, narrative synthesis was performed. All investigated elements in the included studies were categorized. The elements in social media posts that influenced user engagement metrics were identified.

Results: A total of 19 studies were included in this review. Investigated elements were grouped into 9 categories, with 35 elements found to influence user engagement. The 3 predominant categories of elements influencing user engagement were communication using supportive or emotive elements, communication aimed toward behavioral changes, and the appearance of posts. In contrast, the source of post content, social media platform, and timing of post had less than 3 studies with elements influencing user engagement.

Conclusions: Findings demonstrated that supportive or emotive communication toward behavioral changes and post appearance could increase postlevel interactions, indicating a favorable response from the users toward posts made by HCPs. As social media continues to evolve, these elements should be constantly evaluated through further research.

(J Med Internet Res 2024;26:e59742) doi:10.2196/59742

KEYWORDS

chronic disease; health promotion; internet; primary prevention; social media; systematic reviews; health care professional; health personnel; user engagement; lifestyle; risk

XSL•FO
RenderX

Introduction

Social media is a communication tool that allows the creation and exchange of user-generated content. Facebook (Meta Platforms, Inc), YouTube (Google, Inc), and WhatsApp (Meta Platforms, Inc) are the 3 most widely accessed social networking platforms globally [1].

In recent years, social media has been increasingly used in health promotion through the delivery of health information. Nevertheless, the emergence of the COVID-19 pandemic has given rise to both accurate and misleading information. Thus, it is important to ensure the public has access to accurate information from reliable health sources [2]. Health care professionals (HCPs) are, therefore, responsible for the delivery of trustworthy information on social media either as individuals or as part of an organization [3].

Noncommunicable diseases (NCDs) are a major global health concern due to their high disease burden and the large number of deaths, which is estimated to be around 41 million people yearly [4]. The World Health Organization (WHO) has identified 4 key lifestyle risk factors that contribute to NCDs—tobacco use, harmful use of alcohol, unhealthy diet, and physical inactivity [4,5]. These risk factors could be reduced through the practice of healthy lifestyle behaviors. Health information on positive lifestyle behaviors can be effectively delivered through social media, making it accessible to larger populations at a lower cost [6,7]. Findings in the United States have found that approximately 70% of health organizations have used social media in community engagement through patient education and delivery of health-related news and information [3].

While HCPs have used these platforms, their effectiveness in promoting healthy behaviors and engaging users remains uncertain. The initial indication of users' acceptance of social media posts promoting healthy behaviors can be assessed using user engagement metrics. These metrics provide a quantifiable and measurable representation of users' interactions with social media posts, including likes, comments, and shares [8]. High user engagement indicates the resonance between posts and the audience's interests, often leading to extensive sharing within their respective networks [9].

To achieve elevated user engagement in social media posts, it is important to identify and prioritize the key elements in driving user interactions. These elements can be elicited through the examination of social media posts. Indeed, it is crucial to recognize that the elements in social media posts are complex, often relying on combinations of elements to influence user engagement [10]. For example, Hales et al [11] found that elements, such as polls and posts asking for suggestions contributed to increased user engagement in posts related to weight management. Given the complexity of elements present in social media posts, it is necessary to cautiously outline the elements that may influence user engagement metrics. This need was reinforced by Campbell and Rudan [12] on social media health campaigns, which emphasized the significance of tracking user engagement metrics through an investigation of elements like video posting features.

Numerous individual studies have examined the elements in social media posts made by HCPs that influenced user engagement metrics [10,11,13-16]. Each study has focused on different elements that affect user engagement metrics. For example, Kite et al [14] explored the effects of post timing on user engagement, which were not investigated in other similar studies that examined posts on dietary habits [11,15]. Therefore, it would be beneficial to collate the findings from these individual studies into a comprehensive review of the elements that influence user engagement metrics. This would assist HCPs in prioritizing their health promotion strategies on social media to achieve favorable user engagement. However, there have been no known reviews that identified the elements in the context of social media posts related to the risk reduction for NCDs. Therefore, a systematic review was conducted to identify elements influencing user engagement metrics in social media posts aimed at reducing lifestyle risk factors.

Methods

Study Protocol

This systematic review was conducted according to the Cochrane recommendations and was reported in accordance with the updated guidelines of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [17,18]. The review protocol was registered in the International Prospective Register of Systematic Review (PROSPERO; February 27, 2023, registration number CRD42023400177) [19].

Study Design

This review included all types of study designs published in the English language, which are original research reported in peer-reviewed journals.

Eligibility Criteria

Studies were selected according to PICO (Population, Intervention, Comparison, Outcome) criteria [18]. The selection of studies is outlined in Table 1.

XSL•FO
RenderX

Table 1. Selection of studies according to PICO[a] criteria.

Criteria	Description
Population	• The population included social media users of all age groups. They were either population with or without health conditions, or population with health conditions not mentioned.
Intervention	• Inclusion criteria of social media posts: • Posts delivered on existing, commercial social media platforms (eg, Facebook, Instagram, and Twitter). • Posts that were aimed to reduce any of the 4 key lifestyle risk factors under WHO's[b] health priorities (ie, tobacco use, harmful use of alcohol, unhealthy diet, and physical inactivity). • Post creators were health care professionals, represented individually or as part of an organization. They should be part of a health workforce in various health settings, such as hospitals, clinics, community health centers, research institutions, academic institutions, and health organizations. • Posts were examined for elements that were linked to user engagement metrics. • Exclusion criteria of social media posts: • Posts on existing, commercial social media platforms but engagement functions were disabled (eg, Facebook posts with disabled comment functions). • Post creators who were not formally trained in health care (eg, media companies and celebrities).
Comparison	• Studies with (ie, a traditional control group or an alternative intervention) or without (ie, analysis of existing social media posts) comparators were included.
Outcome	• Inclusion criteria of study outcomes: • Outcomes that were user engagement metrics involving direct post interactions and were reported numerically: • Number of post interactions such as likes, shares, comments, or emojis (eg, a post had 10 likes, 2 emojis, 2 comments, and 1 share). • Number of unique users performing a series of post interactions such as likes, emojis, comments, or shares (eg, a post had 30 interactions in total with 10 unique users, some users may like and comment in the same post). • Exclusion criteria of study outcomes: • Outcomes that did not involve direct post interactions, such as reach (ie, number of people viewing a post) or impressions (ie, number of post views). • Outcomes that were self-reported by respondents (eg, surveys).

[a]PICO: Population, Intervention, Comparison, Outcome.

[b]WHO: World Health Organization.

The population included social media users of all age groups who assessed the social media posts that were delivered on existing, commercial social media platforms such as Facebook and Instagram (Meta Platforms). Social media posts were aimed at reducing any of the 4 key lifestyle risk factors, and the posts were created by HCPs. In our study, HCPs represented part of the health workforce in various health settings such as hospitals, clinics, community health centers, research institutions, academic institutions, and health organizations. The social media posts in the included studies were examined for elements that were linked to user engagement metrics. The term "elements" in this review refers to all the components found in social media posts that could be deduced either from the outlook of the post itself (eg, image and poll) or from its content (eg, informative post). User engagement metrics included direct interactions performed by social media users toward the posts that were reported numerically [20,21]. User engagement metrics were restricted to quantifiable postlevel interactions as they represent the most objective and interpretable measures to compare study outcomes over time across published studies [22].

Search Strategy

A search strategy comprising controlled vocabulary (eg, MeSH [Medical Subject Headings]) and free text terms informed by previous literature [21,23,24] was developed and reviewed by 2 authors (YYY and WWC). The search strategy was structured into 4 concept headings which are elements in social media posts and their derivative terms, social media platforms, lifestyle risk factors, and outcome measures.

A literature search was conducted in 4 electronic health databases (MEDLINE or OVID, Scopus, Web of Science, and CINAHL) using the designated search strategy to identify relevant studies for inclusion. Multimedia Appendix 1 outlines the search strategies for all 4 databases. Given the focus on peer-reviewed studies, a gray literature search was not conducted. The search was restricted to studies published in English, between January 2006 (the year when X (previously known as Twitter; Twitter, Inc) and Facebook were publicly accessible, according to the review by Chen and Wang [23] until June 2023. Additionally, OVID auto alerts were used to monitor and include any newly published papers until March 31, 2024. The references from included studies and related

XSL•FO
RenderX

systematic reviews were also screened to identify further eligible studies.

Study Selection

The initial systematic literature search involved a single researcher (YYY) who screened and reviewed the titles and abstracts. Full-text papers from potentially relevant studies were retrieved and assessed for eligibility based on the inclusion criteria. Subsequently, another researcher (WWC) reviewed all included studies. Any discrepancies or disagreements were resolved through collaborative discussions among the study authors. The title, abstract, and full-text screening were completed on Rayyan (Rayyan Systems, Inc) [25].

Data Extraction

A standardized Microsoft Excel sheet was developed for data extraction. Extracted data included (1) publication details (author, publication year, and country); (2) study design; (3) target population; (4) sample size; (5) description of posts (lifestyle risk factor, social media and post creator, and number and duration of posts); (6) a brief description of methods involving delivery of posts; (7) investigated elements and user engagement metrics; and (8) elements that influenced user engagement metrics.

A single researcher (YYY) extracted data from the included studies. Accuracy was ensured by cross-checking the extracted data with another researcher (WWC). Any discrepancies were discussed and resolved through consensus among the study authors.

Data Analysis

The investigated elements encompassed all elements in included studies that examined social media posts, that may or may not have influenced user engagement. Due to the variability of the investigated elements, a narrative synthesis was performed to meticulously synthesize the findings from the included studies. All investigated elements were categorized based on their descriptions as reported in the studies. The initial categorization was carried out by the first researcher (YYY) and was further refined by a second researcher (WWC). The finalized categories were collectively reviewed and assessed by all authors, with any discrepancies resolved through consensus.

The included studies were then reviewed to identify the elements in social media posts that influenced user engagement metrics. The identified elements that influenced user engagement were determined based on 2 criteria—those demonstrating the highest measured user engagement and those for which user engagement was reported as significant during univariate or multivariate analysis.

Quality Assessment

As the studies varied in research design, methodological quality assessment was conducted using appraisal tools according to each research design. Joanna Briggs Institute (JBI) critical appraisal tools were used for cross-sectional studies, randomized trials, and quasi-experimental studies [26] whereas, the Mixed Methods Assessment Tool (MMAT) was used for mixed methods studies [27].

The methodological quality of included studies was tabulated with responses for each item assigned (yes, no, unclear, or not applicable for JBI critical appraisal tools and yes, no, or cannot tell for MMAT). Quality assessment was completed by the first researcher and was checked by a second researcher.

Results

Overview

A total of 2458 studies were identified from the database search, OVID auto alerts and citation search. The full text of 242 papers was assessed for eligibility. A total of 19 studies met the inclusion criteria and were included in the systematic review. Figure 1 outlines the selection of eligible studies.

XSL•FO
RenderX

Figure 1. PRISMA flowchart for selection of eligible studies. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

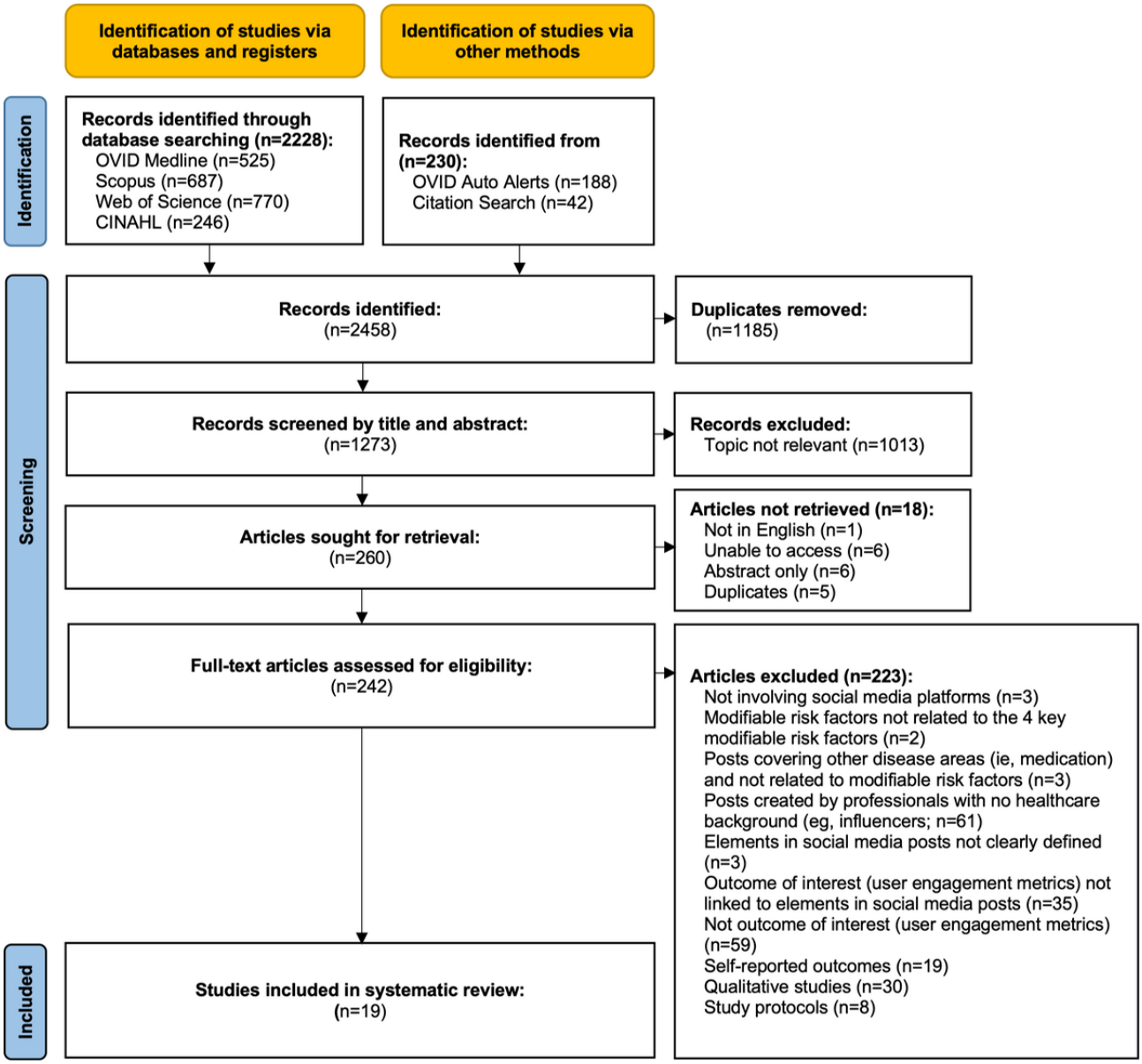

Description of Studies

The reviewed studies are summarized in Multimedia Appendix 2. The study designs for included studies were cross-sectional (n=10) [14,28-36], mixed methods (n=3) [15,37,38], quasi-experimental (n=2) [39,40], and randomized trials (n=4) [11,41-43]. A total of 3 out of 4 randomized trials [11,41,43] were subgroup analyses of previously published trials [44-46]. Studies included were from North America (n=10) [11,15,31,34,35,37,39,40,42,43], Australia (n=3) [14,29,41], Europe (n=3) [28,33,36], Great Britain (n=1) [38], South America (n=1) [32], and Asia (n=1) [30]. All 19 studies depicted posts promoting healthy lifestyle behaviors in reducing risk factors for NCDs. In total, 18 studies targeted healthy populations and 1 study aimed at reducing NCD-related complications among diabetic patients [28]. The examined topics included tobacco-related posts (n=11) [29-36,39,40,42], behaviors to reduce obesity (n=4) [11,14,37,43], posts promoting physical activities (n=2) [28,41], dietary habits (n=1) [38] and multiple lifestyle behaviors (n=1) [15]. The social media platforms used were Facebook (n=15) [11,14,15,29,31-33,35-37,39-43], Weibo (Sina Corporation; n=1) [30], and Instagram (n=1) [38] with 2 studies using multiple platforms of Facebook, Instagram, and Twitter [28,34]. The average number of posts examined was 710 (range 2-3515). In terms of the delivery of posts, 10 studies had posts with elements categorized before being posted on social media [11,15,29,34,37,38,40-43], whereas 7 studies collected existing social media posts and subsequently categorized the elements in the posts [14,28,30,31,33,35,36]. Two studies had posts with elements categorized before being posted as paid Facebook ads [32,39].

Study Quality

Quality assessment for included studies is outlined in Multimedia Appendix 3. All studies had high or moderate quality, having at least 50% of items labeled as “Yes” (4 out of 8, 50% items for cross-sectional studies; 7 out of 13, 54% items for randomized trials; 5 out of 9, 56% items for quasi-experimental studies; and 9 out of 17, 43% items for

XSL•FO
RenderX

mixed methods studies). All cross-sectional studies had valid and reliable measures of exposures and outcomes, with appropriate statistical analysis used. A total of 4 cross-sectional studies addressed confounders through the conduct of multivariate regression analyses [14,29,31,34]. Almost all randomized trials did not blind their participants, investigators, and outcome accessors, with the exception of Tomayko et al [43] that blinded its participants. Both quasi-experimental studies did not have any control groups present [39,40]. Among the mixed methods studies, only 1 study addressed confounders [37].

Study Findings

Table 2 describes the categories of elements in social media posts and user engagement metrics. The elements were grouped into 9 categories based on their characteristics. These categories were communication using supportive or emotive elements, post appearance, communication toward behavioral changes, post topics, requests for direct interaction with the post, tailoring of post content toward the targeted audience, source of post content, social media platform, and day and time of post. A total of 6 types of user engagement metrics were identified, which are likes, comments, shares, emojis, clicks, and votes.

Table 2. Description of categories of elements in social media posts and user engagement metrics.

Term	Description
Categories of elements in social media posts	
Communication using supportive or emotive elements	• Communication using supportive elements that involves the provision of assistance or comfort by post creators to users. Support may arise from relationships within users' social network. These supportive elements can be informational (eg, informative), companionship (eg, networking support), emotional (eg, empathy and concern), tangible assistance (eg, diabetic health tools), or intangible assistance (eg, counseling) [47]. • Communication utilising emotive elements are posts that express a fact or situation which elicits positive or negative feelings among users. It may involve the inclusion of humorous elements (sarcasm, jokes, meme, and popular trending culture) to generate positive feelings or fearful statements to generate negative feelings [10].
Post appearance	• The way a post is presented to users. It indicates the outward appearance of the post (eg poll, video, image, and text).
Communication toward behavioral changes	• Communication between post creators and users that is driven toward users' outcomes (ie, behavioral changes). Statements on intended behavioral changes should be clearly mentioned in the posts. Elements may also incorporate existing behavioral models or theories to elicit changes among users.
Post topics	• Elements that are thematic related, exhibiting topics on lifestyle behaviors which reduces risk factors for noncommunicable diseases.
Requests for direct interaction with the post	• Elements that prompt users to interact directly with the social media post through performing an action related to user engagement (eg, post likes, shares, and comment).
Tailoring of post content toward targeted audience	• Elements that allow content of the post to be aimed at the targeted population group.
Source of post content	• Source of the content of the post. Describes where the content of the post is obtained from (either completely created from scratch or adopted from other sources with the source mentioned).
Social media platform	• Platform used for posting.
Day and time of post	• The day and time when the post is posted on social media.
User engagement metrics	
Likes	• Represents users' actions of showing approval or appreciation of the post, through a "thumbs up."
Comments	• Users' responses to the post in the form of comments or replies.
Shares	• Users' actions of distributing or sharing the post.
Clicks	• Encompass 2 actions—clicking on a post that leads to an external link unrelated to the social media platform and clicking on a video post to initiate playback.
Emojis	• Emoticon reactions provided to the post, offering alternative reactions (eg, sad, haha, wow, love, and angry).
Votes	• Users' reaction toward a poll post, by selecting one of the options available.

XSL•FO
RenderX

Multimedia Appendix 4 shows the categories of elements and user engagement metrics tabulated according to the study. A total of 14 studies [11,14,15,28-31,33,36-39,41,43] reported likes as outcomes of user engagement metrics, with the other metrics involved being comments (n=13) [11,14,15,28,30,31,33,35,37,39-42], shares (n=8) [14,28-31,33,37,39], emojis (n=4) [14,29,36,39], clicks (n=4) [14,32,34,39], and votes (n=3) [11,37,41] (see Multimedia Appendix 4).

The breakdown of investigated elements for each study with the subsequent elements that influenced user engagement metrics is outlined in Multimedia Appendix 5. The PRISMA 2020 checklist for the systematic review is provided in Multimedia Appendix 6. A total of 35 elements from the 9 categories were found to influence user engagement metrics. Table 3 summarizes the definitions of all 35 elements that influenced user engagement metrics.

XSL•FO
RenderX

Table 3. Definition of elements in social media posts that influenced user engagement metrics.

Element	Definition	Studies with elements that influenced user engagement
Category 1: communication using supportive or emotive elements		
Informative post	• Providing information or advice, which might be directed toward the user (eg, some of the difficulties of quitting smoking include lack of peer support) or directed to others (eg, protect your loved ones from smoking) [28,35].	[15,30,31,35,38]
Provide networking support	• Allowance of communication among users with similar concerns and experiences [28].	[28,30,31,38]
Provide assistance	• Providing or offering to provide assistance in the form of tangible assistance in the form of goods and services (eg, diabetic health tools) [28] or intangible assistance (eg, live counseling support) [42,47].	[28,42]
Post tagging health organization	• A post with social media accounts of health organizations tagged, allowing users to obtain support from the tagged health organizations [29].	[29]
Post carrying humor	• Information with the inclusion of contemporary humorous elements such as sarcasm, jokes, and memes that may leverage on popular culture or trending events [38].	[38]
Post carrying negative emotional appeal	• A post expressing a fact, situation, or experience which elicits unpleasant and negative feelings (fear and anger) among users (eg, banning flavored vapes only pushes sales to black market) [14,31].	[31]
Category 2: post appearance		
Poll	• A post that allows votes to be casted on social media.	[11,15,37,41]
Video	• A post where moving visual images with an audio component are present.	[15,29,31,39]
Image	• A post where static images in the form of photos (eg, photographed by camera) or graphics (eg, visualization) are present.	[14,37]
Category 3: communication toward behavioral changes—the usage of behavioral models, with constructs listed are those that influence user engagement		
Health Belief Model (HBM; focuses on user's belief in negative consequences together with beliefs in the effectiveness of the recommended health behavior or action will predict the likelihood the person will adopt the behavior [30])	• Perceived risks: user's subjective perception of the negative consequences of the behavior. • Self-efficacy: level of a user's confidence in his or her ability to successfully perform a behavior.	[30]
Theory of Planned Behavior (TPB; focuses on user's intention to engage in a behavior at a specific time and place [30])	• Subjective norms: the belief about whether most users approve or disapprove of the behavior (eg, netizen's voice which opposes a behavior).	[30]
Transtheoretical Model (TTM; focuses on user's decision-making and is a model of intentional change [40,42])	• Decisional balance: pros and cons of behavior and behavioral changes. • Consciousness raising: increasing awareness, through new facts, ideas, and tips about the healthy behavior. • Dramatic relief: experiencing negative emotions which go along with old behaviors and positive emotions which go along with new behaviors. • Self-liberation: making a firm commitment to change behavior based on the belief that achievement of the healthy behavior is possible.	[40[a],42]
Category 3: communication toward behavioral changes—other elements		
Post exhibiting call-to-action	• A post emphasizing or prompting user to undertake a specific behavior (eg, let's exercise, please eat healthily) or a specific action (eg, call a helpline, register for an event, and join a webinar) [14].	[14]

XSL•FO
RenderX

Element	Definition	Studies with elements that influenced user engagement
Loss-framed post	• A post that is framed toward the loss of an outcome due to user's behavioral action (eg "You will die sooner if you do not quit smoking" indicates users losing their life) [32].	[32]
Motivational interviewing strategies	• A person-centered counseling approach that prepares users for changes by helping users to resolve their ambivalence toward a problem [33].	[33]
Category 4: post topics—risk reduction		
General well-being	• A post focusing on a variety of topics promoting healthy lifestyle.	[38]
Category 4: post topics—healthy diet		
Diet or recipe	• A post that shares recipe or meal ideas toward a healthy diet [43].	[43]
Drinking water	• A post that indicates water as the drink of choice and suggests to decrease sugar-sweetened beverage consumption [14].	[14]
Nutrition news	• A post that shares nutrition fact or nutrition news items, such as results from a nutrition study or food containing certain vitamins or supplements (eg, kale having vitamin B12) [11].	[11]
Weight loss	• A post that asks users undergoing a diet program to discuss weight loss tips and progress [11].	[11]
Category 4: post topics—physical activity		
Physical activity promotion	• A post that promotes physical activity [28].	[28,38,43]
Category 5: requests for direct interaction with the post		
Suggestion	• A post that asks users to directly post a tip (as comments under the post) to help other users [11].	[11]
Discussion question	• A post that encourage direct discussion or sharing of responses by asking questions [15,35].	[15,35]
Statement with "engagement bait" (post action clearly stated)	• Post strategies that lead users to interact through likes, shares, and comments (eg, "Like this!" or "Share this!") [36].	[36][a]
Category 6: tailoring of post content toward targeted audience		
Usage of localized branding	• A post that targets a specific community, by using local branding, such as the logo of a nearby clinic or the language and dialect commonly used in a community as compared to generalized branding, such as the logo of the ministry of health of a country, or the English language in a community where it is not the primary language [29].	[29]
Usage of hashtags	• A post that uses metadata tags, prefaced by the hash symbol, # to allow cross-referencing of a data, and encourage user searchability by a topic or theme.	[29][a]
Paid post	• A boosted post on social media platforms through paid advertising, which uses algorithms to target specific demographics [32,39] and interests or to target a wider audience by increasing post visibility [14].	[14]
Organic post	• A post that is shared on social media platform without payment to the platform.	[34]
Category 7: source of post content		

XSL•FO
RenderX

Element	Definition	Studies with elements that influenced user engagement
Original content not published before	• Content of the post is created entirely from scratch by the health individual or organization involved in posting the post to social media platform [30].	[30]
Content adopted from other sources	• Content of the post is adopted from other sources that have been previously published. Content was either posted as is, or modified with or without attribution to the original source stated in the post [29].	[29]
Category 8: social media platform		
Instagram	• A post that is posted on the Instagram platform.	[28,34]
Facebook	• A post that is posted on the Facebook platform.	[28]
Category 9: date and time of post		
Monday	• A post that is posted on Monday.	[14][a]
Friday	• A post that is posted on Friday.	[14]
8 AM to 5 PM	• A post that is posted between 8 AM and 5 PM.	[14]

[a]Elements for which user engagement was reported as significant during univariate or multivariate analysis, with the elements showing a significant decrease in user engagement.

Communication using supportive or emotive elements and the appearance of posts were the 2 most featured categories with elements influencing user engagement, each appearing in 8 studies [11,14,15,28-31,35,37-39,41,42]. Among the elements under supportive or emotive communication, the informative post was the most popular element (n=5) [15,30,31,35,38]. Networking with other users was also an effective way of delivering support (n=4) [28,30,31,38], as was providing tangible and intangible assistance (n=2) [28,42]. Regarding the appearance of posts, polls (n=4) [11,15,37,41] and videos (n=4) [15,29,31,39] were the 2 most featured elements (n=4), followed by images (n=2) [14,37].

Communication elements leading to behavioral changes influenced user engagement in 6 studies [14,30,32,33,40,42]. The elements were predominantly based on various behavioral models, including the Health Belief Model (HBM) [30], the Theory of Planned Behavior (TPB) [30], and the Transtheoretical Model (TTM) [40,42]. Additionally, 3 studies used non-model elements such as call-to-action, message framing toward loss of outcomes, and person-centered motivational interviewing approaches [14,32,33].

The results of 5 studies show that topic-based elements focusing on positive lifestyle behaviors for risk reduction can increase user engagement [11,14,28,38,43]. These topics include encouraging dietary habits, such as drinking water (n=3) [11,14,43], promoting physical activity through exercise (n=3) [28,38,43], and general well-being topics (n=1) [38].

Social media posts that requested users to interact directly with the post influenced user engagement in 4 studies [11,15,35,36]. Posts that asked questions and encouraged discussion were found to promote engagement in 3 studies (n=3) [11,15,35]. However, engagement baits such as “like this post!” were found to decrease engagement in 1 study [36].

The tailoring of post content to targeted audience groups was shown to influence user engagement in 3 studies [14,29,34]. According to Hefler et al [29], posts that incorporated local branding to generate trust among the local community resulted in increased user engagement. However, the use of hashtags marked by “#” to enhance post searchability showed an unusual trend of decreasing user engagement. On the other hand, amplifying post reaches through payment, as demonstrated by Kite et al [14] led to increased user engagement. In Reuter et al [34], unpaid, organic posts were preferred by users.

The source of post content has been found to impact user engagement in 2 studies [29,30]. According to Jiang and Beaudoin [30], original and unpublished content has a positive influence on user engagement. Hefler et al [29] observed positive user engagement in posts containing externally sourced content from previously published material, whether presented unchanged or with minor modifications.

The selection of social media platforms was found to increase user engagement in 2 recent studies examining similar posts on multiple platforms [28,34]. Facebook and Instagram were found to be the preferred platforms in these studies. Additionally, the timing of postings was shown to have an impact on user engagement. Kite et al [14] found that posts made on Fridays between 8 AM and 5 PM received higher user engagement, whereas posts made on Mondays generated lower user engagement.

Discussion

Principal Findings

To the best of our knowledge, this is the first systematic review to identify elements in social media posts that influenced user engagement metrics. This study addressed the current knowledge gaps related to how HCPs can create posts for optimal user

XSL•FO
RenderX

engagement. Our analysis has identified 35 elements across 9 categories that influence user engagement metrics. Communication elements with supportive and emotive elements that encourage behavioral changes, as well as the appearance of posts were the dominant categories that have a positive impact on user engagement. By prioritizing these elements, we can potentially maximize the effects of health promotion by HCPs through social media. However, the categories of source of post content, social media platforms, and post timings had less than 3 studies showing elements that affect user engagement. Therefore, more studies are needed to confirm the findings in relation to these elements.

It is worth noting that at least three-quarters of the studies on social media posts were conducted in high-income countries, which is not surprising since these countries have more developed digital information infrastructure [48]. Our findings mirrored the review by Elaheebocus et al [49] that focused on targeted behaviors on social media, where approximately half of the studies examined tobacco-related behaviors, while the other half focused on physical activity promotion, healthy dietary habits, and weight control. Interestingly, there were no studies on alcohol-related posts, despite the fact that marketing brands tend to promote alcohol more frequently on social media [50]. This is concerning because HCPs are already sharing limited health-related information on social media platforms [51]. It may be important to increase health promotion pertaining to reducing and abstaining from alcohol consumption on social media platforms.

Elements Influencing User Engagement Metrics

Communication using supportive or emotive elements was one of the categories with the most studies on elements that affect user engagement. In our review, most of the supportive elements that led to high user engagement were provided through informative posts, audience support, and assistance such as live counseling support [42] and diabetic health tools [28]. Tagging other health organizations in posts increases users' trust and allows them to reach out for external assistance. Emotive elements such as humorous posts that adopted contemporary memes also yielded higher user engagement. However, they need to be used alongside informative elements [38]. In addition, posts that evoke negative or unpleasant feelings were also found to increase user engagement. This is supported by evidence in experimental psychology, whereby the presence of negativity bias leads to greater user responsiveness toward negative stimuli [52] resulting in higher postlevel interactions. The context in these posts may still be aligned with positive lifestyle behaviors. For example, the post "Banning vape leads to black market sales which is worrying" shows that the user is worried despite agreeing with the vape ban [31].

The appearance of posts is the second most studied factor influencing user engagement. Polls, videos, and images were the 3 elements grouped under this category. The conduct of polls positively influenced user engagement as users can view the immediate results and participate in discussions under the poll post, leading to higher comments. Photos are static visuals with straightforward messages that draw users' immediate attention and positively influence user engagement [39]. Despite videos requiring more attentive processing, they still have a positive influence on user engagement [14,29,31,39]. With the introduction of shorter videos that require less attention span, it is expected that videos will continue to increase user engagement.

Elements of communication toward behavioral changes that influenced user engagement were mostly based on existing behavioral models. The 3 models studied included the HBM, the TPB, and the TTM [30,40,42]. The health behavioral models were used in the creation of posts on smoking prevention and cessation, with 2 studies using the TTM [40,42]. In the TTM, decisional balance focuses on the advantages and disadvantages of behavioral changes. Decisional balance is consistent with motivational interviewing techniques, whereby mixed feelings are acknowledged before users are guided toward the advantages of behavioral changes [33,40]. Such elements are effective among younger smokers with lower motivation to quit [40,53]. Posts based on consciousness raising in the TTM were more effective among smokers in the midst of quitting and who required additional health information in quitting [40]. Although TTM elements of dramatic relief (ie, eliciting negative emotional responses to old behaviors of smoking and positive emotional responses to new behaviors of quitting smoking) and self-liberation (ie, firm commitments to quitting) reduced engagement levels, authors in the study were optimistic that reduced engagement could be reversed by eliciting positive emotional responses to new behaviors instead of negative emotional responses and restrategizing self-liberation elements into more organized, incremental methods [40].

Other elements of communication toward behavioral changes that positively influenced user engagement included call-to-action and loss-framed posts [14,32]. Findings of loss-framed posts influencing user engagement by mentioning the losses of a behavioral outcome were supported by Graham et al's [54] study assessing smoking cessation advertisements delivered through websites. However, gain-framed messages were still effective in studies exploring messages that convinced active smokers to quit smoking [55,56]. The effects of framed messages should, therefore, be explored further.

In several studies, 3 categories of elements—post topics, requests to interact directly with the post and targeted content have been found to influence user engagement. First, among the post topics, instructional posts delivered directly (eg, exercise more) or indirectly (eg, recipe ideas toward a healthy diet) have been shown to increase user engagement [11,14,28,38,43]. Instructional posts simplify users' comprehension of what actions are required and how to perform them [57]. Second, questions asking for suggestions or discussion act as a cue for users to directly comment on the post [11,15,35]. Users are also inclined to further engage with the post by liking or sharing them. In contrast, "engagement bait" strategies deployed on Facebook using phrases such as "like this!" have been found to reduce user engagement. The authors hypothesized that lower engagement was due to Facebook demoting such posts from users' newsfeeds, causing them to appear less frequently on users' social media timelines [36,58]. Third, tailoring content toward targeted audience groups has also been found to increase user engagement. Information

XSL•FO
RenderX

delivery to a specific community should include a personalized touch by incorporating local elements, such as community logos or local dialects [29]. The inclusion of hashtags in posts can increase their discoverability, making it easier for users to search for them. However, Hefler et al [29] observed that adding hashtags to posts shared from other pages or profiles might lead to lower user engagement. This is because hashtags on shared posts may increase the length of a post, which might reduce its legibility and aesthetic appeal. As both organic and paid posts positively impacted user engagement [14,34], selection for payment should be based on budget availability and target audience group.

The source of post content and social media platforms were found by 2 studies to influence user engagement metrics. The adoption of post content from previously published sources would mean that a meticulous selection of high-quality content has been made, which generated higher user engagement [29]. A study by Waters and Jamal [59] has shown a higher reliance for government and nonprofit organizations to adopt content from external sources, as they are usually less creative in generating their own content as compared to corporate and consumer-driven companies. This postulation is applicable to posts by health organizations which are usually either government or nonprofit-based. In contrast to the study by Hefler et al [29], Jiang and Beaudoin [30] likened an increase in engagement with original, unpublished content created entirely by the health care individual or team involved in posting the content. The preference for original content is due to the novelty of the posts as they have not been viewed by users previously. Despite the potential impact of both original and adopted content on user engagement, original content is preferred when a creative team workforce is available.

Regarding social media platforms, Facebook and Instagram were both favored due to their function as networking sites, which are ideal for sharing ideas. Other microblog-based platforms, such as Twitter were less preferred as only short updates were allowed with limitations in post characters [60,61]. As for the timing of social media posts, only Kite et al [14] found elements that affected user engagement metrics. According to their study, postings made on Fridays generated more user engagement than those made on Mondays, possibly due to users spending more time browsing social media toward the end of the workweek. This finding is consistent with previous studies [62,63]. Interestingly, Kite et al [14] found that users preferred to view social media postings during work hours (between 8 AM and 5 PM), which contradicts previous research suggesting that users are more active on social media during the night [63]. Kite et al [14] suggested that users may feel more comfortable browsing health-related information during work hours.

Limitations

This systematic review is the first of its kind to focus on social media posts that were delivered by HCPs to identify the elements that influenced user engagement metrics. However, some limitations should be considered. First, there was a variability of study designs for included studies. Thus, a meta-analysis was not conducted. Nevertheless, quality assessment was conducted using critical appraisal tools according to the study design to evaluate the methodological quality of all included studies.

Second, the decision to only include studies that reported user engagement metrics as outcome measures, in the form of a direct action toward a post was made to facilitate objective deduction of measures and allowed comparisons to be made across published studies. We acknowledged that our exclusion of studies with outcome measures reliant on subjective assessments, such as psychometric scales might have resulted in overlooking certain findings. However, such subjective outcomes may vary according to individualized studies [64], and the data are prone to response bias [22].

Third, the information pertaining to the elements and the post creators for social media posts was based on the information provided in the papers. Some elements in the social media posts may not have been sufficiently explained, which may have caused limitations when elements were grouped into categories. For example, the element “informative post” categorized under communication using supportive or emotive elements might also have underlying elements of communication toward behavioral changes. The categorization of elements was refined through a collective discussion with all authors.

Fourth, social media posts from Facebook groups may only be accessed by the social media users who are subscribed to the groups. Despite the difference, studies that incorporated such posts were included in the review as the functionalities resemble those of Facebook pages, allowing users to interact directly with the posts.

Implication and Further Research

The review focused on studies examining social media posts on the reduction of lifestyle risk factors that were created by HCPs. Identifying elements that influence user engagement metrics would allow HCPs to have a greater understanding of the post features that are potentially favored by users. Implementation of such elements into future social media posts would empower the delivery of public health messages by HCPs.

Further research could be proposed to help strengthen the interpretations of elements that influence user engagement. Findings from this review mostly came from countries with highly developed digital media infrastructure. We may want to conduct similar experimental studies in less developed countries, to examine if similar elements would affect user engagement metrics. Furthermore, we recommend conducting more studies in areas that are still underresearched, based on the findings from this review. These areas include the source of post content, the choice of social media platform, and the timing of posts.

Conclusions

The systematic review outlined the prospects of effective health promotion by HCPs on social media in future postings. This is made possible by incorporating elements that have a positive impact on user engagement metrics. Positive user engagement metrics serve as an indication of a favorable response from users to the posts made by HCPs. Communication techniques that either used supportive or emotive elements, or techniques that

XSL•FO
RenderX

emphasized behavior changes were 2 of the most dominant categories of elements that could potentially maximize postlevel interactions. These communication elements should also be supported by paying attention to the appearance of each post. As social media continues to evolve, the elements in social media posts should be continuously evaluated, providing adjustments when required.

Acknowledgments

This research was funded by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2020/SS0/UKM/02/11). The funders played no role in study design, collection, analysis, interpretation of data, or writing of the report.

Data Availability

All data generated or analyzed during this study are included in this published article (and Multimedia Appendices 1-6).

Authors' Contributions

YYY and WWC have contributed to the conception or design of the study. YYY contributed to data collection and data analysis. All authors have contributed to data interpretation and have provided scientific inputs and technical improvement. YYY drafted the paper while WWC guided the revisions. All authors read and approved the final version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies for systematic review.
[DOCX File , 36 KB - jmir_v26i1e59742_app1.docx]

Multimedia Appendix 2
Summary of studies.
[DOCX File , 32 KB - jmir_v26i1e59742_app2.docx]

Multimedia Appendix 3
Quality assessment for included studies.
[DOCX File , 27 KB - jmir_v26i1e59742_app3.docx]

Multimedia Appendix 4
Categories of elements in social media posts and user engagement metrics.
[DOCX File , 27 KB - jmir_v26i1e59742_app4.docx]

Multimedia Appendix 5
Investigated elements and elements in social media posts that influenced user engagement metrics.
[DOCX File , 23 KB - jmir_v26i1e59742_app5.docx]

Multimedia Appendix 6
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 94 KB - jmir_v26i1e59742_app6.pdf]

References

1. Global social media statistics. Data Reportal. URL: https://datareportal.com/social-media-users [accessed 2024-04-21]
2. Gunasekeran DV, Chew A, Chandrasekar EK, Rajendram P, Kandarpa V, Rajendram M, et al. The impact and applications of social media platforms for public health responses before and during the COVID-19 pandemic: systematic literature review. J Med Internet Res 2022;24(4):e33680 [FREE Full text] [doi: 10.2196/33680] [Medline: 35129456]
3. Househ M. The use of social media in healthcare: organizational, clinical, and patient perspectives. Stud Health Technol Inform 2013;183:244-248. [Medline: 23388291]
4. WHO. Noncommunicable diseases. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2022-12-06]

XSL•FO
RenderX

5. Reducing modifiable risk factors for noncommunicable diseases. WHO. 2022. URL: https://www.who.int/westernpacific/activities/reducing-modifiable-risk-factors-for-noncommunicable-diseases [accessed 2022-12-06]
6. Islam SMS, Tabassum R, Liu Y, Chen S, Redfern J, Kim S, et al. The role of social media in preventing and managing non-communicable diseases in low-and-middle income countries: hope or hype? Health Policy Technol 2019;8(1):96-101. [doi: 10.1016/j.hlpt.2019.01.001]
7. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014;39(7):491-520 [FREE Full text] [Medline: 25083128]
8. Drivas IC, Kouis D, Kyriaki-Manessi D, Giannakopoulou F. Social media analytics and metrics for improving users engagement. Knowledge 2022;2(2):225-242. [doi: 10.3390/knowledge2020014]
9. Carah N, Angus D. Algorithmic brand culture: participatory labour, machine learning and branding on social media. Media, Culture Soc 2018;40(2):178-194. [doi: 10.1177/0163443718754648]
10. Kite J, Foley BC, Grunseit AC, Freeman B. Please like me: Facebook and public health communication. PLoS One 2016;11(9):e0162765 [FREE Full text] [doi: 10.1371/journal.pone.0162765] [Medline: 27632172]
11. Hales SB, Davidson C, Turner-McGrievy GM. Varying social media post types differentially impacts engagement in a behavioral weight loss intervention. Transl Behav Med 2014;4(4):355-362 [FREE Full text] [doi: 10.1007/s13142-014-0274-z] [Medline: 25584084]
12. Campbell IH, Rudan I. Effective approaches to public engagement with global health topics. J Glob Health 2020;10(1):01040901 [FREE Full text] [doi: 10.7189/jogh.10.010901] [Medline: 32257175]
13. Rahim AIA, Ibrahim MI, Salim FNA, Ariffin MAI. Health information engagement factors in Malaysia: a content analysis of Facebook use by the Ministry of Health in 2016 and 2017. Int J Environ Res Public Health 2019;16(4):591 [FREE Full text] [doi: 10.3390/ijerph16040591] [Medline: 30781691]
14. Kite J, Grunseit A, Li V, Vineburg J, Berton N, Bauman A, et al. Generating engagement on the make healthy normal campaign Facebook page: analysis of Facebook analytics. JMIR Public Health Surveill 2019;5(1):e11132 [FREE Full text] [doi: 10.2196/11132] [Medline: 31344679]
15. Lawton K, Hess L, McCarthy H, Marini M, McNitt K, Savage JS. Feasibility of using Facebook to engage SNAP-Ed eligible parents and provide education on eating well on a budget. Int J Environ Res Public Health 2022;19(3):1457 [FREE Full text] [doi: 10.3390/ijerph19031457] [Medline: 35162478]
16. Zhang Y, Xia T, Huang L, Yin M, Sun M, Huang J, et al. Factors influencing user engagement of health information disseminated by Chinese provincial centers for disease control and prevention on WeChat: observational study. JMIR Mhealth Uhealth 2019;7(6):e12245 [FREE Full text] [doi: 10.2196/12245] [Medline: 31250833]
17. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Chichester (UK): John Wiley & Sons; 2019:1119536626.
18. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. PROSPERO. Social media content by healthcare professionals on modifiable risk factors for non-communicable diseases: a systematic review of elements influencing user engagement. 2023. URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=400177 [accessed 2024-07-04]
20. Hootsuite. How to increase social media engagement: a guide for marketers. 2020. URL: https://blog.hootsuite.com/social-media-engagement/ [accessed 2022-12-06]
21. Trunfio M, Rossi S. Conceptualising and measuring social media engagement: a systematic literature review. Ital J Mark 2021;2021(3):267-292. [doi: 10.1007/s43039-021-00035-8]
22. O'Brien H. A holistic approach to measuring user engagement. In: Filimowicz M, Tzankova V, editors. New Directions in Third Wave Human-Computer Interaction: Volume 2 - Methodologies. Cham: Springer International Publishing; 2018:81-102.
23. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
24. Mita G, Ni Mhurchu C, Jull A. Effectiveness of social media in reducing risk factors for noncommunicable diseases: a systematic review and meta-analysis of randomized controlled trials. Nutr Rev 2016;74(4):237-247 [FREE Full text] [doi: 10.1093/nutrit/nuv106] [Medline: 26946250]
25. Faster systematic reviews. Rayyan. 2022. URL: https://rayyan.ai/ [accessed 2022-12-06]
26. Critical appraisal tools. JBI. 2022. URL: https://jbi.global/critical-appraisal-tools [accessed 2022-12-06]
27. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
28. Gabarron E, Larbi D, Årsand E, Wynn R. Engaging social media users with health education and physical activity promotion. Stud Health Technol Inform 2021;281:789-793. [doi: 10.3233/SHTI210283] [Medline: 34042686]
29. Hefler M, Kerrigan V, Grunseit A, Freeman B, Kite J, Thomas DP. Facebook-based social marketing to reduce smoking in Australia's first nations communities: an analysis of reach, shares, and likes. J Med Internet Res 2020;22(12):e16927 [FREE Full text] [doi: 10.2196/16927] [Medline: 33300883]

XSL•FO
RenderX

30. Jiang S, Beaudoin CE. Smoking prevention in China: a content analysis of an anti-smoking social media campaign. J Health Commun 2016;21(7):755-764. [doi: 10.1080/10810730.2016.1157653] [Medline: 27232655]
31. Lin SY, Cheng X, Zhang J, Yannam JS, Barnes AJ, Koch JR, et al. Social media data mining of antitobacco campaign messages: machine learning analysis of Facebook posts. J Med Internet Res 2023;25:e42863 [FREE Full text] [doi: 10.2196/42863] [Medline: 36780224]
32. Machado NM, Gomide HP, Bernardino HS, Ronzani TM. Facebook recruitment of smokers: comparing gain- and loss-framed ads for the purposes of an internet-based smoking cessation intervention. Cad Saude Publica 2019;35(10):e00151318 [FREE Full text] [doi: 10.1590/0102-311X00151318] [Medline: 31596400]
33. Pócs D, Óvári T, Watti J, Hamvai C, Kelemen O. How to create social media contents based on motivational interviewing approach to support tobacco use cessation? A content analysis. J Subst Use 2021;27(6):591-597. [doi: 10.1080/14659891.2021.1967484]
34. Reuter K, Wilson ML, Moran M, Le N, Angyan P, Majmundar A, et al. General audience engagement with antismoking public health messages across multiple social media sites: comparative analysis. JMIR Public Health Surveill 2021;7(2):e24429 [FREE Full text] [doi: 10.2196/24429] [Medline: 33605890]
35. Strekalova YA, Damiani RE. Message design and audience engagement with tobacco prevention posts on social media. J Cancer Educ 2018;33(3):668-672. [doi: 10.1007/s13187-016-1135-x] [Medline: 27832508]
36. Watti J, Millner M, Siklósi K, Hamvai C, Kelemen O, Pócs D. How to avoid lower priority for smoking cessation support content on Facebook: an analysis of engagement bait. Int J Environ Res Public Health 2023;20(2):958 [FREE Full text] [doi: 10.3390/ijerph20020958] [Medline: 36673713]
37. Merchant G, Weibel N, Patrick K, Fowler JH, Norman GJ, Gupta A, et al. Click "like" to change your behavior: a mixed methods study of college students' exposure to and engagement with Facebook content designed for weight loss. J Med Internet Res 2014;16(6):e158 [FREE Full text] [doi: 10.2196/jmir.3267] [Medline: 24964294]
38. O'Kane N, McKinley MC, Gough A, Hunter RF. Investigating the feasibility and acceptability of using Instagram to engage post-graduate students in a mass communication social media-based health intervention, #WeeStepsToHealth. Pilot Feasibility Stud 2022;8(1):254 [FREE Full text] [doi: 10.1186/s40814-022-01207-9] [Medline: 36510310]
39. Miller CA, Jung Kim S, Schwartz-Bloom RD, Bloom PN, Murphy SK, Fuemmeler BF. Informing women about the risks of exposing babies to tobacco smoke: outreach and education efforts using Facebook "boost posts". Transl Behav Med 2022;12(5):714-720 [FREE Full text] [doi: 10.1093/tbm/ibab158] [Medline: 35532323]
40. Thrul J, Klein AB, Ramo DE. Smoking cessation intervention on Facebook: which content generates the best engagement? J Med Internet Res 2015;17(11):e244 [FREE Full text] [doi: 10.2196/jmir.4575] [Medline: 26561529]
41. Edney S, Looyestyn J, Ryan J, Kernot J, Maher C. Posts, pics, or polls? Which post type generates the greatest engagement in a Facebook physical activity intervention? Transl Behav Med 2018;8(6):953-957. [doi: 10.1093/tbm/iby006] [Medline: 29660093]
42. Thrul J, Meacham MC, Tice C, Kelly O, Ramo DE. Live counselor contact in a Facebook intervention predicts smoking cessation outcomes. Psychol Addict Behav 2020;34(2):360-369 [FREE Full text] [doi: 10.1037/adb0000541] [Medline: 31829664]
43. Tomayko EJ, Webber EJ, Cronin KA, Prince RJ, Adams AK. Use of text messaging and Facebook groups to support the healthy children, strong families 2 healthy lifestyle intervention for American Indian families. Curr Dev Nutr 2021;5(Suppl 4):32-39 [FREE Full text] [doi: 10.1093/cdn/nzaa110] [Medline: 34222765]
44. Looyestyn J, Kernot J, Boshoff K, Maher C. A web-based, social networking beginners' running intervention for adults aged 18 to 50 years delivered via a facebook group: randomized controlled trial. J Med Internet Res 2018;20(2):e67 [FREE Full text] [doi: 10.2196/jmir.7862] [Medline: 29483065]
45. Tomayko EJ, Prince RJ, Cronin KA, Parker T, Kim K, Grant VM, et al. Healthy children, strong families 2: a randomized controlled trial of a healthy lifestyle intervention for American Indian families designed using community-based approaches. Clin Trials 2017;14(2):152-161 [FREE Full text] [doi: 10.1177/1740774516685699] [Medline: 28064525]
46. Turner-McGrievy GM, Davidson CR, Wilcox S. Does the type of weight loss diet affect who participates in a behavioral weight loss intervention? A comparison of participants for a plant-based diet versus a standard diet trial. Appetite 2014;73:156-162 [FREE Full text] [doi: 10.1016/j.appet.2013.11.008] [Medline: 24269507]
47. APA dictionary of psychology - social support. APA. 2023. URL: https://dictionary.apa.org/social-support [accessed 2023-10-07]
48. Hagg E, Dahinten VS, Currie LM. The emerging use of social media for health-related purposes in low and middle-income countries: a scoping review. Int J Med Inform 2018;115:92-105 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.04.010] [Medline: 29779724]
49. Elaheebocus SMRA, Weal M, Morrison L, Yardley L. Peer-based social media features in behavior change interventions: systematic review. J Med Internet Res 2018;20(2):e20 [FREE Full text] [doi: 10.2196/jmir.8342] [Medline: 29472174]
50. Lim MS, Hare JD, Carrotte ER, Dietze PM. An investigation of strategies used in alcohol brand marketing and alcohol-related health promotion on Facebook. Digit Health 2016;2:2055207616647305 [FREE Full text] [doi: 10.1177/2055207616647305] [Medline: 29942554]

XSL•FO
RenderX

51. Room R, O'Brien P. Alcohol marketing and social media: a challenge for public health control. Drug Alcohol Rev 2021;40(3):420-422 [FREE Full text] [doi: 10.1111/dar.13160] [Medline: 33111397]
52. Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. Pers Soc Psychol Rev 2001;5(4):296-320. [doi: 10.1207/S15327957PSPR0504_2]
53. Brown RA, Ramsey SE, Strong DR, Myers MG, Kahler CW, Lejuez CW, et al. Effects of motivational interviewing on smoking cessation in adolescents with psychiatric disorders. Tob Control 2003;12 Suppl 4(Suppl 4):IV3-I10 [FREE Full text] [doi: 10.1136/tc.12.suppl_4.iv3] [Medline: 14645934]
54. Graham AL, Fang Y, Moreno JL, Streiff SL, Villegas J, Muñoz RF, et al. Online advertising to reach and recruit Latino smokers to an internet cessation program: impact and costs. J Med Internet Res 2012;14(4):e116 [FREE Full text] [doi: 10.2196/jmir.2162] [Medline: 22954502]
55. Arendt F, Bräunlein J, Koleva V, Mergen M, Schmid S, Tratner L. Effects of gain- and loss-framed quit messages on smokers: test of the ability to process the health message as a moderator. J Health Commun 2018;23(8):800-806. [doi: 10.1080/10810730.2018.1527878] [Medline: 30300104]
56. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
57. Blaich C, Wise K, Pascarella ET, Roksa J. Instructional clarity and organization: it's not new or fancy, but it matters. Change: Mag Higher Learn 2016;48(4):6-13. [doi: 10.1080/00091383.2016.1198142]
58. Meta. Meta business help centre: how to avoid posting engagement bait on Facebook. 2023. URL: https://www.facebook.com/business/help/259911614709806?id=208060977200861 [accessed 2023-03-30]
59. Waters RD, Jamal JY. Tweet, tweet, tweet: a content analysis of nonprofit organizations' Twitter updates. Public Relat Rev 2011;37(3):321-324. [doi: 10.1016/j.pubrev.2011.03.002]
60. Paul M, Dredze M. You are what you Tweet: analyzing Twitter for public health. Proc Int AAAI Conf Weblogs Soc Media 2021;5(1):265-272.
61. Zhou L, Zhang D, Yang C, Wang Y. Harnessing social media for health information management. Electron Commer Res Appl 2018;27:139-151 [FREE Full text] [doi: 10.1016/j.elerap.2017.12.003] [Medline: 30147636]
62. Whiting A, Williams D. Why people use social media: a uses and gratifications approach. Qualitative Mrkt Res: An Int J 2013;16(4):362-369.
63. Michael K. Australian social media habits. 2020. URL: https://www.sensis.com.au/australian-social-media-habits [accessed 2022-12-06]
64. Sigerson L, Cheng C. Scales for measuring user engagement with social network sites: a systematic review of psychometric properties. Comput Hum Behav 2018;83:87-105. [doi: 10.1016/j.chb.2018.01.023]

Abbreviations

HBM: Health Belief Model
HCP: health care professional
JBI: Joanna Briggs Institute
MeSH: Medical Subject Headings
MMAT: Mixed Methods Assessment Tool
NCD: noncommunicable disease
PICO: Population, Intervention, Comparison, Outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Review
TPB: Theory of Planned Behavior
TTM: Transtheoretical Model
WHO: World Health Organization

Edited by A Coristine; submitted 26.04.24; peer-reviewed by A Al Hamid, P Iranmanesh; comments to author 04.07.24; revised version received 15.07.24; accepted 26.08.24; published 22.11.24.

Please cite as:
Yip YY, Makmor-Bakry M, Chong WW
Elements Influencing User Engagement in Social Media Posts on Lifestyle Risk Factors: Systematic Review
J Med Internet Res 2024;26:e59742
URL: https://www.jmir.org/2024/1/e59742
doi:10.2196/59742
PMID:

XSL•FO
RenderX

©Yan Yee Yip, Mohd Makmor-Bakry, Wei Wen Chong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Artificial Intelligence Applications to Measure Food and Nutrient Intakes: Scoping Review

Jiakun Zheng[1*], PhD; Junjie Wang[2*], PhD; Jing Shen[3], PhD; Ruopeng An[4], MPP, PhD

[1]School of Economics and Management, Shanghai University of Sport, Shanghai, China
[2]School of Kinesiology and Health Promotion, Dalian University of Technology, Dalian, China
[3]Department of Physical Education, China University of Geosciences (Beijing), Beijing, China
[4]Silver School of Social Work, New York University, New York, NY, United States
[*]these authors contributed equally

Corresponding Author:
Jiakun Zheng, PhD
School of Economics and Management
Shanghai University of Sport
399 Changhai Road, Yangpu District
Shanghai, 200438
China
Phone: 86 13817507993
Email: zhengjiakun07@163.com

Abstract

Background: Accurate measurement of food and nutrient intake is crucial for nutrition research, dietary surveillance, and disease management, but traditional methods such as 24-hour dietary recalls, food diaries, and food frequency questionnaires are often prone to recall error and social desirability bias, limiting their reliability. With the advancement of artificial intelligence (AI), there is potential to overcome these limitations through automated, objective, and scalable dietary assessment techniques. However, the effectiveness and challenges of AI applications in this domain remain inadequately explored.

Objective: This study aimed to conduct a scoping review to synthesize existing literature on the efficacy, accuracy, and challenges of using AI tools in assessing food and nutrient intakes, offering insights into their current advantages and areas of improvement.

Methods: This review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. A comprehensive literature search was conducted in 4 databases—PubMed, Web of Science, Cochrane Library, and EBSCO—covering publications from the databases' inception to June 30, 2023. Studies were included if they used modern AI approaches to assess food and nutrient intakes in human subjects.

Results: The 25 included studies, published between 2010 and 2023, involved sample sizes ranging from 10 to 38,415 participants. These studies used a variety of input data types, including food images (n=10), sound and jaw motion data from wearable devices (n=9), and text data (n=4), with 2 studies combining multiple input types. AI models applied included deep learning (eg, convolutional neural networks), machine learning (eg, support vector machines), and hybrid approaches. Applications were categorized into dietary intake assessment, food detection, nutrient estimation, and food intake prediction. Food detection accuracies ranged from 74% to 99.85%, and nutrient estimation errors varied between 10% and 15%. For instance, the RGB-D (Red, Green, Blue-Depth) fusion network achieved a mean absolute error of 15% in calorie estimation, and a sound-based classification model reached up to 94% accuracy in detecting food intake based on jaw motion and chewing patterns. In addition, AI-based systems provided real-time monitoring capabilities, improving the precision of dietary assessments and demonstrating the potential to reduce recall bias typically associated with traditional self-report methods.

Conclusions: While AI demonstrated significant advantages in improving accuracy, reducing labor, and enabling real-time monitoring, challenges remain in adapting to diverse food types, ensuring algorithmic fairness, and addressing data privacy concerns. The findings suggest that AI has transformative potential for dietary assessment at both individual and population levels, supporting precision nutrition and chronic disease management. Future research should focus on enhancing the robustness of AI models across diverse dietary contexts and integrating biological sensors for a holistic dietary assessment approach.

(J Med Internet Res 2024;26:e54557) doi:10.2196/54557

XSL•FO
RenderX

KEYWORDS

food; nutrient; diet; artificial intelligence; machine learning; deep learning; neural networks; computer vision; natural language processing; measurement; AI; food intake; systematic literature; dietary assessments; AI-based; disease management; mobile phone

Introduction

Measuring food and nutrient intake is foundational in nutrition research, dietary surveillance, and clinical practice [1]. Traditional methods, such as 24-hour dietary recalls, food diaries, and food frequency questionnaires, have been the cornerstones of such endeavors [2]. However, these self-reported tools frequently encounter issues associated with recall error, where individuals inadvertently omit, underreport, or exaggerate certain food items or quantities [3]. Furthermore, social desirability bias further complicates matters, with respondents potentially altering their reports to reflect what they perceive as more socially acceptable or healthier dietary habits [4]. While clinical measures in controlled environments, such as laboratories, offer higher accuracy, they have drawbacks [5]. These objective measures often entail labor-intensive processes, significant costs, and potential intrusiveness for participants [6]. Such constraints render them less suitable for large-scale, population-level studies or individuals seeking to personally monitor their food and nutrient intake for disease management and other health-related objectives [6]. In light of these challenges, there is an escalating interest in leveraging artificial intelligence (AI) to enhance the accuracy and feasibility of dietary intake assessment [7].

AI, a branch of computer science focusing on developing algorithms that simulate human cognitive functions, has shown transformative potential across diverse sectors [8]. In health-related research, AI's ability to process vast amounts of data at incredible speeds and its adeptness at pattern recognition has made substantial strides in medical imaging, predictive modeling of disease outbreaks, and personalized medicine [9,10]. In the context of dietary assessment, AI offers several distinct advantages. First, it can potentially mitigate the biases inherent in self-reported methods by using image recognition to identify and quantify food items with minimal input from the user [11]. Advanced machine learning algorithms can analyze photographs of meals and provide instant, objective assessments of portion sizes and nutrient content [11,12]. In addition to image-based methods, AI techniques also use sound, jaw motion from wearable devices, and text data for dietary assessment. These methods provide diverse approaches to capture dietary intake, enhancing the accuracy and comprehensiveness of assessments. Second, AI can offer continuous, real-time monitoring, bridging the temporal gap in methods like 24-hour recalls [13]. Finally, while laboratory-based clinical measures are costly and labor-intensive, once developed, AI-driven tools can be scaled up relatively inexpensively, making them more feasible for large population studies and individual dietary tracking [14]. Given these attributes, AI emerges as a promising candidate to revolutionize the landscape of food and nutrient intake measurement.

While numerous reviews have covered objective measures of dietary intake, our review specifically focuses on the application of AI technologies in this field. This scoping review provides a comprehensive synthesis of recent advancements, highlights the unique challenges faced by AI methodologies, and identifies critical gaps that future research should address. Our work adds to the existing literature by providing a detailed analysis of AI's role in improving the accuracy and efficiency of dietary assessment.

To the best of our knowledge, a comprehensive scoping review that delves into the applications of AI for measuring food and nutrient intakes has not yet been conducted. This gap in the literature underlines the novelty and urgency of our investigation. The primary objective of this review is to explore and map out the current landscape of AI applications in dietary assessment, detailing methodologies, tools, and their associated findings.

This endeavor holds transformative potential for several reasons. First, by consolidating and synthesizing the vast yet dispersed body of knowledge, researchers, clinicians, and policy makers can gain a cohesive understanding of the current state-of-the-art and its implications for the future. Second, the review will spotlight any existing limitations or gaps in the current AI methodologies, paving the way for targeted advancements in technology and research design. Finally, given the paramount importance of accurate dietary assessment in myriad health outcomes and policy decisions, our findings can directly inform best practices, promote technology adoption in clinical and research settings, and guide future funding and priorities in technological and nutritional research sectors.

Methods

Overview

This scoping review followed the guidelines of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; see Multimedia Appendix 1) [15].

Study Selection Criteria

Predefined inclusion and exclusion criteria were established and applied to all identified studies during the screening process. Textbox 1 provides a detailed overview of the inclusion and exclusion criteria, outlining the study characteristics considered for eligibility in this review.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for study selection.

Inclusion criteria:

- Study design: Experimental studies (eg, randomized controlled trials [RCTs], pre-post interventions) and observational studies (eg, cross-sectional, longitudinal).
- Analytic approach: Modern AI approaches, including machine learning (ML), deep learning (DL), and reinforcement learning (RL).
- Participants: Individuals of all ages.
- Data type: Input data, including food images, plate images, etc.
- Outcome: Measures on food and nutrient intakes.
- Article type: Original, empirical, peer-reviewed journal publications.
- Language: Articles written in English.
- Search time frame: From the inception of electronic bibliographic databases to June 30, 2023.

Exclusion criteria

- Study design: Studies that do not involve human subjects, observational or experimental design.
- Analytic approach: Studies using rule-based ("hard-coded") approaches instead of example-based ML, DL, or RL.
- Participants: Non-human subjects.
- Data type: Studies not using dietary input data.
- Outcome: Studies without outcomes related to food and nutrient intakes.
- Article type: Letters, editorials, study or review protocols, case reports, or review articles.
- Language: Non–English-language articles.
- Search time frame: Studies published after June 30, 2023.

Search Strategy

A comprehensive search was performed in 4 electronic bibliographic databases: PubMed, Web of Science, Cochrane Library, and EBSCO. The search strategy used a combination of controlled vocabulary (eg, MeSH terms in PubMed) and free-text keywords. The search terms were structured around two main concepts: (1) AI and (2) nutrition or dietary intake. The AI-related terms included: "artificial intelligence," "machine learning," "deep learning," "neural networks," "natural language processing," "computer vision," "algorithms," "data mining," "big data," "predictive modeling," and "automated pattern recognition." The nutrition-related terms included: "nutrition," "dietetics," "nutritional sciences," "diet," "dietary behavior," "beverage intake," "food intake," "nutrient intake," and "healthy eating." These keywords were combined using Boolean operators (AND, OR) to ensure a comprehensive search. The complete search strategy, including database-specific modifications and detailed search strings, is provided in Multimedia Appendix 2. After the initial search, 2 coauthors independently screened the titles and abstracts for the articles found through the keyword search, obtained potentially relevant articles, and reviewed their full texts. The inter-rater agreement between these two authors was evaluated using Cohen κ (κ=0.85). Disagreements were settled through conversation.

Data Extraction and Synthesis

The following methodological and outcome variables were collected from each study using a standardized data extraction form: authors, year of publication, country or region, study objective, sample size, sample characteristics, AI models used, tasks and applications, type of input data, outcome measures, and perceived usefulness of AI technologies. No meta-analysis was feasible, given the substantial heterogeneity of the models, outcome measures, and applications. Therefore, we synthesized the study findings narratively and categorized them into distinct themes.

Results

Identification of Studies

Figure 1 illustrates the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram, outlining the structured literature search and selection procedure. The initial database search identified 6132 articles. After removing duplicates, 5499 unique articles were retained for preliminary screening based on their titles and abstracts. From this collection, 5456 articles were evaluated as irrelevant and, consequently, excluded from the review. Applying the study selection criteria to the remaining 43 articles resulted in the further exclusion of 18 studies due to various reasons, including lack of AI technology adoption (n=7), absence of food and nutrient intake measurements (n=6), being a commentary rather than original empirical research (n=3), and a focus on smartphone-based apps (n=2). Ultimately, 25 studies met the relevance criteria and were included in the review [12,14,16-38].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram illustrating the study selection process.

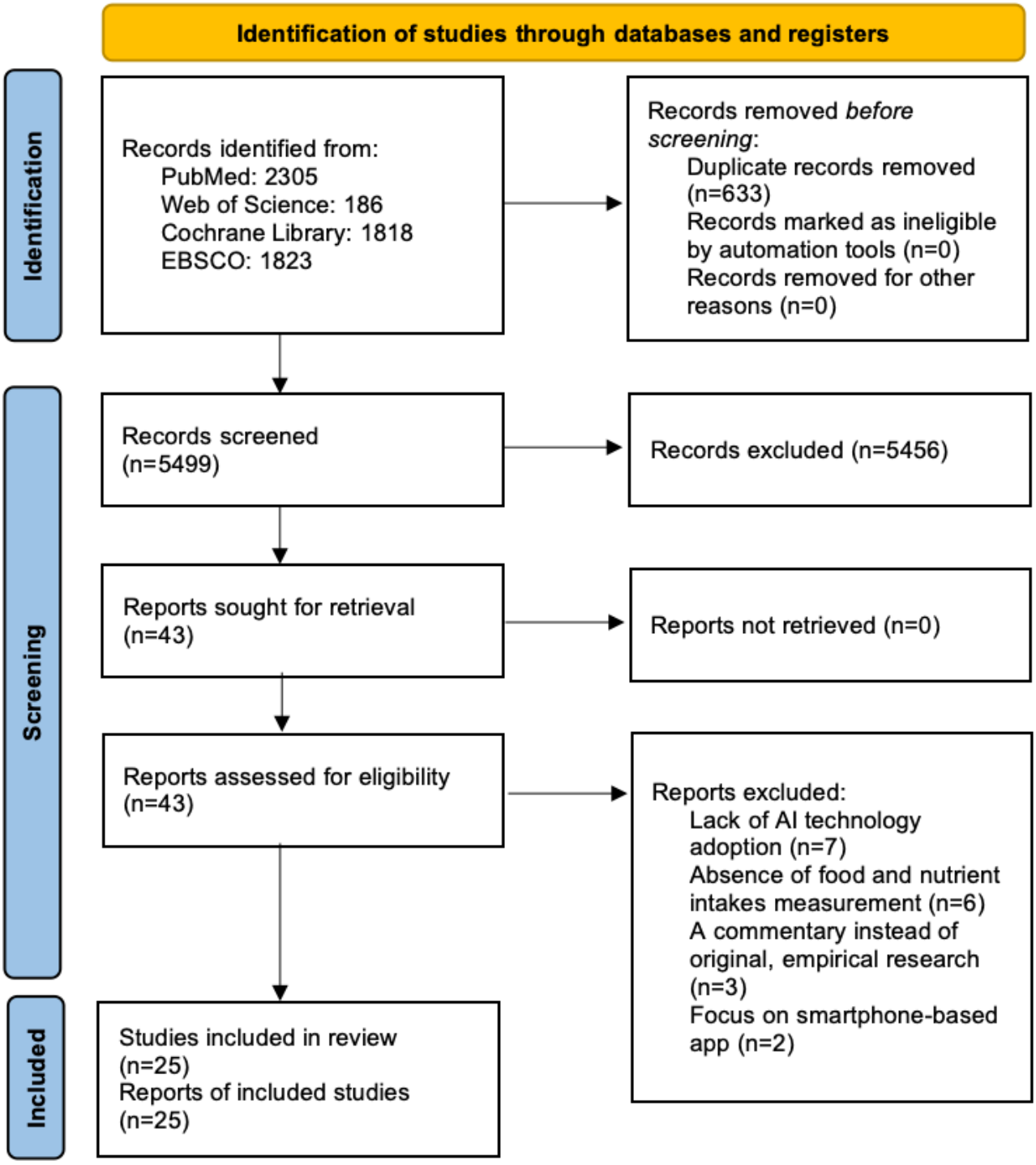

Study Characteristics

Table 1 reports the characteristics, type of input data, outcome measures, and main findings of the 25 studies incorporated in the review (more details in Multimedia Appendix 3). The studies spanned a range of publication years, with the earliest appearing in 2010 [16] and singular studies being published in 2013 [17], 2015 [18], and 2023 [38]. Publications in the following years were more frequent, with 2 studies each in 2016 [19,20], 2018 [21,22], and 2020 [14,27], 3 in 2021 [28-30], 5 in 2019 [12,23-26], and 7 in 2022 [31-37]. The geographical spread of the studies was diverse, with research conducted in several different countries: 14 in the United States [12,16,17,19-22,24-27,30,31,37], 4 in Switzerland [14,18,23,29], 2 in France [32,36], and 1 each in Canada [33], China [38], Denmark [34], Philippines [35], and Slovenia [28].

XSL•FO
RenderX

Table 1. Geographic location, sample size, sample characteristics, artificial intelligence models, type of input data, task, outcome measures, and main findings in the studies included in the review.

Author, Year	Country or Region	Sample size	Sample characteristics	AI[a] models	Type of input data	Task	Outcome measures
Lopez-Meyer et al, 2010 [16]	United States	18	Healthy adults (BMI: 28.01, SD 6.35)	SVM[b], RBFk[c]	Sound, strain signal	Classification	Food intake
Fontana et al, 2013 [17]	United States	12	Healthy adults (BMI: 24.39, SD 3.81)	RF[d]	Jaw motion signal, hand gesture signal, body acceleration	Classification	Food intake
Anthimopoulos et al, 2015 [18]	Switzerland	144	Images of dishes	CV[e], SVM	Image	Regression	Carbohydrates counting
Farooq and Sazonov, 2016 [19]	United States	10	Healthy adults (BMI: 27.87, SD 5.51)	SVM, DT[f]	Jaw motion signal, body acceleration signal	Classification	Food intake
Hezarjaribi et al, 2016 [20]	United States	10	—[g]	SRM[h], NLP[i], SMM[j]	Audio signal	Regression	Calorie intake
Goldstein et al, 2018 [21]	United States	12	Adults with overweight/obesity (BMI: 33.60, SD 5.66)	RF, DT, Logit.Boot, BN[k], Bagging, Random subspace	Text	Regression	Dietary lapses
Hezarjaribi et al, 2018 [22]	United States	30	—	NLP, LA[l]	Audio signal	Regression	Calorie intake
Lu et al, 2019 [23]	Switzerland	644	Meal images (pixel: 640×480)	MTNnet[m], DTM[n], RANSAC algorithm	Image	Regression	Nutrient intake
Fang et al, 2019 [12]	United States	4190	Food images (pixel: 224×224)	GAN[o], CNN[p]	Image	Regression	Food energy
Jia et al, 2019 [24]	United States	38,415	Images (pixel: 640×480)	CNN	Image	Regression	Dietary assessment
Chin et al, 2019 [25]	United States	567	Food descriptions	LASSO[q], Ridge, FFNN[r], XGB[s] models	Text	Regression	Amount of lactose
Farooq et al, 2019 [26]	United States	40	Healthy adults (BMI: 26.1, SD 5.2)	NNC[t]	Hand gesture, jaw motion, body acceleration	Classification	Food intake
Heremans et al, 2020 [27]	United States	126	Adults with dyspepsia	ANN[u]	Heart rate variability signal	Classification	Food intake
Lu et al, 2021 [14]	Switzerland	644	Meal images (pixel: 640×480)	MTCNet[v], FSLBC[w], 3D-SCA[x]	Image	Regression	Nutrient intake
Mezgec and Koroušić Seljak, 2021 [28]	Slovenia	520	Food images (pixel: 512×512)	DNN[y]	Image	Classification	Dietary assessment
Papathanail et al, 2021 [29]	Switzerland	866	Meal images (pixel: 640×480)	CNN, PSPNet[z], DeepLabv3 network	Image	Regression	Energy, nutrient intake
Taylor et al, 2021 [30]	United States	34	Healthy adults (mean BMI: 24)	CNN, SMM	Text, voice data	Regression	Energy intake
Ghosh and Sazonov, 2022 [31]	United States	17	Adolescents and adults	Time-CNN, ResNet[aa], FCN[ab], IM[ac], MLP[ad]	Accelerometer, optical sensor data	Classification	Food intake

XSL•FO
RenderX

Author, Year	Country or Region	Sample size	Sample characteristics	AI[a] models	Type of input data	Task	Outcome measures
Van Wymelbeke-Delannoy et al, 2022 [32]	France	22,544	Dishes images	DNN	Food image	Regression	Food item
Pfisterer et al, 2022 [33]	Canada	689	Plate images (pixel: 640×480)	Deep-CNN	Plate image	Regression	Food intake
Pedersen et al, 2022 [34]	Denmark	100	Adults with normal weight	RF	Psychophysiological responses	Regression	Food intake
Siy Van et al, 2022 [35]	Philippines	618	Children	RF, SVM, LDA[ae], LR[af]	Text	Regression	Under-nutrition
Granal et al, 2022 [36]	France	375	Adults with chronic kidney disease	BN, BTANN[ag]	Text	Regression	Dietary potassium intake
Nguyen et al, 2022 [37]	United States	36	Adolescent	Pop-socket	Image, text	Regression	Dietary intake
Shao et al, 2023 [38]	China	5920	Food images	RGB-D[ah] fusion network	Image	Regression	Energy, nutrient intake

[a]AI: artificial intelligence.
[b]SVM: support vector machine.
[c]RBFk: radial basis function kernels.
[d]RF: random forest.
[e]CV: computer vision.
[f]DT: decision tree.
[g]Not applicable.
[h]SRM: speech recognition model.
[i]NLP: natural language processing.
[j]SMM: string matching module.
[k]BN: Bayesian network.
[l]LA: levenshtein algorithm.
[m]MTNnet: multi-task neural network.
[n]DTM: delaunay triangulation method.
[o]GAN: generative adversarial networks.
[p]CNN: convolutional neural network.
[q]LASSO: least absolute shrinkage and selection operator.
[r]FFNN: feed forward neural network.
[s]XGB: eXtreme gradient boosting.
[t]NNC: neural network classifier.
[u]ANN: artificial neural network.
[v]MTCNet: multi-task contextual network.
[w]FSLBC: few-shot learning-based classifier.
[x]SCA: surface construction algorithm.
[y]DNN: deep neural network.
[z]PSPNet: pyramid scene parsing network.
[aa]ResNet: residual neural network.
[ab]FCN: fully convolutional neural network.
[ac]IM: inception network.
[ad]MLP: multilayer perceptron.
[ae]LDA: linear discriminant analysis.
[af]LR: logistic regression.
[ag]BTANN: bayesian tree augmented naive network.
[ah]RGB-D: Red, Green, Blue-Depth.

The studies varied in sample sizes, ranging from 10 to 38,415. Specifically, 10 studies had sample sizes between 10 and 99 [16,17,19-22,26,30,31,37], 3 had between 100 and 199 [18,27,34], and the remaining 12 had sample sizes exceeding 300. Among the 25 studies, while all involved human subjects, 10 studies focused on analyzing food images, dish images, or

XSL•FO
RenderX

plate images to estimate dietary intake [12,14,18,23,24,28,29,32,33,38], 4 targeted healthy adults dealing with obesity or overweight [16,19,21,26], 3 focused on adults with normal weight [17,30,34], 3 engaged with children and adolescents [31,35,37], and 2 addressed adults with diseases [27,36]. Over the years, there have been notable advancements in AI-based dietary assessment. Early studies primarily focused on developing basic image recognition algorithms. More recent studies have integrated advanced machine learning models, such as deep learning and convolutional neural networks, which have significantly improved the accuracy of food recognition and nutrient estimation.

Among the 25 studies, 10 used image data, 9 used sound or jaw motion data from wearable devices, 4 used text data, and the remaining 2 combined multiple types of input data for dietary assessment. We classified the applications into 4 categories: dietary intake assessment, food detection, nutrient estimation, and food intake prediction.

Applications in Dietary Intake Assessment

Our review identified several critical steps involved in the processing of dietary intake assessment systems, specifically for image-based methods. These steps include (1) identifying images with food, (2) identifying the foods, (3) separating the foods into separate parts, (4) estimating portion sizes served and remaining to estimate intake, and (5) estimating nutrient intake. Each of these steps involves distinct AI methodologies with varying degrees of accuracy and potential errors.

Identifying Images With Food

AI models, particularly convolutional neural networks (CNNs), are widely used for recognizing the presence of food in images. Studies, such as those by Fang et al (2019) [12] and Jia et al (2019) [24], have demonstrated high accuracy in detecting food presence using end-to-end image-based automatic food energy estimation techniques and real-world egocentric images, respectively.

Identifying the Foods

Once food is identified in an image, the next step is to classify and recognize different food items. Techniques such as support vector machines (SVMs) and deep learning models, including GANs (Generative Adversarial Networks) and advanced CNNs, are used for this purpose. For example, the GoCARB system developed by Anthimopoulos et al [18] used computer vision to estimate carbohydrate content by recognizing different food items from smartphone images.

Separating Foods Into Separate Parts

Segmenting individual food items within an image is crucial for accurate portion size estimation. Techniques such as image segmentation using deep neural networks (DNNs) have been effective in this regard. The study by Mezgec and Koroušić Seljak [28] showcased the use of DNNs for image-based dietary assessment with a classification accuracy of 86.72%.

Estimating Portion Sizes Served and Remaining

Estimating the portion sizes of served and remaining food requires precise volume and area measurements, which can be challenging due to varying presentation and occlusion of food items. AI models using RGB-D (Red, Green, Blue-Depth) imagery, as seen in the work by Shao et al [38], have shown promise in improving the precision of such estimations by using depth information to enhance the accuracy of food volume assessments.

Estimating Nutrient Intake

The final step involves estimating the nutrient intake based on the identified and quantified food items. This step often leverages databases such as the US Department of Agriculture (USDA) nutritional database to map food items to their nutrient profiles. The integration of AI for this purpose is exemplified by systems like the S2NI platform, which combines speech recognition and natural language processing to monitor dietary composition from spoken data, achieving high accuracy in nutrient computation.

Non–image-based dietary assessment methods, including those using sound, jaw motion from wearable devices, and text analysis, can also be categorized similarly. These methods contribute to various steps, particularly in identifying food intake and estimating nutrient content. For instance, the use of jaw motion signals analyzed by SVMs, as studied by Lopez-Meyer et al [16], provides high accuracy in detecting food intake.

Applications in Food Detection

Food detection refers to the identification and recognition of food items using AI technologies. AI applications have become increasingly important in automating food detection, providing foundational advancements crucial for accurate nutrient estimation and food intake prediction. SVM and random forests are highlighted as prevalent machine learning models across the studies, aiming to achieve high food detection accuracy [16,17,19]. Random forest classification emphasizes the importance of time and frequency domain features in food intake detection with wearable sensor systems, focusing predominantly on jaw motion and accelerometer signals [17].

Another essential facet in this AI-infused dietary landscape is the integration of image-based assessments [28,33]. The development and validation of deep neural networks like NutriNet for food and beverage image recognition have showcased the ability of image-based approaches to identify multiple food or beverage items in a single image. Moreover, incorporating FCNs and deep residual networks (ResNet) magnifies the efficacy of segmenting food images, presenting a robust method in automated dietary assessments. Notably, Pfisterer et al [33] offered insights into the application of deep convolutional encoder-decoder food networks with depth-refinement (EDFN-D) in long-term care settings, providing an automated imaging system for quantifying food intake with high precision and objectivity, addressing the existing limitations in these settings [33].

A noticeable trend across the studies is the use of wearable and mobile devices, demonstrating the integration of technology with daily human activities for real-time and accurate data collection [19,24,32,37]. Wearable devices, such as the Automatic Ingestion Monitor (AIM) and other novel devices with sensors on the temporalis muscle and accelerometers, have

XSL•FO
RenderX

shown potential in reducing the influence of motion artifacts and speech on food intake detection accuracy [19]. Furthermore, mobile AI technologies, such as FRANI (Food Recognition Assistance and Nudging Insights), illustrate their feasibility and reliability in resource-constrained settings, offering a comparable alternative to traditional methods like weighed records (WRs) [37].

DNN and CNN are central in recognizing and detecting food items from images, providing an automated approach to food detection and segmentation. The FoodIntech system, using a DNN-based approach, has demonstrated reliability in recognizing a variety of dishes and assessing food consumption [32]. Similarly, algorithms designed for egocentric images from wearable cameras have achieved substantial accuracy in food detection, addressing concerns related to data processing burdens and privacy [24].

Combining AI with RGB-D imagery is an evolving approach, showing promise in refining the precision of food nutrition estimation. The use of RGB-D fusion networks has revealed advancements in performing multimodal and multiscale feature fusion, offering a refined accuracy in nutrient analysis [38]. This approach successfully estimated calories and mass with a lower percentage mean absolute error and effectively visualized the estimation results of 4 nutrients [38].

Despite the advancements, there is a discernable disparity in the reported accuracy and reliability among the studies, with accuracy ranging from 74% to 99.85% [19,24]. This variance reflects the diverse methodologies, sensor modalities, ML algorithms, and the nature of features extracted for analysis. The ongoing refinements in methods and technologies showcase the evolving nature of AI applications in food detection, signaling a step forward in automating dietary assessment in varied environments and demographic settings.

Applications in Nutrient Estimation

AI has been used to address the challenges associated with accurate nutrient intake assessment and dietary management for various medical conditions and patient demographics. The GoCARB system [18] exemplifies how AI can assist individuals with type 1 diabetes in carbohydrate counting, using computer vision to automate the estimation process using smartphones, hence aiding in optimal insulin dosage estimations. This application relies on the segmentation and recognition of food items, calculating the carbohydrate content based on food volumes and the USDA nutritional database, demonstrating a mean absolute percentage error in carbohydrate estimation of approximately 10%.

In addressing the nutrition assessment needs of hospitalized patients, an AI-based system has been developed [23,29] that uses RGB-D image pairs to estimate nutrient intake. These applications offer a means to counter malnutrition risks in hospital settings by delivering more accurate and automated nutrient intake assessments. The systems segment images into different food components, estimate the volume consumed, and calculate energy and macronutrient intake, showing a 15% estimation error [23] and improved agreement with expert estimations compared to standard clinical procedures [29].

Efforts have also been made to estimate food energy values using GAN architecture [12]. By mapping food images to their energy distributions, the technology has shown promise in improving the accuracy of dietary assessments, with an average error of 209 kcal per eating occasion in a real-world study setting.

In the context of 24-hour food recalls, machine learning models and database matching have been instrumental in estimating nutrients not directly outputted by specific dietary assessment tools [25]. For instance, lactose was relatively accurately estimated using models like XGB regressor and database matching methods.

Meanwhile, studies on the interplay between behavioral and physiologic variables in predicting food intake [34] have provided foundational insights. However, the predictive capability of combined or separate measures of food reward or biometric responses has not outperformed traditional models in clinical settings. The approach, however, lays the groundwork for further exploration of behavioral nutrition and personalized nutrition strategies.

Furthermore, the development of predictive tools leveraging AI for patients with chronic kidney disease has exhibited the potential to estimate dietary potassium intake, emphasizing the role of AI in clinical and therapeutic management [36]. This application has been noteworthy for its ability to classify potassium diet in 3 classes of potassium excretion with 74% accuracy, focusing more on clinical characteristics and renal pathology than on the potassium content of the ingested food.

Using mobile platforms that incorporate speech and natural language processing to convert spoken data to nutrient information offers a lens into the transformative potential of voice-based solutions [20,22,30]. These solutions, such as S2NI, Speech2Health, and the COCO Nutritionist app, achieve substantial accuracy in computing calorie intake, emphasizing the importance of real-time and pervasive monitoring. They demonstrate an integrated approach to capture dietary information more frequently, revealing the user preference toward voice-based interfaces over text-based and image-based nutrition monitoring due to their ease of use and accessibility.

Applications in Food Intake Prediction

Food intake prediction involves estimating the amount and type of food consumed based on detected items. Advancements in AI are significantly shaping the landscape of food intake prediction by offering various innovative solutions and techniques. For instance, ML techniques in predicting dietary lapses during weight loss interventions have demonstrated the potential to augment adherence to dietary guidelines and offer real-time interventions, providing a comprehensive perspective on combining individual and group-level data to enrich predictions [21].

The adaptability and efficiency of ML are further highlighted in the studies focusing on detecting food intake using various sensor technologies and algorithms. Developing and validating sensor-based food intake detection methods, such as AIM, have illustrated high accuracy and reliability, presenting a promising future for food intake monitoring in unconstrained environments

XSL•FO
RenderX

[26,31]. SVMs have been effectively used in monitoring ingestive behavior, yielding up to 94% accuracy in detecting food intake by analyzing chews and swallows [16].

In particular, the utility of DL algorithms, like ResNet and Fully Convolutional Neural Network (FCN), is revealed to be paramount in differentiating food intake from other activities using sensor signals. The competitive performance of these algorithms indicates the significance of selecting appropriate methods for precise classifications in real-world scenarios, establishing their importance in the evolving field of dietary monitoring and health interventions [31].

The exploration of DNN in automatic food intake detection through dynamic analysis of heart rate variability has opened avenues for addressing meal-related disorders. The notable accuracy of DNN, especially in neuromodulation treatments for conditions like obesity and diabetes, establishes the potential of ML in contributing to varied health care settings [27].

Furthermore, the studies using ML algorithms like the random forest have provided a robust method for identifying and comparing nutritional risk, offering valuable insights into developing targeted nutritional interventions and effectively addressing undernutrition. Such approaches are crucial in considering local dietary culture and delivering more nuanced and culturally competent health care solutions [35].

Another essential facet in this AI-infused dietary landscape is the integration of image-based assessments [28,33]. The development and validation of deep neural networks like NutriNet for food and beverage image recognition have showcased the ability of image-based approaches to identify multiple food or beverage items in a single image. Furthermore, incorporating FCNs and deep residual networks (ResNet) magnifies the efficacy of segmenting food images, presenting a robust method in automated dietary assessments. Notably, Pfisterer et al [33] offered insights into the application of deep convolutional EDFN-D in long-term care settings, providing an automated imaging system for quantifying food intake with high precision and objectivity, addressing the existing limitations in these settings [33].

Discussion

Principal Findings

The increasing intersection of AI with dietary assessment has emerged as a transformative trend, as evidenced by our scoping review. Our literature search revealed 25 pertinent studies published between 2010 and 2023. These studies spanned several nations, diverse demographics, and a spectrum of methodologies. At its core, AI has primarily been used in 3 domains: food detection, nutrient estimation, and food intake prediction. Machine learning models like SVMs and random forests and deep learning models like CNNs have proved instrumental in enhancing the accuracy of food detection and nutrient estimation, often integrated with wearable devices and mobile platforms. Another observation was the use of AI in designing user-friendly interfaces, such as voice-based inputs, to improve adherence to dietary tracking. User experience with AI-based dietary assessment tools varies, but studies indicate generally positive feedback regarding ease of use and convenience. Users appreciate the real-time feedback and reduced burden of manual input. However, there are concerns about accuracy and privacy. Enhanced user training and transparent data privacy policies could improve user trust and interaction with these tools. The collective findings underscore the potential of AI to revolutionize dietary assessment, providing robust accuracy and user-centric solutions. This amalgamation of technology and nutrition research addresses the inherent limitations of traditional methods and charts a path for more personalized, accurate, and real-time dietary assessments in varied settings.

As illustrated by the reviewed studies, integrating AI into food and nutrient intake assessments showcases a marked advancement over traditional methodologies commonly used in nutritional science [14,37]. Historically, methods such as 24-hour recalls, food frequency questionnaires, and dietary records have been the mainstream of dietary assessments [2]. While these methods have provided invaluable insights, they have inherent limitations like recall bias, inaccuracies stemming from self-reporting, and the logistical challenges of frequent, detailed data recording [3]. The reviewed studies, however, highlighted the significant potential of AI to alleviate some of these concerns. For instance, AI-backed systems such as FRANI have been shown to offer a reliable alternative to weighed records, which, although thorough, can be burdensome for participants [37]. Similarly, tools like the GoCARB system automate carbohydrate counting, which, if done manually, demands meticulous attention and can be prone to errors, especially for individuals with conditions like diabetes [18].

Furthermore, the versatility of AI applications across various nutritional assessments is evident from the reviewed literature. For instance, SVMs and random forests, when deployed in monitoring ingestive behaviors, have demonstrated high accuracy in detecting food intake by analyzing nuances such as chews and swallows [16]. This level of precision is difficult to attain through manual observation or self-reports. Applying DNNs to recognize food items from images underscores another leap, automating a process that traditionally demands human expertise. Furthermore, the intersection of AI with RGB-D imagery suggests an improved accuracy in nutrient analysis, an area where traditional methods may not always yield precise results [38]. However, it is crucial to note the variability in reported accuracy among studies, which underscores the importance of refining methodologies and recognizing the evolving nature of AI applications. Despite this, the current trajectory indicates that AI is poised to bring a paradigm shift in automating dietary assessment, melding accuracy with efficiency [36,37]. Wearable technology that detects food intake based on chews and swallows offers significant benefits in real-time dietary monitoring, particularly in clinical and research settings. These devices can be integrated with mobile applications and other wearable sensors to provide comprehensive dietary assessments. While continuous camera use may not be practical for all users, advancements in discreet wearable sensors and intermittent image capture can enhance user compliance and accuracy.

XSL•FO
RenderX

While AI's promise in food and nutrient intake measurement is evident, its application comes with intrinsic challenges and limitations. The reviewed studies, as well as the broader literature, highlight some consistent concerns. First, the AI models heavily depend on the quality and breadth of training data [39]. A model trained on a limited dataset may not recognize diverse food items, particularly those from various global cuisines or those prepared using unique methods [40]. This can lead to inaccuracies in nutrient estimation. Common biases include algorithmic biases resulting from non-diverse training datasets that fail to represent global food diversity. In addition, limitations in image-based recognition systems often stem from varying image quality and presentation, which can affect the accuracy of food and nutrient estimations. The variability in food presentation, portion sizes, and the physical environment in which the food is captured (eg, lighting conditions) can pose challenges for image-based recognition systems [41,42]. Furthermore, while tools like FRANI and GoCARB show promise, they also underscore the current limitations in recognizing mixed dishes or deciphering layered foods with multiple ingredients [18,37]. It is also worth noting that AI systems while reducing human biases, introduce computational biases that may arise from algorithmic designs or training datasets [43,44]. These challenges highlight the need for more comprehensive datasets and improved image processing techniques to enhance AI model reliability. Finally, a potential digital divide exists, where populations without access to advanced technology or those not adept at using it might be excluded from AI-based dietary assessments, thereby limiting its universal applicability [45,46].

Many AI-based dietary assessment tools rely on dietitians to validate and estimate dietary intake from images due to the complexities involved in accurate food identification and portion size estimation. With the constant addition of new food items, maintaining up-to-date nutrient databases is challenging. Some studies have focused narrowly on estimating energy intake or working with a limited set of foods under controlled conditions, which limits the generalizability of their findings. Future research should focus on developing scalable AI models that can handle a broader range of foods and integrate real-time updates to nutrient databases. In addition, enhancing the collaboration between AI technologies and dietitians can help improve the accuracy and applicability of these tools.

Current objective methods face significant limitations, including inaccuracies in nutrient composition tables, the complexity of multi-ingredient dishes, and variability in nutrient composition of commercially available foods. In addition, these methods do not account for individual metabolic differences in nutrient processing. Integrating biological sensors with AI technologies could offer a more definitive approach by providing real-time data on circulating nutrients and individual metabolic responses, thereby improving the accuracy of dietary assessments.

The sequential nature of AI-based dietary assessment introduces cumulative errors, where inaccuracies at each stage—from food detection to nutrient estimation—can compound, leading to significant overall errors. Biological sensors that measure circulating nutrients in real-time offer a promising solution to overcome these limitations, as they provide direct data on nutrient absorption and metabolism, reducing reliance on intermediate estimations and improving overall accuracy.

Our search strategy, while comprehensive, may not have captured all studies involving AI and dietary assessment. Despite significant advancements, several gaps remain in the application of AI for dietary assessment. Future research should focus on enhancing the diversity of training datasets to reduce algorithmic biases and improve the accuracy of AI models in recognizing a wide variety of food items. In addition, integrating real-time metabolic data with dietary assessments could offer more comprehensive insights into individual nutritional statuses. Among the AI tools evaluated, image-based recognition systems like the GoCARB system are highly effective for carbohydrate counting in diabetes management, while wearable devices monitoring jaw motion offer promising real-time intake data, particularly useful in clinical settings.

Ethical considerations in AI-based dietary assessment are paramount. Data privacy concerns arise from the extensive personal data required for accurate assessments, necessitating robust security measures and transparent consent processes. Algorithmic biases can lead to inaccuracies and unfair outcomes, highlighting the need for diverse training datasets. In addition, the digital divide poses a significant challenge, as populations without access to advanced technologies may be excluded from the benefits of AI. Addressing these issues requires comprehensive strategies, including inclusive technology design and stringent ethical standards in data handling and algorithm development.

As AI continues to evolve, there is vast potential for revolutionary enhancements in dietary and nutrient intake measurement. Based on current trajectories in nutrition science and AI advancements, we might anticipate a future where AI systems can recognize food items with high precision and factor in variables like cooking methods, regional variations, and the bioavailability of nutrients. These AI systems could be trained on increasingly diverse datasets, capturing the nuances of global diets and potentially integrating real-time metabolic and physiological data from wearable devices to provide a more comprehensive view of an individual's nutrient absorption [47,48]. AI could facilitate large-scale dietary assessment studies on a population level, helping researchers discern dietary patterns, nutrient deficiencies, and even epidemiological correlations faster and more accurately [49,50]. With the rise of precision nutrition, AI might enable personalized dietary recommendations, considering an individual's genetic, metabolic, and health profile [51]. This tailored approach could radically improve disease management, particularly for conditions like diabetes or cardiovascular diseases, where dietary interventions play a pivotal role [52].

Conclusion

In conclusion, the scoping review highlighted the burgeoning role of AI in advancing the measurement of food and nutrient intakes, with notable advancements in accuracy and efficiency compared to traditional methods. However, while the potential of AI in this domain is substantial, it is imperative to acknowledge its current limitations and areas requiring refinement. As the nexus between nutrition science and

XSL•FO
RenderX

technology continues to strengthen, future research must focus on refining AI methodologies, ensuring their applicability across diverse populations, and integrating them into broader nutritional and health studies. This interdisciplinary collaboration promises a future where dietary assessments are accurate and instrumental in shaping individual and public health outcomes.

Acknowledgments

This research received no external funding.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

JZ and RA contributed to conceptualization. JW contributed to the methodology. JS handled the software. JZ, JW, and JS performed validation. JW and JS conducted the formal analysis. JZ and RA conducted the investigation. RA handled the resources. JW and JS performed data curation. JZ and RA contributed to writing—original draft preparation. JW and JS contributed to writing—review and editing. JW performed visualization. RA performed supervision. RA contributed to project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Note that our paper is a scoping review rather than a systematic review, so some criteria in the checklist may not apply and are thus omitted).
[PDF File (Adobe PDF File), 80 KB - jmir_v26i1e54557_app1.pdf]

Multimedia Appendix 2
Database search algorithms.
[DOCX File , 15 KB - jmir_v26i1e54557_app2.docx]

Multimedia Appendix 3
Main findings in the studies included in the review.
[DOCX File , 21 KB - jmir_v26i1e54557_app3.docx]

References

1. Kirkpatrick SI, Collins CE. Assessment of nutrient intakes: introduction to the special issue. Nutrients 2016;8(4):184 [FREE Full text] [doi: 10.3390/nu8040184] [Medline: 27023598]
2. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009 [FREE Full text] [doi: 10.4178/epih/e2014009] [Medline: 25078382]
3. Ravelli MN, Schoeller DA. Traditional self-reported dietary instruments are prone to inaccuracies and new approaches are needed. Front Nutr 2020;7:90 [FREE Full text] [doi: 10.3389/fnut.2020.00090] [Medline: 32719809]
4. Hebert JR, Clemow L, Pbert L, Ockene IS, Ockene JK. Social desirability bias in dietary self-report may compromise the validity of dietary intake measures. Int J Epidemiol 1995;24(2):389-398. [doi: 10.1093/ije/24.2.389] [Medline: 7635601]
5. Masterton S, Hardman CA, Boyland E, Robinson E, Makin HE, Jones A. Are commonly used lab-based measures of food value and choice predictive of self-reported real-world snacking? An ecological momentary assessment study. Br J Health Psychol 2023;28(1):237-251 [FREE Full text] [doi: 10.1111/bjhp.12622] [Medline: 36000399]
6. Jobarteh ML, McCrory MA, Lo B, Sun M, Sazonov E, Anderson AK, et al. Development and validation of an objective, passive dietary assessment method for estimating food and nutrient intake in households in low- and middle-income countries: a study protocol. Curr Dev Nutr 2020;4(2):nzaa020 [FREE Full text] [doi: 10.1093/cdn/nzaa020] [Medline: 32099953]
7. Oliveira Chaves L, Gomes Domingos AL, Louzada Fernandes D, Ribeiro Cerqueira F, Siqueira-Batista R, Bressan J. Applicability of machine learning techniques in food intake assessment: a systematic review. Crit Rev Food Sci Nutr 2023;63(7):902-919. [doi: 10.1080/10408398.2021.1956425] [Medline: 34323627]
8. Collins C, Dennehy D, Conboy K, Mikalef P. Artificial intelligence in information systems research: a systematic literature review and research agenda. International Journal of Information Management 2021;60:102383 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2021.102383]

XSL•FO
RenderX

9. Schork NJ. Artificial intelligence and personalized medicine. Cancer Treat Res 2019;178:265-283 [FREE Full text] [doi: 10.1007/978-3-030-16391-4_11] [Medline: 31209850]
10. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
11. Sudo K, Murasaki K, Kinebuchi T, Kimura S, Waki K. Machine learning-based screening of healthy meals from image analysis: system development and pilot study. JMIR Form Res 2020;4(10):e18507 [FREE Full text] [doi: 10.2196/18507] [Medline: 33104010]
12. Fang S, Shao Z, Kerr DA, Boushey CJ, Zhu F. An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: protocol and methodology. Nutrients 2019;11(4):877 [FREE Full text] [doi: 10.3390/nu11040877] [Medline: 31003547]
13. Folson GK, Bannerman B, Atadze V, Ador G, Kolt B, McCloskey P, et al. Validation of mobile artificial intelligence technology-assisted dietary assessment tool against weighed records and 24-Hour recall in adolescent females in Ghana. J Nutr 2023;153(8):2328-2338 [FREE Full text] [doi: 10.1016/j.tjnut.2023.06.001] [Medline: 37276939]
14. Lu Y, Stathopoulou T, Vasiloglou MF, Christodoulidis S, Stanga Z, Mougiakakou S. An artificial intelligence-based system to assess nutrient intake for hospitalised patients. IEEE Trans. Multimedia 2021;23:1136-1147. [doi: 10.1109/tmm.2020.2993948]
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Lopez-Meyer P, Schuckers S, Makeyev O, Sazonov E. Detection of periods of food intake using Support Vector Machines. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:1004-1007. [doi: 10.1109/IEMBS.2010.5627796] [Medline: 21096991]
17. Fontana JM, Farooq M, Sazonov E. Estimation of feature importance for food intake detection based on random forests classification. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:6756-6759. [doi: 10.1109/EMBC.2013.6611107] [Medline: 24111294]
18. Anthimopoulos M, Dehais J, Shevchik S, Ransford BH, Duke D, Diem P, et al. Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones. J Diabetes Sci Technol 2015;9(3):507-515 [FREE Full text] [doi: 10.1177/1932296815580159] [Medline: 25883163]
19. Farooq M, Sazonov E. A novel wearable device for food intake and physical activity recognition. Sensors (Basel) 2016;16(7):1067 [FREE Full text] [doi: 10.3390/s16071067] [Medline: 27409622]
20. Hezarjaribi N, Reynolds CA, Miller DT, Chaytor N, Ghasemzadeh H. S2NI: a mobile platform for nutrition monitoring from spoken data. Annu Int Conf IEEE Eng Med Biol Soc 2016;2016:1991-1994. [doi: 10.1109/EMBC.2016.7591115] [Medline: 28268720]
21. Goldstein SP, Zhang F, Thomas JG, Butryn ML, Herbert JD, Forman EM. Application of machine learning to predict dietary lapses during weight loss. J Diabetes Sci Technol 2018;12(5):1045-1052 [FREE Full text] [doi: 10.1177/1932296818775757] [Medline: 29792067]
22. Hezarjaribi N, Mazrouee S, Ghasemzadeh H. Speech2Health: a mobile framework for monitoring dietary composition from spoken data. IEEE J Biomed Health Inform 2018;22(1):252-264. [doi: 10.1109/JBHI.2017.2709333] [Medline: 29300701]
23. Lu Y, Stathopoulou T, Vasiloglou MF, Christodoulidis S, Blum B, Walser T, et al. An artificial intelligence-based system for nutrient intake assessment of hospitalised patients. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:5696-5699. [doi: 10.1109/EMBC.2019.8856889] [Medline: 31947145]
24. Jia W, Li Y, Qu R, Baranowski T, Burke LE, Zhang H, et al. Automatic food detection in egocentric images using artificial intelligence technology. Public Health Nutr 2019;22(7):1168-1179 [FREE Full text] [doi: 10.1017/S1368980018000538] [Medline: 29576027]
25. Chin EL, Simmons G, Bouzid YY, Kan A, Burnett DJ, Tagkopoulos I, et al. Nutrient estimation from 24-hour food recalls using machine learning and database mapping: a case study with lactose. Nutrients 2019;11(12):3045 [FREE Full text] [doi: 10.3390/nu11123045] [Medline: 31847188]
26. Farooq M, Doulah A, Parton J, McCrory MA, Higgins JA, Sazonov E. Validation of sensor-based food intake detection by multicamera video observation in an unconstrained environment. Nutrients 2019;11(3):609 [FREE Full text] [doi: 10.3390/nu11030609] [Medline: 30871173]
27. Heremans ERM, Chen AS, Wang X, Cheng J, Xu F, Martinez AE, et al. Artificial neural network-based automatic detection of food intake for neuromodulation in treating obesity and diabetes. Obes Surg 2020;30(7):2547-2557. [doi: 10.1007/s11695-020-04511-6] [Medline: 32103435]
28. Mezgec S, Koroušić Seljak B. Deep neural networks for image-based dietary assessment. J Vis Exp 2021(169):e61906. [doi: 10.3791/61906] [Medline: 33779595]
29. Papathanail I, Brühlmann J, Vasiloglou MF, Stathopoulou T, Exadaktylos AK, Stanga Z, et al. Evaluation of a novel artificial intelligence system to monitor and assess energy and macronutrient intake in hospitalised older patients. Nutrients 2021;13(12):4539 [FREE Full text] [doi: 10.3390/nu13124539] [Medline: 34960091]

XSL•FO
RenderX

30. Taylor S, Korpusik M, Das S, Gilhooly C, Simpson R, Glass J, et al. Use of natural spoken language with automated mapping of self-reported food intake to food composition data for low-burden real-time dietary assessment: method comparison study. J Med Internet Res 2021;23(12):e26988 [FREE Full text] [doi: 10.2196/26988] [Medline: 34874885]
31. Ghosh T, Sazonov E. A comparative study of deep learning algorithms for detecting food intake. Annu Int Conf IEEE Eng Med Biol Soc 2022;2022:2993-2996. [doi: 10.1109/EMBC48229.2022.9871278] [Medline: 36085821]
32. Van Wymelbeke-Delannoy V, Juhel C, Bole H, Sow AK, Guyot C, Belbaghdadi F, et al. A cross-sectional reproducibility study of a standard camera sensor using artificial intelligence to assess food items: the foodIntech project. Nutrients 2022;14(1):221 [FREE Full text] [doi: 10.3390/nu14010221] [Medline: 35011096]
33. Pfisterer KJ, Amelard R, Chung AG, Syrnyk B, MacLean A, Keller HH, et al. Automated food intake tracking requires depth-refined semantic segmentation to rectify visual-volume discordance in long-term care homes. Sci Rep 2022;12(1):83 [FREE Full text] [doi: 10.1038/s41598-021-03972-8] [Medline: 34997022]
34. Pedersen H, Diaz LJ, Clemmensen KKB, Jensen MM, Jørgensen ME, Finlayson G, et al. Predicting food intake from food reward and biometric responses to food cues in adults with normal weight using machine learning. J Nutr 2022;152(6):1574-1581 [FREE Full text] [doi: 10.1093/jn/nxac053] [Medline: 35325189]
35. Siy Van VT, Antonio VA, Siguin CP, Gordoncillo NP, Sescon JT, Go CC, et al. Predicting undernutrition among elementary schoolchildren in the Philippines using machine learning algorithms. Nutrition 2022;96:111571. [doi: 10.1016/j.nut.2021.111571] [Medline: 35063809]
36. Granal M, Slimani L, Florens N, Sens F, Pelletier C, Pszczolinski R, et al. Prediction tool to estimate potassium diet in chronic kidney disease patients developed using a machine learning tool: the univerSel study. Nutrients 2022;14(12):2419 [FREE Full text] [doi: 10.3390/nu14122419] [Medline: 35745151]
37. Nguyen PH, Tran LM, Hoang NT, Tr ng DTT, Tran THT, Huynh PN, et al. Relative validity of a mobile AI-technology-assisted dietary assessment in adolescent females in Vietnam. Am J Clin Nutr 2022;116(4):992-1001 [FREE Full text] [doi: 10.1093/ajcn/nqac216] [Medline: 35945309]
38. Shao W, Min W, Hou S, Luo M, Li T, Zheng Y, et al. Vision-based food nutrition estimation via RGB-D fusion network. Food Chem 2023;424:136309. [doi: 10.1016/j.foodchem.2023.136309] [Medline: 37207601]
39. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-Thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. Applied Sciences 2023;13(12):7082 [FREE Full text] [doi: 10.3390/app13127082]
40. Mezgec S, Eftimov T, Bucher T, Koroušić Seljak B. Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment. Public Health Nutr 2019;22(7):1193-1202 [FREE Full text] [doi: 10.1017/S1368980018000708] [Medline: 29623869]
41. Allegra D, Battiato S, Ortis A, Urso S, Polosa R. A review on food recognition technology for health applications. Health Psychol Res 2020;8(3):9297 [FREE Full text] [doi: 10.4081/hpr.2020.9297] [Medline: 33553793]
42. Tahir GA, Loo CK. A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment. Healthcare (Basel) 2021;9(12):1676 [FREE Full text] [doi: 10.3390/healthcare9121676] [Medline: 34946400]
43. Detopoulou P, Voulgaridou G, Moschos P, Levidi D, Anastasiou T, Dedes V, et al. Artificial intelligence, nutrition, and ethical issues: a mini-review. Clinical Nutrition Open Science 2023;50:46-56 [FREE Full text] [doi: 10.1016/j.nutos.2023.07.001]
44. Belenguer L. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI Ethics 2022;2(4):771-787 [FREE Full text] [doi: 10.1007/s43681-022-00138-8] [Medline: 35194591]
45. Papathanail I, Abdur Rahman L, Brigato L, Bez NS, Vasiloglou MF, van der Horst K, et al. The nutritional content of meal images in free-living conditions-automatic assessment with goFOOD. Nutrients 2023;15(17):3835 [FREE Full text] [doi: 10.3390/nu15173835] [Medline: 37686866]
46. Zhou J, Wang Z, Liu Y, Yang J. Research on the influence mechanism and governance mechanism of digital divide for the elderly on wisdom healthcare: the role of artificial intelligence and big data. Front Public Health 2022;10:837238 [FREE Full text] [doi: 10.3389/fpubh.2022.837238] [Medline: 36062111]
47. Romero-Tapiador S, Lacruz-Pleguezuelos B, Tolosana R, Freixer G, Daza R, Fernández-Díaz CM, et al. AI4FoodDB: a database for personalized e-Health nutrition and lifestyle through wearable devices and artificial intelligence. Database (Oxford) 2023;2023:baad049 [FREE Full text] [doi: 10.1093/database/baad049] [Medline: 37465917]
48. Shei RJ, Holder IG, Oumsang AS, Paris BA, Paris HL. Wearable activity trackers-advanced technology or advanced marketing? Eur J Appl Physiol 2022;122(9):1975-1990 [FREE Full text] [doi: 10.1007/s00421-022-04951-1] [Medline: 35445837]
49. Sak J, Suchodolska M. Artificial intelligence in nutrients science research: a review. Nutrients 2021;13(2):322 [FREE Full text] [doi: 10.3390/nu13020322] [Medline: 33499405]
50. Kirk D, Kok E, Tufano M, Tekinerdogan B, Feskens EJM, Camps G. Machine learning in nutrition research. Adv Nutr 2022;13(6):2573-2589 [FREE Full text] [doi: 10.1093/advances/nmac103] [Medline: 36166846]
51. de Toro-Martín J, Arsenault BJ, Després JP, Vohl MC. Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome. Nutrients 2017;9(8):913 [FREE Full text] [doi: 10.3390/nu9080913] [Medline: 28829397]

XSL•FO
RenderX

52. Livingstone KM, Ramos-Lopez O, Pérusse L, Kato H, Ordovas JM, Martínez JA. Precision nutrition: a review of current approaches and future endeavors. Trends in Food Science & Technology 2022;128:253-264 [FREE Full text] [doi: 10.1016/j.tifs.2022.08.017]

Abbreviations

AI: artificial intelligence
AIM: Automatic Ingestion Monitor
CNN: convolutional neural network
DNN: deep neural network
EDFN-D: encoder-decoder food networks with depth-refinement
FCN: Fully Convolutional Neural Network
FRANI: Food Recognition Assistance and Nudging Insights
GAN: Generative Adversarial Network
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
ResNet: residual networks
RGB-D: Red, Green, Blue-Depth
SVM: support vector machine
USDA: US Department of Agriculture
WR: weighed record

Edited by A Mavragani; submitted 14.11.23; peer-reviewed by TAR Sure, S Kommireddy, M Elbattah, T Davies, T Baranowski; comments to author 30.05.24; revised version received 18.07.24; accepted 08.10.24; published 28.11.24.

Please cite as:
Zheng J, Wang J, Shen J, An R
Artificial Intelligence Applications to Measure Food and Nutrient Intakes: Scoping Review
J Med Internet Res 2024;26:e54557
URL: https://www.jmir.org/2024/1/e54557
doi:10.2196/54557
PMID:

©Jiakun Zheng, Junjie Wang, Jing Shen, Ruopeng An. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Interventions for Symptoms of Borderline Personality Disorder: Systematic Review and Meta-Analysis

Julia A B Lindsay[1], BSc; Niall M McGowan[1], PhD; Thomas Henning[2], BA; Eli Harriss[3], MSc; Kate E A Saunders[1,2], MRCPsych, DPhil

[1]Department of Psychiatry, University of Oxford, Oxford, United Kingdom

[2]John Radcliffe Medical School, University of Oxford, Oxford, United Kingdom

[3]Bodleian Health Care Libraries, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Julia A B Lindsay, BSc
Department of Psychiatry
University of Oxford
Warneford Hospital
Warneford Ln
Oxford, OX3 7JX
United Kingdom
Phone: 44 01865613196
Email: julia.lindsay@psych.ox.ac.uk

Abstract

Background: Borderline personality disorder (BPD) is a mental health condition with insufficient care availability worldwide. Digital mental health interventions could reduce this treatment gap. Persuasive system design (PSD) is a conceptual framework outlining elements of digital interventions that support behavior change.

Objective: This systematic review aims to characterize digital interventions targeting BPD symptoms, assess treatment efficacy, and identify its association with intervention features, including PSD elements.

Methods: A systematic review of automated digital interventions targeting symptoms of BPD was conducted. Eligible studies recruited participants aged ≥18 years, based on a diagnosis of BPD or one of its common comorbidities, or as healthy volunteers. OVID Embase, OVID MEDLINE, OVID PsycINFO, and the Cochrane Central Register for Controlled Trials were searched on July 19, 2022, and February 28, 2023. Intervention characteristics were tabulated. A meta-analysis of randomized controlled trials (RCTs) determined treatment effects separately for each core symptom of BPD using Hedges *g*. Associations between the treatment effect and intervention features, including PSD elements, were assessed by subgroup analysis (Cochran *Q* test). Risk of bias was assessed using the Cochrane Risk of Bias 2 tool for RCTs and the National Institutes of Health Quality Assessment Tool for pre-post studies.

Results: A total of 40 (0.47%) publications out of 8520 met the inclusion criteria of this review, representing 6611 participants. Studies comprised examinations of 38 unique interventions, of which 32 (84%) were RCTs. Synthesis found that included interventions had the following transdiagnostic treatment targets: severity of BPD symptoms (4/38, 11%), suicidal ideation (17/38, 45%), paranoia (5/38, 13%), nonsuicidal self-injury (5/38, 13%), emotion regulation (4/38, 11%), and anger (3/38, 8%). Common therapeutic approaches were based on dialectical behavioral therapy (8/38, 21%), cognitive behavioral therapy (6/38, 16%), or both (5/38, 13%). Meta-analysis found significant effects of digital intervention for both symptoms of paranoia (Hedges g=–0.52, 95% CI –0.86 to –0.18; P=.01) and suicidal ideation (Hedges g=–0.13, 95% CI –0.25 to –0.01; P=.03) but not overall BPD symptom severity (Hedges g=–0.17, 95% CI –0.42 to 0.10; P=.72). Subgroup analysis of suicidal ideation interventions found that evidence-based treatments such as cognitive behavioral therapy and dialectical behavior therapy were significantly more effective than alternative modalities (Cochran Q=4.87; P=.03). The degree of human support was not associated with the treatment effect. Interventions targeting suicidal ideation that used reminders, offered self-monitoring, and encouraged users to rehearse behaviors were associated with a greater reduction in ideation severity.

Conclusions: Evidence suggests that digital interventions may reduce the symptoms of suicidal ideation and paranoia and that the design of digital interventions may impact the efficacy of treatments targeting suicidal ideation. These results support the use of transdiagnostic digital interventions for paranoia and suicidal ideation.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD42022358270; https://tinyurl.com/3mz7uc7k

(J Med Internet Res 2024;26:e54941) doi:10.2196/54941

KEYWORDS

borderline personality disorder; BPD; digital intervention; digital health; digital therapeutics; persuasive system design; systematic review; meta-analysis; suicidal ideation; paranoia; single symptom; mental health; behavior change; treatment; effectiveness; symptom; suicide; mobile phone

Introduction

Borderline Personality Disorder

Borderline personality disorder (BPD) is a mental health condition characterized by 9 core symptoms usually involving intense and volatile emotional reactivity, identity disturbance, and interpersonal difficulties; however, symptom presentation is highly heterogeneous [1]. People seeking treatment for BPD represent a significant proportion of those referred to psychiatric clinical services: about 22% of inpatients and 12% of outpatients globally [2]. The demand for BPD treatment vastly outstrips availability, with nearly 6000 people seeking treatment for the condition per evidence-based care provider in the United States. Furthermore, the cost per patient is estimated to be more than double that of patients with depression [3,4].

This treatment gap could be reduced by digital interventions, which inform users and support behavior changes to reduce symptom severity and functional impairment [5,6]. Although few digital interventions are specifically designed for BPD, many address transdiagnostic symptoms, making them potentially effective for BPD treatment due to the disorder's diverse symptom presentation. For example, many digital interventions have a treatment target of emotion regulation, a core symptom of BPD that is also common in mood, substance use, and eating disorders [7]. People with BPD tend to use less effective emotion regulation strategies than healthy controls [8]. Poor emotion regulation is associated with behavioral control and drives maladaptive behaviors such as impulsive anger and nonsuicidal self-injury (NSSI), which are also symptoms of BPD [9-11]. Paranoia is another transdiagnostic symptom that can be treated effectively, although few interventions focus specifically on paranoia in BPD [12]. Symptom-based care offers treatment without focusing on a specific diagnosis, which could help reduce pathologization—or the unnecessary labeling of illness—in individuals diagnosed with BPD, many of whom are opposed to the label of personality disorder [13]. Digital interventions offer advantages such as access without diagnosis, lower cost, greater anonymity, and continuous availability. Despite concerns about limited clinician contact, self-guided digital interventions can be as effective as traditional face-to-face alternatives [14]. However, several challenges remain, such as nonadherence, limited assessment of efficacy in controlled trials, and poor user experience [15,16].

Persuasive System Design

The importance of user experience in digital health has been demonstrated by Kelders et al [17], who showed that persuasive system design (PSD) was associated with increased adherence to digital interventions. The concept of PSD proposes that user experience and interface elements, detailed in Tables 1 and 2, support behavior change in users [18]. For example, reminders to engage with the intervention and opportunities to rehearse new skills may help users reach their treatment goals by maintaining their adherence to behavioral and cognitive interventions. Research on adherence typically compares characteristics between adherers and nonadherers, thereby putting the burden of adherence onto the user [19]. However, given the evidence for PSD, adherence may be a product of the interaction between the user and the interface rather than a reflection of the users' characteristics.

XSL•FO
RenderX

Table 1. Coding scheme for the persuasive system design elements that offer primary task support[a].

Primary task support	Concept	Implementation example
Reduction	Breaking goals down into smaller tasks, for example, skills modules, makes them easier to achieve	DBT[b] Coach [20]: presents the user with a series of tasks depending on the DBT skill being practiced in each session
Tunneling	Guiding the user through tasks in a predefined order reduces the mental burden of learning by making it obvious what should be done next	Life Buoy [21]: the user must complete each module in the sequence to unlock the next
Tailoring	The interventions' content and user interface is adapted to suit different groups of users. This content-matching means users receive more relevant information or guidance to help them reach their goals	Man Therapy [22]: external resources signposted depend on the US state in which the user is located
Personalization	The interventions' content and user interface is adapted for each user or provides options that users may choose between. Personalization may increase engagement by helping the user feel seen	Be a Mom [23]: intervention uses the names of the user (a mother) and her baby
Self-monitoring	Providing users with either visual indicators of their progress or opportunities for self-reflecting on progress supports their motivation to reach their goals	mDiary [24]: 10 mood-related variables are logged daily, and charts showing change over time are presented to the user
Simulation	Estimating the cause and effect of different behavioral trajectories can motivate users to change or maintain their behavior	Sleep Scholar [25]: shares information about the consequences of poor sleep strategies
Rehearsal	Practicing skills in a comfortable environment builds the habit of using them as needed	Johnson et al [26]: Users are encouraged to practice relaxation skills and create an implementation intention to use their skills

[a]The scheme and table are adapted from the study by Kelders et al [17].

[b]DBT: dialectical behavior therapy.

Table 2. Coding scheme for the persuasive system design elements that offer dialogue support[a].

Dialogue support	Concept	Implementation example
Rewards	Incentivizes engagement with the intervention. In this analysis, participant compensation is counted as a reward	TEC[b] [27]: awards points for each TEC trial that the user completes. In the trial, participant compensation was based on the number of points earned
Reminders	Prompting users may increase the likelihood of their sustained engagement with the intervention and prevent forgetful nonadherence. Reminders could be related to intervention use or the target behavior	priovi [28]: users can choose to register for daily e-mails and SMS text messages
Suggestion	Describing and endorsing target behavior may encourage change in users	Bernstein et al [29]: throughout the day, the user is prompted to use emotion regulation skills toward any "current distressing emotions"
Similarity	Users may resonate more with interventions that feature people like the users or environments like their own	FitMindKit [30]: modules are presented by characters who have the same mental health challenges as the user
Liking	Appealing and cohesive user interfaces may promote engagement with the intervention	Life Buoy [21]: features a sailing-themed interface in which each module is represented by an island. The interface was designed in consultation with a lived-experience group
Social role	Interventions that take on a familiar anthropomorphic role (eg, coach, instructor, or buddy) may be more naturally adopted by users	FitMindKit [30]: features an expert narrator who takes on the role of the instructor

[a]The scheme and table are adapted from the study by Kelders et al [17].

[b]TEC: therapeutic evaluative conditioning.

The proposed mechanism of action is that PSD elements increase user adherence, which in turn supports greater treatment effect. Different elements have different effects: some may reduce goal behaviors into smaller, manageable changes, while others promote frequent use or lend increased credibility to the intervention [17]. Accordingly, this review aims to (1) identify and describe the therapeutic approach, duration, intended frequency of use, and user interface of digital interventions treating BPD symptoms in community and secondary care settings; (2) conduct a meta-analysis to determine the efficacy of digital interventions compared to active or passive control groups; and (3) determine using subgroup analysis whether there were associations between treatment efficacy and characteristics of the intervention, such as the therapeutic

XSL•FO
RenderX

modality, the degree of human support, and the use of PSD elements. Identifying characteristics associated with treatment efficacy will help optimize the design of future digital psychological interventions.

Novel Contributions

Existing reviews of digital interventions in this area have not thoroughly considered the implementation and user experience of digital interventions, nor have they determined treatment effects for individual symptoms of BPD. Frias et al [31] conducted a scoping review of 15 studies of digital interventions in participants with BPD, which were primarily adjunctive to dialectical behavior therapy (DBT) but did not calculate their treatment effect. A meta-analysis of smartphone apps for BPD-related symptoms did not find a significant treatment effect. However, this meta-analysis pooled outcome measures across symptoms, potentially obscuring item-level changes for individual BPD symptoms [32]. Other reviews have focused on the user experience of these interventions, with a scoping review of 8 interventions, primarily DBT tools, reporting positive user feedback, while another review found serious issues with functionality and interface design, such as the exclusion of DBT diary cards, technical issues with enrollment, and difficulties with navigation [33,34]. This review provides a comprehensive evaluation of digital interventions that target both specific and transdiagnostic symptoms of BPD. Treatment effects for individual symptom measures and an analysis of the effect of user experience features and PSD elements on treatment efficacy are also examined.

Methods

Search Strategy

OVID Embase, OVID MEDLINE, OVID PsycINFO, and the Cochrane Central Register for Controlled Trials were searched for literature. The search strategy [35], further discussed in Multimedia Appendix 1, was developed in consultation with coauthor EH and consisted of three themes: (1) the core symptoms of BPD, (2) digital health, and (3) intervention studies. The digital health theme was based on a National Institute for Health and Care Excellence (NICE)–validated filter for health care applications, with additional terms added for web-based interventions [36]. Automatic filters were used to remove non-English language studies and those published before the year 2000, as the technology of interest did not exist before that time. The search was first executed on July 19, 2022, and was then re-executed on February 28, 2023, to check for new literature. Furthermore, we performed forward reference checking on the included publications to enhance the literature search.

Eligibility Criteria

All records were screened by coauthors JABL and TH independently, and disagreements were settled in consultation with NMM. Rayyan (Rayyan Systems, Inc) was used to manage the screening process [37]. Studies were eligible if they were peer-reviewed, published in English after the year 2000, and assessed an automated digital intervention with a treatment target of BPD or one of its symptoms. Here, *automated* indicates an intervention for which content is primarily delivered without human support. Given the possibility of preliminary studies in this area, both randomized controlled trials (RCTs) and nonrandomized studies were eligible for inclusion in the descriptive analysis. All participants were aged ≥18 years. They were either healthy volunteers or individuals with a diagnosis of BPD or its common comorbidities of depressive disorders, anxiety disorders, substance abuse disorders, posttraumatic stress disorder, or complex posttraumatic stress disorder. Studies drawing on community samples were included because they tended to recruit participants with mild to moderate symptom severity, which corresponded with the degree of BPD symptom severity that would be reasonable to treat with an automated digital intervention. This aligns with the NICE guidelines for digital interventions, suggesting professional oversight should increase proportionally with clinical risk [38]. Exclusion criteria were as follows: (1) interventions without a primary digital component; (2) nonautomated interventions (eg, therapist-centered treatment via telehealth); (3) interventions requiring equipment inaccessible to the public (eg, functional magnetic resonance imaging or professional-caliber virtual or augmented reality); (4) participants recruited based on a mental health or neurodevelopmental disorder not mentioned in the inclusion criteria; or (5) studies focusing on participants belonging to special groups such as combat veterans, people who are displaced, or law enforcement officers. These groups were excluded because the life experiences of these participants were unlikely to be representative of a community sample of people with BPD.

Quality Assessment

The risk of bias was assessed with Cochrane's Risk of Bias 2 for RCTs including cluster-randomized trials and the National Institutes of Health Quality Assessment Tool for pre-post studies for all single-arm studies [39,40]. All studies were separately evaluated by coauthors JABL and TH, and discrepancies were settled by consensus.

Extracted Data

The following features were extracted for each intervention: method of delivery, therapeutic approach, overall intended treatment duration and frequency of use, degree of human support, and any PSD elements present. Data were extracted from the published literature and, where possible, supplemented by direct access to publicly available interventions and interviews with authors. A reliability check of the coding of PSD elements was carried out in a random sample of 5 papers, resulting in 91% interrater reliability. Therapeutic approaches were coded individually to describe each study. To facilitate the subgroup analysis, therapeutic approaches were then classified into evidence-based treatments (EBTs) and non-EBTs. EBTs comprised DBT, cognitive behavioral therapy (CBT), mentalization-based treatment, acceptance and commitment therapy, schema therapy, and transference-focused therapy. The degree of human support of the intervention was binary coded for either full automation or some degree of support based on the descriptions provided in the publications. Facilitated interventions ranged from technical support to therapeutic support to adjunct to in-person care.

XSL•FO
RenderX

PSD Elements

We used the coding scheme by Kelders et al [17] for PSD elements, based on a framework developed by Oinas-Kukkonen and Harjumaa [18] (Tables 1 and 2). All interventions were coded dichotomously, indicating the presence or absence of primary task, dialogue, and social support elements. In cases where the presence of PSD elements was unclear in the publication, we contacted the authors and accessed the intervention directly where possible. Failing these means, the element was coded as absent.

Effect Sizes and Meta-Analysis

Treatment effects for the RCTs were calculated using between-group standardized mean differences (SMDs) at the study end point, with Hedges small sample correction applied [41]. Hedges *g* values of 0.0-0.2 were interpreted as small, 0.21-0.8 as moderate, and 0.81-1.00 as large. Outcome data were extracted directly from the publications or requested from the study lead author. In cases (2/32, 6%) where the authors did not reply or were unable to provide the requested data, the study was excluded from the meta-analysis [42,43].

Six separate meta-analyses were used to assess the effect of treatment for which there were at least 2 studies reporting these as outcomes. These involved the following outcomes: BPD symptom severity assessed via the Zanarini Rating Scale for Borderline Personality Disorder [44] and Borderline Personality Disorder Severity Index [45]; suicidal ideation via the Beck Scale for Suicidal Ideation (BSSI) [46], Suicidal Ideation Attributes Scale [47], Self-Injurious Thoughts and Behaviours Interview–Suicidal Ideation [48], Columbia-Suicide Severity Rating Scale [49], Suicidal Behaviours Questionnaire [50], and Suicide Status Form [51]; NSSI via the Self-Injurious Thoughts and Behaviours Interview–Nonsuicidal Self-Injury Episodes [48]; paranoia via the Paranoia Scale [52] and Adapted Paranoia Checklist [53]; and anger via the Modified Overt Aggression Scale [54] and Trait Anger Scale [55]. As emotion regulation is considered a central mechanism in BPD, a meta-analysis was also conducted with the following outcome measures: the Difficulties with Emotion Regulation Scale in both long and short forms [56] and the German version of the Emotion Regulation Skills Questionnaire [57]. Following the Cochrane Collaboration recommended procedures [58], in cases where multiple different outcome measures of the same symptom were used in a single study, the effect sizes of each measure were calculated and then aggregated before meta-analysis using the approach by Borenstein et al [59].

Random effects models were chosen over fixed-effects as considerable between-study heterogeneity was expected due to differences in control arms [60]. I^2 was used to estimate between-study heterogeneity, and restricted maximum likelihood was used to estimate heterogeneity variance τ^2 [61]. The Knapp-Hartung adjustment was applied, given the small number of studies in some of the meta-analyses [62]. Funnel plots were visually inspected for risk of publication bias. Egger tests were conducted in meta-analyses with sufficient studies [63].

Subgroup Analysis

Following recommendations from Schwarzer et al [64] and Fu et al [65], in meta-analyses with ≥10 studies, subgroup analyses were conducted to determine the effects of specific intervention elements, provided there were a minimum of 4 studies per subgroup. Intervention features assessed using subgroup analysis included PSD elements, degree of human support, and use of EBT versus non-EBT interventions. τ^2 was pooled across subgroups because of the small number of studies in some subgroups [59]. The Cochran *Q* test was used to evaluate the significance in subgroup differences with $P<.10$ considered significant for the *Q* test owing to low event rates [66-68]. The Benjamini-Hochberg false discovery rate was used to correct for multiple comparisons [69].

All meta-analytic procedures were done using the *R* (R Foundation for Statistical Computing) packages: *meta* (version 6.2.1), *dmetar* (version 0.1.0), and *Mad* (version 0.8-3) packages [64,66,70].

Registration

This systematic review was preregistered with PROSPERO (CRD42022358270). The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist can be found in Multimedia Appendix 2. Some deviations were made from the registered protocol: first, gray literature was not included. This decision was made after reviewing the gray literature, as we believed including it could lower the average quality of the studies and increase their overall risk of bias. Furthermore, most of the gray literature was not peer-reviewed and thus did not meet the inclusion criteria for this review. Adherence was not used as an outcome measure of the review because only 10% (4/42) of the included studies described adherence as the number of participants who completed the intervention. Other adherence metrics were reported, but these were inconsistent and could not be organized into a meaningful measure. Finally, some studies listed multiple outcome measures without identifying a primary measure. For this review, all studies reporting at least 1 appropriate outcome measure were included. Studies of interventions with multiple treatment targets and outcome measures were eligible to be included in multiple syntheses. For example, some interventions addressed both suicidal ideation and NSSI and thus are included in both syntheses.

Results

Search Results

Figure 1 illustrates the PRISMA flowchart of the initial and repeated searches. The first search, executed on July 19, 2022, returned 7531 records, of which 1645 (21.84%) were duplicates. A further 428 (5.68%) were automatically removed due to their year of publication (before 2000) or initial publication in a non-English language. Of the remaining 5458 records, 69 (1.26%) were retrieved for full-text screening and 36 (0.66%) of these were included. A repeated search before analysis with publication dates restricted from the initial search date to February 28, 2023, returned 989 records, of which 665 (67.24%) were duplicates. The remaining 324 records were screened, 22

XSL•FO
RenderX

(6.79%) were selected for full-text screening, and 4 (1.23%) of those were included. The details of all included studies are provided in Tables S1-S6 in Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the search and study selection process. For simplicity, the initial and repeat searches have been combined into 1 diagram.

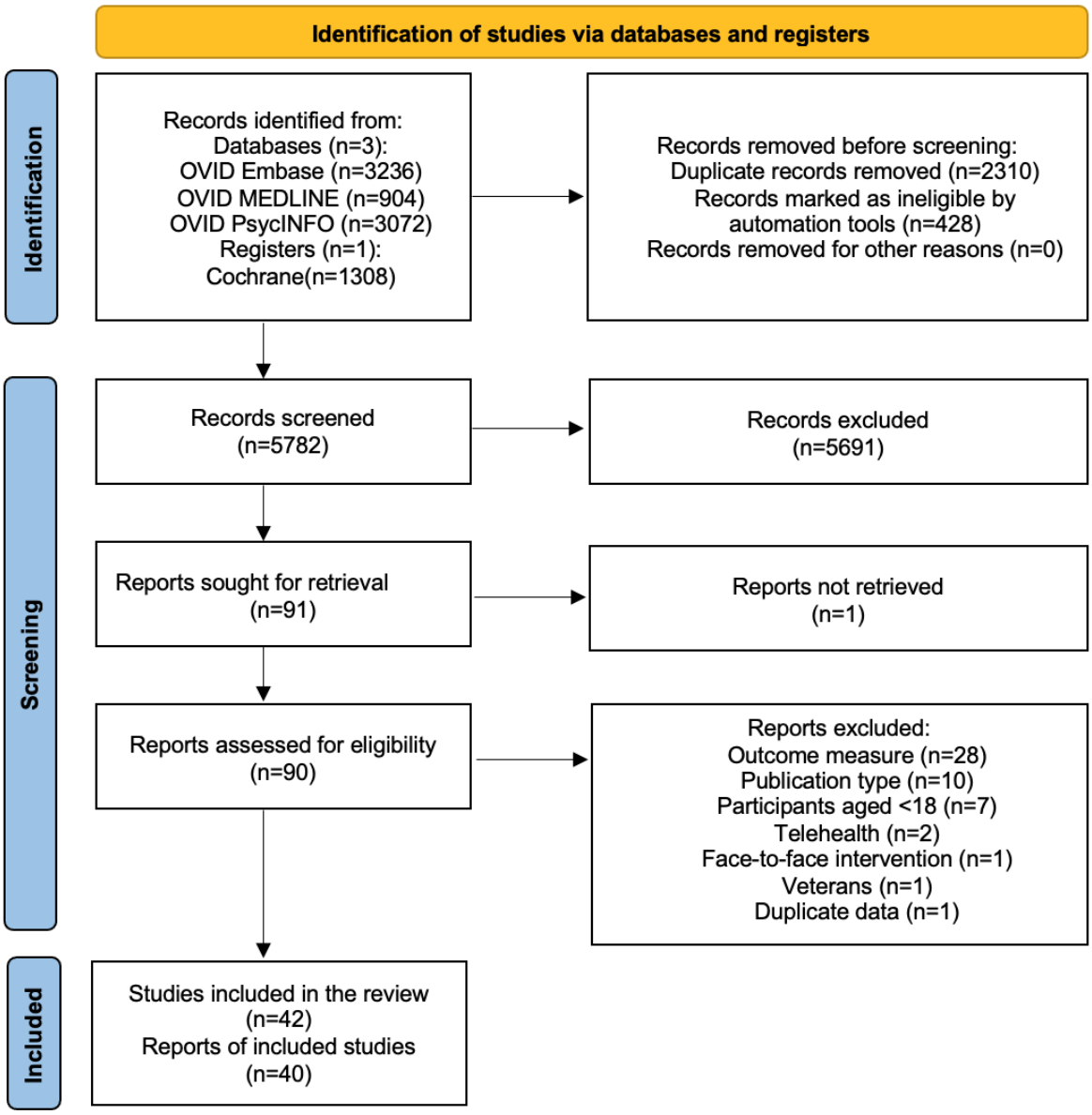

Search Outcomes

The most common reason for exclusion was an inappropriate outcome measure, including unvalidated measures and subscales. Other reasons for exclusion were lack of peer review, enrollment of participants aged <18 years, interventions that were delivered in person or via telehealth meeting software, and research about combat veterans. A total of 40 reports met the eligibility requirements of this review, 1 (3%) of which detailed 3 studies in the same report, such that 42 unique studies were included. Across these studies, 38 unique digital interventions were assessed and discussed in this review. Some studies assessed the same intervention, for example, translations of Living with Deadly Thoughts into multiple languages [71-75]. One study had 3 arms, each assigned a different intervention [76]. In total, 38 unique interventions are discussed in this review. Of the 42 studies included in the review, 32 (76%) were RCTs, 8 (19%) were single-arm studies (pre-post), 1 (2%) was an open-label trial, and 1 (2%) was a cluster RCT. Four studies reported BPD symptoms or key mechanisms as outcome measures, 20 reported suicidal ideation and suicidal thinking, 5 studies each reported paranoia and NSSI, 4 reported emotion regulation, and 4 reported anger and hostility measures. Some interventions listed multiple outcome measures without indicating a primary. No interventions were identified targeting identity disturbance, impulsivity, feelings of emptiness, fears

XSL•FO
RenderX

of abandonment, or relationship instability. Participants were recruited based on a variety of conditions: of the 6611 participants, 3372 (51%) were recruited based on suicidal ideation, 661 (10%) based on NSSI, 397 (6%) on a diagnosis of BPD, 331 (5%) on trait anger or hostility, and 198 (3%) on paranoia. In total, 19% (1256/6611) of the participants were recruited without specific baseline symptoms and 6% (396/6611) were recruited to studies with multiple possible symptom thresholds (eg, they met symptom thresholds for either suicidal ideation or NSSI).

Risk of Bias

The results of the Cochrane's Risk of Bias 2.0 and the National Institutes of Health Quality Assessment Tool for each study can be found in Figures S1-S5 in Multimedia Appendix 1. In brief, there was a widespread risk of bias: only 9% (3/32) of the RCTs and 13% (1/8) of the single-arm studies were deemed low risk.

Intervention Features and System Design

The included studies were primarily conducted in the United States (16/42, 38%), the United Kingdom (6/42, 14%), or Australia (5/42, 12%). The methods of intervention delivery included websites (27/42, 64%), mobile phone apps (10/42, 24%), both (3/42, 7%), and email (2/42, 5%). The most common therapeutic approaches used were based on DBT (8/42, 19%), CBT (6/42, 14%), a combination of CBT and DBT (5/42, 12%), mindfulness (3/42, 7%), and acceptance and commitment therapy (2/42, 5%). Out of the 42 studies, 18 (43%) evaluated other or unspecified therapeutic approaches.

The mean course of treatment, unweighted by the number of participants, was 55 (SD 75) days, with a mean unweighted recommended frequency of use of 5 (SD 2.7) times per week. This frequency calculation excluded 26% (11/42) of the studies with open dosage or intended frequency of use not stated, as well as 10% (4/42) of single-session interventions.

Explanations for each PSD element and examples of relevant interventions are given in Tables 1 and 2, and the frequency of PSD elements used in each of the digital interventions is illustrated in Figure S6 in Multimedia Appendix 1. The presence of PSD elements varied from *suggestion*, used in 86% (36/42) of the studies to *praise*, for which we only found evidence in 10% (4/42) of the interventions. Elements of social support were not analyzed, as "e-motion" was the only intervention that allowed users to interact with one another [77].

BPD Interventions

Study Characteristics

BPD psychopathology outcome measures were identified in 4 studies, including 3 RCTs, as outlined in Table S1 in Multimedia Appendix 1 [28,78-80]. Published between 2017 and 2021, these 4 studies recruited 376 participants in total, of which 303 (74%) provided follow-up data. Three of the studies had a moderate risk of bias [28,78,79], while 1 had a high risk of bias [80].

RCT Effect Sizes (Posttreatment SMDs, Hedges g)

Data from the 3 RCTs was pooled, and a random effects model was fitted. Figure 2 [78-80] shows a small and nonsignificant treatment effect for digital interventions on BPD symptoms as a whole (N=3; Hedge g=–0.17, 95% CI –0.42 to 0.10; P=.11). The between-study heterogeneity variance was estimated at τ^2=0 (95% CI 0.0-0.54) and I^2=0% (95% CI 0%-89.6%). The funnel plot does not show evidence of publication bias (Figure 3 [78-80]).

Figure 2. Forest plot of the treatment effect on borderline personality disorder symptom severity. SMD: standardized mean difference.

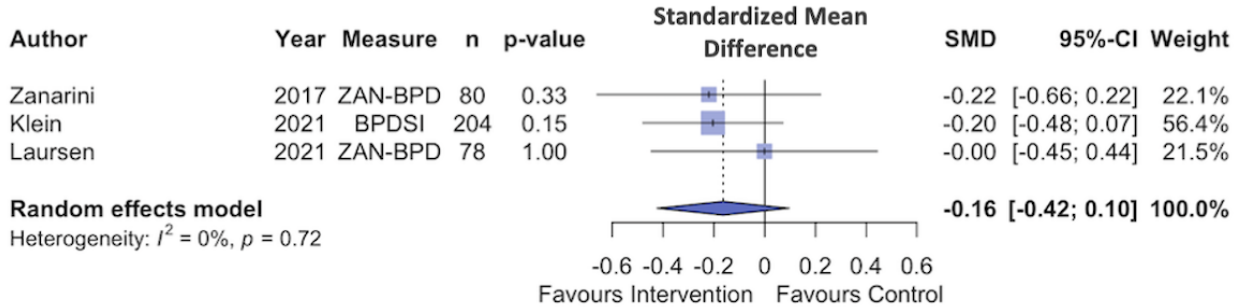

XSL•FO
RenderX

Figure 3. Funnel plot shows the treatment effect and SE of the interventions treating borderline personality disorder (BPD) psychopathology.

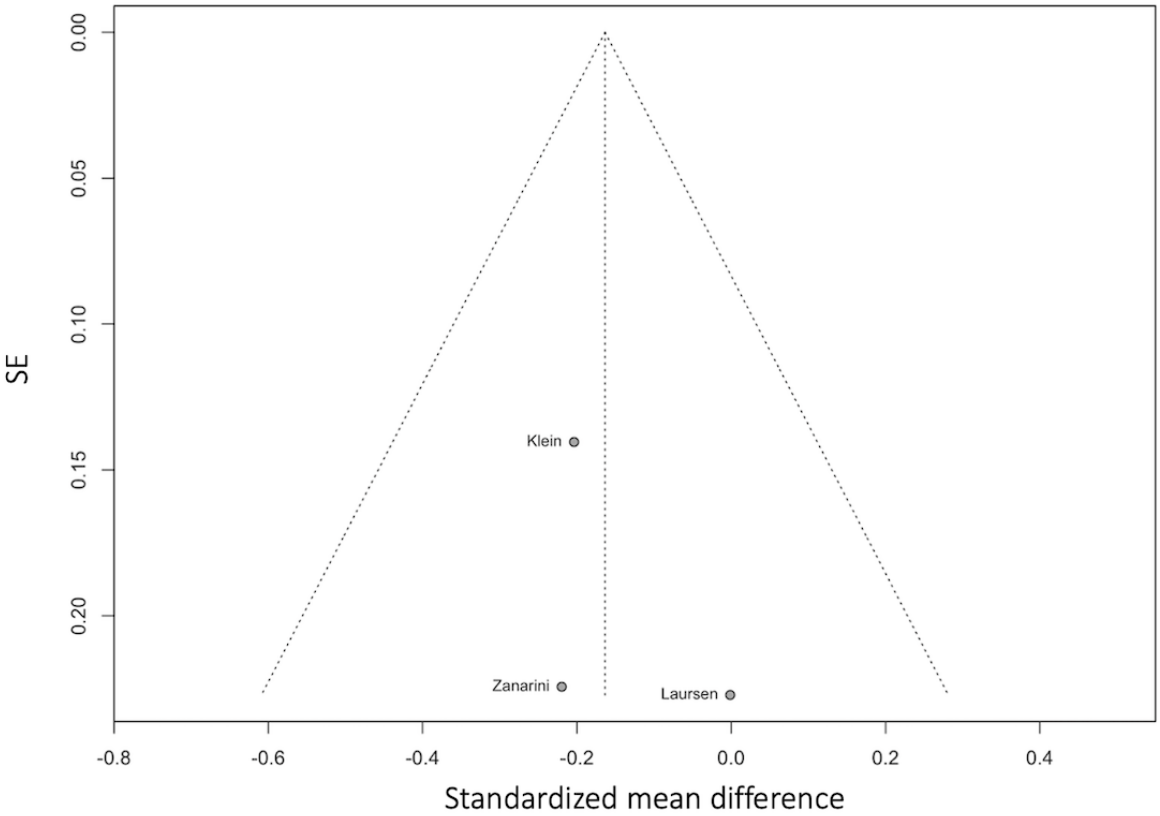

Suicidal Ideation Interventions

Study Characteristics

A total of 20 studies (18 RCTs) of interventions for suicidal ideation symptoms were identified, with publication years ranging from 2014 to 2023 and enrolling a total of 3983 participants (Table S2 in Multimedia Appendix 1 [21,22,25,27,30,71-76,80-86]). Of the 3983 participants, 2623 (66%) provided posttreatment data. Two (10%) of the studies were assessed at low risk of bias [75,81], 13 (65%) were at moderate risk of bias [21,25,27,71-74,76,82,83,86], and 5 (25%) were at high risk of bias [22,30,80,84,85].

Meta-Analysis

As shown in Figure 4 [21,22,27,30,71-76,80-84,86], the random effects model fitted to the data found that digital interventions elicited a small, significant reduction in suicidal ideation (N=18; Hedges g=–0.13, 95% CI –0.25 to –0.01; P=.03). Between-study heterogeneity variance was estimated at τ^2=0.3 (95% CI 0.01-0.10). Between-study heterogeneity was estimated at I^2=64.7% (95% CI 41.6%-78.6%). As referenced in the Methods section, effect sizes in studies with multiple outcome measures of suicidal ideation were aggregated before the meta-analysis. The funnel plot in Figure 5 [21,22,27,30,71-76,80-84,86] does not show evidence of publication bias, and the Egger test also did not show evidence of bias (intercept=1.33, 95% CI –0.2 to 2.86; 2-tailed t_{16}=1.70; P=.11).

XSL•FO
RenderX

Figure 4. Forest plot of the treatment effect for suicidal ideation interventions. BSSI: Beck Scale for Suicidal Ideation; C-SSRS: Columbia-Suicide Severity Rating Scale; DSI-SS: Depressive Symptom Index–Suicidality Subscale; SBQ: Suicidal Behaviours Questionnaire; SIDAS: Suicidal Ideation Attributes Scale; SITBI-SI: Self-Injurious Thoughts and Behaviours Interview–Suicidal Ideation; SSF: Suicide Status Form.

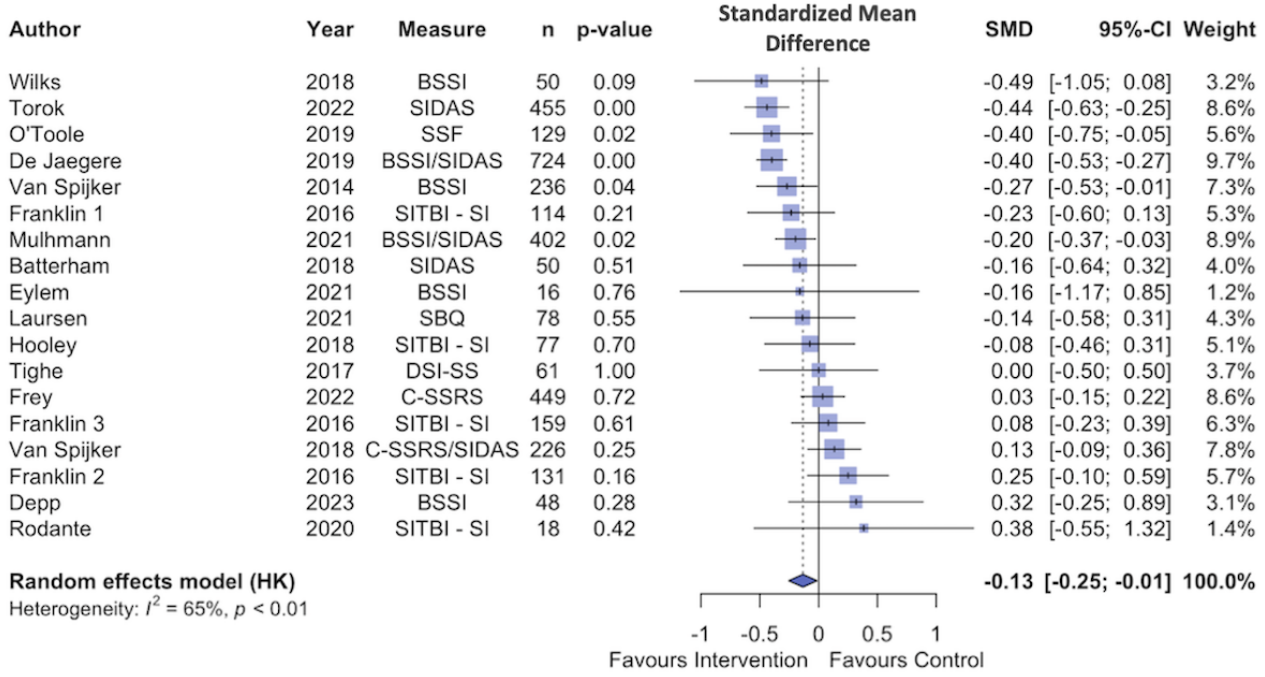

Figure 5. Funnel plot of the treatment effect and SE for suicidal ideation interventions.

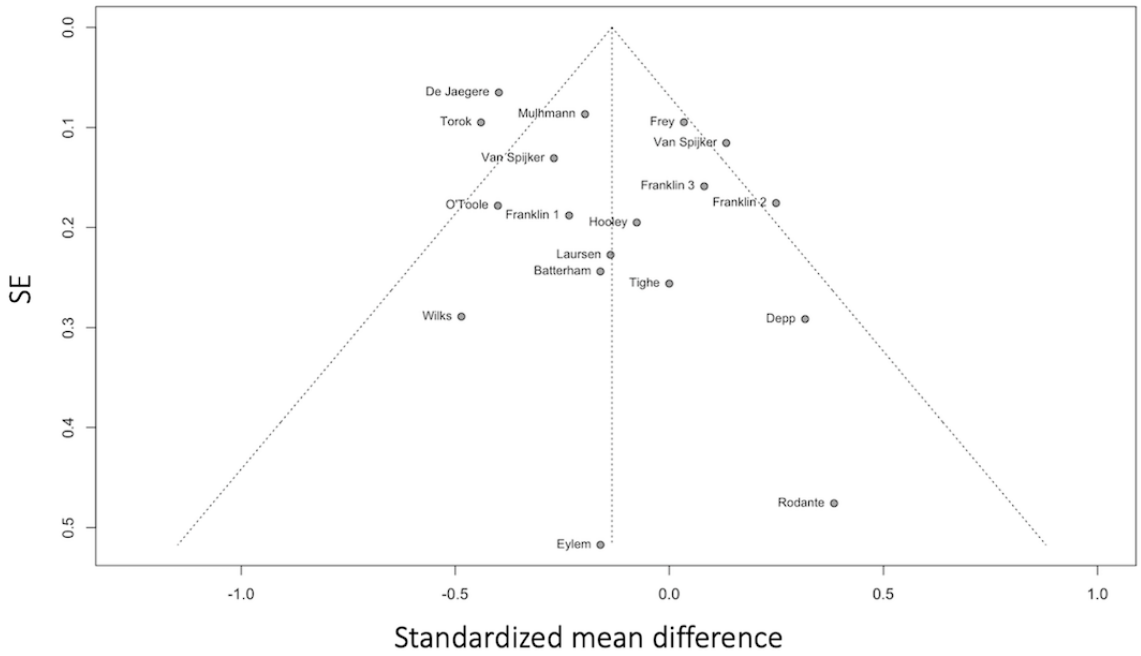

Subgroup Analysis: Effects of Suicidal Ideation Intervention Features

We first compared interventions targeting suicidal ideation that used EBTs (12/18, 67%) to those that did not (6/18, 33%). EBT interventions were associated with significantly greater treatment effect than non-EBT interventions (SMD –0.21 vs 0; Cochran Q=4.87; *P*=.03). We then compared suicidal ideation interventions that were fully automated to those that involved at least some degree of human support. There was not a significant difference in treatment effect between these 2 subgroups (automated SMD –0.09 vs facilitated SMD –0.16; Cochran Q=0.37; *P*=.54). Table 3 reports these results in detail, as well as subgroup analysis comparing treatment effect between interventions that did or did not use each PSD feature. For statistical power reasons, this analysis was only conducted for PSD elements that were present in at least 4 interventions and absent in at least 4 others. Interventions using reminders, opportunities for self-monitoring, and opportunities for rehearsal were associated with significantly greater reductions in suicidal ideation than those without. No significant effects were identified for other PSD elements.

XSL•FO
RenderX

Table 3. Difference in treatment effects between interventions using different therapeutic approaches, degrees of human support, and persuasive system design elements.

	Studies (n=18), n (%)	SMD[a] (95% CI)	*P* value	I^2 (%; 95% CI)	*P* value[b]
Therapeutic approach					.03[c]
EBT[d]	12 (67)	–0.21 (–0.37 to –0.06)	.01	61.4 (27.7-79.4)	
Non-EBT	6 (33)	0.00 (–0.17 to 0.18)	.98	0.0 (0-74.6)	
Support					.54
Human	12 (67)	–0.16 (–0.32 to –0.01)	.04	52.1 (7.67-75.2)	
None	6 (33)	–0.09 (–0.35 to 0.17)	.42	80.1 (56.8-90.8)	
Reduction					.18
Present	14 (78)	–0.18 (–0.32 to –0.04)	.02	65.3 (33.8-80.3)	
Absent	4 (22)	0.01 (–0.31 to 0.34)	.90	23.5 (0-88.3)	
Tunneling					.29
Present	5 (28)	–0.26 (–0.61 to 0.08)	.10	33 (0-74.6)	
Absent	13 (72)	–0.10 (–0.24 to 0.04)	.13	68 (43.0-82.1)	
Self-monitoring					.095[c]
Present	13 (72)	–0.20 (–0.35 to –0.06)	.01	56.1 (24.9-77.9)	
Absent	5 (28)	0.01 (–0.21 to 0.24)	0.87	50.3 (0-80.4)	
Rehearsal					.08[c]
Present	13 (72)	–0.21 (–0.36 to –0.07)	.01	58.2 (22.7-77.4)	
Absent	5 (28)	0.02 (–0.19 to 0.23)	.80	0 (0-79.2)	
Rewards					.21
Present	6 (33)	–0.01 (–0.29 to 0.28)	.95	37.4 (0-75.1)	
Absent	12 (67)	–0.19 (–0.33 to –0.05)	.01	66.8 (39.1-81.9)	
Reminders					.08[c]
Present	13 (72)	–0.21 (–0.35 to –0.07)	.01	58 (22.3-77.3)	
Absent	5 (28)	0.03 (–0.17 to 0.24)	.67	0 (0-79.2)	
Suggestion					.18
Present	14 (78)	–0.18 (–0.32 to –0.04)	.02	65.3 (38.8-80.3)	
Absent	4 (22)	0.01 (–0.31 to 0.34)	.90	23.5 (0-88.3)	
Liking					.99
Present	6 (33)	–0.13 (–0.39 to 0.12)	.23	65.6 (17.6-85.7)	
Absent	12 (67)	–0.13 (–0.29 to 0.03)	.10	67.1 (39.6-82.0)	

[a]SMD: standardized mean difference.

[b]*P* value for Cochran *Q* (significant at *P*<.10) [58] comparing SMD between subgroups. *P* value has been corrected for multiple comparisons using the Benjamini-Hochberg method.

[c]Significance at *P*<.10.

[d]EBT: evidence-based treatment.

Paranoia Interventions

Study Characteristics

From 42 studies, 5 (12%) RCTs of digital interventions were identified for paranoia, with publication years ranging from 2017 to 2021, totaling 413 enrolled participants (Table S3 in Multimedia Appendix 1 [87-91]). Many of the paranoia interventions were single sessions, so loss to follow-up percentage was not applicable. A total of 3 (60%) were at moderate risk of bias [87-89], while 2 (40%) were at high risk [90,91].

Meta-Analysis

A random effects model fitted to the data found a moderate, significant effect of digital interventions for paranoia (N=4;

XSL•FO
RenderX

Hedges g=–0.52, 95% CI –0.86 to –0.18; *P*=.01). The between-study heterogeneity variance was estimated at τ2=0.04 (95% CI 0-0.55), and between-study heterogeneity was estimated at I2=51.3% (95% CI 0%-82.1%; Figure 6 [87-91]). The funnel plot in Figure 7 [87-91] shows possible evidence of publication bias, but there were too few studies to conduct the test.

Figure 6. Forest plot showing standardized mean differences (SMDs) for interventions targeting paranoia. APC: Adapted Paranoia Checklist; PS: Paranoia Scale.

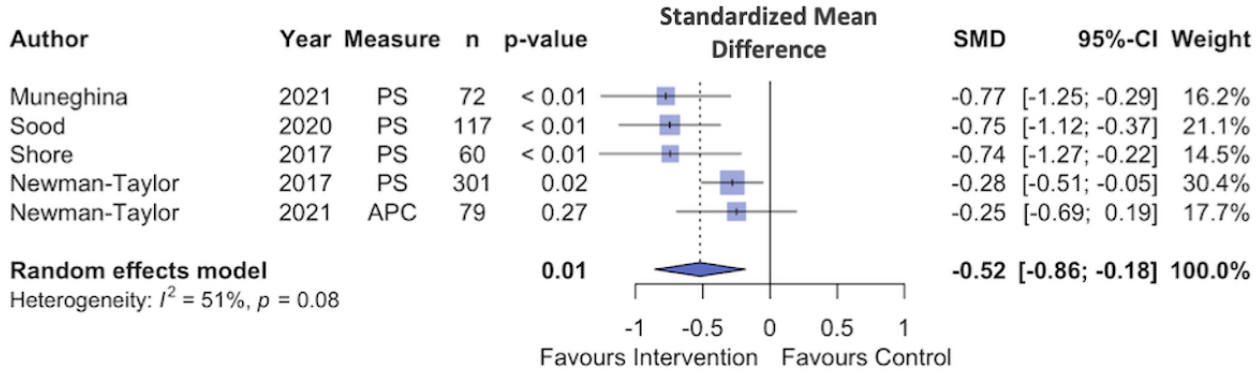

Figure 7. Funnel plot of the digital interventions targeting paranoia.

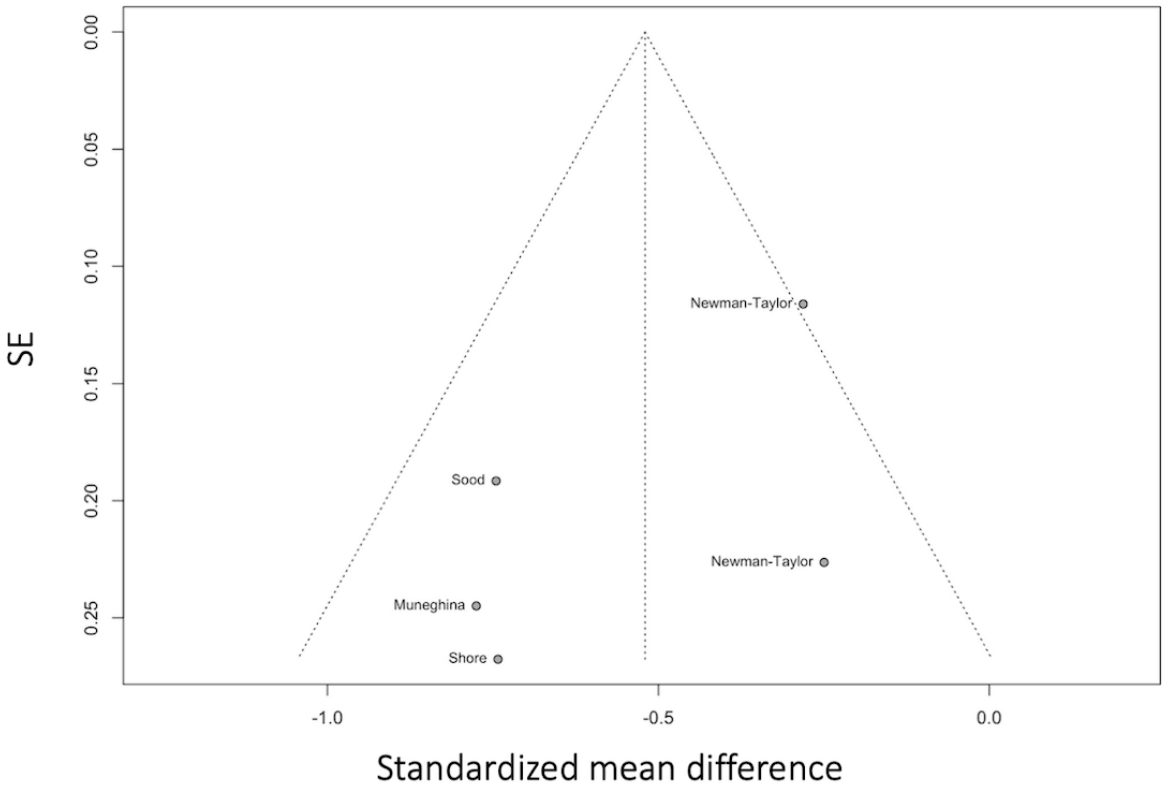

Other BPD Symptoms

Meta-analysis of interventions targeting other symptoms (NSSI, emotion regulation, and anger) did not suggest a significant treatment effect of digital interventions. Full meta-analytic results for these interventions can be found in Figures S7-S12 in Multimedia Appendix 1. Identified interventions are tabulated in Multimedia Appendix 1: NSSI interventions are in Table S4 [27,43,76,86], emotion regulation interventions are in Table S5 [23,29,77,92-94], and anger interventions are in Table S6 [26,95,96].

Discussion

Principal Findings

This review suggests that digital interventions may be efficacious in reducing BPD symptoms of paranoia and suicidal ideation. By contrast, digital interventions addressing emotional regulation, anger, and NSSI did not demonstrate a significant treatment effect. Among interventions targeting suicidal ideation, those based on evidence-based therapeutic modalities for BPD were associated with greater improvement, but effectiveness was not found to be dependent on the degree of human support. Importantly, these findings highlight which

XSL•FO
RenderX

symptoms may be better suited for digital interventions and could be candidates for further investigation and implementation. This is particularly relevant in clinically confirmed BPD, where limited access to in-person therapy and high service costs may be barriers to receiving timely care.

A secondary objective of this review was to assess the effect of PSD on treatment efficacy. There is an overlap between the aims of psychological interventions and the PSD framework: both support change in behavior and beliefs. At least some PSD element involvement was nearly ubiquitous among the studies reviewed; changes to thought and behavior (suggestion) and reduction of these changes into component processes (*reduction*) were used in 79% (33/42) and 86% (36/42) of the reviewed interventions, respectively. Importantly, our meta-analytic review of interventions targeting suicidal ideation found that the specific PSD elements *rehearsal*, *self-monitoring*, and *reminders* were associated with significantly greater treatment effect. Each element may confer different benefits to the design of digital interventions for suicidal ideation and per se parallel the therapeutic components of EBTs for suicidal ideation.

Rehearsal refers to practicing skills in a comfortable environment with the objective of building the habit of using them as needed. Rehearsing a response to periods of intense suicidal ideation may reduce the duration and risk of these events; this is a component in the Collaborative Assessment and Management of Suicidality approach [97]. Indeed, several evidence-based psychological interventions that address suicidal ideation such as DBT and CBT for suicide prevention involve rehearsal of applying the skills learned in therapy in advance of a challenging situation [98,99].

Self-monitoring in the context of PSD refers to keeping users motivated by encouraging reflection on progress toward goals. Both processes may address feelings of hopelessness and difficulty maintaining a course of action, which are associated with suicidal ideation and BPD [100,101]. Furthermore, self-monitoring, often with an emphasis on mood monitoring, is also a core aspect of interventions targeting emotional regulation and suicidality, such as DBT and CBT for suicide prevention.

Reminders are a design element that help reduce forgetful nonadherence [102]. Further work is necessary to understand which types of reminders are most effective for digital interventions. This PSD element is most likely to be digital-intervention specific, with no immediate comparative components in in-person EBTs.

In the context of existing therapeutic frameworks, these PSD features could be leveraged in the development of future digital interventions for suicidal ideation for BPD. Due care should be taken in the design of digital mental health interventions; the inclusion of PSD elements is a good starting point, but user experience should be evaluated and iterated upon, a process with established merits [103].

Implications

As in face-to-face settings, digital EBT interventions for suicidal ideation, such as CBT and DBT, were more effective than alternative therapeutic approaches [104]. EBTs appear to be effective for reasons that transcend delivery method, perhaps by introducing active components that directly facilitate treatment goals, such as behavioral activation in CBT, rather than general nondirective support [105]. Despite the demonstrated efficacy of EBTs, some included interventions used promising alternative therapeutic approaches, such as stimulus pairing or journaling [27,76].

Efficacy of digital intervention may also vary depending on the treatment target. Included visualization interventions were effective, at least temporarily, in reducing symptoms of paranoia. Paranoia in BPD is transient, meaning immediate symptom reduction is relevant even if temporary. Recent work with virtual reality in community and psychosis samples has found that compassionate imagery, cognitive therapy, and mental relaxation exercises all appear to be effective in the reduction of paranoia symptoms [106,107]. The commonality between these approaches is reduced anxious cognitions, which likely underpin paranoia and form a suitable target for digital interventions [108,109]. All included paranoia interventions were brief (ranging from 1 session to 2 weeks). The NICE guidance for BPD advises against brief interventions for BPD partially because few brief EBTs were available at the time of the original 2009 publication, and partially because of the possibility of negative reaction upon withdrawal of therapist support [110]. The latter issue is less applicable to automated digital interventions, particularly if they remain accessible after treatment. Given the efficacy of paranoia interventions in this review, and as brief interventions minimize the risk of user attrition over time, we suggest that further research should evaluate brief digital interventions for symptom relief and teaching specific skills. Such interventions could be deployed in adjunct or anticipation of in-person treatment.

Suicidal ideation also appears feasible as a digital treatment target. For example, Living with Deadly Thoughts, an intervention based on the principles of CBT and DBT that has been translated into multiple languages, has been found effective in several large RCTs. Collectively, the studies of Living with Deadly Thoughts contributed 35% (1604/4583) of the meta-analytical suicidal ideation results. The intervention focuses on reducing negative automatic thoughts, which are associated with suicidal ideation through several mediation pathways [111]. Reduction in negative automatic thoughts after CBT is associated with reduced suicidal ideation [112], a treatment pathway that appears to have been successfully digitally adapted in this series of interventions. For interventions targeting paranoia and suicidal ideation, cognitive therapy components may be best suited to digital formats. These components can help reduce anxious thoughts in paranoia and negative automatic thoughts in suicidal ideation.

Unlike paranoia and suicidal ideation, meta-analytic evidence did not show improvement in emotion regulation, anger, or NSSI after digital intervention. DBT, the gold standard treatment for BPD, has been shown to improve these symptoms likely by improving behavioral control [9,10]. Behavioral control supports emotion regulation, including impulsive anger, which is common in BPD, and, in turn, predicts NSSI behaviors [9,11,113]. DBT typically involves protracted membership in a therapeutic group, through which patients learn to maintain relationships with

XSL•FO
RenderX

peers. This process of supervised exposure and interaction with others is not replicable in automated digital interventions but may be essential to developing behavioral control. As a result, treatment targets of emotion regulation, anger, and NSSI may be less suited to the affordances of digital interventions than targets that can be effectively addressed through cognitive therapies, such as paranoia and suicidal ideation [12,114].

Previous research has found that human-supported digital mental health interventions are more effective than fully automated ones [14,115]. This was not replicated in our subgroup analysis of suicidal ideation interventions; however, human support could deter engagement with suicidal ideation treatments due to stigma and shame, thus reducing treatment efficacy for human-supported interventions [116]. This may not be true for other treatment targets; nonetheless, the effects of anonymity and human support should be weighed carefully in the development of digital interventions.

Our results suggest that effective digital interventions focus on building specific skills through transdiagnostic evidence-based approaches or provide symptom management tools [47]. This review did not find evidence that digital interventions should attempt comprehensive BPD treatment, as these interventions likely require support from a trained clinician and repeated opportunities to build emotion regulation skills among peers. The affordances of digital interventions are likely better suited to targeted, single-symptom treatments that could be deployed adjunctively to in-person care. For people with BPD, this type of intervention may be less pathologizing as the goal is to treat the symptoms of the individual rather than the disorder. The single-symptom approach is increasingly relevant as the diagnosis of personality disorder moves toward a dimensional model [117]. For users with and without personality disorders, the brevity and content relevance of transdiagnostic single-symptom interventions may improve adherence to digital interventions relative to extended courses of comprehensive treatment. Change in single-symptom severity may also be easier to measure because single-symptom measures provide more granularity than composite diagnostic measures [118].

Limitations

There are several limitations to this review, the first being the limited applicability of findings across symptoms. For example, the association between PSD elements and treatment effect was only analyzed in interventions for suicidal ideation and may not hold for all treatment targets. In addition, while the scope of this review includes interventions for BPD symptoms, many of the included studies draw on non-BPD samples. As a result, our findings address the optimization of digital interventions for symptom reduction, rather than BPD psychopathology specifically. These results remain relevant to BPD given that treatment mechanisms in the disorder are poorly understood, and there is little evidence to suggest that efficacy in existing treatments is specific to BPD rather than a general reduction in psychiatric distress [119]. This work may also have been restricted by a lack of statistical power in some analyses, particularly the smaller meta-analyses of treatment effect and the subgroup analyses [66]. In addition, the binary coding of interventions into human-supported and fully automated subgroups may have obscured some of the effect of human support. For example, interventions that only offered as-needed support were grouped with interventions that included regular support, although the former might not impact the treatment effect as much as the latter. All negative subgroup results should be interpreted as an absence of evidence for efficacy rather than evidence of inefficacy. The meta-analyses and subgroup analyses also depend on the quality, completeness, and accuracy of the data in the included studies. As discussed, many studies were at risk of bias and had considerable loss to follow-up. Finally, when intervention descriptions were ambiguous or absent and authors did not reply to our queries, PSD elements were coded conservatively. This may have lowered the frequency estimates of PSD elements and the accuracy of the subgroup analyses.

Conclusions

This review has found meta-analytic evidence that digital interventions, particularly cognitive interventions, are effective at reducing the BPD symptoms of paranoia and suicidal ideation. The most effective suicidal ideation interventions were those that used evidence-based therapies and certain features of PSD. Automated digital interventions are unlikely to replace face-to-face care as primary interventions for BPD because they do not facilitate sufficient opportunities to develop emotion regulation skills. However, digital interventions still represent an opportunity for highly accessible symptom-specific BPD treatment.

Acknowledgments

The authors would like to thank Edoardo Ostinelli and Kevin Matlock for their guidance as well as Shireen Rizvi, Virginie Salamin, Phil Batterham, Ana Fonseca, Jill Hooley, Shirley Wang, Federico Daray, Sukriti Drabu, and Jodi Frey for their willingness to share data and discuss their intervention studies.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

JABL, KEAS, and NMM received nonfinancial support from Big Health Ltd in the form of no-cost access to the digital sleep improvement program, Sleepio, for use in clinical research, outside the submitted work. KEAS is supported by the National

XSL•FO
RenderX

Institute for Health Research Oxford Biomedical Research Centre based at Oxford University Hospitals National Health Service Trust and the University of Oxford. NMM is a current employee of and holds shares, restricted share units, and/or share options in Compass Pathways plc; this work is not related to Compass Pathways plc.

Multimedia Appendix 1
Details of the search strategy, studies included, nonsignificant meta-analyses of treatment effect, and subgroup analysis of population type (community vs outpatient).
[DOCX File , 1453 KB - jmir_v26i1e54941_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 656 KB - jmir_v26i1e54941_app2.pdf]

References

1. Bohus M, Stoffers-Winterling J, Sharp C, Krause-Utz A, Schmahl C, Lieb K. Borderline personality disorder. Lancet 2021 Oct 23;398(10310):1528-1540. [doi: 10.1016/S0140-6736(21)00476-1] [Medline: 34688371]
2. Ellison WD, Rosenstein LK, Morgan TA, Zimmerman M. Community and clinical epidemiology of borderline personality disorder. Psychiatr Clin North Am 2018 Dec;41(4):561-573. [doi: 10.1016/j.psc.2018.07.008] [Medline: 30447724]
3. van Asselt AD, Dirksen CD, Arntz A, Severens JL. The cost of borderline personality disorder: societal cost of illness in BPD-patients. Eur Psychiatry 2007 Sep;22(6):354-361. [doi: 10.1016/j.eurpsy.2007.04.001] [Medline: 17544636]
4. Iliakis EA, Sonley AK, Ilagan GS, Choi-Kain LW. Treatment of borderline personality disorder: is supply adequate to meet public health needs? Psychiatr Serv 2019 Sep 01;70(9):772-781. [doi: 10.1176/appi.ps.201900073] [Medline: 31138059]
5. Lakhtakia T, Torous J. Current directions in digital interventions for mood and anxiety disorders. Curr Opin Psychiatry 2022 Mar 01;35(2):130-135. [doi: 10.1097/YCO.0000000000000772] [Medline: 34966117]
6. Barak A, Klein B, Proudfoot JG. Defining internet-supported therapeutic interventions. Ann Behav Med 2009 Aug;38(1):4-17. [doi: 10.1007/s12160-009-9130-7] [Medline: 19787305]
7. Sloan E, Hall K, Moulding R, Bryce S, Mildred H, Staiger PK. Emotion regulation as a transdiagnostic treatment construct across anxiety, depression, substance, eating and borderline personality disorders: a systematic review. Clin Psychol Rev 2017 Nov;57:141-163. [doi: 10.1016/j.cpr.2017.09.002] [Medline: 28941927]
8. Daros AR, Williams GE. A meta-analysis and systematic review of emotion-regulation strategies in borderline personality disorder. Harv Rev Psychiatry 2019;27(4):217-232. [doi: 10.1097/HRP.0000000000000212] [Medline: 31219881]
9. Neacsiu AD, Rizvi SL, Linehan MM. Dialectical behavior therapy skills use as a mediator and outcome of treatment for borderline personality disorder. Behav Res Ther 2010 Sep;48(9):832-839 [FREE Full text] [doi: 10.1016/j.brat.2010.05.017] [Medline: 20579633]
10. Axelrod SR, Perepletchikova F, Holtzman K, Sinha R. Emotion regulation and substance use frequency in women with substance dependence and borderline personality disorder receiving dialectical behavior therapy. Am J Drug Alcohol Abuse 2011 Jan 22;37(1):37-42 [FREE Full text] [doi: 10.3109/00952990.2010.535582] [Medline: 21091162]
11. Briones-Buixassa L, Alí Í, Schmidt C, Nicolaou S, Pascual JC, Soler J, et al. Predicting non-suicidal self-injury in young adults with and without borderline personality disorder: a multilevel approach combining ecological momentary assessment and self-report measures. Psychiatr Q 2021 Sep 21;92(3):1035-1054. [doi: 10.1007/s11126-020-09875-7] [Medline: 33475912]
12. Freeman D. Persecutory delusions: a cognitive perspective on understanding and treatment. Lancet Psychiatry 2016 Jul;3(7):685-692. [doi: 10.1016/S2215-0366(16)00066-3] [Medline: 27371990]
13. Stalker K, Ferguson I, Barclay A. ‘It is a horrible term for someone’: service user and provider perspectives on ‘personality disorder’. Disabil Soc 2005 Jun;20(4):359-373. [doi: 10.1080/09687590500086443]
14. Cuijpers P, Donker T, van Straten A, Li J, Andersson G. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. Psychol Med 2010 Dec;40(12):1943-1957. [doi: 10.1017/S0033291710000772] [Medline: 20406528]
15. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
16. Larsen ME, Huckvale K, Nicholas J, Torous J, Birrell L, Li E, et al. Using science to sell apps: evaluation of mental health app store quality claims. NPJ Digit Med 2019 Mar 22;2(1):18 [FREE Full text] [doi: 10.1038/s41746-019-0093-1] [Medline: 31304366]
17. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
18. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009;24(1):28. [doi: 10.17705/1cais.02428]

XSL•FO
RenderX

19. Wildeboer G, Kelders SM, van Gemert-Pijnen JE. The relationship between persuasive technology principles, adherence and effect of web-based interventions for mental health: a meta-analysis. Int J Med Inform 2016 Dec;96:71-85. [doi: 10.1016/j.ijmedinf.2016.04.005] [Medline: 27117057]
20. Rizvi SL, Hughes CD, Thomas MC. The DBT Coach mobile application as an adjunct to treatment for suicidal and self-injuring individuals with borderline personality disorder: a preliminary evaluation and challenges to client utilization. Psychol Serv 2016 Nov;13(4):380-388. [doi: 10.1037/ser0000100] [Medline: 27797571]
21. Torok M, Han J, McGillivray L, Wong Q, Werner-Seidler A, O'Dea B, et al. The effect of a therapeutic smartphone application on suicidal ideation in young adults: findings from a randomized controlled trial in Australia. PLoS Med 2022 May 31;19(5):e1003978 [FREE Full text] [doi: 10.1371/journal.pmed.1003978] [Medline: 35639672]
22. Frey JJ, Osteen PJ, Sharpe TL, Mosby AO, Joiner T, Ahmedani B, et al. Effectiveness of man therapy to reduce suicidal ideation and depression among working-age men: A randomized controlled trial. Suicide Life Threat Behav 2023 Feb;53(1):137-153 [FREE Full text] [doi: 10.1111/sltb.12932] [Medline: 36373571]
23. Fonseca A, Monteiro F, Alves S, Gorayeb R, Canavarro MC. Be a mom, a web-based intervention to prevent postpartum depression: the enhancement of self-regulatory skills and its association with postpartum depressive symptoms. Front Psychol 2019 Feb 18;10:265 [FREE Full text] [doi: 10.3389/fpsyg.2019.00265] [Medline: 30873060]
24. Helweg-Joergensen S, Schmidt T, Lichtenstein MB, Pedersen SS. Using a mobile diary app in the treatment of borderline personality disorder: mixed methods feasibility study. JMIR Form Res 2019 Sep 30;3(3):e12852 [FREE Full text] [doi: 10.2196/12852] [Medline: 31573910]
25. Crosby ES, Witte TK. A pilot study of sleep scholar: a single-session, internet-based insomnia intervention for college students with a history of suicide ideation. J Am Coll Health 2023 Oct 20;71(7):1984-1998. [doi: 10.1080/07448481.2021.1953028] [Medline: 34283711]
26. Johnson SL, Zisser MR, Sandel DB, Swerdlow BA, Carver CS, Sanchez AH, et al. Development of a brief online intervention to address aggression in the context of emotion-related impulsivity: evidence from a wait-list controlled trial. Behav Res Ther 2020 Nov;134:103708. [doi: 10.1016/j.brat.2020.103708] [Medline: 32896743]
27. Franklin JC, Fox KR, Franklin CR, Kleiman EM, Ribeiro JD, Jaroszewski AC, et al. A brief mobile app reduces nonsuicidal and suicidal self-injury: evidence from three randomized controlled trials. J Consult Clin Psychol 2016 Jun;84(6):544-557. [doi: 10.1037/ccp0000093] [Medline: 27018530]
28. Jacob GA, Hauer A, Köhne S, Assmann N, Schaich A, Schweiger U, et al. A schema therapy-based eHealth program for patients with borderline personality disorder (priovi): naturalistic single-arm observational study. JMIR Ment Health 2018 Dec 17;5(4):e10983 [FREE Full text] [doi: 10.2196/10983] [Medline: 30559092]
29. Bernstein EE, Bentley KH, Nock MK, Stein MB, Beck S, Kleiman EM. An ecological momentary intervention study of emotional responses to smartphone-prompted CBT skills practice and the relationship to clinical outcomes. Behav Ther 2022 Mar;53(2):267-280 [FREE Full text] [doi: 10.1016/j.beth.2021.09.001] [Medline: 35227403]
30. Batterham PJ, Calear AL, Farrer L, McCallum SM, Cheng VW. FitMindKit: randomised controlled trial of an automatically tailored online program for mood, anxiety, substance use and suicidality. Internet Interv 2018 Jun;12:91-99 [FREE Full text] [doi: 10.1016/j.invent.2017.08.002] [Medline: 30135773]
31. Frías Á, Solves L, Navarro S, Palma C, Farriols N, Aliaga F, et al. Technology-based psychosocial interventions for people with borderline personality disorder: a scoping review of the literature. Psychopathology 2020 Nov 9;53(5-6):254-263 [FREE Full text] [doi: 10.1159/000511349] [Medline: 33166964]
32. Ilagan GS, Iliakis EA, Wilks CR, Vahia IV, Choi-Kain LW. Smartphone applications targeting borderline personality disorder symptoms: a systematic review and meta-analysis. Borderline Personal Disord Emot Dysregul 2020 Jun 16;7(1):12 [FREE Full text] [doi: 10.1186/s40479-020-00127-5] [Medline: 32549987]
33. van der Boom B, Boumparis N, Donker T, de Beurs D, Arntz A, Riper H. Internet-delivered interventions for personality disorders - a scoping review. Internet Interv 2022 Apr;28:100525 [FREE Full text] [doi: 10.1016/j.invent.2022.100525] [Medline: 35450140]
34. Michaels TI, Singal S, Stone E, Braider L, Kane J. Digital tools for the treatment of borderline personality disorder: a critical review of smartphone applications. In: Proceedings of the ACM IUI 2021 Workshops Co-located with 26th ACM Conference on Intelligent User Interfaces. 2021 Presented at: ACMIUI-WS '21; April 13-17, 2021; College Station, TX p. 1-3 URL: https://ceur-ws.org/Vol-2903/IUI21WS-HEALTHI-8.pdf
35. Embase search strategy. Ovid. 2022. URL: https://ovidsp.ovid.com/ovidweb.cgi?T=JS&NEWS=N&PAGE=main&SHAREDSEARCHID=1YSooNQvXUDGmVHOvJjpB1RZ8MQArFNsok6rDtXiw2FS4Tb5soRNwlzkb0LoZKdGB [accessed 2024-04-29]
36. Ayiku L, Hudson T, Glover S, Walsh N, Adams R, Deane J, et al. The NICE MEDLINE and Embase (Ovid) health apps search filters: development of validated filters to retrieve evidence about health apps. Int J Technol Assess Health Care 2020 Oct 27;37(1):e16. [doi: 10.1017/S026646232000080X] [Medline: 33107420]
37. Faster systematic review. Rayyan. URL: https://www.rayyan.ai/ [accessed 2024-04-29]
38. Evidence standards framework (ESF) for digital health technologies. National Institute for Health and Care Excellence. 2019. URL: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [accessed 2024-04-29]

XSL•FO
RenderX

39. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
40. Study quality assessment tools. National Heart, Lung and Blood Institute. URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2024-04-29]
41. Hedges LV. Distribution theory for glass's estimator of effect size and related estimators. J Educ Stat 2016 Nov 23;6(2):107-128. [doi: 10.3102/10769986006002107]
42. Hasking P, Chen NT, Chiu V, Gray N, Gross JJ, Boyes M. "Managing emotion": open label trial and waitlist controlled trial of an emotion regulation program for university students. J Am Coll Health (Forthcoming) 2023 Jan 26:1-11. [doi: 10.1080/07448481.2022.2155468] [Medline: 36701430]
43. Drabu S, Sündermann O, Hong RY. A one-week online self-compassion training reduces self-criticism and pain endurance in adults with non-suicidal self-injury ideation: A randomized-waitlist controlled study. Mindfulness 2022 Apr 06;13(5):1232-1245. [doi: 10.1007/S12671-022-01870-2]
44. Zanarini MC, Vujanovic AA, Parachini EA, Boulanger JL, Frankenburg FR, Hennen J. Zanarini rating scale for borderline personality disorder (ZAN-BPD): a continuous measure of DSM-IV borderline psychopathology. J Pers Disord 2003 Jun;17(3):233-242. [doi: 10.1521/pedi.17.3.233.22147] [Medline: 12839102]
45. Arntz A, van den Hoorn M, Cornelis J, Verheul R, van den Bosch WM, de Bie AJ. Reliability and validity of the borderline personality disorder severity index. J Pers Disord 2003 Feb;17(1):45-59. [doi: 10.1521/pedi.17.1.45.24053] [Medline: 12659546]
46. Beck AT, Kovacs M, Weissman A. Assessment of suicidal intention: the Scale for Suicide Ideation. J Consult Clin Psychol 1979 Apr;47(2):343-352. [doi: 10.1037//0022-006x.47.2.343] [Medline: 469082]
47. van Spijker BA, Batterham PJ, Calear AL, Farrer L, Christensen H, Reynolds J, et al. The suicidal ideation attributes scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. Suicide Life Threat Behav 2014 Aug 24;44(4):408-419. [doi: 10.1111/sltb.12084] [Medline: 24612048]
48. Nock MK, Holmberg EB, Photos VI, Michel BD. Self-injurious thoughts and behaviors interview: development, reliability, and validity in an adolescent sample. Psychol Assess 2007 Sep;19(3):309-317. [doi: 10.1037/1040-3590.19.3.309] [Medline: 17845122]
49. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011 Dec;168(12):1266-1277 [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
50. Osman A, Bagge CL, Gutierrez PM, Konick LC, Kopper BA, Barrios FX. The Suicidal Behaviors Questionnaire-Revised (SBQ-R): validation with clinical and nonclinical samples. Assessment 2001 Dec;8(4):443-454. [doi: 10.1177/107319110100800409] [Medline: 11785588]
51. Conrad AK, Jacoby AM, Jobes DA, Lineberry TW, Shea CE, Arnold Ewing TD, et al. A psychometric investigation of the suicide status form II with a psychiatric inpatient sample. Suicide Life Threat Behav 2009 Jun 31;39(3):307-320. [doi: 10.1521/suli.2009.39.3.307] [Medline: 19606922]
52. Fenigstein A, Vanable PA. Paranoia and self-consciousness. J Pers Soc Psychol 1992 Jan;62(1):129-138. [doi: 10.1037//0022-3514.62.1.129] [Medline: 1538311]
53. Schlier B, Moritz S, Lincoln TM. Measuring fluctuations in paranoia: validity and psychometric properties of brief state versions of the Paranoia Checklist. Psychiatry Res 2016 Jul 30;241:323-332. [doi: 10.1016/j.psychres.2016.05.002] [Medline: 27227702]
54. Sorgi P, Ratey JJ, Knoedler DW, Markert RJ, Reichman M. Rating aggression in the clinical setting. A retrospective adaptation of the Overt Aggression Scale: preliminary results. J Neuropsychiatry Clin Neurosci 1991;3(2):S52-S56. [Medline: 1687961]
55. Spielberger CD. State-trait anger expression inventory. In: Weiner IB, Craighead WE, editors. The Corsini Encyclopedia of Psychology. Totowa, NJ: Wiley-Blackwell; 2010.
56. Kaufman EA, Xia M, Fosco G, Yaptangco M, Skidmore CR, Crowell SE. The Difficulties in Emotion Regulation Scale Short Form (DERS-SF): validation and replication in adolescent and adult samples. J Psychopathol Behav Assess 2015 Nov 23;38(3):443-455. [doi: 10.1007/S10862-015-9529-3]
57. Berking M, Znoj H. Entwicklung und Validierung eines Fragebogens zur standardisierten Selbsteinschätzung emotionaler Kompetenzen (SEK-27). Z fur Psychiatr Psychol Psychother 2008 Jan;56(2):141-153. [doi: 10.1024/1661-4747.56.2.141]
58. Higgins JP, Eldridge S, Li T. Including variants on randomized trials. In: Higgins PT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. Burlington, MA: Wiley-Blackwell; 2019:56942-56993.
59. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta - Analysis. Hoboken, NJ: John Wiley & Sons; 2021.
60. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods 1998;3(4):486-504. [doi: 10.1037//1082-989x.3.4.486]

XSL•FO
RenderX

61. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods 2016 Mar 02;7(1):55-79 [FREE Full text] [doi: 10.1002/jrsm.1164] [Medline: 26332144]
62. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
63. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
64. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R. Cham, Switzerland: Springer; 2015.
65. Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care program. J Clin Epidemiol 2011 Nov 1;64(11):1187-1197 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.08.010] [Medline: 21477993]
66. Harrer M, Cuijpers P, Furukawa T, Ebert D. Doing Meta-Analysis with R: A Hands-On Guide. New York, NY: Taylor & Francis Group; 2012.
67. Deeks JJ, Higgins JP, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins J, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2022.
68. West SL, Gartlehner G, Mansfield AJ, Poole C, Tant E, Lenfestey N, et al. Table 7, summary of common statistical approaches to test for heterogeneity. Agency for Healthcare Research and Quality. 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK53317/table/ch3.t2/ [accessed 2024-04-29]
69. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57(1):300. [doi: 10.1111/J.2517-6161.1995.tb02031.X]
70. Del Re AC, Hoyt WT. MAd: meta-analysis with mean differences. Cran R. 2014. URL: https://cran.r-project.org/web/packages/MAd/index.html [accessed 2024-10-18]
71. Eylem O, van Straten A, de Wit L, Rathod S, Bhui K, Kerkhof AJ. Reducing suicidal ideation among Turkish migrants in the Netherlands and in the UK: the feasibility of a randomised controlled trial of a guided online intervention. Pilot Feasibility Stud 2021 Jan 25;7(1):30 [FREE Full text] [doi: 10.1186/s40814-021-00772-9] [Medline: 33494831]
72. De Jaegere E, van Landschoot R, van Heeringen K, van Spijker BA, Kerkhof AJ, Mokkenstorm JK, et al. The online treatment of suicidal ideation: a randomised controlled trial of an unguided web-based intervention. Behav Res Ther 2019 Aug;119:103406. [doi: 10.1016/j.brat.2019.05.003] [Medline: 31176889]
73. Mühlmann C, Madsen T, Hjorthøj C, Forman JL, Kerkhof AJ, Nordentoft M, et al. Effectiveness of an internet-based self-help therapy program for suicidal ideation with follow-up at 6 months: results of a randomized controlled trial. J Clin Psychiatry 2021 Aug 31;82(5):1-11. [doi: 10.4088/JCP.20m13803] [Medline: 34464522]
74. van Spijker BA, Werner-Seidler A, Batterham PJ, Mackinnon A, Calear AL, Gosling JA, et al. Effectiveness of a web-based self-help program for suicidal thinking in an Australian community sample: randomized controlled trial. J Med Internet Res 2018 Feb 14;20(2):e15 [FREE Full text] [doi: 10.2196/jmir.8595] [Medline: 29444769]
75. van Spijker BA, van Straten A, Kerkhof AJ. Effectiveness of online self-help for suicidal thoughts: results of a randomised controlled trial. PLoS One 2014 Feb 27;9(2):e90118 [FREE Full text] [doi: 10.1371/journal.pone.0090118] [Medline: 24587233]
76. Hooley JM, Fox KR, Wang SB, Kwashie AN. Novel online daily diary interventions for nonsuicidal self-injury: a randomized controlled trial. BMC Psychiatry 2018 Aug 22;18(1):264 [FREE Full text] [doi: 10.1186/s12888-018-1840-6] [Medline: 30134866]
77. Salamin V, Ray P, Gothuey I, Corzani S, Martin-Soelch C. An internet-based intervention for the relatives of people with mental illnesses. Swiss J Psychol 2019 Apr;78(1-2):15-27. [doi: 10.1024/1421-0185/a000219]
78. Zanarini MC, Conkey LC, Temes CM, Fitzmaurice GM. Randomized controlled trial of web-based psychoeducation for women with borderline personality disorder. J Clin Psychiatry 2018 Jun 27;79(3):16065. [doi: 10.4088/jcp.16m11153]
79. Klein JP, Hauer-von Mauschwitz A, Berger T, Fassbinder E, Mayer J, Borgwardt S, et al. Effectiveness and safety of the adjunctive use of an internet-based self-management intervention for borderline personality disorder in addition to care as usual: results from a randomised controlled trial. BMJ Open 2021 Sep 08;11(9):e047771 [FREE Full text] [doi: 10.1136/bmjopen-2020-047771] [Medline: 34497078]
80. Laursen SL, Helweg-Jørgensen S, Langergaard A, Søndergaard J, Sørensen SS, Mathiasen K, et al. Mobile diary app versus paper-based diary cards for patients with borderline personality disorder: economic evaluation. J Med Internet Res 2021 Nov 11;23(11):e28874 [FREE Full text] [doi: 10.2196/28874] [Medline: 34762057]
81. Tighe J, Shand F, Ridani R, Mackinnon A, De La Mata N, Christensen H. Ibobbly mobile health intervention for suicide prevention in Australian indigenous youth: a pilot randomised controlled trial. BMJ Open 2017 Jan 27;7(1):e013518 [FREE Full text] [doi: 10.1136/bmjopen-2016-013518] [Medline: 28132007]
82. Wilks CR, Lungu A, Ang SY, Matsumiya B, Yin Q, Linehan MM. A randomized controlled trial of an internet delivered dialectical behavior therapy skills training for suicidal and heavy episodic drinkers. J Affect Disord 2018 May;232:219-228 [FREE Full text] [doi: 10.1016/j.jad.2018.02.053] [Medline: 29499504]

XSL•FO
RenderX

83. Depp CA, Parrish EM, Chalker SA, Ehret BC, Kamarsu S, Perivoliotis D, et al. Pilot feasibility trial of a brief mobile-augmented suicide prevention intervention for serious mental illness. Psychiatr Rehabil J 2023 Mar;46(1):74-82. [doi: 10.1037/prj0000547] [Medline: 36809018]
84. O'Toole MS, Arendt MB, Pedersen CM. Testing an app-assisted treatment for suicide prevention in a randomized controlled trial: effects on suicide risk and depression. Behav Ther 2019 Mar;50(2):421-429. [doi: 10.1016/j.beth.2018.07.007] [Medline: 30824256]
85. Pauwels K, Aerts S, Muijzers E, De Jaegere E, van Heeringen K, Portzky G. BackUp: development and evaluation of a smart-phone application for coping with suicidal crises. PLoS One 2017 Jun 21;12(6):e0178144 [FREE Full text] [doi: 10.1371/journal.pone.0178144] [Medline: 28636617]
86. Rodante DE, Kaplan MI, Olivera Fedi R, Gagliesi P, Pascali A, José Quintero PS, et al. CALMA, a mobile health application, as an accessory to therapy for reduction of suicidal and non-suicidal self-injured behaviors: a pilot cluster randomized controlled trial. Arch Suicide Res 2022 Oct 18;26(2):801-818. [doi: 10.1080/13811118.2020.1834476] [Medline: 33073745]
87. Muneghina O, Van Gordon W, Barrows P, Richardson M. A novel mindful nature connectedness intervention improves paranoia but not anxiety in a nonclinical population. Ecopsychology 2021 Dec 01;13(4):248-256. [doi: 10.1089/eco.2020.0068]
88. Sood M, Newman-Taylor K. Cognitive fusion mediates the impact of attachment imagery on paranoia and anxiety. Cogn Ther Res 2020 Jul 02;44(6):1150-1161. [doi: 10.1007/S10608-020-10127-Y]
89. Newman-Taylor K, Sood M, Rowe AC, Carnelley KB. The impact of repeated attachment priming on paranoia, mood and help-seeking intentions in an analogue sample. Brain Sci 2021 Sep 22;11(10):1257 [FREE Full text] [doi: 10.3390/brainsci11101257] [Medline: 34679322]
90. Shore R, Strauss C, Cavanagh K, Hayward M, Ellett L. A randomised controlled trial of a brief online mindfulness-based intervention on paranoia in a non-clinical sample. Mindfulness (N Y) 2018 Jul 14;9(1):294-302 [FREE Full text] [doi: 10.1007/s12671-017-0774-2] [Medline: 29387267]
91. Newman-Taylor K, Kemp A, Potter H, Au-Yeung SK. An online investigation of imagery to attenuate paranoia in college students. J Child Fam Stud 2017 Nov 20;27(3):853-859. [doi: 10.1007/s10826-017-0934-y]
92. Flujas-Contreras JM, García-Palacios A, Gómez I. Effectiveness of a web-based intervention on parental psychological flexibility and emotion regulation: a pilot open trial. Int J Environ Res Public Health 2021 Mar 13;18(6):2958 [FREE Full text] [doi: 10.3390/ijerph18062958] [Medline: 33805835]
93. Stappenbeck CA, Gulati NK, Jaffe AE, Blayney JA, Kaysen D. Initial efficacy of a web-based alcohol and emotion regulation intervention for college women with sexual assault histories. Psychol Addict Behav 2021 Nov;35(7):852-865 [FREE Full text] [doi: 10.1037/adb0000762] [Medline: 34291957]
94. Glück TM, Maercker A. A randomized controlled pilot study of a brief web-based mindfulness training. BMC Psychiatry 2011 Nov 08;11(1):175 [FREE Full text] [doi: 10.1186/1471-244X-11-175] [Medline: 22067058]
95. Howie AJ, Malouff JM. Effects of online cognitive treatment for problematic anger: a randomized controlled trial. Cogn Behav Ther 2014 Jul 28;43(4):310-318. [doi: 10.1080/16506073.2014.939103] [Medline: 25065452]
96. Osgood JM, Kase SE, Zaroukian EG, Quartana PJ. Online intervention reduces hostile attribution bias, anger, aggressive driving, and cyber-aggression, results of two randomized trials. Cogn Ther Res 2020 Sep 11;45(2):310-321. [doi: 10.1007/S10608-020-10147-8]
97. Jobes DA. The collaborative assessment and management of suicidality (CAMS): an evolving evidence‐based clinical approach to suicidal risk. Suicide Life Threat Behav 2012 Sep 12;42(6):640-653. [doi: 10.1111/j.1943-278x.2012.00119.x]
98. Zalewski M, Walton CJ, Rizvi SL, White AW, Gamache Martin C, O'Brien JR, et al. Lessons learned conducting dialectical behavior therapy via telehealth in the age of COVID-19. Cogn Behav Pract 2021 Nov;28(4):573-587 [FREE Full text] [doi: 10.1016/j.cbpra.2021.02.005] [Medline: 34629837]
99. Stanley B, Brown G, Brent DA, Wells K, Poling K, Curry J, et al. Cognitive-behavioral therapy for suicide prevention (CBT-SP): treatment model, feasibility, and acceptability. J Am Acad Child Adolesc Psychiatry 2009 Oct;48(10):1005-1013. [doi: 10.1097/chi.0b013e3181b5dbfe]
100. Witte TK, Fitzpatrick KK, Warren KL, Schatschneider C, Schmidt NB. Naturalistic evaluation of suicidal ideation: variability and relation to attempt status. Behav Res Ther 2006 Jul 1;44(7):1029-1040 [FREE Full text] [doi: 10.1016/j.brat.2005.08.004] [Medline: 16188225]
101. Loranger AW, Janca A, Sartorius N. Assessment and Diagnosis of Personality Disorders: The ICD-10 International Personality Disorder Examination (IPDE). Cambridge, MA: Cambridge University Press; 1997.
102. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016 Jan 08;18(1):e6 [FREE Full text] [doi: 10.2196/jmir.4790] [Medline: 26747176]
103. Petrie H, Bevan N. The evaluation of accessibility, usability, and user experience. In: Stepanidis C, editor. The Universal Access Handbook. Boca Raton, FL: CRC Press; 2009:1-16.
104. D'Anci KE, Uhl S, Giradi G, Martin C. Treatments for the prevention and management of suicide: a systematic review. Ann Intern Med 2019 Sep 03;171(5):334-342 [FREE Full text] [doi: 10.7326/M19-0869] [Medline: 31450239]

XSL•FO
RenderX

105. Jacobson NS, Gortner ET. Can depression be de-medicalized in the 21st century: scientific revolutions, counter-revolutions and the magnetic field of normal science. Behav Res Ther 2000 Feb 1;38(2):103-117 [FREE Full text] [doi: 10.1016/s0005-7967(99)00029-7] [Medline: 10660997]
106. Brown P, Waite F, Rovira A, Nickless A, Freeman D. Virtual reality clinical-experimental tests of compassion treatment techniques to reduce paranoia. Sci Rep 2020 May 22;10(1):8547 [FREE Full text] [doi: 10.1038/s41598-020-64957-7] [Medline: 32444619]
107. Freeman D, Lister R, Waite F, Galal U, Yu L, Lambe S, et al. Automated virtual reality cognitive therapy versus virtual reality mental relaxation therapy for the treatment of persistent persecutory delusions in patients with psychosis (THRIVE): a parallel-group, single-blind, randomised controlled trial in England with mediation analyses. Lancet Psychiatry 2023 Nov;10(11):836-847 [FREE Full text] [doi: 10.1016/S2215-0366(23)00257-2] [Medline: 37742702]
108. Contreras A, Valiente C, Vázquez C, Trucharte A, Peinado V, Varese F, et al. The network structure of paranoia dimensions and its mental health correlates in the general population: the core role of loneliness. Schizophr Res 2022 Aug;246:65-73 [FREE Full text] [doi: 10.1016/j.schres.2022.06.005] [Medline: 35717741]
109. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023 Jan 28;53(2):567-579 [FREE Full text] [doi: 10.1017/S0033291721001999] [Medline: 34047264]
110. Borderline personality disorder: treatment and management. National Collaborating Centre for Mental Health (UK). 2019. URL: https://www.ncbi.nlm.nih.gov/books/NBK55403/ [accessed 2024-04-29]
111. Shen Y, Chen D, Ying J, Wang C, You J. The association between negative automatic thoughts and suicidal ideation: a longitudinal serial mediation model. Curr Psychol 2023 Aug 26;43(11):9909-9919. [doi: 10.1007/S12144-023-05133-8]
112. Coleman D, Casey JT. Therapeutic mechanisms of suicidal ideation: the influence of changes in automatic thoughts and immature defenses. Crisis 2007 Jul;28(4):198-203. [doi: 10.1027/0227-5910.28.4.198] [Medline: 18265740]
113. Goodman M, New A. Impulsive aggression in borderline personality disorder. Curr Psychiatry Rep 2000 Feb;2(1):56-61. [doi: 10.1007/s11920-000-0043-1] [Medline: 11122933]
114. Hawton K, Pirkis J. Suicide prevention: reflections on progress over the past decade. Lancet Psychiatry 2024 Jun;11(6):472-480. [doi: 10.1016/S2215-0366(24)00105-6] [Medline: 38754457]
115. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 06;25:e42864 [FREE Full text] [doi: 10.2196/42864] [Medline: 36745497]
116. Ebert DD, Van Daele T, Nordgreen T, Karekla M, Compare A, Zarbo C, et al. Internet- and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health Efficacy, and Potential for Improving Mental Health. Eur Psychol 2018 Apr;23(2):167-187. [doi: 10.1027/1016-9040/a000318]
117. Bach B, Sellbom M. Continuity between DSM-5 categorical criteria and traits criteria for borderline personality disorder. Can J Psychiatry 2016 Aug 10;61(8):489-494 [FREE Full text] [doi: 10.1177/0706743716640756] [Medline: 27310232]
118. Lindhiem O, Bennett CB, Orimoto TE, Kolko DJ. A meta-analysis of personalized treatment goals in psychotherapy: a preliminary report and call for more studies. Clin Psychol (New York) 2016 Jun;23(2):165-176 [FREE Full text] [doi: 10.1111/cpsp.12153] [Medline: 27325908]
119. Bateman AW, Gunderson J, Mulder R. Treatment of personality disorder. Lancet 2015 Feb 21;385(9969):735-743. [doi: 10.1016/S0140-6736(14)61394-5] [Medline: 25706219]

Abbreviations

BPD: borderline personality disorder
CBT: cognitive behavioral therapy
DBT: dialectical behavior therapy
EBT: evidence-based treatment
NICE: National Institute for Health and Care Excellence
NSSI: nonsuicidal self-injury
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSD: persuasive system design
RCT: randomized controlled trial
SMD: standardized mean difference

XSL•FO
RenderX

Edited by A Coristine; submitted 28.11.23; peer-reviewed by S Mangelsdorf, W Deng; comments to author 07.06.24; revised version received 27.06.24; accepted 26.08.24; published 29.11.24.

Please cite as:
Lindsay JAB, McGowan NM, Henning T, Harriss E, Saunders KEA
Digital Interventions for Symptoms of Borderline Personality Disorder: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e54941
URL: https://www.jmir.org/2024/1/e54941
doi:10.2196/54941
PMID:

©Julia A B Lindsay, Niall M McGowan, Thomas Henning, Eli Harriss, Kate E A Saunders. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Virtual Primary Care for People With Opioid Use Disorder: Scoping Review of Current Strategies, Benefits, and Challenges

Shawna Narayan[1], MSc; Ellie Gooderham[2], MA; Sarah Spencer[2], MSc, MPH; Rita K McCracken[1,3], MD, PhD; Lindsay Hedden[2], PhD

[1]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[2]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

[3]Department of Family Medicine, Providence Health Care, Vancouver, BC, Canada

Corresponding Author:
Lindsay Hedden, PhD
Faculty of Health Sciences
Simon Fraser University
8888 University Drive
Burnaby, BC, V5A 1S6
Canada
Phone: 1 778 782 7559
Email: lindsay_hedden@sfu.ca

Abstract

Background: There is a pressing need to understand the implications of the rapid adoption of virtual primary care for people with opioid use disorder. Potential impacts, including disruptions to opiate agonist therapies, and the prospect of improved service accessibility remain underexplored.

Objective: This scoping review synthesized current literature on virtual primary care for people with opioid use disorder with a specific focus on benefits, challenges, and strategies.

Methods: We followed the Joanna Briggs Institute methodological approach for scoping reviews and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for reporting our findings. We conducted searches in MEDLINE, Web of Science, CINAHL Complete, and Embase using our developed search strategy with no date restrictions. We incorporated all study types that included the 3 concepts (ie, virtual care, primary care, and people with opioid use disorder). We excluded research on minors, asynchronous virtual modalities, and care not provided in a primary care setting. We used Covidence to screen and extract data, pulling information on study characteristics, health system features, patient outcomes, and challenges and benefits of virtual primary care. We conducted inductive content analysis and calculated descriptive statistics. We appraised the quality of the studies using the Quality Assessment With Diverse Studies tool and categorized the findings using the Consolidated Framework for Implementation Research.

Results: Our search identified 1474 studies. We removed 36.36% (536/1474) of these as duplicates, leaving 938 studies for title and abstract screening. After a double review process, we retained 3% (28/938) of the studies for extraction. Only 14% (4/28) of the studies were conducted before the COVID-19 pandemic, and most (15/28, 54%) used quantitative methodologies. We summarized objectives and results, finding that most studies (18/28, 64%) described virtual primary care delivered via phone rather than video and that many studies (16/28, 57%) reported changes in appointment modality. Through content analysis, we identified that policies and regulations could either facilitate (11/28, 39%) or impede (7/28, 25%) the provision of care virtually. In addition, clinicians' perceptions of patient stability (5/28, 18%) and the heightened risks associated with virtual care (10/28, 36%) can serve as a barrier to offering virtual services. For people with opioid use disorder, increased health care accessibility was a noteworthy benefit (13/28, 46%) to the adoption of virtual visits, whereas issues regarding access to technology and digital literacy stood out as the most prominent challenge (12/28, 43%).

Conclusions: The available studies highlight the potential for enhancing accessibility and continuous access to care for people with opioid use disorder using virtual modalities. Future research and policies must focus on bridging gaps to ensure that virtual primary care does not exacerbate or entrench health inequities.

(J Med Internet Res 2024;26:e54015) doi:10.2196/54015

XSL•FO
RenderX

KEYWORDS

telehealth; virtual care; primary care; opioid use disorder; opioid agonist therapy; COVID-19; mobile phone

Introduction

Background

The opioid crisis remains a significant public health challenge [1], highlighted by the rising number of deaths worldwide [2-4]. In North America, illicit drug poisoning deaths can be attributed to a contaminated drug supply saturated with synthetic opioids, particularly fentanyl [5]. Simultaneously, opioid use disorder (OUD) is on the rise [6], with 21.4 million cases worldwide in 2019 [7]. Effective medications for OUD, such as opioid agonist therapies (OATs), reduce illicit drug poisoning deaths [8,9]. Given that primary care services are often at the forefront of diagnosing and treating OUD, evidence shows that accessible primary care services can help achieve better health outcomes for people with OUD [10,11].

The COVID-19 pandemic introduced new barriers to traditional approaches to OUD care, including in-person primary care services, due to the need for social distancing and reduced in-person interactions [12,13]. In response to COVID-19 public health measures, health care systems implemented remote care solutions, transforming the delivery of health care services [14] and accelerating the adoption and use of virtual health care [15-17]. During this same period, much of the progress that had been made to address the opioid crisis was reversed [18-20]; illicit drug poisoning deaths increased as a result of limited health services [21], poisoned supply [22], and solitary drug use [23]. In the context of these challenges, and as a result of pandemic-related public health measures, virtual care emerged as a potential solution to address the unique needs of people with OUD by providing essential health care services remotely [24].

The widespread shift to virtual care facilitated continuous access to care during the pandemic. This shift may have been helpful for individuals with OUD [25,26], many of whom experience barriers to health care due to intersecting effects of criminalization, discrimination, and social marginalization [13]. Virtual care allows clinicians to deliver essential health care services to people with OUD [27,28], offering improved accessibility [29] and the ability to deliver OAT [30] while reducing barriers related to transportation, mobility limitations, and stigma [31]. However, virtual visits may limit clinicians' ability to fully assess patients' physical and emotional well-being [27] and may introduce technological barriers for those with limited internet access and low digital literacy [32,33]. In addition, virtual care may not be suitable for all aspects of OUD care as some interventions, such as injectable OAT [34], require regular in-person management [27]. The integration of virtual care into the management of OUD within the primary care context may improve health care delivery, but its limitations, which make in-person visits necessary, must be acknowledged to maximize its efficacy.

Objectives

Virtual platforms have become increasingly common for primary care services—a trend that is expected to persist beyond the pandemic [35]. Therefore, it is essential to understand the benefits and limitations of current virtual care interventions to shape future health care delivery models and optimize the care provided to patients with social and clinical complexities. This scoping review aimed to synthesize the current knowledge landscape and research gaps and offer insights into the potential of virtual primary care for people with OUD.

Methods

Overall Study Design

This review informs an ongoing mixed methods study investigating changes to primary care access and patient outcomes following the rapid introduction of virtual care for people with OUD in British Columbia [36]. We conducted the review in accordance with the Joanna Briggs Institute Reviewer's Manual [37,38] and the framework suggested by Arksey and O'Malley [39]. We registered our protocol on the Open Science Framework [40] and report our findings using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [41] and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for Abstracts checklists [42] (Multimedia Appendix 1).

Eligibility Criteria

We established our eligibility criteria a priori and present these in Textbox 1. We included studies that focused on the concepts of "virtual care," "primary care," and "people with OUD" and excluded studies that did not include all 3 concepts. We defined "primary care" as "the provision of integrated, accessible health care services by physicians and their health care teams who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients, and practicing in the context of family and community" [43]. Therefore, primary care settings are those where patients could receive comprehensive care (ie, treatment for any type of health issue) [44], including physician's offices, community health centers, and specific outpatient clinics. We included some settings that may not traditionally be considered primary care—such as certain opioid treatment programs, syringe service programs, or veterans' health association programs and clinics—when there was evidence that the setting also attended to health care issues beyond OUD [45-47].

XSL•FO
RenderX

Textbox 1. Eligibility criteria.

Inclusion criteria

- Virtual care modality: synchronous visits, telephone calls, and video calls
- Health service setting: primary care, comprehensive care in physicians' offices, community health centers, and some outpatient clinics
- Population: people with opioid use disorder with or without comorbidities
- Study type: abstracts, viewpoints, protocols, observational studies, qualitative data, quantitative data, and systematic reviews
- Language: English
- Geographic region: all

Exclusion criteria

- Virtual care modality: asynchronous visits, text-message or email-based visits, mobile apps, or unspecified
- Health service setting: psychiatric clinics, substance use disorder clinics, noncomprehensive care provided through opioid treatment programs or syringe services, in-hospital care, prisons, or unspecified
- Population: people aged <18 years and studies in which data specific to opioid use disorder were not distinguished from data for other substance use disorders
- Study type: none
- Language: all other (non-English) languages
- Geographic region: none

Search Strategy

We developed search strategies [48-52] by combining Medical Subject Heading terms (eg, *analgesics*, *opioids*, *telemedicine*, and *primary health care*) and free-text keywords (eg, *methadone*, *virtual*, and *family doctor*) using Boolean (eg, AND, OR, and NOT), truncation, and wildcard operators. Librarians from the University of British Columbia and Simon Fraser University verified the search strategies. Our final search strategy is included in Multimedia Appendix 2. From December 6, 2022, to December 8, 2022, one reviewer (SN) systematically searched MEDLINE (Ovid) [49], CINAHL Complete (EBSCOhost) [51], Embase (Ovid) [50], and the Web of Science Core Collection [48] with no end date. From May 2023 to June 2023, we searched gray literature using the Canada Commons database [52], Google Scholar, Trip, Grey Matters, and select organization websites (ie, Canada Health Infoway, British Columbia Centre on Substance Use, Centre for Addiction and Mental Health, and Canadian Research Initiative in Substance Misuse). Following the search and extraction, the bibliographies of the included studies were systematically checked for additional references [53].

Screening

One reviewer (SN) identified and removed all duplicates using Covidence (Veritas Health Innovation) [54], a systematic review software, after which 2 reviewers (SN and EG) screened the identified studies in a 2-stage screening process (title and abstract screening and full-text screening). We achieved a 75% minimum agreement in a pilot test of our screening process and reviewed any conflicts. During screening, some studies did not have well-defined care settings, leaving raters to apply the inclusion criteria differently. Furthermore, the studies spanned a range of methodologies and outcomes, introducing additional challenges in achieving consistent ratings. We reconciled discrepancies at both stages through discussion with the research team and recorded the reasons for study exclusion at each step.

Data Extraction and Analysis

We (EG, LH, RKM, and SN) created a data extraction chart (Multimedia Appendix 3) to capture relevant information (ie, setting, benefits, challenges, virtual modality, and patient population) in Covidence. To ensure its efficacy, we first piloted this chart on a small sample of studies. On the basis of insights from this pilot test, we refined the chart to better capture the nuances of health service features specific to primary care settings. This revised chart became the cornerstone for the subsequent data extraction from the studies. In parallel, EG and SN conducted quality appraisals of the studies using the Quality Assessment With Diverse Studies (QuADS) tool [55]. The QuADS tool assesses the quality of the methods, evidence, and reporting of studies included in systematic reviews. Our appraisal process involved scoring the studies on a scale from 0 to 3, where the score is derived from an evaluation across several areas, including the clarity of the research aims and robustness of the study design (Multimedia Appendix 4). No study that reached the data extraction stage was excluded due to their QuADS score. Following data extraction and quality appraisal, SN and EG reviewed the data to reach a consensus. We transferred the harmonized data to Microsoft Excel (Microsoft Corp) to conduct inductive content analysis, which led us to identify patterns, which we subsequently organized into themes.

For theme categorization, we used the Consolidated Framework for Implementation Research (CFIR) [56], a widely recognized meta-theoretical model previously applied in other virtual care reviews [57-59]. The CFIR comprises five major domains (Figure 1):

XSL•FO
RenderX

1. Innovation—the change being implemented. In the context of our scoping review, we equate this with synchronous virtual visits in primary care.
2. Outer setting—the context in which the inner setting exists. For this review, the outer setting is the state, including laws, policies, and institutions that inform the provision of and influence access to primary care.
3. Inner setting—the more immediate context where the innovation is applied. While virtual care was implemented throughout the health care system, we were specifically interested in how it was applied and experienced in primary care.
4. Individuals—the actors involved in the provision and experience of the given innovation. Within the primary care milieu, there are 2 principal actors—"innovation receivers" (in this study, people with OUD) and "innovation deliverers" (clinicians offering virtual primary care).
5. Implementation process—the methods and strategies used to implement the innovation. This includes facilitators and barriers that influence the implementation of virtual care.

We discuss facilitators of and barriers to the implementation process (domain 5) within each of the other 4 domains.

Figure 1. Consolidated Framework for Implementation Research domains as contextualized for this review.

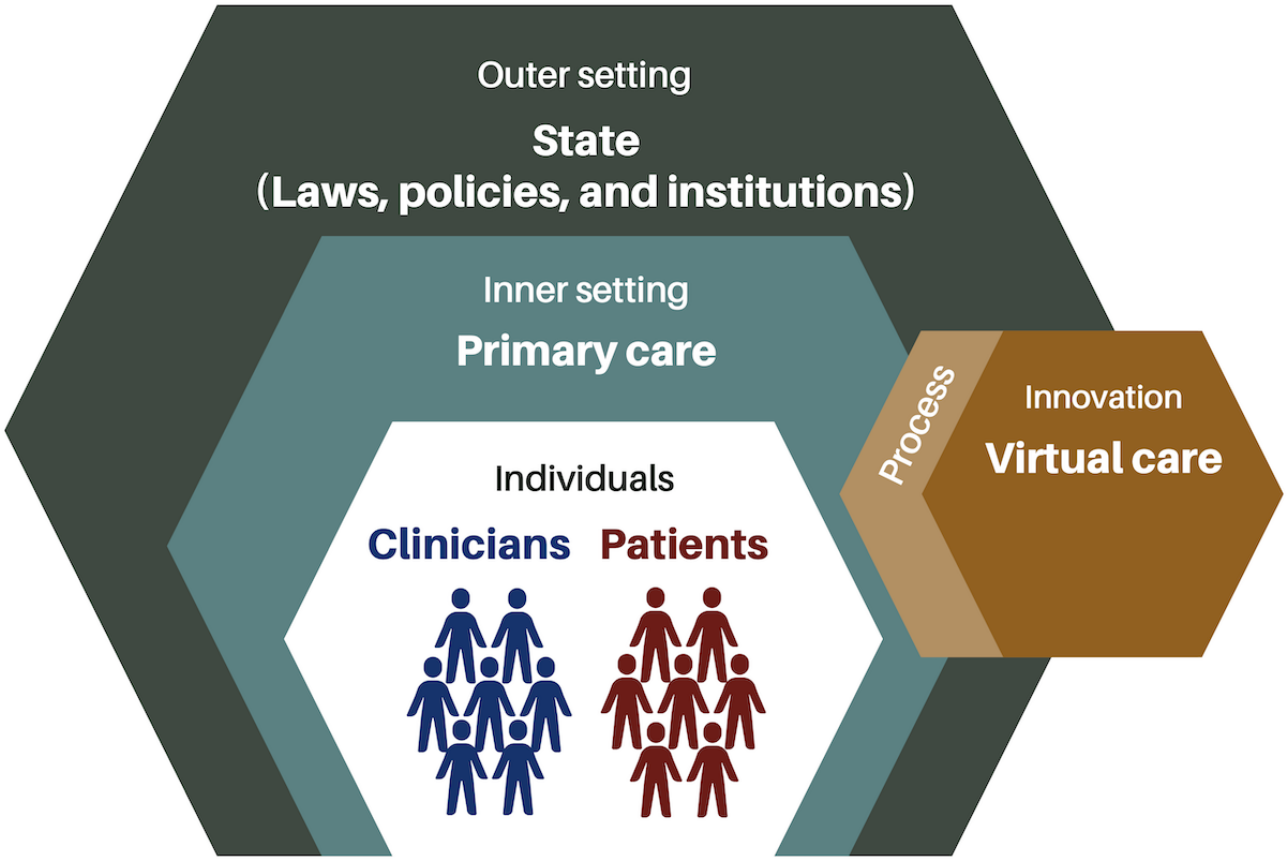

Results

Overview

We identified a total of 1474 studies from our searches of the Embase (Ovid; n=780, 52.92%), Web of Science Core Collection (n=283, 19.2%), Ovid MEDLINE (n=263, 17.84%), and CINAHL Complete (EBSCOhost; n=142, 9.63%) databases and through citation searching (n=6, 0.41%). After removing 36.36% (536/1474) of the identified sources due to duplication, we screened the titles and abstracts of 938 studies, retaining 155 (16.5%) for full-text review. After a double review process, we identified 28 papers for data extraction and quality assessment. A PRISMA flow diagram details the exclusions (Figure 2). In this section, numbers within parentheses indicate the frequency and percentage of the studies included in the review that support the statements.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart—identification, screening, and inclusion process. OUD: opioid use disorder.

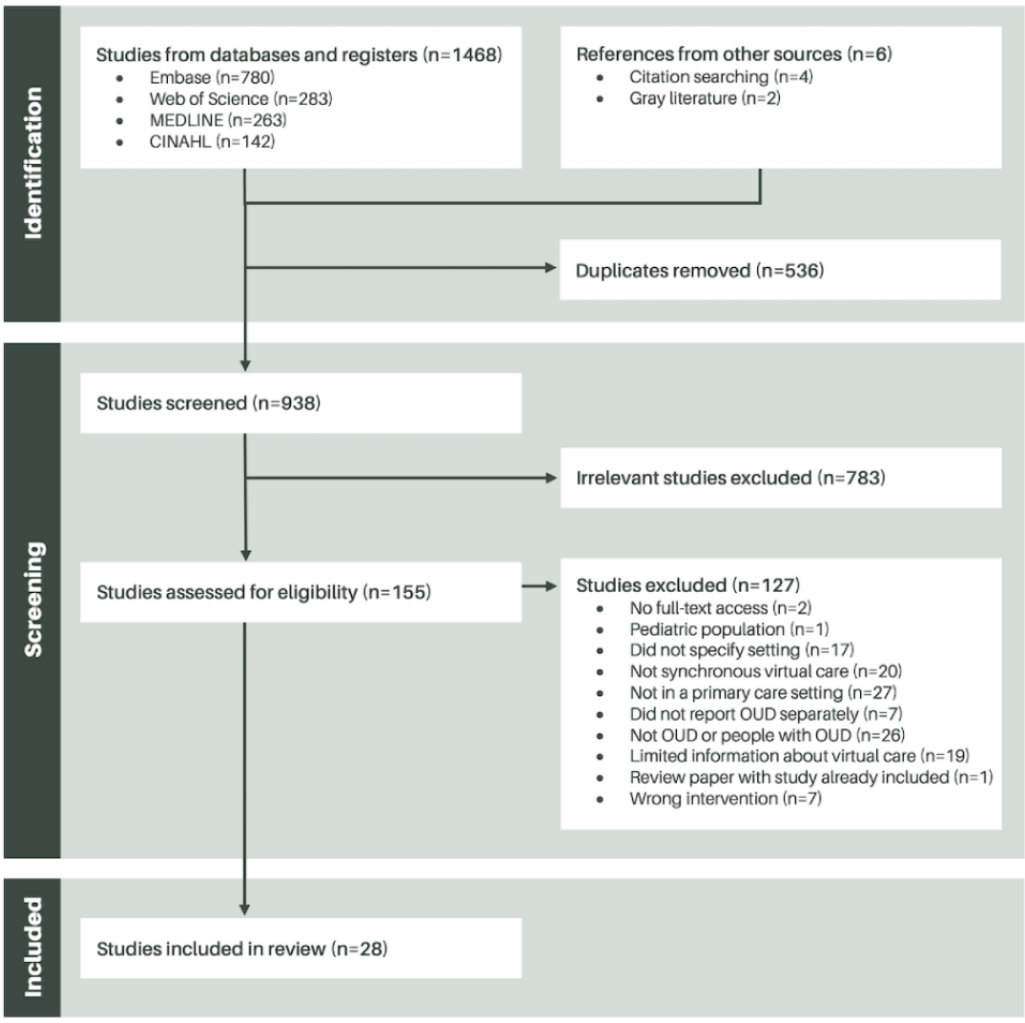

Study Characteristics

Table 1 provides a summary of the characteristics of each study with its corresponding citation. Most studies (25/28, 89%) were based in the United States and were published after the emergence of COVID-19. Only 14% (4/28) of the studies were conducted before the pandemic [60-63], emphasizing the paucity of research on virtual primary care for people with OUD conducted before 2020. Original research constituted the largest group of studies (11/28, 39%), using various study designs. Most studies had either quantitative (15/28, 54%) or qualitative (4/28, 14%) designs, whereas 4% (1/28) of the studies had mixed methods designs. Other studies reported their findings as commentaries (7/28, 25%) or economic evaluations (1/28, 4%).

The key objectives and results identified in each study are summarized in Table 2. A total of 68% (19/28) of the studies identified incorporating virtual care as part of their primary study purpose or design, whereas the remaining studies (9/28, 32%) included virtual care as an ancillary component of their research. The studies commonly reported the introduction of virtual modalities due to COVID-19 (17/28, 61%), with a few of these papers (6/28, 21%) also examining changes to or increases in existing virtual modality use in response to the pandemic. Outcomes included reports of shifts in health service use, such as changes in appointment modalities (16/28, 57%), frequency or demand (4/28, 14%), and differences in prescription duration (4/28, 14%). We reviewed the number of articles that reported on demographic factors given our interest in equitable practices and understanding disparities in research and health care delivery. Given that it is uncommon for commentaries to report demographics, we excluded them in the following analysis. Of the remaining 22 articles, 11 (50%) reported on sex or gender, 13 (59%) reported on race or ethnicity, and 10 (46%) reported the age of the participants. A total of 18% (4/22) of the studies specified collecting data on participants' gender [63,64,74,81], whereas 23% (5/22) specified the participants' sex [36,62,76,77,80]. In total, 9% (2/22) of the studies were unclear on whether they were reporting on sex or gender (eg, *female*/*male* were used, but the study did not specify whether sex or gender were reported) [60,82]. Benefits and challenges of using virtual primary care for people with OUD are grouped within the CFIR domains in the following sections and in Table 3.

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Study	Year	Location	Paper type	Study design
Akoto [61]	2017	United States	Conference abstract	Cohort
Beharie et al [63]	2021	United States	Original research	Qualitative
Behrends et al [64]	2022	United States	Original research	Cross-sectional
Calandra et al [65]	2022	United States	Commentary	Text and opinion
Caton et al [66]	2021	United States	Original research	Cross-sectional
Crowley and Delargy [67]	2020	Republic of Ireland	Brief report	Text and opinion
Griffin et al [68]	2021	United States	Conference abstract	Cross-sectional
Hedden et al [36]	2022	Canada	Protocol	Mixed methods
Hodgkin et al [69]	2021	United States	Review	Economic evaluation
Hser and Mooney [70]	2021	United States	Commentary	Text and opinion
Hser et al [71]	2021	United States	Commentary	Text and opinion
Huskamp et al [72]	2022	United States	Original research	Cross-sectional
Incze et al [73]	2022	Not specified	Conference abstract	Qualitative
Jones et al [74]	2021	United States	Original research	Cross-sectional
Lee et al [60]	2009	United States	Original research	Nonrandomized experimental
Leo et al [75]	2021	United States	Commentary	Text and opinion
Lin et al [62]	2022	United States	Original research	Cohort
O'Gurek [76]	2021	United States	Brief report	Cohort
Patel et al [77]	2022	United States	Brief report	Cohort
Riedel et al [78]	2021	United States	Original research	Cross-sectional
Sahu et al [79]	2022	United States	Conference abstract	Case-control
Sivakumar et al [80]	2022	United States	Original research	Cohort
Snell-Rood et al [81]	2021	United States	Original research	Qualitative
Taylor et al [82]	2021	United States	Conference abstract	Cross-sectional
Thompson et al [83]	2021	United States	Conference abstract	Cross-sectional
Uscher-Pines et al [84]	2020	United States	Original research	Qualitative
Wang et al [85]	2021	United States	Commentary	Text and opinion
Wilson et al [86]	2021	United States	Commentary	Text and opinion

XSL•FO
RenderX

Table 2. Objectives, focus of virtual care findings, and results specific to virtual primary care identified during data extraction.

Study	Objective	Virtual care findings	Results
Akoto [61]	Measure adherence to in-person versus video appointments in outpatient primary care	Primary	Of 228 scheduled appointments, 149 (65%) were in person, and 79 (35%) were through video. A total of 93 (60%) in-person and 79 (100%) video appointments were conducted. Virtual care offered benefits such as cost savings and reduced travel time for patients.
Beharie et al [63]	Understand patient perspectives and experiences of care provided by nurse managers	Ancillary	Frequent nurse care manager engagement through phone calls could encourage patient retention. Offering educational and technical assistance for clinicians facilitated virtual care. The Buprenorphine Nurse Care Manager Initiative facilitated patient screening, appointment coordination, pharmacy and medication navigation, provision of care management, and advocacy to enhance treatment engagement and retention. Patients identified barriers such as inflexible schedules for virtual care and unfulfilled promises of phone follow-ups.
Behrends et al [64]	Measure how syringe service programs changed between 2019 and 2020 due to the COVID-19 pandemic	Ancillary	The proportion of syringe service programs offering virtual OAT[a] prescribing increased from 3% in 2019 to 8% by the end of 2020. Virtual primary care increased from 0% to 8% in the same period. The increase in programs offering virtual mental health care or primary care was statistically significant in facilitating extended virtual service provision. Relaxed prescribing regulations by the DEA[b] and SAMHSA[c] facilitated virtual care provision. The rise in virtual care demand may be attributed to disruptions in drug access during the COVID-19 pandemic and increased OAT needs in areas with high fentanyl use.
Calandra et al [65]	Describe how a family medicine center provides care to a clinically and socially complex rural patient population	Ancillary	COVID-19 accelerated the adoption of virtual care to enhance patient accessibility. Virtual care benefits included better antiviral adherence, flexible communication with care coordinators, eased billing for clinicians, reduced travel, fewer work disruptions, enhanced technology access through community agency partnerships, improved attendance, and consistent care. Many clinicians alternated between in-person appointments and telehealth for stable patients with substance use disorder. Challenges included limited broadband access, technology glitches, inadequate technology knowledge, outdated contacts leading to more administrative tasks, mandatory paperwork and laboratory tests to start visits, restricted physical examination scopes, and reduced reimbursement rates compared to in-person visits.
Caton et al [66]	Understand how appointment frequency and type and patient management changed due to the COVID-19 pandemic	Primary	Most (91%) clinics modified care access to either a hybrid of in-person and virtual appointments or all-virtual appointments. Many medical (40%) and mental health care (54%) appointments were virtual. In-person appointments were mostly used for induction of OAT, although some clinics offered this virtually. No clinics required in-person visits for follow-up OAT care. The frequency of follow-up appointments was unchanged. Buprenorphine prescriptions were longer; rates of injectable buprenorphine or naltrexone were unchanged. Larger clinics wrote prescriptions of a longer duration. There was less urine drug screening conducted at most (67%) clinics. Clinics varied in their success with patient retention. Clinicians indicated an increased demand for OAT and mental health care appointments and perceived an increased preference for virtual appointments from patients. Benefits of virtual care included provision of follow-up counseling, medical and behavioral health sessions for stable patients, decreased transportation issues, modified billing for virtual visits, proactive patient outreach, easier medication initiation, and fewer no-shows. However, a challenge to virtual care was clinicians' reluctance to onboard new patients virtually, perceiving it as high risk and posing a greater liability.
Crowley and Delargy [67]	Describe how to provide virtual OAT care to people with OUD[d]	Primary	COVID-19 transmission risk was reduced. Benefits of virtual care included less travel, easier access for rural patients, enhanced OAT availability, shorter wait times, and the provision of remote psychological support from local drug treatment services, with visit modalities tailored to each patient. A key facilitator of virtual visits was supplying phones to patients in need. However, challenges involved risks from quicker assessments, fewer pretreatment drug screenings, and decreased patient supervision.
Griffin et al [68]	Identify the barriers to virtual care during the COVID-19 pandemic and the feasibility of using community health workers to facilitate virtual care	Primary	A total of 67 community health worker–facilitated virtual appointments were held with 23 unique patients (2.9 mean visits per patient) between April 2020 and December 2020. Approximately half of the appointments were scheduled, and the remaining were walk-in–style virtual appointments. Providing community health workers with smartphones may be a viable strategy to enhance telehealth access and engagement for people experiencing homelessness who might otherwise lack the essential digital tools.

XSL•FO
RenderX

Study	Objective	Virtual care findings	Results
Hedden et al [36]	Understand the impact of expanded virtual primary care on provision of and access to primary care and health outcomes for people with OUD	Primary	The protocol described a mixed methods study with three components focused on virtual primary care for OUD: (1) qualitative interviews with 20 family physicians and 30 people with OUD to understand their experiences with virtual visits, (2) quantitative analysis of population-based administrative data to describe the use of virtual care and its impact on access to services and patient outcomes, and (3) deliberative dialogue sessions to jointly develop resources that promote equitable access to high-quality virtual primary care.
Hodgkin et al [69]	Evaluate the feasibility of office-based treatment	Ancillary	The DEA has allowed buprenorphine to be prescribed virtually, and some insurance companies have temporarily begun covering virtual OAT prescriptions, a service that they previously did not reimburse. Although there are various billing codes, some virtual services receive lower reimbursements than in-person visits, leading to revenue loss for clinics. The continuation of these policies, introduced due to the pandemic to improve OAT access, remains uncertain beyond the COVID-19 pandemic.
Hser and Mooney [70]	Describe high-quality virtual care and opportunities to expand access to virtual care	Primary	Virtual delivery of OAT management was as effective as in-person delivery. Video appointments could be used beyond OAT management, providing opportunities for remote drug screening and mental health care as well. Facilitators of prescribing OAT virtually included more lenient virtual regulations. The collaboration between primary care and seasoned virtual clinicians allowed for more comprehensive access. Factors driving virtual referrals included a shortage of X-waivered clinicians, an absence of behavioral health services, patients' logistical issues, and managing clinically complex cases. Virtual care offered benefits such as eliminating travel, crucial for rural areas, and provided clinic flexibility and broader patient access. Several telemedicine firms had infrastructure in place, with 24/7 services from licensed clinicians and diverse payment options spanning many areas in the United States. This addressed challenges in rural areas, better catering to patient preferences regarding stigma and privacy. However, rural communities faced challenges such as limited economies of scale, dependency on public payers, and low patient numbers. Many rural residents lacked proper internet and smartphone access. Moreover, some patients were skeptical of digital health care quality. Regulations posed challenges to clinicians in terms of licensing and reimbursement, whereas patients struggled with technology use.
Hser et al [71]	Describe the implementation and effectiveness of virtual care in rural settings	Primary	Clinics provided virtual care (before the COVID-19 pandemic) to address capacity limitations, improve access to mental health care for complex or nonadherent patients, provide more options for patients, and reduce waitlist times. In clinics offering virtual appointments to all patients, most people with OUD did not accept a referral to the virtual program. Rural health centers could expand their capacity and improve patient care by collaborating with established telemedicine vendors. Through a structured protocol, they defined services, referral methods, and communication strategies to address patient retention. To facilitate this, clinics provided spaces and devices for patients, initiated community outreach, and offered additional training for clinicians. However, challenges remained. Rural areas often lacked adequate internet and device access. There were initial setup delays, insurance inconsistencies, regulatory hurdles, and uncertainty about virtual care aligning with treatment philosophies. Some patients also expressed hesitancy toward unfamiliar web-based platforms and the absence of private spaces for virtual visits, whereas clinics grappled with managing patient relationships remotely.

XSL•FO
RenderX

Study	Objective	Virtual care findings	Results
Huskamp et al [72]	Understand clinician use of virtual modalities and comfort with initiating OAT remotely	Primary	Note: the results include primary care physicians (40%), psychiatrists, NPs[e], and physician assistants. Of 602 clinicians, 58% did not use virtual care for their patients with OUD in 2019. Of the 506 clinicians who provided OAT initiation in the previous month, 60% used virtual modalities, which accounted for 34% of all of their initiation visits. The remaining 40% of OAT-initiating clinicians only provided in-person appointments. In total, 56% of clinicians used virtual modalities for all OUD appointments. Of the 303 clinicians who provided at least one virtual initiation in the previous month, 33% used video for most of their initiation visits versus 8% who used predominately phone. Among this same subset of clinicians who offered virtual initiation in the previous month, 42% were primary care physicians. Facilitators of virtual care included relaxed OAT regulations, experienced clinicians, use of home urine screening kits, remote patient monitoring, Veterans Administration EMR[f] guidance, clinician support initiatives, and virtual care training during residency. Challenges included clinician hesitancy toward virtual OAT initiations, liability concerns, safety issues, perceptions of diminished care quality with phone visits, constraints previously imposed by the Ryan Haight Act, clinician discomfort, and the "digital divide" that limits access to video visit technologies for many vulnerable patients.
Incze et al [73]	Describe patient experiences of receiving OUD care in primary care	Ancillary	Respectful, team-based care facilitated flexibility for virtual appointments. Virtual care for people with OUD offered benefits such as flexibility in scheduling appointments. Furthermore, it reduced feelings of discrimination and stigma. Participants felt that, when accessing virtual care for substance use disorder, it was akin to consulting a physician for any other medical concern.
Jones et al [74]	Assess adaptations to care for people with OUD, including the use of virtual care, by clinicians during the COVID-19 pandemic	Primary	Before the COVID-19 pandemic, 21% of clinicians prescribed OAT virtually for their established patients. Since the COVID-19 pandemic, clinicians reported either no change (60%) or increased demand (29%) for treatment from patients. In total, 33% of clinicians prescribed virtually to new patients, among whom 87% reported no difficulties with virtual buprenorphine induction. Where clinicians did encounter problems with virtual induction, these included withdrawal symptoms (52%), oversedation (8%), allergic reaction (2%), or other challenges (41%). Clinicians used several virtual modalities with new patients, including computer with video (54%), phone (54%), phone with video (46%), tablet with video (19%), and others (3%). For existing patients, virtual modalities included phone (64%), computer with video (51%), phone with video (47%), tablet with video (21%), and others (8%). The most common applications used for virtual appointments were Zoom (36%), FaceTime (15%), and Skype (6%). Clinician strategies for virtually monitoring patients included prescribing naloxone (50%), mental health care with video (51%) and without video (47%), group counseling with video (12%) and without video (6%), and pill and film checks with video (9%). Since the COVID-19 pandemic, most (72%) clinicians did not change the length of buprenorphine prescriptions, whereas some (24%) increased the number of days. Most clinicians (73%) did not change their use of long-lasting or extended buprenorphine, whereas for others (18%), it was not available. Virtual care was particularly embraced by younger clinicians, many of whom had previous experience with virtual prescribing and handled a large number of OAT patients. These services could be found across various practice settings, but there were notable challenges such as initiating buprenorphine treatments virtually and managing withdrawal symptoms from a distance. Technological barriers such as unreliable systems for both patients and clinicians posed significant problems, compounded by concerns about security and privacy. Financial challenges related to accessing technology and inadequate reimbursements for virtual services further complicated the situation. Some physicians preferred in-person consultations, and some clinics continued to operate primarily in person. Additional barriers included physician hesitation to treat remotely without a previous in-person interaction and specific challenges tied to home inductions. The lack of awareness of DEA exemptions was also a hindrance. Clinicians were less likely to prescribe buprenorphine virtually to new patients without an in-person examination if they were male (OR[g] 0.83) compared to female, practiced in the Midwest (OR 0.74) or the South (OR 0.68) as opposed to the Northeast, worked in suburban (OR 0.80) or rural areas (OR 0.67) rather than urban settings, were physician assistants (OR 0.52) compared to physicians, or worked in an emergency department (OR 0.52) rather than in an office-based solo practice.

XSL•FO
RenderX

Study	Objective	Virtual care findings	Results
Lee et al [60]	Describe the outcomes of a home-only OAT induction program for patients accessing office-based primary care OUD treatment	Ancillary	Of the 103 participants who received home induction for buprenorphine between August 2006 and January 2008, a total of 73% returned for their scheduled week 1 visit, 17% did not return but did communicate with staff, and 11% had no contact with the clinic following induction. There were no reports of issues with sublingual dosing, no participants reported severe withdrawal or adverse events, and few (5%) reported buprenorphine-prompted withdrawal symptoms. In a convenience subsample of 17 patients, 18% contacted staff as instructed on induction days 1-3, a total of 42% contacted staff on days 4-6, a total of 24% did not contact staff but did return messages, and 35% did not respond to follow-up. These postinduction support calls lasted an average of 3-4 minutes and focused on the induction experience. These calls were considered infrequent due to low need (attributed to the patient education pamphlets provided), low rates of complications, and patients' previous experience with buprenorphine.
Leo et al [75]	Describe changes in access to care for people experiencing homelessness resulting from a community partnership that used virtual care during the COVID-19 pandemic	Primary	Working with existing frontline organizations to provide mobile health van deliveries of buprenorphine to patients following a virtual care appointment ensured continued access to medication. Benefits of the program included weekly multidisciplinary meetings to ensure patient follow-up, medication delivery via a mobile health van, and prioritizing team members who had long-standing therapeutic relationships with patients. In addition, the mobile clinic was equipped with phones for virtual visits and, due to DEA regulatory changes, mobile health organizations could prescribe buprenorphine via video or phone. Barriers experienced were related to generational distrust toward health care systems among racialized communities.
Lin et al [62]	Understand how virtual care use changed over time, which patients are using virtual care, and any differences in health outcomes between in-person and virtual care	Primary	Virtual prescribing of buprenorphine increased 3.5 times between 2012 and 2019, from 2% to 8% of patients receiving buprenorphine. Individual mental health care for OAT patients also increased, whereas group mental health care did not. Primary care only accounted for 0.3% of clinic settings in which virtual OAT management was provided. Patients receiving virtual buprenorphine management were more likely to have depressive or anxiety disorders but less likely to have psychotic, alcohol use, stimulant use, or cannabis use disorders than patients receiving buprenorphine management in-person. Health service use was also higher for patients receiving buprenorphine management virtually (a median of 39 appointments, 4 of which were virtual) than for those receiving buprenorphine management in-person (a total of 20 appointments). Patients accessing their OAT management virtually also had a higher median number of days on their buprenorphine prescription (722) than in-person patients (295). Patients receiving buprenorphine management virtually accessed an average of 18 mental health-related visits (1 of which was provided virtually) compared to 15 mental health-related visits for individuals receiving buprenorphine management in-person. Benefits of virtual care included enhanced access to care in rural areas, increased treatment retention, greater accessibility overall, and regulatory changes that favored telehealth. Especially during the COVID-19 pandemic, with barriers to in-person treatment prevalent, virtual care may have improved engagement and use of buprenorphine. Facilitators of virtual care included the introduction of the DEA waiver to prescribe virtually and in community-based settings. Younger, female, and White patients were more likely to use virtual care. Conversely, challenges involved the necessity for patients to be more stable for this approach. Black patients were less inclined to receive their buprenorphine management virtually.
O'Gurek [76]	Understand how appointment adherence is affected by a disaster management plan and virtual care protocols	Primary	Following a change to virtual modalities (video or phone based on patient preference), virtual appointment adherence rates from March 16, 2020, to April 30, 2020, were 92% compared to 74% adherence to in-person appointments between January 1, 2020, and March 13, 2020. There was a statistically significant difference between no-show rates in the pre–COVID-19 (26%) and COVID-19 (8%) periods studied. Benefits of virtual care included cost savings for patients, efficient use of clinician time evidenced by lower no-show rates, and reduced expenses on office space and staff. Virtual care also enhanced access to services. It was also convenient and decreased the need for travel. Guidelines for ongoing care during the pandemic supported its implementation, and the introduction of web-based recovery resources to support behavioral health therapy provided comprehensive recovery care. Conversely, barriers involved initial costs for HIPAA[h]-compliant technologies, restricted broadband access especially in rural regions, and challenges integrating virtual health services that comply with Title 42 (Part 2) of the Code of Federal Regulations–Confidentiality of Substance Use Disorder Patient Records.

XSL•FO
RenderX

Study	Objective	Virtual care findings	Results
Patel et al [77]	Describe the characteristics of patients and clinicians who used virtual care during the COVID-19 pandemic	Primary	Between March 13, 2020, and May 30, 2021, a total of 15% of inductions were provided virtually. The highest use of virtual inductions (34%) was in April 2020, and this dropped to 14% in May 2021. During this time, primary care physicians and NPs conducted 31% and 16% of virtual inductions, respectively; a psychiatrist was more likely to conduct a virtual induction (OR 2.01) than a primary care clinician. Facilitators for virtual care adoption included the Ryan Haight Act waiver, community broadband accessibility, community programs educating patients on virtual visit procedures, and financial assistance for practices to cover initial virtual care setup costs. Larger practice settings facilitated the implementation of virtual care. Conversely, barriers involved the "digital divide," where a lack of reliable technology tools and internet access impeded virtual care use for marginalized populations, potentially widening care disparities. Solo practitioners or primary care clinicians may not have the necessary resources to offer OUD care using virtual modalities. Furthermore, equity-deserving groups, including Black patients, rural residents, and those with lower income, were less likely to experience virtual OAT induction. Compared to solo practitioners, clinicians in larger practices were more likely to use virtual care for OAT induction.
Riedel et al [78]	Understand clinician perspectives on virtual care during the COVID-19 pandemic, including the frequency, barriers to, effectiveness of, comfort level with, and compensation of virtual appointments	Primary	In 2019, most clinicians (58%) did not use virtual care for people with OUD. Clinicians reported in the fall of 2020 that, in the previous month, they had conducted most (57%) of their appointments virtually, both using video (37%) and audio-only (20%) modalities. Primary care clinicians conducted 50% of their appointments in-person, whereas they used video (31%) and audio-only (20%) modalities for virtual appointments. NPs and certified nurse specialists saw patients virtually using video (38%) and audio-only (20%) modalities. Physician assistants saw patients virtually using video (30%) and audio-only (14%) modalities. If reimbursement rates did not change, most (80%) of the clinicians said that they would continue with virtual care. If reimbursement rates were 25% lower for virtual than for in-person appointments, only some (40%) said that they would offer virtual care. Facilitators of virtual care included the Ryan Haight Act waiver and clinician comfort using video modalities with clinically stable patients. Furthermore, clinicians with more patients with OUD used virtual appointments and felt more comfortable using video with less stable patients. Certain practices, especially solo and unspecified "other" types, leaned into using video, and a substantial portion of practitioners (63%) affirmed the efficacy of virtual visits compared to in-person visits. Conversely, barriers included preference for audio-only visits due to technological barriers and a lack of patient readiness or resources. Challenges also emerged at the clinician level, with technological and infrastructural issues (eg, lack of HIPAA-compliant technology), licensing regulations, and a perception among some (33%) that virtual care was less effective than in-person care. In addition, after the pandemic, a significant number of clinicians (70%) expressed a preference for returning to in-person visits, revealing a nuanced landscape of clinician perspectives on the sustained application of virtual care in OUD management. Approximately 33% of respondents believed that licensing regulations restricted them from using virtual care for people with OUD. Approximately 49% were comfortable with video consultations for patients who are less stable. Only 3% did not want any virtual visits, whereas 28% wanted them in special situations, such as for patients with mobility challenges or those who were traveling. Meanwhile, 39% favored virtual care often but solely for their existing patients, and another 29% were amenable to using it regularly for both new and established patients.
Sahu et al [79]	Evaluate how virtual care affected patient continuity for individuals receiving buprenorphine	Primary	Among patients seen virtually, most (61%) received follow-up care, which was higher than the rate for patients seen in person (49%). Most patients (80%) experienced at least one socially-determined barrier to their health (eg, access to housing and transportation). Benefits included improved continuity and accessibility of care. Telehealth could also mitigate challenges such as transportation, work commitments, childcare, and other demands, making care more attainable.

XSL•FO
RenderX

Study	Objective	Virtual care findings	Results
Sivakumar et al [80]	Assess whether a differentiated care model for combined HCV[i] care and OUD screening that used virtual modalities reduced clinical demands during the COVID-19 pandemic	Primary	Combining HCV and OUD care can be successful, providing timely treatment, whether via phone call, video call (for buprenorphine), or in-person (for methadone). Of those patients with HCV and OUD, 45% were initiated on buprenorphine, and 19.4% were initiated on methadone as well as direct-acting antivirals in addition to those already receiving OAT at baseline. In total, 84% of those with HCV and OUD received treatment for both conditions. Telehealth was beneficial for individuals with unstable access to regular in-person care due to socioeconomic or structural issues such as housing instability, lack of insurance, or stigma from health care settings. Limited clinician oversight was required for virtual OAT prescriptions, and in-person contact was not mandatory for buprenorphine prescriptions. Clinically stable patients could opt for take-home dosing with methadone, and phone communication proved sufficient for patient engagement. However, there were concerns about medication adherence. Regulations still required in-person induction for methadone treatment during the pandemic. In many settings, virtual care reimbursement required broadband internet and videoconferencing. To promote health equity, a human rights approach over a reimbursement-driven approach must be prioritized wherein standard phones rather than smartphones can be used for virtual care.
Snell-Rood et al [81]	Capture clinician perspectives on the facilitators of and barriers to rural care	Ancillary	Virtual care was not sufficiently available despite efforts for its integration. The use of virtual modalities, which allows some care providers to work remotely, can impede relationship building between care providers and with patients; this can result in medication errors due to unfamiliarity with patients and insufficient chart reviews. The loosened telehealth regulations may not have sufficiently aided rural organizations, which could require support in establishing lasting connections with remote clinicians. Rural residents often lacked essential broadband and technology for telehealth. Virtual care was often described by participants as an option that their organizations were exploring rather than implementing. Furthermore, virtual care implementation within organizations fluctuated due to staff turnover or equipment changes.
Taylor et al [82]	Capture patient perceptions of a new program offering primary care–based Suboxone induction during the COVID-19 pandemic	Ancillary	A total of 76% of the participants reported being able to call in for an appointment. Benefits of virtual care included lower COVID-19 exposure risk.
Thompson et al [83]	Understand the barriers to care provision associated with OAT	Ancillary	A total of 70% of primary care clinicians with a buprenorphine waiver used virtual appointments for people with OUD. A significant benefit was that most waivered clinicians were using virtual care to support patients in rural areas, filling gaps caused by the absence of OUD services in those regions.
Uscher-Pines et al [84]	Describe how virtual and in-person care were used during the COVID-19 pandemic, as well as the barriers and quality-of-care implications	Primary	As of April 2020, almost all clinicians (94%) used virtual care for people with OUD, and most (72%) did not offer any in-person visits. Most clinicians (83%) offered either phone or video appointments. While most clinicians indicated a preference for video over phone appointments, a median of 20% of appointments were conducted via phone. Common video platforms included Zoom, FaceTime, and Google Meet. Virtual appointments took place in patients' homes or outside the clinic following an in-person urine toxicology screening, although the frequency of these tests decreased or ceased. For some clinicians, appointment frequency decreased, and the length of prescriptions increased. Overall, clinicians felt that transitioning to virtual care went well and that it would be retained in the future in conjunction with in-person care. Benefits of virtual care included enhanced patient access and convenience as it eliminated the need for travel and reduced no-show rates. Patients had flexible options for care outside typical business hours, which decreased wait times and aided in more efficient clinic operations. Virtual care also alleviated pandemic-induced anxiety and provided clinicians with a unique insight into patients' environments, fostering a deeper connection. Adjustments due to COVID-19, such as reduced urine screening and an increased supply of OAT, were received positively. Many found virtual visits comparable to in-person visits and anticipated their continued use after the pandemic for their convenience. However, challenges persisted, including a perceived lack of structure and accountability, technological barriers, shorter visit durations, and physical requirements such as laboratory tests. In addition, there was hesitation in onboarding new patients remotely, especially those with recent relapse. The future of virtual care remained uncertain given potential changes to reimbursement and regulations.

XSL•FO
RenderX

Study	Objective	Virtual care findings	Results
Wang et al [85]	Describe how virtual care was used during the COVID-19 pandemic	Primary	Examples from the clinics highlighted how virtual appointments facilitated a low-barrier, faster, and more flexible OAT induction and access to naloxone. Virtual care for OUD reduced gaps in treatment and mortality. Regulatory modifications streamlined the initiation process for patients seeking buprenorphine as in-person evaluations were not mandatory. This low-barrier approach, aided by the remote guidance of medical students and patient navigators, enabled patients to conveniently access virtual care. Phones with comprehensive plans were distributed to those in need, ensuring that community health workers could facilitate virtual care during outreach activities. This augmented approach enhanced access to OUD treatment, breaking down barriers ranging from limited clinician availability; racial disparities; and other patient challenges such as stigma, transportation issues, and competing priorities such as housing and food. However, in-person interactions were still frequently viewed as more engaging and motivational, with additional benefits for clinicians, such as urine toxicology tests. Virtual care also introduced new disparities as rural residents, racial minority groups, older adults, and low-income individuals may lack the requisite technological equipment or digital literacy. Connectivity was a persistent issue, especially in rural areas. The absence of reliable smartphone access among older patients, those with limited English proficiency, or those of a lower socioeconomic background further exacerbated this challenge. Clinicians sought evidence-backed virtual care protocols to better understand patient management in this digital era amid concerns that virtual care may compromise the treatment structure or increase medication diversion.
Wilson et al [86]	Describe how virtual care was used during the COVID-19 pandemic to provide patient-centered care	Primary	This clinic took a harm reduction approach during the COVID-19 pandemic by spacing out OAT appointments when appropriate, electronically prescribing "buprenorphine/naloxone," and offering phone appointments when possible and based on patient needs. Clinicians also ensured that all patients had an active prescription for naloxone. They noted that virtual care reduced the risk of COVID-19 exposure, which has been supported by insurance changes to reimbursement for such visits. Guidance from the DEA on electronic prescribing, along with a centralized hub for resources on billing, regulations, and compliance, further facilitated virtual care adoption. Virtual care was particularly beneficial for low- to moderate-acuity patients who did not require in-office visits or urine drug screens, offering added appointment flexibility. Conversely, inconsistent cell phone service and limited broadband internet could impede virtual care access. Insurance reimbursement could pose difficulties, and there were concerns that the absence of urine screenings might lead to relapse among some patients. Some clinicians deemed virtual care unsuitable for high-acuity patients such as recent drug users, pregnant women in their third trimester, or those new to care, necessitating in-person appointments.

[a]OAT: opioid agonist therapy.

[b]DEA: Drug Enforcement Administration.

[c]SAMHSA: Substance Abuse and Mental Health Services Administration.

[d]OUD: opioid use disorder.

[e]NP: nurse practitioner.

[f]EMR: electronic medical record.

[g]OR: odds ratio.

[h]HIPAA: Health Insurance Portability and Accountability Act.

[i]HCV: hepatitis C virus.

XSL•FO
RenderX

Table 3. Benefits and challenges of virtual primary care for people with opioid use disorder organized by Consolidated Framework for Implementation Research (CFIR) domains and constructs (N=28).

CFIR domain	CFIR construct	Studies, n (%)	References
Innovation: virtual primary care			
Benefits			
Modality options	Innovation and adaptability	7 (25)	• Sivakumar et al [80] • Uscher-Pines et al [84] • Huskamp et al [72] • Jones et al [74] • Hser et al [71] • Wilson et al [86] • Crowley and Delargy [67]
Challenges			
—[a]	—	—	—
Outer setting: state			
Benefits			
Regulation changes	Policies and laws	11 (39)	• Lin et al [62] • Behrends et al [64] • Patel et al [77] • Huskamp et al [72] • Hser and Mooney [70] • Jones et al [74] • Wang et al [85] • Riedel et al [78] • Wilson et al [86] • Hodgkin et al [69] • Leo et al [75]
COVID-19 public health measures	Critical incidents	4 (14)	• Lin et al [62] • Uscher-Pines et al [84] • Wilson et al [86] • Taylor et al [82]
Billing changes	Policies and laws	4 (14)	• Calandra et al [65] • Hser and Mooney [70] • Caton et al [66] • Hodgkin et al [69]
Insurance reimbursement	Policies and laws	2 (7)	• Wilson et al [86] • Hodgkin et al [69]
Support for up-front costs	Financing	1 (4)	• Patel et al [77]
Community engagement	Partnerships and connections	1 (4)	• Hser et al [71]
Changes in the drug market during the pandemic	Critical incidents	1 (4)	• Behrends et al [64]
Challenges			
Regulations	Policies and laws	7 (25)	• Sivakumar et al [80] • Huskamp et al [72] • Hser and Mooney [70] • O'Gurek [76] • Jones et al [74] • Hser et al [71] • Riedel et al [78]
Billing	Policies and laws	4 (14)	• Beharie et al [63] • Sivakumar et al [80] • Calandra et al [65] • Jones et al [74]

XSL•FO
RenderX

CFIR domain	CFIR construct	Studies, n (%)	References
Insurance reimbursement	Policies and laws	4 (14)	• Hser et al [71] • Riedel et al [78] • Wilson et al [86] • Hodgkin et al [69]
Inner setting: primary care			
Benefits			
Clinical resources available	Available resources	6 (21)	• Huskamp et al [72] • Hser and Mooney [70] • O'Gurek [76] • Wang et al [85] • Hser et al [71] • Wilson et al [86]
Technological supports	Available resources	5 (18)	• Beharie et al [63] • Patel et al [77] • Hser and Mooney [70] • Wang et al [85] • Hser et al [71]
Technology provided	Available resources	5 (18)	• Griffin et al [68] • Wang et al [85] • Hser et al [71] • Crowley and Delargy [67] • Leo et al [75]
Clinical setting	Structural characteristics	4 (14)	• Lin et al [62] • Patel et al [77] • Jones et al [74] • Riedel et al [78]
Mobile medication delivery	Available resources	1 (4)	• Leo et al [75]
Challenges			
Clinical setting	Structural characteristics	4 (14)	• Lin et al [62] • Patel et al [77] • Jones et al [74] • Riedel et al [78]
Lack of guidelines and resources	Available resources	4 (14)	• Jones et al [74] • Wang et al [85] • Hser et al [71] • Snell-Rood et al [81]
Need for more administrative support	Structural characteristics	3 (11)	• Calandra et al [65] • Hser et al [71] • Snell-Rood et al [81]
Cost	Cost	1 (4)	• O'Gurek [76]
Slow uptake of virtual primary care	Capability	3 (11)	• Jones et al [74] • Hser et al [71] • Snell-Rood et al [81]
Individual: innovation deliverer (clinicians)			
Benefits			

XSL•FO
RenderX

CFIR domain	CFIR construct	Studies, n (%)	References
Patient stability	Opportunity	5 (18)	• Sivakumar et al [80] • Calandra et al [65] • Riedel et al [78] • Caton et al [66] • Wilson et al [86]
Higher patient attendance	Need	5 (18)	• Uscher-Pines et al [84] • Calandra et al [65] • O'Gurek [76] • Hser et al [71] • Caton et al [66]
Ability to provide care virtually	Capability	3 (11)	• Uscher-Pines et al [84] • Huskamp et al [72] • Riedel et al [78]
Experience with OAT[b] or virtual primary care	Capability	3 (11)	• Huskamp et al [72] • Jones et al [74] • Riedel et al [78]
Continuity of care	Need	3 (11)	• Calandra et al [65] • Hser et al [71] • Leo et al [75]
Desire for virtual care	Motivation	2 (7)	• Uscher-Pines et al [84] • Riedel et al [78]
Patient bias	Need	1 (4)	• Wang et al [85]
Convenience and efficiency	Need	1 (4)	• O'Gurek [76]
Video appointments for higher-risk patients	Need	1 (4)	• Riedel et al [78]
Younger clinician age	Capability	1 (4)	• Jones et al [74]
Managing clinical complexity	Need	1 (4)	• Hser and Mooney [70]
Challenges			

XSL•FO
RenderX

CFIR domain	CFIR construct	Studies, n (%)	References
Higher risk and liability	Need	10 (36)	• Lin et al [62] • Sivakumar et al [80] • Uscher-Pines et al [84] • Huskamp et al [72] • Jones et al [74] • Wang et al [85] • Riedel et al [78] • Caton et al [66] • Wilson et al [86] • Crowley and Delargy [67]
Hinders patient relations	Need	3 (11)	• Beharie et al [63] • Uscher-Pines et al [84] • Hser et al [71]
Phone insufficient for care needed	Need	3 (11)	• Sivakumar et al [80] • Huskamp et al [72] • Riedel et al [78]
Clinician preference for in-person care	Motivation	3 (11)	• Huskamp et al [72] • Jones et al [74] • Riedel et al [78]
Limited support or guidelines available	Need	3 (11)	• Wang et al [85] • Hser et al [71] • Snell-Rood et al [81]
Lack of technology knowledge or equipment	Capability	3 (11)	• Uscher-Pines et al [84] • Jones et al [74] • Riedel et al [78]
Patient needing in-person care	Need	2 (7)	• Uscher-Pines et al [84] • Calandra et al [65]
Future of virtual primary care unclear	Motivation	2 (7)	• Uscher-Pines et al [84] • Hodgkin et al [69]
Shorter appointment times	Opportunity	1 (4)	• Uscher-Pines et al [84]
Male clinician	Other	1 (4)	• Jones et al [74]
Individual: innovation recipient (patients)			
Benefits			

XSL•FO
RenderX

CFIR domain	CFIR construct	Studies, n (%)	References
Care access	Need	13 (46)	• Lin et al [62] • Sivakumar et al [80] • Uscher-Pines et al [84] • Calandra et al [65] • Hser and Mooney [70] • O'Gurek [76] • Wang et al [85] • Hser et al [71] • Sahu et al [79] • Riedel et al [78] • Caton et al [66] • Crowley and Delargy [67] • Leo et al [75]
Travel and transportation limitations	Need	10 (36)	• Akoto [61] • Uscher-Pines et al [84] • Calandra et al [65] • Hser and Mooney [70] • O'Gurek [76] • Wang et al [85] • Sahu et al [79] • Riedel et al [78] • Caton et al [66] • Crowley and Delargy [67]
Rurality	Need	6 (21)	• Lin et al [62] • Calandra et al [65] • Hser and Mooney [70] • Hser et al [71] • Crowley and Delargy [67] • Taylor et al [82]
Concerns about stigma	Need	4 (14)	• Sivakumar et al [80] • Incze et al [73] • Hser and Mooney [70] • Wang et al [85]
Flexibility of appointments	Need	4 (14)	• Incze et al [73] • Calandra et al [65] • Wang et al [85] • Sahu et al [79]
Continuity of care	Need	3 (11)	• Lin et al [62] • Sahu et al [79] • Caton et al [66]
Convenience and efficiency	Need	2 (7)	• Sivakumar et al [80] • Uscher-Pines et al [84]
Patient preference	Need	1 (4)	• Hser and Mooney [70]
Younger patient age	Capability	1 (4)	• Lin et al [62]
White ethnicity	Other	1 (4)	• Lin et al [62]
Female sex	Other	1 (4)	• Lin et al [62]
Challenges			

XSL•FO
RenderX

CFIR domain	CFIR construct	Studies, n (%)	References
Digital divide	Capability	12 (43)	• Griffin et al [68] • Sivakumar et al [80] • Patel et al [77] • Uscher-Pines et al [84] • Huskamp et al [72] • Calandra et al [65] • Hser and Mooney [70] • O'Gurek [76] • Wang et al [85] • Hser et al [71] • Snell-Rood et al [81] • Wilson et al [86]
Mistrust of virtual care	Motivation	3 (11)	• Huskamp et al [72] • Hser and Mooney [70] • Hser et al [71]
Black ethnicity	Other	2 (7)	• Lin et al [62] • Patel et al [77]
Lack of private and secure spaces	Need	2 (7)	• Jones et al [74] • Hser et al [71]
Older patient age	Capability	1 (4)	• Wang et al [85]
Minority ethnicity	Other	1 (4)	• Wang et al [85]
Hinders patient relations	Need	1 (4)	• Beharie et al [63]
Cost	Cost	1 (4)	• Jones et al [74]
Virtual care not aligned with treatment philosophy	Need	1 (4)	• Hser et al [71]
Longer wait for appointment	Need	1 (4)	• Hser et al [71]
English as a second language	Capability	1 (4)	• Wang et al [85]

[a]No corresponding data.

[b]OAT: opioid agonist therapy.

Innovation: Virtual Care

Primary care (including but not limited to OAT initiation and management) was often delivered through telephone appointments (18/28, 64%), with fewer papers focusing on video modalities (14/28, 50%) or not specifying the type of synchronous virtual visits (9/28, 32%; Table 4). Some studies (7/28, 25%) emphasized the importance of providing multiple modality options, such as virtual and in-person appointments (ie, hybrid).

XSL•FO
RenderX

Table 4. Primary care setting, care provided, and care modality in the included studies.

Study	Primary care setting	Care provided	Care modality
Akoto [61]	Primary care clinic and outpatient treatment	OAT[a] follow-up	Video and in person
Beharie et al [63]	Primary care clinic and outpatient treatment	OAT initiation, OAT follow-up, and other	Phone and in person
Behrends et al [64]	Community health; syringe program	Nonspecific OAT, mental health, harm reduction, and primary care	Nonspecific telemedicine[b]
Calandra et al [65]	Community health and academic training center	Nonspecific OAT, mental health, and STBBI[c]	Video, phone, and in person
Caton et al [66]	Primary care clinic, rural, and community health	OAT initiation, OAT follow-up, and mental health	Video, phone, and in person
Crowley and Delargy [67]	Addiction management in primary care program	OAT initiation, OAT follow-up, and mental health	Video, phone, and in person
Griffin et al [68]	Outreach	OAT initiation, OAT follow-up, mental health, STBBI, and primary care	Phone
Hedden et al [36]	Not specified	Nonspecific OAT, mental health, STBBI, and other	Video and phone
Hodgkin et al [69]	Primary care clinic, community health, and federally qualified health centers	Nonspecific OAT	Nonspecific telemedicine
Hser and Mooney [70]	Primary care clinic and rural	Nonspecific OAT	Video and phone
Hser et al [71]	Primary care clinic and rural	Nonspecific OAT and mental health	Video, phone, and in person
Huskamp et al [72]	Primary care clinic, rural, and outpatient treatment	Nonspecific OAT and OAT initiation	Phone and in person
Incze et al [73]	Primary care clinic	Nonspecific OAT, STBBI, and primary care	Nonspecific telemedicine and in person
Jones et al [74]	Primary care clinic, rural, outpatient treatment, community health, veteran affairs, urgent care, and long-term care facilities	OAT initiation, OAT follow-up, mental health, and harm reduction	Video and phone
Lee et al [60]	Primary care clinic	OAT initiation, mental health, STBBI, and primary care	Phone and in person
Leo et al [75]	Primary care clinic, community health, and outreach	Nonspecific OAT and OAT initiation	Nonspecific telemedicine
Lin et al [62]	Veteran affairs	Nonspecific OAT, mental health, primary care, and other	Nonspecific telemedicine and in person
O'Gurek [76]	Primary care clinic and outpatient treatment	Nonspecific OAT, mental health, and primary care	Video, phone, and in person
Patel et al [77]	Not specified	OAT initiation	Nonspecific telemedicine
Riedel et al [78]	Primary care clinic	Nonspecific OAT	Video, phone, and in person
Sahu et al [79]	Primary care clinic	Nonspecific OAT	Nonspecific telemedicine and in person
Sivakumar et al [80]	Outreach and syringe program	OAT initiation, OAT follow-up, mental health, harm reduction, and STBBI	Video, phone, and in person
Snell-Rood et al [81]	Primary care clinic and rural	Nonspecific OAT and mental health	Nonspecific telemedicine
Taylor et al [82]	Community health	Nonspecific OAT	Phone
Thompson et al [83]	Primary care clinic and rural	Nonspecific OAT	Nonspecific telemedicine and in person
Uscher-Pines et al [84]	Primary care clinic and community health	OAT initiation and OAT follow-up	Video, phone, and in person
Wang et al [85]	Primary care clinic and rural	OAT initiation	Video and phone
Wilson et al [86]	Primary care clinic, rural, and residency programs	OAT initiation, OAT follow-up, and mental health	Phone and in person

[a]OAT: opioid agonist therapy.

[b]Studies that did not explicitly state the use of synchronous telemedicine for opioid use disorder care were assumed to use this method. This assumption

XSL•FO
RenderX

was based on the waiver stipulations of the Ryan Haight Act and the associated Substance Abuse and Mental Health Services Administration guidelines, which mandate the use of "real-time" (ie, audio or visual) modalities for the provision of buprenorphine [87].

[c]STBBI: sexually transmitted and blood-borne infections.

Where studies described primary care prescription of OAT, most (16/28, 57%) referred to provision generally without specific mention of the stage or length of care. Others described OAT initiation (13/28, 46%) or follow-up and stabilization (9/28, 32%) for management of OUD virtually. Outside of providing OAT, mental health care was most commonly provided (14/28, 50%), followed by treatments for sexually transmitted and blood-borne infections (6/28, 21%) [68]. In addition, harm reduction services were provided (such as syringe exchange and naloxone distribution with the support of phone consultations and home deliveries [80]), although to a lesser extent (3/28, 11%). Few studies (6/28, 21%) described other types of primary care provided (eg, on-site vaccinations [64] and in-person diabetes management [73]) and lacked specific details on the care provided (eg, it was listed as "other health needs" [63] or "other clinics" [62]).

Outer Setting: State

The most common facilitator described in the included studies was regulatory changes (11/28, 39%) to support the delivery of virtual primary care for OUD, such as the waiver of the Ryan Haight Act in the United States, which allowed clinicians to prescribe buprenorphine without the previously required first in-person appointment [77]. Another common facilitator was COVID-19 public health measures (4/28, 14%). Due to the increased COVID-19 risk to patients and clinicians associated with in-person appointments, people with OUD could use virtual primary care to adhere to social distancing measures and reduce their COVID-19 exposure risk [84]. The reduction of COVID-19 exposure risk was also described as a benefit of virtual care.

Regulatory barriers were less commonly described (7/28, 25%) but included uncertainty in the permanence of the COVID-19 regulatory changes [72] and limited flexibility in providing methadone via virtual modalities [80]. Billing structures were described as both facilitators and challenges. For example, new billing structures or fee codes introduced during the COVID-19 pandemic (5/28, 18%) facilitated virtual care delivery. However, some clinicians experienced challenges with billing (5/28, 18%), stating that payment models did not equally compensate virtual and in-person care [65] or that virtual engagement was not included as a billable service [63]. Insurance policies yielded negative perspectives (4/28, 14%), where inconsistent [71] or insufficient [69,78] coverage of virtual care and challenges with reimbursement [86] were reported. Urine drug screenings were mentioned in 39% (11/28) of the articles [60,62,63,66,70-72,74,80,85,86]. Some articles described that changes to federal mandates during the COVID-19 pandemic reduced (3/28, 11%) [66,80,86] and facilitated innovative approaches to urine drug screens (3/28, 11%) [70-72], such as take-home kits or remote viewing.

Inner Setting: Primary Care

Several studies (6/28, 21%) described clinical resources or guidelines as essential factors for facilitating virtual primary care implementation, ensuring that clinicians and patients could navigate virtual appointments. Such guidelines ranged from documents produced by the Substance Abuse and Mental Health Services Administration to guide clinics through policy changes to providing OAT virtually [70] or individual clinic protocols for virtual delivery of OUD care [76]. Other resources included the provision of communication devices to clinicians and patients (5/28, 18%) and training or education for patients to prepare for their virtual appointments (5/28, 18%). A total of 14% (4/28) of the studies noted that certain primary care clinics were better positioned to adopt virtual primary care. For example, clinics in urban or community-based settings [74] or larger clinics [77] were more likely to offer virtual OAT induction, although the reasoning behind these differences was not explained in the research.

Challenges affecting implementation at the clinic level included the need for more administrative support for physicians (4/28, 14%) and a lack of guidelines and resources (4/28, 14%) to support clinicians in providing evidence-based care (eg, lack of protocols for virtual primary care [85]). The studies also highlighted the challenges associated with specific clinical settings (4/28, 14%), noting that accessing OUD care without an in-person examination was more difficult in primary care solo practices versus specialty substance use facilities, opioid treatment programs, community clinics, and federally managed health administrations [74]. Furthermore, primary care practices in rural settings were less likely to provide virtual primary care services, possibly due to the lower demand for virtual services as COVID-19 risk was lower than in urban areas [74]. A few studies (3/28, 11%) described low demand for virtual care in general from both patients and clinicians [71,74,81], resulting in slower uptake in some clinics.

Individual Characteristics (Recipient): People With OUD

The most common benefit for patients was the high accessibility of virtual care (13/28, 46%), particularly for rural patients (6/28, 21%). This advantage was linked to other benefits, such as alleviating travel challenges (10/28, 36%) and flexibility of appointments (4/28, 14%). Factors such as confidence in virtual care (2/28, 7%), its perceived convenience and efficiency (2/28, 7%), improved continuity of care (3/28, 11%), and the ability of virtual modalities to address patients' stigma concerns (4/28, 14%) also emerged as pivotal factors. In addition, younger age (1/28, 4%) and cost savings (2/28, 7%) positively influenced patients' engagement with virtual primary care.

The gap between people who can easily use and access technology and those who cannot (ie, "the digital divide") remains the primary obstacle to the uptake of virtual care for people with OUD (12/28, 43%). This is compounded by a lack of technological knowledge or equipment (6/28, 21%), mistrust of virtual care (3/28, 11%), and absence of a private or secure space for appointments (2/28, 7%). Intersectionality between specific marginalizing patient attributes, such as being an older adult (1/28, 4%) and racialization (3/28, 11%), resulted in inequity in virtual primary care access due to structural barriers of the health care system.

XSL•FO
RenderX

Individual Characteristics (Provider): Clinicians

Clinicians found virtual care beneficial, particularly when managing stable patients (5/28, 18%) and patients who could attend appointments consistently (5/28, 18%). Clinicians also found virtual care beneficial for continuity of care (3/28, 11%). Furthermore, previous experience with OAT or using virtual modalities (3/28, 11%) and a genuine interest in delivering virtual primary care (2/28, 7%) facilitated its adoption among clinicians.

Conversely, apprehension about liability or heightened risks associated with offering virtual primary care for people with OUD (10/28, 36%) were seen as challenges among clinicians. Concerns also emerged over the adequacy of phone-based relative to video-based appointments (3/28, 11%), challenges of fostering a therapeutic relationship through virtual visits (3/28, 11%), and the belief that in-person care was necessary for patients (2/28, 7%). Some clinicians also reported a preference for in-person care (3/28, 11%). Additional barriers included technological challenges (3/28, 11%) such as low digital literacy and inadequate technology equipment.

Quality Appraisal

Over a quarter of the studies (8/28, 29%), most of which (7/8, 88%) were original research, received high scores through the QuADS tool. In contrast, commentaries (5/28, 18%) and conference abstracts (5/28, 18%) predominantly received lower scores. The scores obtained through our quality assessment using the QuADS tool are grouped by study type (Table 5).

Table 5. Quality appraisal of the included studies grouped by study type and Quality Assessment With Diverse Studies score [55].

	Low quality (score of 0-12)	Medium quality (score of 13-25)	High quality (score of 26-39)
Original research	__[a]	• Lee et al [60] • Sivakumar et al [80] • Uscher-Pines et al [84] • Snell-Rood et al [81]	• Lin et al [62] • Beharie et al [63] • Behrends et al [64] • Huskamp et al [72] • Jones et al [74] • Riedel et al [78] • Caton et al [66]
Conference abstract	• Griffin et al [68] • Incze et al [73] • Sahu et al [79] • Taylor et al [82] • Thompson et al [83]	• Akoto [61]	—
Commentary	• Calandra et al [65] • Hser and Mooney [70] • Wang et al [85] • Wilson et al [86] • Leo et al [75]	• Hser et al [71]	—
Brief report	• Crowley and Delargy [67]	• Patel et al [77] • O'Gurek [76]	—
Protocol	—	—	• Hedden et al [36]
Review	• Hodgkin et al [69]	—	—

[a]No corresponding data.

Discussion

Principal Findings

In reviewing 28 studies on the experiences of people with OUD and primary care clinicians with virtual modalities in primary care, we sought to explore the current state of the literature, identify research gaps, and provide insights into the potential benefits and challenges of virtual approaches. The CFIR constructs were instrumental in our investigation, guiding us in identifying the organizational-, structural-, and individual-level factors that influence the implementation and ongoing use of virtual primary care for people with OUD. Our findings underscore that, with careful consideration of the implementation factors at all levels, virtual modalities in primary care hold promise as a feasible, acceptable, and effective option for people with OUD as complements to in-person primary care, aligning with research in other settings [88-90].

Our review highlights that research related to virtual primary care for the management of OUD is limited and even more so for the broad range of primary care provided for people with OUD beyond OAT despite the rapid increase in the availability of virtual care resulting from the COVID-19 pandemic. The available details on the range of primary care services offered virtually were scarce, but they did include areas such as mental health care, sexual health care, and harm reduction services. For example, people with OUD are known to frequently experience comorbid conditions (eg, HIV/AIDS, hepatitis C, and hepatitis B [91]), infections (eg, endocarditis and septic arthritis [92]), and other frequent or episodic health care needs (eg, wound care [93]) that require timely and ongoing primary

XSL•FO
RenderX

care. However, the extant literature on virtual primary care for people with OUD focuses on management of OUD and not on other primary care needs, and few of the included studies (6/28, 21%) reported how or whether these conditions were addressed. This remains a core knowledge gap in the literature.

Relying on virtual modalities to increase access to health care was a common finding [94]. Virtual visits remove the necessity for physical travel to medical appointments and reduce associated costs (eg, transit or gas, taking time off work, and childcare). By offering the convenience of accessing care from a smartphone or other communication devices, virtual care can also support continuous access to and retention in OAT [95]. Patient lifestyle and stability play a significant role in driving their use of virtual care. Individuals with childcare responsibilities [79,85], limited flexibility during work hours [65,79], or those who are more stable on OAT [65,78,86] may find virtual care well-suited to meeting their care needs, while minimizing disruptions to their day-to-day lives.

The digital divide, the gap between those who can use and access technology easily and those who cannot, emerged as a central theme in multiple studies (12/28, 43%). Disparities in the components of the digital divide—such as access to technology, internet connectivity, and digital literacy—can affect the ability of people with OUD to use virtual care and its effectiveness. Furthermore, clinicians' belief that virtual primary care was unsuitable for some patients, too risky due to patients' health status, or associated with high liability also hindered its adoption, aligning with concerns found in other (ie, non–primary care) clinical settings [96,97]. With the challenges associated with patient stability and social determinants of health (eg, digital technology access, longevity of OAT treatment, and housing), providing care to people with OUD using solely virtual modalities may unintentionally limit equitable health care access. However, clinicians who believe that virtual primary care is unsuitable for unstable patients may be ignoring patient preferences [70]. While virtual primary care for less stable patients is a nuanced topic and needs concerted research efforts and practice approaches, evidence is emerging that virtual primary care can provide a safe supplement to in-person visits for unstable or new patients [72,78,98]. To enhance the effectiveness of virtual primary care for this population, offering diverse appointment types can accommodate patient preferences, socioeconomic factors, and health circumstances, fostering patient autonomy and promoting patient-centered care.

Virtual primary care for people with OUD is promising and feasible. However, this review also highlights that there may be disparities based on patient age [62,85]; language [85]; race or ethnicity [62,77,85]; sex [62]; and experience of the digital divide [65,68,70-72,76,77,80,81,84-86] due to rurality, socioeconomic status, or limited access to technology. These disparities pose challenges for implementing and accessing virtual care [99]. Without deliberate efforts to address these access barriers, virtual primary care could perpetuate inequities and worsen existing disparities in OAT access [100-102].

Limitations

There are several limitations to this scoping review. Scoping reviews inherently possess limitations in terms of depth of analysis and generalizability [39], and the possibility of missing studies needs to be acknowledged. This is attributable to multiple factors; among them is our stringent inclusion criteria, which not only resulted in excluding studies with unclear or inadequately described settings but also affected our interrater reliability score. Furthermore, our definition of virtual visits excluded asynchronous forms, and we did not consider care settings that provided only OAT, which excludes the recent growth of literature on digital services for treatment of OUD as they take place within stand-alone addiction treatment services. This criterion was in place to align with our broader research project [36] and maintain a manageable review size. We acknowledge that there are valuable lessons from virtual care models provided for other populations and conditions, which may enrich our understanding and approach in this area but were not captured in this review. In addition, studies published in languages other than English were not captured. There was also a variance in the quality of the included studies. Some of the included studies (12/28, 43%) received low QuADS scores due to limitations in study information (eg, conference abstracts with restricted word counts). Furthermore, despite the growing demand for research that distinguishes between experiences associated with sex and gender [103,104], many of the studies we reviewed appear to conflate these distinct experiences (eg, reporting participant gender as female or male) or did not report any sex- or gender-based analyses. In addition, across the studies we reviewed, the heterogeneous research methods complicated data comparison and aggregation. Most studies (25/28, 89%) were conducted in the United States, which can limit or challenge generalizability to other health system contexts. For example, regulations and billing, part of the CFIR's outer setting domain, differ dramatically across settings [105-107]. While qualified physicians in the United States and Canada can prescribe buprenorphine, the United States previously required a waiver under the Drug Addiction Treatment Act, whereas there is no such requirement in Canada [108]. Similarly, methadone treatment requires access to specialized clinics (ie, opioid treatment programs) in the United States [109], but in Canada, it can be prescribed in various health care settings [110]. Nonetheless, this scoping review presents novel results as the first of its kind covering virtual primary care for people with OUD. Our methodologically rigorous approach and the quality appraisal of the included studies make this review a valuable contribution to the literature given the sparse research in this field.

Conclusions

This scoping review highlights the promise of virtual primary care as a feasible, acceptable, and effective option for people with OUD, particularly in enhancing accessibility and continuous access to care. However, addressing the digital divide and mitigating disparities based on demographic and socioeconomic factors will be critical in realizing the full potential of virtual care. Future research and policy initiatives should prioritize efforts to bridge these gaps, ensuring that virtual primary care becomes an inclusive and equitable platform for delivering comprehensive care to and for people with OUD. In addition, there is a need for future research to explore the use of virtual modalities for primary care needs beyond OAT

XSL•FO
RenderX

management to fully understand and enhance virtual care for people with OUD.

Acknowledgments

This study is funded by a grant from the Canadian Institutes of Health Research (VR41 72756). The authors would like to thank the study team investigators for their valuable contributions and commitment to the project. Manuscript editing was supported through a large language model (ie, ChatGPT [111]) to improve clarity and conciseness [112]. The authors reviewed the generated text to ensure that it retained the original meaning.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Contributions to the paper are described using the Contributor Role Taxonomy (CRediT) taxonomy [113]. SN contributed to conceptualization, methodology, formal analysis, investigation, data curation, visualization, writing the original draft, and reviewing and editing the manuscript. EG contributed to conceptualization, methodology, formal analysis, investigation, data curation, visualization, writing the original draft, and reviewing and editing the manuscript. SS contributed to conceptualization, visualization, and reviewing and editing the manuscript. RKM contributed to conceptualization, methodology, funding acquisition, supervision, and reviewing and editing the manuscript. LH contributed to conceptualization, methodology, supervision, funding acquisition, and reviewing and editing the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 96 KB - jmir_v26i1e54015_app1.pdf]

Multimedia Appendix 2
Final search strategy.
[PDF File (Adobe PDF File), 67 KB - jmir_v26i1e54015_app2.pdf]

Multimedia Appendix 3
Data extraction template.
[PDF File (Adobe PDF File), 36 KB - jmir_v26i1e54015_app3.pdf]

Multimedia Appendix 4
Quality assessment template.
[PDF File (Adobe PDF File), 77 KB - jmir_v26i1e54015_app4.pdf]

References

1. Vadivelu N, Kai AM, Kodumudi V, Sramcik J, Kaye AD. The Opioid crisis: a comprehensive overview. Curr Pain Headache Rep 2018 Feb 23;22(3):16. [doi: 10.1007/s11916-018-0670-z] [Medline: 29476358]
2. Opioid- and stimulant-related harms in Canada. Federal, provincial, and territorial Special Advisory Committee on the Epidemic of Opioid Overdoses. Ottawa: Public Health Agency of Canada URL: https://health-infobase.canada.ca/substance-related-harms/opioids-stimulants/ [accessed 2024-04-26]
3. Ahmad FB, Cisewski JA, Rossen LM, Sutton P. Provisional drug overdose death counts. National Center for Health Statistics. URL: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm [accessed 2024-04-29]
4. United Nations Office on Drugs and Crime. World drug report 2022.: United Nations; 2022. URL: https://www.un-ilibrary.org/content/books/9789210019545 [accessed 2024-04-29]
5. Zoorob M. Fentanyl shock: the changing geography of overdose in the United States. Int J Drug Policy 2019 Aug;70:40-46. [doi: 10.1016/j.drugpo.2019.04.010] [Medline: 31079029]
6. Shoff C, Yang TC, Shaw BA. Trends in Opioid use disorder among older adults: analyzing Medicare data, 2013-2018. Am J Prev Med 2021 Jun;60(6):850-855 [FREE Full text] [doi: 10.1016/j.amepre.2021.01.010] [Medline: 33812694]
7. Institute for Health Metrics and Evaluation (IHME). Global prevalence of opioid use disorders, all ages, all sexes, 2019, 2020. Global Burden of Disease Results. 2021. URL: https://vizhub.healthdata.org/gbd-results/ [accessed 2023-08-02]

XSL•FO
RenderX

8. Santo Jr T, Clark B, Hickman M, Grebely J, Campbell G, Sordo L, et al. Association of opioid agonist treatment with all-cause mortality and specific causes of death among people with opioid dependence: a systematic review and meta-analysis. JAMA Psychiatry 2021 Sep 01;78(9):979-993 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0976] [Medline: 34076676]
9. Schwartz RP, Gryczynski J, O'Grady KE, Sharfstein JM, Warren G, Olsen Y, et al. Opioid agonist treatments and heroin overdose deaths in Baltimore, Maryland, 1995–2009. Am J Public Health 2013 May;103(5):917-922 [FREE Full text] [doi: 10.2105/ajph.2012.301049] [Medline: 23488511]
10. O'Connor PG, Oliveto A, Shi J, Triffleman E, Carroll K, Kosten T, et al. A randomized trial of buprenorphine maintenance for heroin dependence in a primary care clinic for substance users versus a methadone clinic. Am J Med 1998 Aug;105(2):100-105 [FREE Full text] [doi: 10.1016/s0002-9343(98)00194-6] [Medline: 9727815]
11. Ahamad K, Hayashi K, Nguyen P, Dobrer S, Kerr T, Schütz CG, et al. Effect of low-threshold methadone maintenance therapy for people who inject drugs on HIV incidence in Vancouver, BC, Canada: an observational cohort study. Lancet HIV 2015 Oct;2(10):e445-e450 [FREE Full text] [doi: 10.1016/S2352-3018(15)00129-0] [Medline: 26423652]
12. Tyndall M. Safer opioid distribution in response to the COVID-19 pandemic. Int J Drug Policy 2020 Sep;83:102880 [FREE Full text] [doi: 10.1016/j.drugpo.2020.102880] [Medline: 32739148]
13. MacKinnon L, Socías ME, Bardwell G. COVID-19 and overdose prevention: challenges and opportunities for clinical practice in housing settings. J Subst Abuse Treat 2020 Dec;119:108153 [FREE Full text] [doi: 10.1016/j.jsat.2020.108153] [Medline: 33032862]
14. Webster P. Virtual health care in the era of COVID-19. Lancet 2020 Apr 11;395(10231):1180-1181 [FREE Full text] [doi: 10.1016/S0140-6736(20)30818-7] [Medline: 32278374]
15. Glazier RH, Green ME, Wu FC, Frymire E, Kopp A, Kiran T. Shifts in office and virtual primary care during the early COVID-19 pandemic in Ontario, Canada. CMAJ 2021 Feb 08;193(6):E200-E210 [FREE Full text] [doi: 10.1503/cmaj.202303] [Medline: 33558406]
16. Mohammed HT, Hyseni L, Bui V, Gerritsen B, Fuller K, Sung J, et al. Exploring the use and challenges of implementing virtual visits during COVID-19 in primary care and lessons for sustained use. PLoS One 2021 Jun 24;16(6):e0253665 [FREE Full text] [doi: 10.1371/journal.pone.0253665] [Medline: 34166441]
17. Breton M, Deville-Stoetzel N, Gaboury I, Smithman MA, Kaczorowski J, Lussier MT, et al. Telehealth in primary healthcare: a portrait of its rapid implementation during the COVID-19 pandemic. Healthc Policy 2021 Aug;17(1):73-90 [FREE Full text] [doi: 10.12927/hcpol.2021.26576] [Medline: 34543178]
18. DiGennaro C, Garcia GG, Stringfellow EJ, Wakeman S, Jalali MS. Changes in characteristics of drug overdose death trends during the COVID-19 pandemic. Int J Drug Policy 2021 Dec;98:103392 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103392] [Medline: 34325184]
19. Rezaeiahari M, Fairman BJ. Impact of COVID-19 on the characteristics of opioid overdose deaths in Arkansas. Int J Drug Policy 2022 Nov;109:103836 [FREE Full text] [doi: 10.1016/j.drugpo.2022.103836] [Medline: 36027767]
20. Vieson J, Yeh AB, Lan Q, Sprague JE. During the COVID-19 pandemic, opioid overdose deaths revert to previous record levels in Ohio. J Addict Med 2022;16(2):e118-e122 [FREE Full text] [doi: 10.1097/ADM.0000000000000874] [Medline: 34172625]
21. Mason M, Arukumar P, Feinglass J. The pandemic stay-at-home order and opioid-involved overdose fatalities. JAMA 2021 Jun 22;325(24):2495-2496 [FREE Full text] [doi: 10.1001/jama.2021.6700] [Medline: 33890982]
22. Niles JK, Gudin J, Radcliff J, Kaufman HW. The Opioid epidemic within the COVID-19 pandemic: drug testing in 2020. Popul Health Manag 2021 Feb;24(S1):S43-S51 [FREE Full text] [doi: 10.1089/pop.2020.0230] [Medline: 33031013]
23. Russell C, Ali F, Nafeh F, Rehm J, LeBlanc S, Elton-Marshall T. Identifying the impacts of the COVID-19 pandemic on service access for people who use drugs (PWUD): a national qualitative study. J Subst Abuse Treat 2021 Oct;129:108374 [FREE Full text] [doi: 10.1016/j.jsat.2021.108374] [Medline: 34080545]
24. Krawczyk N, Fawole A, Yang J, Tofighi B. Early innovations in opioid use disorder treatment and harm reduction during the COVID-19 pandemic: a scoping review. Addict Sci Clin Pract 2021 Nov 13;16(1):68 [FREE Full text] [doi: 10.1186/s13722-021-00275-1] [Medline: 34774106]
25. Langabeer II JR, Yatsco A, Champagne-Langabeer T. Telehealth sustains patient engagement in OUD treatment during COVID-19. J Subst Abuse Treat 2021 Mar;122:108215 [FREE Full text] [doi: 10.1016/j.jsat.2020.108215] [Medline: 33248863]
26. Rollston R, Gallogly W, Hoffman L, Tewari E, Powers S, Clear B. Collaborative, patient-centred care model that provides tech-enabled treatment of opioid use disorder via telehealth. BMJ Innov 2022 Feb 24;8(2):117-122 [FREE Full text] [doi: 10.1136/bmjinnov-2021-000816]
27. Li E, Tsopra R, Jimenez G, Serafini A, Gusso G, Lingner H, et al. General practitioners' perceptions of using virtual primary care during the COVID-19 pandemic: an international cross-sectional survey study. PLOS Digit Health 2022 May 16;1(5):e0000029 [FREE Full text] [doi: 10.1371/journal.pdig.0000029] [Medline: 36812543]
28. Kim SH, Tesmer O. Employing telehealth strategies for opioid addiction during COVID-19: implications for social work health care. Soc Work Health Care 2021 Aug 09;60(6-7):499-508. [doi: 10.1080/00981389.2021.1953207] [Medline: 34278979]

XSL•FO
RenderX

29. Kruse CS, Lee K, Watson JB, Lobo LG, Stoppelmoor AG, Oyibo SE. Measures of effectiveness, efficiency, and quality of telemedicine in the management of alcohol abuse, addiction, and rehabilitation: systematic review. J Med Internet Res 2020 Jan 31;22(1):e13252 [FREE Full text] [doi: 10.2196/13252] [Medline: 32012048]
30. Mahmoud H, Naal H, Whaibeh E, Smith A. Telehealth-based delivery of medication-assisted treatment for opioid use disorder: a critical review of recent developments. Curr Psychiatry Rep 2022 Sep;24(9):375-386 [FREE Full text] [doi: 10.1007/s11920-022-01346-z] [Medline: 35895282]
31. Stringer KL, Langdon KJ, McKenzie M, Brockmann B, Marotta P. Leveraging COVID-19 to sustain regulatory flexibility in the treatment of opioid use disorder. J Subst Abuse Treat 2021 Apr;123:108263 [FREE Full text] [doi: 10.1016/j.jsat.2020.108263] [Medline: 33612196]
32. Wong A, Bhyat R, Srivastava S, Boissé Lomax L, Appireddy R. Patient care during the COVID-19 pandemic: use of virtual care. J Med Internet Res 2021 Jan 21;23(1):e20621 [FREE Full text] [doi: 10.2196/20621] [Medline: 33326410]
33. Rashid M. Virtual inequity: do virtual visits risk leaving some people behind? CFP News.: The College of Family Physicians of Canada; 2020 Sep 25. URL: https://www.cfp.ca/news/2020/09/25/09-25 [accessed 2024-04-29]
34. Magel T, Matzinger E, Blawatt S, Harrison S, MacDonald S, Amara S, et al. How injectable opioid agonist treatment (iOAT) care could be improved? service providers and stakeholders' perspectives. Drugs Educ Prev Policy 2023 Feb 07;31(2):200-211 [FREE Full text] [doi: 10.1080/09687637.2023.2176287]
35. Glauser W. Virtual care is here to stay, but major challenges remain. CMAJ 2020 Jul 27;192(30):E868-E869 [FREE Full text] [doi: 10.1503/cmaj.1095884] [Medline: 32719027]
36. Hedden L, McCracken RK, Spencer S, Narayan S, Gooderham E, Bach P, et al. Advancing virtual primary care for people with opioid use disorder (VPC OUD): a mixed-methods study protocol. BMJ Open 2022 Sep 26;12(9):e067608 [FREE Full text] [doi: 10.1136/bmjopen-2022-067608] [Medline: 36167365]
37. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146 [FREE Full text] [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
38. Peters MD, Godfrey C, McInerney P, Munn Z, Trico AC, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2020.
39. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
40. Gooderham E, Narayan S, McCracken R, Hedden L. People with opioid use disorder and the use of video and telemedicine in primary care: a scoping review protocol. Center for Open Science. 2022. URL: https://archive.org/details/osf-registrations-y7qug-v1 [accessed 2023-08-02]
41. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/m18-0850]
42. Beller EM, Glasziou PP, Altman DG, Hopewell S, Bastian H, Chalmers I, PRISMA for Abstracts Group. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med 2013 Apr 9;10(4):e1001419 [FREE Full text] [doi: 10.1371/journal.pmed.1001419] [Medline: 23585737]
43. Primary care. American Academy of Family Physicians. URL: https://www.aafp.org/about/policies/all/primary-care.html [accessed 2023-08-22]
44. Institute of Medicine (US) Committee on the Future of Primary Care, Donaldson MS, Yordy KD, Lohr KN, Vanselow NA. Primary Care: America's Health in a New Era. Washington, DC: National Academies Press; 1996.
45. Certification of opioid treatment programs (OTPs). Substance Abuse and Mental Health Services Administration (SAMHSA). URL: https://www.samhsa.gov/medications-substance-use-disorders/become-accredited-opioid-treatment-program [accessed 2023-08-22]
46. Veterans health administration: health programs for veterans. U.S. Department of Veterans Affairs. URL: https://www.va.gov/health/programs/index.asp [accessed 2023-08-22]
47. Syringe services programs (SSPs). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/ssp/index.html [accessed 2023-08-22]
48. Narayan S, Gooderham E, Spencer S, McCracken R, Hedden L. Virtual primary care for people with opioid use disorder: a scoping review of current strategies, benefits, and challenges (Web of Science Core Collection). searchRxiv 2023 Oct 06 [FREE Full text] [doi: 10.1079/searchRxiv.2023.00332]
49. Narayan S, Gooderham E, Spencer S, McCracken R, Hedden L. Virtual primary care for people with opioid use disorder: a scoping review of current strategies, benefits, and challenges (Medline, Ovid). searchRxiv 2023 Oct 06 [FREE Full text] [doi: 10.1079/searchRxiv.2023.00331]
50. Narayan S, Gooderham E, Spencer S, McCracken R, Hedden L. Virtual primary care for people with opioid use disorder: a scoping review of current strategies, benefits, and challenges (Embase, Ovid). searchRxiv 2023 Oct 06 Search strategy posted online [FREE Full text] [doi: 10.1079/searchrxiv.2023.00330]
51. Narayan S, Gooderham E, Spencer S, McCracken R, Hedden L. Virtual primary care for people with opioid use disorder: a scoping review of current strategies, benefits, and challenges (CINAHL). searchRxiv 2023 Oct 06 Search strategy posted online [FREE Full text] [doi: 10.1079/searchrxiv.2023.00329]

XSL•FO
RenderX

52. Narayan S, Gooderham E, Spencer S, McCracken R, Hedden L. Virtual primary care for people with opioid use disorder: a scoping review of current strategies, benefits, and challenges (Canada Commons). searchRxiv 2023 Oct 06 Search strategy posted online [FREE Full text] [doi: 10.1079/searchrxiv.2023.00333]
53. Horsley T, Dingwall O, Sampson M. Checking reference lists to find additional studies for systematic reviews. Cochrane Database Syst Rev 2011 Aug 10(8):MR000026 [FREE Full text] [doi: 10.1002/14651858.MR000026.pub2] [Medline: 21833989]
54. Better systematic review management. Covidence. URL: https://www.covidence.org [accessed 2023-09-08]
55. Harrison R, Jones B, Gardner P, Lawton R. Quality assessment with diverse studies (QuADS): an appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. BMC Health Serv Res 2021 Feb 15;21(1):144 [FREE Full text] [doi: 10.1186/s12913-021-06122-y] [Medline: 33588842]
56. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
57. Xyrichis A, Iliopoulou K, Mackintosh NJ, Bench S, Terblanche M, Philippou J, et al. Healthcare stakeholders' perceptions and experiences of factors affecting the implementation of critical care telemedicine (CCT): qualitative evidence synthesis. Cochrane Database Syst Rev 2021 Feb 18;2(2):CD012876 [FREE Full text] [doi: 10.1002/14651858.CD012876.pub2] [Medline: 33599282]
58. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
59. Rangachari P, Mushiana SS, Herbert K. A scoping review of applications of the Consolidated Framework for Implementation Research (CFIR) to telehealth service implementation initiatives. BMC Health Serv Res 2022 Nov 30;22(1):1450 [FREE Full text] [doi: 10.1186/s12913-022-08871-w] [Medline: 36447279]
60. Lee JD, Grossman E, DiRocco D, Gourevitch MN. Home buprenorphine/naloxone induction in primary care. J Gen Intern Med 2009 Feb;24(2):226-232 [FREE Full text] [doi: 10.1007/s11606-008-0866-8] [Medline: 19089508]
61. Akoto K. Comparison of office versus telemedicine video visits in suboxone maintenance [Abstract]. J Addict Med 2017;11(3):E19. [doi: 10.1097/ADM.0000000000000319]
62. Lin LA, Fortney JC, Bohnert AS, Coughlin LN, Zhang L, Piette JD. Comparing telemedicine to in-person buprenorphine treatment in U.S. veterans with opioid use disorder. J Subst Abuse Treat 2022 Feb;133:108492 [FREE Full text] [doi: 10.1016/j.jsat.2021.108492] [Medline: 34175175]
63. Beharie N, Kaplan-Dobbs M, Urmanche A, Paone D, Harocopos A. "I didn't feel like a number": the impact of nurse care managers on the provision of buprenorphine treatment in primary care settings. J Subst Abuse Treat 2022 Jan;132:108633 [FREE Full text] [doi: 10.1016/j.jsat.2021.108633] [Medline: 34688496]
64. Behrends CN, Lu X, Corry GJ, LaKosky P, Prohaska SM, Glick SN, et al. Harm reduction and health services provided by syringe services programs in 2019 and subsequent impact of COVID-19 on services in 2020. Drug Alcohol Depend 2022 Mar 01;232:109323 [FREE Full text] [doi: 10.1016/j.drugalcdep.2022.109323] [Medline: 35124386]
65. Calandra J, Shukla S, Grandy R. Addressing equity and access to care: how an academic family medicine practice manages human immunodeficiency virus, hepatitis C, and substance use disorders in rural Appalachia. J Am Coll Clin Pharm 2022 Jun 24;5(8):921-927 [FREE Full text] [doi: 10.1002/jac5.1670]
66. Caton L, Cheng H, Garneau HC, Fisher T, Harris-Mills B, Hurley B, et al. COVID-19 adaptations in the care of patients with opioid use disorder: a survey of California primary care clinics. J Gen Intern Med 2021 Apr 28;36(4):998-1005 [FREE Full text] [doi: 10.1007/s11606-020-06436-3] [Medline: 33511572]
67. Crowley D, Delargy I. A national model of remote care for assessing and providing opioid agonist treatment during the COVID-19 pandemic: a report. Harm Reduct J 2020 Jul 17;17(1):49 [FREE Full text] [doi: 10.1186/s12954-020-00394-z] [Medline: 32680520]
68. Griffin J, Waldman J, Dankanich M. Engaging people experiencing homelessness in community health worker-facilitated telehealth services in response to COVID-19 [Abstract]. J Gen Intern Med 2021;36(Suppl 1):S387 [FREE Full text] [doi: 10.1007/s11606-021-06830-5] [Medline: 34297318]
69. Hodgkin D, Horgan C, Bart G. Financial sustainability of payment models for office-based opioid treatment in outpatient clinics. Addict Sci Clin Pract 2021 Jul 05;16(1):45 [FREE Full text] [doi: 10.1186/s13722-021-00253-7] [Medline: 34225785]
70. Hser YI, Mooney LJ. Integrating telemedicine for medication treatment for opioid use disorder in rural primary care: beyond the COVID pandemic. J Rural Health 2021 Jan;37(1):246-248 [FREE Full text] [doi: 10.1111/jrh.12489] [Medline: 32579751]
71. Hser YI, Ober AJ, Dopp AR, Lin C, Osterhage KP, Clingan SE, et al. Is telemedicine the answer to rural expansion of medication treatment for opioid use disorder? Early experiences in the feasibility study phase of a national drug abuse treatment clinical trials network trial. Addict Sci Clin Pract 2021 Apr 20;16(1):24 [FREE Full text] [doi: 10.1186/s13722-021-00233-x] [Medline: 33879260]

XSL•FO
RenderX

72. Huskamp HA, Riedel L, Uscher-Pines L, Busch AB, Barnett ML, Raja P, et al. Initiating opioid use disorder medication via telemedicine during COVID-19: implications for proposed reforms to the Ryan Haight act. J Gen Intern Med 2022 Jan 28;37(1):162-167 [FREE Full text] [doi: 10.1007/s11606-021-07174-w] [Medline: 34713386]
73. Incze MA, Chen D, Galyean P, Kimball E, Zickmund SL. Examining the primary care experience of people with opioid use disorder [Abstract]. J Gen Intern Med 2022;37(Suppl 2):S144 [FREE Full text] [doi: 10.1007/S11606-022-07653-8]
74. Jones CM, Diallo MM, Vythilingam M, Schier JG, Eisenstat M, Compton WM. Characteristics and correlates of U.S. clinicians prescribing buprenorphine for opioid use disorder treatment using expanded authorities during the COVID-19 pandemic. Drug Alcohol Depend 2021 Aug 01;225:108783 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108783] [Medline: 34049102]
75. Leo P, Gastala N, Fleurimont J, Messmer S, Maes P, Richardson J, et al. A community partnership to improve access to buprenorphine in a homeless population. Ann Fam Med 2021;19(1):85 [FREE Full text] [doi: 10.1370/afm.2636] [Medline: 33431401]
76. O'Gurek DT. Designing and evaluating COVID-19 protocols for an office-based opioid treatment program in an urban underserved setting. J Am Board Fam Med 2021 Feb;34(Suppl):S136-S140 [FREE Full text] [doi: 10.3122/jabfm.2021.S1.200207] [Medline: 33622828]
77. Patel SY, Ortiz EG, Barsky BA, Huskamp HA, Busch AB, Mehrotra A. Patient and clinician characteristics associated with use of telemedicine for buprenorphine induction among Medicare beneficiaries. J Gen Intern Med 2022 Nov 29;37(14):3758-3761 [FREE Full text] [doi: 10.1007/s11606-022-07633-y] [Medline: 35488099]
78. Riedel L, Uscher-Pines L, Mehrotra A, Busch AB, Barnett ML, Raja P, et al. Use of telemedicine for opioid use disorder treatment - perceptions and experiences of opioid use disorder clinicians. Drug Alcohol Depend 2021 Nov 01;228:108999 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108999] [Medline: 34517225]
79. Sahu N, Chen PH, Shimoni NA. Telehealth to improve continuity for patients receiving buprenorphine treatment for opioid use disorder. Ann Fam Med 2022 Apr 01;20(20 Suppl 1):2936 [FREE Full text] [doi: 10.1370/afm.20.s1.2936] [Medline: 35947491]
80. Sivakumar A, Madden L, DiDomizio E, Eller A, Villanueva M, Altice FL. Treatment of Hepatitis C virus among people who inject drugs at a syringe service program during the COVID-19 response: the potential role of telehealth, medications for opioid use disorder and minimal demands on patients. Int J Drug Policy 2022 Mar;101:103570 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103570] [Medline: 34954493]
81. Snell-Rood C, Pollini RA, Willging C. Barriers to integrated medication-assisted treatment for rural patients with co-occurring disorders: the gap in managing addiction. Psychiatr Serv 2021 Aug 01;72(8):935-942 [FREE Full text] [doi: 10.1176/appi.ps.202000312] [Medline: 33530734]
82. Taylor NK, Saharan A, Grady B, Aboelata N, Brown-Johnson C, Asch SM, et al. Assessing perceptions of MOUD during COVID-19 in a black and homeless population [Abstract]. J Addict Med 2021;15(5):E4-E5. [doi: 10.1097/ADM.0000000000000902] [Medline: 34609360]
83. Thompson RA, Kizewski AL, Johnson D, Fletcher E, Clancy GL, Bunn T. Assessing barriers to serving patients with substance use disorders: a Kentucky pilot [Abstract]. J Addict Med 2021;15(5):E3-E4. [doi: 10.1097/ADM.0000000000000902] [Medline: 34609360]
84. Uscher-Pines L, Sousa J, Raja P, Mehrotra A, Barnett M, Huskamp HA. Treatment of opioid use disorder during COVID-19: experiences of clinicians transitioning to telemedicine. J Subst Abuse Treat 2020 Nov;118:108124 [FREE Full text] [doi: 10.1016/j.jsat.2020.108124] [Medline: 32893047]
85. Wang L, Weiss J, Ryan EB, Waldman J, Rubin S, Griffin JL. Telemedicine increases access to buprenorphine initiation during the COVID-19 pandemic. J Subst Abuse Treat 2021 May;124:108272 [FREE Full text] [doi: 10.1016/j.jsat.2020.108272] [Medline: 33771276]
86. Wilson CG, Ramage M, Fagan EB. A primary care response to COVID-19 for patients with an opioid use disorder. J Rural Health 2021 Jan 02;37(1):169-171 [FREE Full text] [doi: 10.1111/jrh.12438] [Medline: 32277732]
87. Telemedicine prescribing of controlled substances when the practitioner and the patient have not had a prior in-person medical evaluation. The Office of the Federal Register. 2020 Mar 31. URL: https://www.federalregister.gov/documents/2023/03/01/2023-04248/telemedicine-prescribing-of-controlled-substances-when-the-practitioner-and-the-patient-have-not-had [accessed 2024-04-29]
88. Connolly SL, Gifford AL, Miller CJ, Bauer MS, Lehmann LS, Charness ME. Provider perceptions of virtual care during the coronavirus disease 2019 pandemic: a multispecialty survey study. Med Care 2021 Jul 01;59(7):646-652 [FREE Full text] [doi: 10.1097/MLR.0000000000001562] [Medline: 34009880]
89. Singh S, Fletcher GG, Yao X, Sussman J. Virtual care in patients with cancer: a systematic review. Curr Oncol 2021 Sep 08;28(5):3488-3506 [FREE Full text] [doi: 10.3390/curroncol28050301] [Medline: 34590602]
90. Bulkes NZ, Davis K, Kay B, Riemann BC. Comparing efficacy of telehealth to in-person mental health care in intensive-treatment-seeking adults. J Psychiatr Res 2022 Jan;145:347-352 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.11.003] [Medline: 34799124]
91. Degenhardt L, Peacock A, Colledge S, Leung J, Grebely J, Vickerman P, et al. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic

XSL•FO
RenderX

review. Lancet Glob Health 2017 Dec;5(12):e1192-e1207 [FREE Full text] [doi: 10.1016/S2214-109X(17)30375-3] [Medline: 29074409]
92. Altice FL, Kamarulzaman A, Soriano VV, Schechter M, Friedland GH. Treatment of medical, psychiatric, and substance-use comorbidities in people infected with HIV who use drugs. Lancet 2010 Jul 31;376(9738):367-387 [FREE Full text] [doi: 10.1016/S0140-6736(10)60829-X] [Medline: 20650518]
93. Smith ME, Robinowitz N, Chaulk P, Johnson KE. High rates of abscesses and chronic wounds in community-recruited injection drug users and associated risk factors. J Addict Med 2015;9(2):87-93 [FREE Full text] [doi: 10.1097/ADM.0000000000000093] [Medline: 25469653]
94. Sugarman DE, Busch AB, McHugh RK, Bogunovic OJ, Trinh CD, Weiss RD, et al. Patients' perceptions of telehealth services for outpatient treatment of substance use disorders during the COVID-19 pandemic. Am J Addict 2021 Sep;30(5):445-452 [FREE Full text] [doi: 10.1111/ajad.13207] [Medline: 34405475]
95. Eibl JK, Gauthier G, Pellegrini D, Daiter J, Varenbut M, Hogenbirk JC, et al. The effectiveness of telemedicine-delivered opioid agonist therapy in a supervised clinical setting. Drug Alcohol Depend 2017 Jul 01;176:133-138 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.01.048] [Medline: 28535455]
96. McClellan MJ, Florell D, Palmer J, Kidder C. Clinician telehealth attitudes in a rural community mental health center setting. J Rural Ment Health 2020 Jan;44(1):62-73 [FREE Full text] [doi: 10.1037/rmh0000127]
97. Aronowitz SV, Engel-Rebitzer E, Dolan A, Oyekanmi K, Mandell D, Meisel Z, et al. Telehealth for opioid use disorder treatment in low-barrier clinic settings: an exploration of clinician and staff perspectives. Harm Reduct J 2021 Nov 25;18(1):119 [FREE Full text] [doi: 10.1186/s12954-021-00572-7] [Medline: 34823538]
98. Cowan KE, McKean AJ, Gentry MT, Hilty DM. Barriers to use of telepsychiatry: clinicians as gatekeepers. Mayo Clin Proc 2019 Dec;94(12):2510-2523 [FREE Full text] [doi: 10.1016/j.mayocp.2019.04.018] [Medline: 31806104]
99. Poulsen MN, Santoro W, Scotti R, Henderson C, Ruddy M, Colistra A. Implementation of telemedicine delivery of medications for opioid use disorder in Pennsylvania treatment programs during COVID-19. J Addict Med 2023;17(2):e110-e118 [FREE Full text] [doi: 10.1097/ADM.0000000000001079] [Medline: 36129690]
100. Lagisetty PA, Ross R, Bohnert A, Clay M, Maust DT. Buprenorphine treatment divide by race/ethnicity and payment. JAMA Psychiatry 2019 Sep 01;76(9):979-981 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0876] [Medline: 31066881]
101. Nguemeni Tiako MJ. Addressing racial and socioeconomic disparities in access to medications for opioid use disorder amid COVID-19. J Subst Abuse Treat 2021 Mar;122:108214 [FREE Full text] [doi: 10.1016/j.jsat.2020.108214] [Medline: 33248862]
102. Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Public Health 2021 Apr 01;42(1):463-481 [FREE Full text] [doi: 10.1146/annurev-publhealth-090519-093711] [Medline: 33798406]
103. Day S, Mason R, Lagosky S, Rochon PA. Integrating and evaluating sex and gender in health research. Health Res Policy Syst 2016 Oct 10;14(1):75 [FREE Full text] [doi: 10.1186/s12961-016-0147-7] [Medline: 27724961]
104. Gahagan J, Gray K, Whynacht A. Sex and gender matter in health research: addressing health inequities in health research reporting. Int J Equity Health 2015 Jan 31;14:12 [FREE Full text] [doi: 10.1186/s12939-015-0144-4] [Medline: 25637131]
105. Panwala V, Joudrey P, Kowalski M, Bach P, Amram O. Changes to methadone maintenance therapy in the United States, Canada, and Australia during the COVID-19 pandemic: a narrative review. J Subst Use Addict Treat 2023 Sep;152:209086 [FREE Full text] [doi: 10.1016/j.josat.2023.209086] [Medline: 37270103]
106. Ridic G, Gleason S, Ridic O. Comparisons of health care systems in the United States, Germany and Canada. Mater Sociomed 2012;24(2):112-120 [FREE Full text] [doi: 10.5455/msm.2012.24.112-120] [Medline: 23678317]
107. Jin H, Marshall BD, Degenhardt L, Strang J, Hickman M, Fiellin DA, et al. Global opioid agonist treatment: a review of clinical practices by country. Addiction 2020 Dec;115(12):2243-2254 [FREE Full text] [doi: 10.1111/add.15087] [Medline: 32289189]
108. Priest KC, Gorfinkel L, Klimas J, Jones AA, Fairbairn N, McCarty D. Comparing Canadian and United States opioid agonist therapy policies. Int J Drug Policy 2019 Dec;74:257-265 [FREE Full text] [doi: 10.1016/j.drugpo.2019.01.020] [Medline: 30765118]
109. Methadone take-home flexibilities extension guidance. Substance Abuse and Mental Health Services Administration (SAMHSA). URL: https://www.samhsa.gov/medications-substance-use-disorders/statutes-regulations-guidelines/methadone-guidance [accessed 2023-09-15]
110. Kleinman RA, Brothers TD, Danilewitz M, Bahji A. Office-based methadone prescribing for opioid use disorder: the Canadian model. J Addict Med 2022;16(5):499-504. [doi: 10.1097/ADM.0000000000000950] [Medline: 35020695]
111. ChatGPT. Open AI. URL: https://openai.com/chatgpt/overview/ [accessed 2024-04-29]
112. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res 2023;13(4):1148-1154 [FREE Full text] [Medline: 37168339]
113. Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learn Publ 2015 Apr;28(2):151-155. [doi: 10.1087/20150211]

XSL•FO
RenderX

Abbreviations

CFIR: Consolidated Framework for Implementation Research
OAT: opioid agonist therapy
OUD: opioid use disorder
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QuADS: Quality Assessment With Diverse Studies

Edited by A Mavragani; submitted 26.10.23; peer-reviewed by M Poulsen, D McCarty; comments to author 26.02.24; revised version received 11.04.24; accepted 06.10.24; published 02.12.24.

Please cite as:
Narayan S, Gooderham E, Spencer S, McCracken RK, Hedden L
Virtual Primary Care for People With Opioid Use Disorder: Scoping Review of Current Strategies, Benefits, and Challenges
J Med Internet Res 2024;26:e54015
URL: https://www.jmir.org/2024/1/e54015
doi:10.2196/54015
PMID:

©Shawna Narayan, Ellie Gooderham, Sarah Spencer, Rita K McCracken, Lindsay Hedden. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness and Acceptability of Asynchronous Digital Health in Asthma Care: Mixed Methods Systematic Review

Nazim Uzzaman[1], MBBS; Victoria Hammersley[1], PhD; Kirstie McClatchey[2,3], PhD; Jessica Sheringham[4], PhD; Diksha Singh[1], MA; GM Monsur Habib[1,5], MBBS, PhD; Hilary Pinnock[1], MBChB, MD

[1]Usher Institute, The University of Edinburgh, Edinburgh, United Kingdom

[2]Directorate of Public Health, NHS Tayside, Dundee, United Kingdom

[3]School of Medicine, University of Dundee, Dundee, United Kingdom

[4]Institute of Epidemiology and Health Care, University College London, London, United Kingdom

[5]Community Respiratory Centre, Bangladesh Primary Care Respiratory Society (BPCRS), Khulna, Bangladesh

Corresponding Author:
Hilary Pinnock, MBChB, MD
Usher Institute
The University of Edinburgh
5-7 Little France Road
The Usher Building
Edinburgh, EH16 4UX
United Kingdom
Phone: 44 01316517869
Email: hilary.pinnock@ed.ac.uk

Abstract

Background: Asynchronous digital health (eg, web-based portal, text, and email communication) can overcome practical barriers associated with in-person and remote synchronous (real-time) consultations. However, little is known about the effectiveness and acceptability of asynchronous digital health to support care for individuals with asthma (eg, asthma reviews).

Objective: We aimed to systematically review the qualitative and quantitative evidence on the role of asynchronous digital health for asthma care.

Methods: Following Cochrane methodology, we searched 6 databases (January 2001-July 2022; search update: September 2023) for quantitative, qualitative, or mixed methods studies supporting asthma care using asynchronous digital health. Screening and data extraction were duplicated. We assessed the risk of bias in the clinical outcomes of randomized controlled trials included in the meta-analysis using the revised Cochrane risk of bias tool. For the remaining studies, we evaluated the methodological quality using the Downs and Black checklist, critical appraisal skills program, and mixed methods appraisal tool for quantitative, qualitative, and mixed methods studies, respectively. We determined the confidence in the evidence using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) criteria. We conducted a meta-analysis of trial data and a thematic analysis of qualitative data.

Results: We included 30 studies (20 quantitative, 6 qualitative, and 4 mixed methods) conducted in 9 countries involving individuals with asthma, their caregivers, and health care professionals. Asynchronous digital consultations linked with other functionalities, compared to usual care, improved asthma control (standardized mean difference 0.32, 95% CI 0.02-0.63; P=.04) and reduced hospitalizations (risk ratio 0.36; 95% CI 0.14-0.94; P=.04). However, there were no significant differences in quality of life (standardized mean difference 0.16; 95% CI –0.12 to 0.43; P=.26) or emergency department visits (risk ratio 0.83; 95% CI 0.33-2.09; P=.69). Patients appreciated the convenience of asynchronous digital health, though health care professionals expressed concerns. Successful implementation necessitated an organizational approach. Integrative synthesis underscored the ease of asking questions, monitoring logs, and medication reminders as key digital functionalities.

Conclusions: Despite low confidence in evidence, asynchronous consultation supported by digital functionalities is an effective and convenient option for nonemergency asthma care. This type of consultation, well accepted by individuals with asthma and their caregivers, offers opportunities for those facing challenges with traditional synchronous consultations due to lifestyle or geographic constraints. However, efficient organizational strategies are needed to manage the associated workload.

Trial Registration: PROSPERO CRD42022344224; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=344224

XSL•FO
RenderX

International Registered Report Identifier (IRRID): RR2-10.1371/journal.pone.0281538

(J Med Internet Res 2024;26:e57708) doi:10.2196/57708

KEYWORDS

digital health; asthma; asynchronous; asthma care; effectiveness; acceptability; mixed-methods review; systematic review; barrier; remote synchronous; chronic respiratory disease; self-management; digital technology; asynchronous consultation; caregiver; PRISMA

Introduction

Asthma, with an estimated 262 million cases worldwide, is the most prevalent chronic respiratory disease [1]. While some countries have witnessed a decline in asthma-related hospitalizations and deaths [1], asthma still poses an unacceptable burden on health care systems and society at large, disrupting both work and family life [2]. An asthma review helps assess asthma control, adjust management strategies, support self-management education [2,3], and understand patients' thoughts and concerns [4]. While well-controlled individuals should be reviewed at least once a year [3], those with poor asthma control or newly diagnosed cases will require more frequent review. However, challenges such as poor attendance at asthma clinics, time constraints, limited resources, and competing agendas during consultations hinder effective reviews [5]. Given the increasing access to the internet and mobile technology [6,7], digital health emerges as an innovative approach that could improve health outcomes [8] while reducing avoidable clinic visits [9].

Digital health interventions can be delivered either synchronously (real-time interaction) or asynchronously (no real-time interaction, for example, sharing clinical information from patients through email that allows a health care professional to review the data and provide feedback later), potentially offering convenient and accessible health care [10]. Synchronous remote asthma reviews conducted via telephone have been in use for about 2 decades [11], receiving widespread acceptance, while videoconferencing experienced a substantial increase, especially during the COVID-19 pandemic [12,13]. Although asynchronous digital health can overcome the time constraints linked to in-person and synchronous remote reviews and has the potential to support care of the large number of people with asthma, it remains underresearched [14].

Existing systematic reviews have synthesized evidence on remote asthma monitoring, medication adherence, and self-management support using a broad range of digital technologies [15-20]. However, none of these reviews focused on the role of 2-way asynchronous communication between patients and health care professionals. Specifically, they have not synthesized the evidence of the effectiveness of asynchronous digital health (either as an isolated intervention or in combination with other modalities) nor explored patients' and professional stakeholders' perspectives on asynchronous consulting. This is timely given the increasing use of questionnaires promoted to support primary care practices reviewing patients with long-term conditions [21,22]. We, therefore, aimed to synthesize the quantitative and qualitative evidence to derive recommendations for policy and practice on the use of asynchronous digital health for asthma care. Our objectives were to (1) assess the effects of interventions using asynchronous digital health on clinical outcomes (ie, asthma control, quality of life, emergency department visits, and hospitalizations) compared to usual care; (2) describe the digital functionalities used for asynchronous consulting; (3) explore the views and experience of patients and health care professionals on asynchronous digital health; and (4) integrate the quantitative and qualitative synthesis to derive implications for clinical practice and policymaking.

Methods

Overview

We conducted a mixed methods systematic review following results-based convergent design [23]. We adhered to Cochrane methodology [24] and used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to report our review findings [25]. The review is registered at PROSPERO (CRD42022344224), and the protocol is published [26].

Search Strategy and Selection Criteria

We developed a search strategy involving the review authors and a senior librarian from the University of Edinburgh. We searched MEDLINE, Embase, Scopus, PsycInfo, CINAHL, and Cochrane Library (S1 in Multimedia Appendix 1) from 2001 (when internet accessibility surged with third-generation cellular technologies) [27] to July 2022. We updated the search in September 2023. We placed no language restrictions on the database search and kept the option to translate relevant quantitative studies to English [28]. However, we only considered qualitative and mixed-methods studies written in English due to the potential loss of nuance in translation [29].

After the search, all identified records were saved in EndNote 20 (Clarivate Analytics), with duplicates removed using SRA Deduplicator software [30]. In total, 2 authors (NU and MH) independently screened titles, abstracts, and full-text papers for eligibility against the criteria (S2 in Multimedia Appendix 1), using Covidence [31]. Any discrepancies were settled through team discussion (HP, VH, KM, and JS). The search results are presented in a PRISMA flow diagram.

Quantitative Outcome Measurements

Our primary (clinical) outcomes from quantitative data were asthma control or quality of life measured with any validated tools and acute attacks (eg, emergency department visits, hospitalizations, or unscheduled care) [32-34]. We were interested in between-group differences at the first follow-up assessment postintervention. Where multiple assessment tools

XSL•FO
RenderX

for an outcome were reported, we used the most frequently reported validated measure (eg, asthma control test [ACT] [32], childhood-asthma control test [C-ACT] [35]) in the meta-analysis.

Qualitative Phenomena of Interest

We included studies that explored the views and experiences of patients and professional stakeholders on asynchronous digital health (with or without other modes of communication) for asthma care.

Data Extraction and Management

NU and MH independently extracted quantitative data, and NU and DS independently extracted qualitative data from the included studies. HP checked the accuracy of data across text, tables, and meta-analyses.

Data Analysis Addressing Our Review Objectives

Textbox 1 summarizes the methods of data analysis used to address our review objectives.

Textbox 1. Data analysis for our 4 objectives.

- Effectiveness of asynchronous digital health. Clinical outcomes (asthma control, quality of life, emergency department visits, and hospitalizations) from eligible randomized control trials were pooled in the meta-analysis. Cochrane's Review Manager (RevMan 2020, version 5.4.1) was used for conducting the meta-analysis. The sample size of each study in the meta-analysis was based on the reported number of participants in the analysis as provided by the study authors. For asthma control and quality of life, standardized mean difference (for different scale metrics) with 95% CI was calculated using the inverse variance method. For emergency department visits and hospitalizations, we calculated risk ratio (RR) with 95% CI using the Mantel-Haenszel method. A *P* value less than .05 was considered statistically significant for the overall effect.
- Digital functionalities. We identified the digital functionalities used in the included quantitative and mixed methods studies and illustrated these graphically.
- Views and experiences. We used thematic synthesis to combine the findings of studies that described the views and experiences of patients and health care professionals on asynchronous digital health for asthma care, following recognized methodology [36] (see S3 in Multimedia Appendix 1).
- Integrative synthesis. We integrated the quantitative and qualitative findings following the Cochrane Handbook and presented them in a matrix [24,37], illustrating findings from the qualitative synthesis (ie, preferred digital functionalities, and implementation facilitators) aligned with intervention effects on the clinical outcomes.

Heterogeneity and Reporting Bias

We assessed statistical heterogeneity through visual inspection of forest plots, chi-square, and I^2 tests for our clinical outcome [38]. We used a random-effects model to account for heterogeneity in the study intervention, population, and settings. The limited number of studies (<10) in the meta-analysis precluded creation of a funnel plot to assess publication bias [39].

Subgroups and Sensitivity Analysis

Our *a priori* planned subgroups were age (child or adult), high- or low-income countries, asthma severity, and intensity of intervention, and we planned to include an additional subgroup if that were suggested by the qualitative synthesis. We undertook a sensitivity analysis of our meta-analysis comparing the intervention with usual care, excluding studies at high risk of bias and those that could potentially influence results due to study design or inadequacies in data reporting.

Dealing With Missing Data

We contacted the author(s) of included studies to collect any incomplete or missing data but did not perform any statistical calculation to impute missing data into the meta-analysis.

Methodological Quality Assessment

We used the revised Cochrane risk-of-bias tool [40] and assessed risk of bias for clinical outcomes within the randomized control trials (RCTs) that were included in the meta-analysis. We assessed the methodological quality of the remaining quantitative studies with diverse study designs using the Downs and Black checklist [41-43]. We used the critical appraisal skills program for qualitative studies [44,45] and mixed methods appraisal tool for mixed methods studies to appraise their methodological quality [46].

Assessment of Confidence in Evidence

Using Cochrane's GRADEpro GDT software [47], we assessed confidence in evidence for the clinical outcomes following the 5 GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) criteria (risk of bias, consistency of effect, imprecision, indirectness, and publication bias) [24]. We used the Grades of Recommendation, Assessment, Development, and Evaluation-Confidence in the Evidence from Qualitative Reviews (CERQual) approach to assess confidence in synthesized qualitative findings using the interactive summary of qualitative findings tool [48,49].

Results

Study Selection

We identified 11,034 records from 6 databases (Figure 1). After deduplication, 5662 titles and abstracts were screened, leading to the assessment of 85 full-text articles. We included 30 studies (31 unique papers) after combining 2 reports from the same study [50-80]. Out of the 1446 records screened in the search update, 4 full-text articles were assessed, but none met the eligibility criteria for inclusion in the review. The list of excluded articles and reasons is detailed in S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Study selection flow.

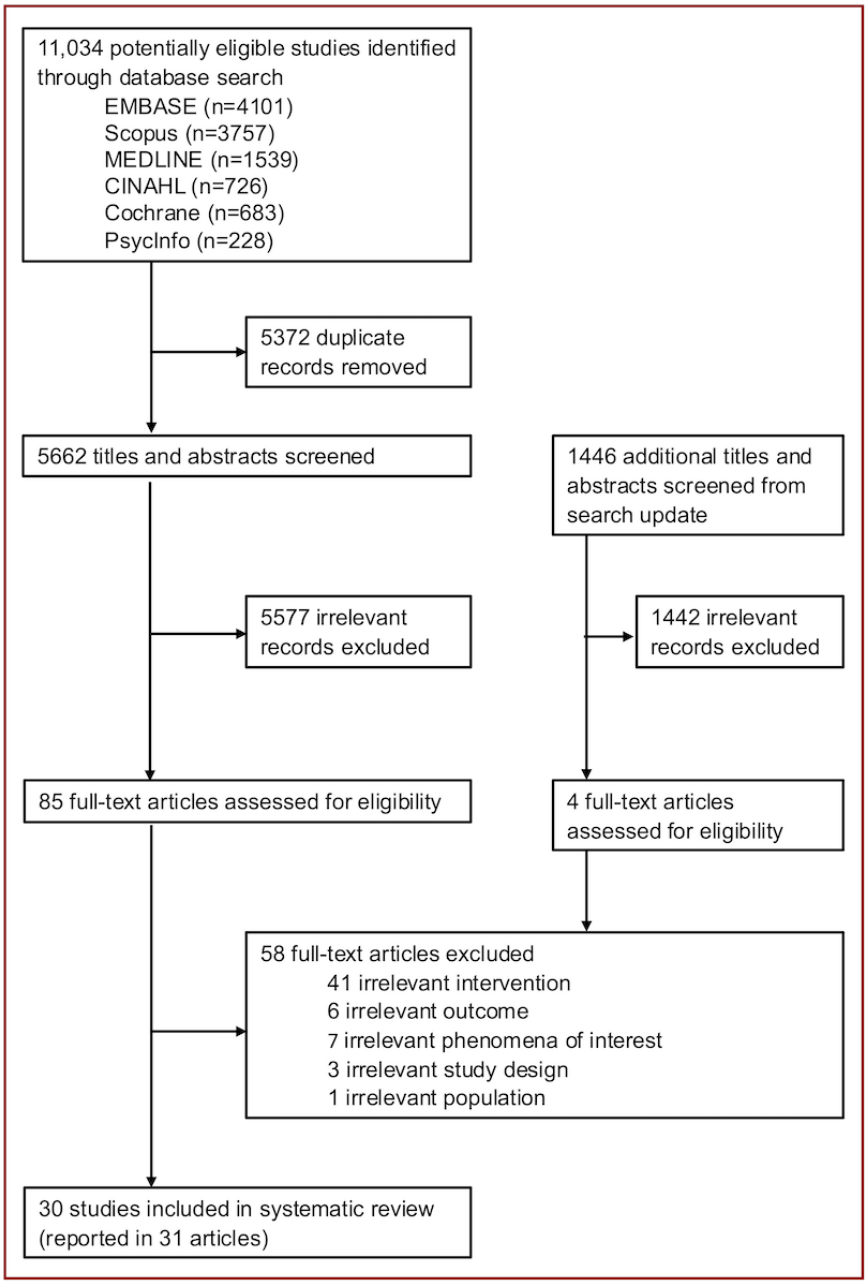

Characteristics of Included Studies

Of the 30 included studies, 20 were quantitative [50,51,53-57,60,61,64,65,67-69,72-74,76,78-80], 6 were qualitative [52,58,66,70,71,77], and 4 were mixed-methods studies [59,62,63,75]. These studies, involving individuals with asthma, their caregivers, and health care professionals, were conducted in 9 countries, predominantly from the United States (n=14) [50,52,53,55,56,58-62,66,69,70,79] and the Netherlands (n=9) [65,71-78]. The quantitative studies encompassed various designs, including RCTs (n=14) [51,53,55,57,60,61,65,67,68,72,74,78-80]. Of the RCTs, a cluster trial was reported in 2 papers [64,65], detailing the main trial outcomes in 1 paper [65], and the process evaluation in the other [64]. The summary of included studies is presented in S5 and S6 in Multimedia Appendix 1. The pool of studies varies for the different objectives, so further characteristics are described by objective in Table 1.

XSL•FO
RenderX

Table 1. Objective-wise evidence summary.

Variables	Values
Studies addressing objective 1: effectiveness of asynchronous digital health	16 studies [51,53,55,57,60,61,65,67-69,72-74,76,78,79]; 2328 participants (1217 children, 1111 adults)
Outcomes included in the meta-analysis	
Asthma control	7 studies [57,60,61,65,72,74,78]
Quality of life	5 studies [51,55,65,74,78]
Emergency department visits	5 studies [55,57,60,68,72]
Hospitalizations	4 studies [55,60,67,72]
Outcomes included in narrative synthesis	
Asthma control	5 studies [51,53,69,73,76]
Quality of life	5 studies [53,60,68,73,76]
Emergency department visits	4 studies [51,53,69,79]
Studies addressing objective 2: digital health functionalities used	21 [50,51,53-57,59-61,65,67-69,72-74,76,78-80]; 2793 participants (1614 children, 1179 adults)
Quantitative studies	
RCT[a]	14 studies [51,53,55,57,60,61,65,67,68,72,74,78-80]
Pre-post	2 studies [56,69]
Quasi-experimental	1 study [73]
Observational comparative	1 study [54]
Quantitative survey	1 study [50]
Long-term follow-up of an RCT	1 study [76]
Mixed-methods study	1 study [59]
Studies addressing objective 3: views and experiences	10 studies [52,58,59,62,63,65,70,71,75,77]; 450 participants (173 parents, 151 patients, 87 physicians, 35 nurses, and 4 pharmacists)
Qualitative studies	6 studies [52,58,66,70,71,77]
Interviews (individual or group or dyadic)	3 studies [52,58,66]
Interviews and focus groups	2 studies [70,77]
Qualitative survey	1 study [71]
Mixed methods studies	4 studies [59,62,63,75]
Sequential exploratory	2 studies [63,75]
Implementation study	1 study [59]
Mixed-methods survey	1 study [62]
Studies addressing objective 4: integrative synthesis	17 studies [51,52,55,57-61,63,65-68,71,72,74,78]; 1886 participants (1576 from quantitative, 310 from qualitative studies)
Quantitative evidence (RCTs in meta-analysis)	11 studies [51,55,57,60,61,65,67,68,72,74,78]
Qualitative evidence	6 studies [52,58,59,63,66,71]

[a]RCT: randomized controlled trial.

Effectiveness of Asynchronous Digital Health (Objective 1)

Overview

No studies reported asynchronous digital consultations as an isolated intervention. Therefore, we assessed the effectiveness of asynchronous digital consultations supported by other digital health functionalities on the clinical outcomes compared to usual care. The meta-analyses are in Figure 2 [51,55,57,60,61,65,67,68,72,74,78], and the sensitivity analysis is in S7 in Multimedia Appendix 1. The risk of bias for individual clinical outcomes is illustrated in a figure in S8 in Multimedia Appendix 1. Quantitative studies excluded from the meta-analysis due to heterogeneity in study design or data reporting are detailed in S9 in Multimedia Appendix 1 through narrative synthesis.

XSL•FO
RenderX

Figure 2. Forest plot of studies comparing the intervention with usual care on asthma control [57,60,61,65,72,74], quality of life [51,55,65,74,78], emergency department visits [55,57,60,68,72], and hospitalizations [55,60,67,72].

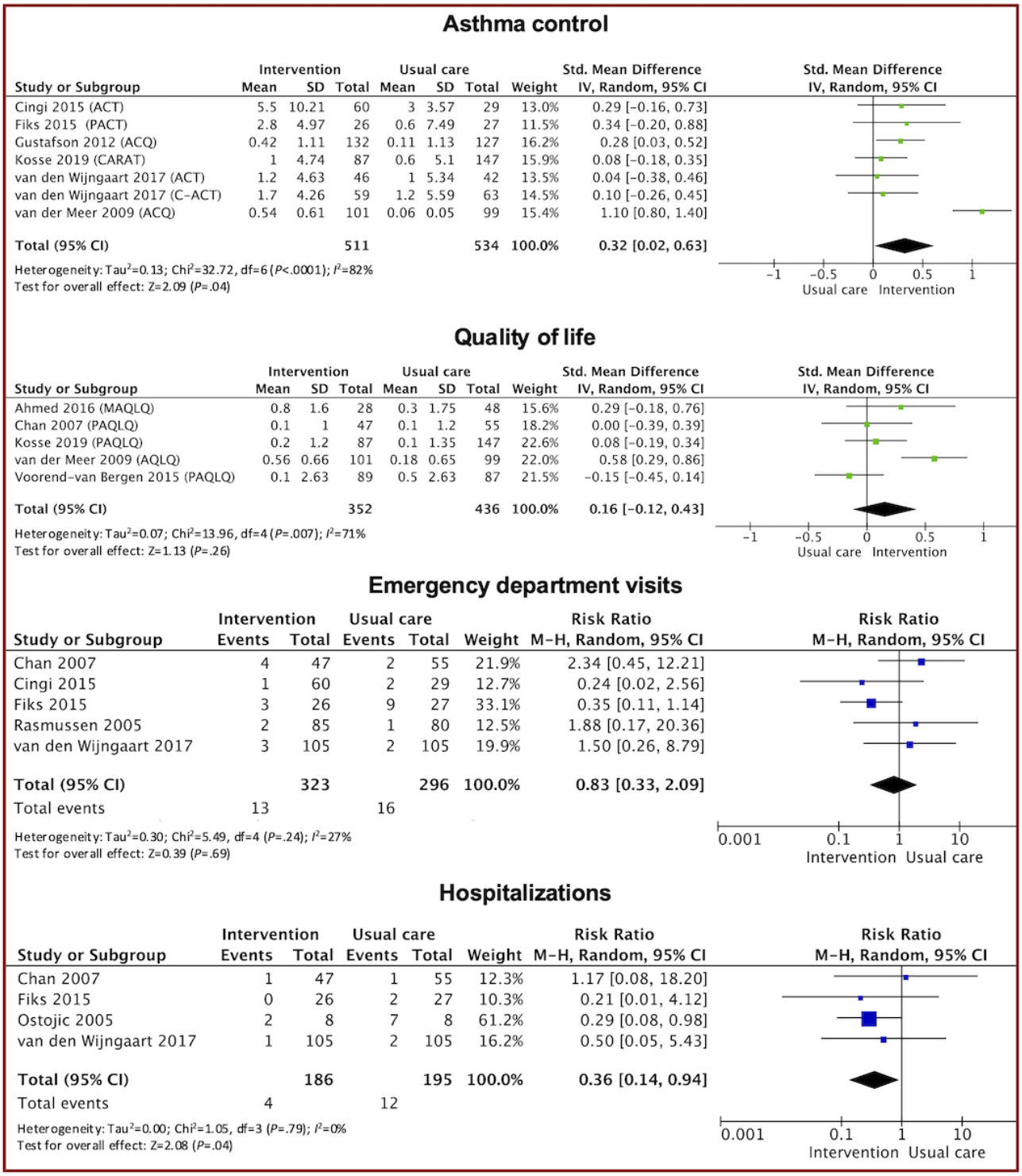

Asthma Control

We were able to retrieve data from 7 trials for meta-analysis comparing asthma control between intervention and usual care groups [57,60,61,65,72,74,78]. In total, 2 trials used both ACT and C-ACT [72,78], 1 used only ACT [57], and the others used asthma control questionnaire (ACQ) [61,74], pediatric asthma control tool [60], and control of allergic rhinitis and asthma test [65].. Of these 7 trials, we excluded a study from the meta-analysis as it reported undifferentiated ACT and C-ACT scores (which are not interchangeable) but included it in the sensitivity analysis [78]. The pooled meta-analysis revealed a statistically significant improvement in asthma control favoring the intervention group (standardized mean difference (SMD) 0.32; 95% CI 0.02-0.63; *P*=.04).

A sensitivity analysis, excluding the study at high risk of bias [57], the cluster RCT [65], and the study that induced substantial heterogeneity (I^2>50%) [74], did not change the conclusion for asthma control (SMD 0.20, 95% CI 0.03-0.37; I^2=0%; *P*=.02).

XSL•FO
RenderX

However, the *P* value edged just beyond the point of statistical significance (SMD 0.14, 95% CI −0.01 to 0.29; I^2=0%; *P*=.06) when we included the study that reported undifferentiated ACT and C-ACT scores for asthma control [78].

Quality of Life

Data from 5 trials were obtained for the meta-analysis [51,55,65,74,78]. Among these, 3 trials used pediatric asthma quality of life questionnaire (PAQLQ) [55,65,78], while the others used mini asthma quality of life questionnaire (MAQLQ) [51] and asthma quality of life questionnaire (AQLQ) [74]. The pooled estimate indicated no statistically significant difference in quality of life when comparing intervention with usual care (SMD 0.16; 95% CI –0.12 to 0.43; *P*=.26). We could not perform sensitivity analysis for quality of life due to a lack of studies following the exclusion of those that could potentially bias results [51,55,65,74].

Emergency Department Visits

In total, 5 trials were incorporated into the meta-analysis for emergency department visits [55,57,60,68,72]. The pooled estimate revealed no statistically significant difference in emergency department visits between intervention and usual care (RR 0.83; 95% CI 0.33-2.09; *P*=.69). The sensitivity analysis did not change the conclusion (RR 1.01; 95% CI 0.37-2.77; I^2=33%; *P*=.99).

Hospitalizations

In total, 4 trials were included in the meta-analysis for hospitalization [55,60,67,72]. The pooled estimate showed that the intervention group had a statistically significant reduced risk of hospitalization compared to the usual care group (RR 0.36; 95% CI 0.14-0.94; *P*=.04). However, the statistical significance of the risk of hospitalization became inconclusive (RR 0.52; 95% CI 0.11-2.42; I^2=0%; *P*=.40) after the sensitivity analysis.

Digital Health Functionalities Used (Objective 2)

In total, 21 studies documented the use of digital functionalities, with all of them incorporating either online chat (n=12) [50,53,54,57,59,60,65,72-74,76,79] and email (n=12) [51,55,56,59-61,68,69,74,76,78,80], and SMS (n=4) [67,69,74,76] for asynchronous communication between patients or their caregivers and health care professionals. The response time ranged from within 24 hours [51,54] to 48 hours [61,72] and 72 hours [78]. No studies explicitly addressed the role of non-digital support in using digital functionalities. However, in an RCT where both groups used a web portal, similar usage was observed, despite the intervention group receiving additional support through home visits by community health workers [53]. Figure 3 [50,51,53-57,59-61,65,67-69,72-74,76,78-80] illustrates the digital health functionalities used in the included studies, with additional details available in S10 in Multimedia Appendix 1.

Figure 3. Digital functionalities reported in 21 studies [50,51,53-57,59-61,65,67-69,72-74,76,78-80].

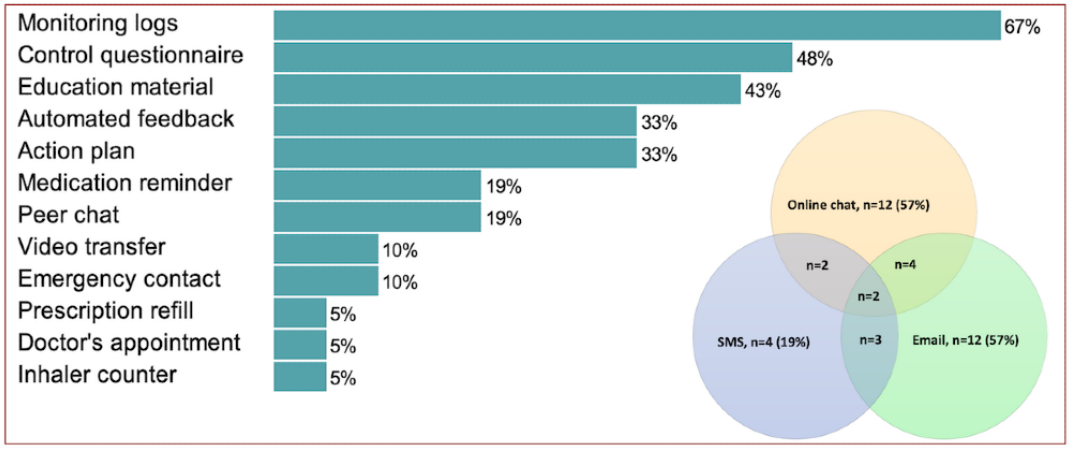

Views and Experiences of Patients and Health Care Professionals (Objective 3)

The thematic synthesis yielded four themes: (1) acceptability of routine asthma reviews, (2) advantages and disadvantages of asynchronous digital health, (3) implementation barriers and facilitators, and (4) preferred digital functionalities.

Theme 1: Acceptability of Routine Asthma Reviews

In total, 3 studies explored the views of patients or their carers and health care professionals on routine asthma reviews [63,66,75]. Most adolescents with well-controlled asthma perceived in-person routine reviews as unnecessary and preferred seeking medical help only when their symptoms worsened [75]. A few individuals with poorly controlled asthma were bothered by regular face-to-face medical reviews, mentioning that they learned to live with their symptoms. Both groups were enthusiastic about internet-based reviews and found email communication and electronic consultation useful [75].

Most patients or caregivers found asynchronous digital health feasible and accessible [63,66,75]; however, none wanted to completely replace traditional face-to-face consultations [66]. Health care professionals echoed convenience but noted the clinical limitations of remote reviews [63].

Theme 2: Advantages and Disadvantages of Asynchronous Digital Health

In total, 6 studies reported the benefits of asynchronous digital health [52,62,66,70,71,75]. Most parents of children with asthma found asynchronous digital health to be a convenient method

XSL•FO
RenderX

for reviewing and managing their child's asthma while managing work and family responsibilities [52,66].

Parents felt that asynchronous digital health improved day-to-day care management and reduced unplanned provider visits. They expressed a sense of reassurance in being able to monitor their child's condition, receive remote guidance from health care professionals, and support self-management [66].

Health care professionals viewed text messaging and emails as effective means to connect with patients, enhancing communication, and fostering compliance and treatment adherence [62,70].

In total, 4 studies reported limitations of asynchronous digital health [59,66,70,71]. Health care professionals perceived that asynchronous digital health was not suitable for patients with poor perception of their asthma symptoms or needing emergency care [70]. They also expressed concerns that the lack of physical examination and reduced face-to-face contact associated with this approach could negatively affect clinical decision-making [71]. Some parents and children with asthma found frequent monitoring data sharing to be burdensome [59,66].

Theme 3: Implementation Barriers and Facilitators

In total, 6 studies reported the barriers to implementing asynchronous digital health in routine practice [59,66,70,71,75,77]. The major barriers, as perceived by most health care professionals, included lack of integration with electronic health records (EHRs) so that data have to be transferred manually [59,71,77], poorly coordinated workflow [59,71,77], high workload [70,71,77], and the absence of financial reimbursement [59,70,77].

Some health care professionals also mentioned that a lack of internet access (eg, because of rurality) and the unavailability of electronic devices or the lack of skills using them by some patients acted as barriers [59,71,77]. Other barriers include data security [70,71,77], negative attitudes (not convinced of the added value of digital health in daily practice) [71,75,77], missing interpersonal relationships [71,77] and a lack of training for practices [59].

In total, 4 studies reported implementation facilitators [52,59,66,71]. Both health care professionals and parents of children with asthma recognized that accessible “easy” 2-way communication, prompt responsiveness from practices and families, as well as a positive attitude and commitment towards adopting asynchronous digital health were facilitators [52,59,71].

Health care professionals also emphasized the importance of efficient task allocation among the practice staff in facilitating successful implementation [59,71].

Other facilitators included user-friendliness of the digital system [66,71], shared decision-making, positive attitude or commitment, and freeing up time for complex patients [71].

Theme 4: Preferred Digital Functionalities

In total, 5 studies reported preferred digital functionalities [52,58,63,66,71]. Most patients wanted the flexibility to ask quick questions [63], receive tailored information about asthma [63,66,71], log and visualize the trend of their peak flows, symptom scores, and medication usage [63], and receive medication reminders [52,58]. Health care professionals acknowledged the importance of a flexible approach to patient reviews but expressed concerns about available resources [63]. They also thought that logs (especially symptoms, but also peak flows) could engage patients in “thoughtfully” reporting their asthma status and would inform the assessment of control and management strategies.

In contrast to the priority attached by professionals to logging, most patients acknowledged that in reality, they checked their peak flow “rarely” or only when their asthma was getting worse [63]. The reasons for not measuring every day were varied. Some forgot, while others felt uncomfortable doing it in front of people, but many suggested it was unnecessary as they knew their asthma and could assess status by how they felt.

Most parents reported checking their emails “frequently,” but this ranged from once a week to several times a day. Some participants preferred to be contacted by SMS text message or smartphone application. Others preferred a phone call or stated they were “okay” with any method of contact (text, application, email, and call) [52]. Some participants highlighted that too frequent messaging was counterproductive [58].

Integration of Qualitative and Quantitative Synthesis (Objective 4)

Implications from qualitative synthesis (preferred digital functionalities and implementation facilitators) were juxtaposed with trial findings for clinical outcomes and integrated to see if those implications had a role in improving clinical outcomes (Table 2). Out of the 6 RCTs [57,60,61,65,72,74] that showed positive improvement in asthma control, all had monitoring logs [57,60,61,65,72,74], 5 had ease of asking quick questions [57,60,65,72,74], 5 provided tailored asthma information [60,61,65,72,74], 3 had medication reminders [57,61,65], 3 had digital action plans [60,61,72], 1 was linked with the EHR, and 4 had an organized workflow with a specific person designated to respond to patients’ queries [61,65,72,74]. One study in which asthma control deteriorated had monitoring logs only [78]. Three RCTs that reported improvement in quality of life, all had ease of asking quick questions, monitoring logs, tailored asthma information, and organized workflow [51,65,74]. One RCT that reported no change in quality of life also had these features [55]. One study, where quality of life deteriorated, had monitoring logs only [78]. Of 2 of the 5 RCTs [55,57,60,68,72] that reported reduced risk of emergency department visits, 1 had medication reminders [57], and the other was linked with the EHR [60]. Of the 3 RCTs [55,68,72] that reported increased risk of emergency department visits, none had medication reminders nor were linked with EHR. However, 2 out of these 3 RCTs had ease of asking quick questions, monitoring logs, tailored asthma information, and organized workflow [55,72]. There was no specific pattern to differentiate the trials that reported increased risk of hospitalization [55] compared with those that reported reduced risk of hospitalization [60,67,72]. Overall, the ease of asking quick questions, medication reminders, tailored asthma information, and organized workflow emerged as important factors that might positively affect the intervention outcomes.

XSL•FO
RenderX

Table 2. Integrative synthesis.

Study	Implications from qualitative synthesis							Trial outcomes from meta-analyses			
	Preferred functionalities					Facilitators		Asthma control	HRQoL[a]	ED visits[b]	Admissions
	Quick Q[c]	Logs[d]	Reminder[e]	PAAP[f]	Info[g]	EHR link[h]	Responder[i]	Measure: MD (95% CI)[j]	Measure: MD (95% CI)	RR (95% CI)[k]	RR (95% CI)
Ahmed et al (2016) [51]	✓	✓	x	✓	✓	✓	Nurse	—	MAQLQ[l]: 0.50 (–0.27 to 1.27)	—	—
Chan et al (2007) [55]	✓	✓	x	x	✓	x	Nurse	—	PAQLQ[m]: 0.00 (–0.43 to 0.43)	2.34 (0.45 to 12.21)	1.17 (0.08 to 18.20)
Cingi et al (2015) [57]	✓	✓	✓	x	x	x	NR[n]	ACT[o]: 2.50 (–0.39 to 5.39)	—	0.24 (0.02 to 2.56)	—
Fiks et al (2015) [60]	✓	✓	x	✓	✓	✓	NR	PACT[p]: 2.20 (–1.21 to 5.61)	—	0.35 (0.11 to 1.14)	0.21 (0.01 to 4.12)
Gustafson et al (2012) [61]	x	✓	✓	✓	✓	x	Study team	ACQ[q]: –0.31 (–0.58 to –0.04)	—	—	—
Kosse et al (2019) [65]	✓	✓	✓	x	✓	x	Pharmacist	CARAT[r]: 0.40 (–0.89 to 1.69)	PAQLQ: 0.10 (–0.23 to 0.43)	—	—
Ostojic et al (2005) [67]	x	✓	✓	x	✓	x	Asthma specialist	—	—	—	0.29 (0.08 to 0.98)
Rasmussen et al (2005) [68]	x	✓	x	✓	x	x	NR	—	—	1.88 (0.17 to 20.36)	—
van den Wijngaart et al (2017) [72]	✓	✓	x	✓	✓	x	Asthma team	ACT: 0.20 (–1.90 to 2.30); C–ACT[s]: 0.50 (–1.26 to 2.26)	—	1.50 (0.26 to 8.79)	0.50 (0.05 to 5.43)
van der Meer et al (2009) [74]	✓	✓	x	x	✓	x	Asthma nurse	ACQ: –0.48 (–0.60 to –0.36)	AQLQ[t]: 0.38 (0.20 to 0.56)	—	—
Voorend-van Bergen et al (2015) [78]	x	✓	x	x	x	x	Nurse	ACT or C–ACT: –0.20 (–1.72 to 1.32)	PAQLQ: –0.40 (–1.18 to 0.38)	—	—

[a]HRQoL: health-related quality of Life.
[b]ED visits: emergency department visits.
[c]Quick Q: ease of asking quick questions.
[d]Logs: monitoring logs.
[e]Reminder: medication reminder.
[f]PAAP: personalized asthma action plan.
[g]Info: tailored asthma information.
[h]EHR link: linked with electronic health record.
[i]Responder: assigned responder to patient queries.
[j]MD (95% CI): mean difference (95% CI).
[k]RR (95% CI): risk ratio (95% CI).
[l]MAQLQ: mini asthma quality of life questionnaire.
[m]PAQLQ: pediatric asthma quality of life questionnaire.
[n]NR: not reported.
[o]ACT: asthma control test.
[p]PACT: pediatric asthma control tool.
[q]ACQ: asthma control questionnaire.

XSL•FO
RenderX

[r]CARAT: control of allergic rhinitis and asthma test.

[s]C-ACT: childhood asthma control test.

[t]AQLQ: asthma quality of life questionnaire.

Methodological Quality

The Downs and Black scores ranged from 6 to 24 for the other quantitative studies that were not included in the meta-analysis. Methodological quality was good in 3 studies (range: 20-25) [53,79,80], fair (range: 15-19) in 2 studies [69,76] and poor (≤14) in the other 4 studies [50,54,56,73]. Among qualitative studies, 1 had high quality [77], while 5 raised some concerns [52,58,66,70,71]. Of the 4 mixed-method studies, 1 demonstrated high quality [63], and 3 raised some concerns [59,62,75]. Methodological quality assessments for qualitative and mixed-methods studies are detailed in S11 and S12 in Multimedia Appendix 1.

Confidence in Evidence

Using the GRADE approach, the overall certainty of evidence for the clinical outcomes was judged as very low for asthma control and low for quality of life, emergency department visits, and hospitalization. Downgrading of the certainty was mostly due to the risk of bias and imprecision (S13 in Multimedia Appendix 1). The GRADE-CERQual assessment of confidence revealed that summary review findings of the qualitative studies ranged from low to high quality. Perspectives of routine reviews, advantages and disadvantages of using asynchronous digital health, implementation facilitators, and preferred digital functionalities had moderate confidence. Limitations of using asynchronous digital health revealed low confidence, and implementation barriers revealed high confidence (S14 in Multimedia Appendix 1).

Discussion

Summary of Findings

Our systematic review identified 30 studies involving people with asthma (and their caregivers) and health care professionals from 9 countries. Overall, we identified a statistically significant improvement in asthma control and a reduced risk of hospitalization when comparing interventions using asynchronous digital health with usual care. However, no statistically significant differences were observed in quality of life and emergency department visits. Certainty of evidence was very low for asthma control and low for quality of life, emergency department visits, and hospitalizations. Patients liked the convenience of asynchronous digital health, while health care professionals had some reservations. Effective implementation requires an organizational approach. Integrative synthesis highlighted the importance of functionalities such as ease of asking quick questions, monitoring logs, and medication reminders.

Strength and Limitations

A strength of this review is its inclusive design, encompassing various study types and integrating quantitative and qualitative findings. We conducted a comprehensive database search, aided by a senior librarian, and the meta-analysis was reviewed by a senior statistician from the University of Edinburgh. Heterogeneity was a concern for some clinical outcomes in the meta-analysis, but our sensitivity analysis increased the reliability of the pooled estimates. We maintained an openness to non-English papers in quantitative studies and included 1 in Russian [54]. We did not deviate from the methodology outlined in the published protocol [26]. While our GRADE assessment reflected low to very low confidence in clinical outcomes, this was influenced by self-reported outcome measures and diverse tools for asthma control and quality of life. This underscores the need for standardized outcomes in trials [81]. In total, 2 independent reviewers screened titles, abstracts, and full texts, resolving discrepancies through discussion with other authors as needed. While data extraction and analysis were duplicated (NU and MH/DS), the precision of the data was verified by another author (HP). We were aware of the potential impact of reflexivity while analyzing and interpreting the qualitative data. Nonetheless, the involvement of a multidisciplinary author group proficient in qualitative evidence synthesis helped ensure a balanced interpretation of the data.

Interpretation in the Light of Published Literature

Effectiveness of Asynchronous Digital Health

Our study indicates that asynchronous digital health can effectively complement or replace other consultation approaches for asthma care in diverse settings. This extends the findings of a recent Cochrane review on digital interventions specifically for improving asthma treatment adherence [15]. All studies included in the meta-analysis demonstrated improved asthma control despite substantial heterogeneity. Our sensitivity analysis confirmed that the heterogeneity was likely attributable to 1 study, possibly because the reported values were derived from model estimates [74]. Another study, not included in the meta-analysis but considered in the sensitivity analysis, revealed a decline in asthma control [78]. This incongruity could be linked to the study's combined reporting of ACT and C-ACT measures despite the known differences in their performance [78].

Our study showed no significant difference in quality of life compared to usual care. In contrast, a scoping review focused on a different health condition using asynchronous digital health demonstrated improved self-efficacy and quality of life [82]. Similarly, we found no significant change in emergency department visits, consistent with a systematic review indicating that health care utilization, specifically physician visits, did not decrease significantly. However, this review did observe a decline in visits among back pain and asthma patients, although these differences were not statistically significant [83]. Our study indicated a significant reduction in hospitalization risk, consistent with other reviews on conditions like rheumatoid arthritis, diabetes, and skin diseases [82-84].

XSL•FO
RenderX

Views and Experiences of Patients and Health Care Professionals

In our review, asynchronous digital health was favored by most parents of children with asthma and some adult patients, offering convenience for asthma management alongside work and family commitments. They valued remote monitoring, guidance from health care professionals, and self-management support. A study involving individuals with diabetes highlighted the benefits of asynchronous communication that complemented clinic visits and boosted patients' sense of responsibility for managing their conditions [85]. Health care professionals in our study believed asynchronous digital health could enhance communication and improve compliance and treatment adherence. Although they acknowledged convenience, they noted that remote reviews might not always be clinically suitable. Our findings align with a Cochrane review indicating that while 2-way text-based communication strengthens patient-provider relations [86], face-to-face consultations remain vital for certain cases, such as patients with complex symptoms or new cases [87]. The potential for misunderstandings in written digital communication was also recognized by clinicians, particularly for diabetes and young individuals with long-term conditions [88,89]. Similar sentiments were echoed in another study, emphasizing that asynchronous communication is not suitable for everyone and everything [90]. Therefore, it is important to align priorities between patients and health care professionals to establish a therapeutic partnership in digital health [91]. In our study, health care professionals perceived barriers to implementation, including workflow issues, lack of EHR integration, high workload, and absence of financial reimbursement. These barriers and challenges are consistent with other qualitative studies [87,92].

Process of Care Outcomes

A systematic review reported that several publications, mostly on the use of telehealth in dermatology and some that assessed multiple medical specialties, reported a positive impact on process of care outcomes, including shorter wait times and less time to perform a consultation [84]. Most asynchronous digital health cases had a total turnaround time of less than 72 hours, which is similar to our review finding [93].

Implications for Clinical Practice and Policy Making

This systematic review gives confidence that asynchronous digital health can be an effective adjunct to other modes of consultation for asthma care. This could be especially valuable for patients and health care professionals with existing familiarity with digital technologies. Some health care systems have introduced guidelines, such as requiring at least 1 in-person visit before an e-visit, to establish patient-clinician relationships [94]. Successful integration of asynchronous digital health into routine practice must be designed to benefit patients, health care professions and organizations. Efficient task allocation and organization of care are crucial implementation facilitators. User-friendly digital features are essential, including tailored functionalities for 2-way asynchronous communication within set time frames, enabling questionnaire completion, and uploading images and videos (eg, inhaler technique) as necessary. Thus, patients, health care professionals and organizations can harness the benefit of asynchronous digital health for asthma care.

Future Research

Further research is necessary to understand the organizational context and track the process of arranging, deploying and sustaining asynchronous consultations for asthma care. Moreover, there is a need for research to establish the best practices for implementing asynchronous digital health, engaging patient and public involvement groups, preventing misuse, and assessing the suitability of individuals with asthma or their caregivers and health care professionals across various health care settings. Identifying a low-risk group of individuals who could be managed appropriately through asynchronous digital consultations is another critical aspect requiring investigation.

Conclusions

Our review concludes with low confidence that asynchronous consultation supported by digital functionalities is effective and convenient and can be considered as an option for non-emergency asthma care, showing good acceptability among individuals with asthma and their caregivers. Thus, asynchronous digital consultations present an opportunity for those whose lifestyles or geographical locations impede synchronous consultations but require organization strategies to manage workload.

Acknowledgments

NU is supported by a University of Edinburgh College of Medicine PhD Studentship within the Asthma UK Centre for Applied Research (AUKCAR), and JS is supported by the National Institute for Health Research Applied Research Collaboration (ARC) North Thames. This report is independent research supported by the National Institute for Health and Care Research ARC North Thames. The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. We are thankful to Marshall Dozier, College Lead for Library Academic Support at the University of Edinburgh, for her valuable contribution to developing the search strategy. We express our gratitude to Richard Parker, Statistician and Senior Research Fellow at the Edinburgh Clinical Trials Unit, University of Edinburgh, for his insightful review of the meta-analysis and sensitivity analysis.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

NU, HP, and VH conceptualized the study. NU designed the study with support from HP, VH, KM, and JS. Study selection was done by NU and MH, while quantitative and qualitative data extraction was independently performed by NU and MH, and NU and DS, respectively. NU performed data analysis with support from MH and DS for the quantitative and qualitative analysis, respectively. HP ensured data accuracy across text, tables, and figures. NU, HP, and KM assessed methodological quality and confidence in evidence. NU drafted the manuscript, and all authors provided critical reviews for intellectually significant content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[PDF File (Adobe PDF File), 1537 KB - jmir_v26i1e57708_app1.pdf]

Multimedia Appendix 2
PRISMA checklist.
[PDF File (Adobe PDF File), 200 KB - jmir_v26i1e57708_app2.pdf]

References

1. GBD 2019 Chronic Respiratory Diseases Collaborators. Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019. EClinicalMedicine 2023 May;59:101936 [FREE Full text] [doi: 10.1016/j.eclinm.2023.101936] [Medline: 37229504]
2. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention, 2023. URL: https://ginasthma.org [accessed 2023-07-01]
3. British guideline on the management of asthma. URL: https://www.brit-thoracic.org.uk [accessed 2019-07-01]
4. Pinnock H, Fletcher M, Holmes S, Keeley D, Leyshon J, Price D, et al. Setting the standard for routine asthma consultations: a discussion of the aims, process and outcomes of reviewing people with asthma in primary care. Prim Care Respir J 2010;19(1):75-83. [doi: 10.4104/pcrj.2010.00006] [Medline: 20119630]
5. Morrow S, Daines L, Wiener-Ogilvie S, Steed L, McKee L, Caress A, et al. Exploring the perspectives of clinical professionals and support staff on implementing supported self-management for asthma in UK general practice: an IMPART qualitative study. NPJ Prim Care Respir Med 2017;27(1):45 [FREE Full text] [doi: 10.1038/s41533-017-0041-y] [Medline: 28720751]
6. How many smartphones are in the world?. URL: https://www.bankmycell.com/blog/how-many-phones-are-in-theworld [accessed 2023-12-25]
7. Digital around the world. URL: https://datareportal.com/global-digital-overview [accessed 2023-12-25]
8. Beasley R, Hancox RJ. Reducing the burden of asthma: time to set research and clinical priorities. Lancet Respir Med 2020;8(10):943-944. [doi: 10.1016/S2213-2600(20)30400-8] [Medline: 32910896]
9. Kew KM, Cates CJ. Remote versus face-to-face check-ups for asthma. Cochrane Database Syst Rev 2016;4(4):CD011715. [doi: 10.1002/14651858.CD011715.pub2] [Medline: 27087257]
10. Mechanic OJ, Persaud Y, Kimball AB. Telehealth Systems. Treasure Island (FL): StatPearls Publishing LLC; 2022.
11. Pinnock H, Bawden R, Proctor S, Wolfe S, Scullion J, Price D, et al. Accessibility, acceptability, and effectiveness in primary care of routine telephone review of asthma: pragmatic, randomised controlled trial. Br Med J 2003;326(7387):477-479. [doi: 10.1136/bmj.326.7387.477] [Medline: 12609944]
12. Pinnock H, Murphie P, Vogiatzis I, Poberezhets V. Telemedicine and virtual respiratory care in the era of COVID-19. ERJ Open Res 2022;8(3):00111-02022. [doi: 10.1183/23120541.00111-2022] [Medline: 35891622]
13. Kinley E, Skene I, Steed E, Pinnock H, McClatchey K. Delivery of supported self-management in remote asthma reviews: a systematic rapid realist review. Health Expect 2022;25(4):1200-1214. [doi: 10.1111/hex.13441] [Medline: 35411670]
14. Murray E, Burns J, See TS, Lai R, Nazareth I. Interactive health communication applications for people with chronic disease. Cochrane Database Syst Rev 2005(4):CD004274. [doi: 10.1002/14651858.CD004274.pub4] [Medline: 16235356]
15. Chan A, De Simoni A, Wileman V, Holliday L, Newby C, Chisari C, et al. Digital interventions to improve adherence to maintenance medication in asthma. Cochrane Database Syst Rev 2022;6(6):CD013030. [doi: 10.1002/14651858.CD013030.pub2] [Medline: 35691614]
16. Mosnaim G, Safioti G, Brown R, DePietro M, Szefler SJ, Lang DM, et al. Digital health technology in asthma: a comprehensive scoping review. J Allergy Clin Immunol Pract 2021;9(6):2377-2398. [doi: 10.1016/j.jaip.2021.02.028] [Medline: 33652136]

XSL•FO
RenderX

17. Chongmelaxme B, Lee S, Dhippayom T, Saokaew S, Chaiyakunapruk N, Dilokthornsakul P. The effects of telemedicine on asthma control and patients' quality of life in adults: a systematic review and meta-analysis. J Allergy Clin Immunol Pract 2019;7(1):199-216.e11. [doi: 10.1016/j.jaip.2018.07.015] [Medline: 30055283]
18. Jeminiwa R, Hohmann L, Qian J, Garza K, Hansen R, Fox BI. Impact of eHealth on medication adherence among patients with asthma: a systematic review and meta-analysis. Respir Med 2019;149:59-68. [doi: 10.1016/j.rmed.2019.02.011] [Medline: 30803887]
19. Kew KM, Cates CJ. Home telemonitoring and remote feedback between clinic visits for asthma. Cochrane Database Syst Rev 2016;2016(8):CD011714. [doi: 10.1002/14651858.CD011714.pub2] [Medline: 27486836]
20. McLean S, Chandler D, Nurmatov U, Liu J, Pagliari C, Car J, et al. Telehealthcare for asthma. Cochrane Database Syst Rev 2010;2010(10):CD007717 [FREE Full text] [doi: 10.1002/14651858.CD007717.pub2] [Medline: 20927763]
21. Accurx: The simple way to communicate about patient care. URL: https://www.accurx.com/ [accessed 2023-11-24]
22. Ardens Healthcare Informatics: eConsultations. URL: https://www.ardens.org.uk/ccg-benefits/#consultations [accessed 2023-11-24]
23. Noyes J, Booth A, Moore G, Flemming K, Tunçalp Ö, Shakibazadeh E. Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. BMJ Glob Health 2019;4(Suppl 1):e000893. [doi: 10.1136/bmjgh-2018-000893] [Medline: 30775016]
24. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.3. Cochrane, 2022. URL: https://www.training.cochrane.org/handbook [accessed 2023-12-25]
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Br Med J 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Uzzaman MN, Hammersley V, McClatchey K, Sheringham J, Habib GMM, Pinnock H. Asynchronous digital health interventions for reviewing asthma: a mixed-methods systematic review protocol. PLoS One 2023;18(2):e0281538. [doi: 10.1371/journal.pone.0281538] [Medline: 36758039]
27. Wikipedia 3G History. URL: https://en.wikipedia.org/wiki/3G [accessed 2023-08-12]
28. Fung ICH. Citation of non-English peer review publications--some Chinese examples. Emerg Themes Epidemiol 2008;5:12. [doi: 10.1186/1742-7622-5-12] [Medline: 18826566]
29. van Nes F, Abma T, Jonsson H, Deeg D. Language differences in qualitative research: is meaning lost in translation? Eur J Ageing 2010;7(4):313-316. [doi: 10.1007/s10433-010-0168-y] [Medline: 21212820]
30. Systematic Review Accelerator-Deduplicator. URL: https://sr-accelerator.com/#/deduplicator [accessed 2023-08-12]
31. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. URL: https://www.covidence.org [accessed 2022-07-05]
32. Nathan RA, Sorkness CA, Kosinski M, Schatz M, Li JT, Marcus P, et al. Development of the asthma control test: a survey for assessing asthma control. J Allergy Clin Immunol 2004;113(1):59-65. [doi: 10.1016/j.jaci.2003.09.008] [Medline: 14713908]
33. Juniper EF, Guyatt GH, Cox FM, Ferrie PJ, King DR. Development and validation of the mini asthma quality of life questionnaire. Eur Respir J 1999;14(1):32-38. [doi: 10.1034/j.1399-3003.1999.14a08.x] [Medline: 10489826]
34. Holguin F, Cardet JC, Chung KF, Diver S, Ferreira DS, Fitzpatrick A, et al. Management of severe asthma: a European respiratory society/American thoracic society guideline. Eur Respir J 2020;55(1):1900588 [FREE Full text] [doi: 10.1183/13993003.00588-2019] [Medline: 31558662]
35. Liu AH, Zeiger R, Sorkness C, Mahr T, Ostrom N, Burgess S, et al. Development and cross-sectional validation of the childhood asthma control test. J Allergy Clin Immunol 2007;119(4):817-825. [doi: 10.1016/j.jaci.2006.12.662] [Medline: 17353040]
36. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
37. Harden A, Thomas J, Cargo M, Harris J, Pantoja T, Flemming K, et al. Cochrane qualitative and implementation methods group guidance series-paper 5: methods for integrating qualitative and implementation evidence within intervention effectiveness reviews. J Clin Epidemiol 2018;97:70-78. [doi: 10.1016/j.jclinepi.2017.11.029] [Medline: 29242095]
38. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Br Med J 2003;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Br Med J 1997;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. Br Med J 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
41. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52(6):377-384. [doi: 10.1136/jech.52.6.377] [Medline: 9764259]
42. Hooper P, Jutai JW, Strong G, Russell-Minda E. Age-related macular degeneration and low-vision rehabilitation: a systematic review. Can J Ophthalmol 2008;43(2):180-187. [doi: 10.3129/i08-001] [Medline: 18347620]

XSL•FO
RenderX

43. Pinnock H, Epiphaniou E, Pearce G, Parke H, Greenhalgh T, Sheikh A, et al. Implementing supported self-management for asthma: a systematic review and suggested hierarchy of evidence of implementation studies. BMC Med 2015;13:127. [doi: 10.1186/s12916-015-0361-0] [Medline: 26032941]
44. CASP Checklist: 10 questions to help you make sense of a Qualitative research. URL: https://casp-uk.net/wp-content/uploads/2018/01/CASP-Qualitative-Checklist-2018.pdf [accessed 2023-08-12]
45. Dennett EJ, Janjua S, Stovold E, Harrison SL, McDonnell MJ, Holland AE. Tailored or adapted interventions for adults with chronic obstructive pulmonary disease and at least one other long-term condition: a mixed methods review. Cochrane Database Syst Rev 2021;7(7):CD013384. [doi: 10.1002/14651858.CD013384.pub2] [Medline: 34309831]
46. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/efi-180221]
47. GRADEpro GDT: GRADEpro guideline development tool [software]. McMaster University and Evidence Prime, 2021. URL: https://www.gradepro.org/ [accessed 2024-10-30]
48. Lewin S, Glenton C, Munthe-Kaas H, Carlsen B, Colvin CJ, Gülmezoglu M, et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 2015;12(10):e1001895. [doi: 10.1371/journal.pmed.1001895] [Medline: 26506244]
49. GRADE-CERQual Interactive Summary of Qualitative Findings (iSoQ) Version 1.0. Oslo, Norway: Norwegian Institute of Public Health (developed by the Epistemonikos Foundation, Megan Wainwright Consulting and the Norwegian Institute of Public Health for the GRADE-CERQual Project Group). URL: https://isoq.epistemonikos.org/ [accessed 2023-12-25]
50. Acharya S, Sarraf R. A novel preventative solution for effective asthma management: a practical evaluation. Netw Model Anal Health Inform Bioinforma 2017;6(1):15. [doi: 10.1007/s13721-017-0156-7]
51. Ahmed S, Ernst P, Bartlett SJ, Valois M, Zaihra T, Paré G, et al. The effectiveness of web-based asthma self-management system, my asthma portal (MAP): a pilot randomized controlled trial. J Med Internet Res 2016;18(12):e313 [FREE Full text] [doi: 10.2196/jmir.5866] [Medline: 27908846]
52. Andrews AL, Nitchie HL, Harvey JB. Parent preferences for methods and content of mobile technology-based asthma medication adherence intervention. Hosp Pediatr 2019;9(3):209-215. [doi: 10.1542/hpeds.2018-0122] [Medline: 30737249]
53. Apter AJ, Localio AR, Morales KH, Han X, Perez L, Mullen AN, et al. Home visits for uncontrolled asthma among low-income adults with patient portal access. J Allergy Clin Immunol 2019;144(3):846-853.e11 [FREE Full text] [doi: 10.1016/j.jaci.2019.05.030] [Medline: 31181221]
54. Arimova PS, Namazova-Baranova LS, Levina JG, Kalugina VG, Vishneva EA, Kharitonova EY. Mobile technologies in achieving and maintaining asthma control in children: first results of MedQuizBot chat bot. Pediatr farmakol 2021;18(3):214-220. [doi: 10.15690/pf.v18i3.2279]
55. Chan DS, Callahan CW, Hatch-Pigott VB, Lawless A, Proffitt HL, Manning NE, et al. Internet-based home monitoring and education of children with asthma is comparable to ideal office-based care: results of a 1-year asthma in-home monitoring trial. Pediatrics 2007;119(3):569-578. [doi: 10.1542/peds.2006-1884] [Medline: 17332210]
56. Chan DS, Callahan CW, Sheets SJ, Moreno CN, Malone FJ. An Internet-based store-and-forward video home telehealth system for improving asthma outcomes in children. Am J Health Syst Pharm 2003;60(19):1976-1981. [doi: 10.1093/ajhp/60.19.1976] [Medline: 14531243]
57. Cingi C, Yorgancioglu A, Cingi CC, Oguzulgen K, Muluk NB, Ulusoy S, et al. The "physician on call patient engagement trial" (POPET): measuring the impact of a mobile patient engagement application on health outcomes and quality of life in allergic rhinitis and asthma patients. Int Forum Allergy Rhinol 2015;5(6):487-497. [doi: 10.1002/alr.21468] [Medline: 25856270]
58. Doyle R, Albright K, Hurley LP, Chávez C, Stowell M, Dircksen S, et al. Patient perspectives on a text messaging program to support asthma management: a qualitative study. Health Promot Pract 2019;20(4):585-592. [doi: 10.1177/1524839918770209] [Medline: 29732922]
59. Fiks AG, DuRivage N, Mayne SL, Finch S, Ross ME, Giacomini K, et al. Adoption of a portal for the primary care management of pediatric asthma: a mixed-methods implementation study. J Med Internet Res 2016;18(6):e172 [FREE Full text] [doi: 10.2196/jmir.5610] [Medline: 27357835]
60. Fiks AG, Mayne SL, Karavite DJ, Suh A, O'Hara R, Localio AR, et al. Parent-reported outcomes of a shared decision-making portal in asthma: a practice-based RCT. Pediatrics 2015;135(4):e965-e973. [doi: 10.1542/peds.2014-3167] [Medline: 25755233]
61. Gustafson D, Wise M, Bhattacharya A, Pulvermacher A, Shanovich K, Phillips B, et al. The effects of combining web-based eHealth with telephone nurse case management for pediatric asthma control: a randomized controlled trial. J Med Internet Res 2012;14(4):e101 [FREE Full text] [doi: 10.2196/jmir.1964] [Medline: 22835804]
62. Haze KA, Lynaugh J. Building patient relationships: a smartphone application supporting communication between teenagers with asthma and the RN care coordinator. Comput Inform Nurs 2013;31(6):266-271; quiz 272. [doi: 10.1097/NXN.0b013e318295e5ba] [Medline: 23728445]
63. Hui CY, McKinstry B, Fulton O, Buchner M, Pinnock H. Patients' and clinicians' visions of a future Internet-of-things system to support asthma self-management: mixed methods study. J Med Internet Res 2021;23(4):e22432. [doi: 10.2196/22432] [Medline: 33847592]

XSL•FO
RenderX

64. Kosse RC, Bouvy ML, Belitser SV, de Vries TW, van der Wal PS, Koster ES. Effective engagement of adolescent asthma patients with mobile health-supporting medication adherence. JMIR Mhealth Uhealth 2019;7(3):e12411 [FREE Full text] [doi: 10.2196/12411] [Medline: 30916664]
65. Kosse RC, Bouvy ML, de Vries TW, Koster ES. Effect of a mHealth intervention on adherence in adolescents with asthma: a randomized controlled trial. Respir Med 2019;149:45-51. [doi: 10.1016/j.rmed.2019.02.009] [Medline: 30803885]
66. Nichols M, Miller S, Treiber F, Ruggiero K, Dawley E, Teufel Ii R. Patient and parent perspectives on improving pediatric asthma self-management through a mobile health intervention: pilot study. JMIR Form Res 2020;4(7):e15295 [FREE Full text] [doi: 10.2196/15295] [Medline: 32442127]
67. Ostojic V, Cvoriscec B, Ostojic SB, Reznikoff D, Stipic-Markovic A, Tudjman Z. Improving asthma control through telemedicine: a study of short-message service. Telemed J E Health 2005;11(1):28-35. [doi: 10.1089/tmj.2005.11.28] [Medline: 15785218]
68. Rasmussen LM, Phanareth K, Nolte H, Backer V. Internet-based monitoring of asthma: a long-term, randomized clinical study of 300 asthmatic subjects. J Allergy Clin Immunol 2005;115(6):1137-1142. [doi: 10.1016/j.jaci.2005.03.030] [Medline: 15940125]
69. Rasulnia M, Burton BS, Ginter RP, Wang TY, Pleasants RA, Green CL, et al. Assessing the impact of a remote digital coaching engagement program on patient-reported outcomes in asthma. J Asthma 2018;55(7):795-800. [doi: 10.1080/02770903.2017.1362430] [Medline: 28800273]
70. Schneider T, Panzera AD, Martinasek M, McDermott R, Couluris M, Lindenberger J, et al. Physicians' perceptions of mobile technology for enhancing asthma care for youth. J Child Health Care 2016;20(2):153-163. [doi: 10.1177/1367493514556555] [Medline: 25427556]
71. van den Wijngaart LS, Geense WW, Boehmer AL, Brouwer ML, Hugen CA, van Ewijk BE, et al. Barriers and facilitators when implementing web-based disease monitoring and management as a substitution for regular outpatient care in pediatric asthma: qualitative survey study. J Med Internet Res 2018;20(10):e284. [doi: 10.2196/jmir.9245] [Medline: 30377147]
72. van den Wijngaart LS, Roukema J, Boehmer ALM, Brouwer ML, Hugen CAC, Niers LEM, et al. A virtual asthma clinic for children: fewer routine outpatient visits, same asthma control. Eur Respir J 2017;50(4):1700471 [FREE Full text] [doi: 10.1183/13993003.00471-2017] [Medline: 28982775]
73. van der Kamp M, Reimering Hartgerink P, Driessen J, Thio B, Hermens H, Tabak M. Feasibility, efficacy, and efficiency of eHealth-supported pediatric asthma care: six-month quasi-experimental single-arm pretest-posttest study. JMIR Form Res 2021;5(7):e24634. [doi: 10.2196/24634] [Medline: 34309568]
74. van der Meer V, Bakker MJ, van den Hout WB, Rabe KF, Sterk PJ, Kievit J, SMASHING (Self-Management in Asthma Supported by Hospitals, ICT, NursesGeneral Practitioners) Study Group. Internet-based self-management plus education compared with usual care in asthma: a randomized trial. Ann Intern Med 2009;151(2):110-120. [doi: 10.7326/0003-4819-151-2-200907210-00008] [Medline: 19620163]
75. van der Meer V, van Stel HF, Detmar SB, Otten W, Sterk PJ, Sont JK. Internet-based self-management offers an opportunity to achieve better asthma control in adolescents. Chest 2007;132(1):112-119. [doi: 10.1378/chest.06-2787] [Medline: 17400674]
76. van Gaalen JL, Beerthuizen T, van der Meer V, van Reisen P, Redelijkheid GW, Snoeck-Stroband JB, SMASHING Study Group. Long-term outcomes of internet-based self-management support in adults with asthma: randomized controlled trial. J Med Internet Res 2013;15(9):e188 [FREE Full text] [doi: 10.2196/jmir.2640] [Medline: 24028826]
77. van Gaalen JL, van Bodegom-Vos L, Bakker MJ, Snoeck-Stroband JB, Sont JK. Internet-based self-management support for adults with asthma: a qualitative study among patients, general practitioners and practice nurses on barriers to implementation. BMJ Open 2016;6(8):e010809. [doi: 10.1136/bmjopen-2015-010809] [Medline: 27566627]
78. Voorend-van Bergen S, Vaessen-Verberne AA, Brackel HJ, Landstra AM, van den Berg NJ, Hop WC, et al. Monitoring strategies in children with asthma: a randomised controlled trial. Thorax 2015;70(6):543-550. [doi: 10.1136/thoraxjnl-2014-206161] [Medline: 25825006]
79. Wiecha JM, Adams WG, Rybin D, Rizzodepaoli M, Keller J, Clay JM. Evaluation of a web-based asthma self-management system: a randomised controlled pilot trial. BMC Pulm Med 2015;15:17 [FREE Full text] [doi: 10.1186/s12890-015-0007-1] [Medline: 25885418]
80. Xu C, Jackson M, Scuffham PA, Wootton R, Simpson P, Whitty J, et al. A randomized controlled trial of an interactive voice response telephone system and specialist nurse support for childhood asthma management. J Asthma 2010;47(7):768-773. [doi: 10.3109/02770903.2010.493966] [Medline: 20716014]
81. Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, et al. The COMET Handbook: version 1.0. Trials 2017;18(Suppl 3):280. [doi: 10.1186/s13063-017-1978-4] [Medline: 28681707]
82. Seppen BF, den Boer P, Wiegel J, Ter Wee MM, van der Leeden M, de Vries R, et al. Asynchronous mHealth interventions in rheumatoid arthritis: systematic scoping review. JMIR Mhealth Uhealth 2020;8(11):e19260. [doi: 10.2196/19260] [Medline: 33151161]
83. de Jong CC, Ros WJ, Schrijvers G. The effects on health behavior and health outcomes of Internet-based asynchronous communication between health providers and patients with a chronic condition: a systematic review. J Med Internet Res 2014;16(1):e19. [doi: 10.2196/jmir.3000] [Medline: 24434570]

XSL•FO
RenderX

84. Deshpande A, Khoja S, Lorca J, McKibbon A, Rizo C, Husereau D, et al. Asynchronous telehealth: a scoping review of analytic studies. Open Med 2009;3(2):e69-e91. [Medline: 19946396]
85. Eschler J, Liu LS, Vizer LM, McClure JB, Lozano P, Pratt W, et al. Designing asynchronous communication tools for optimization of patient-clinician coordination. AMIA Annu Symp Proc 2015;2015:543-552 [FREE Full text] [Medline: 26958188]
86. Odendaal WA, Anstey Watkins J, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020;3(3):CD011942. [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
87. Banks J, Farr M, Salisbury C, Bernard E, Northstone K, Edwards H, et al. Use of an electronic consultation system in primary care: a qualitative interview study. Br J Gen Pract 2018;68(666):e1-e8. [doi: 10.3399/bjgp17X693509] [Medline: 29109115]
88. Lie SS, Karlsen B, Graue M, Oftedal B. The influence of an eHealth intervention for adults with type 2 diabetes on the patient-nurse relationship: a qualitative study. Scand J Caring Sci 2019;33(3):741-749. [doi: 10.1111/scs.12671] [Medline: 30866066]
89. Griffiths F, Bryce C, Cave J, Dritsaki M, Fraser J, Hamilton K, et al. Timely digital patient-clinician communication in specialist clinical services for young people: a mixed-methods study (the LYNC study). J Med Internet Res 2017;19(4):e102 [FREE Full text] [doi: 10.2196/jmir.7154] [Medline: 28396301]
90. Entezarjou A, Bolmsjö BB, Calling S, Midlöv P, Milos Nymberg V. Experiences of digital communication with automated patient interviews and asynchronous chat in Swedish primary care: a qualitative study. BMJ Open 2020;10(7):e036585. [doi: 10.1136/bmjopen-2019-036585] [Medline: 32709650]
91. Matricardi PM, Dramburg S, Alvarez-Perea A, Antolín-Amérigo D, Apfelbacher C, Atanaskovic-Markovic M, et al. The role of mobile health technologies in allergy care: an EAACI position paper. Allergy 2020;75(2):259-272. [doi: 10.1111/all.13953] [Medline: 31230373]
92. Cowie J, Calveley E, Bowers G, Bowers J. Evaluation of a digital consultation and self-care advice tool in primary care: a multi-methods study. Int J Environ Res Public Health 2018;15(5):896. [doi: 10.3390/ijerph15050896] [Medline: 29724040]
93. Williams T, May C, Esmail A, Ellis N, Griffiths C, Stewart E, et al. Patient satisfaction with store-and-forward teledermatology. J Telemed Telecare 2001;7 Suppl 1:45-46. [doi: 10.1177/1357633X010070S118] [Medline: 11576488]
94. Nguyen OT, Alishahi Tabriz A, Huo J, Hanna K, Shea CM, Turner K. Impact of asynchronous electronic communication-based visits on clinical outcomes and health care delivery: systematic review. J Med Internet Res 2021;23(5):e27531 [FREE Full text] [doi: 10.2196/27531] [Medline: 33843592]

Abbreviations

ACQ: asthma control questionnaire
ACT: asthma control test
AQLQ: asthma quality of life questionnaire
C-ACT: childhood-ACT
EHR: electronic health record
GRADE: Grading of Recommendations, Assessment, Development, And Evaluations
GRADE-CERQual: GRADE-Confidence in the Evidence from Qualitative Reviews
MAQLQ: mini AQLQ
PAQLQ: pediatric AQLQ
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized control trial
RR: risk ratio
SMD: standardized mean difference

Edited by N Cahill; submitted 24.02.24; peer-reviewed by D Ryan, C Almonacid; comments to author 11.05.24; revised version received 20.05.24; accepted 30.09.24; published 03.12.24.

Please cite as:
Uzzaman N, Hammersley V, McClatchey K, Sheringham J, Singh D, Habib GMM, Pinnock H
Effectiveness and Acceptability of Asynchronous Digital Health in Asthma Care: Mixed Methods Systematic Review
J Med Internet Res 2024;26:e57708
URL: https://www.jmir.org/2024/1/e57708
doi:10.2196/57708
PMID:

XSL•FO
RenderX

©Nazim Uzzaman, Victoria Hammersley, Kirstie McClatchey, Jessica Sheringham, Diksha Singh, GM Monsur Habib, Hilary Pinnock. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Application of Chatbots to Help Patients Self-Manage Diabetes: Systematic Review and Meta-Analysis

Yibo Wu[1*], PhD; Jinzi Zhang[2,3*], MS; Pu Ge[4], MS; Tingyu Duan[5], BA; Junyu Zhou[6], BS; Yiwei Wu[7], BA; Yuening Zhang[8], BS; Siyu Liu[9], BS; Xinyi Liu[10], BS; Erya Wan[11], MS; Xinying Sun[1], PhD

[1]School of Public Health, Peking University, Beijing, China
[2]School of Humanities and Social Sciences, Harbin Medical University, Harbin, China
[3]Graduate School, Harbin Medical University, Harbin, China
[4]School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China
[5]School of Journalism and Communication, Hebei Institute of Communications,, Shijiazhuang, China
[6]School of Humanities and Social Sciences, University of Science and Technology of China, Hefei, China
[7]Institute of Communication Studies, Communication University of China, Beijing, China
[8]College of Nursing, North Sichuan Medical college, Nanchong, China
[9]College of Stomatology, Shandong University, Jinan, China
[10]School of Nursing and Rehabilitation, Xi'an Medical University, Xi'an, China
[11]School of Global Health, Shanghai Jiao Tong University, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Xinying Sun, PhD
School of Public Health
Peking University
38 Xueyuan Road, Haidian District
Beijing, 100191
China
Phone: 86 13691212050
Email: xysun@bjmu.edu.cn

Abstract

Background: The number of people with diabetes is on the rise globally. Self-management and health education of patients are the keys to control diabetes. With the development of digital therapies and artificial intelligence, chatbots have the potential to provide health-related information and improve accessibility and effectiveness in the field of patient self-management.

Objective: This study systematically reviews the current research status and effectiveness of chatbots in the field of diabetes self-management to support the development of diabetes chatbots.

Methods: A systematic review and meta-analysis of chatbots that can help patients with diabetes with self-management was conducted. PubMed and Web of Science databases were searched using keywords around diabetes, chatbots, conversational agents, virtual assistants, and more. The search period was from the date of creation of the databases to January 1, 2023. Research articles in English that fit the study topic were selected, and articles that did not fit the study topic or were not available in full text were excluded.

Results: In total, 25 studies were included in the review. In terms of study type, all articles could be classified as systematic design studies (n=8, 32%), pilot studies (n=8, 32%), and intervention studies (n=9, 36%). Many articles adopted a nonrandomized controlled trial design in intervention studies (n=6, 24%), and there was only 1 (4%) randomized controlled trial. In terms of research strategy, all articles can be divided into quantitative studies (n=10, 40%), mixed studies (n=6, 24%), and qualitative studies (n=1, 4%). The evaluation criteria for chatbot effectiveness can be divided into technical performance evaluation, user experience evaluation, and user health evaluation. Most chatbots (n=17, 68%) provided education and management focused on patient diet, exercise, glucose monitoring, medications, and complications, and only a few studies (n=2, 8%) provided education on mental health. The meta-analysis found that the chatbot intervention was effective in lowering blood glucose (mean difference 0.30, 95% CI 0.04-0.55; P=.02) and had no significant effect in reducing weight (mean difference 1.41, 95% CI –2.29 to 5.11; P=.46) compared with the baseline.

XSL•FO
RenderX

Conclusions: Chatbots have potential for the development of self-management for people with diabetes. However, the evidence level of current research is low, and higher level research (such as randomized controlled trials) is needed to strengthen the evidence base. More use of mixed research in the research strategy is needed to fully use the strengths of both quantitative and qualitative research. Appropriate and innovative theoretical frameworks should be used in the research to provide theoretical support for the study. In addition, researchers should focus on the personalized and user-friendly interactive features of chatbots, as well as improvements in study design.

(J Med Internet Res 2024;26:e60380) doi:10.2196/60380

KEYWORDS

artificial intelligence; chatbot; diabetes; health education; self-management; systematic review

Introduction

Background

Diabetes, as a significant global public health problem, has resulted in a heavy burden on the health care system. Approximately 530 million people worldwide currently have diabetes, and by 2050, more than 1.31 billion people will have diabetes globally [1]. In 2021, the global health expenditure caused by diabetes was about US $966 billion, and it is estimated that this figure will exceed US $1054 billion by 2045 [2]. Currently, there is no effective cure for diabetes, but it is somewhat controllable and preventable. The American Diabetes Association Professional Practice Committee suggests that diabetes may be diagnosed based on hemoglobin A_{1C} (HbA_{1c}) criteria or plasma glucose criteria, either the fasting plasma glucose value, 2-hour glucose value during a 75-g oral glucose tolerance test, or random glucose value accompanied by classic hyperglycemic symptoms or hyperglycemic crises. Therefore, the main control of diabetes mellitus is the blood glucose level [3]. Diabetes control can reduce and delay the onset and progression of diabetes complications. Good diabetes control depends on a combination of diet, exercise, medication, diabetes education, and blood glucose monitoring [4]. However, high-quality treatment of diabetes and its complications also depends on education about diabetes and active management of the patient's own behavior (also known as self-management) [5]. Health education is vital for the self-management of diabetes [6]. The purpose of diabetes education is to help patients improve their self-care ability, that is, to increase knowledge about self-care of diabetes and to improve the ability of self-care practice [5]. Diabetes self-management and support have been provided face-to-face (one-to-one or group based) for decades, with many trials and real-world studies showing improved diabetes outcomes [7]. However, high costs and resource requirements have limited the scope and scalability of face-to-face programs [8].

Digital therapeutics is a new concept: it is technology produced by the digital integration of medicine and health. It combines disease, data, and algorithms [9]. The essence of digital therapeutics is a disease intervention program based on evidence-based medicine and presented by software programs that provide core functions such as the prevention, treatment, or management of diseases and high-risk factors [5]. Digital therapy applications (Blue Star, Diabeo System, etc) use web-based applications or cloud platforms to provide evidence-based, personalized, and rapid care management for chronic and behavioral-changeable diseases, including diabetes [10]. Digital therapy applications can improve patients' compliance, treatment success rate, and economic achievements in diabetes management by enhancing patients' participation, changing patients' lifestyles, providing comprehensive medical care, and regular glucose monitoring [11]. At present, the increase of patients with diabetes has brought a heavy burden to medical professionals and the medical care system. With the development of artificial intelligence (AI) and digital therapies, a convenient and effective diabetes health education system is gradually being developed [12]. It meets the individual needs of patients with diabetes to acquire knowledge about diabetes and self-management, helps patients improve their daily lifestyles, and supports them in the self-management of diabetes.

The chatbot is one of the products of scientific and technological progress and digital therapy, which has great potential in providing health education for patients and helping patients manage themselves. The chatbot can also be called a conversational agent or system, defined as a nonhuman virtual conversational bot, a computer program, or an AI program [13]. It participates in the conversation via audio or text input to answer the user's questions, and the user can easily access the desired information by sending messages to the chatbot [14]. AI is now gradually being integrated into chatbots. When a user asks a question or issues a command in human language, the chatbot understands the contextual information sent by the user and responds accordingly. Google Assistant, Apple's Siri, and ChatGPT are common chatbots with voice-activated interfaces [15,16]. Chatbots can play an essential role in many areas, especially in health care. They can help patients anytime via a smartphone app or the web [17]. Chatbots can provide fascinating health-related information and services and provide a personalized lifestyle [18]. Chatbots can usually provide better answers to health care questions than other health care resources [19]. For health care organizations, the use of chatbots significantly improves efficiency and reduces the workload of health care professionals and the burden on the health care system [20,21]. For patients, the use of chatbots can assist in self-management, improving behavioral compliance and quality of life [22]. Previous findings indicate that chatbot interventions effectively improve depression, anxiety, stress, medication adherence, and smoking cessation [23].

Chatbots have the potential to provide personalized health education and assist in self-management for people with diabetes. According to our findings, although some studies have analyzed diabetes and self-management or investigated the use

XSL•FO
RenderX

of chatbots in other diseases [24,25], no study has systematically reviewed the use of chatbots in self-management for patients with diabetes. Therefore, this study is innovative to some extent.

Objectives

In summary, the aim of this study was to conduct a systematic review and meta-analysis of chatbots used in the field of diabetes self-management, exploring the specific characteristics of chatbots, research methods, research strategies, evaluation measures of outcome indicators, and some other information. By analyzing information about chatbots, this study aims to assess the potential and effectiveness of chatbots in diabetes self-management, thus helping researchers to sort out the current state of research and future development trends in this field. This will provide support for improving the development of chatbots in the field of diabetes self-management, which will ultimately help patients with diabetes to stay healthy.

Methods

Search Strategy

A systematic search of the literature was conducted in the PubMed and Web of Science databases, from the date of database establishment until January 1, 2023. PubMed and Web of Science are the 2 most commonly used databases for scholars in the biomedical field. PubMed is the largest biomedical retrieval platform in the world. Web of Science is the most classic and authoritative citation database in the scientific community, and the most important core collection of Web of Science contains more than 12,000 academic journals in more than 250 disciplines. The quality of articles in these 2 databases is very high, and hence, the researcher chose to extract high-quality literature from these 2 databases. Search terms were developed from a combination of keywords and Boolean operators in consultation with a professional medical librarian. The keywords of the search mainly revolve around “diabetes” and “Chatbot” (see Table S1 in Multimedia Appendix 1 for the search strings).

Data Management and Extraction

The data filtering process can be divided into four stages: (1) duplicate data removal, (2) primary filtering (filtering based on titles and abstracts), (3) secondary filtering (filtering based on full text), and (4) cross-checking. Two researchers (JZZ and GP) independently screened the articles according to the inclusion and exclusion criteria (see Table S2 in Multimedia Appendix 1). In the first step, for articles that were both in the Web of Science and PubMed databases, only one of them was retained, and duplicates were removed. In the second step, irrelevant articles were initially filtered out by reading titles and abstracts. In the third step, articles that did not fit the topic of the study were excluded by carefully reading the full text for judgment. Finally, independent screening results from JZZ and GP were cross-checked to ensure the accuracy of article screening. For articles with screening disagreements, the disagreements were resolved through discussion with the senior author, YBW. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram was used to visualize the comprehensive search process and the results obtained.

The quality of the included literature was evaluated by 2 researchers independently using the evaluation criteria for observational studies developed by the Agency for Healthcare Research and Quality [26]. The tool has a total of 11 entries, each of which contains 3 options: “Yes,” “No,” and “Unclear.” A score of 1 was given for “Yes” and 0 for “No” or “Unclear”; the total score was 11, with 0-3 as low quality, 4-7 as medium quality, and 8-11 as high quality. Disagreements were resolved through discussion between authors. For details of the quality evaluation results, refer to Table S3 in Multimedia Appendix 1.

Outcomes of Interest

The basic information and data from the articles were extracted and checked according to a predesigned data extraction table. The data information extracted for each study includes (1) basic information about the article (author, country, and year of publication); (2) specific features of the chatbot (name, consultable content, AI technology, question and answer database data sources, and input and output methods); and (3) research design aspects (research strategy, research type, theoretical framework, evaluation indexes of chatbot effect, etc).

Meta-Analysis Methods

Meta-analysis was planned if the included studies were sufficiently homogeneous in terms of statistical characteristics. RevMan 5.3 software (Cochrane Collaboration) was used for the statistical analysis of data. Mean, SD, and 95% CI were used as study effect indicators. The heterogeneity test was performed on the combined data. If $P>.10$ and $I^2<50\%$, it suggests that the heterogeneity among studies was small, and the fixed-effect model was selected for meta-analysis. If $P<.10$ and $I^2>50\%$, it suggests that there is obvious heterogeneity among the studies, and the random effect model was used for analysis. Sensitivity analysis was performed by comparing the results of the combined effect sizes of the fixed-effects model and random-effects model to ensure the stability and reliability of the results [27].

Results

Search Results

In total, 25 articles from 14 countries were included in the analysis (see Table 1 for details). Authors from Australia and Portugal published 4 articles each, and authors from the United States and India published 3 articles each. Also, high-income countries, including Singapore, Norway, the Netherlands, Spain, Greece, and South Korea, and low-income countries, including China, Bangladesh, South Africa, and Ghana, have published relevant studies. The PRISMA flow diagram is shown in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the study selection process.

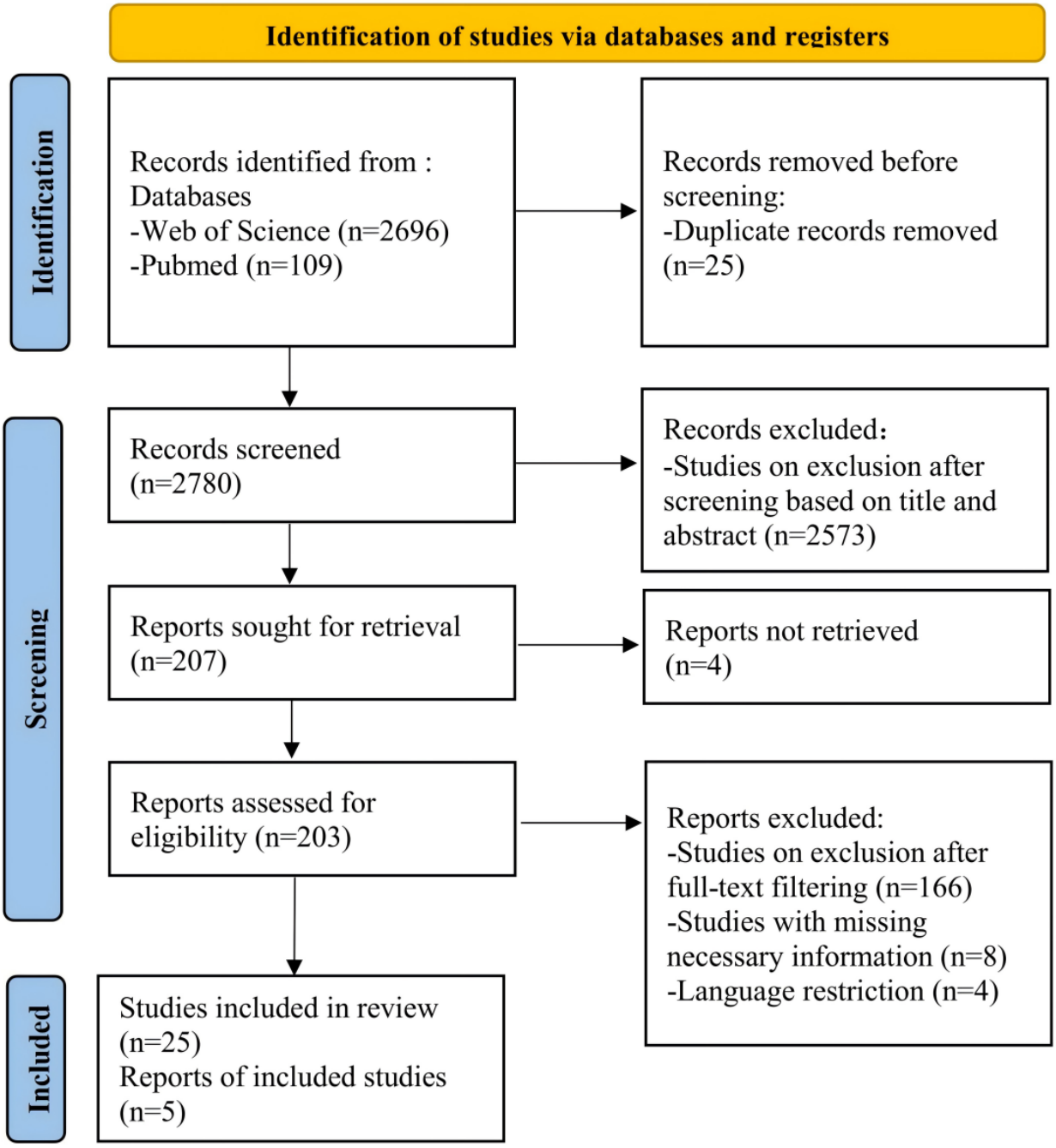

XSL•FO
RenderX

Table 1. Basic information of included articles.

Study	Year	Research type	Country	Consultable content	Evaluation metrics
Krishnakumar et al [28]	2021	Interventional study	India	Diet, exercise, medications, and blood glucose	User experience and user health
Stephens et al [29]	2019	Pilot study	United States	Diet, exercise, and mental health	Technical performance and user experience
Sharma et al [30]	2018	System design	India	General diabetes knowledge	None (theoretical assumptions)
Sowah et al [31]	2020	System design	Ghana	General diabetes knowledge	Technical performance
Anastasiadou et al [32]	2020	System design	Greece	Diet, exercise, medications, blood glucose, treatment, learning resources, and doctor information	User experience
Beaudry et al [33]	2019	Pilot study	United States	Drugs and general knowledge of chronic diseases	User experience
Mitchell et al [34]	2021	Pilot study	United States	Diet and exercise	User experience
Hossain et al [35]	2022	System design	Bangladesh	General diabetes knowledge	Technical performance
Xie et al [6]	2018	System design	China	Diet, medication, and symptoms	Technical performance
Maher et al [36]	2020	Interventional study	Australia	Diet and exercise	User experience and user health
Sagstad et al [37]	2022	Interventional study	Norway	Diet, exercise, blood glucose, and gestational diabetes	User experience
Mash et al [38]	2022	Interventional study	South Africa	Diet, exercise, medications, blood glucose, complications, tobacco, alcohol, and mental health	User experience
Baptista et al [39]	2020	Interventional study	Australia	Diet, exercise, medications, blood glucose, and complications	User experience
Hurmuz et al [40]	2020	Interventional study	Netherlands	Diet, exercise, and social support	User experience and user health
Gong et al [41]	2020	Interventional study	Australia	Diet, exercise, medications, blood glucose, and complications	User experience and user health
Roca et al [42]	2021	Interventional study	Spain	Medicine	User experience and user health
Pimenta et al [43]	2022	System design	Portugal	Diet, exercise, and medication	None (theoretical assumptions)
Balsa et al [44]	2020	Pilot study	Portugal	Diet, exercise, and medication	User experience
Félix et al [45]	2019	System design	Portugal	Diet, exercise, and medication	None (theoretical assumptions)
Saritha et al [46]	2020	System design	India	Diabetes general knowledge	Technical performance
Buinhas et al [47]	2019	Pilot study	Portugal	Diet, exercise, and medication	User experience
Dhinagaran et al [48]	2021	Interventional study	Singapore	Diet, exercise, sleep, and mental health	User experience and user health
Dhinagaran and Car [49]	2022	Pilot study	Singapore	Diet, exercise, sleep, and stress management	User experience
Hussain and Athula [50]	2018	Pilot study	Australia	General knowledge of diabetes	Technical performance and user experience
Rehman et al [51]	2020	Pilot study	South Korea	Disease prediction	Technical performance and user experience

Characteristics of Included Studies

Self-Management Content

Most chatbots provided self-management services (n=17, 68%) around lifestyle interventions (diet and exercise). In addition to providing information on diet and exercise, several studies simultaneously provided information about medications, blood glucose, and complications. In total, 5 (20%) studies provided the management of mental health dimensions such as anxiety, depression, social support, and stress. Two (8%) other studies provided functions in sleep quality, smoking, and alcohol

XSL•FO
RenderX

consumption (see Tables S4-6 in Multimedia Appendix 1 for details). The chatbots included in this study provide education and self-management services to different types of populations (including people with type 2 diabetes, adolescents, older adults, etc; see Tables S4-6 in Multimedia Appendix 1 for details). For example, the Tess chatbot is intended for adolescents. Tess is designed to respond to users with scripted statements that are crafted by mental health professionals to mimic empathy and compassion. Similar to how therapists can adapt their style or approach to address each client's needs, Tess intervenes and responds based on an individual's reported emotions or concerns [29].

Question and Answer Database Source

Researchers have mainly collected and extracted diabetes-related health information as a database source for chatbots through the International Diabetes Federation, clinical guidelines on diabetes published in various countries, authoritative health websites, and medical journal literature (see Tables S4-6 in Multimedia Appendix 1 for details).

Interaction and Input-Output Methods

When users interact with a chatbot to ask a question, the chatbot could provide 3 different forms of question and answer: free input (where the user freely enters a question), button input (where the user clicks on a topic or number button to ask a question), and mixed mode (where both free input and button modes are available). Many chatbots support users' use of unrestricted free speech for queries (n=12, 48%). Ten (40%) studies supported users in making inquiries based on topic buttons provided by the chatbots, and 3 (12%) other studies supported mixed mode. Most chatbots allowed users to enter questions using text (n=19, 76%), and a small number of chatbots allowed users to make inquiries using voice (Tables S4-6 in Multimedia Appendix 1).

Research Strategy

In our study, research strategies can be categorized into 3 types: qualitative research (n=9, 36%), quantitative research (n=1, 4%), and mixed research (n=7, 28%). Qualitative and quantitative research, in particular, have their own strengths and weaknesses, and mixed research allows quantitative and qualitative research to complement each other's strengths (the advantages of quantitative, qualitative, and mixed studies are detailed in Multimedia Appendix 2).

The Theoretical Framework

A total of 10 (40%) articles included in our study reported theoretical frameworks (Tables S4-6 in Multimedia Appendix 1). These theoretical frameworks mainly include the Behavior Change Wheel, Self-Determination Theory, Behavior Change Theory (Cross-Theoretical Model and Social Cognitive Theory), and so on. One of the more applied theoretical frameworks in the study was the Behavior Change Wheel, followed by the Behavior Change Theory and the Self-Determination Theory. The researchers integrated the theoretical frameworks in the design and intervention functions of the chatbot (see Multimedia Appendix 3 for specific explanations and applications of the theoretical frameworks in detail).

Research Type

In our study, researchers classified all articles into 3 types of studies based on research characteristics: system design studies (n=8, 32%), pilot studies (n=8, 32%), and intervention studies (n=9, 36%). In system design studies, chatbots are at an early stage of the study. Most articles at this stage (n=5, 62.5%) consist of theoretical concepts or design chatbots based on theoretical concepts and test their performance. In pilot studies, researchers go deeper in their exploration based on system design. Articles at this stage choose small-scale, short-term experiments to evaluate the effect of chatbots. The intervention studies were a deeper exploration by the researcher based on the pilot study. This study phase has more participants, a longer intervention period, and a more standardized study design. Specific characteristics of chatbots in each study type are detailed in Multimedia Appendices 1 and 4.

Meta-Analysis Results

The studies included in the meta-analysis were pre-post trials. Based on the homogeneity characteristics of the included studies, the researchers used HbA_{1c} and weight change as outcome indicators. In the meta-analysis, the researchers used the diabetes chatbot as an intervention to analyze the changes in HbA_{1c} and body weight of the study participants after a period of intervention, compared to the baseline. The results of the meta-analysis are shown below.

HbA1c

Meta-analysis of HbA_{1c} included 4 articles (Krishnakumar et al [28], Gong et al [41], Baptista et al [39], and Roca et al [42]), totaling 4 groups of pre-post trials and 219 trial participants (Figure 2). The results showed that there was no overall heterogeneity among the 4 groups of included trials (I^2=0%, P=.94). The combined total effect size and 95% CI obtained using a fixed-effects model was (mean difference [MD] 0.30, 95% CI 0.04-0.55; P=.02), which reached the level of significance. Thus, the chatbot intervention had an effect on lowering HbA_{1c} compared with the baseline. This analysis had consistent results under both fixed-effects and random-effects models. Article publication bias was assessed using funnel plots (Figure S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Comparison of fixed-effects and random-effects models in meta-analysis of HbA1c. HbA1c: hemoglobin A1c.

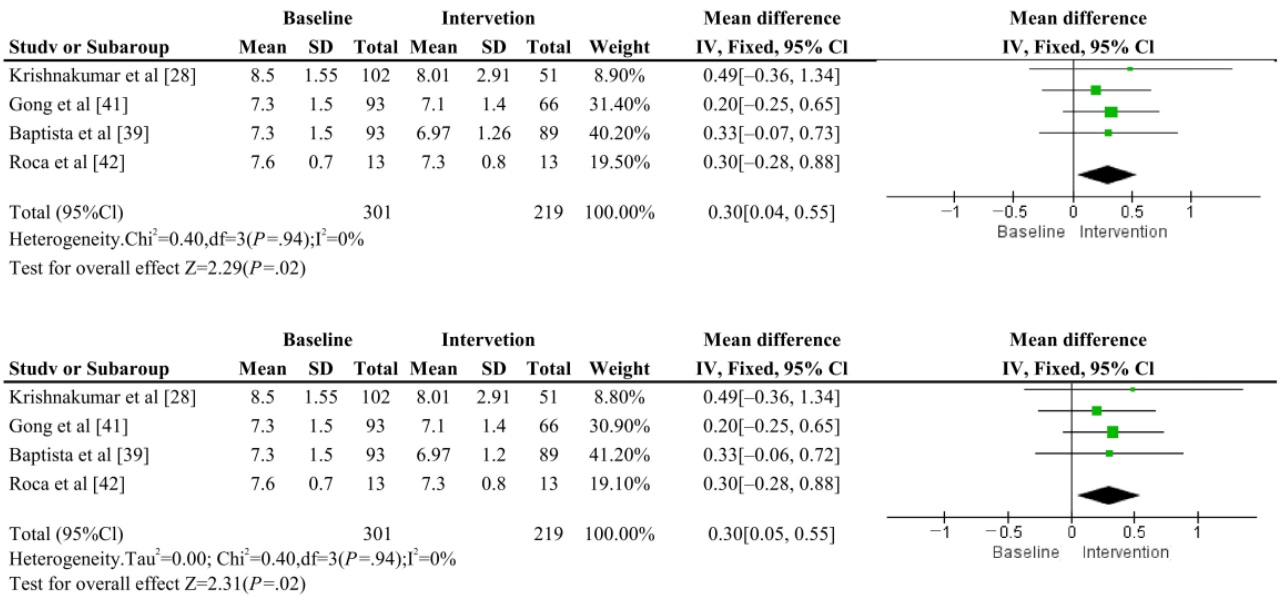

Weight

Meta-analysis of weight included 3 articles (Krishnakumar et al [28], Maher et al [36], and Gong et al [41]), totaling 3 sets of pre-post trials with 176 trial participants (Figure 3). The results showed that there was no overall heterogeneity among the 3 groups of included trials (I^2=0%, P>.99). The combined total effect size and 95% CI obtained using a fixed-effects model was (MD=1.41, 95% CI –2.29 to 5.11; P=.46), which did not reach the level of significance. Thus, there was no significant effect of the chatbot intervention in reducing body weight compared to baseline. This analysis had consistent results under both fixed-effects and random-effects models. Article publication bias was assessed using funnel plots (Figure S2 in Multimedia Appendix 1).

Figure 3. Comparison of fixed-effects and random-effects models in meta-analysis of weight.

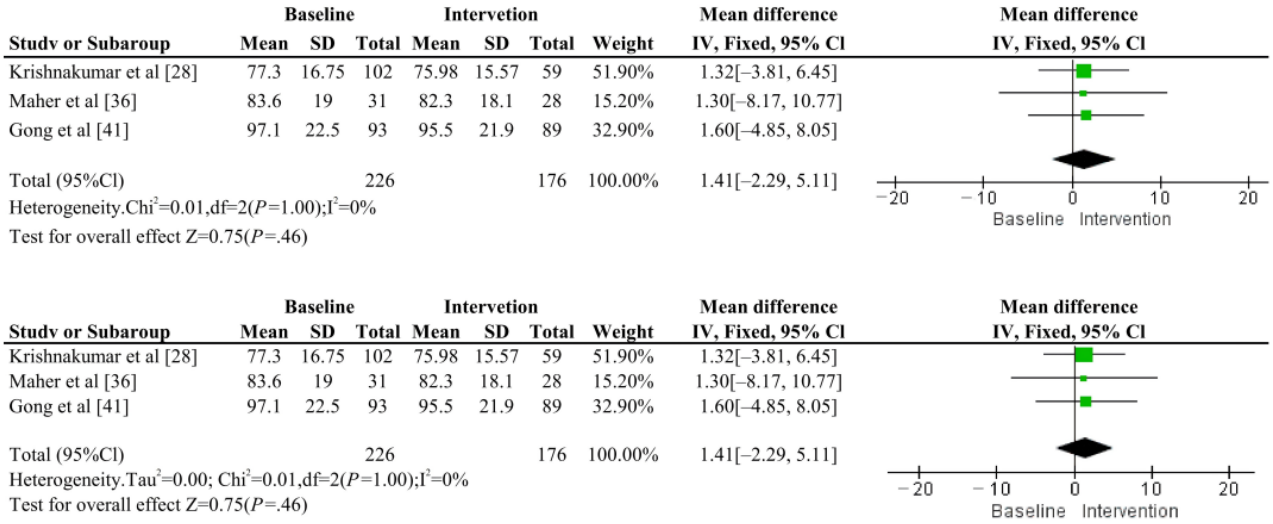

Discussion

Principal Findings

This study is a systematic review and meta-analysis of the effectiveness of chatbots in diabetes self-management. A total of 25 articles were analyzed in this review. Based on the type of study, the 25 studies can be categorized as systematic design type studies, pilot studies, and intervention studies. Chatbots provide services involving both physical and psychological aspects. The criteria for assessing the effectiveness of chatbots can be categorized into technical performance assessment, user experience assessment, and user health assessment. Overall, the results of the evaluation of the chatbot model were good, and the overall acceptance of chatbots by users was high.

General Overview of Study

Chatbots could play an important role in future self-management for patients with diabetes. However, the current use of chatbots in diabetes self-management is still in its infancy. The number of publications in each country was low, and there was room for further research and development. Compared to high-income

XSL•FO
RenderX

countries, there were fewer publications in low-income countries. However, low-income countries have a large population base, and the number of people with diabetes is increasing. Therefore, researchers should pay attention to conducting relevant research in low-income countries in the future.

The chatbot can provide management on physical dimensions such as diet, exercise, medication, blood glucose, and complications management. Only a few articles provided services on psychological dimensions such as anxiety, depression, stress, and social support. More than half of the diabetes chatbots enable users to enter questions freely. A few chatbots offer button input. More than half of the chatbots offer text-based counseling, and a few chatbots allow for voice-to-text counseling. Voice interaction can facilitate users' interaction with chatbots. Therefore, this represents an area for future research. Chatbots need to establish an appropriate rapport with the user through personalized and friendly interactions for sustained and engaging interventions.

Meta-Analysis Results

Conducting a meta-analysis of the interventional studies, the researchers extracted 2 outcome metrics with a high degree of homogeneity—HbA_{1c} and body weight—as criteria for evaluating the effectiveness of the chatbot intervention. Other intervention studies have used measures such as quality of life, BMI, waist circumference, blood pressure, diabetes knowledge, and risk perception. Meta-analysis of the study results showed that the chatbot intervention was effective in lowering blood glucose but had no significant effect on weight reduction. For example, a pilot study conducted over a 4-month period allowed patients to record their meals, weight, physical activity, and blood glucose in a chatbot. Patients received sessions on self-care behaviors and were provided with feedback from the AI-powered chatbot, as well as regular informational interactions with a diabetes educator via voice calls. To increase patient motivation and reduce the difficulty of performing specific tasks, the chatbot facilitates patient action through a variety of forms of brief educational content (e.g., informative lessons, videos, quizzes, and stories). The program uses a gamified approach to build self-management skills and rewards patients with virtual trophies as they complete lessons, tasks, and challenges. As participation in the program increased, patients' HbA_{1c}, fasting blood glucose, and postprandial blood glucose levels gradually decreased [29].

The Role of Chatbots

The results of this study found that as technologies such as semantic analysis and natural language processing continue to be incorporated into chatbots, researchers have built a knowledge base of diabetes questions and answers with logical rules, so that chatbots can provide users with personalized advice and management services to help patients with diabetes improve their self-management.

Chatbots can help people with diabetes manage their diet, exercise, medication, blood glucose monitoring, prevention of complications, and mental health to control blood glucose levels and improve quality of life [33-35,43-45]. In addition, chatbots also have a significant advantage in out-of-hospital diabetes management. Chatbots can save many human medical resources, reducing the work pressure on primary care staff and the burden on the health care system [52]. Therefore, chatbots with AI technology will play a prominent role in diabetes health education and self-management, which is in line with the trend of digital therapy.

Based on the analysis of the effectiveness assessment indicators (technical performance indicators, user experience, and user metrics) of the chatbot, it was found that the technical assessment of the chatbot model was good [6], the overall acceptance of the chatbot by the users was high [29], and the health status of the users improved after using the chatbot for self-management [36,38,41,42]. Overall, the use of chatbots in health education and self-management of diabetic patients is effective.

Lack of High-Quality Research Evidence

Identifying the best evidence is an important part of evidence-based medicine. Evidence grading is an essential tool to help decision-makers obtain the best evidence. Although the criteria and methods for grading research evidence are not uniform, there is some grading of evidence that is generally accepted by scientists. These gradings have consistently endorsed meta-analyses and randomized controlled trials (RCTs) as the highest level of evidence [53]. Most of the articles in our study were system design studies and pilot studies that were in the primary and secondary research phases, while fewer intervention studies were in the analysis. Most of the studies in the system design category tested the model performance of chatbots through algorithms and techniques. However, the actual effectiveness of chatbot applications in the self-management of patients with diabetes has yet to be verified at this stage. In the pilot studies, the number of respondents was small (n=10-33), the duration of the intervention was short (from short-term measurements to 24 weeks), and the respondents included older adults and the general public in addition to people with diabetes. Most of the intervention research articles took a pre-post trial design, and only 1 document used an RCT design with the highest level of evidence. There is still a lack of higher level studies, especially RCTs, to evaluate the effectiveness of chatbots in diabetes self-management. RCTs are important for evaluating the effectiveness of chatbots in diabetes self-management. Therefore, more high-quality studies, such as RCTs, should be conducted in the future to evaluate the effectiveness of chatbots in diabetes self-management, providing stronger evidence that will contribute to the development of the field.

Single Method of Research

Most of the articles included for analysis in our study used quantitative research methods. The data sources mainly relied on data collected from chatbots and user questions and answers, as well as user ratings of the experience, satisfaction, and usability of chatbots collected from questionnaires. In contrast, only a minority of articles used mixed research methods (MMRs) and qualitative research methods (semistructured interviews, focus group discussions, etc). For example, of the pilot studies, most used quantitative research methods, with

XSL•FO
RenderX

only 2 MMR studies and 1 qualitative study. The quantitative studies were mainly based on questionnaires and data collected from chatbots, while MMR studies included interviews with users on top of that. MMR has unique advantages [54]. The original purpose of the MMR was to bridge the gap between the quantitative and qualitative paradigms of analysis, combining the strengths of both to allow a broader and deeper analysis of specific issues. Quantitative research is objective and reproducible but does not facilitate the researcher's observation and communication with the research participants. Qualitative research allows the researcher to observe and communicate with the research participants in a close and more natural setting. Qualitative research makes it easier for the researcher to observe and understand the research participants' behavior, attitudes, and motivations from the participant's perspective, and the research design is more flexible. Adding qualitative research to quantitative research allows researchers to better understand what respondents think about chatbots. Respondents may raise issues that the researcher has not noticed and considered, helping the researcher to improve the research design. Through qualitative research, the researcher can fully play into the motivation of the researcher and interviewees by conducting semistructured interviews and focus group discussions with interviewees. The interview questions centered on the features that users want chatbots to have or their feelings when using chatbots. After the interviews, the researcher organizes the collected data and information and analyzes them to gain an in-depth understanding of the users' experiences, perceptions, and barriers to the use of chatbots so as to provide assistance for subsequent research. Therefore, more MMRs should be used in future research on chatbots and diabetes self-management. By combining the respective strengths and unique perspectives of qualitative and quantitative, a more comprehensive understanding of the impact of chatbots on diabetes self-management can be achieved.

Insufficiently Innovative Theoretical Framework

In our study, it was found that only a minority of studies had a theoretical framework to support them, while most studies did not have a theoretical framework. For example, only 4 of the pilot studies were guided by a theoretical framework. This phenomenon suggests that researchers have not paid enough attention to the theoretical framework. Theories are developed to explain, predict, and understand phenomena. A scientific theoretical framework actively supports and guides research design and experimentation and can inform and guide the selection of research types. A theoretical framework helps the researcher synthesize prior empirical findings within a theoretical framework, clarifying what the concepts are and how they relate to each other. It provides guidance for research at the methodological level. It helps the researcher to build bridges between theories, facilitates the researcher to recognize their own research, and discovers connections to the established body of knowledge.

The main theoretical frameworks used in the field of diabetes self-management and chatbots include Behavior Change Theory, Self-Determination Theory, Cross-Theoretical Models, and Social Cognitive Theory. These theoretical frameworks are relatively well-developed and widely used. Applying a theoretical framework can provide a theoretical foundation for research on chatbots in diabetes self-management. The use of appropriate theoretical frameworks can enhance understanding of the underlying mechanisms between chatbots and diabetes self-management and facilitate the development of effective interventions. Therefore, future research is expected to use more novel theoretical frameworks in studies related to chatbots and diabetes self-management. Or innovatively applying theoretical frameworks from other fields to enrich research in the field of chatbots and diabetes self-management (see Multimedia Appendix 3 for specific explanations and applications of the theoretical frameworks in detail).

Equitable Accessibility and Privacy Protection

In addition, although patients with diabetes and those at risk of developing diabetes are getting younger, older adults still account for a high proportion of patients with diabetes.

Chatbots, a new-era product of mobile health and AI, should consider the needs of different age groups in terms of system design and software interface usage. Especially for older adults, their acceptance of the new system, their learning level, whether the software interface is easy to use, whether the font size is comfortable, and whether the answers are easy to understand need special consideration. In order to ensure that digital health tools such as chatbots are accessible and beneficial to everyone, it is crucial to include diverse populations in future research.

At the same time, chatbots collect a large amount of health and personal data from users. If this information is disclosed, it will violate patients' privacy rights. Lack of privacy protection will affect users' trust. Chatbots can earn users' trust through dialog and prompt them to reveal more personal data [55]. Therefore, it is crucial to safeguard patients' personal privacy. First, regulators must act to ensure user safety and privacy. For example, on March 13, 2024, the AI Bill was passed in the European Union (EU) Parliament and will soon become EU law. The EU's AI Bill will require high-risk AI systems to be assessed and monitored before approval [56]. Second, operators should standardize the management of data usage rights of back-office personnel and strengthen core data positions and departmental prevention. Third, users should be clearly informed of the privacy policy. The privacy policy must be dynamic, and the privacy policy must provide a table of contents to facilitate users' understanding of the content of the privacy policy. The collection of users' personal information shall be based on the principles of clear purpose, explicit consent, and collection of the least necessary data, as well as a detailed description of the type of information to be collected, its purpose, and the manner in which it is to be collected. In addition, operators need to give the correct download uniform resource locator or store. According to their own services, they target the older adults and minors for safety knowledge popularization. When sharing and using personal information, it needs to be deidentified or anonymized so that the subject of the personal information cannot be identified [57].

Strengths and Limitations

This study is the first systematic review of chatbots and self-management for people with diabetes. Researchers searched

XSL•FO
RenderX

for relevant literature from 2 authoritative databases, Web of Science and PubMed. The researchers analyzed and summarized the specific features of chatbots in the field of diabetes, as well as information on the research methodology, research design, and theoretical framework used in the articles. In addition, the articles included from different countries may increase the generalizability of our findings.

However, this article also has some limitations. First, researchers selected more authoritative and comprehensive databases, but some gray literature was not retrieved. Second, in the process of data extraction, the key information extracted was confirmed by experts and extracted and cross-checked by 2 researchers independently, but there was still a possibility of human error. There is some literature that fits the research theme very well, but it was not included in our study because full-text information was not available. Only English-language studies were included in this review, which may lead to a certain degree of publication bias. Finally, the trials in the meta-analysis were all pre-post trials with low evidence validity. Therefore, the effectiveness of chatbot interventions remains to be validated by conducting trials.

Conclusions

This study analyzes and summarizes information on the specific characteristics of chatbots in the field of self-management of patients with diabetes, such as research methods, research strategies, effectiveness evaluation metrics, etc, and provides suggestions for research related to chatbots and the self-management of patients with diabetes. The number of articles published in each country was low, and chatbots were at an early stage of development in the field of self-management of patients with diabetes. Chatbots provide advice and education focused on the physical aspects of patients, and researchers have focused less on the psychological aspects of patients. A number of positive effects of chatbots in supporting the self-management of people with diabetes. The overall acceptance of chatbots among people with diabetes was high. However, most of the current studies were at the stage of system design and pilot studies. Meta-analysis showed that the chatbot intervention was effective in lowering blood glucose but had no significant effect on weight reduction.

Based on the results derived from this study, we make the following appeal to researchers: (1) focus should be placed on the area of chatbots and diabetes self-management, with research to fully demonstrate the effectiveness of chatbot interventions; (2) future studies should address identified research gaps and use higher level evidence, such as RCTs; (3) in the choice of research strategies, more mixed research should be adopted to fully use the advantages of quantitative and qualitative research; (4) use of appropriate and innovative theoretical frameworks in the study to provide theoretical support for the study; and (5) the digital age requires the protection of users' private data. It is hoped that researchers will continue to innovate and enrich the research in this field, which will help to continuously develop effective intervention measures of chat robots in diabetes self-management, and ultimately help patients with diabetes stay healthy.

Acknowledgments

Thanks to the authors of the articles that included analysis in our research. Their articles laid the foundation for our research. The research for this article is funded by the National Key Research and Development Program (2022YFC3600904) and the Capital Health Development Research Project (2022-1G-4252).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YBW involved in conceptualization, data curation, methodology, project administration, and writing – review & editing; JZZ involved in investigation, methodology, writing – original draft, and writing – review & editing; PG, TYD, JYZ, YWW, YNZ, SYL, XYL, and EYW involved in investigation and writing – review & editing; and XYS involved in project administration, supervision, and writing – review & editing. All authors contributed to the article and approved the submitted version. All authors ensure that generative AI was not used in the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 49 KB - jmir_v26i1e60380_app1.docx]

Multimedia Appendix 2
Analysis of the advantages and disadvantages of quantitative, qualitative and mixed studies.
[DOCX File , 14 KB - jmir_v26i1e60380_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Overview of theoretical frameworks and specific applications to chatbots and diabetes.
[DOCX File , 18 KB - jmir_v26i1e60380_app3.docx]

Multimedia Appendix 4
Research type.
[DOCX File , 17 KB - jmir_v26i1e60380_app4.docx]

Multimedia Appendix 5
PRISMA 2020 checklist.
[DOCX File , 30 KB - jmir_v26i1e60380_app5.docx]

References

1. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023;402(10397):203-234 [FREE Full text] [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
2. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022;183:109119 [FREE Full text] [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
3. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S20-S42. [doi: 10.2337/dc24-S002] [Medline: 38078589]
4. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 7. Diabetes technology: standards of care in diabetes-2023. Diabetes Care 2023;46(Suppl 1):S111-S127 [FREE Full text] [doi: 10.2337/dc23-S007] [Medline: 36507635]
5. Association of Diabetes CareEducation Specialists, Kolb L. An effective model of diabetes care and education: the ADCES7 self-care behaviors™. Sci Diabetes Self Manag Care 2021;47(1):30-53. [doi: 10.1177/0145721720978154] [Medline: 34078208]
6. Xie W, Ding R, Yan J, Qu Y. A mobile-based question-answering and early warning system for assisting diabetes management. Wirel Commun Mob Comput 2018;2018(s1):1-14 [FREE Full text] [doi: 10.1155/2018/9163160]
7. Chrvala CA, Sherr D, Lipman RD. Diabetes self-management education for adults with type 2 diabetes mellitus: a systematic review of the effect on glycemic control. Patient Educ Couns 2016;99(6):926-943 [FREE Full text] [doi: 10.1016/j.pec.2015.11.003] [Medline: 26658704]
8. Chatterjee S, Davies MJ, Heller S, Speight J, Snoek FJ, Khunti K. Diabetes structured self-management education programmes: a narrative review and current innovations. Lancet Diabetes Endocrinol 2018;6(2):130-142. [doi: 10.1016/S2213-8587(17)30239-5] [Medline: 28970034]
9. Kim M, Patrick K, Nebeker C, Godino J, Stein S, Klasnja P, et al. The digital therapeutics real-world evidence framework: an approach for guiding evidence-based digital therapeutics design, development, testing, and monitoring. J Med Internet Res 2024;26:e49208 [FREE Full text] [doi: 10.2196/49208] [Medline: 38441954]
10. Hong JS, Wasden C, Han DH. Introduction of digital therapeutics. Comput Methods Programs Biomed 2021;209:106319. [doi: 10.1016/j.cmpb.2021.106319] [Medline: 34364181]
11. Kaufman N. Digital therapeutics: leading the way to improved outcomes for people with diabetes. Diabetes Spectr 2019;32(4):301-303 [FREE Full text] [doi: 10.2337/ds19-0012] [Medline: 31798286]
12. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RTS, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. Ann Med 2024;56(1):2302980. [doi: 10.1080/07853890.2024.2302980] [Medline: 38466897]
13. McGreevey JD, Hanson CW, Koppel R. Conversational agents in health care-reply. J Am Med Assoc 2020;324(23):2444-2445. [doi: 10.1001/jama.2020.21518] [Medline: 33320218]
14. d'Elia A, Gabbay M, Rodgers S, Kierans C, Jones E, Durrani I, et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Fam Med Community Health 2022;10(Suppl 1):e001670 [FREE Full text] [doi: 10.1136/fmch-2022-001670] [Medline: 36450391]
15. Li L, Jiang B, Sun L. HNF1A：from monogenic diabetes to type 2 diabetes and gestational diabetes mellitus. Front Endocrinol (Lausanne) 2022;13:829565 [FREE Full text] [doi: 10.3389/fendo.2022.829565] [Medline: 35299962]
16. Park D, Jeong S, Seo Y. Systematic review on chatbot techniques and applications. J Inf Procress Syst 2022;18(1):26-47 [FREE Full text]
17. Bickmore TW, Kimani E, Trinh H, Pusateri A, Paasche-Orlow MK, Magnani JW. Managing chronic conditions with a smartphone-based conversational virtual agent. 2018 Presented at: Proceedings of the 18th International Conference on Intelligent Virtual Agents; November 05, 2018; Sydney, Australia p. 119-124 URL: https://dl.acm.org/doi/abs/10.1145/3267851.3267908

XSL•FO
RenderX

18. Pereira J, Díaz Ó. Using health chatbots for behavior change: a mapping study. J Med Syst 2019;43(5):135. [doi: 10.1007/s10916-019-1237-1] [Medline: 30949846]
19. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. J Am Med Assoc 2023;330(4):309-310. [doi: 10.1001/jama.2023.9458] [Medline: 37410450]
20. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
21. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
22. Marcuzzi A, Nordstoga AL, Bach K, Aasdahl L, Nilsen TIL, Bardal EM, et al. Effect of an artificial intelligence-based self-management app on musculoskeletal health in patients with neck and/or low back pain referred to specialist care: a randomized clinical trial. JAMA Netw Open 2023;6(6):e2320400 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.20400] [Medline: 37368401]
23. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
24. Alsahli M, Abd-Alrazaq A, Househ M, Konstantinidis S, Blake H. The effectiveness of mobile phone messaging-based interventions to promote physical activity in type 2 diabetes mellitus: systematic review and meta-analysis. J Med Internet Res 2022;24(3):e29663 [FREE Full text] [doi: 10.2196/29663] [Medline: 35258463]
25. Jiang Z, Huang X, Wang Z, Liu Y, Huang L, Luo X. Embodied conversational agents for chronic diseases: scoping review. J Med Internet Res 2024;26:e47134 [FREE Full text] [doi: 10.2196/47134] [Medline: 38194260]
26. Shu X, Mai Q, Blatz M, Price R, Wang XD, Zhao K. Direct and indirect restorations for endodontically treated teeth: a systematic review and meta-analysis, IAAD 2017 consensus conference paper. J Adhes Dent 2018;20(3):183-194. [doi: 10.3290/j.jad.a40762] [Medline: 29984369]
27. Deeks JJ, Higgins JP, Altman DG. Chapter 9: analysing dataundertaking meta-analyses. In: Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Sussex, UK: John Wiley & Sons; 2008:243-296.
28. Krishnakumar A, Verma R, Chawla R, Sosale A, Saboo B, Joshi S, et al. Evaluating glycemic control in patients of South Asian origin with type 2 diabetes using a digital therapeutic platform: analysis of real-world data. J Med Internet Res 2021;23(3):e17908 [FREE Full text] [doi: 10.2196/17908] [Medline: 33764306]
29. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through tess, the AI behavioral coaching chatbot. Transl Behav Med 2019;9(3):440-447. [doi: 10.1093/tbm/ibz043] [Medline: 31094445]
30. Sharma M, Singh G, Singh R. An advanced conceptual diagnostic healthcare framework for diabetes and cardiovascular disorders. EAI Endorsed Trans Scalable Inf Syst 2018;5(18):154828 [FREE Full text]
31. Sowah RA, Bampoe-Addo AA, Armoo SK, Saalia FK, Gatsi F, Sarkodie-Mensah B. Design and development of diabetes management system using machine learning. Int J Telemed Appl 2020;2020:8870141 [FREE Full text] [doi: 10.1155/2020/8870141] [Medline: 32724304]
32. Anastasiadou M, Alexiadis A, Polychronidou E, Votis K, Tzovaras D. A prototype educational virtual assistant for diabetes management. 2020 Presented at: IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE); October 26-28, 2020; Cincinnati, OH p. 999-1004 URL: https://ieeexplore.ieee.org/abstract/document/9288129
33. Beaudry J, Consigli A, Clark C, Robinson KJ. Getting ready for adult healthcare: designing a chatbot to coach adolescents with special health needs through the transitions of care. J Pediatr Nurs 2019;49:85-91. [doi: 10.1016/j.pedn.2019.09.004] [Medline: 31644960]
34. Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. human health coaching: exploring participant and practitioner experiences. Proc ACM Hum Comput Interact 2021;5(CSCW1):99 [FREE Full text] [doi: 10.1145/3449173] [Medline: 36304916]
35. Hossain E, Alshehri M, Almakdi S, Halawani H, Rahman MM, Rahman W, et al. Dm-Health app: diabetes diagnosis using machine learning with smartphone. Comput Mater Continua 2022;72(1):1713-1746 [FREE Full text]
36. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. JMIR Mhealth Uhealth 2020;8(7):e17558 [FREE Full text] [doi: 10.2196/17558] [Medline: 32673246]
37. Sagstad MH, Morken N, Lund A, Dingsør LJ, Nilsen ABV, Sorbye LM. Quantitative user data from a chatbot developed for women with gestational diabetes mellitus: observational study. JMIR Form Res 2022;6(4):e28091 [FREE Full text] [doi: 10.2196/28091] [Medline: 35436213]
38. Mash R, Schouw D, Fischer AE. Evaluating the implementation of the GREAT4Diabetes WhatsApp chatbot to educate people with type 2 diabetes during the COVID-19 pandemic: convergent mixed methods study. JMIR Diabetes 2022;7(2):e37882 [FREE Full text] [doi: 10.2196/37882] [Medline: 35537057]

XSL•FO
RenderX

39. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. Acceptability of an embodied conversational agent for type 2 diabetes self-management education and support via a smartphone app: mixed methods study. JMIR Mhealth Uhealth 2020;8(7):e17038 [FREE Full text] [doi: 10.2196/17038] [Medline: 32706734]
40. Hurmuz MZM, Jansen-Kosterink SM, Op den Akker H, Hermens HJ. User experience and potential health effects of a conversational agent-based electronic health intervention: protocol for an observational cohort study. JMIR Res Protoc 2020;9(4):e16641 [FREE Full text] [doi: 10.2196/16641] [Medline: 32242517]
41. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My diabetes coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]
42. Roca S, Lozano ML, García J, Alesanco Á. Validation of a virtual assistant for improving medication adherence in patients with comorbid type 2 diabetes mellitus and depressive disorder. Int J Environ Res Public Health 2021;18(22):12056 [FREE Full text] [doi: 10.3390/ijerph182212056] [Medline: 34831811]
43. Pimenta N, Félix IB, Monteiro D, Marques MM, Guerreiro MP. Promoting physical activity in older adults with type 2 diabetes an anthropomorphic conversational agent: development of an evidence and theory-based multi-behavior intervention. Front Psychol 2022;13:883354 [FREE Full text] [doi: 10.3389/fpsyg.2022.883354] [Medline: 35903740]
44. Balsa J, Félix I, Cláudio AP, Carmo MB, Silva ICE, Guerreiro A, et al. Usability of an intelligent virtual assistant for promoting behavior change and self-care in older people with type 2 diabetes. J Med Syst 2020;44(7):130. [doi: 10.1007/s10916-020-01583-w] [Medline: 32533367]
45. Félix IB, Guerreiro MP, Cavaco A, Cláudio AP, Mendes A, Balsa J, et al. Development of a complex intervention to improve adherence to antidiabetic medication in older people using an anthropomorphic virtual assistant software. Front Pharmacol 2019;10:680 [FREE Full text] [doi: 10.3389/fphar.2019.00680] [Medline: 31281256]
46. Saritha AK. Medibot A predictive generic diabetic chatbot using bagging ensemble hybrid learning. Int J Eng Adv Technol 2020;9(4):2249-8958 [FREE Full text]
47. Buinhas S, Cláudio AP, Carmo MB, Balsa J, Cavaco A, Mendes A, et al. Virtual assistant to improve self-care of older people with type 2 diabetes: first prototype. In: Gerontechnology. Cham, Switzerland: Springer; 2019:236-248.
48. Dhinagaran DA, Sathish T, Soong A, Theng Y, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. JMIR Form Res 2021;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]
49. Dhinagaran DA, Car LT. Public perceptions of a healthy lifestyle change conversational agent in Singapore: a qualitative study. Digit Health 2022;8:20552076221131190 [FREE Full text] [doi: 10.1177/20552076221131190] [Medline: 36267545]
50. Hussain S, Athula G. Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. 2018 Presented at: 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA); May 16-18, 2018; Krakow, Poland p. 698-703 URL: https://ieeexplore.ieee.org/abstract/document/8418155
51. Rehman UU, Chang DJ, Jung Y, Akhtar U, Razzaq MA, Lee S. Medical instructed real-time assistant for patient with glaucoma and diabetic conditions. Appl Sci 2020;10(7):2216. [doi: 10.3390/app10072216]
52. Andrès E, Meyer L, Zulfiqar A, Hajjam M, Talha S, Bahougne T, et al. Telemonitoring in diabetes: evolution of concepts and technologies, with a focus on results of the more recent studies. J Med Life 2019;12(3):203-214 [FREE Full text] [doi: 10.25122/jml-2019-0006] [Medline: 31666818]
53. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
54. Johnson RB, Onwuegbuzie AJ. Mixed methods research: a research paradigm whose time has come. Educ Res 2004;33(7):14-26. [doi: 10.3102/0013189X033007014]
55. The Lancet. AI in medicine: creating a safe and equitable future. Lancet 2023;402(10401):503. [doi: 10.1016/S0140-6736(23)01668-9] [Medline: 37573071]
56. Gilbert S. The EU passes the AI Act and its implications for digital medicine are unclear. NPJ Digit Med 2024;7(1):135 [FREE Full text] [doi: 10.1038/s41746-024-01116-6] [Medline: 38778162]
57. Liu Q, Liu H, Qin Z, Ma C. Research on users' privacy protection policy of m-health application based on content analysis. Chin Hosp 2019;23(09):20-23.

Abbreviations

AI: artificial intelligence
EU: European Union
HbA1c: hemoglobin A1c
MD: mean difference
MMR: mixed research method
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

RCT: randomized controlled trial

Edited by X Ma; submitted 09.05.24; peer-reviewed by K Rawlings, K Obaideen; comments to author 16.07.24; revised version received 14.08.24; accepted 14.10.24; published 03.12.24.

Please cite as:
Wu Y, Zhang J, Ge P, Duan T, Zhou J, Wu Y, Zhang Y, Liu S, Liu X, Wan E, Sun X
Application of Chatbots to Help Patients Self-Manage Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e60380
URL: https://www.jmir.org/2024/1/e60380
doi:10.2196/60380
PMID:

©Yibo Wu, Jinzi Zhang, Pu Ge, Tingyu Duan, Junyu Zhou, Yiwei Wu, Yuening Zhang, Siyu Liu, Xinyi Liu, Erya Wan, Xinying Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Reviewing the Evidence Base for Topical Steroid Withdrawal Syndrome in the Research Literature and Social Media Platforms: An Evidence Gap Map

Noreen Orr[1], PhD; Morwenna Rogers[1], MSc; Abigail Stein[1], MSc; Jo Thompson Coon[1], PhD; Kenneth Stein[1†], MD

University of Exeter Medical School, Exeter, United Kingdom

[†]deceased

Corresponding Author:
Noreen Orr, PhD
University of Exeter Medical School
South Cloisters
St Luke's Campus
Exeter, EX1 1TE
United Kingdom
Phone: 44 01392 726056
Email: N.Orr@exeter.ac.uk

Abstract

Background: Within the dermatological community, topical steroid withdrawal syndrome (TSWS) is a medically contested condition with a limited research base. Published studies on TSWS indicate that it is a distinct adverse effect of prolonged use of topical corticosteroids, but there is a paucity of high-quality research evidence. Among the "patient community," awareness has been increasing, with rapid growth in social media posts on TSWS and the introduction of online communities such as the International Topical Steroid Awareness Network. This evidence gap map (EGM) was developed in response to recent calls for research to better understand TSWS and aims to be an important resource to guide both researchers and clinicians in the prioritization of research topics for further research.

Objective: This study aims to identify the range, extent, and type of evidence on TSWS in the research literature and social media platforms using an EGM.

Methods: The MEDLINE and Embase (Ovid), CINAHL (EBSCOhost), and ProQuest Dissertations & Theses and Conference Proceedings Citation Index (CPCI-Science and CPCI-Social Science & Humanities via Web of Science) databases were searched. The final search was run in November 2023. Study titles, abstracts, and full texts were screened by 2 reviewers, and a third was consulted to resolve any differences. Blogging sites WordPress, Medium, and Blogspot and Google were searched; Instagram and Reddit were searched for the 100 most recent posts on specific dates in February 2023. Blog titles, Instagram posts, and Reddit posts were screened for relevance by 2 reviewers. A data extraction tool was developed on EPPI-Reviewer, and data extraction was undertaken by one reviewer and checked by a second; any inconsistencies were resolved through discussion. We did not undertake quality appraisal of the included studies. EPPI-Reviewer and EPPI-Mapper were used to generate the interactive EGM.

Results: Overall, 81 academic publications and 223 social media posts were included in the EGM. The research evidence mainly addressed the physical symptoms of TSWS (skin), treatments, and, to a lesser extent, risk factors and disease mechanisms. The social media evidence primarily focused on the physical symptoms (skin and nonskin), mental health symptoms, relationships, activities of everyday living, beliefs and attitudes, and treatments.

Conclusions: The EGM shows that research evidence is growing on TSWS but remains lacking in several important areas: longer-term prospective observational studies to assess the safety of prolonged use of topical corticosteroids and to prevent addiction; qualitative research to understand the lived experience of TSWS; and longitudinal research on the patient's "TSWS journey" to healing. The inclusion of social media evidence is a methodological innovation in EGMs, recognizing the increased presence of #topicalsteroidwithdrawal on social media and how it can be used to better understand the patient perspective and ultimately, provide better care for people with TSWS.

(J Med Internet Res 2024;26:e57687) doi:10.2196/57687

XSL•FO
RenderX

KEYWORDS

topical steroid withdrawal syndrome; evidence gap map; social media; blogs; Instagram; Reddit; topical corticosteroids

Introduction

Background

Topical corticosteroids (TCs) are prescribed for many skin disorders, including chronic conditions such as atopic dermatitis, the most common form of eczema. That they have been the mainstay of treating atopic dermatitis for >60 years is reflected in the many international guidelines published for managing atopic dermatitis [1]. TCs are included in these guidelines as one of the several treatment options, with recommendations provided for dosing, frequency of application, maintenance, and screening for side effects. In a comparison of 14 international guidelines from across the globe published between 2007 and 2018, LePoidevin et al [1] found that the fingertip unit (approximately 0.5 g) as the method to guide TC application was recommended in over half. Most guidelines recommended once to twice daily application to control flares, and most recommended using TCs as maintenance therapy to reduce the incidence of flares. Just over half recommended screening for cutaneous side effects, and none of them recommended screening for systemic side effects of TCs. Despite the variation across the guidelines, importantly, all acknowledged that the prescription of TCs should be tailored to the unique characteristics of individual patients [1].

While TCs have been effective for many people [2-4], they have also been associated with the risk of adverse effects, particularly when used long term [3]. Cutaneous side effects include atrophy, striae, rosacea, perioral dermatitis, acne, and purpura [2,4-7], and systemic side effects include adrenal insufficiency, Cushing syndrome, glaucoma, and cataract [2,3,6,7]. In the recently updated guidelines from the American Academy of Dermatology for the management of adult atopic dermatitis with topical therapies, the “related concepts” of topical steroid addiction (TSA) and topical steroid withdrawal (TSW) are mentioned, along with red face syndrome and red scrotum syndrome, as adverse side effects that may occur after prolonged use of TCs [4]. The European Guideline (EuroGuiDerm) on atopic eczema also has 1 mention of “corticosteroid addiction syndrome” as an important side effect of TCs [7].

The concepts of TSA and TSW are often used interchangeably in the literature [8,9], but they coalesce around 2 elements: physical dependence on TCs and worsening skin symptoms after withdrawal [9-11]. Studies have shown that people use an increased quantity and potency of TCs for months to years to control their skin condition (it is not clear whether the decision to increase the quantity and potency of TCs was made by the individuals themselves or following advice from a health professional) [11,12]. Then, on withdrawal from TCs, individuals experience the rebound phenomenon, often with more extensive and more severe skin manifestations than with the original skin condition [9,13]. An example of a withdrawal symptom is red skin syndrome, characterized by severe erythema in various areas of the body, along with burning and stinging [11]. Withdrawal from TCs can mean complete cessation or tapering use by reducing potency and applying less often and to fewer parts of the body. Following the International Topical Steroid Awareness Network, we use the term topical steroid withdrawal syndrome (TSWS), which is an “umbrella term” for the condition, as it encompasses symptoms before (TSA) and after cessation of TCs (red skin syndrome and TSW).

Despite the recognition of the “addiction” effects of TCs in 1969 [14], TSWS is controversial today [9,15], as “doubts still exist whether this condition is legitimate” [16]. There are no clear or accepted diagnostic criteria for TSWS and for many clinicians, it can be challenging to distinguish it from other conditions (eg, allergic contact dermatitis) [9,17], and some see it as an “exacerbation of the underlying skin disease” [18]. That said, in 2021, the National Eczema Society and the British Association of Dermatologists [19] published a joint statement on TSWS, and in the same year, the Medicines and Healthcare products Regulatory Agency (MHRA) issued a drug safety update advising patients to seek medical help if experiencing redness, burning, itching, or stinging of the skin on stopping TCs [20]. The MHRA requested that warnings and precautions for use on TCs be added to the summary of product characteristics and patient information leaflet for prescription-only and pharmacy-only medicines. The National Eczema Association in the United States has also recognized TSWS as a potentially serious side effect of using TCs [21], as has DermNet in New Zealand [22].

Although the MHRA described TSWS as rare, it is estimated to occur in approximately 12% of people with atopic dermatitis who use TCs [13]. Evidence for the increasing awareness of TSWS is demonstrated by the rapid growth in social media posts and discussions on TSWS [23-25] and the introduction of online communities, such as the International Topical Steroid Awareness Network and Scratch That, United Kingdom. It suggests that people living with TSWS are using social media for information and support, particularly as patients have reported feeling ignored, dismissed, or blamed by their physicians for inappropriate use of TCs [26,27]. However, for many dermatologists and clinicians, social media is perceived negatively as a key source of misinformation [15,28] and a driver of patient self-diagnosis [29]. In contrast, Bowe et al [24] suggested that an awareness of the influence of social media platforms on perceptions of TCs could help “bridge the doctor-patient gap”, and arguably, it offers one way to gain a greater understanding of people’s experiences of living with the symptoms of TSWS: “[p]atients want professionals to recognise the impact of symptoms experienced and feel that recognition would be an important first step to better management” [27].

Aim

This study aimed to identify the range, extent, and type of evidence on TSWS in the research literature and compare the topics identified with the subjects of posts on a number of social media platforms. The overall objectives of the study were as follows:

XSL•FO
RenderX

- Identify published evidence on TSWS: reviews and primary research (both quantitative and qualitative)
- Identify evidence on the lived experience of TSWS in blogs and on posts on Instagram (Meta Platforms, Inc) and Reddit (Reddit, Inc)
- Identify gaps in evidence where further primary research is needed.

Methods

Design

An evidence gap map (EGM) is a systematic evidence synthesis product [30] that provides an overview of the available evidence on a particular topic, theme, or policy area. EGMs are used to highlight gaps in the evidence base, show where there is an abundance of evidence, and increase the discoverability and use of the evidence. EGM methodology seeks to show what evidence is available, not what the evidence says. As part of the "Big Picture" review family, Campbell et al [31] assert that "[n]o other review methodology has developed a systematic approach to identifying gaps in the evidence with this level of rigor and transparency." This EGM aimed to identify the range, extent, and type of evidence about TSWS both in the research literature and on social media platforms. The scope of the EGM was defined by a years–research topics framework, with the rows as years and the columns as research topics. The framework was developed by examining the literature, drawing on the building blocks proposed by Howell et al [27] for high-quality research in TSW, and by consulting with our public collaborator (AS) who has lived experience of TSWS. Additional topics were added if identified within the included evidence. The social media evidence from Instagram and Reddit was restricted to 1 year, reflecting our decision to select a sample of the most recent posts using the popular hashtag #topicalsteroidwithdrawal on specific dates. This EGM has been conducted and reported according to Campbell guidelines for EGMs [30], and the EGM protocol was uploaded to Open Research Exeter, the web-based repository of the University of Exeter [32].

Ethical Considerations

We carefully considered the ethical implications of this study and did not submit it for institutional review board approval because the study involved publicly available data and no analysis was undertaken.

Patient and Public Involvement

Our public collaborator (AS), who has lived experience of TSWS, contributed throughout the process, from question development to manuscript preparation.

Search Methods

Searches were conducted to identify published and unpublished literature via academic databases, websites, blogs, and social media (Instagram and Reddit).

Published and Unpublished Research

We searched MEDLINE and Embase (via Ovid), CINAHL (via EBSCOhost), and ProQuest Dissertations & Theses and Conference Proceedings Citation Index (CPCI-Science and CPCI-Social Science & Humanities via Web of Science) between August 17 and 26, 2022, using a combination of subject headings and free-text terms for steroids, topical application, and terms for withdrawal or addiction. The database searches were updated in November 2023. The full search strategies are available in Multimedia Appendix 1. We carried out forward and backward citation chasing of included studies using Epistemonikos (Epistemonikos Foundation; for reviews) and Scopus (Elsevier). We searched for the phrases "topical steroid withdrawal" and "topical steroid addiction" on Google using the approach recommended by Briscoe and Rogers [33] for additional relevant studies. All results were exported into EndNote (version 20; Clarivate) for screening.

Blogs

We searched for blogs using "topical steroid withdrawal" and "topical steroid addiction" in combination with the word "blogs" on Google. We also carried out a targeted search for "topical steroid withdrawal" on the blogging sites WordPress, Medium, and Blogspot. Blog searches were carried out in October 2022. There were no date or other limits applied to the blog searches. Searching for blogs as opposed to blog posts offered the possibility of understanding the experience of TSWS from a longitudinal perspective and following individuals through the withdrawal process [34].

Instagram and Reddit

We searched for the 100 most recent posts featuring #TopicalSteroidWithdrawal that included text on February 2, 2023. These were copied into an Excel (Microsoft Corp) spreadsheet for screening and coding, along with the date of the post and a URL link. Posts not relating to TSW, posts advertising products, posts that were not in English, or posts that only included links or hashtags were excluded. We searched for the 100 most recent and relevant posts under the subreddit Topical Steroid Withdrawal (r/TS_Withdrawal). The text, date, and URL were entered into an Excel spreadsheet for screening and coding. Searches on Reddit were carried out on February 7 and 8, 2023. Searching for the most recent and relevant posts on specific dates was chosen to simulate "real-world" Instagram and Reddit viewing behavior [35,36].

Inclusion Criteria

Studies, blogs, and social media posts were included if they met the following criteria:

- Population: people experiencing TSWS (includes infants, children, and adolescents), either via complete "sudden" cessation of TCs or complete cessation of TCs via a "tapering" approach
- Exposure: TCs
- Outcomes: effects of withdrawal: physical, psychosocial, knowledge, and attitudes (among both health care professionals and people experiencing withdrawal) and information seeking and sharing by people experiencing withdrawal
- Study type/posts: any studies (quantitative or qualitative) that investigate TSW were included. Letters, case reports, commentaries, and opinion pieces were also included. Blogs

XSL•FO
RenderX

or social media posts were included if they described any aspect of the experience of living with TSWS or caring for someone with TSWS.

Exclusion Criteria

Studies were excluded if they reported side effects of corticosteroids that were not specific to withdrawal. Blogs and social media posts were excluded if they did not describe personal experience relating to TSWS. They were also excluded if they described and promoted treatments for TSWS that were not related to personal experience, for example, posts advertising products.

Study Selection

Studies located via the database searches were exported into EndNote (version 9). Titles, abstracts, and full texts were screened independently by 2 reviewers (NO and MR). Disagreements were resolved by discussion. Posts from Instagram and Reddit were copied and pasted into an Excel spreadsheet and were screened independently by 2 reviewers. Details about blogs (URL, date coverage, number of posts, and usernames) were pasted into an Excel spreadsheet. Short summaries of the blogs were compiled and checked by 2 reviewers (NO and MR). Decisions about the inclusion of blogs were carried out during the data extraction and coding stage.

Data Extraction and Mapping

Data extraction was conducted using EPPI-Reviewer [37]. The data extraction tool was informed by the research question and the structure of the map; it was piloted on a sample of included academic publications and social media posts and modified through team discussion. Data extraction was undertaken by 1 reviewer (NO) and checked by a second reviewer (MR), and any inconsistencies were resolved through discussion. Data were extracted on the type of evidence, location of evidence, type of review, study design, population (age), steroids used, duration of TC use, length of time “off” TCs, areas of body affected, reasons for stopping TCs, diagnosis, names of conditions, and research topics. Following best practice guidelines for conducting EGMs, we did not undertake quality appraisal [31].

EPPI-Reviewer [37] and EPPI-Mapper software [38] were used to generate the interactive EGM. The EGM uses a years–research topics framework (years as rows and research topics as columns), and within each cell of the grid, the evidence is presented as bubbles according to type, with the color and size of the bubble indicating the type and amount of evidence available within that cell of the grid, respectively. The filters of the map enable map users to change the type of evidence displayed based on, for example, authors, population age, and the location of evidence. Map users are also able to download lists of the studies from the map in total or by study type. Due to the nature of the included studies (ie, a study may comprise a review of the literature and a case report), a study may be represented in multiple places in the EGM.

Results

Results of the Search

The database searches identified 1914 records, of which, after duplicates were removed, 1248 titles and abstracts were screened. After excluding irrelevant records, 116 full texts were screened, with another 36 being screened following citation searching. Finally, 81 of the records were included in the map [5,9-18,23-27,29,39-101]).

Of the 100 Instagram posts captured, 76 (76%) contained relevant information about the experience of TSW. Of the 100 Reddit posts captured, 77 (77%) contained relevant information about the experience of TSW. Of the 77 blogs identified through blog searching and internet searches, 71 (92%) met the inclusion criteria. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram shows the flow of records throughout the screening process (Figure 1) [102].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram.

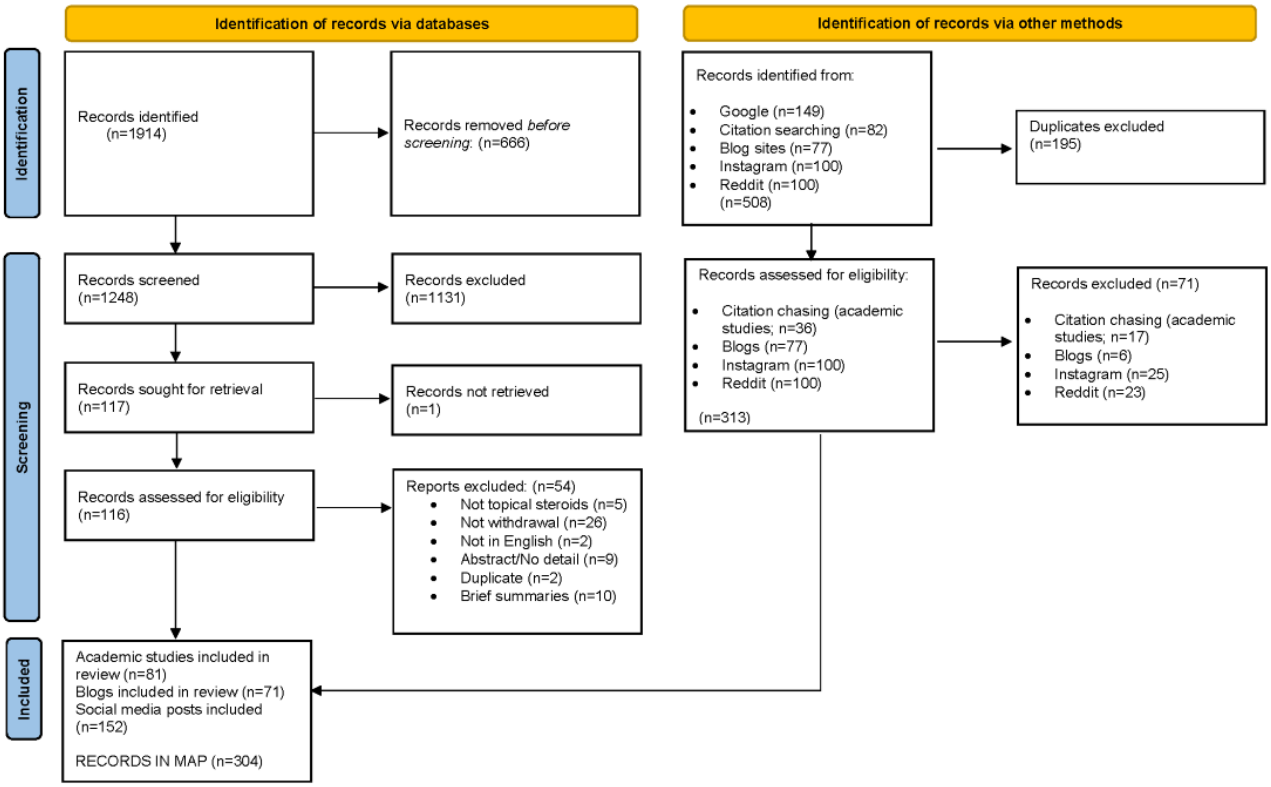

Results of the Map

The interactive EGM [103] comprises two distinct types of evidence: (1) research evidence published in academic journals and (2) evidence published by people (or their carers and family members) living with TSWS on social media. There were 81 publications by academics and clinicians and 223 social media posts (including Instagram posts, Reddit posts, and blogs) by people living with TSWS or by carers and family members.

The 81 publications by academics and clinicians were published across a period of 55 years from 1968 to 2023 (Figure 2), with 1 (1%) study published in 1969 [14] and increasing to 20 (24%) in 2022 to 2023. The research evidence published by academics and clinicians included reviews (16/81, 20%); quantitative studies (54/81, 66%); qualitative studies (2/81, 2%); and guidelines, reports, discussions, or editorials (10/81, 12%). Of the 16 reviews, only 2 (12%) were systematic reviews and 14 (88%) were nonsystematic reviews. The quantitative studies comprised mostly case reports or histories (34/54, 63%) and no randomized controlled trials. The research evidence was located in 14 countries: the United States (37/81, 46%), the United Kingdom (14/81, 17%), Australia (7/81, 9%), Japan (6/81, 7%), India (4/81, 5%), Taiwan (3/81, 4%), Singapore (2/81, 2%), Ireland (2/81, 2%), China (2/81, 2%), France (2/81, 2%), Canada (2/81, 2%), South Africa (1/81, 1%), Tunisia (1/81, 1%), and Germany (1/81, 1%).

Of 223 social media posts, there were 71 (31.8%) blogs and 152 (68.2%) Instagram and Reddit posts. The blogs were published between 2011 and 2022, and the Instagram and Reddit posts were published between 2022 and 2023, which reflects the decision to sample Instagram and Reddit posts on specific dates. The social media evidence was located in 11 countries: the United States (21/223, 9.4%), the United Kingdom (19/223, 8.5%), Australia (7/223, 3.1%), Canada (5/223, 2.2%), New Zealand (2/223, 0.9%), Singapore (2/223, 0.9%), Ireland (1/223, 0.4%), Taiwan (1/223, 0.4%), South Africa (1/223, 0.4%), Portugal (1/223, 0.4%), and Belgium (1/223, 0.4%). However, it was not always possible to know in which country the social media posts originated.

Most research evidence focused on adults (51/81, 63%), followed by children aged 4 to 12 years (18/81, 22%), infants aged 0 to 3 years (13/81, 16%), and young people aged 13 to 18 years (9/81, 11%). Likewise, the adult experience of TSWS was more highly represented in the social media evidence (125/223, 56.1%), with only 5.8% (13/223) of posts focusing on young people, children, and infants. Both the research evidence and the social media evidence did not always report the TCs used (40/81, 49% and 118/223, 52.9%, respectively), and Table 1 shows the TCs that were most used when specified.

It was not always possible to determine how long people had been using TCs, but 33 (41%) out of 81 publications reported people using TCs for up to 10 years and 13 (16%) reported people using TCs for >10 years. In the social media evidence, there were 25 (11.2%) out of 223 posts that indicated up to 10 years of use and 40 (17.9%) posts that indicated>10 years of use. The main areas of the body affected, as reported in the research evidence, were the face (40/81, 49%), legs (16/81, 20%), and arms (15/81, 19%). In the social media evidence, the main areas of the body affected were the face (50/223, 22.4%) and legs (25/223, 11.2%), but 22.4% (50/223) of the posts reported that the entire body was affected. The main reason for stopping TCs, according to the research evidence, was the physician's advice to stop (10/81, 12%), followed by lack of durable benefit (6/81, 7%), lack of effectiveness (6/81, 7%),

XSL•FO
RenderX

and the requirement to participate in a study (4/81, 5%). The main reason for stopping TCs, according to the social media evidence, was the awareness of TSWS as a condition (33/223, 14.8%); other reasons were lack of effectiveness (16/223, 7.2%), awareness of side effects (eg, skin thinning; 9/223, 4%), and lack of durable benefit (4/223, 1,8%). Diagnosis by a clinician or dermatologist was recorded mostly in the research evidence (26/81, 32%), while it was less common in the social media evidence (9/223, 4%). In contrast, self-diagnosis was most prevalent in social media (38/223, 17%), followed by diagnosis supported by social media (25/223, 11.2%). Patch testing was reported in both the research evidence (11/81, 14%) and social media evidence (2/223, 0.9%). There were also cases where a diagnosis of TSWS had not been confirmed in the social media evidence (9/223, 4%), where individuals reported that they were unsure whether they were experiencing TSWS or something else.

Figure 2. Number of academic publications and publication years.

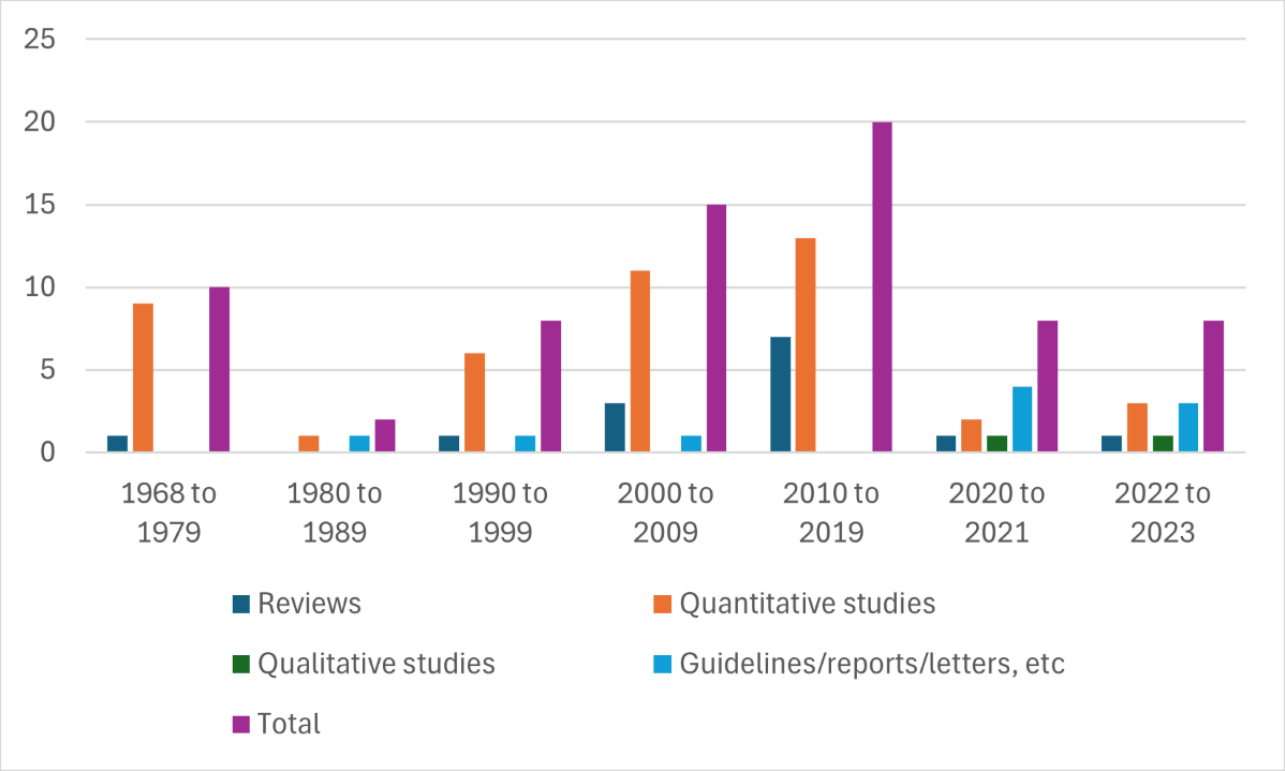

Table 1. Most used topical corticosteroids reported in research evidence and social media evidence (N=304).

Steroids used	Research evidence (n=81), n (%)	Social media evidence (n=223), n (%)
Betamethasone	28 (34.6)	22 (9.9)
Clobetasol	19 (23.5)	20 (9)
Hydrocortisone	16 (19.8)	36 (16.1)
Triamcinolone	15 (18.5)	13 (5.8)
Fluocinolone	13 (16)	0 (0)
Mometasone	13 (16)	18 (8.1)
Systemic steroids	13 (16)	16 (7.2)

Overview

Risk Factors

The risk factors for TSWS featured in the research evidence and highlighted “prolonged use of moderate to high potency TCs” (34/81, 42%) and “prolonged and frequent use of TCs” (32/81, 40%) as the most recognized risk factors. History of atopy (15/81, 19%) and inflammatory skin conditions (12/81, 15%) were also reported. In contrast, risk factors did not feature strongly in the social media evidence, with only 4.9% (11/223) of posts referring to prolonged and frequent use of TCs.

Physical Symptoms (Skin and Nonskin)

A variety of skin symptoms of TSWS were reported in the academic publications: rash (including redness, papules, and erythema; 61/81, 75%), itch (39/81, 48%), burning or stinging (36/81, 44%), scaling (34/81, 42%), pustules (33/821 41%), telangiectasias (30/81, 37%), edema (30/81, 37%), and dryness and flaking (24/81, 30%). The social media evidence reported

XSL•FO
RenderX

similar symptoms but with some variations: rash (9/223, 4%), itch (83/223, 37.2%), dryness or flaking (80/223, 35.9%), oozing or crustiness (69/223, 30.9%), burning or stinging (42/223, 18.8%), cuts (40/223, 17.9%), edema (38/223, 17%), skin pain (other than burning pain; 38/223, 17%), thickening of outer layer of skin (27/223, 12.1%), and smell (25/223, 11.2%).

The most common nonskin symptoms of TSWS reported in the research evidence were increased hair growth (11/81, 14%), sleep loss (9/81, 11%), temperature regulation (9/81, 11%), and lymphadenopathy (9/81, 11%). A wider range of symptoms were reported in the social media evidence with sleep loss (54/223, 24.2%), temperature regulation (38/223, 17%), fatigue and muscle weakness 29/223, 13%), pain (including nerve pain; 20/223, 9%), mobility issues (16/223, 7.2%), and lymphadenopathy (9/223, 4%).

Mental Health Symptoms

In the research evidence, 12% (10/82) of the academic publications reported depression, anxiety, and stress; 7% (6/81) reported suicidal thoughts; and 6% (5/81) reported emotional fluctuations. Interestingly, the social media evidence captured 2 additional aspects of mental health, including self-image (both negative and positive; 45/223, 20.2%) and resilience (45/223, 20.2%). These were followed by depression, anxiety, and stress (42/223, 18.8%); emotional fluctuations (25/223, 11.2%); and suicidal thoughts 4/223, 1.8%).

Relationships

How living with TSWS impacted adults' and children's relationships with others was not addressed in the research evidence, apart from 1 qualitative study that acknowledged its impact on relationships within the home and with physicians and dermatologists. In contrast, this featured strongly within the social media evidence, including relationships at home (45/223, 20.2%), at work, and in education (7/223, 3.1%), as well as relationships with physicians and dermatologists (45/223, 20.2%), with other therapists (9/223, 4%), and with other people with TSWS often via social media (31/223, 13.9%).

Beliefs and Attitudes

The beliefs and attitudes toward TCs, TSWS, and information on TCs and TSWS featured most strongly in the social media evidence. The belief in the importance of perseverance for coping with TSWS was emphasized (49/223, 22%), followed by the need for "alternative" sources of information on TSWS (31/223, 13.9%) and the belief that the information on TCs from physicians was inadequate (20/223, 9%). Mistrust of medical professionals featured (20/223, 9%), along with the perception that physicians do not believe that TSWS exists (16/223, 7.2%) and fears of steroids and medications (11/223, 4.9%). The research evidence was sparse but fear of steroids and medications (6/81, 7%), alternative sources of information on TSWS (5/81, 6%), and belief that TSWS does not exist (5/81, 6%) were reported.

Activities of Everyday Living

Living with TSWS affected people's everyday lives, and not surprisingly, this was captured by the social media evidence across all the domains: social life (31/223, 13.9%), work (31/223, 13.9%), self-care (25/223, 11.2%), holidays and leisure (13/223, 5.8%), and college or school (11/223, 4.9%). In contrast, the research evidence gave little attention to how TSWS impacted work (3/81, 4%), college or school (3/81, 4%), and social life (2/81, 2%).

Treatments

Complete cessation of TCs was the favored strategy in both the research evidence (48/81, 59%) and the social media evidence (51/223, 22.9%). Tapering of TCs featured more in the research evidence (22/81, 27%) than in the social media evidence (11/223, 4.9%). Using systemic steroids as a treatment was reported in the research evidence (11/81, 14%) but less in the social media evidence (4/223, 1.8%), and restarting TCs was reported in the social media (8/223, 3.6%) but to a lesser extent in the research evidence (5/81, 6%). Some of the pharmacological treatments mentioned in the research evidence were niche (mentioned only once), such as Montelukast, adrenocorticotropic hormone, platelet-rich plasma, intravenous immunoglobulin, and Nicardipine, and were not mentioned in the social media evidence at all. Not surprisingly, the research evidence tended to report greater use of pharmacological treatments, while the social media evidence highlighted the treatments most used by people living with TSWS, as presented in Table 2.

XSL•FO
RenderX

Table 2. Treatments for topical steroid withdrawal syndrome in research evidence and social media evidence (N=304).

Treatments used	Research evidence (n=81), n (%)	Social media evidence (n=223), n (%)
Antihistamines	23 (28)	23 (10.3)
Antibiotics	33 (40.7)	20 (9)
Cyclosporine	8 (9.9)	9 (4)
Dupilumab	8 (9.9)	11 (4.9)
Methotrexate	4 (4.9)	4 (1.8)
Emollients and/or moisturizers	25 (30.9)	58 (26)
Alternative remedies	8 (9.9)	51 (22.9)
Bathing interventions	5 (6.2)	47 (21.1)
Diet changes	10 (12.3)	29 (13)
Clothing interventions	1 (1.2)	20 (9)
Hot and/or cold treatments	15 (18.5)	20 (9)
Psychological therapy and/or support	18 (22)	1 (0.5)
Online support	4 (4.9)	36 (16.1)

Treatment Costs and Outcome Measures

Treatment costs were only mentioned in a small proportion of the social media evidence (16/223, 7.2%) and research evidence (2/81, 2%). There was little on outcome measures, with 5% using the Dermatology Life Quality Index and 1% using self-assessed questionnaires. The social media evidence showed that people were recording their feelings of progress on the "TSW journey" to recovery (22/223, 9.9%), and a smaller proportion (9/223, 4%) referred to "being healed" as reaching the end of the TSW journey.

Discussion

Principal Findings

This EGM presented the developing body of research evidence on TSWS, enabling a timely insight into the topics of interest to the medical and dermatological community. The research evidence was distributed across a number of topics, but there were many where evidence was clearly lacking, such as diagnosis, prevention, epidemiology, relationships, activities of everyday living, beliefs and attitudes, treatment costs, and outcome measures. The EGM also presented a sample of the social media evidence on TSWS, which concentrated on the physical symptoms (skin and nonskin), mental health symptoms, relationships, activities of everyday living, beliefs and attitudes, and treatments. This, by implication, highlights additional topics that need greater attention from academics and clinicians and confirms the need for research on the patient's lived experience of TSWS.

The lack of research evidence published across the topics of diagnosis, prevention, and epidemiology is not surprising, given that the legitimacy of the condition is still questioned by many in the medical and dermatological communities [15,29]. The need for consensus on diagnostic criteria has been recognized as a "priority" by a few authors [27,39], and arguably, without these, it is not possible to understand the incidence, prevalence, and distribution of TSWS. In 2014, Fukaya et al [13] stated that there were "no statistics regarding the prevalence of TSA," and this remains the case. The challenge of diagnosing TSWS was reiterated [12,16,17,20,40,41], and the main differential diagnoses were atopic dermatitis itself, allergic contact dermatitis, and infection. Sheary [12] observed that many of the symptoms of TSWS may also be seen in severe atopic dermatitis but argued that the process of differential diagnosis relies on a thorough understanding of patient history and physical examination. Unsurprisingly, most cases of TSWS are self-diagnosed rather than by a clinician or dermatologist [17], which was confirmed by the social media evidence. Sheary [12] proposed a set of diagnostic criteria as a starting point for discussion and future research; it comprised (1) essential criteria, (2) key diagnostic criteria, and (3) additional supporting features that may be present. The essential criteria are history of long-term regular use of TCs, itch, and erythema. Lio and Chandan's [17] key diagnostic criteria (equivalent to the "essential" criteria proposed by Sheary [12]) were burning, confluent erythema, and a history of frequent and prolonged TC use. These proposals, along with the work on diagnostic criteria proposed by Guo et al [104] using a modified Delphi approach, could provide a way forward for the medical and dermatological communities to engage with the issue of agreeing on recognized diagnostic criteria.

Given that addiction precedes withdrawal, prevention of addiction to TCs is of paramount importance. According to Sheary [18], it is "not widely accepted" within the medical community that there may be withdrawal symptoms associated with TC cessation, and she argues that this may be because TCs have been traditionally recommended for "intermittent flares of symptoms" rather than long-term use. However, the EGM indicates that patients can use TCs for years. For example, in a recent survey, people with eczema reported using TCs for >15 years [42], and in another study, the length of use of TCs for some people was reported to be ≥40 years [12]. Sheary [18] observed that TC "overuse" has never been defined, and arguably, neither has "prolonged use." That research has not investigated prolonged use of TCs has been shown by a recent

XSL•FO
RenderX

"umbrella review" on the safety of TCs in adults and children with eczema [106], which found limited high-quality evidence for long-term use of TCs. The data on prolonged use were limited as many randomized controlled trials were of short duration and did not include follow-up beyond 2 to 4 weeks of treatment. This echoes Eichenfield et al's [2] observations in 2014 that "...most studies fail to follow up patients for potential complications," and despite their recommendation that treatment sites should be monitored regularly for adverse effects, just over half of the international guidelines in the review by LePoidevin et al [1] recommended screening for cutaneous side effects. Axon et al [105] recommended that "...longer-term prospect observational studies are better placed to exploit longer-term safety of TCs and should be designed with years rather than months of follow-up to add useful information to the field." Understanding the impact and safety of prolonged use of TCs has not been prioritized, as argued by Peacock [106], who contends that large-scale research funding has, for decades, primarily focused on product development for pharmaceutical companies.

From the existing research evidence in this EGM, prolonged and frequent use of TCs and prolonged use of moderate or high potency TCs were identified as risk factors for TSWS. Yet, this has to be juxtaposed with the dominant focus of the published medical literature on the safety of TCs, reporting on patient "misuse/abuse," "inappropriate use," or "underuse" [107]; and how to encourage "correct use" [108]. The "reluctance to use topical steroids as recommended" [109] has been dismissed by many dermatologists as "steroid phobia," as has TSWS [15]. Arguably, this narrative of blame filters through to general practitioners (GPs) who can "struggle" [106] to support people experiencing TSWS and need guidance from dermatologists on the safe use of TCs. As already noted, there is considerable diversity across the international management guidelines for atopic dermatitis on the use of TCs. For example, the guidance on optimal dosage is variable [1], not least because the evidence base lacks "...studies that examine a range of TCS doses in large numbers of patients" [2]. In the UK context, the MHRA provided advice for health care professionals in 2021 [20] on prescribing TCs in terms of potency, quantity to be applied, and area, frequency, and duration of application. The advice included reporting suspected adverse drug reactions to the Yellow Card Scheme (a UK scheme for monitoring suspected side effects to medicines and medical devices) after cessation of TCs. A recent update to the advice [110] details that TCs will be labeled with information on their potency to help with counseling patients. Given the key role GPs have in prescribing TCs, there is a case to be made for research on how to implement training and guidance for GPs and other health care professionals as an important part of building the evidence base on the safe use of TCs.

The research evidence devoted considerable attention to the physical "skin" symptoms of TSWS, which concurred, in the main, with the social media evidence. However, the social media evidence expanded on "other" physical symptoms, mental health symptoms of TSWS, treatments used, and how it impacted activities of everyday living and relationships. This gap in the research evidence can be explained by the lack of qualitative research with only 2 qualitative studies on TSWS published in the last 50 years. The research evidence also addressed treatments to manage TSWS, which tended to be pharmacological, with some of these treatments also mentioned in social media. The social media evidence revealed that people with TSWS were using a range of nonpharmacological treatments, which included emollients and moisturizers and bathing interventions, and then "alternative remedies," which could encompass acupuncture, hypnotherapy, and homeopathy, for example. Interestingly, the research evidence on pharmacological treatments such as Dupilumab reported treatment improvements after relatively short periods, such as 8, 13, and 31 weeks [43]. "Successful" treatments were less "linear" in the social media evidence, particularly in the blogs, where people with TSWS often reported a "roller coaster" experience of recovery for many months and even years, which could be punctuated with debilitating flares of symptoms such as itch, flaking, and oozing. With only 1 prospective cohort study in the EGM [23], there is clearly a need for more longitudinal research on the patient's "TSWS journey" to healing.

Psychological therapy and support was one of the treatments that clinicians recommended for people with TSWS, but this hardly featured in the social media evidence. Online support was important for people with TSWS, which is possibly an artifact of social media being a form of online support and a reflection of the quality of the physician-patient relationship. Those who felt that their experiences and concerns were disregarded by their physicians and dermatologists often turned to social media for support and information [26].

Arguably, regardless of whether dermatologists and clinicians accept TSWS as a distinct clinical entity, there are important reasons why they need to engage with patients' stories of living with TSWS from social media. To dismiss people living with TSWS as "misusers" and "misinformed" means that they may be left alone to deal with the symptoms of TSWS and consequently may "seek inappropriate alternative therapies" [39] that could be detrimental to their care and impact their recovery from TSWS [9]. Furthermore, dismissal by medical professionals has been found to introduce another element of distress to the lives of people with TSWS: in short, people living with TSWS "need to be heard and acknowledged by the medical community" [44].

Strengths and Limitations

This study represents the first EGM in TSWS and responds to calls for a better understanding of TSWS. It aims to be an important resource to guide both researchers and clinicians in the prioritization of research topics for further research. Support from our public collaborator meant that this study was grounded in "real" lived experience, which is of great importance for ensuring that research is "meaningful to patients" [27]. Additional collaborative input from a researcher or clinician with direct dermatological expertise would also have been beneficial. Mapping both research evidence and social media evidence is a methodological innovation in the production of EGMs. The inclusion of social media evidence was in recognition of the increased presence of

XSL•FO
RenderX

#topicalsteroidwithdrawal on social media [24] and how social media content could contribute to understanding the patient perspective of TSWS. Social media is increasingly used by qualitative health researchers, and there has been some interest in social media within the dermatological community [111]. In the EGM, there is only 1 example of researchers reviewing social media blogs to understand children's experiences of TSWS [45]. Due to the amount of social media content available on TSWS, we decided that a sample of social media evidence was the only feasible approach for this EGM, and therefore, it provides a single snapshot of the social media evidence. Furthermore, bloggers and people using Instagram and Reddit are a self-selecting sample who are clearly proficient with technology, so it is difficult to know how reflective they are of those living with TSWS. This does not devalue their accounts but means that there are likely to be many people's experiences of TSWS that are not shared on social media. There were also some cases, bloggers in particular, who presented with a self-diagnosis of TSWS but were unsure whether their symptoms were indicative of TSWS, and it was challenging to judge the veracity of their accounts. The ephemeral nature of social media data should also be noted, and some of our selected Instagram and Reddit posts were removed after inclusion in the EGM. As an EGM dealing with a range of academic publications and a sample of social media evidence, we did not conduct a critical appraisal [31]. The low quality of the existing research evidence has been noted [9,16,25], and there is undoubtedly an urgent need for high-quality qualitative and quantitative research. However, this EGM offers an overview of the TSWS research landscape and an insight into what people living with TSWS are discussing on social media.

Conclusions

TSA, as an adverse effect of TCs, was first identified in 1969; and yet, >50 years later, TSWS remains controversial and contested in the dermatological and medical communities. This EGM shows that TSWS has attracted increased research attention over the years, but high-quality research is lacking. The evidence gaps highlight priorities for future primary research, which are longer-term prospective observational studies to assess the safety of prolonged use of TCs and help prevent addiction, qualitative research to understand the lived experience of TSWS, and longitudinal research on the patient's "TSWS journey" to healing. However, it is crucial that future research in TSWS is underpinned by work that determines agreed diagnostic criteria for TSWS and by collaboration between researchers and clinicians with the TSWS patient community.

Acknowledgments

This evidence gap map was funded by KS's National Institute for Health Research (NIHR) Senior Investigator Award (NIHR201377). This work was supported by the University of Exeter Medical School, University of Exeter, and NIHR Applied Research Collaboration South West Peninsula. Views expressed are those of the authors and not necessarily those of the University of Exeter or the NIHR. The authors would like to thank Dr Rebecca Abbott and Ms Rebecca Whear for their constructive comments on the draft and the reviewers for their helpful feedback.

Authors' Contributions

KS, JTC, NO, MR, and AS conceived the project and co-developed the framework for the map; MR carried out the systematic search of the literature and social media; and MR and NO selected the academic publications and social media evidence. MR and NO extracted and checked the data. NO and MR drafted the manuscript, and all authors revised and commented on the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 18 KB - jmir_v26i1e57687_app1.docx]

References

1. LePoidevin LM, Lee DE, Shi VY. A comparison of international management guidelines for atopic dermatitis. Pediatr Dermatol 2019 Jan;36(1):36-65. [doi: 10.1111/pde.13678] [Medline: 30303557]
2. Eichenfield LF, Tom WL, Berger TG, Krol A, Paller AS, Schwarzenberger K, et al. Guidelines of care for the management of atopic dermatitis: section 2. Management and treatment of atopic dermatitis with topical therapies. J Am Acad Dermatol 2014 Jul;71(1):116-132 [FREE Full text] [doi: 10.1016/j.jaad.2014.03.023] [Medline: 24813302]
3. Kulthanan K, Tuchinda P, Nitiyarom R, Chunharas A, Chantaphakul H, Aunhachoke K, et al. Clinical practice guidelines for the diagnosis and management of atopic dermatitis. Asian Pac J Allergy Immunol 2021 Sep;39(3):145-155 [FREE Full text] [doi: 10.12932/AP-010221-1050] [Medline: 34246205]

XSL•FO
RenderX

4. Sidbury R, Alikhan A, Bercovitch L, Cohen DE, Darr JM, Drucker AM, et al. Executive summary: American Academy of Dermatology guidelines of care for the management of atopic dermatitis in adults with topical therapies. J Am Acad Dermatol 2023 Jul;89(1):128-129. [doi: 10.1016/j.jaad.2022.08.068] [Medline: 36623556]
5. Hengge UR, Ruzicka T, Schwartz RA, Cork MJ. Adverse effects of topical glucocorticosteroids. J Am Acad Dermatol 2006 Jan;54(1):1-15; quiz 16. [doi: 10.1016/j.jaad.2005.01.010] [Medline: 16384751]
6. Saeki H, Ohya Y, Furuta J, Arakawa H, Ichiyama S, Katsunuma T, et al. Executive summary: Japanese guidelines for atopic dermatitis (ADGL) 2021. Allergol Int 2022 Oct;71(4):448-458 [FREE Full text] [doi: 10.1016/j.alit.2022.06.009] [Medline: 36064654]
7. Wollenberg A, Kinberger M, Arents B, Aszodi N, Avila Valle G, Barbarot S, et al. European guideline (EuroGuiDerm) on atopic eczema - part II: non-systemic treatments and treatment recommendations for special AE patient populations. J Eur Acad Dermatol Venereol 2022 Nov;36(11):1904-1926 [FREE Full text] [doi: 10.1111/jdv.18429] [Medline: 36056736]
8. What is TSW syndrome? ITSAN. URL: https://www.itsan.org/what-is-tsw-syndrome/#tab-id-2 [accessed 2024-10-26]
9. Tan SY, Chandran NS, Choi EC. Steroid phobia: is there a basis? A review of topical steroid safety, addiction and withdrawal. Clin Drug Investig 2021 Oct;41(10):835-842 [FREE Full text] [doi: 10.1007/s40261-021-01072-z] [Medline: 34409577]
10. Kligman AM, Frosch PJ. Steroid addiction. Int J Dermatol 1979;18(1):23-31. [doi: 10.1111/j.1365-4362.1979.tb01905.x] [Medline: 153891]
11. Rapaport MJ, Rapaport V. The red skin syndromes: corticosteroid addiction and withdrawal. Expert Rev Dermatol 2014 Jan 10;1(4):547-561. [doi: 10.1586/17469872.1.4.547]
12. Sheary B. Steroid withdrawal effects following long-term topical corticosteroid use. Dermatitis 2018;29(4):213-218. [doi: 10.1097/DER.0000000000000387] [Medline: 29923852]
13. Fukaya M, Sato K, Sato M, Kimata H, Fujisawa S, Dozono H, et al. Topical steroid addiction in atopic dermatitis. Drug Healthc Patient Saf 2014 Oct 14;6:131-138 [FREE Full text] [doi: 10.2147/dhps.s6920] [Medline: 25378953]
14. Sneddon I. Adverse effect of topical fluorinated corticosteroids in rosacea. Br Med J 1969 Mar 15;1(5645):671-673 [FREE Full text] [doi: 10.1136/bmj.1.5645.670] [Medline: 4237682]
15. Finnegan P, Murphy M, O'Connor C. #corticophobia: a review on online misinformation related to topical steroids. Clin Exp Dermatol 2023 Feb 02;48(2):112-115. [doi: 10.1093/ced/llac019] [Medline: 36730502]
16. Hwang J, Lio PA. Topical corticosteroid withdrawal ('steroid addiction'): an update of a systematic review. J Dermatolog Treat 2022 May;33(3):1293-1298. [doi: 10.1080/09546634.2021.1882659] [Medline: 33499686]
17. Lio PA, Chandan N. Topical steroid withdrawal in atopic dermatitis. Practical Dermatology. 2019 Dec 8. URL: https://practicaldermatology.com/topics/atopic-dermatitis/topical-steroid-withdrawal-in-atopic-dermatitis/23079/ [accessed 2024-10-26]
18. Sheary B. Topical steroid withdrawal: a case series of 10 children. Acta Derm Venereol 2019 May 01;99(6):551-556 [FREE Full text] [doi: 10.2340/00015555-3144] [Medline: 30734047]
19. National Eczema Society and British Association of Dermatologists joint position statement on Topical Steroid Withdrawal. British Association of Dermatologists. URL: https://www.bad.org.uk/national-eczema-society-and-british-association-of-dermatologists-joint-position-statement-on-topical-steroid-withdrawal/ [accessed 2024-10-26]
20. Medicines and Healthcare products Regulatory Agency. Topical corticosteroids: information on the risk of topical steroid withdrawal reactions. Drug Ther Bull 2021 Dec;59(12):184. [doi: 10.1136/dtb.2021.000062] [Medline: 34824131]
21. Ballard A. Topical steroid withdrawal: what the eczema community needs to know, now. National Eczema Association. 2022 Aug 24. URL: https://nationaleczema.org/blog/tsw-need-to-know/ [accessed 2024-10-26]
22. Lee HK, Reiche L. Topical corticosteroid withdrawal. DermNet. 2023. URL: https://dermnetnz.org/topics/topical-corticosteroid-withdrawal [accessed 2024-10-26]
23. Sheary B, Harris MF. Cessation of long-term topical steroids in adult atopic dermatitis: a prospective cohort study. Dermatitis 2020;31(5):316-320. [doi: 10.1097/DER.0000000000000602] [Medline: 32404621]
24. Bowe S, Masterson S, Murray G, Haugh I. Topical steroid withdrawal through the lens of social media. Clin Exp Dermatol 2022 Aug;47(8):1554-1557. [doi: 10.1111/ced.15194] [Medline: 35340034]
25. Hajar T, Leshem YA, Hanifin JM, Nedorost ST, Lio PA, Paller AS, et al. A systematic review of topical corticosteroid withdrawal ("steroid addiction") in patients with atopic dermatitis and other dermatoses. J Am Acad Dermatol 2015 Mar;72(3):541-9.e2. [doi: 10.1016/j.jaad.2014.11.024] [Medline: 25592622]
26. Tan S, Phan P, Law JY, Choi E, Chandran NS. Qualitative analysis of topical corticosteroid concerns, topical steroid addiction and withdrawal in dermatological patients. BMJ Open 2022 Mar 16;12(3):e060867 [FREE Full text] [doi: 10.1136/bmjopen-2022-060867] [Medline: 35296492]
27. Howells L, Broome H, Burleigh A, Hammond H, Ismail F, Proctor A, et al. Topical corticosteroid withdrawal syndrome: the patient community call for high-quality research, clear definitions and diagnostic criteria. Br J Dermatol 2023 Feb 10;188(2):288-289. [doi: 10.1093/bjd/ljac067] [Medline: 36763872]
28. Li AW, Yin ES, Antaya RJ. Topical corticosteroid phobia in atopic dermatitis: a systematic review. JAMA Dermatol 2017 Oct 01;153(10):1036-1042. [doi: 10.1001/jamadermatol.2017.2437] [Medline: 28724128]

XSL•FO
RenderX

29. Guckian J, Hughes O, Nikooam Y, Nair R, Brown J, Bewley A. PS03 Dermatologist perspectives on topical steroid withdrawal: distrust, disinformation and distance? Br J Dermatol 2023;188(Supplement_4):ljad113.355. [doi: 10.1093/bjd/ljad113.355]
30. White H, Albers B, Gaarder M, Kornør H, Littell J, Marshall Z, et al. Guidance for producing a Campbell evidence and gap map. Campbell Syst Rev 2020 Dec 19;16(4):e1125 [FREE Full text] [doi: 10.1002/cl2.1125] [Medline: 37016607]
31. Campbell F, Tricco AC, Munn Z, Pollock D, Saran A, Sutton A, et al. Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): the same but different- the "Big Picture" review family. Syst Rev 2023 Mar 15;12(1):45 [FREE Full text] [doi: 10.1186/s13643-023-02178-5] [Medline: 36918977]
32. Orr N, Rogers M, Stein A, Thompson Coon J, Stein K. What is the range, extent and type of evidence about topic steroid withdrawal in research literature and social media platforms? A protocol for an evidence and gap map. University of Exeter, College of Medicine and Health. 2023 Jun 8. URL: https://ore.exeter.ac.uk/repository/handle/10871/133327 [accessed 2024-10-26]
33. Briscoe S, Rogers M. An alternative screening approach for Google Search identifies an accurate and manageable number of results for a systematic review (case study). Health Info Libr J 2024 Jun;41(2):149-155. [doi: 10.1111/hir.12409] [Medline: 34734655]
34. Keim-Malpass J, Steeves RH, Kennedy C. Internet ethnography: a review of methodological considerations for studying online illness blogs. Int J Nurs Stud 2014 Dec;51(12):1686-1692. [doi: 10.1016/j.ijnurstu.2014.06.003] [Medline: 24985757]
35. Goh AQ, Lo NY, Davis C, Chew EC. #EatingDisorderRecovery: a qualitative content analysis of eating disorder recovery-related posts on Instagram. Eat Weight Disord 2022 May;27(4):1535-1545. [doi: 10.1007/s40519-021-01279-1] [Medline: 34537927]
36. Karimkhani C, Connett J, Boyers L, Quest T, Dellavalle RP. Dermatology on Instagram. Dermatol Online J 2014 Jul 15;20(7):13030/qt71g178w9 [FREE Full text] [Medline: 25046455]
37. Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M, et al. EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. London, UK: UCL Social Research Institute; 2020.
38. Digital Solution Foundry and EPPI Centre. EPPI-Mapper, Version 2.2.4. London, UK: EPPI Centre, UCL Social Research Institute, University College London; 2023.
39. Cotter C, Burton T, Proctor A, Moss C, Flohr C. Topical steroid withdrawal syndrome: time to bridge the gap. Br J Dermatol 2022 Nov;187(5):780-781. [doi: 10.1111/bjd.21770] [Medline: 35833942]
40. Chen AY, Zirwas MJ. Steroid-induced rosacealike dermatitis: case report and review of the literature. Cutis 2009 Apr;83(4):198-204. [Medline: 19445310]
41. Sheary B. Topical corticosteroid addiction and withdrawal - an overview for GPs. Aust Fam Physician 2016 Jun;45(6):386-388 [FREE Full text] [Medline: 27622228]
42. Barta K, Fonacier LS, Hart M, Lio P, Tullos K, Sheary B, et al. Corticosteroid exposure and cumulative effects in patients with eczema: results from a patient survey. Ann Allergy Asthma Immunol 2023 Jan;130(1):93-9.e10 [FREE Full text] [doi: 10.1016/j.anai.2022.09.031] [Medline: 36191848]
43. Arnold KA, Treister AD, Lio PA. Dupilumab in the management of topical corticosteroid withdrawal in atopic dermatitis: a retrospective case series. JAAD Case Rep 2018 Oct;4(9):860-862 [FREE Full text] [doi: 10.1016/j.jdcr.2018.06.012] [Medline: 30306110]
44. Brookes TS, Barlow R, Mohandas P, Bewley A. Topical steroid withdrawal: an emerging clinical problem. Clin Exp Dermatol 2023 Aug 25;48(9):1007-1011. [doi: 10.1093/ced/llad161] [Medline: 37119282]
45. Juhász ML, Curley RA, Rasmussen A, Malakouti M, Silverberg N, Jacob SE. Systematic review of the topical steroid addiction and topical steroid withdrawal phenomenon in children diagnosed with atopic dermatitis and treated with topical corticosteroids. J Dermatol Nurses Assoc 2017;9(5):233-240. [doi: 10.1097/JDN.0000000000000331]
46. No authors listed. Difficulties encountered on discontinuing topical corticosteroids: Patients' experiences. Prescrire International 2021;30(230):509-511.
47. Topical corticosteroids and the risk of topical withdrawal reactions. Health Product InfoWatch, Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/health-product-infowatch/july-2022.html#a6.1.1 [accessed 2024-10-26]
48. No authors listed. Topical corticosteroid withdrawal reactions. Drug Ther Bull 2021 Dec;59(12):184. [doi: 10.1136/dtb.2021.000062] [Medline: 34824131]
49. Aouinti I, Daly W, Lakhoua G, Kaabi W, Kastalli S, Zaiem A, et al. Topical steroid-induced Cushing syndrome and adrenal insufficiency. In: Proceedings of the Annual Meeting of French Society of Pharmacology and Therapeutics. 2023 Presented at: SFPT 2023; June 12-14, 2023; Limoges, France. [doi: 10.1111/fcp.12908]
50. Bartunek C, Brodell RT, Brodell LP. Self-assessment quiz: dermatomyositis. J Ophthalmic Nurs Technol 1997;16(4):199-200. [Medline: 9335734]
51. Béné J, Durieux F, Bergeron S, Potey C, Lebrun-Vignes B, Staumont-Sallé D, et al. About two cases of possible topical steroid withdrawal reaction. Therapie 2023;78(4):448-451. [doi: 10.1016/j.therap.2022.08.002] [Medline: 36096918]

XSL•FO
RenderX

52. Bhatti RS, Flynn MD. Adrenal insufficiency secondary to inappropriate oral administration of topical exogenous steroids presenting with hypercalcaemia. BMJ Case Rep 2012 Jun 21;2012:bcr0320125983 [FREE Full text] [doi: 10.1136/bcr.03.2012.5983] [Medline: 22729340]
53. Boxley JD, Dawber RP, Summerly R. Generalized pustular psoriasis on withdrawal of clobetasol propionate ointment. Br Med J 1975 May 03;2(5965):255-256 [FREE Full text] [doi: 10.1136/bmj.2.5965.255-a] [Medline: 1131576]
54. Brodell RT, O'Brien MJJ. Topical corticosteroid-induced acne. Three treatment strategies to break the 'addiction' cycle. Postgrad Med 1999 Nov;106(6):225-6, 229. [doi: 10.3810/pgm.1999.11.796] [Medline: 10576013]
55. Burry JN. Topical drug addiction: adverse effects of fluorinated corticosteroid creams and ointments. Med J Aust 1973 Feb 24;1(8):393-396. [doi: 10.5694/j.1326-5377.1973.tb118065.x] [Medline: 4266903]
56. Calnan CD. Use and abuse of topical steroids. Dermatologica 1976;152 Suppl 1:247-251. [doi: 10.1159/000257896] [Medline: 986326]
57. Chu CY. The use of 1% pimecrolimus cream for the treatment of steroid-induced rosacea. Br J Dermatol 2005 Feb;152(2):396-399. [doi: 10.1111/j.1365-2133.2005.06404.x] [Medline: 15727676]
58. Chu CY. An open-label pilot study to evaluate the safety and efficacy of topically applied pimecrolimus cream for the treatment of steroid-induced rosacea-like eruption. J Eur Acad Dermatol Venereol 2007 Apr;21(4):484-490. [doi: 10.1111/j.1468-3083.2006.01991.x] [Medline: 17373975]
59. Dhossche J, Simpson E, Hajar T. Topical corticosteroid withdrawal in a pediatric patient. JAAD Case Rep 2017 Sep;3(5):420-421 [FREE Full text] [doi: 10.1016/j.jdcr.2017.06.006] [Medline: 28932784]
60. Feschuk AM, Pratt ME. Topical steroid withdrawal syndrome in a mother and son: a case report. SAGE Open Med Case Rep 2023;11:2050313X231164268 [FREE Full text] [doi: 10.1177/2050313X231164268] [Medline: 37051261]
61. Fourie DT. Benign intracranial hypertension after topical steroid withdrawal. S Afr Med J 1977 Aug 13;52(8):301. [Medline: 897931]
62. Franco HL, Weston WL. Steroid rosacea in children. Pediatrics 1979 Jul;64(1):36-38. [Medline: 156343]
63. Fukaya M. Improvement of atopic dermatitis after discontinuation of topical corticosteroid treatment. Arch Dermatol 2000 May;136(5):679-680. [doi: 10.1001/archderm.136.5.679] [Medline: 10815874]
64. Fukaya M. Histological and immunohistological findings using anti-cortisol antibody in atopic dermatitis with topical steroid addiction. Dermatol Ther (Heidelb) 2016 Mar;6(1):39-46 [FREE Full text] [doi: 10.1007/s13555-016-0096-7] [Medline: 26838582]
65. Gardner DD, Barta K, Tullos KZ, Hart MK, Winders T. Survey of caregivers of children with eczema regarding corticosteroid exposure and cumulative effects. Pediatr Dermatol 2023;40(Supplement 1):20. [doi: 10.1111/pde.15272]
66. Ghosh A, Sengupta S, Coondoo A, Jana AK. Topical corticosteroid addiction and phobia. Indian J Dermatol 2014 Sep;59(5):465-468 [FREE Full text] [doi: 10.4103/0019-5154.139876] [Medline: 25284851]
67. Gust P, Jacob SE. The role of delayed-delayed corticosteroid contact dermatitis in topical steroid withdrawal. J Am Acad Dermatol 2016 Oct;75(4):e167. [doi: 10.1016/j.jaad.2016.05.048] [Medline: 27646768]
68. Hodge BD, Huynh TN, Brodell RT. "Folly" à deux: topical corticosteroid addiction in mother and son. JAAD Case Rep 2019 Jan;5(1):82-85 [FREE Full text] [doi: 10.1016/j.jdcr.2018.09.014] [Medline: 30581940]
69. Horiuchi Y. Difficulties in treating steroid withdrawal: intermittent ACTH and low dose systemic cyclosporin used to treat a senile erythroderma patient. J Dermatol 2000 Jan;27(1):44-48. [doi: 10.1111/j.1346-8138.2000.tb02117.x] [Medline: 10692825]
70. Hosking GP, Elliston H. Benign intracranial hypertension in a child with eczema treated with topical steroids. Br Med J 1978 Mar 04;1(6112):550-551 [FREE Full text] [doi: 10.1136/bmj.1.6112.550-a] [Medline: 630215]
71. Kligman AM. Topical steroid addicts. JAMA 1976 Apr 12;235(15):1550. [doi: 10.1001/jama.1976.03260410014003]
72. Lio PA. Addressing the challenge of topical steroid withdrawal. Practical Dermatology. 2015 Sep. URL: https://assets.bmctoday.net/practicaldermatology/pdfs/PD0915_clinfocus.pdf [accessed 2024-10-26]
73. Litt JZ. Steroid-induced rosacea. Am Fam Physician 1993 Jul;48(1):67-71. [Medline: 8322645]
74. Liu ZH, Du XH. Quality of life in patients with facial steroid dermatitis before and after treatment. J Eur Acad Dermatol Venereol 2008 Jun;22(6):663-669. [doi: 10.1111/j.1468-3083.2008.02639.x] [Medline: 18331297]
75. Majid I, Timungpi R. Platelet rich plasma in treatment of topical steroid damaged face: a retrospective analytical study. Dermatol Ther 2022 Apr;35(4):e15356. [doi: 10.1111/dth.15356] [Medline: 35118757]
76. Martin DL, Turner ML, Williams CM. Recent onset of smooth, shiny, erythematous papules on the face. Steroid rosacea secondary to topical fluorinated steroid therapy. Arch Dermatol 1989 Jun;125(6):828, 831. [doi: 10.1001/archderm.125.6.828b] [Medline: 2525007]
77. Monroe JR. What caused this woman's addiction to steroids? JAAPA 2011 May;24(5):20. [doi: 10.1097/01720610-201105000-00002] [Medline: 21574526]
78. Moss C, Ross K, Proctor A. Topical steroid withdrawal is not a myth. Comment on '#corticophobia: a review on online misinformation related to topical steroids'. Clin Exp Dermatol 2023 Jun 05;48(6):697. [doi: 10.1093/ced/llad062] [Medline: 36787284]

XSL•FO
RenderX

79. Nguyen AP, Kong JS, Teuber SS. Severe topical corticosteroid withdrawal syndrome from over-the-counter steroids. J Allergy Clin Immunol Pract 2021 Nov;9(11):4147-4148 [FREE Full text] [doi: 10.1016/j.jaip.2021.07.026] [Medline: 34489186]
80. Nickles MA, Coale AT, Henderson WJ, Brown KE, Morrell DS, Nieman EL. Steroid phobia on social media platforms. Pediatr Dermatol 2023 Feb 23;40(3):479-482. [doi: 10.1111/pde.15269] [Medline: 36815409]
81. Pabby A, An KP, Laws RA. Combination therapy of tetracycline and tacrolimus resulting in rapid resolution of steroid-induced periocular rosacea. Cutis 2003 Aug;72(2):141-142. [Medline: 12953939]
82. Panda M, Debata I, Behera D, Raj C, Jena A. Topical steroid addiction among paramedical staff: a cross sectional study in a tertiary health care centre. Int J Res Dermatol 2019 Oct 21;5(4). [doi: 10.18203/issn.2455-4529.intjresdermatol20194679]
83. Raimer S, Subrt P. Topical steroids: their use -- and abuse. Consultant (00107069) 1983;23(4):243-253.
84. Rapaport MJ. Topical steroid addiction in dermatology today. Practical Dermatology. 2019 Jun 11. URL: https://practicaldermatology.com/topics/atopic-dermatitis/topical-steroid-addiction-in-dermatology-today/23140/ [accessed 2024-10-26]
85. Rapaport MJ, Lebwohl M. Corticosteroid addiction and withdrawal in the atopic: the red burning skin syndrome. Clin Dermatol 2003 May;21(3):201-214. [doi: 10.1016/S0738-081X(02)00365-6]
86. Rapaport MJ, Rapaport V. Prolonged erythema after facial laser resurfacing or phenol peel secondary to corticosteroid addiction. Dermatol Surg 1999 Oct;25(10):781-4; discussion 785. [doi: 10.1046/j.1524-4725.1999.99097.x] [Medline: 10594579]
87. Rapaport MJ, Rapaport V. Eyelid dermatitis to red face syndrome to cure: clinical experience in 100 cases. J Am Acad Dermatol 1999 Sep;41(3 Pt 1):435-442. [doi: 10.1016/s0190-9622(99)70118-0] [Medline: 10459119]
88. Rathi SK, Kumrah L. Topical corticosteroid-induced rosacea-like dermatitis: a clinical study of 110 cases. Indian J Dermatol Venereol Leprol 2011;77(1):42-46 [FREE Full text] [doi: 10.4103/0378-6323.74974] [Medline: 21220878]
89. Ruzicka T, Assmann T, Lebwohl M. Potential future dermatological indications for tacrolimus ointment. Eur J Dermatol 2003;13(4):331-342. [Medline: 12948911]
90. Saeki H, Watanabe A, Tada Y, Kakinuma T, Komine M, Ihn H, et al. Juvenile pustular psoriasis associated with steroid withdrawal syndrome due to topical corticosteroid. J Dermatol 2008 Sep;35(9):601-603. [doi: 10.1111/j.1346-8138.2008.00531.x] [Medline: 18837708]
91. Sheary B. Topical corticosteroid addiction and withdrawal in a 6 year old. J Prim Health Care 2017 Mar;9(1):90-93. [doi: 10.1071/HC16049] [Medline: 29530193]
92. Sheary B, Tyson C, Harris MF. A qualitative study exploring the views and experiences of adults who decided to cease long-term topical steroid use. Dermatitis 2021;32(6):e131-e133. [doi: 10.1097/DER.0000000000000720] [Medline: 33654020]
93. Sheu HM, Chang CH. Alterations in water content of the stratum corneum following long-term topical corticosteroids. J Formos Med Assoc 1991 Jul;90(7):664-669. [Medline: 1681017]
94. Smith MC, Nedorost S, Tackett B. Facing up to withdrawal from topical steroids. Nursing 2007 Sep;37(9):60-61. [doi: 10.1097/01.NURSE.0000287732.08659.83] [Medline: 17728660]
95. Stegman SJ. Letter: pustular dermatosis on withdrawal of topically applied steroids. Arch Dermatol 1974 Jan;109(1):100. [Medline: 4271968]
96. Tullos K, Barta K, Lio P, Winders T. 171 International survey: effects of cumulative exposure to corticosteroids in patients with eczema, including topical steroid withdrawal syndrome (TSWS). J Invest Dermatol 2022 Aug;142(8):S29. [doi: 10.1016/j.jid.2022.05.178]
97. Velasco RC, Simpson E, Grinich E. 777 Treatment of topical corticosteroid withdrawal. J Invest Dermatol 2023 May;143(5):S133. [doi: 10.1016/j.jid.2023.03.786]
98. Wells K, Brodell RT. Topical corticosteroid 'addiction'. A cause of perioral dermatitis. Postgrad Med 1993 Apr;93(5):225-230. [doi: 10.1080/00325481.1993.11701671] [Medline: 8460079]
99. Weston WL, Morelli JG. Steroid rosacea in prepubertal children. Arch Pediatr Adolesc Med 2000 Jan;154(1):62-64. [Medline: 10632252]
100. Xiao X, Xie H, Jian D, Deng Y, Chen X, Li J. Rebounding triad (severe itching, dryness and burning) after facial corticosteroid discontinuation defines a specific class of corticosteroid-dependent dermatitis. J Dermatol 2015 Jul;42(7):697-702. [doi: 10.1111/1346-8138.12877] [Medline: 25875502]
101. Yoshida A, Sugita K, Yamamoto O. Infiltrative erythemas and nodules on a unilateral cheek following inappropriate use of a topical steroid. Yonago Acta Med 2022 Feb;65(1):88-89 [FREE Full text] [doi: 10.33160/yam.2022.02.003] [Medline: 35221764]
102. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
103. Orr N, Rogers M, Stein A, Thompson CJ, Stein K. Topical steroid withdrawal syndrome: an evidence gap map. EPPI-Mapper. URL: https://eppi.ioe.ac.uk/cms/Portals/35/Maps/tsws.html [accessed 2024-10-26]

XSL•FO
RenderX

104. Guo L, Hsu C, Yu J, Kwatra S, Serrao R, Hanna D, et al. 44453 Identification of diagnostic criteria of topical steroid withdrawal syndrome using a modified Delphi approach. J Am Acad Dermatol 2023 Sep;89(3):AB174. [doi: 10.1016/j.jaad.2023.07.698]
105. Axon E, Chalmers JR, Santer M, Ridd MJ, Lawton S, Langan SM, et al. Safety of topical corticosteroids in atopic eczema: an umbrella review. BMJ Open 2021 Jul 07;11(7):e046476 [FREE Full text] [doi: 10.1136/bmjopen-2020-046476] [Medline: 34233978]
106. Peacock D. TCS overuse and withdrawal reactions: a UK lay-person's perspective. Br J Gen Pract 2023 Mar 30;73(729):174-175. [doi: 10.3399/bjgp23x732453]
107. Teasdale EJ, Muller I, Santer M. Carers' views of topical corticosteroid use in childhood eczema: a qualitative study of online discussion forums. Br J Dermatol 2017 Jun;176(6):1500-1507. [doi: 10.1111/bjd.15130] [Medline: 27753076]
108. Bewley A, Dermatology Working Group. Expert consensus: time for a change in the way we advise our patients to use topical corticosteroids. Br J Dermatol 2008 May;158(5):917-920. [doi: 10.1111/j.1365-2133.2008.08479.x] [Medline: 18294314]
109. Watson J, Crichton M, Affleck AG. Steroid phobia in the ward. J Clin Exp Dermatol Res 2014;05(06):1000252 [FREE Full text] [doi: 10.4172/2155-9554.1000252]
110. Topical steroids: introduction of new labelling and a reminder of the possibility of severe side effects, including Topical Steroid Withdrawal Reactions. Medicines and Healthcare products Regulatory Agency, United Kingdom Government. 2024 May 29. URL: https://www.gov.uk/drug-safety-update/topical-steroids-introduction-of-new-labelling-and-a-reminder-of-the-possibility-of-severe-side-effects-including-topical-steroid-withdrawal-reactions [accessed 2024-10-26]
111. Thelwall S, Thelwall M. Anthropomorphizing atopy: tweeting about eczema. J Dermatol Nurses Assoc 2020;12(2):74-77. [doi: 10.1097/JDN.0000000000000524]

Abbreviations

EGM: evidence gap map
GP: general practitioner
MHRA: Medicines and Healthcare products Regulatory Agency
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TC: topical corticosteroid
TSA: topical steroid addiction
TSW: topical steroid withdrawal
TSWS: topical steroid withdrawal syndrome

Edited by A Schwartz; submitted 29.02.24; peer-reviewed by B Sheary, J Brown, D Peacock, E Choi; comments to author 23.05.24; revised version received 08.08.24; accepted 27.09.24; published 06.12.24.

Please cite as:
Orr N, Rogers M, Stein A, Thompson Coon J, Stein K
Reviewing the Evidence Base for Topical Steroid Withdrawal Syndrome in the Research Literature and Social Media Platforms: An Evidence Gap Map
J Med Internet Res 2024;26:e57687
URL: https://www.jmir.org/2024/1/e57687
doi:10.2196/57687
PMID:

©Noreen Orr, Morwenna Rogers, Abigail Stein, Jo Thompson Coon, Kenneth Stein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Health Insurance Payment for Telehealth Services: Scoping Review and Narrative Synthesis

Qingqing Li[1*], MSc; Feng Cheng[2*], PhD; Huatang Zeng[3], PhD; Junfang Xu[1*], PhD

[1]School of Public Health, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, China

[2]Vanke School of Public Health, Institute for Healthy China, Tsinghua University, Beijing, China

[3]Shenzhen Health Development Research and Data Management Center, Shenzhen, China, China

[*]these authors contributed equally

Corresponding Author:
Junfang Xu, PhD
School of Public Health, Second Affiliated Hospital
Zhejiang University School of Medicine
School of Public Health, the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
Hangzhou, China,
China
Phone: 86 18801230482
Email: junfangxuhappy1987@163.com

Abstract

Background: As telehealth services have demonstrated significant advantages in providing qualified and accessible care, health insurance payments for telehealth services have been issued by various countries. However, the optimization of health insurance payments for telehealth services remains uncertain.

Objective: We conducted a scoping review of the current situation regarding health insurance payments for telehealth services, with the aim of providing evidence to enhance policies related to health insurance payments for such services.

Methods: This scoping review was conducted by comprehensively retrieving data from 6 electronic bibliographic databases from inception to October 2023. The databases included China National Knowledge Infrastructure, Wan Fang, Weipu, Web of Science, PubMed, and Embase, following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Two authors independently assessed search results, extracted data, and evaluated the quality of the included studies using the Critical Appraisal Skills Programme checklist. After the initial screening of titles and abstracts, full texts were obtained and examined. The data regarding the first author, date of publication, country, type of telehealth services introduced in health insurance, health insurance reimbursement providers, reimbursement standards for telehealth (eg, the condition for the reimbursement and reimbursement rate), and key findings of studies were extracted and analyzed. Moreover, we also conducted a narrative synthesis to summarize and report the findings.

Results: A total of 7232 papers were retrieved. Following quality assessment, 23 papers were finally included, with the covered countries including China, the United States, Australia, France, Japan, the United Kingdom, and Germany. The forms of the services vary across different regions, mainly including consultation services, medical monitoring services, mental health services, health education services, and other services. Payment standards are generally categorized into fee-for-service and global budget payment, with clear health insurance payment quotas or proportions and certain restrictions (eg, specifying the location of medical visits and setting the limitation on distance from home to hospitals). The paying entities for health insurance payment include national health insurance and commercial health insurance. In addition, there are 2 kinds of reimbursement rates—a comparable rate for both telehealth and in-person health care services, and a lower rate for telehealth services compared to in-person health care services.

Conclusions: To enhance the accessibility of telehealth services through health insurance payment, it is crucial to further refine the design of health insurance payment for telehealth and strengthen the supervision of services quality, bridging the gap between telehealth and in-person health care services. Additionally, this review did not include studies from all countries, and we recommend that future reviews should include a broader range of countries to provide a more comprehensive view of global telehealth insurance systems.

(J Med Internet Res 2024;26:e56699) doi:10.2196/56699

XSL•FO
RenderX

KEYWORDS

telehealth; internet-based health care; reimbursement; health insurance payment; health care; scoping review; narrative synthesis; mobile phone

Introduction

Telehealth was first proposed by the World Health Organization (WHO) in the 1970s and was defined as the use of video and communication technologies by medical practitioners for the prevention, diagnosis, and treatment of diseases, as well as health counseling, with the aim of promoting the health of individuals and the community. Common forms include telehealth monitoring, telehealth education, telenursing, teleconsultation, and telecare [1]. In 2022, the global telehealth market was valued at approximately US $87.9 billion. By 2023, this valuation had increased to around US $120.4 billion. The market is projected to grow to about US $98.5 billion by 2028, with a compound annual growth rate of 23.2% [2]. Telehealth services have multiple advantages. First, internet technology breaks through the limitations of time and geography, promoting the decentralization of high-quality medical resources to grassroots, remote, and impoverished areas. This facilitates timely and high-quality medical services for patients, addressing issues arising from the uneven distribution of national medical resources, which is significantly important in promoting the fairness of medical resources. Second, it further enhances the mobility of medical resources by leveraging digitalization and informatization. This facilitates a more flexible and efficient circulation of medical resources by promoting the sharing of medical technology knowledge. Third, it meets patients' medical care needs with greater convenience and at a lower cost by providing them with medical services without long-term travel. Especially during the COVID-19 pandemic, the scale of telehealth services has expanded significantly considering the implementation of alleviating the need for face-to-face contact in person (eg, home isolation and lockdowns) to prevent the transmission of COVID-19 [3]. Telehealth services possess several key advantages that enable them to respond actively to sudden public health emergencies (PHE). For example, they can meet the demand for timely and professional medical care of patients while diminishing the risk of virus cross-infection [4]. In 2020, telehealth consultations in hospitals affiliated with China's National Health Commission increased 17-fold compared to the same period in 2019. In the United States, the number of telehealth visits surged by 63 times in 2020 [5]. In the United Kingdom, the usage of video consultations in Scotland experienced a 1000% growth in a 2-week period in March 2020 [6]. Additionally, in Australia, the proportion of consultations conducted through videoconferencing increased from 0.2% in February 2020 to 35% in April 2020 [7].

With the widespread broader adoption of telehealth during the COVID-19 pandemic, public expectations of these services are evolving. Many countries are seizing this opportunity to emphasize the pivotal role of telehealth in the future of health care delivery. This is evidenced by initiatives like expanding internet health care infrastructure, forging partnerships with communication technology companies, diversifying service offerings (eg, telehealth for chronic diseases), and expanding the range of services (eg, telemonitoring and management of chronic diseases), as well as incorporating telehealth services into the reimbursement scope of health insurance.

Different health care systems have been implemented by various countries based on their management capabilities and market needs. The United Kingdom and Australia exemplify the national health insurance model, characterized by government funding and universal coverage. In contrast, Germany, France, and China represent the social health insurance model, where government-regulated funds provide coverage, typically funded through a combination of employer and employee contributions. The United States represents the commercial insurance model, where health care coverage is primarily provided by private insurers. In the process of advancing the implementation of telehealth services, the matter of health insurance payment has been recognized as an obstacle [8-11]. Health insurance payment serves as a compensatory mechanism, empowering more patients to afford the expenses associated with telehealth services. This is particularly crucial for individuals with limited financial means or those in need of long-term medical services. Such support not only stimulates the demand for telehealth services but also fosters an equitable distribution of health care resources, enhancing the efficiency of their use. However, comprehensive evaluations of health insurance payment policies for telehealth remain limited.

To address this gap, we conducted a scoping review to provide an overview of the existing telehealth reimbursement policies across different countries, highlighting the various models and key considerations in designing these policies, such as the scope of coverage and payment standards. Additionally, with the advancement of internet technology, the global use of telehealth is steadily increasing, and many countries are actively promoting telehealth to enhance health care accessibility and strengthen their health care insurance systems. Introducing telehealth into health insurance reimbursement has become a significant trend. This scoping review offers insights into the major trends, challenges, and gaps in global telehealth reimbursement policies. By reviewing the current characteristics of telehealth insurance coverage, payment standards, and other relevant factors, this study provides a theoretical foundation for the development of telehealth reimbursement policies. It also serves as a valuable reference for best practices and policy recommendations.

Methods

Overview

We conducted this scoping review following the methodological framework proposed by Arksey and O'Malley [12], which comprises five stages: (1) identifying research questions; (2) relevant studies; (3) study selection; (4) data charting; and (5) collating, summarizing, and reporting the results. The reporting of the scoping review was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [13].

XSL•FO
RenderX

Narrative synthesis is a method for evaluating and synthesizing the results of multiple studies, primarily relying on the use of descriptive text to summarize and interpret research findings [14]. The application of narrative synthesis enables a clear presentation of the intricacies of telehealth reimbursement policies, thereby providing a robust foundation for the design of comprehensive reimbursement frameworks for telemedicine services.

Search Strategy and Study Selection

We conducted a comprehensive search across China National Knowledge Infrastructure, Wan Fang, Weipu Database, Web of Science, PubMed, and Embase. Our search strategy used both Medical Subject Headings terms and free words from the inception of these databases until October 2023. The search algorithms (see Multimedia Appendix 1) were (telehealth OR telehealth OR Virtual medicine OR eHealth OR mHealth OR online diagnosis OR mobile health) AND (health insurance OR payment OR medicare OR reimbursement). References of the included studies were also reviewed to identify additional eligible studies.

The inclusion criteria for this scoping review included research related to telehealth research, involving topics such as telemedical services, internet hospitals, and diagnosis and treatment; research focused on the health insurance payment policies for telehealth services; research with clear health insurance payment methods including the scope, mechanisms, and entities involved in health insurance payments; and research with quantitative or qualitative methods. The exclusion criteria for this scoping review included nonresearch studies like news reprints and opinion pieces, inability to access the full text, research that only pertains to telehealth without specific mention of health insurance payment policies or practices, and research whose perspective is the combination of internet technology and health insurance payment.

The search process involved 3 stages—collection of studies, scanning of titles and abstracts, and reading of the full texts. Titles and abstracts retrieved from searches were independently screened by 2 authors. Full-text papers were also reviewed independently by the same 2 authors. Any disagreements were resolved through a joint meeting with a senior author to reach a consensus.

Data Extraction and Charting the Data

The data were extracted from studies that met the eligibility and inclusion criteria for this review and recorded using a standardized Microsoft Excel data charting table. The data regarding the first author, date of publication, country, type of telehealth services introduced in health insurance, health insurance reimbursement providers, reimbursement standards for telehealth (eg, the condition for the reimbursement and reimbursement rate), and key findings of studies (see Multimedia Appendix 2) were extracted and analyzed.

Collating, Summarizing, and Reporting the Results

We assessed the quality of each study using the PRISMA-ScR checklist (see Multimedia Appendix 3) and the Critical Appraisal Skills Programme (CASP) checklist (see Multimedia Appendix 4) [15]. Each criterion was scored as 1 (fully meeting the checklist standard) or 0 (not meeting or unclear about the checklist standard), resulting in a quality score for each study based on all criteria. Each scoping analysis received a total quality score of 8 points, and the components include title analysis, abstract, introduction, methods analysis, results analysis, and discussion analysis. Two researchers independently conducted the quality assessment of all papers, consulting a third researcher when necessary to reach a consensus. We used descriptive narrative synthesis combined with tabular summarization to present and interpret the research findings.

Basic Definitions

The coverage area of health insurance payments refers to the geographical coverage of telehealth health insurance. The covered services under health insurance payment represent the types of services eligible for health insurance reimbursement. Standards of health insurance reimbursement for telehealth include the reimbursement rates and eligibility criteria for insurance coverage.

Results

Characteristics of the Studies

A total of 7232 studies were initially identified in the database, consisting of 594 in Chinese and 6635 in English. Following the application of PRISMA-ScR criteria and deduplication using reference management software, a total of 5260 studies were retained (Figure 1). After strict adherence to inclusion and exclusion criteria, 23 studies were ultimately included [16-38], with 8 in Chinese and 15 in English. The selected studies spanned research conducted in the United States (n=8, 34%), China (n=6, 26%), Australia (n=3, 23%), France (n=3, 23%), Japan (n=2, 9%), Germany (n=2, 9%), and United Kingdom (n=1, 4%).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart for the selection of the included studies. CNKI: China National Knowledge Infrastructure; VIP: Weipu Database.

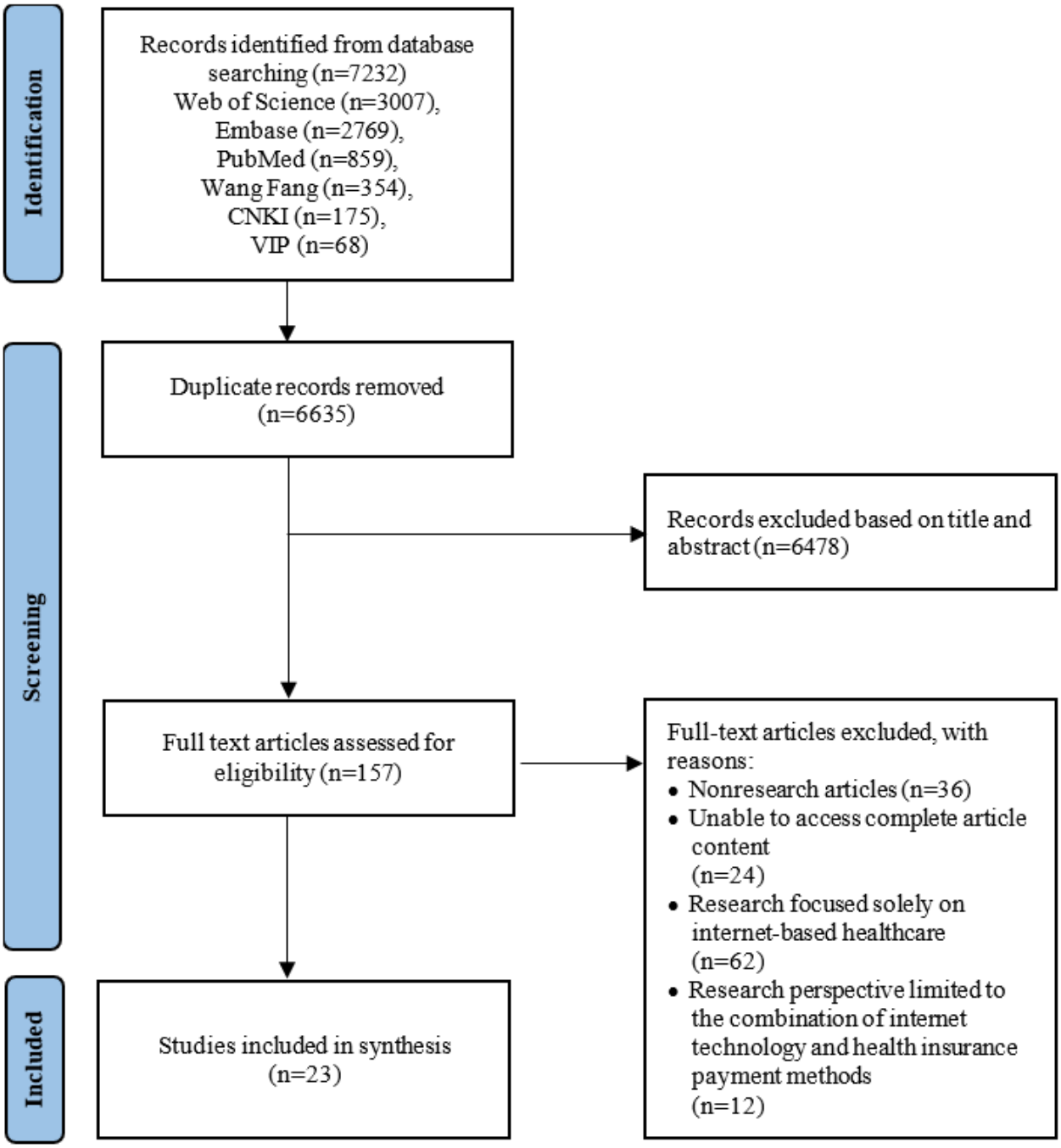

Coverage Area of Health Insurance Payments

A total of 10 (43%) studies provided insights into the coverage of health insurance payments for telehealth across the examined regions. In China, the majority of provinces, cities, and districts have achieved fundamental coverage of telehealth under basic health insurance [16-19].

In the United States, the introduction of telehealth services in Medicare reimbursement categories dates back to the Balanced Budget Act of 1997. Initially, Medicare stipulated that the originating site (where patients receive telehealth services) must be located in health professional shortage areas or outside metropolitan statistical areas. However, in 2018, the Bipartisan Budget Act and the Patients and Communities Act lifted restrictions on the originating site, allowing it to be in any location. Currently, both Medicare and Medicaid reimburse telehealth services in all 50 states and the District of Columbia, with Medicaid allowing 14 states to consider homes as originating sites and 16 states allowing schools as originating sites. In 43 states and the District of Columbia, private payment laws for telehealth services have been enacted [4,20-22].

In Australia, telehealth services were introduced in the scope of health insurance reimbursement in 2011. Before the COVID-19 pandemic, only insured individuals residing in remote areas 2-5 could use telehealth services and receive

XSL•FO
RenderX

reimbursement under the National Medicare Benefits Schedule (MBS). However, during the pandemic, reimbursement expanded to all insured individuals [23,24].

France introduced telehealth services in the scope of health insurance reimbursement in 2018. Insured individuals can be reimbursed for telehealth services just like regular health care. To ensure the treatment of telehealth services, patients who have undergone at least 1 in-person consultation within a year are permitted to use telehealth services. After the outbreak of the COVID-19 pandemic in 2020, all patients became eligible to access telehealth services [24,25].

In Germany, the Electronic Health Law in 2015 supported the adoption of telehealth, explicitly allowing reimbursement for telehealth services. In 2019, the Digital Healthcare Act (Digitale-Versorgung-Gesetz) granted all individuals covered by statutory health insurance the right to be reimbursed for eligible digital health applications, including video consultations, teleconsultations, remote monitoring, and remote secondary diagnoses. Digitale-Versorgung-Gesetz continues to promote the use of telehealth, benefiting all insured patients [26].

Covered Services Under Health Insurance Payment

A total of 18 (78%) studies provided insights into the covered items under health insurance payments for telehealth services in the studied countries (as shown in Tables 1 and 2). In China, the definition of health insurance reimbursement scope varies across different regions based on local actions [16-18,27,28]. Covered services mainly included follow-up visits, remote consultations, remote monitoring, and remote diagnosis. Most regions introduced “electronic follow-up” services into health insurance payment, followed by “electronic consultation” services.

Table 1. Telehealth services covered by health insurance.

Region	Provider	Telehealth services
China	Basic Health Insurance	• Telehealth follow-up services • Remote consultation: single-discipline, multidisciplinary, synchronous, and asynchronous pathology consultation; traditional Chinese medicine syndrome differentiation; and treatment consultation • Remote monitoring: fetal heart monitoring, remote ECG[a] monitoring, pacemaker monitoring, and defibrillator monitoring • Remote diagnosis: imaging diagnosis, pathology diagnosis, ECG diagnosis, and ultrasonography diagnosis
United States	Medicare, Medicaid, Private, Payer, and MA[b] plans	• Medicare: remote physiologic monitoring and treatment management, care management services (transitional care and chronic care), interprofessional health record consultations, psychological therapy, diagnosis and treatment of kidney diseases, medication assessment and management, and stroke telehealth • Medicaid: store-and-forward (33 states), remote patient monitoring (34 states), audio-only calls (37 states and Washington, DC), live video (all 50 states and Washington, DC), and transmission or facility fee (35 states). Specific services limitations in some states • Private payers: telehealth services reimbursed similarly to in-person services under parity state laws. Some states expanding commercial reimbursement policies • MA plans: alternative to traditional Medicare, allowing private insurers an option to offer telehealth services under the PFS[c] and following PFS usage rules. In addition, MA plans provide coverage for services via telehealth beyond what is required in original Medicare as additional telehealth benefits (eg, remote monitoring services for urban patients with multiple chronic diseases)
Australia	MBS[d]	• 2006-2019: sequential introduction of telehealth services into MBS. Video consultations for remote areas, mental health support for drought and bushfire-affected communities and advanced care for aged, cancer, and palliative patients • COVID-19 epidemic: expansion of telehealth services in MBS. Early intervention, care for older people, pediatric, psychiatric, psychological, and dental diagnostic and treatment services
France	Assurance Maladie	• Remote consultation: no professional restrictions on telehealth services • Tele-expertise: a remote consultation among practicing physicians to remotely seek advice on patient treatment • Remote monitoring: use of follow-up, particularly for remote monitoring of organ transplant patients
Japan	National health insurance	• Remote consultation: among medical institutions and practicing physicians

[a]ECG: electrocardiogram.

[b]MA: Medicare Advantage.

[c]PFS: Physician Fee Schedule.

[d]MBS: Medicare Benefits Schedule.

XSL•FO
RenderX

Table 2. Regions covered by health insurance.

Region and telehealth services	Covered regions
China	
Follow-up services	Beijing, Zhejiang, Jiangxi, Jilin, and 27 other provinces and municipalities
Remote consultation	Guangdong, Shanxi, Fujian, and 16 other provinces and municipalities
Remote monitoring	Sichuan, Tianjin, Shandong, and 16 other provinces and municipalities
Remote diagnosis	Jiangsu, Guizhou, Chongqing, and 7 other provinces and municipalities
United States	
Medicare	Nationwide
Medicaid	Nationwide
Private payers	40 states and the District of Columbia
Medicare advantage plans	Nationwide
Australia	
2006-2019	Nationwide
COVID-19 epidemic	Nationwide
France	
Remote consultation	Nationwide
Tele-expertise	Nationwide
Remote monitoring	Nationwide
Japan	
Remote consultation	Nationwide

In the United States, health care insurance reimbursement is primarily through Medicare, Medicaid, private payer, and Medicare advantage plans, with specific payment items taking different forms between insurance types and regions. The Centers for Medicare and Medicaid Services (CMS) in the United States has established a list of Medicare telehealth services for reimbursement, which includes 268 permanent or temporary internet health care service items. These services typically cover consultations, remote monitoring, and mental health services. The list is updated annually and was expanded during the COVID-19 pandemic to include telehealth reimbursement items, such as remote cardiac and pulmonary rehabilitation therapies, enabling insured individuals to receive treatment services provided by clinicians from homes [17,20,29,30].

Australia primarily uses the MBS for health insurance payments, incorporating 11 specialized services (geriatrics, psychiatry, neurosurgery, obstetrics, gynecology, etc), 6 primary health care consultations, and 23 patient-end services. To address the COVID-19 pandemic, MBS gradually introduced internet health care service items and introduced approximately 300 MBS project numbers for remote health care consultations. Among them, 56 are for primary health care and 73 temporarily covered health care services during the pandemic have been permanently included in the health insurance plan [23,24,31].

In France, telehealth services are based on the scope of national health insurance, without specific subspecialty restrictions. During the pandemic, the insurance coverage for remote health care services was expanded. The Ministry of Health allows national health insurance (Assurance Maladie) to reimburse remote consultations and remote specialized knowledge for patients diagnosed with COVID-19 without the need for prior registration [24,25,32].

In the United Kingdom, the development of telehealth was relatively slow before the COVID-19 pandemic. After the outbreak, the surge in demand for telehealth prompted the National Health Service (NHS) to request supporting institutions such as health trusts and general practitioners to provide reimbursement support for telephone or digital video remote consultations [33].

Japan introduced remote health care services between medical service providers in health insurance payment, such as remote pathology and remote radiology diagnoses. For instance, in cases where a patient's pathological tissue sample is received at the destination facility (affiliated hospital), and the examination requires the use of remote diagnostic technology from the sending facility (other medical institution), health insurance will cover the costs incurred by the sending facility for the rapid microscopic examination of the pathological tissue. Remote medical care for specific diseases is eligible for health insurance reimbursement, and some services of medical-patient remote medical care have gradually been introduced in health insurance payments after the COVID-19 pandemic [34].

Standards of Health Insurance Reimbursement for Telehealth

A total of 10 (43%) studies provided insights into the standards of health insurance reimbursement for telehealth across various

XSL•FO
RenderX

countries and health insurance payment standards for telehealth exhibit the following characteristics in different nations (as shown in Table 3).

The first characteristic is the establishment of health care insurance reimbursement quotas or ratios. In China, different provincial-level administrative regions have set varying insurance quotas or ratios. Some provinces and municipalities have set reimbursement amounts according to the level of hospitals or set a certain percentage of reimbursement according to the price of the service for health insurance [16,17,27]. In the United States, the CMS has established different payment systems for various entities, including the Physician Fee Schedule (PFS) and Fee-For-Service (FFS). Each system has distinct health insurance reimbursement conditions for telehealth medical services. The specific payment standards for each internet medical service are explicitly outlined in the PFS. Medicare relies on the payment standards of the PFS, compensating providers based on the unit of service delivered [4,20-22,30]. In France, national health insurance reimburses 70% of the costs for telehealth medical services, with individuals covering the remaining 30%. Certain populations, such as pregnant women and long-term care patients, receive full reimbursement from national insurance. During the COVID-19 pandemic period, the national health insurance fully covers diagnosed and suspected cases [24]. In Australia, the charging standard for telehealth services is set at 50% of the fee for in-person consultations. Physicians receive compensation through a bulk billing system, which means that patients incur no charges for receiving telehealth consultations. Physicians directly bill the MBS, and the MBS covers the full amount. Additionally, after the COVID-19 pandemic, there has been an expansion of full reimbursement for telehealth services, particularly benefiting the disadvantaged [24,31].

Table 3. The health insurance payment standards for telehealth services.

Region	Reimbursement quotas or ratios	Specific conditions
China	• Reimbursement expenses per visit are claimed based on the hospital level. Beijing: follow-up services after reimbursement, ¥19 (approximately US $2.6) for primary hospital, ¥28 (approximately US $3.9) for secondary hospital, and ¥40 (approximately US $5.5) for tertiary hospital. Jiangxi: primary and tertiary hospitals can be reimbursed ¥13 (approximately US $1.8) and ¥9 (approximately US $1.2), respectively, for the follow-up services. • According to the basic health insurance category A and category B list. Zhejiang: follow-up visits are in accordance with category A list. Henan: in accordance with category B list. • Reimbursement is based on a certain percentage of the services price. Fujian: the reimbursement of consultations is set at 30% of the services price. Guizhou: the reimbursement of consultations is set at between 70% and 85% of services price for patients in relatively impoverished counties.	• Reimbursement of specific services: Sichuan: reimbursement for remote ECGa and pacemaker monitoring is limited to patients with arrhythmia. • Payment limits: Shaanxi and Guizhou: set annual payment limits. Fujian: a maximum of no more than ¥90 (approximately US $12.4) per visit. • Visit duration limits: Shandong: reimbursement for remote monitoring service (excluding fetal heart monitoring) is limited to 24 hours per visit.
United States	• The payment standards are listed in the physician fee schedule and are updated on an annual basis	• Payment recipients: Medicare: aged 65 years and older, all patients with end-stage renal disease, and some individuals with disabilities. Medicaid: children, pregnant women, and the impoverished. • Service locations: the originating site: within a Health Professional Shortage Area or a county outside of a metropolitan statistical area. In 2019, Centers for Medicare and Medicaid Services removed the geographic requirement for originating sites for patients with substance use disorder or co-occurring mental health.
Australia	• Bulk-billing: service providers have the option to directly bill Medicare for their services without charging additional fees to patients.	• Service locations: at least 15 km away from medical institutions
France	• Under normal circumstances, the national health insurance covers 70% of the cost of telehealth services. Special populations (eg, pregnant women, long-term care patients, and covered by supplemental health insurance payment support plans) receive full coverage. • During crises like the COVID-19 pandemic, the national health insurance covers 100% of the cost of telehealth services.	• Patients who had in-person consultations within the past year services. • Remote consultation equipment. Computer, tablet, or smartphone equipped with a webcam and connected to the internet, transmitting video data through a secure video platform. • During the COVID-19 pandemic, allowance for the use of telephone for suspected patients with COVID-19, pregnant women, residents in “white zones” (areas without mobile phone or internet coverage), patients with chronic disease, and individuals aged older than 70 years.

[a]ECG: electrocardiogram.

XSL•FO
RenderX

The second characteristic is setting specific conditions. In some provinces and cities in China, restrictive conditions (eg, specific services and payment limits) for health insurance payment for telehealth services are designed to ensure the quality and safety of medical care. In the United States, there are conditions on populations and locations for health care [20,22,29]. France specifies that patients who have received in-person treatment or have been referred by primary care physicians can use reimbursed telehealth services. Patients must use communication devices with cameras and internet connectivity. However, conditions were relaxed after the COVID-19 pandemic, eliminating the mandatory requirement for patients to have in-person initial consultations and allowing certain populations to consult via telephone to address inequalities in digital access [24,31]. In Australia, telehealth services are not accepted for hospitalized patients. Moreover, eligibility for health insurance reimbursement is limited to patients located at least 15 kilometers away from medical institutions who receive such services [23,24].

Differences in Health Insurance Payments for Telehealth Services and In-Person Services

A total of 4 (17%) studies reported the differences in health insurance payment between telehealth and in-person health care services in various countries. First, some regions in China such as Sichuan and Shandong provinces, adopt the same health insurance payment standard for telehealth as for in-person services [27]. Similarly, countries like France, Luxembourg, the Netherlands, and Switzerland maintain equivalent payment standards for telehealth consultations compared to in-person services [32]. Additionally, states in the United States, including Tennessee and Michigan, have enacted laws establishing parity for telehealth, guiding reimbursement for telehealth consultations [36]. When formulating parity policies, factors such as resource use and costs related to telehealth services, demand for alternative payment models, and differences in outcomes between telehealth care and in-person care are typically taken into consideration [37]. Second, telehealth services often adopt reimbursement levels lower than those for in-person services [35]. A survey of telepsychiatry consultations provided in 17 countries and territories found that the degree of telepsychiatry reimbursement rates from public health insurance was lower than in-person service in many countries, such as Japan, even after the outbreak of the COVID-19 pandemic [38].

Discussion

Principal Findings

The review identified the current status of health insurance payments for telehealth services. Our study found that an increasing number of countries are incorporating telehealth into their health insurance systems. However, there are significant differences in reimbursement standards and conditions across different countries.

The COVID-19 pandemic has significantly accelerated the global adoption of telehealth, profoundly impacting health care service models. Many countries have expanded their health insurance policies to include telehealth, recognizing its importance in maintaining the continuity of care during the pandemic. As detailed in the results, several key changes occurred during the pandemic [10,11,22,24,33,34]. The pandemic prompted many health insurance systems to expand their coverage to include a broader range of telehealth services, thereby addressing the increased demand for remote consultations. Additionally, geographical restrictions were often relaxed, allowing patients in more regions to access reimbursed telehealth services.

Moreover, the reimbursement rates for these services were generally increased to ensure they were on par with the in-person consultations. These adjustments were crucial in maintaining access to essential health care services while minimizing the risk of COVID-19 transmission. Overall, the pandemic has accelerated the integration of telehealth into mainstream health care delivery, highlighting the need for adaptable health insurance policies capable of responding to PHE.

The study found that some countries, such as the United States and Australia, have more detailed distinctions in reimbursement items, considering further stratification for specialized services. This is somewhat related to the diagnosis-related groups (DRG) health insurance payment policies implemented by these countries. Such reimbursement strategies help achieve standardization of medical services and rational use of medical resources, ensuring that patients can receive appropriate medical care. However, there are few researches on the design of health care insurance reimbursement frameworks for telehealth services. Many countries lack unified standard regulations, which grants local governments significant flexibility to integrate telehealth services into the existing health insurance systems. Determining reasonable payment standards has always been a topic of keen interest across various sectors of society.

We found that some regions adopt reimbursement methods identical to in-person services, integrating telehealth insurance reimbursement into the overall in-person health care insurance budget for unified management [39]. This approach can promote the development of telehealth and better integrate telehealth services with in-person medical services.

The results also revealed that telehealth services included in health insurance have gradually diversified and the number of disciplines covered has increased, which is a reflection of the digital transformation of health care services and the adaptive adjustment of health insurance policies. It is worth noting that some regions are incorporating telehealth follow-up services for chronic diseases into health insurance reimbursement [27,30]. Encouraging patients with chronic diseases to choose telehealth follow-up appointments can save their time, as well as reduce the wastage of hospital resources. Receiving services is the first part of health insurance payment, but there still leaves many services unaddressed such as some prescription drugs and maternity care, which are not covered by health insurance reimbursement.

Some studies have emphasized the barriers faced in the actual implementation of health insurance reimbursement for telehealth. First, it faces risks to medical quality, safety, and information security, necessitating further protection of patient

XSL•FO
RenderX

rights. Moreover, in the open environment of the internet, there is a risk of third-party theft and leakage of patient consultation information, health insurance records, and other personal privacy. Second, in the absence of targeted supervision and information asymmetry, it may lead to excessive health care consumption and fraudulent claims for health insurance reimbursement, which can result in unreasonable growth in medical expenses, leading to waste of health insurance funds and inefficiency in the allocation of medical resources [40].

Comparisons to Existing Literature

Our results are consistent with existing findings. Policy makers often introduce additional flexibility to facilitate patient reimbursement for telehealth services during the PHE. These adjustments were widely adopted during the pandemic, leading to the rapid expansion of telehealth, which has now become an integral part of routine health care in many regions. The lessons learned from these flexible policies have prompted policy makers in various countries to reevaluate telehealth reimbursement strategies, with considerations to maintain or further enhance these measures in the post-pandemic era to support the long-term growth of telehealth [41]. The differing definitions of telehealth may lead to inconsistencies in subsequent regulations, guidelines, and reimbursement policies [42]. As our study shows, the variation in telehealth definitions across regions results in differences in the services covered by insurance and the associated payment policies. These disparities could lead to significant differences in the development and adoption of telehealth services across regions, presenting challenges for the implementation of telehealth services across regional and national borders.

For payers of health insurance, a key concern is whether telehealth will increase expenditures [43]. The impact of telemedicine on costs is also influenced by the differences in expenses between telehealth and in-person services. Our research shows that most countries or regions have aligned the reimbursement levels for telehealth with those for equivalent in-person visits. The level of reimbursement for telehealth depends on factors such as the patient population, health care environment, specific conditions, and modes of interaction. The ultimate goal of health insurance payment policies is to integrate telehealth and in-person services effectively, aiming to balance and harmonize these 2 service models to enhance accessibility to medical services while managing the balance of insurance revenues and expenditures.

Policy Implications

Our study had several policy implications. First, it is necessary to consider various factors such as medical demand, fairness, efficiency, and the financial capacity of the health insurance fund to establish scientific and feasible admission standards and formulate a unified and reasonable catalog of telehealth services covered by health insurance. Second, a well-implemented health insurance payment mechanism is crucial for promoting the widespread adoption of telehealth services with a lower economic burden on patients. Telehealth, as an emerging form of health care service, relies on the national health insurance policy framework for its insurance policy formulation. Existing payment methods include FFS, capitation, global budget payment, DRG, and diagnosis-intervention packet. Each payment method has its advantages and disadvantages in different countries and health care systems. FFS can encourage doctors to provide medical services and take on risks [44] but may lead to inefficient over-provision or under-provision of some medical services, as well as rapid increases in health care costs [45]. Capitation and global budgets can effectively control the growth of health care costs and improve efficiency but may impact the quality of medical services [44]. DRG and diagnosis-intervention packets help control health care costs and improve service quality but can inevitably lead to issues such as upcoding and cherry-picking patients [46,47]. Different countries are at varying stages of telehealth development, and the suitable health insurance payment methods differ accordingly. In the initial stage, where telehealth is in its infancy, the technical foundation is weak, regulations are underdeveloped, and public awareness and acceptance are low. At this stage, FFS payments, prepaid card systems, and single payment channels are recommended as they are easy to manage and understand. As the technical foundation of the internet improves, telehealth gradually becomes more widespread. The number of users increases, and overall health care costs rise, placing higher demands on the management and use of health insurance funds. Therefore, we recommend implementing capitation, global budget payment methods, and DRG, during this stage, as these can effectively control the rapid growth of health care expenses. In the advanced stage, where internet technology is mature and telehealth is highly developed, the goal is to reduce the growth rate of health care costs while improving medical quality and health outcomes. Therefore, we suggest adopting pay-for-performance methods and linking payment amounts to medical quality and performance to establish a more comprehensive and integrated payment system. Currently, many countries are exploring value-based payment methods to optimize existing health insurance payment models. Overall, we think that the enhancement of telehealth technology is crucial for the improvement of the health insurance payment system. Furthermore, policy makers should integrate the above payment methods according to the characteristics of different telehealth services and consider adopting mixed payment models to optimize deductibles, maximum limits, and copayment ratios in telehealth insurance payments to achieve better results. Third, enhancing the connectivity between telehealth and in-person health insurance payments, and reducing the disparity in reimbursement ratios between telehealth and in-person services to encourage the development of telehealth services are essential. Finally, we recommend promoting the construction of intelligent supervision systems for health insurance, establishing monitoring rules and indicators, and using big data to conduct real-time supervision and control of diagnosis and treatment behaviors, which will ensure the secure use of health insurance funds.

Limitations

Due to limitations in resources and data availability, this review did not include the health care insurance systems for some countries. The inclusion of studies was limited to those available in English or Chinese, which may have led to the omission of some relevant studies published in other languages. Finally, the

XSL•FO
RenderX

review only conducted a qualitative analysis of the included studies, and the quality of the methods reported may introduce bias into the research findings.

Conclusions

The variability in reimbursement practices we observed highlights the need for more standardized guidelines to support the development of telehealth services. We recommend that future scoping reviews broaden the scope and perspective of research to include a wider range of countries, providing a more comprehensive understanding of global telehealth insurance systems. This approach would offer more precise insights for improving telehealth insurance policies, particularly in countries and regions with varying income levels.

Acknowledgments

This work was supported by Special Research Program of Pharmacoeconomics and Health Technology Assessment Committee of Zhejiang Pharmaceutical Association (2022ZYJ06 and 2023ZYJ02) and Sanming Project of Medicine in Shenzhen (SZSM202111001). The authors are grateful to all study participants for their valuable participation.

Authors' Contributions

QL conducted the literature review and drafted the paper. FC, HZ, and JX contributed to the interpretation of the data and critical revisions of the paper. The authors all approved the final versions for submission. All authors have agreed with the content and approved the submission of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 18 KB - jmir_v26i1e56699_app1.docx]

Multimedia Appendix 2
Characteristic of the included studies and data extraction.
[DOCX File , 40 KB - jmir_v26i1e56699_app2.docx]

Multimedia Appendix 3
PRISM-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 108 KB - jmir_v26i1e56699_app3.docx]

Multimedia Appendix 4
Critical Appraisal Skills Programme (CASP) checklist.
[DOCX File , 26 KB - jmir_v26i1e56699_app4.docx]

References

1. Guan X, Liu L, Zhu H. The enlightenment of telehealth practice in the United States to China. Chin Health Serv Manag 2019;36(8):565-568 [FREE Full text]
2. Telehealth market - global forecast to 2028. Markets and Markets. URL: https://www.marketsandmarkets.com/Market-Reports/telehealth-market-201868927.html [accessed 2024-11-05]
3. Benjenk I, Franzini L, Roby D, Chen J. Disparities in audio-only telemedicine use among medicare beneficiaries during the coronavirus disease 2019 pandemic. Med Care 2021;59(11):1014-1022 [FREE Full text] [doi: 10.1097/MLR.0000000000001631] [Medline: 34534186]
4. Bajowala SS, Milosch J, Bansal C. Telemedicine pays: billing and coding update. Curr Allergy Asthma Rep 2020;20(10):60. [doi: 10.1007/s11882-020-00956-y] [Medline: 32715353]
5. Seshamani M. Medicare and telehealth: delivering on innovation's promise for equity, quality, access, and sustainability. Health Aff (Millwood) 2022;41(5):651-653. [doi: 10.1377/hlthaff.2022.00323] [Medline: 35500188]
6. Webster P. Virtual health care in the era of COVID-19. Lancet 2020;395(10231):1180-1181. [doi: 10.1016/S0140-6736(20)30818-7] [Medline: 32278374]
7. Thomas EE, Haydon HM, Mehrotra A, Caffery LJ, Snoswell CL, Banbury A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2022;28(4):301-308. [doi: 10.1177/1357633X20960638] [Medline: 32985380]

XSL•FO
RenderX

8. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
9. Weinstein RS, Lopez AM, Joseph BA, Erps KA, Holcomb M, Barker GP, et al. Telemedicine, telehealth, and mobile health applications that work: opportunities and barriers. Am J Med 2014;127(3):183-187. [doi: 10.1016/j.amjmed.2013.09.032] [Medline: 24384059]
10. Lieneck C, Weaver E, Maryon T. Outpatient telehealth implementation in the United States during the COVID-19 global pandemic: a systematic review. Medicina (Kaunas) 2021;57(5):462 [FREE Full text] [doi: 10.3390/medicina57050462] [Medline: 34065050]
11. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020;26(5):309-313 [FREE Full text] [doi: 10.1177/1357633X20916567] [Medline: 32196391]
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
13. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
14. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme. Lancaster, United Kindgom: Lancaster University; 2006.
15. Critical Appraisal Skills Programme (CASP) checklists. Critical Appraisal Skills Programme. 2023. URL: http://www.casp-uk.net/ [accessed 2024-11-05]
16. Li G, Yao Q, Liu X. Research on medical insurance payment policy of internet plus medical service in China. Health Econ Res 2023;40(5):43-46 [FREE Full text]
17. Ding T, Qian A, Tan Z. Comparative analysis of domestic and foreign internet medical service pricing and medical insurance payment policies at home and abroad. Chin Hosp 2022;26(9):10-13 [FREE Full text]
18. Liang HM, Xu JJ, Peng Y. Analysis on the core issues of pricing and payment of internet medical services in China. Chin Health Econ 2020;39(12):52-56 [FREE Full text]
19. Cui W, Zhang K, Gu S. Policy suggestions on inclusion of internet medical services into medical insurance payment. Chin Health Resour 2020;23(2):102-105 [FREE Full text]
20. Liu Y, Li Y, Jiang X. Analysis and enlightenment of health insurance payment policy for internet medical services in the United States. Chin J Hosp Administration 2021;37(8):642-646 [FREE Full text]
21. Park S, Langellier BA, Burke RE. Telehealth benefits offered by medicare advantage plans in 2020. Med Care 2021;59(1):53-57. [doi: 10.1097/MLR.0000000000001408] [Medline: 32925464]
22. Abbasi-Feinberg F. Telemedicine coding and reimbursement - current and future trends. Sleep Med Clin 2020;15(3):417-429. [doi: 10.1016/j.jsmc.2020.06.002] [Medline: 32762974]
23. Fisher K, Davey AR, Magin P. Telehealth for Australian general practice: the present and the future. Aust J Gen Pract 2022;51(8):626-629 [FREE Full text] [doi: 10.31128/AJGP-11-21-6229] [Medline: 35908757]
24. Suk KJ, Mi LS. A comparative analysis on current status of telemedicine policy before and after COVID-19: focused on France and Australia. J Converg Soc Public Policy 2022;16(3):129-160. [doi: 10.37582/cspp.2022.16.3.129]
25. Seon JK. A study on the legislation of telemedicine in France. Korean Soc Law Med 2022;23(2):141-169 [FREE Full text]
26. Gerke S, Stern AD, Minssen T. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. NPJ Digit Med 2020;3:94 [FREE Full text] [doi: 10.1038/s41746-020-0306-7] [Medline: 32685700]
27. Cui W, Zhang K, Gu S. Research on internet medical service included in medical insurance reimbursement. Chin Hosp 2020;03:24-33 [FREE Full text] [doi: 10.19660/j.issn.1671-0592.2020.03.02]
28. Ma S, Cao Y. Relevant research and policy suggestions on the inclusion of internet plus medical in health insurance payment. Modern Business Trade Ind 2021;35:35-42 [FREE Full text]
29. Xiao GJ, Xu HF. Analysis and enlightenment of medicare payment Internet medical service policy in America. Soft Sci Health 2022;36(8):90-96 [FREE Full text]
30. Bogulski CA, Acharya M, Pro G, Ali MM, Rabbani M, Hayes CJ, et al. A state profile of disparities in telehealth utilization among medicare beneficiaries: an intersection between race/ethnicity, rurality, and chronic conditions-Arkansas, 2019. Telemed J E Health 2023;29(12):1759-1768. [doi: 10.1089/tmj.2023.0012] [Medline: 37074340]
31. Hall Dykgraaf S, Desborough J, de Toca L, Davis S, Roberts L, Munindradasa A, et al. "A decade's worth of work in a matter of days": the journey to telehealth for the whole population in Australia. Int J Med Inform 2021;151:104483. [doi: 10.1016/j.ijmedinf.2021.104483] [Medline: 33984625]
32. Raes S, Trybou J, Annemans L. How to pay for telemedicine: a comparison of ten health systems. Health Syst Reform 2022;8(1):2116088. [doi: 10.1080/23288604.2022.2116088] [Medline: 36084277]

XSL•FO
RenderX

33. Fisk M, Livingstone A, Pit SW. Telehealth in the context of COVID-19: changing perspectives in Australia, the United Kingdom, and the United States. J Med Internet Res 2020;22(6):e19264 [FREE Full text] [doi: 10.2196/19264] [Medline: 32463377]
34. Kim JS, Lim SM. A comparative analysis on current status of telemedicine policy before and after COVID-19: focused on France and Australia. CSPP 2022 Nov 30;16(3):129-160. [doi: 10.37582/cspp.2022.16.3.129]
35. Wilson FA, Rampa S, Trout KE, Stimpson JP. Reimbursements for telehealth services are likely to be lower than non-telehealth services in the United States. J Telemed Telecare 2017;23(4):497-500. [doi: 10.1177/1357633X16652288] [Medline: 27260264]
36. Neufeld JD, Doarn CR, Aly R. State policies influence medicare telemedicine utilization. Telemed J E Health 2016;22(1):70-74. [doi: 10.1089/tmj.2015.0044] [Medline: 26218148]
37. Khera N, Knoedler M, Meier SK, TerKonda S, Williams RD, Wittich CM, et al. Payment and coverage parity for virtual care and in-person care: how do we get there? Telemed Rep 2023;4(1):100-108. [doi: 10.1089/tmr.2023.0014] [Medline: 37283856]
38. Kinoshita S, Cortright K, Crawford A, Mizuno Y, Yoshida K, Hilty D, et al. Changes in telepsychiatry regulations during the COVID-19 pandemic: 17 countries and regions' approaches to an evolving healthcare landscape. Psychol Med 2022;52(13):2606-2613. [doi: 10.1017/S0033291720004584] [Medline: 33243311]
39. Jia HB, Lu JQ. Challenges and countermeasures of basic medical insurance payment in "internet plus" medical services. Health Econ Res 2023;06:40-46 [FREE Full text]
40. Cui W, Gu S, Cun D. Risks and management suggestions after? Internet +? Medical services are included in medical insurance payments J. Chin Hosp 2020;03:24 [FREE Full text]
41. Saravanakumar S, Ostrovsky A. Evaluation of telehealth services that are clinically appropriate for reimbursement in the US medicaid population: mixed methods study. J Med Internet Res 2024;26:e46412 [FREE Full text] [doi: 10.2196/46412] [Medline: 38546706]
42. McCauley JL, Swartz AE. Reframing telehealth: regulation, licensing, and reimbursement in connected care. Obstet Gynecol Clin North Am 2020;47(2):227-239. [doi: 10.1016/j.ogc.2020.02.001] [Medline: 32451014]
43. Mehrotra A, Uscher-Pines L. Informing the debate about telemedicine reimbursement - what do we need to know? N Engl J Med 2022;387(20):1821-1823. [doi: 10.1056/NEJMp2210790] [Medline: 36373824]
44. Robinson JC. Theory and practice in the design of physician payment incentives. Milbank Q 2001;79(2):149-177, III. [doi: 10.1111/1468-0009.00202] [Medline: 11439463]
45. McClellan M. Reforming payments to healthcare providers: the key to slowing healthcare cost growth while improving quality? J Econ Perspect 2011;25(2):69-92. [doi: 10.1257/jep.25.2.69] [Medline: 21595326]
46. Peng HR. Provider payment reforms and the pursuit of healthy China. J Sun Yat-sen Univ 2023;63(2):14-22 [FREE Full text]
47. Peng Y, Jin C, Wang H. Experience and enlightenment of DRG payment system reform in the United States. Chin Health Econ 2018;37(7):93-96 [FREE Full text]

Abbreviations

CASP: Critical Appraisal Skills Programme
CMS: Centers for Medicare and Medicaid Services
DRG: diagnosis-related group
FFS: Fee-For-Service
MBS: Medicare Benefits Schedule
NHS: National Health Service
PFS: Physician Fee Schedule
PHE: public health emergencies
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Schwartz; submitted 24.01.24; peer-reviewed by L Guo, A Trivedi; comments to author 09.04.24; revised version received 03.06.24; accepted 17.10.24; published 09.12.24.

Please cite as:
Li Q, Cheng F, Zeng H, Xu J
Health Insurance Payment for Telehealth Services: Scoping Review and Narrative Synthesis
J Med Internet Res 2024;26:e56699
URL: https://www.jmir.org/2024/1/e56699
doi:10.2196/56699
PMID:

©Qingqing Li, Feng Cheng, Huatang Zeng, Junfang Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Psychological Factors That Contribute to the Use of Video Consultations in Health Care: Systematic Review

Helen M Haydon[1,2], BPsy, Registered Psychologist, AFHEA, DipDementiaCare, PhD; James A Fowler[3], BPsychSc(Hons); Monica L Taylor[1,2], MPH, BSc; Anthony C Smith[1,2,4], BN, MEd, RN, PhD; Liam J Caffery[1,2], BInfoTech, DipAppSci, PhD

[1]Centre for Online Health, The University of Queensland, Woolloongabba, Australia

[2]Centre for Health Services Research, The University of Queensland, Woolloongabba, Australia

[3]School of Public Health, Faculty of Medicine, The University of Queensland, Brisbane, Australia

[4]Centre for Innovative Medical Technology, University of Southern Denmark, Odense, Denmark

Corresponding Author:
Helen M Haydon, BPsy, Registered Psychologist, AFHEA, DipDementiaCare, PhD
Centre for Online Health
The University of Queensland
Ground Floor, Building 33
Princess Alexandra Hospital
Woolloongabba, 4102
Australia
Phone: 61 7 3176 4462
Email: h.haydon@uq.edu.au

Abstract

Background: There are numerous benefits to delivering care via video consultations (VCs). Yet, the willingness of health care professionals (HCPs) to use video as a modality of care is one of the greatest barriers to its adoption. Decisions regarding whether to use video may be based on assumptions and concerns that are not necessarily borne of evidence. To effectively address psychological barriers to VC, it is essential to gain a better understanding of specific factors (eg, attitudes, beliefs, and emotions) that influence HCPs' VC use.

Objective: This study's aim was to conduct a systematic literature review of psychological factors in HCPs that impair or promote VC use.

Methods: Databases were searched in February 2023 for peer-reviewed primary research papers on VCs that discussed psychological factors of health professionals affecting the use of video to deliver health services. A psychological factor was defined as an intraindividual influence related to, or in reaction to, VC use—in this case, the individual being an HCP. Search terms included variations on "telehealth," "clinician," and psychological factors (eg, attitude and beliefs) in combination. Peer-reviewed papers of all methodological approaches were included if they were in an Australian setting and the full text was available in English. Studies where the main intervention was another digital health modality (eg, remote monitoring and telephone) were excluded. Studies were also excluded if they only reported on extrinsic factors (eg, environmental or economic). Information extracted included author, year, medical specialty, psychological component mentioned, explanation as to why the psychological factor was related to VC use, and exemplar quotes from the paper that correspond to a psychological component. Each extracted psychological factor was classified as a positive, negative, ambivalent, or neutral perspective on VC, and a thematic analysis then generated the factors and themes. Theories of behavior were considered and discussed to help frame the interaction between themes.

Results: From 4592 studies, data were extracted from 90 peer-reviewed papers. Cognitive and emotional motivators and inhibitors, such as emotional responses, self-efficacy, attitudes, and perceived impact on the clinician as a professional, all interact to influence HCP engagement in VCs. These factors were complex and impacted upon one another. A cyclical relationship between these factors and intention to engage in VCs and actual use of VCs was found. These findings were used to form the psychological attributes of VC engagement (PAVE) model. Evidence suggests that HCPs fall within 4 key user categories based on the amount of cognitive and practical effort needed to deliver VCs.

XSL•FO
RenderX

Conclusions: Although further research is needed to validate the current findings, this study provides opportunity for more targeted interventions that address psychological factors impeding effective use of VCs.

(J Med Internet Res 2024;26:e54636) doi:10.2196/54636

KEYWORDS

telehealth; video consultations; psychological factors; health professionals; health services research; psychological; video; systematic review; review; barriers; engagement; health services; cognitive; consultations; psychological barrier

Introduction

Telehealth can provide advantages over traditional care delivery, including greater access to care; increased convenience; less disruption to activities of daily living; and lower costs for the patient and, in some cases, the health care professionals (HCPs) [1,2]. Additionally, there are societal benefits and environmental benefits to delivering care by telehealth [3,4]. Furthermore, when practiced appropriately, telehealth is noninferior and, in some instances, superior to usual care [5]. Hence, it would seem logical that telehealth would be a mainstay of health care delivery. However, this is not the case [6].

Clinician barriers are acting as impediments to the widespread adoption of telehealth. The seminal paper by Wade et al [7] concluded that clinician willingness to use telehealth was the key factor of successful telehealth. Building on this work, further research has identified that avoidance or hesitancy with telehealth, specifically the use of video consultations (VCs), results from HCPs' worries, assumptions, or preconceived frames of reference (eg, assuming the patient will not have access to technology) as opposed to evidence-based on empirical research [8,9]. These authors have proposed pragmatic strategies to redress adverse cognitive and emotional reactions to VC, thereby increasing VC use. We know from previous literature that HCPs have mixed feelings about VCs, and attitudes can be quite context specific [10,11]. Some feel it improves patient prognosis and successfully supports their decision-making, and some face challenges with physical exam capabilities and the ability to order further tests [10]. However, a lack of willingness when there are no physical or practical barriers infers psychological factors at play [7]. To effectively address these psychological hurdles, it is essential to comprehend the specific factors (eg, attitudes, beliefs, and emotions) that influence HCPs' VC use.

To the best of our knowledge, there have been no studies that have synthesized the psychological factors that influence HCPs' use of VC. Common theoretical frameworks used to explain technology adoption include the technology acceptance model (TAM), unified theory of acceptance and use of technology, and technology organization environment frameworks [12-14]. Although used in the field of health and well-being, these frameworks are not specific to HCPs or health care organizations, and they include external variables influencing adoption. Australia is one country where there has been particularly limited uptake of VCs in health care compared with other high-income nations. For example, in 2022, while large health services in the United States, such as the Veterans Health Administration, were using video for over 50% of their mental health visits [15], less than one-quarter of Medicare-funded mental health appointments were done by phone or video in Australia [16]. The aim of this study was to specifically identify psychological factors in HCPs in Australia that impair or promote the use of VCs.

Methods

Operational Definition of a Psychological Factor

Psychological factors were defined as intraindividual influences related to, or in reaction to, VC use. We considered a range of psychological definitions, theories, and frameworks and scoped the literature for factors related to concepts such as cognitions, attitudes, beliefs, self-efficacy, intentions to use, motivation, preferences, and emotions.

Search Strategy and Eligibility Criteria

This review protocol was not registered. Searches were run in the Embase, MEDLINE, CINAHL, and PsycINFO databases from inception to February 2, 2023. Inclusion and exclusion criteria were based on our predefined PICO (population, intervention, context, outcome) framework. Search terms included variations on "telehealth" (virtual and video), "clinician" (nurse and physician), and psychological factors (attitude and beliefs) combined. Peer-reviewed papers of all methodological approaches were included if they were in an Australian setting and the full text was available in English. Other inclusion factors were that the intervention involved synchronous video appointments between a patient and HCPs, and that the outcomes reported intrapersonal or intrinsic psychological factors that related to an HCP's use of VC. The full search strategy for each database is included in Multimedia Appendix 1.

Studies where the main intervention was another digital health modality such as remote monitoring or phone appointments were excluded. Studies were also excluded if they only reported on extrinsic factors such as environmental or economic factors affecting HCPs use of VC. If the results of a study did not clearly differentiate between HCPs and patient factors, or between video and other modalities of care, it was excluded.

Screening and Data Extraction

Search results were uploaded into Covidence software (Veritas Health Innovation) to perform the screening stage. Each paper's title and abstract were screened by 2 reviewers independently and excluded if they did not meet the criteria. The full text of the remaining papers was also screened by 2 reviewers, resulting in a final list of included studies. Disagreements between reviewers (eg, whether the included factors were psychological in nature) were resolved by the remainder of the research team at regular meetings. Given the high volume of papers and

XSL•FO
RenderX

in-depth level of extracting required, the included full-text papers were divided between authors for manual data extraction. First, a random sample of 5 papers was chosen for data extraction and completed by all reviewers independently. Results were then compared and checked to ensure consistency in the process. The remaining papers were then divided across reviewers with regular meetings to check for consistency and discuss any anomalies. Information extracted included author, year, medical specialty, psychological component mentioned, explanation as to why the psychological factor was related to VC use, and exemplar quotes from the paper that correspond to a psychological component. An example of a psychological component could have been "positive attitude toward using video consultations as they are patient centred." One paper may have multiple psychological factors mentioned. Notes were also made if papers reported how much VC experience staff had, and whether staff underwent specific VC training.

Data Analysis

For each psychological factor extracted, it was noted whether it represented a positive, negative, ambivalent, or neutral perspective on VC. Psychological component data points were then sorted according to this classification. Common themes across papers were categorized and summarized descriptively. A separate study team member was responsible for each classification. Primarily, the thematic analysis, which followed Braun and Clarke's [17] approach, was inductive, with constant comparison of journal paper text and several peer debriefing meetings, resulting in the themes generated [18] and the proposed relationships between them. In the final stage analysis, during these peer debriefing meetings, theories of behavior were considered and discussed to help frame the interaction between themes and further explore patterns in the findings. Several theories were discussed including the theory of planned behavior (TPB) [19], the stages of change [20], and the normalization process theory [21] that potentially aligned with the inductive findings. Reporting followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [22].

Risk of Bias Assessment

Each included paper was critically appraised using the Mixed Methods Appraisal Tool [23]. This checklist was selected as it can be used for most study designs, and we expected multiple study types to be included in our review. It was specifically developed for use during the study assessment stage of systematic reviews and has been evaluated and validated. The rating outcomes were examined by criterion and results were described descriptively as calculating an overall score is not advised by the tool creators.

Results

Summary of Included Studies

An initial yield of 4592 studies was refined into 90 included papers [7,24-112]. Figure 1 contains an overview of the search and screening processes. Studies were published between 2001 and 2023, with more than two-thirds (62/90, 69%) being published in the last 3 years. The papers were most often focused on allied health topics including mental health (n=11), speech-language pathology (n=8), physiotherapy (n=6), and overall rehabilitation (n=7). There were also many studies that included a variety of HCPs across different contexts and disciplines (n=12). Staff positions varied both within and between studies, such as nurses, general practitioners, and specialists. The study designs included 43 qualitative studies, 34 mixed methods studies, 8 nonrandomized controlled trials, 4 quantitative descriptive studies, and 1 randomized controlled trial.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the screening process.

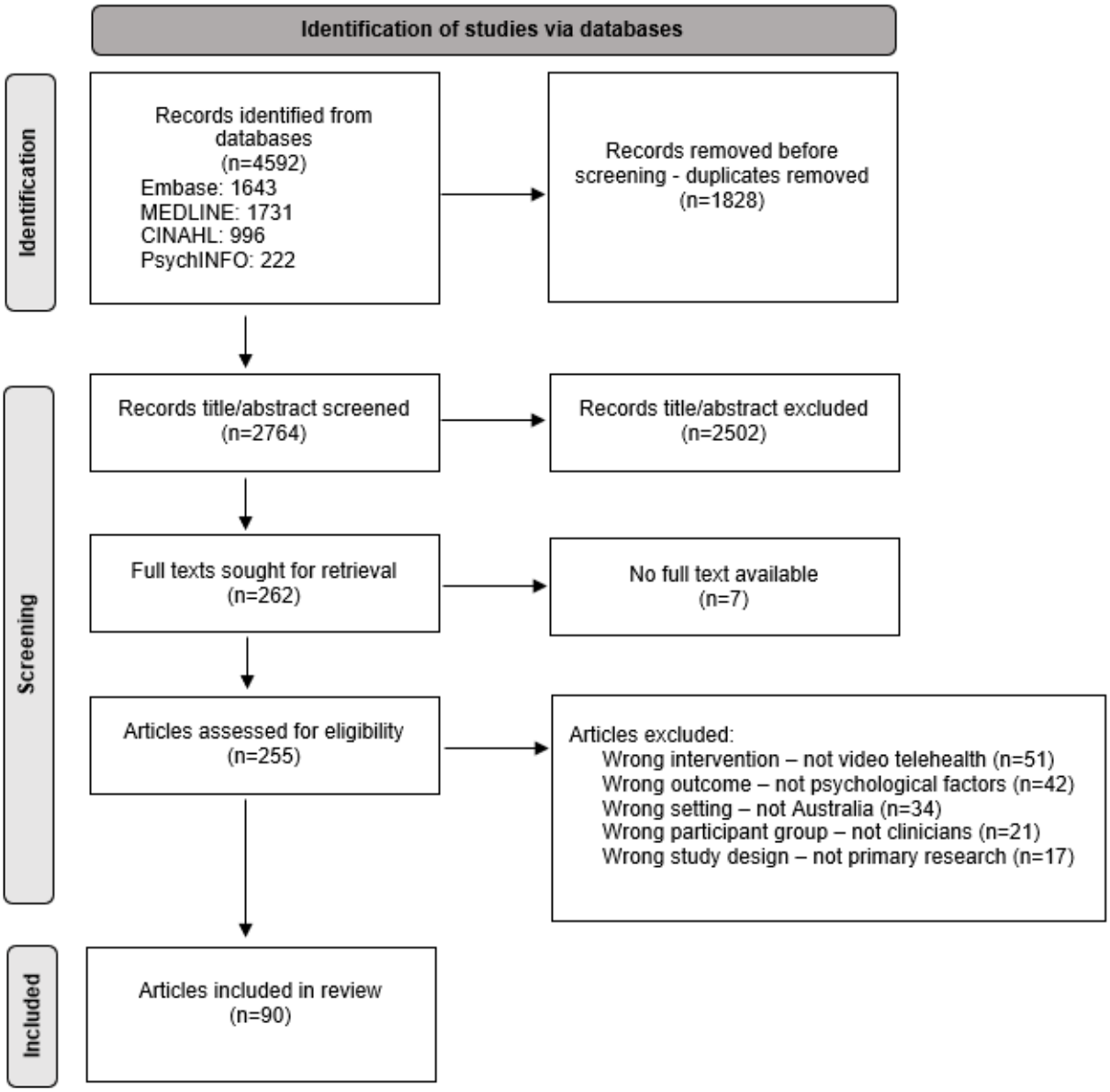

The Psychological Attributes of VC Engagement

Cognitions refer to mental processes involved with memory, attention, problem-solving, and perceptions, whereas emotions refer to feelings or affective states [113]. Literature suggests that the constructs of emotion and cognition (eg, affective attitudes) are highly related but can be considered independent factors [113,114]. Unsurprisingly, our analysis found that cognitive and emotional motivators or inhibitors to engage in VCs were complex, overlapping, and impacting upon one another. Using the findings from this study, we developed the psychological attributes of video consultation engagement (PAVE) model (Figure 2), which highlights the cyclical relationship between cognitive and emotional contributors (including those pertaining to clinician identity and professional role) and engaging in VCs. Intention to use VCs and its relationship with habit and the actual behavior of engaging in VCs are also presented. It should be noted that examples included in the diagram represent the most dominant themes and what was most common in the systematic review but do not necessarily account for all that was found.

The following sections describe the psychological factors and themes that were generated, but it must be emphasized that these constructs are interactive and related. Further, a central psychological factor (eg, attitudes) could be both a motivator (pro-VC) or inhibitor (anti-VC). For instance, fear (anxiety) of doing VCs is influenced by one's attitudes toward engaging in that behavior, which can, in turn, reinforce negative perceptions of VCs and diminish confidence to deliver VCs. Please note that when describing the psychological factors present in the literature, we do not critically examine or cross-reference evidence to support or refute such perspectives.

XSL•FO
RenderX

Figure 2. The psychological attributes of video consultation engagement (PAVE) model showing interacting cognitive and emotional contributors to engaging in video consultations and their perceived impacts on the clinician. Examples included (eg, emotions and attitudes) in this diagram represent the most dominant themes and common examples from the systematic review and do not account for all that was found.

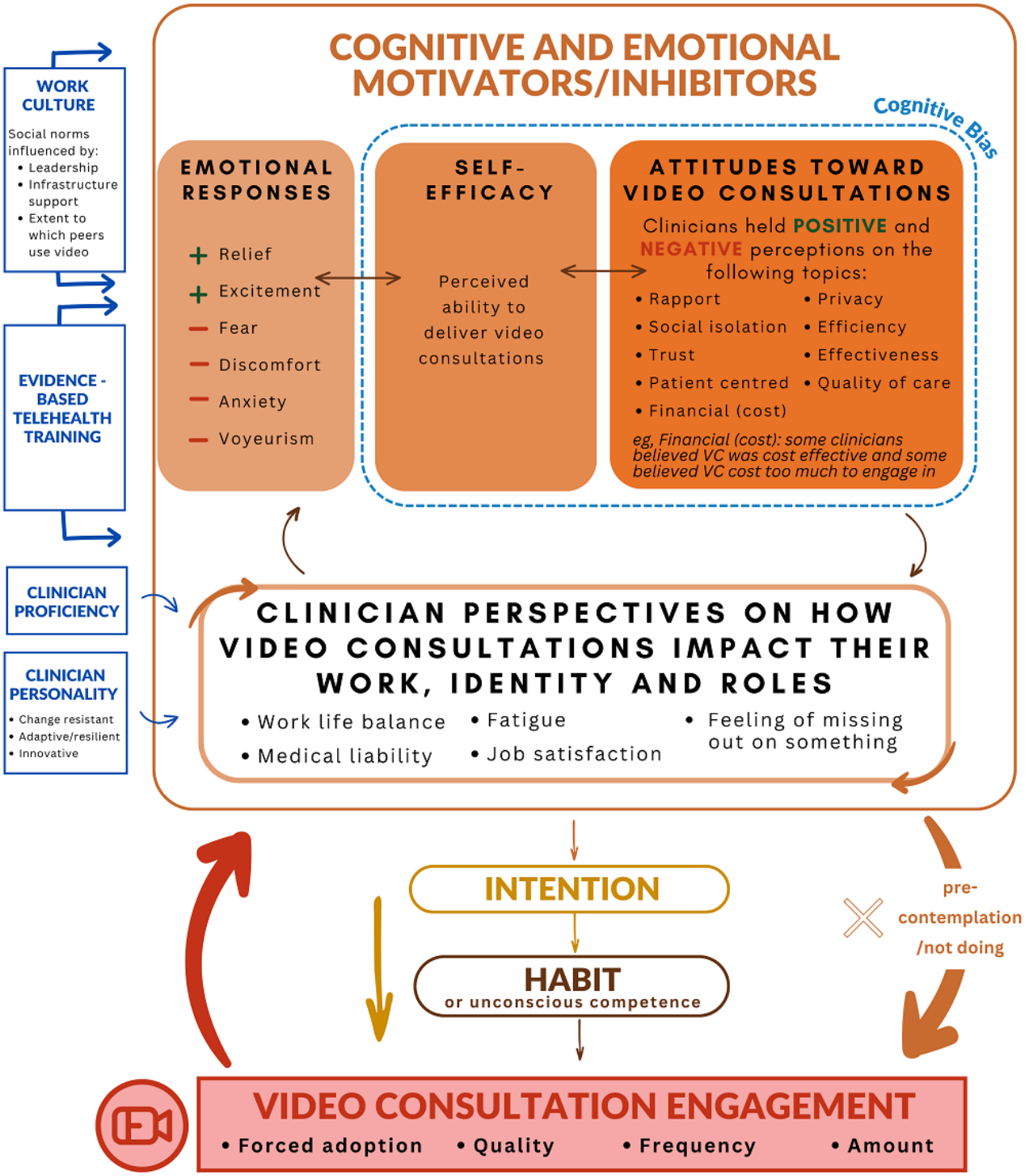

Cognitive and Emotional Motivators or Inhibitors

Four main themes or intraindividual factors embody the majority of the literature results: emotional responses; self-efficacy; attitudes toward VCs; and subjective impact on clinician—role and identity (top rectangle in PAVE model). As suggested by the title, these factors can either motivate or inhibit the use of VCs or the intention to use video.

Emotional Responses

Both positive and negative emotional responses to using VCs were expressed by HCPs across the literature. HCPs' fear, anxiety, and discomfort were often mentioned or alluded to, across the research [47,53,56,71,79,85,87,91,95-97,99,101-104,115]. Conversely, excitement and relief were also responses to doing VCs [49,86]. The cyclical nature was evident when HCPs initially found it uncomfortable to do VCs but after

XSL•FO
RenderX

persisting, they were relieved that it was something they no longer needed to think about (reduced cognitive effort) [55,73,78,87]. An unexpected finding was the sense of voyeurism that HCPs sometimes reported in the literature [56,78]. For example, HCPs uncomfortable "watching" a patient through a screen as they reveal sensitive information or body parts (eg, genitals) [56]. Feelings of voyeurism were also reported when HCPs felt like they were invading a patient's privacy, due to seeing into their living quarter [78].

Attitudes Toward VCs

Attitudes refer to positive or negative personal evaluations of providing care via video [19]. As seen in Figure 2, a range of attitudes toward VCs were found across the literature mostly pertaining to the 9 topics listed in the figure. Of note, HCPs' attitudes, often about the same topic, were polarized and not necessarily evidence based. The following examples highlight the disparity in findings but do not include the full range of attitudes identified within the literature. Some HCPs believe you cannot build rapport with consumers via VCs (which is contrary to consumer perceptions) [26,29,35,51,54,70,73,76,78,81,83,84,86,106,107,111,112,116], while other HCPs perceive VCs as being beneficial to rapport-building [61,87,91,92,97,107]. Some HCPs believe VCs increase social and professional isolation [32,51,53,70,74,84,115], while others perceive it to decrease isolation and increase clinical collaboration [24,59,61,93,95,97,98,102,110,117]. Some HCPs view VCs as a barrier to health care for consumers [26,47,58,93,99,103,104,106,108,111], while others perceive VCs as patient centered and promoting health care access [26,40,46,48,49,51,52,54,55,60,73,78,81,84,92,96,97,101,107]. Some described an attitude that telehealth would reduce the quality of care provided to patients, would be disruptive, or is inferior to physical care [26,33,35,36,48,52,56,58,60,63,70,71,75,76,80-82,85,86,93,96-98,103,104,107,108,111,112,118], whereas others felt it would improve the quality of care provided and the lives of patients [7,26,33,45,51,77,78,84,85,87,92,100,101,107,110-112]. As described by Sutherland et al [99], attitudes can be dependent on the context of use, with their findings showing HCPs were comfortable using video broadly, but not for completing assessments.

Self-Efficacy

Self-efficacy was identified as a pivotal construct that interacted with a range of attitudes toward and emotional responses from engaging in or delivering VCs. We define self-efficacy as HCPs' beliefs that they could provide high-quality telehealth care and overcome technological challenges that may inhibit this [119]. Many HCPs remarked that they did not feel confident in delivering VCs [53,56,79,84,87,95,96,98,101,102,104,107,112,115]. For example, Pitt et al [87] described participants as being unable to manage technology or hardware problems if they were to emerge. However, persevering through barriers to providing VCs (such as some of the emotional factors described above) can build self-efficacy and evoke a positive emotional response [73,78,85,87,115]. For example, one HCP described at the beginning using video seemed "impossible" (low self-efficacy), but over time they became so equipped in its use they "no longer think about it" (high self-efficacy) [78]. Other research has shown that positive attitudes towards telehealth can also predict whether HCPs were willing to build self-efficacy to problem-solve technology failures [40].

Cognitive Bias

Considering the interaction between these themes and the way in which decisions to use VCs are not always based on evidence but upon assumptions, cognitive bias is illustrated as a filter around the cognitive motivators and inhibitors. Cognitive biases are described as simplified information-processing strategies that may cause individuals to form incorrect perspectives [120]. A previous review has highlighted that within medical contexts, HCPs may be influenced by biases such as confirmation bias (intentionally selecting information that meets prior attitudes), or the bandwagon bias (a tendency to believe what the majority believe) [121]. Recent work by Cook et al [9] (published outside of this review period) has highlighted that many HCPs felt telehealth to be clinically inappropriate and not meeting consumer demands, despite evidence suggesting the contrary. This suggests that cognitive biases around telehealth being inappropriate were subconsciously driving telehealth uptake, rather than empirical evidence. Therefore, while the role of cognitive biases was not explored within the scope of the included papers, broader psychological theory highlights the role that they may have in VC uptake, particularly in driving the polarized attitude towards which appears disconnected from available evidence.

How Clinicians Perceive VCs Impact Them Directly, Their Identity, and Roles

The final theme within the cognitive and emotional factors is the perceived impacts on the clinician, their identity, and professional roles. Although the elements within this theme can be considered cognitions or emotions, there was sufficient evidence within the literature to warrant a theme that was specifically related to the profession of being a clinician. While other factors were paradoxical, generally factors related to clinician impact were negative. For example, there was some evidence suggesting that clinicians feared that telehealth would take away their independence, the need for their roles, and broadly modify models of care to make clinicians less important [33,98,104,107,108,110,112,117]. There was evidence that some clinicians perceived VCs as increasing their workload [29,49,62,70,73,75,79,95,97,99,101,103,104,108,110,111]. However, some participants considered this positive, as workload increase stemmed from increased patient loads [59]. VCs also increase fatigue or cognitive load [35,73,104,106,111,116] and medical liability [75,107]. Some also expressed loss of job fulfillment because of videoconferencing [26,112,115], such as a genetic counselor who no longer had the opportunity to deliver good news as this component was reassigned to a different clinician within their model of care [112]. Finally, there are mixed results regarding job satisfaction. Some studies found VCs increased job satisfaction and a sense of achievement [52,59,98], and others reported decreased job satisfaction [104,112,115].

XSL•FO
RenderX

Intention and Habit

Intentions, from a TPB vantage point, are direct antecedents to behavior and are influenced by an individual's perceptions of social pressures, perceived control, and attitudes toward providing VCs [19]. Numerous studies assessed intention [28,31,72,92,93,101]. Findings were mixed regarding the reliability of intentions to predict actual engagement (behavior) in delivering VCs. For example, participants in the study by Swales et al [101] found that despite high interest in using VC, few provided VC services. Another study by Brunelli et al [34], which explored the closely related construct of attitudes, found similar results. As such, we drew upon TPB models that included habits that could account for such results. The role of habit is evident when HCPs revert to traditional and practiced behaviors when providing care via telephone to avoid increased emotional or cognitive effort. For example, in the study by Moffatt and Eley [79], many HCPs reported using phone calls rather than videoconferencing because they were quicker and required less cognitive effort to learn. Other participants in this study described having preferences for the "traditional approach" because they lacked the confidence (emotional effort) to learn videoconferencing.

Engaging in VCs

Overall, the cognitive and emotional factors outlined above seem to have a direct impact on whether HCPs use video conferencing to provide consumers with health care that would otherwise be provided in person. They also have an indirect impact (as described above) through intentions. That is, research showed 4 main categories, and potentially 4 main "user categories" (Table 1) as follows: first, HCPs influenced by their attitudes (often negative), self-efficacy (low), and emotions (negative) do not seriously contemplate (are precontemplative) delivering care via video and therefore do not engage [33,79,102]. For example, participants in the study by Moffatt and Eley [79] described "just not wanting to learn technology." Second, HCPs who exert cognitive effort and plan and think about doing VCs (intentions) but for varied reasons do not engage in VCs [71,111]. For example, participants in the study by White et al [111] described how VCs require "preparation." Third, HCPs who exert cognitive effort and plan and think about doing VCs (intentions) before engaging in VCs. These HCPs are at a crossroads. If they use video so infrequently [36,111] like many skilled behaviors, the effort and planning remain because it is not practiced, and there is the potential for the behavior to stop. For example, White et al [111] described HCPs exerting effort to plan and deliver videoconferencing, but after ongoing technology issues they "abandoned" the format altogether. However, if HCPs have good experiences and engage in VCs regularly, they can progress to type 4. Fourth, HCPs who are so practiced and confident at delivering VCs that it takes minimal cognitive and practical effort (eg, high self-efficacy and effective supporting infrastructure) and has become business as usual (unconscious competence or habit) [49,52,55,78]. For example, Gelber [55] described participants as saying telehealth is "now second nature."

Table 1. Four user categories based on the amount of cognitive and practical effort needed to deliver video consultations.

	Low/no effort (eg, cognitive and practical)	High effort (eg, cognitive and practical)
Not engaged	• Type 1. Characterized by one or more of the following: negative attitudes, low self-efficacy, and negative emotional reactions to use • Precontemplation: No engagement	• Type 2. Cognitive effort and thinking about doing video consultations (intentions) but does not translate to behavior sometimes due to infrequent use and loss of skill and confidence • No engagement or limited or infrequent use means reverting to habit eventuating in extinguished behavior
Engaged	• Type 4. Practiced and confident clinicians push through barriers with minimal cognitive and practical effort (eg, high self-efficacy and effective supporting infrastructure) • Business as usual or unconscious competence	• Type 3. Cognitive effort and think about doing video consultations (intentions) before engaging in video consultations • Engage or persist

The extent to which clinicians engage in VCs can impact the emotional responses, self-efficacy, attitudes, and perceived professional impact. As described by Ayres et al [26], practitioners struggled due to their variations of experiences and an inability to "predict" the next consultation outcome. Specifically, the amount and frequency of engaging in VCs, as shown by studies where HCPs were initially reluctant, but after use, continued to use it more [85,87,94,102,109]. Technology factors, such as the quality of the technology, can create frustration and contribute to technology anxiety or fears [49,55,56,71,100,104,111]. For example, one study reported that audiovisual interruptions can make practitioners "anxious" [53]. However, these can also be influenced by self-efficacy and how confident HCPs were to work with technology issues [7,73]. For example, Lawson et al [73] described technology issues as a way to "break the ice", and Wade et al [7] described how HCPs may accept VC despite technology issues. The audiovisual quality of VCs is important if HCPs are anxious about appearing incompetent to their patient [56,102]. Forced adoption or pressure to use video can lead to resentment toward VCs [102-104]. Similarly, the rapid uptake of telehealth as part of COVID-19 social distancing requirements [6] meant that HCPs were unprepared, untrained, and unsupported, which led to negative responses and low self-efficacy [71,90,91,101,102,104]. In comparison, the "forced adoption" as a result of COVID-19, also exposed HCPs to the benefits of VCs and led to acceptance and more regular use [102].

XSL•FO
RenderX

Other Factors Impacting the Cognitive and Emotional Motivators or Inhibitors

Finally, several factors were found in the literature that are important to consider and overlap or influence the cognitive and emotional factors. Work culture and the social norms influenced by leadership, infrastructure support, and peer influence can influence attitudes toward VCs and self-efficacy [70,76,85,91,102,103]. Similarly, adequate evidence-based VC training can increase self-efficacy and elevate attitudes toward VC and emotional responses [70,86]. More stable factors, like personality, are represented by research that discusses HCPs being change-resistant or adaptive and innovative [26,61,66,86,102,111], all of which may impact how a person responds to or views VCs.

Quality Assessment Results

Overall, there was good methodological quality in the included studies. The majority (50/90, 56%) were positively rated on all appraisal categories for their respective study types. For qualitative studies, there was not always coherence between the data collected, analysis, and interpretation. Coding processes and qualitative theories could have been expanded upon in more detail. In many mixed methods studies (18/34, 53%), the quantitative and qualitative data were not actually integrated, so it was not clear what the benefit was of a mixed methods study over just 2 separate studies. Nonrandomized studies often did not clearly discuss confounders. Full quality assessment rating results are available in Multimedia Appendix 2.

Discussion

Overall Findings

The aim of this systematic review was to identify HCP psychological factors that influence the use of VC for the delivery of health care. Research literature describes a range of complex and interacting psychological factors that influence whether HCPs engage in VCs. Cognitive and emotional motivators and inhibitors, such as emotional responses, self-efficacy, attitudes, and perceived impact on the clinician as a professional, all interact to influence HCP engagement in VCs. In turn, experiences of engaging in VCs impact and shape the cognitive and emotional motivators and inhibitors. Potentially mediating this cyclical relationship are intention and habit.

Managing Attitudes and Emotions, and Other Technology Acceptance Literature

Attitudes, self-efficacy, and intention to use have been considered in other reviews [122] and are central to established models of technology acceptance, such as the TAM [123-125]. The TAM is based on the TPB [19], and like the underpinning behavioral theory, has been modified to examine potential psychological contributors to behavior [126]. For instance, extensions have included an examination of technology-related self-efficacy and its relationship to attitudes and intention to use a technology [127,128]. Interventions and training to increase knowledge and self-efficacy and ultimately VC use have been developed and have contributed to some increases in use [129]. The factor or theme we label as attitudes toward VCs, especially when considering the extent to which VCs are perceived to be effective or efficient, seems to overlap with the TAM construct of perceived usefulness. The research literature presents a range of HCP attitudes toward VCs and how these may be barriers or motivators to do VCs. Attitudes are predominantly polarized being for or against VCs. The lack of consistency across attitudes highlights their complexity and interaction with other factors. For instance, HCPs were divided on whether or not the use of VCs: could reduce or increase social isolation; was patient centered or prohibitive to patients; saved money or was costly; was effective or efficient; was trustworthy; increased or decreased privacy; and reduced or improved quality of care. Finally, there was a range of perspectives regarding the ability to build rapport via video. Regardless of direction, attitudes toward VCs impact and are informed by emotional responses and self-efficacy, which then impact the intention to use VCs or the actual use.

Although there is varied terminology and conceptualizations of emotions (eg, as affective attitudes), the bidirectional interplay between attitude and emotion is well established [130]. Emotions are integral in the processing of information and the formation of attitudes [131,132] and, in turn, attitudes impact emotional responses [133]. Specifically, there is a large body of work that demonstrates the relationship between emotions, especially technology anxiety, and its impact on willingness to engage in technology-related activities such as VCs [79,86,134,135]. Exposure therapy is one way to mitigate HCP anxiety [136], but simply repeated positive experiences and support engaging with VCs can reduce discomfort and lead to relief and sometimes even excitement [49,86,87].

Addressing Health Professional Confidence and Identity

Overlapping with attitudes, emotions, and self-efficacy, there were factors that focused specifically on roles and identity as an HCP. There were numerous papers mentioning fears that telehealth would reduce professional independence and make their roles redundant. Concerns also extend to feelings they could not properly control the patient or the consultation, which influences their willingness to engage in telehealth and the perceived usefulness of health technologies [137]. Therefore, to facilitate future VC uptake, it is important to use appropriate training and education. Training should upskill HCPs to improve their ability to conduct and control VCs effectively. Education should relay the vast evidence demonstrating that technology does not necessarily lead to skill reduction and job loss, but rather, it is most likely to augment workforce capability, freeing up HCPs to conduct advanced scope work [138]. Most importantly, mitigating negative impacts on workload and professional roles can be achieved through comprehensive implementation strategies such as amending workflows and increasing administration time.

Changing HCPs' Biases

Our findings suggest that the HCP cognitions and emotions are filtered through cognitive biases. As described by Cook et al [9], how a clinician perceives telehealth is a key driver of whether telehealth is provided regardless of any empirical evidence supporting telehealth use that may be present.

XSL•FO
RenderX

Decisions regarding whether or not to offer VCs, are shrouded in assumptions. All humans make decisions based on simplified information processing called heuristics, which often leads to inaccurate judgments and systematic assumptions called cognitive biases [139,140]. The role of cognitive bias in technology-related decision-making is emerging [139]. Although there are numerous types of cognitive biases, Oschinsky et al [139] examines the status-quo bias. The status-quo bias explains people's tendency to maintain the "status-quo" or existing behaviors [141]. The potential of status-quo bias to increase resistance to technology use is evident [139,141]. The role of cognitive bias was not explicitly mentioned within the systematic review data but provided an explanation of the varying attitudes across the literature and relationships across psychological factors. However, further research is needed to examine which types of cognitive biases [142,143] may be impacting HCPs' decisions to use VCs and where cognitive biases may be present. One possible example pertains to the cognitive bias called confirmation bias (a tendency to favor information that aligns with personal values or perceptions). Confirmation bias is at play when an HCP has one bad experience with audio on a call, which confirms their belief that VCs are inefficient and provide low-quality care. In contrast, an HCP with a positive attitude toward VCs will minimize the impact of a bad audio experience. A recent review [144] indicates that key strategies to change HCP cognitive biases are reflection and education. However, how to integrate this training in VC contexts specifically warrants further research. Approaches that target cognitive and emotional factors alongside cognitive biases may provide an optimal structure for behavior change.

The "Intention-Behaviour Gap" and Unconscious Competence

Our PAVE model suggests that a range of emotions and cognitions can influence the intention to engage in VCs, but intentions do not always predict actual engagement [145,146]. For example, a previous review [147] suggests that intentions explain between 18% and 23% of the variance in actual behavior engagement. The inconsistent relationship between intention and actual behavior, termed the "intention-behaviour gap," has long been recognized and examined across multiple behaviors [148,149]. Variations to the TPB and testing of moderating effects continue. One such moderator when considering the relationship between intention to engage in VCs and actual engaging in the behavior of conducting VCs, is habit.

Research indicates that when an individual frequently engages in a behavior, the predictive value of intentions, attitudes, and self-efficacy diminishes [150]. This is because frequently engaging in behavior forms a habit and reduces the reliance on cognitive effort to elicit the behavior and it becomes automatic. Evidence within our literature review indicated that, over time, with frequent use, many HCPs no longer consciously thought about how to use VC, because it had become automatic. "Unconscious competence" is cited frequently within knowledge acquisition studies [151] and is consistent with the findings within the current review. Conversely, influenced by attitudes, self-efficacy and other emotional reactions to VCs, the cognitive and practical effort needed to make VCs business as usual can be overridden by habit. If experiences are negative or the use of VCs is infrequent, there is the potential to "fall back into old habits" of using phone or traditional in-person care. Examining information technology use, de Guinea and Markus [152] argue that habit may moderate the relationship between intention and behavior, especially when repeated IT use (behavior) becomes habitual and so the need for the cognitive effort associated with intention diminishes. Research dating back to 2009 has not found conclusive evidence and, as such, there is a need to further examine the role of habit among other psychological factors.

User Categories and Targeting Interventions

Finally, our findings suggest that HCPs may fit into 4 different user categories (Table 1). These include (1) individuals with negative cognitions and emotions and, therefore, do not seriously contemplate engaging in VCs; (2) those who exert cognitive effort and have intentions but for varied reasons do not progress to actually carrying out any VCs; (3) those who intend to use VC and exert cognitive effort to plan VCs, before engaging with VCs. This is a turning point. If the HCP does not regularly engage or has a bad experience, it could produce negative emotions and attitudes and the HCP stops using video. Alternatively, they may progress to the type (4) who use VC regularly and VC becomes automatic or habitual.

Although further examination of these 4 user categories is needed, it may also be prudent to examine them in alignment with the stages of change model (Figure 3) [153]. If future research supports the categorization of these 4 types of alignment with the stages of change model, targeted interventions based on this model could be developed. The stages of change model, traditionally based on addictive behaviors, is aligned with a range of interventions dependent on the stage of change. For instance, different interventions would be needed for someone who is not even contemplating using VC (eg, motivational interviewing exploring concerns and promoting the benefits) compared with someone who has progressed to using it once or twice (increase skills and self-efficacy through practice). One complexity that may counter the alignment with stages of change is that the choice and use of VCs may be context dependent, with providers showing different levels of enthusiasm depending on the clinical requirement.

XSL•FO
RenderX

Figure 3. Use or nonuse categories as they align with The Stages of Change Model.

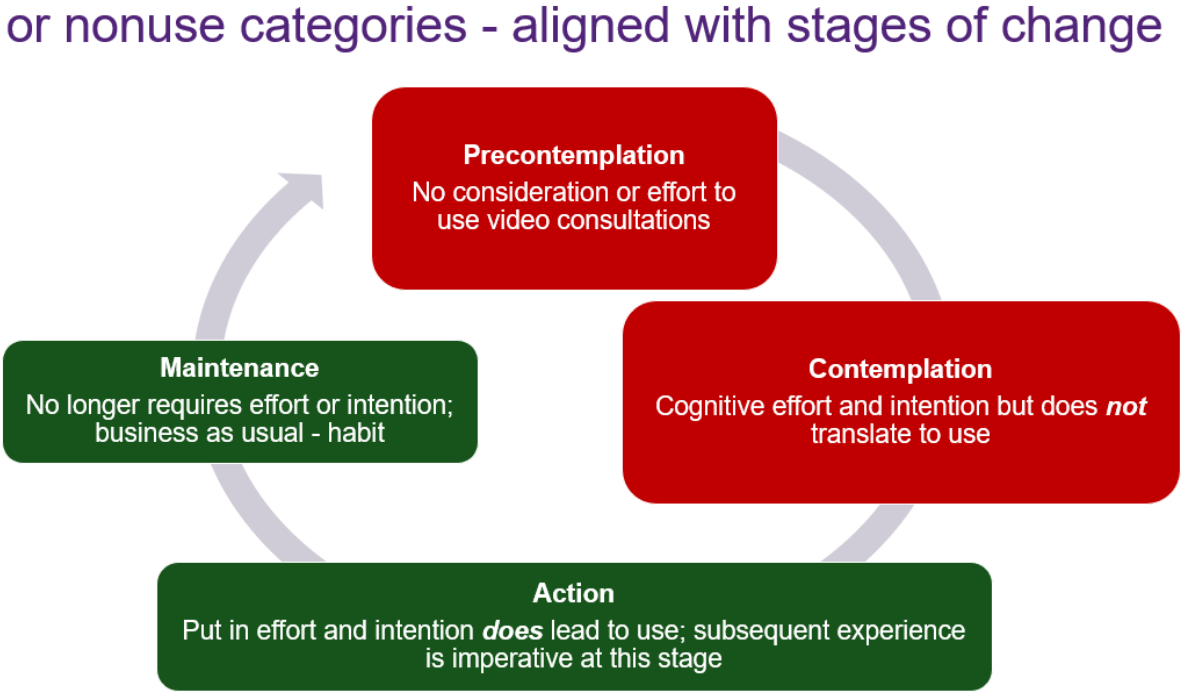

Strengths and Limitations

The authorship team has worked for considerable years within telehealth contexts, including directly with HCPs, which strengthens the data analysis. A limitation is the focus of the literature predominantly on intentions and not actual behavior. As described above, the influence of intention on behavior is not always consistent. The inductive analysis and interpretation of the relationship between psychological factors and the 4 user categories need to be further examined. In addition, individual study quality as detailed in the risk of bias assessment section may affect the validity of findings. It is also a limitation that due to the number of included papers our research team only had the capacity for 1 reviewer to initially code the data. However, a sample of 5 papers was analyzed by all researchers to improve interrater reliability, a second researcher was involved in collating the themes by classification (positive, negative, neutral, ambivalent), a third researcher reviewed overall themes, and regular peer debriefing meetings occurred with all researchers attending.

Heterogeneity in terms of clinical specialties and settings was both a strength in terms of generalizability, but also a limitation in aggregation of data, potentially increasing bias from studies. Profession was not always clearly identified in the journal papers; for instance, some studies simply stated that “health professionals” were interviewed. Further research focusing on individual clinician specialties (eg, do allied health professionals have more motivating attitudes?) could also increase understanding of this topic. However, 5 reviewers and the implementation of rigorous methods (eg, multiple peer meetings) instill confidence in the findings, especially as these factors or themes are widespread due to the shared experience reflected in human behavior. Finally, although the current aim and scope of this research were focused on Australia, it is expected that these results are somewhat generalizable beyond Australia.

Conclusion

This review identified a range of factors that interact to influence the use of VCs. Cognitive and emotional factors that motivate (positive perceptions) or inhibit (negative perceptions) engagement in VCs were found. These include but are not limited to emotional responses (eg, feeling relief or anxiety), varying attitudes toward rapport-building, trust, and patient-centeredness, as well as effectiveness and quality of care. The impact of VC use on an HCP’s professional role and identity was also mentioned throughout the literature analyzed, alongside their perceived ability or self-efficacy. Our PAVE model (Figure 2) highlights the potential cyclical nature and relationships between the cognitive and emotional factors and intention to use or engage in VCs. The PAVE model highlights psychological factors and the relationships between them, which may be important when developing strategies that support clinicians in the use of VCs. Finally, HCPs may fall within 4 key user categories, which can help with targeting solutions when there is low VC uptake. These categories are as follows: (1) negative cognitions and emotions and no VC contemplation; (2) cognitive effort and intentions but do not progress to VC use; (3) planning and engaging with VC; and (4) use VC regularly, automatically or habitually. Further research validating the findings of this review can lead to interventions such as training, education, and reflective practices that address these psychological factors with the aim of delivering care via video when it is clinically appropriate.

XSL•FO
RenderX

Acknowledgments

We would like to acknowledge the work of Nicole Rayner, Outreach Librarian at The University of Queensland, for assistance in developing the search string and conducting database searches for this review. We also acknowledge the work of Roshni Mendis in helping create a visual for the psychological attributes of video consultation engagement (PAVE) model. Generative artificial intelligence was not used in any portion of the manuscript writing. This research was supported by Clinical Excellence Queensland, Queensland Health.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HH, AS, and LC conceptualized this study. MT and HH conducted the searches, and all authors extracted data. LC, MT, and JF conducted the initial coding and analysis of data while HH supervised. HH collated findings and wrote the initial draft with assistance from JF. Review and editing of the manuscript was done by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database search strategies.
[PDF File (Adobe PDF File), 123 KB - jmir_v26i1e54636_app1.pdf]

Multimedia Appendix 2
Quality assessment of included studies.
[XLSX File (Microsoft Excel File), 130 KB - jmir_v26i1e54636_app2.xlsx]

Multimedia Appendix 3
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 90 KB - jmir_v26i1e54636_app3.pdf]

References

1. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
2. Eze ND, Mateus C, Cravo Oliveira Hashiguchi T. Telemedicine in the OECD: an umbrella review of clinical and cost-effectiveness, patient experience and implementation. PLoS One 2020;15(8):e0237585 [FREE Full text] [doi: 10.1371/journal.pone.0237585] [Medline: 32790752]
3. Yellowlees PM, Chorba K, Burke Parish M, Wynn-Jones H, Nafiz N. Telemedicine can make healthcare greener. Telemed J E Health 2010;16(2):229-232. [doi: 10.1089/tmj.2009.0105] [Medline: 20156125]
4. Dullet NW, Geraghty EM, Kaufman T, Kissee JL, King J, Dharmar M, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017;20(4):542-546 [FREE Full text] [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
5. Snoswell CL, Chelberg G, de Guzman KR, Haydon HH, Thomas EE, Caffery LJ, et al. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. J Telemed Telecare 2023;29(9):669-684. [doi: 10.1177/1357633X211022907] [Medline: 34184580]
6. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020;26(5):309-313 [FREE Full text] [doi: 10.1177/1357633X20916567] [Medline: 32196391]
7. Wade VA, Eliott JA, Hiller JE. Clinician acceptance is the key factor for sustainable telehealth services. Qual Health Res 2014;24(5):682-694. [doi: 10.1177/1049732314528809] [Medline: 24685708]
8. Sherrill AM, Wiese CW, Abdullah S, Arriaga RI. Overcoming clinician technophobia: what we learned from our mass exposure to telehealth during the COVID-19 pandemic. J Technol Behav Sci 2022;7(4):547-553 [FREE Full text] [doi: 10.1007/s41347-022-00273-3] [Medline: 36034538]
9. Cook R, Haydon HM, Thomas EE, Ward EC, Ross J, Webb C, et al. Digital divide or digital exclusion? Do allied health professionals' assumptions drive use of telehealth? J Telemed Telecare 2023:1357633X231189846. [doi: 10.1177/1357633X231189846] [Medline: 37543369]

XSL•FO
RenderX

10. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1):bjgpopen20X101020 [FREE Full text] [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
11. Wanderås MR, Abildsnes E, Thygesen E, Martinez SG. Video consultation in general practice: a scoping review on use, experiences, and clinical decisions. BMC Health Serv Res 2023;23(1):316 [FREE Full text] [doi: 10.1186/s12913-023-09309-7] [Medline: 36997997]
12. Rwashana AS, Namatovu HK, Kyanda S, Magumba A. An EHealth adoption framework for developing countries: a systematic review. HIIJ 2021;10(3):1-16. [doi: 10.5121/hiij.2021.10301]
13. Rouidi M, Elouadi AE, Hamdoune A, Choujtani K, Chati A. TAM-UTAUT and the acceptance of remote healthcare technologies by healthcare professionals: a systematic review. Inform Med Unlocked 2022;32:101008. [doi: 10.1016/j.imu.2022.101008]
14. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding clinicians' adoption of mobile health tools: a qualitative review of the most used frameworks. JMIR Mhealth Uhealth 2020;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
15. Nguyen P. Policy brief: telehealth use and availability in VHA outpatient mental health care. Partnered Evidence-based Policy Resource Center. 2023. URL: https://www.peprec.research.va.gov/PEPRECRESEARCH/docs/Policy_Brief_21_Telehealth_Mental_Health.pdf [accessed 2024-10-31]
16. Snoswell CL, Caffery LJ, Taylor ML, Haydon HM, Thomas E, Smith AC. Telehealth and coronavirus: Medicare Benefits Schedule (MBS) activity in Australia. The University of Queensland. 2022. URL: https://coh.centre.uq.edu.au/telehealth-and-coronavirus-medicare-benefits-schedule-mbs-activity-australia [accessed 2024-10-31]
17. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
18. Lincoln Y, Guber E. Naturalistic Inquiry. London: SAGE Publications; 1985.
19. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
20. Prochaska JO, Norcross JC. Stages of change. Psychother Theory Res Pract Train 2001;38(4):443-448. [doi: 10.1037//0033-3204.38.4.443]
21. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med 2010;8:63 [FREE Full text] [doi: 10.1186/1741-7015-8-63] [Medline: 20961442]
22. Welcome to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) website. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). 2024. URL: http://prisma-statement.org/ [accessed 2024-10-31]
23. Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M. Mixed Methods Appraisal Tool (MMAT) version 2018 user guide. National Collaborating Centre for Methods and Tools. 2018. URL: https://www.nccmt.ca/knowledge-repositories/search/232 [accessed 2024-10-31]
24. Allan J, Webster E, Chambers B, Nott S. "This is streets ahead of what we used to do": staff perceptions of virtual clinical pharmacy services in rural and remote Australian hospitals. BMC Health Serv Res 2021;21(1):1306 [FREE Full text] [doi: 10.1186/s12913-021-07328-w] [Medline: 34863164]
25. Aung E, Pasanen L, LeGautier R, McLachlan S, Collins A, Philip J. The role of telehealth in oncology care: a qualitative exploration of patient and clinician perspectives. Eur J Cancer Care (Engl) 2022;31(2):e13563. [doi: 10.1111/ecc.13563] [Medline: 35150180]
26. Ayres L, Pelkowitz L, Simon P, Thompson SC. Necessity as the catalyst of change: exploring client and provider perspectives of accelerated implementation of telehealth by a regional Australian community service organisation during COVID-19 restrictions. Int J Environ Res Public Health 2021;18(21):11433 [FREE Full text] [doi: 10.3390/ijerph182111433] [Medline: 34769949]
27. Bagot K, Moloczij N, Arthurson L, Hair C, Hancock S, Bladin CF, et al. Nurses' role in implementing and sustaining acute telemedicine: a mixed-methods, pre-post design using an extended technology acceptance model. J Nurs Scholarsh 2020;52(1):34-46. [doi: 10.1111/jnu.12509] [Medline: 31508882]
28. Bagot KL, Moloczij N, Barclay-Moss K, Vu M, Bladin CF, Cadilhac DA. Sustainable implementation of innovative, technology-based health care practices: a qualitative case study from stroke telemedicine. J Telemed Telecare 2020;26(1-2):79-91. [doi: 10.1177/1357633X18792380] [Medline: 30193566]
29. Baker FA, Tamplin J. Music therapy service provision via telehealth in response to COVID-19 restrictions: a survey of Australian practitioners and consumers. J Music Ther 2021;32(1):1-24 [FREE Full text]
30. Banbury A, Smith AC, Mehrotra A, Page M, Caffery LJ. A comparison study between metropolitan and rural hospital-based telehealth activity to inform adoption and expansion. J Telemed Telecare 2023;29(7):540-551. [doi: 10.1177/1357633X21998201] [Medline: 33765879]

XSL•FO
RenderX

31. Bennell KL, Lawford BJ, Metcalf B, Mackenzie D, Russell T, van den Berg M, et al. Physiotherapists and patients report positive experiences overall with telehealth during the COVID-19 pandemic: a mixed-methods study. J Physiother 2021;67(3):201-209 [FREE Full text] [doi: 10.1016/j.jphys.2021.06.009] [Medline: 34147399]
32. Bradford NK, Young J, Armfield NR, Herbert A, Smith AC. Home telehealth and paediatric palliative care: clinician perceptions of what is stopping us? BMC Palliat Care 2014;13:29 [FREE Full text] [doi: 10.1186/1472-684X-13-29] [Medline: 24963287]
33. Brown AD, Kelso W, Velakoulis D, Farrand S, Stolwyk RJ. Understanding clinician's experiences with implementation of a younger onset dementia telehealth service. J Geriatr Psychiatry Neurol 2023;36(4):295-308. [doi: 10.1177/08919887221141653] [Medline: 36433702]
34. Brunelli VN, Fox JA, Langbecker DH. Disparity in cancer survivorship care: a cross-sectional study of telehealth use among cancer nurses in Australia. Collegian 2021;28(5):498-505. [doi: 10.1016/j.colegn.2021.01.002]
35. Butt Z, Kirsten L, Beatty L, Kelly B, Dhillon H, Shaw JM. Barriers and enablers to implementing telehealth consultations in psycho-oncology. Psychooncology 2022;31(8):1365-1373 [FREE Full text] [doi: 10.1002/pon.5939] [Medline: 35460322]
36. Caffery LJ, Taylor M, North JB, Smith AC. Tele-orthopaedics: a snapshot of services in Australia. J Telemed Telecare 2017;23(10):835-841. [doi: 10.1177/1357633X17732800] [Medline: 28950754]
37. Carlisle K, Warren R. A qualitative case study of telehealth for in-home monitoring to support the management of type 2 diabetes. J Telemed Telecare 2013;19(7):372-375. [doi: 10.1177/1357633X13506512] [Medline: 24218347]
38. Cartledge S, Rawstorn JC, Tran M, Ryan P, Howden EJ, Jackson A. Telehealth is here to stay but not without challenges: a consultation of cardiac rehabilitation clinicians during COVID-19 in Victoria, Australia. Eur J Cardiovasc Nurs 2022;21(6):548-558 [FREE Full text] [doi: 10.1093/eurjcn/zvab118] [Medline: 34935940]
39. Chapman JE, Ponsford J, Bagot KL, Cadilhac DA, Gardner B, Stolwyk RJ. The use of videoconferencing in clinical neuropsychology practice: a mixed methods evaluation of neuropsychologists' experiences and views. Aust Psychol 2021;55(6):618-633. [doi: 10.1111/ap.12471]
40. Chatterton ML, Marangu E, Clancy EM, Mackay M, Gu E, Moylan S, et al. Telehealth service delivery in an Australian regional mental health service during COVID-19: a mixed methods analysis. Int J Ment Health Syst 2022;16(1):43 [FREE Full text] [doi: 10.1186/s13033-022-00553-8] [Medline: 35986332]
41. Clay-Williams R, Baysari M, Taylor N, Zalitis D, Georgiou A, Robinson M, et al. Service provider perceptions of transitioning from audio to video capability in a telehealth system: a qualitative evaluation. BMC Health Serv Res 2017;17(1):558 [FREE Full text] [doi: 10.1186/s12913-017-2514-7] [Medline: 28806903]
42. Collier A, Morgan DD, Swetenham K, To THM, Currow DC, Tieman JJ. Implementation of a pilot telehealth programme in community palliative care: a qualitative study of clinicians' perspectives. Palliat Med 2016;30(4):409-417. [doi: 10.1177/0269216315600113] [Medline: 26290500]
43. Cottrell M, Burns CL, Jones A, Rahmann A, Young A, Sam S, et al. Sustaining allied health telehealth services beyond the rapid response to COVID-19: learning from patient and staff experiences at a large quaternary hospital. J Telemed Telecare 2021;27(10):615-624 [FREE Full text] [doi: 10.1177/1357633X211041517] [Medline: 34726993]
44. Cottrell MA, Hill AJ, O'Leary SP, Raymer ME, Russell TG. Clinicians' perspectives of a novel home-based multidisciplinary telehealth service for patients with chronic spinal pain. Int J Telerehabil 2018;10(2):81-88 [FREE Full text] [doi: 10.5195/ijt.2018.6249] [Medline: 30588279]
45. de Guzman KR, Snoswell CL, Giles CM, Smith AC, Haydon HH. GP perceptions of telehealth services in Australia: a qualitative study. BJGP Open 2022;6(1):BJGPO.2021.0182 [FREE Full text] [doi: 10.3399/BJGPO.2021.0182] [Medline: 34819294]
46. Dham P, Gupta N, Alexander J, Black W, Rajji T, Skinner E. Community based telepsychiatry service for older adults residing in a rural and remote region- utilization pattern and satisfaction among stakeholders. BMC Psychiatry 2018;18(1):316 [FREE Full text] [doi: 10.1186/s12888-018-1896-3] [Medline: 30261845]
47. Eastman P, Dowd A, White J, Carter J, Ely M. Telehealth: rapid adoption in community palliative care due to COVID-19: patient and professional evaluation. BMJ Support Palliat Care 2021:bmjspcare-2021-002987. [doi: 10.1136/bmjspcare-2021-002987] [Medline: 34162584]
48. Edirippulige S, Reyno J, Armfield NR, Bambling M, Lloyd O, McNevin E. Availability, spatial accessibility, utilisation and the role of telehealth for multi-disciplinary paediatric cerebral palsy services in Queensland. J Telemed Telecare 2016;22(7):391-396. [doi: 10.1177/1357633X15610720] [Medline: 26519377]
49. Erickson S, Bridgman K, Furlong L, Stark H. Speech-language pathologist perspectives of the implementation of telepractice-delivered stuttering treatment for school-age children. Lang Speech Hear Serv Sch 2022;53(1):30-43. [doi: 10.1044/2021_LSHSS-20-00167] [Medline: 34752153]
50. Ervin K, Weller-Newton J, Phillips J. Primary healthcare clinicians' positive perceptions of the implementation of telehealth during the COVID-19 pandemic using normalisation process theory. Aust J Prim Health 2021;27:158-162. [doi: 10.1071/PY20182] [Medline: 33653506]
51. Esther C, Natalie O, Diana B, Marie Antoinette H, Suzi D, Marcia W, et al. Telehealth in a paediatric developmental metropolitan assessment clinic: perspectives and experiences of families and clinicians. Health Expect 2022;25(5):2557-2569 [FREE Full text] [doi: 10.1111/hex.13582] [Medline: 35978461]

XSL•FO
RenderX

52. Fogarty A, Jones A, Seymour M, Savopoulos P, Evans K, O'Brien J, et al. The parenting skill development and education service: telehealth support for families at risk of child maltreatment during the COVID‐19 pandemic. Child Fam Soc Work 2021;27(3):392-404. [doi: 10.1111/cfs.12890]
53. Fogarty A, Savopoulos P, Seymour M, Cox A, Williams K, Petrie S, et al. Providing therapeutic services to women and children who have experienced intimate partner violence during the COVID-19 pandemic: challenges and learnings. Child Abuse Negl 2022;130(Pt 1):105365 [FREE Full text] [doi: 10.1016/j.chiabu.2021.105365] [Medline: 34686360]
54. Fomiatti R, Shaw F, Fraser S. 'It's a different way to do medicine': exploring the affordances of telehealth for hepatitis C healthcare. Int J Drug Policy 2022;110:103875. [doi: 10.1016/j.drugpo.2022.103875] [Medline: 36257085]
55. Gelber H. The experience in Victoria with telepsychiatry for the child and adolescent mental health service. J Telemed Telecare 2001;7 Suppl 2:32-34. [doi: 10.1258/1357633011937065] [Medline: 11747653]
56. Green T, Hartley N, Gillespie N. Service provider's experiences of service separation. J Serv Res 2016;19(4):477-494. [doi: 10.1177/1094670516666674]
57. Haines KJ, Sawyer A, McKinnon C, Donovan A, Michael C, Cimoli C, et al. Barriers and enablers to telehealth use by physiotherapists during the COVID-19 pandemic. Physiotherapy 2023;118:12-19 [FREE Full text] [doi: 10.1016/j.physio.2022.09.003] [Medline: 36308980]
58. Halcomb EJ, Ashley C, Dennis S, McInnes S, Morgan M, Zwar N, et al. Telehealth use in Australian primary healthcare during COVID-19: a cross-sectional descriptive survey. BMJ Open 2023;13(1):e065478 [FREE Full text] [doi: 10.1136/bmjopen-2022-065478] [Medline: 36604135]
59. Haydon HM, Snoswell CL, Thomas EE, Broadbent A, Caffery LJ, Brydon J, et al. Enhancing a community palliative care service with telehealth leads to efficiency gains and improves job satisfaction. J Telemed Telecare 2021;27(10):625-630. [doi: 10.1177/1357633X211048952] [Medline: 34726990]
60. Henry A, Yang J, Grattan S, Roberts L, Lainchbury A, Shanthosh J, et al. Effects of the COVID-19 pandemic and telehealth on antenatal screening and services, including for mental health and domestic violence: an australian mixed-methods study. Front Glob Womens Health 2022;3:819953 [FREE Full text] [doi: 10.3389/fgwh.2022.819953] [Medline: 35814835]
61. Hines M, Lincoln M, Ramsden R, Martinovich J, Fairweather C. Speech pathologists' perspectives on transitioning to telepractice: what factors promote acceptance? J Telemed Telecare 2015;21(8):469-473. [doi: 10.1177/1357633X15604555] [Medline: 26377120]
62. Hinman RS, Nelligan RK, Bennell KL, Delany C. "Sounds a bit crazy, but it was almost more personal:" a qualitative study of patient and clinician experiences of physical therapist-prescribed exercise For knee osteoarthritis via skype. Arthritis Care Res (Hoboken) 2017;69(12):1834-1844. [doi: 10.1002/acr.23218] [Medline: 28217864]
63. James S, Ashley C, Williams A, Desborough J, Mcinnes S, Calma K, et al. Experiences of Australian primary healthcare nurses in using telehealth during COVID-19: a qualitative study. BMJ Open 2021;11(8):e049095 [FREE Full text] [doi: 10.1136/bmjopen-2021-049095] [Medline: 34362804]
64. Jhaveri D, Emeto TI, Alele FO, Strom A, Benham H. Use of telemedicine for rheumatology practice in queensland, Australia: experiences before and during the COVID-19 pandemic. Intern Med J 2022;52(10):1685-1690. [doi: 10.1111/imj.15706] [Medline: 35112769]
65. Jhaveri D, Larkins S, Kelly J, Sabesan S. Remote chemotherapy supervision model for rural cancer care: perspectives of health professionals. Eur J Cancer Care (Engl) 2016;25(1):93-98. [doi: 10.1111/ecc.12309] [Medline: 25871852]
66. Keenan AJ, Tsourtos G, Tieman J. Promise and peril-defining ethical telehealth practice from the clinician and patient perspective: a qualitative study. Digit Health 2022;8:20552076211070394 [FREE Full text] [doi: 10.1177/20552076211070394] [Medline: 35024158]
67. Knott V, Habota T, Mallan K. Attitudes of Australian psychologists towards the delivery of therapy via video conferencing technology. Aust Psychol 2021;55(6):606-617. [doi: 10.1111/ap.12464]
68. Kocanda L, Fisher K, Brown LJ, May J, Rollo ME, Collins CE, et al. Informing telehealth service delivery for cardiovascular disease management: exploring the perceptions of rural health professionals. Aust Health Rev 2021;45(2):241-246. [doi: 10.1071/AH19231] [Medline: 33715764]
69. Kozica-Olenski SL, Soldatos G, Marlow L, Cooray SD, Boyle JA. Exploring the acceptability and experience of receiving diabetes and pregnancy care via telehealth during the COVID-19 pandemic: a qualitative study. BMC Pregnancy Childbirth 2022;22(1):932 [FREE Full text] [doi: 10.1186/s12884-022-05175-z] [Medline: 36514010]
70. Kozica-Olenski SL, Garth B, Boyle JA, Vincent AJ. Menopause care delivery in the time of COVID-19: evaluating the acceptability of telehealth services for women with early and usual age menopause. Climacteric 2023;26(1):34-46. [doi: 10.1080/13697137.2022.2127351] [Medline: 36279887]
71. Lakeman R, Hurley J, Campbell K, Hererra C, Leggett A, Tranter R, et al. High fidelity dialectical behaviour therapy online: learning from experienced practitioners. Int J Ment Health Nurs 2022;31(6):1405-1416 [FREE Full text] [doi: 10.1111/inm.13039] [Medline: 35789190]
72. Lawford BJ, Bennell KL, Kasza J, Hinman RS. Physical therapists' perceptions of telephone- and internet video-mediated service models for exercise management of people with osteoarthritis. Arthritis Care Res (Hoboken) 2018;70(3):398-408. [doi: 10.1002/acr.23260] [Medline: 28437566]

XSL•FO
RenderX

73. Lawson DW, Stolwyk RJ, Ponsford JL, Baker KS, Tran J, Wong D. Acceptability of telehealth in post-stroke memory rehabilitation: a qualitative analysis. Neuropsychol Rehabil 2022;32(1):1-21. [doi: 10.1080/09602011.2020.1792318] [Medline: 32677539]
74. Luscombe GM, Hawthorn J, Wu A, Green B, Munro A. 'Empowering clinicians in smaller sites': a qualitative study of clinician's experiences with a rural virtual paediatric feeding clinic. Aust J Rural Health 2021;29(5):742-752. [doi: 10.1111/ajr.12781] [Medline: 34490941]
75. Mackenzie L, Noble N, Proietto A, Jones J, Norton G, Palazzi K. Acceptability and feasibility of telehealth outpatient video-link consultations: a national cross-sectional survey of surgeons prior to the COVID-19 pandemic. Aust J Rural Health 2023;31(2):244-255 [FREE Full text] [doi: 10.1111/ajr.12940] [Medline: 36326168]
76. Malliaras P, Merolli M, Williams CM, Caneiro JP, Haines T, Barton C. 'It's not hands-on therapy, so it's very limited': telehealth use and views among allied health clinicians during the coronavirus pandemic. Musculoskelet Sci Pract 2021;52:102340 [FREE Full text] [doi: 10.1016/j.msksp.2021.102340] [Medline: 33571900]
77. Martin R, Mandrusiak A, Russell T, Forbes R. New-graduate physiotherapists' training needs and readiness for telehealth. Physiother Theory Pract 2022;38(13):2788-2797. [doi: 10.1080/09593985.2021.1955423] [Medline: 34282699]
78. McKeon G, Fitzgerald C, Furzer B, Rosenbaum S, Stanton R, Lederman O, et al. A qualitative exploration of the experience and attitudes of exercise professionals using telehealth for people with mental illness. J Ment Health Train Educ Pract 2023;18(1):14-29. [doi: 10.1108/jmhtep-07-2021-0084]
79. Moffatt JJ, Eley DS. Barriers to the up-take of telemedicine in Australia--a view from providers. Rural Remote Health 2011;11(2):1581 [FREE Full text] [Medline: 21385004]
80. Mozer R, Bradford NK, Caffery LJ, Smith AC. Identifying perceived barriers to videoconferencing by rehabilitation medicine providers. J Telemed Telecare 2015;21(8):479-484. [doi: 10.1177/1357633X15607136] [Medline: 26556061]
81. Nicholas J, Bell IH, Thompson A, Valentine L, Simsir P, Sheppard H, et al. Implementation lessons from the transition to telehealth during COVID-19: a survey of clinicians and young people from youth mental health services. Psychiatry Res 2021;299:113848 [FREE Full text] [doi: 10.1016/j.psychres.2021.113848] [Medline: 33725578]
82. O'Sullivan B, Rann H, McGrail M. Outreach specialists' use of video consultations in rural Victoria: a cross-sectional survey. Rural Remote Health 2019;19(1):4544 [FREE Full text] [doi: 10.22605/RRH4544] [Medline: 30922059]
83. Owen N. Feasibility and acceptability of using telehealth for early intervention parent counselling. Adv Mental Health 2019;18(1):39-49. [doi: 10.1080/18387357.2019.1679026]
84. Ownsworth T, Theodoros D, Cahill L, Vaezipour A, Quinn R, Kendall M, et al. Perceived usability and acceptability of videoconferencing for delivering community-based rehabilitation to individuals with acquired brain injury: a qualitative investigation. J Int Neuropsychol Soc 2020;26(1):47-57. [doi: 10.1017/S135561771900078X] [Medline: 31983367]
85. Parker J, Robinson J, MugicaCox B, Foy A, Kepu K, HarrisRoxas B. How COVID-19 shaped new models of care for a child and family health nursing service. Aust J Child Fam Health Nurs 2022;19(1):6-14. [doi: 10.33235/ajcfhn.19.1.6-14]
86. Paul SS, Hubbard A, Johnson J, Dennis SM. Transition to a virtual model of physiotherapy and exercise physiology in response to COVID-19 for people in a rural Australia: is it a viable solution to increase access to allied health for rural populations? PLoS One 2023;18(1):e0280876 [FREE Full text] [doi: 10.1371/journal.pone.0280876] [Medline: 36662817]
87. Pitt R, Hill AJ, Theodoros D, Russell T. “I definitely think it’s a feasible and worthwhile option”: perspectives of speech-language pathologists providing online aphasia group therapy. Aphasiology 2018;32(9):1031-1053. [doi: 10.1080/02687038.2018.1482403]
88. Raatz M, Ward EC, Marshall J, Burns CL. Evaluating the use of telepractice for bottle-feeding assessments. Children (Basel) 2021;8(11):989 [FREE Full text] [doi: 10.3390/children8110989] [Medline: 34828701]
89. Raatz MK, Ward EC, Marshall J. Telepractice for the delivery of pediatric feeding services: a survey of practice investigating clinician perceptions and current service models in Australia. Dysphagia 2020;35(2):378-388. [doi: 10.1007/s00455-019-10042-9] [Medline: 31363846]
90. Randall L, Raisin C, Waters F, Williams C, Shymko G, Davis D. Implementing telepsychiatry in an early psychosis service during COVID-19: experiences of young people and clinicians and changes in service utilization. Early Interv Psychiatry 2023;17(5):470-477 [FREE Full text] [doi: 10.1111/eip.13342] [Medline: 35943177]
91. Ross MH, Nelson M, Parravicini V, Weight M, Tyrrell R, Hartley N, et al. Staff perspectives on the key elements to successful rapid uptake of telerehabilitation in medium-sized public hospital physiotherapy departments. Physiother Res Int 2023;28(3):e1991. [doi: 10.1002/pri.1991] [Medline: 36540908]
92. Ryan M, Ward EC, Burns CL, Carrington C, Cuff K, Mackinnon M, et al. An evaluation of telephone versus videoconference consults for pre-treatment medication history taking by cancer pharmacists. J Telemed Telecare 2022;28(10):750-756. [doi: 10.1177/1357633X221122140] [Medline: 36346932]
93. Shannon MM, Callum SM, Callisaya ML. Uncovering healthcare staff attitudes to the rapid deployment of telehealth in Victoria, 2020-2021: a 12-month telehealth experience. Intern Med J 2023;53(6):1018-1026. [doi: 10.1111/imj.15750] [Medline: 35289486]
94. Shulver W, Killington M, Crotty M. 'Massive potential' or 'safety risk'? health worker views on telehealth in the care of older people and implications for successful normalization. BMC Med Inform Decis Mak 2016;16(1):131 [FREE Full text] [doi: 10.1186/s12911-016-0373-5] [Medline: 27733195]

XSL•FO
RenderX

95. Starling J, Foley S. From pilot to permanent service: ten years of paediatric telepsychiatry. J Telemed Telecare 2006;12(3_suppl):80-82. [doi: 10.1258/135763306779380147]
96. Starling J, Rosina R, Nunn K, Dossetor D. Child and adolescent telepsychiatry in New South Wales: moving beyond clinical consultation. Australas Psychiatry 2003;11(1_suppl):S117-S121. [doi: 10.1046/j.1038-5282.2003.02016.x]
97. Sunner C, Giles MT, Kable A, Foureur M. Experiences of nurses working in RACFs and EDs utilising visual telehealth consultation to assess the need for RACF resident transfer to ED: a qualitative descriptive study. J Clin Nurs 2023;32(15-16):4694-4709. [doi: 10.1111/jocn.16529] [Medline: 36081333]
98. Sutarsa IN, Kasim R, Steward B, Bain-Donohue S, Slimings C, Hall Dykgraaf S, et al. Implications of telehealth services for healthcare delivery and access in rural and remote communities: perceptions of patients and general practitioners. Aust J Prim Health 2022;28(6):522-528. [doi: 10.1071/PY21162] [Medline: 35918783]
99. Sutherland R, Hodge A, Chan E, Silove N. Barriers and facilitators: clinicians' opinions and experiences of telehealth before and after their use of a telehealth platform for child language assessment. Int J Lang Commun Disord 2021;56(6):1263-1277. [doi: 10.1111/1460-6984.12666] [Medline: 34455670]
100. Sutherland R, Hodge A, Chan E, Silove N. Clinician experiences using standardised language assessments via telehealth. Int J Speech Lang Pathol 2021;23(6):569-578. [doi: 10.1080/17549507.2021.1903079] [Medline: 34000937]
101. Swales M, Theodoros D, Hill AJ, Russell T. Speech-language pathologists' perceptions of the use of telepractice in the delivery of services to people with Parkinson's disease: a national pilot survey. Int J Speech Lang Pathol 2020;22(4):387-398. [doi: 10.1080/17549507.2019.1650110] [Medline: 31416341]
102. Taylor A, Caffery LJ, Gesesew HA, King A, Bassal A, Ford K, et al. How Australian health care services adapted to telehealth during the COVID-19 pandemic: a survey of telehealth professionals. Front Public Health 2021;9:648009 [FREE Full text] [doi: 10.3389/fpubh.2021.648009] [Medline: 33718325]
103. Thomas EE, de Camargo Catapan S, Haydon HM, Barras M, Snoswell C. Exploring factors of uneven use of telehealth among outpatient pharmacy clinics during COVID-19: a multi-method study. Res Social Adm Pharm 2022;18(9):3602-3611 [FREE Full text] [doi: 10.1016/j.sapharm.2022.02.003] [Medline: 35183460]
104. Thomas EE, Taylor ML, Ward EC, Hwang R, Cook R, Ross J, et al. Beyond forced telehealth adoption: a framework to sustain telehealth among allied health services. J Telemed Telecare 2024;30(3):559-569 [FREE Full text] [doi: 10.1177/1357633X221074499] [Medline: 35130099]
105. Thrum M, Driscoll C, Keogh T. Investigating the satisfactionof clinicians and clientsin a teleaudiology trial. J Hear Sci 2018;8(4):34-47. [doi: 10.17430/1003211]
106. Venville A, O'Connor S, Roeschlein H, Ennals P, McLoughlan G, Thomas N. Mental health service user and worker experiences of psychosocial support via telehealth through the COVID-19 pandemic: qualitative study. JMIR Ment Health 2021;8(8):e29671 [FREE Full text] [doi: 10.2196/29671] [Medline: 34182461]
107. Wade VA, Eliott JA, Hiller JE. A qualitative study of ethical, medico-legal and clinical governance matters in Australian telehealth services. J Telemed Telecare 2012;18(2):109-114. [doi: 10.1258/jtt.2011.110808] [Medline: 22267306]
108. Wade VA, Taylor AD, Kidd MR, Carati C. Transitioning a home telehealth project into a sustainable, large-scale service: a qualitative study. BMC Health Serv Res 2016;16:183 [FREE Full text] [doi: 10.1186/s12913-016-1436-0] [Medline: 27185041]
109. Ward EC, Burns CL, Theodoros DG, Russell TG. Evaluation of a clinical service model for dysphagia assessment via telerehabilitation. Int J Telemed Appl 2013;2013:918526 [FREE Full text] [doi: 10.1155/2013/918526] [Medline: 24381589]
110. Westbrook JI, Coiera EW, Brear M, Stapleton S, Rob MI, Murphy M, et al. Impact of an ultrabroadband emergency department telemedicine system on the care of acutely ill patients and clinicians' work. Med J Aust 2008;188(12):704-708. [doi: 10.5694/j.1326-5377.2008.tb01850.x] [Medline: 18558892]
111. White J, Byles J, Walley T. The qualitative experience of telehealth access and clinical encounters in Australian healthcare during COVID-19: implications for policy. Health Res Policy Syst 2022;20(1):9 [FREE Full text] [doi: 10.1186/s12961-021-00812-z] [Medline: 35033107]
112. Zilliacus E, Meiser B, Lobb E, Barlow-Stewart K, Tucker K. A balancing act--telehealth cancer genetics and practitioners' experiences of a triadic consultation. J Genet Couns 2009;18(6):598-605. [doi: 10.1007/s10897-009-9247-7] [Medline: 19798555]
113. Liu Y, Fu Q, Fu X. The interaction between cognition and emotion. Chin Sci Bull 2009;54(22):4102-4116. [doi: 10.1007/s11434-009-0632-2]
114. Storbeck J, Clore GL. On the interdependence of cognition and emotion. Cogn Emot 2007;21(6):1212-1237 [FREE Full text] [doi: 10.1080/02699930701438020] [Medline: 18458789]
115. Haines KJ, Sawyer A, McKinnon C, Donovan A, Michael C, Cimoli C, et al. Barriers and enablers to telehealth use by physiotherapists during the COVID-19 pandemic. Physiotherapy 2023;118:12-19 [FREE Full text] [doi: 10.1016/j.physio.2022.09.003] [Medline: 36308980]
116. Randall L, Raisin C, Waters F, Williams C, Shymko G, Davis D. Implementing telepsychiatry in an early psychosis service during COVID-19: experiences of young people and clinicians and changes in service utilization. Early Interv Psychiatry 2023;17(5):470-477 [FREE Full text] [doi: 10.1111/eip.13342] [Medline: 35943177]

XSL•FO
RenderX

117. Philip J, Wawryk O, Pasanen L, Wong A, Schwetlik S, Collins A. Telehealth in outpatient delivery of palliative care: a prospective survey evaluation by patients and clinicians. Intern Med J 2022;52(7):1144-1153 [FREE Full text] [doi: 10.1111/imj.15721] [Medline: 35189018]
118. Fomiatti R, Shaw F, Fraser S. 'It's a different way to do medicine': exploring the affordances of telehealth for hepatitis C healthcare. Int J Drug Policy 2022;110:103875. [doi: 10.1016/j.drugpo.2022.103875] [Medline: 36257085]
119. Zulkosky K. Self-efficacy: a concept analysis. Nurs Forum 2009;44(2):93-102. [doi: 10.1111/j.1744-6198.2009.00132.x]
120. Korteling JEH, Paradies GL, Sassen-van Meer JP. Cognitive bias and how to improve sustainable decision making. Front Psychol 2023;14:1129835 [FREE Full text] [doi: 10.3389/fpsyg.2023.1129835] [Medline: 37026083]
121. Whelehan DF, Conlon KC, Ridgway PF. Medicine and heuristics: cognitive biases and medical decision-making. Ir J Med Sci 2020;189(4):1477-1484. [doi: 10.1007/s11845-020-02235-1] [Medline: 32409947]
122. Connolly SL, Miller CJ, Lindsay JA, Bauer MS. A systematic review of providers' attitudes toward telemental health via videoconferencing. Clin Psychol (New York) 2020;27(2):e12311 [FREE Full text] [doi: 10.1111/cpsp.12311] [Medline: 35966216]
123. Hu PJ, Chau PY, Sheng ORL, Tam KY. Examining the technology acceptance model using physician acceptance of telemedicine technology. J Manag Inf Syst 2015;16(2):91-112. [doi: 10.1080/07421222.1999.11518247]
124. Davis FD, Davis F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319. [doi: 10.2307/249008]
125. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 1989;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
126. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2015;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
127. Chow M, Herold DK, Choo T, Chan K. Extending the technology acceptance model to explore the intention to use second life for enhancing healthcare education. Comput Educ 2012;59(4):1136-1144. [doi: 10.1016/j.compedu.2012.05.011]
128. Taylor S, Todd PA. Understanding information technology usage: a test of competing models. Inf Syst Res 1995;6(2):144-176. [doi: 10.1287/isre.6.2.144]
129. van Houwelingen T, Ettema RGA, Bleijenberg N, van Os-Medendorp H, Kort HSM, Ten Cate O. Educational intervention to increase nurses' knowledge, self-efficacy and usage of telehealth: a multi-setting pretest-posttest study. Nurse Educ Pract 2021;51:102924 [FREE Full text] [doi: 10.1016/j.nepr.2020.102924] [Medline: 33583723]
130. Haddock G, Maio GR. Chapter Two - Inter-individual differences in attitude content: Cognition, affect, and attitudes, in Advances in Experimental Social Psychology. Cambridge, MA: Academic Press; 2019:53-102.
131. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. Adv Consum Res 1984:673-675.
132. Maio GR, Haddock G, Verplanken B. Psychology of Attitudes and Attitude Change. Thousand Oaks, CA: SAGE; 2019.
133. Allen CT, Machleit KA, Kleine SS, Notani AS. A place for emotion in attitude models. J Bus Res 2005;58(4):494-499. [doi: 10.1016/s0148-2963(03)00139-5]
134. Beaudry A, Pinsonneault A. The other side of acceptance: studying the direct and indirect effects of emotions on information technology use. MIS Q 2010;34(4):689-710. [doi: 10.2307/25750701]
135. Saadé RG, Kira D. The emotional state of technology acceptance. IISIT 2006;3:529-539. [doi: 10.28945/913]
136. Sherrill AM, Wiese CW, Abdullah S, Arriaga RI. Overcoming clinician technophobia: what we learned from our mass exposure to telehealth during the COVID-19 pandemic. J Technol Behav Sci 2022;7(4):547-553 [FREE Full text] [doi: 10.1007/s41347-022-00273-3] [Medline: 36034538]
137. Bhattacherjee A, Hikmet N. Physicians' resistance toward healthcare information technology: a theoretical model and empirical test. Eur J Inf Syst 2017;16(6):725-737. [doi: 10.1057/palgrave.ejis.3000717]
138. Hak F, Guimarães T, Santos M. Towards effective clinical decision support systems: a systematic review. PLoS One 2022;17(8):e0272846 [FREE Full text] [doi: 10.1371/journal.pone.0272846] [Medline: 35969526]
139. Oschinsky FM, Stelter A, Niehaves B. Cognitive biases in the digital age – how resolving the status quo bias enables public-sector employees to overcome restraint. Gov Inf Q 2021;38(4):101611. [doi: 10.1016/j.giq.2021.101611]
140. Berthet V. The impact of cognitive biases on professionals' decision-making: a review of four occupational areas. Front Psychol 2022;12:802439 [FREE Full text] [doi: 10.3389/fpsyg.2021.802439] [Medline: 35058862]
141. Kim H, Kankanhalli A. Investigating user resistance to information systems implementation: a status quo bias perspective. MIS Q 2009;33(3):567. [doi: 10.2307/20650309]
142. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
143. Korteling JEH, Paradies GL, Sassen-van Meer JP. Cognitive bias and how to improve sustainable decision making. Front Psychol 2023;14:1129835 [FREE Full text] [doi: 10.3389/fpsyg.2023.1129835] [Medline: 37026083]
144. Thompson J, Bujalka H, McKeever S, Lipscomb A, Moore S, Hill N, et al. Educational strategies in the health professions to mitigate cognitive and implicit bias impact on decision making: a scoping review. BMC Med Educ 2023;23(1):455 [FREE Full text] [doi: 10.1186/s12909-023-04371-5] [Medline: 37340395]
145. Faries MD. Why we don't "just do it": understanding the intention-behavior gap in lifestyle medicine. Am J Lifestyle Med 2016;10(5):322-329 [FREE Full text] [doi: 10.1177/1559827616638017] [Medline: 30202289]

XSL•FO
RenderX

146. Conner M, Norman P. Understanding the intention-behavior gap: the role of intention strength. Front Psychol 2022;13:923464 [FREE Full text] [doi: 10.3389/fpsyg.2022.923464] [Medline: 35992469]
147. Armitage CJ, Conner M. Efficacy of the theory of planned behaviour: a meta-analytic review. Br J Soc Psychol 2001;40(Pt 4):471-499. [doi: 10.1348/014466601164939] [Medline: 11795063]
148. Sheppard BH, Hartwick J, Warshaw PR. The theory of reasoned action: a meta-analysis of past research with recommendations for modifications and future research. J Consum Res 1988;15(3):325. [doi: 10.1086/209170]
149. Bhattacherjee A, Sanford C. The intention–behaviour gap in technology usage: the moderating role of attitude strength. Behav Inf Technol 2009;28(4):389-401. [doi: 10.1080/01449290802121230]
150. Sheeran P, Maki A, Montanaro E, Avishai-Yitshak A, Bryan A, Klein WMP, et al. The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behavior: a meta-analysis. Health Psychol 2016;35(11):1178-1188. [doi: 10.1037/hea0000387] [Medline: 27280365]
151. Franz H, Kaletka C, Pelka B, Sarcina R. Growing Experience: From Unconscious Incompetence to Unconscious Competence, in Building Leadership in Project and Network Management: A Facilitator's Toolset. Cham, Switzerland: Springer; 2018:209-226.
152. de Guinea AO, Markus ML. Why break the habit of a lifetime? rethinking the roles of intention, habit, and emotion in continuing information technology use. MIS Q 2009;33(3):433-444. [doi: 10.2307/20650303]
153. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]

Abbreviations

HCP: health care professional
PAVE: psychological attributes of video consultation engagement
PICO: population, intervention, context, outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TAM: technology acceptance model
TPB: theory of planned behavior
VC: video consultation

Edited by T de Azevedo Cardoso; submitted 17.11.23; peer-reviewed by M Kolotylo-Kulkarni, K De Santis; comments to author 12.02.24; revised version received 01.04.24; accepted 30.09.24; published 11.12.24.

Please cite as:
Haydon HM, Fowler JA, Taylor ML, Smith AC, Caffery LJ
Psychological Factors That Contribute to the Use of Video Consultations in Health Care: Systematic Review
J Med Internet Res 2024;26:e54636
URL: https://www.jmir.org/2024/1/e54636
doi:10.2196/54636
PMID:

©Helen M Haydon, James A Fowler, Monica L Taylor, Anthony C Smith, Liam J Caffery. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Factors Associated With Digital Intervention Engagement and Adherence in Patients With Cancer: Systematic Review

Lucile Montalescot[1], PhD; Louise Baussard[1], PhD; Elodie Charbonnier[1], PhD

Laboratoire de Psychopathologie et Processus de Santé, Université Paris-Cité, Boulogne-Billancourt, France

Corresponding Author:
Lucile Montalescot, PhD
Laboratoire de Psychopathologie et Processus de Santé
Université Paris-Cité
71 avenue Edouard Vaillant
Boulogne-Billancourt, 92100
France
Phone: 33 1 76 53 29 81
Email: lucile.montalescot@u-paris.fr

Abstract

Background: Digital interventions offer vital support for patients with cancer through education, behavior change, and monitoring. Despite their potential, patient adherence to and engagement with these self-help interventions is challenging. Factors like user characteristics, technology, and intervention design influence adherence and engagement. Existing reviews have gaps in exploring diverse factors associated with adherence in cancer care.

Objective: This systematic review aims to identify factors influencing adherence to and engagement with digital interventions with self-help components in cancer care. It examined sociodemographic, psychosocial, health-related, and intervention-related factors that affect patients' adherence to and engagement with these digital health solutions.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a search was conducted across PubMed, Embase, Cochrane Library, and PsycINFO to find studies published from January 2010 to September 2021. The studies included in this review focused on adult patients with cancer using digital interventions with self-help features. Data were extracted and synthesized using a standardized approach. Factors associated with adherence were synthesized according to their type—sociodemographic factors, psychosocial factors, health-related factors, technology-related factors, and intervention-related factors.

Results: Among 9386 studies initially screened, 61 (0.6%) were eligible for analysis. These studies covered diverse eHealth intervention types, cancer types, and outcome measures. Investigating the determinants of adherence to and engagement with digital interventions was the main objective for 43% (26/61) of the included studies. Adherence and engagement were gauged using varied measures, such as dropout rates, log-ins, and self-reported measures. Results regarding factors associated with adherence and engagement were inconsistent across studies. Most sociodemographic (eg, age) and health-related factors (eg, cancer stage) yielded mixed outcomes. However, comorbidity consistently predicted lower adherence and engagement. Results regarding psychosocial factors were more stable across studies. Specifically, higher social support was associated with lower adherence and engagement. Finally, intervention-related factors like intervention type or human support showed conflicting results. Adopting an intersectional perspective revealed that specificities vary according to intervention goals and the operationalization of adherence versus engagement, with women being more adherent and engaged than men in interventions targeting distress. When focusing on adherence rather than engagement, older patients were more adherent than younger patients.

Conclusions: This review highlights the complexity of adherence to and engagement with digital interventions in cancer care. While some factors, notably comorbidities and low social support, were consistently linked to adherence and engagement, others displayed mixed associations. The review underscores the need for standardizing measures, investigating specific intervention features, and enhancing study quality to optimize digital interventions for patients with cancer. Further research is crucial to better understand and improve adherence to digital health solutions in cancer care.

Trial Registration: PROSPERO CRD42021281028; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=281028

(J Med Internet Res 2024;26:e52542) doi:10.2196/52542

XSL•FO
RenderX

KEYWORDS

adherence; engagement; eHealth; mHealth; cancer; mobile health; app; eHealth interventions; patient; cancer care; digital health; health-related; intervention-related; sociodemographic; behavior; systematic review

Introduction

Background

Digital interventions include a range of technologies, such as telehealth, mobile health, and web-based platforms that provide health-related information, self-help, support, and monitoring [1]. They have become increasingly popular in recent years as a means of delivering health care services, promoting patient self-care, and improving their health [2], especially for patients with cancer. Indeed, the intricate trajectory of the treatment journey of patients with cancer, spanning diverse health care settings, can be significantly enhanced through the use of digital interventions [3]. By acknowledging the needs encountered by patients with cancer, these interventions proficiently cater to their requirements by providing educational materials, behavior change support, and access to self-help resources [4-6]. Furthermore, digital interventions are pivotal in fostering long-term survivorship care by facilitating the formulation of personalized treatment plans, vigilant monitoring, and advocating for healthy lifestyle choices [7]. More precisely, many digital interventions have been created and tested in the context of cancer care, to improve the patient's quality of life and symptoms as well as promote healthy behaviors [7-9].

Digital health encompasses a wide range of interventions, from forums to websites, with various objectives (eg, information, sharing of experience, self-assessment, behavior change). Numerous eHealth interventions rely on self-help [10,11] (ie, interventions that can be worked through independently by patients themselves). Patients with cancer hold a positive attitude both toward self-help and eHealth self-help, specifically [12]. Digital interventions with a self-help component seem to be cost-effective in cancer care [13] and research has shown that they could be as efficacious as in-person interventions [14]. This specific type of digital intervention seems to both address patients' needs and be efficacious in improving their quality of life and overall well-being. However, several studies have reported difficulties in the implementation of digital interventions, several of them being related to patients' engagement and adherence [15]. Although interrelated, both concepts encompass different experiences and behaviors. Engagement has been defined by Perski et al [16] as both a subjective experience characterized by focused attention, interest, and affect as well as the behaviors associated with this experience. These behaviors include the frequency and duration of use of the digital health intervention. By contrast, adherence to eHealth could be defined as the congruence between the intended use of technology and the effective use by an individual. Moreover, the justification for the intended use should be supported by theory or rationale [17].

Adherence and engagement are often used interchangeably in studies on digital interventions. For example, some operationalize "adherence" to refer to "the more use, the better," without specifying an intended use. This, of course, brings up measurement issues and disparities [17,18]. Engagement is typically measured with log-ins, time spent on an intervention, or number of clicks [16]. Owing to the confusion between engagement and adherence, the latter is often measured with the same indicators as engagement. However, Sieverink et al [17] noted that previous studies that presented adherence, mostly used a measure of completion of the intervention (eg, number of modules accessed and completed). Indeed, previous work has shown that patients were less likely to be adherent to self-help interventions in comparison with interventions involving real-time interactions [19]. Engagement is an important predictor of the effectiveness of these interventions [14,20-22]. Indeed, the more patients use eHealth, the more it is effective. Although digital interventions have the potential to improve patient outcomes and increase access to health care services, adherence to and engagement with these interventions remain a challenge [23]. It is important to note that the patient's view of the intervention is essential to its successful implementation [15]. Understanding the factors that influence adherence to and engagement with digital interventions is critical for developing effective interventions.

Few reviews have examined the factors associated with adherence to and engagement with digital interventions. They showed that engagement depended on users' characteristics, technological aspects, and intervention features [24]. Furthermore, they highlighted that components such as personalized content, push notifications, and quizzes were associated with increased adherence and engagement [25,26]. However, the focus of these systematic reviews limits their reach. Indeed, the 3 systematic reviews that specifically explored adherence to or engagement with digital interventions focused on the intervention features that increased the said adherence and engagement or were conducted in very heterogeneous populations (eg, patients with gynecological problems, caregivers of disabled children) [24-26], without addressing the specificities of cancer care. Another problem lies in the difficulty to define adherence in the context of digital health. In most definitions, use, engagement, and adherence are considered as synonyms [17,27]. While some studies investigate digital health engagement exclusively through log data (eg, number of logins, number of clicks, time spent on a module), other authors argue that adherence should be conceptualized as the degree to which users followed the program as it was designed [27]. In summary, the diversity in the definitions of adherence and engagement makes it challenging to investigate the topic.

In sum, to date, adherence to and engagement with digital health, and particularly to self-help interventions, remains difficult to define, and the factors involved in engagement and adherence have been inadequately investigated, specifically for cancer. Indeed, most systematic reviews published to date on digital health interventions focus on their efficacy [9,28,29] or users' experience with such interventions [30,31], with a small portion of them mentioning engagement as a secondary objective [19,32].

XSL•FO
RenderX

This Review

In this systematic review, we aim to identify and synthesize the existing literature on the factors associated with adherence or engagement with digital interventions presenting a self-help component in cancer care. Specifically, we examine the sociodemographic, psychosocial, health-related, and intervention-related factors that influence digital health adherence. We also investigate specificities according to intervention goal, operationalization of adherence and engagement, and intervention type.

Methods

Overview

The review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. This systematic review has been registered in the PROSPERO database (CRD42021281028).

Search Strategy

A comprehensive search strategy was developed by EC and LM to identify relevant studies. The following databases were searched: PubMed, Embase, Cochrane Library, and PsycINFO. The search terms included cancer, eHealth interventions, adherence, and related synonyms. The search strategy is presented in Multimedia Appendix 1. The last systematic review with a similar objective was published in 2011 [24]. Therefore, the search was limited to studies published in English from January 2010 to September 2021.

Study Selection and Data Extraction

Two reviewers (EC and LM) independently screened titles and abstracts to identify potentially eligible studies through Rayyan software (Rayyan Systems, Inc). Once the blind was off, disagreement was resolved through consensus. Full-text articles were retrieved for studies that met the inclusion criteria. Studies were included if the following criteria were met: (1) they investigated factors associated with adherence to or engagement with digital interventions, (2) said interventions included a self-help component (ie, components that could be worked through independently by patients themselves and implied an active participation from them), and (3) included adult patients with cancer. Studies were excluded if they (1) were not original research (eg, reviews, editorials, and commentaries), (2) did not report empirical data, (3) were qualitative studies, (4) were not intervention trial, (5) only included information modules, (6) included symptom reporting only, without an active component, or (7) included a communication module with health care providers, without an active component.

Three reviewers extracted data from eligible studies using a standardized data extraction form. The following data were extracted: authors, year, country, title, study design, population, intervention type, aim of the intervention, operationalization of adherence and engagement, adherence or engagement measure, intervention duration, study length, primary outcome of interest, analysis, and factors associated with adherence (Multimedia Appendix 2).

Data Synthesis

A narrative synthesis was conducted owing to the heterogeneity of the included studies in terms of study design, intervention type, and outcome measures. Factors associated with adherence and engagement were synthesized according to their type: sociodemographic factors, psychosocial factors, health-related factors, technology-related factors, and intervention-related factors. We also performed subgroup synthesis according to the intervention goal, operationalization of adherence and engagement, and intervention type.

The quality of the included studies could not be assessed with a standardized evaluation grid because of the variety of designs and objectives. However, we proceeded to a narrative critical appraisal of the methods used across studies. We also quantified the studies for which we identified missing summary statistics.

Results

Overview

The search identified 9386 potentially eligible studies, of which 61 (0.6%) [33-93] were included in the final review. A detailed flowchart is available (Figure 1).

XSL•FO
RenderX

Figure 1. Studies flowchart.

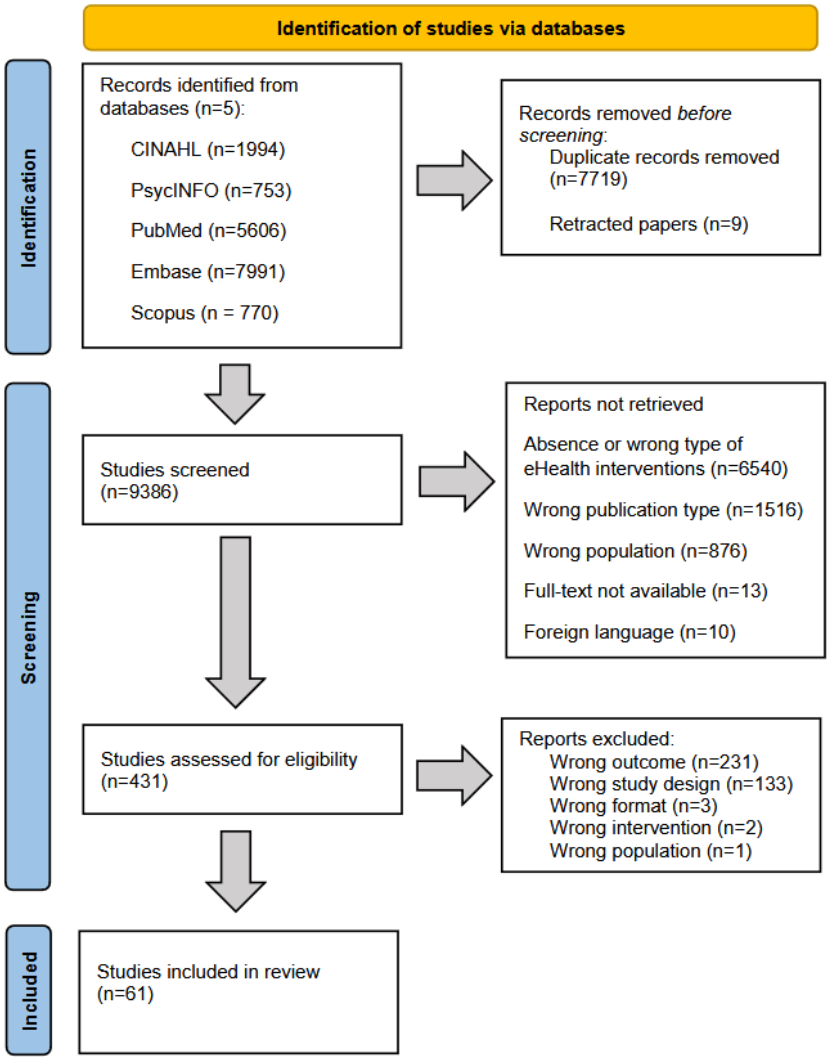

The interventions included web-based interventions (39/61, 64%), mobile-based interventions (18/61, 30%), other types of digital health interventions, and a combination of different types of technologies (4/61, 7%). The goals of the interventions varied (eg, decreasing distress and behavior change). Regarding the population, breast cancer was the most common cancer type across studies (41/61, 67%). Furthermore, a large portion of studies targeted specifically cancer survivors (25/61, 41%). Adherence and engagement were assessed with several measures in 57% (35/61) of the included studies. The measures used included dropout rates (20/61, 33%), time spent on the intervention (20/61, 33%), number of log-ins (18/61, 30%), number of patients' actions within the intervention (eg, message sent, clicks; 15/61, 25%), number of pages or modules viewed (12/61, 20%), completion of the intervention or its modules (10/61, 16%), self-reported measures of use (7/61, 12%), number of active days or weeks (6/61, 10%), specific measures linked to the use of wearables (3/61, 5%), and intention to use the application (2/61, 3%). About 10% (6/61) of the included studies used adherence measures specific to their intervention (eg, doing several specific actions, such as creating a user profile and posting >2 messages to a group). Even within these categories, discrepancies in measures are of note (eg, binarization of log-ins measures, completion of the intervention as a whole vs its modules, and number of pages viewed vs view of a specific page). Finally, investigating the determinants of adherence to eHealth was the main objective for 43% (26/61) of the included studies. Factors associated with adherence and engagement included sociodemographic factors, psychosocial factors, health-related factors, and intervention-related factors. Unsurprisingly, adherence was highly associated with other measures of engagement [33-37]. Table 1 summarizes the characteristics of the included studies while Multimedia Appendix 2 presents each study individually in greater detail.

XSL•FO
RenderX

Table 1. Descriptive statistics for the included studies (n=61).

	Studies, n (%)
Type of study	
RCT[a]	31 (51)
Secondary analysis of RCT	15 (25)
Observational	15 (25)
Studies where most of the patients had breast cancer	41 (67)
Studies on cancer survivors	25 (41)
Adherence operationalization	
Intention	2 (3)
Engagement	32 (52)
Adherence without justification	25 (41)
Adherence with justification	2 (3)
Measures	
Combination	35 (57)
Dropout rate	20 (33)
Time spent on the intervention	20 (33)
Number of log-ins	18 (30)
Number of actions within the intervention	15 (25)
Number of page views	12 (20)
Completion of intervention or modules	10 (16)
Self-reported measures of use	7 (11)
Number of active days or weeks	6 (10)
Measures of wearables	3 (5)
Intention to use the app	2 (3)
Specific measures	6 (10)
Type of intervention	
Web-based intervention	39 (64)
Mobile-based intervention	18 (30)
Others or combination	4 (7)
Goal of the intervention	
Distress	15 (25)
Nutrition and physical activity	10 (165)
Quality of life and symptom management	10 (15)
Others	14 (23)
Combination	12 (20)

[a]RCT: randomized controlled trial.

Sociodemographic Factors Associated With Adherence to and Engagement With Digital Interventions in Patients With Cancer

We found that 38 of 61 (62%) studies investigated the links between sociodemographic factors and engagement with and adherence to digital interventions, with many inconsistent results between studies. First, of the 14 studies concerning age, 6 (43%) showed that older patients tended to be more adherent and engaged than younger ones [34,38-42], 4 (29%) concluded that younger patients were more adherent and engaged [43-46], while 3 (21%) showed no significant associations between age and adherence or engagement [33,47,48]. Finally, 1 (7%) of these studies showed a differential association between age and engagement with digital health interventions; participants' age was not significantly related to smartwatch-wearing compliance

XSL•FO
RenderX

but was significantly and positively correlated with higher symptom rating compliance [49].

Second, of the 13 studies concerning education level, 5 (38%) studies reported that a higher level of education was associated with higher engagement or adherence [38,41,43,50,51], 3 (23%) reported the contrary [34,44,52], and 5 (38%) yielded nonsignificant results [33,39,40,47,53]. One (8%) study reported that patients with some college degree showed the highest decline in engagement compared with high school graduates and college graduates without reporting any statistical tests [39]. Reliable trends regarding economic factors were difficult to determine because of inconsistent results. Among the 5 studies that investigated these links, 2 (40%) showed that being from a privileged background was associated with better adherence and engagement [41,54], while 3 (60%) showed nonsignificant results [33,35,47]. One (20%) study highlighted how income and employment status were associated with the differential use of specific modules [41].

Third, of the 9 studies concerning employment, 4 (44%) studies showed that employed patients were also found to be more adherent and engaged than those who did not work [39,43,50,55], 2 (22%) reported they were less likely to be adherent and engaged [56,57], and 3 (33%) reported nonsignificant results [33,39,53].

Fourth, of the 7 studies concerning gender, 3 (43%) studies showed that women were more adherent or engaged than men [43,50,58], 1 (14%) showed the contrary [34], while 3 (43%) showed no significant association [47-49]. One (14%) study showed that women were more likely to use the printed materials of a hybrid intervention than men [48].

Fifth, of the 5 studies that examined the association between marital status and adherence, 2 (40%) studies suggested that married patients may be more adherent and engaged [38,43], 2 (40%) studies yielded nonsignificant results [39,47], and 1 (20%) showed that married patients were less adherent and engaged [59]. One (20%) study showed that a higher number of people in the household was associated with lower engagement. This same article highlighted that widowed or divorced patients and participants in single households showed a greater decline in engagement, but no statistical test was reported [39].

Sixth, of the 3 studies on race, 2 (67%) studies yielded nonsignificant associations [33,35], and 1 (33%) reported a significant association that showed White women were more likely to use a discussion group module [44].

Finally, 3 studies reported that patients who were more experienced with technology were found to be more adherent and engaged. Although experience with technology was found to be associated with higher use in 2 (67%) studies [33,43], it was not associated with higher use of specific modules [60]. Beyond experience, access to technology was associated with better digital health engagement in 1 (33%) study [35].

Health-Related Factors Associated With Adherence to and Engagement With Digital Interventions in Patients With Cancer

We found that 30 of 61 (49%) studies reported the links between health-related factors and adherence to digital interventions, with, again, many inconsistent results between studies. First, of the 7 (23%) studies that described the stage of cancer, 3 (43%) studies showed that patients living with a more advanced cancer stage were more adherent and engaged [44,61,62], while 3 (43%) did not show any significant associations [33,41,48]. The last 1 (14%) showed that patients living with cancer stage II had the lowest engagement rate. This same study argued that patients who had breast cancer stage III showed a greater decline in engagement, without reporting any statistical test [39].

Second, of the 5 (8%) studies concerning symptoms, 2 (40%) studies showed that patients with more symptoms tended to be more adherent and engaged [34,63], while 3 (60%) reported no significant associations [39,64,65].

Third, of the 4 (7%) studies concerning comorbidity, patients with comorbidities were found to be less adherent and engaged in 2 (50%) studies [38,60], 1 (25%) study reported nonsignificant associations [41], and in 1 (25%) study, the significance of this association depended on the time of the intervention and the comorbidity measure used [39].

Fourth, of the 3 (5%) studies concerning weight or other related measures, 1 (33%) study showed that patients with a higher percentage of body fat were less adherent and engaged [66], 2 (67%) reported no significant association [39,41], and 1 (33%) study highlighted a greater decline in engagement in participants with high BMI without reporting statistical tests [39].

Fifth, among the 3 (5%) studies that investigated the links between diagnosis date and adherence or engagement, 1 (33%) showed that patients with an older date of diagnosis date had higher engagement [47] and the other 2 (67%) reported no significant results [39,67]. The moment patients are proposed to use the intervention also seems to play a role in their adherence and engagement [35,44,45,52,62]. Only 2 (40%) studies showed no significant associations [33,67]; however, authors used very different indicators (eg, summertime, time since diagnosis, and pre or postsurgery status), which makes it difficult to identify trends.

Sixth, of the 2 studies that examined the association between cancer type, 1 (50%) study showed that patients with breast cancer had higher engagement [62], while 1 (50%) study showed no significant differences in adherence according to cancer type [48]. Finally, 1 (2%) study among the 61 included, although not directly having cancer type as a predictor, demonstrated that the predictors of engagement differed between patients with breast cancer and those with prostate cancer [38].

Seventh, 5 (8%) studies investigated the role of treatments. Four (80%) showed that treatments and medical services could be associated with adherence or engagement [41,42,50,53], but the type of treatments investigated varied greatly (eg, surgery, sleeping medication) and the operationalization of these variables differed across studies (eg, medical service, cycles of

XSL•FO
RenderX

chemotherapy). One (20%) study did not show any difference in engagement between patients who underwent chemotherapy and those who did not [39].

Finally, 1 (2%) study showed that patients with cancer had higher engagement scores than participants who did not have cancer [68]. However, in this study, no significant differences emerged in other adherence markers, such as homework completion.

Psychosocial Factors Associated With Adherence to and Engagement With Digital Interventions in Patients With Cancer

Despite fewer studies examining psychosocial factors (14/61, 23%), they yielded more consistent results. First, of the 13 (93%) studies concerning distress, 6 (46%) studies showed that patients who were distressed were more adherent or engaged [33,44,57,58,69,70], 4 (31%) showed the contrary [35,47,50,71], and 3 (23%) did not report any significant results [36,39,72]. All the studies that showed a positive association used a measure of cancer-related distress [33,44,57,58,69-71]. In comparison, among those that showed the contrary, only 1 (25%) used a cancer-specific measure [47]. Moreover, 1 (8%) of these studies showed how distress was associated with engagement with different modules depending on patients' gender [71]. One (8%) study showed a differential association depending on the operationalization of engagement (eg, binarization of use vs use of specific modules) [71]. Finally, 1 (8%) study showed that although distress was a predictor of continuous app use, when controlling for gender, this association was no longer significant [58].

Second, of the 7 (50%) studies that investigated the link between quality of life and adherence or engagement, 3 (43%) studies showed that patients with better quality of life were more adherent or more engaged [65,71,73], 1 (14%) reported the contrary [44], and 3 (43%) did not report significant results [39,56,60].

Third, of the 6 (43%) studies investigating social support, 4 (67%) studies showed that greater social support was associated with decreased adherence or engagement [44,60,69,70] and 1 (17%) showed the contrary [34]. Moreover, 1 (17%) of these studies showed a differential association between social constraints and engagement with digital interventions; participants' social constraints were positively associated with duration of use but negatively with unique module views [33]. As with distress, 1 (17%) study highlighted a different pattern of association between social support and adherence depending on the gender of the participants [60].

Fourth, although self-efficacy was investigated in only 3 (21%) studies, it was found to be positively associated with adherence and engagement in 2 (67%) studies [47,65], while the remaining 1 (33%) reported nonsignificant results [60].

Fifth, positive perceptions of digital interventions (eg, perceived ease of use, perceived usefulness) were consistently associated with better adherence and higher engagement in the 5 (8%) studies that investigated this factor [37,46,59,74,75]. However, 1 (20%) of these studies showed a differential association between intervention perception and adherence; perceived usefulness was positively associated with the intention to use the intervention, but perceived ease of use was not [74].

Sixth, the 3 (5%) studies that investigated the link between information competence and adherence and engagement reported nonsignificant results [33,44,76]. However, 1 (33%) of these 3 studies showed a differential association between information competence and engagement with specific modules; the more patients had high information competence, the less they used "ask the expert" and interactive services, but not information services [44].

Seventh, 3 (5%) studies investigated the links between resistance (such as lack of motivation or difficulty with change) and adherence or engagement. One (33%) of the 3 studies showed a positive association [58], another 1 (33%) showed a negative association [35], and the last 1 (33%) had no significant results [77]. In 1 (33%) study, although resistance to change was a predictor of continuous app use, when controlling for gender, this association was no longer significant [58].

Finally, other psychosocial factors have been found to be associated with adherence or engagement in single studies, such as decisional conflict [76], health perceptions [35], coping (helplessness and anxious preoccupation) [72], personality (openness) [58], previous experience with a similar intervention (mindfulness) [50], fatigue [55], and unmet sexual and physical needs [57]. Other studies investigated different factors without highlighting significant results like therapeutic alliance [78]. One (7%) study stated that patients reporting fatigue showed a greater decline in adherence without reporting statistical tests [39].

Intervention-Related Factors Associated With Adherence to and Engagement With eHealth Interventions in Patients With Cancer

Twenty-six of 61 (43%) studies investigated the links between intervention-related factors and adherence to digital interventions. First, of the 11 (42%) studies concerning intervention type, most studies compared adherence or engagement with digital intervention with other interventions which made it difficult to summarize these results. Four (36%) studies compared digital interventions with paper pamphlets; 3 (75%) of them showed that participants were more adherent and engaged to the digital version of the intervention [45,79,80], and the last 1 (25%) showed no significant differences between the 2 types of intervention [54]. Two (18%) studies compared digital interventions with usual care; 1 (50%) showed that participants in the digital group engaged more in survivorship care plans than people who did not use the digital intervention [37], while the other 1 (50%) showed no significant differences in dropout rates between the digital group and the usual care group [69]. Two (18%) studies compared interactive digital interventions with information-only portals; 1 (50%) yielded inconsistent results depending on the measure used [81], and the other 1 (50%) had nonsignificant results [72]. Two (18%) studies compared eHealth interventions with face-to-face ones; 1 (50%) showed that participants who participated in the eHealth intervention were less adherent than those who participated in

XSL•FO
RenderX

its face-to-face version [78], and the other 1 (50%) did not show any significant differences [34]. Two (18%) studies compared phone interventions with digital ones showing no significant differences [54,66]. Finally, 4 (36%) studies compared 2 different interactive digital interventions; 3 (75%) did not show any significant differences [39,82,83], while 1 (25%) highlighted inconsistent results depending on the chosen measure of engagement [51]. Interestingly, the possibility of interactions between patients yielded nonsignificant associations in the 2 (100%) studies that investigated this topic [74,84].

Second, the link between human support (ie, help from a human with the use of the intervention) and adherence has been studied in only 2 (8%) studies, and the results were conflicting. Of the 2 studies, 1 (50%) showed that human support was associated with increased adherence [85], while the other 1 (50%) showed the contrary [64]. However, it is worth mentioning that 1 (50%) of these studies showed a differential association between human support and engagement and adherence; participants in the technician-guided group logged in more frequently than the self-help group but no significant differences were observed in the proportion of participants who completed lesson 5 between both groups [85].

Third, of the 2 (8%) studies that investigated the effect of time on engagement, 1 (50%) study showed that log-in attrition was significant across the 3 months of the study [33], while the other 1 (50%) showed no significant effect of time on engagement [86].

Finally, although most studies included in this review did not examine the differential use of modules (only 1 study reported no significant differences in module use [87]), a subset of studies investigated the impact of specific features on adherence and engagement. Findings from these studies suggest that a tunneled intervention, in which modules are presented in a fixed sequence, may lead to higher engagement than a free-choice intervention [51]. In addition, reminders were found to be effective in improving adherence and engagement [88,89], regardless of the type of reminder used [90]. Finally, patients were found to consult modules more frequently when informed through interventions tailored to meet their needs [59].

Intersectional Approach to Adherence to and Engagement With Digital Health Interventions

Given the high heterogeneity of studies and interventions as well as the contrasting results presented above, we examined these data with an intersectional approach. Indeed, we chose to investigate these factors by type of intervention, specifically the types of goals the included studies focused on (eg, physical activity and psychological distress), operationalization of the outcome (ie, engagement and adherence with or without justification for dose), and type of intervention (ie, web-based vs mobile-based). We reported on factors investigated by at least 2 studies in each subcategory.

Intervention Aims

Quality of Life and Symptom Management

A total of 13 (21%) interventions centered on quality of life or symptom management, including 3 (23%) that focused on other aims as well (eg, distress). For interventions that targeted quality of life, 2 (15%) studies examined the association between marital status and adherence and engagement, leading to inconsistent results. One (8%) suggested that married patients had a higher engagement [38], and another 1 (8%) showed contrary results [59]. Regarding comorbidities, patients with comorbidities had a lower engagement in the 2 (15%) studies that investigated this factor [38,60].

Among the 3 studies that investigated the link between quality of life and adherence and engagement, the results were inconsistent. One (33%) of the 3 studies showed that better quality of life was associated with higher adherence [73], another 1 (33%) reported the contrary [44], and the last 1 (33%) did not report significant results [60]. The 2 (67%) studies that investigated social support showed that greater social support was associated with decreased adherence or engagement [44,60].

Regarding intervention-related factors, reminders were found to be effective in improving engagement [88,89]. This was the only factor investigated by >2 studies.

Psychological Distress

Twenty-one out of 61 (34%) interventions centered on psychological distress, including 6 (10%) that focused on other aims as well (eg, decisional conflict, self-efficacy). Regarding sociodemographic factors, 1 (33%) study reported that a higher level of education was associated with higher engagement [43], while 2 (67%) others yielded nonsignificant results [33,47]. The 2 (100%) studies that investigated economic background showed nonsignificant results [33,47]. Concerning employment, a reliable trend could not be identified as 1 (33%) study found that employed patients were more adherent than those who did not work [43], another 1 (33%) reported they were less likely to be adherent [57], and the last 1 (33%) reported nonsignificant results [33]. Regarding gender, 2 (67%) studies showed that women were more adherent and engaged than men [43,58], and 1 (33%) showed no significant association [47]. No reliable trend could be identified regarding marital status as 1 (33%) study suggested that married patients may be more adherent [43], another 1 (33%) yielded nonsignificant results [47], and the last 1 (33%) showed that married patients were less adherent [59]. Finally, experience with technology was found to be associated with higher use in 2 (100%) studies [33,43].

Regarding health-related factors, 2 (67%) studies showed that patients living with a more advanced cancer stage were more adherent [61], while 1 (33%) did not show any significant associations [33]. No reliable trend could be identified regarding the time the intervention was proposed to the patients. Indeed, 1 (50%) study showed significant results [62] while the other 1 (50%) showed no significant associations [33].

Concerning distress, 5 (56%) studies showed that patients who were distressed had higher levels of adherence or engagement [33,57,58,69,70], 2 (22%) showed the contrary [47,71], and 2 (22%) did not report any significant results [36,72]. Regarding social support, 2 (100%) studies showed that greater social support was associated with decreased engagement [69,70].

Regarding intervention-related factors, the 2 (100%) studies that compared digital interventions with paper pamphlets showed

XSL•FO
RenderX

that participants were more adherent to the digital version of the intervention [79,80]. Two (10%) studies compared interactive digital interventions with information-only portals; 1 (50%) yielded inconsistent results depending on the measure used [81], and the other 1 (50%) yielded nonsignificant results [72].

Physical Activity and Nutrition

Twelve out of 61 (57%) interventions centered on physical activity and nutrition, including 2 (17%) that also focused on other aims as well (eg, smoking). First, concerning age, 1 (25%) study showed that older patients tended to engage with the intervention more than younger ones [39], 2 (50%) showed the contrary [45,46], while 1 (25%) study showed no significant associations between age and adherence [48]. No reliable trend could be identified for education; 1 (33%) study reported that a higher level of education was associated with higher engagement [51], 1 (33%) reported the contrary [52], and the last 1 (33%) yielded nonsignificant results [39]. Regarding gender, the 3 (100%) studies that investigated the topic showed no significant association [48,49].

Regarding health-related factors, 1 (50%) study out of 2 did not show any significant associations between cancer stage and engagement [48], while 1 (50%) showed that patients living with cancer stage II had the lowest engagement rate [39]. Concerning symptoms, 1 (50%) study showed that patients with more symptoms tended to have higher engagement while the other 1 (50%) reported no significant associations [39]. Regarding weight or other related measures, 1 (50%) study showed that patients with a higher percentage of body fat had lower engagement [66] and the other 1 (50%) reported no significant association [39]. Finally, the moment patients are proposed to use the intervention also seems to play a role in their engagement among the 3 (100%) studies that investigated this topic [35,45,52].

Regarding psychological factors, no reliable trend regarding distress could be established as it was investigated by only 2 (17%) studies with contradicting results; 1 (50%) showed that more distressed patients had lower engagement [35], and the other 1 (50%) did not report any significant results [39]. However, positive perceptions of digital interventions were associated with better adherence in the 2 (100%) studies that investigated this factor within this subcategory [46,59].

Operationalization of Adherence and Engagement

Overview

We found that 2 out of 61 (3%) studies used measures of intention to continue using the intervention as an outcome. The others were classified according to the classification proposed by Sieverink et al [17]: engagement, adherence with justification for intended use, and adherence without justification for intended use [17]. However, only 2 out of 59 (3%) studies justified the intended use specified in the article. No similarity could be reported as their methods and results differed greatly. Therefore, we only report the results for studies that investigated engagement versus the ones that investigated adherence without justification for the intended use.

Engagement

We found that 32 out of 61 (52%) articles used measures of engagement according to the classification proposed by Sieverink et al [17]. First, concerning age, 10 studies investigated this topic: 4 (40%) studies showed that older patients tended to engage more than younger ones [38-41], 3 (30%) studies showed that younger patients engaged more [44-46], while 3 (30%) studies showed no significant associations between age and engagement [33,47,48]. Second, concerning the level of education, 3 (33%) studies reported that higher education was associated with higher engagement [38,41,51], 2 (22%) reported the contrary [44,52], and 4 (44%) yielded nonsignificant results [33,39,40,47]. Among the 4 (13%) studies that investigated economic factors, 1 (25%) showed how being from a privileged background was associated with better engagement [41], while the other 3 (75%) showed nonsignificant results [33,35,47]. Concerning employment, no reliable trend could be identified as the 3 (9%) studies that investigated the topic reached different conclusions; 1 (33%) showed that employed patients were found to have higher engagement than those who did not work [39], 1 (33%) reported they were less likely to engage in the intervention [56], and the last 1 (33%) reported nonsignificant results [33]. The 3 (100%) studies that investigated gender showed no significant association [47,48]. Regarding marital status, 1 (33%) study suggested that married patients may have higher engagement [38], and 2 (67%) studies yielded nonsignificant results [39,47]. For race, 2 (67%) studies yielded nonsignificant associations [33,35], and only 1 (33%) reported a significant association and showed that White women were more likely to use a discussion group module [44]. Finally, experience with technology was found to be associated with higher use in 1 (50%) study [33] but was not associated with higher use of specific modules in another (n=1, 50%) study [60].

Regarding health-related factors, 3 (50%) studies showed that patients living with a more advanced cancer stage had higher engagement [44,61,62], while 3 (50%) did not show any significant associations [33,41,48]. Patients with comorbidities were found to be less engaged in 2 (50%) studies [38,60], 1 (25%) reported nonsignificant associations [41], and in the other 1 (25%), the significance of this association depended on the time of the intervention and the comorbidity measure used [39]. Concerning symptoms, 3 (100%) reported no significant associations [39,65]. One (33%) study showed that patients with a higher percentage of body fat had lower engagement [66], while 2 (67%) reported no significant association [39,41]. No reliable trend could be identified regarding the date of diagnosis as 1 (50%) study showed that patients with an older date of diagnosis had higher engagement [47], and the other 1 (50%) reported no significant results [39]. The moment patients are proposed to use the intervention also seems to play a role in their engagement as shown in 5 studies (83%) [35,44,45,52,62]. Only 1 (17%) study showed no significant associations [33]. Regarding cancer type, 1 (50%) study showed that patients with breast cancer had higher engagement [62], while the other 1 (50%) showed no significant differences in engagement according to cancer type [48]. The association between engagement and treatments or medical services yielded

XSL•FO
RenderX

inconsistent results. One (50%) study showed significant differences [41], while another 1 (50%) did not show any difference in engagement between patients who underwent chemotherapy and those who did not [39].

Regarding psychological factors, 4 (40%) studies showed that distressed patients had higher engagement [33,44,69,70], 4 (40%) showed the contrary [35,47,50,71], and 2 (20%) did not report any significant results [39,72]. Two (33%) studies showed that patients with a better quality of life had higher engagement [65,71], 1 (17%) reported the contrary [44], and 3 (50%) did not report significant results [39,56,60]. Concerning social support, 4 (100%) studies showed that greater social support was associated with decreased engagement [44,60,69,70]. Self-efficacy was found to be positively associated with engagement in 2 (67%) studies [47,65], while the remaining 1 (33%) reported nonsignificant results [60]. Positive perceptions of digital interventions were associated with better engagement in the 2 (%) studies that investigated this factor in this subcategory [37,46]. Finally, the 3 (100%) studies that investigated the links between information competence and engagement reported nonsignificant results [33,44,76].

Relative to intervention-related factors, 2 (100%) studies showed that participants had better engagement with a digital version of the intervention compared with paper pamphlets [45,79]. Two (6%) studies compared digital interventions with usual care; 1 (50%) showed that participants in the digital intervention group were more engaged in survivorship care plans than people who did not use the intervention [37], while 1 (50%) study showed no significant differences in dropout rates between the digital group and the usual care group [69]. Two (6%) studies compared interactive digital interventions with information-only portals; 1 (50%) yielded inconsistent results depending on the measure of engagement used [81], and the other 1 (50%) yielded nonsignificant results [72]. Finally, 4 (13%) studies compared 2 different interactive digital interventions; 3 (75%) did not show any significant differences [39,82,83], while the other 1 (25%) highlighted inconsistent results depending on the chosen measure of engagement [51]. Two (6%) studies investigated the effect of time on engagement; 1 (50%) study showed that log-in attrition was significant across the 3 months of the study [33], while the second 1 (50%) showed no significant effect of time on engagement [86]. In addition, reminders were found to be effective in improving engagement [88,89].

Adherence Without Justification for Intended Use

We found that 25 out of 61 (41%) articles used measures of adherence without a justification for intended use, according to the classification proposed by Sieverink et al [17].

Regarding sociodemographic factors, 2 (100%) studies showed that older patients tended to be more adherent than younger ones [34,42]. Two (50%) studies showed that employed patients were more adherent than those who did not work [43,55], 1 (25%) reported they were less likely to be adherent [57], and 1 (25%) reported nonsignificant results [53]. Regarding gender, 2 (50%) studies showed that women were more adherent than men [43,58], 1 (25%) showed that men were more adherent [34], while 1 (25%) showed no significant association [49].

Regarding health-related factors, 2 (67%) studies showed that patients with more symptoms tended to be more adherent [34,63], while 1 (33%) reported no significant associations [64]. Two (100%) studies showed that treatments and medical services could be associated with adherence [42,53].

Regarding psychological factors, 2 (50%) studies showed that patients who were distressed were more adherent [58,69], 1 (25%) showed the contrary [71], and 1 (25%) did not report any significant results [36]. Two (100%) studies showed that patients with a better quality of life were more adherent [71,73]. Concerning social support, 1 (50%) study showed that greater social support was associated with decreased adherence [69], and 1 (50%) showed the contrary [34]. One (50%) study showed a positive association between resistance and adherence [58], and the second 1 (50%) showed no significant results [77].

Concerning intervention type, 1 (50%) study showed that participants were more adherent to a digital version of the intervention compared with a paper pamphlet [80], while the other 1 (50%) showed no significant differences between the 2 types of intervention [54]. Two (8%) studies compared eHealth interventions with face-to-face ones; 1 (25%) showed that participants who participated in the eHealth intervention were less adherent than those who participated in its face-to-face version [78], while the other 1 (25%) did not show any significant differences [34]. Second, the results regarding the links between human support and adherence were conflicting. One (50%) study showed that human support was associated with increased adherence [85], and the other 1 (50%) showed the contrary [64].

Type of Intervention

We chose to investigate factors associated with adherence and engagement according to the type of intervention (web-based or mobile-based). Two interventions could not be classified in either category.

Web-Based Intervention

We found that 41 out of 61 (67%) articles reported on web-based interventions, including 2 (5%) that also included other components (ie, CD-ROM, wearables). Regarding sociodemographic variables, 4 (40%) showed that older patients tended to be more adherent and engaged than younger ones [34,39,40,42], 3 (30%) showed the contrary [44-46], while 3 (30%) studies showed no significant associations [33,47,48]. Regarding the level of education, 3 (30%) studies reported that a higher level of education was associated with higher engagement [43,50,51], 2 (20%) reported the contrary [34,44] and 5 (50%) yielded nonsignificant results [33,39,40,47,53]. The studies that investigated economic factors showed nonsignificant results [33,47]. Concerning employment, 4 (50%) studies showed that employed patients were found to be more adherent or had higher engagement than those who did not work [39,43,50,55], 1 (13%) reported the contrary [57], and 3 (38%) reported nonsignificant results [33,39,53]. Two (40%) studies showed that women had higher engagement than men [43,50], 1 (20%) showed the contrary [34], while 2 (40%) showed no significant association [47,48]. One (25%) study suggested that married patients may be more adherent [43], 2 (50%) studies

XSL•FO
RenderX

yielded nonsignificant results [39,47], and another 1 (25%) showed the contrary [59]. Regarding race, 1 (50%) study yielded nonsignificant associations [33] and the only 1 (50%) that reported a significant association showed that White women were more likely to use a discussion group module [44]. Although technology experience was found to be associated with higher use in 2 (100%) studies [33,43], it was not associated with higher use of specific modules in another [60].

Regarding health-related factors, 3 (50%) studies showed that patients living with a more advanced cancer stage were more adherent [44,61,62], while 2 (33%) did not show any significant associations [33,48]. The last 1 (17%) showed that patients living with cancer stage II had the lowest adherence rate [39]. Patients with comorbidities were found to be less adherent in 1 (50%) study [60], and in another 1 (50%), the significance of this association depended on the time of the intervention and the comorbidity measure used [39]. Concerning symptoms, 2 (50%) studies showed that patients with more symptoms tended to be more adherent or more engaged in the intervention [34,63], while 2 (50%) reported no significant associations [39,64]. Concerning weight or other related measures, 1 (50%) study showed that patients with a higher percentage of body fat had lower engagement [66], and 1 (50%) reported no significant association [39]. Among the 3 (7%) studies that investigated the link between diagnosis date and adherence and engagement, 1 (33%) showed that patients with an older date of diagnosis date had higher engagement [47], and the other 2 (67%) reported no significant results [39,67]. The moment patients are proposed to use the intervention also seems to play a role in their engagement (3/5, 60%) [44,45,62]. Only 2 (40%) studies showed no significant associations [33,67]. Two (5%) study examined the association between cancer type, 1 (50%) study showed that patients with breast cancer had higher engagement [62], while 1 (50%) study showed no significant differences in engagement according to cancer type [48]. Three (75%) studies showed that treatments and medical services could be associated with engagement [42,50,53], but 1 (25%) study did not show any difference in engagement between patients who underwent chemotherapy and those who did not [39].

Regarding psychological factors, 5 (50%) studies showed that patients who were distressed had higher engagement or adherence [33,44,57,69,70], 2 (20%) showed the contrary [47,50], and 3 (30%) did not report any significant results [36,39,72]. Of the 3 (7%) studies that investigated the links between quality of life and adherence, 1 (33%) reported that patients with a lower quality of life had lower engagement [44], and 2 (67%) did not report significant results [39,60]. Concerning social support, 4 (80%) studies showed that greater social support was associated with decreased adherence or engagement [44,60,69,70], and 1 (20%) showed the contrary [34]. Self-efficacy was found to be positively associated with engagement in 1 (50%) study [47], while the remaining 1 (50%) reported nonsignificant results [60]. Moreover, the 3 (100%) studies that investigated the links between information competence and adherence or engagement reported nonsignificant results [33,44,76]. Regarding resistance, 1 (33%) study showed a positive association with adherence [58], another 1 (33%) showed a negative one [35], and the last 1 (33%) showed no significant results [77].

Concerning intervention-related factors, the 2 (100%) studies that compared web-based interventions with paper pamphlets showed that participants had higher engagement or adherence with the digital version of the intervention [45,80]. Two (5%) studies compared interactive digital interventions with information-only portals. One (50%) study yielded inconsistent results depending on the measure used [81], and the other 1 (50%) yielded nonsignificant results [72]. Two (5%) studies compared eHealth interventions with face-to-face ones; 1 (50%) showed that participants who participated in the eHealth intervention were less adherent than those who participated in its face-to-face version [78], while the other 1 (50%) did not show any significant differences [34]. Finally, 2 (5%) studies compared 2 different interactive digital interventions; 1 (50%) did not show any significant differences [39], while the other 1 (50%) highlighted inconsistent results depending on the chosen measure [51]. Regarding human support the results were conflicting. One (50%) study showed that human support was associated with increased adherence [85], and the other 1 (50%) showed the contrary [64]. Finally, reminders were found to be effective in improving engagement [88,89].

Mobile-Based Intervention

We found that 18 out of 61 (30%) articles reported on mobile-based interventions. Regarding sociodemographic factors, 2 (67%) studies showed that older patients tended to engage less with the interventions than younger ones [38,41], and 1 (33%) showed the contrary [46]. Two (50%) studies reported that higher education was associated with higher engagement [38,41], 1 (25%) reported the contrary [52], and 1 (25%) yielded nonsignificant results [39]. Moreover, 2 (100%) studies showed how being from a privileged background was associated with better adherence or engagement [41,54]. Reliable trends could not be established concerning employment as 1 (33%) study showed that employed patients were found to be more engaged than those who did not work [39], 1 (33%) reported the contrary [56], and 1 (33%) reported nonsignificant results [39]. One (50%) study showed that women were more adherent than men [58], and the other 1 (50%) showed no significant association [49]. One (50%) study suggested that married patients had higher engagement [38], while the other 1 (50%) yielded nonsignificant results [39].

Among health-related factors, patients with comorbidities had a lower engagement in 1 (50%) study [38] and 1 (50%) reported nonsignificant associations [41]. The moment patients are proposed to use the intervention also seems to play a role in their adherence [35,52].

One (33%) showed that distressed patients were more adherent [58], while 2 (67%) showed the contrary [35,71]. Two (67%) studies showed that patients with better quality of life had higher engagement [65,71], and 1 (33%) did not report significant results [56]. Positive perceptions of digital interventions were consistently associated with better adherence and engagement in the 4 (100%) studies that investigated this factor [37,46,74,75]. One (50%) study showed a positive association between resistance and adherence [58], while another 1 (50%)

XSL•FO
RenderX

showed a negative association [35]. Regarding intervention-related factors, 1 (25%) study showed that participants were more adherent to the digital version of the intervention than to a paper pamphlet [79], and the other 1 (25%) showed no significant differences between the 2 types of intervention [54]. Finally, 2 (50%) studies did not show any significant differences [82,83].

Quality of the Included Studies

The quality of the included studies could not be assessed with a standardized evaluation grid, because for several studies, examining the factors associated with adherence was a secondary objective. Therefore, the quality of the method used to assess this objective may not reflect the overall quality of the articles examined.

The results reported in the previous sections varied greatly, including in the same study, depending on the operationalization of the factors under study (eg, social support vs social constraints), their measures (eg, dichotomization of a variable), the way adherence was conceptualized (eg, active app days, percent of messages read) [37], as well as the composition of the sample under investigation (eg, samples composed of people aged ≥65 years [40]). Moreover, some of these results may be imputable to other confounding variables (eg, cancer type and gender).

Finally, it is noteworthy that 15 of the 61 (25%) articles reported a nonsignificant association between some variables under study and adherence, without specifying which ones. This may lead to an overestimation of the effects of said variables, especially sociodemographic and health-related ones [27,35,36,54,56,60,61,69,70,72,77,78,91-93].

Discussion

Principal Findings

This review aimed to highlight factors associated with adherence to digital interventions in cancer care. Despite the large number of articles included in this review, most results were heterogeneous across studies. For illustration, sociodemographic factors were the most investigated, but our review showed that these had the most inconsistent results. This may be due to the heterogeneity in the types of intervention under study (ie, duration, objective, and format), the diversity of the populations included (eg, adolescents and young adults, patients with breast cancer), as well as the inconsistency in the measures of adherence (eg, log-ins, time spent, self-reported measures), and its predictors. However, it is also important to note that some results were consistent across most studies. For example, the most adherent patients to digital health were those who were older, without comorbidities, with a positive perception of the intervention, and a low level of social support. Given the large number of factors covered by this review and the heterogeneity of the results, this discussion will focus exclusively on the most surprising results as well as those that were most widely agreed upon in the included studies.

Regarding health-related factors, the factor on which there is the greatest consensus is the presence of comorbidity. Indeed, the association between comorbidities and low adherence to digital health is noteworthy. Similar results have been found in previous studies in older adults and patients living with rheumatoid arthritis [94,95]. The authors highlight that some comorbidities may prevent access to digital interventions (eg, cognitive functioning and vision problems) [95]. Another explanation for these results may be that current digital interventions in cancer care do not address the complexity of these patients' health and care journeys [96]. These data suggest that the presence or absence of comorbidities must absolutely be taken into account in the conception of adequate digital interventions for patients with cancer.

Regarding psychosocial factors, the results concerning the effect of distress level on adherence were surprising. While distress was found to be positively associated with adherence to digital interventions in little more than half the studies, the results remained inconsistent across other studies. These contradictory results could be explained by the variability of definitions and measures of distress. In patients presenting a mental illness, adherence to digital interventions for depression and anxiety is often low to moderate and predicted by their severity [97-100]. Symptoms of such disorders (eg, anhedonia, lack of motivation) could hinder adherence to these interventions. In the context of cancer, we highlighted that all the studies that showed that higher distress was associated with increased adherence used cancer-specific measures of the concept. Here, the association with distress may be more of an indicator of the need for support regarding their cancer experience than severe symptoms of psychological distress, or even mental illness. In other words, cancer-related distress is an indicator of cancer-related needs; the more patients report such needs, the more likely they are to use digital interventions. The association between distress and adherence to digital health highly depends on how distress is conceptualized (eg, depression, cancer-related distress, anxiety). Consequently, identifying the type and level of distress of patients is crucial to addressing their needs and improving their adherence to the digital interventions offered.

Still among the psychosocial factors, one predictor of adherence that seemed to enjoy consensus in the studies included in this review was social support. More specifically, most studies have shown that low social support was associated with better adherence. This is congruent with studies that showed that cancer patients with low social support were more likely to seek health information on the web [101]. Connection with peers is part of the experience of digital health users, and some interventions include components to foster interactions between them [31]. Indeed, they can successfully improve perceived social support in patients with cancer [102]. Meeting similar people through a shared digital intervention could create a sense of community and decrease feelings of isolation [103,104]. However, not everyone is comfortable with social network features, so these aspects may be less relevant for people who already have a satisfying level of social support [104]. Finally, social support has been linked to better psychosocial outcomes and self-management behaviors in patients [105-107]. Consequently, patients with high levels of social support may be less inclined to use such interventions.

The adoption of an intersectional lens allowed us to highlight specificities according to intervention goals and

XSL•FO
RenderX

operationalization of adherence versus engagement. For interventions targeting distress in patients with cancer, gender might play a more important role than in other types of intervention, with women being more adherent and engaged than men. This might be explained by more negative attitudes toward mental health interventions in men [108,109]. Similarly, patients who are distressed seem to be more adherent to distress-focused interventions than other types of interventions. This may be explained by the relevance of such interventions to their needs. However, these results remain inconsistent across studies. Finally, when focusing on adherence rather than engagement, older patients were more adherent than younger patients. This aligns with previous research highlighting that older patients are more adherent than their younger counterparts [110].

Limitations

This systematic review has limitations worth mentioning. The main one is the significant heterogeneity of the articles included in this review, both conceptually and methodologically. From a conceptual point of view, the absence of a consensus on the definition of adherence in eHealth is a major obstacle. This echoes previous research highlighting a lack of consistency in the definition of this concept in the context of digital health [17]. Due to the wide variation of terms used to refer to adherence (eg, engagement, use, and usage), some relevant studies may not have been included. From a methodological point of view, the analysis of the included studies relies on the content of the articles, yet some of them presented incomplete data (eg, the details of the results and the nonsignificant associations with adherence were not reported). This may lead to an overestimation of the effect of some factors. Examining factors associated with adherence was the main objective of less than half of these studies which may explain the lack of details in some articles. In the future, it seems essential to better conceptualize adherence and to deepen the research into its determinants. Finally, our review included studies with patients with all types of cancers and at different stages of cancer. However, breast cancer was overrepresented in our review which may affect the external validity of our results. In future research, it could be interesting to target a particular type of cancer.

Despite these limitations, we may provide recommendations for the development of future digital interventions targeting patients with cancer. Personalization, specifically, seems to be essential. The platform should consider individual needs, including age, comorbidities, distress type (ie, cancer-specific or nonspecific), and levels. It should target isolated patients to provide tailored support and address gender-specific preferences (eg, regarding mental health). Usability and accessibility are essential, with continuous evaluation for ongoing improvement. Such an intervention would enhance patients' adherence and engagement, and ultimately, patient outcomes.

Conclusions

In summary, this systematic review examined factors associated with digital health adherence, aiming to provide a comprehensive understanding of the current state of research in this field. Our analysis revealed several key findings that shed light on the complexity of eHealth adherence. The results underscore the importance of health-related factors and psychosocial factors in predicting adherence. More specifically, the presence of comorbidities and the level of social support appear to be important factors to consider in ensuring patient adherence to digital interventions. However, our review also highlights the need for further investigation in this area, particularly by studying the effects of promising but poorly considered factors, such as self-efficacy. Finally, to gain a clear understanding of the factors involved in adherence to digital interventions for patients with cancer, it seems essential that future research should pay more attention to investigating the effects of specific features (eg, gamification, peer-support modules), standardizing other factors (eg, human support, comparison to other interventions), and homogenization of adherence measurements to enhance study quality. For example, a significant number of studies did not report which variables they investigated when they did not yield significant results. By addressing these gaps and limitations, future research can contribute to improving digital interventions and ultimately enhancing patient outcomes in this digital health care era.

Acknowledgments

This work was funded by myCharlotte.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File , 19 KB - jmir_v26i1e52542_app1.docx]

Multimedia Appendix 2
Descriptive table of the included studies.
[XLSX File (Microsoft Excel File), 72 KB - jmir_v26i1e52542_app2.xlsx]

XSL•FO
RenderX

Multimedia Appendix 3
PRISMA 2020 Checklist.
[PDF File (Adobe PDF File), 969 KB - jmir_v26i1e52542_app4.pdf]

References

1. Shaw T, McGregor D, Brunner M, Keep M, Janssen A, Barnet S. What is eHealth (6)? Development of a conceptual model for eHealth: qualitative study with key informants. J Med Internet Res 2017 Oct 24;19(10):e324 [FREE Full text] [doi: 10.2196/jmir.8106] [Medline: 29066429]
2. Meier CA, Fitzgerald MC, Smith JM. eHealth: extending, enhancing, and evolving health care. Annu Rev Biomed Eng 2013;15:359-382. [doi: 10.1146/annurev-bioeng-071812-152350] [Medline: 23683088]
3. den Bakker CM, Schaafsma FG, Huirne JA, Consten EC, Stockmann HB, Rodenburg CJ, et al. Cancer survivors' needs during various treatment phases after multimodal treatment for colon cancer - is there a role for eHealth? BMC Cancer 2018 Dec 04;18(1):1207 [FREE Full text] [doi: 10.1186/s12885-018-5105-z] [Medline: 30514325]
4. Kapoor A, Nambisan P, Baker E. Mobile applications for breast cancer survivorship and self-management: a systematic review. Health Informatics J 2020 Dec;26(4):2892-2905 [FREE Full text] [doi: 10.1177/1460458220950853] [Medline: 32842830]
5. Putranto D, Rochmawati E. Mobile applications for managing symptoms of patients with cancer at home: a scoping review. Int J Nurs Pract 2020 Aug;26(4):e12842. [doi: 10.1111/ijn.12842] [Medline: 32347599]
6. Keaver L, Loftus A, Quinn L. A review of iPhone and Android apps for cancer patients and survivors: assessing their quality, nutrition information and behaviour change techniques. J Hum Nutr Diet 2021 Jun 16;34(3):572-584. [doi: 10.1111/jhn.12857] [Medline: 33453133]
7. Adam R, McMichael D, Powell D, Murchie P. Publicly available apps for cancer survivors: a scoping review. BMJ Open 2019 Sep 30;9(9):e032510 [FREE Full text] [doi: 10.1136/bmjopen-2019-032510] [Medline: 31575584]
8. Rincon E, Monteiro-Guerra F, Rivera-Romero O, Dorronzoro-Zubiete E, Sanchez-Bocanegra CL, Gabarron E. Mobile phone apps for quality of life and well-being assessment in breast and prostate cancer patients: systematic review. JMIR Mhealth Uhealth 2017 Dec 04;5(12):e187 [FREE Full text] [doi: 10.2196/mhealth.8741] [Medline: 29203459]
9. Haberlin C, O'Dwyer T, Mockler D, Moran J, O'Donnell DM, Broderick J. The use of eHealth to promote physical activity in cancer survivors: a systematic review. Support Care Cancer 2018 Oct;26(10):3323-3336. [doi: 10.1007/s00520-018-4305-z] [Medline: 29909476]
10. Triberti S, Savioni L, Sebri V, Pravettoni G. eHealth for improving quality of life in breast cancer patients: a systematic review. Cancer Treat Rev 2019 Mar;74:1-14. [doi: 10.1016/j.ctrv.2019.01.003] [Medline: 30658289]
11. Morrison LG, Yardley L, Powell J, Michie S. What design features are used in effective e-health interventions? A review using techniques from critical interpretive synthesis. Telemed J E Health 2012 Mar;18(2):137-144 [FREE Full text] [doi: 10.1089/tmj.2011.0062] [Medline: 22381060]
12. Jansen F, van Uden-Kraan CF, van Zwieten V, Witte BI, Verdonck-de Leeuw IM. Cancer survivors' perceived need for supportive care and their attitude towards self-management and eHealth. Support Care Cancer 2015 Jun 26;23(6):1679-1688. [doi: 10.1007/s00520-014-2514-7] [Medline: 25424520]
13. van der Hout A, Jansen F, van Uden-Kraan CF, Coupé VM, Holtmaat K, Nieuwenhuijzen GA, et al. Cost-utility of an eHealth application 'Oncokompas' that supports cancer survivors in self-management: results of a randomised controlled trial. J Cancer Surviv 2021 Feb 12;15(1):77-86 [FREE Full text] [doi: 10.1007/s11764-020-00912-9] [Medline: 32656739]
14. Matis J, Svetlak M, Slezackova A, Svoboda M, Šumec R. Mindfulness-based programs for patients with cancer via eHealth and mobile health: systematic review and synthesis of quantitative research. J Med Internet Res 2020 Nov 16;22(11):e20709 [FREE Full text] [doi: 10.2196/20709] [Medline: 33196452]
15. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
16. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
17. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017 Dec 06;19(12):e402 [FREE Full text] [doi: 10.2196/jmir.8578] [Medline: 29212630]
18. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
19. Ownsworth T, Chan RJ, Jones S, Robertson J, Pinkham MB. Use of telehealth platforms for delivering supportive care to adults with primary brain tumors and their family caregivers: a systematic review. Psychooncology 2021 Jan;30(1):16-26. [doi: 10.1002/pon.5549] [Medline: 32915517]

XSL•FO
RenderX

20. Norman GJ, Zabinski MF, Adams MA, Rosenberg DE, Yaroch AL, Atienza AA. A review of eHealth interventions for physical activity and dietary behavior change. Am J Prev Med 2007 Oct;33(4):336-345 [FREE Full text] [doi: 10.1016/j.amepre.2007.05.007] [Medline: 17888860]
21. Shams F, Wong JS, Nikoo M, Outadi A, Moazen-Zadeh E, Kamel MM, et al. Understanding eHealth cognitive behavioral therapy targeting substance use: realist review. J Med Internet Res 2021 Jan 21;23(1):e20557 [FREE Full text] [doi: 10.2196/20557] [Medline: 33475520]
22. Coley N, Andre L, Hoevenaar-Blom MP, Ngandu T, Beishuizen C, Barbera M, et al. Factors predicting engagement of older adults with a coach-supported eHealth intervention promoting lifestyle change and associations between engagement and changes in cardiovascular and dementia risk: secondary analysis of an 18-month multinational randomized controlled trial. J Med Internet Res 2022 May 09;24(5):e32006 [FREE Full text] [doi: 10.2196/32006] [Medline: 35385395]
23. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
24. Hardiker NR, Grant MJ. Factors that influence public engagement with eHealth: a literature review. Int J Med Inform 2011 Jan;80(1):1-12 [FREE Full text] [doi: 10.1016/j.ijmedinf.2010.10.017] [Medline: 21112244]
25. Xie LF, Itzkovitz A, Roy-Fleming A, Da Costa D, Brazeau AS. Understanding self-guided web-based educational interventions for patients with chronic health conditions: systematic review of intervention features and adherence. J Med Internet Res 2020 Aug 13;22(8):e18355 [FREE Full text] [doi: 10.2196/18355] [Medline: 32788152]
26. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-health: a scoping review. Telemed J E Health 2022 Jun;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
27. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
28. Knowles SR, Mikocka-Walus A. Utilization and efficacy of internet-based eHealth technology in gastroenterology: a systematic review. Scand J Gastroenterol 2014 Apr;49(4):387-408. [doi: 10.3109/00365521.2013.865259] [Medline: 24494974]
29. Hammersley ML, Jones RA, Okely AD. Parent-focused childhood and adolescent overweight and obesity eHealth interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jul 21;18(7):e203 [FREE Full text] [doi: 10.2196/jmir.5893] [Medline: 27443862]
30. Robinson A, Oksuz U, Slight R, Slight S, Husband A. Digital and mobile technologies to promote physical health behavior change and provide psychological support for patients undergoing elective surgery: meta-ethnography and systematic review. JMIR Mhealth Uhealth 2020 Dec 01;8(12):e19237 [FREE Full text] [doi: 10.2196/19237] [Medline: 33258787]
31. Darley A, Coughlan B, Furlong E. People with cancer and their family caregivers' personal experience of using supportive eHealth technology: a narrative review. Eur J Oncol Nurs 2021 Oct;54:102030 [FREE Full text] [doi: 10.1016/j.ejon.2021.102030] [Medline: 34531122]
32. Cox A, Lucas G, Marcu A, Piano M, Grosvenor W, Mold F, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 09;19(1):e11 [FREE Full text] [doi: 10.2196/jmir.6575] [Medline: 28069561]
33. Lally RM, Kupzyk K, Gallo S, Berry D. Use of an unguided, web-based distress self-management program after breast cancer diagnosis: sub-analysis of CaringGuidance pilot study. J Med Internet Res 2020 Jul 06;22(7):e19734 [FREE Full text] [doi: 10.2196/19734] [Medline: 32628117]
34. Schover LR, Canada AL, Yuan Y, Sui D, Neese L, Jenkins R, et al. A randomized trial of internet-based versus traditional sexual counseling for couples after localized prostate cancer treatment. Cancer 2012 Jan 15;118(2):500-509 [FREE Full text] [doi: 10.1002/cncr.26308] [Medline: 21953578]
35. Psihogios AM, King-Dowling S, O'Hagan B, Darabos K, Maurer L, Young J, et al. Contextual predictors of engagement in a tailored mHealth intervention for adolescent and young adult cancer survivors. Ann Behav Med 2021 Nov 18;55(12):1220-1230 [FREE Full text] [doi: 10.1093/abm/kaab008] [Medline: 33674863]
36. van den Berg SW, Peters EJ, Kraaijeveld JF, Gielissen MF, Prins JB. Usage of a generic web-based self-management intervention for breast cancer survivors: substudy analysis of the BREATH trial. J Med Internet Res 2013 Aug 19;15(8):e170 [FREE Full text] [doi: 10.2196/jmir.2566] [Medline: 23958584]
37. King-Dowling S, Psihogios AM, Hill-Kayser C, Szalda D, O'Hagan B, Darabos K, et al. Acceptability and feasibility of survivorship care plans and an accompanying mobile health intervention for adolescent and young adult survivors of childhood cancer. Pediatr Blood Cancer 2021 Mar;68(3):e28884 [FREE Full text] [doi: 10.1002/pbc.28884] [Medline: 33416214]
38. Crafoord MT, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]

XSL•FO
RenderX

39. Ferrante JM, Lulla A, Williamson JD, Devine KA, Ohman-Strickland P, Bandera EV. Patterns of Fitbit use and activity levels among African American breast cancer survivors during an eHealth weight loss randomized controlled trial. Am J Health Promot 2022 Jan;36(1):94-105 [FREE Full text] [doi: 10.1177/08901171211036700] [Medline: 34344171]
40. Schrijvers J, Vanderhaegen J, Van Poppel H, Haustermans K, Van Audenhove C. How do patients between the age of 65 and 75 use a web-based decision aid for treatment choice in localized prostate cancer? J Evid Based Med 2013 Aug;6(3):167-172. [doi: 10.1111/jebm.12051] [Medline: 24325373]
41. Zhu H, Chen X, Yang J, Wu Q, Zhu J, Chan SW. Mobile breast cancer e-support program for Chinese women with breast cancer undergoing chemotherapy (part 3): secondary data analysis. JMIR Mhealth Uhealth 2020 Sep 16;8(9):e18896 [FREE Full text] [doi: 10.2196/18896] [Medline: 32936087]
42. Cuypers M, Lamers RE, Kil PJ, van de Poll-Franse LV, de Vries M. Impact of a web-based prostate cancer treatment decision aid on patient-reported decision process parameters: results from the Prostate Cancer Patient Centered Care trial. Support Care Cancer 2018 Nov;26(11):3739-3748 [FREE Full text] [doi: 10.1007/s00520-018-4236-8] [Medline: 29752528]
43. Høybye MT, Dalton SO, Deltour I, Bidstrup PE, Frederiksen K, Johansen C. Effect of internet peer-support groups on psychosocial adjustment to cancer: a randomised study. Br J Cancer 2010 Apr 27;102(9):1348-1354 [FREE Full text] [doi: 10.1038/sj.bjc.6605646] [Medline: 20424614]
44. Han JY, Wise M, Kim E, Pingree R, Hawkins RP, Pingree S, et al. Factors associated with use of interactive cancer communication system: an application of the comprehensive model of information seeking. J Comput Mediat Commun 2010 Apr;15(3):367-388 [FREE Full text] [doi: 10.1111/j.1083-6101.2010.01508.x] [Medline: 21760702]
45. Golsteijn RH, Bolman C, Peels DA, Volders E, de Vries H, Lechner L. A web-based and print-based computer-tailored physical activity intervention for prostate and colorectal cancer survivors: a comparison of user characteristics and intervention use. J Med Internet Res 2017 Aug 23;19(8):e298 [FREE Full text] [doi: 10.2196/jmir.7838] [Medline: 28835353]
46. Lozano-Lozano M, Melguizo-Rodríguez L, Fernández-Lao C, Galiano-Castillo N, Cantarero-Villanueva I, Martín-Martín L, et al. Association between the use of a mobile health strategy app and biological changes in breast cancer survivors: prospective pre-post study. J Med Internet Res 2019 Aug 14;21(8):e15062 [FREE Full text] [doi: 10.2196/15062] [Medline: 31414667]
47. Fang CY, Galloway TJ, Egleston BL, Bauman JR, Ebersole B, Chwistek M, et al. Development of a web-based supportive care program for patients with head and neck cancer. Front Oncol 2020;10:602202 [FREE Full text] [doi: 10.3389/fonc.2020.602202] [Medline: 33384959]
48. Webb J, Peel J, Fife-Schaw C, Ogden J. A mixed methods process evaluation of a print-based intervention supported by internet tools to improve physical activity in UK cancer survivors. Public Health 2019 Oct;175:19-27 [FREE Full text] [doi: 10.1016/j.puhe.2019.06.013] [Medline: 31374452]
49. Low CA, Danko M, Durica KC, Kunta AR, Mulukutla R, Ren Y, et al. A real-time mobile intervention to reduce sedentary behavior before and after cancer surgery: usability and feasibility study. JMIR Perioper Med 2020 Mar 23;3(1):e17292 [FREE Full text] [doi: 10.2196/17292] [Medline: 33393915]
50. Bruggeman Everts FZ, van der Lee ML, de Jager Meezenbroek E. Web-based individual mindfulness-based cognitive therapy for cancer-related fatigue — a pilot study. Internet Interv 2015 May;2(2):200-213. [doi: 10.1016/j.invent.2015.03.004]
51. Finlay A, Evans H, Vincent A, Wittert G, Vandelanotte C, Short CE. Optimising web-based computer-tailored physical activity interventions for prostate cancer survivors: a randomised controlled trial examining the impact of website architecture on user engagement. Int J Environ Res Public Health 2020 Oct 28;17(21):7920 [FREE Full text] [doi: 10.3390/ijerph17217920] [Medline: 33126692]
52. Ferrante JM, Devine KA, Bator A, Rodgers A, Ohman-Strickland PA, Bandera EV, et al. Feasibility and potential efficacy of commercial mHealth/eHealth tools for weight loss in African American breast cancer survivors: pilot randomized controlled trial. Transl Behav Med 2020 Oct 08;10(4):938-948 [FREE Full text] [doi: 10.1093/tbm/iby124] [Medline: 30535101]
53. Berry DL, Blonquist TM, Patel RA, Halpenny B, McReynolds J. Exposure to a patient-centered, web-based intervention for managing cancer symptom and quality of life issues: impact on symptom distress. J Med Internet Res 2015 Jun 03;17(6):e136 [FREE Full text] [doi: 10.2196/jmir.4190] [Medline: 26041682]
54. Casillas JN, Schwartz LF, Crespi CM, Ganz PA, Kahn KL, Stuber ML, et al. The use of mobile technology and peer navigation to promote adolescent and young adult (AYA) cancer survivorship care: results of a randomized controlled trial. J Cancer Surviv 2019 Aug;13(4):580-592 [FREE Full text] [doi: 10.1007/s11764-019-00777-7] [Medline: 31350681]
55. Dozeman E, Verdonck-de Leeuw IM, Savard J, van Straten A. Guided web-based intervention for insomnia targeting breast cancer patients: feasibility and effect. Internet Interv 2017 Sep;9:1-6 [FREE Full text] [doi: 10.1016/j.invent.2017.03.005] [Medline: 30135831]
56. Baik SH, Oswald LB, Buscemi J, Buitrago D, Iacobelli F, Perez-Tamayo A, et al. Patterns of use of smartphone-based interventions among Latina breast cancer survivors: secondary analysis of a pilot randomized controlled trial. JMIR Cancer 2020 Dec 08;6(2):e17538 [FREE Full text] [doi: 10.2196/17538] [Medline: 33289669]
57. Chambers SK, Ritterband LM, Thorndike F, Nielsen L, Aitken JF, Clutton S, et al. Web-delivered cognitive behavioral therapy for distressed cancer patients: randomized controlled trial. J Med Internet Res 2018 Jan 31;20(1):e42 [FREE Full text] [doi: 10.2196/jmir.8850] [Medline: 29386173]

XSL•FO
RenderX

58. Mikolasek M, Witt CM, Barth J. Adherence to a mindfulness and relaxation self-care app for cancer patients: mixed-methods feasibility study. JMIR Mhealth Uhealth 2018 Dec 06;6(12):e11271 [FREE Full text] [doi: 10.2196/11271] [Medline: 30522990]
59. Kanera IM, Willems RA, Bolman CA, Mesters I, Zambon V, Gijsen BC, et al. Use and appreciation of a tailored self-management eHealth intervention for early cancer survivors: process evaluation of a randomized controlled trial. J Med Internet Res 2016 Aug 23;18(8):e229 [FREE Full text] [doi: 10.2196/jmir.5975] [Medline: 27554525]
60. Børøsund E, Cvancarova M, Ekstedt M, Moore SM, Ruland CM. How user characteristics affect use patterns in web-based illness management support for patients with breast and prostate cancer. J Med Internet Res 2013 Mar 01;15(3):e34 [FREE Full text] [doi: 10.2196/jmir.2285] [Medline: 23454601]
61. Harris LN, Cleary EH, Stanton AL. Project connect online: user and visitor experiences of an internet-based intervention for women with breast cancer. Psychooncology 2015 Sep;24(9):1145-1151 [FREE Full text] [doi: 10.1002/pon.3734] [Medline: 25521661]
62. Ruland CM, Maffei RM, Børøsund E, Krahn A, Andersen T, Grimsbø GH. Evaluation of different features of an eHealth application for personalized illness management support: cancer patients' use and appraisal of usefulness. Int J Med Inform 2013 Jul;82(7):593-603. [doi: 10.1016/j.ijmedinf.2013.02.007] [Medline: 23507561]
63. Short CE, Rebar A, James EL, Duncan MJ, Courneya KS, Plotnikoff RC, et al. How do different delivery schedules of tailored web-based physical activity advice for breast cancer survivors influence intervention use and efficacy? J Cancer Surviv 2017 Feb;11(1):80-91. [doi: 10.1007/s11764-016-0565-0] [Medline: 27498099]
64. Schover LR, Yuan Y, Fellman BM, Odensky E, Lewis PE, Martinetti P. Efficacy trial of an internet-based intervention for cancer-related female sexual dysfunction. J Natl Compr Canc Netw 2013 Nov;11(11):1389-1397 [FREE Full text] [doi: 10.6004/jnccn.2013.0162] [Medline: 24225972]
65. Zhu J, Ebert L, Liu X, Wei D, Chan SWC. Mobile breast cancer e-support program for Chinese women with breast cancer undergoing chemotherapy (part 2): multicenter randomized controlled trial. JMIR Mhealth Uhealth 2018 Apr 30;6(4):e104 [FREE Full text] [doi: 10.2196/mhealth.9438] [Medline: 29712622]
66. Cox M, Basen-Engquist K, Carmack CL, Blalock J, Li Y, Murray J, et al. Comparison of internet and telephone interventions for weight loss among cancer survivors: randomized controlled trial and feasibility study. JMIR Cancer 2017 Sep 27;3(2):e16 [FREE Full text] [doi: 10.2196/cancer.7166] [Medline: 28954716]
67. Zhou ES, Recklitis CJ. Internet-delivered insomnia intervention improves sleep and quality of life for adolescent and young adult cancer survivors. Pediatr Blood Cancer 2020 Sep;67(9):e28506. [doi: 10.1002/pbc.28506] [Medline: 32568460]
68. Mihuta ME, Green HJ. The implementation of web-based cognitive rehabilitation in adult cancer survivors: examining participant engagement, attrition and treatment fidelity. Support Care Cancer 2018 Feb;26(2):499-506. [doi: 10.1007/s00520-017-3855-9] [Medline: 28866765]
69. Lally RM, Bellavia G, Gallo S, Kupzyk K, Helgeson V, Brooks C, et al. Feasibility and acceptance of the CaringGuidance web-based, distress self-management, psychoeducational program initiated within 12 weeks of breast cancer diagnosis. Psychooncology 2019 Apr;28(4):888-895. [doi: 10.1002/pon.5038] [Medline: 30803084]
70. Lally RM, Kupzyk K, Mills A, Gallo S, Meneses K. Effects of social constraints and web-based psychoeducation on cancer-related psychological adjustment early-after breast cancer diagnosis. J Psychosoc Oncol 2019;37(6):677-698. [doi: 10.1080/07347332.2018.1546787] [Medline: 31631813]
71. Børøsund E, Varsi C, Clark MM, Ehlers SL, Andrykowski MA, Sleveland HR, et al. Pilot testing an app-based stress management intervention for cancer survivors. Transl Behav Med 2020 Aug 07;10(3):770-780 [FREE Full text] [doi: 10.1093/tbm/ibz062] [Medline: 31330023]
72. Beatty L, Koczwara B, Wade T. Evaluating the efficacy of a self-guided web-based CBT intervention for reducing cancer-distress: a randomised controlled trial. Support Care Cancer 2016 Mar;24(3):1043-1051. [doi: 10.1007/s00520-015-2867-6] [Medline: 26248651]
73. Head BA, Keeney C, Studts JL, Khayat M, Bumpous J, Pfeifer M. Feasibility and acceptance of a telehealth intervention to promote symptom management during treatment for head and neck cancer. J Support Oncol 2011 Jan 01;9(1):e1-11 [FREE Full text] [doi: 10.1016/j.suponc.2010.12.006] [Medline: 21499540]
74. Benedict C, Lazard AJ, Smith SM, Agrawal A, Collins MK, Love B. User experiences, usability, and social presence of a peer-to-peer support app: survey of young adults affected by cancer. J Appl Commun Res 2021 Sep 06;49(5):497-514. [doi: 10.1080/00909882.2021.1971737]
75. Petrocchi S, Filipponi C, Montagna G, Bonollo M, Pagani O, Meani F. A breast cancer smartphone app to navigate the breast cancer journey: mixed methods study. JMIR Form Res 2021 May 10;5(5):e28668 [FREE Full text] [doi: 10.2196/28668] [Medline: 33970120]
76. Sivell S, Edwards A, Manstead AS, Reed MW, Caldon L, Collins K, et al. Increasing readiness to decide and strengthening behavioral intentions: evaluating the impact of a web-based patient decision aid for breast cancer treatment options (BresDex: www.bresdex.com). Patient Educ Couns 2012 Aug;88(2):209-217. [doi: 10.1016/j.pec.2012.03.012] [Medline: 22541508]
77. Zachariae R, Amidi A, Damholdt MF, Clausen CD, Dahlgaard J, Lord H, et al. Internet-delivered cognitive-behavioral therapy for insomnia in breast cancer survivors: a randomized controlled trial. J Natl Cancer Inst 2018 Aug 01;110(8):880-887 [FREE Full text] [doi: 10.1093/jnci/djx293] [Medline: 29471478]

XSL•FO
RenderX

78. Bisseling E, Cillessen L, Spinhoven P, Schellekens M, Compen F, van der Lee M, et al. Development of the therapeutic alliance and its association with internet-based mindfulness-based cognitive therapy for distressed cancer patients: secondary analysis of a multicenter randomized controlled trial. J Med Internet Res 2019 Oct 18;21(10):e14065 [FREE Full text] [doi: 10.2196/14065] [Medline: 31628791]
79. Kim HJ, Kim SM, Shin H, Jang JS, Kim YI, Han DH. A mobile game for patients with breast cancer for chemotherapy self-management and quality-of-life improvement: randomized controlled trial. J Med Internet Res 2018 Oct 29;20(10):e273 [FREE Full text] [doi: 10.2196/jmir.9559] [Medline: 30578205]
80. Manne SL, Topham N, D'Agostino TA, Myers Virtue S, Kirstein L, Brill K, et al. Acceptability and pilot efficacy trial of a web-based breast reconstruction decision support aid for women considering mastectomy. Psychooncology 2016 Dec;25(12):1424-1433 [FREE Full text] [doi: 10.1002/pon.3984] [Medline: 26383833]
81. Beatty L, Kemp E, Coll JR, Turner J, Butow P, Milne D, et al. Finding My Way: results of a multicentre RCT evaluating a web-based self-guided psychosocial intervention for newly diagnosed cancer survivors. Support Care Cancer 2019 Jul;27(7):2533-2544. [doi: 10.1007/s00520-018-4526-1] [Medline: 30411239]
82. Bricker JB, Watson NL, Heffner JL, Sullivan B, Mull K, Kwon D, et al. A smartphone app designed to help cancer patients stop smoking: results from a pilot randomized trial on feasibility, acceptability, and effectiveness. JMIR Form Res 2020 Jan 17;4(1):e16652 [FREE Full text] [doi: 10.2196/16652] [Medline: 31951215]
83. Yanez B, Oswald LB, Baik SH, Buitrago D, Iacobelli F, Perez-Tamayo A, et al. Brief culturally informed smartphone interventions decrease breast cancer symptom burden among Latina breast cancer survivors. Psychooncology 2020 Jan;29(1):195-203 [FREE Full text] [doi: 10.1002/pon.5281] [Medline: 31693265]
84. Duffecy J, Sanford S, Wagner L, Begale M, Nawacki E, Mohr DC. Project onward: an innovative e-health intervention for cancer survivors. Psychooncology 2013 Apr;22(4):947-951 [FREE Full text] [doi: 10.1002/pon.3075] [Medline: 22438297]
85. Dirkse D, Hadjistavropoulos HD, Alberts NA, Karin E, Schneider LH, Titov N, et al. Making internet-delivered cognitive behaviour therapy scalable for cancer survivors: a randomized non-inferiority trial of self-guided and technician-guided therapy. J Cancer Surviv 2020 Apr;14(2):211-225. [doi: 10.1007/s11764-019-00810-9] [Medline: 31853727]
86. Birkhoff SD, Cantrell MA, Moriarty H, Lustig R. The usability and acceptability of a patient-centered mobile health tracking app among a sample of adult radiation oncology patients. ANS Adv Nurs Sci 2018;41(3):243-259. [doi: 10.1097/ANS.0000000000000202] [Medline: 29474226]
87. Lynch SM, Stricker CT, Brown JC, Berardi JM, Vaughn D, Domchek S, et al. Evaluation of a web-based weight loss intervention in overweight cancer survivors aged 50 years and younger. Obes Sci Pract 2017 Mar;3(1):83-94 [FREE Full text] [doi: 10.1002/osp4.98] [Medline: 28392934]
88. Graetz I, Anderson JN, McKillop CN, Stepanski EJ, Paladino AJ, Tillmanns TD. Use of a web-based app to improve postoperative outcomes for patients receiving gynecological oncology care: a randomized controlled feasibility trial. Gynecol Oncol 2018 Aug;150(2):311-317. [doi: 10.1016/j.ygyno.2018.06.007] [Medline: 29903391]
89. Graetz I, McKillop CN, Stepanski E, Vidal GA, Anderson JN, Schwartzberg LS. Use of a web-based app to improve breast cancer symptom management and adherence for aromatase inhibitors: a randomized controlled feasibility trial. J Cancer Surviv 2018 Aug;12(4):431-440 [FREE Full text] [doi: 10.1007/s11764-018-0682-z] [Medline: 29492753]
90. Murphy KM, Burns J, Victorson D. Consider the source: examining attrition rates, response rates, and preliminary effects of ehHealth mindfulness messages and delivery framing in a randomized trial with young adult cancer survivors. J Adolesc Young Adult Oncol 2021 Jun;10(3):272-281 [FREE Full text] [doi: 10.1089/jayao.2020.0102] [Medline: 33347390]
91. Børøsund E, Cvancarova M, Moore SM, Ekstedt M, Ruland CM. Comparing effects in regular practice of e-communication and web-based self-management support among breast cancer patients: preliminary results from a randomized controlled trial. J Med Internet Res 2014 Dec 18;16(12):e295 [FREE Full text] [doi: 10.2196/jmir.3348] [Medline: 25525672]
92. Diefenbach MA, Mohamed NE, Butz BP, Bar-Chama N, Stock R, Cesaretti J, et al. Acceptability and preliminary feasibility of an internet/CD-ROM-based education and decision program for early-stage prostate cancer patients: randomized pilot study. J Med Internet Res 2012 Jan 13;14(1):e6 [FREE Full text] [doi: 10.2196/jmir.1891] [Medline: 22246148]
93. Bruggeman-Everts FZ, Wolvers MD, van de Schoot R, Vollenbroek-Hutten MM, Van der Lee ML. Effectiveness of two web-based interventions for chronic cancer-related fatigue compared to an active control condition: results of the "Fitter na kanker" randomized controlled trial. J Med Internet Res 2017 Oct 19;19(10):e336 [FREE Full text] [doi: 10.2196/jmir.7180] [Medline: 29051138]
94. Magnol M, Eleonore B, Claire R, Castagne B, Pugibet M, Lukas C, et al. Use of eHealth by patients with rheumatoid arthritis: observational, cross-sectional, multicenter study. J Med Internet Res 2021 Jan 29;23(1):e19998 [FREE Full text] [doi: 10.2196/19998] [Medline: 33512320]
95. Fang ML, Siden E, Korol A, Demestihas MA, Sixsmith J, Sixsmith A. A scoping review exploration of the intended and unintended consequences of eHealth on older people: a health equity impact assessment. Human Technol 2018 Nov 30;14(3):297-323. [doi: 10.17011/ht/urn.201811224835]
96. Runz-Jørgensen SM, Schiøtz ML, Christensen U. Perceived value of eHealth among people living with multimorbidity: a qualitative study. J Comorb 2017;7(1):96-111 [FREE Full text] [doi: 10.15256/joc.2017.7.98] [Medline: 29359124]

XSL•FO
RenderX

97. Christensen H, Reynolds J, Griffiths KM. The use of e-health applications for anxiety and depression in young people: challenges and solutions. Early Interv Psychiatry 2011 Feb;5 Suppl 1:58-62. [doi: 10.1111/j.1751-7893.2010.00242.x] [Medline: 21208393]
98. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
99. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
100. Chan M, Jiang Y, Lee CY, Ramachandran HJ, Teo JY, Seah CW, et al. Effectiveness of eHealth-based cognitive behavioural therapy on depression: a systematic review and meta-analysis. J Clin Nurs 2022 Nov 20;31(21-22):3021-3031. [doi: 10.1111/jocn.16212] [Medline: 35060252]
101. Kim SC, Shah DV, Namkoong K, McTavish FM, Gustafson DH. Predictors of online health information seeking among women with breast cancer: the role of social support perception and emotional well-being. J Comput Mediat Commun 2013 Jan;18(2):98-118 [FREE Full text] [doi: 10.1111/jcc4.12002] [Medline: 24634575]
102. Bouma G, Admiraal JM, de Vries EG, Schröder CP, Walenkamp AM, Reyners AK. Internet-based support programs to alleviate psychosocial and physical symptoms in cancer patients: a literature analysis. Crit Rev Oncol Hematol 2015 Jul;95(1):26-37. [doi: 10.1016/j.critrevonc.2015.01.011] [Medline: 25701515]
103. Fulford H, McSwiggan L, Kroll T, MacGillivray S. Exploring the use of information and communication technology by people with mood disorder: a systematic review and metasynthesis. JMIR Ment Health 2016 Jul 01;3(3):e30 [FREE Full text] [doi: 10.2196/mental.5966] [Medline: 27370327]
104. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019 Jul 10;7(7):e13817 [FREE Full text] [doi: 10.2196/13817] [Medline: 31293246]
105. Eom C, Shin DW, Kim SY, Yang HK, Jo HS, Kweon SS, et al. Impact of perceived social support on the mental health and health-related quality of life in cancer patients: results from a nationwide, multicenter survey in South Korea. Psychooncology 2013 Jun;22(6):1283-1290. [doi: 10.1002/pon.3133] [Medline: 22833521]
106. Applebaum AJ, Stein EM, Lord-Bessen J, Pessin H, Rosenfeld B, Breitbart W. Optimism, social support, and mental health outcomes in patients with advanced cancer. Psychooncology 2014 Mar 07;23(3):299-306 [FREE Full text] [doi: 10.1002/pon.3418] [Medline: 24123339]
107. Gallant MP. The influence of social support on chronic illness self-management: a review and directions for research. Health Educ Behav 2003 Apr 01;30(2):170-195. [doi: 10.1177/1090198102251030] [Medline: 12693522]
108. Gonzalez JM, Alegria M, Prihoda TJ. How do attitudes toward mental health treatment vary by age, gender, and ethnicity/race in young adults? J Community Psychol 2005 Sep;33(5):611-629. [doi: 10.1002/jcop.20071]
109. Kwon M, Lawn S, Kaine C. Understanding men's engagement and disengagement when seeking support for mental health. Am J Mens Health 2023 Mar 07;17(2):15579883231157971 [FREE Full text] [doi: 10.1177/15579883231157971] [Medline: 36880329]
110. Ghidei L, Simone MJ, Salow MJ, Zimmerman KM, Paquin AM, Skarf LM, et al. Aging, antiretrovirals, and adherence: a meta analysis of adherence among older HIV-infected individuals. Drugs Aging 2013 Oct 20;30(10):809-819 [FREE Full text] [doi: 10.1007/s40266-013-0107-7] [Medline: 23959913]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by T de Azevedo Cardoso; submitted 07.09.23; peer-reviewed by C Valle, M Svetlak; comments to author 30.11.23; revised version received 28.02.24; accepted 11.07.24; published 11.12.24.

Please cite as:
Montalescot L, Baussard L, Charbonnier E
Factors Associated With Digital Intervention Engagement and Adherence in Patients With Cancer: Systematic Review
J Med Internet Res 2024;26:e52542
URL: https://www.jmir.org/2024/1/e52542
doi:10.2196/52542
PMID:

©Lucile Montalescot, Louise Baussard, Elodie Charbonnier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Large Language Models and Empathy: Systematic Review

Vera Sorin[1], MD; Dana Brin[2,3], MD; Yiftach Barash[2,3,4], MD; Eli Konen[2,3], MD; Alexander Charney[5,6], MD; Girish Nadkarni[5,6], MD; Eyal Klang[5,6], MD

[1]Department of Radiology, Mayo Clinic, Rochester, MN, United States
[2]Department of Diagnostic Imaging, Sheba Medical Center, Ramat Gan, Israel
[3]The Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
[4]DeepVision Lab, Chaim Sheba Medical Center, Tel Hashomer, Israel
[5]Division of Data-Driven and Digital Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[6]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Vera Sorin, MD
Department of Radiology
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 5072842511
Email: verasrn@gmail.com

Abstract

Background: Empathy, a fundamental aspect of human interaction, is characterized as the ability to experience another being's emotions within oneself. In health care, empathy is a fundamental for health care professionals and patients' interaction. It is a unique quality to humans that large language models (LLMs) are believed to lack.

Objective: We aimed to review the literature on the capacity of LLMs in demonstrating empathy.

Methods: We conducted a literature search on MEDLINE, Google Scholar, PsyArXiv, medRxiv, and arXiv between December 2022 and February 2024. We included English-language full-length publications that evaluated empathy in LLMs' outputs. We excluded papers evaluating other topics related to emotional intelligence that were not specifically empathy. The included studies' results, including the LLMs used, performance in empathy tasks, and limitations of the models, along with studies' metadata were summarized.

Results: A total of 12 studies published in 2023 met the inclusion criteria. ChatGPT-3.5 (OpenAI) was evaluated in all studies, with 6 studies comparing it with other LLMs such GPT-4, LLaMA (Meta), and fine-tuned chatbots. Seven studies focused on empathy within a medical context. The studies reported LLMs to exhibit elements of empathy, including emotions recognition and emotional support in diverse contexts. Evaluation metric included automatic metrics such as Recall-Oriented Understudy for Gisting Evaluation and Bilingual Evaluation Understudy, and human subjective evaluation. Some studies compared performance on empathy with humans, while others compared between different models. In some cases, LLMs were observed to outperform humans in empathy-related tasks. For example, ChatGPT-3.5 was evaluated for its responses to patients' questions from social media, where ChatGPT's responses were preferred over those of humans in 78.6% of cases. Other studies used subjective readers' assigned scores. One study reported a mean empathy score of 1.84-1.9 (scale 0-2) for their fine-tuned LLM, while a different study evaluating ChatGPT-based chatbots reported a mean human rating of 3.43 out of 4 for empathetic responses. Other evaluations were based on the level of the emotional awareness scale, which was reported to be higher for ChatGPT-3.5 than for humans. Another study evaluated ChatGPT and GPT-4 on soft-skills questions in the United States Medical Licensing Examination, where GPT-4 answered 90% of questions correctly. Limitations were noted, including repetitive use of empathic phrases, difficulty following initial instructions, overly lengthy responses, sensitivity to prompts, and overall subjective evaluation metrics influenced by the evaluator's background.

Conclusions: LLMs exhibit elements of cognitive empathy, recognizing emotions and providing emotionally supportive responses in various contexts. Since social skills are an integral part of intelligence, these advancements bring LLMs closer to human-like interactions and expand their potential use in applications requiring emotional intelligence. However, there remains room for improvement in both the performance of these models and the evaluation strategies used for assessing soft skills.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52597) doi:10.2196/52597

KEYWORDS

empathy; LLMs; AI; ChatGPT; review methods; review methodology; systematic review; scoping; synthesis; foundation models; text-based; human interaction; emotional intelligence; objective metrics; human assessment; emotions; healthcare; cognitive; PRISMA

Introduction

Empathy, a fundamental aspect of human interaction, can be characterized as the ability to experience the emotions of another being within oneself. The origin of the word "empathy" dates back to the 1880s, when Theodore Lipps determined the word "einfuhlung" ("in-feeling") to describe the emotional appreciation of another's feelings [1]. Empathy involves recognition of others' feelings, the causes of these feelings, and the ability to participate in an emotional experience of an individual without becoming part of it [1].

Empathy is described as "the ability to see the world through someone else's eyes," having the ability to imagine what someone else is thinking and feeling in a given situation [2]. It is commonly understood to encompass cognitive and affective components: the ability to understand another's feelings (cognitive empathy) and to experience emotions in response to others (affective empathy) [1,3].

In health care, empathy has an important role in patient care, improving patient satisfaction and treatment adherence. Empathy allows health care professionals to understand the emotional and psychological states of patients, fostering better communication and trust [4].

Large language models (LLMs) have demonstrated remarkable capabilities across various tasks, including text summarization, question-answering, and text generation [5]. There are numerous studies on potential applications in health care, as an educational tool and as a support tool in clinical work [6,7]. These models are already being integrated into practice. For instance, Epic has integrated GPT4 in its electronic health record software [8,9].

While LLMs have the potential to improve and automate some medical tasks, there are significant limitations to these models and their integration [10,11]. Despite the promising natural language processing capabilities, these models make errors and their performance in clinical tasks is challenging to evaluate on a large scale [12]. Many studies thus rely on multiple-choice questions assessment, which do not reflect real-world clinical applications [13]. These models can introduce bias [14], and can be susceptible to cyberattacks [15]. Some studies that evaluated these models for medical tasks suggested that despite impressive capabilities, LLMs lack empathy, a quality that is unique to humans and is imperative in health care [16-19].

Recent studies discuss and evaluate LLMs performance in tasks related to emotional intelligence, theory of mind, and empathy [20-25]. Some evidence suggests that these models may show aspects of cognitive empathy, including emotions recognition and providing supportive responses [17,26-28]. Furthermore, commercial LLM-based applications are being developed to offer emotional support to patients [29]. Given these developments, the aim of our study was to systematically review the literature on the capacity of LLMs in demonstrating empathy.

Methods

We searched the literature on LLMs and empathy using MEDLINE, Google Scholar, PsyArXiv, medRxiv, and arXiv. Studies published between December 2022 and February 2024 were included. The search query was "((("large language models") OR (llms) OR (gpt) OR (chatgpt)) AND ((empathy) OR ("emotional awareness") OR ("emotional intelligence") OR (emotion)) OR (("social robots") OR ("artificial emotional intelligence") OR ("emotional artificial intelligence") OR ("emotional chatbots") OR ("affective computing") OR (HRI) OR ("Human robot interaction"))." We also searched the references lists of relevant studies, including some key studies from major medical journals, for any additional studies that may have been missed during the initial search.

The inclusion and exclusion criteria are detailed in Table 1. Two reviewers (VS and EK) independently performed the search and screened the titles and abstract of the articles resulting from the search. Differences in search results were resolved through discussion to reach a consensus. The reviewers then screened selected articles' full text for final inclusion. Ultimately, 12 publications were included in this review. The results of the included studies including the LLMs used, performance in empathy tasks, and limitations of the models, along with studies' publication details, authors, and other relevant information were systematically summarized in a table.

XSL•FO
RenderX

Table 1. Inclusion and exclusion criteria. This table outlines the inclusion and exclusion criteria applied to select studies for this review for evaluating empathy within large language models.

Criteria	Inclusion	Exclusion
Article type	Full-length original articles	Nonoriginal articles including but not limited to perspectives, opinions, and reviews
Language	English	Non-English
Focus of study	Articles that evaluated empathy within LLMs'[a] outputs	Studies focusing only on emotion recognition or theory of mind, without explicit empathy evaluation
Model	Only LLMs[a]	Any other NLP[b] algorithms

[a]LLM: large language model.

[b]NLP: natural language processing.

Results

Figure 1 presents a flow diagram of the screening and inclusion process. All 12 studies included in this review were published in 2023. Six studies compared ChatGPT-3.5 with other LLMs including GPT-4, versions of LLaMA, and fine-tuned chatbots. Six studies evaluated only ChatGPT-3.5. Seven studies evaluated empathy in ChatGPT in medical context. The results of the studies included are summarized in Table 2. This table provides a detailed summary of studies included in this review that evaluate aspects of empathy exhibited by large language models. The table outlines each study's objectives, the specific large language model used key findings from the evaluations, sample sizes, and the methods used to assess empathy. It also highlights whether the reviewers were blinded to whether responses came from large language models or humans. The limitations of the LLMs as detailed in the different studies are detailed in Table S1 in Multimedia Appendix 1.

Figure 1. Flow diagram of the inclusion process based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

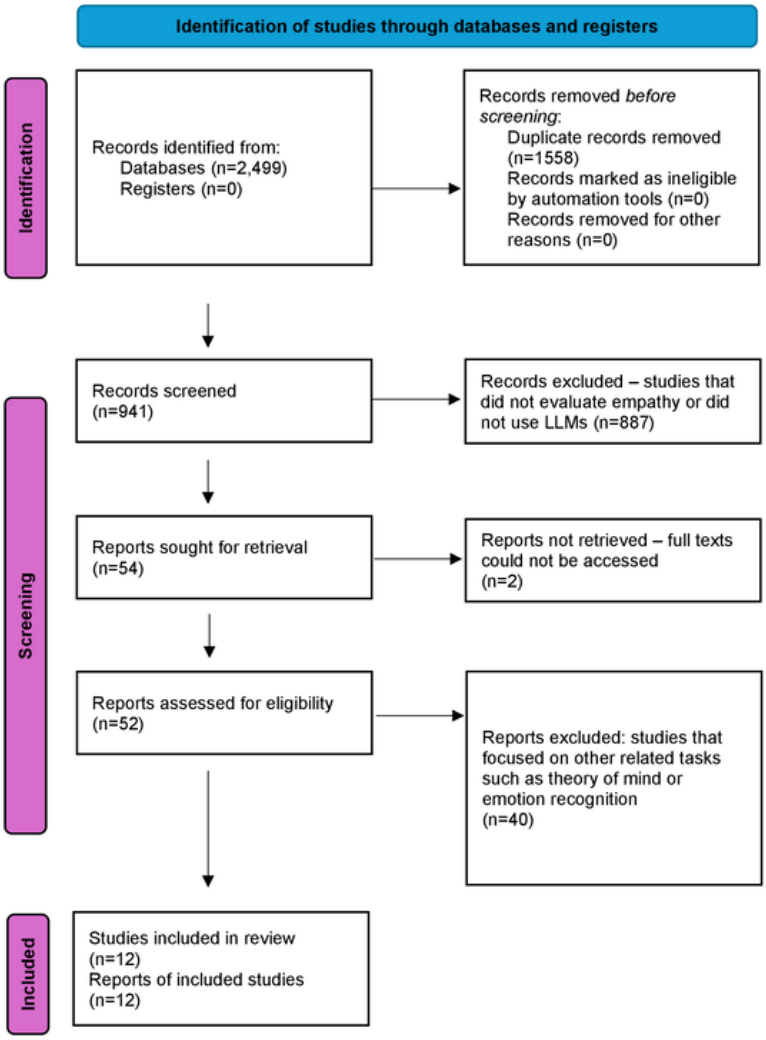

XSL•FO
RenderX

Table 2. Overview of studies evaluating empathy in large language models.

Study	Objective	LLM[a]	Key findings	Sample size	Methods for assessing empathy	Reviewers were blinded to LLM versus human responses
Webb [23]	Breaking bad news in emergency medicine	ChatGPT-3.5	ChatGPT facilitated realistic scenario design, active role-play, and effective feedback through the application of the SPIKES[b] framework for breaking bad news.	1 example	Simulating a patient role in an emergency department setting	No
Ayers et al [26]	Empathetic responses to patient questions	ChatGPT-3.5	ChatGPT responses were preferred by evaluators over physicians in 78.6% evaluations and were rated of significantly higher quality and empathy.	195 questions from social media	Health care professionals rated LLM and physician responses	Yes
Chen et al [27]	Simulating psychiatrists and patients in clinical psychiatric scenarios, and evaluating the expression of empathy in the interactions	Chatbots based on ChatGPT	ChatGPT-powered chatbots showed feasibility in simulating some aspects of empathy in psychiatric interactions, achieving a score of up to 3.43/4 when evaluated by humans for empathetic responses.	14 patients and 11 psychiatrists interacting with an LLM[a]	The participants interacted with the chatbots and scored their responses for empathy	No
Zhao et al [30]	Evaluate emotional dialogue understanding and generation and compare it with other supervised models	ChatGPT-3.5	Supervised models surpassed ChatGPT in emotion recognition. ChatGPT produced longer responses, but responses were also more specific to the context of the conversation compared with other models. When evaluating empathy within responses, humans preferred ChatGPT responses over EmpSOA in 54.33% of cases. When compared with MISC however, ChatGPT responses were preferred in 16% of cases.	100	3 readers rated the responses of different models for empathy	No
Yeo et al [31]	Emotional support for patients with cirrhosis and those with HCC[c]	ChatGPT-3.5	ChatGPT emulated empathetic responses and offered actionable recommendations for patients and caregivers.	4 prompts with 4 scenarios related to emotional support	The authors' subjective description and assessment	No
Elyoseph et al [28]	Emotional awareness performance compared with the general population norms	ChatGPT-3.5	ChatGPT demonstrated significantly higher emotional awareness performance than general population norms, with improvements over time. LEAS[d] scores were significantly higher than those of the general population (both men's and women's), on all the scales.	20 scenarios	LEAS compared with the general population norms. Two psychologists assessed the responses	No
Liu et al [32]	Fine-tuning an LLM to generate responses to patient questions	LLM based on LLaMA-65B; ChatGPT-3.5, GPT-4	GPT-4 and ChatGPT-3.5 outperformed the fine-tuned model. Both ChatGPT models and a fine-tuned LLaMA outperformed physician-generated responses.	10 questions	Physicians rated the responses of the chatbots and actual health care provider responses	Yes

XSL•FO
RenderX

Study	Objective	LLM[a]	Key findings	Sample size	Methods for assessing empathy	Reviewers were blinded to LLM versus human responses
Brin et al [33]	Evaluate ChatGPT and GPT-4 on USMLE[e] soft-skill questions	ChatGPT, GPT-4	GPT-4 correctly answered 90% of questions, outperforming ChatGPT and humans.	80 multiple-choice USMLE soft skills questions	Correctness of responses and comparison of ChatGPT and GPT-4 performance to that of past users from the AMBOSS question bank	No
Huang et al [34]	Evaluate emotional responses to various situations	Text-davinci-003 (a variant of GPT-3), GPT-3.5-turbo, GPT-4, LLaMA-2 7B and LLaMA-2 13B	The different LLMs generally demonstrate appropriate emotional responses. None of the models exhibit strong alignment with human references.	400 situations	Measuring the change in LLMs' different evoked emotions (overall 8 negative emotions) in response to situations compared with human benchmark	No
Chen et al [25]	Evaluate empathy, listening and comfort abilities of a fine-tuned LLM, compared with other LLMs	SoulChat, ChatGLM-6B, ChatGPT, MeChat	The fine-tuned model (SoulChat) outperformed the 3 other models in automatic metrics (ROUGE[f] and BLEU[g]), and based on human evaluation. The mean empathy score ranged between 1.84-1.90 (on a scale of 0-2), compared with 1.62-1.65 for ChatGPT.	10,000 samples for automatic evaluation and 100 samples for manual evaluation	Automatic evaluation tools were used, as well as manual rating by three experts in psychology	No
Belkhir and Sadat [35]	Evaluate whether prompt engineering and an external emotion classifier can improve ChatGPT's empathetic responses	ChatGPT	Prompt engineering and the use of an external emotion classifier improved ChatGPT performance, increasing accuracy for emotion labeling from 28.64% to 39.55%.	25,000 human dialogues	Labeling dialogues with emotion labels	No
Qian et al [36]	Evaluate the performance of LLMs in generating empathetic responses compared with other deep learning available models	GPT-3, GPT-3.5, ChatGPT	ChatGPT outperformed the other models in empathetic response generation, with a mean score of 4.64 (on a scale of 1-5).	100 dialogues for human evaluation	Automatic evaluation tools and three human raters	No

[a]LLMs: large language models.

[b]SPIKES: Setting up, Perception, Invitation, Knowledge, Emotions with Empathy, and Strategy or Summary.

[c]HCC: hepatocellular carcinoma.

[d]LEAS: Levels of Emotional Awareness Scale.

[e]USMLE: United States Medical Licensing Examination.

[f]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[g]BLEU: Bilingual Evaluation Understudy.

Empathy is essential in medicine, particularly when breaking bad news to patients. It allows physicians to deliver difficult information in a manner that respects the patient's emotions and perspective. Webb [23] used ChatGPT to simulate a role play for breaking bad news in the emergency department. The chatbot successfully set up a training scenario, role played as a patient and provided clear feedback through the application of the SPIKES (Setting up, Perception, Invitation, Knowledge, Emotions with Empathy, and Strategy or Summary) framework for breaking bad news [23]. In another study, Yeo et al [31] tested ChatGPT's ability to provide emotional support to patients diagnosed with hepatocellular carcinoma, and their caregivers. ChatGPT was able to acknowledge the likely emotional response of the patient to their diagnosis. Furthermore, the chatbot provided clear and actionable starting points for a newly diagnosed patient and offered motivational responses encouraging proactive steps. For caregivers, ChatGPT provided psychological and practical recommendations [31].

XSL•FO
RenderX

Ayers et al [26] compared the quality and empathy of responses given by ChatGPT and physicians with 195 randomly drawn patient questions from a social media forum. The study found that patients preferred the chatbot's responses over physician responses in 78.6% of cases. ChatGPT's responses were rated significantly higher for both quality and empathy, while physician responses were 41% less empathetic than the chatbot responses. The authors noted that ChatGPT tended to provide more lengthy responses, which could potentially be erroneously associated with greater empathy. They concluded that the chatbot may have potential in aiding drafting responses to patient questions [26].

Another study also assessed empathy in chatbot's responses to patient's questions. Liu et al [32] developed a model based on a pretrained LLaMA-65B and finetuned to generate physician-like responses that are professional and empathetic. They evaluated the model on 10 actual patient questions in primary care and compared the responses with those generated by ChatGPT-3.5 and GPT-4, rating them based on empathy, responsiveness, accuracy, and usefulness. When evaluating empathy, GPT-4 and ChatGPT-3.5 outperformed their model. Interestingly, all language models outperformed physician-generated responses significantly [32].

Understanding and addressing patients' emotions is fundamental in mental health. Chen et al [27] used ChatGPT-powered chatbots to simulate psychiatrists and patients in clinical psychiatric scenarios. The chatbots showed potential in simulating some aspects of empathy. However, they sometimes forgot initial instructions and repeated general empathy phrases too often. They also asked fewer in-depth questions about symptoms compared with physicians, potentially affecting their ability to fully understand the patient's condition. When simulating patients, the chatbots reported symptoms inaccurately [27].

The Levels of Emotional Awareness Scale (LEAS) is a psychological tool that assesses an individual's capacity to identify and describe emotions in themselves and others, a fundamental aspect of empathy [37]. Elyoseph et al [28] compared the LEAS score of ChatGPT to the general population norms. They found that ChatGPT demonstrated significantly higher emotional awareness performance. When repeating the test following 1 month interval, the chatbot's performance further improved, almost reaching the maximum possible LEAS score. The authors propose that ChatGPT could be helpful for cognitive training of people with emotional awareness impairment, as well as for psychiatric assessment support [28].

Zhao et al [30] compared ChatGPT with supervised models in terms of emotional dialogue understanding and generation. The tasks they assessed included emotion recognition, emotion cause recognition, dialog act classification, empathetic response generation, and emotional support conversation. The authors found that while supervised models surpassed ChatGPT in emotion recognition, ChatGPT produced longer, more diverse, and context-specific responses, especially when interacting with users in negative emotional states. Interestingly, Zhao et al [30] also observed a repetitive pattern in ChatGPT's empathy expressions, similar to the results described by Chen et al [27].

Brin et al [33] evaluated ChatGPT and GPT-4 on USMLE (United States Medical Licensing Examination) questions involving communication skills, ethics, empathy, and professionalism. They have used questions from the USMLE website and the AMBOSS question bank and compared the performance of the LLMs with the reported performance at the AMBOSS website. GPT-4 correctly answered 90% of questions, outperforming ChatGPT and humans [33]. Huang et al [34] evaluated emotional responses of 5 different LLMs to various situations designed to evoke emotions, The LLMs' responses were compared with human responses collected from 1266 participants worldwide. The authors reported for each model the changes in emotion scores relative to human benchmarks. They conclude that the different LLMs generally demonstrate appropriate emotional responses. However, none of the models exhibited strong alignment with human references. GPT-3.5-turbo demonstrated the highest alignment in the scores after imagining being in the situations. The 13B version of LLaMA-2 exhibited the strongest comprehension of human emotions.

Chen et al [25] constructed an empathetic conversation dataset of over 2 million samples and used it to fine-tune an LLM to provide empathetic responses. Their finetuned LLM outperformed other LLMs including ChatGPT in responses' coherence and relevancy, as well as empathy, helpfulness and safety.

Blekhir and Sadat [35] evaluated whether prompt engineering and external emotion classifier can enhance empathy in ChatGPT's responses. The study evaluated 2 versions of ChatGPT: 1 incorporating user emotions with an emotion classifier and another adapting to emotions without external tools. They evaluated these versions against the standard ChatGPT, demonstrating that tailored emotional responses significantly improve ChatGPT's empathetic capabilities [35].

Qian et al [36] evaluated ChatGPT compared with other deep-learning models trained for empathetic interactions. They also propose 3 improvement methods including semantically similar in-context learning, 2-stage interactive generation, and combination with knowledge base. These methods improved the quality of responses generated by ChatGPT, which outperformed other models evaluated [36].

Lee et al [38] used Chain-of-Empathy prompting to reason emotion and situational factors that may assist the model to infer the emotional experience. They evaluated GPT-3.5 and compared 4 unique prompts that used Chain-of-Empathy in generating empathetic responses to Reddit posts. The Chain-of-Empathy strategy resulted in improved the model's empathy expression [38].

Discussion

Principal Findings

This review shows that LLMs demonstrate aspects of cognitive empathy, including recognition of emotions, and generation of emotionally supportive responses. Most studies focused on LLMs' performance in medical contexts, assessing their ability to provide empathetic responses in clinical and nonclinical

XSL•FO
RenderX

scenarios. Notable, LLMs were reported in the majority of the studies to perform comparably or even surpass human responses in certain empathy-related tasks. The review also identifies limitations, including the subjective nature of empathy evaluation, the risk of overestimating empathy due to lengthier responses, and the models' inherent lack of emotional experience.

Empathy and Social Intelligence in LLMs

LLMs have shown impressive abilities in semantic understanding and logical reasoning [5]. The ability of LLMs to emulate empathy, especially cognitive empathy, mirrors the growing body of research demonstrating that artificial intelligence can replicate certain aspects of social intelligence. This review supports the idea that LLMs may demonstrate some abilities that resemble social intelligence. Theory of mind involves the understanding of others' thoughts and emotions, and predicting or explaining their behaviors based on these inferences. This concept is fundamental to social interactions, and it is a complex task, as it involves understanding not just the literal meaning of words in a conversation, but the underlying intentions, beliefs, and emotions [39]. Several studies evaluated LLMs on theory of mind tasks, with varied performance, depending on the tasks and the models used [20-22,39,40].

Cognitive Versus Affective Perspectives

The definition of empathy varies among researchers and practitioners in social sciences [1]. One of the debates is whether it is a cognitive or affective concept, and most definitions of empathy include both [1]. Cognitive empathy involves the ability to understand another's feelings, closely related to theory of mind [3]. Affective empathy relates to experiencing emotions in response to an emotional stimulus [1]. The ability of LLMs to demonstrate empathy in various fields as highlighted in this review, seems to align more with the cognitive aspect. It is nevertheless surprising that in some cases the LLM outperformed humans in empathy-related tasks.

Research suggests that cognitive and affective empathy are distinct. For instance, people with autism often struggle with cognitive empathy but have normal levels of affective empathy, while psychopathic individuals typically show the reverse pattern [3]. Neurological studies demonstrated distinct brain regions associated with each type of empathy, which further supports this notion [41,42]. It is worth questioning if demonstrating cognitive empathy alone is sufficient, or whether affective empathy is imperative for achieving human-like emotional intelligence.

Historically, empathy has been viewed as a uniquely human trait, with definitions focused on interactions between humans [1]. The complexity of empathy, influenced by personality, culture, and context, has led to ambiguous definitions of the term [1,20]. Empathy exhibited by AI fundamentally differs from human empathy because an algorithm does not engage in a human's emotional experience. Consequently, human-centric definitions of empathy may not apply to LLMs. This warrants a reevaluation of how empathy is measured. The question arises whether observable responses alone can be considered empathetic if they meet human expectations or preferences. If humans cannot distinguish between responses generated by humans and LLMs, or if they prefer AI-generated responses as demonstrated in the study by Ayers et al [26], perhaps emulating such empathy may be sufficient.

Implications in Health Care

Numerous studies support the remarkable performance of LLMs in clinical reasoning [6,7]. These models can be applied to enhance the medical care patients receive, while decreasing the workload of health care providers [43]. Yet, empathy is a key factor in patient care. Empathy in health care communication is linked to improved patient satisfaction, adherence to treatment plans, and better outcomes [4]. It allows for a more nuanced understanding of patients' emotional states and experiences, facilitating more compassionate and person-centered care. As such, the ability of LLMs to integrate empathy can significantly enhance the role of AI in health care, for both patients and health care providers.

Empathy in health care aligns more with cognitive rather than affective empathy, involving the ability to understand the pain and suffering of patients, and the capability to communicate this understanding [44,45]. Using this perspective, tools like the Jefferson Scale can assess empathy within health care settings [45,46]. There may be scenarios where LLMs might demonstrate more fitting empathy, especially in contexts where cognitive empathy is predominant. However, the lack of standardized methods for assessing empathy in LLMs, as also seen in this review, challenges the ability to compare their empathetic capacities across different models and tasks.

Furthermore, LLMs' empathy is influenced by cultural factors, norms, and contexts, affecting how empathy is perceived by individuals from diverse backgrounds. Awareness of interacting with an AI could bias perceptions of empathy, potentially undermining its authenticity [47]. Conversely, LLMs possess the potential to overcome cultural divides, offering empathetic responses appropriate across various backgrounds.

Limitations

This review has several limitations. First, as all but 1 study evaluated empathy based on subjective assessment, we could not perform a meta-analysis. Second, we only assessed studies directly discussing empathy, while there are many more that evaluate theory of mind tasks that are closely related to "cognitive" empathy. Third, all studies assessed ChatGPT-3.5, and only 1 study evaluated a model based on LLaMA and GPT-4. This can potentially limit the generalizability of findings to other LLMs. It is possible that alternative LLMs may present different empathy characteristics. Furthermore, LLMs are evolving fast, and possibly newer LLMs will present higher cognitive-like abilities.

Conclusion

To conclude, this review demonstrates that LLMs exhibit elements of cognitive empathy, being able to recognize emotions and provide emotionally supportive responses in various contexts. Given that social skills are foundational to the concept of "intelligence," further research is warranted to further develop

XSL•FO
RenderX

that aspect in AI. The ability to simulate empathetic responses could enhance patient experiences, improving patient satisfaction and adherence to treatment plans. However, there remain critical questions regarding whether LLMs' cognitive empathy is sufficient in scenarios that require deeper emotional engagement. Ultimately, as we continue to refine these models, we approach closer to bridging the gap between artificial and human-like interactions, opening opportunities for empathetic AI applications.

Data Availability

Data extracted from the studies are available in Table 2, and upon request from authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental Table 1. Limitation of Large Language Models Detailed in the Studies Included.
[DOCX File , 18 KB - jmir_v26i1e52597_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 82 KB - jmir_v26i1e52597_app2.pdf]

References

1. Cuff BMP, Brown SJ, Taylor L, Howat DJ. Empathy: A review of the concept. Emotion Review 2014;8(2):144-153. [doi: 10.1177/1754073914558466]
2. Hajibabaee F, A Farahani M, Ameri Z, Salehi T, Hosseini F. The relationship between empathy and emotional intelligence among Iranian nursing students. Int J Med Educ 2018;9:239-243 [FREE Full text] [doi: 10.5116/ijme.5b83.e2a5] [Medline: 30244237]
3. Blair RJR. Fine cuts of empathy and the amygdala: dissociable deficits in psychopathy and autism. Q J Exp Psychol (Hove) 2008;61(1):157-170. [doi: 10.1080/17470210701508855] [Medline: 18038346]
4. Moudatsou M, Stavropoulou A, Philalithis A, Koukouli S. The role of empathy in health and social care professionals. Healthcare (Basel) 2020;8(1):26 [FREE Full text] [doi: 10.3390/healthcare8010026] [Medline: 32019104]
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
6. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023;9(1):44 [FREE Full text] [doi: 10.1038/s41523-023-00557-8] [Medline: 37253791]
7. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023;20(10):998-1003. [doi: 10.1016/j.jacr.2023.06.009] [Medline: 37423350]
8. Cool stuff now: epic and generative AI.: Epic; 2023. URL: https://tinyurl.com/4xvcjbh5 [accessed 2024-08-13]
9. Microsoft and epic expand strategic collaboration with integration of azure openAI service.: Epic; 2023. URL: https://www.epic.com/epic/post/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/ [accessed 2024-08-17]
10. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141 [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
11. Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST, et al. Medical ethics of large language models in medicine. NEJM AI 2024;1(7). [doi: 10.1056/aira2400038]
12. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84 [FREE Full text] [doi: 10.1038/s41746-024-01083-y] [Medline: 38570554]
13. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health 2024;6(9):e662-e672 [FREE Full text] [doi: 10.1016/S2589-7500(24)00124-9] [Medline: 39179311]
14. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. medRxiv 2024:2024. [doi: 10.1101/2024.09.09.24313295]
15. Sorin V, Soffer S, Glicksberg BS, Barash Y, Konen E, Klang E. Adversarial attacks in radiology - A systematic review. Eur J Radiol 2023;167:111085. [doi: 10.1016/j.ejrad.2023.111085] [Medline: 37699278]

XSL•FO
RenderX

16. Ilicki J. A framework for critically assessing ChatGPT and other large language artificial intelligence model applications in health care. Mayo Clinic Proceedings: Digital Health 2023;1(2):185-188. [doi: 10.1016/j.mcpdig.2023.03.006]
17. Nashwan AJ, Abujaber AA, Choudry H. Embracing the future of physician-patient communication: GPT-4 in gastroenterology. Gastroenterology & Endoscopy 2023;1(3):132-135. [doi: 10.1016/j.gande.2023.07.004]
18. Sun YX, Li ZM, Huang JZ, Yu NZ, Long X. GPT-4: The future of cosmetic procedure consultation? Aesthet Surg J 2023;43(8):NP670-NP672. [doi: 10.1093/asj/sjad134] [Medline: 37154801]
19. Carlbring P, Hadjistavropoulos H, Kleiboer A, Andersson G. A new era in internet interventions: the advent of Chat-GPT and AI-assisted therapist guidance. Internet Interv 2023;32:100621 [FREE Full text] [doi: 10.1016/j.invent.2023.100621] [Medline: 37273936]
20. Kosinski M. Theory of mind may have spontaneously emerged in large language models. Stanford Graduate School of Business 2023 [FREE Full text]
21. Moghaddam SR, Honey CJ. Boosting theory-of-mind performance in large language models via prompting. ArXiv Preprint posted online on April 22, 2023 [FREE Full text]
22. Sap M, LeBras R, Fried D, Choi Y. Neural theory-of-mind? on the limits of social intelligence in large lms. arXiv:221013312 2022. [doi: 10.18653/v1/2022.emnlp-main.248]
23. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]
24. Weizenbaum J. Empathic AI can't get under the skin. Nat Mach Intell 2024;6(5):495. [doi: 10.1038/s42256-024-00850-6]
25. Chen Y, Xing X, Lin J, Zheng H, Wang Z, Liu Q, et al. Improving llms? empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. 2023 Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 December 10; Singapore. [doi: 10.18653/v1/2023.findings-emnlp.83]
26. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
27. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. ArXiv Preprint posted online on May 23, 2023 [FREE Full text]
28. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058 [FREE Full text] [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
29. LIV: pioneering AI in mental health through strategic partnerships.: Sheba Global; 2024. URL: https://sheba-global.com/transforming-mental-health-care-with-ai-liv-the-next-step-in-psychiatric-innovation/ [accessed 2024-07-11]
30. Zhao W, Zhao Y, Lu X, Wang S, Tong Y, Qin B. Is ChatGPT Equipped with Emotional Dialogue Capabilities? ArXiv Preprint posted online on April 19, 2023 [FREE Full text]
31. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
32. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages. medRxiv 2023 [FREE Full text] [doi: 10.1101/2023.07.14.23292669] [Medline: 37503263]
33. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
34. Huang JT, Lam MH, Li EJ, Ren S, Wang W, Jiao W, et al. Emotionally numb or empathetic? Evaluating how llms feel using emotionbench. ArXiv Preprint posted online on August 07, 2023 [FREE Full text]
35. Belkhir A, Sadat F. Beyond information: Is chatgpt empathetic enough? 2023 Presented at: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing; 2024 November 05; Bulgaria, Varna, Bulgaria,. [doi: 10.26615/978-954-452-092-2_018]
36. Qian Y, Zhang WN, Liu T. Harnessing the power of large language models for empathetic response generation: empirical investigations and improvements. arXiv:231005140 2023. [doi: 10.18653/v1/2023.findings-emnlp.433]
37. Lane RD, Smith R. Levels of emotional awareness: theory and measurement of a socio-emotional skill. J Intell 2021;9(3):42 [FREE Full text] [doi: 10.3390/jintelligence9030042] [Medline: 34449662]
38. Lee YK, Lee I, Shin M, Bae S, Hahn S. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. ArXiv Preprint posted online on November 02, 2023 [FREE Full text]
39. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. ArXiv Preprint posted online on March 22, 2023 [FREE Full text]
40. Marchetti A, Di Dio C, Cangelosi A, Manzi F, Massaro D. Developing ChatGPT's theory of mind. Front Robot AI 2023;10:1189525 [FREE Full text] [doi: 10.3389/frobt.2023.1189525] [Medline: 37377631]
41. Shamay-Tsoory SG, Aharon-Peretz J, Perry D. Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. Brain 2009;132(Pt 3):617-627. [doi: 10.1093/brain/awn279] [Medline: 18971202]

XSL•FO
RenderX

42. Zaki J, Weber J, Bolger N, Ochsner KN. The neural bases of empathic accuracy. Proc Natl Acad Sci U S A 2009;106(27):11382-11387 [FREE Full text] [doi: 10.1073/pnas.0902666106] [Medline: 19549849]
43. Sorin V, Barash Y, Konen E, Klang E. Large language models for oncological applications. J Cancer Res Clin Oncol 2023;149(11):9505-9508. [doi: 10.1007/s00432-023-04824-w] [Medline: 37160626]
44. Hojat M. Empathy in Health Professions Education and Patient Care. Cham: Springer; 2016.
45. Hojat M, Mangione S, Nasca TJ, Cohen MJM, Gonnella JS, Erdmann JB, et al. The jefferson scale of physician empathy: development and preliminary psychometric data. Educational and Psychological Measurement 2001;61(2):349-365. [doi: 10.1177/00131640121971158]
46. Hojat M, DeSantis J, Shannon SC, Mortensen LH, Speicher MR, Bragan L, et al. The jefferson scale of empathy: a nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students. Adv Health Sci Educ Theory Pract 2018;23(5):899-920 [FREE Full text] [doi: 10.1007/s10459-018-9839-9] [Medline: 29968006]
47. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 2018;20(6):e10148 [FREE Full text] [doi: 10.2196/10148] [Medline: 29945856]

Abbreviations

LEAS: Levels of Emotional Awareness Scale
LLM: large language model
SPIKES: Setting up, Perception, Invitation, Knowledge, Emotions with Empathy, and Strategy or Summary
USMLE: United States Medical Licensing Examination

Edited by A Mavragani; submitted 09.09.23; peer-reviewed by X Long, S Pandey, J Ilicki, A Tabaie; comments to author 01.02.24; revised version received 27.02.24; accepted 20.10.24; published 11.12.24.

Please cite as:
Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, Klang E
Large Language Models and Empathy: Systematic Review
J Med Internet Res 2024;26:e52597
URL: https://www.jmir.org/2024/1/e52597
doi:10.2196/52597
PMID:

©Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, Eyal Klang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Current Implementation of Digital Health in Chronic Disease Management: Scoping Review

Candelyn Pong[1*], BSocSci; Rachel Marjorie Wei Wen Tseng[2,3], BA; Yih Chung Tham[2,4,5], PhD; Elaine Lum[1,6*], PhD

[1]Health Services and Systems Research, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore
[2]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
[3]Duke-NUS Medical School, National University of Singapore, Singapore, Singapore
[4]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[5]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[6]Centre for Population Health Research and Implementation, SingHealth, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Elaine Lum, PhD
Health Services and Systems Research
Duke-NUS Medical School
National University of Singapore
8 College Rd
Singapore, 169857
Singapore
Phone: 65 66017213
Email: elaine_lum@duke-nus.edu.sg

Abstract

Background: Approximately 1 in 3 adults live with multiple chronic diseases. Digital health is being harnessed to improve continuity of care and management of chronic diseases. However, meaningful uptake of digital health for chronic disease management remains low. It is unclear how these innovations have been implemented and evaluated.

Objective: This scoping review aims to identify how digital health innovations for chronic disease management have been implemented and evaluated: what implementation frameworks, methods, and strategies were used; how successful these strategies were; key barriers and enablers to implementation; and lessons learned and recommendations shared by study authors.

Methods: We used the Joanna Briggs Institute methodology for scoping reviews. Five databases were searched for studies published between January 2015 and March 2023: PubMed, Scopus, CINAHL, PsycINFO, and IEEE Xplore. We included primary studies of any study design with any type of digital health innovations for chronic diseases that benefit patients, caregivers, or health care professionals. We extracted study characteristics; type of digital health innovation; implementation frameworks, strategies, and outcome measures used; barriers and enablers to implementation; lessons learned; and recommendations reported by study authors. We used established taxonomies to synthesize extracted data. Extracted barriers and enablers were grouped into categories for reporting. Descriptive statistics were used to consolidate extracted data.

Results: A total of 252 studies were included, comprising mainly mobile health (107/252, 42.5%), eHealth (61/252, 24.2%), and telehealth (97/252, 38.5%), with some studies involving more than 1 innovation. Only 23 studies (23/252, 9.1%) reported using an implementation science theory, model, or framework; the most common were implementation theories, classic theories, and determinant frameworks, with 7 studies each. Of 252 studies, 144 (57.1%) used 2 to 5 implementation strategies. Frequently used strategies were “obtain and use patient or consumer feedback” (196/252, 77.8%); “audit and provide feedback” (106/252, 42.1%); and piloting before implementation or “stage implementation scale-up” (85/252, 33.7%). Commonly measured implementation outcomes were acceptability, feasibility, and adoption of the digital innovation. Of 252 studies, 247 studies (98%) did not measure service outcomes, while patient health outcomes were measured in 89 studies (35.3%). The main method used to assess outcomes was surveys (173/252, 68.7%), followed by interviews (95/252, 37.7%). Key barriers impacting implementation were data privacy concerns and patient preference for in-person consultations. Key enablers were training for health care workers and personalization of digital health features to patient needs.

XSL•FO
RenderX

Conclusions: This review generated a summary of how digital health in chronic disease management is currently implemented and evaluated and serves as a useful resource for clinicians, researchers, health system managers, and policy makers planning real-world implementation. Future studies should investigate whether using implementation science frameworks, including how well they are used, would yield better outcomes compared to not using them.

(J Med Internet Res 2024;26:e53576) doi:10.2196/53576

KEYWORDS

digital health; telemedicine; chronic disease; noncommunicable disease; implementation science; evidence-based practice; mobile phone

Introduction

Background

Chronic diseases are among the biggest health threats, causing approximately 41 million deaths yearly [1]. Chronic disease management is especially difficult because these diseases do not occur in isolation; approximately 1 in 3 adults have multiple chronic conditions [2], and the probability of developing multiple comorbidities increases with age [3].

Digital health is a rapidly growing industry that could potentially enhance health outcomes, with its growth accelerated by the COVID-19 pandemic. The use of digital health enabled continuity of care during the COVID-19 pandemic, especially for chronic disease management, when in-person care was limited [4,5]. An estimated 20% or US $1.8 trillion of the world's health care spending is wasteful, and digital health has the potential to lower health care spending by tackling this waste [6]. Despite its benefits, meaningful uptake of digital health for chronic disease management has been relatively low [7]. Some known barriers to adoption include lack of access and availability of digital health [8], poor user interface, suboptimal clinical integration, and lack of transparency about the datasets used for the development of digital health tools [9].

The uptake of health care innovations can be addressed using implementation science. Implementation science is defined as the scientific investigation of ways to increase the adoption of evidence-based practices and research findings into routine clinical use [10]. It is underpinned by theoretical frameworks collectively known as implementation science theories, models, and frameworks (henceforth, for simplicity, implementation science frameworks). Nilsen's taxonomy of implementation science frameworks shows 5 categories (Textbox 1): process models, determinant frameworks, classic theories, implementation theories, and evaluation frameworks [11].

Textbox 1. Descriptions and examples of theories, models, and frameworks used in implementation science.

- Process models: These models are used to detail and facilitate the process of bridging the gap between research and routine clinical practice. Examples are the Knowledge To Action framework [12] and the Ottawa Model [13].
- Determinant frameworks: These are frameworks that provide a list of factors for analysis that could affect implementation outcomes. Examples are Theoretical Domains Framework [14] and Consolidated Framework for Implementation Research [15].
- Classic theories: These are theories used to analyze the factors that shape implementation outcomes, usually drawn from other disciplines, such as psychology and organizational theory [11]. Examples are the Theory of Planned Behavior [16], Social Cognitive Theory [17], and Situated Change Theory [18].
- Implementation theories: These are theories generated from implementation research to understand and address elements that shape implementation outcomes. An example is Normalization Process Theory [19].
- Evaluation frameworks: These are frameworks to guide the evaluation of implementation efforts. Examples are the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) framework [20] and the framework for Outcomes in Implementation Research [21].

Prospective use of implementation science frameworks can aid the development, implementation, and evaluation of digital health innovations [11]. Retrospective application of implementation science frameworks, although less common, can be used to understand an innovation's success or failure [22]. Given the plethora of implementation science frameworks available, the choice of framework largely depends on the study aim. For example, if the study aims to prospectively explore potential barriers and enablers to the implementation of an innovation, a determinant framework, such as the Consolidated Framework for Implementation Research, would be appropriate. The advantage of underpinning real-world implementation efforts with an appropriate implementation science framework is the ability to generate reliable translation, spread, and scale-up for evidence-based innovations [23].

Objectives

Despite the usefulness of implementation science, there is a scarcity of reviews on the implementation [24] and evaluation of digital health innovations for chronic disease management in real-world or clinical settings. Most reviews are relatively narrow, focusing on a specific element of implementation such as barriers, a specific chronic disease, or a specific type of digital health innovation [25-28]. In light of the current knowledge gaps, this scoping review aimed to identify how digital health innovations for the management of chronic diseases have been implemented. Specifically, to understand what implementation frameworks, methods, and strategies were used; how successful these strategies were; what were the key barriers and enablers to implementation; what lessons were learned; and what recommendations were shared by the respective study authors.

XSL•FO
RenderX

Findings will be useful to clinicians and researchers planning to implement digital health innovations.

Methods

Definitions

For the purposes of this scoping review, we defined digital health as the branch of study on the advancement and use of IT to enhance health [29]. This includes eHealth, mobile health (mHealth), telehealth, wearables, and artificial intelligence.

eHealth refers to the provision of health care services with the support of information and communication technology (ICT), for example, computers and phones; mHealth is defined as the use of smart and portable or mobile wireless devices in health care; and telehealth is defined as the use of digital technologies to provide health services remotely [29,30]. Wearables are devices worn by individuals, usually to monitor personal health metrics and their environment [31]. Artificial intelligence in health care refers to the use of machine learning, including natural language processing and deep learning, to make predictions of health outcomes or support clinical decision-making [32].

Search Strategy

This study was carried out according to the Joanna Briggs Institute methodology for scoping reviews [33]. The protocol was published on Open Science Framework [34,35] and briefly detailed in this paper. We developed a search strategy comprising the following key concepts: digital health, implementation, and chronic diseases, and refined it using the PRESS (Peer Review of Electronic Search Strategies) guidelines (Table S1 in Multimedia Appendix 1 [36-198]) [199]. The full search strategy is provided in Multimedia Appendix 1.

We searched the following 5 databases, PubMed, Scopus, CINAHL, PsycINFO, and IEEE Xplore, for literature published between January 2015 and March 2023. We chose to start from 2015 as there was a steep escalation in the number of published studies on digital health innovations from 2016 onward [200].

Eligibility Criteria

We included primary studies of any study design reporting on the preimplementation or implementation of any type of digital health innovations for chronic diseases that benefits either patients, caregivers, or health care professionals directly, with or without the use of an implementation science framework.

Studies that did not include implementation or were not reported in the English language, meta-analyses, systematic reviews, conference proceedings, short reports, study protocols, commentaries, and dissertations were excluded from this review.

Selection of Studies

Three researchers (CP, RMWWT, and EL) conducted title and abstract screening followed by full-text screening, with each study independently screened by 2 researchers. Conflicts at both stages of screening were resolved through discussion, and any unresolved conflicts were mediated by a third researcher. The reference lists of studies that partially met the inclusion criteria but were excluded because these were meta-analyses, systematic reviews, or study protocols were examined by 1 researcher (CP) to identify any additional relevant studies. Covidence (Veritas Health Innovation), a web-based collaboration platform for reviews, and Endnote (version 20; Clarivate Analytics) were used for screening and managing citations, respectively.

Data Extraction and Data Analysis

A standardized form was developed for data extraction using Google Forms (Google LLC). The following data were extracted: publication year, author, country of study, type of study, characteristics of the digital health innovations, definitions of digital health used by study authors, implementation frameworks used, implementation strategies, and outcome measures used to evaluate the implementation. We also extracted the key barriers and enablers for successful implementation, lessons learned, and recommendations shared by the respective study authors. How we operationalized data extraction is presented in Multimedia Appendix 1. For example, textual data points, such as “key barriers,” were summarized as “patients’ lack of motivation and time,” “increased workload for health care workers,” and so forth.

The extraction form was piloted by 2 researchers (CP and RMWWT) using 5 included studies and subsequently refined. One reviewer (CP) completed data extraction for the remaining studies. Data extraction of a random 10% (25/252, 9.9%) of included articles was verified by a second researcher (EL) to ensure rigor and trustworthiness. Descriptive statistics were used to consolidate the extracted data in Excel (version 1808; Microsoft Corporation).

We used the following taxonomies to inform our analysis and summary of extracted data: the categorization for countries and regions by the World Health Organization for the country of study [201]; Nilsen’s taxonomy for implementation theories, models, and frameworks [11]; Expert Recommendations for Implementing Change (ERIC) taxonomy for implementation strategies [202]; and Proctor’s outcomes in implementation research [21]. Logic models and pathways were not considered an implementation science framework in this review.

A Note About Outcomes

Overview

Outcomes were grouped into 3 categories, namely, implementation outcomes, service outcomes, and patient outcomes, following Proctor et al [21]. Implementation outcomes indicate the success (or otherwise) of implementing or embedding the digital health innovation. Service outcomes and patient outcomes indicate the effectiveness of digital health innovation in impacting service delivery or patient care and patient health or well-being, respectively. These 3 categories of outcomes are detailed in the *Implementation Outcomes* and *Service Outcomes and Patient Outcomes* sections.

Implementation Outcomes

Proctor’s outcomes for implementation research comprise 8 types of implementation outcomes, namely, acceptability, adoption, appropriateness, cost, feasibility, fidelity, penetration, and sustainability [21]. Acceptability is the impression among stakeholders that the specific innovation is agreeable, adoption

XSL•FO
RenderX

is the initial desire to test or use a given innovation, and appropriateness is the perceived suitability of the innovation in a particular context. Cost refers to both the cost of the innovation and the cost of implementation. Feasibility is the extent to which an innovation can be effectively implemented in a specific context, and fidelity is the extent to which the innovation was implemented as intended. Penetration is the integration of the innovation into health care services, and sustainability is the degree to which an innovation and its ensuing benefits can be effectively maintained in a specific context [21].

In Proctor's outcomes for implementation research, various types of stakeholders are recognized; for example, administrators, payers, health care providers, and patients or consumers and their family members, to name a few [21]. Hence, an implementation outcome, such as "acceptability," would hold different salience for each type of stakeholder [21]. For clarity, in this review, we chose to foreground the perspectives of target users for the implementation outcome "acceptability."

Hence, we can group the 8 outcomes proposed by Proctor et al [21] into the following three groups: (1) outcomes from a user perspective that are a function of innovation design (acceptability, adoption, appropriateness, and feasibility); (2) the implementation process (fidelity); and (3) outcomes that foreshadow embedment in routine practice from an organizational perspective (cost and cost-effectiveness, penetration, and sustainability).

Service Outcomes and Patient Outcomes

Service outcomes and patient outcomes indicate the effectiveness of digital health innovation in impacting service delivery or patient care and patient health or well-being, respectively. Examples of these outcomes include patient safety indicators, quantifiable health outcomes, patient satisfaction, health-related quality of life, patient empowerment, and patient knowledge.

Results

Search Yield

The search generated 7970 studies. After removing duplicates, 96.16% (7664/7970) studies remained. After title and abstract screening and full-text screening of 7664 and 754 studies, respectively, 3.22% (247/7664) studies remained. Manual examination of reference lists of 5 study protocols and 9 meta-analyses and systematic reviews excluded during screening yielded 5 additional studies, bringing the total number of included studies to 252 (Figure 1). The list of included studies is provided in Multimedia Appendix 1.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Review) flow diagram.

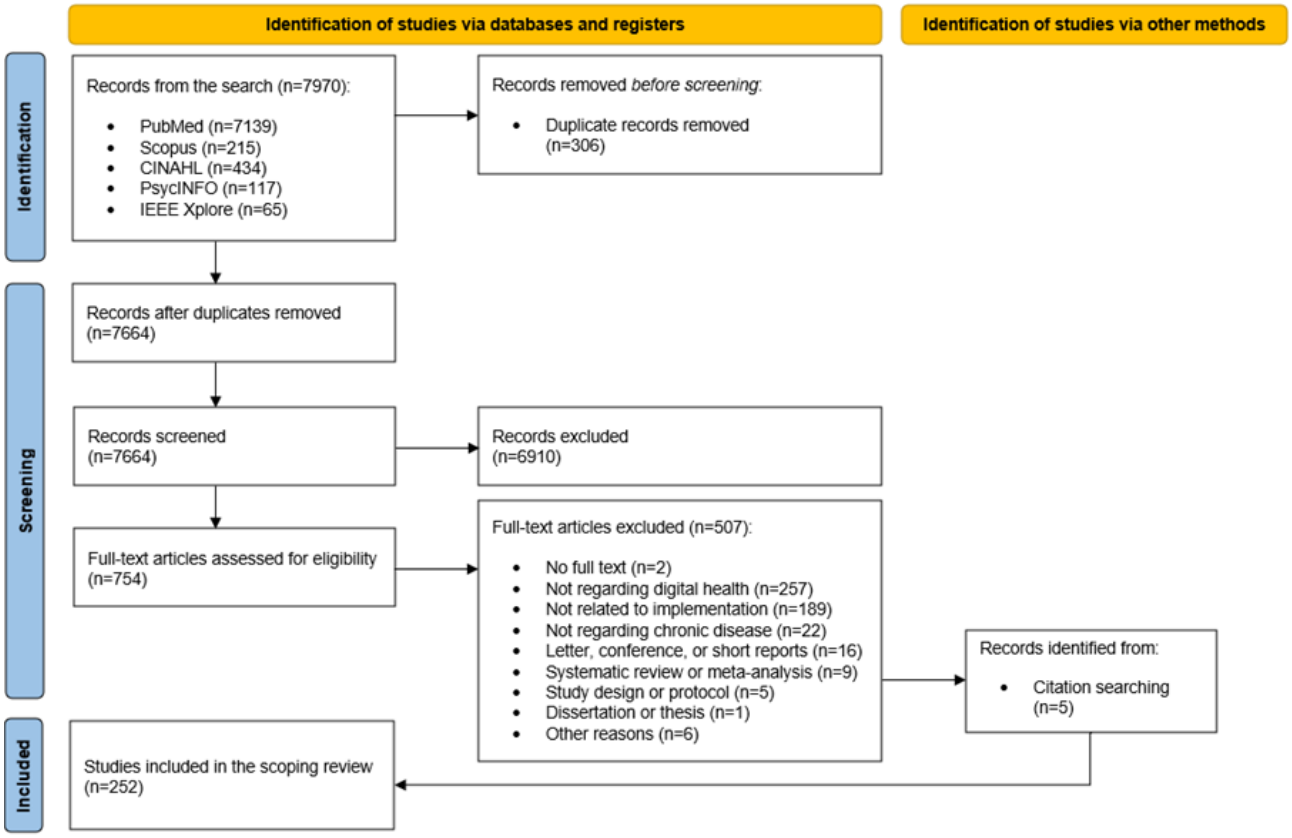

XSL•FO
RenderX

The most common reason for excluding full-text articles was that the innovation used was outside our scope and definition of digital health (257/507, 50.7%). Some examples include robot-assisted therapies and surgeries, minimally invasive and noninvasive medical devices, and online health education classes. Reasons for excluding full-text articles where digital health innovations were used include implementation outcomes not reported (189/507, 37.3%), disease type out of scope (22/507, 4.3%), conference proceedings or short report (16/507, 3.2%), systematic review or meta-analysis (9/507, 1.8%), study protocol (5/507, 1.0%), and dissertation (1/507, 0.2%). Other reasons for exclusion (6/507, 1.2%) were papers describing digital health implementation efforts in general and the creation of causal loop diagrams from published literature. Finally, records for which full text was not available upon reasonable search, for example, Google Scholar (Google LLC), university libraries, and contacting study authors via email or ResearchGate with no reply within 2 weeks, were excluded (2/507, 0.4%).

Study Characteristics

Key characteristics of the included studies are presented in Table 1. The majority were single-country studies (243/252, 96.4%). Most of the studies originated from the Region of the Americas (106/252, 42.1%), the European Region (89/252, 35.3%), and the Western Pacific Region (35/252, 13.9%). About half of the included studies were published between 2015 and 2019 before the emergence of the COVID-19 pandemic (123/252, 48.8%), with the other half (129/252, 51.2%) published between 2020 and 2023, during and post–COVID-19 pandemic. The 3 most common types of chronic diseases managed were chronic respiratory diseases (64/252, 25.4%), cardiovascular diseases (33/252, 13.1%), and neurological disorders (29/252, 11.5%).

The top 3 digital health innovations implemented were mHealth (107/252, 42.5%), telehealth (97/252, 38.5%), and eHealth (61/252, 24.2%). Most innovations were intended for patients (241/252, 95.6%). The top 3 study designs were mixed methods (57/252, 22.6%), qualitative description studies (54/252, 21.4%), and randomized controlled trials (47/252, 18.7%; Multimedia Appendix 1). The sample size varied widely, ranging from 11 to as large as 23,282 (Table 2). Of the 252 studies, 82 (32.5%) reported the study's duration, which ranged from 0.5 to 106 months, with a mean of 18.1 (SD 16.8) months.

XSL•FO
RenderX

Table 1. Characteristics of included studies.

Study characteristics	Papers, n (%)
Single-country study[a]	243 (96.4)
Region of the Americas	106 (42.1)
European Region	89 (35.3)
Western Pacific Region	34 (13.5)
South-East Asia Region	9 (3.6)
Eastern Mediterranean Region	3 (1.2)
African Region	1 (0.4)
Unclassified geographical region: Taiwan	1 (0.4)
Multicountry study[a]	9 (3.6)
European Region	6 (2.4)
European Region and Region of the Americas	1 (0.4)
European Region, Region of the Americas, and Western Pacific Region	1 (0.4)
European Region and Western Pacific Region	1 (0.4)
Publication year	
Before COVID-19 pandemic, 2015-2019	123 (48.8)
After COVID-19 pandemic, 2020-2023	129 (51.2)
Type of chronic disease[b]	
Chronic respiratory diseases	64 (25.4)
Cardiovascular diseases	33 (13.1)
Neurological disorders	29 (11.5)
Diabetes	26 (10.3)
Chronic diseases in general	20 (7.9)
Mental illnesses	20 (7.9)
Chronic musculoskeletal diseases	19 (7.5)
Chronic kidney diseases	14 (5.6)
Hypertension	13 (5.2)
Cancer	12 (4.8)
HIV or AIDS	6 (2.4)
Rheumatic diseases	6 (2.4)
Chronic hematologic disorders	5 (2)
Chronic liver diseases	5 (2)
Chronic skin diseases	5 (2)
Chronic gastrointestinal diseases	4 (1.6)
Obesity	3 (1.2)
Eye diseases	2 (1.2)
Ear diseases	1 (0.4)
Type of digital health innovation[c]	
mHealth[d]	107 (42.5)
Telehealth or telemedicine	97 (38.5)
eHealth	61 (24.2)
Wearables	17 (6.7)

XSL•FO
RenderX

Study characteristics	Papers, n (%)
Big data or deep learning or machine learning	4 (1.6)
Digital health in general	1 (0.4)
Target user of the innovation[e]	
Patients	241 (95.6)
Physicians	58 (23)
Nurses	46 (18.3)
Allied health professionals	26 (10.3)
Caregivers	25 (9.9)
Health care professionals in general[f]	10 (4)
Pharmacists	5 (2)
Ancillary or support staff	4 (1.6)
Midwives	1 (0.4)
Framework used	
No framework	216 (85.7)
Implementation science frameworks[g,h]	23 (9.1)
Implementation theories	7 (2.8)
Classic theories	7 (2.8)
Determinant frameworks	7 (2.8)
Evaluation frameworks	3 (1.2)
Process models	1 (0.4)
Other frameworks	13 (5.2)
Health frameworks	7 (2.8)
Technology adoption frameworks	3 (1.2)
Frameworks developed by medical and health organizations	2 (0.8)
Education frameworks	1 (0.4)
Number of implementation strategies	
0	5 (2)
1	103 (40.9)
2	102 (40.5)
3	40 (15.9)
4	1 (0.4)
5	1 (0.4)
Types of outcome[i]	
Implementation outcomes[j]	
Acceptability	171 (67.9)
Feasibility	38 (15.1)
Adoption	32 (12.7)
Cost[k]	6 (2.4)
Fidelity	4 (1.6)
Appropriateness	3 (1.2)
Sustainability	3 (1.2)

XSL•FO
RenderX

Study characteristics	Papers, n (%)
Penetration	1 (0.4)
Patient outcomes	
Health outcomes	81 (32.1)
Satisfaction	40 (15.9)
Quality of life	24 (9.5)
Patient empowerment	3 (1.2)
Patient knowledge	2 (0.8)
Service outcomes	
Safety	5 (2)
Assessment method used to evaluate outcome[l]	
Surveys	173 (68.7)
Interviews	95 (37.7)
Observations	39 (15.5)
Focus group discussions	18 (7.1)
Think-aloud protocol	7 (2.8)

[a]Countries are categorized based on the World Health Organization's country groupings [201].
[b]Percentages do not add up to 100% as some studies addressed >1 chronic disease.
[c]Percentages do not add up to 100% as some studies discussed >1 innovation.
[d]mHealth: mobile health.
[e]Percentages do not add up to 100% as some studies have >1 target user.
[f]The type of health care professionals is not specified.
[g] Some studies used >1 implementation science framework.
[h]Adapted from Nilsen taxonomy for implementation theories, models, and frameworks [11].
[i]Percentages do not add up to 100% as some studies have >1 outcome.
[j]Adapted from Proctor outcomes in implementation research [21].
[k]Most studies only reported the cost of innovation; only Raeside et al [36] reported the total cost of implementation.
[l]Percentages do not add up to 100% as some studies used >1 assessment method to evaluate the outcome.

XSL•FO
RenderX

Table 2. Sample size of included studies.

Innovation	Sample size, n
Single innovation	
Telehealth or telemedicine	23,282
mHealth[a]	9370
eHealth	5751
Wearables	232
Digital health in general	32
Big data or deep learning or machine learning	11
Multiple innovations	
eHealth and mHealth	9783
eHealth, mHealth, and wearables	457
mHealth and big data or deep learning or machine learning	385
mHealth and wearables	374
mHealth and telehealth or telemedicine	198
eHealth and telehealth or telemedicine	158
eHealth, mHealth, and big data or deep learning or machine learning	120
Telehealth or telemedicine, mHealth, and wearables	93
Telehealth or telemedicine, mHealth, eHealth, and big data or deep learning or machine learning	60
Telehealth or telemedicine and wearables	15

[a]mHealth: mobile health.

Definition of Digital Health

Of the 252 studies, only 24 (9.5%) included a definition of digital health. Of these 24 studies, a general definition of digital health was provided by 3 studies (12.5%): the use of ICTs for health care that includes both eHealth and mHealth [37] or as incorporating disruptive and medical technologies [38,39]. Definitions of a specific type of digital health innovation were provided by the remaining studies (21/24, 88%), of which 9 (43%) studies defined mHealth, 9 (43%) telehealth, 2 (10%) telemonitoring, and 1 (5%) defined eHealth. mHealth was defined as the use of mobile technology for services related to health care [40-48], and telehealth was defined as the use of ICT for internet-based provision of health care [49-57]. Of the 2 studies on telemonitoring, 1 (50%) described telemonitoring as an automated system [58], and the other (50%) defined it as a noninvasive patient-monitoring system [59] that uses ICT for the dissemination of patients' clinical data from their homes to their respective health care providers [58,59]. The sole study with a definition on eHealth defined it as the use of ICT for health care [60].

Type of Framework

Frameworks used to implement digital health innovations are presented in Table 1. Most studies (216/252, 85.7%) did not indicate or specify the framework used to guide the implementation. Of those that did, there was a good mix of implementation science (23/252, 9.1%) and other frameworks used (13/252, 5.2%).

Implementation models and theories used were the capability, opportunity, motivation, behavior (COM-B) model with the associated Behavior Change Wheel (4/252, 1.6%) and the Normalization Process Theory (3/252, 1.2%). Classic theories used were the self-determination theory (4/252, 1.6%), social cognitive theory (2/252, 0.8%), and Bandura's self-efficacy theory (1/252, 0.4%). The determinant frameworks used were the Consolidated Framework for Implementation Research (4/252, 1.6%); Exploration, Preparation, Implementation, and Sustainment framework (1/252, 0.4%); the integrated Promoting Action on Research Implementation in Health Services framework (1/252, 0.4%); and the Theoretical Domains Framework (1/252, 0.4%). The sole process framework used was the model by Grol and Wensing [203] (1/252, 0.4%), although the Exploration, Preparation, Implementation, and Sustainment framework arguably straddles both determinant and process frameworks and could be included here.

Implementation Strategies

Most studies used 1 (103/252, 40.9%) or 2 (102/252, 40.5%) strategies to implement digital health (Table 1). Table 3 shows the implementation strategies used, mapped to the ERIC taxonomy [202,204]. The top 3 strategies used in terms of frequency were collecting feedback from target users (ERIC 46: obtain and use patients or consumers and family feedback; 196/252, 77.8%); reviewing clinical performance details and providing feedback (ERIC 5: audit and provide feedback; 106/252, 42.1%); and conducting pilot studies before implementation (ERIC 61: stage implementation scale-up; 85/252, 33.7%).

XSL•FO
RenderX

Table 3. Implementation strategies (N=252).

Implementation strategy	Papers, n (%)[a]
ERIC[b] 46: obtain and use patients or consumers and family feedback	196 (77.8)
ERIC 5: audit and provide feedback	106 (42.1)
ERIC 61: stage implementation scale-up	85 (33.7)
ERIC 41: involve patients or consumers and family members	19 (7.5)
ERIC 4: assess for readiness and identify barriers and facilitators	10 (4)
ERIC 17: conduct local consensus discussions	3 (1.2)
ERIC 18: conduct local needs assessment	3 (1.2)
ERIC 56: purposefully re-examine the implementation	3 (1.2)
ERIC 19: conduct ongoing training	2 (0.8)
ERIC 35: identify and prepare champions	2 (0.8)
ERIC 55: provide ongoing consultation	2 (0.8)
ERIC 64: use advisory boards and workgroups	2 (0.8)
ERIC 65: use an implementation adviser	2 (0.8)
ERIC 69: use mass media	2 (0.8)
ERIC 15: conduct educational meetings	1 (0.4)
ERIC 26: develop and implement tools for quality monitoring	1 (0.4)
ERIC 27: develop and organize quality monitoring systems	1 (0.4)
ERIC 29: develop educational materials	1 (0.4)
ERIC 54: provide local technical assistance	1 (0.4)
ERIC 63: tailor strategies	1 (0.4)
ERIC 67: use data experts	1 (0.4)

[a]Percentages do not add up to 100% as some studies used >1 implementation strategy.

[b]ERIC refers to the Expert Recommendations for Implementing Change taxonomy for implementation strategies.

Outcomes in Implementation Research

Overview

The main assessment method used to evaluate outcomes was surveys (173/252, 68.7%; Table 1). Outcome measures used to evaluate the implementation and effectiveness of digital health innovations were sorted into 3 categories, following Proctor et al [21] (Table 1 and Multimedia Appendix 1): implementation outcomes, service outcomes, and patient outcomes. While favorable implementation outcomes indicate successful implementation, the effectiveness of the digital health innovation is determined by service and patient outcomes. The 3 most common outcomes measured were acceptability (171/252, 67.9%), which is an implementation outcome, and 2 patient outcomes, namely health outcomes (81/252, 32.1%) and patient satisfaction (40/252, 15.9%). Service outcomes were largely not measured, with only a handful of studies (5/252, 2%) monitoring safety or adverse events.

Implementation Outcomes

Most studies that measured the acceptability, feasibility, and adoption of eHealth found it acceptable (38/39, 97%), feasible (7/10, 70%), and with a high rate of adoption (10/10, 100%). The sole study on eHealth that assessed fidelity reported that it was high among target users [61]. Similarly, most studies that evaluated the acceptability, feasibility, and adoption of mHealth found it to be acceptable (70/74, 95%), feasible (16/18, 89%), and with a high rate of adoption (14/15, 93%). However, the 2 studies that investigated the appropriateness of mHealth reported that it was rated low by target users [42,62].

Likewise, most studies that assessed the acceptability and feasibility of telehealth found it acceptable (59/61, 97%) and feasible (13/14, 93%). The studies that measured the adoption of telehealth (4/5, 80%) reported that the adoption rate was high. All the studies that evaluated the fidelity of telehealth concluded that it was high among target users (3/3, 100%), and the 2 studies that analyzed the appropriateness of telehealth found it appropriate [56,58]. However, the sole study on telehealth that measured penetration observed that it varied widely across different health systems [63], and the 2 studies on telehealth that assessed sustainability received mixed reviews [63,64].

As for wearables, most studies (7/8, 88%) that evaluated its acceptability found it acceptable. The sole study focusing on wearables that measured feasibility concluded that it was feasible [65], and the 2 studies that assessed the adoption reported that the adoption rate was high (Multimedia Appendix 1) [66,67].

XSL•FO
RenderX

Service Outcomes

Most studies (247/252, 98%) did not measure service outcomes, such as efficiency, safety, effectiveness, equity, patient-centeredness, and timeliness. Of the studies that did (5/252, 2%), study authors measured safety and concluded that there were no major adverse incidents caused by the innovation [68-71] or that the adverse events only occurred in a small number of users [72].

Patient Outcomes

Of 252 studies, 89 (35.3%) measured patients' health or quality of life. Most of these studies (65/89, 73%) reported an improvement in at least 1 aspect of patients' health or quality of life after the implementation of eHealth, mHealth, or telehealth, although 1 study (1%) saw a drop in patients' health status but an improvement in quality of life. Of the 24 remaining studies, 22 observed no substantial difference (22/89, 25%), while 2 noticed a reduction in patients' health status (2/89, 2%).

Of 252 studies, 40 (15.9%) measured patient satisfaction with the innovation. Most of these studies (39/40, 98%) reported that patient satisfaction with eHealth, mHealth, and telehealth was generally high. Improvements in patient empowerment, specifically self-management and self-care were seen after the implementation of eHealth and mHealth innovations, respectively [73-75]. In addition, improvements in patient knowledge were seen after the implementation of eHealth [76,77]. None of the articles on wearables investigated patient outcomes (Multimedia Appendix 1).

Barriers and Enablers to Implementation

Of the 252 studies, 123 (48.8%) addressed barriers and enablers to implementation. Most of these studies elicited barriers and facilitators as part of the study aims (110/123, 89.4%), with the exception of 13 studies (10.6%) that reported barriers and facilitators as part of the findings or discussion. The barriers and enablers have been grouped into 4 categories: external factors, factors related to health care workers, patient-related factors, and factors pertinent to both patients and health care workers.

The categories were generated based on the barriers and enablers identified in this review. Barriers and enablers related to recipients (target users) of the digital health innovations were categorized as "factors related to health care workers," "patient-related factors," or "factors pertinent to both patients and health care workers." Barriers and enablers related to systems, policies, and infrastructure were categorized as "external factors."

Barriers

Overview

Common barriers affecting the implementation of eHealth, mHealth, and telehealth are presented in Figure 2, with specific barriers for the individual innovations detailed in Table 4.

Figure 2. Common barriers affecting the implementation of eHealth, mobile health (mHealth), and telehealth.

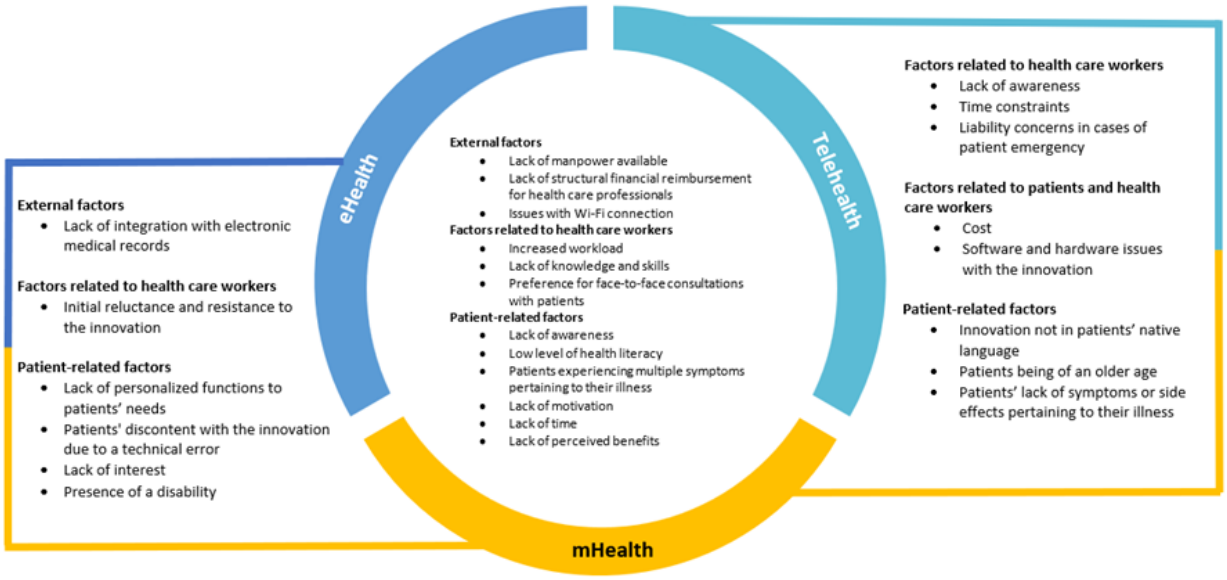

XSL•FO
RenderX

Table 4. Specific barriers pertaining to individual innovations.

Innovation	External factors	Factors related to health care workers	Factors related to patients and health care workers	Patient-related factors
Digital health in general[a]	• Lack of resources to manage the data generated by the digital health tool • Poor quality of health data collected	• Lack of strong evidence for health care workers to safely implement digital health	—[b]	—
eHealth	• Lack of resources to manage privacy and safety concerns	• Lack of access to eHealth tool • Cost • Initial skepticism pertaining to the innovation • Concerns about the eHealth tool's reliability	• Insufficient support and feedback provided by health care workers to patients • Lack of ease of use	• Difficulties in loading and logging into the eHealth portal • Patients being reminded of being sick when using eHealth • Patients experiencing unexpected life events • Patients not being able to use their own personal device as the eHealth tool • Lack of training and support provided
mHealth[c]	• Lack of support from senior physicians at the initial stages of implementation • Lack of systematic assessment and reporting of patients' health records or lack of integrated care for patients • Mismatch between the desire of the clinic director to implement mHealth and the buy-in from staff carrying out the implementation	• Time constraints when mHealth was first implemented	—	• High level of fatigue • Lack of access to smartphones • Presence of too many tasks to complete on the mHealth app • Preference or switching to nontraditional medicine
Telehealth	—	• Difficulties assessing patients via telehealth • Lack of buy-in and engagement • Lack of communication among the different health care professionals	• Lack of technical support • Lack of space and telehealth equipment	• Lack of family support
Wearables	• Technical difficulties linking wearable to a web-based server	—	—	• Discomfort from using wearables • Lack of trust • Poor rapport with clinicians

[a]Type of digital health innovation not specified.

[b]Not applicable.

[c]mHealth: mobile health.

Common Barriers Affecting Implementation

Data privacy emerged as the foremost common barrier among various digital health innovations [37,38,43,45,78-92]. Patient preference for in-person consultations [53,66,83,84,89,93-95] and their level of comfort with digital health or technology in general [66,96-99] were common barriers to the implementation of eHealth, mHealth, telehealth, and wearables. The lack of manpower [63,80,85,100-104] and the preference of health care workers for in-person consultations with patients [43,60,78,85,90,105-107] were shared barriers for eHealth, mHealth, and telehealth innovations. The lack of integration

XSL•FO
RenderX

with electronic medical records [60,80,81,90,108-113] was a reported barrier for both eHealth and mHealth. Health care workers' time constraints were an important barrier to the implementation of both mHealth and telehealth [100,104,114,115].

Barriers Unique to Type of Technology

Some barriers unique to eHealth included the lack of ease of use for both patients and health care workers [79,91] and the initial skepticism of health care workers about eHealth [81]. Barriers distinctive to mHealth were the lack of support from senior management at the initial phase of implementation [101] and patients' lack of access to smartphones [40,96]. For telehealth, unique barriers included health care workers facing difficulties evaluating patients over the internet [104] and the lack of space and equipment for patients and health care workers to attend and conduct telehealth visits, respectively [63,83,98,104,105,115,116]. Barriers unique to wearables include the presence of discomfort after wearing the item [41] and patients' lack of trust in the technology [65]. For the sole study that evaluated digital health in general, some unique barriers were the lack of resources to handle the information collected by digital health tools and the lack of solid evidence to implement digital health safely [37] (Table 4).

Enablers

Common enablers affecting the implementation of eHealth, mHealth, and telehealth are depicted in Figure 3, while specific enablers of each type of innovation are presented in Table 5.

Figure 3. Common enablers affecting the implementation of eHealth, mobile health (mHealth), and telehealth.

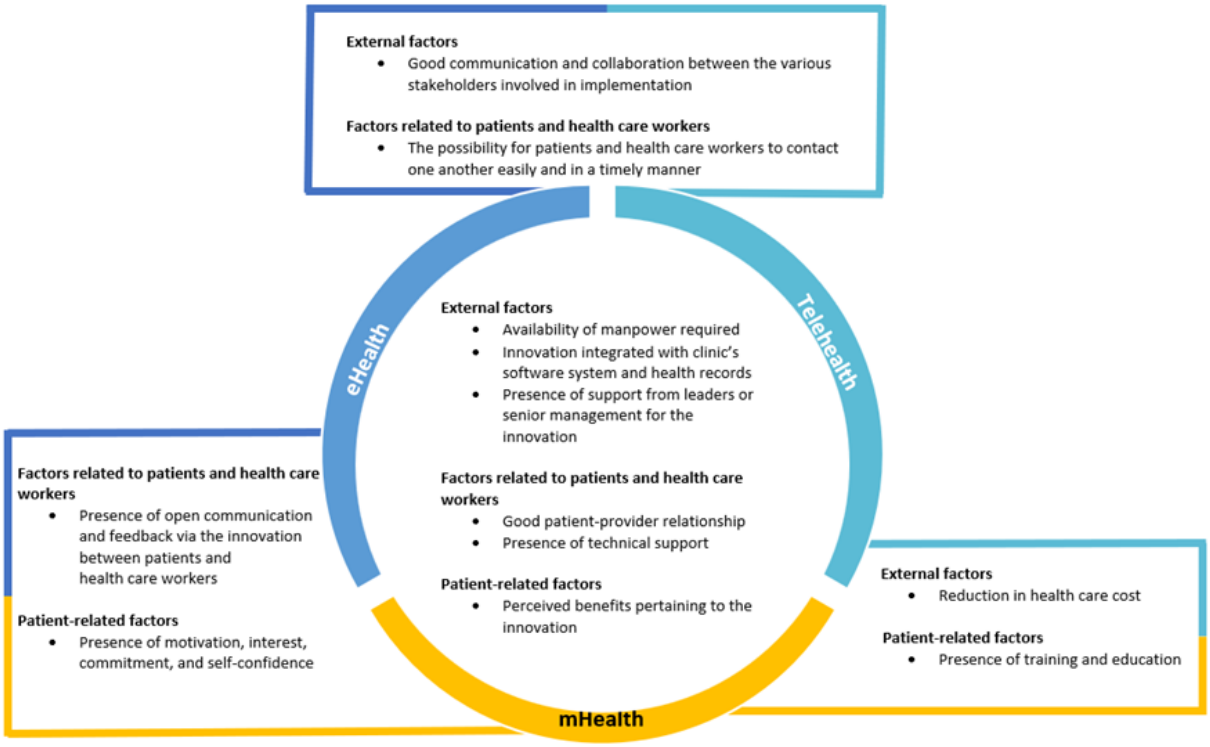

XSL•FO
RenderX

Table 5. Specific enablers pertaining to individual innovations.

Innovation	External factors	Factors related to health care workers	Factors related to patients and health care workers	Patient-related factors
Digital health in general[a]	—[b]	• High level of digital literacy	—	—
eHealth	—	• Awareness of eHealth • Positive perception of eHealth • Perception of having control over patient management • Increased productivity • Presence of incentives	—	• Ability to engage and use the innovation
mHealth[c]	—	—	• mHealth app having a comprehensive set of features • Having a say in the development of mHealth • Presence of health care workers monitoring patients' use of mHealth	• Being of an older age • Familiarity with smartphone • High level of self-efficacy • Perception that the innovation is credible • Presence of reminders to use mHealth
Telehealth	• Presence of adequate funding needed for the implementation • Telehealth as a complement, not a replacement of in-person consultations	• Presence of engagement with health care workers on telehealth • Perception that telehealth is credible • High confidence level with regard to using telehealth • Presence of administrative support	—	• High confidence and trust in the health care system • Presence of a program champion • Presence of encouragement from health care workers to patients to adopt telehealth • Presence of flexibility to attend telehealth sessions at a preferred place and time • Presence of social support • Comfort with telehealth
Wearables	—	—	—	• Being of a younger age • Innovation easily worn under clothing without snagging • Presence of severe symptoms

[a]Type of digital health innovation not specified.

[b]Not applicable.

[c]mHealth: mobile health.

Common Enablers Affecting Implementation

The common enablers between eHealth, mHealth, telehealth, and digital health in general were the presence of training for health care workers [37,45,60,78,80,100,101,117] and personalization of digital health features to patients' needs [37,42,58,80,82,115,118-121,205]. The sole common enabler among eHealth, mHealth, telehealth, and wearables was the ease of use of the respective technologies [36,43,60,63,66,78,80,82,84,87,96,108,109,111,115,118,121,205-212]. With regard to eHealth, mHealth, and telehealth, some common enablers among them included a good relationship between the patient and the health care worker [79,104,118,205,213,214] and the availability of manpower to implement the innovation [98,101,215]. The presence of motivation to use the innovation was particular to both eHealth and mHealth [119,206]. The presence of education and training for patients was an enabler for both mHealth and telehealth [43,45,78,87,206]. Good communication between the implementation stakeholders was an enabler for both telehealth and eHealth [58,102,104,105,108,114,215].

XSL•FO
RenderX

Enablers Unique to Type of Technology

Enablers unique to eHealth included the ability of health care workers to evaluate more patients without an increase in time commitment [90,216] and the presence of incentives for health care workers, such as the provision of financial reimbursement and continuing medical education points [108]. Enablers distinctive to mHealth included patients' familiarity with smartphones [217] and a high level of self-efficacy [216]. For telehealth, unique enablers included the presence of encouragement from health care workers to patients to use telehealth [98,114,218] and flexibility for patients to attend telehealth consultations at their convenience [83,85,93,105,114]. With regard to wearables, examples of unique enablers were the ability to wear the innovation easily under clothing and users being of a younger age group, as wearable designs often do not consider the requirements and needs of older adults [66]. For the sole study that evaluated digital health in general, a high level of digital literacy among health care workers was a specific enabler [37] (Table 5).

Lessons Learned Shared by Study Authors

Of the 252 studies, 62 (24.6%) specified lessons learned. Many patients are generally open to digital health [207], and the initial uptake is usually high [219]. However, health care workers often perceive it as an increase in workload [110]. For digital health innovations to be successful, it is important for them to be tailored to the patient's needs [37,41,67,83,84,86,87,89,111,115,119,121,212,220-227], to enhance the patient-provider relationship [90], and to complement physical appointments [57,105,222,228]. Effective communication between implementation stakeholders is required [87,98,229], and ongoing evaluation is needed to ensure that the uptake remains high [60]. In addition, the benefits of digital health need to be evident to target users [230].

To increase the uptake of digital health, support from relevant authorities is needed [109], along with raising awareness of its benefits [45,78] and embedding it into existing workflows [60,90]. Other ways to increase uptake include the provision of resources required for implementation [98,225,231], the presence of technical support [37,60,225,231], and training for target users [60,213,225]. In addition, digital health innovations should be user-friendly [41,45,219,231,232], target users should be included in its development and implementation [60,66,110,221,233,234], and a transition stage should be included for target users to adapt to the innovation [60].

Recommendations Shared by Study Authors

Of the 252 studies, 122 (48.4%) studies shared recommendations. Study authors recommended conducting a study of a longer duration [37,51,230,235-240] or conducting a larger study [74,96,120,210,224,232,241-244] for future studies. Many authors also recommended using controlled trials for future studies [54,97,110,112,113,207,208,211,230,231,245-256]. Some suggested larger or long-term trials [97,110,112,113,207,211,245,246,248,250,253,255-257], while others proposed clinical trials in general [258,259] or having a control group [46,56,96,260-263].

Furthermore, some authors recommended future research in other contexts, such as different countries or cities [43,59,68,88], different health care systems [59,68] or settings [47,211,264-266], different stakeholder groups involved in the implementation [43,80,96,267], different demographics in terms of age [268] or socioeconomic status [222,269], more diverse demographics [89,91,104,107,110,232,243,257,258,264,266,270], different illnesses [271-274], or severity of disease [99]. In contrast, other authors suggested focusing on patients with similar illness classification [242,275].

Other proposals pertinent to implementation include using different implementation strategies [76], monitoring different outcomes [36,107,225,276], collecting outcomes more frequently [277], or understanding the association between different features of the innovation and their corresponding outcomes [55,278,279]. Other suggestions include personalizing the innovation to individual users [88,95,121,258,280], having more extensive user testing [115,273,281], or making relevant changes to current features of the innovation based on the research findings [213,219,241,282,283].

Moreover, several authors recommended focusing on efficacy [43,45,49,70,217,247,256,284-287], validation [229], effectiveness [65,210,221,258,288-290], cost-effectiveness [45,96,115,247,281], clinical applicability [291], or usability of the innovation [292]. Other proposals include using a mixed methods approach for future studies [43,45,218,248,293] or qualitative research methods to understand target users' experiences with digital health [42,43,45,59,230].

Discussion

Principal Findings

This review bridges a critical gap in the deployment of digital health innovations for the management of chronic diseases by eliciting how such innovations have been implemented and evaluated to date. We conducted the review through the lens of implementation science to generate actionable findings for future real-world implementation of digital health innovations.

First, >90% of the 252 studies included in this review did not report using an implementation science framework for planning, guiding, or evaluating the real-world deployment of digital health technologies, despite the availability of many suitable frameworks to guide implementation [11]. This reportedly low use of implementation science frameworks may be due to implementation science being an emerging field. Hence, awareness and the potential usefulness of such frameworks for implementing digital health technologies are not yet widespread. Of the studies that reported using an implementation science framework, implementation theories, classic theories, and determinant frameworks were the most frequently used. Implementation science frameworks developed specifically for the deployment of digital health technologies were rarely used; for example, only 3 studies used the Normalization Process Theory and none used the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability framework [294].

Of 252 studies, only 23 (9.1%) leveraged an implementation science framework. Study authors used these frameworks to

XSL•FO
RenderX

underpin qualitative data collection, data analysis, or interpretation of findings; to inform the design of behavior change and implementation strategies to enable successful deployment; and for the evaluation of implementation outcomes. Although most included studies (229/252, 90.9%) did not formally use implementation science frameworks to guide implementation, these studies have, nevertheless, addressed or incorporated some elements of these frameworks. For example, 48.8% (123/252) of the studies assessed barriers and enablers affecting implementation. The low use of implementation science frameworks represents missed opportunities to generate reliable translation and scale-up of evidence-based innovations through a deeper understanding of contextual influences and eliciting mechanisms of implementation, as well as productive theorizing of implementation research [23,295].

Second, of the studies that measured implementation or patient outcomes, most (188/227, 82.8%) reported positive implementation or patient outcomes despite the low use of implementation science frameworks in the deployment of digital health innovations. Target users generally found digital health to be acceptable and feasible, and study authors reported high adoption. Only 6 (2.4%) of the 252 studies included appropriateness and sustainability as outcome measures to evaluate implementation, and the results were mixed. Of the studies that measured patients' health outcomes or quality of life, 77% (64/83) reported improvements, with patients being generally satisfied with digital health. These outcomes were mainly assessed using surveys (173/252, 68.7%); this is unsurprising as surveys are commonly used in health services research [296]. Most surveys contained at least 1 validated scale to assess outcomes (103/173, 59.5%) or were created by study authors (85/173, 49.1%).

Third, the most commonly used implementation strategies were evaluative and iterative strategies as per the ERIC taxonomy [202]; this is not surprising, given the relatively higher importance and feasibility of these strategies [202]. Studies focused on innovation development used a variety of co-design strategies with target users, such as patients and health care professionals. Strategies such as obtaining user feedback and conducting pilots before implementation were necessary, considering that most of the digital health innovations were in the early stages of real-world deployment.

Comparison to Prior Work

Overall, the findings from this review are similar to previous scoping reviews in emphasizing the usefulness of digital health innovations, and they are viewed positively by target users [297-299]. Willis et al [297] observed that many included studies found statistically significant improvements in implementation outcomes (eg, adoption and acceptability) and health care performance outcomes (eg, validated health measures), congruent with the findings of our review. Patel et al [298] highlighted that numerous studies indicated that many participants were willing to use digital health innovations, especially if they are personalized to target users' needs and preferences, which is a key enabler found in our review. Likewise, Lim et al [299] argued that most participants found digital health innovations highly acceptable and perceived them to be convenient, user-friendly, and practical.

Despite the above, many studies in this review indicated numerous barriers that could affect implementation. Notable barriers include concerns over data privacy [37,38,43,45,78-92] and patients' preference for physical consultations [53,55,66,83,84,89,93-95]. While it remains to be seen whether the convenience and flexibility of accessing health care via digital modalities would trump patient preference for physical consultations, assuming the quality of care is not compromised, ensuring data privacy is nonnegotiable. Not being able to assure data privacy is arguably the most formidable barrier affecting successful implementation [300,301].

In our review, two-thirds (171/252, 67.9%) of the included studies examined acceptability as an implementation outcome, followed by feasibility (38/252, 15.1%), and adoption (32/252, 12.7%). Other implementation outcomes, such as penetration, sustainability, appropriateness, fidelity, and cost, were infrequently reported (number of studies ranged between 1 and 6 studies). Similarly, a recent scoping review by Proctor et al [302] investigating the progress of implementation outcomes research found that 52.5% (210/400) of their included studies examined acceptability [302]. Fidelity was the next most commonly examined outcome (157/400, 39.3%), followed by feasibility (154/400, 38.5%), adoption (106/400, 26.5%), and appropriateness (87/400, 21.8%). Implementation outcomes such as penetration (64/400, 16%), sustainability (63/400, 15.8%), and cost (31/400, 7.8%) were relatively less frequently examined [302], albeit in higher proportions of included studies compared to our review. Our finding that service outcomes were rarely reported (5/252, 2%) contrasts with that reported in the scoping review by Proctor et al [302], where a small percentage (22/400, 5.5%) of included studies not only reported service outcomes but also examined the relationship between implementation outcomes and service outcomes.

Limitations and Strengths

There are several limitations to this review. First, the identification and classification of implementation strategies and outcomes for included studies were not without challenges, as studies used various terms to describe these constructs. To reduce errors and to ensure consistency in the interpretation of terms for the purposes of data extraction, we piloted the data extraction form and resolved discrepancies before actual extraction.

Second, we were not able to judge how well implementation strategies were carried out, regardless of whether an implementation science framework was used or not, apart from documenting the outcomes reported by the study authors. We cannot rule out that in some included studies, the outcomes were suboptimal due to poorer execution of implementation strategies rather than the lack of an implementation science framework to underpin the work per se.

Third, we did not assess the quality of reporting against the Standards for Reporting Implementation Studies (StaRI) [303], as not all included studies explicitly claimed to be an implementation study. StaRI comprises a 27-item checklist

XSL•FO
RenderX

spanning both the implementation strategy and the clinical, health care, or public health intervention (the innovation) being implemented. According to StaRI, study authors should describe the scientific background and explain their rationale for selecting the underpinning implementation theory, model, or framework and implementation strategies. Study authors should do the same for the innovation being implemented and include a description of the evidence of its effectiveness [303]. Adhering to StaRI would also require a description of how the "selected strategy is expected to achieve its effects"; for example, a logic model or pathway showing hypothesized mechanisms of action for how each implementation strategy is expected to bring about desired patient outcomes [303]. The logic model or pathway does not replace the use of an implementation theory, model, or framework to underpin the work of implementation.

Fourth, we did not include preprints or unpublished literature, which might have affected our findings. Digital health is a rapidly growing field, and pioneering innovations may have yet to be published in peer-reviewed journals. However, preprints are more likely to capture the development and validation phases of digital health innovations rather than last-mile real-world implementation, which is the scope of this review.

Strengths of this review include a comprehensive search strategy and broad inclusion criteria to ensure that we capture as many relevant studies as possible. The included studies are not limited to a particular geographical region or type of digital health innovation, thus allowing a representative overview to be generated. We categorized the findings on barriers and enablers to understand differences related to patients, health care workers, and external factors across different digital health innovations, which will be useful for planning and designing future implementation studies.

Future Directions

This scoping review generated a summary of how digital health in chronic disease management is currently implemented. The benefit of using implementation science frameworks was not unequivocal and was out of the scope of this review. Nevertheless, our findings established a basis for future studies to investigate whether using an implementation science framework (and how well it is used) compared to not using one would yield better or more consistent outcomes. In addition, we agree with Proctor et al [302] that future studies testing the relationships between implementation strategies and implementation, service, and patient outcomes are needed.

The digital health innovation being implemented should not unduly burden target users, whether patients or health care providers. Investing in co-design with target users as early as practicable is a smart way to ensure that the innovation is fit for purpose and to pre-emptively identify user issues. However, as users and contexts continually evolve and change, acceptability of digital health innovation may change with time. Hence, the acceptability of the innovation should be assessed at several time points, and the findings should be used to inform iterative improvements in innovation design.

The process of deployment is as important as the features of the innovation. Target users require adequate preparation and support throughout the implementation. It should be ensured that process measures and adverse events are monitored consistently and addressed. Planning for and monitoring sustainability should be a key outcome measure, given the not-insignificant investment of resources in digital health.

Conclusions

Digital health has changed how health care is viewed and managed; nonetheless, how it is implemented in real-world settings can be further optimized. This implementation science–guided scoping review generated a comprehensive summary of the various ways digital health innovations have been implemented and evaluated for chronic disease management. Findings serve as a useful resource for physicians, researchers, health system managers, and policy makers when designing the successful implementation of digital health innovations.

Acknowledgments

No funding was provided for this study.

Authors' Contributions

EL conceptualized and supervised the study. CP and EL constructed and refined the search strategy. CP, RMWWT, and EL were involved in the screening and development of the data extraction form. CP and RMWWT piloted the data extraction form. CP conducted data extraction for the remaining papers and conducted data analysis. CP and EL codrafted the manuscript. CP revised the draft with EL, and coauthors (RMWWT and YCT) provided further critical inputs to the manuscript. The authors have viewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials which include the PRESS (Peer Review of Electronic Search Strategies) checklist, search strategies, data extraction, list of studies in the scoping review, types of study, and the outcome measures used to evaluate implementation success and effectiveness of the digital health innovation.

XSL•FO
RenderX

[DOCX File , 122 KB - jmir_v26i1e53576_app1.docx]

Multimedia Appendix 2
PRISMA-ScR Checklist.
[PDF File (Adobe PDF File), 84 KB - jmir_v26i1e53576_app2.pdf]

References

1. Noncommunicable diseases. World Health Organization (WHO). 2022. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2023-01-18]
2. Foo KM, Sundram M, Legido-Quigley H. Facilitators and barriers of managing patients with multiple chronic conditions in the community: a qualitative study. BMC Public Health 2020 Feb 27;20(1):273 [FREE Full text] [doi: 10.1186/s12889-020-8375-8] [Medline: 32106838]
3. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 2011 Sep;10(4):430-439. [doi: 10.1016/j.arr.2011.03.003] [Medline: 21402176]
4. Seixas AA, Olaye IM, Wall SP, Dunn P. Optimizing healthcare through digital health and wellness solutions to meet the needs of patients with chronic disease during the COVID-19 era. Front Public Health 2021;9:667654 [FREE Full text] [doi: 10.3389/fpubh.2021.667654] [Medline: 34322469]
5. Gunasekeran DV, Tham YC, Ting DS, Tan GS, Wong TY. Digital health during COVID-19: lessons from operationalising new models of care in ophthalmology. Lancet Digit Health 2021 Feb;3(2):e124-e134 [FREE Full text] [doi: 10.1016/S2589-7500(20)30287-9] [Medline: 33509383]
6. Transforming healthcare: navigating digital health with a value-driven approach. World Economic Forum. URL: https://www3.weforum.org/docs/WEF_Transforming_Healthcare_2024.pdf [accessed 2024-08-14]
7. Shah N, Adusumalli S. Nudges and the meaningful adoption of digital health. Per Med 2020 Nov;17(6):429-433. [doi: 10.2217/pme-2020-0071] [Medline: 33026296]
8. Peterson HB, Dube Q, Lawn JE, Haidar J, Bagenal J, Horton R, Lancet Commission on Evidence-Based Implementation in Global Health. Achieving justice in implementation: the Lancet commission on evidence-based implementation in global health. Lancet 2023 Jul 15;402(10397):168-170. [doi: 10.1016/S0140-6736(23)00870-X] [Medline: 37354916]
9. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. Lancet 2022 Feb 12;399(10325):620. [doi: 10.1016/S0140-6736(22)00235-5] [Medline: 35151388]
10. Eccles MP, Mittman BS. Welcome to implementation science. Implement Sci 2006 Feb 22;1(1):5. [doi: 10.1186/1748-5908-1-1]
11. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10:53 [FREE Full text] [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
12. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24. [doi: 10.1002/chp.47] [Medline: 16557505]
13. Logan JO, Graham ID. Toward a comprehensive interdisciplinary model of health care research use. Sci Commun 1998 Dec 01;20(2):227-246. [doi: 10.1177/1075547098020002004]
14. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
15. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
16. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
17. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs, NJ: Prentice-Hall; 1986.
18. Orlikowski WJ. Improvising organizational transformation over time: a situated change perspective. Inf Syst J 1996 Mar;7(1):63-92. [doi: 10.1287/isre.7.1.63]
19. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554. [doi: 10.1177/0038038509103208]
20. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
21. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]

XSL•FO
RenderX

22. Hill JN, Guihan M, Hogan TP, Smith BM, LaVela SL, Weaver FM, et al. Use of the PARIHS framework for retrospective and prospective implementation evaluations. Worldviews Evid Based Nurs 2017 Apr;14(2):99-107. [doi: 10.1111/wvn.12211] [Medline: 28273407]
23. Damschroder LJ. Clarity out of chaos: use of theory in implementation research. Psychiatry Res 2020 Jan;283:112461 [FREE Full text] [doi: 10.1016/j.psychres.2019.06.036] [Medline: 31257020]
24. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022 Oct 05;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
25. Shi C, Dumville JC, Juwale H, Moran C, Atkinson R. Evidence assessing the development, evaluation and implementation of digital health technologies in wound care: a rapid scoping review. J Tissue Viability 2022 Nov;31(4):567-574 [FREE Full text] [doi: 10.1016/j.jtv.2022.09.006] [Medline: 36192302]
26. Whitelaw S, Pellegrini DM, Mamas MA, Cowie M, Van Spall HG. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021 Mar;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
27. Palacholla RS, Fischer N, Coleman A, Agboola S, Kirley K, Felsted J, et al. Provider- and patient-related barriers to and facilitators of digital health technology adoption for hypertension management: scoping review. JMIR Cardio 2019 Mar 26;3(1):e11951 [FREE Full text] [doi: 10.2196/11951] [Medline: 31758771]
28. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
29. Global strategy on digital health 2020-2025. World Health Organization (WHO). URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2021-11-08]
30. mHealth: use of appropriate digital technologies for public health. World Health Organization (WHO). 2018. URL: https://apps.who.int/gb/ebwha/pdf_files/wha71/a71_20-en.pdf [accessed 2021-11-08]
31. Brancati MC, Curtarelli M, Riso S, Baiocco S. How digital technology is reshaping the art of management. European Commission. 2022. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC130808 [accessed 2024-10-15]
32. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization (WHO). 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2023-03-25]
33. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2020.
34. There's a better way to manage your research. Open Science Framework. URL: https://osf.io/ [accessed 2023-01-18]
35. Implementation of digital health innovations for chronic disease management: A scoping review protocol. Open Science Framework. URL: https://osf.io/a846w/ [accessed 2024-06-04]
36. Raeside R, Singleton AC, Todd A, Partridge SR, Hyun KK, Kulas H, et al. Lung support service: implementation of a nationwide text message support program for people with chronic respiratory disease during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Dec 19;19(24):17073 [FREE Full text] [doi: 10.3390/ijerph192417073] [Medline: 36554954]
37. Slevin P, Kessie T, Cullen J, Butler MW, Donnelly SC, Caulfield B. Exploring the barriers and facilitators for the use of digital health technologies for the management of COPD: a qualitative study of clinician perceptions. QJM 2020 Mar 01;113(3):163-172. [doi: 10.1093/qjmed/hcz241] [Medline: 31545374]
38. Tran C, Dicker A, Leiby B, Gressen E, Williams N, Jim H. Utilizing digital health to collect electronic patient-reported outcomes in prostate cancer: single-arm pilot trial. J Med Internet Res 2020 Mar 25;22(3):e12689 [FREE Full text] [doi: 10.2196/12689] [Medline: 32209536]
39. Kvedarienė V, Biliute G, Didziokaitė G, Kavaliukaite L, Savonyte A, Rudzikaite-Fergize G, et al. Mobile health app for monitoring allergic rhinitis and asthma in real life in Lithuanian MASK-air users. Clin Transl Allergy 2022 Sep;12(9):e12192 [FREE Full text] [doi: 10.1002/clt2.12192] [Medline: 36178186]
40. Nahar P, Kannuri NK, Mikkilineni S, Murthy GV, Phillimore P. mHealth and the management of chronic conditions in rural areas: a note of caution from southern India. Anthropol Med 2017 Apr;24(1):1-16 [FREE Full text] [doi: 10.1080/13648470.2016.1263824] [Medline: 28292206]
41. Bentley CL, Powell L, Potter S, Parker J, Mountain GA, Bartlett YK, et al. The use of a smartphone app and an activity tracker to promote physical activity in the management of chronic obstructive pulmonary disease: randomized controlled feasibility study. JMIR Mhealth Uhealth 2020 Jun 03;8(6):e16203 [FREE Full text] [doi: 10.2196/16203] [Medline: 32490838]
42. Kettlewell J, Phillips J, Radford K, dasNair R. Informing evaluation of a smartphone application for people with acquired brain injury: a stakeholder engagement study. BMC Med Inform Decis Mak 2018 May 30;18(1):33 [FREE Full text] [doi: 10.1186/s12911-018-0611-0] [Medline: 29848312]
43. Alwashmi MF, Fitzpatrick B, Davis E, Farrell J, Gamble JM, Hawboldt J. Features of a mobile health intervention to manage chronic obstructive pulmonary disease: a qualitative study. Ther Adv Respir Dis 2020;14:1753466620951044 [FREE Full text] [doi: 10.1177/1753466620951044] [Medline: 32894025]

XSL•FO
RenderX

44. Song T, Liu F, Deng N, Qian S, Cui T, Guan Y, et al. A comprehensive 6A framework for improving patient self-management of hypertension using mHealth services: qualitative thematic analysis. J Med Internet Res 2021 Jun 21;23(6):e25522 [FREE Full text] [doi: 10.2196/25522] [Medline: 34152272]
45. Alwashmi MF, Fitzpatrick B, Farrell J, Gamble JM, Davis E, Nguyen HV, et al. Perceptions of patients regarding mobile health interventions for the management of chronic obstructive pulmonary disease: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 23;8(7):e17409 [FREE Full text] [doi: 10.2196/17409] [Medline: 32706697]
46. Racioppi A, Dalton T, Ramalingam S, Romero K, Ren Y, Bohannon L, et al. Assessing the feasibility of a novel mHealth app in hematopoietic stem cell transplant patients. Transplant Cell Ther 2021 Feb;27(2):181.e1-181.e9 [FREE Full text] [doi: 10.1016/j.jtct.2020.10.017] [Medline: 33830035]
47. Rodriguez Hermosa JL, Fuster Gomila A, Puente Maestu L, Amado Diago CA, Callejas González FJ, Malo De Molina Ruiz R, et al. Compliance and utility of a smartphone app for the detection of exacerbations in patients with chronic obstructive pulmonary disease: cohort study. JMIR Mhealth Uhealth 2020 Mar 19;8(3):e15699 [FREE Full text] [doi: 10.2196/15699] [Medline: 32191213]
48. O'Connor SR, Treanor C, Ward E, Wickens RA, O'Connell A, Culliford LA, Monarch Study Group. Patient acceptability of home monitoring for neovascular age-related macular degeneration reactivation: a qualitative study. Int J Environ Res Public Health 2022 Oct 21;19(20):13714 [FREE Full text] [doi: 10.3390/ijerph192013714] [Medline: 36294292]
49. Al Rajeh A, Steiner MC, Aldabayan Y, Aldhahir A, Pickett E, Quaderi S, et al. Use, utility and methods of telehealth for patients with COPD in England and Wales: a healthcare provider survey. BMJ Open Respir Res 2019;6(1):e000345 [FREE Full text] [doi: 10.1136/bmjresp-2018-000345] [Medline: 30956795]
50. Nair PP, Aghoram R, Thomas B, Bharadwaj B, Chinnakali P. Video teleconsultation services for persons with epilepsy during COVID-19 pandemic: an exploratory study from public tertiary care hospital in Southern India on feasibility, satisfaction, and effectiveness. Epilepsy Behav 2021 Apr;117:107863 [FREE Full text] [doi: 10.1016/j.yebeh.2021.107863] [Medline: 33677363]
51. D'Haeseleer M, Eelen P, Sadeghi N, D'Hooghe MB, Van Schependom J, Nagels G. Feasibility of real time internet-based teleconsultation in patients with multiple sclerosis: interventional pilot study. J Med Internet Res 2020 Aug 13;22(8):e18178 [FREE Full text] [doi: 10.2196/18178] [Medline: 32447274]
52. Vijayasundaram S, Karthikeyan P, Mehta SD. Proficiency of virtual follow-up amongst tinnitus patients who underwent intratympanic steroid therapy amidst COVID 19 pandemic. Am J Otolaryngol 2020;41(6):102680 [FREE Full text] [doi: 10.1016/j.amjoto.2020.102680] [Medline: 32861124]
53. Gordon HS, Solanki P, Bokhour BG, Gopal RK. "I'm not feeling like I'm part of the conversation" patients' perspectives on communicating in clinical video telehealth visits. J Gen Intern Med 2020 Jun;35(6):1751-1758 [FREE Full text] [doi: 10.1007/s11606-020-05673-w] [Medline: 32016705]
54. Kao DP, Lindenfeld J, Macaulay D, Birnbaum HG, Jarvis JL, Desai US, et al. Impact of a telehealth and care management program on all-cause mortality and healthcare utilization in patients with heart failure. Telemed J E Health 2016 Jan;22(1):2-11 [FREE Full text] [doi: 10.1089/tmj.2015.0007] [Medline: 26218252]
55. Dhakal R, Baniya M, Solomon RM, Rana C, Ghimire P, Hariharan R, et al. TEleRehabilitation Nepal (TERN) for people with spinal cord injury and acquired brain injury: a feasibility study. Rehabil Process Outcome 2022;11:11795727221126070 [FREE Full text] [doi: 10.1177/11795727221126070] [Medline: 36278119]
56. Fritz JM, Minick KI, Brennan GP, McGee T, Lane E, Skolasky RL, et al. Outcomes of telehealth physical therapy provided using real-time, videoconferencing for patients with chronic low back pain: a longitudinal observational study. Arch Phys Med Rehabil 2022 Oct;103(10):1924-1934. [doi: 10.1016/j.apmr.2022.04.016] [Medline: 35667399]
57. Kumar A, Lall N, Pathak A, Joshi D, Mishra VN, Chaurasia RN, et al. A questionnaire-based survey of acceptability and satisfaction of virtual neurology clinic during COVID-19 lockdown: a preliminary study. Acta Neurol Belg 2022 Oct;122(5):1297-1304 [FREE Full text] [doi: 10.1007/s13760-022-02004-0] [Medline: 35796999]
58. Li J, Varnfield M, Jayasena R, Celler B. Home telemonitoring for chronic disease management: perceptions of users and factors influencing adoption. Health Informatics J 2021;27(1):1460458221997893 [FREE Full text] [doi: 10.1177/1460458221997893] [Medline: 33685279]
59. Jaana M, Sherrard H, Paré G. A prospective evaluation of telemonitoring use by seniors with chronic heart failure: adoption, self-care, and empowerment. Health Informatics J 2019 Dec;25(4):1800-1814 [FREE Full text] [doi: 10.1177/1460458218799458] [Medline: 30247080]
60. van den Wijngaart LS, Geense WW, Boehmer AL, Brouwer ML, Hugen CA, van Ewijk BE, et al. Barriers and facilitators when implementing web-based disease monitoring and management as a substitution for regular outpatient care in pediatric asthma: qualitative survey study. J Med Internet Res 2018 Oct 30;20(10):e284 [FREE Full text] [doi: 10.2196/jmir.9245] [Medline: 30377147]
61. Peiris D, Praveen D, Mogulluru K, Ameer MA, Raghu A, Li Q, et al. SMARThealth India: a stepped-wedge, cluster randomised controlled trial of a community health worker managed mobile health intervention for people assessed at high cardiovascular disease risk in rural India. PLoS One 2019;14(3):e0213708 [FREE Full text] [doi: 10.1371/journal.pone.0213708] [Medline: 30913216]

XSL•FO
RenderX

62. Najm A, Lempp H, Gossec L, Berenbaum F, Nikiphorou E. Needs, experiences, and views of people with rheumatic and musculoskeletal diseases on self-management mobile health apps: mixed methods study. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e14351 [FREE Full text] [doi: 10.2196/14351] [Medline: 32310151]
63. Bauer MS, Krawczyk L, Tuozzo K, Frigand C, Holmes S, Miller CJ, et al. Implementing and sustaining team-based telecare for bipolar disorder: lessons learned from a model-guided, mixed methods analysis. Telemed J E Health 2018 Jan;24(1):45-53. [doi: 10.1089/tmj.2017.0017] [Medline: 28665773]
64. Brouwers RW, Kemps HM, Herkert C, Peek N, Kraal JJ. A 12-week cardiac telerehabilitation programme does not prevent relapse of physical activity levels: long-term results of the FIT@Home trial. Eur J Prev Cardiol 2022 May 25;29(7):e255-e257. [doi: 10.1093/eurjpc/zwac009] [Medline: 35040993]
65. Shin S, Yeom C, Shin C, Shin JH, Jeong JH, Shin JU, et al. Activity monitoring using a mHealth device and correlations with psychopathology in patients with chronic schizophrenia. Psychiatry Res 2016 Dec 30;246:712-718. [doi: 10.1016/j.psychres.2016.10.059] [Medline: 27836243]
66. AlMahadin G, Lotfi A, Zysk E, Siena FL, Carthy MM, Breedon P. Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. BMC Neurol 2020 Nov 18;20(1):419 [FREE Full text] [doi: 10.1186/s12883-020-01996-7] [Medline: 33208135]
67. Fisher JM, Hammerla NY, Rochester L, Andras P, Walker RW. Body-worn sensors in Parkinson's disease: evaluating their acceptability to patients. Telemed J E Health 2016 Jan;22(1):63-69 [FREE Full text] [doi: 10.1089/tmj.2015.0026] [Medline: 26186307]
68. Alwakeel AJ, Sicondolfo A, Robitaille C, Bourbeau J, Saad N. The accessibility, feasibility, and safety of a standardized community-based tele-pulmonary rehab program for chronic obstructive pulmonary disease: a 3-year real-world prospective study. Ann Am Thorac Soc 2022 Jan;19(1):39-47. [doi: 10.1513/AnnalsATS.202006-638OC] [Medline: 34170802]
69. Piotrowicz E, Pencina MJ, Opolski G, Zareba W, Banach M, Kowalik I, et al. Effects of a 9-week hybrid comprehensive telerehabilitation program on long-term outcomes in patients with heart failure: the telerehabilitation in heart failure patients (TELEREH-HF) randomized clinical trial. JAMA Cardiol 2020 Mar 01;5(3):300-308 [FREE Full text] [doi: 10.1001/jamacardio.2019.5006] [Medline: 31734701]
70. Seo NJ, Enders LR, Fortune A, Cain S, Vatinno AA, Schuster E, et al. Phase I safety trial: extended daily peripheral sensory stimulation using a wrist-worn vibrator in stroke survivors. Transl Stroke Res 2020 Apr;11(2):204-213 [FREE Full text] [doi: 10.1007/s12975-019-00724-9] [Medline: 31444692]
71. Yeh CH, Kawi J, Ni A, Christo P. Evaluating auricular point acupressure for chronic low back pain self-management using technology: a feasibility study. Pain Manag Nurs 2022 Jun;23(3):301-310. [doi: 10.1016/j.pmn.2021.11.007] [Medline: 34961729]
72. Sulkowski M, Luetkemeyer AF, Wyles DL, Martorell C, Muir A, Weisberg I, et al. Impact of a digital medicine programme on hepatitis C treatment adherence and efficacy in adults at high risk for non-adherence. Aliment Pharmacol Ther 2020 Jun;51(12):1384-1396. [doi: 10.1111/apt.15707] [Medline: 32352586]
73. Jongen PJ, Ter Veen G, Lemmens W, Donders R, van Noort E, Zeinstra E. The interactive web-based program MSmonitor for self-management and multidisciplinary care in persons with multiple sclerosis: quasi-experimental study of short-term effects on patient empowerment. J Med Internet Res 2020 Mar 09;22(3):e14297 [FREE Full text] [doi: 10.2196/14297] [Medline: 32149713]
74. Fortuna KL, Myers AL, Ferron J, Kadakia A, Bianco C, Bruce ML, et al. Assessing a digital peer support self-management intervention for adults with serious mental illness: feasibility, acceptability, and preliminary effectiveness. J Ment Health 2022 Dec;31(6):833-841 [FREE Full text] [doi: 10.1080/09638237.2021.2022619] [Medline: 35088619]
75. Park SK, Bang CH, Lee SH. Evaluating the effect of a smartphone app-based self-management program for people with COPD: a randomized controlled trial. Appl Nurs Res 2020 Apr;52:151231. [doi: 10.1016/j.apnr.2020.151231] [Medline: 31955942]
76. Palermo TM, de la Vega R, Murray C, Law E, Zhou C. A digital health psychological intervention (WebMAP Mobile) for children and adolescents with chronic pain: results of a hybrid effectiveness-implementation stepped-wedge cluster randomized trial. Pain 2020 Dec;161(12):2763-2774 [FREE Full text] [doi: 10.1097/j.pain.0000000000001994] [Medline: 32658147]
77. Patzer RE, McPherson L, Basu M, Mohan S, Wolf M, Chiles M, et al. Effect of the iChoose kidney decision aid in improving knowledge about treatment options among transplant candidates: a randomized controlled trial. Am J Transplant 2018 Aug;18(8):1954-1965 [FREE Full text] [doi: 10.1111/ajt.14693] [Medline: 29446209]
78. Alwashmi MF, Fitzpatrick B, Davis E, Gamble J, Farrell J, Hawboldt J. Perceptions of health care providers regarding a mobile health intervention to manage chronic obstructive pulmonary disease: qualitative study. JMIR Mhealth Uhealth 2019 Jun 10;7(6):e13950 [FREE Full text] [doi: 10.2196/13950] [Medline: 31199330]
79. Ariens LF, Schussler-Raymakers FM, Frima C, Flinterman A, Hamminga E, Arents BW, et al. Barriers and facilitators to eHealth use in daily practice: perspectives of patients and professionals in dermatology. J Med Internet Res 2017 Sep 05;19(9):e300 [FREE Full text] [doi: 10.2196/jmir.7512] [Medline: 28874336]

XSL•FO
RenderX

80. Cohn WF, Canan CE, Knight S, Waldman AL, Dillingham R, Ingersoll K, et al. An implementation strategy to expand mobile health use in HIV care settings: rapid evaluation study using the Consolidated Framework for Implementation Research. JMIR Mhealth Uhealth 2021 Apr 28;9(4):e19163 [FREE Full text] [doi: 10.2196/19163] [Medline: 33908893]
81. Gómez-Restrepo C, Cepeda M, Torrey W, Castro S, Uribe-Restrepo JM, Suárez-Obando F, et al. The DIADA project: a technology-based model of care for depression and risky alcohol use in primary care centres in Colombia. Rev Colomb Psiquiatr (Engl Ed) 2021 Jul;50 Suppl 1(Suppl 1):4-12 [FREE Full text] [doi: 10.1016/j.rcpeng.2020.11.005] [Medline: 34244120]
82. Haynes SC, Kim KK. A mobile system for the improvement of heart failure management: evaluation of a prototype. AMIA Annu Symp Proc 2017;2017:839-848 [FREE Full text] [Medline: 29854150]
83. Jansen-Kosterink S, Dekker-van Weering M, van Velsen L. Patient acceptance of a telemedicine service for rehabilitation care: a focus group study. Int J Med Inform 2019 May;125:22-29. [doi: 10.1016/j.ijmedinf.2019.01.011] [Medline: 30914177]
84. Jiang Y, Sun P, Chen Z, Guo J, Wang S, Liu F, et al. Patients' and healthcare providers' perceptions and experiences of telehealth use and online health information use in chronic disease management for older patients with chronic obstructive pulmonary disease: a qualitative study. BMC Geriatr 2022 Jan 03;22(1):9 [FREE Full text] [doi: 10.1186/s12877-021-02702-z] [Medline: 34979967]
85. Jordan DN, Jessen CM, Ferucci ED. Views of patients and providers on the use of telemedicine for chronic disease specialty care in the Alaska native population. Telemed J E Health 2021 Jan;27(1):82-89. [doi: 10.1089/tmj.2019.0284] [Medline: 32286156]
86. Kondylakis H, Bucur A, Crico C, Dong F, Graf N, Hoffman S, et al. Patient empowerment for cancer patients through a novel ICT infrastructure. J Biomed Inform 2020 Jan;101:103342 [FREE Full text] [doi: 10.1016/j.jbi.2019.103342] [Medline: 31816400]
87. Lee JY, Chan CK, Chua SS, Paraidathathu T, Lee KK, Tan CS, et al. Using telemedicine to support care for people with type 2 diabetes mellitus: a qualitative analysis of patients' perspectives. BMJ Open 2019 Oct 22;9(10):e026575 [FREE Full text] [doi: 10.1136/bmjopen-2018-026575] [Medline: 31640990]
88. Schneider T, Panzera AD, Martinasek M, McDermott R, Couluris M, Lindenberger J, et al. Physicians' perceptions of mobile technology for enhancing asthma care for youth. J Child Health Care 2016 Jun;20(2):153-163. [doi: 10.1177/1367493514556555] [Medline: 25427556]
89. Son YJ, Oh S, Kim EY. Patients' needs and perspectives for using mobile phone interventions to improve heart failure self-care: a qualitative study. J Adv Nurs 2020 Sep;76(9):2380-2390. [doi: 10.1111/jan.14455] [Medline: 32672374]
90. Taylor S, Allsop MJ, Bekker HL, Bennett MI, Bewick BM. Identifying professionals' needs in integrating electronic pain monitoring in community palliative care services: an interview study. Palliat Med 2017 Jul;31(7):661-670 [FREE Full text] [doi: 10.1177/0269216316677470] [Medline: 27836943]
91. Wakefield BJ, Alexander G, Dohrmann M, Richardson J. Design and evaluation of a web-based symptom monitoring tool for heart failure. Comput Inform Nurs 2017 May;35(5):248-254. [doi: 10.1097/CIN.0000000000000324] [Medline: 28045703]
92. Wright S, Thompson N, Yadrich D, Bruce A, Bonar JR, Spaulding R, et al. Using telehealth to assess depression and suicide ideation and provide mental health interventions to groups of chronically ill adolescents and young adults. Res Nurs Health 2021 Feb;44(1):129-137 [FREE Full text] [doi: 10.1002/nur.22089] [Medline: 33305830]
93. Barenfeld E, Fuller JM, Wallström S, Fors A, Ali L, Ekman I. Meaningful use of a digital platform and structured telephone support to facilitate remote person-centred care - a mixed-method study on patient perspectives. BMC Health Serv Res 2022 Apr 04;22(1):442 [FREE Full text] [doi: 10.1186/s12913-022-07831-8] [Medline: 35379247]
94. Scriven H, Doherty DP, Ward EC. Evaluation of a multisite telehealth group model for persistent pain management for rural/remote participants. Rural Remote Health 2019 Mar;19(1):4710 [FREE Full text] [doi: 10.22605/RRH4710] [Medline: 30884951]
95. Wannheden C, Stenfors T, Stenling A, von Thiele Schwarz U. Satisfied or frustrated? A qualitative analysis of need satisfying and need frustrating experiences of engaging with digital health technology in chronic care. Front Public Health 2020;8:623773 [FREE Full text] [doi: 10.3389/fpubh.2020.623773] [Medline: 33585391]
96. Barakat-Johnson M, Kita B, Jones A, Burger M, Airey D, Stephenson J, et al. The viability and acceptability of a virtual wound care command centre in Australia. Int Wound J 2022 Nov;19(7):1769-1785 [FREE Full text] [doi: 10.1111/iwj.13782] [Medline: 35607997]
97. Jakobsen AS, Laursen LC, Rydahl-Hansen S, Østergaard B, Gerds TA, Emme C, et al. Home-based telehealth hospitalization for exacerbation of chronic obstructive pulmonary disease: findings from "the virtual hospital" trial. Telemed J E Health 2015 May;21(5):364-373 [FREE Full text] [doi: 10.1089/tmj.2014.0098] [Medline: 25654366]
98. Martinez RN, Hogan TP, Balbale S, Lones K, Goldstein B, Woo C, et al. Sociotechnical perspective on implementing clinical video telehealth for veterans with spinal cord injuries and disorders. Telemed J E Health 2017 Jul;23(7):567-576 [FREE Full text] [doi: 10.1089/tmj.2016.0200] [Medline: 28067586]
99. Singh G, Nimmon L, Sawatzky B, Ben Mortenson W. Barriers and facilitators to eHealth technology use among community-dwelling individuals with spinal cord injury: a qualitative study. Top Spinal Cord Inj Rehabil 2022;28(2):196-204 [FREE Full text] [doi: 10.46292/sci21-00016] [Medline: 35521060]

XSL•FO
RenderX

100. Diez-Canseco F, Toyama M, Ipince A, Perez-Leon S, Cavero V, Araya R, et al. Integration of a technology-based mental health screening program into routine practices of primary health care services in Peru (the Allillanchu Project): development and implementation. J Med Internet Res 2018 Mar 15;20(3):e100 [FREE Full text] [doi: 10.2196/jmir.9208] [Medline: 29588272]
101. Jindal D, Roy A, Ajay VS, Yadav SK, Prabhakaran D, Tandon N. Strategies for stakeholder engagement and uptake of new intervention: experience from state-wide implementation of mHealth technology for NCD care in Tripura, India. Glob Heart 2019 Jun;14(2):165-172 [FREE Full text] [doi: 10.1016/j.gheart.2019.06.002] [Medline: 31324371]
102. Johnson EE, MacGeorge C, King KL, Andrews AL, Teufel RJ, Kruis R, et al. Facilitators and barriers to implementation of school-based telehealth asthma care: program champion perspectives. Acad Pediatr 2021;21(7):1262-1272. [doi: 10.1016/j.acap.2021.04.025] [Medline: 33940203]
103. Lapão LV, Peyroteo M, Maia M, Seixas J, Gregório J, Mira da Silva M, et al. Implementation of digital monitoring services during the COVID-19 pandemic for patients with chronic diseases: design science approach. J Med Internet Res 2021 Aug 26;23(8):e24181 [FREE Full text] [doi: 10.2196/24181] [Medline: 34313591]
104. Martinez RN, Hogan TP, Lones K, Balbale S, Scholten J, Bidelspach D, et al. Evaluation and treatment of mild traumatic brain injury through the implementation of clinical video telehealth: provider perspectives from the Veterans Health Administration. PM R 2017 Mar;9(3):231-240. [doi: 10.1016/j.pmrj.2016.07.002] [Medline: 27423365]
105. Damhus CS, Emme C, Hansen H. Barriers and enablers of COPD telerehabilitation - a frontline staff perspective. Int J Chron Obstruct Pulmon Dis 2018;13:2473-2482 [FREE Full text] [doi: 10.2147/COPD.S167501] [Medline: 30154650]
106. Hanley J, Fairbrother P, McCloughan L, Pagliari C, Paterson M, Pinnock H, et al. Qualitative study of telemonitoring of blood glucose and blood pressure in type 2 diabetes. BMJ Open 2015 Dec 23;5(12):e008896 [FREE Full text] [doi: 10.1136/bmjopen-2015-008896] [Medline: 26700275]
107. Wali S, Guessi Margarido M, Shah A, Ware P, McDonald M, O'Sullivan M, et al. Expanding telemonitoring in a virtual world: a case study of the expansion of a heart failure telemonitoring program during the COVID-19 pandemic. J Med Internet Res 2021 Jan 22;23(1):e26165 [FREE Full text] [doi: 10.2196/26165] [Medline: 33444153]
108. Bello AK, Molzahn AE, Girard LP, Osman MA, Okpechi IG, Glassford J, et al. Patient and provider perspectives on the design and implementation of an electronic consultation system for kidney care delivery in Canada: a focus group study. BMJ Open 2017 Mar 02;7(3):e014784 [FREE Full text] [doi: 10.1136/bmjopen-2016-014784] [Medline: 28255097]
109. de Batlle J, Massip M, Vargiu E, Nadal N, Fuentes A, Ortega Bravo M, CONNECARE-Lleida Group. Implementing mobile health-enabled integrated care for complex chronic patients: patients and professionals' acceptability study. JMIR Mhealth Uhealth 2020 Nov 20;8(11):e22136 [FREE Full text] [doi: 10.2196/22136] [Medline: 33216004]
110. Kooij L, Vos PJ, Dijkstra A, van Harten WH. Effectiveness of a mobile health and self-management app for high-risk patients with chronic obstructive pulmonary disease in daily clinical practice: mixed methods evaluation study. JMIR Mhealth Uhealth 2021 Feb 04;9(2):e21977 [FREE Full text] [doi: 10.2196/21977] [Medline: 33538699]
111. Lin JL, Huber B, Amir O, Gehrmann S, Ramirez KS, Ochoa KM, et al. Barriers and facilitators to the implementation of family-centered technology in complex care: feasibility study. J Med Internet Res 2022 Aug 23;24(8):e30902 [FREE Full text] [doi: 10.2196/30902] [Medline: 35998021]
112. Ong SW, Jassal SV, Miller JA, Porter EC, Cafazzo JA, Seto E, et al. Integrating a smartphone-based self-management system into usual care of advanced CKD. Clin J Am Soc Nephrol 2016 Jun 06;11(6):1054-1062 [FREE Full text] [doi: 10.2215/CJN.10681015] [Medline: 27173169]
113. Wei KS, Ibrahim NE, Kumar AA, Jena S, Chew V, Depa M, et al. Habits heart app for patient engagement in heart failure management: pilot feasibility randomized trial. JMIR Mhealth Uhealth 2021 Jan 20;9(1):e19465. [doi: 10.2196/19465] [Medline: 33470941]
114. Liu Y, Zupan NJ, Swearingen R, Jacobson N, Carlson JN, Mahoney JE, et al. Identification of barriers, facilitators and system-based implementation strategies to increase teleophthalmology use for diabetic eye screening in a rural US primary care clinic: a qualitative study. BMJ Open 2019 Feb 18;9(2):e022594 [FREE Full text] [doi: 10.1136/bmjopen-2018-022594] [Medline: 30782868]
115. Walsh DM, Moran K, Cornelissen V, Buys R, Cornelis N, Woods C. Electronic health physical activity behavior change intervention to self-manage cardiovascular disease: qualitative exploration of patient and health professional requirements. J Med Internet Res 2018 May 08;20(5):e163 [FREE Full text] [doi: 10.2196/jmir.9181] [Medline: 29739740]
116. Kayser MZ, Valtin C, Greer M, Karow B, Fuge J, Gottlieb J. Video consultation during the COVID-19 pandemic: a single center's experience with lung transplant recipients. Telemed J E Health 2021 Jul;27(7):807-815. [doi: 10.1089/tmj.2020.0170] [Medline: 33054671]
117. Rothgangel A, Braun S, Smeets R, Beurskens A. Feasibility of a traditional and teletreatment approach to mirror therapy in patients with phantom limb pain: a process evaluation performed alongside a randomized controlled trial. Clin Rehabil 2019 Oct;33(10):1649-1660. [doi: 10.1177/0269215519846539] [Medline: 31066315]
118. Doyle J, Murphy E, Gavin S, Pascale A, Deparis S, Tommasi P, et al. A digital platform to support self-management of multiple chronic conditions (ProACT): findings in relation to engagement during a one-year proof-of-concept trial. J Med Internet Res 2021 Dec 15;23(12):e22672 [FREE Full text] [doi: 10.2196/22672] [Medline: 34914612]

XSL•FO
RenderX

119. Jonassaint CR, Kang C, Prussien KV, Yarboi J, Sanger MS, Wilson JD, et al. Feasibility of implementing mobile technology-delivered mental health treatment in routine adult sickle cell disease care. Transl Behav Med 2020 Feb 03;10(1):58-67 [FREE Full text] [doi: 10.1093/tbm/iby107] [Medline: 30508141]
120. Lalloo C, Nishat F, Zempsky W, Bakshi N, Badawy S, Ko YJ, et al. Characterizing user engagement with a digital intervention for pain self-management among youth with sickle cell disease and their caregivers: subanalysis of a randomized controlled trial. J Med Internet Res 2022 Aug 30;24(8):e40096 [FREE Full text] [doi: 10.2196/40096] [Medline: 36040789]
121. Monteiro-Guerra F, Signorelli GR, Rivera-Romero O, Dorronzoro-Zubiete E, Caulfield B. Breast cancer survivors' perspectives on motivational and personalization strategies in mobile app-based physical activity coaching interventions: qualitative study. JMIR Mhealth Uhealth 2020 Sep 21;8(9):e18867 [FREE Full text] [doi: 10.2196/18867] [Medline: 32955446]
122. Adeagbo O, Herbst C, Blandford A, McKendry R, Estcourt C, Seeley J, et al. Exploring people's candidacy for mobile health-supported HIV testing and care services in rural KwaZulu-Natal, South Africa: qualitative study. J Med Internet Res 2019 Nov 18;21(11):e15681 [FREE Full text] [doi: 10.2196/15681] [Medline: 31738174]
123. Ahonle ZJ, Kreider CM, Hale-Gallardo J, Castaneda G, Findley K, Ottomanelli L, et al. Implementation and use of video tele-technologies in delivery of individualized community-based vocational rehabilitation services to rural veterans. J Vocat Rehabil 2021 Sep 07;55(2):227-233. [doi: 10.3233/JVR-211159]
124. Ammenwerth E, Woess S, Baumgartner C, Fetz B, van der Heidt A, Kastner P, et al. Evaluation of an integrated telemonitoring surveillance system in patients with coronary heart disease. Methods Inf Med 2015;54(5):388-397. [doi: 10.3414/ME15-02-0002] [Medline: 26395147]
125. Amparo F, Dana R. Web-based longitudinal remote assessment of dry eye symptoms. Ocul Surf 2018 Apr;16(2):249-253. [doi: 10.1016/j.jtos.2018.01.002] [Medline: 29409963]
126. Andersen TO, Langstrup H, Lomborg S. Experiences with wearable activity data during self-care by chronic heart patients: qualitative study. J Med Internet Res 2020 Jul 20;22(7):e15873 [FREE Full text] [doi: 10.2196/15873] [Medline: 32706663]
127. Anderson LM, Leonard S, Jonassaint J, Lunyera J, Bonner M, Shah N. Mobile health intervention for youth with sickle cell disease: impact on adherence, disease knowledge, and quality of life. Pediatr Blood Cancer 2018 Aug;65(8):e27081. [doi: 10.1002/pbc.27081] [Medline: 29693797]
128. Anglada-Martínez H, Martin-Conde M, Rovira-Illamola M, Sotoca-Momblona JM, Sequeira E, Aragunde V, et al. Feasibility and preliminary outcomes of a web and smartphone-based medication self-management platform for chronically ill patients. J Med Syst 2016 Apr;40(4):99. [doi: 10.1007/s10916-016-0456-y] [Medline: 26872781]
129. Ben-Zeev D, Brian RM, Jonathan G, Razzano L, Pashka N, Carpenter-Song E, et al. Mobile health (mHealth) versus clinic-based group intervention for people with serious mental illness: a randomized controlled trial. Psychiatr Serv 2018 Sep 01;69(9):978-985. [doi: 10.1176/appi.ps.201800063] [Medline: 29793397]
130. Bender C, Hangaard S, Kronborg T, Hejlesen OK, Secher PH. Preliminary qualitative evaluation of patient-related perspectives related to the implementation of a predictive algorithm in a telehealth system for COPD. Stud Health Technol Inform 2021 May 27;281:545-549. [doi: 10.3233/SHTI210230] [Medline: 34042635]
131. Benzo RP, Kramer KM, Hoult JP, Anderson PM, Begue IM, Seifert SJ. Development and feasibility of a home pulmonary rehabilitation program with health coaching. Respir Care 2018 Feb;63(2):131-140 [FREE Full text] [doi: 10.4187/respcare.05690] [Medline: 29066590]
132. Bhatia A, Kara J, Janmohamed T, Prabhu A, Lebovic G, Katz J, et al. User engagement and clinical impact of the manage my pain app in patients with chronic pain: a real-world, multi-site trial. JMIR Mhealth Uhealth 2021 Mar 04;9(3):e26528 [FREE Full text] [doi: 10.2196/26528] [Medline: 33661130]
133. Bonacini M, Kim Y, Pitney C, McKoin L, Tran M, Landis C. Wirelessly observed therapy to optimize adherence and target interventions for oral Hepatitis C treatment: observational pilot study. J Med Internet Res 2020 Feb 24;22(2):e15532 [FREE Full text] [doi: 10.2196/15532] [Medline: 32352385]
134. Botros A, Schütz N, Camenzind M, Urwyler P, Bolliger D, Vanbellingen T, et al. Long-term home-monitoring sensor technology in patients with Parkinson's disease-acceptance and adherence. Sensors (Basel) 2019 Nov 26;19(23):5169 [FREE Full text] [doi: 10.3390/s19235169] [Medline: 31779108]
135. Bruggeman-Everts FZ, Wolvers MD, van de Schoot R, Vollenbroek-Hutten MM, Van der Lee ML. Effectiveness of two web-based interventions for chronic cancer-related fatigue compared to an active control condition: results of the "Fitter na kanker" randomized controlled trial. J Med Internet Res 2017 Oct 19;19(10):e336 [FREE Full text] [doi: 10.2196/jmir.7180] [Medline: 29051138]
136. Chan C, Inskip JA, Kirkham AR, Ansermino JM, Dumont G, Li LC, et al. A smartphone oximeter with a fingertip probe for use during exercise training: usability, validity and reliability in individuals with chronic lung disease and healthy controls. Physiotherapy 2019 Sep;105(3):297-306. [doi: 10.1016/j.physio.2018.07.015] [Medline: 30876715]
137. Cingi C, Yorgancioglu A, Cingi CC, Oguzulgen K, Muluk NB, Ulusoy S, et al. The "physician on call patient engagement trial" (POPET): measuring the impact of a mobile patient engagement application on health outcomes and quality of life in allergic rhinitis and asthma patients. Int Forum Allergy Rhinol 2015 Jun;5(6):487-497. [doi: 10.1002/alr.21468] [Medline: 25856270]

XSL•FO
RenderX

138. Costa Stutzel M, Filippo MP, Sztajnberg A, da Costa RM, da Silva Brites Q, da Motta LB, et al. Multi-part quality evaluation of a customized mobile application for monitoring elderly patients with functional loss and helping caregivers. BMC Med Inform Decis Mak 2019 Jul 22;19(1):140 [FREE Full text] [doi: 10.1186/s12911-019-0839-3] [Medline: 31331309]
139. Davis LE, Harnar J, LaChey-Barbee LA, Pirio Richardson S, Fraser A, King MK. Using teleneurology to deliver chronic neurologic care to rural veterans: analysis of the first 1,100 patient visits. Telemed J E Health 2019 Apr;25(4):274-278. [doi: 10.1089/tmj.2018.0067] [Medline: 30016207]
140. Dodakian L, McKenzie AL, Le V, See J, Pearson-Fuhrhop K, Burke Quinlan E, et al. A home-based telerehabilitation program for patients with stroke. Neurorehabil Neural Repair 2017;31(10-11):923-933 [FREE Full text] [doi: 10.1177/1545968317733818] [Medline: 29072556]
141. Doyle N, Murphy M, Brennan L, Waugh A, McCann M, Mellotte G. The "Mikidney" smartphone app pilot study: empowering patients with Chronic Kidney Disease. J Ren Care 2019 Sep;45(3):133-140. [doi: 10.1111/jorc.12281] [Medline: 31355536]
142. Farmer A, Williams V, Velardo C, Shah SA, Yu LM, Rutter H, et al. Self-management support using a digital health system compared with usual care for chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2017 May 03;19(5):e144 [FREE Full text] [doi: 10.2196/jmir.7116] [Medline: 28468749]
143. Fernon A, Nguyen A, Baysari M, Day R. A user-centred approach to designing an eTool for gout management. Stud Health Technol Inform 2016;227:28-33. [Medline: 27440285]
144. Forbes G, Newton S, Cantalapiedra Calvete C, Birch J, Dodds J, Steed L, et al. MEMPHIS: a smartphone app using psychological approaches for women with chronic pelvic pain presenting to gynaecology clinics: a randomised feasibility trial. BMJ Open 2020 Mar 12;10(3):e030164 [FREE Full text] [doi: 10.1136/bmjopen-2019-030164] [Medline: 32165549]
145. Ford AR, Gibbons CM, Torres J, Kornmehl HA, Singh S, Young PM, et al. Access to dermatological care with an innovative online model for psoriasis management: results from a randomized controlled trial. Telemed J E Health 2019 Jul;25(7):619-627 [FREE Full text] [doi: 10.1089/tmj.2018.0160] [Medline: 30222518]
146. Gabbard J, McLouth CJ, Brenes G, Claudel S, Ongchuan S, Burkart J, et al. Rapid electronic capturing of patient-reported outcome measures in older adults with end-stage renal disease: a feasibility study. Am J Hosp Palliat Care 2021 May;38(5):432-440 [FREE Full text] [doi: 10.1177/1049909120954805] [Medline: 32935548]
147. Geirhos A, Domhardt M, Lunkenheimer F, Temming S, Holl RW, Minden K, et al. Feasibility and potential efficacy of a guided internet- and mobile-based CBT for adolescents and young adults with chronic medical conditions and comorbid depression or anxiety symptoms (youthCOACH): a randomized controlled pilot trial. BMC Pediatr 2022 Jan 29;22(1):69 [FREE Full text] [doi: 10.1186/s12887-022-03134-3] [Medline: 35093047]
148. Guo X, Gu X, Jiang J, Li H, Duan R, Zhang Y, et al. A hospital-community-family-based telehealth program for patients with chronic heart failure: single-arm, prospective feasibility study. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13229 [FREE Full text] [doi: 10.2196/13229] [Medline: 31833835]
149. Hale TM, Jethwani K, Kandola MS, Saldana F, Kvedar JC. A remote medication monitoring system for chronic heart failure patients to reduce readmissions: a two-arm randomized pilot study. J Med Internet Res 2016 Apr 17;18(5):e91 [FREE Full text] [doi: 10.2196/jmir.5256] [Medline: 27154462]
150. Halterman JS, Fagnano M, Tajon RS, Tremblay P, Wang H, Butz A, et al. Effect of the school-based telemedicine enhanced asthma management (SB-TEAM) program on asthma morbidity: a randomized clinical trial. JAMA Pediatr 2018 Mar 05;172(3):e174938 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.4938] [Medline: 29309483]
151. Huygens MW, Voogdt-Pruis HR, Wouters M, Meurs MM, van Lettow B, Kleijweg C, et al. The uptake and use of telemonitoring in chronic care between 2014 and 2019: nationwide survey among patients and health care professionals in the Netherlands. J Med Internet Res 2021 May 03;23(5):e24908 [FREE Full text] [doi: 10.2196/24908] [Medline: 33938808]
152. Jiménez-Reguera B, Maroto López E, Fitch S, Juarros L, Sánchez Cortés M, Rodríguez Hermosa JL, et al. Development and preliminary evaluation of the effects of an mhealth web-based platform (HappyAir) on adherence to a maintenance program after pulmonary rehabilitation in patients with chronic obstructive pulmonary disease: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e18465 [FREE Full text] [doi: 10.2196/18465] [Medline: 32513646]
153. Kadiri SB, Kerr AP, Oswald NK, Budacan AM, Flanagan S, Golby C, et al. Fit 4 surgery, a bespoke app with biofeedback delivers rehabilitation at home before and after elective lung resection. J Cardiothorac Surg 2019 Jul 05;14(1):132 [FREE Full text] [doi: 10.1186/s13019-019-0951-6] [Medline: 31277671]
154. Kamei T, Yamamoto Y, Kanamori T, Nakayama Y, Porter SE. Detection of early-stage changes in people with chronic diseases: a telehome monitoring-based telenursing feasibility study. Nurs Health Sci 2018 Sep;20(3):313-322. [doi: 10.1111/nhs.12563] [Medline: 30252192]
155. Kazankov K, Novelli S, Chatterjee DA, Phillips A, Balaji A, Raja M, et al. Evaluation of CirrhoCare® - a digital health solution for home management of individuals with cirrhosis. J Hepatol 2023 Jan;78(1):123-132. [doi: 10.1016/j.jhep.2022.08.034] [Medline: 36087864]
156. Kenealy TW, Parsons MJ, Rouse AP, Doughty RN, Sheridan NF, Hindmarsh JK, et al. Telecare for diabetes, CHF or COPD: effect on quality of life, hospital use and costs. A randomised controlled trial and qualitative evaluation. PLoS One 2015;10(3):e0116188 [FREE Full text] [doi: 10.1371/journal.pone.0116188] [Medline: 25768023]

XSL•FO
RenderX

157. Kouri A, Yamada J, Sale JE, Straus SE, Gupta S. Primary care pre-visit electronic patient questionnaire for asthma: uptake analysis and predictor modeling. J Med Internet Res 2020 Sep 18;22(9):e19358 [FREE Full text] [doi: 10.2196/19358] [Medline: 32945779]
158. Lalloo C, Hundert A, Harris L, Pham Q, Campbell F, Chorney J, et al. Capturing daily disease experiences of adolescents with chronic pain: mHealth-mediated symptom tracking. JMIR Mhealth Uhealth 2019 Jan 17;7(1):e11838 [FREE Full text] [doi: 10.2196/11838] [Medline: 30664472]
159. Li WY, Chiu FC, Zeng JK, Li YW, Huang SH, Yeh HC, et al. Mobile health app with social media to support self-management for patients with chronic kidney disease: prospective randomized controlled study. J Med Internet Res 2020 Dec 15;22(12):e19452 [FREE Full text] [doi: 10.2196/19452] [Medline: 33320101]
160. Lopez JJ, Svetanoff WJ, Rosen JM, Carrasco A, Rentea RM, Comprehensive Colorectal Center of Children's Mercy Hospital, Kansas City, MO. Leveraging collaboration in pediatric multidisciplinary colorectal care using a telehealth platform. Am Surg 2022 Sep;88(9):2320-2326. [doi: 10.1177/00031348211023428] [Medline: 34105388]
161. Mammen JR, Schoonmaker JD, Java J, Halterman J, Berliant MN, Crowley A, et al. Going mobile with primary care: smartphone-telemedicine for asthma management in young urban adults (TEAMS). J Asthma 2022 Jan;59(1):132-144. [doi: 10.1080/02770903.2020.1830413] [Medline: 33064038]
162. Martínez García MA, Fernández Rosales MS, López Domínguez E, Hernández Velázquez Y, Domínguez Isidro S. Telemonitoring system for patients with chronic kidney disease undergoing peritoneal dialysis: usability assessment based on a case study. PLoS One 2018;13(11):e0206600 [FREE Full text] [doi: 10.1371/journal.pone.0206600] [Medline: 30399164]
163. Mattocks KM, LaChappelle KM, Krein SL, DeBar LL, Martino S, Edmond S, et al. Pre-implementation formative evaluation of cooperative pain education and self-management expanding treatment for real-world access: a pragmatic pain trial. Pain Pract 2023 Apr;23(4):338-348. [doi: 10.1111/papr.13195] [Medline: 36527287]
164. Morano JP, Clauson K, Zhou Z, Escobar-Viera CG, Lieb S, Chen IK, et al. Attitudes, beliefs, and willingness toward the use of mHealth tools for medication adherence in the Florida mHealth adherence project for people living with HIV (FL-mAPP): pilot questionnaire study. JMIR Mhealth Uhealth 2019 Jul 03;7(7):e12900 [FREE Full text] [doi: 10.2196/12900] [Medline: 31271150]
165. Morisada MV, Hwang J, Gill AS, Wilson MD, Strong EB, Steele TO. Telemedicine, patient satisfaction, and chronic rhinosinusitis care in the era of COVID-19. Am J Rhinol Allergy 2021 Jul;35(4):494-499. [doi: 10.1177/1945892420970460] [Medline: 33115249]
166. Muroff J, Robinson W, Chassler D, López LM, Gaitan E, Lundgren L, et al. Use of a smartphone recovery tool for Latinos with co-occurring alcohol and other drug disorders and mental disorders. J Dual Diagn 2017;13(4):280-290. [doi: 10.1080/15504263.2017.1348649] [Medline: 28692420]
167. Nancarrow S, Banbury A, Buckley J. Evaluation of a national broadband network-enabled telehealth trial for older people with chronic disease. Aust Health Rev 2016 Jan;40(6):641-648. [doi: 10.1071/AH15201] [Medline: 27028234]
168. Nohra RG, Chaaban T, Sacre H, Salameh P, Aoun Bacha Z, Le Bon Chami B, et al. Evaluating the feasibility and pretesting the impact of an educational and telemonitoring program for COPD patients in Lebanon. Int J Chron Obstruct Pulmon Dis 2022;17:949-965 [FREE Full text] [doi: 10.2147/COPD.S339592] [Medline: 35510164]
169. Nyberg A, Tistad M, Wadell K. Can the COPD web be used to promote self-management in patients with COPD in swedish primary care: a controlled pragmatic pilot trial with 3 month- and 12 month follow-up. Scand J Prim Health Care 2019 Mar;37(1):69-82 [FREE Full text] [doi: 10.1080/02813432.2019.1569415] [Medline: 30700230]
170. Paldán K, Steinmetz M, Simanovski J, Rammos C, Ullrich G, Jánosi RA, et al. Supervised exercise therapy using mobile health technology in patients with peripheral arterial disease: pilot randomized controlled trial. JMIR Mhealth Uhealth 2021 Aug 16;9(8):e24214 [FREE Full text] [doi: 10.2196/24214] [Medline: 34398800]
171. Panagopoulos C, Malli F, Menychtas A, Smyrli EP, Georgountzou A, Daniil Z, et al. Utilizing a homecare platform for remote monitoring of patients with idiopathic pulmonary fibrosis. Adv Exp Med Biol 2017;989:177-187. [doi: 10.1007/978-3-319-57348-9_15] [Medline: 28971426]
172. Pariser P, Pham TN, Brown JB, Stewart M, Charles J. Connecting people with multimorbidity to interprofessional teams using telemedicine. Ann Fam Med 2019 Aug 12;17(Suppl 1):S57-S62 [FREE Full text] [doi: 10.1370/afm.2379] [Medline: 31405877]
173. Picton P, Wiljer D, Urowitz S, Cafazzo JA. Engaging patients in online self-care technologies for chronic disease management. Healthc Q 2016;18(4):55-61. [doi: 10.12927/hcq.2016.24549] [Medline: 27009709]
174. Pratap A, Grant D, Vegesna A, Tummalacherla M, Cohan S, Deshpande C, et al. Evaluating the utility of smartphone-based sensor assessments in persons with multiple sclerosis in the real-world using an app (elevateMS): observational, prospective pilot digital health study. JMIR Mhealth Uhealth 2020 Oct 27;8(10):e22108 [FREE Full text] [doi: 10.2196/22108] [Medline: 33107827]
175. Quaedackers L, De Wit J, Pillen S, Van Gilst M, Batalas N, Lammers GJ, et al. A mobile app for longterm monitoring of narcolepsy symptoms: design, development, and evaluation. JMIR Mhealth Uhealth 2020 Jan 07;8(1):e14939 [FREE Full text] [doi: 10.2196/14939] [Medline: 31909723]

XSL•FO
RenderX

176. Raghu A, Praveen D, Peiris D, Tarassenko L, Clifford G. Engineering a mobile health tool for resource-poor settings to assess and manage cardiovascular disease risk: SMARThealth study. BMC Med Inform Decis Mak 2015 Apr 29;15:36 [FREE Full text] [doi: 10.1186/s12911-015-0148-4] [Medline: 25924825]
177. Roca S, Lozano ML, García J, Alesanco Á. Validation of a virtual assistant for improving medication adherence in patients with comorbid type 2 diabetes mellitus and depressive disorder. Int J Environ Res Public Health 2021 Nov 17;18(22):12056 [FREE Full text] [doi: 10.3390/ijerph182212056] [Medline: 34831811]
178. Rollo ME, Ash S, Lyons-Wall P, Russell AW. Evaluation of a mobile phone image-based dietary assessment method in adults with type 2 diabetes. Nutrients 2015 Jun 17;7(6):4897-4910 [FREE Full text] [doi: 10.3390/nu7064897] [Medline: 26091234]
179. Rudolf I, Pieper K, Nolte H, Junge S, Dopfer C, Sauer-Heilborn A, et al. Assessment of a mobile app by adolescents and young adults with cystic fibrosis: pilot evaluation. JMIR Mhealth Uhealth 2019 Nov 21;7(11):e12442 [FREE Full text] [doi: 10.2196/12442] [Medline: 31750841]
180. Santo K, Singleton A, Rogers K, Thiagalingam A, Chalmers J, Chow CK, et al. Medication reminder applications to improve adherence in coronary heart disease: a randomised clinical trial. Heart 2019 Feb;105(4):323-329. [doi: 10.1136/heartjnl-2018-313479] [Medline: 30150326]
181. Schougaard LM, Larsen LP, Jessen A, Sidenius P, Dorflinger L, de Thurah A, et al. AmbuFlex: tele-patient-reported outcomes (telePRO) as the basis for follow-up in chronic and malignant diseases. Qual Life Res 2016 Mar;25(3):525-534 [FREE Full text] [doi: 10.1007/s11136-015-1207-0] [Medline: 26790427]
182. Sönnerfors P, Wadell K, Dohrn IM, Nyberg A, Runold M, Halvarsson A. Use of an eHealth tool for exercise training and online contact in people with severe chronic obstructive pulmonary disease on long-term oxygen treatment: a feasibility study. Health Informatics J 2020 Dec;26(4):3184-3200 [FREE Full text] [doi: 10.1177/1460458220945429] [Medline: 32880209]
183. Steare T, Giorgalli M, Free K, Harju-Seppänen J, Akther S, Eskinazi M, et al. A qualitative study of stakeholder views on the use of a digital app for supported self-management in early intervention services for psychosis. BMC Psychiatry 2021 Jun 19;21(1):311 [FREE Full text] [doi: 10.1186/s12888-021-03317-9] [Medline: 34147075]
184. Stonbraker S, Cho H, Hermosi G, Pichon A, Schnall R. Usability testing of a mHealth app to support self-management of HIV-associated non-AIDS related symptoms. Stud Health Technol Inform 2018;250:106-110 [FREE Full text] [Medline: 29857399]
185. Stroupe KT, Martinez R, Hogan TP, Evans CT, Scholten J, Bidelspach D, et al. Health care utilization and costs of veterans evaluated for traumatic brain injury through telehealth. Telemed J E Health 2019 Dec;25(12):1144-1153. [doi: 10.1089/tmj.2018.0182] [Medline: 30874499]
186. Talal AH, Markatou M, Sofikitou EM, Brown LS, Perumalswami P, Dinani A, et al. Patient-centered HCV care via telemedicine for individuals on medication for opioid use disorder: Telemedicine for Evaluation, Adherence and Medication for Hepatitis C (TEAM-C). Contemp Clin Trials 2022 Jan;112:106632 [FREE Full text] [doi: 10.1016/j.cct.2021.106632] [Medline: 34813962]
187. Talboom-Kamp EP, Holstege MS, Chavannes NH, Kasteleyn MJ. Effects of use of an eHealth platform e-Vita for COPD patients on disease specific quality of life domains. Respir Res 2019 Jul 10;20(1):146 [FREE Full text] [doi: 10.1186/s12931-019-1110-2] [Medline: 31291945]
188. Tassorelli C, Jensen R, Allena M, De Icco R, Katsarava Z, Miguel Lainez J, COMOESTAS Consortium. The added value of an electronic monitoring and alerting system in the management of medication-overuse headache: a controlled multicentre study. Cephalalgia 2017 Oct;37(12):1115-1125. [doi: 10.1177/0333102416660549] [Medline: 27440251]
189. Taylor AH, Taylor RS, Ingram WM, Anokye N, Dean S, Jolly K, et al. Adding web-based behavioural support to exercise referral schemes for inactive adults with chronic health conditions: the e-coachER RCT. Health Technol Assess 2020 Nov;24(63):1-106 [FREE Full text] [doi: 10.3310/hta24630] [Medline: 33243368]
190. Tillis W, Bond WF, Svendsen J, Guither S. Implementation of activity sensor equipment in the homes of chronic obstructive pulmonary disease patients. Telemed J E Health 2017 Nov;23(11):920-929. [doi: 10.1089/tmj.2016.0201] [Medline: 28557641]
191. Trosini-Désert V, Lafoeste H, Regard L, Malrin R, Galarza-Jimenez MA, Amarilla CE, et al. A telemedicine intervention to ensure the correct usage of inhaler devices. Telemed J E Health 2020 Nov;26(11):1336-1344. [doi: 10.1089/tmj.2019.0246] [Medline: 32302518]
192. Turino C, Benítez ID, Rafael-Palou X, Mayoral A, Lopera A, Pascual L, et al. Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system based on machine learning aiming to improve continuous positive airway pressure treatment compliance: randomized controlled trial. J Med Internet Res 2021 Oct 18;23(10):e24072 [FREE Full text] [doi: 10.2196/24072] [Medline: 34661550]
193. Tyrrell JS, Redshaw CH. Physical activity in ankylosing spondylitis: evaluation and analysis of an eHealth tool. J Innov Health Inform 2016 Jul 04;23(2):169 [FREE Full text] [doi: 10.14236/jhi.v23i2.169] [Medline: 27869581]
194. Urech C, Grossert A, Alder J, Scherer S, Handschin B, Kasenda B, et al. Web-based stress management for newly diagnosed patients with cancer (STREAM): a randomized, wait-list controlled intervention study. J Clin Oncol 2018 Mar 10;36(8):780-788 [FREE Full text] [doi: 10.1200/JCO.2017.74.8491] [Medline: 29369731]

XSL•FO
RenderX

195. Verwey R, van der Weegen S, Spreeuwenberg M, Tange H, van der Weijden T, de Witte L. Process evaluation of physical activity counselling with and without the use of mobile technology: a mixed methods study. Int J Nurs Stud 2016 Jan;53:3-16. [doi: 10.1016/j.ijnurstu.2015.10.008] [Medline: 26518108]
196. Welzel FD, Bär J, Stein J, Löbner M, Pabst A, Luppa M, et al. Using a brief web-based 5A intervention to improve weight management in primary care: results of a cluster-randomized controlled trial. BMC Fam Pract 2021 Apr 02;22(1):61 [FREE Full text] [doi: 10.1186/s12875-021-01404-0] [Medline: 33794781]
197. Wheeler TS, Michael Vallis T, Giacomantonio NB, Abidi SR. Feasibility and usability of an ontology-based mobile intervention for patients with hypertension. Int J Med Inform 2018 Nov;119:8-16. [doi: 10.1016/j.ijmedinf.2018.08.002] [Medline: 30342690]
198. Zetterqvist V, Gentili C, Rickardsson J, Sörensen I, Wicksell RK. Internet-delivered acceptance and commitment therapy for adolescents with chronic pain and their parents: a nonrandomized pilot trial. J Pediatr Psychol 2020 Oct 01;45(9):990-1004. [doi: 10.1093/jpepsy/jsaa060] [Medline: 32974656]
199. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
200. Ahmadvand A, Kavanagh D, Clark M, Drennan J, Nissen L. Trends and visibility of "digital health" as a keyword in articles by JMIR publications in the new millennium: bibliographic-bibliometric analysis. J Med Internet Res 2019 Dec 19;21(12):e10477 [FREE Full text] [doi: 10.2196/10477] [Medline: 31855190]
201. Countries. World Health Organization (WHO). 2022. URL: https://www.who.int/countries [accessed 2022-10-29]
202. Waltz TJ, Powell BJ, Matthieu MM, Damschroder LJ, Chinman MJ, Smith JL, et al. Use of concept mapping to characterize relationships among implementation strategies and assess their feasibility and importance: results from the Expert Recommendations for Implementing Change (ERIC) study. Implement Sci 2015 Aug 07;10:109 [FREE Full text] [doi: 10.1186/s13012-015-0295-0] [Medline: 26249843]
203. Grol R, Wensing M. What drives change? Barriers to and incentives for achieving evidence-based practice. Med J Aust 2004 Mar 15;180(S6):S57-S60. [doi: 10.5694/j.1326-5377.2004.tb05948.x] [Medline: 15012583]
204. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
205. Nordin C, Michaelson P, Eriksson MK, Gard G. It's about me: patients' experiences of patient participation in the web behavior change program for activity in combination with multimodal pain rehabilitation. J Med Internet Res 2017 Jan 18;19(1):e22 [FREE Full text] [doi: 10.2196/jmir.5970] [Medline: 28100440]
206. Duan H, Wang Z, Ji Y, Ma L, Liu F, Chi M, et al. Using goal-directed design to create a mobile health app to improve patient compliance with hypertension self-management: development and deployment. JMIR Mhealth Uhealth 2020 Feb 25;8(2):e14466 [FREE Full text] [doi: 10.2196/14466] [Medline: 32130161]
207. Infarinato F, Jansen-Kosterink S, Romano P, van Velsen L, Op den Akker H, Rizza F, et al. Acceptance and potential impact of the eWALL platform for health monitoring and promotion in persons with a chronic disease or age-related impairment. Int J Environ Res Public Health 2020 Oct 28;17(21):7893 [FREE Full text] [doi: 10.3390/ijerph17217893] [Medline: 33126506]
208. Kwok JY, Lee JJ, Choi EP, Chau PH, Auyeung M. Stay mindfully active during the coronavirus pandemic: a feasibility study of mHealth-delivered mindfulness yoga program for people with Parkinson's disease. BMC Complement Med Ther 2022 Feb 07;22(1):37 [FREE Full text] [doi: 10.1186/s12906-022-03519-y] [Medline: 35130894]
209. Mucchi L, Jayousi S, Gant A, Paoletti E, Zoppi P. Tele-monitoring system for chronic diseases management: requirements and architecture. Int J Environ Res Public Health 2021 Jul 13;18(14):7459 [FREE Full text] [doi: 10.3390/ijerph18147459] [Medline: 34299910]
210. Sherwin LB, Deroche CB, Yevu-Johnson J, Matteson-Kome M, Bechtold M, Jahnke I, et al. Usability evaluation of a smartphone medication reminder application in patients treated with short-term antibiotic. Comput Inform Nurs 2021 Apr 30;39(10):547-553. [doi: 10.1097/CIN.0000000000000747] [Medline: 33935203]
211. Tucker S, Abbott L, Anderson R, Eppen K, Laroche H, Paelmo E, et al. Implementing follow-along physical activity videos with people living with chronic conditions: a feasibility study. Worldviews Evid Based Nurs 2019 Oct;16(5):352-361. [doi: 10.1111/wvn.12392] [Medline: 31380602]
212. Vatnøy TK, Thygesen E, Dale B. Telemedicine to support coping resources in home-living patients diagnosed with chronic obstructive pulmonary disease: Patients' experiences. J Telemed Telecare 2017 Jan;23(1):126-132. [doi: 10.1177/1357633X15626854] [Medline: 26853926]
213. Lundell S, Modig M, Holmner Å, Wadell K. Perceptions of home telemonitoring use among patients with chronic obstructive pulmonary disease: qualitative study. JMIR Mhealth Uhealth 2020 Jun 03;8(6):e16343 [FREE Full text] [doi: 10.2196/16343] [Medline: 32490844]
214. Vest BM, Hall VM, Kahn LS, Heider AR, Maloney N, Singh R. Nurse perspectives on the implementation of routine telemonitoring for high-risk diabetes patients in a primary care setting. Prim Health Care Res Dev 2017 Jan;18(1):3-13. [doi: 10.1017/S1463423616000190] [Medline: 27269513]

XSL•FO
RenderX

215. Vasi S, Advocat J, Adaji A, Russell G. Building quality chronic illness care: implementation of a web-based care plan. Aust J Prim Health 2020 Apr;26(2):173-177. [doi: 10.1071/PY19039] [Medline: 32027816]
216. Floch J, Vilarinho T, Zettl A, Ibanez-Sanchez G, Calvo-Lerma J, Stav E, et al. Users' experiences of a mobile health self-management approach for the treatment of cystic fibrosis: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 08;8(7):e15896 [FREE Full text] [doi: 10.2196/15896] [Medline: 32673237]
217. Zhang J, Mihai C, Tüshaus L, Scebba G, Distler O, Karlen W. Wound image quality from a mobile health tool for home-based chronic wound management with real-time quality feedback: randomized feasibility study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e26149 [FREE Full text] [doi: 10.2196/26149] [Medline: 34328440]
218. Prevodnik K, Hvalič-Touzery S, Dolničar V, Zaletel J, Laznik J, Petrovčič A. Experience of patients with chronic conditions with telemedicine in primary care: a focus group analysis. Obzor Zdrav Neg 2022 Dec 17;56(4):246-263. [doi: 10.14528/snr.2022.56.4.3150]
219. Lehmann J, Buhl P, Giesinger JM, Wintner LM, Sztankay M, Neppl L, et al. Using the computer-based health evaluation system (CHES) to support self-management of symptoms and functional health: evaluation of hematological patient use of a web-based patient portal. J Med Internet Res 2021 Jun 08;23(6):e26022 [FREE Full text] [doi: 10.2196/26022] [Medline: 34100765]
220. Ali HI, Attlee A, Alhebshi S, Elmi F, Al Dhaheri AS, Stojanovska L, et al. Feasibility study of a newly developed technology-mediated lifestyle intervention for overweight and obese young adults. Nutrients 2021 Jul 26;13(8):2547 [FREE Full text] [doi: 10.3390/nu13082547] [Medline: 34444707]
221. Amann J, Fiordelli M, Brach M, Bertschy S, Scheel-Sailer A, Rubinelli S. Co-designing a self-management app prototype to support people with spinal cord injury in the prevention of pressure injuries: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 09;8(7):e18018 [FREE Full text] [doi: 10.2196/18018] [Medline: 32673241]
222. Lambert J, Taylor A, Streeter A, Greaves C, Ingram WM, Dean S, et al. A process evaluation, with mediation analysis, of a web-based intervention to augment primary care exercise referral schemes: the e-coachER randomised controlled trial. Int J Behav Nutr Phys Act 2022 Sep 29;19(1):128 [FREE Full text] [doi: 10.1186/s12966-022-01360-7] [Medline: 36175996]
223. Morita PP, Yeung MS, Ferrone M, Taite AK, Madeley C, Stevens Lavigne A, et al. A patient-centered mobile health system that supports asthma self-management (breathe): design, development, and utilization. JMIR Mhealth Uhealth 2019 Jan 28;7(1):e10956 [FREE Full text] [doi: 10.2196/10956] [Medline: 30688654]
224. Sloots J, Bakker M, van der Palen J, Eijsvogel M, van der Valk P, Linssen G, et al. Adherence to an eHealth self-management intervention for patients with both COPD and heart failure: results of a pilot study. Int J Chron Obstruct Pulmon Dis 2021;16:2089-2103 [FREE Full text] [doi: 10.2147/COPD.S299598] [Medline: 34290502]
225. Smithson R, Roche E, Wicker C. Virtual models of chronic disease management: lessons from the experiences of virtual care during the COVID-19 response. Aust Health Rev 2021 Jun;45(3):311-316. [doi: 10.1071/AH20190] [Medline: 33583487]
226. Sobrinho A, da Silva LD, Perkusich A, Pinheiro ME, Cunha P. Design and evaluation of a mobile application to assist the self-monitoring of the chronic kidney disease in developing countries. BMC Med Inform Decis Mak 2018 Jan 12;18(1):7 [FREE Full text] [doi: 10.1186/s12911-018-0587-9] [Medline: 29329530]
227. Souza-Silva MV, Domingues ML, Chagas VS, Pereira DN, de Sá LC, Almeida MS, et al. Implementation of a text messaging intervention to patients on warfarin therapy in Brazilian primary care units: a quasi-experimental study. BMC Prim Care 2022 Mar 23;23(1):54 [FREE Full text] [doi: 10.1186/s12875-022-01647-5] [Medline: 35321654]
228. Le Goff-Pronost M, Mourgeon B, Blanchère JP, Teot L, Benateau H, Dompmartin A. Real-world clinical evaluation and costs of telemedicine for chronic wound management. Int J Technol Assess Health Care 2018 Jan;34(6):567-575. [doi: 10.1017/S0266462318000685] [Medline: 30369340]
229. Ong SW, Kaushal A, Pariser P, Chan CT. An integrated kidney care eConsult practice model: results from the iKinect project. Am J Nephrol 2019;50(4):262-271. [doi: 10.1159/000502602] [Medline: 31473734]
230. Herrmann S, Power B, Rashidi A, Cypher M, Mastaglia F, Grace A, et al. Supporting patient-clinician interaction in chronic HIV care: design and development of a patient-reported outcomes software application. J Med Internet Res 2021 Jul 30;23(7):e27861 [FREE Full text] [doi: 10.2196/27861] [Medline: 34328442]
231. Jakubowski KP, Jhamb M, Yabes J, Gujral S, Oberlin LE, Bender FH, et al. Technology-assisted cognitive-behavioral therapy intervention for end-stage renal disease. Transl Behav Med 2020 Aug 07;10(3):657-663 [FREE Full text] [doi: 10.1093/tbm/ibz077] [Medline: 31131853]
232. Carlsen K, Jakobsen C, Houen G, Kallemose T, Paerregaard A, Riis LB, et al. Self-managed eHealth disease monitoring in children and adolescents with inflammatory bowel disease: a randomized controlled trial. Inflamm Bowel Dis 2017 Mar;23(3):357-365. [doi: 10.1097/MIB.0000000000001026] [Medline: 28221247]
233. Katz IJ, Pirabhahar S, Williamson P, Raghunath V, Brennan F, O'Sullivan A, et al. iConnect CKD - virtual medical consulting: a web-based chronic kidney disease, hypertension and diabetes integrated care program. Nephrology (Carlton) 2018 Jul;23(7):646-652. [doi: 10.1111/nep.13070] [Medline: 28474361]
234. Nagel T, Sweet M, Dingwall KM, Puszka S, Hughes JT, Kavanagh DJ, et al. Adapting wellbeing research tools for Aboriginal and Torres strait islander people with chronic kidney disease. BMC Nephrol 2020 Apr 15;21(1):130 [FREE Full text] [doi: 10.1186/s12882-020-01776-y] [Medline: 32293331]

XSL•FO
RenderX

235. Eisner E, Drake RJ, Berry N, Barrowclough C, Emsley R, Machin M, et al. Development and long-term acceptability of ExPRESS, a mobile phone app to monitor basic symptoms and early signs of psychosis relapse. JMIR Mhealth Uhealth 2019 Mar 29;7(3):e11568 [FREE Full text] [doi: 10.2196/11568] [Medline: 30924789]
236. Hicks TA, Thomas SP, Wilson SM, Calhoun PS, Kuhn ER, Beckham JC. A preliminary investigation of a relapse prevention mobile application to maintain smoking abstinence among individuals with posttraumatic stress disorder. J Dual Diagn 2017;13(1):15-20 [FREE Full text] [doi: 10.1080/15504263.2016.1267828] [Medline: 27918881]
237. Kock A, Kaya RS, Müller C, Andersen B, Langer T, Ingenerf J. Design, implementation, and evaluation of a mobile application for patient empowerment and management of long-term follow-up after childhood cancer. Klin Padiatr 2015 May;227(3):166-170. [doi: 10.1055/s-0035-1548840] [Medline: 25985450]
238. Reading Turchioe M, Grossman LV, Baik D, Lee CS, Maurer MS, Goyal P, et al. Older adults can successfully monitor symptoms using an inclusively designed mobile application. J Am Geriatr Soc 2020 Jun;68(6):1313-1318 [FREE Full text] [doi: 10.1111/jgs.16403] [Medline: 32157679]
239. Sumino K, Locke ER, Magzamen S, Gylys-Colwell I, Humblet O, Nguyen HQ, et al. Use of a remote inhaler monitoring device to measure change in inhaler use with chronic obstructive pulmonary disease exacerbations. J Aerosol Med Pulm Drug Deliv 2018 Jun;31(3):191-198. [doi: 10.1089/jamp.2017.1383] [Medline: 29035120]
240. Zahid M, Gallant NL, Hadjistavropoulos T, Stroulia E. Behavioral pain assessment implementation in long-term care using a tablet app: case series and quasi-experimental design. JMIR Mhealth Uhealth 2020 Apr 22;8(4):e17108 [FREE Full text] [doi: 10.2196/17108] [Medline: 32319955]
241. Flujas-Contreras JM, Ruiz-Castañeda D, Gómez I. Promoting emotional well-being in hospitalized children and adolescents with virtual reality: usability and acceptability of a randomized controlled trial. Comput Inform Nurs 2020 Feb;38(2):99-107. [doi: 10.1097/CIN.0000000000000586] [Medline: 31895058]
242. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 03;8(4):e15806 [FREE Full text] [doi: 10.2196/15806] [Medline: 32242820]
243. Jain YS, Garg A, Jhamb DK, Jain P, Karar A. Preparing India to leverage power of mobile technology: development of a bilingual mobile health tool for heart patients. Cardiovasc Hematol Agents Med Chem 2019;17(2):125-134. [doi: 10.2174/1871525717666190912152938] [Medline: 31512999]
244. Knox L, Dunning M, Davies CA, Mills-Bennet R, Sion TW, Phipps K, et al. Safety, feasibility, and effectiveness of virtual pulmonary rehabilitation in the real world. Int J Chron Obstruct Pulmon Dis 2019;14:775-780 [FREE Full text] [doi: 10.2147/COPD.S193827] [Medline: 31040656]
245. Ahmedani BK, Crotty N, Abdulhak MM, Ondersma SJ. Pilot feasibility study of a brief, tailored mobile health intervention for depression among patients with chronic pain. Behav Med 2015;41(1):25-32. [doi: 10.1080/08964289.2013.867827] [Medline: 24313728]
246. Amorim AB, Pappas E, Simic M, Ferreira ML, Jennings M, Tiedemann A, et al. Integrating mobile-health, health coaching, and physical activity to reduce the burden of chronic low back pain trial (IMPACT): a pilot randomised controlled trial. BMC Musculoskelet Disord 2019 Feb 11;20(1):71 [FREE Full text] [doi: 10.1186/s12891-019-2454-y] [Medline: 30744606]
247. Andrén P, Aspvall K, Fernández de la Cruz L, Wiktor P, Romano S, Andersson E, et al. Therapist-guided and parent-guided internet-delivered behaviour therapy for paediatric Tourette's disorder: a pilot randomised controlled trial with long-term follow-up. BMJ Open 2019 Feb 15;9(2):e024685 [FREE Full text] [doi: 10.1136/bmjopen-2018-024685] [Medline: 30772854]
248. Cooke M, Richards J, Tjondronegoro D, Raj Chakraborty P, Jauncey-Cooke J, Andresen E, et al. myPainPal: Co-creation of a mHealth app for the management of chronic pain in young people. Inform Health Soc Care 2021 Sep 02;46(3):291-305. [doi: 10.1080/17538157.2021.1892697] [Medline: 33784952]
249. Crawford MR, Luik AI, Espie CA, Taylor HL, Burgess HJ, Jones AL, Rush University Sleep Research Team, et al. Digital cognitive behavioral therapy for insomnia in women with chronic migraines. Headache 2020 May;60(5):902-915 [FREE Full text] [doi: 10.1111/head.13777] [Medline: 32112436]
250. de Jong M, van der Meulen-de Jong A, Romberg-Camps M, Degens J, Becx M, Markus T, et al. Development and feasibility study of a telemedicine tool for all patients with IBD: MyIBDcoach. Inflamm Bowel Dis 2017 Apr;23(4):485-493. [doi: 10.1097/MIB.0000000000001034] [Medline: 28267047]
251. Imtiaz R, Atkinson K, Guerinet J, Wilson K, Leidecker J, Zimmerman D. A pilot study of OkKidney, a phosphate counting application in patients on peritoneal dialysis. Perit Dial Int 2017;37(6):613-618. [doi: 10.3747/pdi.2017.00050] [Medline: 28970367]
252. Martínez P, Guajardo V, Gómez VE, Brandt S, Szabo W, Soto-Brandt G, et al. Technology-assisted collaborative care program for people with diabetes and/or high blood pressure attending primary health care: a feasibility study. Int J Environ Res Public Health 2021 Nov 15;18(22):12000 [FREE Full text] [doi: 10.3390/ijerph182212000] [Medline: 34831756]
253. Maxwell LG, McFarland MS, Baker JW, Cassidy RF. Evaluation of the impact of a pharmacist-led telehealth clinic on diabetes-related goals of therapy in a veteran population. Pharmacotherapy 2016 Mar;36(3):348-356. [doi: 10.1002/phar.1719] [Medline: 26877253]

XSL•FO
RenderX

254. Ng G, Tan N, Bahadin J, Shum E, Tan SW. Development of an automated healthcare kiosk for the management of chronic disease patients in the primary care setting. J Med Syst 2016 Jul;40(7):169. [doi: 10.1007/s10916-016-0529-y] [Medline: 27240840]
255. Or C, Tao D. A 3-month randomized controlled pilot trial of a patient-centered, computer-based self-monitoring system for the care of type 2 diabetes mellitus and hypertension. J Med Syst 2016 Apr;40(4):81. [doi: 10.1007/s10916-016-0437-1] [Medline: 26802011]
256. Zand A, Nguyen A, Reynolds C, Khandadash A, Esrailian E, Hommes D. Patient experience and satisfaction with an e-Health care management application for inflammatory bowel diseases. Int J Environ Res Public Health 2021 Nov 09;18(22):11747 [FREE Full text] [doi: 10.3390/ijerph182211747] [Medline: 34831502]
257. Wallace T, Morris JT, Glickstein R, Anderson RK, Gore RK. Implementation of a mobile technology-supported diaphragmatic breathing intervention in military mTBI with PTSD. J Head Trauma Rehabil 2022;37(3):152-161 [FREE Full text] [doi: 10.1097/HTR.0000000000000774] [Medline: 35703895]
258. Deng N, Chen J, Liu Y, Wei S, Sheng L, Lu R, et al. Using mobile health technology to deliver a community-based closed-loop management system for chronic obstructive pulmonary disease patients in remote areas of China: development and prospective observational study. JMIR Mhealth Uhealth 2020 Nov 25;8(11):e15978 [FREE Full text] [doi: 10.2196/15978] [Medline: 33237036]
259. Ho K, Newton L, Boothe A, Novak-Lauscher H. Mobile digital access to a web-enhanced network (mDAWN): assessing the feasibility of mobile health tools for self-management of type-2 diabetes. AMIA Annu Symp Proc 2015;2015:621-629 [FREE Full text] [Medline: 26958197]
260. Khan F, Granville N, Malkani R, Chathampally Y. Health-related quality of life improvements in systemic lupus erythematosus derived from a digital therapeutic plus tele-health coaching intervention: randomized controlled pilot trial. J Med Internet Res 2020 Oct 20;22(10):e23868 [FREE Full text] [doi: 10.2196/23868] [Medline: 33079070]
261. Lewis A, Knight E, Bland M, Middleton J, Mitchell E, McCrum K, et al. Feasibility of an online platform delivery of pulmonary rehabilitation for individuals with chronic respiratory disease. BMJ Open Respir Res 2021 Mar;8(1):33762360 [FREE Full text] [doi: 10.1136/bmjresp-2021-000880] [Medline: 33762360]
262. Niendam TA, Tully LM, Iosif AM, Kumar D, Nye KE, Denton JC, et al. Enhancing early psychosis treatment using smartphone technology: a longitudinal feasibility and validity study. J Psychiatr Res 2018 Jan;96:239-246. [doi: 10.1016/j.jpsychires.2017.10.017] [Medline: 29126059]
263. Passardi A, Foca F, Caffo O, Tondini CA, Zambelli A, Vespignani R, et al. A remote monitoring system to optimize the home management of oral anticancer therapies (ONCO-TreC): prospective training-validation trial. J Med Internet Res 2022 Jan 26;24(1):e27349 [FREE Full text] [doi: 10.2196/27349] [Medline: 35080505]
264. Easton K, Potter S, Bec R, Bennion M, Christensen H, Grindell C, et al. A virtual agent to support individuals living with physical and mental comorbidities: co-design and acceptability testing. J Med Internet Res 2019 May 30;21(5):e12996 [FREE Full text] [doi: 10.2196/12996] [Medline: 31148545]
265. Rudin RS, Fanta CH, Qureshi N, Duffy E, Edelen MO, Dalal AK, et al. A clinically integrated mHealth app and practice model for collecting patient-reported outcomes between visits for asthma patients: implementation and feasibility. Appl Clin Inform 2019 Oct;10(5):783-793 [FREE Full text] [doi: 10.1055/s-0039-1697597] [Medline: 31618782]
266. Talal AH, Jaanimägi U, Davis K, Bailey J, Bauer BM, Dharia A, et al. Facilitating engagement of persons with opioid use disorder in treatment for hepatitis C virus infection via telemedicine: stories of onsite case managers. J Subst Abuse Treat 2021 Aug;127:108421. [doi: 10.1016/j.jsat.2021.108421] [Medline: 34134875]
267. Frederix I, Hansen D, Coninx K, Vandervoort P, Vandijck D, Hens N, et al. Medium-term effectiveness of a comprehensive internet-based and patient-specific telerehabilitation program with text messaging support for cardiac patients: randomized controlled trial. J Med Internet Res 2015 Jul 23;17(7):e185 [FREE Full text] [doi: 10.2196/jmir.4799] [Medline: 26206311]
268. Puig J, Echeverría P, Lluch T, Herms J, Estany C, Bonjoch A, et al. A specific mobile health application for older HIV-infected patients: usability and patient's satisfaction. Telemed J E Health 2021 Apr;27(4):432-440. [doi: 10.1089/tmj.2020.0098] [Medline: 32667858]
269. Peterson S. Telerehabilitation booster sessions and remote patient monitoring in the management of chronic low back pain: a case series. Physiother Theory Pract 2018 May;34(5):393-402. [doi: 10.1080/09593985.2017.1401190] [Medline: 29125371]
270. Donald M, Beanlands H, Straus S, Smekal M, Gil S, Elliott MJ, et al. An eHealth self-management intervention for adults with chronic kidney disease, My Kidneys My Health: a mixed-methods study. CMAJ Open 2022;10(3):E746-E754 [FREE Full text] [doi: 10.9778/cmajo.20210332] [Medline: 35973709]
271. Banerjee A, Ramanujan RA, Agnihothri S. Mobile health monitoring: development and implementation of an app in a diabetes and hypertension clinic. In: Proceedings of the 49th Hawaii International Conference on System Sciences. 2016 Presented at: HICSS '16; January 5-8, 2016; Koloa, HI p. 5-8 URL: https://ieeexplore.ieee.org/document/7427611 [doi: 10.1109/hicss.2016.427]
272. Guo X, Yang Y, Takiff HE, Zhu M, Ma J, Zhong T, et al. A comprehensive app that improves tuberculosis treatment management through video-observed therapy: usability study. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e17658 [FREE Full text] [doi: 10.2196/17658] [Medline: 32735222]

XSL•FO
RenderX

273. Sjöström M, Umefjord G, Stenlund H, Carlbring P, Andersson G, Samuelsson E. Internet-based treatment of stress urinary incontinence: 1- and 2-year results of a randomized controlled trial with a focus on pelvic floor muscle training. BJU Int 2015 Dec;116(6):955-964 [FREE Full text] [doi: 10.1111/bju.13091] [Medline: 25683075]
274. Song CE, An M. The self-management smartphone application for cancer survivors, ReLive: development and usability testing. Comput Inform Nurs 2021 Jun;39(6):312-320. [doi: 10.1097/CIN.0000000000000690] [Medline: 33214389]
275. Choi YH, Park HK, Paik N. A telerehabilitation approach for chronic aphasia following stroke. Telemed J E Health 2016 May;22(5):434-440. [doi: 10.1089/tmj.2015.0138] [Medline: 26544535]
276. Cella M, Okruszek Ł, Lawrence M, Zarlenga V, He Z, Wykes T. Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. Schizophr Res 2018 May;195:537-542 [FREE Full text] [doi: 10.1016/j.schres.2017.09.028] [Medline: 28986005]
277. Looman WS, Antolick M, Cady RG, Lunos SA, Garwick AE, Finkelstein SM. Effects of a telehealth care coordination intervention on perceptions of health care by caregivers of children with medical complexity: a randomized controlled trial. J Pediatr Health Care 2015;29(4):352-363 [FREE Full text] [doi: 10.1016/j.pedhc.2015.01.007] [Medline: 25747391]
278. Jonathan G, Carpenter-Song EA, Brian RM, Ben-Zeev D. Life with FOCUS: a qualitative evaluation of the impact of a smartphone intervention on people with serious mental illness. Psychiatr Rehabil J 2019 Jun;42(2):182-189 [FREE Full text] [doi: 10.1037/prj0000337] [Medline: 30589278]
279. Pludwinski S, Ahmad F, Wayne N, Ritvo P. Participant experiences in a smartphone-based health coaching intervention for type 2 diabetes: a qualitative inquiry. J Telemed Telecare 2016 Apr;22(3):172-178. [doi: 10.1177/1357633X15595178] [Medline: 26199275]
280. Ahmed S, Ernst P, Bartlett SJ, Valois M, Zaihra T, Paré G, et al. The effectiveness of web-based asthma self-management system, my asthma portal (MAP): a pilot randomized controlled trial. J Med Internet Res 2016 Dec 01;18(12):e313 [FREE Full text] [doi: 10.2196/jmir.5866] [Medline: 27908846]
281. Donesky D, Selman L, McDermott K, Citron T, Howie-Esquivel J. Evaluation of the feasibility of a home-based TeleYoga intervention in participants with both chronic obstructive pulmonary disease and heart failure. J Altern Complement Med 2017 Sep;23(9):713-721. [doi: 10.1089/acm.2015.0279] [Medline: 28654302]
282. Claes J, Cornelissen V, McDermott C, Moyna N, Pattyn N, Cornelis N, et al. Feasibility, acceptability, and clinical effectiveness of a technology-enabled cardiac rehabilitation platform (physical activity toward health-I): randomized controlled trial. J Med Internet Res 2020 Feb 04;22(2):e14221 [FREE Full text] [doi: 10.2196/14221] [Medline: 32014842]
283. Selman L, McDermott K, Donesky D, Citron T, Howie-Esquivel J. Appropriateness and acceptability of a Tele-Yoga intervention for people with heart failure and chronic obstructive pulmonary disease: qualitative findings from a controlled pilot study. BMC Complement Altern Med 2015 Feb 07;15:21 [FREE Full text] [doi: 10.1186/s12906-015-0540-8] [Medline: 25887324]
284. Kim A, Yun SJ, Sung KS, Kim Y, Jo JY, Cho H, et al. Exercise management using a mobile app in patients with Parkinsonism: prospective, open-label, single-arm pilot study. JMIR Mhealth Uhealth 2021 Aug 31;9(8):e27662 [FREE Full text] [doi: 10.2196/27662] [Medline: 34463635]
285. Peters D, Davis S, Calvo RA, Sawyer SM, Smith L, Foster JM. Young people's preferences for an asthma self-management app highlight psychological needs: a participatory study. J Med Internet Res 2017 Apr 11;19(4):e113 [FREE Full text] [doi: 10.2196/jmir.6994] [Medline: 28400353]
286. Thomas RM, Locke ER, Woo DM, Nguyen EH, Press VG, Layouni TA, et al. Inhaler training delivered by internet-based home videoconferencing improves technique and quality of life. Respir Care 2017 Nov;62(11):1412-1422 [FREE Full text] [doi: 10.4187/respcare.05445] [Medline: 28720676]
287. Wickerson L, Helm D, Gottesman C, Rozenberg D, Singer LG, Keshavjee S, et al. Telerehabilitation for lung transplant candidates and recipients during the COVID-19 pandemic: program evaluation. JMIR Mhealth Uhealth 2021 Jun 17;9(6):e28708 [FREE Full text] [doi: 10.2196/28708] [Medline: 34048354]
288. Ajayi TA, Salongo L, Zang Y, Wineinger N, Steinhubl S. Mobile health-collected biophysical markers in children with serious illness-related pain. J Palliat Med 2021 Apr;24(4):580-588 [FREE Full text] [doi: 10.1089/jpm.2020.0234] [Medline: 33351729]
289. Glattacker M, Boeker M, Anger R, Reichenbach F, Tassoni A, Bredenkamp R, et al. Evaluation of a mobile phone app for patients with pollen-related allergic rhinitis: prospective longitudinal field study. JMIR Mhealth Uhealth 2020 Apr 17;8(4):e15514 [FREE Full text] [doi: 10.2196/15514] [Medline: 32301735]
290. Seneviratne MG, Hersch F, Peiris DP. HealthNavigator: a mobile application for chronic disease screening and linkage to services at an urban primary health network. Aust J Prim Health 2018 May;24(2):116-122. [doi: 10.1071/PY17070] [Medline: 29576044]
291. Krkoska P, Vlazna D, Sladeckova M, Minarikova J, Barusova T, Batalik L, et al. Adherence and effect of home-based rehabilitation with telemonitoring support in patients with chronic non-specific low back pain: a pilot study. Int J Environ Res Public Health 2023 Jan 13;20(2):1504 [FREE Full text] [doi: 10.3390/ijerph20021504] [Medline: 36674258]
292. Ayatollahi H, Hasannezhad M, Fard HS, Haghighi MK. Type 1 diabetes self-management: developing a web-based telemedicine application. Health Inf Manag 2016 Apr;45(1):16-26. [doi: 10.1177/1833358316639456] [Medline: 28691565]

XSL•FO
RenderX

293. Ware P, Shah A, Ross HJ, Logan AG, Segal P, Cafazzo JA, et al. Challenges of telemonitoring programs for complex chronic conditions: randomized controlled trial with an embedded qualitative study. J Med Internet Res 2022 Jan 26;24(1):e31754 [FREE Full text] [doi: 10.2196/31754] [Medline: 35080502]
294. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
295. Kislov R, Pope C, Martin GP, Wilson PM. Harnessing the power of theorising in implementation science. Implement Sci 2019 Dec 11;14(1):103 [FREE Full text] [doi: 10.1186/s13012-019-0957-4] [Medline: 31823787]
296. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. Int J Qual Health Care 2003 Jun;15(3):261-266. [doi: 10.1093/intqhc/mzg031] [Medline: 12803354]
297. Willis VC, Thomas Craig KJ, Jabbarpour Y, Scheufele EL, Arriaga YE, Ajinkya M, et al. Digital health interventions to enhance prevention in primary care: scoping review. JMIR Med Inform 2022 Jan 21;10(1):e33518 [FREE Full text] [doi: 10.2196/33518] [Medline: 35060909]
298. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 06;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
299. Lim S, Tan A, Madden S, Hill B. Health professionals' and postpartum women's perspectives on digital health interventions for lifestyle management in the postpartum period: a systematic review of qualitative studies. Front Endocrinol (Lausanne) 2019;10:767 [FREE Full text] [doi: 10.3389/fendo.2019.00767] [Medline: 31787932]
300. Grande D, Luna Marti X, Feuerstein-Simon R, Merchant RM, Asch DA, Lewson A, et al. Health policy and privacy challenges associated with digital technology. JAMA Netw Open 2020 Jul 01;3(7):e208285 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.8285] [Medline: 32644138]
301. Paul M, Maglaras L, Ferrag MA, Almomani I. Digitization of healthcare sector: a study on privacy and security concerns. ICT Express 2023 Aug;9(4):571-588. [doi: 10.1016/j.icte.2023.02.007]
302. Proctor EK, Bunger AC, Lengnick-Hall R, Gerke DR, Martin JK, Phillips RJ, et al. Ten years of implementation outcomes research: a scoping review. Implement Sci 2023 Jul 25;18(1):31 [FREE Full text] [doi: 10.1186/s13012-023-01286-z] [Medline: 37491242]
303. Pinnock H, Barwick M, Carpenter CR, Eldridge S, Grandes G, Griffiths CJ, StaRI Group. Standards for Reporting Implementation Studies (StaRI): explanation and elaboration document. BMJ Open 2017 Apr 03;7(4):e013318. [doi: 10.1136/bmjopen-2016-013318] [Medline: 28373250]

Abbreviations

COM-B: capability, opportunity, motivation, behavior
ERIC: Expert Recommendations for Implementing Change
ICT: information and communication technology
mHealth: mobile health
PRESS: Peer Review of Electronic Search Strategies
StaRI: Standards for Reporting Implementation Studies

Edited by N Cahill; submitted 11.10.23; peer-reviewed by A Bucher, N Gudi, KY Hsieh; comments to author 10.02.24; revised version received 26.03.24; accepted 28.10.24; published 12.12.24.

Please cite as:
Pong C, Tseng RMWW, Tham YC, Lum E
Current Implementation of Digital Health in Chronic Disease Management: Scoping Review
J Med Internet Res 2024;26:e53576
URL: https://www.jmir.org/2024/1/e53576
doi:10.2196/53576
PMID:

©Candelyn Pong, Rachel Marjorie Wei Wen Tseng, Yih Chung Tham, Elaine Lum. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Prehabilitation With Advanced Technologies in Patients With Musculoskeletal Diseases Waiting for Surgery: Systematic Review and Meta-Analysis

Stefania Guida[1,2], MSc; Jacopo Antonino Vitale[3], PhD; Eva Swinnen[2], Prof Dr; David Beckwée[2,4], Prof Dr; Silvia Bargeri[1], MSc; Federico Pennestrì[5], PhD; Greta Castellini[1*], PhD; Silvia Gianola[1*], PhD

[1]Unità di Epidemiologia Clinica, IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

[2]Rehabilitation Research Group (RERE), Department of Physiotherapy, Human Physiology and Anatomy, Vrije Universiteit Brussel, Brussel, Belgium

[3]Spine Center, Schulthess Klinik, Zurich, Switzerland

[4]Department of Rehabilitation Sciences and Physiotherapy, University of Antwerp, Antwerpen, Belgium

[5]Direzione Scientifica, IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

[*]these authors contributed equally

Corresponding Author:
Stefania Guida, MSc
Unità di Epidemiologia Clinica
IRCCS Istituto Ortopedico Galeazzi
Via Cristina Belgioioso, 173
Milan, 20157
Italy
Phone: 39 0283500010
Email: stefania.guida@grupposandonato.it

Abstract

Background: Prehabilitation delivered with advanced technologies represents a great chance for patients to optimize pre- and postoperative outcomes, reduce costs, and overcome travel-related barriers.

Objective: We aim to evaluate the effects of prehabilitation delivered with advanced technologies on clinically relevant outcomes among patients affected by musculoskeletal diseases and waiting for surgery.

Methods: We searched the PubMed, EMBASE, Cochrane Library, PEDro, and CINAHL databases up to February 2, 2023. ClinicalTrials.gov was also searched for registered protocols. Randomized controlled trials and nonrandomized intervention studies with adult participants of both sexes, affected by any musculoskeletal disease, and undergoing prehabilitation with advanced technologies or standard care were included. Study selection, data extraction, and critical appraisal were conducted in duplicate. Data were pooled for meta-analysis using random-effects models. Certainty of evidence was assessed for the primary outcome with the GRADE (Grading of Recommendations Assessment, Development and Evaluation) system. The primary outcome was function. Secondary outcomes were pain, strength, risk of fall, autonomy in the activities of daily living, patient satisfaction, health-related quality of life, adverse events, and adherence to treatment.

Results: Six studies (7 reports), focusing on patients undergoing total knee or hip arthroplasty and primary meniscal tear and spine surgery were included. We found different prehabilitation programs: mindfulness-based stress reduction, exercise, education, or a combination thereof. Prehabilitation delivered with advanced technologies proved to be more effective in improving function in candidates for knee or hip replacement (Western Ontario McMaster Osteoarthritis Index "function" subscale before surgery: mean difference [MD] –7.45, 95% CI –10.71 to –4.19; I^2=0%; after surgery: MD –7.84, 95% CI –11.80 to –3.88; I^2=75.3%), preoperative pain (MD –1.67, 95% CI –2.50 to –0.48; I^2=0%), risk of fall (MD –2.54, 95% CI –3.62 to –1.46; I^2=0%), and postoperative stiffness (MD –2.00, 95% CI –2.01 to –1.99; I^2=87%). No differences were found in pain 1 month after surgery. Data from studies including participants undergoing primary meniscal tear and spinal surgery could not be pooled.

Conclusions: Prehabilitation delivered with advanced technologies may be better than standard care in improving pre- and postoperative function among candidates for knee or hip arthroplasty. No quantitative results have been achieved on spine surgery candidates or other musculoskeletal diseases.

Trial Registration: PROSPERO CRD42022345811; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=345811

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52943) doi:10.2196/52943

KEYWORDS

prehabilitation; advanced technologies; musculoskeletal diseases; joint arthroplasty; elective surgery; patient outcomes; mobile health; postoperative function

Introduction

Musculoskeletal diseases are one of the biggest contributors to global disability [1] and the volume of related surgeries continues to increase worldwide [2]. Highly invasive surgery imposes substantial stress on the body's physiology [3], affecting not only physical aspects but also physiological, cognitive, and psychosocial factors [3,4]. Undergoing major interventions results in a high systemic inflammatory response and increases catabolism and oxygen demand in the immediate postoperative period [3,5]. Moreover, patients commonly experience preoperative anxiety, sleep disturbance, depression, and pain, which predict higher postoperative disability, poor function, and health-related quality of life [6,7]. Therefore, the waiting period before elective surgery should provide the opportunity for patients to identify and manage these internal contextual factors optimizing their chances of postoperative recovery [5].

The term "prehabilitation" refers to actively preparing patients for surgery through interventions such as exercise, nutritional support, psycho-cognitive training, or a combination thereof, to enhance functional capacity and improve tolerance to upcoming surgical stress [3,5]. Although prehabilitation is highly valued by patients, the evidence supporting its benefits in the musculoskeletal field is varied and uncertain [3,8]. Some studies demonstrate benefits per function, pain, quality of life, and length of hospital stay [8-10], while others show comparable outcomes to no treatment [11-16]. Moreover, prehabilitation is a noninvasive and low-cost intervention [17,18]. Given all these potential benefits, it would be worth recommending [13], especially for patients who are frail or present risk factors for poor recovery [13,19]. However, the evidence from these trials and reviews is relatively weak, often with a high risk of bias (RoB), with critical heterogeneity in populations, interventions, and outcome measures [20]. Additionally, patients' compliance varies widely [21]. Digital health interventions hold significant potential to enhance the availability of health care services, facilitating better access and adherence to treatment [22,23].

Following orthopedic surgery, telerehabilitation using advanced technology systems has demonstrated comparable effects to standard treatment, often with lower costs, high patient satisfaction, and adherence [24-28]. Advanced technologies in rehabilitation refer to a broad spectrum of devices and software that integrate hardware and software capabilities to support and improve the assessment, training, or recovery processes in patients with physical or cognitive impairments. Common examples include virtual reality, wearable sensors, robotic exoskeletons, and telerehabilitation platforms [29]. These technologies provide personalized, interactive, and data-driven rehabilitation experiences, often making use of real-time feedback, data analytics, and remote connectivity to enhance the efficiency and effectiveness of rehabilitation programs [30]. In particular, the remote delivery of prehabilitation through technological devices or web-based modalities offers the double advantage of lower costs compared to traditional prehabilitation [31-33] and facilitates monitoring of adherence and results. Nevertheless, to the best of our knowledge, no systematic reviews investigated the effects of prehabilitation with advanced technologies (PrAT) before orthopedic surgery yet. In this context, we aim to evaluate the effects, per function, of PrAT compared to standard care to provide patients, clinicians, and stakeholders evidence to better manage the preoperative phase of candidates.

Methods

Study Design

This systematic review was conducted following the Cochrane Handbook guidance and it is reported per the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [32] (section S1 in Multimedia Appendix 1). This study protocol was registered on PROSPERO (CRD42022345811). Differences between study protocol and publication are reported in section S1 in Multimedia Appendix 1.

Eligibility Criteria

We considered eligible randomized controlled trials (RCT) and nonrandomized intervention studies (NRISs) comparing the effects of PrAT (eg, synchronous or asynchronous sessions of physical exercise, education to autonomous exercise execution, or activities suggestion) and standard care or no intervention, in adult patients affected by a musculoskeletal disease waiting for orthopedic surgery. More details on inclusion criteria are reported in section S2 in Multimedia Appendix 1.

Outcomes

We considered function as the primary outcome, of physical structures integrity, physiology, activity, and participation [34]. If studies used multiple metrics to assess similar constructs, we prioritized the extraction of outcome measures according to existing core outcome sets. Specifically, we chose Outcome Measures in Rheumatology–Osteoarthritis Research Society International [35] core domain set (WOMAC [Western Ontario and McMaster Universities Osteoarthritis Index] function for the primary outcome) for osteoarthritis and Chiarotto et al's [36] outcome set for lumbar degenerative diseases (Oswestry Disability Index for the primary outcome). We prioritized multiple outcome measures according to the aforementioned core outcome sets post hoc, as this alignment was necessary to reflect the specific population focus of the included studies.

Secondary outcomes were pain, strength, risk of fall, autonomy in the activities of daily living, patient satisfaction, health-related quality of life, adverse events, and adherence to treatment.

XSL•FO
RenderX

Literature Search

PubMed, EMBASE, Cochrane Library, PEDro, CINAHL, and ClinicalTrial.gov were searched in July 2022 and updated on February 2, 2023, with no further results. Details on the search strategy are reported in section S2 in Multimedia Appendix 1. Titles and abstracts of the records were screened by 2 independent reviewers (S Guida and JAV), and the full texts were obtained if they met the inclusion criteria. We also screened for eligibility for all references from relevant systematic reviews. A third reviewer was involved in case of disagreement (S Gianola). Rayyan [37] software was used to manage this study's selection phase.

Data Extraction

Two reviewers (S Guida and JAV) extracted the following data: study generalities, study methods, population characteristics, intervention and comparator characteristics, outcomes, and possible other important variables. As described in the Outcomes section, we prioritized the extraction of multiple outcome measures for similar constructs according to existing core outcome sets (ie, Outcome Measures in Rheumatology–Osteoarthritis Research Society International and Chiarotto et al's [36] core outcome set). By adhering to these core outcome sets, we ensure that the outcome measures extracted were both relevant and standardized, facilitating a more coherent and comparable synthesis of findings across the included studies.

For interventions, we applied the reporting checklist of the 12-item Template for Intervention Description and Replication (TIDieR) checklist (section S2 in Multimedia Appendix 1) [38].

Methodological Quality

Two independent authors (S Guida and JAV) assessed the RoB using The Cochrane RoB tool for RCTs [39] and The Newcastle Ottawa Scale (NOS) for NRISs [40]. Discrepancies were solved by discussion. The description of the domains evaluated by RoB and NOS and the rules for judgment are reported in section S2 in Multimedia Appendix 1. For these phases, we used DistillerSR.

Data Synthesis and Statistics

The treatment effect has been evaluated using mean difference (MD) or standardized mean difference as appropriate. As we supposed a certain degree of clinical heterogeneity between studies (eg, delivered programs of interventions), we opted for random effects models with 95% CIs. We assessed the presence of statistical heterogeneity using the *I*-squared statistic (I^2). When I^2 was higher than 75% we planned to explore the sources of heterogeneity by subgroup or sensitivity analyses, depending on study findings [41,42]. Analyses were done with Review Manager (RevMan5) software (version 5.2; The Cochrane Collaboration) and Stata13 (StataCorp LLC) [43].

Certainty of evidence was assessed for all the meta-analyzed outcomes with the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) system (section S2 in Multimedia Appendix 1).

Results

Study Findings

The searches retrieved 451 records. After removing duplicates, 316 references were screened by titles and abstracts. Twenty studies were considered for eligibility and assessed by full text. Finally, 7 studies were included (Figure 1) [44-50]. Two references belong to the same study, reporting different time points for interested outcomes [45,46].

Figure 1. PRISMA flowchart. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

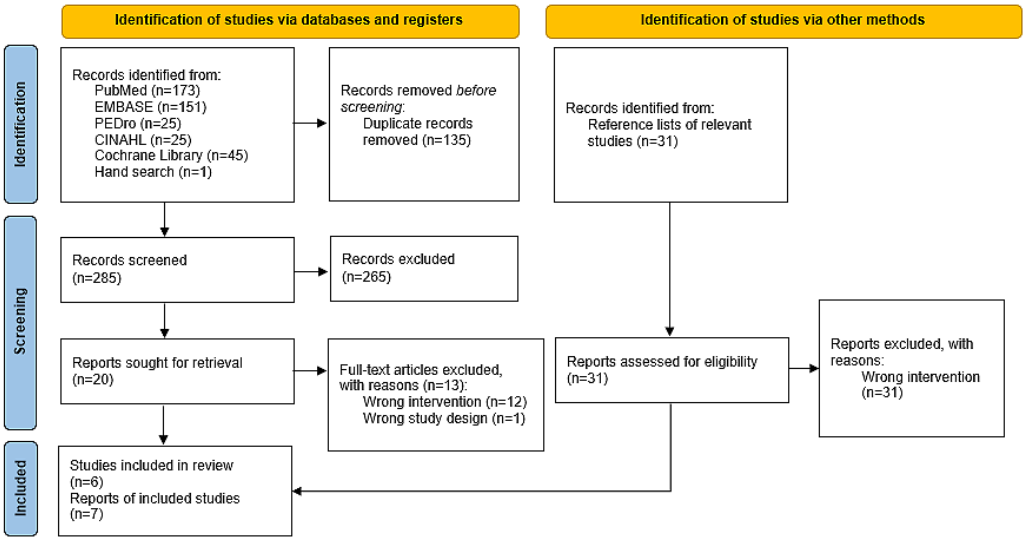

Characteristics of the Included Studies

The 7 included studies were published between 2015 and 2021. Four studies were RCTs [44,47-50] and 2 NRISs [45,46]. Most studies were conducted in the United States (n=4) [45-48] followed by Canada (n=2) [44,49] and Korea (n=1) [50]. The main studies' features are provided in Table 1.

XSL•FO
RenderX

Table 1. General characteristics of the included studies.

First author (year)	Study design	Diagnosis	Sample size (N)	Intervention modality	Intervention dosage and characteristics	Weeks (n): (1) of intervention (2) before surgery	Comparator characteristics	Outcomes measures	Assessment time points
Yi et al (2019) [45]	NRIS[a]	Lumbar degenerative disease	48	Asynchronous	• Total volume: 960 minutes. • Mindfulness-based stress reduction course. • 8 audio-video lessons + 6 h audio-video retreat (about 16 h overall)	• Not reported (at least 1 session before surgery)	No prehabilitation	MED[b] per day, VAS-BP[c], VAS-LP[d], ODI[e], EQ-5D-QUALY, EQ-5D-VAS, PROMIS-PF[f], and PROMIS-PI[g]	Baseline (preoperative visit), 30 days after surgery
Chavez et al (2020) [46]	NRIS	Lumbar degenerative disease	48	Asynchronous	• Total volume: 960 minutes. • Mindfulness-based stress reduction course. • 8 audio-video lessons + 6 h audio-video retreat (about 16 h overall)	• Not reported (at least 1 session before surgery)	No prehabilitation	MED per day, VAS-BP, VAS-LP, ODI, EQ-5D-QUALY, EQ-5D-VAS, PROMIS-PF, and PROMIS-PI	Baseline (preoperative visit), 3 months and 12 months after surgery
Yin et al (2015) [47]	RCT[h]	Primary meniscal tear	64	Asynchronous	• Total volume: 20 minutes. • Web-based tutorial covering anatomy, pathology, and general perioperative instructions	• 1 week • 1-2 weeks	Standard counselling	Ad hoc surveys for patient's perioperative experience preparedness and satisfaction)	Baseline (preoperative visit), the day of surgery (preoperative), first visit after surgery
Culliton et al (2018) [49]	RCT	Knee osteoarthritis	416	Asynchronous	• Total volume: 50 minutes. • Digital version of "My Guide to Total Knee Replacement" + 32 educational videos (patient's expectations on pain, function, restrictions, or exercises) and 2 medical animations of TKA[i] surgery (about 50 minutes overall).	• Not reported (from preadmission visit to surgery)	Paper-printed "My Guide to Total Knee Replacement"	Preoperative expectation survey, KSS[j], Knee Injury and Osteoarthritis Outcome Score, SF-12[k], Hospital Anxiety and Depression Scale, Pain Catastrophizing Scale, Social Role Participation Questionnaire, PASS[l]	Baseline (preoperative visit), 6 weeks, 3 months, and 1 year after surgery
Soeters et al (2018) [48]	RCT	Knee or hip osteoarthritis	126	Asynchronous and synchronous	• Total volume: 85 minutes. • Microsites for information and preparation (with also images and videos) + one face-to-face session	• 1-2 weeks • 2 weeks	One face-to-face session	WOMAC[m] and length of stay	Baseline (preoperative visit), 4-6 weeks after surgery

XSL•FO
RenderX

First author (year)	Study design	Diagnosis	Sample size (N)	Intervention modality	Intervention dosage and characteristics	Weeks (n): (1) of intervention (2) before surgery	Comparator characteristics	Outcomes measures	Assessment time points
An et al (2021) [50]	RCT	Knee osteoarthritis	60	Asynchronous and synchronous	• Total volume group A: 900 minutes. • Group A: 2-way video call (therapeutic exercise). • Group B: Guidebook for exercises and education.	• 3 weeks • 4 weeks	Standard counseling	ROM[n], PPT[o], TUG[p], WOMAC, quadriceps peak torque 60°/s and 180°/s	Baseline (4 weeks before surgery), the day of surgery (preoperative), 6 weeks after surgery
Doiron-Cardin et al (2020) [44]	RCT	Knee or hip osteoarthritis	34	Asynchronous and synchronous	• Total volume group A: 3780 minutes. • Group A: 2/7 sessions through 2-way video call and 5/7 unsupervised home sessions (therapeutic exercise and education). • Group B: 2/7 in-person sessions and 5/7 unsupervised home sessions (therapeutic exercise and education).	• 12 weeks • Not reported	One face-to-face session (home-based)	LEFS[q], WOMAC, SF-36[r], TUG, ST[s], SPW[t], satisfaction survey	Baseline (preoperative visit), 12 weeks after baseline (preoperative)

[a]NRIS: nonrandomized intervention study.
[b]MED: morphine-equivalent dosing.
[c]VAS-BP: visual analogue scale for back pain.
[d]VAS-LP: visual analogue scale for leg pain.
[e]ODI: Oswestry Disability Index.
[f]PROMIS-PF: Patient-Reported Outcome Measurement Information System Physical Function.
[g]PROMIS-PI: Patient-Reported Outcomes Measurement Information System Pain Interference.
[h]RCT: randomized controlled trial.
[i]TKA: total knee arthroplasty.
[j]KSS: Knee Society Score.
[k]SF-12: Short Form Health Survey.
[l]PASS: patient acceptable symptom state.
[m]WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.
[n]ROM: range of motion.
[o]PPT: pressure pain threshold.
[p]TUG: timed-up and go.
[q]LEFS: Lower Extremity Functional Scale.
[r]SF-36: Short-Form 36 Health Survey.
[s]ST: stair test.
[t]SPW: self-paced walk.

Overall, 748 participants were considered, ranging from 34 to 416 per study.

Interventions included therapeutic exercise, mindfulness-based exercises for stress reduction, and preoperative information and education. Treatments were mainly focused on (1) online mindfulness-based exercises for stress reduction, including indications for gentle stretching, yoga, and walking (n=2) [45,46]; (2) online educational content, including exercise suggestions (n=3) [47-49]; (3) remotely supervised exercises (n=1) [50]; and (4) remotely supervised and unsupervised exercises plus educational contents (n=1) [44]. The total prehabilitation dosage ranges from 20 to 3780 minutes, delivered over 1 to 12 weeks using an asynchronous (n=4) [45,46,48,49], synchronous (n=1) [50], or mixed (n=1) [44] modality. Comparisons groups were “no intervention” (n=2) [45,46] or “standard care” (n=5; ie, standard counseling, standard education, or one face-to-face session) [44,47-50] as defined by the authors.

As assessed by the TIDieR checklist, all trials reported: complete information on the provided interventions on rationale, setting,

XSL•FO
RenderX

and materials; incomplete information on procedures [48], intervention providers [45-48], posology [45,46], tailoring, and planning (all studies). No studies reported information on intervention modification. TIDieR scoring details are provided in Table S1 in section S3 in Multimedia Appendix 1. Most studies reported pain outcomes. Details about outcome reporting are in Table S2 in section S3 in Multimedia Appendix 1.

Methodological Quality

The RoB assessment of RCTs is reported in Figure 2 [44,47-50]. All studies were judged as having a high RoB. However, the final judgment is impacted by the nature of the investigated intervention that, for its intrinsic features, does not allow blinding of participants and personnel to the delivered treatment. All the RCTs used blinded procedures to randomize and allocate participants and blinded outcomes assessors. Two studies deviated from what the protocol stated per primary outcomes.

The 2 NRISs [45,46] were both assessed with a score of 3/9 of the NOS. The "poor quality" was imputed, for both studies, to a self-reported ascertainment of exposure to the intervention and to the presence of the outcome of interest at the start of the study. One study had too short a follow-up and the other had a high dropout rate. NOS scores are reported in Table 2, with further details in Table S3 in section S3 in Multimedia Appendix 1.

Figure 2. Risk of bias graph and summary.

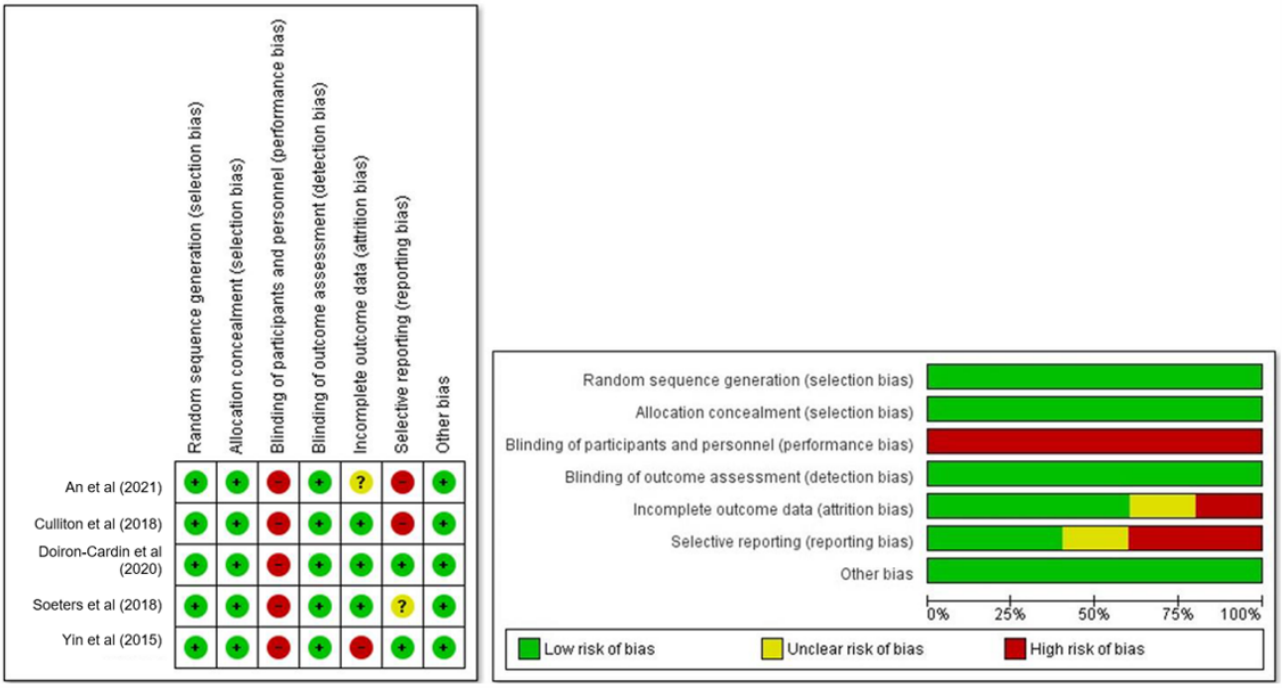

Table 2. Newcastle Ottawa Scale quality score summary.

Author and year	Selection (out of 4)	Comparability (out of 2)	Outcome (out of 3)	Overall quality score
Yi et al (2019) [45]	✖		✖	✖
Chavez et al (2020) [46]	✖		✖	✖

Data Synthesis on Knee or Hip Prehabilitation

Overview

Four studies investigated the effect of PrAT on patients affected by knee or hip osteoarthritis and waiting for arthroplasty.

Primary Outcome: Function

Three out of four studies measured the lower limb function using the WOMAC function subscale. With very low certainty of evidence, meta-analyses show a statistically significant better effect of PrAT compared to standard, both at the end of the intervention (2 studies, n=94, MD –7.45, 95% CI –10.71 to –4.19, I^2=0%) and a month after surgery (2 studies, n=186, MD –7.84, 95% CI –11.80 to –3.88, I^2=75.3%; Figure 3 [45,49,51]). The main reason for the downgrade was imprecision (clinical effect and sample size; section S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 3. Forest plot WOMAC function subscale. WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

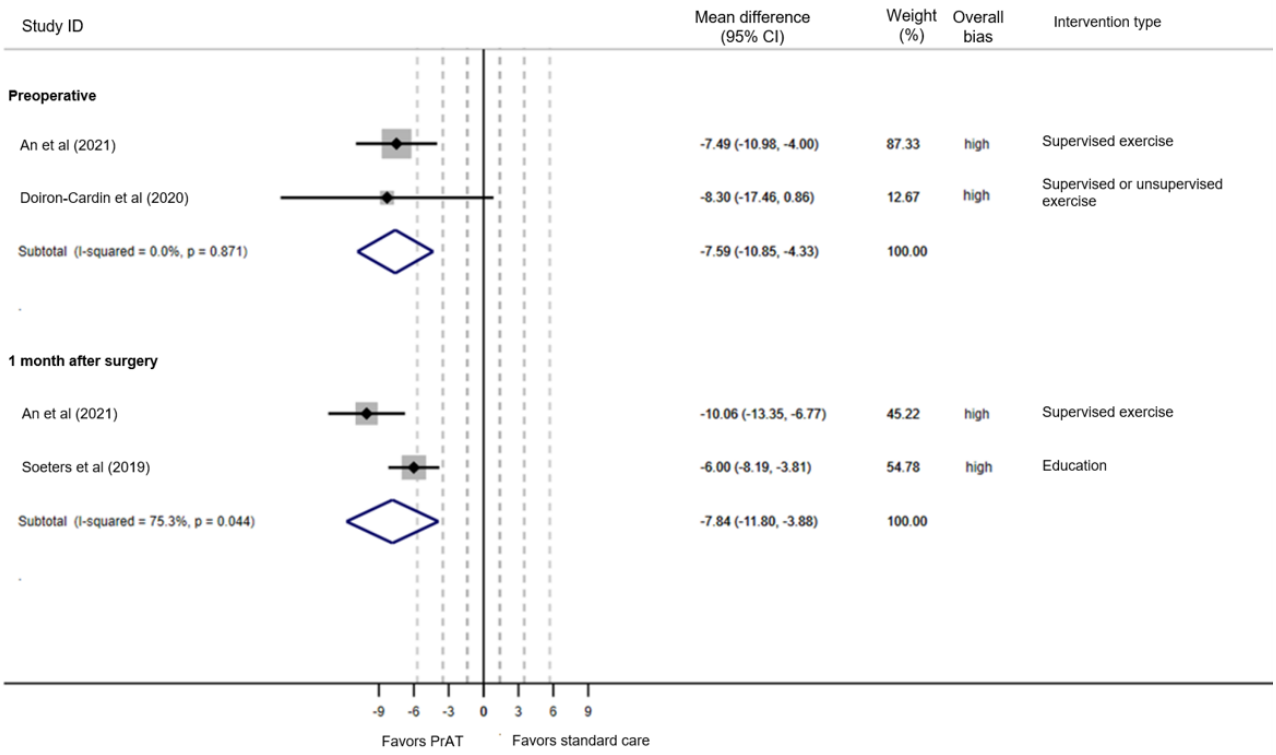

Secondary Outcomes

Risk of Fall

Meta-analysis of 2 studies (n=94) showed a significant improvement in favor of PrAT on the TUG test at the end of the intervention (MD –2.54, 95% CI –3.62 to –1.46, I^2=0%) compared to standard care, with low certainty of evidence (section S4 in Multimedia Appendix 1).

Stiffness

Meta-analysis of two studies (n=186) showed a significant improvement in favor of PrAT on the WOMAC stiffness subscale one month after surgery (MD –2, 95% CI –2.01 to –1.99, I^2=87%), with low certainty of evidence. The forest plot is reported in Figure S2 in section S4 in Multimedia Appendix 1.

Pain

Meta-analysis of two studies (n=94) showed a significant improvement in favor of PrAT on the WOMAC pain subscale at the preoperative time point (MD –1.67, 95% CI –2.50 to –0.84, I^2=0%), with very low certainty of evidence. No statistically significant difference in pain was found one month after surgery (2 studies, n=186; standardized mean difference –1.06, 95% CI –2.24 to 0.12, I^2=88%). Forest plots are reported in Figures S3 and S4 in section S4 in Multimedia Appendix 1.

Strength

No meta-analysis could be performed on this outcome. Quadriceps strength was investigated by one study [50] showing a statistically significant difference in favor of PrAT, at different time points, for participants undergoing total knee arthroplasty (TKA). Significant differences were observed across pre and postoperative time points (P<.001), between groups, as well as a significant time-by-group interaction.

Autonomy in the Activities of Daily Living

No meta-analysis could be performed on this outcome. Activity limitations were investigated by one study [49] showing a statistically significant difference in favor of PrAT one year after surgery, for participants undergoing TKA (P=.04).

Patient's Experience and Satisfaction

No meta-analysis could be performed on this outcome. One study [49] assessed patient satisfaction and experience reporting no differences between groups. Another study [44] reported high satisfaction with tele-prehabilitation but without any information on the usual care group.

Adherence to Treatment

No meta-analysis could be performed on this outcome. Three studies assessed adherence (number of sessions performed versus planned). Two studies [47,48] reported that most PrAT participants completed the intervention (70.83% and 97%, respectively). Another study reported similar adherence rates for both the intervention (between 77% and 73%) and comparison groups (between 80% and 86%) [44].

No studies including participants affected by knee or hip osteoarthritis reported information on possible adverse events related to treatment or PrAT effect on health-related quality of life.

Data Synthesis on Other Musculoskeletal Diseases

The effects of PrAT were assessed by 1 study (2 reports) [45,46] on patients affected by lumbar degenerative disease and by 1 study [47] on patients with primary meniscal tear.

XSL•FO
RenderX

Primary Outcome: Function

No meta-analysis could be performed on this outcome. One study (2 reports) [45,46], investigating PrAT in participants affected by lumbar degenerative disease, assessed function by the Oswestry Disability Index. No statistically significant differences were found between groups 30 days after surgery (*P*=.12, while better function was found at the 3-month follow-up (*P*=.03).

Secondary Outcomes

Pain

No meta-analysis could be performed on this outcome. One study (2 reports) [45,46], investigating PrAT in participants affected by lumbar degenerative disease, assessed function by the visual analogue scale for back pain (VAS-BP) and the visual analogue scale for leg pain. VAS-BP was better in the PrAT group 30 days after surgery (*P*=.004). No statistically significant differences were found for VAS-BP and visual analogue scale for leg pain both at the 30 days (*P*=.495) and 3 months (*P*=.06) after surgery time points.

Patient's Experience and Satisfaction

No meta-analysis could be performed on this outcome. One study [47], investigating PrAT in participants with primary meniscal tears, assessed patients' experience and satisfaction by an ad hoc survey. Participants undergoing PrAT felt significantly more prepared for surgery and satisfied (*P*=.04), and satisfied (*P*=.03) with the received assistance compared to those undergoing standard counseling.

Health-Related Quality of Life

No meta-analysis could be performed on this outcome. One study (2 reports) [45,46], investigating PrAT in participants affected by lumbar degenerative disease, assessed quality of life by the EQ-5D-5L questionnaire after surgery. No statistically significant difference between groups was found in both 30 days (*P*=.19) and 3 months (*P*=.27) after surgery.

Adherence to Treatment

No meta-analysis could be performed on this outcome. One study (2 reports) [45,46], including participants affected by lumbar degenerative disease, reported an adherence rate to PrAT (number of patients that completed at least 1 of 6 preoperative sessions) of 70.83%.

No studies including participants affected by other musculoskeletal diseases reported information on possible adverse events related to treatment or PrAT effect on strength, risk of falls, and autonomy in the activities of daily living.

Discussion

Main Findings

This systematic review aimed to evaluate if PrAT is better than standard care in improving function for patients with musculoskeletal disease and waiting for surgery. We included TKA, total hip arthroplasty, and patients with lumbar degenerative disease undergoing different PrAT interventions per modality (synchronous or asynchronous), prehabilitation main body (exercise, education, instruction, or suggestions), and dosage with a duration of intervention ranging from 1 to 12 weeks. Our analysis suggests that PrAT may increase function, but the evidence is very uncertain at both preoperative (supervised and unsupervised exercises) and postoperative time points (education and exercise) in participants undergoing TKA or total hip arthroplasty. One study assessed the function of spine surgery candidates, with better outcomes 3 months after surgery but not at the 30-day time point. No study assessed the function of other musculoskeletal diseases. The magnitude of the effect for the WOMAC function subscale was critically important for patients undergoing knee or hip arthroplasty. However, the estimated effects were very wide, including both large and moderate effects resulting in an imprecision of estimates.

Little evidence was found in secondary outcomes assessed on participants undergoing knee or hip arthroplasty, suggesting that PrAT may reduce pain and risk of fall before surgery, and pain and stiffness one month after surgery, but in this case, the evidence is very uncertain.

Interventions' heterogeneity was high in the main program body (exercise or education), delivery mode (synchronous or asynchronous), and dosage (from 20 to 3780 min). Since benefits and harms can be associated with volumes of exercise interventions, and a relationship between dosing and effect size exists [51], guidelines for exercise dosage in prehabilitation are needed to compare different intervention types and pool results of primary studies. Compliance with PrAT programs was very high, suggesting great patient engagement and motivation. Nevertheless, no information has been reported on participants' compliance per the intensity of the performed sessions, which could affect the true effect size of the treatment.

Comparison With Previous Studies

Literature on PrAT in musculoskeletal diseases is underdeveloped and mainly focused on knee and hip, with little evidence of spine surgery. Despite the lack of evidence, prehabilitation is gaining relevance in orthopedic surgery [52] and, in other fields of medicine (eg, major surgery or metabolic dysfunctions), yet is widely delivered through advanced technology systems [53,54] to properly meet the care needs of the target population, encourage adherence, and facilitate monitoring, with similar findings. Several systematic reviews exist assessing the effects of traditional prehabilitation programs for patients undergoing joint replacement [9,11,55], but they mostly focussed on the postoperative intervention effect.

Clinical and Research Implications

Given that poor function and high preoperative pain levels are associated with worse postoperative outcomes and higher length of stay [11,16,56], PrAT could be an important intervention for patients presenting with these preoperative clinical characteristics. The high compliance observed in the included studies indicates strong patient engagement, which supports the feasibility of integrating PrAT into preoperative protocols. However, the variability in intervention types and dosages underscores the need for guidelines on exercise volume and intensity to optimize outcomes in prehabilitation interventions.

XSL•FO
RenderX

Indeed, in our systematic review, the interventions could include low-impact warm-up, mobility, strengthening, proprioceptive exercises, and cool-down [44,50] or less structured interventions such as exercise suggestions, walking, stretching, and yoga practicing [45-49]. Exercise administration has proved to be effective in improving outcomes [16]. Thus, also educating patients to exercise properly to improve their conditioning before surgery could be a straightforward and effective strategy. However, to investigate its effectiveness, overall and of the individual components of multimodal interventions, it is necessary to collect data on what participants actually did after receiving education, such as through prehabilitation diaries. This information was not collected in any of the studies where prehabilitation included an educational component.

From a research perspective, further rigorous and well-designed primary studies are needed to provide strong evidence on the effects dimension provided by PrAT, both before and after surgery, extending the investigation to a wider range of musculoskeletal diseases. Additionally, monitoring patient compliance, per attendance and session intensity, and accurately reporting these data, will be crucial for understanding the true impact of PrAT. Developing standardized protocols and robust evidence will ultimately guide clinical practice and enhance the prehabilitation of patients with musculoskeletal issues awaiting surgery.

Strengths and Limitations

To the best of our knowledge, this study has the strength of being the first systematic review assessing the effectiveness of PrAT for patients waiting for musculoskeletal surgery. It also analyzed the effect of intervention at pre and postoperative time points, collecting data from trials providing prehabilitation which included a component of exercise. However, limitations could be assessed at both review and study levels. At the review level, we did not assess grey literature and we planned GRADE assessment only for the primary outcome, as a proxy of clinical importance, because we expected that a patient undergoing PrAT could mainly improve their function during motor tasks [9,11]. Additionally, all the meta-analyses included a small number of studies. While this is methodologically reasonable, readers should consider it when interpreting the robustness of the results. Another limitation is that we did not consider overall WOMAC function but the dimensions presented in the WOMAC scale, considering function as primary outcome and pain and stiffness as secondary outcomes.

At the study level, we found high heterogeneity of the delivered programs, including exercise as the main body of prehabilitation [50], in equal shares with education [44], or as a very marginal part in the form of instructions or suggestions [45,48]. Moreover, the modality of exercise delivery (instructions or suggestions) does not allow for objective verification of what participants actively did, if they had, to perform physical exercise as intended. Indeed, collecting information on the compliance related to the “educational contents” vision or reading does not give any information about whether patients had applied what was suggested. Comparators’ heterogeneity should also be considered when interpreting the results because we pooled together studies assessing PrAT compared to different types of standard care.

Conclusion

PrAT may be more effective than standard care in improving function both pre- and postoperatively, in candidates for knee or hip replacement. Little and uncertain evidence was found in secondary outcomes, with some trend of positive findings except for pain after surgery.

No quantitative results could be achieved on spine surgery candidates or for other musculoskeletal diseases. Intervention heterogeneity was high and adherence was often underassessed.

Acknowledgments

This study was supported and funded by the Italian Ministry of Health—Ricerca Corrente. The Article Processing Fee (APC) was funded by the Italian Ministry of Health—Ricerca Corrente.

Authors' Contributions

S Guida and JAV conceptualized and designed this study, drafted the initial paper, and coordinated and supervised the data selection and collection. FP conducted data selection and provided feedback to this paper. S Gianola, GC, SB, and S Guida did the statistical analysis and contributed to the interpretation of the results. ES and DB contributed to the results’ interpretation and reviewed this paper. All authors approved this final paper as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional materials.
[DOCX File , 268 KB - jmir_v26i1e52943_app1.docx]

References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic

XSL•FO
RenderX

analysis for the global burden of disease study 2017. Lancet 2018;392(10159):1789-1858. [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Weiser TG, Haynes AB, Molina G, Lipsitz SR, Esquivel MM, Uribe-Leitz T, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. Lancet 2015 Apr 27;385 Suppl 2:S11. [doi: 10.1016/S0140-6736(15)60806-6] [Medline: 26313057]
3. Kow AW. Prehabilitation and its role in geriatric surgery. Ann Acad Med Singap 2019 Nov;48(11):386-392 [FREE Full text] [Medline: 31960020]
4. Kow AW. Prehabilitation and its role in geriatric surgery. Ann Acad Med Singap 2019;48(11):386-392 [FREE Full text] [Medline: 31960020]
5. Punnoose A, Claydon-Mueller LS, Weiss O, Zhang J, Rushton A, Khanduja V. Prehabilitation for patients undergoing orthopedic surgery: a systematic review and meta-analysis. JAMA Netw Open 2023;6(4):e238050 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.8050] [Medline: 37052919]
6. Alattas SA, Smith T, Bhatti M, Wilson-Nunn D, Donell S. Greater pre-operative anxiety, pain and poorer function predict a worse outcome of a total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc 2017;25(11):3403-3410 [FREE Full text] [doi: 10.1007/s00167-016-4314-8] [Medline: 27734110]
7. Vitale JA, Banfi G, Tivolesi V, Pelosi C, Borghi S, Negrini F. Rest-activity daily rhythm and physical activity levels after hip and knee joint replacement: the role of actigraphy in orthopedic clinical practice. Chronobiol Int 2021;38(12):1692-1701. [doi: 10.1080/07420528.2021.1939365] [Medline: 34121554]
8. Perry R, Herbert G, Atkinson C, England C, Northstone K, Baos S, et al. Pre-admission interventions (prehabilitation) to improve outcome after major elective surgery: a systematic review and meta-analysis. BMJ Open 2021;11(9):e050806 [FREE Full text] [doi: 10.1136/bmjopen-2021-050806] [Medline: 34593498]
9. Moyer R, Ikert K, Long K, Marsh J. The value of preoperative exercise and education for patients undergoing total hip and knee arthroplasty: a systematic review and meta-analysis. JBJS Rev 2017;5(12):e2. [doi: 10.2106/JBJS.RVW.17.00015] [Medline: 29232265]
10. Burgess LC, Arundel J, Wainwright TW. The effect of preoperative education on psychological, clinical and economic outcomes in elective spinal surgery: a systematic review. Healthcare (Basel) 2019;7(1):48 [FREE Full text] [doi: 10.3390/healthcare7010048] [Medline: 30901875]
11. Wu Z, Wang Y, Li C, Li J, Chen W, Ye Z, et al. Preoperative strength training for clinical outcomes before and after total knee arthroplasty: a systematic review and meta-analysis. Front Surg 2022;9:879593 [FREE Full text] [doi: 10.3389/fsurg.2022.879593] [Medline: 35937597]
12. Potts G, Reid D, Larmer P. The effectiveness of preoperative exercise programmes on quadriceps strength prior to and following anterior cruciate ligament (ACL) reconstruction: a systematic review. Phys Ther Sport 2022;54:16-28. [doi: 10.1016/j.ptsp.2021.12.004] [Medline: 34933208]
13. Delgado-López PD, Rodríguez-Salazar A, Castilla-Díez JM. "Prehabilitation" in degenerative spine surgery: a literature review. Neurocirugia (Astur : Engl Ed) 2019;30(3):124-132. [doi: 10.1016/j.neucir.2018.11.008] [Medline: 30612856]
14. Vasta S, Papalia R, Torre G, Vorini F, Papalia G, Zampogna B, et al. The influence of preoperative physical activity on postoperative outcomes of knee and hip arthroplasty surgery in the elderly: a systematic review. J Clin Med 2020;9(4):969 [FREE Full text] [doi: 10.3390/jcm9040969] [Medline: 32244426]
15. Giesche F, Niederer D, Banzer W, Vogt L. Evidence for the effects of prehabilitation before ACL-reconstruction on return to sport-related and self-reported knee function: a systematic review. PLoS One 2020;15(10):e0240192 [FREE Full text] [doi: 10.1371/journal.pone.0240192] [Medline: 33112865]
16. Konnyu KJ, Thoma LM, Bhuma MR, Cao W, Adam GP, Mehta S. Prehabilitation and Rehabilitation for Major Joint Replacement. Rockville, MD: Agency for Healthcare Research and Quality; 2021.
17. Rombey T, Eckhardt H, Kiselev J, Silzle J, Mathes T, Quentin W. Cost-effectiveness of prehabilitation prior to elective surgery: a systematic review of economic evaluations. BMC Med 2023;21(1):265 [FREE Full text] [doi: 10.1186/s12916-023-02977-6] [Medline: 37468923]
18. Fernandes L, Roos EM, Overgaard S, Villadsen A, Søgaard R. Supervised neuromuscular exercise prior to hip and knee replacement: 12-month clinical effect and cost-utility analysis alongside a randomised controlled trial. BMC Musculoskelet Disord 2017;18(1):5. [doi: 10.1186/s12891-016-1369-0] [Medline: 28061841]
19. Hermus JPS, van Kuijk SMJ, Spekenbrink-Spooren A, Witlox MA, Poeze M, van Rhijn LW, et al. Risk factors for total ankle arthroplasty failure: a Dutch arthroplasty register study. Foot Ankle Surg 2022;28(7):883-886 [FREE Full text] [doi: 10.1016/j.fas.2021.12.001] [Medline: 34949541]
20. Adebero T, Omana H, Somerville L, Lanting B, Hunter SW. Effectiveness of prehabilitation on outcomes following total knee and hip arthroplasty for osteoarthritis: a systematic review and meta-analysis of randomized controlled trials. Disabil Rehabil 2024:1-20. [doi: 10.1080/09638288.2024.2313128] [Medline: 38349251]
21. Lobo D, Skořepa P, Gomez D, Greenhaff P. Prehabilitation: High-Quality Evidence is Still Required. Amsterdam, The Netherlands: Elsevier Ltd; 2023.

XSL•FO
RenderX

22. Lang S, McLelland C, MacDonald D, Hamilton DF. Do digital interventions increase adherence to home exercise rehabilitation? A systematic review of randomised controlled trials. Arch Physiother 2022;12(1):24 [FREE Full text] [doi: 10.1186/s40945-022-00148-z] [Medline: 36184611]
23. Shaw J, Abejirinde IOO, Agarwal P, Shahid S, Martin D. Digital health and equitable access to care. PLOS Digit Health 2024;3(9):e0000573. [doi: 10.1371/journal.pdig.0000573] [Medline: 39321176]
24. Wijnen A, Hoogland J, Munsterman T, Gerritsma CL, Dijkstra B, Zijlstra WP, et al. Effectiveness of a home-based rehabilitation program after total hip arthroplasty driven by a tablet app and remote coaching: nonrandomized controlled trial combining a single-arm intervention cohort with historical controls. JMIR Rehabil Assist Technol 2020;7(1):e14139 [FREE Full text] [doi: 10.2196/14139] [Medline: 32338621]
25. Shukla H, Nair SR, Thakker D. Role of telerehabilitation in patients following total knee arthroplasty: evidence from a systematic literature review and meta-analysis. J Telemed Telecare 2017;23(2):339-346. [doi: 10.1177/1357633X16628996] [Medline: 26843466]
26. Jansson MM, Rantala A, Miettunen J, Puhto AP, Pikkarainen M. The effects and safety of telerehabilitation in patients with lower-limb joint replacement: a systematic review and narrative synthesis. J Telemed Telecare 2022;28(2):96-114. [doi: 10.1177/1357633X20917868] [Medline: 32316837]
27. Russell TG, Buttrum P, Wootton R, Jull GA. Internet-based outpatient telerehabilitation for patients following total knee arthroplasty: a randomized controlled trial. J Bone Joint Surg Am 2011;93(2):113-120. [doi: 10.2106/JBJS.I.01375] [Medline: 21248209]
28. Fascio E, Vitale JA, Sirtori P, Peretti G, Banfi G, Mangiavini L. Early virtual-reality-based home rehabilitation after total hip arthroplasty: a randomized controlled trial. J Clin Med 2022;11(7):1766 [FREE Full text] [doi: 10.3390/jcm11071766] [Medline: 35407373]
29. Kuroda Y, Young M, Shoman H, Punnoose A, Norrish AR, Khanduja V. Advanced rehabilitation technology in orthopaedics-a narrative review. Int Orthop 2021;45(8):1933-1940 [FREE Full text] [doi: 10.1007/s00264-020-04814-4] [Medline: 33051693]
30. Arntz A, Weber F, Handgraaf M, Lällä K, Korniloff K, Murtonen K, et al. Technologies in home-based digital rehabilitation: scoping review. JMIR Rehabil Assist Technol 2023;10:e43615 [FREE Full text] [doi: 10.2196/43615] [Medline: 37253381]
31. Paganini S, Lin J, Kählke F, Buntrock C, Leiding D, Ebert DD, et al. A guided and unguided internet- and mobile-based intervention for chronic pain: health economic evaluation alongside a randomised controlled trial. BMJ Open 2019;9(4):e023390 [FREE Full text] [doi: 10.1136/bmjopen-2018-023390] [Medline: 30967405]
32. Fatoye F, Gebrye T, Fatoye C, Mbada CE, Olaoye MI, Odole AC, et al. The clinical and cost-effectiveness of telerehabilitation for people with nonspecific chronic low back pain: randomized controlled trial. JMIR mHealth uHealth 2020;8(6):e15375 [FREE Full text] [doi: 10.2196/15375] [Medline: 32357128]
33. Prvu Bettger J, Green CL, Holmes DN, Chokshi A, Mather RC, Hoch BT, et al. Effects of virtual exercise rehabilitation in-home therapy compared with traditional care after total knee arthroplasty: VERITAS, a randomized controlled trial. J Bone Joint Surg Am 2020;102(2):101-109. [doi: 10.2106/JBJS.19.00695] [Medline: 31743238]
34. Stucki G, Cieza A, Melvin J. The international classification of functioning, disability and health (ICF): a unifying model for the conceptual description of the rehabilitation strategy. J Rehabil Med 2007;39(4):279-285 [FREE Full text] [doi: 10.2340/16501977-0041] [Medline: 17468799]
35. Smith TO, Hawker GA, Hunter DJ, March LM, Boers M, Shea BJ, et al. The OMERACT-OARSI core domain set for measurement in clinical trials of hip and/or knee osteoarthritis. J Rheumatol 2019;46(8):981-989 [FREE Full text] [doi: 10.3899/jrheum.181194] [Medline: 30647185]
36. Chiarotto A, Maxwell L, Terwee CB, Wells GA, Tugwell P, Ostelo RW. Roland-Morris Disability Questionnaire and Oswestry Disability Index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? Systematic review and meta-analysis. Phys Ther 2016;96(10):1620-1637. [doi: 10.2522/ptj.20150420] [Medline: 27081203]
37. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
38. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
39. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
40. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-605. [doi: 10.1007/s10654-010-9491-z] [Medline: 20652370]
41. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
42. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. GRADEpro. 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2024-11-19]

XSL•FO
RenderX

43. Stata Statistical Software Release 13. StataCorp. URL: https://www.sciepub.com/reference/117495 [accessed 2024-11-19]
44. Doiron-Cadrin P, Kairy D, Vendittoli PA, Lowry V, Poitras S, Desmeules F. Feasibility and preliminary effects of a tele-prehabilitation program and an in-person prehablitation program compared to usual care for total hip or knee arthroplasty candidates: a pilot randomized controlled trial. Disabil Rehabil 2020;42(7):989-998. [doi: 10.1080/09638288.2018.1515992] [Medline: 30638076]
45. Yi JL, Porucznik CA, Gren LH, Guan J, Joyce E, Brodke DS, et al. The impact of preoperative mindfulness-based stress reduction on postoperative patient-reported pain, disability, quality of life, and prescription opioid use in lumbar spine degenerative disease: a pilot study. World Neurosurg 2019;121:e786-e791. [doi: 10.1016/j.wneu.2018.09.223] [Medline: 30312812]
46. Chavez JL, Porucznik CA, Gren LH, Guan J, Joyce E, Brodke DS, et al. The impact of preoperative mindfulness-based stress reduction on postoperative outcomes in lumbar spine degenerative disease: 3-month and 12-month results of a pilot study. World Neurosurg 2020;139:e230-e236. [doi: 10.1016/j.wneu.2020.03.186] [Medline: 32278820]
47. Yin B, Goldsmith L, Gambardella R. Web-based education prior to knee arthroscopy enhances informed consent and patient knowledge recall: a prospective, randomized controlled study. J Bone Joint Surg Am 2015;97(12):964-971. [doi: 10.2106/JBJS.N.01174] [Medline: 26085529]
48. Soeters R, White PB, Murray-Weir M, Koltsov JCB, Alexiades MM, Ranawat AS, HipKnee Surgeons Writing Committee. Preoperative physical therapy education reduces time to meet functional milestones after total joint arthroplasty. Clin Orthop Relat Res 2018;476(1):40-48 [FREE Full text] [doi: 10.1007/s11999.0000000000000010] [Medline: 29529614]
49. Culliton SE, Bryant DM, MacDonald SJ, Hibbert KM, Chesworth BM. Effect of an e-Learning tool on expectations and satisfaction following total knee arthroplasty: a randomized controlled trial. J Arthroplasty 2018;33(7):2153-2158. [doi: 10.1016/j.arth.2018.02.040] [Medline: 29555496]
50. An J, Ryu HK, Lyu SJ, Yi HJ, Lee BH. Effects of preoperative telerehabilitation on muscle strength, range of motion, and functional outcomes in candidates for total knee arthroplasty: a single-blind randomized controlled trial. Int J Environ Res Public Health 2021;18(11):6071 [FREE Full text] [doi: 10.3390/ijerph18116071] [Medline: 34199913]
51. Gallois M, Davergne T, Ledinot P, Ravaud P, Regnaux JP. Dosage of preventive or therapeutic exercise interventions: review of published randomized controlled trials and survey of authors. Arch Phys Med Rehabil 2017;98(12):2558-2565.e10. [doi: 10.1016/j.apmr.2017.03.030] [Medline: 28465221]
52. Vasileiadis D, Drosos G, Charitoudis G, Dontas I, Vlamis J. Does preoperative physiotherapy improve outcomes in patients undergoing total knee arthroplasty? A systematic review. Musculoskeletal Care 2022;20(3):487-502. [doi: 10.1002/msc.1616] [Medline: 35122455]
53. Shelton E, Barreto NB, Bidwell S, Folk-Tolbert M, Shelton A, Trickey AW, et al. Engagement and adherence with a web-based prehabilitation program for patients awaiting abdominal colorectal surgery. J Gastrointest Surg 2021;25(12):3198-3207. [doi: 10.1007/s11605-021-05171-2] [Medline: 34668165]
54. Robinson A, Husband AK, Slight RD, Slight SP. Digital support for patients undergoing bariatric surgery: narrative review of the roles and challenges of online forums. JMIR Perioper Med 2020;3(2):e17230 [FREE Full text] [doi: 10.2196/17230] [Medline: 33393926]
55. Widmer P, Oesch P, Bachmann S. Effect of prehabilitation in form of exercise and/or education in patients undergoing total hip arthroplasty on postoperative outcomes-a systematic review. Medicina (Kaunas) 2022;58(6):742 [FREE Full text] [doi: 10.3390/medicina58060742] [Medline: 35744005]
56. Walker LC, Clement ND, Deehan DJ. Predicting the outcome of total knee arthroplasty using the WOMAC score: a review of the literature. J Knee Surg 2019;32(8):736-741. [doi: 10.1055/s-0038-1666866] [Medline: 29991079]

Abbreviations

GRADE: Grading of Recommendations Assessment, Development and Evaluation
NOS: Newcastle Ottawa Scale
NRIS: nonrandomized intervention study
PrAT: prehabilitation with advanced technologies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB: risk of bias
TIDier: Template for Intervention Description and Replication
TKA: total knee arthroplasty
VAS-BP: visual analogue scale for back pain
WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 20.09.23; peer-reviewed by M Pritchard, M Cioeta, LA Gomes, N Ramadanov; comments to author 30.05.24; revised version received 24.07.24; accepted 31.10.24; published 12.12.24.

Please cite as:
Guida S, Vitale JA, Swinnen E, Beckwée D, Bargeri S, Pennestrì F, Castellini G, Gianola S
Effects of Prehabilitation With Advanced Technologies in Patients With Musculoskeletal Diseases Waiting for Surgery: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e52943
URL: https://www.jmir.org/2024/1/e52943
doi:10.2196/52943
PMID:

©Stefania Guida, Jacopo Antonino Vitale, Eva Swinnen, David Beckwée, Silvia Bargeri, Federico Pennestrì, Greta Castellini, Silvia Gianola. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Prevention and Health Promotion Interventions for Young People in the Context of Digital Well-Being: Rapid Systematic Review

Michelle Colder Carras[1,2], PhD; Dahlia Aljuboori[1], MBChB, MHS; Jing Shi[3], MSc (OT), PhD; Mayank Date[1], BDS, MPH; Fatima Karkoub[1], MHS; Karla García Ortiz[4], MD; Fasika Molla Abreha[1], MSc, MPH, MD; Johannes Thrul[1,5,6], MS, PhD

[1]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[2]Carras Colder Carras, Ellicott City, MD, United States

[3]Health and Social Sciences, Singapore Institute of Technology, Singapore, Singapore

[4]Department of Family and Preventive Medicine, Emory University School of Medicine, Atlanta, GA, United States

[5]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, United States

[6]Centre for Alcohol Policy Research, La Trobe University, Melbourne, Australia

Corresponding Author:
Michelle Colder Carras, PhD
Carras Colder Carras
3375 Ellicott Center Dr Unit 1675
Ellicott City, MD, 21043
United States
Phone: 1 410 412 0797
Email: michelle@carrasandco.com

Abstract

Background: Increasing digital technology and media use among young people has raised concerns about problematic use and negative consequences. The formal recognition of a technology addiction (eg, gaming disorder) requires an understanding of the landscape of interventions designed to prevent this disorder and related technology addictions.

Objective: We conducted a rapid systematic review to investigate the current evidence on approaches to prevent problematic technology use and promote digital well-being, defined as the healthy use of digital media and technology and the absence of problems resulting from excessive use.

Methods: We used a pragmatic and rapid approach to systematically review and synthesize recent literature with a focus on contextual factors that can aid in understanding translatability, making trade-offs appropriate for rapid reviews per the Cochrane Collaboration guidelines. We searched multiple databases, including gray literature, for primary studies and systematic reviews of prevention interventions targeting children, adolescents, and youth. We extracted data on study characteristics, quality, and translatability and synthesized evidence through narrative description and vote counting of controlled trials. Data are openly available on our Open Science Framework website.

Results: We found 6416 citations, of which 41 (0.64%) were eligible for inclusion (6 reviews and 35 primary studies of 33 interventions). Most interventions (26/33, 79%) combined intervention approaches and included an education component. Synthesis through vote counting showed benefits for all forms of digital well-being. Both included meta-analyses reported small positive effects on reductions of screen time. However, study reporting was overall lacking, impairing the ability to draw conclusions.

Conclusions: As digital technology use increases, interventions to prevent problematic technology use and promote digital well-being continue to proliferate. Understanding context factors that influence healthy technology use and understanding the limitations of the current evidence are vital for informing future research. This review demonstrates positive findings for the effectiveness of prevention interventions and describes factors that may contribute to translation and implementation. Future research would benefit from following appropriate reporting guidelines, reporting both the benefits and harms of interventions, and including greater detail on factors informing translation.

Trial Registration: PROSPERO CRD42023444387; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=444387

(J Med Internet Res 2024;26:e59968) doi:10.2196/59968

XSL•FO
RenderX

KEYWORDS

digital well-being; internet addiction; gaming disorder; social media; screen time; prevention; children; adolescents; mobile phone; PRISMA

Introduction

Background

The proliferation of digital technologies has ushered new possibilities for connection, communication, and facilitation of life tasks such as work and studying but also concerns about potential issues associated with time spent in front of screens, including excessive gaming and social media use. Problematic use of digital media and technologies is a multifaceted issue encompassing different behaviors and different impacts for various age groups. Although professional organizations have recommendations regarding the use of various types of digital media [1,2], clinically diagnosable problematic digital media use has been characterized as a disorder only regarding video games [3,4]. However, research continues to address other problematic digital behaviors, including addiction to the internet broadly [5,6], social media [7-11], and smartphone use [11-13]. Although there are disagreements about what constitutes normative use and how it differs between media types, media affordances, and generations of users [14], most researchers agree that, for some users, technology- and digital media–related behaviors can cause significant life interference at times, leading to the need for treatment and intervention approaches to promote healthy digital media use.

A growing body of research has started to address this concern through interventions to prevent problematic use and promote digital well-being, which we define as the healthy use of digital media and technology and the absence of problems resulting from excessive use, consistent with other definitions that focus on balancing the benefits and drawbacks to various life domains, such as mental, physical, and social [15,16]. However, the diverse nature of problematic digital media use and the heterogeneity of these interventions make it challenging to evaluate their efficacy [5,17]. While much of the earlier research summarized in previous reviews focuses on excessive involvement in online video games, multiple forms of digital media are now used extensively in daily life, providing a new population norm that differs from that in earlier research on technology addictions.

Previous reviews synthesizing evidence on interventions to prevent problematic use and promote digital well-being face several challenges. Reviewers must contend with a field in which new measures of problematic use continue to proliferate even with 2 very specifically defined technology or media use disorders [17]. In addition, research often takes place outside of disciplines such as public health that have rigid standards for evidence synthesis, leading to incomplete or selective outcome reporting [18]. Context is also a challenge; public and scientific concern about the extent of problematic digital media use varies across global regions, leading to differences in intervention targets and approaches. For example, policy approaches differ between global regions [17,19], and studies about problematic digital media use often downplay contextual factors that may contribute to the feasibility of interventions in various settings [20,21]. Successful public health interventions hinge on a deep understanding of the social and environmental factors that contribute not only to health but also to intervention implementation [22]. As such, understanding context is vital for evaluating what might work to prevent problematic digital media use and promote digital well-being in global settings.

Using frameworks that account for constellations of potential risk factors can provide insights into evidence on prevention interventions [23,24]. To better contextualize potentially modifiable factors that affect the development of digital well-being or problematic use across the life span, we adapted the development of a digital well-being framework [25], which starts from a social ecological model and incorporates cognitive theories [26], media context [27], traditional approaches to the development of addictions (ie, incentive sensitization theory [28]), and uses and gratifications [29].

Objectives

The aim of this rapid review was to describe the recent literature on interventions used in children, adolescents, and youth (aged <25 years) to prevent problematic digital media and technology use and promote digital well-being, with a focus on understanding the context of interventions and theoretical approaches from a developmental framework. The review methods, including the research question (RQ), search strategy, inclusion and exclusion criteria, and risk-of-bias assessment, were developed a priori and described in the registered protocol (PROSPERO CRD42023444387); these are also available in Multimedia Appendix 1 [8-13,30-58].

Methods

Overview

The registered PROSPERO protocol; description of the population, intervention, comparator, outcome, and study design (PICOS) framework; search strategies; and other supporting documentation are available in Multimedia Appendix 1 [8-13,30-58]. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist is available in Multimedia Appendix 2. Data extraction forms and databases, including a list of articles excluded at the full-text screening stage with reasons for exclusion and all extracted data, extraction of all data, and results of the tool named A Measurement Tool to Assess Systematic Reviews (AMSTAR) checklists, are available on our Open Science Framework (OSF) website [59]. As this study is a systematic review of the literature, it is not considered human participant research.

Study Design and RQs

Our review included both systematic reviews (henceforth called *reviews*) and primary studies (henceforth called *studies*). As this was not human participant research, institutional ethics approval was not necessary. We used a pragmatic approach to systematically review and synthesize recent literature rapidly

XSL•FO
RenderX

(approximately 9 months) with a focus on contextual factors that can aid in understanding translatability, making trade-offs in our methods appropriate for rapid reviews as per recommendations from the Cochrane Rapid Reviews Methods Group [60]. For example, we limited the number of outcomes to two (first, digital wellbeing or problematic use, and second, time spent on digital media); we restricted the search dates; and we used a single-extractor approach to data extraction, with 10% checked by a second team member. We adapted a conceptual model that combines theoretical frameworks focusing on individual development of addictions and media effects as well as models of behavior that combine individual-level and contextual factors [25-29,61].

We addressed the following RQs developed with stakeholder input:

1. What health promotion and prevention interventions have been used to promote digital well-being and prevent problematic digital media and technology use in children, adolescents, and youth?
2. What theoretical or treatment models and approaches inform the development of these interventions?
3. How effective are these interventions when compared with other interventions or no intervention?
4. What characteristics of intervention setting or delivery may limit or promote translation to other contexts?
5. What quality issues should be addressed in future empirical studies of these interventions?

Search Strategy and Screening

We conducted searches in March 2023 combining terms as described in this section. Citations were downloaded into Zotero (Corporation for Digital Scholarship) [62] and then uploaded into Covidence (Veritas Health Innovation) [63] for screening and inclusion. To update previous literature reviews and allow for a rapid review focus on methodological rigor in data extraction, synthesis, and risk-of-bias assessment, we searched for empirical studies and systematic reviews published since 2017 in all languages in PsycINFO, Web of Science, and PubMed, as well as the gray literature databases of the World Health Organization (IRIS database) and ClinicalTrials.gov. We also hand searched the reference lists of review articles. We piloted combinations of various search terms that included subject and text keywords to retrieve papers related to the following: (1) children, adolescents, and young adults (aged ≤25 years) who did not meet the criteria for technology use disorders (participants); (2) interventions related to promoting healthy digital media use or preventing excessive or problematic use (interventions or exposures); (3) any comparator or no comparator (comparators); (4) digital well-being or the prevention of problematic or disordered use, including symptoms of problematic use and time spent on digital media (outcomes); and (5) primary studies and reviews of interventions (study type).

The final search terms can be found in Multimedia Appendix 1 [8-13,30-58]. We included intervention studies of various designs, protocols, and systematic reviews from all settings that (1) addressed primary or secondary prevention (ie, used a sample that was not formally diagnosed with digital media use disorders, such as internet gaming disorder, where the aim was prevention rather than treatment) *and* (2) contained outcomes related to time using digital media or technology or problematic or excessive use even if these were not primary outcomes. We excluded studies that (1) focused on clinical treatments or a sample with disordered levels of behavior or symptoms (eg, focused on "addicted" individuals), (2) focused primarily on other behavioral addictions, (3) lacked outcomes measuring digital media or technology use or addiction symptoms, or (4) had a mean sample age of >25 years. Titles and abstracts and then full texts were screened independently by 2 raters; discrepancies were resolved through discussion. In cases where study samples seemed to overlap, we contacted the study authors to ensure that the studies were distinct.

Data Extraction

We developed and pilot-tested a data extraction form based on the standards for rapid reviews [64] and evaluation of complex interventions [65], AMSTAR 2 criteria [66], and the Cochrane Eyes and Vision work group [67]. Data extracted from primary studies included PICOS criteria; theoretical approach, study aims, and design; facilitators or barriers for the study context; possible sources of bias; harms reported; authors' conclusions; and evaluation of quality and translatability using adapted forms of the Critical Appraisal Skills Programme (CASP) checklists [68,69]. One investigator extracted data from each study, and then a second investigator reviewed the data extraction for all studies, including reviews.

We classified the outcome as the broadest category when categories were combined; for example, "excessive media use" ("exzessive Mediennutzung" [8]) and "GD [gaming disorder] and unspecified IUD [internet use disorder]" [30] were designated as problematic internet use. Where social networking–related categories were combined with another category but were the subject of the intervention, we designated social networking as the focus, for example, "Internet addiction *to the social networks*" (emphasis added) [31].

Efficacy outcome data were extracted from tables, text, and figures in Microsoft Excel (Microsoft Corp) and Stata (StataCorp) to create counts for vote counting for controlled trials only. Where possible, use of digital media was converted to minutes per day (eg, from hours per day) to facilitate comparisons. For studies reporting univariate and multivariate outcomes, we extracted all outcomes.

Data extracted from reviews included citation information, type of review, aims and objectives, target population and setting, eligibility criteria for inclusion and exclusion, methods for specification of outcomes chosen for inclusion, measurement of outcomes, results, authors' conclusions, and quality and risk-of-bias criteria. The AMSTAR 2 checklist was completed for each review and can be found on our OSF website along with the review data extraction database. Due to time constraints, missing data were not requested from the authors.

XSL•FO
RenderX

Quality and Reliability Assessment

Primary Studies

Because this was a rapid review focusing on context and translatability, we used critical appraisal rather than formal risk-of-bias assessment and scoring to evaluate. CASP checklists [68,69] were adapted for controlled trials, before-and-after studies, and observational studies to focus on addressing quality and translatability outside the original study context. Data were extracted by 2 reviewers independently, and discrepancies were resolved through discussion.

Reviews

We adapted an existing approach to quality and risk-of-bias assessment based on a definition of reliability developed by Cochrane Eyes and Vision [67]. This definition specifies that reviews can be defined as *reliable* when they define eligibility criteria for study inclusion; conduct a comprehensive literature search, including articles not in English and multiple search methods; assess the risk of bias of individual studies; use appropriate methods for meta-analysis where conducted; and present conclusions that are supported by the evidence reported in the review. To address potential selective outcome reporting, we added the specification that review authors should have clearly specified in the methods or protocol which outcomes from their eligible studies were included. Reviews were classified as reliable only when all the aforementioned 6 criteria were met. We also conducted a full assessment of review quality using the AMSTAR 2 checklist [66].

Data Synthesis

Consistent with Cochrane recommendations for a rapid review approach [60], we synthesized evidence through multiple tables and figures and provided a narrative synthesis of findings structured around intervention characteristics and approach, level of prevention, and population and setting. Due to the rapid nature of this review, we conducted vote counting to aggregate results from randomized controlled trials (RCTs) and other controlled studies that directly compared an intervention to no intervention. We compared the number of outcomes favoring the intervention versus control groups (ie, beneficial vs harmful effects), focusing on the direction of the effect rather than on statistical significance per revised Cochrane guidelines (a departure from our registered protocol) [70]. We then conducted a binomial proportion test using the *bitesti* command in Stata to determine the probability of observing the distribution of beneficial versus harmful effects by chance. We used the *cii proportions* command in Stata to estimate the SE and Jeffrey CI for this proportion.

Results

The search resulted in 6416 citations, of which 41 (0.64%) were eligible for inclusion (n=35, 85% primary studies and n=6, 15% reviews; Figure 1). The time from completing the search to analyzing and writing up the manuscript was 6.5 months.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. WB: well-being; WHO: World Health Organization.

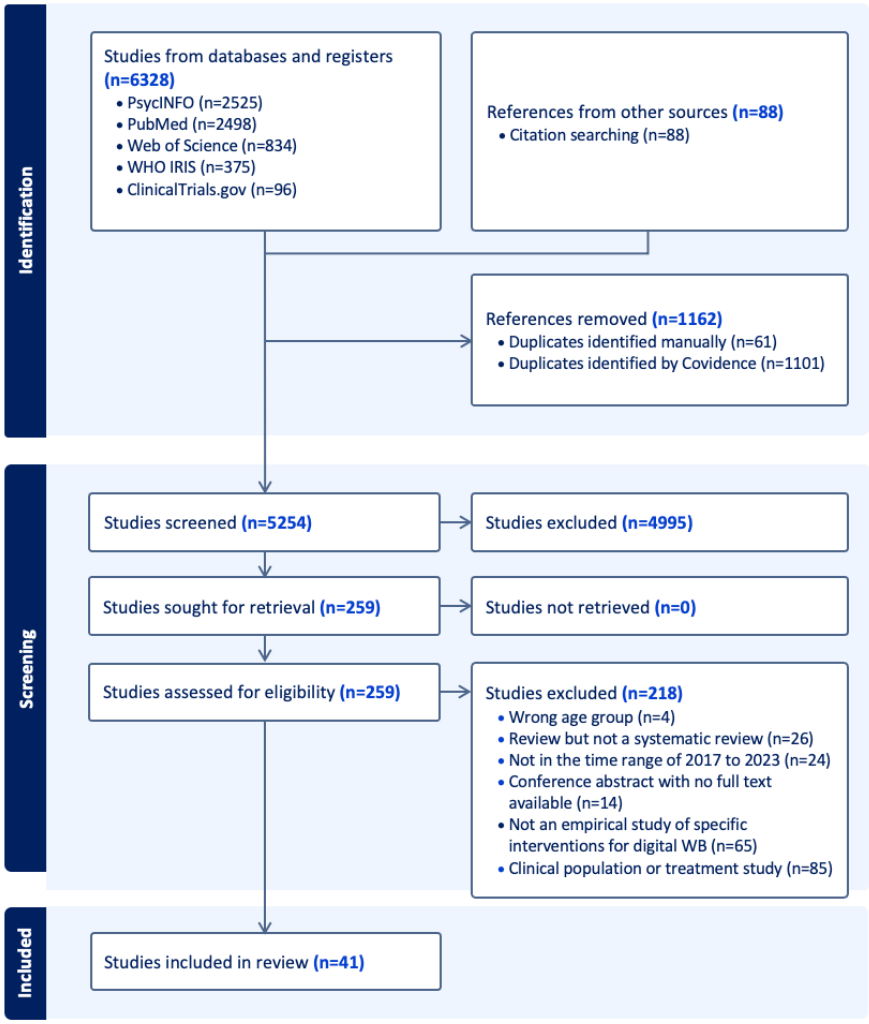

Participants

The studies took place in 19 countries and included between 9 and 243,957 participants, with an average of 8282 (SD 43,744). One national study in South Korea accounted for 243,957 of the participants [32]; when that study was excluded, the maximum number of participants was 2997, and the average was 426 (SD 613). The systematic reviews included between 11 and 204 studies (mean 58, SD 73); 33% (2/6) of the systematic reviews conducted meta-analyses. The systematic reviews included a range of countries, and the total numbers of participants in the systematic reviews ranged from 5627 (calculated) to 162,494 (calculated). The studies focused on children aged 0 to 12 years (3/35, 9% studies), adolescents aged 13 to 18 years (8/35, 23%), and youth aged 19 to 25 years (5/35, 14%), whereas an additional 31% (11/35) of the studies combined more than one age group, 17% (6/35) focused on parents or children and parents, and 6% (2/35) focused on other groups [8,13].

Study Designs and Comparators

Table 1 provides an abbreviated overview of the characteristics of the studies of individual interventions included in our review; a full summary-of-findings table can be found in Multimedia Appendix 1 [8-13,30-58]. Overall, we found 33 different interventions in 35 studies. A total of 54% (19/35) of the studies were controlled trials that compared interventions to a control group (waitlist or no intervention), 20% (7/35) were before-and-after studies, 23% (8/35) were observational studies or protocols, and 3% (1/35) were unclear [31]. Of the controlled trials, some (3/19, 16%) did not compare extracted outcomes directly between intervention and control groups using statistical tests (eg, compared only within-group changes) and were not included in the vote counting. Of the 27 studies (77%) using causal language in their aims (eg, aiming to “analyze or examine effectiveness or investigate or examine [the] effects” of interventions), several (12/27, 44%) did not use a before-and-after design that compared an intervention group and a control group directly. Most reviews (4/6, 67%) required prospective study designs, but 33% (2/6) included protocols [5,71]; no review included only RCTs.

XSL•FO
RenderX

Table 1. Overview of the characteristics of the included studies.

Study	Population and study design	Intervention description	Relevant outcomes
Controlled studies			
Primary prevention			
Education			
Apisitwasana et al [33], 2018	Cluster RCT[a] of 310 fourth- and fifth-grade students in 2 schools in Bangkok, Thailand	Classroom activities and educational material provided to parents	Game addiction
Bickham et al [34], 2018	Controlled before-and-after study of 412 sixth- to eighth-grade students in the United States	Take the Challenge—a school-based media education and use education program	Television viewing, video gaming, and internet use
Bonnaire et al [35], 2019	Controlled before-and-after study of 384 middle school students in Paris, France	Single-session prevention intervention designed to increase knowledge and skills	Proportion of gamers with IGD[b] and minutes per week on the internet and video games
Esmaeili Rad and Ahmadi [36][c], 2018	RCT comparing intervention to control (no intervention) conducted with 200 college students in Urmia, Iran	Use of a mobile app to decrease social networking addiction	Online social networking addiction and time spent on social networks
Gui et al [13], 2023	Cluster-randomized trial of the impact of a teacher training course on 2997 tenth-grade students in 171 classes in 18 schools in northern Italy	DWB-S[d], a hybrid web-based and in-person training intervention for teachers	Smartphone addiction and social network use
Hansen et al [8], 2022	Postintervention survey of 834 students in participating schools in 3 federal states in Germany	The “Net-Piloten” (Net Pilots) peer project for the prevention of social media addiction	Video game play duration, social media chatting time, streaming time, and problematic internet use
Li et al [37], 2019	Cluster RCT comparing intervention schools (n=35) to control schools (n=35); 362 parents of fourth- to sixth-grade gamers in Hong Kong, China	GOI[e], a single-session, 4-hour intervention for parents	Gaming time and gaming disorder symptoms
Marco and Choliz [38][c], 2017	RCT of 1110 primary and secondary school students in 9 public and 4 private schools in Valencia, Spain	PrevTec 3.1 for video games	Frequency and duration of video gaming and video game dependence
Mumcu et al [39][c], 2021	RCT of 110 schoolchildren in Turkey	Recreational exercises in school	Game addiction
Ortega-Barón et al [40], 2021	Controlled pretest-posttest study of 165 middle schoolers in 3 regions of Spain	The Safety.net program	Problematic internet use
Weaver [41], 2022	Pretest-posttest study of 65 students in a single high school in the United States	Mindful Connections	Problematic social media use
Community			
Sanders et al [42], 2018	RCT of parents of children between the ages of 5 and 12 years	Single-session, in-person intervention delivered by a therapist to address media parenting	Screen time
Online			
Throuvala et al [11], 2020	RCT of 143 university students in the United Kingdom who used mobile phones and social media	An intervention designed to promote mindfulness, raise awareness of media and smartphone use, enhance self-regulation, and reduce distractions and time spent on smartphones and social media	Hours per day of smartphone use, problematic social media use, and hours per day of social media use
Secondary prevention			
Education			

XSL•FO
RenderX

Study	Population and study design	Intervention description	Relevant outcomes
Affouneh et al [43], 2021	RCT with 30 university students in West Bank, Palestine, who "engaged in excessive Internet usage"	Group training (CBT[f]-based) to improve social skills and reduce internet-addictive behaviors	Internet addiction
Lindenberg et al [30], 2022	Multicenter cluster RCT of 422 high school students in Germany	PROTECT[g] preventive group intervention	Combined gaming disorder and unspecified internet use disorder (measured using a single modified scale) and intervention harms (number of participants who developed combined gaming disorder and unspecified internet use disorder)
Tang et al [44], 2021	Cluster RCT of 775 students at 2 middle schools in Zizhong County, a rural area of Sichuan Province, China	The intervention was primarily provided by teachers in health education courses, whereas the final phase of the intervention included sports equipment that could be used during class	Proportion of excessive users, proportion of those gaming or spending time on the internet for ≥2 hours a day, and proportion of those being on the internet overnight
Online			
Brailovskaia et al [12], 2022	RCT of 619 young adult smartphone users (at least 75 minutes per day) in Germany	Abstinence vs reduction of smartphone use by 1 hour (vs control)	Daily smartphone use and problematic smartphone use
Ko et al [9][h], 2021	Controlled pretest-posttest study of 16 young adults who perceived themselves as social media addicts	Social Media Addiction Coach	Social media use
Thai et al [10], 2023	RCT of 260 (220 completed) college students who regularly used social media (>45 minutes per day) in Canada who had symptoms of depression or anxiety	IG[i] was asked to limit social media use to 1 hour per day for 3 weeks	Smartphone use
Pretest-posttest studies without control groups			
Primary prevention			
Education			
Chau et al [45], 2019	Before-and-after study of 248 primary school students in Hong Kong, China	Wise IT use program	Distribution of normal, at-risk, and high-risk gamers and levels of IGD symptoms
Ke and Wong [46], 2018	Pretest-posttest study of 45 teenage students from government secondary schools in Malaysia	PIP-IU-Y[j]	PIU[k]
Ke and Wong [47], 2018	Pretest-posttest study of 157 teenage students from government secondary schools in Malaysia	PIP-IU-Y	PIU
Kent et al [48], 2021	Pretest-posttest study of 10 undergraduate students in the United Kingdom	An intervention based on the stages of change model was delivered via a smartphone app	Problematic mobile phone use and internet addiction measured by IAT[l]
Online			
King et al [49], 2017	27 young adult MMO[m] players	Participants were asked to refrain from gaming over a single weekend	Gaming hours and IGD symptoms
Secondary prevention			
Community			
Männikkö et al [50], 2022	Pretest-posttest study of 37 young adults recruited from 5 cities in Finland who self-reported excessive gaming	The Limitless Gaming Bootcamp	Problematic gaming, gaming time, internet use, and television viewing time
Other			

XSL•FO
RenderX

Study	Population and study design	Intervention description	Relevant outcomes
Heffler et al [51], 2022	Pretest-posttest study of 9 preschool-aged children diagnosed with autism who viewed screens for at least 2 hours a day	Parents viewed a 40-minute educational video and then received weekly 1-hour in-home support visits from a trained therapist	Parents' reports of children's screen time and adverse effects of keeping screens off
Observational studies and protocols			
Primary prevention			
Education			
Barker [52], 2021	Protocol only; program will target preadolescent students in the United States who play internet games	Mindfulness and DBT[n] group psychoeducation and skill building to prevent problematic internet gaming	—[o]; protocol only
Hansen et al [53], 2021	A cross-sectional survey conducted with 542 "multipliers" who had been trained across Germany to deliver the program	"Net-Piloten" (Net Pilots) peer project for the prevention of social media addiction	Implementation of the intervention, barriers, and potential solutions
Neverkovich et al [31][p], 2017	Experimental study of 657 high school, college, and medical school students in Moscow and Irkutsk, Russia	Psychological and educational support programs providing motivational, cognitive, practice-oriented, and reflective units	Internet addiction and social network addiction
Tang et al [54], 2021	Protocol for a cluster-randomized RCT with a proposed sample of 240 level-5 primary school students in Hong Kong, China	MBCP[q] aimed at improving adolescents' resilience, changing their smartphone use behavior, and reducing smartphone addiction symptoms	—; protocol only
Community			
Krossbakken et al [55], 2018	A posttest-only experimental study of 1762 (1657 analyzed) guardians of young children (aged 8-12 years) in Norway	The intervention consisted of a single mailed brief parental guide on "how to regulate video game behavior in children"	Gaming time and video game problems (IGD criteria)
Sela [56], 2021	Protocol only	Training and assistance in setting parental controls (TPM[r]) vs training in PVC[s] vs both (TPM+PVC)	—; protocol only
Online			
Hayes and Jones [57], 2022	Protocol only; the study was terminated due to inability to recruit and retain sufficient participants	A 6-session web-based group intervention targeting problematic smartphone use	—; study was terminated
Schmuck [58], 2020	Cross-sectional web-based survey of 500 young adults in Austria who used a smartphone	Self-reported existing use of one of several digital detoxification apps for smartphone, such as iOS Screen Time, Android Digital Wellbeing, Moment, Forest, QualityTime, Detox, Space, Offtime, RealizeD, or similar	Problematic smartphone use
Other			

XSL•FO
RenderX

Study	Population and study design	Intervention description	Relevant outcomes
Choi et al [32], 2018	Panel survey of a representative sample of 243,957 middle and high schoolers in South Korea	Article 26 of the Juvenile Protection Act (shutdown policy): "Internet games should not be offered to those under the age of 16 from 12:00 am to 6:00 am"	Problematic use (>300 minutes per day, called "addiction" in the study) and weekly internet use

[a]RCT: randomized controlled trial.
[b]IGD: internet gaming disorder.
[c]Excluded from vote counting because of lack of direct comparison between the intervention and control groups.
[d]DWB-S: Digital Well-Being–Schools.
[e]GOI: Game Over Intervention.
[f]CBT: cognitive behavioral therapy.
[g]PROTECT: Professioneller Umgang mit technischen Medien (Professional Use of Technical Media).
[h]Excluded from vote counting because of unclear analysis reporting (direction of effect unclear).
[i]IG: intervention group.
[j]PIP-IU-Y: Psychological Intervention Program–Internet Use for Youth.
[k]PIU: problematic internet use.
[l]IAT: Internet Addiction Test.
[m]MMO: massively multiplayer online game.
[n]DBT: dialectic and behavioral therapy.
[o]Not applicable.
[p]Excluded from vote counting due to unclear study design and analysis.
[q]MBCP: mindfulness-based cognitive program.
[r]TPM: Technological Parental Monitoring.
[s]PVC: Parental Vigilant Care.

Outcomes

Overview

Outcomes were assessed using a variety of validated and ad hoc measures related to problematic use, including validated scales (eg, Internet Addiction Test [43]), subscales or questions from validated scales (eg, "two sub-dimensions of disturbance of adaptive functions and withdrawal from the Smartphone Addiction Scale" [13]), modified scales (eg, "a modified version of the brief Bergen Social Media Addiction Scale" [12]), and ad hoc questions (eg, "Technology-Related Parenting Strategies" [41]). Parent or school counselor reports were sometimes used for recruitment and eligibility criteria or for outcome measurement. No review limited measures to specific scales.

Synthesized Findings

RQ 1: What Health Promotion and Prevention Interventions Have Been Used to Promote Digital Well-Being and Prevent Problematic Digital Media and Technology Use in Children, Adolescents, and Youth?

This section describes the level of prevention, sector of implementation, and behavior or symptom targets. In total, 60% (21/35) of the primary studies described primary or universal prevention interventions, 37% (13/35) described secondary or selective prevention approaches, and 3% (1/35) of the studies [49] targeted both primary and secondary prevention. Of the 21 studies describing primary interventions, 17 (81%) were conducted in educational settings, 2 (10%) were conducted in community settings, 1 (5%) was online, and 1 (5%) was a national policy. Of the 13 secondary intervention studies, 8 (62%) were conducted in educational settings, 1 (8%) was conducted in a community setting, 1 (8%) was conducted in a health care setting, 2 (15%) were online, and 1 (8%) was unclear. The study targeting both primary and secondary prevention was online [49]. Of the 6 reviews, 1 (17%) included only universal or primary prevention studies [72], 1 (17%) included universal and selective or secondary prevention studies [73], and 1 (17%) included universal and indicated prevention studies [5]; 3 (50%) reviews did not indicate a prevention level.

Figure 2 illustrates the sample sizes by country and prevention level for all primary studies except a study of South Korea's Juvenile Protection Act [32], which had 243,957 participants.

XSL•FO
RenderX

Figure 2. Sample size and prevention level by country. This figure does not include a national study of >200,000 adolescents in South Korea, as this represents an outlier value that makes the visualization of other sample sizes impossible.

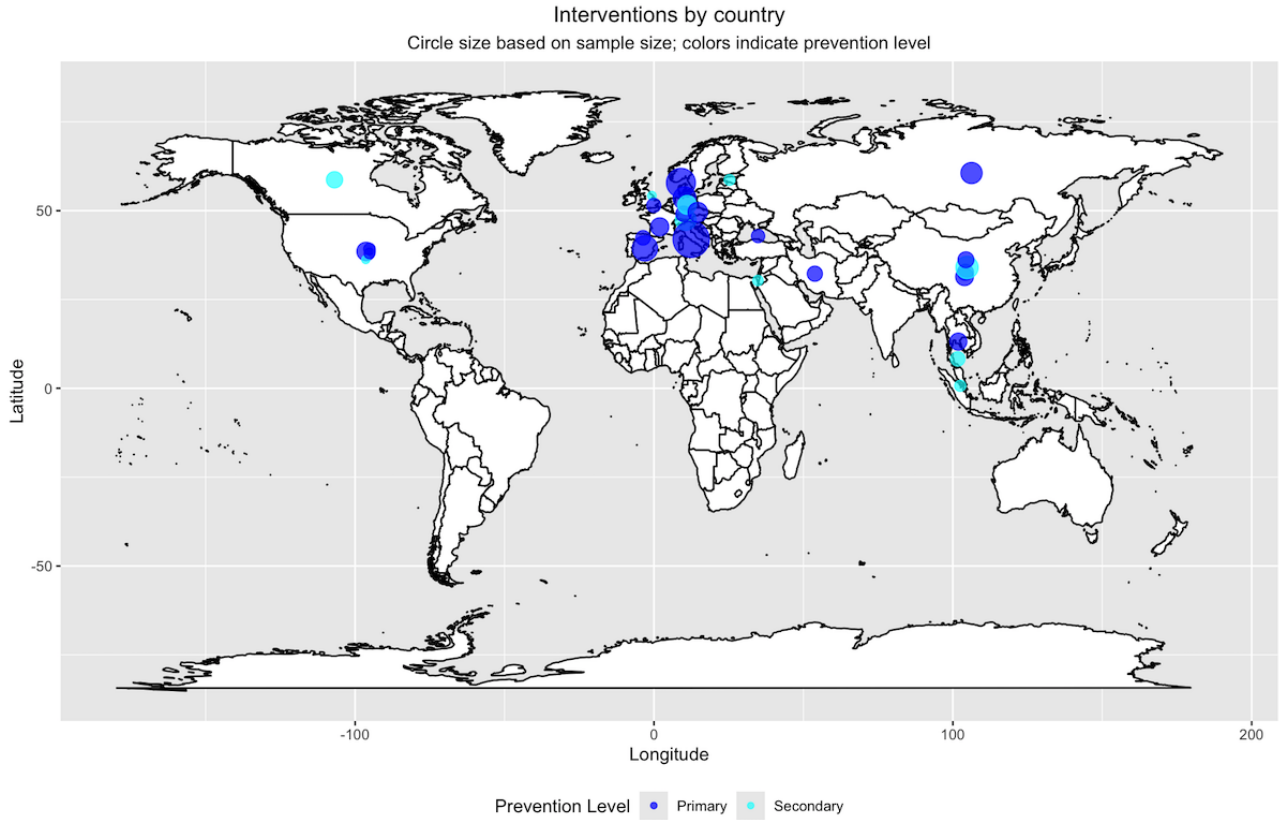

Sample sizes ranged from 32 to 243,957 (mean 13,386.42, SD 55,839.74; median 362) for primary prevention studies and 9 to 775 (mean 216.36, SD 272.75; median 45) for secondary prevention studies. A total of 67% (4/6) of the primary prevention studies with >1000 participants took place in Europe. Of the 9 studies with <100 participants, 5 (56%) were delivered by health care professionals or school counselors, and 3 (33%) were self-administered (eg, apps or online abstinence interventions). Only 6% (2/33) of the interventions were delivered without interaction with the target population (or teachers and parents): South Korea implemented the Juvenile Protection Act, which was designed to limit internet access for adolescents [32], and a study in Norway [55] mailed a brief parental guide with advice and strategies for regulating children's video gaming.

Various behaviors and problems associated with digital well-being were targeted as outcomes of the interventions. Internet use and game use were the most common intervention targets—39 outcomes were measured across 20% (7/35) of the studies for internet use, and 33 outcomes were measured across 29% (10/35) of the studies for game use. All reviews (Table 2) defined intervention targets regarding behaviors and addictions broadly. The reviews generally synthesized literature on screen time (4/6, 67%), although 17% (1/6) of the reviews examined prevention of internet addiction, and 17% (1/6) focused on abstinence interventions for various behavioral addiction outcomes [5,74].

XSL•FO
RenderX

Table 2. Characteristics of the included reviews

Review	Review type, aims, and population	Studies; participants[a], N	Results	Conclusions	Study quality criteria met; AMSTAR 2[b] rating
Fernandez et al [74], 2020	Systematic review of short-term abstinence interventions across potential behavioral addictions, including benefits or counterproductive consequences of abstinence	47; 8245[c]	Voluntary abstinence may be useful for specific problematic behaviors, especially gaming, pornography use, mobile phone use, and social media use	The paper describes the benefits of short-term abstinence but also discusses that harms were not addressed systematically in the studies. There were issues of reporting, including quality review (eg, outcome-level limitations were not addressed) and outcome reporting (it was unclear how "key findings" were selected for inclusion in the review).	No; critically low
Jones et al [73], 2021	Systematic review and meta-analysis of behavioral interventions to reduce children's screen time	204; 162,494[c]	The review found a small positive effect of interventions, with interventions targeting younger children and interventions with a shorter duration showing larger effects	This high-quality meta-analysis found overall benefits of screen time interventions and was systematic and complete about reporting how outcomes were chosen and describing limitations and other study quality criteria. However, potential harms were not examined. The authors note the importance of "determining the active ingredients to optimize interventions along the translational continuum."	Yes; critically low
Krafft et al [75], 2021	Systematic review of intervention strategies to reduce screen time among children from birth to 12 years of age	11; 5627[c]	Interventions can promote awareness of and sometimes reduce screen time	The findings were not consistent, and screen time "awareness" was sometimes described as a benefit. The paper lacked a clear description of how outcomes would be reported and did not discuss limitations at the outcome level or potential harms.	No; critically low
Oh et al [72], 2022	Systematic review and meta-analysis of interventions to promote healthy screen time and reduce sedentary behavior in school-aged children and adolescents in all settings	51; 16,418	The review found moderate quality of evidence for reduction in television viewing but low quality of evidence for overall screen time and no effects for gaming and studies with quasi-experimental designs.	This high-quality review showed the benefits of nondigital screen time interventions, particularly for television watching. The review did not address harms or discuss limitations at the study level but did systematically address bias and quality by grading the evidence.	Yes; critically low
Throuvala et al [5], 2019	Systematic review of prevention programs for internet addiction within the school context	20; 226,762[c]	This review highlighted the diversity in program scope and outcomes and mixed results in reducing internet and gaming use.	The review found very diverse interventions and results, drawing no overall conclusions as to the benefits or lack thereof. The review did not describe how outcomes would be reported, assess bias or quality, discuss harms, or address limitations other than at the study level.	No; critically low

XSL•FO
RenderX

Review	Review type, aims, and population	Studies; participants[a], N	Results	Conclusions	Study quality criteria met; AMSTAR 2[b] rating
Throuvala et al [71], 2021	Systematic review of intervention strategies to reduce screen time among adolescents	15; 17,241[c]	Interventions are more likely to reduce excessive screen time if they include strategies targeting other factors that drive internet use behaviors and consider activities separately rather than lumping them together in a single "screen time" construct.	The review highlighted the overall mixed effectiveness of the studies and the way in which screen time behaviors were measured in interventions aiming to reduce sedentary behavior and pointed out appropriately that interventions need to consider content, context, and reasons for use rather than solely assessing changes in amount of screen time. It did not evaluate harms from the interventions or describe how outcomes were chosen.	No; critically low

[a]The number of included studies in a review was taken from the review's PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (where possible) or from reports in the text or tables of each review.

[b]AMSTAR 2: A Measurement Tool to Assess Systematic Reviews-2.

[c]The total number of participants was calculated from a summary table or from the included studies.

RQ 2: What Theoretical or Treatment Models and Approaches Inform the Development of These Interventions?

Characteristics and approaches used in interventions were usually described in tables or appendices, but some studies reported information in a way that was difficult to interpret. We extracted intervention approaches and mapped these to our proposed conceptual framework (Figure 3), resulting in 11 mapped intervention approaches.

For example, we mapped the intervention approaches "gain more knowledge about gaming addiction and its effects" [33] and "raise awareness about the consequences of excessive use of video games on sleep, school investment, and family" [35] to digital well-being education (ie, education on the potential long-term problems and benefits of digital media or technology use) but mapped modules focusing on mental health symptoms, including "learning ways to reduce social anxiety" [47], to psychoeducation. Interventions including a self-regulation approach were most prevalent (22/33, 67%) and included aims related to regulation and time management directly [30,33,38].

Of the 33 interventions described in the 35 included studies, most (26/33, 79%) combined approaches. The most common components focused on education or training components, such as education about self-regulation, digital well-being education, media literacy and effects, or mental health symptoms and coping (psychoeducation). Combined interventions such as Net Pilots [8] focused on educating about self-regulation of digital technology use through critical examination of one's own use but were also delivered by peer "net pilots" (peer influence). Most interventions that combined approaches used 2 approaches (11/26, 42%), followed by 3 approaches (6/26, 23%) and 4 approaches (5/26, 19%). The most commonly combined intervention approaches were self-regulation and digital well-being education, followed by self-regulation and media literacy and effects. Interventions with a single approach were most often delivered via mobile phone and encouraged temporary abstinence from or reduction of social media or mobile phone use. Other single-approach interventions included an informational pamphlet to help parents promote healthy gaming among their children [55], a study of a national policy in South Korea [32], a protocol for a study comparing various types of parent training [56], and a protocol for a study of a mindfulness-based intervention [10].

Most review articles (4/6, 67%) specifically addressed intervention components that might promote effectiveness, looking for "active ingredients" associated with success [73]. The detailed meta-analysis by Jones et al [73] aimed to "...identify the behavior change techniques and study characteristics associated with effectiveness in behavioral interventions to reduce children's (0-18 years) screen time." This review used the Behavior Change Technique Taxonomy by Abraham and Michie [76] to analyze studies by behavior change technique, finding that interventions that included goals, feedback, and planning were more likely to be effective and that interventions delivered within less than a year and with smaller samples had larger effects than those delivered over a longer period or that had larger samples. Other reviews described intervention components, methods, and frameworks in tables or text.

XSL•FO
RenderX

Figure 3. Development of problematic digital media use and digital well-being (DWB) with mapped intervention approaches.

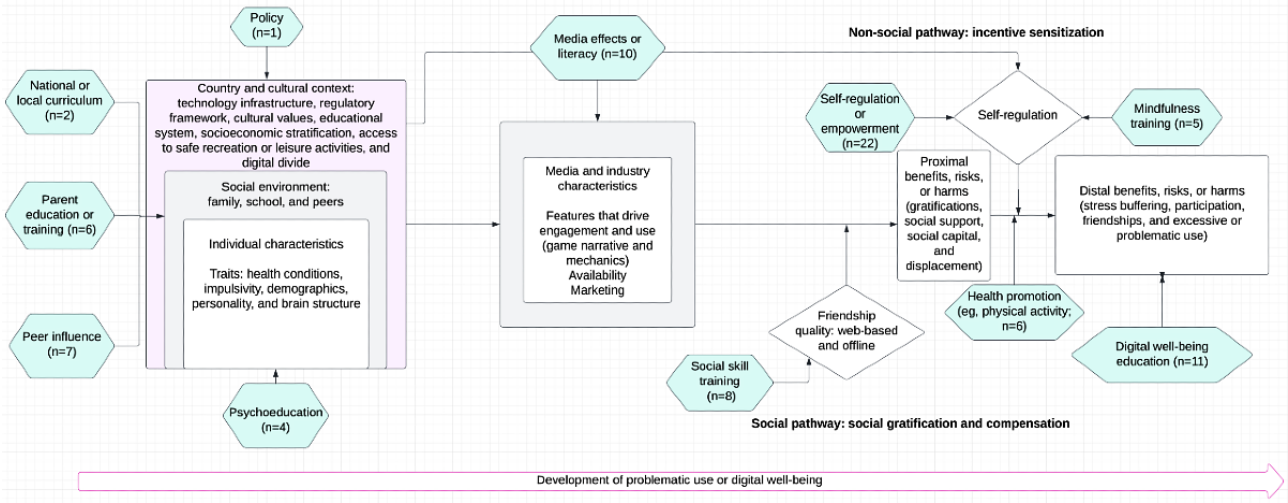

RQ 3: How Effective Are These Interventions When Compared to Other Interventions or No Intervention?

To determine the effectiveness of the interventions, we extracted relevant outcome data from 19 controlled trials (ie, RCTs and controlled before-and-after studies), as described previously, and then conducted vote counting for controlled trials of interventions that directly compared an intervention and a control group (14/19, 74%). A total of 71% (10/14) of the studies took place in the education sector, whereas 29% (4/14) of the studies were online. None of the included controlled studies took place in health care, the community, or a policy setting. No controlled trials delivered interventions to children only. Due to the differing characteristics of the interventions, their different target and delivery populations, and the rapid nature of our review, we did not synthesize outcomes quantitatively or conduct subgroup analyses. However, data are available on our OSF website.

Table 3 presents results for the 14 included controlled trials that examined our outcomes of interest. Of the 51 outcomes comparing intervention and control groups, most measured gaming time (n=14, 27% of the outcomes across 5/14, 36% of the studies). Internet use time, problematic gaming, and problematic internet use accounted for 12% (6/51) of the outcomes each. Interventions delivered to study populations other than young people (eg, parents or teachers) had, in general, longer follow-up periods (up to 3 months after baseline), although one intervention delivered to adolescents and youth [30] had follow-ups at 1, 4, and 12 months after the intervention.

Vote counting showed that, of the 51 outcomes, 44 (86%) favored the intervention group. Under the null hypothesis of no difference between groups, the probability of observing this result would be <0.001. The estimated Jeffrey 95% CI for this proportion ranged from 0.75 to 0.94, supporting improvements in digital well-being. Of those comparisons favoring the control group (7/51, 14%), 86% (6/7) reflected likely baseline differences [30,37]. Just 1 outcome from 14% (1/7) of the studies that favored the control group had an absence of apparent baseline differences [30].

Effects on minutes of daily use varied widely. The minimum absolute difference in mean minutes after the intervention was 2.6 (favoring the control group), whereas the maximum absolute difference was 110.5. The average difference in minutes per day for all interventions was 22.5. Considering only RCTs (10/15, 67%) made little impact—the average reduction in minutes of use across platforms fell to 18.6, and the range remained the same. The lowest impacts were found for gaming time and screen time, which showed an average reduction of 18.9 minutes per day (12 outcomes across 5/15, 33% of the studies for gaming time and 4 outcomes across 2/15, 13% of the studies for screen time). In contrast, interventions measuring time on social media showed an average reduction of 40.9 minutes per day (4 outcomes across 4/15, 27% of the studies), whereas interventions measuring smartphone use time showed an average reduction of 34.0 minutes per day, and those measuring internet use time showed an average reduction of 25.1 minutes per day.

Harms or unintended consequences of an intervention were not discussed in any of the assessed primary studies. One review [74] discussed harms and found that, generally, there was a lack of discussion about the potential adverse effects and unintended consequences of interventions (eg, rebound effects and more harmful compensatory behaviors) in reports of primary studies (see the OSF website [59], review study data extraction).

A total of 33% (2/6) of the reviews analyzed intervention effects using statistical analyses. A meta-regression of 204 studies [73] found a small positive effect of interventions to reduce screen time, with interventions including goals, feedback, and planning having slightly higher effect sizes (overall standard difference of the mean=0.116; interventions including goals, feedback, and planning=0.145). The other meta-analysis [72] found an overall small reduction in television viewing time (12.45 minutes) for nondigital interventions versus control groups in 23 studies with moderate quality of evidence but high uncertainty in effects on minutes per day of computer or video gaming screen time or screen time overall, with mean differences ranging from approximately 5 minutes to <30 minutes (Multimedia Appendix 1 [8-13,30-58]).

XSL•FO
RenderX

Table 3. Controlled studies by outcome type and target[a].

Study and outcome	Study type	Participants, N	Time point of assessment	Intervention group value	Control group value	Statistical test results	*P* value
Children and adolescents combined							
Time using media							
Gaming time							
Hansen et al [8], 2022: minutes per weekday	CS[b]	210	After the intervention	143	166	F_{1832}=6.45	.01
Hansen et al [8], 2022: minutes per weekend	CS	210	After the intervention	209	240	F_{1832}=4.06	.04
Internet use time							
Hansen et al [8], 2022: minutes per weekday	CS	210	After the intervention	102	121	F_{1832}=5.83	.02
Hansen et al [8], 2022: minutes per weekend	CS	210	After the intervention	152	180	F_{1832}=6.73	.001
Social media use time							
Hansen et al [8], 2022: minutes per weekday	CS	210	After the intervention	74	87	F_{1832}=2.97	.09
Hansen et al [8], 2022: minutes per weekend	CS	210	After the intervention	99	121	F_{1832}=4.08	.04
Problematic use							
Problematic gaming							
Ortega-Barón et al [40], 2021: IGDS9-SF[c]	CS	165	After the intervention	3.88	5.56	F=5.39	<.05
Problematic internet use							
Hansen et al [8], 2022: CIUS[d]	CS	210	After the intervention	__[e]	—	F_{1832}=14.76	.001
Ortega-Barón et al [40], 2021: GPIUS2[f]	CS	165	After the intervention	16.71	25.83	F=8.29	<.01
Adolescents and youth							
Time using media							
Gaming time							
Bickham et al [34], 2018: minutes after school	CS	412	After the intervention	22	42.5	F_{1460}=4.8	.03
Bickham et al [34], 2018: minutes on Saturday	CS	412	After the intervention	33.5	42.9	F_{1463}=0.8	.37
Bonnaire et al [35], 2019: minutes per weekday	CS	384	After the intervention	119.8	188.9	F_2=18.57	<.001
Bonnaire et al [35], 2019: minutes per weekend	CS	384	After the intervention	177.8	213.1	F_2=31.06	<.001
Internet use time							
Bickham et al [34], 2018: minutes after school	CS	412	After the intervention	36.4	65.8	F_{1460}=2.9	.09
Bickham et al [34], 2018: minutes on Saturday	CS	412	After the intervention	42.3	55	F_{1459}=4.7	.03
Bonnaire et al [35], 2019: minutes per weekday	CS	384	After the intervention	300.2	325.5	F_2=17.68	<.001
Bonnaire et al [35], 2019: minutes per weekend	CS	384	After the intervention	317.9	353.1	F_2=6.90	.001
Screen time							

XSL•FO
RenderX

Study and outcome	Study type	Participants, N	Time point of assessment	Intervention group value	Control group value	Statistical test results	*P* value
Bickham et al [34], 2018: minutes after school	CS	412	After the intervention	49	82.2	F_{1459}=28.1	<.001
Bickham et al [34], 2018: minutes on Saturday	CS	412	After the intervention	63.3	78.3	F_{1457}=6.1	.01
Social media use time							
Thai et al [10], 2023	RCT[g]	260	After the intervention	78.25	188.76	F_{3648}=94.05	<.001
Throuvala et al [11], 2020	RCT	143	After the intervention	130.2	148.2	$F_{1, 140}$=3.697	.06
Smartphone time							
Brailovskaia et al [12], 2022	RCT	619	1 month	157.43	198.2	$F_{4, 1187}$=6.754	<.001
Brailovskaia et al [12], 2022	RCT	619	4 months	162.02	187.33	$F_{4, 1187}$=6.754	<.001
Throuvala et al [11], 2020	RCT	143	After the intervention	210.6	246.6	$F_{1, 140}$=4.43	<.001
Problematic use							
Problematic gaming							
Bonnaire et al [35], 2019: GAS-7[h] (proportion of gamers screening positive for IGD[i])	CS	384	After the intervention	10	23	Not given	.03
Problematic internet use							
Affouneh et al [43], 2021: IAT[j]	RCT	30	After the intervention	2.83	3.88	F_1=304.44	<.001
Lindenberg et al [30], 2022: modified CSAS[k]	RCT	422	1 month	*14.46* [l]	*12.48* [l]	γ_{11}=–0.128	.03
Lindenberg et al [30], 2022: modified CSAS	RCT	422	4 months	12.09	12.74	γ_{11}=–0.128	.03
Lindenberg et al [30], 2022: modified CSAS	RCT	422	12 months	9.2	10.07	γ_{11}=–0.128	.03
Problematic social media use							
Throuvala et al [11], 2020: Bergen SMAS[m]	RCT	143	After the intervention	15.12	17.24	$F_{1, 140}$=6.96	<.001
Weaver [41], 2021: SMUQ[n]	CS	54	After the intervention	15.6	22	$F_{2, 52}$=6.02	<.05
Problematic smartphone use							
Brailovskaia et al [12], 2022: modified Bergen SMAS	RCT	619	After the intervention	*12.8* [o]	*12.14* [o]	$F_{6, 1806}$=8.40	<.001
Brailovskaia et al [12], 2022: modified Bergen SMAS	RCT	619	1 month	11.22	12.28	$F_{6, 1806}$=8.40	<.001
Brailovskaia et al [12], 2022: modified Bergen SMAS	RCT	619	4 months	11.4	12.16	$F_{6, 1806}$=8.40	<.001
Other							
Time using media							
Gaming time							

XSL•FO
RenderX

Study and outcome	Study type	Participants, N	Time point of assessment	Intervention group value	Control group value	Statistical test results	*P* value
Apisitwasana et al [33][p], 2018: days per week	RCT	310	After the intervention	2.83	3.44	—[q]	<.001
Apisitwasana et al [33], 2018: days per week	RCT	310	3 months	3.57	4.03	—	<.05
Apisitwasana et al [33], 2018: minutes per weekday	RCT	310	After the intervention	54.6	64.8	—	<.05
Apisitwasana et al [33], 2018: minutes per weekday	RCT	310	3 months	72.6	84.6	—	NS[r]
Apisitwasana et al [33], 2018: minutes per weekend day	RCT	310	After the intervention	103.8	118.2	—	NS
Apisitwasana et al [33], 2018: minutes per weekend day	RCT	310	3 months	129	146.4	—	NS
Li et al [37][s], 2019 (parent report)	RCT	362	After the intervention	*45.2* [o]	*32.6* [o]	$F_{1.72, 516.22}$=4.86	.01
Li et al [37], 2019 (parent report)	RCT	362	3 months	*38.8* [o]	*36.3* [o]	$F_{1.72, 516.22}$=4.86	.01
Screen time							
Sanders et al [42], 2018: parental assessment of daily time	RCT	32	After the intervention	*180.6* [o]	*160.8* [o]	Not conducted	N/A[t]
Sanders et al [42], 2018: parent diary assessment of daily time	RCT	32	After the intervention	133.8	181.2	Not conducted	N/A
Social media use time							
Gui et al [13][u], 2023: interactive use index	RCT	2997	After the intervention	28.8	30.5	*t*=–1.7	>.05
Problematic use							
Problematic gaming							
Apisitwasana et al [33], 2018: GAST[v]	RCT	310	After the intervention	9.34	17.1	$F_{1, 823}$=7.79	.001
Apisitwasana et al [33], 2018: GAST	RCT	310	3 months	11.48	20.57	$F_{1, 823}$=7.79	.001
Li et al [37], 2019: parent version of adapted KSAS[w]	RCT	362	After the intervention	*25.48* [o]	*24.44* [o]	$F_{1.93, 520.5}$=1.27	.28
Li et al [37], 2019: parent version of adapted KSAS	RCT	362	3 months	*25.05* [o]	*24.61* [o]	$F_{1.93, 520.5}$=1.27	.28
Problematic smartphone use							
Gui et al [13], 2023: disturbance	RCT	2997	After the intervention	34.2	36.2	*t*=–2.0	<.01

XSL•FO
RenderX

Study and outcome	Study type	Participants, N	Time point of assessment	Intervention group value	Control group value	Statistical test results	*P* value
Gui et al [13], 2023: withdrawal	RCT	2997	After the intervention	27.3	28.5	*t*=–1.3	>.05

[a]The table describes only controlled trials, excluding subgroup analyses, that had estimates after the intervention or at a follow-up time point. Intervention and control values represent predicted or observed means, totals, or proportions as reported by the authors of the primary studies. Values for time are shown in minutes per day, either given or converted from other measures, unless otherwise specified.

[b]CS: other controlled study.

[c]IGDS9-SF: Internet Gaming Disorder Scale—Short-Form.

[d]CIUS: Compulsive Internet Use Scale.

[e]Not reported.

[f]GPIUS2: Generalized Problematic Internet Use Scale.

[g]RCT: randomized controlled trial.

[h]GAS-7: Game Addiction Scale-7 (item).

[i]IGD: internet gaming disorder.

[j]IAT: Internet Addiction Test.

[k]CSAS: Video Game Addiction Scale.

[l]Studies favoring the control group in the absence of baseline differences.

[m]SMAS: Social Media Addiction Scale.

[n]SMUQ: Social Media Use Questionnaire.

[o]Studies favoring the control group in the presence of baseline differences.

[p]This intervention was delivered to both children and parents.

[q]Not reported.

[r]NS: the authors of the study reported significant *P* values with asterisks in the manuscript, and these comparisons were not reported with asterisks.

[s]This intervention was delivered to parents.

[t]N/A: not applicable.

[u]This intervention was delivered to teachers.

[v]GAST: Game Addiction Screening Test.

[w]KSAS: Korean Internet Addiction Scale for Adolescents, adapted to fit the video gaming context.

RQ 4: What Other Characteristics of Intervention Setting or Delivery May Limit or Promote Translation to Other Contexts?

To synthesize evidence in a way that might inform translation to other contexts, we used the CASP tools to capture data about study quality and intervention implementation. In this review, some studies (7/35, 20%) included information about factors important to implementation, such as attrition and dropout rates [50], participant adherence [12,43,51], and fidelity [13,42,53]. One intervention [48] reported 100% participant engagement with no attrition. Others (3/35, 9%) evaluated the perceived competence of the person delivering the intervention [42] or success in implementation in the school setting [13,53]. However, most studies (22/35, 63%) provided little or no information on implementation or fidelity.

Our review noted several other characteristics of the studies that could affect the ability to translate the findings to other contexts. First, we found considerable diversity in availability of information about costs and funding sources. One protocol [52] calculated the cost of their intervention in a very detailed manner; however, such data were absent for all other studies. Studies disclosed varying sources of financial support, including support from national governments [9,13,31], public health organizations [49,53], charitable foundations [48,49], and other research organizations [37,45,51].

The availability of intervention manuals is also an important factor for translation. Few studies (3/35, 9%) used an intervention manual, and descriptions of the interventions were sometimes lacking. Some studies (5/35, 14%) incorporated a table with the elements and purpose for each module in either the body of the paper or an appendix. In contrast, some studies (2/35, 6%) made explicit references to the use of intervention manuals, such as a “gaming addiction prevention manual” for parents [33] or “manual-based CBT intervention” [47].

Factors contributing to scalability and resource use were often more clearly described and varied widely. For example, one intervention [51] used weekly 1-hour one-to-one parent visits by a therapist to promote parental restrictions on screen time and improved social interactions between parents and preschoolers with autism. Other interventions (17/35, 49%) were delivered by (or included) psychologists, school counselors, or school prevention officers [13,30,35,43,47,50] or teachers [13,33,34,40,54], usually over many weeks. Some interventions were delivered by peers with or without a professional colead, such as the Net Pilots intervention by Hansen et al [8,53], which was delivered by adolescents, or the joint peer coach (“a young adult volunteer with a personal history of problematic gaming”) and therapist-delivered intervention by Männikkö et al [50]. In contrast, some interventions (7/35, 20%) were online, requiring no direct interaction between intervention providers and participants,

XSL•FO
RenderX

including planned abstinence interventions [12,49] or interventions using apps [9,11].

The reviews provided some information about study and intervention elements that could promote translation, such as intervention intensity, components, and approaches. For example, Jones et al [73] found larger effect sizes for interventions with smaller sample sizes and shorter durations. However, no review discussed funding sources for interventions, and only 17% (1/6) of the reviews systematically examined harms, making it difficult to evaluate the costs, risks, and benefits of implementation in different contexts.

RQ 5: What Quality Issues Should Be Addressed in Future Empirical Studies of These Interventions?

Using the analytical questions of the CASP framework for both controlled trials and observational studies (see the OSF website), we found several design- and reporting-related issues. Most studies (32/35, 91%) reported information about PICOS with a few exceptions, notably, 6% (2/35) of the studies, which did not provide details on the population studied and the comparator [36,47], and 3% (1/35) of the studies, which lacked a clear RQ and intervention description [31]. Second, many of the included experimental and quasi-experimental studies did not report adequate detail about randomization (12/27, 44%) and blinding (7/27, 26%). Third, not all studies (15/27, 56%) addressed loss to follow-up clearly, and power calculations were rare (4/27, 15%).

Of the 6 reviews, only the 2 (33%) meta-analyses clearly specified which outcomes would be extracted from the studies. Other reviews did not specify how they would select outcomes, using language such as “critical outcomes” [71] and “key outcome measures” [74], resulting in a designation of low reliability according to our prespecified criteria. In addition, all reviews received an AMSTAR 2 score of “critically low” (see the OSF website), with no reviews reporting on funding sources of individual studies and few (4/35, 11%) reporting preregistered protocols or duplicate data extraction.

Discussion

Principal Findings

We conducted a rapid systematic review of the recent literature (since 2017) to investigate the current evidence on prevention approaches to promote digital well-being, defined as the healthy use of digital media and technology and the absence of problems resulting from excessive use, among children, adolescents, and youth. The final selection included 41 studies (n=35, 85% primary studies and n=6, 15% reviews). In the primary studies, internet and game use were the most common outcomes, and most interventions (22/33, 67%) included education about self-regulation. Digital well-being education, media literacy and effects, or mental health symptoms and coping (psychoeducation) were other common intervention approaches. Most interventions (26/33, 79%) also combined several approaches. Vote counting of controlled trials showed benefits for all interventions. A total of 33% (2/6) of the reviews conducted statistical analyses and found that the interventions studied had a small but beneficial impact on decreasing the amount of time that people spent on screens.

Similar to previous reviews in this area [5,17], this review also found that most programs for universal and selective intervention to promote digital well-being and prevent problematic technology use were implemented in educational settings. Many of these interventions can be classified as complex interventions that contain multiple active ingredients and tap into different potential behavior change strategies. Intervention approaches that aimed at improving the self-regulation of participants were the most common in the included studies, and this self-regulation closely resembled goals, feedback, and planning in a widely used taxonomy of behavior change strategies [76]. However, a close investigation of the individual studies showed that self-regulation, even if mentioned using that term, could look very different. For example, the 6-week intervention by Apisitwasana et al [33] contained 5 weeks of instruction specific to self-regulation, whereas the study by Affouneh et al [43] delivered 1 week of “self-discipline” instruction for avoidance of addictions specifically, with other weeks targeting “problem solving strategies for dealing with problems to prevent addictive behaviors” and “effective use of time to prevent addictive behaviors.” While grouping of intervention approaches is important for systematic investigation, these limitations in comparability of approaches across studies should be kept in mind, and detailed reporting of intervention approaches in future studies is needed.

Moreover, the interventions included in this review varied widely in intensity, with some interventions requiring a single session and others requiring 1:1 weekly clinician visits over weeks. Peer-delivered interventions may reduce overall costs. These have been successful for other addictions and are a common method of task sharing for psychosocial interventions, especially in middle-income settings [77]. However, this depends on the intensity of training; if extensive training of teachers is needed for them to be able to train peers and teenage peers then age out, these extensive trainings may need to be repeated with high frequency. Although our review was not designed to provide a quantitative evaluation of the studies, the included review by Jones et al [73] found that interventions with shorter durations were more likely to be effective in reducing screen time. Interventions with shorter durations are more likely to be feasibly implemented, and future studies should aim to examine associations among intervention intensity, feasibility, and efficacy.

Harms are not usually assessed in the reporting of studies of interventions for behavioral addictions [74], but it is important for digital well-being. Interventions involving abstinence from addictive behaviors may result in rebound effects after the intervention or may lead to engagement in compensatory behaviors that could be more harmful (eg, searching for gaming-related pornography when restricted from gaming [78]). We found that some comparisons included in our vote counting (7/51, 14%) had results that favored the control group; however, these were usually due to baseline differences (6/7, 86%). In contrast, individuals struggling with more harmful addictive behaviors such as alcohol or drug use may use digital

XSL•FO
RenderX

behaviors—even excessive ones—as a less harmful compensatory behavior to deal with cravings and withdrawal symptoms [79]. Findings on both the potential benefits and negative consequences of digital well-being interventions underscore the need for RCTs with appropriate treatment of any baseline differences as well as comprehensive and consistent reporting of outcomes in future research.

As suggested by other scholars [17], our review included an explicit aim to report aspects of published work that impact how the findings can be translated and applied to different contexts (eg, countries and areas of the world that are starting to see technology use problems but have not developed or systematically evaluated digital well-being interventions themselves). Many studies included in this review (20/35, 57%) were conducted in Europe and the Global North. Whether these studies can translate to other contexts is another question. Our review focused on identifying factors that can help interventionists decide whether an intervention will work in their setting. Things to consider are information on intervention costs, intervention manuals, and information on scalability. Only a minority of studies (7/35, 20%) reported some of these aspects. Future studies would benefit from including greater detail on factors that would help inform translation to other contexts.

Moreover, formulating a well-defined RQ; using a clear study design; and providing comprehensive information about the studied population, intervention, comparator, and measured outcomes are crucial for understanding study quality. Reporting guidelines for empirical studies and systematic reviews provide clear recommendations for elements of study design, methods, and outcomes to report and should be used in future work. We hope that, by explicitly focusing on these topics, our review can inspire future studies to report this crucial information and, thus, make a contribution to developing the international evidence base in this important and growing area of research.

Limitations

This review has some limitations. As this was designed to be a rapid review, we deferred some elements that are common in other types of reviews (eg, double data extraction, contacting authors for missing data, performing meta-analyses or otherwise evaluating effect sizes, and conducting a formal risk-of-bias assessment [64]). Moreover, we included the most recent literature published since 2017, which is appropriate for a rapid review and quickly changing technology platforms and use patterns. We excluded studies that described “treatment” of “individuals with disorder” even if the screening instrument scores of the included samples fell below the cutoffs used for the disorders. Thus, we may have missed other interventions that could be helpful to improve digital well-being. Although we searched 5 databases, we missed interventions such as the Digital Balance survey and website and Ithra Sync, which were delivered through websites and social media and did not show up in our searches [16,80]. Outcome-level limitations include that we did not limit our review to studies with specific outcome measures or scales or even to studies with validated outcome measures. This allowed us to include a wide range of studies but runs the risk of including weak outcome measures (eg, inconsistent measures for the same outcome across the studies and problems with self-reporting screen time). At the study level, we found that poor reporting of key information, including fidelity of intervention implementation or missing data, presented significant limitations to our ability to synthesize data into evidence. In some cases, study design and statistical analysis were difficult to ascertain [44] or missing [31]. Only a subset of RCTs (14/19, 74%; Table 3) were included in vote counting because they reported direct comparison between the intervention and control groups on the outcomes selected for this review. In addition, the dates of our review, 2017 to 2023, included the 3 years of the COVID-19 pandemic (ie, March 2020 to May 2023), making it possible that empirical studies occurring during this time could contain cohort effects. Finally, most reviews (4/6, 66%) lacked protocols or outcome specification, leading to low quality scores.

The strengths of our review include that all methods were guided by standards, including the Cochrane rapid review recommendations and AMSTAR 2 checklist. To enhance the ability of interventionists to use these findings, we focused on critical appraisal of quality and translatability through CASP checklists. We preregistered our protocol, and all accompanying data are supplied on an OSF website, which promotes reproducibility.

Conclusions

As digital technology use is increasing, interventions to prevent problematic use and promote digital well-being will be of growing importance. This review demonstrates positive findings for effectiveness of prevention interventions for behavioral outcomes and describes factors that may contribute to translation and implementation. Digital well-being is a global need, and understanding contextual factors that influence healthy use as well as limitations of the current evidence is vital for informing future research. Future research would benefit from following appropriate reporting guidelines, reporting both the benefits and harms of interventions, and including greater detail on factors informing translation.

Acknowledgments

This research was supported by a grant from Aramco Services Company, a subsidiary of Saudi Aramco. The funders had no role in conducting the review or reporting the results. The authors wish to acknowledge input from the King Abdulaziz Center for World Culture (Ithra) in developing research questions for this project and Janardan Devkota for his contribution to data visualizations.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are available in the Open Science Framework repository [59].

Authors' Contributions

Authors MCC, JS, and JT contributed to the study concept and design, authors MCC, DA, MD, JS, FK, KGO, FMA, and JT contributed to the analysis and interpretation of the data, authors MCC and JT obtained funding, and authors MCC, DA, JS, and JT contributed to study supervision. All authors had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Conflicts of Interest

MCC conducts consulting related to video games and well-being and has held leadership positions in nonprofit organizations related to the promotion of healthy video gaming. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 1900 KB - jmir_v26i1e59968_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e59968_app2.docx]

References

1. Social media and teens. American Academy of Child and Adolescent Psychiatry. URL: https://tinyurl.com/565yxe3e [accessed 2024-02-20]
2. Council on Communications and Media. Media use in school-aged children and adolescents. Pediatrics 2016 Nov;138(5):e20162592. [doi: 10.1542/peds.2016-2592] [Medline: 27940794]
3. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). San Francisco, CA: Booksmith Publishing LLC; 2013.
4. ICD-11 for mortality and morbidity statistics. World Health Organization. URL: https://tinyurl.com/594s5amn [accessed 2020-10-11]
5. Throuvala MA, Griffiths MD, Rennoldson M, Kuss DJ. School-based prevention for adolescent internet addiction: prevention is the key. A systematic literature review. Curr Neuropharmacol 2019;17(6):507-525 [FREE Full text] [doi: 10.2174/1570159X16666180813153806] [Medline: 30101714]
6. Young KS. The evolution of internet addiction. Addict Behav 2017 Jan;64:229-230. [doi: 10.1016/j.addbeh.2015.05.016] [Medline: 26059165]
7. van den Eijnden RJ, Meerkerk GJ, Vermulst AA, Spijkerman R, Engels RC. Online communication, compulsive internet use, and psychosocial well-being among adolescents: a longitudinal study. Dev Psychol 2008 May;44(3):655-665. [doi: 10.1037/0012-1649.44.3.655] [Medline: 18473634]
8. Hansen J, Hanewinkel R, Goecke M, Morgenstern M. Prevention of excessive media use in childhood and adolescence. Monatsschr Kinderheilkd 2021 Jun 21;170:435-442. [doi: 10.1007/s00112-021-01220-x]
9. Ko JL, Yuan C, Malherbe B, Yang CH, Kuo PY. Exploring the effect of activity intervention on reducing social media use: lessons learned in a field study. In: Proceedings of the 23rd HCI International Conference on Late Breaking Posters. 2021 Presented at: HCII 2021; July 24-29, 2021; Virtual Event. [doi: 10.1007/978-3-030-90179-0_47]
10. Thai H, Davis CG, Mahboob W, Perry S, Adams A, Goldfield GS. Reducing social media use improves appearance and weight esteem in youth with emotional distress. Psychol Pop Media 2024;13(1):162-169. [doi: 10.1037/ppm0000460]
11. Throuvala MA, Griffiths MD, Rennoldson M, Kuss DJ. Mind over matter: testing the efficacy of an online randomized controlled trial to reduce distraction from smartphone use. Int J Environ Res Public Health 2020 Jul 05;17(13):4842 [FREE Full text] [doi: 10.3390/ijerph17134842] [Medline: 32635650]
12. Brailovskaia J, Delveaux J, John J, Wicker V, Noveski A, Kim S, et al. Finding the "sweet spot" of smartphone use: reduction or abstinence to increase well-being and healthy lifestyle?! An experimental intervention study. J Exp Psychol Appl 2023 Mar;29(1):149-161. [doi: 10.1037/xap0000430] [Medline: 35389685]
13. Gui M, Gerosa T, Argentin G, Losi L. Mobile media education as a tool to reduce problematic smartphone use: results of a randomised impact evaluation. Comput Educ 2023 Mar;194:104705. [doi: 10.1016/j.compedu.2022.104705]
14. van Rooij AJ, Ferguson CJ, Colder Carras M, Kardefelt-Winther D, Shi J, Aarseth E, et al. A weak scientific basis for gaming disorder: let us err on the side of caution. J Behav Addict 2018 Mar 01;7(1):1-9 [FREE Full text] [doi: 10.1556/2006.7.2018.19] [Medline: 29529886]

XSL•FO
RenderX

15. Vanden Abeele MM. Digital wellbeing as a dynamic construct. Commun Theory 2021 Nov;31(4):932-955. [doi: 10.1093/ct/qtaa024]
16. Digitale balans homepage. Digitale Balans. URL: https://digitalebalans.nl/digitale-balans/ [accessed 2021-03-27]
17. King DL, Delfabbro PH, Doh YY, Wu AM, Kuss DJ, Pallesen S, et al. Policy and prevention approaches for disordered and hazardous gaming and internet use: an international perspective. Prev Sci 2018 Feb;19(2):233-249. [doi: 10.1007/s11121-017-0813-1] [Medline: 28677089]
18. Colder Carras M, Shi J, Hard G, Saldanha IJ. Evaluating the quality of evidence for gaming disorder: a summary of systematic reviews of associations between gaming disorder and depression or anxiety. PLoS One 2020 Oct 26;15(10):e0240032 [FREE Full text] [doi: 10.1371/journal.pone.0240032] [Medline: 33104730]
19. Consumer protection in online video games: a European single market approach. European Parliament. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0008_EN.pdf [accessed 2023-01-18]
20. Saquib N, Saquib J, AlSalhi A, Colder Carras M, Labrique AB, Al-Khani AM, et al. The associations between family characteristics and problematic internet use among adolescents in Saudi Arabia. Int J Adolesc Youth 2023 Sep 18;28(1). [doi: 10.1080/02673843.2023.2256826]
21. Xiao LY. Reserve your judgment on "Draconian" Chinese video gaming restrictions on children •. J Behav Addict 2022 May 27;11(2):249-255 [FREE Full text] [doi: 10.1556/2006.2022.00022] [Medline: 35895617]
22. Porter CM. Revisiting precede–proceed: a leading model for ecological and ethical health promotion. Health Educ J 2016;75(6):753-764. [doi: 10.1177/0017896915619645]
23. Kellam SG, Langevin DJ. A framework for understanding "evidence" in prevention research and programs. Prev Sci 2003 Sep;4(3):137-153. [doi: 10.1023/a:1024693321963] [Medline: 12940466]
24. Lassi ZS, Salam RA, Das JK, Bhutta ZA. The conceptual framework and assessment methodology for the systematic reviews of community-based interventions for the prevention and control of infectious diseases of poverty. Infect Dis Poverty 2014 Jul 31;3:22 [FREE Full text] [doi: 10.1186/2049-9957-3-22] [Medline: 25105014]
25. Colder Carras M. Video game play, social interactions and friendship quality in adolescents: a latent class analysis. The Johns Hopkins University. 2015 Oct. URL: https://tinyurl.com/bdfck9a7 [accessed 2017-06-06]
26. Young KS, Brand M. Merging theoretical models and therapy approaches in the context of internet gaming disorder: a personal perspective. Front Psychol 2017 Oct 20;8:1853 [FREE Full text] [doi: 10.3389/fpsyg.2017.01853] [Medline: 29104555]
27. Elson M, Breuer J, Ivory JD, Quandt T. More than stories with buttons: narrative, mechanics, and context as determinants of player experience in digital games. J Commun 2014 Jun;64(3):521-542. [doi: 10.1111/jcom.12096]
28. Robinson MJ, Zumbusch AS, Anselme P. The incentive sensitization theory of addiction. In: Oxford Research Encyclopedia of Psychology. Oxford, UK: Oxford University Press; May 2023.
29. Ruggiero T. Uses and gratifications theory in the 21st century. In: Refining Milestone Mass Communications Theories for the 21st Century. Milton Park, UK: Routledge; 2016.
30. Lindenberg K, Kindt S, Szász-Janocha C. Effectiveness of cognitive behavioral therapy-based intervention in preventing gaming disorder and unspecified internet use disorder in adolescents: a cluster randomized clinical trial. JAMA Netw Open 2022 Feb 01;5(2):e2148995 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.48995] [Medline: 35179587]
31. Neverkovich SD, Bubnova IS, Kosarenko NN, Sakhieva RG, Sizova ZM, Zakharova VL, et al. Students' internet addiction: study and prevention. EURASIA J Math Sci Tech Ed 2018 Apr;14(4):1483-1495 [FREE Full text] [doi: 10.29333/ejmste/83723]
32. Choi J, Cho H, Lee S, Kim J, Park EC. Effect of the online game shutdown policy on internet use, internet addiction, and sleeping hours in Korean adolescents. J Adolesc Health 2018 May;62(5):548-555. [doi: 10.1016/j.jadohealth.2017.11.291] [Medline: 29434003]
33. Apisitwasana N, Perngparn U, Cottler LB. Effectiveness of school- and family-based interventions to prevent gaming addiction among grades 4–5 students in Bangkok, Thailand. Psychol Res Behav Manag 2018 Apr;Volume 11:103-115. [doi: 10.2147/prbm.s145868]
34. Bickham DS, Hswen Y, Slaby RG, Rich M. A preliminary evaluation of a school-based media education and reduction intervention. J Prim Prev 2018 Jun;39(3):229-245. [doi: 10.1007/s10935-018-0510-2] [Medline: 29721652]
35. Bonnaire C, Serehen Z, Phan O. Effects of a prevention intervention concerning screens, and video games in middle-school students: influences on beliefs and use. J Behav Addict 2019 Sep 01;8(3):537-553 [FREE Full text] [doi: 10.1556/2006.8.2019.54] [Medline: 31537087]
36. Esmaeili Rad M, Ahmadi F. A new method to measure and decrease the online social networking addiction. Asia Pac Psychiatry 2018 Dec 03;10(4):e12330. [doi: 10.1111/appy.12330] [Medline: 30175904]
37. Li AY, Chau C, Cheng C. Development and validation of a parent-based program for preventing gaming disorder: the game over intervention. Int J Environ Res Public Health 2019 Jun 04;16(11):1984 [FREE Full text] [doi: 10.3390/ijerph16111984] [Medline: 31167457]
38. Marco C, Choliz M. Effectiveness of impulsivity control techniques to videogame addiction prevention. Terapia Psicologica 2017 Jan;35(1):57-69.

XSL•FO
RenderX

39. Mumcu HE, Yazici Ö, Yilmaz O. Effect of 12-week recreational activity program on digital game addiction and peer relationships qualities in children. Acta Medica Mediterranea 2021 Sep;37(5):2919 [FREE Full text] [doi: 10.19193/0393-6384_2021_5_451]
40. Ortega-Barón J, González-Cabrera J, Machimbarrena JM, Montiel I. Safety.Net: a pilot study on a multi-risk internet prevention program. Int J Environ Res Public Health 2021 Apr 16;18(8):4249 [FREE Full text] [doi: 10.3390/ijerph18084249] [Medline: 33923779]
41. Weaver JL, Swank JM. A mindfulness-based intervention for adolescent social media users: a quasi-experimental study. J Child Adolesc Couns 2023 Oct 18;10(1):3-14 [FREE Full text] [doi: 10.1080/23727810.2023.2266895]
42. Sanders W, Parent J, Forehand R. Parenting to reduce child screen time: a feasibility pilot study. J Dev Behav Pediatr 2018 Jan;39(1):46-54 [FREE Full text] [doi: 10.1097/DBP.0000000000000501] [Medline: 28937450]
43. Affouneh S, Mahamid FA, Berte DZ, Shaqour AZ, Shayeb M. The efficacy of a training program for social skills in reducing addictive internet behaviors among Palestinian university students. Psicol Reflex Crit 2021 Jun 26;34(1):19 [FREE Full text] [doi: 10.1186/s41155-021-00185-w] [Medline: 34175995]
44. Tang C, Raat H, Yan M, Zhang Q, Li K, Jiang M, et al. Application of the health action process approach model for reducing excessive internet use behaviors among rural adolescents in China: a school-based intervention pilot study. BMC Public Health 2021 May 26;21(1):986 [FREE Full text] [doi: 10.1186/s12889-021-10999-z] [Medline: 34039318]
45. Chau CL, Tsui YY, Cheng C. Gamification for internet gaming disorder prevention: evaluation of a Wise IT-Use (WIT) program for Hong Kong primary students. Front Psychol 2019 Nov 1;10:2468 [FREE Full text] [doi: 10.3389/fpsyg.2019.02468] [Medline: 31736842]
46. Ke GN, Wong SF. A healthy mind for problematic internet use. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):637-645. [doi: 10.1089/cyber.2018.0072] [Medline: 30256674]
47. Ke GN, Wong SF. Outcome of the psychological intervention program: internet use for youth. J Rational Emot Cognit Behav Ther 2017 Nov 10;36(2):187-200. [doi: 10.1007/s10942-017-0281-3]
48. Kent S, Masterson C, Ali R, Parsons CE, Bewick BM. Digital intervention for problematic smartphone use. Int J Environ Res Public Health 2021 Dec 14;18(24):13165 [FREE Full text] [doi: 10.3390/ijerph182413165] [Medline: 34948774]
49. King DL, Kaptsis D, Delfabbro PH, Gradisar M. Effectiveness of brief abstinence for modifying problematic internet gaming cognitions and behaviors. J Clin Psychol 2017 Dec 02;73(12):1573-1585. [doi: 10.1002/jclp.22460] [Medline: 28152189]
50. Männikkö N, Mustonen T, Tanner N, Vähänikkilä H, Kääriäinen M. Effectiveness of a brief group intervention program for young adults with gaming-related problems. Int J Ment Health Addict 2021 Jun 14;20(5):2956-2972. [doi: 10.1007/s11469-021-00559-2]
51. Heffler KF, Frome LR, Garvin B, Bungert LM, Bennett DS. Screen time reduction and focus on social engagement in autism spectrum disorder: a pilot study. Pediatr Int 2022 Jan 08;64(1):e15343. [doi: 10.1111/ped.15343] [Medline: 36348519]
52. Barker AE. Promoting healthy internet gaming among preadolescent youth: a DBT-informed program. Capella University. 2020. URL: https://tinyurl.com/yc3fufdw [accessed 2024-12-09]
53. Hansen J, Janssen J, Hanewinkel R, Goecke M, Morgenstern M. Results of the process evaluation of the school-based prevention program “Net-Piloten” to avoid excessive use of computer games and the internet in adolescence. Präv Gesundheitsf 2020 Aug 21;16:256-262. [doi: 10.1007/s11553-020-00808-7]
54. Tang AC, Lee RL. Effects of a group mindfulness-based cognitive programme on smartphone addictive symptoms and resilience among adolescents: study protocol of a cluster-randomized controlled trial. BMC Nurs 2021 Jun 05;20(1):86 [FREE Full text] [doi: 10.1186/s12912-021-00611-5] [Medline: 34090443]
55. Krossbakken E, Torsheim T, Mentzoni RA, King DL, Bjorvatn B, Lorvik IM, et al. The effectiveness of a parental guide for prevention of problematic video gaming in children: a public health randomized controlled intervention study. J Behav Addict 2018 Mar 01;7(1):52-61 [FREE Full text] [doi: 10.1556/2006.6.2017.087] [Medline: 29313731]
56. Sela Y. Examining the efficacy of "parental vigilant care" for reducing problematic internet usage. National Institutes of Health National Library of Medicine. 2021. URL: https://clinicaltrials.gov/ct2/show/NCT05117216 [accessed 2023-03-02]
57. Can you reduce problematic smartphone use through a group intervention: a single case experimental design. National Institute for Health National Library of Medicine. 2022. URL: https://clinicaltrials.gov/ct2/show/NCT05384548 [accessed 2023-03-02]
58. Schmuck D. Does digital detox work? Exploring the role of digital detox applications for problematic smartphone use and well-being of young adults using multigroup analysis. Cyberpsychol Behav Soc Netw 2020 Aug;23(8):526-532. [doi: 10.1089/cyber.2019.0578] [Medline: 32354288]
59. Rapid review of interventions for adolescents' use of digital technologies. OSF Home. 2023 Mar 8. URL: https://osf.io/5fe3z/?view_only=0127c110808744ad9753b52020eddcbc [accessed 2024-12-09]
60. Garritty C, Gartlehner G, Nussbaumer-Streit B, King VJ, Hamel C, Kamel C, et al. Cochrane rapid reviews methods group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol 2021 Feb;130:13-22 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.10.007] [Medline: 33068715]
61. EU Kids Online: findings, methods, recommendations (deliverable D1.6). EU Kids Online, The London School of Economics and Political Science. 2014. URL: https://eprints.lse.ac.uk/60512/ [accessed 2023-06-29]

XSL•FO
RenderX

62. Zotero homepage. Zotero. URL: https://www.zotero.org/about/ [accessed 2024-03-16]
63. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2024-03-16]
64. Garritty C, Hamel C, Trivella M, Gartlehner G, Nussbaumer-Streit B, Devane D, et al. Updated recommendations for the Cochrane rapid review methods guidance for rapid reviews of effectiveness. BMJ 2024 Feb 06;384:e076335. [doi: 10.1136/bmj-2023-076335] [Medline: 38320771]
65. Coly A, Parry G. Evaluating complex health interventions: a guide to rigorous research designs. AcademyHealth. 2017 Jun. URL: https://academyhealth.org/sites/default/files/AH_Evaluation_Guide_FINAL.pdf [accessed 2023-06-26]
66. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
67. Saldanha IJ, Lindsley KB, Lum F, Dickersin K, Li T. Reliability of the evidence addressing treatment of corneal diseases: a summary of systematic reviews. JAMA Ophthalmol 2019 Jul 01;137(7):775-785 [FREE Full text] [doi: 10.1001/jamaophthalmol.2019.1063] [Medline: 31070698]
68. Critical appraisal checklists. Critical Appraisal Skills Programme. URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2024-12-09]
69. CASP checklist: CASP cohort study checklist. Critical Appraisal Skills Programme. URL: https://casp-uk.net/casp-tools-checklists/cohort-study-checklist/ [accessed 2024-12-09]
70. McKenzie JE, Brennan SE. Chapter 12: Synthesizing and presenting findings using other methods. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions, Second Edition. London, UK: The Cochrane Collaboration; 2019.
71. Throuvala MA, Griffiths MD, Rennoldson M, Kuss DJ. The role of recreational online activities in school-based screen time sedentary behaviour interventions for adolescents: a systematic and critical literature review. Int J Ment Health Addict 2020 Feb 28;19(4):1065-1115. [doi: 10.1007/s11469-019-00213-y]
72. Oh C, Carducci B, Vaivada T, Bhutta ZA. Interventions to promote physical activity and healthy digital media use in children and adolescents: a systematic review. Pediatrics 2022 May 01;149(Suppl 5):e2021053852I [FREE Full text] [doi: 10.1542/peds.2021-053852I] [Medline: 35503334]
73. Jones A, Armstrong B, Weaver RG, Parker H, von Klinggraeff L, Beets MW. Identifying effective intervention strategies to reduce children's screen time: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Sep 16;18(1):126 [FREE Full text] [doi: 10.1186/s12966-021-01189-6] [Medline: 34530867]
74. Fernandez DP, Kuss DJ, Griffiths MD. Short-term abstinence effects across potential behavioral addictions: a systematic review. Clin Psychol Rev 2020 Mar;76:101828. [doi: 10.1016/j.cpr.2020.101828] [Medline: 32062303]
75. Krafft H, Boehm K, Schwarz S, Eichinger M, Büssing A, Martin D. Media awareness and screen time reduction in children, youth or families: a systematic literature review. Child Psychiatry Hum Dev 2023 Jun;54(3):815-825 [FREE Full text] [doi: 10.1007/s10578-021-01281-9] [Medline: 34855040]
76. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
77. Singla DR, Kohrt BA, Murray LK, Anand A, Chorpita BF, Patel V. Psychological treatments for the world: lessons from low- and middle-income countries. Annu Rev Clin Psychol 2017 May 08;13:149-181 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-045217] [Medline: 28482687]
78. Castro-Calvo J, Ballester-Arnal R, Potenza MN, King DL, Billieux J. Does "forced abstinence" from gaming lead to pornography use? Insight from the April 2018 crash of Fortnite's servers. J Behav Addict 2018 Sep 01;7(3):501-502 [FREE Full text] [doi: 10.1556/2006.7.2018.78] [Medline: 30203695]
79. Colder Carras M, Kalbarczyk A, Wells K, Banks J, Kowert R, Gillespie C, et al. Connection, meaning, and distraction: a qualitative study of video game play and mental health recovery in veterans treated for mental and/or behavioral health problems. Soc Sci Med 2018 Nov;216:124-132 [FREE Full text] [doi: 10.1016/j.socscimed.2018.08.044] [Medline: 30257787]
80. Sync - digital wellbeing programme. Sync Ithra. URL: https://sync.ithra.com/ [accessed 2023-11-01]

Abbreviations

AMSTAR: A Measurement Tool to Assess Systematic Reviews
CASP: Critical Appraisal Skills Programme
OSF: Open Science Framework
PICOS: population, intervention, comparator, outcome, and study design
RCT: randomized controlled trial
RQ: research question

XSL•FO
RenderX

Edited by G Eysenbach; submitted 30.04.24; peer-reviewed by E Selkie, HK Kang; comments to author 04.06.24; revised version received 15.08.24; accepted 31.08.24; published 18.12.24.

Please cite as:
Colder Carras M, Aljuboori D, Shi J, Date M, Karkoub F, García Ortiz K, Abreha FM, Thrul J
Prevention and Health Promotion Interventions for Young People in the Context of Digital Well-Being: Rapid Systematic Review
J Med Internet Res 2024;26:e59968
URL: https://www.jmir.org/2024/1/e59968
doi:10.2196/59968
PMID:39693138

©Michelle Colder Carras, Dahlia Aljuboori, Jing Shi, Mayank Date, Fatima Karkoub, Karla García Ortiz, Fasika Molla Abreha, Johannes Thrul. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

When and Why Adults Abandon Lifestyle Behavior and Mental Health Mobile Apps: Scoping Review

Patrick G Kidman[1], BClinExPhys (Hons); Rachel G Curtis[1], PhD; Amanda Watson[1,2], PhD; Carol A Maher[1], PhD

[1]Alliance for Research in Exercise, Nutrition and Activity, Allied Health and Human Performance, University of South Australia, Adelaide, Australia
[2]Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, Australia

Corresponding Author:
Carol A Maher, PhD
Alliance for Research in Exercise, Nutrition and Activity
Allied Health and Human Performance
University of South Australia
GPO Box 2471
Adelaide, 5001
Australia
Phone: 61 883026558
Email: carol.maher@unisa.edu.au

Abstract

Background: With 1 in 3 adults globally living with chronic conditions and the rise in smartphone ownership, mobile health apps have become a prominent tool for managing lifestyle-related health behaviors and mental health. However, high rates of app abandonment pose challenges to their effectiveness.

Objective: We explored the abandonment of apps used for managing physical activity, diet, alcohol, smoking, and mental health in free-living conditions, examining the duration of app use before abandonment and the underlying reasons.

Methods: A scoping review was conducted based on the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines and eligibility criteria were designed according to the SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research type) framework. In total, 4 databases were searched (MEDLINE, Scopus, Embase, and PsycINFO) to identify quantitative and qualitative studies with outcome measures related to app abandonment in adults with free-living conditions, including reasons for abandonment and duration of use, for mobile apps related to WHO (World Health Organization) modifiable health behaviors and mental health. The included studies' risk of bias was appraised based on the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) and COREQ (Consolidated Criteria for Reporting Qualitative Research) checklists. To enable data synthesis across different methodologies, app domains, demographic data, and outcome measures were categorized. Results are presented in 2 sections: quantitatively in a scatterplot to understand when users abandon apps and qualitatively through basic qualitative content analysis to identify the underlying reasons.

Results: Eighteen eligible studies (525,824 participants) published between 2014 and 2022, predominantly from the United States, Canada, the United Kingdom, and Germany, were identified. Findings revealed a curvilinear pattern of app abandonment, with sharper abandonment soon after acquisition, followed by a slowing rate of abandonment over time. Taken together, a median of 70% of users discontinued use within the first 100 days. The abandonment rate appeared to vary by app domain, with apps focusing on alcohol and smoking exhibiting faster abandonment, and physical activity and mental health exhibiting longer usage durations. In total, 22 unique reasons for abandonment were organized into six categories: (1) technical and functional issues, (2) privacy concerns, (3) poor user experience, (4) content and features, (5) time and financial costs, and (6) evolving user needs and goals.

Conclusions: This study highlights the complex nature of health app abandonment and the need for an improved understanding of user engagement over time, underscoring the importance of addressing various factors contributing to abandonment, from technical issues to evolving user needs. Our findings also emphasize the need for longitudinal studies and a consistent definition of app abandonment to better understand and mitigate this phenomenon, thereby enhancing the effectiveness of health apps in supporting public health initiatives.

(J Med Internet Res 2024;26:e56897) doi:10.2196/56897

XSL•FO
RenderX

KEYWORDS

mobile health apps; smartphone applications; app abandonment; app attrition; user engagement; health behavior; user retention; lifestyle management; quantitative analysis; qualitative analysis; mobile phone

Introduction

A total of 1 in 3 adults worldwide live with multiple chronic conditions [1], a situation exacerbated by modifiable health behaviors such as tobacco use, alcohol consumption, diet, sleep patterns, physical activity, and sedentary habits [2]. Furthermore, mental illnesses such as depression and anxiety are significant sources of global disability [3]. Mobile health apps offer a promising avenue for enhancing both physical [4] and mental health outcomes [5] by facilitating improvements in these health behaviors.

As of 2023, an estimated 6.7 billion people around the world own smartphones [6]. In 2022, the global mobile health market was valued at US $50 billion and is projected to grow to US $466 billion by 2032 [7]. The health and fitness app market is substantial, with a reported 350,000 health apps available on the Apple and Google Play Stores, and around 90,000 new health apps added in 2020 alone (an average of 250 per day) [8]. Despite their undoubted potential, industry data shows that 66% of health apps and 69% of fitness apps are abandoned within 90 days, which is worse than the average of 52% for all app categories [9]. This represents a tremendous waste of time and resources, particularly given that, on average, development takes 7-12 months and costs US $270,000 [10].

Numerous studies have examined health app adoption and usage [11-14], suggesting that individuals select apps based on specific health needs and preferences. A desire to improve health, through health behaviors like diet [12] or physical activity [15], is a common driver. App reviews and ratings, the credibility of the source, and recommendations from health care practitioners, friends, or family members have all been identified as important determinants of app adoption [13]. Specific features, such as the ability to track outcomes over time, data visualization, and gamification elements also play a role [11]. Cost and accessibility also impact the choice of health apps, with free apps generally favored, although subscription models can be successful if they are perceived as offering value [11]. Two 2022 systematic reviews, by Jakob et al [16], and Amagai et al [14] examined the factors influencing health app adherence and engagement. They identified a total of 99 and 62 previous studies, respectively, which, taken together, suggested that user-friendliness, tailoring, reminders, in-app support from health care professionals, gamification, and financial incentives are key factors in improving app adherence. While these studies provide insight into reasons for the adoption and usage of health apps, less is known about app abandonment.

In 2020, Meyerowitz-Katz et al [17] published a systematic review specifically examining attrition in app-based intervention studies for chronic diseases. They identified 17 studies, including apps designed for a cast array of chronic diseases and management approaches, including HIV management, management of menstrual pain, low back pain, and chronic kidney disease, including intervention approaches such as pain management, medication adherence, and more generalized health behaviors such as healthy eating and physical activity. They noted that most studies have focused on reporting attrition rates quantitatively, without attempting to determine the reasons behind the attrition. They also reported that attrition rates appeared to vary depending on whether it was studied under randomized controlled trial conditions (lower dropout) or under “real-world” conditions (higher dropout). This fits with a wider commentary about the limited real-world applicability of results obtained under RCT conditions being reported across a range of health and medical research contexts [18,19].

This study set out to address these gaps. The objective of this review is to explore the abandonment of apps used for managing lifestyle behaviors (physical activity, diet, alcohol, and smoking) and mental health (henceforth referred to as “health apps” for simplicity), collected under free-living conditions. Furthermore, we will bring together abandonment data with qualitative research exploring the reasons why people abandon health apps. We hope that this will provide enhanced understanding to guide the design of health apps that are less likely to be abandoned in the future.

Methods

Protocol

This scoping review was conducted with reference to and reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; refer to Multimedia Appendix 1 for checklist) [20].

Eligibility Criteria

The eligibility criteria (Table 1) were developed using the Sample, Phenomenon of Interest, Design, Evaluation, and Research type (SPIDER) framework [21].

XSL•FO
RenderX

Table 1. Eligibility criteria according to the Sample, Phenomenon of Interest, Design, Evaluation, and Research type framework.

SPIDER[a]	Inclusion	Exclusion
Sample	Adults (aged ≥18 years), smartphone users, general population with no restrictions on health status.	Pediatric samples, populations other than general smartphone users (eg, app developers).
Phenomenon of Interest	Mobile health apps related to the WHO[b] modifiable health behaviors (physical activity, diet, weight loss, sedentary behavior, smoking, sleep, and alcohol use) [2] or mental health.	Mobile health apps focusing on managing chronic conditions (eg, tracking health symptoms such as glucose, blood pressure, or pain and managing medication), making medical appointments, or accessing telehealth services. Research involving the use of wearable technology, such as fitness trackers.
Design	All qualitative (eg, interviews, focus groups) and quantitative (eg, surveys, app use data) observational research designs in free-living conditions (ie, where participants used a publicly available app before enrolling in the research study).	Research in non-free-living conditions (eg, clinical trials, app development, and testing) and secondary research (reviews).
Evaluation	Outcome measures relating to app abandonment such as reasons for app abandonment (eg, technical issues, privacy) and duration of app use.	No app abandonment outcome measures are present.
Research type	Qualitative, quantitative, or mixed methods research designs. Peer-reviewed journal articles published in English from 2007 onwards to coincide with the first iPhone (Apple Inc) release. We focused on peer-reviewed literature to ensure a high standard of evidence quality and maintain feasibility within our resource constraints.	Research in non-free-living conditions. Gray literature and studies not published in English.

[a]SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, Research type.

[b]WHO: World Health Organization.

Information Sources and Search Strategy

A database search was conducted on December 5, 2022, and included 4 databases: MEDLINE, Scopus, Embase, and PsycINFO. The search strategy was devised with the academic team and finalized with the support of an academic librarian.

The searches consisted of combining synonyms for the phenomenon of interest (mobile health apps), app types (diet, physical activity, etc), and evaluation (app abandonment). The search was also restricted to adult humans, English, and studies from 2007 (when the first iPhone was released) to the present. Terms were mapped to MeSH (Medical Subject Headings) as appropriate. Reference lists of included studies were scanned, to determine additional studies that met the inclusion criteria. The full search strategy can be found in Multimedia Appendix 2.

Evidence Selection and Data Charting

Database search results were collated and imported into EndNote V.x9 (Clarivate). Duplicates were removed in EndNote and checked manually. Studies were screened in Covidence (Veritas Health Innovation) based on title and abstracts in duplicate by 2 independent reviewers (PGK and AW), as per best practice to reduce bias [22]. The results of the screening were compared and discussed until a consensus was reached. Another independent reviewer (CAM and RGC) was consulted as needed, to assist in resolving any discrepancies.

Next, full-text screening was conducted to define eligible studies. This was conducted by 3 independent reviewers. PGK, RGC, and AW completed a pilot full-text screening of 5 studies to test the criteria. RGC and PGK completed 20% (13/63) of the full-text screening in duplicate, with a Cohen κ score of 0.82. PGK completed the screening of the remaining studies.

A custom document was generated and used for data charting (Multimedia Appendix 3). Charted data included country of origin, study design, health app domain, aims or objectives, sample size, population, app abandonment definition, and outcome measures (reasons for abandonment or duration). If unclear, the authors were contacted to confirm the data. Pilot data charting of 5 studies (28%, 5/18) was conducted to test the quality of the form. PGK completed data charting and, if there were any concerns, other members of the team were contacted to reach a unanimous verdict before proceeding.

Critical Appraisal

Two tools were used to determine the risk of bias based on the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [23] checklist for observational studies and the COREQ (Consolidated Criteria for Reporting Qualitative Research) [24] checklist for qualitative studies. These 2 tools were needed due to the heterogeneity of the methodologies of the included studies. Mixed methods studies were evaluated using the tool that captured the majority of charted data. If the study met the criteria, it received a 1. If the item was not applicable or not satisfied, it was scored 0. PGK completed data charting and risk of bias, consulting the team for advice in the case of uncertainty. While not required for scoping reviews, we conducted a critical appraisal to enhance the rigor of our review and provide context for interpreting findings. This aligns with PRISMA-ScR guidelines and helps identify methodological gaps in the field.

Synthesis of Results

In order to facilitate data synthesis across diverse research designs, overarching categories were constructed to describe the health domains, study design, and outcomes using qualitative content analysis. After data extraction was completed, a list of all reported health domains and outcomes were compiled. After

XSL•FO
RenderX

familiarization with the data, the first author sorted them into similar categories using an inductive approach (ie, directed by the data with no preconceived categories) [25]. These categories were reviewed with the senior author, refined, and named.

A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the search and screening process was presented along with a descriptive discussion of these results. A study characteristics table used descriptive statistics (n and %) to describe the year, country, health app domain, outcome measures, and study design. The risk of bias for the included studies based on either the STROBE or COREQ reporting guidelines was reported descriptively.

Due to the review aims and study heterogeneity, the results were presented in 2 sections. First, quantitative findings were presented in a scatterplot depicting duration for use for different health app domains over time, before being discussed descriptively with the range of abandonment percentage intervention lengths calculated.

For the synthesis of qualitative data, a basic qualitative content analysis using an inductive approach was used, enabling the organization of categories and subcategories from the data [25,26]. To do this, codes were assigned to the extracted qualitative data on reasons for app abandonment. We then organized the codes into categories and subcategories, to reach an integrated understanding of why people abandoned apps. Coding was primarily undertaken by PGK and reviewed by the other 3 authors.

Results

Selection of Sources of Evidence

The database search discovered 1805 studies (see Figure 1 for PRISMA flow diagram), with 1058 unique studies sent to title and abstract screening after the removal of duplicates (n=747). In total, 63 studies were assessed during full-text screening, with 45 being excluded. Reasons included not having an outcome measure related to app abandonment (n=23), nonpeer-reviewed research (n=6), not original research (n=6), wrong health app domain (n=5), non–free-living conditions (n=4) and not being written in English (n=1). As a result, 18 studies were deemed eligible for this review [11,12,27-42].

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart.

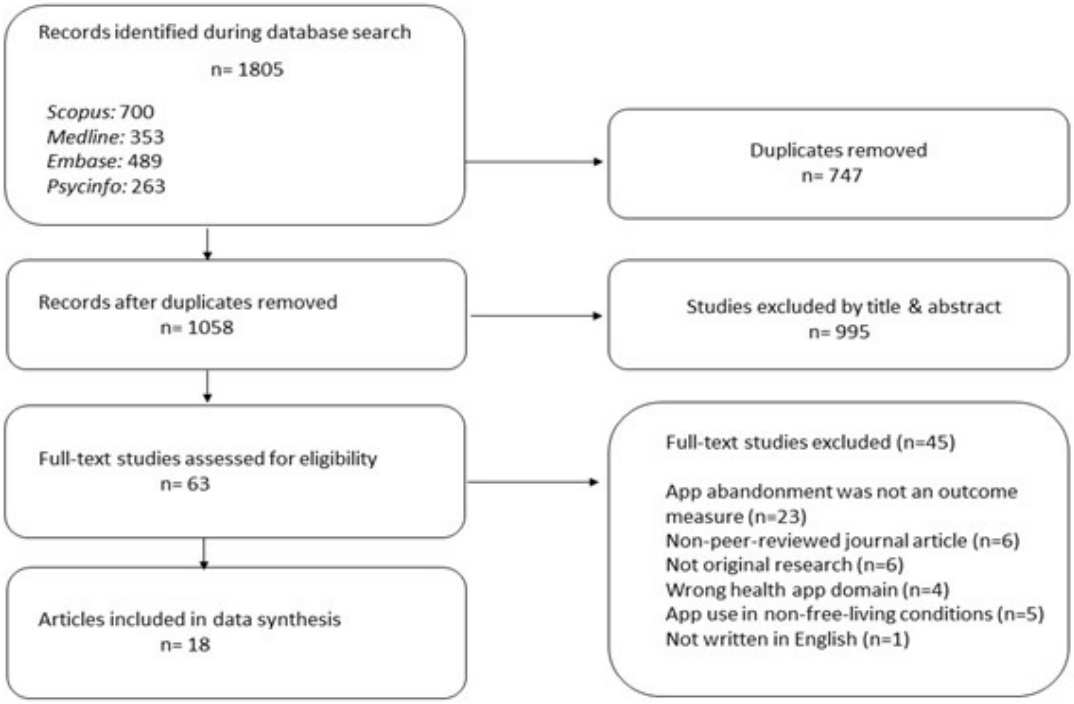

Characteristics for Sources of Evidence

All 18 studies [11,12,27-42] included in this review were written in English and published between 2014 and 2022, with half being published between 2019 and 2022. The total sample size was 525,824 participants, with the range being 10-189,770 participants. Most studies were conducted in the United States (n=6), Canada (n=3), the United Kingdom (n=2), and Germany (n=2). There were 11 quantitative studies [11,30-34,36-38,40], 4 qualitative studies [27,35,39], and 3 mixed methods studies [12,29,41,42]. Some studies reported singular app domains while others reported multiple. Of the 18 studies included in the review, 13 reported on time to app abandonment [11,29-38,40,42], and one-third (n=6) described reasons for abandonment [12,27-29,39,41], with 1 study exploring both [11]. A full summary of study characteristics can be found in Table 2.

XSL•FO
RenderX

Table 2. Summary of characteristics of studies included in the scoping review (n=18).

Study characteristic	Studies, n (%)
Year	
2019-2022	12 (69)
2015-2018	5 (28)
2011-2014	1 (6)
2007-2010	__[a]
Country	
United States	5 (28)
Canada	3 (17)
United Kingdom	2 (11)
Germany	2 (11)
Other	6 (36)
Health domain	
Mental health	4 (22)
Physical activity	3 (17)
Diet	3 (17)
Physical activity and diet	3 (17)
Alcohol	2 (11)
Diet and physical activity and mindfulness	1 (6)
Alcohol consumption	1 (6)
Sport and fitness	1 (6)
Study design	
Quantitative	11 (61)
Qualitative	4 (22)
Mixed methods	3 (17)
Outcome category	
Time to abandonment	13 (72)
Reasons for abandonment	5 (28)
Both	1 (6)

[a]Not applicable.

Critical Appraisal Within Sources of Evidence

Thirteen studies [11,29-34,36-38,40-42] were assessed using the STROBE tool, with scores ranging from 16-21 out of a possible 26 (refer to Multimedia Appendix 4 for full results). All studies satisfied the criteria for having a structured abstracted summary, study rationale, stating key objectives and design elements, defining outcome variables, reporting demographic data, and setting information. Criteria less likely to be satisfied included indicating the study design in the title, rationale for the study size, any efforts to address bias, explanation of how missing data was handled, reasons for nonparticipation, and follow-up times (if relevant).

Five studies [12,27,28,35,39] were assessed using the COREQ tool, with scores ranging from 10-18 out of a possible 25 (refer to Multimedia Appendix 5 for full results). All studies satisfied the criteria for having a structured abstract summary, and study rationale, stating key objectives, design elements, definition of outcome measures, and participant characteristics. Criteria less likely to be satisfied included participant knowledge about the interviewers, evidence of interview pilot testing, presence of nonparticipants, details of when field notes were made, return of transcripts for correction, discussion of data saturation, participant feedback on findings, and duration of interviews.

The key findings table for the included studies can be found in Multimedia Appendix 6. Definitions of abandonment varied widely across the studies. For example, terminology included various terms such as abandonment, churn, and disengagement as well as retention. One study defined abandonment as no use for at least 7 days [31], while others defined it as no use for at least 14 days [33] and at least 4 weeks [40]. In addition,

XSL•FO
RenderX

abandonment was measured at different time points across studies.

Duration of App Use

Figure 2 graphically represents the duration apps were used before their abandonment across the examined studies. A curvilinear relationship can be seen, where health apps tend to be abandoned by many soon after they are acquired, with the rate of abandonment slowing with time. Though this general pattern is consistent across studies, the abandonment rate differed notably among studies. For example, Attwood et al [29] and Owen et al [42] found that half of the users stopped using the app within the initial days. In contrast, Lau et al [38] observed that users took approximately 240 days before half of them abandoned the app. Most studies (8 out of 10, assuming the studies which followed their users for <100 days continued on the same trajectory) reported that 50% or more of users had abandoned their app within the first 100 days of initial use, with a median abandonment rate at 100 days of approximately 70%.

While abandonment increased with time, no study reported complete abandonment by its participants. Attwood et al [29] and Bell et al [31] reported that over 95% of users had ceased using their apps by their last recorded follow-up timepoint, ranging from 80 to 300 days post initiation. Visual inspection suggested that app abandonment may occur more quickly for alcohol and smoking apps, whereas physical activity and possibly mental health apps appeared to achieve more prolonged use before abandonment.

Figure 2. Duration of app use before abandonment. Four studies could not be included in the graph due to the way the data were reported.

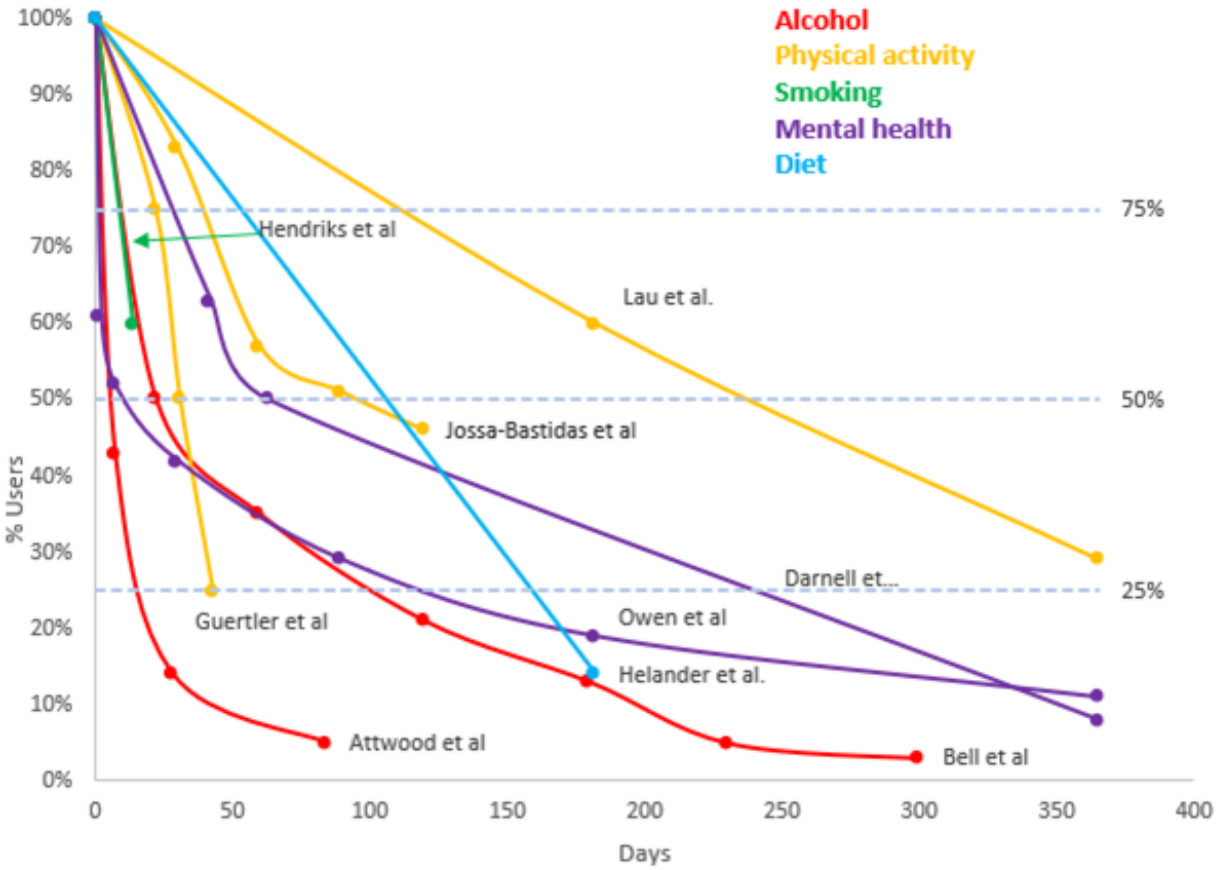

Reasons for Abandonment

In total, 6 studies collected data on reasons for app abandonment [11,12,27,28,39,41]. Table 3 summarizes the reasons for abandonment. A total of 22 unique reasons were identified, which were arranged into 12 subcategories and 6 overarching abandonment categories.

XSL•FO
RenderX

Table 3. Reasons for app abandonment were identified in the scoping review.

Category and subcategory	Codes
Technical and functional issues	
Device compatibility	• It no longer works on my phone [11]
Data management	• Data loss [27] • Needing to create an account before using the app [27]
Privacy	
General privacy	• Privacy [27,41]
Data sharing discomfort	• Did not like sharing data with the app [11]
User experience	
Effectiveness and engagement	• Poor app effectiveness [11] • The app was not engaging [41]
Design and usability challenges	• Confusing or difficult [11,28,39,41] • Annoying notifications [28]
Content and features	
Lacking features	• Lack of personalization [28] • Lack of accountability [28] • Did not have the desired features [41] • Lack of variety or options [27,39]
Lacking content	• Lack of knowledge within the app [12] • Lack of “professionals” [12]
Time and financial costs	
Unexpected expenses and obligations	• Time to enter data [11,28] • Hidden costs [11]
Evolving user needs and goals	
Achievement and transition	• No need to track anymore [12] • I achieved my fitness goal [11,12,41]
Interest and motivation shift	• Fading motivation [11,12,41] • Got bored [11,41]
App selection and replacement	• Identified the most suitable app and uninstalled the rest [11,41]

Technical and functional issues led to abandonment across mobile health apps [11,27]. For example, some users reported abandoning apps due to phone compatibility issues [11], while others described abandoning an app after a technical error and resultant data loss, and due to the need to create an account before using an app [27].

Concerns regarding privacy [27,41] along with user discomfort with and concern regarding data sharing [11] were contributing factors to mobile health app abandonment.

Poor user experience played a role in users abandoning mobile health apps [11,12,27,28,39,41]. Users reported that their apps were not engaging [41] or were perceived to lack effectiveness [11]. In addition, design and usability concerns contributed to abandonment, particularly where the apps were considered confusing to use [11,28,39,41]. In addition, annoyance from notifications [28] contributed to abandonment.

Participants reported lack of desired content and features contributed to them abandoning their apps. This included a lack of personalization, accountability [28], and other desired features [41], lacking variety [27,39]' insufficient knowledge content [12], and a lack of professional advice from experts [12].

Time and financial costs also contributed to abandonment. This was particularly the case where there were unexpected hidden costs within apps [11] or burdensome data entry [11,28].

Finally, many users cited their evolving goals and needs as reasons for abandoning mobile health apps [11,12,41]. Entering a state of transition where users no longer needed to track their data [12] or had achieved their goals [11,12,41] led to user abandonment. Moreover, loss of motivation [11,12,41] and users reporting feeling bored with their app [11,41] resulted in mobile health app abandonment. Finally, some users reported

XSL•FO
RenderX

downloading several potential apps, identifying the most suitable, and uninstalling the rest [11,41].

Discussion

Principal Findings

This study set out to examine the abandonment of health apps for lifestyle management in free-living conditions and to understand the underlying reasons for this abandonment. The duration of app use before abandonment varied across studies, but a common trend showed many users abandoning apps soon after the acquisition, with the abandonment rate then slowing over time. Most studies (7/9) reported more than 50% of users abandoning their apps within the first 100 days. The review identified several reasons for app abandonment, categorized into 6 overarching categories, that are technical and functional issues, privacy concerns, poor user experience, content and features, time and financial costs, and evolving user needs and goals. It appeared that physical activity and mental health apps might be used longer while smoking and alcohol apps appear to be abandoned more swiftly. Furthermore, our review highlights a considerable methodological diversity across studies, particularly in the measurement and terminology used to describe app abandonment.

Our findings reveal a common trend where users tend to abandon health apps soon after acquisition, with a slower rate of abandonment among the remaining users over time. The abandonment rates for different app domains varied, with physical activity apps (n=3) showing a 54%-75% abandonment rate, alcohol apps (n=2) at 95%-97%, smoking apps (n=1) at 40%, mental health apps (n=2) at 89%-92%, and diet apps (n=1) at 86%. These rates are notably higher than the 43% attrition rate reported in the meta-analysis by Meyerowitz-Katz et al [17], likely due to their inclusion of both controlled trial and free-living data. Controlled trials typically feature rigorous screening and follow-up processes that help reduce attrition. Conversely, our review found that in the majority of studies, over half of the users discontinued app use within the first 100 days, with a median abandonment rate of approximately 70%. This is consistent with industry reports indicating that around 69% of fitness and nutrition apps are abandoned within the first 90 days, followed by a slower rate of abandonment resulting in 81% abandonment over a year [9].

Our synthesis of qualitative data suggested that there are many and wide-ranging potential reasons for app abandonment. We identified 22 different reasons, which fell into 6 overarching categories, that are technical and functional issues, privacy concerns, poor user experience, content and features, time and financial costs, and evolving user needs and goals. Some of these findings inversely mirror the factors identified in the Jakob et al [16] and Amagai et al [14] reviews as influencing positive engagement with health apps. For instance, while Jakob et al [16] found that personalization enhances adherence, our study noted that a lack of personalization contributes to abandonment. Similarly, support from health professionals was seen as beneficial in Amagai and colleagues' [14] study, whereas our findings indicate that the absence of expert content can lead to app discontinuation. This pattern is also observed with technical stability, where Jakob et al [16] identified it as a factor for adherence, contrasting with our identification of technical issues as a reason for abandonment. User experience issues such as being difficult or confusing to use were most consistently identified as being crucial to app abandonment in our review. This finding aligns with the results of a previous study, which extracted over 5 million user reviews from 278 apps and found that most complaints were linked to user experience issues [43].

The reasons for abandonment may be viewed in the context of the curvilinear relationship between app abandonment and time, suggesting there may be 2 key periods of abandonment, that consist of an initial phase characterized by early failure or disappointment, followed by a more gradual decline in interest. Early abandonment might be attributed to immediate deterrents such as device incompatibility, technical errors, or barriers like mandatory account creation before app usage. For example, research examining app user reviews found that complaints related to the sign-up process, such as complex sign-up forms and not receiving verification codes, were not uncommon [43]. These could potentially cause users to abandon apps before they have even begun using them. In contrast, factors like annoying notifications, a lack of desired features, or evolving user motivations could contribute to a more gradual disengagement over time. Research suggests that examining user experience over time is important to identify critical issues at different stages of usage [44].

Our findings suggested that the time to abandon mobile health apps may vary for different health app domains, with a seemingly sharper rate of abandonment for apps targeting smoking and alcohol, and more gradual abandonment for apps targeting physical activity and mental health. However, this finding should be interpreted with caution, as it is derived from visual inspection of data from a relatively small number of studies. The Meyerowitz-Katz et al [17] and Jakob et al [16] reviews each attempted to consider whether attrition and adherence (respectively) were associated with app domains and suggested that findings were inconclusive due to rates of attrition and adherence varying widely between different studies and methodological heterogeneity. Certainly, there are logical reasons for why app abandonment may vary by health domain. For example, smoking and alcohol apps target addictive behaviors, which can be particularly challenging to change [45,46]. Users might download these apps during a moment of motivation but find it difficult to maintain the commitment, leading to early abandonment. Also, the immediate withdrawal symptoms or cravings associated with quitting smoking or alcohol may lead to relapse and subsequent app abandonment. Other factors, like social features, stigma, and privacy concerns are also likely to differ between different health domains. For example, physical activity apps often include social features like sharing progress with friends or participating in challenges, which might enhance user engagement [47]. Conversely, there may be greater privacy concerns or stigma associated with using apps for smoking or alcohol cessation, particularly when using social support features [48,49].

Key strengths of our study include its comprehensive scope, covering a broad range of health app domains and contexts, to provide a holistic understanding of app abandonment. The

XSL•FO
RenderX

adherence to the PRISMA-ScR protocol ensured a systematic and transparent review process, enhancing the reliability of our findings. We included methodological features that went beyond those expected for scoping reviews, including undertaking study screening in duplicate. The use of an inductive approach to content analysis allowed for the organization of categories directly from the data, offering insights that are grounded in empirical evidence.

The study's most important limitations arose from the limitations of the evidence base itself, in particular, the heterogeneity in included studies' methods and their definitions of abandonment, making comparison across studies challenging. Focusing on peer-reviewed journal articles and excluding gray literature might have introduced publication bias, as studies with significant or positive findings are more likely to be published. Limiting the review to English-language studies may exclude relevant research published in other languages, potentially introducing language bias. We also note that the use of negatively framed terminology to describe app abandonment (eg, abandonment, disengagement) might have resulted in the exclusion of studies that described how long participants used an app with only positively-framed terminology (eg, usage). Furthermore, while inductive content analysis is a strength, it also carries the risk of subjective interpretation of qualitative data.

This study highlights that to date, there has been limited research focused on app abandonment, an area that is crucial to optimizing the health benefits of apps. More longitudinal studies are needed to understand app abandonment patterns across various health behaviors. Furthermore, our field needs to establish a consistent definition and reporting approach to app abandonment within the literature to enable comparative analysis across studies.

Conclusion

In conclusion, this study sheds light on the often-overlooked issue of health app abandonment in real-world settings, revealing key patterns and underlying reasons for this phenomenon. Our findings indicate that a large number of users abandon health apps shortly after downloading. The reasons for abandonment are multifaceted, encompassing technical and functional issues, privacy concerns, poor user experience, content inadequacies, financial considerations, and evolving user needs. These insights highlight critical areas for future research and development in health app design and implementation. There is a pressing need for more longitudinal studies to understand app abandonment patterns across various health behaviors and to establish a consistent definition and reporting approach for app abandonment. Such efforts are essential for optimizing the health-enhancing potential of health apps, guiding future development, and ensuring these digital tools effectively contribute to public health and individual well-being.

Acknowledgments

CAM conceived the study. PGK, RGC, AW, and CAM design the study. PGK conducted database searches. PGK, RGC, and AW conducted screening. PGK conducted data charting, analyzed the data, and drafted the manuscript. All authors interpreted the data, revised the manuscript, and reviewed and approved the final version. This study received no funding. CAM is supported by a Medical Research Future Fund Emerging Leader Grant (GNT1193862).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[PDF File (Adobe PDF File), 517 KB - jmir_v26i1e56897_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 24 KB - jmir_v26i1e56897_app2.docx]

Multimedia Appendix 3
Data charting form.
[XLSX File (Microsoft Excel File), 9 KB - jmir_v26i1e56897_app3.xlsx]

Multimedia Appendix 4
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Risk of bias results.
[DOCX File , 33 KB - jmir_v26i1e56897_app4.docx]

Multimedia Appendix 5
COREQ (Consolidated Criteria for Reporting Qualitative Research) Risk of bias results.
[DOCX File , 22 KB - jmir_v26i1e56897_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Study characteristics and key findings.
[DOCX File , 23 KB - jmir_v26i1e56897_app6.docx]

References

1. Hajat C, Stein E. The global burden of multiple chronic conditions: a narrative review. Prev Med Rep 2018;12:284-293 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.008] [Medline: 30406006]
2. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the global burden of disease study 2010. Lancet 2012;380(9859):2224-2260 [FREE Full text] [doi: 10.1016/S0140-6736(12)61766-8] [Medline: 23245609]
3. Castelpietra G, Knudsen AKS, Agardh EE, Armocida B, Beghi M, Iburg KM, Madureira-Carvalho, et al. The burden of mental disorders, substance use disorders and self-harm among young people in Europe, 1990-2019: findings from the global burden of disease study 2019. Lancet Reg Health Eur 2022;16:100341 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100341] [Medline: 35392452]
4. Lee JA, Choi M, Lee SA, Jiang N. Effective behavioral intervention strategies using mobile health applications for chronic disease management: a systematic review. BMC Med Inform Decis Mak 2018;18(1):12 [FREE Full text] [doi: 10.1186/s12911-018-0591-0] [Medline: 29458358]
5. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002 [FREE Full text] [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
6. Number of smartphone mobile network subscriptions worldwide from 2016 to 2022, with forecasts from 2023 to 2028. Statista. 2023. URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide [accessed 2024-11-12]
7. The Brainy Insights. mHealth Market Size: Global Industry Analysis, Share, Growth, Trends, and Forecast 2023 to 2032. Pune, India: The Brainy Insights; 2023.
8. IQVIA Inc. Digital health trends 2021. IQVIA Inc. 2021. URL: https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/digital-health-trends-2021/iqvia-institute-digital-health-trends-2021.pdf?_=1628089218603 [accessed 2024-11-12]
9. Alchemer. Healthcare apps: 2021 engagement benchmarks. Louisville, US URL: https://www.alchemer.com/resources/blog/healthcare-apps-2021-engagement-benchmarks/ [accessed 2024-11-12]
10. Kinvey. CIO & mobile leader: State of enterprise mobility survey. Boston, US: Kinvey Inc; 2014. URL: https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/KINVEYUS/K141112K.pdf [accessed 2024-11-12]
11. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]
12. Lu X, Chen Y, Epstein DA. A model of socially sustained self-tracking for food and diet. Proc. ACM Hum.-Comput. Interact 2021;5(CSCW2):1-32. [doi: 10.1145/3479595]
13. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
14. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022;24(4):e35120 [FREE Full text] [doi: 10.2196/35120] [Medline: 35471414]
15. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
16. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
17. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
18. Kim HS, Lee S, Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. J Korean Med Sci 2018;33(34):e213 [FREE Full text] [doi: 10.3346/jkms.2018.33.e213] [Medline: 30127705]
19. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ Digit Med 2020;3:67 [FREE Full text] [doi: 10.1038/s41746-020-0277-8] [Medline: 32411828]

XSL•FO
RenderX

20. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
21. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012;22(10):1435-1443. [doi: 10.1177/1049732312452938] [Medline: 22829486]
22. Peters MDJ, Marnie C, Colquhoun H, Garritty CM, Hempel S, Horsley T, et al. Scoping reviews: reinforcing and advancing the methodology and application. Syst Rev 2021;10(1):263 [FREE Full text] [doi: 10.1186/s13643-021-01821-3] [Medline: 34625095]
23. Ghaferi AA, Schwartz TA, Pawlik TM. STROBE reporting guidelines for observational studies. JAMA Surg 2021;156(6):577-578. [doi: 10.1001/jamasurg.2021.0528] [Medline: 33825815]
24. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
26. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, de Moraes, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
27. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020;26(3):2042-2066 [FREE Full text] [doi: 10.1177/1460458219896492] [Medline: 31920160]
28. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR Mhealth Uhealth 2019;7(8):e12983 [FREE Full text] [doi: 10.2196/12983] [Medline: 31469081]
29. Attwood S, Parke H, Larsen J, Morton KL. Using a mobile health application to reduce alcohol consumption: a mixed-methods evaluation of the drinkaware track & calculate units application. BMC Public Health 2017 May 17;17(1):394 [FREE Full text] [doi: 10.1186/s12889-017-4358-9] [Medline: 28511698]
30. Jossa-Bastidas O, Zahia S, Fuente-Vidal A, Sánchez Férez N, Roda Noguera O, Montane J, et al. Predicting physical exercise adherence in fitness apps using a deep learning approach. Int J Environ Res Public Health 2021;18(20):10769 [FREE Full text] [doi: 10.3390/ijerph182010769] [Medline: 34682515]
31. Bell L, Garnett C, Qian T, Perski O, Williamson E, Potts HW. Engagement with a behavior change app for alcohol reduction: data visualization for longitudinal observational study. J Med Internet Res 2020;22(12):e23369 [FREE Full text] [doi: 10.2196/23369] [Medline: 33306026]
32. Giraldo‐O'Meara M, Doron G. Can self-esteem be improved using short daily training on mobile applications? Examining real world data of GG Self-esteem users. Clinical Psychologist 2021;25(2):131-139. [doi: 10.1080/13284207.2021.1923126]
33. Guertler D, Vandelanotte C, Kirwan M, Duncan MJ. Engagement and nonusage attrition with a free physical activity promotion program: the case of 10,000 steps Australia. J Med Internet Res 2015;17(7):e176 [FREE Full text] [doi: 10.2196/jmir.4339] [Medline: 26180040]
34. Helander E, Kaipainen K, Korhonen I, Wansink B. Factors related to sustained use of a free mobile app for dietary self-monitoring with photography and peer feedback: retrospective cohort study. J Med Internet Res 2014;16(4):e109 [FREE Full text] [doi: 10.2196/jmir.3084] [Medline: 24735567]
35. Hendriks Y, Peek S, Kaptein M, Bongers I. Process and information needs when searching for and selecting apps for smoking cessation: qualitative study using contextual inquiry. JMIR Hum Factors 2022;9(2):e32628 [FREE Full text] [doi: 10.2196/32628] [Medline: 35436217]
36. König LM, Sproesser G, Schupp HT, Renner B. Describing the process of adopting nutrition and fitness apps: behavior stage model approach. JMIR Mhealth Uhealth 2018;6(3):e55 [FREE Full text] [doi: 10.2196/mhealth.8261] [Medline: 29535078]
37. Kwon H, Kim HH, An J, Lee J, Park YR. Lifelog data-based prediction model of digital health care app customer churn: retrospective observational study. J Med Internet Res 2021;23(1):e22184 [FREE Full text] [doi: 10.2196/22184] [Medline: 33404511]
38. Lau EY, Mitchell MS, Faulkner G. Long-term usage of a commercial mHealth app: a "multiple-lives" perspective. Front Public Health 2022;10:914433 [FREE Full text] [doi: 10.3389/fpubh.2022.914433] [Medline: 36438245]
39. Lupton D. Australian women's use of health and fitness apps and wearable devices: a feminist new materialism analysis. Feminist Media Studies 2019;20(7):983-998. [doi: 10.1080/14680777.2019.1637916]
40. Darnell D, Pullmann MD, Hull TD, Chen S, Areán P. Predictors of disengagement and symptom improvement among adults with depression enrolled in talkspace, a technology-mediated psychotherapy platform: naturalistic observational study. JMIR Form Res 2022;6(6):e36521 [FREE Full text] [doi: 10.2196/36521] [Medline: 35731563]
41. Mustafa AS, Ali N, Dhillon JS, Alkawsi G, Baashar Y. User engagement and abandonment of mHealth: a cross-sectional survey. Healthcare (Basel) 2022;10(2):221 [FREE Full text] [doi: 10.3390/healthcare10020221] [Medline: 35206837]
42. Owen JE, Jaworski BK, Kuhn E, Makin-Byrd KN, Ramsey KM, Hoffman JE. mHealth in the wild: using novel data to examine the reach, use, and impact of PTSD coach. JMIR Ment Health 2015;2(1):e7 [FREE Full text] [doi: 10.2196/mental.3935] [Medline: 26543913]

XSL•FO
RenderX

43. Haggag O, Grundy J, Abdelrazek M, Haggag S. A large scale analysis of mHealth app user reviews. Empir Softw Eng 2022;27(7):196 [FREE Full text] [doi: 10.1007/s10664-022-10222-6] [Medline: 36246486]
44. Feng L, Wei W. An epirical study on user experience evaluation and identification of critical UX issues. Sustainability 2019;11(8):2432. [doi: 10.3390/su11082432]
45. Chaiton M, Diemert L, Cohen JE, Bondy SJ, Selby P, Philipneri A, et al. Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers. BMJ Open 2016;6(6):e011045 [FREE Full text] [doi: 10.1136/bmjopen-2016-011045] [Medline: 27288378]
46. Garbusow M, Sebold M, Beck A, Heinz A. Too difficult to stop: mechanisms facilitating relapse in alcohol dependence. Neuropsychobiology 2014;70(2):103-110. [doi: 10.1159/000362838] [Medline: 25359490]
47. Maher CA, Olds T, Vandelanotte C, Plotnikoff R, Edney SM, Ryan JC, et al. Gamification in a physical activity app: what gamification features are being used, by whom, and does it make a difference? Games Health J 2022;11(3):193-199. [doi: 10.1089/g4h.2021.0207] [Medline: 35501981]
48. Perski O, Blandford A, Ubhi HK, West R, Michie S. Smokers' and drinkers' choice of smartphone applications and expectations of engagement: a think aloud and interview study. BMC Med Inform Decis Mak 2017;17(1):25 [FREE Full text] [doi: 10.1186/s12911-017-0422-8] [Medline: 28241759]
49. Struik LL, Bottorff JL, Baskerville NB, Oliffe JL. The crush the crave quit smoking app and young adult smokers: qualitative case study of affordances. JMIR Mhealth Uhealth 2018;6(6):e134 [FREE Full text] [doi: 10.2196/mhealth.9489] [Medline: 29884602]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, Research type
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by A Mavragani; submitted 01.02.24; peer-reviewed by P Voorheis, F Burstein; comments to author 12.06.24; revised version received 05.08.24; accepted 17.09.24; published 18.12.24.

Please cite as:
Kidman PG, Curtis RG, Watson A, Maher CA
When and Why Adults Abandon Lifestyle Behavior and Mental Health Mobile Apps: Scoping Review
J Med Internet Res 2024;26:e56897
URL: https://www.jmir.org/2024/1/e56897
doi:10.2196/56897
PMID:

©Patrick G Kidman, Rachel G Curtis, Amanda Watson, Carol A Maher. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Investigating eHealth Lifestyle Interventions for Vulnerable Pregnant Women: Scoping Review of Facilitators and Barriers

Ashley JP Smit[1], MSc; Isra Al-Dhahir[2], MSc; Lieke Schiphof-Godart[3], PhD; Linda D Breeman[2], PhD; Andrea WM Evers[2,4], Prof Dr; Koen FM Joosten[1], MD, Prof Dr

[1]Department of Neonatal and Pediatric Intensive Care, Erasmus Medical Center, Sophia Children's Hospital, Rotterdam, Netherlands
[2]Faculty of Social and Behavioral Sciences, Health, Medical and Neuropsychology Unit, Leiden University, Leiden, Netherlands
[3]Department of Medical Informatics, Erasmus University Medical Center Rotterdam, Rotterdam, Netherlands
[4]Medical Delta, Leiden University, Delft University of Technology, Erasmus University, Delft, Netherlands

Corresponding Author:
Ashley JP Smit, MSc
Department of Neonatal and Pediatric Intensive Care
Erasmus Medical Center
Sophia Children's Hospital
Dr. Molewaterplein 40
Rotterdam, 3015 GD
Netherlands
Phone: 31 10 704 0704
Email: a.j.p.smit@erasmusmc.nl

Abstract

Background: The maintenance of a healthy lifestyle significantly influences pregnancy outcomes. Certain pregnant women are more at risk of engaging in unhealthy behaviors due to factors such as having a low socioeconomic position and low social capital. eHealth interventions tailored to pregnant women affected by these vulnerability factors can provide support and motivation for healthier choices. However, there is still a lack of insight into how interventions for this target group are best designed, used, and implemented and how vulnerable pregnant women are best reached.

Objective: This review aimed to identify the strategies used in the design, reach, use, and implementation phases of eHealth lifestyle interventions for vulnerable pregnant women; assess whether these strategies acted as facilitators; and identify barriers that were encountered.

Methods: We conducted a search on MEDLINE, Embase, Web of Science, CINAHL, and Google Scholar for studies that described an eHealth intervention for vulnerable pregnant women focusing on at least one lifestyle component (diet, physical activity, alcohol consumption, smoking, stress, or sleep) and provided information on the design, reach, use, or implementation of the intervention.

Results: The literature search identified 3904 records, of which 29 (0.74%) met our inclusion criteria. These 29 articles described 20 eHealth lifestyle interventions, which were primarily delivered through apps and frequently targeted multiple lifestyle components simultaneously. Barriers identified in the design and use phases included financial aspects (eg, budgetary constraints) and technological challenges for the target group (eg, limited internet connectivity). In addition, barriers were encountered in reaching vulnerable pregnant women, including a lack of interest and time constraints among eligible participants and limited support from health care providers. Facilitators identified in the design and use phases included collaborating with the target group and other stakeholders (eg, health care providers), leveraging existing eHealth platforms for modifications or extensions, and adhering to clinical and best practice guidelines and behavior change frameworks. Furthermore, tailoring (eg, matching the content of the intervention to the target groups' norms and values) and the use of incentives (eg, payments for abstaining from unhealthy behavior) were identified as potential facilitators to eHealth use. Facilitators in the interventions' reach and implementation phases included stakeholder collaboration and a low workload for the intervention deliverers involved in these phases.

Conclusions: This scoping review offers a comprehensive overview of strategies used in different phases of eHealth lifestyle interventions for vulnerable pregnant women, highlighting specific barriers and facilitators. Limited reporting on the impact of the strategies used and barriers encountered hinders a complete identification of facilitators and barriers. Nevertheless, this review

XSL•FO
RenderX

sheds light on how to optimize the development of eHealth lifestyle interventions for vulnerable pregnant women, ultimately enhancing the health of both future mothers and their offspring.

(J Med Internet Res 2024;26:e54366) doi:10.2196/54366

KEYWORDS

eHealth; pregnancy; vulnerability; socioeconomic status; lifestyle intervention; intervention development; barriers; facilitators; mobile phone; PRISMA

Introduction

Background

Maintaining a healthy lifestyle during pregnancy benefits pregnancy outcomes and the health of the developing fetus in particular. For example, the maternal diet plays a significant role in embryonic growth and development [1,2], and engaging in physical activity during pregnancy has been associated with a decreased risk of conditions such as excessive maternal weight gain, preeclampsia, and gestational diabetes mellitus [3]. On the other hand, maternal smoking and alcohol use are associated with an increased risk of preterm birth, among many other detrimental outcomes [4-6]. In addition, high levels of stress during pregnancy have been associated with various adverse outcomes for mother and child, among them preterm birth [7]. However, not all pregnant women maintain healthy lifestyles. Various studies have found that most women in the periconceptional period have inadequate dietary intake, 40% to 78% engage in insufficient physical activity, and 5% to 14% smoke [8,9]. Furthermore, 45% of pregnant women experience stress [8]. These lifestyle behaviors are often negatively affected by nonmedical vulnerability factors such as a low level of education, a low socioeconomic position (SEP), and low social capital [10]. Therefore, pregnant women affected by these vulnerability factors (henceforth referred to as “vulnerable pregnant women”) are more likely to engage in unhealthy behaviors. Acknowledging the intersectionality of these vulnerability factors is crucial as they often compound each other’s effects, enhancing health disparities [11]. For instance, a pregnant woman with a low income may have limited access to nutritious food and simultaneously feel stressed due to financial problems. Similarly, a pregnant woman with a low level of education may encounter barriers in understanding health-related information and accessing appropriate support systems. In addition, cultural beliefs and practices can significantly influence lifestyle behaviors during pregnancy. For instance, certain cultural norms dictate dietary preferences or restrictions during pregnancy, influencing the nutritional intake of pregnant women [12].

The pregnancy period offers a unique opportunity to improve maternal health and, consequently, fetal health, rendering it a crucial time in which vulnerable pregnant women should be encouraged to adopt a healthier lifestyle [13]. Considering that many pregnant women use the internet and smartphone apps as a source of information on pregnancy [14], eHealth interventions targeting the lifestyle of vulnerable pregnant women have the potential to support them in making healthier choices. For example, a tablet delivered lifestyle intervention for underserved pregnant women (those who lack access to essential resources and support during pregnancy, often due to socioeconomic barriers) significantly reduced the number of risk behaviors, particularly in the areas of stress and smoking, and increased fruit and vegetable consumption [15]. The use of digital tools to deliver interventions offers many advantages. First, given the widespread access to the internet among the population [16], eHealth lifestyle interventions are highly accessible. Furthermore, the ability to tailor interventions toward the specific needs and values of their users can increase their effectiveness by enhancing user participation and engagement [17]. These advantages enable these interventions to reach diverse populations. However, despite the many advantages of eHealth lifestyle interventions for vulnerable pregnant women, challenges might be encountered in various phases of their development. For instance, when attempting to recruit participants for a smoking cessation app, researchers encountered difficulties due to a lack of interest among Medicaid-eligible pregnant women [18]. Furthermore, difficulties associated with downloading an app deterred pregnant women from using the eHealth lifestyle intervention Health-e Babies [19]. In addition, despite adapting the content to a level accessible to early-stage readers, a study by Song et al [20] revealed that a third of their participants had difficulty understanding the information provided through SMS text messages. This emphasizes the necessity for developers of eHealth lifestyle interventions targeting vulnerable pregnant women to consider the particular needs and skills of their target group. Research has underscored the significance of eHealth users possessing adequate literacy levels and proficient digital skills [21]. However, it is important to recognize that these prerequisites may present additional challenges for vulnerable groups [22,23].

This Study

Despite an increasing number of eHealth lifestyle interventions developed for vulnerable pregnant women, there is still a lack of insight into how these interventions are designed, used, and implemented and how vulnerable pregnant women are reached by these interventions. To address this gap, this study sought to extract insights from studies on existing eHealth lifestyle interventions developed for vulnerable pregnant women. These studies on developed interventions offer valuable information regarding their components, challenges faced, and strategies used in each phase (design, reach, use, and implementation) and, therefore, can serve as a valuable resource to guide future researchers in the development or adaptation of eHealth lifestyle interventions for vulnerable pregnant women. Therefore, this scoping review aimed to (1) identify the strategies used and barriers encountered in the design, reach, use, and implementation phase of existing eHealth lifestyle interventions

XSL•FO
RenderX

for vulnerable pregnant women; and (2) determine whether these strategies acted as facilitators in the aforementioned phases to provide future developers with an overview of the available knowledge regarding the impact of these strategies.

Methods

Design: Scoping Review

As the research area of eHealth lifestyle interventions for vulnerable pregnant women is still in its infancy, a scoping review was chosen as the appropriate method to summarize and disseminate research findings, allowing for the inclusion of literature with varying types of methodological designs. We conducted the search for this scoping review in February 2023 and conducted an updated search in June 2023. We did not publish a review protocol. The conduct of the scoping review was guided by the methodological framework for scoping reviews by Arksey and O'Malley [24] and the 2018 PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Multimedia Appendix 1) [25].

Search Strategy

An experienced librarian from Erasmus Medical Center formulated a search strategy together with the first author (AJPS) and conducted a literature search addressing the research objectives (Multimedia Appendix 2). The search strategy included key terms for pregnancy, digital interventions, and lifestyle components. Vulnerability was not part of the search strategy. Instead, the titles and abstracts were screened for vulnerability factors. Vulnerability was defined as characteristics that stratify health opportunities and outcomes based on the PROGRESS-Plus framework by Cochrane [26] and can refer to place of residence, race, ethnicity, culture, and language (henceforth referred to as "ethnicity"), educational level, SEP, social capital, and age. We included low health and digital literacy as additional vulnerability factors [27,28]. In addition, the authors of the included articles had to specifically mention the characteristic as contributing to disparities in health opportunities or outcomes. The databases searched included MEDLINE (1946-present), Embase (1971-present), Web of Science (1975-present), CINAHL (1982-present), and Google Scholar. Duplicate findings were removed. Furthermore, the reference lists of relevant reviews and of the included articles after full-text screening were examined to identify additional relevant articles.

Eligibility Criteria and Screening

Abstract screening was conducted using ASReview (version 1.1). ASReview is a free open-source screening assistant tool that uses machine learning to assist the reviewer in literature screening. When articles are included or excluded by the reviewer within this software, the ASReview algorithm learns which articles are relevant for the reviewer and adjusts the order of the articles to present the most relevant first. In this way, ASReview allows for a more efficient and time-saving manner of screening articles. The predefined stopping rules for screening entailed screening a minimum of 36.1% of the articles and encountering 25 consecutive nonrelevant articles. With these criteria, it was expected that no more relevant articles would be identified among the remaining unscreened articles [29]. The first author (AJPS) screened the articles based on titles and abstracts against the inclusion and exclusion criteria (Textbox 1). To ensure that the interventions included in this scoping review were relevant for populations from high-income countries, we limited the studies to those conducted in high-income countries as defined in the World Economic Situation and Prospects 2023 report by the United Nations [30]. When uncertainty existed about the relevance of the article based on the title and abstract, or when no abstract was available, the article was included for full-text screening as well. After the initial abstract screening, AJPS performed a full-text screening in which articles were assessed for a second time against the inclusion and exclusion criteria. Any uncertainty was discussed with the other authors. The reasons for exclusion of articles after full-text screening were recorded.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria

Inclusion criteria

- Description of an eHealth intervention that aims to change modifiable behaviors
- Focus on at least one lifestyle component (physical activity, smoking, alcohol consumption, diet, stress, or sleep)
- Targeting vulnerable pregnant women, where vulnerability is defined as characteristics that stratify health opportunities and outcomes based on the PROGRESS-Plus framework; authors also had to specifically mention the characteristic as contributing to disparities in health opportunities or outcomes
- Presenting information on the design, reach, use, or implementation of the intervention
- Interventions taking place in a high-income country
- Full text available in English

Exclusion criteria

- Interventions targeting medically vulnerable women (eg, women with diabetes gravidarum or preeclampsia)
- Interventions not primarily administered during pregnancy
- Interventions consisting of <50% of eHealth components
- Interventions that were online advertisements or campaigns
- Case reports and reviews

Data Extraction and Synthesis

We defined 4 phases based on 2 frameworks used in the process of eHealth intervention development, which we used as guides for data extraction and analysis. These were the design and use phases, derived from the Centre for eHealth Research Roadmap [31], and the reach and implementation phases, derived from the Reach, Effectiveness, Adoption, Implementation, and Maintenance framework [32]. For each article, AJPS extracted all strategies used and barriers encountered in the design, reach, use, or implementation phases, which was informed by the aforementioned frameworks. These strategies and barriers were then organized and presented into key themes for each phase. A strategy was classified as a facilitator if it was explicitly mentioned by the authors as beneficial to one or more phases of the intervention. If the authors suggested the strategy was possibly beneficial, it was referred to as a possible facilitator. Similarly, a factor was classified as a barrier if the authors explicitly mentioned it as detrimental to one or more phases of the intervention. When the authors suggested the factor as possibly detrimental, it was referred to as a possible barrier. If uncertainties arose regarding the phase that the strategy or encountered barrier belonged to or regarding its classification as a (possible) facilitator or barrier, discussions were conducted among the authors to address these uncertainties. In addition, we extracted information on the intervention components and theoretical frameworks used in the interventions. Furthermore, the study characteristics from the selected articles were organized in a Microsoft Excel file (Multimedia Appendix 3 [15,18-20,33-57]). We used a narrative synthesis to address our research question.

Results

Study Selection

The systematic search across the databases revealed 3904 potentially relevant citations. After screening 1409 titles and abstracts using ASReview, 73 (5.18%) articles were retained for full-text screening. A total of 36% (26/73) of these articles met the inclusion criteria and were included in this review, along with 3 articles that were identified through reference checking, resulting in 29 included articles (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of the article screening process.

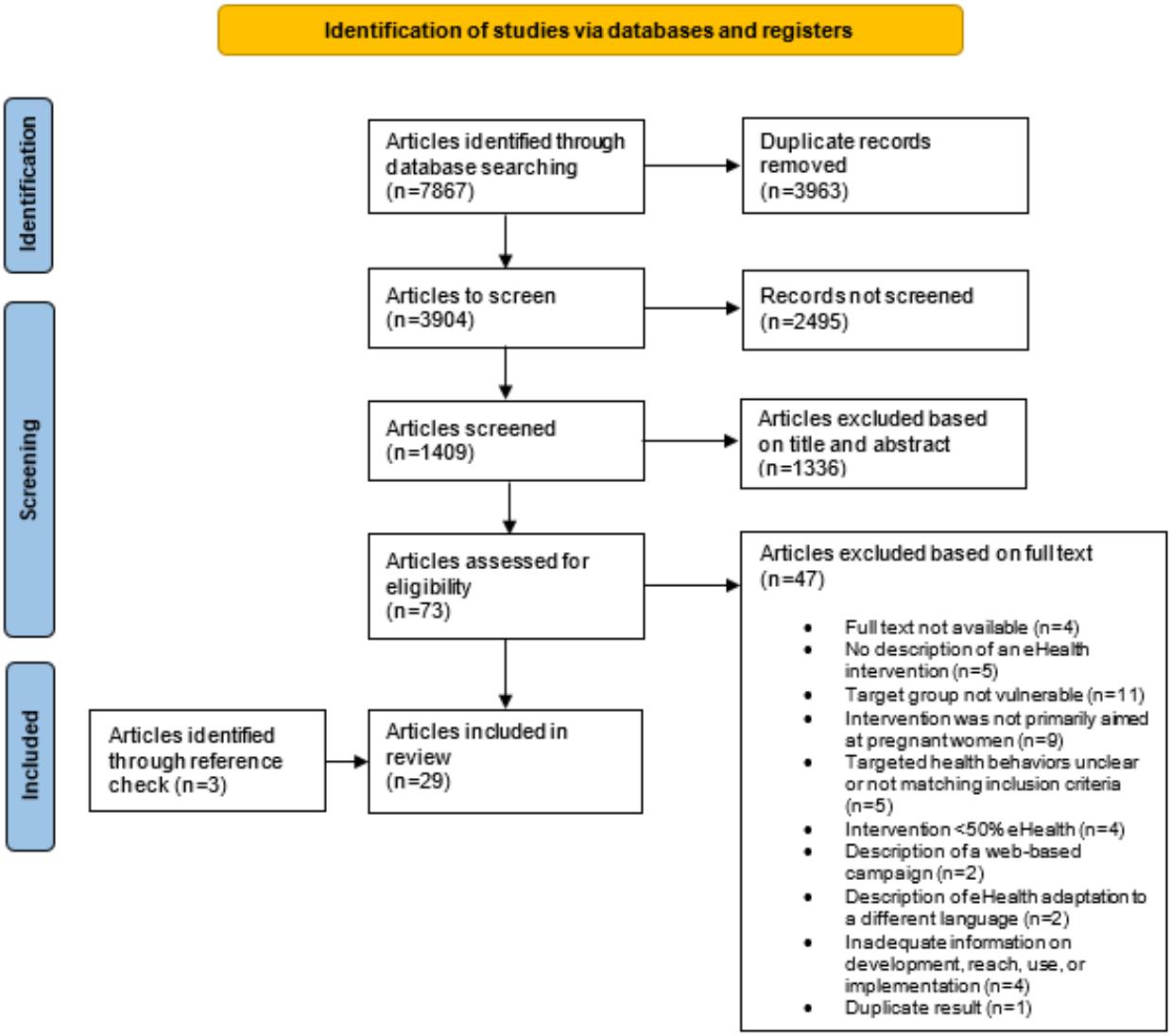

Description of the Included Studies

A summary of the characteristics of the 29 included articles is presented in Table 1. These 29 articles described 20 different interventions and included design papers (n=6, 21%), (pilot) randomized controlled trials (n=7, 24%), quasi-experimental studies (n=3, 10%), observational studies (n=4, 14%), protocols (n=2, 7%), a pilot evaluation (n=1, 3%), or a combination of study designs (n=6, 21%). These 20 interventions were delivered via apps (n=8, 40%), SMS text message (n=4, 20%), websites or web applications (n=3, 15%), tablets or computers (n=1, 5%), or a combination of modalities (n=3, 15%). For 5% (1/20) of the interventions, the modality had not been determined yet, but it was described as an app or digital tool [33]. A few eHealth lifestyle interventions (4/20, 20%) were combined with face-to-face or telephone coaching [34-38].

XSL•FO
RenderX

Table 1. Study characteristics (N=29).

Intervention name	eHealth modality used	Targeted health topic	Vulnerability factor and target group	Study and study design
Baby Steps to Healthier Habits or Baby Buddy	App	Diet and physical activity	SEP[a]—from economically and socially disadvantaged communities	• Rhodes et al [45], 2023 (design paper)
Health-e Babies	App	Diet, physical activity, and stress	SEP—from socially disadvantaged communities	• Dalton et al [19], 2018 (observational study)
Healthy Babies	SMS text messaging, web based, application and social media	Diet, physical activity, stress, and sleep	Ethnicity[b]—African American participants	• Herring et al [37], 2019 (design paper and RCT[c])
Healthy Pregnancy: Step by Step	Tablet or computer based program	Diet, stress, and smoking	SEP and ethnicity—underserved pregnant women	• Mauriello et al [46], 2011 (quasi-experimental study) • Prochaska et al [47], 2011 (design paper) • Mauriello et al [15], 2016 (RCT)
I-PREGNO	App	Diet, physical activity, and stress	SEP, educational level, social capital, and age—low SEP or psychosocially burdened[d]	• Vogel et al [38], 2023 (RCT protocol)
MAMA-EMPOWER	App	Diet, physical activity, stress, smoking, and alcohol consumption	Ethnicity—Aboriginal and Torres Strait Islander participants	• Kennedy et al [48], 2021 (design paper, observational study and quasi-experimental study)
momHealth	Tablet and SMS text messaging	Diet, physical activity, and stress	Age—adolescents	• Wambach et al [34], 2021 (quasi-experimental study) • Wambach et al [35], 2022 (pilot RCT)
Mums and Bubs Deadly Diets	App or digital tool	Diet	Ethnicity—Aboriginal and Torres Strait Islander participants	• Gilbert et al [33], 2023 (design paper)
My Healthy Pregnancy	App	Diet, stress, smoking, and alcohol consumption	SEP, ethnicity, and social capital—from a hard-to-reach population	• Krishnamurti et al [49], 2017 (design paper, observational study and quasi-experimental study)
Quit4baby	SMS text messaging	Smoking	SEP—low income	• Abroms et al [50], 2015 (observational study and quasi-experimental study) • Abroms et al [42], 2017 (RCT) • Leavitt et al [51], 2017 (observational study)
S.M.A.S.H. Out Cigarettes	Web based	Smoking	Age—adolescents	• Comer and Grassley [52], 2010 (design paper)
SmartMoms in WIC[e] or Healthy Beginnings	App	Diet, physical activity, stress, and sleep	SEP—low income and receiving WIC benefits	• Flanagan et al [36], 2020 (design paper and RCT protocol)
SmokeBeat	App with wearable	Smoking	SEP—Medicaid eligible	• Joyce et al [18], 2021 (pilot RCT)
Text4baby	SMS text messaging	Diet, physical activity, smoking, and alcohol consumption	SEP—low income	• Whittaker et al [53], 2012 (design paper) • Evans et al [39], 2012 (pilot evaluation) • Remick and Kendrick [54], 2013 (design paper) • Huberty et al [55], 2016 (design paper and observational study) • Huberty et al [43], 2017 (RCT)

XSL•FO
RenderX

Intervention name	eHealth modality used	Targeted health topic	Vulnerability factor and target group	Study and study design
Together with Eva	App	Stress and smoking	SEP—low SEP	• Van Dijk et al [56], 2021 (RCT protocol)
N/A[f]	Web based	Diet and physical activity	SEP—socially disadvantaged area	• Carolan-Olah et al [57], 2021 (design paper and observational study)
N/A	Web based	Smoking	Place of residence and SEP—rural, economically depressed region	• Harris and Reynolds [41], 2015 (pilot RCT)
N/A	SMS text messaging	Diet and physical activity	SEP—low income	• Holmes et al [44], 2020 (RCT)
N/A	SMS text messaging (2 way)	Stress	SEP and ethnicity—low income and minority population	• Song et al [20], 2013 (quasi-experimental study)
N/A	Social media and SMS text messaging	Diet, physical activity, and stress	SEP and age—adolescents and adults with a low income	• Vander Wyst et al [40], 2019 (quasi-experimental study)

[a]SEP: socioeconomic position.

[b]Ethnicity, race, culture, and language.

[c]RCT: randomized controlled trial.

[d]Meeting at least one of the psychosocial burden factors as described in the study by Vogel et al [38].

[e]WIC: women and infant center.

[f]N/A: not applicable or unknown.

Most interventions (14/20, 70%) targeted multiple health behaviors. The most commonly targeted health behavior was diet (14/20, 70%), followed by physical activity (11/20, 55%), stress (11/20, 55%), smoking (9/20, 45%), alcohol consumption (3/20, 15%), and sleep (2/20, 10%). The included interventions addressed various vulnerabilities in pregnant women, and some (6/20, 30%) targeted multiple vulnerabilities. The most commonly targeted vulnerability was socioeconomic status (eg, low income; 15/20, 75%), followed by ethnicity (eg, Aboriginal and Torres Strait Islander participants; 6/20, 30%), age (adolescents; 3/20, 15%), social capital (eg, no partner; 2/20, 10%), place of residence (rural; 1/20, 5%), and educational level (low educational level; 1/20, 5%). The researchers of these interventions did not use low health literacy or limited digital skills as criteria for identifying the target group as vulnerable. Instead, low health literacy occasionally seemed to be considered as a mediator between vulnerability and the adoption of a healthy lifestyle and, therefore, necessitates attention in intervention design. Only 3% (1/29) of the studies reported the health literacy levels of their recruited participants [37].

Of the 29 studies, 10 (34%) investigated the efficacy or effectiveness of the intervention (Multimedia Appendix 3 [15,18-20,33-57]). These studies improved pregnancy-related knowledge [20], motherhood preparedness [39], eating behaviors [15], physical activity [35], and stress [20]. Some studies (3/29, 10%) showed mixed results in improving eating behaviors [40] and smoking cessation [41,42]. In addition, some interventions (5/29, 17%) failed to significantly improve their target group's perceived uncertainty level [20], eating behaviors [35], physical activity [43], gestational weight gain [44], and smoking cessation [18]. To enhance effectiveness, several recommendations were proposed, including initiating the intervention earlier in pregnancy [44], increasing intervention duration [44], and integrating SMS text messaging as part of a multilevel intervention rather than relying exclusively on it [43].

Design and Use

Overview

Table 2 highlights the strategies and barriers identified in the different phases of the eHealth lifestyle interventions. These phases involve design decisions, developing intervention content, and the willingness of individuals to use the intervention. In this section, we elaborate on these strategies to provide a detailed overview of the findings.

XSL•FO
RenderX

Table 2. Strategies and barriers per phase. Symbols are used to indicate whether authors mentioned a strategy as a facilitator (++), possible facilitator (+), possible barrier (–), barrier (– –), or facilitator and possible barrier (++ -). The absence of a symbol means that the strategy was used, but that the authors did not mention the strategy to be a (possible) facilitator or (possible) barrier.

Phase and theme	Example	Studies
Design and use		
Adaptation of or extension to an existing (eHealth) intervention	Adaptations were made to the previous intervention to better meet the needs of economically disadvantaged women [36].	• Flanagan et al [36] • Vogel et al [38]++ • Abroms et al [42]++ • Rhodes et al [45]++ • Mauriello et al [46] • Abroms et al [50]+ • Huberty et al [55]+ • Leavitt et al [51]
Collaborating with stakeholders	Research began by consulting with 4 medical expert informants in the field of maternal-fetal medicine and community informants from a diverse set of groups (eg, churches, nonprofit organizations, women's shelters, and doula groups) [49].	• Mauriello et al [15]++ • Joyce et al [18]++ • Dalton et al [19]++ • Song et al [20]++ • Gilbert et al [33]++ • Wambach et al [34]++ • Flanagan et al [36]++ – • Herring et al [37]++ • Vogel et al [38]++ • Abroms et al [42]++ • Rhodes et al [45]++ • Mauriello et al [46]++ • Prochaska et al [47]++ • Kennedy et al [48]++ • Krishnamurti et al [49]++ • Abroms et al [50]++ • Whittaker et al [53]++ – • Remick and Kendrick [54]++ • Huberty et al [55]++ • Van Dijk et al [56]++ • Carolan-Olah et al [57]++
Financial aspects	Limitations on app development and research were budgetary constraints due to the level of pilot funding [48].	• Mauriello et al [46]– – • Kennedy et al [48]– –
Preventing attrition	To address attrition after enrollment, the recruitment staff was trained to focus on clear, unrushed explanations of the study requirements during the invitation and consent processes [34].	• Mauriello et al [15]++ • Wambach et al [34]+
Providing devices	To assist with adherence, participants were provided with a digital "bathroom" scale for self-weighing [37].	• Joyce et al [18] • Song et al [20] • Wambach et al [34]– – • Wambach et al [35] • Flanagan et al [36] • Herring et al [37]+ • Harris and Reynolds [41] • Huberty et al [43]
Technical problems	Technology issues included limited internet connectivity during teleconference meetings [34].	• Joyce et al [18]– – • Dalton et al [19]– – • Wambach et al [34]– – • Kennedy et al [48]– –
Offering technological support	The study coordinator checked how the smartwatch and app were functioning and helped with any technical issues encountered [18].	• Joyce et al [18]+ • Dalton et al [19]+ • Wambach et al [34,35]+

XSL•FO
RenderX

Phase and theme	Example	Studies
Tailoring	If the app detected a decrease in self-reported cigarette use, it provided encouraging messages in addition to quitting resources [49].	• Mauriello et al [15]+ • Joyce et al [18] • Song et al [20]+ • Gilbert et al [33] • Flanagan et al [36]+ • Herring et al [37]+ • Vander Wyst et al [40] • Holmes et al [44]+ • Mauriello et al [46]+ • Prochaska et al [47] • Kennedy et al [48]+ • Krishnamurti et al [49] • Comer and Grassley [52]+ • Whittaker et al [53]+ • Remick and Kendrick [54]+ • Van Dijk et al [56]+ • Carolan-Olah et al [57]+ –
Using incentives	Incentive payments were earned for consistent smartband wearing and abstaining from smoking [18].	• Mauriello et al [15] • Joyce et al [18] • Flanagan et al [36] • Harris and Reynolds [41]+ • Krishnamurti et al [49]
Using theoretical frameworks	The website was developed using elements of social cognitive theory [57].	• Mauriello et al [15]++ • Gilbert et al [33]++ • Wambach et al [34,35]++ • Herring et al [37]++ • Evans et al [39]++ • Vander Wyst et al [40]++ • Harris and Reynolds [41] • Abroms et al [42]++ • Huberty et al [43] • Holmes et al [44]++ • Rhodes et al [45]++ • Mauriello et al [46]++ • Prochaska et al [47]++ • Kennedy et al [48] • Krishnamurti et al [49]++ • Abroms et al [50]++ • Comer and Grassley [52]++ • Carolan-Olah et al [57]++
Using clinical and best practice guidelines	One of the modules was based on the most recent best practice guidelines on physical activity and exercise during pregnancy [57].	• Harris and Reynolds [41]++ • Abroms et al [42,50]++ • Comer and Grassley [52]++ • Carolan-Olah et al [57]++
Reach		

XSL•FO
RenderX

Phase and theme	Example	Studies
Access to the internet and devices	Women with non-Android mobile phones were excluded due to the app having been developed for Android smartphones only [19].	• Dalton et al [19]– – • Evans et al [39]– – • Harris and Reynolds [41] • Prochaska et al [47]
Collaborating with stakeholders	Limited clinical staff support was experienced for assisting research personnel during recruitment visits [34].	• Mauriello et al [15]++ • Joyce et al [18]++ • Song et al [20]++ • Wambach et al [34]– – • Herring et al [37]++ • Vogel et al [38]+ • Evans et al [39]– – • Harris and Reynolds [41]++ • Mauriello et al [46]++ • Prochaska et al [47]++ • Kennedy et al [48]++ • Whittaker et al [53]++ • Van Dijk et al [56]++
Reluctance to participate	Some potential participants had misgivings about enrolling in a service that involved providing their mobile phone number and other personal information, such as their baby's due date [39].	• Joyce et al [18]– – • Wambach et al [34]– – • Evans et al [39]– –
Limited number of eligible participants	Due to the high rate of preterm labor in this population, recruiting women in their second and third trimesters proved to be challenging [20].	• Mauriello et al [15]– – • Joyce et al [18]– – • Song et al [20]– – • Wambach et al [34]– –
Using incentives	Removal of the reference to the incentive in the recruitment message significantly reduced response and enrollment [51].	• Joyce et al [18] • Song et al [20] • Gilbert et al [33] • Flanagan et al [36] • Herring et al [37] • Vogel et al [38] • Abroms et al [42] • Huberty et al [43] • Rhodes et al [45] • Mauriello et al [46] • Krishnamurti et al [49] • Abroms et al [50] • Leavitt et al [51]++ • Huberty et al [55] • Van Dijk et al [56] • Carolan-Olah et al [57]
Implementation		

XSL•FO
RenderX

Phase and theme	Example	Studies
Collaborating with stakeholders	The involvement of diverse national public, private, and local partners was vitally important for national uptake [53].	• Mauriello et al [15]+ • Flanagan et al [36]+ • Vogel et al [38]+ • Prochaska et al [47]+ • Whittaker et al [53]++
Financial aspects	The funding from and involvement of high-profile national partners made an aggressive timeline to national launch possible [53].	• Harris and Reynolds [41]– • Mauriello et al [46]++ • Whittaker et al [53]++ • Remick and Kendrick [54]++
Integration into health care	Once developed, the intervention was easy to implement in the health care system [56].	• Flanagan et al [36] • Vogel et al [38] • Mauriello et al [46] • Prochaska et al [47] • Krishnamurti et al [49] • Comer and Grassley [52] • Van Dijk et al [56]
Low workload	The intervention did not require much from WIC[a] staff as this is commonly acknowledged as a roadblock for the implementation of long-standing programs [36].	• Flanagan et al [36]++ • Harris and Reynolds [41]++ • Mauriello et al [46]++
Mobile optimization	Mobile optimization allowed the program to be distributed via any internet-enabled device.	• Mauriello et al [15]++

[a]WIC: women and infant center.

Adaptation of or Extension to an Existing (eHealth) Intervention

Some of the eHealth interventions included components of existing interventions or were adaptations of existing interventions. For instance, SmartMoms in WIC [36], Quit4baby [42], I-PREGNO [38], and Healthy Pregnancy: Step by Step [46] made adaptations to existing interventions to better meet the specific needs of their vulnerable target groups. They made these adaptations, such as creating a sense of community, by including support groups through Facebook based on recommendations provided by stakeholders [36,46]. In addition, Rhodes et al [45] and Huberty et al [55] developed their interventions as an extension to existing eHealth interventions or modalities and referred to existing eHealth modalities as well suited for the implementation and evaluation of novel eHealth lifestyle interventions.

Theoretical Frameworks and Guidelines

Over half (17/29, 59%) of the studies used theoretical frameworks for behavior change, which contributed substantially to the design of their interventions. The frameworks used were social cognitive theory [37,39,40,42-44,50,52,57], the transtheoretical model of behavior change [15,39,41,46,47], the multiple health behavior change paradigm [34,35], the Behavior Change Wheel [48], the Fogg behavior model [45], and the health belief model [39]. Most studies (24/29, 83%) did not provide justifications for their selection of a particular behavior change framework. However, Kennedy et al [48] and Mauriello et al [15] selected their frameworks based on their effectiveness in previous studies. In addition, the Behavior Change Wheel was used for its ability to advance understanding of features in need of improvement [48]. The multiple health behavior change paradigm was chosen as a guiding framework due to its unique approach in addressing multiple health behaviors simultaneously through one intervention [34,35]. In addition, the researchers used clinical and best practice guidelines to inform the content of their interventions [41,42,50,52,57]. Furthermore, theoretical frameworks and guidelines were used to guide the development of the eHealth tools, including the Centre for eHealth Research Roadmap [48], the behavioral decision research paradigm [49], the Behavior Change Wheel [45], the Sanders and Stappers co-design framework [33], the Kaupapa Māori framework [33], and Noorbergen's guidelines for co-design of mobile health (mHealth) systems [33].

Collaborating With Stakeholders

Stakeholders provided valuable guidance to researchers in terms of the design and content of the eHealth lifestyle interventions. Stakeholders included, among others, the target group, research centers, academics specialized in different health domains, social service providers, literacy experts, and mHealth companies. For co-creation with the target group, the researchers used multiple methods, such as surveys [20,55], interviews [18,37,45,47-49,55,57], focus groups [19,37,38,45-47,53,56], user research [45-49], and monthly meetings [36], to ensure that their target group's needs were met in terms of design, literacy, content, and usability of the eHealth lifestyle intervention. For instance, findings from interviews with the target group can indicate barriers that women encounter related to healthy lifestyle practices during pregnancy [57]. In turn, this knowledge can inform the design of an eHealth lifestyle intervention that overcomes these barriers.

However, collaborating with stakeholders can also entail challenges. For example, Flanagan et al [36] experienced a disparity between their mothers advisory group's request to

XSL•FO
RenderX

include more health markers in their trial and the limited enthusiasm for this adaptation from the scientific review panel. Similarly, Whittaker et al [53] encountered challenges in information sharing and estimating the in-kind costs of their initiative. They also faced confusion regarding defined roles and responsibilities among their partners and differing perspectives and priorities in transitioning to the next stage of their partnerships. Establishing and adhering to a set of guiding principles, comprising key intervention design objectives and features, could ensure that all stakeholders involved in the intervention design work toward a common vision [45].

Tailoring

Overview

Many studies (17/29, 59%) tailored their eHealth lifestyle intervention to the characteristics and skills of their target group, for example, by matching the content of their intervention to their target groups' norms and values or literacy level. Some interventions (9/29, 31%) provided individualized tailored information or feedback based on gestation [53], collected participant data (eg, body weight measurements) [18,36,46,49], stage of change [15,37,46], or the Behavior Change Wheel [48]. In some studies (2/29, 7%), the researchers mentioned tailoring of tips, recipes, and feedback without providing details about what the tailoring was based on [37,47].

Language and Culture

Certain researchers made sure that their intervention was available or would become available in multiple languages [15,46,53,54], whereas others paid special attention to the level of (health) literacy of their target group [18,20,36,40,44,46,52,54,56,57]. The importance of health literacy and language availability in eHealth interventions was highlighted by several studies. For instance, Carolan-Olah et al [57] used multiple photos and illustrations, limited textual content, and maintained a single idea per slide in their intervention to improve access for women with low levels of health literacy. However, some of the less educated women found that there was still too much information provided in the modules. In addition, their intervention was solely available in English, and the women expressed a preference for it to be available in other languages as well. In addition, in their intervention, Evans et al [39] found that educational level was an important factor for health belief outcomes and suggested that this could be a result of differences in literacy levels and message comprehension.

Culture was incorporated into the content and design of the interventions to increase its relevance to or acceptability by the target group [15,36,46,52], align their messages with the norms and values of the target group [33,36,54], and cater to the preferences expressed by the target group [48]. However, most of these studies (4/7, 57%) did not describe in detail how culture was incorporated into their intervention. Nevertheless, in some interventions (3/7, 43%), researchers included images of women in the same age categories and from the same ethnic backgrounds as their target group [46,48,52]. Furthermore, recipes and links were provided from specific cuisines to align with the high proportion of particular ethnic groups in the area in which recruitment took place [57].

Providing Devices

Researchers provided devices to their participants for various purposes. Some researchers provided devices necessary for the delivery of the intervention [18,34,35], whereas others provided devices to collect data for evaluating the intervention's effectiveness. The latter included a Fitbit to measure physical activity [36,43], a pedometer to track step counts [37], a piCO Smokerlyzer to measure breath carbon monoxide [41], and a digital scale to monitor gestational weight gain [27,37]. In addition to evaluating effectiveness, Herring et al [37] highlighted that providing devices might facilitate adherence to the intervention. However, personal reasons can influence the use of devices, such as not being able to wear a smartwatch during work [18].

Technical Problems and Offering Support

Technical problems were expressed by participants in several studies (4/29, 14%) [18,19,34,48], often negatively influencing engagement. Multiple studies (4/29, 14%) provided technological support to prevent or help with technological issues [18,19,34,35]. For example, information was provided on how to use the device, download the app, and synchronize devices, and contact information for study coordinators was provided to help when the participant encountered technical issues. However, offering contact details for technological support alone may not be enough. For example, Dalton et al [19] provided phone numbers for technological support but found that 9% of the participants failed to report the problems they encountered while downloading the app. Leaving participants to manage the app on their own to assess its usability might have influenced the high dropout rate in their study [19].

Preventing Attrition

In a few studies (4/29, 14%), the researchers mentioned attrition of participants after enrollment [18,19,34,35]. Reasons that might have contributed to this attrition were participants' employment status [19], financial constraints [19], anxiety levels [19], and time constraints [18,19,34]. To retain participants, Mauriello et al [15] had participants engage with the intervention during their appointments, contacted participants who could not be reached in various ways to make them complete final assessments, and used incentives for completing a session. These strategies resulted in an impressive retention. The attrition in the study by Wambach et al [34] prompted the development of strategies to limit attrition in future research. These included enhancing staff training to focus on a clear and unrushed explanation of study requirements during recruitment, improving the description of study requirements on an advertising flyer, and including larger incentives. This resulted in a lower attrition rate, although it remained high at 25.8% [35].

Features Used in eHealth Interventions

The features (components that make up the eHealth intervention) included in the eHealth interventions are summarized in Table 3. Many interventions (6/20, 30%) included a feature in which participants could interact with others in their intervention, such as a chat room or Facebook page [19,34-37,48,52]. However,

XSL•FO
RenderX

in one intervention, peer support was considered the least useful component by most of the participants [34]. Some interventions (2/20, 10%) actively included partners as a way of providing social support to pregnant women [38,45]. In addition, links to external sources of information were provided, which prevented the inclusion of too much information in the content of the intervention but also catered to those who were looking for more information about a certain topic [19,34,35,40,41,43,45,52,55,57]. Furthermore, although some interventions included weight trackers, Rhodes et al [45] decided not to include weight monitoring or weight-related messages based on their potential to demotivate their participants.

Some of the incorporated features used were specific to a limited number of interventions. For example, Krishnamurti et al [49] included Uber services in their intervention to provide free transportation to prenatal care appointments as transportation was revealed to be a barrier for their target group. Providing Uber transportation prevented missed appointments and was found to be cost saving. Song et al [20] were the only ones to include an automated, 2-way SMS text messaging system in their intervention to distribute pregnancy and health-related information and foster patient–health care provider interaction. Despite some frustrations regarding its ability to answer participants' questions, the SMS text messaging system could promote health communication while offering psychological benefits as well [20]. However, it was suggested that the addition of more system-initiated SMS text messages could benefit women who are less comfortable with asking questions [20].

XSL•FO
RenderX

Table 3. Features of the eHealth lifestyle interventions.

Features	Studies
Calculator	• Mauriello et al [15] • Kennedy et al [48] • Carolan-Olah et al [57]
Diary	• Vogel et al [38] • Van Dijk et al [56]
Feedback	• Mauriello et al [15] • Flanagan et al [36] • Joyce et al [18] • Rhodes et al [45] • Mauriello et al [46] • Kennedy et al [48] • Krishnamurti et al [49] • Comer and Grassley [52]
Food serving size measurements	• Carolan-Olah et al [57]
Fictitious peer offering advice	• Abroms et al [42]
Game or quiz component	• Wambach et al [35] • Flanagan et al [36] • Herring et al [37] • Abroms et al [42] • Carolan-Olah et al [57]
Goal setting	• Mauriello et al [15] • Herring et al [37] • Vogel et al [38] • Abroms et al [42] • Rhodes et al [45] • Kennedy et al [48] • Carolan-Olah et al [57]
Motivation from peers	• Abroms et al [42]
Multimedia	• Wambach et al [34,35] • Flanagan et al [36] • Vogel et al [38] • Vander Wyst et al [40] • Rhodes et al [45] • Mauriello et al [46] • Prochaska et al [47] • Kennedy et al [48] • Comer and Grassley [52] • Carolan-Olah et al [57]
Links to external sources of information	• Dalton et al [19] • Wambach et al [34,35] • Vander Wyst et al [40] • Harris and Reynolds [41] • Huberty et al [43] • Rhodes et al [45] • Comer and Grassley [52] • Huberty et al [55] • Carolan-Olah et al [57]
Messages of support	• Mauriello et al [15]
Pregnancy-tracking features	• Dalton et al [19] • Kennedy et al [48] • Krishnamurti et al [49]
Real-time alerts to medical staff	• Krishnamurti et al [49]

XSL•FO
RenderX

Features	Studies
Recipes	• Mauriello et al [15] • Flanagan et al [36] • Herring et al [37] • Vander Wyst et al [40] • Harris and Reynolds [41] • Rhodes et al [45] • Kennedy et al [48] • Carolan-Olah et al [57]
Reminders	• Dalton et al [19] • Abroms et al [42] • Rhodes et al [45] • Krishnamurti et al [49] • Remick and Kendrick [54]
Self-assessment and self-monitoring	• Flanagan et al [36] • Herring et al [37] • Vogel et al [38] • Rhodes et al [45] • Kennedy et al [48] • Krishnamurti et al [49]
Social component	• Wambach et al [34,35] • Flanagan et al [36] • Herring et al [37] • Vogel et al [38] • Rhodes et al [45] • Dalton et al [19] • Kennedy et al [48] • Comer and Grassley [52]
Stress-reducing exercises	• Dalton et al [19] • Vogel et al [38] • Mauriello et al [46] • Kennedy et al [48] • Van Dijk et al [56]
Free transportation	• Krishnamurti et al [49]
Two-way SMS text messaging system	• Song et al [20]
Weight tracker	• Flanagan et al [36]

Reach

Table 2 highlights the strategies and encountered barriers identified in the reach phase, which involves recruitment methods and the willingness of individuals to participate in the intervention. In this section, we elaborate on the strategies found and barriers encountered.

Recruitment Characteristics

In the included studies, the researchers used multiple sites and strategies and involved various key persons to recruit vulnerable pregnant women (Table 4). Often, recruitment took place in a health care setting (19/29, 66%), and many studies (12/29, 41%) received support from health professionals for recruitment. Apart from face-to-face or phone recruitment (22/29, 76%), many studies used printed materials (11/29, 38%) or the internet (5/29, 17%) to promote their interventions. Printed materials were placed at sites frequently visited by pregnant women, such as schools and children's retail stores.

XSL•FO
RenderX

Table 4. Recruitment characteristics.

Recruitment characteristic	Studies
Individuals involved	
Health care professionals	• Mauriello et al [15] • Herring et al [37] • Vogel et al [38] • Song et al [20] • Evans et al [39] • Harris and Reynolds [41] • Mauriello et al [46] • Prochaska et al [47] • Comer and Grassley [52] • Whittaker et al [53] • Van Dijk et al [56]
Researchers	• Gilbert et al [33] • Wambach et al [34] • Flanagan et al [36] • Herring et al [37] • Holmes et al [44] • Kennedy et al [48] • Whittaker et al [53] • Huberty et al [55] • Carolan-Olah et al [57]
Staff from non–health care organizations	• Joyce et al [18] • Kennedy et al [48] • Whittaker et al [53]
Professional recruitment services	• Rhodes et al [45]
Recruitment sites	
Health care setting	• Mauriello et al [15] • Joyce et al [18] • Dalton et al [19] • Gilbert et al [33] • Wambach et al [34,35] • Herring et al [37] • Vogel et al [38] • Evans et al [39] • Vander Wyst et al [40] • Harris and Reynolds [41] • Huberty et al [43] • Mauriello et al [46] • Prochaska et al [47] • Kennedy et al [48] • Krishnamurti et al [49] • Huberty et al [55] • Van Dijk et al [56] • Carolan-Olah et al [57]
Federal benefit and assistance clinics for low-income women and families	• Flanagan et al [36] • Herring et al [37] • Holmes et al [44]
Public places (schools, family support centers, churches, retail stores, and events)	• Wambach et al [35] • Huberty et al [43] • Krishnamurti et al [49]
Homes of the target group	• Song et al [20]
Community organizations	• Gilbert et al [33] • Kennedy et al [48]
Recruitment strategies	

XSL•FO
RenderX

Recruitment characteristic	Studies
Personal contact or phone call	• Mauriello et al [15] • Joyce et al [18] • Dalton et al [19] • Gilbert et al [33] • Wambach et al [34,35] • Flanagan et al [36] • Herring et al [37] • Vogel et al [38] • Evans et al [39] • Vander Wyst et al [40] • Harris and Reynolds [41] • Holmes et al [44] • Mauriello et al [46] • Prochaska et al [47] • Krishnamurti et al [49] • Whittaker et al [53] • Huberty et al [55] • Van Dijk et al [56] • Carolan-Olah et al [57]
Online (social media, websites, or discussion boards)	• Huberty et al [43] • Rhodes et al [45] • Whittaker et al [53] • Huberty et al [55] • Van Dijk et al [56]
Printed materials (flyers, posters, and brochures)	• Gilbert et al [33] • Wambach et al [34,35] • Flanagan et al [36] • Vander Wyst et al [40] • Harris and Reynolds [41] • Huberty et al [43] • Kennedy et al [48] • Whittaker et al [53] • Huberty et al [55] • Van Dijk et al [56]
Email (listserve) or SMS text messaging	• Huberty et al [43,55]
Word-of-mouth or grassroots strategies	• Wambach et al [34] • Harris and Reynolds [41] • Huberty et al [43] • Kennedy et al [48] • Huberty et al [55]
Existing eHealth interventions	• Abroms et al [42] • Rhodes et al [45] • Leavitt et al [51]
A clinical study or pregnancy support program	• Joyce et al [18] • Vander Wyst et al [40]
Home visits	• Song et al [20]
Personal network	• Kennedy et al [48]
Survey	• Rhodes et al [45]

Recruitment Challenges and Strategies

Challenges related to participant recruitment were frequently encountered. One of these challenges was a limited number of eligible participants [15,18,20,34]. For example, Song et al [20] experienced challenges in recruiting pregnant women in their second and third trimesters due to a high rate of preterm labor within the low-income minority pregnant population. A second challenge in recruitment was due to eligible candidates' reluctance to participate because of a lack of interest, time constraints, and concerns about sharing personal information [18,34,39]. Health care professionals and staff from non–health

XSL•FO
RenderX

care organizations were often mentioned as facilitators in recruiting the target group, either providing assistance or taking full responsibility for the recruitment process [15,18,20,37,41,46-48,53,56], but sometimes this posed challenges as well [34,39]. For example, research personnel in the study by Wambach et al [34] received limited support from clinical staff during recruitment visits. Finally, resource aspects, including limited phone ownership [39], lack of access to an internet service [19,47], and phone operating systems that were incompatible with the eHealth intervention [19], hindered recruitment. In some studies (6/29, 21%), recruitment difficulties resulted in a small sample size [20,34,39-41,48].

Certain strategies were implemented to overcome recruitment challenges. For example, in response to the limited number of pregnant adolescents, Wambach et al [34] expanded their age range and the number of recruitment sites and included word-of-mouth recruitment. Second, in the study by Harris and Reynolds [41], research personnel conducted home visits to set up equipment and provide detailed training on its use, thereby enhancing accessibility for rural pregnant smokers. Furthermore, to overcome participants' limited access to an internet or telephone service, Prochaska et al [47] encouraged providers of the intervention to have a computer kiosk at their centers for women to access the program.

Incentives

Many studies (18/29, 62%) offered incentives to encourage participation in their trial, interview, focus group, or workshop [15,20,33,43,45,46,51,55,57]; attend study visits or video check-ins [18,36]; complete assessments, questionnaires, or interviews [20,38,49,50,56]; submit self-monitoring data [37] or bodily samples [42]; engage with the intervention [15,18,36,49]; or abstain from unhealthy behaviors [18,41]. The incentives provided in these studies included gift cards [15,20,33,36,37,41,42,46,50,57]; monetary compensation [38,41,45,49,55]; devices such as smartphones [20,49], scales [49], and smartwatches [43]; and health-related items such as yoga mats and prenatal vitamins [36].

Few studies (2/18, 11%) discussed the impact of the incentives they used. Nevertheless, Leavitt et al [51] observed a significant decline in the response rate after removing the reference to their incentive in their recruitment message. Furthermore, Harris and Reynolds [41] believed that their incentives enhanced participants' motivation to quit smoking by the specified quit date. However, in the study by Joyce et al [18], qualitative interviews revealed mixed feelings toward financial incentives, where one participant stated that the financial incentive was encouraging to quit smoking, whereas another participant expressed that a motivational tailored message would be rewarding enough without financial rewards.

Implementation

Overview

Assessment of implementation includes factors such as the successfulness and costs of intervention delivery. Most studies (18/29, 62%) did not describe how their eHealth lifestyle intervention was or would be implemented. However, some strategies in the implementation phase of the eHealth lifestyle interventions were identified, which are highlighted in Table 2.

Financial Aspects

A limited number of studies (5/29, 17%) described whether costs and funding acted as barriers or facilitators in the design or implementation phase of their eHealth lifestyle intervention. Nevertheless, in 7% (2/29) of the studies, budget constraints limited intervention development [46,48]. Furthermore, costs from financial incentives and loaned devices could be barriers to intervention implementation [41]. Implementing the intervention in a clinical setting was seen as a way to reduce costs [41]. Whittaker et al [53] mentioned how funding from and involvement of well-known national partners facilitated the national launch of their intervention. In addition, providing the intervention free of charge to vulnerable pregnant women could facilitate implementation [53,54].

Dissemination and Integration Into Health Care Structures

A few strategies were highlighted in the studies to ensure the dissemination of their interventions. First, collaboration with stakeholders was frequently emphasized as vital for the successful implementation of the interventions. For example, the adoption of an intervention by a governmental program or by existing perinatal care services can facilitate widespread implementation and distribution [36,38]. The enthusiasm of stakeholders and the intervention's relevance to them were mentioned as characteristics that should contribute to the ease of the dissemination of their interventions [15,47]. Second, a low workload for intervention deliverers was mentioned in multiple studies (3/29, 10%) as an important aspect of a feasible implementation [36,41,46]. Third, providing the intervention in a format that can be distributed via any internet-enabled device, either in health care settings or via a personal internet-enabled device at the convenience of its user, contributes to dissemination [15]. Furthermore, media appearance can result in a considerable increase in rates of enrollment in the intervention [53]. Finally, to develop an intervention that could be easily integrated into health care structures, dissemination issues should be considered from the start of the project [47].

A few research groups developed their interventions with the aim of integrating them into the current health care structures [36,38,46,47,49,52,56]. Several suggestions were given for this integration, such as women using the intervention before their consult and sharing a printed report of their results with their health care provider [47]. For this integration, it was important that the program be self-directed, require little to no staff training, and allow for low-cost and consistent delivery [46]. However, taking health care providers' time and resource constraints into consideration, an intervention can be intentionally developed to be used outside of health care as well [37].

XSL•FO
RenderX

Discussion

Principal Findings

Overview

This scoping review provides a comprehensive overview of the applied strategies in 4 phases (design, reach, use, and implementation) of eHealth lifestyle interventions for vulnerable pregnant women. In addition, it highlights which barriers researchers encountered and which strategies acted as facilitators for these interventions. By identifying barriers and facilitators in current eHealth lifestyle interventions aimed at vulnerable pregnant women, our study generated insights into how to optimize eHealth lifestyle interventions for this population. As this scoping review covered a wide range of interventions, study designs, targeted health behaviors, and vulnerabilities, it also contributes to a nuanced understanding of the landscape of available eHealth lifestyle interventions for vulnerable pregnant women.

We included 29 articles describing 20 eHealth interventions in this review. These interventions were delivered through different modalities, targeted different lifestyle components, and were aimed at pregnant women with different vulnerabilities. The studies examining the effectiveness of the eHealth lifestyle interventions showed potential as some managed to significantly change health behaviors, but the results were inconsistent. This finding aligns with those of previous research on eHealth lifestyle interventions for the low-SEP population [58], showing that effect sizes are small and differ among the interventions. This enhances the need for insights into the experienced barriers and facilitators in the different phases of the development of an eHealth intervention. The insights gained from our review will be explored separately per phase.

Intervention Design and Use

For the design of eHealth lifestyle interventions, researchers were guided by stakeholders, existing health interventions, guidelines, and theoretical frameworks. Researchers that collaborated with their target group and other stakeholders (eg, health care professionals) often referred to these collaborations as facilitating to their intervention design [15,18-20,33,34,36-38,42,45]. While co-creating with the target group was mostly limited to content design of the intervention, other stakeholders were often involved in multiple phases of intervention development. Engaging in qualitative research with intended users can shape the foundational guiding principles of the intervention, and user feedback can help refine the content and functionality of the intervention [45]. Existing health interventions were examined to identify components that could be valuable for the design of new eHealth interventions [36,38,42,45,46,50,55,56]. For clinical content, researchers additionally consulted guidelines from organizations such as the American Congress of Obstetricians and Gynecologists [41]. In addition, various theoretical frameworks for behavior change were used to guide the content of the interventions. However, it often remained unclear why researchers chose the frameworks they used. Furthermore, even though frameworks could help guide health intervention design, evaluation, adaptation, and implementation, only some studies (4/29, 59%) mentioned using a framework to guide the development of their eHealth tool.

Various features and strategies were included in the eHealth lifestyle interventions to increase user engagement. First, including social components (eg, virtual peer support sessions and partner involvement) in the intervention was (expected to be) appreciated by end users [19,34-38,45,48,52]. Research indicates that pregnant women value partner involvement and support in eHealth lifestyle interventions and its benefits could extend beyond improving health outcomes [59]. Nevertheless, this review revealed a limited number of interventions (2/20, 10%) that actively involved partners to support maternal health behavior. Future eHealth lifestyle interventions should explore ways of engaging partners of vulnerable pregnant women as their involvement seems promising. Furthermore, although it was somewhat unclear whether tailoring led to better results in the included studies, tailoring the intervention to the characteristics and needs of the target group or individuals is expected to lead to increased engagement [60]. The same expectation applies to the use of incentives [61]. Technological support was provided to assist end users with any technological difficulty encountered. Although positive assumptions about pregnant women's digital literacy due to their age may be made, one study highlighted an instance in which technical challenges adversely affected user engagement [19]. In addition, digital literacy has previously been identified as a barrier to mHealth adoption among people of a low socioeconomic status [62]. These results highlight the importance of addressing digital literacy when developing eHealth interventions for vulnerable pregnant women.

Reaching Vulnerable Pregnant Women

Recruitment of vulnerable groups for study purposes frequently presents challenges, as has been emphasized in a previous review [58] and was once more highlighted in this review, where most researchers encountered difficulties in the recruitment of vulnerable pregnant women for participation in their studies to test their eHealth lifestyle interventions. In a few studies (6/29, 21%), difficulties with recruitment resulted in a small sample size [20,34,39-41,48]. However, the researchers did not always explicitly discuss the barriers that kept them from achieving an adequate sample size. Interestingly, our findings show that recruitment barriers were not solely attributed to factors such as low patient volumes or lack of interest by participants. Health care providers, although often identified as facilitators in the recruitment process, were also identified as barriers, either because the research staff received limited support from health care providers or because recruitment was not feasible for health care providers operating in their natural setting [34,39]. Finally, resource aspects (eg, lack of access to internet and limited phone ownership) were mentioned as a barrier to the recruitment of vulnerable pregnant women [19,39,47].

The study by Mauriello et al [46] was the only one that not only met but also exceeded its recruitment goals within a short time frame, and they attributed this success to the willingness of the prenatal care staff and eagerness of the pregnant women attending the health center where they recruited from. However, it remains unclear what exactly caused this eagerness. Many

XSL•FO
RenderX

studies (11/29, 38%) involved health care professionals in the recruitment process, whereas community-led recruitment was less prevalent. However, using a community-based participatory research approach has shown to be a promising strategy for conducting health disparity–related research in minority populations [63] and, therefore, might be considered in the design of future interventions targeting vulnerable pregnant women. In addition, including incentives could facilitate recruitment, although we only found one study that clearly showed that their reference to an incentive increased recruitment numbers [42,51].

Overall, to increase the successful recruitment of vulnerable pregnant women, researchers should consider all 4 conditions that were identified as barriers to recruitment in this review. First, it is important to identify the prevalence of the target group, and recruitment sites should be adjusted accordingly based on this prevalence. Second, researchers ought to incorporate methods to ensure that participation in eHealth interventions is appealing and minimally time-consuming for their target group. Including incentives could be particularly helpful in addressing this concern. Third, researchers should aim to either handle recruitment themselves or involve dedicated health care providers in the recruitment process, ensuring it does not become resource intensive or time-consuming for them or disrupt the delivery of health care. Finally, resource-related barriers to recruitment can be handled by enabling access to internet services during recruitment and providing the devices necessary for the use of the eHealth lifestyle intervention.

Implementation

The insights from the facilitators and barriers encountered in previous eHealth implementation processes could help future researchers, health care professionals, and eHealth developers devise more effective strategies for forthcoming implementations. Nevertheless, most studies (18/29, 62%) failed to outline the implementation process of their eHealth lifestyle intervention. This could be attributed to the prevalence of pilot and design papers among the included studies, where the interventions had often not been implemented beyond these initial stages. However, a few facilitators could be identified related to implementation. First, collaboration with stakeholders was frequently emphasized as necessary for a successful implementation of the intervention [15,36,38,47,53]. Collaborating with stakeholders enhances the likelihood that interventions are designed and implemented in a manner that is relevant, acceptable, and feasible within real-world settings [64,65]. In addition, providing the intervention free of charge was important for uptake by end users. Furthermore, as eHealth lifestyle interventions often include the involvement of health care professionals, for whom time constraints can be a barrier to involvement, a low workload came forward as an important facilitator to eHealth implementation. In contrast, the costs associated with financial incentives and loaned devices were found to be a potential barrier to implementation.

Limitations

This scoping review is the first to identify strategies in the development of eHealth lifestyle interventions for vulnerable pregnant women. The barriers and facilitators that were identified can guide researchers, health care professionals, and eHealth developers in the development of future eHealth tools for this target group. However, this review has some limitations. Although certain strategies emerged as clear facilitators, the authors did not consistently report on the impact of other applied strategies. This lack of reporting on facilitators and barriers within specific interventions was also noted in an earlier review about eHealth lifestyle interventions in the low-SEP population [58]. Identifying strategies and obstacles encountered as possible barriers and facilitators allowed us to partly solve this limitation. Guidelines that ensure complete and accurate documentation of eHealth development and implementation, such as the Guidelines and Checklist for the Reporting on Digital Health Implementation, foster the transparency necessary for future developers and, therefore, should be used when reporting on eHealth development and implementation [66]. In addition, facilitators were mentioned more often than barriers in the articles, which might be related to publication bias, in which articles about interventions that failed are not written or published. Furthermore, the heterogeneity of the lifestyle behaviors and eHealth modalities in the included studies might limit the generalizability of the barriers and facilitators to other lifestyle behaviors and eHealth modalities.

Conclusions

This scoping review provides a comprehensive overview of the strategies used and the challenges faced in developing and implementing eHealth lifestyle interventions for vulnerable pregnant women throughout different phases of development. Specifically, our findings in the design and use phases of eHealth lifestyle interventions highlight the importance of stakeholder engagement, a user-centered design, theoretical frameworks, tailoring to the needs and skills of the target group, and providing technological support. Furthermore, the challenges and strategies related to recruitment underscore the complexities involved in engaging vulnerable populations in research and interventions. By identifying barriers such as limited access to resources and health care provider support, this scoping review offers practical recommendations for improving the reach of vulnerable pregnant women. The insights into implementation facilitators and barriers highlight the importance of stakeholder collaboration, a low workload for intervention deliverers, and financial considerations. These findings provide valuable guidance for researchers, health care professionals, and eHealth developers seeking to implement eHealth interventions effectively within health care settings and broader community contexts.

Despite the considerable insights derived from this review, more detailed reporting on the impact of the strategies used and barriers encountered in eHealth lifestyle interventions for vulnerable pregnant women is warranted. Nevertheless, our insights will pave the way for the development of more impactful eHealth interventions for vulnerable pregnant women, ultimately enhancing the health of both mothers and their offspring.

XSL•FO
RenderX

Acknowledgments

The authors wish to thank W Bramer from the Erasmus Medical Center Medical Library for developing and updating the search strategies. This publication is part of the Our Smart Family Buddy project (KICH1.GZ01.20.016) of the research program "Knowledge and Innovation Convenant Lifestyle and Living Environment, which is financed by the Dutch Research Council.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 549 KB - jmir_v26i1e54366_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 18 KB - jmir_v26i1e54366_app2.docx]

Multimedia Appendix 3
Study characteristics.
[XLSX File (Microsoft Excel File), 18 KB - jmir_v26i1e54366_app3.xlsx]

References

1. Smit AJ, Hojeij B, Rousian M, Schoenmakers S, Willemsen SP, Steegers-Theunissen RP, et al. A high periconceptional maternal ultra-processed food consumption impairs embryonic growth: the Rotterdam periconceptional cohort. Clin Nutr 2022 Aug;41(8):1667-1675 [FREE Full text] [doi: 10.1016/j.clnu.2022.06.006] [Medline: 35772220]
2. Parisi F, Rousian M, Huijgen NA, Koning AH, Willemsen SP, de Vries JH, et al. Periconceptional maternal 'high fish and olive oil, low meat' dietary pattern is associated with increased embryonic growth: the Rotterdam Periconceptional Cohort (Predict) study. Ultrasound Obstet Gynecol 2017 Dec;50(6):709-716 [FREE Full text] [doi: 10.1002/uog.17408] [Medline: 28078758]
3. Davenport MH, Ruchat S, Poitras VJ, Jaramillo Garcia A, Gray CE, Barrowman N, et al. Prenatal exercise for the prevention of gestational diabetes mellitus and hypertensive disorders of pregnancy: a systematic review and meta-analysis. Br J Sports Med 2018 Nov;52(21):1367-1375. [doi: 10.1136/bjsports-2018-099355] [Medline: 30337463]
4. Avşar TS, McLeod H, Jackson L. Health outcomes of smoking during pregnancy and the postpartum period: an umbrella review. BMC Pregnancy Childbirth 2021 Mar 26;21(1):254 [FREE Full text] [doi: 10.1186/s12884-021-03729-1] [Medline: 33771100]
5. Broccia M, Hansen BM, Winckler JM, Larsen T, Strandberg-Larsen K, Torp-Pedersen C, et al. Heavy prenatal alcohol exposure and obstetric and birth outcomes: a Danish nationwide cohort study from 1996 to 2018. Lancet Public Health 2023 Jan;8(1):e28-e35 [FREE Full text] [doi: 10.1016/S2468-2667(22)00263-8] [Medline: 36603908]
6. Broccia M, Munch A, Hansen BM, Sørensen KK, Larsen T, Strandberg-Larsen K, et al. Heavy prenatal alcohol exposure and overall morbidities: a Danish nationwide cohort study from 1996 to 2018. Lancet Public Health 2023 Jan;8(1):e36-e46 [FREE Full text] [doi: 10.1016/S2468-2667(22)00289-4] [Medline: 36603909]
7. Staneva A, Bogossian F, Pritchard M, Wittkowski A. The effects of maternal depression, anxiety, and perceived stress during pregnancy on preterm birth: a systematic review. Women Birth 2015 Sep;28(3):179-193. [doi: 10.1016/j.wombi.2015.02.003] [Medline: 25765470]
8. Badon SE, Enquobahrie DA, Wartko PD, Miller RS, Qiu C, Gelaye B, et al. Healthy lifestyle during early pregnancy and risk of gestational diabetes mellitus. Am J Epidemiol 2017 Aug 01;186(3):326-333 [FREE Full text] [doi: 10.1093/aje/kwx095] [Medline: 28481961]
9. van Zundert SK, van Rossem L, Willemsen SP, van der Meer L, Ernst-Smelt HE, Steegers-Theunissen RP. Periconceptional maternal social, lifestyle and medical risk factors impair embryonic growth: the Rotterdam Periconceptional Cohort. Reprod Biomed Online 2022 Jun;44(6):1123-1133 [FREE Full text] [doi: 10.1016/j.rbmo.2022.02.011] [Medline: 35361546]
10. Zhao G, Ford ES, Tsai J, Li C, Ahluwalia IB, Pearson WS, et al. Trends in health-related behavioral risk factors among pregnant women in the United States: 2001-2009. J Womens Health (Larchmt) 2012 Mar;21(3):255-263. [doi: 10.1089/jwh.2011.2931] [Medline: 22047097]
11. Moreno-Juste A, Gimeno-Miguel A, Poblador-Plou B, Calderón-Larrañaga A, Cano Del Pozo M, Forjaz MJ, et al. Multimorbidity, social determinants and intersectionality in chronic patients. Results from the EpiChron Cohort. J Glob Health 2023 Feb 03;13:04014 [FREE Full text] [doi: 10.7189/13.04014] [Medline: 36757132]

XSL•FO
RenderX

12. de Diego-Cordero R, Rivilla-Garcia E, Diaz-Jimenez D, Lucchetti G, Badanta B. The role of cultural beliefs on eating patterns and food practices among pregnant women: a systematic review. Nutr Rev 2021 Aug 09;79(9):945-963. [doi: 10.1093/nutrit/nuaa119] [Medline: 33119079]
13. Kapur A. Pregnancy: a window of opportunity for improving current and future health. Int J Gynaecol Obstet 2011 Nov;115 Suppl 1:S50-S51. [doi: 10.1016/S0020-7292(11)60015-5] [Medline: 22099444]
14. Wallwiener S, Müller M, Doster A, Laserer W, Reck C, Pauluschke-Fröhlich J, et al. Pregnancy eHealth and mHealth: user proportions and characteristics of pregnant women using web-based information sources-a cross-sectional study. Arch Gynecol Obstet 2016 Nov;294(5):937-944. [doi: 10.1007/s00404-016-4093-y] [Medline: 27084763]
15. Mauriello LM, Van Marter DF, Umanzor CD, Castle PH, de Aguiar EL. Using mHealth to deliver behavior change interventions within prenatal care at community health centers. Am J Health Promot 2016 Sep;30(7):554-562. [doi: 10.4278/ajhp.140530-QUAN-248] [Medline: 26305603]
16. Households - level of internet access. Eurostat. URL: https://tinyurl.com/57xw993f [accessed 2001-05-23]
17. Lustria ML, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
18. Joyce CM, Saulsgiver K, Mohanty S, Bachireddy C, Molfetta C, Steffy M, et al. Remote patient monitoring and incentives to support smoking cessation among pregnant and postpartum Medicaid members: three randomized controlled pilot studies. JMIR Form Res 2021 Sep 30;5(9):e27801 [FREE Full text] [doi: 10.2196/27801] [Medline: 34591023]
19. Dalton JA, Rodger D, Wilmore M, Humphreys S, Skuse A, Roberts CT, et al. The Health-e Babies App for antenatal education: feasibility for socially disadvantaged women. PLoS One 2018 May 16;13(5):e0194337 [FREE Full text] [doi: 10.1371/journal.pone.0194337] [Medline: 29768407]
20. Song H, May A, Vaidhyanathan V, Cramer EM, Owais RW, McRoy S. A two-way text-messaging system answering health questions for low-income pregnant women. Patient Educ Couns 2013 Aug;92(2):182-187. [doi: 10.1016/j.pec.2013.04.016] [Medline: 23711635]
21. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
22. Hargittai E. Digital Na(t)ives? Variation in internet skills and uses among members of the “net generation”*. Sociol Inq 2010 Feb;80(1):92-113. [doi: 10.1111/j.1475-682x.2009.00317.x]
23. Scheerder A, van Deursen A, van Dijk J. Determinants of internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telemat Inform 2017 Dec;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]
24. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
26. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014 Jan;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]
27. Schillinger D. Social determinants, health literacy, and disparities: intersections and controversies. Health Lit Res Pract 2021 Jul;5(3):e234-e243 [FREE Full text] [doi: 10.3928/24748307-20210712-01] [Medline: 34379549]
28. Cullen R. The digital divide: a global and national call to action. Electron Libr 2003;21(3):247-257. [doi: 10.1108/02640470310480506]
29. Ferdinands G, Schram R, de Bruin J, Bagheri A, Oberski DL, Tummers L, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. Syst Rev 2023 Jun 20;12(1):100. [doi: 10.1186/s13643-023-02257-7] [Medline: 37340494]
30. World economic situation and prospects. United Nations. New York: United Nations Publication; 2023. URL: https://www.un-ilibrary.org/content/books/9789210024617/read [accessed 2024-04-29]
31. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
32. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
33. Gilbert S, Irvine R, D'or M, Adam MT, Collins CE, Marriott R, et al. Indigenous women and their nutrition during pregnancy (the Mums and Bubs Deadly Diets Project): protocol for a co-designed mHealth resource development study. JMIR Res Protoc 2023 Jul 06;12:e45983 [FREE Full text] [doi: 10.2196/45983] [Medline: 37147188]
34. Wambach KA, Davis AM, Nelson EL, Snow K, Yungmeyer A, Muzzy R, et al. momHealth: a feasibility study of a multibehavioral health intervention for pregnant and parenting adolescent mothers. Kans J Med 2021;14:176-181 [FREE Full text] [doi: 10.17161/kjm.vol1414738] [Medline: 34262638]

XSL•FO
RenderX

35. Wambach K, Davis AM, Nelson EL, Romine RS, Romero K, Muzzy R, et al. A health behavior and lifestyle intervention pilot trial for childbearing adolescents. Am J Lifestyle Med 2022 Mar 23:155982762210803. [doi: 10.1177/15598276221080367]
36. Flanagan EW, Altazan AD, Comardelle NR, Gilmore LA, Apolzan JW, St Romain J, et al. The design of a randomized clinical trial to evaluate a pragmatic and scalable eHealth intervention for the management of gestational weight gain in low-income women: protocol for the SmartMoms in WIC trial. JMIR Res Protoc 2020 Sep 10;9(9):e18211 [FREE Full text] [doi: 10.2196/18211] [Medline: 32909954]
37. Herring SJ, Albert JJ, Darden N, Bailer B, Cruice J, Hassan S, et al. Targeting pregnancy-related weight gain to reduce disparities in obesity: baseline results from the Healthy Babies trial. Contemp Clin Trials 2019 Dec;87:105822 [FREE Full text] [doi: 10.1016/j.cct.2019.105822] [Medline: 31400513]
38. Vogel L, Färber T, Hölzl I, Deliens T, Henning C, Liel C, et al. I-PREGNO - prevention of unhealthy weight gain and psychosocial stress in families during pregnancy and postpartum using an mHealth enhanced intervention: a study protocol of two cluster randomized controlled trials. BMC Pregnancy Childbirth 2023 Jun 06;23(1):418 [FREE Full text] [doi: 10.1186/s12884-023-05735-x] [Medline: 37280529]
39. Evans WD, Wallace JL, Snider J. Pilot evaluation of the text4baby mobile health program. BMC Public Health 2012 Nov 26;12:1031 [FREE Full text] [doi: 10.1186/1471-2458-12-1031] [Medline: 23181985]
40. Vander Wyst KB, Vercelli ME, O'Brien KO, Cooper EM, Pressman EK, Whisner CM. A social media intervention to improve nutrition knowledge and behaviors of low income, pregnant adolescents and adult women. PLoS One 2019;14(10):e0223120 [FREE Full text] [doi: 10.1371/journal.pone.0223120] [Medline: 31647852]
41. Harris M, Reynolds B. A pilot study of home-based smoking cessation programs for rural, appalachian, pregnant smokers. J Obstet Gynecol Neonatal Nurs 2015;44(2):236-245. [doi: 10.1111/1552-6909.12547] [Medline: 25656503]
42. Abroms LC, Johnson PR, Leavitt LE, Cleary SD, Bushar J, Brandon TH, et al. A randomized trial of text messaging for smoking cessation in pregnant women. Am J Prev Med 2017 Dec;53(6):781-790 [FREE Full text] [doi: 10.1016/j.amepre.2017.08.002] [Medline: 28982527]
43. Huberty JL, Buman MP, Leiferman JA, Bushar J, Hekler EB, Adams MA. Dose and timing of text messages for increasing physical activity among pregnant women: a randomized controlled trial. Transl Behav Med 2017 Jun;7(2):212-223 [FREE Full text] [doi: 10.1007/s13142-016-0445-1] [Medline: 27800565]
44. Holmes H, Palacios C, Wu Y, Banna J. Effect of a short message service intervention on excessive gestational weight gain in a low-income population: a randomized controlled trial. Nutrients 2020 May 15;12(5):1428 [FREE Full text] [doi: 10.3390/nu12051428] [Medline: 32429069]
45. Rhodes A, Pimprikar A, Baum A, Smith AD, Llewellyn CH. Using the person-based approach to develop a digital intervention targeting diet and physical activity in pregnancy: development study. JMIR Form Res 2023 May 26;7:e44082 [FREE Full text] [doi: 10.2196/44082] [Medline: 37234026]
46. Mauriello L, Dyment S, Prochaska J, Gagliardi A, Weingrad-Smith J. Acceptability and feasibility of a multiple-behavior, computer-tailored intervention for underserved pregnant women. J Midwifery Womens Health 2011;56(1):75-80. [doi: 10.1111/j.1542-2011.2010.00007.x] [Medline: 21323854]
47. Prochaska JM, Mauriello L, Dyment S, Gökbayrak S. Designing a health behavior change program for dissemination to underserved pregnant women. Public Health Nurs 2011;28(6):548-555. [doi: 10.1111/j.1525-1446.2011.00959.x] [Medline: 22092465]
48. Kennedy M, Kumar R, Ryan NM, Bennett J, La Hera Fuentes G, Gould GS. Codeveloping a multibehavioural mobile phone app to enhance social and emotional well-being and reduce health risks among Aboriginal and Torres Strait Islander women during preconception and pregnancy: a three-phased mixed-methods study. BMJ Open 2021 Nov 24;11(11):e052545 [FREE Full text] [doi: 10.1136/bmjopen-2021-052545] [Medline: 34819285]
49. Krishnamurti T, Davis AL, Wong-Parodi G, Fischhoff B, Sadovsky Y, Simhan HN. Development and testing of the My Healthy Pregnancy app: a behavioral decision research-based tool for assessing and communicating pregnancy risk. JMIR Mhealth Uhealth 2017 Apr 10;5(4):e42 [FREE Full text] [doi: 10.2196/mhealth.7036] [Medline: 28396302]
50. Abroms LC, Johnson PR, Heminger CL, Van Alstyne JM, Leavitt LE, Schindler-Ruwisch JM, et al. Quit4baby: results from a pilot test of a mobile smoking cessation program for pregnant women. JMIR Mhealth Uhealth 2015 Jan 23;3(1):e10 [FREE Full text] [doi: 10.2196/mhealth.3846] [Medline: 25650765]
51. Leavitt L, Abroms L, Johnson P, Schindler-Ruwisch J, Bushar J, Singh I, et al. Recruiting pregnant smokers from Text4baby for a randomized controlled trial of Quit4baby. Transl Behav Med 2017 Jun 1;7(2):157-165 [FREE Full text] [doi: 10.1007/s13142-016-0450-4] [Medline: 27909881]
52. Comer L, Grassley JS. A smoking cessation website for childbearing adolescents. J Obstet Gynecol Neonatal Nurs 2010;39(6):695-702. [doi: 10.1111/j.1552-6909.2010.01179.x] [Medline: 20880049]
53. Whittaker R, Matoff-Stepp S, Meehan J, Kendrick J, Jordan E, Stange P, et al. Text4baby: development and implementation of a national text messaging health information service. Am J Public Health 2012 Dec;102(12):2207-2213. [doi: 10.2105/AJPH.2012.300736] [Medline: 23078509]
54. Remick AP, Kendrick JS. Breaking new ground: the text4baby program. Am J Health Promot 2013;27(3 Suppl):S4-S6 [FREE Full text] [doi: 10.4278/ajhp.27.3.c2] [Medline: 23286662]

XSL•FO
RenderX

55. Huberty J, Rowedder L, Hekler E, Adams M, Hanigan E, McClain D, et al. Development and design of an intervention to improve physical activity in pregnant women using Text4baby. Transl Behav Med 2016 Jun;6(2):285-294 [FREE Full text] [doi: 10.1007/s13142-015-0339-7] [Medline: 27356999]
56. van Dijk W, Oosterman M, Jansen I, de Vente W, Huizink A. Stress- and smoke free pregnancy study protocol: a randomized controlled trial of a personalized eHealth intervention including heart rate variability-biofeedback to support pregnant women quit smoking via stress reduction. BMC Public Health 2021 May 12;21(1):905 [FREE Full text] [doi: 10.1186/s12889-021-10910-w] [Medline: 33980201]
57. Carolan-Olah M, Vasilevski V, Nagle C, Stepto N. Overview of a new eHealth intervention to promote healthy eating and exercise in pregnancy: initial user responses and acceptability. Internet Interv 2021 Sep;25:100393 [FREE Full text] [doi: 10.1016/j.invent.2021.100393] [Medline: 34026565]
58. Al-Dhahir I, Reijnders T, Faber JS, van den Berg-Emons RJ, Janssen VR, Kraaijenhagen RA, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022 Aug 24;24(8):e34229 [FREE Full text] [doi: 10.2196/34229] [Medline: 36001380]
59. Rhodes A, Smith AD, Llewellyn CH, Croker H. Investigating partner involvement in pregnancy and identifying barriers and facilitators to participating as a couple in a digital healthy eating and physical activity intervention. BMC Pregnancy Childbirth 2021 Jun 28;21(1):450 [FREE Full text] [doi: 10.1186/s12884-021-03917-z] [Medline: 34182953]
60. Kreuter MW, Farrell DL, Olevitch L, Brennan LK. Tailoring Health Messages: Customizing Communication With Computer Technology. New York, NY: Routledge; 2013.
61. Mitchell MS, Faulkner GE. On supplementing "Foot in the door" incentives for eHealth program engagement. J Med Internet Res 2014 Jul 25;16(7):e179 [FREE Full text] [doi: 10.2196/jmir.3701] [Medline: 25092221]
62. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023 Dec;9(12):20552076231198702 [FREE Full text] [doi: 10.1177/20552076231198702] [Medline: 37691766]
63. De las Nueces D, Hacker K, DiGirolamo A, Hicks LS. A systematic review of community-based participatory research to enhance clinical trials in racial and ethnic minority groups. Health Serv Res 2012 Jun 21;47(3 Pt 2):1363-1386 [FREE Full text] [doi: 10.1111/j.1475-6773.2012.01386.x] [Medline: 22353031]
64. Dugstad J, Eide T, Nilsen ER, Eide H. Towards successful digital transformation through co-creation: a longitudinal study of a four-year implementation of digital monitoring technology in residential care for persons with dementia. BMC Health Serv Res 2019 Jun 10;19(1):366 [FREE Full text] [doi: 10.1186/s12913-019-4191-1] [Medline: 31182093]
65. Haluza D, Jungwirth D. ICT and the future of health care: aspects of health promotion. Int J Med Inform 2015 Jan;84(1):48-57. [doi: 10.1016/j.ijmedinf.2014.09.005] [Medline: 25293532]
66. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. Correction: iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 18;25(12):e49027 [FREE Full text] [doi: 10.2196/49027] [Medline: 37201181]

Abbreviations

mHealth: mobile health
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SEP: socioeconomic position

Edited by A Mavragani; submitted 09.11.23; peer-reviewed by A Pimprikar, RT Villarino; comments to author 28.02.24; revised version received 23.04.24; accepted 01.10.24; published 20.12.24.

Please cite as:
Smit AJP, Al-Dhahir I, Schiphof-Godart L, Breeman LD, Evers AWM, Joosten KFM
Investigating eHealth Lifestyle Interventions for Vulnerable Pregnant Women: Scoping Review of Facilitators and Barriers
J Med Internet Res 2024;26:e54366
URL: https://www.jmir.org/2024/1/e54366
doi:10.2196/54366
PMID:

©Ashley JP Smit, Isra Al-Dhahir, Lieke Schiphof-Godart, Linda D Breeman, Andrea WM Evers, Koen FM Joosten. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Large Language Models in Gastroenterology: Systematic Review

Eun Jeong Gong[1,2,3], MD, PhD; Chang Seok Bang[1,2,3], MD, PhD; Jae Jun Lee[3], MD, PhD; Jonghyung Park[4], MS; Eunsil Kim[4], MS; Subeen Kim[4], MS; Minjae Kimm[5], PhD; Seoung-Ho Choi[6], MS

[1]Department of Internal Medicine, Hallym University College of Medicine, Chuncheon, Republic of Korea
[2]Institute for Liver and Digestive Diseases, Hallym University, Chuncheon, Republic of Korea
[3]Institute of New Frontier Research, Hallym University College of Medicine, Chuncheon, Republic of Korea
[4]Meninblox, Inc, Gwangju, Republic of Korea
[5]Tech University of Korea, Siheung, Republic of Korea
[6]Hansung University, Seoul, Republic of Korea

Corresponding Author:
Chang Seok Bang, MD, PhD
Department of Internal Medicine, Hallym University College of Medicine
sakjuro, 77
Chuncheon
Republic of Korea
Phone: 82 332405000
Email: csbang@hallym.ac.kr

Abstract

Background: As health care continues to evolve with technological advancements, the integration of artificial intelligence into clinical practices has shown promising potential to enhance patient care and operational efficiency. Among the forefront of these innovations are large language models (LLMs), a subset of artificial intelligence designed to understand, generate, and interact with human language at an unprecedented scale.

Objective: This systematic review describes the role of LLMs in improving diagnostic accuracy, automating documentation, and advancing specialist education and patient engagement within the field of gastroenterology and gastrointestinal endoscopy.

Methods: Core databases including MEDLINE through PubMed, Embase, and Cochrane Central registry were searched using keywords related to LLMs (from inception to April 2024). Studies were included if they satisfied the following criteria: (1) any type of studies that investigated the potential role of LLMs in the field of gastrointestinal endoscopy or gastroenterology, (2) studies published in English, and (3) studies in full-text format. The exclusion criteria were as follows: (1) studies that did not report the potential role of LLMs in the field of gastrointestinal endoscopy or gastroenterology, (2) case reports and review papers, (3) ineligible research objects (eg, animals or basic research), and (4) insufficient data regarding the potential role of LLMs. Risk of Bias in Non-Randomized Studies—of Interventions was used to evaluate the quality of the identified studies.

Results: Overall, 21 studies on the potential role of LLMs in gastrointestinal disorders were included in the systematic review, and narrative synthesis was done because of heterogeneity in the specified aims and methodology in each included study. The overall risk of bias was low in 5 studies and moderate in 16 studies. The ability of LLMs to spread general medical information, offer advice for consultations, generate procedure reports automatically, or draw conclusions about the presumptive diagnosis of complex medical illnesses was demonstrated by the systematic review. Despite promising benefits, such as increased efficiency and improved patient outcomes, challenges related to data privacy, accuracy, and interdisciplinary collaboration remain.

Conclusions: We highlight the importance of navigating these challenges to fully leverage LLMs in transforming gastrointestinal endoscopy practices.

Trial Registration: PROSPERO 581772; https://www.crd.york.ac.uk/prospero/

(J Med Internet Res 2024;26:e66648) doi:10.2196/66648

KEYWORDS

large language model; LLM; deep learning; artificial intelligence; AI; endoscopy; gastroenterology; clinical practice; systematic review; diagnostic; accuracy; patient engagement; emotional support; data privacy; diagnosis; clinical reasoning

XSL•FO
RenderX

Introduction

Background

In the rapidly evolving landscape of health care, the convergence of medicine and technology has opened new avenues for improving patient care and operational efficiency. Among the most promising technological advancements is the development of artificial intelligence (AI), particularly large language models (LLMs), which have the potential to transform various medical specialties. The advent of AI has ushered in a new era of innovation across various sectors, with health care being a primary beneficiary [1-3].

Gastrointestinal endoscopy, a critical field for diagnosing and treating digestive diseases, stands on the cusp of significant advancements with the integration of these technologies. This, pivotal in diagnosing and treating digestive tract diseases, faces challenges like diagnostic variability and labor-intensive documentation. Gastrointestinal endoscopy, traditionally reliant on the expertise of specialists to interpret complex visual data and execute precise interventions, can greatly benefit from the automation and analytical capabilities provided by LLMs [4,5]. While traditional convolutional neural network (CNN)–based lesion detection or lesion diagnosis models in endoscopy have addressed the limitations of endoscopists' visual diagnosis, LLMs capable of processing massive datasets are expected to address a wider range of clinical unmet needs by enhancing diagnostic support, automating report generation, and improving educational tools [1-3]. The AI models built thus far are narrow AI models designed to tackle a specific task for a specific purpose, and they have demonstrated very high performance within the scope of the problem they are intended to solve. Nevertheless, performance drops when the task's objective or the type of data changes, necessitating a repeat of the data collection and model-improvement procedure in order to address the issue. Rapid advances in generative AI and the emergence of foundation models have led to the possibility of performing a wide range of tasks with no or minimal additional training.

Brief History of LLMs

The inception of LLMs traces back to the early developments in AI and natural language processing. Initially, these models were designed to understand and generate human-like text by learning from vast datasets of text on the internet. The breakthrough came with the introduction of the Transformer architecture in 2017, which enabled models to handle long-range dependencies in text, significantly improving their understanding and generation capabilities. This led to the development of models like GPT and Bidirectional Encoder Representations from Transformers (BERT), which demonstrated unprecedented performance in a wide range of natural language processing tasks. LLM has been trained with tens to hundreds of billions of parameters using very large amounts of data, and it is mainly trained based on a pretext task that predicts the next word or token using a list of given words or tokens as input. One of the most defining characteristics of the LLM is the emergent ability of the model to grow in size, even when it is not trained for a specific task [6]. It has the ability to perform zero-shot learning, where it performs a task without examples, as well as few-shot learning, where its performance increases when a few examples are provided. These advancements paved the way for the integration of LLMs into various sectors, including health care, where they have been instrumental in enhancing diagnostic accuracy, automating documentation, and facilitating patient care. The evolution of LLMs has been marked by continuous improvements in model architecture, training techniques, and dataset quality, culminating in the current generation of models that are capable of complex reasoning and generating coherent, contextually relevant text.

Prompt Engineering and Fine-Tuning

LLMs are pretrained models, and the learning process can be broadly divided into 2 parts: pretraining and fine-tuning. In the pretraining phase, the model is trained using large amounts of unstructured textual data (eg, news articles and novels), while in the fine-tuning phase, the pretrained model is fine-tuned for a specific task (model updates). In this fine-tuning phase, a small amount of labeled data is used to tune the model. In this step, the parameters of the model are fine-tuned to achieve optimal results for a specific task.

In-context learning is achieved through a technique called "prompt engineering" (no model updates). When a user enters a prompt, an LLM analyzes the text to understand the context and generate relevant output, that is, it performs the task you want it to do based on the content of the prompt alone. Literally, it means that the model understands (learns) the contextual meaning of the prompt (in-context) and generates an answer to it [6]. In-context learning does not update the weight of the model like pretraining or fine-tuning, and there is no separate model training process. Therefore, the importance of prompt engineering is emphasized because well-written prompts lead to good results. Depending on the number of examples, it can be divided into zero-shot, one-shot, and few-shot learning. One hypothesis for this unsupervised learning outcome is that the process of "inference" can be a form of "optimization" (maximum likelihood estimation) [7]. This systematic review aims to explore the emerging role of LLMs in gastroenterology, especially for gastrointestinal endoscopy, offering insights into how they can support specialists in improving diagnostic accuracy, streamlining documentation, enhancing training, and engaging patients.

Methods

Study Design

A systematic review was performed to check the recent research trend of the potential role of LLMs in gastrointestinal endoscopy. The protocol of this systematic review was registered at PROSPERO (581772) before the initiation of this study. The PICO (population, intervention, comparator, and outcome) was as follows: population: general population or patients with gastrointestinal disorders; intervention: application of LLMs; comparator: none; and outcome: clinical benefits or improvement of performance. This systematic review was performed in accordance with the statement of the PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analyses) for systematic review [8] (Multimedia Appendix 1).

XSL•FO
RenderX

Databases

MEDLINE (through PubMed), Embase, and CENTRAL in the Cochrane Library were searched using common keywords (from inception to April 2024). Medical Subject Headings terminology or Emtree keywords were used for the search strategy. Two evaluators (EJG and CSB) independently performed literature searching, and disagreements between the 2 evaluators were resolved by discussion or consultation with a third author (JJL).

In detail, duplicated papers were first removed from the retrieved papers (through searching of 3 databases and hand searching) by using the find duplicate function of the EndNote software program (Clarivate). Next, we selected papers by looking at the abstracts and titles and then checked for full-text paper eligibility. Full-text papers were evaluated for the final systematic review based on predefined inclusion and exclusion criteria. The eligibility of all studies was assessed by 2 blinded independent raters (EJG and CSB), and discrepancies were resolved by consultation with a third author (JJL).

Inclusion and Exclusion Criteria

Studies were included if they satisfied the following criteria: (1) any type of studies that investigated the potential role of LLMs in the field of gastrointestinal endoscopy or gastroenterology, (2) studies published in English, and (3) studies in full-text format. The exclusion criteria were as follows: (1) studies that did not report the potential role of LLMs in the field of gastrointestinal endoscopy or gastroenterology, (2) case reports and review papers, (3) ineligible research objects (eg, animals or basic research), and (4) insufficient data regarding the potential role of LLMs. Only publications conducted on human participants were searched, and the bibliographies of relevant papers were also reviewed to identify additional studies. The search strategy to find the relevant papers is described in Textbox 1.

Textbox 1. Search strategy to find the relevant papers.

Database: MEDLINE (through PubMed; April 12, 2024)

- #1 “large language model”[tiab] OR “LLM”[tiab] OR “foundation model”[tiab] OR “language vision model”[tiab] OR “GPT”[tiab] OR “ChatGPT”[tiab] OR “BERT”[tiab] OR “Claude”[tiab] OR “transformer”[tiab] OR “generative AI”[tiab]: 18148
- #2 “gastroenterology”[tiab] OR “gastrointestinal”[tiab] OR “endoscopy”[tiab] OR “gastroscopy”[tiab] OR “colonoscopy”[tiab]: 425845
- #3 #1 AND #2: 178
- #4 #3 AND English[Lang]: 132

Database: Embase

- #1 “large language model”:ab,ti,kw OR “LLM”:ab,ti,kw OR “foundation model deep”:ab,ti,kw OR “language vision model”:ab,ti,kw OR “GPT”:ab,ti,kw OR “ChatGPT”:ab,ti,kw OR “BERT”:ab,ti,kw OR “Claude”:ab,ti,kw OR “transformer”:ab,ti,kw OR “generative AI”: 20956
- #2 “gastoenterology”:ab,ti,kw OR “gastrointestinal”:ab,ti,kw OR “endoscopy”:ab,ti,kw OR “gastroscopy”:ab,ti,kw OR “colonoscopy”:ab,ti,kw: 624510
- #3 #1 AND #2: 242
- #4 #3 AND ([article]/lim OR [article in press]/lim OR [review]/lim) AND [English]/lim: 118

Database: Cochrane Library

- #1 (large language model):ab,ti,kw OR LLM:ab,ti,kw OR (foundation model):ab,ti,kw OR (language vision model):ab,ti,kw OR GPT:ab,ti,kw OR ChatGPT:ab,ti,kw OR BERT:ab,ti,kw OR Claude:ab,ti,kw OR transformer:ab,ti,kw OR (generative AI): 2550
- #2 gastoenterology:ab,ti,kw OR gastrointestinal:ab,ti,kw OR endoscopy:ab,ti,kw OR gastroscopy:ab,ti,kw OR colonoscopy:ab,ti,kw: 72568
- #3 #1 and #2: 123

Data Extraction

Two evaluators (EJG and CSB) independently extracted the outcomes of all the finally included studies using Microsoft Excel sheet form (knowledge-based response evaluation, document summary or AI-generated draft response, overcome language barriers, identifying research questions, and combining multiple tasks including causal inference) and disagreements between the 2 evaluators were resolved by discussion or consultation with a third author (JJL).

Methodological Quality Assessment

Risk of Bias in Non-Randomized Studies—of Interventions (ROBINS-I) was used to evaluate the quality of the identified studies [8]. Seven domains make up the ROBINS-I tool: “bias due to confounding,” “bias in selection of participants into the study,” “bias in classification of intervention,” “bias due to deviations from intended interventions,” “bias due to missing data,” “bias in measurement outcomes,” and “bias in selection of the reported result.” It is established that there is a “low,” “moderate,” “serious,” or “critical risk of bias” in each domain. The evaluation of each domain level determines the overall risk of bias judgment; a low risk suggests that the study is equivalent to a well-conducted randomized trial for all domains under consideration. A nonrandomized study with a “moderate risk of bias” has solid evidence, but it cannot be compared to a randomized trial. When there is a “serious risk of bias” in 1 or more domains but not a “critical risk of bias” in any 1 domain, it suggests that there are “serious risk of bias.” “Critical risk of bias” in at least 1 domain signifies that there is a significant

XSL•FO
RenderX

danger of bias in the study, making it difficult to draw any meaningful conclusions [30]. Two evaluators (EJG and CSB) independently performed the methodological quality assessment, and disagreements between the 2 evaluators were resolved by discussion or consultation with a third author (JJL).

Data Synthesis

Narrative synthesis was done because of heterogeneity in the specified aims and methodology in each included study.

Results

Potential Role of LLMs in Gastrointestinal Endoscopy

LLMs can transform gastrointestinal endoscopy by improving diagnostic accuracy, streamlining documentation, and enriching education and patient engagement [1]. By analyzing endoscopic imagery with precision and automating report generation, LLMs offer a layer of analysis that could reduce diagnostic errors and administrative burdens. Medical information retrieval is another potential role of LLMs [1]. They can quickly access and interpret large volumes of medical data and provide accurate answers to complex medical queries. This is particularly useful for answering rare or obscure medical questions and staying updated on the latest medical research. Through the same process stated earlier, they can support the diagnosis process and treatment recommendation. This can be used for clinical reasoning using LLMs with real medical cases. Furthermore, their application in creating interactive training materials and personalizing patient education presents an exciting frontier for the field [31]. One important benefit that should not be overlooked is emotional support [32]. By providing emotional support, answering patient queries, and assisting with daily tasks, they can help improve the overall patient experience and satisfaction [32].

Systematic Review

A total of 373 studies were identified from the literature searching process on the 3 databases. Nine studies were additionally identified by manual screening of references. After excluding duplicate studies, additional papers were excluded after reviewing their titles and abstracts. Full-text versions of the remaining 97 studies were obtained and thoroughly reviewed based on the aforementioned inclusion and exclusion criteria. Among these, 76 papers were excluded because these papers did not meet the inclusion criteria (narrative review: n=1, study with incomplete data: n=65, systematic review or meta-analysis: n=3, study protocol: n=2, and editorial or comment or letter: n=5). Finally, 21 studies [9-29] for the potential role of LLMs in gastrointestinal disorders were included in the systematic review. A flowchart of the study selection process is shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process.

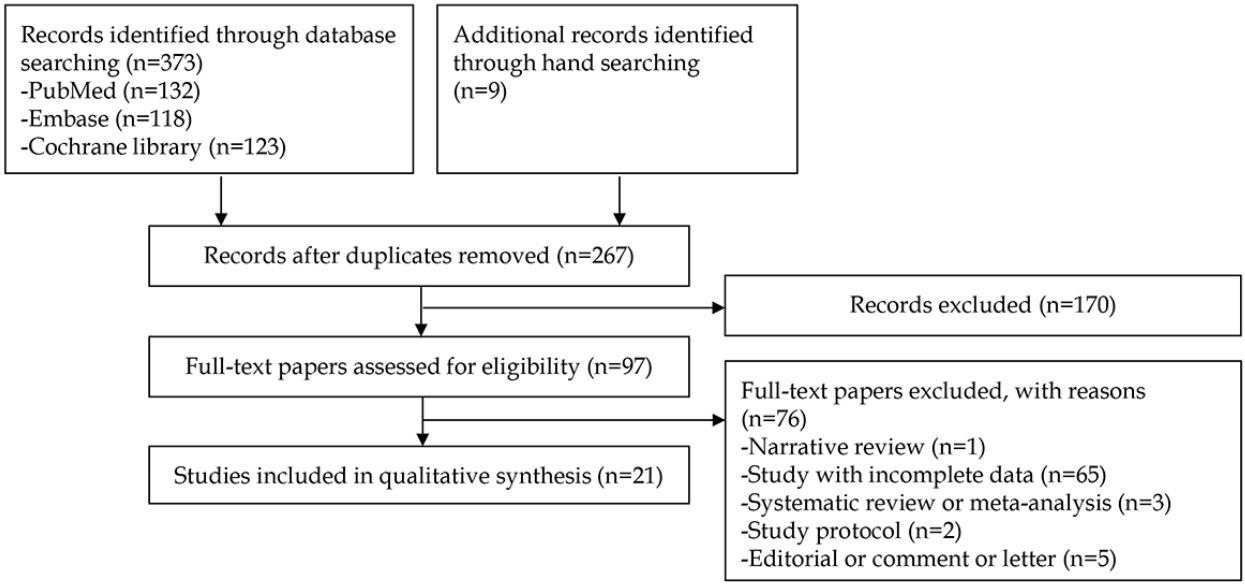

Table 1 shows the summary of this systematic review. Each retrieved study [9-29] can be categorized by several topics, including knowledge-based response evaluation [9-22], document summary or AI-generated draft responses [23,24], overcome language barrier [25,26], identifying research questions [27], or combining multiple tasks including causal inference [28,29].

XSL•FO
RenderX

Table 1. Clinical summary of the included studies.

Outcome and study (year)	Nationality (institution)	Type of AI[a] model	Study topic	Evaluation	Rating
Knowledge-based response evaluation					
Lim et al (2024) [9]	Singapore	GPT-4	To evaluate whether the contextualized GPT model (using guidelines) could provide correct advice for screening and surveillance intervals for colonoscopy (62 example case scenarios)	3 gastroenterology fellows under the supervision of 2 senior gastroenterologists	The contextualized GPT-4 model outperformed the standard GPT-4 in all domains. No high-risk features were missed, and only 2 cases had hallucinations of additional high-risk features. A correct interval to colonoscopy was provided in the majority of cases. Guidelines were appropriately cited in almost all cases.
Munir et al (2024) [10]	United States	ChatGPT in June 2023	To assess the quality and perceived utility of chat-based AI responses related to 3 common gastrointestinal surgical procedures (cholecystectomy, pancreaticoduodenectomy, and colectomy)	45 surgeons	Overall, the most commonly assigned quality grade was "fair" or "good" for most responses (622/1080, 57.6%). Most of the 1080 total utility grades were "fair" (n=279, 25.8%) or "good" (n=344, 31.9%), whereas only 129 utility grades (11.9%) were "poor." Overall, only 20% of the experts deemed ChatGPT to be an accurate source of information, whereas 15.6% of the experts found it unreliable. Moreover, 1 in 3 surgeons deemed ChatGPT responses as not likely to reduce patient-physician correspondence (31.1%) or not comparable to in-person surgeon responses (35.6%).
Rammohan et al (2024) [11]	United States	ChatGPT-4.0 and Google Bard (2023)	To assess the reliability and accuracy of LLMs[b] in answering gastroenterology-related queries	2 independent reviewers	ChatGPT-4.0 demonstrated higher reliability and accuracy in its responses than Google Bard, as indicated by higher mean ratings and statistically significant *P* values in hypothesis testing. However, limitations in the data structure, such as the inability to conduct detailed correlation analysis, were noted.
Tariq et al (2024) [12]	United States	ChatGPT-3.5, ChatGPT-4, and Bard (version July 2023, queries ran on July 17, 2023)	47 common patient inquiries related to colonoscopy	2 reviewer gastroenterologists	In total, 43 of 47 (91.4%) responses graded as completely correct, whereas 4 of 47 (8.6%) responses by ChatGPT-4 were graded as correct but incomplete.
Gravina et al (2024) [13]	Italy	ChatGPT-3.5 and Perplexity AI	Performance in responding to questions from the 2023 Italian national residency admission examination (SSM23[c]) and comparing results and chatbots' concordance with previous years' SSMs	N/A[d]	In SSM23, ChatGPT-3.5 outperforms Perplexity AI with 94.11% of correct responses, demonstrating consistency across years. Concordance weakened in 2023 (κ=0.203; *P*=.148), but ChatGPT consistently maintains a high standard compared to Perplexity AI.
Atarere et al (2024) [14]	United States	ChatGPT, BingChat, and YouChat (April 2023)	15 questions on important colorectal cancer screening concepts and 5 common questions asked by patients	2 board-certified internal medicine physicians	ChatGPT and YouChat provided reliably appropriate responses to all 15 (100%) questions, while BingChat provided reliably appropriate responses to 13 of 15 (86.7%) questions and unreliable responses to 2 of 15 (13.3%) questions.
Gorelik et al (2024) [15]	Israel	Customized GPT	A custom GPT was developed to provide guideline-based management advice for pancreatic cysts	2 gastroenterologists (pancreatobiliary specialists) and a hepatobiliary surgeon	The custom GPT aligned with expert recommendations in 87% of scenarios. Initial expert recommendations were correct in 97% and 87% of cases, respectively. No significant difference was observed between the accuracy of custom GPT and the experts.
Cankurtaran et al (2023) [16]	Turkey	ChatGPT-4 (March 2023)	20 specific questions regarding IBD[e]	2 experts	Reliability and usefulness score as follows: mean 5.00 (SD 1.21) and mean 5.15 (SD 1.08), respectively (7-point Likert scale).

XSL•FO
RenderX

Outcome and study (year)	Nationality (institution)	Type of AI[a] model	Study topic	Evaluation	Rating
Gorelik et al (2023) [17]	Israel	ChatGPT (GPT-4)	Compliance with guidelines and accuracy about 20 clinical scenarios relevant to post-colonoscopy patient management	2 senior gastroenterologists	ChatGPT exhibited 90% compliance with guidelines and 85% accuracy, with a very good interrater agreement (Fleiss κ coefficient of 0.84; P<.01).
Ali et al (2023) [18]	United States	ChatGPT (launched in November 2022)	113 questions related to EGD[f], colonoscopy, EUS[g], and ERCP[h]	At least 2 board-certified or eligible advanced endoscopists	Moderate precision in answering questions about EGD (57.9% comprehensive), colonoscopy (47.6% comprehensive), EUS (48.1% comprehensive), and ERCP (44.4% comprehensive). Medical accuracy was highest for EGD (52.6% fully accurate) and lowest for EUS (40.7% fully accurate).
Lahat et al (2023) [19]	Israel	ChatGPT (November 2022)	To evaluate the performance of ChatGPT in answering patients' 110 real-life questions regarding gastrointestinal health	3 experienced gastroenterologists	About treatments, the average (SD) accuracy, clarity, and efficacy scores (1 to 5) were 3.9 (0.8), 3.9 (0.9), and 3.3 (0.9), respectively. For symptom questions, the average (SD) accuracy, clarity, and efficacy scores were 3.4 (0.8), 3.7 (0.7), and 3.2 (0.7), respectively. For diagnostic test questions, the average (SD) accuracy, clarity, and efficacy scores were 3.7 (1.7), 3.7 (1.8), and 3.5 (1.7), respectively.
Lee et al (2023) [20]	United States	ChatGPT (January 30, 2023, version)	To evaluate the answers about 8 common questions about colonoscopy (compared to publicly available web pages of 3 randomly selected hospitals from the top 20 list of the US News & World Report's Best Hospitals for Gastroenterology and Gastrointestinal Surgery)	4 gastroenterologists (2 senior gastroenterologists and 2 fellows)	Gastroenterologists rated ChatGPT answers similarly to non-AI answers in ease of understanding (AI: 5.0-6.4 vs non-AI: 4.8-5.8), with the AI mean scores higher than non-AI scores. Scientific adequacy scores were also similar (AI: 5.4-6.5 vs non-AI: 5.1-6.3; nonsignificant), with the AI mean score higher than non-AI 63% of the time. AI and non-AI answers received similar ratings regarding satisfaction with the answers (AI: 4.9-6.3 vs non-AI: 4.8-5.8; nonsignificant).
Samaan et al (2023) [21]	United States	March 14, 2023, version of GPT-4	To examine the accuracy and reproducibility of responses by GPT-4 to 88 patient nutrition questions related to IBD	2 IBD-focused registered dieticians	The model provided correct responses to 73 of 88 (83%) questions, with 61 (69%) graded as comprehensive. A total of 15 of 88 (17%) responses were graded as mixed with correct and incorrect or outdated data.
Henson et al (2023) [22]	United States	ChatGPT (version March 14, 2023)	Ability to respond appropriately to questions regarding gastroesophageal reflux disease (23 question prompts)	3 gastroenterologists and 8 patients	Appropriate responses (91.3%), although with some inappropriateness (8.7%) and inconsistency. Most responses (78.3%) contained at least some specific guidance. Patients considered this a useful tool (100%).
Document summary or AI-generated draft response					
Garcia et al (2024) [23]	United States	GPT-3.5 Turbo and GPT-4 (July to August 2023)	AI-generated draft response utilization rate across clinicians	162 clinicians	The mean AI-generated draft response utilization rate across clinicians was 20%. There were statistically significant reductions in the 4-item physician task load score derivative and work exhaustion scores.
Syed et al (2022) [24]	United States	Hybrid artificial neural network to concatenate and fine-tune BERT[i] and Flair embeddings	To extract comprehensive clinical concepts from the consolidated colonoscopy documents	Validated using 300 colonoscopy procedures (the chart review was done by 4 reviewers [1 medical student and 3 trained data warehouse analysts] under the guidance of domain expert)	F_1-scores of 91.76%, 92.25%, and 88.55% for colonoscopy, pathology, and radiology reports, respectively (5-fold cross-validation).
Overcome language barriers					

XSL•FO
RenderX

Outcome and study (year)	Nationality (institution)	Type of AI[a] model	Study topic	Evaluation	Rating
Yeo et al (2023) [25]	United States	ChatGPT and GPT-4	Evaluates ChatGPT and GPT-4's ability to comprehend and respond to cirrhosis-related questions in English, Korean, Mandarin, and Spanish, addressing language barriers that may impact patient care	Native-speaking hepatologists	GPT-4 showed a marked improvement in the proportion of comprehensive and correct answers compared to ChatGPT across all 4 languages (P<.05). GPT-4 demonstrated enhanced accuracy and avoided erroneous responses evident in ChatGPT's output.
Samaan et al (2023) [26]	United States	ChatGPT (January 30, 2023, version)	ChatGPT's accuracy in responding to cirrhosis-related questions in Arabic and compared its performance to English (91 questions in Arabic and English were graded. Accuracy of responses was assessed using the scale.)	A transplant hepatologist fluent in both languages	The model provided 22 (24.2%) comprehensive, 44 (48.4%) correct but inadequate, 13 (14.3%) mixed with correct and incorrect or outdated data, and 12 (13.2%) completely incorrect Arabic responses. When comparing the accuracy of Arabic and English responses, 9 (9.9%) of the Arabic responses were graded as more accurate, 52 (57.1%) similar in accuracy, and 30 (33%) as less accurate compared to English.
Identifying research questions					
Lahat et al (2023) [27]	Israel	ChatGPT (December 15, 2023)	To evaluate the potential of ChatGPT for identifying research priorities in gastroenterology and provide a starting point for further investigation, we queried ChatGPT on 4 key topics in gastroenterology: IBD, microbiome, AI in gastrointestinal, and advanced endoscopy in gastroenterology	3 experienced gastroenterologists	On average, the questions were rated 3.6 (SD 1.4), with interrater reliability ranging from 0.80 to 0.98 (P<.001). The mean (SD) grades for relevance, clarity, specificity, and originality were 4.9 (0.1), 4.6 (0.4), 3.1 (0.2), and 1.5 (0.4), respectively (1-5 scale).
Combining multiple tasks including causal inference					
Zhou et al (2023) [28]	China	ChatGPT and GPT-4	To explore ChatGPT's potential in disseminating gastric cancer knowledge, providing consultation recommendations, and interpreting endoscopy reports	Pre-established ground truth	GPT-4 model of ChatGPT achieved an appropriateness of 91.3% and a consistency of 95.7% in a gastric cancer knowledge test.
Wang et al (2022) [29]	United States	BioBERT	Causal inference of idiosyncratic DILI[j] based on LiverTox	Domain experts	Accuracy of 0.92 and an F_1-score of 0.84 for the DILI prediction. High concordance of 0.91 between the severity scores generated by model and domain experts.

[a]AI: artificial intelligence.
[b]LLM: large language model.
[c]SSM: Scuole Specializzazione Medicina.
[d]N/A: not applicable.
[e]IBD: inflammatory bowel disease.
[f]EGD: esophagogastroduodenoscopy.
[g]EUS: endoscopic ultrasound.
[h]ERCP: endoscopic retrograde cholangiopancreatography.
[i]BERT: Bidirectional Encoder Representations from Transformers.
[j]DILI: drug-induced liver injury.

Most retrieved studies [9-22] have measured reliability by asking LLMs about their common or specified medical knowledge, such as common gastrointestinal disorders or gastrointestinal procedures (gastroesophageal reflux disease management, nutrition questions related to inflammatory bowel disease, screening and surveillance intervals for colonoscopy, guideline-based management advice for pancreatic cysts, or board examination tests). Evaluation of the performance in LLM was rated by expert endoscopists or gastroenterologists, and most of the studies have shown real-world applicability.

Another topic was the document summary or AI-generated draft responses [23,24], and these studies showed the potential for usability and improvement in assessments of the burden and burnout of medical specialists. Although the LLMs have the

XSL•FO
RenderX

potential to serve as an adjunct source of information for patients, language barriers could impact the quality of response, and 2 studies [25,26] have pointed out this issue and the need for applications in diverse linguistic contexts for LLMs.

In the context of medical research, LLMs have been used in medical research to streamline literature reviews, enhance drug discovery processes, and assist in the design and analysis of clinical trials. Additionally, they support personalized medicine, biomedical data mining, and the interpretation of complex clinical information for improved decision-making and patient care. In this systematic review, one study was identified for this issue, and Lahat et al [27] tried to evaluate the potential of ChatGPT for identifying research priorities in gastroenterology and provide a starting point for further investigation. They showed that LLMs may be a useful tool for identifying research priorities in the field of gastroenterology, although more work is needed to improve the novelty of the generated research questions.

Since LLMs are capable of performing a wide range of tasks, studies have been conducted to align these functions in order to create a streamline and assess their performance. This necessitates a more complicated or occasionally customized LLM models. Wang et al [29] tried to establish a causal inference model of idiosyncratic drug-induced liver injury (DILI) based on the LiverTox database. BioBERT (fine-tuned model with biomedical-specific corpora, including PubMed abstracts and PubMed Central full-text papers) was used as a backbone model. To make BioBERT more specific for the DILI application, they further fine-tuned the BioBERT model with the extracted sentences from LiverTox [27]. Since the presumptive diagnosis of exclusion is the main diagnosis of DILI, it is time-consuming; however, this model has the potential to help differentiate it.

Methodological Quality

The primary limitation was the question-and-answer interaction of LLMs to evaluate the measurement of outcomes in the “knowledge-based response evaluation” [9-22] or “overcome language barriers” studies [25,26]. This may differ from actual patient-physician conversations or practice situations and is subject to evaluator bias. Therefore, all the studies were rated in a moderate risk of bias in the “bias in measurement of outcomes” domain. Otherwise, all the other remaining domains were rated as “low risk of bias”: in the ROBINS-I tool methodology, evaluation was in the “knowledge-based response evaluation” studies (Table 2).

XSL•FO
RenderX

Table 2. Risk of bias evaluation (ROBINS-I[a] assessment tool).

Study	Bias preintervention and at intervention domains			Risk of bias postintervention domains				Overall assessment of bias
	Bias due to confounding	Bias in selection of participants into the study	Bias in classification of intervention	Bias due to deviations from intended intervention	Bias due to missing data	Bias in measurement of outcomes	Bias in selection of the reported result	
Lim et al (2024) [9]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Munir et al (2024) [10]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Rammohan et al (2024) [11]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Tariq et al (2024) [12]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Gravina et al (2024) [13]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Atarere et al (2024) [14]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Gorelik et al (2024) [15]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Cankurtaran et al (2023) [16]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Gorelik et al (2023) [17]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Ali et al (2023) [18]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Lahat et al (2023) [19]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Lee et al (2023) [20]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Samaan et al (2023) [21]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Henson et al (2023) [22]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Garcia et al (2024) [23]	Low	Low	Low	Low	Low	Low	Low	Low
Syed et al (2022) [24]	Low	Low	Low	Low	Low	Low	Low	Low
Yeo et al (2023) [25]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Samaan et al (2023) [26]	Low	Low	Low	Low	Low	Moderate	Low	Moderate
Lahat et al (2023) [27]	Low	Low	Low	Low	Low	Low	Low	Low
Zhou et al (2023) [28]	Low	Low	Low	Low	Low	Low	Low	Low
Wang et al (2022) [29]	Low	Low	Low	Low	Low	Low	Low	Low

[a]ROBINS-I: Risk of Bias in Non-Randomized Studies—of Interventions.

In terms of the studies with "document summary or AI-generated draft response" [23,24], "identifying research questions" [27], or "combining multiple tasks including causal inference" [28,29], the risk of "bias in measurement of outcomes" was minimal because pre-established ground truth exists or LLM's

XSL•FO
RenderX

answer was not the primary outcome. Therefore, all the studies were rated in a low risk of bias in these domains (Table 2).

Discussion

Principal Findings

This study explored the emerging role of LLMs in gastroenterology, especially for gastrointestinal endoscopy, providing a summary of recently published relevant papers through the systematic review process. In total, 21 studies from systematic review revealed the potential of LLMs for disseminating general medical knowledge, providing consultation recommendations, automatic generation of procedure reports, or causal inference of presumptive diagnosis of complex medical disorders.

Benefits and Limitations

The integration of LLMs promises improved diagnostic accuracy, efficiency, enhanced education, and better patient engagement. In addition, LLMs can be used as a source of knowledge for medical staff and as a source of medical knowledge for patients [33]. It can be used as a questionnaire system to help medical staff and patients communicate, and it can also be used to generate papers and organize data for research purposes. This can be also used for quality improvement in medical practice. For example, LLMs can analyze electronic health records and better identify specified patients for alerts [34]. It is already proven to be good at arithmetic reasoning [35] and can also be used for the analysis of quality metrics, such as adenoma detection rate or polyp detection rate through the analysis of pathology reports of the patients. However, challenges such as ensuring data privacy, overcoming biases in training data, and maintaining human oversight highlight the need for careful implementation [1]. Additionally, integrating these technologies into clinical practice requires overcoming technical and cultural hurdles. A study that was recently published revealed that the LLM was found to amplify negative societal biases; overrepresent stereotypes, including problematic representations of minority groups; and exaggerate known disease prevalence differences between groups [36]. In their study, the authors noted that explicitly instructing the model to avoid bias or perform equitably is unlikely to produce the desired result and may even cause the model to overcorrect, resulting in an even worse bias [36,37]. Rather than relying solely on recommendations produced by the model, it appears essential that the model be connected to an independent, verifiable source of bias-free knowledge via retrieval-augmented generation [36,37]. Another major consideration is the likelihood of hallucinations. LLM is the general-purpose model. Prompt engineering, in which explicit instructions meant to exploit the optimal capabilities of LLMs are incorporated in addition to the question within the LLM input, can significantly improve LLM performance for specific tasks [38]. However, hallucinations may be common if we simply ask the question without any specific instructions. Fine-tuning or retrieval-augmented generation improves the goal-directedness of LLM [1]. Another limitation is the lack of clinical validation in the retrieved studies. Most selected studies are single-center pilot application formats with no large-scale studies or multicenter performance validation. Given that the performance of LLMs is constantly improving, this is likely to happen in the near future when they are ready for real-world clinical use.

Future Directions

LLMs have the potential to help both patients and doctors in the future, such as communicating professionally with patients in clinical settings or participating in the process of summarizing and reasoning about various multimodal information during clinical diagnosis. In addition, it is likely to be applied to medical research, such as hypothesizing, testing, and planning experiments, and finally, in education, it can play a role in training medical staff by acting as a fictitious patient or taking questions on various medical situations. It is necessary to prepare for wise use in the future by overcoming the limitations of misinformation, hallucination, and ethical issues.

In terms of the technical aspects, although, the transformer-based vision model (language vision model) needs much more data compared to CNN, this does not have inductive bias, which is the limitation of CNN [39]. There is a potential for LLM or language vision model to cover all the current performance of CNN's vision task in gastrointestinal endoscopy. These days, even LLMs with a focus on medicine are available [40,41]. Considering that medical practice is basically a multimodal task including history taking, visual diagnosis, data interpretation, and diagnosis reasoning, LLM-based foundation models that have multimodal function would be the next generation mainstream of AI model in medical practice [1]. LLMs, as a form of generative models, along with other generative models like generative adversarial networks, diffusion models, and variational autoencoders, are evolving to include creative features that are being integrated into the LLM framework. It is important to acknowledge that while LLMs are making strides in medical applications, issues such as model hallucination and unexpected biases remain challenges that require vigilant attention and ongoing research to mitigate. As we continue to witness the introduction of new LLM models with larger parameters and optimized performance, it is clear that these technologies are becoming indispensable tools in real-world clinical settings. Their potential to enhance medical practice is immense; yet, it is crucial to approach their integration with care to ensure patient safety and uphold ethical standards. The future directions of LLMs in gastroenterology, powered by advancements in processing units and innovative platforms, hold the promise of transforming the landscape of medical diagnostics and treatment, making health care more effective, personalized, and accessible to patients around the globe.

Limitation of This Systematic Review

Due to the different objectives and primary outcomes of each retrieved study, a meta-analysis (quantitative synthesis) was not feasible, which is a drawback of this systematic review process. Consequently, it was impossible to evaluate publication bias. A systematic review, however, showed that the main outcomes of each study divided by several categories, which LLMs are currently in use for the clinical research, which primary tasks are assessed, and what the limitations of each study are. However, in terms of future research perspectives,

XSL•FO
RenderX

the outcomes of this study can assist in comprehending the current state of gastroenterology research using LLMs and designing a study that complements the existing limitations.

Conclusions

The potential of LLMs to revolutionize gastrointestinal endoscopy is immense, offering improvements in diagnostic accuracy, operational efficiency, and patient care. The successful integration of LLMs hinges on addressing data privacy concerns, ensuring quality data, and fostering interdisciplinary collaboration. As we embrace these technological advancements, the future of endoscopy looks toward a more informed, efficient, and patient-centered approach.

Acknowledgments

This work was supported by the Gangwon Information & Culture Industry Promotion Agency and the National IT Industry Promotion Agency, a grant funded by the Korean government (MSIT; Diagnosis and prediction of digestive diseases and Development of precision medical AI questionnaire solution).

Data Availability

All data generated or analyzed during this study are included in this published paper.

Authors' Contributions

CSB conceptualized this study and participated in data curation, formal analysis, investigation, funding acquisition, methodology synthesis, project administration, gathering resources, writing the original draft, and reviewing and editing of the final manuscript. EJG participated in data curation, formal analysis, and writing the original draft. JJL, JP, EK, SK, SHC, and MK participated in data curation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analyses) 2020 checklist.
[DOCX File , 67 KB - jmir_v26i1e66648_app1.docx]

References

1. Kim HJ, Gong EJ, Bang CS. Application of machine learning based on structured medical data in gastroenterology. Biomimetics (Basel) 2023;8:512 [FREE Full text] [doi: 10.3390/biomimetics8070512] [Medline: 37999153]
2. Gong EJ, Bang CS. Interpretation of medical images using artificial intelligence: current status and future perspectives. Korean J Gastroenterol 2023;82:43-45. [doi: 10.4166/kjg.2023.071]
3. Bang CS. Artificial intelligence in the analysis of upper gastrointestinal disorders. Korean J Helicobacter Up Gastrointest Res 2021;21:300-310. [doi: 10.7704/kjhugr.2021.0030]
4. Gong EJ, Bang CS, Lee JJ, Jeong HM, Baik GH, Jeong JH, et al. Clinical decision support system for all stages of gastric carcinogenesis in real-time endoscopy: model establishment and validation study. J Med Internet Res 2023;25:e50448 [FREE Full text] [doi: 10.2196/50448] [Medline: 37902818]
5. Gong EJ, Bang CS, Lee JJ, Baik GH, Lim H, Jeong JH, et al. Deep learning-based clinical decision support system for gastric neoplasms in real-time endoscopy: development and validation study. Endoscopy 2023;55(8):701-708. [doi: 10.1055/a-2031-0691] [Medline: 36754065]
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. ArXiv. Preprint posted online on July 22, 2020 . [doi: 10.48550/arXiv.2005.14165]
7. Xie SM, Raghunathan A, Liang P, Ma T. An explanation of in-context learning as implicit Bayesian inference. ArXiv. Preprint posted online on July 21, 2022 [FREE Full text] [doi: 10.48550/arXiv.2111.02080]
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021;18(3):e1003583 [FREE Full text] [doi: 10.1371/journal.pmed.1003583] [Medline: 33780438]
9. Lim DYZ, Tan YB, Koh JTE, Tung JYM, Sng GGR, Tan DMY, et al. ChatGPT on guidelines: providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals. J Gastroenterol Hepatol 2024;39(1):81-106. [doi: 10.1111/jgh.16375] [Medline: 37855067]
10. Munir MM, Endo Y, Ejaz A, Dillhoff M, Cloyd JM, Pawlik TM. Online artificial intelligence platforms and their applicability to gastrointestinal surgical operations. J Gastrointest Surg 2024;28(1):64-69. [doi: 10.1016/j.gassur.2023.11.019] [Medline: 38353076]

XSL•FO
RenderX

11. Rammohan R, Joy MV, Magam SG, Natt D, Magam SR, Pannikodu L, et al. Understanding the landscape: the emergence of artificial intelligence (AI), ChatGPT, and Google Bard in gastroenterology. Cureus 2024;16(1):e51848 [FREE Full text] [doi: 10.7759/cureus.51848] [Medline: 38327910]
12. Tariq R, Malik S, Khanna S. Evolving landscape of large language models: an evaluation of ChatGPT and Bard in answering patient queries on colonoscopy. Gastroenterology 2024;166(1):220-221. [doi: 10.1053/j.gastro.2023.08.033] [Medline: 37634736]
13. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: cross-sectional evaluation of ChatGPT and Perplexity AI in medical residency exam. Dig Liver Dis 2024;56(8):1304-1311. [doi: 10.1016/j.dld.2024.02.019] [Medline: 38503659]
14. Atarere J, Naqvi H, Haas C, Adewunmi C, Bandaru S, Allamneni R, et al. Applicability of online chat-based artificial intelligence models to colorectal cancer screening. Dig Dis Sci 2024;69(3):791-797. [doi: 10.1007/s10620-024-08274-3] [Medline: 38267726]
15. Gorelik Y, Ghersin I, Arraf T, Ben-Ishay O, Klein A, Khamaysi I. Using a customized GPT to provide guideline-based recommendations for management of pancreatic cystic lesions. Endosc Int Open 2024;12(4):E600-E603 [FREE Full text] [doi: 10.1055/a-2289-9334] [Medline: 38681146]
16. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. Cureus 2023;15(10):e46736. [doi: 10.7759/cureus.46736] [Medline: 38022227]
17. Gorelik Y, Ghersin I, Maza I, Klein A. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. Gastrointest Endosc 2023;98(4):639-641.e4. [doi: 10.1016/j.gie.2023.06.025] [Medline: 37385548]
18. Ali H, Patel P, Obaitan I, Mohan BP, Sohail AH, Smith-Martinez L, et al. Evaluating the performance of ChatGPT in responding to questions about endoscopic procedures for patients. iGIE 2023;2(4):553-559. [doi: 10.1016/j.igie.2023.10.001]
19. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics (Basel) 2023;13(11):1950 [FREE Full text] [doi: 10.3390/diagnostics13111950] [Medline: 37296802]
20. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT answers common patient questions about colonoscopy. Gastroenterology 2023;165(2):509-511.e7 [FREE Full text] [doi: 10.1053/j.gastro.2023.04.033] [Medline: 37150470]
21. Samaan J, Issokson K, Feldman E. Artificial intelligence and patient education: examining the accuracy and reproducibility of responses to nutrition questions related to inflammatory bowel disease by GPT-4. medRxiv 2023 [FREE Full text] [doi: 10.1101/2023.10.28.23297723]
22. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA. Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management. Am J Gastroenterol 2023 Jul 10;118(12):2276-2279. [doi: 10.14309/ajg.0000000000002397] [Medline: 37410934]
23. Garcia P, Ma SP, Shah S, Smith M, Jeong Y, Devon-Sand A, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024;7(3):e243201 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
24. Syed S, Angel AJ, Syeda HB, Jennings CF, VanScoy J, Syed M, et al. The h-ANN model: comprehensive colonoscopy concept compilation using combined contextual embeddings. Biomed Eng Syst Technol Int Jt Conf BIOSTEC Revis Sel Pap 2022;5:189-200 [FREE Full text] [doi: 10.5220/0010903300003123] [Medline: 35373222]
25. Yeo YH, Samaan JS, Ng WH, Ma X, Ting PS, Kwak MS, et al. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. medRxiv 2023. [doi: 10.1101/2023.05.04.23289482]
26. Samaan JS, Yeo YH, Ng WH, Ting P, Trivedi H, Vipani A, et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab J Gastroenterol 2023;24(3):145-148. [doi: 10.1016/j.ajg.2023.08.001] [Medline: 37673708]
27. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci Rep 2023;13(1):4164 [FREE Full text] [doi: 10.1038/s41598-023-31412-2] [Medline: 36914821]
28. Zhou J, Li T, Fong SJ. Exploring ChatGPT's potential for consultation, recommendations and report diagnosis: gastric cancer and gastroscopy reports' case. Int J Interact Multimedia Artif Intell 2023;8:7-13 [FREE Full text]
29. Wang X, Xu X, Tong W, Liu Q, Liu Z. DeepCausality: a general AI-powered causal inference framework for free text: a case study of LiverTox. Front Artif Intell 2022;5:999289 [FREE Full text] [doi: 10.3389/frai.2022.999289] [Medline: 36561659]
30. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
31. Buzzaccarini G, Degliuomini RS, Borin M, Fidanza A, Salmeri N, Schiraldi L, et al. The promise and pitfalls of AI-generated anatomical images: evaluating midjourney for aesthetic surgery applications. Aesthetic Plast Surg 2024;48(9):1874-1883. [doi: 10.1007/s00266-023-03826-w] [Medline: 38238569]

XSL•FO
RenderX

32. Chin H, Song H, Baek G, Shin M, Jung C, Cha M, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023;25:e51712 [FREE Full text] [doi: 10.2196/51712] [Medline: 37862063]
33. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024 Nov 01;80(5):1158-1168 [FREE Full text] [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
34. Savage T, Wang J, Shieh L. A large language model screening tool to target patients for best practice alerts: development and validation. JMIR Med Inform 2023;11:e49886 [FREE Full text] [doi: 10.2196/49886] [Medline: 38010803]
35. Trinh TH, Wu Y, Le QV, He H, Luong T. Solving olympiad geometry without human demonstrations. Nature 2024;625(7995):476-482 [FREE Full text] [doi: 10.1038/s41586-023-06747-5] [Medline: 38233616]
36. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6(1):e12-e22 [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
37. Hastings J. Preventing harm from non-conscious bias in medical generative AI. Lancet Digit Health 2024;6(1):e2-e3 [FREE Full text] [doi: 10.1016/S2589-7500(23)00246-7] [Medline: 38123253]
38. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
39. Gong EJ, Bang CS. Revolutionizing gastrointestinal endoscopy: the emerging role of large language models. Clin Endosc 2024;57(6):759-762 [FREE Full text] [doi: 10.5946/ce.2024.039] [Medline: 39206500]
40. Chen Z, Hernández-Cano A, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. ArXiv. Preprint posted online on November 27, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.16079]
41. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. ArXiv. Preprint posted online on May 16, 2023 . [doi: 10.48550/arXiv.2305.09617]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
CNN: convolutional neural network
DILI: drug-induced liver injury
LLM: large language model
PICO: population, intervention, comparator, and outcome
PRISMA: Preferred Reporting Items for a Systematic Review and Meta-Analyses
ROBINS-I: Risk of Bias in Non-Randomized Studies—of Interventions

Edited by T de Azevedo Cardoso; submitted 19.09.24; peer-reviewed by D Shung, H Zhang; comments to author 01.11.24; revised version received 04.11.24; accepted 27.11.24; published 20.12.24.

Please cite as:
Gong EJ, Bang CS, Lee JJ, Park J, Kim E, Kim S, Kimm M, Choi SH
Large Language Models in Gastroenterology: Systematic Review
J Med Internet Res 2024;26:e66648
URL: https://www.jmir.org/2024/1/e66648
doi:10.2196/66648
PMID:

©Eun Jeong Gong, Chang Seok Bang, Jae Jun Lee, Jonghyung Park, Eunsil Kim, Subeen Kim, Minjae Kimm, Seoung-Ho Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Tracking Upper Limb Motion via Wearable Solutions: Systematic Review of Research From 2011 to 2023

Eirini Karoulla[1], MEng, MSc; Maria Matsangidou[2], MA, PhD; Fotos Frangoudes[2,3], BS, MSc; Panayiotis Paspalides[2], MEng; Kleanthis Neokleous[2], MEng, MSc, PhD; Constantinos S Pattichis[2,3,4], HTI, BSc, MSc, PhD

[1]CEA-List, Université Paris-Saclay, Gif-sur-Yvette, France
[2]CYENS - Centre of Excellence, Nicosia, Cyprus
[3]Department of Computer Science, University of Cyprus, Nicosia, Cyprus
[4]Biomedical Engineering Research Centre, University of Cyprus, Nicosia, Cyprus

Corresponding Author:
Maria Matsangidou, MA, PhD
CYENS - Centre of Excellence
Lellou Demetriades, Plateia Dimarchou 1
Nicosia, 1016
Cyprus
Phone: 357 22747575
Email: m.matsangidou@cyens.org.cy

Abstract

Background: The development of wearable solutions for tracking upper limb motion has gained research interest over the past decade. This paper provides a systematic review of related research on the type, feasibility, signal processing techniques, and feedback of wearable systems for tracking upper limb motion, mostly in rehabilitation applications, to understand and monitor human movement.

Objective: The aim of this article is to investigate how wearables are used to capture upper limb functions, especially related to clinical and rehabilitation applications.

Methods: A systematic literature search identified 27 relevant studies published in English from 2011 to 2023, across 4 databases: ACM Digital Library, IEEE Xplore, PubMed, and ScienceDirect. We included papers focusing on motion or posture tracking for the upper limbs, wearable devices, feedback given to end users, and systems having clinical or rehabilitation purposes. We excluded papers focusing on exoskeletons, robotics, prosthetics, orthoses, or activity recognition systems; reviews; and books.

Results: The results from this research focus on wearable devices that are designed to monitor upper limb movement. More specifically, studies were divided into 2 distinct categories: clinical motion tracking (15/27, 56%) and rehabilitation (12/27, 44%), involving healthy individuals and patients, with a total of 439 participants. Among the 27 studies, the majority (19/27) used inertial measurement units to track upper limb movement or smart textiles embedded with sensors. These devices were attached to the body with straps (mostly Velcro), providing flexibility and stability. The developed wearable devices positively influenced user motivation through the provided feedback, with visual feedback being the most common owing to the high level of independence provided. Moreover, a variety of signal processing techniques, such as Kalman and Butterworth filters, were applied to ensure data accuracy. However, limitations persist and include sensor positioning, calibration, and battery life, as well as a lack of clinical data on the effectiveness of these systems. The sampling rate of the data collection ranged from 50 Hz to 2000 Hz, which notably affected data quality and battery life. In addition, several findings were inconclusive, and thus, further future research is needed to understand and improve upper limb posture to develop progressive wearable systems.

Conclusions: This paper offers a comprehensive overview of wearable monitoring systems, with a focus on upper limb motion tracking and rehabilitation. It emphasizes the various types of available solutions; their efficacy, wearability, and feasibility; and proposed processing techniques. Finally, it presents robust findings regarding feedback accuracy derived from experiments and outlines potential future research directions.

(J Med Internet Res 2024;26:e51994) doi:10.2196/51994

XSL•FO
RenderX

KEYWORDS

motion tracking; motion sensing; posture monitoring; wearable devices; upper limb rehabilitation; interactive feedback; real-time feedback; wearble technology; upper limb motion

Introduction

Two out of the three leading conditions driving the need for rehabilitation are musculoskeletal and neurological disorders, with 1.71 billion and 255 million people, respectively, being affected globally. Musculoskeletal disorders have a significant impact on both individuals and society as a whole and incur a substantial economic cost. Thus, their effective and timely treatment through traditional rehabilitation approaches or digital health interventions is very important [1]. Carpal tunnel syndrome, rotator cuff tendonitis, and trigger finger are among the most common upper limb musculoskeletal dysfunctions that require methodical rehabilitation treatment [2]. People experiencing such disorders are often exposed to physical activities that require repetitive work [3].

Neurological diseases, on the other hand, not only affect the physical movement of patients but also influence their independence, living conditions, and quality of life. Usually, these disorders induce upper limb impairments, which often require conventional interventions to recover, such as physical or occupational therapy [4], as their rehabilitation depends on the duration, intensity, onset, and task orientation [5]. Stroke is the leading neurological disorder [6] and can potentially require long-term rehabilitation, with the affected person sometimes never achieving full recovery [7]. The high prevalence of stroke and the unique challenges it introduces have led to many studies about how technology can be used to assist affected populations and how to meet their needs [8].

Over the last 2 decades, technological devices have been included in rehabilitation programs within hospitals and treatment centers, which provide assistance and quantitative analysis during the rehabilitation process. Wearable devices have seen rapid development recently, with new technologies emerging constantly. The sensors used with wearables are getting smaller, more portable, and more energy efficient while, at the same time, having improved accuracy [9]. The prevention and rehabilitation of upper limb disorders are major problems that can benefit from the use of advanced wearable recovery systems. However, their integration and adoption in clinical practice are limited, and further actions are required toward this goal [6,10].

Therefore, efforts should be focused on developing systems that ensure that researchers can get reliable data for clinical evaluations. In this way, the variations in therapy and functional recovery can be better understood, aiming for the development of intervention strategies and the comprehension of the neuromuscular system [6]. Wearable devices are developed to solve these problems by providing affordable home-based solutions. It has been argued that inertial measurement units (IMUs) and surface electromyography (EMG) sensors are among the best options from the wide range of sensors currently available on the market that are used for the collection of data from wearable devices. These sensors offer a good balance between unobtrusiveness, robustness, and data quality [7].

This paper aims to provide a thorough review of how wearable-based monitoring systems are used for upper limb motion tracking and rehabilitation. More specifically, we aim to provide new insights into this area and analyze them in the following aspects: (1) assess the type and effectiveness of each wearable system; (2) assess the wearability and feasibility of the sensing technology; (3) clarify the signal processing techniques and extracted features; (4) classify the type and accuracy of the feedback according to the experiment results; and (5) review the findings, discuss limitations, and propose future directions.

Methods

Review Phases

The review was conducted based on the Bargas-Avila and Hornbæk approach [11] and Cochrane methodology [12], and it included 5 phases.

Phase 1: Potentially Relevant Publications Identified

Electronic Libraries

We searched 4 electronic libraries, which cover a balanced range of disciplines, including computer science and engineering, medical research, and multidisciplinary sources. The following libraries were included in the review: (1) ACM Digital Library, (2) IEEE Xplore, (3) PubMed, and (4) ScienceDirect. We restricted the search to a timeframe of 13 years (2011 to 2023).

Search Terms

The following queries were used: (1) “upper limb rehabilitation” AND “wearables,” and (2) (“posture monitoring” OR “motion monitoring”) AND “wearables.”

Search Procedure

The search terms were applied to the publication’s title, abstract, and keywords.

Inclusion Criteria

The inclusion criteria were as follows: (1) the paper concerns motion or posture tracking of the upper limbs, (2) the study focuses on wearable devices, (3) feedback is given to the end users, (4) the system considers clinical or rehabilitation purposes, and (5) the paper has been published in the last 13 years and is written in English.

Exclusion Criteria

The exclusion criteria were as follows: (1) robotic and exoskeleton systems, (2) prosthetics and orthoses, (3) activity recognition systems (activity/gesture or motion capture), and (4) reviews and books.

XSL•FO
RenderX

Search Results

The total number of search results from phase 1 was 1564 papers. More detailed results are presented in Table 1.

Table 1. Search results per library.

Query	Library (N=1564), n				
	ACM[a]	IEEE[b]	PM[c]	SD[d]	Total
Query 1[e]	338	145	309	78	870
Query 2[f]	228	99	67	300	694

[a]ACM: ACM Digital Library.
[b]IEEE: IEEE Xplore.
[c]PM: PubMed.
[d]SD: ScienceDirect.
[e]Query 1: “upper limb rehabilitation” AND “wearables.”
[f]Query 2: (“posture monitoring” OR “motion monitoring”) AND “wearables.”

Phase 2: Publications Retrieved for Detailed Evaluation

First Exclusion

All 1564 search results from phase 1 were imported into the software “Paperpile.” Duplicate entities were excluded manually. Overall, 96 duplicate publications were removed, and 1468 papers remained.

Second Exclusion

Publications with incomplete or restricted entries, those with no available full text, and those considered irrelevant based on the abstract were excluded manually. As a result, 1110 papers were removed.

Third Exclusion

We narrowed the entries to original full papers written in English. Part of this third exclusion was to remove entities that were not original full papers, such as workshops, posters, speeches, reviews, magazine articles, and generally grey literature without formal peer review. As a result, 51 papers were excluded. The remaining 307 papers included 255 journal articles, 45 conference papers, and 7 book chapters.

Phase 3: Publications to be Included in the Analysis

Final Exclusion

The focus of this review is on tracking upper limb motion via wearable solutions. Consequently, this final exclusion phase excluded studies that were not relevant based on a full-text review. Based on the exclusion criteria, we removed 280 irrelevant publications (eg, exoskeleton, robotics, and orthosis), and finally, 27 papers were selected for analysis.

Phase 4: Data Gathering

The screening process and data extraction were performed independently by 3 researchers (EK, MM, and PP). Discrepancies between reviewers were resolved through regular meetings and detailed discussions, addressing all the disagreements to minimize bias. In this phase, relevant information was extracted from the selected papers to conduct the analysis. From each study, the following information was extracted: target population, sample size of participants, sensor placement, type of the wearable system, type and feasibility of wearability, data sampling rate, energy consumption/battery characteristics, type and accuracy of feedback according to statistical analysis results, methodology, measurement techniques, instruments, key findings, limitations, and future directions.

Phase 5: Data Analysis

The data collected in phase 4 were analyzed using descriptive statistics. We then reviewed the literature to support and enhance the additional knowledge that this paper provides. Thematic analysis was used as an extra methodology to categorize our findings based on themes: (1) type and effectiveness of wearable systems, (2) wearability of sensing technology, (3) data processing and measurement techniques, and (4) type and accuracy of feedback.

Results

Study Characteristics

The review identified 27 studies (Figure 1). All the reviewed studies tracked upper limb motion using wearable devices and were divided into 2 main categories according to the purpose of the study: (1) clinical motion tracking and (2) rehabilitation. Some studies examined both applications. More specifically, the majority of papers that were selected (15/27, 56%) [13-27] focused on wearable devices that monitor human movement, and data were collected to be employed for general use in a clinical setting. The remaining papers (12/27, 44%) [28-39] emphasized systems for upper limb rehabilitation. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist is provided in Multimedia Appendix 1.

The key features of the studies are summarized and compared in Table 2, according to the purpose of the study, the type and number of sensors used, the placement location of the sensors, the placement method for the sensors, and the measurement method, with division into the 2 aforementioned categories.

XSL•FO
RenderX

Figure 1. Identification and selection flow diagram.

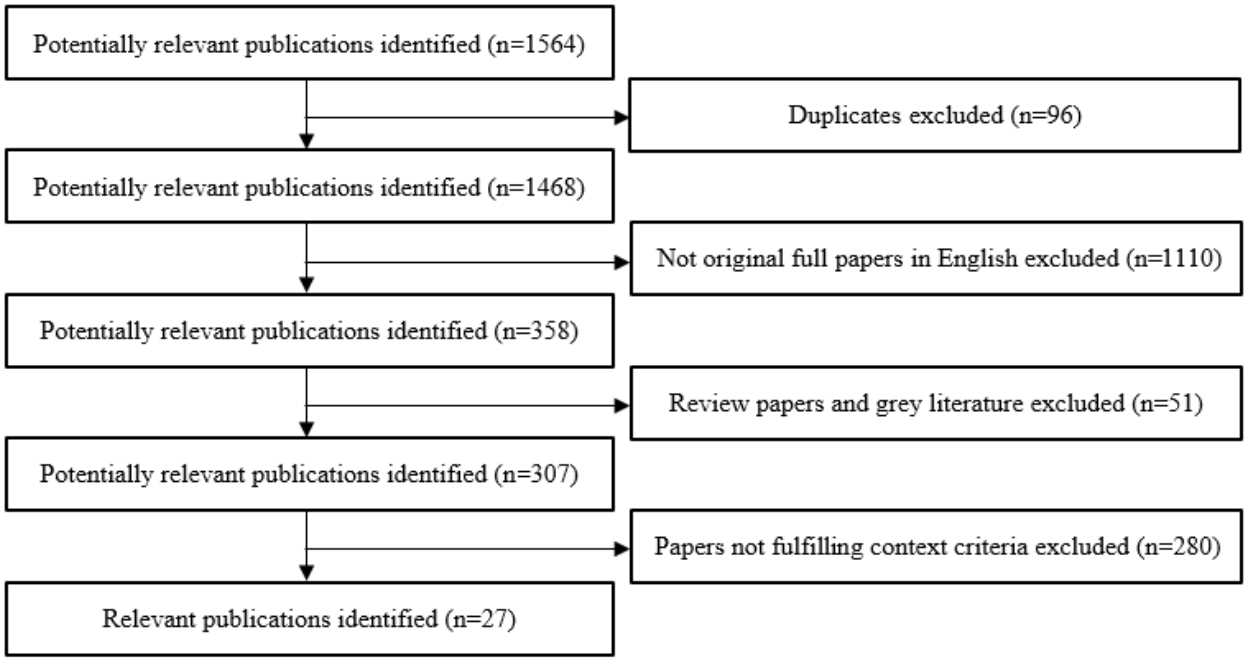

XSL•FO
RenderX

Table 2. Key features of the studies.

Content and study	Purpose	Technology (number of sensors)	Placement location	Placement method	Measurement/methodology
Motion tracking					
Yu et al [27]	Monitor upper limb motion function of stroke patients	Accelerometer (n=2)	Forearm, upper arm	Straps	ROM[a]: Bobath's handshake and shoulder touch
Wang et al [24]	Monitor compensatory movements and evaluate their applicability in a clinical setting	9-DOF[b] IMU[c] (n=2)	Shoulder, torso	Zipped vest, Velcro straps	Compensatory movement of the shoulder girdle
Li et al [18]	Evaluate upper limb motor function in hemiplegic patients	6-DOF IMU (n=2), EMG[d] (n=10)	Wrist, forearm, upper arm	Wristband, armband	Movements of major upper limb joints: shoulder, elbow, wrist, and finger joints
Repnik et al [20]	Quantify upper limb movement for muscle activity analysis in stroke patients	9-DOF IMU (n=7), EMG (n=2)	Hand, wrist, forearm, upper arm, sternum	Wristband, armband, straps	Movement quantified: hand smoothness, trajectory, trunk stability, and muscle activity
Tolvanen et al [23]	General motion tracking	Piezoresistive strain sensor (n=1)	Hand, wrist, forearm, upper arm (biceps, triceps)	Reusable adhesive layer	Opening-closing cycles (hand), muscle tension of the flexor sternum, biceps, upper arm, bicep curl, peck fly, and triceps
Bai et al [13]	Monitor different body movements, muscle contraction, and relaxation	GYS[e] sensor	Upper arm, fingers	Direct winding	ROM: instant tensing, bending, static motions of fingers, varied contractions of the bicep
Gu et al [15]	Identify hand motions, joint bending, hand posture, gesture, and sign language	Hydrogel-elastomer hybrid ionic sensor (n=10)	Hand, fingers	Water-borne adhesive	ROM: finger bending/extending and hand gestures
Lee et al [17]	Analyze data from neurologically intact individuals and the free-living environment, and develop a system to monitor stroke survivors	Accelerometer (n=4)	Wrists, fingers	Rings, wristbands	General quantification of the amount of use of upper limb function
Zhang P et al [26]	Monitor human movement with the use of a flexible resistance strain sensor with a porous structure	Strain sensor (n=1)	Upper arm, forearm, wrist, fingers	Velcro straps	ROM: finger wrist and elbow bending; responses to breathing
Lee et al [16]	Facilitate the clinically fitted measurement of fine-motor finger and wrist joint movements. Characterize age-related changes in hand functions	9-DOF IMU (n=7)	Wrist, hand, fingers	Clip-on straps	ROM: finger movement (index finger, thumb flexion/extension) and wrist movement (ulnar/radial flexion)
Zhang J et al [25]	Implement 3D motion velocity measurement, and propose a functional link artificial neural network model (FLANN)	Microthermal flow sensor (n=2)	Wrist	Straps	Trunk velocity, relative limb velocity, and absolute limb velocity
Schwarz et al [21]	Evaluate spatiotemporal kinematic metrics for the assessment of upper limb movements after stroke	6-DOF IMU (n=8)	Sternum, shoulder, upper arm, forearm, hand, fingers, thumb	Medical tape/3D-printed flexible straps	ROM: shoulder, elbow, thumb, index flexion/extension, and wrist supination/pronation
Formstone et al [14]	Develop a comprehensive system designed for the clinical environment, and quantify hand/wrist movement	9-DOF IMU (n=3), MMG[f] (n=2)	Sternum, upper arm, wrist, forearm	3D-printed housing cases attached with 3D-printed flexible resin straps	Shoulder twist angle (range), abduction, flexion, elbow twist angle, wrist flexion, and circumduction muscle activity
Little et al [19]	Analyze kinematic and physiological features for predicting elbow motion intention	9-DOF IMU (n=3), EMG (n=4), stretch sensor (n=1)	Forearm, upper arm, torso	Straps, direct winding	Muscle activity, elbow flexion angle, and custom-made changes in muscle volume

XSL•FO
RenderX

Content and study	Purpose	Technology (number of sensors)	Placement location	Placement method	Measurement/methodology
Schwerz de Lucena et al [22]	Real-time quantification of the effect of wearable feedback on hand counts for increasing hand activity	6-DOF IMU (n=1), magnetometer (n=4)	Wrist, fingers	Wristband, ring	"Hand counts": finger flexion/extension, wrist flexion/extension, and wrist radial/ulnar deviation movement
Rehabilitation					
Ding et al [39]	Measure orientation and correct arm posture using vibrotactile actuators for stroke rehabilitation patients and therapists	9-DOF IMU (n=2)	Forearm, upper arm	Velcro straps	Body segment posture; forearm and upper arm orientation; trajectory of upper arm's yaw, pitch, and roll; elbow angle; and forearm roll
Kim et al [31]	Wearable upper limb motion tracking method for stroke rehabilitation therapy at home	6-DOF IMU (n=2)	Wrist, upper arm	Velcro straps	ROM: position and orientation of the wrist and elbow joints. Accuracy of motion estimation and motion matching
Mohammadzadeh et al [34]	Develop and evaluate the feasibility of a wearable sensor-based motion-tracking system	6-DOF IMU (n=3)	Forearm, upper arm, sternum	Velcro straps	ROM: elbow joint angle
Ploderer et al [35]	Patient monitoring system to support occupational therapists in upper limb rehabilitation work with stroke patients	9-DOF IMU (n=3)	Shoulder, upper arm, wrist	Velcro straps, medical tape	ROM of each degree of freedom
Wang et al [28]	Evaluate garments equipped with sensors that support posture monitoring; used in upper extremity rehabilitation training of stroke patients	9-DOF IMU (n=3)	Scapula (shoulder blade), torso	Vest with Velcro straps	Analytical shoulder flexion, and analytical and functional elevation in the scapular plane
Salchow-Hömmen et al [37]	Part of a feedback-controlled hand neuroprosthesis for the rehabilitation of patients who experience motor impairment of the hand	9-DOF IMU (n=16)	Hand, fingers, forearm	Skin-friendly tape	ROM: combined abduction and flexion motion
Semjonova et al [38]	Evaluate the impact of the Double Aid (DAid) smart shirt; training process of patients with subacromial pain syndrome	Strain sensors (n=2)	Scapula	Commercial elastane-based fitness shirt	Perform exercise without moving the shoulders; detect movement or no movement of the shoulders
Friedman et al [29]	Nonobtrusive option for monitoring wrist and hand movement; needed for stroke rehabilitation and other applications	Triaxial magnetometer (n=2), accelerometer (n=1)	Wrist, fingers	Watch-like enclosure, small neodymium ring worn on the index finger	Accuracy of monitoring finger motion, wrist flexion/extension, and wrist ulnar/radial deviation. Accuracy in estimating different levels of movement activity
Kortier et al [32]	Ambulatory system using inertial sensors for hand kinematics, and evaluation of hand functioning	6-DOF IMU (n=15), 9-DOF IMU (n=6)	Hand, fingers, thumb	Double-sided adhesive tape/mounted on polyamide/elastane-fabricated glove	Static accuracy (ROM: flexion/extension, dynamic range, and repeatability)
Kim et al [30]	Identify optimal sensor locations	Bending sensor (n=2)	Thumb, hand	Flexible fabric straps partially made of Lycra sewed on a glove structure	ROM; circumduction motion
Liu et al [33]	Primary use of conductive stretchable fabrics to sense skin deformation during joint motion and infer the joint rotational angle	Strain sensor (n=2)	Forearm	Fabric	ROM: elbow flexion by various degrees; repeat motion at 3 levels of speed. Repeat each motion and perform free-form motions

XSL•FO
RenderX

Content and study	Purpose	Technology (number of sensors)	Placement location	Placement method	Measurement/methodology
Pregnolato et al [36]	Define the clinical features of stroke patients while performing hand movements for rehabilitation training	9-DOF IMU (n=1), EMG (n=1)	Forearm	Armband	Detect the total muscle activity of the forearm circumference

[a]ROM: range of motion.

[b]DOF: degrees of freedom.

[c]IMU: inertial measurement unit.

[d]EMG: electromyography.

[e]GYS: graphene thin-film yarn sensor.

[f]MMG: mechanomyography.

Study Type and Effectiveness

As mentioned above, wearable devices are used to capture upper limb function for general health purposes to obtain information and data about physiological parameters to assist rehabilitation patients and therapists. The experiments in most of the studies (22/27) [13-19,21,22,24-26,28,29,31-37,39] were noncontrolled experiments, where participants were either healthy able-bodied individuals or patients who experienced upper limb disorder, and they were recruited for participation in the experimental evaluation. In some studies (4/27) [20,27,30,38], the experiments were controlled, where both patients and healthy participants were included and the results were compared statistically. The number of participants was not specified in 1 study [23].

The findings revealed that most of the studies (22/27) [13-19,23-29,31,32,34-39] had a positive outcome. In 2 cases [22,30], further studies need to be conducted. In the first study, the maximum estimated error exceeded the required accuracy for a typical clinical assessment (over 5 degrees). In the second study, the hand feedback for stroke patients was briefly modifiable, indicating no therapeutic benefit over a short period. Thus, these studies are considered neutral. The outcomes of the remaining 3 studies [20,21,33] were negative as they do not offer any major findings [20] on motion tracking and the application is difficult to implement [21,33] due to limitations such as the system's measurement accuracy and large errors.

From the above information, it can be concluded that motion-tracking wearable devices have an overall positive impact on the interpretation of the data, and they provide useful assessments that can be used in future systems for clinical and rehabilitation purposes. The aforementioned data are presented in Table 3, categorized according to the purpose of the study.

XSL•FO
RenderX

Table 3. Study type and target population.

Content and study	Study type	Target population
Motion tracking		
Yu et al [27]	Controlled	23 stroke patients, 4 physicians
Wang et al [24]	Noncontrolled	8 musculoskeletal shoulder pain patients
Li et al [18]	Noncontrolled	16 healthy adults
Repnik et al [20]	Controlled	28 stroke patients, 14 healthy adults
Tolvanen et al [23]	Not specified	Not specified
Bai et al [13]	Noncontrolled	1 healthy adult
Gu et al [15]	Noncontrolled	1 healthy adult
Lee et al [17]	Noncontrolled	35 healthy adults
Zhang P et al [26]	Noncontrolled	1 healthy adult
Lee et al [16]	Noncontrolled	34 healthy adults
Zhang J et al [25]	Noncontrolled	1 healthy adult
Schwarz et al [21]	Noncontrolled	9 stroke patients
Formstone et al [14]	Noncontrolled	3 healthy adults
Little et al [19]	Noncontrolled	3 healthy adults
Schwerz de Lucena et al [22]	Noncontrolled	20 chronic stroke patients
Rehabilitation		
Ding et al [39]	Noncontrolled	5 healthy adults
Kim et al [31]	Noncontrolled	4 healthy adults
Mohammadzadeh et al [34]	Noncontrolled	8 healthy adults
Ploderer et al [35]	Noncontrolled	1st study: 8 occupational therapists; 2nd study: 1 healthy participant; 3rd study: 2 occupational therapists
Salchow-Hömmen et al [37]	Noncontrolled	4 healthy adults
Semjonova et al [38]	Controlled	17 primary subacromial pain syndrome patients and 17 healthy adults
Wang et al [28]	Noncontrolled	17 stroke patients
Friedman et al [29]	Noncontrolled	7 healthy adults
Kortier et al [32]	Noncontrolled	1st study: 1 participant; 2nd study: 1 participant; 3rd study: 5 participants
Kim et al [30]	Noncontrolled	10 participants for the optimal sensor location and 4 participants for experimental evaluation
Liu et al [33]	Noncontrolled	10 healthy adults
Pregnolato et al [36]	Noncontrolled	117 stroke adults

Wearability and Feasibility of Sensing Technologies

Categorization of Wearable Sensors

Capturing the motion of the upper limbs through sensor technologies has been essential for the development of interactive wearable devices for rehabilitation and clinical setting purposes. From the reviewed papers, 4 main categories of sensing technologies were identified: (1) inertial-based sensors (accelerometer, magnetometer, and IMU); (2) bending, force, and strain sensors (bending/stretch sensor and strain sensor [piezoresistive strain sensor and hydrogel-elastomer ionic sensor]); (3) myography sensors (EMG and mechanomyography [MMG]); and (4) other sensors (graphene thin-film yarn sensor [GYS] and microthermal flow sensor).

Figure 2 provides an overview of the number of sensing technologies that were used in the studies, according to the type of sensor.

XSL•FO
RenderX

Figure 2. Categorization of sensing technology. DOF: degrees of freedom; EMG: electromyography; GYS: graphene thin-film yarn sensor; IMU: inertial measurement unit; MMG: mechanomyography.

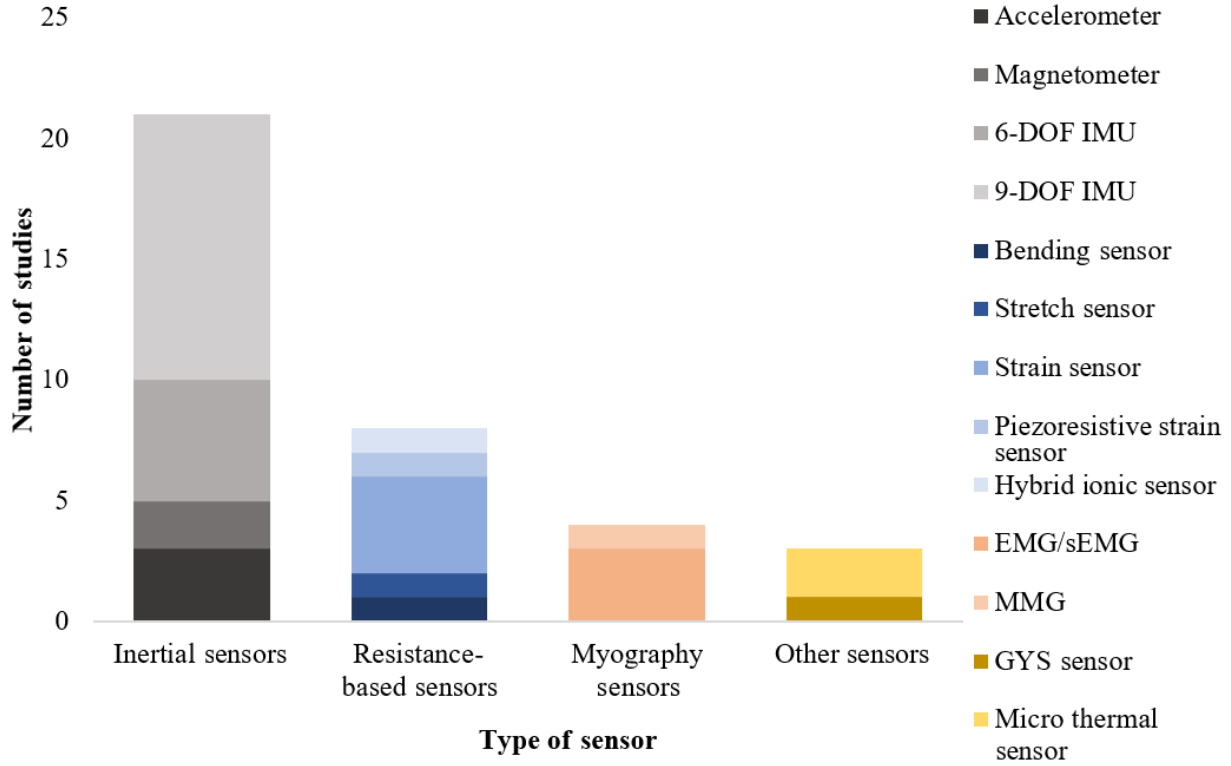

As Figure 2 shows, inertial sensors are the most common type of sensor used for data acquisition (19/27) [14,16-22,24,27-29,31,32,34-37,39]. An inertial sensor is an electronic device that measures the force and the angular rate of a body, which can be achieved by a combination of 3 embedded sensors: accelerometer, gyroscope, and magnetometer. It can also be used to calculate the orientation of the body. The accelerometer measures the proper acceleration, the gyroscope is used for measuring orientation and angular velocity, and finally, the magnetometer measures the strength and sometimes the direction of the magnetic field. Of the 19 studies, 14 [16,17,21,24,27-29,31,32,34,35,37-39] used a combination of the 3 sensors to measure upper limb posture. More specifically, in 6 studies [16,24,28,35,37,39], 9 degrees of freedom (DOF) inertial sensors were used, including a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer. In 3 studies [21,31,34], the magnetometer was excluded from the data fusion, and 6-DOF inertial sensors were proposed. Two of the studies [17,27] used only accelerometers as their sensing technology, while in 1 study [29], only a magnetometer was included. Moreover, a combination of 6-DOF and 9-DOF IMUs was used [32] to measure finger movement for the assessment of hand kinematics while using inertial sensors. Another way to capture wrist and finger motion was introduced by Schwerz de Lucena et al [22], where magnetometers were placed on the wrist to capture the magnetic field changes of the index finger, and the orientation of the wrist was quantified by a 6-DOF IMU. Furthermore, Pregnolato et al [36] used a combination of a 9-DOF IMU with EMG. Only the gyroscope of the inertial sensor was used to place the device on the patient's forearm, and EMG was used to detect the overall muscle activity in the circumference. The remaining 4 studies [14,18-20], proposed a fusion of inertial sensors with EMG or MMG [14,18,20] and a stretch sensor [19] to measure myographic data and changes in muscle volume, respectively.

In addition, the use of bending sensors and strain sensors was proposed in 2 papers [30,38]. This type of sensor was introduced owing to the limited available information on the actual impact of smart garments on clinical outcomes in physiotherapy [38] and the inaccurate measurement of thumb carpometacarpal joint movements [30]. Semjonova et al [38] proposed a purely textile-based smart shirt for the training process of patients with subacromial pain syndrome, while another study [30] presented a novel approach to identify optimal sensor locations to properly measure carpometacarpal joint configurations. Low-cost everyday fabrics were also introduced [33], which consist of stretchable conductive fabrics as strain sensors to sense skin deformations during elbow joint motion and infer the joint rotation angle. Furthermore, 1 study [23] proposed strain sensors along with a highly functional piezoresistive strain sensor, which was designed and fabricated exclusively because it provides excellent durability in human motion monitoring. Strain [26] and hydrogel-elastomer ionic [15] sensors are also used as they provide flexibility when they are worn, allowing the precise monitoring of upper limb movement and respiratory changes.

An alternative approach was also provided [13,25] as microthermal flow sensors and GYSs were introduced. The former method calculates velocity without integral calculation, and thus, accumulated errors are excluded. In the latter, the degree of resistance recovery and the gauge sensitivity can be

XSL•FO
RenderX

well controlled and modulated, providing high evaluability for developing next-generation wearable electronics.

The placement of each sensor can vary greatly, including the core of the body, across the arm, and on the fingers, according to the intended application. Table 4 provides information about the number and placement of sensors, and Figure 3 graphically shows their distribution for the studies examined in this paper.

Table 4. Studies that used sensors in each upper limb location.

Location	Studies	Number of studies
Scapula	[21,28]	2
Torso	[19,24,28]	3
Sternum	[14,20,21,34]	4
Shoulder	[21,24,35]	3
Upper arm	[13,14,18-21,23,26,27,31,34,35,39]	13
Forearm	[14,18-21,23,27,33,34,36,37,39]	13
Wrist	[14,16-18,20,22,23,25,26,29,31,35]	12
Hand	[15,16,20,21,23,30,32,37]	8
Thumb	[28,30,32]	3

Figure 3. Infograph of sensor placement. This model summarizes the placement of sensors on the upper body; however, the overall number of sensors or whether they are placed on the right or left hand has not been considered. DOF: degrees of freedom; EMG: electromyography; GYS: graphene thin-film yarn sensor; IMU: inertial measurement unit; MMG: mechanomyography.

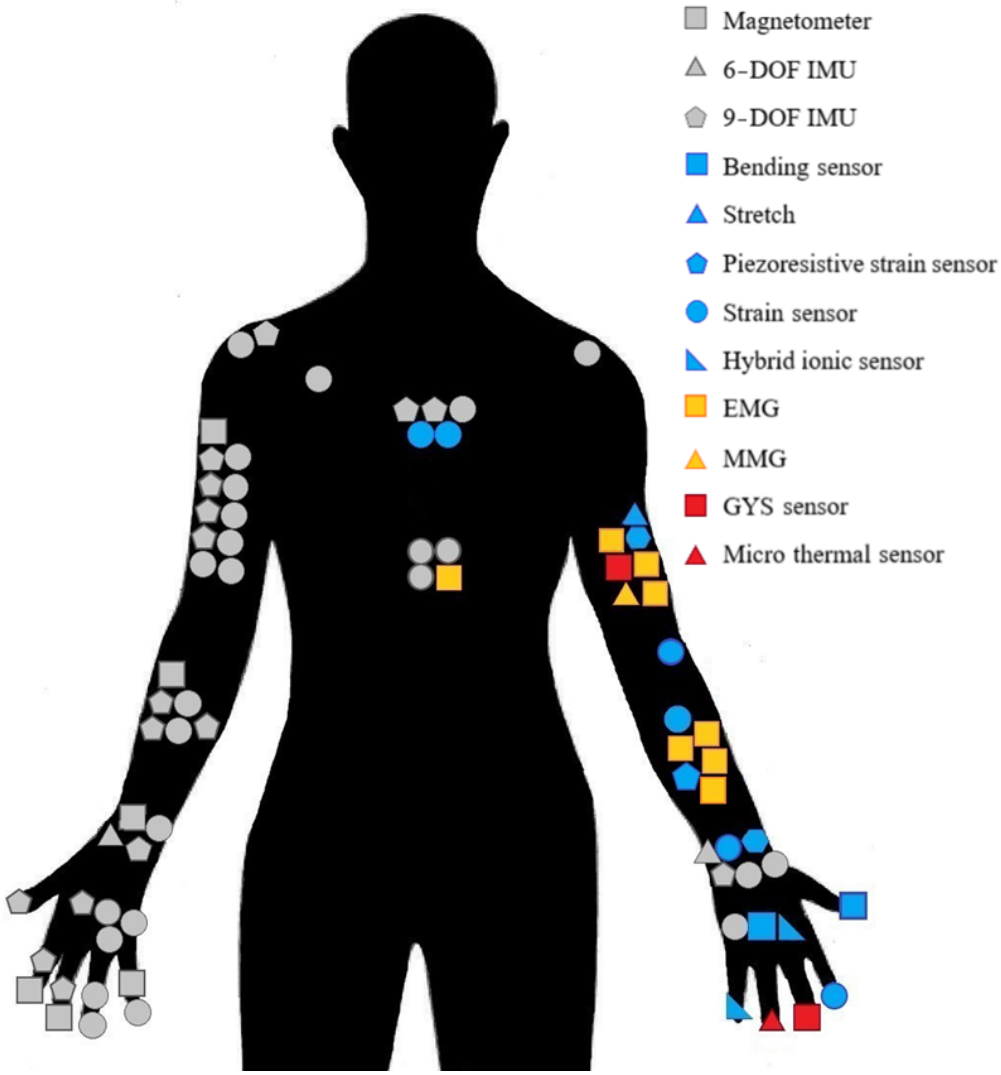

XSL•FO
RenderX

Among the 27 studies, 13 [13,14,18-21,23,26,27,31,34,35,39] involved sensor placement in the upper arm, 12 [14,16-18,20,22,23,25,26,29,31,35] involved placement on the wrist, and 13 [14,18-21,23,27,33,34,36,37,39] involved placement on the forearm. Many studies focused on finger movement monitoring. Sensors were placed on the hand in 8 studies [15,16,20,21,23,30,32,37], on the thumb in 3 studies [28,30,32], and on the shoulder in 3 studies [21,24,35]. In some cases, complex body posture and position were monitored. Sensors were placed at the sternum in 4 papers [14,20,21,34], at the torso in 3 papers [19,24,28], and at the scapula in 2 papers [21,28].

Placement Method and Feasibility of Sensors

One of the key factors considered for the development of upper limb wearable devices is the feasibility and wearability of the sensors, as in some cases, long-term monitoring is required. Consequently, the design of each system must be examined because it can greatly influence the feasibility of the system. From the reviewed papers, wearable systems can be classified into 4 categories, according to the attachment method of the sensors: (1) hook and loop straps (Velcro straps) and fastened straps; (2) bands; (3) adhesive bonding; and (4) other methods.

With regard to the methods used for placing the sensors, 13 out of the 27 studies [14,16,19,21,24-28,30,34,35,39] selected hook and loop straps. More specifically, 4 studies [26,31,34,39] used only a Velcro strap, which is a fastener that adheres to itself. Straps are generally a preferred type of placement method as they can be placed in convenient places on the arm [27] and they are simple, lightweight, and easy to use [25]. Clip-on straps have also been developed, which are flexible, and they allow inertial sensors to be placed on the fingers [16]. Furthermore, 5 studies [19,21,24,28,35] used a combination of hook and loop straps. Medical tape with 3D-printed flexible straps [21] or Velcro straps [35] were designed for upper limb assessment and rehabilitation of stroke patients, respectively. The former application was characterized by participants as being "comfortable to wear;" however, 3 of them reported impedance to grasp because of the finger sensors. In the latter, the design was a major issue as there is still work in the development of the sensors and their alignment. Straps along with direct winding of sensors in the armband [19] were also used as a method of placement; however, there is no discussion about the feasibility of the system. Another way to improve the feasibility of wearability is by designing a vest where different garment parts are attached with Velcro straps [24]. A zipped vest of soft material was developed, which makes it easier for patients to put on and take off. The precision of the sensor is guaranteed by a predefined position, and it is sewn by coated conductive yarn on a soft elastic strap with a Velcro strap fastened at the end. Overall, this system is perceived as highly usable, and the patients were motivated to train with it. Similarly, a garment embedded with smart textiles, conductive points, yarns, and sensors attached with elastic Velcro straps was also introduced [28]. This system is adjustable and more precise as sensors are placed in different positions, and it was rated as having high usability by users because it resembles everyday clothing in appearance and comfort while accurately tracking posture. Finally, flexible resin straps were also used [14] for attaching the housing cases of the sensors; however, the wearability of the system was not characterized.

Stretchable bands have also been designed [18] as they provide convenience and comfort around the arm without disturbing the subject's movement [20]. Furthermore, for hand and finger tracking, wristbands and rings were developed [16], and they were attached to the wrist and finger, respectively; however, the feasibility of the application was not discussed. For tracking the forearm, a study [36] used an armband to secure the IMU position; thus, the surface EMG acquisition remained the same for all patients. Feasibility was also not examined for this system.

Another method of attaching sensors to the body is adhesive bonding [15,23,32,37]. In this case, conductive sensors are included in the fabric and support wearable technologies. Their main function is to sense the physical movement of the arm and then transform it into electrical signals. Assessment of hand movement is necessary when evaluating hand function, and many glove-sensing systems lack rotational observability, hand orientation estimation, and user customization. Sensors can be mounted on a double-sided adhesive tape [23] as well as on a polyamide/elastane-fabricated glove [32], which consists of multiple printed circuit board strings that are attached to each finger segment. However, the feasibility of the system was not specified. Skin-friendly tape that attaches individual sensor strips adhesively to the finger segments and a silicon fixture that attaches the base unit of the system to the back of the hand make the system compact and portable. The sensor strips can be removed and replaced, thus increasing the flexibility in different therapy settings and different hand sizes, which makes the system more practical and easier to maintain [37]. A waterborne adhesive is an alternative method of attaching sensors to the hands as the system is fully integrated, but the result is not considered very feasible [15].

Regular elastic fabrics have also been used [33], where a prototype provided an acceptable comfort level and could be adapted to target users with different figures. A commercial elastane-based fitness shirt with elasticity was also used as a "vest," where sensors were attached by polychloroprene-based adhesive [38]. Comfort and elasticity were the key requirements for the base shirt to conform with the shape of the body of the user.

From the papers included in this review, only 1 study was conducted by using direct winding (the only placement method for attaching sensors to the body) [13]. GYSs were attached by direct winding on varied portions of the human body for monitoring different movements such as finger bending and muscle contraction and relaxation. A new direct-wearing mode, where the GYS can be directly attached to the skin, was introduced, which resulted in the improvement of sensing accuracy. Moreover, flexible fabric straps made of Lycra have been partially sewn on a glove structure [31] to fix the sensor in an optimal position. The feasibility and flexibility of this application are not discussed, and thus, the usability of the system cannot be concluded. A wearable device for tracking the daily use of the wrist and a finger was developed by designing a watch-like enclosure that is worn on the wrist with

XSL•FO
RenderX

a small neodymium magnetic ring on the index finger [29]. This design is described as "nonobtrusive" as the ring provides a reliable wireless signal without a power source, and thus, the need for bulky wiring or a battery is eliminated. Because of its design, the "Manumeter" is described as "socially acceptable," and it can be worn for long durations. In a more recent study [22], a newer version of the "Manumeter" was developed, where a jewelry-like device was fastened to the wrist with a band to monitor its movement. For finger movement, an ipsilateral finger ring was placed on the index finger. However, feasibility was not discussed.

Signal Processing Techniques and Extracted Features

Wearable upper limb devices provide important information about limb motion through motion analysis. Different signal processing techniques were used, and features were extracted to provide feedback to end users or therapists for better data assessment and interpretation of movement.

Signal processing is essential for such devices to reduce noise signals, transform data, and extract meaningful motion features. A variety of filters are used depending on the data provided by the sensors and the desired output of the system. In 13 studies [14,16-22,26,29,31,32,35], filters were used to remove noise from sensor signals. More specifically, the Kalman filter was used in 2 studies [20,32], which is an algorithm that uses observed measurements over time to produce estimated unknown variables that tend to be more accurate by estimating a joint probability distribution over the variable of each timeframe.

In 3 studies [18,21,32], the Butterworth filter was used to minimize low-frequency noise and high-frequency interferences. Another method to extract data from upper limb devices involves the infinite impulse response filter (IIR) [14,26]. This filter is digital with infinite impulse response and can be designed as a low-pass or high-pass filter. The Madgwick filter is an algorithm that reduces integration errors by the magnetometer and accelerometer, and it was used in 1 study [16]. A combination of 2 of the aforementioned filters was implemented in 1 study [21]. The accelerometer and gyroscope data were low-pass filtered using the Butterworth filter and then passed through the Madgwick filter to integrate the drift of the angular velocity. A complementary filter was implemented in 1 study [31] to reduce the integration drift by providing correct data on the orientation and position of each upper limb segment. More recently, research has been conducted on machine learning (ML) techniques to assist with predicting motion and improving tracking accuracy. For example, the Block Sparse Bayesian Learning (BSBL) algorithm was used in 1 study to reconstruct the accelerometer signal from compressed data [27]. Little et al [19], on the other hand, compared 10 different ML algorithms to predict angle trajectories through the fusion of physiological and kinematic features.

Another important feature of wearable devices is the sampling rates of the sensors, as they determine the quality of the captured data, providing a better understanding of upper limb motion. High sampling rates provide more precise data acquisition; however, battery life is significantly reduced as power consumption is increased. More specifically, in the revised studies, the smallest sampling rate was mentioned in the research conducted by Tolvanen et al [23], where a pressure sensitivity test for light finger touch by a stretch sensor was performed. In contrast, the highest sampling rate was at 2000 Hz in a paper by Pregnolato et al [36], where hand movements of stroke patients for rehabilitation were captured by a 9-DOF IMU with a surface EMG sensor. While both studies provide valuable insights into their respective applications, the battery life was not mentioned. With regard to the battery life, the shortest battery duration was observed in a study by Gu et al [15], where 10 hydrogel-elastomer hybrid ionic sensors were used to capture hand motion, joint bending, hand posture, and gesture. The battery life in this study was 1-2 hours, with an approximate sampling rate of 333.33 Hz, attributed to a 3.3-V rechargeable Li-ion battery. In contrast, the longest battery life was reported by Friedman et al [29], with a duration of 21.5 hours. Power was delivered to 2 triaxial magnetometers and 1 accelerometer to monitor wrist and hand movements, and the sensors were powered by a 3.7-V 450-mAh lithium polymer battery.

The relationship between battery size and sampling rate is crucial as it can affect the efficiency of the wearable sensors. Specifically, in the study conducted by Lee et al [17], a 170-mAh battery (sized 1-2 cm) was used to power an accelerometer with a sampling rate of 67 Hz, and the battery life was approximately 6 hours. In contrast, Wang et al [24] employed a larger 10-cm 3-V battery to power two 9-DOF IMUs with a sampling rate of 50 Hz. This variety of sampling rates and data acquisition reflects the diverse design choices between consumption and device portability in wearable sensor technology.

Type of Feedback and Accuracy of the System

Feedback is a key feature of the rehabilitation process and motion tracking as it provides important and meaningful information to therapists to interpret the performance of the system. Moreover, it plays an important role in informing patients of their progress, which leads to better recovery chances. The majority of studies (23/27) [13-23,25-27,29-35,37,38] used visual feedback. Three studies provided visual and auditory information [24,36] and haptic feedback along with visual and auditory information [24,28], and 1 study [39] provided only haptic feedback.

More specifically, visual feedback provides therapists and users with more precise information about their tasks and training instructions to achieve the desired position through direct visual analysis of the movement. Haptic feedback is provided by vibrotactile actuators, and small transducers are designed to optimize skin response to vibration. The vibrotactile actuators must have a minimum critical distance for their vibration to be identifiable, and the subjects of the experiment had to rely exclusively on it to accurately rehabilitate their posture and eventually regain their lost muscular abilities. Wearable devices that used a combination of feedback approaches (visual, audio, and haptic) provided an objective outcome that contributed to increasing the effectiveness of training. Moreover, these approaches provide support to therapists, giving them additional information about the patients' motions, and lastly, the quality of training is improved by detailed feedback signals.

XSL•FO
RenderX

Accuracy of system measurements is necessary for the effectiveness of proposed wearable devices as it plays a vital role in the interpretation of output data and the extraction of user features. Consequently, feedback should be quick enough to improve the operator's performance in terms of reducing mental effort and informing therapists about movement characteristics. One of the most common ways to present the accuracy of the system and evaluate position and velocity based on a kinematic model is through the calculation of the mean error between the bending joint angle of the user and the proposed validated algorithm with good accuracy, which is used as a reference, under different speeds and magnitudes [14,16,17,19,22,24,25,27,29-37]. Moreover, a statistical analysis is conducted by calculating the "root mean square deviation" (RMSD) or "root mean square error" (RMSE), correlation coefficient (CC), mean absolute error (MAE), and *P* value. In some papers, accuracy was either not measured or the circumstances under which accuracy was calculated were not mentioned [13,15,18,20,21,23,26,28,35,38,39].

Discussion

Principal Findings

This paper provides a systematic literature review of upper limb wearable device technologies that were developed in the past decade. The screening of papers from 4 different electronic libraries yielded a total of 27 relevant papers that were included for analysis. The papers were classified into 2 major categories, according to the purpose of the study: clinical motion tracking and rehabilitation. The analysis of the findings suggests that upper limb motion is most often tracked using inertial sensors owing to their accuracy, compact size, and effectiveness in assessing range of motion (ROM). Advanced data processing techniques, such as Kalman and Madgwick filters, were used for data fusion, ensuring data accuracy. One of the key elements of wearable technologies is usability, which can be affected by how wearables are placed on a user, their form factor, and their energy consumption. Even though most studies used straps for placing sensors, the other characteristics have greater variability, with no clear consensus among the research community.

The papers were categorized according to the type of study (controlled and uncontrolled). Most of them were uncontrolled studies, which may lead to bias because of the absence of randomly selected control groups and a comparison between them.

The goal of upper limb wearable devices is to provide assistance to therapists and researchers through either monitoring upper limb function for clinical assessment or aiding rehabilitation training and reducing the recovery time [40]. Overall, the developed wearable devices can positively influence the motivation of users, while in the case of rehabilitation, patients can undergo treatment at home by aiming for a high level of independence [41].

Various sensing technologies were used for tracking upper limb motion. Inertial sensors tend to be the most used sensors as they are used to estimate joint angles of the upper limbs. These technologies provide data accuracy, and because of their size, they are used to monitor and provide feedback to patients and therapists on ROM and rehabilitation performance. However, the placement method needs to be considered owing to its essential role in the ROM assessment, as the interpretation of data influences the development of rehabilitation treatment. Moreover, a flexible and well-fitted design can improve signal quality and reduce measurement noise [42]. Consequently, the majority of devices were attached to the body with straps (mostly Velcro straps), as they provide flexibility and great strength (eliminating data bias) and require low maintenance. The hold that Velcro straps provide can be generally characterized as "firm." Although this placement method is easy and noninvasive [39] and there is no need for external cameras, emitters, or markers [31], the feasibility of the system is not guaranteed for an extended period [25,39]. Similarly, clip-on straps can make the system modular in many aspects. This design can be adopted regardless of hand dimensions and the presence of deformities or inflammation. These straps are manufactured by using a stretchable and flexible material that yields extra comfort and causes minimum disruptions to movement [16].

The placement of wearable sensors significantly impacts their performance, user acceptance, and engineering demands. As sensor technology progresses from wearable to implantable and ingestible forms, challenges arise across regulatory, technical, and translational domains. Misplacement or misalignment of wearable sensors can reduce classifier accuracy; however, some approaches have maintained precision (97%) and recall (98%) at high levels during movement classification [7,9].

Despite the utility of straps, challenges remain in optimal sensor placement for maintaining user comfort and mobility and managing interference between sensors. Overcoming these challenges includes performing comprehensive user studies and data analyses to determine the best sensor placement, as well as designing an ergonomic and adaptable sensor housing for enhancing user comfort [43]. Thus, the ergonomic aspects of the system, such as dimension, weight, and undesirable contact, should be considered to accommodate various hand sizes and deformations [16]. Recent innovations have focused on achieving body compliance, ensuring comfort for the wearer, and maintaining accurate sensing performance [43].

Furthermore, some studies used smart textiles or e-textiles with embedded sensors as the sensing technology, which provided great feasibility regarding wearability. The use of textiles in health care and wellness applications has increased over the last few years and is expected to grow further in the future [41]. The primary benefit of using textile-based electrodes is that there is no direct contact with the skin, and this prevents problems like allergy and skin irritation [44]. Nevertheless, additional research needs to be conducted for integrating accuracy, improving usability, and implementing clinical validation. We expect that the benefits provided by textiles, especially related to ease of use and flexibility, and the advances in technology will make them essential for tracking motion and muscle activity and improving rehabilitation outcomes [44,45].

Wearable devices capture upper limb motion through data acquisition and processing. The examined studies used various

XSL•FO
RenderX

measurement methods to record data such as body segment posture, amount of use, and ROM. Undoubtedly, data processing is necessary for developing wearable devices as the signals captured have to be interpreted. The most common filters are Butterworth, low-pass, and band-pass, which are used to weaken potential high-frequency noise in the accelerometer and gyroscope. Moreover, Kalman, Madgwick, and complementary filters are used to fuse the sensor readings and overcome potential biomechanical constraints. These algorithms combine the sensor readings to indicate the rotation and orientation of the arm; however, they tend to require high computational power.

Sampling rates across the analyzed studies varied significantly and ranged from as low as 2 Hz for pressure-sensing technologies to as high as 2000 Hz for motion-sensing applications. Balancing high sampling rates and energy efficiency remains a challenge for upper limb wearable devices, as higher rates enhance tracking precision but can reduce battery life. The wide range of rates indicates the adaptability and flexibility of the sensors used for upper limb applications, as they capture muscle movements as well as rapid motion movements with precision. However, the selection of the right battery to support the desired sampling rates remains a challenge, and there is a need for careful selection to ensure prolonged and uninterrupted data acquisition.

Different techniques are being developed that aim to improve energy efficiency and extend the battery life of wearable devices. For example, compressed sensing [46] allows the reduction of the sampling rate of a signal, which can then be transmitted using a compressed sparse representation and reconstructed with minimal loss compared to the original signal [47,48]. Another algorithm that can be used with time series data, such as motion and muscle data, and can improve energy efficiency is change point detection [49]. The algorithm is used to detect the point when a signal changes and is often used to assist with action recognition [50]. Therefore, by detecting the time point when a signal changes, the sampling rate can be adapted and energy consumption can be reduced during nonrelevant activities (eg, transition between motions, resting, etc) [49]. At the same time, recently, research has been conducted on extending battery life through energy harvesting [51]. The process relies on capturing energy either directly from the human body, including motion [43], or from the environment and converting it to power.

Feedback plays a crucial role in the overall system performance as it provides useful information to therapists and researchers and affects therapy outcomes by influencing motivation. Feedback is given frequently to users who are less proficient or whose posture needs to be improved. This is beneficial for future applications as users should not rely on external feedback, but instead follow their intrinsic feedback mechanics. Visual information on a computer or smartphone/tablet is usually provided, which is very useful, especially for systems that are remotely monitored. Ultimately, the main aim of feedback is to help users improve their performance while providing useful data to therapists for better information processing and future reference.

The improved computation power of electronics and advances in ML algorithms have increased the use of these algorithms for not only motion trajectory prediction [52] but also assessing motion quality and providing relevant feedback during rehabilitation [53]. Using ML with wearable devices through Tiny Machine Learning (TinyML) has gained popularity recently; however, many challenges remain [54]. Further research in this area can revolutionize the use of wearable technology in health care applications by providing greater accuracy, and improved and more personalized feedback. Feedback can also be further reinforced using haptic devices [55]. However, current haptic devices can be cumbersome, and more portable devices need to be developed to make such technologies easy to use in a clinical setting [56].

Limitations and Conclusion

In this review, an effort was made to cover studies related to upper limb wearable devices, including study purpose, sensing technology, feasibility and wearability of the system, sensor placement, methodology, and feedback received. However, because of the variety of studies conducted in this area, every aspect could not be covered, and hence, a summary was provided for aspects with relatively more research. Limitations, such as the positioning, number, and possible disruption of sensors, are challenges that need to be overcome as they affect the limb's computed trajectory. Moreover, when a magnetometer is used, ferromagnetic materials can affect its calibration, and the data may not be transmitted with accuracy. Additionally, the lack of uniformity of battery specifications (compared to sampling rates) and sensor specifications highlights the challenges in comparisons between these characteristics and standardized data in this field. Although the studies reviewed indicated a positive influence regarding the motivation of users, more clinical trials need to be conducted, as they are important to assess the effectiveness of the system. Another limitation is the search strategy employed in this review. The search was performed in only 4 specific databases: ACM Digital Library, IEEE Xplore, PubMed, and ScienceDirect. The search may have excluded relevant papers, and some studies might have been overlooked. The low number of studies analyzed might not fully capture the diversity of the current research in the field, which limits the comprehensiveness of the review.

Future studies should aim to reduce the weight and dimensions of the system and increase the sampling rate, which can facilitate quick motion tracking with high accuracy. Additionally, efforts should be made to fabricate upper limb devices that are more flexible, powerful, and compact. The progress of ML algorithms would be beneficial, particularly in IMU- and EMG-based devices, as the rehabilitation process is automatically guided in real-life settings while being able to provide remote intervention [55]. Moreover, researchers should focus on the further development of feedback as it should be more adaptable and provide more options to users and therapists. In addition, long-term clinical trials are essential for establishing the effectiveness of wearable devices in real-world rehabilitation settings. These trials should focus on larger and more diverse patient groups to better establish the system's efficacy. Moreover, in broader health care frameworks, standardized protocols and consistent measurement methods should be

XSL•FO
RenderX

developed to ensure the results are accurately compared, thus ensuring the reliability of the system. Behavioral measurements are also important, especially in training sessions. In conclusion, future research should focus on integrating sensors and improving usability and feasibility as upper limb wearable devices are predictably becoming powerful tools, enabling innovative rehabilitation treatments while improving the quality of health care.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement number 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

Data Availability

All data generated or analyzed during this study are included in the published articles mentioned in the review.

Authors' Contributions

EK contributed to data collection, data analysis, data interpretation, and drafting the article. MM contributed to conception of the work, data collection, data analysis, data interpretation, and critical revision of the article. FF contributed to drafting and revising the article. PP contributed to data collection and analysis. KN contributed to critical revision of the article. CP contributed to critical revision of the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 550 KB - jmir_v26i1e51994_app1.pdf]

References

1. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020 Jun 05;22(6):e15617 [FREE Full text] [doi: 10.2196/15617] [Medline: 32501277]
2. Balogh I, Arvidsson I, Björk J, Hansson G, Ohlsson K, Skerfving S, et al. Work-related neck and upper limb disorders - quantitative exposure-response relationships adjusted for personal characteristics and psychosocial conditions. BMC Musculoskelet Disord 2019 Apr 01;20(1):139 [FREE Full text] [doi: 10.1186/s12891-019-2491-6] [Medline: 30935374]
3. Srinivasan D, Mathiassen SE. Motor variability in occupational health and performance. Clin Biomech (Bristol) 2012 Dec;27(10):979-993 [FREE Full text] [doi: 10.1016/j.clinbiomech.2012.08.007] [Medline: 22954427]
4. Orfanidis C, Darwich A, Cheong R, Fafoutis X. Monitoring neurological disorders with AI-enabled wearable systems. In: DigiBiom '22: Proceedings of the 2022 Workshop on Emerging Devices for Digital Biomarkers. 2022 Presented at: 20th Annual International Conference on Mobile Systems, Applications and Services; July 1, 2022; Portland, Oregon. [doi: 10.1145/3539494.3542755]
5. Maciejasz P, Eschweiler J, Gerlach-Hahn K, Jansen-Troy A, Leonhardt S. A survey on robotic devices for upper limb rehabilitation. J NeuroEngineering Rehabil 2014 Jan 09;11(1):3 [FREE Full text] [doi: 10.1186/1743-0003-11-3]
6. Vélez-Guerrero M, Callejas-Cuervo M, Mazzoleni S. Artificial intelligence-based wearable robotic exoskeletons for upper limb rehabilitation: a review. Sensors (Basel) 2021 Mar 18;21(6):2146 [FREE Full text] [doi: 10.3390/s21062146] [Medline: 33803911]
7. Maceira-Elvira P, Popa T, Schmid A, Hummel FC. Wearable technology in stroke rehabilitation: towards improved diagnosis and treatment of upper-limb motor impairment. J Neuroeng Rehabil 2019 Nov 19;16(1):142 [FREE Full text] [doi: 10.1186/s12984-019-0612-y] [Medline: 31744553]
8. van Ommeren AL, Smulders LC, Prange-Lasonder GB, Buurke JH, Veltink PH, Rietman JS. Assistive technology for the upper extremities after stroke: systematic review of users' needs. JMIR Rehabil Assist Technol 2018 Nov 29;5(2):e10510 [FREE Full text] [doi: 10.2196/10510] [Medline: 30497993]
9. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on wearable technology for connected health: scoping review of research trends, advances, and barriers. J Med Internet Res 2019 Sep 05;21(9):e14017 [FREE Full text] [doi: 10.2196/14017] [Medline: 31489843]
10. IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics. Piscataway, NJ: IEEE; 2012.

XSL•FO
RenderX

11. Bargas-Avila J, Hornbæk K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011 Presented at: SIGCHI Conference on Human Factors in Computing Systems; May 7-12, 2011; Vancouver, BC, Canada. [doi: 10.1145/1978942.1979336]
12. Becker L, Oxman A. Chapter 22: Overviews of reviews. Cochrane Handbook for Systematic Reviews of Interventions. URL: https://handbook-5-1.cochrane.org/chapter_22/22_overviews_of_reviews.htm [accessed 2024-11-27]
13. Bai M, Zhai Y, Liu F, Wang Y, Luo S. Stretchable graphene thin film enabled yarn sensors with tunable piezoresistivity for human motion monitoring. Sci Rep 2019 Dec 09;9(1):18644 [FREE Full text] [doi: 10.1038/s41598-019-55262-z] [Medline: 31819146]
14. Formstone L, Huo W, Wilson S, McGregor A, Bentley P, Vaidyanathan R. Quantification of motor function post-stroke using novel combination of wearable inertial and mechanomyographic sensors. IEEE Trans Neural Syst Rehabil Eng 2021;29:1158-1167. [doi: 10.1109/TNSRE.2021.3089613] [Medline: 34129501]
15. Gu G, Xu H, Peng S, Li L, Chen S, Lu T, et al. Integrated soft ionotronic skin with stretchable and transparent hydrogel-elastomer ionic sensors for hand-motion monitoring. Soft Robot 2019 Jun;6(3):368-376. [doi: 10.1089/soro.2018.0116] [Medline: 30848994]
16. Lee S, Lee H, Lee J, Ryu H, Kim IY, Kim J. Clip-on IMU system for assessing age-related changes in hand functions. Sensors (Basel) 2020 Nov 05;20(21):6313 [FREE Full text] [doi: 10.3390/s20216313] [Medline: 33167512]
17. Lee SI, Liu X, Rajan S, Ramasarma N, Choe EK, Bonato P. A novel upper-limb function measure derived from finger-worn sensor data collected in a free-living setting. PLoS One 2019;14(3):e0212484 [FREE Full text] [doi: 10.1371/journal.pone.0212484] [Medline: 30893308]
18. Li Y, Zhang X, Gong Y, Cheng Y, Gao X, Chen X. Motor function evaluation of hemiplegic upper-extremities using data fusion from wearable inertial and surface EMG sensors. Sensors (Basel) 2017 Mar 13;17(3):582 [FREE Full text] [doi: 10.3390/s17030582] [Medline: 28335394]
19. Little K, K Pappachan B, Yang S, Noronha B, Campolo D, Accoto D. Elbow motion trajectory prediction using a multi-modal wearable system: a comparative analysis of machine learning techniques. Sensors (Basel) 2021 Jan 12;21(2):498 [FREE Full text] [doi: 10.3390/s21020498] [Medline: 33445601]
20. Repnik E, Puh U, Goljar N, Munih M, Mihelj M. Using inertial measurement units and electromyography to quantify movement during action research arm test execution. Sensors (Basel) 2018 Aug 22;18(9):2767 [FREE Full text] [doi: 10.3390/s18092767] [Medline: 30135413]
21. Schwarz A, Bhagubai MMC, Wolterink G, Held JPO, Luft AR, Veltink PH. Assessment of upper limb movement impairments after stroke using wearable inertial sensing. Sensors (Basel) 2020 Aug 24;20(17):2767 [FREE Full text] [doi: 10.3390/s20174770] [Medline: 32846958]
22. Schwerz de Lucena D, Rowe JB, Okita S, Chan V, Cramer SC, Reinkensmeyer DJ. Providing real-time wearable feedback to increase hand use after stroke: a randomized, controlled trial. Sensors (Basel) 2022 Sep 14;22(18):6938 [FREE Full text] [doi: 10.3390/s22186938] [Medline: 36146287]
23. Tolvanen J, Hannu J, Jantunen H. Stretchable and washable strain sensor based on cracking structure for human motion monitoring. Sci Rep 2018 Sep 05;8(1):13241. [doi: 10.1038/s41598-018-31628-7] [Medline: 30185926]
24. Wang Q, De Baets L, Timmermans A, Chen W, Giacolini L, Matheve T, et al. Motor control training for the shoulder with smart garments. Sensors (Basel) 2017 Jul 22;17(7):1687 [FREE Full text] [doi: 10.3390/s17071687] [Medline: 28737670]
25. Zhang J, Liu S, Zhang Y, Zhu R. A method to extract motion velocities of limb and trunk in human locomotion. IEEE Access 2020;8:120553-120561. [doi: 10.1109/access.2020.3006336]
26. Zhang P, Chen Y, Li Y, Zhang Y, Zhang J, Huang L. Correction: Zhang, P., et al. A flexible strain sensor based on the porous structure of a carbon black/carbon nanotube conducting network for human motion detection. Sensors 2020, 20, 1154. Sensors (Basel) 2020 May 20;20(10):2901 [FREE Full text] [doi: 10.3390/s20102901] [Medline: 32443818]
27. Yu L, Xiong D, Guo L, Wang J. A compressed sensing-based wearable sensor network for quantitative assessment of stroke patients. Sensors (Basel) 2016 Feb 05;16(2):202 [FREE Full text] [doi: 10.3390/s16020202] [Medline: 26861337]
28. Wang Q, Timmermans A, Chen W, Jia J, Ding L, Xiong L, et al. Stroke patients' acceptance of a smart garment for supporting upper extremity rehabilitation. IEEE J. Transl. Eng. Health Med 2018;6:1-9. [doi: 10.1109/jtehm.2018.2853549]
29. Friedman N, Rowe JB, Reinkensmeyer DJ, Bachman M. The manumeter: a wearable device for monitoring daily use of the wrist and fingers. IEEE J Biomed Health Inform 2014 Nov;18(6):1804-1812. [doi: 10.1109/JBHI.2014.2329841] [Medline: 25014974]
30. Kim DH, Lee SW, Park H. Improving kinematic accuracy of soft wearable data gloves by optimizing sensor locations. Sensors (Basel) 2016 May 26;16(6):766 [FREE Full text] [doi: 10.3390/s16060766] [Medline: 27240364]
31. Kim J, Yang S, Gerla M. StrokeTrack: wireless inertial motion tracking of human arms for stroke telerehabilitation. In: mHealthSys '11: Proceedings of the First ACM Workshop on Mobile Systems, Applications, and Services for Healthcare. 2011 Presented at: 9th ACM Conference on Embedded Network Sensor Systems; November 1, 2011; Seattle, Washington. [doi: 10.1145/2064942.2064948]
32. Kortier HG, Sluiter VI, Roetenberg D, Veltink PH. Assessment of hand kinematics using inertial and magnetic sensors. J Neuroeng Rehabil 2014 Apr 21;11:70 [FREE Full text] [doi: 10.1186/1743-0003-11-70] [Medline: 24746123]

XSL•FO
RenderX

33. Liu R, Shao Q, Wang S, Ru C, Balkcom D, Zhou X. Reconstructing human joint motion with computational fabrics. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Mar 29;3(1):1-26. [doi: 10.1145/3314406]
34. Mohammadzadeh FF, Liu S, Bond KA, Nam CS. Feasibility of a wearable, sensor-based motion tracking system. Procedia Manufacturing 2015;3:192-199. [doi: 10.1016/j.promfg.2015.07.128]
35. Ploderer B, Fong J, Withana A, Klaic M, Nair S, Crocher V, et al. ArmSleeve: A Patient Monitoring System to Support Occupational Therapists in Stroke Rehabilitation. In: DIS '16: Proceedings of the 2016 ACM Conference on Designing Interactive Systems. 2016 Presented at: DIS '16: Designing Interactive Systems Conference; June 4-8, 2016; Brisbane, QLD, Australia. [doi: 10.1145/2901790.2901799]
36. Pregnolato G, Rimini D, Baldan F, Maistrello L, Salvalaggio S, Celadon N, et al. Clinical features to predict the use of a sEMG wearable device (REMO) for hand motor training of stroke patients: a cross-sectional cohort study. Int J Environ Res Public Health 2023 Mar 14;20(6):5082 [FREE Full text] [doi: 10.3390/ijerph20065082] [Medline: 36981992]
37. Salchow-Hömmen C, Callies L, Laidig D, Valtin M, Schauer T, Seel T. A tangible solution for hand motion tracking in clinical applications. Sensors (Basel) 2019 Jan 08;19(1):208 [FREE Full text] [doi: 10.3390/s19010208] [Medline: 30626130]
38. Semjonova G, Vetra J, Cauce V, Oks A, Katashev A, Eizentals P. Improving the recovery of patients with subacromial pain syndrome with the DAid Smart Textile Shirt. Sensors (Basel) 2020 Sep 15;20(18):5277 [FREE Full text] [doi: 10.3390/s20185277] [Medline: 32942730]
39. Ding ZQ, Luo ZQ, Causo A, Chen IM, Yue KX, Yeo SH, et al. Inertia sensor-based guidance system for upperlimb posture correction. Med Eng Phys 2013 Feb;35(2):269-276. [doi: 10.1016/j.medengphy.2011.09.002] [Medline: 21978912]
40. Simpson LA, Menon C, Hodgson AJ, Ben Mortenson W, Eng JJ. Clinicians' perceptions of a potential wearable device for capturing upper limb activity post-stroke: a qualitative focus group study. J Neuroeng Rehabil 2021 Sep 08;18(1):135 [FREE Full text] [doi: 10.1186/s12984-021-00927-y] [Medline: 34496894]
41. Toh SFM, Gonzalez PC, Fong KNK. Usability of a wearable device for home-based upper limb telerehabilitation in persons with stroke: A mixed-methods study. Digit Health 2023 Feb 07;9:20552076231153737 [FREE Full text] [doi: 10.1177/20552076231153737] [Medline: 36776407]
42. Khan Y, Ostfeld AE, Lochner CM, Pierre A, Arias AC. Monitoring of vital signs with flexible and wearable medical devices. Adv Mater 2016 Jun;28(22):4373-4395. [doi: 10.1002/adma.201504366] [Medline: 26867696]
43. Beniwal R, Kalra S, Beniwal NS, Mazumdar H, Singhal AK, Singh SK. Walk - to - charge technology: exploring efficient energy harvesting solutions for smart electronics. Journal of Sensors 2023 Oct 10;2023(1):614658. [doi: 10.1155/2023/6614658]
44. Vidhya CM, Maithani Y, Singh JP. Recent advances and challenges in textile electrodes for wearable biopotential signal monitoring: a comprehensive review. Biosensors (Basel) 2023 Jun 26;13(7):679 [FREE Full text] [doi: 10.3390/bios13070679] [Medline: 37504078]
45. Meena JS, Choi SB, Jung S, Kim J. Electronic textiles: New age of wearable technology for healthcare and fitness solutions. Mater Today Bio 2023 Apr;19:100565 [FREE Full text] [doi: 10.1016/j.mtbio.2023.100565] [Medline: 36816602]
46. Candes E, Wakin M. An introduction to compressive sampling. IEEE Signal Process Mag 2008 Mar;25(2):21-30. [doi: 10.1109/msp.2007.914731]
47. Pagan J, Fallahzadeh R, Pedram M, Risco-Martin J, Moya J, Ayala J, et al. Toward ultra-low-power remote health monitoring: an optimal and adaptive compressed sensing framework for activity recognition. IEEE Trans on Mobile Comput 2019 Mar 1;18(3):658-673. [doi: 10.1109/tmc.2018.2843373]
48. Kerdjidj O, Ramzan N, Ghanem K, Amira A, Chouireb F. Fall detection and human activity classification using wearable sensors and compressed sensing. J Ambient Intell Human Comput 2019 Jan 31;11(1):349-361. [doi: 10.1007/s12652-019-01214-4]
49. Culman C, Aminikhanghahi S, J Cook D. Easing power consumption of wearable activity monitoring with change point detection. Sensors (Basel) 2020 Jan 06;20(1):310 [FREE Full text] [doi: 10.3390/s20010310] [Medline: 31935907]
50. Bermejo U, Almeida A, Bilbao-Jayo A, Azkune G. Embedding-based real-time change point detection with application to activity segmentation in smart home time series data. Expert Systems with Applications 2021 Dec;185:115641. [doi: 10.1016/j.eswa.2021.115641]
51. Rong G, Zheng Y, Sawan M. Energy solutions for wearable sensors: a review. Sensors (Basel) 2021 May 31;21(11):3806 [FREE Full text] [doi: 10.3390/s21113806] [Medline: 34072770]
52. Rudenko A, Palmieri L, Herman M, Kitani K, Gavrila D, Arras K. Human motion trajectory prediction: a survey. The International Journal of Robotics Research 2020 Jun 07;39(8):895-935. [doi: 10.1177/0278364920917446]
53. Frangoudes F, Matsangidou M, Schiza E, Neokleous K, Pattichis C. Assessing human motion during exercise using machine learning: a literature review. IEEE Access 2022;10:86874-86903. [doi: 10.1109/access.2022.3198935]
54. Diab M, Rodriguez-Villegas E. Embedded machine learning using microcontrollers in wearable and ambulatory systems for health and care applications: a review. IEEE Access 2022;10:98450-98474. [doi: 10.1109/access.2022.3206782]
55. Huang Y, Yao K, Li J, Li D, Jia H, Liu Y, et al. Recent advances in multi-mode haptic feedback technologies towards wearable interfaces. Materials Today Physics 2022 Jan;22:100602. [doi: 10.1016/j.mtphys.2021.100602]
56. Yin J, Hinchet R, Shea H, Majidi C. Wearable soft technologies for haptic sensing and feedback. Adv Funct Materials 2020 Dec 31;31(39):7428. [doi: 10.1002/adfm.202007428]

XSL•FO
RenderX

Abbreviations

DOF: degrees of freedom
EMG: electromyography
GYS: graphene thin-film yarn sensor
IMU: inertial measurement unit
ML: machine learning
MMG: mechanomyography
ROM: range of motion

Edited by T de Azevedo Cardoso; submitted 20.08.23; peer-reviewed by S Okita, L Guo; comments to author 17.03.24; revised version received 31.05.24; accepted 16.10.24; published 23.12.24.

Please cite as:
Karoulla E, Matsangidou M, Frangoudes F, Paspalides P, Neokleous K, Pattichis CS
Tracking Upper Limb Motion via Wearable Solutions: Systematic Review of Research From 2011 to 2023
J Med Internet Res 2024;26:e51994
URL: https://www.jmir.org/2024/1/e51994
doi:10.2196/51994
PMID:

©Eirini Karoulla, Maria Matsangidou, Fotos Frangoudes, Panayiotis Paspalides, Kleanthis Neokleous, Constantinos S Pattichis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Characteristics and Outcomes of mHealth Interventions in Psychosis: Systematic Mapping Review

Pei Yi Loh[1]; Laura Martinengo[1,2], MD, PhD; Creighton Heaukulani[3], PhD; Xin Yang Tan[3], MBBS, MRCS; Moses Hng[4], BSc; Yong Yin Cheah[4], BA; Robert J T Morris[3,5], PhD; Lorainne Tudor Car[6*], MD, PhD; Jimmy Lee[1,4*], MBBS, MMed, MCI

[1]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

[2]Centre for Behavioural and Implementation Sciences Interventions, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Office for Healthcare Transformation, Ministry of Health, Singapore, Singapore

[4]Institute of Mental Health, Singapore, Singapore

[5]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[6]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Jimmy Lee, MBBS, MMed, MCI
Institute of Mental Health
10 Buangkok View
Singapore, 539747
Singapore
Phone: 65 63892000
Email: jimmy_lee@imh.com.sg

Abstract

Background: Mobile health (mHealth) interventions have gained popularity in augmenting psychiatric care for adults with psychosis. Interest has grown in leveraging mHealth to empower individuals living with severe mental illness and extend continuity of care beyond the hospital to the community. However, reported outcomes have been mixed, likely attributed in part to the intervention and adopted outcomes, which affected between-study comparisons.

Objective: This study aimed to critically review outcome measures used to evaluate mHealth interventions for adults with psychosis in relation to the characteristics of mHealth interventions.

Methods: A systematic mapping review was conducted. We searched PubMed, CINAHL, Embase, PsycINFO, and Cochrane Libraries from 1973 to the present. Selection criteria included randomized controlled studies of mHealth interventions in adults diagnosed with schizophrenia spectrum disorders. Reviewers worked in pairs to screen and extract data from included studies independently using a standardized form; disagreements were resolved by consensus with an independent reviewer. We report our findings in line with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines.

Results: A total of 1703 citations were screened; 29 publications reporting on 23 studies were included in this review. mHealth interventions for psychosis span a wide range, with psychological therapy being the most-deployed intervention (12/23, 52%), followed by psychoeducation (8/23, 35%) and active self-monitoring (8/23, 35%). Several mHealth interventions for psychosis targeted multiple pillars of biopsychosocial well-being (10/23, 43%); the bulk of interventions (16/23, 70%) incorporated features promoting users' self-management. The majority of mHealth interventions were delivered through applications (14/23, 61%) as the main medium and smartphones (17/23, 74%) as the main channel of delivery. Interventions were primarily administered in the outpatient and community settings (16/23, 70%); many were also blended with in-person sessions (11/23, 48%) or guided remotely (6/23, 26%) by persons, including health care providers or trained peer supporters. The severity of psychosis-related symptoms (21/23, 91%) was the most prevalent outcome, of which positive symptoms (13/23, 57%), mood and anxiety (10/23, 43%), and overall psychopathology severity (9/23, 39%) were most commonly measured. Patient-centric outcomes, including well-being (17/23, 74%)—particularly quality of life (10/23, 43%)—and user experience (15/23, 65%), including feasibility

XSL•FO
RenderX

(7/23, 30%), acceptability (7/23, 30%), and engagement (7/23, 26%). Notably, outcome choices remained diverse despite stratification by type of mHealth intervention.

Conclusions: mHealth interventions for psychosis encompass a wide range of modalities and use outcome measures that probe various social and behavioral determinants of health. These should be considered complex interventions, and a holistic evaluation approach combining clinical and patient-centric outcomes is recommended.

(J Med Internet Res 2024;26:e55924) doi:10.2196/55924

KEYWORDS

mHealth; digital health; mobile applications; psychosis; schizophrenia; schizophrenia spectrum; psychotic disorders; mapping review

Introduction

Psychotic disorders are a group of serious mental illnesses consisting of "abnormalities in one or more of the following five domains: delusions, hallucinations, disorganized thinking (speech), grossly disorganized or abnormal motor behavior (including catatonia), and negative symptoms" [1]. The global burden of psychotic disorders on individuals, communities, and health care systems cannot be understated. Schizophrenia, the most well-known psychotic disorder, has been ranked among the top 20 causes of years lived with disability (YLDs) among all illnesses and injuries and afflicts 23.6 million individuals worldwide [2]. Schizophrenia also has an early age of onset in adolescence [3] and has a propensity for relapsing. Such acute relapses can leave deleterious impacts on the individual yet are difficult to predict. The chronic, relapsing-remitting illness trajectory of schizophrenia not only contributes to a high disability weight during acute psychotic episodes [4] but also incurs hefty societal costs—upwards of US $300 billion in the United States alone [5].

The debilitating nature of psychosis thus spells the need for scalable, cost-effective, and accessible solutions to augment traditional psychiatric management, which remains underpinned by intensive human-delivered care. Mobile health (mHealth), defined as the use of wireless mobile technologies for public health [6], stands out as an emerging possibility. This is especially so with smartphone ownership becoming more pervasive in the 21st century, including among persons living with psychotic disorders [7]. In an era where different smart devices permeate everyday life, mHealth can enable data-driven assessment of individuals' lifestyles and well-being. The repertoire of personalized mHealth interventions has similarly been expanding, ranging from medication adherence tools [8] to internet-based cognitive behavioral therapy (iCBT) [9]. mHealth interventions, therefore, harbor much optimism in empowering persons living with psychosis toward proactive self-care with timely symptom management and targeted interventions [10].

With the exponential increase in popularity of mHealth in recent years [11], there is substantial interest in evaluating the efficacy of mHealth interventions for adult patients with psychosis. However, to the best of our knowledge, there are no systematic reviews evaluating mHealth interventions for adult patients with psychosis. A systematic meta-review of mHealth interventions for mental health in general [12] found that none of the included meta-analyses studied their effects on psychotic disorders. There have been a few reviews looking into mHealth technologies for psychosis, but these focused on the scope of technologies rather than the outcomes these technologies seek to achieve [13-15]. Reviews attempting to quantify the impact of mHealth interventions on psychosis outcomes included few articles at the time of publication [16], which may no longer represent the current body of mHealth interventions. Furthermore, Firth and Torous [17] assessed the impact of mHealth intervention only for feasibility, while Clarke et al [18] focused on their effectiveness in reducing psychotic symptoms without assessing other patient-centric outcomes.

It is therefore important to critically review prevailing outcome measures used to evaluate mHealth interventions for adults with psychosis, whether in terms of traditional clinical outcomes (eg, relapse prevention, reduction of hospitalization, or mortality) or person-centric attributes (eg, quality of life, subjective well-being, and various psychological constructs). This can be achieved through a systematic mapping review to "collate, describe and catalogue available evidence" while following "the same rigorous, objective and transparent processes as do systematic reviews" [19]. Although systematic reviews of randomized controlled trials (RCTs) are the perceived "gold standard," there are significant hurdles in doing so, given the heterogeneity of mHealth interventions and study designs and lack of agreement over choices of outcomes and their measurement instruments. Against this backdrop, this systematic mapping review seeks to answer the following questions: (1) What are the characteristics of mHealth interventions for adults with psychosis? (2) What type of outcomes are assessed and reported in RCTs of mHealth interventions for adults with psychosis?

Methods

Overview

This mapping review was performed according to the methodology proposed by James et al [19] and reported in line with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guidelines (Multimedia Appendix 1). The protocol was registered in Open Science Framework Registries [20] in May 2023.

Setting Inclusion Criteria for Studies

Studies included in this mapping review were randomized controlled studies of any design, reporting the use of mHealth

XSL•FO
RenderX

interventions for adults with psychosis. Our detailed inclusion and exclusion criteria are reflected in Textbox 1.

Textbox 1. Inclusion and exclusion criteria for randomized controlled trials (RCTs) reporting psychosis-focused mobile health (mHealth) interventions.

Inclusion criteria:

- Article type:
  - RCTs.
  - Cluster RCTs.
  - Quasi-RCTs.
  - Randomized controlled feasibility studies.
- Language: English-language studies.
- Population:
  - Adults above 18 years old diagnosed with schizophrenia spectrum disorders.
  - Any inpatient, outpatient, or community setting.
  - Any gender, ethnicity, or cultural background.
- Intervention:
  - Any kind of mHealth intervention that is intended to alter, manage, or prevent changes in a patient's behavior, emotions, cognition, functioning, or well-being.
  - Examples of possible intervention types include delivering psychological therapy, psychoeducation, emergency assistance, self-monitoring, personalized recommendations on coping strategies, or medication adherence.
- Comparison: Studies with any type of control, such as:
  - Another mHealth intervention.
  - A non-mHealth intervention.
  - Treatment-as-usual.
  - A sham comparison or placebo.
  - No intervention.
- Outcomes: All studies that evaluate the effectiveness of the mHealth intervention in terms of patient-related outcomes in psychosis, whether in terms of objective measures or patient-reported data.

Exclusion criteria:

- Article type:
  - All other study designs, such as qualitative studies, review articles (including meta-analyses and scoping, literature, and systematic reviews), commentaries, editorials, opinion pieces, protocols, and observational studies (including cross-sectional studies, cohort studies).
  - Conference abstracts, proceedings, and letters will be excluded unless a control group is present and there is sufficient data for extraction.
- Language: non–English language studies.
- Population:
  - Organic psychosis, substance-induced psychosis, psychosis secondary to other medical conditions, and postnatal psychosis.
  - Individuals with ultra-high risk (UHR) of psychosis.
  - Caregivers or next-of-kin of patients and health care professionals.
- Intervention:
  - Interventions that do not involve mHealth.
  - Interventions that exclusively collect passive data.
  - Interventions that only contain teleconferencing or virtual reality.
- Comparison: Single-arm interventional studies without a relevant control.
- Outcomes: Studies that exclusively assessed user experience-related or technical outcomes.

XSL•FO
RenderX

Searching for Evidence

An electronic literature search was performed on September 20, 2022, across PubMed, CINAHL (EBSCO), PsycINFO (EBSCO), Embase (Ovid), and Cochrane Library. We included English language reports published from January 1973 onwards, aligning with the advent of mobile technologies. The search strategy was developed in PubMed and adapted to other databases in consultation with a medical librarian. Search terms included a comprehensive list of words and phrases representing the intersection between mHealth interventions and psychosis (Multimedia Appendix 2). During the screening process, the citations of reports identified from bibliographical databases were also searched to include any other eligible publications that met all the inclusion criteria.

Screening Evidence

The search results from all databases were imported into a single EndNote (version 20; Clarivate) library, and duplicate records were removed. Subsequently, reviewers worked in pairs to select studies independently and in parallel using the online screening tool Covidence. This was performed in 2 stages: initial screening of title and abstract, followed by a second round of full-text screening. Discrepancies in any screening stage were resolved through a stepwise approach of mutual discussion, followed by engaging a third reviewer. The search and screening process was documented in a study selection flowchart [21].

Coding

Data were coded using a standardized data extraction form on Microsoft Excel developed by the review team. Variables that were extracted included the following: study methodology, participant baseline characteristics, study setting, characteristics of mHealth intervention and delivery, choice of outcomes, and corresponding measurement instruments. The data extraction form was piloted in 2 studies and amended based on feedback before it was used for data extraction. Furthermore, reviewers met up regularly to ensure concordance in the data extraction process. Like the screening stage, reviewers worked in pairs to extract data from included studies independently and in parallel. The extracted data was compared, and any discrepancies were resolved through mutual discussion or involving a third reviewer acting as the arbiter.

Describing and Visualizing the Findings

Data were visualized in a diagrammatic or tabular form accompanied by a narrative summary. Descriptive statistics were used to ascertain the choice of outcomes and measurement instruments reported in the included studies. Furthermore, data were classified and mapped according to the delivery medium (the digital platform used, eg, app, website, SMS), delivery channel (the physical device used, eg, smartphone, tablet), and delivery format. Based on recommendations by Lattie et al for digital mental health interventions, we modified their approach to subclassify delivery formats of mHealth interventions within this study based on the level of human support incorporated. Whereas self-guided interventions were defined to be fully automated, we consider guided interventions to “include human support as part of their delivery” (whether dyssynchronous or synchronous), while blended interventions involve the delivery of the digital modules “as part of face-to-face mental health interventions” [22].

Outcomes were grouped into 8 categories, namely severity of psychosis-related symptoms, functioning, well-being, medication adherence, adverse events, user experience, technical, and all other outcomes reported. These were derived directly or indirectly based on the International Consortium for Health Outcomes Measurement (ICHOM) patient-centered outcome measures for psychosis (eg, symptoms and functioning) [23], as well as the Core Outcome Measures in Effectiveness Trials (COMET) Initiative’s 38-item medical research outcome taxonomy (eg, 28: Emotional functioning/well-being, 32: Delivery of care, 38: Adverse events/effects) [24]. Instruments measuring any of these outcomes were also classified as objective or subjective, in line with the COSMIN (Consensus-based Standards for the selection of health Measurement INstruments) definition. [25] Objective measures include data passively obtained from phone usage or embedded sensors. Subjective measurement instruments include either observer-rated or self-report questionnaires that can be written or administered digitally, whether regarding their health status or experience using the mHealth intervention.

Results

Overview of Search Strategy

The initial search of databases yielded 2537 papers, of which a total of 1699 titles and abstracts were screened after removing duplicate records. Thereafter, 98 reports were assessed in full for eligibility, including 4 additional records, which were retrieved by searching the reference list of papers screened. This culminated in 29 publications reporting 23 studies being included in this review (PRISMA flow diagram in Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA flow diagram of study selection.

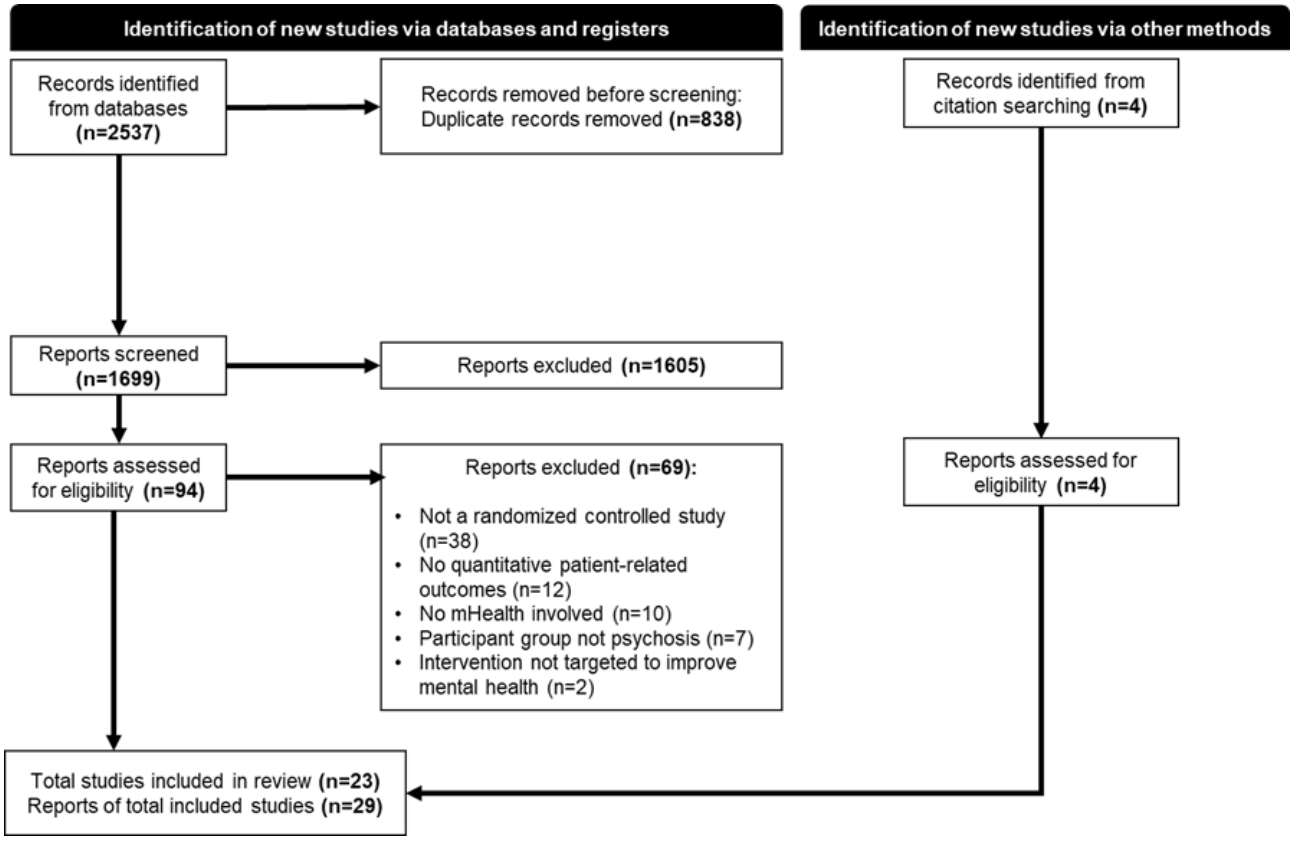

Characteristics of Included Studies

All included studies were published from 2017 onwards. Of these, nearly three-quarters of the included studies (17/23, 74%) were published from 2020 onwards [26-44]. Based on the 2021 World Bank Country and Landing Groups classification system, all except 2 studies [36,43] were conducted in high-income countries; 8/23 (35%) of the studies were conducted in the United States [32,38,44-52] and another 10/23 (43%) of the studies (including multicenter studies) were conducted in Europe [26,28,29,33-35,37,39-41,53,54]. The majority of the studies (16/23, 70%) were conducted in the outpatient and community setting [26,27,30-33,35-37,39-42,45-49,53,54]. Study designs included mostly RCTs with a variety of comparisons used. Particularly, two of the studies were 3-arm RCTs that featured both positive and negative controls [41,45].

There was also a mix of psychiatric diagnoses among recruited participants of the included studies. While over half (13/23, 57%) of included studies recruited only participants with psychotic disorders [26,28-32,34,36,39,40,42-44,47-49,52,53], 6 studies also recruited individuals with mood disorders (such as bipolar disorder or major depressive disorder) [27,33,38,45,46,50,51] and 1 study [35] included individuals with “Ultra-High Risk” of psychosis. By the American Psychiatric Association’s definition of adulthood, the mean participant age of most studies (16/23, 70%) 26-34,38,41-50,52] corresponded to middle adulthood (35 years and older) while that of remaining studies corresponded to young adulthood (35 years and younger). Table 1 presents a summary of the characteristics of the included studies, and further details can be found in Multimedia Appendix 3.

XSL•FO
RenderX

Table 1. Characteristics of included randomized controlled studies (N=23) reporting the use of mobile health (mHealth) interventions in psychosis.

Study characteristics	Studies, n (%)
Year of latest publication	
Before 2020	6 (26)
2020 or after	17 (74)
Country	
United States	8 (35)
United Kingdom	4 (17)
Australia	1 (4)
China	1 (4)
Denmark	1 (4)
France	1 (4)
Poland	1 (4)
South Korea	1 (4)
Netherlands	1 (4)
More than 1 country	4 (17)
Setting	
Outpatient and community	16 (70)
Inpatient	1 (4)
Not specified	6 (26)
Study design	
Randomized controlled trial (RCT)	20 (87)
Quasi-RCT	2 (9)
Cluster RCT	1 (4)
Type of control	
Treatment as usual (TAU)	9 (39)
Another mHealth intervention	3 (13)
Different version of mHealth intervention	3 (13)
Waitlist control	3 (13)
Non-mHealth intervention	2 (9)
Placebo	1 (4)
TAU + another mHealth intervention	1 (4)
TAU + non-mHealth intervention	1 (4)
Type of participant diagnosis	
Psychotic disorders only	13 (57)
Psychotic disorders + mood disorders	6 (26)
Ultra-High Risk + psychotic disorders	1 (4)
Not specified	3 (13)
Participant mean age	
<35	7 (30)
35 and older	16 (70)

XSL•FO
RenderX

Characteristics of mHealth Interventions

A wide range of features were incorporated into the mHealth interventions investigated in included studies, spanning from psychological therapy (12/23, 52%) [26,28,29,32,38,40,43,45-49,52-54], psychoeducation (8/23, 35%) [32,36,38,40,42,45,50-53], active self-monitoring (8/23, 35%) [27,30,31,33-35,37,43,50,51], medication adherence (7/23, 30%) [33,34,36,40,41,44,45,53], personalized recommendations on coping strategies (5/23, 22%) [27,30,31,33-35] to peer support (3/23, 13%) [32,43,52]. Different types of psychological therapy were used, such as cognitive behavioral therapy (CBT) [26,28,29,32,45,46,52,54], cognitive training [38,40,53], social cognition training [43], or a mix of techniques [47-49]. Less frequently incorporated features included emergency assistance [33], behavioral activation and mindfulness [52], shared decision-making [39], and teleconsultations [40,53].

To further characterize the type of mHealth interventions, these features were, in turn, grouped into distinct themes based on the pillars of Engel's biopsychosocial model [55]: medication-related features (eg, medication reminders, teleconsultations) representing the biological aspect, psychological therapy, and social support (eg, peer support, emergency assistance). A fourth intervention type—self-management—was also identified, encompassing psychoeducation, active self-monitoring and personalized recommendations on coping strategies. This was found to be a highly prevalent modality targeted by mHealth interventions (16/23, 70%) [27,30-38,40,42,43,45,50-53]. Many mHealth interventions were found to be multimodal, combining features targeting multiple pillars of Engel's biopsychosocial model (10/23, 43%) [32-34,36,38,40,43,45,52,53]. Even among unimodal interventions, a combination of multiple features was commonly used, such as active symptom monitoring with correspondingly tailored coping strategies [27,34,35], or medication reminders along with teleconsultations [40,53]. Of these, there was an even spread between self-management, psychological therapy, and medication or treatment-related interventions without any interventions that exclusively delivered social support.

The majority of mHealth interventions were delivered through applications (14/23, 61%) as the main medium [27,30,31,34,35,37-42,46,50-54], and smartphones (17/23, 74%) as the main channel of delivery [27,30,31,34-45,50-54]. Most of the interventions were also blended with in-person sessions (11/23, 48%) [26,27,32,33,35,39,40,42-45,53] or guided remotely (6/23, 26%) [28-31,36,37,50-52] by persons such as health care providers or trained peer supporters (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of mobile health (mHealth) interventions for psychosis from included studies (n=23).

mHealth intervention characteristics	Studies, n (%)
Intervention type	
Self-management only	6 (26)
Psychological therapy only	5 (22)
Medication-related only	2 (9)
Social support only	0 (0)
More than 1 type	10 (43)
Features involved[a]	
Psychological therapy	12 (52)
Cognitive behavioral therapy	7 (30)
Cognitive training	2 (9)
Social cognition training	1 (4)
Social cognition + cognitive training	1 (4)
Psychoeducation	8 (35)
Active self-monitoring	8 (35)
Medication adherence	7 (30)
Personalized recommendations on coping strategies	5 (22)
Peer support	
Emergency assistance	3 (13)
Behavioral activation and mindfulness	1 (4)
Shared decision making	1 (4)
Teleconsultations	1 (4)
Delivery medium	
Application	14 (61)
SMS or messaging applications	2 (9)
Website	2 (9)
More than 1 delivery medium	4 (17)
Not specified	1 (4)
Delivery channel	
Smartphone	17 (74)
Mobile phone	1 (4)
Tablet	1 (4)
Smartphone + computer	3 (13)
Tablet + computer	1 (4)
Format of delivery	
Blended	11 (48)
Guided	6 (26)
Self-guided	4 (17)
Not specified	2 (9)

[a]For this category, we note that the total sum of percentages will exceed 100% as some interventions come with multiple features.

XSL•FO
RenderX

Outcome Measures

The most common primary outcome was the severity of psychosis-related symptoms, which was chosen in 11/23 (48%) of included studies [26-29,32,34,35,37,40,50-53]. This was also the most prevalently measured outcome overall, being reflected in almost all (21/23, 91%) included studies [26-32,34-43,45-54]. Notably, psychosis-related symptoms were consistently reported in conjunction with other outcomes, such as well-being (17/23, 74%) [26-33,36-39,42,43,46-52,54], user experience (15/23, 65%) [27-31,33,34,37-40,43,46-54] and functioning (12/23, 52%) [34,35,37-39,41-43,45,47-49,52,54].

Figure 2 illustrates the choice of outcome measures and their relative frequencies in included studies. The most frequent measures of psychosis-related symptoms were positive symptoms (13/23, 57%) [26,27,34,35,37-40,42,50-54], mood and anxiety (10/23, 43%) [26,27,34,35,37,38,40,50-54], and overall psychopathology severity (9/23, 39%) [30,31,35,40-43,45,47-51,53]. Less commonly were negative symptoms (7/23, 30%) [34,35,39,40,45,47-49,52,53], other psychosis-related symptoms such as rehospitalization, relapse, and insight (7/23, 30%) [28-32,36,39-41,53], and cognitive symptoms (3/23, 13%) [26,46-49]. Measurement instruments for positive symptoms were highly varied with a mix of observer-rated questionnaires [27-29,35,37,39,40,42,50-54], self-report questionnaires [26,28,29,34,38], and ecological momentary assessments (EMAs) [35] to gauge aspects such as overall positive symptom severity, hallucinations, paranoia, and intensity and distress of psychotic experiences. Similar diversity was observed in measurement instruments for negative symptoms, including overall negative symptom severity [34,35,39,40,53], motivation [47-49,52], and defeatist beliefs [45,52]. On top of self-report questionnaires used to rate mood and anxiety symptoms, EMAs were also leveraged to measure momentary mood [34,35]. Finally, overall psychopathology severity was uniformly observer-rated.

Figure 2. Treemap of reported outcomes in randomized controlled studies of mHealth interventions for psychosis. The size of individual squares is proportional to the frequency of outcome measures.

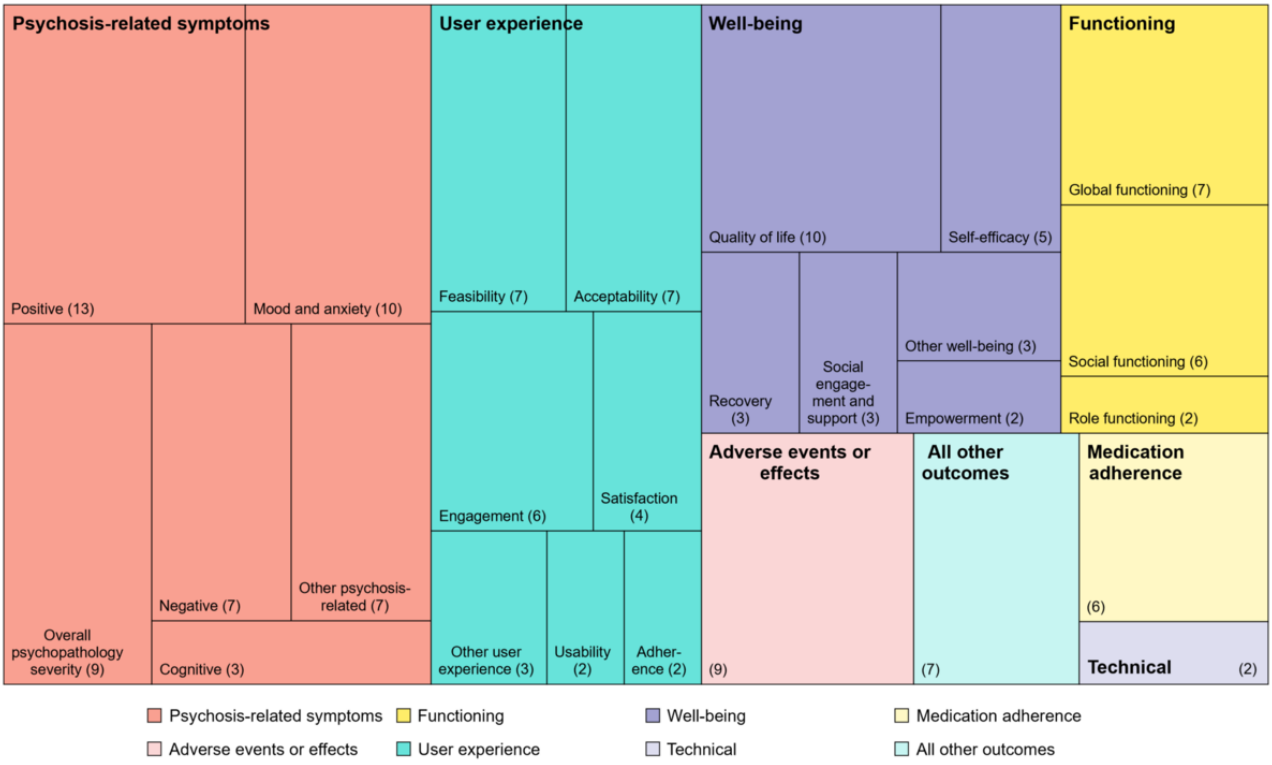

User experience was another commonly measured outcome, with the main domains of interest being feasibility (7/23, 30%) [27,30,31,34,37,40,46,52,53], acceptability (7/23, 30%) [27,30,31,37-39,52,54], and engagement (6/23, 26%) [30,31,38,43,47-51,54]. Feasibility was mostly understood as an objective construct among studies that reported it [27,31,34,37,52,54], with the main measurement instrument being passive data collected from devices. Interpretation of acceptability was slightly more pleomorphic, with most studies using self-report questionnaires to ascertain participants' views on the intervention [27,30,31,38,39,52,54] and some using objective phone data [37,39] and study retention rates [52]. Of the studies measuring engagement, most studied objective app engagement [30,31,38,40,43,45,50-54], although there were instances of measuring attendance of physical sessions accompanying the mHealth intervention [43,50,51] or subjective service engagement [30,31]. Other measures of user experience include satisfaction, usability, adherence, perceived usefulness, user-friendliness, and negative experiences.

Well-being was uniformly measured through self-report questionnaires. Although quality of life was the most measured construct of well-being (10/23, 43%) [26,28-33,37,43,47-51,54], there was marked diversity in the choice of outcomes within this category, including self-efficacy, recovery, social engagement and support, empowerment, hope, self-esteem, emotional distress, stigma, use of coping strategies and

XSL•FO
RenderX

metacognitive beliefs. On the other hand, functioning was mostly measured with observer-rated questionnaires, though other types of measurement instruments were also used [34,35,38,43]. Most instruments measured global (7/23, 30%) [31,35,38,39,41,45,54] and social functioning (6/23, 26%) [34,35,39,42,43,54]. Almost all studies measuring medication adherence used the self-reported Medication Adherence Rating Scale [30,31,33,44,54] except for one study using the Medication Adherence Questionnaire [36] and another collecting data from sensors embedded in medication bottle caps [41]. Contrary to most outcomes, which were compared before and after the intervention, adverse events and technical outcomes were predominantly monitored throughout the study as and when they arose [26-31,37,40,43,45,46,53,54].

Choices of outcomes remained multidimensional even after stratifying by the type of intervention studied, as displayed in Figure 3. The severity of psychosis-related symptoms, functioning, and medication adherence were categories of reported outcomes that were common to all intervention types. In contrast, well-being, adverse events, and user experience outcomes were outcome choices common to psychotherapy-only, self-management-only, and mixed modality interventions but not reported in exclusively medication-related mHealth interventions for psychosis. Other outcomes refer to intervention-specific outcomes that cannot be encompassed in the other categories, such as participants' recall of specific coping heuristics, auditory processing, and motivational incongruence.

Figure 3. Types of reported outcomes according to the type of mHealth intervention for psychosis.

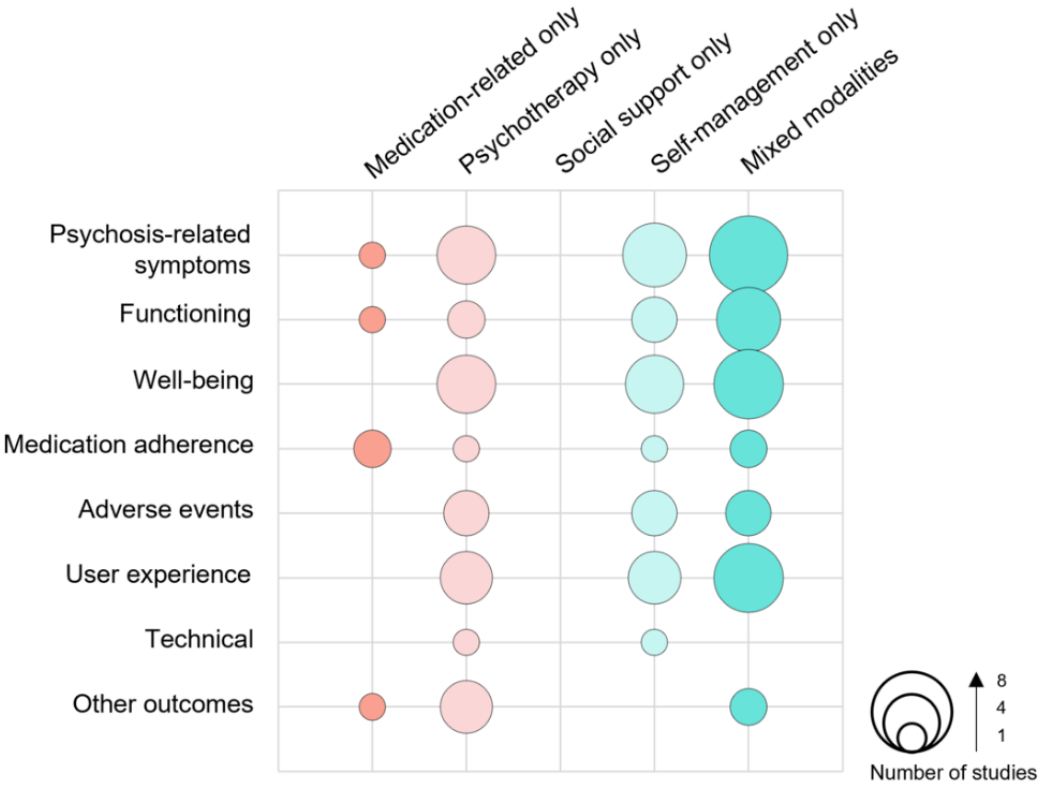

Discussion

Principal Findings

In this study, we rigorously reviewed the choice of outcomes and measurement instruments in randomized controlled studies evaluating the use of mHealth interventions for adults with psychosis in relation to intervention and control characteristics. The recency of the 23 included studies, particularly the surge in studies published from 2020 onwards, is concordant with the dramatic growth of research in digital mental health in general [12,56], with the COVID-19 pandemic probably being a key catalyst of further development [57,58]. It is encouraging that most mHealth interventions studied are tailored toward outpatient and community care, where sustaining accessibility to mental health services remains a challenge: World Health Organization statistics reveal that almost one-third of countries worldwide still have fewer than 1 psychiatrist in the workforce per 100,000 individuals [59]. The emergence of transdiagnostic mHealth interventions also aligns with a previous systematic review [60], which found a marked increase in studies on mental health applications capable of benefitting multiple groups of patients with different diagnoses. The focus on participants in mid-adulthood may impress upon the possibility of mHealth to manage psychosis in the long term. That being said, younger populations may arguably be a bigger benefactor of mHealth interventions with more pervasive smartphone use and lower engagement with traditional health services [61,62].

The findings of this mapping review strongly suggest that mHealth interventions for adults with psychosis are complex and typically multidimensional in nature. When reviewing the intervention features in isolation, psychotherapy was the most frequently deployed intervention, followed by psychoeducation and active symptom monitoring. Unsurprisingly, CBT was the most common type of psychotherapy delivered, in keeping with

XSL•FO
RenderX

its established evidence base and showing its translation toward the mobile realm [63]. Upon further classifying intervention features by domain, most mHealth interventions for psychosis either traversed multiple pillars of Engel's biopsychosocial model or minimally combined multiple features within the same pillar. Notably, supporting users' self-management was found to be the most common single-mode intervention. With previous evidence demonstrating how self-management can improve outcomes of people living with severe mental illness [64], this corroborates the role of mHealth interventions in helping users navigate the chronic trajectory of psychosis. Furthermore, we found that the most common medium and channel of delivery were applications and smartphones. This is within expectations given the growing penetration of smartphone usage day-to-day and smartphone uptake among individuals with psychosis [65].

Outcomes reported were similarly numerous, comparable with findings from previous literature [14]. The severity of psychosis-related symptoms—the most reported outcome—was always measured together with other patient-centric measures such as well-being, user experience, or functioning. This bears testament to the transition toward patient-centered care [66] in both the physical and digital spheres and the ensuing importance of evaluating user-related factors [60,67,68] in adopting digital mental health tools. Overall, there needs to be better concordance between measurement instruments of the same outcome categories. We observed multiple studies using mHealth-enabled nascent tools to pick up objective data, such as EMAs [27,30,31,34,35,37,43,50,51] and embedded sensors [41]. This is exciting for refining our understanding of how users' clinical pictures evolve temporally, but it would require proper validation against traditional measurement instruments [17,69]. Furthermore, included studies vastly differed in the choice and modality of questionnaires, whether observer-rated or self-reported, for variables including hallucinations and social functioning. This might further hinder the reproducibility of results, on top of the myriad of control designs in included studies. Such heterogeneity may directly impact subsequent evaluations of efficacy—for instance, Goldberg et al [12] noted that effect sizes of mHealth-based mental health interventions were smaller when compared with controls with therapeutic intent.

Interestingly, our review dispels the notion that the heterogeneity of outcomes directly reflects intervention diversity—in fact, the spectrum of reported outcomes remained broad despite stratifying by intervention type. This may put into question whether deciding outcomes to be studied based on interventions' putative mechanisms (like what the ecological interventionist-causal model approach proposes [70]) might still apply when evaluating mHealth interventions. Taken from another perspective, this finding might just be another portrayal of how multifaceted mHealth interventions can be with its integration of multiple components. Their effects may transcend illness severity to simultaneously impact patients' experiences of their illnesses. In turn, this would strengthen the impetus for holistically assessing clinical and functional outcomes when evaluating mHealth interventions for psychosis. Overall, the plethora of interventions, comparisons, outcomes, and outcome measurement instruments reinforces the need for more high-quality studies with standardized controls and evaluation frameworks to enable validity in further data synthesis. Considering the complexity of between-feature interaction and synergistically affected patient outcomes, the uncertainty still lingers if RCTs are the best evaluation framework for mHealth interventions despite being conventionally perceived as the gold standard of interventional trials.

Strengths and Limitations

This systematic mapping review has notable strengths. We used a comprehensive search strategy across 5 major bibliographic databases to identify publications reporting on the use of mHealth interventions in adults with psychosis. We have also attempted to minimize missing eligible records by searching the references of included studies. A robust methodology to screen identified articles and perform data extraction in pairs was used to ensure reliable review findings.

There are also some limitations. While including only randomized controlled study designs allowed greater strength in interpreting effects on outcomes, we may not have included nascent interventions in the early stages of development in the process. Furthermore, some potentially relevant studies may have been omitted due to the gamut of terms used in this area.

Conclusions

mHealth interventions for psychosis should be considered complex interventions probing multiple social and behavioral determinants of health. Randomized controlled studies in the field often report a remarkable breadth of outcomes regardless of the domains they seek to target. Marked variation in the choice of measurement instruments for said outcomes and comparison groups persist amongst studies included in this review. Hence, we recommend a holistic evaluation approach combining clinical and patient-centric outcomes to adequately account for the assortment of mHealth interventions and outcome measures. This would advance our understanding of individuals' interactions with mHealth interventions that can enhance the person-centered delivery of mental health care for people with psychosis.

Acknowledgments

We would like to acknowledge Mdm Rebecca David for her guidance toward developing and translating the search strategy.

This project was supported by Nanyang Technological University under the URECA Undergraduate Research Programme. JL was supported by the Singapore Ministry of Health's National Medical Research Council under its Clinician Scientist Award (Grant number: CSAINV17nov-0004).

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL and LTC conceived the idea of the review. PYL, CH, XYT, MH, and YYC screened the studies and extracted data from the included studies. PYL analyzed the data from eligible studies. PYL wrote the review, and JL, LTC, and LM provided methodological guidance. All authors critically reviewed and approved the final manuscript.

Conflicts of Interest

JL had received honoraria, served as a consultant or advisory board member from Otsuka, Janssen, Lundbeck, Sumitomo Pharmaceuticals, Boehringer Ingelheim, and ThoughtFull World Pte. Ltd. LTC is an Associate Editor for *JMIR Medical Education*. The remaining authors have no conflicts of interest to declare.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 109 KB - jmir_v26i1e55924_app1.pdf]

Multimedia Appendix 2
PubMed search strategy.
[DOCX File , 18 KB - jmir_v26i1e55924_app2.docx]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File , 26 KB - jmir_v26i1e55924_app3.docx]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. USA: American Psychiatric Association; 2013.
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. Tandon R, Nasrallah HA, Keshavan MS. Schizophrenia, "just the facts" 4. Clinical features and conceptualization. Schizophr Res 2009;110(1-3):1-23. [doi: 10.1016/j.schres.2009.03.005] [Medline: 19328655]
4. Whiteford H, Ferrari A, Degenhardt L. Global burden Of disease studies: implications for mental and substance use disorders. Health Aff (Millwood) 2016;35(6):1114-1120. [doi: 10.1377/hlthaff.2016.0082] [Medline: 27269030]
5. Kadakia A, Fan A, Marden J, Dembek C, Catillon M, Anderson A, et al. The economic burden of schizophrenia in the United States in 2019. CNS Spectr 2022;27(2):227-227. [doi: 10.1017/s1092852922000207]
6. World Health Organization. mHealth: use of mobile wireless on public health. Exec Board 139th Session. 2016. URL: https://apps.who.int/gb/ebwha/pdf_files/EB139/B139_8-en.pdf [accessed 2023-06-23]
7. Firth J, Cotter J, Torous J, Bucci S, Firth JA, Yung AR. Mobile phone ownership and endorsement of "mHealth" Among people with psychosis: a meta-analysis of cross-sectional studies. Schizophr Bull 2016;42(2):448-455 [FREE Full text] [doi: 10.1093/schbul/sbv132] [Medline: 26400871]
8. Ahmed I, Ahmad NS, Ali S, Ali S, George A, Saleem Danish H, et al. Medication adherence apps: review and content analysis. JMIR Mhealth Uhealth 2018;6(3):e62 [FREE Full text] [doi: 10.2196/mhealth.6432] [Medline: 29549075]
9. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. J Med Internet Res 2017;19(11):e399 [FREE Full text] [doi: 10.2196/jmir.8598] [Medline: 29187342]
10. Achtyes ED, Gega L, Linnaranta O. Editorial: mHealth: self-management and complementary psychiatric treatment. Front Psychiatry 2021;12:777090 [FREE Full text] [doi: 10.3389/fpsyt.2021.777090] [Medline: 34744848]
11. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
12. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002 [FREE Full text] [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
13. Chivilgina O, Wangmo T, Elger BS, Heinrich T, Jotterand F. mHealth for schizophrenia spectrum disorders management: a systematic review. Int J Soc Psychiatry 2020;66(7):642-665. [doi: 10.1177/0020764020933287] [Medline: 32571123]

XSL•FO
RenderX

14. Gire N, Farooq S, Naeem F, Duxbury J, McKeown M, Kundi PS, et al. mHealth based interventions for the assessment and treatment of psychotic disorders: a systematic review. Mhealth 2017;3:33 [FREE Full text] [doi: 10.21037/mhealth.2017.07.03] [Medline: 28894743]
15. Camacho E, Levin L, Torous J. Smartphone apps to support coordinated specialty care for prodromal and early course schizophrenia disorders: systematic review. J Med Internet Res 2019;21(11):e16393 [FREE Full text] [doi: 10.2196/16393] [Medline: 31714250]
16. Alvarez-Jimenez M, Alcazar-Corcoles MA, González-Blanch C, Bendall S, McGorry PD, Gleeson JF. Online, social media and mobile technologies for psychosis treatment: a systematic review on novel user-led interventions. Schizophr Res 2014;156(1):96-106. [doi: 10.1016/j.schres.2014.03.021] [Medline: 24746468]
17. Firth J, Torous J. Smartphone apps for schizophrenia: a aystematic review. JMIR Mhealth Uhealth 2015;3(4):e102 [FREE Full text] [doi: 10.2196/mhealth.4930] [Medline: 26546039]
18. Clarke S, Hanna D, Mulholland C, Shannon C, Urquhart C. A systematic review and meta-analysis of digital health technologies effects on psychotic symptoms in adults with psychosis. Psychosis Routledge 2019;11(4):362-373. [doi: 10.1080/17522439.2019.1632376]
19. James KL, Randall NP, Haddaway NR. A methodology for systematic mapping in environmental sciences. Environ Evid 2016;5(1):7. [doi: 10.1186/s13750-016-0059-6]
20. Loh PY. OSF Registry. mHealth Interventions for Adult Patients With Psychosis: A Systematic Mapping Review. 2023. URL: https://doi.org/10.17605/OSF.IO/DSPHU [accessed 203-05-08]
21. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
22. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
23. McKenzie E, Matkin L, Sousa Fialho L, Emelurumonye IN, Gintner T, Ilesanmi C, Psychotic Disorders Working Group of the International Consortium for Health Outcomes Measurement. Developing an international standard set of patient-reported outcome measures for psychotic disorders. Psychiatr Serv 2022;73(3):249-258 [FREE Full text] [doi: 10.1176/appi.ps.202000888] [Medline: 34369809]
24. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. J Clin Epidemiol 2018;96:84-92 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.12.020] [Medline: 29288712]
25. Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials 2016;17(1):449 [FREE Full text] [doi: 10.1186/s13063-016-1555-2] [Medline: 27618914]
26. Garety P, Ward T, Emsley R, Greenwood K, Freeman D, Fowler D, et al. Effects of SlowMo, a blended digital therapy targeting reasoning, on paranoia among people with psychosis: a randomized clinical trial. JAMA Psychiatry 2021;78(7):714-725 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0326] [Medline: 33825827]
27. Bell IH, Rossell SL, Farhall J, Hayward M, Lim MH, Fielding-Smith SF, et al. Pilot randomised controlled trial of a brief coping-focused intervention for hearing voices blended with smartphone-based ecological momentary assessment and intervention (SAVVy): Feasibility, acceptability and preliminary clinical outcomes. Schizophr Res 2020;216:479-487. [doi: 10.1016/j.schres.2019.10.026] [Medline: 31812327]
28. Westermann S, Rüegg N, Lüdtke T, Moritz S, Berger T. Internet-based self-help for psychosis: findings from a randomized controlled trial. J Consult Clin Psychol 2020;88(10):937-950. [doi: 10.1037/ccp0000602] [Medline: 32790453]
29. Lüdtke T, Platow-Kohlschein H, Rüegg N, Berger T, Moritz S, Westermann S. Mindfulness mediates the effect of a psychological online intervention for psychosis on self-reported hallucinations: a secondary analysis of voice hearers from the EviBaS trial. Front Psychiatry 2020;11:228 [FREE Full text] [doi: 10.3389/fpsyt.2020.00228] [Medline: 32308631]
30. Gumley AI, Bradstreet S, Ainsworth J, Allan S, Alvarez-Jimenez M, Aucott L, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in scotland and Australia. Lancet Psychiatry 2022;9(6):477-486 [FREE Full text] [doi: 10.1016/S2215-0366(22)00103-1] [Medline: 35569503]
31. Gumley AI, Bradstreet S, Ainsworth J, Allan S, Alvarez-Jimenez M, Birchwood M, et al. Digital smartphone intervention to recognise and manage early warning signs in schizophrenia to prevent relapse: the EMPOWER feasibility cluster RCT. Health Technol Assess 2022;26(27):1-174 [FREE Full text] [doi: 10.3310/HLZE0479] [Medline: 35639493]
32. Homan P, Schooler NR, Brunette MF, Rotondi A, Ben-Zeev D, Gottlieb JD, et al. Relapse prevention through health technology program reduces hospitalization in schizophrenia. Psychol. Med 2022;53(9):4114-4120. [doi: 10.1017/s0033291722000794]
33. Röhricht F, Padmanabhan R, Binfield P, Mavji D, Barlow S. Simple mobile technology health management tool for people with severe mental illness: a randomised controlled feasibility trial. BMC Psychiatry 2021;21(1):357 [FREE Full text] [doi: 10.1186/s12888-021-03359-z] [Medline: 34271902]
34. Hanssen E, Balvert S, Oorschot M, Borkelmans K, van Os J, Delespaul P, et al. An ecological momentary intervention incorporating personalised feedback to improve symptoms and social functioning in schizophrenia spectrum disorders. Psychiatry Res 2020;284:112695 [FREE Full text] [doi: 10.1016/j.psychres.2019.112695] [Medline: 31831201]

XSL•FO
RenderX

35. Myin-Germeys I, van Aubel E, Vaessen T, Steinhart H, Klippel A, Lafit G, et al. Efficacy of acceptance and commitment therapy in daily life in early psychosis: results from the multi-center INTERACT randomized controlled trial. Psychother Psychosom 2022;91(6):411-423. [doi: 10.1159/000522274] [Medline: 35306502]
36. Zhu X, Li M, Liu P, Chang R, Wang Q, Liu J. A mobile health application-based strategy for enhancing adherence to antipsychotic medication in schizophrenia. Arch Psychiatr Nurs 2020;34(6):472-480. [doi: 10.1016/j.apnu.2020.08.001] [Medline: 33280669]
37. Lewis S, Ainsworth J, Sanders C, Stockton-Powdrell C, Machin M, Whelan P, et al. Smartphone-enhanced symptom management in psychosis: open, randomized controlled trial. J Med Internet Res 2020;22(8):e17019 [FREE Full text] [doi: 10.2196/17019] [Medline: 32788150]
38. Ben-Zeev D, Chander A, Tauscher J, Buck B, Nepal S, Campbell A, et al. A smartphone intervention for people with serious mental illness: fully remote randomized controlled trial of CORE. J Med Internet Res 2021;23(11):e29201 [FREE Full text] [doi: 10.2196/29201] [Medline: 34766913]
39. Vitger T, Hjorthøj C, Austin SF, Petersen L, Tønder ES, Nordentoft M, et al. A smartphone app to promote patient activation and support shared decision-making in people with a diagnosis of schizophrenia in outpatient treatment settings (Momentum Trial): randomized controlled assessor-blinded trial. J Med Internet Res 2022;24(10):e40292 [FREE Full text] [doi: 10.2196/40292] [Medline: 36287604]
40. Krzystanek M, Krysta K, Borkowski M, Skałacka K, Przybyło J, Pałasz A, et al. The effect of smartphone-based cognitive training on the functional/cognitive markers of schizophrenia: a one-year randomized study. J Clin Med 2020;9(11):3681 [FREE Full text] [doi: 10.3390/jcm9113681] [Medline: 33207811]
41. Tessier A, Dupuy M, Baylé FJ, Herse C, Lange A, Vrijens B, et al. Brief interventions for improving adherence in schizophrenia: a pilot study using electronic medication event monitoring. Psychiatry Res 2020;285:112780 [FREE Full text] [doi: 10.1016/j.psychres.2020.112780] [Medline: 31954541]
42. Han M, Lee K, Kim M, Heo Y, Choi H. Effects of a metacognitive smartphone intervention with weekly mentoring sessions for individuals with schizophrenia: a quasi-experimental study. J Psychosoc Nurs Ment Health Serv 2023;61(2):27-37 [FREE Full text] [doi: 10.3928/02793695-20220706-01] [Medline: 35858205]
43. Dabit S, Quraishi S, Jordan J, Biagianti B. Improving social functioning in people with schizophrenia-spectrum disorders via mobile experimental interventions: results from the CLIMB pilot trial. Schizophr Res Cogn 2021;26:100211 [FREE Full text] [doi: 10.1016/j.scog.2021.100211] [Medline: 34381699]
44. Lisinge E. Improving medication adherence in patients diagnosed with schizophrenia using cell phone apps in addition to attending focus group therapy. Diss Abstr Int Sect B Sci Eng 2020;81(6-B):1-17.
45. Depp CA, Perivoliotis D, Holden J, Dorr J, Granholm EL. Single-session mobile-augmented intervention in serious mental illness: a three-arm randomized controlled trial. Schizophr Bull 2019;45(4):752-762 [FREE Full text] [doi: 10.1093/schbul/sby135] [Medline: 30281086]
46. Roberts DL, Liu PYT, Busanet H, Maples N, Velligan D. A tablet-based intervention to manipulate social cognitive bias in schizophrenia. American Journal of Psychiatric Rehabilitation 2017;20(2):143-155. [doi: 10.1080/15487768.2017.1302897]
47. Biagianti B, Fisher M, Howard L, Rowlands A, Vinogradov S, Woolley J. Feasibility and preliminary efficacy of remotely delivering cognitive training to people with schizophrenia using tablets. Schizophr Res Cogn 2017;10:7-14 [FREE Full text] [doi: 10.1016/j.scog.2017.07.003] [Medline: 28824850]
48. Fisher M, Nahum M, Howard E, Rowlands A, Brandrett B, Kermott A, et al. Supplementing intensive targeted computerized cognitive training with social cognitive exercises for people with schizophrenia: an interim report. Psychiatr Rehabil J 2017;40(1):21-32 [FREE Full text] [doi: 10.1037/prj0000244] [Medline: 28368179]
49. Miley K, Fisher M, Nahum M, Howard E, Rowlands A, Brandrett B, et al. Six month durability of targeted cognitive training supplemented with social cognition exercises in schizophrenia. Schizophr Res Cogn 2020;20:100171 [FREE Full text] [doi: 10.1016/j.scog.2019.100171] [Medline: 31908976]
50. Ben-Zeev D, Brian RM, Jonathan G, Razzano L, Pashka N, Carpenter-Song E, et al. Mobile health (mHealth) versus clinic-based group intervention for people with serious mental illness: a randomized controlled trial. Psychiatr Serv 2018;69(9):978-985. [doi: 10.1176/appi.ps.201800063] [Medline: 29793397]
51. Ben-Zeev D, Buck B, Chu PV, Razzano L, Pashka N, Hallgren KA. Transdiagnostic mobile health: smartphone intervention reduces depressive symptoms in people with mood and psychotic disorders. JMIR Ment Health 2019;6(4):e13202 [FREE Full text] [doi: 10.2196/13202] [Medline: 30977736]
52. Schlosser DA, Campellone TR, Truong B, Etter K, Vergani S, Komaiko K, et al. Efficacy of PRIME, a mobile app intervention designed to improve motivation in young people with schizophrenia. Schizophr Bull 2018;44(5):1010-1020 [FREE Full text] [doi: 10.1093/schbul/sby078] [Medline: 29939367]
53. Krzystanek M, Borkowski M, Skałacka K, Krysta K. A telemedicine platform to improve clinical parameters in paranoid schizophrenia patients: results of a one-year randomized study. Schizophr Res 2019;204:389-396. [doi: 10.1016/j.schres.2018.08.016] [Medline: 30154027]
54. Bucci S, Barrowclough C, Ainsworth J, Machin M, Morris R, Berry K, et al. Actissist: proof-of-concept trial of a theory-driven digital intervention for psychosis. Schizophr Bull 2018;44(5):1070-1080 [FREE Full text] [doi: 10.1093/schbul/sby032] [Medline: 29566206]

XSL•FO
RenderX

55. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977;196(4286):129-136. [doi: 10.1126/science.847460] [Medline: 847460]
56. Drissi N, Ouhbi S, Janati Idrissi MA, Fernandez-Luque L, Ghogho M. Connected mental health: systematic mapping study. J Med Internet Res 2020;22(8):e19950 [FREE Full text] [doi: 10.2196/19950] [Medline: 32857055]
57. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
58. Baños RM, Herrero R, Vara MD. What is the current and future status of digital mental health interventions? Span. J. Psychol 2022;25:e5. [doi: 10.1017/sjp.2022.2]
59. World Health Organization. Mental Health Workforce Statistics. 2019. URL: https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/mental-health-workers [accessed 2023-06-26]
60. Miralles I, Granell C, Díaz-Sanahuja L, Van Woensel W, Bretón-López J, Mira A, et al. Smartphone apps for the treatment of mental disorders: systematic review. JMIR Mhealth Uhealth 2020;8(4):e14897 [FREE Full text] [doi: 10.2196/14897] [Medline: 32238332]
61. Burns J, Birrell E. Enhancing early engagement with mental health services by young people. Psychol Res Behav Manag 2014;7:303-312 [FREE Full text] [doi: 10.2147/PRBM.S49151] [Medline: 25473320]
62. Rus-Calafell M, Schneider S. Are we there yet?!-a literature review of recent digital technology advances for the treatment of early psychosis. Mhealth 2020;6:3-3 [FREE Full text] [doi: 10.21037/mhealth.2019.09.14] [Medline: 32190614]
63. Kumar V, Sattar Y, Bseiso A, Khan S, Rutkofsky IH. The effectiveness of internet-based cognitive behavioral therapy in treatment of psychiatric disorders. Cureus 2017;9(8):e1626 [FREE Full text] [doi: 10.7759/cureus.1626] [Medline: 29098136]
64. Lean M, Fornells-Ambrojo M, Milton A, Lloyd-Evans B, Harrison-Stewart B, Yesufu-Udechuku A, et al. Self-management interventions for people with severe mental illness: systematic review and meta-analysis. Br J Psychiatry 2019;214(5):260-268 [FREE Full text] [doi: 10.1192/bjp.2019.54] [Medline: 30898177]
65. Buck B, Chander A, Tauscher J, Nguyen T, Monroe-DeVita M, Ben-Zeev D. mHealth for young adults with early psychosis: user preferences and their relationship to attitudes about treatment-seeking. J Technol Behav Sci 2021;6(4):667-676 [FREE Full text] [doi: 10.1007/s41347-021-00223-5] [Medline: 34604506]
66. Ellen ME, Shach R, Balicer RD. Helping patients help themselves: supporting the healthcare journey. Patient Educ Couns 2018;101(9):1708-1711. [doi: 10.1016/j.pec.2018.04.005] [Medline: 29691111]
67. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022;9(6):e35591 [FREE Full text] [doi: 10.2196/35591] [Medline: 35671081]
68. Kaveladze BT, Wasil AR, Bunyi JB, Ramirez V, Schueller SM. User experience, engagement, and popularity in mental health apps: secondary analysis of app analytics and expert app reviews. JMIR Hum Factors 2022;9(1):e30766. [doi: 10.2196/30766]
69. Kirkpatrick B, Luther L, Strauss GP. Negative symptoms in the clinic: we treat what we can describe. Br J Psychiatry 2023;223(1):271-272 [FREE Full text] [doi: 10.1192/bjp.2023.68] [Medline: 37424200]
70. Reininghaus U, Depp CA, Myin-Germeys I. Ecological interventionist causal models in psychosis: targeting psychological mechanisms in daily life. Schizophr Bull 2016;42(2):264-269 [FREE Full text] [doi: 10.1093/schbul/sbv193] [Medline: 26707864]

Abbreviations

CBT: cognitive behavioral therapy
COMET: Core Outcome Measures in Effectiveness Trials
COSMIN: Consensus-based Standards for the selection of health Measurement INstruments
EMA: ecological momentary assessment
iCBT: internet-based cognitive behavioral therapy
ICHOM: International Consortium for Health Outcomes Measurement
mHealth: mobile health
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
YLDs: years lived with disability

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.01.24; peer-reviewed by X B, M Han; comments to author 10.04.24; revised version received 28.06.24; accepted 06.10.24; published 23.12.24.

Please cite as:
Loh PY, Martinengo L, Heaukulani C, Tan XY, Hng M, Cheah YY, Morris RJT, Tudor Car L, Lee J
Characteristics and Outcomes of mHealth Interventions in Psychosis: Systematic Mapping Review
J Med Internet Res 2024;26:e55924
URL: https://www.jmir.org/2024/1/e55924
doi:10.2196/55924
PMID:39714907

©Pei Yi Loh, Laura Martinengo, Creighton Heaukulani, Xin Yang Tan, Moses Hng, Yong Yin Cheah, Robert J T Morris, Lorainne Tudor Car, Jimmy Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning and Deep Learning for Diagnosis of Lumbar Spinal Stenosis: Systematic Review and Meta-Analysis

Tianyi Wang[1*], MD; Ruiyuan Chen[1*], MM; Ning Fan[1*], MD; Lei Zang[1], MD; Shuo Yuan[1], MD; Peng Du[1], MM; Qichao Wu[1], MD; Aobo Wang[1], MD; Jian Li[1], MM; Xiaochuan Kong[1], MB; Wenyi Zhu[1], MD

Beijing Chaoyang Hospital, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Lei Zang, MD
Beijing Chaoyang Hospital
Capital Medical University
5 JingYuan Road, Shijingshan District
Beijing, 100043
China
Phone: 86 51718268
Email: zanglei@ccmu.edu.cn

Abstract

Background: Lumbar spinal stenosis (LSS) is a major cause of pain and disability in older individuals worldwide. Although increasing studies of traditional machine learning (TML) and deep learning (DL) were conducted in the field of diagnosing LSS and gained prominent results, the performance of these models has not been analyzed systematically.

Objective: This systematic review and meta-analysis aimed to pool the results and evaluate the heterogeneity of the current studies in using TML or DL models to diagnose LSS, thereby providing more comprehensive information for further clinical application.

Methods: This review was performed under the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines using articles extracted from PubMed, Embase databases, and Cochrane Library databases. Studies that evaluated DL or TML algorithms assessment value on diagnosing LSS were included, while those with duplicated or unavailable data were excluded. Quality Assessment of Diagnostic Accuracy Studies 2 was used to estimate the risk of bias in each study. The MIDAS module and the METAPROP module of Stata (StataCorp) were used for data synthesis and statistical analyses.

Results: A total of 12 studies with 15,044 patients reported the assessment value of TML or DL models for diagnosing LSS. The risk of bias assessment yielded 4 studies with high risk of bias, 3 with unclear risk of bias, and 5 with completely low risk of bias. The pooled sensitivity and specificity were 0.84 (95% CI: 0.82-0.86; I^2=99.06%) and 0.87 (95% CI 0.84-0.90; I^2=98.7%), respectively. The diagnostic odds ratio was 36 (95% CI 26-49), the positive likelihood ratio (LR+) was 6.6 (95% CI 5.1-8.4), and the negative likelihood ratio (LR–) was 0.18 (95% CI 0.16-0.21). The summary receiver operating characteristic curves, the area under the curve of TML or DL models for diagnosing LSS of 0.92 (95% CI 0.89-0.94), indicating a high diagnostic value.

Conclusions: This systematic review and meta-analysis emphasize that despite the generally satisfactory diagnostic performance of artificial intelligence systems in the experimental stage for the diagnosis of LSS, none of them is reliable and practical enough to apply in real clinical practice. Further efforts, including optimization of model balance, widely accepted objective reference standards, multimodal strategy, large dataset for training and testing, external validation, and sufficient and scientific report, should be made to bridge the distance between current TML or DL models and real-life clinical applications in future studies.

Trial Registration: PROSPERO CRD42024566535; https://tinyurl.com/msx59x8k

(J Med Internet Res 2024;26:e54676) doi:10.2196/54676

KEYWORDS

lumbar spinal stenosis; LSS; machine learning; ML; deep learning; artificial intelligence; AI; diagnosis; spine stenosis; lumbar; predictive model; early detection; diagnostic; older adult;

XSL•FO
RenderX

Introduction

Lumbar spinal stenosis (LSS) is a major cause of pain and disability in older individuals [1]. LSS has become a worldwide public health issue as it is estimated that more than 102 million people are diagnosed with LSS annually, with high incidence in Europe and the United States of America [2,3]. According to the clinical guideline developed by the North American Spine Society, LSS is characterized as a condition of diminished space available for the neural and vascular elements in the lumbar spine, secondary to degenerative changes in the spinal canal [4]. An accurate LSS diagnosis is essential for treatment options and effectiveness. Currently, clinicians diagnose LSS based on a comprehensive evaluation combined with the patient's history, physical examination, and spinal imaging tests such as x-ray, computed tomography (CT), and magnetic resonance imaging (MRI) [1,2]. As a superior radiographic screening tool for soft tissues, MRI plays a crucial role in detecting the presence, classification, and grading of LSS [5-7]. However, detailing numerous information in spinal MRI is time-consuming and repetitive, which causes laborious clinical workloads [7]. Furthermore, existing LSS grading systems are mainly qualitative or semiquantitative, which highly depend on expertise and suffer from high interobserver variations because of the complexity of the spinal canal and foramen [5,6,8-11]. Therefore, more intelligent radiographic diagnostic and grading methods of LSS are warranted.

Machine learning (ML), a subdiscipline of artificial intelligence (AI), has shown great advantages in analyzing medical imaging and predicting outcome decisions [12-14]. ML begins with algorithms trained with a set of data, such as image features, to establish the prediction or diagnosis by extracting and classifying relevant information. More recently, a crucial branch of ML, named deep learning (DL), was standing out rapidly. DL algorithms were designed with multiple processing layers, which can learn more complex image features than traditional ML methods [15]. Although DL is still challenged by the demand for large-scale datasets and the difficulty of interpretation, it owns the incomparable advantage of automatic feature extraction, minimizing the bias by manual intervention [12,14]. In 2016, He et al [16] attempted to use traditional ML (TML) methods based on their newly proposed synchronized superpixel representation model to recognize the presence of radiographic lumbar foraminal stenosis (LFS). Subsequently, increasing studies of TML and DL were conducted in the field of diagnosing and grading LSS and gained prominent results [16-34]. However, most of these studies focus either on algorithm development or clinical validation, causing great variations in experimental settings and incompleteness of evaluation parameters of accuracy and reliability. Hence, a systematic review and meta-analysis were believed to be necessary to evaluate the heterogeneity and provide comprehensive results from these studies. However, to our knowledge, no systematic review and meta-analysis was previously conducted to address this issue.

Therefore, this systematic review and meta-analysis aimed to evaluate the heterogeneity and pool the results of the current studies in using ML or DL models to diagnose LSS, thereby providing more comprehensive information for further clinical application.

Methods

Study Design and Registration

This systematic literature review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and flowchart [35,36] and the PRISMA diagnostic test accuracy checklist (Multimedia Appendix 1) [37]. The protocol for this systematic review was registered in PROSPERO (ID: CRD42024566535). Ethical approval was not required because this systematic literature review focused on retrospective studies.

Search Strategy

This review collected the records from 3 major databases up to October 2023. A second search was performed in February 2024 to complement newly published studies. Those databases include PubMed, Embase, and the Cochrane Library (CENTRAL), which are recommended academic search systems for systematic reviews and meta-analyses [38]. We used the MeSH (Medical Subject Headings) and Emtree headings in several combinations and supplemented them with free text to increase sensitivity. In addition, we searched references contained in the included studies to supplement the relevant literature. An experienced librarian designed and implemented the search strategy. The following MeSH terms were used for PubMed: "Spinal Stenosis," "Intervertebral Disc Degeneration," "Lumbar Vertebrae," "Machine Learning," "Deep Learning," and "Neural Networks, Computer*." The details of the search strategy are stated in Multimedia Appendix 2.

Inclusion and Exclusion Criteria

We included studies that evaluated the assessment value of DL or TML algorithms for diagnosing LSS and that were available in English. The included studies in the meta-analysis should provide or could be reconstructed as a 2×2 confusion matrix from sensitivity, specificity, and precision. Applied statistical, non–artificial intelligence, and general AI methods are not considered DL or TML. Articles with duplicated or unavailable data were excluded. Furthermore, abstracts from protocols, case reports, editorials, and review articles were excluded.

Review Process

A total of 2 reviewers (TW and NF) independently performed an initial screening of the titles and abstracts of the remaining articles to determine potential eligibility after removing duplicates. We reviewed the full texts of the remaining articles and excluded those that did not meet the inclusion criteria. We searched and screened a list of references for all relevant studies and a systematic review of potentially relevant studies. Disagreements were resolved by discussion and by third-party adjudication when necessary. For studies enrolled in systematic review while lack of available data for meta-analysis, an email was sent to the corresponding authors for acquisition of the necessary data.

XSL•FO
RenderX

Data Extraction

A total of 2 reviewers independently extracted, summarized, and tabulated the following data using a standard form: baseline characteristics of studies, including the publication year, study type, model type, algorithms used, LSS classifications, number of participants, validation strategy, imaging modality, and diagnosis criteria of LSS. Any discrepancies in the extracted data were resolved by discussion. For the studies that provided multiple contingency tables based on different classifier algorithms, datasets, LSS types, or label strategies, we assumed these to be independent of each other. For the studies that provided multiple contingency tables based on different preprocessing strategies, we selected the best-performing result. If there was no preprocessing strategy that performed significantly better than the others, we also enrolled each strategy as an individual study and collected the corresponding results. For repeat test results based on the same classifier algorithms, datasets, and so on, we calculated the average values of metrics as the final results.

Quality Assessment

A total of 2 reviewers (TW and NF) used the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2), which is a tool for assessing the quality of primary diagnostic accuracy studies, to independently assess the risk of bias for each eligible study [39]. The QUADAS-2 criteria assessed the risk of bias in 4 domains: patient selection, index test, reference standard, and flow and timing. Any disagreements were resolved by discussion with a third author.

Statistical Analysis

We used the MIDAS module and the METAPROP module [40] of Stata (version 17.0; StataCorp) for statistical analysis. Postestimation procedures for model diagnostic were used to assess heterogeneity using the I^2 statistic. The following metrics were used: 0%-40% (low heterogeneity), 30%-60% (moderate heterogeneity), 50%-90% (substantial heterogeneity), and 75%-100% (considerable heterogeneity). Bivariate mixed-effects logistic regression modeling was conducted, and forest plots were used to compare the sensitivity and specificity of DL or TML models for diagnosing LSS. We used summary receiver operating characteristic (SROC) curves to assess overall diagnostic accuracy. We used the Fagan nomogram to explore the relationship between pretest probability, likelihood ratio (LR), and posttest probability. LR dot plots were divided into 4 quadrants according to the strength of the evidence threshold, which was used to determine DL or TML model exclusion and confirmation. Finally, subgroup analyses were performed to examine whether the estimated sensitivity, specificity, and associated I^2 differed by several moderators when each subgroup included ≥4 datasets.

Results

Study Selection and Characteristics

The initial search identified 934 titles and abstracts, of which 269 were duplicates. After screening, 567 articles were excluded following this study's inclusion and exclusion criteria. In addition, 98 studies were reviewed for full text, of which 19 and 12 studies were included in the systematic review and meta-analysis, respectively (Figure 1). Table 1 summarizes the characteristics of the studies in the systematic review and meta-analyses, including study type, model type, algorithms used, LSS classifications, number of participants, validation strategy, imaging modality, and diagnosis criteria of LSS. The 19 studies included in the systematic review were published from 2016 to 2024. The 12 studies included in the meta-analysis were all retrospective and included 21 external tests [17,25-27,30,32,34] and 35 internal tests [17,20,22,24,26-28,33,34]. Therefore, the meta-analysis included 56 datasets and completely different data sources. Among the 56 datasets, 32 identified LSS on MRI [17,20,22,24,25,27,28,30,32], 20 on x-ray [26,34], and 4 on CT [33]. Furthermore, 29 datasets have developed and internally tested DL models [17,20,24,26-28,33,34], 6 datasets internally tested TML models [22,28], and 21 datasets externally tested the DL models [17,25-27,30,32,34].

XSL•FO
RenderX

Figure 1. Flowchart depicting PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) search strategy.

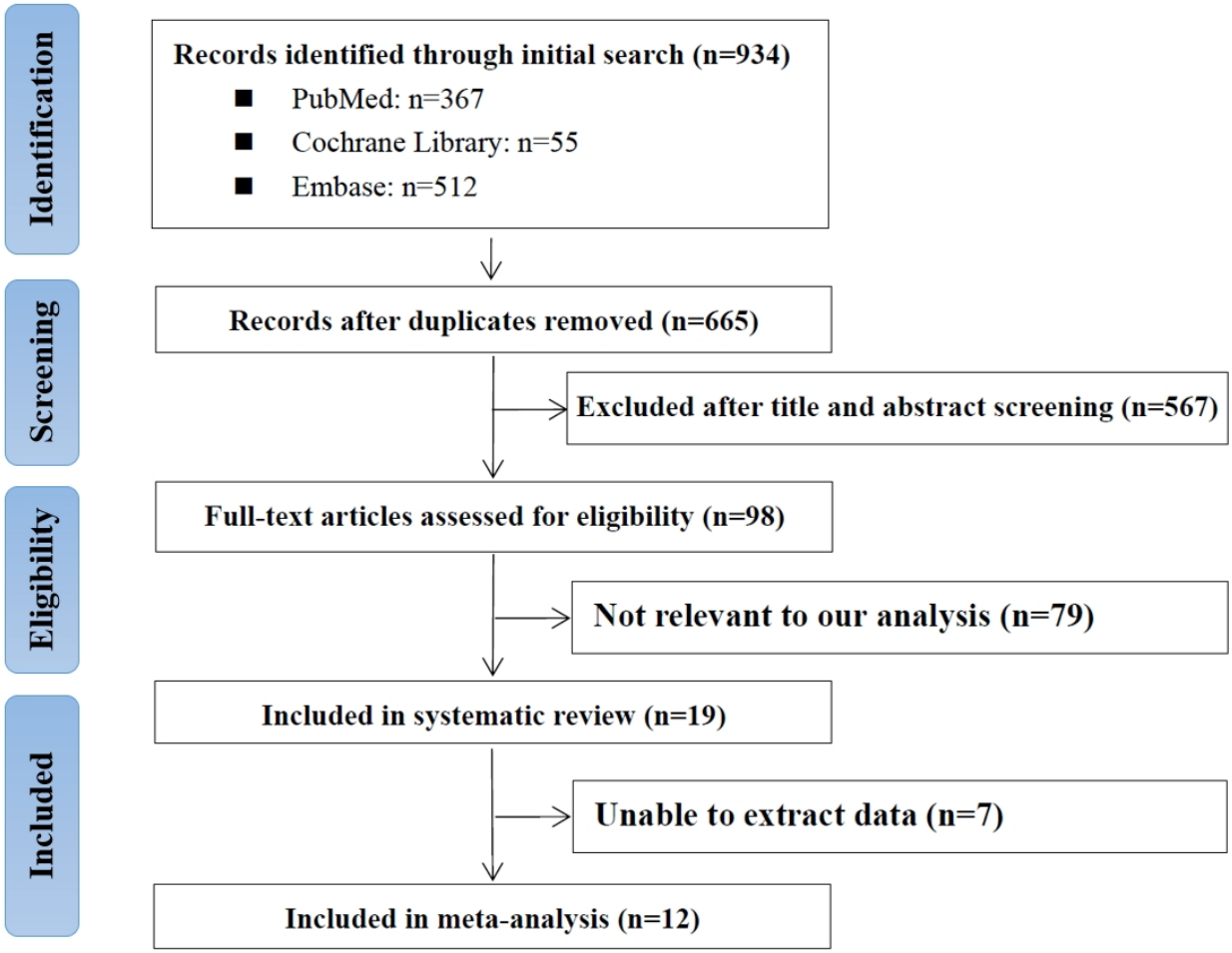

XSL•FO
RenderX

Table 1. Characteristics of the included studies in the systemic review and meta-analysis.

Study	Study type	Model type	Algorithms used[a]	LSS[b] type	Number of participants, n	Validation strategy	Imaging modality	Diagnosis criteria
He et al [16]	Development study and internal test	TML[c]	KNN[d] and SVM[e] and LDA[f]	LFS[g]	110	Cross-validation	MRI[h]	Lee et al [5]
Jamaludin et al [18]	Development study and internal test	DL[i]	CNN[j] (SpineNet[k])	LCS[l]	2009	Hold-out validation	MRI	No reference
Zhang et al [19]	Development study and internal test	TML	SVM and Decision Tree	LCS and LFS	600	Hold-out validation	MRI	No reference
Lu et al [21]	Development study and internal test	DL	ResNeXt-50	LCS and LFS	4075	Hold-out validation	MRI	No reference
Han et al [20][m]	Development study and internal test	DL	CNN and FCN and SegNet and DeepLabv3+ and U-Net	LFS	253	Cross-validation	MRI	Lee et al [5]
Huber et al [22][m]	Development study and internal test	TML	Decision Tree	LCS	82	Cross-validation	MRI	Lee et al [9] and Schizas et al [6]
Ishimoto et al [23]	Replication study and internal test	DL	CNN (SpineNet[k])	LCS	971	Hold-out validation	MRI	Lurie et al [8]
Won et al [24][m]	Development study and internal test	DL	VGG	LCS	542	Cross-validation	MRI	Schizas et al [5]
Hallinan et al [17][m]	Development study and internal test and external test	DL	CNN	LCS and LRS[n] and LFS	446/100	Hold-out validation	MRI	Lurie et al [8] and Bartynski and Lin [10]
Lehnen et al [25][m]	External test	DL	CNN (CoLumbo[k])	LCS	146	—[o]	MRI	Lee et al [9]
Grob et al [30][m]	External test	DL	CNN (SpineNet[k])	LCS	882	—	MRI	Lurie et al [8]
Kim et al [26][m]	Development study and internal test and external test	DL	VGG19 and VGG16 and ResNet50 and Efficient1	LCS	4644/199	Cross-validation	X-Ray	Lee et al [9]
Su et al [27][m]	Development study and internal test and external test	DL	ResNet-50	LCS	1015/100	Hold-out validation	MRI	Lee et al [9] and Park et al [11]
Altun et al [28][m]	Development study and internal test	TML and DL	RF and SVM and VGG16 and ResNet and MobileNet and InceptionNet	LSS	1030	Cross-validation	MRI	No reference
Bharadwaj et al [29]	Development study and internal test	TML and DL	Decision Tree and BiTCNN	LCS and LFS	200	Hold-out validation	MRI	Schizas et al [6] and Lee et al [5]
Tumko et al [32][m]	Development study and external test	DL	RegNetY32GF	LCS and LRS and LFS	1635/150	—	MRI	Schizas et al [6]
Shahzadi et al [31]	Development study and internal test	DL	CNN	LRS and LFS	515	Cross-validation	MRI	No reference
Li et al [33][m]	Development study and internal test	DL	VGG11 and ResNet-18	LCS and LRS	236	Hold-out validation	CT	Lurie et al [8] and Bartynski and Lin [10]
Park et al [34][a]	Development study and internal test and extra-internal test and external test	DL	ResNet50 and VGG19 and VGG16 and EfficientNet-B1	LCS	3831/199/100	Cross-validation (validation) and Hold-out validation (Internal test)	X-Ray	No reference

XSL•FO
RenderX

[a]Algorithms for only classifiers.
[b]LSS: lumbar spinal stenosis.
[c]TML: traditional machine learning.
[d]KNN: k-nearest neighbors.
[e]SVM: support vector machine.
[f]LDA: linear discriminant analysis.
[g]LFS: lumbar foraminal stenosis.
[h]MRI: magnetic resonance imaging.
[i]DL: deep learning.
[j]CNN: convolutional neural network.
[k]Name of software.
[l]LCS: lumbar central stenosis.
[m]Studies included in meta-analysis (confusion matrix available or can be reconstructed).
[n]LRS: lateral recess stenosis.
[o]Not applicable.

Methodological Quality

Regarding the QUADAS-2 risk of bias assessment (Figure 2 [17,20,22,24-28,30,32-34]), we revealed 4 studies with a high risk of bias [26,28,30,33], 3 with an unclear risk of bias [20,25,34], and 5 with a completely low risk of bias [17,22,24,27,32]. In particular, 2 of the included studies reported no details of patient selection [26,28], causing a high bias in patient selection. Furthermore, 1 study provided unclear information on how to perform the index test [28], thereby causing an unclear risk of bias. Furthermore, 1 study used the improper reference standard, which was not likely to correctly classify the target condition [30], causing a high risk of bias in the reference standard. Besides, 1 study showed a high risk of bias with regard to flow and timing issues [33].

XSL•FO
RenderX

Figure 2. Methodological assessment by Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2). (A) The proportion of risk of bias for all domains and proportion of applicability concerns in three domains. (B) Summary of the risk of bias for each study. Green, yellow, and red circles indicate low, unclear, and high risk of bias, respectively [8,20,22,24-28,30,32-34].

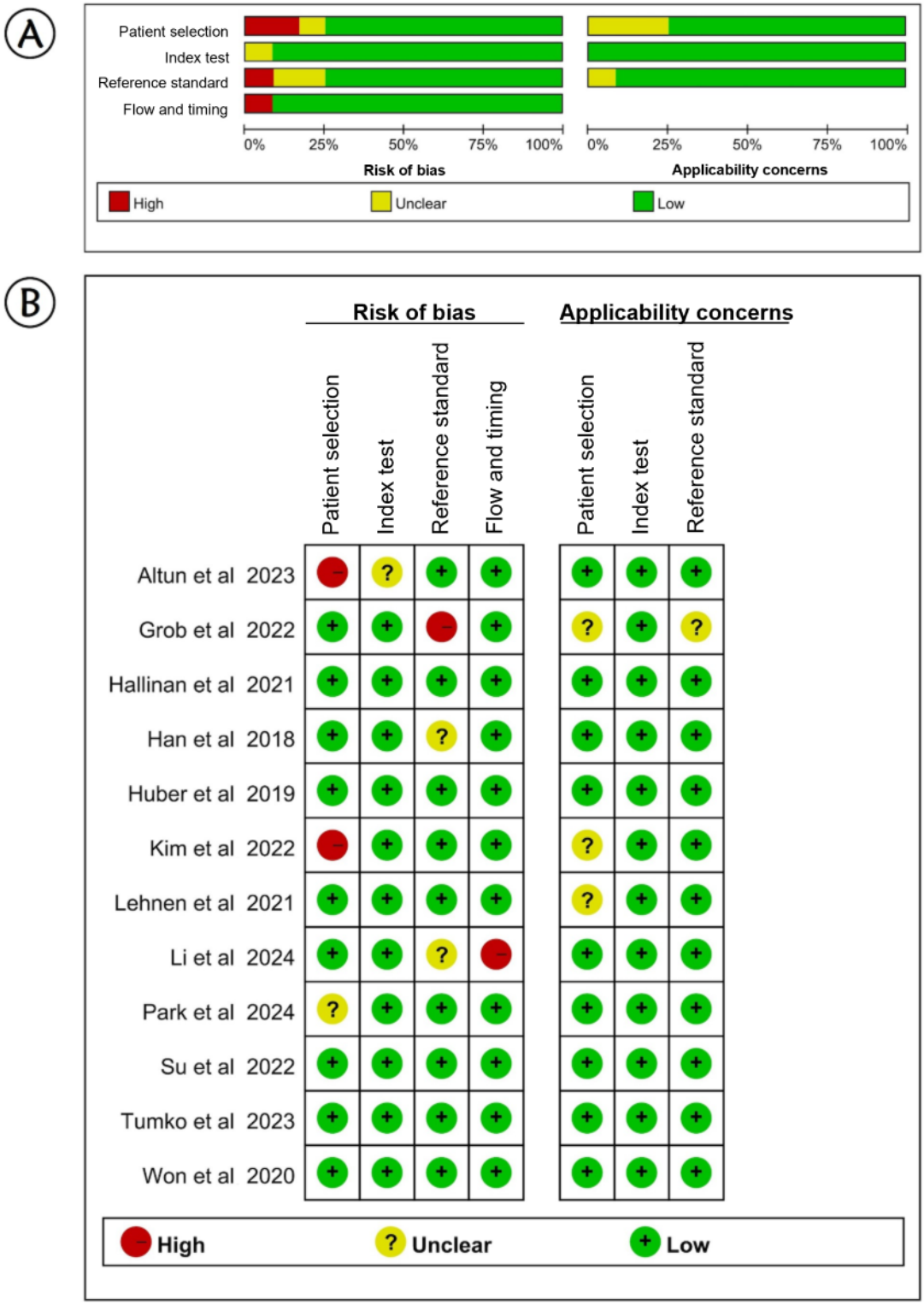

Performance of TML and DL Models for LSS

A total of 12 studies with 15,044 patients reported the assessment value of TML or DL models for diagnosing LSS. The pooled sensitivity was 0.84 (95% CI 0.82-0.86; I^2=99.06%), and specificity was 0.87 (95% CI 0.84-0.90; I^2=98.7%; Figure 3). The diagnostic odds ratio was 36 (95% CI 26-49). The SROC curve (Figure 4) revealed that the area under the curve of TML or DL models for diagnosing LSS was 0.92 (95% CI 0.89-0.94), indicating a high diagnostic value.

XSL•FO
RenderX

Figure 3. Forest plots in sensitivity and specificity of traditional machine learning (TML) or deep learning (DL) models. The pooled sensitivity and specificity were 0.84 (95% CI 0.82-0.86) and 0.87 (95% CI 0.84-0.90), respectively.

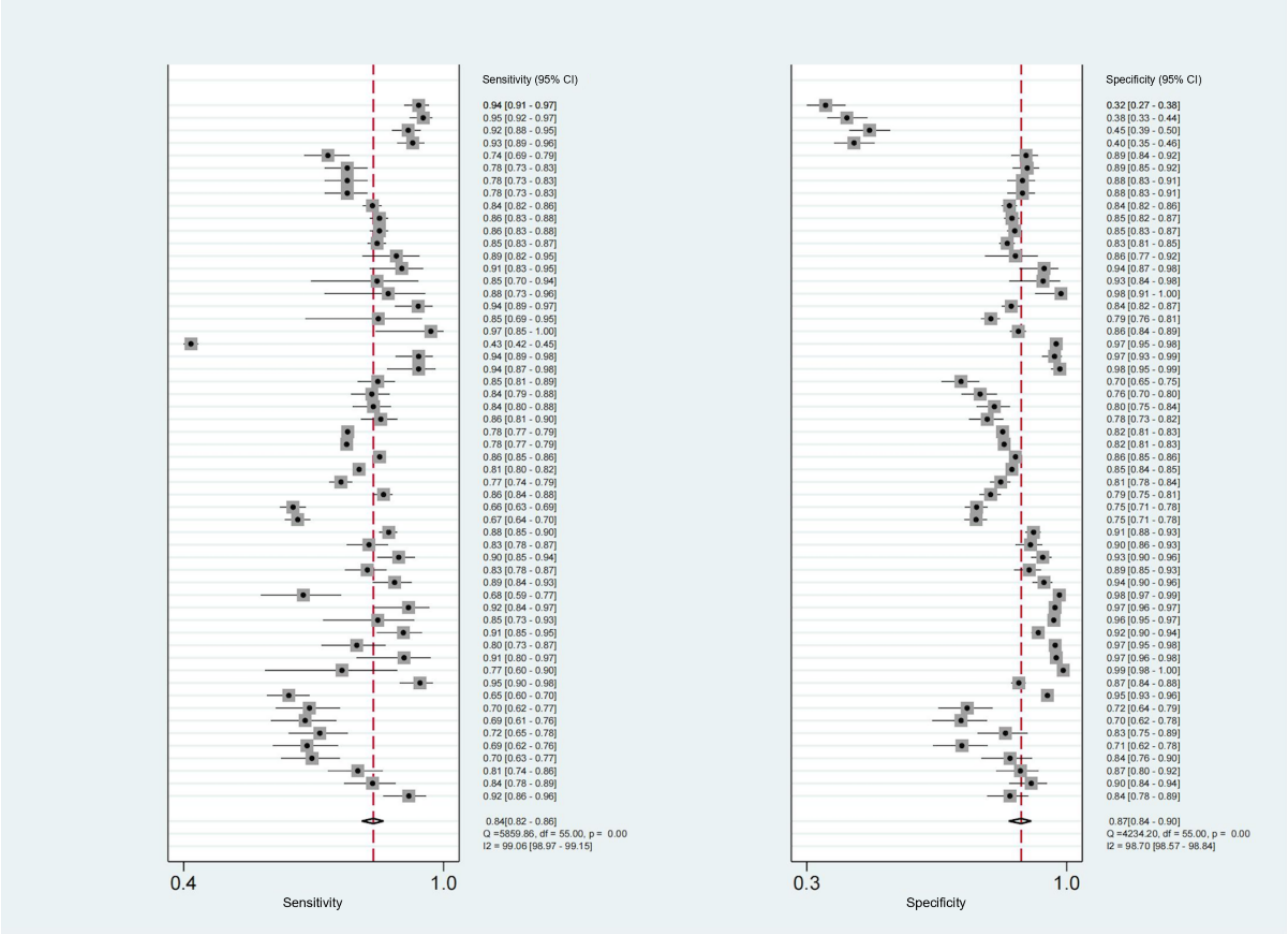

XSL•FO
RenderX

Figure 4. The SROC graph for the studies. The AUC of summary receiver operating characteristic (TML) or deep learning (DL) models for diagnosing LSS was 0.92 (95% CI 0.89-0.94). SROC: summary receiver operating characteristic; AUC: area under the curve; SENS: sensitivity; SPEC: specificity.

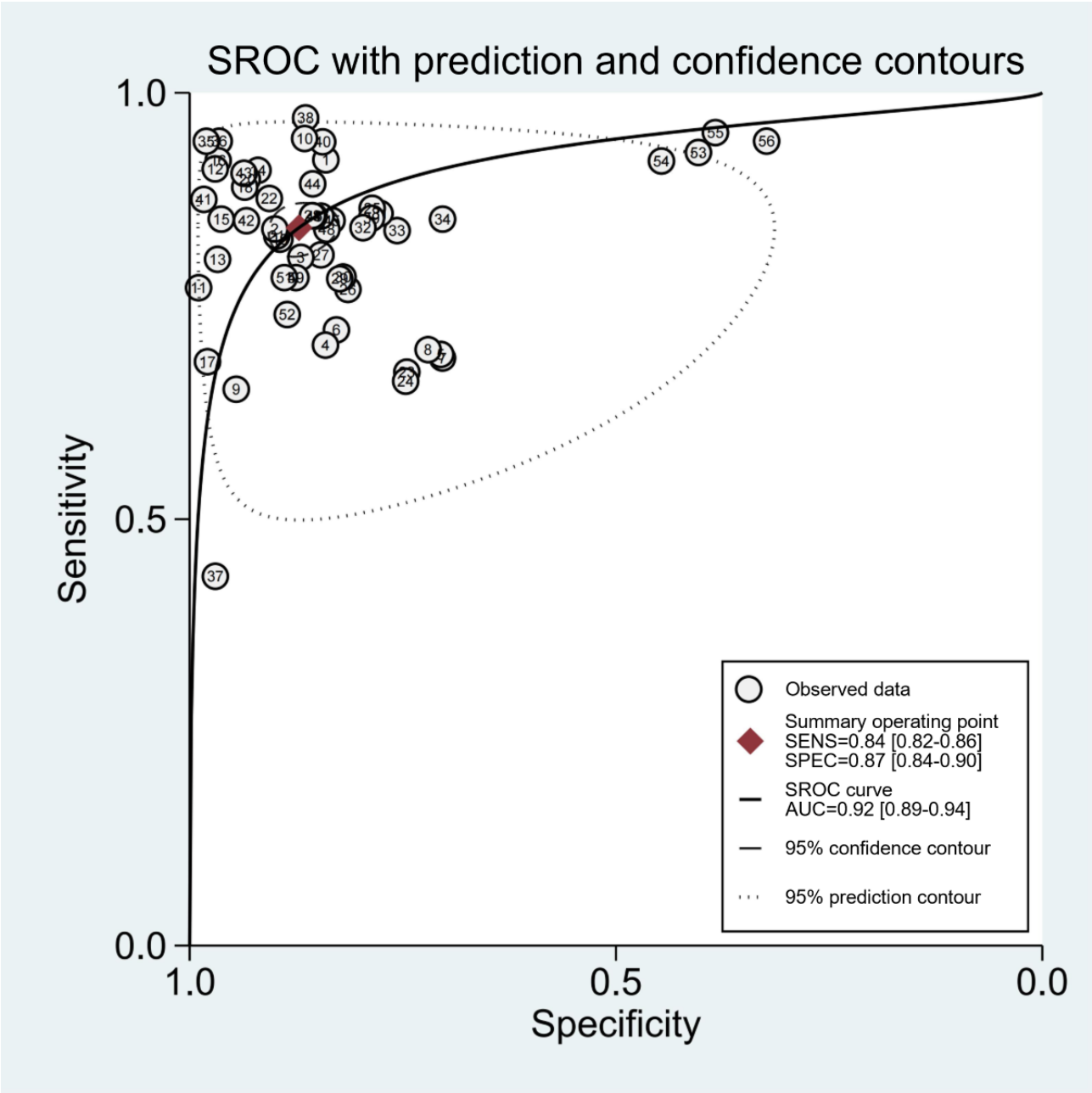

We set the pretest probability to 50% based on the pretest probability of disease. At this point, true positives accounted for 87% when patients were diagnosed with LSS by the TML or DL model, and false negatives accounted for 15% when the diagnosis was not LSS (Figure 5). Furthermore, the models showed a positive likelihood ratio (LR+) of 6.6 (95% CI 5.1-8.4) and a negative likelihood ratio (LR–) of 0.18 (95% CI 0.16-0.21), respectively (Figure 5). However, the summary likelihood ratio plot of TML or DL models was in the right lower quadrant (LR+<10 and LR–>0.1: no exclusion or confirmation), and the individual plots were scattered and distributed (Figure 6). The results indicated that although the TML or DL models achieved an acceptable performance generally, it was still insufficient enough for diagnosing or excluding LSS, and the current models suffered from certain performance variations.

XSL•FO
RenderX

Figure 5. Fagan nomogram of traditional machine learning (TML) or deep learning (DL) models for diagnosing lumbar spinal stenosis (LSS). The first column of this nomogram represents the pretest probability, the second column represents the likelihood ratio, and the third shows the posttest probability.

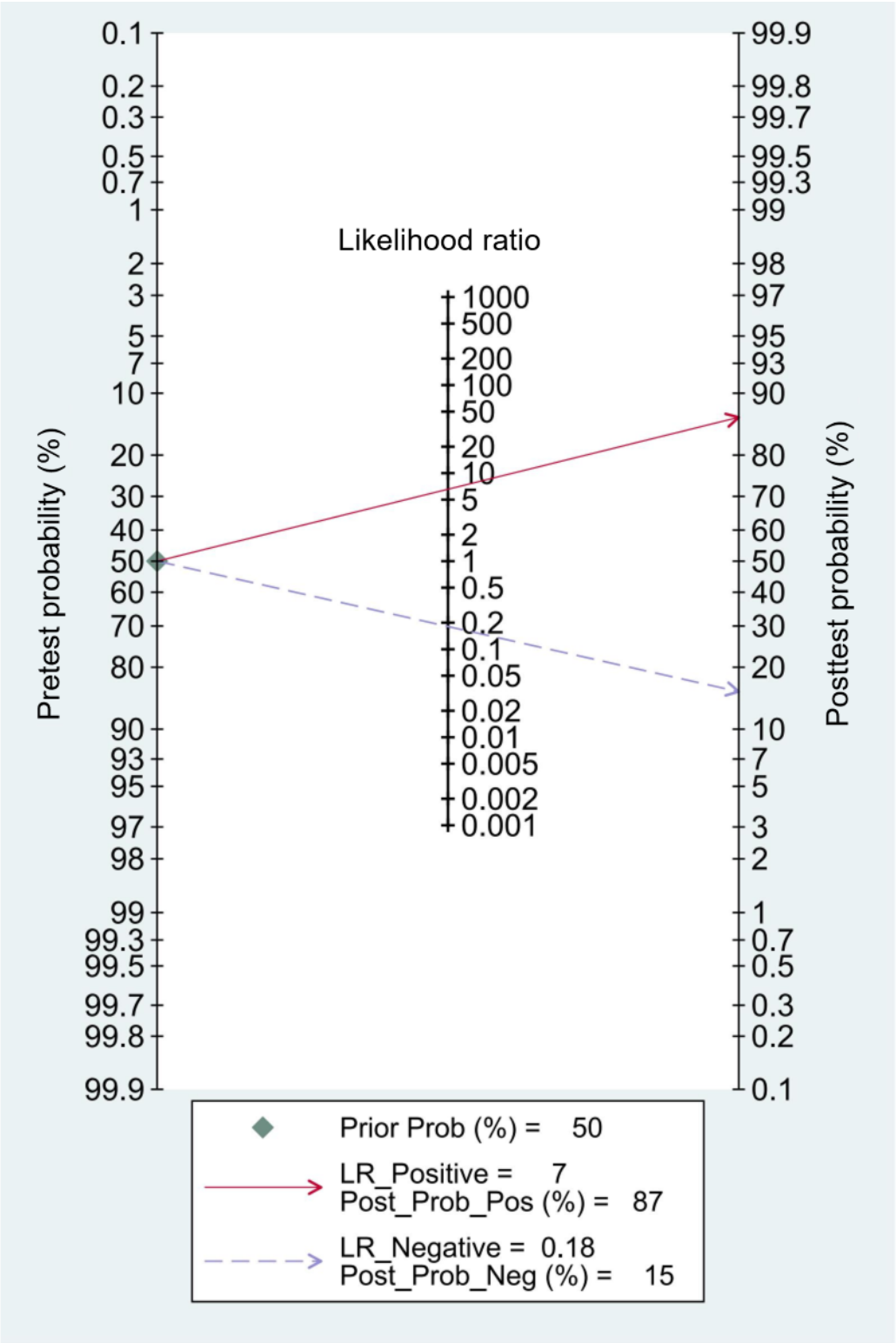

XSL•FO
RenderX

Figure 6. Likelihood ratio (LR) dot plot of traditional machine learning (TML) or deep learning (DL) prediction models. The summary point of TML or DL models was in the right lower quadrant (LR+<10 and LR–>0.1: no exclusion or confirmation). LRN: negative likelihood ratio; LRP: positive likelihood ratio; LUQ: left upper quadrant; RLQ: right lower quadrant; RUQ: right upper quadrant.

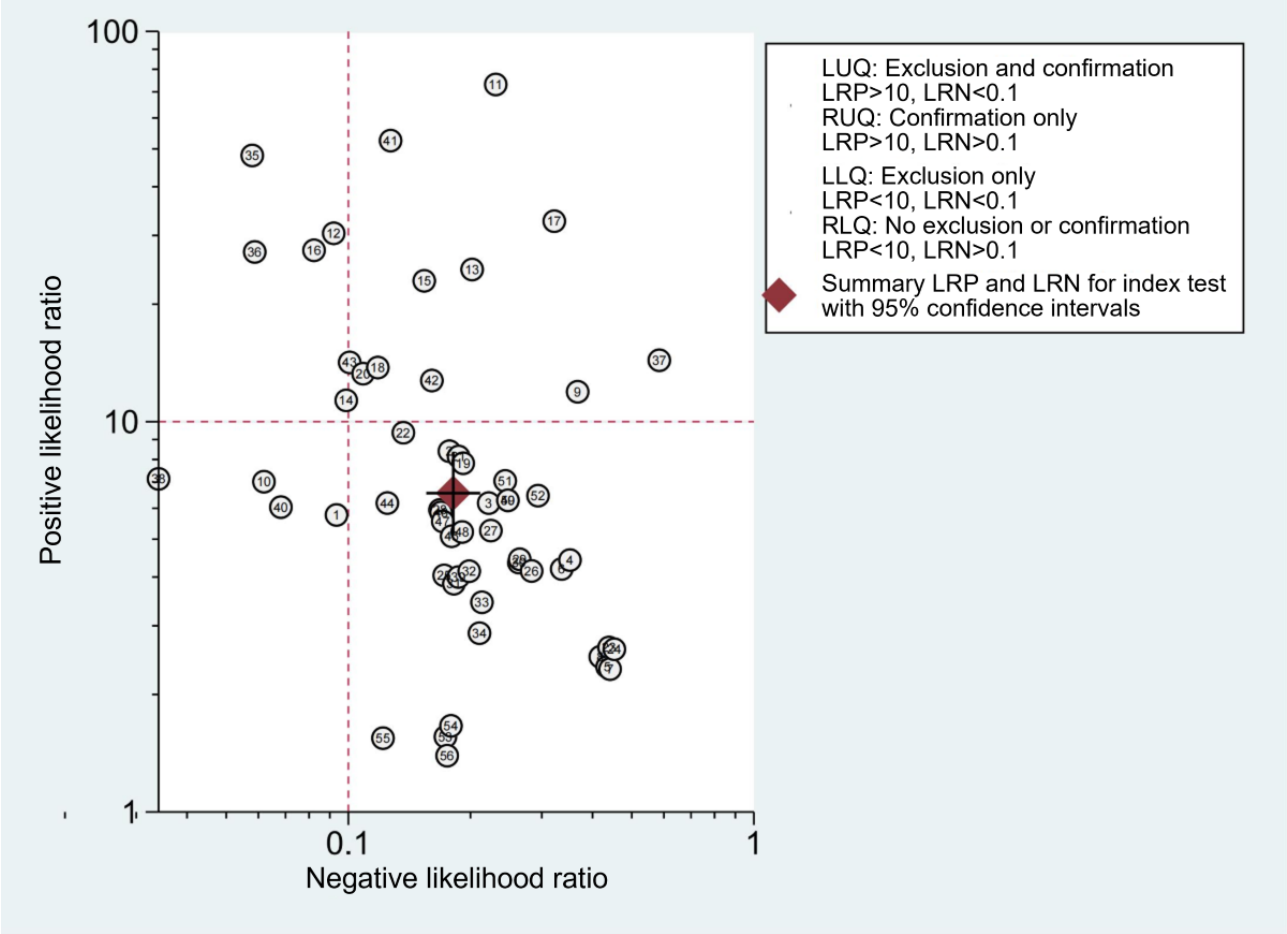

In total, 4 studies [17,23,29,32] simultaneously provided the performance of reliability both of observers and TML or DL models, including 3 studies [17,29,32] that performed a direct comparison between their reliabilities based on the same assessment datasets (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of the studies available for the agreement between models and observers and reference standard.

Study	Number of participants, n	Agreement assessment strategy	Control group	Model type	Type of classification	LSS[a] type	Model results	Control group results
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL[b]	Binary	LCS[c]	0.96	0.98/0.98
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Binary	LRS[d]	0.92	0.92/0.95
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Binary	LFS[e]	0.89	0.94/0.95
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Multigrading	LCS	0.82	0.89/0.89
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Multigrading	LRS	0.72	0.71/0.79
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Multigrading	LFS	0.75	0.80/0.87
Bharadwaj et al [29]	200	Cohen κ	2 radiologists	DL	Multigrading	LCS	0.54	0.80/0.86
Bharadwaj et al [29]	200	Cohen κ	2 radiologists	TML[f]	Multigrading	LCS	0.80	0.80/0.86
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Binary	LCS	0.431	Average 0.372
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Binary	LRS	0.315	Average 0.323
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Binary	LFS	0.672	Average 0.596
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Multigrading	LCS	0.310	Average 0.376
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Multigrading	LRS	0.199	Average 0.359
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Multigrading	LFS	0.637	Average 0.620

[a]LSS: lumbar spinal stenosis.
[b]DL: deep learning.
[c]LCS: lumbar central stenosis.
[d]LRS: lateral recess stenosis.
[e]LFS: lumbar foraminal stenosis.
[f]TML: traditional machine learning.

Subgroup Analysis

We conducted the subgroup analyses in 3 areas, including data partition (internal test or external test), model networks (TML or DL), and image (MRI or x-ray), to effectively understand how the 3 different types affected the performance of the algorithm for LSS assessment (Table 3). The internal test group demonstrated a lower sensitivity ($P<.01$) yet higher specificity ($P<.01$) than the external test group. Besides, the MRI group showed a lower sensitivity ($P<.01$) yet higher specificity ($P<.01$) than the x-ray group. The sensitivity in the DL group achieved 0.85, which was significantly higher than that (0.80) in the TML group ($P<.01$). Meanwhile, the DL group showed a more stable performance on specificity than the TML group ($P=.04$).

Table 3. Results of subgroup analysis.

Categories	Studies, n	Sensitivity (95% CI[a])	*P* value (HBG[b] of sensitivity)	Specificity (95% CI)	*P* value (HBG of specificity)
Data partition			<.001		<.001
Internal test	35	0.83 (0.80-0.86)		0.89 (0.85-0.92)	
External test	21	0.86 (0.82-0.90)		0.85 (0.79-0.91)	
Model networks			<.001		.04
TML[c]	6	0.80 (0.72-0.89)		0.87 (0.77-0.97)	
DL[d]	50	0.85 (0.82-0.87)		0.87 (0.84-0.91)	
Image			<.001		<.001
MRI	32	0.83 (0.79-0.86)		0.91 (0.88-0.93)	
X-ray	20	0.85 (0.82-0.89)		0.77 (0.70-0.84)	

[a]CI: confidence interval.
[b]HBG: heterogeneity between groups.
[c]TML: traditional machine learning.
[d]DL: deep learning.

XSL•FO
RenderX

Discussion

Principal Findings

In recent years, there has been a boom in assessing the diagnosis and grading of LSS by TML or DL methods. After systemically reviewing the available evidence, we revealed that all related studies were published after 2016 and increased annually. It can also be said that TML and DL algorithms have been showing promising potential in this field. To the best of our knowledge, this is the first systematic review and meta-analysis for addressing this issue. Our pooled results showed an overall sensitivity of 0.84 and a specificity of 0.87 for diagnosing LSS by TML or DL models. The area under the SROC was 0.92, indicating a high diagnostic value. Subgroup analysis revealed a better diagnostic performance in internal validation than in external validation, while DL algorithms demonstrated higher sensitivity and specificity than TML algorithms. However, 37% of studies enrolled in the systemic review were unavailable in the meta-analysis, which may have caused a discrepancy between pooled results and reality. Therefore, the results should be interpreted with caution.

A permanent debate focuses on whether the diagnostic performance of ML or DL algorithms surpassed that of clinicians. High-level evidence showed that the performance of AI diagnostic systems is equivalent to health care professionals, and AI-assistance systems improve clinician diagnostic performance [41-43]. However, in the field of LSS, there were few studies designed to directly compare the performance of additional radiologists or orthopedic surgeons with ML or DL algorithms in the same dataset. Hallinan et al [17] developed a DL method for diagnosing different LSS and compared the sensitivity and specificity of the DL model with 2 independent clinicians (a neuroradiologist and a musculoskeletal radiologist) with less than 10 years of experience. The study revealed that the sensitivity of DL in detecting LSS was on par with clinicians in general, with even slightly higher in lumbar central stenosis (LCS) and lateral recess stenosis (LRS), but with lower specificity of DL. It is reasonable because pursuing sensitivity to reduce false-negative results on the premise of maximizing the accuracy and AUC may be an alternative and beneficial method for clinical demands [44]. Compared with the complete replacement of clinicians, AI diagnostic systems are more expected to be assisted screening tools to use in areas with poor medical resources without experts or to reduce the workload of clinicians and missed diagnoses, followed by high-level medical team screening of image marked positive by the automatic diagnosis [45].

Although the general performance of diagnostic models was satisfactory, it was still insufficient enough for diagnosing or excluding LSS according to the summary likelihood ratio plot. Besides, our systemic review and meta-analysis found that ML or DL models showed similar, even slightly lower, sensitivity compared with specificity in general, especially in the MRI modality. There may be several reasons. First, the complexity and variety of pathological structures in individuals with LSS result in no broadly accepted quantitative radiologic evidence for diagnosis, even in expert evaluation [46], which makes automatic detection by MRI difficult. Furthermore, we cannot exclude that the results may be influenced by heterogeneity. Consideration should be taken for developers to optimize models prone to higher sensitivity than specificity for diagnosis and grading of LSS, which may be more beneficial to clinical workflow.

Notably, a consensus of reference standards in determining ML or DL performance for diagnosing LSS has not been reached till now. The reference standards in almost all included studies were labeled by qualitative or semiquantitative expert evaluation, which suffered from considerable heterogeneity due to the different amount, specialties, and years of experience of experts. Huber et al [22] combined texture analysis and decision trees to detect LSS based on the cross-sectional area (CSA) as a quantitative reference standard. However, a CSA of <130mm^2 was not a widely accepted criterion, and it is only appropriate for LCS, while quantitative radiological criteria remain unavailable for diagnosing LRS or LFS [46,47]. More comprehensive and rigorous criteria for reference standards should be developed in future work. In addition, the diagnosis of LSS should combine the imaging findings with history and clinical presentation because LSS is a clinical syndrome, and solely radiographic LSS may be symptom-free [2]. However, the diagnosis criteria in all reviewed studies were only based on radiographic criteria or reports, which means that the current TML or DL models were developed for the diagnosis of radiographic LSS objectively. Yet, it is not said that radiographic evaluation is valueless for LSS. On the one hand, it can provide details in pathological anatomy, which guides further treatment options and surgical approaches. On the other hand, a potentially imperceptible relationship between radiographic characters and clinical LSS may be explored with the help of AI models. Therefore, we suggest attempting to label the data by clinical LSS as golden standards on the premise of model interpretability and eliminating confounding factors. Furthermore, developers can set multiple data types, such as crucial details of patient's history, physical examination, and imaging tests, as inputs to build a multimodal to improve the clinical value of LSS diagnosis and grading by AI approaches [48].

Overall, our meta-analysis revealed a better performance for diagnosing LSS in DL than in TML. Whereas results should be interpreted with caution because of the limited number of enrolled studies on TML in meta-analysis. Only 2 included studies in the systemic review designed a direct comparison of the capability of DL and TML models for diagnosing LSS, yet showed contradictory results. Altun et al [28] found that VGG16 and 3 other DL techniques performed better in addressing the issue of binary LSS classification compared with random forest and support vector machines. Conversely, Bharadwaj et al [29] combined segmentation with DL and TML classifiers to conduct multiclass and binary LSS grading. Both accuracy, AUC, and reproducibility were higher in the TML group [29]. The inconsistency may be attributed to the scale of training data. DL was generally acknowledged as the most outstanding ML technique for automatic medical image analysis [49]. However, DL is restricted to a stronger data dependency compared with any other ML, as it is designed with a more complex architecture [14,15,50]. In particular, there is an extreme need for DL to be

XSL•FO
RenderX

trained with a sufficiently large sample set, particularly considering the complexity of spinal MRI. In the 2 studies above, the scales of training samples were more than 5 times higher in the study of Altun et al [28] than that of Bharadwaj et al [29] (927 vs 170). Hence, we recommend that a larger scale dataset for training both TML and DL models is beneficial for exploring their capabilities in diagnosing and grading LSS in order to reduce data overfitting and improve the performance of models.

Any AI diagnostic systems should be clinically oriented instead of technically oriented. This poses a challenge for developers in developing ML or DL models more appropriately for clinical practice rather than more technically challenging. Currently, although promising results in exactitude (accuracy, sensitivity, etc) have been widely reported, other aspects of great clinical significance, like reliability, usability, and safety, were rarely assessed. A good agreement between the model and reference standard can verify the validity and reliability of the model. However, only 3 included studies performed a direct comparison of reliabilities between models and observers [17,29,32], and a generally higher consistency of diagnosis was achieved by clinicians than that of ML or DL systems. Besides, external validation is a valuable approach to validate the generalizability of ML or DL algorithms, testing their capabilities for adapting the differences between initial settings of data collection, imaging tests, and imaging processing with replication or real-world settings [45]. Inspiringly, this meta-analysis showed the performances of external validation were generally on par with that of the internal validation, with better sensitivity but worse specificity. However, the results may be inconclusive because only 37% of studies (7/19) enrolled in this systemic review tested the models by external validation (separate datasets for model validation only) [17,25-27,30,32,34].

Anyway, a gap remains between current ML or DL algorithms for diagnosing and grading LSS with real clinical applications. A recent review highlighted the importance of large-scale and mixed-source datasets, clinician collaboration, and a clear statement of data collection to facilitate DL in clinical applications [15]. Furthermore, only a few software, such as SpineNet (University of Oxford) [18,23,30] and CoLumbo (SmartSoft Ltd) [25], were introduced into public view despite several AI models having been developed in this field. We urged that the exploration of software design may be beneficial to extend the application of AI diagnostic models.

Limitations

Several limitations exist in this systematic review and meta-analysis. First, most of the enrolled studies were conducted under small sample sizes, and only 6 studies (32%) had a sample size >1000 [18,21,26,27,32,34]. However, a large-scale dataset is warranted for both training and validation in AI diagnostic algorithms, especially for DL algorithms [14,15,50]. Second, few models performed external validation to test the reproducibility and extensibility. Thus, the reported performance should be interpreted with caution. Third, only a few studies provided a contingency table, while the incompleteness of reported performance metrics made it difficult to conduct a comprehensive meta-analysis, which a recent systematic review and meta-analysis in the spine field also mentioned [51]. This may cause a discrepancy between pooled results and reality. Finally, the risk of bias in this study was identified by the QUADAS-2, which is more suitable for traditional diagnostic models [52]. A more specific and practical guideline for diagnostic AI models remains under development [53].

Conclusions

This systematic review and meta-analysis emphasize that despite the generally satisfactory diagnostic performance of artificial intelligence systems in the experimental stage for the diagnosis of LSS, none of them is reliable and practical enough to apply in real clinical practice. Further efforts, including optimization of model balance, widely accepted objective reference standards, multimodal strategy, large dataset for training and testing, external validation, and sufficient and scientific report, should be made to bridge the distance between current TML or DL models and real-life clinical applications in future studies.

Authors' Contributions

TW, RC, and NF contributed equally to this work. LZ contributed to the conception and design. TW, RC, and NF contributed to the acquisition of data. TW and NF contributed to the analysis and interpretation of data. TW and NF contributed to drafting the article. RC, NF, SY, PD, QW, AW, JL, XK, and WZ contributed to critically revising the article. LZ reviewed the submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for diagnostic test accuracy) checklist.
[DOCX File , 32 KB - jmir_v26i1e54676_app1.docx]

Multimedia Appendix 2
Details of search strategy.
[DOCX File , 13 KB - jmir_v26i1e54676_app2.docx]

XSL•FO
RenderX

References

1. Katz JN, Zimmerman ZE, Mass H, Makhni MC. Diagnosis and management of lumbar spinal stenosis: a review. JAMA 2022;327(17):1688-1699. [doi: 10.1001/jama.2022.5921] [Medline: 35503342]
2. Lurie J, Tomkins-Lane C. Management of lumbar spinal stenosis. BMJ 2016;352:h6234 [FREE Full text] [doi: 10.1136/bmj.h6234] [Medline: 26727925]
3. Ravindra VM, Senglaub SS, Rattani A, Dewan MC, Härtl R, Bisson E, et al. Degenerative lumbar spine disease: estimating global incidence and worldwide volume. Global Spine J 2018;8(8):784-794 [FREE Full text] [doi: 10.1177/2192568218770769] [Medline: 30560029]
4. Kreiner DS, Shaffer WO, Baisden JL, Gilbert TJ, Summers JT, Toton JF, North American Spine Society. An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis (update). Spine J 2013;13(7):734-743. [doi: 10.1016/j.spinee.2012.11.059] [Medline: 23830297]
5. Lee S, Lee JW, Yeom JS, Kim KJ, Kim HJ, Chung SK, et al. A practical MRI grading system for lumbar foraminal stenosis. AJR Am J Roentgenol 2010;194(4):1095-1098. [doi: 10.2214/AJR.09.2772] [Medline: 20308517]
6. Schizas C, Theumann N, Burn A, Tansey R, Wardlaw D, Smith FW, et al. Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. Spine (Phila Pa 1976) 2010;35(21):1919-1924. [doi: 10.1097/BRS.0b013e3181d359bd] [Medline: 20671589]
7. Sigmundsson FG, Kang XP, Jönsson B, Strömqvist B. Correlation between disability and MRI findings in lumbar spinal stenosis: a prospective study of 109 patients operated on by decompression. Acta Orthop 2011;82(2):204-210 [FREE Full text] [doi: 10.3109/17453674.2011.566150] [Medline: 21434811]
8. Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino JA, Kaiser J, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine (Phila Pa 1976) 2008;33(14):1605-1610 [FREE Full text] [doi: 10.1097/BRS.0b013e3181791af3] [Medline: 18552677]
9. Lee GY, Lee JW, Choi HS, Oh K, Kang HS. A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. Skeletal Radiol 2011;40(8):1033-1039. [doi: 10.1007/s00256-011-1102-x] [Medline: 21286714]
10. Bartynski WS, Lin L. Lumbar root compression in the lateral recess: MR imaging, conventional myelography, and CT myelography comparison with surgical confirmation. AJNR Am J Neuroradiol 2003;24(3):348-360 [FREE Full text] [Medline: 12637281]
11. Park HJ, Kim SS, Lee YJ, Lee SY, Park NH, Choi YJ, et al. Clinical correlation of a new practical MRI method for assessing central lumbar spinal stenosis. Br J Radiol 2013;86(1025):20120180 [FREE Full text] [doi: 10.1259/bjr.20120180] [Medline: 23426848]
12. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505-515 [FREE Full text] [doi: 10.1148/rg.2017160130] [Medline: 28212054]
13. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020;13(1):69-76 [FREE Full text] [doi: 10.1007/s12178-020-09600-8] [Medline: 31983042]
14. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020;9(2):14 [FREE Full text] [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
15. Chen X, Wang X, Zhang K, Fung KM, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal 2022;79:102444 [FREE Full text] [doi: 10.1016/j.media.2022.102444] [Medline: 35472844]
16. He X, Yin Y, Sharma M, Brahm G, Mercado A, Li S. Automated diagnosis of neural foraminal stenosis using synchronized superpixels representation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI. Cham: Springer International Publishing; 2016.
17. Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. Radiology 2021;300(1):130-138. [doi: 10.1148/radiol.2021204289] [Medline: 33973835]
18. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. Med Image Anal 2017;41:63-73. [doi: 10.1016/j.media.2017.07.002] [Medline: 28756059]
19. Zhang Q, Bhalerao A, Hutchinson C. Weakly-supervised evidence pinpointing and description. In: Information Processing in Medical Imaging. Cham: Springer International Publishing; 2017.
20. Han Z, Wei B, Mercado A, Leung S, Li S. Spine-GAN: semantic segmentation of multiple spinal structures. Med Image Anal 2018;50:23-35. [doi: 10.1016/j.media.2018.08.005] [Medline: 30176546]
21. Lu JT, Pedemonte S, Bizzo B, Doyle S, Andriole KP, Michalski MH, et al. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. 2018 Presented at: Proceedings of the 3rd Machine Learning for Healthcare Conference; August 17-18, 2018; Palo Alto, California URL: https://proceedings.mlr.press/v85/lu18a.html
22. Huber FA, Stutz S, Vittoria de Martini I, Mannil M, Becker AS, Winklhofer S, et al. Qualitative versus quantitative lumbar spinal stenosis grading by machine learning supported texture analysis-Experience from the LSOS study cohort. Eur J Radiol 2019;114:45-50. [doi: 10.1016/j.ejrad.2019.02.023] [Medline: 31005175]

XSL•FO
RenderX

23. Ishimoto Y, Jamaludin A, Cooper C, Walker-Bone K, Yamada H, Hashizume H, et al. Could automated machine-learned MRI grading aid epidemiological studies of lumbar spinal stenosis? Validation within the wakayama spine study. BMC Musculoskelet Disord 2020;21(1):158 [FREE Full text] [doi: 10.1186/s12891-020-3164-1] [Medline: 32164627]
24. Won D, Lee HJ, Lee SJ, Park SH. Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks. Spine (Phila Pa 1976) 2020;45(12):804-812. [doi: 10.1097/BRS.0000000000003377] [Medline: 31923125]
25. Lehnen NC, Haase R, Faber J, Rüber T, Vatter H, Radbruch A, et al. Detection of degenerative changes on MR images of the lumbar spine with a convolutional neural network: a feasibility study. Diagnostics (Basel) 2021;11(5):902 [FREE Full text] [doi: 10.3390/diagnostics11050902] [Medline: 34069362]
26. Kim T, Kim YG, Park S, Lee JK, Lee CH, Hyun SJ, et al. Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs. J Neurosurg Spine 2022;37(1):104-111. [doi: 10.3171/2021.11.SPINE211136] [Medline: 35061993]
27. Su ZH, Liu J, Yang MS, Chen ZY, You K, Shen J, et al. Automatic grading of disc herniation, central canal stenosis and nerve roots compression in lumbar magnetic resonance image diagnosis. Front Endocrinol (Lausanne) 2022;13:890371 [FREE Full text] [doi: 10.3389/fendo.2022.890371] [Medline: 35733770]
28. Altun S, Alkan A, Altun İ. LSS-VGG16: diagnosis of lumbar spinal stenosis with deep learning. Clin Spine Surg 2023;36(5):E180-E190. [doi: 10.1097/BSD.0000000000001418] [Medline: 36727890]
29. Bharadwaj UU, Christine M, Li S, Chou D, Pedoia V, Link TM, et al. Deep learning for automated, interpretable classification of lumbar spinal stenosis and facet arthropathy from axial MRI. Eur Radiol 2023;33(5):3435-3443 [FREE Full text] [doi: 10.1007/s00330-023-09483-6] [Medline: 36920520]
30. Grob A, Loibl M, Jamaludin A, Winklhofer S, Fairbank JCT, Fekete T, et al. External validation of the deep learning system "SpineNet" for grading radiological features of degeneration on MRIs of the lumbar spine. Eur Spine J 2022;31(8):2137-2148. [doi: 10.1007/s00586-022-07311-x] [Medline: 35835892]
31. Shahzadi T, Ali MU, Majeed F, Sana MU, Diaz RM, Samad MA, et al. Nerve root compression analysis to find lumbar spine stenosis on MRI using CNN. Diagnostics (Basel) 2023;13(18):2975 [FREE Full text] [doi: 10.3390/diagnostics13182975] [Medline: 37761342]
32. Tumko V, Kim J, Uspenskaia N, Honig S, Abel F, Lebl DR, et al. A neural network model for detection and classification of lumbar spinal stenosis on MRI. Eur Spine J 2024;33(3):941-948. [doi: 10.1007/s00586-023-08089-2] [Medline: 38150003]
33. Li KY, Weng JJ, Li HL, Ye HB, Xiang JW, Tian NF. Development of a dep-learning model for diagnosing lumbar spinal stenosis based on CT images. Spine (Phila Pa 1976) 2024;49(12):884-891. [doi: 10.1097/BRS.0000000000004903] [Medline: 38112156]
34. Park S, Kim JH, Ahn Y, Lee CH, Kim YG, Yuh WT, et al. Multi-pose-based convolutional neural network model for diagnosis of patients with central lumbar spinal stenosis. Sci Rep 2024;14(1):203 [FREE Full text] [doi: 10.1038/s41598-023-50885-9] [Medline: 38168665]
35. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
36. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022;18(2):e1230 [FREE Full text] [doi: 10.1002/cl2.1230] [Medline: 36911350]
37. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
38. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods 2020;11(2):181-217 [FREE Full text] [doi: 10.1002/jrsm.1378] [Medline: 31614060]
39. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
40. Nyaga VN, Arbyn M, Aerts M. Metaprop: a stata command to perform meta-analysis of binomial data. Arch Public Health 2014;72(1):39 [FREE Full text] [doi: 10.1186/2049-3258-72-39] [Medline: 25810908]
41. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
42. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open 2021;4(3):e211276 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1276] [Medline: 33704476]
43. Xue P, Si M, Qin D, Wei B, Seery S, Ye Z, et al. Unassisted clinicians versus deep learning-assisted clinicians in image-based cancer diagnostics: systematic review with meta-analysis. J Med Internet Res 2023;25:e43832 [FREE Full text] [doi: 10.2196/43832] [Medline: 36862499]

XSL•FO
RenderX

44. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149(12):889-897 [FREE Full text] [doi: 10.7326/0003-4819-149-12-200812160-00008] [Medline: 19075208]
45. Lex JR, Di Michele J, Koucheki R, Pincus D, Whyne C, Ravi B. Artificial intelligence for hip fracture detection and outcome prediction: a systematic review and meta-analysis. JAMA Netw Open 2023;6(3):e233391 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.3391] [Medline: 36930153]
46. Mamisch N, Brumann M, Hodler J, Held U, Brunner F, Steurer J, Lumbar Spinal Stenosis Outcome Study Working Group Zurich. Radiologic criteria for the diagnosis of spinal stenosis: results of a delphi survey. Radiology 2012;264(1):174-179. [doi: 10.1148/radiol.12111930] [Medline: 22550311]
47. Andreisek G, Deyo RA, Jarvik JG, Porchet F, Winklhofer SFX, Steurer J, LSOS working group. Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. Eur Radiol 2014;24(12):3224-3232 [FREE Full text] [doi: 10.1007/s00330-014-3346-z] [Medline: 25079488]
48. Ramachandram D, Taylor GW. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag 2017;34(6):96-108. [doi: 10.1109/msp.2017.2738401]
49. Wu JH, Liu TYA, Hsu WT, Ho JHC, Lee CC. Performance and limitation of machine learning algorithms for diabetic retinopathy screening: meta-analysis. J Med Internet Res 2021;23(7):e23863 [FREE Full text] [doi: 10.2196/23863] [Medline: 34407500]
50. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020;1213:3-21 [FREE Full text] [doi: 10.1007/978-3-030-33128-3_1] [Medline: 32030660]
51. Compte R, Granville Smith I, Isaac A, Danckert N, McSweeney T, Liantis P, et al. Are current machine learning applications comparable to radiologist classification of degenerate and herniated discs and Modic change? A systematic review and meta-analysis. Eur Spine J 2023;32(11):3764-3787 [FREE Full text] [doi: 10.1007/s00586-023-07718-0] [Medline: 37150769]
52. Zhang Z, Yang L, Han W, Wu Y, Zhang L, Gao C, et al. Machine learning prediction models for gestational diabetes mellitus: meta-analysis. J Med Internet Res 2022;24(3):e26634 [FREE Full text] [doi: 10.2196/26634] [Medline: 35294369]
53. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7):e048008 [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]

Abbreviations

AI: artificial intelligence
CSA: cross-sectional area
CT: computed tomography
DL: deep learning
LCS: lumbar central stenosis
LFS: lumbar foraminal stenosis
LR: likelihood ratio
LRS: lateral recess stenosis
LSS: lumbar spinal stenosis
MeSH: Medical Subject Headings
ML: machine learning
MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
SROC: summary receiver operating characteristic
TML: traditional machine learning

Edited by Q Jin; submitted 18.11.23; peer-reviewed by Z Li, Y Wang, S Mao, A Claus; comments to author 31.05.24; revised version received 23.10.24; accepted 11.11.24; published 23.12.24.

Please cite as:
Wang T, Chen R, Fan N, Zang L, Yuan S, Du P, Wu Q, Wang A, Li J, Kong X, Zhu W
Machine Learning and Deep Learning for Diagnosis of Lumbar Spinal Stenosis: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e54676
URL: https://www.jmir.org/2024/1/e54676
doi:10.2196/54676
PMID:

XSL•FO
RenderX

©Tianyi Wang, Ruiyuan Chen, Ning Fan, Lei Zang, Shuo Yuan, Peng Du, Qichao Wu, Aobo Wang, Jian Li, Xiaochuan Kong, Wenyi Zhu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Exploring the Applications of Explainability in Wearable Data Analytics: Systematic Literature Review

Yasmin Abdelaal[1*], MSc; Michaël Aupetit[2*], PhD; Abdelkader Baggag[2*], PhD; Dena Al-Thani[1*], PhD

[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[2]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

[*]all authors contributed equally

Corresponding Author:
Yasmin Abdelaal, MSc
College of Science and Engineering
Hamad Bin Khalifa University
Penrose Building
Doha
Qatar
Phone: 974 74024682
Email: y.abdelaal98@gmail.com

Abstract

Background: Wearable technologies have become increasingly prominent in health care. However, intricate machine learning and deep learning algorithms often lead to the development of "black box" models, which lack transparency and comprehensibility for medical professionals and end users. In this context, the integration of explainable artificial intelligence (XAI) has emerged as a crucial solution. By providing insights into the inner workings of complex algorithms, XAI aims to foster trust and empower stakeholders to use wearable technologies responsibly.

Objective: This paper aims to review the recent literature and explore the application of explainability in wearables. By examining how XAI can enhance the interpretability of generated data and models, this review sought to shed light on the possibilities that arise at the intersection of wearable technologies and XAI.

Methods: We collected publications from ACM Digital Library, IEEE Xplore, PubMed, SpringerLink, JMIR, Nature, and Scopus. The eligible studies included technology-based research involving wearable devices, sensors, or mobile phones focused on explainability, machine learning, or deep learning and that used quantified self data in medical contexts. Only peer-reviewed articles, proceedings, or book chapters published in English between 2018 and 2022 were considered. We excluded duplicates, reviews, books, workshops, courses, tutorials, and talks. We analyzed 25 research papers to gain insights into the current state of explainability in wearables in the health care context.

Results: Our findings revealed that wrist-worn wearables such as Fitbit and Empatica E4 are prevalent in health care applications. However, more emphasis must be placed on making the data generated by these devices explainable. Among various explainability methods, post hoc approaches stand out, with Shapley Additive Explanations as a prominent choice due to its adaptability. The outputs of explainability methods are commonly presented visually, often in the form of graphs or user-friendly reports. Nevertheless, our review highlights a limitation in user evaluation and underscores the importance of involving users in the development process.

Conclusions: The integration of XAI into wearable health care technologies is crucial to address the issue of black box models. While wrist-worn wearables are widespread, there is a notable gap in making the data they generate explainable. Post hoc methods such as Shapley Additive Explanations have gained traction for their adaptability in explaining complex algorithms visually. However, user evaluation remains an area in which improvement is needed, and involving users in the development process can contribute to more transparent and reliable artificial intelligence models in health care applications. Further research in this area is essential to enhance the transparency and trustworthiness of artificial intelligence models used in wearable health care technology.

(J Med Internet Res 2024;26:e53863) doi:10.2196/53863

XSL•FO
RenderX

KEYWORDS

explainable artificial intelligence; XAI; wearable; machine learning; deep learning; health informatics; wearable sensors; user experience; wearable data; analytics; interpretation

Introduction

Background

Wearable technologies have become indispensable and dominant in the health care landscape [1]. A notable recent shift in their use involves leveraging their capabilities for the continuous monitoring of users, which proves particularly beneficial in patient care scenarios, such as monitoring patients with diabetes to preempt hypertension [2] or in the case of athletes, analyzing heart rate (HR) data to tailor personalized exercise regimes for enhanced progress [3]. However, to develop such sophisticated models, machine learning (ML) or deep learning (DL) algorithms are used, which often function as "black boxes," lacking transparency and making them challenging for medical professionals and end users to comprehend. In this context, explainability becomes crucial in ensuring the responsible and ethical use of wearable technologies in health care. By providing transparent insights into the inner workings of complex algorithms, explainable artificial intelligence (XAI) empowers medical professionals and end users to trust and confidently use these technologies for improved patient outcomes and personalized interventions [4]. This bridges the gap between ML experts and health care professionals, empowering the latter with actionable insights derived from these models.

Despite the growing interest and research on explainability in 2017, the association between artificial intelligence (AI) and explainability dates back to the mid-80s [5-7]. Over time, the significance of explainability has grown, recognizing its immense potential across various domains. In 2018, to further promote the importance of explainability, organizations such as the Association for Computing Machinery issued statements emphasizing algorithmic transparency and accountability [8], encouraging researchers and institutions to prioritize explainability when designing AI systems. International institutes such as the Defense Advanced Research Projects Agency have also contributed to the focus on explainability by funding the Explainable AI (XAI) Program [9]. This initiative emphasizes the importance of transparency and interpretability in AI systems, further highlighting the growing recognition of explainability's significance in the field of AI. As a result of this, there has been a notable increase in research on explainability, emphasizing its growing importance in promoting the ethical use of AI in various domains.

In parallel, the recent emphasis on XAI has sparked interest in integrating it with wearable technologies. Wearables have demonstrated significant potential and effectiveness in health monitoring, paving the way for innovative health care applications. What sets XAI apart when applied to wearable data compared to other datasets lies in the unique characteristics of wearables. Wearables capture highly personalized, granular data, often in dynamic real-world settings. This personal and real-time nature introduces complexities that demand a specialized XAI approach. The interpretability and transparency of AI models become even more critical as users must understand not just the decisions but also their impact on health and well-being. Moreover, wearables often integrate diverse data types, from physiological signals to activity tracking, requiring XAI techniques capable of handling multimodal data. Thus, delving into XAI within the domain of wearables is essential to address these distinct challenges and harness the full potential of wearable technology in health care, fitness, and personal well-being. However, the incorporation of XAI techniques into wearables remains an emerging research frontier. A recent review, which analyzed papers from 2011 to 2022, underscored an existing gap in the field: the lack of XAI research specifically focused on interpreting 1D biosignals obtained from wearable devices [10]. To address this gap, this paper aimed to review the recent literature, exploring the application of explainability in wearables. By examining how XAI can enhance the interpretability of generated data and models, this review sought to shed light on the possibilities that arise at the intersection of wearable technologies and XAI.

Related Studies

In the current AI era, a notable transformation is being witnessed in health care [1]. Various applications are powered by AI systems, leading to the emergence of ML and DL. As AI complexity increases, the demand for enhanced transparency is being recognized. This demand is met by the implementation of XAI, which allows AI model workings to be understood. The following paragraphs provide a more comprehensive exploration of the terminology associated with AI, ML, DL, and XAI. Following this, the paper proceeds to delve into specific research questions (RQs) that will be addressed.

AI involves the creation of systems and machines designed to replicate human intelligence, enabling them to perform real-world tasks effectively. AI systems are trained on data, allowing them to learn from experience and solve specific problems. They continuously refine their performance based on the information they receive. AI applications are diverse and include advanced web search engines, self-driving cars, gaming, speech recognition, recommendation systems, and health care AI systems and applications. AI essentially emulates human cognitive processes, making it invaluable when dealing with extensive datasets [1].

ML, a subset of AI, empowers computers to recognize patterns, make highly accurate predictions, and self-improve through experiential learning without the need for explicit programming. Building AI-driven applications relies heavily on ML methodologies. These models undergo training using extensive datasets, enabling them to deliver precise predictions [2]. ML is further categorized into supervised and unsupervised learning, with supervised learning branching into semisupervised learning [4] and reinforcement learning [5].

DL, a specialized field within ML, draws inspiration from the human brain's structure and functionality. DL effectively uses both structured and unstructured data for model training. It plays

XSL•FO
RenderX

a vital role in predicting life-threatening diseases in medical research, with deep neural networks achieving remarkable predictive capabilities [11]. Prominent DL models include convolutional neural networks (CNNs), residual networks, fully convolutional networks, and long short-term memory (LSTM) [12]. Figure 1 [13] depicts the relationship among AI, ML, DL, and XAI.

Figure 1. Relationship among artificial intelligence (AI), machine learning (ML), deep learning (DL), and explainable AI (XAI) [13].

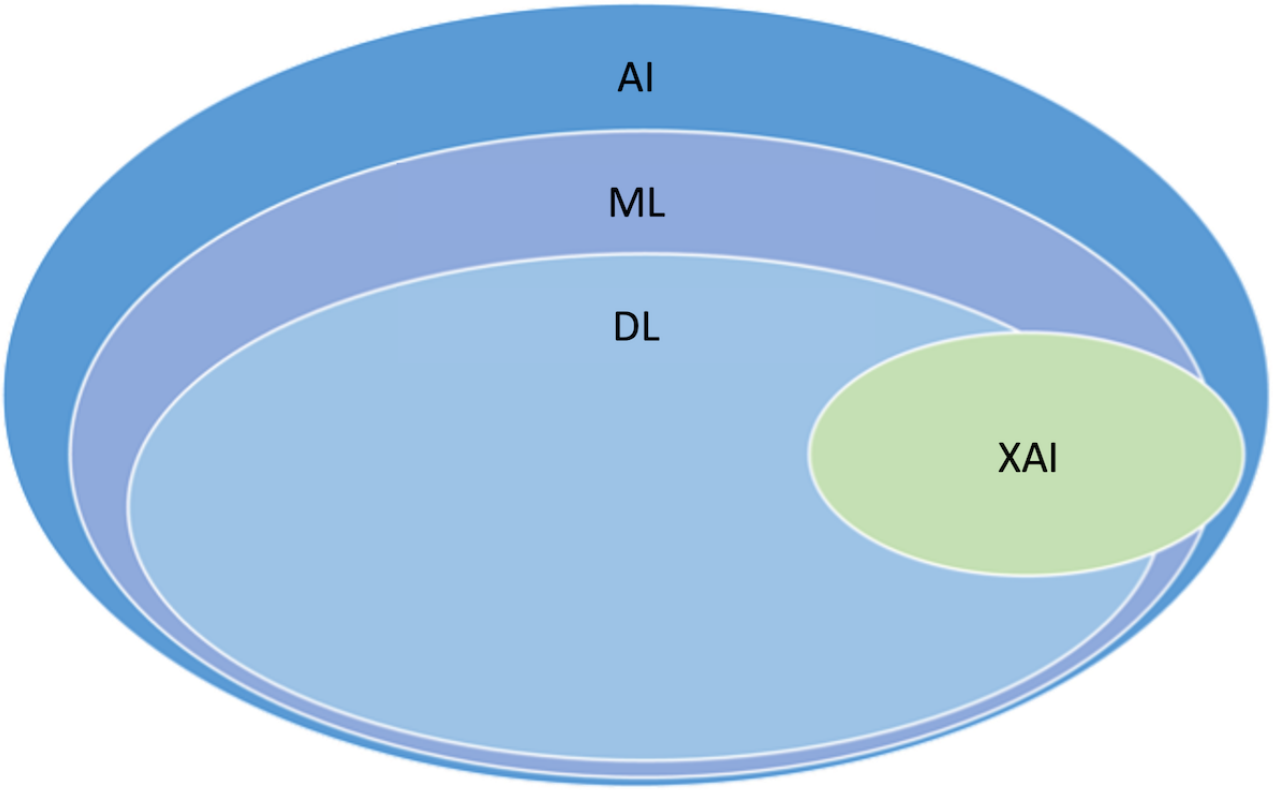

XAI enriches AI models with information comprehensible to the end users. While AI algorithms enable users to make informed business decisions, the opacity of these algorithms often leaves users uninformed about the decision-making processes [14]. This lack of transparency is where XAI comes into play. XAI strives to elucidate the inner workings of AI models, offering users comprehensible explanations of the methodologies, procedures, and outputs. The term *XAI* is often referred to as the "white box" approach due to its emphasis on revealing the model's processes.

In the field of XAI, training data serve as input, and users select the prediction methodology and XAI techniques based on specific application requirements. The input data vary depending on their source. For example, they can be electronic health records, vital sign recordings, medical scans, and wearable data. This review focuses on wearable data as the training data. Wearable data vary among physiological, activity, environmental, behavioral, biometric, and social interaction data. These techniques shed light on the model's internal operations and provide an explanatory interface, as shown in Figure 2, adapted from the study by Saranya and Subhashini [13]. In the modified diagram, the input data are changed to "wearable data" to be more specific to this review. This transparency empowers users with insights into AI model outputs, fostering trust [15]. Armed with this understanding, users can enhance output accuracy and identify model shortcomings, facilitating informed decision-making to improve the model.

Figure 2. Process of explainable artificial intelligence (AI) from wearable data, adapted from Saranya and Subhashini [13].

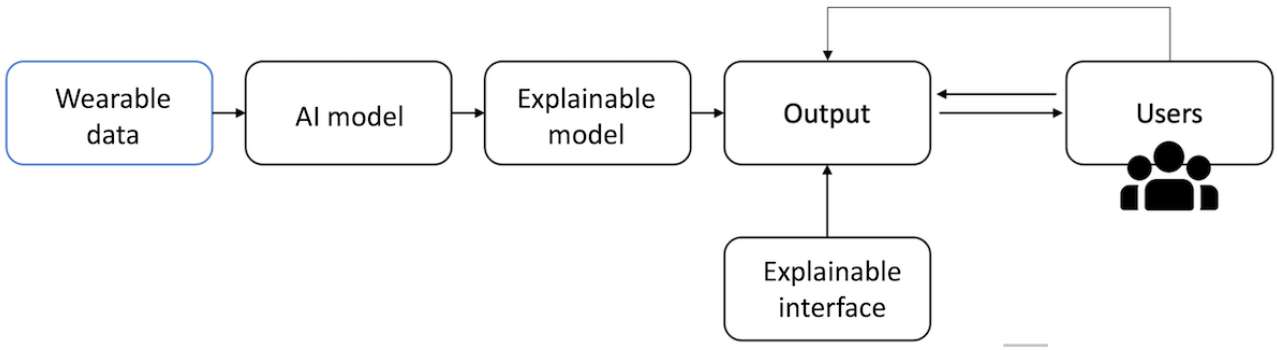

XSL•FO
RenderX

Despite the recent surge in wearable devices within the health care sector, there is a notable gap in research on XAI applied to wearable data. The application of XAI in wearable data analysis is crucial due to several key factors. First, the high granularity and personal nature of the data collected via wearables emphasize the need for transparency. Users must understand how their data are processed and interpreted to trust the insights generated. Second, given the inherently personal nature of wearables, establishing trust becomes paramount. Users need confidence in the accuracy and privacy of the data collected, making transparency in algorithms and data handling essential. Finally, wearables operate in dynamic environments and integrate diverse data types. XAI can help unravel complex relationships within these data streams, enabling better insights and decisions in various health care and well-being applications. Although recent reviews have explored the broader applications of XAI in health care, the potential of wearables in this context has remained somewhat overshadowed. Some of these reviews have touched on the relevance of wearable biosensors in health care applications, yet they often fail to fully uncover the diverse opportunities presented by both commercial and noninvasive smartwatches [16,17]. Similarly, certain reviews have focused on health care Internet of Things devices, but they tend to only scratch the surface regarding the intricate field of wearable technology [18,19]. The limited number of studies on XAI from wearable data highlights a significant research gap in this emerging field. While wearables have gained traction in health care and other domains for data collection and analysis, the incorporation of XAI principles to ensure the interpretability of AI models is an area that holds great potential but has seen limited exploration. This review aimed to explore the use of XAI within wearables (RQ 1).

In the domain of wearable technology, these devices offer a diverse array of data, encompassing activity metrics such as calorie expenditure and step counts as well as physiological signals such as electrodermal activity (EDA) and electrocardiography (ECG). Furthermore, wearables extend their capabilities to encompass behavioral patterns and environmental cues, thereby enriching the contextual understanding derived from the data they collect. Recent reviews in the health care field have illuminated the widespread adoption of wearables across various anatomical regions, including the head, limbs, and torso [20]. These versatile devices have been instrumental in monitoring a range of medical conditions, spanning stroke and poststroke rehabilitation to Parkinson disease, among others [21]. Hence, the adoption of wearables in health care is evident, yet some studies have delved deeper and adopted XAI for wearable data. For example, a study used EDA for pain recognition using the gradient-weighted class activation mapping technique [22], and another adopted accelerometer data for fall detection using local interpretable model-agnostic explanations (LIME) as the explainability method [23]. This review sought to address the question of which specific wearable data types have been explored within the context of explainability (RQ 2). By shedding light on this aspect, this review sought to bridge the gap and provide insights into the areas of wearable data that warrant greater attention from researchers and practitioners in the field of XAI. In addition, this review sheds light on the various explainability methods deployed specifically for each type of wearable data (RQ 3).

Previous reviews on XAI in the health care field have focused on explainability. They have highlighted the dominance of feature explanations over textual and example-based explanation methods. In addition, most explainability methods use the post hoc approach while focusing on the local rather than the global scope of the data. This review aimed to identify whether wearable data follow this trend from previous reviews (RQ 4).

The evaluation of AI models after applying an explainability method is crucial for ensuring transparency, accountability, and user understanding. Several researchers have stressed the need for formal evaluation metrics and a more systematic evaluation of the XAI methods [5,24]. Evaluation allows for a formal comparison of the available explanation methods [25] and offers a formal method to assess whether explainability is achieved in an application [25]. A previous review focused on the importance of assessing the explanation, and the results revealed that only 1 in 3 studies solely relied on anecdotal evidence for evaluation, whereas 1 in 5 studies incorporated user evaluations [26]. This highlights the gap in evaluating the XAI outcomes through either anecdotal evidence or user evaluations. To address this gap with wearable data, this review aimed to identify the evaluation methods used for the XAI techniques in the field of wearable data (RQ 5).

These aspects led to the following RQs, which guided our survey:

1. In the health care domain, how is XAI being used within the context of wearables? (RQ 1)
2. What is the predominant data type used in building XAI models from wearables? (RQ 2)
3. How is the explainability of AI models represented to the users? (RQ 3)
4. To what extent do model-agnostic, post hoc, and global explainability methods prevail compared to other approaches in line with existing literature? (RQ 4)
5. What are the evaluation methods used for various explainability techniques in the context of wearables and XAI in health care? (RQ 5)

A Typology of XAI Features

Overview

To guide the exploration of XAI within the domain of wearable technology, recent reviews on explainability have provided categorizations of XAI methods. These reviews have extensively examined various dimensions of XAI. One review focused on 6 dimensions, namely, the type of explanation, type of task, type of data, type of explainability method, type of problem, and type of model to be explained [26]. Another review focused on the application domain, model type, stage, scope, and output format [16]. We opted to build our survey on those previous XAI taxonomies [27] as this informs the different forms of explainability deployed in wearable technology, encompassing stage, scope, problem type, input data, and output format. This taxonomy is recent, covers relevant concepts, and has been highly cited by other researchers.

XSL•FO
RenderX

The following sections provide a brief definition of each explainability dimension.

Input Data of the XAI System

The term *input data* within the context of XAI systems refers to the data used to train the AI model, as shown in Figure 2 [13]. The nature of these input data can vary based on their source. In the context of this review, which focuses on wearable data, the primary source of input data is wearable devices. Wearable data encompass various forms, including physiological signals such as HR, ECG, and electroencephalography. In addition, wearables provide data in the form of activity metrics such as step and calorie counts. Moreover, wearable technology captures behavioral and emotional states, allowing users to log factors such as stress levels. Environmental data, including temperature and ambient noise levels, can also be collected via wearables. Furthermore, wearables can capture social data, such as monitoring the time spent on specific social media apps and related interactions.

Output Format of the XAI System

Output format of explainability refers to how the explanations are presented. This can vary from *visual* representations such as graphs and images to *textual*, *numerical*, *rule-based*, or *mixed* formats [27]. The presentation of explainability encompasses diverse approaches that depend on several factors, such as the target population and the nature of the input data. For instance, when the target population consists of lay users, explainability must be presented in a simple manner that aligns with their level of understanding. Conversely, when explainability is intended for medical professionals or researchers, a more detailed and in-depth approach is necessary as this population possesses a higher level of expertise and requires a comprehensive understanding of the underlying mechanisms and processes driving the model's outcomes. In addition, each type of input data may require specific methods of explanation. The nature and characteristics of the input data play a role in determining the most effective approach for conveying their insights and interpretations.

Stage of Explainability

Stage of explainability refers to the point during the XAI process (Figure 2 [13]) at which a method generates explanations. The stage when the explainability is introduced can be *post hoc* or *ante hoc*. Ante hoc methods aim to consider the explainability of a model from the beginning and during training to make it naturally explainable. In contrast, post hoc methods maintain pretrained models without any structural modifications and introduce explainability mechanisms after the model's training phase. The post hoc methods frequently use external explainer techniques during testing. An example of a post hoc method is Shapley Additive Explanations (SHAP) [28], which aims to provide a way to attribute the contribution of each feature to a model's prediction. Conversely, an example of an ante hoc method is recurrent lexicon networks [29], a method that models lexicons as naïve gated recurrent networks while seamlessly integrating explainability principles throughout the training process. This approach ensures that the model inherently possesses explainability characteristics from the beginning.

Scope of Explainability

The scope of explainability refers to the extent to which an explanation clarifies the inner workings of the AI model. The scope of explainability can be either local or global [30]. Local explainability focuses on clarifying the reasoning behind individual predictions, offering insights into why a specific decision was made for a particular instance. For instance, by using techniques such as LIME [31], a model may reveal that it diagnosed a rare medical condition for a patient based on a combination of relevant biomarkers and wearable data. Conversely, global explainability seeks to provide a broader perspective, offering a holistic understanding of the model's behavior and feature importance across an entire dataset. Using methods such as SHAP [28], one can analyze the model's tendencies in an entire patient population. Thus, local explainability delves into explaining individual predictions, whereas global explainability offers insights into the model's behavior and feature importance across an entire dataset.

Problem Types Addressed by AI Models

Each AI model that is constructed is tailored to address a specific underlying problem type, which could fall into either *regression* or *classification* categories. Studies that use regression aim to predict a continuous numerical value, whereas classification focuses on categorizing data into distinct classes or groups. Regression models are valuable for predicting patient outcomes and estimating essential health parameters. For instance, these models can be used to predict blood pressure levels in patients with diabetes based on factors such as food intake, exercise, and insulin dosage [32]. On the other hand, classification models are instrumental in disease diagnosis and treatment planning. For example, ML models can analyze physiological data to classify an individual's stress levels [33].

Methods

Study Design

We followed a systematic review design using qualitative methods. We adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [34].

Data Sources and Search Strategy

At the time of conducting our study, we captured the most recent publications in the rapidly evolving fields of XAI and wearable devices. Since the Defense Advanced Research Projects Agency introduced the XAI program in 2017, there has been an increase in papers focusing on XAI. Recent reviews have shown a significant rise in publications in the field of XAI during the period from 2018 to 2022, highlighting the growing research interest and developments [13] in the field, and this trend continues today. In this review, we collected publications covering 3 years, from January 1, 2018, to December 31, 2022, and then conducted the analysis and reporting process. We deem the qualitative outcomes of our review also representative of more recent publications in the field. We collected the publications from ACM Digital Library, IEEE Xplore, PubMed, SpringerLink, JMIR, Nature, and Scopus. The selected publishers are renowned for their high-quality and impactful

XSL•FO
RenderX

research across computer science, engineering, health care, and interdisciplinary studies. In addition, our preliminary search indicated the potential of these publishers to provide research on XAI and wearable data. The search strategy encompassed a comprehensive range of terms from various domains. These included explainability concepts, AI (Medical Subject Heading; MeSH) terminology, target population (MeSH) descriptors, wellness (MeSH) terms, and technology (MeSH) keywords. XAI, ML, and DL were considered alongside target populations such as patients and physicians. Wellness-related terms such as *physical activity*, *sleep*, and *exercise* were also incorporated. In addition, technology aspects encompassed wearable electronic devices and sensors. Multimedia Appendix 1 provides detailed keywords for each database.

Eligibility Criteria

The inclusion criteria for this review encompassed technology-based research involving wearable devices, sensors, and mobile phones. The studies were required to incorporate explainability or ML or DL techniques and use quantified self data. The quantified self domain, which began in 2007, uses technology such as apps and wearable smart devices to monitor, measure, and quantify different aspects of daily life [35]. In addition, the focus of the studies needed to be medical, and they had to be proceedings, book chapters, or peer-reviewed journals; written in English; and published between 2018 and 2022. On the other hand, the exclusion criteria included duplicate studies, review articles, and books, as well as workshop papers, courses, tutorials, and talks. Textbox 1 provides detailed information.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Technology-based research (eg, wearable devices, sensors, and mobile phones)
- Studies incorporating explainability
- Studies incorporating machine learning or deep learning techniques
- Studies using quantified self data
- Medical-based studies
- Studies published in peer-reviewed journals
- Proceedings or book chapters
- Studies written in English
- Studies published from 2018 to 2022

Exclusion criteria

- Duplicate studies and review articles
- Books
- Workshop papers
- Courses
- Tutorials
- Talks

Study Screening

Screening of potentially eligible studies was performed in 3 steps: duplicate removal, title and abstract screening, and full-text screening. Duplicates were removed using Rayyan (Rayyan Systems Inc) [36]. Additional duplicates that were not removed during this process were removed manually. Two review authors (YA and MA) independently screened the titles and abstracts for inclusion using the predefined inclusion and exclusion criteria specified previously. The other 2 review authors (DA-T and AB) independently screened a random 30% (n=197) of the included studies. Agreement among all authors was confirmed using a Cohen κ test. Table 1 provides more details.

XSL•FO
RenderX

Table 1. Cohen κ test of reviewer agreement.

Author initials	YA	DA-T	MA	AB
YA				
Cohen κ	—[a]	0.63	1	0.78
Percentage of agreement	—	94.17	100	96.58
DA- T				
Cohen κ	0.63	—	0.63	0.63
Percentage of agreement	94.17	—	94.17	94.17
MA				
Cohen κ	1	0.63	—	0.78
Percentage of agreement	100	94.17	—	96.57
AB				
Cohen κ	0.78	0.63	0.78	—
Percentage of agreement	96.58	94.17	96.57	—

[a]Not applicable.

All studies not discarded through this process were then screened by 1 review author (YA) in a full-text review process from which studies were identified for inclusion. Subsequently, a backward and forward referencing approach was used to further uncover potentially relevant papers. Conflicts among the 4 review authors were addressed through a majority vote system where a 3-versus-1 decision was made. Only 3 papers were found to have 2-versus-2 conflicts, which were subsequently resolved through discussions with each review author. Full data extraction, categorization, and labeling of papers were performed by 1 author (YA) and validated by 2 authors (MA and AB).

Feature Extraction

During the analysis, various features were extracted from different perspectives. *Metadata features* provided information about the publication and dissemination of the selected studies. *Explainability features* delved into the specific problem, the AI model used, and the input data used, following the taxonomy [27]. *Evaluation of explainability* included whether it was evaluated through user studies. For studies involving user evaluations, additional features such as the materials used, participant count, data collected, and outcome of the evaluation were extracted. *Technology features* focused on the type of wearable device used and its placement on the body. Table 2 provides detailed information. All features except the explainability features were created by the authors to fit the topic.

XSL•FO
RenderX

Table 2. Feature extraction.

Viewpoint and feature	Description
Metadata (PRISMA Flowchart and Selection Statistics and Geographic and Time Statistics sections)	
Country	Country of the article based on the affiliation of the first author
Domain	Targeted domain of the study, such as a specific disease or medical condition (eg, diabetes or stress management)
Publication date	When the article was published
Explainability [27] (Analysis Using the Typology of XAI Features section)	
Model	Type of ML[a] or DL[b] model that was used for performing the primary task
Problem type	Methods for explainability can vary according to the underlying problem (classification or regression)
Output format	Form of explanation generated for the model's outcome
Target group	The group targeted by the explainability generated
Input data	Type of quantified self data used, such as step count or calories
Stage	Refers to the stage at which a method generates explanations; can be either ante hoc or post hoc
Scope	Refers to the scope of an explanation; can be either global or local
Evaluation (Human-Centered Evaluation section)	
Evaluation method	Method of evaluation; can be with or without end users
Study design	The study design used for verifying the explainability
Study method	The method used, such as qualitative, quantitative, or mixed methods
Participants for building the model	Number of participants recruited for collecting data to build the model
Participants for testing the model	Number of participants recruited for collecting data to test the model
Type of participants	Participant types, such as healthy or disease or condition specific
Data collection methods	Methods used to collect the data for the user study
Duration of the study	Duration of the user study
Medium of interaction	Medium used to communicate the explainability
Study procedure	The intervention of the user study
Data analysis	Techniques used for analyzing the user study
Technology (Technologies for Capturing Quantified Self Data section)	
Type of technology	Different types of wearable technology used, such as Empatica E4 wristband or Fitbit wearables
Position of the wearable	Position of the wearable device, such as on the wrist or arm

[a]ML: machine learning.

[b]DL: deep learning.

Data Synthesis and Analysis

We used descriptive statistics to describe the metadata, explainability features, human-centered evaluation, and technology used.

Results

PRISMA Flowchart and Selection Statistics

The study selection sequence is outlined in a PRISMA flowchart, which is shown in Figure 3. Our search yielded 690 articles, of which 32 (4.6%) were identified as duplicates and removed. After screening abstracts, a further 609 (88.3%) were excluded, leaving 49 (7.1%) assessed for eligibility via full-text review. Among these, 29 (59.2%) were excluded. Forward and backward citation searching yielded an additional 6 included papers. A total of 25 studies met the inclusion criteria after completing the backward and forward citation check. Refer to Multimedia Appendix 2 for the PRISMA checklist.

XSL•FO
RenderX

Figure 3. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the systematic review. DL: deep learning; ML: machine learning.

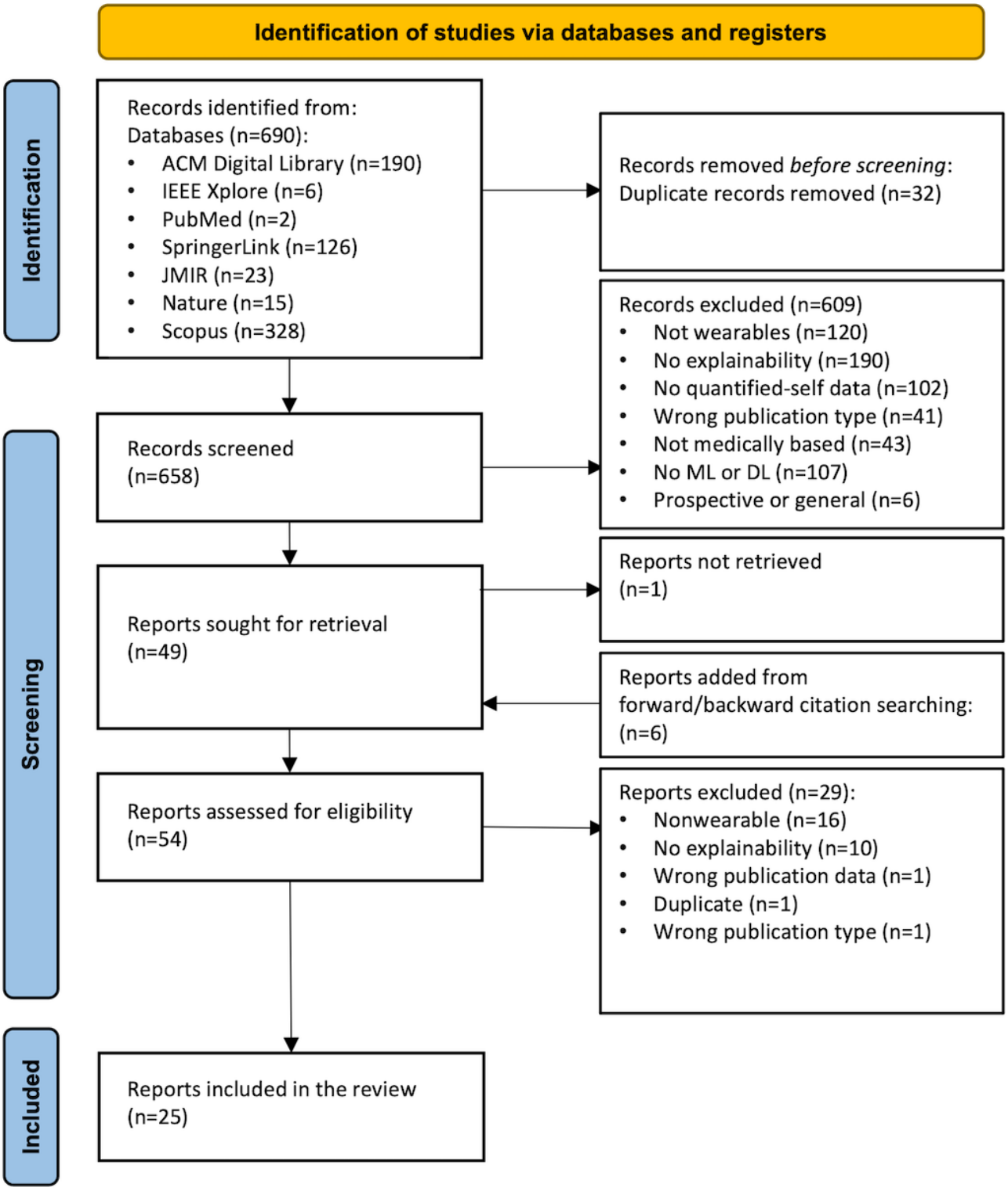

Geographic and Time Statistics

Table 3 illustrates the distribution of the 25 included papers, with 10 (40%) originating from the United States, 8 (32%) originating from various European countries, 1 (4%) originating from China, and 6 (24%) originating from South Korea. This widespread interest and research focus on explainability in wearable devices indicates its global significance and relevance in diverse regions. Furthermore, Figure 4 highlights that the selected papers were primarily published between 2019 and 2022, indicating a recent uptrend in research activity in this field. This recent attention highlights the growing recognition and importance of integrating explainability with wearable technologies for advancing health care applications.

XSL•FO
RenderX

Table 3. Distribution of studies per country (N=25).

Country	Publications, n (%)
United States	10 (40)
South Korea	6 (24)
The Netherlands	2 (8)
Switzerland	2 (8)
China	1 (4)
Italy	1 (4)
Norway	1 (4)
Turkey	1 (4)
United Kingdom	1 (4)

Figure 4. Distribution of included papers per year.

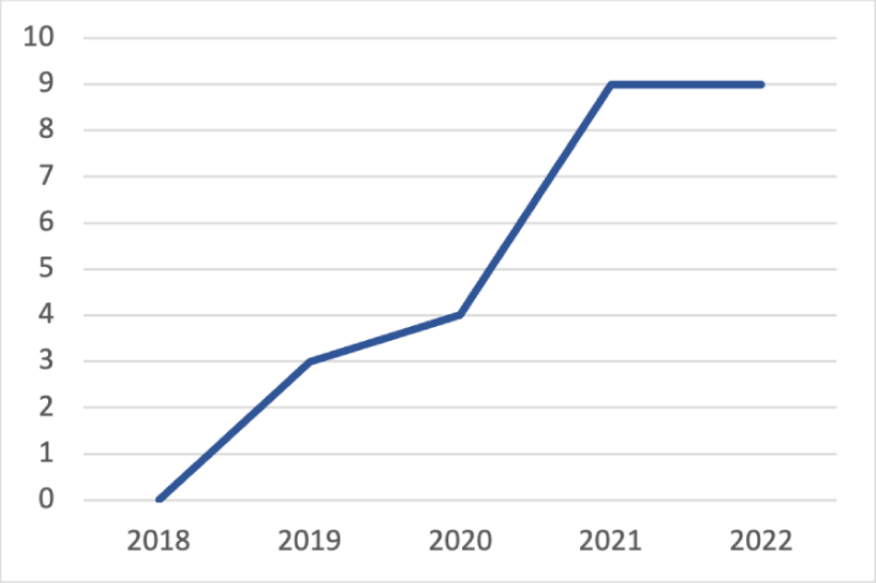

Application Domains

In this systematic review, certain domains received more exploration and attention in the context of explainability in health care monitoring compared to others, as shown in Table 4. When assessing the depth of investigation, certain domains stand out, with health conditions and diseases being the most extensively studied, accounting for 24% (6/25) of the research focus. Notably, these studies delved into conditions such as kidney disease, multiple sclerosis, influenza, sarcopenia, osteopenia, and the ongoing COVID-19 pandemic. Similarly, sleep and activity monitoring attracted some research interest, constituting 24% (6/25) of the reviewed studies. These investigations provided valuable insights into individuals' sleep patterns, personal health, and activity levels. Vital sign and health monitoring, encompassing parameters such as blood pressure and blood glucose levels, shared a comparable portion of the research landscape, also constituting 24% (6/25) of the total. This attention stems from the critical role of vital sign and health monitoring in the early detection and management of various health conditions, including diabetes. In comparison, domains such as mental health, stress management, and weight management collectively accounted for 16% (4/25) of the research focus. While these areas received some attention, they stood slightly behind the aforementioned domains. Finally, domains such as activity recognition, neurological monitoring (specifically brain signals), and substance abuse detection with a focus on opioid detection collectively constituted 12% (3/25) of the reviewed papers. These domains, while important, received relatively less exploration within the context of explainability in health care monitoring.

The explainability of wearable devices was applied in various health domains. It was widely used in the physical activity and health domain, such as predicting user-specific health risks [35] and identifying effective representations of fitness goals to enhance user physical activity and trust in the system. Wearable devices were also extensively used for diabetes control. For example, some studies (2/25, 8%) used the wearable device Empatica E4 and glucose monitoring to detect hypoglycemia [37] and hyperglycemia [2] with a lead time of up to 60 minutes. Another study focused on detecting eating moments and explaining glucose levels using wearables [32]. Similarly, other studies (2/25, 8%) focused on blood pressure monitoring and generated personalized lifestyle recommendations based on the user's blood pressure [38,39]. It is worth noting that there were

XSL•FO
RenderX

not many studies exploring XAI for weight management [40]. This is interesting because it could have practical applications for a large user base, including nonexperts. One possible reason for this gap in research might be that weight-related data are relatively easy to understand for most people and something that can be easily measured. In contrast, more complex data types received more attention in the field of XAI.

Table 4. The different wearable data applications using explainability.

Applications	Studies, n (%)
Vital sign and health monitoring	6 (24)
Sleep and activity monitoring	6 (24)
Health conditions or diseases	6 (24)
Mental health	3 (12)
Opioid abuse and detection	1 (4)
Weight management	1 (4)
Neurological and brain signals	1 (4)
Human activity recognition	1 (4)

Stress detection and management also received attention in wearable research. A study focused on stress detection and coping strategies [33], whereas another explored the prediction of next-day perceived stress using physiological signals such as ECG [41]. Comparative studies were also conducted to compare the effectiveness of different wearables in stress detection [42]. Furthermore, wearables were used in medical applications such as opioid detection [35] and identifying COVID-19 [43]. These studies highlight the diverse applications and potential of explainable wearable devices in health care.

Analysis Using the Typology of XAI Features

Overview

The following sections aim to provide concise descriptions of the primary categories of explainability methods identified in this systematic review (Multimedia Appendix 3 [2,3,33,38-53]). This is followed by a summarization of their stage, scope, problem type, input data, and output format. Figure 5 provides a summary of the 25 included articles, categorizing them into the five explainability features identified in the study.

XSL•FO
RenderX

Figure 5. Overview of the explainability features.

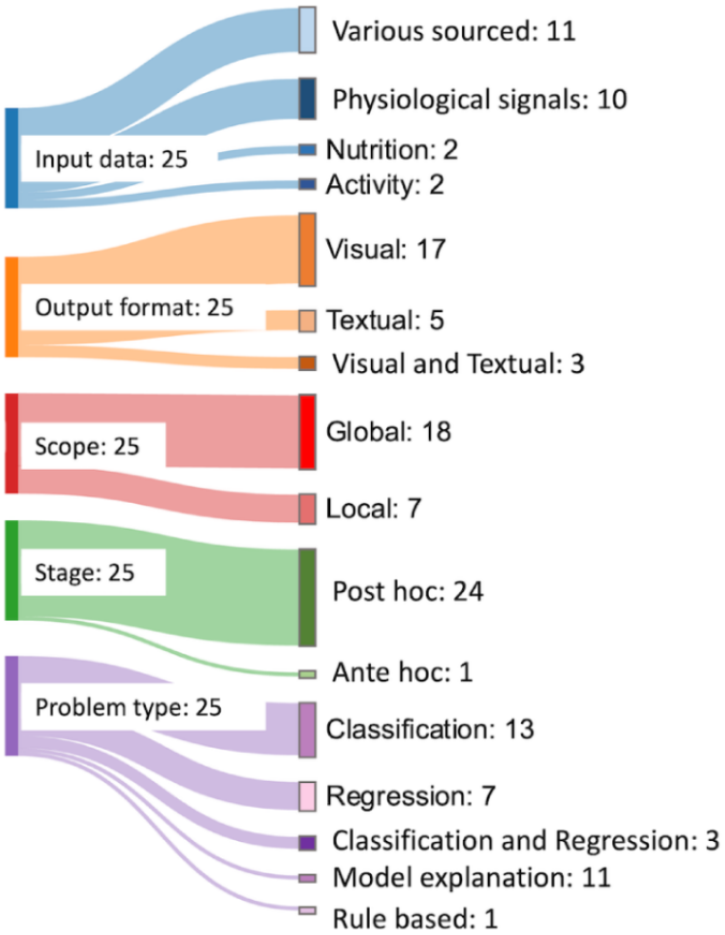

Input Data of the XAI System

This systematic review identified 5 primary categories of input data used in the research, including physiological signals (eg, HR), activity data (eg, steps), sleep data (eg, sleep duration and stages), nutritional data (eg, calorie intake), and mood (measured through surveys). Among these categories, physiological signals emerged as the most used input data for developing AI models, as shown in Figure 6 [44,54]. This prevalence can be attributed to the widespread use of wearable technology for collecting quantified self data. Commercial wearables such as Fitbit and Apple Watch excel in capturing physiological information such as HR, whereas medical wearables such as Empatica E4 offer advanced capabilities for gathering signals such as EDA, photoplethysmography, HR variability, and accelerometer data. Several studies in the reviewed literature (3/25, 12%) opted for a multisource approach, using various input data types to obtain a comprehensive understanding of the user's health, as shown in Figure 6 [44,54]. For instance, researchers incorporated a combination of physiological signals, activity data, and sleep metrics into their analyses [43].

XSL•FO
RenderX

Figure 6. UpSet diagram [45,46] of the various combinations of input data.

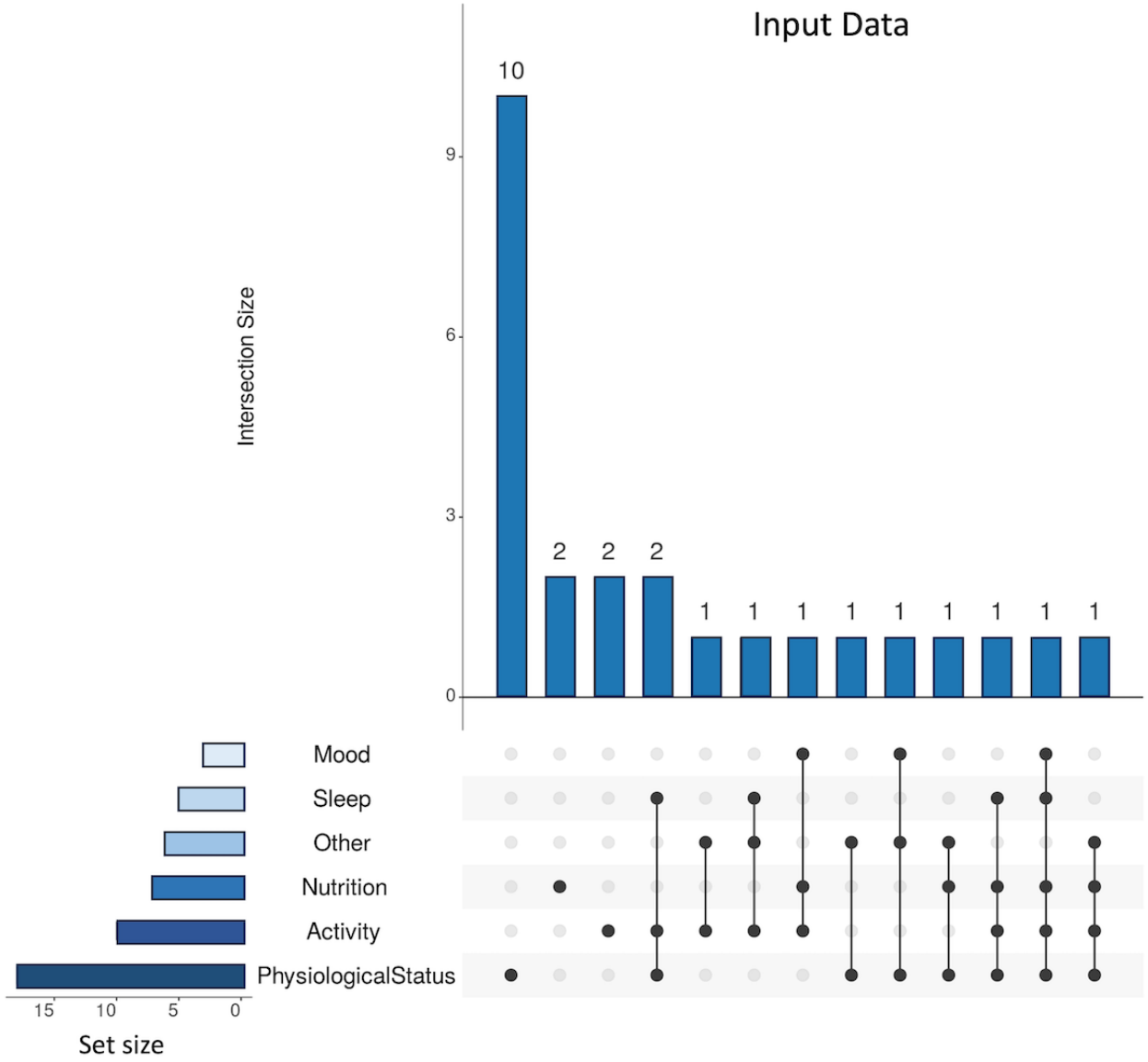

The input data in this review can also be classified into manual and automatic data collection methods, as shown in Textbox 2. Automatic data collection involves using technological devices such as wearables or mobile apps to gather data without direct user involvement, whereas manual data collection requires users to actively provide information or respond to specific queries. For instance, nutritional data are manually entered by users to estimate calorie intake [32], and mood data are collected through questionnaires to identify stress levels [33]. By considering both manual and automatic data collection, researchers can obtain a comprehensive and diverse set of input data, leading to a more holistic understanding [39]. Automatic data collection occurs with devices such as continuous glucose monitoring systems, which continuously measure and record glucose levels. Some studies in this review (3/25, 12%) used devices such as Freestyle Libre for blood glucose monitoring [2,32,37].

XSL•FO
RenderX

Textbox 2. Categorization of input data into automatically and manually captured.

Automatic (sensor signals)

- Physiological signals
- Activity
- Mobility
- App use
- Environmental data
- Sleep

Manual (logs)

- Weight
- Food intake
- Mood

Output Format of the XAI System

Overview

The following sections delve into the various explainability methods found in this review. Figure 7 summarizes the various output formats in this review. Among the reviewed studies, 64% (16/25) primarily used visualizations for explanations, whereas 16% (4/25) relied solely on text. A total of 12% (3/25) of the studies combined visual and textual explanations, and only 4% (1/25) used rule-based explanations, highlighting the diversity in interpretability approaches for health care–monitoring AI models.

Figure 7. Overview of the explainability formats used in the studies included in this review.

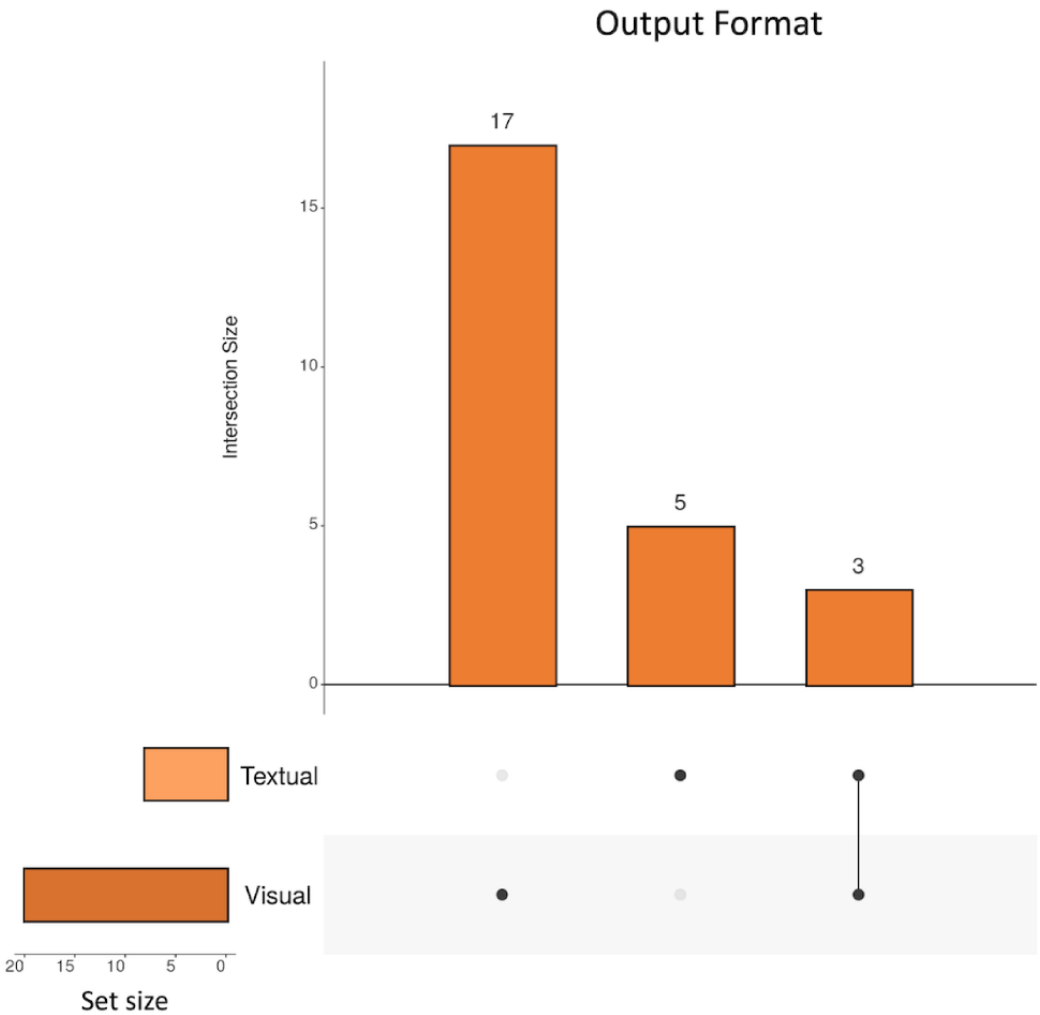

XSL•FO
RenderX

Visual Explanations

Visual explanations provide a natural and engaging means of conveying information and are highly effective in facilitating understanding. They are particularly useful in explaining complex concepts or processes as they can leverage the power of visual imagery to enhance comprehension. In the context of explainability, visual explanations can play a crucial role in explaining the "black-box" models in simpler means. Figure 7 shows the prevalence of visual explainability methods, highlighting their significance in the field. These methods use graphical tools and visual representations to provide insights into how a model operates and arrives at its predictions or decisions.

In some cases, additional visual aids such as graphs and scatter plots are used to generate visual explanations. For instance, a study focusing on COVID-19 detection used a gradient boosting prediction model based on decision trees. To illustrate the importance of different variables in the detection process, a bar plot was used, allowing researchers to visually identify the feature significance, as shown in Figure 8 [43]. Similarly, frequency bubble plots were used in certain studies to visualize glucose levels in healthy individuals, as shown in Figure 9 [32]. This visual representation assists in supporting individuals' self-management of their health by providing clear and easily interpretable information about their glucose levels [32]. Overall, visual explanations offer a powerful and accessible approach to communicating complex concepts, system behavior, and feature importance, enabling better understanding and informed decision-making.

Figure 8. Explainable gradient boosting of COVID-19 symptoms using bar charts [43].

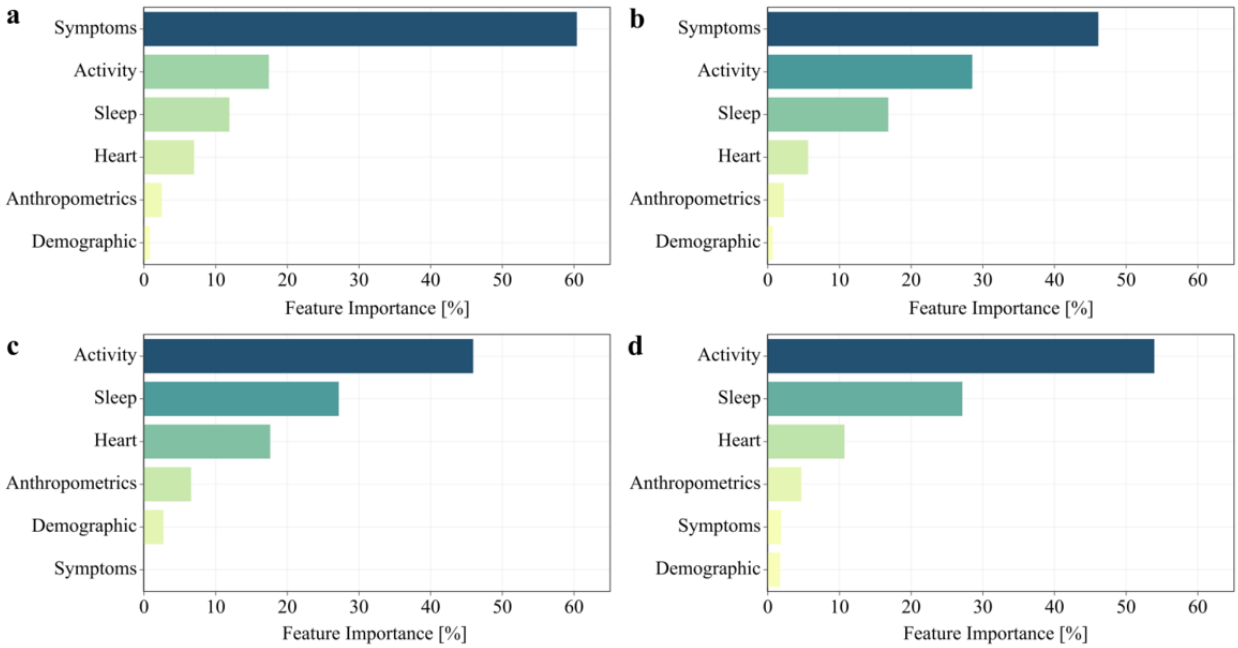

XSL•FO
RenderX

Figure 9. Shapley Additive Explanations explainability using a bubble frequency plot [32].

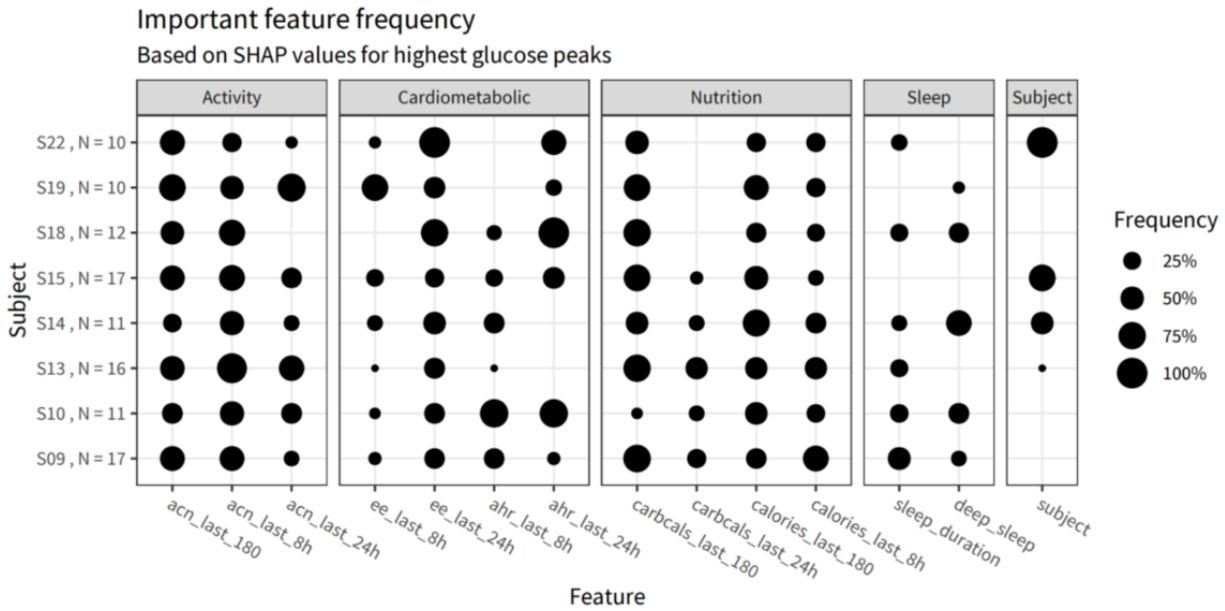

Textual Explanations

Textual explanations are another intuitive and widely used method for providing explanations. They are presented as natural language statements, whether written or spoken. Figure 10 [55] shown the significant use of text in the explainability of AI models. In various studies (3/25, 12%) [39,55-57], textual explanations enhanced transparency and trust in different domains.

Figure 10. Fitness tracker goals displayed as plain text [47].

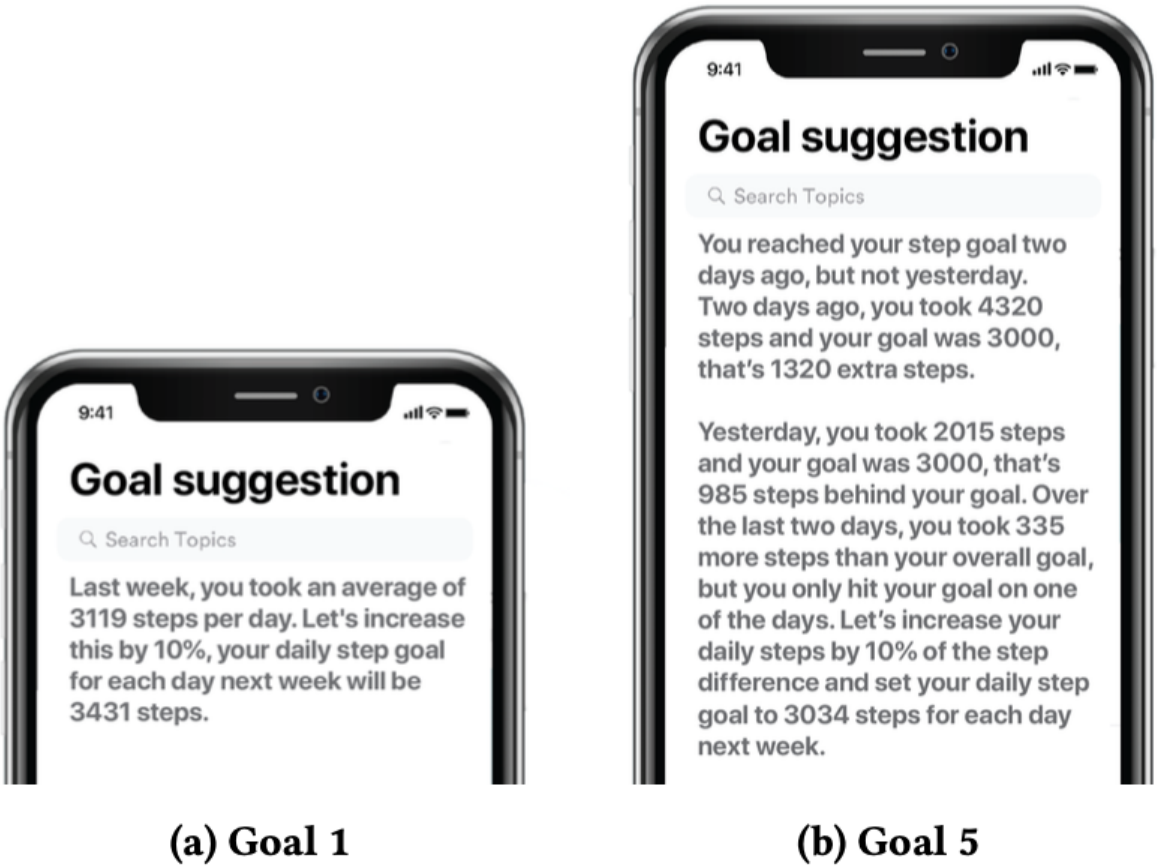

(a) Goal 1 (b) Goal 5

XSL•FO
RenderX

For example, one study compared 2 textual formats for suggesting fitness tracker goals, focusing on improving the transparency of step goal computation, thereby fostering user–AI model understanding and trust [55]. Furthermore, textual explanations were used to generate comprehensive summaries of personal health data, enhancing their explainability [56]. These summaries involve generating natural language descriptions of temporal trends and patterns from time-series health data. Such summaries help users evaluate their health data and compare them to their goals or general health guidelines. In addition, pattern-based summaries help identify hidden behaviors and provide explanations to the user. A linguistic summarization approach leveraging time-series protoforms was used to generate these summaries [56]. Protoforms act as sentence templates with placeholders automatically selected to reflect trends and patterns supported by the data.

Multimodal Explanations

In explainability research, tailoring the presentation to end users is crucial for effective comprehension and engagement. One approach involves leveraging familiar and user-friendly graphics to convey information in an accessible manner. Figure 11A [37] shows the use of activity rings from the Apple Watch, which are already well known to users [39]. In the context of hypoglycemia detection for patients with diabetes, the nearly closed violet ring in the visual representation serves as a warning sign for low blood glucose levels. This intuitive approach incorporates the individual's physiological state as a significant factor in identifying hypoglycemia.

Figure 11. (A) Stress-monitoring app [37] using a ring chart; (B) detecting hypoglycemia using wearables [33] displayed as text and emoticons; (C) blood pressure monitoring and lifestyle recommendations given as plain text and bar and line charts [38].

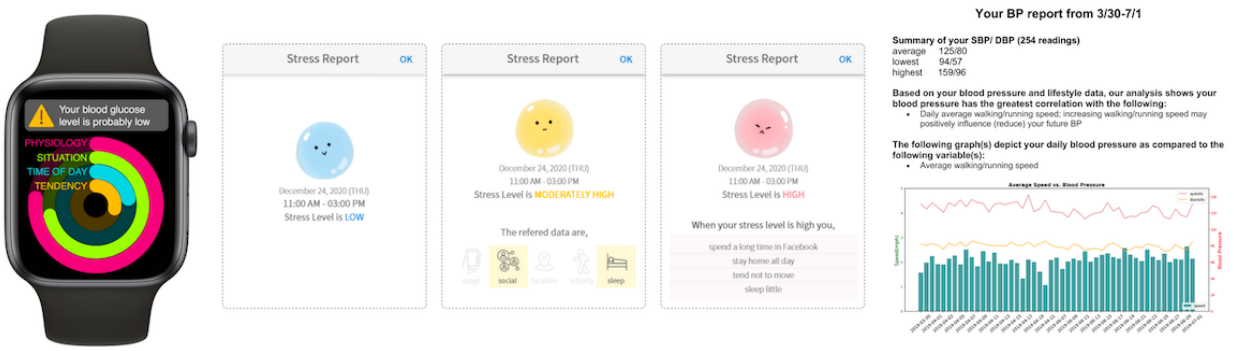

Similarly, when developing an app for stress prediction, it is essential to ensure that end users can easily interpret the results of the AI model. Visual representations prove valuable in illustrating the level of stress experienced by the user. Figure 11B [33] showcases how different colors and graphical elements can be used to enhance the explanation of the model's outcomes and ultimately support the user in managing their stress levels effectively [33]. Another notable application of multimodal explainability lies in providing personalized recommendations to end users. Researchers introduced a multimodal explainability approach in a study investigating the relationship between blood pressure and lifestyle factors [38]. This approach generates a report summarizing the relevant blood pressure features and offers tailored lifestyle modifications to improve blood pressure. An example of such a report can be observed in Figure 11C [38], which combines textual and graphical elements to present comprehensive and actionable information to the user. By using user-friendly graphics, visual representations, and multimodal explanations, researchers aim to optimize end-user understanding, engagement, and decision-making processes. These approaches leverage familiar and intuitive formats to convey complex information effectively, empowering users to monitor their health, manage stress, and make informed lifestyle choices.

Stage of Explainability

In this systematic review, the application of different explainability methods in health care monitoring varied, with some methods being more extensively explored than others. A total of 88% (22/25) of the reviewed studies used post hoc methods, whereas only 12% (3/25) used ante hoc methods. Model-agnostic methods were used in 60% (15/25) of the reviewed studies, and the rest were distributed between specific-domain (3/25, 12%), attention-based (3/25, 12%), gradient-based (3/25, 12%), and general explainability (1/25, 4%) methods, as seen in Table 5. Among the model-agnostic methods, SHAP was the most widely applied method, appearing in 48% (12/25) of the studies. SHAP aims to provide interpretable explanations for individual predictions made by complex ML models [28]. This also highlights its prevalence as a post hoc method in explainability research. In contrast, only 4% (1/25) of the studies explored an ante hoc explainability approach, specifically the explainable gradient boosting method. This method involves modifying or enhancing the original Extreme Gradient Boosting algorithm to enhance interpretability, indicating a limited use of ante hoc techniques in the reviewed studies.

XSL•FO
RenderX

Table 5. Distribution of explainability methods.

	Studies, n (%)
Model-agnostic methods	15 (60)
Specific-domain methods	3 (12)
Attention-based methods	3 (12)
Gradient-based methods	3 (12)
General model explanation methods	1 (4)

[a]Not applicable.

Gradient-based methods garnered attention in 12% (3/25) of the studies, encompassing explainable gradient boosting [43], deep learning important features [58], and layer-wise relevance propagation [45], with each technique being featured once. Explainable gradient boosting represents an advanced iteration of the gradient boosting algorithm, integrating mechanisms to render transparent insights into its decision-making process and feature significance [59]. Deep learning important features, on the other hand, serves as a technique that dissects a neural network's output prediction for a specific input, unraveling the contributions of all network neurons to each feature of the input [60]. Layer-wise relevance propagation, functioning as a framework, facilitates the deconstruction of deep neural network predictions on a sample into relevance scores [61]. These methods collectively serve the purpose of unveiling the significance of individual features and comprehending the intricate decision-making of complex models. While their numbers are limited, these techniques remain instrumental in providing invaluable insights into model behavior.

Attention-based methods were also studied in 12% (3/25) of the studies, involving interpretable LSTM-attention [46], CAM [47], and interpretable recurrent neural network [40]. In the case of interpretable LSTM-attention, an attention model is used to assign varying weights to input features of financial time series at each time step. This attention feature is then used to effectively select relevant feature sequences for input into the LSTM neural network, aiding prediction in subsequent time frames [62]. Meanwhile, CAM is an explainability technique used to identify significant regions within an input image that contribute to specific class predictions within CNNs [63]. Furthermore, the concept of an interpretable RNN emerges as a variant of the RNN architecture aimed at providing transparent insights into its decision-making process and internal representations, thereby enhancing its interpretability [64]. These methods excel in capturing salient regions and identifying sequential dependencies in health care data. Their use is driven by the need to understand how the model focuses on specific areas or patterns when making predictions.

Specific-domain methods such as GNNExplainer for graph neural networks and temporal summaries for time-series data were investigated in 12% (3/25) of the studies [47,48,56]. These methods cater to the unique characteristics of specific domains, allowing for domain-specific insights. Temporal summaries, in particular, were applied twice [56,57], emphasizing the significance of time-series data in health care monitoring. Finally, general model explanation methods were explored in a single study [65]. This broader category encompasses various techniques, but its limited application in this review suggests that researchers focused more on domain-specific or model-specific approaches in health care monitoring. The variation in the number of studies that applied each method indicate the varying levels of interest and emphasis placed on different explainability techniques. Researchers may choose certain methods based on their effectiveness in providing understandable explanations, compatibility with the data type, or the specific requirements of the health care monitoring domain being studied.

Scope of Explainability

This review encompassed studies using 15 local explainability methods and 10 global explainability methods applied in the context of health care monitoring. The inclusion of a larger number of local explainability methods, accounting for 60% (15/25) of the studies, indicates the significance of understanding individual predictions and the specific factors influencing them. These methods provide granular insights into model behavior and help build trust by explaining the rationale behind individual decisions. However, the presence of global explainability methods is also noteworthy as they offer a broader perspective on model behavior, identifying general patterns and highlighting features that consistently contribute to predictions across the entire dataset. The balance between local and global explainability methods in this review demonstrates the importance of both instance-level interpretability and a comprehensive understanding of model behavior in the context of health care monitoring. Researchers recognize the need for a multifaceted approach to ensure transparency, reliability, and generalizability in interpreting the outcomes of health care models.

Problem Types Addressed by AI Models

Overall, classification emerged as the most prevalent underlying problem in the reviewed studies, accounting for 52% (13/25), followed by regression at 28% (7/25). In addition, 4% (1/25) of the studies used a rule-based approach [56], whereas another (1/25, 4%) used a model explanation [65] as the underlying problem type. It is noteworthy that the combined total of regression and classification studies exceeded the total number of reviewed papers (N=25). This disparity arises from certain studies using models for both classification and regression tasks (3/25, 12%). The predominance of classification problems (16/25, 64%) in the reviewed studies indicates the significance of accurately categorizing health care data for diagnostic, predictive, or decision-making purposes. Gradient boosting emerged as the most frequently used algorithm for classification

XSL•FO
RenderX

(5/25, 20%), highlighting its effectiveness in achieving high predictive performance and capturing complex relationships within health care datasets. Support vector machine (SVM) was also commonly used (3/25, 12%), known for its ability to handle both linear and nonlinear classification problems [66]. SVM generally exhibits improved performance when applied to smaller datasets [67]. In addition, decision trees, Neural Structured Learning, graph neural networks, and CNNs were selected for specific classification tasks (5/25, 20%), reflecting the diversity of approaches used to tackle different health care monitoring challenges.

In regression problems (13/25, 52%), the reviewed studies used various algorithms to predict continuous or numerical outcomes relevant to–health care. Gradient boosting, known for its powerful ensemble learning capabilities, was the most prevalent algorithm (3/25, 12%), indicating its effectiveness in capturing nonlinear relationships and providing accurate regression predictions. Random forest (5/25, 20%) and SVM (1/25, 4%) were also used for regression tasks, leveraging their ability to handle complex datasets and capture intricate relationships between input features and target variables. Furthermore, CNNs, fully convolutional networks, ExtraTree, and RNNs were each used in specific regression studies (4/25, 16%), showcasing their suitability for capturing temporal or spatial patterns and making accurate predictions in health care monitoring contexts. It is worth noting that decision trees are generally considered more interpretable or transparent compared to some other ML models. The distribution of problem types observed in this review reflects the complexity and diversity of health care monitoring tasks (Figure 12). By leveraging these algorithms effectively, researchers can develop robust and accurate health care–monitoring models that cater to different types of problems.

Figure 12. (A) Overview of problem types in the reviewed papers; (B) regression and classification problem types with the corresponding machine learning or deep learning model application. CNN: convolutional neural network; FCN: fully convolutional network; GNN: graph neural network; LSTM: long short-term memory; NSL: neural structured learning; RNN: recurrent neural network; SVM: support vector machine; TCN: temporal convolutional network.

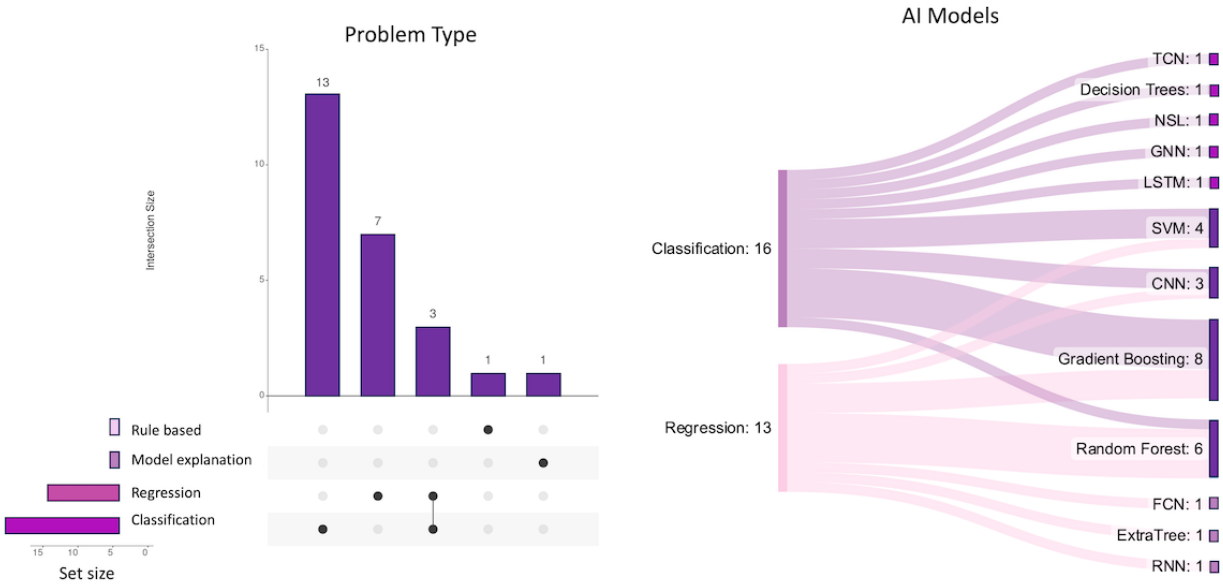

Training Dataset

In the development of AI models, they can be constructed using either user datasets, which are derived from real-world user interactions and activities and capture user behaviors and characteristics, or benchmark datasets, which are widely recognized standards for comparison and evaluation within the field [68]. Irrespective of the dataset used to train the AI model, some explainability models were evaluated with end users, whereas others were not. As the primary goal of explainability is to enhance user interaction with AI models, real-world evaluation involving lay users is essential.

Table 6 provides an overview of the datasets used in training AI models for health care monitoring, where 80% (20/25) of the studies used user datasets in comparison to 20% (5/25) of the studies, which used a benchmark dataset. Researchers adopted various strategies when selecting these datasets. Some chose to create custom datasets gathered by having users wear devices such as the Empatica E4 wristband to capture physiological signals while performing specific tasks. Conversely, others used benchmark datasets, exemplified by the MyFitnessPal food log dataset, comprising 587,187 days of food log data across 9000 users for 180 days [69]. The decision to use benchmark datasets often stems from resource constraints, such as limited time or access to participants. Conversely, the creation of custom datasets was preferred when specific testing conditions such as unique population groups or environments were not covered by existing datasets. Of the 25 studies reviewed, 20 (80%) opted to construct models using a dataset created by the authors, whereas 5 (20%) chose established benchmark datasets, such as the Floodlight Proof-of-Concept dataset used in the study by Creagh et al [45], the MyFitnessPal dataset used in the study by Harris et al [56], the MHEALTH dataset used in the study by Uddin and Soylu [49], the Continuous Glucose Monitoring Intervention in Teens and

XSL•FO
RenderX

Young Adults With Type 1 Diabetes study dataset [70], and the Human Activity Recognition database [71].

An analysis of the experimental setups for training datasets revealed distinct approaches, as shown in Table 6. It is noteworthy that the evaluation environment, whether conducted in the real world or in a controlled setting, plays a role in shaping the evaluation process. In this review, data collection occurred in real-world settings in 80% (20/25) of the studies, reflecting the natural environmental conditions in which users typically interact with the technology, such as registering the food they consume in a day [56]. Conversely, 20% (5/25) of the studies opted for a controlled environment for their data collection. This controlled setup involves specific conditions designed to eliminate external influences and noise. For instance, a study collected gait signals using sensors embedded in shoe insoles along a 27-meter straight corridor, ensuring a controlled and consistent testing environment [50]. In addition, physiological signals such as photoplethysmography and EDA were gathered using wearable wristbands such as Empatica E4 or Samsung Gear within soundproof rooms, eliminating auditory and visual distractions [42]. In another instance, during a private study, EDA and HR signals were collected using Empatica E4 devices within the controlled environment of a hospital setting [58]. After training the model, some studies went a step further and tested the explainability model with lay users. Of the 25 studies, only 5 (20%) tested the explainability output with potential end users. The Human-Centered Evaluation section explores the different aspects of the human-centered evaluation methods.

Table 6. Experimental setup of the user and benchmark datasets.

Training dataset	Experimental setup, n (%)		Total datasets, n (%)
	In the wild	Controlled	
User dataset	>16 (64)	4 (16)	20 (80)
Benchmark dataset	5 (20)	0 (0)	5 (20)
Total experimental setup	20 (80)	5 (20)	__[a]

[a]Not applicable.

Human-Centered Evaluation

Overview

The following sections focus on the 20% (5/25) of the studies (Table 7) included in this review that evaluated the user perception of explainability. They examined key aspects, such as the studied population, materials and methods, study duration, design, medium of interaction, and type of interaction. Table 7 summarizes the results of the 20% (5/25) of user studies.

Table 7. Human-centered evaluation of the explainability model.

Study	Studied population	Materials	Study duration	Study design	Method	Medium of interaction	Type of interaction
Kim et al [33]	Healthy participants	Interviews, questionnaire, and use logs	30-min introductory session, 25-day MindScope use, and 50-min follow-up interview	Pretest-posttest study	Mixed methods	Apps	Interactive
Wozniak et al [55]	Healthy participants	Survey	3 min 35 s to 5 min 43 s	Between subject	Mixed methods	Apps	Passive
Leitner et al [39]	Patients with prehypertension or hypertension	Survey and data logs	6 mo	Between subject	Quantitative	SMS text messages	Passive
Harris et al [56]	Healthy participants	Survey	Not mentioned	Within subject	Quantitative	No instruments	Passive
Chiang et al [38]	Patients with hypertension	Data log	1 mo	Between-subject and pretest-posttest study	Quantitative	Emails	Interactive

Studied Population

In all the studies (5/5, 100%), the primary focus was on providing explainability to end users who were not medical experts, researchers, or AI experts. The emphasis was on ensuring that the explainability methods used were simple and easily understandable for the target audience. Of these 5 studies, 3 (60%) specifically targeted healthy individuals [33,55,56]. The focus of these studies was on monitoring and addressing aspects such as stress levels, promoting trust in AI-generated step goals, and providing fitness summaries. The goal was to provide simple and interpretable explanations to help individuals monitor and improve their overall well-being. The remaining studies focused on patients with stage-1 hypertension [38,39], which is when the blood pressure of an individual ranges from

XSL•FO
RenderX

130 to 139 mm Hg systolic or 80 to 89 mm Hg diastolic most of the time. In these studies, the primary objective was to provide explainability for lifestyle recommendations aimed at managing and improving blood pressure. The goal was to empower patients with understandable explanations regarding the recommended lifestyle changes, facilitating their active participation in their health management. While many explainability papers center on creating models for clinicians, such as opioid detection [35] and multiple sclerosis ambulatory characteristics [45], none of the reviewed studies specifically addressed this target audience. This observation highlights a gap in the current research landscape as there is potential for further exploration and development of explainability models tailored to the needs and understanding of clinicians and other health care professionals.

Instruments for Delivering Explainability

In the reviewed studies, explainability was presented and communicated to the users through various instruments and platforms. Most of the studies (7/25, 28%) opted for a mobile app to deliver the explainability components to the users [33,39,55]. Only 20% (1/5) of the studies used email as the medium for presenting the explainability [38]. This suggests that alternative communication channels can be used to deliver explainability, such as computers or wearable wristbands, depending on the specific study context and target audience. By using different instruments for delivering explainability, the researchers demonstrated their flexibility in tailoring the presentation of information to suit the needs and preferences of users.

Type of Data Collected

In the user studies, various types of data were collected to gather insights and perspectives from the participants. These methods included interviews, questionnaires, and use logs, which provided qualitative and quantitative information about user experiences, preferences, and use patterns [33]. In addition, surveys were used to gather structured feedback and opinions from the participants [55,56]. Examples of surveys include the Perceived Stress Scale [72] used in the study by Kim et al [33], the Goal Commitment Scale [73], and the Trust Scale [74] used in the study by Wozniak et al [55]. In one study, a combination of surveys and data logs was used to capture both subjective responses and objective use data [39]. Certain studies (1/5, 20%) also used data logs to record and analyze user interactions and behavior [38]. It is worth noting that this diverse range of data collection methods reflects the adaptability of researchers in selecting appropriate tools for evaluating explainability models, with no discernible pattern emerging from the reviewed studies in terms of preferred data types.

Duration of the Study

The data collection duration in the user studies varied. In one study, the time for data collection per participant ranged from 3 minutes and 35 seconds to 5 minutes and 43 seconds, indicating brief interactions or tasks [55]. Another study involved a 30-minute introductory session followed by a 25-day MindScope use study and concluded with a 40-minute follow-up interview, allowing for a more extensive examination of participant experiences [33]. In addition, data collection occurred over 1 month in a specific study [38], providing a medium-term perspective on user engagement. One study spanned a longer duration of 6 months [39], enabling researchers to comprehensively assess user experiences and outcomes over an extended period. These variations in data collection duration highlight the diverse approaches used in user studies and provide insights into the range of participant engagement and the depth of data collection achieved in each study.

Study Design

The user studies incorporated various study designs to evaluate the effectiveness of the explainability models. These designs included pretest-posttest studies where data were collected before and after participant interactions with the system to assess any changes or improvements [33]. Between-subject and between-subject vignette designs involved dividing participants into different groups to compare the impact of different conditions or scenarios [39,55]. Within-subject designs allowed participants to experience multiple conditions or treatments, enabling comparisons within individual contexts [56]. In addition, between-subject designs and pretest-posttest studies were combined in some studies (1/25, 20%) to evaluate the system's impact across different groups comprehensively [38]. The review should reflect a clear preference or consistent pattern in the selection of a specific study design. However, the choice of study design seems to be driven by the researchers' individual preferences and the specific objectives of their investigations.

Challenges Reported

Researchers reported several challenges while conducting user studies. One notable issue pertained to the experimental design, wherein participants had limited interaction with each explanation type, restricted to 5-day intervals [33]. This abbreviated time frame may not have been sufficient to adequately assess the varying impacts and distinctions among different explanation methods. To address this limitation, a longitudinal study could be considered to offer a more comprehensive understanding. Another challenge centered on the need for a control group [33]. The inclusion of a control group would enable a clearer differentiation of the effects of prediction and explanation on stress reduction and management outcomes. Moreover, the task of counterbalancing the diverse explanations also posed a challenge [33]. Ensuring that each participant receives explanations in a balanced manner could enhance the validity of the findings. Furthermore, some studies (2/5, 40%) had a restricted participant demographic, such as solely involving college students [33] or relying on a single platform such as Mechanical Turk [75] for data collection [55]. In addition, some studies (2/5, 40%) had to contend with a limited number of participants [38,39]. These challenges collectively underline the importance of carefully addressing methodological constraints to ensure the robustness and generalizability of the study's conclusions.

Technologies for Capturing Quantified Self Data

Among the wearable devices reviewed, the wrist was the most common placement location (13/25, 52%), as illustrated in Figure 13. Some studies (4/25, 16%) used commercially

XSL•FO
RenderX

available devices such as Fitbit and Galaxy Watch [43,46,48,55], which gather data such as step count, distance traveled, calories burned, weight, HR, sleep stages, active minutes, and even location. On the other hand, research-based wristbands such as the Empatica E4 focus on capturing data such as EDA, blood volume pulse, HR, and interbeat interval [35,37,58]. In addition, wearables were predominantly utilized in the studies (16/25, 64%), while mobile applications were used in a smaller proportion (7/25, 28%), as shown in Figure 14.

Figure 13. Sources of the data collected.

Figure 14. Data collection medium.

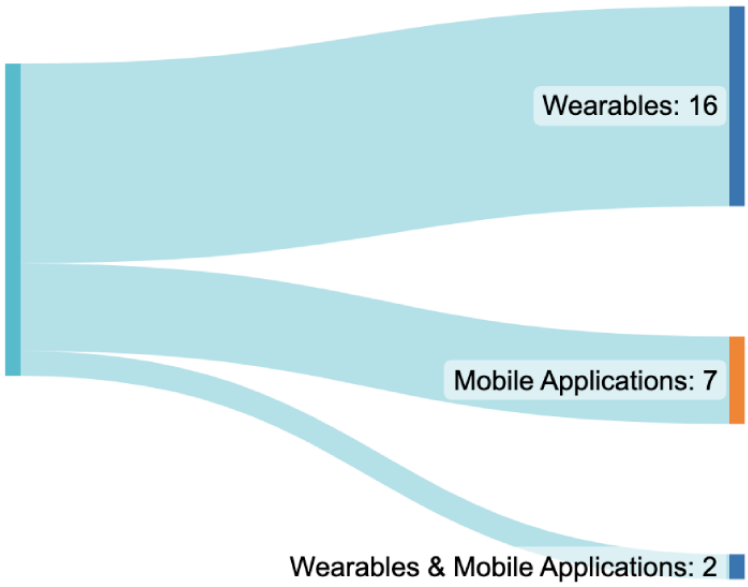

In addition, other wearables were identified, specifically arm-based devices (6/25, 24%) such as glucose monitoring devices (eg, Freestyle Libre and BioStampRC) [2,41]. These devices serve the purpose of monitoring glucose levels. Furthermore, arm-based devices were also found to collect blood pressure data, with the Omron EVOLV wireless blood pressure monitor being one notable example [39]. In a couple of studies (2/25, 8%), the researchers explored the use of sensors embedded within the shoe insole to capture quantified self data [50,51]. This approach allowed for the collection of data related to foot movement and pressure distribution. In several other studies (7/25, 28%), the focus shifted toward using mobile apps for gathering quantified self data [40,47,52,56,57]. These apps often leveraged GPS data and logs to obtain information about user activities and movement patterns. Overall, this systematic review revealed a wide range of technologies used to capture quantified self data, with wristbands, arm-based devices, shoe insole sensors, and mobile apps playing prominent roles in the collection process.

XSL•FO
RenderX

Discussion

Principal Findings

Overview

The emergence of wearable technologies has revolutionized the field of health care, enabling the collection of vast amounts of quantified self data and opening new possibilities for AI models. However, the inherent "black box" nature of AI models has hindered their seamless integration into medical practices and public acceptance. This emphasizes the critical role of explainability in bridging the gap between health care professionals and AI experts. In this review, we delved into the adoption of explainable models in health care, with a specific focus on wearable technologies and quantified self data. The following sections provide a comprehensive exploration of the 3 main themes: wearables, explainability, and human-centered evaluation.

Wearables and Data

To address RQs 1 and 2, the widespread application of wearables in the health care field has been extensively explored in various reviews [20,76]. These wearable applications encompass a diverse range, from assessing wearables' usability [77] to using wearable cameras for self-management [78], revealing their limitless potential. However, despite their prevalence and versatility, the aspect of explainability in wearable technologies has received less attention. Previous reviews on wearables in health care have attracted the interest of researchers, covering a wide range of research papers. Reviews focusing on wearables have yielded many papers, such as 82 [20] or 73 [79] reviewed papers, whereas reviews on explainability have produced even more extensive results, such as 91 [13], 93 [80], or 196 [81] reviewed papers. In contrast, this review specifically delved deeper into the investigation of explainability in wearables, and thus, the number of relevant results decreased significantly, with only 25 papers. This discrepancy highlights the relatively limited focus on explainability within the realm of wearable technologies despite their vast potential and relevance in the health care domain. Notably, explainability is particularly vital for the lay public and health care professionals, who form the primary target audience for wearable devices.

Despite the limited focus on explainability in wearables, this review revealed that some of the findings align with general trends observed about wearables. Notably, wrist-worn wearables have emerged as the most commonly used technology, as seen in the Technologies for Capturing Quantified Self Data section, which is in line with previous research [82,83]. Specifically, Fitbit and Empatica E4 wristbands were identified as the most used wearables, with Fitbit being more dominant than Empatica E4 in the literature [82,83]. This can be due to the high cost of Empatica E4 in comparison to Fitbit devices [84] as it costs an average of US $1700, whereas the Fitbit Sense costs US $170, which is 10 times cheaper. However, the reason for choosing a specific wearable depends on the use case and aim of the study as some signals are not captured by Fitbit, such as those from temperature sensors and accelerometers. For example, the Empatica E4 wristband stands out for its promising potential in the unobtrusive measurement of HR variability [85]. The data collected from these wearables predominantly consisted of physiological signals, making them the most used input data for developing AI models. User-generated datasets significantly outnumbered benchmark datasets. When examining the experimental setups, it became evident that most studies (20/25, 80%) operated within uncontrolled real-world settings. A recent review underscored the constraints associated with imbalanced datasets primarily stemming from publicly available sources rather than real-world data [13], which is contrary to the findings of this review.

While wearables in health care have shown immense promise, there are critical considerations and recommendations to ensure their effective use while aligning with the goals of explainability. First, there is an urgent need to prioritize the development of XAI models and user interfaces for wearables. This is essential to bridge the gap in understanding between users and the insights derived from wearable data, thereby enhancing trust, reliability, and usability—key components of explainability. Second, when choosing wearables for health care applications, researchers and practitioners must carefully assess their specific data requirements against factors such as cost, comfort, and data capture capabilities. This strategic selection of wearables can significantly impact the quality and relevance of collected data, contributing to transparency and fairness in data collection. Moreover, it is crucial to continue emphasizing the use of real-world data in wearable studies. Such data better reflect the complexity of health care scenarios and enhance the practicality of research findings, ensuring that causal relationships and privacy concerns are appropriately addressed. By adhering to these recommendations and aligning them with the goals of explainability, we can ensure that wearable technology in health care not only realizes its potential but also serves users effectively, ethically, and transparently.

Explainability for Wearable Data

To address RQs 3 and 4, various explainability methods were explored, and a noticeable trend emerged toward using model-agnostic approaches (15/25, 60%) such as SHAP, as seen in the Stage of Explainability section. Model-agnostic methods have gained popularity due to their ability to be applied to a wide range of ML models regardless of complexity or type, unlike other methods such as gradient-based or attention-based methods. These methods provide explanations independently of the model's internal workings [86]. Interpretable models face a trade-off between accuracy and interpretability [87]. Model-agnostic interpretability treats the model as a black box, creating a separation between interpretability and the model. This approach allows the model to be flexible and versatile, accommodating diverse ML methods, including complex ones such as deep neural networks [86]. Decoupling interpretability from the model enables a balance to be achieved between accuracy and interpretability, providing a valuable tool for understanding and explaining the model's behavior [86]. Furthermore, this approach empowers medical professionals to gain valuable insights into the decision-making process and understand the impact of different features on predictions for conditions or diseases. By understanding how the AI models

XSL•FO
RenderX

arrive at their conclusions, medical professionals can build trust and confidence in using AI technology [88].

SHAP stands out as a widely used model-agnostic method (12/25, 48%), offering feature importance values based on the Shapley value from the cooperative game theory [28]. This framework helps interpret the impact of individual features on model predictions [28]. The trends in explainability identified in this review align with previous findings in the literature. A previous review categorized explainability into different types, such as feature, example-based, and textual explanations [89]. Feature explainability involves techniques such as SHAP, LIME, and CAM. In this review, a similar pattern emerged, with feature explainability being more prevalent. However, the choice of explainability method depends on the specific use case, the model in question, and desired interpretability levels. While model-agnostic methods such as SHAP have gained popularity, other techniques such as rule-based models, decision trees, and feature importance analysis may be more suitable for certain scenarios.

In this review, it was observed that post hoc explainability methods were more commonly used (22/25, 88%) than ante hoc explainability methods (3/25, 12%) in the field of wearables, as seen in the Stage of Explainability section. This trend is attributed to the fact that post hoc methods can be applied to any existing model regardless of its initial explainability capabilities. On the other hand, ante hoc methods require models to be designed and trained with explicit explainability mechanisms, making these methods less prevalent. These findings align with those of previous studies, which have reported a higher prevalence of post hoc methods [81,89]. However, compared to reviews generally focusing on explainability, ante hoc methods are used more frequently than post hoc methods [26,90]. This discrepancy may be attributed to different research contexts and objectives.

Furthermore, it was observed that most of the explainability methods reviewed in this study had a local scope (15/25, 60%) rather than a global one (10/25, 40%), as seen in the Scope of Explainability section. Local explainability methods focus on providing explanations for individual predictions or instances, allowing for a detailed understanding of how the model arrived at a specific decision. On the other hand, global explainability methods aim to provide insights into the model's overall behavior and decision-making process. The prevalence of local explainability methods in this review suggests a more targeted approach to understanding specific model predictions, which can be valuable in practical applications in which users may be more interested in individual predictions rather than a holistic view of the entire model. When discussing local and global explainability, it is valuable to consider and present both types together rather than focusing on only one [91].

The explainability of models was often conveyed through visual formats (17/25, 68%) using different types of graphs. Visual explainability has demonstrated greater prominence when compared to other outputs such as numerical, rule-based, and textual explanations [26,81,92]. This preference for visual representations highlights the effectiveness of using visualizations to enhance the interpretability and understanding of complex AI models. The level of explainability output is linked to the end-user audience. For instance, when the end user comprises the public, the output is tailored to be visually engaging, interactive, and user-friendly, as seen in the Multimodal Explanations section. In contrast, when the intended audience consists of researchers or medical professionals, the explanation may be presented through more technical means, such as data-rich visualizations (eg, heat maps), as seen in the Visual Explanations section.

Model explainability is paramount in the evolving landscape of AI research. The decisions in this realm are critical. Model-agnostic methods, notably SHAP, have garnered attention for their adaptability. A significant proportion of studies (24/25, 96%) adopted post hoc explainability as it offers the convenience of integration after model development. While there is a discernible tilt toward elucidating local predictions, a holistic approach encompassing both local and global perspectives is essential. Visual presentations stand out as the preferred mode to enhance the comprehension of AI models. From the findings, given that the datasets used in building the AI models were mostly from users (20/25, 80%) rather than benchmark datasets (5/25, 20%) and the emphasis on using local (15/25, 60%) rather than global (10/25, 40%) explainability, this showed a distinct move toward personalization. This shift is anticipated given the growing momentum of precision medicine on the global stage whether in research or practice and the emphasis on the importance of patient-centered health care.

Human-Centered XAI Needs More User-Based Evaluation

To address RQ 5, the primary objective of explainability is to create models that are transparent and understandable for end users, making it crucial to involve them in the validation process. However, in this review, of the 25 papers analyzed, only 5 (20%) reported user studies, indicating a limited focus on user evaluations. This trend aligns with findings in the broader literature, where only 1 in 5 papers included user evaluations [26]. Furthermore, the user evaluations conducted in the reviewed studies primarily targeted healthy individuals and patients with hypertension, limiting the representation of other populations. Despite the diversity of domains in the studies, the scope of user evaluation still needs to be narrower. Expanding the range of user evaluations across various populations is essential to enhance the real-world applicability and acceptance of AI models in the health care domain.

Assessing the explainability of wearables is crucial. As wearables rapidly evolve into essential tools for persistent health monitoring, they produce massive amounts of data daily. These complex data necessitate clear interpretation to make well-informed decisions [93]. As these devices become more woven into a person's daily life, the decisions based on their data significantly affect health and well-being. Such decisions are deeply personal. Hence, it is imperative not just to have accurate data interpretation but also for users to easily understand, use, and trust the results. If users, including medical professionals, cannot comprehend or have confidence in these insights, they might hesitate to act on them or could make incorrect choices.

XSL•FO
RenderX

Considering the personal and intimate nature of data from wearables, user studies are essential. Furthermore, gauging XAI's effectiveness with users is paramount to foster trust, customize explanations to the users' needs, understand the users' context, and ensure overall usability [93]. This assessment not only supports and evaluates the alignment with AI's ethical and regulatory standards but also enhances its overall acceptability. However, in-the-wild user studies remain a notable gap in this area [94]. The design, development, and evaluation of AI and ML systems must transition to real-world applications. This shift calls for interdisciplinary collaboration and multiple design and evaluation iterations [95]. While real-world evidence is beneficial for patients, caregivers, medical professionals, and society at large, adopting a human-centric approach can be challenging due to the sociotechnical aspects involved [53].

Limitations

While our literature search was comprehensive, the application of strict inclusion and exclusion criteria resulted in a relatively low number of studies being included in the review. We also stopped the collection and started the analysis and reporting work on December 31, 2022. This selection process, while ensuring the quality of the included studies, may limit the generalizability of our findings to a broader population of studies. In addition, the limited number of studies involving end users in our review underscores the need for caution when concluding this subset of the data. To address these limitations and enhance the robustness of the findings, further research with more extensive and diverse samples is recommended.

Implications for Practice and Future Development

The findings of this review have several implications for both current practice and future development in the field of XAI in wearable technologies for health care. For practice, this review highlights the importance of incorporating explainability into the design and implementation of wearable technologies used in health care. As these technologies become more prevalent in medical settings, medical professionals and end users must understand how AI models arrive at their decisions. By providing interpretable and transparent explanations, medical professionals can gain insights into the decision-making process of complex models and build trust and confidence in using AI technology. This, in turn, can lead to more effective and informed decision-making in patient care and treatment.

Moreover, this review emphasizes the need for more user evaluation and involvement in the development of XAI models. Understanding the perspectives and needs of end users is essential to ensure that the explanations provided by the models are meaningful, useful, and user-friendly. Future development should focus on integrating user feedback into the design process, enabling personalized explanations. In addition, this review highlights the significance of visual outputs for presenting explainability. While the current research landscape in XAI primarily relies on visualization and text-based explanations, there needs to be more exploration into alternative modalities for XAI, such as audio or gamification. To advance future development in this field, it is imperative to shift our focus from existing trends toward addressing these gaps. Instead of solely concentrating on established approaches, we should actively investigate novel methods that may offer more effective means of conveying complex model behavior to diverse audiences, including medical professionals and patients. By diversifying our approach to XAI and considering alternatives such as audio or gamification, we can strive to create simpler and more user-friendly explanations that not only foster greater understanding but also enhance the acceptance and adoption of AI technologies in health care.

Conclusions

This review highlights the nascent and growing significance of XAI in wearable technologies for health care, with 25 research papers included over the last 5 years. The results highlighted that, while wrist-worn wearables such as Fitbit and Empatica E4 are commonly used, the focus on explainability is relatively limited. Post hoc methods such as SHAP emerged as popular choices due to their versatility. Visual outputs are commonly used for user-friendly representations. However, this review highlighted the limitations in user evaluation and the importance of involving users in the process. The small number of human-centered evaluation studies limits the generalizability of the results. Overall, further research in the area of XAI and wearables can pave the way for more transparent and reliable AI models in health care applications.

Acknowledgments

This contribution was made possible by grant NPRP11C-0115-180010 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy used for the databases.
[DOCX File , 14 KB - jmir_v26i1e53863_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e53863_app2.docx]

Multimedia Appendix 3
Overview of explainability features per reviewed paper.
[DOCX File , 49 KB - jmir_v26i1e53863_app3.docx]

References

1. Iqbal MH, Aydin A, Brunckhorst O, Dasgupta P, Ahmed K. A review of wearable technology in medicine. J R Soc Med 2016 Oct 11;109(10):372-380 [FREE Full text] [doi: 10.1177/0141076816663560] [Medline: 27729595]
2. Duckworth C, Guy MJ, Kumaran A, O'Kane AA, Ayobi A, Chapman A, et al. Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations. J Diabetes Sci Technol 2024 Jan 13;18(1):113-123 [FREE Full text] [doi: 10.1177/19322968221103561] [Medline: 35695284]
3. Barricelli BR, Casiraghi E, Gliozzo J, Petrini A, Valtolina S. Human digital twin for fitness management. IEEE Access 2020;8:26637-26664. [doi: 10.1109/access.2020.2971576]
4. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl Based Syst 2023 Mar;263:110273. [doi: 10.1016/j.knosys.2023.110273]
5. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]
6. Nagahisarchoghaei M, Nur N, Cummins L, Nur N, Karimi MM, Nandanwar S, et al. An empirical survey on explainable AI technologies: recent trends, use-cases, and categories from technical and application perspectives. Electronics 2023 Feb 22;12(5):1092. [doi: 10.3390/electronics12051092]
7. Neches R, Swartout WR, Moore JD. Enhanced maintenance and explanation of expert systems through explicit models of their development. IEEE Trans Softw Eng 1985 Nov;SE-11(11):1337-1351. [doi: 10.1109/tse.1985.231882]
8. USACM issues statement on algorithmic transparency and accountability. Association for Computing Machinery. 2017. URL: https://www.acm.org/articles/bulletins/2017/january/usacm-statement-algorithmic-accountability [accessed 2024-04-29]
9. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. AI Mag 2019 Jun 24;40(2):44-58. [doi: 10.1609/aimag.v40i2.2850]
10. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). Comput Methods Programs Biomed 2022 Nov;226:107161. [doi: 10.1016/j.cmpb.2022.107161] [Medline: 36228495]
11. Bai X, Wang X, Liu X, Liu Q, Song J, Sebe N, et al. Explainable deep learning for efficient and robust pattern recognition: a survey of recent developments. Pattern Recognit 2021 Dec;120:108102. [doi: 10.1016/j.patcog.2021.108102]
12. Watson C, Cooper N, Palacio DN, Moran K, Poshyvanyk D. A systematic literature review on the use of deep learning in software engineering research. ACM Trans Softw Eng Methodol 2022 Mar 04;31(2):1-58 [FREE Full text] [doi: 10.1145/3485275]
13. Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. Decis Anal J 2023 Jun;7:100230. [doi: 10.1016/j.dajour.2023.100230]
14. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 2020 Jul;14(2):627-660. [doi: 10.5465/annals.2018.0057]
15. Ahmad MA, Teredesai A, Eckert C. Interpretable machine learning in healthcare. In: Proceedings of the 2018 IEEE International Conference on Healthcare Informatics. 2018 Presented at: ICHI '18; June 4-7, 2018; New York, NY p. 447 URL: https://ieeexplore.ieee.org/document/8419428 [doi: 10.1109/ichi.2018.00095]
16. Islam MR, Ahmed MU, Barua S, Begum S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl Sci 2022 Jan 27;12(3):1353. [doi: 10.3390/app12031353]
17. Qureshi R, Irfan M, Ali H, Khan A, Nittala AS, Ali S, et al. Artificial intelligence and biosensors in healthcare and its clinical relevance: a review. IEEE Access 2023;11:61600-61620. [doi: 10.1109/access.2023.3285596]
18. Chaudhari A, Sarode V, Udtewar S, Moharkar L, Patil L, Barreto F. A review of artificial intelligence for predictive healthcare analytics and healthcare IoT applications. In: Proceedings of the 2023 Conference on Intelligent Computing and Networking. 2023 Presented at: IC-ICN '23; December 16-17, 2023; Bhubaneswar, India p. 555-562 URL: https://link.springer.com/chapter/10.1007/978-981-99-3177-4_42 [doi: 10.1007/978-981-99-3177-4_42]
19. Jagatheesaperumal SK, Pham QV, Ruby R, Yang Z, Xu C, Zhang Z. Explainable AI over the internet of things (IoT): overview, state-of-the-art and future directions. IEEE Open J Commun Soc 2022;3:2106-2136. [doi: 10.1109/ojcoms.2022.3215676]
20. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 09;8(11):e18907 [FREE Full text] [doi: 10.2196/18907] [Medline: 33164904]

XSL•FO
RenderX

21. Vijayan V, Connolly JP, Condell J, McKelvey N, Gardiner P. Review of wearable devices and data collection considerations for connected health. Sensors (Basel) 2021 Aug 19;21(16):5589 [FREE Full text] [doi: 10.3390/s21165589] [Medline: 34451032]
22. Gouverneur P, Li F, Shirahama K, Luebke L, Adamczyk WM, Szikszay TM, et al. Explainable artificial intelligence (XAI) in pain research: understanding the role of electrodermal activity for automated pain recognition. Sensors (Basel) 2023 Feb 09;23(4):1959 [FREE Full text] [doi: 10.3390/s23041959] [Medline: 36850556]
23. Mankodiya H, Jadav D, Gupta R, Tanwar S, Alharbi A, Tolba A, et al. XAI-fall: explainable AI for fall detection on wearable devices using sequence models and XAI techniques. Mathematics 2022 Jun 09;10(12):1990. [doi: 10.3390/math10121990]
24. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2018 Aug 22;51(5):1-42. [doi: 10.1145/3236009]
25. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021 Jan;113:103655 [FREE Full text] [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
26. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput Surv 2023 Jul 13;55(13s):1-42. [doi: 10.1145/3583558]
27. Vilone G, Longo L. Classification of explainable artificial intelligence methods through their output formats. Mach Learn Knowl Extr 2021 Aug 04;3(3):615-661. [doi: 10.3390/make3030032]
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 4768-4777 URL: https://dl.acm.org/doi/10.5555/3295222.3295230
29. Clos J, Wiratunga N, Massie S. Towards explainable text classification by jointly learning lexicon and modifier terms. Semantic Scholar. URL: https://www.semanticscholar.org/paper/Towards-Explainable-Text-Classification-by-Jointly-Clos-Wiratunga/ca68496cb928636b8d735b4b3ad864daad50d386 [accessed 2024-09-25]
30. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv. Preprint posted online June 16, 2020 2020 [FREE Full text]
31. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2016 Presented at: NAACL '16; June 12-17, 2016; San Diego, CA p. 97-101 URL: https://aclanthology.org/N16-3020.pdf [doi: 10.1145/2939672.2939778]
32. van den Brink WJ, van den Broek TJ, Palmisano S, Wopereis S, de Hoogh IM. Digital biomarkers for personalized nutrition: predicting meal moments and interstitial glucose with non-invasive, wearable technologies. Nutrients 2022 Oct 24;14(21):4465 [FREE Full text] [doi: 10.3390/nu14214465] [Medline: 36364728]
33. Kim T, Kim H, Lee HY, Goh H, Abdigapporov S, Jeong M, et al. Prediction for retrospection: integrating algorithmic stress prediction into personal informatics systems for college students’ mental health. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-20 URL: https://dl.acm.org/doi/10.1145/3491102.3517701 [doi: 10.1145/3491102.3517701]
34. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
35. Gullapalli BT, Carreiro S, Chapman BP, Ganesan D, Sjoquist J, Rahman T. OpiTrack: a wearable-based clinical opioid use tracker with temporal convolutional attention networks. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Sep 14;5(3):1-29 [FREE Full text] [doi: 10.1145/3478107] [Medline: 35291374]
36. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
37. Maritsch M, Föll S, Lehmann V, Bérubé C, Kraus M, Feuerriegel S, et al. Towards wearable-based hypoglycemia detection and warning in diabetes. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI EA '20; April 25-30, 2020; Honolulu, HI p. 1-8 URL: https://dl.acm.org/doi/abs/10.1145/3334480.3382808 [doi: 10.1016/j.cmpb.2021.106461]
38. Chiang PH, Wong M, Dey S. Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure. IEEE J Transl Eng Health Med 2021;9:1-13. [doi: 10.1109/jtehm.2021.3098173]
39. Leitner J, Chiang PH, Khan B, Dey S. An mHealth lifestyle intervention service for improving blood pressure using machine learning and IoMTs. In: Proceedings of the 2022 IEEE International Conference on Digital Health. 2022 Presented at: ICDH '22; July 10-16, 2022; Barcelona, Spain p. 142-150 URL: https://ieeexplore.ieee.org/document/9861082 [doi: 10.1109/icdh55609.2022.00030]

XSL•FO
RenderX

40. Kim HH, Kim Y, Park YR. Interpretable conditional recurrent neural network for weight change prediction: algorithm development and validation study. JMIR Mhealth Uhealth 2021 Mar 29;9(3):e22183 [FREE Full text] [doi: 10.2196/22183] [Medline: 33779574]
41. Ng A, Wei B, Jain J, Ward EA, Tandon SD, Moskowitz JT, et al. Predicting the next-day perceived and physiological stress of pregnant women by using machine learning and explainability: algorithm development and validation. JMIR Mhealth Uhealth 2022 Aug 02;10(8):e33850 [FREE Full text] [doi: 10.2196/33850] [Medline: 35917157]
42. Chalabianloo N, Can YS, Umair M, Sas C, Ersoy C. Application level performance evaluation of wearable devices for stress classification with explainable AI. Pervasive Mob Comput 2022 Dec;87:101703. [doi: 10.1016/j.pmcj.2022.101703]
43. Gadaleta M, Radin JM, Baca-Motes K, Ramos E, Kheterpal V, Topol EJ, et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. NPJ Digit Med 2021 Dec 08;4(1):22. [doi: 10.1038/S41746-021-00533-1]
44. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. IEEE Trans Visual Comput Graphics 2014 Dec 31;20(12):1983-1992. [doi: 10.1109/tvcg.2014.2346248]
45. Creagh AP, Lipsmeier F, Lindemann M, Vos MD. Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones. Sci Rep 2021 Jul 12;11(1):14301 [FREE Full text] [doi: 10.1038/s41598-021-92776-x] [Medline: 34253769]
46. Park S, Li CT, Han S, Hsu C, Lee SW, Cha M. Learning sleep quality from daily logs. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019 Presented at: KDD '19; August 4-8, 2019; Anchorage, AK p. 2421-2429 URL: https://dl.acm.org/doi/10.1145/3292500.3330792 [doi: 10.1145/3292500.3330792]
47. Dong G, Cai L, Datta D, Kumar S, Barnes LE, Boukhechba M. Influenza-like symptom recognition using mobile sensing and graph neural networks. In: Proceedings of the 2021 Conference on Health, Inference, and Learning. 2021 Presented at: ACM CHIL '21; April 8-10, 2021; Virtual Event p. 291-300 URL: https://dl.acm.org/doi/10.1145/3450439.3451880 [doi: 10.1145/3450439.3451880]
48. Tang M, Dong G, Zoellner J, Bowman B, Abel-Rahman E, Boukhechba M. Using ubiquitous mobile sensing and temporal sensor-relation graph neural network to predict fluid intake of end stage kidney patients. In: Proceedings of the 21st ACM/IEEE International Conference on Information Processing in Sensor Networks. 2022 Presented at: IPSN '22; May 4-6, 2022; Milano, Italy p. 398-309 URL: https://ieeexplore.ieee.org/document/9825934 [doi: 10.1109/ipsn54338.2022.00031]
49. Uddin MZ, Soylu A. Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. Sci Rep 2021 Aug 12;11(1):16455 [FREE Full text] [doi: 10.1038/s41598-021-95947-y] [Medline: 34385552]
50. Kim JK, Bae MN, Lee KB, Hong SG. Identification of patients with sarcopenia using gait parameters based on inertial sensors. Sensors (Basel) 2021 Mar 04;21(5):1786 [FREE Full text] [doi: 10.3390/s21051786] [Medline: 33806525]
51. Kim JK, Bae MN, Lee K, Kim J, Hong SG. Explainable artificial intelligence and wearable sensor-based gait analysis to identify patients with osteopenia and sarcopenia in daily life. Biosensors (Basel) 2022 Mar 07;12(3):167 [FREE Full text] [doi: 10.3390/bios12030167] [Medline: 35323437]
52. Guthrie NL, Berman MA, Edwards KL, Appelbaum KJ, Dey S, Carpenter J, et al. Achieving rapid blood pressure control with digital therapeutics: retrospective cohort and machine learning study. JMIR Cardio 2019 Mar 12;3(1):e13030 [FREE Full text] [doi: 10.2196/13030] [Medline: 31758792]
53. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018;4:2055207618770325 [FREE Full text] [doi: 10.1177/2055207618770325] [Medline: 29942629]
54. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 2017 Sep 15;33(18):2938-2940 [FREE Full text] [doi: 10.1093/bioinformatics/btx364] [Medline: 28645171]
55. Woźniak PW, Kucharski PP, de Graaf MM, Niess J. Exploring understandable algorithms to suggest fitness tracker goals that foster commitment. In: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society. 2020 Presented at: NordiCHI '20; October 25-29, 2020; Tallinn, Estonia p. 1-12 URL: https://dl.acm.org/doi/10.1145/3419249.3420131 [doi: 10.1145/3419249.3420131]
56. Harris JJ, Chen CH, Zaki MJ. A framework for generating summaries from temporal personal health data. ACM Trans Comput Healthcare 2021 Jul 15;2(3):1-43. [doi: 10.1145/3448672]
57. Harris J, Zaki MJ. Towards neural numeric-to-text generation from temporal personal health data. arXiv. Preprint posted online July 11, 2022 2022 [FREE Full text]
58. Hatteland AH, Marcinkevičs R, Marquis M, Frick T, Hubbard I, Vogt JE. Exploring relationships between cerebral and peripheral biosignals with neural networks. In: Proceedings of the 2021 IEEE International Conference on Digital Health. 2021 Presented at: ICDH '21; September 5-10, 2021; Chicago, IL p. 103-113 URL: https://ieeexplore.ieee.org/document/9581203 [doi: 10.1109/icdh52753.2021.00022]
59. Delgado-Panadero Á, Hernández-Lorca B, García-Ordás MT, Benítez-Andrades JA. Implementing local-explainability in gradient boosting trees: feature contribution. Inf Sci 2022 Apr;589:199-212. [doi: 10.1016/j.ins.2021.12.111]

XSL•FO
RenderX

60. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning. 2017 Presented at: ICML '17; August 6-11, 2017; Sydney, Australia p. 3145-3153 URL: https://dl.acm.org/doi/10.5555/3305890.3306006
61. Binder A, Montavon G, Bach S, Müller KR, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers. arXiv. Preprint posted online April 4, 2016 2016 [FREE Full text] [doi: 10.1007/978-3-319-44781-0_8]
62. Zhang X, Liang X, Zhiyuli A, Zhang S, Xu R, Wu B. AT-LSTM: an attention-based LSTM model for financial time series prediction. IOP Conf Ser Mater Sci Eng 2019 Jul 01;569(5):052037. [doi: 10.1088/1757-899x/569/5/052037]
63. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. arXiv. Preprint posted online December 14, 2015 2015 [FREE Full text] [doi: 10.1109/cvpr.2016.319]
64. Wisdom S, Powers T, Pitton J, Atlas L. Interpretable recurrent neural networks using sequential sparse recovery. arXiv. Preprint posted online November 22, 2016 2016 [FREE Full text] [doi: 10.48550/arXiv.1611.07252]
65. Alarcón Á, Madrid NM, Seepold R, Ortega JA. The role of digital twins in personalized sleep medicine. In: Proceedings of the 2021 German-Italian Workshop on Social Innovation in Long-Term Care Through Digitalization. 2021 Presented at: WS-LTC '21; November 2-4, 2021; Berlin, Germany p. 71-79 URL: https://link.springer.com/chapter/10.1007/978-3-031-16855-0_8 [doi: 10.1007/978-3-031-16855-0_8]
66. Support vector machine (SVM) algorithm. GeeksforGeeks. URL: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989 [accessed 2024-08-25]
67. Yu H, Yang J, Han J. Classifying large data sets using SVMs with hierarchical clusters. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003 Presented at: KDD '03; August 24-27, 2003; Washington, DC p. 306-316 URL: https://dl.acm.org/doi/10.1145/956750.956786 [doi: 10.1145/956750.956786]
68. Sarkar A, Yang Y, Vihinen M. Variation benchmark datasets: update, criteria, quality and applications. Database (Oxford) 2020 Jan 01;2020:2969 [FREE Full text] [doi: 10.1093/database/baz117] [Medline: 32016318]
69. Weber I, Achananuparp P. Insights from machine-learned diet success prediction. Pac Symp Biocomput 2016;21:540-551 [FREE Full text] [Medline: 26776216]
70. Shinoda I, Tokida N, Kurobe M, Furukawa S, Hayashi K. Demonstration of a considerable amount of mouse epidermal growth factor in aqueous humor. Biochem Int 1988 Aug;17(2):243-248. [Medline: 3263860]
71. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. A public domain dataset for real-life human activity recognition using smartphone sensors. Comput Intell 2013;20:21 [FREE Full text]
72. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
73. Seijts GH, Latham GP. The construct of goal commitment: measurement and relationships with task performance. In: Goffin RD, Helmes E, editors. Problems and Solutions in Human: Assessment Honoring Douglas N. Jackson at Seventy. Cham, Switzerland: Springer; 2012:315-332.
74. Cramer H, Evers V, Ramlal S, van Someren M, Rutledge L, Stash N, et al. The effects of transparency on trust in and acceptance of a content-based art recommender. User Model User Adapt Interact 2008 Aug 20;18(5):455-496. [doi: 10.1007/S11257-008-9051-3]
75. Crowston K. Amazon Mechanical Turk: a research tool for organizations and information systems scholars. In: Proceedings of the 2012 Conference on Shaping the Future of ICT Research: Methods and Approaches. 2012 Presented at: IFIP WG '12; December 13-14, 2012; Tampa, FL p. 210-221 URL: https://link.springer.com/chapter/10.1007/978-3-642-35142-6_14 [doi: 10.1007/978-3-642-35142-6_14]
76. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 09;8(11):e18907 [FREE Full text] [doi: 10.2196/18907] [Medline: 33164904]
77. Andrade E, Quinlan L, Harte R, Byrne D, Fallon E, Kelly M, et al. Augmenting critical care patient monitoring using wearable technology: review of usability and human factors. JMIR Hum Factors 2021 May 25;8(2):e16491 [FREE Full text] [doi: 10.2196/16491] [Medline: 34032574]
78. Maddison R, Cartledge S, Rogerson M, Goedhart NS, Ragbir Singh T, Neil C, et al. Usefulness of wearable cameras as a tool to enhance chronic disease self-management: scoping review. JMIR Mhealth Uhealth 2019 Jan 03;7(1):e10371 [FREE Full text] [doi: 10.2196/10371] [Medline: 30609985]
79. Kristoffersson A, Lindén M. A systematic review on the use of wearable body sensors for health monitoring: a qualitative synthesis. Sensors (Basel) 2020 Mar 09;20(5):1502 [FREE Full text] [doi: 10.3390/s20051502] [Medline: 32182907]
80. Giuste F, Shi W, Zhu Y, Naren T, Isgut M, Sha Y, et al. Explainable artificial intelligence methods in combating pandemics: a systematic review. IEEE Rev Biomed Eng 2023;16:5-21. [doi: 10.1109/rbme.2022.3185953]
81. Vilone G, Longo L. Explainable artificial intelligence: a systematic review. arXiv. Preprint posted online May 29, 2020 2020 [FREE Full text]
82. Huhn S, Axt M, Gunga HC, Maggioni MA, Munga S, Obor D, et al. The impact of wearable technologies in health research: scoping review. JMIR Mhealth Uhealth 2022 Jan 25;10(1):e34384 [FREE Full text] [doi: 10.2196/34384] [Medline: 35076409]

XSL•FO
RenderX

83. Mattison G, Canfell O, Forrester D, Dobbins C, Smith D, Töyräs J, Sullivan C. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 01;24(7):e36690 [FREE Full text] [doi: 10.2196/36690] [Medline: 35776492]
84. Ronca V, Martinez-Levy AC, Vozzi A, Giorgi A, Aricò P, Capotorto R, et al. Wearable technologies for electrodermal and cardiac activity measurements: a comparison between Fitbit sense, Empatica E4 and Shimmer GSR3. Sensors (Basel) 2023 Jun 23;23(13):5847. [doi: 10.3390/s23135847] [Medline: 37447697]
85. Schuurmans AA, de Looff P, Nijhof KS, Rosada C, Scholte RH, Popma A, et al. Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J Med Syst 2020 Sep 23;44(11):190 [FREE Full text] [doi: 10.1007/s10916-020-01648-w] [Medline: 32965570]
86. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv. Preprint posted online June 16, 2016 2016 [FREE Full text]
87. Freitas AA. Comprehensible classification models: a position paper. SIGKDD Explor Newsl 2014 Mar 17;15(1):1-10. [doi: 10.1145/2594473.2594475]
88. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
89. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors (Basel) 2023 Jan 05;23(2):634 [FREE Full text] [doi: 10.3390/s23020634] [Medline: 36679430]
90. Bharati S, Mondal MR, Podder P. A review on explainable artificial intelligence for healthcare: why, how, and when? IEEE Trans Artif Intell 2024 Apr;5(4):1429-1442. [doi: 10.1109/tai.2023.3266418]
91. Radensky M, Downey D, Lo K, Popović Z, Weld DS. Exploring the role of local and global explanations in recommender systems. arXiv. Preprint posted online September 27, 2021 2021 [FREE Full text] [doi: 10.1145/3491101.3519795]
92. Chakrobartty S, El-Gayar OF. Explainable artificial intelligence in the medical domain: a systematic review. In: Proceedings of the 2021 AMCIS Conference on Digital Innovation and Entrepreneurship. 2021 Presented at: AMCIS '21; August 9-13, 2021; Virtual Event p. 2 URL: https://www.researchgate.net/publication/361361000_Explainable_Artificial_Intelligence_in_the_Medical_Domain_A_Systematic_Review
93. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
94. Andersen TO, Nunes F, Wilcox L, Coiera E, Rogers Y. Introduction to the special issue on human-centred AI in healthcare: challenges appearing in the wild. ACM Trans Comput Hum Interact 2023 Jun 30;30(2):1-12. [doi: 10.1145/3589961]
95. Colusso L, Jones R, Munson SA, Hsieh G. A translational science model for HCI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Scotland, UK p. 1-13 URL: https://dl.acm.org/doi/10.1145/3290605.3300231 [doi: 10.1145/3290605.3300231]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural networkDL: deep learning
ECG: electrocardiographyEDA: electrodermal activityHR: heart rate
LIME: local interpretable model-agnostic explanations
LSTM: long short-term memory
MeSH: Medical Subject Heading
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question
SHAP: Shapley Additive Explanations
SVM: support vector machine
XAI: explainable artificial intelligence

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 23.10.23; peer-reviewed by TAR Sure, G Vos, GA Tsihrintzis, CI Nwakanma; comments to author 06.05.24; revised version received 31.08.24; accepted 06.11.24; published 24.12.24.

Please cite as:
Abdelaal Y, Aupetit M, Baggag A, Al-Thani D
Exploring the Applications of Explainability in Wearable Data Analytics: Systematic Literature Review
J Med Internet Res 2024;26:e53863
URL: https://www.jmir.org/2024/1/e53863
doi:10.2196/53863
PMID:39718820

©Yasmin Abdelaal, Michaël Aupetit, Abdelkader Baggag, Dena Al-Thani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Consideration of Cybersecurity Risks in the Benefit-Risk Analysis of Medical Devices: Scoping Review

Oscar Freyer[1], Approbierter Arzt; Fatemeh Jahed[1], BSc; Max Ostermann[1], MSc; Christian Rosenzweig[2], Dipl-Ing; Pascal Werner[3], MSc, MEng; Stephen Gilbert[1], PhD

[1]Else Kröner Fresenius Center for Digital Health, Dresden University of Technology, Dresden, Germany

[2]Johner Institute, Konstanz, Germany

[3]Regulatory.me, Mebane, NC, United States

Corresponding Author:
Oscar Freyer, Approbierter Arzt
Else Kröner Fresenius Center for Digital Health
Dresden University of Technology
Fetscherstraße
Dresden, 01307
Germany
Phone: 49 351 458 7558
Email: oscar.freyer@tu-dresden.de

Abstract

Background: The integration of connected medical devices (MDs) into health care brings benefits but also introduces new, often challenging-to-assess risks related to cybersecurity, which have the potential to harm patients. Current regulations in the European Union and the United States mandate the consideration of these risks in the benefit-risk analysis (BRA) required for MD approval. This important step in the approval process weighs all the defined benefits of a device with its anticipated risks to ensure that the product provides a positive argument for use. However, there is limited guidance on how cybersecurity risks should be systematically evaluated and incorporated into the BRA.

Objective: This scoping review aimed to identify current legal frameworks, guidelines, and standards in the United States, Canada, South Korea, Singapore, Australia, the United Kingdom, and the European Union on how cybersecurity risks should be considered in the BRA of MDs.

Methods: This scoping review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework. A systematic literature search of 10 databases was conducted in two phases on July 3, 2024 and September 30, 2024, including the guidance databases of the Food and Drug Administration, the Medical Device Coordination Group, and other International Medical Device Regulators Forum members; the International Medical Device Regulators Forum database; PubMed; and Scopus. Search terms included “cybersecurity,” “security,” “benefit/risk,” “benefit-risk,” and “risk-benefit.” Additional references were identified via citation searching and expert interviews. Inclusion criteria were met if a document was a guideline or standard in force that provided guidance on the BRA or cybersecurity risks of MDs. Documents were excluded when they were not relevant to MDs, they were limited to a subclass of devices, they were about in vitro diagnostic MDs or investigational devices, and the content of the source was insufficient to undertake a scientific analysis. Data were extracted and analyzed using MAXQDA 2022, and the findings were narratively summarized and visualized in figures and tables.

Results: The search identified 150 documents, with 34 (22.7%) meeting the inclusion criteria. These 34 documents included 4 (12%) regulations, 5 (15%) standards, 6 (18%) technical reports, and 19 (56%) guidance documents. While cybersecurity risks were acknowledged in most documents, detailed methods for their integration into the BRA were lacking. Some standards and guidelines provided examples of how to consider cybersecurity risks in the BRA, but a comprehensive and standardized approach was lacking.

Conclusions: This review highlights a substantial gap between the recognition of cybersecurity risks in MDs and the guidance on their incorporation into the BRA. Standardized frameworks are needed to provide clear methods for evaluating cybersecurity risks and their impact on the safety and security of MDs.

(J Med Internet Res 2024;26:e65528) doi:10.2196/65528

XSL•FO
RenderX

KEYWORDS

cybersecurity; connected medical devices; benefit-risk analysis; risk management; patient safety; regulation; vulnerability assessment

Introduction

Background

Connected medical devices (cMDs) have become an integral part of modern health care and play a crucial role in diagnosing, monitoring, and treating a wide range of medical conditions [1]. These devices have integrated software or are entirely software based, allowing them to connect to other devices or networks to exchange, transfer, or receive commands and data [1]. Examples include smartwatches, implanted devices often referred to as the Internet of Medical Things [2], and stationary devices such as computed tomography scanners [1,3]. While their connectivity offers multiple advantages, such as real-time disease and physiology monitoring [1,4,5] along with the potential for remote device management and over-the-air updating [6], it also introduces new risks for patients, particularly related to cybersecurity [4,5,7,8]. These risks are not just theoretical but could harm patients [9,10]. "Cybersecurity risks" is used as an umbrella term for risks specifically arising from cybersecurity vulnerabilities and measures that could affect the security (that is, a state in which information assets are protected) and safety (the absence of unacceptable risks to the patient's health) of a medical device (MD) [11,12]. However, those 2 areas overlap to some degree as security risks could exist that affect the MD's safety [13]. Thus, cybersecurity risks can lead to harm "to people, property, and the environment" [12].

The relevance of cybersecurity risks is underscored by a rising number of cyberattacks on health care infrastructure [9,14] while, at the same time, the US Food and Drug Administration (FDA) reports a substantial number of adverse events related to cybersecurity vulnerabilities [15] and has even recalled several devices because of them [16,17]. Regulatory frameworks such as the Medical Device Regulation (EU_MDR) in the European Union (EU) and the Federal Food, Drug, and Cosmetic Act (FD&C Act) in the United States mandate that manufacturers secure cMDs against these vulnerabilities to ensure that the risks for patients are as minimal as possible to guarantee a high level of health and safety protection [18-20]. Those high-level requirements are often further specified in guidance documents provided by authorities such as the Medical Device Coordination Group (MDCG) for the EU, the FDA for the United States, or the Therapeutic Goods Administration (TGA) for Australia [13,21,22].

To assess whether these requirements for health and safety protection are met, manufacturers are often obligated under legislation in multiple regions, including the United States, Australia, and the EU, to carry out a benefit-risk analysis (BRA), which determines whether the benefits of an MD outweigh its risks [18-20,23]. This important step in the approval process weighs all defined benefits of a device with its anticipated risks (including the risk detected for similar devices) considering the device's intended use to ensure that the product as a whole provides a positive argument for use [24].

As guidance documents often prescribe a BRA but rarely prescribe a precise methodology, there are multiple qualitative, quantitative, and semiquantitative methods in use [25,26], such as the multicriteria decision analysis (MCDA) [27], the quantitative BRA [28], or the FDA's Benefit-Risk Framework (BRF) [24]. Despite the prevalence of qualitative methods among manufacturers [29], there is currently no established standard for their use. Qualitative methods are often critiqued for their subjectivity and lack of rigor, whereas quantitative methods are often regarded as superior by multiple researchers [28,30,31]. However, the applicability of quantitative methods, which often originate from the pharmaceutical industry, is limited for MDs due to the inherent challenges in quantifying risks associated with aspects of their use (eg, cybersecurity risks) [25]. The importance of the consideration of cybersecurity risks in the BRA is addressed in several guidance documents [13,22,32]. However, the guidance fails to provide clear instructions on how to evaluate and incorporate these often challenging-to-assess risks into the BRA [11,12,25].

Objectives

To address the existing ambiguities regarding the BRA and cybersecurity, this review aimed to identify the current legal frameworks, guidelines, and standards in the United States, Canada, South Korea, Singapore, Australia, the United Kingdom, and the EU on how cybersecurity risks should be considered in the BRA of MDs.

Methods

This scoping review was conducted according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [33]. The completed PRISMA-ScR checklist can be found in Table S1 in Multimedia Appendix 1 [33].

Search Strategy

The first literature search was conducted on July 3, 2024, and a second search was conducted on September 30, 2024, in the regulatory guidance databases of 7 International Medical Device Regulators Forum (IMDRF) member states that provide relevant guidance in English: the US FDA guidance database, the EU MDCG document database, the Australian TGA guidance database, the Health Canada guidance document database, the South Korean Ministry of Food and Drug Safety (MFDS) regulation database, the UK guidance and regulation database (with a filter for Medicines and Healthcare Products Regulatory Agency [MHRA] documents), and the Singaporean Health Sciences Authority (HSA) database of guidance documents for MDs. In addition, the IMDRF guidance document database, Scopus, and PubMed were searched. Each guidance database was searched using the terms "cybersecurity," "security," "benefit-risk," "benefit/risk," and "risk-benefit." Scopus was

XSL•FO
RenderX

searched using the following search string: *TITLE-ABS-KEY ((cybersecurity OR "information security" OR "cyber security") AND ("medical device*" OR "health device*") AND ("benefit-risk" OR "risk-benefit" OR "benefit/risk"))*. PubMed was searched using the following search string: (*cybersecurity OR "information security" OR "cyber security") AND ("medical device*" OR "health device*") AND ("benefit-risk" OR "risk-benefit" OR "benefit/risk")*. In addition, a reference search of the included guidelines was conducted to identify any other relevant standards or guidelines referenced in the official guidance documents, and 2 experts on the regulation of cybersecurity in MDs (authors CR and PW) were asked to provide the most recent guidance documents that are not yet referenced in the official guidelines. The data from all the identified guidelines were exported and gathered in an electronic database (Microsoft Excel for Mac; version 16.86.3).

Inclusion and Exclusion Criteria

Sources were included in this review if they met the following inclusion criteria: (1) they were a regulation, guideline, or standard or an academic paper describing regulations, guidelines, or standards; (2) they were in force (applicable for regulations, guidelines, and standards); and (3) they delivered guidance for the BRA of MDs or cybersecurity risks of MDs. Sources were excluded when (1) they were not relevant to MDs, (2) they were limited to a specific subclass of MDs and not generalizable (eg, digital diabetes devices), (3) they were about in vitro diagnostic MDs or investigational devices, and (4) the content of the source was insufficient to undertake a scientific analysis.

Study Selection

The titles and summaries of the identified sources were screened by 2 independent researchers between July 3, 2024, and October 6, 2024, to evaluate whether they met the criteria for inclusion. In case of disagreements, a third independent reviewer was consulted. After the initial screening, 1 reviewer screened the full text of the included guidelines for eligibility.

Data Extraction and Analysis

Data extraction and analysis were conducted using MAXQDA 2022 (VERBI GmbH). One researcher identified keywords related to the BRA and cybersecurity risks within each document. Thereafter, any guidance and recommendations related to both topics and to the intersection of both were extracted. The extracted data and the relationships between different data sources were then synthesized using figures, tables, and narrative summaries. Recommendations regarding cybersecurity risks in the BRA provided in the documents were listed.

Results

Search Results

The systematic search of the 10 databases resulted in 150 documents. Of those 150 documents, 16 (10.7%) were included in this review. The reference search and expert interviews retrieved another 30 documents, of which 18 (60%) were included in this review. In total, 34 documents were included in this review. Figure 1 [34] shows the flowchart of the screening process and was prepared in accordance with the template provided by Page et al [34].

Figure 1. Flowchart of the screening process. This flowchart follows the template provided by Page et al [34]. BRA: benefit-risk analysis; EU: European Union; FDA: Food and Drug Administration; HSA: Health Sciences Authority; IMDRF: International Medical Device Regulators Forum; MD: medical device; MDCG: Medical Device Coordination Group; MFDS: Ministry of Food and Drug Safety; MHRA: Medicines and Healthcare Products Regulatory Agency; TGA: Therapeutic Goods Administration.

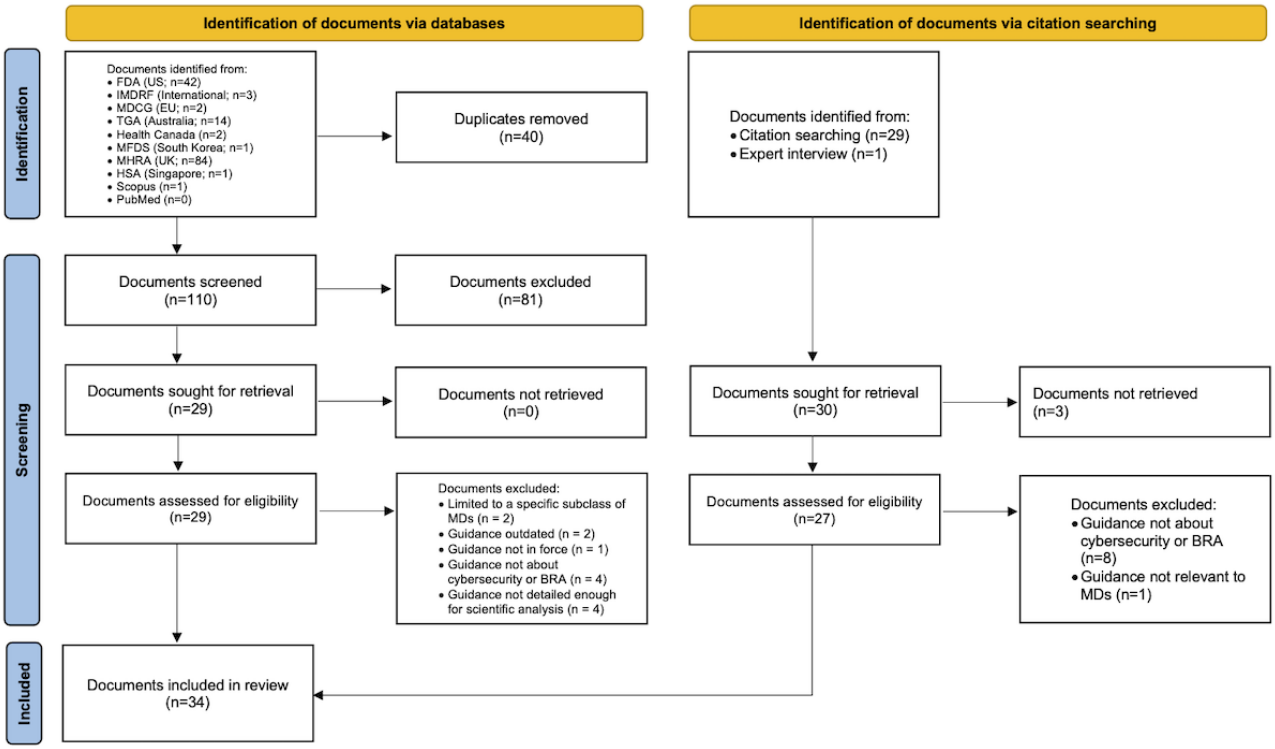

XSL•FO
RenderX

Source Characteristics

Among the 34 documents included in the study were 4 (12%) regulations, 5 (15%) standards, 6 (18%) technical reports, and 19 (56%) guidance documents. In total, 6% (2/34) of the documents were applicable only in the EU, 18% (6/34) were applicable in the United States, 12% (4/34) were applicable in Australia, 6% (2/34) were applicable in Canada, 6% (2/34) were applicable in Singapore, 3% (1/34) were applicable in South Korea, 3% (1/34) were applicable in the United Kingdom, and 47% (16/34) were applicable internationally. A total of 15% (5/34) of the documents only described the concept of BRA without mentioning cybersecurity explicitly, 18% (6/34) described cybersecurity risks without mentioning the BRA, and 68% (23/34) described the intersection of both. In total, 24% (8/34) of the documents provided examples or methods on how to consider cybersecurity risks in the BRA. Table 1 provides an overview of the included documents, whereas Figure 2 shows a visual representation of the relationship between the documents. To improve readability, documents from the same source describing similar content were grouped together (eg, FDA_Cyber encompasses all Food and Drug Administration cybersecurity guidance and IEC_80001 encompasses all substandards of this standard family mentioned in this paper) and only the short titles of the included documents are used. A list with the URLs to the included documents can be found in Table S2 in Multimedia Appendix 1.

Figure 2. The relationships among the included documents. The nodes are color grouped by type. The edge color is inherited from the node it originates from. FDA_BRA: Benefit Risk Analysis Guidance from the Food and Drug Administration; FDA_Cyber: Cybersecurity Guidance from Food and Drug Administration HC_Cyber: Pre-market Requirements for Medical Device Cybersecurity; HSA_Cyber: Regulatory Guidelines for Software Medical Devices—A Life Cycle Approach; HSA_GN-20: GN-20: Guidance on Clinical Evaluation; IMDRF_Cyber: Cybersecurity Guidance from the International Medical Device Regulators Forum; MFDS_Cyber: Guideline on Review and Approval for Cybersecurity of Medical Devices (For industry); NIST_CSF: National Institute of Standards and Technology Cybersecurity Framework; TGA_Cyber: Cybersecurity Guidance from the Therapeutic Goods Administration.

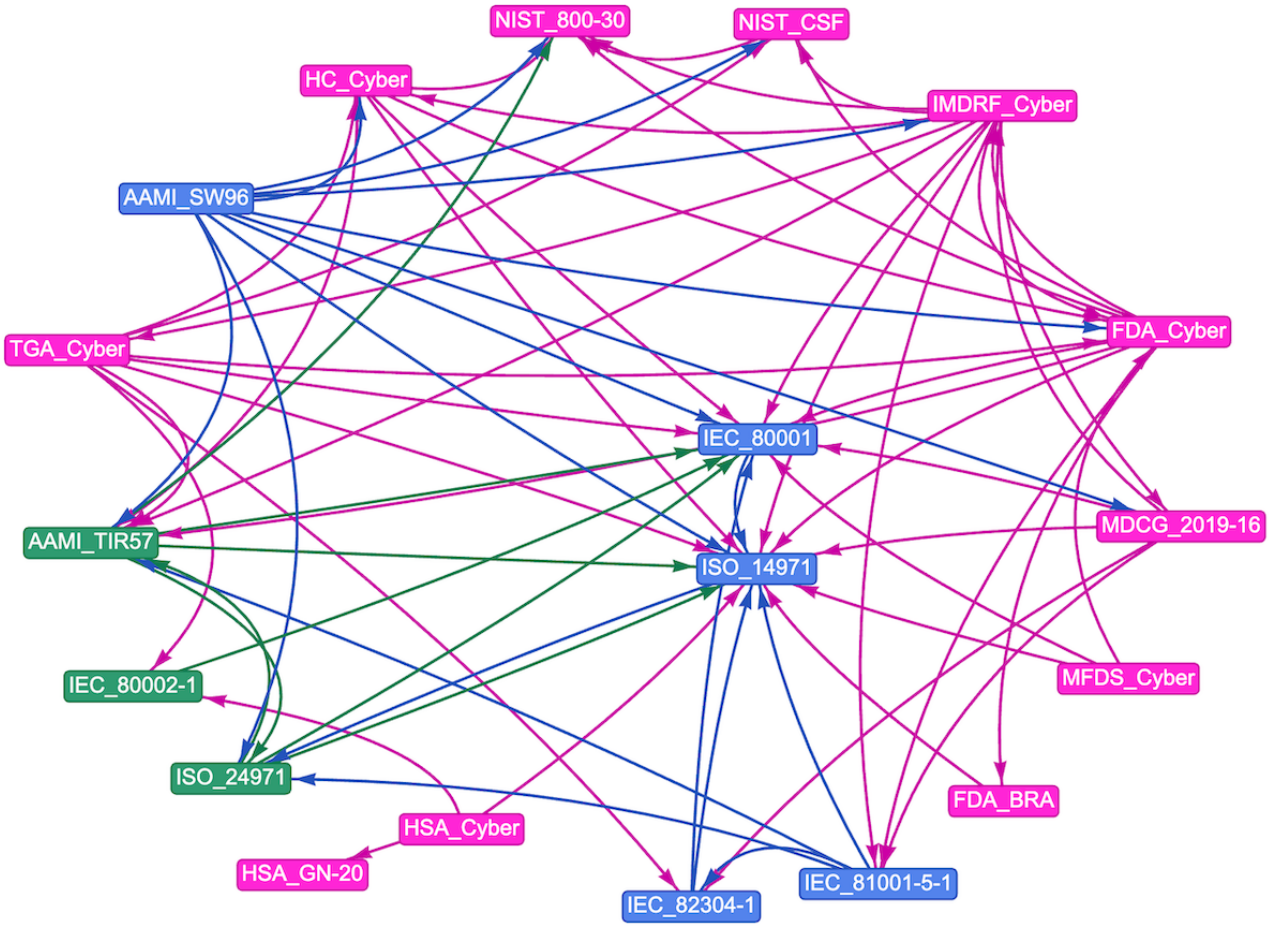

XSL•FO
RenderX

Table 1. Characteristics of the included documents. The table provides the complete name of the sources and the abbreviations used in this paper, if applicable. The year of publication and the year of the last update, if applicable, are provided.

Name	Abbreviation	Type	Organization	Year of publication	Year of last update	Description
Federal Food, Drug, and Cosmetic Act [19]	US_FD&C	Regulation	US House of Representatives	1938	2024	This is the law that grants the FDA[a] the power to regulate MDs[b]. It establishes high-level rules for MDs and their approval process and defines requirements for many aspects of MDs, including cybersecurity and the BRA[c].
Medical Devices Regulations (SOR/98-282) [23]	CA_MDR	Regulation	Parliament of Canada	1985	2024	This Canadian regulation defines basic principles and high-level requirements for the approval of MDs in Canada.
Therapeutic Goods (Medical Devices) Regulations 2002 [20]	AU_MDR	Regulation	Parliament of Australia	2002	2024	This Australian regulation defines basic principles and high-level requirements for the approval of MDs in Australia. The requirements are organized into 15 essential principles.
Regulation (EU[d]) 2017/745 of the European Parliament and of the Council of 5 April 2017, on MDs, amending Directive 2001/83/EC, Regulation (EC[e]) 178/2002, and Regulation (EC) 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (text with EEA[f] relevance) [18]	EU_MDR	Regulation	EU Parliament and Council	2017	2024	This European regulation came into force in 2017. It provides high-level rules and requirements for MDs that are or will be placed on the European market.
NIST[g] SP[h] 800-30 [35]	NIST_800-30	Guidance	NIST	2012	__[i]	This special publication by the NIST defines principles for how to conduct risk assessments for information systems. It is not specifically designed for MDs.
NIST Cybersecurity Framework [36]	NIST_CSF2.0	Guidance	NIST	2014	2024	This document by the NIST provides general guidance on cybersecurity risk management. It is not specifically designed for MDs.
Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions [37]	FDA_BRA_AC & ED	Guidance	FDA	2016	2018	This document provides guidance for FDA staff and industry on the factors considered in the BRA regarding availability, compliance, and enforcement decisions.
Postmarket Management of Cybersecurity in Medical Devices [32]	FDA_Cyber_Post	Guidance	FDA	2016	2018	Provides guidance on managing cybersecurity vulnerabilities in MDs that are already on the market, emphasizing continuous monitoring and identifying and addressing cybersecurity threats to ensure safety and effectiveness throughout their life cycle.
GN-20: Guidance on Clinical Evaluation [38]	HSA_GN-20	Guidance	HSA[j]	2017	2022	This document provides guidance for the clinical evaluation required for the registration of MDs in Singapore.

XSL•FO
RenderX

Name	Abbreviation	Type	Organization	Year of publication	Year of last update	Description
Consideration of Uncertainty in Making Benefit-Risk Determinations in Medical Device Premarket Approvals, De Novo Classifications, and Humanitarian Device Exemptions [39]	FDA_BRA_Uncertainty	Guidance	FDA	2019	—	This document provides guidance on how the FDA considers uncertainty in MD premarket approvals and de novo classifications and describes factors influencing this uncertainty.
Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and De Novo Classifications [40]	FDA_BRA	Guidance	FDA	2019	—	This document provides guidance on how the FDA conducts the BRA, explaining the factors considered and how they should be applied.
MDCG[k] 2019-16, revision 1 [13]	MDCG_2019-16	Guidance	MDCG	2019	2020	Provides guidance for MD manufacturers on how to address the cybersecurity requirements stated by the MDR[l]. It outlines the necessary steps that manufacturers should take to ensure that their devices are secured against cyber threats throughout the product life cycle.
Pre-market Requirements for Medical Device Cybersecurity [22]	HC_Cyber	Guidance	Health Canada	2019	—	This document provides nonbinding guidance for the cybersecurity of MDs and about the information that should be provided when applying for a license in Canada.
Medical device cyber security guidance for industry [21]	TGA_Cyber_Dev	Guidance	TGA[m]	2019	2022	This document provides guidance on how to comply with the essential principles for MDs in Australia, focusing on cybersecurity.
Medical device cyber security information for users [41]	TGA_Cyber_User	Guidance	TGA	2019	2022	This document provides recommendations for users of MDs regarding cybersecurity practices. It emphasizes the importance of users and operating environments for secure MDs.
Principles and Practices for Medical Device Cybersecurity [42]	IMDRF_Cyber	Guidance	IMDRF[n]	2020	—	It provides principles and recommendations regarding the cybersecurity of MDs for all stakeholders and emphasizes the relevance of a risk management process for cybersecurity risks.
Regulatory Guidelines for Software Medical Devices—A Life Cycle Approach [43]	HSA_Cyber	Guidance	HSA	2020	2024	The guideline details the regulatory requirements for software in MDs or as an MD over the product's life cycle. This includes aspects of cybersecurity, including risk management.
Guideline on Review and Approval for Cybersecurity of Medical Devices (For industry) [44]	MFDS_Cyber	Guidance	MFDS[o]	2020	—	This nonbinding guideline for manufacturers provides an overview of the requirements and how they can be met for the approval of MDs in South Korea.
Software and AI as a Medical Device Change Programme [45]	MHRA_SaMD	Guidance	MHRA[p]	2021	2023	This document provides an overview of upcoming changes to the regulation of MDs in the United Kingdom regarding software and AI[q]. It does not deliver clear guidance in its current state.

XSL•FO
RenderX

Name	Abbreviation	Type	Organization	Year of publication	Year of last update	Description
Complying with the Essential Principles on the safety and performance of medical devices [46]	TGA_Safety	Guidance	TGA	2022	2024	This guidance describes how the 15 essential principles for MD approval in Australia defined in the "Therapeutic Goods (Medical Devices) Regulations 2002" should be met and addressed.
Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions [22]	FDA_Cyber_Pre	Guidance	FDA	2023	—	Provides guidance for MD manufacturers on addressing cybersecurity in their premarket submissions, emphasizing the importance of integrating robust cybersecurity measures throughout the device's life cycle to ensure safety and effectiveness.
Principles and Practices for Software Bill of Materials for Medical Device Cybersecurity [47]	IMDRF_SBOM	Guidance	IMDRF	2023	—	Outlines principles and recommendations for the implementation of an SBOM[r] to enhance cybersecurity for MDs.
Principles and Practices for the Cybersecurity of Legacy Medical Devices [48]	IMDRF_LegMD	Guidance	IMDRF	2023	—	Provides principles and recommendations for maintaining the cybersecurity of legacy MDs.
IEC[s] 82304-1:2016 [49]	IEC_82304-1	Standard	IEC	2016	—	This document provides requirements for safety and security over the entire life cycle of MDs that constitute or contain software.
ISO[t] 14971:2019 [11]	ISO_14971	Standard	ISO	2019	—	Describes how the risk management process for MDs should be conducted by MD manufacturers.
IEC 80001-1:2021 [50]	IEC_80001-1	Standard	IEC	2021	—	This standard describes the requirements for different stakeholders in the risk management process of MDs connected to a health care infrastructure. It covers multiple risk types, including cybersecurity risks.
IEC 81001-5-1:2021 [51]	IEC_81001-5-1	Standard	IEC	2021	—	This standard defines life cycle requirements for MDs regarding cybersecurity, including best practices and security risk management.
ANSI[u]/AAMI[v] SW96:2023 [52]	AAMI_SW96	Standard	AAMI	2023	—	This standard applies the framework provided within ISO 14971:2019 to security risk management.
IEC 80002-1:2009 [53]	IEC_80002-1	Technical report	IEC	2009	—	This technical report details how to apply the risk management framework of ISO 14971 to software as an MD or in MDs.
IEC/TR[w] 80001-2-2:2012 [54]	IEC_80001-2-2	Technical report	IEC	2012	—	This document provides a framework for disclosing security-related capabilities and risks associated with the risk management of MDs connected to IT networks.
IEC/TR 80001-2-1:2012 [55]	IEC_80001-2-1	Technical report	IEC	2012	—	This document provides guidance on how to implement the risk management process described in IEC 80001-1:2021.
ISO/TR 80001-2-7:2015 [56]	ISO_80001-2-7	Technical report	ISO	2015	—	This document provides guidance on how to assess MD conformity with IEC 80001-1:2021.

XSL•FO
RenderX

Name	Abbreviation	Type	Organization	Year of publication	Year of last update	Description
AAMI TIR57:2016/(R)2023 Principles for medical device security—Risk Management [57]	AAMI_TIR57	Technical report	AAMI	2016	2023	Provides guidance on methods for performing cybersecurity risk management for MDs incorporating principles from ISO 14971:2019 and IEC 80001-1:2021.
ISO TR 24971:2020 [12]	ISO_24971	Technical report	ISO	2020	—	Provides guidance on the application of ISO 14971:2019.

[a]FDA: Food and Drug Administration.
[b]MD: medical device.
[c]BRA: benefit-risk analysis.
[d]EU: European Union.
[e]EC: European Community.
[f]EEA: European Economic Area.
[g]NIST: National Institute of Standards and Technology.
[h]SP: Special Publications.
[i]The document was not updated.
[j]HSA: Health Sciences Authority.
[k]MDCG: Medical Device Coordination Group.
[l]MDR: Medical Device Regulation.
[m]TGA: Therapeutic Goods Administration.
[n]IMDRF: International Medical Device Regulators Forum.
[o]MFDS: Ministry of Food and Drug Safety.
[p]MHRA: Medicines and Healthcare Products Regulatory Agency.
[q]AI: artificial intelligence.
[r]SBOM: software bill of materials.
[s]IEC: International Electrotechnical Commission.
[t]ISO: International Organization for Standardization.
[u]ANSI: American National Standards Institute.
[v]AAMI: Association for the Advancement of Medical Instrumentation.
[w]TR: Technical Report.

Regulations in the EU, the United States, Canada, and Australia

The scope and requirements for MD manufacturers in the EU are determined at a high level by the Regulation (EU) 2017/745 of the European Parliament and of the Council on MDs (EU_MDR) [18] and by the Regulation (EU) 2017/746 of the European Parliament and of the Council on in vitro diagnostic MDs [58]. The latter was excluded from this scoping review. The provided rules are further defined by guidance documents issued by the MDCG or by referring to existing harmonized standards developed or adapted from existing international standards by a European standard organization [18]. In the United States, the FD&C Act defines foundational requirements for MDs and empowers the FDA to establish detailed rules for MDs and enforce them [19]. The situation is similar in the other regions analyzed. While the fundamental regulation is usually passed by the legislature, it is specified by executive organizations and authorities such as the TGA in Australia or the HSA in Singapore [20,21,23,43,45,59].

In the EU, the EU_MDR addresses BRA requirements in multiple sections, requiring weighing all known risks of a device against its benefits to the patient [18] without further specifying the types of risks. While not mentioned explicitly, it is generally assumed that cybersecurity risks should be considered. Cybersecurity risks are not mentioned in detail; instead, the EU_MDR defines 2 broad requirements using the terms “information security” and “IT security”: MDs that include software or are software must be developed following the state of the art and must have protective measures against unauthorized access [18]. This state of the art is not further defined.

In the United States, the FD&C Act section 515(a) states that the effectiveness and safety of an MD are determined by “weighing any probable benefit to health from the use of the device against any probable risk of injury or illness from such use” [19]. High-level cybersecurity requirements are defined in section 524B of the FD&C Act to monitor and address postmarket vulnerabilities, ensure device and system security, provide software updates and a software bill of materials (SBOM), and comply with additional regulations for cybersecurity assurance [19]. Similarly to the EU, the intersection of cybersecurity risks and BRA is not mentioned explicitly within the regulation, but the wording in both the EU_MDR and the FD&C Acts suggests a consideration.

XSL•FO
RenderX

The Canadian Medical Device Regulations require conducting a BRA to determine whether a product is safe and should be put on the market. It is not specified which risks should be considered or how the BRA should be conducted. Cybersecurity is not mentioned [23].

In the Australian Therapeutic Goods (Medical Devices) Regulations (AU_MDR) 2002, the requirement of the BRA and high-level cybersecurity requirements are mentioned separately. The performance of a BRA is prescribed in "Essential Principle 6," with the requirement that the benefits of an MD must outweigh any undesirable effects arising from its use [20]. While cybersecurity is not mentioned directly, additional guidance clarifies that cybersecurity risks are within the scope of this principle [21]. Additional cybersecurity requirements are defined in "Essential Principle 12," including the implementation of a risk management process [20].

Guidelines for the BRA and Cybersecurity

Among the 34 documents, we identified 19 (56%) guidelines currently used within the included regions that describe the concept of the BRA or cybersecurity risks. Table 2 provides an overview of the identified regulations, guidelines and standards, and their recommendations for including cybersecurity risks in the BRA.

One guideline was issued by the MDCG, expanding and framing the cybersecurity requirements provided in the EU_MDR. In addition to implementing cybersecurity measures, manufacturers must establish a process for managing cybersecurity risks that considers the device's safety, security, and effectiveness. The residual security risks (which could include cybersecurity ones) should then be considered in a newly named concept called the "security benefit-risk analysis" [13]. However, it remains unclear when the BRA should be performed as the corresponding passage is ambiguous: the BRA could be divided into subcategories identified in the risk assessment (eg, cybersecurity risks, electromagnetic risks, and usability risk), or all risks, including cybersecurity ones, may be considered within the overall BRA [13]. Further details or guidance about the method to use are not provided.

A total of 9% (3/34) of the documents were issued by the FDA, providing nonbinding but recommended guidance on how to conduct a BRA for MD approval in the United States, including practical examples [37,39,40]. The qualitative method described in those guidance documents is known as the FDA's BRF and could be adapted for the consideration of cybersecurity risks; however, in its current state, it provides no specific guidance on how this can be done. The other 2 FDA guidelines describe cybersecurity requirements for MDs, mentioning the relevance of cybersecurity risks for BRA multiple times [22,32] and partly referring to the FDA's BRF.

The cybersecurity guidance provided by Health Canada recommends incorporating cybersecurity risks into the risk management process, whereas the BRA is only mentioned superficially [59].

One guidance document about the clinical evaluation of MDs issued by the HSA describes the BRA as an ongoing process that includes any risks associated with using the MD without mentioning cybersecurity directly [38]. Another HSA guidance document about cybersecurity recommends that the framework set out in the International Organization for Standardization (ISO) standard 14971 for cybersecurity risk management be followed, which includes the BRA as an integral step of this process [43].

The MHRA guidance on cybersecurity acknowledges the relevance of cybersecurity and provides a road map for future developments without defining cybersecurity requirements or mentioning the BRA in its current state [45].

A TGA guidance document describes the implementation of the 15 "Essential Principles" of the AU_MDR, containing recommendations for both the BRA and cybersecurity [46]. The BRA is described as a holistic approach that assures that the benefits from the use of an MD outweigh any undesirable effects while providing no operational method [46]. The cybersecurity requirements, which are described at a high level in this guidance, are further detailed in 2 additional guidance documents [21,41]. These recommend a continuous risk assessment and management over the MD's life cycle that includes cybersecurity risks and the consideration of cybersecurity risks in the BRA without providing a detailed methodology or description of the impact of cybersecurity on the BRA [21,41].

The MFDS guidance defines high-level requirements for the cybersecurity risk management process without mentioning the BRA [44].

In total, 9% (3/34) of the guidelines were issued by the IMDRF, a joint group of multiple regulatory authorities, including the European Commission and the FDA. While the provided documents are not legally binding, they provide a good overview of regulatory practices and are acknowledged as part of the overall regulatory state of the art by the FDA as well as by EU regulatory bodies [13,22]. These 3 guidelines describe cybersecurity recommendations for MDs, acknowledging the need to conduct a BRA that considers cybersecurity risks [42,47,48].

A total of 6% (2/34) of the guidelines were issued by the National Institute of Standards and Technology (NIST) but are referenced in multiple guidance documents originating from sources outside the United States [21,42,59]. The NIST_800-30 provides guidance on the implementation of a risk management process, describing different methods while addressing the concept of the BRA only slightly [35]. Similarly, the NIST Cybersecurity Framework provides guidance for cybersecurity risk management with a focus on organizations but does not describe the BRA at all [36].

XSL•FO
RenderX

Table 2. The consideration of cybersecurity risks in the benefit-risk analysis (BRA) as defined in different regulations, guidelines, and standards.

Name	Mention of BRA	Mention of cybersecurity	Provides a method or example	Intersection between BRA and cybersecurity
US_FD&C[a]	Yes	Yes	No	The FD&C Act addresses the BRA and cybersecurity independently. However, the broadly defined risks in the BRA include safety risks that could be caused by cybersecurity vulnerabilities, ensuring comprehensive risk management without explicitly mentioning cybersecurity.
CA_MDR[b]	Yes	No	No	The CA_MDR requires conducting a BRA to determine whether a product is safe and effective. It acknowledges the uncertainty that lies in this process but provides no further detail on methods. It does not mention cybersecurity.
AU_MDR[c]	Yes	Yes	No	The AU_MDR mentions cybersecurity requirements and the BRA independently in its "Essential Principles." Principle 6 defines that the benefits of an MD[d] must outweigh any undesirable effects arising from its use. Guidance documents by the TGA[e] clarify that cybersecurity is within the scope of this principle. No detailed method is provided. In principle 12, the document defines requirements for the cybersecurity of MDs, including the implementation of a risk management process.
EU_MDR[f]	Yes	Yes	No	The EU_MDR addresses the BRA and cybersecurity independently. However, the broadly defined risks in the BRA include safety risks that could be caused by cybersecurity vulnerabilities, ensuring comprehensive risk management without explicitly mentioning cybersecurity.
NIST_800-30	Yes	Yes	No	The document provides guidance on the cybersecurity risk assessment process and describes the benefits and problems of qualitative, quantitative, and semiquantitative methods, acknowledging the limitations of personal judgment and uncertainties of the process. Due to the non–MD-specific nature of the document, the concept of a BRA is only slightly addressed, whereas the main focus is on cost and benefit trade-offs.
NIST_CSF2.0[h]	No	Yes	No	The document provides a framework for cybersecurity risk management with a focus on organizations. The BRA is not mentioned.
FDA_BRA_AC&ED[i]	Yes	No	Yes (not specifically designed for cybersecurity)	This document supplements the FDA's[j] BRA framework by incorporating additional benefits and risks. It specifically describes the impact of availability, compliance, and enforcement decisions on the BRA.
FDA_Cyber_Post[k]	Yes	Yes	No	This guidance document advises manufacturers to monitor, identify, and address cybersecurity vulnerabilities and exploits as part of their postmarket management of MDs. It also recommends evaluating residual risks, the results of the BRA, and any risks introduced through remediation efforts as part of the Postmarket Cybersecurity Program after detecting a vulnerability.
HSA_GN-20[l]	Yes	No	No	This guidance defines an important objective of clinical evaluation as determining whether the risks associated with the use of the MD are acceptable when weighed against the benefits for the patient. This BRA is also seen as an ongoing process that includes any risks associated with the use of the MD without mentioning cybersecurity directly.
FDA_BRA_Uncertainty[m]	Yes	No	No	This guidance document addresses the inherent uncertainty in MDs' premarket decision-making regarding their benefits and risks. It outlines the need to consider this uncertainty in the BRA and recommends collecting postmarket data to address and reduce this uncertainty over an MD's life cycle.

XSL•FO
RenderX

Name	Mention of BRA	Mention of cybersecurity	Provides a method or example	Intersection between BRA and cybersecurity
FDA_BRA[n]	Yes	No	Yes (not specifically designed for cybersecurity)	This guidance document states that the BRA can be conducted using both clinical and nonclinical data, recognizing that some MDs' attributes cannot be tested using clinical methods. It emphasizes that clinical benefits are usually measured directly through factors such as magnitude, probability, and duration. In the BRA, all risks must be considered and weighed against the clinical benefits, both based on the totality of the evidence. The document also highlights the importance of incorporating patient perspectives and accounting for uncertainty. It provides an assessment tool similar to a checklist and examples and notes that the BRA is an essential step that should always be conducted following the risk assessment and management processes that are outlined in ISO[o] standard 14971:2019.
MDCG_2019-16[p]	Yes	Yes	Yes	This guidance document emphasizes the need to consider the relationship between safety and security in the context of risk. It highlights that security risks can be caused by both weak and restrictive security measures. Rather than conducting a BRA for each individual security risk, an overall BRA should be executed based on the device's intended use and the potential impacts on safety and performance using the safety risk assessment, which includes security hazard categories. In addition, the original security BRA should be updated with PMS[q] data.
HC_Cyber[r]	Yes	Yes	No	In this document, cybersecurity is seen as an integral component of an MD over its complete life cycle. The overlap between security and safety is acknowledged, and it is recommended to incorporate cybersecurity into the risk management process. The relevance of the BRA is only mentioned superficially in the introduction without providing further details on the impact of cybersecurity risks.
TGA_Cyber_Dev[s]	Yes	Yes	No	This document recommends a continuous risk assessment and management over the MD's life cycle that includes cybersecurity risks. Devices should be secure by design and by default to minimize risks to patients. The overlap of security and safety is acknowledged. It is stated that, to comply with principle 6 of the Therapeutic Goods (Medical Devices) Regulations 2002, cybersecurity risks must be considered in the BRA. The document does not provide a detailed methodology or description of the impact of cybersecurity on the BRA. Instead, it offers general recommendations, such as maintaining an SBOM[t], and suggests that new features could potentially increase cybersecurity risks but that they should also provide benefits.
TGA_Cyber_User[u]	Yes	Yes	No	This document emphasizes the role of users to maintain security in MDs. The BRA is mentioned in the context of communication activities, which should include discussions on the benefits of a device versus its cybersecurity risks.
IMDRF_Cyber[v]	Yes	Yes	No	This guidance document emphasizes the need to assess the impact of security risk mitigation measures on the management of other risks (eg, considering the benefits and risks associated with deploying updates). HCPs[w] are encouraged to conduct a BRA of manufacturers' proposed mitigations before implementing them. According to this document, a cybersecurity-informed BRA is an ongoing process.
HSA_Cyber[x]	Yes	Yes	No	This document recommends the implementation of an ongoing risk management process following ISO standard 14971, which also takes cybersecurity risks into account. The overlap of security and safety risks is acknowledged. The BRA is mentioned in the context of clinical evaluation and new risks arising from the use of a device without specifying a method or describing the impact of cybersecurity risks on the BRA.

XSL•FO
RenderX

Name	Mention of BRA	Mention of cybersecurity	Provides a method or example	Intersection between BRA and cybersecurity
MFDS_Cyber[y]	No	Yes	No	This document acknowledges an overlap of security and safety risks and defines high-level requirements for the cybersecurity risk management process. The concept of a BRA is not mentioned.
MHRA_SaMD[z]	No	Yes	No	The relevance of cybersecurity risks and a gap in the current regulation are recognized. A road map for future developments is presented but without providing details on risk management in the area of cybersecurity or on the BRA.
TGA_Safety[aa]	Yes	Yes	No	In line with principle 6 of the Therapeutic Goods (Medical Devices) Regulations 2002, this guidance emphasizes the relevance of the BRA. The BRA is described as a holistic method that assures that the benefits from the use of an MD outweigh any undesirable effects. A process for risk assessment and risk management must be implemented throughout the life cycle of the product. The impact of cybersecurity risks on the BRA is not described in detail, and no method is given for how a BRA must be performed.
FDA_Cyber_Pre[ab]	Yes	Yes	No	This guidance document emphasizes the unpredictable nature of cybersecurity risks, which cannot be easily assessed or quantified using historical data or modeling. Thus, qualitative methods could also be used, which is similarly acknowledged in the most recent version of ISO standard 14971:2019. This document also acknowledges the interconnected yet distinct nature of safety and security risk management. Manufacturers should assess identified risks based on the level of risk posed by the device and its operational system, with continuous risk identification throughout the device life cycle. Cybersecurity should be integrated into the device from the beginning. Responses to security events should consider the BRA to determine the appropriateness of actions given that updates may limit device availability. In addition, this document recognizes the challenges in updating devices already on the market.
IMDRF_SBOM[ac]	Yes	Yes	No	This guidance document highlights that providing an SBOM enables regulators to perform a more comprehensive BRA. It aids in estimating and addressing the impact of threats, vulnerabilities, and exploits, thereby enhancing overall risk management.
IMDRF_LegMD[ad]	Yes	Yes	No	This guidance document advises HCPs and MD manufacturers to monitor the risk profile of devices throughout their life cycle. HCPs should perform regular clinical BRAs comparing the risks of the use of legacy devices beyond their EOS[ae] date with acquiring new or upgraded devices.
IEC_82304-1 standard[af]	No	Yes	No	The standard defines that a risk management process should be in place to fulfill its risk-benefit approach. The risk assessment should also include the network environment. The impact of cybersecurity risks on risk management is only described at a high level without providing details on the BRA.
ISO_14971 standard[ag]	Yes	Yes	Yes (not specifically designed for cybersecurity)	ISO 14971:2019 acknowledges that quantitative data are often unavailable for cybersecurity risks, allowing for qualitative risk estimation. The standard specifically includes risks related to data and system security within its scope. It highlights that breaches in data and system security can lead to harm, such as loss of data, uncontrolled access to data, corruption or loss of diagnostic information, or software corruption resulting in device malfunction. If a residual risk is deemed unacceptable based on the risk management plan's criteria and further risk control is impractical, manufacturers may review data and literature to determine whether the benefits of the intended use outweigh this residual risk. In addition, a general BRA of all residual risks should be conducted by weighing them against the overall benefits provided by the MD's intended use. This overall residual risk could be stronger than each individual risk.

XSL•FO
RenderX

Name	Mention of BRA	Mention of cybersecurity	Provides a method or example	Intersection between BRA and cybersecurity
IEC_80001-1 standard[ah]	Yes	Yes	Yes	This standard specifies that, if risks cannot be reduced to an acceptable level through risk management, a BRA should be conducted. When no suitable risk control measures are possible, a holistic BRA should be performed, weighing the overall residual risk against the system benefit. The analysis should involve judgment by experienced and knowledgeable individuals considering technical, clinical, regulatory, economic, sociological, and political contexts. It also recommends conducting a general BRA weighing all unaccepted risks against the net clinical benefit of deployment, aligning with the guidance in ISO TR[ai] 24971:2020.
IEC_81001-5-1 standard[aj]	No	Yes	No	This standard mandates that handling residual security risks should be done in cooperation with product risk management. It also requires considering the impacts on safety caused by the degradation of security over time.
AAMI_SW96 standard[ak]	Yes	Yes	Yes	If a security residual risk is deemed unacceptable, a BRA should be performed. Manufacturers should balance the residual security risk against the benefits provided by the device's design capabilities or security controls. The overall BRA should consider all security residual risks as well as the impact of the implementation of the device into the IT infrastructure, intersecting with the requirements defined in IEC 80001-1:2021.
IEC_80002-1[al]	Yes	Yes	No	This document has a high overlap with ISO 14971 as it describes the application of ISO 14971 for software. The BRA is described as an integral part of the risk management process for all residual risks. Cybersecurity is mentioned as a relevant risk, whereas further details of its impact on the BRA are not provided.
IEC_80001-2-2	No	Yes	No	This technical report emphasizes that residual cybersecurity risks should be considered in the BRA.
IEC_80001-2-1	Yes	Yes	No	This technical report specifies that, if risks cannot be reduced to an acceptable level through the risk management process, the device should be changed or the risk should be outweighed by the expected benefits, determined in a BRA.
ISO_80001-2-7	Yes	Yes	No	This technical report specifies that individual and overall residual risks should be assessed for acceptability. If the required risk reduction is impractical, the responsible organization must conduct and document a BRA of the residual risk. In addition, when the residual risk remains unacceptable, a BRA should be conducted to weigh the overall residual risk against the health benefits emerging from incorporating the MD into the IT network.
AAMI_TIR57[am]	Yes	Yes	Yes	This technical report advises that, if a security residual risk is deemed unacceptable, a BRA should be performed. Manufacturers should balance the residual security risk against the benefits provided by the device's design capabilities or security controls. It warns against using "security by obscurity" as a risk reduction strategy. Security risks impacting safety must be evaluated in the safety risk assessment following ISO 14971:2019. Effective communication with patients and HCPs is crucial, ensuring that they understand how to manage residual risk without providing attackers with a blueprint. Manufacturers should also balance usability, device safety, and security to ensure appropriate security controls for intended users and connected systems.

XSL•FO
RenderX

Name	Mention of BRA	Mention of cybersecurity	Provides a method or example	Intersection between BRA and cybersecurity
ISO_24971[an]	Yes	Yes	Yes (not specifically designed for cybersecurity)	This technical report provides examples of cybersecurity-related safety risks and emphasizes the importance of evaluating security risks by considering confidentiality, integrity, and availability in the context of the device's intended use. It states that the BRA is used to determine whether an individual residual risk is outweighed by the expected benefits of the device's intended use. Manufacturers are encouraged to consider technical, regulatory, economic, and sociological contexts in their risk management decisions, acknowledging that implementing risk control measures might introduce new risks or exacerbate existing ones. The report emphasizes the complexity of directly comparing benefits and risks and suggests that the overall residual risk should be viewed from a broad perspective, ensuring that all identified hazardous situations have been evaluated and risks have been reduced to an acceptable level or accepted based on a BRA.

[a]FD&C: Federal Food, Drug, and Cosmetic Act.

[b]CA_MDR: Canadian Medical Device Regulations.

[c]AU_MDR: Australian Therapeutic Goods (Medical Devices) Regulations.

[d]MD: medical device.

[e]TGA: Therapeutic Goods Administration.

[f]EU_MDR: European Union Medical Device Regulation.

[g]NIST_800-30: National Institute of Standards and Technology Special Publication 800-30.

[h]NIST_CSF2.0: National Institute of Standards and Technology Cybersecurity Framework.

[i]FDA_BRA_AC&ED: Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions.

[j]FDA: Food and Drug Administration.

[k]FDA_Cyber_Post: Postmarket Management of Cybersecurity in Medical Devices.

[l]HSA_GN-20: GN-20: Guidance on Clinical Evaluation.

[m]FDA_BRA_Uncertainty: Consideration of Uncertainty in Making Benefit-Risk Determinations in Medical Device Premarket Approvals, De Novo Classifications, and Humanitarian Device Exemptions.

[n]FDA_BRA: Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and De Novo Classifications.

[o]ISO: International Organization for Standardization.

[p]MDCG_2019-16: Medical Device Coordination Group Document 2019-16–Guidance on Cybersecurity for medical devices.

[q]PMS: postmarket surveillance.

[r]HC_Cyber: Pre-market Requirements for Medical Device Cybersecurity.

[s]TGA_Cyber_Dev: Medical device cybersecurity guidance for industry.

[t]SBOM: software bill of materials.

[u]TGA_Cyber_User: Medical device cybersecurity information for users.

[v]IMDRF_Cyber: Principles and Practices for Medical Device Cybersecurity.

[w]HCP: health care provider.

[x]HSA_Cyber: Regulatory Guidelines for Software Medical Devices—A Life Cycle Approach.

[y]MFDS_Cyber: Guideline on Review and Approval for Cybersecurity of Medical Devices (for industry).

[z]MHRA_SaMD: Software and AI as a Medical Device Change Programme.

[aa]TGA_Safety: Complying with the Essential Principles on the safety and performance of medical devices.

[ab]FDA_Cyber_Pre: Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions.

[ac]IMDRF_SBOM: Principles and Practices for Software Bill of Materials for Medical Device Cybersecurity.

[ad]IMDRF_LegMD: Principles and Practices for the Cybersecurity of Legacy Medical Devices.

[ae]EOS: end of service.

[af]IEC_82304-1 standard: International Electrotechnical Commission Standard 82304-1:2016 Health software Part 1: General requirements for product safety.

[ag]ISO_14971 standard: International Organization for Standardization Standard 14971:2019 Medical devices—Application of risk management to medical devices.

[ah]IEC_80001-1 standard: International Electrotechnical Commission Standard IEC 80001-1:2021 Application of risk management for IT-networks incorporating medical devices; Part 1: Safety, effectiveness and security in the implementation and use of connected medical devices or connected health software.

[ai]TR: Technical Report.

XSL•FO
RenderX

[aj]IEC_81001-5-1 standard: International Electrotechnical Commission Standard 81001-5-1:2021 Health software and health IT systems safety, effectiveness and security; Part 5-1: Security — Activities in the product life cycle.

[ak]AAMI_SW96 standard: American National Standards Institute and Association for the Advancement of Medical Instrumentation Standard SW96:2023; Standard for medical device security—Security risk management for device manufacturers.

[al]IEC_80002-1: International Electrotechnical Commission Standard 80002-1:2009 Medical device software; Part 1: Guidance on the application of ISO 14971 to medical device software.

[am]AAMI_TIR57: Association for the Advancement of Medical Instrumentation Technical Report TIR57:2016/(R)2023; Principles for medical device security—Risk management.

[an]ISO_24971: International Organization for Standardization Technical Report 24971:2020 Medical devices — Guidance on the application of ISO 14971.

Standards for the BRA and Cybersecurity

Among the 34 documents, 5 (15%) were standards and 6 (18%) were technical reports. All of them describe cybersecurity considerations for the BRA process. The ISO 14971:2019 standard explicitly mentions that security risks should be considered in the risk assessment and the BRA [11] as they can lead to harm [12]. This standard is the recognized consensus standard of the FDA for the risk management of MDs and was harmonized for the EU as the EN ISO 14971:2019 standard [60]. It also lists basic requirements for the BRA, mandating an evaluation of whether a device's benefits outweigh its residual risks and necessary modifications if the benefits do not justify the risks [11]. The ISO Technical Report 24971:2020 standard, a technical report on how to implement ISO 14971:2019, provides examples and detailed guidance on which risks to consider within the security risk management process. In line with ISO 14971:2019, it proposes that the overall residual risk should be viewed from a broad perspective considering all residual risks, ensuring that all identified hazardous situations have been evaluated and the risks have been reduced to an acceptable level or accepted based on a BRA [12].

The International Electrotechnical Commission (IEC) 80001-1:2021 standard and the associated technical reports describe that, if single individual risks cannot be reduced to an acceptable level through mitigation measures in the risk management process, a BRA should be conducted for each of those risks individually [50]. In addition, when residual risks remain, the aggregated residual risk should be weighed against the benefits emerging from the deployment of an MD [50]. The IEC 81001-5-1:2021 standard mandates that handling residual security risks should be done in cooperation with product risk management without mentioning a BRA explicitly [51].

The Association for the Advancement of Medical Instrumentation (AAMI) TIR57 standard is referred to in multiple cybersecurity-related FDA guidance documents and explains in detail how information security risk management for MDs should be performed and how the identified risks should be considered in the BRA [57]. It states that, if a security-related residual risk is assessed as unacceptable, a BRA should be performed. Manufacturers should balance the residual risk against the benefits provided by the device's design capabilities or security controls. This report also provides an example of an unacceptable cybersecurity-related residual risk and how this affects the BRA outcome: an unacceptable risk of exposing personally identifiable information should not be justified solely because the device provides life-saving therapy. Instead, the benefit of storing the information on the device should be weighed against the risk of compromising confidentiality [57].

The FDA-recognized AAMI SW96 standard applies the risk management framework provided in the ISO 14971:2019 standard to security risks and defines requirements similar to those of the AAMI TIR57 standard [52]. In addition, it provides an example of how a security benefit could outweigh a security risk and adds an infrastructure view to an overall security BRA: not only the device itself must be considered but also the impact of the implementation of the device into the IT infrastructure, intersecting with the requirements defined in the IEC 80001-1:2021 standard [52].

The IEC 80002-1:2009 standard describes the application of ISO standard 14971 for MD software and acknowledges the relevance of cybersecurity risks without adding information to the framework set out in ISO standard 14791 [53].

The IEC 82304-1:2016 standard describes the implementation of risk management that considers cybersecurity risks as relevant for its benefit-risk approach. Further details on the BRA are not provided [49].

Recommendations and Best Practices

Within the identified documents, several recommendations and best practices on cybersecurity risks in a BRA for manufacturers were identified. Table 3 provides an overview of those recommendations, including the provision of an SBOM [19,21,43,47]; recommendations for a BRA that considers the overall residual risks, including cybersecurity-related ones in the context of clinical benefits [11-13,21,37,39,40,46], and is conducted on a regular basis as part of postmarket surveillance (PMS) activities [13,21,32,38,43,51,59]; recommendations for considering the ISO 14971 framework [21,22,43,52,53,57]; and general development requirements to minimize the risk to patients [13,18,20-22].

XSL•FO
RenderX

Table 3. Recommendations on how to conduct a benefit-risk analysis (BRA) that considers cybersecurity risks. In addition to the recommendation and its description, the phase of the medical device (MD) life cycle in which a recommendation is relevant for the BRA is provided.

Recommendation	Document it was included in	Description	Phase of the MD life cycle
Prepare an SBOM[a].	IMDRF_SBOM[b], HSA_Cyber[c], TGA_Cyber_Dev[d], and US_FD&C[e]	Manufacturers should provide an SBOM to facilitate the BRA for regulators. The SBOM should be updated on a regular basis.	Premarket and postmarket phases
A BRA should always be conducted.	TGA_Safety[f], AU_MDR[g], MDCG_2019-16[h], and FDA_BRA[i]	The BRA should always be conducted as part of the overall risk management process and not only if unacceptable residual risks remain as a combination of risks could have a higher impact than each individual risk.	Premarket phase
An overall BRA should be conducted.	MDCG_2019-16, FDA_BRA, FDA_BRA_Uncertainty[j], FDA_BRA_AC&ED[k], ISO_14971[l], ISO_24971[m], TGA_Safety, and TGA_Cyber_Dev	Instead of conducting a BRA for each cybersecurity risk individually, an overall BRA should be conducted considering the overall residual risks, including cybersecurity-related ones, in the context of clinical benefits.	Premarket and postmarket phases
The entire IT infrastructure should be considered.	TGA_Cyber_Dev, IEC_82304-1, and MDCG_2019-16	In the BRA, not only the cybersecurity risks for the individual device must be considered but also the impact of the implementation of the device into the IT infrastructure as the device might introduce new vulnerabilities that influence all other devices in the network.	Premarket and postmarket phases
Effective communication of residual risks	AAMI_TIR57[o], TGA_Cyber_User[p], and IMDRF_Cyber[q]	If residual risks remain, manufacturers should maintain good communication with patients and health care providers, describing those risks without delivering a blueprint for an attacker.	Postmarket phase
Security event responses and updates should be guided by a BRA.	FDA_BRA_AC&ED, FDA_Cyber_Pre[r], FDA_Cyber_Post[s], and IMDRF_Cyber	The response to security events and the deployment of updates should be guided by a BRA as both could cause new risks for patients.	Postmarket phase
The cybersecurity BRA should be part of postmarket surveillance activities.	HC_Cyber[t], HSA_Cyber, HSA_GN-20[u], TGA_Cyber_Dev, FDA_Cyber_Post, MDCG_2019-16, and IEC_81001-5-1	As previously unknown vulnerabilities could emerge in cMDs[v] that can change the benefit-risk ratio, the BRA must be conducted regularly as part of postmarket surveillance activities. For this, a continuous threat analysis and response plan is required.	Postmarket phase
A device's EOS[w] should be considered.	TGA_Cyber_Dev, IMDRF_LegMD[x], and HSA_Cyber	For devices after the EOS, the BRA conducted by the health care provider should consider the benefits and risks of using old devices versus acquiring new ones.	Postmarket phase
The framework set out in ISO[y] standard 14971 should be considered.	HSA_Cyber, TGA_Cyber_Dev, IEC_80002-1, AAMI_TIR57, AAMI_SW96, and FDA_Cyber_Pre	ISO standard 14971 provides a standardized and recognized framework for risk management, of which the BRA is a part. ISO standards 14971 and 24971 state that the framework could also be used for cybersecurity risks.	Premarket and postmarket phases
Devices should be developed to minimize risks.	TGA_Cybyer_Dev, MDCG_2019-16, AU_MDR, EU_MDR[z], and FDA_Cyber_Pre	To minimize the risk to patients, devices should follow the state of the art; be secure by design; and be able to respond to future, as yet unknown threats.	Premarket phase

[a]SBOM: software bill of materials.

[b]IMDRF_SBOM: Principles and Practices for Software Bill of Materials for Medical Device Cybersecurity.

[c]HSA_Cyber: Regulatory Guidelines for Software Medical Devices—A Life Cycle Approach.

[d]TGA_Cyber_Dev: Medical device cybersecurity guidance for industry.

[e]FD&C: Federal Food, Drug, and Cosmetic Act.

[f]TGA_Safety: Complying with the Essential Principles on the safety and performance of medical devices.

[g]AU_MDR: Australian Therapeutic Goods (Medical Devices) Regulations.

[h]MDCG_2019-16: Medical Device Coordination Group Document 2019-16 – Guidance on Cybersecurity for medical devices.

[i]FDA_BRA: Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and De Novo Classifications.

[j]FDA_BRA_Uncertainty: Consideration of Uncertainty in Making Benefit-Risk Determinations in Medical Device Premarket Approvals, De Novo Classifications, and Humanitarian Device Exemptions.

[k]FDA_BRA_AC&ED: Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions.

XSL•FO
RenderX

[l]ISO_14971: International Organization for Standardization Standard 14971:2019 Medical devices—Application of risk management to medical devices.

[m]ISO_24971: International Organization for Standardization Technical Report 24971:2020 Medical devices — Guidance on the application of ISO 14971.

[n]IEC_82304-1: International Electrotechnical Commission Standard 82304-1:2016 Health software Part 1: General requirements for product safety.

[o]AAMI_TIR57: Association for the Advancement of Medical Instrumentation Technical Report TIR57:2016/(R)2023; Principles for medical device security—Risk management.

[p]TGA_Cyber_User: Medical device cybersecurity information for users.

[q]IMDRF_Cyber: Principles and Practices for Medical Device Cybersecurity.

[r]FDA_Cyber_Pre: Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions.

[s]FDA_Cyber_Post: Postmarket Management of Cybersecurity in Medical Devices.

[t]HC_Cyber: Pre-market Requirements for Medical Device Cybersecurity.

[u]HSA_GN-20: GN-20: Guidance on Clinical Evaluation.

[v]cMD: connected medical device.

[w]EOS: end of service.

[x]IMDRF_LegMD: Principles and Practices for the Cybersecurity of Legacy Medical Devices.

[y]ISO: International Organization for Standardization.

[z]EU_MDR: European Union Medical Device Regulation.

Discussion

Principal Findings

The goal of this scoping review was to identify the current legal frameworks, guidelines, and standards in the United States and EU on how cybersecurity should be considered in the BRA of MDs. Among the 34 documents, we identified 4 (12%) regulations (the EU_MDR, AU_MDR, Canadian Medical Device Regulations, and FD&C Act); 5 (15%) standards issued by the ISO, IEC, and AAMI; 6 (18%) technical documents; and 19 (56%) guidelines issued by the FDA, MDCG, Health Canada, HSA, NIST, MHRA, MFDS, and IMDRF.

The regulations in the United States, Australia, Canada, and the EU provide high-level requirements for the BRA but do not explicitly address the intersection of cybersecurity and BRA [18-20,23]. At the guideline level, MDCG 2019-16 provides an overview of cybersecurity requirements for the EU, acknowledges the relevance of cybersecurity considerations in the BRA, and loosely defines how cybersecurity risks should influence the BRA [13]. Other guidelines (12/34, 35%) underscore the relevance of the BRA without detailing methods for the consideration of cybersecurity risks [21,22,32,37-40,42,43,47,48,59].

Multiple standards and technical reports describing the intersection of cybersecurity and the BRA (3/34, 9%) mandate the execution of a BRA only when residual risks, including cybersecurity-related ones, cannot be reduced to an acceptable level through risk management [11,50,57]. Some of these documents (3/34, 9%) provide methods and examples explicitly for a security BRA [50,52,57], whereas others (2/34, 6%) only loosely describe methods that could be adapted to consider cybersecurity risks without providing more details [11,12].

BRA and Cybersecurity

In our scoping review, we identified a notable gap between the acknowledgment of the relevance of cybersecurity risks to the BRA and the actual guidance or methods on how to consider these risks within it. Thus, and through the lack of concrete examples, there remains some amount of uncertainty within those guidelines on how and when to conduct a BRA in general and how to conduct a security BRA.

Many of the included documents (9/34, 26%) consider BRA an explicit requirement, necessitating the description and coverage of all safety risks, including those caused by cybersecurity threats [11-13,18,19,21,37,40,46]. Some of the included documents (8/34, 24%) provide sophisticated overviews of cybersecurity risks and the security risk management process, but they do not go into detail regarding the BRA [35,36,44,49,51,59] or refer to generic methods [50,53]. Others (5/34, 15%) provide methods that could be adapted to consider cybersecurity while not containing specific information about the relationship between cybersecurity risks and the BRA [11,12,37,39,40].

Of the 34 documents, 3 (9%) guidance documents separate security and safety risk management into 2 processes [13,52,57]. While the intersection of both processes is acknowledged, they include separate BRAs: one for safety risk management, which considers all residual risks, including security-related ones, comparing them to the clinical benefits to patients and to the health system, and another one for security alone [13,52,57], where manufacturers should appropriately balance the residual security risk against the benefit gained by the design capabilities or security controls of the device [57]. While a lot of guidance exists for the identification of security risks, it remains unclear how security benefits are defined and how they influence the security BRA as well as the safety BRA. The SW96:2023 standard provides a brief hypothetical example, where the benefit of accurate patient identification outweighs the residual risk of storing sensitive patient data on a device [52]. However, this approach somewhat contradicts the understanding of a BRA described in other documents, which mandate a more holistic approach weighing the clinical benefits of the device against the residual cybersecurity risks and conducting an additional general BRA that considers all residual risks [11,12,37,39,40,50].

Some standards and official guidance documents (10/34, 29%) differ in their description of when a BRA must be carried out. One of them only calls for a BRA if residual risks remain

XSL•FO
RenderX

unacceptable [51], which contradicts the requirements of the MDCG, TGA, and FDA, which see the BRA as a final step in deciding whether an MD should be put on the market [13,18,19,21,37,39-41,46]. A middle-ground position is provided by multiple documents (5/34, 15%) that require a separate BRA for any individual residual risk that is not judged as acceptable, whereas a BRA should be conducted to evaluate the overall residual risk, which should consider cybersecurity-related residual risks [11,12,50,52,57].

BRA as Part of the PMS

Traditionally, BRA is often seen as primarily a premarket activity in the MD development and approval process. However, multiple guidelines (6/34, 18%) clarify that BRA is an ongoing process [13,22,32,37,40,48], and especially when initial data are limited, PMS should be used to further define the risk profile of a device and update the BRA accordingly [37]. This is particularly important for cybersecurity risks as these are constantly changing, for example, when previously unknown vulnerabilities in software or communication protocols are discovered and exploited by malicious actors. Therefore, cybersecurity risk monitoring and management is seen as an integral part of PMS by many authorities [13,21,32]. Maintaining an SBOM is seen as supportive for this effort [21,47]. On the basis of newly identified risks, and guided by the BRA as part of risk management, mitigation actions (eg, through patching or device recalls) could be necessary [13,32,42]. In addition, a BRA should be conducted by the health care provider and the MD manufacturer before deploying updates and after the end of service of devices [22,42,48].

Implementation Challenges

In the included documents, multiple issues were identified that bring to light particular technical challenges to the process of conducting a BRA. First, cybersecurity risks could be difficult to assess, especially in a quantitative manner, due to the unpredictable nature of vulnerabilities [11,22,32,51]. Those vulnerabilities appear over time, for example, through proactive attacks targeting specific aspects of devices or, more commonly, through bugs in standard software libraries or components [61] and zero-day exploits (unknown vulnerabilities that can be exploited before they are publicly known and mitigated [62]). The ISO Technical Report 24971:2020 standard acknowledges the existence of such hard-to-quantify risks, which could also be connected to other new technologies such as artificial intelligence (AI) [63-65], gamification [66], or virtual reality [67,68], but considers them still as relevant for the BRA [12]. Second, updating devices could also lead to harm, primarily regarding the availability of products during the update phase, which needs to be considered in the BRA [22,32,37,42]. Third, there is the need to balance security measures as both too weak and overly restrictive security measures could pose risks [13]. Weak measures do not provide adequate protection (eg, for personal data), whereas overly restrictive measures hinder device usability. Fourth, implementing a risk control measure to reduce one risk can introduce new risks, for example, when adding a higher authentication standard (such as 2-factor authentication) to address the risk of the disclosure of patient data could limit the accessibility of the device in case of an emergency [12]. Finally, for cMDs, the entire IT infrastructure needs to be considered when assessing cybersecurity risks [50,52] (ie, with a network-level risk assessment) as cMDs usually exchange data with multiple other devices and services, some of which are MDs whereas some are not. In addition, the implementation of a new device into an existing infrastructure might introduce vulnerabilities for the entire system.

Future Research

Innovative approaches for the assessment and management of cybersecurity risks could help overcome these challenges and improve the understanding of the evolving cybersecurity risk landscape of MDs. This includes the use of risk management ontologies to help developers identify existing vulnerabilities in the risk assessment phase [69] and the use of AI-based intrusion detection systems for postmarket risk management, which use machine and deep learning approaches to detect suspicious network behavior, untypical user patterns, malware, and other malicious activities [70,71]. The pre- and postmarket risk assessment could be supported by automatic systems linked to vulnerability databases, which continuously assess a given infrastructure, alert in case of newly defined risks, and propose mitigation measures [72], and AI-based systems that assess and predict vulnerabilities [73,74]. Another approach relevant to cMDs is device monitoring, a standard in industries other than health care, which could help provide an overview of the overall attack surface and vulnerable devices [75]. While some of these approaches are already used in non–health care environments (eg, AI-based intrusion detection systems or device management), others are still in early development phases (eg, automatic risk management and mitigation systems). However, these innovative approaches often do not cover all aspects relevant to the BRA. Thus, further development of methods for the assessment of cybersecurity risks and the incorporation of these risks into the BRA process is necessary. While current methods were often not developed with cybersecurity risks in mind, some, especially qualitative frameworks such as the FDA's BRF, might be capable of considering cybersecurity risks as they are more flexible [25]. A quantitative framework frequently used for the BRA, the MCDA, is also used for cybersecurity decision-making and evaluation in health care [76,77]. In particular, the MCDA method "Technique for Order Preference by Similarity to Ideal Solution" could be adopted for cybersecurity considerations in the BRA, for example, to define best practices with a favorable benefit-risk profile for authentication, device monitoring, or network architecture, which could take a broad set of criteria, including patient factors, into account. In addition, models such as the Gordon-Loeb model are currently used for the economic analysis of benefits and costs of cybersecurity risk mitigations [78,79]. Future research should explore the feasibility of adapting these models to weigh cybersecurity risks against MD benefits, although this is beyond the scope of this review. The absence of standardized frameworks for BRA in general and for the evaluation of cybersecurity risks likely contributes to the current discrepancies among existing standards, technical documents, and guidance on these themes.

In addition, the interaction among clinical benefits, cybersecurity risks, safety, and security should be explored further. The

XSL•FO
RenderX

limitations of current guidance are demonstrated by the following example. A “traditional” predigital laser device for surgical interventions in the eye without any software running on the device and without any interfaces poses only safety risks. However, when adding a digital component (eg, for a cloud connection), new security and safety risks, partly related to cybersecurity, arise. At this time, it remains unclear how those new risks would impact the security BRA as well as the overall BRA.

Understanding how these elements influence each other will help develop more effective risk management strategies and a better understanding of what risks are tolerable in innovative MDs. As the technology develops, new attack vectors will emerge (eg, through quantum computing, AI, or previously undetected dormant faults). Research is needed to identify these potential threats and develop innovative mitigation strategies to counter and proactively approach them. This research should not focus on single devices but also consider entire infrastructures as MDs and non-MDs often exist in the same system.

Limitations

The results of this scoping review could be limited by several factors. First, our review focused exclusively on IMDRF member states that provide guidance in English (the EU, the United States, Canada, Australia, South Korea, Singapore, and the United Kingdom). This geographic limitation may reduce the applicability of our findings to other regions. Second, the analysis in this scoping review was limited to guidance documents and standards that were relevant to MDs. This could have led to relevant principles described in standards from other areas not being considered. This has no limitation on the validity of the review of the current state of the art for MD cybersecurity risk assessment and BRA, but it could have limited the ability to suggest future directions for the further developments of standards.

Conclusions

Today, cybersecurity vulnerabilities in cMDs pose a growing risk to patients and health care providers. Therefore, the consideration of these risks in the BRA is essential to decide whether an MD is safe and secure and should be made available on the market. While the recognition of cybersecurity risks in cMDs has increased, there is a substantial gap between this acknowledgment and practical actions. Current regulations, guidelines, and standards mandate the consideration of cybersecurity risks but lack detailed methods for incorporating them into the BRA. To bridge this gap, it is essential for manufacturers and regulators to develop standardized frameworks that provide clear guidance on evaluating the impact of cybersecurity risks on the device’s safety and security, ultimately enhancing patient safety and device effectiveness.

Acknowledgments

This work was supported by the European Commission under the Horizon Europe program as part of the Enhanced cybersecurity for networked medical devices through optimisation of guidelines, standards, risk management and security by design (CYMEDSEC) project (grant 101094218). The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authorities can be held responsible for them. Responsibility for the information and views expressed herein lies entirely with the authors. This work was supported by the German Federal Ministry of Education and Research (BMBF) as part of the Zunkuftscluster SEMECO (03ZU1210BA). During the preparation of this work, the authors used DeepL (DeepL SE), Grammarly (Grammarly, Inc), and ChatGPT (in versions GPT-3.5, GPT-4, and GPT-4o; OpenAI, Inc) to improve the grammar, spelling, and readability of the manuscript. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Authors' Contributions

OF developed the design of the study. FJ, MO, and OF conducted the screening. OF conducted data extraction. FJ, MO, and OF conducted the analysis. OF wrote the first draft of the manuscript. FJ, OF, MO, CR, PW, and SG contributed to the writing, interpretation of the content, and editing of the manuscript, revising it critically for important intellectual content. All authors had final approval of the completed version and take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

SG is an advisory group member of the EY (Ernest and Young Global Limited)–coordinated Study on Regulatory Governance and Innovation in the field of Medical Devices conducted on behalf of the Directorate-General for Health and Food Safety of the European Commission. The authors declare the following competing interests. SG has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd, Flo Ltd, Thymia Ltd, FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, and Ada Health GmbH, and he holds share options in Ada Health GmbH. OF has worked as a freelance physician for CRS Clinical Research Services Berlin GmbH and has a leadership role and holds stock in WhalesDontFly GmbH. CR is an employee of Johner Institut GmbH and works and has worked as a consultant for multiple medical device manufacturers. PW works as a regulatory consultant at regulatory.me and declares no competing interests associated with this research. FJ and MO declare no competing interests.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist and a detailed table with URLs for the searched organizations and search results.
[PDF File (Adobe PDF File), 1907 KB - jmir_v26i1e65528_app1.pdf]

References

1. Machal ML. An overview about connected medical devices and their risks. Stud Health Technol Inform 2023 Jun 29;305:119-122. [doi: 10.3233/SHTI230438] [Medline: 37386972]
2. Huang C, Wang J, Wang S, Zhang Y. Internet of medical things: a systematic review. Neurocomputing 2023 Nov;557:126719. [doi: 10.1016/j.neucom.2023.126719]
3. Haghi Kashani M, Madanipour M, Nikravan M, Asghari P, Mahdipour E. A systematic review of IoT in healthcare: applications, techniques, and trends. J Netw Comput Appl 2021 Oct 15;192:103164. [doi: 10.1016/j.jnca.2021.103164]
4. Dwivedi R, Mehrotra D, Chandra S. Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: a systematic review. J Oral Biol Craniofac Res 2022 Mar;12(2):302-318 [FREE Full text] [doi: 10.1016/j.jobcr.2021.11.010] [Medline: 34926140]
5. Perakslis E, Ginsburg GS. Digital health-the need to assess benefits, risks, and value. JAMA 2021 Jan 12;325(2):127-128. [doi: 10.1001/jama.2020.22919] [Medline: 33369627]
6. Bauwens J, Ruckebusch P, Giannoulis S, Moerman I, Poorter ED. Over-the-air software updates in the internet of things: an overview of key principles. IEEE Commun Mag 2020 Feb;58(2):35-41. [doi: 10.1109/mcom.001.1900125]
7. Bracciale L, Loreti P, Bianchi G. Cybersecurity vulnerability analysis of medical devices purchased by national health services. Sci Rep 2023 Nov 09;13(1):19509 [FREE Full text] [doi: 10.1038/s41598-023-45927-1] [Medline: 37945583]
8. He Y, Aliyu A, Evans M, Luo C. Health care cybersecurity challenges and solutions under the climate of COVID-19: scoping review. J Med Internet Res 2021 Apr 20;23(4):e21747 [FREE Full text] [doi: 10.2196/21747] [Medline: 33764885]
9. Munoz Cornejo G, Lee J, Russell BA. A thematic analysis of ransomware incidents among United States hospitals, 2016–2022. Health Technol 2024 Jul 09;14(6):1059-1070. [doi: 10.1007/s12553-024-00890-3]
10. Ralston W. The untold story of a cyberattack, a hospital and a dying woman. Wired. 2020 Nov 11. URL: https://www.wired.com/story/ransomware-hospital-death-germany/ [accessed 2024-07-08]
11. ISO 14971:2019 medical devices — application of risk management to medical devices. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/72704.html [accessed 2024-07-03]
12. ISO/TR 24971:2020 medical devices — guidance on the application of ISO 14971. International Organization for Standardization. 2020. URL: https://www.iso.org/standard/74437.html [accessed 2024-07-10]
13. MDCG 2019-16 rev. 1 - Guidance on cybersecurity for medical devices. European Commission. 2020 Jul. URL: https://health.ec.europa.eu/document/download/b23b362f-8a56-434c-922a-5b3ca4d0a7a1_en?filename=md_cybersecurity_en.pdf [accessed 2024-07-03]
14. Neprash HT, McGlave CC, Cross DA, Virnig BA, Puskarich MA, Huling JD, et al. Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016-2021. JAMA Health Forum 2022 Dec 02;3(12):e224873 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.4873] [Medline: 36580326]
15. Manufacturer and user facility device experience (MAUDE) database. U.S. Food & Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm [accessed 2024-07-03]
16. Klonoff D, Han J. The first recall of a diabetes device because of cybersecurity risks. J Diabetes Sci Technol 2019 Sep 17;13(5):817-820 [FREE Full text] [doi: 10.1177/1932296819865655] [Medline: 31313589]
17. Medical device recalls. U.S. Food & Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfres/res.cfm [accessed 2024-07-03]
18. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (text with EEA relevance). European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32017R0745 [accessed 2024-07-03]
19. 21 U.S. Code chapter 9 - Federal Food, Drug, and Cosmetic Act. Legal Informative Institute. URL: https://www.law.cornell.edu/uscode/text/21/chapter-9 [accessed 2024-07-03]
20. Therapeutic goods (medical devices) regulations 2002. Australian Government. URL: https://www.legislation.gov.au/F2002B00237/latest/text [accessed 2024-10-02]
21. Medical device cyber security guidance for industry. Australian Government, Department of Health and Aged Care. 2022 Nov. URL: https://www.tga.gov.au/sites/default/files/medical-device-cyber-security-guidance-industry.pdf [accessed 2024-10-02]
22. Cybersecurity in medical devices: quality system considerations and content of premarket submissions. U.S. Food & Drug Administration. 2023 Sep 27. URL: https://www.fda.gov/media/119933/download [accessed 2024-07-03]
23. Medical devices regulations (SOR/98-282). Government of Canada. URL: https://laws-lois.justice.gc.ca/eng/regulations/sor-98-282/ [accessed 2024-10-02]

XSL•FO
RenderX

24. Su G, Deng D. Regulatory requirements and optimization of multiple criteria decision analysis to quantify the benefit-risk assessment of medical devices. Expert Rev Med Devices 2023 Apr;20(4):273-281. [doi: 10.1080/17434440.2023.2190021] [Medline: 36896851]
25. Freyer O, Jahed F, Ostermann M, Feig M, Gilbert S. Methodologies for the benefit-risk analysis of medical devices: a systematic review. Research Square Preprint posted online on August 5, 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-4832842/v1]
26. Mt-Isa S, Ouwens M, Robert V, Gebel M, Schacht A, Hirsch I. Structured benefit-risk assessment: a review of key publications and initiatives on frameworks and methodologies. Pharm Stat 2016 Jul 15;15(4):324-332 [FREE Full text] [doi: 10.1002/pst.1690] [Medline: 25981683]
27. Agapova M, Devine EB, Bresnahan BW, Higashi MK, Garrison LPJ. Applying quantitative benefit-risk analysis to aid regulatory decision making in diagnostic imaging: methods, challenges, and opportunities. Acad Radiol 2014 Sep;21(9):1138-1143. [doi: 10.1016/j.acra.2014.05.006] [Medline: 25107866]
28. Fu B, Li X, Scott J, He W. A new framework to address challenges in quantitative benefit-risk assessment for medical products. Contemp Clin Trials 2020 Aug;95:106073. [doi: 10.1016/j.cct.2020.106073] [Medline: 32622973]
29. Gebel M, Renz C, Rodriguez L, Simonetti A, Yang H, Edwards B, et al. A survey to assess the current status of structured benefit-risk assessment in the global drug and medical device industry. Ther Innov Regul Sci 2024 Jul 22;58(4):756-765. [doi: 10.1007/s43441-024-00650-5] [Medline: 38649524]
30. Sun S, Heske S, Mercadel M, Wimmer J. Predicting regulatory product approvals using a proposed quantitative version of FDA's benefit-risk framework to calculate net-benefit score and benefit-risk ratio. Ther Innov Regul Sci 2021 Jan 08;55(1):129-137 [FREE Full text] [doi: 10.1007/s43441-020-00197-1] [Medline: 32643080]
31. Tervonen T, Veldwijk J, Payne K, Ng X, Levitan B, Lackey LG, et al. Quantitative benefit-risk assessment in medical product decision making: a good practices report of an ISPOR task force. Value Health 2023 Apr;26(4):449-460 [FREE Full text] [doi: 10.1016/j.jval.2022.12.006] [Medline: 37005055]
32. Postmarket management of cybersecurity in medical devices. U.S. Food & Drug Administration. 2016 Dec. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/postmarket-management-cybersecurity-medical-devices [accessed 2024-07-03]
33. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
35. NIST SP 800-30. National Institute of Standards and Technology. URL: https://www.nist.gov/privacy-framework/nist-sp-800-30 [accessed 2024-10-02]
36. The NIST Cybersecurity Framework (CSF) 2.0. National Institute of Standards and Technology. 2023. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.ipd.pdf [accessed 2024-10-02]
37. Factors to consider regarding benefit risk in medical device product availability, compliance, and enforcement decisions. U.S. Food & Drug Administration. 2016 Dec 27. URL: https://tinyurl.com/2s4arbhn [accessed 2024-07-03]
38. GN-20: guidance on clinical evaluation. Health Science Authority. 2022 Nov. URL: https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/gn-20-r2-guidance-on-clinical-evaluation(2022-nov)_pub.pdf?sfvrsn=426a4e43_2 [accessed 2024-10-02]
39. Consideration of uncertainty in making benefit-risk determinations in medical device premarket approvals, de novo classifications, and humanitarian device exemptions. U.S. Food & Drug Administration. 2019 Aug 30. URL: https://www.fda.gov/media/115672/download [accessed 2024-07-03]
40. Factors to consider when making benefit-risk determinations in medical device premarket approval and de novo classifications. U.S. Food & Drug Administration. 2019 Aug 30. URL: https://www.fda.gov/media/99769/download [accessed 2024-07-03]
41. Medical device cyber security information for users. Commonwealth of Australia. 2022 Nov 24. URL: https://www.tga.gov.au/resources/resource/reference-material/medical-device-cyber-security-information-users [accessed 2024-10-02]
42. Principles and practices for medical device cybersecurity. International Medical Device Regulators Forum. 2020 Apr 20. URL: https://www.imdrf.org/documents/principles-and-practices-medical-device-cybersecurity [accessed 2024-07-10]
43. Guidance documents for medical devices. Health Sciences Authority. URL: https://www.hsa.gov.sg/medical-devices/guidance-documents [accessed 2024-10-02]
44. Guideline on review and approval for cybersecurity of medical devices (for industry). Ministry of Food and Drug Safety. URL: https://tinyurl.com/mr3jp2t9 [accessed 2024-10-02]
45. Software and AI as a medical device change programme - roadmap. Medicines & Healthcare products Regulatory Agency, United Kingdom Government. 2023 Jun 14. URL: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap [accessed 2024-10-02]

XSL•FO
RenderX

46. Complying with the Essential Principles on the safety and performance of medical devices. The Department of Health and Aged Care, Australian Government. 2022 Aug 25. URL: https://www.tga.gov.au/resources/guidance/complying-essential-principles-safety-and-performance-medical-devices [accessed 2024-10-02]
47. Principles and practices for software bill of materials for medical device cybersecurity. International Medical Device Regulators Forum. 2023 Apr 13. URL: https://www.imdrf.org/documents/principles-and-practices-software-bill-materials-medical-device-cybersecurity [accessed 2024-07-10]
48. Principles and practices for the cybersecurity of legacy medical devices. International Medical Device Regulators Forum. 2023 Apr 11. URL: https://www.imdrf.org/documents/principles-and-practices-cybersecurity-legacy-medical-devices [accessed 2024-07-10]
49. IEC 82304-1:2016 Health software part 1: general requirements for product safety. International Electrotechnical Commission. 2016. URL: https://www.iso.org/standard/59543.html [accessed 2024-10-02]
50. IEC 80001-1:2021: application of risk management for IT-networks incorporating medical devices - part 1: Safety, effectiveness and security in the implementation and use of connected medical devices or connected health software. International Electrotechnical Commission. 2021 Sep. URL: https://www.iso.org/standard/72026.html [accessed 2024-07-10]
51. IEC 81001-5-1:2021: health software and health IT systems safety, effectiveness and security. International Electrotechnical Commission. 2021. URL: https://www.iso.org/standard/76097.html [accessed 2024-07-10]
52. ANSI/AAMI SW96:2023; standard for medical device security—security risk management for device manufacturers. Association for the Advancement of Medical Instrumentation. 2022. URL: https://array.aami.org/doi/book/10.2345/9781570208621 [accessed 2024-07-03]
53. IEC/TR 80002-1:2009: medical device software part 1: guidance on the application of ISO 14971 to medical device software. International Electrotechnical Commission. 2009 Sep. URL: https://www.iso.org/standard/54146.html [accessed 2024-10-02]
54. IEC/TR 80001-2-2:2012: application of risk management for IT-networks incorporating medical devices: part 2-2: guidance for the communication of medical device security needs, risks and controls. International Electrotechnical Commission. 2012. URL: https://www.iso.org/standard/57939.html [accessed 2024-07-10]
55. IEC/TR 80001-2-1:2012: application of risk management for IT-networks incorporating medical devices: part 2-1: step by step risk management of medical IT-networks; practical applications and examples. International Electrotechnical Commission. 2012. URL: https://www.iso.org/standard/57934.html [accessed 2024-07-10]
56. ISO/TR 80001-2-7:2015: application of risk management for IT-networks incorporating medical devices — application guidance. International Organization for Standardization. 2015. URL: https://www.iso.org/standard/63509.html [accessed 2024-07-10]
57. AAMI TIR57:2016 (R2023) principles for medical device security - risk management. Association for the Advancement of Medical Instrumentation (AAMI). 2023. URL: https://webstore.ansi.org/standards/aami/aamitir572016r2023 [accessed 2024-07-10]
58. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (Text with EEA relevance.). European Parliament, European Council. 2017. URL: http://data.europa.eu/eli/reg/2017/746/oj/eng [accessed 2024-07-25]
59. Guidance document: pre-market requirements for medical device cybersecurity. Government of Canada. 2019. URL: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/cybersecurity/document.html [accessed 2024-10-02]
60. DIN EN ISO 14971:2022-04. Deutsches Institut für Normung. 2022. URL: https://www.dinmedia.de/de/norm/din-en-iso-14971/348493801 [accessed 2024-07-03]
61. Lins M, Mayrhofer R, Roland M, Hofer D, Schwaighofer M. On the critical path to implant backdoors and the effectiveness of potential mitigation techniques: early learnings from XZ. arXiv Preprint posted online on April 13, 2024 [FREE Full text] [doi: 10.48550/arXiv.2404.08987]
62. Bilge L, Dumitraş T. Before we knew it: an empirical study of zero-day attacks in the real world. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security. 2012 Presented at: CCS '12; October 16-18, 2012; Raleigh, NC. [doi: 10.1145/2382196.2382284]
63. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
64. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023 Oct 30;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
65. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. The Lancet Digit Health 2024 Sep;6(9):e662-e672. [doi: 10.1016/s2589-7500(24)00124-9]
66. Freyer O, Wrona KJ, de Snoeck Q, Hofmann M, Melvin T, Stratton-Powell A, et al. The regulatory status of health apps that employ gamification. Sci Rep 2024 Sep 09;14(1):21016 [FREE Full text] [doi: 10.1038/s41598-024-71808-2] [Medline: 39251786]
67. Freyer O, Gilbert S. Bridging between hype and implementation in medical extended reality. NPJ Digit Med 2023 Dec 07;6(1):227 [FREE Full text] [doi: 10.1038/s41746-023-00972-y] [Medline: 38062115]

XSL•FO
RenderX

68. Souchet AD, Lourdeaux D, Pagani A, Rebenitsch L. A narrative review of immersive virtual reality's ergonomics and risks at the workplace: cybersickness, visual fatigue, muscular fatigue, acute stress, and mental overload. Virtual Reality 2022 Jul 16;27(1):19-50. [doi: 10.1007/s10055-022-00672-0]
69. Bughio KS, Cook DM, Shah SA. Developing a novel ontology for cybersecurity in internet of medical things-enabled remote patient monitoring. Sensors (Basel) 2024 Apr 27;24(9):2804 [FREE Full text] [doi: 10.3390/s24092804] [Medline: 38732910]
70. Sowmya T, Mary Anita EA. A comprehensive review of AI based intrusion detection system. Meas Sens 2023 Aug;28:100827. [doi: 10.1016/j.measen.2023.100827]
71. Salem AH, Azzam SM, Emam OE, Abohany AA. Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. J Big Data 2024 Aug 04;11:105. [doi: 10.1186/s40537-024-00957-y]
72. Ksibi S, Jaidi F, Bouhoula A. A comprehensive study of security and cyber-security risk management within e-health systems: synthesis, analysis and a novel quantified approach. Mobile Netw Appl 2022 Sep 29;28(1):107-127. [doi: 10.1007/s11036-022-02042-1]
73. Abdulsatar M, Ahmad H, Goel D, Ullah F. Towards deep learning enabled cybersecurity risk assessment for microservice architectures. arXiv Preprint posted online on March 22, 2024 [FREE Full text]
74. Sarker IH, Furhad MH, Nowrozy R. AI-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Comput Sci 2021 Mar 26;2:173. [doi: 10.1007/s42979-021-00557-0]
75. Best IoT device management platforms: a comprehensive comparison. Novotech. 2023 Jul 26. URL: https://novotech.com/blogs/news/best-iot-device-management-platforms-a-comprehensive-comparison [accessed 2024-10-09]
76. Alzahrani FA, Ahmad M, Ansari MT. Towards design and development of security assessment framework for internet of medical things. Appl Sci 2022 Aug 15;12(16):8148. [doi: 10.3390/app12168148]
77. Wang L, Ali Y, Nazir S, Niazi M. ISA evaluation framework for security of internet of health things system using AHP-TOPSIS methods. IEEE Access 2020;8:152316-152332. [doi: 10.1109/access.2020.3017221]
78. Gordon LA, Loeb MP, Zhou L. Integrating cost–benefit analysis into the NIST Cybersecurity Framework via the Gordon–Loeb Model. J Cybersecurity 2020;6(1):tyaa005 [FREE Full text] [doi: 10.1093/cybsec/tyaa005]
79. Krutilla K, Alexeev A, Jardine E, Good D. The benefits and costs of cybersecurity risk reduction: a dynamic extension of the Gordon and Loeb Model. Risk Anal 2021 Oct;41(10):1795-1808. [doi: 10.1111/risa.13713] [Medline: 33586801]

Abbreviations

AAMI: Association for the Advancement of Medical Instrumentation
AI: artificial intelligence
AU_MDR: Australian Therapeutic Goods (Medical Devices) Regulations
BRA: benefit-risk analysis
BRF: Benefit-Risk Framework
cMD: connected medical device
EU: European Union
EU_MDR: European Union Medical Device Regulation
FD&C Act: Federal Food, Drug, and Cosmetic Act
FDA: Food and Drug Administration
HSA: Health Sciences Authority
IEC: International Electrotechnical Commission
IMDRF: International Medical Device Regulators Forum
ISO: International Organization for Standardization
MCDA: multicriteria decision analysis
MD: medical device
MDCG: Medical Device Coordination Group
MFDS: Ministry of Food and Drug Safety
MHRA: Medicines and Healthcare Products Regulatory Agency
NIST: National Institute of Standards and Technology
PMS: postmarket surveillance
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SBOM: software bill of materials
TGA: Therapeutic Goods Administration

XSL•FO
RenderX

Edited by A Mavragani; submitted 18.08.24; peer-reviewed by J Dykstra, N Hakiem, S Vistini; comments to author 12.09.24; revised version received 11.10.24; accepted 13.11.24; published 24.12.24.

Please cite as:
Freyer O, Jahed F, Ostermann M, Rosenzweig C, Werner P, Gilbert S
Consideration of Cybersecurity Risks in the Benefit-Risk Analysis of Medical Devices: Scoping Review
J Med Internet Res 2024;26:e65528
URL: https://www.jmir.org/2024/1/e65528
doi:10.2196/65528
PMID:39718821

©Oscar Freyer, Fatemeh Jahed, Max Ostermann, Christian Rosenzweig, Pascal Werner, Stephen Gilbert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Platform-Based Patient-Clinician Digital Health Interventions for Care Transitions: Scoping Review

Chantal Backman[1*], RN, MHA, PhD; Rosie Papp[1*], BSc; Aurelie Tonjock Kolle[1], RN, MScN; Steve Papp[2], MSc, MD, FRCSC; Sarah Visintini[1], MLS; Ana Lúcia Schaefer Ferreira de Mello[3], PhD; Gabriela Marcellino de Melo Lanzoni[3], PhD; Anne Harley[4], MD, CCFP, COE, FCFP

[1]University of Ottawa, Ottawa, ON, Canada
[2]The Ottawa Hospital, Ottawa, ON, Canada
[3]Universidade Federal de Santa Catarina, Florianópolis, Brazil
[4]Bruyère Continuing Care, Ottawa, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Chantal Backman, RN, MHA, PhD
University of Ottawa
200 Lees Ave
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6133240838
Email: chantal.backman@uottawa.ca

Abstract

Background: Care transitions are complex and can make patients vulnerable to adverse events. Poor communication among clinicians, patients, and their caregivers is a critical gap during these periods of transition. Technology solutions such as platform-based patient-clinician digital health interventions (DHIs) can provide support and education to patients.

Objective: The aims of this scoping review were to explore the literature on platform-based patient-clinician DHIs specific to hospital-to-home care transitions and identify the barriers to and enablers of the uptake and implementation of these DHIs.

Methods: A scoping review was conducted. A total of 4 databases (MEDLINE, CINAHL, Embase, and the Cochrane Central Register of Controlled Trials) were searched on July 13, 2022. Studies involving patients aged >18 years who used platform-based DHIs during their hospital-to-home transition were included. In total, 2 reviewers independently screened the articles for eligibility using a 2-stage process of title and abstract and full-text screening. Eligible studies underwent data extraction, and the results were analyzed using descriptive and narrative methods.

Results: We screened 8322 articles, of which 97 (1.17%) met our inclusion criteria. DHIs were implemented using a mobile app (59/97, 61%), a web-based platform (28/97, 29%), or a combination of both (10/97, 10%). The 2 most common health conditions related to the DHIs were cardiac disease (22/97, 23%) and stroke (11/97, 11%). Outcomes varied greatly but were grouped by health care use, complications, and wellness outcomes. The top 2 barriers were lack of interest (13/97, 13%) and time constraints to use the DHIs (10/97, 10%), and the top 2 enablers were the ability to use the DHIs (17/97, 18%) and their ease of use (11/97, 11%). The main conflicting theme was access (enabler; 28/97, 29%) or limited access (barrier; 15/97, 15%) to technology or the internet.

Conclusions: Platform-based DHIs could help improve communication, coordination, and information sharing between clinicians and patients during transition periods. Further research is needed to assess the effectiveness of these platform-based DHIs on patient outcomes.

(J Med Internet Res 2024;26:e55753) doi:10.2196/55753

KEYWORDS

platform based; patient-clinician; digital health intervention; care transition; mobile phone

XSL•FO
RenderX

Introduction

Care transitions occur when a patient moves from one health care setting to another, such as from hospital to home. Effective care transitions are critical for ensuring that patients receive safe, high-quality care and for reducing the risk of adverse events, such as medication errors or unnecessary hospital readmissions [1]. However, care transitions are often complex and often involve multiple health care providers. In addition, recently hospitalized patients typically experience a temporary period of generalized risk of a wide range of adverse health events, also called "posthospital syndrome" [2]. Consequently, care transitions can be a significant source of errors, delays, and gaps in care [3,4]. One critical gap that has been identified is poor communication among clinicians, patients, and their caregivers [5,6], which is the one aspect that we hypothesize can be addressed through better technology solutions.

Technological solutions that can help improve the transition of patients between different care settings or health care providers are often referred to as "platform-based patient-clinician digital health interventions (DHIs)." These interventions can be used to address challenges during care transitions by leveraging technology to improve communication, coordination, and information sharing between clinicians and patients. These interventions can include but are not limited to providing patients with (1) personalized care plans, reminders, and educational resources to support self-management during transitions [7]; (2) remote monitoring to help health care providers track patient vital signs and symptoms to detect early warning signs of complications [8]; and (3) exchange and communication of information between clinicians and patients [9].

Previous research has identified 10 key elements in an ideal care transition [10]. The key domains of the Ideal Transition of Care framework [10] comprise (1) discharge planning; (2) complete communication of information; (3) availability, timeliness, clarity, and organization of information; (4) medication safety; (5) patient education to promote self-management; (6) social and community supports; (7) advanced care planning; (8) coordination of care among team members; (9) symptom monitoring and management after discharge; and (10) outpatient follow-ups. This framework was created to establish the necessary considerations pertaining to safe transitions in care. Ideally, any DHIs being developed would consider these proposed domains.

Platform-based patient-clinician DHIs can also facilitate continuity of care. Continuity of care is an important aspect that is present when a patient experiences coherent and linked care over time or when discrete elements of care that endure over time are maintained and supported [11]. Although continuity of care can be interpreted differently between care providers, there are 3 types that are agreed upon. The first type is "informational continuity," which refers to the use of data from previous events in a patient's medical history to inform the appropriate care of the patient's current encounter [11]. The second type is "management continuity," which occurs when care from multiple health care providers is linked in a coherent manner [11]. The third type is "relational continuity," which acknowledges the importance of the relationship between patients and providers [11].

In addition, platform-based patient-clinician DHIs for care transitions have the potential to encourage individual behavior change and reduce health care costs [12]. It is important to ensure that these interventions are evidence based, user-friendly, and aligned with patient preferences and needs. There are limited reviews on DHIs for discharge and transitional care use, and these reviews focus mainly on a specific health aspect or condition, such as self-care associated with surgery [9] or postsurgery care for patients with hip fracture [13]. One other review identified specific health care provider roles and functions related to the use of DHIs [14]. Given that patients often have multiple health conditions, there are, to our knowledge, no reviews to date examining the use of DHIs during care transitions across various conditions to reduce siloed care. The aims of this scoping review were to explore the literature on platform-based patient-clinician DHIs specific to hospital-to-home care transitions and identify the barriers to and enablers of the uptake and implementation of these platform-based patient-clinician DHIs.

Methods

Design

We conducted a scoping review based on the work by Arksey and O'Malley [15], Levac et al [16], and the Joanna Briggs Institute scoping review methodologies [17]. The study is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [18] (Multimedia Appendix 1). The methods involved five steps: (1) identifying the research question; (2) identifying relevant studies; (3) selecting the studies; (4) charting the data; and (5) collating, synthesizing, and reporting the results. The detailed protocol has been published previously [19].

Search

A peer-reviewed search [20] was conducted on July 13, 2022, in MEDLINE, CINAHL, Embase, and the Cochrane Central Register of Controlled Trials (Multimedia Appendix 2). The main search concepts comprised terms related to "hospital to home transition," "patient discharge," "transitional care," "internet-based interventions," "mobile applications," "mhealth," and "digital health" platforms and were informed by previously conducted systematic searches [14,21,22]. No limit to language was applied; however, the results were limited by a publication date from 2012 onward. We chose to review results from the previous 10 years, recognizing the rapid pace of technological advancements. Search results were exported to Covidence (Veritas Health Innovation), a systematic review software [23], and duplicates were removed using the platform's duplicate identification features.

Screening

Studies were screened in 2 steps (title and abstract, and full text) based on the eligibility criteria (Textbox 1).

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Population: adult patients (aged >18 y) discharged from hospital to home
- Concept: digital-based platforms that support a hospital-to-home transition, including web-based digital health interventions (DHIs), defined as programs that were delivered via the internet and accessed through a website link (URL) [24], and mobile apps, defined as software programs developed for smartphones [25]
- Context: hospital-to-home transitions
- Type of evidence: studies published after 2012 and identified as randomized controlled trials, quasi-experimental studies, pilot studies, feasibility studies, observational studies (case-control, cohort, cross-sectional, and descriptive studies), or qualitative studies

Exclusion criteria

- Concept: non–platform-based DHIs, including but not limited to wearable devices if the intervention was stand-alone (eg, to track activity), prosthetics, robotics, medical imaging technology (eg, x-rays and ultrasounds), interventions using only a standard telephone, machine learning, and telehealth
- Type of evidence: studies in the design stage at the time of screening or if the platform-based patient-clinician DHIs were tailored specifically to and solely focused on mental health or cancer as these are unique clinical areas

Data Charting

Studies that met all the inclusion criteria were extracted using the Google Forms platform. The pilot-tested Google Form used a survey format for the extraction of the relevant information, including the following: lead author, year of publication, country, objectives, study design, participants, patient health condition, name of DHI, rationale of intervention, theory guiding the intervention, content of the intervention, elements of postcare, digital health tools, function of digital health tool, management or informational or relational continuity of care [11], who provided the intervention, number of days after discharge, number and duration of sessions, tailoring or modification, adherence or attrition, results, and barriers or enablers. To ensure a high-quality description of the interventions, the information extracted from each study was guided by the recommendations made by the 12-item Template for Intervention Description and Replication [26]. After data charting was completed, the Google Form was used to generate a Microsoft Excel spreadsheet to analyze the findings using descriptive and narrative methods.

Results

Overview

A total of 12,752 records were retrieved from the search, of which 4430 (34.74%) were duplicates and 7837 (61.46%) were rejected at the abstract review stage, leaving 485 (3.8%) records selected for full-text review. A total of 34.2% (166/485) of the reports could not be retrieved as the full text was unavailable. In addition, 222 full-text articles were excluded for the following reasons: incorrect intervention (not DHIs related to care transitions; n=196, 88.3%), language other than French or English (n=15, 6.8%), incorrect study design (n=9, 4.1%), and incorrect patient population (n=2, 0.9%). The list of articles with reasons for exclusion can be found in Multimedia Appendix 3. The final review included a total of 97 studies after the assessment process was completed (Figure 1 [27]).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram—digital heath interventions (DHIs) for care transitions.

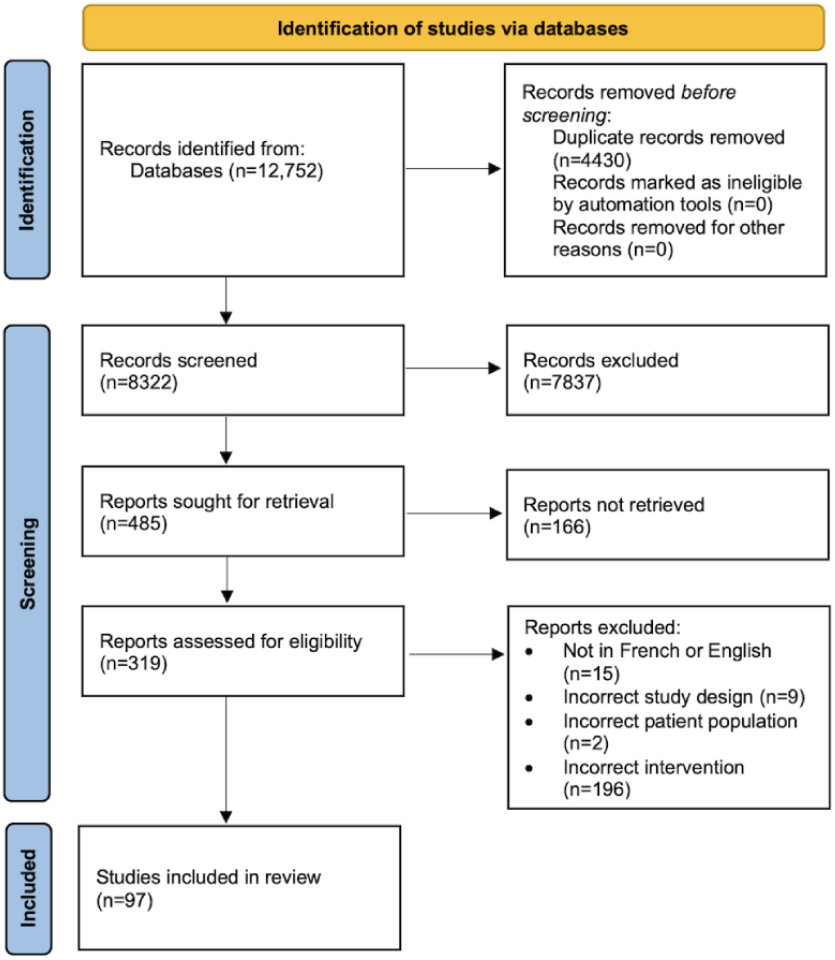

Characteristics of the Included Studies

Of the 97 included studies [28-124], 34 (35%) were conducted in the United States [31,38,43,48,49,53,54,56,59,61-63,67,71,73-75,78,79,87,89,90,92,93,102,104-108,111,113,116,118], 16 (16%) were conducted in Canada [30,33,36,51,52,55,57,60,70,80,81,94,96,97,100,117], 10 (10%) were conducted in China [41,50,82-85,91,109,121,122], 8 (8%) were conducted in the Netherlands [39,40,76,95,103,112,119,120], and 5 (5%) were conducted in Australia [42,66,68,101,115]. The remaining 25% (24/97) of the studies were conducted in 14 other countries [28,29,32,34,35,37,44-47,58,64,65,69,72,77,86,88,98,99,110,114,123,124]. The 2 most common health conditions relevant to the platform-based patient-clinician DHIs were cardiac disease (22/97, 23%) [29,31,36,38,49,50,58,59,65-67,72,79,83,87,89,90,97,109,114,116,123] and stroke (11/97, 11%) [32,44,71,96,98,104,107,110,115,119,124].

A wide variety of elements of postcare were implemented in the platform-based patient-clinician DHIs, including but not limited to detecting postoperative issues, assessing patients' needs, improving patient understanding, rehabilitation, improving symptom management, and increasing self-care (Table 1).

XSL•FO
RenderX

Table 1. Study characteristics.

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Agri et al [28], 2020	Switzerland	Retrospective monocentric cohort study	Community or home or retirement home	Patients—intervention: n=43	Colorectal surgery (colorectal resections, ostomy procedures, and stoma closures)	Maela	Mobile app	Enhancing patient-provider communication; ability to identify symptoms of poor wound healing; enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; pain control or management; wound care
Antypas and Wangberg [29], 2014	Norway	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=7; control: n=12	Cardiovascular disease	Internet- and mobile-based tailored intervention to enhance maintenance of physical activity after cardiac rehabilitation	Mobile app and web app	Improving physical activity
Armstrong et al [30], 2017	Canada	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=32; control: n=33	Ambulatory breast reconstruction surgery	QoC[b] Health Inc mobile app	Mobile app	Ability to identify symptoms of poor wound healing; detecting any postoperative issues; improving symptom management; follow-up appointments with primary care provider; pain control or management; wound care
Athilingam et al [31], 2017	United States	Feasibility (pilot) study; randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=9; control: n=9	Congestive HF[c]	Mobile app to improve self-care behaviors and quality of life for patients with HF	Mobile app	Increasing self-care; medication management; fostering treatment adherence; improving quality of life; enhancing HF-specific knowledge
Avci and Gozum [32], 2018	Turkey	Descriptive study	Community, home, or retirement home	Caregivers—intervention: n=62	Stroke	Supportive website for the caregivers of patients with stroke after discharge	Web app	Enhancing caregiver preparedness; enhancing caregiver knowledge and skills
Backman et al [33], 2020	Canada	Feasibility (pilot) study	Community, home, or retirement home	Patients: n=34; caregivers: n=19; clinicians: n=37	Hip fracture	MyPath to Home	Web app	Promoting communication among patients, caregivers, and clinicians
Bäcker et al [34], 2021	Germany	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=20; control: n=15	Knee arthroplasty	GenuSport	Mobile app	Rehabilitation (physical therapy and occupational therapy)
Bauwens et al [35], 2022	France	Case-control	Community, home, or retirement home	Patients—intervention: n=32; control: n=101	ACL[d] reconstruction surgery	Doct-Up	Mobile app	Encouraging ambulation; pain control or management; rehabilitation (physical therapy and occupational therapy)

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Ben-Ali et al [36], 2021	Canada	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=1108	Cardiac surgery	SeamlessMD	Mobile app	Enhancing patients' knowledge, skills, and confidence; improving patient understanding; increasing self-care; detecting any postoperative issues; encouraging lifestyle changes; improving symptom management; pain control or management; wound care; providing PROs[e], surveys, and feedback
Birkhäuser et al [37], 2020	Switzerland	Prospective nonrandomized pilot clinical trial	Community, home, or retirement home	Patients—intervention: n=18	Radical cystectomy	Cellphone-based health care app	Mobile app and web app	Monitor progression of patient recovery
Blewer et al [38], 2020	United States	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=699; control: n=626	Coronary artery disease	mApp CPR[f] training app	Mobile app	Providing effective CPR
Bouwsma et al [39], 2018	The Netherlands	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=227; control: n=206	Hysterectomy and laparoscopic adnexal surgery	eHealth intervention	Web app	Ability to identify symptoms of poor wound healing; enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; improving symptom management; guidance in the process of resuming work activities
Bouwsma et al [40], 2018	The Netherlands	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=227; control: n=206	Hysterectomy or laparoscopic adnexal surgery	eHealth intervention	Web app	Enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; encouraging return to work
Cheng et al [41], 2022	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=19; control: n=20	Hip fracture	Home-based rehabilitation mobile app	Mobile app	Enhancing caregiver knowledge and skills; enhancing patients' knowledge, skills, and confidence; improving physical activity; rehabilitation (physical therapy and occupational therapy); fostering treatment adherence; progress summary (can track completion of tasks), reminders, and support information
Cox et al [42], 2015	Australia	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=10	Cystic fibrosis	ActivOnline	Web app	Improving physical activity; encouraging lifestyle changes; improving symptom management; improving quality of life
Davis et al [43], 2020	United States	Retrospective review	Community, home, or retirement home	Patients—intervention: n=47	Total shoulder arthroplasty	Force Therapeutics	Mobile app	Enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; rehabilitation (physical therapy and occupational therapy)

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Davoody and Hägglund [44], 2016	Sweden	Qualitative study	Community, home, or retirement home	Care professionals—intervention: n=8	Stroke	eHealth for postdischarge stroke	Web app	Improve patient understanding; rehabilitation (physical therapy and occupational therapy)
De Batlle et al [45], 2021	Spain	Prospective, pragmatic, 2-arm, parallel-implementation trial	Community, home, or retirement home	Patients—intervention: n=48; control: n=28	Chronic obstructive pulmonary disease and HF	CONNECARE	Mobile app and wearable device	Enhancing patient-provider communication; monitoring pulse, oxygen, HR[g], BP[h], and weight at home; enhancing patients' knowledge, skills, and confidence; improving physical activity; assessing patients' needs
Debono et al [46], 2016	France	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=60	Ambulatory lumbar discectomy	Mobile app for postoperative monitoring after outpatient lumbar discectomy	Mobile app	Enhancing patient-provider communication; assessing rates of complications; pain control or management
Debono et al [47], 2019	France	Retrospective analysis	Community, home, or retirement home	Patients—intervention: n=1920; control: n=1563	Spinal cord injury and lumbar disc herniation	e-fitback	Mobile app	Enhancing patient-provider communication; detecting any postoperative issues; pain control or management; wound care
Devito Dabbs et al [48], 2016	United States	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=99; control: n=102	Lung transplant	Pocket PATH[i]	Mobile app	Detecting any postoperative issues; assessing patients' needs; assessing rates of complications; pain control or management
Dorsch et al [49], 2021	United States	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=42; control: n=41	HF	Manage-HF4Life	Mobile app and monitoring devices	Enhancing patients' knowledge, skills, and confidence; increasing self-care; encouraging lifestyle changes; improving symptom management; assessing patients' needs; medication management; nutrition support; improving quality of life
Duan et al [50], 2018	China	Pilot randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=44; control: n=39	Coronary artery disease	Health behavior intervention for patients with coronary heart disease through the web	Web app	Improving physical activity; improving food consumption
Dukeshire et al [51], 2012	Canada	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=31	Hysterectomy	SAFER[j] project	Web app	Improving symptom management; ability to identify symptoms; postsurgical care
Eustache et al [52], 2023	Canada	Cohort study	Community, home, or retirement home	Patients—intervention: n=94; matched cohort: n=256	Colorectal surgery	Same-day discharge mHealth[k] app (CareSense)	Mobile app	Enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; enhancing patient-provider communication

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Felbaum et al [53], 2018	United States	Prospective cohort study	Community, home, or retirement home	Patients—intervention: n=56	Spinal cord injury, lumbar disc herniation, and neurosurgery (spinal and cranial surgery)	TrackMyRecovery	Mobile app	Enhancing patient-provider communication; ability to identify symptoms of poor wound healing; detecting any postoperative issues; pain control or management; wound care
Ganapathy et al [54], 2017	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=40; caregivers—intervention: n=40	Cirrhosis and hepatic encephalopathy	Patient Buddy	Mobile app	Enhancing caregiver preparedness; enhancing caregiver knowledge and skills; enhancing patient-provider communication; monitoring pulse, oxygen, HR, BP, and weight at home; enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; delirium screening and management; fall prevention; medication management; entering grams of sodium consumed; assessing cognition (EncephalApp); Timed Up and Go test
Gollish et al [55], 2019	Canada	Feasibility (pilot) study; qualitative study	Community, home, or retirement home	Patients—intervention: n=629	Total hip replacement and knee arthroplasty	myHip&Knee	Mobile app	Enhancing patient-provider communication; increasing self-care; detecting any postoperative issues; encouraging lifestyle changes; improving symptom management; medication management; pain control or management
Gunter et al [56], 2018	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=40	Vascular surgery	WoundCheck	Mobile app	Ability to identify symptoms of poor wound healing; detecting any postoperative issues; improving symptom management; wound care
Habib et al [57], 2021	Canada	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=23; control: n=26	All health conditions	SAM[l]	Mobile app	Medication management; promoting adherence to medication
Hägglund et al [58], 2015	Sweden	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=32; control: n=40	HF	Home intervention system (OPTILOGG[m])	Web app and monitoring devices	Enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; improving patient understanding; increasing self-care; improving symptom management

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Heiney et al [59], 2020	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=12	HF	Healthy Heart	Mobile app	Enhancing patient-provider communication; monitoring pulse, oxygen, HR, BP, and weight at home; ability to identify symptoms of HF; enhancing patients' knowledge, skills, and confidence; encouraging lifestyle changes; medication management; managing emotional changes; improving quality of life
Heuser et al [60], 2021	Canada	Retrospective cohort study	Community, home, or retirement home	Patients—intervention: n=396; control: n=458	Obesity and bariatric surgery	SeamlessMD	Mobile app	Managing mood and anxiety; enhancing patients' knowledge, skills, and confidence; increasing self-care; improving physical activity; encouraging lifestyle changes; improving symptom management; medication management; nutrition support; wound care
Heyworth et al [61], 2014	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=60	All health conditions	SMMRT[n]	Web app	Medication management
Highland et al [62], 2019	United States	Feasibility (pilot) study; randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=24; control: n=26	Peripheral nerve block affecting one or more of the limbs	mCare system	Mobile app	Detecting any postoperative issues; assessing patients' needs
Holzer et al [63], 2022	United States	Feasibility (pilot) study	Community, home, or retirement home, rehabilitation, long-term care and nursing home (24-hour care)	Patients—intervention: n=89; control: n=128	Acute venous thromboembolism	HealthFlo	Mobile app	Enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; medication management
Houchen-Wolloff et al [64], 2021	United Kingdom	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=100	Chronic obstructive pulmonary disease	SPACE[o] for chronic obstructive pulmonary disease	Web app	Enhancing patient-provider communication; improving patient understanding; increasing self-care; improving physical activity
İlaslan and Özer [65], 2022	Turkey	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=32; control: n=32	Congestive HF	Web app for training and telephone follow-up for patients with HF	Web app	Ability to identify symptoms of HF; improving symptom management; meeting the informational needs of patients
Indraratna et al [66], 2022	Australia	Feasibility (pilot) study; randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=81; control: n=83	Cardiac disease	TeleClinical Care	Mobile app and wearable device	Monitoring pulse, oxygen, HR, BP, and weight at home

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Johnson et al [67], 2022	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=16; control: n=15	Decompensated HF	HF-SMART[p]	Web app	Ability to identify symptoms of HF; enhancing patients' knowledge, skills, and confidence; improving patient understanding; encouraging lifestyle changes; improving symptom management; improving quality of life
Kang et al [68], 2022	Australia	Feasibility (pilot) study; randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=43; control: n=42	General surgery	Web-based discharge education program	Web app	Increasing self-care
Kargar et al [69], 2020	Iran	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=30; control: n=30	Burns	Self-care educational mobile app for burns	Mobile app and web app	Improving quality of life
Keng et al [70], 2020	Canada	Cross-sectional study	Community, home, or retirement home	Patients—intervention: n=106	Colorectal surgery	Home to Stay digital program after colorectal surgery	Mobile app	Monitoring patient recovery at home
Kersey et al [71], 2022	United States	Feasibility (pilot) study; randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=16; control: n=15	Stroke	mHealth platform for strategy training in inpatient stroke rehabilitation (iADAPT[q])	Mobile app	Assessing the feasibility of inpatient stroke rehabilitation
Khan et al [72], 2018	Denmark	Mixed methods study	Community, home, or retirement home	Patients—intervention: n=33	Cardiac surgery	Activeheart portal	Web app	Not reported
Khanwalkar et al [73], 2019	United States	Case-control	Community, home, or retirement home	Patients—intervention: n=208	Septosplasty and endoscopic sinus surgery	DPE[r] platform	Mobile app	Detecting any postoperative issues; pain control or management; PROs; monitoring pain; collecting data on the postoperative day when the patient returned to work
Kim et al [74], 2016	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=13	Knee arthroplasty	iGB[s] program	Mobile app	Enhancing patients' knowledge, skills, and confidence; increasing self-care; detecting any postoperative issues; improving physical activity; encouraging lifestyle changes; rehabilitation (physical therapy and occupational therapy); improving quality of life

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Knapp et al [75], 2021	United States	Quantified patient engagement	Community, home, or retirement home	Patients—intervention: n=207	Total hip replacement and knee arthroplasty	PeerWell	Mobile app, web app, and SMS text messaging	Promoting mental well-being; enhancing patients' knowledge, skills, and confidence; improving physical activity; nutrition support; rehabilitation (physical therapy and occupational therapy); improving quality of life
Kooij et al [76], 2021	The Netherlands	Feasibility (pilot) study; mixed methods study	Community, home, or retirement home	Patients: n=39	Chronic obstructive pulmonary disease	Self-management app for high-risk patients with chronic obstructive pulmonary disease	Mobile app	Increasing self-care; improving symptom management; medication management
Kristjánsdóttir et al [77], 2013	Norway	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=48; control: n=64	Chronic widespread pain	Smartphone-based intervention for chronic widespread pain	Mobile app and web app	Promoting self-management of chronic pain
Kummerow et al [78], 2015	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=50	General surgery	MHAV[t]	Web app	Enhancing patient-provider communication; improving symptom management; assessing patients' needs; wound care
Layton et al [79], 2014	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=16	Coronary artery disease and congestive HF	Smartphone-based app to monitor outpatient discharge instructions of patients with cardiac disease	Mobile app	Enhancing patient-provider communication; improving patient understanding; increasing self-care; follow-up appointments with primary care provider; medication management; encouraging activity
Lee et al [80], 2022	Canada	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=70; control: n=35	Colorectal surgery	Mobile app follow-up for same-day discharge	Mobile app	Detecting any postoperative issues; pain control or management
Lee et al [81], 2022	Canada	Cohort study	Community, home, or retirement home	Patients—intervention: n=48; control: n=73	Colorectal surgery	mHealth remote postdischarge monitoring	Mobile app	Enhancing patient-provider communication; improving symptom management; assessing rates of complications
Liu et al [82], 2021	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=49; control: n=49	Spinal cord injury	Together	Mobile app	Enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; follow-up appointments with primary care provider; improving quality of life

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Li et al [83], 2022	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=143; control: n=147	Coronary artery disease	DTx[u]	Mobile app	Monitoring pulse, oxygen, HR, BP, and weight at home; enhancing patients' knowledge, skills, and confidence; encouraging lifestyle changes; improving symptom management; medication management
Lou et al [84], 2022	China	Quasi-experimental study	Community, home, or retirement home	Patients—intervention: n=101; control: n=60	Not reported	mVS[v] intervention to enhance spiritual well-being	Mobile app	Fostering spiritual well-being
Lyu et al [85], 2021	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=58; control: n=58	Diabetes (type 2)	Nurse-led web-based transitional care program	Web app	Increasing self-care; fostering treatment adherence; improving quality of life
María Gómez et al [86], 2022	Colombia	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=39; control: n=42	Diabetes (type 2)	mHealth app for patients with type 2 diabetes	Mobile app and web app	Managing glycemic control
Marvel et al [87], 2021	United States	Nonrandomized controlled trial	Community, home, or retirement home	Patients—intervention: n=200; control: n=864	Acute myocardial infarction	Acute myocardial infarction DHI	Mobile app and wearable device	Enhancing patient-provider communication; monitoring pulse, oxygen, HR, BP, and weight at home; increasing self-care; follow-up appointments with primary care provider; medication management
Metilda et al [88], 2021	India	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=50; control: n=50	Brain injury	Aimeo	Mobile app	Enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; detect any postoperative issues
Park et al [89], 2019	United States	Feasibility and adoptability study	Community, home, or retirement home	Patients—intervention: n=58	Congestive HF	Digital health monitoring for patients with HF	Mobile app and web app	Ability to identify symptoms of HF; increasing self-care; improving symptom management
Paruchuri et al [90], 2021	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=118; control: n=343	Coronary artery disease	Wellframe	Mobile app	Enhancing patients' knowledge, skills, and confidence; encouraging lifestyle changes; improving symptom management; improving quality of life

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Peng et al [91], 2022	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=47; control: n=47	Hepato-pancreato-biliary surgery and biliary tract disease	Mobile continuous nursing platform	Mobile app	Enhancing caregiver knowledge and skills; enhancing patient-provider communication; ability to identify symptoms of poor wound healing; enhancing patients' knowledge, skills, and confidence; improving caregiver understanding; detecting any postoperative issues; assessing patients' needs; medication management; managing emotional changes; nutrition support; improving quality of life; wound care; T-tube placement and fixation method; observation of bile-related traits; treatment method of T-tube slippage; selection, fixation, or replacement of drainage bag
Pickens et al [92], 2019	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=122	Hepato-pancreato-biliary surgery	SeamlessMD	Mobile app	Managing mood and anxiety; enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; improving physical activity; improving symptom management; nutrition support; improving quality of life; collecting PROs
Ponder et al [93], 2020	United States	Descriptive study; feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=47	Spinal cord injury	Smartphone app with a digital care pathway for patients undergoing spine surgery	Mobile app and web app	Improving patient engagement; facilitating shared decision-making between patients and caregivers
Pooni et al [94], 2022	Canada	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=128; control: n=125	Colorectal surgery	Postdischarge Home to Stay mobile app	Mobile app and web app	Ability to identify symptoms of poor wound healing; improving patient understanding; detecting any postoperative issues
Pronk et al [95], 2020	The Netherlands	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=38; control: n=33	Knee arthroplasty	PainCoach app	Mobile app	Medication management; pain control or management; rehabilitation (physical therapy and occupational therapy)
Pugliese et al [96], 2019	Canada	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=30	Stroke	RecoverNow	Mobile app	Enhancing patient-provider communication; rehabilitation (physical therapy and occupational therapy)
Reid et al [97], 2012	Canada	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=115; control: n=118	Cardiac disease	CardioFit internet-based expert system	Web app	Improving physical activity

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Requena et al [98], 2019	Spain	2-arm open-label nonrandomized study	Community, home, or retirement home	Patients—intervention: n=107; control: n=52	Stroke	FAR-MALARM	Mobile app	Enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; improving physical activity; medication management; controlling vascular risk factors
Rian et al [99], 2022	Norway	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=134	Knee arthroplasty	Eir (Eir Solutions AS)	Web app	Enhancing patient-provider communication; detecting any postoperative issues; medication management
Rosner et al [100], 2018	Canada	Cohort study	Community, home, or retirement home	Patients—intervention: n=371	Orthopedic fracture (any)	Internet-based orthopedic patient self-reports of post-discharge complications	Mobile app	Enhancing patient-provider communication; improving symptom management; assessing rates of complications
Saunders et al [101], 2021	Australia	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=50; control: n=49	Hip osteoarthritis	My Hip Journey	Web app and email reminders	Enhancing caregiver preparedness; enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; improving patient understanding; increasing self-care; improving physical activity; encouraging lifestyle changes; improving quality of life
Schenkel et al [102], 2020	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=28; control: n=28	Lung transplantation	ActiCare	Web app	Enhancing caregiver knowledge and skills; enhancing patient-provider communication; monitoring pulse, oxygen, HR, BP, and weight at home; enhancing patients' knowledge, skills, and confidence; increasing self-care; detecting any postoperative issues; encouraging lifestyle changes; medication management; nutrition support; rehabilitation (physical therapy and occupational therapy); tracking appointments
Scheper et al [103], 2019	The Netherlands	Cohort study	Community, home, or retirement home	Patients—intervention: n=69	Total hip replacement and knee arthroplasty	Woundcare	Mobile app	Enhancing patient-provider communication; ability to identify symptoms of poor wound healing; detecting any postoperative issues; pain control or management; wound care; prevention of prosthetic joint infection

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Schneider and Howard [104], 2017	United States	Descriptive study	Community, home, or retirement home	Patients—intervention: n=44; control: n=42	Stroke	Technology-improved coping for patients after stroke	Mobile app	Improving patient understanding; improving symptom management; follow-up appointments with primary care provider; medication management; managing emotional changes
Schubart [105], 2012	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=14	Spinal cord injury	e-Learning program to prevent pressure ulcers in adults with spinal cord injury	Web app	Pressure ulcer prevention; pressure ulcer management
Scott et al [106], 2017	United States	Mixed methods study	Community, home, or retirement home	Patients—intervention: n=20	Colorectal surgery	Postoperative mHealth app	Mobile app	Not reported
Siegel et al [107], 2016	United States	Feasibility (pilot) study	Not reported	Patients—intervention: n=3	Stroke	PHA[w] stroke app	Mobile app	Enhancing patient-provider communication; follow-up appointments with primary care provider; medication management
Stapler et al [108], 2022	United States	Case-control, retrospective analysis	Community, home, or retirement home	Patients—preintervention group: n=1052; postintervention group: n=668	Elective colon and rectal surgery; colorectal neoplasia, diverticulitis, IBD[x], and other diseases of the colon and rectum	St. Joseph's Health App	Mobile app	Detecting any postoperative issues; nutrition support; rehabilitation (physical therapy and occupational therapy); wound care
Su and Yu [109], 2021	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=73; control: n=73	Coronary heart disease	NeCR[y]	Web app	Promoting mental well-being; enhancing patients' knowledge, skills, and confidence; increasing self-care; detecting any postoperative issues; improving physical activity; encouraging lifestyle changes; assessing patients' needs; improving quality of life; cardiac rehabilitation
Sureshkumar et al [110], 2016	India	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=60	Stroke	Care for Stroke intervention	Mobile app and web app	Enhancing functional skills and activities of daily living
Symer et al [111], 2017	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=31	Major abdominal surgery	Gastrointestinal mHealth app	Mobile app and wearable device	Managing mood and anxiety; ability to identify symptoms of poor wound healing; detecting any postoperative issues; improving symptom management; wound care; decreasing length of hospital stay

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Timmers et al [112], 2019	The Netherlands	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=114; control: n=99	Knee arthroplasty	The Patient Journey App	Mobile app	Enhancing patients' knowledge, skills, and confidence; increasing self-care; pain control or management; rehabilitation (physical therapy and occupational therapy); improving quality of life; wound care
Tolentino [113], 2020	United States	Case-control, retrospective study	Community, home, or retirement home	Patients—intervention: n=50; control: n=50	Multiple conditions	Meducation	Mobile app	Medication management
Torri et al [114], 2018	Italy	Quasi-experimental study	Community, home, or retirement home	Patients—intervention: n=26; control: n=27	Coronary artery disease, cardiac surgery, congestive HF, percutaneous coronary revascularization, or acute ischemic events	CRMP[z]	Mobile app	Improving physical activity; encouraging lifestyle changes; medication management; improving quality of life
Van den Berg et al [115], 2016	Australia	Proof-of-concept trial	Community, home, or retirement home	Patients—intervention: n=31; control: n=32	Stroke	CARE4STROKE	Mobile app and wearable device	Enhancing patient-provider communication; improving physical activity; early weight bearing (weight bearing as tolerated); encouraging ambulation; rehabilitation (physical therapy and occupational therapy); improving quality of life
Venkatraman et al [116], 2022	United States	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=69	Cardiac surgery and aortic stenosis	MMS[aa]	Mobile app	Ability to identify symptoms of HF; enhancing patients' knowledge, skills, and confidence; increasing self-care; recording baseline and postoperative PROs
Vincent et al [117], 2021	Canada	Qualitative study; qualitative usability study	Community, home, or retirement home	Patients and caregivers—intervention: n=17	Hip fracture	My-HF[ab]	Mobile app	Enhancing patient-provider communication; managing mood and anxiety; enhancing patients' knowledge, skills, and confidence
Visperas et al [118], 2021	United States	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=204; control: n=195	Total hip replacement and knee arthroplasty	JointCOACH	Web app	Enhancing patient-provider communication; enhancing patients' knowledge, skills, and confidence; detecting any postoperative issues; medication management; pain control or management; rehabilitation (physical therapy and occupational therapy)

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Vloothuis et al [119], 2019	The Netherlands	Randomized controlled trial	Community, home, or retirement home	Intervention: n=32; control: n=34—patient-caregiver dyads	Stroke	CARE4STROKE digital intervention	Web app	Managing mood and anxiety; increasing self-care; improving physical activity; improving quality of life; improving motor impairment, strength, walking ability, balance, mobility, and (extended) activities of daily living of patients; reducing caregiver strain; reducing fatigue; improving quality of life of both patients and caregivers
Vonk Noordegraaf et al [120], 2014	The Netherlands	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=110; control: n=105	Gynecological surgery	Personalized eHealth program after gynecological surgery	Web app	Enhancing patient-provider communication; improving symptom management; achieving self-empowerment
Wang et al [121], 2017	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=55; control: n=65	Chronic obstructive pulmonary disease	Web-based coaching program using EHRs[ac]	Web app	Enhancing patient-provider communication; increasing self-care; improving symptom management; assessing patients' needs
Wang et al [122], 2018	China	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=100; control: n=103	Stoma	Stoma home care mobile app	Mobile app	Enhancing psychosocial adjustment; fostering self-efficacy; assessing stoma complication incidence
Werhahn et al [123], 2019	Germany	Feasibility (pilot) study	Community, home, or retirement home	Patients—intervention: n=10	HF	CPMP[ad]	Mobile app and wearable device	Monitoring pulse, oxygen, HR, BP, and weight at home; ability to identify symptoms of HF; improving physical activity; encouraging lifestyle changes; improving symptom management; assessing patients' needs; assessing rates of complications; medication management; monitoring physical activity (daily step count); PRO measures

XSL•FO
RenderX

Study	Country	Study design	Setting	Participants	Health condition	Name of DHI[a]	Digital health tool	Elements of postcare
Willeit et al [124], 2020	Austria	Randomized controlled trial	Community, home, or retirement home	Patients—intervention: n=1438; control: n=711	Stroke	MyStrokecard	Web app	Enhancing patient-provider communication; ability to identify symptoms of HF; enhancing patients' knowledge, skills, and confidence; encouraging lifestyle changes; improving symptom management; improving quality of life; risk factor monitoring

[a]DHI: digital health intervention.
[b]QoC: quality of care.
[c]HF: heart failure.
[d]ACL: anterior cruciate ligament.
[e]PRO: patient-reported outcome.
[f]CPR: cardiopulmonary resuscitation.
[g]HR: heart rate.
[h]BP: blood pressure.
[i]PATH: Personal Assistant for Tracking Health.
[j]SAFER: Studying Adverse Events From Elective Surgery Research.
[k]mHealth: mobile health.
[l]SAM: Smart About Meds.
[m]OPTILOGG: home intervention system.
[n]SMMRT: Secure Messaging for Medication Reconciliation Tool.
[o]SPACE: Self-Management Program of Activity Coping and Education.
[p]HF-SMART: Heart Failure Self-Management And Readmission Prevention Technique.
[q]iADAPT: mobile health platform for strategy training in inpatient stroke rehabilitation.
[r]DPE: digital patient engagement.
[s]IGB: iGetBetter.
[t]MHAV: My Health at Vanderbilt.
[u]DTx: digital therapeutics.
[v]mVS: mHealth-supported volunteer-assisted self-help.
[w]PHA: personal health assistant.
[x]IBD: inflammatory bowel disease.
[y]NeCR: nurse-led eHealth cardiac rehabilitation.
[z]CRMP: cardiac rehabilitation maintenance program.
[aa]MMS: ManageMySurgery.
[ab]My-HF: My Hip Fracture.
[ac]EHR: electronic health record.
[ad]CPMP: cardio patient monitoring platform.

Platform-based patient-clinician DHIs were implemented using a mobile app (59/97, 61%) [28,30,31,34-36,38,41,43,45-49,52-57,59,60,62,63,66,70,71,73,74,76,79-84,87,88,90-92,95,96,98,100,103,104,106-108,111-117,122,123], a web-based platform (28/97, 29%) [32,33,39,40,42,44,50,51,58,61,64,65,67,68,72,78,85,97,99,101,102,105,109,118-121,124], or a combination of both (10/97, 10%) [29,37,69,75,77,86,89,93,94,110].

Figure 2 illustrates the frequency distribution of platform-based DHIs by year of publication.

Only 5% (5/97) of the studies [83,91,96,109,121] included all 3 types of continuity of care [11]. Informational continuity was frequently implemented through patient education and facilitating patient-provider communication. The most common ways in which management continuity was implemented was providing an assessment of the patient, monitoring the patient's health status after discharge, and facilitating follow-up care. Relational continuity was the least implemented, with only 6% (6/97) of the studies in which the interventions included counseling or rapport building [50,82,91,96,109,121]. A more detailed description of each of the interventions using the Template for Intervention Description and Replication [26] is available in Multimedia Appendix 4 [28-124].

XSL•FO
RenderX

Figure 2. Frequency distribution of platform-based digital health interventions used for care transitions by year.

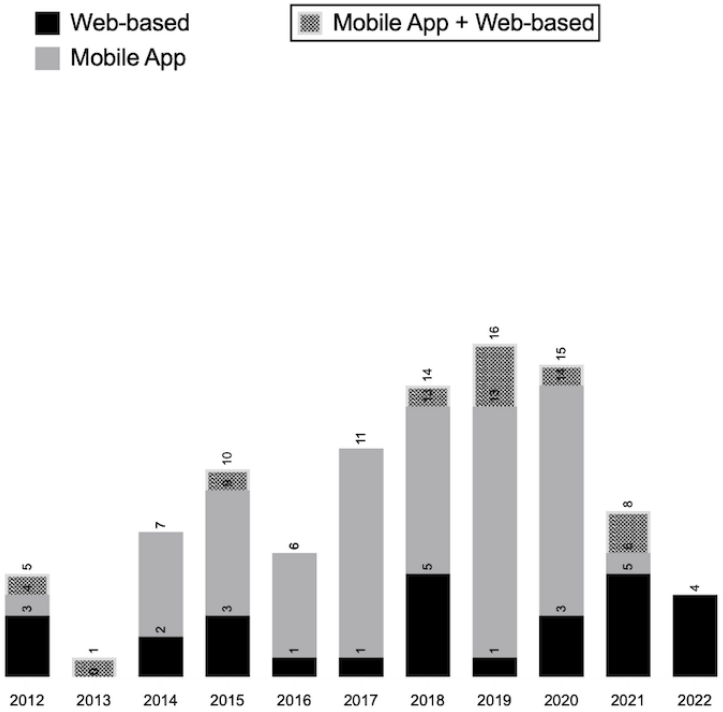

Outcome Measures

The included studies reported various outcomes measures. The most common were grouped as health care use, complications, and wellness outcomes.

Health Care Use

A total of 21% (20/97) of the studies reported on readmission rates [36,52-54,57,58,60,66,67,70,80,81,86,87,90,94,102,108,111,118]. Of the 20 studies, only 3 (15%) [58,102,108] showed a significant improvement, 10 (50%) reported their results as nonsignificant [36,52,60,67,80,81,86,90,94,118], 1 (5%) reported mixed results [66], and 6 (30%) [53,54,57,70,87,111] did not perform a statistical test.

In total, 8% (8/97) of the studies [36,52,60,80,81,94,108,118] reported on emergency department visits. Of the 8 studies, 2 (25%) [36,108] showed a significant improvement, whereas the remaining 6 (75%) showed no significant improvements.

Complications

A total of 8 studies [28,37,52,53,61,80,81,91] reported on complication rates, and only 1 (12%) [91] showed significant improvement in the complication rates compared to the control group; 3 (38%) [52,80,81] showed nonsignificant results, whereas the other 4 (50%) [28,37,53,61] did not perform a statistical test (Table 2).

XSL•FO
RenderX

Table 2. Studies reporting on health care use and complications.

Study	Name of DHI[a]	Participants	Readmission	ED[b] visits	Complications	Direction and magnitude of effect
Agri et al [28], 2020	Maela	Intervention: n=43	—[c]	—	NA[d]	Among the 43 patients, the app detected 12 adverse events, and 10 (83%) were handled through the app.
Ben-Ali et al [36], 2021	SeamlessMD	Intervention: n=1108	NS[e]	Significant	—	ED visits: negative, *P*=.03, and magnitude not reported; 30-day readmissions: negative, NS, and magnitude not reported
Birkhäuser et al [37], 2020	Cellphone-based health care app	Intervention: n=18	—	—	NA	In total, 2 patients required readmission within the study period of 90 days because of postoperative complications.
Eustache et al [52], 2023	Same-day discharge mHealth[f] app (CareSense)	Intervention: n=94; control: n=256	NS	NS	NS	30-day complications: negative, *P*=.18, and magnitude not reported; 30-day ED visits: no difference and *P*=.59; readmissions: positive, *P*=.35, and magnitude not reported
Felbaum et al [53], 2018	TrackMyRecovery	Intervention: n=56	NA	—	NA	There was 1 postoperative complication. There were no readmissions.
Ganapathy et al [54], 2017	Patient Buddy	Patients: n=40; caregivers: n=40	NA	—	—	A total of 17 patients (42.5%) were readmitted within 30 days.
Habib et al [57], 2021	Medication adherence mobile app	Intervention: n=23; control: n=26	NA	—	—	Hospital readmissions: negative (8.7% for the intervention group vs 15.4% for the control group); ED visits: positive (21.7% for the intervention group vs 19.2% for the control group)
Hägglund et al [58], 2015	Home intervention system (OPTILOGG[g])	Intervention: n=32; control: n=40	Significant	—	—	HF[h]-related days in hospital (readmissions): negative, *P*<.005, and magnitude not reported
Heuser et al [60], 2021	SeamlessMD	Intervention: n=396; control: n=458	NS	NS	—	ED visits without subsequent readmission: no difference and *P*=.65; ED visits with subsequent readmission: no difference and *P*=.99; readmissions: no difference and *P*=.97
Heyworth et al [61], 2014	SMMRT[i]	Intervention: n=60	—	—	NA	23 potential adverse drug events observed
Indraratna et al [66], 2022	TeleClinical Care	Intervention: n=81; control: n=83	Mixed	—	—	Unplanned 30-day readmissions: no difference, *P*=.97, and magnitude not reported; total readmissions at 6 months: negative, *P*=.02, and magnitude not reported; cardiac readmissions at 6 months: negative, *P*=.03, and magnitude not reported
Johnson et al [67], 2022	HF-SMART[j]	Intervention: n=16; control: n=15	NS	—	—	30-day readmissions: positive, *P*=.65, and magnitude not reported; 90-day readmissions: positive, *P*=.70, and magnitude not reported
Keng et al [70], 2020	Home to Stay digital program after colorectal surgery	Intervention: n=106	NA	—	—	The 30-day readmission rate was 6% and lower than the 30-day readmission rate of 18% reported for the 4 months before the start of the study.
Lee et al [80], 2022	Mobile app follow-up for same-day discharge	Intervention: n=48; control: n=73	NS	NS	NS	30-day complications: positive, *P*=.81, and magnitude not reported; 30-day ED visits: positive, *P*=.66, and magnitude not reported; 30-day readmissions: positive, *P*=.68, and magnitude not reported
Lee et al [81], 2022	mHealth remote postdischarge monitoring	Intervention: n=70; control: n=35	NS	NS	NS	ED visits: no difference and *P*>.99; readmissions: negative, *P*=.37, and magnitude not reported; incidence of 30-day complications: negative, *P*=.58, and magnitude not reported

XSL•FO
RenderX

Study	Name of DHI[a]	Participants	Readmission	ED[b] visits	Complications	Direction and magnitude of effect
María Gómez et al [86], 2022	mHealth app for patients with type 2 diabetes transitioning from inpatient to outpatient care	Intervention: n=41; control: n=45	NS	—	—	Hospitalization for diabetes: negative, *P*=.06, and magnitude not reported
Marvel et al [87], 2021	Acute myocardial infarction DHI	Intervention: n=200; control: n=864	NA	—	—	Risk of readmission within 30 days after discharge: negative, *P*=.02, and magnitude not reported
Paruchuri et al [90], 2021	Wellframe	Intervention: n=118; historical control group: n=343	NS	—	—	All-cause readmission within 30 days: negative, *P*=.70, and magnitude not reported; all-cause readmission within 90 days: negative, *P*=.39, and magnitude not reported
Peng et al [91], 2022	Mobile continuous nursing platform	Intervention: n=47; control: n=47	—	—	Significant	Total complication rates: negative, *P*<.05, and magnitude not reported
Pooni et al [94], 2022	Home to Stay app	Intervention: n=128; control: n=125	NS	NS	—	30-day ED visits: negative, *P*=.49, and magnitude not reported; 30-day readmissions: negative, *P*=.55, and magnitude not reported
Schenkel et al [102], 2020	ActiCare	Intervention: n=28; control: n=28	Significant	—	—	Hospital readmissions (events): negative, *P*<.001, and magnitude not reported
Stapler et al [108], 2022	St. Joe's Health App	Patients (preintervention group: n=1052; postintervention group: n=668)	Significant	Significant	—	Readmissions: negative, *P*<.001, and magnitude not reported; ED visit rate: negative, *P*<.001, and magnitude not reported
Symer et al [111], 2017	Gastrointestinal mHealth app	Intervention: n=31	NA	—	—	One patient was readmitted.
Visperas et al [118], 2021	JointCOACH	Intervention: n=204; control: n=195	NS	NS	—	ED visits: no difference and NS; readmissions: no difference and NS

[a]DHI: digital health intervention.

[b]ED: emergency department.

[c]Missing data or not applicable.

[d]NA: not available.

[e]NS: nonsignificant.

[f]mHealth: mobile health.

[g]OPTILOGG home intervention system.

[h]HF: heart failure.

[i]SMMRT: Secure Messaging for Medication Reconciliation Tool.

[j]HF-SMART: Heart Failure Self-Management And Readmission Prevention Technique.

Wellness Outcomes

A total of 14% (14/97) of the studies reported on quality of life [31,45,48,50,59,65,67,69,82,85,91,97,109,120]. Of the 14 studies, 8 (57%) [48,50,65,69,85,91,97,120] showed a significant improvement, whereas the other 6 (43%) reported their results to be nonsignificant. In total, 8 studies [31,51,68,76,82,91,109,115] were conducted on self-care, and 4 (50%) [82,91,109,115] showed significant results. In addition, 6 studies [31,51,65,94,119,122] reported on mental health outcomes, with 4 (67%) [65,94,119,122] showing significant improvements. A total of 8 studies [29,41,42,50,97,109,114,123] were identified for physical activity, with 5 (62%) [41,97,109,114,123] showing significant results (Table 3).

XSL•FO
RenderX

Table 3. Wellness outcomes.

Study	Name of DHI[a]	Participants	Quality of life	Self-care	Mental health	Physical activity	Direction and magnitude of effect
Antypas and Wangberg [29], 2014	Internet- and mobile-based tailored intervention to enhance maintenance of physical activity after cardiac rehabilitation	Intervention: n=7; control: n=12	__[b]	—	—	Mixed	Physical activity 1 month after discharge: positive, Kolmogorov-Smirnov Z=0.823, and P=.38; physical activity 3 months after discharge: positive, Kolmogorov-Smirnov Z=1.397, and P=.02
Athilingam et al [31], 2017	Mobile app to improve self-care behaviors and quality of life for patients with HF[c]	Intervention: n=9; control: n=90	NS[d]	Mixed	NS	—	Self-care maintenance: positive, t_{11}=0.083 and P=.93; self-care management: positive, t_{11}=3.38 and P=.01; self-care confidence: positive, t_{11}=2.53 and P=.28; depression: negative, t_{11}=1.97 and P=.07; quality of life: negative, t_{11}=−1.43 and P=.18
Cheng et al [41], 2022	Mobile app for home-based rehabilitation after hip fracture	Intervention: n=19; control: n=20	—	—	—	Significant	First-month exercise adherence: positive, P=.03, and magnitude not reported
Cox et al [42], 2015	ActivOnline	Intervention: n=10	—	—	—	NA[e]	Participants recorded a mean of 35 (range 15-57) physical activity sessions during the intervention period, equating to a mean of 4 recorded sessions of physical activity each week.
De Batlle et al [45], 2021	CONNECARE	Intervention: n=48; control: n=28	NS	—	—	—	Quality of life (SF-12[f]): positive, P=.10, and magnitude not reported
Devito Dabbs et al [48], 2016	Pocket PATH[g]	Intervention: n=99; control: n=102	Significant	—	—	—	Self-care: positive, group effect size=1.67, and P=.59
Duan et al [50], 2018	Health behavior intervention for patients with coronary heart disease through the web	Intervention: n=44; control: n=39	Significant	—	—	NS	Quality of life: positive, $F_{1, 79}$=16.36, and P<.001; physical activity: positive, $F_{1, 81}$=1.33, and P=.25
Dukeshire et al [51], 2012	Website tailored to women recovering at home after hysterectomy	Intervention: n=31	—	—	NA	—	The website reduced anxiety and worry for patients.
Heiney et al [59], 2020	Healthy Heart	Intervention: n=12	NS	NS	—	—	Quality of life: positive, P=.15, and magnitude not reported; Self-Care of Heart Failure Index—maintenance: difference score=9.37 and P=.15; Self-Care of Heart Failure Index—management: difference score=15.00 and not applicable (presample too small); Self-Care of Heart Failure Index—confidence: difference score=7.04 and P=.17
İlaslan and Özer [65], 2022	Web-based training and follow-up for patients with HF	Intervention: n=32; control: n=32	Significant	—	Significant	—	Quality of life—LVD-36[h]: negative, F=77.01, and P<.001
Johnson et al [67], 2022	HF-SMART[i]	Intervention: n=16; control: n=15	NS	—	—	—	Quality of life at 30 days: positive, P=.09, and magnitude not reported; quality of life at 90 days; negative, P=.10, and magnitude not reported
Kang et al [68], 2022	Web-based discharge education program	Intervention: n=43; control: n=42	—	NS	—	—	Self-care ability over time: positive, $F_{1, 60}$=8.934, and P=.004 (significant); self-care ability—group and time interaction: positive, $F_{1, 60}$=3.007, and P=.09
Kargar et al [69], 2020	Self-care educational mobile app for burns	Intervention: n=30; control: n=30	Significant	—	—	—	Quality of life: positive, P<.001, and magnitude not reported

XSL•FO
RenderX

Study	Name of DHI[a]	Participants	Quality of life	Self-care	Mental health	Physical activity	Direction and magnitude of effect
Kooij et al [76], 2021	Self-management app for high-risk patients with chronic obstructive pulmonary disease	Patients: n=39	—	NS	—	—	Self-management knowledge and coping: positive, *P*=.75, and magnitude not reported
Liu et al [82], 2021	Together	Intervention: n=49; control: n=49	NS	Significant	—	—	Self-efficacy: positive, *F*=8.506, and *P*=.004; quality of life: positive, *F*=0.082, and *P*=.78
Lyu et al [85], 2021	Nurse-led web-based transitional care program	Intervention: n=58; control: n=58	Significant	—	—	—	Quality of life: positive, *d*=0.52, and *P*<.01; self-efficacy: positive, *d*=0.50, and *P*<.01
Peng et al [91], 2022	Mobile continuous nursing platform	Intervention: n=47; control: n=47	Significant	Significant	—	—	Self-care ability: positive, *P*<.05, and magnitude not reported; quality of life (SF-36[j]): positive, *P*<.05, and magnitude not reported
Pooni et al [94], 2022	Postdischarge Home to Stay mobile app	Intervention: n=128; control: n=125	—	—	Significant	—	Feeling worried or anxious: negative and *P*<.001
Reid et al [97], 2012	CardioFit internet-based expert system	Intervention: n=115; control: n=118	Significant	—	—	Significant	Pedometer-measured steps per day: positive, *F*=5.226, and *P*=.02; heart disease health-related quality of Life (27-item MacNew instrument): positive, *F*=1.785, and *P*=.11
Su and Yu [109], 2021	NeCR[k] system	Intervention: n=73; control: n=73	NS	Significant	—	Significant	Mean daily steps 6 weeks after the intervention: positive, *P*=.02, and magnitude not reported; mean daily steps 12 weeks after the intervention: positive, *P*=.006, and magnitude not reported; self-efficacy: positive, *P*=.005, and magnitude not reported; MacNew health-related quality of life, positive, *P*=.06, and magnitude not reported
Torri et al [114], 2018	CRMP[l]	Intervention: n=26; control: n=27	—	—	—	Significant	Self-reported physical activity: positive, *P*=.35, and magnitude not reported
Van den Berg et al [115], 2016	CARE4STROKE	Intervention: n=31; control: n=32	—	Significant	—	—	Self-efficacy: positive, *P*=.008, and magnitude not reported
Vloothuis et al [119], 2019	CARE4STROKE digital intervention	Intervention: n=32; control: n=34	—	—	Significant	—	Patient anxiety: negative, *P*=.02, and magnitude not reported; caregiver depression: negative, *P*=.003, and magnitude not reported
Vonk Noordegraaf et al [120], 2014	Personalized eHealth program after gynecological surgery	Intervention: n=110; control: n=105	Significant	—	—	—	Quality of life: positive, between-group mean total score difference=30 (95% CI 4-57), and *P*=.02 (significant)
Wang et al [122], 2018	Stoma home care mobile app	Intervention: n=100; control: n=103	—	—	Significant	—	Psychosocial adjustment (1-, 3-, and 6-month follow-ups): positive, *F*=81.21, and *P*<.001

XSL•FO
RenderX

Study	Name of DHI[a]	Participants	Quality of life	Self-care	Mental health	Physical activity	Direction and magnitude of effect
Werhahn et al [123], 2019	CPMP[m]	Patients: n=10	—	—	—	Significant	Mean daily step count: positive, $P<.001$, and magnitude not reported

[a]DHI: digital health intervention.
[b]Missing data or not applicable.
[c]HF: heart failure.
[d]NS: nonsignificant.
[e]NA: not available.
[f]SF-12: 12-Item Short Form Health Survey.
[g]PATH: Personal Assistant for Tracking Health.
[h]LVD-36: left ventricular dysfunction questionnaire.
[i]HF-SMART: Heart Failure Self-Management And Readmission Prevention Technique.
[j]SF-36: 36-Item Short Form Health Survey.
[k]NeCR: nurse-led eHealth cardiac rehabilitation.
[l]CRMP: cardiac rehabilitation maintenance program.
[m]CPMP: cardio patient monitoring platform.

Patient, Caregiver, and Health Care Provider Barriers

Eight unique barriers were identified in the included studies: (1) lack of interest (13/97, 13%) [28,34,56,63,64,70,76,89,95,106,107,118,124], (2) time constraints (10/97, 10%) [44,54,56,76,78,79,96,106,109,115], (3) technological issues (7/97, 7%) [41,54,62,79,88,96,111], (4) usability issues (7/97, 7%) [42,62,71,76,96,107,117], (5) language barrier (4/97, 4%) [28,66,89,96], (6) irrelevant content of DHIs (3/97, 3%) [96,106,117], (7) lack of comfort (3/97, 3%) [78,115,117], and (8) lack of support and engagement (1/97, 1%) [106].

Patient, Caregiver, and Health Care Provider Enablers

Seven unique enablers were identified: (1) ability to use the DHI (17/97, 18%) [28,36,62,64,67,70,76,78,80,81,89,96,99,101,111,116,117], (2) ease of use (11/97, 11%) [49,55,57,58,68,76,85,101,105,111,121], (3) ability to collaborate with patients (1/97, 1%) [44], (4) caregiver support (1/97, 1%) [115], (5) confidence in the technology (1/97, 1%) [51], (6) convenience of using the DHIs (1/97, 1%) [62], and (7) participation in the development and implementation processes (1/97, 1%) [45] (Multimedia Appendix 5 [10,28,31,33,34,36,38-42,44,45,49,51,54-59,61-68,70,71,75,76,78-82,85,87-90,93,95,96,99,101,105-107,109,111,112,115-118,121,123,124]).

Conflicting Themes

In total, 3 themes were identified as both barriers and enablers. Many studies (15/97, 15%) reported that patients or caregivers were limited by their access to technology or the internet (barrier) [28,36,51,54,56,61,63,70,75,90,99,103,107,111,118], whereas other studies (28/97, 29%) reported that the patients or caregivers had access to these resources (enabler) [28,39,42,51,54,58,61,64,65,67,70,75,76,79-82,90,93,95,99,101,105,109,112,116,118,121]. Similarly, a few studies (4/97, 4%) reported that participants had difficulty with understanding the DHIs (barrier) [28,49,64,117], whereas other studies (19/97, 20%) reported that the DHIs were easy to understand (enabler) [31,33,34,38,49,55,57,58,68,76,80,81,85,87,93,101,105,111,121]. Finally, a recurring theme that acted as both a barrier (11/97, 11%) [31,36,59,63,64,89,99,101,107,112,118] and an enabler (17/97, 18%) [39,40,42,58,59,64,65,76,78,82,90,99,101,109,116,118,123] was whether the participants had digital literacy to use the DHIs.

Discussion

Summary

In this scoping review, we summarized the current evidence on platform-based patient-clinician DHIs specific to hospital-to-home care transitions and the reported barriers to and enablers of the uptake and implementation of these platform-based patient-clinician DHIs.

Mobile Apps Versus Web-Based Platforms

Most of the included studies used either a mobile app (59/97, 61%) or a combination of a mobile app and a web-based platform (10/97, 10%). Apps are unique because they are software programs that have been developed to run on a mobile device and are tailored to achieve a specific goal [125]. This is interesting because many studies have used other digital health tools, including e-charts [126], telehealth [127], and monitoring devices [128]. This trend may illustrate the benefits of using a mobile app over other types of digital health tools. This can include convenience, such as portability; effective communication; using a point of care for many different purposes; and immediate up-to-date information, guidelines, or medical literature [129]. Mobile apps have multiple uses in health care and have demonstrated numerous benefits, such as improved accuracy of patient documentation, improved workflow patterns or efficiency, and increased productivity of health care providers [129]. More specifically, the continuity of care between the hospital and the home.

Effectiveness of Platform-Based DHIs

The included studies had a broad range of outcome measures, and overall, these outcomes showed mixed results. The studies on platform-based DHIs did not show a significant improvement

XSL•FO
RenderX

in readmission rates (only 3/20, 15% showed significance), emergency department visits (only 2/8, 25% showed significance), or complication rates (only 1/8, 12% showed significance). However, the studies reported promising results for quality of life (8/14, 57% of the studies), self-care (4/8, 50% of the studies), mental health (4/6, 67% of the studies), and physical activity (5/8, 62% of the studies). Further research is needed to better plan and evaluate the overall effectiveness of these specific DHIs by clearly linking outcomes with specific interventions.

Barriers to and Enablers of the Uptake and Implementation of Platform-Based DHIs

The most prominent barriers were lack of interest and time constraints to use the DHIs, and the most prominent enablers reported were the ability to use the DHIs and their ease of use. This reveals the importance of simple, user-friendly DHIs as the patient's confidence in using them will determine how engaged they are throughout the intervention. Another important factor that plays a role in whether DHIs will be successful is whether the patient has access to the proper technological resources. This came up as a prominent barrier if they lacked the appropriate resources or as an enabler if they possessed what they needed. This reveals an important factor when considering the implementation of DHIs as the target population must have access to the correct resources to allow the intervention to take place.

Strengths and Limitations

Our study had several strengths and limitations. We designed an in-depth a priori protocol. The search strategy was developed and peer reviewed by a research librarian with extensive knowledge of scoping and systematic review methodologies. This scoping review was unique in that it specifically examined platform-based patient-clinician DHIs and not all types of DHIs. This allowed us to examine interventions that implemented elements that may promote more patient engagement, foster better communication between patients and health care providers, and integrate everything needed into one convenient program. The wide variety of studies included in this review led to a wide range of outcomes. We focused on the outcomes of health care use, complications, and wellness during the transition from hospital to home. Some other outcomes were excluded as they were specific to the disease or procedure performed. In addition, this review did not limit the inclusion to one type of health condition. However, this allowed us to examine DHIs across multiple areas of research and evaluate barriers to and enablers of DHI implementation across health conditions.

We identified a substantial body of literature on platform-based DHIs and their role in supporting patient care transitions from hospital to home. Most studies (95/97, 98%) primarily focused on patients' use of these DHIs. While patients are central to health care delivery, it is equally important to evaluate the effectiveness of platform-based DHIs from the health care providers' perspective. If these systems are not user-friendly for providers, widespread adoption is unlikely. Therefore, a deeper understanding of how providers interact with DHIs is essential for their successful implementation.

The transition of patients from hospital to home is a critical process that must be carried out safely and efficiently. However, this process is inherently complex due to factors such as unclear provider roles, suboptimal communication, and the patient's ability to manage their own care [130]. When transitions are not carried out effectively, patient care can be compromised, leading to negative outcomes [131]. Platform-based DHIs offer a promising solution to help streamline care during this vulnerable period, potentially improving the quality and safety of transitions. The findings of this work can inform future work on DHIs and, more specifically, the "MyPath to Home" DHI previously piloted for the population with hip fracture during their transition from hospital to home [33].

Conclusions

There is a lot of potential for using DHIs for care transitions; however, the specific elements that will improve patient outcomes need to be further explored. Specifically, further work is needed to involve all key stakeholders in the design, development, and implementation of these DHIs and understand their effectiveness to embed them in practice more broadly.

Acknowledgments

The authors thank Valentina Ly, master of library and information studies (Health Sciences Library, University of Ottawa), for peer review of the MEDLINE search strategy. This work received no funding.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Concept and design were conducted by CB, SP, SV, ALSFdM, GMdML, and AH. Data acquisition and analysis were conducted by RP, ATK, and CB. Drafting of the manuscript was conducted by RP and CB. All authors critically revised the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 509 KB - jmir_v26i1e55753_app1.pdf]

Multimedia Appendix 2
Search strategies.
[DOCX File , 38 KB - jmir_v26i1e55753_app2.docx]

Multimedia Appendix 3
List of excluded studies.
[XLS File (Microsoft Excel File), 87 KB - jmir_v26i1e55753_app3.xls]

Multimedia Appendix 4
Description of the interventions.
[DOCX File , 96 KB - jmir_v26i1e55753_app4.docx]

Multimedia Appendix 5
Barriers and enablers.
[DOCX File , 32 KB - jmir_v26i1e55753_app5.docx]

References

1. Enhancing the continuum of care report of the avoidable hospitalization advisory panel, submitted to the Ministry of Health and Long-Term Care. Ontario Ministry of Health and Long-Term Care. 2011 Nov. URL: http://www.health.gov.on.ca/en/common/ministry/publications/reports/baker_2011/baker_2011.pdf [accessed 2023-05-11]
2. Krumholz HM. Post-hospital syndrome--an acquired, transient condition of generalized risk. N Engl J Med 2013 Jan 10;368(2):100-102 [FREE Full text] [doi: 10.1056/NEJMp1212324] [Medline: 23301730]
3. Forster AJ, Clark HD, Menard A, Dupuis N, Chernish R, Chandok N, et al. Adverse events among medical patients after discharge from hospital. CMAJ 2004 Feb 03;170(3):345-349 [FREE Full text] [Medline: 14757670]
4. Cook RI, Render M, Woods DD. Gaps in the continuity of care and progress on patient safety. BMJ 2000 Mar 18;320(7237):791-794 [FREE Full text] [doi: 10.1136/bmj.320.7237.791] [Medline: 10720370]
5. Backman C, Stacey D, Crick M, Cho-Young D, Marck PB. Use of participatory visual narrative methods to explore older adults' experiences of managing multiple chronic conditions during care transitions. BMC Health Serv Res 2018 Jun 20;18(1):482 [FREE Full text] [doi: 10.1186/s12913-018-3292-6] [Medline: 29925369]
6. Ghazzawi A, Kuziemsky C, O'Sullivan T. Using a complex adaptive system lens to understand family caregiving experiences navigating the stroke rehabilitation system. BMC Health Serv Res 2016 Oct 01;16(1):538 [FREE Full text] [doi: 10.1186/s12913-016-1795-6] [Medline: 27716177]
7. Wayne N, Perez DF, Kaplan DM, Ritvo P. Health coaching reduces HbA1c in type 2 diabetic patients from a lower-socioeconomic status community: a randomized controlled trial. J Med Internet Res 2015 Oct 05;17(10):e224 [FREE Full text] [doi: 10.2196/jmir.4871] [Medline: 26441467]
8. Wolf A, Fors A, Ulin K, Thorn J, Swedberg K, Ekman I. An eHealth diary and symptom-tracking tool combined with person-centered care for improving self-efficacy after a diagnosis of acute coronary syndrome: a substudy of a randomized controlled trial. J Med Internet Res 2016 Feb 23;18(2):e40 [FREE Full text] [doi: 10.2196/jmir.4890] [Medline: 26907584]
9. Wikström L, Schildmeijer K, Nylander EM, Eriksson K. Patients' and providers' perspectives on e-health applications designed for self-care in association with surgery - a scoping review. BMC Health Serv Res 2022 Mar 23;22(1):386 [FREE Full text] [doi: 10.1186/s12913-022-07718-8] [Medline: 35321707]
10. Burke RE, Kripalani S, Vasilevskis EE, Schnipper JL. Moving beyond readmission penalties: creating an ideal process to improve transitional care. J Hosp Med 2013 Feb 26;8(2):102-109 [FREE Full text] [doi: 10.1002/jhm.1990] [Medline: 23184714]
11. Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R. Continuity of care: a multidisciplinary review. BMJ 2003 Nov 22;327(7425):1219-1221 [FREE Full text] [doi: 10.1136/bmj.327.7425.1219] [Medline: 14630762]
12. Elbert NJ, van Os-Medendorp H, van Renselaar W, Ekeland AG, Hakkaart-van Roijen L, Raat H, et al. Effectiveness and cost-effectiveness of ehealth interventions in somatic diseases: a systematic review of systematic reviews and meta-analyses. J Med Internet Res 2014 Apr 16;16(4):e110 [FREE Full text] [doi: 10.2196/jmir.2790] [Medline: 24739471]
13. Zhang J, Yang M, Ge Y, Ivers R, Webster R, Tian M. The role of digital health for post-surgery care of older patients with hip fracture: a scoping review. Int J Med Inform 2022 Apr;160:104709 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104709] [Medline: 35144101]

XSL•FO
RenderX

14. Singh H, Tang T, Steele Gray C, Kokorelias K, Thombs R, Plett D, et al. Recommendations for the design and delivery of transitions-focused digital health interventions: rapid review. JMIR Aging 2022 May 19;5(2):e35929 [FREE Full text] [doi: 10.2196/35929] [Medline: 35587874]
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
16. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
17. Peters M, Godfrey C, Mclnerney P, Munn Z, Tricco A, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2024.
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Backman C, Papp S, Harley A, Tonjock Kolle A, Visintini S, Shah S, et al. Platform-based patient-clinician digital health interventions for care transitions: protocol for a scoping review. JMIR Res Protoc 2023 Apr 05;12:e42056 [FREE Full text] [doi: 10.2196/42056] [Medline: 37018041]
20. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016 Jul;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
21. Gonçalves-Bradley DC, J Maria AR, Ricci-Cabello I, Villanueva G, Fønhus MS, Glenton C, et al. Mobile technologies to support healthcare provider to healthcare provider communication and management of care. Cochrane Database Syst Rev 2020 Aug 18;8(8):CD012927 [FREE Full text] [doi: 10.1002/14651858.CD012927.pub2] [Medline: 32813281]
22. Backman C, Papp S, Harley A, Skidmore B, Green M, Shah S, et al. Patient-clinician digital health interventions for the hip fracture population: a scoping review. BMC Health Serv Res 2023 Oct 02;23(1):1052 [FREE Full text] [doi: 10.1186/s12913-023-09784-y] [Medline: 37784118]
23. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2024-04-29]
24. Ritterband LM, Andersson G, Christensen HM, Carlbring P, Cuijpers P. Directions for the International Society for Research on Internet Interventions (ISRII). J Med Internet Res 2006 Sep 29;8(3):e23 [FREE Full text] [doi: 10.2196/jmir.8.3.e23] [Medline: 17032639]
25. Payne HE, Lister C, West JH, Bernhardt JM. Behavioral functionality of mobile apps in health interventions: a systematic review of the literature. JMIR Mhealth Uhealth 2015 Feb 26;3(1):e20 [FREE Full text] [doi: 10.2196/mhealth.3335] [Medline: 25803705]
26. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348(mar07 3):g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
28. Agri F, Hahnloser D, Demartines N, Hübner M. Gains and limitations of a connected tracking solution in the perioperative follow-up of colorectal surgery patients. Colorectal Dis 2020 Aug 27;22(8):959-966. [doi: 10.1111/codi.14998] [Medline: 32012423]
29. Antypas K, Wangberg SC. An internet- and mobile-based tailored intervention to enhance maintenance of physical activity after cardiac rehabilitation: short-term results of a randomized controlled trial. J Med Internet Res 2014 Mar 11;16(3):e77 [FREE Full text] [doi: 10.2196/jmir.3132] [Medline: 24618349]
30. Armstrong KA, Coyte PC, Brown M, Beber B, Semple JL. Effect of home monitoring via mobile app on the number of in-person visits following ambulatory surgery: a randomized clinical trial. JAMA Surg 2017 Jul 01;152(7):622-627 [FREE Full text] [doi: 10.1001/jamasurg.2017.0111] [Medline: 28329223]
31. Athilingam P, Jenkins B, Johansson M, Labrador M. A mobile health intervention to improve self-care in patients with heart failure: pilot randomized control trial. JMIR Cardio 2017 Aug 11;1(2):e3 [FREE Full text] [doi: 10.2196/cardio.7848] [Medline: 31758759]
32. Avci YD, Gozum S. The frequency of utilizing the supportive web site by stroke patients' caregivers after discharge. Int J Caring Sci 2018:1499-1508 [FREE Full text]
33. Backman C, Harley A, Kuziemsky C, Mercer J, Peyton L. MyPath to home web-based application for the geriatric rehabilitation program at Bruyère continuing care: user-centered design and feasibility testing study. JMIR Form Res 2020 Sep 14;4(9):e18169 [FREE Full text] [doi: 10.2196/18169] [Medline: 32924953]
34. Bäcker HC, Wu CH, Schulz MR, Weber-Spickschen TS, Perka C, Hardt S. App-based rehabilitation program after total knee arthroplasty: a randomized controlled trial. Arch Orthop Trauma Surg 2021 Sep;141(9):1575-1582 [FREE Full text] [doi: 10.1007/s00402-021-03789-0] [Medline: 33547927]

XSL•FO
RenderX

35. Bauwens PH, Fayard JM, Tatar M, Abid H, Freychet B, Sonnery-Cottet B, et al. Evaluation of a smartphone application for self-rehabilitation after anterior cruciate ligament reconstruction during a COVID-19 lockdown. Orthop Traumatol Surg Res 2023 Feb;109(1):103342 [FREE Full text] [doi: 10.1016/j.otsr.2022.103342] [Medline: 35660080]
36. Ben-Ali W, Lamarche Y, Carrier M, Demers P, Bouchard D, El-Hamamsy I, et al. Use of mobile-based application for collection of patient-reported outcomes in cardiac surgery. Innovations (Phila) 2021 Dec 09;16(6):536-544. [doi: 10.1177/15569845211045677] [Medline: 34882492]
37. Birkhäuser FD, Moltzahn F, Huber PM, Zehnder JL, Flückiger S, Hasler D, et al. Intensified and standardized digital communication with cystectomy patients as a potentially simple and effective modality for early detection of postoperative complications: results from a pilot study. Eur Urol Open Sci 2020 Dec;22:3-8 [FREE Full text] [doi: 10.1016/j.euros.2020.08.005] [Medline: 34337472]
38. Blewer AL, Putt ME, McGovern SK, Murray AD, Leary M, Riegel B, CHIP Study Group. A pragmatic randomized trial of cardiopulmonary resuscitation training for families of cardiac patients before hospital discharge using a mobile application. Resuscitation 2020 Jul;152:28-35. [doi: 10.1016/j.resuscitation.2020.04.026] [Medline: 32376347]
39. Bouwsma EV, Bosmans JE, van Dongen JM, Brölmann HA, Anema JR, Huirne JA. Cost-effectiveness of an internet-based perioperative care programme to enhance postoperative recovery in gynaecological patients: economic evaluation alongside a stepped-wedge cluster-randomised trial. BMJ Open 2018 Jan 21;8(1):e017782 [FREE Full text] [doi: 10.1136/bmjopen-2017-017782] [Medline: 29358423]
40. Bouwsma EV, Huirne JA, van de Ven PM, Vonk Noordegraaf A, Schaafsma FG, Schraffordt Koops SE, et al. Effectiveness of an internet-based perioperative care programme to enhance postoperative recovery in gynaecological patients: cluster controlled trial with randomised stepped-wedge implementation. BMJ Open 2018 Jan 30;8(1):e017781 [FREE Full text] [doi: 10.1136/bmjopen-2017-017781] [Medline: 29382673]
41. Cheng KC, Lau KM, Cheng AS, Lau TS, Lau FO, Lau MC, et al. Use of mobile app to enhance functional outcomes and adherence of home-based rehabilitation program for elderly with hip fracture: a randomized controlled trial. Hong Kong Physiother J 2022 Dec;42(2):99-110 [FREE Full text] [doi: 10.1142/S101370252250010X] [Medline: 37560168]
42. Cox NS, Alison JA, Button BM, Wilson JW, Holland AE. Feasibility and acceptability of an internet-based program to promote physical activity in adults with cystic fibrosis. Respir Care 2015 Mar;60(3):422-429 [FREE Full text] [doi: 10.4187/respcare.03165] [Medline: 25425703]
43. Davis DE, Cox R, Patel MS, Lazarus M, Ramsey M, Namdari S. Successful outcomes achieved via web-based, home program after total shoulder arthroplasty. Arch Bone Jt Surg 2020 Nov;8(6):661-667 [FREE Full text] [doi: 10.22038/abjs.2020.42832.2164] [Medline: 33313345]
44. Davoody N, Hägglund M. Care professionals' perceived usefulness of eHealth for post-discharge stroke patients. Stud Health Technol Inform 2016;228:589-593. [Medline: 27577452]
45. de Batlle J, Vargiu E, Torres G, Massip M, Michel M, Matthes F, et al. Implementation of an integrated care platform for the management of complex chronic patients in Lleida, Spain. Am J Respir Crit Care Med 2019;199:A6246 [FREE Full text] [doi: 10.1164/ajrccm-conference.2019.199.1_meetingabstracts.a6246]
46. Debono B, Bousquet P, Sabatier P, Plas JY, Lescure JP, Hamel O. Postoperative monitoring with a mobile application after ambulatory lumbar discectomy: an effective tool for spine surgeons. Eur Spine J 2016 Nov 27;25(11):3536-3542. [doi: 10.1007/s00586-016-4680-4] [Medline: 27349754]
47. Debono B, Corniola MV, Pietton R, Sabatier P, Hamel O, Tessitore E. Benefits of enhanced recovery after surgery for fusion in degenerative spine surgery: impact on outcome, length of stay, and patient satisfaction. Neurosurg Focus 2019 Apr 01;46(4):E6. [doi: 10.3171/2019.1.FOCUS18669] [Medline: 30933923]
48. DeVito Dabbs A, Song MK, Myers BA, Li R, Hawkins RP, Pilewski JM, et al. A randomized controlled trial of a mobile health intervention to promote self-management after lung transplantation. Am J Transplant 2016 Jul;16(7):2172-2180 [FREE Full text] [doi: 10.1111/ajt.13701] [Medline: 26729617]
49. Dorsch MP, Farris KB, Rowell BE, Hummel SL, Koelling TM. The effects of the ManageHF4Life mobile app on patients with chronic heart failure: randomized controlled trial. JMIR Mhealth Uhealth 2021 Dec 07;9(12):e26185 [FREE Full text] [doi: 10.2196/26185] [Medline: 34878990]
50. Duan YP, Liang W, Guo L, Wienert J, Si GY, Lippke S. Evaluation of a web-based intervention for multiple health behavior changes in patients with coronary heart disease in home-based rehabilitation: pilot randomized controlled trial. J Med Internet Res 2018 Nov 19;20(11):e12052 [FREE Full text] [doi: 10.2196/12052] [Medline: 30455167]
51. Dukeshire S, Gilmour D, MacDonald N, MacKenzie K. Development and evaluation of a web site to improve recovery from hysterectomy. Comput Inform Nurs 2012 Mar;30(3):164-176. [doi: 10.1097/NCN.0b013e31823eb8f9] [Medline: 22266705]
52. Eustache J, Latimer EA, Liberman AS, Charlebois P, Stein B, Fiore JF, et al. A mobile phone app improves patient-physician communication and reduces emergency department visits after colorectal surgery. Dis Colon Rectum 2023 Jan 01;66(1):130-137. [doi: 10.1097/DCR.0000000000002187] [Medline: 34933314]
53. Felbaum DR, Stewart JJ, Anaizi AN, Sandhu FA, Nair MN, Voyadzis J. Implementation and evaluation of a smartphone application for the perioperative care of neurosurgery patients at an academic medical center: implications for patient

XSL•FO
RenderX

satisfaction, surgery cancelations, and readmissions. Oper Neurosurg (Hagerstown) 2018 Mar 01;14(3):303-311. [doi: 10.1093/ons/opx112] [Medline: 28541569]
54. Ganapathy D, Acharya C, Lachar J, Patidar K, Sterling RK, White MB, et al. The patient buddy app can potentially prevent hepatic encephalopathy-related readmissions. Liver Int 2017 Dec;37(12):1843-1851. [doi: 10.1111/liv.13494] [Medline: 28618192]
55. Gollish J, Pereira L, MacLeod A, Wainwright A, Kennedy D, Robarts S, et al. myHipandKnee: improving patient engagement and self-management through mobile technology. Healthc Q 2019 Jul 31;22(2):63-67. [doi: 10.12927/hcq.2019.25902] [Medline: 31556382]
56. Gunter RL, Fernandes-Taylor S, Rahman S, Awoyinka L, Bennett KM, Weber SM, et al. Feasibility of an image-based mobile health protocol for postoperative wound monitoring. J Am Coll Surg 2018 Mar;226(3):277-286 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2017.12.013] [Medline: 29366555]
57. Habib B, Buckeridge D, Bustillo M, Marquez SN, Thakur M, Tran T, et al. Smart About Meds (SAM): a pilot randomized controlled trial of a mobile application to improve medication adherence following hospital discharge. JAMIA Open 2021 Jul;4(3):ooab050 [FREE Full text] [doi: 10.1093/jamiaopen/ooab050] [Medline: 34345805]
58. Hägglund E, Lyngå P, Frie F, Ullman B, Persson H, Melin M, et al. Patient-centred home-based management of heart failure. Findings from a randomised clinical trial evaluating a tablet computer for self-care, quality of life and effects on knowledge. Scand Cardiovasc J 2015 Aug 04;49(4):193-199 [FREE Full text] [doi: 10.3109/14017431.2015.1035319] [Medline: 25968968]
59. Heiney SP, Donevant SB, Arp Adams S, Parker PD, Chen H, Levkoff S. A smartphone app for self-management of heart failure in older African Americans: feasibility and usability study. JMIR Aging 2020 Apr 03;3(1):e17142 [FREE Full text] [doi: 10.2196/17142] [Medline: 32242822]
60. Heuser J, Maeda A, Yang L, Masino C, Duggal S, Jackson T, et al. Impact of a mobile app to support home recovery of patients undergoing bariatric surgery. J Surg Res 2021 May;261:179-184. [doi: 10.1016/j.jss.2020.12.005] [Medline: 33444947]
61. Heyworth L, Paquin AM, Clark J, Kamenker V, Stewart M, Martin T, et al. Engaging patients in medication reconciliation via a patient portal following hospital discharge. J Am Med Inform Assoc 2014 Feb;21(e1):e157-e162 [FREE Full text] [doi: 10.1136/amiajnl-2013-001995] [Medline: 24036155]
62. Highland KB, Tran J, Edwards H, Bedocs P, Suen J, Buckenmaier CC. Feasibility of app-based postsurgical assessment of pain, pain impact, and regional anesthesia effects: a pilot randomized controlled trial. Pain Med 2019 Aug 01;20(8):1592-1599. [doi: 10.1093/pm/pny288] [Medline: 30726985]
63. Holzer H, Goodlev ER, Pearson JM, Engelman S, Sperber D, Dunn AS, et al. Engagement and outcomes with mobile health technology among patients hospitalized with acute venous thromboembolism. J Gen Intern Med 2022;17(2):5-14 [FREE Full text]
64. Houchen-Wolloff L, Orme M, Barradell A, Clinch L, Chaplin E, Gardiner N, et al. Web-based self-management program (SPACE for COPD) for individuals hospitalized with an acute exacerbation of chronic obstructive pulmonary disease: nonrandomized feasibility trial of acceptability. JMIR Mhealth Uhealth 2021 Jun 11;9(6):e21728 [FREE Full text] [doi: 10.2196/21728] [Medline: 34114960]
65. İlaslan E, Özer Z. Web-based training and telephone follow-up of patients with heart failure: randomized controlled trial. Comput Inform Nurs 2021 Sep 27;40(2):82-89. [doi: 10.1097/CIN.0000000000000833] [Medline: 34570004]
66. Indraratna P, Biswas U, McVeigh J, Mamo A, Magdy J, Vickers D, et al. A smartphone-based model of care to support patients with cardiac disease transitioning from hospital to the community (TeleClinical Care): pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Feb 28;10(2):e32554 [FREE Full text] [doi: 10.2196/32554] [Medline: 35225819]
67. Johnson AE, Routh S, Taylor CN, Leopold M, Beatty K, McNamara DM, et al. Developing and implementing an mHealth heart failure self-care program to reduce readmissions: randomized controlled trial. JMIR Cardio 2022 Mar 21;6(1):e33286 [FREE Full text] [doi: 10.2196/33286] [Medline: 35311679]
68. Kang E, Chaboyer W, Tobiano G, Gillespie B. Evaluating the feasibility of a web-based discharge education programme to improve general surgical patients' postdischarge recovery: a pilot randomised controlled trial. BMJ Open 2022 Feb 09;12(2):e054038 [FREE Full text] [doi: 10.1136/bmjopen-2021-054038] [Medline: 35140156]
69. Kargar N, Deldar K, Ahmadabadi A, Froutan R, Mazlom SR. Can a self-care educational mobile application improve the quality of life of victims with hand burns? A randomized controlled trial. Crescent J Med Biol Sci 2020;7(4):497-502 [FREE Full text]
70. Keng CJ, Goriawala A, Rashid S, Goldstein R, Schmocker S, Easson A, et al. Home to stay: an integrated monitoring system using a mobile app to support patients at home following colorectal surgery. J Patient Exp 2020 Dec 12;7(6):1241-1246 [FREE Full text] [doi: 10.1177/2374373520904194] [Medline: 33457571]
71. Kersey J, Kringle E, Setiawan IM, Parmanto B, Skidmore ER. Pilot RCT examining feasibility and disability outcomes of a mobile health platform for strategy training in inpatient stroke rehabilitation (iADAPT). Top Stroke Rehabil 2023 Jul 18;30(5):512-521 [FREE Full text] [doi: 10.1080/10749357.2022.2077522] [Medline: 35583268]
72. Khan D, Fjerbæk A, Andreasen JJ, Thorup CB, Dinesen B. Cardiac surgery patients' e-health literacy and their use of a digital portal. Health Educ J 2018 Feb 19;77(4):482-494. [doi: 10.1177/0017896918756435]

XSL•FO
RenderX

73. Khanwalkar AR, Shen J, Kern RC, Welch KC, Smith SS, Tan BK, et al. Utilization of a novel interactive mobile health platform to evaluate functional outcomes and pain following septoplasty and functional endoscopic sinus surgery. Int Forum Allergy Rhinol 2019 Apr 26;9(4):345-351. [doi: 10.1002/alr.22273] [Medline: 30586227]
74. Kim K, Pham D, Schwarzkopf R. Mobile application use in monitoring patient adherence to perioperative total knee arthroplasty protocols. Surg Technol Int 2016 Apr;28:253-260. [Medline: 27042789]
75. Knapp PW, Keller RA, Mabee KA, Pillai R, Frisch NB. Quantifying patient engagement in total joint arthroplasty using digital application-based technology. J Arthroplasty 2021 Sep;36(9):3108-3117 [FREE Full text] [doi: 10.1016/j.arth.2021.04.022] [Medline: 33965282]
76. Kooij L, Vos PJ, Dijkstra A, van Harten WH. Effectiveness of a mobile health and self-management app for high-risk patients with chronic obstructive pulmonary disease in daily clinical practice: mixed methods evaluation study. JMIR Mhealth Uhealth 2021 Feb 04;9(2):e21977. [doi: 10.2196/21977] [Medline: 33538699]
77. Kristjánsdóttir Ó, Fors EA, Eide E, Finset A, Stensrud TL, van Dulmen S, et al. A smartphone-based intervention with diaries and therapist-feedback to reduce catastrophizing and increase functioning in women with chronic widespread pain: randomized controlled trial. J Med Internet Res 2013 Jan 07;15(1):e5 [FREE Full text] [doi: 10.2196/jmir.2249] [Medline: 23291270]
78. Kummerow Broman K, Oyefule OO, Phillips SE, Baucom RB, Holzman MD, Sharp KW, et al. Postoperative care using a secure online patient portal: changing the (Inter)face of general surgery. J Am Coll Surg 2015 Dec;221(6):1057-1066 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2015.08.429] [Medline: 26453260]
79. Layton AM, Whitworth J, Peacock J, Bartels MN, Jellen PA, Thomashow BM. Feasibility and acceptability of utilizing a smartphone based application to monitor outpatient discharge instruction compliance in cardiac disease patients around discharge from hospitalization. Int J Telemed Appl 2014;2014:415868-415810 [FREE Full text] [doi: 10.1155/2014/415868] [Medline: 25574165]
80. Lee L, Eustache J, Baldini G, Liberman AS, Charlebois P, Stein B, et al. Enhanced recovery 2.0 - same day discharge with mobile app follow-up after minimally invasive colorectal surgery. Ann Surg 2022 Dec 01;276(6):e812-e818. [doi: 10.1097/SLA.0000000000004962] [Medline: 34091514]
81. Lee L, Eustache J, Tran-McCaslin M, Basam M, Baldini G, Rudikoff AG, et al. North American multicentre evaluation of a same-day discharge protocol for minimally invasive colorectal surgery using mHealth or telephone remote post-discharge monitoring. Surg Endosc 2022 Dec 13;36(12):9335-9344. [doi: 10.1007/s00464-022-09208-8] [Medline: 35419638]
82. Liu T, Xie S, Wang Y, Tang J, He X, Yan T, et al. Effects of app-based transitional care on the self-efficacy and quality of life of patients with spinal cord injury in China: randomized controlled trial. JMIR Mhealth Uhealth 2021 Apr 01;9(4):e22960 [FREE Full text] [doi: 10.2196/22960] [Medline: 33792555]
83. Li Y, Gong Y, Zheng B, Fan F, Yi T, Zheng Y, et al. Effects on adherence to a mobile app-based self-management digital therapeutics among patients with coronary heart disease: pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Feb 15;10(2):e32251 [FREE Full text] [doi: 10.2196/32251] [Medline: 34906924]
84. Lou VW, Cheng CY, Ng DK, Chan FH, Mo SS, Kung EK, et al. A mHealth-supported volunteer-assisted spiritual well-being intervention for discharged older patients: a tripartite collaboration. J Gerontol Soc Work 2023;66(2):189-207. [doi: 10.1080/01634372.2022.2068722] [Medline: 35477345]
85. Lyu QY, Huang JW, Li YX, Chen QL, Yu XX, Wang JL, et al. Effects of a nurse led web-based transitional care program on the glycemic control and quality of life post hospital discharge in patients with type 2 diabetes: a randomized controlled trial. Int J Nurs Stud 2021 Jul;119:103929. [doi: 10.1016/j.ijnurstu.2021.103929] [Medline: 33901941]
86. María Gómez A, Cristina Henao D, León Vargas F, Mauricio Muñoz O, David Lucero O, García Jaramillo M, et al. Efficacy of the mHealth application in patients with type 2 diabetes transitioning from inpatient to outpatient care: a randomized controlled clinical trial. Diabetes Res Clin Pract 2022 Jul;189:109948 [FREE Full text] [doi: 10.1016/j.diabres.2022.109948] [Medline: 35700926]
87. Marvel FA, Spaulding EM, Lee MA, Yang WE, Demo R, Ding J, et al. Digital health intervention in acute myocardial infarction. Circ Cardiovasc Qual Outcomes 2021 Jul;14(7):e007741 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.121.007741] [Medline: 34261332]
88. Metilda CJ, Sharma KK, Sinha AP, Agrawal D. Effectiveness of nurse-driven discharge teaching using mobile application for home-based health care practices among postoperative neurosurgical patients or caregivers in a tertiary care hospital, New Delhi: a randomized control study. Indian J Neurotrauma 2021 Mar 15;18(02):119-125. [doi: 10.1055/s-0041-1724143]
89. Park C, Otobo E, Ullman J, Rogers J, Fasihuddin F, Garg S, et al. Impact on readmission reduction among heart failure patients using digital health monitoring: feasibility and adoptability study. JMIR Med Inform 2019 Nov 15;7(4):e13353 [FREE Full text] [doi: 10.2196/13353] [Medline: 31730039]
90. Paruchuri K, Finneran P, Marston NA, Healy EW, Andreo J, Lynch R, et al. Outcomes of a smartphone-based application with live health-coaching post-percutaneous coronary intervention. EBioMedicine 2021 Oct;72:103593 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103593] [Medline: 34657825]
91. Peng Y, Wan H, Hu X, Xiong F, Cao Y. Internet+Continuous nursing mode in home nursing of patients with T-Tube after hepatolithiasis surgery. Comput Math Methods Med 2022 May 31;2022:9490483-9490489 [FREE Full text] [doi: 10.1155/2022/9490483] [Medline: 35685900]

XSL•FO
RenderX

92. Pickens R, Cochran A, Tezber K, Berry R, Bhattacharya E, Koo D, et al. Using a mobile application for real-time collection of patient-reported outcomes in hepatopancreatobiliary surgery within an ERAS® pathway. Am Surg 2019 Aug 01;85(8):909-917. [Medline: 32051068]
93. Ponder M, Ansah-Yeboah AA, Charalambous LT, Adil SM, Venkatraman V, Abd-El-Barr M, et al. A smartphone app with a digital care pathway for patients undergoing spine surgery: development and feasibility study. JMIR Perioper Med 2020 Oct 16;3(2):e21138 [FREE Full text] [doi: 10.2196/21138] [Medline: 33393924]
94. Pooni A, Brar MS, Anpalagan T, Schmocker S, Rashid S, Goldstein R, et al. Home to stay: a randomized controlled trial evaluating the effect of a postdischarge mobile app to reduce 30-day readmission following elective colorectal surgery. Ann Surg 2023 May 01;277(5):e1056-e1062. [doi: 10.1097/SLA.0000000000005527] [Medline: 35815882]
95. Pronk Y, Peters MC, Sheombar A, Brinkman JM. Effectiveness of a mobile eHealth app in guiding patients in pain control and opiate use after total knee replacement: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 13;8(3):e16415 [FREE Full text] [doi: 10.2196/16415] [Medline: 32167483]
96. Pugliese M, Ramsay T, Shamloul R, Mallet K, Zakutney L, Corbett D, et al. RecoverNow: a mobile tablet-based therapy platform for early stroke rehabilitation. PLoS One 2019;14(1):e0210725 [FREE Full text] [doi: 10.1371/journal.pone.0210725] [Medline: 30682076]
97. Reid RD, Morrin LI, Beaton LJ, Papadakis S, Kocourek J, McDonnell L, et al. Randomized trial of an internet-based computer-tailored expert system for physical activity in patients with heart disease. Eur J Prev Cardiol 2012 Dec;19(6):1357-1364. [doi: 10.1177/1741826711422988] [Medline: 21903744]
98. Requena M, Montiel E, Baladas M, Muchada M, Boned S, López R, et al. Farmalarm. Stroke 2019 Jul;50(7):1819-1824. [doi: 10.1161/STROKEAHA.118.024355] [Medline: 31167621]
99. Rian T, Sand K, Skogvoll E, Klepstad P, Wik TS. A web-based communication tool for postoperative follow-up and pain assessment at home after primary knee arthroplasty: feasibility and usability study. JMIR Form Res 2022 Apr 28;6(4):e34543 [FREE Full text] [doi: 10.2196/34543] [Medline: 35482392]
100. Rosner BI, Gottlieb M, Anderson WN. Accuracy of internet-based patient self-report of postdischarge health care utilization and complications following orthopedic procedures: observational cohort study. J Med Internet Res 2018 Jul 20;20(7):e10405 [FREE Full text] [doi: 10.2196/10405] [Medline: 30030212]
101. Saunders R, Seaman K, Emery L, Bulsara M, Ashford C, McDowall J, et al. Comparing an eHealth program (My Hip Journey) with standard care for total hip arthroplasty: randomized controlled trial. JMIR Rehabil Assist Technol 2021 Mar 03;8(1):e22944 [FREE Full text] [doi: 10.2196/22944] [Medline: 33656449]
102. Schenkel FA, Barr ML, McCloskey CC, Possemato T, O'Conner J, Sadeghi R, et al. Use of a bluetooth tablet-based technology to improve outcomes in lung transplantation: a pilot study. Am J Transplant 2020 Dec;20(12):3649-3657 [FREE Full text] [doi: 10.1111/ajt.16154] [Medline: 32558226]
103. Scheper H, Derogee R, Mahdad R, van der Wal R, Nelissen R, Visser L, et al. A mobile app for postoperative wound care after arthroplasty: ease of use and perceived usefulness. Int J Med Inform 2019 Sep;129:75-80. [doi: 10.1016/j.ijmedinf.2019.05.010] [Medline: 31445292]
104. Schneider MA, Howard KA. Using technology to enhance discharge teaching and improve coping for patients after stroke. J Neurosci Nurs 2017 Jun;49(3):152-156. [doi: 10.1097/JNN.0000000000000275] [Medline: 28346312]
105. Schubart J. An e-learning program to prevent pressure ulcers in adults with spinal cord injury: a pre- and post- pilot test among rehabilitation patients following discharge to home. Ostomy Wound Manage 2012 Oct;58(10):38-49 [FREE Full text] [Medline: 23037331]
106. Scott AR, Alore EA, Naik AD, Berger DH, Suliburk JW. Mixed-methods analysis of factors impacting use of a postoperative mHealth app. JMIR Mhealth Uhealth 2017 Feb 08;5(2):e11 [FREE Full text] [doi: 10.2196/mhealth.6728] [Medline: 28179215]
107. Siegel J, Edwards E, Mooney L, Smith C, Peel JB, Dole A, et al. A feasibility pilot using a mobile personal health assistant (PHA) app to assist stroke patient and caregiver communication after hospital discharge. Mhealth 2016 Aug 09;2:31 [FREE Full text] [doi: 10.21037/mhealth.2016.08.02] [Medline: 28293604]
108. Stapler SJ, Brockhaus KK, Battaglia MA, Mahoney ST, McClure AM, Cleary RK. A single-institution analysis of targeted colorectal surgery enhanced recovery pathway strategies that decrease readmissions. Dis Colon Rectum 2022 Jul 01;65(7):e728-e740. [doi: 10.1097/DCR.0000000000002129] [Medline: 34897213]
109. Su JJ, Yu DS. Effects of a nurse-led eHealth cardiac rehabilitation programme on health outcomes of patients with coronary heart disease: a randomised controlled trial. Int J Nurs Stud 2021 Oct;122:104040. [doi: 10.1016/j.ijnurstu.2021.104040] [Medline: 34333211]
110. Sureshkumar K, Murthy G, Natarajan S, Naveen C, Goenka S, Kuper H. Evaluation of the feasibility and acceptability of the 'Care for Stroke' intervention in India, a smartphone-enabled, carer-supported, educational intervention for management of disability following stroke. BMJ Open 2016 Feb 02;6(2):e009243 [FREE Full text] [doi: 10.1136/bmjopen-2015-009243] [Medline: 26839011]
111. Symer MM, Abelson JS, Milsom J, McClure B, Yeo HL. A mobile health application to track patients after gastrointestinal surgery: results from a pilot study. J Gastrointest Surg 2017 Sep;21(9):1500-1505. [doi: 10.1007/s11605-017-3482-2] [Medline: 28685388]

XSL•FO
RenderX

112. Timmers T, Janssen L, van der Weegen W, Das D, Marijnissen W, Hannink G, et al. The effect of an app for day-to-day postoperative care education on patients with total knee replacement: randomized controlled trial. JMIR Mhealth Uhealth 2019 Oct 21;7(10):e15323 [FREE Full text] [doi: 10.2196/15323] [Medline: 31638594]
113. Evaluating the impact of Meducation®, a SMART application to medication adherence: the need to evaluate EHR apps post-implementation. Healthcare Information and Management Systems Society (HIMSS). URL: https://login.proxy.bib.uottawa.ca/login?url=https://www.proquest.com/scholarly-journals/evaluating-impact-meducation®-smart-application/docview/2621676922/se-2?accountid=14701 [accessed 2024-04-29]
114. Torri A, Panzarino C, Scaglione A, Modica M, Bordoni B, Redaelli R, et al. Promotion of home-based exercise training as secondary prevention of coronary heart disease: a pilot web-based intervention. J Cardiopulm Rehabil Prev 2018 Jul;38(4):253-258. [doi: 10.1097/HCR.0000000000000316] [Medline: 29738378]
115. van den Berg M, Crotty M, Liu E, Killington M, Kwakkel G, van Wegen E. Early supported discharge by caregiver-mediated exercises and e-health support after stroke: a proof-of-concept trial. Stroke 2016 Jul;47(7):1885-1892. [doi: 10.1161/STROKEAHA.116.013431] [Medline: 27301941]
116. Venkatraman V, Ponder M, Gellad ZF, Lad SP, Christy E, Plichta R, et al. Feasibility study of a novel digital health platform for patients undergoing transcatheter aortic valve replacement. J Card Surg 2022 Jul 17;37(7):2017-2022. [doi: 10.1111/jocs.16515] [Medline: 35434823]
117. Vincent C, Wegier P, Chien V, Kurahashi AM, Ginsburg S, Molla Ghanbari H, et al. Qualitative evaluation of a novel educational tool to communicate individualized hip fracture prognostic information to patients and surrogates: my hip fracture (My-HF). Geriatr Orthop Surg Rehabil 2021;12:21514593211050513 [FREE Full text] [doi: 10.1177/21514593211050513] [Medline: 34707914]
118. Visperas AT, Greene KA, Krebs VE, Klika AK, Piuzzi NS, Higuera-Rueda CA. A web-based interactive patient-provider software platform does not increase patient satisfaction or decrease hospital resource utilization in total knee and hip arthroplasty patients in a single large hospital system. J Arthroplasty 2021 Jul;36(7):2290-6.e1. [doi: 10.1016/j.arth.2021.01.037] [Medline: 33581971]
119. Vloothuis JD, Mulder M, Nijland RH, Goedhart QS, Konijnenbelt M, Mulder H, et al. Caregiver-mediated exercises with e-health support for early supported discharge after stroke (CARE4STROKE): a randomized controlled trial. PLoS One 2019 Apr 8;14(4):e0214241 [FREE Full text] [doi: 10.1371/journal.pone.0214241] [Medline: 30958833]
120. Vonk Noordegraaf A, Anema JR, van Mechelen W, Knol DL, van Baal WM, van Kesteren PJ, et al. A personalised eHealth programme reduces the duration until return to work after gynaecological surgery: results of a multicentre randomised trial. BJOG 2014 Aug;121(9):1127-1136. [doi: 10.1111/1471-0528.12661] [Medline: 24511914]
121. Wang L, He L, Tao Y, Sun L, Zheng H, Zheng Y, et al. Evaluating a web-based coaching program using electronic health records for patients with chronic obstructive pulmonary disease in China: randomized controlled trial. J Med Internet Res 2017 Jul 21;19(7):e264 [FREE Full text] [doi: 10.2196/jmir.6743] [Medline: 28733270]
122. Wang Q, Zhao J, Huo X, Wu L, Yang L, Li J, et al. Effects of a home care mobile app on the outcomes of discharged patients with a stoma: a randomised controlled trial. J Clin Nurs 2018 Oct 10;27(19-20):3592-3602. [doi: 10.1111/jocn.14515] [Medline: 29775491]
123. Werhahn SM, Dathe H, Rottmann T, Franke T, Vahdat D, Hasenfuß G, et al. Designing meaningful outcome parameters using mobile technology: a new mobile application for telemonitoring of patients with heart failure. ESC Heart Fail 2019 Jun 13;6(3):516-525 [FREE Full text] [doi: 10.1002/ehf2.12425] [Medline: 30868756]
124. Willeit P, Toell T, Boehme C, Krebs S, Mayer L, Lang C, STROKE-CARD study group. STROKE-CARD care to prevent cardiovascular events and improve quality of life after acute ischaemic stroke or TIA: a randomised clinical trial. EClinicalMedicine 2020 Aug;25:100476 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100476] [Medline: 32954239]
125. Wallace S, Clark M, White J. 'It's on my iPhone': attitudes to the use of mobile computing devices in medical education, a mixed-methods study. BMJ Open 2012 Aug 24;2(4):e001099 [FREE Full text] [doi: 10.1136/bmjopen-2012-001099] [Medline: 22923627]
126. Griffin A, Skinner A, Thornhill J, Weinberger M. Patient portals: who uses them? What features do they use? And do they reduce hospital readmissions? Appl Clin Inform 2016;7(2):489-501 [FREE Full text] [doi: 10.4338/ACI-2016-01-RA-0003] [Medline: 27437056]
127. Kane LT, Thakar O, Jamgochian G, Lazarus MD, Abboud JA, Namdari S, et al. The role of telehealth as a platform for postoperative visits following rotator cuff repair: a prospective, randomized controlled trial. J Shoulder Elbow Surg 2020 Apr;29(4):775-783 [FREE Full text] [doi: 10.1016/j.jse.2019.12.004] [Medline: 32197766]
128. Ejofodomi O, Zara JM, Ofualagba G. MEDLINK: a low-cost, portable, verbally interactive and programmable remote patient monitoring (RPM) device. BMJ Innov 2020 Jun 30;6(4):151-158 [FREE Full text] [doi: 10.1136/bmjinnov-2019-000371]
129. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
130. Arbaje AI, Kansagara DL, Salanitro AH, Englander HL, Kripalani S, Jencks SF, et al. Regardless of age: incorporating principles from geriatric medicine to improve care transitions for patients with complex needs. J Gen Intern Med 2014 Jun;29(6):932-939 [FREE Full text] [doi: 10.1007/s11606-013-2729-1] [Medline: 24557511]

XSL•FO
RenderX

131. Earl T, Katapodis N, Schneiderman S. Care transitions. In: Hall KK, Shoemaker-Hunt S, Hoffman L, Richard S, Gall E, Schoyer E, et al, editors. Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices. New York, NY: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; Sep 2020:S1-S2.

Abbreviations

DHI: digital health intervention

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Mavragani; submitted 22.12.23; peer-reviewed by K Beatty, M Sadiq; comments to author 18.09.24; revised version received 16.10.24; accepted 07.11.24; published 30.12.24.

Please cite as:
Backman C, Papp R, Tonjock Kolle A, Papp S, Visintini S, Schaefer Ferreira de Mello AL, de Melo Lanzoni GM, Harley A
Platform-Based Patient-Clinician Digital Health Interventions for Care Transitions: Scoping Review
J Med Internet Res 2024;26:e55753
URL: https://www.jmir.org/2024/1/e55753
doi:10.2196/55753
PMID:

©Chantal Backman, Rosie Papp, Aurelie Tonjock Kolle, Steve Papp, Sarah Visintini, Ana Lúcia Schaefer Ferreira de Mello, Gabriela Marcellino de Melo Lanzoni, Anne Harley. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Applying AI to Structured Real-World Data for Pharmacovigilance Purposes: Scoping Review

Stella Dimitsaki[1], MSc; Pantelis Natsiavas[2], PhD; Marie-Christine Jaulent[1], PhD

[1]Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé - LIMICS, Inserm, Université Sorbonne Paris-Nord, Sorbonne Université, Paris, France

[2]Centre for Research and Development Hellas, Institute of Applied Biosciences, Thessaloniki, Greece

Corresponding Author:
Stella Dimitsaki, MSc
Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé - LIMICS
Inserm, Université Sorbonne Paris-Nord
Sorbonne Université
15 Rue de l'École de Médecine
Paris, 75006
France
Phone: 33 767968072
Email: Stella.Dimitsaki@etu.sorbonne-universite.fr

Abstract

Background: Artificial intelligence (AI) applied to real-world data (RWD; eg, electronic health care records) has been identified as a potentially promising technical paradigm for the pharmacovigilance field. There are several instances of AI approaches applied to RWD; however, most studies focus on unstructured RWD (conducting natural language processing on various data sources, eg, clinical notes, social media, and blogs). Hence, it is essential to investigate how AI is currently applied to structured RWD in pharmacovigilance and how new approaches could enrich the existing methodology.

Objective: This scoping review depicts the emerging use of AI on structured RWD for pharmacovigilance purposes to identify relevant trends and potential research gaps.

Methods: The scoping review methodology is based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology. We queried the MEDLINE database through the PubMed search engine. Relevant scientific manuscripts published from January 2010 to January 2024 were retrieved. The included studies were "mapped" against a set of evaluation criteria, including applied AI approaches, code availability, description of the data preprocessing pipeline, clinical validation of AI models, and implementation of trustworthy AI criteria following the guidelines of the FUTURE (Fairness, Universality, Traceability, Usability, Robustness, and Explainability)-AI initiative.

Results: The scoping review ultimately yielded 36 studies. There has been a significant increase in relevant studies after 2019. Most of the articles focused on adverse drug reaction detection procedures (23/36, 64%) for specific adverse effects. Furthermore, a substantial number of studies (34/36, 94%) used nonsymbolic AI approaches, emphasizing classification tasks. Random forest was the most popular machine learning approach identified in this review (17/36, 47%). The most common RWD sources used were electronic health care records (28/36, 78%). Typically, these data were not available in a widely acknowledged data model to facilitate interoperability, and they came from proprietary databases, limiting their availability for reproducing results. On the basis of the evaluation criteria classification, 10% (4/36) of the studies published their code in public registries, 16% (6/36) tested their AI models in clinical environments, and 36% (13/36) provided information about the data preprocessing pipeline. In addition, in terms of trustworthy AI, 89% (32/36) of the studies followed at least half of the trustworthy AI initiative guidelines. Finally, selection and confounding biases were the most common biases in the included studies.

Conclusions: AI, along with structured RWD, constitutes a promising line of work for drug safety and pharmacovigilance. However, in terms of AI, some approaches have not been examined extensively in this field (such as explainable AI and causal AI). Moreover, it would be helpful to have a data preprocessing protocol for RWD to support pharmacovigilance processes. Finally, because of personal data sensitivity, evaluation procedures have to be investigated further.

(J Med Internet Res 2024;26:e57824) doi:10.2196/57824

XSL•FO
RenderX

KEYWORDS

pharmacovigilance; drug safety; artificial intelligence; machine learning; real-world data; scoping review

Introduction

Background

Pharmacovigilance is defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem" [1]. Pharmacovigilance plays a crucial role in ensuring the safety of medications and protecting the health of patients because it mostly focuses on the identification of potential adverse drug reactions (ADRs) after medicinal products have been licensed and released to the public.

ADRs can range from mild and tolerable side effects to severe and life-threatening events. They constitute 5% to 7% of emergency department consultations [2]. Their impact in terms of public health is significant because there are estimates concluding that ADRs can cause an increase in the duration of hospitalization stays for outpatient (mean 9.2, SD 0.2 d) and inpatient (mean 6.1, SD 2.3 d) settings [3]. Typically, pharmacovigilance professionals analyze data from individual case safety report (ICSR) databases (such as the Food and Drug Administration Adverse Event Reporting System, the database maintained by the US Food and Drug Administration) to identify potential pharmacovigilance signals, namely potential causal relationships between an ADR and a drug. ICSRs are typically submitted either by patients or by health care or pharmacovigilance professionals, and they are the main data source used today for pharmacovigilance. However, ICSR databases are subject to many biases; in addition, underreporting has been identified as a huge issue [4]. Moreover, such databases frequently lack information that could make a significant difference in the examination of a potential signal (eg, patients' medical history). Hence, the early detection of potential pharmacovigilance signals by collecting and analyzing data from various sources is critical to prevent serious side effects as soon as possible.

The term "real-world data" (RWD) refers to data collected outside of the controlled environment of clinical trials, such as electronic health records (EHRs), patient registries, insurance claims databases, electronic prescription systems, and so on. There is a growing interest in using RWD for pharmacovigilance signal management to facilitate faster and more efficient postmarketing surveillance [5]. The significance of RWD in pharmacovigilance lies in its potential for representing longitudinal real-world patient experiences and health care practices that can provide insights into drug safety under real-life conditions. Analyzing RWD could also enrich and consolidate the already existing knowledge on ADRs (eg, by detecting new cofounders). Indicatively, a federated RWD network was used recently to validate the value of RWD in terms of pharmacovigilance signal management [6].

To this end, the European Medicines Agency and the US Food and Drug Administration have established infrastructures for the leverage of RWD for drug safety purposes, called Data Analysis and Real World Interrogation Network (DARWIN) [7] and the Sentinel Initiative [8], respectively. RWD are also being actively investigated for purposes beyond drug safety (eg, epidemiology) [9]. It should be noted that although RWD could in principle provide a good overview of patients' clinical course, two major challenges are preventing their use: (1) these datasets typically come with significant data quality risks and usually contain a high proportion of null values and errors; and (2) because of legal, ethical, and regulatory issues (eg, patient privacy issues), it is difficult to access these data sources.

Rationale

Artificial intelligence (AI) is widely acknowledged as a potentially very useful technical breakthrough that could be used to support decisions in health care (eg, clinical decision support systems) due to its ability to efficiently process big data to seek useful information. AI could be used to identify patterns and associations within large amounts of data (eg, RWD) where traditional statistical methods of data analysis may struggle to extract because of the amount and complexity (eg, nonlinear relationships between variables) of the data. AI has been widely investigated regarding its applications in health care (eg, personalized medicine) with promising results [10,11]; however, it is not yet widely applied in clinical practice. In the context of pharmacovigilance, AI could potentially support multiple aspects (eg, the identification of patient subpopulations who may be more vulnerable to specific ADRs), contributing to the vision of personalized drug safety management.

Objectives

The objective of this scoping review (SR) was to identify and characterize the current research trends regarding the use of AI on structured RWD for pharmacovigilance and identify relevant gaps.

Methods

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [12] methodology or rationale was applied. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) statement is a road map for authors to describe more precisely the state of the art and the findings of the literature search, as well as discuss the results.

Eligibility Criteria

Journal and conference articles written in English were selected if they focused on pharmacovigilance and reported the use of symbolic and nonsymbolic AI approaches applied to RWD, specifically EHRs, insurance claims databases, and administrative health data (Textbox 1).

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for the scoping review.

Inclusion criteria

- Article type: research
- Language: English
- Data type: tabular
- Data analysis method: symbolic artificial intelligence (AI) and nonsymbolic AI

Exclusion criteria

- Article type: review and opinion articles
- Language: other
- Data type: image and text
- Data analysis method: statistical

Review and opinion articles were excluded from the final manuscript selection. Furthermore, research articles focusing on image and text data (eg, social media and clinical notes) were also excluded. In addition, AI methods focusing on the use of natural language processing (NLP), natural language understanding, image processing, or object detection were considered beyond the scope of this work.

A key issue that came up during this SR was the lack of a clear distinction between plain statistical methods and machine learning (ML) approaches because these 2 domains frequently overlap, and these 2 terms are sometimes used interchangeably. In this manuscript, we acknowledge that the difference between AI and statistical methods is that AI creates models that can "learn" from data during iterative training processes, while statistical methods deal with finding relationships between variables. Thus, we considered the iterative "learning" part of an algorithm as the key feature to classify the algorithm as AI and ML. We excluded papers that were based on algorithms with no iterative "learning" scheme because we considered them to be part of the "plain statistical methods" approaches. Finally, we excluded papers that focused on adverse drug events related to medical devices.

Information Sources and Search Strategy

A search query was developed and executed on January 31, 2024, to include research articles from 2010 to 2024 exclusively from the MEDLINE scientific library, given that it is the oldest and biggest repository of journal articles in life sciences. Textbox 2 presents the query structure.

Textbox 2. Query structure ([pharmacovigilance terms with OR] AND [artificial intelligence (AI) terms with OR] AND [real-world data (RWD) terms with OR]).

Pharmacovigilance (keywords relevant to known adverse drug reaction [ADR] categories, synonyms of drug safety, pharmacovigilance terminology, and known individual case safety report [ICSR] databases)

V OR "pharmacovigil*" OR "pharmaco-vigil*" OR "side effect*" OR "adverse reaction*" OR "Product Surveillance" OR "postmarket*" OR pharmacoepidemiol* OR pharmaco-epidemiol* OR "drug safety" OR "drug event*" OR "toxicit*" OR "drug reaction*" OR "adverse drug*" OR "allerg*" OR "post-market*" OR "post market*" OR vaccinovigil* OR vaccino-vigil* OR eudravigilance OR "individual case safety report*" OR ICSR OR VAERS OR FAERS OR AERS OR vigibase OR "adverse effect*" OR "adverse event*" OR hypersensitiv* OR "spontaneous report*" OR "yellow card" OR "yellow-card" OR ADR OR "personalized pharmacovigilance" OR "precision pharmacovigilance" OR "pharmacosurveillance" OR "pharmaco-surveillance"

AI (categories of AI, terms that are used in the development of an AI model, explainable and interpretable AI methods, and different AI architectures)

"artificial intelligence" OR AI OR "machine learning" OR ML OR "neural network*" OR NN* OR "deep learning" OR DL OR ontolog* OR "knowledge engineering" OR KE OR reasoning OR inference OR "semantic web" OR "OWL" OR "Web Ontology Language" OR SWRL OR "RDF" OR "Resource Description Framework" OR "prediction" OR "estimation" OR "XAI" OR "SHAP" OR "Shapley value" OR "LIME" OR "Local Interpretable Model-agnostic Explanations" OR "DeepSHAP" OR "DeepLIFT" OR "CXplain" OR "Explainable Artificial Intelligence" OR "Explainable machine learning" OR "Interpretable artificial intelligence" OR "Interpretable machine learning"

RWD or real-world evidence (categories of RWD and data models that are used to store RWD)

"Real World Evidence" OR "Real World Data" OR RWE OR RWD OR "Observational Medical Outcomes Partnership" OR "OMOP" OR "Electronic Healthcare Record*" OR "EHR" OR "Electronic Medical Record*" OR "EMR*" OR EHDEN OR OHDSI OR i2b2 OR Sentinel OR DARWIN OR "Data Analysis and Real World Interrogation Network" OR administrative OR claim* OR "Observational Health Data Sciences and Informatics" OR "European Health Data Evidence Network" OR "multimodal data" OR "multimodal drug data" OR "multidimensional data" OR "multidimensional drug data" OR "multi-modal data" OR "multi-modal drug data" OR "multi-dimensional data" OR "multi-dimensional drug data"

XSL•FO
RenderX

Selection Process

The initial phase (phase 1) focused on screening the titles and abstracts of the articles retrieved from the search query (Textbox 2) to map those that potentially met our inclusion criteria and exclude irrelevant studies using the Rayyan tool (Rayyan Systems Inc) [13]. Rayyan is an AI tool designed to facilitate remote collaboration among researchers when conducting systematic literature reviews. The platform gathers the titles and abstracts of all articles selected for the study, and reviewers can evaluate the eligibility (ie, "include," "exclude," or "maybe") of every article based on their review's objectives in blind mode, that is, each reviewer assesses the articles without prior knowledge of the other reviewers' decisions. We resolved any conflicts that arose during this process through consensus meetings involving all reviewers.

The second phase focused on the full-text review of the papers selected during phase 1 to decide on the final set for inclusion in this study. In the full-text review of the studies selected based on titles and abstracts, we excluded research papers that did not meet ≥1 of the inclusion criteria (ie, strong focus on AI, RWD, and pharmacovigilance) as well as studies that met the exclusion criteria (eg, studies related to image and text data or those following only statistical approaches).

Data-Charting Process

A standard data extraction form was used to obtain an overview of the 36 selected studies (Tables S1 and S2 in Multimedia Appendix 1). For each study, we extracted information about the authors; journal name (where the study was published); publication year; country of origin (where the study was conducted); the objective of the study; types of organizations that participated in the study (based on the authors' affiliations); and key findings that relate to the scoping review question, which are described in the next subsection (Data Collection Process and Mapping). Any inconsistencies were discussed and resolved among the reviewers.

Data Collection Process and Mapping

The selected studies were further elaborated and mapped against evaluation criteria using a spreadsheet. The main categories of mapping criteria were as follows: pharmacovigilance objectives (drug safety core activities and drug safety special topics), data provenance (data source categories and data sources), countries of origin, AI algorithm categories, data preprocessing methods, the use of explainable AI (XAI) methods, code availability, the use of models in clinical practice, ethical AI, and so on. Table 1 presents an external description of the mapping criteria.

Table 1. The categories and subcategories used in the risk-of-bias assessment of included studies designed to characterize artificial intelligence (AI) studies on structured real-world data in pharmacovigilance.

Category of bias	Explanation	Subcategories
Selection bias	The bias that occurs when the input data of an AI model underrepresent the target population	• Underrepresentation of certain demographic groups • Overrepresentation of adverse drug events from specific health care systems or regions
Measurement bias	How the different features are collected and measured	• Inconsistent adverse drug event reporting practices • Variations in diagnostic criteria or coding practices for medical conditions
Temporal bias	How the study processes time-dependent features	• Changes in prescribing patterns or drug formulations over time • Seasonal variations in disease prevalence or reporting behaviors
Algorithmic bias	The biases produced form AI model outputs	• Differential performance in adverse drug event detection across patient subgroups • Biased risk assessments for certain medications or populations
Implicit bias	How stereotypes influence the AI model design and interpretation	• Overlooking potential drug interactions more common in specific ethnic groups • Underestimating the severity of side effects reported by certain demographics
Confounding bias	How unaccounted-for confounders influence prediction	• Failing to consider comorbidities when assessing drug safety profiles • Not accounting for polypharmacy effects in adverse-event analysis
Automation bias	This refers to the tendency to overly rely on automated systems	• Overlooking rare or unusual adverse drug events not flagged by AI systems • Reduced critical evaluation of AI-generated safety signals by human experts

XSL•FO
RenderX

Studying Risk-of-Bias Assessment

To effectively map the risk of bias in each included study, we considered selection, measurement, temporal, implicit, confounding and automation biases. Furthermore, we translated these categories into more specific categories according to our study (Table 1).

Synthesis Methods

The mapping strategy was designed based on the 3 main pillars of the objective; in addition, we included general information about the research papers (Table 2). Furthermore, we included free-text fields in the mapping Microsoft Excel file to add significant extra details that cannot be easily classified. These fields included "objective," "methods," "assessment," and "interesting results." The criteria encompassed specific attributes (eg, drug safety core activities) that were defined based on previous experience of conducting an SR in the field [14] and key interest aspects identified during the review.

Furthermore, in terms of ethical AI, the included studies were evaluated based on trustworthy AI guidelines for solutions in medicine and health care from the FUTURE (Fairness, Universality, Traceability, Usability, Robustness, and Explainability)-AI initiative [15]. These guidelines are separated into 7 categories (fairness, universality, traceability, usability, robustness, explainability, and a general category). For our evaluation procedure, we included only the highly recommended subcategories from each of the 7 main categories for proof of concept (low technology readiness levels for ML models) [16]. Table 3 presents the selected criteria and their description.

XSL•FO
RenderX

Table 2. Mapping criteria architecture for each different category in the query. There are 2 types of criteria: textual and binary (yes or no).

Categories and critera	Subcriteria
General information	
PubMed and MEDLINE ID (number)	ID number of articles
Authors (text)	List of authors
Title (text)	Article title
Journal (text)	ID number of articles
Year published (number)	Year of article publication
Types of organizations (text)	Types of organizations based on the authors' affiliation; possible values: health care, government, academia, industry, pharmacovigilance monitoring
Country (text)	Country where the research was conducted based on the authors' affiliations
Pharmacovigilance	
Drug safety core activities (text)	Possible values: ADE[a] detection, ADE monitoring, ADE prevention, ADE assessment, ADE information collection, and ADE reporting
Drug safety special topics (text)	Possible values: comparative drug analysis, drug interactions, MoA[b] identification and analysis, personalized drug safety, signal detection, specific (class of) disease, specific (class of) drugs, specific adverse effect, and vaccine safety
Drug (text)	Drugs being examined in the research papers
Reaction (text)	Reactions being examined in the research papers
Indication (text)	Indications being examined in the research papers
Reference terminologies (text)	Known health informatics terminologies that are detected in the research papers
AI[c]	
AI categories (text)	Possible values: nonsymbolic AI and symbolic AI
Nonsymbolic AI (text)	Possible values: classification and regression
Classification (text)	Possible values: random forest, logistic regression, artificial neural network, XGBoost[d], support vector machine, decision tree, knowledge graph, k-nearest neighbors, gradient boost, naïve Bayes, random survival forest, and extra tree
Regression (text)	Possible values: logistic regression, linear regression, LASSO[e], and regularized Cox regression
Data preprocessing type (text)	Possible values: dimensionality reduction, feature engineering, null imputation, and data cleansing
Data cleansing (text)	Possible values: data normalization and remove null values
Feature engineering (text)	Possible values: one-hot encoding, binning, splitting, and calculated features
Null imputation (text)	Possible values: regression or classification imputation
Explainable AI methods (text)	Possible values: LIME[f] and SHAP[g]
Knowledge representation formalism (text)	Possible values: OWL[h] and RDF[i]
Knowledge engineering core activities (text)	Possible values: knowledge extraction, knowledge integration, and knowledge representation
Real-world data	
Data source categories (text)	Possible values: ADE databases, clinical narratives, clinical trials drug information databases, drug regulation documentation, EHRs[j], genetics and biochemical databases, spontaneous reporting systems, dispensing records from pharmacies, and administrative claims data

XSL•FO
RenderX

Categories and critera	Subcriteria
Data source or sources (text)	Possible values: proprietary closed data sources (eg, specific hospital EHR), FAERS[k], SIDER[l], SMILES[m], UK Biobank, Osteoarthritis Initiative dataset, PharmGKB[n], TwoSIDES, EU-ADR[o] reference set, Stockholm Electronic Patient Record Corpus, MIMIC[p], OMIM[q], DisGeNet[r], and AEOLUS[s]
Data model (text)	Possible values: OMOP-CDM[t], Sentinel, and custom
Evaluation criteria	
Code availability (text)	The availability of the code in an open registry; possible values: yes and no
Data preprocessing	Information about the data preprocessing procedures; possible values: yes and no
Clinical use	Information about the evaluation of the produced work pipeline in clinical environments; possible values: yes and no

[a]ADE: adverse drug event.
[b]MoA: mechanism of action.
[c]AI: artificial intelligence.
[d]XGBoost: extreme gradient boosting.
[e]LASSO: least absolute shrinkage and selection operator.
[f]LIME: local interpretable model-agnostic explanations.
[g]SHAP: Shapley additive explanations.
[h]OWL: Web Ontology Language.
[i]RDF: resource description framework.
[j]EHR: electronic health record.
[k]FAERS: Food and Drug Administration Adverse Event Reporting System.
[l]SIDER: Side Effect Resource.
[m]SMILES: Simplified Molecular Input Line Entry System.
[n]PharmGKB: Pharmacogenomics Knowledge Base.
[o]EU-ADR: European Union Adverse Drug Reaction.
[p]MIMIC: Medical Information Mart for Intensive Care.
[q]OMIM: Online Mendelian Inheritance in Man.
[r]DisGeNet: gene-disease association network.
[s]AEOLUS: Adverse Event Open Learning through Universal Standardization.
[t]OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model.

XSL•FO
RenderX

Table 3. Detailed description of the FUTURE (Fairness, Universality, Traceability, Usability, Robustness, and Explainability)-AI highly recommended and proof-of-concept machine learning guidelines used for this study, along with a general category (Table S3 Multimedia Appendix 1).

Categories and recommendations	Description
Fairness	
Define sources of bias	Identification of possible types and sources of bias for the AI[a] tool during the design phase (eg, sex, gender, age, ethnicity, socioeconomics, geography, comorbidities or disability of patients, and human biases during data labeling)
Universality	
Define clinical settings	Specification of the clinical settings in which the AI tool will be applied (eg, primary health care centers, hospitals, home care, low- vs high-resource settings, and 1 country or multiple countries)
Evaluate using external data	Testing of the developed AI model to an external dataset with different characteristics from the training set
Traceability	
Provide documentation (eg, technical and clinical)	Creation of documentation files that provide technical (eg, public repositories) and clinical information (eg, bias of the model based on its use)
Usability	
Define user requirements	Specification of the model's use from health care professionals
Robustness	
Define sources of data variation	Specification of data sources' variation that may impact the AI tool's robustness in the real world (differences in equipment, technical fault of the machine, data heterogeneities during data acquisition or annotation, or adversarial attacks)
Train with representative data	Data for the training process should represent the population based on the case study for which the AI model has been developed
Evaluate and optimize robustness	Risk mitigation measures should be implemented to optimize the robustness of the AI model, such as regularization, data augmentation, data harmonization, or domain adaptation
Explainability	
Define explainability needs	Use of interpretable or explainable models
General	
Engage interdisciplinary stakeholders throughout the AI lifecycle	—[b]
Implement measures for data privacy and security	—
Define adequate evaluation plan (eg, datasets, metrics, and reference methods)	—

[a]AI: artificial intelligence.

[b]Not applicable.

Reporting Risk-of-Bias Assessment

The selection of only studies written in English and the exclusion of AI studies focused on text mining or NLP, image processing, and statistical analysis could be identified as potential risks for bias. Furthermore, the selection of papers only from the MEDLINE database could be identified as a potential bias risk because it potentially leads to the omission of papers from other databases (eg, AI databases).

Results

Study Selection

The PubMed search query originally returned 4264 studies. During the abstract and title screening process (phase 1), we selected 93 (2.18%) of the 4264 articles for full-text screening (phase 2). During phase 2, based on the inclusion criteria, of these 93 research papers, 36 (39%) were selected. The PRISMA-ScR flowchart (Figure 1) presents a detailed overview of the selection procedure. The PRISMA-ScR checklist is presented in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. ADR: adverse drug reaction; AI: artificial intelligence; NLP: natural language processing.

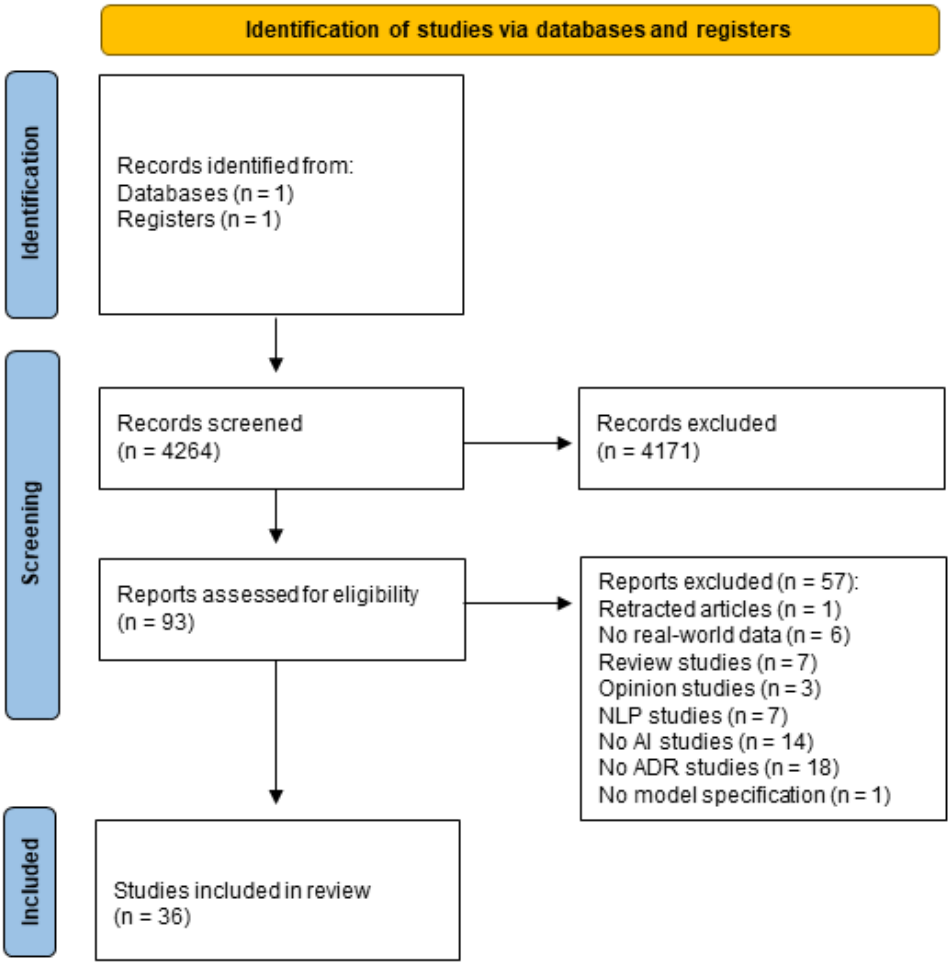

Study Characteristics

The included studies were published between 2015 and 2023, with a notable increase in the number of studies after 2019 (Table 4).

Of the 36 studies, 19 (53%) originated from the United States, 4 (11%) from Korea, and 4 (11%) from the United Kingdom, while the rest of the studies (n=9, 25%) were distributed across a variety of other countries (Table 5).

Most of the studies (30/36, 83%) were conducted from academia (Table 6).

Table 4. The distribution of studies through the years (n=36).

Years	Studies, n (%)
2015	1 (3)
2016	2 (5)
2017	3 (8)
2018	2 (5)
2019	2 (5)
2020	4 (11)
2021	10 (29)
2022	7 (20)
2023	5 (14)

[a]Includes studies conducted in multiple countries.

XSL•FO
RenderX

Table 5. Country of origin of the included studies (n=36)[a].

Countries	Studies, n (%)
United States	19 (53)
South Korea	4 (11)
United Kingdom	4 (11)
Canada	3 (8)
Sweden	3 (8)
China	3 (8)
France	3 (8)
Australia	2 (6)
Netherlands	2 (6)
Bangladesh	1 (3)
Israel	1 (3)
Belgium	1 (3)
Denmark	1 (3)
Taiwan	1 (3)
Ireland	1 (3)
Switzerland	1 (3)

[a]Includes studies that involved >1 type of organization.

Table 6. Types of organizations that participated in the including studies (n=36)[a].

Organizations	Studies, n (%)
Academia	30 (83)
Health care	9 (25)
Industry	6 (17)
Government	2 (6)
Regulatory bodies	1 (3)

[a]Includes studies that involved multiple databases.

[b]EHR: electronic health record.

[c]SRS: spontaneous reporting system.

[d]ADE: adverse drug event.

In terms of AI, of the 36 studies, 34 (94%) applied only nonsymbolic AI, and 1 (3%) used only symbolic AI, while 1 (3%) study combined the symbolic and nonsymbolic AI technical paradigms. Of the 34 nonsymbolic AI articles, 29 (85%) used classification tasks, whereas 3 (9%) selected regression algorithms, 3 (9%) applied causality algorithms (causal inference: n=2, 67%; causal discovery: n=1, 33%), and only 1 (3%) applied an association rule mining technique. The association rule mining study [17] followed a mathematical framework called formal concept analysis to create association rules between drugs and phenotypes to detect possible ADRs. Moreover, of the 29 studies that used classification tasks, 6 (21%) used XAI techniques, of which 4 (67%) used Shapley additive explanations, 1 (17%) used local interpretable model-agnostic explanations, and 1 (17%) tested both approaches.

Regarding RWD (Table 7), of the 36 articles, 28 (78%) focused on the use of EHRs (from local hospital databases), 4 (11%) used data from pharmacy dispensing records, and 3 (8%) used administrative claims data, while 2 (6%) focused on patient registries and 1 (3%) on insurance claims. In addition, a variety of other sources were used, including RWD such as drug information databases (3/36, 8%), spontaneous reports (3/36, 8%), adverse drug event databases (2/36, 6%), electronic prescription data (2/36, 6%), and genetics and biochemical databases (1/36, 3%).

Of the 36 studies, 23 (64%) used AI for ADR detection, 4 (11%) examined ADR assessment, 2 (6%) focused on ADR monitoring, 7 (19%) investigated ADR prevention, and 2 (6%) used AI to collect information about ADRs (Table 8).

XSL•FO
RenderX

Table 7. Variety of data used in the development of artificial intelligence models in the included studies (n=36)[a].

Type of database	Studies, n (%)
EHRs[b]	28 (78)
Drug information databases	4 (11)
Dispensing records from pharmacies	4 (11)
SRSs[c]	3 (8)
Administrative claims data	3 (8)
Patient registries	2 (6)
Electronic prescription data	2 (6)
ADE[d] databases	2 (6)
Insurance claims	1 (3)

[a]Includes studies that examined multiple pharmacovigilance core activities.
[b]ADR: adverse drug reaction.

Table 8. Description of pharmacovigilance core activities in the included studies (n=36)[a].

Pharmacovigilance core activities	Studies, n (%)
ADR[b] detection	23 (64)
ADR prevention	7 (19)
ADR assessment	4 (11)
ADR monitoring	2 (6)
ADR information collection	2 (6)

[a]Includes studies that involved multiple AI algorithms.
[b]XGBoost: extreme gradient boosting.
[c]LASSO: least absolute shrinkage and selection operator.
[d]No algorithms.

The classification studies (29/36, 81%; Table 9) tested several AI techniques, with random forest (RF) being the most frequently used algorithm (17/29, 59%). However, the regression studies (3/36, 8%) developed AI models only with extreme gradient boosting (1/3, 33%) and logistic regression (2/3, 67%).

Finally, for the evaluation of AI models (Table 10), most of the studies (24/36, 67%) reported area under the receiver operating characteristic curve as the primary metric.

Of the 36 studies, 32 (89%) investigated specific drug safety topics: 16 (50%) on specific adverse effects, 14 (44%) on specific class of drugs, 8 (25%) on specific (class of) diseases, 6 (19%) on signal detection, 3 (9%) on drug interactions, 2 (6%) on personalized drug safety, and 1 (3%) on vaccine safety (Table 11).

XSL•FO
RenderX

Table 9. The types of artificial intelligence (AI) algorithms that the models developed in the included studies (n=36).

AI models and algorithms	Studies, n (%)
Classification (n=29) [a]	
Random forest	17 (59)
XGBoost[b]	10 (34)
Artificial neural network	8 (28)
Logistic regression	8 (28)
Support vector machine	7 (24)
Decision tree	5 (17)
K-nearest neighbor	2 (7)
Gradient boost	2 (7)
LASSO[c]	2 (7)
Extra tree	1 (3)
Naïve Bayes	1 (3)
Random survival forest	1 (3)
Linear regression	1 (3)
Regularized Cox regression	1 (3)
Regression (n=3)	
XGBoost	1 (33)
Logistic regression	2 (67)
Causality[d] (n=3)	3 (100)

[a]Includes studies that involved multiple model evaluation metrics.

Table 10. Evaluation metrics of artificial intelligence (AI) models developed in the included studies (n=36)[a].

AI model evaluation metrics	Studies, n (%)
Area under the receiver operating characteristic curve	24 (67)
Accuracy	10 (28)
F_1-score	8 (22)
Precision	11 (31)
Recall	10 (28)
Negative predictive value	6 (17)
Sensitivity	9 (25)
Specificity	7 (19)
Other (≤2)	20 (56)

[a]Includes studies that examined multiple pharmacovigilance topics.

XSL•FO
RenderX

Table 11. Specialized pharmacovigilance topics presented in the included studies (n=36)[a].

Specialization of pharmacovigilance topics	Studies, n (%)
Specific adverse effect	16 (50)
Specific (class of) drugs	14 (44)
Specific (class of) disease	8 (25)
Signal detection	6 (19)
Drug interactions	3 (9)
Personalized drug safety	2 (6)
Vaccine safety	1 (3)

[a]Includes studies that used multiple data sources.

[b]SIDER: Side Effect Resource.

[c]FAERS: Food and Drug Administration Adverse Event Reporting System.

Table 12 presents the diversity in the data sources used in the included studies. Of the 36 studies, 29 (81%) chose proprietary closed data sources (eg, specific hospital EHRs) for their experiments. Along with EHR data, other data sources were also used (eg, Food and Drug Administration Adverse Event Reporting System and Side Effect Resource). Of the 36 studies, 2 (6%) selected the Stockholm Electronic Patient Record Corpus. The remaining RWD sources (Medical Information Mart for Intensive Care and the Osteoarthritis Initiative dataset) are represented in only 2 (6%) of the 36 studies (n=1, 50% for every database).

In terms of data models, of the 36 studies, 27 (75%) used proprietary data models, 3 (8%) did not mention any data model, 5 (14%) used the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM), and 1 (3%) used the Sentinel model (Table 13).

Figure 2 presents the case studies examined in the included articles. Notably, an important number of studies (20/36, 55%) did not work in specific ADR case studies. Another significant outcome is the diversity of case studies; the articles do not focus on a specific drug, indication, or reaction. It can be observed that chemotherapy drugs and their associated reactions in various types of cancers emerge as slightly more prominent categories in this review (Figure 3).

Table 12. Variety of data sources used in the included studies (n=36)[a].

Data sources	Studies, n (%)
Proprietary closed data sources	29 (81)
Other	12 (33)
SIDER[b]	3 (8)
FAERS[c]	2 (6)
Stockholm Electronic Patient Record Corpus	2 (6)

[a]OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model.

Table 13. Included studies' distribution based on data models that the data are stored (n=36).

Data models	Studies, n (%)
Custom	27 (75)
OMOP-CDM[a]	5 (14)
Unknown	3 (8)
Sentinel	1 (3)

XSL•FO
RenderX

Figure 2. Association pathways between artificial intelligence models, data sources, and drug safety categories in the included studies. ADE: adverse drug event; EHR: electronic health record; LASSO: least absolute shrinkage and selection operator; SRS: spontaneous reporting system; XGBoost: extreme gradient boosting.

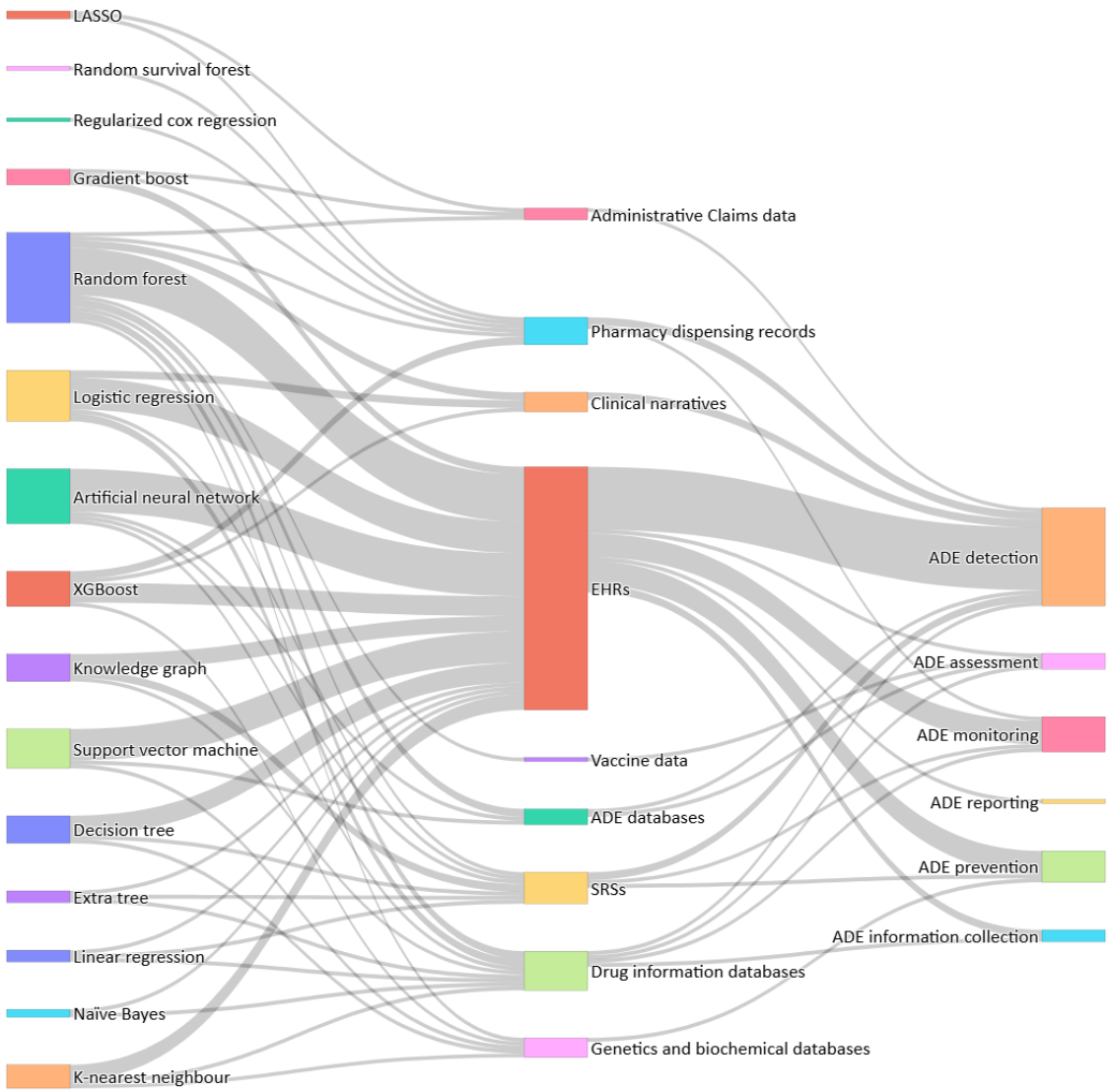

XSL•FO
RenderX

Figure 3. Drugs, indications, and reactions in the included studies. (A) Drugs. (B) Indications. (C) Reactions.

Although most of the studies (21/36, 58%) used complex AI algorithms (black boxes), such as RF (an ensemble method) and artificial neural networks (ANNs), to construct their prediction models in all ADR categories, many studies (15/36, 42%) used simple interpretable ML approaches such as logistic regression. Moreover, it is important to highlight that all studies worked on EHR databases, except for the adverse drug event assessment category in which we detected a single study with a vaccine database.

RWD databases were also used alongside other types of data; for example, EHRs were mostly combined with spontaneous reporting systems and drug information databases, vaccine data with adverse drug event databases, and administrative claims data with spontaneous reporting systems. Furthermore, some of the studies (3/36, 8%) integrated different types of observational data to develop AI models, combining pharmacy dispensing records with EHRs and administrative claims data.

XSL•FO
RenderX

Evaluation Results

Only 3 (8%) of the 36 studies included in this SR openly provided their code. In addition, only 16 (44%) of the 36 studies included a detailed description of data preprocessing pipelines for RWD. Moreover, just 4 (11%) of the 36 studies evaluated their methodology within a clinical environment (Table 14).

Table 14. Summary of code availability, data preprocessing, and clinical validation evaluation criteria (n=36).

Evaluation metrics	Studies, n (%)	
	Yes	No
Code availability	3 (8)	33 (92)
Data preprocessing	16 (44)	20 (56)
Clinical validation	5 (14)	31 (86)

In terms of trustworthy AI, only 5 (14%) of the 36 studies scored <50% on the Fairness, Universality, Traceability, Usability, Robustness, and Explainability–AI (FUTURE-AI) criteria (Tables 15 and 16). Among the studies that achieved scores of >75% [18-20], 3 (75%) out of 4 used external data to evaluate their models, addressing the Universality criterion (Table S4 in Multimedia Appendix 1).

Table 15. Distribution of included studies according to the FUTURE (Fairness, Universality, Traceability, Usability, Robustness, and Explainability)-AI guidelines (n=36)[a].

Data models	Studies, n (%)
Custom	27 (75)
OMOP-CDM[a]	5 (14)
Unknown	3 (8)
Sentinel	1 (3)

[a]Includes studies that fell into multiple FUTURE-AI and general categories.

XSL•FO
RenderX

Table 16. Evaluation of articles included in the scoping review based on the Fairness, Universality, Traceability, Usability, Robustness, and Explainability–AI (FUTURE-AI) and Code Availability, Data preprocessing and clinical validation criteria.

Study	Year	Code availability	Data preprocessing	Clinical use	FUTURE-AI criteria satisfaction (%)
Anastopoulos et al [18]	2021	No	Yes	No	75
Ward et al [21]	2021	No	Yes	No	67
Zhang et al [22]	2020	No	Yes	No	50
Kim et al [23]	2021	No	No	No	67
Morel et al [24]	2020	Yes	No	Yes	67
Zou et al [25]	2021	No	No	Yes	67
Liu et al [26]	2018	No	No	Yes	42
Zhu et al [27]	2022	No	Yes	No	67
Kidwai-Khan et al [28]	2022	No	Yes	No	58
Sharma et al [29]	2022	No	No	No	58
On et al [30]	2022	No	Yes	No	50
Datta et al [31]	2021	No	No	No	58
Bagattini et al [19]	2019	Yes	Yes	No	75
Gibson et al [32]	2021	No	Yes	No	58
Jeong et al [33]	2018	No	Yes	No	67
Zhao et al [34]	2015	No	Yes	No	58
Zhao and Henriksson [35]	2016	No	Yes	No	42
Segal et al [20]	2019	No	No	Yes	75
Boland et al [36]	2017	No	No	No	67
Wang et al [37]	2021	No	No	No	58
Li et al [38]	2022	No	No	No	58
Jin et al [39]	2020	No	No	No	67
Hansen et al [40]	2016	No	No	No	50
Mosa et al [41]	2021	No	No	No	67
Herrin et al [42]	2021	No	No	No	50
Pichardo et al [43]	2022	No	No	No	67
Puzhko et al [44]	2021	No	No	No	58
Souissi et al [45]	2017	No	No	Yes	42
Personeni et al [17]	2017	No	No	No	42
Zhou et al [46]	2020	Yes	No	No	67
Goyal et al [47]	2023	No	No	No	58
Wang et al [48]	2023	No	Yes	No	58
Hughes et al [49]	2023	No	Yes	No	58
Sharma et al [50]	2023	No	Yes	No	67
Akimoto et al [51]	2023	No	Yes	No	58
Zhang et al [52]	2022	No	Yes	No	58

Risk of Bias in the Included Studies

Table 17 provides an overview of the distribution of biases across the included studies. Notably, of the 36 studies, 21 (58%) included selection biases, and 17 (47%) included confounding biases. Algorithmic biases were identified in 14 (39%) of the 36 studies. Table S5 in Multimedia Appendix 1 presents the detailed categorization of the included studies in the different risk-of-bias categories.

XSL•FO
RenderX

Table 17. Results of the risk-of-bias assessment for each included study organized into subsections that better characterize the artificial intelligence (AI) models developed using structural real-world data in pharmacovigilance.

Bias categories and subcategories	References
Data-related biases	
Selection bias	
Underrepresentation of certain demographic groups	[18,19,21,22,26,28-30,33,38,41,42,44-47,49-52]
Overrepresentation of adverse drug events from specific health care systems or regions	[18,19,21,22,24,26,28-30,33,38,42,45-47,49-52]
Measurement bias	
Inconsistent adverse drug event reporting practices	[18,20,21,24,32,38,41,44,45]
Variations in diagnostic criteria or coding practices for medical conditions	[20,24,32,38,46,51,52]
Temporal bias	
Changes in prescribing patterns or drug formulations over time	[17,18,21,24,34,39,40,44,45,51,52]
Seasonal variations in disease prevalence or reporting behaviors	[34,39,40,44,45,51]
Algorithm-related biases	
Algorithmic bias	
Differential performance in adverse drug event detection across patient subgroups	[17,18,27,32,36,37-39,41,45,46,48-50]
Biased risk assessments for certain medications or populations	[17,18,27,32,36-39,41,45,46,48-50]
Implicit bias	
Overlooking potential drug interactions more common in specific ethnic groups	[17,18,24,31,41,44]
Underestimating the severity of side effects reported by certain demographics	[17,24,31,38,41,44]
Deployment and interpretation biases	
Confounding bias	
Failing to consider comorbidities when assessing drug safety profiles	[18,21,22,25,27,31,34,36,44,46-49,51]
Not accounting for polypharmacy effects in adverse drug event analysis	[18,21,22,25,27,28,31,34,36,40,46-49,51]
Automation bias	
Overlooking rare or unusual adverse drug events not flagged by AI systems	[19,25,49,50]
Reduced critical evaluation of AI-generated safety signals by human experts	[19,24,49,50]

Interesting Results of Individual Studies

It is important to mention that 6% (2/36) of the studies successfully combined AI and a self-controlled case series (SCCS) model for ADR detection. Morel et al [24] introduced the convolutional SCCS (ConvSCCS) model in which the SCCS model is enriched with a convolutional neural network. This allows the ConvSCCS model to consider a few longitudinal data dimensions (eg, drug exposure) from observational data and predict a potential ADR without a prior definition of risk windows, which is mandatory in SCCS models. The ConvSCCS model was tested in glucose-lowering drugs and the risk of bladder cancer case study. Another interesting advantage shown by the results is that the ConvSCCS model is useful for analyzing high-dimensional data while requiring minimal data preprocessing. Zhang et al [22] developed the neural SCCS (NSCCS) model to detect probable drug interactions and control for time-invariant confounders. The NSCCS model was tested in the OMOP-CDM reference dataset [53]. Both the ConvSCCS and NSCCS models outperformed traditional SCCS statistical models in comparative analyses; the ConvSCCS model demonstrated superior precision and computational speed, while the NSCCS model achieved an area under the receiver operating characteristic curve score of 0.779 [22].

Furthermore, Kidwai-Khan et al [28] focused on improving the prediction of preventable adverse events by integrating in an AI decision support tool, EHRs with genetic data (the presence or absence of genes contraindicated with a person's medication). This is the only study in this review that combined EHRs with genetic data and 1 (25%) of the 4 studies that used XAI methods. In addition, all AI models achieved high evaluation scores (>95%).

A few ADR prediction studies (3/36, 8%) introduced innovative ideas on feature preprocessing. Jeong et al [33] developed an ML prediction model in which the features are calculated from algorithms such as the "comparison of extreme laboratory test" results, "comparison of extreme abnormality ratio", and "prescription pattern around clinical events" to help determine whether a drug-laboratory event pair is associated. A different approach was proposed by Wang et al [37] who addressed problems with low-quality observational data (eg, missing data) by creating patient embeddings and treating patients "as bags with the various number of feature-value pairs, called instances."

XSL•FO
RenderX

This method led to the development of the final AI model (AMI-Net3), which achieved exceptional performance. Chen et al [54] also proposed an embedding methodology, called "physiological signal embeddings." This study proved that training deep embedding models on physiological signals could lead to better forecasts of adverse outcomes. In addition, this methodology enables data transferability through the physiological signal embeddings models.

Only 1 (3%) of the 36 studies developed an application for the prediction of ADRs. Mosa et al [41] leveraged the interoperability of a decision tree ML model and, based on their results, designed a rule-based mobile app to assess the risk of specific ADRs and indications.

A completely different approach to ADR prediction was introduced by Liu et al [26]. In this study, the authors applied an ML method to develop a prediction model for osteoarthritis ADRs in analgesic drugs. Afterward, the authors used explainability techniques to identify patients who might be prescribed analgesic drugs without the risk of osteoarthritis ADRs. The diversity of this study is addressed in a different scope: instead of predicting ADRs based on a patient's medical history, the model focuses on identifying the characteristics that make the patient suitable for a medication, specifically by considering the presence or absence of an ADR.

Recently, the causal ML paradigm was introduced into pharmacovigilance through the studies of Wang et al [48] and Zhang et al [52], who applied causal inference with average treatment effects and causal discovery with directed acyclic graphs, respectively. Wang et al [48] used causal ML models to make a representation of a randomized clinical trial with EHR data. Their results successfully identified both well-known and new medications that could cause the suspected ADR in their case study. Furthermore, Zhang et al [52] created a causal graph for a drug-event combination and compared the results from 2 causal discovery algorithms. Their results showcased the causal discovery algorithms' abilities to explore the mechanisms of the suspect drug that could lead to a potential ADR, uncovering previously unknown causal links.

Only 1 (3%) study focused on the use of symbolic AI [43] compared to those investigating the use of ML (35/36, 97%) and 1 study combine symbolic and nonsymbolic AI. Notably, Pichardo et al [43] stand out for integrating ontologies and ML, namely combining symbolic and nonsymbolic AI. The objective of this study is to examine the performance of a clinically informed framework for the prediction of short-term ADRs.

Furthermore, very few studies (5/36, 14%) focused on the clinical evaluation of the proposed ML approach. Segal et al [20] presented a clinical decision support system designed to provide medication error alerts to prevent ADRs, demonstrating significant results—40% of the prescriptions were altered based on these alerts. Herrin et al [42] compared the effectiveness of their proposed ML scheme to that of an established clinical practice, specifically the HAS-BLED approach, to evaluate a patient's risk of gastrointestinal bleeding.

Synthesized Findings

Although the included studies followed widely used approaches in AI and pharmacovigilance to predict potential ADRs, primarily using EHR data, such as predicting ADR outcomes using well-known ML model architectures, the aforementioned studies follow fundamental methodologies. In terms of data, it describes several interesting processes such as the creation of patient embeddings and the application of the "comparison of extreme laboratory test" results, "comparison of extreme abnormality ratio", and "prescription pattern around clinical event" algorithms for calculating input data for the ML model. Furthermore, it is described a variety of innovative AI algorithms such as SCCS models, causal ML, and symbolic AI. Finally, out-of-the-box pharmacovigilance approaches were followed, such as identifying patients suitable for specific treatments based on their ADR profiles.

Discussion

The findings of this study enabled us to identify innovative ideas, spot existing limitations, and propose potential directions for future work in this field.

Principal Findings

In summary, the number of studies on AI methodologies applied to RWD for pharmacovigilance purposes has significantly increased in the last 5 years—most of the included studies (28/36, 78%) were published after 2019, with the United States contributing the most publications in this field (19/36, 53%).

Comparing this review study with 3 recent reviews from 2022 to 2023 in the same field, we conclude that only the study by Kaas-Hansen et al [55] could be considered a study with the exact focus with our review. Their review included only 7 scientific papers because they restricted their selection to studies published between 2015 and 2021 and involving >1000 patient records. Their findings are similar to ours in terms of dominant AI solutions (classification) and the type of RWD used (EHR). Finally, it is essential to note that their review, like ours, highlights the limited adoption of widely used common data models, such as the OMOP-CDM.

In terms of risk of bias, selection biases were due to the fact that most of the studies (15/36, 42%) did not include patients from >1 data source (regional hospital [46] or insurance claims database [24]) in their models. Regarding confounding, while high-dimensional RWD offer a significant amount of information, they also contain substantial noise and null values. Sequentially, features that could be potential confounders are eliminated from the final dataset.

Pharmacovigilance

The reviewed papers focused on using ML to detect ADRs to confirm whether previously known ADRs could have been identified using RWD. Another major theme identified was the prediction of ≥1 ADRs based on the classification of patients with different characteristics, ultimately aiming to support personalized ADR prevention. However, there was a lack of studies investigating new, potential pharmacovigilance signals. Regarding the investigated ADRs, there was a slightly higher

XSL•FO
RenderX

interest in chemotherapy drugs for different types of cancers due to the high incidence of serious reactions associated with this treatment.

Finally, it is important to highlight that only 4 (11%) of the 36 studies in this review were tested in real-world clinical environments, which leads us to conclude that AI models may lack generalizability or that health care professionals may lack trust in AI models. By contrast, the trustworthy AI evaluation based on the FUTURE-AI guidelines proved that only a few studies (4/36, 11%) failed to satisfy half of the criteria, indicating relatively high research quality.

RWD Preprocessing

In terms of RWD, EHRs were the most commonly used data source. EHRs are multidimensional, offering data that could be crucial for detecting postmarketing ADRs. At least in principle, EHRs could serve as an invaluable data source for investigating potential drug synergies or interactions across diverse populations. Furthermore, the variety of information in patient records could be an advantage in the creation of multimodal datasets, for example, by integrating biological, signaling pathway, and drug information databases.

However, the use of EHRs comes with significant burdens because they contain sensitive personal data, leading to limited access. Medical Information Mart for Intensive Care is the only openly available EHR dataset for researchers, but it is not commonly used in pharmacovigilance (only 2/36, 6% studies used it).

Moreover, it is important to mention that RWD preprocessing is challenging due to its complexity and real-world nature (biases, errors, gaps, noise, etc). Consequently, less than half of the articles (17/36, 47%) described in detail the data preprocessing step in their pipelines.

Another noteworthy outcome is that widely adopted data models such as the OMOP-CDM, Informatics for Integrating Biology & the Bedside, and Sentinel appeared sporadically in the studies. This could be attributed to the fact that the use of EHR data and AI models is relatively new. However, it should be noted that initiatives in this direction are emerging (eg, the Assessment of Pre-trained Observational Large Longitudinal models in Observational Health Data Sciences and Informatics initiative [56]).

Finally, it should be noted that RWD have a very substantial longitudinal dimension and questionable quality (due to gaps, errors, etc). As such, leveraging RWD for pharmacovigilance purposes requires the development of new approaches that focus on using time-related sequential information. While several attempts have been made to exploit this temporal aspect of RWD [19,33-35], validating AI and ML algorithms focusing on time-series rationale for pharmacovigilance signal detection remains a critical issue.

AI Models

The detection of potential pharmacovigilance signals is a challenging procedure. As a result, the development of ML models to support the detection of ADR signals could have a significant impact. We can outline 2 major approaches for ADR signal detection. The first focuses on creating an AI tool that could discover unknown relationships between drugs and conditions, highlighting potential causal associations. The second approach emphasizes AI pipelines tailored to specific ADRs, where the input data for the final ML model are preprocessed based on prior medical knowledge of the drug-event combination.

The AI models identified in this review are generally complex, with ensemble methods such as RF being the most commonly used. A significant number of studies also applied ANNs. As RWD contain a substantial amount of diverse information, the relationships between different features may not be linear. Hence, the use of black box models (eg, ensemble methods and ANNs) is essential for discovering more complicated associations in a dataset beyond linear relationships.

On the basis of the review papers, there is a noticeable lack of use of XAI models (ie, local interpretable model-agnostic explanations and Shapley additive explanations models). Health care professionals highlight the necessity to understand the motifs between AI models' tasks to accept the decisions made by the algorithms. This could not only lead to the biological translation of the results based on existing knowledge but also reveal new information about a disease, a medication, and so on. In terms of pharmacovigilance, XAI models applied to RWD could bring evidence about unknown confounders in an ADR and provide more informative results for pharmacovigilance experts regarding the causes of a potential pharmacovigilance signal. Although XAI methods' results are tested extensively in the health care domain, we found that only 6 recent studies (4/36, 11% are included in this review) had applied them in the pharmacovigilance domain [21,23,27,28,54,57] (4 in 2021, 2 in 2022). Another novel approach discussed extensively in the explainability field is the newly introduced causal ML or causal deep learning algorithm, which combines AI and causal inference to uncover underlying cause-and-effect relationships between variables. The complexity of RWD presents a challenge that causal ML could potentially solve more efficiently by providing meaningful explanations of the causal relationships between variables [58]. These innovative AI models could serve as a good hypothesis for future work because they seek and present the relationships between different variables in RWD sources with a more informative structure than traditional AI models. They have already been applied efficiently in pharmacological treatment patterns [59]. In this review, only 3 (8%) of the 36 studies applied causal deep learning to EHR data [25,48,52].

Finally, a major problem identified based on the SR findings is the lack of code availability. This issue hinders the reproducibility of the models, preventing further testing on different datasets and raising questions about the developed AI models' robustness and generalizability.

Strengths and Limitations

The strengths of this review include the use of a considerable number of studies (n=36), providing a thorough knowledge of the specific scientific field. Besides, we compared our findings with those of the most recent review in the same field and analyzed the differences.

XSL•FO
RenderX

Nevertheless, this systematic review has several limitations. First, we only included articles from the MEDLINE database. As such, we may have excluded other existing AI approaches to structural RWD in the field of pharmacovigilance that could be available in AI databases. Second, because of the variation in the articles' methodologies, we were unable to conduct a meta-analysis of the quantitative results. Finally, it is important to mention the limited number of symbolic AI studies in this review (4/36, 11%) [17,45,60,61]. The construction of knowledge graphs (KGs) usually requires the use of text mining procedures such as NLP and focused on real text such as clinical notes. As we excluded NLP studies from our query, we assume that this contributed to the small number of symbolic AI articles (n=36) included in our review.

Current Gaps and Potential Future Work Paths

Detecting new pharmacovigilance signals using ML approaches requires evidence of a causal association between the suspect drug and the reaction. XAI models can assist pharmacovigilance professionals in this process. To this end, further investigation into causal ML and causal deep learning approaches could be a highly impactful line of research for identifying pharmacovigilance signals from RWD.

Another gap identified in this SR that could indicate future work paths could be the use of multitask learning approaches. Multitask learning is an ML methodology that takes as input 1 dataset to execute multiple prediction tasks. RWD, such as EHR data, are rich data sources that could support >1 task (eg, pharmacovigilance and pharmacoepidemiology); for instance, a multitask learning model could predict an adverse drug event, the severity of an adverse drug event, and the likelihood of the same adverse drug event occurring with other drugs in a patient.

Furthermore, combining ML approaches with symbolic AI is a line of work that offers further potential for exploration. Combining ML with ontologies and automatic reasoning upon KGs could enable new AI approaches (eg, neurosymbolic AI) and provide new insights based on well-established expert knowledge formed as a KG. Moreover, using ontologies and KGs could support integration with other kinds of data sources (eg, data sources containing low-level biochemical or pharmacokinetics and pharmacodynamics information and signaling pathway information).

Finally, exploiting the currently formed federated data networks could also be an interesting area for future research; for example, the European Health Data & Evidence Network is currently setting up a network of >180 data partners across Europe, using the OMOP-CDM as the main data model [62]. The adoption of the OMOP-CDM and the potential exploitation of such data networks would significantly enhance the prospects of potential AI models used for pharmacovigilance.

Conclusions

In this paper, we reviewed scientific papers focusing on AI approaches to structured RWD for pharmacovigilance purposes. It should be noted, as a key finding, that most models are designed not for pharmacovigilance signal detection but for personalized ADR prediction. Furthermore, XAI methods and causal ML and causal deep learning are not investigated in depth. Moreover, there are no identified gold standard methodologies for data preprocessing of structured RWD for pharmacovigilance. Finally, an evaluation of the already developed AI models in external data is difficult because of code unavailability and a lack of data access.

Therefore, there is an essential need for more informative XAI models that can be validated on external datasets and for a more detailed description of RWD preprocessing pipelines and methods to examine potential pharmacovigilance signals in clinical practice. Implementing AI approaches in RWD analysis could tackle the problems of pharmacovigilance signaling underreporting and support the vision of personalized ADR management.

Acknowledgments

This work was funded by the Agence nationale de sécurité du médicament et des produits de santé (2016s076) and was supported by a PhD contract with Sorbonne University.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[XLSX File (Microsoft Excel File), 5587 KB - jmir_v26i1e57824_app1.xlsx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[DOCX File , 84 KB - jmir_v26i1e57824_app2.docx]

XSL•FO
RenderX

References

1. Pharmacovigilance: ensuring the safe use of medicines. World Health Organization. URL: https://www.who.int/publications/i/item/WHOEDM2004.8 [accessed 2024-04-29]
2. Just KS, Dormann H, Böhme M, Schurig M, Schneider KL, Steffens M, et al. Personalising drug safety-results from the multi-centre prospective observational study on Adverse Drug Reactions in Emergency Departments (ADRED). Eur J Clin Pharmacol 2020 Mar 12;76(3):439-448. [doi: 10.1007/s00228-019-02797-9] [Medline: 31832731]
3. Formica D, Sultana J, Cutroneo PM, Lucchesi S, Angelica R, Crisafulli S, et al. The economic burden of preventable adverse drug reactions: a systematic review of observational studies. Expert Opin Drug Saf 2018 Jul;17(7):681-695. [doi: 10.1080/14740338.2018.1491547] [Medline: 29952667]
4. Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf 2017 Jul 09;16(7):761-767. [doi: 10.1080/14740338.2017.1323867] [Medline: 28447485]
5. Patadia VK, Schuemie MJ, Coloma PM, Herings R, van der Lei J, Sturkenboom M, et al. Can electronic health records databases complement spontaneous reporting system databases? A historical-reconstruction of the association of Rofecoxib and acute Myocardial infarction. Front Pharmacol 2018;9:594 [FREE Full text] [doi: 10.3389/fphar.2018.00594] [Medline: 29928230]
6. Gauffin O, Brand JS, Vidlin SH, Sartori D, Asikainen S, Català M, et al. Supporting pharmacovigilance signal validation and prioritization with analyses of routinely collected health data: lessons learned from an EHDEN network study. Drug Saf 2023 Dec 07;46(12):1335-1352 [FREE Full text] [doi: 10.1007/s40264-023-01353-w] [Medline: 37804398]
7. Data analysis and real world interrogation network (DARWIN EU). European Medicines Agency. URL: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu [accessed 2024-04-29]
8. The FDA's sentinel initiative. HealthAffairs. URL: https://www.healthaffairs.org/content/briefs/fda-s-sentinel-initiative#:~:text=As%20of%20February%202015%2C%20queries,communication%20in%20just%20four%20cases [accessed 2024-04-29]
9. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, N3C Consortium. Identifying who has long COVID in the USA: a machine learning approach using N3C data. Lancet Digit Health 2022 Jul;4(7):e532-e541 [FREE Full text] [doi: 10.1016/S2589-7500(22)00048-6] [Medline: 35589549]
10. Yang Y, Yuan Y, Zhang G, Wang H, Chen YC, Liu Y, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nat Med 2022 Oct 22;28(10):2207-2215 [FREE Full text] [doi: 10.1038/s41591-022-01932-x] [Medline: 35995955]
11. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov 22;24(11):1716-1720 [FREE Full text] [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
12. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
13. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
14. Natsiavas P, Malousi A, Bousquet C, Jaulent MC, Koutkias V. Computational advances in drug safety: systematic and mapping review of knowledge engineering based approaches. Front Pharmacol 2019 May 17;10:415 [FREE Full text] [doi: 10.3389/fphar.2019.00415] [Medline: 31156424]
15. Lekadir K, Feragen A, Fofanah AJ, Frangi AF, Buyx A, Emelie A, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. arXiv Preprint posted online August 11, 2023 [FREE Full text]
16. Lavin A, Gilligan-Lee CM, Visnjic A, Ganju S, Newman D, Ganguly S, et al. Technology readiness levels for machine learning systems. Nat Commun 2022 Oct 20;13(1):6039 [FREE Full text] [doi: 10.1038/s41467-022-33128-9] [Medline: 36266298]
17. Personeni G, Bresso E, Devignes MD, Dumontier M, Smaïl-Tabbone M, Coulet A. Discovering associations between adverse drug events using pattern structures and ontologies. J Biomed Semantics 2017 Aug 22;8(1):29 [FREE Full text] [doi: 10.1186/s13326-017-0137-x] [Medline: 28830518]
18. Anastopoulos IN, Herczeg CK, Davis KN, Dixit AC. Multi-drug featurization and deep learning improve patient-specific predictions of adverse events. Int J Environ Res Public Health 2021 Mar 05;18(5):2600 [FREE Full text] [doi: 10.3390/ijerph18052600] [Medline: 33807714]
19. Bagattini F, Karlsson I, Rebane J, Papapetrou P. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. BMC Med Inform Decis Mak 2019 Jan 10;19(1):7 [FREE Full text] [doi: 10.1186/s12911-018-0717-4] [Medline: 30630486]

XSL•FO
RenderX

20. Segal G, Segev A, Brom A, Lifshitz Y, Wasserstrum Y, Zimlichman E. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Inform Assoc 2019 Dec 01;26(12):1560-1565 [FREE Full text] [doi: 10.1093/jamia/ocz135] [Medline: 31390471]
21. Ward IR, Wang L, Lu J, Bennamoun M, Dwivedi G, Sanfilippo FM. Explainable artificial intelligence for pharmacovigilance: what features are important when predicting adverse outcomes? Comput Methods Programs Biomed 2021 Nov;212:106415. [doi: 10.1016/j.cmpb.2021.106415] [Medline: 34715520]
22. Zhang W, Peissig P, Kuang Z, Page D. Adverse drug reaction discovery from electronic health records with deep neural networks. Proc ACM Conf Health Inference Learn (2020) 2020 Apr;2020:30-39 [FREE Full text] [doi: 10.1145/3368555.3384459] [Medline: 33283213]
23. Kim Y, Jang JH, Park N, Jeong NY, Lim E, Kim S, et al. Machine learning approach for active vaccine safety monitoring. J Korean Med Sci 2021 Aug 09;36(31):e198 [FREE Full text] [doi: 10.3346/jkms.2021.36.e198] [Medline: 34402232]
24. Morel M, Bacry E, Gaïffas S, Guilloux A, Leroy F. ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection. Biostatistics 2020 Oct 01;21(4):758-774. [doi: 10.1093/biostatistics/kxz003] [Medline: 30851046]
25. Zou B, Mi X, Tighe PJ, Koch GG, Zou F. On kernel machine learning for propensity score estimation under complex confounding structures. Pharm Stat 2021 Jul 22;20(4):752-764 [FREE Full text] [doi: 10.1002/pst.2105] [Medline: 33619894]
26. Liu L, Yu Y, Fei Z, Li M, Wu FX, Li HD, et al. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. BMC Syst Biol 2018 Nov 22;12(Suppl 6):105 [FREE Full text] [doi: 10.1186/s12918-018-0624-4] [Medline: 30463545]
27. Zhu X, Hu J, Xiao T, Huang S, Shang D, Wen Y. Integrating machine learning with electronic health record data to facilitate detection of prolactin level and pharmacovigilance signals in olanzapine-treated patients. Front Endocrinol (Lausanne) 2022 Oct 13;13:1011492 [FREE Full text] [doi: 10.3389/fendo.2022.1011492] [Medline: 36313772]
28. Kidwai-Khan F, Rentsch CT, Pulk R, Alcorn C, Brandt CA, Justice AC. Pharmacogenomics driven decision support prototype with machine learning: a framework for improving patient care. Front Big Data 2022 Nov 15;5:1059088 [FREE Full text] [doi: 10.3389/fdata.2022.1059088] [Medline: 36458283]
29. Sharma V, Kulkarni V, Jess E, Gilani F, Eurich D, Simpson SH, et al. Development and validation of a machine learning model to estimate risk of adverse outcomes within 30 days of opioid dispensation. JAMA Netw Open 2022 Dec 01;5(12):e2248559 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.48559] [Medline: 36574245]
30. On J, Park HA, Yoo S. Development of a prediction models for chemotherapy-induced adverse drug reactions: a retrospective observational study using electronic health records. Eur J Oncol Nurs 2022 Feb;56:102066. [doi: 10.1016/j.ejon.2021.102066] [Medline: 34861529]
31. Datta A, Flynn NR, Barnette DA, Woeltje KF, Miller GP, Swamidass SJ. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from a retrospective electronic health record (EHR) cohort. PLoS Comput Biol 2021 Jul 6;17(7):e1009053 [FREE Full text] [doi: 10.1371/journal.pcbi.1009053] [Medline: 34228716]
32. Gibson TB, Nguyen MD, Burrell T, Yoon F, Wong J, Dharmarajan S, et al. Electronic phenotyping of health outcomes of interest using a linked claims-electronic health record database: findings from a machine learning pilot project. J Am Med Inform Assoc 2021 Jul 14;28(7):1507-1517 [FREE Full text] [doi: 10.1093/jamia/ocab036] [Medline: 33712852]
33. Jeong E, Park N, Choi Y, Park RW, Yoon D. Machine learning model combining features from algorithms with different analytical methodologies to detect laboratory-event-related adverse drug reaction signals. PLoS One 2018;13(11):e0207749 [FREE Full text] [doi: 10.1371/journal.pone.0207749] [Medline: 30462745]
34. Zhao J, Henriksson A, Kvist M, Asker L, Boström H. Handling temporality of clinical events for drug safety surveillance. AMIA Annu Symp Proc 2015;2015:1371-1380 [FREE Full text] [Medline: 26958278]
35. Zhao J, Henriksson A. Learning temporal weights of clinical events using variable importance. BMC Med Inform Decis Mak 2016 Jul 21;16 Suppl 2(Suppl 2):71 [FREE Full text] [doi: 10.1186/s12911-016-0311-6] [Medline: 27459993]
36. Boland MR, Polubriaginof F, Tatonetti NP. Development of a machine learning algorithm to classify drugs of unknown fetal effect. Sci Rep 2017 Oct 09;7(1):12839 [FREE Full text] [doi: 10.1038/s41598-017-12943-x] [Medline: 28993650]
37. Wang Z, Poon J, Wang S, Sun S, Poon S. A novel method for clinical risk prediction with low-quality data. Artif Intell Med 2021 Apr;114:102052. [doi: 10.1016/j.artmed.2021.102052] [Medline: 33875163]
38. Li C, Chen L, Chou C, Ngorsuraches S, Qian J. Using machine learning approaches to predict short-term risk of cardiotoxicity among patients with colorectal cancer after starting fluoropyrimidine-based chemotherapy. Cardiovasc Toxicol 2022 Feb 18;22(2):130-140. [doi: 10.1007/s12012-021-09708-4] [Medline: 34792740]
39. Jin S, Kostka K, Posada JD, Kim Y, Seo SI, Lee DY, et al. Prediction of major depressive disorder following beta-blocker therapy in patients with cardiovascular diseases. J Pers Med 2020 Dec 18;10(4):288 [FREE Full text] [doi: 10.3390/jpm10040288] [Medline: 33352870]
40. Hansen PW, Clemmensen L, Sehested TS, Fosbøl EL, Torp-Pedersen C, Køber L, et al. Identifying drug-drug interactions by data mining: a pilot study of warfarin-associated drug interactions. Circ Cardiovasc Qual Outcomes 2016 Nov;9(6):621-628. [doi: 10.1161/CIRCOUTCOMES.116.003055] [Medline: 28263937]
41. Mosa AS, Rana MK, Islam H, Hossain AK, Yoo I. A smartphone-based decision support tool for predicting patients at risk of chemotherapy-induced nausea and vomiting: retrospective study on app development using decision tree induction. JMIR Mhealth Uhealth 2021 Dec 02;9(12):e27024 [FREE Full text] [doi: 10.2196/27024] [Medline: 34860677]

XSL•FO
RenderX

42. Herrin J, Abraham NS, Yao X, Noseworthy PA, Inselman J, Shah ND, et al. Comparative effectiveness of machine learning approaches for predicting gastrointestinal bleeds in patients receiving antithrombotic treatment. JAMA Netw Open 2021 May 03;4(5):e2110703 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10703] [Medline: 34019087]
43. Pichardo D, Michael R, Mercer M, Korina N, Onukwugha E. Utility of a clinically guided data-driven approach for predicting breast cancer complications: an application using a population-based claims data set. JCO Clin Cancer Inform 2022 Nov;6:e2100191 [FREE Full text] [doi: 10.1200/CCI.21.00191] [Medline: 36417684]
44. Puzhko S, Schuster T, Barnett TA, Renoux C, Munro K, Barber D, et al. Difference in patterns of prescribing antidepressants known for their weight-modulating and cardiovascular side effects for patients with obesity compared to patients with normal weight. J Affect Disord 2021 Dec 01;295:1310-1318. [doi: 10.1016/j.jad.2021.08.018] [Medline: 34706445]
45. Souissi SB, Abed M, Elhiki L, Fortemps P, Pirlot M. Reducing the toxicity risk in antibiotic prescriptions by combining ontologies with a multiple criteria decision model. AMIA Annu Symp Proc 2017;2017:1625-1634 [FREE Full text] [Medline: 29854233]
46. Zhou Y, Hou Y, Hussain M, Brown SA, Budd T, Tang WH, et al. Machine learning-based risk assessment for cancer therapy-related cardiac dysfunction in 4300 longitudinal oncology patients. J Am Heart Assoc 2020 Dec;9(23):e019628 [FREE Full text] [doi: 10.1161/JAHA.120.019628] [Medline: 33241727]
47. Goyal J, Ng DQ, Zhang K, Chan A, Lee J, Zheng K, et al. Using machine learning to develop a clinical prediction model for SSRI-associated bleeding: a feasibility study. BMC Med Inform Decis Mak 2023 Jun 11;23(1):105 [FREE Full text] [doi: 10.1186/s12911-023-02206-3] [Medline: 37301967]
48. Wang Y, Ma J, Ma S, Wang J, Li J. Causal evaluation of post-marketing drugs for drug-induced liver injury from electronic health records. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10340721] [Medline: 38083643]
49. Hughes JH, Tong DM, Burns V, Daly B, Razavi P, Boelens JJ, et al. Clinical decision support for chemotherapy-induced neutropenia using a hybrid pharmacodynamic/machine learning model. CPT Pharmacometrics Syst Pharmacol 2023 Nov 10;12(11):1764-1776 [FREE Full text] [doi: 10.1002/psp4.13019] [Medline: 37503916]
50. Sharma V, Joon T, Kulkarni V, Samanani S, Simpson SH, Voaklander D, et al. Predicting 30-day risk from benzodiazepine/Z-drug dispensations in older adults using administrative data: a prognostic machine learning approach. Int J Med Inform 2023 Oct;178:105177. [doi: 10.1016/j.ijmedinf.2023.105177] [Medline: 37591010]
51. Akimoto H, Hayakawa T, Nagashima T, Minagawa K, Takahashi Y, Asai S. Detection of potential drug-drug interactions for risk of acute kidney injury: a population-based case-control study using interpretable machine-learning models. Front Pharmacol 2023 May 23;14:1176096 [FREE Full text] [doi: 10.3389/fphar.2023.1176096] [Medline: 37288110]
52. Zhang J, Kummerfield E, Hultman G, Drawz PE, Adam TJ, Simon G, et al. Application of causal discovery algorithms in studying the nephrotoxicity of remdesivir using longitudinal data from the EHR. AMIA Annu Symp Proc 2022;2022:1227-1236 [FREE Full text] [Medline: 37128413]
53. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010 Nov 02;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
54. Chen H, Lundberg SM, Erion G, Kim JH, Lee SI. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. NPJ Digit Med 2021 Dec 08;4(1):167 [FREE Full text] [doi: 10.1038/s41746-021-00536-y] [Medline: 34880410]
55. Kaas-Hansen BS, Gentile S, Caioli A, Andersen SE. Exploratory pharmacovigilance with machine learning in big patient data: a focused scoping review. Basic Clin Pharmacol Toxicol 2023 Mar 03;132(3):233-241 [FREE Full text] [doi: 10.1111/bcpt.13828] [Medline: 36541054]
56. Schuemie M, Chen Y, Fridgeirsson E, Kim C, Reps J, Suchard M, et al. Assessment of pre-trained observational large longitudinal models in OHDSI (APOLLO). Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/wp-content/uploads/2023/10/110_Schuemie-BriefReport.pdf [accessed 2024-04-29]
57. Sharma V, Kulkarni V, Eurich DT, Kumar L, Samanani S. Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada. BMJ Open 2021 May 26;11(5):e043964 [FREE Full text] [doi: 10.1136/bmjopen-2020-043964] [Medline: 34039572]
58. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2020 Jul 13;2(7):369-375. [doi: 10.1038/S42256-020-0197-Y]
59. Belthangady C, Giampanis S, Jankovic I, Stedden W, Alves P, Chong S, et al. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. Nat Commun 2022 Nov 14;13(1):6921 [FREE Full text] [doi: 10.1038/s41467-022-33732-9] [Medline: 36376286]
60. Jiang G, Liu H, Solbrig HR, Chute CG. Mining severe drug-drug interaction adverse events using semantic web technologies: a case study. BioData Min 2015 Mar 25;8(1):12 [FREE Full text] [doi: 10.1186/s13040-015-0044-6] [Medline: 25829948]
61. Pacaci A, Gonul S, Sinaci AA, Yuksel M, Laleci Erturkmen GB. A semantic transformation methodology for the secondary use of observational healthcare data in Postmarketing safety studies. Front Pharmacol 2018 Apr 30;9:435 [FREE Full text] [doi: 10.3389/fphar.2018.00435] [Medline: 29760661]
62. Home page. The European Health Data & Evidence Network. URL: https://www.ehden.eu/ [accessed 2024-04-29]

XSL•FO
RenderX

Abbreviations

ADR: adverse drug reaction
AI: artificial intelligence
ANN: artificial neural network
ConvSCCS: convolutional self-controlled case series
DARWIN: Data Analysis and Real World Interrogation Network
EHR: electronic health record
FUTURE-AI: Fairness, Universality, Traceability, Usability, Robustness, and Explainability–Artificial Intelligence
ICSR: individual case safety report
KG: knowledge graph
ML: machine learning
NLP: natural language processing
NSCCS: neural self-controlled case series
OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RF: random forest
RWD: real-world data
SCCS: self-controlled case series
XAI: explainable artificial intelligence

Edited by A Coristine; submitted 28.02.24; peer-reviewed by J Fossouo Tagne, T Goto, C Hohl; comments to author 15.08.24; revised version received 03.10.24; accepted 27.10.24; published 30.12.24.

Please cite as:
Dimitsaki S, Natsiavas P, Jaulent MC
Applying AI to Structured Real-World Data for Pharmacovigilance Purposes: Scoping Review
J Med Internet Res 2024;26:e57824
URL: https://www.jmir.org/2024/1/e57824
doi:10.2196/57824
PMID:

©Stella Dimitsaki, Pantelis Natsiavas, Marie-Christine Jaulent. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Use of Extrinsic Motivators to Improve the BMI of Obese or Overweight Adolescents: Systematic Review

Ana Gonçalves[1,2], MD; Pedro Augusto Simões[2,3], MD, PhD; Bernardo Sousa-Pinto[4,5], MD, PhD; Tiago Taveira-Gomes[5], MD, PhD

[1]Faculty of Medicine, Universidade do Porto, Porto, Portugal
[2]Faculty of Health Sciences, University of Beira Interior, Covilha, Portugal
[3]Health Sciences Research Centre - University of Beira Interior, University of Beira Interior, Covilha, Portugal
[4]Department of Community Medicine, Information and Health Decision Sciences, Faculty of Medicine of the University of Porto, Porto, Portugal
[5]Centre for Health Technology and Services Research, Health Research Network, Faculty of Medicine of the University of Porto, Porto, Portugal

Corresponding Author:
Ana Gonçalves, MD
Faculty of Medicine
Universidade do Porto
Alameda Prof. Hernâni Monteiro
Porto, 4200-319
Portugal
Phone: 351 225513600
Email: an.goncalves@ubi.pt

Abstract

Background: The prevalence of overweight and obesity is increasing at an alarming rate in children and adolescents worldwide. Given the dimensions of the problem, the treatment of childhood obesity is considered extremely important. Current evidence indicates that behavioral and cognitive behavioral strategies combined with diet and physical activity approaches may assist in reducing adolescent obesity.

Objective: The purpose of this systematic review was to evaluate the use of extrinsic motivators for improving the BMI of obese or overweight adolescents.

Methods: The inclusion criteria were as follows: (1) overweight or obese adolescents, (2) intervention using extrinsic motivators, and (3) outcome variables related to weight status. The exclusion criterion was the presence of an associated chronic disease. The search process was conducted in PubMed and Web of Science (last searched on April 23, 2023). The risk of bias was evaluated independently by 2 authors using Cochrane tools (RoB2 [randomized controlled trials], ROBINS-I, and ROBINS-E).

Results: From 3163 studies identified, 20 articles (corresponding to 18 studies) were included in the analysis. The studies differed in terms of study design, sample size, follow-up duration, outcomes reported, and extrinsic motivators used. Most of the studies had videogames or apps as interventions. Of the 18 studies, 9 (50%) reported a statistically significant decrease in BMI. The most used extrinsic motivators were “motivation” (n=13), “feedback” (n=10), and “rewards” (n=9). Among the motivators, “reminders” (100%) and “peer-support” (80%) appeared to have high impacts on BMI reduction.

Conclusions: The heterogeneity of the included studies made analysis difficult. No study evaluated extrinsic motivators in isolation. Most of the studies had a moderate or high risk of bias. The extrinsic motivators “reminders” and “peer-support” appeared to be useful. Further studies are needed, and these should include well-designed randomized controlled trials and studies involving homogeneity in BMI measures, consistent extrinsic motivator definitions, and longer durations to better understand the long-term impacts of extrinsic motivators on weight management success.

(J Med Internet Res 2024;26:e57458) doi:10.2196/57458

KEYWORDS

adolescents; obesity; overweight; extrinsic motivators; body mass index

XSL•FO
RenderX

Introduction

The prevalence of overweight and obesity is increasing at an alarming rate in children and adolescents worldwide, and the rise has occurred similarly among both boys and girls [1,2]. Over 340 million children and adolescents aged 5-19 years were overweight or obese in 2016 [2]. In the United States, in 2020-2021, 17% of children and adolescents aged 10-17 years had obesity, with rates significantly higher for non-Hispanic Black (22.9%), Hispanic (22.4%), and non-Hispanic American Indian/Alaska Native (20.5%) individuals [3]. Similarly, in China, in 2020, the prevalence of overweight (including obesity) was 19% for children aged 6-17 years [4], and in Europe, in 2018, 19% of children aged 15 years were either overweight or obese [5]. The data available in Portugal are different, with a prevalence of overweight (including obesity) between 20% and 40% [6].

The widespread obesity among young people is generating significant social concern for public health, including not only the number of children affected but also the consequences [7]. A raised BMI is a major risk factor for cardiovascular diseases, diabetes, musculoskeletal disorders, and some cancers, and the risk for these diseases increases with an increase in BMI [2]. Obesity also has psychological consequences, such as low self-esteem, sadness, nervousness, and a negative self-image [7]. Overweight and obesity, as well as their related diseases, are largely preventable [2]. The global prevalence of obesity in children and adolescents almost doubled between 1980 and 2015 [8]. In view of the high and increasing prevalence, an early and effective intervention is urgently required [9].

Adolescence is a vulnerable period for the development of obesity, and adolescent weight tracks strongly into adulthood. Adolescence is also a unique period of development, with puberty resulting in considerable physical, hormonal, and psychosocial changes, which need to be considered when evaluating the efficacy of treatment programs [10].

Given the dimensions of the problem, the treatment of childhood obesity is considered extremely important. Fundamental approaches involve changing food intake and quality, increasing effort in physical activity, and reducing sedentary habits [7]. Current evidence indicates that behavioral and cognitive behavioral strategies combined with diet and physical activity approaches may assist in reducing adolescent obesity [10].

According to the self-determination theory, extrinsic motivation is a type of motivation in which the behavior is a means of achieving external outcomes. There are 4 different types of extrinsic motivations, which vary on a spectrum according to the degree of internalization and autonomy: external regulation (motivated by external controls prescribed by others, like gaining rewards or avoiding punishment); introjected regulation (motivated by internal pressure from internalized constructs of external controls, such as feeling self-approval or avoiding feeling guilty); identified regulation (motivated because behavior is perceived as important and useful, like exercise to get healthy or lose weight); and integrated regulation (motivated because the behavior is in concordance with one's values and sense of self, such as feeling of identification) [11]. Some examples of extrinsic motivators are rewards/points, avatars, challenges, and feedback [11-13].

To our knowledge and research, no review has studied extrinsic motivators as a whole in weight loss among obese adolescents. Most studies have analyzed the impact of specific interventions (eg, active video gaming) on weight or BMI. Byrne et al [14] found that motivation was also affected by feedback from a mobile app that involved caring for a virtual pet. Andrade et al [15] found that "owing to the fun, pleasure, joy, and innovative technology provided by exergames, they have the potential to attract children and adolescents who are overweight or obese."

The purpose of this systematic review was to evaluate the use of extrinsic motivators for improving the BMI of obese or overweight adolescents.

Methods

Reporting Guidelines

This review has been reported according to the recommendations of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Eligibility Criteria

The review question was framed using the PICO framework, and the following eligibility criteria were used: (1) Population: overweight (P85-95 for age and gender) or obese (≥P95 for age and gender) adolescents (age range 10-19 years; according to the World Health Organization [WHO]); (2) Intervention: extrinsic motivators; (3) Comparison: usual care or other interventions; and (4) Outcome: variables related to weight status (weight or BMI z-scores).

We only included published primary research studies in Portuguese, Spanish, English, French, and Italian. The exclusion criterion was presence of associated chronic disease.

Information Sources

The search process was conducted in PubMed and Web of Science (last searched on April 23, 2023).

Search Strategy

The databases PubMed and Web of Science were searched using the following terms: adolescent (a validated search filter) [16], obesity, overweight, pediatric obesity, weight loss, gamification, and telemedicine.

The search on PubMed was carried out for the title, abstract, and MeSH terms, and the search on Web of Science was carried out for the title, abstract, and keywords. The full strategy used is presented in Multimedia Appendix 1.

The references of the included studies and other reviews were analyzed for eligibility.

Selection Process

Titles and abstracts identified from the search were transferred to the site "Rayyan" for the relevance assessment process and deduplication. Potentially eligible studies were first screened independently by 2 authors (AG and PAS) according to the title and abstract to evaluate if they met the inclusion criteria. Then,

XSL•FO
RenderX

the full texts were read independently by the same 2 authors, and a decision was made about inclusion in the review. Disagreements were resolved by consensus among the authors.

Data Collection Process

The data were collected from each report independently by 2 authors (AG and PAS) with Excel (Microsoft Corp). Disagreements were resolved by consensus among the authors. A set of variables was chosen from consensus among all authors, and the following data were systematically extracted for each eligible article: bibliographic information, study design, target population, inclusion criteria, exclusion criteria, sample size, demographic characteristics, duration of the study/follow-up/time of data collection, purpose, intervention, extrinsic motivator, and outcome BMI.

Study Risk of Bias Assessment

The risk of bias was evaluated with Cochrane tools: RoB2 for randomized controlled trials (RCTs), ROBINS-I for interventional studies, and ROBINS-E for exposure studies. Two authors (AG and PAS) independently assigned the risk of bias to each study, and disagreements were resolved by consensus among the authors.

Results

Included Studies

In total, 3163 studies were identified from the searched databases, and after removing duplicates, 2653 articles were left for the following steps. Moreover, 2558 articles were deleted after analyzing the title and abstract. Among the remaining articles, 92 were retrieved for more detailed assessment (3 articles could not be retrieved after contacting the authors), and they were evaluated considering the inclusion and exclusion criteria. After the evaluation of the full text, 72 articles were excluded for not meeting the inclusion criteria (population, intervention, outcome, and language). Finally, 20 articles (corresponding to 18 studies) met the inclusion criteria and were included in the analysis [17-36].

Figure 1 shows the steps of the study selection process following the PRISMA statement. The PRISMA checklist is presented in Multimedia Appendix 2.

Figure 1. Flow diagram showing the article selection process.

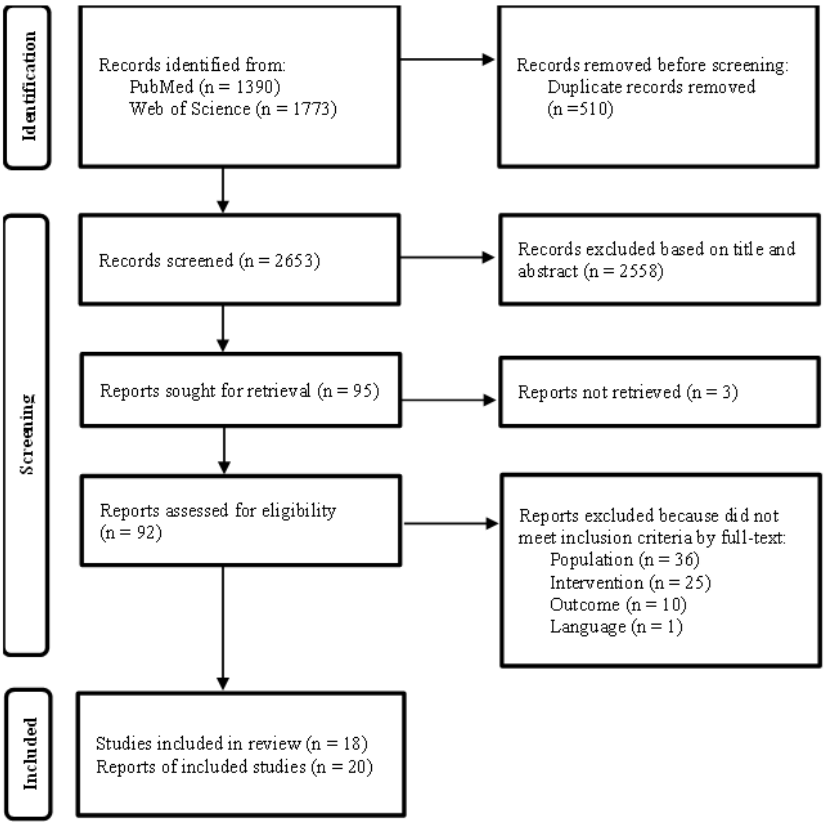

Study Characteristics

The studies included were published between 2006 and 2022 (12 in the last 10 years) and were performed in 8 different countries: 11 in the United States, 1 in China, 1 in Singapore, 1 in Korea, 1 in Canada, 1 in Switzerland, 1 in Italy, and 1 in New Zealand.

Risk of Bias in the Studies

There was heterogeneity in the design of the studies: 13 studies were clinical trials (9 RCTs), 4 were cohort studies, and 1 was a retrospective observational study. Detailed findings concerning the risk of bias according to the ROB-2, ROBINS-I, and ROBINS-E tools are presented in Figure 2 [17-23,28,31], Figure 3 [29,30,32,33], and Figure 4 [24-27,34], respectively. Most of the RCTs presented a moderate risk of bias, and all other study designs had a high or serious risk.

XSL•FO
RenderX

Figure 2. Risk of bias according to the ROB-2 tool.

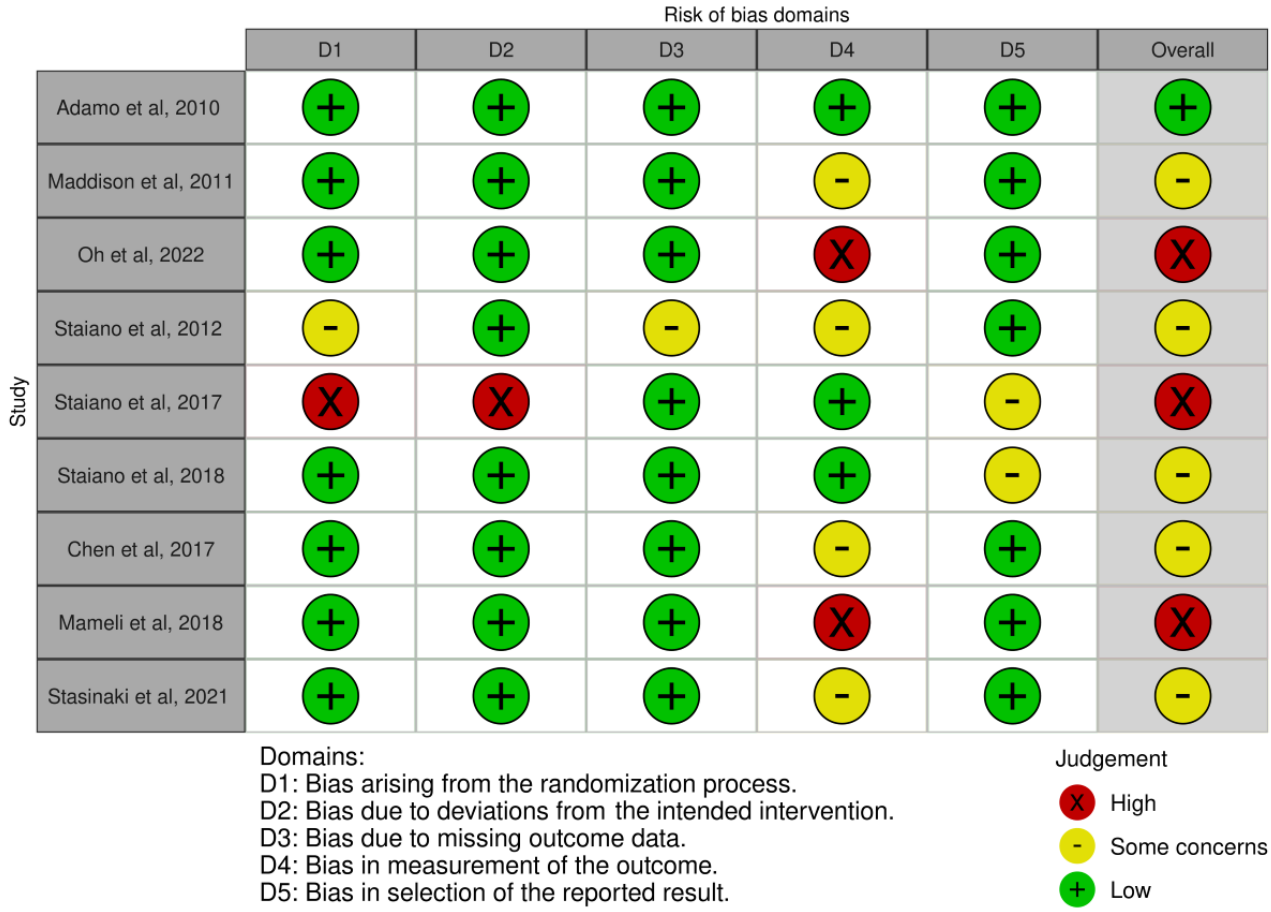

Figure 3. Risk of bias according to the ROBINS-I tool.

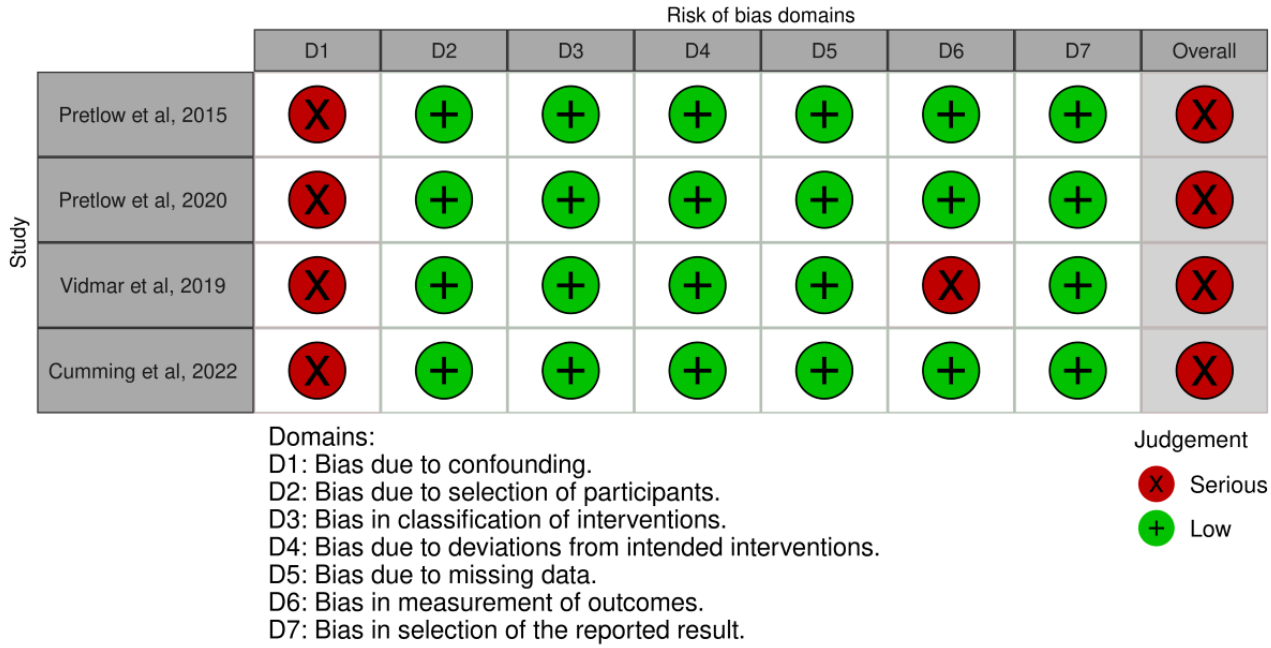

XSL•FO
RenderX

Figure 4. Risk of bias according to the ROBINS-E tool.

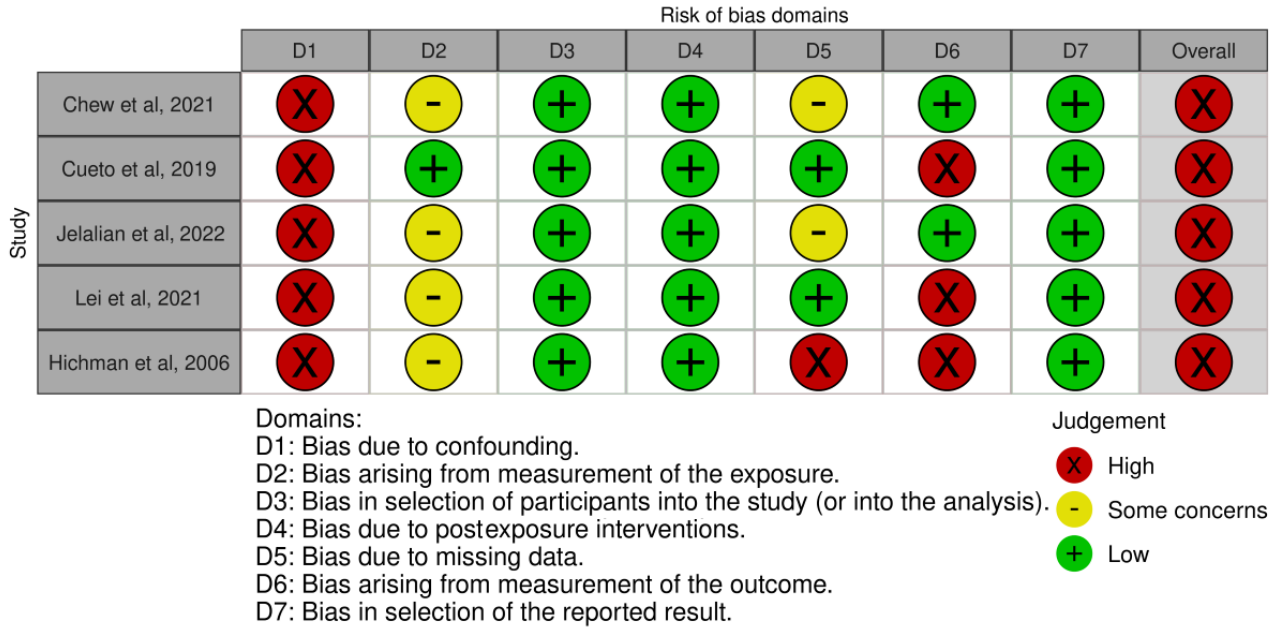

Results of Individual Studies

The characteristics of the included studies are summarized in Table 1.

Regarding the sample size, we verified that 14 studies had less than 100 participants (minimum 24 and maximum 85), 1 had 322 participants, and 3 had more than 1000 participants (minimum 1120 and maximum 3500). The follow-up of the studies varied from 3 weeks to more than 18 months. However, most of the studies had a follow-up of 3 to 6 months.

XSL•FO
RenderX

Table 1. Main characteristics of the selected studies.

Author, year, and country	Study design	Target population	Subjects	Follow-up	Intervention	Control	Extrinsic motivator
Adamo et al [17], 2010; Canada	RCT[a]	• IC[b]: 12-17 years and pBMI[c] >95 or pBMI >85 with high risk of type 2 diabetes or cardiovascular disease • EC[d]: Medication or supplement that affects body composition, regular physical activity, psychiatric illness, and contraindication to physical activity	• Sample: 26 (13 I[e], 13 C[f]) • Gender: 53.8% (I, C) female • Mean age: 13.9 (SD 1.4) (I), 15.1 (SD 1.8) (C) years • pBMI: 97.9 (SD 1.5) (I), 97.8 (SD 2.7) (C) • Ethnicity: Not mentioned	2.5 months	Active videogame cycling	Stationary cycling to music	• Feedback • Select games
Maddison et al [18], 2011; New Zealand	RCT	• IC: 10-14 years and overweight/obese • EC: Own active videogames or contraindication for physical activity	• Sample: 322 (160 I, 162 C) • Gender: 72.5% (I), 73.5% (C) male • Mean age: 11.6 (SD 1.1) years (I, C) • zBMI[g]: 1.3 (SD 1.1) (I, C) • Ethnicity: 57.5% (I), 56.2% (C) NZ[h] European	6 months	Active videogame	Nonactive videogames	• Avatar • Select games
Oh et al [19], 2022; Korea	RCT	• IC: 10-17 years and pBMI >85 • EC: Pharmacologic treatment, sign of hormonal/orthopedic/metabolic/cardiovascular illness, and physical activity outside school	• Sample: 24 (12 I, 12 C) • Gender: 83.33% (I, C) male • Mean age: 13.58 (SD 1.44) (I), 13.08 (SD 2.35) (C) years • BMI: 24.99 (SD 1.35) (I), 25.03 (SD 1.92) (C) • Ethnicity: Not mentioned	3 weeks	Active videogame SUIKA (Super Kids Adventure)	Active videogame NINS (Nintendo Switch "Ring Fit Adventure")	• Avatar • Feedback • Reminder • Rewards
Staiano et al [20], 2012; United States	RCT	• IC: African American, 15-19 years, and pBMI ≥75 • EC: Not mentioned	• Sample: 54 (19 CP[i], 19 CO[j], 16 C) • Gender: 55.6% female • Mean age: Not mentioned • pBMI: 94.7 (SD 6.0) • Ethnicity: African American	5 months	Active videogame (Nintendo Wii Active): cooperative or competitive	Usual daily activities	• Motivation • Peer-support • Rewards
Staiano et al [21], 2017; United States	RCT	• IC: 14-18 years, pBMI ≥85, female, and postmenarcheal period • EC: Pregnancy, contraindication for physical activity, recent hospitalization for mental illness, and inability to comply with study procedures	• Sample: 41 (22 I, 19 C) • Gender: 100% female • Mean age: 15.3 (SD 1.2) (I), 16.1 (SD 1.4) (C) years • pBMI: 97.4 (SD 2.9) (I), 97.1 (SD 3.3) (C) • Ethnicity: 64% (I), 61% (C) African American	3.5 months	Active videogame	Self-directed care control	• Motivation • Rewards • Select games

XSL•FO
RenderX

Author, year, and country	Study design	Target population	Subjects	Follow-up	Intervention	Control	Extrinsic motivator
Staiano et al [22], 2018; United States	RCT	• IC: 10-12 years and pBMI ≥85 • EC: Pregnancy and contraindication for physical activity	• Sample: 46 (23 I, 23 C) • Gender: 54% male • Mean age: 11.2 (SD 0.8) years • zBMI: 2.06 (SD 0.46) (I), 2.10 (SD 0.42) (C) • Ethnicity: 57% African American, 41% White	6 months	Active videogame	Usual level of physical activity	• Challenges • Feedback • Motivation • Select games
Chen et al [23], 2017; United States	RCT	• IC: 13-18 years and BMI ≥P85 • EC: Not mentioned	• Sample: 40 (23 I, 17 C) • Gender: 58% (I), 53% (C) male • Mean age: 15 (SD 1.69) (I), 14.77 (SD 1.60) (C) years • zBMI: 1.6 (SD 0.50) (I), 1.54 (SD 0.42) (C) • Ethnicity: 90% Chinese American	6 months	App (iStart Smart for Teens online program)	Online program's content with information from an app	• Motivation
Chew et al [24], 2021; Singapore	Cohort	• IC: 10-17 years and pBMI >85 • EC: Secondary causes of obesity, parents are non-English speakers, and logistic problems	• Sample: 40 • Gender: 58% male • Mean age: 13.8 (SD 1.7) years • BMI: 80% obesity • Ethnicity: 45% Chinese, 33% Malay	6 months	App (Kurbo)	No	• Feedback • Motivation
Cueto et al [25], 2019; United States	Cohort	• IC: 5-19 years • EC: pBMI <85 at baseline and data measurement errors	• Sample: 1120 • Gender: 68.04% female • Mean age: 12.0 (SD 2.5) years • pBMI: 96.6 (SD 3.1) • Ethnicity: Not mentioned	4 weeks, 3-4 months, and 6 months	App (Kurbo)	No	• Feedback • Motivation
Jelalian et al [26], 2022; United States	Cohort	• IC: 5-17 years and pBMI >85 • EC: Missing data, follow-up age >240.5 months, and measurement errors	• Sample: 3500 • Gender: 71.3% female • Mean age: 12.79 (SD 2.41) years • BMI: 29.65 (SD 5.56); %BMIp95: 116.31 (SD 19.57) • Ethnicity: Not mentioned	0-3 months, 3-6 months, 6-12 months, 12-18 months, and >18 months	App (Kurbo)	No	• Feedback • Motivation
Lei et al [27], 2021; China	Observational	• IC: 10-17 years and pBMI ≥85 • EC: Secondary causes of obesity, pharmacologic/cirurgic intervention for weight reduction, logistic problems, and outliers	• Sample: 2825 • Gender: 54.8% female • Mean age: 14.4 (SD 2.2) years • pBMI: 48% ≥97 • Ethnicity: Not mentioned	4 months	App (MetaWell)	No	• Challenges • Feedback
Mameli et al [28], 2018; Italy	RCT	• IC: 10-17 years, pBMI ≥95, and Caucasian ethnic group • EC: Genetic/syndromic obesity, psychiatric disease, and any condition compromising the ability to walk		3 months	App (MeTeDa)	Mediterranean diet plus physical activity	• Challenges • Feedback • Motivation

XSL•FO
RenderX

Author, year, and country	Study design	Target population	Subjects	Follow-up	Intervention	Control	Extrinsic motivator
			• Sample: 30 (16 I, 14 C) • Gender: 68.8% (I), 57.1% (C) male • Mean age: 12.6 (SD 1.7) (I), 12.4 (SD 2.2) (C) years • BMI-SDS[k]: 2.20 (SD 0.47) (I), 2.09 (SD 0.34) (C) • Ethnicity: Caucasian				
Pretlow et al [29], 2015; United States	Pilot clinical trial without control	• IC: Child/adolescent with overweight/obesity • EC: Motivation scores <50/100 and logistic problems	• Sample: 43 • Gender: 65.1% female • Mean age: 16 (SD 0.43) years • pBMI: 37.2% >98 • Ethnicity: 83.7% White	5 months	App	No	• Motivation • Peer-support • Rewards
Pretlow et al [30], 2020; United States	Clinical trial without control	• IC: 10-20 years, pBMI ≥95, and motivation score >50/100 • EC: Not mentioned	• Sample: 35 • Gender: 51.4% female • Mean age: 13.8 (SD 0.52) years • pBMI: 34.3% ≥99 • Ethnicity: 65.7% Caucasian	4 months	App	No	• Motivation • Peer-support • Reminder • Rewards
Stasinaki et al [31], 2021; Switzerland	RCT	• IC: 10-18 years, pBMI >97 or >90 with ≥1 risk factor or co-morbidity • EC: Not controlled psychiatric disease, medication that affects weight, and participation in another obesity treatment program in the past year	• Sample: 31 (18 I, 13 C) • Gender: 61.1% (I), 53.8% (C) male • Mean age: 12.6 (I), 13.7 (C) years • BMI-SDS: 2.5 (I, C) • Ethnicity: Not mentioned	12 months	App (PathMate2) + multicomponent behavior change intervention	Multicomponent behavior change intervention	• Challenges • Feedback • Motivation • Rewards
Vidmar et al [32], 2019; United States	Pilot clinical trial	• IC: 12-18 years and YFAS-c[l] + • EC: Obesity comorbidities, psychiatric disease or developmental delay, and inability to read English	• Sample: 35 (18 I, 17 C) • Gender: 72.22% (I), 47.06% (C) female • Mean age: 14.44 (SD 1.65) (I), 14.35 (SD 1.77) (C) years • BMI: Not mentioned • Ethnicity: 61.11% (I) vs 64.71% (C) Hispanic	6 months	App (W8Loss2Go)	EMPOWER program	• Motivation • Peer-support • Rewards
Cummings et al [33], 2022; United States	Clinical trial without control	• IC: 13-18 years and pBMI >90 • EC: Contraindication for use scanner, psychiatric medication, pregnancy, comorbidities, and visual acuity that could not be corrected	• Sample: 26 • Gender: 54% male • Mean age: 14.81 (SD 1.59) years • pBMI: 97.07 (SD 1.85) • Ethnicity: 93% White	3 months	Digital health program	No	• Challenges • Feedback • Motivation • Rewards
Hinchman et al [34], 2006; United States	Cohort	• IC: 11-17 years, completion of Operation Zero in 2001-2002, at risk of overweight, and overweight • EC: Not mentioned		12 months	Operation Zero	Members who never attended Operation Zero	• Challenges • Peer-support • Rewards

XSL•FO
RenderX

Author, year, and country	Study design	Target population	Subjects	Follow-up	Intervention	Control	Extrinsic motivator
			• Sample: 85 (43 I, 42 C) • Gender: 56% (I), 45% (C) female • Age: 56% (I), 57% (C) ≥13 years • pBMI: 98% (I), 85% (C) >95 • Ethnicity: Not mentioned				

[a]RCT: randomized controlled trial.
[b]IC: inclusion criteria.
[c]pBMI: BMI percentile.
[d]EC: exclusion criteria.
[e]I: intervention.
[f]C: control.
[g]zBMI: BMI z-score.
[h]NZ: New Zealand.
[i]CP: competitive arm.
[j]CO: cooperative arm.
[k]BMI-SD: BMI standard deviation.
[l]YFAS-c: Yale Food Addiction Scale for Children.

Regarding the demographic characteristics of the participants, 10 studies had a sample composed mainly of female adolescents (1 of them had only female participants). The mean age reported by the studies varied from 11.2 to 16 years. Moreover, 11 studies reported information about ethnicity (8 in the United States, 1 in Italy, 1 in Singapore, and 1 in New Zealand), and the majority of participants were Caucasian and African American.

Interventions of the Studies

Adamo et al [17] assessed an intervention that involved 60-minute sessions 2 times per week with a GameBike interactive video gaming system (with a handlebar-mounted game controller) interfaced with a Sony PlayStation 2. It was designed to determine the participant's speed by cycling cadence, and faster pedaling was associated with faster movement in the virtual world on screen. The participants were allowed to select videogames to play while cycling and were permitted to switch games during the exercise sessions. There was no nutritional intervention or parental involvement. The participants were weighed in a lab at the end of the sessions.

In the study by Maddison et al [18], participants assigned to the intervention received a Sony PlayStation EyeToy, a dance mat, and a selection of active videogames (eg, Play3, Kinetic, Sport, and Dance Factory), from which they could choose 5 different games. The motion camera placed a picture of the gamer on the screen, which the gamer then interacted with. The participants were asked to achieve the recommended 60 minutes of moderate-to-vigorous physical activity on most days of the week by supplementing periods of inactivity with active videogame play and substituting periods of traditional nonactive videogame play with the active version. There was no nutritional intervention or parental involvement. The participants were weighed at a central location (at baseline, and 12 and 24 weeks) by the researchers.

Oh et al [19] created an intervention using the Super Kids Adventure game that provides upper extremity stretching, lunging, boxing, side-bending, squatting, and arm and jumping exercises. They asked participants to perform activities for 30 minutes 5 times per week at an intensity of about 4-6 metabolic equivalents of task. There was no intervention regarding nutrition. The participants could select various characters to play, and there was real-time visual and auditory feedback given by the game. When the participants performed the correct posture for any of the exercises, the character moved and scored. There was an alarm function that reminded the participants to exercise regularly. They were compensated according to the quantity of training they performed and maintained to progress. There was no parental involvement. The participants were weighed after all the sessions.

Staiano et al [20] created an intervention that consisted of 30-60 minutes of active videogame play every school day. The study included a group that performed the activity in a cooperative way (earn the most points and expend the most calories as a team) and another that performed it in a competitive way (compete against their opponent to earn the most points and expend the most calories). There was no nutritional intervention. During the sessions, the coordinators encouraged completion of each daily exergame routine through periodic verbal reinforcement. There was no parental involvement. Weight was measured at the school-based wellness clinic.

Staiano et al [21] assessed an intervention that consisted of three 1-hour exergaming dancing sessions per week. The participants were allowed to self-select the games, songs, dance mode, intensity level, and dance partner. Gaming coaches provided ongoing motivation, and gifts were provided throughout the 12 weeks to encourage attendance and motivate participants to exercise at a high level. There was no nutritional intervention

XSL•FO
RenderX

or parental involvement. The participants were weighed at clinic visits (baseline and end points).

Staiano et al [22] provided the intervention group with a Kinect and Xbox 360 gaming console, a 24-week Xbox Live subscription, 4 exergames (Your Shape: Fitness Evolved 2012, Just Dance 3, Disneyland Adventures, and Kinect Sports Season 2), and a Fitbit Zip to wear. The participants were encouraged to meet a goal of 60 minutes per day of moderate-to-vigorous physical activity for 3 days a week with a family member or friend. The participants could choose between the 4 active videogames provided. They received 3 challenges each week with increasing intensity, difficulty, and duration. They also had telehealth coaching: meeting with a fitness coach (video chat or exergame console) on a weekly basis for the first 6 weeks and biweekly thereafter with individualized feedback, motivation, and encouragement to meet the goal and help to create solutions to barriers for physical activity. There was no nutritional or parental involvement. The participants were weighed at the clinic (at baseline and end point).

In the study by Chen et al [23], the intervention group received a Fitbit Flex and downloaded an app and a link to the iStart Smart for Teens program to their mobile phone. The Fitbit Flex recorded and tracked their physical activity, sedentary activity, and food intake in a diary. The iStart Smart for Teens online educational program had 8 modules (participants were asked to perform 1 module per week) that focused on nutrition and physical activity. The participants also received instructions by mobile phone or computer and supplementary information and tips by app messages. There was no parental involvement. The participants were weighed in the clinic (at baseline, and 3 and 6 months).

Chew et al [24], Cueto et al [25], and Jelalian et al [26] used the Kurbo intervention, which is a mobile app developed to aid adolescents and their families with weight management through dietary self-monitoring (promotion of the gradual reduction of high-calorie foods over time by using the traffic light diet to categorize foods), physical activity behaviors (recommended 60 minutes of moderate-to-vigorous physical activity each day), and weekly individualized coaching sessions (via video, phone, or text) with feedback. Chew et al [24] asked the participants to weigh themselves at least weekly, and they were weighed at clinic visits (at 3 and 6 months). In the other 2 studies, the participants were weighed at baseline and the end point.

Lei et al [27] created an intervention that consisted of a nutritional program based on calorie restriction with meal replacement and individualized low-calorie meal plans. Physical activity was encouraged but with no specific exercise recommendations. The participants had access to their weight loss progress and snapshots of their current health data and optimal measures of BMI. There was parental involvement. The participants were asked to weigh themselves on a daily basis (wireless scale).

In the study by Mameli et al [28], the intervention group received a wristband that they were required to use for at least 5 days per week (to measure energy expenditure) and an app that allowed real-time recording of food consumption (to measure energy intake) by asking them to enter the raw foods consumed into the app. It had a visual database of foods and 3 portion sizes (small, medium, and large) for each food. The participants received personalized lifestyle programs based on previous week energy intake (for nutritional recommendations) and expenditure (for physical activity recommendations). There was weekly feedback on the adequacy of the diet and physical activity (compliance with the diet, energy gap, sedentary time, physical activity level, and quality of the diet) via text messages. The participants received suggestions on how to reach each of the 5 goals. Positive feedback was included in the text messages every time a participant reached at least one goal, with specification of the reached goals. There was parental involvement. The participants were weighed at clinic visits (at 1, 2, and 3 months).

Pretlow et al [29] and Vidmar et al [32] focused on sequential withdrawal from problem foods, snacking, and excessive food amounts at meals. The participants had 4 face-to-face group meetings, weekly phone meetings, and text messages 5 days per week. The participants also had access to the mentor's contact details and peer support through app bulletin boards and "weight loss buddies" (chat). There was a monetary incentive proportional to the completion of the requirements of the study. There was no physical activity intervention or parental involvement. The participants were asked to weigh themselves daily (wireless scale), and they were also weighed during face-to-face meetings.

Pretlow et al [30] developed an intervention similar to that in another study [29], in which they added a motor addiction component that consisted of (1) viewing aversive photos/videos or snapping a rubber band against the wrist to quell eating urges, (2) stress reduction, (3) avoiding triggers, (4) relaxation techniques, (5) competing behaviors, (6) distractions, and (7) distress tolerance. Another motor addiction method was the Worry List, a stress reduction feature that prompted participants to journal their current worries and create an action plan for each worry. The participants received daily notifications to login and answer questions about their eating behavior and if they used any sensory and motor addiction treatment methods (if they reported addictive eating behaviors, the app asked why this had happened and what was the participant's plan to keep this from happening again). Moreover, they received weekly app prompts to update their worry lists and plans.

Stasinaki et al [31] created individual multi-component behavior changing interventions following Swiss guidelines, including handouts on nutritional education and physical activity, and used the support of the PathMate2 app. The participants had daily challenges (eg, number of steps per day). The app included 2 chat channels: virtual coach (daily interaction, encouraging them to achieve challenges to earn virtual rewards) and human coach (maximum 10 minutes during all interventions). In the app dashboard, the participants could also see their progress. There were 4 on-site visits and 2 remote counseling sessions (via telephone). There was no parental involvement. The participants were weighed during clinical visits (at 6 and 12 months).

Cummings et al [33] gave participants a Fitbit device to track daily goals of >60 active minutes or ≥10,000 steps of exercise

XSL•FO
RenderX

(participants could choose which one they wanted to achieve). Weekly goals depended upon the level of activity in the prior weeks. The participants received feedback on daily and weekly progress with daily texts about whether they met their goal the previous day (praising them if they met their goal or reminding them to do so if they did not meet their goal) and received weekly texts about whether they met their weekly goal (praising them if they did and providing encouragement if they did not). These texts also informed adolescents of the amount of incentives earned that week and the goal for the upcoming week. There was monetary reinforcement (debit card once per week). There was no nutritional intervention or parental involvement. The participants were weighed during clinic visits (at baseline and end point).

Hinchman et al [34] used a family-oriented approach and incorporated behavior change strategies to address the behaviors, knowledge, attitudes, and self-efficacy of patients and their parents regarding nutrition and physical activity. In the nutritional intervention, they aimed to increase milk consumption until drinking 4 glasses a day, decrease milk fat until drinking fat-free milk, increase fruit and vegetable servings until eating 5 servings a day, and promote eating breakfast every morning. In the physical activity intervention, they aimed to increase the number of days being physically active for 60 minutes until activity for 5 days a week, decrease sedentary behavior to less than 1 hour per day, and increase the number of steps per week on a pedometer until taking 70,000 steps per week. The participants and their parents had weekly 1-hour group appointments (with a health educator/clinician, nursing staff, and a dietitian or chef) that included pedometer games, interactive learning, competitions (for prizes), cooking demonstrations, and group exercises, and involved defining the weekly goals for nutrition and physical activity. The participants and their parents received the Operation Zero manual (with health education, activities, and recipes). The participants were weighed weekly during the appointments.

In short, the interventions focused on only physical activity in 7 studies, only nutrition in 4 studies, and both in 7 studies. Seven studies had parental involvement, and in 3 studies, participants took their own weights.

Regarding extrinsic motivators, the most used was “motivation” (by coaches, messages, and sessions), followed by “feedback” and “rewards” (that included points and money). In the studies that used videogames as the intervention, the most used motivators were “select games,” “rewards,” and “feedback.” In the studies that used an app as the intervention, the most used motivators were “motivation,” “feedback,” and “rewards.”

Outcomes

The outcomes related to BMI reported for interventions and controls are summarized in Table 2.

XSL•FO
RenderX

Table 2. Outcomes related to BMI reported for participants and controls in the selected studies.

Study and country, and outcome	Intervention group		Control group		Intervention group vs control group	
	Outcome data	*P* value	Outcome data	*P* value	Outcome data	*P* value
Adamo et al [17], 2010; Canada						
BMI (B[a] vs E[b])	35.5 (SD 9.3) vs 35.5 (SD 9.7)	>.05	39.3 (SD 8.9) vs 39.4 (SD 8.9)	>.05	__[c]	—
pBMI[d] (B vs E)	97.8 (SD 1.4) vs 97.5 (SD 1.8)	>.05	97.8 (SD 2.7) vs 97.8 (SD 2.3)	>.05	—	—
Maddison et al [18], 2011; New Zealand						
BMI (Δ|B vs E)	25.6 (SD 4.1) vs 24.8 (SD 3.6)	—	25.8 (SD 4.3) vs 25.8 (SD 4.2)	—	–0.24 (95% CI –0.44 to –0.04)	.02
zBMI[e] (Δ|B vs E)	1.3 (SD 1.1) vs 1.1 (SD 1.1)	—	1.3 (SD 1.1) vs 1.3 (SD 1.0)	—	–0.06 (95% CI –0.12 to –0.03)	.03
Oh et al [19], 2022; Korea						
BMI (B vs E)	24.99 (SD 1.35) vs 23.93 (SD 1.83)	.02	25.03 (SD 1.92) vs 24.29 (SD 1.83)	.03	—	—
Staiano et al [20], 2012; United States						
Weight adjusted to growth curve (B vs E)	CO[f]: 93.93 (SD 26.02) vs 84.74 (SD 14.23); CP[g]: 96.22 (SD 17.92) vs 95.17 (SD 20.94)	—	95.48 (SD 22.72) vs 94.23 (SD 20.88)	—	CO (–1.65, SD 4.52) vs C[h] (0.86, SD 3.01), *P*=.02; CP (0.04, SD 3.46) did not differ from the others	—
Staiano et al [21], 2017; United States						
ΔzBMI	–0.002 (SE 0.02)	>.05	0.004 (SE 0.02)	>.05	–0.01 (SE 0.03)	>.05
ΔpBMI	–0.1 (SE 0.2)	>.05	0.1 (SE 0.2)	>.05	–0.2 (SE 0.3)	>.05
Staiano et al [22], 2018; United States						
ΔzBMI	–0.06 (SE 0.03)	—	0.03 (SE 0.03)	—	–0.06 (SE 0.03) vs 0.03 (SE 0.03)	.02
Δ%BMIp95[i]	–2.2 (SE 1.1)	—	0.9 (SE 1.1)	—	–2.1 (SE 1.1) vs 0.9 (SE 1.1)	.07
Chen et al [23], 2017; United States						
BMI (B vs E)	27.37 (SD 3.26) vs 26.93 (SD 3.43)	—	28.35 (SD 4.36) vs 29.18 (SD 3.88)	—	z=–4.37	.001
zBMI (B vs E)	1.60 (SD 0.50) vs 1.42 (SD 0.38)	—	1.54 (SD 0.42) vs 1.80 (SD 0.50)	—	z=–4.36	.001
Chew et al [24], 2021; Singapore						
ΔzBMI	0.045 (SD 0.15; 95% CI –0.024 to 0.114)	.19	—	—	—	—
Cueto et al [25], 2019; United States						

XSL•FO
RenderX

Study and country, and outcome	Intervention group		Control group		Intervention group vs control group	
	Outcome data	*P* value	Outcome data	*P* value	Outcome data	*P* value
Δ%BMIp95	4 weeks: –5.4 (95% CI –6.2 to –4.5), *P*<.001; 12-16 weeks: –4.8 (95% CI –5.3 to –4.3), *P*<.001; 24 weeks: –6.9 (95% CI –8.3 to –5.6), *P*<.001; Differences between age groups (*P*=.09): 5-11 years: –5.6 (SD 7.9), 12-14 years: –4.7 (SD 5.9), 15-18 years: –5.2 (SD 5.6)	—	—	—	—	—
Jelalian et al [26], 2022; United States						
ΔBMI	–0.7 (SD 2.19)	—	—	—	—	—
Δ%BMIp95	–4.45 (SD 8.5)	—	—	—	—	—
Lei et al [27], 2021; China						
ΔBMI	–3.13 (MOE[j]: 0.21)	<.001	—	—	—	—
ΔzBMI	–0.42 (MOE: 0.03)	>.001	—	—	—	—
Δ%BMIp95	–11.51 (MOE: 0.77)	>.001	—	—	—	—
Mameli et al [28], 2018; Italy						
ΔBMI-SDS[k]	–0.03 (95% CI 0.14 to 0.09)	—	–0.04 (95% CI 0.16 to 0.08)	—	0.01 (95% CI 0.15 to 0.18)	.87
Pretlow et al [29], 2015; United States						
Δ%overBMI	–0.051	<.01	—	—	—	—
Pretlow et al [30], 2020; United States						
ΔBMI	–1.6	—	—	—	—	—
ΔzBMI	–0.22	<.001	—	—	—	—
Stasinaki et al [31], 2021; Switzerland						
ΔBMI-SDS	–0.09 (range –0.4 to 0.4)	.33	–0.16 (range –1.9 to 0.3)	>.05	—	—
Vidmar et al [32], 2019; United States						
zBMI (Δ|B vs E)	–0.09 (95% CI –0.13 to –0.05)	<.001	2.39 (SD 0.34) vs 2.33 (SD 0.38)	.004	–0.02 (95% CI –0.04 to 0.01)	.32
%BMIp95 (Δ|B vs E)	127.17 (SD 21.10) vs 118.83 (SD 22.73)	<.001	136.45 (SD 22.76) vs 132.20 (SD 22.49)	.002	–2.04 (95% CI –4.16 to 0.08)	.06
Cummings et al [33], 2022; United States						
%BMIp95 (B vs E)	110.2 (SD 11.42) vs 110.46 (SD 10.33)	.80	—	—	—	—
Hinchman et al [34], 2006; United States						
ΔBMI	1.22 (SD 2.8)	<.05	1.60 (SD 2.29)	<.05	No differences between groups at 1 year	—

XSL•FO
RenderX

Study and country, and outcome	Intervention group		Control group		Intervention group vs control group	
	Outcome data	*P* value	Outcome data	*P* value	Outcome data	*P* value
ΔpBMI	0.22 (SD 1.2)	>.05	0.76 (SD 1.86)	<.05	No differences between groups at 1 year	—

[a]B: baseline.

[b]E: end point.

[c]Not reported or not applicable.

[d]pBMI: BMI percentile.

[e]zBMI: BMI z-score.

[f]CO: cooperative arm.

[g]CP: competitive arm.

[h]C: control.

[i]%BMIp95: percent over the 95th percentile.

[j]MOE: margin of error.

[k]BMI-SDS: BMI standard deviation score.

Of the 18 studies, 9 [19,20,22,23,25,27,29,30,32] reported a statistically significant decrease in BMI and 9 [17,18,21,24,26,28,31,33,34] did not. One of the studies that reported a statistically significant decrease [20] only showed this decrease in 1 of the 2 intervention arms (the cooperative arm).

Among the 9 RCTs, 4 [19,20,22,23] reported a statistically significant decrease in BMI and 5 [17,18,21,28,31] did not. The clinical trial with controls [32] reported a statistically significant decrease in BMI. Among the 3 clinical trials without controls, 2 [29,30] reported a statistically significant decrease in BMI and 1 [33] did not. Among the 4 cohort studies, 1 [25] reported a statistically significant decrease in BMI and 3 [24,26,34] did not. The observational study [27] reported a statistically significant decrease in BMI.

Of the 18 studies, 15 had a follow-up of ≤6 months, and among these, 9 [19,20,22,23,25,27,29,30,32] showed a statistically significant decrease in BMI and 6 [17,18,21,24,28,33] did not. The 3 studies that had a follow-up higher than 6 months [26,31,34] did not report a statistically significant decrease in BMI.

Among the 10 studies that used an app as an intervention, 6 [23,25,27,29,30,32] reported a statistically significant decrease in BMI and 4 [24,26,28,31] did not. Among the 6 studies that used a videogame as an intervention, 3 [19,20,22] reported a statistically significant decrease in BMI and 3 [17,18,21] did not. The other 2 studies [33,34] that used different interventions did not report a statistically significant decrease in BMI.

The only study restricted to the postmenarcheal period did not report a statistically significant decrease in BMI [21].

Regarding the analysis of extrinsic motivators, of the 2 studies that used “avatar” as a motivator, 1 [19] reported a statistically significant decrease in BMI and the other [18] did not. Of the 6 studies that used “challenges” as a motivator, 2 [22,27] reported a statistically significant decrease in BMI and 4 [28,31,33,34] did not. Of the 10 studies that used “feedback” as a motivator, 4 [19,22,25,27] reported a statistically significant decrease in BMI and 6 [17,24,26,28,31,33] did not. Of the 13 studies that used “motivation” as a motivator, 6 [22,23,25,29,30,32] reported a statistically significant decrease in BMI, 6 [21,24,26,28,31,33] did not report a decrease, and 1 [20] reported a statistically significant decrease in the cooperative arm but not in the competitive arm. Of the 5 studies that used “peer-support” as a motivator, 3 [29,30,32] reported a statistically significant decrease in BMI, 1 [34] did not report a decrease, and 1 [20] reported a statistically significant decrease in the cooperative arm (the only arm that had peer-support). The 2 studies that used “reminders” as a motivator [19,30] reported a statistically significant decrease in BMI. Of the 9 studies that used “rewards” as a motivator, 4 [19,29,30,32] reported a statistically significant decrease in BMI, 4 [21,31,33,34] did not report a decrease, and 1 [20] reported a statistically significant decrease in the cooperative arm. Among the studies, 5 used “money incentives” (3 [29,30,32] reported a decrease in BMI and 2 [33,34] did not) and 3 used “points” (2 [19,32] reported a decrease in BMI and 1 [20] reported a statistically significant decrease in the cooperative arm but not in the competitive arm). Of the 4 that used “select games” as a motivator, 1 [22] reported a statistically significant decrease in BMI and 3 [17,18,21] did not.

According to the number of extrinsic motivators applied, 1 study [23] only applied 1 motivator (motivation), and it reported a statistically significant decrease in BMI. Of the 7 studies that applied 2 extrinsic motivators, with 1 being the competitive arm in the study by Staiano et al [20], 2 studies [25,27] reported a statistically significant decrease in BMI and 5 [17,18,20,24,26] did not. Of the 6 studies that applied 3 extrinsic motivators, with 1 being the cooperative arm in the study by Staiano et al [20], 3 studies [20,29,32] reported a statistically significant decrease in BMI and 3 [21,28,34] did not. Of the 5 studies that applied 4 extrinsic motivators, 3 [19,22,30] reported a statistically significant decrease in BMI and 2 [31,33] did not.

Analysis of Studies That Compared the Impact of the Intervention With Controls

Among the 8 studies that compared the impact of the intervention with controls, 4 [20,22,23,32] reported a statistically significant decrease in BMI (3 RCTs and 1 clinical trial with

XSL•FO
RenderX

controls) and 4 [18,21,28,34] did not (3 RCTs and 1 cohort study).

Among 7 studies [18,20-23,28,32] that had a follow-up of ≤6 months, 4 reported a decrease in BMI and 3 did not, and 1 study that had a follow-up of >6 months [34] did not report a decrease in BMI.

Three studies [23,28,32] used an app as an intervention (2 reported a decrease in BMI and 1 did not), 4 [18,20-22] used a videogame (2 reported a decrease in BMI and 2 did not), and 1 [34] used Operation Zero (did not report a decrease in BMI).

One study [18] used "avatar" as a motivator (did not report a decrease in BMI), 3 [22,28,34] used "challenges" (1 reported a decrease in BMI and 2 did not), 2 [22,28] used "feedback" (1 reported a decrease in BMI and 1 did not), 6 [20-23,28,32] used "motivation" (3 reported a decrease in BMI, 2 did not report a decrease, and 1 reported a decrease only in the cooperative arm), 3 [20,32,34] used "peer-support" (2 reported a decrease in BMI and 1 did not), 3 [20,21,32] used "rewards" (1 reported a decrease in BMI, 1 did not report a decrease, and 1 reported a decrease only in the cooperative arm), 2 [21,22] used "select games" (1 reported a decrease in BMI and 1 did not).

One study [23] only applied 1 motivator (reported a decrease in BMI), 2 (with 1 being the competitive arm) [18,20] applied 2 motivators (both reported no decrease in BMI), 5 (with 1 being the cooperative arm) [20,21,28,32,34] applied 3 motivators (2 reported a decrease in BMI and 3 did not), and 1 [22] applied 4 motivators (reported a decrease in BMI).

Discussion

This review aimed to systematically assess the evidence regarding the use of extrinsic motivators for improving BMI in obese or overweight adolescents.

Half of the studies reported a statistically significant decrease in BMI, and the findings were similar in the subanalysis of studies that compared the impact of the intervention with controls. This indicates how varied the impact on BMI is in this population.

None of the studies that had a follow-up of >6 months reported a statistically significant decrease in BMI. This is in line with the literature [37,38] showing no significant effect of financial incentives on weight loss or maintenance in the long term. One hypothesis for this is that extrinsic motivators can work as a boost for initial change, but the effect decreases over time.

According to the intervention, among the studies that used an app as an intervention, 60% (6/10) reported a statistically significant decrease in BMI. The literature [39,40] has limited and mixed evidence on the impact of mobile app use on motivation and goal-setting behaviors and obesity-related outcomes. Regarding the use of a videogame as an intervention, only half reported a statistically significant decrease in BMI. This is in line with another systematic review [41] that found that using only videogames for weight management does not deliver satisfying results.

Regarding extrinsic motivators, among the studies that used "select games," only 25% (1/4) reported a statistically significant decrease in BMI. To our knowledge, there are no other studies that analyzed this aspect. One hypothesis for the low impact of the extrinsic motivator "select games" is that in most interventions, the games were similar, and thus, adolescents did not feel that they had much control and that it was really their choice. As for "challenges," only 33% (2/6) of studies reported a statistically significant decrease in BMI. One hypothesis for this is that most of the "challenges" were imposed and not self-chosen, particularly in studies that used an app as the intervention, which could negatively influence the motivation for completion.

On the other hand, regarding the use of "rewards" as a motivator, among studies that used "money incentives," 60% (3/5) reported a statistically significant decrease in BMI. This is in line with the literature [38] showing that financial incentives produce significant weight loss during the intervention, but weight is usually regained following cessation of the incentive [37,38]. Moreover, among studies that used "points," 75% (3/4) reported a statistically significant decrease in BMI. The mechanism of influence appears to be similar to that for "money incentives" as both work as rewards [42]. Among studies that used "peer-support," 80% (4/5) reported a statistically significant decrease in BMI. On analyzing the study by Staiano et al [20], we can see that this extrinsic motivator has a major impact on the outcome because only the intervention arm with it (cooperative arm) showed a statistically significant difference compared with the control group. This is in line with a meta-analysis [43] showing that peer-support appears to be associated with decreased weight and BMI levels in individuals with overweight and obesity. One hypothesis is that peers can personalize weight control interventions for individuals with overweight or obesity in a way that medical professionals often fail to achieve. With regard to "reminders," all studies showed a statistically significant decrease in BMI. One hypothesis is that these "reminders" can help adolescents to not forget about their activities and goals.

The number of extrinsic motivators used was not related to the outcome. This may be due to the different definitions applied to each extrinsic motivator between different studies. For example, "motivation" could involve coaches, messages, sessions, or a mixture of these, and thus, the impact could be different.

Among studies where intrinsic motivation was already present, all showed a statistically significant decrease in BMI. Comparing academic performance [44], students with primarily intrinsic motivators had better academic performance than those with primarily extrinsic motivators. Extrinsic motivation can become an essential strategy for only less interesting or enjoyable tasks [44,45]. Therefore, having intrinsic motivation could be more important than having extrinsic motivation for achieving the goal, and extrinsic motivation can be a boost for more boring tasks.

The trends observed in the subanalysis of studies that evaluated the impact of interventions with controls were overlapping with those observed in the general analysis in most cases.

XSL•FO
RenderX

The studies included had many limitations, and the heterogeneities in study design, sample size, follow-up duration, outcomes reported (BMI, BMI z-score, BMI percentile, percent over the 95th percentile, and BMI SD score), outcome analysis (some reported baseline and end point values and others reported variations in outcomes), and extrinsic motivators used make the comparisons and result analyses difficult. Furthermore, studies did not evaluate extrinsic motivators in isolation and did not report the stage of pubertal development among participants; therefore, each study had numerous confounding factors. Finally, most of the studies had a moderate or high risk of bias.

As a limitation of the review process, we could not obtain 3 studies, although we contacted the authors. Therefore, these studies had to be excluded. We also excluded 1 study owing to language limitations (Korean). We could not perform a subanalysis by age or sex as we did not have access to individual data. We only searched 2 databases, and thus, there may be some publication bias and some relevant grey literature may have been missed. Finally, new studies may have been published after April 2023.

Future studies, including RCTs, should be well designed to minimize the heterogeneity of BMI measures and extrinsic motivator definitions. They should also focus on the impact of specific extrinsic motivators on BMI reduction. Studies having a longer duration are also needed to better understand the long-term impact of extrinsic motivators on weight management success.

In conclusion, the extrinsic motivators “reminders” and “peer-support” appear to have high impacts on BMI reduction. Further studies involving specific extrinsic motivator definitions are needed to closely assess their impacts on BMI reduction, including their long-term impacts. Moreover, future studies should evaluate the impacts of extrinsic motivators on BMI according to other important factors such as the level of sexual development.

Acknowledgments

This study was supported by National Funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., within CINTESIS, R&D Unit (reference UIDB/4255/2020). This study is the result of a master’s degree work [46].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 21 KB - jmir_v26i1e57458_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 55 KB - jmir_v26i1e57458_app2.pdf]

References

1. Ebbeling CB, Pawlak DB, Ludwig DS. Childhood obesity: public-health crisis, common sense cure. The Lancet 2002 Aug;360(9331):473-482. [doi: 10.1016/s0140-6736(02)09678-2]
2. Obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2023-08-27]
3. Explore childhood obesity related data by demographic (ages 10-17). State of Childhood Obesity. URL: https://stateofchildhoodobesity.org/demographic-data/ages-10-17/ [accessed 2023-08-27]
4. Healthy Weight Among Children. UNICEF. URL: https://www.unicef.cn/en/reports/healthy-weight-among-children [accessed 2023-08-22]
5. Health at a Glance: Europe. OECD iLibrary. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2020_82129230-en [accessed 2024-12-04]
6. Frade F, Carteiro D, Pereira F, Marques J, Frade J. Prevalence of childhood obesity in Portugal: a narrative review of the literature. Port J Public Health 2020 Dec 3;38(2):119-128. [doi: 10.1159/000511792]
7. Borrello M, Pietrabissa G, Ceccarini M, Manzoni GM, Castelnuovo G. Motivational interviewing in childhood obesity treatment. Front Psychol 2015 Nov 12;6:1732 [FREE Full text] [doi: 10.3389/fpsyg.2015.01732] [Medline: 26617555]
8. The GBD 2015 Obesity Collaborators. Health effects of overweight and obesity in 195 countries over 25 years. N Engl J Med 2017 Jul 06;377(1):13-27. [doi: 10.1056/nejmoa1614362]
9. Schiel R, Kaps A, Bieber G. Electronic health technology for the assessment of physical activity and eating habits in children and adolescents with overweight and obesity IDA. Appetite 2012 Apr;58(2):432-437. [doi: 10.1016/j.appet.2011.11.021] [Medline: 22155072]

XSL•FO
RenderX

10. Tsiros MD, Sinn N, Coates AM, Howe PRC, Buckley JD. Treatment of adolescent overweight and obesity. Eur J Pediatr 2008 Jan 1;167(1):9-16. [doi: 10.1007/s00431-007-0575-z] [Medline: 17973118]
11. Woo S, Park KH. Motivating children and adolescents in obesity treatment. J Obes Metab Syndr 2020 Dec 30;29(4):260-269 [FREE Full text] [doi: 10.7570/jomes20026] [Medline: 32843587]
12. Yoshida-Montezuma Y, Ahmed M, Ezezika O. Does gamification improve fruit and vegetable intake in adolescents? a systematic review. Nutr Health 2020 Dec 23;26(4):347-366. [doi: 10.1177/0260106020936143] [Medline: 32703067]
13. Sardi L, Idri A, Fernández-Alemán J. A systematic review of gamification in e-Health. J Biomed Inform 2017 Jul;71:31-48 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
14. Byrne S, Gay G, Pollack JP, Gonzales A, Retelny D, Lee T, et al. Caring for mobile phone-based virtual pets can influence youth eating behaviors. Journal of Children and Media 2012 Feb;6(1):83-99. [doi: 10.1080/17482798.2011.633410]
15. Andrade A, Correia CK, Coimbra DR. The psychological effects of exergames for children and adolescents with obesity: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2019 Nov 01;22(11):724-735. [doi: 10.1089/cyber.2019.0341] [Medline: 31697604]
16. Adolescents et jeunes adultes. Santecom. URL: https://extranet.santecom.qc.ca/wiki/!biblio3s/doku.php?id=concepts:adolescents-et-jeunes-adultes [accessed 2024-12-04]
17. Adamo KB, Rutherford JA, Goldfield GS. Effects of interactive video game cycling on overweight and obese adolescent health. Appl. Physiol. Nutr. Metab 2010 Dec;35(6):805-815. [doi: 10.1139/h10-078]
18. Maddison R, Foley L, Ni Mhurchu C, Jiang Y, Jull A, Prapavessis H, et al. Effects of active video games on body composition: a randomized controlled trial. Am J Clin Nutr 2011 Jul;94(1):156-163 [FREE Full text] [doi: 10.3945/ajcn.110.009142] [Medline: 21562081]
19. Oh W, An Y, Min S, Park C. Comparative effectiveness of artificial intelligence-based interactive home exercise applications in adolescents with obesity. Sensors (Basel) 2022 Sep 28;22(19):7352 [FREE Full text] [doi: 10.3390/s22197352] [Medline: 36236451]
20. Staiano AE, Abraham AA, Calvert SL. Adolescent exergame play for weight loss and psychosocial improvement: a controlled physical activity intervention. Obesity (Silver Spring) 2013 Mar;21(3):598-601 [FREE Full text] [doi: 10.1002/oby.20282] [Medline: 23592669]
21. Staiano AE, Marker AM, Beyl RA, Hsia DS, Katzmarzyk PT, Newton RL. A randomized controlled trial of dance exergaming for exercise training in overweight and obese adolescent girls. Pediatr Obes 2017 Apr 26;12(2):120-128 [FREE Full text] [doi: 10.1111/ijpo.12117] [Medline: 26918815]
22. Staiano AE, Beyl RA, Guan W, Hendrick CA, Hsia DS, Newton RL. Home-based exergaming among children with overweight and obesity: a randomized clinical trial. Pediatr Obes 2018 Nov 20;13(11):724-733 [FREE Full text] [doi: 10.1111/ijpo.12438] [Medline: 30027607]
23. Chen J, Guedes CM, Cooper BA, Lung AE. Short-term efficacy of an innovative mobile phone technology-based intervention for weight management for overweight and obese adolescents: pilot study. Interact J Med Res 2017 Aug 02;6(2):e12 [FREE Full text] [doi: 10.2196/ijmr.7860] [Medline: 28768612]
24. Chew CSE, Davis C, Lim JKE, Lim CMM, Tan YZH, Oh JY, et al. Use of a mobile lifestyle intervention app as an early intervention for adolescents with obesity: single-cohort study. J Med Internet Res 2021 Sep 28;23(9):e20520 [FREE Full text] [doi: 10.2196/20520] [Medline: 34581672]
25. Cueto V, Wang CJ, Sanders LM. Impact of a mobile app-based health coaching and behavior change program on participant engagement and weight status of overweight and obese children: retrospective cohort study. JMIR Mhealth Uhealth 2019 Nov 15;7(11):e14458 [FREE Full text] [doi: 10.2196/14458] [Medline: 31730041]
26. Jelalian E, Darling K, Foster GD, Runyan T, Cardel MI. Effectiveness of a scalable mHealth intervention for children with overweight and obesity. Child Obes 2023 Dec 01;19(8):552-559 [FREE Full text] [doi: 10.1089/chi.2022.0154] [Medline: 36576892]
27. Lei S, Inojosa JRM, Kumar S, Lee AT, Scott CG, Lerman A, et al. Effectiveness of a weight loss program using digital health in adolescents and preadolescents. Child Obes 2021 Jul 01;17(5):311-321 [FREE Full text] [doi: 10.1089/chi.2020.0317] [Medline: 33826417]
28. Mameli C, Brunetti D, Colombo V, Bedogni G, Schneider L, Penagini F, et al. Combined use of a wristband and a smartphone to reduce body weight in obese children: randomized controlled trial. Pediatr Obes 2018 Feb 29;13(2):81-87. [doi: 10.1111/ijpo.12201] [Medline: 27900849]
29. Pretlow RA, Stock CM, Allison S, Roeger L. Treatment of child/adolescent obesity using the addiction model: a smartphone app pilot study. Child Obes 2015 Jun;11(3):248-259 [FREE Full text] [doi: 10.1089/chi.2014.0124] [Medline: 25760813]
30. Pretlow RA, Stock CM, Roeger L, Allison S. Treatment of the sensory and motor components of urges to eat (eating addiction?): a mobile-health pilot study for obesity in young people. Eat Weight Disord 2020 Dec 14;25(6):1779-1787 [FREE Full text] [doi: 10.1007/s40519-019-00836-z] [Medline: 31939105]
31. Stasinaki A, Büchter D, Shih CI, Heldt K, Güsewell S, Brogle B, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 09;21(1):308 [FREE Full text] [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]

XSL•FO
RenderX

32. Vidmar AP, Pretlow R, Borzutzky C, Wee CP, Fox DS, Fink C, et al. An addiction model-based mobile health weight loss intervention in adolescents with obesity. Pediatr Obes 2019 Feb 16;14(2):e12464 [FREE Full text] [doi: 10.1111/ijpo.12464] [Medline: 30117309]
33. Cummings C, Crochiere R, Lansing AH, Patel R, Stanger C. A digital health program targeting physical activity among adolescents with overweight or obesity: open trial. JMIR Pediatr Parent 2022 Mar 28;5(1):e32420 [FREE Full text] [doi: 10.2196/32420] [Medline: 35343903]
34. Hinchman J, Beno L, Mims A. Kaiser Permanente Georgia's experience with operation zero: a group medical appointment to address pediatric overweight. TPJ 2006 Dec;10(3):66-71. [doi: 10.7812/tpp/05-126]
35. Maddison R, Mhurchu C, Jull A, Prapavessis H, Foley LS, Jiang Y. Active video games: the mediating effect of aerobic fitness on body composition. Int J Behav Nutr Phys Act 2012 May 03;9(1):54 [FREE Full text] [doi: 10.1186/1479-5868-9-54] [Medline: 22554052]
36. Foley L, Jiang Y, Ni Mhurchu C, Jull A, Prapavessis H, Rodgers A, et al. The effect of active video games by ethnicity, sex and fitness: subgroup analysis from a randomised controlled trial. Int J Behav Nutr Phys Act 2014 Apr 03;11(1):46 [FREE Full text] [doi: 10.1186/1479-5868-11-46] [Medline: 24694082]
37. Paul‐Ebhohimhen V, Avenell A. Systematic review of the use of financial incentives in treatments for obesity and overweight. Obesity Reviews 2007 Oct 23;9(4):355-367. [doi: 10.1111/j.1467-789x.2007.00409.x]
38. John LK, Loewenstein G, Troxel AB, Norton L, Fassbender JE, Volpp KG. Financial incentives for extended weight loss: a randomized, controlled trial. J Gen Intern Med 2011 Jun 20;26(6):621-626 [FREE Full text] [doi: 10.1007/s11606-010-1628-y] [Medline: 21249462]
39. Quelly SB, Norris AE, DiPietro JL. Impact of mobile apps to combat obesity in children and adolescents: A systematic literature review. J Spec Pediatr Nurs 2016 Jan 22;21(1):5-17. [doi: 10.1111/jspn.12134] [Medline: 26494019]
40. Lee J, Piao M, Byun A, Kim J. A systematic review and meta-analysis of intervention for pediatric obesity using mobile technology. Stud Health Technol Inform 2016;225:491-494. [Medline: 27332249]
41. Mack I, Bayer C, Schäffeler N, Reiband N, Brölz E, Zurstiege G, et al. Chances and limitations of video games in the fight against childhood obesity-a systematic review. Eur Eat Disord Rev 2017 Jul 03;25(4):237-267. [doi: 10.1002/erv.2514] [Medline: 28467004]
42. Peña S, Carranza M, Cuadrado C, Parra DC, Villalobos Dintrans P, Castillo C, et al. Effectiveness of a gamification strategy to prevent childhood obesity in schools: a cluster controlled trial. Obesity (Silver Spring) 2021 Nov 17;29(11):1825-1834 [FREE Full text] [doi: 10.1002/oby.23165] [Medline: 34533295]
43. Chen Y, Li Z, Yang Q, Yang S, Dou C, Zhang T, et al. The effect of peer support on individuals with overweight and obesity: a meta-analysis. Iran J Public Health 2021 Dec 06;50(12):2439-2450 [FREE Full text] [doi: 10.18502/ijph.v50i12.7926] [Medline: 36317042]
44. Stelling AG, Mastenbroek NJ, Kremer WD. Predictive value of three different selection methods for admission of motivated and well-performing veterinary medical students. J Vet Med Educ 2019 Aug;46(3):289-301. [doi: 10.3138/jvme.0417-050r1] [Medline: 30565980]
45. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
46. Gonçalves A. Use of extrinsic motivators to improve the BMI of obese or overweight adolescents. Open Repository of the University of Porto. URL: https://repositorio-aberto.up.pt/bitstream/10216/156041/2/653138.pdf [accessed 2024-12-04]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by A Mavragani; submitted 17.02.24; peer-reviewed by U Ellis, A Chwałczyńska; comments to author 09.04.24; revised version received 26.05.24; accepted 14.10.24; published 30.12.24.

Please cite as:
Gonçalves A, Simões PA, Sousa-Pinto B, Taveira-Gomes T
Use of Extrinsic Motivators to Improve the BMI of Obese or Overweight Adolescents: Systematic Review
J Med Internet Res 2024;26:e57458
URL: https://www.jmir.org/2024/1/e57458
doi:10.2196/57458
PMID:39576963

XSL•FO
RenderX

©Ana Gonçalves, Pedro Augusto Simões, Bernardo Sousa-Pinto, Tiago Taveira-Gomes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effectiveness of Video Animations as a Tool to Improve Health Information Recall for Patients: Systematic Review

Steffen Hansen[1], MPH; Tue Secher Jensen[2,3,4], MSc, PhD; Anne Mette Schmidt[5], PhD; Janni Strøm[6], RN, PhD; Peter Vistisen[7], PhD; Mette Terp Høybye[1,8], MSc, PhD

[1]University Clinic for Interdisciplinary Orthopedic Pathways (UCOP), Elective Surgery Center, Silkeborg Regional Hospital, Silkeborg, Denmark
[2]Diagnostic Imaging, Silkeborg Regional Hospital, Silkeborg, Denmark
[3]Department of Sport Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark
[4]University Clinic for Innovative Patient Pathways, Medical Diagnostic Centre, Silkeborg Regional Hospital, Silkeborg, Denmark
[5]University Research Clinic for Innovative Patient Pathways, Medical Diagnostic Centre, Silkeborg Regional Hospital, Silkeborg, Denmark
[6]Quality Section, Capio Private Hospitals, Viborg, Denmark
[7]Department of Communication & Psychology, Aalborg University, Aalborg, Denmark
[8]Interacting Minds Centre, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

Corresponding Author:
Mette Terp Høybye, MSc, PhD
University Clinic for Interdisciplinary Orthopedic Pathways (UCOP)
Elective Surgery Center
Silkeborg Regional Hospital
Falkevej 1-3
Silkeborg, 8600
Denmark
Phone: 45 61106066
Email: hoybye@clin.au.dk

Abstract

Background: Access to clear and comprehensible health information is crucial for patient empowerment, leading to improved self-care, adherence to treatment plans, and overall health outcomes. Traditional methods of information delivery, such as written documents and oral communication, often result in poor memorization and comprehension. Recent innovations, such as animation videos, have shown promise in enhancing patient understanding, but comprehensive investigations into their effectiveness across various health care settings are lacking.

Objective: This systematic review aims to investigate the effectiveness of animation videos on health information recall in adult patients across diverse health care sectors, comparing their impact to usual information delivery methods on short-term and long-term recall of health information.

Methods: We conducted systematic searches in PubMed, CINAHL, and Embase databases, supplemented by manual searches of reference lists. Included studies were randomized controlled trials involving adult participants (≥18 years) that focused on the use of animation videos to provide health information measured against usual information delivery practice. There were no language restrictions. Out of 2 independent reviewers screened studies, extracted data, and assessed the risk of bias using the Revised Cochrane risk-of-bias tool for randomized trials (RoB2), Covidence was used to handle screening and risk of bias process. A narrative synthesis approach was applied to present results.

Results: A total of 15 randomized controlled trials—3 in the United States, 2 in France, 2 in Australia, 2 in Canada, and 1 in the United Kingdom, Japan, Singapore, Brazil, Austria, and Türkiye, respectively—met the inclusion criteria, encompassing 2,454 patients across various health care settings. The majority of studies (11/15, 73%) reported statistically significant improvements in health information recall when animation videos were used, compared with usual care. Animation videos ranged from 1 to 15 minutes in duration with the most common length ranging from 1 to 8 minutes (10/15) and used various styles including 2D cartoons, 3D computers, and whiteboard animations. Most studies (12/15) assessed information recall immediately after intervention, with only 3 studies including longer follow-up periods. Most studies exhibited some concerns related to the risk of bias, particularly in domains related to deviations from intended interventions and selection of reported results.

XSL•FO
RenderX

Conclusions: Animation videos appear to significantly improve short-term recall of health information among adult patients across various health care settings compared with usual care. This suggests that animation videos could be a valuable tool for informing patients in different health care settings. However, further research is needed to explore the long-term efficacy of these interventions, their impact on diverse populations, and how different animation styles might affect information recall. Future studies should also address methodological limitations identified in current research, including the use of validated outcome measures and longer follow-up periods.

Trial Registration: PROSPERO CRD42022380016; http://crd.york.ac.uk/prospero/display_record.php?RecordID=380016

(J Med Internet Res 2024;26:e58306) doi:10.2196/58306

KEYWORDS

public health; health information; patient information; animation video; digital health; visualization; memory; recall; education; synthesis; review methods; review methodology; systematic; PRISMA

Introduction

Accessible, accurate, and understandable health information is crucial in health care, empowering patients to make informed decisions, take appropriate actions, and actively participate in their own health management [1-3]. It enables patients to understand their conditions, treatments, and health behaviors, leading to improved self-care, better adherence to treatment plans, and overall better health outcomes [4-6].

Studies have shown that lacking or inaccurate knowledge of health conditions or treatment reduces the ability of the patient to engage in treatment and preventive management behaviors [7-9] with lower compliance leading to poorer health outcomes, such as physical disabilities and decreased mental health [4,10].

Information about illness and relevant treatments is traditionally provided through written documents, for example, information pamphlets handed out by a health care professional, or oral information given during a consultation. However, these types of information often lead to poor memorization and comprehension, resulting in potential mistrust and confusion. Furthermore, these types of information were strongly correlated with level of education and health literacy levels [11,12].

Over the past decade, there has been a growing interest in using innovative tools to improve patient information, such as animation videos, to enhance patients' knowledge and skills in managing their illnesses. Combining verbal and pictorial information in a coordinated way has been shown to enhance learning outcomes, for individuals with high and low health literacy [13-15].

While previous research has delved into the use of animated videos and shown promising results in selected health areas [15], a notable gap exists in the comprehensive examination of animations' precise influence as a tool for improving the recall of health-related information in general. More knowledge is particularly needed when exploring more technical parameters when using animation videos [16,17]. Factors such as the type of animation video used, the specific health topics addressed, the duration of the animation's impact, and other relevant variables need to be further investigated.

Therefore, the objective of this systematic review is to investigate the effectiveness of animation videos on health information recall in patients as they engage with the health care system across diverse health care sectors.

Methods

Protocol and Registration

The review protocol was registered at PROSPERO (International Prospective Register of Systematic Reviews; ID: CRD42022380016; November 2022). PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed when reporting the search results [18] (Multimedia Appendix 1). Our original protocol specified the inclusion of studies investigating only patients with musculoskeletal disorders. This criterion was later broadened to include studies examining all types of disorders.

Eligibility Criteria

Participants

This systematic review included studies involving adult participants aged 18 years and older in a health care setting. No restrictions were imposed based on ethnicity, gender, or socioeconomic status. However, studies on patients with personality disorders or conditions known to affect cognitive functionality were excluded.

Intervention

Animation videos were only considered if they were objective and factual and provided information about a patient's specific health condition, such as the course of treatment or physiological and anatomical facts. In addition, the animation videos should enhance patients' understanding of their health condition and how to manage it. The delivery mode of interventions should involve either the use of CD-ROMs, websites, tablet devices, computers, or mobile phones within a health care context. The animation was eligible if it consisted of, for example, cartoons, avatars, "whiteboard" animation, or animated 2D or 3D models. The intervention was only eligible if sufficient details about the animation video were provided, either by providing detailed descriptions, screenshots, or links; authors were contacted for more information if details were scarce. We used the exclusion definition of insufficient details about the intervention when it was unclear if the animation video component was the main component in the intervention or just a bi-part of the intervention as a whole. Furthermore, we used the definition could not

XSL•FO
RenderX

contact the author if we were uncertain if the studies included what we defined as animation or not, and did not get any response when trying to contact authors to provide sufficient details about the animation component. All exclusion reasoning is displayed in Figure 1.

Comparator

Usual care encompasses conventional information delivered by health care professionals, such as the use of written printed materials or verbal communication.

Outcome

Health information recall refers to the participant's ability to remember and recall the information provided by the animation video immediately after receiving it or during a follow-up period. Therefore, all studies that aimed to measure information recall, knowledge gain, or acquisition of information about the specific disorders as a primary or secondary outcome were included.

Types of Studies

This review included randomized controlled trials (RCTs) that were peer-reviewed. No specific criteria were set for the language used in the studies. If necessary, a translation tool (DeepL) was used to review and assess a study. The time frame was not restricted, as it was anticipated that research related to animation videos would primarily be limited to the last few decades.

Search Strategy

A 3-step search was performed with the assistance of an experienced research librarian. First, an initial search was conducted and was ongoing from July 2022 to August 2022. This search was limited to PubMed, CINAHL, and Embase using preliminary subject headings and keywords based on experience and knowledge of the field. Articles that appeared appropriate to the aim were viewed. Notes were made of relevant keywords contained in the title, abstract, and index terms in each relevant article.

The preliminary subject headings and keywords were revised in accordance with the findings obtained in the initial search. Articles found relevant in the first search were set aside to confirm that the second search identified these.

The second search was performed between September 12, 2022, and September 16, 2022, using the revised subject headings and keywords in PubMed, CINAHL, Embase, and Web of Science. The search was divided into blocks consisting of main keywords and additional variables and is shown in Multimedia Appendix 2. The third and final search was performed under and after the screening process from January 24, 2023, to September 18, 2023, and was conducted using reference lists. Reference lists were manually consulted to identify any additional studies, and finally, a crosscheck was done to confirm that previously identified articles were included in the search.

Study Selection

A 3-step selection and assessment process were conducted. First, all studies were imported into Covidence, a web-based collaboration software for managing systematic reviews and also to remove duplicates [19].

Second, all titles and abstracts were independently screened twice by 2 reviewers (SH, TSJ, AMS, and MTH). Articles identified for full-text underwent a similar process, being independently screened twice by the 2 reviewers (SH, TSJ, AMS, and MTH). Discrepancies between the reviewers' decisions in the inclusion and exclusion process were clarified by involving a third reviewer who had not been part of the initial review (SH, TSJ, AMS, and MTH; Figure 1). Finally, all studies reviewed in full-text that did not meet the inclusion criteria were excluded.

Data Collection

Characteristics and outcomes of the studies were extracted using a data collection form developed in Covidence [19]. This was done by 2 separate reviewers (SH, TSJ, AMS, and MTH), and in case of conflicting understandings, the subject was resolved by discussion or handled by the third reviewer. Data were extracted on author and date, setting (health care sector) and country, sample size, mean age of participants, intervention and comparator, follow-up period, and outcome, including type of measurement instrument.

Study Quality Assessment

The studies that met the eligibility criteria were assessed for their methodological validity using the Revised Cochrane risk-of-bias tool for randomized trials (RoB2) template [20]. A total of 2 independent reviewers (SH, TSJ, AMS, and MTH) appraised the studies, and in cases of disagreement, a third reviewer was consulted.

Data Analysis

Due to the diverse nature, use of measurement instruments, and the heterogeneity of articles in this field, it was not planned to perform a meta-analysis. Therefore, a narrative synthesis approach was planned. Detailed information on the study characteristics, and methodological quality, as well as a summary of the outcome measures and the results, was obtained. Also, the narrative synthesis summarized the findings according to three pre-identified outcome-related categories (length of the animation videos and time for outcome assessment; animation styles and health-related topics; and comparators and the role of the animation video) inspired by the data extraction template made for this study. We have used $P<.05$ as an indicator of effect.

Results

Study Selection

A total of 2505 studies were identified, yielding 79% (1985/2505) studies after removal of duplicates. After reviewing the 1985 titles and abstracts, 1825 studies were excluded, resulting in a total of 160 (8.1%) eligible for full-text assessment. After full-text assessment, 145 additional studies were excluded, resulting in a total of 15 included studies (Figure 1). All inconsistencies in data assessment and extraction were resolved by consensus among the reviewers. A total of 2454

XSL•FO
RenderX

patients were included, and details of the study selection are displayed in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

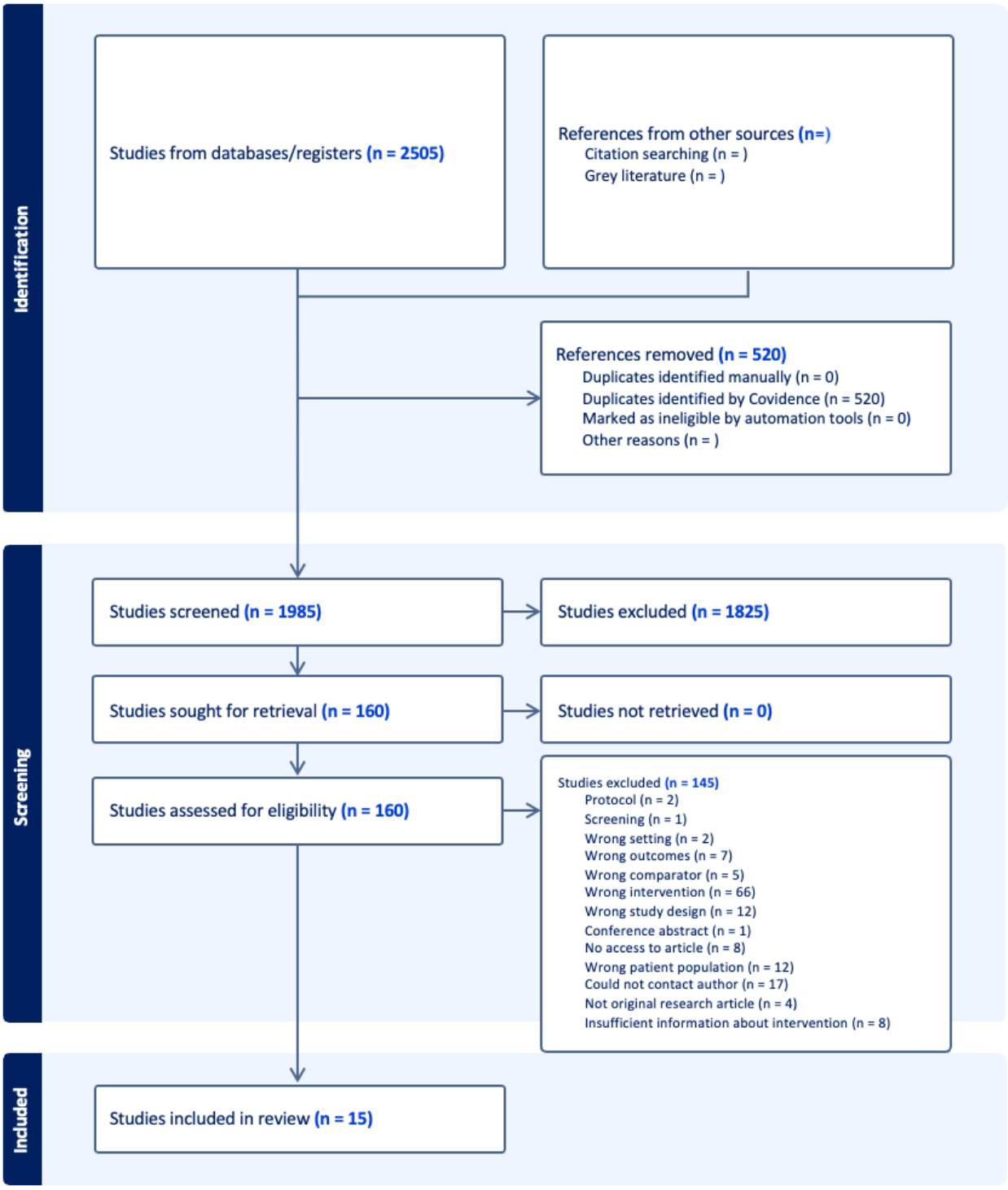

Study Characteristics

Detailed characteristics and results of the included studies are shown in Multimedia Appendix 3. All studies were conducted in the last 10 years, with most studies 75% (10/15) published within the past 5 years [21-30]. Of the 15 studies included the majority 12 (80%) were conducted in high-income countries [21,22,25-27,29-35], and 3 (20%) were conducted in upper- to middle-income countries [23,24,28]. Most studies originated from the United States (20%) [21,26,29], Australia (13.3%) [32,33], Canada (13.3%) [27,34], and France (13.3%) [22,30], and one (6.7%) each from Japan [31], Austria [35], United Kingdom [25], Singapore [24], Brazil [23], and Türkiye [28].

Length of the Animation Videos and Time for Outcome Assessment

The animation videos exhibited a diverse range of durations, spanning from 1 to 15 minutes. Dividing the length into

XSL•FO
RenderX

percentiles revealed distinct patterns among the studies. A third (5/15) of the studies used short animation videos lasting 1 to 5 minutes [24,26,30,31,34]. An equal proportion featured middle-length animation videos ranging from 6 to 8 minutes [21,23,25,27,29]. In total, 3 studies used longer animation videos, falling within the 13- to 15-minute range [22,32,33]. In addition, 2/15 had animation videos with unspecified durations [28,35]. Most of the studies (12/15) assessed the outcome immediately after the patients were given the intervention [22-24,26,27,29-35] or on the day of treatment [25]. Only 3 of 15 studies had follow-up, one after 1 month [28], one after 6 weeks [32], and one after 6 and 8 weeks [21].

Animation Styles and Health-Related Topics

The studies used a variety of animation styles on different health-related topics. 2D cartoon animated videos were used in 7 studies, on the subjects of diabetes food care [28]; imaging and inevitable consequences of lower back pain [23]; living with atrial fibrillation [27]; anesthesia during surgery [31]; the etiology, symptoms, and treatments for stress and urgency urinary incontinence [21]; and preoperative retention undergoing bowel resection [33] and postoperative risk with benign parotid tumor [22]. 3D computer animation was used in 4 studies, with topics regarding; informed consent before operative laparoscopy [32]; informed content before elective inpatient coronary angiography [30]; knowledge regarding Mohs surgery [25]; and age-related cataract surgery [35]. A total of 3 studies used narrated whiteboard animation the subjects presented by this form of animation were opioid safety and proper usage, storage, and disposal [29]; first-time intravenous fluorescein angiography [34]; and patients undergoing coronary angiography [24]. Finally, one study used a mixture of 2D and 3D animation on the subject of Mohs surgery [26].

Comparators and the Role of the Animation Video

A variety of comparators, commonly referred to as “usual care,” were used across the studies. Notably, the majority of the studies (12/15) evaluated animation videos as a supplementary tool alongside usual care. Among these 12 studies, 6 used usual care that combined written materials with verbal information provided by health care professionals [22,24,25,29,30,35]. Out of 4 other studies within this subset described usual care as either a routine verbal consent process [26,32] or individual face-to-face conversations [27,31]. A more flexible approach to usual care was noted in 1 study, which incorporated a mix of verbal explanations, drawings, or picture imagery [21]. The remaining study relied solely on written information sheets as usual care [33]. In contrast, the last 3 studies out of the 15 had a different approach, examining animation videos as an alternative, rather than a supplement, to their usual care. In total, 2 of these studies used a combination of verbal and written information as their usual care [28,34] while the last study only used written materials [23].

Outcomes

Health Information Recall

The 15 studies used different outcome measurements to assess health information recall. All the studies used self-developed questionnaires, which were based on the information provided during the consultation. The knowledge measurement tools included multiple-choice answers, open-ended questions, and statement questions with true or false and yes or no options. The studies included in the review used varying numbers of questions related to the specific medical subject, with the number of questions ranging from 4 to 16.

Positive Effects Using Animation Videos

In total, 11 studies reported statistically significant improvements in health information recall when compared with control groups [22,24,26-32,34,35].

Notably when looking at animation style, the 3 studies that used whiteboard animation as an educational tool demonstrated a significant positive effect over usual care [24,29,34]. A total of 2 out of the 3 studies using 3D animation videos had a positive effect [32,35], and 5 of the total 8 studies using 2D cartoon animation [22,27,28,31] including the study using a mixture of 2D and 3D [26] had a positive effect.

When looking at the time of outcome assessment 10 out of the 11 studies that found effect was measured immediately after intervention [22,24,26,27,29-32,34,35]. One of those studies did not have a persistent effect at the 6-week follow-up assessment [32]. Conversely, the study conducted by Dincer and Bahçecik [28] in 2021 observed a positive effect 4 weeks post intervention [28].

Equal Effect of Using Animation Video

In total, 3 out of the 15 studies found equal benefits between the use of animation video and usual care [21,25,33]. Two of the studies used 2D cartoon animation [21,33] and 1 used a 3D animation video [25]. In the cases of these 3 studies, the usual care alone achieved the same positive effect as supplementing usual care with the animation video. The studies’ usual care consisted of leaflets and verbal information [25], a combination of verbal communication, drawing or pictorial imagery, using a more open and flexible information style [21], and finally a simple information sheet [33].

Negative Effect of Using Animation Video

In 1 study [23], it was observed that usual care achieved significantly greater benefits than the use of animation videos when assessing the accuracy of patients’ responses regarding the correct use of imaging. Notably, the animation video was introduced as an alternative to conventional consent information. The comparators to animation were in this case (1) a 2-page written summary on an A4 paper and (2) an infographic solution combining both imagery and written summary, both these solutions had equal positive effects.

Study Quality Assessment

An overview of the risk of bias assessment for the included studies is presented in Figure 2 [21-35]. In general, most studies exhibited “some concerns” related to the risk of bias, with the most prominent issues being observed in domain D2: Bias due to deviations from intended interventions. The primary reason for this was our inability to access trial protocols or get a sufficient description of the protocol, which would have provided crucial insights into the trial context. Without access or information in the protocols, it was difficult to determine if

XSL•FO
RenderX

the intended interventions were consistently followed, which could have potentially introduced bias into the results. In some cases, there was a lack of transparency regarding blinding procedures, further contributing to concerns in this domain.

Moreover, the domain "D5: Bias in selection of the reported results," also raised some concerns. These issues were partly due to the unavailability of information on study protocols, making it difficult to determine if the reported results align with a prespecified analysis plan.

To a lesser extent, some concerns were identified in the domain "D4: Risk of bias in the measurement of the outcome." Many studies of interventions in this area face a common challenge of using self-developed outcome measurement tools instead of validated instruments. This is because the outcomes being measured are often highly specific and directly related to the particular disorder targeted by the animation. However, using nonvalidated tools raises questions about the reliability and validity of the outcome measurements.

Figure 2. Risk of bias of included studies in this review of the effectiveness of animation as a tool for recall of health information, Denmark 2022-2024 [21-35].

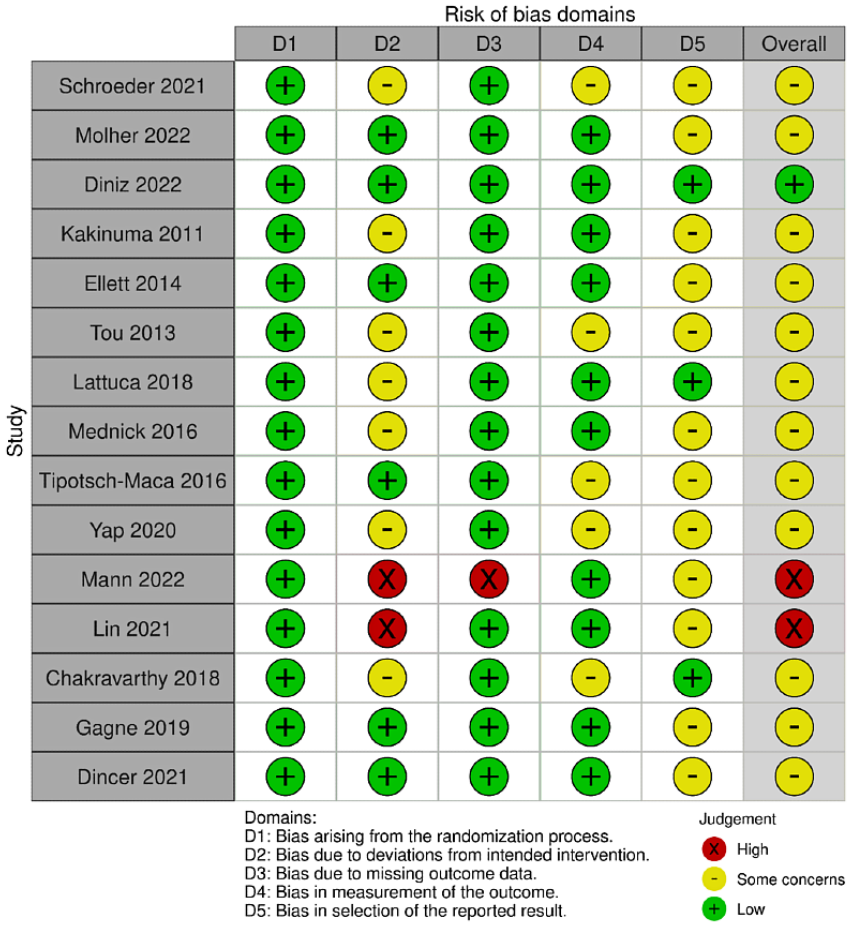

Discussion

Principal Findings

The primary objective of this systematic review was to investigate the effectiveness of animation videos as informative tools in the health care system, particularly in enhancing patient health information recall. The main finding is that 11 of the 15 studies reported a significant positive impact of animation videos on health information recall. The studies predominantly originated from high-income countries and were mostly conducted within the last decade, with a significant concentration (75%) published in the past 5 years. The studies demonstrated a diversity in the length and style of animation videos, ranging from short (1-5 minutes) to longer durations (up to 15 minutes). Different animation styles were used, including 2D cartoon animation, 3D computer animation, and narrated whiteboard animation. This variety underscores the adaptable nature of animation as a medium for conveying health information.

Comparison to Previous Work

Similar results are found in a recent systematic review investigating the effectiveness of animations as an information medium for patients and the general public; they found that across 30 trials assessing knowledge, 19 of those had greater knowledge compared with a range of comparators. However, the evidence base was highly variable, with mostly small trials [36]. The same tendencies are shown in a systematic review and meta-analysis by Feeley et al [37], which found an overall weighted effect size of 0.38, which indicates that improvements are modest, but not less reliable across 20 out of 21 included studies [37]. This indicates an overall positive effect in

XSL•FO
RenderX

increasing patient understanding when using animation videos across different health care settings.

Contrasting our work with the review by Moe-Byrne et al [36], several distinctions become apparent. First, our review focuses exclusively on animations designed as a tool for patient information, whereas the review by Moe-Byrne et al [36] includes public interventions, such as health promotion campaigns. Second, we considered solely RCTs, while the previous review also encompassed quasi-experimental studies. Finally, their analysis was not limited to health information recall but also examined changes in attitudes and behaviors [36]. Interestingly, Feeley et al [37] found that animations work better with patients than they do with samples from the general population or mixed groups of patients and nonpatients [37]. This finding supports our focus on patient populations and suggests that animations tailored specifically for certain patient groups may be particularly effective, possibly due to increased motivation and relevance of the information presented.

Notably, in our systematic review, all studies using whiteboard animation showed a significant positive effect, suggesting its effectiveness as an educational tool [25,29,34]. Even though only 3 studies used whiteboard animation, this can indicate that simple forms of animations are efficient in their ability to inform patients. This finding aligns with previous studies investigating the impact of details in animation videos, and their ability to be informing. For example, it was found that realism in animation, such as the level of visual detail, did not significantly affect cognitive learning performance, suggesting different levels of animation details can be equally effective for instructional purposes [38]. Furthermore, a study found that increasing detail can result in overloading the viewer's visual information processing, resulting in worse performance tests regarding information retention [39].

Most studies assessed recall immediately after intervention [22-24,26,27,29-35], with only a few extending the follow-up period [21,28,32]. This raises questions about the long-term efficacy of animations in patient recall since only 3 out of the 15 studies had follow-up, and only one of them found a significant effect at 4 weeks [28]. When comparing these results with the systematic review by Moe-Byrne et al [36] they conclude similar findings, indicating that the effect tends to be most present immediately after intervention [36], suggesting that while animations are effective in the short term, the sustainability of this effect warrants further investigation.

Our systematic review observed a diverse range of animation durations, from 1 to 15 minutes. Interestingly, Feeley et al [37] found that video length was a significant moderate of effectiveness, but only at an α=.06 criterion. Their analysis suggested that longer videos yielded smaller learning effects across the 16 studies that reported video length [37]. This aligns with our observation that a third of our included studies used short animation videos lasting 1 to 5 minutes, all of which demonstrated positive effects. These findings collectively suggest that concise, well-designed animations might be more effective in conveying health information, possibly due to their ability to maintain viewer attention and avoid cognitive overload.

Finally, our review found that most studies (80%) were conducted in high-income countries, Feeley et al [37] reported that 9 out of 21 studies were conducted in the United States, with the remaining 12 done in countries Germany, Japan, the Netherlands, Australia, Taiwan, Singapore, and Saudi Arabia. This similarity in geographical distribution underscores the need for more research in diverse global settings, particularly in low- and middle-income countries.

Strengths and Limitations

For this systematic review, we used different processes to reduce the potential risk of bias. First, it was a strength that the review protocol was registered with PROSPERO, and the PRISMA guidelines were adhered to, ensuring transparency and consistency in our reporting. Our collaboration with an experienced research librarian enabled a thorough 3-step search across multiple databases and manual searches, enhancing the breadth and relevance of the literature covered. The use of dual independent reviewers for screening, selection, data extraction, and risk of bias assessment, along with the systematic use of Covidence software, helped minimize biases and ensured a rigorous selection process. Using the Revised Cochrane risk-of-bias tool for randomized trials added robustness to our methodology, with discrepancies resolved through consultation with a third researcher.

This systematic review had several limitations. First, our search strategy may have missed some relevant papers. While we conducted a comprehensive search across multiple databases (PubMed, CINAHL, and Embase) and performed manual searches of reference lists, we did not have access to all potentially relevant databases, such as PsycINFO or CENTRAL. To mitigate this, we worked closely with an experienced research librarian to develop and refine our search strategy. However, future reviews could expand the search to include additional databases and consider using a wider range of search terms to capture all relevant studies.

Second, we faced challenges regarding the accessibility and evaluation of the animation video content used in the included studies. Only 8 out of 15 studies provided direct links to the animation videos, while the remaining studies offered only screenshots or written explanations. This limited our ability to fully assess the quality and effectiveness of the animations. Furthermore, we had to exclude several studies from the screening process due to this problem. To address this, we attempted to contact all authors to gain access to the animation videos or sufficient information about the intervention, but with limited success. This limitation may have influenced our results by preventing a comprehensive analysis of the animation characteristics that contribute to effective health information recall. Future studies in this area should focus on elaborating the design of the animation video and its role in the intervention.

Third, our decision to include only RCTs in this systematic review may have limited the breadth of evidence we were able to consider. While this choice was made to focus on the highest quality evidence available, it potentially excluded valuable insights from well-designed quasi-experimental studies or other study types. To mitigate this, we ensured a thorough search and screening process for RCTs. However, this limitation may have

XSL•FO
RenderX

influenced our results by potentially overestimating the effectiveness of animation videos, as RCTs with positive outcomes are more likely to be published. In addition, we may have missed nuanced findings from real-world implementations that are often captured in quasi-experimental designs. For future reviews in this area, researchers could consider including a wider range of study designs, such as high-quality quasi-experimental studies, while using appropriate tools to assess their risk of bias. This approach could provide a more comprehensive understanding of the effectiveness of animation videos in various real-world health care settings, while still maintaining a focus on methodological rigor. Finally, the final search was conducted in mid-2023 and additional studies may have been published in the interim given the progressive nature of this topic.

Future Directions

First, our review and Feeley et al [37] suggest that shorter animations may be more effective. Future interventions should focus on creating concise, targeted animations, typically around 5-8 minutes in length, to maintain viewer engagement and minimize cognitive overload. Second, the effectiveness of whiteboard animations in our review suggests that simple, clear animations can be highly effective. Future work should explore how to maximize the impact of these simpler animation styles across various health topics.

Third, given the limited long-term follow-up in current studies, future interventions should incorporate strategies to reinforce information over time. This could include providing patients with access to animations for repeated viewing or developing complementary materials that build on the animated content. Our systematic review is a great stepping stone for organizations, institutions, and so on, that want to develop their own animation videos on the subject of health care information.

Conclusion

In conclusion, this systematic review finds an overall positive impact of animation videos on short-term health information recall. Further studies should investigate the identified research gaps, particularly the long-term efficacy, impact on diverse populations, and the impact of animation style on the effectiveness of health information recall. Furthermore, future studies should focus on including a detailed description of the type of animation used in the intervention, or provide a link, to enable full transparency and usability of the studies.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 332 KB - jmir_v26i1e58306_app1.pdf]

Multimedia Appendix 2
Search Strategy.
[PDF File (Adobe PDF File), 279 KB - jmir_v26i1e58306_app2.pdf]

Multimedia Appendix 3
Overview/characteristics of included studies.
[PDF File (Adobe PDF File), 161 KB - jmir_v26i1e58306_app3.pdf]

References

1. Corp N, Mansell G, Stynes S, Wynne-Jones G, Morsø L, Hill JC, et al. Evidence-based treatment recommendations for neck and low back pain across europe: a systematic review of guidelines. Eur J Pain 2021;25(2):275-295 [FREE Full text] [doi: 10.1002/ejp.1679] [Medline: 33064878]
2. Howell D, Harth T, Brown J, Bennett C, Boyko S. Self-management education interventions for patients with cancer: a systematic review. Support Care Cancer 2017;25(4):1323-1355. [doi: 10.1007/s00520-016-3500-z] [Medline: 28058570]
3. Boyde M, Turner C, Thompson DR, Stewart S. Educational interventions for patients with heart failure: a systematic review of randomized controlled trials. J Cardiovasc Nurs 2011;26(4):E27-E35. [doi: 10.1097/JCN.0b013e3181ee5fb2] [Medline: 21076308]
4. Linn AJ, van Dijk L, Smit EG, Jansen J, van Weert JC. May you never forget what is worth remembering: the relation between recall of medical information and medication adherence in patients with inflammatory bowel disease. J Crohns Colitis 2013;7(11):e543-e550. [doi: 10.1016/j.crohns.2013.04.001] [Medline: 23660489]
5. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implement Sci 2012;7:50 [FREE Full text] [doi: 10.1186/1748-5908-7-50] [Medline: 22651257]
6. Feng YY, Chaves GSS, Shi W, Pakosh M, Zhang L, Gallagher R, et al. Education interventions in Chinese cardiac patients on health behaviours, disease-related knowledge, and health outcomes: a systematic review and meta-analysis. Patient Educ Couns 2021;104(5):1018-1029. [doi: 10.1016/j.pec.2020.12.001] [Medline: 33349505]

XSL•FO
RenderX

7. Greene J, Hibbard JH, Sacks R, Overton V, Parrotta CD. When patient activation levels change, health outcomes and costs change, too. Health Aff (Millwood) 2015;34(3):431-437. [doi: 10.1377/hlthaff.2014.0452] [Medline: 25732493]
8. Leiner M, Peinado J, Baylon A, Lopez I, Pathak I. Divide and conquer: improving parental understanding of health-related instructions using sequential pictorial instructions. Health Educ Res 2018;33(2):104-113. [doi: 10.1093/her/cyy004] [Medline: 29579193]
9. Muruganandam P, Neelamegam S, Menon V, Alexander J, Chaturvedi SK. COVID-19 and severe mental illness: impact on patients and its relation with their awareness about COVID-19. Psychiatry Res 2020;291:113265 [FREE Full text] [doi: 10.1016/j.psychres.2020.113265] [Medline: 32763536]
10. Juergens MC, Seekatz B, Moosdorf RG, Petrie KJ, Rief W. Illness beliefs before cardiac surgery predict disability, quality of life, and depression 3 months later. J Psychosom Res 2010;68(6):553-560. [doi: 10.1016/j.jpsychores.2009.10.004] [Medline: 20488272]
11. Tille F, Weishaar H, Gibis B, Schnitzer S. Patients' understanding of health information in Germany. Patient Prefer Adherence 2019;13:805-817 [FREE Full text] [doi: 10.2147/PPA.S202748] [Medline: 31190762]
12. Laccourreye L, Werner A, Cauchois R, Laccourreye O. [Interpretation and memorization of information provided to the patient for parotidectomy]. Ann Otolaryngol Chir Cervicofac 2009;126(2):53-60. [doi: 10.1016/j.aorl.2009.02.003] [Medline: 19345932]
13. Schnitman G, Gomes D, Deckelbaum D, Utiyama EM. Feasibility of multimedia animations as preoperative guides for urgent abdominal surgeries in a public hospital in Brazil. Health Educ Res 2022;37(5):333-354. [doi: 10.1093/her/cyac023] [Medline: 36125090]
14. Rosaasen N, Mainra R, Kukha-Bryson A, Nhin V, Trivedi P, Shoker A, et al. Development of a patient-centered video series to improve education before kidney transplantation. Patient Educ Couns 2018;101(9):1624-1629. [doi: 10.1016/j.pec.2018.04.014] [Medline: 29729857]
15. Meppelink CS, van Weert JC, Haven CJ, Smit EG. The effectiveness of health animations in audiences with different health literacy levels: an experimental study. J Med Internet Res 2015 Jan 13;17(1):e11 [FREE Full text] [doi: 10.2196/jmir.3979] [Medline: 25586711]
16. Dahodwala M, Geransar R, Babion J, de Grood J, Sargious P. The impact of the use of video-based educational interventions on patient outcomes in hospital settings: a scoping review. Patient Educ Couns 2018;101(12):2116-2124. [doi: 10.1016/j.pec.2018.06.018] [Medline: 30087021]
17. Keller M, Lucas T, Zachariah M, Feeley T, Tumiel Berhalter L, Kayler L. A scoping review of animated video?s effect on individual health knowledge. Austin Medical Sciences 2021;6(1):1048.
18. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
19. Covidence. Covidence systematic review software. Melbourne, Australia Veritas Health Innovation. 2022. URL: https://www.covidence.org/ [accessed 2024-11-07]
20. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
21. Schroeder M, Plotner EA, Sharma S, Hunter K, Spector S, Lipetskaia L. A randomized controlled trial of a multimedia patient education tool for stress versus urgency urinary incontinence. Obstetrical & Gynecological Survey 2021;76(9):531-532. [doi: 10.1097/OGX.0000000000000966] [Medline: 32925421]
22. Molher J, Sagardoy T, Dassé R, Champel S, Castetbon V, de Monès E. Educational video to improve memorization of preoperative information in parotidectomy: a randomized controlled study. Eur Ann Otorhinolaryngol Head Neck Dis 2022;139(5):281-284 [FREE Full text] [doi: 10.1016/j.anorl.2022.04.002] [Medline: 35538000]
23. Diniz LM, Oliveira CB, Machado GC, Maher CG, Verhagen AP, Fernandes DAM, et al. Effectiveness of brief patient information materials for promoting correct beliefs about imaging and inevitable consequences of low back pain: a randomised controlled trial. Clin Rehabil 2022;36(4):527-537. [doi: 10.1177/02692155211065974] [Medline: 34931854]
24. Yap J, Teo TY, Foong P, Binte Hussin N, Wang H, Shen T, et al. A randomized controlled trial on the effectiveness of a portable patient education video prior to coronary angiography and angioplasty. Catheter Cardiovasc Interv 2020;96(7):1409-1414. [doi: 10.1002/ccd.28655] [Medline: 31867778]
25. Mann J, Li L, Kulakov E, Bassett P, Birnie A. Home viewing of educational video improves patient understanding of mohs micrographic surgery. Clin Exp Dermatol 2022;47(1):93-97. [doi: 10.1111/ced.14845] [Medline: 34260092]
26. Lin MJ, Dubin DP, Younessi S, Khorasani H. Brief educational video improves patient understanding of mohs surgery: a randomized controlled trial. Dermatol Surg 2021;47(3):323-326. [doi: 10.1097/DSS.0000000000002843] [Medline: 33337731]
27. Gagné M, Legault C, Boulet LP, Charbonneau L, Lemyre M, Giguere AMC, et al. Impact of adding a video to patient education on quality of life among adults with atrial fibrillation: a randomized controlled trial. Patient Educ Couns 2019;102(8):1490-1498. [doi: 10.1016/j.pec.2019.03.015] [Medline: 30956021]
28. Dincer B, Bahçecik N. The effect of a mobile application on the foot care of individuals with type 2 diabetes: a randomised controlled study. Health Education Journal 2020;80(4):425-437. [doi: 10.1177/0017896920981617]

XSL•FO
RenderX

29. Chakravarthy B, Somasundaram S, Mogi J, Burns R, Hoonpongsimanont W, Wiechmann W, et al. Randomized pilot trial measuring knowledge acquisition of opioid education in emergency department patients using a novel media platform. Subst Abus 2018;39(1):27-31. [doi: 10.1080/08897077.2017.1375061] [Medline: 28873050]
30. Lattuca B, Barber-Chamoux N, Alos B, Sfaxi A, Mulliez A, Miton N, et al. Impact of video on the understanding and satisfaction of patients receiving informed consent before elective inpatient coronary angiography: a randomized trial. Am Heart J 2018;200:67-74. [doi: 10.1016/j.ahj.2018.03.006] [Medline: 29898851]
31. Kakinuma A, Nagatani H, Otake H, Mizuno J, Nakata Y. The effects of short interactive animation video information on preanesthetic anxiety, knowledge, and interview time: a randomized controlled trial. Anesth Analg 2011;112(6):1314-1318. [doi: 10.1213/ANE.0b013e31820f8c18] [Medline: 21346166]
32. Ellett L, Villegas R, Beischer A, Ong N, Maher P. Use of a multimedia module to aid the informed consent process in patients undergoing gynecologic laparoscopy for pelvic pain: randomized controlled trial. J Minim Invasive Gynecol 2014;21(4):602-611. [doi: 10.1016/j.jmig.2014.01.002] [Medline: 24462856]
33. Tou S, Tou W, Mah D, Karatassas A, Hewett P. Effect of preoperative two-dimensional animation information on perioperative anxiety and knowledge retention in patients undergoing bowel surgery: a randomized pilot study. Colorectal Dis 2013;15(5):e256-e265. [doi: 10.1111/codi.12152] [Medline: 23350660]
34. Mednick Z, Irrcher I, Hopman WM, Sharma S. Assessing a narrated white board animation as part of the consent process for intravenous fluorescein angiography: a randomized educational study. Can J Ophthalmol 2016;51(6):471-475. [doi: 10.1016/j.jcjo.2016.04.023] [Medline: 27938960]
35. Tipotsch-Maca SM, Varsits RM, Ginzel C, Vecsei-Marlovits PV. Effect of a multimedia-assisted informed consent procedure on the information gain, satisfaction, and anxiety of cataract surgery patients. J Cataract Refract Surg 2016;42(1):110-116. [doi: 10.1016/j.jcrs.2015.08.019] [Medline: 26948785]
36. Moe-Byrne T, Evans E, Benhebil N, Knapp P. The effectiveness of video animations as information tools for patients and the general public: a systematic review. Front Digit Health 2022;4:1010779 [FREE Full text] [doi: 10.3389/fdgth.2022.1010779] [Medline: 36777110]
37. Feeley TH, Keller M, Kayler L. Using animated videos to increase patient knowledge: a meta-analytic review. Health Educ Behav 2023;50(2):240-249. [doi: 10.1177/10901981221116791] [Medline: 35950326]
38. Lucas T. Exploring the effect of realism at the cognitive stage of complex motor skill learning. E-Learning and Digital Media 2019;16(4):242-266. [doi: 10.1177/2042753019835893]
39. Mayer RE, Heiser J, Lonn S. Cognitive constraints on multimedia learning: when presenting more material results in less understanding. Journal of Educational Psychology 2001;93(1):187-198. [doi: 10.1037//0022-0663.93.1.187]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
RoB2: Revised Cochrane risk-of-bias tool for randomized trials

Edited by N Cahill; submitted 12.03.24; peer-reviewed by I Mircheva, M Mahmic Kaknjo, T Moe-Byrne, B Assaye; comments to author 02.07.24; revised version received 26.09.24; accepted 09.10.24; published 30.12.24.

Please cite as:
Hansen S, Jensen TS, Schmidt AM, Strøm J, Vistisen P, Høybye MT
The Effectiveness of Video Animations as a Tool to Improve Health Information Recall for Patients: Systematic Review
J Med Internet Res 2024;26:e58306
URL: https://www.jmir.org/2024/1/e58306
doi:10.2196/58306
PMID:

©Steffen Hansen, Tue Secher Jensen, Anne Mette Schmidt, Janni Strøm, Peter Vistisen, Mette Terp Høybye. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Limitations of the Cough Sound-Based COVID-19 Diagnosis Artificial Intelligence Model and its Future Direction: Longitudinal Observation Study

Jina Kim[1], MSc; Yong Sung Choi[1], MD, PhD; Young Joo Lee[1], MD, PhD; Seung Geun Yeo[1], MD, PhD; Kyung Won Kim[2], MD, PhD; Min Seo Kim[3], MD, PhD; Masoud Rahmati[4,5], PhD; Dong Keon Yon[6,7*], MD, PhD; Jinseok Lee[1,8*], PhD

[1]Department of Biomedical Engineering, Kyung Hee University, Seoul, Republic of Korea

[2]Department of Radiology and Research Institute of Radiology, Asan Image Metrics, Clinical Trial Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[3]Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, United States

[4]Department of Physical Education and Sport Sciences, Faculty of Literature and Human Sciences, Lorestan University, Khoramabad, Iran

[5]Department of Physical Education and Sport Sciences, Faculty of Literature and Humanities, Vali-E-Asr University of Rafsanjan, Rafsanjan, Iran

[6]Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[7]Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[8]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Jinseok Lee, PhD
Department of Biomedical Engineering, Kyung Hee University
1732 Deogyeong-daero, Giheung-gu
Seoul, 17104
Republic of Korea
Phone: 82 2 6935 2476
Fax: 82 504 478 0201
Email: gonasago@khu.ac.kr

Abstract

Background: The outbreak of SARS-CoV-2 in 2019 has necessitated the rapid and accurate detection of COVID-19 to manage patients effectively and implement public health measures. Artificial intelligence (AI) models analyzing cough sounds have emerged as promising tools for large-scale screening and early identification of potential cases.

Objective: This study aimed to investigate the efficacy of using cough sounds as a diagnostic tool for COVID-19, considering the unique acoustic features that differentiate positive and negative cases. We investigated whether an AI model trained on cough sound recordings from specific periods, especially the early stages of the COVID-19 pandemic, were applicable to the ongoing situation with persistent variants.

Methods: We used cough sound recordings from 3 data sets (Cambridge, Coswara, and Virufy) representing different stages of the pandemic and variants. Our AI model was trained using the Cambridge data set with subsequent evaluation against all data sets. The performance was analyzed based on the area under the receiver operating curve (AUC) across different data measurement periods and COVID-19 variants.

Results: The AI model demonstrated a high AUC when tested with the Cambridge data set, indicative of its initial effectiveness. However, the performance varied significantly with other data sets, particularly in detecting later variants such as Delta and Omicron, with a marked decline in AUC observed for the latter. These results highlight the challenges in maintaining the efficacy of AI models against the backdrop of an evolving virus.

Conclusions: While AI models analyzing cough sounds offer a promising noninvasive and rapid screening method for COVID-19, their effectiveness is challenged by the emergence of new virus variants. Ongoing research and adaptations in AI methodologies are crucial to address these limitations. The adaptability of AI models to evolve with the virus underscores their potential as a

XSL•FO
RenderX

foundational technology for not only the current pandemic but also future outbreaks, contributing to a more agile and resilient global health infrastructure.

(J Med Internet Res 2024;26:e51640) doi:10.2196/51640

KEYWORDS

COVID-19 variants; cough sound; artificial intelligence; diagnosis; human lifestyle; SARS-CoV-2; AI model; cough; sound-based; diagnosis; sounds app; development; COVID-19; AI

Introduction

In 2019, the outbreak of SARS-CoV-2 set off a global pandemic, substantially impacting human lifestyles and posing an unprecedented challenge to health care systems worldwide [1]. Given its high infectivity, the urgent need for rapid and accurate COVID-19 detection has become paramount for effective patient management, timely isolation, and implementation of appropriate public health measures. Consequently, numerous studies have been undertaken to devise efficient methods for detecting infection status quickly and easily. Among these studies, artificial intelligence (AI) models that leverage prominent symptoms, such as coughing, have shown potential [2,3]. The ability to discern COVID-19 solely based on analyzing recorded cough sounds presents a straightforward and swift approach to screening individuals. These AI models offer hope for rapid, large-scale screening and early identification of potential cases. However, it is essential to acknowledge certain limitations in these research trends, particularly regarding variations in symptoms attributed to different SARS-CoV-2 variants [4]. The emergence of new variants has introduced complexities in symptom profiles, making it more challenging to rely solely on cough sounds for accurate COVID-19 detection. As the virus continues to evolve, ongoing research and adaptations in AI models will be crucial to address these challenges and improve the accuracy and effectiveness of diagnostic approaches. Particularly, patients with the Omicron variant have shown a decreased occurrence of characteristic symptoms such as loss of taste and smell, while cold-like symptoms such as sneezing and a blocked nose have increased [5]. In essence, the symptoms of COVID-19 have been increasingly resembling those of a common cold. This implies that the previous AI models designed to detect COVID-19 using cough sounds may not work as effectively or may not be as reliable both currently and in the future.

Methods

Overview

In this study, we first used cough sounds as a diagnostic tool for COVID-19 stems from distinctive acoustic features that can differentiate between positive and negative cases. Building on prior research that used mel spectrograms to discern variance in cough sound spectra, our study adopts the variable frequency complex demodulation technique. Variable frequency complex demodulation offers a higher resolution than mel spectrograms, thus amplifying the discernibility of these patterns [6]. As depicted in Figure S1 in Multimedia Appendix 1, the cough sound spectra of patients with COVID-19 exhibit a unique irregular spectral intensity distribution, characterized by an initial decline followed by a subsequent rise over time. In contrast, the spectra from non–COVID-19 groups display a more uniform pattern, typically showing either a consistent distribution or a gradual decline in intensity. Furthermore, our analysis extends to the differentiation in audio signal features between the 2 patient groups. The examination of 5 selected audio signal features indicates clear distinctions, which are consistent across different data sets. Illustrated in Figure S2 in Multimedia Appendix 1 [6-9], we observe that the mean spectral roll-off for patients with SARS-CoV-2 positive trends toward lower values compared with that of negative individuals. Conversely, the mean spectral bandwidth is generally higher for those with COVID-19. Additional features, such as the SD of the spectral centroid, present lower values for positive cases, whereas the SD of spectral bandwidth is elevated, and the SD of the zero-crossing rate is reduced when compared with negative cases.

Furthermore, we aimed to demonstrate that an AI model trained on cough sound recordings from a specific period, especially the early stages of the COVID-19 pandemic, is not applicable to the ongoing situation with persistent variants. To illustrate these findings, we used 3 cough data sets named Cambridge, Coswara, and Virufy [5,10,11]. The Cambridge data set consists of audio recordings of variable-length cough audio collected through the COVID-19 Sounds App, developed by the University of Cambridge. The data measurement period encompasses the early stages of the wild type and Alpha variant occurrences (April 30, 2020, to April 26, 2021). We trained our AI model using the Cambridge data set. More specifically, we split the data into 8:2 for train and test data. Subsequently, for the train data, we performed 3-fold cross-validation and evaluated the model using the test data. The Virufy data set also includes crowdsourced cough sounds to identify patterns that signify respiratory diseases, such as COVID-19. The data measurement period primarily covers the early stages when the wild type was present (April 9, 2020, to November 26, 2020). We used the data set as additional test data. The Coswara data set also includes worldwide crowdsourced data collected through a website app, including cough, breath, and voice recordings for COVID-19 diagnosis. The data measurement period extended from the Alpha variant to the Omicron variant occurrences. We also used the data set as additional test data. To evaluate the performance of our AI model based on different variants, we categorized the data into 3 periods, each dominated by the Alpha, Delta, and Omicron variants, respectively. We summarized the data sets, along with the periods and COVID-19 variant status in which they were measured, in Figure 1 and Tables S1 and S2 in Multimedia Appendix 1 [6-9].

XSL•FO
RenderX

Figure 1. Area under the receiver operating curve (AUC) values for each data set and measurement period.

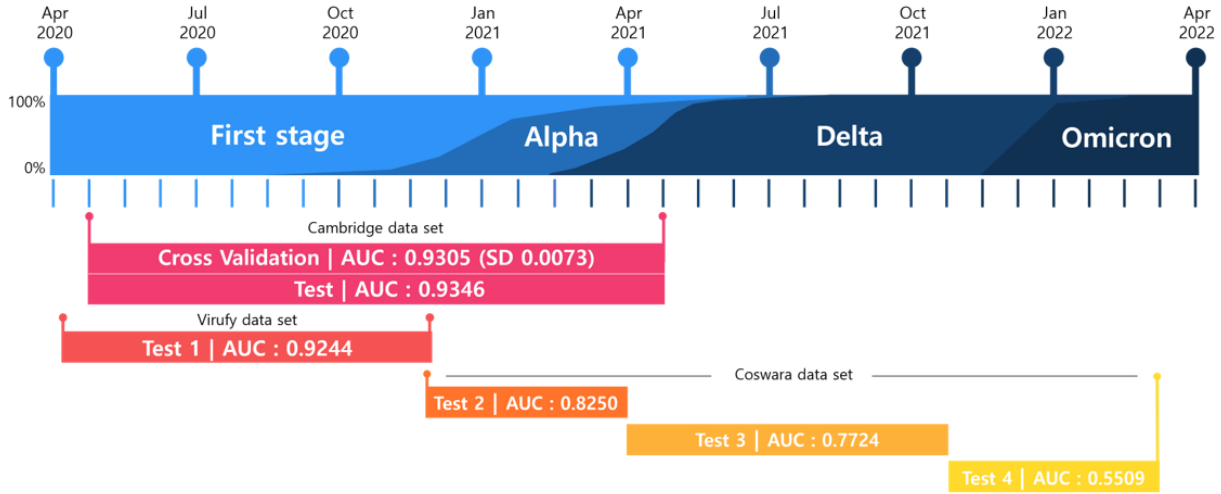

Ethical Considerations

The protocol of this study was approved by the institutional review board of Kyung Hee University Hospital (KHUH 2022-06-042).

Results

After the cough sound preprocessing, we developed an AI model to detect COVID-19 from extracted features and time-frequency spectrum using variable frequency complex demodulation (Multimedia Appendix 1 [6-9]). Table S2 in Multimedia Appendix 1 [6-9] summarizes the performance of our proposed AI model according to each data set and data measurement period. The results show that the cross-validation and test results from the Cambridge data set overlap, with an area under the receiver operating curve (AUC) of 0.93. Furthermore, the Virufy data set, which was measured around a similar period as the Cambridge data set, also demonstrated a similar performance: the AUC was 0.92. On the other hand, in the case of the Coswara data set, divided into 3 groups based on the variants, it was observed that the AUC for Alpha was 0.83; for Delta, 0.77; and for Omicron, 0.55; this indicates a decline in performance as the variants progressed. Figures 1 and 2 summarize the overall performance for each data set and measurement period.

Figure 2. Our proposed model architecture. BN: batch normalization; FC: fully connected; PD: prediction; SB: spectral bandwidth; SD SC: SD of spectral centroid; SD ZCR: SD of zero crossing rate; SD SB: SD of spectral bandwidth; SR: spectral rolloff.

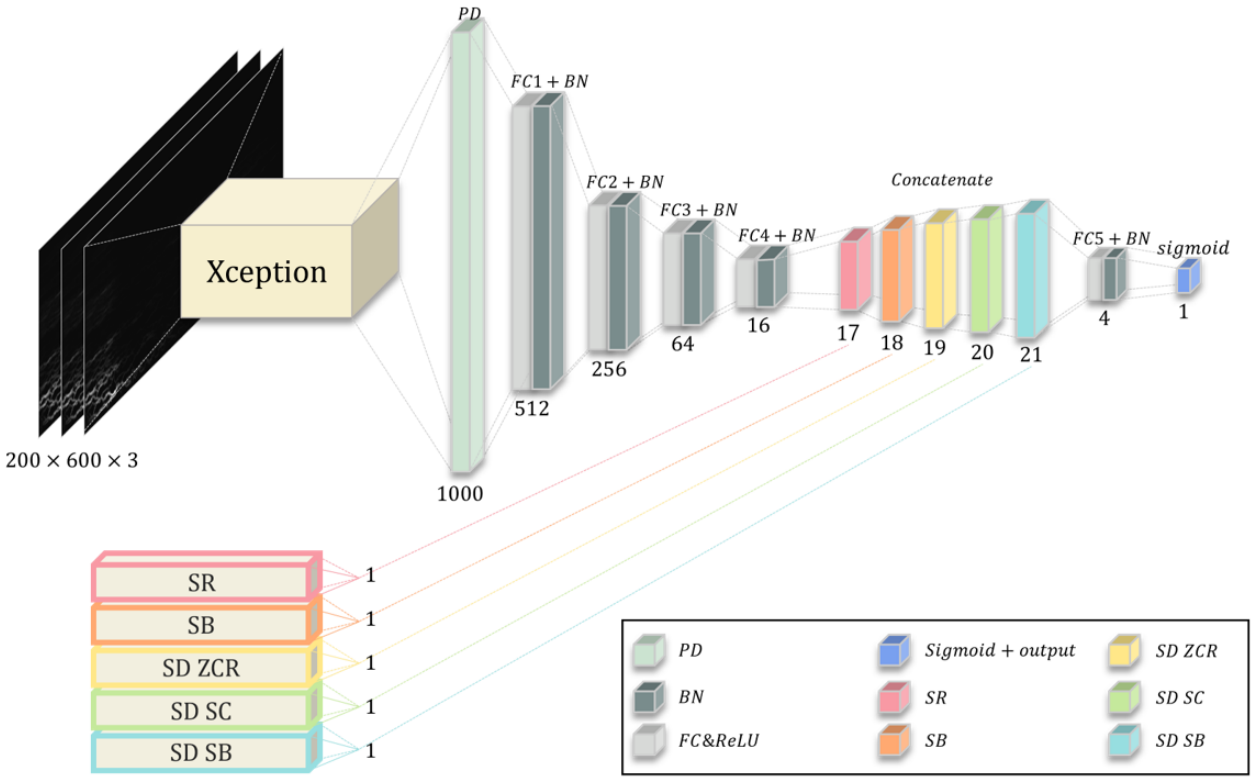

XSL•FO
RenderX

Discussion

The fight against COVID-19 has been marked by the deployment of various traditional diagnostic methods, each serving as a critical pillar in our response to the pandemic. Central among these is the reverse transcription-polymerase chain reaction test, widely regarded as the gold standard for its high specificity and sensitivity [12,13]. Rapid antigen tests and serological assays have also been instrumental, offering quick screening and insight into past infections, respectively. Despite their strengths, these methods have limitations, including resource dependency, time constraints, and varying degrees of accuracy. In this context, AI has emerged as a powerful aid, augmenting traditional diagnostic methods and addressing their shortcomings. By facilitating enhanced data analysis and interpretation, AI algorithms have the potential to streamline reverse transcription-polymerase chain reaction workflows, improve the reliability of antigen tests, and provide a more nuanced understanding of serological data. The integration of AI extends further to the analysis of medical imaging, where it aids in identifying patterns indicative of COVID-19, thus offering a valuable complement to molecular testing [14-16].

Our research has focused on the novel application of AI in analyzing cough sounds, a symptom-based method with promise for noninvasive, rapid screening. We demonstrated that AI models, when trained on cough sounds, could provide a swift approach to preliminary screening, though their performance varied with the emergence of new variants. Despite the performance variability, the inherent adaptability of AI models ensures their lasting relevance. These models can be reconfigured as pretrained frameworks ready to be fine-tuned against emerging viral strains. Such adaptability allows for the rapid deployment of AI tools in response to evolving pathogens, showcasing the potential of AI to serve not only the current pandemic but also as a foundational technology for future outbreaks. The capacity of these AI models to evolve in step with the virus paves the way for an agile diagnostic ecosystem that can quickly adapt to new threats, ultimately contributing to a more resilient global health infrastructure.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (HV22C0233), by a National Research Foundation of Korea grant funded by the Korea government (RS-2023-00253081), and by the Hyundai Motor Chunag Mong-Koo Foundation. The authors did not use generative AI tools in writing the manuscript.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JL had full access to all of the data in the study and took responsibility for the integrity of the AI model and performance analysis. All authors approved the final version before submission. The study's concept and design were devised by DKY and JL. The acquisition was led by JK and JL. Model development was spearheaded by JK. Analysis and interpretation involved DKY and JL. Validation was performed by YSC, YJL, SGY, and KWK. Drafting of the manuscript was done by JK, DKY, and JL. Critical revision of the manuscript for important intellectual content was a collective effort of all authors. Study supervision was undertaken by DKY and JL. JL and DKY supervised the study and are guarantors for this study. JL and DKY contributed equally as corresponding authors. The corresponding authors attest that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables and figures.
[PDF File (Adobe PDF File), 589 KB - jmir_v26i1e51640_app1.pdf]

References

1. Cha Y, Jung W, Seo M, Rahmati M. The emerging pandemic recent: SARS-CoV-2. Life Cycle 2023;3:e2 [FREE Full text] [doi: 10.54724/lc.2023.e2]
2. Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J Eng Med Biol 2020;1:275-281 [FREE Full text] [doi: 10.1109/OJEMB.2020.3026928] [Medline: 34812418]
3. Andreu-Perez J, Perez-Espinosa H, Timonet E, Kiani M, Giron-Perez MI, Benitez-Trinidad AB, et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need Covid-19 test and severity levels. IEEE Trans Serv Comput 2022;15(3):1220-1232 [FREE Full text] [doi: 10.1109/TSC.2021.3061402] [Medline: 35936760]

XSL•FO
RenderX

4. Choi YJ, Acharya KP. How serious is the Omicron variant? transmissibility, genomics, and responses to COVID-19 vaccines, and 'Stealth' Omicron variants. Life Cycle 2022;2:e7 [FREE Full text] [doi: 10.54724/lc.2022.e7]
5. Whitaker M, Elliott J, Bodinier B, Barclay W, Ward H, Cooke G, et al. Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England. Nat Commun 2022;13(1):6856 [FREE Full text] [doi: 10.1038/s41467-022-34244-2] [Medline: 36369151]
6. Wang H, Siu K, Ju K, Chon KH. A high resolution approach to estimating time-frequency spectra and their amplitudes. Ann Biomed Eng 2006;34(2):326-338. [doi: 10.1007/s10439-005-9035-y] [Medline: 16463086]
7. Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. Applied Acoustics 2020 Jan;158:107020. [doi: 10.1016/j.apacoust.2019.107020]
8. Chang AB. The physiology of cough. Paediatr Respir Rev 2006 Mar;7(1):2-8. [doi: 10.1016/j.prrv.2005.11.009] [Medline: 16473809]
9. Chollet F. Xception: deep learning with depthwise separable convolutions. 2017 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI p. 1800-1807. [doi: 10.1109/CVPR.2017.195]
10. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. Sci Data 2021;8(1):156 [FREE Full text] [doi: 10.1038/s41597-021-00937-4] [Medline: 34162883]
11. Bhattacharya D, Sharma NK, Dutta D, Chetupalli SR, Mote P, Ganapathy S, et al. Coswara: a respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. Sci Data 2023;10(1):397 [FREE Full text] [doi: 10.1038/s41597-023-02266-0] [Medline: 37349364]
12. Park M, Won J, Choi BY, Lee CJ. Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. Exp Mol Med 2020;52(6):963-977 [FREE Full text] [doi: 10.1038/s12276-020-0452-7] [Medline: 32546849]
13. Tombuloglu H, Sabit H, Al-Khallaf H, Kabanja JH, Alsaeed M, Al-Saleh N, et al. Multiplex real-time RT-PCR method for the diagnosis of SARS-CoV-2 by targeting viral N, RdRP and human RP genes. Sci Rep 2022;12(1):2853 [FREE Full text] [doi: 10.1038/s41598-022-06977-z] [Medline: 35181721]
14. Chung H, Ko H, Lee H, Yon DK, Lee WH, Kim TS, et al. Development and validation of a deep learning model to diagnose COVID-19 using time-series heart rate values before the onset of symptoms. J Med Virol 2023;95(2):e28462. [doi: 10.1002/jmv.28462] [Medline: 36602055]
15. Chung H, Ko H, Kang WS, Kim KW, Lee H, Park C, et al. Prediction and feature importance analysis for severity of COVID-19 in South Korea using artificial intelligence: model development and validation. J Med Internet Res 2021;23(4):e27060 [FREE Full text] [doi: 10.2196/27060] [Medline: 33764883]
16. Ko H, Chung H, Kang WS, Park C, Kim DW, Kim SE, et al. An artificial intelligence model to predict the mortality of COVID-19 patients at hospital admission time using routine blood samples: development and validation of an ensemble model. J Med Internet Res 2020;22(12):e25442 [FREE Full text] [doi: 10.2196/25442] [Medline: 33301414]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating curve

Edited by T Leung, T de Azevedo Cardoso; submitted 06.08.23; peer-reviewed by L Hua, C Bossley; comments to author 27.10.23; revised version received 10.11.23; accepted 02.01.24; published 06.02.24.

Please cite as:
Kim J, Choi YS, Lee YJ, Yeo SG, Kim KW, Kim MS, Rahmati M, Yon DK, Lee J
Limitations of the Cough Sound-Based COVID-19 Diagnosis Artificial Intelligence Model and its Future Direction: Longitudinal Observation Study
J Med Internet Res 2024;26:e51640
URL: https://www.jmir.org/2024/1/e51640
doi:10.2196/51640
PMID:38319694

©Jina Kim, Yong Sung Choi, Young Joo Lee, Seung Geun Yeo, Kyung Won Kim, Min Seo Kim, Masoud Rahmati, Dong Keon Yon, Jinseok Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Patterns of Ownership and Usage of Wearable Devices in the United States, 2020-2022: Survey Study

Ashwini Nagappan[1], MBE; Adriana Krasniansky[2], MTS; Madelyn Knowles[2], MPH

[1]Department of Health Policy and Management, University of California, Los Angeles, Los Angeles, CA, United States
[2]Rock Health, San Francisco, CA, United States

Corresponding Author:
Ashwini Nagappan, MBE
Department of Health Policy and Management
University of California, Los Angeles
650 Charles E Young Dr S
31-269 CHS
Los Angeles, CA, 90095
United States
Phone: 1 3108252594
Email: ashwininagappan@ucla.edu

Abstract

Background: Although wearable technology has become increasingly common, comprehensive studies examining its ownership across different sociodemographic groups are limited.

Objective: The aims of this study were to (1) measure wearable device ownership by sociodemographic characteristics in a cohort of US consumers and (2) investigate how these devices are acquired and used for health-related purposes.

Methods: Data from the Rock Health Digital Health Consumer Adoption Survey collected from 2020 to 2022 with 23,974 US participants were analyzed. The sample was US Census–matched for demographics, including age, race/ethnicity, gender, and income. The relationship between sociodemographic factors and wearable ownership was explored using descriptive analysis and multivariate logistic regression.

Results: Of the 23,974 respondents, 10,679 (44.5%) owned wearables. Ownership was higher among younger individuals, those with higher incomes and education levels, and respondents living in urban areas. Compared to those aged 18-24 years, respondents 65 years and older had significantly lower odds of wearable ownership (odds ratio [OR] 0.18, 95% CI 0.16-0.21). Higher annual income (≥US $200,000; OR 2.27, 95% CI 2.01-2.57) and advanced degrees (OR 2.23, 95% CI 2.01-2.48) were strong predictors of ownership. Living in rural areas reduced ownership odds (OR 0.65, 95% CI 0.60-0.72). There was a notable difference in ownership based on gender and health insurance status. Women had slightly higher ownership odds than men (OR 1.10, 95% CI 1.04-1.17). Private insurance increased ownership odds (OR 1.28, 95% CI 1.17-1.40), whereas being uninsured (OR 0.41, 95% CI 0.36-0.47) or on Medicaid (OR 0.75, 95% CI 0.68-0.82) decreased the odds of ownership. Interestingly, minority groups such as non-Hispanic Black (OR 1.14, 95% CI 1.03-1.25) and Hispanic/Latine (OR 1.20, 95% CI 1.10-1.31) respondents showed slightly higher ownership odds than other racial/ethnic groups.

Conclusions: Our findings suggest that despite overall growth in wearable ownership, sociodemographic divides persist. The data indicate a need for equitable access strategies as wearables become integral to clinical and public health domains.

(J Med Internet Res 2024;26:e56504) doi:10.2196/56504

KEYWORDS

digital health; health equity; adoption; usage patterns; wearable devices; United States; adoption; technology; sociodemographic; survey; health insurance; public health

Introduction

Consumers are increasingly integrating wearable technology, which refers to devices that can be worn on the body to capture data [1], into their daily lives [2]. Wearable devices now range from smartwatches to fertility trackers and sleep trackers, serving a variety of health monitoring applications [3,4]. While prior studies have focused on the adoption of wearables within

XSL•FO
RenderX

specific populations such as adolescents, older adults, and underserved populations [5-7], large-scale studies examining differences in wearable adoption across sociodemographic characteristics remain limited.

This study aimed to measure wearable ownership by sociodemographic characteristics in a cohort of US consumers. Specifically, a respondent was considered a wearable owner if, at the time of the survey, they reported currently owning a wearable device, even if they did not currently use the device. Additionally, we examined how respondents received their wearable devices and explored relevant health-related uses. The findings provide important insights to guide wearable product development and identify priorities to improve ownership among underrepresented groups.

Methods

Study Design

This study aggregated data from 3 consecutive years (2020-2022) of the Rock Health Digital Health Consumer Adoption Survey [8]. The total sample consisted of 23,974 US-based respondents (7,980 in 2020, 7,980 in 2021, and 8,014 in 2022). Survey respondents were not excluded from participating in the study in subsequent years and repeat respondents comprised 3.3% (n=784) of the total cohort. The survey was administered by Toluna, a survey management organization. Toluna used its existing panel, initially recruited via online advertising, to identify participants. Eligible participants were 18 years or older, and the sample was Census-matched each year according to age, race/ethnicity, geographic region, gender, and annual household income.

Ethical Considerations

In accordance with the Common Rule (Federal Policy for the Protection of Human Subjects, 82 Federal Regulation 7259, January 19, 2017 [9]), this study was exempt from regulations for research with human subjects as the data were deidentified.

Data Collection and Survey Questions

Surveys were digitally administered through Toluna's platform, and respondents used their personal desktop, laptop, smartphone, or tablet to complete the survey in English. The survey encompassed 4 main domains: (1) sociodemographic factors, (2) health status, (3) adoption of digital health tools (eg, telemedicine, wearables), and (4) attitudes toward and perceptions of digital health technology. For this study, we focused on questions about wearable technology adoption.

Measures

The primary outcome was wearable ownership as measured by the survey question, "Do you own a wearable device or smartwatch (note: this excludes smartphones) that helps you track your health? This could include: number of steps/exercise, sleep, heart rate, or blood pressure." Covariates included sociodemographic characteristics (eg, age, gender, race/ethnicity, income level, and educational level) and self-reported health status. The secondary analysis was exploratory, delving into the cohort of wearable owners to investigate the source of their devices and their reasons for using wearables. Respondents were asked: "Do or did you use your wearable device or smartwatch for any of the purposes listed below? Select all that apply."

Data Analysis

We performed a descriptive analysis of the pooled sample and examined associations with wearable ownership. The χ^2 test was used to assess variation across covariates and multivariate logistic regression was used to identify factors associated with wearable ownership. Additional analyses explored how respondents acquired their wearables and what they used the devices for. Analyses were conducted in Stata v.16.

Results

The analytic sample included 23,974 survey respondents, including 13,295 wearable nonowners and 10,679 wearable owners (Table 1). Younger respondents (18-44 years) exhibited higher ownership rates, while nonowners skewed toward the older demographic (55+ years). An income gradient emerged, with those earning less than US $25,000 annually representing a smaller fraction of owners (10.5%) than nonowners (25.8%), with ownership rates increasing with income. Advanced degree holders represented 31.4% of wearable device owners versus 13.5% of nonowners.

Ownership varied by rurality, with 45.6% of owners in urban areas compared to 12.6% in rural areas. Nonowners were more evenly distributed, with 27.8% in urban areas and 23.4% in rural areas. Men represented a slightly higher proportion of owners than women. Health insurance coverage emerged as another differentiator; 41.2% of wearable owners had employment-based insurance and 21.5% had private insurance, while nonowners were more likely to be covered by Medicare, Medicaid, or to be uninsured. Ownership also varied across racial and ethnic groups, with higher proportions of non-Hispanic Black, non-Hispanic Asian, and Hispanic/Latine respondents in the owner group. Nearly one-third of owners self-reported excellent health, which was over twice the rate among nonowners.

Multivariable logistic regression was used to model the relationship between sociodemographics and wearable ownership, controlling for covariates (Table 2). Wearable device ownership increased from 2020 to 2022. Compared to the age group of 18-24 years, all other age groups showed decreased odds of ownership, especially those 65 years and older. Higher income and education significantly increased ownership odds, with those earning US $200,000 and above per year and those with advanced degrees showing more than double the odds compared to those of the comparator groups.

XSL•FO
RenderX

Table 1. Respondent characteristics, by wearable ownership, 2020-2022 (N=23,974).

Characteristic	Nonowners (n=13,295), n (%)	Owners (n=10,679), n (%)	*P* value[a]
Age (years)			<.001
18-24	1340 (10.1)	1424 (13.3)	
25-34	1743 (13.1)	2574 (24.1)	
35-44	1457 (11.0)	2586 (24.2)	
45-54	2112 (15.9)	1778 (16.6)	
55-64	2677 (20.1)	1196 (11.2)	
65+	3966 (29.8)	1121 (10.5)	
Annual household income (US $)			<.001
<25,000	3428 (25.8)	1117 (10.5)	
25,000-34,999	1401 (10.5)	758 (7.1)	
35,000-49,999	1854 (13.9)	1019 (9.5)	
50,000-74,999	2365 (17.8)	1627 (15.2)	
75,000-99,999	1381 (10.4)	1493 (14.0)	
100,000-149,999	1501 (11.3)	2072 (19.4)	
150,000-199,999	549 (4.1)	1128 (10.6)	
≥200,000	717 (5.4)	1443 (13.5)	
Prefer not to say	99 (0.7)	22 (0.2)	
Education			<.001
Less than high school	430 (3.2)	151 (1.4)	
High school graduate (includes equivalency)	3376 (25.4)	1560 (14.6)	
Some college, no degree	3232 (24.3)	1680 (15.7)	
Associate degree	1519 (11.4)	1227 (11.5)	
Bachelor degree	2897 (21.8)	2677 (25.1)	
Master degree	1244 (9.4)	2297 (21.5)	
PhD	160 (1.2)	306 (2.9)	
Graduate or professional degree (eg, MD, JD)	386 (2.9)	749 (7.0)	
Prefer not to say	51 (0.4)	32 (0.3)	
Rurality			<.001
Rural	3105 (23.4)	1346 (12.6)	
Suburban	6496 (48.9)	4461 (41.8)	
Urban	3694 (27.8)	4872 (45.6)	
Gender			<.001
Woman	6840 (51.4)	5116 (47.9)	
Man	6330 (47.6)	5476 (51.3)	
Other	77 (0.6)	74 (0.7)	
Prefer not to disclose	48 (0.4)	13 (0.1)	
Health insurance			<.001
Employment-based	3559 (26.8)	4405 (41.2)	
Private purchase	1375 (10.3)	2294 (21.5)	
Medicare (over the age of 65)	3995 (30.0)	1559 (14.6)	
Medicaid	2105 (15.8)	1287 (12.1)	
Other public	454 (3.4)	375 (3.5)	

XSL•FO
RenderX

Characteristic	Nonowners (n=13,295), n (%)	Owners (n=10,679), n (%)	*P* value[a]
Other	253 (1.9)	150 (1.4)	
Uninsured	1052 (7.9)	374 (3.5)	
I don't know	502 (3.8)	235 (2.2)	
Race/ethnicity			<.001
NH[b]-White	8941 (67.3)	6443 (60.3)	
NH-Black or African-American	1373 (10.3)	1310 (12.3)	
NH-American Indian or Alaska Native	53 (0.4)	58 (0.5)	
NH-Asian	603 (4.5)	570 (5.3)	
NH-Hawaiian Native or Pacific Islander	38 (0.3)	59 (0.6)	
NH-Other	69 (0.5)	24 (0.2)	
Multiracial, NH	257 (1.9)	230 (2.2)	
Hispanic/Latine	1881 (14.1)	1962 (18.4)	
Prefer not to say	80 (0.6)	23 (0.2)	
Self-reported health status			<.001
Very poor	160 (1.2)	36 (0.3)	
Poor	958 (7.2)	238 (2.2)	
Moderate	3384 (25.5)	1609 (15.1)	
Good	6913 (52.0)	5386 (50.4)	
Excellent	1880 (14.1)	3410 (31.9)	

[a]*P* values were calculated using χ^2 tests to compare the distributions of each variable across the two groups. Due to the large sample sizes, statistically significant *P* values may be observed even for minor differences; therefore, *P* values should be interpreted with caution and in the context of practice relevance beyond statistical significance.

[b]NH: non-Hispanic.

XSL•FO
RenderX

Table 2. Multivariable analysis of wearable ownership and sociodemographic characteristics, 2020-2022 (N=23,974).

Variable	Adjusted odds ratio (95% CI)	*P* value
Year (reference: 2020)		
2021	1.16 (1.08-1.25)	<.001
2022	1.27 (1.18-1.36)	<.001
Age (years) (reference: 18-24)		
25-34	0.98 (0.88-1.09)	.68
35-44	0.90 (0.80-1.01)	.06
45-54	0.52 (0.46-0.58)	<.001
55-64	0.32 (0.28-0.36)	<.001
65+	0.18 (0.16-0.21)	<.001
Annual income (US $) (reference: <50,000)		
50,000-99,999	1.62 (1.50-1.75)	<.001
100,000-149,999	2.02 (1.83-2.23)	<.001
150,000-199,999	2.73 (2.40-3.12)	<.001
≥200,000	2.27 (2.01-2.57)	<.001
Education (reference: ≤high school diploma)		
Some college	1.27 (1.17-1.38)	<.001
Bachelor's degree	1.44 (1.32-1.59)	<.001
Advanced degree	2.23 (2.01-2.48)	<.001
Rurality (reference: urban)		
Rural	0.65 (0.60-0.72)	<.001
Suburban	0.72 (0.68-0.77)	<.001
Gender: women (reference: men)	1.10 (1.04-1.17)	.001
Health insurance (reference: employment-based)		
Private purchase	1.28 (1.17-1.40)	<.001
Medicare (over the age of 65)	1.06 (0.95-1.19)	.29
Medicaid	0.75 (0.68-0.82)	<.001
Uninsured	0.41 (0.36-0.47)	<.001
Race/ethnicity (reference: NH[a]-White)		
NH-Black or African-American	1.14 (1.03-1.25)	.008
NH-American Indian or Alaska Native	1.04 (0.68-1.57)	.87
NH-Asian	0.82 (0.72-0.94)	.004
NH-Hawaiian Native or Pacific Islander	1.59 (1.03-2.45)	.04
NH-Other	0.62 (0.38-1.03)	.07
Multiracial, NH	1.02 (0.84-1.25)	.82
Hispanic/Latine	1.20 (1.10-1.31)	<.001
Health status (reference: excellent)		
Very poor	0.30 (0.21-0.45)	<.001
Poor	0.36 (0.31-0.43)	<.001
Moderate	0.53 (0.48-0.58)	<.001
Good	0.66 (0.62-0.72)	<.001

[a]NH: non-Hispanic.

XSL•FO
RenderX

Living in rural areas was associated with 35% lower ownership odds, and women demonstrated slightly higher odds than men. Compared to employment-based plans, private insurance increased ownership odds. Conversely, being uninsured significantly reduced the odds of wearable ownership, as did being covered by Medicaid. Relative to non-Hispanic White respondents, non-Hispanic Black and Hispanic/Latine respondents had higher ownership odds, whereas non-Hispanic Asian respondents showed lower odds. Excellent self-reported health predicted the highest odds of ownership.

The majority of owners purchased their devices (61.1%), followed by 24.1% who received them as gifts (Table 3). Smaller proportions obtained devices from health care providers, employers, or insurance companies. Wearable owners primarily used their devices for fitness and wellness; top-use cases included physical activity (56.6%), fitness training (55.4%), and losing weight (43.8%). Approximately one-third of respondents reported using their wearable to manage a diagnosed condition (Table 3).

Table 3. Source of wearable device and top health-related uses of wearable devices among wearable owners (N=10,679).

Source and use	Respondents, n (%)
Source of wearable device	
I purchased it myself	6521 (61.1)
It was a gift	2574 (24.1)
It was offered to me by my doctor/clinician	862 (8.1)
It was offered to me by my insurance company	376 (3.5)
It was offered to me by my employer	123 (1.2)
None of these	223 (2.1)
Top health-related uses of wearable devices	
More physical activity	6049 (56.6)
Fitness training	5911 (55.4)
Lose weight	4682 (43.8)
Sleep better	3834 (35.9)
Manage diagnosed condition	3643 (34.1)
Manage stress	3075 (28.8)

Discussion

Principal Findings

Despite steadily increasing wearable ownership from 2020 to 2022 among US consumers, sociodemographic disparities persist related to age, income, education, and residence. Ownership skews toward younger, more affluent, and highly educated respondents living in urban areas. This is consistent with prepandemic findings [2,10], suggesting continuity in the digital divide [11]. This unequal access impacts the potential benefits of wearables on health promotion and health outcomes [12-15].

Surprisingly, non-Hispanic Black and Hispanic/Latine respondents have slightly higher ownership odds than non-Hispanic White respondents. If corroborated by future research indicating active usage among these communities, this finding suggests that wearables can be harnessed to monitor chronic conditions prevalent in these communities, such as hypertension and diabetes [16,17]. The inclination of wearable ownership among these groups presents an opportunity to inform the design and development of new wearable technologies [18].

Reliance on direct-to-consumer channels (eg, direct purchase or gifts) for acquiring wearables underscores the discretionary nature of wearable purchase and signals a possible relationship between health consciousness and wearable ownership [19]. The self-motivated nature of acquisition points to the importance of consumer preferences in uptake and indicates a potential barrier to access among those with limited discretionary means. This signals a need to improve equitable access through alternative distribution pathways.

As wearables are considered for clinical care [20] and public health surveillance applications [21-23] such as communicable respiratory (eg, COVID-19) and infectious (eg, dengue) diseases, it will be critical to continue to track ownership patterns and address biases in the wearable user population related to age, income, education, and area of residence [24,25]. Thus, effective wearable-based public health surveillance efforts must actively mitigate such biases within the current wearable-owning population and increase wearable uptake among subgroups reporting lower rates of ownership. In the near term, acknowledging and adjusting for biases is essential. This requires recognizing barriers that lead to uneven ownership. Over the longer term, there is an opportunity to address barriers to wearable ownership, which can allow the user cohort to gradually better represent the general population.

Limitations

Study limitations include sampling bias, as the survey undersamples those without digital devices (eg, smartphones, tablets, or computers), regular internet connectivity, and non-English speakers. While the study used a pooled sample across 3 years, a small proportion of respondents were surveyed

XSL•FO
RenderX

in multiple waves, which may affect comprehensiveness. Further, the study focuses on ownership, not wearable adoption, impacting findings related to active usage patterns of wearables. Additionally, respondents who indicated ownership may or may not own more than one device, which could impact the responses to subsequent questions.

Conclusion

This study provides an updated benchmark on wearable device ownership among US consumers, highlighting sociodemographic disparities in ownership across age, income, education, and residence. While some traditionally disadvantaged groups demonstrate higher ownership rates, other findings reflect persisting disparities, painting a mixed picture of progress toward a more equitable wearable distribution. This underscores the need for further research, particularly into higher ownership rates among minority groups. Future research may also explore reasons for nonownership among wearable nonowners, such as lack of interest, privacy concerns, cost, or other barriers. Understanding these differences is critical for designing inclusive wearable products and business models, and for ensuring the appropriate use of wearable data in health care and public health research by accounting for biases in sample representation.

Acknowledgments

AN was supported by a National Institutes of Health (NIH)/National Center for Advancing Translational Science (NCATS) University of California, Los Angeles (UCLA) Clinical and Translational Science Institute (CTSI) grant (TL1TR001883). The content is solely the responsibility of the authors and does not represent the official views of the NIH or Rock Health.

Authors' Contributions

AN was involved in conceptualization; data curation, analysis, validation, and visualization; writing the original draft; and writing, review, and editing of the manuscript. AK was involved in data validation and writing, review, and editing of the manuscript. MK was involved in data curation, analysis, and validation; writing the original draft; and writing, review, and editing of the manuscript. All authors had full access to the data and accept responsibility to submit for publication.

Conflicts of Interest

AN reported fees from Rock Health as a Research Fellow. AK and MK are employed by Rock Health. No other disclosures are reported.

References

1. Friend SH, Ginsburg GS, Picard RW. Wearable digital health technology. N Engl J Med 2023 Nov 30;389(22):2100-2101. [doi: 10.1056/NEJMe2303219] [Medline: 38048193]
2. Vogels EA. About one-in-five Americans use a smart watch or fitness tracker. Pew Research Center. 2020 Jan 09. URL: https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/ [accessed 2023-10-21]
3. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 09;8(11):e18907 [FREE Full text] [doi: 10.2196/18907] [Medline: 33164904]
4. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018 Sep;15(5):429-448. [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
5. Kim EH, Jenness JL, Miller AB, Halabi R, de Zambotti M, Bagot KS, et al. Association of demographic and socioeconomic indicators with the use of wearable devices among children. JAMA Netw Open 2023 Mar 01;6(3):e235681 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.5681] [Medline: 36995714]
6. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
7. Holko M, Litwin TR, Munoz F, Theisz KI, Salgin L, Jenks NP, et al. Wearable fitness tracker use in federally qualified health center patients: strategies to improve the health of all of us using digital health devices. NPJ Digit Med 2022 Apr 25;5(1):53. [doi: 10.1038/s41746-022-00593-x] [Medline: 35469045]
8. Knowles M, Krasniansky A, Nagappan A. Consumer adoption of digital health in 2022: moving at the speed of trust. Rock Health. 2023 Feb 21. URL: https://rockhealth.com/insights/consumer-adoption-of-digital-health-in-2022-moving-at-the-speed-of-trust/ [accessed 2023-08-10]
9. Federal Policy for the Protection of Human Subjects. Common Rule. Federal Register National Archives. URL: https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects [accessed 2024-07-18]
10. Chandrasekaran R, Katthula V, Moustakas E. Patterns of use and key predictors for the use of wearable health care devices by US adults: insights from a national survey. J Med Internet Res 2020 Oct 16;22(10):e22443 [FREE Full text] [doi: 10.2196/22443] [Medline: 33064083]

XSL•FO
RenderX

11. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 01;27(7):1147-1148 [FREE Full text] [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
12. Berryhill S, Morton CJ, Dean A, Berryhill A, Provencio-Dean N, Patel SI, et al. Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. J Clin Sleep Med 2020 May 15;16(5):775-783 [FREE Full text] [doi: 10.5664/jcsm.8356] [Medline: 32043961]
13. Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022 Aug;4(8):e615-e626 [FREE Full text] [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]
14. Mattison G, Canfell O, Forrester D, Dobbins C, Smith D, Töyräs J, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 01;24(7):e36690 [FREE Full text] [doi: 10.2196/36690] [Medline: 35776492]
15. Lee MA, Song M, Bessette H, Roberts Davis M, Tyner TE, Reid A. Use of wearables for monitoring cardiometabolic health: a systematic review. Int J Med Inform 2023 Nov;179:105218. [doi: 10.1016/j.ijmedinf.2023.105218] [Medline: 37806179]
16. Aggarwal R, Chiu N, Wadhera RK, Moran AE, Raber I, Shen C, et al. Racial/ethnic disparities in hypertension prevalence, awareness, treatment, and control in the United States, 2013 to 2018. Hypertension 2021 Dec;78(6):1719-1726 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.17570] [Medline: 34365809]
17. Rodríguez JE, Campbell KM. Racial and ethnic disparities in prevalence and care of patients with type 2 diabetes. Clin Diabetes 2017 Jan;35(1):66-70 [FREE Full text] [doi: 10.2337/cd15-0048] [Medline: 28144049]
18. Walter JR, Xu S, Rogers JA. From lab to life: how wearable devices can improve health equity. Nat Commun 2024 Jan 02;15(1):123. [doi: 10.1038/s41467-023-44634-9] [Medline: 38167483]
19. Yang Meier D, Barthelmess P, Sun W, Liberatore F. Wearable technology acceptance in health care based on national culture differences: cross-country analysis between Chinese and Swiss consumers. J Med Internet Res 2020 Oct 22;22(10):e18801 [FREE Full text] [doi: 10.2196/18801] [Medline: 33090108]
20. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45. [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
21. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health 2020 Feb;2(2):e85-e93 [FREE Full text] [doi: 10.1016/S2589-7500(19)30222-5] [Medline: 33334565]
22. Mason AE, Hecht FM, Davis SK, Natale JL, Hartogensis W, Damaso N, et al. Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict Study. Sci Rep 2022 Mar 02;12(1):3463. [doi: 10.1038/s41598-022-07314-0] [Medline: 35236896]
23. Ming DK, Sangkaew S, Chanh HQ, Nhat PTH, Yacoub S, Georgiou P, et al. Continuous physiological monitoring using wearable technology to inform individual management of infectious diseases, public health and outbreak responses. Int J Infect Dis 2020 Jul;96:648-654 [FREE Full text] [doi: 10.1016/j.ijid.2020.05.086] [Medline: 32497806]
24. Zinzuwadia A, Singh JP. Wearable devices-addressing bias and inequity. Lancet Digit Health 2022 Dec;4(12):e856-e857 [FREE Full text] [doi: 10.1016/S2589-7500(22)00194-7] [Medline: 36335031]
25. Canali S, Schiaffonati V, Aliverti A. Challenges and recommendations for wearable devices in digital health: data quality, interoperability, health equity, fairness. PLOS Digit Health 2022 Oct;1(10):e0000104 [FREE Full text] [doi: 10.1371/journal.pdig.0000104] [Medline: 36812619]

Abbreviations

OR: odds ratio

Edited by S Ma; submitted 19.01.24; peer-reviewed by C Sullivan, L Philpot; comments to author 14.03.24; revised version received 31.03.24; accepted 25.06.24; published 26.07.24.

Please cite as:
Nagappan A, Krasniansky A, Knowles M
Patterns of Ownership and Usage of Wearable Devices in the United States, 2020-2022: Survey Study
J Med Internet Res 2024;26:e56504
URL: https://www.jmir.org/2024/1/e56504
doi:10.2196/56504
PMID:

XSL•FO
RenderX

©Ashwini Nagappan, Adriana Krasniansky, Madelyn Knowles. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Investigating How the Use of Technology Can Reduce Missed Appointments: Quantitative Case Study at a General Practitioner Surgery

Teresa Sides[1,2*], BSc, MSc; Dhouha Kbaier[1*], MEng, PhD

[1]The Open University, Milton Keynes, United Kingdom
[2]Glyndwr University, Wrexham, United Kingdom
[*]all authors contributed equally

Corresponding Author:
Teresa Sides, BSc, MSc
The Open University
Walton Hall
Kents Hill
Milton Keynes, MK7 6AA
United Kingdom
Phone: 44 0300 303 5303
Email: Teresa.Sides@open.ac.uk

Abstract

Background: SMS texting systems have been considered a potential solution to reduce missed appointments in primary care. Existing research in this area focuses on qualitative studies investigating the attitudes of SMS text users and receivers.

Objective: This study aimed to examine appointment data from an independent general practitioner (GP) surgery in Wrexham, United Kingdom, with approximately 15,000 patients, to determine the impact of text messaging systems on reducing missed appointments. The objective of this study was to investigate whether the use of text messages can effectively reduce missed appointments.

Methods: To collect data for the study, SQL reports were run on EMIS Web, the United Kingdom's most widely used clinical system. The data spanned 10 years, from September 1, 2010, to March 31, 2020. Data accuracy was verified by cross-referencing with appointment diary records. Mann-Whitney and Kruskal-Wallis tests, chosen for their suitability in comparing groups in nonparametric settings, were conducted in Microsoft Excel due to its accessibility.

Results: Statistical analyses were conducted to compare data before and after implementation of the text messaging system. The results revealed a significant 42.8% reduction in missed appointments (before: 5848; after: 3343; $P<.001$). Further analysis of demographic characteristics revealed interesting trends, with no significant difference in missed appointments between genders, and variations observed across different age groups. The median number of missed appointments was not significantly different between genders (women: 1.55, IQR 1.11-2.16; men: 1.61, IQR 1.08-2.12; $P=.73$). Despite the prevalence of mobile phone use among young adults aged 20-25 years, the highest rates of missed appointments (848/7256, 11.7%) were noted in this group, whereas the lowest rates were noted in the 75-80 years age group (377/7256; 5.2%; $P<.001$). Analysis by age and gender indicated inconsistencies: women aged 20-25 years (571/4216) and men aged 35-40 years (306/3040) had the highest rates of missed appointments, whereas women aged 70-75 years (177/4216) and men aged 75-80 years (129/3040) had the lowest rates ($P<.001$ for both).

Conclusions: This study demonstrates that SMS text messaging in primary care can significantly reduce missed appointments. Implementing technology such as SMS text messaging systems enables patients to cancel appointments on time, leading to improved efficiency in primary care settings.

(J Med Internet Res 2024;26:e43894) doi:10.2196/43894

KEYWORDS

National Health Service; primary care; SMS text messaging; SMS reminders; missed appointments; quantitative research; Kruskal-Wallis test; Mann-Whitney test

XSL•FO
RenderX

Introduction

In recent years, primary care has adopted SMS text messaging systems for patient communication, yet their efficacy in achieving positive outcomes remains unknown. For general practitioner (GP) practices, the National Health Service (NHS) sets annual targets that determine their funding. Since 2019, in addition to clinical outcomes, there has been a shift toward enhancing accessibility and appointment availability [1].

The widespread use of mobile phones in daily lives across all age groups [2] has facilitated the development of cost-effective, bulk text messaging systems for businesses [3]. Text messages are universally compatible and accessible across various network speeds, and do not require upgrades [4]. While bulk text messaging has been widely explored in business, its applications and potential impact within health care remain underexplored, which is the focus of this study.

Numerous scholarly articles have delved into the perspectives of patients about text message reminders [5-7]. Sherman et al [8] found that while text messages were deemed useful, patients wanted the option to reply to the messages. Maslakpak and Safaie [9] compared treatment compliance between groups receiving text messages, reminder cards, or no reminders; they found no significant difference between reminder cards and text messages.

Hirst et al [10] reported that patients received a postal reminder with or without a text reminder (generated by the iPlato texting system [11] integrated with EMIS Web [12]). They reported a slight improvement in screening uptake from 39.9% to 40.5%, but the difference was not statistically significant.

In a study involving 18,138 participants, Fisher et al [13] evaluated the effectiveness and utility of text messages in health care, noting that text messages increased appointment attendance and patient satisfaction. However, nonrandom sampling may compromise validity. Since participants volunteered, their willingness to engage might influence appointment attendance, potentially introducing bias to the results.

Research on GPs using text messaging systems primarily investigates patient preferences for receiving texts and doctors' willingness to send them [14]. The literature predominantly discusses the advantages and disadvantages of such systems from the GP standpoint, rather than assessing their impact on patient appointment attendance. Trials evaluating text reminder systems have yielded mixed results, with no significant overall improvement in attendance rates [5,6]. Even trials with substantial cohorts exhibited only a modest 0.6% attendance increase, suggesting limited benefits of using text reminders [13].

This study adds to the expanding knowledge base in the field and offers insights into implementing effective appointment management strategies in primary care.

Methods

Overview

This study analyzed missed appointment data from an independent GP practice in Wrexham, United Kingdom, spanning a decade (September 1, 2010, to March 31, 2020). Data were gathered from EMIS web, and the iPlato texting system was implemented on September 15, 2015. iPlato messages contained standard appointment information and allowed patients to respond or cancel the appointment. Before iPlato, patients often relied on appointment cards or memory. Approximately 74% of appointments were associated with a mobile number.

Data included adults aged 18-79 years, excluding children due to parental responsibility for missed appointments and older adults aged 80 years and above due to frailty or care home residence. Nurse data were not integrated into the system until September 2015, so only missed GP appointments were analyzed.

Missed appointments data over a decade were gathered using SQL reports to compare 5 years before and after text reminders were implemented. The collected figures were cross-referenced with appointment diary records to ensure accuracy.

Mann-Whitney and Kruskal-Wallis tests were used for quantitative analysis due to their suitability for nonparametric comparisons [15,16]. Mann-Whitney test, known for its efficiency and robustness [17], was used to assess gender categories, and Kruskal-Wallis test was applied across various age groups due to its suitability for comparing more than 2 groups.

Ethics Approval

The Open University Human Research Ethics Committee approved this project (HREC/4180/Sides).

Results

This study aimed to analyze missed appointment frequencies and demographic characteristics of patients who missed appointments. Of the 572,794 appointments booked, 9191 were missed, with 5286 (57.5%) attributed to women and 3905 (42.5%) to men (Table 1).

Mann-Whitney test results (Table 2) indicated higher rates of missed appointments before text messaging was implemented compared to after implementation, suggesting that text messaging can be an effective strategy for reducing missed appointments, particularly among women.

To assess the data more accurately, missed appointment rates were calculated for each age group (Figure 1). Significant differences in median missed appointment numbers across different age groups were observed (Table 3). The 20-25 years age group had the highest median number of missed appointments, whereas the 65-80 years age group had the lowest. Further information and statistical data are available in Multimedia Appendix 1.

XSL•FO
RenderX

Data analysis showed that text message reminders reduced missed appointments at the GP practice. Missed appointment frequency decreased before text messaging was introduced in September 2015 (Figure 2). Then, shortly after text messaging implementation, it increased, followed by a further decline to the lowest level noted in 10 years. Several factors may have contributed to this trend, including increased attention and improved data recording practices. Despite this factor, texting resulted in a statistically significant decrease in nonattendance at appointments (Table 1). These findings demonstrate the effectiveness of text messaging in reducing missed appointments.

Table 1. Population and gender descriptive statistics before and after text messaging implementation (N=9191 missed appointments).

Study Population	Missed appointments, n (%)		Reduction in missed appointments (%)
	Before text messaging implementation	After text messaging implementation	
Total population	5848 (63.6)	3343 (36.3)	42.8
Women	3435 (37.4)	1851 (20.1)	46.1
Men	2413 (26.3)	1492 (16.2)	38.2

Table 2. Mann-Whitney test results (N=9191 missed appointments; 5286 by women and 3905 by men).

Patient population and frequency of missed appointments	Before text messaging implementation, median (IQR)	After text messaging implementation, median (IQR)	U^a	*z* score	*P* value
Total population					
Quarterly	1.8 (1.4-2.5)	1.1 (0.9-1.6)	74	3.1	.002
Monthly	1.9 (1.5-2.7)	1.2 (0.8-1.6)	719	5.2	<.001
Women					
Quarterly	1.9 (1.5-2.3)	1.1 (0.8-1.5)	72	3.2	.002
Monthly	1.9 (1.5-2.5)	1.2 (0.7-1.6)	694	5.3	<.001
Men					
Quarterly	1.8 (1.4-2.1)	1.1 (1.0-1.5)	80	2.9	.003
Monthly	1.9 (1.5-2.5)	1.3 (1.0-1.7)	833	4.6	<.001

[a]Mann-Whitney U (Wilcoxon) statistic.

Figure 1. Missed appointment frequency across 5-year age groups.

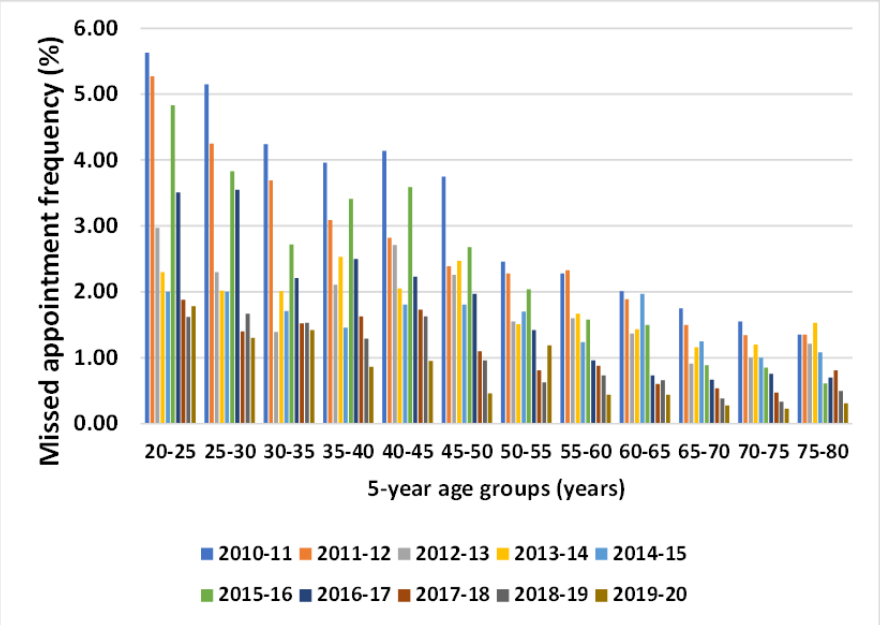

XSL•FO
RenderX

Table 3. Statistical results for missed appointments across 5-year age ranges (N=7256 missed appointments; 4216 by women and 3040 by men).

Study population and age range (years)	Patients, n (%)	*r*	Mean (SD)	s^2	Median (IQR)
Total population					
20-25	848 (11.7)	1.0	3.2 (1.5)	2.4	2.6 (1.9-4.5)
25-30	832 (11.5)	0.9	2.7 (1.3)	1.8	2.2 (1.8-3.8)
30-35	665 (9.2)	0.7	2.2 (1)	1.0	1.9 (1.5-2.6)
35-40	628 (8.7)	1.0	2.3 (1)	1.0	2.3 (1.5-3)
40-45	734 (10.1)	0.9	2.4 (0.9)	0.9	2.1 (1.8-2.8)
45-50	690 (9.5)	1.0	2.0 (0.9)	0.9	2.1 (1.3-2.5)
50-55	603 (8.3)	0.9	1.6 (0.6)	0.4	1.5 (1.3-2)
55-60	564 (7.8)	0.7	1.4 (0.6)	0.4	1.4 (0.9-1.7)
60-65	554 (7.6)	0.6	1.3 (0.6)	0.4	1.4 (0.7-1.8)
65-70	403 (5.6)	0.9	0.9 (0.5)	0.2	0.9 (0.6-1.2)
70-75	358 (4.9)	0.9	0.9 (0.4)	0.2	0.9 (0.5-1.2)
75-80	377 (5.2)	0.1	0.9 (0.4)	0.2	0.9 (0.6-1.3)
Women					
20-25	571 (13.5)	1.0	3.3 (1.6)	2.9	2.8 (2-4.9)
25-30	570 (13.5)	0.9	2.7 (1.3)	2	2.3 (1.8-3.6)
30-35	416 (9.9)	0.8	2.2 (0.9)	0.9	1.9 (1.6-2.4)
35-40	322 (7.6)	0.8	2 (0.9)	0.8	1.8 (1.4-2.5)
40-45	401 (9.5)	0.9	2.5 (1.5)	2.5	2 (1.4-3.5)
45-50	406 (9.6)	0.7	2 (0.9)	1	2.3 (1.5-2.5)
50-55	334 (7.9)	0.8	1.5 (0.6)	0.4	1.5 (1-2)
55-60	300 (7.1)	0.8	1.2 (0.6)	0.4	1.3 (0.7-1.7)
60-65	258 (6.1)	0.3	1 (0.5)	0.3	1.1 (0.7-1.3)
65-70	213 (5.1)	0.9	0.9 (0.5)	0.3	0.9 (0.4-1.2)
70-75	177 (4.2)	0.9	0.8 (0.4)	0.2	0.9 (0.5-1)
75-80	248 (5.8)	0.2	1.3 (1.2)	1.5	1.1 (0.7-1.3)
Men					
20-25	277 (9.1)	0.7	3 (1.3)	1.9	2.5 (2-3.8)
25-30	262 (8.6)	0.9	2.7 (1.3)	1.8	2.4 (1.7-3.7)
30-35	249 (8.2)	0.5	2.3 (1.1)	1.2	1.7 (1.5-2.9)
35-40	306 (10.1)	0.9	2.8 (1.2)	1.7	2.6 (1.5-3.8)
40-45	333 (11)	0.5	2.5 (0.7)	0.6	2.4 (2.1-2.7)
45-50	284 (9.3)	0.8	1.9 (1)	1.1	1.9 (1.2-2.5)
50-55	269 (8.8)	0.9	1.8 (0.6)	0.4	1.8 (1.4-2.1)
55-60	264 (8.7)	0.7	1.5 (0.7)	0.5	1.4 (1-2.1)
60-65	296 (9.7)	0.9	1.4 (0.7)	0.5	1.7 (0.7-1.9)
65-70	190 (6.3)	0.9	0.9 (0.4)	0.1	0.8 (0.7-1)
70-75	181 (6)	0.9	0.9 (0.5)	0.3	0.9 (0.6-1.2)
75-80	129 (4.2)	0.4	0.8 (0.3)	0.1	0.8 (0.6-1.1)

XSL•FO
RenderX

Figure 2. Missed appointment frequency by gender (with trend lines) before and after implementation of the SMS text messaging system.

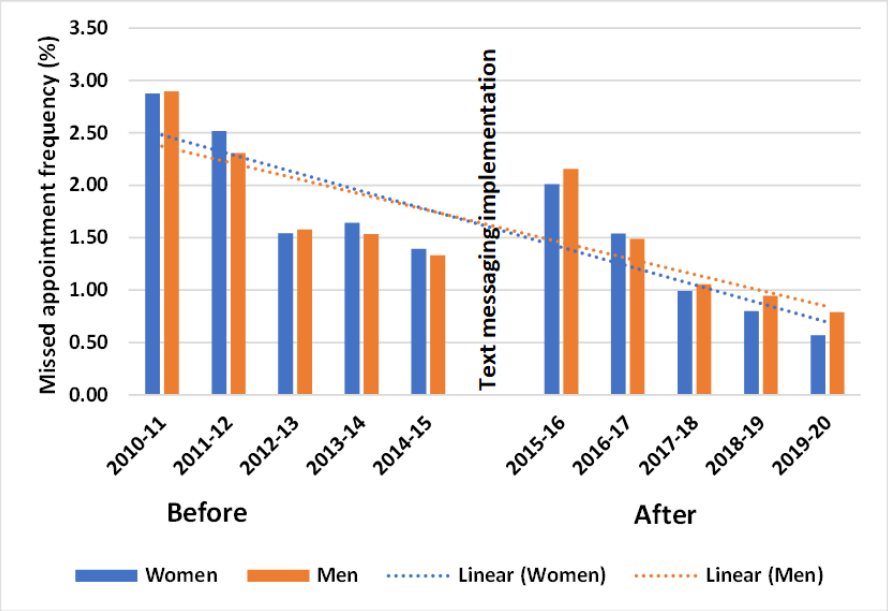

Discussion

This study explores the impact of SMS text messaging systems on reducing missed appointments in primary care, focusing on a single GP practice. While a significant 42.8% decrease in missed appointments was noted after implementation of text messaging reminders, caution is warranted in generalizing these results to broader health care settings. This 42.8% reduction is significant when compared with the findings of Maslakpak and Safaie [9] who found no significant difference.

Age group analysis revealed interesting findings regarding missed appointments. Surprisingly, the 20-25 age group emerged as significant overall, with a higher percentage of missed appointments, contrary to expectations given their high mobile phone use. De-Sola et al [18] observed increased phone use for the 16-25 years age group. However, despite prolific phone usage, the 20-25 years age group patients did not exhibit reduced missed appointments. These findings highlight the positive impact of text reminder systems on missed appointments overall, yet further exploration is needed to understand factors influencing attendance in specific demographics.

In conclusion, text messaging reminders can effectively reduce missed appointments, enabling more GP consultations. However, it is unlikely that missed appointments can be completely avoided. Health care technology offers opportunities for further improvement, including timely text reminders and systematic patient data maintenance. By embracing technology in health care, we can overcome challenges, maximize appointment availability, and enhance patient care.

Acknowledgments

We thank Prof Dr Vic Grout and Bindu Jose (BEng) for their valuable suggestions.

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study data and Mann-Whitney and Kruskal-Wallis test results.
[XLSX File (Microsoft Excel File), 205 KB - jmir_v26i1e43894_app1.xlsx]

References

1. NHS. NHS Long Term Plan. 2019. URL: https://www.longtermplan.nhs.uk [accessed 2022-05-22]
2. Plateauing at the peak. Deloitte. URL: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology-media-telecommunications/deloitte-uk-plateauing-at-the-peak-the-state-of-the-smartphone.pdf [accessed 2024-07-15]

XSL•FO
RenderX

3. Cheng C, Brown CR, Venugopalan J, Wang MD. Towards an effective patient health engagement system using cloud-based text messaging technology. IEEE J Transl Eng Health Med 2020;8:1-7. [doi: 10.1109/jtehm.2018.2868358]
4. Siang BK, Ramli AR, Prakash V, Bin Syed Mohamed SAR. SMS gateway interface remote monitoring and controlling via GSM SMS. In: NCTT 2003 Proceedings. 2003 Presented at: 4th National Conference of Telecommunication Technology; Jan 14-15, 2023; Shah Alam, Malaysia p. 84-87. [doi: 10.1109/nctt.2003.1188308]
5. Liu JX, Modrek S. Evaluation of SMS reminder messages for altering treatment adherence and health seeking perceptions among malaria care-seekers in Nigeria. Health Policy Plan 2016 Dec 16;31(10):1374-1383 [FREE Full text] [doi: 10.1093/heapol/czw076] [Medline: 27315831]
6. Domek GJ, Contreras-Roldan IL, O'Leary ST, Bull S, Furniss A, Kempe A, et al. SMS text message reminders to improve infant vaccination coverage in Guatemala: a pilot randomized controlled trial. Vaccine 2016 May 05;34(21):2437-2443 [FREE Full text] [doi: 10.1016/j.vaccine.2016.03.065] [Medline: 27026145]
7. Owolabi EO, Goon DT, Ajayi AI. Impact of mobile phone text messaging intervention on adherence among patients with diabetes in a rural setting: A randomized controlled trial. Medicine (Baltimore) 2020 Mar;99(12):e18953 [FREE Full text] [doi: 10.1097/MD.0000000000018953] [Medline: 32195927]
8. Sherman E, Clauson KA, Elrod S, Eckardt P, Alkhateeb F, Jacobs R. Patient perceptions of text messaging to improve antiretroviral therapy adherence: a qualitative study of patients in a Ryan White Clinic. J Int Assoc Provid AIDS Care 2018;17:2325958218759209 [FREE Full text] [doi: 10.1177/2325958218759209] [Medline: 29473483]
9. Maslakpak MH, Safaie M. A comparison between the effectiveness of short message service and reminder cards regarding medication adherence in patients with hypertension: a randomized controlled clinical trial. Int J Community Based Nurs Midwifery 2016 Jul;4(3):209-218 [FREE Full text] [Medline: 27382587]
10. Hirst Y, Kerrison R, Kobayashi LC, Counsell N, Djedovic N, Ruwende J, et al. Text reminders in colorectal cancer screening (TRICCS): protocol for a randomised controlled trial. BMC Public Health 2016 Jan 25;16(1):74 [FREE Full text] [doi: 10.1186/s12889-016-2733-6] [Medline: 26809344]
11. iPlato. URL: https://www.iplato.com/ [accessed 2024-07-15]
12. EMIS Web. URL: https://www.emishealth.com/products/emis-web [accessed 2024-07-15]
13. Fischer HH, Moore SL, Johnson TL, Everhart RM, Batal H, Davidsoni AJ. Appointment reminders by text message in a safety net health care system: a pragmatic investigation. EGEMS (Wash DC) 2017 Sep 04;5(1):20 [FREE Full text] [doi: 10.5334/egems.215] [Medline: 29881740]
14. Leahy D, Lyons A, Dahm M, Quinlan D, Bradley C. Use of text messaging in general practice: a mixed methods investigation on GPs' and patients' views. Br J Gen Pract 2017 Sep 25;67(664):e744-e750. [doi: 10.3399/bjgp17x693065]
15. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947 Mar;18(1):50-60. [doi: 10.1214/aoms/1177730491]
16. Fay MP, Malinovsky Y. Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. Stat Med 2018 Nov 30;37(27):3991-4006 [FREE Full text] [doi: 10.1002/sim.7890] [Medline: 29984411]
17. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952 Dec;47(260):583-621. [doi: 10.1080/01621459.1952.10483441]
18. de-Sola J, Talledo H, Rodríguez de Fonseca F, Rubio G. Prevalence of problematic cell phone use in an adult population in Spain as assessed by the Mobile Phone Problem Use Scale (MPPUS). PLoS One 2017 Aug 3;12(8):e0181184 [FREE Full text] [doi: 10.1371/journal.pone.0181184] [Medline: 28771626]

Abbreviations

GP: general practitioner
NHS: National Health Service

Edited by T Leung, A Mavragani; submitted 28.10.22; peer-reviewed by M Blanker, A Fosse, A Singleton; comments to author 09.03.23; revised version received 23.06.23; accepted 30.05.24; published 29.07.24.

Please cite as:
Sides T, Kbaier D
Investigating How the Use of Technology Can Reduce Missed Appointments: Quantitative Case Study at a General Practitioner Surgery
J Med Internet Res 2024;26:e43894
URL: https://www.jmir.org/2024/1/e43894
doi:10.2196/43894
PMID:39073855

XSL•FO
RenderX

©Teresa Sides, Dhouha Kbaier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceptions of HIV-Related Comorbidities and Usability of a Virtual Environment for Cardiovascular Disease Prevention Education in Sexual Minority Men With HIV: Formative Phases of a Pilot Randomized Controlled Trial

S Raquel Ramos[1,2,3*], RN, MSN, MBA, PhD; Harmony Reynolds[4*], MD; Constance Johnson[5,6*], RN, MS, PhD; Gail Melkus[7*], EDD, ANP; Trace Kershaw[2,3*], PhD; Julian F Thayer[8*], MA, PhD; Allison Vorderstrasse[9*], APRN, DNSc

[1]School of Nursing, Yale University, Orange, CT, United States

[2]School of Public Health, Social and Behavioral Sciences, Yale University, New Haven, CT, United States

[3]Center for Interdisciplinary Research on AIDS, Yale University, New Haven, NY, United States

[4]Cardiovascular Clinical Research Center, Leon H. Charney Division of Cardiology, Department of Medicine, NYU Grossman School of Medicine, New York, NY, United States

[5]Czik School of Nursing, The University of Texas Health Science Center at Houston, Houston, TX, United States

[6]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

[7]Rory Myers College of Nursing, New York University, New York, NY, United States

[8]School of Social Ecology, Psychological Science, University of California, Irvine, CA, United States

[9]Elaine Marieb College of Nursing, University of Massachusetts Amherst, Amherst, MA, United States

[*]all authors contributed equally

Corresponding Author:
S Raquel Ramos, RN, MSN, MBA, PhD
School of Nursing
Yale University
400 West Campus Drive
Orange, CT, 06477
United States
Phone: 1 2037372339
Email: raquel.ramos@yale.edu

Abstract

Background: Sexual minority men with HIV are at an increased risk of cardiovascular disease (CVD) and have been underrepresented in behavioral research and clinical trials.

Objective: This study aims to explore perceptions of HIV-related comorbidities and assess the interest in and usability of a virtual environment for CVD prevention education in Black and Latinx sexual minority men with HIV.

Methods: This is a 3-phase pilot behavioral randomized controlled trial. We report on formative phases 1 and 2 that informed virtual environment content and features using qualitative interviews, usability testing, and beta testing with a total of 25 individuals. In phase 1, a total of 15 participants completed interviews exploring HIV-related illnesses of concern that would be used to tailor the virtual environment. In phase 2, usability testing and beta testing were conducted with 10 participants to assess interest, features, and content.

Results: In phase 1, we found that CVD risk factors included high blood pressure, myocardial infarction, stroke, and diabetes. Cancer (prostate, colon, and others) was a common concern, as were mental health conditions. In phase 2, all participants completed the 12-item usability checklist with favorable feedback within 30 to 60 minutes. Beta-testing interviews suggested (1) mixed perceptions of health and HIV, (2) high risk for comorbid conditions, (3) virtual environment features were promising, and (4) the need for diverse avatar representations.

Conclusions: We identified several comorbid conditions of concern, and findings carry significant implications for mitigating barriers to preventive health screenings, given the shared risk factors between HIV and related comorbidities. Highly rated aspects of the virtual environment were anonymity; meeting others with HIV who identify as gay or bisexual; validating lesbian, gay,

XSL•FO
RenderX

bisexual, transgender, queer, and others (LGBTQ+) images and content; and accessibility to CVD prevention education. Critical end-user feedback from beta testing suggested more options for avatar customization in skin, hair, and body representation. Our next phase will test the virtual environment as a new approach to advancing cardiovascular health equity in ethnic and racial sexual minority men with HIV.

Trial Registration: ClinicalTrials.gov NCT04061915; https://clinicaltrials.gov/study/NCT05242952

International Registered Report Identifier (IRRID): RR2-10.2196/38348

(J Med Internet Res 2024;26:e57351) doi:10.2196/57351

KEYWORDS

virtual environment; digital health; gamification; eHealth; sexual minorities; cardiovascular disease; HIV; cardiometabolic risk; mental health; lesbian, gay, bisexual, transgender, and queer; LGBTQ health; HIV care; prevention; virtual; minority; men; Latin; Black; men who have sex with men; intervention; high blood pressure; myocardial infarction; preventive health screenings; gay; bisexual; patients; cancer

Introduction

Background

In the United States, cardiovascular disease (CVD) has been the leading contributor to mortality for >100 years [1]. Heart disease risk is especially prevalent among persons with HIV [2-4]. Antiretroviral therapy has increased life expectancy, permitting the link between HIV and premature-onset CVD to come to light. Persons with HIV are 1.5 to 2 times more likely to develop CVD than persons without HIV [4]. This increase in CVD risk is similar to the increase in CVD risk due to diabetes [4,5]. It is estimated that 78% of persons with HIV will have CVD by the year 2030 [6,7].

The risk of HIV, as well as CVD as a comorbidity, disproportionately impacts Black and Latinx sexual minority men [8], as they experience higher rates of HIV-related conditions, such as high blood pressure and diabetes [9]. CVD risk is intersectional and exacerbated by social determinants and structural forces. When these multilevel stressors coalesce, they create barriers to education, safe neighborhoods, resources, and quality health care for those who are economically, racially, and ethnically marginalized. In sexual minoritized individuals, this is intensified and manifests in the forms of repeated experiences of racism, discrimination, homophobia, and violence [10,11]. CVD risk is heightened by racism and discrimination [10,11] and their persistent effects on allostatic load [12] and weathering [13] (the ongoing, cumulative burden of stress that predisposes individuals to morbidity and mortality). Allostatic load has adverse impacts on the immune, cardiovascular, and metabolic systems over time [14], which can increase stress responses and activate the hypothalamic-pituitary axis. This creates an increase in cortisol and other inflammatory markers, resulting in physiological disruptions [15], leading to cardiometabolic and mental health morbidities [16]. Thayer et al [10] and Hill and Thayer [17] identified a “cardiovascular conundrum,” in which perpetual discrimination in Black and sexual minoritized individuals leads to both elevated systemic vascular resistance, which can heighten CVD morbidity and mortality risk, and higher vagally mediated heart rate variability (HRV), typically associated with lower risk. Exposure to discrimination and racism lowers HRV and may elicit maladaptive coping mechanisms, such as rumination, which further lowers HRV [17-19]. Racism and discrimination toward racial, ethnic, and sexual minoritized individuals have resulted in longstanding disadvantages in achieving health equity. These are root causes of medical mistrust, are catalysts for avoiding care engagement [20], and reduce opportunities for CVD screening and preventive counseling [21]. A review examining nonpharmacologic behavioral interventions to prevent CVD in persons with HIV suggested that over the past decade, there has been an increase in behavioral CVD interventions [22]. However, the extant literature is limited in its representation of populations that continue to be disproportionately affected by low cardiovascular health [22]. Current research also identified the interplay among an individual’s race, gender, and sexual orientation [23] that negatively affects health outcomes and further increases the risk of chronic conditions [24], creating an urgent need to mitigate health risks in disadvantaged, underserved, and underresourced populations. Clinical research on behavioral and lifestyle interventions to prevent heart disease that are culturally salient and tailored is essential to mitigating CVD disparities in minoritized populations. The standard approach to prevention education has been conventional patient teaching within the clinical setting. However, the literature confirms the existence of an equity gap in who has access to timely [25], quality care [26] and who has access to participation in clinical trials [27,28]. The health care and research enterprise has rapidly shifted toward ubiquitous computing as a common source of data collection for research (eg, electronic health records, artificial intelligence, wearables, mobile health, data processing via virtual environments, and augmented reality) [29-34]. Behavioral research using virtual environments and augmented reality has been gaining popularity [35-38]. However, the level of accessibility to this type of research has remained limited in diverse and marginalized populations. The novelty and relevance of technology-driven behavioral research have not been fully realized in low-income and urban communities where large numbers of minoritized individuals reside.

This Study

The American Heart Association has called for an increase in interventions that focus on enhancing cardiovascular health in underserved communities by addressing social determinants and modifiable risk factors of CVD [39]. Specifically, the American Heart Association has called for more cardiovascular health interventions to be taken from the clinical setting and

XSL•FO
RenderX

implemented into the community and community-serving programs. Community-based interventions are sought for their value in engaging individuals and families from disadvantaged or marginalized groups [40]. This provides an exceptional opportunity to bridge the equity gap that currently exists—the reason why clinical interventions do not reach marginalized individuals at increased risk. To mitigate longstanding barriers to health equity and advance understanding about the social determinates that impede cardiovascular health and perceptions of other HIV-related comorbidities of concern, qualitative, community-engaged approaches are needed. This study aimed to explore perceptions of HIV-related comorbidities and assess the usability and functionality of a virtual environment for CVD prevention education in Black and Latinx sexual minority men with HIV.

Methods

Study Design

This study is a 3-phase pilot behavioral randomized controlled trial. The purpose of the intervention is to test the acceptability and feasibility of a virtual environment for CVD prevention education in Black and Latinx sexual minority men with HIV using the Diabetes Learning in a Virtual Environment (LIVE) platform [35,37]. The protocol for the study is reported elsewhere [33] and is registered on Clinicaltrials.gov (NCT04061915). Formative phases 1 and 2 informed virtual environment content and features using qualitative interviews, usability testing, and beta testing with a total of 25 individuals. This paper reports on phases 1 and 2 that were necessary to inform the phase-3 clinical trial.

Participants

We wanted to better understand chronic illness health needs from the perspectives of both HIV health experts and individuals with HIV. We partnered with 2 community-based organizations that provide supportive services to ethnic, racial, immigrant, and sexual minoritized individuals. The organizations were located in New York City and served individuals who were economically marginalized. We refer to “clients” of the community-based organizations as study “participants.”

We engaged several experts in HIV care with clinical and community health backgrounds. Eligibility criteria for the HIV care experts were as follows: (1) engagement in health care or care services with persons living with HIV for at least 5 years, (2) any gender, and (3) age >18 years.

Client participants were all adults and identified as being of either Latinx or Black heritage. Eligibility criteria for participants were as follows: (1) self-identification as gay or bisexual, (2) HIV serostatus positive, (3) ages 30 to 65 years, and (4) access to a laptop or desktop. Participants were ineligible if they (1) identified as heterosexual; (2) were HIV negative; or (3) had a history of myocardial infarction, stroke, or cognitive impairment.

Recruitment

Before recruitment, we convened with administrators at the community-based organizations to discuss study objectives and approaches to data collection, describe the types of data that would be collected, and outline logistics for obtaining the data in ways that would be comfortable for the participants. Program managers and administrators from the community-serving organizations posted flyers and used word of mouth to provide information about the study. We worked with the community-serving organizations to contact individuals who conveyed interest in learning more about the study. Using Zoom (Zoom Video Communications), interested persons were screened, provided consent (if eligible), and were enrolled.

Ethical Considerations

This study was approved by the Yale University Institutional Review Board (2000031403) in 2022, the New York University Institutional Review Board (IRB-FY2018-2284) in 2019, and the University of Texas Health Science Center at Houston (HSC-SN-20-1143) in 2020. All procedures performed in this study involving human participants are in accordance with the ethical standards of the institutional and national research committees, the 1964 Helsinki Declaration and its later amendments, or comparable ethical standards. All study data were deidentified to protect the identity of participants. Verbal informed consent was obtained from all individuals who participated in the study. We provided a US $30 gift card as compensation for participation.

Study Procedures

Overview

Phases 1 and 2 were conducted using Zoom. Zoom is a Health Insurance Portability and Accountability Act–compliant, web-based videoconferencing platform that uses end-to-end encryption. During the interviews and focus groups, cameras were turned off. Only audio recordings using pseudonyms were used during data collection.

Phase 1: Qualitative Interviews

Phase-1 data were collected from June 2021 to May 2022 with a total of 15 participants. The purpose of phase 1 was to learn about health concerns that were important to participants and use this information to tailor the content and design of the virtual environment, which would be piloted in phase 3. Our interview questions were derived from an open-ended interview guide. The interview guide included 5 questions with probes for clarity. An example question is as follows: “Tell me about any medical conditions, other than HIV, that you are concerned about?” Example questions from the guide were reviewed with the program managers at the community-serving organizations to ensure that the language and content were appropriate for their clients. To ensure participant accessibility and convenience, semistructured interviews and focus groups (for participants who preferred to have an in-person, group interview at their community-serving organization) were conducted. Before starting the focus groups, ground rules were established for respect and the confidentiality of the information to be shared. The first author (SRR), an academic researcher, had a preexisting relationship with the community-based organizations and facilitated qualitative data collection. For both semistructured interviews and focus groups, pseudonyms were created before the interviews to maintain anonymity. After the

XSL•FO
RenderX

interviews were completed, we used the data to embed the virtual environment with content that was reported as important, relevant, and culturally salient.

Phase 2: Usability and Beta Testing

After we had embedded the virtual environment with content from phase 1, we conducted usability and beta testing. Data for phase 2 were collected from August 2022 to October 2022 with a total of 10 participants. A 12-item usability checklist (Textbox 1) was developed and pilot-tested by members of the study team. The purpose of the usability checklist was to conduct a task analysis to assess the practical functionality of the virtual environment. Example checklist items include the following: (1) "customize your avatars hair, skin, clothes, and shoes" and (2) "use the microphone to say hello to others in the virtual environment." Participants were tasked with moving through all 12 items, which would convey they were able to navigate all areas, engage with features, and customize their avatars. After usability testing, we conducted beta testing, which comprised several questions to assess: (1) avatar customization, (2) functionality and features, (3) quest content, and (4) educational health content. We also asked for any additional feedback for tailoring the virtual environment content before the initiation of the clinical trial. In phase 3, we will test the new content that is embedded and assess outcomes on usability and acceptability.

Textbox 1. Virtual environment features that will be accessible to individuals enrolled in the clinical trial.

Features and key activity examples

- Avatar
  - Customize avatar
  - Engage anonymously with other participants
- Quests
  - Build knowledge about chronic illness prevention
  - Test knowledge
  - Earn tickets to customize avatar
- Districts
  - Lobby (pharmacy and bookstore 1)
  - Plaza (fitness center, clothing store, and bookstore 2)
  - Food court (restaurant, grocery, and farmers market)
  - Reflection garden (tranquil area)
- Nurse-led health sessions (not part of phases 1 and 2)
  - Virtual or video presentations focusing on heart health
- Technical support (not part of phases 1 and 2)
  - Instructional guidebook on virtual environment use
  - Videos to enhance the use of the virtual environment
  - Web-based newsletter

Data Analysis

Qualitative data were analyzed using content analysis in NVivo (version 12; Lumivero). Content analysis is an established set of stepwise techniques used in qualitative research to identify, code, and describe themes and then assign themes to open-ended narratives. In qualitative research, data are contextual and focused on maximizing the understanding of the phenomenon under study and not necessarily on a fixed sample size. A total of 3 coders analyzed the data (SRR, Lauren Chin, and Maame-Owusua Boateng), and discrepancies were resolved through discussion until a consensus was reached. Data saturation was achieved after 15 phase-1 participants were interviewed, and there were no new emerging data [41,42]. In phase-2 usability testing, metrics of content navigation and the duration of time to completion were tracked using the 12-item checklist. Beta-testing qualitative data were pooled, and common themes were assigned thematic categories to contextualize the findings using representative quotes.

Measures of Rigor for Qualitative Data

Reliability was demonstrated by providing a detailed description and rationale for the study purpose, participant sampling approach, and role of the researchers and study team [43]. Dependability was ensured through using consistent approaches to data collection, such as the use of a prespecified interview guide [44]. Validity was assessed by checking for the accuracy or credibility of the findings through member checking and peer

XSL•FO
RenderX

debriefing [43,45]. Debriefings were held with the study team following data collection to strengthen the validity of data interpretations and not leave them subject to recall [43]. Transferability [44] of our study findings may be applicable to diverse and marginalized populations who are at a risk of comorbid conditions. Standards for Reporting Qualitative Research was used to report findings [46].

Results

Sample Characteristics

We collected qualitative data about health concerns outside of HIV from 11 participants who are clients at the community-based organizations. Their characteristics are presented in Table 1.

On average, participants reported having been diagnosed with HIV for 5.8 (SD 6.62) years with the range of years spanning from 1 to 21 years. The study participants with HIV reflect the demographic characteristics of individuals who have been traditionally underrepresented in clinical research and disproportionately face significant health disparities.

Of the 15 participants, 4 (27%) were HIV care experts with clinical and community health backgrounds: 1 holistic advanced practice nurse practitioner with a master's degree and 5 years of experience, 1 access-to-care specialist with a bachelor's degree and 5 years of experience, 1 community health educator with a master's degree and 15 years of experience, and 1 HIV senior program director with a bachelor's degree and 6 years of experience. All 4 HIV care experts identified as being from an ethnic or racial minoritized background. These individuals were highly trusted by participants. Their engagement during phases 1 and 2 facilitated open lines of communication from our participants with HIV, which added to the quality of the data.

The phase-2 sample consisted of 10 participants, which aligned with published recommendations for conducting user testing [47,48]. Overall, 45% (5/11) of client participants with HIV who participated in phase-1 qualitative interviews also participated in phase 2. Educational attainment ranged from 11th grade to a bachelor's degree.

In phase 2, we enrolled 5 HIV care experts who had backgrounds in medicine and community health. There was an overlap where 1 HIV care expert also participated in the phase-1 qualitative interview. HIV care experts had completed a college level education ranging from an associate's degree to a Doctor of Medicine degree.

Table 1. Phase 1 qualitative interview participant characteristics (n=11)[a].

Characteristics	Values
Race and ethnicity, n (%)	
African American (non-Hispanic or Latinx)	5 (45)
Latinx	4 (36)
Mixed race ethnicity or biracial	2 (18)
Age (y), mean (SD)	32.27 (3.04)
Highest level of education, n (%)	
Master's degree	__[b]
Bachelor's degree	2 (18)
Some college or associate's degree	2 (18)
HS[c] diploma or GED[d]	7 (64)

[a]HIV care experts (n=4) were not included in the table due to the small sample size. Their characteristics are narrated in the *Results* section.

[b]Not applicable.

[c]HS: high school.

[d]GED: General Educational Development.

Phase 1: Focus Group and Semistructured Interviews

Overview

Qualitative data collection for each interview took 60 to 90 minutes. A total of 9 semistructured interviews and 1 focus group were conducted with 6 participants. Four (67%) of the 6 participants were HIV care experts with clinical and community health backgrounds, and 2 (33%) were client participants with HIV who were receiving services from the community-based organizations. Before the focus group, participants requested a program manager be a part of the group because some participants with HIV were new to participating in research. The program manager contributed ease during the group, which resulted in a robust discussion about HIV comorbidities of concern. Phase 1 resulted in a total of 84 pages of transcribed data.

Theme 1: Mixed Perceptions of Health and HIV

Overall, most participants rated their health as "decent" to "really healthy." They conveyed being very in tune with their health issues and that having a diagnosis of HIV was really the catalyst for being more cognizant about their health behaviors. Of the 11 client participants, 7 (63%) were diagnosed with HIV in the last 5 years and had mixed perceptions regarding how

XSL•FO
RenderX

they perceived their overall health. Some of the sentiments conveyed about health perceptions were due to the effects of acquiring HIV during the COVID-19 pandemic (Textbox 2).

Of the 4 HIV care experts, 3 conveyed that mixed perceptions of health-related behaviors (Textbox 3) result from HIV diagnosis in newly diagnosed young people when compared to people who have had HIV for an extended duration.

Textbox 2. Sentiments conveyed by participants about health perceptions.

Sentiments

- "I think me testing positive...being diagnosed with HIV...have helped me open up my mind and tell myself now it's time to get your health in order...not just your physical health, but your mental health as well...I've changed my diet trying to eat healthier, drink healthier, trying to exercise more to help with my physical health, my mental health. Sometimes depending on the weather, instead of taking transportation somewhere I walk just to get a little bit of exercise; or when the weather is warm out, I'll bring out my bike and I'll go bike riding around the city just for the exercise" (Pseudonym: Eric).
- "...overall, I feel pretty healthy...I want to live a good long healthy life with the least amount of complications as possible. So yes, if there is ever something, I go to my doctor for, and he's like, hey, you need to fix this – I'm already on it. For example, like for vitamin D, I think it was my vitamin D, it was really low, and I said okay. Right after the visit, I went right to the store bought those vitamins, started taking those. Three months later, we did my blood work, he's like, oh, you're good. I said, thanks" (Pseudonym: Roman).
- "I got diagnosed [with HIV] during the beginning of COVID, like in March...It's been a process, so to speak, because I would've expected for me to be a ho [promiscuous] and then get it [acquire HIV]. But I trusted somebody, you know, that was supposed to be mine and he fell in love with somebody else that was positive. To even know that he got [HIV] positive, and yeah, that's what, to me, the whole world went...you know issues. Because, I have trust issues already as a person, you know. People show their true colors. It's common for people to have trust issues after you go through a very traumatic situation" (Pseudonym: Arai).
- "My diagnosis was actually August of 2020, so almost two years. It was a shock. This is on the period of now COVID, and then this happening...This has been a journey in that...from receiving my diagnosis to now has been...as a person also of color, trying to find the right physician that has listened to my needs and my concerns. Living with a compromised immune system especially because of everything, post-diagnosis...it's been one of those things where I have had to...I've had...I felt dismissed in the beginning. I've had some concerns as related to other health issues related to diagnosis. So, I've had it where some providers have thought I was being a little overly anxious, so they dismissed my concerns or they spoke down to me. At the time until I spoke to other friends, I would have stayed [in care with the same provider] with it because part of it were issues were...maybe financially or economic reasons...I just didn't have the access to get a better physician. So, I'm stuck with what was familiar with even when I didn't feel like most of my concerns and medical needs were met. It took some time to navigate that and go out of my comfort or familiarity and then seek out providers who assess my needs. Even if they were a little bit further geographically than I was used to, that overall my experience with them would be a little bit better than what I was already getting" (Pseudonym: Zoe).

Textbox 3. HIV care experts' selected utterances about clients' health-related behaviors.

Theme 1: mixed perceptions of health and HIV

- "...people who are living with HIV tend to be more adherent to taking their medications, you know, as prescribed, simply because they want to achieve undetectable viral load. And in addition, in my opinion, they tend to be a little more healthier than people who are not living with HIV" (HIV care expert pseudonym: Tyler).
- "...people that have been already living with HIV for many, many years, they are more conscious about how to take care about themselves, the ways of getting HIV or any other comorbidities, how to take care of themselves. But, I am not very sure that evolves with young people to take care of their issues, their conditions in the same way even though we have more information right now, and more access to treatment. I think that the attitude is a little bit different" (HIV care expert pseudonym: Lloyd).
- "I've noticed that over my experience with patients, people with HIV is that there are ebbs and flows, and it depends on where they are in their life and how the perceive their HIV. I do notice that younger folks sometimes think about it like, I have to take this from now until forever for my HIV, and it becomes overwhelming for them at times. I do find that the ones who are diagnosed years ago tend to be more stabilized with their ideas and their thoughts about HIV, and being able to keep them adherent to medication and so forth" (HIV care expert pseudonym: Eric).

Theme 2: High Risk for Comorbid Conditions

Participants with HIV perceived that they were susceptible to other illnesses as a result of having HIV, stemming from environmental conditions, such as diabetes and asthma. They spoke about lung function and the effects of the weather on their breathing. CVD and CVD risk factors were consistently expressed concerns in relation to HIV and health, including hypertension, myocardial infarction, stroke, and diabetes. Additional well-described comorbidities of HIV, such as kidney and liver diseases, were also identified by participants as important. Cancer (prostate, colon, and others) was a common concern among participants. Participants with HIV also voiced concerns about mental health conditions. Participants conveyed the following:

> *I heard that sometimes complications can come with, you know living with HIV, like cancer and other health problems that can come up, that I'm not really...I'm not aware of all of them. But I don't know; cancer is just something that I thought that could be. And especially too, with heart problems or you know, different things like that...Heart attacks, strokes, palpitations, all kinds of stuff.* [Pseudonym: Kori]

> *I have high blood pressure, so I take high blood pressure medicine. I also have asthma, so I usually*

XSL•FO
RenderX

have a pump just in case. That's pretty much it. [Pseudonym: Ralph]

Cancer is a concern for me because it runs in my family. It's something that's kind of, I guess is hereditary. We've lost a lot of family members due to that. So, for people who probably have hereditary illnesses or diseases that run in their family, that should be of concern to them. [Pseudonym: Omar]

I don't know why, but colon cancer, and that's not it, because I remember hearing someone talk about it, and I was just like, this can really happen! I never knew that. [Pseudonym: Roman]

The HIV care experts described chronic diseases and mental health conditions that were affecting participants (Textbox 4), such as anxiety and depression due to the stress of living with HIV and, for some, HIV comorbidities.

Textbox 4. HIV care experts' selected utterances about their clients' risk for comorbid conditions.

Theme 2: high risk for comorbid conditions

- "I've dealt with a lot of diabetes and hypertension dominating those areas. And, I deal with a lot of patients who are HIV positive and diabetes, and like the secondary factor that is a big concern to them...They haven't learned yet...most of the people say that they've not learned what are the best options, as far as nutrition, to take care of their diabetes on top of their HIV. So, they're focused on what they should or should not be eating " (HIV care expert pseudonym: Eric).
- "...but another health issue that still didn't come out. And it's about liver disease. The population that I work with that are mostly over 45 to 50-years old, this sort of thing has developed. And, in the different part of the body or even cirrhosis, even those these people do not drink any alcohol, okay. So, I feel that it's another point to keep in mind" (HIV care expert pseudonym: Lloyd).

Theme 2.1: high risk for mental health conditions

- "I know we're talking a lot about more physical, although I do think depression is physical too. But, more physical illnesses, but depression is probably the, really, most common one I've heard about, you know, being either diagnosed with depression or anxiety disorder" –(HIV care expert pseudonym: Eric).
- "I work with people who become, at least clinically depressed after their diagnoses. I was going to say that earlier as well. So, there may not actually be like a physical thing, but it certainly can develop into a physical thing, as someone falls into, I guess, a deep depression. Or, like they have these drug induced psychosis and stuff like that, which is quite possible for an HIV induced psychosis as well, you know, because you were recently, you know, diagnosed, that can kick up some things in that person that may have not necessarily been focused on earlier or identified earlier" (HIV care expert pseudonym: Tyler).

Theme 3: Virtual Environment Features

We tested out features of the virtual environment, such as avatar customization, quests, and navigation to various districts (Textbox 1). There was excitement around the privacy and anonymity of the platform and the customization of the skin tone, hair color, and clothing of the avatar. Participants were excited about the potential of meeting other participants with HIV who identified as gay or bisexual. Information within the virtual environment was deemed interesting, engaging, and supportive of their identity. Participants commented about naming the bookstore after Marsha P Johnson and the representation of pictures and content that validated race and ethnicity. Overall, the accessibility of the virtual environment to ethnic and racial sexual minority men with HIV, who have not had prior exposure to clinical trials addressing heart disease prevention, was deemed as appealing. The participants' views are listed in Textbox 5.

HIV care experts had favorable sentiments about the virtual environment. They focused on nutrition, avatar diversity, engagement with others, and pushing the envelope to extend the virtual environment to new areas in preventive screenings (Textbox 6).

Textbox 5. Participants' views about virtual environment features.

Participants' views

- "I don't even think I could add anything else. I think just within itself I feel like it would be amazing because it's so different. Say maybe somebody wants information or wants to meet other people to discuss things, sometimes it can be nerve wracking or sometimes people aren't ready to really put themselves out there like that. So, I think with having the avatar and having different spaces, like the library and the relaxation area is good because you could be who you would want to be with the avatar. So, you're comfortable to really put yourself out there. I think the concept is really good because some people aren't as up front. It's like they still get to maintain their sense of privacy, but also get to interact with other people and get more knowledge and stuff about things. Yes, I think it's great the way it is" (Pseudonym: Ralph).
- "I think that [privacy and anonymity] is important for the simple fact that that people do enjoy their privacy. I would say that's great, because I'm someone who plays online a lot, and I never am giving out my real name. So, it's always, this is my gamer tag, and if you want to chat, we can talk on this court where it makes things more private. You don't have my number, but you can message me on there" (Pseudonym: Roman).
- "I would love that [to use a virtual environment for heart disease prevention], and honestly, for me, like I said, I'm very much a hands-on person. I love the idea of the experiencing new things. So, you had me at, you know, the start of the [virtual environment platform] pictures" (Pseudonym: Arai).
- "...I think that giving tips to us, how to take care of ourselves, better care of ourselves to prevent certain illnesses. I think it should be information about the type of foods to eat, you know vitamins, different tips of how to keep ourselves, you know, healthy" (Pseudonym: Kori).

XSL•FO
RenderX

Textbox 6. HIV care experts' feedback on beta testing the virtual environment.

Feedback

- "I liked the avatar. I liked how it was easy to change to whether I wanted to be male or female. Skin tone is important. If I wanted to wear something short, something long, it gave the options. So, it wasn't limited to just I have to be a girl or have to be a boy...And I think that the areas are very colorful. The Reflection area, it looked like a Tiki bar. So it makes you want to meet with people. So it's a good, it's visually appealing to say; hey, let's meet here...Some people are very uncomfortable talking about health messages or their own personal. But if you have an avatar and your avatar can ask questions, or a trivia questions pops up and you're in a group, and they're now competing with another group, it may be better to get messages out" (HIV care expert pseudonym: Emma).
- "...I definitely liked the areas. I think all the areas was nice. I think the game was well-developed. However, I think maybe the avatars should be maybe a little more reflective of the people using the games. Also, I think that...And I know that you said that this is something that you guys are working on. But changing the menu options and the pharmacy things, the things that people actually use...I think it's definitely something that we would use with our clients. It's easy to navigate. The information is there. I mean I work with a lot of people who are undocumented. They have a hard time understanding English. The visual aspect of it is great. It is also something that we can use to keep our clients on track. So if we were able to customize it, I would customize it to questions that reflect the communities that we serve. Like, have you been taking your medication in the past five days?" (HIV care expert pseudonym: Sequan)
- "For example, talking about aging with HIV. How does that look? The other comorbidities that can accompany HIV as well. As we get older, we have so many things that can...We are not as potent as we used to be when we were younger. So we are coming with other comorbidities. And plus, HIV adds into that as well. So how does that look? So what are the other problems that you may have? Also the screenings, what screenings are recommended based on particular times of your life? For example, Are the males doing their colonoscopy on time? It's recommended to do it at 45 now. It used to be 50. Are you receiving some vaccines? For example, you have Shingrix that protects you against Shingles. Will you have to take that a second time? And you have vaccines against pneumonia, and flu shots. And also, cognition. As we get older, you can have cases of dementia. You can have isolation. How do you meet other people of your age? How do you socialize? There are other things that can...I mean if you want to really push it to the limit, there are so many things that you can identify that can also make a difference" (HIV care expert pseudonym: Mario).
- "I think that's a good use for it [the virtual environment], just being able to look at the calories and the ingredients in certain food items and such. But also, what I think my favorite part was the gym and the health tips. So like in terms of basic, how much water intake you should have in a day, and information walking. I think one of the things it talked about was how to select a shoe and how to do things like that, for exercise. These are questions that people don't think about until they have to think about it. And so going there and finding that information I think is helpful" (HIV care expert pseudonym: Kayla).
- "So by adding those varieties of cuisine, for example, like Italian, Caribbean, it's a way you're helping the person to make a different choice, if they ever get a chance in the real life. For example, there is a food that you might not know about. And by presenting it to you in the virtual, it's all right. I mean they might be more willing to try this. And also if you used to hear something about like a food, now that you have the facts, would you want to try it? So I think that's going to be pretty helpful" (HIV care expert pseudonym: Porter).

Phase 2: Usability and Beta Testing

All 10 participants were able to complete all 12 action items on the checklist (Figure 1).

The reflection garden was the most enjoyed area by participants because it was the most serene. The fitness center received the most feedback, especially that it needs more equipment to become more realistic. The duration of time to completion of the checklist ranged from 30 to 60 minutes. Those with more technology experience were able to navigate the checklist faster. Information within the virtual environment was deemed interesting, engaging, and supportive of their identity. Participants commented about naming the bookstore after Marsha P Johnson and the representation of pictures and content that validated race and ethnicity (Figure 2). Overall, feedback from the usability and beta testing of the virtual environment was favorable but also contained constructive feedback for representation and diversity.

Selected utterances from the beta testers are presented in Textbox 7.

XSL•FO
RenderX

Figure 1. 12-item usability checklist. FAQ: frequently asked question.

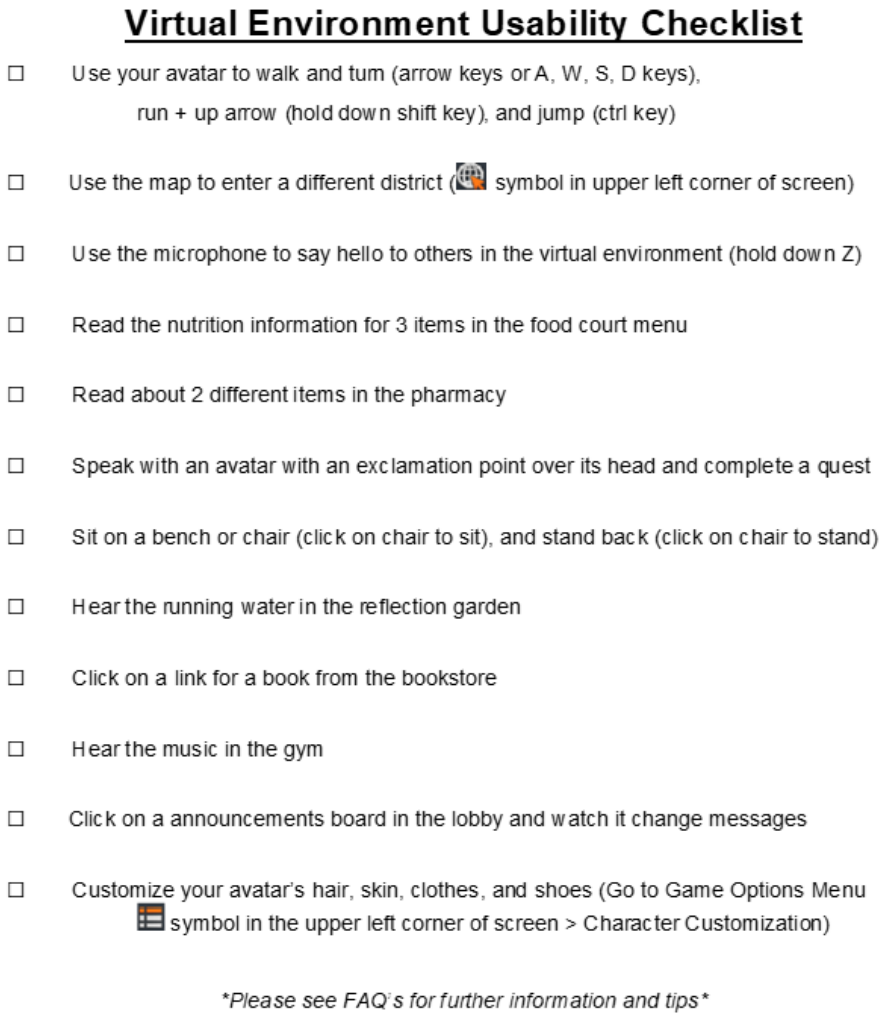

Virtual Environment Usability Checklist

- ☐ Use your avatar to walk and turn (arrow keys or A, W, S, D keys), run + up arrow (hold down shift key), and jump (ctrl key)
- ☐ Use the map to enter a different district (symbol in upper left corner of screen)
- ☐ Use the microphone to say hello to others in the virtual environment (hold down Z)
- ☐ Read the nutrition information for 3 items in the food court menu
- ☐ Read about 2 different items in the pharmacy
- ☐ Speak with an avatar with an exclamation point over its head and complete a quest
- ☐ Sit on a bench or chair (click on chair to sit), and stand back (click on chair to stand)
- ☐ Hear the running water in the reflection garden
- ☐ Click on a link for a book from the bookstore
- ☐ Hear the music in the gym
- ☐ Click on a announcements board in the lobby and watch it change messages
- ☐ Customize your avatar's hair, skin, clothes, and shoes (Go to Game Options Menu symbol in the upper left corner of screen > Character Customization)

Please see FAQ's for further information and tips

XSL•FO
RenderX

Figure 2. The LEARN virtual environment's 4 districts (lobby, plaza, food court, and reflection garden).

Textbox 7. Selected quotes obtained during best testing.

Selected quotes

- "I think the names [of the stores and restaurants] are really creative. And I like how some of them pay homage [to historical lesbian, gay, bisexual, transgender, queer, and others (LGBTQ+) leaders] I think you guys can get creative with the graphics on these graphics here, because you've got a bookstore called the Marsha B. Johnson Bookstore. So it'd be nice if you could have maybe pictures of her on the slides, or historical figures on the slides. That would be cool. Yeah. That's pretty much it. I would just add that" (Pseudonym: Justin).
- "Gay's Anatomy! I'm dying. I'm so down with this game. This is really cute. Okay...Marsha P. Johnson Bookstore? Marsha P? Come on, honey. Stone wall and all, girl!...But the avatars, they need to be able to customize the body, too. Like make them bigger, if they want them to be bigger. Or make them skinnier if they want them to be skinnier. Because some people might come on something like this not just for the information, but like I said, to meet other people and maybe talk to other people. And I feel like they should be able to emulate themselves in this world like however they want" (Pseudonym: Beta4).
- "I would change the texture of the hair [on the avatars] and have more options of the shoes and colors of pants. His [the avatar's] hair is more like one of them Backstreet Boys hair. What about curly? What about afro? What about braids" (Pseudonym: Felix).
- "I would be more interested in the trivia, simply because I can learn something new. And it's easy to figure out, because I get explanations on why I'm right or wrong. And if I have questions about something, I can always go to a different section and be informed about my questions." (Pseudonym: Silver).

Discussion

Principal Findings

The purpose of this study was to explore perceptions of HIV-related comorbidities and assess the usability of a virtual environment for CVD prevention education in Black and Latinx sexual minority men with HIV. We found that perceptions of health among participants were mixed and based on life context. Those who perceived their health in a positive light were motivated to practice enhanced health behaviors, such as enhancing physical activity, medication adherence, and following clinician health recommendations, as a result of acquiring HIV. Negative sentiments toward health perceptions were reported due to personal relationships, being diagnosed with HIV during the COVID-19 pandemic, and socioeconomic determinants that limited accessibility to lesbian, gay, bisexual, transgender, queer, and others (LGBTQ+) health care specialists.

XSL•FO
RenderX

Cancer was identified as a predominant illness of concern. There is an increased risk of cancers, other than AIDS-defining illnesses, among people with HIV, and concern may be due to reported heredity factors as well as potential risk due to shared cancer-related risk factors. In a recent study by Bell et al [49], CVD was associated with a higher incidence of cancer, especially in persons with atherosclerotic CVD, which carries significant implications for mitigating barriers to preventive health screenings and raises considerations for cancer and multimorbidity in Black and Latinx sexual minority men with HIV.

With regard to the virtual environment, almost all phase-1 client participants with HIV (10/11, 91%) were enthusiastic. We believe that hesitancy in the remaining participant was related to not having prior access to or experience with evolving technologies and unfamiliarity with the premise of using a virtual environment. Participants with HIV were particularly open to the use of a virtual environment because they would assume the identity of an avatar, with a pseudonym, and would meet others who had shared characteristics and health goals. Participants were pleased about the potential to learn with others who were a part of the study, which is an important consideration in behavioral interventions, as individuals build community around shared life experiences. We found that the participants were interested and found the content desirable because we named the bookstore after Marsha P Johnson (a Black transgender woman who was a pioneer and activist for gay and transgender rights) and included positive visualizations of ethnic and racial sexual minority men, queer love, and pride. This was important to validate and affirm the existence of the participants. These types of visualizations are uncommon in public areas in the United States and are banned in certain states. Providing a virtual environment that celebrates ethnic and racial LGBTQ+ individuals and legacy activists was deemed as acceptable and wanted.

HIV care experts are trusted individuals who work with the community and have keen insights into their needs, challenges, and opportunities to enhance cardiovascular health. Their lived and professional experiences working with ethnic, racial, and sexual minoritized individuals enhanced our understanding about the intersectionality of race, ethnicity, HIV serostatus, and CVD risk in relation to community needs.

The prevention of comorbidities in persons with HIV is complicated due to historic and ongoing barriers to health equity in health care and research [20] that disproportionately target racial, ethnic, sexual minority populations and other marginalized populations [11]. CVD prevention is also challenging because individuals are required to become proficient in a variety of lifestyle behaviors to enhance cardiovascular health and incorporate these into their everyday lives [50]. Our study findings was favorable toward the use of a virtual environment for CVD prevention education in ethnic and racial sexual minority men with HIV. These findings are in alignment with a 2019 survey conducted by the AARP formerly known as the American Association of Retired Persons, which suggested that 50% of African American and 42% of Latino persons aged >50 years played video games [51,52]. The 2 prominent reasons for engaging in video game play were entertainment and the maintenance of mental sharpness [51,52]. Moreover, other studies have used virtual environments (defined as 3D computerized environments with avatars) for risk reduction in adolescents [53,54], diabetes self-management [37,55], and cardiac rehabilitation [56]; however, none have focused on CVD prevention in persons with HIV. Recommendations from a systematic review identified the need for future research on gamification for CVD prevention to include (1) larger sample sizes, (2) experimental designs, (3) objective measures, and (4) behavior change frameworks [57]. Our qualitative study, used to inform our future clinical trial [33], directly responds to the need to develop novel, widely applicable strategies to address cardiovascular health equity using a virtual environment for individuals from a variety of minoritized racial and ethnic groups with HIV, particularly those who might otherwise lack access to clinical trials research.

Limitations

First, although our sample size was in alignment with qualitative guidelines, small sample sizes limit generalizability. However, the diverse participant sample and inclusion of HIV care experts provided insights into perceived CVD and other comorbid risks and the potential for the use of a virtual environment for CVD prevention education. This type of data has not been captured in previous studies from an exclusive participant sample of economically marginalized, ethnic and racial sexual minority men with HIV. Second, using qualitative methods, we were unable to demonstrate causality [58]. However, we followed established criteria for ensuring rigor during data collection [58] and obtained >80 pages of qualitative data. A strength of qualitative data collection was the ability to gain high-quality contextual insights and details into HIV comorbidities of concern and assess the usability of and verbal reactions about the virtual environment, which would not be obtainable using survey methods. Third, recruitment bias is a potential limitation. We minimized recruitment bias by adhering to our prespecified inclusion criteria.

Conclusions

We identified several comorbid conditions of concern, and findings carry significant implications for mitigating barriers to preventive health screenings in Black and Latinx sexual minority men with HIV, given the shared risk factors between HIV and related comorbidities. Highly rated aspects of the virtual environment were anonymity, the ability to meet others with HIV who identify as gay or bisexual, validating LGBTQ+ inclusive images and content, and accessibility to CVD prevention education. Critical end-user feedback from beta testing suggested more options for avatar customization in skin, hair, and body representation. We will apply these recommendations and others in phase 3 to formally test the virtual environment in the subsequent clinical trial.

XSL•FO
RenderX

Acknowledgments

This study was funded by the National Heart, Lung, and Blood Institute (K01HL145580) and the Yale Fund for Lesbian and Gay Studies (principal investigator: SRR). The funding organizations had no role in the design or conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

The authors would like to acknowledge and thank their community partners and study participants for their engagement and feedback during the phases of data collection. They would also like to thank Maame-Owusua Boateng (Master of Public Health) and Lauren Chin (Master of Public Health) for their assistance with coding and data analysis and Lauren Chin and Harry Kim (Registered Nurse and Bachelor of Science in Nursing) for their assistance with usability checklist pilot testing.

Data Availability

The data sets generated during and analyzed during this study are not publicly available or available for sharing given the small sample size and even the combination of indirect identifiers could possibly identify a participant.

Conflicts of Interest

SRR is an associate editor for the *Journal of Medical Internet Research*. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Standards for Reporting Qualitative Research (SRQR) checklist.
[PDF File (Adobe PDF File), 873 KB - jmir_v26i1e57351_app1.pdf]

References

1. Martin SS, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. 2024 Heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2024 Feb 20;149(8):e347-e913 [FREE Full text] [doi: 10.1161/CIR.0000000000001209] [Medline: 38264914]
2. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
3. Feinstein MJ, Bahiru E, Achenbach C, Longenecker CT, Hsue P, So-Armah K, et al. Patterns of cardiovascular mortality for HIV-infected adults in the United States: 1999 to 2013. Am J Cardiol 2016 Jan 15;117(2):214-220 [FREE Full text] [doi: 10.1016/j.amjcard.2015.10.030] [Medline: 26639041]
4. Losina E, Hyle EP, Borre ED, Linas BP, Sax PE, Weinstein MC, et al. Projecting 10-year, 20-year, and lifetime risks of cardiovascular disease in persons living with human immunodeficiency virus in the United States. Clin Infect Dis 2017 Oct 15;65(8):1266-1271 [FREE Full text] [doi: 10.1093/cid/cix547] [Medline: 28605504]
5. Althoff KN, McGinnis KA, Wyatt CM, Freiberg MS, Gilbert C, Oursler KK, et al. Comparison of risk and age at diagnosis of myocardial infarction, end-stage renal disease, and non-AIDS-defining cancer in HIV-infected versus uninfected adults. Clin Infect Dis 2015 Feb 15;60(4):627-638 [FREE Full text] [doi: 10.1093/cid/ciu869] [Medline: 25362204]
6. Smit M, Brinkman K, Geerlings S, Smit C, Thyagarajan K, Sighem AV, et al. Future challenges for clinical care of an ageing population infected with HIV: a modelling study. Lancet Infect Dis 2015 Jul;15(7):810-818 [FREE Full text] [doi: 10.1016/S1473-3099(15)00056-0] [Medline: 26070969]
7. Hsue PY, Waters DD. Time to recognize HIV infection as a major cardiovascular risk factor. Circulation 2018 Sep 11;138(11):1113-1115 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.118.036211] [Medline: 30354392]
8. Diaz CL, Shah NS, Lloyd-Jones DM, Khan SS. State of the nation's cardiovascular health and targeting health equity in the United States: a narrative review. JAMA Cardiol 2021 Aug 01;6(8):963-970 [FREE Full text] [doi: 10.1001/jamacardio.2021.1137] [Medline: 34009231]
9. Caceres BA, Ancheta AJ, Dorsen C, Newlin-Lew K, Edmondson D, Hughes TL. A population-based study of the intersection of sexual identity and race/ethnicity on physiological risk factors for CVD among U.S. adults (ages 18-59). Ethn Health 2022 Apr;27(3):617-638 [FREE Full text] [doi: 10.1080/13557858.2020.1740174] [Medline: 32159375]
10. Thayer JF, Carnevali L, Sgiofo A, Williams DP. Angry in America: psychophysiological responses to unfair treatment. Ann Behav Med 2020 Dec 01;54(12):924-931. [doi: 10.1093/abm/kaaa094] [Medline: 33416840]
11. Caceres BA, Streed CGJ, Corliss HL, Lloyd-Jones DM, Matthews PA, Mukherjee M, et al. Assessing and addressing cardiovascular health in LGBTQ adults: a scientific statement from the American Heart Association. Circulation 2020 Nov 10;142(19):e321-e332 [FREE Full text] [doi: 10.1161/CIR.0000000000000914] [Medline: 33028085]
12. McEwen BS, Stellar E. Stress and the individual. Mechanisms leading to disease. Arch Intern Med 1993 Sep 27;153(18):2093-2101. [Medline: 8379800]

XSL•FO
RenderX

13. Geronimus AT. The weathering hypothesis and the health of African-American women and infants: evidence and speculations. Ethn Dis 1992;2(3):207-221. [Medline: 1467758]
14. Seeman TE, Singer BH, Rowe JW, Horwitz RI, McEwen BS. Price of adaptation--allostatic load and its health consequences. MacArthur studies of successful aging. Arch Intern Med 1997 Oct 27;157(19):2259-2268. [Medline: 9343003]
15. Duong MT, Bingham BA, Aldana PC, Chung ST, Sumner AE. Variation in the calculation of allostatic load score: 21 examples from NHANES. J Racial Ethn Health Disparities 2017 Jun;4(3):455-461 [FREE Full text] [doi: 10.1007/s40615-016-0246-8] [Medline: 27352114]
16. Fazeli PL, Waldrop-Valverde D, Yigit I, Turan B, Edberg J, Kempf M, et al. An exploratory study of correlates of allostatic load in older people living with HIV. J Acquir Immune Defic Syndr 2020 Apr 15;83(5):441-449 [FREE Full text] [doi: 10.1097/QAI.0000000000002293] [Medline: 31914006]
17. Hill LK, Thayer JF. The autonomic nervous system and hypertension: ethnic differences and psychosocial factors. Curr Cardiol Rep 2019 Feb 28;21(3):15 [FREE Full text] [doi: 10.1007/s11886-019-1100-5] [Medline: 30820683]
18. Hill LK, Hoggard LS, Richmond AS, Gray DL, Williams DP, Thayer JF. Examining the association between perceived discrimination and heart rate variability in African Americans. Cultur Divers Ethnic Minor Psychol 2017 Jan;23(1):5-14 [FREE Full text] [doi: 10.1037/cdp0000076] [Medline: 28045306]
19. Williams DP, Joseph N, Hill LK, Sollers JJ3, Vasey MW, Way BM, et al. Stereotype threat, trait perseveration, and vagal activity: evidence for mechanisms underpinning health disparities in Black Americans. Ethn Health 2019 Nov;24(8):909-926 [FREE Full text] [doi: 10.1080/13557858.2017.1378803] [Medline: 28922935]
20. Ramos SR, Warren R, Shedlin M, Melkus G, Kershaw T, Vorderstrasse A. A framework for using eHealth interventions to overcome medical mistrust among sexual minority men of color living with chronic conditions. Behav Med 2019;45(2):166-176 [FREE Full text] [doi: 10.1080/08964289.2019.1570074] [Medline: 31343963]
21. Muncan B. Cardiovascular disease in racial/ethnic minority populations: illness burden and overview of community-based interventions. Public Health Rev 2018;39:32 [FREE Full text] [doi: 10.1186/s40985-018-0109-4] [Medline: 30524764]
22. Ramos SR, O'Hare OM, Hernandez Colon A, Kaplan Jacobs S, Campbell B, Kershaw T, et al. Purely behavioral: a scoping review of nonpharmacological behavioral and lifestyle interventions to prevent cardiovascular disease in persons living with HIV. J Assoc Nurses AIDS Care 2021;32(5):536-547 [FREE Full text] [doi: 10.1097/JNC.0000000000000230] [Medline: 33481464]
23. Fredriksen-Goldsen KI, Simoni JM, Kim HJ, Lehavot K, Walters KL, Yang J, et al. The health equity promotion model: reconceptualization of lesbian, gay, bisexual, and transgender (LGBT) health disparities. Am J Orthopsychiatry 2014 Nov;84(6):653-663 [FREE Full text] [doi: 10.1037/ort0000030] [Medline: 25545433]
24. Ramos SR, Lardier DTJ, Opara I, Turpin RE, Boyd DT, Gutierrez JIJ, et al. Intersectional effects of sexual orientation concealment, internalized homophobia, and gender expression on sexual identity and HIV risk among sexual minority men of color: a path analysis. J Assoc Nurses AIDS Care 2021;32(4):495-511 [FREE Full text] [doi: 10.1097/JNC.0000000000000274] [Medline: 34101701]
25. Caraballo C, Ndumele CD, Roy B, Lu Y, Riley C, Herrin J, et al. Trends in racial and ethnic disparities in barriers to timely medical care among adults in the US, 1999 to 2018. JAMA Health Forum 2022 Oct 07;3(10):e223856 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.3856] [Medline: 36306118]
26. Commodore-Mensah Y, Loustalot F, Himmelfarb CD, Desvigne-Nickens P, Sachdev V, Bibbins-Domingo K, et al. Proceedings from a National Heart, Lung, and Blood Institute and the Centers for Disease Control and Prevention Workshop to control hypertension. Am J Hypertens 2022 Mar 08;35(3):232-243 [FREE Full text] [doi: 10.1093/ajh/hpab182] [Medline: 35259237]
27. Entrup P, Garcia S, Hall OT. Bilingual care in clinical trials-understanding the need for expanded research in Latino communities. JAMA Netw Open 2023 May 01;6(5):e2314860 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.14860] [Medline: 37219909]
28. Turner BE, Steinberg JR, Weeks BT, Rodriguez F, Cullen MR. Race/ethnicity reporting and representation in US clinical trials: a cohort study. Lancet Reg Health Am 2022 Jul;11:100252 [FREE Full text] [doi: 10.1016/j.lana.2022.100252] [Medline: 35875251]
29. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin A, Park H, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023 Sep 05;330(9):818-820. [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
30. Dhingra LS, Shen M, Mangla A, Khera R. Cardiovascular care innovation through data-driven discoveries in the electronic health record. Am J Cardiol 2023 Sep 15;203:136-148. [doi: 10.1016/j.amjcard.2023.06.104] [Medline: 37499593]
31. Johnson AE, Routh S, Taylor CN, Leopold M, Beatty K, McNamara DM, et al. Developing and implementing an mHealth heart failure self-care program to reduce readmissions: randomized controlled trial. JMIR Cardio 2022 Mar 21;6(1):e33286 [FREE Full text] [doi: 10.2196/33286] [Medline: 35311679]
32. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018 Sep;15(5):429-448. [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
33. Ramos SR, Johnson C, Melkus G, Kershaw T, Gwadz M, Reynolds H, et al. Cardiovascular disease prevention education using a virtual environment in sexual-minority men of color with HIV: protocol for a sequential, mixed method, waitlist

XSL•FO
RenderX

randomized controlled trial. JMIR Res Protoc 2022 May 17;11(5):e38348 [FREE Full text] [doi: 10.2196/38348] [Medline: 35579928]
34. Vorwerg S, Stamm O, Müller-Werdan U. Movement training in augmented reality for older adults with essential hypertension: a qualitative study determining training requirements. Games Health J 2020 Dec;9(6):436-445. [doi: 10.1089/g4h.2019.0133] [Medline: 33301387]
35. Johnson C, Feinglos M, Pereira K, Hassell N, Blascovich J, Nicollerat J, et al. Feasibility and preliminary effects of a virtual environment for adults with type 2 diabetes: pilot study. JMIR Res Protoc 2014 Apr 08;3(2):e23 [FREE Full text] [doi: 10.2196/resprot.3045] [Medline: 24713420]
36. Ruggiero L, Moadsiri A, Quinn LT, Riley BB, Danielson KK, Monahan C, et al. Diabetes island: preliminary impact of a virtual world self-care educational intervention for African Americans with type 2 diabetes. JMIR Serious Games 2014;2(2):e10 [FREE Full text] [doi: 10.2196/games.3260] [Medline: 25584346]
37. Vorderstrasse AA, Melkus GD, Pan W, Lewinski AA, Johnson CM. Diabetes learning in virtual environments: testing the efficacy of self-management training and support in virtual environments (randomized controlled trial protocol). Nurs Res 2015;64(6):485-493 [FREE Full text] [doi: 10.1097/NNR.0000000000000128] [Medline: 26505161]
38. Stamm O, Vorwerg S, Haink M, Hildebrand K, Buchem I. Usability and acceptance of exergames using different types of training among older hypertensive patients in a simulated mixed reality. Appl Sci 2022 Nov 10;12(22):11424. [doi: 10.3390/app122211424]
39. Lloyd-Jones DM, Allen NB, Anderson CA, Black T, Brewer LC, Foraker RE, et al. Life's essential 8: updating and enhancing the American Heart Association's construct of cardiovascular health: a presidential advisory from the American Heart Association. Circulation 2022 Aug 02;146(5):e18-e43 [FREE Full text] [doi: 10.1161/CIR.0000000000001078] [Medline: 35766027]
40. Ramos SR, Fraser M, Araya F, Kim HY, Parrilla JA, Sy KM, et al. Community-engaged intervention mapping for cardiovascular disease prevention in Black and Latinx sexual minority men with HIV in New York City: protocol for a web-based mixed methods study. JMIR Res Protoc 2022 Oct 11;11(10):e41602 [FREE Full text] [doi: 10.2196/41602] [Medline: 36130735]
41. Sandelowski M. Sample size in qualitative research. Res Nurs Health 1995 Apr;18(2):179-183. [doi: 10.1002/nur.4770180211] [Medline: 7899572]
42. Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, et al. What is an adequate sample size? operationalising data saturation for theory-based interview studies. Psychol Health 2010 Dec;25(10):1229-1245. [doi: 10.1080/08870440903194015] [Medline: 20204937]
43. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications; 2017.
44. Speziale HS, Carpenter DR. Qualitative Research in Nursing: Advancing the Humanistic Imperative. New York, NY: Wolters Kluwer; 2011.
45. Creswell JW, Poth CN. Qualitative Inquiry & Research Design: Choosing Among Five Approaches. Thousand Oaks, CA: SAGE Publications; 2017.
46. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
47. Valacich J, George JF, Hoffer JA. Essentials of Systems Analysis and Design, Global Edition. London, UK: Pearson Education; 2014.
48. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008 Apr;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]
49. Bell CF, Lei X, Haas A, Baylis RA, Gao H, Luo L, et al. Risk of cancer after diagnosis of cardiovascular disease. JACC CardioOncol 2023 Aug;5(4):431-440 [FREE Full text] [doi: 10.1016/j.jaccao.2023.01.010] [Medline: 37614573]
50. Riegel B, Lee CS, Dickson VV, Medscape. Self care in patients with chronic heart failure. Nat Rev Cardiol 2011 Jul 19;8(11):644-654. [doi: 10.1038/nrcardio.2011.95] [Medline: 21769111]
51. Gaming gains. AARP. URL: https://www.aarp.org/content/dam/aarp/research/surveys_statistics/technology/2020/hispanics-video-games-trends.doi.10.26419-2Fres.00328.004.pdf [accessed 2021-07-24]
52. Gaming - more than just entertainment. AARP. 2020. URL: https://www.aarp.org/content/dam/aarp/research/surveys_statistics/technology/2020/african-american-video-games-trends.doi.10.26419-2Fres.00328.003.pdf [accessed 2022-07-16]
53. Fiellin LE, Kyriakides TC, Hieftje KD, Pendergrass TM, Duncan LR, Dziura JD, et al. The design and implementation of a randomized controlled trial of a risk reduction and human immunodeficiency virus prevention videogame intervention in minority adolescents: PlayForward: Elm City Stories. Clin Trials 2016 Aug;13(4):400-408 [FREE Full text] [doi: 10.1177/1740774516637871] [Medline: 27013483]
54. Fiellin LE, Hieftje KD, Pendergrass TM, Kyriakides TC, Duncan LR, Dziura JD, et al. Video game intervention for sexual risk reduction in minority adolescents: randomized controlled trial. J Med Internet Res 2017 Sep 18;19(9):e314 [FREE Full text] [doi: 10.2196/jmir.8148] [Medline: 28923788]

XSL•FO
RenderX

55. Vorderstrasse A, Melkus GD, Feinglos M, Pan W, Chang J, Pereira K, et al. Virtual environment for diabetes self-management education and support: preliminary RCT outcomes. Circulation 2017 Nov 14;136:A17519 [FREE Full text]
56. Brewer LC, Kaihoi B, Zarling KK, Squires RW, Thomas R, Kopecky S. The use of virtual world-based cardiac rehabilitation to encourage healthy lifestyle choices among cardiac patients: intervention development and pilot study protocol. JMIR Res Protoc 2015 Apr 08;4(2):e39 [FREE Full text] [doi: 10.2196/resprot.4285] [Medline: 25857331]
57. Davis AJ, Parker HM, Gallagher R. Gamified applications for secondary prevention in patients with high cardiovascular disease risk: a systematic review of effectiveness and acceptability. J Clin Nurs 2021 Oct;30(19-20):3001-3010. [doi: 10.1111/jocn.15808] [Medline: 33872436]
58. Raskind IG, Shelton RC, Comeau DL, Cooper HL, Griffith DM, Kegler MC. A review of qualitative data analysis practices in health education and health behavior research. Health Educ Behav 2019 Feb;46(1):32-39 [FREE Full text] [doi: 10.1177/1090198118795019] [Medline: 30227078]

Abbreviations

CVD: cardiovascular disease
HRV: heart rate variability
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and others
LIVE: Learning in a Virtual Environment

Edited by A Mavragani; submitted 13.02.24; peer-reviewed by ASW Chan , C Ni, X Tao; comments to author 08.03.24; revised version received 24.04.24; accepted 12.06.24; published 22.08.24.

Please cite as:
Ramos SR, Reynolds H, Johnson C, Melkus G, Kershaw T, Thayer JF, Vorderstrasse A
Perceptions of HIV-Related Comorbidities and Usability of a Virtual Environment for Cardiovascular Disease Prevention Education in Sexual Minority Men With HIV: Formative Phases of a Pilot Randomized Controlled Trial
J Med Internet Res 2024;26:e57351
URL: https://www.jmir.org/2024/1/e57351
doi:10.2196/57351
PMID:38924481

©S Raquel Ramos, Harmony Reynolds, Constance Johnson, Gail Melkus, Trace Kershaw, Julian F Thayer, Allison Vorderstrasse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Retention in HIV Primary Care Using a Web-Based Patient Engagement Platform: Multistate Case-Control Study

Adam Carl Sukhija-Cohen[1], MPH, PhD; Henna Patani[2], MPH; Michael Foxworth Blasingame[3], PhD; Kathy Linh Vu[4], MPH; Ramin Bastani[3], BA

[1]Sutter Health, Palo Alto Medical Foundation Research Institute, Palo Alto, CA, United States
[2]Public Health Division, AIDS Healthcare Foundation, Los Angeles, CA, United States
[3]Healthvana, Los Angeles, CA, United States
[4]Research Triangle Institute, Research Triangle Park, CA, United States

Corresponding Author:
Adam Carl Sukhija-Cohen, MPH, PhD
Sutter Health
Palo Alto Medical Foundation Research Institute
795 El Camino Real
Ames Building
Palo Alto, CA, 94301
United States
Phone: 1 (650) 330 5963
Email: adam.sukhija-cohen@sutterhealth.org

Abstract

Background: Digital interventions to improve retention in HIV care are critical to ensure viral suppression and prevent further transmission. AIDS Healthcare Foundation Healthcare Centers are centers across the United States that provide primary HIV care. Traditionally, the Healthcare Centers conduct phone calls with patients to schedule and confirm appointments, as well as share laboratory results. In 2017, Healthvana piloted a digital platform at AIDS Healthcare Foundation Healthcare Centers to send patients SMS text message appointment reminders and allow patients to review their upcoming appointment and view their laboratory results in the web-based patient portal.

Objective: A national implementation in 15 US states and Washington, DC, of this digital intervention pilot by Healthvana aims to determine whether SMS appointment reminders and web-based patient portal logins improved retention in care compared to traditional methods.

Methods: A retrospective analysis of 40,028 patients living with HIV was conducted at the 61 AIDS Healthcare Foundation Healthcare Centers between January 2, 2017, and May 22, 2018. Patients were invited to enroll in Healthvana's digital intervention pilot, allowing for a natural, organization-wide case-control study. Separate binary logistic regression models evaluated the relationship between receiving SMS appointment reminders and completing scheduled appointments, as well as the relationship between logging into the web-based patient portal and completing scheduled appointments. Four scheduled consecutive appointments for each patient were included in the analysis to account for 1 full year of data per patient.

Results: Patients who received the SMS appointment reminder were 1.7 times more likely to complete appointment 1 compared to patients who did not receive the SMS appointment reminder ($P<.001$). In addition, patients who received the SMS appointment reminder were 1.6 times more likely to complete appointment 2 ($P<.001$), 1.7 times more likely to complete appointment 3 ($P<.001$), and 1.8 times more likely to complete appointment 4 ($P<.001$) compared to patients who did not receive the SMS appointment reminder. Patients who logged in to the web-based patient portal prior to their scheduled appointment were 7.4 times more likely to complete appointment 1 compared to patients who did not log in ($P<.001$). In addition, patients who logged in to the web-based patient portal prior to their scheduled appointment were 3.6 times more likely to complete appointment 2 ($P<.001$), 3.2 times more likely to complete appointment 3 ($P<.001$), and 2.8 times more likely to complete appointment 4 ($P<.001$) compared to patients who did not log in.

Conclusions: HIV primary care appointment completion was higher when patients engaged with Healthvana's digital platform. Digital technology interventions to ensure patients complete their scheduled HIV care appointments are imperative to curb the HIV epidemic.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e64092) doi:10.2196/64092

KEYWORDS

HIV; primary health care; retention in care; digital technology; appointments and schedules

Introduction

There are 1.2 million people living with HIV in the United States [1]. A public health model called the HIV care continuum represents 4 important stages of HIV care [2]. Across the continuum, 87% of people living with HIV are diagnosed, 66% have received HIV medical care, 50% are retained in care, and 57% have achieved viral suppression.

People living with HIV retained in care—that is, those who have continuously scheduled medical care, defined as having 2 or more cluster of differentiation 4 (CD4) or viral load tests at least 3 months apart—are at the lowest rate among the 4 steps [1]. People living with HIV who are aware of their HIV status and are either not in care or not virally suppressed account for nearly two-thirds of new infections [3]. Therefore, implementing innovative interventions is important to retain people living with HIV in care, increasing their likelihood of achieving viral suppression and preventing further transmissions [2-5].

AIDS Healthcare Foundation (AHF) provides HIV primary care at 61 AHF Healthcare Centers in 15 states (California, Florida, Georgia, Illinois, Indiana, Louisiana, Maryland, Mississippi, Nevada, New York, Ohio, Pennsylvania, South Carolina, Texas, and Washington) and Washington, DC. These are all priority states with a disproportionate burden of new HIV diagnoses according to the Ending the HIV Epidemic in the US initiative [6]. The centers operate in major metropolitan areas and offer standardized care.

Traditionally, AHF Healthcare Centers call patients to schedule and confirm appointments and share laboratory results. Beginning in 2017, an AHF/Healthvana collaboration pilot-tested a digital platform that sends patients SMS text message appointment reminders and allows patients to access a web-based patient portal. Within the portal, Healthvana features customized laboratory result delivery, 2-way communication between patients and health care providers, and appointment management. This type of digital intervention has been shown to improve HIV care across the care continuum [7-11]. However, few studies have used large sample sizes to evaluate the effect of digital intervention on HIV care retention. This pilot study aims to determine whether SMS appointment reminders and patient portal logins improved retention in care nationally compared to traditional methods.

Methods

A retrospective analysis of all AHF patients was conducted at the 61 AHF Healthcare Centers between January 2, 2017, and May 22, 2018. AHF recommendations call for patients to complete an appointment every 3 months. Therefore, 4 scheduled consecutive appointments for each patient are included in the analysis to account for 1 full year of data per patient.

Patients were invited to enroll in Healthvana's digital intervention pilot test, allowing for a natural, organization-wide case-control study. Demographics retrospectively extracted from the entire patient population included age group, gender, race, and health insurance type. The 2 predictor variables were SMS appointment reminders and web-based patient portal login. Patients in the "SMS appointment reminder" category were those who affirmed receiving an SMS reminder before their scheduled appointment. Only patients who consented to receiving SMS reminders received them; those who did not consent or opted out had SMS reminders removed entirely. Patients in the "web-based patient portal login" category were those who logged in to the portal before their scheduled appointment. As part of AHF Healthcare Center services, all patients create a username and password for Healthvana's patient portal; however, patients are not required to log in as part of their appointment. The outcome variable, "scheduled appointment," is categorized as "completed" versus "not completed." Logistic regression models were built to determine the odds of patients completing their appointments.

Ethical Considerations

Healthvana securely collects, stores, and maintains AHF health data in encrypted format both in transit and at rest and is Health Insurance Portability and Accountability Act–compliant. For this evaluation, data were extracted and deidentified from Healthvana's digital platform by Healthvana staff. All analyses were conducted by the AHF research team in SAS (version 9.4; SAS Institute). As a quality improvement pilot study requested by AHF leadership, this retrospective analysis received exemption from the Western Institutional Review Board.

Results

Overview

Table 1 depicts demographic characteristics. Most patients scheduled appointment 1 (n=40,028), followed by appointment 2 (n=37,878), appointment 3 (n=34,866), and appointment 4 (n=31,550). Most patients in appointment 1 were aged 50-54 (n=5418, 13.5%) or 25-29 years (n=5226, 13.1%), male (n=32,756, 81.8%), and White (n=21,633, 54%) or Black/African American (n=17,361, 43.4%). Most patients had private insurance (n=22,401, 56%), followed by insurance provided by the Ryan White HIV/AIDS Program (n=9121, 22.8%) or Medicare/Medicaid (n=8506, 21.3%). Distributions of age group, gender, and race did not change significantly across the 4 appointments (χ^2 distribution tests; $P>.05$ for all groups). However, insurance type distribution differed significantly across the 4 appointments: private insurance decreased, and Medicare/Medicaid and Ryan White increased over time (χ^2 distribution test; $P<.001$).

Table 2 shows appointment outcomes based on SMS reminders before scheduled appointments. A total of 1.7% (664/40,028)

XSL•FO
RenderX

of patients received reminders for appointment 1, 13.3% (5037/37,878) for appointment 2, 22.1% (7698/34,866) for appointment 3, and 28.3% (8944/31,550) for appointment 4.

Table 3 describes the patients' appointment outcomes based on whether they logged in to the web-based patient portal; 19.2% (7705/40,028) logged in for appointment 1, 22.7% (8591/37,878) for appointment 2, 25.6% (8913/34,866) for appointment 3, and 26.8% (8469/31,550) for appointment 4. Overall, 71.3% (28,540/40,028) of patients who scheduled appointment 1 completed the appointment. A total of 71.1% (28,004/39,364) of patients who did not receive the SMS appointment reminder completed appointment 1; in contrast, 80.7% (536/664) of patients who received the SMS reminder completed appointment 1. By patient portal, 66% (21,338/32,323) of patients who did not log in to the patient portal completed appointment 1; on the other hand, 93.5% (7202/7,705) of patients who logged in to the patient portal prior to their scheduled appointment completed appointment 1. For each subsequent scheduled appointment, patients who either received the SMS appointment reminder or who logged in to the patient portal prior to their scheduled appointment had a higher proportion of scheduled appointment completions compared to patients who either did not receive the SMS appointment reminder or did not log in to the patient portal.

In separate logistic regression models (Table 4), patients who received the SMS appointment reminder before their scheduled appointment were 1.7 times more likely to complete appointment 1 compared to patients who did not receive the reminder (odds ratio [OR] 1.7, 95% CI 1.4-2.1; $P<.001$); the likelihood increased 1.6 times for appointment 2 (OR 1.6, 95% CI 1.5-1.7; $P<.001$), 1.7 times for appointment 3 (OR 1.7, 95% CI 1.6-1.8; $P<.001$), and 1.8 times for appointment 4 (OR 1.8, 95% CI 1.6-1.8; $P<.001$). Patients who logged in to the patient portal before their scheduled appointment were 7.4 times more likely to complete appointment 1 compared to patients who did not log in (OR 7.4, 95% CI 6.7-8.1; $P<.001$); the likelihood increased 3.6 times for appointment 2 (OR 3.6, 95% CI 3.4-3.8; $P<.001$), 3.2 times for appointment 3 (OR 3.2, 95% CI 3.0-3.4; $P<.001$), and 2.8 times for appointment 4 (OR 2.8, 95% CI 2.7-3.0; $P<.001$).

XSL•FO
RenderX

Table 1. AIDS Healthcare Foundation (AHF) patient demographics by scheduled AHF Healthcare Center appointment, January 2, 2017, to May 22, 2018.

	Appointment 1 (n=40,028), n (%)	Appointment 2 (n=37,878), n (%)	Appointment 3 (n=34,866), n (%)	Appointment 4 (n=31,550), n (%)
Age group (years)				
18-24	2407 (6)	2219 (5.9)	1956 (5.6)	1715 (5.4)
25-29	5226 (13.1)	4885 (12.9)	4405 (12.6)	3886 (12.3)
30-34	4993 (12.5)	4686 (12.4)	4296 (12.3)	3886 (12.3)
35-39	4415 (11)	4180 (11)	3822 (11)	3443 (10.9)
40-44	3929 (9.8)	3709 (9.8)	3436 (9.9)	3116 (9.9)
45-49	4614 (11.5)	4399 (11.6)	4093 (11.7)	3713 (11.8)
50-54	5418 (13.5)	5186 (13.7)	4843 (13.9)	4420 (14)
55-59	4254 (10.6)	4070 (10.7)	3810 (10.9)	3503 (11.1)
60-64	2606 (6.5)	2510 (6.6)	2328 (6.7)	2156 (6.8)
≥65	2166 (5.4)	2034 (5.4)	1877 (5.4)	1712 (5.4)
Gender				
Female	7228 (18.1)	6738 (17.8)	6177 (17.7)	5592 (17.7)
Male	32,756 (81.8)	31,096 (82.1)	28,646 (82.2)	25,917 (82.1)
Transgender	44 (0.1)	44 (0.1)	43 (0.1)	41 (0.1)
Race				
American Indian/Alaska Native	87 (0.2)	86 (0.2)	83 (0.2)	72 (0.2)
Asian	790 (2)	728 (1.9)	659 (1.9)	579 (1.8)
Black/African American	17,361 (43.4)	16,586 (43.8)	15,361 (44)	14,016 (44.4)
Multiracial	71 (0.2)	65 (0.2)	57 (0.2)	50 (0.2)
Native Hawaiian/Pacific Islander	86 (0.2)	83 (0.2)	80 (0.2)	72 (0.2)
White	21,633 (54)	20,330 (53.7)	18,626 (53.4)	16,761 (53.1)
Insurance type				
Medicare/Medicaid	8506 (21.3)	8189 (21.6)	7693 (22.1)	7090 (22.5)
Private insurance	22,401 (56)	20,889 (55.1)	18,973 (54.4)	16,973 (53.8)
Ryan White HIV/AIDS Program	9121 (22.8)	8800 (23.2)	8200 (23.5)	7487 (23.7)

XSL•FO
RenderX

Table 2. Outcome of AIDS Healthcare Foundation Healthcare Center patient appointments by SMS appointment reminder and appointment number, January 2, 2017, to May 22, 2018.

Appointment	Did not receive the SMS appointment reminder, n/N (%)	Received the SMS appointment reminder, n/N (%)	Total, n/N (%)
Appointment 1			
Completed	28,004/39,364 (71.1)	536/664 (80.7)	28,540/40,028 (71.3)
Not completed	11,360/39,364 (28.9)	128/664 (19.3)	11,488/40,028 (28.7)
Total	39,364/40,028 (98.3)	664/40,028 (1.7)	__[a]
Appointment 2			
Completed	20,510/32,841 (62.5)	3643/5037 (72.3)	24,153/37,878 (63.8)
Not completed	12,331/32,841 (37.5)	1394/5037 (27.7)	13,725/37,878 (36.2)
Total	32,841/37,878 (86.7)	5037/37,878 (13.3)	—
Appointment 3			
Completed	15,578/27,168 (57.3)	5355/7698 (69.6)	20,933/34,866 (60)
Not completed	11,590/27,168 (42.7)	2343/7698 (30.4)	13,933/34,866 (40)
Total	27,168/34,866 (77.9)	7698/34,866 (22.1)	—
Appointment 4			
Completed	12,317/22,606 (54.5)	6054/8944 (67.7)	18,371/31,550 (58.2)
Not completed	10,289/22,606 (45.5)	2890/8944 (32.3)	13,179/31,550 (41.8)
Total	22,606/31,550 (71.7)	8944/31,550 (28.3)	—

[a]Not applicable.

Table 3. Outcome of AIDS Healthcare Foundation Healthcare Center patient appointments by web-based patient portal login and appointment number, January 2, 2017, to May 22, 2018.

Appointment	Did not log in to the web-based patient portal, n/N (%)	Logged in to the web-based patient portal, n/N (%)	Total, n/N (%)
Appointment 1			
Completed	21,338/32,323 (66)	7202/7705 (93.5)	28,540/40,028 (71.3)
Not completed	10,985/32,323 (34)	503/7705 (6.5)	11,488/40,028 (28.7)
Total	32,323/40,028 (80.8)	7705/40,028 (19.2)	__[a]
Appointment 2			
Completed	17,004/29,287 (58.1)	7149/8591 (83.2)	24,153/37,878 (63.8)
Not completed	12,283/29,287 (41.9)	1442/8591 (16.8)	13,725/37,878 (36.2)
Total	29,287/37,878 (77.3)	8591/37,878 (22.7)	—
Appointment 3			
Completed	13,929/25,953 (53.7)	7004/8913 (78.6)	20,933/34,866 (60)
Not completed	12,024/25,953 (46.3)	1909/8913 (21.4)	13,933/34,866 (40)
Total	25,953/34,866 (74.4)	8913/34,866 (25.6)	—
Appointment 4			
Completed	11,993/23,081 (52)	6378/8469 (75.3)	18,371/31,550 (58.2)
Not completed	11,088/23,081 (48)	2091/8469 (24.7)	13,179/31,550 (41.8)
Total	23,081/31,550 (73.2)	8469/31,550 (26.8)	—

[a]Not applicable.

XSL•FO
RenderX

Table 4. Odds ratios (ORs) for completing scheduled appointments at AIDS Healthcare Foundation Healthcare Centers by SMS appointment reminder, web-based patient portal login, and appointment number, January 2, 2017, to May 22, 2018.

Appointment completed	Received the SMS appointment reminder		Logged in to the web-based patient portal	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Appointment 1	1.7 (1.4-2.1)	<.001	7.4 (6.7-8.1)	<.001
Appointment 2	1.6 (1.5-1.7)	<.001	3.6 (3.4-3.8)	<.001
Appointment 3	1.7 (1.6-1.8)	<.001	3.2 (3.0-3.4)	<.001
Appointment 4	1.8 (1.6-1.8)	<.001	2.8 (2.7-3.0)	<.001

Discussion

Healthvana's digital platform was associated with AHF patients completing their scheduled AHF Healthcare Center appointments. Patients who received the SMS appointment reminder and patients who logged into the web-based patient portal prior to their scheduled appointments were more likely to complete their scheduled appointments compared to patients who did not receive the SMS appointment reminder or did not log in to the patient portal. These findings suggest engaging patients in HIV care using a digital platform can help improve retention—a critical step in the HIV care continuum.

There are 2 strengths in this study. First, data across 15 states and Washington, DC, were analyzed, increasing generalizability and reducing geographic bias. Second, this study includes a very large sample size of patients, enabling us to rule out spurious associations. Within the current body of research, studies are often limited to 1 or 2 jurisdictions and have small sample sizes [7-11].

This study includes important limitations. Healthvana's digital platform was pilot-tested at different time points during the study duration; this study did not distinguish whether patients had the option to engage with the digital platform. In addition, this study did not include socioeconomic factors that may affect an individual's access to a mobile phone to receive SMS appointment reminders or a smartphone or computer to log in to the patient portal. Lastly, this study did not assess the reasons patients' appointments were not completed.

Clinic-level barriers, including understaffing, can undermine HIV prevention and care at all stages of the HIV care continuum [12,13]. The capacity of staff to engage with people living with HIV in timely and consistent intervals is an important structural factor digital platforms can address. This study found significant improvements in scheduled appointment completion with a large patient population. Future HIV research should continue to focus on digital platforms that can improve retention in care and viral suppression across HIV care settings and insurance coverage (ie, federally qualified health centers and private primary care clinics).

Data Availability

The data set analyzed during this study is not publicly available due to data protections for people living with HIV. However, the data are available on reasonable request.

Conflicts of Interest

None declared.

References

1. U.S. statistics. HIV.gov. 2024 Aug 15. URL: https://www.hiv.gov/hiv-basics/overview/data-and-trends/statistics [accessed 2024-09-26]
2. HIV care continuum. HIV.gov. 2022 Oct 28. URL: https://www.hiv.gov/federal-response/policies-issues/hiv-aids-care-continuum [accessed 2024-09-09]
3. Li Z, Purcell DW, Sansom SL, Hayes D, Hall HI. Vital signs: HIV transmission along the continuum of care - United States, 2016. MMWR Morb Mortal Wkly Rep 2019 Mar 22;68(11):267-272 [FREE Full text] [doi: 10.15585/mmwr.mm6811e1] [Medline: 30897075]
4. Gopalsamy S, Shah N, Marconi V, Armstrong W, Del Rio C, Pennisi E, et al. The impact of churn on HIV outcomes in a southern United States clinical cohort. Open Forum Infect Dis 2022 Jul;9(7):ofac338 [FREE Full text] [doi: 10.1093/ofid/ofac338] [Medline: 35899283]
5. López Rios J, Lentz C, Balán IC, Grosskopf N, D'Angelo A, Stief M, et al. Engagement in care among newly diagnosed HIV-positive gay, bisexual, and other men who have sex with men in the United States: results from the Together 5,000 study. AIDS Educ Prev 2022 Oct;34(5):349-364 [FREE Full text] [doi: 10.1521/aeap.2022.34.5.349] [Medline: 36181497]
6. EHE overview. Office of Infectious Disease and HIV/AIDS Policy, U.S. Department of Health and Human Services. 2023 Dec 04. URL: https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview [accessed 2024-09-09]

XSL•FO
RenderX

7. Ngowi K, Pima F, Mmbaga BT, Aarnoutse RE, Reiss P, Nieuwkerk PT, et al. "I wish to continue receiving the reminder short messaging service": a mixed methods study on the acceptability of digital adherence tools among adults living with HIV on antiretroviral treatment in Tanzania. Patient Prefer Adherence 2021;15:559-568 [FREE Full text] [doi: 10.2147/PPA.S290079] [Medline: 33727801]
8. Che Pa MF, Makmor-Bakry M, Islahudin F. Digital health in enhancing antiretroviral therapy adherence: a systematic review and meta-analysis. AIDS Patient Care STDS 2023 Nov;37(11):507-516. [doi: 10.1089/apc.2023.0170] [Medline: 37956244]
9. Canan CE, Waselewski ME, Waldman ALD, Reynolds G, Flickinger TE, Cohn WF, et al. Long term impact of PositiveLinks: clinic-deployed mobile technology to improve engagement with HIV care. PLoS One 2020;15(1):e0226870 [FREE Full text] [doi: 10.1371/journal.pone.0226870] [Medline: 31905209]
10. Wang Z, Zhu Y, Cui L, Qu B. Electronic health interventions to improve adherence to antiretroviral therapy in people living with HIV: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Oct 16;7(10):e14404 [FREE Full text] [doi: 10.2196/14404] [Medline: 31621641]
11. Arayasirikul S, Turner C, Trujillo D, Le V, Wilson EC. Efficacy and impact of digital HIV care navigation in young people living with HIV in San Francisco, California: prospective study. JMIR Mhealth Uhealth 2020 May 08;8(5):e18597 [FREE Full text] [doi: 10.2196/18597] [Medline: 32383680]
12. Dutta S, Pulsifer BH, Dance KV, Leue EP, Beaupierre M, Lowman K, et al. Clinic-level complexities prevent effective engagement of people living with HIV who are out-of-care. PLoS One 2024;19(5):e0304493 [FREE Full text] [doi: 10.1371/journal.pone.0304493] [Medline: 38820362]
13. Ritchwood TD, Ba A, Ingram L, Atujuna M, Marcus R, Ntlapo N, et al. Community perspectives of South African adolescents' experiences seeking treatment at local HIV clinics and how such clinics may influence engagement in the HIV treatment cascade: a qualitative study. AIDS Care 2020 Jan;32(1):83-88 [FREE Full text] [doi: 10.1080/09540121.2019.1653442] [Medline: 31402674]

Abbreviations

AHF: AIDS Healthcare Foundation
CD4: cluster of differentiation 4
OR: odds ratio

Edited by A Mavragani; submitted 08.07.24; peer-reviewed by HJ Gilmore, M Stoner; comments to author 02.08.24; revised version received 13.08.24; accepted 16.09.24; published 02.10.24.

Please cite as:
Sukhija-Cohen AC, Patani H, Blasingame MF, Vu KL, Bastani R
Retention in HIV Primary Care Using a Web-Based Patient Engagement Platform: Multistate Case-Control Study
J Med Internet Res 2024;26:e64092
URL: https://www.jmir.org/2024/1/e64092
doi:10.2196/64092
PMID:

©Adam Carl Sukhija-Cohen, Henna Patani, Michael Foxworth Blasingame, Kathy Linh Vu, Ramin Bastani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Accuracy, Quality, and Misinformation of YouTube Abortion Procedural Videos: Cross-Sectional Study

Nicole Acero[1,2], MD; Emma Herrero[3], BS; Juanita Foncham[4], MD, MPH, MBA; Jamie McIlvaine[5], MD; Emre Kayaalp[6], MD, PhD; Melissa Figueora[1,4], MD, MPH; Antonia Francis Oladipo[1,4], MD, MSCI

[1]Hackensack Meridian School of Medicine, Nutley, NJ, United States

[2]Department of Family Medicine, Boston University School of Medicine, Boston Medical Center, Boston, MA, United States

[3]Syracuse University, Syracuse, NY, United States

[4]Department of Obstetrics and Gynecology, Hackensack University Medical Center, Hackensack, NJ, United States

[5]Department of Obstetrics and Gynecology, Jersey City Medical Center, Rutgers Health, Jersey City, NJ, United States

[6]Department of Obstetrics and Gynecology, Overlook Medical Center, Atlantic Health System, Summit, NJ, United States

Corresponding Author:
Nicole Acero, MD
Department of Family Medicine
Boston University School of Medicine
Boston Medical Center
Dowling Bldg, 5th Fl
One Boston Medical Center Pl
Boston, MA, 02116
United States
Phone: 1 617 638 8000
Email: nicole.acero@bmc.org

Abstract

Background: The internet is often the first source patients turn to for medical information. YouTube is a commonly used internet-based resource for patients seeking to learn about medical procedures, including their risks, benefits, and safety profile. Abortion is a common yet polarizing medical procedure. People interested in obtaining an abortion are likely to use the internet to learn more about abortion procedures and may encounter misinformed and biased information. This is troubling as information found on the internet can significantly alter perceptions and understanding of these procedures. There is no current research that evaluates the accuracy, quality, and misinformation of instructional abortion videos available to patients.

Objective: The purpose of this study was to assess if any given video can deliver accurate and quality information about this topic in an unbiased manner and to assess the level of factually incorrect, distorted, or medically irrelevant information in any given video.

Methods: Procedural methods of abortion were queried on YouTube on August 22, 2022. The videos were screened with strict exclusion criteria. Videos were categorized into “video slants” based on the language and attitudes expressed in each video. Video accuracy was calculated using the Surgical Curriculum in Obstetrics and Gynecology (SCOG) checklist for each corresponding procedure. Video quality was calculated using the Laparoscopic Surgery Video Educational Guidelines (LAP-VEGaS) criteria. The level of misinformation was assessed with the evidence-based Anti-Choice Rubric, which scores the amount of factually incorrect, distorted, or medically irrelevant information in each video.

Results: A total of 32 videos were analyzed and categorized into 3 “video slant” groups: neutral (n=23, 72%), antichoice (n=4, 12%), and prochoice (n=5, 16%). Using the SCOG checklist, neutral videos had the highest median accuracy (45.9%), followed by antichoice videos (24.6%) and prochoice videos (18.5%). None of the videos met the LAP-VEGaS quality control criteria, (score>11, indicating adequate quality). Neutral videos had a median score of 8.8 out of 18, with antichoice videos scoring 10.75 and prochoice videos scoring 6.2. Using the Anti-Choice Rubric, neutral videos mentioned only 1 factually incorrect piece of information. Antichoice videos mentioned 12 factually incorrect pieces of information, 8 distortions, and 3 medically irrelevant pieces of information. Prochoice videos did not mention any of the 3 themes.

Conclusions: Using the SCOG checklist, the accuracy of instructional videos were inconsistent across the 3 identified “video slants.” Using LAP-VEGaS criteria, the quality of educational videos were also inconsistent across the 3 “video slants.” Prochoice

XSL•FO
RenderX

videos had the lowest level of misinformation, with no mentions of any of the 3 themes. Antichoice videos had the highest levels of misinformation, with mentions in all 3 themes. Health care professionals should consider this when counseling patients who may watch YouTube videos for information regarding abortion procedures.

(J Med Internet Res 2024;26:e50099) doi:10.2196/50099

KEYWORDS

abortion; YouTube; social media; accuracy; quality; misinformation; reliability; obstetrics; women's health; reproductive; patient education; health information; prochoice

Introduction

The internet is often the first source patients turn to for medical information. One in 3 American adults have used the internet to better understand a medical condition, with nearly half seeking information about specific procedures [1]. Abortion is one of the most common medical procedures globally [2]. People interested in obtaining an abortion are likely to use the internet to learn more about abortion procedures, including their safety and complications [3]. Internet searches on abortion-related terms spiked following the leaked Supreme Court opinion in Dobbs v. Jackson Women's Health Organization in May 2022 [4].

Procedural abortion includes manual vacuum aspiration, electric vacuum aspiration (colloquially referred to as dilation and curettage ["D&C"]), and dilation and evacuation ("D&E"). Scientific evidence demonstrates that induced abortion is safe. While the risks of abortion increase with gestational age, abortion at any gestational age is safer than childbirth [5]. Many antichoice arguments rest on misinformation regarding the safety and health consequences of abortion. Additionally, false or misleading information on the internet about abortion exerts an influence on public debates and policies [6].

Abortion knowledge in the United States is low overall. There is a large gap between the scientific evidence of abortion safety and public perception, and political beliefs are not significantly associated with abortion knowledge [7]. People seeking information about the safety and potential risks of abortion are likely to encounter a substantial volume of misinformed and biased information. This is troubling as information found on the internet can significantly alter perceptions and understanding of procedures [3].

YouTube is commonly used as a freely accessible internet-based resource for patients seeking information about medical procedures, including their risks, benefits, and safety profiles. One study found that 7% of participants surveyed used YouTube to aid their search on abortion safety [3]. There is no current research that evaluates the accuracy, quality, or level of misinformation of instructional videos available to patients. Given the reliance of information gathered on the internet to inform medical decision-making, this study was designed to assess whether any particular video on YouTube can deliver accurate, quality, and factually correct information about this topic in an unbiased manner.

Methods

YouTube Search

Surgical methods of abortion were used as a search query on YouTube in incognito mode, so the searcher's own algorithm would not affect the results. Search terms included "surgical abortion," "video demonstration + manual vacuum aspiration/D&C/D&E," "how to perform abortion manual vacuum aspiration OR D&C," "how to perform a D&E," and "manual vacuum aspiration/D&C/D&E + surgery." Videos were excluded if they were not in English, duplicate of another video, noninstructional in nature, or unrelated in topic.

Ethical Considerations

The creators of the videos analyzed in this study have been deidentified. No ethics approval was applied for, as this study did not involve human subjects.

Video Categorization

Videos were categorized into 3 "video slant" groups: neutral, antichoice, and prochoice, based on the methodology by Han et al [2]. Videos were determined to have a prochoice or antichoice bias if either (1) information was given in a biased or dramatic fashion or (2) the video displayed an opinion regarding the provision of abortion and its legality. Otherwise, the video was considered neutral. We assigned a slant clarity rating based on how easy it was to discern the video slant: (1) obvious, (2) in between, or (3) difficult to tell.

Data Analysis

Each video's accuracy was calculated based on how many steps on the corresponding Surgical Curriculum in Obstetrics and Gynecology (SCOG) checklist were included in each video (Multimedia Appendix 1). Each video's quality was calculated using the LAP-VEGaS criteria (Multimedia Appendix 2) [8]. There is significant correlation between a score of ≥11 out of 18 on the LAP-VEGaS tool and recommended acceptance for publication, indicating adequate quality of educational content. The Anti-Choice Rubric was created to grade the level of misinformation within each video. The rubric was created based on talking points gathered from antichoice organizations and categorized into 3 themes: factually incorrect, distortion, and medically irrelevant information (Multimedia Appendix 3). Videos were scored for the number of times a given viewpoint was mentioned.

XSL•FO
RenderX

Results

Video Characteristics

After removing duplicates, our search yielded 37 videos. Of those, 32 videos met inclusion criteria and were analyzed (Table 1). Videos were categorized into 3 "video slant" groups: neutral (n=23, 72%), antichoice (n=4, 12%), and prochoice (n=5, 16%). Clarity was obvious (n=26, 81%), difficult to tell (n=4, 12%), or in between (n=2, 6%). Of the 4 videos deemed "difficult to tell," 3 (75%) had a prochoice slant. The 2 videos deemed "in between" were neutral.

Table 1. YouTube video characteristics.

Characteristics	All videos (n=32)	Neutral videos (n=23)	Antichoice videos (n=4)	Prochoice videos (n=5)
Creator type, n (%)				
Physician	21 (66)	15 (65)	3 (75)	4 (80)
Nonphysician	5 (16)	4 (17)	1 (25)	1 (20)
Unspecified	6 (19)	4 (17)	0 (0)	0 (0)
Video type, n (%)				
Live	7 (22)	7 (30)	0 (0)	0 (0)
Digital reconstruction	5 (16)	3 (13)	2 (50)	0 (0)
Model or simulation	3 (9)	2 (9)	1 (25)	0 (0)
Combined	4 (13)	4 (17)	0 (0)	0 (0)
Lecture with images	12 (38)	7 (30)	1 (25)	4 (80)
Lecture without images	1 (3)	0 (0)	0 (0)	1 (20)
Channel type, n (%)				
Education	10 (31)	7 (30)	0 (0)	3 (60)
Physician	8 (25)	7 (30)	0 (0)	1 (20)
Personal	7 (22)	6 (26)	1 (25)	0 (0)
Clinic	1 (3)	1 (4)	0 (0)	0 (0)
Medical	1 (3)	1 (4)	0 (0)	0 (0)
Other	5 (16)	1 (4)	3 (75)	1 (20)
Country, n (%)				
United States	8 (25)	4 (17)	3 (75)	1 (20)
England	2 (6)	1 (4)	0 (0)	1 (20)
Pakistan	2 (6)	2 (9)	0 (0)	0 (0)
India	3 (9)	2 (9)	0 (0)	1 (20)
New Zealand	1 (3)	0 (0)	1 (25)	0 (0)
Moldova	1 (3)	0 (0)	0 (0)	1 (20)
Ethiopia	1 (3)	1 (4)	0 (0)	0 (0)
Kenya	1 (3)	1 (4)	0 (0)	0 (0)
Unspecified	13 (41)	12 (52)	0 (0)	1 (20)
Views, median (IQR)	1,476,321 (3689.5-5,061,341)	94,465 (22,108-608,745)	977,863.5 (56,619-5,111,341)	526 (207-2928)
Likes, median (IQR)	270.5 (22-1100)	492 (165-460)	373.5 (29.5-5844)	7 (0-55,008.5)

Evaluation Outcomes

Using the SCOG checklist, neutral videos had the highest median accuracy (45.9%), followed by antichoice videos (24.6%) and prochoice videos (18.5%).

Of the 32 videos, none met the LAP-VEGaS quality control criteria. Neutral videos had a median score of 8.8 out of 18, with antichoice videos scoring 10.75 and prochoice videos scoring 6.2.

Using the Anti-Choice Rubric, neutral videos mentioned 1 factually incorrect piece of information. Antichoice videos

XSL•FO
RenderX

mentioned 12 factually incorrect pieces of information, 8 distortions, and 3 medically irrelevant pieces of information. Prochoice videos did not mention any of the 3 themes.

Discussion

Principal Findings

Our study of YouTube videos providing information about procedural abortions found that a majority (23/32, 72%) were neutral. Based on the SCOG checklist, neutral videos were the most accurate, followed by antichoice videos. Prochoice videos were the least accurate. Based on the LAP-VEGaS criteria, antichoice videos had the highest quality, followed by neutral videos. Prochoice videos had the lowest quality. While prochoice videos had the lowest accuracy and quality, they contained no misinformation according to our Anti-Choice Rubric. Antichoice videos contained the highest level of misinformation. The neutral videos had 1 mention of factually incorrect information.

Based on these findings, we can conclude that neutral videos are more reliable than either antichoice or prochoice videos. Antichoice videos may be of higher quality and more accurately reflect a procedure compared to prochoice videos. However, they also contained the largest amount of misinformation. Prochoice videos, in contrast, omitted more steps involved in a given procedure but contained no misinformation. Therefore, accuracy and quality alone cannot be used to determine whether any given video is trustworthy.

One limitation of our analysis was the small sample size, which limited the ability to perform association analysis. Additionally, this study offers a snapshot of the videos available at the time of our analysis. Furthermore, most of our researchers are involved in abortion-related research. We recognize that nonexperts may have rated and classified these videos differently.

The strengths of this study include the use of validated tools, namely, the SCOG checklist and the LAP-VEGaS criteria, to measure accuracy and quality, respectively.

Comparison to Prior Work

To our knowledge, there is no published research evaluating the trustworthiness of individual YouTube procedural abortion videos for their informational content. Quality analyses of YouTube videos have been conducted in urology and obstetrics/gynecology, although there are no known studies on abortion procedures [9,10]. A similar work by Han et al [2] sought to describe the trustworthiness of web-based information on abortion safety and risks of depression and infertility following abortion. This study determined trustworthiness by factors such as content and source of content. Our study sought to determine the trustworthiness of any given video based on accuracy, quality, and level of misinformation. Accuracy and quality were determined using 2 validated tools (the SCOG checklist and LAP-VEGaS criteria, respectively). The level of misinformation was determined by an evidence-based rubric. This study shows that accuracy, creator type, references, or affiliation with health care organizations or health care professionals alone cannot determine whether any given video is trustworthy.

Conclusions

Our results show that the accuracy or quality of a video alone are not reliable factors in determining trustworthiness. Videos that appear trustworthy based on these factors alone may contain language or imagery that provides misinformation or lends itself to misinterpretation, despite also containing accurate technical information. Furthermore, a given video's slant may provide motivation to either downplay elements of a procedure to not scare a patient away from a procedure or emphasize the details of the procedure for the opposite reason.

A shared decision-making model combines provider clinical expertise with patient values. It is the job of clinical providers to educate their patients, and it is the patients' responsibility to use that information in combination with their values to make an informed decision about their health. Therefore, it is important for providers to be aware of the type of information patients may encounter when making decisions to address patient concerns. It is also important for providers to be able to direct patients toward accurate and quality information to assist patients in their own self-education processes. In the future, we hope to use the information from these findings to develop a tool for patients and physicians to grade the trustworthiness of informational content in real time. We hope the use of this tool will encourage content creators to develop content that is not just impactful but—more importantly—truthful.

Acknowledgments

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. *JMIR Publications* provided article processing fee (APF) support for the publication of this article.

Data Availability

The dataset used in this study is available in Multimedia Appendix 4.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
The Surgical Curriculum in Obstetrics and Gynecology (SCOG) checklist for procedural abortions, which was used to determine the accuracy of a YouTube video.
[PDF File (Adobe PDF File), 60 KB - jmir_v26i1e50099_app1.pdf]

Multimedia Appendix 2
The Laparoscopic Surgery Video Educational Guidelines (LAP-VEGaS) criteria, which were used to determine the quality of a video.
[PNG File , 132 KB - jmir_v26i1e50099_app2.png]

Multimedia Appendix 3
The evidence-based Anti-Choice Rubric, which was used to determine the level of abortion misinformation in a given video.
[PDF File (Adobe PDF File), 48 KB - jmir_v26i1e50099_app3.pdf]

Multimedia Appendix 4
Dataset of YouTube procedural abortion videos included in this cross-sectional study.
[PDF File (Adobe PDF File), 142 KB - jmir_v26i1e50099_app4.pdf]

References

1. Fox S, Duggan M. Health online 2013. Pew Research Center. 2013 Jan 15. URL: https://tinyurl.com/56yjn4x6 [accessed 2024-10-01]
2. Han L, Boniface ER, Han LY, Albright J, Doty N, Darney BG. The abortion web ecosystem: cross-sectional analysis of trustworthiness and bias. J Med Internet Res 2020 Oct 26;22(10):e20619 [FREE Full text] [doi: 10.2196/20619] [Medline: 33104002]
3. Rupert DJ, Gard Read J, Amoozegar JB, Moultrie RR, Taylor OM, O'Donoghue AC, et al. Peer-generated health information: the role of online communities in patient and caregiver health decisions. J Health Commun 2016 Nov 02;21(11):1187-1197 [FREE Full text] [doi: 10.1080/10810730.2016.1237592] [Medline: 27805496]
4. Gupta S, Perry B, Simon K. Trends in abortion- and contraception-related internet searches after the US Supreme Court overturned constitutional abortion rights: how much do state laws matter? JAMA Health Forum 2023 Apr 07;4(4):e230518 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.0518] [Medline: 37115538]
5. Raymond EG, Grimes DA. The comparative safety of legal induced abortion and childbirth in the United States. Obstet Gynecol 2012 Feb;119(2 Pt 1):215-219. [doi: 10.1097/AOG.0b013e31823fe923] [Medline: 22270271]
6. Conti J, Cahill E. Abortion in the media. Curr Opin Obstet Gynecol 2017 Dec;29(6):427-430. [doi: 10.1097/GCO.0000000000000412] [Medline: 28915157]
7. Forbes M, Darney BG, Ramanadhan S, Earp M, Waldner-James L, Han L. How do women interpret abortion information they find online? Contraception 2021 Apr;103(4):276-281. [doi: 10.1016/j.contraception.2021.01.005] [Medline: 33454372]
8. Celentano V, Smart N, Cahill RA, Spinelli A, Giglio MC, McGrath J, et al. Development and validation of a recommended checklist for assessment of surgical videos quality: the Laparoscopic Surgery Video Educational Guidelines (LAP-VEGaS) video assessment tool. Surg Endosc 2021 Mar 06;35(3):1362-1369 [FREE Full text] [doi: 10.1007/s00464-020-07517-4] [Medline: 32253556]
9. Collà Ruvolo C, Califano G, Tuccillo A, Tolentino S, Cancelliere E, Di Bello F, et al. YouTube™ as a source of information on placenta accreta: a quality analysis. Eur J Obstet Gynecol Reprod Biol 2022 May;272:82-87. [doi: 10.1016/j.ejogrb.2022.03.015] [Medline: 35290877]
10. Morra S, Napolitano L, Collà Ruvolo C, Celentano G, La Rocca R, Capece M, et al. Could YouTubeTM encourage men on prostate checks? a contemporary analysis. Arch Ital Urol Androl 2022 Sep 26;94(3):285-290 [FREE Full text] [doi: 10.4081/aiua.2022.3.285] [Medline: 36165471]

Abbreviations

LAP-VEGaS: Laparoscopic Surgery Video Educational Guidelines
SCOG: Surgical Curriculum in Obstetrics and Gynecology

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.06.23; peer-reviewed by K Biswas, L Napolitano; comments to author 25.04.24; revised version received 14.05.24; accepted 20.05.24; published 22.10.24.

Please cite as:
Acero N, Herrero E, Foncham J, McIlvaine J, Kayaalp E, Figueora M, Oladipo AF
Accuracy, Quality, and Misinformation of YouTube Abortion Procedural Videos: Cross-Sectional Study
J Med Internet Res 2024;26:e50099
URL: https://www.jmir.org/2024/1/e50099
doi:10.2196/50099
PMID:

©Nicole Acero, Emma Herrero, Juanita Foncham, Jamie McIlvaine, Emre Kayaalp, Melissa Figueora, Antonia Francis Oladipo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Use of ChatGPT to Explore Gender and Geographic Disparities in Scientific Peer Review

Paul Sebo[1], MSc, MD

University of Geneva, Geneva, Switzerland

Corresponding Author:
Paul Sebo, MSc, MD
University of Geneva
Rue Michel-Servet 1
Geneva, 1211
Switzerland
Phone: 41 223794390
Email: paul.seboe@unige.ch

Abstract

Background: In the realm of scientific research, peer review serves as a cornerstone for ensuring the quality and integrity of scholarly papers. Recent trends in promoting transparency and accountability has led some journals to publish peer-review reports alongside papers.

Objective: ChatGPT-4 (OpenAI) was used to quantitatively assess sentiment and politeness in peer-review reports from high-impact medical journals. The objective was to explore gender and geographical disparities to enhance inclusivity within the peer-review process.

Methods: All 9 general medical journals with an impact factor >2 that publish peer-review reports were identified. A total of 12 research papers per journal were randomly selected, all published in 2023. The names of the first and last authors along with the first author's country of affiliation were collected, and the gender of both the first and last authors was determined. For each review, ChatGPT-4 was asked to evaluate the "sentiment score," ranging from –100 (negative) to 0 (neutral) to +100 (positive), and the "politeness score," ranging from –100 (rude) to 0 (neutral) to +100 (polite). The measurements were repeated 5 times and the minimum and maximum values were removed. The mean sentiment and politeness scores for each review were computed and then summarized using the median and interquartile range. Statistical analyses included Wilcoxon rank-sum tests, Kruskal-Wallis rank tests, and negative binomial regressions.

Results: Analysis of 291 peer-review reports corresponding to 108 papers unveiled notable regional disparities. Papers from the Middle East, Latin America, or Africa exhibited lower sentiment and politeness scores compared to those from North America, Europe, or Pacific and Asia (sentiment scores: 27 vs 60 and 62 respectively; politeness scores: 43.5 vs 67 and 65 respectively, adjusted P=.02). No significant differences based on authors' gender were observed (all P>.05).

Conclusions: Notable regional disparities were found, with papers from the Middle East, Latin America, and Africa demonstrating significantly lower scores, while no discernible differences were observed based on authors' gender. The absence of gender-based differences suggests that gender biases may not manifest as prominently as other forms of bias within the context of peer review. The study underscores the need for targeted interventions to address regional disparities in peer review and advocates for ongoing efforts to promote equity and inclusivity in scholarly communication.

(J Med Internet Res 2024;26:e57667) doi:10.2196/57667

KEYWORDS

Africa; artificial intelligence; discrimination; peer review; sentiment analysis; ChatGPT; disparity; gender; geographic; global south; inequality; woman; assessment; researcher; communication; consultation; gender bias

Introduction

The peer-review process plays a pivotal role in validating the quality and integrity of scholarly papers. With an increasing emphasis on transparency and accountability, some journals adopted the practice of publishing peer-review reports alongside papers.

XSL•FO
RenderX

Sentiment analysis, the identification and categorization of opinions, attitudes, and emotions conveyed in text data as positive, negative, or neutral [1], finds ChatGPT (OpenAI) standing out with unique advantages compared to "traditional methods" [1-5]. Unlike lexicon-based approaches that may overlook nuanced expressions, ChatGPT understands natural language, capturing subtle nuances and context-specific sentiments. Unlike machine learning methods requiring labeled datasets, ChatGPT adapts to diverse domains without explicit training, reducing costs and time. Its humanlike responses facilitate intuitive sentiment interpretation. Additionally, artificial intelligence (AI)–driven sentiment analysis, including ChatGPT, ensures efficient, scalable (able to accommodate increasing data volumes without significant performance degradation), consistent, objective (devoid of human biases and preconceptions), and adaptable analyses. ChatGPT showed promise for sentiment analysis in several recent studies [6-10].

Verharen [11] showed that ChatGPT was accurate in determining sentiment and politeness scores in peer reviews of scientific papers. The study also showed gender inequalities, with woman authors receiving less polite reviews than men. The study, limited to papers from a single journal (*Nature Communications*), aligns with the broader issue of gender discrimination in academic medicine [12-17]. While sentiment and politeness metrics may not directly measure bias, they serve as useful proxies for identifying potential biases in peer review. Biased reviewers may exhibit tendencies toward overly positive or negative sentiments, as well as varying levels of politeness toward authors based on factors such as gender or geographic region.

Building upon Verharen's [11] work, ChatGPT-4 was used to quantitatively assess sentiment and politeness in peer-review reports from 9 high-impact medical journals, exploring gender and geographical disparities to enhance inclusivity within the peer review process. By leveraging the capabilities of AI, this study sheds light on the potential of AI technologies to mitigate biases and promote fairness in scholarly communication.

Methods

Selection of Journals, Papers, and Peer-Review Reports

The Clarivate and ASAPbio websites were searched to identify all 9 general medical journals with a journal citation reports impact factor >2 that publish peer-review reports (Table 1). None of these journals uses double-blind peer review. Using simple randomization, which involves randomly selecting papers using a random number generator, 12 research papers per journal were selected, all published in 2023, and all peer-review reports from the initial round for these 108 papers were retrieved. This sample size was chosen to ensure robust and manageable analysis.

Table 1. List of high-impact general medical journals included in this cross-sectional study using ChatGPT-4 to quantitatively assess sentiment and politeness in 291 peer-review reports for 108 papers published in 2023, as well as number of papers and reviews per journal, and median sentiment and politeness scores per journal.

Journal	2022 impact factor, n	Papers (n=108), n (%)	Reviews (n=291), n (%)	Sentiment score, median (IQR)	Politeness score, median (IQR)
BMJ	107.7	12 (11.1)	51 (17.5)	68 (53-78)	73 (60-80)
PLOS Medicine	15.8	12 (11.1)	41 (14.1)	60 (50-73)	70 (62-78)
BMC Medicine	9.3	12 (11.1)	29 (10.0)	63 (48-70)	73 (60-78)
Journal of Clinical Medicine	3.9	12 (11.1)	31 (10.7)	43 (17-70)	50 (25-67)
Diagnostics	3.6	12 (11.1)	28 (9.6)	37 (3-60)	57 (28.5-68.5)
Journal of Personalized Medicine	3.4	12 (11.1)	27 (9.3)	57 (17-70)	68 (27-82)
BMJ Open	2.9	12 (11.1)	27 (9.3)	63 (40-72)	60 (45-70)
BMC Primary Care	2.9	12 (11.1)	25 (8.6)	57 (10-67)	55 (30-70)
Medicina	2.6	12 (11.1)	32 (11.0)	48.5 (20-65)	57 (43.5-66)

Data Collection and Gender Determination

The names of the first and last authors along with the first author's country of affiliation were collected, and the gender of both the first and last authors was determined. The authors' genders were categorized in 2 steps. The gender was determined based on names alone, classifying names as man or woman accordingly. For authors whose gender could not be inferred from their names, professional networks and university websites were searched for photos or text containing gender-specific pronouns. This method enabled to assign genders to all authors. A gender detection tool (Gender API; Markus Perl IT Solutions) was also used to confirm the classifications [18]. Gender API demonstrated high accuracy in previous studies [19]. Both approaches yielded similar results, with high agreement between them (first authors: percentage agreement=0.9725, Cohen κ=0.9450; last authors: percentage agreement=0.9691, Cohen κ=0.9295).

Sentiment and Politeness Scores

For each review, ChatGPT-4 was asked to evaluate the "sentiment score," ranging from –100 (negative) to +100 (positive), and the "politeness score," ranging from –100 (rude) to +100 (polite). The measurements were repeated 5 times and the minimum and maximum values were removed. The sentiment score measures how favorable the review is, and the politeness score how polite a review's language is. The same prompt as Verharen [11] was used:

XSL•FO
RenderX

> *Below you will find a scientific peer review. Can you score this peer review on the sentiment, on a scale from –100 (negative) to 0 (neutral) to 100 (positive), and politeness of language use, on a scale of –100 (rude) to 0 (neutral) to 100 (polite)?*

All the data were collected in January 2024. The data associated with this paper are available in the Open Science Framework [20]. The database is provided as a "dta" file for use with Stata. Multimedia Appendix 1 provides a detailed list and description of the variables included in the Stata file.

Statistical Analyses

The mean sentiment and politeness scores for each review were computed and rounded to the nearest whole number. These scores were then summarized using the median and IQR, both overall and stratified by journal, gender, and affiliation. Affiliation countries were categorized into 3 regions (North America, Europe, or Pacific; Asia; and Latin America, Middle East, or Africa), following prior research [21]. Comparisons were conducted using Wilcoxon rank-sum tests (for gender) and Kruskal-Wallis rank tests (for affiliation). Negative binomial regressions were performed, adjusting for journal, affiliation, and intracluster correlation within papers [22,23]. To accommodate the requirement of nonnegative outcome variables in negative binomial regressions, 100 were added to the sentiment and politeness scores, resulting in a scale from 0 to 200. Quadratic weighted agreement coefficients were calculated to assess the agreement between the 3 sentiment and politeness score measurements. Fleiss κ was used instead of Cohen κ due to the presence of more than 2 measurements [24]. All analyses were conducted using Stata (version 15.1; StataCorp).

Ethical Considerations

Since this study did not involve the collection of personal health-related data, it did not require ethical review, according to the current Swiss law.

Results

There were 291 reviews for the 108 papers selected for the study (Multimedia Appendix 2). Men were the first and last authors of 61 (56.5%) and 75 (69.4%) papers, respectively. The 5 most represented countries of affiliation were the United Kingdom (n=14), Germany (n=13), the United States (n=10), China (n=9), and Italy (n=6). The 3 main regions of affiliation were Western Europe (n=56), Asia (n=16), and North America (n=15).

Overall, the median sentiment and politeness scores were 58 (IQR 30-72; range –70 to 90) and 63 (IQR 47-75; range –73 to 92), respectively, but there were notable variations by the journal (Table 1). The 3 journals with the highest impact factor tended to have higher sentiment and politeness scores. There was no significant difference in scores between men and women (all $P>.05$; Table 2). The results were almost identical when using Gender API to determine gender (first authors: median sentiment or politeness scores=58, IQR 37-72 and 65, IQR 47-77 for women, 57.5, IQR 27-70 and 63, IQR 47-73 for men; last authors: median sentiment or politeness scores=55, IQR 33-68 and 63, IQR 50-75 for women, 58, IQR 27-72 and 63, IQR 45-75 for men). By contrast, papers authored by scholars from countries in the Middle East, Latin America, or Africa exhibited significantly lower sentiment and politeness scores compared to those from the other 2 regions, with differences exceeding 30 and 20 points in absolute value, respectively (adjusted $P=.02$; Table 2).

In light of the *BMJ*'s notable impact factor compared to the other journals in the study, additional analyses were conducted by excluding the 51 peer-review reports for *BMJ*. The results remained consistent (Multimedia Appendix 3). The interrater agreement between the 3 measurements was high (sentiment scores: percentage agreement=0.9958, 95% CI 0.9954-0.9962; Fleiss κ=0.9496, 95% CI 0.9395-0.9598; politeness scores: percentage agreement=0.9962, 95% CI 0.9958-0.9966; Fleiss κ=0.9463, 95% CI 0.9316-0.9610; all $P<.001$).

XSL•FO
RenderX

Table 2. Associations between sentiment or politeness scores and first or last authors' gender and first authors' affiliation in this cross-sectional study using ChatGPT-4 to quantitatively assess sentiment and politeness in 291 peer-review reports for 108 papers published in 2023 in 9 high-impact general medical journals.

Variable	Papers (n=108), n (%)	Reviews (n=291), n (%)	Sentiment score, median (IQR)	Crude *P* value[a]	Adjusted *P* value[b]	Politeness score, median (IQR)	Crude *P* value[a]	Adjusted *P* value[b]
First authors' gender				.49	.48		.37	.68
Woman	47 (43.5)	127 (43.6)	58 (33-72)			65 (47-77)		
Man	61 (56.5)	164 (56.4)	57.5 (27-70)			63 (48.5-73)		
Last authors' gender				.52	.88		.74	.86
Woman	33 (30.6)	91 (31.3)	57 (33-68)			63 (50-75)		
Man	75 (69.4)	200 (68.7)	60 (27.5-72)			63 (45-75)		
First authors' affiliation				.001	.02[c]		<.001	.02[d]
North America, Europe, and Pacific	82 (75.9)	220 (75.6)	60 (33-72)			67 (53-77)		
Asia	16 (14.8)	43 (14.8)	62 (40-70)			65 (50-73)		
Middle East, Latin America, and Africa	10 (9.3)	28 (9.6)	27 (–3 to 55)			43.5 (17.5-57)		

[a]Wilcoxon rank-sum test (for gender) and Kruskal-Wallis equality-of-populations rank test (for affiliation).

[b]Multivariable negative binomial regression, adjusted for journal, affiliation, and intracluster correlation within papers (for first or last authors' gender), and adjusted for journal and intracluster correlation within papers (for first authors' affiliation).

[c]Incidence rate ratio: Asia versus Middle East, Latin America, and Africa: 1.27 (95% CI 1.06-1.51); North America, Europe, and Pacific versus Middle East, Latin America, and Africa: 1.23 (95% CI 1.02-1.47).

[d]Incidence rate ratio: Asia versus Middle East, Latin America, and Africa: 1.30 (95% CI 1.07-1.57); North America, Europe, and Pacific versus Middle East, Latin America, and Africa: 1.27 (95% CI 1.04-1.54).

Discussion

Principal Findings

ChatGPT-4 was used to analyze sentiment and politeness in 291 peer-review reports from 9 general medical journals. The study unveiled notable regional disparities, with papers from the Middle East, Latin America, or Africa demonstrating significantly lower scores, while no discernible differences were observed based on the authors' gender.

Comparison With Existing Literature

The gender disparities experienced by women in academic medicine are widely recognized [12-17]. Consequently, significant gender discrepancies were anticipated in this study to the detriment of women. The absence of discernible differences between genders contrasts with Verharen's [11] findings, where woman first authors typically received less polite reviews compared to men. Importantly, the study focused on a different discipline (neuroscience) and was confined to a single journal (*Nature Communications*).

The regional disparities highlighted in the study resonate with prior research illustrating the challenges faced by researchers, particularly those from countries in the Global South [21,25]. These findings align with those of earlier studies [26-28] and must be understood within broader sociocultural, economic, and institutional contexts. Factors such as limited funding opportunities, language barriers, and cultural differences may contribute to biases against authors from underrepresented regions. Additionally, institutional biases within academic publishing systems may further exacerbate these discrepancies, highlighting the need for interventions to foster equity and inclusivity in scholarly discourse.

The observed disparities across geographic regions, contrasted with the absence of significant differences based on gender, could indicate that within the context of peer review, gender biases may not manifest as prominently or uniformly as other forms of bias, such as those influenced by geographic factors. In addition, existing diversity or inclusion initiatives within the scholarly community may have been more effective in mitigating gender disparities compared to other forms of bias.

Strengths and Limitations

This study built on the findings of Verharen [11], who demonstrated through several validation methods the accuracy of ChatGPT in estimating sentiment and politeness in peer-review reports, surpassing that of human evaluation and traditional lexicon-based language models. This study has several limitations, including its exclusive focus on high-impact general medical journals, reliance on binary gender determination without consideration for nonbinary or transgender identities, and uncertainty about gender or geographic distributions of rejected papers. Future research should explore alternative approaches for gender determination such as self-identification for accurately capturing gender diversity. In addition, the sample size (108 papers and 291 peer-review reports) is smaller than that of prior research by Verharen [11], potentially limiting the generalizability of the findings. The author also acknowledges the challenges associated with using sentiment or politeness metrics to capture

XSL•FO
RenderX

the nuanced biases inherent in peer review, and the dependence on algorithms like ChatGPT may introduce potential inaccuracies. Furthermore, manual scoring was not conducted for comparison. Finally, the observed association between sentiment and politeness metrics and affiliation regions suggests an alternative explanation: the possibility of higher scientific merit in papers from certain authors. The methodology lacks the capability to discern between the 2 hypotheses.

Conclusions

ChatGPT-4 demonstrated effectiveness in this study by consistently evaluating sentiment and politeness in peer-review reports. The study underscores the need for targeted interventions to address regional disparities in peer review and advocates for ongoing efforts to promote equity and inclusivity in scholarly communication.

Data Availability

The datasets generated and analyzed during this study are available in the Open Science Framework (OSF) repository [20].

Conflicts of Interest

None declared.

Multimedia Appendix 1
List and description of variables available in the Stata file uploaded to the Open Science Framework (OSF).
[DOCX File , 25 KB - jmir_v26i1e57667_app1.docx]

Multimedia Appendix 2
List of papers and reviews.
[XLSX File (Microsoft Excel File), 25 KB - jmir_v26i1e57667_app2.xlsx]

Multimedia Appendix 3
Associations between sentiment or politeness scores and first or last authors' gender and first authors' affiliation in this cross-sectional study using ChatGPT-4, to quantitatively assess sentiment and politeness in 240 peer review reports for 96 articles published in 2023 (same data as in Table 2 but without taking into account the 51 peer-review reports for the 12 articles published in BMJ).
[DOCX File , 16 KB - jmir_v26i1e57667_app3.docx]

References

1. Mehta P, Pandya DS. A review on sentiment analysis methodologies, practices and applications. Int J Sci Technol Res 2020;9:601-609 [FREE Full text]
2. Bordoloi M, Biswas SK. Sentiment analysis: a survey on design framework, applications and future scopes. Artif Intell Rev 2023;56:12505-12560 [FREE Full text] [doi: 10.1007/s10462-023-10442-2] [Medline: 37362892]
3. Cui J, Wang Z, Ho SB, Cambria E. Survey on sentiment analysis: evolution of research methods and topics. Artif Intell Rev 2023;56:8469-8510 [FREE Full text] [doi: 10.1007/s10462-022-10386-z] [Medline: 36628328]
4. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 2022;55(7):1-50. [doi: 10.1007/s10462-022-10144-1]
5. Islam T, Sheakh MA, Sadik MR, Tahosin MS, Foysal MMR, Ferdush J, et al. Lexicon and deep learning-based approaches in sentiment analysis on short texts. JCC 2024;12(01):11-34. [doi: 10.4236/jcc.2024.121002]
6. Fu Z, Hsu YC, Chan CS, Lau CM, Liu J, Yip PSF. Efficacy of ChatGPT in Cantonese sentiment analysis: comparative study. J Med Internet Res 2024;26:e51069 [FREE Full text] [doi: 10.2196/51069] [Medline: 38289662]
7. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, et al. A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. JMIR Ment Health 2024;11:e50150 [FREE Full text] [doi: 10.2196/50150] [Medline: 38271138]
8. Tabone W, de Winter J. Using ChatGPT for human-computer interaction research: a primer. R Soc Open Sci 2023;10(9):231053 [FREE Full text] [doi: 10.1098/rsos.231053] [Medline: 37711151]
9. Kim S, Kim K, Wonjeong Jo C. Accuracy of a large language model in distinguishing anti- and pro-vaccination messages on social media: the case of human papillomavirus vaccination. Prev Med Rep 2024;42:102723 [FREE Full text] [doi: 10.1016/j.pmedr.2024.102723] [Medline: 38659997]
10. Wang Z, Xie Q, Feng Y, Ding Z, Yang Z, Xia R. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv. Preprint posted online on April 10, 2023 . [doi: 10.48550/arXiv.2304.04339]
11. Verharen JPH. ChatGPT identifies gender disparities in scientific peer review. Elife 2023;12(RP90230) [FREE Full text] [doi: 10.7554/eLife.90230] [Medline: 37922198]

XSL•FO
RenderX

12. Richter KP, Clark L, Wick JA, Cruvinel E, Durham D, Shaw P, et al. Women physicians and promotion in academic medicine. N Engl J Med 2020;383(22):2148-2157. [doi: 10.1056/NEJMsa1916935] [Medline: 33252871]
13. Sebo P, de Lucia S, Vernaz N. Gender gap in medical research: a bibliometric study in Swiss university hospitals. Scientometrics 2020;126(4):741-755. [doi: 10.1007/s11192-020-03741-w]
14. Hart KL, Perlis RH. Trends in proportion of women as authors of medical journal articles, 2008-2018. JAMA Intern Med 2019;179(9):1285-1287 [FREE Full text] [doi: 10.1001/jamainternmed.2019.0907] [Medline: 31135815]
15. Sebo P, Clair C. Gender gap in authorship: a study of 44,000 articles published in 100 high-impact general medical journals. Eur J Intern Med 2022;97:103-105. [doi: 10.1016/j.ejim.2021.09.013] [Medline: 34598855]
16. Filardo G, da Graca B, Sass DM, Pollock BD, Smith EB, Martinez MAM. Trends and comparison of female first authorship in high impact medical journals: observational study (1994-2014). BMJ 2016;352:i847 [FREE Full text] [doi: 10.1136/bmj.i847] [Medline: 26935100]
17. Sebo P, Clair C. Gender inequalities in citations of articles published in high-impact general medical journals: a cross-sectional study. J Gen Intern Med 2023;38(3):661-666 [FREE Full text] [doi: 10.1007/s11606-022-07717-9] [Medline: 35794309]
18. Gender API. URL: https://gender-api.com [accessed 2023-12-31]
19. Sebo P. Performance of gender detection tools: a comparative study of name-to-gender inference services. J Med Libr Assoc 2021;109(3):414-421 [FREE Full text] [doi: 10.5195/jmla.2021.1185] [Medline: 34629970]
20. Sebo P. Use of ChatGPT to explore gender and geographic disparities in scientific peer review. Open Science Framework. URL: https://osf.io/WNRZU/ [accessed 2024-11-29]
21. Sebo P. Gender and geographical inequalities among highly cited researchers: a cross-sectional study (2014-2021). Intern Emerg Med 2023;18(4):1227-1231. [doi: 10.1007/s11739-023-03240-9] [Medline: 36877434]
22. Negative binomial regression | Stata annotated output.: UCLA: Statistical Consulting Group URL: https://stats.idre.ucla.edu/stata/output/negative-binomial-regression/ [accessed 2023-12-31]
23. Negative binomial regression | Stata data analysis examples. UCLA: Statistical Consulting Group. URL: https://stats.idre.ucla.edu/stata/dae/negative-binomial-regression/ [accessed 2023-12-31]
24. Klein D. Implementing a general framework for assessing interrater agreement in Stata. Sage J 2018;18(4):871-901. [doi: 10.1177/1536867x1801800408]
25. Pouris A, Pouris A. The state of science and technology in Africa (2000–2004): a scientometric assessment. Scientometrics 2008;79(2):297-309. [doi: 10.1007/s11192-009-0419-x]
26. Skopec M, Issa H, Reed J, Harris M. The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis. Res Integr Peer Rev 2020;5:2. [doi: 10.1186/s41073-019-0088-0] [Medline: 31956434]
27. Lor P. Scholarly publishing and peer review in the Global South: the role of the reviewer. JLIS.it 2022;14(1):10-29. [doi: 10.36253/jlis.it-512]
28. Schneider S, Morgan C, Magill C, Hersh M, Sang K, MacIntosh R. Evidence Review: Peer Review Bias in the Funding Process: Main Themes and Interventions. Edinburgh, Scotland: Heriot-Watt University; 2024.

Abbreviations

AI: artificial intelligence

Edited by A Mavragani; submitted 22.02.24; peer-reviewed by L Zhu, A Hassan, GK Gupta, H Fang; comments to author 09.04.24; revised version received 21.04.24; accepted 12.06.24; published 09.12.24.

Please cite as:
Sebo P
Use of ChatGPT to Explore Gender and Geographic Disparities in Scientific Peer Review
J Med Internet Res 2024;26:e57667
URL: https://www.jmir.org/2024/1/e57667
doi:10.2196/57667
PMID:

©Paul Sebo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Commentary

Commentary on "Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line"

Timothy D Reierson[1], BSCE
Yakima, WA, United States

Corresponding Author:
Timothy D Reierson, BSCE
6604 Appleview Rd
Yakima, WA, 98908
United States
Phone: 1 509 965 7175
Email: holdspacefree@proton.me

Related Article:

Comment on: http://www.jmir.org/2019/1/e11507/

(J Med Internet Res 2024;26:e42144) doi:10.2196/42144

KEYWORDS
bias; data ethics; data sharing; consent; crisis intervention; text messaging; commercial use; conflict of interest

This commentary is unconventional because I am not an academic peer with the authors, and I am not a credentialed practitioner in the field of mental health. Instead, as a former volunteer for the nonprofit Crisis Text Line, Inc., I am a human subject of study. My words and the words of those I conversed with are in the data being shared.

The paper [1] discusses a range of issues for sharing sensitive data for research purposes. Here, I focus on two: consent for research, and safeguards against commercial use. I provide facts and invite reconsideration of the paper's treatment of consent and data safeguards (sharing, use, and commercialization) under the lens of potential bias and exploitation.

"Bias is any trend or deviation from the truth in data collection, data analysis, interpretation and publication which can cause false conclusions. Bias can occur either intentionally or unintentionally" [2]. I invite others to decide for themselves about the potential direction, intention, and influence of bias based on the facts presented here.

The Research Enclave and Prohibition on Commercial Use

The paper [1] presents itself as a guide to technology companies for ethical sharing of data for noncommercial research purposes. The study was an 18-month pilot of a data-sharing program led by Crisis Text Line. The authors reviewed the submitted protocols and, in cooperation with Crisis Text Line, made deidentified crisis conversation data available for research purposes from enclave data. From February 2016 to July 2017, Crisis Text Line posted instructions for researchers to apply on its official website, which included a Frequently Asked Questions (FAQ) page. The FAQ promised prohibition of commercial use: "In addition, we will NEVER share data…For commercial use…" and "Are you selling this? Nope. Heck no. Not gonna happen. Yuck. Gross. (Read: no commercial use. Never ever ever.)" [3]. By July 29, 2017, however, the research application page redirected to a data philosophy page, with all prohibition language about commercial use removed and research applications for data access closed [4].

Data Enclave Discontinued

The Robert Wood Johnson Foundation granted primary funding of US $962,080 for the study's December 31, 2015 – August 30, 2017, timeframe. According to emails provided to me, on August 3, 2017, Crisis Text Line notified its Data Ethics Committee that the data enclave program would be discontinued as of September 30, 2017, and that a closed research model would be used instead [5]. By email the same day, the corresponding author wrote to ethics committee members and others that "with support of Crisis Text Line leadership, we plan to move forward with an article describing the data sharing program…We hope that those who previously expressed interest will continue to participate" [5].

XSL•FO
RenderX

Crisis Text Line, Inc. and Loris.ai, Inc.: Licensing Data for Commercial Use

Public records show that on November 16, 2017, Loris.ai, Inc. was created as a for-profit company, and on January 10, 2018, it was granted a licensing agreement to use Crisis Text Line's deidentified conversation transcript data. Loris.ai, Inc., with funding from private investors, claimed it would use "both the Crisis Text Line training and large sentiment-rich data corpus" from crisis conversations to create customer service software for commercial use. Loris.ai, Inc. launched its website in March 2018, offering a beta version of its software [6]. The chief executive officer (CEO) of Crisis Text Line, Inc. and Loris.ai, Inc. were the same person, and they made deals with themselves which they benefited from, between a nonprofit and a for-profit entity.

Conflicts of Interest and Affiliation Disclosures

When the paper was published in 2019, only two authors were identified as affiliated with Crisis Text Line, Inc. These were staff members Bob Filbin and Dr Shairi Turner. Mr Filbin was a cofounder and served as chief data scientist for Crisis Text Line until 2021. The Data Ethics Committee reported to him. One other author was a Crisis Text Line staff member—the study's open data collaborations manager Nitya Kanuri. Her staff member affiliation is undeclared in the paper. The following affiliations to Crisis Text Line, Inc. were also undeclared: the corresponding author [Anthony Pisani] and one other author [Madelyn Gould] were serving on the clinical advisory board to Crisis Text Line. Seven other authors were members of the Data Ethics Committee to the study [Brian Pascal, David Rousseau, John E Marcotte, Lisa S Lehmann, Megan L Ranney, Robert Levine (deceased), and Shirley Yen]. The Data Ethics Committee was listed on Crisis Text Line's official website as an advisory panel of experts to the organization. None of the authors made any disclosures or acknowledged the existence of the for-profit or its licensing agreement with Crisis Text Line. Loris.ai, Inc. had been in operation for months by the time the paper was submitted for publication.

Characterization of Consent

The paper discusses consent within Table 1. "CTL [Crisis Text Line] provides texters with a link to an easy-to-understand Terms of Service, including a disclosure of potential future data use, before every crisis conversation." See referenced Terms of Service [7]. There are several problems with this, but intuitively the claim is implausible, and appears contrary to the paper's own citation No. 7 (Pisani et al, 2016) [8]. The very same terms contain language granting Crisis Text Line broad control and ownership of data for any purpose [7].

Research consent, however, is more than a simple agreement to a Terms of Service. Crisis Text Line understood this and required the enclave studies to obtain institutional review board (IRB) approval to have access to the data. These IRBs are prepared to assess research involving vulnerable populations, because individuals in an active crisis texting with Crisis Text Line volunteers are vulnerable. Additionally, the conversation is between volunteers and those reaching out to the service. The volunteers are research subjects as well.

While preparing this commentary, in July 2022, I asked the Crisis Text Line legal department directly whether it took the position that volunteers had provided consent for research, but did not receive any reply [9]. In October 2022 and to my knowledge, for the first time, Crisis Text Line posted a Terms of Service & Privacy Policy specifically for volunteers [10]. The new Volunteer Terms, by their own terms, only apply going forward. This further indicates a lack of informed consent from volunteers for the entire prior conversation transcript history.

When obtaining consent from adult subjects, it is common to disclose any potential future commercial use of the data. Disclosures about commercial use are typically found in data sharing, data retention, and data use sections of consent forms. All of this relates to protecting the privacy of participants, in this case the volunteers and those in crisis reaching out to Crisis Text Line.

Minors and adults under guardianship contact Crisis Text Line as well. The regulatory environment for research on children is different. It is more strict, with additional requirements. The paper does not address this issue in any meaningful way.

Crisis Text Line's unequivocal position, before shuttering the enclave program in 2017, was that the deidentified conversation transcripts would only be shared with researchers who had IRB-approved research. In addition, there would be no commercial use of the data, or licensing and sale of the transcript data.

Commercial Use

Under "Supplementary Management Tools," the paper discusses an option to create a separate company dedicated to sharing data for research, but "CTL [Crisis Text Line] ultimately decided that their needs could be met through the other revisions made to their program and, thus, opted not to pursue the spin-off company structure." The paper does not mention the antithetical and contradictory existence of Loris.ai, Inc., a for-profit company formed by Crisis Text Line with a data-sharing agreement allowing commercial use.

One of the enclave research papers mentions commercial uses as a possible extension of its research, specifically "customer service" and "business interactions" [11].

POLITICO Response: Crisis Text Line's Cite to Paper

On January 29, 2022, the day after POLITICO published its widely read story about Loris.ai, Inc. [12], Crisis Text Line referenced the paper directly in response to criticism of its *commercial* use of deidentified data [13]. The US Federal Communications Commission has asked the Federal Trade

XSL•FO
RenderX

Commission to investigate Crisis Text Line regarding its consent practices [14]. Yet the paper continues to be cited [15].

More Documentation, and Transparency Needed

It seems likely that key documents in the possession of Crisis Text Line could shed light on these concerns. Chief data scientist and coauthor Bob Filbin likely had first-hand knowledge about Loris.ai, Inc. The change from a commitment to forever prohibit commercial use to a highly sophisticated actual commercial use, by the host of the pilot study, is significant. It warrants disclosure of documents describing the turnabout with sufficient transparency to show the timing and substance truthfully.

Did Crisis Text Line inform the authors and the Data Ethics Committee about its July 2017 website change, a marked change in ethical stance regarding commercial use? Did they inform them about Crisis Text Line's role in creating Loris.ai, Inc.? According to Dr Megan Ranney, they did not [16]. Did they inform the research participants—the volunteers and those who reached out for assistance from Crisis Text Line—about their intent to use the transcript data for commercial purposes and profit from it?

For the organization or any author to withhold information about plans for commercial use from the others would introduce significant potential for bias, and many IRBs would find this an issue of noncompliance and potentially exploitative of a vulnerable population participating in research.

A Self-Regulated Environment

I have learned a few things through all of this:

1. Consent is not a current legal requirement for research on human subject data held by a nonprofit or private corporation, despite it being the overwhelming norm in health care research.

2. IRBs may, in some circumstances, waive consent when human subject data is deidentified, consistent with institutional policies and applicable regulations. If they are at accountable care organizations (ACOs) or if the research involves children, regulations are more strict, and consent requires special consideration.

3. Crisis Text Line's ethical standards are voluntary.

To my reading, the paper presented no safeguards other than public opinion for guardrails on use of the data. For example, a code of ethics prohibiting commercial use could have been recommended to the organization for adoption. Without author or committee knowledge about Crisis Text Line's plan for commercial use, such safeguards may not have seemed necessary. The problem of technology companies being largely self-regulated ethical stewards of the data they collect was not adequately addressed in this paper, as evidenced by the very existence of Loris.ai, Inc. prior to the paper's submittal for publication.

Human Subject Protections Matter

In summary, there are ways to harmonize respect for persons, one of the three pillars of the Belmont Report, with the greater good. The loss of trust in systems of care when consent is waived is a concern that deserves greater attention. The paper provides an example where data are used for purposes other than that for which they were collected. Participants were misled about the data sharing and data use practices. In addition, the participants are vulnerable in the situation where the data are collected, and likely incapable of protecting their own interests. Further, if they were children, additional safeguards and different regulatory requirements may apply. This makes voluntary consent problematic, and no additional safeguards for their protection were proposed.

This resulted in an apparent misrepresentation to IRBs when seeking a waiver of consent for research, a misrepresentation to the participants about commercial use, a lack of safeguards for a vulnerable population, inadequate safeguards for the protection of minors participating in research, and a lack of safeguards against commercial use.

This originated, apparently, from corporate Crisis Text Line, which made explicit claims prohibiting commercial use and then implemented an explicit commercial use plan. These actions cast shadows on the enclave study design and the paper itself. For most authors and ethics committee members, this is likely unintentional, because they may not have known what Crisis Text Line was doing or their intent. For Crisis Text Line and its key personnel, the possibility of this being intentional should be seriously considered, and further documentation requested.

My detailed letter of concern to JMIR is available at my advocacy website [17].

Conflicts of Interest

The author served as uncompensated volunteer for Crisis Text Line April 2020–August 2021.

References

1. Pisani A, Kanuri N, Filbin B, Gallo C, Gould M, Lehmann L, et al. Protecting user privacy and rights in academic data-sharing partnerships: principles from a pilot program at Crisis Text Line. J Med Internet Res;21(1) 2019 Jan 17:e11507 [FREE Full text] [doi: 10.2196/11507]
2. Simundić AM. Bias in research. Biochem Med (Zagreb) 2013;23(1):12-15 [FREE Full text] [doi: 10.11613/bm.2013.003] [Medline: 23457761]
3. Enclave Data FAQ - archive, retrieved February 19, 2016, page capture. Crisis Text Line. URL: https://web.archive.org/web/20160219144924/http://www.crisistextline.org/open-data/faq/ [accessed 2022-08-10]

XSL•FO
RenderX

4. Frequently Asked Questions - archive, retrieved March 10, 2018, page capture with redirect to Data Philosophy page. Crisis Text Line. URL: https://web.archive.org/web/20181027021701/https://www.crisistextline.org/open-data/faq/ [accessed 2022-08-10]
5. Kanuri, Nitya, to various and Pisani, Anthony, to various (August 3, 2017). Two emails, partly redacted, to various parties. Reform Crisis Text Line. URL: https://reformcrisistextline.com/wp-content/uploads/2022/08/emails_2018_CTL_data_enclave_redacted.pdf [accessed 2022-08-10]
6. Loris.ai - archive, retrieved March 1, 2018, page capture. URL: https://web.archive.org/web/20180301081606/https://www.loris.ai/ [accessed 2022-08-10]
7. Terms of Service & Privacy Policy (last updated December 9, 2016) - archive, retrieved June 12, 2017, page capture. Crisis Text Line. URL: https://web.archive.org/web/20170612220236/http://www.crisistextline.org/privacy/ [accessed 2022-08-10]
8. Pisani A, Wyman P, Mohr D, Perrino T, Gallo C, Villamar J, et al. Human subjects protection and technology in prevention science: selected opportunities and challenges. Prev Sci 2016 Aug;17(6):765-778 [FREE Full text] [doi: 10.1007/s11121-016-0664-1] [Medline: 27220838]
9. Reierson TD. Two emails, to Legal Dept. and Rodriguez, Shawn. Reform Crisis Text Line. URL: https://reformcrisistextline.com/wp-content/uploads/2022/11/email_to_crisis_text_line_legal_2022-07-13.pdf [accessed 2023-04-22]
10. Volunteer Terms of Service & Privacy Policy (last updated October 26, 2022) - archive, retrieved April 22, 2023, page capture. Crisis Text Line. URL: http://web.archive.org/web/20230421123722/https://www.crisistextline.org/volunteer-terms-of-service-privacy-policy/ [accessed 2023-04-21]
11. Zhang J, Filbin R, Morrison C, Weiser J, Danescu-Niculescu-Mizil C. Finding your voice: the linguistic development of mental health counselors. arXiv. Preprint posted online Jun 17, 2019 [FREE Full text]
12. Levine AS. Suicide hotline shares data with for-profit spinoff, raising ethical questions. Politico. 2022 Jan 28. URL: https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617 [accessed 2024-12-27]
13. @CrisisTextLine. 3. Leading academics published research in the Journal of Medical Internet Research after spending 18 months evaluating our data sharing practices. They found that our ‘practices yielded key principles and protocols for ethical data sharing’. 2022 Jan 29. URL: https://twitter.com/CrisisTextLine/status/1487515686364712961 [accessed 2022-08-10]
14. Carr B. Letter from Commissioner Brendan Carr to Federal Trade Commission. 2022 Mar 22. URL: https://reformcrisistextline.com/wp-content/uploads/2022/04/2022-03-22_ltr_FCC_to_FTC.pdf [accessed 2022-08-10]
15. Pisani AR, Gould MS, Gallo C, Ertefaie A, Kelberman C, Harrington D, et al. Individuals who text Crisis text Line: key characteristics and opportunities for suicide prevention. Suicide Life Threat Behav 2022 Jun 26;52(3):567-582 [FREE Full text] [doi: 10.1111/sltb.12872] [Medline: 35615898]
16. Levine AS. Suicide hotline left ethics board out of the loop about data-sharing with for-profit spinoff - archive, retrieved December 5, 2022, page capture. Forbes. 2022 Feb 24. URL: https://web.archive.org/web/20221205110730/https://www.forbes.com/sites/alexandralevine/2022/02/24/suicide-hotline-left-ethics-board-out-of-the-loop-about-data-sharing-with-for-profit-spinoff/ [accessed 2022-08-10]
17. Reform Crisis Text Line. URL: https://reformcrisistextline.com/ [accessed 2024-12-24]

Abbreviations

ACO: accountable care organizations
CEO: chief executive officer
FAQ: frequently asked questions
IRB: institutional review board

Edited by T Leung, G Eysenbach; submitted 24.08.22; this is a non–peer-reviewed article;accepted 28.06.24; published 30.12.24.

Please cite as:
Reierson TD
Commentary on “Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line”
J Med Internet Res 2024;26:e42144
URL: https://www.jmir.org/2024/1/e42144
doi:10.2196/42144
PMID:

©Timothy D Reierson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usability of Health Care Price Transparency Data in the United States: Mixed Methods Study

Negar Maleki[1], PhD; Balaji Padmanabhan[2], PhD; Kaushik Dutta[1], PhD

[1]School of Information Systems and Management, Muma College of Business, University of South Florida, Tampa, FL, United States

[2]Decision, Operations & Information Technologies Department, Robert H. Smith School of Business, University of Maryland, College Park, MD, United States

Corresponding Author:
Negar Maleki, PhD
School of Information Systems and Management
Muma College of Business
University of South Florida
4202 E Fowler Avenue
Tampa, FL, 33620
United States
Phone: 1 8139742011
Email: negarmaleki@usf.edu

Abstract

Background: Increasing health care expenditure in the United States has put policy makers under enormous pressure to find ways to curtail costs. Starting January 1, 2021, hospitals operating in the United States were mandated to publish transparent, accessible pricing information online about the items and services in a consumer-friendly format within comprehensive machine-readable files on their websites.

Objective: The aims of this study are to analyze the available files on hospitals' websites, answering the question—is price transparency (PT) information as provided usable for patients or for machines?—and to provide a solution.

Methods: We analyzed 39 main hospitals in Florida that have published machine-readable files on their website, including commercial carriers. We created an Excel (Microsoft) file that included those 39 hospitals along with the 4 most popular services—Current Procedural Terminology (CPT) 45380, 29827, and 70553 and Diagnosis-Related Group (DRG) 807—for the 4 most popular commercial carriers (Health Maintenance Organization [HMO] or Preferred Provider Organization [PPO] plans)—Aetna, Florida Blue, Cigna, and UnitedHealthcare. We conducted an A/B test using 67 MTurkers (randomly selected from US residents), investigating the level of awareness about PT legislation and the usability of available files. We also suggested format standardization, such as master field names using schema integration, to make machine-readable files consistent and usable for machines.

Results: The poor usability and inconsistent formats of the current PT information yielded no evidence of its usefulness for patients or its quality for machines. This indicates that the information does not meet the requirements for being consumer-friendly or machine readable as mandated by legislation. Based on the responses to the first part of the experiment (PT awareness), it was evident that participants need to be made aware of the PT legislation. However, they believe it is important to know the service price before receiving it. Based on the responses to the second part of the experiment (human usability of PT information), the average number of correct responses was not equal between the 2 groups, that is, the treatment group (mean 1.23, SD 1.30) found more correct answers than the control group (mean 2.76, SD 0.58; t_{65}=6.46; P<.001; d=1.52).

Conclusions: Consistent machine-readable files across all health systems facilitate the development of tools for estimating customer out-of-pocket costs, aligning with the PT rule's main objective—providing patients with valuable information and reducing health care expenditures.

(J Med Internet Res 2024;26:e50629) doi:10.2196/50629

KEYWORDS

price transparency; user experiments; schema analysis; health care; patients; algorithms

XSL•FO
RenderX

Introduction

Overview

From 1970 to 2020, on a per capita basis, health care expenditures in the United States have increased sharply from US $353 per person to US $12,531 per person. In constant 2020 dollars, the increase was from US $1875 in 1970 to US $12,531 in 2020 [1]. The significant rise in health care expenses has put policy makers under enormous pressure to find ways to contain these expenditures. Price transparency (PT) in health care is 1 generally proposed strategy for addressing these problems [2] and has been debated for years [3]. Some economists believe that PT in health care will cut health care prices in the same way it has in other industries, while others argue that owing to the specific characteristics of the health care market, PT would not ameliorate rising health care costs. Price elasticity also does not typically apply in health care, since, if a problem gets severe, people will typically seek treatment regardless of cost, with the drawback that individuals learn of their health care costs after receiving treatment [4]. Complex billing processes, hidden insurer-provider contracts, the sheer quantity of third-party payers, and substantial quality differences in health care delivery are other unique aspects of health care that complicate the situation considerably.

The Centers for Medicare & Medicaid Services (CMS) mandated hospitals to post negotiated rates, including payer-specific negotiated costs, for 300 “shoppable services” beginning in January 2021. The list must include 70 CMS-specified services and an additional 230 services each hospital considers relevant to its patient population. Hospitals must include each third-party payer and their payer-specific fee when negotiating multiple rates for the same care. The data must be displayed simply, easily accessible (without requiring personal information from the patient), and saved in a machine-readable manner [5]. These efforts aim to facilitate informed patient decision-making, reduce out-of-pocket spending, and decrease health care expenditures. Former Secretary of Health and Human Services, Alex Azar, expressed a vision of hospital PT when declaring the new legislation “a patient-centered system that puts you in control and provides the affordability you need, the options and control you want, and the quality you deserve. Providing patients with clear, accessible information about the price of their care is a vital piece of delivering on that vision” [6].

Despite the legislation, it is not clear if people are actually engaging in using PT tools. For example, in 2007, New Hampshire’s HealthCost website was established, providing the negotiated price and out-of-pocket costs for 42 commonly used services by asking whether the patient is insured or their insurer and the zip code to post out-of-pocket costs in descending order. Mehrotra et al [7] examined this website over 3 years to understand how often and why these tools have mainly been used. Their analysis suggested that despite the growing interest in PT, approximately 1% of the state’s population used this tool. Low PT tool usage was also seen in other studies [8-10], suggesting that 3% to 12% of individuals who were offered the tool used it during the study period, and in all studies, the duration was at least 12 months. Thus, offering PT tools does not in itself lead to decreased total spending, since few people who have access to them use them to browse for lower-cost services [7,11].

In a recent paper, researchers addressed 1 possible reason for low engagement—lack of awareness. They implemented an extensive targeted online advertising campaign using Google Advertisements to increase awareness and assessed whether it increased New Hampshire’s PT website use. Their findings suggested that although lack of awareness is a possible reason for the low impact of PT tools in health care spending, structural factors might affect the use of health care information [12]. Individuals may not be able to exactly determine their out-of-pocket expenses from the information provided.

Surprisingly, there is little research on the awareness and usability of PT information after the current PT legislation went into effect. A recent study [13] highlighted the nonusability of existing machine-readable files for employers, policy makers, researchers, or consumers, and this paper adds to this literature by answering the question—is PT information as provided usable for patients or machines? Clearly, if it is of value to patients, it can be useful; the reason to take the perspective of machines was to examine whether this information as provided might also be useful for third-party programs that can extract information from the provided data (to subsequently help patients through other ways of presenting this information perhaps). We address this question through a combination of user experiments and data schema analysis. While there are recent papers that have also argued that PT data have deficiencies [13,14], ours is the first to combine user experiments with analysis of data schema from several hospitals in Florida to make a combined claim on value for patients and machines. We hope this can add to the discourse on PT and what needs to be done to extract value for patients and the health care system as a whole.

Background

Impact of PT Tools

The impact of PT tools on consumers and health care facilities has been investigated in the literature. Some studies showed that consumers with access to PT tools are more likely to reduce forgone needed services over time. Moreover, consumers who use tools tend to find the lowest service prices [8,15-17]. A few studies investigated the impact of PT tools on the selection of health care facilities. They illustrated that some consumers tend to change health care facilities pursuing lower prices, while some others prefer to stay with expensive ones, although they are aware of some other facilities that offer lower prices [9,18]. Finally, some research studied the impact of PT tools on cost and showed that some consumers experienced no effect, while others experienced decreases in average consumer expenses [8,17,18]. However, the impact of PT tools on health care facilities is inconclusive, meaning different studies concluded different effects. Some stated that PT tools decrease the prices of imaging and laboratory services, while others said that although public charge disclosure lowers health care facility charges, the final prices remained unchanged [17,18].

XSL•FO
RenderX

Legislation Related Works

In a study, researchers considered 20 leading US hospitals to assess provided chargemasters to understand to what extent patients can obtain information from websites to determine the out-of-pocket costs [19]. Their findings showed that although all hospitals provided chargemasters on their websites, they rarely offered transparent information, making it hard for patients to determine out-of-pocket costs. Their analysis used advanced diagnostic imaging services to assess hospitals' chargemasters since these are the most common services people look for. Mehrotra et al [7] also mentioned that the most common searches belonged to outpatient visits, magnetic resonance imaging (MRI), and emergency department visits. To this end, we used "MRI scan of the brain before and after contrast" as one of the shoppable services in our analysis. Another study examined imaging services in children's hospitals (n=89), restricting the analysis to hospitals (n=35) that met PT requirements—published chargemaster rates, discounted cash prices, and payer-negotiated prices in a machine-readable file, and published costs for 300 common shoppable medical services in a consumer-friendly format. Their study revealed that, in addition to a broad range of imaging service charges, most hospitals lack the machine-readable file requirement [20].

Arvisais-Anhalt et al [21] identified 11 hospitals with available chargemasters in Dallas County to compare the prices of a wide range of available services. They observed significant variations for a laboratory test: partial thromboplastin time, a medication: 5 mg tablet of amlodipine, and a procedure: circumcision. Reddy et al [22] focus on New York State to assess the accessibility and usability of hospitals' chargemasters from patients' viewpoint. They found that 189 out of 202 hospitals had a locatable chargemaster on their home page. However, only 37 hospitals contain the Current Procedural Terminology (CPT) code, which makes those without the CPT code unusable due to the existence of many different descriptions for the same procedure; for example, an elective heart procedure had 34 entries. We add to this considerable literature by examining a subset of Florida hospitals.

In a competitive market, higher-quality goods and services require higher prices [23]. Based on this, Patel et al [24] examined the relationship between the Diagnosis-Related Group (DRG) chargemaster and quality measures. Although prior research found no convincing evidence that hospitals with greater costs also delivered better care [25], they discovered 2 important quality indicators that were linked to standard charges positively and substantially—mortality rate and readmission rates—which both are quality characteristics that are in line with economic theory. Moreover, Patel et al [24] studied the variety of one of the most commonly performed services (vaginal delivery) as a DRG code, which motivated us to select "Vaginal delivery without sterilization or D&C without CC/MCC" as another shoppable service in our analysis.

Methods

Ethical Considerations

All data used in this study, including the secondary data set obtained from hospitals' websites and the data collected during the user experiment, underwent a thorough anonymization process. The study was conducted under protocols approved by the University of South Florida institutional review board (STUDY004145: "Effect of price transparency regulation (PTR) on the public decisions") under HRP-502b(7) Social Behavioral Survey Consent. This approval encompassed the use of publicly available anonymized secondary data from hospitals' websites, as well as a user experiment aimed at assessing awareness of the PT rule and the usability of hospitals' files. No individual-specific data were collected during the experiment, which solely focused on capturing subjects' awareness and opinions regarding the PT rule and associated files. At the onset of the experiment, participants were provided with a downloadable consent form and were allowed to withdraw their participation at any time. Survey participants were offered a US $2 reward, and their involvement was entirely anonymous.

Data Collection

According to CMS, "Starting January 1, 2021, each hospital operating in the United States will be required to provide clear, accessible pricing information online about the items and services they provide in two ways: 1- As a comprehensive machine-readable file with all items and services. 2- In a display of shoppable services in a consumer-friendly format." As stated, files available on hospitals' websites should be consumer-friendly, so the question of whether these files are for users arises. On the other hand, as stated, files should be machine-readable, so again the question of whether these files are for machines arises. Below we try to answer both questions in detail, respectively.

Value for Users: User Experiments

Overview

When a public announcement is disseminated, its efficacy relies on ensuring widespread awareness and facilitating practical use during times of necessity. Previous research on PT announcements has highlighted the challenges faced by patients in accurately estimating out-of-pocket expenses. However, a fundamental inquiry arises—are individuals adequately informed about the availability of tools that enable them to estimate their out-of-pocket costs for desired services? To address this, we conducted a survey to assess public awareness of PT legislation. The survey encompassed a range of yes or no and multiple-choice questions aimed at gauging participants' familiarity with the PT rule in health care and their entitlement to obtain cost information prior to receiving a service. Additionally, we inquired about participants' knowledge of resources for accessing pricing information and whether they were aware of the PT rule. Furthermore, we incorporated follow-up questions to ensure that the survey responses were not provided arbitrarily, thereby securing reliable and meaningful outcomes.

Moreover, considering the previously established evidence of subpar usability associated with the currently available files, we propose streamlining the existing files and developing a user-friendly and comprehensive document for conducting an A/B test. This test aims to evaluate which file better facilitates participants in accurately estimating their out-of-pocket costs.

XSL•FO
RenderX

In collaboration with Florida Blue experts during biweekly meetings throughout the entire process outlined in this paper, the authors determined the optimal design for the summary table. This design, which presents prices in a more user-friendly format, enhancing overall participant comprehension, was used during the A/B testing. Participants were randomly assigned to either access the hospitals' files or a meticulously constructed summary table, manually created in Excel, prominently displaying cost information (Please note that all files, including the hospitals' files and our Excel file, are made available in the same format [Excel] on a cloud-based platform to eliminate any disparities in accessing the files. This ensures equitable ease of finding, downloading, and opening files, as accessing the hospitals' files typically requires significant effort.). The experiment entailed presenting 3 distinct health-related scenarios and instructing participants to locate the price for the requested service. Subsequently, participants were asked to provide the hospital name, service price, insurer name, and insurance plan. Additionally, we sought feedback on the perceived difficulty of finding the requested service and their priority for selecting hospitals [26], followed by Likert scale questions to assess participants' evaluation of the provided file's efficacy in facilitating price retrieval.

The experiments were conducted to investigate the following questions: (1) Are the individuals aware of the PT legislation? and (2) Is the information provided usable for patients? To evaluate the usability of files found on websites, we selected 2 prevalent services based on existing literature and 2 other services recommended as high-demand ones by Florida Blue experts, Table 1. Furthermore, meticulous efforts were made to ensure that both the control and treatment groups encountered identical circumstances, thus allowing for a systematic examination of the disparities solely attributable to variations in data representation.

Table 1. Detail information on selected services for the experiment.

Service code	Code type	Service full name	References
807	DRG[a]	Vaginal delivery without sterilization or D&C[b] without CC/MCC[c]	[24]
29827	CPT[d]	Arthroscopic cuff repair (shoulder arthroscopy)	Suggested by Florida Blue experts as high-demand services.
45380	CPT	Colonoscopy, flexible, with biopsy, single, or multiple	Suggested by Florida Blue experts as high-demand services.
70553	CPT	MRI[e] scan of brain before and after contrast	[7,20]

[a]DRG: Diagnosis-Related Group.

[b]D&C: dilation and curettage.

[c]CC/MCC: complication or comorbidity/major complication or comorbidity.

[d]CPT: Current Procedural Terminology.

[e]MRI: magnetic resonance imaging.

Participants

A total of 67 adults (30 female individuals; mean 41.43, SD 12.39 years) were recruited on the Amazon Mechanical Turk platform, with no specific selection criteria other than being located in the United States.

Data

We focused on 75 main hospitals (ie, the main hospital refers to distinguish a hospital from smaller clinics or specialized medical centers within the same health system) in the state of Florida. When we searched their websites for PT files (machine-readable files), only 89% (67/75) of hospitals included machine-readable files. According to the PT legislation, these files were supposed to contain information about 300 shoppable services. However, only 58% (39/67) of hospitals included information such as insurer prices in their files. Therefore, for the rest of the analysis, we only included the 39 hospitals that have the required information in their machine-readable files on their websites. We created an Excel file that included those 39 hospitals along with the 4 services—CPT 45380, 29827 and 70553 and DRG 807—mentioned in the literature (Table 1) for 4 popular (suggested by Florida Blue experts) commercial carriers (Health Maintenance Organization [HMO] or Preferred Provider Organization [PPO] plans)—Aetna, Florida Blue, Cigna, and UnitedHealthcare.

Procedure

Participants were recruited for the pilot and randomly assigned by the Qualtrics XM platform to answer multiple-choice questions and fill in blanks based on the given scenarios. First, participants responded to questions regarding the awareness of PT and then were divided into 2 groups randomly to answer questions regarding the usability of hospital-provided PT information. One group was assigned hospitals' website links (control group), while the other group was given an Excel file with the same information provided in files on hospitals' websites, but in a manner that was designed to allow easier comparison of prices across hospitals (Multimedia Appendix 1). Participants were given 3 scenarios that asked them to find a procedure's price based on their hospital and insurer selection to compare hospital-provided information with Excel. We provide some examples of hospitals' files and our Excel file in Multimedia Appendix 1 and the survey experiment questions in Multimedia Appendix 2.

XSL•FO
RenderX

Value for Machines: Schema Integration—Machine-Readable Files Representation

Through meticulous investigation of machine-readable files from 39 hospitals, we discovered that these files may vary in formats such as CSV or JSON, posing a challenge for machines to effectively manage the data within these files. Another significant obstacle arises from the lack of uniformity in data representation across these files, rendering them unsuitable for machine use without a cohesive system capable of processing them collectively. Our analysis revealed that hospitals within a single health system exhibit consistent data representation, although service prices may differ (we include both the same and different chargemaster prices in our study), while substantial disparities in data representation exist between hospitals affiliated with different health systems.

Moving forward, we will use the terms "data representation" and "schema" interchangeably, with "schema" denoting its database management context. In this context, a schema serves as a blueprint outlining the structure, organization, and relationships of data within a database system. It encompasses key details such as tables, fields, data types, and constraints that define the stored data. To systematically illustrate schema differences among hospitals associated with different health systems, we adopted the methodology outlined in reference [27] for schema integration, which offers a valid approach for comparing distinct data representations. The concept of schema integration encompasses four common categories: (1) identical: hospitals within the same health system adhere to this concept as their representations are identical; (2) equivalent: while hospitals in health system "A" may present different representations from those in health system "B," they possess interchangeable columns; (3) compatible: in cases where hospitals across different health systems are neither identical nor equivalent, the modeling constructs, designer perception, and integrity constraints do not contradict one another; and (4) incompatible: in situations where hospitals within different health systems demonstrate contradictory representations, distinct columns exist for each health system due to specification incoherence.

Our analysis focused on health systems in Florida that encompassed a minimum of 4 main hospitals, using the most up-to-date data available on their respective websites. Within this scope, we identified 8 health systems with at least 4 main hospitals, of which 88% (7/8) of health systems had published machine-readable files on their websites. Consequently, our analysis included 65% (36/55) of hospitals that possessed machine-readable files available on their websites. To facilitate further investigation by interested researchers, we have made the analyzed data accessible on a cloud-based platform. During our analysis, we meticulously extracted the schema of each health system by closely scrutinizing the hospitals associated with each health system, capturing key details such as tables, fields, and data types. Subsequently, we compiled a comprehensive master field name table trying to have the same data type and field names that make it easier for machines to retrieve information. We elaborate on the master field names table in greater detail within the results section.

Results

Value for Users

Question 1 (PT Awareness)

Based on the responses, it is evident that participants need to be made aware of the PT legislation. Among the participants, 64% (49/76) reported that they had not heard about the legislation. However, they believe it is important to know the service price before receiving it—response charts are provided in Multimedia Appendix 3.

Question 2 (Human Usability of PT Information)

Based on the responses to scenarios, the average number of correct responses is not equal between the 2 groups, that is, the treatment group (mean 1.23, SD 1.30) found more correct answers than the control group (mean 2.76, SD 0.58; t_{65}=6.46; P<.049; d=1.52). The t tests (2-tailed) for the other questions in the experiment are in Multimedia Appendix 4.

These suggest that current files on hospitals' websites are not consumer-friendly, and participants find it challenging to estimate out-of-pocket costs for a desired service. For this reason, in addition to making the files easier to use, this information should also include thorough documentation that explains what each column represents, up to what amount an insurer covers for a specific service, or the stated price covers up to how many days of a particular service, that is, "contracting method." For example, based on consulting with one of the senior network analysts of Florida Blue, some prices for a service like DRG 807 are presented as per diem costs, and based on the current information on these files, it cannot be recognizable without having comprehensive documentation for them.

Value for Machines

After carefully reviewing all machine-readable file schemas, we create a master field name table, including the available field names in machine-readable files (Table 2). According to Table 2, the first column represents master field names that we came up with, and the following columns each represent hospitals within a health system. The "✓" mark shows that hospitals within a health system have identical field names as we consider as master field names and the "written" cells show equivalent field names, meaning that hospitals within that health system use different field names—we write what they use in their representation—while the content is equivalent to what we select as the master field name. The "✱" mark means that although hospitals within health system #2 provide insurer names and plans in their field names, some codes make those columns unusable for machines to recognize them the same as master field names. We also include the type of field names for all representations in parentheses.

XSL•FO
RenderX

Table 2. Master field names that can be used for all hospitals within all health systems.

Master field names[a]	Hospitals within health system #1	Hospitals within health system #2	Hospitals within health system #3	Hospitals within health system #4	Hospitals within health system #5	Hospitals within health system #6	Hospitals within health system #7
Code | (type)	✓[b] | (str[c])	✓ | (int[d])	✓ | (int)	Final CPT[e] | (str)	✓ | (str)	HCPCS[f]/CPT Code| (str)	CPT code/HCPCS | (str)
Description | (type)	✓ | (str)	✓ | (str)	✓ | (str)	Tech name | (str)	Code description | (str)	✓ | (str)	Procedure description | (str)
Code type | (type)	✓ | (str)	✓ | (str)	__[g]	—	✓ | (str)	—	—
Apr[h] DRG[i] | (type)	—	—	—	✓ | (int)	—	—	✓ | (int)
Ms[j] DRG | (type)	—	—	—	✓ | (int)	—	—	✓ | (int)
Ms DRG Title | (type)	—	—	—	✓ | (str)	—	—	—
CDM[k] service description | (type)	—	—	—	—	—	—	✓ | (str)
Patient class | (type)	Type | (str)	—	Type | (str)	—	✓ | (str)	—	—
Revenue description | (type)	—	UB[l] revenue description | (str)	—	Final rev | (str)	—	—	—
Revenue code | (type)	Rev code | (int)	UB revenue code | (int)	—	Final rev | (int)	—	—	✓ | (int)
Package or line level | (type)	—	—	✓ | (str)	—	—	—	—
Procedure ID | (type)	—	—	—	Srvc Prvd Srvc ID | (int)	—	✓ | (int)	Procedure code/charge code | (int)
Payer | (type)	—	—	—	—	✓ | (str)	—	—
Payer specific negotiated charge | (type)	—	—	—	—	✓ | (float[m])	—	—
Gross charge | (type)	✓ | (float)	✓ | (float)	✓ | (float)	✓ | (float)	✓ | (float)	✓ | (float)	Default charge | (int)
Discounted cash price | (type)	✓ | (float)	Cash charge | (float)	Derived contracted rate | (float)	✓ | (float)	✓ | (float)	✓ | (float)	Cash charge | (float)
Min negotiated rate | (type)	Min | (float)	✓ | (float)	Deidentified min contracted rate | (float)	✓ | (float)	Deidentified min contracted charge | (float)	—	Minimum | (int)
Max negotiated rate | (type)	Max | (float)	✓ | (float)	Deidentified max contracted rate | (float)	✓ | (float)	Deidentified max contracted charge | (float)	—	Maximum | (int)
Self pay | (type)	—	—	✓ | (float)	✓ | (float)	—	—	—
Insurer name | (type)	✓ | (str)	✻[n] | (str)	✓ | (str)	✓ | (str)	—	—	✓ | (str)
Insurer plan | (type)	✓ | (str)	✻ | (str)	✓ | (str)	✓ | (str)	—	—	✓ | (str)
Price | (type)	✓ | (float)	✓ | (float)	✓ | (float)	✓ | (float)	—	—	✓ | (float)

[a]As noted previously, since we focus on the health system level instead of the hospital level, our schema does not have hospital-level information;

XSL•FO
RenderX

however, it would be beneficial to add hospital information to the table.

[b]✓: it means the given master field name in that row appears on the given health system file in that column.

[c]str: shows "string" as the data type.

[d]int: shows "integer" as the data type.

[e]CPT: Current Procedural Terminology.

[f]HCPCS: Health care Common Procedure Coding System.

[g]Not applicable.

[h]Apr: all patients refined.

[i]DRG: Diagnosis-Related Group.

[j]Ms: Medicare severity.

[k]CDM: charge description master.

[l]UB: uniform billing.

[m]float: it shows "float" as the data type.

[n]✱: it means that although hospitals within health system #2 provide insurer names and plans in their field names, some codes make those columns unusable for machines to recognize them the same as master field names.

We did reverse engineering and drew entity-relationship diagrams (ERDs) for each hospital based on their data representation. However, as hospitals within the same health system have the same ERDs, we only include 1 ERD for each health system (Figure 1). According to Figure 1, although hospitals have tried to follow an intuitive structure, we can still separate them into three groups: (1) group I: all hospitals within this group have several columns for different insurers. As shown in the ERDs, we decided to have a separate entity, called "Insurance" for this group; (2) group II: all hospitals within this group have many sheets, and each sheet belongs to a specific insurer with a specific plan. As shown in the ERDs, we decided to create an "Insurance_Name" entity for this group's ERD to show the difference in data representation; and (3) group III: all hospitals within this system have a "payer" column which includes the names of insurers without their plans. As shown in the ERDs, we decided to put this column as an attribute in the "Service" entity, and do not have an "Insurance" entity for this group's ERD.

In conclusion, although most hospitals have adopted group I logic for data representation, for full similarity, a standard representation with the same intuitive field names (like what we suggest as the master field name; Table 2) should be proposed so that it can cover all systems' data representations and be used as machine-readable file, for at least machine benefits. Mainly, standardization in the format and semantics of the provided data can help substantially in making the data more machine friendly.

XSL•FO
RenderX

Figure 1. Grouping hospitals' ERDs within each health system. ERD: entity-relationship diagram; PK: primary key.

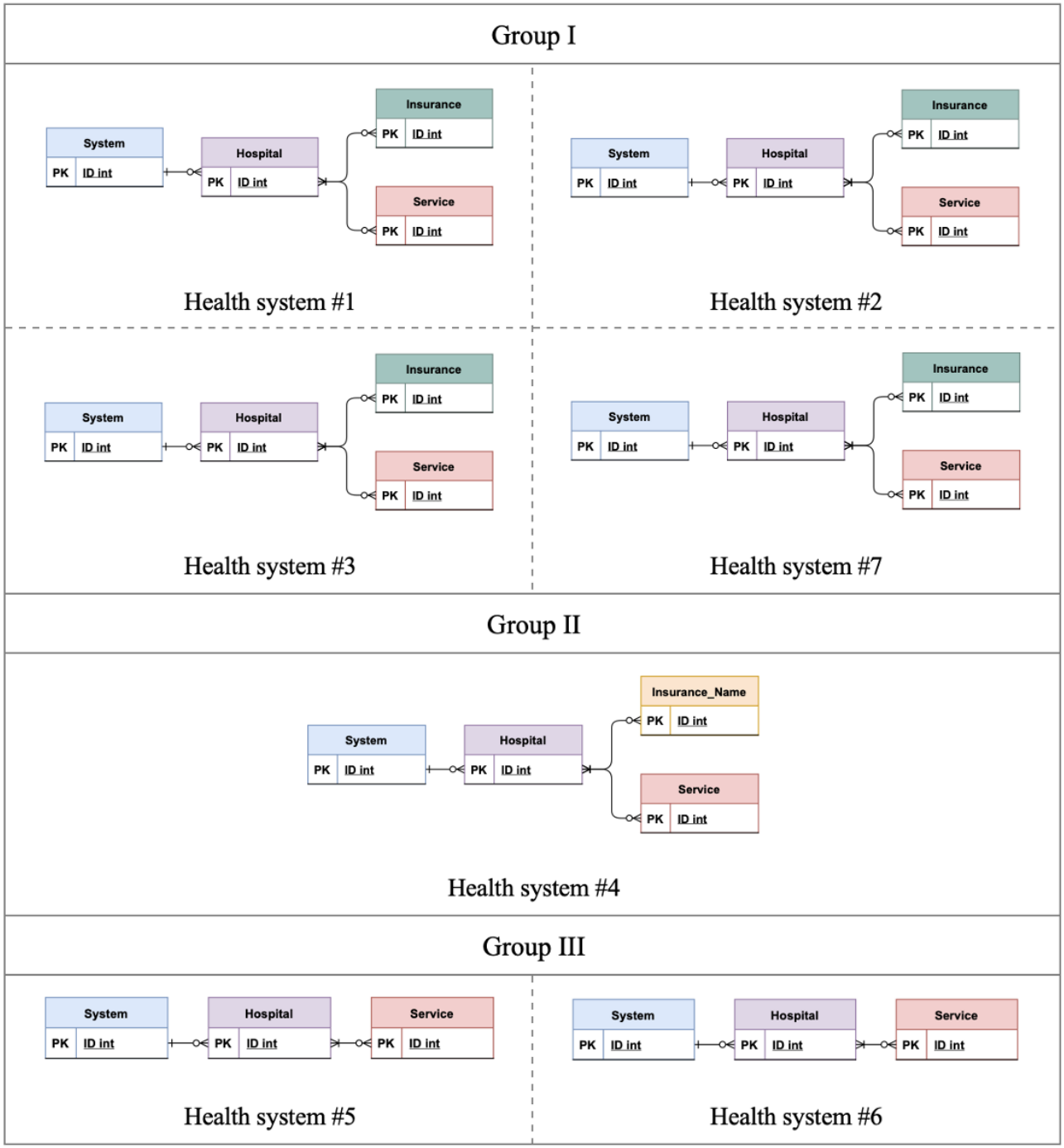

Discussion

Comparison With New CMS Guidelines

Recently, CMS has published guidelines regarding the PT legislation [28]. The most recent CMS guideline is a step forward in ensuring standardization but is still only recommended and is not mandatory. These guidelines exhibit overlaps with our fields in Table 2, with slight differences attributed to granularities. Our observation reveals that hospitals within the same health system adopt a uniform schema. Therefore, our suggested schema operates on the granularity of health systems rather than individual hospitals.

The recent CMS guidelines allocate 24% (6/25) of field names specifically to hospital information, encompassing details such as "Hospital Name," "Hospital File Date," "Version," "Hospital Location," "Hospital Financial Aid Policy," and "Hospital Licensure Information." These details, absent in current hospital files, are crucial for informed decision-making. As noted previously, since we focus on the health system level instead of the hospital level, our schema does not have hospital-level information; however, it would be beneficial to add hospital information to the tables.

Our analysis reveals that the 11 field names in Table 2 align with the field names in the new CMS guidelines, demonstrating a substantial overlap of 58% (11/19). The corresponding CMS

XSL•FO
RenderX

field names (compatible with our schema) include "Item or Service Description (Description or CDM Service Description)," "Code (Code)," "Code Type (Code Type)," "Setting (Patient Class)," "Gross Charge (Gross Charge)," "Discounted Cash Price (Discounted Cash Price)," "Payer Name (Insurer Name)," "Plan Name (Insurer Plan)," "Payer Specific Negotiated Charge: Dollar Amount (Price)," "De-identified Minimum Negotiated Charge (Min Negotiated Rate)," and "De-identified Maximum Negotiated Charge (Max Negotiated Rate)." Additionally, both our schema and the new CMS guidelines propose data types for each field name.

In our schema, which represents current hospitals' files, there are 5 field names absent in the new CMS guidelines "Revenue Description," "Revenue Code," "Package/Line Level," "Procedure ID," and "Self Pay." Conversely, the new CMS guidelines introduce 8 additional field names "Billing Class," "Drug Unit of Measurement," "Drug Type of Measurement," "Modifiers," "Payer Specific Negotiated Charge: Percentage," "Contracting Method," "Additional Generic Notes," and "Additional Payer-Specific Notes." We regard these new field names as providing further detailed information and enhancing consumer decision-making. If hospitals within a health system adopt consistent formats and can map their formats to the new CMS guidelines clearly in a mapping document they also provide, this can be more useful than the current optional guideline that is suggested.

In summary, since our analysis is based on the current data schema that hospitals have in place, we believe the schema we put out is easier to implement with minimal change to what the hospitals are currently doing. However, given the recent CMS guidelines, we recommend adding 8 additional fields as well as hospital-specific information.

Implications

The PT legislation aims to enable informed decision-making, reduce out-of-pocket expenses, and decrease overall health care expenditures. This study investigates the usage of current files by individuals and machines. Our results, unfortunately, suggest that PT data—as currently reported—appear to be neither useful for patients nor machines, raising important questions as to what these appear to be achieving today. Moreover, the findings indicate that even individuals with basic computer knowledge struggle with the usability of these files, highlighting the need for significant revisions to make them consumer-friendly and accessible to individuals of all technical proficiency levels. Additionally, inconsistencies in data representation between hospitals affiliated with different health systems pose challenges for machines, necessitating schema design improvements and the implementation of a standardized data representation. By addressing these concerns, PT legislation can achieve consistency and enhance machine readability, thus improving its effectiveness in promoting informed decision-making and reducing health care costs.

Although the official announcement of PT legislation is recent, prior studies [15-17] have attempted to evaluate the usability of PT, while subsequent studies [19-22] have examined the effectiveness of PT tools following the announcement. However, despite the introduction of PT rules, it appears that the usability of these files has not undergone significant improvements, indicating the necessity for proactive measures from responsible executives to ensure the effectiveness of this legislation. Our analysis of this matter emphasizes 2 primary factors—a lack of awareness among stakeholders and the challenges associated with using files due to inconsistencies in their format and representation.

As of April 2023, the CMS has issued over 730 warning notices and 269 requests for Corrective Action Plans. A total of 4 hospitals have faced Civil Monetary Penalties for noncompliance, and these penalties are publicly disclosed on the CMS website. The remaining hospitals subjected to comprehensive compliance reviews have either rectified their deficiencies or are actively engaged in doing so. While we acknowledge these efforts to comply with PT rules, our research revealed a notable disparity in data representation among hospitals affiliated with different health systems. Consequently, we focused on schema design and proposed the implementation of a master field name that encompasses a comprehensive data representation derived from an analysis of 36 hospitals. Standardizing the data representation across all health systems' machine-readable files will effectively address concerns about consistency. Therefore, significant modifications are required for the PT legislation to enhance machine readability and provide clearer guidance on the design and structure of the files' schema. If the hospital-provided information is consistent and of high quality, PT tools provided by health insurers may be able to estimate an individual's total expenses more accurately.

Limitations

Our objective was to have an equal number in both groups. However, in the case of the group tasked with obtaining information from the hospitals' websites, most did not finish the task and dropped out without completing it. This occurred because the task of retrieving the cost from the hospitals' websites in its current form is complex, as indicated by feedback from some participants. Only 19% (13/67) completed the task in that group (control group). Although this is a limitation of the study, it also highlights the complexity of obtaining cost information from hospitals' websites in the current form. In the treatment group, 81% (54 out of 67) of participants completed the task of retrieving the data, and the completion percentage was much higher.

Conclusions

Due to the poor usability and inconsistency of the formats, we, unfortunately, did not find evidence that the PT rule as implemented currently is useful to consumers, researchers, or policy makers (despite the legislation's goals that files are "consumer-friendly" and "machine-readable"). As 1 solution, we suggest a master field name for the data representation of machine-readable files to make them consistent, at least for the machines. Building tools that enable customers to estimate out-of-pocket costs is facilitated by having consistent machine-readable files across all health systems, which can be considered as future work for researchers and companies to help the PT rule reach its main goal, which is providing useful information for patients and reducing health care expenditures. In addition, another worthwhile approach to reducing some of

XSL•FO
RenderX

the exorbitant health care costs in the United States would be to integrate clinical decision support tools into the providers' workflow, triggered by orders for medications, diagnostic testing, and other billable services. In this regard, Bouayad et al [29] conducted experiments with physicians to demonstrate that PT, when included as part of the system they interact with, such as clinical decision support integrated into electronic health record systems, can significantly aid in cost reduction. This is a promising direction for practice but needs to be implemented carefully to avoid unanticipated consequences, such as scenarios where cost is incorrectly viewed as a proxy for quality, or where the use of this information introduces new biases for physicians and patients.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of Excel format of hospitals' files and our created Excel file.
[PDF File (Adobe PDF File), 1509 KB - jmir_v26i1e50629_app1.pdf]

Multimedia Appendix 2
Survey questions and experiment scenarios.
[PDF File (Adobe PDF File), 344 KB - jmir_v26i1e50629_app2.pdf]

Multimedia Appendix 3
Participants' responses chart regarding price transparency awareness.
[PDF File (Adobe PDF File), 253 KB - jmir_v26i1e50629_app3.pdf]

Multimedia Appendix 4
The *t* test analysis regarding human usability of price transparency information based on participants' responses.
[PDF File (Adobe PDF File), 120 KB - jmir_v26i1e50629_app4.pdf]

References

1. McGough M, Winger A, Rakshit S, Amin K. How has U.S. spending on healthcare changed over time? Health System Tracker. 2022. URL: https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/ [accessed 2024-03-13]
2. Christensen HB, Floyd E, Maffett M. The only prescription is transparency: the effect of charge-price-transparency regulation on healthcare prices. Manag Sci 2020;66(7):2861-2882. [doi: 10.1287/mnsc.2019.3330]
3. Muir MA, Alessi SA, King JS. Clarifying costs: can increased price transparency reduce healthcare spending? UC Hastings Research Paper No. 38 (SSRN) 2013 Feb 26:319-367 [FREE Full text] [doi: 10.2139/ssrn.2224151]
4. Reinhardt UE. Health care price transparency and economic theory. JAMA 2014;312(16):1642-1643. [doi: 10.1001/jama.2014.14276] [Medline: 25335145]
5. CY 2020 hospital Outpatient Prospective Payment System (OPPS) policy changes: hospital price transparency requirements (CMS-1717-F2). CMS.gov. 2020. URL: https://tinyurl.com/mrafxtvd [accessed 2024-03-13]
6. Secretary Azar statement on proposed rule for hospital price transparency. HHS.gov. 2020. URL: https://tinyurl.com/yc4dx2vx [accessed 2024-03-13]
7. Mehrotra A, Brannen T, Sinaiko AD. Use patterns of a state health care price transparency web site: what do patients shop for? Inquiry 2014;51:0046958014561496 [FREE Full text] [doi: 10.1177/0046958014561496] [Medline: 25466414]
8. Desai S, Hatfield LA, Hicks AL, Sinaiko AD, Chernew ME, Cowling D, et al. Offering a price transparency tool did not reduce overall spending among California public employees and retirees. Health Aff (Millwood) 2017;36(8):1401-1407 [FREE Full text] [doi: 10.1377/hlthaff.2016.1636] [Medline: 28784732]
9. Sinaiko AD, Joynt Maddox KE, Rosenthal MB. Association between viewing health care price information and choice of health care facility. JAMA Intern Med 2016;176(12):1868-1870 [FREE Full text] [doi: 10.1001/jamainternmed.2016.6622] [Medline: 27775768]
10. Desai S, Hatfield LA, Hicks AL, Chernew ME, Mehrotra A. Association between availability of a price transparency tool and outpatient spending. JAMA 2016;315(17):1874-1881 [FREE Full text] [doi: 10.1001/jama.2016.4288] [Medline: 27139060]
11. Sinaiko AD, Rosenthal MB. Examining a health care price transparency tool: who uses it, and how they shop for care. Health Aff (Millwood) 2016;35(4):662-670 [FREE Full text] [doi: 10.1377/hlthaff.2015.0746] [Medline: 27044967]

XSL•FO
RenderX

12. Desai SM, Shambhu S, Mehrotra A. Online advertising increased New Hampshire residents' use of provider price tool but not use of lower-price providers. Health Aff (Millwood) 2021;40(3):521-528 [FREE Full text] [doi: 10.1377/hlthaff.2020.01039] [Medline: 33646866]
13. Kona M, Corlette S. Hospital and insurer price transparency rules now in effect but compliance is still far away. Health Affairs Forefront. 2022. URL: hhttps://tinyurl.com/3x6ymxf2 [accessed 2024-03-13]
14. Wheeler C, Taylor R. New year, new CMS price transparency rule for hospitals. Health Affairs Forefront. 2021. URL: https://www.healthaffairs.org/content/forefront/new-year-new-cms-price-transparency-rule-hospitals [accessed 2024-03-13]
15. Chernew M, Cooper Z, Larsen-Hallock E, Morton FS. Are health care services shoppable? Evidence from the consumption of lower-limb MRI scans. National Bureau of Economic Research. 2021. URL: https://www.nber.org/papers/w24869 [accessed 2024-03-13]
16. Mehrotra A, Dean KM, Sinaiko AD, Sood N. Americans support price shopping for health care, but few actually seek out price information. Health Aff (Millwood) 2017;36(8):1392-1400 [FREE Full text] [doi: 10.1377/hlthaff.2016.1471] [Medline: 28784731]
17. Brown ZY. Equilibrium effects of health care price information. Rev Econ Stat 2019;101(4):699-712 [FREE Full text] [doi: 10.1162/rest_a_00765]
18. Wu SJ, Sylwestrzak G, Shah C, DeVries A. Price transparency for MRIs increased use of less costly providers and triggered provider competition. Health Aff (Millwood) 2014;33(8):1391-1398. [doi: 10.1377/hlthaff.2014.0168] [Medline: 25092841]
19. Glover M, Whorms DS, Singh R, Almeida RR, Prabhakar AM, Saini S, et al. A radiology-focused analysis of transparency and usability of top U.S. hospitals' chargemasters. Acad Radiol 2020;27(11):1603-1607. [doi: 10.1016/j.acra.2019.11.021] [Medline: 32014405]
20. Hayatghaibi SE, Alves VV, Ayyala RS, Dillman JR, Trout AT. Transparency and variability in pricing for pediatric outpatient imaging in US children's hospitals. JAMA Netw Open 2022;5(3):e220736 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.0736] [Medline: 35234885]
21. Arvisais-Anhalt S, McDonald S, Park JY, Kapinos K, Lehmann CU, Basit M. Survey of hospital chargemaster transparency. Appl Clin Inform 2021;12(2):391-398 [FREE Full text] [doi: 10.1055/s-0041-1729168] [Medline: 33951742]
22. Reddy S, Daly G, Baban S, Kadesh A, Block AE, Grimes CL. Accessibility and usability of hospital chargemasters in New York state. J Gen Intern Med 2022;37(8):2130-2131 [FREE Full text] [doi: 10.1007/s11606-021-07173-x] [Medline: 34713387]
23. Robinson JC. Hospital quality competition and the economics of imperfect information. Milbank Q 1988;66(3):465-481. [Medline: 3211089]
24. Patel KN, Mazurenko O, Ford E. Analysis of hospital quality measures and web-based chargemasters, 2019: cross-sectional study. JMIR Form Res 2021;5(8):e26887 [FREE Full text] [doi: 10.2196/26887] [Medline: 34420914]
25. Batty M, Ippolito B. Mystery of the chargemaster: examining the role of hospital list prices in what patients actually pay. Health Aff (Millwood) 2017;36(4):689-696 [FREE Full text] [doi: 10.1377/hlthaff.2016.0986] [Medline: 28373335]
26. Muhlestein DB, Wilks CEA, Richter JP. Limited use of price and quality advertising among American hospitals. J Med Internet Res 2013;15(8):e185 [FREE Full text] [doi: 10.2196/jmir.2660] [Medline: 23988296]
27. Batini C, Lenzerini M, Navathe SB. A comparative analysis of methodologies for database schema integration. ACM Comput Surv 1986;18(4):323-364 [FREE Full text] [doi: 10.1145/27633.27634]
28. Voluntary hospital price transparency machine-readable file sample format data dictionary (version 1.1). CMS.gov. URL: https://www.cms.gov/files/document/hospital-price-transparency-machine-readable-data-dictionary-tall.pdf [accessed 2024-03-13]
29. Bouayad L, Padmanabhan B, Chari K. Can recommender systems reduce healthcare costs? the role of time pressure and cost transparency in prescription choice. MIS Q 2020;44(4):1859-1903. [doi: 10.25300/misq/2020/14435/]

Abbreviations

CMS: Centers for Medicare & Medicaid Services
CPT: Current Procedural Terminology
DRG: Diagnosis-Related Group
ERD: entity-relationship diagram
HMO: Health Maintenance Organization
MRI: magnetic resonance imaging
PPO: Preferred Provider Organization
PT: price transparency
PTR: price transparency regulation

XSL•FO
RenderX

Edited by S He; submitted 07.07.23; peer-reviewed by KN Patel, R Marshall, G Deckard; comments to author 03.12.23; revised version received 21.01.24; accepted 26.02.24; published 29.03.24.

Please cite as:
Maleki N, Padmanabhan B, Dutta K
Usability of Health Care Price Transparency Data in the United States: Mixed Methods Study
J Med Internet Res 2024;26:e50629
URL: https://www.jmir.org/2024/1/e50629
doi:10.2196/50629
PMID:38442238

©Negar Maleki, Balaji Padmanabhan, Kaushik Dutta. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cost Evaluation of the Ontario Virtual Urgent Care Pilot Program: Population-Based, Matched Cohort Study

Jean-Eric Tarride[1,2,3], PhD; Justin N Hall[4,5,6], MD, MSc, MPH, MBA; Shawn Mondoux[7,8,9], MSc, MD; Katie N Dainty[6,10], PhD; Joy McCarron[11], BScN; J Michael Paterson[6,12], MSc; Lesley Plumptre[12], PhD; Emily Borgundvaag[12], MPH; Howard Ovens[13,14], MD; Shelley L McLeod[13,14], MSc, PhD

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

[2]Centre for Health Economics and Policy Analysis, McMaster University, Hamilton, ON, Canada

[3]Programs for Assessment of Technology in Health, The Research Institute of St. Joe's Hamilton, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada

[4]Department of Emergency Medicine, Sunnybrook Health Sciences Centre, Toronto, ON, Canada

[5]Division of Emergency Medicine, Department of Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[6]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[7]Division of Emergency Medicine, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada

[8]Department of Medicine, McMaster University, Hamilton, ON, Canada

[9]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[10]North York General Hospital, Toronto, ON, Canada

[11]Ontario Health, Toronto, ON, Canada

[12]ICES, Toronto, ON, Canada

[13]Schwartz/Reisman Emergency Medicine Institute, Sinai Health, Toronto, ON, Canada

[14]Division of Emergency Medicine, Department of Family and Community Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Jean-Eric Tarride, PhD
Centre for Health Economics and Policy Analysis
McMaster University
Communication Research Laboratory 227
1280 Main Street West
Hamilton, ON, L8S 4K1
Canada
Phone: 1 905 522 9140 ext 20205
Email: tarride@mcmaster.ca

Abstract

Background: In 2020, the Ministry of Health (MoH) in Ontario, Canada, introduced a virtual urgent care (VUC) pilot program to provide alternative access to urgent care services and reduce the need for in-person emergency department (ED) visits for patients with low acuity health concerns.

Objective: This study aims to compare the 30-day costs associated with VUC and in-person ED encounters from an MoH perspective.

Methods: Using administrative data from Ontario (the most populous province of Canada), a population-based, matched cohort study of Ontarians who used VUC services from December 2020 to September 2021 was conducted. As it was expected that VUC and in-person ED users would be different, two cohorts of VUC users were defined: (1) those who were promptly referred to an ED by a VUC provider and subsequently presented to an ED within 72 hours (these patients were matched to in-person ED users with any discharge disposition) and (2) those seen by a VUC provider with no referral to an in-person ED (these patients were matched to patients who presented in-person to the ED and were discharged home by the ED physician). Bootstrap techniques were used to compare the 30-day mean costs of VUC (operational costs to set up the VUC program plus health care expenditures) versus in-person ED care (health care expenditures) from an MoH perspective. All costs are expressed in Canadian dollars (a currency exchange rate of CAD $1=US $0.76 is applicable).

XSL•FO
RenderX

Results: We matched 2129 patients who presented to an ED within 72 hours of VUC referral and 14,179 patients seen by a VUC provider without a referral to an ED. Our matched populations represented 99% (2129/2150) of eligible VUC patients referred to the ED by their VUC provider and 98% (14,179/14,498) of eligible VUC patients not referred to the ED by their VUC provider. Compared to matched in-person ED patients, 30-day costs per patient were significantly higher for the cohort of VUC patients who presented to an ED within 72 hours of VUC referral ($2805 vs $2299; difference of $506, 95% CI $139-$885) and significantly lower for the VUC cohort of patients who did not require ED referral ($907 vs $1270; difference of $362, 95% CI 284-$446). Overall, the absolute 30-day costs associated with the 2 VUC cohorts were $18.9 million (ie, $6.0 million + $12.9 million) versus $22.9 million ($4.9 million + $18.0 million) for the 2 in-person ED cohorts.

Conclusions: This costing evaluation supports the use of VUC as most complaints were addressed without referral to ED. Future research should evaluate targeted applications of VUC (eg, VUC models led by nurse practitioners or physician assistants with support from ED physicians) to inform future resource allocation and policy decisions.

(J Med Internet Res 2024;26:e50483) doi:10.2196/50483

KEYWORDS

virtual urgent care; health care expenditures; Canada; virtual care; economic evaluation; pilot program; pilot; Ontario; urgent care; care; emergency department; users; patient; patients; resources; resource allocation; policy decision; decision-making; policy

Introduction

Although virtual care has been available for several decades, it was not widely used before the COVID-19 pandemic. In Ontario, the most populous province in Canada (2021 population: 14.2 million [1]), the Ontario Telemedicine Network (OTN) was the only fully funded virtual care model prior to the COVID-19 pandemic. Historically, two-thirds of OTN use has been for mental health and substance use [2]. In the second quarter of 2019, OTN accounted for approximately 1.6% [3] of all ambulatory care visits. Following the first wave of the COVID-19 pandemic and the introduction of physician billing codes to support virtual care in Ontario, the percentage of ambulatory care visits in Ontario delivered by virtual care in the second quarter of 2020 has increased to 70.6% [3]. In parallel, between the pre–COVID-19 period (March 12 to July 29, 2019) and the COVID-19 period (March 11 to July 28, 2020), the number of in-person physician office visits decreased by approximately 80% [4].

In the fall of 2020, the Ontario Ministry of Health (MoH) introduced a virtual urgent care (VUC) pilot program to provide alternative access to urgent care services and reduce the need for in-person emergency department (ED) visits for patients with low acuity health concerns [5]. Hospitals interested in offering VUC services were invited to submit a funding application to Ontario Health, Ontario's health care administrative agency. Following a review and feedback process, the MoH approved $4 million of operational funding to support 14 ED-led VUC programs in the 5 health regions of Ontario. As part of this funding, participating sites were expected to launch their VUC programs within 1 to 2 months of funding approval and participate in a mixed methods evaluation of the VUC pilot program to inform future policy and funding decisions. The characteristics of the VUC sites [6] and patient characteristics [7] are described elsewhere but are briefly summarized here. The 14 VUC sites included a mix of pediatric, northern, urban, academic, and community sites across Ontario. Since sites were responsible for the design and implementation of the VUC program to meet their local needs, models of care differed between sites (eg, clinical and operational governance, triage, technology, staffing, operating hours, marketing, and communications) and sites launched their VUC services at different start dates [6].

There were 22,278 VUC encounters between December 2020 and September 2021, and the median (IQR) duration of the virtual visit was 14 (11-18) minutes [7]. VUC users were more likely to be younger, female, English speaking, and have a postsecondary or higher education and a primary care physician. Most complaints were managed by the VUC provider without the need to refer the patient to an in-person ED visit [7]. Other results based on 1:1 propensity score matching (PSM) methods indicated that 30-day health care use (eg, hospitalization, ED, and physician visits) was similar between 2129 VUC users who presented to an ED within 72 hours of VUC referral and 2129 matched patients presenting to the ED in person [8]. VUC users not promptly referred to an ED (N=14,179) were more likely to have an ED visit within 72 hours, 7 days, and 30 days following the index VUC visit compared to matched in-person ED individuals [8]. However, 30-day hospital admissions were lower for the VUC group not promptly referred to the ED, but the length of hospital admission was longer than for patients presenting to the ED in person [8].

To inform future investments in VUC in Canada and elsewhere, here, we present the results from the costing analysis of the VUC pilot program that consider both funding provided by the MoH in Ontario to operate the VUC program and direct costs associated with health care resource use during the 30 days following each VUC visit. Our research hypotheses were that (1) VUC users promptly referred to an in-person ED by their VUC provider incur higher costs than matched in-person ED users with any type of discharge and (2) VUC users who were not referred to an in-person ED by their VUC provider incur lower 30-day costs than matched in-person ED users discharged home by the ED physician.

Methods

Study Design

We conducted [8] a population-based, matched cohort study of Ontarians who used VUC services provided by 12 of the 14

XSL•FO
RenderX

ED-led pilot sites from December 2020 to September 2021. Records from 2 sites were excluded as 1 site had a provider-to-provider VUC model (ie, physicians or nurses consulting with ED physicians) and the other site was delayed in launching the VUC program. These data sets were linked using unique encoded identifiers and analyzed at ICES (formerly the Institute for Clinical Evaluative Sciences). Patient-level VUC encounter data were linked to administrative health care databases (Multimedia Appendix 1) held at ICES to compare the 30-day costs of VUC versus in-person ED care from an MoH perspective. ICES data contain all health care services covered by the Ontario Health Insurance Plan, which include all hospital care, ED visits, hospital-based outpatient specialty clinics, physician visits for all residents of Ontario (40% of the Canadian population), and prescription drugs for residents aged 65 years and older and social assistance recipients. The analyses and reporting follow the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [9].

Ethical Considerations

The study protocol for the collection of VUC data was reviewed and approved by the Research Ethics Board at Sinai Health (21-003E), Toronto, Ontario, Canada. Individual resident consent was not required to access and use provincial health care records for this study, as it was conducted at ICES, an independent, not-for-profit health services research institute whose legal status under Ontario's Personal Health Information Protection Act authorizes it to collect personal health information, without consent, for the purposes of analysis or compiling statistical information with respect to the management of, evaluation or monitoring of, the allocation of resources to or planning for all or part of the health system. All data analyses were performed at ICES, where the data are securely held in linked, coded (deidentified) form.

Population

The study population included both pediatric and adult Ontario residents visiting 1 of the 12 included VUC pilot programs or any Ontario in-person ED from December 2020 to September 2021. Non-Ontario residents, individuals without a valid Ontario Health Insurance Plan number, and those who left without seeing a VUC provider were excluded from the analyses. We also excluded individuals who did not show up to an in-person ED within 72 hours after being promptly referred to the ED as it was deemed that an ED visit after 72 hours was less likely to be directly related to the initial visit issue (or at least the extent or severity of the issue at that time). As some individuals may have multiple VUC encounters, the first VUC encounter was considered the index VUC visit.

Since VUC users were expected to be different from patients who presented in person to the ED, VUC users were split into those who presented to an ED within 72 hours of VUC referral and those who saw a VUC provider with no referral to an in-person ED ("discharged home" by the VUC provider) [8]. Using propensity scoring matching methods, the first group of VUC users who presented to an ED within 72 hours of VUC referral was matched with a comparable group of individuals presenting in-person to an ED with any discharge disposition (eg, admitted, transferred, and discharged home). The second group of VUC users discharged home was matched with a comparable group of individuals presenting in person to an ED and discharged home by the ED physician. In addition, to generate comparable groups of VUC and in-person ED users based on their discharge disposition, this stratification allowed each matched group to have an equal opportunity to consume health care resources and incur costs.

Primary and Secondary Outcomes

The primary outcome was 30-day MoH costs following the first VUC encounter or matched in-person ED encounter. Secondary outcomes were 30-day health care resource use (hospitalizations, ED visits, same-day surgeries, physician visits, outpatient publicly funded drugs, laboratory tests, and nonphysician visits) and associated health care expenditures.

Statistical Analysis

To match our 2 subgroups of VUC patients with comparable patients who presented in person to an ED, we used a greedy nearest-neighbor 1:1 matching based on the index encounter date (±14 days), day (weekday vs weekend), and time of registration (day: 8 AM-8 PM; night: 8 PM-8 AM), presenting Canadian Emergency Department Information System complaint [10] and the logit of a propensity score. Canadian Emergency Department Information System includes 169 presenting complaints divided into 20 categories [10]. The variables included in the propensity score were age; sex; Statistics Canada Census neighborhood income quintile; urban or rural residence status; Ontario Marginalization Index [11] quintiles related to ethnic concentration, residential instability, material deprivation, and dependency; formal rostering with a family physician; the number of major ACG System Aggregated Diagnosis Groups derived from the Johns Hopkins ACG System version 10 (Health Services Research & Development Center, The Johns Hopkins University, Bloomberg School of Public Health) grouped as 0-4, 5-9, 10-14, 15+; medical conditions derived from validated administrative data algorithms (asthma, congestive heart failure, hypertension, chronic obstructive pulmonary disease, diabetes, and dementia); and the number of physician visits, ED visits, and hospitalizations in the 365 days preceding the index date. A caliper width of 0.2 of the SD of the logit of the propensity score was used. Balance in baseline covariates for each cohort was evaluated using standardized differences, with values less than 0.10 indicating that the groups were well matched [12].

Per-patient mean (SD) 30-day health care expenditures (eg, ED, physician, and hospitalization) expressed in 2020-2021 Canadian dollars (a currency exchange rate of CAD $1=US $0.76 is applicable) were calculated for each cohort using standardized costing algorithms for Ontario administrative health care data in Ontario [13]. The costing methodology uses a bottom-up or micro-costing approach to cost services at the individual level and costs represent amounts paid by the Ontario MoH [13]. Operational funding provided by the MoH to establish the VUC programs was added to the 30-day health care expenditures for the VUC cohort using site-specific information (ie, total operational funds received by each hospital divided by the total number of VUC visits to each hospital). Bootstrap techniques were used to generate 95% CIs associated with the difference in 30-day costs per patient between the VUC cohorts

XSL•FO
RenderX

(operational funds plus health care expenditures) and matched in-person ED cohorts (health care expenditures). The same approach was used to analyze differences in secondary outcomes. In addition to the per-patient analyses, we present the total absolute costs for each cohort. All analyses were performed at ICES using linked, encoded data using SAS (version 9.4; SAS Institute).

Results

Study Population

There were 22,278 patient encounters in the pilot program, of which 19,595 patient VUC encounters were available for data linkage. This included 2931 (15%) VUC encounters that resulted in a referral to an in-person ED and 16,664 (85%) VUC encounters that did not result in a referral to an in-person ED. After applying our eligibility criteria, records from 2150 patients promptly referred to the ED and who presented to the ED within 72 hours (Multimedia Appendix 2) and records from 14,498 patients who were not referred to an in-person ED by their VUC provider (Multimedia Appendix 3) were available for the matched analyses. An additional 669 patients promptly referred to the ED by their VUC provider did not present to an ED within 72 hours and were excluded from the matched analysis as per our exclusion criteria. After 1:1 matching, 2129 patients who presented to the ED within 72 hours of a referral by a VUC provider and 14,179 patients seen by a VUC provider with no referral to an in-person ED were matched to an equal number of in-person ED individuals. Our matched VUC cohorts represented 99% (2129/2150) and 98% (14,179/14,498) of all eligible VUC patients, respectively.

VUC Patients Who Were Promptly Referred to the ED and Who Presented to the ED Within 72 Hours and Matched In-Person ED Controls

Pediatric sites accounted for approximately 41% (876/2129) of the cohort of VUC users promptly referred to the ED and who presented to the ED within 72 hours of being referred by their VUC provider. The mean (SD) age of the VUC cohort was 29.1 (24.9) years, and 10% (222/2129) were aged 65 years or older. Approximately 61% (1296/2129) of VUC users were female, 96% (2052/2129) were living in urban areas, and 96% (2050/2129) were rostered to a primary care physician. Upon ED presentation, 80% of acuity scores corresponded to Canadian Triage and Acuity Scale levels 3 (urgent: 1242/2129, 58%), 4 (less urgent: 402/2129, 19%), and 5 (nonurgent: 109/2129, 5%). The most common medical conditions of the VUC cohort were asthma (424/2129, 20%), hypertension (346/2129, 16%), and diabetes (175/2129, 8%). The mean (SD) numbers of in-person ED visits and hospitalizations in the preceding 365 days for this VUC cohort promptly referred to the ED were 1.0 (3.0) and 0.3 (0.7), respectively. Table 1 presents the baseline characteristics of our matched VUC users who presented to an ED within 72 hours of VUC referral and in-person ED cohorts (n=2129 each), which were well-balanced.

The 30-day health care expenditures were similar between the 2 cohorts. However, the 30-day MoH costs per patient were relatively greater for the VUC cohort (mean $2805, SD $7026, including mean $163, SD $99 of operational funds and mean $ 2642, SD $7017 of health care expenditures) compared to the matched in-person ED cohort (mean $2299, SD $6174), resulting in a difference of $506 (95% CI $139-$885) per patient (Table 2).

XSL•FO
RenderX

Table 1. Baseline characteristics of patients referred to an ED[a] by a virtual urgent care provider and who presented to the ED within 72 hours and matched in-person ED controls[b].

Characteristics	Virtual urgent care (n=2129)	In-person ED care (n=2129)	Standardized difference[c]
Age (years), mean (SD)	29.1 (24.9)	28.2 (22.1)	0.04
Age (years), n (%)			
<18	876 (41.2)	781 (36.7)	0.09
18-64	1031 (48.4)	1175 (55.2)	0.14
≥65	222 (10.4)	173 (8.1)	0.08
Female, n (%)	1296 (60.9)	1322 (62.1)	0.03
Living in an urban setting, n (%)	2052 (96.4)	2043 (96)	0.02
Neighborhood income quintile, n (%)			
High (quintiles 4-5)	976 (45.8)	958 (45)	0.02
Other	1153 (54.2)	1171 (55)	0.02
Acuity score,[d] n (%)			
Resuscitation (CTAS[e] 1)	11 (0.5)	11-15[f]	0.01
Emergent (CTAS 2)	365 (17.1)	423 (19.9)	0.07
Urgent (CTAS 3)	1242 (58.3)	1161 (54.5)	0.08
Less-urgent (CTAS 4)	402 (18.9)	417 (19.6)	0.02
Nonurgent (CTAS 5)	109 (5.1)	114 (5.3)	0.01
Unknown	0 (0.0)	1-5[f]	0.04
Selected medical conditions, n (%)			
Asthma	424 (19.9)	397 (18.7)	0.03
Congestive heart failure	60 (2.8)	48 (2.3)	0.04
Hypertension	346 (16.3)	287 (13.5)	0.08
Chronic obstructive pulmonary disease	39 (1.8)	44 (2.1)	0.02
Diabetes	175 (8.2)	167 (7.8)	0.01
Dementia	30 (1.4)	23 (1.1)	0.03
Johns Hopkins ACG System ADGs[g], n (%)			
0-4	687 (32.3)	673 (31.6)	0.01
5-9	925 (43.5)	977 (45.9)	0.05
10-14	407 (19.1)	383 (18)	0.03
15+	110 (5.2)	96 (4.5)	0.03
Rostered to a primary care physician, n (%)	2050 (96.3)	2021 (94.9)	0.07
Number of physician (GP[h] and specialists) visits in preceding 365 days, mean (SD)	9.8 (13.4)	8.9 (11.6)	0.07
Number of ED visits in proceeding 365 days, mean (SD)	1.0 (3.0)	0.9 (1.7)	0.05
Number of hospitalizations in preceding 365 days, mean (SD)	0.3 (0.7)	0.3 (0.8)	0.00

[a]ED: emergency department.

[b]In addition, the cohorts were matched in function of the Ontario Marginalization (ONMARG) Ethnic Diversity quintile, ONMARG Residential Instability quintile, ONMARG Material Deprivation, and ONMARG Dependency quintile (data not shown).

[c]Standardized differences greater than 0.1 are generally considered meaningful.

[d]For the cohort of VUC users referred to an ED and who presented to the ED within 72 hours, CTAS was evaluated at the time of the ED visit.

[e]CTAS: Canadian Triage and Acuity Scale.

[f]Cells are suppressed to protect patient privacy.

XSL•FO
RenderX

[g]ADG: aggregated diagnosis group.

[h]GP: general practitioner.

Table 2. The 30-day costs per patient associated with patients referred to an ED[a] by a virtual urgent care provider and who presented to the ED within 72 hours and matched in-person ED controls.

Costs and health care resource use	VUC[b], mean (SD)	In-person ED care, mean (SD)	VUC versus in-person ED care, mean difference (95% CI)
Hospitalizations (CAD $)[c]	$1367.33 ($6071.73)	$1179.76 ($5119.87)	$187.57 (–$123.25 to $504.81)
Number of hospitalizations	0.2 (0.4)	0.1 (0.4)	0.0 (0.0 to 0.0)
Same-day surgeries (CAD $)	$41.67 ($328.62)	$26.49 ($226.84)	$15.18 (–$1.15 to $31.03)
Number of same-day surgeries	0.0 (0.2)	0.0 (0.2)	0.0 (0.0 to 0.0)
ED visits (CAD $)	$472.99 ($414.14)	$436.67 ($369.78)	$36.32 ($14.91 to $57.95)
Number of ED visits	1.3 (0.7)	1.3 (0.6)	0.0 (–0.0 to 0.1)
Outpatient publicly funded drugs (CAD $)	$53.96 ($344.63)	$62.61 ($357.82)	–$8.65 (–$30.47 to $11.69)
Number of claims	1.2 (5.1)	1.7 (9.5)	–0.5 (–0.9 to –0.1)
Index day (day 1) physician visits[d,e] (CAD $)	$212.08 ($176.23)	$182.39 ($214.57)	$29.69 ($17.99 to $41.51)
Day 1-30 physician visits[f] (CAD $)	$691.49 ($997.00)	$574.90 ($1033.88)	$116.59 ($58.44 to $175.07)
GP[g] visits (CAD $)	$74.75 ($178.08)	$62.35 ($139.29)	$12.40 ($3.60 to $21.55)
Number of GP visits	1.2 (1.6)	0.9 (1.4)	0.3 (0.2 to 0.4)
Specialist visits (CAD $)	$390.31 ($842.63)	$339.00 ($883.46)	$51.31 ($4.57 to $102.53)
Number of specialist visits	1.5 (1.5)	0.7 (1.2)	0.8 (0.7 to 0.9)
Other costs (shadow billing, etc; CAD $)	$215.94 ($345.02)	$165.42 ($348.67)	$50.52 ($29.59 to $71.45)
Capitation costs (CAD $)	$14.54 ($18.26)	$18.39 ($16.40)	–$3.85 (–$4.77 to –$2.97)
Laboratory test (CAD $)	$8.84 ($34.94)	$7.09 ($26.04)	$1.75 ($0.09 to $3.62)
Number of tests	1.4 (4.4)	1.3 (4.5)	0.2 (–0.1 to 0.4)
Nonphysician visits (CAD $)	$1.65 ($36.36)	$1.04 ($8.42)	$0.61 (–$0.53 to $2.63)
Number of nonphysician visits	0.0 (0.3)	0.0 (0.3)	0.0 (–0.0 to 0.0)
Total health care expenditures (CAD $)	$2641.97 ($7016.61)	$2298.82 ($6173.60)	$343.15 (–$23.54 to 719.88)
Operating funds (CAD $)	$163.24 ($98.64)	N/A[h]	N/A
Total costs (CAD $)	$2805.21 ($7,026.03)	$2298.82 ($6173.60)	$506.39 ($139.10 to $885.16)

[a]ED: emergency department.

[b]VUC: virtual urgent care.

[c]A currency exchange rate of CAD $1=US $0.76 is applicable.

[d]Due to the absence of the exact time of the billings, the "index day physician" visits include all physician visits (eg, general practitioners and specialists) that occurred during the index day (day 1).

[e]The costs associated with physician visits during the index day are included in the 30-day costs associated with physician visits (days 1-30).

[f]Physician costs are composed of the prices and quantity of each service billed by physicians to Ontario Health Insurance Plan, as well as shadow billing costs and capitation costs.

[g]GP: general practitioner.

[h]N/A: not applicable.

VUC Patients Not Referred to an In-Person ED and Matched In-Person ED Controls

The baseline characteristics of the VUC cohort not promptly referred to an in-person ED by their VUC provider were similar to VUC users promptly referred to the ED by their VUC provider. Table 3 presents the baseline characteristics of our matched VUC and in-person ED cohorts (n=14,179 each), which were well balanced.

The 30-day mean (SD) MoH health care expenditures per patient were lower with VUC (mean $758, SD $3129) than in-person ED care (mean $1270, SD $3846; $P<.001$), resulting in a

XSL•FO
RenderX

difference of $511 (95% CI $434-$595) per patient in favor of VUC. More than 90% ($480/$511) of this cost difference was explained by a significant reduction in 30-day ED costs ($237/$511, 46% of total cost reduction) and physician visits ($243/$511, 47%). After adding the VUC operational funding (mean $149, SD $95 per VUC user), VUC remained less expensive than in-person ED care. The details are provided in Table 4.

Table 3. Baseline characteristics of patients seen by a virtual urgent care provider with no referral to an in-person ED[a] and matched in-person ED controls[b].

Characteristics	Virtual urgent care (n=14,179)	In-person ED care (n=14,179)	Standardized difference[c]
Age (years), mean (SD)	28.1 (23.5)	27.6 (22.5)	0.02
Age (years), n (%)			
<18	5640 (39.8)	5485 (38.7)	0.02
18-64	7386 (52.1)	7633 (53.8)	0.03
≥65	1153 (8.1)	1061 (7.5)	0.02
Female, n (%)	8482 (59.8)	8704 (61.4)	0.03
Living in an urban setting, n (%)	13,533 (95.4)	13,330 (94)	0.06
Neighborhood income quintile, n (%)			
High (quintiles 4-5)	6210 (43.9)	6202 (43.7)	0.00
Other	7969 (56.2)	7977 (56.3)	0.00
Selected medical conditions, n (%)			
Asthma	2778 (19.6)	2684 (18.9)	0.02
Congestive heart failure	232 (1.6)	252 (1.8)	0.01
Hypertension	1939 (13.7)	1801 (12.7)	0.03
Chronic obstructive pulmonary disease	223 (1.6)	247 (1.7)	0.01
Diabetes	1051 (7.4)	1161 (8.2)	0.03
Dementia	133 (0.9)	83 (0.6)	0.04
Johns Hopkins ADGs[d], n (%)			
0-4	4777 (33.7)	4527 (31.9)	0.04
5-9	6427 (45.3)	6617 (46.7)	0.03
10-14	2460 (17.4)	2468 (17.4)	0.00
≥15	515 (3.6)	567 (4)	0.02
Rostered to a primary care physician, n (%)	13,745 (96.9)	13,708 (96.7)	0.01
Number of physician (GP[e] and specialists) visits in the preceding 365 days, mean (SD)	8.5 (10.8)	8.4 (11.2)	0.01
Number of ED visits in the proceeding 365 days, mean (SD)	0.9 (2.3)	1.0 (1.7)	0.03
Number of hospitalizations in the preceding 365 days, mean (SD)	0.2 (0.7)	0.2 (0.7)	0.03

[a]ED: emergency department.

[b]In addition, the cohorts were matched in function of the Ontario Marginalization (ONMARG) Ethnic Diversity quintile, ONMARG Residential Instability quintile, ONMARG Material Deprivation, and ONMARG Dependency quintile (data not shown).

[c]Standardized differences greater than 0.1 are generally considered meaningful.

[d]ADG: aggregated diagnosis groups.

[e]GP: general practitioner.

XSL•FO
RenderX

Table 4. The 30-day costs per patient associated with patients seen by a VUC[a] provider with no referral to an in-person ED[b] and matched in-person ED controls.

Costs and health care resource use	VUC, mean (SD)	In-person ED care, mean (SD)	VUC versus in-person ED care, mean difference (95% CI)
Hospitalizations (CAD $[c])	$274.70 ($2485.70)	$356.64 ($3076.53)	–$81.94 (–$150.59 to -$19.08)
Number of hospitalizations	0.0 (0.2)	0.0 (0.2)	–0.0 (–0.0 to –0.0)
Same-day surgeries (CAD $)	$23.96 ($252.67)	$47.22 ($367.97)	–$23.26 (–$30.41 to –$15.81)
Number of same-day surgeries	0.0 (0.1)	0.0 (0.2)	–0.0 (–0.0 to –0.0)
ED visits (CAD $)	$127.00 ($288.33)	$364.42 ($356.44)	–$237.42 (–$244.24 to –$230.58)
Number of ED visits	0.4 (0.8)	1.3 (0.8)	–0.9 (–0.9 to –0.9)
Outpatient publicly funded drugs (CAD $)	$45.52 ($394.42)	$65.13 ($445.68)	–$19.61 (–$29.29 to –$10.27)
Number of claims	1.2 (6.0)	1.5 (6.5)	–0.3 (–0.5 to –0.2)
Index day (day 1) physician visits[d,e] (CAD $)	$54.67 ($73.88)	$152.65 ($136.93)	–97.98 (–$100.62 to –$95.59)
Day 1-30 physician visits[f] (CAD $)	$272.65 ($601.14)	$417.20 ($724.25)	–144.55 (–$159.72 to –$129.64)
GP[g] visits (CAD $)	$50.33 ($136.31)	$54.85 ($153.31)	–$4.52 (–$7.69 to –$1.35)
Number of GP visits	1.2 (1.6)	0.9 (1.4)	0.3 (0.2 to 0.3)
Specialist visits (CAD $)	$151.50 ($501.17)	$212.01 ($601.28)	–$60.51 (–$72.95 to –$47.99)
Number of specialist visits	1.0 (1.3)	0.7 (1.3)	0.3 (0.3 to 0.4)
Other costs (CAD $)	$62.00 ($159.04)	$142.06 ($219.03)	–$80.06 (–$84.61 to –$75.50)
Capitation costs (CAD $)	$14.59 ($17.31)	$19.09 ($15.89)	–$4.50 (–$4.88 to –$4.13)
Laboratory test (CAD $)	$7.86 ($28.91)	$7.03 ($26.68)	$0.83 ($0.17 to $1.45)
Number of tests	1.5 (4.7)	1.3 (4.3)	0.2 (0.1 to 0.3)
Nonphysician visits (CAD $)	$0.98 ($9.00)	$1.24 ($16.74)	–$0.26 (–$0.59 to $0.03)
Number of nonphysician visits	0.0 (0.3)	0.0 (0.3)	–0.0 (–0.0 to 0.0)
Total health care expenditures (CAD $)	$758.37 ($3128.81)	$1269.70 ($3846.26)	–$511.33 (–$595.26 to -$433.55)
Operating funds (CAD $)	$149.12 ($94.91)	N/A[h]	N/A
Total costs (CAD $)	$907.49 ($3133.39)	$1269.70 ($3846.26)	–$362.21 (–$446.26 to -$284.24)

[a]VUC: virtual urgent care.

[b]ED: emergency department.

[c]A currency exchange rate of CAD $1=US $0.76 is applicable.

[d]Due to the absence of the exact time of the billings, the "index day physician" visits include all physician visits (eg, GP and specialists) that occurred during the index day (day 1).

[e]The costs associated with physician visits during the index day are included in the 30-day costs associated with physician visits (days 1-30).

[f]Physician costs are composed of the prices and quantity of each service billed by physicians to Ontario Health Insurance Plan, as well as shadow billing costs and capitation costs.

[g]GP: general practitioner.

[h]N/A: not applicable.

Total Absolute Costs for the VUC and In-Person ED Cohorts

At the cohort level, the absolute 30-day cost associated with the cohort of VUC users who presented to an ED within 72 hours of VUC referral and matched in-person ED patients (n=2129 each) were $6.0 million and $4.9 million, respectively. For the cohort of VUC users not referred to the ED by their VUC provider and the matched cohort of individuals attending in-person to the ED and discharged home by their ED physician (n=14,179 each), the absolute 30-day costs were $12.9 million and $18.0 million, respectively. Overall, the absolute 30-day costs associated with the two VUC cohorts were $18.9 million (ie, $6.0 million + $12.9 million) versus $22.9 million ($4.9 million + $18.0 million) for the 2 in-person ED cohorts.

XSL•FO
RenderX

Discussion

Principal Findings

The data from this Ontario VUC pilot program indicate that approximately 15% (2931/19,595) of VUC encounters were promptly referred to an in-person ED. Our costing analysis of 2129 patients who presented to the ED within 72 hours of a referral by a VUC provider indicated that there were no statistically significant differences in 30-day health care expenditures when compared to matched in-person ED patients. However, when the operational costs to set up the VUC programs were considered, the VUC cohort promptly referred to the ED incurred significantly higher costs. Due to differences in ED and physician costs, the 30-day costs of the cohort of VUC users not promptly referred to the ED were significantly lower than those of a matched cohort of in-person ED users.

These findings are important for several reasons. First, the data from the Ontario VUC pilot program evaluation suggest that VUC models may represent a good investment of government funding, especially if the operational costs to set up and maintain VUC models decrease over time with subsequent implementation efficiencies. However, considering that most VUC complaints were dealt with by the VUC provider with no referral to the ED, other models of VUC where nurse practitioners, physician assistants, or family physicians provide first-line VUC instead of emergency physicians may be more attractive, especially in the context of current severe ED workforce shortages. Finally, this study proposes a costing methodology to deal with both the heterogeneity between VUC users and non-VUC users, which allows more meaningful comparisons.

Comparison With Prior Work

To our knowledge, this study is the first to evaluate the economic costs of VUC in Canada. Although several US studies have shown telemedicine or VUC for low acuity patients save costs from a health system perspective [14-17] or patient perspective [18], these results may not be generalizable to the Canadian health care system and these studies predate the COVID-19 pandemic. While recent Canadian data from April 2021 to March 2022 indicate the number of ED visits has almost returned to pre–COVID-19 levels [19], EDs across Canada continue to experience great pressures due to severe staffing shortages [20]. In this sense, VUC models could serve to complement in-person ED care.

Strengths and Limitations

Compared to previous economic studies of virtual care, this study has several strengths. First, we had access to linked records of all Ontarians who used VUC services across the VUC pilot programs, which allowed us to document the 30-day health care use and costs. Second, we were able to match 98% (16,308/16,648) of all eligible VUC users. Third, we separated the VUC users into 2 cohorts based on their discharge status following the VUC encounter ("discharged home" or "referred to in-person ED") and we matched each VUC cohort with comparable in-person ED users (eg, the VUC users not referred to an in-person ED by their VUC provider were matched to comparable in-person ED individuals discharged home by their ED physician), which allowed us to decrease the heterogeneity between patients using VUC and in-person ED care. Finally, we used PSM methods and bootstrap techniques to compare matched VUC and in-person ED cohorts and included VUC infrastructure costs in our cost calculations.

This study also has important limitations as patients who choose VUC may be different from patients who attend the ED in person. Since the decision to use VUC versus attending an ED in person is influenced by multiple factors, many of which are not captured in administrative data (eg, self-perceived symptom severity at the time of decision, time and costs to travel to the ED, and satisfaction with prior ED services), comparing users of VUC with those who attend in-person ED is methodologically challenging. To deal with this important limitation and the heterogeneity between VUC users and between VUC and in-person ED users, we split and matched our VUC users by discharge disposition (eg, VUC presenting to the ED within 72 hours of a VUC referral or discharged home) and used PSM methods to compare with in-person ED patients. Despite matching on many measured confounders, important potential confounders were not available to us and the risk for potential unmeasured confounding remains. For example, the pediatric population and the senior population may require assistance with travel to the ED, which may influence decisions to use VUC rather than attend the ED in person. Although we matched the presenting complaint leading to the encounter, we were not able to match on acuity scores for the cohort of VUC users not referred to the ED by their VUC provider. In the VUC pilot, nearly 80% of VUC users screened themselves using a symptom checklist [7]. As such, caution should be expressed when interpreting the results of the matched analyses for the VUC cohort not referred to ED. We also did not know if VUC users would have presented in person to the ED in the absence of the VUC pilot program. They could have gone, for example, to an outpatient urgent care or walk-in clinic. However, compared to other jurisdictions, urgent care clinics are limited in Ontario. For example, while there are approximately 140 hospitals in Ontario, there are less than 20 urgent care centers in Ontario [21]. While there are more walk-in clinics in Ontario than urgent care centers, walk-in clinics do not have ED physicians on staff, and therefore, walk-in clinics may not be an appropriate comparator. In addition, as opposed to hospitalization or ED visits, reporting outpatient clinic data is not mandatory in Ontario [22]. As such, we did not have access to these data to create a control group of individuals attending urgent care outpatient clinics.

While we had access to several administrative databases, privately funded drug expenditures were not included in our analyses nor did we include patients' time (including economic impact and opportunity costs) or out-of-pocket expenditures (eg, parking and travel), which may have been avoided by the availability of VUC. As this was a pilot project, the population who would use VUC should it become universally available across Ontario may be different. However, we believe the mix of the VUC sites participating in the pilot program is broadly representative of the breadth of ED settings in Ontario. There may be differences in 30-day health care use and costs between

XSL•FO
RenderX

models of care (eg, nurse triage vs self-screening), populations (pediatrics vs adults), or settings (community vs academic sites; rural vs urban), which we did not examine. Due to data limitations, we were not able to isolate the physician costs associated with the VUC or in-person ED encounter. Finally, the MoH operational funds used in our calculations may represent an underestimation of the true costs of setting up VUC programs as it did not include any in-kind support provided by the sites. While the operational costs to set up and maintain VUC models may decrease with efficiencies in implementation, the cost of providing VUC is likely to change if different models of care are implemented (eg, nurse practitioners or physician assistants as first VUC contact before escalating to the ED physician if needed) or if the VUC technology becomes more expensive. Generalizations on health system costs within this study are limited to acute and subacute patient outcomes treated within the 30-day window of this study, and results may not be generalizable to other time periods or health care settings. Finally, due to the lack of data, our analyses were limited to costs and did not include economic outcomes (eg, quality-adjusted life-years). Therefore, the cost-effectiveness of the VUC program is unknown and this is left for future research.

Conclusions

This costing analysis supports the use of VUC for low acuity presenting complaints as most patient concerns were addressed without an in-person referral to the ED. However, additional research should evaluate alternative models of VUC to inform future resource allocation and policy decisions.

Acknowledgments

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MoH) and the Ministry of Long-Term Care. This document used data adapted from the Statistics Canada Postal Code^{OM} Conversion File, which is based on data licensed from Canada Post Corporation, and data adapted from the Ontario MoH Postal Code Conversion File, which contains data copied under license from Canada Post Corporation and Statistics Canada. Parts of this material are based on data and information compiled and provided by the Canadian Institute for Health Information and the Ontario MoH. The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred. We thank the Toronto Community Health Profiles Partnership for providing access to the Ontario Marginalization Index. The authors wish to thank the members of the Provincial Virtual Urgent Care Evaluation Committee.

Data Availability

The data sets analyzed during this study are not publicly available due to legal data sharing agreements between ICES and data providers. However, confidential access may be granted to those who meet prespecified ICES criteria by contacting the corresponding author.

Authors' Contributions

JET, SLM, SM, JMP, LP, EB, and JNH contributed to the conception and design of the work. JMP, LP, and EB contributed to the acquisition and analysis of the data. All of the authors contributed to the interpretation of data. JET drafted the manuscript. All the authors revised it critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

JM is a paid employee of Ontario Health. HO is a paid advisor to the Ministry of Health and in that capacity has provided executive sponsorship for the virtual urgent care (VUC) evaluation. SM reports research funding from the Juravinski Research Institute and PSI Foundation, as well as travel support from Mitacs. Ontario Health provided operational funding to each of the VUC pilot sites. There was no role of the funding body in the design, analysis, interpretation, writing, or approval of the manuscript. The views expressed in this manuscript are the views of the authors and do not necessarily reflect the views of the province.

Multimedia Appendix 1
ICES databases.
[DOCX File , 14 KB - jmir_v26i1e50483_app1.docx]

Multimedia Appendix 2
Cohort selection of virtual urgent care (VUC) patients referred to the emergency department after their VUC encounter and before 1:1 matching.
[DOCX File , 16 KB - jmir_v26i1e50483_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Cohort selection of virtual urgent care (VUC) patients not referred to an in-person emergency department after their VUC encounter and before 1:1 matching.
[DOCX File , 16 KB - jmir_v26i1e50483_app3.docx]

References

1. Census Profile, 2021 Census of Population. Statistics Canada. 2023. URL: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E [accessed 2024-01-10]
2. O'Gorman LD, Hogenbirk JC, Warry W. Clinical telemedicine utilization in Ontario over the Ontario telemedicine network. Telemed J E Health 2016 Jun;22(6):473-479 [FREE Full text] [doi: 10.1089/tmj.2015.0166] [Medline: 26544163]
3. Bhatia RS, Chu C, Pang A, Tadrous M, Stamenova V, Cram P. Virtual care use before and during the COVID-19 pandemic: a repeated cross-sectional study. CMAJ Open 2021;9(1):E107-E114 [FREE Full text] [doi: 10.9778/cmajo.20200311] [Medline: 33597307]
4. Glazier RH, Green ME, Wu FC, Frymire E, Kopp A, Kiran T. Shifts in office and virtual primary care during the early COVID-19 pandemic in Ontario, Canada. CMAJ 2021;193(6):E200-E210 [FREE Full text] [doi: 10.1503/cmaj.202303] [Medline: 33558406]
5. Virtual urgent care funding opportunities. Ontario Health. URL: https://www.ontariohealth.ca/sites/ontariohealth/files/2021-08/FundingCriteria-VirtualUrgentCare.pdf [accessed 2022-06-24]
6. Hall JN, Ackery AD, Dainty KN, Gill PS, Lim R, Masood S, et al. Designs, facilitators, barriers, and lessons learned during the implementation of emergency department led virtual urgent care programs in Ontario, Canada. Front Digit Health 2022;4:946734 [FREE Full text] [doi: 10.3389/fdgth.2022.946734] [Medline: 36093385]
7. McLeod SL, Mondoux S, Hall JN, Dainty K, McCarron J, Tarride J, et al. Demographic characteristics, outcomes and experience of patients using virtual urgent care services from 14 emergency department led sites in Ontario. CJEM 2023;25(1):65-73 [FREE Full text] [doi: 10.1007/s43678-022-00407-9] [Medline: 36380242]
8. McLeod SL, Tarride J, Mondoux S, Paterson JM, Plumptre L, Borgundvaag E, et al. Health care utilization and outcomes of patients seen by virtual urgent care versus in-person emergency department care. CMAJ 2023;195(43):E1463-E1474 [FREE Full text] [doi: 10.1503/cmaj.230492] [Medline: 37931947]
9. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 2007;45(4):247-251. [doi: 10.1016/j.ypmed.2007.08.012] [Medline: 17950122]
10. Grafstein E, Bullard MJ, Warren D, Unger B, CTAS National Working Group. Revision of the Canadian Emergency Department Information System (CEDIS) presenting complaint list version 1.1. CJEM 2008;10(2):151-173. [doi: 10.1017/s1481803500009878] [Medline: 18371253]
11. Matheson F, Dunn J, Smith K, Moineddin R, Glazier R. Development of the Canadian Marginalization Index: a new tool for the study of inequality. Can J Public Health 2012;103(8 Suppl 2):S12-S16 [FREE Full text] [doi: 10.1007/BF03403823] [Medline: 23618065]
12. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28(25):3083-3107 [FREE Full text] [doi: 10.1002/sim.3697] [Medline: 19757444]
13. Wodchis WP, Bushmeneva K, Nikitovic M, McKillop I. Guidelines on Person-Level Costing Using Administrative Databases in Ontario. Working Paper Series Volume 1. Toronto: Health System Performance Research Network; 2013.
14. Nord G, Rising KL, Band RA, Carr BG, Hollander JE. On-demand synchronous audio video telemedicine visits are cost effective. Am J Emerg Med 2019;37(5):890-894. [doi: 10.1016/j.ajem.2018.08.017] [Medline: 30100333]
15. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-to-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130] [Medline: 28264950]
16. Lovell T, Albritton J, Dalto J, Ledward C, Daines W. Virtual vs traditional care settings for low-acuity urgent conditions: an economic analysis of cost and utilization using claims data. J Telemed Telecare 2021;27(1):59-65. [doi: 10.1177/1357633X19861232] [Medline: 31342852]
17. Gordon AS, Adamson WC, DeVries AR. Virtual visits for acute, nonurgent care: a claims analysis of episode-level utilization. J Med Internet Res 2017;19(2):e35 [FREE Full text] [doi: 10.2196/jmir.6783] [Medline: 28213342]
18. Khairat S, Lin X, Liu S, Man Z, Zaman T, Edson B, et al. Evaluation of patient experience during virtual and in-person urgent care visits: time and cost analysis. J Patient Exp 2021;8:2374373520981487 [FREE Full text] [doi: 10.1177/2374373520981487] [Medline: 34189260]
19. NACRS emergency department visits and lengths of stay [release summary]. CIHI. URL: https://www.cihi.ca/en/nacrs-emergency-department-visits-and-lengths-of-stay [accessed 2023-03-25]
20. Wallace K. ‘Staggering’ number of Ontario emergency department closures revealed by Star analysis. The Toronto Star. 2023. URL: https://tinyurl.com/38d5mudm [accessed 2023-03-27]
21. Urgent care centres bring ER relief. Hospital News. URL: https://hospitalnews.com/urgent-care-centres-bring-er-relief/ [accessed 2024-05-29]

XSL•FO
RenderX

22. Data Quality Documentation, National Ambulatory Care Reporting System—Current-Year Information, 2020–2021. Ottawa, ON: Canadian Institute for Health Information; 2021.

Abbreviations

ED: emergency department
MoH: Ministry of Health
OTN: Ontario Telemedicine Network
PSM: propensity score matching
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VUC: virtual urgent care

Edited by T de Azevedo Cardoso; submitted 02.07.23; peer-reviewed by L Nijmeh, S Mehta, S Cressman, J Sharp; comments to author 12.12.23; revised version received 31.01.24; accepted 10.05.24; published 15.07.24.

Please cite as:
Tarride JE, Hall JN, Mondoux S, Dainty KN, McCarron J, Paterson JM, Plumptre L, Borgundvaag E, Ovens H, McLeod SL
Cost Evaluation of the Ontario Virtual Urgent Care Pilot Program: Population-Based, Matched Cohort Study
J Med Internet Res 2024;26:e50483
URL: https://www.jmir.org/2024/1/e50483
doi:10.2196/50483
PMID:39008348

©Jean-Eric Tarride, Justin N Hall, Shawn Mondoux, Katie N Dainty, Joy McCarron, J Michael Paterson, Lesley Plumptre, Emily Borgundvaag, Howard Ovens, Shelley L McLeod. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Policy Guidance for Direct-to-Consumer Genetic Testing Services: Framework Development Study

Suzanne Maria Onstwedder[1,2,3], MSc; Marleen Elizabeth Jansen[1,2,3], PhD; Martina Cornelia Cornel[2,3], MD, PhD; Tessel Rigter[1,2,3], PhD

[1]Department of Public Health Genomics and Screening, Centre for Health Protection, Dutch National Institute for Public Health and the Environment, Bilthoven, Netherlands

[2]Section Community Genetics, Department of Human Genetics, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[3]Personalized Medicine Programme, Amsterdam Public Health Research Institute, Amsterdam, Netherlands

Corresponding Author:
Suzanne Maria Onstwedder, MSc
Department of Public Health Genomics and Screening
Centre for Health Protection
Dutch National Institute for Public Health and the Environment
Antonie van Leeuwenhoeklaan 9
Bilthoven, 3721 MA
Netherlands
Phone: 31 088 689 8989
Email: suzanne.onstwedder@rivm.nl

Abstract

Background: The online offer of commercial genetic tests, also called direct-to-consumer genetic tests (DTC-GTs), enables citizens to gain insight into their health and disease risk based on their genetic profiles. DTC-GT offers often consist of a combination of services or aspects, including advertisements, information, DNA analysis, and medical or lifestyle advice. The risks and benefits of DTC-GT services have been debated and studied extensively, but instruments that assess DTC-GT services and aid policy are lacking. This leads to uncertainty among policy makers, law enforcers, and regulators on how to ensure and balance both public safety and autonomy and about the responsibilities these 3 parties have toward the public.

Objective: This study aimed to develop a framework that outlines aspects of DTC-GTs that lead to policy issues and to help provide policy guidance regarding DTC-GT services.

Methods: We performed 3 steps: (1) an integrative literature review to identify risks and benefits of DTC-GT services for consumers and society in Embase and Medline (January 2014-June 2022), (2) structuring benefits and risks in different steps of the consumer journey, and (3) development of a checklist for policy guidance.

Results: Potential risks and benefits of DTC-GT services were mapped from 134 papers and structured into 6 phases. In summary, these phases were called the consumer journey: (1) exposure, (2) pretest information, (3) DNA analysis, (4) data management, (5) posttest information, and (6) individual and societal impact. The checklist for evaluation of DTC-GT services consisted of 8 themes, covering 38 items that may raise policy issues in DTC-GT services. The themes included the following aspects: general service content, validity and quality assurance, potential data and privacy risks, scientific evidence and robustness, and quality of the provided information.

Conclusions: Both the consumer journey and the checklist break the DTC-GT offer down into key aspects that may impact and compromise individual and public health, safety, and autonomy. This framework helps policy makers, regulators, and law enforcers develop methods to interpret, assess, and act in the DTC-GT service market.

(J Med Internet Res 2024;26:e47389) doi:10.2196/47389

KEYWORDS

genetic testing; direct-to-consumer testing; health policy; genetic privacy; online market; informed consent; public health genomics; policy decision; mHealth; mobile health; privacy

XSL•FO
RenderX

Introduction

Background

Since the development and marketing of commercial genetic tests, citizens have been able to buy DNA tests from the internet without any direct involvement of a medical professional [1]. Using these DNA tests, consumers—here viewed as individuals who have purchased and taken a commercial genetic test or potential buyers who are exposed to these offers—are now able to gain insight into their health characteristics and risks of disease. These digital health consumer products are an example of those technological innovations that the private sector has picked up from research. Their availability fits with the trends of the commercial market, which reflect the consumerization of mobile health (mHealth) worldwide in numerous areas of health care, fitness, and wellness. The rise of the consumer DNA test market has mainly been fueled by a drop in the cost of collecting and analyzing genomic data [2]. Meanwhile, the public is increasingly showing an interest in genomics [3].

These DNA tests are also called direct-to-consumer genetic tests (DTC-GTs) and, at a minimum, entail a service to collect a saliva sample (sending a kit, giving instructions on how to collect and return the sample), conduct a genetic test of the sample, and provide a report of outcomes of the test. Because the DTC-GT process entails multiple aspects, in this study, we described them as "DTC-GT services." There are many types of DTC-GT services, including tests for ancestry and paternity. This paper focused on DTC-GT services aimed at health. Additional digital services are often provided by means of personal health information and lifestyle advice in apps or reports, online consultation with genetic experts, and privacy-ensured storage of DNA data [4]. In this way, DTC-GT services combine aspects of many mHealth products that assess and monitor health. After purchase, consumers are often able to download their raw genetic data and explore the possibilities of multiple online third-party analyzers that will run an additional analysis on their data [5]. In this way, these third-party analyzers further connect DTC-GT users to the scientific literature and promise to make novel research insights accessible to users [6]. This illustrates that DTC-GTs are part of a complex international online market [5,6].

DTC-GT companies generally state that their services can inform health decisions and minimize disease risk by offering insight into personal health information and providing lifestyle advice [7,8]. Although these statements sound promising in relation to health improvement, their validity and utility in this field is under debate [9]. There is also further uncertainty about issues such as privacy, psychological impact, and the effects on health care systems [10].

Consequently, it can be argued that the risks and benefits of DTC-GTs are valid reasons for developing policies on this topic. However, ensuring that these issues are addressed at the right level is challenging. The fact that these tests are offered internationally via the web complicates any effort to address their potential risks or benefits. Some of these issues are regulated at European or international levels, while others are regulated at national or regional levels [11]. Addressing these issues raises questions regarding the roles and responsibilities of policy makers, regulators, and law enforcers, both within and outside the health care system. Insight is needed into the potential risks and benefits of DTC-GT services that would lead to the development of policies. A structured overview of potential risks and benefits could guide policy on DTC-GT services. Such an overview would enable policy makers, regulators, and law enforcers to identify their responsibilities. For example, policy makers at the national government level, professional associations, grant providers, and patient and consumer organizations would then be able to define the outstanding questions for research and needs for publicly available information to ultimately enable responsible availability and use of DTC-GTs.

Risks and Benefits of DTC-GT Services

Genomics holds the potential to improve diagnosis, personalize treatment, and determine groups at risk of developing a disease. Researchers are continuously extending their knowledge of genomics and its potential impact on individual health, personalized medicine, prevention [12-14], public health, and precision public health [15]. It is argued that gaining insight into personal health risks, which are impacted by genetic risks, lifestyle factors, and environmental effects, could ultimately help empower citizens to gain more personal control over their health. Having said that, it seems promising that consumers can autonomously impact their health through DTC-GT services. Yet, genetic and ethics experts have expressed their concerns about DTC-GTs [9,10]. Experts argue that the information provided can be misleading and that results that indicate an increased or reduced disease risk can easily be misinterpreted, potentially leading to misguided health decisions and medical risks [16,17]. This leads to the question of whether DTC-GT services are an effective and safe way for the public to pursue health improvement.

However, it could be argued that DTC-GT services should not be judged solely as a medical product and that arguments about personal utility and autonomy should also be part of the equation [18-21]. One could make the case that a consumer product may serve a purpose other than a medical one, an educational one, for instance. Therefore, the risks and benefits of DTC-GT services as a consumer product could be weighed differently. When purchasing a test, aspects of a product, such as ease of use and accessibility, may be a stronger factor in consumer decision-making, than the test being 100% accurate.

Policy and Regulatory Issues

Due to a lack of consensus about the net benefit or risk of DTC-GTs, uncertainty has arisen among policy makers, regulators, and law enforcers about their roles and responsibilities in the DTC-GT market. DTC-GT services fall into a gap of regulatory structure in both Europe and the United States [11,22]. Currently, no specific DTC-GT legislative instruments are being implemented in most of the member states of Europe or in the United States. A variety of laws, all on different levels of legislation (state or country specific, European, and international laws), are effective in the DTC-GT market. These laws cover aspects such as medical device safety, laboratory quality assurance, medical supervision, genetic

XSL•FO
RenderX

counseling, and informed consent [11,22-24]. Yet, most of these laws only apply to genetic testing within the conventional health care system. Applying these laws to a commercial product may not be sufficient to protect citizens from potentially harmful tests.

Further regulatory difficulties are caused by the international and dynamic character of DTC-GT services. A great variety of tests for health and other purposes are available online. Examples include cancer risk tests, nutrigenetic testing, and ancestry testing, combined with insights into health characteristics and other traits [4,22]. Moreover, the additional services that are offered, such as apps or reports containing personal health and lifestyle advice, add further variety to their offers [4]. Consumers can purchase any of these DTC-GT services on the web from many places in the world. The global and dynamic nature of the market hinders regulatory bodies in implementing effective policies and in law enforcement.

Ambiguity Calls for Policy Guidance

The ambiguity in the legislative framework applicable to DTC-GTs has resulted in challenges for policy makers, regulators, and law enforcers in protecting citizens from potentially harmful services. Meanwhile, the impact of DTC-GTs on citizens and their health continues to be a topic of debate. To balance safety and individual autonomy, all aspects of DTC-GT services should be considered when regulatory decisions are required [11]. For this reason, it is important that the current body of knowledge on the risks and benefits of all key aspects of DTC-GTs is translated into policy guidance. To ensure effective policy making, researchers have found the following issues to be relevant in regulating the DTC-GT market: autonomy and welfarism, informed decision-making, privacy, clinical validity and utility, perspectives of the public and health care professionals, the role of regulatory organizations, legislation on genetic testing, and laws protecting against genetic discrimination [25]. All these issues are interlinked in DTC-GT services. The existing literature has focused on gaining insight into these issues and has started to collect empirical evidence on the impact of the market. Yet, instruments that assess those aspects of DTC-GT services that are likely to need regulation are lacking.

To ensure that policy guidance covers all aspects of DTC-GT services, it is important to include the risks and benefits of each distinct service element offered by commercial companies. To this end, in this study, we described those aspects of DTC-GT services that may give rise to policy issues. Subsequently, we synthesized a framework for policy makers, law enforcers, and regulators to develop methods to interpret, assess, and act in the DTC-GT service market.

The framework development process consisted of 3 steps:

- Performing an integrative literature review to identify the risks and benefits of DTC-GT services to consumers and society
- Structuring the benefits and risks of each step of the consumer journey
- Developing a checklist for policy guidance

Methods

Study Design

This study focused on DTC-GT services that advertise that they deliver insight into disease risks and lifestyles that promise to impact customer health through behavior change. This includes anticipated lifestyle changes, changes to medication intake, and recommendations to seek medical help from professionals. We viewed consumers of the DTC-GT market as both individuals who have purchased DTC-GT services and as potential consumers who are exposed to offers of these services. These services are offered via webshops, including primary DTC-GT service companies and secondary selling points, such as Amazon, and at high street stores.

Study Layout

To distill key aspects of DTC-GTs, those services that lead to policy issues, potential risks, and benefits for consumers and society, including and beyond medical implications, were systematically mapped by means of an integrative literature review (step 1). To do this, the complete DTC-GT service pathway was studied, from exposure to the offer of DTC-GTs to the ultimate individual and societal impacts of DTC-GT services.

In step 2, the risks and benefits were structured into phases, following each step the consumer takes along the DTC-GT service pathway. Each of these phases triggered distinct policy issues, including issues regarding informed decision-making, privacy, clinical validity and utility, and legislation for genetic testing [25].

Step 3 involved developing a checklist that can ultimately offer guidance on each distinct DTC-GT policy issue. To ensure adequate policy guidance, checklist items that help evaluate all phases of the DTC-GT service pathway were included. Every checklist item reflects a potential risk or benefit found in the literature review.

Study Identification and Eligibility Check

The risks and opportunities associated with DTC-GTs were determined by means of a literature search that focused on health, including lifestyle and disease risk. The literature search was performed systematically using Embase.com, which combines the databases of Embase and Medline. The search strategy included the following primary search terms with corresponding synonyms: “direct-to-consumer genetic testing,” “impact,” “risks,” “opportunities,” “clinical study,” “evidence based practice,” “accuracy,” and “ethical, legal, and social issues” (see Multimedia Appendix 1). In addition, 5 experts in the field of genomics, public health, and biomedical sciences with knowledge of DTC-GTs were asked to provide key papers discussing the risks and benefits of DTC-GTs.

Publications were included if they matched the publication date restriction: January 2014-June 2022. Papers from March 2020 to June 2022 were included as part of a study update (see the *Literature Update* section). We excluded studies dating from before 2014, as it is argued that around 2015, the DTC-GT market transformed into a second generation in response to a

XSL•FO
RenderX

shutdown of the Food and Drug Administration (FDA) of DTC-GT companies [26]. Language requirements were English or Dutch. The included study types were meta-analyses, reviews and systematic reviews, case-control studies, case reports and series, cohort studies, randomized controlled trials, interview and Delphi studies, expert opinion papers, and commentaries. Conference posters and abstracts were excluded.

Studies were included if they described the following: population (P), intervention (I), comparison (C), and outcome (O; PICO; see Table 1) [27]. Citizens (P), excluding animals and microorganisms, who took a DTC-GT for health or lifestyle purposes (I1) or considered taking a DTC-GT for health or lifestyle purposes (I2) were included. In addition, professionals (P2) with expertise in DTC-GTs (I3) were included, among others: health care professionals, ethics experts, and researchers. Studies were restricted to DTC-GTs that purport to improve the consumers' own health by analyzing their DNA for health and lifestyle purposes. Genealogy and ethnicity testing were excluded. DTC-GTs for prenatal and carrier-screening purposes were also excluded, since the primary aims and subsequent actions of these tests go beyond improving personal health.

Insights were included in this study if they indicated an increased risk or benefit to those citizens who took DTC-GTs when compared with those citizens who did not take a DTC-GT (C1) or those citizens who underwent genetic testing within the conventional health care system. Lastly, the studies must have reported insight into the impact of DTC-GTs along the complete DTC-GT service pathway, including exposure to the offer of DTC-GTs, insight into the validity and utility of DTC-GTs, and the individual and societal impact of a DTC-GT (O). Information about the legal framework of DTC-GTs was not included, since this topic has been reviewed by others and is country specific.

The following steps were taken during the review process: duplicate removal, title and abstract screening using the abovementioned criteria, and full-text screening using the same criteria. The titles and abstracts were screened by 1 researcher (SMO) using the abovementioned criteria. The results were discussed with a second researcher (TLA). Next, 2 researchers (SMO and IKK) divided and screened the remaining full-text papers. When in doubt about inclusion, the researchers discussed this with an additional researcher (MEJ). Those key aspects that determine policy issues were analyzed by 1 researcher (SMO) and later discussed with 3 researchers (MCC, MEJ, and TR). No generative artificial intelligence was used in any portion of the manuscript writing.

Table 1. Description of PICO[a] in this study.

PICO [28]	Included	Excluded
Population	• Citizens • Experts	• Animals • Microorganisms
Intervention	• Citizens who took a DTC-GT[b] for health or lifestyle purposes • Citizens who considered taking a DTC-GT for health and lifestyle purposes • Experts with expertise in DTC-GT services	• DTC-GTs for genealogy and ethnicity • DTC-GTs for prenatal and carrier screening
Comparison	• Citizens who did not take a DTC-GT • Citizens who underwent genetic testing within the conventional health care system	• N/A[c]
Outcome	• The individual and societal impact of DTC-GT services throughout the complete process of the DTC-GT service pathway	• Legislation and regulations

[a]PICO: population, intervention, comparison, and outcome.

[b]DTC-GT: direct-to-consumer genetic test.

[c]N/A: not applicable.

Data Collection and Analysis

The risks and benefits were structured into phases, following the steps that consumers take along the complete DTC-GT service pathway. Studies were analyzed, and the data thus derived were grouped (SMO and IKK; step 2) by means of an iterative process based on the findings of the literature review itself and the model of Mosdøl et al [28]. This model has previously been used to research the effects of direct-to-consumer advertisements, specifically for prescription medicines [29]. The data were divided into the following themes: (1) consumer demographics; (2) test quality, validity, and utility; (3) company features; (4) diseases tested for and health aspects; (5) consumer effects; (6) impact on health care; (7) informational content and advertising; (8) ethics; and (9) general and any remaining information about DTC-GTs.

Information about these themes was gathered to determine those key aspects that lead to policy issues. The key aspects were defined when new themes no longer emerged, that is, when data saturation was achieved. These findings were shared with 5 experts on DTC-GTs, genetic tests, or genomics. Their input was used to further interpret the risks and benefits of DTC-GTs. Next, a framework was composed that structured the key aspects of DTC-GTs that affect potential risks and benefits and engender

XSL•FO
RenderX

policy issues (SMO). This was then discussed with 3 researchers (MCC, MEJ, and TR).

Literature Update

A literature update was performed between March 2020 and June 2022. Novel studies were identified and selected using the same strategy, as described before, differing by only 1 step. During the literature update, we first assessed whether the studies researched themes or topics that were already included in the primary search. If this was the case, they were excluded from the sample. After that, full-text documents were retrieved. The data from the initial literature review (January 2014-March 2020) were supplemented with the potential risks and benefits gleaned from this literature update. Potential risks and benefits that had already been reported in the initial search were viewed as saturated findings, as they did not lead to new policy issues. Those data were therefore not further collected.

Checklist Development

Ultimately, a checklist was designed to offer guidance in DTC-GT policy issues (SMO and TR) and to evaluate those key aspects that determine product safety, quality, and utility.

Five previously developed and published checklists and frameworks [7,30-33] formed the basis for this checklist. These checklists and frameworks separately do not cover the whole DTC-GT service pathway. Therefore, checklist items were selected from these publications to encompass the entire DTC-GT service pathway. The content of all published checklists and frameworks was combined and listed in 1 document. We evaluated each item on this list to see whether it appraised a risk or a benefit identified by our literature review. That way, checklist items were included to match the risks and benefits found in the literature review. Duplicate checklist items were combined into 1 item. When necessary, items were adjusted or supplemented to ensure maximum fit for evaluation of DTC-GT services. The checklist items were further subdivided into themes to improve ease of use.

Questions that evaluate the quality of DTC-GT service elements and informational content on DTC-GT websites were formulated based on the checklist items. For ease of application, the questionnaire was piloted by 3 researchers who analyzed 3 DTC-GT company websites. These 3 companies were selected to reflect the broad variety of services on the market, including 1 international leading DTC-GT company offering a variety of genetic health tests, 1 big international company with a Dutch website offering a variety of genetic health tests, and 1 small and relatively new national Dutch company offering a DTC-GT focused on sport performance and health.

Results

Integrative Literature Review

The literature search resulted in a primary yield of 323 papers (Figure 1). In addition, 15 key papers were received from the 5 experts consulted; however, most of these papers (n=13, 86.7%) were duplicates of the literature search yield. After removal of duplicates (n=24), a total of 314 (97.2%) papers were included for screening. Title and abstract screening and the inability to retrieve full-text papers led to the exclusion of 106 (33.8%) papers. Ultimately, data were collected from 127 (40.4%) papers. The other 81 (25.8%) papers were excluded as they were incorrect document types (Congress posters and abstracts or study protocols), were not published in English or Dutch, or lacked relevance, and this had not been picked up during the first screening. In addition, the literature update yielded 7 papers with novel findings that had not been reported in the initial literature review (see Multimedia Appendix 2 for a list of included papers) [34-40].

XSL•FO
RenderX

Figure 1. PRISMA diagram describing the reviewing process of the literature search to identify relevant publications between January 2014 and March 2020 and between March 2020 and June 2022. *Papers from the literature update performed between March 2020 and June 2022. †This step was performed after checking whether the studies included in the literature update introduced themes that were not included in the primary search. All 7 papers that yielded novel findings could be retrieved. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

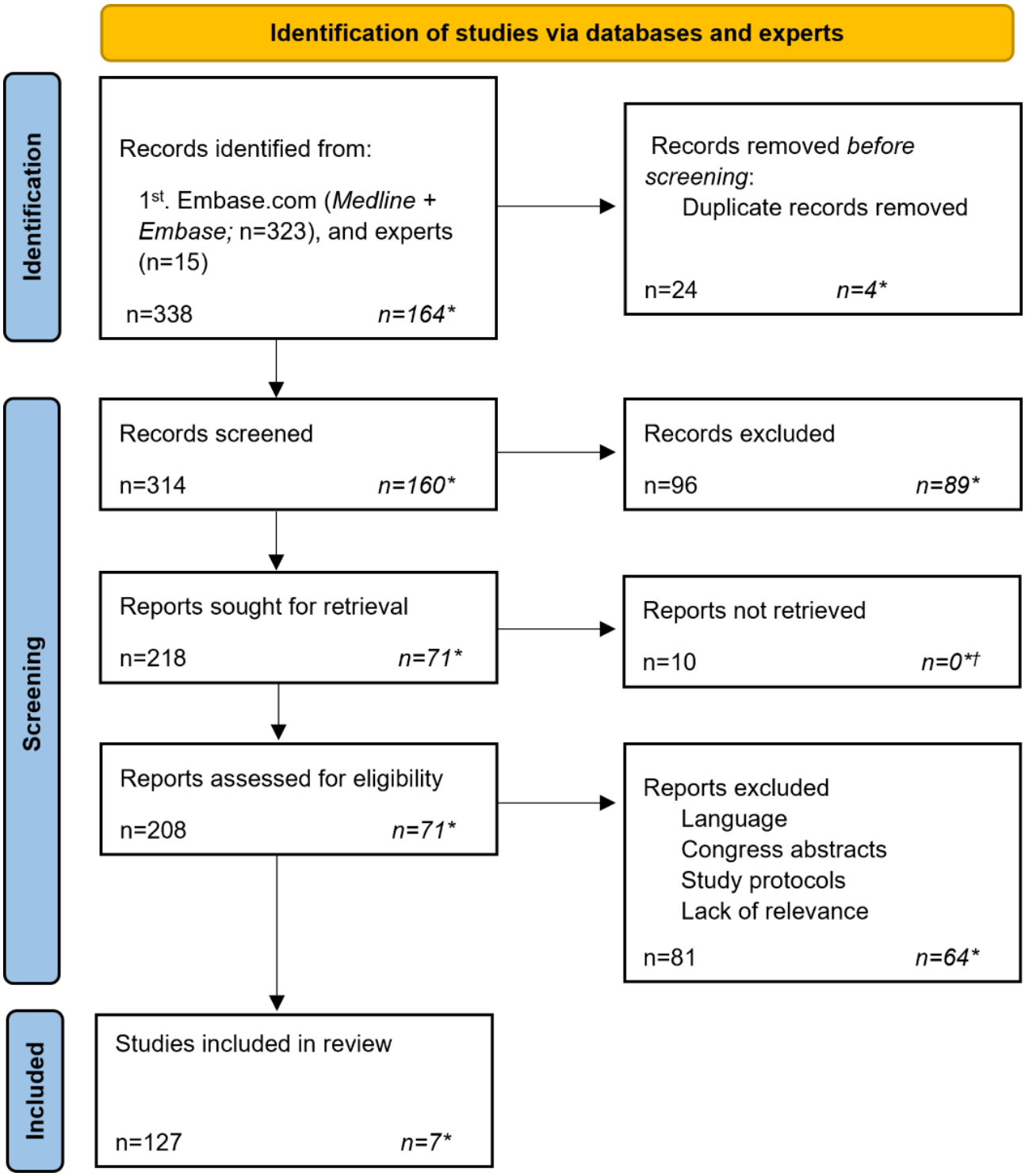

General Findings

DTC-GT services have been researched and discussed in both the descriptive literature, such as expert opinions, and empirical studies, such as questionnaires and interview studies with consumers. The taking of DTC-GTs in case-control and cohort studies was rare. Research mainly focused on various aspects of DTC-GTs, including but not limited to quality and validity, attitudes and perceptions of the public and consumers, the content of the product on offer, experiences of health professionals with DTC-GTs, the utility of the tests, the regulatory landscape surrounding DTC-GTs, and the ethical appraisal of both the advantages and disadvantages of DTC-GTs.

The individual and societal impact of DTC-GT services, caused by their potential risks and benefits, seemed wide ranging. Overall, potential risks were reported more frequently in the literature than potential benefits. The risks and benefits reported

XSL•FO
RenderX

included both medical risks, such as inadequately informed health decisions, and nonmedical risks, such as invasion of privacy or obtaining information about as yet unknown family relations. These risks and benefits appeared to be instigated by distinct aspects of DTC-GT services, such as information provision, DNA analysis, and reporting of results.

The Consumer Journey

Both risks and benefits were structured into phases, following the steps that consumers take when taking DTC-GT services. This led to 6 different phases, collectively called the *consumer journey.* These phases covered the complete process of taking DTC-GT services, from the first exposure to an offer of a DTC-GT service up to the ultimate individual and societal impact of DTC-GT services (depicted in Figure 2). Distinguishing and analyzing the phases of the consumer journey helped obtain insight into the consumer audience, distinct service elements that DTC-GT companies offer their consumers, and the individual and societal impacts of these services.

Each of these service elements may have distinct risks and benefits (see Figure 3). For example, in phase 1, consumers with common motivations, such as curiosity or the desire to improve health, could benefit from easy access to genetic tests, which may promote autonomy. Interest in DTC-GT services appeared to be more prominent among individuals of a higher socioeconomic status (SES) and also among adoptees; this may negatively impact the health equity and privacy of the biological parents of the latter group.

Phase 2 concerns access to information about health and genetics gathered before purchase of a DTC-GT. This could facilitate health literacy; however, research also points out that information about the potential impact of a DTC-GT, for example, may be unbalanced or misleading. This then impairs the process of informed decision-making among consumers, thereby limiting their autonomy and personal utility.

One benefit of phase 3 DNA analysis may be that private companies are investing in DNA research and innovation, which could boost DNA test innovation and reduce test costs. Yet, aspects such as limited analytical and clinical validity, as well as limited clinical utility and insight into quality assurance, are risks of phase 3 of DTC-GT service pathways. This may cause test results to be erroneous, thus providing consumers with incorrect information about their health risks.

How sensitive genetic data are handled by DTC-GT companies (phase 4), for example, privacy-ensured storage, or potentially shared with third parties can safeguard or impede the privacy of consumers. Genetic data are not unique in the way that they engender privacy issues, yet the data content may be extensive and sometimes concern delicate issues. This makes DNA data sensitive data, which is acknowledged in the General Data Protection Regulation [11].

Figure 2. Flow diagram describing the complete process of DTC-GT services, called the consumer journey. As depicted, the consumer journey is broken down into 6 phases. In this study, consumers of the DTC-GT market were both individuals who purchased a DTC-GT or potential consumers who were considering buying the test. DTC-GT: direct-to-consumer genetic test.

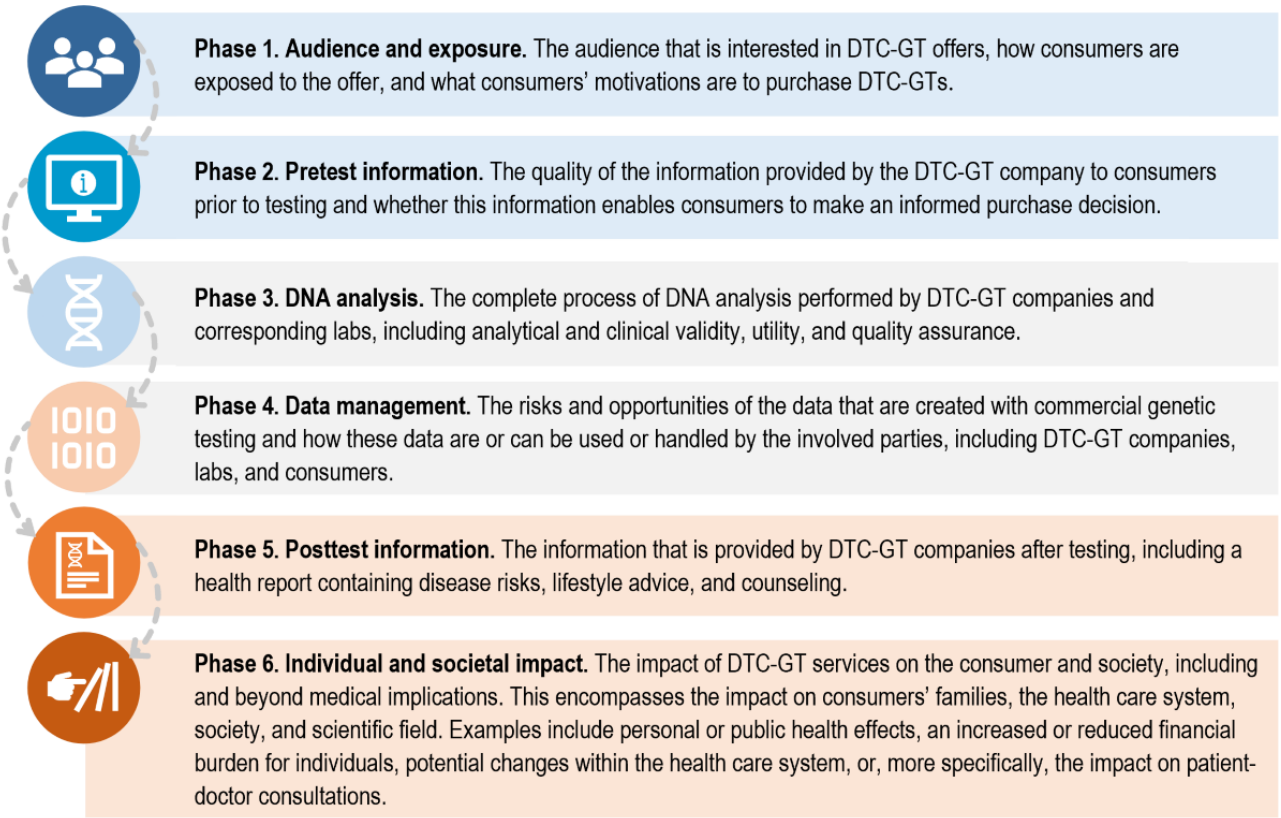

XSL•FO
RenderX

Figure 3. Summary of findings, describing the benefits and risks (darker boxes) of DTC-GTs and their potential impact (lighter boxes) throughout the different phases of the consumer journey. DTC-GT: direct-to-consumer genetic test; PGx: pharmacogenetic tests (DNA tests that predict the patients' response to specific pharmaceuticals, for example, adverse reactions or insufficient uptake of drugs); SES: socioeconomic status.

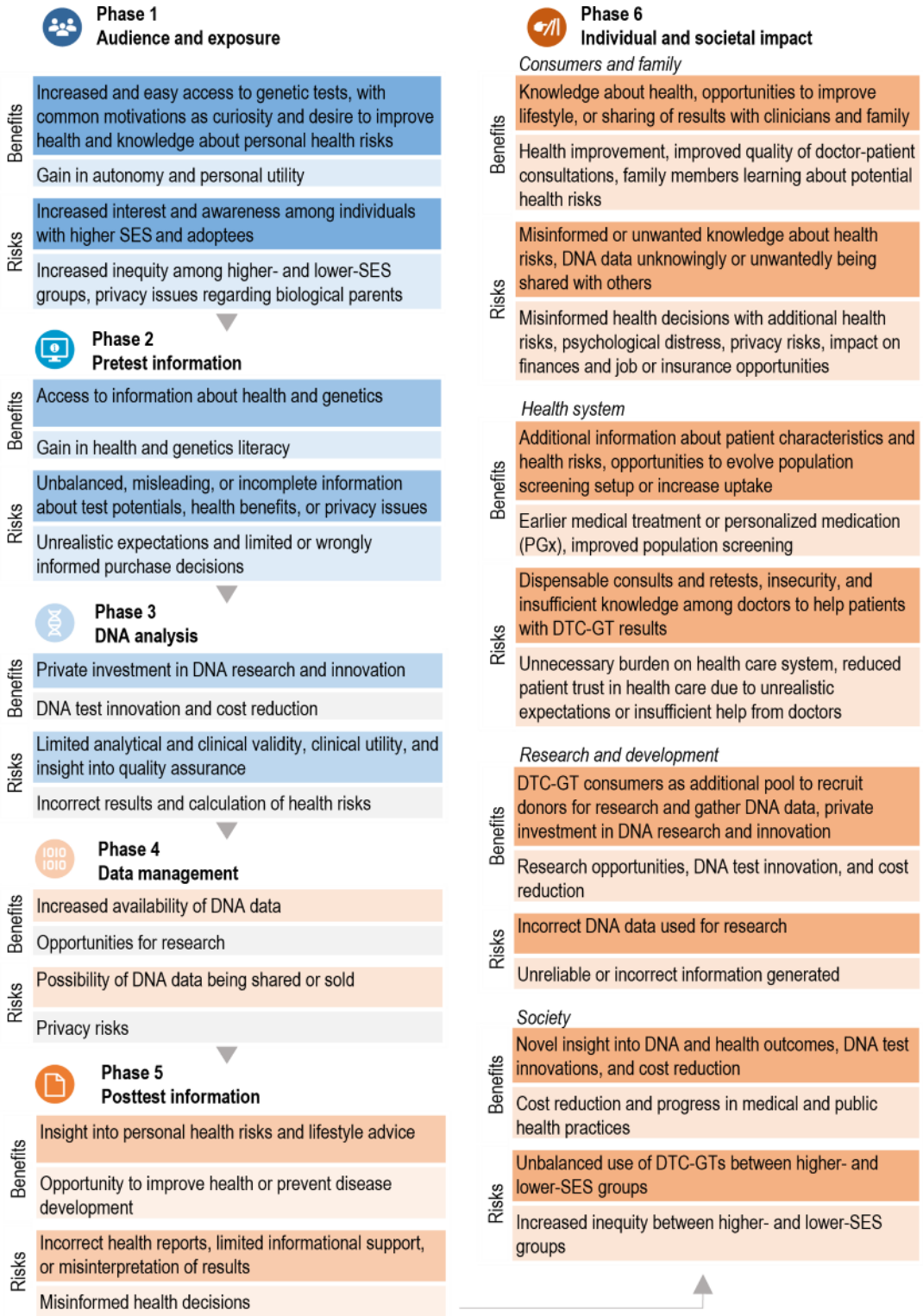

Although in phase 5, the information provided with the test results may help in making decisions about improving health, thereby improving autonomy, if this information is not clear or the health risks have been calculated based on erroneous results (phase 3), consumers may misinterpret their results and make misinformed health decisions.

Consequently, purchasing and taking DTC-GTs may impact individuals and society (phase 6). Consumers may be empowered to improve their health. However, health risks might arise if sensitive health decisions are based on results of low validity. Moreover, consumers may decide to purchase a test without first fully considering the potential risks and limitations of testing, and upon receiving DTC-GT reports, their

XSL•FO
RenderX

expectations may be unmet. Potential risks beyond medical effects include psychological distress as results might be worrying or have an impact on finances and job or insurance opportunities. DTC-GT services may also impair the autonomy of relatives, as DNA is shared within families. The breakthrough in the well-known Golden State Killer cold case, a murder case that was solved by matching the DNA at a crime scene to family members through DTC-GT databanks, is an example of how personal genetic data of relatives, obtained through DTC-GTs, can be used for purposes other than consumers might have opted for [41]. In terms of societal impact, consumers may decide to consult their health care professionals about their results or join large research projects, which may further impact the health care system and research (eg, increased burden, impact on patient trust in health care when medical and DTC-GT results are conflicting, and more DNA data available for research opportunities).

It is important to note that the phases of the consumer journey are often intertwined. For instance, the quality of the information provided after testing (eg, health reports, phase 5) may lead to uncertainties among consumers (individual impact, phase 6), causing them to seek help from professionals in the health care system (societal impact, phase 6). Furthermore, where data storage by DTC-GT companies is either not applicable or guaranteed and executed with adequate safety measurements (phase 4), the privacy of the consumer (individual impact, phase 6) and their family members is preserved (societal impact, phase 6).

Checklist for Evaluating DTC-GTs

After creating an overview of risks and benefits in each phase of the consumer journey, a checklist for policy guidance based on key aspects of the consumer journey was synthesized. Building on these phases ensured that all steps taken by consumers and all DTC-GT service elements were included in the checklist.

In total, 38 checklist items were derived from published evaluation frameworks and papers [7,30-33]. These items were subdivided into 8 themes (see Table 2):

- Theme I (general DTC-GT service features): 7 checklist items that cover the overall content of a DTC-GT service
- Theme II (DNA analysis and quality assurance): 4 checklist items about the validity and quality assurance of DTC-GT services
- Theme III (privacy and data management): 4 checklist items concerning how consumer privacy and data are managed by DTC-GT companies
- Theme IV (scientific evidence): 3 checklist items relating to the scientific evidence and robustness of DTC-GT services
- Theme V (information about results, interpretation, and consultation): 7 checklist items regarding the information provided about health outcomes assessed, genetics, and results of DTC-GTs
- Theme VI (information about potential consequences of taking a DTC-GT): 6 checklist items that cover information provided about the potential consequences of purchasing DTC-GT services for the consumer, including and beyond medical implications
- Theme VII (presentation of information): 4 checklist items about how information about the previous themes is presented, such as highlighting benefits over risks and vice versa or incorporating opinions of experts or public figures about the usefulness of DTC-GTs
- Theme VIII (informed decision-making): 3 checklist items that help evaluate whether consumers can make informed purchase decisions

Every checklist item reflects a potential risk or benefit found in the integrative literature review. For example, the cost of the test (theme I) applies to the potential risk of increased health disparity between individuals with a higher and a lower SES (consumer journey phase 1). Another example is the DNA test type (theme II), which correlates to the analytical and clinical validity of a DTC-GT and affects the validity of the DTC-GT results (consumer journey phase 3).

Subsequently, 3 researchers tested and administered the checklist in the form of a questionnaire to assess DTC-GT service webshops. This allowed the checklist to be further refined. Questions were formulated for all checklist items. Almost all questions were formulated as closed questions. Potential answers included “yes,” “no,” and “unclear” or “detailed,” “incomplete,” and “missing” when asked to evaluate the information provided about certain key aspects. Open answers were only permitted where the provided closed answer options were deficient (eg, “other, namely…”). Every answer included room for notes and expansion. For details on the layout of the checklist in the form of a questionnaire, see Multimedia Appendix 3.

The 3 researchers reviewed 2 well-known DTC-GT webshops that sell health-related DTC-GT services. On reviewing the webshops, the researchers often came up with different answers when evaluating the same company website. This could imply that the appraisal of webshop content is subjective. Further refinement of the questionnaire may help, although it may not fully compensate for individual differences in how the content of a webshop is perceived.

XSL•FO
RenderX

Table 2. Checklist with key items for DTC-GT[a] service evaluation, grouped by theme.

Theme	Key items
Theme I: general DTC-GT service features (consumer journey phases 1 and 2)	• Test costs • Assessed health features (eg, disease risks, sports performance, pharmacogenetics) • Type of generated report (eg, lifestyle advice, diagnosis) • Option to download raw DNA data after testing • Option to opt out from receiving certain results • Requirement of referral of health professional • Residency of the company and the labs
Theme II: DNA analysis and quality assurance (consumer journey phase 3)	• DNA test type (eg, sequencing, single-nucleotide polymorphism [SNP] array) • Quality assurance in labs that perform DTC-GTs (eg, International Standardization Organization (ISO)/Clinical Laboratory Improvement Amendments (CLIA) certification) • Information about how much and which genes or DNA variants will be analyzed • All relevant background information of consumers considered
Theme III: privacy and data management (consumer journey phase 4)	• Presence of a privacy policy on the website • Management of consumers' personal and genetic data by the company (eg, duration of storage, will the data be shared with or sold to third parties) • Management of consumers' samples after DNA analysis by the company (eg, duration of storage, what will happen to the samples upon bankruptcy) • Option to withhold storage, sharing, and selling of samples and data
Theme IV: scientific evidence and robustness (consumer journey phase 3)	• Methods based on reliable scientific evidence (DNA test type) • Robustness of scientific evidence underlying the chosen methods (eg, novel gene-disease interactions vs well-studied gene-disease interactions) • Collaboration with scientific partners (eg, academic research groups, prominent researchers from universities)
Theme V: information about results, interpretation, and consultation (consumer journey phase 5)	• Information about the assessed health feature (eg, disease burden, prevalence) • Information about how to interpret results • Information about the robustness of results • Possible actions in the case of a positive finding (eg, treatment options) • Option to consult a genetic professional at the company • Recommendation to consult a health care professional to discuss test results • Referral to companies or websites for additional analysis on raw DNA data (ie, third-party analyzers)
Theme VI: information about potential consequences of taking a DTC-GT (consumer journey phases 3, 5, and 6)	• DTC-GT results informing and impacting future health decisions and health behavior • DTC-GT results impacting insurance policies (eg, health insurance) • DTC-GT results impacting family (eg, revelation of family health risks or unknown family relations) • DTC-GT having other consequences beyond medical purposes (eg, feeling of anxiety or relief following test results) • DTC-GT results altering due to future technological innovations (eg, new health results) • DTC-GT service company investing in or performing own research
Theme VII: presentation of information (consumer journey phases 3 and 5)	• Usage of fear • Usage of public figures or medical experts to convince consumers into buying the test • Balanced information about advantages, usefulness, disadvantages, and risks of taking a DTC-GT • Clear distinction between health outcomes: (1) for which (proven effective) treatments or interventions are available, (2) whose outcome can be partially improved, (3) for which treatment is not yet available
Theme VIII: informed decision-making (consumer journey phases 3 and 5)	• Information that analyzing the DNA of others without consent is ethically irresponsible or punishable or both • Active confirmation of informed decision (eg, via a pop-up) • Generally clear, understandable, complete, and easy to retrieve information provided to make an informed decision (eg, privacy policy)

[a]DTC-GT: direct-to-consumer genetic test.

XSL•FO
RenderX

Discussion

Principal Findings

Following the results of an integrative literature review, this paper described the potential risks and benefits of DTC-GT services that may cause policy issues. The risks and benefits are wide ranging. They include medical risks and benefits, such as inadequately informed health decisions and improved knowledge about genetics and health, as well as nonmedical risks and benefits, such as invasion of privacy or obtaining information about as yet unknown family relations. These risks and benefits seem to be linked to distinct aspects of DTC-GT services, such as information provision, DNA analysis, and reporting of results. By structuring the risks and benefits into phases, following the consumers' steps when taking a DTC-GT service, 6 phases can be distinguished, which are collectively called the consumer journey. The consumer journey provides insight into the consumer audience, distinct service elements that DTC-GT companies provide for their consumers, and the individual and societal impacts of these services. Distinguishing different steps in the DTC-GT service process reveals multiple aspects of the offers of DTC-GT services that engender policy issues and may require policy guidance in the future. The checklist that was introduced in this study provides a backbone for policy guidance for DTC-GT services. To ensure coverage across all DTC-GT service elements, checklist items were included to evaluate all phases of the consumer journey. Each checklist item covers a risk or a benefit found in the literature review. By reviewing and extensively discussing the medical, ethical, social, and technological issues of DTC-GTs, the checklist translates current scientific knowledge into a concise and helpful instrument for policy makers, regulators, and law enforcers.

Implications

In its current form, the checklist summarizes crucial items that may affect individuals or society and that can trigger policy issues and may call for intervention. Combined with insight into the phases of the consumer journey, this could help different stakeholders comprehend the potential risks and benefits of the DTC-GT service market. Horton et al [42] structured the process of DTC-GT services similarly to the consumer journey. In their model, the risks in the DTC-GT process were described by presenting fictitious cases and what to discuss with patients in a clinical setting. Yet, their study focused mainly on the impact of technical and quality limitations after a consumer chooses to purchase a test, while we included the impact of all aspects of DTC-GT services, for example, information provision and informed purchase decisions, data management and privacy, and the impact on family members. The current combination of the checklist and the consumer journey could be helpful for a broader audience, in particular regulators, law enforcers, and policy makers within professional or patient organizations who are expected to safeguard their citizens or patients. To ensure a clear strategy on how to act in the DTC-GT market, it may be fruitful to design a process that incorporates the checklist and helps assign specific responsibility holders to specific policy issues. Policy makers at the national level may facilitate this by delegating tasks. If there is no national party to take on this directorial role, regulators, law enforcers, and policy makers may use the checklist to set their own agenda, preferably in collaboration with relevant stakeholders.

By using the checklist, policy makers can structurally review the DTC-GT services on offer to the residents of their country. This will help them determine societal and public health hazards and reveal aspects of DTC-GT services or highlight specific providers that call for intervention. Furthermore, it could aid in monitoring potentially useful genetic testing applications for uptake in health care and thereby strive to reduce the risk of inequity. Considering the roles and responsibilities of policy makers, this checklist may prove especially useful in monitoring the risks of the current market and protecting citizens from harm. For instance, public health agencies, patient and consumer organizations, and professional associations may choose to take responsibility for monitoring the tests on offer, as well as consumer issues, such as uninformed purchase decisions leading to unexpected results, or results of limited validity.

Furthermore, policy experts may review the legislation within a country with the checklist items to check for leads for law enforcement or gaps in regulations that call for novel policy. This is expected to be country specific, as the diverse service elements provided by DTC-GT companies, such as data protection and in vitro diagnostic medical devices, fall within distinct regulatory areas in Europe and the United States [11,22-24,43].

Several steps can still be taken to refine the checklist. Although all items on the checklist can lead to potential risks for individuals and society, the size of the risk or benefit and its subsequent impact may differ. Consequently, the next step in the development and application of this checklist is to prioritize checklist items. In doing this, it is important to collect and compare empirical evidence on the risks of DTC-GT services. Important factors that may influence prioritizing the checklist items include:

- The probability that a risk will occur—rare versus often
- The impact of the risk—wrongly informed health decisions versus learning about as yet unknown family members
- The availability of clear regulations that enable law enforcement—available versus absent

Additionally, the checklist could be used to determine crucial aspects of an informed purchase decision. To achieve this, there should be more focus on improving the current version of the checklist to facilitate ease of use for consumers. This could include items such as a visual online tool or decision aid and the use of accessible language. Information regarding the potential impact of gaining health insights through DTC-GT services may help empower consumers decide whether they, indeed, want to purchase and use a DTC-GT service, from which provider, and which health outcomes they do or do not want to learn more about. It may also help them understand how to interpret results, the impact of the results and consequent health decisions, and how their data are handled. This is essential, as the information on DTC-GT service websites is often misleading or incomplete [44,45].

XSL•FO
RenderX

Future Directions

Deciding on how to act in the online, complex, and diverse DTC-GT service market and to strike a balance between ensuring public safety and maintaining autonomy is challenging. To move forward, it is key that stakeholders maintain an ongoing discussion about the potential impact of these products on the public and about what impact should be desired or should be avoided. Stakeholders, such as policy makers, law enforcers, regulators, health care professionals, and members of the public, should be informed by ongoing research.

DTC-GT services combine elements of many mHealth products [46]. Specific elements of DTC-GT services can be seen in many mHealth technologies, such as easily retrievable health information, insight into personal health and disease risk, and products that enable preventive actions and self-management and empower citizens. Due to the sensitive nature of genetic data, DTC-GT services present additional challenges. These include issues regarding the reidentification and sharing of DNA within families, leading to privacy issues and concerns regarding the "right not to know" among consumers and their family members, and the challenges of comprehending and acting on complex personal risk information.

DTC-GT services and mHealth products in general challenge traditional distinctions between clinical care and self-promoted well-being: what is a medical matter, and what are personal lifestyle choices? These services raise questions about reimbursement policies, for instance, insurance companies that use devices to monitor customers' lifestyle and ultimately to adapt their premium. Another challenge lies in defining which devices require special certification because of specific safety issues [47]. Unraveling existing international policies on diverse mHealth products and how these policies affect the uptake and impact of mHealth products may aid policy makers in deciding how to respond to the market. Furthermore, comparing technologies and their potential consequences helps determine whether and how the public might be empowered or impeded.

The internet enables easy communication and the dissemination of information, which is a strong suit of DTC-GT service companies. To empower citizens and ensure informed purchase decisions, it is key that consumers be well informed about the potential impact of service aspects in all phases of the consumer journey. Aspects such as privacy and data management, scientific evidence, and information about how to use DTC-GTs and interpret their results are relevant to ensuring safety and maintaining autonomy. Studies that show variance in result comprehension [48], emphasis on positive aspects and negligence of potential risks by companies [45], limited consideration of potential risks by consumers prior to testing [49], and unmet expectations of DTC-GT service potential [49] suggest that the quality of the information currently available on DTC-GTs should be improved [44]. Further research is required to understand the communication strategies of companies, to point out information needs, and to determine how these needs should be met in order to support consumers in their decisions about DTC-GT services. The framework presented here could aid in distinguishing the various aspects of a service about which information can be provided, which information empowers consumers in their purchase decisions, and which information can be easily misunderstood.

In understanding the need for information provision, an important question is who should provide this support. This could include general practitioners [50], clinical geneticists and genetic counselors [51], nurses [52], pharmacists [53], government and public health institutes, and DTC-GT service providers. Several factors may impact the choice of which strategy would be fitting for dissemination. Examples of these factors include the media via which consumers are often exposed to offers of tests, such as TikTok [34] and YouTube [54], and consumer motivations to purchase, including the motivation to improve their own health [55] or an adoptee's health without knowledge of the family disease history [56]. Additional research should help clarify how consumers can be empowered in their decisions prior to and after purchasing a DTC-GT service [46].

Limitations

This study outlined the potential risks and opportunities in the consumer journey of DTC-GT services based on an integrative literature review. However, we were unable to determine the extent of either the risks or the opportunities or to point out which risks and opportunities should have priority over others. Furthermore, due to the complexities involved, the study lacks in-depth insight into which legislation applies to the diverse aspects of the services in each specific country. Answers to these questions will provide insight into policy decisions related to the DTC-GT service market. To gain knowledge of this, more empirical evidence is needed on items such as the validity of DTC-GTs, the wanted or unwanted effects of taking DTC-GTs, or factors that codetermine whether risks and opportunities will occur.

Second, the synthesized checklist is not a tool that policy makers, law enforcers, and regulators can directly apply to any policy. Therefore, the checklist is not yet a ready-made instrument. However, in their current forms, the checklist and consumer journey may together serve as a framework for multiple policy needs. The checklist has room for further refinement, depending on policy needs, such as insight into information provision on DTC-GT websites or appraisal of legislation governing distinct aspects of the service. Potentially, future checklists could be designed to focus specifically on 1 jurisdiction or 1 aspect of the consumer journey governed by specific legislation (eg, clinical validity under In Vitro Diagnostics Regulation [11]).

Third, the findings of our literature search may be biased toward the risks of the DTC-GT market. This may be reflected in the summary of findings and the checklist. The publications included indicate a greater focus on the risks of the DTC-GT service market than on the opportunities. This could be partially due to the fact that experts are rather skeptical about the potential health benefits of the DTC-GT market because of the immaturity of scientific evidence on the validity and utility of the tests offered by companies, for instance.

A shift in the study approach toward the effects of DNA testing for disease prevention, personalized medicine, or increased

XSL•FO
RenderX

public education could reveal more potential opportunities for the DTC-GT market. For example, research on pharmacogenetic testing and health outcomes may provide further insight into the potential benefits of the commercial genetic testing market [57]. Moreover, the findings of these studies could also enhance knowledge of those elements that are key to improving the effectiveness and safety of commercially available genetic testing services.

Fourth, this study excluded DTC-GTs for preconception carrier testing and prenatal testing. However, these tests could impact public health indirectly, as they could affect reproductive decisions. Couples may decide to not have children or seek additional medical care in conceiving a child. In addition, other issues may result from carrier or prenatal genetic testing, such as informed reproductive decisions, sex selection, and stigmatization or discrimination against disabled individuals [58,59]. Insight into issues of DTC-GT services for reproductive purposes may be relevant for policy experts. Many issues found in this study may be applicable for carrier or prenatal DTC-GTs, such as questions regarding privacy, validity, and learning about unexpected disease risks. Presumably, the need for informational support may be of comparable importance, or even greater importance, to consumers who are considering taking carrier or prenatal DTC-GTs and then using the results in their reproductive decisions.

Conclusion

This study summarized the risks and benefits associated with distinct DTC-GT aspects at a personal and a public level. Both the consumer journey and the checklist break the offer of a DTC-GT service down into key aspects that may impact and compromise the safety and autonomy of the individual and of public health. This framework helps policy makers, regulators, and law enforcers develop methods to interpret, assess, and act in the DTC-GT service market. This could serve as a backbone for future research into key themes such as the need for provision of information. The checklist could be further developed into a decision support tool to empower consumers in their decisions both before and after purchasing a DTC-GT service.

Acknowledgments

The authors wish to thank Dr Inge van Klink-de Kruijff (IKK) and Dr Teresa Leonardo Alves (TLA) for their contributions to the study design. Furthermore, we would like to thank the 5 experts who provided key papers and reviewed the findings and aided in result interpretation.

This study would not have been possible without the request for research and funding from the Dutch Ministry of Health, Welfare and Sports. This study was partly financed by the Empowering Citizens for Responsible Use of Direct-to-Consumer Genetic Testing (ERUDIGIT) project, which received funding from the Netherlands Organization for Health Research and Development (grant number 05550402110010). An open access publication would not have been possible without the support of the Dutch National Institute for Public Health and the Environment, Centre for Health Protection.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request. A list of all the included papers can be found in Multimedia Appendix 2.

Authors' Contributions

MCC, MEJ, SMO, and TR contributed to conception and design of the study. SMO organized the database, analyzed the data, and wrote the first draft of the manuscript. All authors have contributed to manuscript revision and have read and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for DTC-GTs in Embase.com (Medline + Embase). DTC-GT: direct-to-consumer genetic test.
[DOCX File , 18 KB - jmir_v26i1e47389_app1.docx]

Multimedia Appendix 2
List of all the articles included in the literature study. The list combines the articles from both the initial literature review and the update.
[DOCX File , 49 KB - jmir_v26i1e47389_app2.docx]

Multimedia Appendix 3
DTC-GT webshop questionnaire. DTC-GT: direct-to-consumer genetic test.

XSL•FO
RenderX

[DOCX File , 51 KB - jmir_v26i1e47389_app3.docx]

References

1. Hudson K, Javitt G, Burke W, Byers P. ASHG statement* on direct-to-consumer genetic testing in the United States. Am J Hum Genet 2007 Sep;81(3):635-637. [doi: 10.1086/521634]
2. Wetterstrand K. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). National Human Genome Research Institute. 2016. URL: https://www.genome.gov/sequencingcostsdata/ [accessed 2024-06-21]
3. Ramos E, Weissman SM. The dawn of consumer-directed testing. Am J Med Genet C Semin Med Genet 2018 Mar;178(1):89-97. [doi: 10.1002/ajmg.c.31603] [Medline: 29512889]
4. Thiebes S, Toussaint PA, Ju J, Ahn J, Lyytinen K, Sunyaev A. Valuable genomes: taxonomy and archetypes of business models in direct-to-consumer genetic testing. J Med Internet Res 2020 Jan 21;22(1):e14890 [FREE Full text] [doi: 10.2196/14890] [Medline: 31961329]
5. Badalato L, Kalokairinou L, Borry P. Third party interpretation of raw genetic data: an ethical exploration. Eur J Hum Genet 2017 Nov 23;25(11):1189-1194 [FREE Full text] [doi: 10.1038/ejhg.2017.126] [Medline: 28832567]
6. Nelson SC, Fullerton SM. "Bridge to the literature"? Third-party genetic interpretation tools and the views of tool developers. J Genet Couns 2018 Aug 07;27(4):770-781. [doi: 10.1007/s10897-018-0217-9] [Medline: 29411211]
7. Lachance CR, Erby LA, Ford BM, Allen VC, Kaphingst KA. Informational content, literacy demands, and usability of websites offering health-related genetic tests directly to consumers. Genet Med 2010 May;12(5):304-312. [doi: 10.1097/gim.0b013e3181dbd8b2]
8. Hoe het werkt. AnalyseMe. URL: https://www.analyseme.nl/hoe-het-werkt [accessed 2024-06-27]
9. Fears R, ter Meulen V, EASAC–FEAM Working Group. The perspective from EASAC and FEAM on direct-to-consumer genetic testing for health-related purposes. Eur J Hum Genet 2013 Jul 21;21(7):703-707 [FREE Full text] [doi: 10.1038/ejhg.2012.238] [Medline: 23169492]
10. Majumder MA, Guerrini CJ, McGuire AL. Direct-to-consumer genetic testing: value and risk. Annu Rev Med 2021 Jan 27;72(1):151-166. [doi: 10.1146/annurev-med-070119-114727] [Medline: 32735764]
11. Hoxhaj I, Stojanovic J, Sassano M, Acampora A, Boccia S. A review of the legislation of direct-to-consumer genetic testing in EU member states. Eur J Med Genet 2020 Apr;63(4):103841. [doi: 10.1016/j.ejmg.2020.103841] [Medline: 31923586]
12. Ziegler A, Koch A, Krockenberger K, Grosshennig A. Personalized medicine using DNA biomarkers: a review. Hum Genet 2012 Oct 1;131(10):1627-1638 [FREE Full text] [doi: 10.1007/s00439-012-1188-9] [Medline: 22752797]
13. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 2015 Jun 27;8(1):33 [FREE Full text] [doi: 10.1186/s12920-015-0108-y] [Medline: 26112054]
14. Bland JS. Cardiology meets personalized lifestyle medicine. Integr Med (Encinitas) 2017 Dec;16(6):12-16 [FREE Full text] [Medline: 30936810]
15. Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, et al. From public health genomics to precision public health: a 20-year journey. Genet Med 2018 Jun;20(6):574-582 [FREE Full text] [doi: 10.1038/gim.2017.211] [Medline: 29240076]
16. European Society of Human Genetics. Statement of the ESHG on direct-to-consumer genetic testing for health-related purposes. Eur J Hum Genet 2010 Dec;18(12):1271-1273 [FREE Full text] [doi: 10.1038/ejhg.2010.129] [Medline: 20736974]
17. Dinulos MBP, Vallee SE. The impact of direct-to-consumer genetic testing on patient and provider. Clin Lab Med 2020 Mar;40(1):61-67. [doi: 10.1016/j.cll.2019.11.003] [Medline: 32008640]
18. Sheehan M. The right to know and genetic testing. J Med Ethics 2015 Apr 23;41(4):287-288. [doi: 10.1136/medethics-2015-102767] [Medline: 25802356]
19. Bunnik EM, Janssens ACJW, Schermer MHN. Personal utility in genomic testing: is there such a thing? J Med Ethics 2015 Apr 28;41(4):322-326. [doi: 10.1136/medethics-2013-101887] [Medline: 24872596]
20. Bunnik EM. Do genomic tests enhance autonomy? J Med Ethics 2015 Apr 30;41(4):315-316. [doi: 10.1136/medethics-2014-102171] [Medline: 24980546]
21. Vayena E. Direct-to-consumer genomics on the scales of autonomy. J Med Ethics 2015 Apr 05;41(4):310-314 [FREE Full text] [doi: 10.1136/medethics-2014-102026] [Medline: 24797610]
22. Phillips AM. Only a click away - DTC genetics for ancestry, health, love…and more: a view of the business and regulatory landscape. Appl Transl Genom 2016 Mar;8:16-22 [FREE Full text] [doi: 10.1016/j.atg.2016.01.001] [Medline: 27047755]
23. Kalokairinou L, Howard HC, Slokenberga S, Fisher E, Flatscher-Thöni M, Hartlev M, et al. Legislation of direct-to-consumer genetic testing in Europe: a fragmented regulatory landscape. J Community Genet 2018 Apr 18;9(2):117-132 [FREE Full text] [doi: 10.1007/s12687-017-0344-2] [Medline: 29150824]
24. Wright CF, Hall A, Zimmern RL. Regulating direct-to-consumer genetic tests: what is all the fuss about? Genet Med 2011 Apr;13(4):295-300. [doi: 10.1097/gim.0b013e3181f69dd2]
25. Cernat A, Bashir NS, Ungar WJ. Considerations for developing regulations for direct-to-consumer genetic testing: a scoping review using the 3-I framework. J Community Genet 2022 Apr 16;13(2):155-170 [FREE Full text] [doi: 10.1007/s12687-022-00582-3] [Medline: 35171498]

XSL•FO
RenderX

26. Allyse MA, Robinson DH, Ferber MJ, Sharp RR. Direct-to-consumer testing 2.0: emerging models of direct-to-consumer genetic testing. Mayo Clin Proc 2018 Jan;93(1):113-120. [doi: 10.1016/j.mayocp.2017.11.001] [Medline: 29304915]
27. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995;123(3):A12-A13. [Medline: 7582737]
28. Mosdøl A. Targeted mass media interventions promoting healthy behaviours to reduce risk of non-communicable diseases in adult, ethnic minorities. Cochrane Database Syst Rev 2015;2(2):1-67. [doi: 10.1002/14651858.cd011683]
29. Alves TL, Mantel‐Teeuwisse AK, Paschke A, Leufkens HGM, Puil L, Poplavska E, et al. Unbranded advertising of prescription medicines to the public by pharmaceutical companies. Cochrane Database Syst Rev 2017;2017(7):CD012699. [doi: 10.1002/14651858.CD012699]
30. Leonardo Alves T, Martins de Freitas AF, van Eijk MEC, Mantel-Teeuwisse AK. Compliance of disease awareness campaigns in printed Dutch media with national and international regulatory guidelines. PLoS One 2014;9(9):e106599 [FREE Full text] [doi: 10.1371/journal.pone.0106599] [Medline: 25198776]
31. Giacomini M, Miller F, Browman G. Confronting the "gray zones" of technology assessment: evaluating genetic testing services for public insurance coverage in Canada. Int J Technol Assess Health Care 2003;19(2):301-316. [doi: 10.1017/s0266462303000278] [Medline: 12862188]
32. Haddow JE, Palomaki GE. An introduction to assessing genomic screening and diagnostic tests. Nutrit Today 2011 Jul;46(4):162-168. [doi: 10.1097/NT.0b013e3182261d7f]
33. Hall JA, Gertz R, Amato J, Pagliari C. Transparency of genetic testing services for 'health, wellness and lifestyle': analysis of online prepurchase information for UK consumers. Eur J Hum Genet 2017 Aug;25(8):908-917 [FREE Full text] [doi: 10.1038/ejhg.2017.75] [Medline: 28703222]
34. Basch CH, Fera J, Quinones N. A content analysis of direct-to-consumer DNA testing on TikTok. J Community Genet 2021 Jul 15;12(3):489-492 [FREE Full text] [doi: 10.1007/s12687-021-00526-3] [Medline: 33860464]
35. Burke S, Mork M, Qualmann K, Woodson A, Jin Ha M, Arun B, et al. Genetic counselor approaches to BRCA1/2 direct-to-consumer genetic testing results. J Genet Couns 2021 Jun;30(3):803-812. [doi: 10.1002/jgc4.1380] [Medline: 33550665]
36. Lee H, Vogel RI, LeRoy B, Zierhut HA. Adult adoptees and their use of direct-to-consumer genetic testing: searching for family, searching for health. J Genet Couns 2021 Feb;30(1):144-157. [doi: 10.1002/jgc4.1304] [Medline: 32602181]
37. Narod SA, Gojska N, Sun P, Tryon A, Kotsopoulos J, Metcalfe K, et al. The Screen Project: guided direct-to-consumer genetic testing for breast cancer susceptibility in Canada. Cancers (Basel) 2021 Apr 15;13(8) [FREE Full text] [doi: 10.3390/cancers13081894] [Medline: 33920818]
38. Nhan H, Zhan J, Kosinski C, Hampel H, Wu S, 23andMe Health Team. Allele frequencies of two MUTYH variants, Y179C, and G396D, in a direct-to-consumer genetic database. JCO 2020;38:63. [doi: 10.1200/JCO.2020.38.4_suppl.63]
39. Ryan MM, Cox CG, Witbracht M, Hoang D, Gillen DL, Grill JD. Using direct-to-consumer genetic testing results to accelerate Alzheimer disease clinical trial recruitment. Alzheimer Dis Assoc Disord 2021;35(2):141-147 [FREE Full text] [doi: 10.1097/WAD.0000000000000421] [Medline: 33060367]
40. Watanabe M, Hosono S, Nakagawa-Senda H, Yamamoto S, Aoyama M, Hattori S, et al. Does direct-to-consumer personal genetic testing improve gynecological cancer screening uptake among never-screened attendees? A randomized controlled study. Int J Environ Res Public Health 2021 Nov 24;18(23) [FREE Full text] [doi: 10.3390/ijerph182312333] [Medline: 34886058]
41. Ram N, Guerrini CJ, McGuire AL. Genealogy databases and the future of criminal investigation. Science 2018 Jun 08;360(6393):1078-1079 [FREE Full text] [doi: 10.1126/science.aau1083] [Medline: 29880677]
42. Horton R, Crawford G, Freeman L, Fenwick A, Wright CF, Lucassen A. Direct-to-consumer genetic testing. BMJ 2019 Oct 16;367:l5688 [FREE Full text] [doi: 10.1136/bmj.l5688] [Medline: 31619392]
43. Malgorzata M, Maria S, Michał W. Genetic testing—whether to allow complete freedom? Direct to consumer tests versus genetic tests for medical purposes. J Appl Genet 2022 Feb 26;63(1):119-126 [FREE Full text] [doi: 10.1007/s13353-021-00670-z] [Medline: 34826052]
44. Schaper M, Schicktanz S. Medicine, market and communication: ethical considerations in regard to persuasive communication in direct-to-consumer genetic testing services. BMC Med Ethics 2018 Jun 05;19(1):56 [FREE Full text] [doi: 10.1186/s12910-018-0292-3] [Medline: 29871685]
45. Covolo L, Rubinelli S, Ceretti E, Gelatti U. Internet-based direct-to-consumer genetic testing: a systematic review. J Med Internet Res 2015 Dec 14;17(12):e279 [FREE Full text] [doi: 10.2196/jmir.4378] [Medline: 26677835]
46. Davis TL, DiClemente R, Prietula M. Taking mHealth forward: examining the core characteristics. JMIR Mhealth Uhealth 2016 Aug 10;4(3):e97 [FREE Full text] [doi: 10.2196/mhealth.5659] [Medline: 27511612]
47. Lucivero F, Prainsack B. The lifestylisation of healthcare? 'Consumer genomics' and mobile health as technologies for healthy lifestyle. Appl Transl Genom 2015 Mar;4:44-49 [FREE Full text] [doi: 10.1016/j.atg.2015.02.001] [Medline: 26937349]
48. Ostergren JE, Gornick MC, Carere DA, Kalia SS, Uhlmann WR, Ruffin MT, PGen Study Group. How well do customers of direct-to-consumer personal genomic testing services comprehend genetic test results? Findings from the Impact of

XSL•FO
RenderX

Personal Genomics Study. Public Health Genomics 2015 Jun 16;18(4):216-224 [FREE Full text] [doi: 10.1159/000431250] [Medline: 26087778]
49. Roberts JS, Gornick MC, Carere DA, Uhlmann WR, Ruffin MT, Green RC. Direct-to-consumer genetic testing: user motivations, decision making, and perceived utility of results. Public Health Genomics 2017 Jan 10;20(1):36-45 [FREE Full text] [doi: 10.1159/000455006] [Medline: 28068660]
50. Cohidon C, Cardinaux R, Cornuz J, Chenal R, Desvergne B, Guessous I, et al. May direct-to-consumer genetic testing have an impact on general practitioners' daily practice? A cross-sectional study of patients' intentions towards this approach. BMC Fam Pract 2021 Apr 26;22(1):79 [FREE Full text] [doi: 10.1186/s12875-021-01428-6] [Medline: 33902442]
51. Martins MF, Murry LT, Telford L, Moriarty F. Direct-to-consumer genetic testing: an updated systematic review of healthcare professionals' knowledge and views, and ethical and legal concerns. Eur J Hum Genet 2022 Dec 12;30(12):1331-1343 [FREE Full text] [doi: 10.1038/s41431-022-01205-8] [Medline: 36220915]
52. Flowers E, Leutwyler H, Shim JK. Direct-to-consumer genomic testing: are nurses prepared? Nursing 2020 Aug;50(8):48-52 [FREE Full text] [doi: 10.1097/01.NURSE.0000684200.71662.09] [Medline: 32618766]
53. Gammal RS, Smith DM, Wiisanen KW, Cusimano JM, Pettit RS, Stephens JW, et al. The pharmacist's responsibility to ensure appropriate use of direct‐to‐consumer genetic testing. J Am Coll Clin Pharm 2021 May 16;4(5):652-658. [doi: 10.1002/jac5.1437]
54. Basch CH, Hillyer GC, Wahrman MZ, Garcia P, Basch CE. DNA testing information on YouTube: inadequate advice can mislead and harm the public. J Genet Couns 2021 Jun 31;30(3):785-792. [doi: 10.1002/jgc4.1375] [Medline: 33382174]
55. Schleit J, Naylor LV, Hisama FM. First, do no harm: direct-to-consumer genetic testing. Genet Med 2019 Feb;21(2):510-511 [FREE Full text] [doi: 10.1038/s41436-018-0071-z] [Medline: 29904164]
56. Baptista NM, Christensen KD, Carere DA, Broadley SA, Roberts JS, Green RC. Adopting genetics: motivations and outcomes of personal genomic testing in adult adoptees. Genet Med 2016 Sep;18(9):924-932 [FREE Full text] [doi: 10.1038/gim.2015.192] [Medline: 26820063]
57. Swen JJ, van der Wouden CH, Manson LE, Abdullah-Koolmees H, Blagec K, Blagus T, Ubiquitous Pharmacogenomics Consortium. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. Lancet 2023 Feb 04;401(10374):347-356. [doi: 10.1016/S0140-6736(22)01841-4] [Medline: 36739136]
58. Zaami S, Orrico A, Signore F, Cavaliere AF, Mazzi M, Marinelli E. Ethical, legal and social issues (ELSI) associated with non-invasive prenatal testing: reflections on the evolution of prenatal diagnosis and procreative choices. Genes (Basel) 2021 Jan 30;12(2):204 [FREE Full text] [doi: 10.3390/genes12020204] [Medline: 33573312]
59. Kraft SA, Duenas D, Wilfond BS, Goddard KA. The evolving landscape of expanded carrier screening: challenges and opportunities. Genet Med 2019 Apr;21(4):790-797 [FREE Full text] [doi: 10.1038/s41436-018-0273-4] [Medline: 30245516]

Abbreviations

DTC-GT: direct-to-consumer genetic test
mHealth: mobile health
PICO: population, intervention, comparison, and outcome
SES: socioeconomic status

Edited by A Mavragani; submitted 17.03.23; peer-reviewed by S Persky, E Onaku, M Siermann; comments to author 07.09.23; revised version received 10.11.23; accepted 09.05.24; published 17.07.24.

Please cite as:
Onstwedder SM, Jansen ME, Cornel MC, Rigter T
Policy Guidance for Direct-to-Consumer Genetic Testing Services: Framework Development Study
J Med Internet Res 2024;26:e47389
URL: https://www.jmir.org/2024/1/e47389
doi:10.2196/47389
PMID:

©Suzanne Maria Onstwedder, Marleen Elizabeth Jansen, Martina Cornelia Cornel, Tessel Rigter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Public Health Dashboards in Overdose Prevention: The Rhode Island Approach to Public Health Data Literacy, Partnerships, and Action

Jesse Yedinak[1], MPA; Maxwell S Krieger[1], BS; Raynald Joseph[2]; Stacey Levin[3], CPRS, CCHW; Sarah Edwards[4]; Dennis A Bailer[5]; Jonathan Goyer[6], BA; Colleen Daley Ndoye[5], MS; Cathy Schultz[7], MPH; Jennifer Koziol[4], MPH; Rachael Elmaleh[4]; Benjamin D Hallowell[4], PhD; Todd Hampson[1], BA; Ellen Duong[8], MSc; Abdullah Shihipar[1], MPH; William C Goedel[1], PhD; Brandon DL Marshall[1], PhD

[1]Department of Epidemiology, Brown University School of Public Health, Providence, RI, United States
[2]AIDS Care Ocean State, Providence, RI, United States
[3]Parent Support Network, Warwick, RI, United States
[4]Rhode Island Department of Health, Providence, RI, United States
[5]Project Weber/RENEW, Providence, RI, United States
[6]Rhode Island College, Providence, RI, United States
[7]State of Rhode Island Executive Office of Health and Human Services, Cranston, RI, United States
[8]Center for Computation and Visualization, Brown University, Providence, RI, United States

Corresponding Author:
Jesse Yedinak, MPA
Department of Epidemiology
Brown University School of Public Health
Box G S 121-3
121 South Main Street, 3rd Floor
Providence, RI, 02912
United States
Phone: 1 401 863 3375
Email: jesse_yedinak@brown.edu

Abstract

As the field of public health rises to the demands of real-time surveillance and rapid data-sharing needs in a postpandemic world, it is time to examine our approaches to the dissemination and accessibility of such data. Distinct challenges exist when working to develop a shared public health language and narratives based on data. It requires that we assess our understanding of public health data literacy, revisit our approach to communication and engagement, and continuously evaluate our impact and relevance. Key stakeholders and cocreators are critical to this process and include people with lived experience, community organizations, governmental partners, and research institutions. In this viewpoint paper, we offer an instructive approach to the tools we used, assessed, and adapted across 3 unique overdose data dashboard projects in Rhode Island, United States. We are calling this model the "Rhode Island Approach to Public Health Data Literacy, Partnerships, and Action." This approach reflects the iterative lessons learned about the improvement of data dashboards through collaboration and strong partnerships across community members, state agencies, and an academic research team. We will highlight key tools and approaches that are accessible and engaging and allow developers and stakeholders to self-assess their goals for their data dashboards and evaluate engagement with these tools by their desired audiences and users.

(J Med Internet Res 2024;26:e51671) doi:10.2196/51671

KEYWORDS

community engagement; data dashboards; data literacy; health literacy; overdose; public health communication; public health surveillance

XSL•FO
RenderX

Introduction

Public health data dashboards are a critical communication tool for governments, health organizations, and academic research teams to bring rapidly emerging health surveillance information to a broader audience [1,2]. The rise of data dashboards in the United States during the COVID-19 pandemic resulted in a rapid proliferation of interactive maps, charts, and graphs and the wide dissemination of disease surveillance data by governmental organizations and research teams [3,4]. However, such data dashboards were in their infancy, and the governmental and research organizations creating them had urgent and multidimensional needs. Dashboards were expected to convey rapidly changing disease patterns, monitor emerging issues in local communities, and communicate about the risk of individual exposure [4,5]. Data dashboards are now a common platform for communicating policy changes and displaying trends for decision makers, researchers, and the news media [6]. Public health data dashboards created to monitor the overdose crisis and other ongoing public health issues across the United States are no different. Given their growing presence and appeal, the landscape of public health data dashboards leaves room for focused development and improvement in engaging users through accessible data storytelling, sustainable community partnerships, and data-driven decision-making [4,5,7-10]. The aim of this viewpoint paper is to describe the Rhode Island approach for data dashboard development, an instructional approach to overdose dashboard development driven by our experiences to date. First, we will illustrate our approach using 3 overdose data dashboard examples from Rhode Island; then, we will outline key considerations for dashboard data and accessibility, reciprocal partnerships, and pathways for action.

Methods

Public Health Data Literacy: PreventOverdoseRI

Rhode Island's public data dashboard and educational information hub, PreventOverdoseRI [11] (nicknamed PORI), was launched in 2016 through a collaborative effort across state agencies, community organizations, and an academic research team. Making timely overdose surveillance data publicly accessible was critical to understanding the scale and scope of the growing overdose crisis and effectively distributing additional resources across the state during a period of rapidly rising fatal overdoses [9,12]. The first goal was to increase awareness of the overdose crisis by reaching a statewide public audience. Our next goal was to engage a subset of that audience as active dashboard users (ie, those who would stay on the website to interact with simple surveillance and resource maps and read educational infographics about overdose and the statewide policy response) [9,12].

Given the wide audience for this content, PORI was designed for accessibility and clear action steps by the users. This included descriptive headlines, simple data visualizations, and careful use of evidence-based tools such as the Center for Disease Control and Prevention (CDC) Clear Communication Index to assess content (Figure 1) [13]. Additional assessments included user testing of educational content and stakeholder involvement in data collection and dashboard dissemination [9,14]. Overdose surveillance data are visualized on PORI using the Tableau server [15]. Dashboard and educational content for PORI is hosted on a WordPress 6.3 website, using Google Analytics 4 (Google LLC) to monitor web traffic [16,17]. Evaluation of this type of data dashboard includes measuring levels of daily or monthly users and page views; duration of time and engagement on priority pages; and user pathway testing for understanding, ease of navigation, and follow-through to action, as demonstrated in Figure 2 [14]. Content development and improvement are ongoing, and user engagement with PORI continues to grow each year (with over 13,000 visits a month from 2018 to 2020 and 17,000 visits a month from 2020 to 2022).

XSL•FO
RenderX

Figure 1. PreventOverdoseRI: wireframing and headlining content.

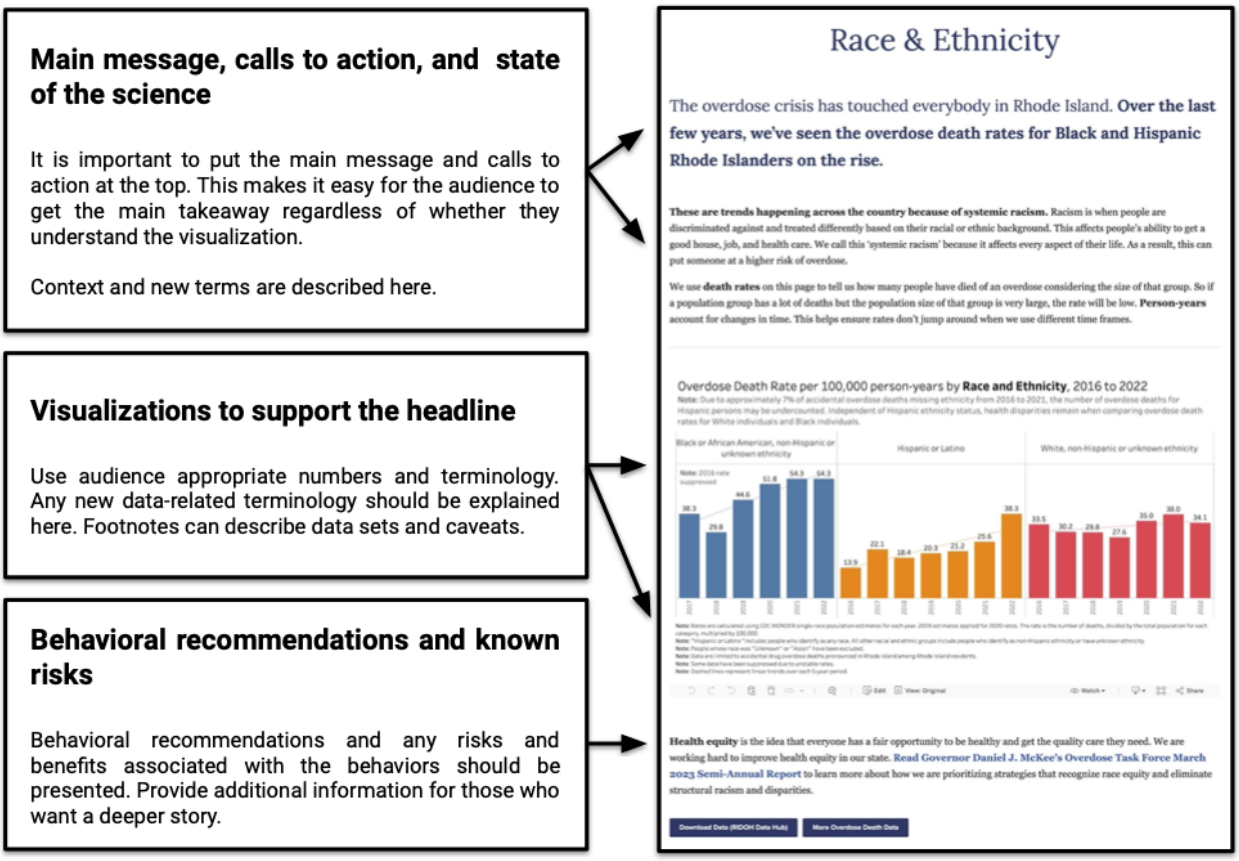

Figure 2. PreventOverdoseRI: evaluation metrics for public data dashboards.

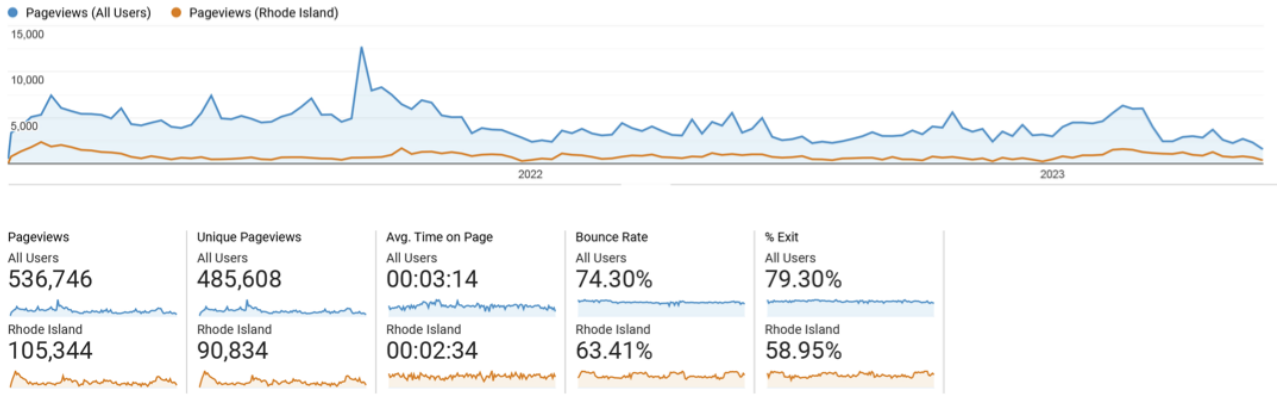

Meaningful Partnerships: The Vulnerability Investigation of Underlying Local Risk and Geographic Events Project

In 2018, our academic research team conducted the Vulnerability Investigation of Underlying Local Risk and Geographic Events (VILLAGE) project, an evaluative public health project that aimed to identify areas at higher risk of increased drug-related harms relative to other areas of the state. For this type of data dashboard partnership, community organizations played a central role in conveying key messages and the dissemination of information to the informed public. The emphasis was on creating a simple, responsive data dashboard mapping tool for community organizations to inform harm reduction resource distribution at the community level. The goal was to create maps with census-tract level granularity to convey community and structural vulnerability to drug-related harms, such as overdose and HIV or hepatitis C outbreaks, as compared to other census tracts in the state [18,19]. Given the more specialized end users and emphasis on structural data and predictive modeling results visualized on a map, the data dashboard involved stronger user testing and evaluation of the maps from the Rhode Island Department of Health and community stakeholders. Our academic research team provided support for technical training on machine learning uses and worked with partners to draft language to convey the results of the predive model, named "vulnerability scores," that prioritized census tracts for intervention. Community stakeholders helped lead the translation

XSL•FO
RenderX

and dissemination of the VILLAGE project and accompanying materials (Figure 3) [18,19]. Results were visualized with ESRI ArcMap (version 10.4; Environmental Systems Research Institute, Inc) [20]. The final map and findings were printed for local distribution as well as hosted on the internet in PDF format. Evaluation of this work included the level of community engagement with dissemination activities, monitoring the relative allocation of resources as compared to vulnerability scores, and use of the census-tract-level vulnerability maps to secure additional grant funding.

Figure 3. The Vulnerability Investigation of Underlying Local Risk and Geographic Events Project: contextualizing relative risk levels across the state.

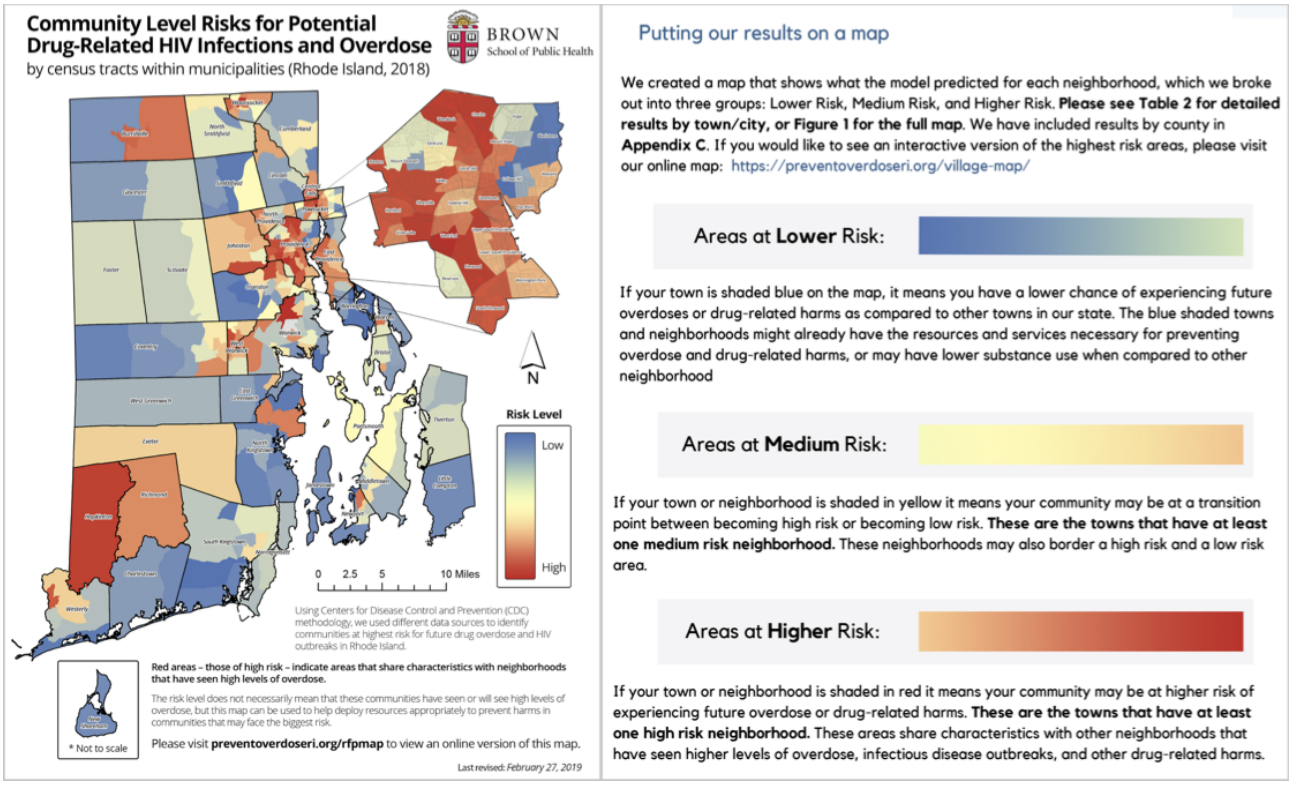

Pathways for Action: Preventing Overdose Using Information and Data From the Environment Community Randomized Trial

In 2021, our academic research team launched Preventing Overdose Using Information and Data From the Environment (PROVIDENT), a statewide, randomized, community-level trial designed to assess the impacts of proactive harm reduction resource allocation to census block groups prioritized based on a predictive analytics model [21]. While this data dashboard features increased complexity of available information for end users, the data dashboard is still organized for rapid usability and uptake, as illustrated in Figure 4. The PROVIDENT dashboard places a higher emphasis on tablet-friendly or field use of the dashboard maps, and the predictive model results are visualized by shading at-risk neighborhoods. The dashboard was developed in *Vue 3*, a JavaScript library, using Vega (University of Washington Interactive Data Lab) for visualizations, Google Firestore (Google LLC) for data storage, and Google Analytics 4 for monitoring web traffic and engagement [17,22-24]. Evaluation of the implementation work for this community trial includes structured focus groups by end users, pre- and posttest assessments of workshops and public health data literacy materials, and a better understanding of the multitude of reasons community stakeholders may engage with the data dashboard and how they use data to drive action.

XSL•FO
RenderX

Figure 4. Preventing Overdose Using Information and Data From the Environment (PROVIDENT) community randomized trial: wireframing and mapping.

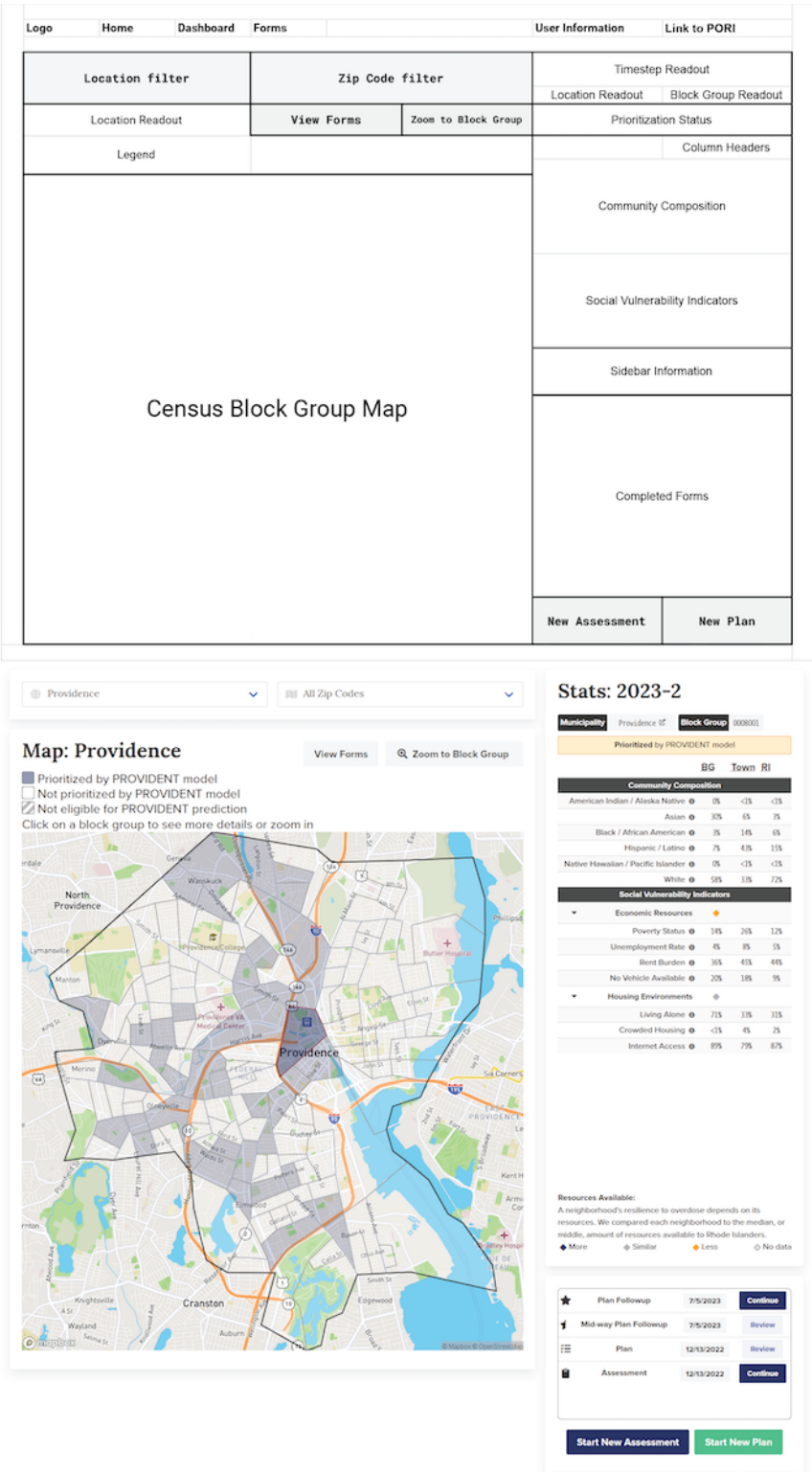

Discussion

Building Context Using Positionality and Structural Data

A broad goal of public health data dashboards is to increase awareness of emerging issues and build trust through data sharing and transparency, while ultimately telling the "public health stories" of communities and connecting data to decision-making and results [1,25,26]. In communicating public health concerns, the organizations creating a data dashboard are in a powerful position to determine their audiences and users; define and narrate the issue for their selected groups; and make key decisions regarding what data are collected, how they are analyzed, and how they are visualized and interpreted [10,27]. The volume of data made available on the internet can vary widely, from a municipal or state "open data" portal that allows

XSL•FO
RenderX

end users to view large collections of row-level data to the more traditional epidemiological surveillance report on a governmental website [26]. Critically, the decisions about which data to curate and include in a dashboard affect how risk factors and disease drivers are explained, whether structural determinants are identified or measured, and if disparate health outcomes are assessed or contextualized for the intended audience and presented in an accessible way [6,10,27,28].

Building trust is a prerequisite for any entity that wishes to develop a public health data dashboard. Data dashboards are now a crucial platform to scale up regional capacity for conveying critical public health messages and emerging changes, and they require continuous work to ensure audience engagement and dashboard usability [29]. As an academic research group, an important step in trust-building for these data dashboards was to name our positionality and connection to institutional resources in the partnership development process of each new dashboard. Positionality is a process used in community-engaged and qualitative research methods, where researchers and investigators explicitly name their affiliations, privileges, identities, and other relations to power structures as they engage in community partnerships and work to address gaps in resources across partnerships [30]. Examples of building trust and transparency include engaging local partners to help increase local information represented on the dashboard and improve the granularity of maps as part of that process (Table 1 contains additional examples). As representatives of research institutions, governmental organizations, or other structures of power, we must be careful not to repeat, minimize, ignore, or perpetuate long-standing harms of systemic, institutional, and personal racism against Black and other minoritized populations in the medical, research, governmental, and public health fields [31-35]. Rhode Island has a system of work groups sponsored by the Governor's Overdose Prevention and Intervention Task Force (Task Force), including a racial equity work group whose work brings attention to disparate health outcomes in access to care and service delivery by highlighting gaps in data. Specifically, these stakeholders brought awareness to how underreported racial disparities in existing data sets, including methadone uptake and increases in death rates among Black and Hispanic Rhode Islanders, needed to be contextualized and displayed on PORI. Compounding the disparate health outcomes of the overdose crisis is the history of racialized criminalization of drug use, excessive use of policing, and overincarceration of Black and Hispanic or Latine Americans who use drugs [32,36,37]. Central to the work of addressing substance use and overdose through a public health data dashboard is the acknowledgment and identification of how racist structural and political forces create our current climate of health inequities and disparate health outcomes, drive our interpretation of data, and influence local action and resource allocation [36,37].

XSL•FO
RenderX

Table 1. Self-assessing public health data literacy and user engagement.

Design considerations	Self-assessment and tools
Capture the attention of your intended audience	Who is your intended audience? • Bite, Snack, Meal: a tool to curate and edit content using the intended audience to adjust the level of detail. • Headlining: the main headline clearly explains and interprets the main issue and key figures to engage your intended audience (Figure 1).
Making decisions about content to engage people	What is the story? • Wireframing: organize information on the page so that the most impactful information is first, ending with the recommended action (Figures 1 and 4). • Representation: community strengths, voice, and expertise are included and centered. Stories about overdose and drug use avoid stigmatizing terms and narratives.
Increase accessibility for intended users.	How is the story being told? • Increase accessibility for intended users. Does the story use personal narrative, counts, percentages, or rates? What type of data storytelling resonates most with your audience? • Use the CDC[a] Clear Communication Index tool to assess health literacy and prioritization of health information (Figures 1 and 3) [13]. • Make content accessible. Use readability tools (sixth- to eighth-grade US reading level). The Simple Measure of Gobbledygook was created for health information in English and has open-access web-based tools, or the Fleish-Kinkaid tool, which is included in common word processing applications [38,39].
Increase transparency with context and multiple types of data	Is the story the same across all groups and categories? • How does our story change when looking at race, gender, geography, or incarceration experiences? What visualizations clearly illustrate these differences in health outcomes? • Build transparency with detailed footnotes to clarify definitions, types of analyses, and limits of the data or metrics. Increase the likelihood of engagement with the data and content by choosing data that reflect the diversity of race, gender, age, and other key determinants of health.
Build user engagement with dashboard elements	Is there more to the story? • Define structural racism and structural determinants of health and offer context about redlining or patterns of historical and structural investments or disinvestments. • Create data that are interactive, and allow users to adjust timeframes, toggle between counts and rates, and map that zoom to zip code or census tract level if possible. • Add local information, interactive maps, and relevant policy timelines to guide decisions and actions.
Create pathways for action	What does this mean for me? • Make it personal for end users with a call to action on each page. Examples include connecting to additional education, encouraging harm reduction behavior change, offering referrals to care, or advocating for policy action.

[a]CDC: Centers for Disease Control and Prevention.

For PORI, where the goal is to engage a wide audience with public health data stories, the focus is on showing improved reporting on race and ethnicity across key data points related to health care engagement and outcomes during public Task Force meetings and visualizing both counts and rates on the dashboard whenever possible. Conversely, with the VILLAGE and PROVIDENT dashboards, the goal is to engage partners and create pathways for data-driven policy response and resource distribution. Therefore, those dashboards include additional neighborhood-level context, such as community demographics and social vulnerability indicators, which are chosen in collaboration with community stakeholders and can act as a proxy for identifying communities facing increased structural vulnerabilities (Figure 4).

In particular, creators of data dashboards have a responsibility to define, adjust, and disaggregate the types of data they collect from Black, Indigenous, Asian, Hispanic or Latin, and other marginalized racial and ethnic communities and contextualize why and how such data are collected and analyzed [27,28,35,40]. Reporting on health inequities and outcomes to a broad audience should include naming socially constructed categories of race, addressing the historical drivers of persistent and unequal health outcomes, and including more community-level reporting [28,40,41]. This is one such process that reveals how racism (rather than race) is the root cause of the differences we might observe by race and allows for disparities and inequities that may otherwise be hidden in the data to come forward. Ponce et al [27] speak to a wide range of data-driven approaches to improving data equity, which include disaggregating and contextualizing “racialized” identities in our US governmental data sets and the importance of visibility of additional ethnic identities as a means toward visibility and health outcomes. They draw out this concept further, citing the rapid growth of “multiracial” identities, and suggest other contextual questions

XSL•FO
RenderX

to address the intersection of health and race: “depicting the health needs of multiracial people is more complex than a better measurement of group membership” [27].

Public health data dashboards also present the opportunity to increase the visibility of some structural systems of racism that create harm in the United States, particularly through the use of maps, data storytelling, and prioritization of resources in housing, education, health care, and transportation within a community. Overlays on maps highlight historical forces of segregation and structural disinvestment, such as through redlining maps, and reveal areas to be prioritized for resources and intervention [31,32,42]. Therefore, to begin addressing the structural harms and outcomes of systemic and institutionalized racism, data dashboard developers can create shared community power and oversight over definitions of race and other identifiers, the process of data collection, and increased use of data on structural drivers of health [40,43,44].

Reciprocal Partnerships and Engaging Users

The diverse components of a data dashboard have led to our curated approach to data sharing and intentional oversight by community stakeholders and users [9]. This system has been made possible by funding, collaboration, iteration, and strong community engagement and oversight from the extended overdose prevention and intervention community, including peer advocates, community organizations, governmental agencies, and our academic research center [9,14]. Assessing partnerships and defining and engaging an audience of potential dashboard users are among the first steps to creating a data dashboard, as outlined broadly in Figure 5. In Table 2, we detail 3 groups of decision makers who can share oversight for developing and creating a community-driven data dashboard and potential pathways for partnerships and cofunding. The broad decision-making groups include: (1) community stakeholders, including people with lived experience from a diversity of racial, ethnic, and cultural backgrounds; local nonprofits and leaders; and local policy makers closest to the issue who can intervene directly, cocreate, or improve data collection and interpretation, and advise on policy; (2) academic partners, or academic research teams, who have access to analytic and data security resources, partnership funding, national and international context, and methods for research and can develop evidence for interventions and policy change; and (3) governmental agencies who represent or own health and surveillance data, can help shape policy directives and resource distribution, and increase access to national funding.

Figure 5. Establishing partnerships and engaging users.

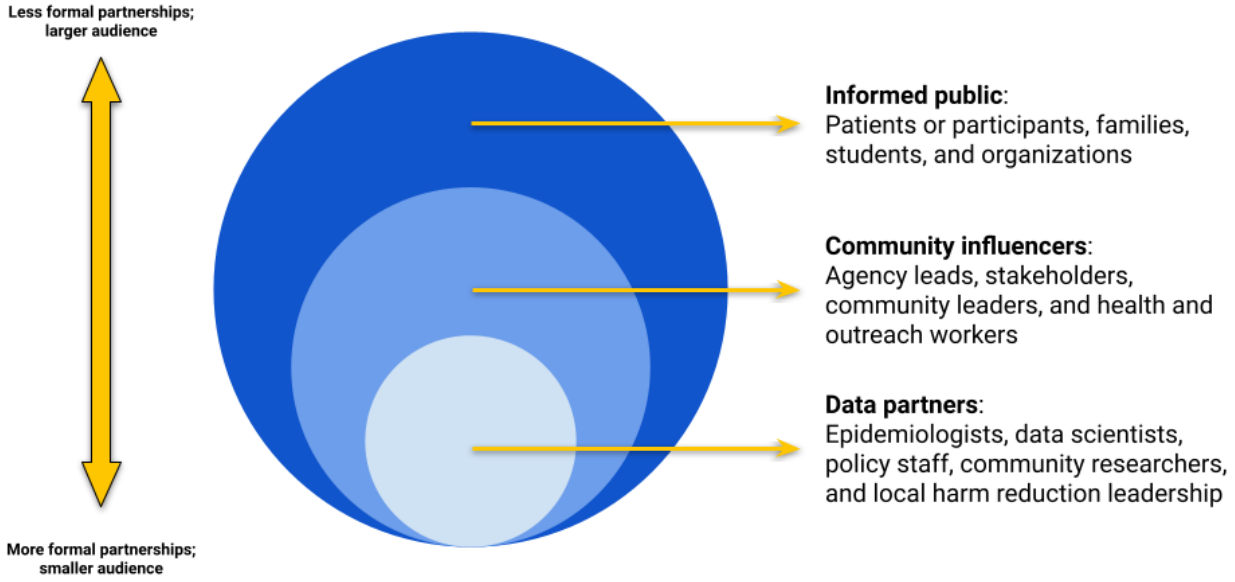

XSL•FO
RenderX

Table 2. Defining data dashboard goals and partners.

Dashboard type	User goals	Decision-making partners
Public health data literacy (PORI[a])	Dashboard users may interact with the headlines, engage in data visualizations such as maps and bar charts, review educational content, and use resource maps.	• Community stakeholders advocate for their information needs and priorities. Give input on data governance, the interpretation of data, how to contextualize with mapping, and messaging to connect with the audience. • Academic partners may host data dashboards, data visualization, and data-sharing security. Helps frame messaging and interpretation of data trends in the context of national or international trends. Provide analytic support and rigor. • Governmental agencies coordinate volunteer working groups on data governance and support purchasing web resources. Helps frame messaging and interpretation of data trends from a policy leadership point of view. Determines analytic processes and partners. Secures funding.
Meaningful partnerships (the VILLAGE[b])	Dashboard users interact heavily with the data visualizations and maps and offer expertise on education and resources.	• Community stakeholders establish contracts with key community organizations, accountable to the government. Community organizations participate actively in the creation of data dashboards and are compensated for their time, meeting attendance, survey completion, and agreement to participate. • Academic partners further develop data security measures, which may include protected log-ins for selective access. Data are readily available and contextualized, and headlines clearly interpret results and footnotes to help drive policy decisions and resources for intervention. Curated maps with increased detail are tested with community stakeholders. • Governmental agencies establish contracts with key community organizations, are accountable to the government, and have resources to contribute. Funders may organize meetings with stakeholders to interpret or translate data collectively.
Pathways for action (PROVIDENT[c])	Dashboard users influence how and what data are collected, analyzed, and visualized. They may request highly specific interactive features to improve the functionality and utility of the dashboard.	• Community stakeholders: community organizations participate actively in the creation of data dashboards and are compensated for their time. Formal contracts and scopes of work are required to define each partner's role. • Academic partners may be a stronger leader in this role, providing dedicated research funding and engagement opportunities. Engaging community influencers by building shared language, training support for interpreting and contextualizing the data, and making data products that are highly actionable and usable for the community. • Governmental agencies: governmental organizations engage with all partners to discuss the interpretation of data, refine data collection processes, and address missing information.

[a]PORI: PreventOverdoseRI.

[b]VILLAGE: Vulnerability Investigation of Underlying Local Risk and Geographic Events.

[c]PROVIDENT: Preventing Overdose Using Information and Data From the Environment.

The ways in which the decision-making groups interact and govern each project vary by dashboard type, as noted in Table 2. With PORI, the academic partners are highly responsive to the feedback and input of the community stakeholders as a means to increase the accessibility of the dashboard [9]. Content creation and data oversight from community stakeholders occur in public or semipublic meetings, which are often organized and facilitated by government agencies (ie, the Task Force and related working groups). For the VILLAGE project, community stakeholders advised on data set selection and helped create the dissemination plan, which included a community white paper, brief "1-pagers" to describe high-level results in lay terms, and a statewide conference with sessions created and led by the stakeholders [18,19]. Finally, for our PROVIDENT study, oversight of the dashboard is led by academic partners but governed through structured community advisory meetings and project team meetings with contracted community stakeholders and government agencies. Community stakeholders also help cofacilitate workshops and develop "1-pagers" as we scale up and disseminate the study results.

Next, in Figure 5, we define 3 types of dashboard partnerships to help guide editorial decision-making and content development, which should inform how data are displayed and contextualized for users and determine action steps. The first audience we focus on is the informed public, which includes the local audience that needs access to emerging public health headlines and visualizations that develop an overall data story or narrative. The second audience we identify is community influencers, key community organizations, and leaders in the public health response who have existing partnerships with

XSL•FO
RenderX

governmental agencies. This audience can engage with maps and other interactive visualizations as active dashboard users and provide ongoing insights to help prioritize resources at the town or neighborhood level and identify structural disparities or emerging needs. Our last audience is the data partners, who include all stakeholders involved in cocreating and iterating on data dashboard tools, acting as users of the data to get ahead of emerging issues, address systemic disparities, highlight community strengths, and lead strategic planning and policy change.

Operationalizing Public Health Data Literacy

We use a self-assessment process to think about engaging potential dashboard users with key headlines but also with interactive elements such as data visualizations and maps (Table 1). Some data elements work best when the intended user has some exposure to statistical concepts such as rates and percentages, geographical units, and an understanding of 1 or more public health data sets [45,46]. Data visualizations can quickly become complex and may use a number of analytic methods and software types that rely on an understanding of how to navigate and operate a user interface [47,48]. Best practices in user interface and data dashboard research suggest such products should be created in partnership with the intended users and monitored or revised continuously for improvement and engagement metrics, but resource constraints and hesitancy to use data dashboards for broader communication have limited the use of these tools in overdose prevention and intervention at the local level [4,7,45,47,49-52].

We identified a key set of skills to support wider audience engagement and accessibility, which we refer to as public health data literacy. Health literacy is defined in the United States as "when a society provides accurate health information and services that people can easily find, understand, and use to inform their decisions and actions" [53-56]. The concept and field of health literacy have been explored, researched, and further defined in the past 20 years to reveal differences in health outcomes based on individual "levels" of health and health systems literacy, highlighting the importance of accessible health communication and reliable information on accessing systems of care [56-58]. To contextualize this as it applies to public health dashboards, we suggest that numeracy, a component of health literacy, should be elevated as a core skill of data dashboard engagement. Numeracy is "the degree to which individuals have the capacity to access, process, interpret, communicate, and act on numerical, quantitative, graphical, biostatistical, and probabilistic health information needed to make effective health decisions" [46,59]. Establishing health literacy levels, including numeracy, can be achieved using evidence-based tools for assessing web-based content. We recommend reviewing individual dashboard pages using the CDC Clear Communication Index [13]. Designing a data dashboard with these skills in mind affects how the user engages with the dashboard and their ability to act on data-based recommendations, not just at the individual level with dashboards such as PORI, but at the organizational, community, and policy levels using data from forecasting-driven dashboards such as the VILLAGE and PROVIDENT. Therefore, we think of public health data literacy as how our users collect, interpret, and understand the public health data we need and use it to guide community-level action and policy decisions for better health outcomes and systemic change.

To increase trust in public health data dashboards, we recommend health literacy and accessibility efforts be done in tandem with community stakeholder partnerships for improved creation, interpretation, and prioritization of results and action [9,29,60-62]. As academic partners, we paired public health training sessions with ongoing feedback sessions. These interactive training sessions were first developed as an introduction to dashboard storytelling with PORI and soon evolved to include public health terminologies such as counts, percentages, rates, and relative risk for the VILLAGE project. Ultimately, as the dashboard complexity grew, we created a workshop for the PROVIDENT study to discuss the processes behind prediction and forecasting, as well as the other concepts from previous training sessions. In Table 1, we focus on our latest self-assessment questions to develop and contextualize the data and present it as a narrative or story, outlining key questions the story should answer for optimal user engagement and action. Each question in Table 1 is accompanied by a list of suggested tools for editing, curating, and contextualizing the data. We begin with the self-assessment question, "Who is the audience?" We suggest that most audiences benefit when information is curated to the key conclusions, written as headlines, to drive dashboard usage and engagement with the content. Here we use the marketing tool "Bite, Snack, Meal" to refer to how much of the data story to share with each audience [63]. For example, highly curated data and easy-to-remember headlines are useful for nearly all audiences to broadly understand and engage with the issue and are considered the "bite." On the other hand, the deep dive into the data is the "meal" and should be reserved for the most engaged stakeholders and users. Next, we ask, "What is the story?" We offer specific suggestions for the organization and structure of the story on the data dashboard page for the greatest accessibility and understanding. The following 3 questions—"How is the story being told? Is the story the same across all groups and categories? Is there more to the story?"—each refers to dimensions of the issue and engagement to ensure sufficient context and connection with key issues and methodologies used. There are specific tools listed to address numeracy, health disparities, and historical or political context within each of those questions. Finally, "What does this mean for me?" is our data dashboard action step. Each data dashboard page can lead to the desired action steps for the user to go beyond the data, such as links to education, connections to local care and resources, or opportunities for community and policy engagement.

Summary

The lessons learned across 3 overdose dashboard projects in Rhode Island offer insights into the complexity of providing data to engage communities and using data to drive action on emerging issues. PORI illustrated for us how research and governmental institutions that provide data have a significant role to play in making the data accessible and in addressing the complexity of disparate health outcomes and data equity in our communities. Data partners such as Rhode Island's Executive

XSL•FO
RenderX

Office of Health and Human Services lead with a framework for centering racial justice, stating openly, "We acknowledge and seek to address the historical and ongoing harm caused by systemic oppression, including the role of state and non-state actors in perpetuating these harms." [64]. Furthermore, community engagement and partnerships are essential to building a reciprocal relationship, trust, and accountability with the dashboard users, as we explored in the VILLAGE project. As noted by Ponce and colleagues [27], "For a solid foundation, there must be visibility of marginalized communities throughout the pipeline of data equity, that is, from data collection or capture to data use to how data are used and by whom." [27]. Finally, through the PROVIDENT community trial, we saw the powerful insights brought about through cocreated dashboards and stronger formalized and compensated partnerships. Similar conclusions were drawn from the HEALing Communities Study: "dashboards can be used to foster community-driven solutions to address the opioid overdose epidemic" [52].

Conclusion

Public health and overdose surveillance and dashboarding methods will continue to evolve as new public health crises emerge and software and web tools enable broad dissemination. Given the scale and escalation of changes in the overdose crisis and other emerging public health issues, the need for rapid and effective communication and local response will likely remain rooted in the key concepts learned thus far: accessibility, partnerships, and action. The approach presented herein offers accessible and concise steps that local governments and academic research partners can take to proactively engage and cocreate advanced epidemiological content with communities while preparing for the increasing demand and rapid scale-up of public health data visualization and communication during times of highest need.

Acknowledgments

We would like to acknowledge the contributions of the many organizations and individuals who have contributed to this work in Rhode Island, including people whose anonymized health data have been used to provide critical surveillance information in our state—people who use drugs, people who are on medications for opioid use disorder, and people who identify as being in recovery. We would like to acknowledge the team at the Center for Computation and Visualization (CCV) at Brown University. Part of this work was conducted using the robust visualization resources and services at the CCV. We would like to thank our students at the People Place and Health Collective at Brown University School of Public Health, who have contributed to this work over the years, and the staff who make this work possible, including Jackie Goldman at the People, Place, and Health Collective for their project management and community engagement oversight, and Sam Bessey and Seashore Li of the People, Place, and Health Collective for their programming and analytic expertise.

This work is supported by the National Institute on Drug Abuse (grant R01DA046620) and by the Center for Disease Control and Prevention Overdose to Action (grant RFA-CE15-1501) grants awarded to the Rhode Island Department of Health.

Disclaimer

The Rhode Island Department of Health (RIDOH) and the State of Rhode Island Executive Office of Health and Human Services (EOHHS) are not responsible for any analyses, opinions, or conclusions contained in this document. The views expressed in this report are those of the authors and do not represent the official positions or policies of the RIDOH or EOHHS.

Conflicts of Interest

None declared.

References

1. Lechner B, Fruhling A. Towards public health dashboard design guidelines. In: Nah FFH, editor. HCI in Business: First International Conference, HCIB 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings. Cham: Springer International Publishing; 2014:49-59.
2. Ross DA, Baker EL. A brief history of public health informatics-lessons for leaders and a look into the future. J Public Health Manag Pract 2023;29(1):101-104 [FREE Full text] [doi: 10.1097/PHH.0000000000001672] [Medline: 36448762]
3. Fareed N, Swoboda CM, Chen S, Potter E, Wu DTY, Sieck CJ. U.S. COVID-19 State Government Public dashboards: an expert review. Appl Clin Inform 2021;12(2):208-221 [FREE Full text] [doi: 10.1055/s-0041-1723989] [Medline: 33853140]
4. Ivanković D, Barbazza E, Bos V, Fernandes Ó, Gilmore KJ, Jansen T, et al. Features constituting actionable COVID-19 dashboards: descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. J Med Internet Res 2021;23(2):e25682 [FREE Full text] [doi: 10.2196/25682] [Medline: 33577467]
5. Sorapure M. User perceptions of actionability in data dashboards. J Bus Tech Commun 2023;37(3):253-280 [FREE Full text] [doi: 10.1177/10506519231161611]
6. Thorpe LE, Gourevitch MN. Data dashboards for advancing health and equity: proving their promise? Am J Public Health 2022;112(6):889-892 [FREE Full text] [doi: 10.2105/AJPH.2022.306847] [Medline: 35446603]

XSL•FO
RenderX

7. Claborn K, Creech S, Conway FN, Clinton NM, Brinkley KT, Lippard E, et al. Development of a digital platform to improve community response to overdose and prevention among harm reduction organizations. Harm Reduct J 2022;19(1):62 [FREE Full text] [doi: 10.1186/s12954-022-00636-2] [Medline: 35658871]
8. Garett R, Chiu J, Zhang L, Young SD. A literature review: website design and user engagement. Online J Commun Media Technol 2016;6(3):1-14 [FREE Full text] [Medline: 27499833]
9. Marshall BDL, Yedinak JL, Goyer J, Green TC, Koziol JA, Alexander-Scott N. Development of a statewide, publicly accessible drug overdose surveillance and information system. Am J Public Health 2017;107(11):1760-1763. [doi: 10.2105/AJPH.2017.304007] [Medline: 28933938]
10. Dasgupta N, Kapadia F. The future of the public health data dashboard. Am J Public Health 2022;112(6):886-888 [FREE Full text] [doi: 10.2105/AJPH.2022.306871] [Medline: 35613427]
11. PreventOverdoseRI. URL: https://preventoverdoseri.org/ [accessed 2024-01-19]
12. Lasher L, Hallowell BD, Chambers LC, Koziol J, McDonald J, Elmaleh R, et al. Using timely overdose data to address a spike in nonfatal overdoses and inform a coordinated community-level response in Rhode Island, 2019. Public Health Rep 2021;136(Suppl 1):24S-30S [FREE Full text] [doi: 10.1177/00333549211012407] [Medline: 34726976]
13. Baur C, Prue C. The CDC clear communication index is a new evidence-based tool to prepare and review health information. Health Promot Pract 2014;15(5):629-637. [doi: 10.1177/1524839914538969] [Medline: 24951489]
14. Waye KM, Yedinak JL, Koziol J, Marshall BDL. Action-focused, plain language communication for overdose prevention: a qualitative analysis of Rhode Island's overdose surveillance and information dashboard. Int J Drug Policy 2018;62:86-93. [doi: 10.1016/j.drugpo.2018.08.010] [Medline: 30384027]
15. Tableau server. Tableau. URL: https://www.tableau.com/products/server [accessed 2023-12-04]
16. Documentation version 6.3. Wordpress. 2023. URL: https://wordpress.org/documentation/wordpress-version/version-6-3/ [accessed 2023-12-04]
17. Google Analytics. Google for Developers. URL: https://developers.google.com/analytics [accessed 2023-12-04]
18. Yedinak JL, Li Y, Krieger MS, Howe K, Ndoye CD, Lee H, et al. Machine learning takes a village: assessing neighbourhood-level vulnerability for an overdose and infectious disease outbreak. Int J Drug Policy 2021;96:103395 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103395] [Medline: 34344539]
19. Yedinak JL, Krieger MS, Li Y, Nelson E, Shihipar A, Goldman J, et al. The VILLAGE prevention plan: improving prevention efforts and health outcomes for people who use drugs. Brown University School of Public Health. 2019. URL: https://preventoverdoseri.org/wp-content/uploads/2020/01/VILLAGE-Prevention-Plan.pdf [accessed 2024-01-12]
20. GIS mapping software, location intelligence and spatial analytics. ESRI. URL: https://www.esri.com/en-us/home [accessed 2023-12-04]
21. Marshall BDL, Alexander-Scott N, Yedinak JL, Hallowell BD, Goedel WC, Allen B, et al. Preventing overdose using information and data from the environment (PROVIDENT): protocol for a randomized, population-based, community intervention trial. Addiction 2022;117(4):1152-1162 [FREE Full text] [doi: 10.1111/add.15731] [Medline: 34729851]
22. The Progressive JavaScript Framework: an approachable, performant and versatile framework for building web user interfaces. Vue.js. 2014. URL: https://vuejs.org/ [accessed 2023-12-04]
23. Vega—a visualization grammar. Vega. URL: https://vega.github.io/vega/docs/ [accessed 2023-12-04]
24. Cloud Firestore. Firebase. URL: https://firebase.google.com/docs/firestore [accessed 2023-12-04]
25. Green LW, Ottoson JM, García C, Hiatt RA. Diffusion theory and knowledge dissemination, utilization, and integration in public health. Annu Rev Public Health 2009;30:151-174. [doi: 10.1146/annurev.publhealth.031308.100049] [Medline: 19705558]
26. Lnenicka M, Nikiforova A. Transparency-by-design: what is the role of open data portals? Telemat Inform 2021;61:101605 [FREE Full text] [doi: 10.1016/j.tele.2021.101605]
27. Ponce NA, Shimkhada R, Adkins-Jackson PB. Making communities more visible: equity-centered data to achieve health equity. Milbank Q 2023;101(S1):302-332 [FREE Full text] [doi: 10.1111/1468-0009.12605] [Medline: 37096622]
28. Krieger N. Structural racism, health inequities, and the two-edged sword of data: structural problems require structural solutions. Front Public Health 2021;9:655447 [FREE Full text] [doi: 10.3389/fpubh.2021.655447] [Medline: 33937178]
29. Schulze A, Brand F, Geppert J, Böl GF. Digital dashboards visualizing public health data: a systematic review. Front Public Health 2023;11:999958 [FREE Full text] [doi: 10.3389/fpubh.2023.999958] [Medline: 37213621]
30. Muhammad M, Wallerstein N, Sussman AL, Avila M, Belone L, Duran B. Reflections on researcher identity and power: the impact of positionality on community based participatory research (CBPR) processes and outcomes. Crit Sociol (Eugene) 2015;41(7-8):1045-1063 [FREE Full text] [doi: 10.1177/0896920513516025] [Medline: 27429512]
31. Bailey ZD, Feldman JM, Bassett MT. How structural racism works—racist policies as a root cause of U.S. racial health inequities. N Engl J Med 2021;384(8):768-773 [FREE Full text] [doi: 10.1056/NEJMms2025396] [Medline: 33326717]
32. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. Lancet 2017;389(10077):1453-1463. [doi: 10.1016/S0140-6736(17)30569-X] [Medline: 28402827]
33. Feagin J, Bennefield Z. Systemic racism and U.S. health care. Soc Sci Med 2014;103:7-14. [doi: 10.1016/j.socscimed.2013.09.006] [Medline: 24507906]

XSL•FO
RenderX

34. Paradies Y, Ben J, Denson N, Elias A, Priest N, Pieterse A, et al. Racism as a determinant of health: a systematic review and meta-analysis. PLoS One 2015;10(9):e0138511 [FREE Full text] [doi: 10.1371/journal.pone.0138511] [Medline: 26398658]
35. Griffith DM, Bergner EM, Fair AS, Wilkins CH. Using mistrust, distrust, and low trust precisely in medical care and medical research advances health equity. Am J Prev Med 2021;60(3):442-445 [FREE Full text] [doi: 10.1016/j.amepre.2020.08.019] [Medline: 33208267]
36. Pamplin JR, Rouhani S, Davis CS, King C, Townsend TN. Persistent criminalization and structural racism in US drug policy: the case of overdose good Samaritan laws. Am J Public Health 2023;113(S1):S43-S48 [FREE Full text] [doi: 10.2105/AJPH.2022.307037] [Medline: 36696623]
37. Khan MR, Hoff L, Elliott L, Scheidell JD, Pamplin JR, Townsend TN, et al. Racial/ethnic disparities in opioid overdose prevention: comparison of the naloxone care cascade in White, Latinx, and Black people who use opioids in New York City. Harm Reduct J 2023;20(1):24 [FREE Full text] [doi: 10.1186/s12954-023-00736-7] [Medline: 36841763]
38. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
39. Özdemir B. SMOG readability. Character Calculator (TÜRKPATENT/056777). URL: https://charactercalculator.com/smog-readability/ [accessed 2023-07-21]
40. Boyd RW, Lindo EG, Weeks LD, McLemore MR. On racism: a new standard for publishing on racial health inequities. Health Affairs Forefront. URL: http://tinyurl.com/5bn9w23t [accessed 2024-01-12]
41. Neighbors HW, Mattingly DT, Johnson J, Morse K. The contribution of research to racial health equity? blame and responsibility in navigating the status quo of anti-black systemic racism. Soc Sci Med 2023;316:115209. [doi: 10.1016/j.socscimed.2022.115209] [Medline: 35927144]
42. Nelson RK, Winling L, Marciano R, Connolly N. Mapping inequality: redlining in new deal America. Digital Scholarship Lab. URL: https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58 [accessed 2023-06-09]
43. Jaiswal J, Halkitis PN. Towards a more inclusive and dynamic understanding of medical mistrust informed by science. Behav Med 2019;45(2):79-85 [FREE Full text] [doi: 10.1080/08964289.2019.1619511] [Medline: 31343962]
44. Adkins-Jackson PB, Chantarat T, Bailey ZD, Ponce NA. Measuring structural racism: a guide for epidemiologists and other health researchers. Am J Epidemiol 2022;191(4):539-547 [FREE Full text] [doi: 10.1093/aje/kwab239] [Medline: 34564723]
45. Stieb DM, Huang A, Hocking R, Crouse DL, Osornio-Vargas AR, Villeneuve PJ. Using maps to communicate environmental exposures and health risks: review and best-practice recommendations. Environ Res 2019;176:108518 [FREE Full text] [doi: 10.1016/j.envres.2019.05.049] [Medline: 31202044]
46. Golbeck AL, Ahlers-Schmidt CR, Paschal AM, Dismuke SE. A definition and operational framework for health numeracy. Am J Prev Med 2005;29(4):375-376. [doi: 10.1016/j.amepre.2005.06.012] [Medline: 16242604]
47. Carroll LN, Au AP, Detwiler LT, Fu T, Painter IS, Abernethy NF. Visualization and analytics tools for infectious disease epidemiology: a systematic review. J Biomed Inform 2014;51:287-298 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.006] [Medline: 24747356]
48. Concannon D, Herbst K, Manley E. Developing a data dashboard framework for population health surveillance: widening access to clinical trial findings. JMIR Form Res 2019;3(2):e11342 [FREE Full text] [doi: 10.2196/11342] [Medline: 30946016]
49. Jones CE, Haklay M, Griffiths S, Vaughan L. A less-is-more approach to geovisualization – enhancing knowledge construction across multidisciplinary teams. Int J Geogr Inf Sci 2009;23(8):1077-1093 [FREE Full text] [doi: 10.1080/13658810802705723]
50. Krackov A, Martin EG. Visualizing the opioid epidemic: lessons learned from a dozen years of engaging communities with data. J Public Health Manag Pract 2019;25(4):398-401. [doi: 10.1097/PHH.0000000000001037] [Medline: 31136514]
51. Cheng CKY, Ip DKM, Cowling BJ, Ho LM, Leung GM, Lau EHY. Digital dashboard design using multiple data streams for disease surveillance with influenza surveillance as an example. J Med Internet Res 2011;13(4):e85 [FREE Full text] [doi: 10.2196/jmir.1658] [Medline: 22001082]
52. Wu E, Villani J, Davis A, Fareed N, Harris DR, Huerta TR, et al. Community dashboards to support data-informed decision-making in the HEALing communities study. Drug Alcohol Depend 2020;217:108331 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108331] [Medline: 33070058]
53. Institute of Medicine (US) Committee on Health Literacy, Nielsen-Bohlman L, Panzer AM, Kindig DA. Health Literacy: A Prescription to End Confusion. Washington, DC: National Academies Press (US); 2004.
54. Smith CN, Gorczynski P, Thomas JD. The ever-evolving nature of health literacy in organizations: a commentary on the 2021 JPHMP article "updating health literacy for healthy people 2030". J Public Health Manag Pract 2022;28(6):E804-E807 [FREE Full text] [doi: 10.1097/phh.0000000000001589]
55. Santana S, Brach C, Harris L, Ochiai E, Blakey C, Bevington F, et al. Updating health literacy for healthy people 2030: defining its importance for a new decade in public health. J Public Health Manag Pract 2021;27(Suppl 6):S258-S264 [FREE Full text] [doi: 10.1097/PHH.0000000000001324] [Medline: 33729194]

XSL•FO
RenderX

56. Liu C, Wang D, Liu C, Jiang J, Wang X, Chen H, et al. What is the meaning of health literacy? a systematic review and qualitative synthesis. Fam Med Community Health 2020;8(2):e000351 [FREE Full text] [doi: 10.1136/fmch-2020-000351] [Medline: 32414834]
57. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019;29(Suppl 3):13-17 [FREE Full text] [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
58. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Viera A, Crotty K, et al. Health literacy interventions and outcomes: an updated systematic review. Evid Rep Technol Assess (Full Rep) 2011(199):1-941. [Medline: 23126607]
59. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050] [Medline: 18952344]
60. Cashman SB, Adeky S, Allen AJ, Corburn J, Israel BA, Montaño J, et al. The power and the promise: working with communities to analyze data, interpret findings, and get to outcomes. Am J Public Health 2008;98(8):1407-1417 [FREE Full text] [doi: 10.2105/AJPH.2007.113571] [Medline: 18556617]
61. Bowman S, Engelman A, Koziol J, Mahoney L, Maxwell C, McKenzie M. The Rhode Island community responds to opioid overdose deaths. R I Med J (2013) 2014;97(10):34-37. [Medline: 25271658]
62. Dayton L, Tobin K, Falade-Nwulia O, Davey-Rothwell M, Al-Tayyib A, Saleem H, et al. Racial disparities in overdose prevention among people who inject drugs. J Urban Health 2020;97(6):823-830 [FREE Full text] [doi: 10.1007/s11524-020-00439-5] [Medline: 32514829]
63. O'Flahavan L. The bite, the snack and the meal. Inc. 2001. URL: https://www.inc.com/articles/2001/06/23143.html [accessed 2023-06-09]
64. Centering Diversity, Equity, Inclusion, and Justice (DEIJ) in integrated data. State of Rhode Island, The Executive Office of Health and Human Services. URL: https://eohhs.ri.gov/initiatives/data-ecosystem/centering-racial-justice-integrated-data [accessed 2023-12-10]

Abbreviations

CDC: Centers for Disease Control and Prevention
PORI: PreventOverdoseRI
PROVIDENT: Preventing Overdose Using Information and Data From the Environment
Task Force: Governor's Overdose Prevention and Intervention Task Force
VILLAGE: Vulnerability Investigation of Underlying Local Risk and Geographic Events

Edited by G Greco; submitted 08.08.23; peer-reviewed by E Barbazza, C Baur; comments to author 24.11.23; revised version received 12.12.23; accepted 09.01.24; published 12.02.24.

Please cite as:
Yedinak J, Krieger MS, Joseph R, Levin S, Edwards S, Bailer DA, Goyer J, Daley Ndoye C, Schultz C, Koziol J, Elmaleh R, Hallowell BD, Hampson T, Duong E, Shihipar A, Goedel WC, Marshall BDL
Public Health Dashboards in Overdose Prevention: The Rhode Island Approach to Public Health Data Literacy, Partnerships, and Action
J Med Internet Res 2024;26:e51671
URL: https://www.jmir.org/2024/1/e51671
doi:10.2196/51671
PMID:38345849

©Jesse Yedinak, Maxwell S Krieger, Raynald Joseph, Stacey Levin, Sarah Edwards, Dennis A Bailer, Jonathan Goyer, Colleen Daley Ndoye, Cathy Schultz, Jennifer Koziol, Rachael Elmaleh, Benjamin D Hallowell, Todd Hampson, Ellen Duong, Abdullah Shihipar, William C Goedel, Brandon DL Marshall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Extended Reality—New Opportunity for People With Disability? Practical and Ethical Considerations

Karen Stendal[1*], PhD; Rosemarie D L C Bernabe[2,3*], PhD

[1]Department of Business, Marketing and Law, University of South-Eastern Norway, Honefoss, Norway
[2]Centre for Medical Ethics, University of Oslo, Oslo, Norway
[3]Department of Optometry, Radiography and Lighting Design, University of South-Eastern Norway, Kongsberg, Norway
[*]all authors contributed equally

Corresponding Author:
Karen Stendal, PhD
Department of Business, Marketing and Law
University of South-Eastern Norway
Bredalsveien 14
Honefoss, 3502
Norway
Phone: 47 31009477
Email: karen.stendal@usn.no

Abstract

Since the introduction of virtual environments in the 70s, technologies have moved through virtual reality, mixed reality, and augmented reality into extended reality (XR). This development is promising for various groups. Previous research has shown people with disability benefiting from using technology in social and professional settings. Technology has offered people with disability the opportunity to communicate, interact, participate, and build new relationships. However, we do not know what impact XR has or will have and whether it will offer new opportunities for people with disability. This paper aims to indicate potential opportunities and challenges afforded by XR to people with disability. We offer reflections on the opportunities as well as the ethical considerations needed when introducing immersive technologies to a marginalized group.

(J Med Internet Res 2024;26:e41670) doi:10.2196/41670

KEYWORDS

extended reality; virtual worlds; virtual reality; disability; practical; ethical; technology; virtual; reality; development; research; challenges

Introduction

We live in a society where technology in many ways covers our needs, which may also help people with disability in various ways [1]. Extended reality (XR) is moving into various areas of our lives and in areas of research such as rehabilitation [2], education [3], and tourism [4]. What impact will this technology have on people with disability? We look into XR head-mounted displays, which can become mainstream.

What is XR?

Various technologies used for social interaction have been in rapid development. Virtual worlds have, in countless versions, existed since the late 1970s [5-7]. Today, we see XR as the latest example of this technology.

XR is an umbrella term used to describe technologies such as virtual reality, augmented reality, and mixed reality [8].

XR has been described as technology that blurs the boundary between the physical and virtual worlds and enables users to experience a sense of immersion [8]. The term immersive refers to “the degree to which a virtual environment submerges the perceptual system of the user in computer-generated stimuli” [9]. Being immersive means that individuals wearing head-mounted technology will use several senses, such as vision and hearing. This way, the physical world will be “shut out” and the individual will be totally immersed in XR. This gives a sense of presence, of “being there,” in a way not experienced through virtual technology on a screen [10].

Research exploring XR for people with disability is of current interest since previous research found virtual worlds to be promising for people with disability [5,11-13]. Previous research has also included virtual reality for people with disability;

XSL•FO
RenderX

however, it has in many cases involved virtual reality on a screen from a distance [14,15], where the user sees themselves in the virtual environment. The question continues to be whether XR offers the same or new opportunities for people with disability, as called for by previous literature reviews [16,17].

New Opportunities Come With Practical and Ethical Challenges

People with disability experience various challenges, such as low self-esteem, a lack of support, stigma or discrimination, and a low degree of employment [18], not to mention the persistent exclusion or distortion of the disability narrative and experience in technological development [19]. Being treated as equals and having the same access to social and professional arenas as nondisabled peers are important to minimize or eliminate these challenges. Previous research has introduced virtual worlds as the technology to ensure this [20], and XR is the natural next step.

XR promises to empower people with disability by changing the way we learn, accomplish tasks, and interact with each other [16,21]. Being able to partake in activities not possible in the physical world affords a new opportunity for participation [21] through XR. The technology offers the possibility to visit physical world locations through a headset and ensures a fully immersive experience of the visiting location. Adding haptic technology—a feedback technology—enables a sense of touch by applying forces, vibrations, and motions upon the user [22], which further enhances the immersive experience.

The promises of this technology are abundant. However, there are challenges for people with disability, such as the effect of sensory stimuli, the comfort of wearing the technology, the visual and other physiological side effects of the technology [23], or the emotional outcomes of using XR. Part of being immersed in the technology involves a high level of sensory stimuli [8]. For instance, using a head-mounted display may lead to feeling motion sickness [8]. These results are based on users without disabilities, however, and we have little knowledge about how this technology affects people with disability when they want to use the technology [16]. In addition, there is little knowledge about how vision impairment affects and may further be affected by the use of XR [23]. This technology is made for young adults with perfect vision, which may be a challenge for many people with disability [23]. Thus, there is a need to explore who can take advantage of the technology based on a vision perspective.

Further, previous research has explored the emotional outcomes using XR [8]; however, this is an underexplored area for individuals with disabilities. There are positive emotional outcomes, such as pleasure, fun, happiness, confidence, and hope. In addition, negative emotional outcomes were identified, such as boredom, anxiety, depression, tension, fear, anger, and rage [8]. How these emotional outcomes will affect people with, for example, an intellectual disability should be explored further. XR, similar to other technologies that display avatars, does not sufficiently display body language, such as facial expressions, which makes situations more difficult to understand. Previous research has also stated that people with Asperger spectrum disorder are unable to transfer skills learned in the virtual world to the physical world [24]. Clearly, more research is also needed in this area.

As with any powerful technology, XR has its risks, several of which are relevant for users with disability. For example, XR technology can be highly immersive such that (prolonged) use can lead to social dissociation [25]. This means a user can be so emotionally involved in XR leading to detachment from one's physical reality–based social relationships. There is also the related risk of addiction [26]. The propensity of individuals with disability to this risk is yet to be explored. At the same time, we have very good reasons to believe that individuals with, for example, specific types of psychological disabilities may be more vulnerable to this risk than others. On top of the question of how they are vulnerable, there is the need to explore what technological safeguards must be put in place to lessen this risk.

Still within the topic of risks is the growing concern for deep behavioral manipulation, deep fakes, and malicious designs in XR. XR's power lies in its capacity to separate one's sensory perceptions from the physical world, effectively creating virtual, mixed, or augmented reality [27,28]. When this power is coupled with malicious intent, then the use of XR runs the risk of deep behavioral manipulation, deep fakes, and malicious designs. Especially in the case of individuals with psychological disabilities, but also for other users who may be considered relatively more vulnerable to such malice, the bigger question of what regulatory and technical safeguards must be in place is yet to be answered [25,26,29]. In addition, current metaverses for XR do not consider inclusion and accessibility for people with disability, which also should be considered for future research [30].

By design, XR favors the young White male population without physical or psychological disabilities [31]. This raises the question of how XR can be adapted for universal use, that is, to include as many individuals with disabilities as possible as potential users. This could mean several things such as redesigning hardware and software using the principles of universal design and human-centered design [32] while also considering the special needs of this population for the protection of privacy autonomy, and against the misuse of personal data.

XR technologies bring with them a lot of promises and potential opportunities for individuals with disability. But, as with other potent technologies, some challenges and risks require more research and regulatory attention. There is a need to have a closer look at these emerging technologies, ideally with and by people with disability [19]. There is a tradition of exploring new technology through able-bodied users and making assumptions about the abilities needed to use the technology. There is a need to know the affordances offered by XR to people with disability. Affordances help us understand the technological and human abilities needed to use the technology [21,33] and give us a direction to ensure that more people can take advantage of the technology. In addition, more research is needed on how

XSL•FO
RenderX

XR can be used as a support tool in the treatment of people with intellectual and multiple disabilities [34].

Conclusion

For those who are able, XR may offer many new opportunities. There is, however, a need to understand the limitations, risks, and challenges of the technology. Compared to virtual worlds on a screen, XR may affect multiple senses for the individual, and this may have implications for people with disability. As discussed earlier, they may experience overstimulation, sound issues, or motion sickness or be subjected to risks such as addiction, unintended visual and other physiological effects, malicious design, and privacy issues. Although positive or negative experiences are not exclusively related to people with disability, the potential benefit from XR for people with disability could be significant. By conducting research involving people with disability, we ensure a nuanced picture of the abilities needed to use the technology, without making assumptions.

Future research should explore how people with disability experience XR for several reasons. This technology may empower and enable people with disability to connect increasingly to society and have access to learning environments and social arenas. Through exploring the affordances offered by XR, we may gain knowledge about what facilitating conditions will allow people with disability to take advantage of innovative and new technology.

Conflicts of Interest

None declared.

References

1. Raja DS. Bridging the disability divide through digital technologies. World Bank. 2016. URL: https://thedocs.worldbank.org/en/doc/123481461249337484-0050022016/original/WDR16BPBridgingtheDisabilityDividethroughDigitalTechnologyRAJA.pdf [accessed 2024-01-24]
2. Keshner EA, Kenyon RV. Using immersive technology for postural research and rehabilitation. Assist Technol 2004;16(1):54-62. [doi: 10.1080/10400435.2004.10132074] [Medline: 15357148]
3. Moreira P, de Oliveira EC, Tori R. Impact of immersive technology applied in computer graphics learning. 2016 Presented at: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE 2016); November, 2016; Uberlândia, Minas Gerais, Brazil p. 410-419. [doi: 10.5753/cbie.sbie.2016.410]
4. Yapp A, Kim LC, Kuan GH, Zainudin H. The application of immersive technology, virtual reality in electronic tourism. Int J Adv Res Technol Innov 2019;1(1):8-13 [FREE Full text]
5. Stendal K, Molka-Danielsen J, Munkvold BE, Balandin S. Social affordances for people with lifelong disability through using virtual worlds. 2020 Presented at: 2013 46th Hawaii International Conference on System Sciences; January 07-10, 2013; Wailea, HI, USA p. 873-882.
6. Sanchez J. A social history of virtual worlds. Libr Technol Rep 2009;45(2):9-12.
7. Davies G. Preface. In: Molka-Danielsen J, Deutschmann M, editors. Learning and Teaching in the Virtual World of Second Life. Trodheim: Tapir Academic Press; 2009:3-6.
8. Suh A, Prophet J. The state of immersive technology research: a literature analysis. Comput Hum Behav 2018;86:77-90. [doi: 10.1016/j.chb.2018.04.019]
9. Biocca F, Levy MR. Communication in the Age of Virtual Reality. Hillsdale, NJ: L. Erlbaum Associates; 1995.
10. Cummings JJ, Bailenson JN. How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. Media Psychol 2016;19(2):272-309. [doi: 10.1080/15213269.2015.1015740]
11. Traphagan JWP, Sanchez KBA. Second life and access for the disabled and homebound. Care Manag J 2014;15(3):147-151. [doi: 10.1891/1521-0987.15.3.147]
12. Stewart S, Hansen TS, Carey TA. Opportunities for people with disabilities in the virtual world of second life. Rehabil Nurs 2010;35(6):254-259. [doi: 10.1002/j.2048-7940.2010.tb00056.x] [Medline: 21140720]
13. Standen PJ, Brown DJ. Virtual reality and its role in removing the barriers that turn cognitive impairments into intellectual disability. Virtual Real 2006;10(3-4):241-252. [doi: 10.1007/s10055-006-0042-6]
14. Weiss PLT, Bialik P, Kizony R. Virtual reality provides leisure time opportunities for young adults with physical and intellectual disabilities. Cyberpsychol Behav 2003;6(3):335-342. [doi: 10.1089/109493103322011650] [Medline: 12855092]
15. Yalon-Chamovitz S, Weiss PLT. Virtual reality as a leisure activity for young adults with physical and intellectual disabilities. Res Dev Disabil 2008;29(3):273-287. [doi: 10.1016/j.ridd.2007.05.004] [Medline: 17590313]
16. Maran PL, Daniëls R, Slegers K. The use of extended reality (XR) for people with moderate to severe intellectual disabilities (ID): a scoping review. Technol Disabil 2022;34(2):53-67 [FREE Full text] [doi: 10.3233/tad-210363]
17. Yenioglu BY, Ergulec F, Yenioglu S. Augmented reality for learning in special education: a systematic literature review. Interact Learn Environ 2023;31(7):4572-4588. [doi: 10.1080/10494820.2021.1976802]
18. Lindsay S, Cagliostro E, Carafa G. A systematic review of workplace disclosure and accommodation requests among youth and young adults with disabilities. Disabil Rehabil 2018;40(25):2971-2986. [doi: 10.1080/09638288.2017.1363824] [Medline: 28797182]

XSL•FO
RenderX

19. Shew A. How to get a story wrong: technoableism, simulation, and cyborg resistance. Including Disabil 2022(1):13-36 [FREE Full text] [doi: 10.51357/id.vi1.169]
20. Stendal K, Balandin S, Molka-Danielsen J. Virtual worlds: a new opportunity for people with lifelong disability? J Intellect Dev Disabil 2011;36(1):80-83. [doi: 10.3109/13668250.2011.526597] [Medline: 21070118]
21. Steffen JH, Gaskin JE, Meservy TO, Jenkins JL, Wolman I. Framework of affordances for virtual reality and augmented reality. J Manag Inf Syst 2019;36(3):683-729. [doi: 10.1080/07421222.2019.1628877]
22. Handa M, Aul G, Bajaj S. Immersive technology—uses, challenges and opportunities. Int J Comput Bus Res 2012;6(2):1-11 [FREE Full text]
23. Pladere T, Svarverud E, Krumina G, Gilson SJ, Baraas RC. Inclusivity in stereoscopic XR: human vision first. Front Virtual Real 2022;3:1006021 [FREE Full text] [doi: 10.3389/frvir.2022.1006021]
24. Standen PJ, Brown DJ. Virtual reality in the rehabilitation of people with intellectual disabilities: review. Cyberpsychol Behav 2005;8(3):272-282. [doi: 10.1089/cpb.2005.8.272] [Medline: 15971976]
25. IEEE. Extended reality in A/IS. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2016. URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead/EAD1e_extended_reality.pdf [accessed 2024-01-24]
26. Madary M, Metzinger TK. Real virtuality: a code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. Front Robot AI 2016;3:3 [FREE Full text] [doi: 10.3389/frobt.2016.00003]
27. Aliman NM, Kester L. Facing immersive “post-truth” in AIVR? Philosophies 2020;5(4):45 [FREE Full text] [doi: 10.3390/philosophies5040045]
28. Slater M, Gonzalez-Liencres C, Haggard P, Vinkers C, Gregory-Clarke R, Jelley S, et al. The ethics of realism in virtual and augmented reality. Front Virtual Real 2020;1:1 [FREE Full text] [doi: 10.3389/frvir.2020.00001]
29. Snijders D, Horsman S, Kool L, Van Est R. Responsible VR: protect consumers in virtual reality. Rathenau Instituut. The Hague; 2020. URL: https://www.rathenau.nl/en/digitalisering/responsible-vr [accessed 2024-01-24]
30. Seigneur JM, Choukou MA. How should metaverse augment humans with disabilities? 2022 Presented at: AH2022: 13th Augmented Human International Conference; May 26-27, 2022; Winnipeg, MB, Canada p. 1-6. [doi: 10.1145/3532525.3532534]
31. Stanney K, Fidopiastis C, Foster L. Virtual reality is sexist: but it does not have to be. Front Robot AI 2020;7:4 [FREE Full text] [doi: 10.3389/frobt.2020.00004] [Medline: 33501173]
32. Simon-Liedtke JT, Baraas RC. The need for universal design of eXtended Reality (XR) technology in primary and secondary education: identifying opportunities, challenges, and knowledge gaps from the literature. 2022 Presented at: Virtual, Augmented and Mixed Reality: Applications in Education, Aviation and Industry: 14th International Conference, VAMR 2022, Held as Part of the 24th HCI International Conference, HCII 2022; June 26–July 1, 2022; Virtual Event p. 121-141. [doi: 10.1007/978-3-031-06015-1_9]
33. Cosentino E. Artifacts and affordances. Synthese 2019;198(S17):4007-4026. [doi: 10.1007/s11229-019-02297-4]
34. da Cunha RD, Neiva FW, de Souza da Silva RL. Virtual reality as a support tool for the treatment of people with intellectual and multiple disabilities: a systematic literature review. Revista de Informática Teórica e Aplicada 2018;25(1):67-81 [FREE Full text] [doi: 10.22456/2175-2745.77994]

Abbreviations

XR: extended reality

Edited by A Mavragani; submitted 04.08.22; peer-reviewed by V Kotrajaras, M Mosher, PH Liao, B Lok; comments to author 19.12.22; revised version received 28.06.23; accepted 29.12.23; published 13.02.24.

Please cite as:
Stendal K, Bernabe RDLC
Extended Reality—New Opportunity for People With Disability? Practical and Ethical Considerations
J Med Internet Res 2024;26:e41670
URL: https://www.jmir.org/2024/1/e41670
doi:10.2196/41670
PMID:38349731

©Karen Stendal, Rosemarie D L C Bernabe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Building Pandemic-Resilient Primary Care Systems: Lessons Learned From COVID-19

Yejin Jeong[1*], BA; Trevor Crowell[1*], BA; Anna Devon-Sand[1*], MPH; Theadora Sakata[1*], MD; Amelia Sattler[1*], MD; Shreya Shah[1*], MD; Timothy Tsai[1*], MMCI, DO; Steven Lin[1*], MD

Division of Primary Care and Population Health, Department of Medicine, Stanford University School of Medicine, Redwood City, CA, United States
[*]all authors contributed equally

Corresponding Author:
Timothy Tsai, MMCI, DO
Division of Primary Care and Population Health
Department of Medicine
Stanford University School of Medicine
585 Broadway
Suite 800
Redwood City, CA, 94063
United States
Phone: 1 4086889720
Fax: 1 4086889730
Email: timothy.tsai@stanford.edu

Abstract

On January 30, 2023, the Biden Administration announced its intention to end the existing COVID-19 public health emergency declaration. The transition to a "postpandemic" landscape presents a unique opportunity to sustain and strengthen pandemic-era changes in care delivery. With this in mind, we present 3 critical lessons learned from a primary care perspective during the COVID-19 pandemic. First, clinical workflows must support both in-person and internet-based care delivery. Second, the integration of asynchronous care delivery is critical. Third, planning for the future means planning for everyone, including those with potentially limited access to health care due to barriers in technology and communication. While these lessons are neither unique to primary care settings nor all-encompassing, they establish a grounded foundation on which to construct higher-quality, more resilient, and more equitable health systems.

(J Med Internet Res 2024;26:e47667) doi:10.2196/47667

KEYWORDS

barrier; COVID-19; digital health; implementation; internet-based care; pandemic; primary care; resilience; resilient system; telehealth; telemedicine

Overview

Alongside testing and vaccinations, telehealth was the third leg of the 3-legged stool representing the United States' primary care strategy against COVID-19. Telehealth use surged early in the pandemic as patients and providers sought ways to safely access and deliver care. Improved attitudes toward telehealth, coupled with regulatory changes to reimbursements, enabled investments in internet-based care and digital health to skyrocket [1,2]. With the announcement that the United States planned to end the public health emergency for COVID-19 in May 2023 [3], patients, providers, and health systems are entering a new chapter, and a critical question remains: How will primary care systems in the United States sustain and optimize the positive changes that resulted from pandemic response efforts while continuing to address the challenges? In this paper, we describe 3 important lessons learned from the shift to internet-based care during COVID-19 that will facilitate the implementation of high-quality, resilient primary care systems.

Lesson 1: Clinical Workflows Must Support Both In-Person and Internet-Based Care Delivery

At the onset of the COVID-19 pandemic, public health mandates abruptly shifted the basic structure of how patients interfaced with the health care system at a time when access was paramount. Necessary infection prevention measures, such as

XSL•FO
RenderX

shelter-in-place and social distancing, spurred a remarkable transformation in which new processes for telehealth were designed and implemented rapidly to meet emerging demands [4]. Going forward, clinical workflows must support both in-person and internet-based care delivery, with special attention paid to ensuring consistent preventive care and chronic disease management across all care contexts. For example, traditional in-clinic rooming procedures such as preventive care reminders could be conducted asynchronously through electronic previsit questionnaires or phone calls.

Accounting for a broad spectrum of care delivery models when designing clinical workflows provides flexibility, allowing health systems to be nimble and provide high-quality care even when faced with unanticipated constraints. To accomplish this, systems must outline the essential components of care delivery and explore how to make each component accessible, taking into consideration patient factors. Systems should build pathways that efficiently connect patients with the most appropriate type of encounter (eg, clinic, telehealth, group, e-visit, e-consult, home visit, or referral), hosted by the most relevant team members (eg, physician, advance practice provider, registered nurse, clinical pharmacist, behavioral health provider, social worker, case manager, or medical assistant). Systems should recognize and address tasks that limit flexibility and convenience, such as patient forms that need to be completed on paper, scanned, and faxed. Resilient systems will use structured quality improvement methodologies to identify opportunities for innovation and the implementation of new workflows and technologies that enhance both in-person and internet-based care delivery.

Lesson 2: The Importance of Asynchronous Care Delivery Integration Within Primary Care

Another trend that has significant implications for health care delivery in primary care is the rise in asynchronous care modalities. Methods of asynchronous care include patient portals, remote patient monitoring (RPM), and mobile health (mHealth) apps. Patient portals enable patients to view and schedule appointments, view medical test results, view and request medications, and communicate with their providers through a secure platform [5]. Asynchronous care may also leverage patient-generated health data (PGHD), examples of which include health history, biometric data, and patient-reported outcome measures. PGHD can be transmitted through mHealth apps and connected health devices, including blood pressure cuffs, glucometers, and fitness trackers that use an application programming interface (API), which allows information to be sent back and forth between an app and the primary care team. Some of our most costly chronic diseases, such as hypertension, diabetes, and behavioral health, can benefit from an increased number of telehealth visits with the health care system [6]. RPM applications can be the platform that provides frequent interactions, improved access to care, and improved health equity [7].

Digital health tools such as RPM and health apps are changing how patients interact with primary care. Patients are relying more on their smartphones and devices for prevention, diagnosis, and disease management. This provides a significant opportunity, as these tools can lead to improved quality of care and cost savings [8]. Specific examples in primary care include improving hemoglobin A_{1c} control in patients with diabetes and improving blood pressure control in patients with hypertension [9]. RPM care management and transitional care programs for congestive heart failure can reduce hospitalization days and death [10]. A multidisciplinary team-based approach is an important part of a successful RPM program, with a shift from the role of a single provider-patient relationship to a care team–patient relationship. Care team members can include primary care physicians, advanced practice providers, registered nurses, clinical pharmacists, behavioral health providers, social workers, case managers, or medical assistants. For example, with hypertension, the comparative effectiveness of 8 implementation strategies found that team-based care models including a nonphysician team member who titrates medications are the most effective for blood pressure control [11]. As the volume of available tools continues to rise, it is increasingly important for care teams to determine which technologies to recommend for their patients [12]. Textbox 1 outlines these clinical, usability, and technical considerations.

XSL•FO
RenderX

Textbox 1. Key considerations for integration of asynchronous digital health tools within primary care.

Clinical

- Does the tool align with evidence-based medicine?
- Does the tool improve clinical care?
- Is the tool safe?
- How will the tool be integrated in the clinical workflow?
- Are there protocols for urgent issues or out-of-range remote patient monitoring measurements?
- Is there evidence of closed-loop communication with the patient?

Usability

- Are there training and education needs for patients and staff?
- Is the tool easy to use for patients and staff?
- Does the tool provide knowledgeable insights to patients and staff?
- Can the tool help mitigate increasing health care team burden and burnout?
- Are there any reimbursement considerations?

Technical

- What technology is used?
- What input data is required?
- If artificial intelligence is used, is it explainable?
- Is the tool accurate?
- Is the tool secure with reliable user authentication?
- Does the tool meet industry data security standards?

Lesson 3: Planning for the Future Means Planning for Everyone

A final lesson learned during the pandemic is that while a transition to digital health platforms improved access for some, it exacerbated and created new barriers for others. Patients without a stable internet connection or appropriate hardware to support a telehealth visit faced a “tech gap” that left them without any way of seeing their medical providers if they belonged to one of the many health systems that suspended in-person visits. While the underpinnings of this gap stem from several channels, ranging from affordability to infrastructural shortcomings to global supply chain delays, addressing issues of digital health access before the next pandemic will minimize the number of patients who get left behind. Already, some health care systems, such as the US Veterans Administration, have incorporated a “digital divide consult” into their cadre of services [13]. Through this program, veterans are able to access tablets and cellular data plans at no cost. Though not every system will be able to intervene at this level, including a baseline technological assessment as part of a social history might at least alert providers to which patients will need additional outreach.

Language and communication barriers that already created inequities in health care before the pandemic played an even larger role during the pandemic [14]. As procedures for accessing care evolved rapidly, many clinics alerted patients to new scheduling and visit requirements through audio recordings, websites, and signs on clinic doors. More often than not, these notices were provided in only English, which presented difficulties for patients with limited English proficiency to access and use telehealth services. Additionally, as visits moved into the internet-based space, some providers had limited access to video platforms able to support more than 2 parties, thus barring the ability to use third-party interpretation services to communicate with non–English-speaking or hard-of-hearing patients. In order to optimize our current delivery system, we must not only be able to communicate changes to our non–English-speaking patients or patients with disabilities in real-time, but we must also ensure that interpretation is integrated into the digital platform with additional support before, during, and after their internet-based visits. As efforts are being made to address these barriers, there should be a concurrent effort to gather quality metrics to measure the impact of these efforts.

Conclusion

While these lessons are neither unique to primary care nor all-encompassing, they generate insights into the work set before us. A combination of in-person and internet-based care, adaptable workflows with integrated asynchronous patient care, and an awareness of threats to health equity could address existing gaps and provide opportunities for building better systems. As the public health emergency for COVID-19 ends, let us not waste these valuable lessons in creating a future of

XSL•FO
RenderX

primary care built to withstand future pandemics and other crises yet to come.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Conflicts of Interest

None declared.

References

1. Malouff TD, TerKonda SP, Knight D, Dabrh AMA, Perlman AI, Munipalli B, et al. Physician satisfaction with telemedicine during the COVID-19 pandemic: the Mayo Clinic Florida experience. Mayo Clin Proc Innov Qual Outcomes 2021;5(4):771-782 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2021.06.006] [Medline: 34226884]
2. Weigel G, Ramaswamy A, Sobel L, Salganicoff A, Cubanski J, Freed M. Opportunities and barriers for telemedicine in the U.S. during the COVID-19 emergency and beyond. KFF. 2020. URL: https://www.kff.org/womens-health-policy/issue-brief/opportunities-and-barriers-for-telemedicine-in-the-u-s-during-the-covid-19-emergency-and-beyond/ [accessed 2023-03-09]
3. LaFraniere S, Weiland N. U.S. plans to end public health emergency for Covid in May. The New York Times. 2023. URL: https://www.nytimes.com/2023/01/30/us/politics/biden-covid-public-health-emergency.html [accessed 2023-03-09]
4. Choi K, Gitelman Y, Leri D, Deleener ME, Hahn L, O'Malley C, et al. Insourcing and scaling a telemedicine solution in under 2 weeks: lessons for the digital transformation of health care. Healthc (Amst) 2021;9(3):100568 [FREE Full text] [doi: 10.1016/j.hjdsi.2021.100568] [Medline: 34293616]
5. Avdagovska M, Menon D, Stafinski T. Capturing the impact of patient portals based on the quadruple aim and benefits evaluation frameworks: scoping review. J Med Internet Res 2020;22(12):e24568 [FREE Full text] [doi: 10.2196/24568] [Medline: 33289677]
6. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002;288(14):1775-1779 [FREE Full text] [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
7. Omboni S, McManus RJ, Bosworth HB, Chappell LC, Green BB, Kario K, et al. Evidence and recommendations on the use of telemedicine for the management of arterial hypertension: an international expert position paper. Hypertension 2020;76(5):1368-1383 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.120.15873] [Medline: 32921195]
8. Iribarren SJ, Cato K, Falzon L, Stone PW. What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. PLoS One 2017;12(2):e0170581 [FREE Full text] [doi: 10.1371/journal.pone.0170581] [Medline: 28152012]
9. Salehi S, Olyaeemanesh A, Mobinizadeh M, Nasli-Esfahani E, Riazi H. Assessment of Remote Patient Monitoring (RPM) systems for patients with type 2 diabetes: a systematic review and meta-analysis. J Diabetes Metab Disord 2020;19(1):115-127 [FREE Full text] [doi: 10.1007/s40200-019-00482-3] [Medline: 32550161]
10. Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Kirwan BA, et al. Efficacy of Telemedical Interventional Management in patients with Heart Failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. Lancet 2018;392(10152):1047-1057. [doi: 10.1016/S0140-6736(18)31880-4] [Medline: 30153985]
11. Mills KT, Obst KM, Shen W, Molina S, Zhang HJ, He H, et al. Comparative effectiveness of implementation strategies for blood pressure control in hypertensive patients: a systematic review and meta-analysis. Ann Intern Med 2018;168(2):110-120 [FREE Full text] [doi: 10.7326/M17-1805] [Medline: 29277852]
12. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3(1):14 [FREE Full text] [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
13. Heyworth L, Kirsh S, Zulman D, Ferguson JM, Kizer KW. Expanding access through virtual care: the va's early experience with COVID-19. NEJM Catal Innov Care Deliv 2020;1(4) [FREE Full text]
14. Diamond LC, Jacobs EA, Karliner L. Providing equitable care to patients with limited dominant language proficiency amid the COVID-19 pandemic. Patient Educ Couns 2020;103(8):1451-1452 [FREE Full text] [doi: 10.1016/j.pec.2020.05.028] [Medline: 32571503]

Abbreviations

API: application programming interface
mHealth: mobile health
PGHD: patient-generated health data
RPM: remote patient monitoring

XSL•FO
RenderX

Edited by A Mavragani; submitted 30.03.23; peer-reviewed by B Green, D Alfarra; comments to author 17.01.24; revised version received 25.01.24; accepted 29.01.24; published 23.02.24.

Please cite as:
Jeong Y, Crowell T, Devon-Sand A, Sakata T, Sattler A, Shah S, Tsai T, Lin S
Building Pandemic-Resilient Primary Care Systems: Lessons Learned From COVID-19
J Med Internet Res 2024;26:e47667
URL: https://www.jmir.org/2024/1/e47667
doi:10.2196/47667
PMID:38393776

©Yejin Jeong, Trevor Crowell, Anna Devon-Sand, Theadora Sakata, Amelia Sattler, Shreya Shah, Timothy Tsai, Steven Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Improvement of Cancer Prevention and Control: Reflection on the Role of Emerging Information Technologies

Pei Dong[1], MSc; Ayan Mao[1], MPH, MA; Wuqi Qiu[1], PhD; Guanglin Li[2], PhD

[1]Department of Public Health Strategy Research, Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Chinese Preventive Medicine Association, Beijing, China

Corresponding Author:
Ayan Mao, MPH, MA
Department of Public Health Strategy Research
Institute of Medical Information
Chinese Academy of Medical Sciences and Peking Union Medical College
Number 3 YaBao Road, Chaoyang District
Beijing, 100020
China
Phone: 86 01052328731
Email: mao.ayan@imicams.ac.cn

Abstract

Cancer has become an important public health problem affecting the health of Chinese residents, as well as residents all over the world. With the improvement of cancer prevention and treatment, the growth of the mortality rate of cancers has slowed down gradually, but the incidence rate is still increasing rapidly, and cancers still impose heavy disease and economic burdens. Cancer screening and early cancer diagnosis and treatment are important ways to reduce the burden of cancer-related diseases. At present, various projects for early cancer diagnosis and treatment have been implemented in China. With the expansion of the coverage of these projects, the problems related to project implementation, operation, and management have emerged gradually. In recent years, emerging information technologies have been applied in the field of health and have facilitated health management and clinical decision-making. Meanwhile, China announced multiple policies to encourage and promote the application of information technologies in the field of health. Therefore, combined with the analysis of major problems in cancer prevention and control projects, this paper probes into how to apply information technologies such as biological information mining, artificial intelligence, and electronic information collection technology to various stages of cancer prevention and control. Information technologies realize the integrated management of prevention and control processes, for example, mobilization and preliminary identification, high-risk assessment, clinical screening, clinical diagnosis and treatment, tracking and follow-up, and biological sample management of high-risk groups, and promote the efficient implementation of cancer prevention and control projects in China.

(J Med Internet Res 2024;26:e50000) doi:10.2196/50000

KEYWORDS

emerging information technologies; cancer; prevention and control

Background

Cancer has become the leading cause of death in many countries in the world [1]. It is an important public health problem affecting the health of people all over the world, including Chinese people. In recent years, with the improvement of cancer prevention and treatment, the growth of the mortality rate of cancers in China slowed down gradually. The all-sex mortality rate of cancers increased from 123.11 per 100,000 persons in 1990 to 156.41 per 100,000 persons in 2004 [2] (with an average growth rate of about 2.57%) and increased from 157.98 per 100,000 persons in 2005 to 190.66 per 100,000 persons in 2019 [2] (with an average growth rate of about 1.81%). However, the incidence rate is still increasing rapidly. The all-sex crude incidence of cancers increased from 148.84 per 100,000 persons in 1990 to 212.88 per 100,000 persons in 2004 [2] (with an average growth rate of about 3.45%) and increased from 218.10 per 100,000 persons in 2005 to 334.53 per 100,000 persons in 2019 [2] (with an average growth rate of about 3.56%; Figure 1). In China, cancers still impose a heavy disease burden. The rate of standardized disability-adjusted life years for cancers in China showed a downward trend during the years from 1990 to 2019, but it is still higher than the world average level [3]. The economic burden caused by cancers in China should not

XSL•FO
RenderX

be ignored. Some studies have pointed out that the direct economic burden caused by cancers in China in 2003 and 2010 was CN ¥32.63 billion (US $5.28 billion) and CN ¥100.74 billion (US $163.30 billion), respectively [4], accounting for 4.96% and 5.04% of the total health expenditure in that year respectively.

Figure 1. Incidence and mortality of all-sex cancers in China from 1990 to 2019.

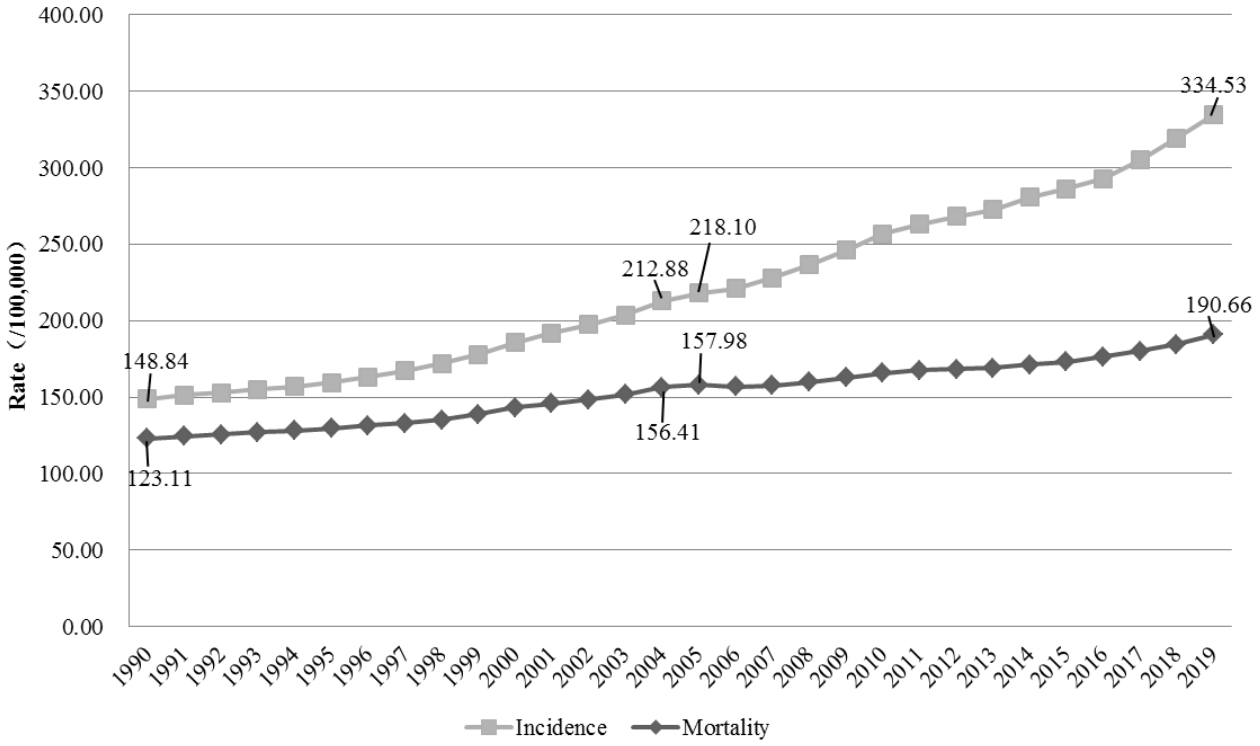

A large number of studies have confirmed that cancer screening and early cancer diagnosis and treatment are important ways to reduce the burden of cancer-related diseases [5]. To increase the rate of early cancer diagnosis and treatment, a variety of early cancer diagnosis and treatment projects have been implemented since 2005 [6], which have played a positive role in reducing the health hazards of cancers. With the expansion of the coverage of the projects; the increase in the number of participants; and the accumulation of massive evaluation, screening, and complicated data, higher requirements have been raised for the implementation and management of cancer prevention and control screening projects. In recent years, China announced multiple policies to encourage and promote the application of information technologies to the field of health [7-9]. The *National Health Plan for the 14th Five-Year Plan Period* requires that emerging information technologies, for example, artificial intelligence (AI) and big data technologies, should be popularized and applied to provide intelligent medical services and chronic disease screening services [10]. At present, most of the projects in China's cancer prevention and control working modes are still based on traditional methods (eg, collecting information by paper questionnaires and manual entry), while emerging information technologies have been more maturely applied to similar scenarios in the health field. We, therefore, aimed to reflect on the combination of emerging information technologies and cancer prevention and control projects in China, build a bridge that may persuade China's tumor prevention and control management agencies to make full use of information technology, promote the efficiency of the management, and benefit more residents and medical staff.

Main Cancer Prevention and Control Projects and Their Working Modes in China

Cancer Prevention and Control Projects in China

China's current cancer prevention and control programs are roughly divided into 3 categories, that is, national-level public welfare projects, local-level public welfare projects, and public welfare projects implemented by charity organizations.

As major specific public health services in the country, national-level cancer screening projects include *Cancer Screening Program in Rural Areas* (2005), *Cancer Screening Program in Huaihe River Area* (2007), *Cervical Cancer and Breast Cancer Screening Program for Women* (2009; included in the basic public health service in 2019), and *Cancer Screening Program in Urban Areas* (2012). These cover such high-incidence cancers as lung cancer, stomach cancer, esophageal cancer, colorectal cancer, liver cancer, breast cancer, and cervical cancer. These projects enable related people, for example, high-incidence populations and special populations in rural, urban, and high-incidence areas, to enjoy cancer screening services free of charge and thus reduce the harm of cancers on the health of residents [11-14] (Table 1).

Local-level cancer screening projects are mainly screening projects launched for local special high-incidence cancers

XSL•FO
RenderX

according to local conditions. For example, Tianjin launched the "Joint Screening for Common Cancers Project in Tianjin" in 2017, which targeted 4 local high-incidence cancers, that is, lung cancer, breast cancer, stomach cancer, and liver cancer [15]. In 2019, Hunan Province launched the "Oral Cancer Screening and Early Diagnosis and Treatment Project," which targeted oral cancer, a local high-incidence cancer [16].

In addition, many charity organizations have implemented some cancer screening projects in China. For example, the "Healthy China - Maternal and Child Health Promotion Action" Central and Western China Cervical Cancer Screening Project, which was implemented jointly by China Population Welfare Foundation and the Maternal and Child Health Center of Chinese Center for Disease Control and Prevention, was launched in March 2023 [17]. The project is intended to explore a new mode of breast cancer and cervical cancer screening services in central and western China and help build the cervical cancer prevention and treatment capacity in central and western China. The "Healthcare Aid Project of Early Gastrointestinal Cancer Screening," initiated jointly by Tencent Foundation, Guangdong Foundation for Poverty Alleviation, and Guangzhou United for Good Development Center, was launched on June 15, 2019 [18]. It provides early cancer screening and treatment assistance to poverty-stricken households and families in key poverty relief and development areas by offering medical AI technical support to some primary hospitals.

Table 1. The 4 national cancer screening programs in China.

Program	Start year	Prevention focus	Coverage
Cancer Screening Program in Rural Areas	2005	High-risk areas of upper gastrointestinal cancer	In 263 counties in 31 provinces (by 2020)
Cancer Screening Program in Huaihe River Area	2007	High-risk population of esophageal cancer, gastric cancer, and liver cancer	By 2019, a total of 32 counties in 4 provinces (Jiangsu, Shandong, Anhui, and Henan)
Cervical Cancer and Breast Cancer Screening Program for Women	2009	Women aged 35-64 years in rural target areas, cervical cancer, and breast cancer	In total, 87% (2644 counties) of counties in China for cervical cancer and 81% (2456 counties) of counties in China for breast cancer (by 2020)
Cancer Screening Program in Urban Areas	2012	Population risk factors survey, high-risk population assessment, cancer screening, and health economic evaluation were carried out for lung, breast, colorectal, esophagus, stomach, and liver	The initial 9 provinces in 2012 and 76 cities in 30 provinces by 2022

Cancer Prevention and Control Working Modes

The working modes and steps of cancer prevention and control projects in China are largely identical but with minor differences [19,20]. They are described roughly as follows (Figure 2).

- Mobilization and preliminary identification of high-risk groups: subdistricts, communities, and local health administration departments organize health education and publicity activities together; specify the conditions for screening service receivers; and mobilize eligible people, especially high-risk groups, to actively participate in assessment and screening activities.
- High-risk assessment: eligible screening service receivers who were willing to participate in a project fill in the questionnaire of risk factors under the guidance of appointed persons. The high-risk group assessment model and its background software are used. People who are assessed as high-risk groups are informed of receiving subsequent clinical cancer screening, while those who are assessed as non–high-risk groups are advised to maintain a healthy lifestyle and undergo regular physical examinations.
- Clinical screening: people who are assessed as high-risk groups go to the designated hospital to undergo relevant clinical cancer screening at the time specified. People who tested positive in a clinical screening will be assisted to go to another medical institution for re-examination and confirmation, and they will be followed up on regularly. Those who are tested negative in clinical screening will be followed up and intervened with regularly.
- Transfer to another medical institution, examination, and confirmation: suspected patients who test positive in clinical cancer screening will be recommended and assisted to move to a medical institution with cancer diagnosis and treatment capacities for physical examination and confirmation. Confirmed patients will be recommended to receive medical treatment and be registered. The project team will visit them regularly. Unconfirmed patients will be recommended to receive further consultation and be followed up regularly.
- Cancer registration: for confirmed cancer patients, per the *Measures for the Administration of Cancer Registration* and the cancer classification specified in the *International Classification of Diseases, 10th Edition*, all the medical institutions under the jurisdiction where the cancer registry is located should include the cancer information of patients in the *Residents' Cancer Information Report Card* and *Cancer Incidence Register*. Further, they should then provide such information quarterly to the cancer registry under the jurisdiction through the information system or the registration report card. The cancer registry will file, code, verify, and analyze the data; make up the deficiency; remove repeated information; and then submit the data step-by-step to the provincial-level cancer prevention office and the National Cancer Center in the form of web-based direct reporting.

XSL•FO
RenderX

Figure 2. Cancer prevention and control working mode in China. ICD-10: International Classification of Diseases, 10th Edition.

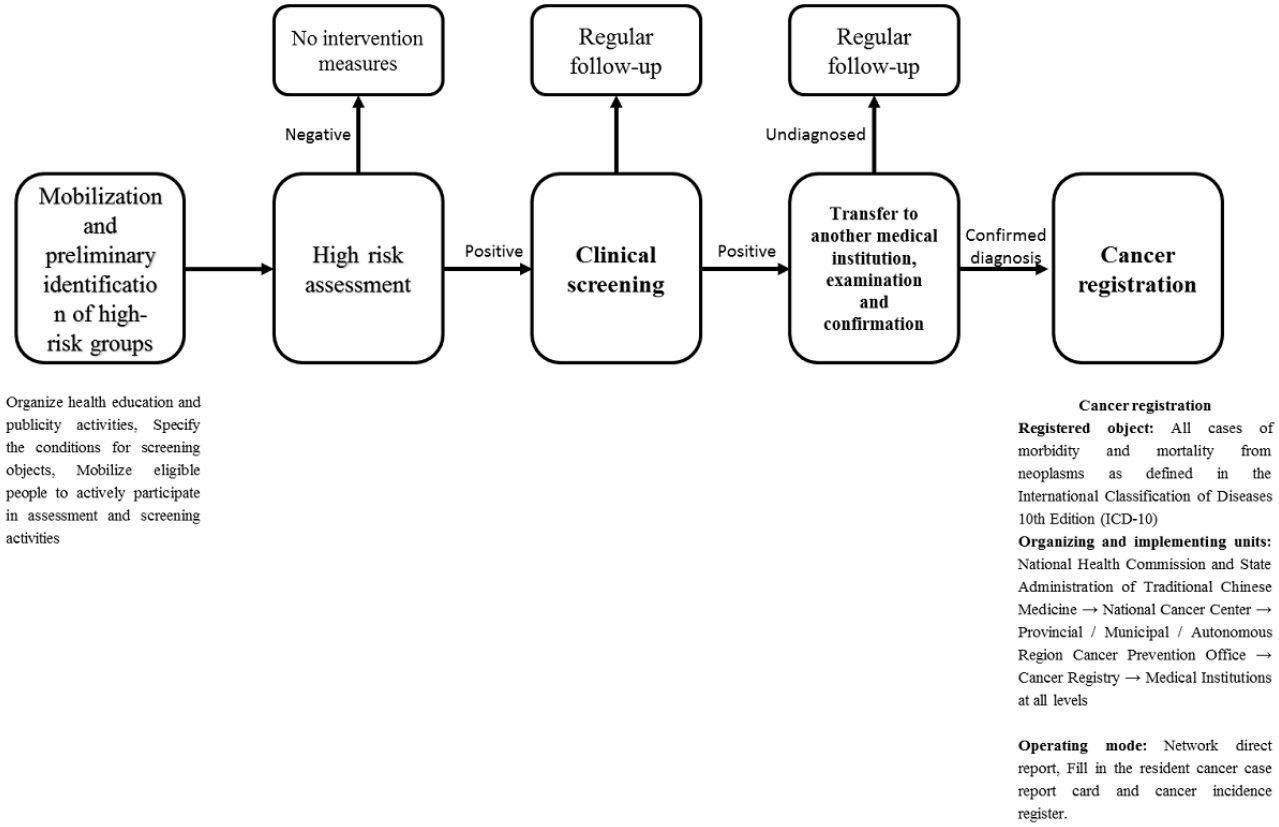

Main Problems in Cancer Prevention and Control Projects at the Present Stage and Countermeasures

Screening Coverage Is Relatively Limited

The cancer screening projects implemented in China currently mainly target common cancers and high-risk cancer areas or groups [21,22], covering limited cancers, regions, and groups. Take the urban cancer project as an example; currently, it covers 76 cities. Compared with the 293 prefecture-level cities in China, the project's coverage is very limited. In addition, employees' physical examinations—organized by some enterprises and public institutions where residents have spontaneous physical examinations and patient clinical examinations and in which cancer lesions are accidentally found—are also important ways to find cancers and precancerous lesions. However, since relevant physical examination, test, and diagnosis information has not been included in the current cancer prevention and control system, there is no way to track and manage such information in a standardized manner. For such problems, AI and other emerging information technologies may be taken into consideration to automatically capture massive health information, for example, in routine physical examination and hospital examination and diagnosis information, to accurately locate high-risk groups and patients and have them managed in a standardized manner. On this basis, cloud computing and intelligent algorithms may also be used to determine target high-risk residents for cancer screening targets based on massive population health data, to break the regional or population boundaries in traditional screening projects and make the projects cover all residents finally.

Technological Level of Screening Services Is Not Balanced, and Screening Service Capacity and Technical Reserve Are Limited

The level of community and economic development varies in different places in China. Therefore, the medical service capacity, clinical screening hardware facilities, and medical personnel's professional capacity in different regions are different. As a result, the level and quality of screening vary greatly [5]. In addition, organizers and managers fail to fully consider the screening service capabilities of screening service providers, for example, the technical level of clinical examinations and tests, the current workload, and workload expansion space. They only make a rough judgment and assign "indicator-based" tasks according to the level, functional positioning, and qualification of medical institutions. Such tasks may have an impact on daily business and work, increase the workload, and thus affect the quality of work [23,24]. How to make full use of cross-regional high-quality medical resources and technological means to assist areas with weak technology and workforce in improving their technical level of screening and service capacity through emerging information technologies, for example, telemedicine and AI-assisted imaging diagnosis, is a key issue that needs to be considered and put into practice.

Efficiency of Defining High-Risk Groups of Cancers Needs to Be Increased

Efficiently and accurately locating high-risk groups of cancers is an important basis for accurate cancer prevention and control.

XSL•FO
RenderX

Community-based questionnaire surveys are still the main channel used to obtain information on high-risk groups in cancer screening projects in China [25,26]. Since a huge population needs to be screened, such screening becomes an extra workload for community workers who are already busy with their routine work. Moreover, the screening is of huge human cost. Therefore, manual screening of high-risk groups by communities becomes a bottleneck restricting the efficient implementation of a whole screening project, and the accuracy and timeliness of collected information directly affect the efficiency and results of screening projects. At present, China has accumulated a certain scale of cancer monitoring data, cancer registration sample information, and cancer follow-up data [27], which lay a good foundation for the concentration of high-risk groups. On this basis, the use of big data technology, for example, data pipeline triggers during data ingestion or conditional logic in data analysis and visualization tools, to accurately identify high-risk groups and further locate high-risk individuals using e-survey can effectively reduce the workload of communities in high-risk group screening, improve the accuracy and effectiveness of defining high-risk groups, break through the bottleneck in current screening projects, and increase work output.

Screening Service Receivers' Insufficient Willingness to Enjoy Services Has an Impact on the Sustainability of Screening Projects

Residents' awareness of the benefits that cancer screening can bring to themselves has an important impact on their participation in and compliance with screening projects. Studies [25,28-32] have found that most Chinese residents still lack awareness of cancer risk factors, knowledge of cancer prevention and control, and consciousness to prevent cancers. Especially, they know little about the importance of cancer screening and cancer prevention and treatment strategies. Due to the pain in the screening process for some cancers, people are reluctant to accept and comply with cancer screening. For these reasons, cancer screening services are underused, which has a direct impact on the overall implementation effect of screening projects and thus affects the long-term sustainability of screening projects. For the screening projects in progress currently, improving residents' awareness of cancer risk factors, consciousness to prevent cancers, and willingness to participate in screening is the key to having more screening services used. In this demand context, efforts should be made to make use of new media and other publicity modes to push multilevel, all-round, and accurate information on cancer risk factors and screening for cancer prevention. This is to further enable more people to accept cancer screening on their own and lay a foundation for people to accept cancer screening stably and sustainably.

Envisagement on Making Use of Emerging Information Technologies to Upgrade Cancer Prevention and Control

Overview

Combined with the current situation of cancer prevention and control projects, we try to think about the application of emerging information technologies in various stages of cancer prevention and control. We describe these below (Tables 2 and 3).

Table 2. Current situation and potential solutions in various stages of cancer prevention and control.

Stage and current situation	Potential solutions
Mobilization of high-risk groups and preliminary identification of high-risk groups	
Traditional publicity methods (for example, notification of doctors and neighborhood associations, poster campaign): the regional restrictions	• Use new media platforms to shape more cancer prevention and control information dissemination channels • Use big data technologies to collect and analyze the massive user interaction information on new media platforms and push personalized cancer prevention knowledge according to users' screening demands to improve the cancer prevention and control information dissemination efficiency • Bring the multiple roles of new media users to enhance the interaction and sense of substitution of publicity and intervention
Traditional screening projects: the restrictions on regulations or populations	• Expand data sources • Use AI[a] algorithm to analyze residents' demands for cancer screening based on multisource data
Cancer risk assessment	
Collect participants' basic information, dietary habits, lifestyle and environment, psychology and emotion, and history and tumor cancer family history, through paper-based questionnaires	• Replacement of paper questionnaires with online electronic questionnaires • Use machine learning and biological information mining technologies to capture, summarize, develop, and use relevant data and information • Identify key tumor cancer biomarkers related to the carcinogenetic process in combination with methods of epidemiology, statistics, and computational biology

[a]AI: artificial intelligence.

XSL•FO
RenderX

Table 3. Current situation and potential solutions in various stages of cancer prevention and control.

Stage	Current situation	Potential solutions
Clinical screening, physical examination, and confirmation	• Clinicians interpret screening results • The shortage of doctors or screening technologies in some areas	• Help clinicians obtain image data and section information comprehensively and assist clinicians to interpret screening results with the help of the powerful biological data analysis ability and the deep machine learning ability of AI[a]-assisted diagnosis technologies • Use the clinical auxiliary cancer decision system built based on AI and other technologies to assist doctors in making a diagnosis and decision on the people who tested positive in the clinical screening of screening projects
Cancer registration	• Population-based cancer registration and hospital-based cancer registration	• Rely on big data mining, AI, and information capture technologies to extract tumor cancer diagnosis, treatment, and other medical record information, as well as follow-up visit and risk factor information, in time and build a monitoring database with more accurate and detailed risk factors, incidence, and death information • Integrating and complementing population-based and hospital-based cancer registration
Follow-up visits	• Absence of a follow-up management system for screening projects	• Build a follow-up visit management system for cancer prevention and control based on internet and AI-based machine learning technologies to realize the information interconnection with hospitals' diagnosis and treatment systems and tumor cancer registration systems

[a]AI: artificial intelligence.

Stage of Mobilization and Preliminary Identification of High-Risk Groups

At the stage of mobilization of high-risk groups, a new cancer screening publicity network ecosystem may be built based on various new media platforms by breaking through the regional restrictions on traditional publicity ways. New media platforms such as mobile communication tools (various apps), internet tools (blogs, microblogs [Sina, Tencent, and Sohu], and forums), and new television technology may be used to provide image, text, video, or audio content to publicize cancer risk factors, the identification of early cancer symptoms, and the important role of individuals in the process of cancer prevention to shape up more cancer prevention and control information dissemination channels. Big data technologies, for example, collaborative filtering, may be used to collect and analyze the massive user interaction information on new media platforms. After such information is optimized and integrated, personalized cancer prevention knowledge may be pushed according to users' screening demands to improve the cancer prevention and control information dissemination efficiency. The multiple roles of new media users as information receivers, re-editors, and publishers may be brought into play to enhance the interaction and sense of substitution of publicity and intervention. Through around-the-clock and full-coverage cancer prevention and control information dissemination by new media, residents can improve their awareness of cancer prevention and more residents will take the initiative to accept cancer screening. Accordingly, more people will enjoy cancer screening services.

At the stage of preliminary identification of high-risk groups, AI technology may be used to extract big data related to residents' health records, physical examination data, medical diagnosis information, and regional health data that are under standardized management. Cloud computing and intelligent algorithms may be used to determine residents' demands for cancer screening. In this way, the restrictions on regulations or populations in traditional screening projects may be broken through and cancer prevention and control may cover more people and regions. AI algorithms were used to automatically extract the human health monitoring data from wearable devices, health consumption data in consumer payments, netizens' demand data that are captured by internet search engines, users' installation and use data that are collected by mobile apps, and residents' physical examination data in the *Tsinghua City Health Index 2022*. The experience of the multisource social big data collection and extraction in the *Tsinghua City Health Index 2022* may be useful for the preliminary determination of high-risk groups in cancer screening projects. By analyzing residents' demands for cancer screening based on multisource data, more people's demands for cancer prevention may be covered.

Stage of Cancer Risk Assessment

At the stage of risk assessment, machine learning and biological information mining technologies may be used to capture, summarize, develop, and use relevant data and information. In combination with methods of epidemiology, statistics, and computational biology, key cancer biomarkers related to the carcinogenetic process may be identified. At the same time, based on the resident health information accumulated by various interconnected and shared platforms, a high-risk cancer group identification model may be built to achieve accurate stratification and management of high-risk groups. Such technology has been applied both at home and abroad. For example, some researchers in the United States built a prostate cancer risk prediction model by using machine learning technologies and methods, and they found that the prediction result was good after verification [33]. Some researchers used biological information data mining technologies and they found specific signals enhancing the migration and invasiveness of

XSL•FO
RenderX

nonsmall cell lung cancers and established a high-risk nonsmall cell lung cancer metastasis risk prediction model [34]. Chinese researchers built risk prediction models for the "onset" and "progress" of esophageal cancer. They found specific indicators affecting the accuracy of esophageal cancer progress risk prediction and achieved accurate and individualized esophageal cancer progress risk assessment [35].

Meanwhile, in the process of collecting data, for example, participants' basic information, dietary habits, lifestyle and environment, psychology and emotion, history, and cancer family history, the screening projects' online electronic questionnaire survey may be carried out. Thus, a variety of out-of-order data may be reasonably classified, simplified, well-organized, and standardized. Automatic sorting can help reduce workload sharply; reduce the screening workload of primary health institutions, that is, specific implementers of screening projects; decrease the service supply problems caused by insufficient service willingness resulting from excessive workload; and further lower the manpower cost for screening.

Stage of Clinical Screening, Physical Examination, and Confirmation

At the stage of clinical screening, with the help of the powerful biological data analysis ability and deep machine learning ability of AI-assisted diagnosis technologies, the massive image and body fluid sample test data accumulated by cancer screening projects as preliminary diagnosis results of screening service receivers may be analyzed, and preliminary conclusions may be provided. These help clinicians obtain image data and section information comprehensively, assist clinicians in interpreting screening results, quickly screen diseases, and make a definite diagnosis to make up for the shortage of doctors or screening technologies in some areas and increase the efficiency and effect of screening. AI-based computer-aided diagnosis technologies have been widely used in many medical fields such as imaging and pathological diagnosis of cancers or precancerous lesions. For example, in terms of pulmonary nodule diagnosis, the AI-based accurate algorithm model can quickly detect nodules in a short time and predict benign and malignant nodules [36]. As such a model is combined with film reading by doctors of the imaging department, work efficiency and diagnostic accuracy have been greatly improved. For pathological sections of lung tissue, the AI model based on deep learning algorithms may be used to assist pathologists of Tongji Hospital of Tongji University and Shanghai Pulmonary Hospital to analyze stained pathological sections of lung lesion tissue [37]. The model has been connected with 2 modules, that is, the division of benign and malignant lesion areas and the classification of pathological subtypes, which can help users more sensitively detect malignant lesions and more accurately classify pathological subtypes. Good diagnostic results have been achieved. To rapidly assess living breast tissue, Loukas et al [38] have established a highly intelligent breast cancer image classification system based on AI-based computer-aided diagnosis technologies from 2 aspects of studying nuclear mechanism characteristics and breast cancer image distribution characteristics. The results showed that this system is entirely possible to replace artificial visual diagnosis.

At the stage of physical examination and diagnosis, the auxiliary consultation and auxiliary diagnosis modules of the clinical auxiliary cancer decision system built based on AI and other technologies may be used to assist doctors in multiple aspects to make a diagnosis and decision on the people who tested positive in the clinical screening of screening projects. These may be used to improve the efficiency of medical diagnosis, shorten the time of confirmation, enhance the diagnosis and treatment ability, alleviate the shortage of medical resources, provide support for standardizing personalized diagnosis and treatment schemes for patients, and realize early diagnosis and treatment of cancers. Similar systems have been built and applied in China. For example, the intelligent auxiliary multiple myeloma decision system built based on AI provides support for doctors' clinical diagnosis and decision-making by mining and analyzing a large number of patients' clinical diagnosis and treatment data and experts' diagnosis and treatment experience [39]. The system has shown that the confirmation rate has increased, the time required for confirmation of patients has been significantly shortened, and the average length of stay and cost have also reduced sharply. Sichuan Cancer Hospital has designed and applied a clinical auxiliary cancer decision system according to its actual needs [40]. The system makes use of data mining technologies to automatically extract data related to patients' medical records and related analysis indicators of medical record data and keeps improving the accuracy and execution efficiency of clinical decisions. In addition, a telemedicine platform for the Oncology Department has been established by using internet and data processing technologies. It is used to perform such functions as remote ultrasound examination, remote pathological examination, expert consultation, cancer diagnosis, and referral for cancer patients. It is also used to diagnose and treat cancers earlier, to provide convenient cancer diagnosis and treatment services for residents in remote and grassroots areas with a low diagnosis and treatment level and to alleviate the problem of uneven distribution of medical resources especially diagnosis and treatment resources of cancer departments in China.

Stage of Cancer Registration

At the stage of cancer registration, efforts should be made to learn from the experience in Yanting in Sichuan Province, Shanghai, Jiaxing, and Ninghai in Zhejiang Province [41]. These are regions that fully rely on big data mining, AI, and information capture technologies to timely extract cancer diagnosis, treatment, and other medical record information, as well as follow-up visit and risk factor information, and build a monitoring database with more accurate and detailed risk factors, incidence, and death information. Separated population-based cancer registration should be integrated with hospital-based cancer registration, and both kinds of cancer registration should complement each other. A connection and information-sharing mechanism between cancer registration platforms and medical institutions in a wider range in the country should be established to further enable the functions of cross-regional information capture, integration, analysis, and release.

XSL•FO
RenderX

Stage of Follow-Up Visits

At the stage of follow-up visits of screening projects, existing follow-up visit management experiences of patients with cancer [42,43] should be used for reference. Further, a follow-up visit management system for cancer prevention and control should be built based on internet and AI-based machine learning technologies to realize the information interconnection with hospitals' diagnosis and treatment systems and cancer registration systems. With the help of the advantages of different platforms such as telecommunication networks, internet platforms, and mobile terminals, relevant information of follow-up recipients should be collected, supplemented, and improved constantly to automatically shape follow-up visit files. Based on the massive data collected, the next measures to be taken for follow-up recipients should be assessed. According to assessment results, relevant content should be intelligently pushed to follow-up recipients and medical personnel. Information of each follow-up visit should be automatically recorded and stored to realize such functions as screening service and process management, information collection, data visualization, and multiterminal support. It should also realize the automatic generation of treatment or follow-up visit arrays based on the pathological examination results of screening service receivers; realize interconnection with medical and health institutions in terms of admission, incidence, diagnosis, and ending information; and promote the complete and efficient management of cancer prevention and control efforts.

Conclusion

Combined with the analysis of major problems in cancer prevention and control projects, this paper probes into how to apply emerging information technologies, for example, AI-based machine learning, data mining, and electronic information collection, to various stages of cancer prevention and control. Information technologies realize the integrated management of prevention and control processes, for example, mobilization and preliminary identification, high-risk assessment, clinical screening, clinical diagnosis and treatment, tracking and follow-up, and biological sample management of high-risk groups, and promote the efficient implementation of cancer prevention and control projects in China.

Acknowledgments

We greatly appreciate the kind assistance of WQ and AM. PD takes responsibility for the integrity of the data. She is also grateful to his family and friends for their support and concern. The opinions reported in this paper are those of the authors alone and are not the official position of any affiliated institutions.

Authors' Contributions

PD and WQ are co-first authors and have contributed equally to this work. AM and GL are co-corresponding authors and have contributed equally to this work. PD, WQ and AM conceived and designed this study. PD contributed to design and the creation of figures, tables, and wrote the manuscript. AM and GL contributed to direct this study. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. World Health Organization. Global health estimates: Leading causes of death. World Health Organization. 2020. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death [accessed 2023-04-21]
2. The 2019 Global Burden of Disease (GBD) study. IHME, Global Health Data Exchange. 2020. URL: https://vizhub.healthdata.org/gbd-results/ [accessed 2023-04-21]
3. Gao B, Chu HC, Lu WL, Wang Y. Changing trend concerning the burden of cancer between 1990 and 2019 in China. Chin J Dis Control Prev 2022;26(4):430-489. [doi: 10.16462/j.cnki.zhjbkz.2022.04.012]
4. Shi JF, Shi CL, Yue XP, Huang HY, Wang L, Li J, et al. Economic burden of cancer in China during 1996-2014: a systematic review. Zhonghua Zhong Liu Za Zhi 2016;38(12):929-941. [doi: 10.3760/cma.j.issn.0253-3766.2016.12.010] [Medline: 27998471]
5. Teng Y, Cao MM, Chen WQ. Development, status and challenges of cancer screening in China. China Cancer 2022;31(7):481-487. [doi: 10.11735/j.issn.1004-0242.2022.07.A001]
6. Cao MM, Chen WQ. Promoting government-supported cancer screening programs in China. China Cancer 2021;30(11):803-805. [doi: 10.11735/j.issn.1004-0242.2021.11.A001]
7. The CPC Central Committee and the state council issued the "health China 2030" planning outline. Xinhua News Agency. 2016. URL: http://www.gov.cn/zhengce/2016-10/25/content_5124174.htm [accessed 2023-04-21]
8. Medium- and long-term plan for the prevention and treatment of chronic diseases in China (2017-2025) State Office [2017]. National Development and Reform Commission of the People's Republic of China. 2017. URL: https://www.ndrc.gov.cn/fggz/fzzlgh/gjjzxgh/201705/t20170511_1196751.html [accessed 2023-04-21]

XSL•FO
RenderX

9. Health China Action (2019-2030). National Health Commission of the People's Republic of China. 2019. URL: http://www.gov.cn/xinwen/2019-07/15/content_5409694.htm [accessed 2023-04-21]
10. Notice on the issuance of the "14th five-year plan" for national health, state office issued [2022] 11. General Office of the State Council of the People's Republic of China. 2022. URL: http://www.gov.cn/zhengce/zhengceku/2022-05/20/content_5691424.htm [accessed 2023-04-21]
11. Letter of response to proposal no. 2941 (medical and sports no. 355) of the fifth session of the 12th national committee of the Chinese people's political consultative conference. National Health Commission of the People's Republic of China. 2018. URL: http://www.nhc.gov.cn/wjw/tia/201801/eec6b89d0ed44c338ead262694dffa3b.shtml [accessed 2023-04-17]
12. Li J, Li H, Zeng H, Zheng R, Cao M, Sun D, et al. Trends in high-risk rates and screening rates for the population-based cancer screening program on esophageal, stomach and liver cancer in China, 2010-2016. J Natl Cancer Cent 2021;1(3):101-107 [FREE Full text] [doi: 10.1016/j.jncc.2021.05.001]
13. Letter of response to proposal no. 0873 (medical and sports no. 094) of the second session of the 13th national committee of the Chinese people's political consultative conference. National Health Commission of the People's Republic of China. 2020. URL: http://www.nhc.gov.cn/wjw/tia/202009/511b9b02b53e4aa2a2b67cdbf6878435.shtml [accessed 2023-04-17]
14. Chen W, Li N, Cao M, Ren J, Shi J, Chen H, et al. Preliminary analysis of cancer screening program in urban China from 2013 to 2017. China Cancer 2020;29(1):1-6. [doi: 10.11735/j.issn.1004-0242.2020.01.A001]
15. The interim data of the joint screening program for common cancers published in Tianjin. Tianjin Municipal People's Government. 2021. URL: https://www.tj.gov.cn/sy/tjxw/202108/t20210806_5526832.html [accessed 2023-04-23]
16. The first oral cancer screening program would be implemented in Hunan, China. Hunan Health Commission. 2019. URL: http://wjw.hunan.gov.cn/wjw/xxgk/gzdt/szdt/201909/t20190916_5484747.html [accessed 2023-04-24]
17. "Cervical cancer screening public welfare project in central and Western China" was successfully launched in Beijing. National center for women and children's health, China CDC. 2023. URL: https://www.chinawch.org.cn/zxgz2021/202303/t20230317_264360.html [accessed 2023-04-23]
18. Health poverty alleviation project of early gastrointestinal cancer screening launched will enable 50,000 people to receive public welfare screening and treatment. 39Health.com. 2019. URL: http://zl.39.net/a/190617/7224422.html [accessed 2023-04-23]
19. Bureau of Disease Control and Prevention, Ministry of Health. Technical Plan of Cancer Screening Program[R]. Beijing: People's Health Publishing House; 2011.
20. Notice on the issuance of cervical cancer screening work program and breast cancer screening work program. General Office of National Health and Health Commission. 2022. URL: http://www.nhc.gov.cn/fys/s3581/202201/cad44d88acca4ae49e12dab9176ae21c.shtml [accessed 2023-04-23]
21. Cao M, Chen W. Epidemiology of cancer in China and the current status of prevention and control. Chin J Clin Oncol 2019;46(3):145-149. [doi: 10.3969/j.issn.1000-8179.2019.03.283]
22. Shi JF, Mao AY, Bai YN, Liu GX, Liu CC, Wang H, et al. Priority setting in scaled-up cancer screening in China: an systematic review of economic evaluation evidences. Zhonghua Yu Fang Yi Xue Za Zhi 2020;54(3):306-313. [doi: 10.3760/cma.j.issn.0253-9624.2020.03.012] [Medline: 32187937]
23. Mao AY, Shi JF, Qiu WQ, Dong P, Sun ZX, Huang HY, et al. Willingness of potential service suppliers to provide cancer screening in urban China. Zhonghua Liu Xing Bing Xue Za Zhi 2018;39(2):150-156. [doi: 10.3760/cma.j.issn.0254-6450.2018.02.004] [Medline: 29495197]
24. Shi JF, Mao AY, Sun ZX, Lei HK, Qiu WQ, Huang HY, et al. Willingness and preferences of actual service suppliers regarding cancer screening programs: a multi-center survey in urban China. Zhonghua Liu Xing Bing Xue Za Zhi 2018;39(2):142-149. [doi: 10.3760/cma.j.issn.0254-6450.2018.02.003] [Medline: 29495196]
25. Lei HK, Dong P, Zhou Q, Qiu WQ, Sun ZX, Huang HY, et al. Potential demand on cancer screening service in urban populations in China: a cross-sectional survey. Zhonghua Liu Xing Bing Xue Za Zhi 2018;39(3):289-294. [doi: 10.3760/cma.j.issn.0254-6450.2018.03.008] [Medline: 29609241]
26. Yan XL, Mao AY, Hu GY, Dong P, Qiu WQ, Liu RM. Acceptability of cancer screening among urban residents in Beijing. Chin J Public Health 2015;31(8):1012-1015. [doi: 10.11847/zgggws2015-31-08-10]
27. He J. Strengthen the cancer surveillance to promote cancer prevention and control in China. Zhonghua Zhong Liu Za Zhi 2018;40(1):1-4. [doi: 10.3760/cma.j.issn.0253-3766.2018.01.001] [Medline: 29365410]
28. Wang K, Liu CC, Mao AY, Shi JF, Dong P, Huang HY, et al. Analysis on the demand, access and related factors of cancer prevention and treatment knowledge among urban residents in China from 2015 to 2017. Zhonghua Yu Fang Yi Xue Za Zhi 2020;54(1):84-91. [doi: 10.3760/cma.j.issn.0253-9624.2020.01.016] [Medline: 31914574]
29. Huang J, Yang XH, Liu A, Zhou WJ. Problems and countermeasures in the implementation of national cervical and breast screening program for women in rural areas. Chin Gen Pract 2020;23(13):1680-1686. [doi: 10.12114/j.issn.1007-9572.2019.00.509]
30. Liang L, Lyu YL, Wang HD. Status quo and prospects of cancer screening in Huaihe River basin of Anhui province. China Cancer 2022;31(9):679-682. [doi: 10.11735/j.issn.1004-0242.2022.09.A001]

XSL•FO
RenderX

31. Dong P, Qiu WQ, Shi JF, Mao AY, Huang HY, Sun ZX, et al. Cancer screening service utilization and willingness-to-pay of urban populations in China: a cross-sectional survey from potential service demander's perspective. Zhonghua Liu Xing Bing Xue Za Zhi 2018;39(2):165-172. [doi: 10.3760/cma.j.issn.0254-6450.2018.02.006] [Medline: 29495199]
32. Zhu J, Huang HY, Mao AY, Sun ZX, Qiu WQ, Lei HK, et al. Preference on screening frequency and willingness-to-pay for multiple-cancer packaging screening programs in urban populations in China. Zhonghua Liu Xing Bing Xue Za Zhi 2018;39(2):157-164. [doi: 10.3760/cma.j.issn.0254-6450.2018.02.005] [Medline: 29495198]
33. Perera M, Mirchandani R, Papa N, Breemer G, Effeindzourou A, Smith L, et al. PSA-based machine learning model improves prostate cancer risk stratification in a screening population. World J Urol 2021;39(6):1897-1902. [doi: 10.1007/s00345-020-03392-9] [Medline: 32747980]
34. Zhong J, Chen JM, Chen SL, Yi YF. Constructing a risk prediction model for lung cancer recurrence by using gene function clustering and machine learning. Comb Chem High Throughput Screen 2019;22(4):266-275. [doi: 10.2174/1386207322666190129111749] [Medline: 30698110]
35. Liu M, Liu Z, Cai H, Guo C, Li X, Zhang C, et al. A model to identify individuals at high risk for esophageal squamous cell carcinoma and precancerous lesions in regions of high prevalence in China. Clin Gastroenterol Hepatol 2017;15(10):1538-1546.e7 [FREE Full text] [doi: 10.1016/j.cgh.2017.03.019] [Medline: 28342951]
36. Zhang ZH, Cai YQ, Han D, Zhou XJ, Huang YL, Li JL. Pulmonary nodule screening and qualitative diagnosis based on deep learning. Cancer Res Prev Treat 2020;47(4):283-287. [doi: 10.3971/j.issn.1000-8578.2020.19.1107]
37. Bi K, Wang Y, Zhang TT, Qian YP, Qian ZB, Chen ZM, et al. An artificial intelligence pathological diagnosis model for lung cancer based on deep learning algorithm: development and application. Acad J Second Mil Med Univ 2020;41(11):1229-1235. [doi: 10.16781/j.0258-879x.2020.11.1229]
38. Loukas C, Kostopoulos S, Tanoglidi A, Glotsos D, Sfikas C, Cavouras D. Breast cancer characterization based on image classification of tissue sections visualized under low magnification. Comput Math Methods Med 2013;2013:829461 [FREE Full text] [doi: 10.1155/2013/829461] [Medline: 24069067]
39. Zhu JF, Zheng XP, Li C, Zhao ZF, Xing Z. Construction and clinical application of assisted decision-making system for multiple myeloma. China Digit Med 2020;15(12):61-64. [doi: 10.3969/j.issn.1673-7571.2020.12.018]
40. Yang PJ. Research and implementation of cancer clinical assistant decision system driven by rule engine. School of Information and Software Engineering. 2021. URL: https://d.wanfangdata.com.cn/thesis/D02414555 [accessed 2024-02-01]
41. Zhang W, Liu XX, Qiao L, Li B. Exploring a new model for the construction of cancer registries in the era of big data. J Cancer Control Treat 2019;32(11):951-954. [doi: 10.3969/j.issn.1674-0904.2019.11.001]
42. Ren J. Application of information technology in tumor follow-up. Equip Manage Med Inf 2019;25(10):181-182 This reference can be reach on the website of "https://d.wanfangdata.com.cn/periodical/zgylqxxx201910092". [doi: 10.3969/j.issn.1006-6586.2019.10.092]
43. Zheng CW. The design and implementation of nasopharyngeal cancer screening and follow-up information management system. Qufu Normal University. 2019. URL: https://d.wanfangdata.com.cn/thesis/D01841987 [accessed 2024-02-01]

Abbreviations

AI: artificial intelligence

Edited by A Mavragani; submitted 16.06.23; peer-reviewed by J Noah, Y Tao, J Shi; comments to author 27.07.23; revised version received 30.10.23; accepted 29.01.24; published 27.02.24.

Please cite as:
Dong P, Mao A, Qiu W, Li G
Improvement of Cancer Prevention and Control: Reflection on the Role of Emerging Information Technologies
J Med Internet Res 2024;26:e50000
URL: https://www.jmir.org/2024/1/e50000
doi:10.2196/50000
PMID:38412009

©Pei Dong, Ayan Mao, Wuqi Qiu, Guanglin Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Promises, Pitfalls, and Clinical Applications of Artificial Intelligence in Pediatrics

Hansa Bhargava[1,2,3], MD; Carmela Salomon[4], PhD; Srinivasan Suresh[5,6], MD; Anthony Chang[7], MBA, MD, MPH, MS; Rachel Kilian[8], PhD; Diana van Stijn[9], MD, PhD; Albert Oriol[10], MBA, MHS; Daniel Low[11], MD; Ashley Knebel[4]; Sharief Taraman[4,12,13], MD

[1]Children's Hospital of Atlanta, Atlanta, GA, United States
[2]School of Medicine, Emory University, Atlanta, GA, United States
[3]Healio, South New Jersey, NJ, United States
[4]Cognoa, Inc, Palo Alto, CA, United States
[5]Division of Health Informatics, Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, United States
[6]UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, United States
[7]Fowler School of Engineering, Chapman University, Orange, CA, United States
[8]SSI Strategy, Parsippany, NJ, United States
[9]Lapsi Health, Amsterdam, Netherlands
[10]Rady Children's Hospital, San Diego, CA, United States
[11]AdaptX, Seattle, WA, United States
[12]Children's Hospital of Orange County, Orange, CA, United States
[13]University of California Irvine School of Medicine, Irvine, CA, United States

Corresponding Author:
Carmela Salomon, PhD
Cognoa, Inc
2185 Park Blvd
Palo Alto, CA, 94306
United States
Phone: 1 8664264622
Email: carmela.salomon@cognoa.com

Abstract

Artificial intelligence (AI) broadly describes a branch of computer science focused on developing machines capable of performing tasks typically associated with human intelligence. Those who connect AI with the world of science fiction may meet its growing rise with hesitancy or outright skepticism. However, AI is becoming increasingly pervasive in our society, from algorithms helping to sift through airline fares to substituting words in emails and SMS text messages based on user choices. Data collection is ongoing and is being leveraged by software platforms to analyze patterns and make predictions across multiple industries. Health care is gradually becoming part of this technological transformation, as advancements in computational power and storage converge with the rapid expansion of digitized medical information. Given the growing and inevitable integration of AI into health care systems, it is our viewpoint that pediatricians urgently require training and orientation to the uses, promises, and pitfalls of AI in medicine. AI is unlikely to solve the full array of complex challenges confronting pediatricians today; however, if used responsibly, it holds great potential to improve many aspects of care for providers, children, and families. Our aim in this viewpoint is to provide clinicians with a targeted introduction to the field of AI in pediatrics, including key promises, pitfalls, and clinical applications, so they can play a more active role in shaping the future impact of AI in medicine.

(J Med Internet Res 2024;26:e49022) doi:10.2196/49022

KEYWORDS

artificial intelligence; pediatrics; autism spectrum disorder; ASD; disparities; pediatric; youth; child; children; autism; autistic; barrier; barriers; clinical application; clinical applications; professional development; continuing education; continuing medical education; CME; implementation

XSL•FO
RenderX

What is Artificial Intelligence in Medicine?

Artificial intelligence (AI) is a broad term describing the use of machine-based learning algorithms and software [1]. Nonmedical applications of AI include text prediction when we write emails or movie suggestions on cable and streaming based on our previous viewing choices. Machine learning (ML) is a specific branch of AI focused on altering programs or algorithms based on exposure to data in order to improve performance over time; in other words, the "machine" is "learning" as it accumulates more data and patterns. ML exists on a continuum. Supervised algorithms, for example, may require a great deal of outside input in order to function, while unsupervised algorithms can function with much greater degrees of autonomy [2]. One ML approach, useful when exploring complex nonlinear patterns, involves convolutional neural networks (CNNs). CNNs apply a special mathematical operation called a convolution across cumulative steps to perform specialized tasks like image processing. Natural language processing (NLP) solutions using ML are also being pioneered to help computers understand, categorize, and extract insights from natural language data sets [1]. In the field of medicine, AI has shown promise to assist with a wide array of clinical tasks, including risk prediction, diagnosis, and augmented decision-making and treatment and monitoring. Intelligent technology may also be leveraged to streamline workflows and automate some routine but historically time-intensive tasks, such as clinical note-taking [3]. NLP solutions, for example, have proven extremely useful in supporting clinicians to more efficiently leverage the large quantities of data collected in electronic medical records (EMRs) every day. NLP "reads" EMRs, attempting to understand medical words and phrases within the specific context of a given patient. IBM Watson, for example, has already successfully been used to examine large data sets of EMRs from diverse populations to create lists of common complications for a given population. Watson has also been used to gather medical literature in response to queries contained within examined EMRs [4].

In the following section of this viewpoint, we explore a selection of pediatric conditions, where AI-based approaches are being pioneered to various support risk prediction, diagnosis, therapy and response monitoring, or clinical workflow efficiency.

AI in Pediatric Asthma Care

Risk Prediction

Despite well-established clinical standards and pathways for asthma management, asthma remains one of the most frequent clinical concerns in pediatric offices and emergency rooms [5]. AI-based innovative solutions are being explored to help predict asthma risk so earlier intervention can occur. One study analyzed over 335,000 asthma events spanning the years 2005 to 2018 and then used ML to predict the risk of hospitalization for patients with asthma [6]. Similarly, another set of ML algorithms integrated clinical data with information on weather, neighborhood characteristics, and community viral load to predict the likelihood of hospitalization for children with asthma [7]. Another study used EHR data from 9934 children to train ML models to accurately predict the likelihood of a child's asthma persisting over time [8].

Monitoring and Medication Management

Several companies are pioneering efforts in this space. One company is optimizing AI for spirometry in a cell phone app [9]. The app captures a patient's exhalation with the phone microphone and then translates these acoustic sounds into a curve representing lung volume. A portable spirometry test allows patients to conveniently track deterioration or therapeutic improvement over time. A separate mobile tool uses smart sensor data to register inhaler usage. This sensor analyzes the inhalation history of a patient in relation to weather conditions in order to recognize potential triggers of asthma attacks [10]. This analysis can also provide suggestions for the treatment dosage based on exacerbating factors present [10]. Another innovative digital therapy (DTx) uses AI to alert patients and parents to the early stages of an asthma attack. DTx continuously analyzes respiratory sounds in order to more quickly recognize changes in the sounds that could signal an oncoming asthmatic event. Ideally, warnings from DTx will help children and caregivers recognize the triggers as well as symptoms of an asthmatic crisis. Predicting outcomes in real time by using patient data and pattern detection algorithms patterns could lead to more precise and personalized asthma management [11].

AI in Pediatric Rare Disease Management: Diagnostics

A leading children's hospital has developed a program that uses pattern recognition and real-time data analysis to efficiently diagnose rare diseases in newborns. Leveraging AI to help automate genomic diagnosis could streamline and expedite the diagnostic process in neonatal intensive care units and pediatric intensive care units (PICUs). While clinical adoption considerations must first be addressed, such technology could potentially reduce delays to targeted life-changing treatments in newborns with diseases of previously unknown etiology [12].

AI in Pediatric Sepsis: Risk Detection

Another serious concern in pediatrics is the early diagnosis of sepsis. In a study of almost 500 PICU patients, AI detected severe sepsis as early as 8 hours prior to traditional EMR-based screening algorithms [13]. This efficiency can have a profound impact on the management of sepsis in the PICU by allowing for earlier intervention and thus potentially reducing morbidity and mortality.

AI and Opioid Dependence: Monitoring and Medication Management

In the United States, approximately 7% of adolescents develop persistent opioid use after surgery [14]. A new company focused on health and wellness for athletes recently made use of AI-powered software, OR Advisor (AdaptX). The use of OR Advisor enables frontline clinicians to perform rapid analysis of real-world data collected in their EMR. OR Advisor is also capable of tracking clinical metrics over time as well as

XSL•FO
RenderX

monitoring treatment outcomes [14]. OR Advisor can help physicians more quickly answer clinically relevant questions, such as how adjusting medication doses might affect pain management. Using OR Advisor, Seattle Children's Bellevue Clinic and Surgery Center was able to reduce opioid administration from 85% of surgeries to less than 1% [15]. Decreasing perioperative opioid use could potentially not only save thousands of lives per year but also reduce complications, thereby reducing the length of stay and improving recovery.

AI in Radiology: Diagnostics

AI in radiology is perhaps the most impressive application of AI in medicine, with CNNs often performing as or even more accurately than clinicians. AI has already been applied in a variety of pediatric settings, such as scoliosis quantification, fracture and injury detection, and cystic fibrosis scoring on chest radiographs [16], but there are many other potential uses. One of the best potential uses of CNNs is as a screening tool for life-threatening diagnoses that need rapid intervention, such as acute ischemic strokes [17]. CNNs can accelerate the timeline of a non–life-threatening diagnosis (such as a small pneumonia) as well. Sophisticated AI image interpretation can also potentially lead to novel discoveries in medical images (such as tumor-stroma interfaces) that can help with subtyping, therapy modification, and prognostication of medical conditions [18]. To date, the majority of AI-based radiology solutions have been developed for adult populations. Radiology imaging advocacy groups have recently begun lobbying the US Congress to develop policies addressing the scarcity of pediatric-specific AI-based innovations [19].

AI in Pediatric Autism: Diagnosis

AI is being leveraged to support improved behavioral and mental health outcomes in pediatric populations, including in children with autism. While autism diagnosis is possible as early as the age of 18 months, the average age of diagnosis in the United States remains older than 4 years [20,21]. The majority of these children are diagnosed in specialty care; however, the increasing demand for evaluations coupled with a shortage of specialists has contributed to lengthy waits for assessments [20,22,23]. Delays in treatment initiation can have lifelong implications for children and their families [24,25]. Applications using AI can potentially reduce these delays. For example, a Food and Drug Administration (FDA)–authorized AI-based diagnostic device, Canvas Dx, uses inputs from a child's health care provider, caregiver, and video analysis to help health care providers rapidly diagnose or rule out autism in young children with concern for developmental delay [26-28]. The diagnostic device can be used in primary care settings, potentially decreasing delays to treatment initiation and thus significantly improving the child's trajectory [27].

AI may also have a role to play in reducing diagnostic bias prevalent in some autism care pathways. Currently, compared to White peers, Black and Latino or Hispanic children receive later autism diagnoses [29,30]. One study found that in comparison to White children, Hispanic and Black children are 65% and 19% less likely to be diagnosed with autism and these racial or ethnic disparities exist regardless of economic class [29]. Similarly, girls are diagnosed with autism an average of 1.5 years later than their male peers [31,32]. Children from rural communities also experience lags in diagnosis and access issues. AI could help address some of these biases and promote diagnostic equity in several ways. Item-level racial or gender bias noted in some existing diagnostic and screening tools [33], for example, could be reduced through careful training of AI-based diagnostic alternatives on large samples of gender and racially diverse accurately labeled data. AI-based remotely administered technologies may also reduce access barriers for children in rural communities who face health care disparities due to geographical isolation, hospital closures, and insufficient clinical workforce coverage [34].

Adolescent Depression and Anxiety: AI and Symptom Management

Mental health management in preteens and adolescents is another important concern in the primary care pediatric setting and can be especially challenging in the context of limited mental health resource availability. An estimated 31% of teens have an anxiety disorder [35]. The severity of these disorders and the scarcity of providers make it imperative that patients and caregivers have easily accessible resources that can be used while waiting to see a mental health provider. Such is the case with a research-based mental health software app, Woebot, that provides digital cognitive behavioral therapy to patients. Using AI, Woebot tailors its conversations to each patient, with the goal of providing personalized and high-quality care [33]. In a study published in *JMIR*, the intervention group experienced a 22% reduction in depressive symptoms over the course of 2 weeks after using this software in comparison with the control group [33]. The product was initially tested with subjects between the ages of 18 and 65 years; however, it may prove to be useful in the pediatric space as well, given ongoing mental health access and resource issues. Table S1 in Multimedia Appendix 1 [6-10,12,13,36,37] summarizes topical studies discussed in this viewpoint including further detail about methodology, performance metrics, and product development.

Clinical Challenges and Concerns

Although AI has the potential to meaningfully benefit pediatric care, a number of regulatory, implementation, and ethical challenges exist. Additionally, poorly conceived AI has the potential to cause harm and therefore requires serious clinical consideration and risk identification and mitigation as part of any workflow implementation process. First, clinicians should be aware that training algorithms often require large data sets. In pediatrics, the small sample sizes of some data sets, especially when split by age group, can pose challenges when building unbiased AI algorithms. Relatedly, in order to build less biased algorithms, training data must be representative of the applicable populations across domains such as race, ethnicity, gender, socioeconomic status, and geographic contexts. Obtaining samples with known demographics and large enough subsamples of intersecting groups can prove challenging.

XSL•FO
RenderX

While we previously touched on the potential of AI to enhance health equity, the reverse may also be true. Without careful and equitable sampling, algorithms can potentially perpetuate rather than reduce bias. Even when this risk is mitigated by training AI algorithms on a diverse patient base, bias can still occur at the data labeling level. This is because the algorithm may be "learning" from the implicit biases of humans who are annotating the input data or validating its output classifications. Clinicians can help mitigate these risks by requesting information on the size and diversity of the training data used and through careful interrogation of any potential labeling biases.

Also considering that even when AI is built on representative data sets and yields generalizable algorithms, implementation challenges may reduce the extent to which such technologies can address inequities in practice. An unbiased algorithm implemented inequitably could, in fact, maintain or perpetuate disparities in pediatric health care. Families without access to a computer or reliable internet connectivity, for example, may not be able to access some digital technologies that could benefit them. As AI-powered tools become more commonplace in clinical practice, we recommend that their implementation be accompanied by physician education, antibias training, and system-wide change aimed at making health care more accessible to marginalized populations. Similarly, disparities in access to technology should be addressed. These goals should be supported through tactics such as the development of consensus ethical frameworks for the use of AI health care, the introduction of robust AI and ethics curriculum in medical education, as well as regulatory frameworks.

There are a number of ethical questions to consider when integrating AI-based technologies into clinical pathways. For example, what are the implications of physicians relying on outputs from complex, multidimensional ML models in clinical decision-making when they may have limited visibility into how such models arrive at diagnostic conclusions? [3,38] Thoughtful use of technology also necessitates weighing the benefits of intelligent technology against the potential scale of harm that could occur if a widely deployed algorithm were to malfunction [3]. Open questions also remain about who should be held accountable for incorrect diagnostic or treatment recommendations produced by devices powered by AI [39]. Some patients have also raised concerns about how their data privacy and security will be maintained within such systems [40]. In response to these questions, a number of ethical frameworks for the use of AI in clinical practice are being developed [41,42]. These include action points related to building institutional capabilities to redress any emerging AI-related harms; distinguishing between appropriate and inappropriate AI task delegation; developing metrics to help measure and monitor the trustworthiness of AI-based technologies; and financially incentivizing the inclusion of ethical, legal, and social considerations in AI research projects [41].

Open questions also remain about how to effectively regulate AI-based medical technologies. To date, the FDA has cleared more than 160 AI-powered devices for use [43]. Existing regulatory frameworks, however, do not account for the ability of intelligent devices to iteratively "learn" from the data they are exposed to and alter their algorithms in response [44]. Should such features be disabled in approved products to avoid drift in performance and function? If not, how should such features be scrutinized and regulated on an ongoing basis? The FDA proposed regulatory framework for software as medical device technologies provide some direction for potential solutions such as predetermined change control plan and algorithm change plan protocol specifications [45].

Future of AI in Medical Education

As clinical AI apps become more commonplace, pediatricians will require fundamental training and orientation to the data science behind AI so they can use AI-enabled tools appropriately and evaluate their strengths and weaknesses. However, AI integration into medical curriculum is currently limited and inconsistent, and no topical content is tested in key licensing examinations [46]. This has left both students and practicing clinicians with knowledge and confidence gaps regarding the application of AI in clinical practice [47-49].

Medical schools should establish introductory AI courses to foster clinical confidence and competence. The future pediatrician will benefit greatly from the synergy between AI and human cognition. A number of proposed curriculum updates are being explored [46,50,51], including MEd2030, an updated medical school curriculum that will add 10 additional modules (eg, biomedical informatics and AI, transdisciplinary collaboration and diversity principles, biomedical entrepreneurship, design innovation) [52]. In the continuing education sector, conferences and programs on digital therapeutics and AI-based health technologies are also growing as more companies introduce FDA-regulated products.

Actively engaging with AI-based technologies will allow pediatricians to play a greater role in the implementation and use of AI in every aspect of health care. Learning to apply developing technologies is one key to optimizing patient care and improving the future of health care systems globally. When EMRs were first introduced in the 2000s, many clinicians knew about them, but few were engaged in the implementation and continual improvement of the technology. EMRs have since infiltrated almost every aspect of a clinician's practice; yet, as a result of limited early-stage clinical engagement, they are often not optimized for the clinical workflow [1]. By seeking to understand and engage with the next wave of technological advances earlier and more proactively, pediatricians can help ensure they will be more tailored to the needs of both care providers and patients.

Conclusions

Health care delivery models are changing rapidly, and AI is poised to greatly impact the future of pediatric care. Efficient diagnostic tools, faster data analysis, predictive outcomes modeling, more streamlined care experiences, and automation of some previously time-intensive tasks are just some of the potential benefits offered by AI. However, widespread clinical integration of AI into pediatric care pathways will also require

XSL•FO
RenderX

thoughtful solutions to complex data quality and privacy issues, and ethical and regulatory implementation barriers. We recommend that tailored AI-focused curriculum and ongoing training opportunities be developed and implemented rapidly to support pediatricians and medical students alike to responsibly use and understand the strengths and limitations of such technologies more deeply. With improved education, we are optimistic about the future potential of AI to enhance clinical efficiencies and outcomes and to support broader access to high-quality pediatric health care in marginalized communities.

Acknowledgments

No generative AI was used in any portion of the manuscript writing.

Authors' Contributions

AO contributed to the manuscript section focusing on using artificial intelligence (AI) to expedite genomic diagnosis and to predict the physiological deterioration of noncardiac patients in the intensive care unit as well as patients in the cardiothoracic intensive care unit. DvS contributed to the case study examples including the use of AI in pediatric respiratory diseases. AK made revisions to each section and helped with the organization of the manuscript. CS contributed to the section on the challenges of AI in health care, made revisions to each section of the manuscript, and helped with the organization of the manuscript. ST contributed to all sections of the manuscript and provided final review and edits. Primary contributions were in the overview of AI, the use of AI in clinical care specifically in regard to autism spectrum disorder and behavioral and mental health, challenges in AI, and future medical education. RK contributed to all sections of the manuscript and provided overall review and edits. SS contributed to all sections of the manuscript and provided overall review and edits. AC contributed to the section titled “What is artificial intelligence in medicine?” and provided overall review and edits. DL contributed to the section of the manuscript discussing the use of AI to reduce opioid use in surgery.

Conflicts of Interest

RK, HB, and AK are consultants for Cognoa, Inc. ST and CS are employees of Cognoa, Inc and have its stock options. ST additionally receives consulting fees for Cognito Therapeutics, volunteers as a board member of the American Academy of Pediatrics, Orange County Chapter, and American Academy of Pediatrics–California, is a paid advisor for MI10 LLC, and owns stock in NTX, Inc, and HandzIn. HB is also an advisor for FreeSpira. DvS is an employee and shareholder of Lapsi Health. DL is an employee of AdaptX. AO is an employee of Rady Children’s Hospital. AO also serves on the board of directors for San Diego Health Connect and the advisory board for KLAS Research.

Multimedia Appendix 1
Summary of reviewed studies.
[DOCX File , 16 KB - jmir_v26i1e49022_app1.docx]

References

1. Chang AC. Intelligence-Based Medicine: Artificial Intelligence and Human Cognition in Clinical Medicine and Healthcare. London, United Kingdom: Academic Press; 2020.
2. Aylward BS, Abbas H, Taraman S, Salomon C, Gal-Szabo D, Kraft C, et al. An introduction to artificial intelligence in developmental and behavioral pediatrics. J Dev Behav Pediatr 2023;44(2):e126-e134 [FREE Full text] [doi: 10.1097/DBP.0000000000001149] [Medline: 36730317]
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
4. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med 2018;131(2):129-133 [FREE Full text] [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
5. Zahran HS, Bailey CM, Damon SA, Garbe PL, Breysse PN. Vital signs: asthma in children - United States, 2001-2016. MMWR Morb Mortal Wkly Rep 2018;67(5):149-155 [FREE Full text] [doi: 10.15585/mmwr.mm6705e1] [Medline: 29420459]
6. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]
7. Patel SJ, Chamberlain DB, Chamberlain JM. A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. Acad Emerg Med 2018;25(12):1463-1470 [FREE Full text] [doi: 10.1111/acem.13655] [Medline: 30382605]
8. Bose S, Kenyon CC, Masino AJ. Personalized prediction of early childhood asthma persistence: a machine learning approach. PLoS One 2021;16(3):e0247784 [FREE Full text] [doi: 10.1371/journal.pone.0247784] [Medline: 33647071]

XSL•FO
RenderX

9. Larson EC, Goel M, Boriello G, Heltshe S, Rosenfeld M, Patel SN. SpiroSmart: using a microphone to measure lung function on a mobile phone. : ACM Press; 2012 Presented at: Ubicomp '12: The 2012 ACM Conference on Ubiquitous Computing; September 5-8, 2012; Pittsburgh, PA p. 280. [doi: 10.1145/2370216.2370261]
10. Van Sickle D, Barrett M, Humblet O, Henderson K, Hogg C. Randomized, controlled study of the impact of a mobile health tool on asthma SABA use, control and adherence. Eur Respir J 2016;48(suppl 60):PA1018 [FREE Full text] [doi: 10.1183/13993003.congress-2016.PA1018]
11. Dimitriades JB. Ep. 85: value of digital therapeutics – Jhonatan Dimitriades, MD (CEO Lapsi Health). Health Podcast Network. 2021. URL: https://healthpodcastnetwork.com/episodes/health-unchained/ep-85-value-of-digital-therapeutics-jhonatan-dimitriades-md-ceo-lapsi-health/ [accessed 2021-11-15]
12. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med 2019;11(489):eaat6177 [FREE Full text] [doi: 10.1126/scitranslmed.aat6177] [Medline: 31019026]
13. Kamaleswaran R, Akbilgic O, Hallman MA, West AN, Davis RL, Shah SH. Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the PICU. Pediatr Crit Care Med 2018;19(10):e495-e503 [FREE Full text] [doi: 10.1097/PCC.0000000000001666] [Medline: 30052552]
14. AdaptX- eliminating opioids from ambulatory care. AdaptX. URL: https://www.adaptx.com/ [accessed 2024-01-31]
15. Low D. Closing the gateway from surgery to persistent opioid use. Institute for Healthcare Improvement. 2019. URL: http://www.ihi.org/communities/blogs/closing-the-gateway-from-surgery-to-persistent-opioid-use [accessed 2021-11-05]
16. Otjen JP, Moore MM, Romberg EK, Perez FA, Iyer RS. The current and future roles of artificial intelligence in pediatric radiology. Pediatr Radiol 2022;52(11):2065-2073 [FREE Full text] [doi: 10.1007/s00247-021-05086-9] [Medline: 34046708]
17. Öman O, Mäkelä T, Salli E, Savolainen S, Kangasniemi M. 3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke. Eur Radiol Exp 2019;3(1):8 [FREE Full text] [doi: 10.1186/s41747-019-0085-6] [Medline: 30758694]
18. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol 2022;35(1):23-32 [FREE Full text] [doi: 10.1038/s41379-021-00919-2] [Medline: 34611303]
19. Stempniak M. Radiology groups urge congress to address scarcity of AI solutions in pediatric care. Radiology Business. 2022. URL: https://www.radiologybusiness.com/topics/artificial-intelligence/radiology-congress-ai-solutions-pediatric-care [accessed 2024-01-31]
20. Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2018. MMWR Surveill Summ 2021;70(11):1-16 [FREE Full text] [doi: 10.15585/mmwr.ss7011a1] [Medline: 34855725]
21. Pierce K, Gazestani VH, Bacon E, Barnes CC, Cha D, Nalabolu S, et al. Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. JAMA Pediatr 2019;173(6):578-587 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.0624] [Medline: 31034004]
22. Maenner MJ, Shaw KA, Baio J, Washington A, Patrick M, DiRienzo M, et al. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2016. MMWR Surveill Summ 2020;69(4):1-12 [FREE Full text] [doi: 10.15585/mmwr.ss6904a1] [Medline: 32214087]
23. Zuckerman KE, Lindly OJ, Sinche BK. Parental concerns, provider response, and timeliness of autism spectrum disorder diagnosis. J Pediatr 2015;166(6):1431-1439.e1 [FREE Full text] [doi: 10.1016/j.jpeds.2015.03.007] [Medline: 25888348]
24. Vivanti G, Dissanayake C, Victorian ASELCC Team. Outcome for children receiving the early start denver model before and after 48 months. J Autism Dev Disord 2016;46(7):2441-2449 [FREE Full text] [doi: 10.1007/s10803-016-2777-6] [Medline: 27020055]
25. MacDonald R, Parry-Cruwys D, Dupere S, Ahearn W. Assessing progress and outcome of early intensive behavioral intervention for toddlers with autism. Res Dev Disabil 2014;35(12):3632-3644 [FREE Full text] [doi: 10.1016/j.ridd.2014.08.036] [Medline: 25241118]
26. CanvasDx. URL: https://canvasdx.com/ [accessed 2021-10-26]
27. Megerian JT, Dey S, Melmed RD, Coury DL, Lerner M, Nicholls CJ, et al. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. NPJ Digit Med 2022;5(1):57 [FREE Full text] [doi: 10.1038/s41746-022-00598-6] [Medline: 35513550]
28. FDA authorizes marketing of diagnostic aid for autism spectrum disorder. U.S. Food & Drug Administration. 2021. URL: https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-diagnostic-aid-autism-spectrum-disorder [accessed 2024-01-31]
29. Durkin MS, Maenner MJ, Baio J, Christensen D, Daniels J, Fitzgerald R, et al. Autism spectrum disorder among US children (2002-2010): socioeconomic, racial, and ethnic disparities. Am J Public Health 2017;107(11):1818-1826 [FREE Full text] [doi: 10.2105/AJPH.2017.304032] [Medline: 28933930]
30. Wiggins LD, Durkin M, Esler A, Lee LC, Zahorodny W, Rice C, et al. Disparities in documented diagnoses of autism spectrum disorder based on demographic, individual, and service factors. Autism Res 2020;13(3):464-473 [FREE Full text] [doi: 10.1002/aur.2255] [Medline: 31868321]

XSL•FO
RenderX

31. McCormick CEB, Kavanaugh BC, Sipsock D, Righi G, Oberman LM, De Luca DM, et al. Autism heterogeneity in a densely sampled U.S. population: results from the first 1,000 participants in the RI-CART study. Autism Res 2020;13(3):474-488 [FREE Full text] [doi: 10.1002/aur.2261] [Medline: 31957984]
32. Harrison AJ, Long KA, Tommet DC, Jones RN. Examining the role of race, ethnicity, and gender on social and behavioral ratings within the autism diagnostic observation schedule. J Autism Dev Disord 2017;47(9):2770-2782 [FREE Full text] [doi: 10.1007/s10803-017-3176-3] [Medline: 28597187]
33. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
34. Digital therapeutics: reducing rural health inequalities. Digital Therapeutics Alliance. 2022. URL: https://dtxalliance.org/wp-content/uploads/2021/01/DTA_Rural-Health_r13_110220.pdf [accessed 2024-04-30]
35. Any anxiety disorder. The National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder [accessed 2021-11-05]
36. Yang J, Zhang K, Fan H, Huang Z, Xiang Y, Yang J, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. Commun Biol 2019;2:390. [doi: 10.1038/s42003-019-0635-8] [Medline: 31667364]
37. Wall DP, Liu-Mayo S, Salomon C, Shannon J, Taraman S. Optimizing a de novo artificial intelligence-based medical device under a predetermined change control plan: improved ability to detect or rule out pediatric autism. Intelligence-Based Med 2023;8:100102. [doi: 10.1016/j.ibmed.2023.100102]
38. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745-e750 [FREE Full text] [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
39. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
40. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021;4(1):140 [FREE Full text] [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
41. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. In: Ethics, Governance, and Policies in Artificial Intelligence. Cham, Switzerland: Springer International Publishing; 2021:19-39.
42. Miller KW. Moral responsibility for computing artifacts: "the rules". IT Prof 2011;13(3):57-59 [FREE Full text] [doi: 10.1109/MITP.2011.46]
43. Katznelson G, Gerke S. The need for health AI ethics in medical school education. Adv Health Sci Educ Theory Pract 2021;26(4):1447-1458 [FREE Full text] [doi: 10.1007/s10459-021-10040-3] [Medline: 33655433]
44. Kagiyama N, Shrestha S, Farjo PD, Sengupta PP. Artificial intelligence: practical primer for clinical research in cardiovascular disease. J Am Heart Assoc 2019;8(17):e012788 [FREE Full text] [doi: 10.1161/JAHA.119.012788] [Medline: 31450991]
45. Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD) - discussion paper and request for feedback. U.S. Food and Drug Administration. 2019. URL: https://www.regulations.gov/document/FDA-2019-N-1185-0001 [accessed 2024-01-31]
46. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
47. Banerjee M, Chiew D, Patel KT, Johns I, Chappell D, Linton N, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021;21(1):429 [FREE Full text] [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
48. Dos Santos DP, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019;29(4):1640-1646 [FREE Full text] [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
49. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020;11(1):14 [FREE Full text] [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
50. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018;1:54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
51. Wartman SA, Combs CD. Reimagining medical education in the age of AI. AMA J Ethics 2019;21(2):E146-E152 [FREE Full text] [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
52. Chang A. Medical education for the future clinician: a proposal for our medical education leaders. Medical Intelligence 10. URL: https://www.mi10.ai/2021/10/06/medical-education-for-the-future-clinician/ [accessed 2024-04-30]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

CNN: convolutional neural network
DTx: digital therapy
EMR: electronic medical record
FDA: Food and Drug Administration
ML: machine learning
NLP: natural language processing
PICU: pediatric intensive care unit

Edited by T de Azevedo Cardoso; submitted 15.05.23; peer-reviewed by V Kaelin, D Riggins; comments to author 11.08.23; revised version received 01.09.23; accepted 22.01.24; published 29.02.24.

Please cite as:
Bhargava H, Salomon C, Suresh S, Chang A, Kilian R, Stijn DV, Oriol A, Low D, Knebel A, Taraman S
Promises, Pitfalls, and Clinical Applications of Artificial Intelligence in Pediatrics
J Med Internet Res 2024;26:e49022
URL: https://www.jmir.org/2024/1/e49022
doi:10.2196/49022
PMID:38421690

©Hansa Bhargava, Carmela Salomon, Srinivasan Suresh, Anthony Chang, Rachel Kilian, Diana van Stijn, Albert Oriol, Daniel Low, Ashley Knebel, Sharief Taraman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Digital Therapeutics Real-World Evidence Framework: An Approach for Guiding Evidence-Based Digital Therapeutics Design, Development, Testing, and Monitoring

Meelim Kim[1,2,3,4], PhD; Kevin Patrick[1,3], MS, MD; Camille Nebeker[1,3,4], PhD; Job Godino[1,3,5], PhD; Spencer Stein[6], MPH; Predrag Klasnja[7], PhD; Olga Perski[1,8], PhD; Clare Viglione[1], MPH; Aaron Coleman[9], BA; Eric Hekler[1,3,4], PhD

[1]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, United States
[2]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[3]The Qualcomm Institute, University of California San Diego, La Jolla, CA, United States
[4]The Design Lab, University of California San Diego, La Jolla, CA, United States
[5]Laura Rodriguez Research Institute, Family Health Centers of San Diego, San Diego, CA, United States
[6]Spiral Health Inc, San Diego, CA, United States
[7]School of Information, University of Michigan, Ann Arbor, MI, United States
[8]Faculty of Social Sciences, Tampere University, Tampere, Finland
[9]Small Steps Labs LLC dba Fitabase Inc, San Diego, CA, United States

Corresponding Author:
Eric Hekler, PhD
Herbert Wertheim School of Public Health and Human Longevity Science
University of California San Diego
9500 Gilman Dr
MC0811
La Jolla, CA, 92093
United States
Phone: 1 858 429 9370
Email: ehekler@ucsd.edu

Abstract

Digital therapeutics (DTx) are a promising way to provide safe, effective, accessible, sustainable, scalable, and equitable approaches to advance individual and population health. However, developing and deploying DTx is inherently complex in that DTx includes multiple interacting components, such as tools to support activities like medication adherence, health behavior goal-setting or self-monitoring, and algorithms that adapt the provision of these according to individual needs that may change over time. While myriad frameworks exist for different phases of DTx development, no single framework exists to guide evidence production for DTx across its full life cycle, from initial DTx development to long-term use. To fill this gap, we propose the DTx real-world evidence (RWE) framework as a pragmatic, iterative, milestone-driven approach for developing DTx. The DTx RWE framework is derived from the 4-phase development model used for behavioral interventions, but it includes key adaptations that are specific to the unique characteristics of DTx. To ensure the highest level of fidelity to the needs of users, the framework also incorporates real-world data (RWD) across the entire life cycle of DTx development and use. The DTx RWE framework is intended for any group interested in developing and deploying DTx in real-world contexts, including those in industry, health care, public health, and academia. Moreover, entities that fund research that supports the development of DTx and agencies that regulate DTx might find the DTx RWE framework useful as they endeavor to improve how DTxcan advance individual and population health.

(J Med Internet Res 2024;26:e49208) doi:10.2196/49208

KEYWORDS

accessible; decision making; decision; decision-based evidence-making; development; digital therapeutics; medication adherence; monitoring; pharmaceuticals; public health; real-world data; real-world evidence; safe; testing; therapeutics

XSL•FO
RenderX

Introduction

Digital therapeutics (DTx) are health software tools designed to prevent, treat, or alleviate a disease, disorder, condition, or injury by delivering interventions that have demonstrable positive therapeutic effects on individual health and produce real-world outcomes [1-3]. DTx are often complex interventions [4] as they include multiple components, such as goal-setting or problem-solving elements, and algorithms that adapt the provision of support to each person's changing needs. Common goals for DTx include improving medication adherence or regular use of medical devices (eg, glucometers), facilitating behavior change, such as improving diet, physical activity, or sleep, or improving mental health, such as care for depression, anxiety, or stress. DTx can also supplement other care, such as additional support in between clinic visits. The aspiration is that DTx can improve a patient's health outcomes, reduce the burden on health care professionals, and increase access to and usability of interventions [1,5], by providing safe, effective, and equitable support for individual and population health [6]. While myriad frameworks exist for DTx development, to date, no single unifying framework guides DTx evidence production and regulatory decision-making [7-11]. By evidence production, we mean the use of scientific methods and processes to produce meaningful data about interventions, such as DTx, both qualitative and quantitative. By regulatory decision-making, we mean the set of oversight activities governing bodies such as the US Food and Drug Administration (FDA) engage in to ensure the products or services in a targeted sector (eg, the pharmaceutical industry) are safe, effective, and aligned with individual and societal needs. We believe that a framework that streamlines linkages between evidence production and regulatory decision-making for DTx will accelerate the development, adoption, and impact of DTx.

For a comprehensive framework to be successful, it must address 2 overarching issues that distinguish DTx from other therapeutics commonly used in health care. The first issue is that DTx are large pieces of software and thus benefit from decades of experience in how software is developed, used, and improved over time. The second issue relates to the relatively new regulatory environment for DTx, which has unique demands likely to evolve further as the field of DTx advances. With respect to the first issue, the basis of DTx in software renders them as dynamic entities that benefit from, and indeed require, periodic upgrades and regular maintenance to ensure they fit with ever-evolving user needs and technological changes. DTx needs to be interoperable with the constantly changing landscape of other software solutions used within health care and requires high levels of software sophistication based on enterprise-grade code embedded within a robust system architecture that supports security, privacy, and ongoing maintenance. Research-grade code is often of good enough quality to enable a novel digital health tool to be tested in small studies and efficacy trials, such as the activities commonly done by academics when engaging in frameworks like the Obesity-Related Behavioral Intervention Trials (ORBIT) [12], but it is rarely of sufficient quality to be sustainably deployed in real-world contexts. Thus, to be successful, a new framework must be able to guide appropriate evidence production that matches the inherent dynamic and often context-dependent nature of DTx.

The second overarching issue a new framework must address relates to regulatory issues. To date, regulation of DTx typically follows one of three approaches: (1) providing relatively limited guidance on evidence production, biasing toward the trustworthiness of DTx companies, as used in the FDA Precertification (Pre-Cert) Program [13]; (2) using emerging standards relevant to real-world data (RWD) and real-world evidence (RWE), such as reliance on data quality standards, use of RWD to efficiently run simulated clinical trials [14-17], and open science practices [18,19]; or (3) simply following variations of the 4-phase model [12] originally created for pharmaceuticals [20-22]. Payors are not providing adequate reimbursement systems, likely in part because of these issues, causing some DTx companies to declare bankruptcy [23]. We contend that a new framework that incorporates elements of these approaches may be helpful to multiple stakeholders.

Our Approach

Based on this background, we had 2 primary objectives for our proposed DTx RWE framework (hereafter called the "Framework"): (1) to create for users a decision-focused flowchart of key steps to develop DTx, with clear go/no-go milestones needed to move between phases of DTx RWE production that maps to the needs of regulatory decision-making (a point we return to in the discussion); and (2) to provide guidance on how to use RWD to develop this RWE. The 4-phase model is adapted from the National Institutes of Health (NIH)–supported ORBIT [12] model for behavioral intervention development, with phases focused on design, development, testing, and monitoring. We also considered it important that the Framework provide guidance on RWD use and evidence production in accordance with safe, timely, effective, efficient, equitable, and patient-centered (STEEEP) targets [24], as well as accessibility [25], sustainability [26], and scalability [27].

To accomplish this, we synthesized frameworks and best practice methods from relevant fields including, but not limited to, behavioral medicine, psychology, public health, medicine, human-centered design, human-computer interaction, bioinformatics, agile software development, computer engineering, health equity research, and community-based participatory research. We drew upon a range of different scoping reviews of frameworks for DTx evidence production (eg, Torous et al [28] and Lagan et al [29]). However, rather than using a scoping review process or other formalized expert consensus approach, first and foremost, we were guided by the issues summarized above because we consider them critical to the development of successful DTx yet underemphasized to date.

While we drew from many sources, the ORBIT model, an NIH-recognized approach for behavioral intervention development that uses 4 phases analogous to pharmaceutical development, was a foundational source. However, for our purposes, the ORBIT model had important limitations. It is set up to be broad and accommodate the development and testing of a wide range of different types of behavioral interventions.

XSL•FO
RenderX

This domain-specificity can make it challenging for those who may have limited familiarity with behavioral sciences to know how to use it for their specific needs. Additionally, the ORBIT model focuses on evidence production in support of the design of phase III efficacy trials. This is well-matched to studying novel interventions or the efficacy of technologies in ideal conditions, but it is very different from our goal of optimizing the development and sustainable deployment of DTx over time.

We also drew from expert consensus recommendations, including those from the World Health Organization (WHO) [30] and consensus statements from relevant workshops hosted by the NIH [31] within the United States, when creating the Framework. When necessary, the authors used first-hand knowledge based on their participation in expert consensus statements in related fields [7,32,33]; experience with both the research methods and community practices delineated in the Framework [8,32,34]; experience teaching graduate-level methods courses on topics covered in the Framework; and as innovators engaged in the development, use, and evaluation of novel methods explicitly created for digital health evidence production [35-37].

In addition, we drew on the FDA Pre-Cert Program, which was created to assess the credibility and readiness of a group to engage in DTx evidence production. The FDA Pre-Cert Program begins with an excellence appraisal, which aims to establish credibility by demonstrating the company's readiness through evaluating organizational excellence and a culture of quality [13]. Following that, the product goes through a streamlined review process to ensure a reasonable level of safety and effectiveness assurance, which leads to a decision on whether commercial distribution is approved. Once the product is on the market, they are asked to provide RWE based on RWD with a limited list of clinical trial designs in a specified period of time. The Pre-Cert Program ensures that companies have high standards of organizational excellence, that they carry out real-world monitoring of the software as it is used, and, critically, provide a mechanism that could be used to allow DTx groups to be reimbursed in some fashion while RWE production occurs.

The DTx RWE Framework

Overview

The Framework (Figure 1) is centered on a flowchart with 4 phases analogous to the ORBIT model but adapted for DTx RWE production (phase I: design; phase II: develop; phase III: test; and phase IV: monitor). Phase I activities correspond to the "double diamond" approach [38] used in human-centered design and related methods where the problem and solution specifications are delineated. Phase II activities are drawn primarily from ORBIT [12], its extensions [39], and the Multiphase Optimization Strategy (MOST) [40,41]. Phase III activities are based on insights from pragmatic clinical trial best practices, including Pragmatic Explanatory Continuum Indicator Summary-2 (PRECIS-2) [42], reach effectiveness, adoption, implementation, and maintenance (RE-AIM) [43,44], and recommendations from an NIH-recognized expert panel on comparator selection for behavioral interventions [45]. Phase IV activities are drawn from implementation science [46,47], and emerging recommendations on RWE use for postmarketing surveillance from the WHO [48]. The approach to RWD in the Framework draws on recent recommendations for the use of RWD for pharmaceuticals [48-51] and on recommendations on open science best practices, which are integrated into each phase, with additional suggested recommendations [52] summarized in the Discussion section.

The Framework has 4 phases as described in detail below. Sufficient resources need to be provided at both the start of this process and as it goes through the development life cycle if it is to be successful. Moreover, analogous to what we outlined above for the FDA Pre-Cert process, we recommend that groups undertaking this process commit to the following, either on their own or through one or more partnerships, as illustrated in the Framework use case provided in Multimedia Appendix 1. With this, there are 3 critical roles that must be present:

1. Designate a DTx implementor with the capability to provide ongoing, sustained deployment of the DTx. Examples could be industry, medical centers, or public health departments with proven software development and management capabilities.
2. Designate a community-serving organization that is working with and serving a target population that can provide RWD. These could be hospitals and clinics, federally qualified health centers, community clinics, or public health departments.
3. Designate a DTx evaluator with expertise in the relevant methods and approaches recommended throughout the Framework, both in terms of the flowchart of research activities and the use of RWD.

The responsibility of key stakeholders across all phases of the Framework is to intentionally consider the relevant ethical, legal, and social implications of the DTx pipeline. Including a consultant on the team who is well versed in thinking about, for example, participant characteristics and enrollment, data management (eg, collection, storage, analysis, and sharing), and related issues of bias and privacy are important throughout the process in the Framework. Engaging with an ethics review board, like a research ethics committee or institutional review board, can also be useful at various points in the process.

XSL•FO
RenderX

Figure 1. Digital therapeutics real-world evidence framework flowchart. Practical, iterative, and milestone-driven guidance to producing real-world evidence (RWE) during the design, development, testing, and monitoring phases of digital therapeutics (DTx). Oval shapes represent research or operation activities and diamond shapes represent decisions or review activities. MVP: minimal viable product.

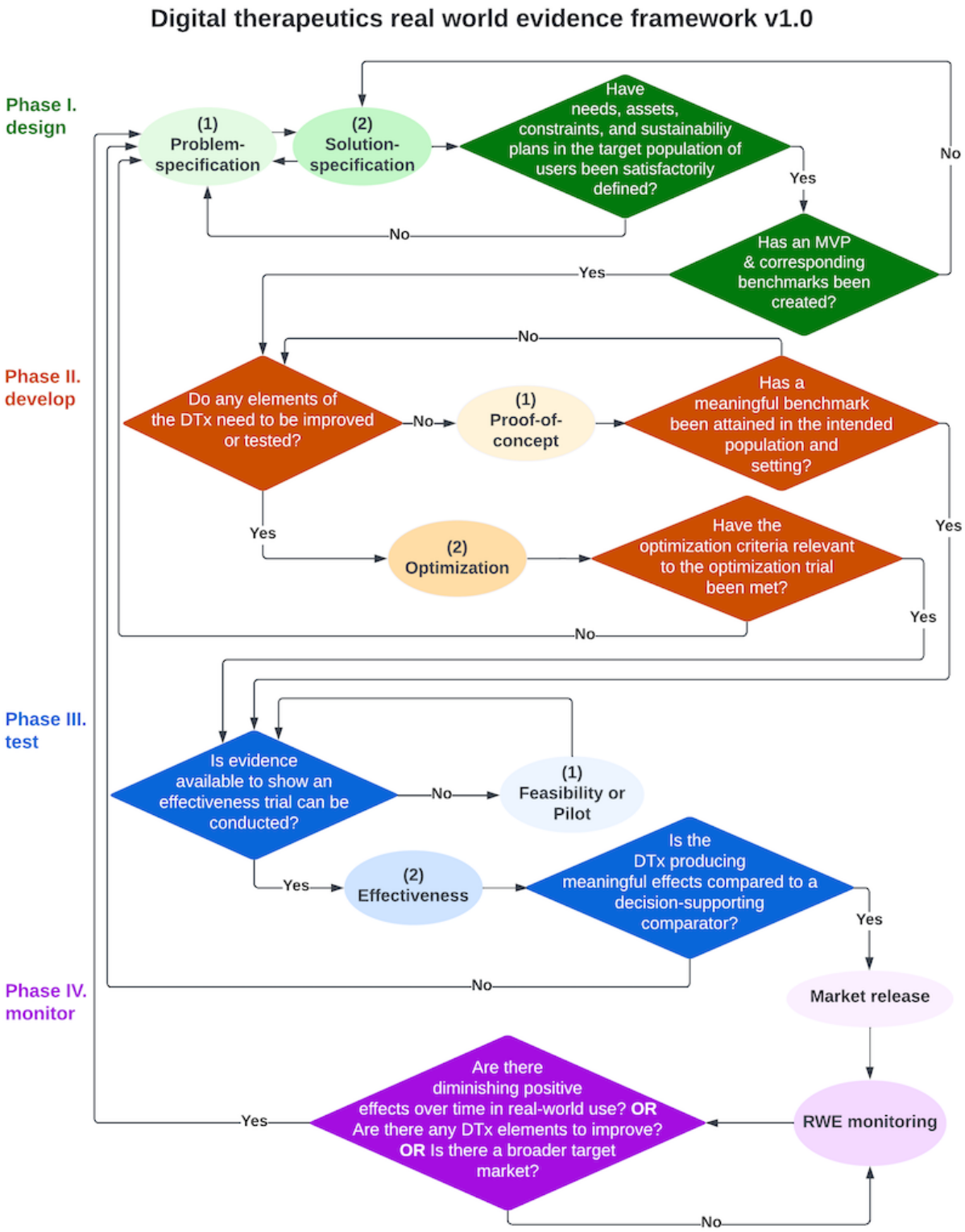

Phase I: Design

Overview

The goal of phase I is to design the DTx product. The two activities work iteratively together: (1) problem-specification and (2) solution-specification, in accordance with the well-recognized "double diamond" used in human-centered design [38]. Problem specification includes delineating real-world needs, constraints, assets and approaches to support future sustainability, both financial and ecological [53,54]. Solution specification focuses on iteratively creating a DTx through agile development and prototyping from low-fidelity (eg, paper concepts and storyboards) to high-fidelity (fully functional code) prototypes, with this iterative work being

XSL•FO
RenderX

situated within ethical practices and exploring the potential for reusing components or functionalities from available DTx where possible, in line with open science practices [55].

Intervention plausibility claims are the key focus of phases I and II. We are explicitly using the word "plausible" instead of "feasible" given emerging nomenclature recommendations. By "intervention plausibility," we are referring to context-dependent probabilistic claims regarding the interaction between the DTx and targeted populations and settings, such as acceptability, demand, capacity to be adapted to a local context, etc. (a full list of plausibility targets is available at [56]; note: it was labeled feasibility in this paper given that it was published before emerging nomenclature recommendations). Progress can be determined in phase I through the specification of benchmarks that conform with the specific, measurable, actionable, realistic, and timely (SMART) goal concept [57] but are applied to establishing benchmarks. The transition from phase I to II is justifiable when benchmarks relevant to, at a minimum, safety, and effectiveness are defined and a minimal viable product (MVP) DTx has been produced that meets basic ethical, functional, and usability requirements.

Question 1a: Have Needs, Assets, Constraints, and Sustainability Plans in the Target Population of Users Been Satisfactorily Defined?

This work specifies the target population and setting and answers other key questions relevant to DTx [7]. Foundational to this is the RWD from a community-serving organization to identify unmet needs, current assets (eg, standard practices and billable activities related to the targeted DTx), and constraints (eg, likely number of billable sessions and the scope of real-world needs, if scaled). Beyond RWD, a mixed methods approach is recommended for this stage. This includes formative research such as ethnographic studies, focus groups, and interviews; community-based participatory and community-driven methods; literature reviews, such as reviewing previous intervention and epidemiological studies; and market research and analysis. Determination of whether these have been satisfactorily defined can be achieved by reaching a consensus to determine if there is overarching agreement among stakeholders, which involves a decision being made when no one objects [58].

Question 1b: Have an MVP and Corresponding Benchmarks Been Created?

The focus is to iteratively build the DTx and finalize benchmarks. In terms of DTx creation, movement to phase II is justified when a group has produced evidence showing that an MVP is functioning according to minimal usability and accessibility requirements and meets a threshold of being plausible as a tool to enable phase II activities. This could be demonstrated through DTx that are free of "bugs" and meet basic usability requirements (eg, good System Usability Score [59]). Ideally, benchmarks are established in relation to STEEEP targets [24] as well as accessibility [25], sustainability [26], and scalability [27]. To guide future implementation, data could be gathered about critical implementation issues [56]. Minimal benchmarks are needed for safety and effectiveness.

When setting benchmarks, the team should balance what is meaningful relative to current best practices with what is plausible to achieve for the target population and setting [60]. These benchmarks should, ideally, be based on RWD and establish a threshold that defines if the proposed DTx could plausibly produce desired effects safely relative to current practices. Examples of benchmarks include "a decrease of 3% in hemoglobin A_{1c} among 60% of our DTx users after 6 months of intervention" for effectiveness and "less than 10% of our DTx users experienced nonserious adverse events associated with digital treatment after a month of intervention" for safety. With a functioning MVP and benchmarks defined, the team has produced the requisite information to transition to phase II. Key approaches here include agile or lean development practices, prototype testing and development, rapid prototype testing, and qualitative methods [61]. See Multimedia Appendix 1 for an example.

Phase II: Develop

Overview

The goal of phase II is to guide DTx development and optimization. There are two types of activities: (1) proof-of-concept trials and (2) optimization trials. Movement from phase II to phase III requires RWE—produced either from a proof-of-concept trial or optimization trial—that demonstrates it is plausible that the DTx can produce clinically meaningful effects in the targeted population and setting while meeting requisite implementation requirements. Given that ORBIT is our primary starting point, it is important to flag that we are rearranging where these trials take place. Specifically, in ORBIT, optimization trials occur late in phase I (ie, phase Ib). In contrast, proof-of-concept trials occur in phase IIa in ORBIT. In the authors' opinion, the approach used in ORBIT creates 2 issues. First, it places many important activities as phase I. Second, it implicitly signals that phase I and II activities are subservient to phase III activities (a point we return to when discussing phase III of the Framework). In the Framework, we unpack activities in phase I of ORBIT to spread them across phases I and II and provide more explicit labels of the key purposes for each phase, namely design and development. With this, our intent was to provide clearer guidance on how one progresses between phases and also to connote the unique value of each phase without any need for one to be subservient to another; instead, the phase is selected based on the type of evidence production needed.

Within this phase, RWD should be used to support targeted recruitment and selection of study participants with a particular eye toward accounting for health equity in defining a target population (eg, targeting a population that explicitly underuses the current standard of care). In addition, RWD can be used to monitor for unintended consequences, both positive and negative, of the use of the DTx within the development trials. RWD could also be used, particularly with deployed DTx, for conducting data-driven algorithm development [62,63].

XSL•FO
RenderX

Question 2a: Do Any Elements of the DTx Need to Be Improved or Tested?

The proof-of-concept trial focuses primarily on intervention plausibility testing about if the overall package is producing meaningful results, as defined through the benchmarks established in phase I. Optimization trials produce evidence to improve elements of the DTx. For example, optimization trials can be used to support the evidence-based selection of DTx components (factorial trials) [64-68], refinement of components, particularly those used across time (microrandomized trials) [35,69,70], and refinement of adaptation algorithms to match the provision of support to context, individual differences, and timing (microrandomized trials [35,71], and sequential multiple assignment randomized trials [72-75]). Note also that optimization trials could also be conducted that are explicitly used to support algorithm development (eg, system identification experiments [76-78] or more data-driven algorithm development from RWD [79,80]). Like phase I, proof-of-concept and optimization trials can be used iteratively. Determining if optimization is needed is based on whether any element of a DTx needs to be improved. If no elements of the DTx package need to be refined, then a proof-of-concept trial is appropriate. If some element of the DTx package needs to be tested or improved, then an optimization trial is needed.

Question 2b: Has a Meaningful Benchmark Been Attained in the Intended Population and Setting?

This is a fundamental question for the proof-of-concept trial, an emerging approach in behavioral trials that tests a DTx in a small group (eg, 10-20) in relation to a benchmark (eg, 70% of the 10 patients shift from hypertensive to systolic blood pressure <120 mm Hg). This approach is used because of 2 known issues with small sample trials within formative work. First, small samples render the use of frequentist inferential statistics problematic [60,81,82]. Second, humans have confirmation bias, which refers to the tendency for individuals to seek out or interpret evidence in ways that align with previous beliefs, expectations, or hopes [83]. A proof-of-concept trial overcomes these challenges without the need for running larger trials, using a clearly specified a priori benchmark that can be tested using descriptive statistics. Its use of benchmarks enables resource-efficient studies with clear go/no go decision-making that reduces the risk of falling prey to confirmatory bias. Quasi-experimental, single-case, or within-participant designs in which the participants serve as their own controls are appropriate design options for a proof-of-concept trial. Further, mixed methods, where both qualitative and quantitative data are relevant to the goals of evidence production and real-world intervention implementation plausibility [56] targets, should be used. Proof-of-concept trials provide clear go/no go milestones established a priori, thus reducing the risk of continuing when not justified, which is common with more traditional piloting study approaches [39]. If benchmarks are met, the group can either shift to an optimization trial or transition to phase III. If the benchmarks are not met, then the team should consider returning to phase I or focusing on DTx optimization.

Question 2c: Have the Optimization Criteria Relevant to the Optimization Trial Been Met?

Optimization is a concept drawn from engineering that emphasizes data-driven improvement for a DTx [41,84]. Optimization supports any problem arising with the DTx, such as the DTx costing too much, not being sufficiently adhered to, being difficult to implement in real-world contexts, being inaccessible to the target population, etc. Like a benchmark, the key logic here is to specify clear optimization criteria, meaning a definition of success that can be tested using an appropriately selected optimization trial. Specification of optimization criteria is a central focus of MOST [41,84]. The goal is to create optimization criteria that are measurable and, ideally, account for real-world constraints. For example, one could establish optimization criteria that a DTx includes only intervention components with demonstrated effectiveness, that the DTx can be deployed for under US $50 per client, or that total interactions per week with the DTx stay below 30 minutes. These criteria can be translated into clear go/no go criteria that can be assessed in an optimization trial. If the optimization criteria are met, then this can often justify movement to phase III. Plausible optimization trials for this phase could include but are not limited to: A/B testing (as used in the technology industry for improving usability) [85-87], factorial trials as used in MOST [64,65,67,68], sequential multiple assigned randomized trials [72-75], microrandomized trials [35,36,69-71], system identification experiments [76-78], studies explicitly designed to support algorithm development [79,80], and control optimization trials [37,88]. Nahum-Shani et al [40] provide guidance on when to use common optimization trial designs. See Multimedia Appendix 1 for an example.

Phase III: Test

Overview

The goal of phase III is to test if the DTx produces meaningful improvements relative to a comparator in real-world contexts. There may be two types of activities: (1) feasibility or pilot studies; or (2) an effectiveness trial. As in phase II, we have shifted phase labeling from the original ORBIT model while still honoring the types of evidence production ORBIT generally advocates for. Specifically, in the ORBIT model, feasibility or pilot studies are conducted in phase IIb. In ORBIT, phase III is reserved purely for an efficacy trial to test if the intervention impacts health outcomes. We relabeled each phase intentionally in the Framework to allow each phase of work to be conducted and produce insights that are valuable alone, with no phase treated as subservient to other phases. Thus, we sought to have development phase activities stand alone in terms of their unique value for evidence production. We recommend this shift in thinking to clearly flag the critical, independent importance of each phase of work, particularly phase II development, which could feasibly be used in perpetuity alone as a rigorous approach to continuous quality improvement. While speculative, we contend that ORBIT and related evidence production models that implicitly or explicitly treat earlier phases as subservient to phase III trials, particularly efficacy trials in ideal conditions, send a message that privileges one type of evidence at the expense of other evidence. This is problematic, as the evidence

XSL•FO
RenderX

from the other phases is particularly important for fostering real-world implementation and health equity. Further, privileging one type of evidence over others establishes the risk of a mono-method bias within scientific knowledge. This creates issues with fostering trustworthy scientific knowledge [89], reducing confidence in any consensus statements from overemphasizing one particular type of evidence, and might slow the pace of learning and progress [90]. This is particularly true when evidence production privileges tests occurring in ideal conditions, which is a valuable focus for novel interventions. When groups are developing novel interventions, they should use the ORBIT model, given its emphasis on ensuring appropriate evidence is produced to complete a high-quality efficacy trial. With phase III of the Framework, our focus is on testing if a DTx package produces meaningful results in real-world contexts to foster evidence production with high ecological validity. Thus, in the Framework we bias toward pragmatism through a focus on effectiveness trials, inclusion of benchmarks added to a modified CONSORT (Consolidated Standards of Reporting Trials) diagram guided by RE-AIM [91] to justify generalization claims, use of decision-oriented comparator selection, and emerging best practices on power calculations [39]. Moreover, the use of hybrid clinical trials, which incorporate the focus on testing both effectiveness and implementation outcomes [92,93] can be another option to be used in phase III testing of the Framework.

Even within the context of RWE, we recognize the classic tension between pragmatism and explanatory knowledge, as illustrated in PRECIS-2 [42] (eg, recruitment, eligibility, and setting). With context-invariant interventions, such as vaccines, explanatory knowledge tends to be highly valuable for producing both robust internal validity and generalizable knowledge. Further, with highly novel interventions, explanatory knowledge is valuable to determine if a signal is present in ideal conditions for the novel intervention. For evidence production to guide regulation of use in the real-world context of DTx, we suggest a bias toward pragmaticism over explanatory knowledge to support ecologically valid knowledge as an approach to increase the likelihood of generalizable knowledge. Recognizing that, just like in traditional trials, this tension must be balanced in each trial conducted based on the goals of the work and what is already known. RWD should be used to support recruitment, with a particular eye toward accounting for health equity in recruiting truly representative samples. Confidence in any claims of representation and, thus, generalizability, from the trial can be supported by clearly defining meaningful while also achievable benchmarks for reach and adoption, which can be measured and reported in a modified CONSORT diagram [91]. Furthermore, once a DTx is widely deployed, RWD could be used to run simulated clinical trials to test effectiveness in real-world contexts [14-17].

Question 3a: Is Evidence Available to Show an Effectiveness Trial Can Be Conducted?

This is a key question for feasibility or pilot studies, which are used to pave the way for a future effectiveness trial. A feasibility study examines whether and how a proposed or planned effectiveness trial can be done, but without a requirement for resemblance to the future trial [39]. Note, we are explicitly using the word “feasibility” only to refer to attributes of a targeted future fully powered trial, in alignment with 2010 CONSORT recommendations on clinical trial nomenclature [94,95]. The word “feasibility” is sometimes used in reference to issues about the intervention, such as if it would be acceptable, have sufficient demand, or could be integrated into real-world contexts [56]. To avoid confusion, we refer to these targets as intervention plausibility. Emerging recommendations from scientific groups, such as the 2010 CONSORT recommendations and others more relevant to digital health [96,97], suggest that the term “feasibility” be used to describe the probability that a particular type of study, such as a phase III clinical trial, can be conducted with sufficient rigor and fidelity by an investigative team in a given context. Based on this, we will honor this emerging consensus and only use the term “feasibility” to refer to context-dependent probabilistic claims in relation to the likelihood that a targeted study can be conducted in a particular setting by a particular group with sufficient fidelity to allow conclusions to be drawn from it. Similarly, we use emerging naming conventions and reserve the term “pilot” to refer to a specific type of feasibility trial that implements the exact eventual full protocol of a fully powered trial, but with fewer participants. The goal of a pilot study, thus, is to gather information about the likelihood that if a full trial were conducted, the data quality would be sufficient to enable trustworthy inferences [39]. If the evidence is available to show the feasibility of an effectiveness trial, the team can proceed directly to the effectiveness trial. If this evidence is not available, then the team should consider conducting a feasibility or pilot study before effectiveness trials to increase the likelihood that, if a trial is run, it will be conducted with sufficient fidelity to provide sufficient quality evidence to guide decisions.

Question 3b: Is the DTx Producing Meaningful Effects Compared to a Decision-Supporting Comparator?

The complexities of DTx establish higher evidentiary standards for generalization claims. By generalization, we specifically focus on transportability, meaning the degree to which insights gleaned from a given study sample are relevant to a stated population and setting. Recent studies of health behaviors, cognitive processes, and emotions, all factors that may influence the effectiveness of a DTx, show that behaviors, cognitions, and emotions have multiple contextual influences, can differ widely from person to person, and fluctuate over time within individuals [50,98-103]. Thus, research studies intended to produce generalizable knowledge about the effectiveness of DTx, including where, for whom, and when a given DTx is useful, need to take these factors into account. Given these complexities with regard to DTx evidence production, transportability is difficult to establish. We suggest, first, the use of a modified CONSORT diagram that integrates insights from the RE-AIM framework, and second, benchmarks be established relevant to the percentage of plausible settings and participants enrolled and completing the study. This modified CONSORT diagram provides an approach to quantifying the degree to which a sample may or may not be representative of a stated population and setting. If, for example, there is a large disparity between an eligible population and the number of patients who are enrolled, then transportability claims would

XSL•FO
RenderX

be questionable. As with other phases in this Framework, it is suggested that achievable benchmarks related to the modified CONSORT diagram be specified a priori (eg, 80% of eligible clinics will take part, 80% of eligible staff will take part, and 50% of eligible participants will be enrolled). As before, these benchmarks need to balance the need to be ambitious while also being achievable based on what is known. Only if the benchmarks are met can transportability claims be justified. These benchmarks on the modified CONSORT diagram should be derived from RWD.

As this is an effectiveness trial, comparator selection should support real-world decision-making. For example, if the clinical or community partner has a current standard of care and they are considering replacing it with the DTx, the standard of care should be the comparator. Alternatively, if the DTx would fulfill a new area of need, a stepped wedge trial [104,105], in which the DTx is released in a phased fashion across clinics, could be considered. For detailed guidance on comparator selection, see NIH expert panel recommendations [45]. For DTx testing, options include but are not limited to between-person RCTs [106,107], including remote RCTs [108,109], cluster RCTs [110,111], stepped wedge trial [104,105], and, when sufficient RWD is available, the use of simulated clinical trials [14-17] could be considered.

We also recommend the use of best-practice recommendations for power calculations that specifically do not rely upon underpowered studies to infer effect sizes [60,112,113]. Instead, what is recommended is to establish 2 effect size estimates, a threshold of clinical significance, and a plausible effect size that could be observed in the trial. The threshold of clinical significance is the smallest effect size of interest [114] that would influence clinical decision-making based on an explicit qualitative determination of a noticeable difference. This threshold of clinical significance should be informed by RWD and can be translated from the benchmarks for effectiveness defined in phase I. The plausible effect size is the most likely effect size to be observed if the trial were conducted. This plausible effect size can be informed, in part, from the results of the proof-of-concept trial, particularly if the benchmark that was met is well matched to the threshold of clinical significance. That said, given the unreliability of small sample sizes, RWD and effect sizes from previous trials most like the proposed study should be used to establish the plausible effect size. If the plausible effect size is at or above the threshold of clinical significance, then a trial is warranted. If the plausible effect size is below the threshold of clinical significance, then the trial should not be conducted, as the results gathered would not be sufficient to make a convincing argument to change clinical practice.

If benchmarks set to the modified CONSORT diagram are not met or there is no clinically significant difference observed between the DTx and comparator, then returning to phase I or phase II activities is appropriate. If benchmark and clinically meaningful differences are observed, then results can be submitted to regulators for official review. If the trial was done ethically and responsibly and the results are positive, then regulating bodies can certify the DTx and allow the DTx to market to the population and setting that was studied within the phase III trial. See Multimedia Appendix 1 for an illustrative example.

Phase IV: Monitor

Overview

The goal of phase IV is to monitor the use of the DTx within the real world, enable DTx implementers to improve the DTx with additional RWD collected from their clinical or community partners and support the expansion of the target market for the DTx through stepwise additional assessments. This is analogous to traditional non-DTx phase IV activities, including RWD use with pharmaceuticals [48].

Question 4: Are There Diminishing Positive Effects Over Time in Real-World Use? Are There Any DTx Elements to Improve? Is There a Broader Target Market?

As described earlier and elsewhere [7,32,115], continuous improvements in the DTx are not only desired but required for a DTx. For example, user interface expectations of technologies and the use of application programming interfaces drive the evolution of technology. A web application designed and tested in the 1990s [116,117], if it was not continually updated to meet changing user expectations and remain up-to-date with related application programming interfaces, would, at best, be perceived as “old” and, at worst, would not work. Thus, prespecified quality control methods for these updates would need to be used, and this could be one area where regulatory guidelines could be helpful. For example, there is active discussion about when the accumulative changes to a DTx warrant running another clinical trial [118].

Given this, any notion of a “definitive” clinical trial, a concept traditionally used, is inappropriate for DTx. Pragmatic ways of gleaning insights about when DTx is meeting expectations across the STEEEP and related criteria listed earlier are critical to monitor over time. To support this, the implementation of science practices, particularly strategies for ongoing monitoring, thoughtful adaptation, and guidance on rigorous continuous quality improvement, can be gleaned from the dynamic sustainability framework [46]. Through ongoing monitoring, issues of potential diminishing benefit (labeled voltage drop) can be observed and used to inspire a response [22]. For example, if effectiveness levels go below some predefined threshold, regulators could provide DTx with a time-limited window for continued marketing while also requiring the DTx company to reestablish a partnership with a clinical or community partner and reengage with earlier stages of the process. With this potential risk looming, it could establish an incentive for the DTx company to engage in continuous improvement, guided by the other 2 questions, and to maintain mutually beneficial partnerships.

These recommendations conform with recommendations from the WHO for monitoring and evaluating digital health interventions [119]. According to the WHO, the 4 major components of digital health monitoring (ie, functionality, stability, fidelity, and quality) should guide ongoing monitoring, with these questions mapping onto our proposed questions:

XSL•FO
RenderX

- Are there diminishing positive effects over time in real-world use?
  - (Quality) Is the content and the delivery of the intervention of high enough quality to yield intended outcomes?
  - (Quality) How well and consistently is the intervention delivered?
- Are there any DTx elements to improve?
  - (Functionality) Does the system operate as intended?
  - (Stability) Does the system consistently operate as intended?
- Is there a broader target market?
  - (Fidelity) Do the realities of field implementation alter the functionality and stability of the system, changing the intervention from that which was intended?

RWE for postmarket surveillance is being explored, and opportunities and pitfalls that are also relevant to DTx are being articulated and should be considered regarding DTx regulation [48]. Monitoring could require benchmarks to be set for all key targets of evidence production (eg, effectiveness, safety, and equity) as one pathway for cultivating more rigor in monitoring efforts and reducing the risk of confirmation bias during phase IV.

If the DTx implementer believes their DTx can support a more diverse market share than what was approved in phase III, RWD collected in phase IV may help the DTx implementer accelerate this expansion. For example, monitoring could be used to identify plausible new populations, settings, or areas for improvement of the DTx, particularly if done with other community-serving organization partners who may have providers prescribing the DTx for "off-label" uses. One plausible way to improve evidence production during phase IV monitoring would be to focus evidence production more on testing and improving the elements of a DTx (eg, intervention components and adaptation algorithms) instead of the DTx package. A more detailed rationale for this is described elsewhere [7,32,47,82,120]. A second opportunity would be to link activities and efforts with ongoing behavioral ontology efforts to foster better knowledge comparison across various DTx [121,122]. With that said, standards for ongoing monitoring of RWD and RWE are rapidly evolving; thus, this is a critical area for continued work. See Multimedia Appendix 1 for an illustrative example.

Discussion

Overview

The Framework provides guidance to groups seeking to sustainably deploy DTx for use in real-world contexts and may be helpful to regulatory and funding entities as they provide support and oversight of DTx. We acknowledge that the Framework has not been rigorously vetted and that additional work is needed to establish its value. This includes determining whether the use of the Framework has greater or lesser utility in specific domains of DTx applications, such as those used in mental health, behavioral health, or as an adjunct to pharmacological and other interventions for chronic diseases like cancer, musculoskeletal disorders, and cardiovascular disease. We know of no specific reasons why such differences should exist, but as published reports of DTx research emerge in the future, these distinctions might become evident.

Another issue pertains to how the Framework can assist with evaluating DTx that are already in the field, including digital wellness tools that do not meet the definition of a DTx. The longer a DTx has been sustainably deployed at scale, the more likely it is that simulated clinical trial methods could be used to study DTx along STEEEP criteria. With this, efficiencies could be further advanced for evidence production through simulated clinical trials through RWD. Future work would benefit from continued focus on the refined development of simulated clinical trial best practices to improve the pace and resource efficiency of learning.

Regarding digital health wellness tools, these tools often build on foundational behavior change techniques, such as self-monitoring and goal setting that have decades of evidence to guide their design and implementation. When situations like this exist, the burden of proof should be to justify why current evidence is not sufficient already. The most likely gaps in research for these may relate to insufficient evidence for their effectiveness in a broad range of users or settings, so a targeted adaptation of the Framework to fill in these gaps might be the best approach. For example, future work could explore ways to adapt the Framework for use with community-based organizations and community-serving well-being institutions such as the YMCA or Jewish Family Services, along with corporate wellness and related wellness programs that are not implemented by or in partnership with the health care system.

With these limitations recognized, we expand below on the need for three areas of future work related to the Framework: (1) how using RWD advances health equity; (2) cultivating trustworthy partnerships that foster the use of Framework, as a secondary pathway to advance health equity; and (3) suggesting next steps with regard to regulation and funding.

Advancing Health Equity Through RWD

Overview

In our view, increased sophistication on the effective use of RWD can become a critical tool to overcome some of the major challenges currently faced in health care, including identifying and addressing health disparities to advance health equity for all, and to foster more targeted and resource-efficient evidence production. RWD provides the information needed to specify unmet needs in general as well as those for individuals, communities, and populations where current practices are not producing desired results. This can help focus resource expenditures and efforts to reduce health disparities.

Future work could advance the use of RWD to drive the development of evidence-based solutions that serve communities most in need. Clinically meaningful benchmarks based on RWD provide an approach for guiding DTx development, both for individual DTxs and for DTxs at large. These would create pressure for DTx not simply to replicate existing standards of care but to improve upon them. Indeed, RWD can be used to

XSL•FO
RenderX

establish benchmarks across the various evidence production targets, such as effectiveness, safety, and equity, to provide the foundational data needed to measure, monitor, and, thus, drive equitable progress in individual and population health.

Cultivating Trustworthy Partnerships

As presented in the section above introducing the Framework, we recommend a tripartite approach to its use, comprising an entity committed to sustaining the DTx, a community-serving organization from which the RWD comes, and an entity with appropriate expertise in RWE evaluation efforts. While these conditions can be met within well-resourced settings such as academic medical centers, we suggest that there are opportunities for implementing the Framework through partnerships among groups that may historically not have worked as closely together, such as industry partners working with federally qualified health centers and supported with an academic partner, as illustrated in Multimedia Appendix 1.

The trustworthiness of all actors involved must be acknowledged as a foundational starting point for any approach to evidence production [123,124]. This includes not merely thinking that trust can be achieved with effective communication but that, at its core, trust involves acknowledging and centering ethics, inclusion, and equity as central guiding principles in the work [89,123]. To do this, we propose the use of best practices in cultivating and maintaining partnerships that have already been delineated like community-based participatory research [125,126], patient-led innovation [127,128], community-driven design [129], community psychology practices [130,131], and ethical digital health research practices [33,34,132,133]. Incorporation of approaches to determining corporate trustworthiness that was formatively tested in the FDA Pre-Cert program can be used, including excellence appraisal, and streamlined review elements (eg, real-world performance plan and review determination information) [13]. The Digital Health Checklist [133] might also be helpful to guide ethical practices for evidence production relevant to DTx pertaining to issues such as accessibility, privacy, data management, balancing risks, and benefits, all grounded in fundamental ethical principles including respect for persons, beneficence, justice, and respect for law and the public interest.

Regulatory and Funding Issues

We encourage regulators and funders of DTx to explore whether the Framework can help guide their efforts. The principles embodied in the Framework could be used to establish generalized regulatory expectations for DTx. Clarifying these expectations could help get multiple DTx developers and purchases “on the same page” with respect to achieving and maintaining appropriate standards of quality throughout the life cycle of DTx use. Similarly, funders of DTx research and development, such as the NIH, the Agency for Healthcare Research and Quality, the Patient-Centered Outcomes Research Institute, and the Health Resources and Services Administration, could encourage applicants to use the Framework, and if they do, then demonstrate how they propose to achieve the benchmarks that it includes.

Conclusion

The Framework is intended to improve evidence production and sustainable deployment of DTx in real-world contexts. The Framework provides guidance on how to design, develop, test, and monitor DTx, both in the early stages of their development and over time as they are used in real-world contexts. Our hope is that the Framework can help address issues commonly seen with DTx, including low DTx uptake, long-term sustainability, and insufficient attention to health disparities. Overall, there is considerable opportunity to improve individual and population health equitably through DTx, and we hope the Framework can contribute to this end.

Acknowledgments

We want to thank Dr Kenneth Freedland for providing expert consultation and review of earlier versions of this manuscript. This study is supported by the National Library of Medicine (R01LM013107) and the National Cancer Institute (R01CA244777) of the National Institute of Health. CN was supported by the Patient-Centered Outcomes Research Institute (PCORI; award ME-2020C3-21310) and the Altman Clinical and Translational Research Institute (ACTRI) at the University of California, San Diego. The ACTRI is funded from awards issued by the National Center for Advancing Translational Sciences (NIH UL1TR001442). PK is supported by the National Cancer Institute (1U01CA229445) of the National Institute of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or PCORI.

Authors' Contributions

MK, KP, CN, and EH contributed to conceptualizing, drafting, reviewing, and finalizing the paper. All authors contributed to critical reviews and revisions of the paper and have given approval for the paper to be submitted for publication. The authors read and approved the final manuscript.

Conflicts of Interest

SS has an ownership interest in Spiral Health, Inc. AC has an ownership interest in Fitabase, Inc. OP acts as an unpaid scientific advisor for the Smoke Free app. All other authors declare no competing interests.

Multimedia Appendix 1

XSL•FO
RenderX

The hypothetical example of DTx RWE Framework.
[DOCX File , 34 KB - jmir_v26i1e49208_app1.docx]

References

1. Understanding DTx: a new category of medicine. Digital Therapeutics Alliance. 2022. URL: https://dtxalliance.org/understanding-dtx/ [accessed 2024-02-02]
2. Digital therapeutics get a brand new definition. Healthcare Brew. 2023. URL: https://tinyurl.com/y4uyr9ww [accessed 2024-02-02]
3. ISO/TR 11147:2023(en) health informatics—personalized digital health—digital therapeutics health software systems. International Organization for Standardization (ISO). 2023. URL: https://tinyurl.com/bddx8vtx [accessed 2024-02-02]
4. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
5. Patel NA, Butte AJ. Characteristics and challenges of the clinical pipeline of digital therapeutics. NPJ Digit Med 2020;3(1):159 [FREE Full text] [doi: 10.1038/s41746-020-00370-8] [Medline: 33311567]
6. Dang A, Arora D, Rane P. Role of digital therapeutics and the changing future of healthcare. J Family Med Prim Care 2020;9(5):2207-2213 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_105_20] [Medline: 32754475]
7. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
8. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
9. Voorheis P, Zhao A, Kuluski K, Pham Q, Scott T, Sztur P, et al. Integrating behavioral science and design thinking to develop mobile health interventions: systematic scoping review. JMIR Mhealth Uhealth 2022;10(3):e35799 [FREE Full text] [doi: 10.2196/35799] [Medline: 35293871]
10. Torous J. Evaluating apps: making informed decisions with available real-world data. Digital health interventions from wellness to therapeutics: development and dissemination. 2022 Presented at: National Institutes of Health, Digital Health Intervention Workshop; July 13, 2022; NA.
11. Miao BY, Arneson D, Wang M, Butte AJ. Open challenges in developing digital therapeutics in the United States. PLOS Digit Health 2022;1(1):e0000008 [FREE Full text] [doi: 10.1371/journal.pdig.0000008] [Medline: 36812515]
12. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015;34(10):971-982. [doi: 10.1037/hea0000161] [Medline: 25642841]
13. The Software Precertification (Pre-Cert) pilot program: tailored total product lifecycle approaches and key findings. U.S. Food & Drug Administration. 2022. URL: https://tinyurl.com/ycwsjh8u [accessed 2024-02-02]
14. Chen Z, Zhang H, Guo Y, George TJ, Prosperi M, Hogan WR, et al. Exploring the feasibility of using real-world data from a large clinical data research network to simulate clinical trials of Alzheimer's disease. NPJ Digit Med 2021;4(1):84 [FREE Full text] [doi: 10.1038/s41746-021-00452-1] [Medline: 33990663]
15. Swift B, Jain L, White C, Chandrasekaran V, Bhandari A, Hughes DA, et al. Innovation at the intersection of clinical trials and real-world data science to advance patient care. Clin Transl Sci 2018;11(5):450-460 [FREE Full text] [doi: 10.1111/cts.12559] [Medline: 29768712]
16. Müller P, Chandra NK, Sarkar A. Bayesian approaches to include real-world data in clinical studies. Philos Trans A Math Phys Eng Sci 2023;381(2247):20220158 [FREE Full text] [doi: 10.1098/rsta.2022.0158] [Medline: 36970825]
17. Wang C, Rosner GL. A bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence. Stat Med 2019;38(14):2573-2588. [doi: 10.1002/sim.8134] [Medline: 30883861]
18. Kokol P. Agile software development in healthcare: a synthetic scoping review. Appl Sci 2022;12(19):9462 [FREE Full text] [doi: 10.3390/app12199462]
19. Van Velthoven MH, Smith J, Wells G, Brindley D. Digital health app development standards: a systematic review protocol. BMJ Open 2018;8(8):e022969 [FREE Full text] [doi: 10.1136/bmjopen-2018-022969]
20. Basic principles on use of registries in approval applications. Pharmaceuticals and Medical Devices Agency. URL: https://www.pmda.go.jp/files/000240810.pdf [accessed 2024-02-02]
21. Guideline on the application of real world evidence of medical device. Ministry of Food and Drug Safety (MFDS). 2019. URL: https://tinyurl.com/mr3zpj85 [accessed 2024-02-02]
22. Framework for FDA's real-world evidence program. U.S. Food & Drug Administration. 2018. URL: https://www.fda.gov/media/120060/download [accessed 2024-02-02]
23. Pear therapeutics assets sold for $6M at auction after bankruptcy. MobiHealthNews. 2023. URL: https://tinyurl.com/4v7b87kk [accessed 2024-02-02]

XSL•FO
RenderX

24. Six domains of healthcare quality. Agency for Healthcare Research and Quality. 2022. URL: https://www.ahrq.gov/talkingquality/measures/six-domains.html [accessed 2024-02-02]
25. Gulliford M, Figueroa-Munoz J, Morgan M, Hughes D, Gibson B, Beech R, et al. What does 'access to health care' mean? J Health Serv Res Policy 2002;7(3):186-188. [doi: 10.1258/135581902760082517] [Medline: 12171751]
26. McCool J, Dobson R, Muinga N, Paton C, Pagliari C, Agawal S, et al. Factors influencing the sustainability of digital health interventions in low-resource settings: lessons from five countries. J Glob Health 2020;10(2):020396 [FREE Full text] [doi: 10.7189/jogh.10.020396] [Medline: 33274059]
27. Milat A, Lee K, Conte K, Grunseit A, Wolfenden L, van Nassau F, et al. Intervention scalability assessment tool: a decision support tool for health policy makers and implementers. Health Res Policy Syst 2020;18(1):1 [FREE Full text] [doi: 10.1186/s12961-019-0494-2] [Medline: 31900230]
28. Torous JB, Chan SR, Gipson SYMT, Kim JW, Nguyen TQ, Luo J, et al. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. Psychiatr Serv 2018;69(5):498-500 [FREE Full text] [doi: 10.1176/appi.ps.201700423] [Medline: 29446337]
29. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open 2021;11(3):e047001 [FREE Full text] [doi: 10.1136/bmjopen-2020-047001] [Medline: 33741674]
30. Consensus statement: role of policy-makers and health care leaders in implementation of the Global Patient Safety Action Plan 2021–2030. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/WHO-UHL-IHS-PSF-2022.3 [accessed 2024-02-02]
31. NIH consensus statement. National Institutes of Health (NIH). 2000. URL: https://tinyurl.com/4txypc3d [accessed 2024-02-02]
32. Patrick K, Hekler EB, Estrin D, Mohr DC, Riper H, Crane D, et al. The pace of technologic change: implications for digital health behavior intervention research. Am J Prev Med 2016;51(5):816-824. [doi: 10.1016/j.amepre.2016.05.001] [Medline: 27745681]
33. Nebeker C, Torous J, Ellis RJB. Building the case for actionable ethics in digital health research supported by artificial intelligence. BMC Med 2019;17(1):137 [FREE Full text] [doi: 10.1186/s12916-019-1377-7] [Medline: 31311535]
34. Nebeker C, Ellis RJB, Torous J. Development of a decision-making checklist tool to support technology selection in digital health research. Transl Behav Med 2020;10(4):1004-1015 [FREE Full text] [doi: 10.1093/tbm/ibz074] [Medline: 31120511]
35. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015;34S:1220-1228 [FREE Full text] [doi: 10.1037/hea0000305] [Medline: 26651463]
36. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019;53(6):573-582 [FREE Full text] [doi: 10.1093/abm/kay067] [Medline: 30192907]
37. Hekler EB, Rivera DE, Martin CA, Phatak SS, Freigoun MT, Korinek E, et al. Tutorial for using control systems engineering to optimize adaptive mobile health interventions. J Med Internet Res 2018;20(6):e214 [FREE Full text] [doi: 10.2196/jmir.8622] [Medline: 29954725]
38. Framework for innovation: Design Council's evolved double diamond. Design Council. 2019. URL: https://tinyurl.com/bdhmmyuw [accessed 2024-02-02]
39. Powell LH, Freedland KE, Kaufmann PG. Behavioral Clinical Trials for Chronic Diseases: Scientific Foundations. Cham, Switzerland: Springer International Publishing; 2021.
40. Nahum-Shani I, Dziak JJ, Wetter DW. MCMTC: a pragmatic framework for selecting an experimental design to inform the development of digital interventions. Front Digit Health 2022;4:798025 [FREE Full text] [doi: 10.3389/fdgth.2022.798025] [Medline: 35355685]
41. Collins LM. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST). Cham, Switzerland: Springer International Publishing; 2018.
42. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ 2015;350:h2147 [FREE Full text] [doi: 10.1136/bmj.h2147] [Medline: 25956159]
43. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
44. Holtrop JS, Estabrooks PA, Gaglio B, Harden SM, Kessler RS, King DK, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126 [FREE Full text] [doi: 10.1017/cts.2021.789] [Medline: 34367671]
45. Freedland KE, King AC, Ambrosius WT, Mayo-Wilson E, Mohr DC, Czajkowski SM, et al. The selection of comparators for randomized controlled trials of health-related behavioral interventions: recommendations of an NIH expert panel. J Clin Epidemiol 2019;110:74-81 [FREE Full text] [doi: 10.1016/j.jclinepi.2019.02.011] [Medline: 30826377]
46. Chambers DA, Glasgow RE, Stange KC. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. Implement Sci 2013;8:117 [FREE Full text] [doi: 10.1186/1748-5908-8-117] [Medline: 24088228]

XSL•FO
RenderX

47. Hekler EB, Klasnja P, Riley WT, Buman MP, Huberty J, Rivera DE, et al. Agile science: creating useful products for behavior change in the real world. Transl Behav Med 2016;6(2):317-328 [FREE Full text] [doi: 10.1007/s13142-016-0395-7] [Medline: 27357001]
48. Beaulieu-Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, et al. Examining the use of real-world evidence in the regulatory process. Clin Pharmacol Ther 2020;107(4):843-852 [FREE Full text] [doi: 10.1002/cpt.1658] [Medline: 31562770]
49. Garrison LP, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. Value Health 2007;10(5):326-335 [FREE Full text] [doi: 10.1111/j.1524-4733.2007.00186.x] [Medline: 17888097]
50. Hekler E, Tiro JA, Hunter CM, Nebeker C. Precision health: the role of the social and behavioral sciences in advancing the vision. Ann Behav Med 2020;54(11):805-826 [FREE Full text] [doi: 10.1093/abm/kaaa018] [Medline: 32338719]
51. Vaithinathan AG, Asokan V. Public health and precision medicine share a goal. J Evid Based Med 2017;10(2):76-80. [doi: 10.1111/jebm.12239] [Medline: 28276633]
52. Burns L, Roux NL, Kalesnik-Orszulak R, Christian J, Hukkelhoven M, Rockhold F, et al. Real-world evidence for regulatory decision-making: guidance from around the world. Clin Ther 2022;44(3):420-437 [FREE Full text] [doi: 10.1016/j.clinthera.2022.01.012] [Medline: 35181179]
53. Chevance G, Hekler EB, Efoui-Hess M, Godino J, Golaszewski N, Gualtieri L, et al. Digital health at the age of the anthropocene. Lancet Digit Health 2020;2(6):e290-e291 [FREE Full text] [doi: 10.1016/S2589-7500(20)30130-8] [Medline: 33328121]
54. Chevance G, Fresán U, Hekler E, Edmondson D, Lloyd SJ, Ballester J, et al. Thinking health-related behaviors in a climate change context: a narrative review. Ann Behav Med 2023;57(3):193-204 [FREE Full text] [doi: 10.1093/abm/kaac039] [Medline: 35861123]
55. Kwasnicka D, Keller J, Perski O, Potthoff S, Hoor GAT, Ainsworth B, et al. White paper: open digital health—accelerating transparent and scalable health promotion and treatment. Health Psychol Rev 2022;16(4):475-491 [FREE Full text] [doi: 10.1080/17437199.2022.2046482] [Medline: 35240931]
56. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
57. SMART goals: a how to guide. University of California. 2016. URL: https://tinyurl.com/mvy8cxzt [accessed 2024-02-02]
58. Rau T. The difference between whole-group consensus and dynamic governance/sociocracy. Sociocracy For All. 2023. URL: https://tinyurl.com/ysy47skf [accessed 2024-02-02]
59. System Usability Scale (SUS). usability.gov. URL: https://tinyurl.com/22pkunfy [accessed 2024-02-02]
60. Freedland KE. Pilot trials in health-related behavioral intervention research: problems, solutions, and recommendations. Health Psychol 2020;39(10):851-862. [doi: 10.1037/hea0000946] [Medline: 32614197]
61. Design thinking 101. Nielsen Norman Group. 2016. URL: https://www.nngroup.com/articles/design-thinking/ [accessed 2024-02-02]
62. Martín CA, Rivera DE, Hekler EB, Riley WT, Buman MP, Adams MA, et al. Development of a control-oriented model of social cognitive theory for optimized mHealth behavioral interventions. IEEE Trans Control Syst Technol 2020;28(2):331-346 [FREE Full text] [doi: 10.1109/tcst.2018.2873538] [Medline: 33746479]
63. Chevance G, Golaszewski NM, Baretta D, Hekler EB, Larsen BA, Patrick K, et al. Modelling multiple health behavior change with network analyses: results from a one-year study conducted among overweight and obese adults. J Behav Med 2020;43(2):254-261 [FREE Full text] [doi: 10.1007/s10865-020-00137-2] [Medline: 31997127]
64. Spring B, Pfammatter AF, Marchese SH, Stump T, Pellegrini C, McFadden HG, et al. A factorial experiment to optimize remotely delivered behavioral treatment for obesity: results of the Opt-IN study. Obesity (Silver Spring) 2020;28(9):1652-1662 [FREE Full text] [doi: 10.1002/oby.22915] [Medline: 32656994]
65. Thomas JG, Goldstein CM, Bond DS, Lillis J, Hekler EB, Emerson JA, et al. Evaluation of intervention components to maximize outcomes of behavioral obesity treatment delivered online: a factorial experiment following the multiphase optimization strategy framework. Contemp Clin Trials 2021;100:106217 [FREE Full text] [doi: 10.1016/j.cct.2020.106217] [Medline: 33197609]
66. Lee H, Choi EH, Shin JU, Kim TG, Oh J, Shin B, et al. The impact of intervention design on user engagement in digital therapeutics research: factorial experiment With a mixed methods study. JMIR Form Res 2024;8:e51225. [doi: 10.2196/51225] [Medline: 38335015]
67. Cipriani A, Barbui C. What is a factorial trial? Epidemiol Psychiatr Sci 2013;22(3):213-215 [FREE Full text] [doi: 10.1017/s2045796013000231]
68. Kip H, Da Silva MC, Bouman YHA, van Gemert-Pijnen LJEWC, Kelders SM. A self-control training app to increase self-control and reduce aggression—a full factorial design. Internet Interv 2021;25:100392 [FREE Full text] [doi: 10.1016/j.invent.2021.100392] [Medline: 33996509]
69. Seewald NJ, Smith SN, Lee AJ, Klasnja P, Murphy SA. Practical considerations for data collection and management in mobile health micro-randomized trials. Stat Biosci 2019;11(2):355-370 [FREE Full text] [doi: 10.1007/s12561-018-09228-w] [Medline: 31462937]

XSL•FO
RenderX

70. Bell L, Garnett C, Qian T, Perski O, Potts HWW, Williamson E. Notifications to improve engagement with an alcohol reduction app: protocol for a micro-randomized trial. JMIR Res Protoc 2020;9(8):e18690 [FREE Full text] [doi: 10.2196/18690] [Medline: 32763878]
71. Battalio SL, Conroy DE, Dempsey W, Liao P, Menictas M, Murphy S, et al. Sense2Stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. Contemp Clin Trials 2021;109:106534 [FREE Full text] [doi: 10.1016/j.cct.2021.106534] [Medline: 34375749]
72. Czyz EK, King CA, Prouty D, Micol VJ, Walton M, Nahum-Shani I. Adaptive intervention for prevention of adolescent suicidal behavior after hospitalization: a pilot sequential multiple assignment randomized trial. J Child Psychol Psychiatry 2021;62(8):1019-1031 [FREE Full text] [doi: 10.1111/jcpp.13383] [Medline: 33590475]
73. De Barros Gonze B, Da Costa Padovani R, Do Socorro Simoes M, Lauria V, Proença NL, Sperandio EF, et al. Use of a smartphone app to increase physical activity levels in insufficiently active adults: feasibility Sequential Multiple Assignment Randomized Trial (SMART). JMIR Res Protoc 2020;9(10):e14322 [FREE Full text] [doi: 10.2196/14322] [Medline: 33094733]
74. Tamura RN, Krischer JP, Pagnoux C, Micheletti R, Grayson PC, Chen YF, et al. A small n sequential multiple assignment randomized trial design for use in rare disease research. Contemp Clin Trials 2016;46:48-51 [FREE Full text] [doi: 10.1016/j.cct.2015.11.010] [Medline: 26586608]
75. Lu X, Nahum-Shani I, Kasari C, Lynch KG, Oslin DW, Pelham WE, et al. Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in SMART studies. Stat Med 2016;35(10):1595-1615 [FREE Full text] [doi: 10.1002/sim.6819] [Medline: 26638988]
76. Freigoun MT, Martín CA, Magann AB, Rivera DE, Phatak SS, Korinek EV, et al. System identification of just walk: a behavioral mHealth intervention for promoting physical activity. 2017 Presented at: 2017 American Control Conference (ACC); May 24-26, 2017; Seattle, WA, USA p. 116-121. [doi: 10.23919/acc.2017.7962940]
77. dos Santos PL, Freigoun MT, Martin CA, Rivera DE, Hekler EB, Romano RA, et al. System identification of just walk: using matchable-observable linear parametrizations. IEEE Trans Control Syst Technol 2020;28(1):264-275 [FREE Full text] [doi: 10.1109/tcst.2018.2884833]
78. Hojjatinia S, Daly ER, Hnat T, Hossain SM, Kumar S, Lagoa CM, et al. Dynamic models of stress-smoking responses based on high-frequency sensor data. NPJ Digit Med 2021;4(1):162 [FREE Full text] [doi: 10.1038/s41746-021-00532-2] [Medline: 34815538]
79. Chevance G, Baretta D, Heino M, Perski O, Olthof M, Klasnja P, et al. Characterizing and predicting person-specific, day-to-day, fluctuations in walking behavior. PLoS One 2021;16(5):e0251659 [FREE Full text] [doi: 10.1371/journal.pone.0251659] [Medline: 33989338]
80. Goldstein SP, Thomas JG, Foster GD, Turner-McGrievy G, Butryn ML, Herbert JD, et al. Refining an algorithm-powered just-in-time adaptive weight control intervention: a randomized controlled trial evaluating model performance and behavioral outcomes. Health Informatics J 2020;26(4):2315-2331 [FREE Full text] [doi: 10.1177/1460458220902330] [Medline: 32026745]
81. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2(8):e124 [FREE Full text] [doi: 10.1371/journal.pmed.0020124] [Medline: 16060722]
82. Freedland KE. Progress in health-related behavioral intervention research: making it, measuring it, and meaning it. Health Psychol 2022;41(1):1-12. [doi: 10.1037/hea0001160] [Medline: 35113583]
83. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 1998;2(2):175-220. [doi: 10.1037/1089-2680.2.2.175]
84. Collins LM, Kugler KC. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics. Cham, Switzerland: Springer International Publishing; 2018.
85. Austrian J, Mendoza F, Szerencsy A, Fenelon L, Horwitz LI, Jones S, et al. Applying A/B testing to clinical decision support: rapid randomized controlled trials. J Med Internet Res 2021;23(4):e16651 [FREE Full text] [doi: 10.2196/16651] [Medline: 33835035]
86. Miller HN, Plante TB, Gleason KT, Charleston J, Mitchell CM, Miller ER, et al. A/B design testing of a clinical trial recruitment website: a pilot study to enhance the enrollment of older adults. Contemp Clin Trials 2021;111:106598 [FREE Full text] [doi: 10.1016/j.cct.2021.106598] [Medline: 34653651]
87. Senderey AB, Kornitzer T, Lawrence G, Zysman H, Hallak Y, Ariely D, et al. It's how you say it: systematic A/B testing of digital messaging cut hospital no-show rates. PLoS One 2020;15(6):e0234817 [FREE Full text] [doi: 10.1371/journal.pone.0234817] [Medline: 32574181]
88. Rivera DE, Pew MD, Collins LM. Using engineering control principles to inform the design of adaptive interventions: a conceptual introduction. Drug Alcohol Depend 2007;88(Suppl 2):S31-S40 [FREE Full text] [doi: 10.1016/j.drugalcdep.2006.10.020] [Medline: 17169503]
89. Hekler E, Anderson CAM, Cooper LA. Is it time to restructure the National Institutes of Health? Am J Public Health 2022;112(7):965-968 [FREE Full text] [doi: 10.2105/AJPH.2022.306830] [Medline: 35728037]
90. Kwasnicka D, Hoor GAT, Hekler E, Hagger MS, Kok G. Proposing a new approach to funding behavioural interventions using iterative methods. Psychol Health 2021;36(7):787-791. [doi: 10.1080/08870446.2021.1945061] [Medline: 34180340]

XSL•FO
RenderX

91. Glasgow RE, Huebschmann AG, Brownson RC. Expanding the CONSORT figure: increasing transparency in reporting on external validity. Am J Prev Med 2018;55(3):422-430. [doi: 10.1016/j.amepre.2018.04.044] [Medline: 30033029]
92. Wolfenden L, Williams CM, Wiggers J, Nathan N, Yoong SL. Improving the translation of health promotion interventions using effectiveness-implementation hybrid designs in program evaluations. Health Promot J Austr 2016;27(3):204-207. [doi: 10.1071/HE16056] [Medline: 29241482]
93. Ullman AJ, Beidas RS, Bonafide CP. Methodological progress note: hybrid effectiveness-implementation clinical trials. J Hosp Med 2022;17(11):912-916 [FREE Full text] [doi: 10.1002/jhm.12936] [Medline: 35934981]
94. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
95. Montgomery P, Grant S, Mayo-Wilson E, Macdonald G, Michie S, Hopewell S, et al. Reporting randomised trials of social and psychological interventions: the CONSORT-SPI 2018 extension. Trials 2018;19(1):407 [FREE Full text] [doi: 10.1186/s13063-018-2733-1] [Medline: 30060754]
96. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022;60(1):95-103 [FREE Full text] [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
97. Pearson N, Naylor PJ, Ashe MC, Fernandez M, Yoong SL, Wolfenden L. Guidance for conducting feasibility and pilot studies for implementation trials. Pilot Feasibility Stud 2020;6(1):167 [FREE Full text] [doi: 10.1186/s40814-020-00634-w] [Medline: 33292770]
98. Heino MTJ, Knittle K, Noone C, Hasselman F, Hankonen N. Studying behaviour change mechanisms under complexity. Behav Sci (Basel) 2021;11(5):77 [FREE Full text] [doi: 10.3390/bs11050077] [Medline: 34068961]
99. Hekler EB, Klasnja P, Chevance G, Golaszewski NM, Lewis D, Sim I. Why we need a small data paradigm. BMC Med 2019;17(1):133 [FREE Full text] [doi: 10.1186/s12916-019-1366-x] [Medline: 31311528]
100. Regal RE, Billi JE, Glazer HM. Phenothiazine-induced cholestatic jaundice. Clin Pharm 1987;6(10):787-794. [Medline: 2905941]
101. Komatsu T. Analysis of a characteristic alteration of glycogen in the liver of alloxan diabetic rat by fasting. Nagoya J Med Sci 1977;39(3-4):59-67. [Medline: 887142]
102. Korinek EV, Phatak SS, Martin CA, Freigoun MT, Rivera DE, Adams MA, et al. Adaptive step goals and rewards: a longitudinal growth model of daily steps for a smartphone-based walking intervention. J Behav Med 2018;41(1):74-86. [doi: 10.1007/s10865-017-9878-3] [Medline: 28918547]
103. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015;34S:1209-1219 [FREE Full text] [doi: 10.1037/hea0000306] [Medline: 26651462]
104. Palermo TM, de la Vega R, Murray C, Law E, Zhou C. A digital health psychological intervention (WebMAP mobile) for children and adolescents with chronic pain: results of a hybrid effectiveness-implementation stepped-wedge cluster randomized trial. Pain 2020;161(12):2763-2774 [FREE Full text] [doi: 10.1097/j.pain.0000000000001994] [Medline: 32658147]
105. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials 2015;16:352 [FREE Full text] [doi: 10.1186/s13063-015-0842-7] [Medline: 26279154]
106. Kim M, Kim Y, Go Y, Lee S, Na M, Lee Y, et al. Multidimensional cognitive behavioral therapy for obesity applied by psychologists using a digital platform: open-label randomized controlled trial. JMIR Mhealth Uhealth 2020;8(4):e14817 [FREE Full text] [doi: 10.2196/14817] [Medline: 32352391]
107. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. Lancet Digit Health 2020;2(4):e168-e178 [FREE Full text] [doi: 10.1016/S2589-7500(20)30017-0] [Medline: 33334505]
108. Luderer H, Chiodo L, Wilson A, Brezing C, Martinez S, Xiong X, et al. Patient engagement with a game-based digital therapeutic for the treatment of opioid use disorder: protocol for a randomized controlled open-label, decentralized trial. JMIR Res Protoc 2022;11(1):e32759 [FREE Full text] [doi: 10.2196/32759] [Medline: 35080499]
109. Kaizer AM, Wild J, Lindsell CJ, Rice TW, Self WH, Brown S, et al. Trial of Early Antiviral Therapies during Non-hospitalized Outpatient Window (TREAT NOW) for COVID-19: a summary of the protocol and analysis plan for a decentralized randomized controlled trial. Trials 2022;23(1):273 [FREE Full text] [doi: 10.1186/s13063-022-06213-z] [Medline: 35395957]
110. Fitzsimmons-Craft EE, Taylor CB, Graham AK, Sadeh-Sharvit S, Balantekin KN, Eichen DM, et al. Effectiveness of a digital cognitive behavior therapy-guided self-help intervention for eating disorders in college women: a cluster randomized clinical trial. JAMA Netw Open 2020;3(8):e2015633 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15633] [Medline: 32865576]
111. Pallejà-Millán M, Rey-Reñones C, Uriarte MLB, Granado-Font E, Basora J, Flores-Mateo G, et al. Evaluation of the Tobbstop mobile app for smoking cessation: cluster randomized controlled clinical trial. JMIR Mhealth Uhealth 2020;8(6):e15951 [FREE Full text] [doi: 10.2196/15951] [Medline: 32589153]

XSL•FO
RenderX

112. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry 2006;59(11):990-996. [doi: 10.1016/j.biopsych.2005.09.014] [Medline: 16368078]
113. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry 2006;63(5):484-489. [doi: 10.1001/archpsyc.63.5.484] [Medline: 16651505]
114. Anvari F, Lakens D. Using anchor-based methods to determine the smallest effect size of interest. J Exp Soc Psychol 2021;96:104159 [FREE Full text] [doi: 10.1016/j.jesp.2021.104159]
115. Wang XQ, Mao SP. Comparison of morphology, pathogenicity and drug response among three isolates of schistosoma japonicum in the mainland of China. Ann Parasitol Hum Comp 1989;64(2):110-119 [FREE Full text] [doi: 10.1051/parasite/1989642110] [Medline: 2735700]
116. 14 iconic websites that show off classic 90s web design. Webflow. 2022. URL: https://webflow.com/blog/90s-website-design [accessed 2024-02-02]
117. Beckman BL. What Apple, Google, and Amazon's websites looked like in 1999. Mashable. 2020. URL: https://mashable.com/article/90s-web-design [accessed 2024-02-02]
118. Perski O. Scientific and ethical challenges to defining what constitutes 'proportionate evidence' for the regulation and accreditation of applications to treat addiction. Addiction 2021;116(12):3285-3287 [FREE Full text] [doi: 10.1111/add.15619] [Medline: 34231264]
119. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Helath Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-02-02]
120. Klasnja P, Hekler EB, Korinek EV, Harlow J, Mishra SR. Toward usable evidence: optimizing knowledge accumulation in HCI research on health behavior change. Proc SIGCHI Conf Hum Factor Comput Syst 2017;2017:3071-3082 [FREE Full text] [doi: 10.1145/3025453.3026013] [Medline: 30272059]
121. Larsen KR, Michie S, Hekler EB, Gibson B, Spruijt-Metz D, Ahern D, et al. Behavior change interventions: the potential of ontologies for advancing science and practice. J Behav Med 2017;40(1):6-22 [FREE Full text] [doi: 10.1007/s10865-016-9768-0] [Medline: 27481101]
122. Larsen KI, Hekler EB, Paul MJ, Gibson BS. Improving usability of social and behavioral sciences‘ evidence: a call to action for a national infrastructure project for mining our knowledge. Commun Assoc Inf Syst 2020;46:1-17 [FREE Full text] [doi: 10.17705/1CAIS.04601]
123. Reardon J, Lee SSJ, Goering S, Fullerton SM, Cho MK, Panofsky A, et al. Trustworthiness matters: building equitable and ethical science. Cell 2023;186(5):894-898 [FREE Full text] [doi: 10.1016/j.cell.2023.01.008] [Medline: 36724788]
124. Espie CA, Torous J, Brennan TA. Digital therapeutics should be regulated with gold-standard evidence. Health Affairs Forefront. 2022. URL: https://www.healthaffairs.org/do/10.1377/forefront.20220223.739329/ [accessed 2024-02-02]
125. Mullins CD, Abdulhalim AM, Lavallee DC. Continuous patient engagement in comparative effectiveness research. JAMA 2012;307(15):1587-1588. [doi: 10.1001/jama.2012.442] [Medline: 22511684]
126. Minkler M, Wallerstein N. Community-Based Participatory Research for Health: From Process to Outcomes, 2nd Edition. San Francisco, CA: Wiley; 2011.
127. Birnbaum F, Lewis D, Rosen RK, Ranney ML. Patient engagement and the design of digital health. Acad Emerg Med 2015;22(6):754-756 [FREE Full text] [doi: 10.1111/acem.12692] [Medline: 25997375]
128. Petersen C, Austin RR, Backonja U, Campos H, Chung AE, Hekler EB, et al. Citizen science to further precision medicine: from vision to implementation. JAMIA Open 2020;3(1):2-8 [FREE Full text] [doi: 10.1093/jamiaopen/ooz060] [Medline: 32607481]
129. Wilson BB. Resilience for All: Striving for Equity Through Community-Driven Design. Washington, DC: Island Press; 2018.
130. Jason L, Glantsman O, O'Brien JF, Ramian KN. Introduction to Community Psychology: Becoming an Agent of Change. Montreal, Quebec, Canada: Rebus Community; 2019.
131. Whittaker A. Research Skills for Social Work, 2nd Edition. London: Sage Publications; 2012.
132. Ellis RB, Wright J, Miller LS, Jake-Schoffman D, Hekler EB, Goldstein CM, et al. Lessons learned: beta-testing the digital health checklist for researchers prompts a call to action by behavioral scientists. J Med Internet Res 2021;23(12):e25414 [FREE Full text] [doi: 10.2196/25414] [Medline: 34941548]
133. Nebeker C, Gholami M, Kareem D, Kim E. Applying a digital health checklist and readability tools to improve informed consent for digital health research. Front Digit Health 2021;3:690901 [FREE Full text] [doi: 10.3389/fdgth.2021.690901] [Medline: 34713167]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DTx: digital therapeutics
FDA: Food and Drug Administration
MOST: Multiphase Optimization Strategy

XSL•FO
RenderX

MVP: minimal viable product
NIH: National Institutes of Health
ORBIT: Obesity-Related Behavioral Intervention Trials
Pre-Cert: Precertification
PRECIS-2: Pragmatic Explanatory Continuum Indicator Summary-2
RE-AIM: reach effectiveness, adoption, implementation, and maintenance
RWD: real-world data
RWE: real-world evidence
SMART: specific, measurable, actionable, realistic, and timely
STEEEP: safe, timely, effective, efficient, equitable, and patient-centered
WHO: World Health Organization

Edited by T Leung; submitted 21.05.23; peer-reviewed by R Barak Ventura, D Boeldt; comments to author 06.12.23; revised version received 13.01.24; accepted 29.01.24; published 05.03.24.

Please cite as:
Kim M, Patrick K, Nebeker C, Godino J, Stein S, Klasnja P, Perski O, Viglione C, Coleman A, Hekler E
The Digital Therapeutics Real-World Evidence Framework: An Approach for Guiding Evidence-Based Digital Therapeutics Design, Development, Testing, and Monitoring
J Med Internet Res 2024;26:e49208
URL: https://www.jmir.org/2024/1/e49208
doi:10.2196/49208
PMID:38441954

©Meelim Kim, Kevin Patrick, Camille Nebeker, Job Godino, Spencer Stein, Predrag Klasnja, Olga Perski, Clare Viglione, Aaron Coleman, Eric Hekler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges

Yan Chen[1*], PhD; Pouyan Esmaeilzadeh[1*], PhD

Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, FL, United States
[*]all authors contributed equally

Corresponding Author:
Pouyan Esmaeilzadeh, PhD
Department of Information Systems and Business Analytics
College of Business
Florida International University
Modesto A Maidique Campus
11200 SW 8th St, RB 261 B
Miami, FL, 33199
United States
Phone: 1 3053483302
Email: pesmaeil@fiu.edu

Abstract

As advances in artificial intelligence (AI) continue to transform and revolutionize the field of medicine, understanding the potential uses of generative AI in health care becomes increasingly important. Generative AI, including models such as generative adversarial networks and large language models, shows promise in transforming medical diagnostics, research, treatment planning, and patient care. However, these data-intensive systems pose new threats to protected health information. This Viewpoint paper aims to explore various categories of generative AI in health care, including medical diagnostics, drug discovery, virtual health assistants, medical research, and clinical decision support, while identifying security and privacy threats within each phase of the life cycle of such systems (ie, data collection, model development, and implementation phases). The objectives of this study were to analyze the current state of generative AI in health care, identify opportunities and privacy and security challenges posed by integrating these technologies into existing health care infrastructure, and propose strategies for mitigating security and privacy risks. This study highlights the importance of addressing the security and privacy threats associated with generative AI in health care to ensure the safe and effective use of these systems. The findings of this study can inform the development of future generative AI systems in health care and help health care organizations better understand the potential benefits and risks associated with these systems. By examining the use cases and benefits of generative AI across diverse domains within health care, this paper contributes to theoretical discussions surrounding AI ethics, security vulnerabilities, and data privacy regulations. In addition, this study provides practical insights for stakeholders looking to adopt generative AI solutions within their organizations.

(J Med Internet Res 2024;26:e53008) doi:10.2196/53008

KEYWORDS

artificial intelligence; AI; generative artificial intelligence; generative AI; medical practices; potential benefits; security and privacy threats

Introduction

Artificial intelligence (AI) is transforming many industries, including health care. AI has the potential to revolutionize health care by enabling the detection of signs, patterns, diseases, anomalies, and risks. From administrative automation to clinical decision support, AI holds immense potential to improve patient outcomes, lower costs, and accelerate medical discoveries [1]. An especially promising subset of AI is generative models, which are algorithms that can synthesize new data, imagery, text, and other content with humanlike creativity and nuance based on patterns learned from existing data [2]. Generative AI could power clinical practices in health care, from generating synthetic patient data to augmenting rare disease research to creating AI-assisted drug discovery systems [3]. Generative AI has the potential to detect signs, patterns, diseases, anomalies, and risks and assist in screening patients for various chronic diseases, making more accurate and data-driven diagnoses and improving clinical decision-making [4]. Generative AI also has

XSL•FO
RenderX

the potential to transform patient care with generative AI virtual health assistants [5].

However, generative AI systems pose acute privacy and security risks along with their transformative potential because of their vast data requirements and opacity [6]. Generative AI models can be trained on sensitive, multimodal patient data, which could be exploited by malicious actors. Therefore, the collection and processing of sensitive patient data, along with tasks such as model training, model building, and implementing generative AI systems, present potential security and privacy risks. Given the sensitive nature of medical data, any compromise can have dire consequences, not just in data breaches but also in patients' trust and the perceived reliability of medical institutions. As these AI systems move from laboratory to clinical deployment, a measured approach is required to map and mitigate their vulnerabilities. Another challenge of using generative AI models is that they can be biased, which could lead to inaccurate diagnoses and treatments [7].

Despite the growing interest in generative AI in health care, there is a gap in the literature regarding a comprehensive examination of the unique security and privacy threats associated with generative AI systems. Our study attempts to provide insights into the different categories of generative AI in health care, including medical diagnostics, drug discovery, virtual health assistants, medical research, and clinical decision support. This study also aims to address the gap by identifying security and privacy threats and mapping them to the life cycle of various generative AI systems in health care, from data collection through model building to clinical implementation. By identifying and analyzing these threats, we can gain insights into the vulnerabilities and risks associated with the use of generative AI in health care. We also seek to contribute to theory and practice by highlighting the importance of addressing these threats and proposing mitigation strategies.

The findings of this study can inform the development of future generative AI systems in health care and help health care organizations better understand the potential benefits and risks of using these systems. The significance of this study lies in its potential to inform policy makers, health care organizations, and AI developers about the security and privacy challenges associated with generative AI in health care. The findings of this study can guide the development of robust data governance frameworks, secure infrastructure, and ethical guidelines to ensure the safe and responsible use of generative AI in health care. With careful governance, the benefits of generative models can be realized while safeguarding patient data and public trust. Ultimately, this study contributes to the advancement of knowledge in the field of AI in health care and supports the development of secure and privacy-preserving generative AI systems for improved patient care and outcomes.

Generative AI Applications in Health Care

Overview

Generative AI models use neural networks to identify patterns and structures within existing data to generate new and original content. Generative AI refers to techniques such as generative adversarial networks (GANs) and large language models (LLMs) that synthesize novel outputs such as images, text, and molecular structures [8]. GANs use 2 neural networks, a generator and a discriminator, that compete against each other to become better at generating synthetic data [9]. LLMs such as GPT-4 (OpenAI) are trained on massive text data and can generate synthetic natural language text, code, and so on [10].

Generative AI has spurred a wide range of applications in health care. This subset of AI has the potential to make a breakthrough in medical diagnostic applications, given its capability to build models using multimodal medical data [5]. Generative AI also promises to accelerate drug discovery by inventing optimized molecular candidates [11]. In research settings, these generative AI techniques can hypothesize promising new directions by creatively combining concepts [12]. Generative AI also has applications in engaging patients through natural conversation powered by LLMs [2]. When integrated into clinical workflows, it may also provide physicians with patient-specific treatment suggestions [13].

The classification of generative AI systems presented in Table 1 was developed based on a careful analysis of the various factors that differentiate these technologies.

XSL•FO
RenderX

Table 1. Categories of generative artificial intelligence (AI) applications in health care.

Category	Example	Setting	User	Input data	Output data	Personalization level	Workflow integration	Validation needed	Impact	Risks	Human involvement
Medical diagnostics	AI-Rad Companion	Radiology	Radiologists	Medical images	Text findings	Individual	Postimaging	High	Improved diagnosis	Reliability and bias	High
Drug discovery	Insilico Medicine	Biotechnology	Research scientists	Target proteins and disease data	Novel molecular structures	Semipersonalized	Early-stage research	High	Faster discoveries	Safety and testing requirements	Moderate
Virtual health assistants	Sensely	Web clinics	Patients	Conversation	Conversation	Semipersonalized	Patient engagement	Moderate	Increased access	Privacy and misinformation	Moderate
Medical research	Anthropic	Laboratories and academia	Researchers	Research concepts and data sets	Hypotheses and questions	Semipersonalized	Idea generation	Low	Research insights	Misdirection	Moderate
Clinical decision support	Glass AI	Point of care	Physicians	Patient data	Treatment suggestions	Individual	Diagnosis and treatment	High	Improved outcomes	Overreliance and bias	High

Differentiating Factors

The goal was to provide a framework for better understanding the diversity of generative AI across health care settings. We leverage several key factors to differentiate the applications and provide insights into this emerging field, described in the following sections.

Setting

The clinical setting categorizes where in the health care workflow the generative AI system is applied, such as diagnostics, treatment planning, drug discovery, clinical decision support, and patient education [14]. This provides insights into the breadth of health care contexts leveraging these technologies.

Users

Generative AI tools are tailored to different types of users in health care, from clinicians to researchers to patients [15]. Categorization by intended user groups reveals how generative AI penetrates various stakeholder groups and which user groups may adopt and interact with generative AI applications.

Input Data

The data sources powering generative AI systems vary significantly, from electronic health records (EHRs) and medical imaging to biomedical literature, laboratory tests, and patient-provided data [16]. Categorization by data inputs illustrates how different data fuel different categories of applications.

Output Data

The outputs produced by the system, such as images, care planning, prescription advice, treatment options, drug molecules, text, risk scores, and education materials [17], demonstrate the wide range of generative AI capabilities in health care.

Personalization Level

The level of personalization to individual patients reveals the precision of the outputs, from generalized to fully patient specific. This provides a perspective on the customizability of the generative AI system.

Workflow Integration

Some generative AI systems are designed as stand-alone applications, whereas others are integrated into clinical workflows via EHRs, order sets, and so on. Categorization by workflow integration sheds light on the level of adoption, implementation practices, and integration of these tools.

Validation Needs

The extent of validation required, from noncritical outputs to those needing rigorous US Food and Drug Administration approval [18], highlights differences in oversight and impact levels.

Impact: profiling the benefits and use cases served by the generative AI technology, such as improving diagnostics, reducing medication errors, or accelerating drug discovery, provides insights into the varied impacts.

Risks

Discussing risks and limitations provides a balanced view of concerns such as algorithmic bias, privacy concerns, security issues, system vulnerability, and clinical integration challenges.

Human-AI Collaboration

Generative AI systems differ in the level of human involvement required, from fully automated to human-in-the-loop (human engagement in overseeing and interacting with the AI's operational process) [19]. Categorization by human-AI partnership provides insights into the changing dynamics between humans and AI across health care.

XSL•FO
RenderX

Aims

This study aims to reveal crucial differences, use cases, adoption levels, various risks, and implementation practices by developing categories based on these key attributes of generative AI systems. The proposed framework clarifies the heterogeneous landscape of generative AI in health care and enables a trend analysis across categories. These factors provide a perspective on how generative AI manifests distinctly for various users, data types, workflows, risk factors, and human-AI partnerships within health care. By systematically analyzing the diverse range of generative AI systems across health care settings using the key factors discussed previously, we can classify the heterogeneous landscape of generative AI in health care into 5 overarching categories: medical diagnostics, drug discovery, virtual health assistants, medical research, and clinical decision support.

Medical Diagnostics

Generative AI techniques can analyze data from wearables, EHRs, and medical images (eg, x-rays, magnetic resonance imaging, and computed tomography scans) to detect signs, patterns, diseases, anomalies, and risks and generate descriptive findings to improve diagnoses. Systems such as AI-Rad Companion leverage natural language generation models to compose radiology reports automatically, highlighting potential abnormalities and issues for clinician review [20]. This assists radiologists by providing initial draft findings more rapidly. However, clinicians must thoroughly validate any generative AI outputs before clinical use. Ongoing challenges include reducing false positives and negatives [21].

Drug Discovery

Generative AI shows promise for expediting and enhancing drug discovery through inventing optimized molecular structures de novo. Techniques such as GANs combined with reinforcement learning allow the intelligent generation of molecular graph representations [22]. Companies such as Insilico Medicine are using these generative chemistry techniques to propose novel target-specific drug candidates with desired properties. This accelerates preclinical pharmaceutical research. However, validating toxicity and efficacy remains critical before human trials.

Virtual Health Assistants

Generative models such as LLMs can power conversational agents that understand and respond to patient questions and concerns [23]. Companies such as Sensely and Woebot Health leverage these techniques to create virtual assistants that explain symptoms, provide health information, and offer screening triage advice through natural dialogue [24]. This increases access and engagement for patients. However, challenges remain around privacy, information accuracy, and integration into provider workflows [25].

Medical Research

In research settings, generative AI can formulate novel hypotheses by making unexpected combinations of concepts, mimicking human creativity and intuition. Claude from Anthropic can read research papers and propose unexplored directions worth investigating [26]. This unique generative capacity could accelerate scientific advancement. However, corroboration by human researchers is crucial to prevent the blind acceptance of AI-generated findings [27].

Clinical Decision Support

Integrating generative AI into clinical workflows could provide patient-specific suggestions to assist physicians in decision-making. Glass AI leverages LLMs such as GPT-3 to generate tailored treatment options based on patient data for physicians to review [15]. This could improve outcomes and reduce errors. However, bias mitigation and high validation thresholds are critical before real-world adoption [28].

By holistically examining all the key factors, we can see how each one contributes to delineating these 5 high-level categories that provide a comprehensive snapshot of the generative AI landscape in health care. Analyzing these 5 categories through the lens of the proposed factors enables our study to reveal crucial differences, use cases, benefits, limitations, and implementation practices of generative AI technologies across major health care domains.

Literature Review

The adoption of AI (powered by various models) is accelerating across health care for applications ranging from medical imaging to virtual assistants. However, the data-intensive nature and complexity of these systems introduce acute privacy and security vulnerabilities that must be addressed to ensure safe and ethical deployment in clinical settings. This literature review covers 2 topics. First, we highlight the dual nature of technological advancements in generative AI within health care, its benefits, and its risks, particularly in terms of privacy and security that it entails. Second, we explain AI regulation and compare the key aspects of the European Union (EU) AI Act and the US AI Bill of Rights.

Generative AI: Balancing Benefits and Risks

Overview

The use of generative AI systems in medicine holds promise for improvements in areas such as patient education and diagnosis support. However, recent studies highlight that privacy and security concerns may slow user adoption. A survey explores the application of GANs toward ensuring privacy and security [29]. It highlights how GANs can be used to address increasing privacy concerns and strengthen privacy regulations in various applications, including medical image analysis. The unique feature of GANs in this context is their adversarial training characteristic, which allows them to investigate privacy and security issues without predetermined assumptions about opponents' capabilities. This is crucial because these capabilities are often complex to determine with traditional attack and defense mechanisms. In the privacy and security models using GANs, the generator can be modeled in two ways: (1) as an attacker aiming to fool a defender (the discriminator) to simulate

XSL•FO
RenderX

an attack scenario and (2) as a defender resisting a powerful attacker (the discriminator) to simulate a defense scenario.

Examples of defense models include generative adversarial privacy [30], privacy-preserving adversarial networks [31], compressive adversarial privacy [32], and reconstructive adversarial network [33]. These GAN-based mechanisms offer innovative ways to enhance privacy and security in various machine learning and data processing scenarios. The examples are described in the subsequent sections.

Protection of Preimage Privacy

The compressive privacy GAN is designed to preprocess private data before the training stage in machine learning as a service scenarios [34]. It includes 3 modules: a generator module (G) as a privatization mechanism for generating privacy-preserving data, a service module (S) providing prediction services, and an attacker module (A) that mimics an attacker aiming to reconstruct the data. The objective is to ensure optimal performance of the prediction service, even in the face of strong attackers, by intentionally increasing the reconstruction error. This method defends against preimage privacy attacks in machine learning as a service by ensuring that the input data of a service module contains no sensitive information.

Privacy in Distributed Learning Systems

In decentralized learning systems, such as distributed selective stochastic gradient descent [35] and federated learning (FL) [36], data are trained locally by different participants without data sharing. This setup can protect data privacy to some extent, but it is not perfect. The GAN-based models in these systems can mimic data distribution and potentially threaten data privacy. The potential risks associated with the application of GAN-based models in decentralized learning systems are multifaceted, highlighting the need for robust privacy protection measures. These risks are explained as the following: an attacker might use GANs to recover sensitive information within the distributed training system, and a malicious server can reveal user-level privacy in distributed learning systems by training a multitask GAN with auxiliary identification.

Protection mechanisms include embedding a “buried point layer” in local models to detect abnormal changes and block attackers and integrating GAN with FL to produce realistic data without privacy leakage.

Differential Privacy in GANs

To address the problem of privacy leakage in the models, two solutions have been proposed: (1) adding a regularization term in a loss function to avoid overfitting and improve robustness; for example, this method can be applied to defend against membership inference attacks, [37] and (2) adding acceptable noise into the model parameters to hinder privacy inference attacks. Such methods have been used for privacy protection, particularly the combination of differential privacy and neural networks [38].

In medical research, the widespread use of medical data, particularly in image analysis, raises significant concerns about the potential exposure of individual identities. An innovative adversarial training method focused on identity-obfuscated segmentation has been proposed to address this challenge [39]. This method is underpinned by a deep convolutional GAN-based framework comprising three key components: (1) a deep encoder network, functioning as the generator, efficiently obscuring identity markers in medical images by incorporating additional noise; (2) a binary classifier serves as the discriminator, ensuring that the transformed images retain a resemblance to their original counterparts; and (3) a convolutional neural network–based network dedicated to medical image analysis, acting as an alternate discriminator responsible for analyzing the segmentation details of the images. This framework integrates an encoder, a binary classifier, and a segmentation analysis network to form a robust approach to safeguard medical data privacy while preserving the integrity and efficacy of medical image segmentation.

The use of EHR medical records has significantly advanced medical research while simultaneously amplifying concerns regarding the privacy of this sensitive information. In response, Choi et al [40] devised the medical GAN (medGAN), an innovative adaptation of the standard GAN framework, aimed at producing synthetic patient records that respect privacy. The medGAN excels at generating high-dimensional discrete variables. Its architecture uses an autoencoder as the generator, which creates synthetic medical data augmented with noise. A binary classifier functions as the discriminator, ensuring the resemblance of these data to real records. The outcome is synthetic medical data suitable for various uses, such as distribution analysis, predictive modeling, and medical expert evaluations, minimizing the privacy risks associated with both identity and attributes. Furthering these advancements, Yale et al [41] conducted an in-depth evaluation of medGAN’s ability to protect privacy in medical records. In a parallel development, Torfi and Fox [42] introduced Correlation-Capturing Convolutional Generative Adversarial Networks (CorGAN), which focuses on the correlations within medical records. Unlike medGAN, CorGAN uses a dual autoencoder in its generator, enabling the creation of sequential EHRs rather than discrete entries. This approach enhances predictive accuracy, providing more effective assistance to medical professionals [43].

Similarly, Nova [14] discusses the transformative impact of generative AI on EHRs and medical language processing, underlining the accompanying privacy concerns. It examines the balance between the utility of GANs in generating health care data and the preservation of privacy. Rane [44] explores the wider privacy and security implications of using generative AI models, such as ChatGPT, in health care within the context of Industry 4.0 and Industry 5.0 transformation. The impact of generative content on individual privacy is further explored by Bale et al [45], emphasizing the ethical considerations in health care.

Ghosheh et al [46] suggest that the use of GANs to create synthetic EHRs creates many privacy challenges (eg, reidentification and membership attacks). Hernandez et al [47] discuss privacy concerns related to synthetic tabular data generation in health care. Various methods and evaluation metrics are used to assess the privacy dimension of the synthetic tabular data generation approaches. These methods include identity disclosure, attribute disclosure, distance to the closest

XSL•FO
RenderX

record, membership attack, maximum real-to-synthetic similarity, differential privacy cost, and GANs. For instance, differential privacy is an approach that adds noise to the data to prevent the identification of individuals. GANs can create new and nonreal data points. Other advanced statistical and machine learning techniques attempt to balance data utility and privacy. Each method has its strengths and limitations, and the choice depends on the specific requirements of the health care application and the sensitivity of the data involved.

The applications and challenges of generative AI in health care, including privacy issues and AI-human collaboration, are explored by Fui-Hoon et al [48]. They discuss several privacy issues related to generative AI, such as the potential disclosure of sensitive or private information by generative AI systems, the widening of the digital divide, and the collection of personal and organizational data by these systems, which raises concerns about security and confidentiality. In addition, they highlight regulatory and policy challenges, such as issues with copyright for AI-generated content, the lack of human control over AI behavior, data fragmentation, and information asymmetries between technology giants and regulatory authorities.

A study discusses the potential of FL as a privacy-preserving approach in health care AI applications [49]. FL is a distributed AI paradigm that offers privacy preservation in smart health care systems by allowing models to be trained without accessing the local data of participants. It provides privacy to end users by only sharing gradients during training. The target of FL in health care AI applications is to preserve the privacy of sensitive patient information communicated between hospitals and end users, particularly through Internet of Medical Things (IoMT) devices. The approach incorporates advanced techniques such as reinforcement learning, digital twin, and GANs to detect and prevent privacy threats in IoMT networks. The potential beneficiaries of FL in health care include patients, health care providers, and organizations involved in collaborative health care research and analysis. However, implementing FL in IoMT networks presents challenges, such as the need for robust FL for diffused health data sets, the integration of FL with next-generation IoMT networks, and the use of blockchain for decentralized and secure data storage. Furthermore, incentive mechanisms are being explored to encourage the participation of IoMT devices in FL, and digital twin technology is being leveraged to create secure web-based environments for remote patient monitoring and health care research. Overall, FL in health care AI applications aims to address privacy and security concerns while enabling collaborative and efficient health care systems.

Another study emphasizes the need for secure and robust machine learning techniques in health care, particularly focusing on privacy and security [50]. Finally, a study addresses the vulnerabilities of generative models to adversarial attacks (eg, evasion attacks and membership inference attacks), highlighting a significant area of concern in health care data security [51]. These studies collectively underscore the need for a balanced approach to leveraging the benefits of AI-driven health care innovations while ensuring robust privacy and security measures.

AI, Legal Challenges, and Regulation

AI, especially generative AI, has presented many legal challenges, raising many profound questions on how AI can be legally, securely, and safely used by businesses and individuals [52]. The EU AI Act, passed in 2023, is the first comprehensive legal framework to specifically regulate AI systems [53]. It categorizes systems by risk level and introduces mandatory requirements for high-risk AI related to data and documentation, transparency, human oversight, accuracy, cybersecurity, and so on. As stated in the act, national authorities will oversee compliance.

The US AI Bill of Rights, unveiled in 2023, takes a different approach as a nonbinding set of principles to guide AI development and use focused on concepts such as algorithmic discrimination awareness, data privacy, notice and explanation of AI, and human alternatives and oversight [54]. Rather than authoritative regulation, it promotes voluntary adoption by organizations.

Although the EU law institutes enforceable accountability around risky AI, the US bill espouses aspirational AI ethics principles. Both identify important issues such as potential bias, privacy risks, and the need for human control but tackle them differently—the EU through compliance requirements and the United States through voluntary principles. Each seeks more responsible AI but via divergent methods that fit their governance models. Despite differences in methods, there is a consensus on fundamental issues such as ensuring transparency, maintaining accuracy, minimizing adverse effects, and providing mechanisms for redressal.

Specifically, for generative AI such as ChatGPT, the EU AI Act mandates transparency requirements, such as disclosing AI-generated content, designing models to prevent illegal content generation, and publishing training data summaries. Although the principles mentioned in the US AI Bill of Rights do not specifically address generative AI, they provide a framework for the ethical and responsible use of all AI technologies, including generative AI. The principles emphasize safety, nondiscrimination, privacy, transparency, and human oversight, all of which are relevant to developing and deploying generative AI systems.

Ultimately, the EU legislates binding rules that companies must follow, whereas the United States issues guidance that organizations may freely adopt. Despite this schism, both highlight growing policy makers' concern over AI's societal impacts and the emergence of either compulsory or optional frameworks aimed at accountability. As leading AI powers craft different but related policy solutions, ongoing collaboration around shared values while allowing varied implementations will be important for setting global AI standards.

XSL•FO
RenderX

Security and Privacy Threats in the Life Cycle of a Generative AI in Health Care System

Overview

Although generative AI in health care holds great promise, substantial validation is required before real-world deployment. Ethical risks around reliability, accountability, algorithmic bias, and data privacy as well as security risks related to confidentiality, integrity, and availability must be addressed through a human-centric approach [55]. Liu et al [56] surveyed the security and privacy attacks related to machine learning and developed a taxonomy. The taxonomy classifies those attacks into three categories: (1) attacks targeting classifiers; (2) attacks violating integrity, availability, and privacy (ie, part of confidentiality); and (3) attacks with or without specificity. They also summarize the defense techniques in the training phase and the testing and inferring phase of the life cycle of machine learning, for example, data sanitization techniques against data poisoning attacks in the training phase and privacy-preserving techniques against privacy attacks in the testing or inferring phase. Similarly, Hu et al [57] present an overall framework of attacks and defense strategies based on the following five phases of the AI life cycle: (1) data collection phase—main security threats include databases, fake data, data breaches, and sensor attacks; defense strategies include data sanitization and data government; (2) data processing phase—image scaling is the main threat; recommended defense strategies include image reconstruction and data randomization; (3) training phase—data poisoning is the main threat; defense strategies focus on techniques that can identify and remove poisoned data (eg, the certified defense technique proposed by Tang et al [58]) and provide robust and reliable AI models; (4) inference phase—this phase mainly faces adversarial example attacks such as white-box, gray-box, and black-box attacks depending on how much the attacker knows about the target model; a variety of defense strategies can be implemented to tackle such attacks, such as adopting strategies in phases 1 to 3 to modify data (eg, data reconstruction and randomization) or modify or enhance models with newer model construction methods resistant to adversarial example attacks (eg, using deep neural networks and GAN-based networks [58,59]); (5) integration phase—AI models face AI biases, confidentiality attacks (eg, model inversion, model extraction, and various privacy attacks), and code vulnerability exploitation; defense strategies in this phase should be comprehensive via integrating various solutions such as fuzz testing and blockchain-based privacy protection.

Generative AI is built upon machine learning and AI techniques and hence faces similar security and privacy threats, as summarized in the studies by Liu et al [56] and Hu et al [57]. Nevertheless, because generative AI, such as LLMs, often requires large volumes of data (eg, large volumes of patient data) to train, it faces many existing and new security and privacy threats. If deployed carelessly, generative models increase the avenues for protected health information (PHI) to be leaked, stolen, or exposed in a breach. For example, deidentifying data for LLMs is challenging [60]. Even anonymized patterns in data could potentially reidentify individuals if models are improperly handled after training. One example is medical image analysis, as deidentified medical images could be reidentified in medical image analysis because of the massive amount of image data used in training [39]. LLMs in health care also face data quality and bias issues, similar to any machine learning model, leading to erroneous medical conclusions or recommendations [61].

Furthermore, hackers could also exploit vulnerabilities in systems hosting generative models to access the sensitive health data used for training. Skilled hackers may be able to feed prompts to models to obtain outputs of specific patient details that allow reidentification even from anonymized data. For example, improperly secured LLMs could enable bad actors to generate fake patient data or insurance claims [62]. In general, generative AI in health care encounters many of the same security and privacy threats as general AI and machine learning systems, along with new threats stemming from its unique context. On the basis of the life cycle in the studies by Liu et al [56] and Hu et al [57], our study presents a 3-phase life cycle for generative AI. It also identifies security and privacy threats and maps them to the life cycle of various generative AI systems in health care (Figure 1). It should be noted that although this study primarily discusses various security and privacy threats associated with generative AI in health care (such as AI hallucination in health care), many of these threats are not unique to generative AI systems and are also prevalent in broader AI systems and machine learning models in health care and other fields.

Figure 1. Artificial intelligence (AI) security and privacy threats in 3 phases of the AI life cycle.

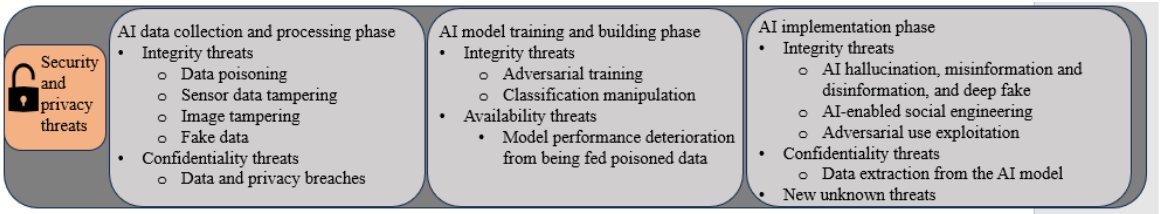

Data Collection and Processing Phase

Similar to AI systems in other fields, almost all types of generative AI in health care face integrity threats. The main integrity threats in this phase are traditionally owing to errors and biases. Unintentionally, the increased data volume and complexity of generative AI threatens data integrity because errors and biases are prone to occur [63]. Errors and biases also depend on the data sources for different types of generative AI in health care. For example, assembling genomic databases and

XSL•FO
RenderX

chemical compound or protein structure databases for drug discovery is extremely challenging and could be error ridden because many genomic and protein databases lack necessary annotations, are inconsistent in formats, and may be poor in data quality [64].

Intentionally, data poisoning can occur when data are collected from various software packages by tampering with data. For example, malicious insiders can tamper with data intentionally when gathering data from various software sources. For example, malicious actors can internationally submit mislabeled genomic sequences and chemical compound protein structures to tamper genomic databases and chemical compound or protein structure databases, leading to fault training models and AI hallucination.

In addition to data poisoning from software, in health care, data may be gathered from sensors embedded in medical devices and equipment. Sensor data can be spoofed [65,66], tampered with, and thus poisoned. Furthermore, medical data contains a large number of images. Adversaries can exploit the difference in cognitive processes between AI and humans and tamper with images during the data collection and processing phase. Image-scaling attacks, in which an adversary manipulates images so that changes are imperceptible to the human eye but recognizable by AI after downscaling, represent one such form of attack [67,68]. Other attacks on data sources of medical images include, but are not limited to, copy-move tampering (ie, copying an area and moving it to another area), classical inpainting tampering (ie, patching a missing area with tampered image slices), deep inpainting tampering (ie, similar to classical inpainting tampering but using highly realistic image slides generated by GANs), sharpening, blurring, and resampling [69]. In scenarios where AI in imaging diagnostics is targeted by such attacks, the image data can be poisoned with malicious information. Furthermore, generative AI, such as GANs, has empowered hackers to generate or change the attributes or content of medical images with high visual realism, making the detection of tampered images extremely difficult [69].

Moreover, many generative AI applications in health care rely on LLMs and are trained on large amounts of internet data without being properly screened and filtered [70]. Adversaries can use AI technologies to automatically generate large quantities of fake data to poison data to be fed into LLMs, resulting in deteriorated performance of the models (eg, accuracy and fairness) and eventually AI hallucination, misinformation or disinformation, and deepfakes. Although some of these threats are not unique to generative AI in health care, they can be particularly risky if false information is used for medical decision-making. Generative AI also carries unique integrity risks. As mentioned before, its capability to create synthetic data leads to a unique integrity risk—AI hallucination. In the health care context, generative AI in health care could be used to create fake medical records or alter existing ones. Fabricated medical data can be fed again into LLMs, further threatening the integrity of medical information. For instance, the malicious use of deepfakes generated by deep generative models could fabricate a patient's medical history to falsely claim insurance or lead to incorrect treatments. Another example is that a generative AI model may create synthetic radiology reports to diagnose nonexistent medical conditions, leading to misdiagnosis or unnecessary treatment.

By contrast, research has used synthetic data in AI for medicine and health care to address the scarcity of annotated medical data in the real world [71]. For instance, deep generative models are used to create synthetic images such as skin lesions, pathology slides, colon mucosa, and chest x-rays, thereby greatly improving the reproducibility of medical data [71]. With the development of generative AI, researchers have increasingly used GANs to synthesize realistic training data for data imputation when the data lacks distribution. Noise-to-image and image-to-image GANs have been used to synthesize realistic training magnetic resonance imaging images to boost the performance of convolutional neural networks for image diagnostic AI [39,72]. CorGAN [42] synthesizes discrete and continuous health care records for model training. From a broader perspective, generative AI is projected to build and use next-generation synthetic gene networks for various AI applications in health care, including medical diagnostics, drug discovery, and medical research [73]. The growth in the use of synthetic data by generative AI also creates new concerns about data integrity and AI hallucination. Nevertheless, given that health care is a heavily regulated field in terms of patient privacy and safety, researchers even claim that synthetic medical data might be promising to overcome data sharing obstacles for health care AI and free developers from sensitive patient information [74]. These applications indicate that there is a fine line between harmful AI hallucinations or deepfakes and beneficial synthetic data use by generative AI in health care. Nevertheless, even the benevolent use of synthetic medical data faces privacy and security challenges as well as integrity challenges. Deep-faked patient face images could violate patient privacy and lead to the leakage or exploitation of PHI [75]. How to navigate this fine line is both a policy and research blind spot. Currently, there are just insufficient use cases, especially for rare use cases, to establish clinical reference standards such as clinical quality measures and evaluation metrics to assess risks and benefits.

Similar to generative AI applications in other fields, almost all types of generative AI in health care face confidentiality threats. Deidentified data may become identifiable during the data collection and processing phase, and confidential proprietary medical information, such as drug development and treatment plans, may be inferred during the data collection and processing phase [76], leading to data and privacy breaches. Research has found that genomic databases are prone to privacy violations. For example, legit researchers obtain or recover the whole or partial genomic sequence of a target individual (privacy violation through reference), link the sequence to a target individual (ie, reidentifying), and identify the group of interest of a target individual (privacy violation through membership reference) when processing data from multiple sources. In addition, the growth of synthetic medical data in health AI systems raises concerns about the vulnerabilities of such systems and the challenges of the current regulations and policies.

Table 2 summarizes the data sources and security or privacy threats for each type of generative AI in health care in the data collection and processing phase.

XSL•FO
RenderX

Table 2. Generative AI[a] in health care categories, data sources, and security or privacy threats in the data collection and processing phase.

AI categories	Data source	Security and privacy threats	
		Unintentional (integrity and privacy threats)	Intentional (availability and integrity attacks)
Medical diagnostics	1. Medical images (eg, x-rays, CT[b] scans, MRI[c] scans, pet scans, and microscopy images) 2. Patient reports and EHRs[d] (eg, laboratory results, comorbidities, and symptoms) 3. Clinical measurements (vital signs, tumor measurements, and fluid output) 4. Patient metadata (demographics and family history) 5. Expert annotations to train models	1-4: Incorrect, missing, or incomplete patient data or images occur owing to hardware or software errors, measurement and label errors, and human errors (eg, distorted images, partial images, and mismatched data or laboratory results or images) 1-4: Data integration errors occur when integrating data from various sources (eg, by mislabeling data attributes and mismatching patient information with their images and laboratory results) 1-4: Organic biases occur because of the nature of the disease and the demographics of patients, and selection biases rise because of human biases 5. Annotation errors and biases occur in all sources of data because of expert mistakes and human biases 1-4: Errors and bias in synthetic data or images 1-7: Privacy breaches (eg, reidentify patients)	1-3: Software tampering, medical sensor spoofing, medical equipment tampering or poisoning (eg, CT and MRI scanning equipment tampering), medical image tampering (eg, image scaling, copy-move tampering, sharpening, blurring, and resampling), generative fake data and images (eg, generative fake CT and MRI images undetectable by both human experts and generative AI), and medial data tampering or poisoning (eg, noise injection and maliciously synthesized data) 5: Annotation errors by intention
Drug discovery	1. Genomic databases (DNA or RNA sequencing data) 2. Chemical compound or protein structure databases 3. Bioactivity assay data (in vivo and in vitro) 4. Disease or treatment knowledge bases (peer-reviewed findings) 5. Patient clinical trial data 6. Toxicity predictions from pharmacokinetic models	1-2: Duplication issues (eg, sequence redundancies or sequence duplications with minor variations), structural errors, and assembly or carried-over errors owing to poor data quality of sources 1-6: Data integration errors occur when integrating data from various sources 4-5: Wrong findings and errors in trials 1-6: Missing and incomplete data, missing or incorrect annotations, and human errors 1-6: Errors and bias in synthetic data 6: Incorrect or inaccurate models 1-7: Privacy breaches (eg, reidentify patients)	1-5: Genomic data tampering or poisoning (eg, maliciously forge and inject structures or sequences, analyses, and findings) 1-5: Annotation errors by intention 6: Model tampering
Virtual health assistants	1. EHRs 2. Insurance claims data 3. Patient symptom reports 4. Mobile health data: data collected from mobile apps 5. Speech and text inputs: data from patient interactions, including spoken dialogue and written communication 6. Digitized medical reference information (guides and protocols) 7. Custom health care knowledge bases	1-5: Incorrect, missing, or incomplete patient data 1-7: Data integration errors occur when integrating data from various sources 1-7: Organic biases occur because of the nature of the disease and the demographics of patients, and selection biases rise because of human biases 2: Errors owing to unknown fraudulent claims 6: Incorrect or inaccurate models 5-7: Errors and bias in synthetic data and AI hallucination 1-7: Privacy breaches (eg, reidentify patients)	1-7: Data or records tampering or poisoning (eg, noise injection using maliciously synthesized data, analyses, and findings) 1-7: Annotation errors by intention 1-7: AI hallucination
Medical research	1. Clinical trial and study data sets 2. Epidemiological data from public health departments 3. Biomedical publications and preprint archives 4. Physician's notes and patient diagnosis histories 5. Genomics databases 6. NIH[e] open-source data repositories 7. Biobanks: collections of biological samples	1-7: All the errors and biases mentioned in the above cells could be applicable	1-7: All the attacks mentioned in the above cells could be applicable

XSL•FO
RenderX

AI categories	Data source	Security and privacy threats	
		Unintentional (integrity and privacy threats)	Intentional (availability and integrity attacks)
Clinical decision support	1. Real-time patient data feeds (vitals, laboratory results, etc) 2. EHRs 3. Population health data 4. Hospital medical reference or treatment protocol guides 5. Custom evidence-based clinical rules or guidelines 6. Medical insurance claims data 7. g. Pharmaceutical reference database	1-7: All the errors and biases mentioned in the above cells could be applicable	1-7: All the attacks mentioned in the above cells could be applicable

[a]AI: artificial intelligence.

[b]CT: computed tomography.

[c]MRI: magnetic resonance imaging.

[d]EHR: electronic health record.

[e]NIH: National Institutes of Health.

Again, it should be noted that although all AI and machine learning systems face many similar threats, as listed in Table 2, generative AI amplifies them because of its generating nature and data source volume and complexity. For example, generative medical research AI may update knowledge and literature databases with "wrong inputs" based on wrong findings in these databases or with synthesized but hallucinated findings. Similarly, generative virtual health assistants may put dangerous advice into knowledge databases based on erroneous data from sources or again put synthesized but hallucinated advice into such databases.

Model Training and Building Phase

Generative AI also encounters integrity issues, leading to phenomena such as AI hallucinations during model training and development phases. This is especially true for generative AI in health care. Prior research found that generative AI created nonfactual or unfaithful data and outputs [72,77]. The growing use of highly synthetic data or images by generative AI, such as CorGAN, exacerbates the situation as it becomes increasingly challenging for human professionals to detect unfaithful data and outputs [69]. This can be a serious integrity and authenticity issue, as both patients and clinicians expect factual, scientific answers or outputs with consistency from such models. Technically speaking, similar to all other AI models, generative AI models in health care, particularly those based on deep learning, are often seen as "black boxes" [78]. The lack of interpretability and explainability can be a significant challenge in health care, where understanding the reasoning behind a diagnosis or treatment recommendation is crucial for integrity and accountability.

Adversarial training is a method to check for the integrity and accountability of AI models. The method uses carefully crafted adversarial examples to attack the training model to check for the integrity and robustness of outputs [57,79]. It is an active AI research area in the health care field. Adversarial training is used to check for fake or realistic features in synthetic medical images created by GANs to avoid fabrication and misleading in the model training process. By contrast, malicious parties also intensively explore this method and use adversarial examples to attack training models to generate incorrect outcomes [57]. Technically, all types of generative AI using GANs and LLMs, particularly those in health care, can be attacked with adversarial examples that compromise the integrity of the training model. For example, adversaries can use image-scaling attacks to feed human-invisible data into an AI model to force it to make a mistake [67,68].

Another example is to feed an AI model with carefully crafted relabeled data to create the wrong classification [80]. When being trained with adversarial examples, a diagnostic AI could make an incorrect diagnosis, a conversational virtual assistant could offer harmful advice to patients, and a clinical decision support AI could make the wrong recommendations, to list a few. Moreover, feeding an AI model with adversarial training examples and other poisonous data can also deteriorate the performance of AI, eventually making the AI model useless and thus unavailable. In general, adversarial attacks can pose long-term risks, such as thwarting AI innovation in health care because of concerns about misdiagnosis, mistreatment, and patient safety.

Implementation Phase

In practice, generative AI systems in health care have been found to experiencing integrity threats, such as generating disinformation and misinformation, and making biased decisions [81]. AI hallucination is a newly-coined term describing the phenomenon wherein generative AI generates fake information that appears authentic [82]. If generative AI in health care is used for diagnostics, personalized medicine, or clinical assistance, AI hallucination can be extremely dangerous and may even harm patients' lives [83]. As discussed before, because GANs and LLMs need large annotated medical data for training, the difficulty of acquiring such data (eg, unwillingness to share because of legal compliance requirements and data paucity resulting from rare medical conditions) leads to the proliferation of synthetic medical data creation. The relationship between AI

XSL•FO
RenderX

hallucination by GANs and LLMs and synthetic data use is an unknown territory in research and practice, leading to unknown vulnerabilities such as adversarial attacks.

Privacy attacks are a grave concern at this stage. The use of GANs for creating synthetic EHRs and its associated privacy challenges are analyzed by Ghosheh et al [46]. Such privacy challenges are as follows: (1) risk of reidentification—although the data are synthetic, there might be a risk of reidentifying individuals if the synthetic data closely resemble real patient data; (2) data leakage—ensuring that the synthetic data do not leak sensitive information from the original data set; (3) model inversion attacks—potential for attackers to use the GAN model to infer sensitive information about the original data set. In this attack, attackers aim to reconstruct the training data using their ability to constantly query the model [84]; (4) membership inference attacks—an attacker gains access to a set of real patient records and tries to determine whether any of the real patients are included in the training set of the GAN model [85]; and (5) attribute disclosure attacks—an attacker can infer additional attributes about a patient by learning a subset of other attributes about the same patient [86].

Generative medical diagnosis and drug discovery AI involving genomic databases and chemical compound or protein structure databases are extremely susceptible to privacy attacks. Fernandes et al [87] pointed out that genomic data such as DNA data are susceptible to inference attacks, reidentification attacks, membership attacks, and recovery attacks. It is extremely concerning when such attacks target high-profile individuals. Moreover, generative AI enhances the ability to profile patients, thereby increasing the risk of privacy violations and attacks, although this capability is not unique to AI.

In addition to AI-specific security and privacy threats, AI systems interfacing with other hardware and software may face new security and privacy threats that have never existed before [57]. Malicious use and exploitation may also threaten the integrity of AI systems. Similar to other AI systems, health care AI systems, especially generative AI systems, are susceptible to code extraction and information extraction (eg, black-box, gray-box, and white-box attacks), leading to security and privacy breaches [57]. The excessive use of prompts may reveal copyright-protective data, proprietary research findings (eg, chemical compounds of a new drug), and training models or algorithms.

Table 3 summarizes the previously discussed security and privacy threats associated with each category of generative AI systems throughout their life cycle in health care.

Table 3. Generative artificial intelligence (AI) in health care categories and security or privacy threats in model training or building and implementation phases.

Category	Model training and building phase		Implementation phase	
	Integrity threats	Availability threats	Integrity threats	Confidentiality threats
Medical diagnostics	Adversarial training and classification manipulation (eg, image classification manipulation)	Model performance deteriorating by feeding poisonous data	AI hallucination (eg, made-up diagnosis), misinformation or disinformation, and adversarial use exploitation	Data extraction from carefully crafted prompts and privacy attacks
Drug discovery	Adversarial training and classification manipulation	Model performance deteriorating by feeding poisonous data	AI hallucination (eg, made-up chemical compound or protein structures), misinformation or disinformation, and adversarial use exploitation	Data extraction from carefully crafted prompts and privacy attacks
Virtual health assistants	Adversarial training and classification manipulation	Model performance deteriorating by feeding poisonous data	AI hallucination (eg, made-up medical advice), misinformation or disinformation, and adversarial use exploitation	Data extraction from carefully crafted prompts and privacy attacks
Medical research	Adversarial training and classification manipulation	Model performance deteriorating by feeding poisonous data	AI hallucination (eg, made-up findings, hypothesis, and citations), misinformation or disinformation, and adversarial use exploitation	Data extraction from carefully crafted prompts and privacy attacks
Clinical decision support	Adversarial training and classification manipulation	Model performance deteriorating by feeding poisonous data	AI hallucination (eg, made-up conclusions, findings, and recommendations), misinformation or disinformation, and adversarial use exploitation	Data extraction from carefully crafted prompts and privacy attacks

Again, it should be noted that some of these threats are unique to generative AI systems, but many of the threats are prevalent in broader AI systems in health care and other fields.

Recommendations

Overview

As security and privacy threats exist in the life cycle of various generative AI systems in health care, from data collection through model building to clinical implementation, a systematic approach to safeguard them is critical. This section provides some recommendations on safeguards. In doing so, we rely on the National Institute of Standards and Technology Privacy Framework and the National Institute of Standards and Technology AI Risk Management Framework as well as the regulatory guidance discussed in the Literature Review section. It should be noted that although the security and privacy threats discussed in this study are significant and some are unique in the context of generative AI in health care, many are also common in other types of AI models and other AI application

XSL•FO
RenderX

contexts. Hence, many of the recommendations we propose in the subsequent section can be applied to AI in non–health care contexts.

Development Protocols of Risk Assessment for Generative AI in Health Care

AI risks, including those of generative AI in health care, can emerge in a variety of ways at any phase of an AI project. Health care organizations need to learn from managing risks for other technologies to develop risk assessment protocols for generative AI in health care, along with risk assessment metrics.

AI Risk Assessment Protocols

To systematically manage AI risks, health care organizations must develop risk assessment protocols that include risk assessment procedures and methodologies by following industrial standards and frameworks as well as best practices [63]. A total of 3 main risk assessment activities are involved in the protocol development: risk identification, risk prioritization, and risk controls. All 3 activities must be conducted throughout the life cycle of a generative AI system in health care.

In the data collection and processing phase, health care organizations can use several methods to identify, prioritize, and control AI risks. As discussed before, health care data are messy and tend to have organic biases (eg, a hospital specializes in serving a particular patient demographic, attending to gender-specific health requirements or offering dedicated care for rare diseases). When collecting data or using GANs to generate synthetic data, the health care field needs to be extremely diligent. One recommendation is to establish data collection or generation policies and procedures. The separation of clinical and nonclinical data is necessary, given the significantly different risks in these 2 types of data. Similarly, the establishment of the metrics and methods to check training data on biases for clinical and nonclinical data is also important. Data provenance and authentication metrics can be used to prevent collecting data from untrustworthy sources; detecting and filtering methods can be used to identify and filter poisoned data; and data standardization improves the quality of data collection [57]. As the frontline defense, these prevention mechanisms can prevent integrity and availability attacks during this phase. Nevertheless, regardless of the mechanisms, data collected from medical sources or generated by GANs should reflect the comprehensive overview of a medical domain and the complexity of the physical and digital dimensions in such a domain to prevent biases and test for risks.

In the model training and building phase, detecting and filtering are also important for identifying and removing adversary training examples. Robustness, generalizability, and other vulnerability tests (eg, black-box and white-box tests) can further prevent integrity and availability attacks and data breaches [88]. Input reconstruction is another mechanism to pinpoint sources of adversary training [89]. Modifying training processes and models as well as training methods may also help to control AI risks in this phase [57]. Given the complexity and variety of AI models in reasoning and learning, we suggest a taxonomy approach. For example, a deep learning model can carry significantly different risks than a probabilistic learning model. By building a taxonomy of AI models and their risks, researchers can systematically identify and control security and privacy risks based on the AI model.

In the model implementation phase, routine verification and validation are key to identifying and controlling AI risks [63]. The implementation contexts of generative AI also matter. In some cases, verification and validation are about not only factual accuracy but also communications and perceptions as well as cultures. A medical chatbot that was thoroughly tested in adult populations may not be very useful in teenage populations. Gesture and face recognition AI for medical diagnosis may need to be culturally sensitive to be useful. When generative AI is integrated and interacts with other systems, for example, to create multiagent systems or medical robotics (eg, companion robots), security tests along with social, philosophical, and ethical tests are a must.

AI Risk Assessment Metrics

Given the complexity of AI security and privacy risks, health care organizations should develop risk assessment metrics for each of the 3 phases of the life cycle of a generative AI project. The following subsections highlight some measures for AI risk assessment metrics.

Security Objectives

AI risk assessment metrics should include well-established security and privacy objectives such as confidentiality, integrity, availability, nonrepudiation, authentication, and privacy protection. In the data collection and processing phase, collection technologies should be evaluated regardless of software- or hardware-based collection to ensure that they meet the security and privacy objectives. The use of synthetic medical data should follow the same security and privacy objectives to ensure that such data capture the factual and scientific truth. In the model training and building phase, vulnerability tests should be conducted to identify known and unknown threats based on security objectives. For example, availability attacks such as denial of service can be used to flood conversational health AI applications to assess their resilience and availability before deployment, and integrity attacks with poisoned data can be used to test the stability of model performance and generalizability [57]. In the implementation phase, all security objectives should be routinely assessed.

Generative AI–Specific Metrics

AI Inscrutability

AI inscrutability refers to the lack of understandability of an AI model and its outcomes [63]. Although AI inscrutability is not directly related to security and privacy, it adds obfuscations to AI risk assessment to identify threats and vulnerabilities as well as biases owing to the lack of transparency and explainability in AI, especially in generative AI based on deep learning. Although we have identified AI inscrutability as a key metric for generative AI assessment, we acknowledge that the challenge of inscrutability is not unique to generative AI and has been a long-standing issue in the broader field of AI, particularly in health care. Various algorithms used in patient matching, diagnosis, and other proprietary applications often lack

XSL•FO
RenderX

transparency because of their closed nature or intellectual property constraints. Therefore, many of them, even those that are not based on generative techniques, face similar scrutiny regarding their lack of transparency. Hence, the call for greater openness and explainability applies broadly across AI applications in health care, reflecting a growing demand for accountable and interpretable AI systems.

Nevertheless, the problem of inscrutability becomes pronounced in the context of generative AI because of its complex and often opaque decision-making processes, which can amplify the challenges already faced in health care AI. Generative AI models, especially when based on deep learning, can operate as "black boxes," making it even more difficult for practitioners to understand how conclusions or recommendations are derived. This opacity is a critical concern in health care, where explainability and trust as well as accountability are paramount for clinical acceptance and ethical practice.

To address these concerns, there is a need for concerted efforts toward developing more interpretable AI models and regulatory frameworks that mandate transparency in AI applications, including those used in patient care. These efforts should be complemented by initiatives to educate health care professionals about the workings and limitations of AI tools, enabling them to make informed decisions while using these technologies in clinical settings. Therefore, although the inscrutability of generative AI presents specific challenges owing to the complexity and novelty of these models, it is a continuation of the broader issue of transparency in health care AI. Recognizing this, our discussion of AI inscrutability not only highlights the unique aspects of generative AI but also situates it within the ongoing discourse on the need for greater transparency and accountability in all AI applications in health care.

AI Trustworthiness

AI trustworthiness is defined as the degree to which stakeholders of an AI system have confidence in its various attributes [63,90]. Trust has been a significant factor in IT adoption. The fundamental argument is that if an IT system automatically runs behind the scenes to assist the work and decisions of human users, a trusting relationship must be established for users to interact with and rely on the system [91]. Nevertheless, trust is a complex concept and is built upon human users' interaction and consequent assessment of the system from cognitive, emotional, and social dimensions [91-93]. Since the emergence of AI, AI trustworthiness has caught significant attention in research, given the foreseeable complexity of human-AI interaction. The rise of generative AI has stimulated more discussions on this topic. The current consensus is that AI trustworthiness itself is a complex measurement with multiple dimensions, such as reliability, resilience, accuracy, and completeness [63,90]. Many other AI metrics or factors, such as transparency, explainability, robustness, fairness, and user interactions or perceptions, can be the antecedents of AI trustworthiness. AI trustworthiness can also be context dependent. For example, explainability and interaction experience can be the determinants of the AI trustworthiness of a chatbot application on the patient portal, whereas reliability, accuracy, and completeness are significant factors in the AI trustworthiness of a radiology diagnosis AI for radiologists. Given the complexity of measuring AI trustworthiness, we recommend developing context-specific AI trustworthiness metrics. Similar to AI inscrutability, although AI trustworthiness is not a direct measure of security and privacy risks, it helps reduce the probability and magnitude of such risks throughout the life cycle of generative AI in health care. For instance, accuracy and reliability help to improve the integrity of an AI system.

AI Responsibility

AI responsibility is another key measure in AI risk assessment. Again, although this measure does not directly evaluate security and privacy risks, it endorses responsible AI practices that facilitate the discovery of the negative consequences and risks of AI, including the security and privacy risks of generative AI. Moreover, this measure is centered on the uniqueness of AI, especially generative AI, in "human centricity, social responsibility, and sustainability" [63]. In other words, AI responsibility is a multifaceted measure depending on many other metrics and factors such as the ethical framework (eg, biases, fairness, and transparency) and legal perspective (eg, accountability and traceability). This is also an emerging concept that is under development. The development and deployment of generative AI add complexity to this measure owing to its possible, unintended, but profound negative consequences and risks to human society. In health care, there is a legal ambiguity related to AI responsibility. Hospitals are still unclear about their legal liability when facing an AI incident. Despite such legal uncertainty, responsible AI use should be the baseline. We recommend that health care organizations use AI for consultation and assistance instead of replacement, given legal ambiguity and uncertainty, while intensively exploring generative AI from the perspectives of patient centricity and social responsibility and asking serious questions. For example, a generative drug discovery AI may find a new molecular formula for a biochemical weapon. How can we responsibly use such AI without crossing the line of no harm to human beings? Such a question leads to another key measure for AI risk assessment—AI harm.

AI Harm

AI harm can occur to individuals, organizations, and societies. For example, AI may cause physical harm to individual patients, damage a hospital's reputation owing to AI incidents, and even endanger society if it is weaponized (eg, being used to disrupt the global drug manufacturing and supply chain). Hence, AI harm is a risk measure highly related to AI responsibility and trustworthiness. Developing trustworthy AI and following responsible AI practices can reduce or avoid AI harm.

It is worth mentioning that some of the metrics we proposed here pass some human characteristics into AI. A crucial philosophical distinction must be made regarding the attribution of human characteristics such as trustworthiness and responsibility to generative AI systems versus the health care organizations and technology partners developing these algorithms. Although metrics aim to make models appear more trustworthy and responsible in reality, trust emerges from human-centered institutional processes, and responsibility stems

XSL•FO
RenderX

from human accountability. It may be challenging to humanize AI systems and transfer attributes such as trustworthiness to the algorithms themselves. Indicators of model transparency, reliability, or accuracy may engender confidence among stakeholders, but public trust fundamentally arises from the ethical data governance, risk communication, and oversight procedures instantiated by organizations. Without robust governance and review processes overseeing development, data practices, and risk monitoring, claims of AI trustworthiness lack substantiation. Similarly, although algorithmic outputs highlighting potential issues such as biases or errors increase awareness, this does not intrinsically amount to AI responsibility. True accountability involves diligent human investigation of problems that surface, enacting appropriate recourse, and continuous authority oversight. Metrics may aim for AI to appear more responsible, but responsibility mainly manifests in organizational commitment to discovering issues, working with experts to properly assess AI harms, and instituting robust redress processes with stakeholder input. Thus, trustworthiness and responsibility are contingent on extensive institutional support structures rather than innate model capabilities. Although progress indicators may serve as signals for these desired attributes, establishing genuine public trust and accountability in health care ultimately falls on the shoulders of health care administrators, innovators, and engaged communities, rather than solely on the algorithms themselves. Clarifying this distinction enables us to properly set expectations and delineate responsibilities as generative AI becomes increasingly prevalent in critical medical settings.

Conclusions

Integrating generative AI systems into health care offers immense potential to transform medical diagnostics, research, treatment planning, and patient care. However, deploying these data-intensive technologies also introduces complex privacy and security challenges that must be proactively addressed to ensure the safe and effective use of these systems. Examining diverse applications of generative AI across medical domains (ie, medical diagnostics, drug discovery, virtual health assistants, medical research, and clinical decision support) helps this study uncover vulnerabilities and threats across the life cycle of these systems, from data collection to model development to clinical implementation. Although generative AI enables innovative use cases, adequate safeguards are needed to prevent breaches of PHI and to maintain public trust. Strategies such as developing AI risk assessment protocols; formulating specific metrics for generative AI such as inscrutability, trustworthiness, responsibility, and harm; and ongoing model monitoring can help mitigate risks. However, developing robust governance frameworks and updates to data privacy regulations are also required to oversee these rapidly evolving technologies. By analyzing the use cases, impacts, and risks of generative AI across diverse domains within health care, this study contributes to theoretical discussions surrounding AI ethics, security vulnerabilities, and data privacy regulations. Future research and development in generative AI systems should emphasize security and privacy to ensure the responsible and trustworthy use of these AI models in health care. Moreover, the security and privacy concerns highlighted in this analysis should serve as a call to action for both the AI community and health care organizations looking to integrate generative AI. Collaborative efforts between AI developers, health care providers, policy makers, and domain experts will be critical to unlocking the benefits of generative AI while also prioritizing ethics, accountability, and safety. By laying the groundwork to make security and privacy the central pillars of generative AI in medicine, stakeholders can work to ensure that these transformative technologies are harnessed responsibly for patients worldwide.

Conflicts of Interest

None declared.

References

1. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. Am J Med 2019 Jul;132(7):795-801 [FREE Full text] [doi: 10.1016/j.amjmed.2019.01.017] [Medline: 30710543]
2. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
4. Li X, Jiang Y, Rodriguez-Andina JJ, Luo H, Yin S, Kaynak O. When medical images meet generative adversarial network: recent development and research opportunities. Discov Artif Intell 2021 Sep 22;1(1):1-20 [FREE Full text] [doi: 10.1007/s44163-021-00006-0]
5. Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science 2023 Sep 15;381(6663):adk6139. [doi: 10.1126/science.adk6139] [Medline: 37708283]
6. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. Opinion paper: “so what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 2023 Aug;71:102642. [doi: 10.1016/j.ijinfomgt.2023.102642]

XSL•FO
RenderX

7. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Alqahtani H, Kavakli-Thorne M, Kumar G. Applications of generative adversarial networks (GANs): an updated review. Arch Computat Methods Eng 2019 Dec 19;28(2):525-552. [doi: 10.1007/s11831-019-09388-y]
9. Jain S, Seth G, Paruthi A, Soni U, Kumar G. Synthetic data augmentation for surface defect detection and classification using deep learning. J Intell Manuf 2020 Nov 18;33(4):1007-1020. [doi: 10.1007/s10845-020-01710-x]
10. Arora A, Arora A. The promise of large language models in health care. Lancet 2023 Feb;401(10377):641. [doi: 10.1016/s0140-6736(23)00216-7]
11. Zeng X, Wang F, Luo Y, Kang S, Tang J, Lightstone FC, et al. Deep generative molecular design reshapes drug discovery. Cell Rep Med 2022 Dec 20;3(12):100794 [FREE Full text] [doi: 10.1016/j.xcrm.2022.100794] [Medline: 36306797]
12. Jiang S, Hu J, Wood KL, Luo J. Data-driven design-by-analogy: state-of-the-art and future directions. J Mech Des 2022;144(2):020801. [doi: 10.1115/1.4051681]
13. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. TBench 2023 Feb;3(1):100105. [doi: 10.1016/j.tbench.2023.100105]
14. Nova K. Generative AI in healthcare: advancements in electronic health records, facilitating medical languages, and personalized patient care. J Adv Anal Healthc Manag 2023;7(1):115-131 [FREE Full text]
15. Zhang P, Kamel Boulos MN. Generative AI in medicine and healthcare: promises, opportunities and challenges. Future Internet 2023 Aug 24;15(9):286. [doi: 10.3390/fi15090286]
16. Byrne DW. Artificial Intelligence for Improved Patient Outcomes: Principles for Moving Forward with Rigorous Science. Philadelphia, PA: Lippincott Williams & Wilkins; 2022.
17. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Amsterdam, The Netherlands: Elsevier Academic Press; 2020.
18. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today 2021 Jan;26(1):80-93 [FREE Full text] [doi: 10.1016/j.drudis.2020.10.010] [Medline: 33099022]
19. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 2022 Aug 17;56(4):3005-3054. [doi: 10.1007/s10462-022-10246-w]
20. Martín-Noguerol T, Oñate Miranda MO, Amrhein TJ, Paulano-Godino F, Xiberta P, Vilanova JC, et al. The role of Artificial intelligence in the assessment of the spine and spinal cord. Eur J Radiol 2023 Apr;161:110726. [doi: 10.1016/j.ejrad.2023.110726] [Medline: 36758280]
21. Ellis RJ, Sander RM, Limon A. Twelve key challenges in medical machine learning and solutions. Intell Based Med 2022;6:100068. [doi: 10.1016/j.ibmed.2022.100068]
22. Martinelli DD. Generative machine learning for de novo drug discovery: a systematic review. Comput Biol Med 2022 Jun;145:105403. [doi: 10.1016/j.compbiomed.2022.105403] [Medline: 35339849]
23. Kasirzadeh A, Gabriel I. In conversation with artificial intelligence: aligning language models with human values. Philos Technol 2023 Apr 19;36(2):1-24. [doi: 10.1007/s13347-023-00606-x]
24. van Bussel MJ, Odekerken-Schröder GJ, Ou C, Swart RR, Jacobs MJ. Analyzing the determinants to accept a virtual assistant and use cases among cancer patients: a mixed methods study. BMC Health Serv Res 2022 Jul 09;22(1):890 [FREE Full text] [doi: 10.1186/s12913-022-08189-7] [Medline: 35804356]
25. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021 Nov 29;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
26. Summerfield C. Natural General Intelligence: How Understanding the Brain Can Help Us Build AI. Oxford, UK: Oxford University Press; 2022.
27. Gesk TS, Leyer M. Artificial intelligence in public services: when and why citizens accept its usage. Gov Inf Q 2022 Jul;39(3):101704. [doi: 10.1016/j.giq.2022.101704]
28. Wang Z, Qinami K, Karakozis IC, Genova K, Nair P, Hata K. Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 Presented at: CVPR '20; June 13-19, 2020; Seattle, WA p. 8916-8925 URL: https://ieeexplore.ieee.org/document/9156668 [doi: 10.1109/cvpr42600.2020.00894]
29. Cai Z, Xiong Z, Xu H, Wang P, Li W, Pan Y. Generative adversarial networks: a survey toward private and secure applications. ACM Comput Surv 2021 Jul 13;54(6):1-38. [doi: 10.1145/3459992]
30. Huang C, Kairouz P, Chen X, Sankar L, Rajagopal R. Context-aware generative adversarial privacy. Entropy 2017 Dec 01;19(12):656. [doi: 10.3390/e19120656]
31. Tripathy A, Wang Y, Ishwar P. Privacy-preserving adversarial networks. In: Proceedings of the 57th Annual Allerton Conference on Communication, Control, and Computing. 2019 Presented at: ALLERTON '19; September 24-27, 2019; Monticello, IL p. 495-505 URL: https://ieeexplore.ieee.org/document/8919758 [doi: 10.1109/allerton.2019.8919758]
32. Chen CS, Chang SF, Liu CH. Understanding knowledge-sharing motivation, incentive mechanisms, and satisfaction in virtual communities. Soc Behav Pers 2012 May 01;40(4):639-647. [doi: 10.2224/sbp.2012.40.4.639]

XSL•FO
RenderX

33. Liu S, Shrivastava A, Du J, Zhong L. Better accuracy with quantified privacy: representations learned via reconstructive adversarial network. arXiv Preprint posted online January 25, 2019 2019 [FREE Full text] [doi: 10.1090/mbk/121/79]
34. Tseng BW, Wu PY. Compressive privacy generative adversarial network. IEEE Trans Inf Forensics Secur 2020;15:2499-2513. [doi: 10.1109/tifs.2020.2968188]
35. Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015 Presented at: CCS '15; October 12-16, 2015; Denver, CO p. 1310-1321 URL: https://dl.acm.org/doi/10.1145/2810103.2813687 [doi: 10.1145/2810103.2813687]
36. McMahan B, Moore E, Ramage D, Hampson S. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017 Presented at: AISTATS '17; April 20-22, 2017; Fort Lauderdale, FL URL: https://proceedings.mlr.press/v54/mcmahan17a?ref=https://githubhelp.com
37. Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018 Presented at: CCS '18; October 15-19, 2018; Toronto, ON p. 634-646 URL: https://dl.acm.org/doi/10.1145/3243734.3243855 [doi: 10.1145/3243734.3243855]
38. Abadir PM, Chellappa R, Choudhry N, Demiris G, Ganesan D, Karlawish J, et al. The promise of AI and technology to improve quality of life and care for older adults. Nat Aging 2023 Jun 25;3(6):629-631 [FREE Full text] [doi: 10.1038/s43587-023-00430-0] [Medline: 37231197]
39. Kim BN, Dolz J, Jodoin PM, Desrosiers C. Privacy-net: an adversarial approach for identity-obfuscated segmentation of medical images. IEEE Trans Med Imaging 2021 Jul 27;40(7):1737-1749 [FREE Full text] [doi: 10.1109/TMI.2021.3065727] [Medline: 33710953]
40. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: Proceedings of the 2017 Machine Learning for Health Care Conference. 2017 Presented at: MLHC '17; August 18-19, 2017; Boston, MA p. 1-20 URL: https://proceedings.mlr.press/v68/choi17a/choi17a.pdf
41. Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP. Generation and evaluation of privacy preserving synthetic health data. Neurocomput 2020 Nov;416:244-255. [doi: 10.1016/j.neucom.2019.12.136]
42. Torfi A, Fox EA. CorGAN: correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. arXiv Preprint posted online January 25, 2020 2020 [FREE Full text]
43. Lee D, Yu H, Jiang X, Rogith D, Gudala M, Tejani M, et al. Generating sequential electronic health records using dual adversarial autoencoder. J Am Med Inform Assoc 2020 Jul 01;27(9):1411-1419 [FREE Full text] [doi: 10.1093/jamia/ocaa119] [Medline: 32989459]
44. Rane N. ChatGPT and similar generative artificial intelligence (AI) for smart industry: role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0. SSRN J 2023 [FREE Full text] [doi: 10.2139/ssrn.4603234]
45. Bale AS, Dhumale R, Beri N, Lourens M, Varma RA, Kumar V, et al. The impact of generative content on individuals privacy and ethical concerns. Int J Intell Syst Appl Eng 2023;12(1):697-703 [FREE Full text]
46. Ghosheh GO, Li J, Zhu T. A survey of generative adversarial networks for synthesizing structured electronic health records. ACM Comput Surv 2024 Jan 22;56(6):1-34. [doi: 10.1145/3636424]
47. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: a systematic review. Neurocomput 2022 Jul;493:28-45. [doi: 10.1016/j.neucom.2022.04.053]
48. Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and ChatGPT: applications, challenges, and AI-human collaboration. J Inf Technol Case Appl Res 2023 Jul 21;25(3):277-304. [doi: 10.1080/15228053.2023.2233814]
49. Ali M, Naeem F, Tariq M, Kaddoum G. Federated learning for privacy preservation in smart healthcare systems: a comprehensive survey. IEEE J Biomed Health Inform 2023 Feb;27(2):778-789. [doi: 10.1109/jbhi.2022.3181823]
50. Khan S, Saravanan V, Lakshmi TJ, Deb N, Othman NA. Privacy protection of healthcare data over social networks using machine learning algorithms. Comput Intell Neurosci 2022 Mar 24;2022:9985933-9985938 [FREE Full text] [doi: 10.1155/2022/9985933] [Medline: 35371203]
51. Sun H, Zhu T, Zhang Z, Jin D, Xiong P, Zhou W. Adversarial attacks against deep generative models on data: a survey. IEEE Trans Knowl Data Eng 2023 Apr 1;35(4):3367-3388. [doi: 10.1109/tkde.2021.3130903]
52. The legal issues presented by generative AI. MIT Sloan School of Management. URL: https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai [accessed 2024-01-29]
53. EU AI Act: first regulation on artificial intelligence. European Parliament. 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2024-02-16]
54. Blueprint for an AI bill of rights: making automated systems work for the American people. The White House. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ [accessed 2024-02-19]
55. Ahmad K, Maabreh M, Ghaly M, Khan K, Qadir J, Al-Fuqaha A. Developing future human-centered smart cities: critical analysis of smart city security, Data management, and Ethical challenges. Comput Sci Rev 2022 Feb;43:100452. [doi: 10.1016/j.cosrev.2021.100452]
56. Liu Q, Li P, Zhao W, Cai W, Yu S, Leung VC. A survey on security threats and defensive techniques of machine learning: a data driven view. IEEE Access 2018;6:12103-12117. [doi: 10.1109/access.2018.2805680]

XSL•FO
RenderX

57. Hu Y, Kuang W, Qin Z, Li K, Zhang J, Gao Y, et al. Artificial intelligence security: threats and countermeasures. ACM Comput Surv 2021 Nov 23;55(1):1-36. [doi: 10.1145/3487890]
58. Tang X, Yin P, Zhou Z, Huang D. Adversarial perturbation elimination with GAN based defense in continuous-variable quantum key distribution systems. Electronics 2023 May 27;12(11):2437. [doi: 10.3390/electronics12112437]
59. Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv Preprint posted online December 11, 2014 2014 [FREE Full text]
60. Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What does it mean for a language model to preserve privacy? In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022 Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea URL: https://dl.acm.org/doi/fullHtml/10.1145/3531146.3534642 [doi: 10.1145/3531146.3534642]
61. Albahri A, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion 2023 Aug;96:156-191. [doi: 10.1016/j.inffus.2023.03.008]
62. Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023 Presented at: FAccT '23; June 12-15, 2023; Chicago, IL p. 1112-1113 URL: https://dl.acm.org/doi/abs/10.1145/3593013.3594067 [doi: 10.1145/3593013.3594067]
63. Artificial Intelligence Risk Management Framework (AIRMF1.0). National Institute of Standards and Technology. 2023. URL: https://doi.org/10.6028/NIST.AI.100-1 [accessed 2023-09-20]
64. Learned K, Durbin A, Currie R, Kephart ET, Beale HC, Sanders LM, et al. Barriers to accessing public cancer genomic data. Sci Data 2019 Jun 20;6(1):98 [FREE Full text] [doi: 10.1038/s41597-019-0096-4] [Medline: 31222016]
65. Park Y, Son Y, Shin H, Kim D. This ain't your dose: sensor spoofing attack on medical infusion pump. In: Proceedings of the 10th USENIX Workshop on Offensive Technologies. 2016 Presented at: WOOT '16; August 8-9, 2016; Austin, TX URL: https://www.usenix.org/system/files/conference/woot16/woot16-paper-park_0.pdf
66. Shoukry Y, Martin P, Tabuada P, Srivastava M. Non-invasive spoofing attacks for anti-lock braking systems. In: Proceedings of the 15th International Workshop on Cryptographic Hardware and Embedded Systems. 2013 Presented at: CHES '13; August 20-23, 2013; Santa Barbara, CA p. 55-72 URL: https://link.springer.com/chapter/10.1007/978-3-642-40349-1_4 [doi: 10.1007/978-3-642-40349-1_4]
67. Quiring E, Klein D, Arp D, Johns M, Rieck K. Adversarial preprocessing: understanding and preventing image-scaling attacks in machine learning. In: Proceedings of the 29th USENIX Security Symposium. 2020 Presented at: USS '20; August 12-14, 2020; Boston, MA p. 1363-1380 URL: https://www.usenix.org/conference/usenixsecurity20/presentation/quiring
68. Xiao Q, Chen Y, Shen C, Chen Y, Li K. Seeing is not believing: camouflage attacks on image scaling algorithms. In: Proceedings of the 28th USENIX Security Symposium. 2019 Presented at: USENIXS '19; August 14-16, 2019; Santa Clara, CA URL: https://www.usenix.org/conference/usenixsecurity19/presentation/xiao
69. Reichman B, Jing L, Akin O, Tian Y. Medical image tampering detection: a new dataset and baseline. In: Proceedings of the 2021 Workshops and Challenges on Pattern Recognition. 2021 Presented at: ICPR '21; January 10-15, 2021; Virtual Event p. 266-277 URL: https://link.springer.com/chapter/10.1007/978-3-030-68763-2_20 [doi: 10.1007/978-3-030-68763-2_20]
70. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023 Apr;90:104512 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
71. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 2021 Jun 15;5(6):493-497 [FREE Full text] [doi: 10.1038/s41551-021-00751-8] [Medline: 34131324]
72. Han C, Rundo L, Araki R, Nagano Y, Furukawa Y, Mauri G, et al. Combining noise-to-image and image-to-image GANs: brain MR image augmentation for tumor detection. IEEE Access 2019;7:156966-156977. [doi: 10.1109/access.2019.2947606]
73. Lu TK, Khalil AS, Collins JJ. Next-generation synthetic gene networks. Nat Biotechnol 2009 Dec 9;27(12):1139-1150 [FREE Full text] [doi: 10.1038/nbt.1591] [Medline: 20010597]
74. Synthetic data is enabling better healthcare tools - here's how. Particle Health. URL: https://www.particlehealth.com/blog/synthetic-data-healthcare-tools [accessed 2024-01-29]
75. Chen D, Yu N, Zhang Y, Fritz M. GAN-Leaks: a taxonomy of membership inference attacks against generative models. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020 Presented at: CCS '20; November 9-13, 2020; Virtual Event p. 343-362 URL: https://dl.acm.org/doi/10.1145/3372297.3417238 [doi: 10.1145/3372297.3417238]
76. Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H. Beyond inferring class representatives: user-level privacy leakage from federated learning. In: Proceedings of the 2019 IEEE Conference on Computer Communications. 2019 Presented at: IEEE INFOCOM '19; April 29-May 2, 2019; Virtual Event p. 2512-2520 URL: https://dl.acm.org/doi/abs/10.1109/infocom.2019.8737416 [doi: 10.1109/infocom.2019.8737416]
77. Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in medicine: a systematic review with large language models and beyond. Res Sq 2023 Dec 04:2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3661764/v1] [Medline: 38106170]

XSL•FO
RenderX

78. McCoy LG, Brenna CT, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol 2022 Feb;142:252-257. [doi: 10.1016/j.jclinepi.2021.11.001] [Medline: 34748907]
79. Mahmood F, Chen R, Durr NJ. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans Med Imaging 2018 Dec;37(12):2572-2581. [doi: 10.1109/tmi.2018.2842767]
80. Shafahi A, Huang W, Najibi M, Suciu O, Studer C, Dumitras TA, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018 Presented at: NIPS'18; December 3-8, 2018; Montréal, QC p. 6106-6116 URL: https://dl.acm.org/doi/10.5555/3327345.3327509
81. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023 Jun 30;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
82. Athaluri S, Manthena SV, Kesapragada VK, Yarlagadda V, Dave T, Duddumpudi RT. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023 Apr;15(4):e37432 [FREE Full text] [doi: 10.7759/cureus.37432] [Medline: 37182055]
83. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/nejmsr2214184]
84. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015 Presented at: CCS '15; October 12-16, 2015; Denver, CO p. 1322-1323. [doi: 10.1145/2810103.2813677]
85. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: Proceedings of the 38th IEEE Symposium on Security and Privacy. 2017 Presented at: SSP '17; May 22-24, 2017; San Jose, CA p. 3-18 URL: https://www.computer.org/csdl/proceedings-article/sp/2017/07958568/12OmNBUAvVc [doi: 10.1109/sp.2017.41]
86. Matwin S, Nin J, Sehatkar M, Szapiro T. A review of attribute disclosure control. In: Advanced Research in Data Privacy. Thousand Oaks, CA: Springer; 2015.
87. Fernandes M, Decouchant J, Couto FM. Security, privacy, and trust management in DNA computing. Adv Comput 2023:129. [doi: 10.1016/bs.adcom.2022.08.009]
88. Mopuri KR, Uppala PK, Babu VR. Ask, acquire, and attack: data-free UAP generation using class impressions. In: proceedings of the 15th European Conference on Computer Vision. 2018 Presented at: ECCV '18; September 8-14, 2018; Munich, Germany p. 20-35 URL: https://link.springer.com/chapter/10.1007/978-3-030-01240-3_2 [doi: 10.1007/978-3-030-01240-3_2]
89. Song Y, Kim T, Nowozin S, Ermon S, Kushman N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. arXiv Preprint posted online October 30, 2017 2017 [FREE Full text] [doi: 10.48550/arXiv.1710.10766]
90. Mattioli J, Sohier H, Delaborde A, Amokrane-Ferka K, Awadid A, Chihani Z, et al. Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation. In: Proceedings of the Safe AI 2023-The AAAI's Workshop on Artificial Intelligence Safety. 2023 Presented at: SafeAI '23; February 13-14, 2023; Washington, DC URL: https://hal.science/hal-04086455
91. Chen Y, Zahedi FM, Abbasi A, Dobolyi D. Trust calibration of automated security IT artifacts: a multi-domain study of phishing-website detection tools. Inf Manag 2021 Jan;58(1):103394. [doi: 10.1016/j.im.2020.103394]
92. Lankton N, McKnight DH, Tripp J. Technology, humanness, and trust: rethinking trust in technology. J Assoc Inf Syst 2015 Oct;16(10):880-918 [FREE Full text] [doi: 10.17705/1jais.00411]
93. Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: an investigation of its components and measures. ACM Trans Manag Inf Syst 2011 Jul;2(2):1-25. [doi: 10.1145/1985347.1985353]

Abbreviations

AI: artificial intelligence
CorGAN: Correlation-Capturing Convolutional Generative Adversarial Networks
EHR: electronic health record
EU: European Union
FL: federated learning
GAN: generative adversarial network
IoMT: Internet of Medical Things
LLM: large language model
medGAN: medical generative adversarial network
PHI: protected health information

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 22.09.23; peer-reviewed by P Williams, M Noman; comments to author 27.11.23; revised version received 12.12.23; accepted 31.01.24; published 08.03.24.

Please cite as:
Chen Y, Esmaeilzadeh P
Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges
J Med Internet Res 2024;26:e53008
URL: https://www.jmir.org/2024/1/e53008
doi:10.2196/53008
PMID:38457208

©Yan Chen, Pouyan Esmaeilzadeh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Mitigating Health-Related Uncertainties During Pregnancy: The Role of Smart Health Monitoring Technologies

Max Tretter[1], Mag Theol

Chair of Systematic Theology (Ethics), Seminar for Systematic Theology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Max Tretter, Mag Theol
Chair of Systematic Theology (Ethics)
Seminar for Systematic Theology
Friedrich-Alexander-Universität Erlangen-Nürnberg
Kochstr 6
Erlangen, 91054
Germany
Phone: 49 9131 85 26078
Email: max.tretter@fau.de

Abstract

Pregnancy is a time filled with uncertainties, which can be challenging and lead to fear or anxiety for expectant parents. Health monitoring technologies that allow monitoring of the vital signs of both the mother and fetus offer a way to address health-related uncertainties. But are smart health monitoring technologies (SHMTs) actually an effective means to reduce uncertainties during pregnancy, or do they have the opposite effect? Using conceptual reasoning and phenomenological approaches grounded in existing literature, this Viewpoint explores the effects of SHMTs on health-related uncertainties during pregnancy. The argument posits that while SHMTs can alleviate some health-related uncertainties, they may also create new ones. This is particularly the case when the abundance of vital data overwhelms pregnant persons, leads to false-positive diagnoses, or raises concerns about the accuracy and analysis of data. Consequently, it is concluded that the use of SHMTs is not a cure-all for overcoming health-related uncertainties during pregnancy. Since the use of such monitoring technologies can introduce new uncertainties, it is important to carefully consider where and for what purpose they are used, use them sparingly, and promote a pragmatic approach to uncertainties.Using conceptual reasoning and phenomenological approaches grounded in existing literature, the effects of SHMTs on health-related uncertainties during pregnancy are explored. The argument posits that while SHMTs can alleviate some health-related uncertainties, they may also create new ones. This is particularly the case when the abundance of vital data overwhelms pregnant persons, leads to false-positive diagnoses, or raises concerns about the accuracy and analysis of data. Consequently, it is concluded that the use of SHMTs is not a cure-all for overcoming health-related uncertainties during pregnancy. Since the use of such monitoring technologies can introduce new uncertainties, it is important to carefully consider where and for what purpose they are used, use them sparingly, and promote a pragmatic approach to uncertainties.

(J Med Internet Res 2024;26:e48493) doi:10.2196/48493

KEYWORDS

stress; anxiety; reproductive technologies; fetal health; epistemology; ethics

Introduction

Pregnancy is a time full of uncertainties [1], starting from uncertainties concerning the health of the fetus and pregnant person, to parental uncertainties about how life will change during pregnancy and after birth, to socioeconomic uncertainties [2,3]. These uncertainties can be very challenging and can cause fear and anxiety among those affected [4]. Smart health monitoring technologies (SHMTs) offer the possibility to continuously monitor the health of both the child and the mother, allowing parents to at least mitigate some health-related uncertainties, thereby alleviating associated stress and anxiety.

But are SHMTs actually an effective means to reduce uncertainties during pregnancy, or do they have the opposite effect? In this Viewpoint, I will explore this question at a conceptual level and argue that while SHMTs can indeed reduce some health uncertainties, their use may also introduce new ones.

The argument is developed as follows: First, I define the concept of uncertainty, drawing on relevant theories of uncertainty.

XSL•FO
RenderX

Then, using conceptual reasoning supported by existing research, I theorize how pregnancy screenings can address certain uncertainties and how SHMTs can alleviate the remaining ones. Following this, I adopt a phenomenological approach to examine how the introduction of these technologies could create new uncertainties. This analysis is underpinned by key studies identified from database searches on PubMed, Embase, Web of Science, and Google Scholar using the keywords "pregnancy" and "uncertainty." Finally, I reflect on the implications and limitations of these insights and provide practical recommendations based on pragmatic considerations.

Uncertainties in the Context of Pregnancy

Both in common sense and in the scientific discussion of this concept, uncertainty is largely understood as an epistemic state [5,6] in which individuals lack knowledge because they either lack information or the available information is inaccurate or incorrect [7]. Many people experience uncertainty as a challenge, which can lead to stress, anxiety, and fear [8]. A well-established method to navigate these uncertainties and alleviate the ensuing negative emotions involves bridging these knowledge gaps [9].

In the realm of health, uncertainty is nearly ubiquitous [10-12]. Even among those with a solid grasp of their bodily functions and biological processes, there remain gaps in knowledge and looming questions: What will my health look like tomorrow, next week, or a year from now [13]? What steps can I take to prolong my well-being? The same can be said for pregnancy [14], as pregnant women undergo significant physical and psychological changes and are constantly confronted with uncertainty [15] regarding whether the fetus is developing healthily, whether there are signs of potential health issues, whether they are behaving correctly, or whether they need to make changes to promote the healthy development of the fetus [3]. Pregnancy, an inherently stressful period, can be further complicated by these uncertainties, potentially escalating stress and anxiety levels in expectant mothers [15], which may adversely affect both the parent and child [16,17]. Therefore, it is crucial to mitigate these uncertainties where feasible, to minimize unnecessary stressors.

Pregnancy Screenings to Address Health-Related Uncertainties

Thankfully, expectant parents are not helpless in the face of these health-related uncertainties. Various prenatal tests and screenings routinely conducted throughout pregnancy can provide insight into the health of both the pregnant individual and the fetus, thereby helping to alleviate existing uncertainties.

Pregnancy screenings typically involve monitoring the vital signs of the pregnant person, including blood pressure, heart rate, and weight [18]. By tracking weight gain, health care providers can assess the health of both the pregnant person and fetus, as inadequate weight gain may be harmful to the fetus, while rapid or excessive weight gain may be a sign of health concerns [19,20], such as gestational diabetes [21,22] or pre-eclampsia [23]. A pregnant person's heart rate and blood pressure can also provide valuable information, with a low maternal heart rate potentially negatively affecting fetal growth and birth weight of the fetus [24,25], while high blood pressure or an abnormal heart rate is associated with increased risks of gestational diabetes [26] and maybe a sign of future cardiovascular disease [27,28]. Monitoring these vital signs is crucial for safeguarding the health of both the pregnant person and fetus, enabling timely interventions if necessary, and ultimately reducing health-related uncertainties during pregnancy.

Smart Health Monitoring Technologies to Address These Uncertainties

Although pregnancy screenings are vital for ensuring the health of both the pregnant person and the fetus, weeks can pass between appointments. During this time, pregnant individuals may find themselves grappling with everyday uncertainties, such as the effects of dietary changes or physical activities on the fetus, and seeking reassurance about both their own well-being and that of their unborn child.

SHMTs such as smartwatches, smart rings, and fitness trackers, especially when integrated within an Internet of Things ecosystem [29,30], can offer a practical solution [31-33]. These devices enable parents to independently measure the vital signs of the pregnant person and, in some cases, the fetus at home. This data can be analyzed to provide parents with an assessment of their fetus's health condition without needing to visit a doctor. The capability to ascertain the health status of their fetus from the comfort of home, or to receive timely alerts if the data points to potential health issues, could provide much-needed reassurance in between medical screenings and counter emerging uncertainties.

New Uncertainties Produced by the Use of Smart Health Monitoring Technologies

The widespread acceptance of SHMTs among expectant parents [34], coupled with the fact that their use is considered extremely feasible by them [35], as well as the empirical evidence suggesting that their use can have a positive impact on the health of pregnant persons [36] and can reduce their stress levels [37], should not obscure the fact that their use also brings some disadvantages. In addition to privacy and data protection issues [38,39], the use of SHMTs can also produce new health-related uncertainties.

First, new information can in itself produce new uncertainties [40], particularly when there are no medical personnel available to help interpret the additional data. In the absence of such guidance, every recorded irregularity, whether it be an irregular heartbeat, temporary high blood pressure, or a decrease in blood oxygen saturation, can cause pregnant persons to question their own health and that of their child. In extreme cases, these uncertainties can lead to unnecessary anxiety and stress symptoms [41,42], which can be detrimental to their mental and emotional well-being.

XSL•FO
RenderX

However, even in professional contexts, more data might produce new uncertainties. This happens, for example, when continuous health monitoring triggers false alarms [43], as, for instance, particular vital sign patterns are erroneously interpreted as symptoms of a specific disease or indicators of an emergency, despite being totally harmless. As Welch et al explain in their publication *Overdiagnosed* [44], the likelihood of such false-positive diagnoses inevitably increases with additional data collection, which can lead to unnecessary or even incorrect diagnoses and create additional health uncertainties.

Last, there is the uncertainty of whether health data are being collected accurately and analyzed correctly. Users of SHMTs must ask themselves if the sensors are properly attached, if smart devices are being used correctly, and if the data are being transmitted and analyzed correctly by the app's algorithms [45]. Particularly for medical and technical "laypeople," there is no way to verify this, and they must live with the uncertainty about the reliability of the data.

Ambivalent Effects of Pregnancy Monitoring Technologies

I have shown how SHMTs can mitigate the daily health uncertainties encountered during pregnancy. However, it is important to note that the use of SHMTs during pregnancy can also create new uncertainties as they make even the slightest abnormalities visible, increase the likelihood of false-positive results, and raise constant questions about the reliability of the data.

In summary, the use of SHMTs should be viewed as a double-edged sword. While they can help alleviate health uncertainties and reassure and relax pregnant persons, they also carry the risk of introducing new uncertainties, potentially exacerbating stress, anxiety, and fear.

Limitations and Opportunities for Further Research

The discussion regarding the correlation between SHMTs and uncertainties during pregnancy highlights 2 significant limitations. While these limitations require careful consideration, they do not invalidate the fundamental insights. Rather, they should be interpreted as avenues for further investigation.

First, I did not present original data to substantiate my arguments. Although I have incorporated recent literature to substantiate key assumptions, and empirical research on uncertainty in pregnancy [46] and on the impact of (health) technologies on uncertainty [47] support my findings, my contribution was strictly conceptual.

Second, the results do not explore the balance between the emergence of new uncertainties and the resolution of preexisting ones. Does the use of SHMTs produce more health uncertainties than it reduces, or vice versa? This question cannot be answered definitively, as shifts in uncertainty resulting from the use of such technologies need to be assessed on a case-by-case basis.

Practical Recommendations

In conclusion, the question arises of what practical recommendations can be derived from the insights gained regarding the use of SHMTs during pregnancy. Considering the preceding discussions, I recommend not to use SHMTs in terms of "the more the better," as their increased use might also produce more uncertainties. Instead, it seems advisable to use them sparingly and only where constant monitoring is actually necessary [10], such as when abnormalities have been identified during pregnancy screening or when there are hereditary medical conditions. In cases where constant monitoring is not necessary, these technologies should only be used upon explicit request from parents and with appropriate training on their use and data interpretation.

Moreover, even if it may sound cynical at first, a pragmatic approach toward uncertainties in health care (not just during pregnancy) can be helpful and should be encouraged [11,48-51]. This involves acknowledging that not everything can be controlled and that sometimes it can be beneficial to simply let things "take their course" [52,53].

Acknowledgments

This work has been funded by Bundesministerium für Bildung und Forschung (grant 01GP2202B) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-SFB 1483, under project ID 442419336, EmpkinS. The author conveys his thanks to Hannah Bleher for her input and discussions, as well as to Isabella Auer, Fiona Bendig, and Tabea Ott. ChatGPT was used as a tool for idea generation in the early stages of this paper and for linguistically refining the final document. The content suggestions provided by ChatGPT were verified for accuracy and relevance by the author and were developed into a coherent manuscript by the author without the help of ChatGPT.

Conflicts of Interest

None declared.

References

1. Coxon K, Scamell M, Alaszewski A. Risk, Pregnancy and Childbirth. Milton Park, NY: Routledge; 2018.

XSL•FO
RenderX

2. Sorenson DL. Uncertainty in pregnancy. NAACOGS Clin Issu Perinat Womens Health Nurs 1990;1(3):289-296. [Medline: 2206747]
3. Affonso DD, Liu-Chiang CY, Mayberry LJ. Worry: conceptual dimensions and relevance to childbearing women. Health Care Women Int 1999;20(3):227-236. [doi: 10.1080/073993399245728] [Medline: 10633621]
4. Reeves N. Common fears and anxiety during pregnancy. Naître et Grandir. 2020. URL: https://naitreetgrandir.com/fr/grossesse/sante-bien-etre/anxiety-pregnancy/ [accessed 2024-03-14]
5. Anker M. The Ethics of Uncertainty: Aporetic Openings. New York, NY: Atropos Press; Jan 15, 2009.
6. Wittgenstein L. In: Anscombe GEM, von Wright GH, editors. On Certainty. Oxford, UK: Basil Blackwell; 1969.
7. Smithson M. Ignorance and Uncertainty: Emerging Paradigms. New York, NY: Springer; 1989.
8. Bauman Z. Liquid Times: Living in an Age of Uncertainty. Malden, MA: Polity; 2007.
9. Dewey J. The Quest for Certainty: a Study of the Relation of Knowledge and Action. London, UK: George Allen & Unwin; 1929.
10. Hatch S. Snowball in a Blizzard: a Physician's Notes on uncertainty in Medicine. New York, NY: Basic Books; 2016.
11. Hatch S. Uncertainty in medicine. BMJ 2017;357:j2180. [doi: 10.1136/bmj.j2180] [Medline: 28495912]
12. Mackintosh N, Armstrong N. Understanding and managing uncertainty in health care: revisiting and advancing sociological contributions. Sociol Health Illn 2020;42(Suppl 1):1-20 [FREE Full text] [doi: 10.1111/1467-9566.13160] [Medline: 32757281]
13. Gillespie C. The experience of risk as 'measured vulnerability': health screening and lay uses of numerical risk. Sociol Health Illn 2012;34(2):194-207 [FREE Full text] [doi: 10.1111/j.1467-9566.2011.01381.x] [Medline: 21848989]
14. Matthias MS. Problematic integration in pregnancy and childbirth: contrasting approaches to uncertainty and desire in obstetric and midwifery care. Health Commun 2009;24(1):60-70. [doi: 10.1080/10410230802607008] [Medline: 19204859]
15. de Wolff MG, Rom AL, Johansen M, Broberg L, Midtgaard J, Tabor A, et al. Worries among pregnant Danish women with chronic medical conditions—a cross sectional study with data from the Copenhagen pregnancy cohort. Sex Reprod Healthc 2021;29:100623. [doi: 10.1016/j.srhc.2021.100623] [Medline: 33984666]
16. Dunkel Schetter C. Psychological science on pregnancy: stress processes, biopsychosocial models, and emerging research issues. Annu Rev Psychol 2011;62:531-558. [doi: 10.1146/annurev.psych.031809.130727] [Medline: 21126184]
17. Krishnamurti T, Davis AL, Simhan HN. Worrying yourself sick? association between pre-eclampsia onset and health-related worry in pregnancy. Pregnancy Hypertens 2019;18:55-57. [doi: 10.1016/j.preghy.2019.09.003] [Medline: 31525710]
18. Routine tests during pregnancy. American College of Obtetricians and Gynecologists. 2021. URL: https://www.acog.org/womens-health/faqs/routine-tests-during-pregnancy [accessed 2023-04-25]
19. Oken E, Kleinman KP, Belfort MB, Hammitt JK, Gillman MW. Associations of gestational weight gain with short- and longer-term maternal and child health outcomes. Am J Epidemiol 2009;170(2):173-180 [FREE Full text] [doi: 10.1093/aje/kwp101] [Medline: 19439579]
20. Champion ML, Harper LM. Gestational weight gain: update on outcomes and interventions. Curr Diab Rep 2020;20(3):11. [doi: 10.1007/s11892-020-1296-1] [Medline: 32108283]
21. LifeCycle Project-Maternal Obesity and Childhood Outcomes Study Group, Voerman E, Santos S, Inskip H, Amiano P, Barros H, et al. Association of gestational weight gain with adverse maternal and infant outcomes. JAMA 2019;321(17):1702-1715 [FREE Full text] [doi: 10.1001/jama.2019.3820] [Medline: 31063572]
22. Peng Y, Han N, Su T, Zhou S, Bao H, Ji Y, et al. Gestational weight gain and the risk of gestational diabetes mellitus: a latent class trajectory analysis using birth cohort data. Diabetes Res Clin Pract 2021;182:109130. [doi: 10.1016/j.diabres.2021.109130] [Medline: 34774643]
23. Perry A, Stephanou A, Rayman MP. Dietary factors that affect the risk of pre-eclampsia. BMJ Nutr Prev Health 2022;5(1):118-133 [FREE Full text] [doi: 10.1136/bmjnph-2021-000399] [Medline: 35814725]
24. Odendaal H, Kieser E, Nel D, Brink L, du Plessis C, Groenewald C, et al. Effects of low maternal heart rate on fetal growth and birthweight. Int J Gynaecol Obstet 2019;146(2):250-256 [FREE Full text] [doi: 10.1002/ijgo.12873] [Medline: 31131885]
25. Huang S, Liu S, Huang M, He JR, Wang C, Wang T, et al. The born in Guangzhou cohort study enables generational genetic discoveries. Nature 2024;626(7999):565-573. [doi: 10.1038/s41586-023-06988-4] [Medline: 38297123]
26. Hedderson MM, Ferrara A. High blood pressure before and during early pregnancy is associated with an increased risk of gestational diabetes mellitus. Diabetes Care 2008;31(12):2362-2367 [FREE Full text] [doi: 10.2337/dc08-1193] [Medline: 18809624]
27. Tooher J, Chiu CL, Yeung K, Lupton SJ, Thornton C, Makris A, et al. High blood pressure during pregnancy is associated with future cardiovascular disease: an observational cohort study. BMJ Open 2013;3(7):e002964 [FREE Full text] [doi: 10.1136/bmjopen-2013-002964] [Medline: 23883883]
28. Lo CCW, Lo ACQ, Leow SH, Fisher G, Corker B, Batho O, et al. Future cardiovascular disease risk for women with gestational hypertension: a systematic review and meta-analysis. J Am Heart Assoc 2020;9(13):e013991 [FREE Full text] [doi: 10.1161/JAHA.119.013991] [Medline: 32578465]
29. Venkatasubramanian S. Ambulatory monitoring of maternal and fetal using deep convolution generative adversarial network for smart health care IoT system. Int J Adv Comput Sci Appl 2022;13(1):214-222 [FREE Full text] [doi: 10.14569/ijacsa.2022.0130126]

XSL•FO
RenderX

30. Ahmed LJ, Madhumitha R, Mehana sree M, Oviya K, Sharmila M. A survey of IoT based pregnancy woman health monitoring system. 2022 Presented at: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS); March 25-26, 2022; Coimbatore, India p. 25-26. [doi: 10.1109/icaccs54159.2022.9785324]
31. Prabhu M, Malha L, Riley LE. Self-monitoring of blood pressure during pregnancy. JAMA 2022;327(17):1651-1652. [doi: 10.1001/jama.2022.2802] [Medline: 35503365]
32. Schramm K, Grassl N, Nees J, Hoffmann J, Stepan H, Bruckner T, et al. Women's attitudes toward self-monitoring of their pregnancy using noninvasive electronic devices: cross-sectional multicenter study. JMIR Mhealth Uhealth 2019;7(1):e11458 [FREE Full text] [doi: 10.2196/11458] [Medline: 30617040]
33. Runkle J, Sugg M, Boase D, Galvin SL, Coulson CC. Use of wearable sensors for pregnancy health and environmental monitoring: descriptive findings from the perspective of patients and providers. Digit Health 2019;5:2055207619828220 [FREE Full text] [doi: 10.1177/2055207619828220] [Medline: 30792878]
34. Li X, Lu Y, Shi S, Zhu X, Fu X. The impact of healthcare monitoring technologies for better pregnancy. 2021 Presented at: 2021 IEEE 4th International Conference on Electronics Technology (ICET); May 07-10, 2021; Chengdu, China p. 7-10. [doi: 10.1109/icet51757.2021.9450980]
35. Zielinska AP, Mullins E, Lees C. The feasibility of multimodality remote monitoring of maternal physiology during pregnancy. Medicine (Baltimore) 2022;101(26):e29566 [FREE Full text] [doi: 10.1097/MD.0000000000029566] [Medline: 35777056]
36. Bertini A, Gárate B, Pardo F, Pelicand J, Sobrevia L, Torres R, et al. Impact of remote monitoring technologies for assisting patients with gestational diabetes mellitus: a systematic review. Front Bioeng Biotechnol 2022;10:819697 [FREE Full text] [doi: 10.3389/fbioe.2022.819697] [Medline: 35310000]
37. Auxier J, Asgari Mehrabadi M, Rahmani AM, Axelin A. A descriptive comparative pilot study: association between use of a self-monitoring device and sleep and stress outcomes in pregnancy. Comput Inform Nurs 2023;41(6):457-466 [FREE Full text] [doi: 10.1097/CIN.0000000000000958] [Medline: 36730074]
38. Willcox JC, van der Pligt P, Ball K, Wilkinson SA, Lappas M, McCarthy EA, et al. Views of women and health professionals on mHealth lifestyle interventions in pregnancy: a qualitative investigation. JMIR Mhealth Uhealth 2015;3(4):e99 [FREE Full text] [doi: 10.2196/mhealth.4869] [Medline: 26510886]
39. Overdijkink SB, Velu AV, Rosman AN, van Beukering MD, Kok M, Steegers-Theunissen RP. The usability and effectiveness of mobile health technology-based lifestyle and medical intervention apps supporting health care during pregnancy: systematic review. JMIR Mhealth Uhealth 2018;6(4):e109 [FREE Full text] [doi: 10.2196/mhealth.8834] [Medline: 29691216]
40. Wallerstein I. The Uncertainties of Knowledge. Philadelphia, PA: Temple University Press; 2004.
41. Wakefield C, Yao L, Self S, Frasch MG. Wearable technology for health monitoring during pregnancy: an observational cross-sectional survey study. Arch Gynecol Obstet 2023;308(1):73-78 [FREE Full text] [doi: 10.1007/s00404-022-06705-y] [Medline: 35831759]
42. Marteau TM. Towards informed decisions about prenatal testing: a review. Prenat Diagn 1995;15(13):1215-1226. [doi: 10.1002/pd.1970151304] [Medline: 8710763]
43. Downey CL, Chapman S, Randell R, Brown JM, Jayne DG. The impact of continuous versus intermittent vital signs monitoring in hospitals: a systematic review and narrative synthesis. Int J Nurs Stud 2018;84:19-27 [FREE Full text] [doi: 10.1016/j.ijnurstu.2018.04.013] [Medline: 29729558]
44. Welch HG, Schwartz LM, Woloshin S. Overdiagnosed: Making People Sick in the Pursuit of Health. Boston, MA: Beacon Press; 2011.
45. Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 2019;6(1):44 [FREE Full text] [doi: 10.1186/s40537-019-0206-3]
46. Matthias MS, Babrow AS. Problematic integration of uncertainty and desire in pregnancy. Qual Health Res 2007;17(6):786-798. [doi: 10.1177/1049732307303241] [Medline: 17582021]
47. Zvonareva O, Popova E, Horstman K. Health, Technologies, and Politics in Post-Soviet Settings: Navigating Uncertainties. London, UK: Palgrave Macmillan; 2018.
48. Christoffersen MG. Living with Risk and Danger: Studies in Interdisciplinary Systematic Theology. Göttingen, Germany: Vandenhoeck and Ruprecht; 2019.
49. Rogers WA, Walker MJ. Fragility, uncertainty, and healthcare. Theor Med Bioeth 2016;37(1):71-83. [doi: 10.1007/s11017-016-9350-3] [Medline: 26906556]
50. Simpkin AL, Schwartzstein RM. Tolerating uncertainty—the next medical revolution? N Engl J Med 2016;375(18):1713-1715. [doi: 10.1056/NEJMp1606402] [Medline: 27806221]
51. Tretter M, Ott T, Dabrock P. AI-produced certainties in health care: current and future challenges. AI Ethics 2023 Nov 21. [doi: 10.1007/s43681-023-00374-6]
52. Seely AJ. Embracing the certainty of uncertainty: implications for health care and research. Perspect Biol Med 2013;56(1):65-77. [doi: 10.1353/pbm.2013.0009] [Medline: 23748527]
53. Han PKJ. Uncertainty in Medicine: a Framework for Tolerance. New York, NY: Oxford University Press; 2021.

XSL•FO
RenderX

Abbreviations

SHMTs: smart health monitoring technologies

Edited by A Mavragani; submitted 25.04.23; peer-reviewed by MDG Pimentel, E Baker; comments to author 11.01.24; revised version received 26.01.24; accepted 23.02.24; published 25.03.24.

Please cite as:
Tretter M
Mitigating Health-Related Uncertainties During Pregnancy: The Role of Smart Health Monitoring Technologies
J Med Internet Res 2024;26:e48493
URL: https://www.jmir.org/2024/1/e48493
doi:10.2196/48493
PMID:38526554

©Max Tretter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Political Economy of Digital Health Equity: Structural Analysis

James Shaw[1,2], PhD; Wiljeana Glover[3], PhD

[1]Department of Physical Therapy, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[2]Joint Centre for Bioethics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[3]Technology, Operations, and Information Management Division, Babson College, Wellesley, MA, United States

Corresponding Author:
James Shaw, PhD
Department of Physical Therapy
Temerty Faculty of Medicine
University of Toronto
500 University Avenue
Toronto, ON, M5G1V7
Canada
Phone: 1 4169780315
Email: jay.shaw@wchospital.ca

Abstract

Digital technologies have produced many innovations in care delivery and enabled continuity of care for many people when in-person care was impossible. However, a growing body of research suggests that digital health can also exacerbate health inequities for those excluded from its benefits for reasons of cost, digital literacy, and structural discrimination related to characteristics such as age, race, ethnicity, and socioeconomic status. In this paper, we draw on a political economy perspective to examine structural barriers to progress in advancing digital health equity at the policy level. Considering the incentive structures and investments of powerful actors in the field, we outline how characteristics of neoliberal capitalism in Western contexts produce and sustain digital health inequities by describing 6 structural challenges to the effort to promote health equity through digital health, as follows: (1) the revenue-first incentives of technology corporations, (2) the influence of venture capital, (3) inequitable access to the internet and digital devices, (4) underinvestment in digital health literacy, (5) uncertainty about future reimbursement of digital health, and (6) justified mistrust of digital health. Building on these important challenges, we propose future immediate and long-term directions for work to support meaningful change for digital health equity.

(J Med Internet Res 2024;26:e46971) doi:10.2196/46971

KEYWORDS

digital health equity; health equity; health policy; telemedicine; digital care; political economy; race; ethnicity; socioeconomic; policy

Introduction

Digital health has become an integral component of health care delivery [1-3]. Although the use of digital technologies to deliver care has been significantly accelerated by the COVID-19 pandemic [3], the trend toward digitally enabled health care has been progressing for several years alongside the growing power of the technology industry and the widespread distribution of digital devices and infrastructure [4-6]. Digital technologies have produced many innovations in care delivery and enabled continuity of care for many populations when in-person care was challenging or impossible. However, a growing body of research suggests that digital health can also exacerbate health inequities for those excluded from its benefits for reasons of cost, digital literacy, and structural discrimination based on age, race, ethnicity, or socioeconomic status [7,8].

Often referred to as the "health-related digital divide," the exacerbation of health inequities by digital health technologies has been widely discussed in health informatics and digital health literature [7,9,10]. Now that the digital divide has widened as a result of the COVID-19 pandemic, strategies to address its impacts on health have been more widely discussed at policy, health system leadership, clinical, and community levels [8,11]. Building on past research bringing critical perspectives to bear on health equity and digital innovation [12], we argue that these recommended practices and policies have not yet made meaningful progress on digital health equity in part because scholars, providers, policy makers, and advocates have not yet adequately confronted the social and structural

XSL•FO
RenderX

influences that reinforce the status quo at the digital health and health care industry nexus.

In our paper, we draw on a political economy perspective that brings attention to the incentives and actions of powerful stakeholders in a given field [13,14], in this case primarily government, professional, and corporate actors who hold the power to shape the digital health landscape. Considering the incentive structures and investments of such powerful actors, we outline how characteristics of neoliberal capitalism in Western contexts produce and sustain digital health inequities by describing 6 structural challenges to the effort to promote health equity through digital health. Acknowledging that the concept of "digital health equity" has been defined in different ways [15,16], we focus here on digital health equity as an aspirational state in which all communities have access to digital health technologies that enable them to meet their health-related needs. Such a state can be defined at different scales, for example, within a patient population, a city, a province or state, an entire country, or worldwide. Achieving the goal of digital health equity requires commitments from diverse actors in various settings, which we explore in detail in our paper.

Instead of reproducing frameworks proposing multiple points of action to redress digital health inequities [10,15,17], in the concluding sections of our paper, we highlight what we understand to be a central tension inherent in efforts to enhance digital health equity. Specifically, we discuss the challenge of supporting commercial actors to invest in immediate actions that enhance digital health equity while acknowledging that the broader health care systems and business models through which they operate can be obstacles to the meaningful advancement of equity and justice for people's health. We conclude by outlining the need for immediate and long-term directions for research, policy, and advocacy for digital health equity, one focused on foundational reform and the other on stronger collaboration to deepen the commitments of commercial organizations to equity in digital health.

Defining Digital Health

In its broadest sense, digital health refers to a diverse collection of technologies and approaches to care delivery that uses digital technologies or the data they generate to inform health care, self-management, and public health [4,18,19]. Digital health can be understood to include at least 4 distinct groups of applications of digital technology to achieve health-related aims, as outlined in Textbox 1 [20,21]. This broad sense of digital health introduces the wide variety of use cases that are considered when policy makers contemplate digital health strategies as solutions to problems in health care and public health.

Textbox 1. Definition of digital health.

Digital health refers to a diverse collection of technologies and their uses to achieve health-related goals. We identify the following 4 groups of digital health use cases [20]:

1. The collection of data about health and health care into electronic patient records to inform service planning and care delivery
2. Digital apps (usually used on smartphones) that support people to engage with their health and self-monitoring in an ongoing way
3. Telemedicine and internet-based care strategies that enable health care to be delivered at a distance
4. Analytics of health-related data from multiple sources, including Internet of Things devices, to generate new health-related insights and applications.

The hopes for what digital health can accomplish are high. One highly cited commentary on evaluation standards for digital health stated that "digital health interventions have enormous potential as scalable tools to improve health and health care delivery by improving effectiveness, efficiency, accessibility, safety, and personalization" [5]. The variety of goals embedded in this statement conveys a general sense that digital health strategies can help to achieve many of the most elusive goals of health care and public health. For example, the notion that digital health apps can lead patients to be more deeply engaged in their own health and well-being, becoming more active participants in their own wellness and care [22,23], or that electronic patient records can enhance collaboration between providers to produce more integrated health care experiences for patients [24]. If successfully implemented, these specific applications of digital health technologies would contribute to improving health care in profound ways.

Digital Health and Health Inequities

Despite these high ambitions, the positive impacts of digital health have not been evenly distributed. Over the past several years, researchers have documented the exclusion of particular communities from the benefits of digital health technologies in both high-income and low-income country settings [8,25-27]. Systematic reviews of this literature have consistently illustrated that communities excluded from the benefits of digital health are those communities facing forms of structural oppression, including economic exclusion, racism, housing injustice, ableism, ageism, and others [9,28-30]. For example, multiple systematic reviews have reported the exclusions of Indigenous peoples in North America and Australia from engaging with digital health technologies as a result of a lack of cultural safety in applications of digital health [17,31]. Communities speaking a first language other than English have also been excluded from the benefits of digital health, for example, in situations that demand people navigate a complex digital interface for mental health support [32]. Communities with lower levels of digital literacy or digital readiness as a result of a lack of educational opportunities also face added barriers to engaging in digital health; for example, older people with less exposure to digital technology and those living with lower incomes who lack the financial resources to afford such technologies in the first place [8,17,25]. The reliance on digital health during the COVID-19 pandemic has only exacerbated these realities,

XSL•FO
RenderX

leading to deeper inequities in access to health care that stand to have important longer-term consequences [33-36]. For example, the accelerated adoption of internet-based care and other digital health technologies has led to a range of inequities in access to care [36], and although not enough time has passed to know the long-term impacts of these technologies, they are very likely to be a permanent element of care delivery.

A number of authors have proposed frameworks for understanding where, in the design, adoption, and use of digital health technologies, these exclusions occur in order to identify likely points of redress [9,17,26]. Such works have emphasized the importance of a multilevel approach to addressing digital health inequities, generally proposing interventions at the levels of policy, health system leadership, clinical practice, and patient or community engagement [9,11,16]. These strategies, some of which are summarized in Table 1, constitute identifiable actions for specific stakeholders to engage in the effort to enhance health equity in the field of digital health. However, beyond a few notable examples [37], there is little indication that meaningful progress has been made in adopting these strategies across digital health stakeholders. In this paper, we explore why that is the case by pointing out structural obstacles to advancing digital health equity in the broader digital health ecosystem.

Table 1. Sample strategies to promote digital health equity (adapted from Shaw et al [9]).

Stakeholder group responsible and domain of intervention	Example strategies
Policy and government	
Government policy	Government policy should clarify standards for the inclusive design of digital health innovations.
Funder (reimbursement)	Ensure payment parity between insurers for video and audio visits.
Access to devices	Promote access to broadband internet, especially among those who cannot afford it.
Health-related messaging	Increase emphasis on and diffusion of culturally relevant public health messages (eg, increase redundancy of important messaging).
Organization and health system	
Health care organizations	Measurement of equity-related outcomes, such as the number of visits using interpreter services.
Clinical practice	Provide training and support to patients seeking to access care on the internet.
Community and patient	
Community engagement	Partner with community organizations to provide peer-led educational support.
Enhancing digital literacy	Mitigate digital literacy and resource barriers (eg, provide patient education to enhance digital literacy skills and inform patients about free or reduced-cost internet access locations).

Political Economic Analysis: A Structural Approach

We approach our analysis from a political economy perspective on digital health inequities. The concept of political economy has been used in many ways in relation to health care and public health [38-40]. For our purposes, we describe political economy as a set of relationships and incentives that characterize the interconnections between state, corporate, and other actors with respect to a given domain of social activity, in our case, digital health. Central to analyses of political economies is the concept of power, understood in this context as the capacity of certain actors to exert influence over states of affairs in order to produce outcomes favorable to their interests [27,39]. Power is a complex and productive phenomenon embedded in systems of policy institutions, economic activity, and globalization that make such influence possible [41].

Analyses of political economy build on a clear understanding of the links between systems of policy or public administration and the economic regimes in place in a given jurisdiction. In North American and European contexts, the impacts of neoliberal capitalism on population health have garnered substantial attention [31,42,43]. In this sense, a core feature of neoliberal capitalism is the favoring of private enterprise and the reduction in capacity and responsibility of government or public services [42]. For example, Friel et al [41] showed how neoliberal policy approaches to health-related reforms in Australia led to the favoring of private enterprise and individualized strategies as opposed to structural- or population-level initiatives intended to improve health.

With respect to digital health, the policy framework and incentive structures of neoliberalism incentivize the development of products and technologies with a maximum capacity for scale because maximal scale generates maximal profits [42,43]. While neoliberalism has apparently led to a population-level decrease in poverty through economic growth, it has also increased financial and power inequities that have important impacts on health [44]. Given their capacity to explain whether and how large-scale change occurs in political and economic regimes, political economy approaches have been particularly useful in understanding why more profound action is not taken to address

XSL•FO
RenderX

the social determinants of health by states, health care funders, and health systems [41,45].

In these ways, a political economy approach to digital health in the context of neoliberalism examines the influence held by powerful actors over investment and innovation in digital health and the capacity of the state or other actors to steer these processes in the public's interest. Reflecting on the political economy of digital health, Storeng et al [27] exclaimed that "new actors, financial instruments, (absent) legal frameworks, and the power structures of capitalism drive the innovation, development, and applications of digital technologies, reshaping health systems, health care, and medicine." We make observations in our analysis informed by these remarks about political economy, describing the impact of active investments by some actors and inaction by others on the possibilities of achieving digital health equity at scale.

In our analysis, we sought to take an innovation ecosystem perspective [46], beginning with a broader set of influences on digital health inequities and concluding with a narrower focus on health care delivery and patients. Specifically, we aimed to document the impacts of powerful political and economic actors that shape the broader ecosystem in which digital health innovation takes place. These actors include large corporations, venture capitalists, digital health innovators, and policy makers who set the terms of engagement in digital health. Although policy and investments in digital health are a central consideration, we also acknowledged that the influence of broader policy related to digital inclusion is essential to consider (eg, decisions regarding where high-speed internet is made available and how education regarding digital technologies will be provided). We then narrowed our perspective to focus on health care delivery systems and the patients with whom they engage, documenting decision-making and circumstances that impact digital health inequities at the interface between provider organizations and the communities they serve. Our observations were directly informed by the political economy perspective outlined here.

Structural Barriers to Digital Health Equity

The Role of Large Technology Corporations

The first challenge we outline relates to the role of large technology corporations in the digital health innovation ecosystem. Large technology companies have come to enjoy immense influence over our everyday lives, and this influence extends deeply into health care [46]. The COVID-19 pandemic witnessed large technology corporations taking advantage of the need for rapid responses to issues such as contact tracing, with the Apple-Google contact tracing application programming interface becoming central to several countries' contact tracing efforts [14,47]. However, investments from technology corporations extend well beyond contact tracing. Meta (Facebook), Apple, Microsoft, Google, and Amazon invested a combined US $6.8 billion in health care between January 2020 and June 2021 [48]. Investments ranged from Microsoft's purchase of a health care chatbot company to Facebook's development of a preventative health tool to promote regular health screening [48]. With the immense resources held by large technology corporations and their persistent investment in acquiring market share for digital technologies in health care, they stand to gain substantial influence over the structure and delivery of health care [49].

The influence of large technology companies may limit the progress that can be made in advancing digital health equity. The incentives of large corporations are fundamentally driven toward the scaling of technology offerings and the profit they generate [6]. Innovations developed in large corporate environments are those that can be marketed to a segment of the population (or some other payer) that is capable of paying market prices [8,14]. The disposable income or connection to other forms of financial capital (eg, generous health insurance plans) required to access these innovations as they arrive on the market excludes communities that are economically marginalized. This means that the incentives characterizing the activity of technology corporations are fundamentally different from those for achieving equitable access to and outcomes of digital health. Even applications that are free to end users, for example, contact-tracing apps, raise issues of privacy, security, access, and equity for cultural minority and other structurally marginalized patient populations because app rollout may not target these populations [50]. By supporting and incentivizing technologies that work well for the economically privileged and creating challenges for other actors aiming to provide free or low-cost access to technologies, the actions and incentives of large technology corporations risk widening the divide between those who benefit from access to digital health and those who do not.

Innovation Policy and the Venture Capital System

Beyond the power and impact of large technology corporations, smaller-scale digital health startups are constantly appearing in the digital health ecosystem [51]. These companies are established and developed through a variety of routes and seek out financial support to fund their activities progressing toward implementation and scale [52]. In some cases, health systems actively invest in digital health startups as a strategy to promote innovation that more directly addresses clinical and operational needs [53]. In these cases, an explicit focus on digital innovations that promote health equity is far more likely [16]. However, external venture capital investment in digital health startups is responsible for the largest portion of venture capital funding. Despite recent declines in new investments due to a market downturn, digital health startups worldwide attracted a combined US $30.7 billion of venture capital investments in 2021, with telemedicine, data analytics, and health-related mobile apps being the highest investment categories [54].

Importantly, venture capital investors have a deep influence over the development and direction of technologies and business models in health-focused start-ups [55,56]. In a series of papers reporting research on the role of venture capital in health care technology companies, Lehoux and colleagues [52,55,56] illustrated the ways that Canadian venture capital investors actively shape the activities of startups and influence the value proposition of the technology being developed. Investors retain a primary incentive to maximize future potential earnings and contribute to shaping activities and priorities in the earlier phases

XSL•FO
RenderX

of a company's development that situate the company to generate the highest potential profits. As was described with respect to large technology corporations, this incentive may close down possibilities for technology that is designed with health equity in mind. Parthasarathy [57] provided a detailed description of the ways in which a focus on economic gain detracts from the possibility of investing in technologies that meaningfully address the needs of structurally marginalized patients and health care systems. When considered at the scale of investments occurring in digital health around the world, this observation raises serious concerns regarding the likelihood of digital health making meaningful contributions to enhancing equity in access to and outcomes of health care.

Inequitable Access to High-Speed Internet and Digital Technologies

Where communities lack access to high-speed internet and digital devices, they will be unable to benefit from digital health in many ways [58]. Greene [59] explained that the reasons for a lack of internet access among marginalized populations relate directly to political and economic investments made during the period during which the internet was being established. The internet, relying on a collection of infrastructures, including physical cables, wireless signals, support personnel, and networks of information exchange, was established in 1983, during a time in the United States that led to a commercialized, free market approach to internet access [59]. Although there has been progress in expanding high-speed internet to rural areas and increasing free and affordable access in urban areas [60], a lack of access to high-speed internet and digital devices remains a major barrier to health equity in digital health [58]. As the centrality of high-speed internet access to health and social services becomes more widely understood, scholars have begun to call for high-speed internet access to be formally recognized as a social determinant of health [58].

Digital device availability also contributes to the digital divide. Roughly a quarter of American adults with household incomes below US $30,000 a year (24%) say they do not own a smartphone [61]. If digital health is going to remain a central feature of health care, then establishing sustainable approaches to making devices available to those without access is going to be as important as high-speed internet access [9,17]. A number of charitable programs arose during the COVID-19 pandemic to make digital devices available to those without access [62], but these have largely been temporary responses to the crisis. The question then becomes whether health care funders will fund digital devices for people without access in order to support their engagement with digital health. These 2 related issues of access to high-speed internet and digital devices raise important questions about perceptions of rightful investments by health care funders, health systems, and governments. Our contention is that without investments in these core infrastructures, digital health will exacerbate existing inequities in access to and outcomes of health care over the long term.

Lack of Investment in Digital Health Literacy

The availability of free or affordable technologies and the internet necessary to use them for health is a crucial starting point for promoting digital health equity. However, the skills and knowledge to use the technology in meaningful ways that promote the attainment of health-related goals are equally important. Digital health literacy has been discussed in different ways, and in this paper, we align with Azzopardi-Muscat and Sørensen [63], who define the concept broadly as "the ability of people to use emerging information and communications technologies to improve or enable health and health care." In this way, digital health literacy is about possessing the skills and knowledge necessary to navigate digital technologies, digest health-related information, and engage with health care systems in ways that contribute to enhancing one's health. Even though supporting access to technology for economically marginalized communities has been used in government strategy internationally [59], a lack of coherent investment in digital health literacy aligns with the neoliberal erosion of public services, public education, and coordinated approaches to government [64].

Although governments have progressively built policies on investments in technology as a strategy to promote upward economic mobility [59], digital health literacy initiatives have appeared to arise only sporadically. Government-sponsored initiatives exist internationally, such as the All of Us Research Program in the United States [65] and Health Education England's collaboration with Libraries Connected [66]. However, the prospect of open-access, well-resourced, and scalable digital health literacy training programs remains elusive internationally. An effort to promote digital health equity at any meaningful scale would require coordinated multistakeholder investments in the delivery of digital health literacy initiatives across communities, health care organizations, and levels of government, and would necessarily include education across proprietary technology offerings.

Uncertainty About the Future of Equitable Reimbursement for Digital Health Care

In the context of the rapid virtualization of care experienced during the COVID-19 pandemic, the question of the sustainability of multiple digital modalities for ambulatory care delivery was much discussed [67]. In the provincial health care system in Ontario, Canada, physician billing for digital care visits was initially permitted on a temporary basis; that permission was extended multiple times [68]. Importantly, these permissions included the capability of billing for telephone-only digital visits, which has been widely regarded as an equity-promoting policy to ensure those without access to internet-connected digital technologies can retain access to care [67]. However, the government of Ontario announced in March 2022 that telephone-only physician visits would be reimbursed at a substantially lower rate than video visits by provincial health insurance, instituting a major blow to sustained efforts to enable equity of access to digital health services. In other words, where audio-only provision of direct care serves to enhance access for those without the digital devices and high-speed internet access required for video visits, reimbursing audio-only visits at a lower rate will disincentivize providers to continue with audio-only digital care. This policy issue represents a substantial challenge for policy makers, seeking to maintain a high level of quality care while at the same time promoting broad accessibility. Understanding how to address this challenge while

XSL•FO
RenderX

maintaining access to telephone-based care options that promote equitable access to services is a central policy issue for the next 10 years [67].

Research in Ontario has shown that the telephone was used appropriately for ambulatory care during the pandemic, and the rise in the use of the telephone was not associated with aberrant or unexpected billing practices among physicians (ie, did not lead to substantial fraudulent billing) [69]. The removal of reimbursement for telephone visits despite this evidence is a symptom of the effort to narrow the scope of government spending, aligning directly with a neoliberal paradigm of public services. However, the question of funding for digital health and digital care is a broader one than reimbursements for telephone visits.

Discussions about funding for telehealth and digital care following the pandemic have emphasized the importance of paying for value, shifting away from volume-based payment toward value-based payment for primary health care [70]. The policy work associated with shifting toward value-based payment as a paradigm for health care systems is immense; as that work proceeds and telephone visits are defunded in the near term, gaps in access to digital care will widen. This reality is compounded by the lack of affordability of internet access and digital devices and low investments in digital health literacy, as discussed earlier. As a future state, a value-based care approach would ostensibly involve health systems or health care providers being paid block payments to manage a patient roster in the way they see fit, using funds to support various modalities of digital health care when such modalities could contribute to enhanced patient outcomes for all [71]. We acknowledge this is an important domain of ongoing research and contend that digital health equity must be a core consideration as value-based care develops in research and policy.

Justified Mistrust and Exclusionary Design

Beyond the issues of the availability and affordability of digital health technologies is the issue of the documented harms some technologies have caused. Examples linked to health care and health promotion are common, including racism in an automated health care resource allocation algorithm [72] and ineffective sensors on heart rate monitors for people with darker skin [73]. These harms are representations of social systems arising from histories of excluding marginalized communities from commercial activity and quality health care, leading to the development of technologies and services that neglect the needs of such communities or cause them outright harm. Under these circumstances, some communities choose not to engage with digital health technologies as a result of a justified mistrust of their safety and impact [16].

Strategies to address justified mistrust of digital health technologies are well-established and include community engagement and inclusive co-design with affected communities to ensure the cultural sensitivity, accuracy, and use of digital health offerings [16]. However, in the current political economic regime, there is a disincentive to adopt inclusive design strategies across types of commercial actors in digital health (ie, from start-ups to large technology companies). These disincentives arise because inclusive design approaches are costly in that they require additional expertise and time and are challenging to accomplish in meaningful ways [74,75]. As such, barriers to inclusive design remain high. Shifting these incentives and lowering barriers to engagement with inclusive design is a primary challenge in the effort to achieve digital health equity.

Discussion

Overview

The challenges to digital health equity arising from a neoliberal political economic regime are multiple and substantial. However, we also find examples across the 6 challenges where efforts are being made within the neoliberal context to prioritize digital health equity by commercial entities, venture capitalists, providers of high-speed internet, digital and health literacy educators, insurers, and inclusive designers. We acknowledge that such examples may be subject to criticism, especially from scholars adopting critical approaches to the digital technology industry. Research in technology ethics, critical data studies, critical digital health studies, and other related fields has emphasized the ways that commercial entities (and especially large technology corporations) use bounded special projects focused on health equity to avoid deeper changes to their business models or the broader health and economic environments [13,18,23,76]. In these instances of “ethics washing” [77,78], or perhaps more accurately, “equity washing,” business models are critiqued for their drive for profits without clear prioritization of health equity, health outcomes, and the quality of health care systems [79].

As scholars anchored by the aim of informing innovations that enhance equity and social justice for people’s health, we are closely attuned to these critiques. However, we have also witnessed the effectiveness of multiple points of intervention in the development of digital health technologies that can orient digital health toward more equitable organization and delivery of care. Furthermore, we are encouraged by the growth in awareness since the onset of the COVID-19 pandemic of the urgent need to promote equitable access to and outcomes of digital health. We are aware of mission-driven digital health initiatives that demonstrate a strong understanding of issues of equitable access, informed consent, and ethical business models. Across our 6 areas, we observe promising opportunities for future work regarding immediate investments in changes to the business models that characterize digital technology under neoliberal capitalism as well as broader changes to health system policies that promote universal access to care.

Large technology corporations like Google are becoming more intentional about their influence on digital health equity. The Google Health Equity team is expanding access to health information through initiatives within Google Search and YouTube and is incorporating insights about the social determinants of health within their Fitbit Health Solutions. Google’s Health Equity Tracker, developed by their health equity team in partnership with the Satcher Health Leadership Institute at the Morehouse School of Medicine [80], 1 of 3 major historically Black medical schools in the United States, provides

XSL•FO
RenderX

a real-time dashboard of progress on various health equity indicators. The platform was designed in collaboration with affected communities and is governed by a diverse team of researchers, community members, providers, and technology leaders. The platform presents the opportunity for governments and payers to make strategic investments in enhancing the health of structurally marginalized communities, offering a pathway for digital health technologies to explicitly enhance health equity. Future research could consider how the revenue models that support these collaborations can more equitably yield financial and health benefits for all involved, including the collaborators and individuals sharing data for such initiatives.

Venture capital has explored the potential of social impact investing as a strategy to enhance health equity at local and global levels [81,82]. Social impact investing involves the development of investments in commercial entities that promise to achieve both a financial return on investment and meaningful progress in enhancing social and economic equity [81]. Such investments have included large initiatives in global health [82]. Health systems have also invested in the development of digital health technologies that more comprehensively meet the needs of diverse patients and communities served, such as those developed by accountable care organizations to enhance access to care and reduce hospital admissions [83]. Future research at the nexus of impact investing and digital health should examine the process of venture selection, social, health, and commercial outcomes, and the role of government in directly investing in or developing policies to support equitable digital health venture development [84].

Government, large corporations, and digital health ventures are contributing to digital infrastructure initiatives to improve inequitable access to high-speed internet and digital technologies. For example, the US Affordable Connectivity Program of 2022 recently gave over 48 million households (40% of the country) access to high-speed internet for US $30 per month [85], with a similar government-subsidized program announced in Canada [86]. Certain health plans in the United States allow for free smartphones for Medicaid patients [87], with similar programs occurring only through charitable organizations in Canada [62]. Securing long-term investments and publicizing these offerings to marginalized communities represent crucial investments in digital health equity. Future research should examine factors that promote and inhibit the system-wide scaling of necessary digital infrastructure.

The need for deeper investment in digital health literacy is closely related to the perennial struggle of governments and professional groups to organize systems of health care in ways that support access to care and health promotion for structurally marginalized communities [41]. In our case, this means fractured investments in digital health literacy, applications, and infrastructure that reinforce the existing fragmentation in health systems, with harmful effects magnified for those who are most marginalized [15]. Where initiatives aimed at supporting equitable digital health literacy do exist, they are often temporary, local, and not coordinated across the parties needed to ensure their effective implementation in practice. Digital health literacy may be bolstered not only through health agencies but also through broader educational initiatives. For example, the digital literacy efforts by the US Department of Education's Literacy Information and Communication System and the Canadian Digital Literacy Exchange program could be expanded to explicitly include digital health literacy. Future research is needed on approaches to the governance of digital health that enable innovation in digital health literacy, with an explicit focus on the needs of structurally marginalized communities.

Questions about inequitable reimbursement for digital health care point to the need for broader changes to health system policies that promote universal access to care. These changes are fundamentally longer-term and deeply intertwined with the distribution of power in the health care and digital technology industries. Given the concentrations of power in these domains, strategies to achieve such high-level change are not immediately clear. Building on the strong critiques provided by critical studies of both the health care and digital technology industries and leveraging insights from diverse fields of work such as institutional change, social movements, and sustainability transitions, urgent work is needed on practical strategies to mobilize change at the industry and policy levels. We suggest that research on these changes in the sphere of digital health has much to learn from interdisciplinary social sciences on the green energy transition, exploring multiple points of intervention to modify business models and social habits that support more sustainable ways of life.

Finally, we see a need for support for the implementation and spread of an inclusive design to address mistrust, and with high urgency. Counteracting our aforementioned disincentives for inclusive design, the US Food and Drug Administration is developing plans to enroll more participants from underrepresented racial and ethnic populations into clinical trials [88]. Design approaches like value-sensitive design are gaining attention from health researchers, such that tensions between stakeholder values can be identified and addressed earlier in design processes [89]. Remaining aware of current critiques and attuned to the need for broader institutional change, this immediate practical work remains focused on collaboration with digital health developers, governments, health systems, and commercial entities to support equity-focused design and culture changes. Future research is needed that acknowledges critiques of co-design in health care while building strong approaches to inclusive design and values-based collaboration with commercial providers of digital health technologies.

Conclusions

Digital health has become a central feature of health system transformation, presenting an opportunity to steer this transformation in directions that are equity-promoting as a fundamental principle. The role of growth and profit motives and fragmented policy contexts complicate efforts to build coordinated approaches to digital health equity. Nevertheless, the community of researchers, clinicians, community members, and health care leaders committed to health equity holds the power to influence the course of digital health.

As outlined in our paper, there is a substantial body of literature that has drawn on community-engaged and other research approaches to generate strategies to combat digital health inequities at multiple levels. The intellectual work to identify

XSL•FO
RenderX

these strategies has already been done and continues to grow. Acknowledging where power wields influence in the broader landscape of digital health, we have proposed future lines of immediate and long-term work aiming to actively promote the advancement of digital health equity. Our recommendations are based on both immediate collaborative work with digital health developers and longer-term efforts to achieve institutional change in the health care and technology industries. Acknowledging the challenges outlined here and investing in the strategies proposed represent important steps toward achieving the goal of digital health equity.

Acknowledgments

Funding for JS was provided by the Canadian Institutes of Health Research Canada Research Chairs program (CRC-2021-00312).

Authors' Contributions

JS and WG contributed to the conceptualization, writing, critical review, and approval of the manuscript.

Conflicts of Interest

None declared.

References

1. Bhatia RS, Chu C, Pang A, Tadrous M, Stamenova V, Cram P. Virtual care use before and during the COVID-19 pandemic: a repeated cross-sectional study. CMAJ Open 2021;9(1):E107-E114 [FREE Full text] [doi: 10.9778/cmajo.20200311] [Medline: 33597307]
2. Zachrison KS, Yan Z, Schwamm LH. Changes in virtual and in-person health care utilization in a large health system during the COVID-19 pandemic. JAMA Netw Open 2021;4(10):e2129973 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.29973] [Medline: 34705016]
3. De Vera K, Challa P, Liu RH, Fuller K, Feroz AS, Gamble A, et al. Virtual primary care implementation during COVID-19 in high-income countries: a scoping review. Telemed J E Health 2022;28(7):920-931. [doi: 10.1089/tmj.2021.0377] [Medline: 34847758]
4. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
5. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
6. Sharon T. When digital health meets digital capitalism, how many common goods are at stake? Big Data Soc 2018;5(2):2053951718819032 [FREE Full text] [doi: 10.1177/2053951718819032]
7. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
8. Weiss D, Rydland HT, Øversveen E, Jensen MR, Solhaug S, Krokstad S. Innovative technologies and social inequalities in health: a scoping review of the literature. PLoS One 2018;13(4):e0195447 [FREE Full text] [doi: 10.1371/journal.pone.0195447] [Medline: 29614114]
9. Shaw J, Brewer LC, Veinot T. Recommendations for health equity and virtual care arising from the COVID-19 pandemic: narrative review. JMIR Form Res 2021;5(4):e23233 [FREE Full text] [doi: 10.2196/23233] [Medline: 33739931]
10. Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing equity in telemedicine for chronic disease management during the COVID-19 pandemic. NEJM Catal Innov Care Deliv 2020;1(3) [FREE Full text]
11. Veinot TC, Ancker JS, Cole-Lewis H, Mynatt ED, Parker AG, Siek KA, et al. Leveling up: on the potential of upstream health informatics interventions to enhance health equity. Med Care 2019;57(Suppl 6 Suppl 2):S108-S114 [FREE Full text] [doi: 10.1097/MLR.0000000000001032] [Medline: 31095048]
12. Cruz TM. Perils of data-driven equity: safety-net care and big data's elusive grasp on health inequality. Big Data Soc 2020;7(1):2053951720928097 [FREE Full text] [doi: 10.1177/2053951720928097]
13. Lupton D. Critical perspectives on digital health technologies. Sociol Compass 2014;8(12):1344-1359 [FREE Full text] [doi: 10.1111/soc4.12226]
14. Wamsley D, Chin-Yee B. COVID-19, digital health technology and the politics of the unprecedented. Big Data Soc 2021;8(1):20539517211019441 [FREE Full text]
15. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
16. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]

XSL•FO
RenderX

17. Budhwani S, Fujioka J, Thomas-Jacques T, De Vera K, Challa P, De Silva R, et al. Challenges and strategies for promoting health equity in virtual care: findings and policy directions from a scoping review of reviews. J Am Med Inform Assoc 2022;29(5):990-999 [FREE Full text] [doi: 10.1093/jamia/ocac022] [Medline: 35187571]
18. Shaw JA, Donia J. The sociotechnical ethics of digital health: a critique and extension of approaches from bioethics. Front Digit Health 2021;3:725088 [FREE Full text] [doi: 10.3389/fdgth.2021.725088] [Medline: 34713196]
19. Kruse CS, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
20. Marent B, Henwood F. Digital health. In: Routledge International Handbook of Critical Issues in Health and Illness. London, UK: Routledge; 2021.
21. Elenko E, Underwood L, Zohar D. Defining digital medicine. Nat Biotechnol 2015;33(5):456-461 [FREE Full text] [doi: 10.1038/nbt.3222] [Medline: 25965750]
22. Klein S, Hostetter M, McCarthy D. A vision for using digital health technologies to empower consumers and transform the US health care system. Candid. 2014. URL: https://www.issuelab.org/resources/25081/25081.pdf [accessed 2024-03-08]
23. Lupton D. The digitally engaged patient: self-monitoring and self-care in the digital health era. Soc Theory Health 2013;11(3):256-270 [FREE Full text] [doi: 10.1057/sth.2013.10]
24. Gray CS, Barnsley J, Gagnon D, Belzile L, Kenealy T, Shaw J, et al. Using information communication technology in models of integrated community-based primary health care: learning from the iCOACH case studies. Implement Sci 2018;13(1):87 [FREE Full text] [doi: 10.1186/s13012-018-0780-3] [Medline: 29940992]
25. Latulippe K, Hamel C, Giroux D. Social health inequalities and eHealth: a literature review with qualitative synthesis of theoretical and empirical studies. J Med Internet Res 2017;19(4):e136 [FREE Full text] [doi: 10.2196/jmir.6731] [Medline: 28450271]
26. Fujioka JK, Budhwani S, Thomas-Jacques T, De Vera K, Challa P, Fuller K, et al. Challenges and strategies for promoting health equity in virtual care: protocol for a scoping review of reviews. JMIR Res Protoc 2020;9(12):e22847 [FREE Full text] [doi: 10.2196/22847] [Medline: 33211020]
27. Storeng KT, Fukuda‐Parr S, Mahajan M, Venkatapuram S. Digital technology and the political determinants of health inequities: special issue introduction. Glob Policy 2021;12(S6):5-11 [FREE Full text] [doi: 10.1111/1758-5899.13001]
28. Parker S, Prince A, Thomas L, Song H, Milosevic D, Harris M, IMPACT Study Group. Electronic, mobile and telehealth tools for vulnerable patients with chronic disease: a systematic review and realist synthesis. BMJ Open 2018;8(8):e019192 [FREE Full text] [doi: 10.1136/bmjopen-2017-019192] [Medline: 30158214]
29. Jones L, Jacklin K, O'Connell ME. Development and use of health-related technologies in Indigenous communities: critical review. J Med Internet Res 2017;19(7):e256 [FREE Full text] [doi: 10.2196/jmir.7520] [Medline: 28729237]
30. Arsenijevic J, Tummers L, Bosma N. Adherence to electronic health tools among vulnerable groups: systematic literature review and meta-analysis. J Med Internet Res 2020;22(2):e11613 [FREE Full text] [doi: 10.2196/11613] [Medline: 32027311]
31. Fraser S, Mackean T, Grant J, Hunter K, Towers K, Ivers R. Use of telehealth for health care of Indigenous peoples with chronic conditions: a systematic review. Rural Remote Health 2017;17(3):4205 [FREE Full text] [doi: 10.22605/RRH4205] [Medline: 28930638]
32. Truong M, Yeganeh L, Cook O, Crawford K, Wong P, Allen J. Using telehealth consultations for healthcare provision to patients from non-Indigenous racial/ethnic minorities: a systematic review. J Am Med Inform Assoc 2022;29(5):970-982 [FREE Full text] [doi: 10.1093/jamia/ocac015] [Medline: 35150266]
33. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post-COVID era. Milbank Q 2021;99(2):340-368 [FREE Full text] [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
34. Ortega G, Rodriguez JA, Maurer LR, Witt EE, Perez N, Reich A, et al. Telemedicine, COVID-19, and disparities: policy implications. Health Policy Technol 2020;9(3):368-371 [FREE Full text] [doi: 10.1016/j.hlpt.2020.08.001] [Medline: 32837888]
35. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020;323(23):2381-2382 [FREE Full text] [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
36. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]
37. Wood BR, Young JD, Abdel-Massih RC, McCurdy L, Vento TJ, Dhanireddy S, et al. Advancing digital health equity: a policy paper of the Infectious Diseases Society of America and the HIV Medicine Association. Clin Infect Dis 2021;72(6):913-919 [FREE Full text] [doi: 10.1093/cid/ciaa1525] [Medline: 33033829]
38. Ghobarah HA, Huth P, Russett B. Comparative public health: the political economy of human misery and well-being. Int Studies Q 2004;48(1):73-94 [FREE Full text] [doi: 10.1111/j.0020-8833.2004.00292.x]
39. McCartney G, Hearty W, Arnot J, Popham F, Cumbers A, McMaster R. Impact of political economy on population health: a systematic review of reviews. Am J Public Health 2019;109(6):e1-e12. [doi: 10.2105/AJPH.2019.305001] [Medline: 31067117]

XSL•FO
RenderX

40. Sell SK, Williams OD. Health under capitalism: a global political economy of structural pathogenesis. Rev Int Polit Econ 2019;27(1):1-25 [FREE Full text] [doi: 10.1080/09692290.2019.1659842]
41. Friel S, Townsend B, Fisher M, Harris P, Freeman T, Baum F. Power and the people's health. Soc Sci Med 2021;282:114173 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114173] [Medline: 34192622]
42. Kickbusch I, Allen L, Franz C. The commercial determinants of health. Lancet Glob Health 2016;4(12):e895-e896 [FREE Full text] [doi: 10.1016/s2214-109x(16)30217-0]
43. McKee M, Stuckler D. Revisiting the corporate and commercial determinants of health. Am J Public Health 2018;108(9):1167-1170. [doi: 10.2105/AJPH.2018.304510] [Medline: 30024808]
44. Jacobs D, Myers L. Union strength, neoliberalism, and inequality. Am Sociol Rev 2014;79(4):752-774 [FREE Full text] [doi: 10.1177/0003122414536392]
45. Scott-Samuel A, Smith KE. Fantasy paradigms of health inequalities: utopian thinking? Soc Theory Health 2015;13(3-4):418-436 [FREE Full text] [doi: 10.1057/sth.2015.12]
46. Thomason J. Big tech, big data and the new world of digital health. Glob Health J 2021;5(4):165-168 [FREE Full text] [doi: 10.1016/j.glohj.2021.11.003]
47. Sharon T. Blind-sided by privacy? Digital contact tracing, the Apple/Google API and big tech's newfound role as global health policy makers. Ethics Inf Technol 2021;23(Suppl 1):45-57 [FREE Full text] [doi: 10.1007/s10676-020-09547-x] [Medline: 32837287]
48. The big tech in healthcare report: how Amazon, Google, Microsoft, Apple, and Oracle are fighting for the $11T market. CB Insights. 2022. URL: https://www.cbinsights.com/research/report/famga-big-tech-healthcare/ [accessed 2022-01-21]
49. Powles J, Hodson H. Google DeepMind and healthcare in an age of algorithms. Health Technol (Berl) 2017;7(4):351-367 [FREE Full text] [doi: 10.1007/s12553-017-0179-1] [Medline: 29308344]
50. Costa A, Milne R. Understanding 'passivity' in digital health through imaginaries and experiences of coronavirus disease 2019 contact tracing apps. Big Data Soc 2022;9(1):20539517221091138 [FREE Full text] [doi: 10.1177/20539517221091138] [Medline: 36819735]
51. The digital health 150: the most promising digital health companies of 2022. CB Insights Res. 2022. URL: https://www.cbinsights.com/research/report/digital-health-startups-redefining-healthcare/ [accessed 2022-01-22]
52. Lehoux P, Daudelin G, Williams-Jones B, Denis JL, Longo C. How do business model and health technology design influence each other? Insights from a longitudinal case study of three academic spin-offs. Research Policy 2014;43(6):1025-1038 [FREE Full text] [doi: 10.1016/j.respol.2014.02.001]
53. Safavi KC, Cohen AB, Ting DY, Chaguturu S, Rowe JS. Health systems as venture capital investors in digital health: 2011-2019. NPJ Digit Med 2020;3(1):103 [FREE Full text] [doi: 10.1038/s41746-020-00311-5] [Medline: 32802968]
54. 5 digital health categories led VC investment in 2021. Digital Health Business Technology. 2022. URL: https://digitalhealth.modernhealthcare.com/finance/5-digital-health-categories-led-vc-investment-2021 [accessed 2022-01-22]
55. Lehoux P, Miller FA, Daudelin G. How does venture capital operate in medical innovation? BMJ Innov 2016;2(3):111-117 [FREE Full text] [doi: 10.1136/bmjinnov-2015-000079] [Medline: 27547447]
56. Lehoux P, Miller FA, Daudelin G, Urbach DR. How venture capitalists decide which new medical technologies come to exist. Sci Public Policy 2015;43(3):375-385 [FREE Full text] [doi: 10.1093/scipol/scv051]
57. Parthasarathy S. Health innovation policy for the people. Democracy Collaborative. URL: https://thenextsystem.org/sites/default/files/2021-11/Health-innovation-policy-FINAL.pdf [accessed 2024-03-08]
58. Benda NC, Veinot TC, Sieck CJ, Ancker JS. Broadband internet access is a social determinant of health!. Am J Public Health 2020;110(8):1123-1125 [FREE Full text] [doi: 10.2105/AJPH.2020.305784] [Medline: 32639914]
59. Greene D. The Promise of Access: Technology, Inequality, and the Political Economy of Hope. Cambridge, MA: MIT Press; 2021.
60. Chen W, Li X. Digital inequalities in American disadvantaged urban communities: access, skills, and expectations for digital inclusion programs. Inf Commun Soc 2021;25(13):1916-1933 [FREE Full text] [doi: 10.1080/1369118x.2021.1907434]
61. Vogels EA. Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. URL: https://www.pewresearch.org/short-reads/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/ [accessed 2022-05-25]
62. Durocher K, Boparai N, Jankowicz D, Strudwick G. Identifying technology industry-led initiatives to address digital health equity. Digit Health 2021;7:20552076211056156 [FREE Full text] [doi: 10.1177/20552076211056156] [Medline: 35340562]
63. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019;29(Supplement_3):13-17 [FREE Full text] [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
64. Hughes OE. Public Management and Administration: An Introduction. London, UK: Macmillan International Higher Education; 2012.
65. Digital health literacy. National Library of Medicine All of Us Program. 2022. URL: https://allofus.nnlm.gov/digital-health-literacy [accessed 2022-03-14]
66. Health and digital literacy partnership provides local learning for better health. NHS England. 2023. URL: https://library.nhs.uk/our-news/ [accessed 2023-01-20]

XSL•FO
RenderX

67. Thomas-Jacques T, Jamieson T, Shaw J. Telephone, video, equity and access in virtual care. NPJ Digit Med 2021;4(1):159 [FREE Full text] [doi: 10.1038/s41746-021-00528-y] [Medline: 34795356]
68. OHIP–bulletins–health care professionals–MOHLTC. Government of Ontario, Ministry of Health and Long-Term Care. URL: https://www.health.gov.on.ca/en/pro/programs/ohip/bulletins/redux/bul210704.aspx [accessed 2022-03-14]
69. Stamenova V, Chu C, Pang A, Tadrous M, Bhatia RS, Cram P. Using administrative data to explore potentially aberrant provision of virtual care during COVID-19: retrospective cohort study of ontario provincial data. J Med Internet Res 2021;23(9):e29396 [FREE Full text] [doi: 10.2196/29396] [Medline: 34313590]
70. Mehrotra A, Bhatia RS, Snoswell CL. Paying for telemedicine after the pandemic. JAMA 2021;325(5):431-432 [FREE Full text] [doi: 10.1001/jama.2020.25706] [Medline: 33528545]
71. Modi PK, Kaufman SR, Portney DS, Ryan AM, Hollenbeck BK, Ellimoottil C. Telemedicine utilization by providers in accountable care organizations. Mhealth 2019;5:10 [FREE Full text] [doi: 10.21037/mhealth.2019.03.02] [Medline: 31143768]
72. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
73. Shcherbina A, Mattsson CM, Waggott D, Salisbury H, Christle JW, Hastie T, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med 2017;7(2):3 [FREE Full text] [doi: 10.3390/jpm7020003] [Medline: 28538708]
74. Costanza-Chock S. Design Justice: towards an intersectional feminist framework for design theory and practice. 2018 Presented at: Design as a Catalyst for Change-DRS International Conference 2018; June 25-28, 2018; Limerick, Ireland. [doi: 10.21606/drs.2018.679]
75. Donia J, Shaw JA. Ethics and values in design: a structured review and theoretical critique. Sci Eng Ethics 2021;27(5):57 [FREE Full text] [doi: 10.1007/s11948-021-00329-2] [Medline: 34410536]
76. Holzmeyer C. Beyond 'AI for Social Good' (AI4SG): social transformations—not tech-fixes—for health equity. Interdiscip Sci Rev 2021;46(1-2):94-125 [FREE Full text] [doi: 10.1080/03080188.2020.1840221]
77. Bietti E. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. 2020 Presented at: FAT* '20: Conference on Fairness, Accountability, and Transparency; January 27-30, 2020; Barcelona, Spain p. 210-219 URL: https://dl.acm.org/doi/abs/10.1145/3351095.3372860 [doi: 10.1145/3351095.3372860]
78. van Maanen G. AI ethics, ethics washing, and the need to politicize data ethics. Digit Soc 2022;1(2):9 [FREE Full text] [doi: 10.1007/s44206-022-00013-3] [Medline: 35971404]
79. Safavi K, Mathews SC, Bates DW, Dorsey ER, Cohen AB. Top-funded digital health companies and their impact on high-burden, high-cost conditions. Health Aff (Millwood) 2019;38(1):115-123 [FREE Full text] [doi: 10.1377/hlthaff.2018.05081] [Medline: 30615535]
80. Advancing health justice. Health Equity Tracker. 2022. URL: https://healthequitytracker.org [accessed 2022-05-25]
81. Höchstädter A, Scheck B. What's in a name: an analysis of impact investing understandings by academics and practitioners. J Bus Ethics 2014;132(2):449-475 [FREE Full text] [doi: 10.1007/s10551-014-2327-0]
82. Brest P, Born K. When can impact investing create real impact. Stanf Soc Innov Rev 2013;11(4):31 [FREE Full text]
83. Cassatly S, Cassatly M. The affordable care act and digital health applications. J Med Pract Manage 2016;32(3):198-201. [Medline: 29944817]
84. Agrawal A, Hockerts K. Impact investing: review and research agenda. J Small Bus Entrep 2019;33(2):153-181 [FREE Full text] [doi: 10.1080/08276331.2018.1551457]
85. Low-wage earners to get high-speed Internet for $30 in Biden program. The Washington Post. 2022. URL: https://www.washingtonpost.com/business/2022/05/09/biden-internet-discount/ [accessed 2022-05-25]
86. Government of Canada announces affordable high-speed Internet to help connect low-income families and seniors. Innovation, Science and Economic Development Canada. 2022. URL: https://tinyurl.com/mrpyc3c8 [accessed 2022-05-25]
87. Medicaid | Lifeline smartphone. Sentara Health Plans. 2022. URL: https://www.optimahealth.com/plans/medicaid/lifeline-smartphone [accessed 2022-05-25]
88. Diversity plans to improve enrollment of participants from underrepresented racial and ethnic populations in clinical trials; draft guidance for industry; availability. Food and Drug Administration. URL: https://tinyurl.com/rfvpd7xj [accessed 2023-06-30]
89. Glover WJ, Hendricks-Sturrup R. Ethics and equity-centred perspectives in engineering systems design. In: Maier A, Oehmen J, Vermaas PE, editors. Handbook of Engineering Systems Design. Cham, Switzerland: Springer; 2023:1-24.

XSL•FO
RenderX

Edited by T Leung; submitted 03.03.23; peer-reviewed by L Husain, C Lyles PhD; comments to author 16.06.23; revised version received 30.06.23; accepted 27.02.24; published 26.03.24.

Please cite as:
Shaw J, Glover W
The Political Economy of Digital Health Equity: Structural Analysis
J Med Internet Res 2024;26:e46971
URL: https://www.jmir.org/2024/1/e46971
doi:10.2196/46971
PMID:38530341

©James Shaw, Wiljeana Glover. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Digital Alcohol Interventions Could Be Part of the Societal Response to Harmful Consumption, but We Know Little About Their Long-Term Costs and Health Outcomes

Katarina Ulfsdotter Gunnarsson[1*], MSc; Martin Henriksson[1*], PhD; Marcus Bendtsen[1*], PhD

Department of Health, Medicine, and Caring Sciences, Linköping University, Linköping, Sweden

[*]all authors contributed equally

Corresponding Author:
Katarina Ulfsdotter Gunnarsson, MSc
Department of Health, Medicine, and Caring Sciences
Linköping University
Campus US
Linköping, 58183
Sweden
Phone: 46 13281000
Email: katarina.ulfsdotter.gunnarsson@liu.se

Abstract

Alcohol consumption causes both physical and psychological harm and is a leading risk factor for noncommunicable diseases. Digital alcohol interventions have been found to support those looking for help by giving them tools for change. However, whether digital interventions can help tackle the long-term societal consequences of harmful alcohol consumption in a cost-effective manner has not been adequately evaluated. In this Viewpoint, we propose that studies of digital alcohol interventions rarely evaluate the consequences of wider dissemination of the intervention under study, and that when they do, they do not take advantage of modeling techniques that allow for appropriately studying consequences over a longer time horizon than the study period when the intervention is tested. We argue that to help decision-makers to prioritize resources for research and dissemination, it is important to model long-term costs and health outcomes. Further, this type of modeling gives important insights into the context in which interventions are studied and highlights where more research is required and where sufficient evidence is available. The viewpoint therefore invites the researcher not only to reflect on which interventions to study but also how to evaluate their long-term consequences.

(J Med Internet Res 2024;26:e44574) doi:10.2196/44574

KEYWORDS

alcohol; health economics; telemedicine; psychological harm; eHealth; digital intervention; decision-making

Beyond the Buzz: Unraveling the Nonexistent Long-Term Outcomes

Consumption of alcohol is in many societies the norm [1]. In Sweden, for instance, alcohol is consumed regularly by 4 of 5 adults, and approximately 1 of 3 adults are classified as harmful drinkers [2,3]. While evidence suggests that there is no safe limit on alcohol consumption [4], harmful drinking represents a marked increased risk of negative health and social consequences [1,4,5]. In 2016, alcohol was estimated to have contributed to approximately 6.2% of disability-adjusted life years (DALYs) among women and 5.1% of DALYs among men in Sweden, and was attributed to 4.5% and 5.7% of deaths (women and men, respectively) [1]. A health economic report showed that in 2017, alcohol consumption cost Swedish society €10 billion [6]. Sweden is not unique in this situation, as many societies worldwide face significant health and economic burdens due to the high prevalence of harmful alcohol consumption.

Reducing the harmful use of alcohol is on the World Health Organization's list of "best buys" for tackling noncommunicable diseases [7]. Recommended actions include excise taxes on alcoholic beverages and restrictions on the retail availability of alcohol. Provisioning of psychosocial interventions for persons with harmful alcohol use is also included on the list of recommended actions, and has been operationalized by, for instance, delivering face-to-face interventions in primary health care [8]. With the ubiquity of internet connectivity in high-income countries, and increasingly in low- and

XSL•FO
RenderX

middle-income countries, there is interest to provide *digital* psychosocial interventions to those who may benefit. Digital interventions—that is, interventions that deliver supportive content through for instance websites, mobile phone apps, text messages, or email—can support those looking for help by giving them tools to increase their motivation, form intentions for change, and give them tools to help support change [9,10]. Digital interventions can scale to large populations and can be designed to be anonymous, which can reduce the stigma of looking for help in face-to-face settings [11-13]. Studies have found that digital alcohol interventions may have positive effects on behavior in a range of different populations [14-17], and evidence from meta-analyses confirms these findings [18-20]. However, whether or not they can help tackle the long-term consequences of harmful alcohol consumption, for instance by reducing the incidence of noncommunicable diseases, and whether or not they can do so in a cost-effective manner, is still uncertain. In contrast, modeling studies have demonstrated that face-to-face interventions hold promise in mitigating health consequences while being cost-effective over the long term [21]. Thus, limiting our assessment only to the short-term effects of digital interventions means that we cannot compare the societal impacts between modalities.

We propose that studies of digital alcohol interventions rarely evaluate the consequences of wider dissemination of the intervention under study, and that when they do, they do not take advantage of modeling techniques that allow for appropriately studying consequences over a longer time horizon than the study's period when the intervention is tested. We support our viewpoint with a pragmatic review of the literature (Multimedia Appendix 1 [22-27]) and make suggestions for how future studies may evaluate the impact of digital alcohol interventions at the societal level—which includes workplace and productivity, social services and nonstatutory care, and the criminal justice sector—and not only consequences related to the health care sector such as costs for treatment. We also invite researchers to reflect on their decision-making when it comes to deciding which intervention research projects should be prioritized; in particular, if formal evaluations of the long-term consequences of dissemination are considered, or even required, in the current decision-making process.

Our View

Health economic evaluations of digital alcohol interventions are scarce. When done, the consequences of interventions are evaluated over a short period, the period over which a randomized controlled trial (RCT) is run, ranging from 4 to 12 months. While the reports identified in our pragmatic review (Multimedia Appendix 1 [22-27]) all concluded that the interventions were cost-effective, they based these findings on short-term follow-up data collected during the trial period—it stands to reason that while there certainly are acute consequences from alcohol consumption, the long-term consequences are substantial and cannot be captured when evaluating such short time horizons [28]. Thus, current evaluations cannot readily answer the question being asked: What are the consequences of disseminating digital alcohol interventions into society? Considering the growth in interest and resources put into studying digital alcohol interventions, it is unfortunate that the literature cannot provide evidence of the long-term consequences of this investment.

The literature can however provide examples of long-term health economic evaluations outside the scope of digital interventions. Studies of the long-term consequences of nondigital alcohol interventions have been conducted, including the development and use of the Sheffield Alcohol Policy Model [29], which evaluates public health strategies for alcohol harm reduction. In addition, a review of nondigital brief alcohol interventions found that 14 out of the 23 included studies used modeling techniques that allowed for estimating consequences over a lifetime horizon. The majority of the 14 studies discovered that they either saved costs while improving health or incurred minimal costs compared to the health benefits [21]. Thus, there exist examples in the literature for evaluating alcohol interventions, which can inspire and be partially adopted by the digital intervention research field. Other literature can be followed to model long-term health economic outcomes, which guides the construction, analysis, and implementation of health economic models [30,31]. In addition, if it is important to model interactions among individuals, their propensity to use support, and relationships between multiple physical and mental health conditions, then agent-based models can be used to capture these complexities and forecast the long-term effects of interventions [32,33].

Importance of Health Economic Research

The importance of health economic evaluations in general, and modeling studies in particular, should not be understated. They provide input to decision-makers faced with difficult prioritization tasks, that is, if resources should be invested in disseminating a specific intervention. In many jurisdictions with a publicly funded health care system, they have long been an integral part of health policy [34] with a broad acceptance that modeling methods are required [35], not least in the evaluation of cancer drugs where intermediate outcomes (eg, progression-free survival) are routinely extrapolated to outcomes such as mortality [36]. It seems prudent that public health care efforts also should provide similar input of long-term outcomes since they often fall under the jurisdiction of public health care systems, however, this is not the case today. Their role can further be extended to also guide the prioritization of research resources. In developing an intervention, one may consider if a hypothetical intervention with a potential effect size even is worth developing and evaluating in an RCT. If such a hypothetical intervention cannot be shown to be cost-effective, then there is no reason to invest resources in its development and evaluation. Researchers are routinely asked by review panels to provide sample size estimates, which are based on anticipated effect sizes. These anticipated effect sizes commonly fall into 2 categories [37-39]: so-called clinical significance, which presents effect sizes large enough that the anticipated sample size is achievable, or minimum relevant effect size, which ensures that even small effect estimates are identified as statistically significant. Interestingly, these anticipated effect sizes are rarely (if ever) put to the test in a health economic evaluation where their long-term costs and health implications

XSL•FO
RenderX

can be estimated; thus, a rationale for a study based on a formal evaluation of the long-term consequences of dissemination of the proposed intervention is not produced. We argue that health economic evaluations before RCTs can help reduce research waste and allow both researchers and other decision-makers to prioritize resources more effectively and, indeed, even identify cost-effective research designs [40].

It should be acknowledged that conducting health economic evaluations is not without barriers. It requires a certain technical skill set by the researcher conducting the evaluation, usually involving statistical software to create models that can extrapolate over time. It also requires epidemiological expertise to decide on input variables, as well as knowledge of the costs and consequences from both the health care and societal perspectives concerning treatment and disease. These are formidable challenges; however, approaching these challenges confers additional benefits. First, it puts into context the intervention being studied, allowing researchers to carefully think about the degree to which interventions will be adopted, their anticipated short- and long-term costs and effects, and how the novel intervention relates to existing interventions. Second, it makes it very clear where evidence is lacking from both an efficacy, epidemiologic, and economic perspective. For instance, when faced with having to decide on the risk of disease given a range of contextual variables, the process of conducting a health economic evaluation highlights where epidemiological studies are required, but also where the evidence is sufficient. Thus, by conducting a health economic evaluation, research prioritizations can become clear beyond the primary objective of the study. These prioritizations can be further refined by conducting sensitivity analyses to show which input variables affect the outcome of the evaluation the most, refining the agenda for which phenomena require further study. Third, by taking our proposed route, one that has been the gold standard for a long time in clinical medicine, decision-makers will be provided with the relevant evidence when prioritizing scarce resources that can be used for the provision of interventions or for further research. Basing evaluations solely on data from 1 RCT will yield a very precise estimate of an irrelevant cost-effectiveness estimate; thus, our advice is to follow the old advice: It is better to be roughly right than precisely wrong.

To conclude, health economic evaluations of digital alcohol interventions lack modeling of long-term societal consequences, leaving a knowledge gap concerning the degree to which they may address the burgeoning societal burdens caused by alcohol consumption. We invite researchers to reflect on their decision-making process when it comes to deciding which studies should be prioritized.

Acknowledgments

This study was conducted under the auspices of the Swedish Research Council for Health, Working Life, and Welfare (2022-00193). The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Conflicts of Interest

MB owns a private company (Alexit AB) that maintains and distributes evidence-based lifestyle interventions to be used by the public and in health care settings. Alexit AB played no role in the study design, data analysis, data interpretation, or writing of this report. KUG and MH declare no competing interests.

Multimedia Appendix 1
A pragmatic review.
[DOCX File , 52 KB - jmir_v26i1e44574_app1.docx]

References

1. GBD 2016 Alcohol Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2018;392(10152):1015-1035 [FREE Full text] [doi: 10.1016/S0140-6736(18)31310-2] [Medline: 30146330]
2. Guttormsson U. Självrapporterade alkoholvanor i Sverige 2004–2019. Centralförbundet för alkohol- och narkotikaupplysning. 2020. URL: https://www.can.se/app/uploads/2020/10/can-rapport-195-sjlvrapporterade-alkoholvanor-i-sverige-2004-2019.pdf [accessed 2023-03-13]
3. Reid MC, Fiellin DA, O'Connor PG. Hazardous and harmful alcohol consumption in primary care. Arch Intern Med 1999;159(15):1681-1689. [doi: 10.1001/archinte.159.15.1681] [Medline: 10448769]
4. Wood AM, Kaptoge S, Butterworth AS, Willeit P, Warnakula S, Bolton T, et al. Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599 912 current drinkers in 83 prospective studies. Lancet 2018;391(10129):1513-1523 [FREE Full text] [doi: 10.1016/S0140-6736(18)30134-X] [Medline: 29676281]
5. Laslett AM, Room R, Waleewong O, Stanesby O, Callinan S, editors. Harm to Others from Drinking: Patterns in Nine Societies. Geneva, Switzerland: World Health Organization; 2019.
6. Alkoholens samhällsekonomiska konsekvenser: en beskrivande samhällsekonomisk studie. Ramboll Management Consulting AB. 2019. URL: https://www.omsystembolaget.se/globalassets/pdf/alkoholproblematik/ramboll---resultatrapport----alkoholens-samhallskonsekvenser---uppdaterad-oktober-2020.pdf [accessed 2022-07-09]

XSL•FO
RenderX

7. Tackling NCDs: “best buys” and other recommended interventions for the prevention and control of noncommunicable diseases. World Health Organization. 2017. URL: https://iris.who.int/handle/10665/259232 [accessed 2023-11-14]
8. O'Donnell A, Anderson P, Newbury-Birch D, Schulte B, Schmidt C, Reimer J, et al. The impact of brief alcohol interventions in primary healthcare: a systematic review of reviews. Alcohol Alcohol 2014;49(1):66-78 [FREE Full text] [doi: 10.1093/alcalc/agt170] [Medline: 24232177]
9. Garnett C, Crane D, Brown J, Kaner E, Beyer F, Muirhead C, et al. Reported theory use by digital interventions for hazardous and harmful alcohol consumption, and association with effectiveness: meta-regression. J Med Internet Res 2018;20(2):e69 [FREE Full text] [doi: 10.2196/jmir.8807] [Medline: 29490895]
10. Garnett CV, Crane D, Brown J, Kaner EFS, Beyer FR, Muirhead CR, et al. Behavior change techniques used in digital behavior change interventions to reduce excessive alcohol consumption: a meta-regression. Ann Behav Med 2018;52(6):530-543 [FREE Full text] [doi: 10.1093/abm/kax029] [Medline: 29788261]
11. Cohen E, Feinn R, Arias A, Kranzler HR. Alcohol treatment utilization: findings from the National Epidemiologic Survey on Alcohol and Related Conditions. Drug Alcohol Depend 2007;86(2-3):214-221. [doi: 10.1016/j.drugalcdep.2006.06.008] [Medline: 16919401]
12. Rehm J, Allamani A, Elekes Z, Jakubczyk A, Manthey J, Probst C, et al. Alcohol dependence and treatment utilization in Europe: a representative cross-sectional study in primary care. BMC Fam Pract 2015;16(1):90 [FREE Full text] [doi: 10.1186/s12875-015-0308-8] [Medline: 26219430]
13. Wallhed Finn S, Bakshi AS, Andréasson S. Alcohol consumption, dependence, and treatment barriers: perceptions among nontreatment seekers with alcohol dependence. Subst Use Misuse 2014;49(6):762-769. [doi: 10.3109/10826084.2014.891616] [Medline: 24601784]
14. McCambridge J, Bendtsen M, Karlsson N, White IR, Nilsen P, Bendtsen P. Alcohol assessment and feedback by email for university students: main findings from a randomised controlled trial. Br J Psychiatry 2013;203(5):334-340 [FREE Full text] [doi: 10.1192/bjp.bp.113.128660] [Medline: 24072758]
15. Bendtsen M, Åsberg K, McCambridge J. Effectiveness of a digital intervention versus alcohol information for online help-seekers in Sweden: a randomised controlled trial. BMC Med 2022;20(1):176 [FREE Full text] [doi: 10.1186/s12916-022-02374-5] [Medline: 35578276]
16. Gustafson DH, McTavish FM, Chih MY, Atwood AK, Johnson RA, Boyle MG, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA Psychiatry 2014;71(5):566-572 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.4642] [Medline: 24671165]
17. Suffoletto B, Kristan J, Chung T, Jeong K, Fabio A, Monti P, et al. An interactive text message intervention to reduce binge drinking in young adults: a randomized controlled trial with 9-month outcomes. PLoS One 2015;10(11):e0142877 [FREE Full text] [doi: 10.1371/journal.pone.0142877] [Medline: 26580802]
18. Riper H, Hoogendoorn A, Cuijpers P, Karyotaki E, Boumparis N, Mira A, et al. Effectiveness and treatment moderators of internet interventions for adult problem drinking: an individual patient data meta-analysis of 19 randomised controlled trials. PLoS Med 2018;15(12):e1002714 [FREE Full text] [doi: 10.1371/journal.pmed.1002714] [Medline: 30562347]
19. Kaner EFS, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
20. Bendtsen M, McCambridge J, Åsberg K, Bendtsen P. Text messaging interventions for reducing alcohol consumption among risky drinkers: systematic review and meta-analysis. Addiction 2021;116(5):1021-1033 [FREE Full text] [doi: 10.1111/add.15294] [Medline: 33047865]
21. Angus C, Latimer N, Preston L, Li J, Purshouse R. What are the implications for policy makers? A systematic review of the cost-effectiveness of screening and brief interventions for alcohol misuse in primary care. Front Psychiatry 2014;5:114 [FREE Full text] [doi: 10.3389/fpsyt.2014.00114] [Medline: 25225487]
22. Drost RMWA, Paulus ATG, Jander AF, Mercken L, de Vries H, Ruwaard D, et al. A web-based computer-tailored alcohol prevention program for adolescents: cost-effectiveness and intersectoral costs and benefits. J Med Internet Res 2016;18(4):e93 [FREE Full text] [doi: 10.2196/jmir.5223] [Medline: 27103154]
23. Deluca P, Coulton S, Alam MF, Boniface S, Cohen D, Donoghue K, et al. Brief interventions to prevent excessive alcohol use in adolescents at low-risk presenting to emergency departments: three-arm, randomised trial of effectiveness and cost-effectiveness. Int J Drug Policy 2021;93:103113 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103113] [Medline: 33487528]
24. Blankers M, Nabitz U, Smit F, Koeter MWJ, Schippers GM. Economic evaluation of internet-based interventions for harmful alcohol use alongside a pragmatic randomized controlled trial. J Med Internet Res 2012;14(5):e134 [FREE Full text] [doi: 10.2196/jmir.2052] [Medline: 23103771]
25. Mujcic A, Blankers M, Boon B, Berman AH, Riper H, van Laar M, et al. Effectiveness, cost-effectiveness, and cost-utility of a digital alcohol moderation intervention for cancer survivors: health economic evaluation and outcomes of a pragmatic randomized controlled trial. J Med Internet Res 2022;24(2):e30095 [FREE Full text] [doi: 10.2196/30095] [Medline: 35103605]

XSL•FO
RenderX

26. Boß L, Lehr D, Schaub MP, Paz Castro R, Riper H, Berking M, et al. Efficacy of a web-based intervention with and without guidance for employees with risky drinking: results of a three-arm randomized controlled trial. Addiction 2018;113(4):635-646 [FREE Full text] [doi: 10.1111/add.14085] [Medline: 29105879]
27. Smit F, Lokkerbol J, Riper H, Majo MC, Boon B, Blankers M. Modeling the cost-effectiveness of health care systems for alcohol use disorders: how implementation of eHealth interventions improves cost-effectiveness. J Med Internet Res 2011;13(3):e56 [FREE Full text] [doi: 10.2196/jmir.1694] [Medline: 21840836]
28. Holmes J, Meier PS, Booth A, Guo Y, Brennan A. The temporal relationship between per capita alcohol consumption and harm: a systematic review of time lag specifications in aggregate time series analyses. Drug Alcohol Depend 2012;123(1-3):7-14 [FREE Full text] [doi: 10.1016/j.drugalcdep.2011.12.005] [Medline: 22197480]
29. Brennan A, Meier P, Purshouse R, Rafia R, Meng Y, Hill-Macmanus D, et al. The Sheffield alcohol policy model: a mathematical description. Health Econ 2015;24(10):1368-1388 [FREE Full text] [doi: 10.1002/hec.3105] [Medline: 25270223]
30. Briggs AH, Claxton K, Sculpher M. Decision Modelling For Health Economic Evaluation. Oxford, UK: Oxford University Press; 2006.
31. Alarid-Escudero F, Krijkamp EM, Enns EA, Yang A, Hunink MGM, Pechlivanoglou P, et al. An introductory tutorial on cohort state-transition models in R using a cost-effectiveness analysis example. Med Decis Making 2023;43(1):3-20 [FREE Full text] [doi: 10.1177/0272989X221103163] [Medline: 35770931]
32. Silverman E, Gostoli U, Picascia S, Almagor J, McCann M, Shaw R, et al. Situating agent-based modelling in population health research. Emerg Themes Epidemiol 2021;18(1):10 [FREE Full text] [doi: 10.1186/s12982-021-00102-7] [Medline: 34330302]
33. Badham J, Chattoe-Brown E, Gilbert N, Chalabi Z, Kee F, Hunter RF. Developing agent-based models of complex health behaviour. Health Place 2018;54:170-177 [FREE Full text] [doi: 10.1016/j.healthplace.2018.08.022] [Medline: 30290315]
34. Claxton K, Sculpher M, Drummond M. A rational framework for decision making by the National Institute For Clinical Excellence (NICE). Lancet 2002 Aug 31;360(9334):711-715. [doi: 10.1016/S0140-6736(02)09832-X] [Medline: 12241891]
35. EUnetHTA Joint Action 2, Work Package 7, Subgroup 3, Heintz E, Gerber-Grote A, Ghabri S, Hamers FF, Rupel VP, et al. Is there a European view on health economic evaluations? results from a synopsis of methodological guidelines used in the EUnetHTA partner countries. Pharmacoeconomics 2016;34(1):59-76. [doi: 10.1007/s40273-015-0328-1] [Medline: 26446858]
36. Latimer NR, Adler AI. Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. BMJ Med 2022;1(1):e000094 [FREE Full text] [doi: 10.1136/bmjmed-2021-000094] [Medline: 36936578]
37. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ 2009;338:b1732 [FREE Full text] [doi: 10.1136/bmj.b1732] [Medline: 19435763]
38. Raju TNK, Langenberg P, Sen A, Aldana O. How much 'better' is good enough? The magnitude of treatment effect in clinical trials. Am J Dis Child 1992;146(4):407-411. [doi: 10.1001/archpedi.1992.02160160027007] [Medline: 1558071]
39. Vickers AJ. Underpowering in randomized trials reporting a sample size calculation. J Clin Epidemiol 2003;56(8):717-720. [doi: 10.1016/s0895-4356(03)00141-0] [Medline: 12954462]
40. Fenwick E, Steuten L, Knies S, Ghabri S, Basu A, Murray JF, et al. Value of information analysis for research decisions-an introduction: report 1 of the ISPOR value of information analysis emerging good practices task force. Value Health 2020;23(2):139-150 [FREE Full text] [doi: 10.1016/j.jval.2020.01.001] [Medline: 32113617]

Abbreviations

DALY: disability-adjusted life year
RCT: randomized controlled trial

Edited by T Leung, T de Azevedo Cardoso; submitted 24.11.22; peer-reviewed by S Bonn, W Campbell, R Davis; comments to author 04.03.23; revised version received 19.04.23; accepted 13.02.24; published 27.03.24.

Please cite as:
Ulfsdotter Gunnarsson K, Henriksson M, Bendtsen M
Digital Alcohol Interventions Could Be Part of the Societal Response to Harmful Consumption, but We Know Little About Their Long-Term Costs and Health Outcomes
J Med Internet Res 2024;26:e44574
URL: https://www.jmir.org/2024/1/e44574
doi:10.2196/44574
PMID:38536228

XSL•FO
RenderX

©Katarina Ulfsdotter Gunnarsson, Martin Henriksson, Marcus Bendtsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Challenges in Using eHealth Decision Resources for Surrogate Decision-Making in the Intensive Care Unit

Wan-Na Sun[1], MSN; Chi-Yin Kao[2], PhD

[1]Department of Nursing, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[2]Department of Nursing, College of Medicine, National Cheng Kung University, Tainan, Taiwan

Corresponding Author:
Chi-Yin Kao, PhD
Department of Nursing, College of Medicine
National Cheng Kung University
1 University Road
Tainan, 701
Taiwan
Phone: 886 6 2353535 ext 5843
Fax: 886 6 2377550
Email: chiyinkao@mail.ncku.edu.tw

Abstract

The mortality rate in intensive care units (ICUs) is notably high, with patients often relying on surrogates for critical medical decisions due to their compromised state. This paper provides a comprehensive overview of eHealth. The challenges of applying eHealth tools, including economic disparities and information inaccuracies are addressed. This study then introduces eHealth literacy and the assessment tools to evaluate users' capability and literacy levels in using eHealth resources. A clinical scenario involving surrogate decision-making is presented. This simulated case involves a patient with a hemorrhagic stroke who has lost consciousness and requires medical procedures such as tracheostomy. However, due to the medical surrogate's lack of familiarity with eHealth devices and limited literacy in using eHealth resources, difficulties arise in assisting the patient in making medical decisions. This scenario highlights challenges related to eHealth literacy and solution strategies are proposed. In conclusion, effective ICU decision-making with eHealth tools requires a careful balance between efficiency with inclusivity. Tailoring communication strategies and providing diverse materials are essential for effective eHealth decision resources in the ICU setting. Health professionals should adopt a patient-centered approach to enhance the decision-making experience, particularly for individuals with limited eHealth literacy.

(J Med Internet Res 2024;26:e47017) doi:10.2196/47017

KEYWORDS

decision-making; eHealth; intensive care unit; literacy; surrogate; mobile phone

Introduction

Decision-Making by Patient Surrogates

The mortality rate of patients in intensive care units (ICUs) exceeds that of those in the general wards. The Society of Critical Care Medicine reported a mortality rate ranging from 10% to 29% in adult ICUs [1]. In Taiwan, the mortality rate in ICUs is 11.6% [2]. The patients in ICUs, grappling with impaired cardiopulmonary functions and severe illness, are in critical condition and are typically physically vulnerable or even in an unconscious state [3]. They depend on the health care team for basic daily activities such as eating and toileting. In addition, given their critical condition, a variety of complex medical treatments such as endotracheal intubation and ventilator use, are often required to maintain their vital signs [4]. Coupled with the administration of multiple medications, patients often find it challenging to express their preferences. If patients have not previously provided an advance directive, they must rely on surrogates for medical decisions. Statistics indicate that 95% of surrogates make at least one major medical decision upon the patient's ICU admission [4]. Another study highlights that surrogates are involved in up to 71% of medical decisions within the first 48 hours of the patient's ICU stay [5,6], underscoring the prevalence of surrogates in guiding medical choice decisions for critically ill patients. However, the decision-making process places significant pressure on surrogates, forcing them to reach medical decisions quickly in collaboration with health professionals [5,6]. This pressure can lead to conflicts in decision-making and potentially disrupt the harmony of the

XSL•FO
RenderX

medical-patient relationship, all exacerbated by the psychological strain of the possible loss of a loved one and time constraints.

Miller et al [4] noted that 30% of surrogates experienced anxiety and depression, and up to 50% even developed posttraumatic stress disorder. In addition, surrogates possess varying levels of information and knowledge regarding medical treatment, which can influence their capacity to assess and participate in medical decision-making, consequently affecting their ability to comprehend information while communicating with health professionals and making decisions on behalf of the patients [7].

However, with the advancement of health care technology and the catalyzing effect of the pandemic, the use of eHealth tools to assist surrogates in decision-making has become a trend. However, it may pose several challenges. For instance, surrogates may require time to learn to use these eHealth tools effectively [8,9]. Health care professionals must provide effective training and support to ensure that surrogates can fully use these tools. During the use of eHealth tools, there should also be a mechanism for immediate clarification if any doubts arise [8,10]. Furthermore, despite the use of eHealth tools, face-to-face communication remains essential for surrogates. Through in-person observations, one can understand and clarify considerations related to medical interventions stemming from different cultures and family backgrounds [10]. This is vital to genuinely assist surrogates in making informed decisions.

Use of eHealth Tools in Decision-Making

The term “eHealth” has gained popularity. In 2016, the World Health Organization (WHO) introduced the term to encompass the latest advancements in medical technology, particularly the use of information and communication technology (ICT) to enhance the function and efficiency of the health care system. eHealth encompasses more than just internet medicine; it also includes telehealth and telemedicine. It is a dynamic field that is in a constant state of evolution, signifying not only technological progress but also shifts in how we envision improving the health care landscape as well [11,12].

Scholars regard eHealth as an emerging domain of this generation. It is considered a highly convenient, easily accessible, and user-friendly tool that captures the attention of users [11,12]. Among the most widely used eHealth services is the dissemination of health care information through means such as QR codes, videos, and websites [9]. Many developing countries are transitioning away from paper-based methods, integrating hospital data through ICT, replacing conventional patient care models, and using technology to alleviate the workload on nursing staff to reduce unnecessary costs for individuals and society and alleviate economic and medical burdens. Research has revealed that over 50% of users seek health- or disease-related information through eHealth while up to 60% make decisions based on the information they find [8]. Moreover, in high-income countries, nearly 80% of adults regularly use the internet to access health information [10].

Furthermore, by disseminating pertinent medical and disease-related information through internet platforms and other channels, health professionals can enhance health care knowledge among the public, patients, and their families. This fosters greater awareness of illnesses and facilitates immediate diagnosis and treatment for health concerns, which significantly augments health literacy and diminishes the number of patients facing serious illnesses [10]. In recent years, there has been extensive use of the patient decision aid tools developed using eHealth technology. Some scholars even highlight the potential for eHealth to enhance equitable medical care for remote areas and among patients with limited mobility [10].

Nevertheless, health professionals should carefully consider whether eHealth is suitable for all groups and diseases as a decision-making tool. This is an issue that warrants our attention, as indiscriminate use across all demographics might not be advisable.

Challenges of Applying eHealth Tools

The application of eHealth tools can present a range of challenges, including economic disparities, information inaccuracies, and the potential exclusion of specific groups. For instance, government agencies should formulate policies and ensure the provision of adequate equipment. Failure to do so may result in unequal access to eHealth tools, favoring higher socioeconomic groups while leaving lower socioeconomic groups without access. This misalignment with the principle of equitable health care is a concern, as some medical institutions may initially refuse to offer services for marginalized groups such as Black communities and low-income individuals [9]. Consequently, a comprehensive support system should be in place before the implementation of eHealth tools.

When conveying medical and disease-related information, it is crucial to consider the risk of patients’ privacy being compromised or the possibility of conveying inaccurate and misleading messages [9]. For instance, Ruppert et al [13] searched for medical information on the social platform YouTube, focusing on terms such as skin cancer and sunscreen. Their findings indicated that up to 40% of the messages contained misinformation and 20% of the messages were potentially misleading [13]. Swartzendruber et al [14] explored the Women’s Resource Center of China and found that up to 53% of the content provided incorrect information about breast cancer treatment, with 84% of the website content being dedicated to advertising programs.

These findings underscore the fact that, despite the eHealth revolution, there are still numerous unresolved issues that require attention. In addition to effectively harnessing technology for caregiving, careful planning is essential to address the potential pitfalls associated with technology. In this era, where eHealth is a prevailing trend, it is crucial to consider how we can make better and more user-friendly use of it. Therefore, there is a need for tools to assess the suitability of patients for using eHealth tools.

eHealth Literacy and Assessment Tools

As medical technology advances, an increasing number of countries are actively promoting national health literacy through ICT. However, the main purpose of using eHealth technologies is to empower users to make informed health decisions by

XSL•FO
RenderX

finding, comprehending, and discerning relevant health information, and then applying this knowledge to address health-related concerns [15]. It is important to note that eHealth differs from traditional health literacy in that it entails using electronic resources to access, understand, and distinguish pertinent health information and then apply this knowledge to deal with health issues [15,16]. It is a dynamic process influenced by factors such as the user's background, knowledge, educational level, and health status, and it can lead to different motivations and strategies in various contexts [15,17].

However, if only a limited number of people participate in or use the eHealth tools, the goal of universalization for these tools is compromised; accordingly, it is important to assess users' capability and literacy level to provide them with appropriate eHealth tools to enhance their health literacy.

Norman and Skinner [15] introduced the eHealth Literacy Lily Model (Figure 1), which consists of 2 categories, analytic and context-specific. The model comprises 6 various components of literacy: traditional, information, media, health, computer, and scientific. The analytic components (traditional, information, and media) mainly evaluate the users' fundamental skills and are not directly tied to the content of health literacy. In contrast, the context-specific components (health, computer, and scientific) are more focused on context-specific and computational skills. The following section elaborates on these 6 components of the eHealth Literacy Lily Model. Table 1 provides more details.

Figure 1. eHealth Literacy Lily Model.

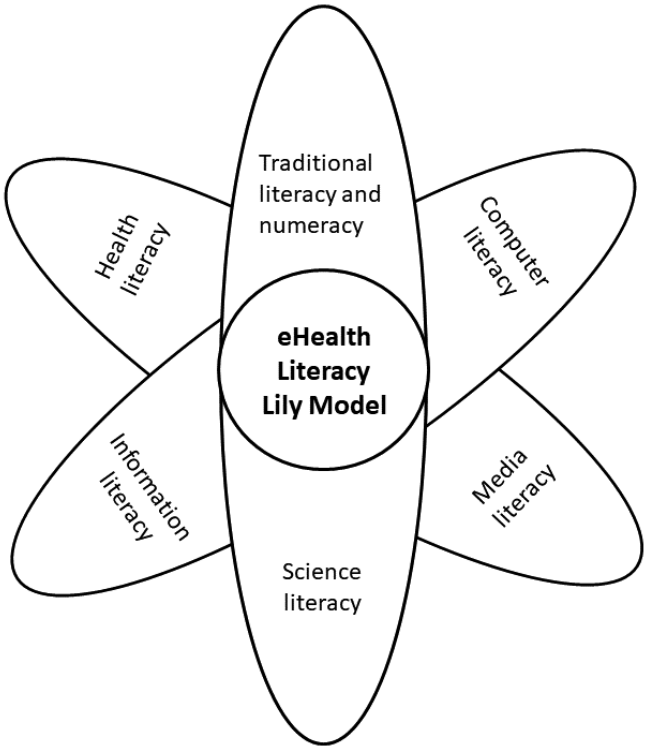

XSL•FO
RenderX

Table 1. Definitions of the eHealth Literacy Lily Model.

Central types and 6 components	Definition
Analytic type	
Traditional literacy	The most fundamental general abilities include reading and writing skills, and the ability to express oneself logically through reading.
Information literacy	The ability to find information online or in other ways to address the health literacy issues required by an individual, and even in such a way that others can learn from them.
Media literacy	The media would consider sociocultural background, market demand, political considerations, and other factors when delivering health literacy messages. Media literacy pertains to whether we can perceive and think critically about such information. If users are unable to understand and organize the content of health information and knowledge from the media, or even develop subjective judgments, they are inadequate in this core skill.
Context-specific type	
Health literacy	The Institute of Medicine in the United States points out that health literacy affects one's ability to engage in appropriate self-care, which is a necessary skill to interact with the health system, including basic reading and calculating skills.
Computer literacy	A necessary skill in eHealth, which refers to problem-solving with the use of computer devices and 3C products (computers, communications, and consumer electronics) currently available on the market, for example, desktops, iPads, and mobile phones and the use of software such as email and social network apps.
Scientific literacy	It pertains to the understanding of natural science as well as the aims, methods, and limitations of scientific research.

Substantiating science-based online health messages in health literacy is a common approach, but it can be a significant challenge for individuals who lack adequate education. This is particularly true since users without these 6 core skills could face difficulties or obstacles when trying to adopt eHealth [9,15]. For this reason, assessing users' health literacy has become an essential step before implementing eHealth, and the availability of quantitative data enables health professionals to provide evidence-based and useful eHealth tools for users.

Norman and Skinner [15] developed the eHealth Literacy Scale in English, although it has been subsequently translated into multiple languages including Amharic, traditional and simplified Chinese, Dutch, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Korean, Persian, Polish, Portuguese, Norwegian, and Swedish [18]. The questionnaire consists of 8 questions rated on a 5-point Likert scale, ranging from 1 for "Not useful at all" to 5 for "Very Useful." A total score range is 0-40, with higher scores indicating a better ability to use eHealth. A score of ≤31 points indicates a lower level of eHealth literacy. Previous studies have reported a high internal consistency coefficient α of .88 [19,20]. Further details are provided in Table 2.

Table 2. The eHealth Literacy Scale.

No	Items	1[a]	2[b]	3[c]	4[d]	5[e]
1	I know how to find helpful health resources on the internet.	☐	☐	☐	☐	☐
2	I know how to use the internet to answer my health questions.	☐	☐	☐	☐	☐
3	I know what health resources are available on the internet.	☐	☐	☐	☐	☐
4	I know where to find helpful health resources on the internet.	☐	☐	☐	☐	☐
5	I know how to use the health information I find on the internet to help me.	☐	☐	☐	☐	☐
6	I have the skills I need to evaluate the health resources I find on the internet.	☐	☐	☐	☐	☐
7	I can tell high quality from low-quality health resources on the internet.	☐	☐	☐	☐	☐
8	I feel confident in using information from the internet to make health decisions.	☐	☐	☐	☐	☐

[a]1: not useful at all.
[b]2: not useful.
[c]3: unsure.
[d]4: useful.
[e]5: very useful.

Clinical Scenario

Overview

In this section, we present a simulated clinical scenario to address the challenges associated with applying an eHealth tool. We then use the eHealth Literacy Lily Model and the eHealth Literacy Scale to analyze the situation and propose potential solutions.

Yuliya, a 68-year-old woman, lives with her husband, Eric. Their 2 children are working abroad, and they communicate via

XSL•FO
RenderX

internet phone calls. Yuliya does not have any underlying medical conditions and regularly attends government-provided health checkups, following the advice of her neighbors. Her primary sources of health information are her children and neighbors.

Eric, on the other hand, has a complex medical history, including diabetes, chronic kidney disease stage IV, and congestive heart failure stage 2 according to the New York Heart Association. In 2020, Yuliya returned home to find Eric unconscious. After Eric was sent to the emergency room, his cognitive alertness was unclear, EVM of Glasgow Coma Scale (GCS) was E1V2M3; he was then diagnosed with COVID-19 (cycle threshold value 19.5 / polymerase chain reaction showed positive) and acute respiratory failure with chest x-ray showing pneumonia and SpO_2 range of 77%-82%. The doctor performed tracheal intubation with a ventilator, and Eric was subsequently transferred to the ICU. The doctor informed Yuliya that Eric's consciousness had not recovered and that he needed a computed tomography (CT) scan. The doctor hoped that Yuliya could make an immediate decision regarding the CT scan after the explanation. However, Yuliya was terrified and did not understand the doctor's explanation; nevertheless, she consented to the CT scan. Sadly, Eric was diagnosed with a hemorrhagic stroke of basal ganglia, and even after 2 weeks, he was still unable to breathe spontaneously due to the stroke and he remained in an unconscious state. The doctor recommended a tracheostomy for Eric, proposing that Eric needed long-term care in the future. However, Yuliya was faced with a difficult decision regarding the procedure.

To facilitate Yuliya's understanding, the nurse provided her with a QR code to access a video. The nurse emphasized that the video contained information about tracheostomy procedures and care. Yuliya was instructed to watch the video and decide whether Eric should undergo the tracheostomy. However, Yuliya encountered challenges as she was unfamiliar with how to use the QR code, and she had limited literacy skills, with her only means of communication being in Taiwanese. Additionally, the video was in Chinese, which further compounded her difficulties. This left her with numerous unanswered questions and uncertainties.

Due to the time difference, Yuliya could not communicate with her children abroad in a timely manner, so upon returning home, Yuliya sought assistance from her neighbors to watch the video, but unfortunately, her neighbors could not operate the QR code either. Instead, they shared their limited knowledge of tracheostomy surgery, drawing from soap opera scenarios. This misconception led Yuliya to believe that patients who had strokes must be bedridden and entirely dependent on others for care.

Meanwhile, the Health Bureau contacted her and requested her to use a social distancing app to monitor Eric's condition. Yuliya informed the contact person that she did not know how to use mobile phone apps or browse the internet.

Analysis of the Problem

After evaluating this case using the eHealth Literacy Lily Model and the eHealth Literacy Scale, it became clear that Yulia faced challenges in 6 components of literacy (Table 3). Additionally, Yuliya's total score on the eHealth Literacy Scale was 8, indicating her limited ability to use eHealth resources effectively.

Table 3. Application of the eHealth Literacy Lily Model for the clinical scenario.

Central types and 6 components	Clinical scenario
Analytic type	
Traditional literacy	Yuliya had limited literacy skills, with her only means of communication being in Taiwanese.
Information literacy	Yuliya often communicated with her children and her neighbors to access health-related information. However, due to the time difference, she could not reach her children promptly. As a result, she discussed tracheostomy surgery with her neighbors, who had limited knowledge, primarily based on soap scenarios.
Media literacy	Yuliya and her neighbors shared their limited knowledge of tracheostomy surgery, drawing from soap opera scenarios. This misconception led Yuliya to believe that patients who had strokes must be bedridden and entirely dependent on others for care.
Context-specific type	
Health literacy	Yuliya does not have any underlying medical conditions and regularly attends government-provided health checkups, following the advice of her neighbors. Her primary sources of health information are her children and neighbors.
Computer literacy	Yuliya did not know how to use mobile phone apps or browse the internet.
Scientific literacy	Yuliya's medical knowledge was primarily derived from her neighbors, soap operas, and her children, though the accuracy of this information could not be guaranteed.

XSL•FO
RenderX

Solution Strategies for This Clinical Scenario

According to the Analytic Type

Overview

For individuals with limited traditional, information, and media literacy skills, it is highly recommended to facilitate communication through family meetings. Family meetings can provide comprehensive opportunities for health professionals, patients, or decision makers to exchange information, explain treatments, and clarify questions [21]. Effective communication skills are crucial in this process, and health professionals should use listening skills, questioning skills, and effective message delivery to enhance understanding and facilitate shared decision-making.

Listening Skills

Health professionals should encourage patients or surrogates to express their thoughts and concerns. By creating a safe space for open dialogue, they can increase the patients' or surrogates' willingness to communicate and provide emotional support. Observing nonverbal cues and summarizing patient or surrogate expectations can help ensure effective communication.

Questioning Skills

The use of open-ended and closed-ended questions is essential to discuss and clarify the patient's preferences and concerns. This approach allows health professionals to gain insight into patients' or surrogates' thoughts and needs.

Sending Messages

Health professionals must provide clear and understandable explanations of treatment options, avoiding the use of medical jargon or technical terms [22]. By offering evidence-based information and using empirical medical research findings to illustrate the advantages and disadvantages of each treatment option, they can empower patients to make informed decisions [23].

In this clinical scenario, the health professionals did not provide an appropriate environment or sufficient time for Yuliya to discuss Eric's medical treatment and address her concerns. Instead, they simply offered a QR code to access a video, assuming it would efficiently convey the necessary information. By using effective communication strategies during family meetings and providing patient-centered explanations and support, health professionals can significantly improve the decision-making process for patients with limited eHealth literacy.

According to the Context-Specific Type

Individuals with health, computer, and scientific literacy deficits might face significant challenges in accessing essential medical information, potentially leading to difficulties and delays in making informed medical decisions. To enhance individuals' health literacy in such cases, the government can use various strategies including television advertisements, hospital promotions, public health clinic campaigns, and even celebrity endorsements to increase public awareness and understanding of health-related topics [2].

In the presented clinical scenario, health professionals could take a proactive approach by providing Yuliya with various prepared materials such as notebooks, pictures, DVDs, videos, QR codes, and apps. They should offer guidance on using these tools effectively and assess Yuliya's understanding of the information provided by eHealth decision resources. The development of decision aids should adhere to specific criteria, including a rigorous and systematic process, the provision of scientific information and data on medical options, assistance in clarifying decision makers' intentions and preferences, transparency regarding potential conflicts of interest, the avoidance of biased language that might lead to the selection of a particular medical option, and the exclusion of medical jargon or technical terms. Decision aids should provide evidence-based information to empower patients or surrogates to make informed choices [24].

Given the current global trend of an aging population, such decision-making aids are suitable for individuals older than 60 years, as they often have a higher demand for health care information. When creating these aids, it is essential to consider 3 critical components, economic costs, health and recovery aspects, and social and environmental support. Despite the increasing availability of technology-based decision-making aids, printed materials in large print, colorful chart cards, or brochures remain the most widely accepted format among patients and their surrogates, especially for more elderly populations [25].

Summaries of This Case

In this clinic scenario, health professionals did not consider Yuliya's eHealth literacy when providing appropriate resources, even though she needed to make immediate decisions. To address this issue in the future, we recommend the following actions: first, health professionals should take the time to explain the conditions for a patient's admission to the ICU thoroughly. This includes providing clear information about the patient's medical condition, the available treatment options, and the potential risks and benefits associated with each choice. Second, when circumstances permit, health professionals should engage in open and transparent communication with patients and their families. They should offer opportunities for patients to ask questions and express their concerns regarding daily urgent medical treatment decisions. Third, it is crucial to provide Yuliya and patients like her with sufficient and understandable information that empowers them to make informed decisions about their health care. This may involve using plain language, visual aids, or written materials to enhance their understanding.

By implementing these recommendations, health professionals could improve the overall patient experience and facilitate more effective shared decision-making, especially for individuals with limited eHealth literacy like Yuliya.

Recommendations for eHealth Issues in ICUs

Based on current development trends, the use of eHealth tools has become an irresistible trend. To address this issue in the future, we recommend the following actions: (1) education and

XSL•FO
RenderX

training: ensure that all health professionals receive appropriate training on how to use eHealth tools; (2) when developing eHealth tools, it is crucial to ensure that the provided content is accurate, aligns with the needs and expectations of surrogates, and enhances usability through a user-friendly interface. Additionally, having real time technical support for clarification is essential to prevent any disruptions for surrogates during usage; and (3) conduct regular assessments of the eHealth tools and establish an evaluation framework that incorporates data and feedback for continuous updates and adjustments based on the surrogates' needs.

In summary, effectively integrating eHealth tools can improve the quality and efficiency of ICU care. However, it is imperative to ensure that these tools not only meet medical standards but also consider patient privacy and data security. Therefore, strategies such as education, interoperability, security, and user-friendliness are key to successfully meeting surrogates' expectations with eHealth tools.

Conclusions

In conclusion, the challenges associated with using eHealth decision resources to assist surrogates in ICU decision-making are multifaceted and require careful consideration. The high mortality rate in ICUs, when coupled with critical conditions, compels swift decision-making by surrogates, and the increasing popularity of eHealth tools adds complexity to this process.

The application of eHealth tools brings both opportunities and challenges. While these tools have the potential to disseminate important medical information and improve health care knowledge, they also present economic, informational, and inclusivity challenges. Economic disparities, information inaccuracies, and the risk of excluding specific groups must be addressed to ensure equitable access and benefit for all.

The eHealth Literacy Lily Model and the eHealth Literacy Scale are valuable tools for assessing individuals' capability and literacy levels in using eHealth resources. To improve decision-making for individuals with limited eHealth literacy, both analytic and context-specific approaches are recommended. Effective communication through family meetings via using listening and questioning skills and providing patient-centered explanations can all enhance understanding and facilitate shared decision-making. Additionally, proactive measures such as using various materials, realia, and decision-making aids, can improve literacy in the health, computer, and scientific dimensions.

In summary, recognizing the time constraints and high-stakes nature of ICU decision-making, it is crucial to strike a balance between leveraging eHealth tools for efficiency and ensuring inclusivity. Health professionals should adopt a patient-centered approach, considering the diverse literacy levels and needs of individuals, especially those with limited eHealth literacy. Tailoring communication strategies, providing diverse materials, and assessing eHealth literacy levels using tools like the eHealth Literacy Scale are essential steps toward enhancing the effectiveness of eHealth decision resources in the ICU setting.

Conflicts of Interest

None declared.

References

1. Critical care statistics. Society of Critical Care Medicine. 2021. URL: https://www.sccm.org/Clinical-Resources/Guidelines/Guidelines [accessed 2024-02-21]
2. Current situation of medical institutions and statistics of hospital medical service volume e-book. In: Ministry of Health and Welfare. Taiwan: Ministry of Health and Welfare; 2020.
3. Abe T, Madotto F, Pham T, Nagata I, Uchida M, Tamiya N, et al. Epidemiology and patterns of tracheostomy practice in patients with acute respiratory distress syndrome in ICUs across 50 countries. Crit Care 2018;22(1):195 [FREE Full text] [doi: 10.1186/s13054-018-2126-6] [Medline: 30115127]
4. Miller JJ, Morris P, Files DC, Gower E, Young M. Decision conflict and regret among surrogate decision makers in the medical intensive care unit. J Crit Care 2016;32:79-84. [doi: 10.1016/j.jcrc.2015.11.023] [Medline: 26810482]
5. Cai X, Robinson J, Muehlschlegel S, White DB, Holloway RG, Sheth KN, et al. Patient preferences and surrogate decision making in neuroscience intensive care units. Neurocrit Care 2015;23(1):131-141 [FREE Full text] [doi: 10.1007/s12028-015-0149-2] [Medline: 25990137]
6. Lilley EJ, Morris MA, Sadovnikoff N, Luxford JM, Changoor NR, Bystricky A, et al. "Taking over somebody's life": experiences of surrogate decision-makers in the surgical intensive care unit. Surgery 2017;162(2):453-460. [doi: 10.1016/j.surg.2017.03.011] [Medline: 28549520]
7. Wu F, Zhuang Y, Chen X, Wen H, Tao W, Lao Y, et al. Decision-making among the substitute decision makers in intensive care units: an investigation of decision control preferences and decisional conflicts. J Adv Nurs 2020;76(9):2323-2335. [doi: 10.1111/jan.14451] [Medline: 32538477]
8. Wynn R, Gabarron E, Johnsen JAK, Traver V. Special issue on e-health services. Int J Environ Res Public Health 2020;17(8):2885 [FREE Full text] [doi: 10.3390/ijerph17082885] [Medline: 32331273]
9. Terrasse M, Gorin M, Sisti D. Social media, e-health, and medical ethics. Hastings Cent Rep 2019;49(1):24-33. [doi: 10.1002/hast.975] [Medline: 30790306]

XSL•FO
RenderX

10. Arvind S, Martin C. What is e-health? Eur Soc Cardiol 2020;18(24):10.
11. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):e20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
12. Azeez NA, Van der Vyver C. Security and privacy issues in e-health cloud-based system: a comprehensive content analysis. Egypt Inform J 2019;20(2):97-108 [FREE Full text] [doi: 10.1016/j.eij.2018.12.001]
13. Ruppert L, Køster B, Siegert A, Cop C, Boyers L, Karimkhani C, et al. YouTube as a source of health information: analysis of sun protection and skin cancer prevention related issues. Dermatol Online J 2017;23(1):1 [FREE Full text] [doi: 10.5070/d3231033669] [Medline: 28329465]
14. Swartzendruber A, Newton-Levinson A, Feuchs AE, Phillips AL, Hickey J, Steiner RJ. Sexual and reproductive health services and related health information on pregnancy resource center websites: a statewide content analysis. Womens Health Issues 2018;28(1):14-20. [doi: 10.1016/j.whi.2017.10.007] [Medline: 29158038]
15. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
16. Posititi D, Cilliers L. An investigation of the level of e-health literacy in South Africa. ICICIS 2019;12:261-271 [FREE Full text]
17. Stellefson M, Hanik B, Chaney B, Chaney D, Tennant B, Chavarria EA. eHealth literacy among college students: a systematic review with implications for eHealth education. J Med Internet Res 2011;13(4):e102 [FREE Full text] [doi: 10.2196/jmir.1703] [Medline: 22155629]
18. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
19. Koo M, Norman CD, Hsiao-Mei C. Psychometric evaluation of a Chinese version of the eHealth Literacy Scale (eHEALS) in school age children. Int Electron J Health Educ 2012;15:29-36 [FREE Full text]
20. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
21. Scheunemann LP, Ernecoff NC, Buddadhumaruk P, Carson SS, Hough CL, Curtis JR, et al. Clinician-family communication about patients' values and preferences in intensive care units. JAMA Intern Med 2019;179(5):676-684 [FREE Full text] [doi: 10.1001/jamainternmed.2019.0027] [Medline: 30933293]
22. Communication skills when providing decision support. The Ottawa Hospital. 2022. URL: https://decisionaid.ohri.ca/docs/decision_coaching_communication.pdf [accessed 2024-02-21]
23. The SHARE approach—essential steps of shared decisionmaking: expanded reference guide with sample conversation starters. Agency for Healthcare Research and Quality. 2020. URL: https://www.ahrq.gov/health-literacy/professional-training/shared-decision/tool/resource-2.html [accessed 2024-02-21]
24. Liu RW, Ceng BL, Ckang JN. Principles of evidence-based decision aid development. J Healthc Qual 2017;11(4):25-30.
25. Baxter K, Mann R, Birks Y, Overton L. A scoping review of evidence on the use and effectiveness of decision aids in adult social care. J Long Term Care 2021:100-113 [FREE Full text] [doi: 10.31389/jltc.56]

Abbreviations

CT: computed tomography
GCS: Glasgow Coma Scale
ICT: information and communication technology
ICU: intensive care unit
WHO: World Health Organization

Edited by A Mavragani; submitted 08.03.23; peer-reviewed by A Van Dongen, Y Chu, C Hudak; comments to author 21.06.23; revised version received 23.11.23; accepted 29.01.24; published 01.04.24.

Please cite as:
Sun WN, Kao CY
The Challenges in Using eHealth Decision Resources for Surrogate Decision-Making in the Intensive Care Unit
J Med Internet Res 2024;26:e47017
URL: https://www.jmir.org/2024/1/e47017
doi:10.2196/47017
PMID:38557504

©Wan-Na Sun, Chi-Yin Kao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A Perspective on Crowdsourcing and Human-in-the-Loop Workflows in Precision Health

Peter Washington[1], PhD

Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI, United States

Corresponding Author:
Peter Washington, PhD
Information and Computer Sciences
University of Hawaii at Manoa
1680 East-West Road
Honolulu, HI, 96822
United States
Email: pyw@hawaii.edu

Abstract

Modern machine learning approaches have led to performant diagnostic models for a variety of health conditions. Several machine learning approaches, such as decision trees and deep neural networks, can, in principle, approximate any function. However, this power can be considered to be both a gift and a curse, as the propensity toward overfitting is magnified when the input data are heterogeneous and high dimensional and the output class is highly nonlinear. This issue can especially plague diagnostic systems that predict behavioral and psychiatric conditions that are diagnosed with subjective criteria. An emerging solution to this issue is crowdsourcing, where crowd workers are paid to annotate complex behavioral features in return for monetary compensation or a gamified experience. These labels can then be used to derive a diagnosis, either directly or by using the labels as inputs to a diagnostic machine learning model. This viewpoint describes existing work in this emerging field and discusses ongoing challenges and opportunities with crowd-powered diagnostic systems, a nascent field of study. With the correct considerations, the addition of crowdsourcing to human-in-the-loop machine learning workflows for the prediction of complex and nuanced health conditions can accelerate screening, diagnostics, and ultimately access to care.

(J Med Internet Res 2024;26:e51138) doi:10.2196/51138

KEYWORDS

crowdsourcing; digital medicine; human-in-the-loop; human in the loop; human-AI collaboration; machine learning; precision health; artificial intelligence; AI

Introduction

Crowdsourcing, a term first coined in 2006 [1], is the use of distributed human workers to accomplish a central task. Crowdsourcing exploits the “power of the crowd” to achieve goals that are only feasible with a distributed group of humans collaborating, either explicitly or implicitly, toward a common goal. Crowdsourcing has often been applied to public health surveillance [2], such as for tracking epidemics [3,4], quantifying tobacco use [5], monitoring water quality [6], tracking misinformation [7], and understanding the black-market price of prescription opioids [8]. In the context of health care, crowdsourcing is most often used for public health, a domain that can clearly benefit from scalable and distributed assessments of health status. Although sampling bias can be an issue in epidemiological uses of crowdsourcing [9], approaches that account for these issues have performed quite robustly.

A smaller but potentially transformative effort to apply crowdsourcing to precision health rather than population health has recently emerged. In precision health contexts, the goal is to provide a diagnosis using information labeled by crowd workers. There are several variations to this basic setup. Crowdsourcing workflows for diagnostics can diverge with respect to the underlying task, worker motivation strategies, worker training, worker filtering, and privacy requirements.

Here, I describe the existing research in the relatively small and early but growing field of crowdsourcing for precision health. I then discuss ongoing challenges and corresponding opportunities that must be addressed as this field matures.

XSL•FO
RenderX

Existing Examples of Crowdsourcing in and Adjacent to Health Care

There are relatively few examples of crowdsourcing in precision health. The vast successes of machine learning for health [10-15] and the human labor costs required for crowdsourcing make purely automated approaches more appealing when they are possible and feasible. However, the crowdsourcing approaches that have been tested tend to perform well for prediction tasks that are beyond the scope of current automated approaches, especially in psychiatry and the behavioral sciences.

I want to begin by highlighting successes in science, as they can often be applied to health and have started to lead to improvements in diagnostics. Framing crowdsourcing tasks as "citizen science" opportunities can be an effective incentive mechanism [16]. Oftentimes, these projects are "gamified." Gamification refers to the incorporation of engaging elements into traditionally burdensome workflows, and in particular game-like affordances, to foster increased participation. A combination of large crowd sizes, worker training procedures, and easy identification tasks have led to previous success in the existing gamified citizen science experiments applied to precision health. For example, in a study involving nearly 100,000 crowd workers who scored images on a citizen science platform, cancer was correctly identified with an area under the receive operating characteristic of around 95% [17]. In the BioGames app, users who performed with greater than 99% accuracy in a training tutorial were invited to diagnose malaria [18,19]. It was discovered that with a large crowd size, the aggregated diagnostic accuracy of nonexpert crowd workers approached that of experts [20]. Another citizen science malaria diagnosis application, MalariaSpot, resulted in 99% accuracy in the diagnosis of malaria from blood films [21]. If the annotation task is relatively simple and nonexperts can be trained with minimal onboarding efforts, then citizen science can be an effective and affordable approach.

"Gamified" crowdsourcing for citizen science has also been successful without explicitly requiring workers to undergo a formal training process. Foldit [22-25] and EteRNA [26-31] are 2 games where players with no biology or chemistry background can explore the design space of protein and RNA folding, respectively. These are both NP-hard (ie, computationally complex) problems, and human players in aggregate have designed solutions that outcompete state-of-the-art computational approaches. These solutions have been used to solve health challenges, such as finding a gene signature for active tuberculosis, which can potentially be used in tuberculosis diagnostics [32]. Other gamified experiences have been used to build training libraries for complex classification tasks in precision psychiatry. Notably, GuessWhat is a mobile charades game played between children with autism and their parents [33,34]. While the game provides therapeutic benefits to the child with autism [35], the game simultaneously curates automatic labels of behaviors related to autism through the metadata associated with gameplay [36,37]. These automatically annotated video data have been used to develop state-of-the-art computer vision models for behaviors related to the diagnosis of autism, such as facial expression evocation [38-41], eye gaze [42], atypical prosody [43], and atypical body movements [44,45].

An alternative incentive mechanism is paid crowdsourcing. The most popular paid crowdsourcing platform, by far, is Amazon Mechanical Turk (MTurk) [46]. While paid crowdsourcing specifically for precision health is a relatively nascent field, the general study of paid crowdsourcing (particularly on MTurk) is quite mature. Studies have explored worker quality management [47], understanding crowd worker demographics [48], the generation of annotations for use in the training of machine learning models [49-53], the rights of crowd workers [54-56], and understanding crowd worker communities and economics [57-59]. Preliminary studies of paid crowdsourcing have yielded mixed success. Around 81% of images were correctly classified on MTurk in a study involving the grading of diabetic retinopathy from images, with workers failing to correctly indicate the level of severity [60]. In a separate binary labeling task for glaucomatous optic neuropathy, workers achieved sensitivity in the 80s but reached a specificity below 50% [61].

In a broader classification task of various medical conditions, workers consistently labeled the "easy" cases while struggling to correctly label and even refusing to label more complicated and nuanced tasks [62]. Clearly, there is a need for extensive innovations to the traditional paid crowdsourcing workflow to translate this methodology to precision health.

I have extensively investigated the utility of paid crowdsourcing for the diagnosis of autism from unstructured home videos, achieving relatively high diagnostic performance [63-66]. In these experiments, untrained annotators watched short videos depicting children with and without autism and answered questions about the behaviors depicted within the videos. These annotations were provided as input into previously developed machine learning models, achieving binary test performance in the 90s across performance metrics due to the reduction of the complex feature space (unstructured videos) into a low-dimensional representation (vectors of a few categorical ordinal values). This pipeline combining crowdsourcing and machine learning can possibly be extended to other diagnostic domains in psychiatry where the input feature space is complex, heterogeneous, and subjective.

Ongoing Challenges and Corresponding Opportunities

Since crowdsourcing for precision health care is an emerging field of study, numerous challenges must be solved for clinical translation to develop. In the proceeding sections, I highlight several areas that are pressing for the field and for which preliminary work has been published.

Worker Identification and Training

While traditional crowdsourcing can work with minimal to no worker training, complex annotation tasks require the identification of qualified workers. I have found that worker identification can occur through the quantification of their

XSL•FO
RenderX

performance on test tasks [67,68] and training promising workers [66]. Such crowd filtration paradigms will require domain-specific procedures. There is ample room to develop new crowdsourcing systems that inherently support natural worker identification and training procedures for crowdsourcing workflows that require well-designed training processes.

Worker Retention

Once proficient workers are identified, continually engaging and retaining these workers is critical. I have found that workers who are repeatedly encouraged by a human (or human-like chatbot) and treated as members of a broader research team tend to enjoy paid work and even ask for more tasks after the completion of the study [69]. Thus, it is possible that the guarantee of job security can lead to long-term worker retention. However, worker retention in unpaid settings that rely on intrinsic motivation will require additional innovations. For example, there exists an opportunity to explore the creation of crowd worker communities to provide a means of intrinsic motivation leading to worker retention.

Task Assignment

Certain workers perform exceptionally well on a subset of tasks while underperforming on other assigned tasks [70,71]. There is an opportunity to develop algorithmic innovations involving the effective and optimal assignment of workers to subtasks in a dynamic manner. Reinforcement learning could be a promising approach but has yet to be explored in such scenarios.

Privacy of Human Participants

Data in psychiatry and behavioral sciences are particularly sensitive. Ensuring that sensitive health information is handled appropriately and that workers' privacy is maintained is essential from an ethical perspective. There are 2 general families of approaches to achieving privacy in crowd-powered health care. First, the data can be modified to obscure sensitive information without removing information required for a diagnosis. I have explored privacy-preserving alterations to video data that obfuscate the identity of participants while maximizing the capacity for workers to annotate behaviors of interest [70,71]. For example, in the case of video analytics on bodily movements, the face can be tracked and blurred, or the body can be converted to a stick figure using a pose-based computer vision library. Sometimes, however, it is impossible to modify the data without severely degrading the diagnostic performance. Therefore, the second family of approaches involves carefully vetting crowd workers, training them, and onboarding them into a secure computing environment. In my previous experiences with this process [40], I discovered that crowd workers were enthusiastic about the prospect of the "job security" that is implied from the thorough vetting procedure and were, therefore, willing to complete extra privacy and security training (in our case, Research, Ethics, Compliance, and Safety training). There is ample room to expand upon these methods and to develop new paradigms and systems for crowdsourcing involving identifiable and protected health information.

Ensuring Reliability and Reproducibility

An intrinsic challenge when incorporating human workers into precision health workflows is the variability in human responses, both within workers and between workers. I have found that while most crowd workers are inconsistent in their annotation patterns, there are workers who provide consistently sensitive and specific annotations across a wide spectrum of data points [67]. It is therefore critical to measure both internal consistency and consistency against a gold standard when recruiting crowd workers for precision health care workflows.

Handling Financial Constraints

The crowdsourcing method with the lowest setup barriers is paid crowdsourcing. In such scenarios, financial constraints can limit the scalability of crowdsourcing workflows. One approach is to migrate from a paid system to a gamified system or another means of providing intrinsic motivation to crowd workers. However, achieving critical mass for large-scale pipelines is likely unattainable for such unpaid solutions. Paid crowd workers who consistently perform well could be recruited as full-time or long-term part-time employees for companies and organizations providing crowd-powered services. Integrating such workflows into a Food and Drug Administration (FDA)–approved process can be challenging, but it is worth exploring if it turns out that crowd-powered solutions for digital psychiatry continue to remain superior to pure-artificial intelligence (AI) approaches in the coming years.

Translation Outside of Research Contexts

While pure machine learning approaches for precision health are beginning to translate to clinical settings through formal FDA approval procedures, the prospect of translating human-in-the-loop methods that integrate crowd workers rather than expert clinicians is daunting, especially in light of the challenges mentioned above. However, if such approaches lead to clinical-grade performance for certain conditions that are challenging to diagnose using machine learning alone, then the extra implementation and regulatory effort required to migrate these methods into production-level workflows are likely to be warranted.

Conclusion

While machine learning for health has enabled and will continue to enable more efficient, precise, and scalable diagnostics for a variety of conditions, such models are unlikely to generalize to more difficult scenarios such as psychiatry and the behavioral sciences, which require the ability to identify complex and nuanced social human behavior. Crowd-powered human-in-the-loop workflows have the potential to mitigate some of these current limitations while still offering a high degree of automation. I invite researchers in the fields of digital phenotyping [72-76], mobile sensing [77-83], affective computing [84-90], and related subjects to consider integrating crowdsourcing and human-in-the-loop approaches into their methods when pure-AI leads to suboptimal performance.

XSL•FO
RenderX

Acknowledgments

This project is funded by the NIH Director's New Innovator Award (DP2) from the National Institutes of Health (award DP2-EB035858).

Conflicts of Interest

None declared.

References

1. Howe J. The rise of crowdsourcing. Wired. 2006 Jun 01. URL: https://sistemas-humano-computacionais.wdfiles.com/local--files/capitulo%3Aredes-sociais/Howe_The_Rise_of_Crowdsourcing.pdf [accessed 2024-03-29]
2. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med 2014;46(2):179-187. [doi: 10.1016/j.amepre.2013.10.016] [Medline: 24439353]
3. Leung GM, Leung K. Crowdsourcing data to mitigate epidemics. Lancet Digit Health 2020;2(4):e156-e157 [FREE Full text] [doi: 10.1016/S2589-7500(20)30055-8] [Medline: 32296776]
4. Stockham N, Washington P, Chrisman B, Paskov K, Jung JY, Wall DP. Causal modeling to mitigate selection bias and unmeasured confounding in internet-based epidemiology of COVID-19: model development and validation. JMIR Public Health Surveill 2022;8(7):e31306 [FREE Full text] [doi: 10.2196/31306] [Medline: 35605128]
5. Kraemer JD, Strasser AA, Lindblom EN, Niaura RS, Mays D. Crowdsourced data collection for public health: a comparison with nationally representative, population tobacco use data. Prev Med 2017;102:93-99 [FREE Full text] [doi: 10.1016/j.ypmed.2017.07.006] [Medline: 28694063]
6. Jakositz S, Pillsbury L, Greenwood S, Fahnestock M, McGreavy B, Bryce J, et al. Protection through participation: crowdsourced tap water quality monitoring for enhanced public health. Water Res 2020;169:115209. [doi: 10.1016/j.watres.2019.115209] [Medline: 31669904]
7. Ghenai A, Mejova Y. Catching Zika fever: application of crowdsourcing and machine learning for tracking health misinformation on Twitter. arXiv Preprint posted online Jul 12, 2017. [doi: 10.48550/arXiv.1707.03778]
8. Dasgupta N, Freifeld C, Brownstein JS, Menone CM, Surratt HL, Poppish L, et al. Crowdsourcing black market prices for prescription opioids. J Med Internet Res 2013;15(8):e178 [FREE Full text] [doi: 10.2196/jmir.2810] [Medline: 23956042]
9. Wazny K. "Crowdsourcing" ten years in: a review. J Glob Health 2017;7(2):020602 [FREE Full text] [doi: 10.7189/jogh.07.020602] [Medline: 29302322]
10. Chen PHC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater 2019;18(5):410-414. [doi: 10.1038/s41563-019-0345-0] [Medline: 31000806]
11. Dua S, Acharya UR, Dua P, editors. Machine Learning in Healthcare Informatics, Volume 56. Berlin, Heidelberg: Springer; 2014.
12. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
13. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191-200 [FREE Full text] [Medline: 32477638]
14. Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: a review. 2018 Presented at: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA); March 29-31, 2018; Coimbatore, India p. 910-914. [doi: 10.1109/iceca.2018.8474918]
15. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
16. Das R, Keep B, Washington P, Riedel-Kruse IH. Scientific discovery games for biomedical research. Annu Rev Biomed Data Sci 2019;2(1):253-279 [FREE Full text] [doi: 10.1146/annurev-biodatasci-072018-021139] [Medline: 34308269]
17. Dos Reis FJC, Lynn S, Ali HR, Eccles D, Hanby A, Provenzano E, et al. Crowdsourcing the general public for large scale molecular pathology studies in cancer. EBioMedicine 2015;2(7):681-689 [FREE Full text] [doi: 10.1016/j.ebiom.2015.05.009] [Medline: 26288840]
18. Mavandadi S, Dimitrov S, Feng S, Yu F, Sikora U, Yaglidere O, et al. Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. PLoS One 2012;7(5):e37245 [FREE Full text] [doi: 10.1371/journal.pone.0037245] [Medline: 22606353]
19. Ozcan A. Educational games for malaria diagnosis. Sci Transl Med 2014;6(233):233ed9 [FREE Full text] [doi: 10.1126/scitranslmed.3009172] [Medline: 24760185]
20. Feng S, Woo M, Chandramouli K, Ozcan A. A game-based platform for crowd-sourcing biomedical image diagnosis and standardized remote training and education of diagnosticians. In: Optics and Biophotonics in Low-Resource Settings, Volume 9314. 2015 Presented at: SPIE BIOS; February 7-12, 2015; San Francisco, CA. [doi: 10.1117/12.2077884]
21. Luengo-Oroz MA, Arranz A, Frean J. Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. J Med Internet Res 2012;14(6):e167 [FREE Full text] [doi: 10.2196/jmir.2338] [Medline: 23196001]

XSL•FO
RenderX

22. Cooper S, Khatib F, Makedon I, Lu H, Barbero J, Baker D, et al. Analysis of social gameplay macros in the Foldit cookbook. 2011 Presented at: FDG '11: Proceedings of the 6th International Conference on Foundations of Digital Games; June 29, 2011 - July 1, 2011; Bordeaux, France p. 9-14. [doi: 10.1145/2159365.2159367]
23. Curtis V. Motivation to participate in an online citizen science game: a study of Foldit. Sci Commun 2015;37(6):723-746. [doi: 10.1177/1075547015609322]
24. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, Shen BW, et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotechnol 2012;30(2):190-192 [FREE Full text] [doi: 10.1038/nbt.2109] [Medline: 22267011]
25. Kleffner R, Flatten J, Leaver-Fay A, Baker D, Siegel JB, Khatib F, et al. Foldit standalone: a video game-derived protein structure manipulation interface using Rosetta. Bioinformatics 2017;33(17):2765-2767 [FREE Full text] [doi: 10.1093/bioinformatics/btx283] [Medline: 28481970]
26. Andreasson JOL, Gotrik MR, Wu MJ, Wayment-Steele HK, Kladwang W, Portela F, et al. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches. Proc Natl Acad Sci U S A 2022;119(18):e2112979119 [FREE Full text] [doi: 10.1073/pnas.2112979119] [Medline: 35471911]
27. Krüger A, Watkins AM, Wellington-Oguri R, Romano J, Kofman C, DeFoe A, et al. Community science designed ribosomes with beneficial phenotypes. Nat Commun 2023;14(1):961 [FREE Full text] [doi: 10.1038/s41467-023-35827-3] [Medline: 36810740]
28. Lee J, Kladwang W, Lee M, Cantu D, Azizyan M, Kim H, et al. RNA design rules from a massive open laboratory. Proc Natl Acad Sci U S A 2014;111(6):2122-2127 [FREE Full text] [doi: 10.1073/pnas.1313039111] [Medline: 24469816]
29. Shi J, Das R, Pande VS. SentRNA: improving computational RNA design by incorporating a prior of human design strategies. arXiv Preprint posted online Mar 06, 2019. [doi: 10.48550/arXiv.1803.03146]
30. Treuille A, Das R. Scientific rigor through videogames. Trends Biochem Sci 2014;39(11):507-509 [FREE Full text] [doi: 10.1016/j.tibs.2014.08.005] [Medline: 25300714]
31. Wayment-Steele HK, Kladwang W, Strom AI, Lee J, Treuille A, Becka A, et al. RNA secondary structure packages evaluated and improved by high-throughput experiments. Nat Methods 2022;19(10):1234-1242 [FREE Full text] [doi: 10.1038/s41592-022-01605-0] [Medline: 36192461]
32. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir Med 2016;4(3):213-224 [FREE Full text] [doi: 10.1016/S2213-2600(16)00048-5] [Medline: 26907218]
33. Kalantarian H, Jedoui K, Washington P, Wall DP. A mobile game for automatic emotion-labeling of images. IEEE Trans Games 2020;12(2):213-218 [FREE Full text] [doi: 10.1109/tg.2018.2877325] [Medline: 32551410]
34. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall D. A gamified mobile system for crowdsourcing video for autism research. 2018 Presented at: 2018 IEEE International Conference on Healthcare Informatics (ICHI); June 4-7, 2018; New York, NY p. 350-352. [doi: 10.1109/ICHI.2018.00052]
35. Penev Y, Dunlap K, Husic A, Hou C, Washington P, Leblanc E, et al. A mobile game platform for improving social communication in children with autism: a feasibility study. Appl Clin Inform 2021;12(5):1030-1040 [FREE Full text] [doi: 10.1055/s-0041-1736626] [Medline: 34788890]
36. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall DP. Guess what? Towards understanding autism from structured video using facial affect. J Healthc Inform Res 2019;3(1):43-66 [FREE Full text] [doi: 10.1007/s41666-018-0034-9] [Medline: 33313475]
37. Kalantarian H, Jedoui K, Washington P, Tariq Q, Dunlap K, Schwartz J, et al. Labeling images with facial emotion and the potential for pediatric healthcare. Artif Intell Med 2019;98:77-86 [FREE Full text] [doi: 10.1016/j.artmed.2019.06.004] [Medline: 31521254]
38. Hou C, Kalantarian H, Washington P, Dunlap K, Wall DP. Leveraging video data from a digital smartphone autism therapy to train an emotion detection classifier. medRxiv Preprint posted online Aug 01, 2021. [doi: 10.1101/2021.07.28.21260646]
39. Kalantarian H, Jedoui K, Dunlap K, Schwartz J, Washington P, Husic A, et al. The performance of emotion classifiers for children with parent-reported autism: quantitative feasibility study. JMIR Ment Health 2020;7(4):e13174 [FREE Full text] [doi: 10.2196/13174] [Medline: 32234701]
40. Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, et al. Improved digital therapy for developmental pediatrics using domain-specific artificial intelligence: machine learning study. JMIR Pediatr Parent 2022;5(2):e26760 [FREE Full text] [doi: 10.2196/26760] [Medline: 35394438]
41. Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, et al. Training an emotion detection classifier using frames from a mobile therapeutic game for children with developmental disorders. arXiv Preprint posted online Dec 16, 2020 . [doi: 10.48550/arXiv.2012.08678]
42. Varma M, Washington P, Chrisman B, Kline A, Leblanc E, Paskov K, et al. Identification of social engagement indicators associated with autism spectrum disorder using a game-based mobile app: comparative study of gaze fixation and visual scanning methods. J Med Internet Res 2022;24(2):e31830 [FREE Full text] [doi: 10.2196/31830] [Medline: 35166683]

XSL•FO
RenderX

43. Chi NA, Washington P, Kline A, Husic A, Hou C, He C, et al. Classifying autism from crowdsourced semistructured speech recordings: machine learning model comparison study. JMIR Pediatr Parent 2022;5(2):e35406 [FREE Full text] [doi: 10.2196/35406] [Medline: 35436234]
44. Lakkapragada A, Kline A, Mutlu OC, Paskov K, Chrisman B, Stockham N, et al. The classification of abnormal hand movement to aid in autism detection: machine learning study. JMIR Biomed Eng 2022;7(1):e33771 [FREE Full text] [doi: 10.2196/33771]
45. Washington P, Kline A, Mutlu OC, Leblanc E, Hou C, Stockham N, et al. Activity recognition with moving cameras and few training examples: applications for detection of autism-related headbanging. 2021 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-7. [doi: 10.1145/3411763.3451701]
46. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgm Decis Mak 2010;5(5):411-419 [FREE Full text] [doi: 10.1017/s1930297500002205]
47. Ipeirotis PG, Provost F, Wang J. Quality management on Amazon Mechanical Turk. 2010 Presented at: HCOMP '10: Proceedings of the ACM SIGKDD Workshop on Human Computation; July 25, 2010; Washington DC p. 64-67. [doi: 10.1145/1837885.1837906]
48. Lee YJ, Arida JA, Donovan HS. The application of crowdsourcing approaches to cancer research: a systematic review. Cancer Med 2017;6(11):2595-2605 [FREE Full text] [doi: 10.1002/cam4.1165] [Medline: 28960834]
49. Ross J, Zaldivar A, Irani L, Tomlinson B. Who are the turkers? Worker demographics in Amazon Mechanical Turk. ResearchGate. 2009. URL: https://www.researchgate.net/publication/268427703_Who_are_the_Turkers_Worker_Demographics_in_Amazon_Mechanical_Turk [accessed 2024-03-22]
50. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-252. [doi: 10.1007/s11263-015-0816-y]
51. Sheng VS, Zhang J. Machine learning with crowdsourcing: a brief summary of the past research and future directions. Proc AAAI Conf Artif Intell 2019;33(01):9837-9843. [doi: 10.1609/aaai.v33i01.33019837]
52. Sorokin A, Forsyth D. Utility data annotation with Amazon Mechanical Turk. 2008 Presented at: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; June 23-28, 2008; Anchorage, AK p. 1-8. [doi: 10.1109/cvprw.2008.4562953]
53. Xintong G, Hongzhi W, Song Y, Hong G. Brief survey of crowdsourcing for data mining. Expert Syst Appl 2014;41(17):7987-7994. [doi: 10.1016/j.eswa.2014.06.044]
54. Irani LC, Silberman MS. Turkopticon: interrupting worker invisibility in Amazon Mechanical Turk. 2013 Presented at: CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 27, 2013 - May 2, 2013; Paris, France p. 611-620. [doi: 10.1145/2470654.2470742]
55. Irani L, Silberman MS. From critical design to critical infrastructure: lessons from turkopticon. Interactions 2014;21(4):32-35. [doi: 10.1145/2627392]
56. Kummerfeld JK. Quantifying and avoiding unfair qualification labour in crowdsourcing. arXiv Preprint posted online May 26, 2021. [doi: 10.48550/arXiv.2105.12762]
57. Hansson K, Ludwig T. Crowd dynamics: conflicts, contradictions, and community in crowdsourcing. Comput Support Coop Work 2019;28:791-794 [FREE Full text] [doi: 10.1007/s10606-018-9343-z]
58. Shen XL, Lee MKO, Cheung CMK. Exploring online social behavior in crowdsourcing communities: a relationship management perspective. Comput Human Behav 2014;40:144-151. [doi: 10.1016/j.chb.2014.08.006]
59. Wu W, Gong X. Motivation and sustained participation in the online crowdsourcing community: the moderating role of community commitment. Internet Res 2021;31(1):287-314. [doi: 10.1108/intr-01-2020-0008]
60. Brady CJ, Villanti AC, Pearson JL, Kirchner TR, Gupta OP, Shah C. Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. J Med Internet Res 2014 Oct 30;16(10):e233 [FREE Full text] [doi: 10.2196/jmir.3807] [Medline: 25356929]
61. Mitry D, Peto T, Hayat S, Blows P, Morgan J, Khaw KT, et al. Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. PLoS One 2015;10(2):e0117401 [FREE Full text] [doi: 10.1371/journal.pone.0117401] [Medline: 25692287]
62. Cheng J, Manoharan M, Zhang Y, Lease M. Is there a doctor in the crowd? Diagnosis needed! (For less than $5). iConference 2015 Proceedings. 2015. URL: https://www.ideals.illinois.edu/items/73844 [accessed 2024-03-22]
63. Leblanc E, Washington P, Varma M, Dunlap K, Penev Y, Kline A, et al. Feature replacement methods enable reliable home video analysis for machine learning detection of autism. Sci Rep 2020;10(1):21245 [FREE Full text] [doi: 10.1038/s41598-020-76874-w] [Medline: 33277527]
64. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: a development and prospective validation study. PLoS Med 2018;15(11):e1002705 [FREE Full text] [doi: 10.1371/journal.pmed.1002705] [Medline: 30481180]
65. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, et al. Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: development and validation study. J Med Internet Res 2019;21(4):e13822 [FREE Full text] [doi: 10.2196/13822] [Medline: 31017583]

XSL•FO
RenderX

66. Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, et al. Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Sci Rep 2021;11(1):7620 [FREE Full text] [doi: 10.1038/s41598-021-87059-4] [Medline: 33828118]
67. Washington P, Leblanc E, Dunlap K, Penev Y, Kline A, Paskov K, et al. Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for video-based autism feature recognition. J Pers Med 2020;10(3):86 [FREE Full text] [doi: 10.3390/jpm10030086] [Medline: 32823538]
68. Washington P, Leblanc E, Dunlap K, Penev Y, Varma M, Jung JY, et al. Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. Pac Symp Biocomput 2021;26:14-25 [FREE Full text] [Medline: 33691000]
69. Washington P, Kalantarian H, Tariq Q, Schwartz J, Dunlap K, Chrisman B, et al. Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. J Med Internet Res 2019;21(5):e13668 [FREE Full text] [doi: 10.2196/13668] [Medline: 31124463]
70. Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, et al. Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Sci Rep 2021;11(1):7620 [FREE Full text] [doi: 10.1038/s41598-021-87059-4] [Medline: 33828118]
71. Washington P, Chrisman B, Leblanc E, Dunlap K, Kline A, Mutlu C, et al. Crowd annotations can approximate clinical autism impressions from short home videos with privacy protections. Intell Based Med 2022;6:100056 [FREE Full text] [doi: 10.1016/j.ibmed.2022.100056] [Medline: 35634270]
72. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
73. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2:88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
74. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021;46(1):45-54 [FREE Full text] [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]
75. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016;41(7):1691-1696 [FREE Full text] [doi: 10.1038/npp.2016.7] [Medline: 26818126]
76. Washington P, Mutlu CO, Kline A, Paskov K, Stockham NT, Chrisman B, et al. Challenges and opportunities for machine learning classification of behavior and mental state from images. arXiv Preprint posted online Jan 26, 2022. [doi: 10.48550/arXiv.2201.11197]
77. Baumeister H, Montag C, editors. Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics. Cham, Switzerland: Springer International Publishing; 2019.
78. Laport-López F, Serrano E, Bajo J, Campbell AT. A review of mobile sensing systems, applications, and opportunities. Knowl Inf Syst 2020;62(1):145-174. [doi: 10.1007/s10115-019-01346-1]
79. Macias E, Suarez A, Lloret J. Mobile sensing systems. Sensors (Basel) 2013;13(12):17292-17321 [FREE Full text] [doi: 10.3390/s131217292] [Medline: 24351637]
80. Nazir S, Ali Y, Ullah N, García-Magariño I. Internet of things for healthcare using effects of mobile computing: a systematic literature review. Wirel Commun Mob Comput 2019;2019:1-20 [FREE Full text] [doi: 10.1155/2019/5931315]
81. Silva BMC, Rodrigues JJPC, de la Torre Díez I, López-Coronado M, Saleem K. Mobile-health: a review of current state in 2015. J Biomed Inform 2015;56:265-272 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.003] [Medline: 26071682]
82. Sim I. Mobile devices and health. N Engl J Med 2019;381(10):956-968. [doi: 10.1056/nejmra1806949]
83. Yürür O, Liu CH, Sheng Z, Leung VCM, Moreno W, Leung KK. Context-awareness for mobile sensing: a survey and future directions. IEEE Commun Surv Tutor 2014;18(1):68-93. [doi: 10.1109/comst.2014.2381246]
84. Picard RW. Affective Computing. Cambridge, MA: MIT Press; 2000.
85. Picard RW. Affective computing: challenges. Int J Hum-Comput Stud 2003;59(1-2):55-64. [doi: 10.1016/s1071-5819(03)00052-1]
86. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 2017;37:98-125. [doi: 10.1016/j.inffus.2017.02.003]
87. Scherer KR, Bänziger T, Roesch E, editors. A Blueprint for Affective Computing: a Sourcebook and Manual. Oxford, United Kingdom: Oxford University Press; 2010.
88. Tao J, Tan T. Affective computing: a review. In: Affective Computing and Intelligent Interaction. Berlin, Heidelberg: Springer; 2005 Presented at: First International Conference, ACII 2005; October 22-24, 2005; Beijing, China p. 981-995. [doi: 10.1007/11573548_125]
89. Wang Y, Song W, Tao W, Liotta A, Yang D, Li X, et al. A systematic review on affective computing: emotion models, databases, and recent advances. Inf Fusion 2022;83-84:19-52. [doi: 10.1016/j.inffus.2022.03.009]
90. Zhao S, Wang S, Soleymani M, Joshi D, Ji Q. Affective computing for large-scale heterogeneous multimedia data: a survey. ACM Trans Multimed Comput Commun Appl 2019;15(3s):1-32. [doi: 10.1145/3363560]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
FDA: Food and Drug Administration
MTurk: Amazon Mechanical Turk

Edited by A Mavragani; submitted 22.07.23; peer-reviewed by E Vashishtha, MO Khursheed, L Guo; comments to author 02.09.23; revised version received 15.11.23; accepted 30.01.24; published 11.04.24.

Please cite as:
Washington P
A Perspective on Crowdsourcing and Human-in-the-Loop Workflows in Precision Health
J Med Internet Res 2024;26:e51138
URL: https://www.jmir.org/2024/1/e51138
doi:10.2196/51138
PMID:38602750

©Peter Washington. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Effect of Everyday-Life Social Contact on Pain

Martin Weiß[1*], PhD; Marthe Gründahl[1*], PhD; Annalena Jachnik[1]; Emilia Caya Lampe[1]; Ishitaa Malik[1]; Heike Lydia Rittner[2], MD; Claudia Sommer[3], MD; Grit Hein[1], PhD

[1]University Hospital Würzburg, Center of Mental Health, Department of Psychiatry, Psychosomatic and Psychotherapy, Translational Social Neuroscience Unit, Würzburg, Germany

[2]University Hospital Würzburg, Center for Interdisciplinary Medicine, Department of Anaesthesiology, Intensive Care, Emergency and Pain Medicine, Würzburg, Germany

[3]University Hospital Würzburg, Department of Neurology, Würzburg, Germany

[*]these authors contributed equally

Corresponding Author:
Martin Weiß, PhD
University Hospital Würzburg, Center of Mental Health, Department of Psychiatry, Psychosomatic and Psychotherapy, Translational Social Neuroscience Unit
Margarete-Höppel-Platz 1
Würzburg, 97080
Germany
Phone: 49 931 201 77412
Email: weiss_m11@ukw.de

Abstract

Pain is a biopsychosocial phenomenon, resulting from the interplay between physiological and psychological processes and social factors. Given that humans constantly interact with others, the effect of social factors is particularly relevant. Documenting the significance of the social modulation of pain, an increasing number of studies have investigated the effect of social contact on subjective pain intensity and pain-related physiological changes. While evidence suggests that social contact can alleviate pain, contradictory findings indicate an increase in pain intensity and a deterioration of pain coping strategies. This evidence primarily stems from studies examining the effect of social contact on pain within highly controlled laboratory conditions. Moreover, pain assessments often rely on one-time subjective reports of average pain intensity across a predefined period. Ecological momentary assessments (EMAs) can circumvent these problems, as they can capture diverse aspects of behavior and experiences multiple times a day, in real time, with high resolution, and within naturalistic and ecologically valid settings. These multiple measures allow for the examination of fluctuations of pain symptoms throughout the day in relation to affective, cognitive, behavioral, and social factors. In this opinion paper, we review the current state and future relevance of EMA-based social pain research in daily life. Specifically, we examine whether everyday-life social support reduces or enhances pain. The first part of the paper provides a comprehensive overview of the use of EMA in pain research and summarizes the main findings. The review of the relatively limited number of existing EMA studies shows that the association between pain and social contact in everyday life depends on numerous factors, including pain syndromes, temporal dynamics, the nature of social interactions, and characteristics of the interaction partners. In line with laboratory research, there is evidence that everyday-life social contact can alleviate, but also intensify pain, depending on the type of social support. Everyday-life emotional support seems to reduce pain, while extensive solicitous support was found to have opposite effects. Moreover, positive short-term effects of social support can be overshadowed by other symptoms such as fatigue. Overall, gathering and integrating experiences from a patient's social environment can offer valuable insights. These insights can help interpret dynamics in pain intensity and accompanying symptoms such as depression or fatigue. We conclude that factors determining the reducing versus enhancing effects of social contact on pain need to be investigated more thoroughly. We advocate EMA as the assessment method of the future and highlight open questions that should be addressed in future EMA studies on pain and the potential of ecological momentary interventions for pain treatment.

(J Med Internet Res 2024;26:e53830) doi:10.2196/53830

KEYWORDS

social contact; pain; ecological momentary assessment

XSL•FO
RenderX

Pain

Pain is associated with impaired physical and mental health and reduced quality of life [1,2]. The International Association for the Study of Pain defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage". Moreover, many pain disorders happen in the absence of tissue damage or any obvious pathophysiological cause [3]. Acute pain is an injury signal and usually recedes when the cause has resolved. Chronic pain is defined as lasting for a minimum of 12 weeks or longer than the expected healing time of an injury. The global prevalence of chronic pain is estimated between 20% and 30% of the population [4-6]. Thus, for many among us, pain is a frequent experience in daily life [2].

The prevalence and consequences of pain call for adequate scientific and clinical measurements and interventions. Yet, there is still a noticeable lack of efficient pain treatments. Among others, this has been traced back to insufficient pain monitoring [6,7] and heterogeneity regarding core outcome measures [8,9]. Pain assessments indicate the severity or quality of pain, enable a diagnosis, and provide indications for medical or therapeutic treatments and their effectiveness [9,10]. The most common pain measure is subjective reports of average pain intensity levels across a predefined period, for example, assessed via visual analog or numerical rating scales [10-12]. However, average and single-time pain measures overlook that (chronic) pain is dynamic rather than static and is characterized by inter- and intraindividual fluctuations in pain intensity, maximum pain levels, and related impairments, which are in turn associated with changes in cognition, affect, behavior, and motivation [2,10,13]. Intraindividual temporal pain variations can be important indicators of pain manageability and overall impairment, and their assessment can thus enhance the understanding and treatment of the pathophysiological, behavioral, and emotional processes related to pain [14,15]. Repeated, fine-grained, temporally precise, and longitudinal pain assessments in a real-life context are needed to adequately and representatively capture the structural and dynamic process of pain [10,13].

Social Contact and Pain

Pain is multifaceted, and pain-related illnesses have heterogeneous symptoms, recoveries, and risks that can strongly differ between individuals [15]. While some patients' daily lives are highly disrupted and impaired, others struggle less in coping with pain. Factors that contribute to these differences are yet poorly understood [15,16]. Researchers are therefore calling for the inclusion of facets other than pain intensity and variability into pain-related research [17]. Notably, pain is often described as a biopsychosocial phenomenon, referring to the interplay between physiological pain-related processes and psychological and social factors [18]. Previous pain studies have shown that pain can impair our social relationships and social functioning [13], for example, by reducing our participation in social life [19,20]. This may in turn increase pain intensity and related symptoms and impairments (eg, negative cognitive processes and impaired quality of life) [2,18]. However, "positive" psychosocial experiences can also decrease pain, for example, social interactions providing social support [1,21]. Social contact has thus been related to either reductions or increases in pain experiences, depending on factors such as the provision of adequate versus extensive social support [22,23], supportive versus unsupportive behavior by others [24,25], or characteristics of the social partners (eg, outgroup vs ingroup membership; sex: female vs male) [26,27]. Protective effects of social contact on pain-related responses are evident in reducing influences on the physiological stress system ("social buffering") [28,29], but also on cognitive and emotional facets such as negative thoughts and negative affect. Consequently, social contact can lead to improved pain reappraisal, less pain-related thoughts, and lower perceived pain intensity [1] and could therefore be an important chess piece in both the development and treatment of pain. Yet, only a few studies have examined how social contact affects the development, resolution, or persistence of pain in a daily life context.

In this paper, we review the status quo of social pain research in daily life and advocate ecological momentary assessments (EMAs) as the assessment method of the future. In particular, we focus on the question of whether everyday-life social support reduces or enhances pain, as laboratory studies provide evidence for both effects. The first part of the paper provides a general overview of the use of EMAs in pain research. Next, we summarize the results of these studies on everyday-life social contact on pain and discuss the ambiguous relationship between social support and pain in more detail. Finally, we highlight open questions and point out future directions.

The literature research for this opinion paper was conducted in June and July 2023 in the databases Google Scholar and PubMed. Searches included the topic "pain" and terms representing EMA, that is, "ecological momentary assessment" OR "ambulatory assessment" OR "experience sampling" OR "diary assessment" OR "intensive longitudinal method" OR "intensive longitudinal study" OR "real-time" OR "daily life" OR "everyday life." The search results were limited to papers that were written in English. The database searches were complemented with manual reviews of the reference lists of relevant papers.

Use of EMAs

EMA (also known as experience sampling or ambulatory assessment) has become increasingly popular in psychological and clinical research over the last decades. In parallel, daily diaries are used as part of clinical diagnostics, as they can provide relevant information (eg, pain intensity ratings) in patients' everyday life. The compliance with daily diaries is comparable to EMAs (approximately 85%) [30,31]. Daily diaries are occasionally categorized among EMA methods [32]. However, they are usually completed only once or twice per day at fixed time points. In contrast, (more advanced) EMA has the capacity to capture various aspects of human behavior and experiences multiple times a day, in real time, in high resolution, and in naturalistic and ecologically valid settings [33-35]. These multiple measures of daily-life experiences can be put into relation to fluctuations of symptoms (eg, pain) throughout the

XSL•FO
RenderX

day, depending on affective, cognitive, behavioral, and—last but not least—social factors [12,36,37]. As ecological momentary interventions (EMIs), EMA methods can even be used for interventional purposes [12,38] and provide treatment once pain levels increase (eg, provision of social support) [39].

EMA offers important advantages to (social) pain research. It reduces the problem of recall bias and thus "purifies" pain measurements, as it captures current experiences rather than retrospective memories [10,40]. Compared with single-time assessments, EMA is less prone to social desirability, cognitive biases, and measurement error [41]. It provides flexibility regarding the selection of representative sampling schedules and intensities, which can be tailored to the study objectives and the demands and capability of specific (pain) samples [17]. In accordance with this, compliance with EMA in patients with different pain conditions was high in previous research [12,42,43]. A key advantage for pain research in particular is EMA's ability to observe and summarize within-person effects and temporal dynamics, including time-lagged relationships, through repeated, longitudinal measurements [17,34]. As summarized by Stone et al [17], EMA can quantify, predict, and potentially influence the ebb and flow of pain, for example, after surgery [44]. What is more, its real-world setting (eg, everyday-life social contact) maximizes the ecological validity of pain-related experiences [17]. Thus, EMA can measure daily-life pain several times a day [13,45], for weeks [9], or even months [46] and capture its relation to daily-life experiences such as social contact [22], work-related stress [47], physical activity [48], and countless other potential influences on pain [17].

EMA can further detect differences between pain conditions, for example, higher variability in fatigue levels in women experiencing fibromyalgia compared with women with rheumatoid arthritis and osteoarthritis [49]. The variability in pain symptomology, related impairments, and treatment response among patients with pain calls for personalized treatments, and continuous, ecological, and momentary pain assessments can advance the design and validation of such interventions [6,50]. EMA could advance the development and monitoring of preventive interventions [51] by observing pain levels of those at risk (eg, patients with acute pain) over time, but also in relation to potential influences in daily life. In the wake of a shift away from on-demand medication and a strong focus on pain in pain treatment, EMA could contribute to tailoring basic analgesia to individual patient needs and patterns [52-54].

Despite these advantages, EMA application in pain research is still relatively rare and, regarding its methodological choices, heterogeneous, unclear, and often outdated [12,17]. By 2020, at least 116 studies had applied EMA for the measurement of pain [12,17]. May et al [12] reviewed 62 research projects that were reported in 105 papers. Only 9 projects (14.5%) used smartphones for data collection, while the majority (39/62, 63%) used other electronic methods, except for phone calls (3/62, 4.8%) and paper booklets (11/62, 17.7%). Stone et al [17] updated this review, focusing on EMA papers published in *PAIN* and *the Journal of Pain* between 2016 and 2020. In addition to those reported in May et al [12], they found 11 papers covering 9 projects, including 2 (22%) projects using smartphone applications and 6 (67%) projects using other full electronic data assessment tools (eg, personal digital assistants such as palmtop computers). Notably, several pain EMA studies used obsolete instead of modern EMA measurement tools such as smartphone-based surveys (eg, telephone calls, pen and paper, or handheld computers) [25,55]. This is a pity, as there is a high acceptance and feasibility of smartphone-based assessments [56-58]. Smartphones are an easy-to-use and broadly available measurement tool with diverse and accurate digital data collection possibilities that almost everyone in our modern world is familiar with [14,59]. Today's technological innovations even enable clinicians to receive in-time feedback on their patients' current (pain) experiences, which could be used for timely interventions. This approach can be integrated into EMIs to deliver individualized, momentary treatments in dependence on current pain-related experiences (eg, predictors of higher pain levels), even without the active participation of a clinician [17]. The prospects and findings outlined above raise the questions of why (social) pain research and practice have not yet shifted to a broader application of (modern) EMA.

Pain and Social Contact in Daily Life

Within and beyond pain research, EMA seems particularly relevant when investigating the effects of an integral part of our daily lives: social contact [1,60]. Notably, social contacts are diverse as they differ in factors like length, content, aim, and tone, but also regarding the number, characteristics, and relationships of social partners [61-63]. Only intensive and ecologically valid measurement tools can adequately capture this diversity. Although the influence of social contact on pain has been acknowledged in the literature [1,21], very few studies have investigated this interplay in daily life settings [22,64,65].

The existing EMA studies on pain and social contact provide promising first insights (see Table 1 for an overview). For example, pain-related social impairment in daily life was evident in a study with 102 adults with multiple sclerosis using wrist-worn monitors. The monitors were worn on the nondominant hand, except during activities such as showering, bathing, or swimming. Participants were asked to rate acute pain, fatigue, depressive mood, and cognitive function on a scale of 0 to 10, 5 times a day using a wrist-worn device [50]. In addition, participants were asked to provide a more detailed report on their social participation with a web-based survey once per day in the evening. However, to uncover relevant features of social contact, researchers usually present several questions to retrieve more comprehensive impressions of the features of the contact (eg, quantity of interaction partners, number of strangers, familiarity and gender of the interaction partner, or perceived personality traits) [31,60,66]. Thus, wrist-worn devices might not be as suited as smartphones when it comes to extensive social contact research in daily life as presenting multiple-choice EMA questions is more burdensome compared with smartphones [67]. In the study described above, higher pain was related to lower same-day social participation [50]. Other EMA results imply ambiguous effects of social contact on pain. In a small-sampled study with older adults with HIV (n=20) and smartphone surveys (5/day for 1 week), social

XSL•FO
RenderX

activity was related to higher levels of fatigue and pain during the day, but also to higher levels of happiness. Looking at temporal relations, higher pain was related to previously being alone, but also to a higher likelihood of not being alone later during the day [65].

Table 1. Selection of studies exemplifying ecological momentary assessment research differing in targeted pain syndrome, measuring method, sample size, social contact measures, and results.

Study	Pain syndrome[a]	Measuring method	Sample size, N	Social contact measure	Results
Kratz et al [50] (2017)	Multiple sclerosis	Wrist-worn monitor	102	• "doing all of the family activities that I want to do" • "doing all of the activities with friends that are really important to me" • "doing all the leisure activities with others that I want to do" • "doing all of the work that I feel I should do (include work at home)"	• Higher pain related to lower same-day social participation
Paolillo et al [65] (2018)	HIV	Smartphone surveys	20	• "Who is with you at this moment?" • "Since the last alarm, how many times did you socialize with someone else (e.g., spent more than 5 min talking or communicating with someone else)?"	• Social activity related to higher levels of fatigue, pain, and happiness • Higher pain related to previously being alone and a higher likelihood of not being alone later on
Herbert et al [64] (2022)	HIV	Smartphone surveys	66	• "Since the last alarm, how many times did you socialize with someone else [e.g., spent more than five minutes talking/communicating with someone else]?"	• Higher frequency of recent social contact related to lower current pain • Higher current pain was linked to a decrease in subsequent social interactions • Higher current negative affect related to higher current pain; this relationship was buffered by increased recent social contact
Rivera et al [68] (2020)	Osteoarthritis of the knee	Telephone calls	268	• "open ended question...asking the participants to indicate what they were doing just prior to receiving the phone call if an interaction was occurring...type of interaction [was assessed] from 'positive' to 'negative', with remaining categories consisting of 'help given', 'help received', 'neutral', and 'professional'."	• In general, more social interactions reduced the association between pain and negative affect • On a within-day level, more social interactions related to more positive affect

[a]Pain syndrome: pain-related disease of study participants.

In a larger sample of older adults with HIV (n=66) and smartphone surveys (4/day for 2 weeks), Herbert et al [64] sought to replicate the results reported by Paolillo et al [65], showing an association between recent social contact and lower current pain. In contrast, Herbert et al [64] found that higher current pain was temporally associated with less subsequent social contact. Interestingly, recent social contact buffered the relation between negative affect and current pain, as only those with low contact frequency exhibited an increase in pain with higher negative affect.

A study with 268 adults with osteoarthritis of the knee investigated the impact of social interactions and pain on daily affect via telephone calls (4/day for 1 week). On a general level, more social interactions reduced the association between pain and negative affect. On a within-day level, more social interactions were related to more positive affect. These results are limited by the minimalistic assessment of daily social interactions: the occurrence of social interactions was coded based on a general question regarding the participant's activity prior to receiving the phone call, and no additional social aspects were assessed [68]. However, similar to other health-related contexts (eg, anxiety-related responses) [60], daily-life pain might change in dependence on the personal characteristics of social partners, such as their gender or familiarity. For instance, there are indications from the laboratory that social support by strangers is less efficient in reducing pain than social support provided by more familiar social partners [21,69]. Similarly, female social partners tend to provide more care and may thus have more pain-reducing effects [70]—or, if providing too much care, pain-enhancing effects (see below). Such influences have rarely been explored in pain research but could be highly relevant for daily-life pain.

Importantly, EMA can expand the researcher's gaze to the experiences and behaviors of others that are part of an

XSL•FO
RenderX

individual's social environment. Interview-based research has shown that chronic pain also impacts a patient's relatives and partners, who for example report changes in leisure activities, sleep disturbances (n=12) [71], feelings of powerlessness, alienation, and emotional distress (n=9) [72]. This can in turn impair the relationships between patients and their social networks [72], thus increasing pain symptoms [73]. It is therefore important to understand when and how pain affects the social environment and vice versa. For instance, interviews with 27 patients with chronic low back pain and their partners indicated that partners' pain responses can be interpreted differently by patients and partners, resulting in misinterpretations. Thus, the interpretation of pain responses may determine whether the behavior in question increases or decreases the patient's pain experiences [74]. Based on these findings, experiences from the family's perspective provide additional insight into factors affecting the development, maintenance, or resolution of pain while also providing indications for intervention. A more fine-grained EMA assessment in daily life, ideally with parallel assessments in patients and their social environments as well as with bigger sample sizes, would be particularly valuable here.

Pain and Social Support: An Ambiguous Relationship

One social factor particularly associated with changes in pain intensity is social support [75,76]. Social support can be characterized as the experience of feeling supported, cared for, and connected to others, contributing to a sense of belonging [77]. Researchers distinguish between perceived and actual (or received) social support [78]. While perceived social support is an individual's subjective evaluation of the emotional and psychological support they believe is available from family and friends in times of need, received social support quantifies the actual support received. Importantly, both types of social support are only moderately correlated [79,80], as the amount of support someone receives does not always align with how much they feel they are supported (for meta-analysis, see [79]). In general, social support is associated with positive emotional states and more effective pain adaptation, which can predict lower pain and improved psychological functioning [81]. It can serve as a coping mechanism, enhance the ability to cope with pain, and facilitate pain management [1,24]. Insufficient and absent social support and social integration have been related to increases in pain [73,82] and even emerged as antecedents or magnifiers of chronic pain [83,84]. Contrasting these pain-buffering effects of social contact, some findings imply the pain-enhancing effects of social contact and support. For instance, a social partner's extensive pain-related concerns in relation to painful stimulation [85] or chronic pain experiences [86,87] can lead to increased pain-related experiences. Solicitous responses (eg, the encouragement to be less active) and negative responses (eg, the expression of frustration or anger about pain) by significant others were previously related to higher pain and disability in patients with chronic pain [86,88,89].

In line with these negative outcomes, operant and cognitive-behavioral models of pain [90,91] assume that people experiencing pain communicate their pain to close others through pain behaviors and that others' (particularly romantic partners') emotional or solicitous responses reinforce these behaviors, while negative responses punish and thus reduce future pain responses [92]. Consequently, the models suggest that social support may compromise self-sufficiency, reinforce maladaptive pain behaviors, foster dependency, and interfere with a patient's ability to cope with pain [24,93]. In contrast, interpersonal process models [94] assume that empathic and validating pain-related responses by spouses lead to positive outcomes by enhancing intimacy and emotion regulation, while negative responses lead to negative pain-related outcomes [25,92]. In relation to these models and partial evidence for each of them, Mogil [23] has suggested a U-shaped relationship between social support and pain experience: in general, the perception of social support versus no support decreases pain [1], but extensive solicitous concern may enhance pain intensity and related factors such as functional disability and pain catastrophizing [95,96] and foster pain expression in patients with pain [87,93].

EMA has provided the first steps to respond to these relevant yet contrasting assumptions with closer-to-life data. In 109 men living with HIV, smartphone-based EMA across 1 week (3 assessments per day) showed that social support related to lower subsequent pain intensity on a within-person level. Results further suggest that this relation may be moderated by between-person factors such as attachment-related avoidance, which was associated with higher pain reduction through social support [22]. In a study of older adults experiencing arthritis, the researchers distinguished between the effects of emotional, solicitous, and negative support. The pain was lower after having received emotional support from one's spouse. However, solicitous support (eg, expression of pain-related concern and comfort) and avoidance behavior by one's spouse were associated with higher pain levels [25]. While highly interesting, this EMA study investigated a very specific sample and used outdated methods (telephone interviews 2 times a day) instead of more feasible and fine-grained smartphone surveys. Similarly, operant models have been particularly investigated in chronic pain couples [25,86,87]. Previous findings should be validated and extended by research with modern methods, diverse pain samples, and other social network members (eg, friends and children).

Note that some research could not find a relationship between social support and pain [97,98]. For example, an EMA study investigating factors contributing to the chronification of pain after potentially traumatic injuries used text message–based questionnaires for 2 weeks in 67 adults. They did not find a relationship between social support and pain on a daily level. However, the study only included 1 assessment per day, and a more fine-grained (and modern) assessment may have provided more insight, as also stated by the authors [98].

In sum, both the effects of social contact on pain, as well as pain-related social impairment, seem to differ within and between days and individuals depending on numerous factors such as pain syndromes, temporal relations, the nature of social interactions, or the social partners and their experiences. A lack of social contact may enhance pain, but it remains unclear under

XSL•FO
RenderX

which conditions pain reduces [50] or fosters [65] social participation. In some, but not all, pain syndromes, short-term symptom-reducing effects of social contact (eg, on fatigue) may become overshadowed by next-day increases in fatigue [99]. Obtaining and integrating experiences from a patient's social environment can provide additional information, for example, regarding the misinterpretation of pain behaviors [74]. Finally, factors determining enhancing versus reducing effects of emotional and solicitous social support need to be investigated more thoroughly [22,25].

Future Directions

There are still some challenges in the EMA application that remain unsolved, as recently pointed out by Stone et al [35]. Despite reports of high feasibility [14,57,58], acceptance [11], and compliance [12,42], EMA can be burdensome to patients with pain [43]). We should therefore strive to make EMA as convenient and appealing as possible, for example, by low weight and size of the assessment tools, easy-to-use interfaces, appealing designs, gamification, or motivational incentives [11,41,100]. There are also issues to be considered from the researchers' and practitioners' side. For instance, EMA protocols require thorough planning and execution, and complex EMA data sets require adequate statistical analyses [33,101]. Other issues include the accessibility of EMA tools, which are currently often limited by factors such as smartphone operating systems (there is a lack of EMA tools running on iOS), and the provision of data protection, particularly when using private smartphones as assessment tools [102]. Moreover, there is a potential selection bias of EMA respondents or the question of whether the interpretation of EMA survey questions is congruent between and within participants (and in accordance with the researcher's intentions) [35]. Solving and standardizing these issues could further advance the field of EMA research [35].

Furthermore, one could inquire whether EMA and its items influence individuals' behavior in social interactions or introduce biases in the perception of their interaction partners. To address this intriguing question, studies would need to assess whether the quality and quantity of social interactions change during the EMA assessment period, that is, with increasing numbers of prompts. This would be an interesting issue for future research. Another relevant aspect of EMA that might be affected by the context of social contact is the participants' compliance [103]. Among the 69 projects involving patients with chronic pain from the review by May et al [12], 39 provided data on completion rates. These projects exhibited an average completion rate of 86%, with individual project rates ranging from approximately 29% to 99%. EMA research examining social interactions with large and healthy samples (≥115 participants) also showed high completion rates of at least 85% [30,31]. Notably, participants in these studies received monetary compensation supplemented by additional bonuses for responding to each prompt or for exceeding a specific number of answered prompts, which may have enhanced compliance. In line with this, a review of pain studies with electronic diaries [42] observed that financial incentives as well as shorter diaries contribute to higher compliance rates (83%). Future research using smartphone-based EMA to study social interactions in patients with pain should thus consider the use of shorter surveys as well as financial incentives per prompt or prompt threshold, contingent upon the availability of funding and ethical approval for compensating patients, to maintain and enhance compliance.

Overall, more homogenous EMA pain study protocols and measurements would strengthen future EMA pain research. As pointed out in a review, previous EMA studies frequently lack detailed reports of factors such as study design decisions and completion rates. They use heterogeneous pain scales, for example, regarding scale labels and points [12]. This decreases the comparability and generalizability of EMA-based results. Future research should aim for a homogenous pain scale to enhance comparability between studies. Numeric rating scales (most commonly anchored from 0 to 10, eg, with 0="no pain" and 10="pain as bad as you can imagine" or "worst pain imaginable") [15,104] are popular, fine-grained, and insightful [10,12] and could serve as a common tool. Based on the experience of our and others' previous research on social interactions, we recommend the consideration of several contextual factors in order to obtain clean and interpretable EMA data [31,38,60,105]: Was the interaction face-to-face or not? How many persons were involved in the interaction? Are questions about an interaction partner's characteristics targeting a single person or the group? How long is or was the interaction? Is it currently ongoing or, if not, how long ago did it end? Up to what time interval between social contact and its assessment should the data be analyzed to avoid recall bias? Are attention checks needed to ensure clean data, particularly in long-term surveys? and Should an alternative, equally long questionnaire be presented in the absence of social contact to prevent participants from avoiding the survey to save time?

It is also noteworthy that a vast majority of previous social and pain-related EMA studies have missed the opportunities of combining subjective momentary pain reports with ambulatory physiological assessments of autonomic measures (eg, heart rate measured with electrocardiogram sensors) and physical activity (eg, through accelerometer sensors). Ambulatory sensors may provide additional information on pain intensity, change, and interference in daily life. There are numerous technologically advanced measurement tools for assessing physiological measures associated with pain experiences and related factors, including wearable electrocardiogram sensors and smartwatches with biosensors. Such ambulatory measurements are objective, unbiased, and can collect data continuously and at a high resolution [33,41]. Previous research has already used ambulatory sensors to investigate social interaction and support in daily life [60,106]. The acceptance and feasibility of mobile measurement devices such as smartwatches are high among the general population [41,107], and the technology is quickly evolving to provide even more opportunities. Next to simply assessing physiological correlates of subjective pain experiences, wearable devices could also be implemented to detect physiological changes and impairments associated with a specific pain syndrome and consequently trigger prompt interventional steps for patients with pain. For instance, in a similar approach as reported by Kim et al [108] on depression, physical activity patterns tracked with ambulatory sensors could be used to objectively estimate momentary pain

XSL•FO
RenderX

levels. Combining smartphone-based pain surveys and portable sensors to measure physiological responses (eg, heart rate, heart rate variability, or other indicators of stress) could provide a more profound understanding of the contextual and proximal causes and consequences of pain and thus advance pain treatment [2,41].

Outlook

This opinion paper requests a drastic expansion of EMA-based pain research, particularly in the context of social contact. The first steps have been made, yet much more remains to be explored regarding the social-pain phenomenon. EMA should not be the sole source of pain-related data, however. Rather, researchers and clinicians should complement clinical assessments (eg, average pain assessments, pain questionnaires) and treatments with EMA as well as other methodological approaches [10,11], such as brain imaging [109], physiological sensors to capture objective measures such as blood pressure or physical activity [33,41], or data from biosamples such as genetics, transcriptomics, and proteomics [110,111]. Moreover, EMI provides great potential for pain research and practice. Future directions may include the personalization of smartphone-based treatments with more extensive and more convenient assessment and treatment tools [11], for example, with respect to the interindividual differences in factors such as pain variability and function [11] or the potential influence of daily-life surroundings such as social contacts [25,64]. Such just-in-time adaptive interventions [112] could help patients with pain identify associations between their current pain level and their momentary activity [113].

Acknowledgments

GH was supported by the German Research Foundation (HE 4566/5-1). MG, AJ, HLR, CS, and GH are members of the Klinische Forschungsgruppe (Clinical Research Unit) 5001 ResolvePAIN, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; Project ID 426503586).

Conflicts of Interest

None declared.

References

1. Che X, Cash R, Ng SK, Fitzgerald P, Fitzgibbon BM. A systematic review of the processes underlying the main and the buffering effect of social support on the experience of pain. Clin J Pain 2018;34(11):1061-1076. [doi: 10.1097/AJP.0000000000000624] [Medline: 29697476]
2. Mills SEE, Nicolson KP, Smith BH. Chronic pain: a review of its epidemiology and associated factors in population-based studies. Br J Anaesth 2019;123(2):e273-e283 [FREE Full text] [doi: 10.1016/j.bja.2019.03.023] [Medline: 31079836]
3. Treede RD. The international association for the study of pain definition of pain: as valid in 2018 as in 1979, but in need of regularly updated footnotes. Pain Rep 2018;3(2):e643 [FREE Full text] [doi: 10.1097/PR9.0000000000000643] [Medline: 29756089]
4. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
5. Pozek JPJ, Beausang D, Baratta JL, Viscusi ER. The acute to chronic pain transition: can chronic pain be prevented? Med Clin North Am 2016;100(1):17-30. [doi: 10.1016/j.mcna.2015.08.005] [Medline: 26614716]
6. Suso-Ribera C, Mesas Á, Medel J, Server A, Márquez E, Castilla D, et al. Improving pain treatment with a smartphone app: study protocol for a randomized controlled trial. Trials 2018;19(1):145 [FREE Full text] [doi: 10.1186/s13063-018-2539-1] [Medline: 29482614]
7. Dansie EJ, Turk DC. Assessment of patients with chronic pain. Br J Anaesth 2013;111(1):19-25 [FREE Full text] [doi: 10.1093/bja/aet124] [Medline: 23794641]
8. Dosenovic S, Kadic AJ, Jeric M, Boric M, Markovic D, Vucic K, et al. Efficacy and safety outcome domains and outcome measures in systematic reviews of neuropathic pain conditions. Clin J Pain 2018;34(7):674-684. [doi: 10.1097/AJP.0000000000000574] [Medline: 29252868]
9. Stone AA, Broderick JE, Schneider S, Schwartz JE. Expanding options for developing outcome measures from momentary assessment data. Psychosom Med 2012;74(4):387-397 [FREE Full text] [doi: 10.1097/PSY.0b013e3182571faa] [Medline: 22582336]
10. Fillingim RB, Loeser JD, Baron R, Edwards RR. Assessment of chronic pain: domains, methods, and mechanisms. J Pain 2016;17(9 Suppl):T10-T20 [FREE Full text] [doi: 10.1016/j.jpain.2015.08.010] [Medline: 27586827]
11. Kaur E, Haghighi PD, Cicuttini FM, Urquhart DM. Smartphone-based ecological momentary assessment for collecting pain and function data for those with low back pain. Sens (Basel) 2022;22(18):7095 [FREE Full text] [doi: 10.3390/s22187095] [Medline: 36146442]

XSL•FO
RenderX

12. May M, Junghaenel DU, Ono M, Stone AA, Schneider S. Ecological momentary assessment methodology in chronic pain research: a systematic review. J Pain 2018;19(7):699-716 [FREE Full text] [doi: 10.1016/j.jpain.2018.01.006] [Medline: 29371113]
13. Mun CJ, Suk HW, Davis MC, Karoly P, Finan P, Tennen H, et al. Investigating intraindividual pain variability: methods, applications, issues, and directions. Pain 2019;160(11):2415-2429. [doi: 10.1097/j.pain.0000000000001626] [Medline: 31145212]
14. Bakshi N, Smith ME, Ross D, Krishnamurti L. Novel metrics in the longitudinal evaluation of pain data in sickle cell disease. Clin J Pain 2017;33(6):517-527. [doi: 10.1097/AJP.0000000000000431] [Medline: 27584817]
15. Schneider S, Junghaenel DU, Keefe FJ, Schwartz JE, Stone AA, Broderick JE. Individual differences in the day-to-day variability of pain, fatigue, and well-being in patients with rheumatic disease: associations with psychological variables. Pain® 2012;153(4):813-822 [FREE Full text] [doi: 10.1016/j.pain.2012.01.001] [Medline: 22349917]
16. de C Williams AC, Fisher E, Hearn L, Eccleston C. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2020;8(8):CD007407 [FREE Full text] [doi: 10.1002/14651858.CD007407.pub4] [Medline: 32794606]
17. Stone AA, Obbarius A, Junghaenel DU, Wen CKF, Schneider S. High-resolution, field approaches for assessing pain: ecological momentary assessment. Pain 2021;162(1):4-9 [FREE Full text] [doi: 10.1097/j.pain.0000000000002049] [Medline: 32833794]
18. Dueñas M, Ojeda B, Salazar A, Mico JA, Failde I. A review of chronic pain impact on patients, their social environment and the health care system. J Pain Res 2016;9:457-467 [FREE Full text] [doi: 10.2147/JPR.S105892] [Medline: 27418853]
19. Leadley RM, Armstrong N, Lee YC, Allen A, Kleijnen J. Chronic diseases in the European Union: the prevalence and health cost implications of chronic pain. J Pain Palliat Care Pharmacother 2012;26(4):310-325. [doi: 10.3109/15360288.2012.736933] [Medline: 23216170]
20. Reid KJ, Harker J, Bala MM, Truyers C, Kellen E, Bekkering GE, et al. Epidemiology of chronic non-cancer pain in Europe: narrative review of prevalence, pain treatments and pain impact. Curr Med Res Opin 2011;27(2):449-462. [doi: 10.1185/03007995.2010.545813] [Medline: 21194394]
21. Krahé C, Springer A, Weinman JA, Fotopoulou A. The social modulation of pain: others as predictive signals of salience—a systematic review. Front Hum Neurosci 2013;7:386 [FREE Full text] [doi: 10.3389/fnhum.2013.00386] [Medline: 23888136]
22. Crockett KB, Turan B. Moment-to-moment changes in perceived social support and pain for men living with HIV: an experience sampling study. Pain 2018;159(12):2503-2511 [FREE Full text] [doi: 10.1097/j.pain.0000000000001354] [Medline: 30074592]
23. Mogil JS. Social modulation of and by pain in humans and rodents. Pain 2015;156(Suppl 1):S35-S41. [doi: 10.1097/01.j.pain.0000460341.62094.77] [Medline: 25789435]
24. Matthias MS, Hirsh AT, Ofner S, Daggy J. Exploring the relationships among social support, patient activation, and pain-related outcomes. Pain Med 2022;23(4):676-685 [FREE Full text] [doi: 10.1093/pm/pnab306] [Medline: 34718764]
25. Pow J, Stephenson E, Hagedoorn M, DeLongis A. Spousal support for patients with rheumatoid arthritis: getting the wrong kind is a pain. Front Psychol 2018;9:1760 [FREE Full text] [doi: 10.3389/fpsyg.2018.01760] [Medline: 30294292]
26. Hein G, Engelmann JB, Tobler PN. Pain relief provided by an outgroup member enhances analgesia. Proc Biol Sci 2018;285(1887):20180501 [FREE Full text] [doi: 10.1098/rspb.2018.0501] [Medline: 30257910]
27. Reddan MC, Young H, Falkner J, López-Solà M, Wager TD. Touch and social support influence interpersonal synchrony and pain. Soc Cogn Affect Neurosci 2020;15(10):1064-1075 [FREE Full text] [doi: 10.1093/scan/nsaa048] [Medline: 32301998]
28. Hennessy MB, Kaiser S, Sachser N. Social buffering of the stress response: diversity, mechanisms, and functions. Front Neuroendocrinol 2009;30(4):470-482. [doi: 10.1016/j.yfrne.2009.06.001] [Medline: 19545584]
29. Kikusui T, Winslow JT, Mori Y. Social buffering: relief from stress and anxiety. Philos Trans R Soc Lond B Biol Sci 2006;361(1476):2215-2228 [FREE Full text] [doi: 10.1098/rstb.2006.1941] [Medline: 17118934]
30. Bernstein MJ, Zawadzki MJ, Juth V, Benfield JA, Smyth JM. Social interactions in daily life: within-person associations between momentary social experiences and psychological and physical health indicators. J Soc Pers Relatsh 2017;35(3):372-394. [doi: 10.1177/0265407517691366]
31. Weiß M, Gründahl M, Jachnik A, Hein G. Who is interacting with whom? assessing the relationship between personality traits and preferences for interaction partners in real life. Collabra: Psychol 2023;9(1):91094 [FREE Full text] [doi: 10.1525/collabra.91094]
32. Lischetzke T, Könen T. Daily diary methodology. In: Maggino F, editor. Encyclopedia of Quality of Life and Well-Being Research. Cham: Springer International Publishing; 2020:1-8.
33. Conner TS, Mehl MR. Ambulatory assessment: methods for studying everyday life. In: Emerging Trends in the Social and Behavioral Sciences. Hoboken, New Jersey: Wiley; 2015:1-15.
34. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
35. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023;19:107-131 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]

XSL•FO
RenderX

36. Colombo D, Fernández-Álvarez J, Patané A, Semonella M, Kwiatkowska M, García-Palacios A, et al. Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: a systematic review. J Clin Med 2019;8(4):465 [FREE Full text] [doi: 10.3390/jcm8040465] [Medline: 30959828]
37. Walz LC, Nauta MH, Rot MAH. Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: a systematic review. J Anxiety Disord 2014;28(8):925-937. [doi: 10.1016/j.janxdis.2014.09.022] [Medline: 25445083]
38. Gründahl M, Deckert J, Hein G. Three questions to consider before applying ecological momentary interventions (EMI) in psychiatry. Front Psychiatry 2020;11:333 [FREE Full text] [doi: 10.3389/fpsyt.2020.00333]
39. Guillory J, Chang P, Henderson CR, Shengelia R, Lama S, Warmington M, et al. Piloting a text message-based social support intervention for patients with chronic pain: establishing feasibility and preliminary efficacy. Clin J Pain 2015;31(6):548-556 [FREE Full text] [doi: 10.1097/AJP.0000000000000193] [Medline: 25565587]
40. Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. Pain 2008;139(1):146-157 [FREE Full text] [doi: 10.1016/j.pain.2008.03.024] [Medline: 18455312]
41. Leroux A, Rzasa-Lynn R, Crainiceanu C, Sharma T. Wearable devices: current status and opportunities in pain assessment and management. Digit Biomark 2021;5(1):89-102 [FREE Full text] [doi: 10.1159/000515576] [Medline: 34056519]
42. Morren M, van Dulmen S, Ouwerkerk J, Bensing J. Compliance with momentary pain measurement using electronic diaries: a systematic review. Eur J Pain 2009;13(4):354-365. [doi: 10.1016/j.ejpain.2008.05.010] [Medline: 18603458]
43. Ono M, Schneider S, Junghaenel DU, Stone AA. What affects the completion of ecological momentary assessments in chronic pain research? an individual patient data meta-analysis. J Med Internet Res 2019;21(2):e11398 [FREE Full text] [doi: 10.2196/11398] [Medline: 30720437]
44. Tighe PJ, Bzdega M, Fillingim RB, Rashidi P, Aytug H. Markov chain evaluation of acute postoperative pain transition states. Pain 2016;157(3):717-728 [FREE Full text] [doi: 10.1097/j.pain.0000000000000429] [Medline: 26588689]
45. Salaffi F, Sarzi-Puttini P, Atzeni F. How to measure chronic pain: new concepts. Best Pract Res Clin Rheumatol 2015;29(1):164-186. [doi: 10.1016/j.berh.2015.04.023] [Medline: 26267010]
46. Garland EL, Hanley AW, Kline A, Cooperman NA. Mindfulness-oriented recovery enhancement reduces opioid craving among individuals with opioid use disorder and chronic pain in medication assisted treatment: ecological momentary assessments from a stage 1 randomized controlled trial. Drug Alcohol Depend 2019;203:61-65 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.07.007] [Medline: 31404850]
47. Bolliger L, Debra G, Lukan J, Peeters R, Colman E, Baele E, et al. The association between day-to-day stress experiences and work-life interference among office workers in academia: an ecological momentary assessment study. Int Arch Occup Environ Health 2023;96(2):201-212 [FREE Full text] [doi: 10.1007/s00420-022-01915-y] [Medline: 36104629]
48. Niermann CYN, Herrmann C, von Haaren B, van Kann D, Woll A. Affect and subsequent physical activity: an ambulatory assessment study examining the affect-activity association in a real-life context. Front Psychol 2016;7:677 [FREE Full text] [doi: 10.3389/fpsyg.2016.00677] [Medline: 27242591]
49. Zautra AJ, Fasman R, Parish BP, Davis MC. Daily fatigue in women with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Pain 2007;128(1-2):128-135. [doi: 10.1016/j.pain.2006.09.004] [Medline: 17055648]
50. Kratz AL, Braley TJ, Foxen-Craft E, Scott E, Murphy JF, Murphy SL. How do pain, fatigue, depressive, and cognitive symptoms relate to well-being and social and physical functioning in the daily lives of individuals with multiple sclerosis? Arch Phys Med Rehabil 2017;98(11):2160-2166 [FREE Full text] [doi: 10.1016/j.apmr.2017.07.004] [Medline: 28729170]
51. Mouraux A, Bannister K, Becker S, Finn DP, Pickering G, Pogatzki-Zahn E, et al. Challenges and opportunities in translational pain research—an opinion paper of the working group on translational pain research of the European pain federation (EFIC). Eur J Pain 2021;25(4):731-756 [FREE Full text] [doi: 10.1002/ejp.1730] [Medline: 33625769]
52. Huijnen IPJ, Rusu AC, Scholich S, Meloto CB, Diatchenko L. Subgrouping of low back pain patients for targeting treatments: evidence from genetic, psychological, and activity-related behavioral approaches. Clin J Pain 2015;31(2):123-132. [doi: 10.1097/AJP.0000000000000100] [Medline: 24681821]
53. Schneider J, Algharably E, Budnick A, Wenzel A, Dräger D, Kreutz R. Deficits in pain medication in older adults with chronic pain receiving home care: a cross-sectional study in Germany. PLoS One 2020;15(2):e0229229 [FREE Full text] [doi: 10.1371/journal.pone.0229229] [Medline: 32084203]
54. Wang R, Wang S, Duan N, Wang Q. From patient-controlled analgesia to artificial intelligence-assisted patient-controlled analgesia: practices and perspectives. Front Med (Lausanne) 2020;7:145 [FREE Full text] [doi: 10.3389/fmed.2020.00145] [Medline: 32671076]
55. Harris RE, Williams DA, McLean SA, Sen A, Hufford M, Gendreau RM, et al. Characterization and consequences of pain variability in individuals with fibromyalgia. Arthritis Rheum 2005;52(11):3670-3674 [FREE Full text] [doi: 10.1002/art.21407] [Medline: 16258905]
56. Bedson J, Hill J, White D, Chen Y, Wathall S, Dent S, et al. Development and validation of a pain monitoring app for patients with musculoskeletal conditions (the keele pain recorder feasibility study). BMC Med Inform Decis Mak 2019;19(1):24 [FREE Full text] [doi: 10.1186/s12911-019-0741-z] [Medline: 30683106]

XSL•FO
RenderX

57. Garcia-Palacios A, Herrero R, Belmonte MA, Castilla D, Guixeres J, Molinari G, et al. Ecological momentary assessment for chronic pain in fibromyalgia using a smartphone: a randomized crossover study. Eur J Pain 2014;18(6):862-872. [doi: 10.1002/j.1532-2149.2013.00425.x] [Medline: 24921074]
58. de Vries LP, Baselmans BML, Bartels M. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J Happiness Stud 2021;22(5):2361-2408 [FREE Full text] [doi: 10.1007/s10902-020-00324-7] [Medline: 34720691]
59. Villegas F, Martínez-Borba V, Suso-Ribera C, Castilla D, Zaragoza I, García-Palacios A, et al. Characterizing breakthrough cancer pain using ecological momentary assessment with a smartphone app: feasibility and clinical findings. Int J Environ Res Public Health 2021;18(11):5991 [FREE Full text] [doi: 10.3390/ijerph18115991] [Medline: 34204871]
60. Gründahl M, Weiß M, Stenzel K, Deckert J, Hein G. The effects of everyday-life social interactions on anxiety-related autonomic responses differ between men and women. Sci Rep 2023;13(1):9498 [FREE Full text] [doi: 10.1038/s41598-023-36118-z] [Medline: 37308494]
61. Argyle M. Social Interaction: Process and Products. Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK: Routledge; 2017.
62. Cohen S. Social relationships and health. Am Psychol 2004;59(8):676-684. [doi: 10.1037/0003-066X.59.8.676] [Medline: 15554821]
63. Uchino BN. Social support and health: a review of physiological processes potentially underlying links to disease outcomes. J Behav Med 2006;29(4):377-387. [doi: 10.1007/s10865-006-9056-5] [Medline: 16758315]
64. Herbert MS, Wooldridge JS, Paolillo EW, Depp CA, Moore RC. Social contact frequency and pain among older adults with HIV: an ecological momentary assessment study. Ann Behav Med 2022;56(2):168-175 [FREE Full text] [doi: 10.1093/abm/kaab037] [Medline: 34057465]
65. Paolillo EW, Tang B, Depp CA, Rooney AS, Vaida F, Kaufmann CN, et al. Temporal associations between social activity and mood, fatigue, and pain in older adults with HIV: an ecological momentary assessment study. JMIR Ment Health 2018;5(2):e38 [FREE Full text] [doi: 10.2196/mental.9802] [Medline: 29759960]
66. Heinig I, Weiß M, Hamm AO, Hein G, Hollandt M, Hoyer J, et al. Exposure traced in daily life: improvements in ecologically assessed social and physical activity following exposure-based psychotherapy for anxiety disorders. J Anxiety Disord 2024;101:102792 [FREE Full text] [doi: 10.1016/j.janxdis.2023.102792] [Medline: 37989038]
67. Ponnada A, Haynes C, Maniar D, Manjourides J, Intille S. Microinteraction ecological momentary assessment response rates: effect of microinteractions or the smartwatch? Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):92 [FREE Full text] [doi: 10.1145/3130957] [Medline: 30198012]
68. Rivera NV, Parmelee PA, Smith DM. The impact of social interactions and pain on daily positive and negative affect in adults with osteoarthritis of the knee. Aging Ment Health 2020;24(1):8-14. [doi: 10.1080/13607863.2018.1506744] [Medline: 30380912]
69. Eisenberger NI. An empirical review of the neural underpinnings of receiving and giving social support: implications for health. Psychosom Med 2013;75(6):545-556 [FREE Full text] [doi: 10.1097/PSY.0b013e31829de2e7] [Medline: 23804014]
70. Liebler CA, Sandefur GD. Gender differences in the exchange of social support with friends, neighbors, and co-workers at midlife. Social Science Research 2002;31(3):364-391. [doi: 10.1016/s0049-089x(02)00006-6]
71. Ojeda B, Salazar A, Dueñas M, Torres LM, Micó JA, Failde I. The impact of chronic pain: the perspective of patients, relatives, and caregivers. Families, Systems, Health 2014;32(4):399-407. [doi: 10.1037/fsh0000069]
72. West C, Usher K, Foster K, Stewart L. Chronic pain and the family: the experience of the partners of people living with chronic pain. J Clin Nurs 2012;21(23-24):3352-3360. [doi: 10.1111/j.1365-2702.2012.04215.x] [Medline: 22834990]
73. Nicolson PJA, Williamson E, Morris A, Sanchez-Santos MT, Bruce J, Silman A, et al. Musculoskeletal pain and loneliness, social support and social engagement among older adults: analysis of the Oxford pain, activity and lifestyle cohort. Musculoskeletal Care 2021;19(3):269-277 [FREE Full text] [doi: 10.1002/msc.1526] [Medline: 33201582]
74. Akbari F, Mohammadi S, Dehghani M, Sanderman R, Hagedoorn M. Interpretations of partners' responses to pain behaviours: perspectives of patients and partners. Br J Health Psychol 2021;26(2):401-418 [FREE Full text] [doi: 10.1111/bjhp.12490] [Medline: 33180996]
75. López-Martínez AE, Esteve-Zarazaga R, Ramírez-Maestre C. Perceived social support and coping responses are independent variables explaining pain adjustment among chronic pain patients. J Pain 2008;9(4):373-379 [FREE Full text] [doi: 10.1016/j.jpain.2007.12.002] [Medline: 18203665]
76. Roberts MH, Klatzkin RR, Mechlin B. Social support attenuates physiological stress responses and experimental pain sensitivity to cold pressor pain. Ann Behav Med 2015;49(4):557-569 [FREE Full text] [doi: 10.1007/s12160-015-9686-3] [Medline: 25623896]
77. Cobb S. Presidential address-1976. Social support as a moderator of life stress. Psychosom Med 1976;38(5):300-314. [doi: 10.1097/00006842-197609000-00003] [Medline: 981490]
78. Eagle DE, Hybels CF, Proeschold-Bell RJ. Perceived social support, received social support, and depression among clergy. J Soc Pers Relatsh 2018;36(7):2055-2073. [doi: 10.1177/0265407518776134]

XSL•FO
RenderX

79. Haber MG, Cohen JL, Lucas T, Baltes BB. The relationship between self-reported received and perceived social support: a meta-analytic review. Am J Community Psychol 2007;39(1-2):133-144 [FREE Full text] [doi: 10.1007/s10464-007-9100-9] [Medline: 17308966]
80. Lakey B, Orehek E, Hain KL, Van Vleet M. Enacted support's links to negative affect and perceived support are more consistent with theory when social influences are isolated from trait influences. Pers Soc Psychol Bull 2010;36(1):132-142. [doi: 10.1177/0146167209349375] [Medline: 19875827]
81. Sturgeon JA, Zautra AJ. Social pain and physical pain: shared paths to resilience. Pain Manag 2016;6(1):63-74 [FREE Full text] [doi: 10.2217/pmt.15.56] [Medline: 26678402]
82. Wolf LD, Davis MC, Yeung EW, Tennen HA. The within-day relation between lonely episodes and subsequent clinical pain in individuals with fibromyalgia: mediating role of pain cognitions. J Psychosom Res 2015;79(3):202-206 [FREE Full text] [doi: 10.1016/j.jpsychores.2014.12.018] [Medline: 25637526]
83. Evers AWM, Kraaimaat FW, Geenen R, Jacobs JWG, Bijlsma JWJ. Pain coping and social support as predictors of long-term functional disability and pain in early rheumatoid arthritis. Behav Res Ther 2003;41(11):1295-1310. [doi: 10.1016/s0005-7967(03)00036-6] [Medline: 14527529]
84. Jacobs JM, Hammerman-Rozenberg R, Cohen A, Stessman J. Chronic back pain among the elderly: prevalence, associations, and predictors. Spine (Phila Pa 1976) 2006;31(7):E203-E207. [doi: 10.1097/01.brs.0000206367.57918.3c] [Medline: 16582841]
85. Hurter S, Paloyelis Y, de C Williams AC, Fotopoulou A. Partners' empathy increases pain ratings: effects of perceived empathy and attachment style on pain report and display. J Pain 2014;15(9):934-944 [FREE Full text] [doi: 10.1016/j.jpain.2014.06.004] [Medline: 24953886]
86. Raichle KA, Romano JM, Jensen MP. Partner responses to patient pain and well behaviors and their relationship to patient pain behavior, functioning, and depression. Pain® 2011;152(1):82-88. [doi: 10.1016/j.pain.2010.09.015] [Medline: 20947249]
87. Romano JM, Jensen MP, Turner JA, Good AB, Hops H. Chronic pain patient-partner interactions: further support for a behavioral model of chronic pain. Behav Ther 2000;31(3):415-440. [doi: 10.1016/s0005-7894(00)80023-4]
88. Jensen MP, Moore MR, Bockow TB, Ehde DM, Engel JM. Psychosocial factors and adjustment to chronic pain in persons with physical disabilities: a systematic review. Arch Phys Med Rehabil 2011;92(1):146-160 [FREE Full text] [doi: 10.1016/j.apmr.2010.09.021] [Medline: 21187217]
89. Lumley MA, Cohen JL, Borszcz GS, Cano A, Radcliffe AM, Porter LS, et al. Pain and emotion: a biopsychosocial review of recent research. J Clin Psychol 2011;67(9):942-968 [FREE Full text] [doi: 10.1002/jclp.20816] [Medline: 21647882]
90. Fordyce WE. Behavioral Methods for Chronic Pain and Illness. Clay County, Missouri, United States: Mosby; 1976.
91. Turk DC, Meichenbaum D, Genest M. Pain and Behavioral Medicine: A Cognitive-Behavioral Perspective. New York City: Guilford Press; 1983.
92. Cano A, de C Williams AC. Social interaction in pain: reinforcing pain behaviors or building intimacy? Pain® 2010;149(1):9-11 [FREE Full text] [doi: 10.1016/j.pain.2009.10.010] [Medline: 19892466]
93. Turk DC, Kerns RD, Rosenberg R. Effects of marital interaction on chronic pain and disability: examining the down side of social support. Rehabil Psychol 1992;37(4):259-274. [doi: 10.1037/0090-5550.37.4.259]
94. Reis HT. Intimacy as an interpersonal process. In: Relationships, Well-Being and Behaviour. Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK: Routledge; 2018:113-143.
95. Keefe FJ, Lipkus I, Lefebvre JC, Hurwitz H, Clipp E, Smith J, et al. The social context of gastrointestinal cancer pain: a preliminary study examining the relation of patient pain catastrophizing to patient perceptions of social support and caregiver stress and negative responses. Pain® 2003;103(1-2):151-156. [doi: 10.1016/s0304-3959(02)00447-5] [Medline: 12749969]
96. McCracken LM. Social context and acceptance of chronic pain: the role of solicitous and punishing responses. Pain 2005;113(1-2):155-159. [doi: 10.1016/j.pain.2004.10.004] [Medline: 15621376]
97. Hanley MA, Raichle K, Jensen M, Cardenas DD. Pain catastrophizing and beliefs predict changes in pain interference and psychological functioning in persons with spinal cord injury. J Pain 2008;9(9):863-871 [FREE Full text] [doi: 10.1016/j.jpain.2008.04.008] [Medline: 18550442]
98. Pacella ML, Girard JM, Wright AGC, Suffoletto B, Callaway CW. The association between daily posttraumatic stress symptoms and pain over the first 14 days after injury: an experience sampling study. Acad Emerg Med 2018;25(8):844-855 [FREE Full text] [doi: 10.1111/acem.13406] [Medline: 29513381]
99. Parrish BP, Zautra AJ, Davis MC. The role of positive and negative interpersonal events on daily fatigue in women with fibromyalgia, rheumatoid arthritis, and osteoarthritis. Health Psychol 2008;27(6):694-702 [FREE Full text] [doi: 10.1037/0278-6133.27.6.694] [Medline: 19025264]
100. Moore RC, Depp CA, Wetherell JL, Lenze EJ. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. J Psychiatr Res 2016;75:116-123 [FREE Full text] [doi: 10.1016/j.jpsychires.2016.01.011] [Medline: 26851494]
101. Wilhelm FH, Grossman P, Müller MI. Bridging the gap between the laboratory and the real world: integrative ambulatory psychophysiology. In: Mehl MR, Conner TS, editors. Handbook of Research Methods for Studying Daily Life. New York City: Guilford Press; 2012:210-234.

XSL•FO
RenderX

102. Doherty K, Balaskas A, Doherty G. The design of ecological momentary assessment technologies. Interact Comput 2020;32(1):257-278 [FREE Full text] [doi: 10.1093/iwcomp/iwaa019]
103. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
104. Kratz AL, Murphy SL, Braley TJ. Ecological momentary assessment of pain, fatigue, depressive, and cognitive symptoms reveals significant daily variability in multiple sclerosis. Arch Phys Med Rehabil 2017;98(11):2142-2150 [FREE Full text] [doi: 10.1016/j.apmr.2017.07.002] [Medline: 28729168]
105. Schwerdtfeger AR, Rominger C, Obser PD. A shy heart may benefit from everyday life social interactions with close others: an ecological momentary assessment trial using bayesian multilevel modeling. Biol Psychol 2020;152:107864. [doi: 10.1016/j.biopsycho.2020.107864] [Medline: 32036003]
106. Gerteis AKS, Schwerdtfeger AR. When rumination counts: perceived social support and heart rate variability in daily life. Psychophysiology 2016;53(7):1034-1043. [doi: 10.1111/psyp.12652] [Medline: 27137911]
107. Lau SCL, Connor LT, King AA, Baum CM. Multimodal ambulatory monitoring of daily activity and health-related symptoms in community-dwelling survivors of stroke: feasibility, acceptability, and validity. Arch Phys Med Rehabil 2022;103(10):1992-2000 [FREE Full text] [doi: 10.1016/j.apmr.2022.06.002] [Medline: 35780826]
108. Kim J, Nakamura T, Kikuchi H, Yoshiuchi K, Sasaki T, Yamamoto Y. Covariation of depressive mood and spontaneous physical activity in major depressive disorder: toward continuous monitoring of depressive mood. IEEE J Biomed Health Inform 2015;19(4):1347-1355 [FREE Full text] [doi: 10.1109/JBHI.2015.2440764] [Medline: 26054079]
109. Davis KD, Moayedi M. Central mechanisms of pain revealed through functional and structural MRI. J Neuroimmune Pharmacol 2013;8(3):518-534. [doi: 10.1007/s11481-012-9386-8] [Medline: 22825710]
110. Liang P, Dai M, Wang X, Wang D, Yang M, Lin X, et al. Efficacy and safety of ciprofol vs. propofol for the induction and maintenance of general anaesthesia: a multicentre, single-blind, randomised, parallel-group, phase 3 clinical trial. Eur J Anaesthesiol 2023;40(6):399-406 [FREE Full text] [doi: 10.1097/EJA.0000000000001799] [Medline: 36647565]
111. Sluka KA, Wager TD, Sutherland SP, Labosky PA, Balach T, Bayman EO, et al. Predicting chronic postsurgical pain: current evidence and a novel program to develop predictive biomarker signatures. Pain 2023;164(9):1912-1926 [FREE Full text] [doi: 10.1097/j.pain.0000000000002938] [Medline: 37326643]
112. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
113. Kaplan DM. Social-ecological measurement of daily life: how relationally focused ambulatory assessment can advance clinical intervention science. Rev Gen Psychol 2023;27(2):206-227. [doi: 10.1177/10892680221142802]

Abbreviations

EMA: ecological momentary assessment
EMI: ecological momentary intervention

Edited by A Mavragani; submitted 20.10.23; peer-reviewed by DM Niddam, K Yoshiuchi; comments to author 19.01.24; revised version received 01.02.24; accepted 13.03.24; published 30.04.24.

Please cite as:
Weiß M, Gründahl M, Jachnik A, Lampe EC, Malik I, Rittner HL, Sommer C, Hein G
The Effect of Everyday-Life Social Contact on Pain
J Med Internet Res 2024;26:e53830
URL: https://www.jmir.org/2024/1/e53830
doi:10.2196/53830
PMID:38687594

©Martin Weiß, Marthe Gründahl, Annalena Jachnik, Emilia Caya Lampe, Ishitaa Malik, Heike Lydia Rittner, Claudia Sommer, Grit Hein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Envisioning the Future of Personalized Medicine: Role and Realities of Digital Twins

Alexandre Vallée[1], MD, PhD

Department of Epidemiology and Public Health, Foch Hospital, Suresnes, France

Corresponding Author:
Alexandre Vallée, MD, PhD
Department of Epidemiology and Public Health
Foch Hospital
40 rue Worth
Suresnes, 92150
France
Phone: 33 0146257352
Fax: 33 0146257352
Email: al.vallee@hopital-foch.com

Abstract

Digital twins have emerged as a groundbreaking concept in personalized medicine, offering immense potential to transform health care delivery and improve patient outcomes. It is important to highlight the impact of digital twins on personalized medicine across the understanding of patient health, risk assessment, clinical trials and drug development, and patient monitoring. By mirroring individual health profiles, digital twins offer unparalleled insights into patient-specific conditions, enabling more accurate risk assessments and tailored interventions. However, their application extends beyond clinical benefits, prompting significant ethical debates over data privacy, consent, and potential biases in health care. The rapid evolution of this technology necessitates a careful balancing act between innovation and ethical responsibility. As the field of personalized medicine continues to evolve, digital twins hold tremendous promise in transforming health care delivery and revolutionizing patient care. While challenges exist, the continued development and integration of digital twins hold the potential to revolutionize personalized medicine, ushering in an era of tailored treatments and improved patient well-being. Digital twins can assist in recognizing trends and indicators that might signal the presence of diseases or forecast the likelihood of developing specific medical conditions, along with the progression of such diseases. Nevertheless, the use of human digital twins gives rise to ethical dilemmas related to informed consent, data ownership, and the potential for discrimination based on health profiles. There is a critical need for robust guidelines and regulations to navigate these challenges, ensuring that the pursuit of advanced health care solutions does not compromise patient rights and well-being. This viewpoint aims to ignite a comprehensive dialogue on the responsible integration of digital twins in medicine, advocating for a future where technology serves as a cornerstone for personalized, ethical, and effective patient care.

(J Med Internet Res 2024;26:e50204) doi:10.2196/50204

KEYWORDS

digital health; digital twin; personalized medicine; prevention; prediction; health care system

Introduction

Digital twins have emerged as a groundbreaking concept in the field of personalized medicine, offering immense potential to transform health care delivery and improve patient outcomes [1-3]. By creating digital replicas of individuals and leveraging advanced technologies, digital twins enable health care professionals to gain a comprehensive understanding of patients' health, personalize treatment plans, and make data-driven decisions [4,5]. In this era of rapid technological advancements, digital twins have proven to be instrumental in enhancing the understanding of patient health, predictive modeling, risk assessment, digital clinical trials, remote patient monitoring, and telemedicine [6,7].

One of the key advantages of digital twins is their ability to integrate and analyze diverse data sets, including electronic health records (EHRs), wearable devices, genetic information, and patient-reported data [8]. This comprehensive data integration empowers health care providers to have a holistic view of a patient's health status, identifying patterns, correlations, and potential health risks that might go unnoticed with isolated data sources. Through real-time monitoring and

XSL•FO
RenderX

feedback, digital twins enable continuous assessment of a patient's health, allowing for timely interventions and adjustments to treatment plans when necessary [1,6].

Digital twins also have a major role in predictive modeling and risk assessment [9]. By leveraging artificial intelligence (AI) algorithms and machine learning techniques, digital twins can simulate disease progression [10], predict treatment responses [11,12], and conduct personalized risk assessments [13]. This personalized approach considers factors such as genetics, lifestyle choices, and environmental influences, enabling health care providers to offer targeted preventive measures, early detection strategies, and personalized interventions to mitigate identified risks [14,15].

Moreover, digital twins have the potential to revolutionize digital clinical trials and drug development [2]. By creating digital patient populations and simulating treatment effects, digital twins streamline the drug discovery process, optimize trial design, and reduce costs and timelines [16,17]. They enable researchers to assess the efficacy and safety of drug candidates before conducting traditional trials, accelerating the availability of new treatments to patients [16].

Moreover, digital twins contribute significantly to remote patient monitoring and telemedicine [18]. By leveraging wearable sensors, Internet of Things (IoT) devices, and remote data transmission, digital twins facilitate continuous monitoring of patient health parameters, personalized interventions, and remote consultations [13]. This approach improves patient access to health care services, reduces hospitalizations, and empowers individuals to actively participate in their own health care management [19].

As the field of personalized medicine continues to evolve, digital twins hold tremendous promise in transforming health care delivery and revolutionizing patient care [20]. By harnessing the power of data integration, predictive modeling, digital simulations, and remote monitoring, digital twins enhance the understanding of patient health, optimize treatment strategies, and pave the way for a more personalized and efficient health care system [1]. Digital twins are enhancing health care by providing health care professionals with a deeper understanding of patients' health, enabling personalized care, and optimizing treatment strategies. They also hold immense potential in advancing drug development and making health care more accessible through remote monitoring and telemedicine. Personalized medicine is at a crossroads, and digital twins represent a path toward a more precise and patient-centric health care paradigm. Drawing from years of research and clinical observations, it is important to argue that the integration of digital twins is not just an advancement but a necessity for modern health care. Thus, this viewpoint critically examines the role of digital twins on personalized medicine across (1) the understanding of patient health, (2) risk assessment, (3) clinical trials and drug development, and (4) patient monitoring.

Digital Twin

In recent years, the concept of digital twins has been receiving increasing attention from both researchers and engineers. As research in the field of digital twins progresses, carried out by both industry and academia, the boundaries between digital twins and other related concepts have started to blur. Initially, the scope of the digital twin included physical and digital products along with their interconnections [21]. This concept has evolved due to the rapid advancements in communication technology, sensor technology, big data analysis, IoTs, and simulation technology [22]. This growth has led to significant research into digital twins, even if digital twins remain theoretical applications.

Subsequently, the digital twin was redefined as a digital replication of living or nonliving physical entities, opening applications in areas such as health and well-being [23]. As a dynamic concept, the digital twin represents a digital replica of human organs, tissues, cells, or microenvironments that continuously adapts to real-time data variations and predicts corresponding future scenarios [24]. However, the digital twin goes beyond being just a digital model linked to its real-life counterpart through emerging technologies. It emerges as a sentient, intelligent, and evolving model capable of optimizing processes and continually forecasting future states, such as defects, damages, and failures, through a closed-loop interaction between the digital twin and its surrounding environment.

Broadly, the technologies crucial for the digital twin can be categorized into 2 groups as follows: one involves a data-driven statistical model, while the other integrates multiscale knowledge and data into a mechanical model [25]. The numerical model calculates structural performance, while the analytical model facilitates structural analysis. An AI model, trained with samples and numerical data, derives real-time structural insights from sensor data.

The impact of the digital twin is profoundly reshaping industries and has been adopted by major corporations to enhance efficiency and identify issues. This transformative technology is also making its way into the health care sector. In this context, the digital twin can treat patients as digitalized stand-alone assets applicable to various health care scenarios [26]. This potential holds significant promise for improving treatment and diagnostics within hospitals and for individual patients.

A digital twin represents a digital replica of a tangible entity or process, such as a patient, their anatomical structure, or the setting of a hospital. Currently, digital twins are designed to dynamically mirror various data sources, including disease registries, "-omics" data (such as genomics, biomics, proteomics, or metabolomics data), as well as physical indicators, demographic information, and lifestyle data pertaining to an individual's progression over time [27]. The evolution of foundational technologies like the IoTs and AI, coupled with the availability of an expanding array of accurate and accessible data types (ranging from biometric and behavioral data to emotional, cognitive, and psychological insights), has sparked increased interest and exploration in the research and potential applications of digital twins within the health care domain [27].

XSL•FO
RenderX

Enhanced Understanding of Patient Health

Digital twins have a major role in providing health care professionals with an in-depth and comprehensive understanding of an individual's health status [28]. By integrating and analyzing data from various sources, digital twins offer a holistic view of a patient's physiological parameters, lifestyle patterns, and genetic predispositions [21,29]. This enhanced understanding enables health care providers to personalize treatment plans and interventions according to an individual's unique needs, ultimately improving patient outcomes [30].

Digital twins integrate data from multiple sources, including EHRs [5], wearable devices [24], genetic information [31], and patient-reported data [2]. This comprehensive data integration allows health care professionals to gain a more complete picture of a patient's health. By aggregating and analyzing diverse data sets, digital twins can identify patterns, correlations, and potential health risks that might not be evident through isolated data sources [32].

Digital twins enable continuous real-time monitoring of patient data, providing health care providers with up-to-date information on vital signs, biomarkers, medication adherence, and other relevant parameters [33]. This real-time feedback allows for immediate assessment of a patient's health status and the ability to make timely interventions or adjustments to treatment plans when necessary [34]. Real-time data can be provided by IoT solutions, and large data flows may be managed and secured by robust digital infrastructures [6]. For example, a hospital's digital transformation team has put forth a plan to create an advanced decision support model that uses real-time data from various health care systems and devices. This model facilitated the evaluation of the effectiveness of current health care delivery systems and the assessment of the potential impact of service modifications, all while seamlessly integrating with the hospital's day-to-day operations. It offered the ability to predict the outcomes of proposed model changes before implementing them in practice [35].

By leveraging AI algorithms and machine learning techniques, digital twins can analyze patient data to assess an individual's risk factors for specific diseases or health conditions with high levels of performance [31,36] (Table 1). This personalized risk assessment considers a patient's genetic profile, lifestyle choices, environmental factors, and other relevant data [37,38]. With this information, health care providers can offer targeted preventive measures, early detection strategies, and personalized interventions to mitigate the identified risks.

XSL•FO
RenderX

Table 1. Models and accuracies of digital twins in research.

Year	Applications	Modeling	Database	Accuracy	Reference
2022	Cancer analyses progression and specific organ identification	Natural language processing labeling with CNN[a], recurrent neural network models, term frequency-inverse document frequency ensemble model	Computed tomography scans (2009-2021)	In validation population: 93.8% for lung cancer, 92.5% for liver cancer, and 96.1% for adrenal cancer with term frequency-inverse document frequency ensemble model; 96.8% for lung, 99% for liver, and 99.7% for adrenal with augmented CNN.	[39]
2022	Personal health care improvement with emotion recognition	Automatic detection of emotion from EEG[b] signals with gradient boosting, k-nearest neighbor, and RF[c] models	EEG images	99.9% with gradient boosting, 98.6% with decision tree classifier, 99.7% with k-nearest neighbor, and 99.6% with RF	[40]
2022	Prostate cancer progression with biochemical recurrence and seminal vesicle	Machine learning methods with support vector machine, RF, NN, recurrent neural network, and long short-term memory models	Clinical data warehouse with patients with cancer	82.7% of accuracy for biochemical recurrence and 83.9% for seminal vesicle	[41]
2021	Abdominal aortic aneurysm severity detection	Inverse analysis with CNN and long short-term memory models	Digital patient's data set	99.91% for detection and 97.79% for severity	[42]
2021	Prevention of stroke and treatment of post-stroke	Support vector machine	EEG data set	76% accuracy and 0.84 for performance	[43]
2019	Fault diagnosis pattern	Deep neural network model with deep-neural network and deep transfer learning (DFDD[d])	Digital patient's data set	98% for accuracy with DFDD and 91.5% with deep-neural network	[44]
2019	Ischemic heart disease detection	Deep neural model	Pulmonary tuberculosis diagnostic electrocardiogram database	85.8% for the implemented model	[45]

[a]CNN: convolutional neural network.

[b]EEG: Electroencephalography.

[c]RF: random forest.

[d]DFDD: fault diagnosis method using deep transfer learning.

Digital twins can simulate and predict the progression of diseases based on real-time patient data [2]. By using historical patient data and integrating it with the current health parameters, digital twins can generate predictive models that help anticipate disease progression, identify potential complications, and estimate treatment outcomes [46]. This predictive modeling enables health care professionals to make informed decisions and develop personalized treatment plans that maximize efficacy and minimize risks.

Moreover, digital twins facilitate longitudinal tracking of patient health data, allowing health care providers to analyze trends and changes over time. This longitudinal view provides insights into disease progression, response to treatments, and the impact of lifestyle modifications on overall health outcomes [47,48]. Thus, by identifying the patterns and trends, digital twins can help health care professionals identify personalized interventions and adjustments to optimize patient care [1].

Predictive Modeling and Risk Assessment

Digital twins can leverage advanced algorithms and machine learning techniques to perform predictive modeling and risk assessment in the field of personalized medicine [49]. By analyzing real-time patient data and historical records, digital twins enable health care professionals to anticipate disease progression, identify potential risks, and optimize treatment strategies [2].

Digital twins can simulate and predict the progression of diseases based on patient-specific data [19]. By integrating various data sources such as genetic information, biomarkers, lifestyle factors, and treatment history, digital twins create digital replicas of patients and model the progression of diseases over time [5]. This modeling allows health care professionals to forecast potential outcomes, anticipate complications, and adjust treatment plans accordingly.

Digital twins enable health care providers to predict how individual patients will respond to different treatment options [21,50]. By analyzing data from similar patient cases and incorporating real-time patient data, digital twins can generate predictive models, taking into consideration uncertainty and confidence intervals, that estimate treatment outcomes [46,51]. This information helps health care professionals make informed decisions about the most effective treatments for individual patients, increasing treatment success rates and reducing trial and error [52,53].

XSL•FO
RenderX

Digital twins assess individual patients' risks for specific diseases or health conditions by analyzing their genetic profiles, lifestyle choices, environmental factors, and other relevant data [1,28]. By integrating this information and leveraging machine learning algorithms, digital twins can identify personalized risk factors and quantify the likelihood of developing certain conditions [54]. This personalized risk assessment enables health care providers to implement preventive measures, early detection strategies, and targeted interventions to mitigate risks [55].

Furthermore, digital twins can serve as early warning systems by continuously monitoring patient data and identifying potential health risks or deviations from normal patterns [56,57]. By establishing baseline parameters for each patient, digital twins can detect anomalies or warning signs that may indicate the onset of a disease or adverse health event [58]. This early detection allows for timely interventions, proactive health care management, and prevention of complications [1]. By considering individual patient characteristics, such as genetics, lifestyle, and response to previous treatments, digital twins can recommend tailored treatment plans that maximize efficacy and minimize side effects [5,21]. This personalized approach helps health care professionals select the most suitable treatments, dosages, and interventions for each patient, leading to improved outcomes.

The use of digital twins for predictive modeling and risk assessment presents several ethical considerations that should be carefully examined and addressed [59]. The collection and storage of data should be done with a strong focus on privacy and data security and ensuring that personal or sensitive information is adequately protected is paramount [57]. Moreover, individuals whose data are used in digital twin models should provide informed consent. They should be aware of how their data are being used and could opt out if they choose [21]. Furthermore, there should be transparency in how digital twins are created and used. Clear explanations of the modeling process and the factors influencing predictions should be provided to stakeholders, including regulators and the public [59].

Digital Clinical Trials and Drug Development

Digital twins are poised to revolutionize the landscape of clinical trials and drug development, offering significant advantages over traditional approaches [60]. By simulating digital patient populations and leveraging advanced technologies, digital twins streamline the drug discovery process, accelerate clinical trials, and enhance the efficacy and safety of new treatments [60].

The creation of digital patient populations that closely resemble real-world patient demographics is now possible with the future development of digital twins [61-63]. These digital populations can be customized to reflect diverse characteristics, such as age, sex, genetics, and comorbidities [30]. By generating digital patients with varying profiles, digital twins provide a comprehensive representation of the population under study, improving the generalizability of trial results [64]. Indeed, digital twins could simulate the effects of potential treatments on digital patients, considering individual patient characteristics and treatment responses [51]. This allows researchers to assess the efficacy and safety of drug candidates before conducting costly and time-consuming clinical trials. Digital twins assist in optimizing the design of clinical trials by providing insights into patient recruitment, trial end points, sample size determination, and treatment protocols [12,27]. By analyzing digital patient populations, digital twins can predict the likely response rates, treatment effects, and potential adverse events, helping researchers tailor trial parameters for maximum efficiency and statistical power [65,66]. Digital clinical trials conducted through digital twins offer substantial cost and time savings [51]. By replacing or supplementing traditional trials, which often involve extensive site visits and patient recruitment, digital twins streamline the data collection process. Additionally, digital trials eliminate the need for physical infrastructure, reduce administrative burdens, and enable remote patient monitoring [67]. These efficiencies translate into reduced trial costs and shorter timelines, expediting the availability of new treatments to patients in need [60].

The contribution of such tools can improve patient safety in drug development [16]. By using digital patient data, digital twins can assess the potential risks and side effects of new treatments, enabling researchers to optimize dosages and identify vulnerable patient subgroups [16]. This approach minimizes the exposure of actual patients to experimental treatments, ensuring ethical considerations and protecting patient well-being [57,68].

The facilitation of real-time monitoring of digital patients by digital twins can allow researchers to collect and analyze data continuously [69-71]. This capability enables adaptive trial designs, where trial parameters can be modified based on ongoing analysis of digital patient responses [72,73]. Adaptive trials enhance the efficiency of clinical research by reducing the number of patients required, optimizing treatment regimens, and maximizing the chances of success [74].

Clinical drug development is an intricate and time-intensive journey, spanning roughly 6-15 years [75,76]. The financial commitment for shepherding a novel drug from its conceptual infancy through the labyrinth of research, development, and eventual market approval teeters at an astounding sum of approximately US $2.6 billion [77]. Shockingly, approximately 85% of potential therapies stumble in the early stages of clinical development, and of the fortunate few that reach phase 3, only half earn the coveted stamp of approval [78]. Another daunting statistic reveals that nearly 80% of trials falter in their attempts to meet their initial enrollment goals and prescribed timelines, which translates into a staggering daily revenue loss of up to US $8 million for pharmaceutical behemoths [75]. Furthermore, an annual expenditure nearing US $6 billion is dedicated to the pursuit of patient recruitment. Remarkably, merely 2% of the eligible population in the United States partakes in these clinical odysseys and those brave souls who do so endure an average of 11 visits to the trial site within the span of 6 months [79].

From an economic standpoint, there looms the tantalizing prospect of savings entwined with digital clinical trials. This progressive approach is geared toward streamlining the study

XSL•FO
RenderX

duration, curtailing the recruitment period, and accelerating the data collection process. Furthermore, it obliterates the conventional need for numerous physical study sites, offering an alluring window into substantial capital savings, with estimated costs of managing a solitary digital clinical trials study site fluctuating between US $1500 and US $2500 per month [75]. Additionally, the demand to reimburse patients for their travel-related expenses dwindles, providing an enticing financial incentive. However, it is imperative to acknowledge that this paradigm shift is contingent upon robust technological support, which is the cornerstone upon which these cost-saving benefits are built.

The creation of digital populations brings forth a fresh perspective, coupled with inventive techniques and methodologies. Over recent years, the concept of generating digital populations through data augmentation of existing data sets has been steadily gaining traction. One notable example is the Synthea project [80], which simulates EHR data based on the fundamental demographics of the Massachusetts population and specific disease models. In a similar vein, endeavors have been made to craft digital patients with precise measurements, such as glucose levels, either through mathematical models [62] or by incorporating highly specific attributes [81]. A digital patient data set is constructed by leveraging real-world data, including demographics, laboratory findings, and anatomical features. Subsequently, using patient-specific disease progression models, diverse vascular models can be generated and explored [82].

In situations where an extensive patient data set is available, clinical trial simulations necessitate the careful selection of a representative subset from the original pool. However, when dealing with limited data sets, the imperative arises to augment the existing data. This process, which introduces digital patients into the original data set, serves as an invaluable technique. Its reliability shines through, particularly in data sets characterized by non-Gaussian distributions of covariates [83].

Conventional clinical trial designs often overlook the intricate patient diversity and intricacies. The inherent heterogeneity of patients enrolled in clinical trials manifests as a correspondingly wide spectrum of responses to stent implantation. Enter the era of digital populations applied to the realm of novel stent design, poised to enhance patient safety, slash the costs of clinical trials, and ultimately usher in a new era of clinical practice [84]. Thus, highly efficient stents with minimized side effects, promise a brighter future for patients.

Remote Patient Monitoring and Telemedicine

Digital twins can enable remote patient monitoring and telemedicine, revolutionizing the way health care is delivered and accessed [21,85]. By leveraging wearable sensors, IoT devices, and advanced technologies, digital twins enable health care providers to remotely monitor patients, offer timely interventions, and enhance the quality of care [86-88]. Thus, digital twins facilitate continuous monitoring of patient health parameters, regardless of the patient's location. By integrating data from wearable devices, sensors, and other IoT-enabled devices, digital twins collect real-time information on vital signs, medication adherence, and other relevant health data. This continuous monitoring enables health care providers to stay informed about a patient's condition, detect anomalies, and intervene promptly when necessary [6,89].

Digital twins enable the seamless transmission of patient data from remote locations to health care providers [23]. Through secure data channels, patient information is sent to the digital twin, which can analyze and interpret the data in real time [90]. This remote data analysis provides health care professionals with valuable insights into a patient's health status, allowing for remote assessment and decision-making [64,91]. By analyzing the data collected through remote monitoring, digital twins can also identify trends, patterns, and potential health risks specific to an individual patient [28,68]. This information enables health care professionals to tailor interventions, adjust treatment plans, and provide personalized guidance to patients, enhancing the effectiveness and efficiency of care delivery.

Remote patient monitoring and telemedicine facilitated by digital twins can help reduce hospitalizations and readmissions [92,93]. By closely monitoring patients' health at home or in nonacute care settings, health care providers can detect early signs of deterioration or complications. Timely interventions can then be initiated, preventing the need for hospitalization, or reducing the length of hospital stays. This approach also supports the transition from hospital to home, ensuring continuity of care and reducing the likelihood of readmissions [5,6].

The promotion of patient engagement and self-management could be enhanced by digital twins by empowering individuals to take an active role in their own health care. By providing access to their own health data, patients can monitor their progress, track their vital signs, and gain insights into their health conditions [2,27,85]. Digital twins can also deliver personalized recommendations, reminders, and educational materials to patients, supporting self-care and adherence to treatment plans [21,24]. Through videoconferencing, secure messaging platforms, and real-time data sharing, digital twins can facilitate digital visits, allowing patients to consult with health care professionals from the comfort of their homes [51]. This remote access to health care services improves convenience and accessibility, and reduces the need for physical appointments, particularly for patients in rural or underserved areas.

Digital twins support long-term disease management by providing continuous monitoring and personalized interventions [56,64]. For patients with chronic conditions, digital twins enable health care providers to remotely track disease progression, assess treatment effectiveness, and adjust therapies as needed. This proactive approach helps patients maintain optimal health, prevent complications, and reduce the burden of frequent hospital visits.

Conversely, a pivotal challenge in the implementation of digital twins lies in the absence of seamless connectivity among various systems and medical devices within the Digital Health Twin framework. The solution to this quandary lies in the widespread adoption of standardized data representation and exchange

XSL•FO
RenderX

protocols. These protocols serve as the linchpin for facilitating smooth communication and the harmonious integration of EHRs, medical devices, and assorted health care systems [94]. This newfound interoperability, in turn, fosters the exchange of patient-specific information, elevating the precision and comprehensiveness of the DHT model [95].

The successful implementation of digital twins necessitates a tapestry of expertise, a multidisciplinary symphony where physicians, radiologists, image processing virtuosos, molecular biologists, geneticists, bioinformaticians, computer scientists, data wizards, and engineers converge. Their collective efforts are pivotal in confronting the intricate challenges inherent in crafting a digital counterpart that is both accurate and dependable [96].

Challenges for Digital Twin Implementation

The digital twin market faces formidable challenges that could impede its growth (Table 2). These obstacles encompass the high deployment costs, surging demands for power and storage, integration issues with existing systems and proprietary software, as well as the intricate nature of its architecture. The implementation of digital twin solutions is a costly endeavor, necessitating substantial investments in technology platforms (comprising sensors and software), infrastructure development, maintenance, data quality control, and security solutions. Digital twin constantly collects, analyses, and accumulates data from physical space to provide sufficient information for decision-making. Thus, the challenge of data integrity and security remains major. This includes measures like data encryption, secure data storage, and regular backups for privacy data used by digital twin models. By real-time data perception of dynamic environment and high accuracy model, digital twin should include regular control processes for performance prediction. Moreover, the upkeep of the digital twin infrastructure incurs significant operational expenses. The high fixed costs and the complexity of digital twin architectures are anticipated to decelerate the adoption of digital twin technologies. Digital twins pose a formidable challenge in their demand for rich, extensive data sets and innovative EHR designs that facilitate data mining and the automated acquisition of pristine data. Currently, one of the major impediments to human digital twins is the glaring heterogeneity and operational intricacies found in EHRs and health care information systems [12]. Furthermore, these data often reside in an unstructured format, necessitating either manual intervention or the deployment of advanced automation through natural language processing technologies to extract the required information.

Table 2. Challenges for digital twin implementation for personalized medicine.

Challenge	Description
Challenge 1	Methodological hazards were a noted challenge to using artificial intelligence for inductive reasoning; for example, the generalizability of findings necessitates external validation with new patient cohorts, or cohorts from different centers or different geographical locations, and across time.
Challenge 2	The Internet of Things as applied to health care presents challenges for devices with processor, memory, and energy limitations.
Challenge 3	Transformational technologies present demands on digital twin software for confidentiality, reliability, safety, and secure coding, with minimal requirements for patching.
Challenge 4	The credibility of digital patient models to predict disease risk and progression in a real patient, and the trust required of the computational processes that deliver these, presents a potential barrier to their uptake into a routine workflow.
Challenge 5	Creating a digital twin of a patient for precision medicine raises considerable ethical questions around its legacy, privacy, and identity; and its termination when the real twin dies.
Challenge 6	Potential regulatory and legal issues for a health digital twin are yet uncertain but are likely to be especially demanding for approval of devices associated with medical cyber-physical systems that contain large amounts of embedded software for sensing and monitoring people's activities.

The quality of the data plays a paramount role. Although sensors are adept at efficiently collecting and transmitting data to human digital twins, the processes involved in gathering hospital data can be both expensive and time-consuming [96]. Presently, many individual data are procured through blood tests, imaging systems, and health scans. Consequently, these hospital data collection procedures place a considerable burden on digital twin processes. For instance, achieving top-tier image quality in computerized tomography scans of cardiac patients is no straightforward task, and the results often hinge on the expertise of radiology personnel, especially those with limited experience. Experts in the field argue that the future milestones in digital twins will not revolve around the advancements in AI research but rather focus on rectifying the issues associated with small-scale, unorganized health care data [12].

While digital twin applications have been portrayed as fully autonomous processes, there is an essential need for interdisciplinary knowledge spanning fields such as biomedicine, mathematics, bioengineering, and computer science, as well as insights derived from people's experiences, given the intricate nature of human beings [96]. Moreover, digital twin software developers should prioritize the creation of user-friendly interfaces for digital twins to facilitate communication between digital twin software, patients, and physicians. These interfaces should enable discussions on optimal treatment based on informed consent. Nevertheless, experts in the field have identified a dearth of user-friendly software for digital twin applications in the realm of health care [97].

Physicians continue to harbor reservations about placing trust in decisions derived from algorithms and big data, primarily

XSL•FO
RenderX

because these predictions often lack a plausible or transparent explanation. Recent research on the integration of AI systems in the hospital environment reveals that many physicians remain skeptical of AI due to the significant risk associated with potential misdiagnoses and inappropriate treatment decisions. The apprehension of clinicians being replaced may surface with the broader use of digital twins in clinical tasks [12]. In certain scenarios, digital twins may surpass clinicians in performance, as it is highly improbable that a clinician could process all patient data and provide a solution during a brief consultation. Nonetheless, given the prevailing mistrust of digital twin decision points and the current state of digital twin applications in health care, digital twins will strive to adapt to the needs and workflows of clinicians in the future. This adaptation aims to enhance their capacity to efficiently consider the entire spectrum of available information when making decisions.

Collaboration among researchers, health care providers, and technology developers is essential for advancing the field of digital twins in cardiovascular disease management. Data sharing initiatives, research consortia, and interdisciplinary collaborations can accelerate innovation, improve model accuracy, and ensure the ethical and responsible use of digital twin technology [46]. The use of digital twins in health gives rise to ethical dilemmas that require several considerations. Upholding principles such as informed consent, data ownership, and patient autonomy is of utmost importance to guarantee the responsible and ethical integration of digital twins. Establishing well-defined directives and regulatory frameworks becomes imperative for effectively addressing the ethical quandaries stemming from the use of individuals' personal health data for digital twin modeling. Maintaining reverence for patients' rights and upholding transparency at all stages is crucial for cultivating a bond of trust between health care providers and their patients.

To date, there has been no comprehensive approach to validating digital twin models. Validating digital twins and simulation models, in general, presents several formidable challenges. While simulations can be validated using retrospective longitudinal data, alternative scenarios are often absent from the ground truth. To instill confidence in the simulation, it becomes imperative to compare the simulated model averages with a separate benchmark. The intricate interplay and diversity within human physiology, disease progression, and individual patient characteristics complicate the task of ensuring an accurate representation of real-world patients. The pursuit of dependable results from digital twins necessitates the implementation of rigorous testing, validation methodologies, and clinical studies. Continual refinement and validation of these models against real-world patient data assume pivotal roles in enhancing their predictive capabilities and expanding their clinical utility [98]. Sensitivity analysis, which model explanations are a subset of, enables the ascription of change in parameters to the outcomes [99]. This empowers a domain expert to identify the most influential factors affecting the outcomes. Subsequently, the domain expert, in our case, the hospital's operational leadership team, can decide which levers to use to optimize the results. Nevertheless, conducting sensitivity analysis on large-scale simulations becomes computationally challenging due to the numerous parameters involved. [100]. By using machine learning models to train simulated data, we can test the global sensitivity of the model parameters through attribution.

Conclusions

Digital twins represent a groundbreaking approach to personalized medicine, leveraging digital replicas of patients to optimize diagnostics, treatment strategies, and health care outcomes. Through enhanced understanding of patient health, predictive modeling, digital clinical trials, and remote patient monitoring, digital twins pave the way for more precise, individualized health care interventions. While challenges exist, the continued development and integration of digital twins hold the potential to revolutionize personalized medicine, ushering in an era of tailored treatments and improved patient well-being. Nevertheless, the use of human digital twins gives rise to ethical dilemmas related to informed consent, data ownership, and the potential for discrimination based on health profiles. As we stand on the brink of a new era in medicine, the integration of digital twins offers both exhilarating possibilities and formidable challenges. It is imperative that we, as a medical community, proactively shape ethical guidelines and regulations to harness their full potential while safeguarding patient welfare.

Authors' Contributions

AV was responsible for conceptualization and writing the original draft of the manuscript. The author has read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Boulos MNK, Zhang P. Digital twins: from personalised medicine to precision public health. J Pers Med 2021;11(8):745-745 [FREE Full text] [doi: 10.3390/jpm11080745] [Medline: 34442389]
2. Venkatesh KP, Raza MM, Kvedar JC. Health digital twins as tools for precision medicine: considerations for computation, implementation, and regulation. NPJ Digit Med 2022;5(1):150-150 [FREE Full text] [doi: 10.1038/s41746-022-00694-7] [Medline: 36138125]

XSL•FO
RenderX

3. Vallée A. Digital twin for healthcare systems. Front Digit Health 2023 Jul;5:1253050-1253050 [FREE Full text] [doi: 10.3389/fdgth.2023.1253050] [Medline: 37744683]
4. Sahal R, Alsamhi SH, Brown KN. Personal digital twin: a close look into the present and a step towards the future of personalised healthcare industry. Sensors (Basel) 2022 Aug;22(15):5918-5953 [FREE Full text] [doi: 10.3390/s22155918] [Medline: 35957477]
5. Armeni P, Polat I, De Rossi LM, Diaferia L, Meregalli S, Gatti A. Digital twins in healthcare: is it the beginning of a new era of evidence-based medicine? a critical review. J Pers Med 2022 Dec;12(8):1255-1269 [FREE Full text] [doi: 10.3390/jpm12081255] [Medline: 36013204]
6. Elkefi S, Asan O. Digital twins for managing health care systems: rapid literature review. J Med Internet Res 2022;24(8):e37641-e37641 [FREE Full text] [doi: 10.2196/37641] [Medline: 35972776]
7. Trenfield SJ, Awad A, McCoubrey LE, Elbadawi M, Goyanes A, Gaisford S, et al. Advancing pharmacy and healthcare with virtual digital technologies. Adv Drug Deliv Rev 2022;182:114098-114098. [doi: 10.1016/j.addr.2021.114098] [Medline: 34998901]
8. Zhou X, Wu X. Teaching mode based on educational big data mining and digital twins. Comput Intell Neurosci 2022;2022:9071944-9071944 [FREE Full text] [doi: 10.1155/2022/9071944] [Medline: 35222637]
9. Zhong D, Xia Z, Zhu Y, Duan J. Overview of predictive maintenance based on digital twin technology. Heliyon 2023;9(4):e14534-e14534 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e14534] [Medline: 37025897]
10. Qiao Q, Wang J, Ye L, Gao RX. Digital twin for machining tool condition prediction. Procedia CIRP 2019;81:1388-1393 [FREE Full text] [doi: 10.1016/j.procir.2019.04.049]
11. Lal A, Li G, Cubro E, Chalmers S, Li H, Herasevich V, et al. Development and verification of a digital twin patient model to predict specific treatment response during the first 24 hours of sepsis. Crit Care Explor 2020;2(11):e0249-e0249 [FREE Full text] [doi: 10.1097/CCE.0000000000000249] [Medline: 33225302]
12. Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, et al. The 'digital twin' to enable the vision of precision cardiology. Eur Heart J 2020;41(48):4556-4564 [FREE Full text] [doi: 10.1093/eurheartj/ehaa159] [Medline: 32128588]
13. Liu Y, Zhang L, Yang Y, Zhou L, Ren L, Wang F, et al. A novel cloud-based framework for the elderly healthcare services using digital twin. IEEE Access 2019;7:49088-49101 [FREE Full text] [doi: 10.1109/access.2019.2909828]
14. Strianese O, Rizzo F, Ciccarelli M, Galasso G, D'Agostino Y, Salvati A, et al. Precision and personalized medicine: how genomic approach improves the management of cardiovascular and neurodegenerative disease. Genes (Basel) 2020;11(7):747 [FREE Full text] [doi: 10.3390/genes11070747] [Medline: 32640513]
15. Pashayan N, Antoniou AC, Ivanus U, Esserman LJ, Easton DF, French D, et al. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. Nat Rev Clin Oncol 2020;17(11):687-705 [FREE Full text] [doi: 10.1038/s41571-020-0388-9] [Medline: 32555420]
16. An G, Cockrell C. Drug development digital twins for drug discovery, testing and repurposing: a schema for requirements and development. Front Syst Biol 2022;2:928387 [FREE Full text] [doi: 10.3389/fsysb.2022.928387] [Medline: 35935475]
17. Moingeon P, Chenel M, Rousseau C, Voisin E, Guedj M. Virtual patients, digital twins and causal disease models: paving the ground for in silico clinical trials. Drug Discov Today 2023;28(7):103605. [doi: 10.1016/j.drudis.2023.103605] [Medline: 37146963]
18. Volkov I, Radchenko G, Tchernykh A. Digital twins, internet of things and mobile medicine: a review of current platforms to support smart healthcare. Program Comput Soft 2021;47(8):578-590 [FREE Full text] [doi: 10.1134/s0361768821080284]
19. Sun T, He X, Li Z. Digital twin in healthcare: recent updates and challenges. Digit Health 2023;9:20552076221149651 [FREE Full text] [doi: 10.1177/20552076221149651] [Medline: 36636729]
20. Haleem A, Javaid M, Singh RP, Suman R. Exploring the revolution in healthcare systems through the applications of digital twin technology. Biomed Technol 2023;4:28-38 [FREE Full text] [doi: 10.1016/j.bmt.2023.02.001]
21. Sun T, He X, Song X, Shu L, Li Z. The digital twin in medicine: a key to the future of healthcare? Front Med (Lausanne) 2022;9:907066 [FREE Full text] [doi: 10.3389/fmed.2022.907066] [Medline: 35911407]
22. Li L, Lei B, Mao C. Digital twin in smart manufacturing. J Ind Inf Integr 2022;26:100289. [doi: 10.1016/j.jii.2021.100289]
23. El Saddik A, Laamarti F, Alja'Afreh M. The potential of digital twins. IEEE Instrum Meas Mag 2021;24(3):36-41. [doi: 10.1109/mim.2021.9436090]
24. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, et al. Digital twins to personalize medicine. Genome Med 2019;12(1):4 [FREE Full text] [doi: 10.1186/s13073-019-0701-3] [Medline: 31892363]
25. Peirlinck M, Costabal FS, Yao J, Guccione JM, Tripathy S, Wang Y, et al. Precision medicine in human heart modeling: perspectives, challenges, and opportunities. Biomech Model Mechanobiol 2021;20(3):803-831 [FREE Full text] [doi: 10.1007/s10237-021-01421-z] [Medline: 33580313]
26. Tao F, Qi Q. Make more digital twins. Nature 2019;573(7775):490-491. [doi: 10.1038/d41586-019-02849-1] [Medline: 31554984]
27. Schwartz SM, Wildenhaus K, Bucher A, Byrd B. Digital twins and the emerging science of self: implications for digital health experience design and “small” data. Front Comput Sci 2020;2:31 [FREE Full text] [doi: 10.3389/fcomp.2020.00031]

XSL•FO
RenderX

28. Popa EO, van Hilten M, Oosterkamp E, Bogaardt MJ. The use of digital twins in healthcare: socio-ethical benefits and socio-ethical risks. Life Sci Soc Policy 2021;17(1):6 [FREE Full text] [doi: 10.1186/s40504-021-00113-x] [Medline: 34218818]
29. Gkouskou K, Vlastos I, Karkalousos P, Chaniotis D, Sanoudou D, Eliopoulos AG. The "virtual digital twins" concept in precision nutrition. Adv Nutr 2020;11(6):1405-1413 [FREE Full text] [doi: 10.1093/advances/nmaa089] [Medline: 32770212]
30. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
31. Coorey G, Figtree GA, Fletcher DF, Snelson VJ, Vernon ST, Winlaw D, et al. The health digital twin to tackle cardiovascular disease-a review of an emerging interdisciplinary field. NPJ Digit Med 2022;5(1):126 [FREE Full text] [doi: 10.1038/s41746-022-00640-7] [Medline: 36028526]
32. Boyes H, Watson T. Digital twins: an analysis framework and open issues. Comput Ind 2022;143:103763 [FREE Full text] [doi: 10.1016/j.compind.2022.103763]
33. Elayan H, Aloqaily M, Guizani M. Digital twin for intelligent context-aware IoT healthcare systems. IEEE Internet Things J 2021;8(23):16749-16757. [doi: 10.1109/jiot.2021.3051158]
34. Hardavella G, Aamli-Gaagnat A, Saad N, Rousalova I, Sreter KB. How to give and receive feedback effectively. Breathe (Sheff) 2017;13(4):327-333 [FREE Full text] [doi: 10.1183/20734735.009917] [Medline: 29209427]
35. Karakra A, Fontanili F, Lamine E, Lamothe J, Taweel A. Pervasive computing integrated discrete event simulation for a hospital digital twin. 2018 Presented at: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA); October 28, 2018-November 01, 2018; Aqaba, Jordan p. 1-6. [doi: 10.1109/aiccsa.2018.8612796]
36. Kaul R, Ossai C, Forkan ARM, Jayaraman PP, Zelcer J, Vaughan S, et al. The role of AI for developing digital twins in healthcare: the case of cancer care. WIREs Data Min Knowl Discov 2022;13(1):e1480 [FREE Full text] [doi: 10.1002/widm.1480]
37. McBride CM, Alford SH, Reid RJ, Larson EB, Baxevanis AD, Brody LC. Characteristics of users of online personalized genomic risk assessments: implications for physician-patient interactions. Genet Med 2009;11(8):582-587 [FREE Full text] [doi: 10.1097/GIM.0b013e3181b22c3a] [Medline: 19606049]
38. McClellan KA, Avard D, Simard J, Knoppers BM. Personalized medicine and access to health care: potential for inequitable access? Eur J Hum Genet 2013;21(2):143-147 [FREE Full text] [doi: 10.1038/ejhg.2012.149] [Medline: 22781088]
39. Batch KE, Yue J, Darcovich A, Lupton K, Liu CC, Woodlock DP, et al. Developing a cancer digital twin: supervised metastases detection from consecutive structured radiology reports. Front Artif Intell 2022;5:826402 [FREE Full text] [doi: 10.3389/frai.2022.826402] [Medline: 35310959]
40. Subramanian B, Kim J, Maray M, Paul A. Digital twin model: a real-time emotion recognition system for personalized healthcare. IEEE Access 2022;10:81155-81165 [FREE Full text] [doi: 10.1109/access.2022.3193941]
41. Kim JK, Lee SJ, Hong SH, Choi IY. Machine-learning-based digital twin system for predicting the progression of prostate cancer. Appl Sci 2022;12(16):8156 [FREE Full text] [doi: 10.3390/app12168156]
42. Chakshu NK, Sazonov I, Nithiarasu P. Towards enabling a cardiovascular digital twin for human systemic circulation using inverse analysis. Biomech Model Mechanobiol 2021;20(2):449-465 [FREE Full text] [doi: 10.1007/s10237-020-01393-6] [Medline: 33064221]
43. Hussain I, Hossain MA, Park SJ. A healthcare digital twin for diagnosis of stroke. 2021 Presented at: 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON); December 04-05, 2021; Dhaka, Bangladesh p. 18-21. [doi: 10.1109/becithcon54710.2021.9893641]
44. Xu Y, Sun Y, Liu X, Zheng Y. A digital-twin-assisted fault diagnosis using deep transfer learning. IEEE Access 2019;7:19990-19999 [FREE Full text] [doi: 10.1109/access.2018.2890566]
45. Martinez-Velazquez R, Gamez R, El Saddik S. Cardio twin: a digital twin of the human heart running on the edge. 2019 Presented at: 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA); June 26-28, 2019; Istanbul, Turkey p. 1-6. [doi: 10.1109/memea.2019.8802162]
46. Stahlberg EA, Abdel-Rahman M, Aguilar B, Asadpoure A, Beckman RA, Borkon LL, et al. Exploring approaches for predictive cancer patient digital twins: opportunities for collaboration and innovation. Front Digit Health 2022;4:1007784 [FREE Full text] [doi: 10.3389/fdgth.2022.1007784] [Medline: 36274654]
47. Rose SMSF, Contrepois K, Moneghetti KJ, Zhou W, Mishra T, Mataraso S, et al. A longitudinal big data approach for precision health. Nat Med 2019;25(5):792-804. [doi: 10.1038/s41591-019-0414-6] [Medline: 31068711]
48. Bodai BI, Nakata TE, Wong WT, Clark DR, Lawenda S, Tsou C, et al. Lifestyle medicine: a brief review of its dramatic impact on health and survival. Perm J 2018;22:17-025 [FREE Full text] [doi: 10.7812/TPP/17-025] [Medline: 29035175]
49. Nagaraj D, Khandelwal P, Steyaert S, Gevaert O. Augmenting digital twins with federated learning in medicine. Lancet Digit Health 2023;5(5):e251-e253 [FREE Full text] [doi: 10.1016/S2589-7500(23)00044-4] [Medline: 37100540]
50. Moztarzadeh O, Jamshidi MB, Sargolzaei S, Jamshidi A, Baghalipour N, Moghani MM, et al. Metaverse and healthcare: machine learning-enabled digital twins of cancer. Bioengineering (Basel) 2023;10(4):455 [FREE Full text] [doi: 10.3390/bioengineering10040455] [Medline: 37106642]
51. Attaran M, Celik BG. Digital twin: benefits, use cases, challenges, and opportunities. Decis Anal J 2023;6:100165 [FREE Full text] [doi: 10.1016/j.dajour.2023.100165]

XSL•FO
RenderX

52. Alotaibi YK, Federico F. The impact of health information technology on patient safety. Saudi Med J 2017;38(12):1173-1180 [FREE Full text] [doi: 10.15537/smj.2017.12.20631] [Medline: 29209664]
53. Masic I, Miokovic M, Muhamedagic B. Evidence based medicine—new approaches and challenges. Acta Inform Med 2008;16(4):219-225. [doi: 10.5455/aim.2008.16.219-225] [Medline: 24109156]
54. van Dinter R, Tekinerdogan B, Catal C. Predictive maintenance using digital twins: a systematic literature review. Inf Softw Technol 2022;151:107008 [FREE Full text] [doi: 10.1016/j.infsof.2022.107008]
55. Brooks JD, Nabi HH, Andrulis IL, Antoniou AC, Chiquette J, Després P, et al. Personalized risk assessment for prevention and early detection of breast cancer: integration and implementation (PERSPECTIVE I and I). J Pers Med 2021;11(6):511 [FREE Full text] [doi: 10.3390/jpm11060511] [Medline: 34199804]
56. Calcaterra V, Pagani V, Zuccotti G. Digital twin: a future health challenge in prevention, early diagnosis and personalisation of medical care in paediatrics. Int J Environ Res Public Health 2023;20(3):2181 [FREE Full text] [doi: 10.3390/ijerph20032181] [Medline: 36767547]
57. Bruynseels K, de Sio FS, van den Hoven J. Digital twins in health care: ethical implications of an emerging engineering paradigm. Front Genet 2018;9:31 [FREE Full text] [doi: 10.3389/fgene.2018.00031] [Medline: 29487613]
58. Aluvalu R, Mudrakola S, Maheswari VU, Kaladevi AC, Sandhya MVS, Bhat CR. The novel emergency hospital services for patients using digital twins. Microprocess Microsyst 2023;98:104794. [doi: 10.1016/j.micpro.2023.104794]
59. Iqbal JD, Krauthammer M, Biller-Andorno N. The use and ethics of digital twins in medicine. J Law Med Ethics 2022;50(3):583-596 [FREE Full text] [doi: 10.1017/jme.2022.97] [Medline: 36398633]
60. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023;29(1):49-58 [FREE Full text] [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
61. Teutonico D, Musuamba F, Maas HJ, Facius A, Yang S, Danhof M, et al. Generating virtual patients by multivariate and discrete re-sampling techniques. Pharm Res 2015;32(10):3228-3237 [FREE Full text] [doi: 10.1007/s11095-015-1699-x] [Medline: 25994981]
62. Resalat N, El Youssef J, Tyler N, Castle J, Jacobs PG. A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model. PLoS One 2019;14(7):e0217301. [doi: 10.1371/journal.pone.0217301] [Medline: 31344037]
63. Urresti-Gundlach M, Tolks D, Kiessling C, Wagner-Menghin M, Härtl A, Hege I. Do virtual patients prepare medical students for the real world? development and application of a framework to compare a virtual patient collection with population data. BMC Med Educ 2017;17(1):174 [FREE Full text] [doi: 10.1186/s12909-017-1013-1] [Medline: 28938884]
64. Voigt I, Inojosa H, Dillenseger A, Haase R, Akgün K, Ziemssen T. Digital twins for multiple sclerosis. Front Immunol 2021;12:669811 [FREE Full text] [doi: 10.3389/fimmu.2021.669811] [Medline: 34012452]
65. Mulder ST, Omidvari AH, Rueten-Budde AJ, Huang PH, Kim KH, Bais B, et al. Dynamic digital twin: diagnosis, treatment, prediction, and prevention of disease during the life course. J Med Internet Res 2022;24(9):e35675 [FREE Full text] [doi: 10.2196/35675] [Medline: 36103220]
66. Spitzer M, Dattner I, Zilcha-Mano S. Digital twins and the future of precision mental health. Front Psychiatry 2023;14:1082598 [FREE Full text] [doi: 10.3389/fpsyt.2023.1082598] [Medline: 36993921]
67. Whitehead D, Conley J. The next frontier of remote patient monitoring: hospital at home. J Med Internet Res 2023;25:e42335 [FREE Full text] [doi: 10.2196/42335] [Medline: 36928088]
68. Huang PH, Kim KH, Schermer M. Ethical issues of digital twins for personalized health care service: preliminary mapping study. J Med Internet Res 2022;24(1):e33081 [FREE Full text] [doi: 10.2196/33081] [Medline: 35099399]
69. Chase JG, Preiser JC, Dickson JL, Pironet A, Chiew YS, Pretty CG, et al. Next-generation, personalised, model-based critical care medicine: a state-of-the art review of in silico virtual patient models, methods, and cohorts, and how to validation them. Biomed Eng Online 2018;17(1):24 [FREE Full text] [doi: 10.1186/s12938-018-0455-y] [Medline: 29463246]
70. Biermann E, Barkhausen K, Standl E. How would patients behave if they were continually informed of their blood glucose levels? a simulation study using a "virtual" patient. Diabetes Technol Ther 2008;10(3):178-187. [doi: 10.1089/dia.2007.0281] [Medline: 18473691]
71. Lamarque G, Ravier P, Dumez-Viou C. A new concept of virtual patient for real-time ECG analyzers. IEEE Trans Instrum Meas 2011;60(3):939-946. [doi: 10.1109/tim.2010.2064610]
72. Plackett R, Kassianos AP, Kambouri M, Kay N, Mylan S, Hopwood J, et al. Online patient simulation training to improve clinical reasoning: a feasibility randomised controlled trial. BMC Med Educ 2020;20(1):245 [FREE Full text] [doi: 10.1186/s12909-020-02168-4] [Medline: 32736583]
73. Liaw SY, Chan SWC, Chen FG, Hooi SC, Siau C. Comparison of virtual patient simulation with mannequin-based simulation for improving clinical performances in assessing and managing clinical deterioration: randomized controlled trial. J Med Internet Res 2014;16(9):e214 [FREE Full text] [doi: 10.2196/jmir.3322] [Medline: 25230684]
74. Inan OT, Tenaerts P, Prindiville SA, Reynolds HR, Dizon DS, Cooper-Arnold K, et al. Digitizing clinical trials. NPJ Digit Med 2020;3(1):101 [FREE Full text] [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]
75. Ali Z, Zibert JR, Thomsen SF. Virtual clinical trials: perspectives in dermatology. Dermatology 2020;236(4):375-382 [FREE Full text] [doi: 10.1159/000506418] [Medline: 32126560]

XSL•FO
RenderX

76. Prasad V, Mailankody S. Research and development spending to bring a single cancer drug to market and revenues after approval. JAMA Intern Med 2017;177(11):1569-1575 [FREE Full text] [doi: 10.1001/jamainternmed.2017.3601] [Medline: 28892524]
77. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R and D costs. J Health Econ 2016;47:20-33. [doi: 10.1016/j.jhealeco.2016.01.012] [Medline: 26928437]
78. Ledford H. Translational research: 4 ways to fix the clinical trial. Nature 2011;477(7366):526-528. [doi: 10.1038/477526a] [Medline: 21956311]
79. Comis RL, Miller JD, Aldigé CR, Krebs L, Stoval E. Public attitudes toward participation in cancer clinical trials. J Clin Oncol 2003;21(5):830-835. [doi: 10.1200/JCO.2003.02.105] [Medline: 12610181]
80. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018;25(3):230-238 [FREE Full text] [doi: 10.1093/jamia/ocx079] [Medline: 29025144]
81. Hejblum BP, Weber GM, Liao KP, Palmer NP, Churchill S, Shadick NA, et al. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. Sci Data 2019;6:180298 [FREE Full text] [doi: 10.1038/sdata.2018.298] [Medline: 30620344]
82. Pleouras D, Sakellarios A, Rigas G, Karanasiou GS, Tsompou P, Karanasiou G, et al. A novel approach to generate a virtual population of human coronary arteries for in silico clinical trials of stent design. IEEE Open J Eng Med Biol 2021;2:201-209 [FREE Full text] [doi: 10.1109/OJEMB.2021.3082328] [Medline: 35402969]
83. Cevenini G, Barbini P. A bootstrap approach for assessing the uncertainty of outcome probabilities when using a scoring system. BMC Med Inform Decis Mak 2010;10:45 [FREE Full text] [doi: 10.1186/1472-6947-10-45] [Medline: 20796275]
84. Chen Z, Zhang H, Guo Y, George TJ, Prosperi M, Hogan WR, et al. Exploring the feasibility of using real-world data from a large clinical data research network to simulate clinical trials of Alzheimer's disease. NPJ Digit Med 2021;4(1):84 [FREE Full text] [doi: 10.1038/s41746-021-00452-1] [Medline: 33990663]
85. Alazab M, Khan LU, Koppu S, Ramu SP, Iyapparaja M, Boobalan P, et al. Digital twins for healthcare 4.0—recent advances, architecture, and open challenges. IEEE Consum Electron Mag 2023;12(6):29-37. [doi: 10.1109/mce.2022.3208986]
86. Singh RP, Javaid M, Haleem A, Suman R. Internet of Things (IoT) applications to fight against COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):521-524. [doi: 10.1016/j.dsx.2020.04.041] [Medline: 32388333]
87. Kelly JT, Campbell KL, Gong E, Scuffham P. The internet of things: impact and implications for health care delivery. J Med Internet Res 2020;22(11):e20135 [FREE Full text] [doi: 10.2196/20135] [Medline: 33170132]
88. Shanbehzadeh M, Nopour R, Kazemi-Arpanahi H. Internet of Things (IoT) adoption model for early identification and monitoring of COVID-19 cases: a systematic review. Int J Prev Med 2022;13:112. [doi: 10.4103/ijpvm.IJPVM_667_20] [Medline: 36247189]
89. Milne-Ives M, Fraser LK, Khan A, Walker D, van Velthoven MH, May J, et al. Life course digital twins-intelligent monitoring for early and continuous intervention and prevention (LifeTIME): proposal for a retrospective cohort study. JMIR Res Protoc 2022;11(5):e35738 [FREE Full text] [doi: 10.2196/35738] [Medline: 35617022]
90. Ruppert T, Abonyi J. Integration of real-time locating systems into digital twins. J Ind Inf Integr 2020;20:100174. [doi: 10.1016/j.jii.2020.100174]
91. Hernandez-Boussard T, Macklin P, Greenspan EJ, Gryshuk AL, Stahlberg E, Syeda-Mahmood T, et al. Digital twins for predictive oncology will be a paradigm shift for precision cancer care. Nat Med 2021;27(12):2065-2066. [doi: 10.1038/s41591-021-01558-5] [Medline: 34824458]
92. Zhong X, Sarijaloo FB, Prakash A, Park J, Huang C, Barwise A, et al. A multidisciplinary approach to the development of digital twin models of critical care delivery in intensive care units. Int J Prod Res 2022;60(13):4197-4213. [doi: 10.1080/00207543.2021.2022235]
93. Lonsdale H, Gray GM, Ahumada LM, Yates HM, Varughese A, Rehman MA. The perioperative human digital twin. Anesth Analg 2022;134(4):885-892 [FREE Full text] [doi: 10.1213/ANE.0000000000005916] [Medline: 35299215]
94. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast healthcare interoperability resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
95. Kukushkin K, Ryabov Y, Borovkov A. Digital twins: a systematic literature review based on data analysis and topic modeling. Data 2022;7(12):173 [FREE Full text] [doi: 10.3390/data7120173]
96. Erol T, Mendi AF, Doğan D. The digital twin revolution in healthcare. 2020 Presented at: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); October 22-24, 2020; Istanbul, Turkey p. 1-7. [doi: 10.1109/ismsit50672.2020.9255249]
97. James L. Digital twins will revolutionise healthcare: digital twin technology has the potential to transform healthcare in a variety of ways—improving the diagnosis and treatment of patients, streamlining preventative care and facilitating new approaches for hospital planning. Eng Technol 2021;16(2):50-53. [doi: 10.1049/et.2021.0210]
98. Turab M, Jamil S. A comprehensive survey of digital twins in healthcare in the era of metaverse. BioMedInformatics 2023;3(3):563-584 [FREE Full text] [doi: 10.3390/biomedinformatics3030039]

XSL•FO
RenderX

99. Niida A, Hasegawa T, Miyano S. Sensitivity analysis of agent-based simulation utilizing massively parallel computation and interactive data visualization. PLoS One 2019;14(3):e0210678. [doi: 10.1371/journal.pone.0210678] [Medline: 30835730]
100. Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 2001;55(1):271-280. [doi: 10.1016/s0378-4754(00)00270-6]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
IoT: Internet of Things

Edited by A Mavragani; submitted 22.06.23; peer-reviewed by S Mitra, KL Mauco; comments to author 19.09.23; revised version received 01.10.23; accepted 29.12.23; published 13.05.24.

Please cite as:
Vallée A
Envisioning the Future of Personalized Medicine: Role and Realities of Digital Twins
J Med Internet Res 2024;26:e50204
URL: https://www.jmir.org/2024/1/e50204
doi:10.2196/50204
PMID:38739913

©Alexandre Vallée. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Power of Rapid Reviews for Bridging the Knowledge-to-Action Gap in Evidence-Based Virtual Health Care

Megan MacPherson[1], PhD; Sarah Rourke[1], MSN

Fraser Health, Surrey, BC, Canada

Corresponding Author:
Megan MacPherson, PhD
Fraser Health
400-13450 102nd Avenue
Surrey, BC, V3T 0H1
Canada
Phone: 1 6045616605
Email: meganmargaretmacpherson@gmail.com

Abstract

Despite the surge in popularity of virtual health care services as a means of delivering health care through technology, the integration of research evidence into practice remains a challenge. Rapid reviews, a type of time-efficient evidence synthesis, offer a potential solution to bridge the gap between knowledge and action. This paper aims to highlight the experiences of the Fraser Health Authority's Virtual Health team in conducting rapid reviews. This paper discusses the experiences of the Virtual Health team in conducting 15 rapid reviews over the course of 1.5 years and the benefit of involving diverse stakeholders including researchers, project and clinical leads, and students for the creation of user-friendly knowledge products to summarize results. The Virtual Health team found rapid reviews to be a valuable tool for evidence-informed decision-making in virtual health care. Involving stakeholders and focusing on implementation considerations are crucial for maximizing the impact of rapid reviews. Health care decision makers are encouraged to consider implementing rapid review processes to improve the translation of research evidence into practice, ultimately enhancing patient outcomes and promoting a culture of evidence-informed care.

(J Med Internet Res 2024;26:e54821) doi:10.2196/54821

KEYWORDS

virtual health care; rapid reviews; evidence synthesis; evidence-informed decision-making; knowledge translation

Introduction

Background

Virtual health care services, which involve the delivery of health care through information and communication technologies, have gained popularity among health care providers, patients, and organizations. In recent decades, several initiatives have been undertaken to implement virtual care and improve the access, quality, and safety of health care delivery in Canada [1]; however, technological advancement and a rapidly expanding evidence base make supporting virtual care with research evidence challenging. Specifically, to adequately support virtual care, health care decision makers are expected to keep up with available technologies, their applications, and evidence of their effectiveness among a variety of health conditions.

Despite decision makers recognizing the need to consider research evidence in the context of public health problems [2,3], there is still a knowledge-to-action (KTA) gap between what is known and what is put into practice clinically [4-6], with health care professionals worldwide demonstrating suboptimal use of research evidence within clinical practice [7-14]. Further, it has been estimated that one-third of patients do not receive treatments that have proven efficacious, one-quarter receive treatments that are potentially harmful, and up to three-quarters of patients and half of clinicians do not receive the information necessary for research-informed decision-making [15]. Clearly, there is a need to improve the translation of research evidence into practice, particularly in the case of virtual care where technological innovations and research evidence are rapidly expanding.

Knowledge Translation

The field of knowledge translation (KT) strives to enhance the usefulness of research evidence through the design and conduct of stakeholder-informed, patient-oriented studies as well as the dissemination and implementation of research findings into practice [16]. The Canadian Institutes for Health Research defines KT as the ethical exchange, synthesis, and application of knowledge among researchers and users to accelerate the

XSL•FO
RenderX

benefits of research for Canadian people [17]. The ultimate goal of KT has been further described as the facilitation of evidence-informed decision-making [18] and the integration of various forms of evidence into public health practice and policy.

The Canadian Institutes for Health Research describes 2 "Death Valleys" on the continuum from research to action, which contributes to the KTA gap [19]. Valley 1 refers to the reduced ability to translate basic biomedical research discoveries from the laboratory to the bedside and to effectively commercialize health innovations. Valley 2 refers to the reduced ability to synthesize, disseminate, and integrate research findings more broadly into clinical practice and clinical decision-making. To improve the utility of biomedical and clinical research, enhance health outcomes, and ensure an evidence-based and sustainable health care system, strategic attempts to bridge these valleys must be made.

Rapid Reviews

One way to help overcome the second valley is through evidence syntheses such as systematic, scoping, and rapid reviews [20]. Evidence syntheses have emerged as valuable methods for KT as they can compile large bodies of evidence into a single knowledge product, making them an essential tool for decision makers to enhance evidence-informed decision-making [21,22]. Systematic reviews offer a comprehensive synthesis of available evidence on a particular topic, playing an ever-expanding role in informing policy making and practice [23,24]; however, the resource-intensive nature of conducting systematic reviews, in terms of both time and cost, presents a significant obstacle to facilitating prompt and efficient decision-making [25].

Given the time constraints health care practitioners and policy makers often face [26], rapid reviews provide a more resource- and time-efficient means to conduct evidence syntheses that offer actionable evidence in a more relevant manner compared to other types of evidence syntheses such as systematic or scoping reviews [20,26-34]. Specifically, rapid reviews are a form of evidence synthesis in which systematic review steps are streamlined to generate actionable evidence within a condensed time frame [35]. To expedite the review process, rapid reviews often compromise on the rigor typically associated with systematic reviews, resulting in a less precise and robust evaluation in comparison [32]. That being said, rapid reviews have gained traction in health systems' policy making, health-related intervention development, and health technology assessment [34-36]. This paper outlines the experiences of the Fraser Health (FH) Authority Virtual Health team in rapidly producing and disseminating rapid review results to date. Rapid reviews were chosen as they are often highly driven by end-user demands [37] and have been highlighted as a viable tool to disseminate knowledge within the rapidly growing field of virtual health [33].

FH Authority Context

As the largest regional health authority in British Columbia, Canada, FH serves more than 1.9 million people in Canada [38]. In recent years, FH has prioritized the expansion of virtual care [39], conducting over 1.9 million virtual visits between January 2019 and 2023 (roughly 27% of all visits). Within the Virtual Health department at FH, the "research and evaluation team" aims to improve the translation of research into practice while engaging in ongoing collaborative evaluation of existing Virtual Health programming. During Virtual Health strategic planning, rapid reviews have emerged as a central tool for knowledge dissemination and have been used to inform the development of frameworks, services, and program scale-up. This paper highlights FH's experience in conducting 15 rapid reviews over the course of 1.5 years. This paper is meant to serve as an overview on the utility and feasibility of rapid reviews within a health authority; for more information on rapid review methods to aid in conducting reviews within a team-based setting, see MacPherson et al [33].

Rapid reviews are used within the Virtual Health team to provide an overview of available evidence addressing a research question related to a single topic produced within a short time frame (typically 1 week to 4 months). From October 2022 until March 2024, the Virtual Health team conducted 15 rapid reviews following published recommendations [33]. Questions posed to date include the following:

1. What are the perspectives on virtual care among immigrant, refugee, and Indigenous people in Canada [40]?
2. What virtual care solutions exist for people with heart failure [41]?
3. What virtual care solutions exist for people with diabetes [41]?
4. What virtual care solutions exist for people with chronic obstructive pulmonary disease (COPD) [41]?
5. What are currently used decision guides or algorithms to inform escalation within remote patient monitoring services for people with heart failure?
6. What barriers, facilitators, and recommendations exist for remote patient monitoring services within the context of respiratory care [42]?
7. What virtual care or digital innovations are used by physicians in acute care [43]?
8. What barriers and facilitators exist for patient-to-provider virtual messaging (eg, SMS text messaging) [44]?
9. What is the existing evidence for centralized remote patient monitoring services [45]?
10. What domains are included within virtual care frameworks targeting appropriateness and safety?
11. What are patient and provider barriers to virtual care [46]?
12. What is the evidence for virtual hospital programs [47]?
13. What KT strategies exist that could be used by the Virtual Health research and evaluation team in their efforts to translate research findings into practice?
14. What is the available evidence on virtual decision-making and clinical judgment?
15. What is the available evidence for, and are there existing validated assessment criteria for nursing assessment frameworks?

Team members assisting with the rapid reviews included researchers, project leads, clinical leads, and students previously unfamiliar with the review process. Knowledge users within the Virtual Health team (eg, clinical leads and clinical directors) were involved throughout the entirety of the review process from developing the research questions to the presentation of

XSL•FO
RenderX

research findings in Virtual Health team meetings and the implementation of findings into Virtual Health practice.

Similar to other rapid reviews [20], results were collated and narratively or visually summarized (eg, through infographics) and presented to Virtual Health team members. The final knowledge products were created to offer a high-level overview of the evidence arranged in a user-friendly manner, aiming to provide VH team members with a high-level understanding of the available evidence [41].

Experiences and Lessons Learned

Overview

The Virtual Health team's journey in conducting 15 rapid reviews over the course of 1.5 years has provided valuable insights into the feasibility and utility of rapid reviews within a health authority setting. These lessons learned are from the perspectives of the authors of this paper. MM is the research and KT lead of the Virtual Health department at the FH Authority. Prior to creating the rapid review program within the Virtual Health department, she has prior experience conducting systematic, scoping, and rapid reviews. SR is a clinical nurse specialist within the Virtual Health department at FH. As a system-level leader, SR leverages evidence to informed clinical and service model changes to optimize patient care and outcomes and support strategic priorities. Prior to her involvement in the Virtual Health rapid review program, SR had no previous experience with conducting evidence reviews.

Importance of Defining a Clear and Actionable Research Question

Throughout this journey, one of the key lessons learned was about the importance of the research question being actionable to ensure that the results of rapid reviews can be readily integrated into practice. Initially, our reviews had broader scopes aimed at informing future Virtual Health service implementations across various populations such as COPD, diabetes, and heart failure. While these reviews were informative, they did not lead to immediate changes in Virtual Health practice and required strategic efforts to disseminate findings and integrate results into practice. Subsequently, we learned that focusing on specific programs or initiatives within the Virtual Health setting yields more actionable results. For instance, a review focused on identifying patient and provider barriers to virtual care was conducted with the explicit purpose of informing the development of a framework to improve video visit uptake among primary care providers. This targeted approach enabled us to directly address the identified barriers through the development of a framework focused on the uptake of safe and appropriate video visits within primary care.

Benefits and Challenges Involving Knowledge Users

The involvement of knowledge users such as clinical leads and directors in the rapid review process proved to be invaluable. First, they helped focus the scope of reviews by providing insights into the practical needs and priorities within the FH context. For example, the reviews focusing on virtual care solutions for patients with heart failure, COPD, and diabetes were initiated by 1 of the directors within Virtual Health and included an occupational therapist and clinical nurse specialist on the review team. The diverse insights offered by clinician team members helped shape the review questions, search strategy, and analysis, ensuring it addressed the practical needs in delivering virtual care to this specific patient population.

Second, the engagement of nonresearchers, students, and health care professionals in the review process not only enhanced the quality and relevance of the rapid reviews but also provided an opportunity for experiential learning and professional development. By participating in the rapid review process, students and other team members developed essential skills such as critical appraisal, evidence synthesis, and scientific communication. This approach has the potential to bridge the gap between research and practice by building a generation of clinicians who are well versed in evidence-based practice and can effectively translate research findings into clinical decision-making. For example, a team of nursing students participated in a rapid review focused on algorithms for care escalation within remote patient monitoring services for patients with heart failure. While they lacked prior review experience, their fresh perspectives and familiarity with health care practice as it relates to heart failure brought unique insights helping to shape the clinician-oriented KT efforts.

While involving knowledge users throughout the review process offers numerous benefits, it can also extend the time required to complete a review. This is often due to the necessity for these individuals to familiarize themselves with new software while simultaneously mastering the intricacies of conducting reviews and adhering to all associated steps. For instance, several Virtual Health team members have observed that during their initial and subsequent reviews, they encountered difficulties in efficiently navigating the study screening phase. The abundance of potentially relevant literature posed a challenge, with concerns arising about potentially overlooking papers containing valuable insights or "hidden gems." This underscores the importance of establishing clear eligibility criteria and providing comprehensive training from the outset to ensure reviewers feel empowered to exclude papers confidently, even those that may initially appear intriguing.

Resources and Staff Time Involved

Readers interested in starting a rapid review program in their own health systems may find it helpful to understand the resources and staff time involved in our process. As the research and KT lead within the Virtual Health team, MM has been responsible for building the rapid review program, training team members, and leading rapid reviews. Her full-time role allows for dedicated focus on these as well as other research and KT-related activities, ensuring the smooth operation of the rapid review process.

Additionally, strong leadership support within the Virtual Health team has been instrumental in fostering a culture of evidence-informed decision-making and facilitating the integration of research evidence into practice. While we do not have a core team with a dedicated full-time equivalent specifically for rapid reviews, a call is put out to the Virtual Health department at the beginning of a review to identify who has the capacity to assist in a review. A testament to the value

XSL•FO
RenderX

of these reviews is that VH team members have begun autonomously conducting rapid reviews with the research and KT lead acting as an advisor, not a lead on the reviews. For example, a nurse who was tasked with creating a framework for a virtual nursing assessment requested assistance in running a search for her team to complete a rapid review, to ensure that the resulting framework did not miss any key components seen in the literature.

Rapid Review Process

The overall process map for our team (an adaptation of MacPherson et al [33,48]) can be found in Figure 1. Our journey in conducting rapid reviews has been accompanied by several challenges and the implementation of quality assurance measures to ensure the integrity of our findings. The overall process of reviews within the Virtual Health team includes Virtual Health team members submitting a request or having an informal meeting with the research and KT lead outlining the scope and purpose of the review, which is then refined to ensure that it will result in actionable evidence relevant to the Virtual Health team and is in alignment with organizational priorities.

Challenges or obstacles encountered during the rapid review process have included resource constraints. When there are not enough people to assist with a review, either the time to complete the review needs to be extended or additional constraints must be placed on the review question. Time limitations have also been a factor, especially when there is an urgent request. Clear communication on how the results will be used is needed to refine the review topic and search strategy to quickly produce actionable evidence. Given the wealth of research, we have started all reviews by first exploring if our questions can be answered by conducting a review of reviews. This has allowed for the timely synthesis of evidence instead of relying on individual studies. We have also found that decision makers value the most up-to-date evidence (especially regarding virtual health care technologies); as such, many of our reviews have imposed limitations to the past 5-10 years to ensure their relevance to decision makers. Additionally, difficulties in accessing relevant literature have been noted, as health authorities often do not have access to the same resources as academic institutions. This results in increased time to secure papers through interlibrary loans, which can be overcome by collaborating with academics.

XSL•FO
RenderX

Figure 1. Virtual Health rapid review process map (adapted from MacPherson et al [33], which is published under Creative Commons Attribution 4.0 International License [48]).

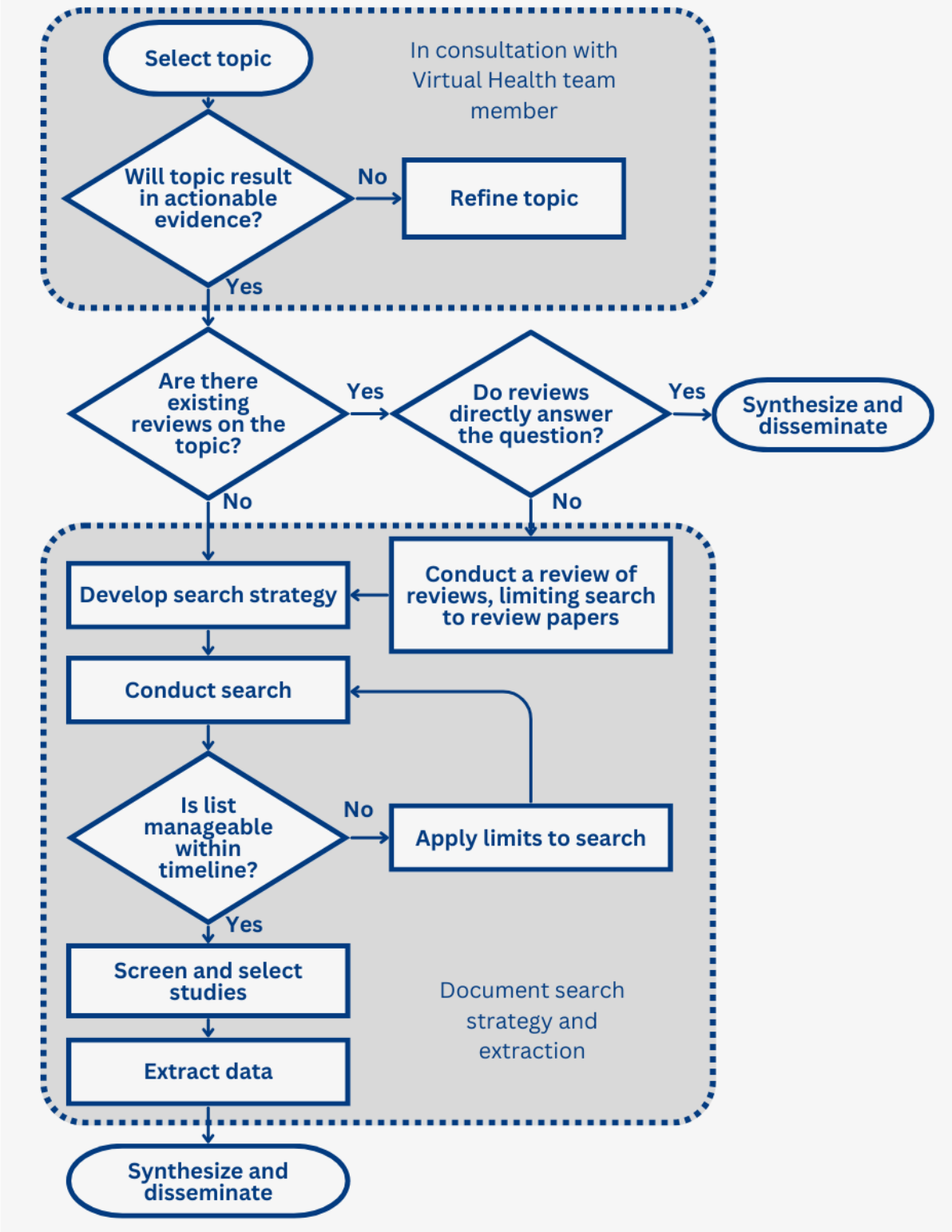

Knowledge Translation

Another strength of the Virtual Health team's rapid review approach was the development of easily digestible knowledge products highlighting key data synthesized in the review. Rather than providing end users with lengthy reports that often go unread, clinicians within the Virtual Health team helped to create brief summaries and infographics highlighting the main findings and recommendations. This approach was aimed at improving the uptake of research evidence into practice by presenting the information in a format that was easily accessible and understandable for clinicians and other stakeholders. By creating visually appealing and user-friendly knowledge products, the Virtual Health team was able to efficiently communicate key takeaways from the rapid reviews, thus facilitating their dissemination and implementation within the

XSL•FO
RenderX

FH context. This approach also helped to overcome a common challenge of KT, where research evidence can be difficult to access, understand, and apply in practice. By presenting the information in a format that was relevant and easily digestible, the Virtual Health team was able to enhance the applicability of the rapid reviews, thereby building clinician capacity and increasing their potential impact on patient outcomes.

Leveraging Rapid Reviews for Clinically Based Tools

Our most recent reviews were focused on developing a virtual nursing assessment and virtual nursing decision-making framework. Unlike traditional KT efforts used within other reviews, where the focus often lies on creating user-friendly summaries and infographics, our approach took a slightly different path. We aimed to directly inform the development of clinical decision support tools (DSTs).

Rather than developing traditional KT products, the raw data extracted from these reviews served as a foundational resource for the development of the clinical DSTs. Each piece of information was carefully referenced and integrated into the tool, providing evidence-based support for specific components and functionalities. This direct integration of research evidence into the tool development process not only strengthened the validity and credibility of the tool but also facilitated the transparent communication of the evidence behind each recommendation or feature.

Within these reviews, the active participation of those who were responsible for the development of the DSTs proved invaluable. Their involvement was crucial in ensuring understanding and confidence in the information as well as in merging research evidence with their own clinical expertise. By involving end users in the review process, we could tailor the outcomes to their specific needs and preferences, ultimately enhancing the relevance and applicability of the extracted evidence. This collaborative approach ensured that the resulting DSTs were not only evidence based but also resonated effectively with the clinical context they were intended for.

Discussion

Principal Findings

The Virtual Health team's experience with conducting 15 rapid reviews over the course of 1.5 years highlights the potential of rapid reviews as a time-efficient tool for improving the translation and uptake of research evidence into Virtual Health programming. Compared to more traditional review types (eg, systematic or scoping), which can take more than a year to complete [49], rapid reviews provide a practical way of synthesizing available evidence to inform clinical decision-making. The ability to produce a high-quality evidence summary in a shorter time frame can be particularly valuable in rapidly evolving areas of health care, such as virtual health. While rapid reviews are not new, our program offers insights into their application in a dynamic and rapidly evolving field such as virtual health. The lessons learned from FH's rapid review program have important implications for evidence-based decision-making and KT within health care settings.

One of our primary lessons learned underscores the importance of establishing clear and actionable research questions. By outlining precise objectives, rapid reviews can ensure the relevance and applicability of their results, thus facilitating their seamless integration into clinical practice. Moreover, our experiences highlight the transformative impact of involving knowledge users throughout the review process. This collaborative approach not only enhances the quality and relevance of the evidence synthesized but also fosters a culture of evidence-informed decision-making within the organization. This type of early and continued engagement of knowledge users in research endeavors has been increasingly recognized as pivotal for establishing research priorities and enhancing the utility of research findings in real-world health care contexts [50,51]. In line with this, the overarching goal of knowledge-user engagement in health research is to coproduce knowledge that directly addresses the needs of decision makers. By involving knowledge users from the outset, research priorities can be aligned with the practical requirements of health care delivery, thereby increasing the relevance and utility of research outputs [52-54].

Limitations of Rapid Reviews

Despite its benefits, the rapid review approach is not without limitations. Loss of rigor, as mentioned earlier in this paper, remains a concern. The rapid nature of the process may compromise the depth and comprehensiveness of the literature search and synthesis, potentially leading to oversights or biases in the evidence presented. Furthermore, within the context of virtual health, the rapid pace of technological advancements poses a challenge. New technologies may outpace the generation of peer-reviewed literature, resulting in a lag between their implementation and the availability of robust evidence.

In response to the challenge posed by rapidly evolving technologies, FH's Virtual Health department has used creative solutions to capture relevant evidence. While peer-reviewed literature remains a primary source, we have also incorporated gray literature, such as news articles, trade publications, and reports, from other health care authorities or departments within the review processes when applicable. Additionally, to supplement reviews and provide more contextual evidence, additional research and evaluation methodologies are used (time permitting) to inform Virtual Health service development such as consulting Patient and Family Advisory Councils within FH, conducting interviews with patient and clinician partners, and conducting analyses on existing data within FH.

Next Steps for FH's Rapid Review Program

We remain committed to advancing the rapid review program to meet the evolving needs of the Virtual Health department at FH. While we have heard anecdotally that knowledge users value the user-friendly knowledge products developed for rapid reviews, the next steps of this program include an evaluation of our knowledge dissemination to assess the reach and impact the reviews are having within the Virtual Health department.

Conclusions

Rapid reviews are a valuable tool for the timely synthesis of available research evidence to inform health care

XSL•FO
RenderX

decision-making. The Virtual Health team's experience with conducting rapid reviews highlights the importance of involving a diverse range of knowledge users in the review process and the need to focus on implementation considerations. By engaging knowledge users beyond designated researchers, and particularly by involving clinicians across the research process, rapid reviews become more robust, applicable, and aligned with the practical needs of health care providers and organizations, which can help to bridge the KTA gap.

Conflicts of Interest

None declared.

References

1. Goodridge D, Marciniuk D. Rural and remote care: overcoming the challenges of distance. Chron Respir Dis 2016;13(2):192-203 [FREE Full text] [doi: 10.1177/1479972316633414] [Medline: 26902542]
2. Bowen S, Zwi AB. Pathways to "evidence-informed" policy and practice: a framework for action. PLoS Med 2005;2(7):e166 [FREE Full text] [doi: 10.1371/journal.pmed.0020166] [Medline: 15913387]
3. Jacobs JA, Jones E, Gabella BA, Spring B, Brownson RC. Tools for implementing an evidence-based approach in public health practice. Prev Chronic Dis 2012;9:E116 [FREE Full text] [doi: 10.5888/pcd9.110324] [Medline: 22721501]
4. Davis D, Evans M, Jadad A, Perrier L, Rath D, Ryan D, et al. The case for knowledge translation: shortening the journey from evidence to effect. BMJ 2003;327(7405):33-35 [FREE Full text] [doi: 10.1136/bmj.327.7405.33] [Medline: 12842955]
5. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. Lancet 2003;362(9391):1225-1230. [doi: 10.1016/S0140-6736(03)14546-1] [Medline: 14568747]
6. Grol R, Jones R. Twenty years of implementation research. Fam Pract 2000;17(Suppl 1):S32-S35 [FREE Full text] [doi: 10.1093/fampra/17.suppl_1.s32] [Medline: 10735266]
7. Straus SE, McAlister FA. Evidence-based medicine: a commentary on common criticisms. CMAJ 2000;163(7):837-841 [FREE Full text] [Medline: 11033714]
8. Mellis C. Evidence-based medicine: what has happened in the past 50 years? J Paediatr Child Health 2015;51(1):65-68 [FREE Full text] [doi: 10.1111/jpc.12800] [Medline: 25536873]
9. Villar J, Carroli G, Gülmezoglu AM. The gap between evidence and practice in maternal healthcare. Int J Gynaecol Obstet 2001;75(Suppl 1):S47-S54. [Medline: 11742642]
10. Grol R. Successes and failures in the implementation of evidence-based guidelines for clinical practice. Med Care 2001;39(8 Suppl 2):II46-II54. [doi: 10.1097/00005650-200108002-00003] [Medline: 11583121]
11. Schuster MA, McGlynn EA, Brook RH. How good is the quality of health care in the United States? Milbank Q 1998;76(4):517-563 [FREE Full text] [doi: 10.1111/1468-0009.00105] [Medline: 9879302]
12. McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003;348(26):2635-2645 [FREE Full text] [doi: 10.1056/NEJMsa022615] [Medline: 12826639]
13. Lauer MS, Skarlatos S. Translational research for cardiovascular diseases at the National Heart, Lung, and Blood Institute: moving from bench to bedside and from bedside to community. Circulation 2010;121(7):929-933 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.109.917948] [Medline: 20177007]
14. Lang ES, Wyer PC, Haynes RB. Knowledge translation: closing the evidence-to-practice gap. Ann Emerg Med 2007;49(3):355-363. [doi: 10.1016/j.annemergmed.2006.08.022] [Medline: 17084943]
15. Kitson AL, Straus SE. Identifying knowledge to action gaps. Knowledge Transl Health Care 2013:97-109 [FREE Full text] [doi: 10.1002/9781118413555.ch09]
16. Graham ID, Logan JL, Harrison MB, Straus SE, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24 [FREE Full text] [doi: 10.1002/chp.47] [Medline: 16557505]
17. Knowledge translation strategy 2004-2009: innovation in action. Canadian Institutes of Health Research. 2004. URL: https://cihr-irsc.gc.ca/e/26574.html [accessed 2024-04-25]
18. Ciliska D, Thomas H, Buffett C. A compendium of critical appraisal tools for public health practice. National Collaborating Centre for Methods and Tools. 2008. URL: https://www.nccmt.ca/uploads/media/media/0001/01/b331668f85bc6357f262944f0aca38c14c89c5a4.pdf [accessed 2024-04-25]
19. Canada's strategy for patient-oriented research. Government of Canada. Canadian Institutes of Health Research. 2011. URL: https://cihr-irsc.gc.ca/e/44000.html#a1.1 [accessed 2023-04-06]
20. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Syst Rev 2012;1(1):10 [FREE Full text] [doi: 10.1186/2046-4053-1-10] [Medline: 22587960]
21. Chambers D, Wilson PM, Thompson CA, Hanbury A, Farley K, Light K. Maximizing the impact of systematic reviews in health care decision making: a systematic scoping review of knowledge-translation resources. Milbank Q 2011;89(1):131-156 [FREE Full text] [doi: 10.1111/j.1468-0009.2011.00622.x] [Medline: 21418315]

XSL•FO
RenderX

22. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implement Sci 2012;7(1):50 [FREE Full text] [doi: 10.1186/1748-5908-7-50] [Medline: 22651257]
23. Bosch-Capblanch X, Lavis JN, Lewin S, Atun R, Røttingen JA, Dröschel D, et al. Guidance for evidence-informed policies about health systems: rationale for and challenges of guidance development. PLoS Med 2012;9(3):e1001185 [FREE Full text] [doi: 10.1371/journal.pmed.1001185] [Medline: 22412356]
24. Oxman AD, Lavis JN, Lewin S, Fretheim A. SUPPORT tools for evidence-informed health policymaking (STP). Norwegian Knowledge Centre for the Health Services. 2010. URL: https://fhi.brage.unit.no/fhi-xmlui/bitstream/handle/11250/2378076/NOKCrapport4_2010.pdf?sequence=1 [accessed 2023-11-22]
25. Oliver K, Innvar S, Lorenc T, Woodman J, Thomas J. A systematic review of barriers to and facilitators of the use of evidence by policymakers. BMC Health Serv Res 2014;14(1):2 [FREE Full text] [doi: 10.1186/1472-6963-14-2] [Medline: 24383766]
26. Ganann R, Ciliska D, Thomas H. Expediting systematic reviews: methods and implications of rapid reviews. Implement Sci 2010;5(1):56 [FREE Full text] [doi: 10.1186/1748-5908-5-56] [Medline: 20642853]
27. Moore G, Redman S, Rudge S, Haynes A. Do policy-makers find commissioned rapid reviews useful? Health Res Policy Syst 2018;16(1):17 [FREE Full text] [doi: 10.1186/s12961-018-0293-1] [Medline: 29482643]
28. Flores EJ, Jue JJ, Giradi G, Schoelles K, Mull NK, Umscheid CA. AHRQ EPC Series on improving translation of evidence: use of a clinical pathway for C. Difficile treatment to facilitate the translation of research findings into practice. Jt Comm J Qual Patient Saf 2019;45(12):822-828. [doi: 10.1016/j.jcjq.2019.10.002] [Medline: 31672660]
29. Hartling L, Guise J, Kato E, Anderson J, Belinson S, Berliner E, et al. A taxonomy of rapid reviews links report types and methods to specific decision-making contexts. J Clin Epidemiol 2015;68(12):1451-1462.e3. [doi: 10.1016/j.jclinepi.2015.05.036] [Medline: 26278023]
30. Hartling L, Guise JM, Kato E, Anderson J, Aronson N, Belinson S, et al. EPC Methods: An Exploration of Methods and Context for the Production of Rapid Reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2015.
31. Hartling L, Guise JM, Hempel S, Featherstone R, Mitchell MD, Motu'apuaka ML, et al. Fit for purpose: perspectives on rapid reviews from end-user interviews. Syst Rev 2017;6(1):32 [FREE Full text] [doi: 10.1186/s13643-017-0425-7] [Medline: 28212677]
32. Featherstone RM, Dryden DM, Foisy M, Guise JM, Mitchell MD, Paynter RA, et al. Advancing knowledge of rapid reviews: an analysis of results, conclusions and recommendations from published review articles examining rapid reviews. Syst Rev 2015;4(1):50 [FREE Full text] [doi: 10.1186/s13643-015-0040-4] [Medline: 25925676]
33. MacPherson MM, Wang RH, Smith EM, Sithamparanathan G, Sadiq CA, Braunizer AR. Rapid reviews to support practice: a guide for professional organization practice networks. Can J Occup Ther 2023;90(3):269-279 [FREE Full text] [doi: 10.1177/00084174221123721] [Medline: 36229992]
34. Watt A, Cameron A, Sturm L, Lathlean T, Babidge W, Blamey S, et al. Rapid reviews versus full systematic reviews: an inventory of current methods and practice in health technology assessment. Int J Technol Assess Health Care 2008;24(2):133-139. [doi: 10.1017/S0266462308080185] [Medline: 18400114]
35. Polisena J, Garritty C, Kamel C, Stevens A, Abou-Setta AM. Rapid review programs to support health care and policy decision making: a descriptive analysis of processes and methods. Syst Rev 2015;4(1):26 [FREE Full text] [doi: 10.1186/s13643-015-0022-6] [Medline: 25874967]
36. Harker J, Kleijnen J. What is a rapid review? A methodological exploration of rapid reviews in health technology assessments. Int J Evid Based Healthc 2012;10(4):397-410 [FREE Full text] [doi: 10.1111/j.1744-1609.2012.00290.x] [Medline: 23173665]
37. Garritty C, Gartlehner G, Nussbaumer-Streit B, King VJ, Hamel C, Kamel C, et al. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol 2021;130:13-22 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.10.007] [Medline: 33068715]
38. Fraser Health. 2023. URL: https://www.fraserhealth.ca/ [accessed 2023-04-06]
39. Virtual Health. Fraser Health. URL: https://www.fraserhealth.ca/patients-and-visitors/virtual-health [accessed 2023-04-06]
40. MacPherson M. Immigrant, refugee, and Indigenous Canadians' experiences with virtual health care services: rapid review. JMIR Hum Factors 2023 09;10:e47288 [FREE Full text] [doi: 10.2196/47288] [Medline: 37812489]
41. MacPherson M. Virtual care in heart failure, chronic obstructive pulmonary disease, and diabetes: a rapid review protocol. OSF Registries. 2023. URL: https://osf.io/xn2pe [accessed 2023-09-11]
42. MacPherson M. Barriers, facilitators, and recommendations to inform the expansion of remote patient monitoring services for respiratory care: a rapid review. OSF Registries. 2022. URL: https://osf.io/asf2v/ [accessed 2024-04-25]
43. MacPherson M. Virtual health services in the context of acute care: a rapid review. OSF Registries. 2023. URL: https://osf.io/ub2d8/ [accessed 2024-04-25]
44. MacPherson MM, Kapadia S. Barriers and facilitators to patient-to-provider messaging using the COM-B model and theoretical domains framework: a rapid umbrella review. BMC Digit Health 2023;1(1):33 [FREE Full text] [doi: 10.1186/s44247-023-00033-0]
45. Chan L, MacPherson M. Remote patient monitoring: an evidence synthesis. OSF Registries. 2023. URL: https://osf.io/7wqb8/ [accessed 2024-04-25]

XSL•FO
RenderX

46. Montenegro M, MacPherson M. Barriers to virtual care experienced by patients and healthcare providers: a rapid umbrella review. OSF Registries. 2023. URL: https://osf.io/nufg4/ [accessed 2024-04-25]
47. Montenegro M, MacPherson M. Virtual hospitals: a rapid review. OSF Registries. 2023. URL: https://osf.io/m3a4b/ [accessed 2024-04-25]
48. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-05-13]
49. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7(2):e012545 [FREE Full text] [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
50. Deverka PA, Lavallee DC, Desai PJ, Esmail LC, Ramsey SD, Veenstra DL, et al. Stakeholder participation in comparative effectiveness research: defining a framework for effective engagement. J Comp Eff Res 2012;1(2):181-194 [FREE Full text] [doi: 10.2217/cer.12.7] [Medline: 22707880]
51. Bragge P, Clavisi O, Turner T, Tavender E, Collie A, Gruen RL. The Global Evidence Mapping Initiative: scoping research in broad topic areas. BMC Med Res Methodol 2011;11(1):92 [FREE Full text] [doi: 10.1186/1471-2288-11-92] [Medline: 21682870]
52. Langlois EV, Montekio VB, Young T, Song K, Alcalde-Rabanal J, Tran N. Enhancing evidence informed policymaking in complex health systems: lessons from multi-site collaborative approaches. Health Res Policy Syst 2016;14(1):20 [FREE Full text] [doi: 10.1186/s12961-016-0089-0] [Medline: 26983405]
53. Vindrola-Padros C, Pape T, Utley M, Fulop NJ. The role of embedded research in quality improvement: a narrative review. BMJ Qual Saf 2017;26(1):70-80 [FREE Full text] [doi: 10.1136/bmjqs-2015-004877] [Medline: 27129492]
54. Ghaffar A, Langlois EV, Rasanathan K, Peterson S, Adedokun L, Tran NT. Strengthening health systems through embedded research. Bull World Health Organ 2017;95(2):87 [FREE Full text] [doi: 10.2471/BLT.16.189126] [Medline: 28250505]

Abbreviations

COPD: chronic obstructive pulmonary disease
DST: decision support tool
FH: Fraser Health
KT: knowledge translation
KTA: knowledge-to-action

Edited by Z Yin; submitted 22.11.23; peer-reviewed by W LaMendola, M Willenbring, Y Zhang, P Blasi; comments to author 10.03.24; revised version received 15.03.24; accepted 13.04.24; published 22.05.24.

Please cite as:
MacPherson M, Rourke S
The Power of Rapid Reviews for Bridging the Knowledge-to-Action Gap in Evidence-Based Virtual Health Care
J Med Internet Res 2024;26:e54821
URL: https://www.jmir.org/2024/1/e54821
doi:10.2196/54821
PMID:38776542

©Megan MacPherson, Sarah Rourke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Efficient Use of Biological Data in the Web 3.0 Era by Applying Nonfungible Token Technology

Guanyi Wang[1*], PhD; Chen Chen[1*], PhD; Ziyu Jiang[1*], MS; Gang Li[1], PhD; Can Wu[1], MD; Sheng Li[1], MD

Department of Urology, Cancer Precision Diagnosis and Treatment and Translational Medicine Hubei Engineering Research Center, Zhongnan Hospital, Wuhan University, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Sheng Li, MD
Department of Urology
Cancer Precision Diagnosis and Treatment and Translational Medicine Hubei Engineering Research Center
Zhongnan Hospital, Wuhan University
169 Donghu Road, Wuchang District
Wuhan, 430062
China
Phone: 86 18086601827
Email: lisheng-znyy@whu.edu.cn

Abstract

CryptoKitties, a trendy game on Ethereum that is an open-source public blockchain platform with a smart contract function, brought nonfungible tokens (NFTs) into the public eye in 2017. NFTs are popular because of their nonfungible properties and their unique and irreplaceable nature in the real world. The embryonic form of NFTs can be traced back to a P2P network protocol improved based on Bitcoin in 2012 that can realize decentralized digital asset transactions. NFTs have recently gained much attention and have shown an unprecedented explosive growth trend. Herein, the concept of digital asset NFTs is introduced into the medical and health field to conduct a subversive discussion on biobank operations. By converting biomedical data into NFTs, the collection and circulation of samples can be accelerated, and the transformation of resources can be promoted. In conclusion, the biobank can achieve sustainable development through “decentralization.”

(J Med Internet Res 2024;26:e46160) doi:10.2196/46160

KEYWORDS

NFTs; biobanks; blockchains; health care; medical big data; sustainability; blockchain platform; platform; tracing; virtual; biomedical data; transformation; development; promoted

Dilemma of Sustainable Development of Biobanks

Growth of Biobanks

Biobanks are biological application systems that are composed of physical samples and databases that centrally store various biological materials and their corresponding information for clinical treatments and life science research. Specific functions of biobank include standardized collection, processing, and storing of physical samples such as tissues, body fluids, and biological derivatives (eg, DNA and RNA), as well as related data storage such as pathological data, clinical diagnosis, treatment records, laboratory omics, ethical approval statements, and informed consent forms [1]. The rudimentary concept of a biobank was born in 1996 [2], and their numbers have increased yearly in the past 20 years with rapid developments in the biomedical industry. Around 636 biobanks opened in the United States in 2013 [3], and 325 institutions in the European-based Biobanking and BioMolecular resources Research Infrastructure were founded until 2018 [4]. However, most biobanks face problems with sustainable development.

Current Dilemma

As biobanks are an essential link in the precision medicine industry, they require large-scale storage and sample and data management. Significant personnel and funding for equipment, technical services, biomedical consumables, and professional knowledge are necessary for early setup or maintenance. Moreover, biobanks undergo active or passive expansion as the storage sizes increase, which inevitably increases the need for funding [5]. However, as biobanks are not directly involved with production, they often act as an accessory to a specific research group or a collaborative department of public hospitals

XSL•FO
RenderX

and have a fuzzy economic operation model [6]. In addition, international rules stipulate that stored human samples should not be used for commercial purposes [7], so establishing a reasonable operational model for biobanks and achieving sustainable development is critical and needs urgent discussion.

Currently, most biobanks in the world, such as the University of California San Francisco AIDS specimen bank, the Wales Cancer Biobank, the Australia Prostate Cancer Bioresource, and the Ontario Cancer Bank [8], recover costs by charging access fees from researchers. Despite this operational model, most biobanks cannot fully recover costs and rely solely on the support of government policies and donations. The Singapore biobank, established in 2002, eventually closed owing to a lack of finances and loss of public and government support [9]. Visibly, programs in which biobanks are developed solely by the government or with funding from a public agency are unrealistic. A path to achieving self-sufficiency is the only solution for long-term survival.

Roots of the Dilemma

As biobanks form a bridge between clinical medicine and medical research, they must connect with patients who are willing to provide sample data resources and researchers of scientific institutes. Compared to researchers, patients have less initiative and are not as engaged in the entire industrial chain despite them being at the initial link of the chain as the sample and data source. One subjective reason is that patients usually do not fully understand and value the meaning of sample donation. The survey also evidenced that biobank participants are sometimes unaware of the potential risks and benefits of a project [10]; therefore, it may be challenging to convince them to donate.

Biobanks may alleviate funding dilemmas by charging samples and associated data from researchers. Nevertheless, this scheme does not completely resolve the financial burden and can cause other troubles. Relevant data show that sample use in biobanks is <10% [6]. The benefit from only a small number of samples and data may not be enough to afford the long-term management and storage expenses incurred by unused samples and data. In addition, it is difficult for biobanks to define the scope and quality of shared services when providing samples and data to researchers. For example, patient data, particularly those with chronic diseases, often dynamically increases with disease duration. Whereas the availability of biological medical data from biobanks cannot be a “one-time deal,” and determining how the access fee is charged is yet to be discussed. Moreover, security hazards such as data leakage during data sharing or informal access to data may affect the direct benefit of biobanks. It is noteworthy that even when biobanks take considerable economic benefits from operating on patient specimens and data, they can raise ethical and legal issues [11].

Besides ensuring adequate sample quantity, high-quality samples with complete data are prerequisites that guarantee subsequent scientific research output. Many larger biobanks are currently involved in converting physical samples into digital data besides storage management services. Thus, through comprehensive research and analysis of biological samples, genomics, proteomics, transcriptomics, and metabolomics data can be generated, which is convenient for storage and sharing. However, complete sample digitalization is unrealistic, and relying solely on biobanks is impractical. Failure in ensuring data integrity and comprehensiveness also affects subsequent sharing services that make it difficult for researchers to mine for valuable scientific information, in addition to other problems such as data standardization and sharing difficulties.

Throughout the entire industrial chain of translational medicine in biobanks, it is not difficult to find intrinsic motivation weak under its promising prospect. “Biobank” seems to exist merely as “storage” and does not reflect the actual connotation of “bank,” that is, to realize the distribution, value-added, and accelerated turnover of “resource wealth.” Perhaps, biobanks should be positioned not only as pipelines connecting upstream and downstream data but also as self-powered accelerators that vigorously drive research output while robustly accumulating data from clinical samples. Therefore, we introduce the concept of a nonfungible token (NFT) for digital assets that envision a new underlying model to address the existing issues.

NFT of Biobank Sample Data

What Is an NFT

In the real world, assets can generally be divided into fungible or nonfungible. Fungible assets such as stocks, game coins, and precious metals tend to have a fixed value that focuses on their quantity rather than their characteristics during any transaction. Unlike fungible assets, nonfungible assets such as houses, buildings, artworks, and game equipment are characterized by uniqueness and irreplaceability (Figure 1).

With the rise and rapid development of blockchain, digital assets have attracted wide attention from academia and industry. These are also classified into fungible and nonfungible assets corresponding to their value in the physical world. The most common fungible digital assets include digital currencies such as Bitcoins and Ethereum. However, nonfungible digital assets include NFTs such as game items and digital artworks [12]. NFTs make it possible to connect the real physical world with the digital network world and have become a core element at the center of an intelligent economy. NFT is a blockchain-based digital asset, which is verifiable, tradable, indivisible, irreplaceable, and unique. Ownership of an NFT can be transferred through intelligent contracts, and this transaction can be recorded by the blockchain; all transaction records can be viewed through any node. The characteristics of blockchain comprise open, transparent, traceable, anticounterfeit, and tamper-proof data that ensure transparency, tamper-proof, and antiduplication of the NFT transaction process [13].

XSL•FO
RenderX

Figure 1. Differences between fungible assets and nonfungible assets. Fungible assets often have a fixed value and are concerned with their quantity rather than their specific attributes in any transaction. NFTs, like real estate, buildings, artworks, and gaming equipment, belong to the category of nonfungible assets, possessing uniqueness and nonreplaceability. NFT: nonfungible token.

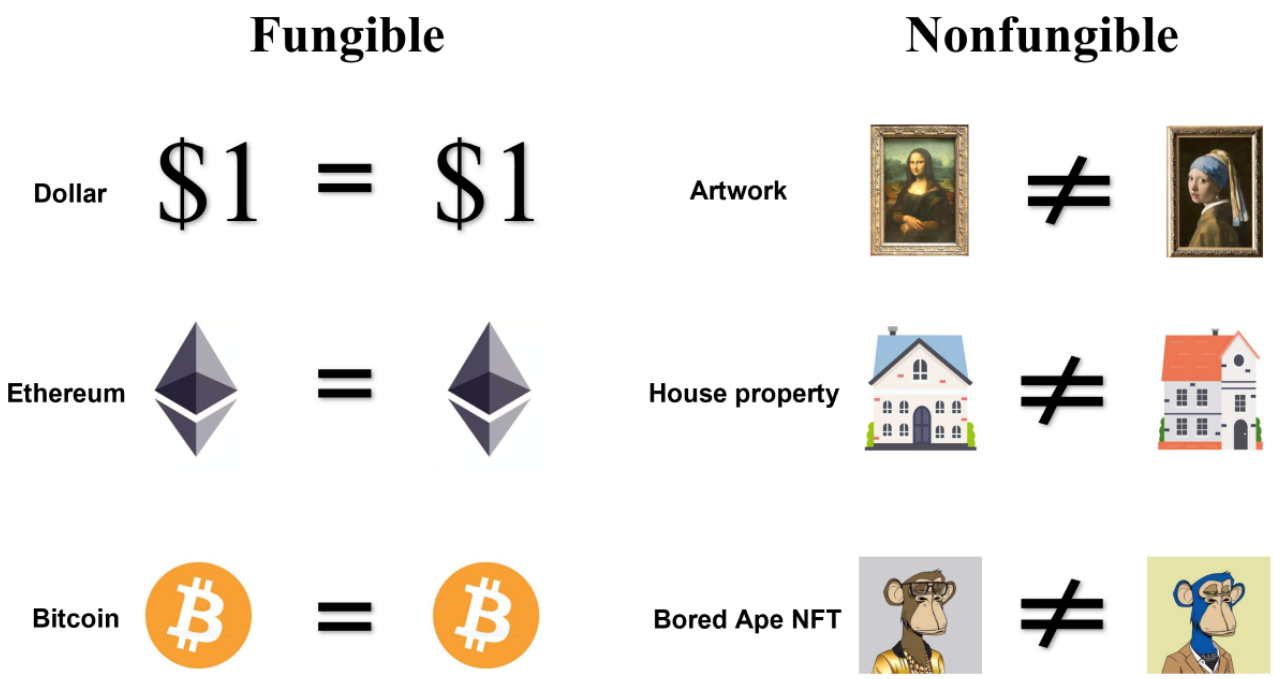

Feasibility of NFT for Biomedical Data

NFT has been successfully applied to many industries. However, it is mainly comprised of digital assets such as artwork, collectibles, video game items, and property rights. Bamakan et al [14] offered a layered conceptual NFT-based framework to create a decentralized market for patent storage and recording that facilitates changing the patent ownership and accelerates their commercialization. In addition, Dos Santos et al [15] proposed the use of NFTs and other tokens to realize transparency of supply chains and trace product supervision from the source and ensure food safety, enhance consumer trust, and realize sustainable agricultural development. With the maturity of blockchain technology and the popularization of its applications, transforming various digital assets and even physical assets into certifications will be a general trend due to the essential advantages of NFT in rights confirmation and transaction.

So far, NFT has not been deployed on a large scale in the field of biomedicine. Ng et al [16] systematically reviewed cutting-edge research on blockchain applications and other digital technologies in medicine and health care. They concluded a strong trend of digitization, such as the application of information technology in the public health response to COVID-19 [17], and blockchain-based patient-centric data management system for managing electronic health records in cancer nursing [18]. Digitalization of medical data will be a good start for NFT construction. NFT relies on a decentralized network that uses advanced encryption to verify the efficacy of transactions. Unlike cryptocurrencies that are seen as currency or digital assets, NFT is equivalent to a digital fingerprint of data or an authenticity certificate. It can hold many files that include transaction IDs and detailed history. Therefore, we can store a large number of biomedical and sample data through an NFT [19]. As envisaged by Skalidis et al [20], tagging and converting patient data (medical information, health status, examination results, and consent forms) into an NFT can enhance privacy and ensure data integrity and confidentiality in clinical practice and research.

NFTs' Potential Applications in the Biobanks

Data Security and Integrity

Biobanks are vital resources for medical research, containing extensive biological samples and related clinical data. The security and integrity of these data are crucial to ensure its credibility and availability. NFTs leverage blockchain technology to provide a decentralized method of data storage and management, effectively safeguarding data from unauthorized access and tampering. Each data transaction is recorded on the blockchain, ensuring data immutability, which helps maintain data integrity and protect patient privacy.

Data Traceability and Credibility

Data in biobanks often require tracing their sources and histories to ensure their credibility. NFTs, with the distributed nature and transparency of blockchain technology, make it easy to trace the origin and modification history of each data point. This is particularly helpful for verifying data accuracy and credibility, especially when validating the source and the quality of data is essential in medical research.

XSL•FO
RenderX

Data Standardization and Sharing

Due to varying standards in data collection, organization, and storage, there is a high degree of heterogeneity among biobanks. Undoubtedly, the lack of standardization leads to difficulties in sharing, which hinders subsequent research translation. By relying on NFTs' standardization and universality, data storage could potentially break the barriers of nonstandardized sharing. Additionally, incidents of patient privacy breaches are continually emerging. The Georgia-based Emory Clinic experienced unauthorized data access in March 2017, resulting in the exposure of personal information for 79,930 individuals. The Alberta Health Service recently faced a privacy breach that leaked confidential health information for 12,848 individuals [11]. These security issues can be resolved through NFT's blockchain-based sharing model.

Application of Smart Contracts

NFT-based smart contracts can programmatically execute contract conditions, providing a way for biobanks to automate the management of data and samples. Unlike other research activities, research using samples from biobanks has specific ethical and social problems owing to their long-term storage. So far, most biobanks have adopted the "broad consent" model, which is an agreement to use samples for current or future research within a specific framework without any subsequent contact with patients [21]. This provides researchers with enough flexibility to adapt to future research topics, but it also means that patients will not receive any feedback from the study. To some extent, "broad consent" cannot fully comply with the principle of "informed consent." The World Medical Association Declaration of Helsinki in 1964 stipulated that every potential subject must be fully informed about any study-related information. Therefore, relying on the programmability of NFTs equipped with the corresponding service platform, we can achieve "dynamic consent" through smart contracts [22,23]. This approach places participants at the center of biobank governance, ensuring their data are only used with explicit consent. Patients can access "real-time" information about research projects and easily grant or revoke consent. This dynamic consent model increases participant control over the research process and enhances engagement while avoiding theoretical and legal loopholes in the "broad consent" model.

Decentralized Power Structure

"Decentralization" is a significant feature of organizations with blockchain applications in the Web 3.0 era, and its impact on biobanks is no exception. Traditional biobanks are typically managed and controlled by central institutions, which can lead to data centralization and potential misuse. Blockchain technology can achieve a decentralized power structure, transferring control of data to data owners, such as patients. This decentralization can increase data owner involvement and control while enhancing data availability. Patients often do not receive direct benefits from biological samples or data donations, and their potential benefit lies in the results or findings of scientific research projects in which they have participated. Patients can benefit earlier from the industrial chain after their data are converted into NFTs. They do not need to fully understand the value of samples in scientific research projects, and donation can shift from a seemingly moral act of "giving" to a direct "benefit exchange" behavior. Depending on the data source, NFTs include clinical data NFTs, owned by patients, experimental data NFTs generated by researchers, and sample information NFTs generated during the storage process. Rights and power relations between research institutions, individual researchers, and participants can be explicitly defined through NFT ownership. This resolves the complex ethical issue of who owns and controls biobank data. As indicated by Kostick-Quenet et al [24], NFTs or NFT-like frameworks can incentivize a more democratic, transparent, and efficient health information exchange. In our vision, biobanks will decentralize their rights and responsibilities to both upstream and downstream data, achieving a beneficial situation (Figure 2).

XSL•FO
RenderX

Figure 2. Achieving sustainable biobanks through NFTs. Clinical data from patients and research data from scientists are stored within the biobank, simultaneously generating 3 types of NFTs: the patient's clinical data NFT, the researcher's research data NFT, and the storage data NFT for the samples. The data within the biobank are subsequently mined by research institutions, ultimately resulting in research outputs such as new drugs or treatment protocols that benefit the patients themselves. Due to the inherent value of NFTs, there is an incentive for the storage of both clinical and research data. The storage data NFT for the samples can be considered as the self-generated value of the biobank, while the other 2 types of NFTs are anchored to it during transactions. Additionally, we envision incentive and economic models for the sustainable development of biobanks based on NFT transactions, including data contribution incentives, data access fees, data storage costs, data quality and credibility rewards, and data sharing rewards. NFT: nonfungible token.

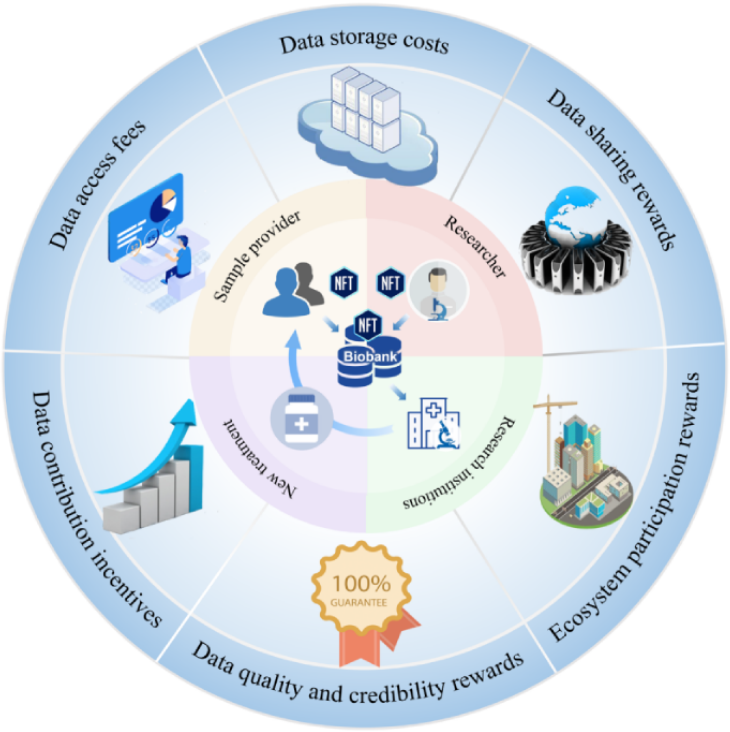

Achieving Sustainable Biobanks Through NFTs

Overview

NFT technology can also improve the economic model of biobanks. Biobanks should be able to operate much like banks and sustain their operations by earning "interest," thus reducing their reliance on government funding. To achieve this, we propose the following incentive and economic models.

Data Contribution Incentives

Biobanks can motivate individuals or research institutions to contribute their biomedical data to the repository through reward mechanisms. When individuals or institutions convert their data into NFTs and store them on the blockchain, they can earn cryptocurrency or other rewards. This incentive model encourages more data contributions, increasing the quantity and diversity of data in the repository.

Data Access Fees

Researchers or health care institutions seeking access to the repository's data can pay data access fees. These fees can be determined based on the value and scarcity of the data and can be automatically processed by smart contracts. Some of the data access fees can be allocated to the maintenance and operation of the repository, ensuring its sustainability.

Data Storage Costs

Individuals or institutions storing their data in the repository may need to pay storage fees. This encourages data owners to update and manage their data promptly, avoiding wastage of storage space. Storage fees can also serve as a source of income for the biobank, supporting its operations.

Data Quality and Credibility Rewards

Data quality and credibility are crucial for medical research. The repository can establish reward mechanisms to incentivize data contributors to maintain high-quality, accurate, and complete data. These rewards can be provided in the form of NFTs or cryptocurrencies to encourage data contributors to provide the best data.

Ecosystem Participation Rewards

Parties participating in the biobank ecosystem, including researchers, health care institutions, developers, and maintainers, can earn ecosystem participation rewards. These rewards can be determined based on their contributions and level of involvement in the ecosystem. This helps build an actively engaged community, driving the growth of the repository.

Data Sharing Rewards

When data are used multiple times or shared across various research projects, data contributors can earn rewards. This incentivizes the reuse and sharing of data, increasing its value.

XSL•FO
RenderX

These incentive and economic models can be automatically executed on the blockchain through smart contracts, ensuring fairness and transparency. They encourage best practices in data contribution, management, access, and quality, thereby increasing the value and sustainability of biobanks. Moreover, these models contribute to the establishment of a shared and collaborative medical research ecosystem, driving advancements in medical research.

NFT-Based Biomedical Data Framework

Overview

Based on the above assumption, we propose a basic framework of biomedical data based on an NFT, which includes 4 primary levels: application layer, management layer, smart contract layer, and blockchain network layer (Figure 3). The specific structure and implementation of the entire framework are as follows.

Figure 3. Proposed architecture for presenting NFT-based biomedical data. A basic framework of biomedical data based on an NFT, which includes 4 primary levels: application layer, management layer, smart contract layer, and blockchain network layer. Creating NFTs for biomedical data and leveraging the properties of NFTs can indeed be beneficial for improving the efficiency of data conversion and sharing within biobanks. NFT: nonfungible token.

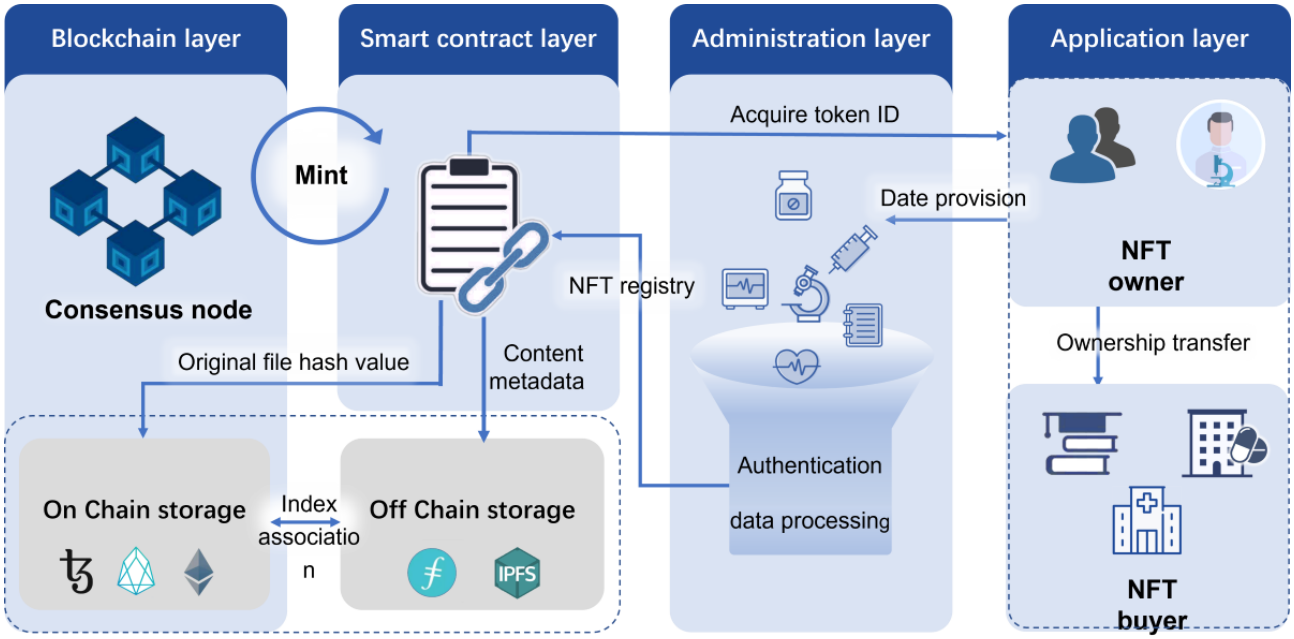

Application Layer

By interacting with decenteralized applications, the biomedical data platform allows data providers such as patients and researchers to input or import data via interfaces. After the data are marked as an NFT, its ownership will be transferred to the data demander (schools, hospitals, pharmaceutical enterprises, and other scientific research institutions). Ownership will be eventually tracked through the blockchain.

Management Level

This will implement user registration and authentication management and use the data processing module to desensitize, clean, and fuse the data to ensure data efficacy and security. Implementation of the interface operation will be done with the NFT smart contract.

Smart Contract Layer

Smart contracts generated related to biomedical information as programs running on the blockchain can receive and send transactions. However, they must comply with NFT protocols like Ethereum's erc721 protocol [25]. After generating a smart contract, the user will obtain a token ID as the NFT ownership certificate. The hash value of the original biomedical information file will be uploaded to the blockchain network for broadcasting, and the metadata of the original file will be stored in the chain [26,27]. The so-called "Mint" process records the ID of an NFT and the address of its owner on the blockchain.

Blockchain Network Layer

Blockchain systems can be mainly divided into public and private blockchains based on their consensus mechanisms. In public blockchains, any node can participate in the point-to-point network, and it is completely decentralized. In a private blockchain, the read permission can be opened to the public or it can be arbitrarily restricted based on the situation. In addition to these, there is also a semidecentralized alliance chain whose consensus process may be controlled by some designated nodes [28,29]. Public chains with good credit can be selected, such as Ethereum, EOS, Flow, Tezos, and so forth.

Off-Chain Storage

Biomedical data are rich in content, and the chain of original files tends to put tremendous pressure on the blockchain network. Thus, the chain of original files is not realistic at present, and off-chain storage can solve the network storage pressure, network pressure, and high maintenance costs [30]. Developers use a function (token uniform resource identifier) to let the app know the location of the metadata of a specific item and establish on-chain and off-chain indexes to realize the

XSL•FO
RenderX

association [25,31]. Common off-chain external storage platforms include interPlanetary File System, Filecoin, and Pinata.

Based on the above framework, NFT technology presents potential solutions for the biobank field, enhancing data security, credibility, and management efficiency while also increasing data owners' participation and control. These technologies hold promise for significant improvements in biobank operations and medical research, providing better data management and use for both patients and researchers.

Future and Development

We are amidst the tremendous transformation of the Web 2.0 era into the Web 3.0 era. This requires joint exploration of multiple technologies, including blockchain, artificial intelligence, and the internet of things. Ultimately, we will embrace a relatively decentralized, automated, and intelligent new era of the internet based on the premise that personal digital identity, digital assets, and data return would be entirely secure for individuals. In addition to changes in the web-based network, continuous interaction between the digital and the real world and developing future technologies will bring new subversive actions to the social systems and institutions in the physical world. In terms of ethical issues, web-based world ethics is usually an extension or reflection of real ethics. The interaction of virtual reality makes personal data more widely collected and shared, and there are more potential unknown risks. In this situation, protecting personal privacy and data security has become crucial. Therefore, it is necessary to establish emerging ethical and legal norms to balance the relationship between technological development and individual rights. Ethics committees or review bodies at the level of hospitals and research institutions, as well as higher level supervisory bodies, should focus more on the authenticity, security, and traceability of data and ensure the transparency and effectiveness of ethical review and supervision agencies. In addition, with the decentralized development model of the Web 3.0 era, after desensitizing relevant data, the supervision and management rights of such relevant institutions can even be delegated to every data participant.

Currently, medical institutions in many countries offer enhanced views of precision 3D imaging to physicians using digitally assisted procedures during the treatment, such as applying Microsoft HoloLens 2 hybrid reality technology in high-finesse surgical procedures [32,33]. Similarly, in the future, with the development of XR (virtual reality, augmented reality, mixed reality) technology, 6G high-speed network transmission, and cloud computing, we could eventually meet all requirements of the "all-round true experience" of the metaverse and can provide a better treatment experience and visits to patients.

In the Web 3.0 era, "all-round true experience" will break through the barriers of time and space, making medical resources more evenly shared. Finally, the importance of individual physician's autonomy and data exchange platform will increase. Perhaps biobanks could be a foundational platform for the convenient sharing of all health care data in the Web 3.0 era. In addition to clinical and experimental data, remote monitoring of health and epidemiology data generated by wearable devices, implanted microchips, and the internet of things will be incorporated into storage and management [34].

We also hypothesize that an open-source, public blockchain platform with an intelligent contract function could also be developed in public health, which could breach the barrier of data silos among various hospitals, medical research institutes, and drug development enterprises. Thus, we can create a cosmopolitan project using human health data and build a new world for revolutionary data management.

Authors' Contributions

GW and SL conceived and designed the study; GW, ZJ, and CW contributed to the writing of the manuscript; GL provided software support; and CC implemented project management. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020 May 26;11(1):2624. [doi: 10.1038/s41467-020-15948-9] [Medline: 32457287]
2. Hewitt R, Watson P. Defining biobank. Biopreserv Biobank 2013;11(5):309-315. [doi: 10.1089/bio.2013.0042] [Medline: 24835262]
3. Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, et al. Characterizing biobank organizations in the U.S.: results from a national survey. Genome Med 2013;5(1):3 [FREE Full text] [doi: 10.1186/gm407] [Medline: 23351549]
4. Paskal W, Paskal AM, Dębski T, Gryziak M, Jaworowski J. Aspects of modern biobank activity—comprehensive review. Pathol Oncol Res 2018;24(4):771-785 [FREE Full text] [doi: 10.1007/s12253-018-0418-4] [Medline: 29728978]
5. McQueen MJ, Keys JL, Bamford K, Hall K. The challenge of establishing, growing and sustaining a large biobank: a personal perspective. Clin Biochem 2014;47(4-5):239-244. [doi: 10.1016/j.clinbiochem.2013.11.017] [Medline: 24316099]

XSL•FO
RenderX

6. Henderson MK, Goldring K, Simeon-Dubach D. Advancing professionalization of Biobank Business Operations: performance and utilization. Biopreserv Biobank 2019;17(3):213-218 [FREE Full text] [doi: 10.1089/bio.2019.0005] [Medline: 31188630]
7. De Souza YG. Sustainability of biobanks in the future. In: Karimi-Busheri F, editor. Biobanking in the 21st Century. Cham: Springer International Publishing; 2015:29-36.
8. Coppola L, Cianflone A, Grimaldi AM, Incoronato M, Bevilacqua P, Messina F, et al. Biobanking in health care: evolution and future directions. J Transl Med 2019;17(1):172 [FREE Full text] [doi: 10.1186/s12967-019-1922-3] [Medline: 31118074]
9. Wai CT. The closure of the National Bio-Bank in Singapore. Asia Pac Biotech News 2012;16:40-43 [FREE Full text]
10. Siminoff LA, Wilson-Genderson M, Mosavel M, Barker L, Trgina J, Traino HM. Confidentiality in biobanking research: a comparison of donor and nondonor families' understanding of risks. Genet Test Mol Biomarkers 2017;21(3):171-177 [FREE Full text] [doi: 10.1089/gtmb.2016.0407] [Medline: 28121471]
11. Caulfield T, Murdoch B. Genes, cells, and biobanks: yes, there's still a consent problem. PLoS Biol 2017;15(7):e2002654 [FREE Full text] [doi: 10.1371/journal.pbio.2002654] [Medline: 28742850]
12. Nadini M, Alessandretti L, Di Giacinto F, Martino M, Aiello LM, Baronchelli A. Mapping the NFT revolution: market trends, trade networks, and visual features. Sci Rep 2021;11(1):20902 [FREE Full text] [doi: 10.1038/s41598-021-00053-8] [Medline: 34686678]
13. Karandikar N, Chakravorty A, Rong C. Blockchain based transaction system with fungible and non-fungible tokens for a community-based energy infrastructure. Sensors (Basel) 2021;21(11):3822 [FREE Full text] [doi: 10.3390/s21113822] [Medline: 34073110]
14. Bamakan SMH, Nezhadsistani N, Bodaghi O, Qu Q. Patents and intellectual property assets as non-fungible tokens; key technologies and challenges. Sci Rep 2022;12(1):2178 [FREE Full text] [doi: 10.1038/s41598-022-05920-6] [Medline: 35140251]
15. Dos Santos RB, Torrisi NM, Pantoni RP. Third party certification of agri-food supply chain using smart contracts and blockchain tokens. Sensors (Basel) 2021;21(16):5307 [FREE Full text] [doi: 10.3390/s21165307] [Medline: 34450749]
16. Ng WY, Tan TE, Movva PVH, Fang AHS, Yeo KK, Ho D, et al. Blockchain applications in health care for COVID-19 and beyond: a systematic review. Lancet Digit Health 2021;3(12):e819-e829 [FREE Full text] [doi: 10.1016/S2589-7500(21)00210-7] [Medline: 34654686]
17. Gunasekeran DV, Tseng RMWW, Tham YC, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021;4(1):40 [FREE Full text] [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
18. Dubovitskaya A, Baig F, Xu Z, Shukla R, Zambani PS, Swaminathan A, et al. ACTION-EHR: patient-centric blockchain-based electronic health record data management for cancer care. J Med Internet Res 2020;22(8):e13598 [FREE Full text] [doi: 10.2196/13598] [Medline: 32821064]
19. Jones N. How scientists are embracing NFTs. Nature 2021;594(7864):481-482. [doi: 10.1038/d41586-021-01642-3] [Medline: 34145410]
20. Skalidis I, Muller O, Fournier S. CardioVerse: the cardiovascular medicine in the era of metaverse. Trends Cardiovasc Med 2023;33(8):471-476 [FREE Full text] [doi: 10.1016/j.tcm.2022.05.004] [Medline: 35568263]
21. Grady C, Eckstein L, Berkman B, Brock D, Cook-Deegan R, Fullerton SM, et al. Broad consent for research with biological samples: workshop conclusions. Am J Bioeth 2015;15(9):34-42 [FREE Full text] [doi: 10.1080/15265161.2015.1062162] [Medline: 26305750]
22. Kaye J, Whitley EA, Kanellopoulou N, Creese S, Lund D, Hughes KJ. Dynamic consent: a solution to a perennial problem? BMJ 2011:343 [FREE Full text]
23. Steinsbekk KS, Myskja BK, Solberg B. Broad consent versus dynamic consent in biobank research: is passive participation an ethical problem? Eur J Hum Genet 2013;21(9):897-902 [FREE Full text] [doi: 10.1038/ejhg.2012.282] [Medline: 23299918]
24. Kostick-Quenet K, Mandl KD, Minssen T, Cohen IG, Gasser U, Kohane I, et al. How NFTs could transform health information exchange. Science 2022;375(6580):500-502 [FREE Full text] [doi: 10.1126/science.abm2004] [Medline: 35113709]
25. Entriken W, Shirley D, Evans J, Sachs N. ERC-721 non-fungible token standard. Ethereum Improvement Proposals, no. 721. 2018. URL: https://eips.ethereum.org/EIPS/eip-721 [accessed 2024-04-10]
26. Dowling M. Is non-fungible token pricing driven by cryptocurrencies? Finance Research Letters 2022 Jan;44:102097. [doi: 10.1016/j.frl.2021.102097]
27. Wang Q, Li R, Wang Q, Chen S. Non-fungible token (NFT): overview, evaluation, opportunities and challenges. arXiv. Preprint posted online on May 16, 2021 [FREE Full text]
28. Saberi S, Kouhizadeh M, Sarkis J, Shen L. Blockchain technology and its relationships to sustainable supply chain management. Int J Prod Res 2018;57(7):2117-2135. [doi: 10.1080/00207543.2018.1533261]
29. Lv L, Yang Z, Zhang L, Huang Q, Tian Z. Multi-party transaction framework for drone services based on alliance blockchain in smart cities. J Inf Secur Appl 2021;58:102792. [doi: 10.1016/j.jisa.2021.102792]
30. Miyachi K, Mackey TK. hOCBS: a privacy-preserving blockchain framework for healthcare data leveraging an on-chain and off-chain system design. Inf Process Manage 2021;58(3):102535. [doi: 10.1016/j.ipm.2021.102535]

XSL•FO
RenderX

31. Benet J. IPFS-content addressed, versioned, P2P file system. Preprint posted online July 14, 2014. Comput Sci :3561 [FREE Full text]
32. Tu P, Gao Y, Lungu AJ, Li D, Wang H, Chen X. Augmented reality based navigation for distal interlocking of intramedullary nails utilizing Microsoft HoloLens 2. Comput Biol Med 2021;133:104402. [doi: 10.1016/j.compbiomed.2021.104402] [Medline: 33895460]
33. Liebmann F, Roner S, von Atzigen M, Scaramuzza D, Sutter R, Snedeker J, et al. Pedicle screw navigation using surface digitization on the Microsoft HoloLens. Int J Comput Assist Radiol Surg 2019;14(7):1157-1165. [doi: 10.1007/s11548-019-01973-7] [Medline: 30993519]
34. Sodhro AH, Zahid N. AI-enabled framework for fog computing driven e-healthcare applications. Sensors (Basel) 2021;21(23):8039 [FREE Full text] [doi: 10.3390/s21238039] [Medline: 34884048]

Abbreviations

NFT: nonfungible token

Edited by G Eysenbach; submitted 01.02.23; peer-reviewed by M Brown, Y Zhang; comments to author 19.07.23; revised version received 26.09.23; accepted 24.03.24; published 28.05.24.

Please cite as:
Wang G, Chen C, Jiang Z, Li G, Wu C, Li S
Efficient Use of Biological Data in the Web 3.0 Era by Applying Nonfungible Token Technology
J Med Internet Res 2024;26:e46160
URL: https://www.jmir.org/2024/1/e46160
doi:10.2196/46160
PMID:38805706

©Guanyi Wang, Chen Chen, Ziyu Jiang, Gang Li, Can Wu, Sheng Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

An Infrastructure Framework for Remote Patient Monitoring Interventions and Research

Jennifer Claggett[1,2], PhD; Stacie Petter[1], PhD; Amol Joshi[1,2], PhD; Todd Ponzio[3], PhD; Eric Kirkendall[2], MD

[1]School of Business, Wake Forest University, Winston-Salem, NC, United States

[2]Center for Healthcare Innovation, School of Medicine, Wake Forest University, Winston-Salem, NC, United States

[3]Health Science Center, University of Tennessee, Memphis, TN, United States

Corresponding Author:
Jennifer Claggett, PhD
School of Business
Wake Forest University
1834 Wake Forest Rd
Winston-Salem, NC, 27109-6000
United States
Phone: 1 3363027991
Email: claggett@gmail.com

Abstract

Remote patient monitoring (RPM) enables clinicians to maintain and adjust their patients' plan of care by using remotely gathered data, such as vital signs, to proactively make medical decisions about a patient's care. RPM interventions have been touted as a means to improve patient care and well-being while reducing costs and resource needs within the health care ecosystem. However, multiple interworking components must be successfully implemented for an RPM intervention to yield the desired outcomes, and the design and key driver of each component can vary depending on the medical context. This viewpoint and perspective paper presents a 4-component RPM infrastructure framework based on a synthesis of existing literature and practice related to RPM. Specifically, these components are identified and considered: (1) data collection, (2) data transmission and storage, (3) data analysis, and (4) information presentation. Interaction points to consider between components include transmission, interoperability, accessibility, workflow integration, and transparency. Within each of the 4 components, questions affecting research and practice emerge that can affect the outcomes of RPM interventions. This framework provides a holistic perspective of the technologies involved in RPM interventions and how these core elements interact to provide an appropriate infrastructure for deploying RPM in health systems. Further, it provides a common vocabulary to compare and contrast RPM solutions across health contexts and may stimulate new research and intervention opportunities.

(J Med Internet Res 2024;26:e51234) doi:10.2196/51234

KEYWORDS

remote patient monitoring; eHealth; telehealth; telemonitoring; telemedicine; digital infrastructure; clinical decision-making

Introduction

Overview

Remote patient monitoring (RPM; sometimes referred to as eHealth, telehealth, telemonitoring, or telemedicine) involves the capture of patient data through sensors or devices outside of a clinical setting, such as at the patient's home or work while the patient is engaging in everyday activities. Ideally, the data captured through RPM devices are analyzed and used to inform clinicians' decisions on patient care. For example, typical decisions include adjusting the recommended dosage or timing of a patient's medication based on observed changes in the patient's vital signs or patterns of activity.

RPM interventions have increased exponentially in the United States of America since 2020 [1]. The COVID-19 pandemic exacerbated the need for remote patient care solutions when there were severe resource shortages of clinicians, equipment, and capacity within health care systems [2-4] and patients were required to socially distance themselves to mitigate the spread of COVID-19. As the United States of America eased regulations and made changes to encourage reimbursements for RPM interventions, health care providers sought to reap RPM's potential benefits along three main dimensions: (1) enhancing quality by offering more personalized care; (2) achieving scale by growing their customer (patient) base; and (3) securing new reimbursement opportunities by evolving in response to shifts in payment policies [1,3].

XSL•FO
RenderX

The excitement and promise of the benefits of RPM to improve patient care while also expanding a health system's market are well-documented in meta-analyses that find evidence of RPM reducing hospital admissions and length of stay for certain conditions, such as cardiovascular disease or chronic obstructive pulmonary disease [5,6]. Decreased travel time, cost savings, and increased access to services are commonly ascribed as benefits for patients, and most eHealth interventions are described as successes [7]. However, other scholars counter that RPM interventions may not live up to the hype. One study finds that RPM interventions do not impact patient health factors, such as weight, body fat percentage, and blood pressure [8], and other related studies raise concerns about the limited evidence that RPM interventions can indeed adequately scale to meaningfully improve patient outcomes and demonstrably reduce health care costs [3].

These mixed results regarding the impact of RPM interventions showcase the current challenge of understanding how to design and effectively implement the infrastructure to support successful RPM programs. Successful RPM programs should meet at least one, but ideally both, of the following standards: (1) improved management of symptoms (evaluated using population-normalized values or patient feedback) and (2) reduced financial costs (evaluated in terms of the health system, payers, and patient out-of-pocket expenses). Previous work reporting on RPM interventions tends to report details on isolated projects and is focused, understandably, on a specific medical condition without offering generalizable advice to a broader audience or a catalog of best practices. Although RPM has been implemented in many different types of contexts, we contend that the key infrastructure points are consistent across interventions. Therefore, we present a framework consisting of 4 core infrastructure components necessary for any RPM intervention and identify common questions across contexts that should influence the RPM intervention design and results. This RPM infrastructure framework is useful to scholars and clinicians implementing RPM projects in that it (1) presents a shared vocabulary and reference point, (2) serves as a resource to guide some of the major decisions associated with an RPM implementation, and (3) provides a logical scaffolding to categorize and disseminate lessons learned within RPM projects to leverage them in other contexts. While the set of considerations nested within the four infrastructure components is not exhaustive, these considerations serve as a useful starting point as RPM research and interventions are planned and developed in the future.

RPM Infrastructure Framework

As an information technology (IT), RPM relies on a combined and layered infrastructure of hardware, software, and networks to support the collection, storage, processing, and management of data. By considering emergent patterns and themes from the literature, cases, and reports, discussing this topic in various panels and workshops, and reflecting on our experiences designing and assessing RPM projects, we propose a four-component infrastructure framework that is necessary in any RPM infrastructure project: (1) data collection, (2) data transmission and storage, (3) algorithmic data analysis, and (4) information presentation. The first RPM infrastructure component, data collection, collects a patient's vital signs and other biometric data remotely through a measurement device such as a wearable sensor. Data transmission and storage, the second infrastructure component, leverages software interface services, networking, and hardware to transfer the data from the patient's device to a centralized data architecture [9,10]. Third, software-based algorithms analyze the stored remote patient data to identify patterns and outliers for a single patient or for a patient population. The final RPM infrastructure component is to present information obtained from the analysis to support clinicians' decision-making processes [11,12]. Figure 1 depicts the RPM infrastructure framework, and each of the following sections describes key considerations for each component.

XSL•FO
RenderX

Figure 1. Remote patient monitoring (RPM) infrastructure framework.

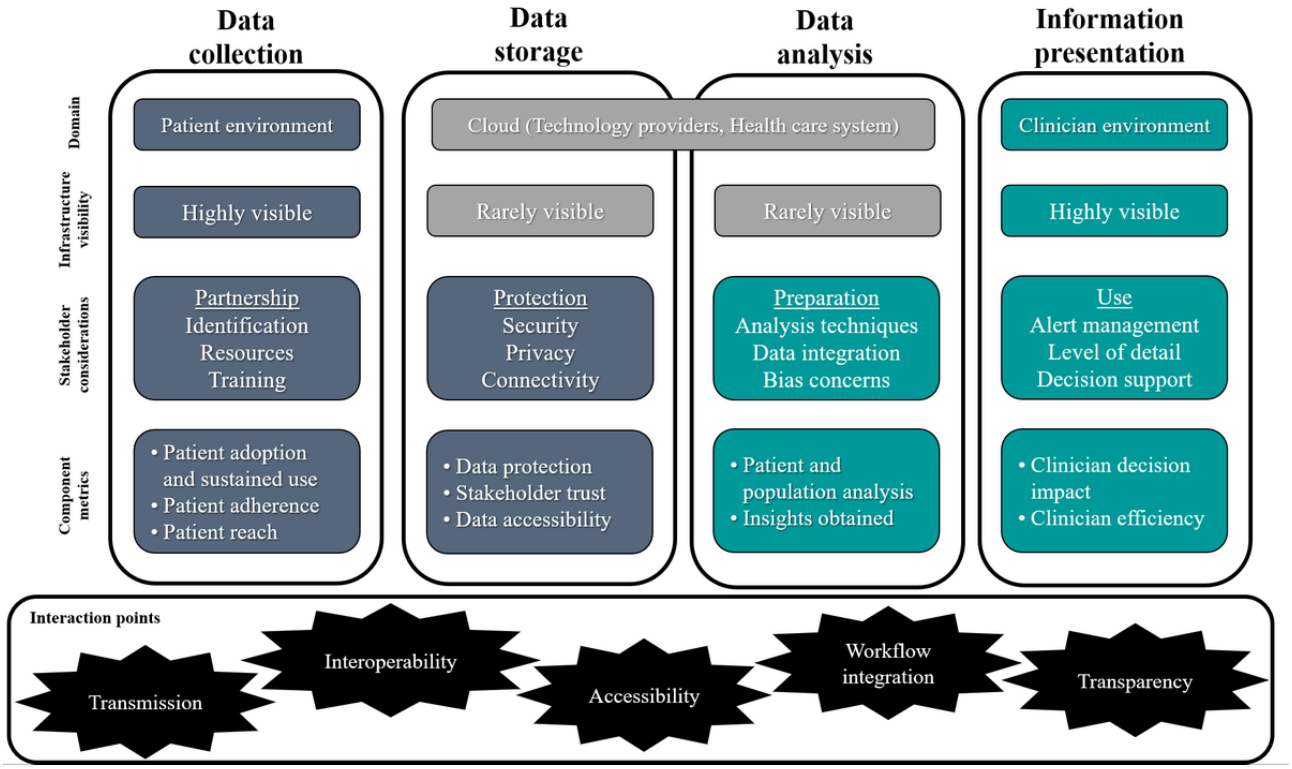

Component 1: Remote Patient Data Collection

Overview

Patients interact with an RPM device to enable the collection of data outside of clinical settings. Some devices are worn continuously throughout a person's day, while other devices are used at specific times to capture health indicators periodically based on the patient's medical condition and the provider's care protocol. Patients may use a specialized RPM device that registers a single form of biometric data (eg, a continuous glucose monitor capturing blood glucose levels) or a device that captures multiple data types (eg, a blood pressure cuff that measures blood pressure, pulse rate, and oxygen saturation). Given the growing number of technologies capable of collecting patient health data along with the need for patients to interact with a device for data collection, several questions must be carefully addressed when considering how to best collect data for an RPM intervention.

How Should Patients be Selected for RPM?

While RPM has the potential to improve patients' quality of care and reduce clinic costs, successful implementation relies on the effective use of the device and the fidelity of the collected data. The existing literature highlights several key considerations and components for identifying patients who are a good match for remote monitoring. Of paramount importance is suitability—is the patient's medical condition one that is likely to actually benefit from the collection and analysis of more data? Patients with chronic diseases such as diabetes, heart failure, hypertension, or chronic obstructive pulmonary disease are often more likely to benefit from RPM, as it can help them better manage their health status and condition over the long-term [13]. Comorbidities also play a significant role in patient selection, as those with multiple chronic conditions or complex health situations might require more comprehensive monitoring [14]. RPM can provide a more holistic view of their health, making it a potentially valuable tool for these patients; however, the complexity of their medical conditions may limit their ability to adhere to the monitoring program and necessitate more immediate and direct medical interventions.

Patients who are noncompliant or have a history of difficulty adhering to their treatment plans might benefit from RPM, as it can help improve compliance and provide additional support [15]. RPM solutions may make a patient feel more engaged, empowered, and informed through messaging systems that interact with patients on a routinely structured basis [16,17]. Patient motivation and engagement are key factors, as patients who are motivated and engaged in managing their health are more likely to actively participate in and adhere to the RPM program [18].

Other patient-specific factors—commonly referred to as the social determinants of health—such as socioeconomic status, age, and social support should be considered when designing RPM interventions [19]. For instance, patients with lower socioeconomic status might benefit from RPM the most, as it can help reduce health care disparities and provide better access to care [20-22]. A disproportionally large number of people affected by chronic conditions are from socioeconomically disadvantaged groups [23]. Communities of color, immigrants, and women are particularly likely to be in distress from undiagnosed chronic diseases, and even when diagnosed, these populations are more likely than their counterparts to face structural and logistical obstacles to obtaining the appropriate

XSL•FO
RenderX

level of intermittent care. So long as they have reliable connectivity to the internet, patients who live in remote or rural areas or have limited access to transportation might benefit from RPM, as it can help overcome geographical barriers to care [18,24]. Age can also play a role in identifying suitable patients for RPM, in that elderly patients or those with age-related conditions may benefit from RPM. The patient's living situation is another important factor. A strong support system, such as family or caregivers, can facilitate device use, data collection, and overall engagement, making these patients more suitable for RPM [25].

Finally, technological competence plays a crucial role in a patient's ability to engage with an RPM device. Patients with some level of technology literacy (eg, "digital natives") are more likely to engage with and effectively use RPM devices and systems [26]. However, patients with lower socioeconomic status or those who are elderly may have lower levels of technological competence or may have other barriers that could limit the effectiveness of an RPM program [27-29]. There is a natural continuum of sophistication and familiarity with devices and the inevitable troubleshooting they often require, and a more "set and forget" approach may be advisable for certain populations.

Which Device and Which Types of Data?

A fundamental characteristic of RPM is the acquisition of data outside of conventional clinical environments. Consequently, patient data must be collected remotely using sensors and equipment such as wearable devices, mobile phones, or portable devices installed at a patient's residence or other environments [30]. One strategy involves using data from off-the-shelf, general-purpose smart health consumer electronics purchased by the patient, while another option is to rely on data from specialized devices or software prescribed or supplied by the health care provider. Technological advancements enable the collection, through devices within the RPM infrastructure, of various types of data, such as electrocardiograms, electroencephalograms, heartbeats and respiration rates, oxygen saturation in the blood or pulse oximetry, nervous system signals, blood pressure, body or skin temperature, blood glucose levels, patient weight, and sleep patterns, among others [31].

A crucial consideration is the optimal combination of metrics to be collected for a specific patient. The US Centers for Disease Control report that 51.8% of US adults have at least one chronic condition, and 27.2% have multiple chronic conditions such as obesity, diabetes, and cardiovascular disease [32]. Emerging evidence indicates that RPM initiatives are more likely to succeed when multiple metrics are evaluated concurrently [33]. For instance, compiling data from various physiological sensors measuring heart rate, blood oxygen saturation, and blood glucose levels simultaneously can offer a more comprehensive overview of a patient's health, which is particularly significant for patients with comorbidities and additional complications. This suggests that the diagnostic value of data can be enhanced by carefully considering what health indicators are needed to manage a patient's care.

How Frequently are Data Collected?

Determining the optimal frequency of data collection in RPM scenarios is a critical consideration, as it can significantly impact the effectiveness of patient care and the efficient use of health care resources. The appropriate frequency for data collection depends on various factors, including the severity and type of the patient's condition, the objectives of monitoring, and the required patient involvement in data collection [34]. For instance, some conditions may necessitate multiple data readings per day, while others may only require weekly monitoring [35]. Passive data collection methods, such as continuous monitoring of vital signs using wearable sensors, can be advantageous for patients requiring frequent monitoring, whereas active data collection methods, which involve patient involvement and interaction, may be suitable for other conditions [16,36]. Passive methods are usually less likely to cause patient burnout and abandonment [37].

Health care providers should consider adopting several best practices to ensure that patients remain engaged and compliant with RPM protocols. These typically include providing personalized and clear instructions, offering training and support to ensure device functionality, improving patients' understanding and comfort with the technology, and fostering regular remote communication between patients and health care providers [38]. Furthermore, involving patients in the decision-making process regarding their monitoring plans and adjusting the frequency and type of data collection based on their individual needs and preferences can lead to increased patient engagement and satisfaction [13,39].

Component 2: Remote Patient Data Transmission and Storage

Overview

Once remote patient data are collected by one or more devices, the data must be transmitted and shared with clinicians, and stored in a data architecture. The manner in which the data are transmitted from an RPM device is dependent on the device and the network access of the patient. RPM data transmission may occur through a network using a wired link, or high-speed wireless link with or without human intervention. In some cases, patients or caregivers may be asked to record readings or values from devices into an app available on their smartphone or computer that will transmit data to the medical provider. Another option could be that a patient must physically visit a clinician's office with the device to upload the data to the patient's electronic medical record. The storage of remote patient data may be in a system that is managed by the device manufacturer and accessed through a web portal, and the data may or may not be integrated within the patient's electronic medical record.

Is There Sufficient Connectivity?

Connectivity plays a vital role in the successful implementation of RPM, as it enables the transmission of patient data from monitoring devices to health care providers and fosters timely interventions and informed decision-making. Addressing the digital divide is crucial to ensuring equitable access to RPM services, as patients with limited internet access or low digital

XSL•FO
RenderX

literacy may face barriers to fully benefiting from RPM [40]. This disparity is particularly concerning for patients from socioeconomically disadvantaged backgrounds, who may experience greater difficulties in accessing health care services and could benefit the most from RPM [40,41]. Some patients may have access to home internet solutions through local internet service providers that include Wi-Fi networks at home, while others may be limited to cellular network access through mobile devices. Often, the latter is subject to slower connections and data caps that place constraints on the patient's connectivity.

Strategies for addressing connectivity for RPM interventions should consider alternatives, such as the constant connectivity approach or using batch or episodic data uploads when data connections are available [42]. Constant connectivity can facilitate real-time monitoring and immediate interventions, which may be especially beneficial for patients with critical or rapidly changing health conditions [43]. However, this approach may not be feasible for patients living in areas with limited or unreliable internet access or for those who cannot afford consistent connectivity. In these cases, episodic data uploads when a connection is possible may provide a more accessible and cost-effective solution, allowing health care providers to track patient progress and identify potential issues while accommodating the patients' connectivity limitations [44]. Additionally, some RPM hardware solutions may include direct cellular network connectivity, where the device sends the data through a connection provided by the wearable device to the provider, bypassing the need for a patient home network. These solutions will incur additional costs related to data transmission and may not naturally provide a patient dashboard or a way for patients to easily view data that may traditionally be housed in a patient application.

Is the Transmission Secure?

The sensitive nature of medical data necessitates robust protection measures to maintain patient privacy and prevent unauthorized access. Data breaches and cyberattacks can have severe consequences for patients and their health care providers, including identity theft, financial loss, and reputational damage [45]. The increasing connectivity of medical devices and the use of cloud-based data storage have created new opportunities for cybercriminals, leading to the emergence of threats such as medjacking [46]. Medjacking, a term coined from "medical device hijacking," refers to the unauthorized access and manipulation of medical devices, such as pacemakers or insulin pumps, to cause harm to patients or extract sensitive data [47]. As RPM technologies rely on a variety of connected devices for data collection across multiple networks, they can be vulnerable to medjacking and other cybersecurity risks. Furthermore, the rapid expansion of the internet of things in health care has amplified these risks, as a larger number of interconnected devices create more potential entry points for attackers [48,49].

Health care providers and technology developers should prioritize the implementation of robust security measures to mitigate the risks associated with medjacking and other security threats in RPM. These may include strong encryption protocols for data transmission ("in flight") and storage ("at rest"), regular security updates, and the development of secure communication channels between devices and health care providers [45,48]. Additionally, health care organizations should adopt a proactive approach to security by conducting regular risk assessments, promoting cybersecurity awareness and training among staff, and fostering a culture of security-mindedness [50].

Can Data Move Across Health Systems Software?

Interoperability is a crucial aspect of RPM projects, as it enables seamless communication and data sharing among different health information systems, devices, and providers. This encompasses not only the technical aspects of data exchange but also the semantic understanding and interpretation of shared data, ensuring that the information can be effectively used by health care providers, patients, and other stakeholders. Effective interoperability contributes to improved patient care by ensuring that clinicians have access to comprehensive and up-to-date medical information, allowing for better decision-making and coordination of care [51]. However, achieving interoperability in RPM poses several challenges, including the need to balance data accessibility with patient privacy and maintain control over personal health information.

One of the primary challenges in achieving interoperability in RPM is the heterogeneity of health information systems and devices used by health care providers. These systems often rely on different (often proprietary) data formats, communication protocols, and standards, which can create barriers to effective information exchange. To address this issue, several major standards have been developed to facilitate interoperability in health IT (eg, [52,53]). For example, the US Department of Health and Human Services Office of the National Coordinator for Health IT released the third version (V3) of the US Core Data for Interoperability in 2022 [54].

Another challenge in achieving interoperability is protecting patient privacy while sharing data freely among authorized health care providers [44]. Using privacy-preserving techniques, such as pseudonymization, which replaces personally identifiable information with unique identifiers to maintain patient anonymity, may reinforce privacy during the transmission of data between systems. However, these approaches must be rigorously tested to systematically mitigate privacy risks [55]. One-way hashing of sensitive identifiers is another technique that can reduce the risk of leakage of personal health identifiers. Additionally, the implementation of access control mechanisms can help ensure that only authorized users can access and share patient data, further safeguarding privacy [56].

A related issue to moving data across health systems is determining the appropriate granularity to share between stakeholders and systems. For example, in a remote blood pressure monitoring project, should each reading be recorded, transmitted, and made available, including any relevant metadata about time, place, and cuff placement, or should only summary data about daily or weekly averages be shared between systems? Like any sensor-based technology, the amount of raw data generated by RPM initiatives may be overwhelming [57]; however, providing only summarized data limits the transparency and future uses of the data.

XSL•FO
RenderX

Component 3: Algorithmic Analysis of Remote Patient Data

Overview

Remote patient data that are stored within an information system and are not analyzed provide no value to the patient or the clinician. After transmitting and storing RPM data, they should be processed and analyzed to identify and summarize patterns and trends in individual patients and patient populations [58]. The process of analyzing raw data to deliver actionable insights could also form the basis for financial reimbursement, which is fundamental to any sustainable RPM program.

What Analysis Techniques are Appropriate?

Data analysis involves the use of algorithms, or a series of steps, to process the data in a meaningful way. Algorithms may use static rule logic, which can be used to draw attention to results over a certain threshold, or they may leverage machine learning techniques to dynamically adapt and learn from large sets of patient data, such as adjusting the threshold based on similar patients with similar conditions recorded in the data [59]. The distinction between static and dynamic rules has implications that need to be explored.

Static rules can be based on established medical guidelines, such as thresholds for vital signs or other clinical parameters, which can help health care providers identify potential health issues and take appropriate actions [60]. While this method can be effective in some cases, it may not account for the unique characteristics and complexities of individual patients, which may limit its ability to provide personalized care [61].

Alternatively, machine learning techniques offer more advanced and adaptable solutions for analyzing RPM data [62]. These techniques use algorithms that can learn from data patterns and make predictions or decisions without being explicitly programmed [63,64]. Machine learning can be used to identify trends, anomalies, and correlations in patient data, enabling health care providers to make more informed decisions and deliver personalized care [65,66]. Adaptive interpretation techniques take RPM data analysis a step further by dynamically adjusting their approach based on real-time patient data. These methods, which often rely on artificial intelligence and machine learning algorithms, can continuously refine their analysis and predictions to better understand the evolving health status of individual patients [63]. This adaptive approach can help health care providers identify subtle changes in patients' conditions that may not be evident through traditional analysis techniques, leading to more proactive and personalized care [67].

Which Comorbidities Should be Included in the Analysis?

This question centers around the appropriate complexity level of analyses of RPM solutions. Incorporating comorbidities into the analysis of RPM data can help health care providers better understand the complex interactions between various conditions and their impact on patients' health. This, in turn, can lead to more accurate and personalized treatment recommendations. Static rules that solely focus on a single condition, such as high blood pressure, may not adequately account for the impact of comorbidities on patients' overall health status. For instance, a patient with both diabetes and hypertension may require a different treatment approach than a patient with hypertension alone, which is why any given individual should be managed holistically with a consolidated approach, rather than divided by symptoms and specialty [68].

This comprehensive monitoring can provide a more accurate representation of the patient's health status, allowing health care providers to make more informed decisions regarding treatment and care management [69,70]. However, these solutions may be so patient-specific that cognitive efficiencies and the ability to scale the solution are compromised in the absence of built-in coordination systems with well-defined decision-making heuristics and robust care protocols.

What Biases Exist Within the Analysis and How Should They be Mitigated?

Biases in the analysis of remote patient data can have a significant impact on the accuracy and effectiveness of health care services. Particularly in machine learning-based analysis techniques, biases can arise from various sources, such as data sampling, measurement errors, or algorithmic design, leading to potentially biased predictions or recommendations [71,72]. It is essential to detect and account for biases to ensure that the solutions provided are equitable and reliable for all patients.

One primary source of bias in data analysis is the data itself. If the training data used to develop machine learning models do not accurately represent the diverse patient population, the resulting models may be skewed toward specific subgroups, leading to suboptimal or even harmful recommendations for other groups [73,74]. For instance, if a model is trained predominantly on data from patients of a particular age, gender, or ethnicity, it may not perform well on patients from other demographics. To mitigate such biases, it is crucial to ensure that the training data are representative of the target patient population, considering factors such as age, gender, ethnicity, and socioeconomic status [75].

Another source of bias can arise from the choice of features or variables used in the analysis. If certain relevant variables are not included, or if irrelevant variables are considered, the resulting predictions or recommendations may be biased or even spurious [76]. Careful feature selection, based on domain knowledge and a thorough understanding of the underlying data, can help address this issue.

Algorithmic biases can also emerge from the choice of machine learning methods or algorithms, as well as their specific implementations. To address this, it is essential to evaluate and compare multiple algorithms and implementations to identify potential biases and select the most appropriate method for the specific application [77]. Patients themselves can serve as their own baselines too, particularly for measurements that do not lend themselves as easily to a population approach (eg, mood and gastric motility).

Lastly, ongoing monitoring and evaluation of the performance of data analysis solutions, including machine learning models, is critical to detecting and addressing biases. Regular

XSL•FO
RenderX

assessments of model performance, particularly with respect to various subgroups within the patient population, can help identify potential biases and ensure that the solutions remain equitable and effective for all patients [78].

Component 4: Presentation of RPM Data to a Clinician

Overview

Once the data have been analyzed, the results need to be presented as information to support clinicians' decision-making. Unless the RPM data are used to inform patient care, the RPM intervention will not yield the intended results. Therefore, it is critical that the information is presented in a manner that is likely to inform clinicians as they make decisions that affect specific patients and patient populations.

Is RPM Information Accessible in the Right Electronic Health Record Software?

Physicians and other clinical decision makers often face significant time constraints and high cognitive workloads in their daily practice, making it challenging for them to manage and monitor patient data effectively. A study by Sinsky et al [79] found that primary care physicians spent nearly half of their workday interacting with EHR systems, leaving them with limited time for direct patient care. The high volume of clinical tasks and responsibilities can lead to cognitive overload, increasing the risk of burnout and negatively impacting the quality of care provided [80]. Given these constraints, it is critical to ensure that RPM data are easily accessible within the existing EHR systems without requiring clinicians to log into additional platforms or apps. Integrating RPM data into EHRs can help streamline clinical workflows and reduce the cognitive burden on health care providers, enabling them to focus on essential tasks such as patient evaluation, diagnosis, and treatment planning [81]. This underscores the importance of seamless integration and interoperability between RPM solutions and EHR systems, ultimately supporting more efficient and effective patient care by easing the pathway of the information being used in decision-making.

One of the key benefits of integrating RPM data into EHR systems is the ability to provide a comprehensive and up-to-date view of a patient's health status. By combining RPM data with other health information such as medical history, laboratory results, and imaging studies, clinicians can gain a more holistic understanding of a patient's condition, enabling them to make more informed decisions about treatment plans and care management strategies [82].

Integration of RPM data into EHR systems can also support the development and implementation of clinical decision support (CDS) tools, which can help health care providers make more informed, evidence-based decisions about patient care [83]. By leveraging RPM data, CDS tools can provide real-time alerts or recommendations to clinicians, assisting them in diagnosing, treating, or managing a patient's condition more effectively.

How Should the Decision Maker Receive Information?

In the context of RPM solutions, there is a delicate balance between providing exception reporting and summary data reporting. Exception reporting involves the generation of alerts or notifications only when specific events or abnormal values are detected, which require immediate attention from health care providers. This yields the advantage of focusing health care providers' attention on situations that need prompt intervention, potentially improving the efficiency and timeliness of care and reducing the number of alerts [84]. However, exception reporting may not always provide sufficient context or information about a patient's overall health status, making it difficult for clinicians to assess the impact of treatment strategies or identify more subtle changes in condition over time. On the other hand, summary data reporting provides a broader overview of a patient's progress over time, allowing clinicians to evaluate trends and assess the overall effectiveness of treatment plans. Both approaches have their merits and challenges, making the choice between them a critical consideration in RPM projects.

Alert fatigue is a critical concern in the context of RPM solutions, as it can have significant implications for the effectiveness of the system and the quality of patient care. Alert fatigue occurs when health care providers are exposed to a high volume of alerts, leading to desensitization and potentially reduced responsiveness to these notifications [85-87]. This phenomenon has been observed in various clinical settings, including electronic health record systems and CDS tools, where excessive alerts can contribute to cognitive overload, increased stress, and the risk of overlooking critical information [88].

In RPM systems, balancing the type and frequency of messaging is essential to minimize alert fatigue. The choice between push and pull messaging strategies can play a significant role in this regard. Push messaging involves automatically sending alerts or notifications to health care providers, whereas pull messaging requires providers to actively request or retrieve the information. Although push messaging can ensure timely delivery of critical information, it may also contribute to alert fatigue if used indiscriminately or too frequently. Solutions to alleviate this tension may involve tailoring alert thresholds based on individual patient needs, incorporating CDS algorithms to filter and prioritize alerts, and using a combination of push and pull messaging to strike the right balance between proactively notifying providers and allowing them to access information on demand.

What is the Right Amount of Information to Provide to Decision Makers?

Balancing transparency and detail in the presentation of RPM data with cognitive ease is crucial for ensuring that health care providers effectively use the information in their decision-making processes. While transparency is essential for building trust and understanding of the underlying data analysis, providing excessive detail can overwhelm clinicians and hinder their ability to quickly assimilate the information [89]. Consequently, it is vital to strike an optimal balance between presenting comprehensive information and ensuring cognitive ease for end users.

XSL•FO
RenderX

One approach to achieving this balance is to use a tiered or "drill-down" presentation of data, which allows health care providers to access additional layers of detail only if they require it [90]. This design can present a high-level summary of the patient's condition and only flag critical alerts, while enabling providers to delve deeper into the data if they desire further context or clarification. This, in turn, helps mitigate information overload and supports more efficient decision-making by prioritizing the most relevant and actionable insights [91]. Moreover, incorporating the principles of cognitive ergonomics and human-centered design can further enhance the usability of RPM solutions. This may involve the use of visual aids, such as graphs, charts, and color-coding, to facilitate rapid comprehension of complex data and even presenting proposed treatment plans based on the algorithmic analysis of the patient's full record [92] and providing reference statistics from the health system's relevant patient population.

Conclusion

The mixed results with RPM interventions have raised concerns about the scalability and value of this technology. This viewpoint paper highlights some of the key questions and core considerations that affect the various infrastructure components of an RPM intervention. Differences between health conditions, metrics, devices, storage, analysis, and information presentation across RPM implementations result in countless permutations. If scholars fail to document and clearly explain the RPM infrastructure and choices made for an RPM implementation, it will be difficult to create an evidence-based research tradition. Having a shared vocabulary and more consistent documentation of the RPM infrastructure can support future literature reviews and meta-analyses seeking to evaluate the outcomes of RPM interventions. The RPM infrastructure framework presented in this article offers scholars a means to describe the different choices and constraints associated with their RPM interventions.

We also identify how each of the infrastructure components can stimulate new research and intervention opportunities in Table 1. While not exhaustive, the list offers a sampling of the many research questions that could be studied to further increase the understanding associated with RPM interventions. The RPM framework offers scholars and clinicians a more comprehensive guide to exploring various aspects of RPM implementation. As a result, they can further optimize the design and functionality of RPM solutions for improved patient care and health care provider experiences.

XSL•FO
RenderX

Table 1. Key questions and considerations per remote patient monitoring (RPM) infrastructure component.

Infrastructure component and relevant questions	Research considerations	Intervention considerations
Data collection		
Which patients?	• Examining lesser-studied chronic and acute health conditions that could benefit from RPM interventions • Exploring which comorbidities benefit from RPM data collection • Determining ways to improve patient compliance with RPM protocols • Understanding how to identify and improve technology competence among patients suitable for RPM interventions	• Using RPM interventions for health conditions with known positive patient outcomes • Collecting additional data through RPM based on the comorbidities of the patient • Identifying patient populations most likely to benefit from RPM intervention outcomes • Developing strategies to increase technology competence among patient populations who can benefit from RPM interventions
Which device and data?	• Comparing the effectiveness of device types (eg, invasive vs noninvasive and prescribed devices vs consumer electronics) for RPM • Identifying which metrics (and combinations of metrics) are most useful (and least obtrusive) to collect that will yield the desired patient outcomes	• Choosing an RPM device by considering the patient's needs and the type of data needed to provide patient care • Determining which metrics are most appropriate to capture for a patient given their health condition and comorbidities • Evaluating the business models for device purchases and maintenance
How frequent?	• Determining best practices for the frequency of RPM data collection while considering the patient's health condition and likelihood of adhering to the schedule	• Reviewing evidence-based practices to determine the frequency of RPM data collection for clinical care pathways
Data transmission and storage		
How to transmit data?	• Identifying systematic bias or limitations in transmitting RPM data due to lack of access to broadband, technology competence, or ability to physically visit a clinician office	• Determining the data transmission approach most feasible for the patient • Providing connectivity through the device (direct cell network transmission) versus leveraging patient networks and device apps
How to secure data?	• Determining means to limit the vulnerability of RPM data when transmitted and stored	• Informing patients of the security, privacy, and risks of their RPM data during transmission and storage of data
How interoperable?	• Creating robust methods to maintain patient privacy when sharing RPM data across health care information systems • Determining the appropriate level of granularity of data to share (raw vs summary)	• Leveraging interoperability standards to ensure RPM data can be shared across systems (eg, from a device's proprietary system to an electronic health record)
Data analysis		
What algorithm?	• Examining the advantages and disadvantages associated with different forms of algorithms and rules (ie, static vs dynamic) used to analyze RPM data	• Understanding the implications associated with the manner in which RPM data are analyzed
Which co-morbidities?	• Identifying the interactions between comorbidities affect clinical care pathways when using RPM interventions	• Determining which comorbidities to control for or consider when implementing RPM for a patient population
What is the potential for bias?	• Benchmarking and validating algorithms to detect and mitigate bias	• Considering how biases within the patient's RPM data (or across multiple patients' RPM data) could influence the findings or recommendations
Information presentation		
How to make information accessible?	• Evaluating the outcomes of patients when RPM information is embedded within electronic health record systems	• Ensuring the RPM information is part of the clinician's workflow

XSL•FO
RenderX

Infrastructure component and relevant questions	Research considerations	Intervention considerations
How to share information for decision-making?	• Determining how to present information effectively to improve decision-making while avoiding cognitive overload • Examining methods of presentation that encourage clinicians to use the results to augment (rather than replace) human judgment	• Designing information presentation to avoid alert fatigue • Developing alternative strategies, such as digital nudges, to encourage clinicians to apply RPM data in patient care decisions
How to present information for decision-making?	• Identifying factors (eg, trust, transparency, and usability) related to RPM information that affect clinician's decision-making	• Exploring presentation methods to enable effective use of RPM information for decision-making

Acknowledgments

This research is supported in part by funding from Wake Forest University through the School of Business and the School of Medicine Center for Healthcare Innovation. These sponsors had no involvement in the research.

Conflicts of Interest

None declared.

References

1. Tang M, Mehrotra A, Stern AD. Rapid growth of remote patient monitoring is driven by a small number of primary care providers. Health Aff (Millwood) 2022;41(9):1248-1254. [doi: 10.1377/hlthaff.2021.02026] [Medline: 36067430]
2. Birkmeyer JD, Barnato A, Birkmeyer N, Bessler R, Skinner J. The impact of the COVID-19 pandemic on hospital admissions in the United States. Health Aff (Millwood) 2020;39(11):2010-2017 [FREE Full text] [doi: 10.1377/hlthaff.2020.00980] [Medline: 32970495]
3. Mecklai K, Smith N, Stern AD, Kramer DB. Remote patient monitoring: overdue or overused? N Engl J Med 2021;384(15):1384-1386. [doi: 10.1056/NEJMp2033275] [Medline: 33853209]
4. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
5. de Farias FAC, Dagostini CM, de Assunção Bicca Y, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
6. Taylor ML, Thomas EE, Snoswell CL, Smith AC, Caffery LJ. Does remote patient monitoring reduce acute care use? A systematic review. BMJ Open 2021;11(3):e040232 [FREE Full text] [doi: 10.1136/bmjopen-2020-040232] [Medline: 33653740]
7. LeBlanc M, Petrie S, Paskaran S, Carson DB, Peters PA. Patient and provider perspectives on eHealth interventions in Canada and Australia: a scoping review. Rural Remote Health 2020;20(3):5754 [FREE Full text] [doi: 10.22605/RRH5754] [Medline: 32949485]
8. Noah B, Keller MS, Mosadeghi S, Stein L, Johl S, Delshad S, et al. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. NPJ Digit Med 2018;1(1):20172 [FREE Full text] [doi: 10.1038/s41746-017-0002-4] [Medline: 31304346]
9. Silva BMC, Rodrigues JJPC, de la Torre Díez I, López-Coronado M, Saleem K. Mobile-health: a review of current state in 2015. J Biomed Inform 2015;56:265-272 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.003] [Medline: 26071682]
10. Verma P, Sood SK. Fog assisted-IoT enabled patient health monitoring in smart homes. IEEE Internet Things J 2018;5(3):1789-1796. [doi: 10.1109/jiot.2018.2803201]
11. Gold R, Bunce A, Cowburn S, Dambrun K, Dearing M, Middendorf M, et al. Adoption of social determinants of health EHR tools by community health centers. Ann Fam Med 2018;16(5):399-407 [FREE Full text] [doi: 10.1370/afm.2275] [Medline: 30201636]
12. Kwon BC, Choi MJ, Kim JT, Choi E, Kim YB, Kwon S, et al. RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Trans Vis Comput Graph 2019;25(1):299-309. [doi: 10.1109/TVCG.2018.2865027] [Medline: 30136973]
13. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]

XSL•FO
RenderX

14. Bogyi P, Vamos M, Bari Z, Polgar B, Muk B, Nyolczas N, et al. Association of remote monitoring with survival in heart failure patients undergoing cardiac resynchronization therapy: retrospective observational study. J Med Internet Res 2019;21(7):e14142 [FREE Full text] [doi: 10.2196/14142] [Medline: 31350836]
15. Lippke S, Gao L, Keller FM, Becker P, Dahmen A. Adherence with online therapy vs face-to-face therapy and with online therapy vs care as usual: secondary analysis of two randomized controlled trials. J Med Internet Res 2021;23(11):e31274 [FREE Full text] [doi: 10.2196/31274] [Medline: 34730541]
16. Midaglia L, Mulero P, Montalban X, Graves J, Hauser SL, Julian L, et al. Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: nonrandomized interventional feasibility study. J Med Internet Res 2019;21(8):e14863 [FREE Full text] [doi: 10.2196/14863] [Medline: 31471961]
17. Rojas G, Guajardo V, Martínez P, Castro A, Fritsch R, Moessner M, et al. A remote collaborative care program for patients with depression living in rural areas: open-label trial. J Med Internet Res 2018;20(4):e158 [FREE Full text] [doi: 10.2196/jmir.8803] [Medline: 29712627]
18. Seto E. Cost comparison between telemonitoring and usual care of heart failure: a systematic review. Telemed J E Health 2008;14(7):679-686. [doi: 10.1089/tmj.2007.0114] [Medline: 18817497]
19. NEJM Catalyst. Social Determinants of Health (SDOH). NEJM Catalyst 2017;3(6) [FREE Full text]
20. Najarian M, Goudie A, Bona JP, Rezaeiahari M, Young SG, Bogulski CA, et al. Socioeconomic determinants of remote patient monitoring implementation among rural and urban hospitals. Telemed J E Health 2023;29(11):1624-1633. [doi: 10.1089/tmj.2022.0412] [Medline: 37010391]
21. Bailey JE, Gurgol C, Pan E, Njie S, Emmett S, Gatwood J, et al. Early patient-centered outcomes research experience with the use of telehealth to address disparities: scoping review. J Med Internet Res 2021;23(12):e28503 [FREE Full text] [doi: 10.2196/28503] [Medline: 34878986]
22. Yong J, Yang O. Does socioeconomic status affect hospital utilization and health outcomes of chronic disease patients? Eur J Health Econ 2021;22(2):329-339. [doi: 10.1007/s10198-020-01255-z] [Medline: 33389255]
23. Veisani Y, Jenabi E, Nematollahi S, Delpisheh A, Khazaei S. The role of socio-economic inequality in the prevalence of hypertension in adults. J Cardiovasc Thorac Res 2019;11(2):116-120 [FREE Full text] [doi: 10.15171/jcvtr.2019.20] [Medline: 31384405]
24. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018;42(11):214 [FREE Full text] [doi: 10.1007/s10916-018-1075-6] [Medline: 30269237]
25. Demiris G, Speedie SM, Finkelstein S. Change of patients' perceptions of TeleHomeCare. Telemed J E Health 2001;7(3):241-248. [doi: 10.1089/153056201316970948] [Medline: 11564360]
26. Birati Y, Yefet E, Perlitz Y, Shehadeh N, Spitzer S. Cultural and digital health literacy appropriateness of app- and web-based systems designed for pregnant women with gestational diabetes mellitus: scoping review. J Med Internet Res 2022;24(10):e37844 [FREE Full text] [doi: 10.2196/37844] [Medline: 36240008]
27. Tseng KC, Hsu CL, Chuang YH. Designing an intelligent health monitoring system and exploring user acceptance for the elderly. J Med Syst 2013;37(6):9967. [doi: 10.1007/s10916-013-9967-y] [Medline: 24037138]
28. Martínez A, Everss E, Rojo-Alvarez JL, Figal DP, García-Alberola A. A systematic review of the literature on home monitoring for patients with heart failure. J Telemed Telecare 2006;12(5):234-241. [doi: 10.1258/135763306777889109] [Medline: 16848935]
29. Rodríguez I, Herskovic V, Gerea C, Fuentes C, Rossel PO, Marques M, et al. Understanding monitoring technologies for adults with pain: systematic literature review. J Med Internet Res 2017;19(10):e364 [FREE Full text] [doi: 10.2196/jmir.7279] [Medline: 29079550]
30. Baig MM, GholamHosseini H, Moqeem AA, Mirza F, Lindén M. A systematic review of wearable patient monitoring systems: current challenges and opportunities for clinical adoption. J Med Syst 2017;41(7):115. [doi: 10.1007/s10916-017-0760-1] [Medline: 28631139]
31. Jagadeeswari V, Subramaniyaswamy V, Logesh R, Vijayakumar V. A study on medical internet of things and big data in personalized healthcare system. Health Inf Sci Syst 2018;6(1):14 [FREE Full text] [doi: 10.1007/s13755-018-0049-x] [Medline: 30279984]
32. Nguyen T, Barefield A, Nguyen GT. Social determinants of health associated with the use of screenings for hypertension, hypercholesterolemia, and hyperglycemia among American adults. Med Sci (Basel) 2021;9(1):19 [FREE Full text] [doi: 10.3390/medsci9010019] [Medline: 33806794]
33. Kalid N, Zaidan AA, Zaidan BB, Salman OH, Hashim M, Muzammil H. Based real time remote health monitoring systems: a review on patients prioritization and related "big data" using body sensors information and communication technology. J Med Syst 2018;42(2):30. [doi: 10.1007/s10916-017-0883-4] [Medline: 29288419]
34. Matthews M, Abdullah S, Gay G, Choudhury T. Tracking mental well-being: balancing rich sensing and patient needs. Computer 2014;47(4):36-43. [doi: 10.1109/MC.2014.107]

XSL•FO
RenderX

35. Maguire R, Connaghan J, Arber A, Klepacz N, Blyth KG, McPhelim J, et al. Advanced symptom management system for patients with malignant pleural mesothelioma (ASyMSmeso): mixed methods study. J Med Internet Res 2020;22(11):e19180 [FREE Full text] [doi: 10.2196/19180] [Medline: 33180025]
36. Simblett S, Greer B, Matcham F, Curtis H, Polhemus A, Ferrão J, et al. Barriers to and facilitators of engagement with remote measurement technology for managing health: systematic review and content analysis of findings. J Med Internet Res 2018;20(7):e10480 [FREE Full text] [doi: 10.2196/10480] [Medline: 30001997]
37. Hunt D. Passive analytics and the future of RPM (remote patient monitoring). Healthcare IT Today. 2021. URL: https://www.healthcareittoday.com/2021/02/12/passive-analytics-and-the-future-of-rpm-remote-patient-monitoring/ [accessed 2023-05-29]
38. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Analysing the role of complexity in explaining the fortunes of technology programmes: empirical application of the NASSS framework. BMC Med 2018;16(1):66 [FREE Full text] [doi: 10.1186/s12916-018-1050-6] [Medline: 29754584]
39. Su D, Michaud TL, Estabrooks P, Schwab RJ, Eiland LA, Hansen G, et al. Diabetes management through remote patient monitoring: the importance of patient activation and engagement with the technology. Telemed J E Health 2019;25(10):952-959. [doi: 10.1089/tmj.2018.0205] [Medline: 30372366]
40. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
41. Leidig M, Teeuw RM. Quantifying and mapping global data poverty. PLoS One 2015;10(11):e0142076 [FREE Full text] [doi: 10.1371/journal.pone.0142076] [Medline: 26560884]
42. Pappas G. Planning for internet connectivity in remote patient monitoring. Telemed J E Health 2010;16(5):639-641. [doi: 10.1089/tmj.2009.0137] [Medline: 20575734]
43. McConnell MV, Turakhia MP, Harrington RA, King AC, Ashley EA. Mobile health advances in physical activity, fitness, and atrial fibrillation: moving hearts. J Am Coll Cardiol 2018;71(23):2691-2701 [FREE Full text] [doi: 10.1016/j.jacc.2018.04.030] [Medline: 29880130]
44. Ancker JS, Witteman HO, Hafeez B, Provencher T, Van de Graaf M, Wei E. The invisible work of personal health information management among people with multiple chronic conditions: qualitative interview study among patients and providers. J Med Internet Res 2015;17(6):e137 [FREE Full text] [doi: 10.2196/jmir.4381] [Medline: 26043709]
45. Kramer DB, Baker M, Ransford B, Molina-Markham A, Stewart Q, Fu K, et al. Security and privacy qualities of medical devices: an analysis of FDA postmarket surveillance. PLoS One 2012;7(7):e40200 [FREE Full text] [doi: 10.1371/journal.pone.0040200] [Medline: 22829874]
46. Cuningkin V, Riley E, Rainey L. Preventing medjacking. Am J Nurs 2021;121(10):46-50. [doi: 10.1097/01.NAJ.0000794252.99183.5e] [Medline: 34554984]
47. Williams PAH, Woodward AJ. Cybersecurity vulnerabilities in medical devices: a complex environment and multifaceted problem. Med Devices (Auckl) 2015;8:305-316 [FREE Full text] [doi: 10.2147/MDER.S50048] [Medline: 26229513]
48. Kelly JT, Campbell KL, Gong E, Scuffham P. The Internet of Things: impact and implications for health care delivery. J Med Internet Res 2020;22(11):e20135 [FREE Full text] [doi: 10.2196/20135] [Medline: 33170132]
49. Djenna A, Harous S, Saidouni DE. Internet of Things meet internet of threats: new concern cyber security issues of critical cyber infrastructure. Appl Sci 2021;11(10):4580 [FREE Full text] [doi: 10.3390/app11104580]
50. Roman R, Zhou J, Lopez J. On the features and challenges of security and privacy in distributed internet of things. Comput Netw 2013;57(10):2266-2279. [doi: 10.1016/j.comnet.2012.12.018]
51. Adler-Milstein J, Jha AK. HITECH act drove large gains in hospital electronic health record adoption. Health Aff (Millwood) 2017;36(8):1416-1422 [FREE Full text] [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]
52. Pathak N, Misra S, Mukherjee A, Kumar N. HeDI: healthcare device interoperability for IoT-based e-Health platforms. IEEE Internet Things J 2021;8(23):16845-16852. [doi: 10.1109/jiot.2021.3052066]
53. Finet P, Gibaud B, Dameron O, Le Bouquin Jeannès R. Interoperable infrastructure and implementation of a health data model for remote monitoring of chronic diseases with comorbidities. IRBM 2018;39(3):151-159. [doi: 10.1016/j.irbm.2018.03.003]
54. United States Core Data for Interoperability (USCDI). Office of the National Coordinator for Health Information Technology. 2022. URL: https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi [accessed 2024-04-19]
55. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071 [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
56. Fernández-Alemán JL, Señor IC, Lozoya P, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
57. Azodo I, Williams R, Sheikh A, Cresswell K. Opportunities and challenges surrounding the use of data from wearable sensor devices in health care: qualitative interview study. J Med Internet Res 2020;22(10):e19542 [FREE Full text] [doi: 10.2196/19542] [Medline: 33090107]
58. Lowery C. What is digital health and what do i need to know about it? Obstet Gynecol Clin North Am 2020;47(2):215-225. [doi: 10.1016/j.ogc.2020.02.011] [Medline: 32451013]

XSL•FO
RenderX

59. El-Rashidy N, El-Sappagh S, Islam SMR, El-Bakry HM, Abdelrazek S. Mobile health in remote patient monitoring for chronic diseases: principles, trends, and challenges. Diagnostics (Basel) 2021;11(4):607 [FREE Full text] [doi: 10.3390/diagnostics11040607] [Medline: 33805471]
60. Zhou N, Wang L, Almirall D. Estimating tree-based dynamic treatment regimes using observational data with restricted treatment sequences. Biometrics 2023;79(3):2260-2271. [doi: 10.1111/biom.13754] [Medline: 36063542]
61. Norton EH, Fleming SM, Daw ND, Landy MS. Suboptimal criterion learning in static and dynamic environments. PLoS Comput Biol 2017;13(1):e1005304 [FREE Full text] [doi: 10.1371/journal.pcbi.1005304] [Medline: 28046006]
62. Kumari J, Kumar E, Kumar D. A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. Arch Comput Methods Eng 2023;30:3673-3701 [FREE Full text] [doi: 10.1007/s11831-023-09915-y] [Medline: 37359744]
63. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
64. Gomez-Garcia CA, Askar-Rodriguez M, Velasco-Medina J. Platform for healthcare promotion and cardiovascular disease prevention. IEEE J Biomed Health Inform 2021;25(7):2758-2767. [doi: 10.1109/JBHI.2021.3051967] [Medline: 33449888]
65. Marimuthu P, Perumal V, Vijayakumar V. Intelligent personalized abnormality detection for remote health monitoring. Int J Intell Inf Technol 2020;16(2):87-109. [doi: 10.4018/ijiit.2020040105]
66. Ukil A, Jara AJ, Marin L. Data-driven automated cardiac health management with robust edge analytics and de-risking. Sensors (Basel) 2019;19(12):2733 [FREE Full text] [doi: 10.3390/s19122733] [Medline: 31216659]
67. Trohman RG, Huang HD, Larsen T, Krishnan K, Sharma PS. Sensors for rate-adaptive pacing: how they work, strengths, and limitations. J Cardiovasc Electrophysiol 2020;31(11):3009-3027. [doi: 10.1111/jce.14733] [Medline: 32877004]
68. Sloan FA, Bethel MA, Ruiz D, Shea AM, Feinglos MN. The growing burden of diabetes mellitus in the US elderly population. Arch Intern Med 2008;168(2):192-199; discussion 199 [FREE Full text] [doi: 10.1001/archinternmed.2007.35] [Medline: 18227367]
69. Mancia G, Grassi G. The autonomic nervous system and hypertension. Circ Res 2014;114(11):1804-1814 [FREE Full text] [doi: 10.1161/CIRCRESAHA.114.302524] [Medline: 24855203]
70. Shimada SL, Brandt CA, Feng H, McInnes DK, Rao SR, Rothendler JA, et al. Personal health record reach in the veterans health administration: a cross-sectional analysis. J Med Internet Res 2014;16(12):e272 [FREE Full text] [doi: 10.2196/jmir.3751] [Medline: 25498515]
71. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
72. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019;21(7):e13659 [FREE Full text] [doi: 10.2196/13659] [Medline: 31293245]
73. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866-872 [FREE Full text] [doi: 10.7326/M18-1990] [Medline: 30508424]
74. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019;9(2):010318 [FREE Full text] [doi: 10.7189/jogh.09.020318] [Medline: 31788229]
75. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
76. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
77. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
78. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight: reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383(9):874-882 [FREE Full text] [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
79. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
80. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med 2018;283(6):516-529 [FREE Full text] [doi: 10.1111/joim.12752] [Medline: 29505159]
81. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
82. Menachemi N, Rahurkar S, Harle CA, Vest JR. The benefits of health information exchange: an updated systematic review. J Am Med Inform Assoc 2018;25(9):1259-1265 [FREE Full text] [doi: 10.1093/jamia/ocy035] [Medline: 29718258]
83. Osheroff JA, Teich J, Levick D, Saldana L, Velasco F, Sittig D, et al. Improving Outcomes with Clinical Decision Support: An Implementer's Guide, Second Edition. New York: HIMSS Publishing; 2012.

XSL•FO
RenderX

84. Weenk M, van Goor H, Frietman B, Engelen LJLPG, van Laarhoven CJHM, Smit J, et al. Continuous monitoring of vital signs using wearable devices on the general ward: pilot study. JMIR Mhealth Uhealth 2017;5(7):e91 [FREE Full text] [doi: 10.2196/mhealth.7208] [Medline: 28679490]
85. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. AACN Adv Crit Care 2013;24(4):378-386; quiz 387-388. [doi: 10.1097/NCI.0b013e3182a903f9] [Medline: 24153215]
86. van der Sijs H, Baboe I, Phansalkar S. Human factors considerations for contraindication alerts. Stud Health Technol Inform 2013;192:132-136. [Medline: 23920530]
87. van der Sijs H, van Gelder T, Vulto A, Berg M, Aarts J. Understanding handling of drug safety alerts: a simulation study. Int J Med Inform 2010;79(5):361-369. [doi: 10.1016/j.ijmedinf.2010.01.008] [Medline: 20171929]
88. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
89. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
90. Miller K, Mosby D, Capan M, Kowalski R, Ratwani R, Noaiseh Y, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. J Am Med Inform Assoc 2018;25(5):585-592 [FREE Full text] [doi: 10.1093/jamia/ocx118] [Medline: 29126196]
91. Rand V, Coleman C, Park R, Karar A, Khairat S. Towards understanding the impact of EHR-related information overload on provider cognition. Stud Health Technol Inform 2018;251:277-280. [Medline: 29968657]
92. Baker J, Jones D, Burkman J. Using visual representations of data to enhance sensemaking in data exploration tasks. J Assoc Inf Syst 2009;10(7):533-559 [FREE Full text] [doi: 10.17705/1jais.00204]

Abbreviations

CDS: clinical decision support
IT: information technology
RPM: remote patient monitoring

Edited by T de Azevedo Cardoso; submitted 26.07.23; peer-reviewed by M Baucum, H Ewald, E Vashishtha, R Williams, A Georgiou, R Bidkar; comments to author 24.08.23; revised version received 12.10.23; accepted 09.04.24; published 30.05.24.

Please cite as:
Claggett J, Petter S, Joshi A, Ponzio T, Kirkendall E
An Infrastructure Framework for Remote Patient Monitoring Interventions and Research
J Med Internet Res 2024;26:e51234
URL: https://www.jmir.org/2024/1/e51234
doi:10.2196/51234
PMID:38815263

©Jennifer Claggett, Stacie Petter, Amol Joshi, Todd Ponzio, Eric Kirkendall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Harnessing ChatGPT for Thematic Analysis: Are We Ready?

V Vien Lee[1*], PhD; Stephanie C C van der Lubbe[1*], PhD; Lay Hoon Goh[1,2], MBBS, MMed, PhD; Jose Maria Valderas[1,2,3], MD, PhD

[1]Division of Family Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2]Department of Family Medicine, National University Health System, Singapore, Singapore
[3]Centre for Research in Health Systems Performance, National University of Singapore, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Jose Maria Valderas, MD, PhD
Division of Family Medicine
Yong Loo Lin School of Medicine
National University of Singapore
NUHS Tower Block, Level 9, 1E Kent Ridge Road
Singapore, 119228
Singapore
Phone: 65 67723874
Email: jmvalderas@nus.edu.sg

Abstract

ChatGPT (OpenAI) is an advanced natural language processing tool with growing applications across various disciplines in medical research. Thematic analysis, a qualitative research method to identify and interpret patterns in data, is one application that stands to benefit from this technology. This viewpoint explores the use of ChatGPT in three core phases of thematic analysis within a medical context: (1) direct coding of transcripts, (2) generating themes from a predefined list of codes, and (3) preprocessing quotes for manuscript inclusion. Additionally, we explore the potential of ChatGPT to generate interview transcripts, which may be used for training purposes. We assess the strengths and limitations of using ChatGPT in these roles, highlighting areas where human intervention remains necessary. Overall, we argue that ChatGPT can function as a valuable tool during analysis, enhancing the efficiency of the thematic analysis and offering additional insights into the qualitative data. While ChatGPT may not adequately capture the full context of each participant, it can serve as an additional member of the analysis team, contributing to researcher triangulation through knowledge building and sensemaking.

(J Med Internet Res 2024;26:e54974) doi:10.2196/54974

KEYWORDS

ChatGPT; thematic analysis; natural language processing; NLP; medical research; qualitative research; qualitative data; technology; viewpoint; efficiency

Introduction

Background

Thematic analysis is a method to analyze qualitative data, commonly obtained through semistructured interviews or focus groups, with the aim of identifying and interpreting patterns of meaning or themes within the data [1]. As a method, thematic analysis is inherently flexible and dependent on the researcher's underlying philosophical assumptions [2]. For instance, positivist approaches may place greater emphasis on coding reliability, while interpretivist approaches may place more significance on reflexivity and the researcher's role (including subjectivity) in knowledge production [2]. Accordingly, thematic analysis may be well-suited to meet varying research needs and requirements [3]. While there are multiple methods for thematic analysis, Braun and Clarke's [1] 6 phases of thematic analysis is one of the most widely used approaches (Figure 1). The phases in Figure 1 were first defined in 2006, and the reflexive nature of their approach was further clarified in their 2019 publication [1,2].

XSL•FO
RenderX

Figure 1. Braun and Clarke's 6 phases of thematic analysis.

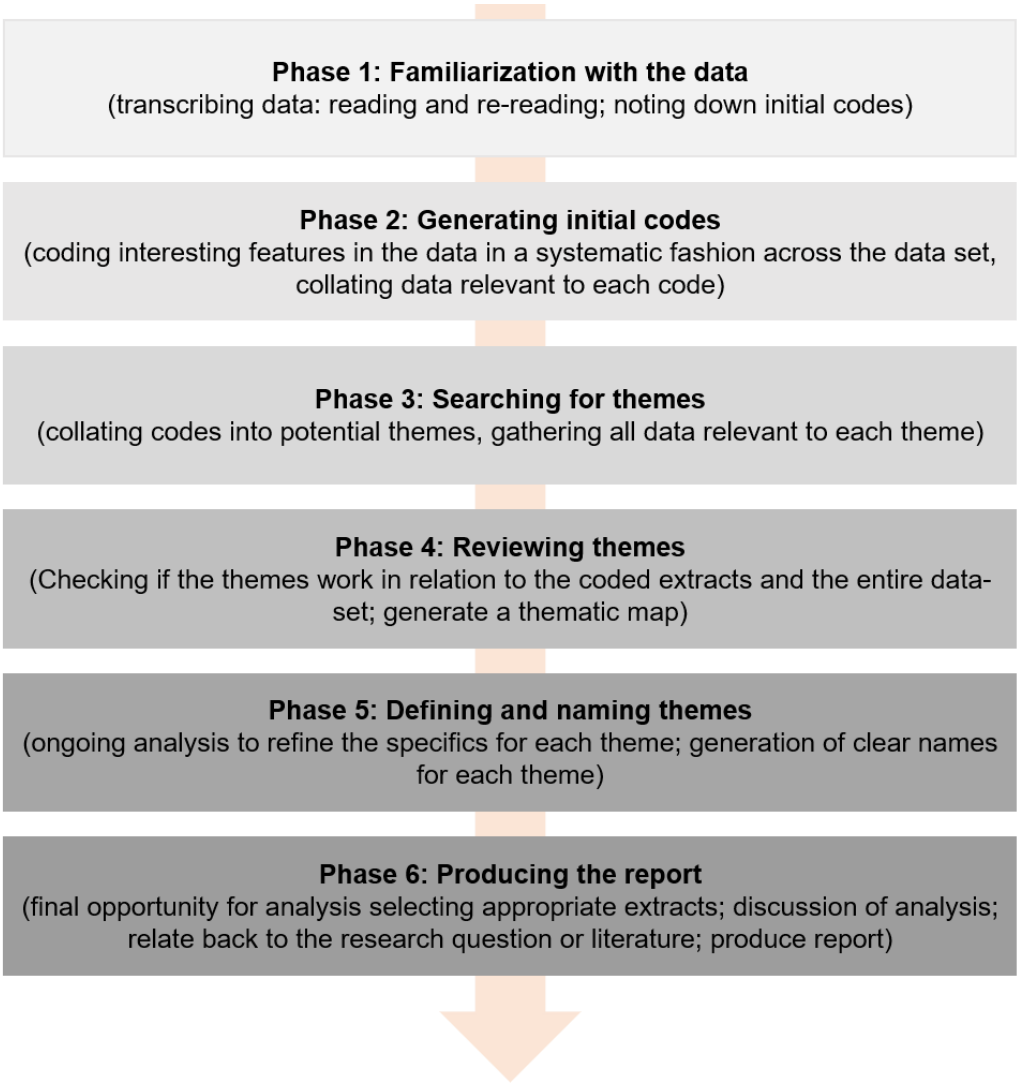

Given the flexibility of thematic analysis, there is room for creativity when engaging with the data and exploring tools that may aid the researcher's analytic process. With the increasing adoption of natural language processing (NLP) in health care research, such as diagnostic evaluation of electronic health records and the prediction of clinical outcomes based on consultation notes [4-6], researchers have begun to explore if there is space for artificial intelligence (AI) within the domain of qualitative research. To date, several AI-based tools, such as AILYZE (James Goh) and MonkeyLearn (Raul Garreta), are available to aid researchers in conducting thematic analysis [7,8]. For instance, AILYZE is able to summarize interview transcripts, provide suggestions for themes, and extract relevant quotes for each theme [7]. Nevertheless, full access to these tools often requires subscription payments, making them inaccessible for researchers and institutions that lack adequate financial resources.

In November 2022, OpenAI released version 3.5 of ChatGPT, a large language model-based chatbot capable of performing a wide range of text-based tasks based on context and past conversations (eg, summarizing research articles, answering domain-specific questions, and generating outlines for manuscripts) [9]. ChatGPT-3.5 is the chatbot adaptation of GPT-3.5 and is specifically optimized for back-and-forth conversations, though GPT-3.5 and ChatGPT-3.5 share the same foundational model. ChatGPT-3.5 is able to process a request and provide a response within a combined limit of 4096 tokens (ie, textual units, equivalent to approximately 3000 words in English), typically within a few seconds and free of charge [9,10].

Due to its electronic accessibility and free-to-use model, there has been a proliferation of discussion in the scientific community about incorporating GPT and ChatGPT into various aspects of research, including literature reviewing, data processing, and manuscript writing [11,12]. Qualitative studies have begun exploring the use of GPT and ChatGPT for conducting various aspects and types of qualitative analysis, from transcription cleaning to theme generation through thematic analysis [13-19]. Table 1 summarizes these studies by describing how GPT and ChatGPT were used in the analysis process, the main findings, and the challenges faced during the process. While all these studies are in preprint form and some are awaiting formal peer review, they provide an early glimpse into the feasibility of harnessing ChatGPT as an assistive tool when conducting qualitative analysis.

XSL•FO
RenderX

Table 1. Summary of studies that used ChatGPT or GPT for different phases of thematic analysis.

Reference	Country	GPT model	Use of ChatGPT or GPT	Findings	Challenges with ChatGPT or GPT
De Paoli (2023) [13]	United Kingdom	GPT-3.5 Turbo	The study used GPT to conduct thematic analysis, including generating initial codes, searching for themes, reviewing themes, and defining and naming themes.	GPT-3.5 Turbo was able to provide themes with synthetic descriptions. However, some inferred themes were not considered relevant by human researchers, and ChatGPT missed out on themes that were reported by human researchers.	• Interviews had to be divided into chunks due to the token limit. • Output is prompt-dependent (eg, asking for a different number of themes produced a different set of themes). • Hallucination (eg, assigned incorrect code to theme).
De Paoli (2023, preprint) [14]	United Kingdom	GPT-3.5 Turbo	The study used GPT to build user personas (ie, fictional yet realistic descriptions of a typical or target user of a product [20]) based on interview transcripts using thematic analysis.	GPT-3.5 Turbo was able to generate 2 relevant personas based on challenges and needs identified during thematic analysis.	• Biased toward creating specific types of user personas (eg, mostly middle-aged, from Italy) • Required human intervention to refine codes and themes generated (eg, generated code with a truncated quote)
Gao et al (2023, preprint) [15]	Singapore and United States of America	GPT-3.5	The study explored the functionality of CollabCoder (a data management prototype incorporating GPT-3.5) in assisting with open coding, iterative discussions, and the development of codebooks.	Participants valued GPT-3.5 for reducing cognitive burden during coding, but some participants cited that summaries generated by GPT-3.5 are too detailed and not relevant.	• Does not consider research questions or the intended direction of analysis if not explicitly instructed.
Mesec (2023, preprint) [16]	Slovenia	Information not available	The study used ChatGPT to conduct qualitative analysis using the grounded theory method.	ChatGPT was able to summarize and balance opposing ideas but tended to express ideas using descriptive terms at a lower level of abstraction compared to human researcher.	• Hallucination (eg, made-up information in a summary of texts) • Codes inadequately capture the content of the transcript • Unproductive repetition of output • Inappropriate use of terms (eg, “we can form some qualitative analyses”)
Tabone and de Winter (2023) [17]	Netherlands	GPT-3.5 Turbo and GPT-4-0613	The study used GPT to (1) conduct sentiment analysis, (2) provide meta-summaries of interviews, and (3) identify differences between 2 think-aloud transcripts.	Ratings (r=.98) and summaries generated with GPT-3.5 were strongly correlated or generally in line with those generated by human researchers. GPT-3.5 was also able to summarize the descriptive differences between 2 transcripts.	• Prompt-dependent (eg, modified prompt increased correlation of ratings) • Summary by GPT-4.0 was richer and touched on more facets than GPT-3.5 Turbo; however, some topics that were identified did not emerge in content analysis conducted by humans.
Taylor (2024) [18]	United States of America	Information not available	The study used ChatGPT to clean interview transcripts after using an artificial intelligence-assisted method to transcribe interviews.	ChatGPT cleaned redundant words and sentence fragments, but transcripts were more difficult to read due to ChatGPT connecting sentence fragments, which resulted in longer words per sentence.	• Word and syntax errors remained in several transcriptions. • Quality of transcription cleaning is dependent on the speech of the speaker (eg, clarity, pauses, and filler words). • Limited input word count (500-600 words as of March 2023)
Xiao et al (2023, preprint) [19]	Canada and France	GPT-3	The study used GPT to conduct deductive coding using an expert-developed codebook.	GPT-3 achieved fair to substantial agreement with human researchers (Cohen =0.38-0.61).	• The model occasionally produced incorrect labels.

Whereas these previous papers have focused on the broader use of ChatGPT in thematic analysis, its integration into medical research has yet to be investigated. From this viewpoint, we therefore explore the use of ChatGPT for thematic analysis in the medical domain while addressing the unique challenges that arise within a medical context. We begin by assessing the use of ChatGPT for generating codes based on an interview transcript, followed by extracting themes from a list of generated codes. Subsequently, we use ChatGPT for tidying quotes for manuscript preparation. Finally, we use ChatGPT to generate interview transcripts, which may be used for various academic

XSL•FO
RenderX

and educational purposes. For each application, we identify the areas where human intervention may still be required.

Using ChatGPT in Thematic Analysis

Given ChatGPT's ability to handle large textual data and provide sets of meaningful codes and themes, as demonstrated by De Paoli [13], ChatGPT has the potential to improve the efficiency of the thematic analysis process. As thematic analysis is typically conducted in a cyclical and iterative manner (eg, data collection and data analysis should occur concurrently, with insights from the analysis informing subsequent rounds of data collection and vice versa) [21], being able to digest and process large amounts of information efficiently (eg, by requesting a summary of an interview or generating an initial set of codes to help breakdown a transcript) can be helpful to researchers streamlining this cyclical process.

We used ChatGPT-3.5, which is free of charge, to illustrate the various ways ChatGPT can be used for thematic analysis. We have also used ChatGPT-4.0, which currently requires a fee, to see whether the results would improve when using a newer version of ChatGPT. Analyses were done on a transcript of the first episode of "Diabetes Discussion: A Diabetes UK Podcast," in which 2 guests share their experiences of living with diabetes [22].

Coding the Transcript

Our starting point is to investigate the capability of ChatGPT to code transcripts directly. To this end, we made the following request in ChatGPT:

> *The following is a transcript of an interview for a scientific paper focused on experiences of living with diabetes. Label the text by codes as is done in thematic analysis. Give the codes in the following format:*
>
> ***CODE***
>
> *- First words of the sentence(s) that was/were labeled with this code*
>
> *[transcript]*

A section of the coded transcript by ChatGPT-3.5 is displayed in Figure 2. ChatGPT-3.5 successfully identified multiple codes in the transcript within a single answer, and the corresponding codes match the textual content. However, as the transcript progresses (Multimedia Appendix 1), the coding by ChatGPT-3.5 becomes less detailed. While the subsequent answer contains multiple topics, ChatGPT-3.5 only assigned a single code, resulting in a loss of information in the coded output. Additionally, the second interview question (Multimedia Appendix 1) is also coded, which is not a common practice in thematic analysis.

Figure 2. Transcript coded by ChatGPT-3.5. A larger portion of the coded transcript is given in Multimedia Appendix 1.

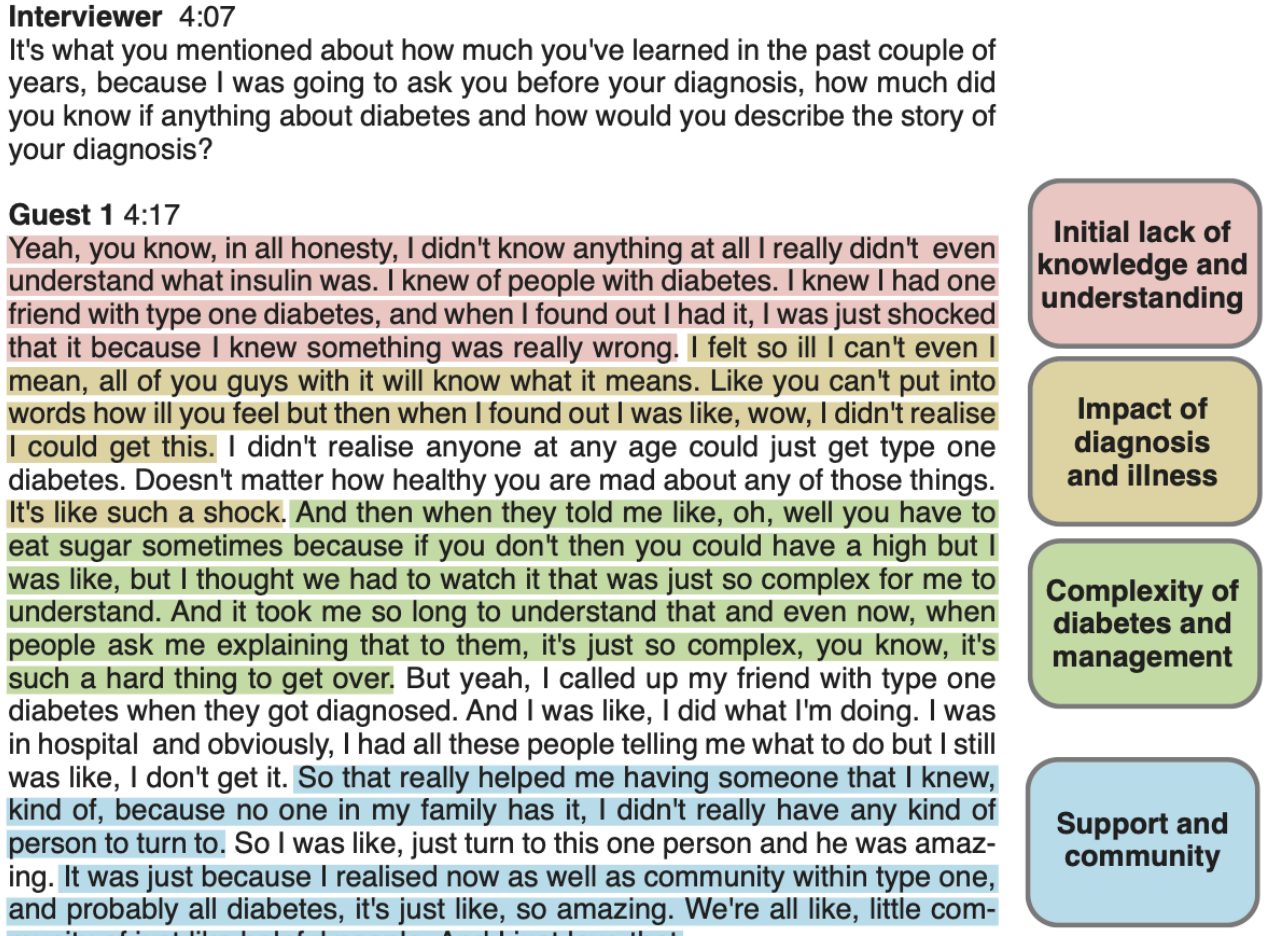

XSL•FO
RenderX

Using ChatGPT-4.0 for the same analysis results in a noticeable improvement in the output (Figure S2 in Multimedia Appendix 1). Not only does ChatGPT-4.0 ignore the transcript of the sections corresponding to the interviewer, but it also captures more details of the transcript. ChatGPT-4.0 assigned 5 different codes for the second question (as opposed to one assigned by ChatGPT-3.5), resulting in a set of codes that give a more complete picture of the transcript.

ChatGPT demonstrates a promising ability to code transcripts, but its performance depends on the GPT model used. Consequently, it is still necessary for a human researcher to review the codes generated and ensure that the codes appropriately capture all essential data. Furthermore, while the codes generated by ChatGPT sufficiently describe important concepts within the transcript, initial codes generated by human researchers during data analysis may continue to evolve to become more specific or reworded to better capture the data as new transcripts are analyzed. Given ChatGPT's limited context window, ChatGPT may not remember the full text of an interview transcript or the codes it previously generated, resulting in a loss of contextual understanding. For these reasons, a human researcher will be required to consolidate the codes generated to ensure that the codes adequately capture concepts or patterns of interest within the context of the whole data set.

Extracting Themes From Codes

Another way that ChatGPT may be used is to extract the themes and subthemes from the generated codes. These codes may be obtained either from ChatGPT or a human analyzer. We used the following request:

The following codes were obtained via coding of a transcript. Please identify the overarching themes and subthemes as is done in thematic analysis. These themes should have as little overlap as possible, and will be used in a scientific paper focused on experiences of living with diabetes. Use the following format:

THEME

**Subtheme*:*

- codes that belong to this subtheme

**Subtheme*:*

- codes

THEME

Subtheme

...

[list of codes]

The resulting themes are shown in Figure 3; the subthemes are tabulated in Table S1 in Multimedia Appendix 1. The themes and subthemes were derived from 81 codes obtained through the coding of the Spotify (Daniel Ek and Martin Lorentzon) transcript. Both ChatGPT and the human analyzer identified 5 unique themes, though their analyses had notable differences. For example, ChatGPT identified the "diet and nutrition management" theme, a subject that the human analyst neither classified as a theme nor a subtheme (Table S1 in Multimedia Appendix 1).

Figure 3. Themes identified by ChatGPT-3.5 (left) and the human analyzer (right) from 81 codes.

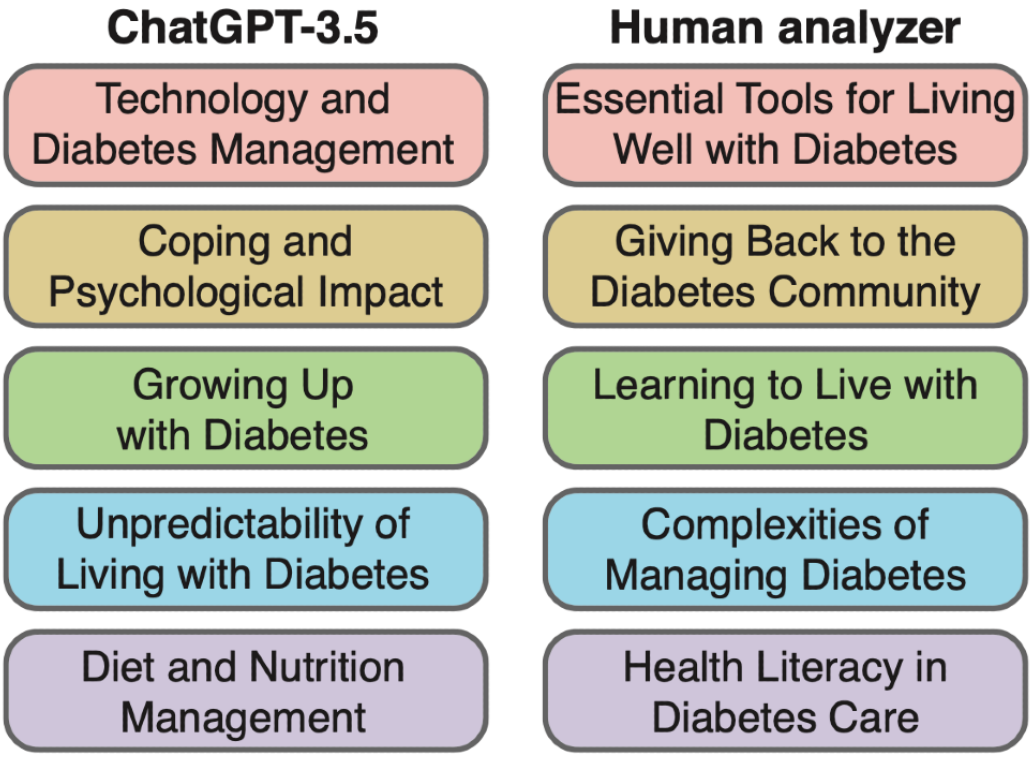

We conducted several rounds of analysis with ChatGPT-3.5 by resubmitting the same prompt, which led to interesting variations in the identified themes. For example, in the second analysis round, ChatGPT identified the "pregnancy and diabetes" theme (Table S1 in Multimedia Appendix 1), whereas pregnancy only emerged as a subtheme in ChatGPT's first analysis round. The

XSL•FO
RenderX

human analyzer, in contrast, did not identify pregnancy as a theme or a subtheme in their analysis.

It is impossible to state which of these themes more accurately reflects the essence of the interview, as there is no absolute truth in thematic analysis. Instead, we see the identification of different themes as an advantage. After all, a greater diversity in themes indicates that the codes were interpreted from different angles, thereby adding more layers to the overall analysis. In this framework, ChatGPT should be viewed as an additional team member when doing thematic analysis by offering fresh perspectives and proposing alternative interpretations of the identified codes.

Similar to the coding process, all themes and subthemes generated by ChatGPT should still be reviewed by human researchers to ensure that the themes and subthemes generated are aligned with the research question or questions and that essential data have been appropriately captured by the themes.

Tidying Up Quotes

While quotes in manuscripts are presented verbatim as much as possible, qualitative researchers often face the need to balance succinctness due to word count limits while maintaining the faithfulness of the quote to the participant's intended meaning. Examples of this include using ellipses to indicate that a portion of the quote (irrelevant to the point being made) has been cut, correction of words to maintain the grammatical integrity of the sentence, or removal of filler words such as "ah" or "um" [23]. We explored the potential use of ChatGPT to tidy up quotes from the interview transcript for manuscript preparation using the following request:

> *Clean the following transcript so it may be used as a direct quote for an academic paper:*
>
> *[quote]*
>
> *Omit all text that is not essential for the main message. Any altered or inserted words must be shown between square brackets, "[]", and omitted text must be replaced by three dots, "…".*

As shown in Figure 4, ChatGPT-3.5 struggled to tidy up the quote for manuscript writing purposes. A substantial portion of the text underwent revision without proper use of square brackets, despite our explicit request to denote any modifications or insertions. Additionally, sections of the text were interpreted by ChatGPT-3.5, which deviated from our initial instructions. In contrast, ChatGPT-4.0 performed the task more effectively (Figure 4). The generated quote contains only one word that should have been added between square brackets, while the remainder of the quote is indeed correctly extracted from the original transcript. Nevertheless, depending on how the quote is used to describe the data, one may argue that too much text has been omitted from the original transcript, filtering out the more personal viewpoints or emotional content from the final quote. Accordingly, quotes tidied up by ChatGPT should be reviewed by human researchers to determine if the quotes provide sufficient information and context to further elaborate on the theme or subtheme discussed. As tidying up quotes during manuscript preparation is less time-consuming than coding and theme generation, researchers should consider if the steps needed to generate a tidy quote outweigh the manual process.

Figure 4. Direct quotes generated by ChatGPT-3.5 (left) and ChatGPT-4.0 (right). The yellow highlights indicate the text that has been included in the quote, whereas the red highlights indicate the text that has been rephrased by ChatGPT without proper use of square brackets.

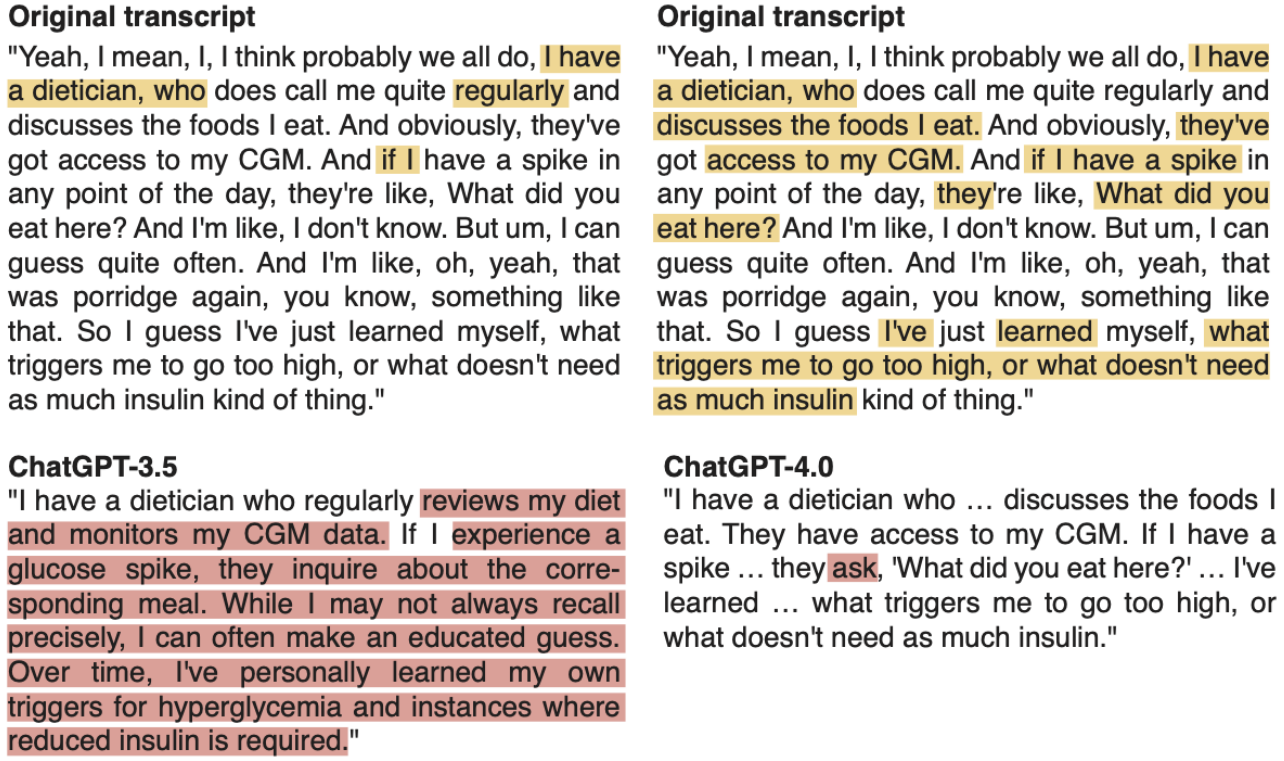

XSL•FO
RenderX

Creating Transcripts

Finally, we have explored the potential of ChatGPT to generate transcripts of interviews. To this end, we made the following request to ChatGPT:

> *Generate a transcript of an interview for a scientific study with a diabetes patient about their experience with living with diabetes. Required length: ~1500 words. Use the following format:*
>
> *Interviewer mm:ss*
>
> *[Question 1]*
>
> *Patient 1 mm:ss*
>
> *[Answer]*
>
> *Include stop words, pauses, and words such as "I think..", "uhh", and "I mean...", so it looks like an authentic transcript.*

This request was made without any additional instructions, such as an example transcript or a list of themes or codes. A snapshot of the generated transcript is shown in Figure 5. The resulting transcript matches the requested format and uses the stopwords that were asked for, making it similar to an actual conversation between 2 people. However, some differences exist between the generated transcripts and those from real-life interviews. The ChatGPT-generated transcript has a very direct question-answer structure, with all answers being on-point and of similar length. In contrast, a real-life interview is often more organic. For example, the interviewee may not understand the question and give answers of various lengths, and the interviewer may go in more depth before moving on to the next question.

Figure 5. Snapshot of the transcript generated by ChatGPT-3.5. The full transcript is given in Multimedia Appendix 1.

Interviewer: 00:30
[...] Can you share what your initial thoughts and feelings were when you received the diagnosis?

Patient 1: 00:36
Oh, absolutely. Well, I remember feeling, you know, a mix of emotions. I mean, first there was this sense of disbelief, like, "Is this really happening to me?" And, uh, I guess there was some fear too, you know, about what it would mean for my future. I think I was also a bit confused about how it all worked, you know, managing blood sugar levels and all that.

Interviewer: 00:54
It's completely understandable to feel that way. Managing diabetes can be complex. How has your daily routine changed since your diagnosis?

Patient 1: 01:02
Oh, my daily routine has changed quite a bit, actually. I mean, now I have to, you know, check my blood sugar levels regularly, and, umm, I have to be more mindful of what I eat. So, I've had to make some adjustments to my diet, like cutting back on sugary stuff and, uhh, watching my portion sizes. And, uh, I've also started incorporating, you know, regular exercise into my routine. It's been a bit challenging to juggle everything, but I think I'm getting the hang of it.

Despite these differences, the ChatGPT-generated transcripts have multiple advantages. By slightly adjusting the ChatGPT request, the generated transcripts may focus on any topic of interest or feature different respondent characteristics. This adaptability of ChatGPT allows for the quick generation of a large and diverse collection of transcripts. Furthermore, generating transcripts with ChatGPT eliminates any privacy and confidentiality concerns, which is a common issue with actual interviews.

Given these advantages, we envision several potential applications for the ChatGPT-generated transcripts. First, these transcripts may be used as instruction material for students learning thematic analysis. A second approach worth investigating is to use the generated transcripts as a training set for NLP models, particularly in topic modeling. As real-life transcripts are often limited or hard to get, ChatGPT offers a practical way to expand the data set, thereby exposing the model to a larger diversity of text. Further research into the feasibility of using generated transcripts as training data for NLP models is essential, taking into account the quality, diversity, and representativeness of the generated transcripts and the potential influence of the training set on the model's performance. It will be crucial to maintain a strict separation between training (ie, generated) data and real transcripts to ensure that any insights are obtained exclusively from real transcripts.

In short, while ChatGPT-generated transcripts cannot (yet) replicate the nuanced complexities of real interactions, they may be a promising source of training data for various academic and educational applications. Unfortunately, these generated transcripts may also open the opportunity for data falsification and be claimed as collected research data, especially since real data may be costly and time-consuming to obtain. To preserve the integrity of research, it is essential to ensure complete transparency regarding the origins of transcripts and their specific use within the analytical framework.

XSL•FO
RenderX

Challenges When Using ChatGPT for Thematic Analysis

While ChatGPT is a promising assistive tool for thematic analysis, previous studies have identified challenges when working with ChatGPT (Table 1) [13-19]. Major challenges relevant to thematic analysis include hallucination (ie, responses produced by the system that are not justified by the data used), the output being prompt-dependent (eg, prompts requesting the same output but phrased differently will lead to different outputs), and missing themes or codes previously reported by researchers [13]. Similarly, we encountered several challenges when using ChatGPT to conduct thematic analysis. When working with patient data, the primary concern is data confidentiality. Inputs to ChatGPT may be used as training data to improve their services, and network activities may be shared with third parties [24]. For this reason, uploading sensitive information, such as patient interview transcripts, to ChatGPT should be avoided. This precaution restricts the use of ChatGPT for coding (ie, phase 2 in Figure 1) unless the transcript holds no confidential information.

A more practical challenge is that ChatGPT has a word limit for each prompt, which may prevent users from inputting full transcripts and very long lists of codes or quotes. One potential solution is to split the input. However, as discussed above, ChatGPT has a limited context window, so it may forget the earlier parts of the interview transcript and/or codes it previously generated. As a result, ChatGPT may not be able to adequately capture patterns of ideas or concepts within the context of the whole data set. When coding, researchers consider existing knowledge (eg, the research question or current information about the topic), knowledge obtained through data collection (eg, interviews and field notes), and existing codes from previously analyzed transcripts. Without further information beyond the input, ChatGPT may adopt a narrower lens and generate results that are highly specific to a singular transcript. Accordingly, at this point in time, it is still essential for human researchers to collate the codes generated for each transcript and review them within the context of the study.

In the context of understanding text, ChatGPT, though advanced, may not capture every nuance that a human analyst would pick up [25]. In certain instances, ChatGPT may overlook underlying emotions or implicit themes that would otherwise be evident to human analysts. It is thus important to review the output of ChatGPT to ensure that all essential aspects of the transcript are captured in its thematic analysis. Without taking into account the context of participants (eg, gender, age, and culture), there is concern that the experiences shared by participants may be diluted when analyzed by ChatGPT. As such, output from ChatGPT should be recognized as complementary rather than a replacement for analysis conducted by human researchers. Beyond that, we also found that ChatGPT sometimes excludes existing codes or introduces new codes when generating themes. Hence, it is advisable to double-check whether all codes have been correctly assigned to the identified themes and subthemes.

Finally, ChatGPT may give different answers to the same questions, leading to nonreproducible results. Yet, in the context of thematic analysis, we do not see this variability as a drawback, as humans would also generate different results when doing thematic analysis. Instead, the different responses from ChatGPT may be seen as an opportunity because they may provide new insights that were not captured during the first round of thematic analysis.

Conclusions and Recommendations

ChatGPT has the potential to enrich the field of qualitative research. In our investigation, ChatGPT demonstrated its ability to code interview transcripts, generate themes from a list of codes, tidy quotes for manuscript preparation, and generate unique transcripts for education and training purposes. Nevertheless, limitations such as the inability to manage multiple transcripts and not fully capturing nuanced data essential to the research question necessitate the involvement of human researchers to collate and review the output generated. At this stage, ChatGPT requires human-AI collaboration, where researchers have to remain in the loop to intervene when necessary [13,15]. We present Figure 6 to show the opportunities available for ChatGPT to assist in thematic analysis and areas where human involvement is still required.

XSL•FO
RenderX

Figure 6. Opportunities for the usage of ChatGPT during the thematic analysis process and where human intervention is needed.

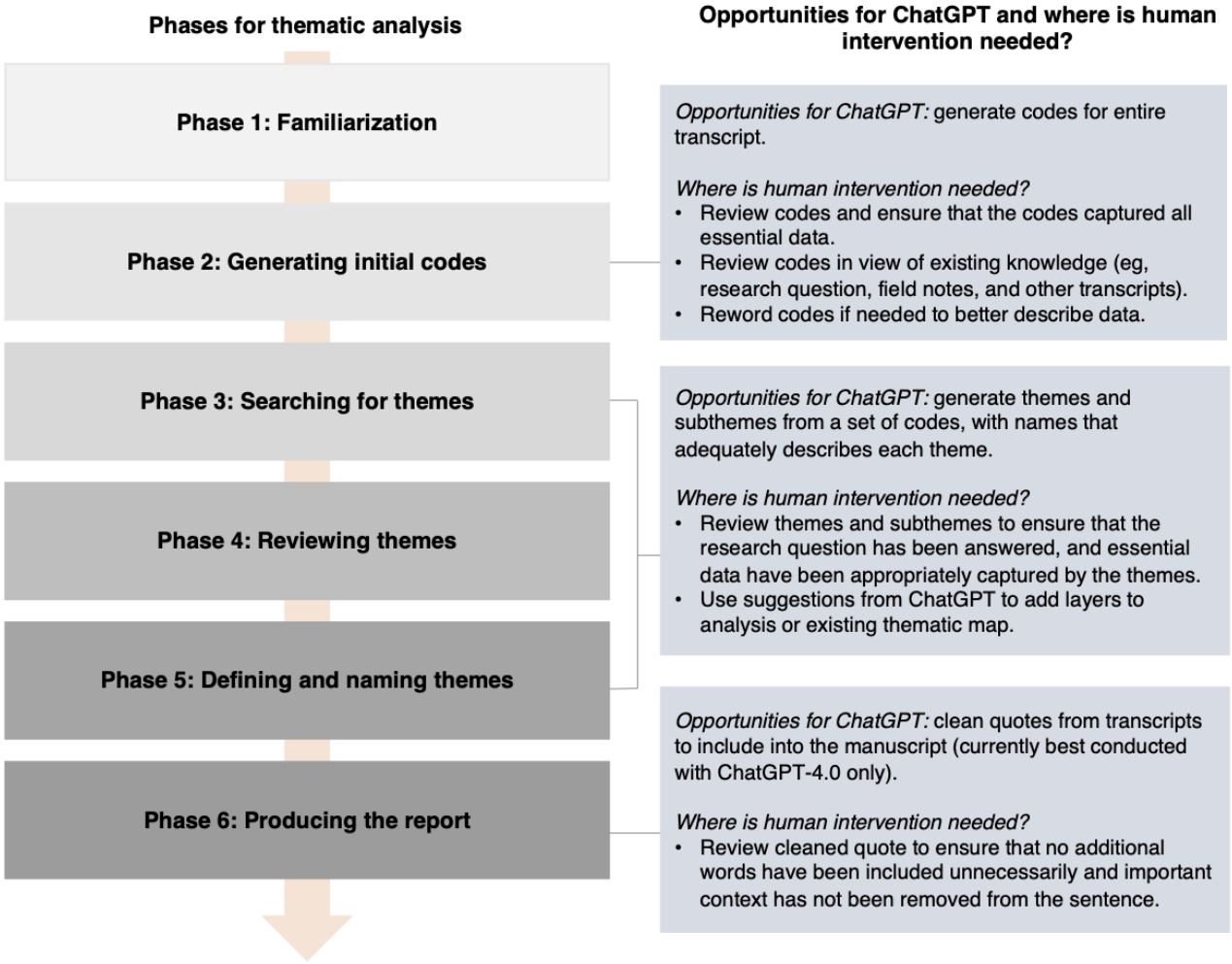

Given the need for considerable interaction between ChatGPT and human researchers, it will be more valuable to recognize ChatGPT as an additional member of the analysis team, contributing to researcher triangulation by adding to knowledge building and sensemaking, rather than a replacement for human researchers. However, with the ongoing progress in the field of natural language processing, the role of ChatGPT in qualitative research will evolve rapidly. This fast-paced development, in combination with the growing use of ChatGPT in research, necessitates further discussions regarding the use of ChatGPT in qualitative research. For example, how should ChatGPT's contribution be acknowledged, and what are the best practices regarding prompt formulation? Another important point of consideration is the confidentiality of the data, especially when working with patient data such as interview transcripts. The recent ChatGPT data breach in March 2023 should encourage researchers to remain mindful of the implications when working with AI tools that store data, regardless of purpose [26].

In summary, ChatGPT has the potential to function as a valuable tool during analysis, enhancing the efficiency of the thematic analysis and offering additional insights into the qualitative data. While this viewpoint remains an exercise to investigate the potential feasibility of using ChatGPT for thematic analysis, findings from the investigation can serve as a starting point for future studies that intend to further push the boundaries of AI involvement within qualitative research.

Authors' Contributions

VVL, SCCvdL, LHG, and JMV conceptualized and outlined the investigation. VVL and SCCvdL wrote the original draft, and all authors reviewed and edited the manuscript. LHG and JMV supervised the investigation. VVL and SCCvdL are considered cofirst authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Transcripts and table.
[DOCX File , 2247 KB - jmir_v26i1e54974_app1.docx]

References

1. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
2. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
3. Braun V, Clarke V. What can "thematic analysis" offer health and wellbeing researchers? Int J Qual Stud Health Well-being 2014;9:26152 [FREE Full text] [doi: 10.3402/qhw.v9.26152] [Medline: 25326092]
4. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019;25(3):433-438 [FREE Full text] [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
5. Dormosh N, Schut MC, Heymans MW, Maarsingh O, Bouman J, van der Velde N, et al. Predicting future falls in older people using natural language processing of general practitioners' clinical notes. Age Ageing 2023;52(4):afad046 [FREE Full text] [doi: 10.1093/ageing/afad046] [Medline: 37014000]
6. Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med 2016;69:53-61 [FREE Full text] [doi: 10.1016/j.artmed.2016.03.003] [Medline: 27085847]
7. AILYZE. 2023. URL: https://www.ailyze.com/ [accessed 2024-04-02]
8. MonkeyLearn. 2023. URL: https://monkeylearn.com/ [accessed 2024-04-02]
9. ChatGPT. OpenAI. 2023. URL: https://chat.openai.com/ [accessed 2024-04-02]
10. What are tokens and how to count them? OpenAI. 2023. URL: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them [accessed 2024-04-02]
11. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
12. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
13. De Paoli S. Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach. Social Science Computer Review 2023 [FREE Full text] [doi: 10.1177/08944393231220483]
14. De Paoli S. Writing user personas with large language models: testing phase 6 of a thematic analysis of semi-structured interviews. arXiv Preprint posted online on May 19, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.18099]
15. Gao J, Guo Y, Lim G, Zhang T, Zhang Z, Li TJJ, et al. CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. arXiv Preprint posted online on January 22, 2024 [FREE Full text] [doi: 10.1145/3584931.3607500]
16. Mesec B. The language model of artificial inteligence chatGPT: a tool of qualitative analysis of texts. Authorea 2023 [FREE Full text] [doi: 10.22541/au.168182047.70243364/v1]
17. Tabone W, de Winter J. Using ChatGPT for human-computer interaction research: a primer. R Soc Open Sci 2023;10(9):231053 [FREE Full text] [doi: 10.1098/rsos.231053] [Medline: 37711151]
18. Taylor Z. Using Chat GPT to clean interview transcriptions: a usability and feasibility analysis. Am J Qual Res 2024;8(2):153-160. [doi: 10.29333/ajqr/14487]
19. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. arXiv Preprint posted online on April 17, 2023 [FREE Full text] [doi: 10.1145/3581754.3584136]
20. Harley A. Personas make users memorable for product team members. Nielsen Norman Group. 2015. URL: https://www.nngroup.com/articles/persona/ [accessed 2024-04-02]
21. Fossey E, Harvey C, McDermott F, Davidson L. Understanding and evaluating qualitative research. Aust N Z J Psychiatry 2002;36(6):717-732. [doi: 10.1046/j.1440-1614.2002.01100.x] [Medline: 12406114]
22. Diabetes discussions: living with diabetes. Diabetes UK. 2023. URL: https://www.diabetes.org.uk/guide-to-diabetes/podcasts/living-with-diabetes [accessed 2023-06-21]
23. Lingard L. Beyond the default colon: effective use of quotes in qualitative research. Perspect Med Educ 2019;8(6):360-364 [FREE Full text] [doi: 10.1007/s40037-019-00550-7] [Medline: 31758490]
24. Privacy policy. OpenAI. 2023. URL: https://openai.com/policies/privacy-policy [accessed 2024-04-02]
25. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Physical Syst 2023;3:121-154 [FREE Full text] [doi: 10.1016/j.iotcps.2023.04.003]
26. March 20 ChatGPT outage: here's what happened. OpenAI. 2023. URL: https://openai.com/index/march-20-chatgpt-outage [accessed 2023-04-02]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
NLP: natural language processing

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 29.11.23; peer-reviewed by J Parsons, C Turvey; comments to author 09.02.24; revised version received 28.02.24; accepted 20.03.24; published 31.05.24.

Please cite as:
Lee VV, van der Lubbe SCC, Goh LH, Valderas JM
Harnessing ChatGPT for Thematic Analysis: Are We Ready?
J Med Internet Res 2024;26:e54974
URL: https://www.jmir.org/2024/1/e54974
doi:10.2196/54974
PMID:38819896

©V Vien Lee, Stephanie C C van der Lubbe, Lay Hoon Goh, Jose Maria Valderas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A Taxonomy for Health Information Systems

Anna Janssen[1], BA, MA, PhD; Candice Donnelly[1], BSc, MPH; Tim Shaw[1], BSc, PhD

Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

Corresponding Author:
Anna Janssen, BA, MA, PhD
Faculty of Medicine and Health
The University of Sydney
Level 2, Charles Perkins Centre D17
Sydney, 2066
Australia
Phone: 61 02 9036 9406
Email: anna.janssen@sydney.edu.au

Abstract

The health sector is highly digitized, which is enabling the collection of vast quantities of electronic data about health and well-being. These data are collected by a diverse array of information and communication technologies, including systems used by health care organizations, consumer and community sources such as information collected on the web, and passively collected data from technologies such as wearables and devices. Understanding the breadth of IT that collect these data and how it can be actioned is a challenge for the significant portion of the digital health workforce that interact with health data as part of their duties but are not for informatics experts. This viewpoint aims to present a taxonomy categorizing common information and communication technologies that collect electronic data. An initial classification of key information systems collecting electronic health data was undertaken via a rapid review of the literature. Subsequently, a purposeful search of the scholarly and gray literature was undertaken to extract key information about the systems within each category to generate definitions of the systems and describe the strengths and limitations of these systems.

(J Med Internet Res 2024;26:e47682) doi:10.2196/47682

KEYWORDS

eHealth; digital health; electronic health data; data revolution; actionable data; mobile phone

Background

The collection and use of electronic health data (EHD) is common in contemporary society due to the high level of digitization. As the amount of electronic data continually increases, many sectors refer to the phenomenon as a data revolution [1]. The health sector mirrors this trend and is increasingly digitized, particularly in response to the COVID-19 pandemic [2,3]. Information and communication technologies (ICTs) that are enabling EHD collection include, but are not limited to, electronic health records (EHRs), patient administrative systems, web-based information from social media and other sources, data collected within apps, and data from wearable devices [4].

The widespread use of these ICTs enables the development of a rich health data ecosystem, which has great potential to enhance existing approaches to patient self-management, service delivery, and new care models. Different types of EHD and their potential have been investigated in several contexts. This includes how EHD is used to improve health and well-being [5,6], manage chronic conditions [7,8], identify and respond to public health concerns [9], and support quality improvement activities [10-12].

Despite the large amount of EHD collected, it continues to be underused by many health care organizations and providers [13]. There is a gap in understanding what constitutes actionable data. Actionable data have been described as data that enable the user to make an informed decision to solve a specific problem [14,15]. In the health sector, actionable data have been described as data that can improve the quality, outcomes, or cost of care [16]. Another challenge in using EHD is the increasingly large amount of data held by private industry, as the organizations that develop the technology often retain custody of its data [17,18].

To transform EHD into actionable data, unique challenges need to be overcome. One of the most widely acknowledged challenges is adequate infrastructure for EHD use. Infrastructure challenges include (1) the inability to access qualified and experienced technical experts for extracting and analyzing data and (2) deeply ingrained interoperability issues between existing

XSL•FO
RenderX

data sources [19]. A further challenge is that EHD is incredibly complicated and fragmented within individual departments and health care organizations [19]. Data are often not centralized within organizations, and a cultural change is required [20,21]. Extracting data from these silos brings new challenges related to privacy and the safety of data transmission within a connected health system [16]. Finally, many health care organizations only have small data, which have significant potential for real-time analytics and presentation in digestible formats, that is, dashboards [13]. However, much of the literature has focused on big data use [22,23], with limited research or understanding of the value of small health data sets for generating meaningful insights.

There are many benefits to increasing the actionability of health data. It can provide a foundation for learning health systems. Such systems transform routinely collected EHD into useful information to improve health care quality and outcomes and support timely decision-making [24]. Another benefit is the increased accessibility and value of EHD for health professionals and organizations. It is particularly important that health professionals see value from their data input. Since the implementation of EHRs, health professionals have been increasingly required to take an active role in data entry. Such data entry has been repeatedly noted as adding a significant workload burden [25]. Although there is a workload burden, it is also postulated that health professional–collected data will be of higher quality than other sources [26]. Demonstrating the benefits of EHD to health professionals will also likely increase their buy-in to data collection, which is essential in improving the completeness and quality of EHD. Expanding the utility of EHD would also benefit consumers in supporting new care models such as virtual care [27]. Furthermore, actionable and transparent EHD can support personalized and patient-centered care [28].

This paper aims to present a classification of health information systems used in health care and describe how the information collected by these systems contributes to the health data ecosystem. Furthermore, the paper aims to better explore what constitutes actionable data in health care and contextualize the role of different data sources for this purpose. A key contribution of this study is the development of a health information system taxonomy to help classify the breadth of data collected about health (Table 1) and a visualization of how the different ICTs that collect EHD can be categorized (Figure 1).

This paper addresses a gap in the current literature by providing an overarching description of the health data ecosystem. Recent research has focused on describing certain types of EHD, that is, imaging, administrative, and genomic data [29], or classifying subcategories of health information systems, particularly clinical information systems [30,31].

XSL•FO
RenderX

Table 1. Overview of the different health information systems, the data they collect, and the examples of how the data have been actioned.

Health information system	Description of the system	Actionability of data
Clinical data sources		
EHRs[a] and EMRs[b]	Repositories of patient health information created by health professionals	• Reducing errors as a single source of truth for health information • Improving governance, organizational processes, or health service delivery • Supporting research, quality improvement, and reflective practice • Providing data for clinical decision support systems and AI[c] technologies
Registries	Information systems that collect uniform data to evaluate outcomes for a specific population, disease, condition, or exposure	• Supporting disease management • Real-time source of information for monitoring disease in the community • Understanding population health trends • Supporting research, quality improvement, and reflective practice
Practice management software	Software designed to manage the everyday activities of medical practices	• Reducing errors as a single source of truth for health information • Improving workflow efficiency • Underpinning clinical decision support systems
Consumer complaints and incident reports	Repositories of data patients report about problems they experience when interacting with a health care organization	• Supporting research, quality improvement, and reflective practice • Improving governance, organizational processes, or health service delivery
Hospital administrative information systems	Platforms that collect data about administrative information and billing information within health care organizations	• Supporting research, quality improvement, and reflective practice • Monitoring adverse events • Improving efficiency of health service delivery
Patient-reported outcome measures	Questionnaires that measure patients' perceptions of a disease or its treatment on their health	• Identifying patient perspectives on issues important to them about their health and health care • Supporting research, quality improvement, and reflective practice • Improving governance and organizational processes
Diagnostic information systems	Information systems that collate and report on results from different diagnostic processes including blood tests, radiology, and imaging	• Supporting research, quality improvement, and reflective practice • Providing data for clinical decision support systems and AI technologies
Electronic prescribing systems	Digital technologies for managing collection, distribution and storage of scripts	• Can improve the efficiencies of prescribing processes
Remote monitoring platforms	Digital technologies that can passively monitor, assess, and potentially manage decisions about care	• Data are collected continuously about wearers' health • Lack of guidelines or tools to inform best practice use of data by health professionals
Bespoke databases	A variety of platforms exist, including repositories of information created by members of the health workforce to collect data that they prioritize and health information collected by medical devices within health care organizations	• Supporting research, quality improvement, and reflective practice • Adapting to the priorities and data collection needs of individual health professions, specialty groups, or organizations
Consumer and community data sources		
WPRs[d]	Information consumers publish on review websites on the internet, sharing reviews of clinical encounters with health professionals or health care organizations	• Identifying patient perspectives of issues important to them about their health and health care • Supporting information-seeking behaviors • Improving governance, organizational processes, or health service delivery

XSL•FO
RenderX

Health information system	Description of the system	Actionability of data
Appointment booking systems	Web-based applications that enable consumers to book appointments with health professionals and manage those bookings	• Reducing health care organization costs • Improving patient-centeredness in health care and health service delivery
Web-based communities	Information shared by individuals on the web via social media and virtual forums, or input into web-based search engines, can be used to understand community health and well-being	• Supporting health communication • Improving consumer knowledge about their health • Identifying patient perspectives on issues important to them about their health and health care
Web-based search engines	Data collected by interacting with web-based search engines.	• Understanding population health trends • A real-time source of information for monitoring disease in the community
Smartphone, web, and desktop apps	Self-contained programs that run on the internet, on smartphones, or on computer operating systems that are designed to improve health and well-being	• Improving the completeness and quality of EHD[e] collection • Support disease management
Technology-collected data sources		
Wearables and devices	Computer hardware that an individual wears as an accessory or by attaching to their clothing, which passively collects data about their activity	• Supporting disease management • Enabling new, more personalized models of care • Collecting information on individual health and well-being behaviors
Direct-to-consumer health care	A range of emerging products and services that provide technology at the first point of care and may use AI to triage patients	• A large amount of data are being collected by these systems • To date, access to these data have largely been controlled by product vendors

[a]EHR: electronic health record.

[b]EMR: electronic medical record.

[c]AI: artificial intelligence.

[d]WPR: web-based patient review.

[e]EHD: electronic health data.

XSL•FO
RenderX

Figure 1. A visualization of the different health information systems that collected electronic health data and how they can be broadly categorized.

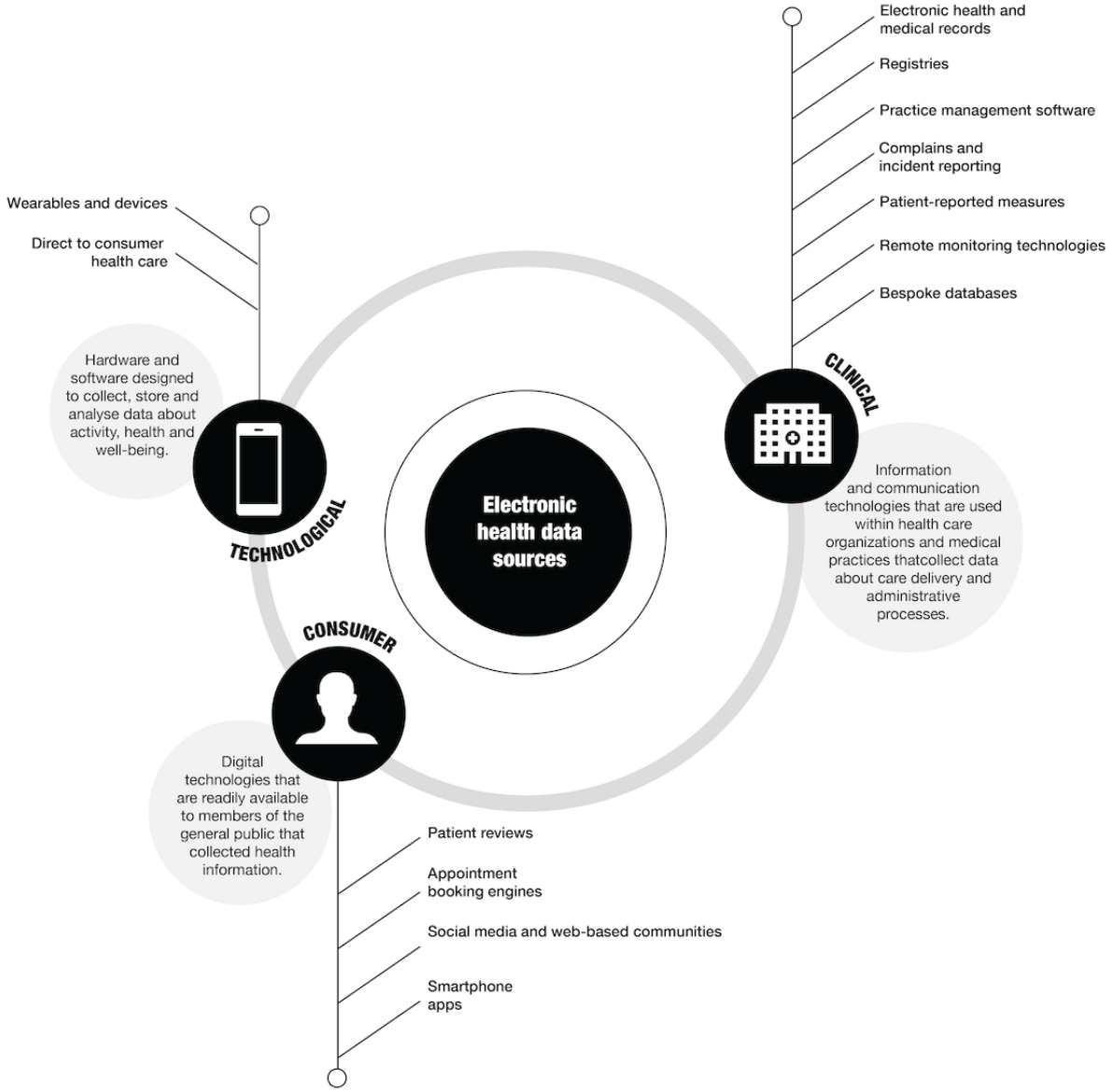

Developing the Taxonomy

The EHD ecosystem underpins much of digital health, but it is complex and difficult to navigate. To tackle this issue, the authors have created a comprehensive taxonomy that categorizes essential health information systems. This taxonomy serves as a valuable resource for the numerous individuals in the digital health workforce who may lack extensive informatics expertise but are bound to encounter such systems in their roles. These individuals in the digital health workforce may regularly *interact* with health data and use actionable insights from the data to inform aspects of their job, or they may be interacting with computer scientists, data analysts, informaticians, or other data analysts regularly [32]. The taxonomy is for these individuals who need a nontechnical overview of key systems and the reasons the systems were designed to understand the strengths and limitations of the data they collect.

A taxonomy is an approach for specifying the characteristics of objects and categorizing them to help understand complex domains [33]. A wide range of methodologies can be used for developing taxonomies, and the literature indicates that inductive, deductive, and intuitive approaches can all be used to good effect [34]. Methodologies for developing taxonomies vary across disciplines. The inherently multidisciplinary nature of digital health means there is not a single, widely recognized methodology for developing taxonomies in this space. The taxonomy *development* described in this paper was guided by a modified version of design science research and an improved method for taxonomy development [33]. This was done by applying the following steps: (1) problem definition, (2) definition of objectives for the taxonomy, (3) specifying end conditions, and (4) design and development.

The problem definition (step 1) and objectives for the taxonomy (step 2) are outlined in earlier sections of this paper. The end conditions for the taxonomy (step 3) were chosen to be the point of adequate conciseness [34], that is, the point at which the taxonomy had become meaningful without becoming unwieldy or overwhelming to end users. Finally, the design and development process for the taxonomy (step 4) was undertaken

XSL•FO
RenderX

via a rapid review of existing taxonomies or classifications of EHD systems. This review was used to create an initial categorization for the objects in the taxonomy, the key information systems for inclusion in the taxonomy, and the higher-level categorizations for these systems. This initial categorization was discussed by the researchers (AJ and CD) and augmented with missing systems. A purposeful search of the scholarly and gray literature was subsequently undertaken to extract key information about the EHD systems within each category, generate definitions, and describe the strengths and limitations of these systems.

A limitation of this taxonomy is that it did not undergo an evaluation process. However, it has been noted in the literature that this is a common limitation of taxonomies, which frequently go through iterative development but do not have a final evaluation stage [33].

Health Information Systems

Summary of EHD Sources

EHD can be broadly classified into 3 categories, described in depth in subsequent sections of this study. Refer to Table 1 for the taxonomy of key health information systems. Figure 1 visualizes the higher-level categorizations of these systems. The three categories are as follows:

1. Clinical data sources: digital technologies that are used by health care organizations to collect EHD. It includes technologies such as electronic health and medical records, registries, practice management software, patient-reported measures, and bespoke databases collecting information from various medical devices used in the clinical setting.
2. Consumer and community data sources: consumer technologies that collect data about the health and well-being of the individuals using the technology. It includes technologies such as desktop and smartphone apps, social media, and web-based communities.
3. Technology-collected data sources: systems that passively collect data about people with digital technology. It largely describes data collected by wearables and other devices.

Clinical Information Systems

Overview

There are a variety of definitions of clinical data. At the highest level, clinical data can be defined as data collected during the care delivery process of health care organizations [35]. Although clinical data mostly refers to data about patients, the data are used for a wide range of applications beyond the delivery of care, including billing [36] and research [37]. The systems that collect clinical data are typically those used by organizations to deliver health services or to monitor processes and outcomes of care. Data in this category are commonly entered by a member of the health workforce. However, some systems may also passively collect metrics on end-user interactions with the technology.

Electronic Health and Medical Records

The terms EHR and electronic medical record (EMR) are frequently used interchangeably, but there are subtle differences between the 2 technologies. EMRs are repositories of patient health information created by health professionals to capture data related to specific clinical encounters [38]. EHRs include EMR functions but also have additional ones. The crucial difference is that EHRs are designed to share information across different health care settings and potentially between health professionals and patients [39]. Typically, EHRs are designed to collect both structured and unstructured data. Data collected by EHRs can be beneficial in many contexts, including at the clinical, organizational, and societal levels [40]. Clinical benefits include a reduction in medical errors and improving the completeness and accuracy of data [41]. EHR and EMR data have also been useful for supporting quality improvement activities such as audits and feedback [42] and can potentially enable workplace learning and reflective practice [43].

Registries

Clinical registries are information systems that collect uniform data to evaluate specified outcomes for a population, disease, condition, or exposure [44]. Registries can be implemented for a range of purposes, including quality improvement [37], disease monitoring [10], device surveillance, and health care services monitoring [10]. Registries are generally designed to collect structured data. Health professionals usually input data retrospectively using strict definitions aligned with informatics infrastructure [45]. However, there is growing interest in the use of e-registries, which use ICTs to enable systematic, automated, and longitudinal collection, retrieval, and analysis of data [46].

Practice Management Software

Practice management software is designed to manage the everyday activities of a medical practice [47]. Historically, practice management software was largely focused on supporting billing [48]. Contemporary systems often incorporate a range of administrative functions such as patient scheduling and registrations, financial reporting, and collections management as well as elements such as patient records and potentially patient communications functionality [48]. Practice management software usually collects a combination of structured and unstructured data. Some practice management software has started incorporating more sophisticated tools for supporting clinical care, such as tracking vaccinations and recording test results.

Consumer Complaints and Incident Reports

Consumer complaints and incident data can include problems as diverse as organization parking issues to physical harms and sentinel events [49]. These data can be captured on a range of different platforms. As such, there is variation in the data collected across organizations, which can influence what data points are captured. Generally, the data collected by these systems are unstructured.

Hospital Administrative Information Systems

Health care organizations use a variety of platforms to collect data about administrative aspects of care and clinical billing [36]. One of the most common types are patient administration systems (PASs), which capture a range of data, including demographic information about patients and interactions

XSL•FO
RenderX

between the health care organization and the patient [50]. These administrative data systems indirectly capture information about care delivered, but it is not their primary purpose [51]. PAS data are often carefully structured, particularly data related to payments, as they may need to be provided by health care organizations with different PAS to health insurance companies with different processing software [52]. This is done by coding PAS and other data collected within a health care organization and providing the coded data to payers, commonly called claims data, for reimbursement purposes [53].

Patient-Reported Outcome Measures

Patient-reported outcome measures are data collected using questionnaires that measure patients' perceptions of disease management and outcomes [54]. They do not have to be collected electronically, but they often are. These data are particularly valuable for assessing the patient-centeredness of care and how effectively care responds to patient needs [55]. They can be reported as individual data or in an aggregated form [56]. These types of information systems are increasingly designed to collect structured data, but this can vary depending on the design of the individual data collection tool.

Diagnostic Information Systems

These systems are designed to collect data generated by diagnostic results, including imaging, radiology, and pathology test results, and to report genomic data to inform diagnosis and guide therapeutic approaches [57]. Data from these systems are often integrated into EHR, EMR, or PAS within health care organizations and practice management software in primary care in addition to being stored in independent information systems [21]. Radiology information systems and their equivalents collect reports on test results for use by the health workforce [58].

Electronic Prescribing

Electronic prescribing systems are used by pharmacies to generate digital prescriptions instead of paper-based documents [59]. These systems can enable the uploading and exchange of dispensed script data [59,60] and support monitoring the dispensing of controlled medicines to minimize misuse of pharmaceuticals [60].

Remote Monitoring

Remote monitoring describes a range of digital technologies that can passively monitor, assess, and potentially manage patient care regardless of whether they are with a health professional [61]. They include technologies such as cardiac devices implanted into a patient on the recommendation of a health professional, but also devices such as pulse oximeters that consumers can access regardless health professional input [62].

Bespoke Databases

In addition to more formalized repositories of clinical information, a portion of EHD is collected by bespoke databases. This can include databases of patient information set up by individual health professionals using free electronic capture tools such as REDCap (Research Electronic Data Capture; Vanderbilt University) [63] to collect data on care quality and outcomes [64,65]. It can also include data from specialized medical devices and hardware such as bioimpedance spectroscopy machines [66,67] and surgical robots [68,69]. The type of data collected by medical hardware and devices varies depending on the machine. These data frequently include metrics on how end users are interacting with the device. As the data collection tools are often custom built, they can also often be flexible and adaptable to the data collection needs of the organization or specialty area [70].

Consumer and Community Data Sources

Overview

The widespread access to digital technologies such as smartphones and internet access has created a plethora of potential sources of EHD generated by consumers and the public. Many of these data sources were not explicitly designed to collect data on the health and well-being of individuals, but researchers and other stakeholders have sought to understand their value. There is still a recognized challenge in harnessing these data, particularly integrating them with clinical data sources [71]. The data in this category are generally input into these ICTs by consumers or individuals in the community rather than the health workforce.

Smartphone, Web, and Desktop Applications

Smartphone, web, and desktop apps can be defined as self-contained programs optimized for these platforms. In health care, these applications are designed to improve health and well-being using different designs and functions [71]. The use of health applications is still mostly community and consumer driven, limiting the potential of these technologies to transform health care [71]. Due to the diversity of health conditions applications have been developed for, there is considerable breadth to the data they could collect. Data collection fits across four broad categories: (1) automatic collection using smartphone features such as GPS to track distance traveled and cameras to collect photo diaries; (2) linkage with Internet of Things devices that collect, record, and transmit data such as weight scales and blood pressure monitors; and (3) manual input of data by users, including recording calories, menstrual cycles, medication compliance, and blood glucose [71]. Despite the breadth of data that could be collected, researchers have noted that most users are using them to collect data on things such as fitness activities, daily activity levels, and tracking sleep [72].

Web-Based Patient Reviews

A diverse array of websites on the internet let consumers publish reviews and provide web-based ratings about clinical encounters with individual health professionals or health care organizations. Feedback provided via web-based patient reviews (WPRs) is unsolicited by the health professional. These data can be structured (ie, ratings) or unstructured (ie, free text comments) [73]. WPRs have been shown to influence health care consumers' health professional selection and decision-making [74]. In addition, some research suggests that specialist clinicians have used WPR data to improve patient communication and workflows [75,76].

XSL•FO
RenderX

Appointment Booking Systems

Appointment booking systems allow patients to book and manage their medical appointments on the web [77]. While traditionally these systems were developed and managed by individual health care practices, they are increasingly being developed and maintained by software companies. Health care practices often embed the booking application into their website, and structured data collected within the system are often stored on the cloud [28].

Web-Based Communities

These data describe various sources of information on health, including social media data and data from forums and other web-based communities. Web-based communities are peer-to-peer communities where people with common interests can gather virtually to share experiences, ask questions, and provide emotional support [78]. The data collected by web-based communities are typically unstructured and text based but can vary based on the platform. Social media describes internet-based applications with a social dimension, typically by enabling users to post content and interact with the content of others almost synchronously [79]. Data collected by social media platforms are varied but include the content in individual posts, metrics on engagement with the content, and information on the network individuals choose to follow [18]. Data generated from web-based sources possess distinct characteristics compared to other web-based information, such as patient reviews. This is because the main objective of the web-based platform is not specifically to gather health data. However, there are many web-based communities for discussion of health conditions, including diabetes management [80], cancer management [81], and general health and wellness [82].

Web-Based Search Engines

Web-based search engine data describe information on specific terms individuals input into search engines to find results relevant to their queries. Data collected by search engines are primarily terms users input into an individual search engine. A widely used source for such information is Google Trends, which provides a database of terms frequently searched on their platform. It allows for comparing individual terms across different regions and languages [83]. This is among the earliest web-based sources of health information, with initial research using this data to monitor public health trends dating as far back as 2008 [84].

Technological Data Sources

Overview

The high level of digitization in contemporary society means that individuals interact with a significant amount of ICT daily. Some of these ICTs are actively sought out by individuals, such as purchasing a smartphone or an activity tracker. Other ICTs in this category are interacted with less overtly, such as search engines. Typically, data in this category are passively collected by technology as end users interact with it rather than actively input by consumers or the health workforce.

Wearables and Devices

Wearable devices are digital technologies that an individual wears as an accessory and that may also use sensors to track health information [7]. The data they collect are typically structured. Although used to collect health data, wearable devices were not originally developed to support health care. They were developed by the fitness industry to track activity to support individual health and well-being [85]. Wearable devices are increasing in popularity in the community [86], translating into an increased interest in their use in health care.

Direct Digital Health Care

Technology is increasingly used to enable direct consumer health care models, which may entirely digitize care delivery. These systems include health check stations that enable individuals to monitor their health without interaction with a traditional health care organization [87]. In addition, there are systems that integrate artificial intelligence (AI) technologies with human expertise, serving as digital-first entry points to care. They offer AI-based triage and referral services to deliver preventative care [88-90]. Finally, large non–health organizations are also entering this space to offer end-to-end health and wellness management services, including home installation of apps and devices and dedicated teams of virtual clinicians and managers [88].

Challenges and Limitations of the Current Digital Health Information Systems

Quality, Completeness, and Interoperability

Ensuring data input into health information systems is complete, of high quality, and able to be linked and shared across different systems within and across organizations is a major challenge with the current digital health information systems. In practice, management software data completeness issues are common, particularly on older systems where administrative and patient records are not linked, potentially leading to loss of information when data are transferred between systems [48]. Similarly, PAS used in hospitals have been shown to be inaccurate for reporting care outcomes [51]. The data can have limitations due to varied coding systems across organizations, data quality and completeness, and other issues [52]. In EMRs and EHRs, interoperability can be a major challenge, particularly a lack of integration across care contexts such as primary and acute care [46]. Registries have also had issues with interoperability due to instances of variation in standards used to define common data elements across different registries [44].

Interoperability is also a major challenge for emerging consumer health information repositories such as smartphone apps. Many apps do not include functionality that allows users to easily export data from apps and share it with health professionals [5]. In addition, the breadth of data collected by these apps could prevent integration into clinical informatics systems [71].

Data quality issues are also a significant limitation of different repositories of web-based patient health information, though the quality issues are different from clinical information systems. WPR data may not be representative, are often skewed toward

XSL•FO
RenderX

a small number of health professionals, and therefore may not be sufficient for consumers to make an informed decision about a provider [91]. There is little research to show that these data can be used to measure clinical quality for individual health care providers, and the results of research in this area show mixed [91] association [73]. Coupled with this, health care organizations and health professionals have reservations about the validity of WPR data [74] and the ability of a single bad review to reach a wide audience [92]. These concerns may limit health professionals' uptake of these data to understand the quality of care.

Social media provides a rich source of real-time health data, although it also has many limitations. These include collecting a large amount of irrelevant data, challenges in standardizing and validating data, and potential biases in the demographics and geographic location of social media users and identifying these biases in the data [84,93]. A noted limitation of search engine data is its potential to dilute cultural differences that shape web-based search behaviors or completely exclude data sets from geographic regions where those search engines are unavailable [93]. As with smartphone apps, wearables have significant data quality issues, with noted inconsistencies in data accuracy [85,94]. Wearable devices also encounter quality and completeness issues, compounded by interfaces that may not effectively highlight these gaps [95], leaving users unaware when data are not being collected [96].

Data Security and Privacy

Data security and privacy are a major concern with contemporary information systems and are often top of mind for health care organizations. When implementing EMRs and EHRs, ensuring data stored within and shared across organizations are governed in a way that provides private and secure data exchange can be challenging [25]. In consumer-facing systems such as appointment booking engines, data have been misappropriated for financial gain [97,98], suggesting issues to be resolved about the privacy and security of these data and the social license for its secondary use.

Complexity of Implementation and Ongoing Maintenance

Implementing any new technology in existing workflows can be challenging, and digital health information systems are no exception. Resistance to changes in workflow can be a notable barrier to the adoption of clinical information systems [28,56]. In the context of certain types of data collection, such as patient-reported outcome measures, inadequate infrastructure to support the collection of these data [99] is a major challenge. System maintenance after implementation can also be a challenge for health information systems. Limitations associated with clinical registries include the expenses related to maintenance, which often require staff with specialized skills to operate [44], and the necessity for sustained funding to support the registry in the long term [10].

Another challenge of modern information systems such as wearable devices is a lack of guidelines or tools to inform the best practice use of remote monitoring devices in clinical care [61]. Currently, very few wearable devices are subject to regulatory standards that govern other medical equipment [86], making it challenging to know which wearables are safe to use in the context of health care. Integration of these devices and the data they produce into health care remains challenging [100], as they currently do not integrate with clinical workflows or other information systems such as EHRs [86]. This same challenge occurs with health data gathered by apps, as there are currently few pathways for use in the clinical setting by health professionals [101].

Impact on Workloads

One of the major limitations of the current digital health information systems is their impact on the workloads of health professionals. The digitization of health data collection parallels an increased administrative load for health professionals. This issue is particularly pronounced with EHRs, as health professionals typically enter data. This has been shown to increase the workloads of health professionals and contribute to burnout [25].

Secondary Use

While considerable time and effort is spent inputting data into contemporary digital health information systems and given the vast amount of data collected, extracting and accessing these data for secondary purposes is notably difficult. Even when data can be extracted, secondary use of data from health information systems has several limitations. In the context of EMRs and EHRs, it can be challenging to obtain longitudinal insights from data as information is captured at each time point a patient interacts with the health system, which for many individuals occurs as sporadic or occasional visits [102]. Similarly, data from consumer complaints and incident reporting have limitations because they capture individual incidents not holistic care [103]. This can make it challenging to identify a strategy for improving quality and can also be an issue for determining whether a complaint is attributed to an individual or a system issue.

Another challenge supporting the secondary use of many digital health information systems is scaffolding the information or supporting end users to interpret it in different contexts from those in which it was collected. For example, registries are often primarily designed for research purposes, and they are not intended to make data readily accessible for use by health services [37]. Patient-reported outcome measure data have a similar limitation. The data are not routinely used by many health care organizations [56]. Health professionals may not use patient-reported measures due to not feeling like they have the capacity to use them, not seeing their value [56], or a lack of knowledge regarding how to meaningfully interpret the data [99]. When using these data for quality improvement, barriers include a lack of timeliness of the data, limitations in determining the cause of a poor outcome, and implementing change based on a poor outcome [104]. It has also been noted that patient-reported outcome measures are only 1 piece of performance data that health professionals have access to, and the data may be ignored if individual providers do not see themselves as outliers in the data set [104].

XSL•FO
RenderX

Data Ownership and Consent

As health information systems increase the amount of EHD collected and move toward better interoperability, patient consent for secondary use of data becomes an increasing concern. In the context of health data, consumers freely share about themselves, such as that on social media and web-based communities, which is another challenge related to consent exists. Although individuals choose to share this content publicly, it remains unclear if users are comfortable having their data used for research and other secondary applications [84].

Another major challenge with contemporary digital health information systems is data ownership. While all the data collected belongs to patients, the data are housed in information systems that consumers do not have access to. The value of these data is being increasingly recognized, and this is acting as a barrier to access. For example, the device manufacturer often controls access to data from remote monitoring devices, which can raise complex legal and ethical concerns [62]. These technologies collect large amounts of data that may be vulnerable to privacy breaches or be coopted for commercial purposes [105], and custodianship of these data is widely contested [18]. Data collected by bespoke databases in devices has value as a commercial asset [17,18], and vendors may not want to make it available to end users. Furthermore, the data collection architecture in these systems is not governed by any standards, so linking and comparing data, even if it can be accessed, is difficult.

Actioning EHD in Practice

Overview

EHD has great potential to transform many aspects of the health sector. There is a growing body of research exploring how data can be used to predict patient outcomes [64,68,106], understand care quality [51,55,107], and personalize treatments [21]. Approaches such as those described in learning health system processes can increase the actionability of the EHD to ensure it improves care processes and outcomes [24]. There are also a growing number of ICTs designed to increase the actionability of EHD from different sources. Much research has been undertaken into visualizing EHD from clinical sources such as PAS and EMRs to uncover insights for end users of the tools [66,108] and how approaches such as learning health system processes can increase the actionability of the EHD. Electronic patient portals use digital technology, such as a website or a smartphone application, to provide patients access to personal health information [109]. These platforms are often linked to EHRs to increase accessibility to data, help involve patients and caregivers in clinical decision-making, and improve communication between patients and their health care team [110,111]. Similarly, initiatives such as OpenNotes, which makes clinical notes available to patients and caregivers, improve transparency and patient-centered care [112,113].

In the following subsections, 4 use cases describe how ICT supports existing care models and how the data collected might be harnessed to transform service provision, health, and well-being. Some of these use cases describe scenarios increasingly realized in the health sector, partly due to the rapid adoption of digital platforms in response to the COVID-19 pandemic. Other use cases are informed by the research but describe hypothetical scenarios requiring evidence to be translated into practice before they can be fully realized by the health sector. These scenarios are designed to illustrate and prompt reflection on how EHD can be harnessed to generate actionable insights and transform health care in the future. Figure 2 presents a visualization of the 4 different use cases.

XSL•FO
RenderX

Figure 2. A visualization of the 4 different use cases and a high-level summary of how electronic health data can underpin it in the future.

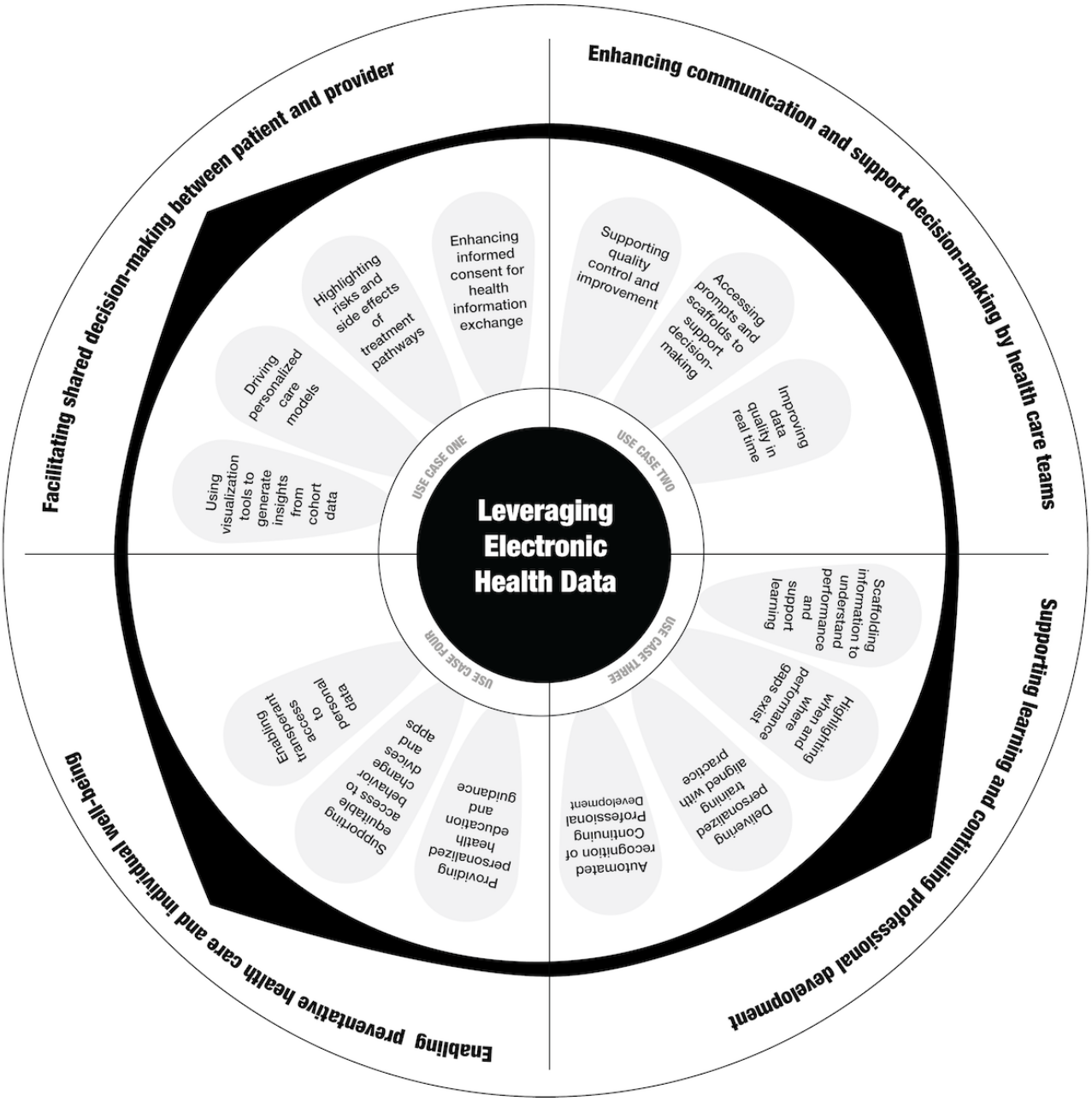

Use Case 1: Informing Decision-Making in Consultations

Digital technologies and the EHD they collect can be used to strengthen patient consultations with health professionals. The value of technology to support patient interactions both within and outside of the consultation is increasingly being recognized as virtual care. Virtual care describes non–face-to-face clinical care professionally enabled through digital mechanisms [27]. In the current clinical practice, health professionals have limited access to clinical decision support tools that harness EHD and support shared decision-making [114]. Genomics data are increasingly available to health professionals to support precision medicine [57] and are likely to be more prominent in future clinical decisions. EHD from patient applications and wearable devices can be challenging for health professionals to access in consultations. When data are available, there is limited guidance on how to incorporate these data [115,116]. In addition to data access challenges, there is some evidence that the use of digital technologies during the consultation may impact patient-provider communication and rapport [117,118].

In the future, access to EHD during consultations may enable shared decision-making between patients and health professionals. This could be achieved by visualizing cohort data within an organization to help patients understand their likely outcomes for a given procedure at that specific care center [66]. It can also support more personalized decision-making in the consultation, such as by harnessing visual interfaces presenting EHD to help patients assess the risk of undergoing a specific procedure [43], supporting more informed consent processes. Finally, patients could consent to sharing health application and wearable device data with their health professional so that the data can be reviewed and discussed in the consultation to

XSL•FO
RenderX

understand patient compliance with treatment plans and tailor care to align best with the patient's lifestyle. Alongside this, the increasing application of disruptive technologies such as AI to EHD could streamline administrative processes in health care. Although not directly beneficial in the consultation, the actionability of EHD for such applications should reduce the administrative burden on health professionals and allow more time for service delivery.

Use Case 2: Health Professional to Health Professional Interactions

Health care teams currently underuse EHD. While digital technologies such as videoconferencing have been successfully adopted as tools to support team meetings [119], the uptake of ICTs for visualizing data has been slower. There may be a need to modify physical meeting spaces to better incorporate technologies for presenting these data [120].

In the future, there will undoubtedly be greater access to EHD by individual health professionals and teams. Given the considerable time health professionals spend on data entry [25], there will likely be an increasing expectation that they will have access to these data in return for the workload burden. Access to technologies to visualize and scaffold EHD will provide health teams with a holistic view of each patient's care journey, transforming all aspects of team interactions, including clinical decision-making, quality improvement, research, and learning. Furthermore, ready access to data by health care teams will likely have an important role in ensuring quality control as health professionals incorporate visualization tools into team meetings to review processes and outcomes of care. As part of these team reviews or as individuals, health professionals will likely use digital technologies presenting these data to review the quality of information, annotate points of interest, and edit and improve data in real time. As the application of AI in health data becomes more reliable, there will also likely be an increase in prompts and scaffolds provided by advanced analytics tools that can support decision-making within health professional teams.

Use Case 3: Personalized Training and Reflective Practice

Health professionals regularly dedicate considerable time to engaging in education and training activities to stay up to date on the latest evidence [121,122]. Health professionals have limited access to EHD to reflect on their practice and inform decisions about professional development. Despite this, regulatory bodies expect that EHD will be used in training and professional development to maintain registration to practice [123,124]. Research suggests that health professionals would like greater access to electronic data for educational purposes [125].

In the future, health professionals, teams, and organizations will be able to leverage the plethora of EHD to understand care delivery and outcomes and enable personalized reflective practice and learning. A new range of educational technologies will be available to engage health professionals in their data to develop a rich understanding of how their behavior can change care quality. This learning approach could create authentic clinical narratives and link it with evidence-based approaches to technology enabled learning. This would focus efforts in health professional knowledge and skill development and enable accreditation bodies to recognize quality in learning experiences rather than quantity of time spent engaging in learning.

Use Case 4: People Using Their data for Preventative Care and Health and Wellness Management

Patients have limited access to data collected by health care organizations about their treatment pathway. While technologies such as patient portals [109] and open notes [113] are making these data more available to patients and their caregivers, uptake of these solutions is limited. However, there may be an increase in patient portal adoption as regulation requires health care providers to increase the accessibility of EHD for individual patients [57]. Coupled with this, vendors from outside of health care are increasingly entering the sector to provide direct-to-consumer care by harnessing AI to enable services to be delivered when and where the consumer desires [88,89].

The health care transparency movement has also explored the value of making cost and quality information about health services available to the public [126,127], but these data are not consistently published for all organizations. There is also a plethora of consumer-centric apps and wearable devices collecting extensive data on individual activities, both general and health-specific, that enable the self-management of health and wellness [6,71,72]. Some of these innovations are beginning to be used to improve patient self-efficacy in managing chronic conditions. The increasing availability of diverse data sets also opens up exciting research opportunities. It will undoubtedly unlock new knowledge about human health and support the discovery of more personalized and adaptive approaches to maintaining health and delivering care. In the future, EHD has the potential to enable personalized preventative health and well-being solutions for members of the public. Digital technologies will increasingly make health care available outside of clinical settings, at home, in the workplace, or in whatever location suits the patient.

Conclusions

The health sector is collecting an ever-increasing quantity of EHD. Navigating the ICTs in health care and the siloed data they collect is an ongoing challenge for stakeholders. A classification system for these data could enable stakeholders to get a high-level understanding of the complex health data ecosystem. EHD could be classified into three broad categories: (1) clinical data, (2) consumer and community data, and (3) technology-enabled data.

Understanding the complex health data ecosystem is essential if EHD is going to be leveraged to generate actionable insights that can be harnessed by ICTs to support new care models. In the last 2 years, health care has undergone considerable accelerated digital disruption as part of the COVID-19 pandemic response, which has expanded the types of EHD the sector collects. This includes a growing repository of EHD collected by new non–health corporations entering health care with direct-to-consumer products that support e-prescribing and other

XSL•FO
RenderX

synchronous and asynchronous communications, which collect large volumes of health data. While these services add a rich source of information to the EHD ecosystem, they also bring challenges, including increasing the amount of data in commercial vendors' custody.

Considerable focus to date has been placed on the value of aggregating large data sets into single repositories, which represents a significant infrastructure achievement.

However, moving forward, it is as important to understand *why* data are being collected as *how* they will be collected to ensure the correct information is available to benefit the health system and support public health and well-being.

Acknowledgments

AJ is undertaking a postdoctoral research fellowship that is funded through the Digital Health Cooperative Research Centre. The Digital Health Cooperative Research Centre is established and supported under the Australian Government's Cooperative Research Centres Program.

Conflicts of Interest

None declared.

References

1. Kitchin B. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Thousand Oaks, CA: SAGE Publications; 2014.
2. Bhambere S, Abhishek B, Sumit H. Rapid digitization of healthcare - a review of COVID-19 impact on our health systems. Int J All Res Educ Sci Method 2021;9(2):1457-1459.
3. Vargo D, Zhu L, Benwell B, Yan Z. Digital technology use during COVID-19 pandemic: a rapid review. Human Behav Emerg Technol 2020 Dec 28;3(1):13-24. [doi: 10.1002/hbe2.242]
4. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
5. McKay FH, Wright A, Shill J, Stephens H, Uccellini M. Using health and well-being apps for behavior change: a systematic search and rating of apps. JMIR Mhealth Uhealth 2019 Jul 04;7(7):e11926 [FREE Full text] [doi: 10.2196/11926] [Medline: 31274112]
6. DiFilippo KN, Huang WH, Andrade JE, Chapman-Novakofski KM. The use of mobile apps to improve nutrition outcomes: a systematic literature review. J Telemed Telecare 2015 Feb 12;21(5):243-253. [doi: 10.1177/1357633x15572203]
7. Kamei T, Kanamori T, Yamamoto Y, Edirippulige S. The use of wearable devices in chronic disease management to enhance adherence and improve telehealth outcomes: a systematic review and meta-analysis. J Telemed Telecare 2020 Aug 20;28(5):342-359. [doi: 10.1177/1357633x20937573]
8. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness: a qualitative systematic review. Qual Health Res 2017 Jan 12;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
9. Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O. A scoping review of the use of Twitter for public health research. Comput Biol Med 2020 Jul;122:103770 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103770] [Medline: 32502758]
10. Pop B, Fetica B, Blaga ML, Trifa AP, Achimas-Cadariu P, Vlad CI, et al. The role of medical registries, potential applications and limitations. Med Pharm Rep 2019 Jan 27;92(1):7-14 [FREE Full text] [doi: 10.15386/cjmed-1015] [Medline: 30957080]
11. A vision for primary health care in the 21st century: towards universal health coverage and the Sustainable Development Goals. World Health Organization and the United Nations Children's Fund (UNICEF). 2018. URL: https://www.who.int/docs/default-source/primary-health/vision.pdf [accessed 2024-04-24]
12. Berner ES, Detmer DE, Simborg D. Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. J Am Med Inform Assoc 2004 Oct 18;12(1):3-7. [doi: 10.1197/jamia.m1664]
13. Wills MJ. Decisions through data: analytics in healthcare. J Healthc Manag 2014;59(4):254-262. [doi: 10.1097/00115514-201407000-00005] [Medline: 25154123]
14. San SY, Chan T. Defining and conceptualizing actionable insight: a conceptual framework for decision-centric analytics. arXiv Preprint posted online June 11, 2016. [FREE Full text]
15. Cooper A. What is analytics? Definition and essential characteristics. Cetis Analytics Series. 2012 Nov. URL: https://www.solaresearch.org/core/what-is-analytics-definition-and-essential-characteristics/ [accessed 2024-04-24]
16. Makeham MA, Ryan A. Sharing information safely and securely: the foundation of a modern health care system. Med J Aust 2019 Apr 30;210 Suppl 6(S6):S3-S4. [doi: 10.5694/mja2.50038] [Medline: 30927463]
17. Graham M. Data for sale: trust, confidence and sharing health data with commercial companies. J Med Ethics 2023 Jul 30;49(7):515-522 [FREE Full text] [doi: 10.1136/medethics-2021-107464] [Medline: 34330796]

XSL•FO
RenderX

18. Hunter P. The big health data sale: as the trade of personal health and medical data expands, it becomes necessary to improve legal frameworks for protecting patient anonymity, handling consent and ensuring the quality of data. EMBO Rep 2016 Aug 11;17(8):1103-1105 [FREE Full text] [doi: 10.15252/embr.201642917] [Medline: 27402546]
19. Turning healthcare big data into actionable clinical intelligence. Health IT Analytics. 2017 May 24. URL: https://healthitanalytics.com/features/turning-healthcare-big-data-into-actionable-clinical-intelligence [accessed 2024-04-24]
20. Flying blind 2: Australian researchers and digital health. Digital Health CRC. URL: https://digitalhealthcrc.com/future-thinking/flyingblind-australian-researchers-and-digital-health-australian-health-data-series-volume-2/ [accessed 2024-04-24]
21. Srinivasan U, Rao S, Ramachandran D, Jonas D. Flying blind: Australian consumers and digital health, volume 1: Australian health data series. Capital Markets Cooperative Research Centre. 2016. URL: https://nla.gov.au/nla.obj-403180437/view [accessed 2024-04-24]
22. Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: accelerating value and innovation. McKinsey & Company. 2013 Apr 1. URL: https://www.mckinsey.com/industries/healthcare/our-insights/the-big-data-revolution-in-us-health-care [accessed 2024-04-24]
23. Groves P, Kayyali B, Knott D, Van Kuiken S. The 'big data' revolution in healthcare: accelerating value and innovation. McKinsey & Company. 2013 Jan. URL: https://tinyurl.com/3nenx2aa [accessed 2024-04-24]
24. Enticott JC, Melder A, Johnson A, Jones A, Shaw T, Keech W, et al. A learning health system framework to operationalize health data to improve quality care: an Australian perspective. Front Med (Lausanne) 2021 Oct 27;8:730021 [FREE Full text] [doi: 10.3389/fmed.2021.730021] [Medline: 34778291]
25. Gawande A. Why doctors hate their computers. The New Yorker. 2018 Nov 5. URL: https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers [accessed 2024-04-24]
26. Farrington C, Burt J, Boiko O, Campbell J, Roland M. Doctors' engagements with patient experience surveys in primary and secondary care: a qualitative study. Health Expect 2017 Jun 28;20(3):385-394 [FREE Full text] [doi: 10.1111/hex.12465] [Medline: 27124310]
27. Baxby L. Australia's health reimagined: the journey to a connected and confident consumer. Deloitte. 2022 Mar 1. URL: https://www.deloitte.com/au/en/Industries/life-sciences-health-care/perspectives/australias-health-reimagined.html [accessed 2024-04-24]
28. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017 Apr 26;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
29. Koopman B, Bradford D, Hansen D. Exemplars of artificial intelligence and machine learning in healthcare: improving the safety, quality, efficiency and accessibility of Australia's healthcare system. Commonwealth Scientific and Industrial Research Organisation. 2020. URL: https://aehrc.csiro.au/wp-content/uploads/2021/10/Exemplars-AI-in-Health-July-2020.pdf [accessed 2024-04-24]
30. National digital health workforce and education roadmap. Australian Digital Health Agency. 2020. URL: https://www.digitalhealth.gov.au/sites/default/files/2020-11/Workforce_and_Education-Roadmap.pdf [accessed 2024-04-24]
31. Vincent A, Kaelber DC, Pan E, Shah S, Johnston D, Middleton B. A patient-centric taxonomy for personal health records (PHRs). AMIA Annu Symp Proc 2008 Nov 06;2008:763-767 [FREE Full text] [Medline: 18998912]
32. Brunner M, McGregor D, Keep M, Janssen A, Spallek H, Quinn D, et al. An eHealth capabilities framework for graduates and health professionals: mixed-methods study. J Med Internet Res 2018 May 15;20(5):e10229 [FREE Full text] [doi: 10.2196/10229] [Medline: 29764794]
33. Omair B, Alturki A. An improved method for taxonomy development in information systems. Int J Adv Comput Sci Appl 2020;11(4). [doi: 10.14569/IJACSA.2020.0110470]
34. Nickerson RC, Varshney U, Muntermann J. A method for taxonomy development and its application in information systems. Eur J Inf Syst 2017 Dec 19;22(3):336-359. [doi: 10.1057/ejis.2012.26]
35. Grossmann C, McGinnis JM, Olsen L, Goolsby WA. Clinical data as the basic staple of health learning: creating and protecting a public good: workshop summary. In: Institute of Medicine Roundtable on Value and Science-Driven Health Care. Washington, DC: National Academies Press; 2011.
36. Lawrence J, Delaney CP. Integrating hospital administrative data to improve health care efficiency and outcomes: "the socrates story". Clin Colon Rectal Surg 2013 Mar 4;26(1):56-62 [FREE Full text] [doi: 10.1055/s-0033-1333662] [Medline: 24436649]
37. Hoque DM, Kumari V, Hoque M, Ruseckaite R, Romero L, Evans SM. Impact of clinical registries on quality of patient care and clinical outcomes: a systematic review. PLoS One 2017 Sep 8;12(9):e0183667 [FREE Full text] [doi: 10.1371/journal.pone.0183667] [Medline: 28886607]
38. Alpert J. The electronic medical record in 2016: advantages and disadvantages. Digit Med 2016;2(2):48. [doi: 10.4103/2226-8561.189504]
39. McMullen PC, Howie WO, Philipsen N, Bryant VC, Setlow PD, Calhoun M, et al. Electronic medical records and electronic health records: overview for nurse practitioners. J Nurs Pract 2014 Oct;10(9):660-665. [doi: 10.1016/j.nurpra.2014.07.013]
40. Menachemi N, Collum. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011 May;4:47-55. [doi: 10.2147/rmhp.s12985]

XSL•FO
RenderX

41. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
42. Donnelly C, Janssen A, Vinod S, Stone E, Harnett P, Shaw T. A systematic review of electronic medical record driven quality measurement and feedback systems. Int J Environ Res Public Health 2022 Dec 23;20(1):200 [FREE Full text] [doi: 10.3390/ijerph20010200] [Medline: 36612522]
43. Janssen A, Shaw T. The attitudes of medical practitioners towards the actionability of performance data. Int J Electron Healthc 2022;12(3). [doi: 10.1504/ijeh.2022.124492]
44. Blumenthal S. The use of clinical registries in the United States: a landscape survey. EGEMS (Wash DC) 2017 Dec 07;5(1):26 [FREE Full text] [doi: 10.5334/egems.248] [Medline: 29930965]
45. 5 things to know about clinical data registries. American Medical Association. 2014 Oct 15. URL: https://ama-assn.org/practice-management/digital/5-things-know-about-clinical-data-registries [accessed 2024-04-24]
46. Frøen JF, Myhre SL, Frost MJ, Chou D, Mehl G, Say L, et al. eRegistries: electronic registries for maternal and child health. BMC Pregnancy Childbirth 2016 Jan 19;16(1):11 [FREE Full text] [doi: 10.1186/s12884-016-0801-7] [Medline: 26791790]
47. Davey S, Davey A. Effect of practice management softwares among physicians of developing countries with special reference to Indian scenario by Mixed Method Technique. J Family Med Prim Care 2015;4(2):208-216 [FREE Full text] [doi: 10.4103/2249-4863.154637] [Medline: 25949969]
48. Baldwin G. How the internet is changing practice management systems. Health Data Manag 2000 Oct;8(10):42-4, 48-50, 52 passim. [Medline: 11140316]
49. Reader TW, Gillespie A, Roberts J. Patient complaints in healthcare systems: a systematic review and coding taxonomy. BMJ Qual Saf 2014 Aug 29;23(8):678-689 [FREE Full text] [doi: 10.1136/bmjqs-2013-002437] [Medline: 24876289]
50. Teasdale GM, Council of the Society of British Neurological Surgeons. Learning from Bristol: report of the public inquiry into children's heart surgery at Bristol Royal Infirmary 1984-1995. Br J Neurosurg 2002 Jun;16(3):211-216. [doi: 10.1080/02688690220148815] [Medline: 12201391]
51. Stey AM, Russell MM, Ko CY, Sacks GD, Dawes AJ, Gibbons MM. Clinical registries and quality measurement in surgery: a systematic review. Surgery 2015 Feb;157(2):381-395. [doi: 10.1016/j.surg.2014.08.097] [Medline: 25616951]
52. Chalmers K, Pearson SA, Badgery-Parker T, Brett J, Scott IA, Elshaug AG. Measuring 21 low-value hospital procedures: claims analysis of Australian private health insurance data (2010-2014). BMJ Open 2019 Mar 05;9(3):e024142 [FREE Full text] [doi: 10.1136/bmjopen-2018-024142] [Medline: 30842110]
53. Konrad R, Zhang W, Bjarndóttir M, Proaño R. Key considerations when using health insurance claims data in advanced data analyses: an experience report. Health Syst (Basingstoke) 2019 Mar 01;9(4):317-325 [FREE Full text] [doi: 10.1080/20476965.2019.1581433] [Medline: 33354323]
54. Valderas JM, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. Qual Life Res 2008 Nov 3;17(9):1125-1135. [doi: 10.1007/s11136-008-9396-4] [Medline: 18836850]
55. Tzelepis F, Sanson-Fisher R, Zucca A, Fradgley E. Measuring the quality of patient-centered care: why patient-reported measures are critical to reliable assessment. Patient Prefer Adher 2015 Jun;9:831-835. [doi: 10.2147/ppa.s81975]
56. Foster A, Croot L, Brazier J, Harris J, O'Cathain A. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. J Patient Rep Outcomes 2018 Dec 3;2(1):46 [FREE Full text] [doi: 10.1186/s41687-018-0072-3] [Medline: 30363333]
57. Mattick JS, Dziadek MA, Terrill BN, Kaplan W, Spigelman AD, Bowling FG, et al. The impact of genomics on the future of medicine and health. Med J Aust 2014 Jul 07;201(1):17-20. [doi: 10.5694/mja13.10920] [Medline: 24999876]
58. Georgiou A, Li J, Hardie RA, Wabe N, Horvath AR, Post JJ, et al. Diagnostic informatics-the role of digital health in diagnostic stewardship and the achievement of excellence, safety, and value. Front Digit Health 2021 Jun 10;3:659652 [FREE Full text] [doi: 10.3389/fdgth.2021.659652] [Medline: 34713132]
59. Porterfield A, Engelbert K, Coustasse A. Electronic prescribing: improving the efficiency and accuracy of prescribing in the ambulatory care setting. Perspect Health Inf Manag 2014 Apr 1;11(Spring):1g [FREE Full text] [Medline: 24808808]
60. Odukoya OK, Chui MA. E-prescribing: a focused review and new approach to addressing safety in pharmacies and primary care. Res Social Adm Pharm 2013 Nov;9(6):996-1003 [FREE Full text] [doi: 10.1016/j.sapharm.2012.09.004] [Medline: 23062769]
61. Gray K, Chapman W, Khan UR, Borda A, Budge M, Dutch M, et al. The rapid development of virtual care tools in response to COVID-19: case studies in three Australian health services. JMIR Form Res 2022 Apr 06;6(4):e32619 [FREE Full text] [doi: 10.2196/32619] [Medline: 35297765]
62. Nielsen JC, Kautzner J, Casado-Arroyo R, Burri H, Callens S, Cowie MR, et al. Remote monitoring of cardiac implanted electronic devices: legal requirements and ethical principles - ESC Regulatory Affairs Committee/EHRA joint task force report. Europace 2020 Nov 01;22(11):1742-1758. [doi: 10.1093/europace/euaa168] [Medline: 32725140]
63. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Posterior approaches in malleolar fracture: when, why and how. J Biol Regul Homeost Agents 2020;34(3 Suppl. 2):89-95. ADVANCES IN MUSCULOSKELETAL DISEASES AND INFECTIONS. [Medline: 32856446]

XSL•FO
RenderX

64. Cochran AR, Raub KM, Murphy KJ, Iannitti DA, Vrochides D. Novel use of REDCap to develop an advanced platform to display predictive analytics and track compliance with Enhanced Recovery After Surgery for pancreaticoduodenectomy. Int J Med Inform 2018 Nov;119:54-60. [doi: 10.1016/j.ijmedinf.2018.09.001] [Medline: 30342686]
65. Pang X, Kozlowski N, Wu S, Jiang M, Huang Y, Mao P, et al. Construction and management of ARDS/sepsis registry with REDCap. J Thorac Dis 2014 Sep;6(9):1293-1299 [FREE Full text] [doi: 10.3978/j.issn.2072-1439.2014.09.07] [Medline: 25276372]
66. Janssen A, Donnelly C, Kay J, Thiem P, Saavedra A, Pathmanathan N, et al. Developing an intranet-based lymphedema dashboard for breast cancer multidisciplinary teams: design research study. J Med Internet Res 2020 Apr 21;22(4):e13188 [FREE Full text] [doi: 10.2196/13188] [Medline: 32314968]
67. Naranjo-Hernández D, Reina-Tosina J, Roa LM, Barbarov-Rostán G, Aresté-Fosalba N, Lara-Ruiz A, et al. Smart bioimpedance spectroscopy device for body composition estimation. Sensors (Basel) 2019 Dec 21;20(1):70 [FREE Full text] [doi: 10.3390/s20010070] [Medline: 31877699]
68. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD. Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot 2018 Feb 29;14(1). [doi: 10.1002/rcs.1850] [Medline: 28660725]
69. Yasar MS, Alemzadeh H. Real-time context-aware detection of unsafe events in robot-assisted surgery. In: Proceedings of the 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). 2020 Presented at: DSN 2020; June 29-July 02, 2020; Valencia, Spain URL: https://ieeexplore.ieee.org/document/9153443 [doi: 10.1109/dsn48063.2020.00054]
70. Hawley S, Yu J, Bogetic N, Potapova N, Wakefield C, Thompson M, et al. Digitization of measurement-based care pathways in mental health through REDCap and electronic health record integration: development and usability study. J Med Internet Res 2021 May 20;23(5):e25656 [FREE Full text] [doi: 10.2196/25656] [Medline: 34014169]
71. Kao CK, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM&R 2017 May 18;9(5S):S106-S115. [doi: 10.1016/j.pmrj.2017.02.018]
72. Feng S, Mäntymäki M, Dhir A, Salmela H. How self-tracking and the quantified self promote health and well-being: systematic review. J Med Internet Res 2021 Sep 21;23(9):e25171 [FREE Full text] [doi: 10.2196/25171] [Medline: 34546176]
73. Placona AM, Rathert C. Are online patient reviews associated with health care outcomes? A systematic review of the literature. Med Care Res Rev 2022 Feb 24;79(1):3-16. [doi: 10.1177/10775587211014534] [Medline: 34027743]
74. Trehan SK, Daluiski A. Online patient ratings: why they matter and what they mean. J Hand Surg Am 2016 Feb;41(2):316-319. [doi: 10.1016/j.jhsa.2015.04.018] [Medline: 26754195]
75. Yang A, Finkelstein M, Palese M. PD16-09 Online patient ratings of urologists are correlated with short-term outcomes following major urologic surgeries. J Urol 2019 Apr;201(Supplement 4):e303-e304. [doi: 10.1097/01.ju.0000555557.53309.fa]
76. Emmert M, Meszmer N, Sander U. Do health care providers use online patient ratings to improve the quality of care? Results from an online-based cross-sectional study. J Med Internet Res 2016 Sep 19;18(9):e254 [FREE Full text] [doi: 10.2196/jmir.5889] [Medline: 27644135]
77. Paré G, Trudel MC, Forget P. Adoption, use, and impact of e-booking in private medical practices: mixed-methods evaluation of a two-year showcase project in Canada. JMIR Med Inform 2014 Sep 24;2(2):e24 [FREE Full text] [doi: 10.2196/medinform.3669] [Medline: 25600414]
78. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 13;328(7449):1166. [doi: 10.1136/bmj.328.7449.1166]
79. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
80. Didegah F, Ghaseminik Z, Alperin JP. Using a diabetes discussion forum and Wikipedia to detect the alignment of public interests and the research literature. J Almetrics 2020 Jul 10;3(1). [doi: 10.29024/joa.26]
81. Bender JL, Jimenez-Marroquin MC, Ferris LE, Katz J, Jadad AR. Online communities for breast cancer survivors: a review and analysis of their characteristics and levels of use. Support Care Cancer 2013 May 23;21(5):1253-1263. [doi: 10.1007/s00520-012-1655-9] [Medline: 23179491]
82. Ba S, Wang L. Digital health communities: the effect of their motivation mechanisms. Decis Support Syst 2013 Nov;55(4):941-947. [doi: 10.1016/j.dss.2013.01.003]
83. Jun SP, Yoo HS, Choi S. Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. Technol Forecast Soc Change 2018 May;130:69-87. [doi: 10.1016/j.techfore.2017.11.009]
84. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. J Biomed Inform 2020 Aug;108:103500 [FREE Full text] [doi: 10.1016/j.jbi.2020.103500] [Medline: 32622833]
85. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 09;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]

XSL•FO
RenderX

86. Mattison G, Canfell O, Forrester D, Dobbins C, Smith D, Töyräs J, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 01;24(7):e36690 [FREE Full text] [doi: 10.2196/36690] [Medline: 35776492]
87. Flitcroft L, Chen WS, Meyer D. The demographic representativeness and health outcomes of digital health station users: longitudinal study. J Med Internet Res 2020 Jun 26;22(6):e14977 [FREE Full text] [doi: 10.2196/14977] [Medline: 32589150]
88. Best Buy Health homepage. Best Buy Health. URL: https://www.bestbuyhealth.com/ [accessed 2024-04-24]
89. We deliver healthcare that puts patients first. Eucalyptus. 2024. URL: https://eucalyptus.vc/ [accessed 2024-05-13]
90. Oliver D. David Oliver: lessons from the Babylon health saga. BMJ 2019 Jun 05;365:l2387. [doi: 10.1136/bmj.l2387] [Medline: 31167923]
91. Hong YA, Liang C, Radcliff TA, Wigfall LT, Street RL. What do patients say about doctors online? A systematic review of studies on patient online reviews. J Med Internet Res 2019 Apr 08;21(4):e12521 [FREE Full text] [doi: 10.2196/12521] [Medline: 30958276]
92. Atherton H, Fleming J, Williams V, Powell J. Online patient feedback: a cross-sectional survey of the attitudes and experiences of United Kingdom health care professionals. J Health Serv Res Policy 2019 Oct 02;24(4):235-244 [FREE Full text] [doi: 10.1177/1355819619844540] [Medline: 31154858]
93. Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (infoveillance): systematic review. J Med Internet Res 2020 Mar 13;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
94. Karmen CL, Reisfeld MA, McIntyre MK, Timmermans R, Frishman W. The clinical value of heart rate monitoring using an Apple watch. Cardiol Rev 2019;27(2):60-62. [doi: 10.1097/CRD.0000000000000243] [Medline: 30628918]
95. Meyer J, Kay J, Epstein DA, Eslambolchilar P, Tang LM. A life of data: characteristics and challenges of very long term self-tracking for health and wellness. ACM Trans Comput Healthcare 2020 Mar 11;1(2):1-4. [doi: 10.1145/3373719]
96. Ceross A, Bergmann J. Evaluating the presence of software-as-a-medical-device in the Australian therapeutic goods register. Prosthesis 2021 Aug 24;3(3):221-228. [doi: 10.3390/prosthesis3030022]
97. Healthcare booking sites must be transparent about use of patient data: RACGP. The Royal Australian College of General Practitioners. 2018. URL: https://www1.racgp.org.au/newsgp/professional/gp-booking-sites-must-be-transparent-about-use-of [accessed 2024-05-13]
98. HealthEngine in court for allegedly misusing patient data and manipulating reviews. Australian Competition & Consumer Commission. 2019 Aug 8. URL: https://tinyurl.com/bdhsnzj8 [accessed 2024-04-24]
99. Nguyen H, Butow P, Dhillon H, Sundaresan P. A review of the barriers to using Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs) in routine cancer care. J Med Radiat Sci 2021 Jun 19;68(2):186-195 [FREE Full text] [doi: 10.1002/jmrs.421] [Medline: 32815314]
100. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
101. Fiore-Gartland B, Neff G. Communication, mediation, and the expectations of data: data valences across health and wellness communities. Int J Commun 2015;9(2015):1466-1484.
102. Holmes JH, Beinlich J, Boland MR, Bowles KH, Chen Y, Cook TS, et al. Why is the electronic health record so challenging for research and clinical care? Methods Inf Med 2021 May 19;60(1-02):32-48 [FREE Full text] [doi: 10.1055/s-0041-1731784] [Medline: 34282602]
103. de Vos MS, Hamming JF, Marang-van de Mheen PJ. The problem with using patient complaints for improvement. BMJ Qual Saf 2018 Sep 03;27(9):758-762. [doi: 10.1136/bmjqs-2017-007463] [Medline: 29298910]
104. Greenhalgh J, Dalkin S, Gibbons E, Wright J, Valderas JM, Meads D, et al. How do aggregated patient-reported outcome measures data stimulate health care improvement? A realist synthesis. J Health Serv Res Policy 2018 Jan 20;23(1):57-65 [FREE Full text] [doi: 10.1177/1355819617740925] [Medline: 29260592]
105. Lupton D. Digital health now and in the future: findings from a participatory design stakeholder workshop. Digit Health 2017 Nov 09;3:2055207617740018 [FREE Full text] [doi: 10.1177/2055207617740018] [Medline: 29942616]
106. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 2015 Jul;19(4):1216-1223. [doi: 10.1109/jbhi.2015.2404829]
107. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016 Feb 30;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
108. Bucalon B, Shaw T, Brown K, Kay J. State-of-the-art dashboards on clinical indicator data to support reflection on practice: scoping review. JMIR Med Inform 2022 Feb 14;10(2):e32695 [FREE Full text] [doi: 10.2196/32695] [Medline: 35156928]
109. Impact of digital health on the safety and quality of health care. Australian Commission on Safety and Quality in Health Care. 2018. URL: https://tinyurl.com/2s3v3try [accessed 2024-04-24]
110. Marthick M, Janssen A, Cheema BS, Alison J, Shaw T, Dhillon H. Feasibility of an interactive patient portal for monitoring physical activity, remote symptom reporting, and patient education in oncology: qualitative study. JMIR Cancer 2019 Nov 28;5(2):e15539 [FREE Full text] [doi: 10.2196/15539] [Medline: 31778123]

XSL•FO
RenderX

111. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019 Apr 11;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]
112. Bell SK, Mejilla R, Anselmo M, Darer JD, Elmore JG, Leveille S, et al. When doctors share visit notes with patients: a study of patient and doctor perceptions of documentation errors, safety opportunities and the patient-doctor relationship. BMJ Qual Saf 2017 Apr 18;26(4):262-270 [FREE Full text] [doi: 10.1136/bmjqs-2015-004697] [Medline: 27193032]
113. Delbanco T, Walker J, Darer JD, Elmore JG, Feldman HJ, Leveille SG, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010 Jul 20;153(2):121-125 [FREE Full text] [doi: 10.7326/0003-4819-153-2-201007200-00008] [Medline: 20643992]
114. Yung A, Kay J, Beale P, Gibson KA, Shaw T. Computer-based decision tools for shared therapeutic decision-making in oncology: systematic review. JMIR Cancer 2021 Oct 26;7(4):e31616 [FREE Full text] [doi: 10.2196/31616] [Medline: 34544680]
115. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
116. West P, Giordano R, Van Kleek M, Shadbolt N. The quantified patient in the doctor's office: challenges and opportunities. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI'16; May 7-12, 2016; San Jose, CA. [doi: 10.1145/2858036.2858445]
117. Abrams J, Sossong S, Schwamm LH, Barsanti L, Carter M, Kling N, et al. Practical issues in delivery of clinician-to-patient telemental health in an academic medical center. Harv Rev Psychiatry 2017;25(3):135-145. [doi: 10.1097/hrp.0000000000000142]
118. Kazmi Z. Effects of exam room EHR use on doctor-patient communication: a systematic literature review. Inform Prim Care 2013;21(1):30-39 [FREE Full text] [doi: 10.14236/jhi.v21i1.37]
119. Janssen A, Brunner M, Keep M, Hines M, Nagarajan SV, Kielly-Carroll C, et al. Interdisciplinary eHealth practice in cancer care: a review of the literature. Int J Environ Res Public Health 2017 Oct 25;14(11):1289 [FREE Full text] [doi: 10.3390/ijerph14111289] [Medline: 29068377]
120. Li J, Robertson T. Physical space and information space: studies of collaboration in distributed multi-disciplinary medical team meetings. Behav Inf Technol 2011 Jul;30(4):443-454. [doi: 10.1080/0144929x.2011.577194]
121. Chaiyachati KH, Shea JA, Asch DA, Liu M, Bellini LM, Dine CJ, et al. Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. JAMA Intern Med 2019 Jun 01;179(6):760-767 [FREE Full text] [doi: 10.1001/jamainternmed.2019.0095] [Medline: 30985861]
122. Westbrook JI, Ampt A, Kearney L, Rob MI. All in a day's work: an observational study to quantify how and with whom doctors on hospital wards spend their time. Med J Aust 2008 May 05;188(9):506-509. [doi: 10.5694/j.1326-5377.2008.tb01762.x] [Medline: 18459920]
123. Building a professional performance framework. Medical Board of Australia. 2024. URL: https://tinyurl.com/e7rxu8wy [accessed 2024-05-12]
124. Archer J, Bloor K, Bojke C, Boyd A, Bryce M, Ferguson J, et al. Evaluating the development of medical revalidation in England and its impact on organisational performance and medical practice: overview report. University of Manchester, University of York and University of Plymouth. 2018 Mar. URL: https://www.alliancembs.manchester.ac.uk/media/ambs/content-assets/images/news/Final-report---the-development-and-impact-of-medical-revalidation---final.pdf [accessed 2024-04-24]
125. Shaw T, Janssen A, Crampton R, O'Leary F, Hoyle P, Jones A, et al. Attitudes of health professionals to using routinely collected clinical data for performance feedback and personalised professional development. Med J Aust 2019 Apr 30;210 Suppl 6(S6):S17-S21. [doi: 10.5694/mja2.50022] [Medline: 30927464]
126. Kirschner N, Ejnes Y, Tape TG, Auted AM, Fihn SD, Hatton D, et al. Healthcare transparency—focus on price and clinical performance information. American College of Physicians. 2010. URL: https://www.acponline.org/sites/default/files/documents/advocacy/current_policy_papers/transparency.pdf [accessed 2024-04-24]
127. Mehrotra A, Schleifer D, Shefrin A, Ducas AM. Defining the goals of health care price transparency: not just shopping around. NEJM Catalyst 2018 Jun 26.

Abbreviations

AI: artificial intelligence
EHD: electronic health data
EHR: electronic health record
EMR: electronic medical record
ICT: information and communication technology
PAS: patient administration system
REDCap: Research Electronic Data Capture
WPR: web-based patient review

XSL•FO
RenderX

Edited by T Leung; submitted 29.03.23; peer-reviewed by M Glöggler, Z Zandesh, R Taggart; comments to author 23.07.23; revised version received 05.10.23; accepted 31.01.24; published 31.05.24.

Please cite as:
Janssen A, Donnelly C, Shaw T
A Taxonomy for Health Information Systems
J Med Internet Res 2024;26:e47682
URL: https://www.jmir.org/2024/1/e47682
doi:10.2196/47682
PMID:38820575

©Anna Janssen, Candice Donnelly, Tim Shaw. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Does an App a Day Keep the Doctor Away? AI Symptom Checker Applications, Entrenched Bias, and Professional Responsibility

Ma'n H Zawati[1*], LLB, LLM, DCL, PhD; Michael Lang[1*], BA, BCL, LLB, LLM

Centre of Genomics and Policy, McGill University, Montreal, QC, Canada

[*]all authors contributed equally

Corresponding Author:
Ma'n H Zawati, LLB, LLM, DCL, PhD
Centre of Genomics and Policy
McGill University
5200-740 Dr Penfield Avenue
Montreal, QC, H3A 0G1
Canada
Phone: 1 5143988155
Email: man.zawati@mcgill.ca

Abstract

The growing prominence of artificial intelligence (AI) in mobile health (mHealth) has given rise to a distinct subset of apps that provide users with diagnostic information using their inputted health status and symptom information—AI-powered symptom checker apps (AISympCheck). While these apps may potentially increase access to health care, they raise consequential ethical and legal questions. This paper will highlight notable concerns with AI usage in the health care system, further entrenchment of preexisting biases in the health care system and issues with professional accountability. To provide an in-depth analysis of the issues of bias and complications of professional obligations and liability, we focus on 2 mHealth apps as examples—Babylon and Ada. We selected these 2 apps as they were both widely distributed during the COVID-19 pandemic and make prominent claims about their use of AI for the purpose of assessing user symptoms. First, bias entrenchment often originates from the data used to train AI systems, causing the AI to replicate these inequalities through a “garbage in, garbage out” phenomenon. Users of these apps are also unlikely to be demographically representative of the larger population, leading to distorted results. Second, professional accountability poses a substantial challenge given the vast diversity and lack of regulation surrounding the reliability of AISympCheck apps. It is unclear whether these apps should be subject to safety reviews, who is responsible for app-mediated misdiagnosis, and whether these apps ought to be recommended by physicians. With the rapidly increasing number of apps, there remains little guidance available for health professionals. Professional bodies and advocacy organizations have a particularly important role to play in addressing these ethical and legal gaps. Implementing technical safeguards within these apps could mitigate bias, AIs could be trained with primarily neutral data, and apps could be subject to a system of regulation to allow users to make informed decisions. In our view, it is critical that these legal concerns are considered throughout the design and implementation of these potentially disruptive technologies. Entrenched bias and professional responsibility, while operating in different ways, are ultimately exacerbated by the unregulated nature of mHealth.

(J Med Internet Res 2024;26:e50344) doi:10.2196/50344

KEYWORDS

artificial intelligence; applications; mobile health; mHealth; bias; biases; professional obligations; professional obligation; app; apps; application; symptom checker; symptom checkers; diagnose; diagnosis; self-diagnose; self-diagnosis; ethic; ethics; ethical; regulation; regulations; legal; law; laws; safety; mobile phone

Background

Smartphone apps with health-related functions are becoming increasingly popular [1]. A large portion of the global population owns smartphones [2] and many smartphone owners are using widely available apps to monitor their fitness, track their health data, and receive information about illness [3]. Smartphone health apps vary greatly in their functioning and technological complexity. In recent years, a number of highly sophisticated smartphone apps powered by artificial intelligence (AI) have emerged [4]. One prominent subset of AI-powered health apps offers diagnostic information by providing users the capacity to input health status and symptom information [5]. The app,

XSL•FO
RenderX

in return, gives users feedback consisting of lists of conditions that might be responsible for the reported symptoms.

Apps of this kind might be capable of increasing access to high-quality medical care, potentially helping to address population inequality in care access. At the same time, symptom checker apps powered by AI also raise serious ethical and legal questions [6]. Chief among them is that AI-powered symptom checker (AISympCheck) apps, to which we will refer in this viewpoint as AISympCheck apps, may work to entrench biases preexisting in the health care system [7] and that they may generate confusion about the legal obligations of health professionals in the course of interacting with these systems [8]. In our view, it is critical that these legal concerns are considered throughout the design and implementation of these potentially disruptive technologies. While it may certainly be the case, AISympCheck apps will, on balance, improve access to preclinical health information in a way that tends to improve health outcomes, this outcome is not assured, particularly if the quality of AISympCheck apps varies wildly. It is important, in other words, to work to ensure that AISympCheck apps are lawfully designed so that their safety and efficacy can be relied upon by users and so that these systems can have beneficial impacts on the broader health care system.

This viewpoint aims to address this idea in 2 parts. First, we provide a brief overview of mobile health (mHealth) and AISympCheck apps by highlighting 2 prominent examples—Babylon and Ada. Second, we outline concerns surrounding (1) the replication and entrenchment of bias and (2) the complication of professional obligations and liability in the use of AISympCheck apps. We conclude by briefly considering how these challenges may be addressed through policy and regulatory reform.

AISympCheck Apps

Overview

This first part of the viewpoint provides an overview of AISympCheck apps. We first describe the broader categories of mHealth apps and symptom checker apps of which AI-powered apps are a part. From there, we specify the unique emerging characteristics of symptom checker apps powered by AI. We outline 2 prominent examples as a way of highlighting how these apps function.

Symptom Checker Apps

Mobile devices are significantly altering the ways individuals interact with health care systems around the world [9]. As we detailed above, smartphone connectivity has become a widespread phenomenon across much of the human population. While there is no universally accepted definition of the smartphone, many commenters track the emergence of the concept to the release of Apple's iPhone in 2007 [10]. In most of the world's highly industrialized countries, an excess of 80% of people own a smartphone [11] and across the globe, nearly 4 billion people were smartphone owners in 2020 [12]. These devices are becoming increasingly technologically sophisticated, with features such as high-definition cameras, multiple gigabytes of storage, relatively precise accelerometers, and GPS connectivity quickly becoming standard. These factors combine to make smartphones potentially revolutionary tools for collecting, storing, and sharing health information [13]. Smartphone apps with health-related functions serve a wide variety of purposes, from recording basic fitness data [14] to measuring heart function [15]. Such apps are becoming incredibly numerous, by 1 estimate, more than 350,000 health apps are presently available across Android and iOS systems [16].

Of the several hundred thousand available mHealth apps, approximately 7% of them perform functions broadly resembling clinical diagnosis [17]. These are what we refer to here as "symptom checker applications." As we outline, this category of mHealth apps, potentially consisting of more than 25,000 examples, is fairly diverse. For our purposes, we follow a conception of the symptom checker category, set out by Lupton and Jutel [18], as a set of apps "by which self-diagnosis can take place." In most cases, symptom checker apps function in a relatively straightforward way. Users are generally prompted to enter information about symptoms or ailments that are afflicting them. Such information may be processed in a freehand text box or from a drop-down list from which specific symptom descriptions may be selected as appropriate. Symptom information is often supplemented with demographic data such as age, weight, height, or ethnic identity. Following internal app processing, users are usually provided a list of candidate conditions and diseases that may be responsible for the reported symptoms. This information is sometimes supplemented with a probability score expressed as a percentage or some analogous measure indicating how likely a given condition is to be responsible for the specific symptom cluster described by the user. Many symptom checker apps convey a written warning to the effect that information communicated by the app does not constitute a diagnosis and should not be used to replace appropriate clinical care [19]. It is not clear how app users interpret these warnings or whether waivers of nondiagnosis have the meaning in the law that they intend. It may be, for example, that a reviewing court will determine that an app's tendency to draw causal associations between descriptions of symptoms and disease will constitute a diagnosis in potential contravention of reservations of professional activity to physicians [19].

Symptom checker apps, like mHealth more generally, are a recently emerging and rapidly developing phenomenon in health care. While a comprehensive history of the evolution of symptom checker apps is well beyond the scope of this viewpoint, it may be useful to provide a broad overview of major developments in this space. In 2013, the United Kingdom's National Health Service published perhaps the most sophisticated early example of a symptom checker app. This app engaged users in a general health assessment, "with users working through a set of questions on health and symptoms, followed by more detailed assessments" [20]. Guidance provided by the app would range from at-home self-care to attending an emergency room [20]. In the period since 2013, symptom checker apps have proliferated. Many of the apps now available are developed and distributed by private entities and businesses, such as the symptom checker app WebMD [18]. This app takes

XSL•FO
RenderX

the same basic form as the NHS (National Health Service) Symptom Checker App, though without direct integration into national health care infrastructure. WebMD's symptom checker interface opens upon download with a disclaimer. Users are warned that while the app provides "useful information," it is "not a substitute for professional medical advice, diagnosis, or treatment" [21]. User symptoms may be identified according to an animated body plan or by using the app's database search function [21]. Available symptom descriptions range from cold hands to congestive heart failure. Upon entering symptom descriptions and information about current and past medication use, the app delivers an overview of conditions that might be responsible for the reported symptoms [21]. Potential conditions are presented according to the strength of the relationship between reported symptoms and the identified condition.

AI-Powered Symptom Checker Apps

Overview

This viewpoint draws a further distinction between symptom checker apps and AISympCheck apps. The essential difference in these categories is not elaborate, AISympCheck apps are the subset of symptom checker apps that make use of artificially intelligent processing techniques to deliver information associating symptom inputs with disease. As we outline, there may be important functional differences between those apps holding themselves out in marketing and instructional materials as AI-powered symptom checkers and those actually applying artificially intelligent processing models. There are ongoing debates about what precisely qualifies as an AI app in the mHealth space, debates into which we will not enter here. For our present purposes, we refer to all apps promoted or described by their developers as using artificially intelligent models as AISympCheck apps. Multiple AISympCheck apps have been developed and released in recent years [18] and many have entered popular and academic discussion. By way of example, we provide a short overview of 2 especially prominent AISympCheck apps—Babylon and Ada. We have elected to focus on these as exemplars of the emerging AISympCheck app category, but do not mean to suggest that these are the only presently available AISympCheck apps, nor that Babylon and Ada are more significant than other AISympCheck apps that may be on the market. We mean only for these to serve as examples that illustrate a novel, emerging concept. Babylon and Ada were widely distributed during the COVID-19 pandemic and make prominent claims about their use of AI for the purpose of assessing user symptoms.

Babylon

Babylon was first released in the United Kingdom in 2017 [6]. Babylon describes itself as a "revolutionary digital health company," combining elements of an AI-powered symptom checker and video consult telemedicine [22]. This app deploys an artificially intelligent interface trained on a data set composed of physician clinic notes, user medical records, probabilistic modeling architecture, and a series of diagnostic simulations [23]. Just as in the case of the symptom checker apps described above, a Babylon assessment begins with users searching for symptom descriptions from a standardized list. Upon selecting those most appropriate to a user's situation, the app interface may pose additional queries. Babylon then provides a list of potentially responsible conditions and advice on the next steps. Such advice might include home care or a physician consult. As Babylon also offers telehealth services, the app may connect users directly with a physician.

Shortly after its release, Babylon claimed its artificially intelligent diagnostic model outperforms human physicians on the Membership of the Royal College of General Practitioners examination, a mandatory standardized test taken by British trainee general practitioners prior to their admission to the practice [24]. But these results have been subject to a certain degree of dispute, with some questioning their validity on methodological grounds [25] and pointing out that Babylon training has been conducted on data primarily collected from young, healthy patients who have been encouraged to register for health services through aggressive direct marketing [26]. With the arrival of the COVID-19 pandemic in early 2020, Babylon's adoption appears to have expanded. In several Canadian jurisdictions, for example, provincial health systems promoted Babylon use as a way of easing pressure on in-person care. In the provinces of British Columbia, Alberta, and Ontario, telemedicine services on the Babylon app were provided by provincial health insurance schemes [27]. Babylon was not intended in these arrangements to diagnose, treat, or manage COVID-19 directly. Academics and activists raised concerns about Babylon's promotion by provincial health ministries. Critics note that Babylon's AI-powered interface has not been independently reviewed for safety and efficacy [28]. Others note that Babylon received initial funding from Saudi Arabia's sovereign wealth fund, potentially raising conflict of interest concerns for some users [29].

Ada

Like Babylon, Ada describes itself as a "world-leading clinical AI" [30]. Launched on the Apple App Store in 2016, Ada reports having conducted more than 20 million health assessments and having more than 13 million active users engaging with the app in 11 product languages [30]. Unsurprisingly, Ada's user interface bears a great deal of resemblance to Babylon and to the other symptom checker apps described. Ada operates by walking users through a series of prompts to illicit symptom information. The app poses a series of increasingly targeted questions and provides a list of potentially responsible conditions. These are organized probabilistically, with descriptions of the proportion of users who are likely to have a specific condition given the symptoms described. Ada makes fairly sweeping claims about the accuracy of its AI-powered model, a trait as we outlined, it shares with Babylon. Ada's website, for example, suggests that it "is the most accurate symptom assessment app" [30]. At least 1 study suggests that Ada's rate of condition suggestion accuracy is consistently just over 70%, as compared with a general practitioner's accuracy of just over 82% [31]. In the same study, Babylon was determined to have a condition suggestion accuracy rate of about 32% [31].

On the whole, Babylon and Ada perform a similar set of functions in a roughly similar way. There is a level of technical disagreement about how these apps function—whether system

XSL•FO
RenderX

modeling can properly be described as AI, how best to measure output accuracy, and how these systems were initially funded and produced. Each of these functional factors raises distinct ethical and legal questions. In the following section, we outline 2 such questions—entrenched bias and professional responsibility.

Legal and Ethical Considerations: Entrenched Bias and Professional Responsibility

Overview

In the previous section, we provided an overview of the AISympCheck app. We summarized the functions and conceptual orientation of 2 such apps—Babylon and Ada. In this second part of the viewpoint, we turn to some of the ethical and legal challenges likely to be associated with the proliferation of AISympCheck apps. While an active and growing literature considers the ethical and legal issues raised by the adoption of generalized mHealth apps [32], considerations specific to symptom checker apps have received considerably less attention. Less still has been written on ethical and legal questions that might be particular to the subset of apps powered by AI, though media and activist attention given to certain high-profile AI-powered apps, such as Babylon and Ada, appears to have stimulated a certain level of interest in the field [33]. We focus here on 2 sets of issues that, while challenges for mHealth generally, may have particularly salient implications for AISympCheck apps—entrenched bias and professional responsibility.

Entrenched Bias

It is well understood that bias, both implicit and explicit, is a persistent and troubling feature of modern health care [34]. Perhaps its most pernicious source is in its relationship to a wide array of social demographic factors, usually called the social determinants of health. These include factors such as race, immigration status, income, and educational attainment, each of which track closely with an individual's capacity to care, benefit from treatment, and experience complications [35]. Demographic factors also correlate closely with health status and the risk that individuals will become ill in the first place, a reality painfully and unambiguously exposed during the COVID-19 pandemic [36]. Any novel health intervention or technology is likely to raise concerns about its intersection with existing lines of oppression and inequality. Those most vulnerable and most in need of access to improvements in health care delivery are also often those least able to benefit from them.

This is almost certain to be the case in the field of mHealth, in which factors such as income, health literacy, and level of comfort with technology, are likely to have an enormous impact on the capacity of individuals to make use of and benefit from even the most effective and widely available tools [37]. Individuals who are unable to afford a smartphone, for example, are, as per the definition, excluded from the possibility of benefitting from most mobile apps. Likewise, persons with a low level of health or technological literacy might, even where mobile technologies are accessible to them, be unable to effectively use these systems in a beneficial or productive way. Notably, AI systems could be designed in such a way that they identify and correct biased decision-making by reporting to users when a training data set is demographically unrepresentative or otherwise flawed. This solution, though, can only do so much. Reporting that a system relies on unrepresentative training data does not repair the data itself, nor does it answer the question of how much bias is acceptable in the relevant decision-making endeavor. These concerns are surely relevant in the context of AISympCheck apps, though they also take on a secondary and confounding character.

Artificially intelligent systems often face what is sometimes described as a "black box" or inscrutability problem [38]. Much of our most technologically sophisticated AI learns and improves in function over time [39]. It does so by virtue of highly complex, layered programming designed to loosely replicate human neuronal processes [40]. AI systems using machine learning techniques, for example, require minimal human instruction and supervision. These systems, when provided sufficient data, will identify and interpret patterns in a manner similar to human reasoning, though usually significantly more efficiently. Numerous commenters note that this manner of functioning raises serious concerns that AI systems will produce biased decisional outcomes [7]. Bias in this context is likely to originate from the data on which AI systems are trained. Owing to bias preexisting in the health care system, it may be that AI models will work to replicate and entrench inequalities that produce disparate outcomes, rather than remedying them. The essential nature of this problem is often summarized by commenters as the "garbage in, garbage out" phenomenon [41]. Artificially intelligent systems trained on biased data are likely to produce decisional outputs that by necessity reflect that bias.

In the context of AISympCheck apps, this kind of biased output might have a particularly worrying salience. Bias embedded in the health care system owing to compounding social factors are likely also to affect the effective use of modern consumer technologies. We know, for example, that smartphone ownership and rate of use is inconsistent across demographic measures. While 96% of Americans with household incomes above US $75,000 own a smartphone, for example, the rate of ownership among those earning less than US $30,000 in household income is only 76% [42]. Likewise, 93% of Americans with a college degree report owning a smartphone. Only 75% of Americans with a high school diploma or less are smartphone owners [42]. Demographic factors such as level of educational attainment and household income are known to be associated with poorer health outcomes and lower levels of access to the health care system. It is potentially worrying that the increasing adoption of AISympCheck apps may turn out to be poorly suited to addressing such disparities and may, in fact, inadvertently make them worse. Interpreted another way, AI system training that is happening now may turn out to have a significant impact on the future shape of the field. As novel tools and systems are developed and commercially distributed, they may be dominantly affected by presently available (biased) data sets. Current users of AISympCheck apps are not likely to be demographically representative of the population at large, potentially distorting the kinds of results such apps are capable

XSL•FO
RenderX

of providing. An AISympCheck app hypothetically trained on data obtained from a data set composed of a disproportionate number of male users, for example, may be inadequately suited to providing diagnostic information addressing conditions primarily affecting nonmale users. This may not be a temporary problem, for even as app use becomes more demographically equitable over time, initial training parameters may have a pernicious and lasting impact.

These considerations raise the possibility that AISympCheck apps will work to entrench and replicate biases preexisting in our health care systems and societies. This is a concern that may limit the capacity of these tools to achieve their ostensible aim of increasing equitable access to health information and care.

Professional Responsibility

Another factor keeping AISympCheck apps from realizing their anticipated clinical potential is the possibility they will integrate poorly into the professional practice of medicine. AISympCheck apps raise a cluster of questions surrounding professional and regulatory responsibility. It is unclear, for example, whether such apps are subject to the same safety and efficacy review usually carried out prior to the regulatory approval of novel medical devices. It is likewise unclear whether physicians ought to recommend AISympCheck apps to their patients. Where prescribing a mobile app could be a clinical benefit, it is unclear how a physician ought to select an appropriate system from the many thousands on the market. Patients, moreover, are likely to use AISympCheck apps outside of their relationships with doctors. They may, from time to time, present app-generated information or diagnoses in the clinic. It is not obvious how a physician ought to conceive of or manage such information. Physicians are generally concerned about the increasing prevalence of mHealth apps in the health care system. In one study, two-thirds of Australian general practitioners had used mHealth apps in some capacity in their practice and more than half had recommended apps to their patients [16]. But their use was far from universal. Physicians might be hesitant to use and recommend mHealth apps for a variety of reasons. Byambasuren et al [16] find 2 major reasons for avoiding the clinical adoption of mHealth apps. First, an absence of adequate knowledge about which apps are most effective (n=372, 60%) and second, the lack of trustworthy sources through which to access health apps (n=96, 15%). Surveyed physicians indicated that increased mHealth training (n=243, 30%) and lists of apps evaluated for safety and efficacy by a reliable source (n=224, 28%) would increase comfort in the clinical adoption of mHealth apps [16]. These findings suggest that many physicians are (perhaps sensibly) fearful of inadvertently adopting ineffective, potentially harmful apps in their practice.

Without proper training and guidelines to follow, it is quite unreasonable that physicians could be expected to determine for themselves which mHealth apps are best suited to serving a patient's interests and how these apps ought to be used in the provision of care. There are simply too many apps with too many features and too many potential complications. Even in highly specific, discrete subspecialties, the number of available apps is rapidly increasing. Against this backdrop, formal law provides little in the way of guidance for clinicians. The mHealth apps, as we suggested, are typically not governed by existing formal regulations. Symptom checker apps in Canada, for example, are explicitly excluded from regulatory oversight [43]. Similar situations can be found in jurisdictions around the world [44]. There are a range of reasons that might explain the relative paucity of regulation surrounding symptom checker apps, including that such apps are exceptionally technically versatile and may be developed in a diverse number of settings, from hospitals to corporations to the homes of individuals [44]. Just as the number and diversity of apps create difficulty for clinicians in deciding whether to use apps in their medical practice, so too might it generate challenges for regulators.

In the absence of formal regulation, advocacy organizations and professional bodies have taken up some of the work of controlling the use of mHealth apps through professional responsibility regimes, by which we mean the quasi-judicial procedures according to which physicians and other health care practitioners are regulated by the professional bodies of which they are members. The United Kingdom's National Health Service, for example, maintains a searchable database of apps reviewed for clinical safety, accessibility, usability, and technical stability [45]. The App Library serves primarily as an advisory resource for clinicians and the public. It does not control clinician activities, does not provide guidance on appropriate clinical applications, and does not purport to serve as an exhaustive list of trustworthy mHealth resources. Taking another approach, the Canadian Medical Association (CMA) in 2015 released its "Guiding Principles for Physicians Recommending Mobile Health Applications to Patients" [46,47]. Among other things, this CMA policy notes that mobile apps may permissibly be used to complement, but should never replace physician care. It further recommends that physicians "prescribe" apps to patients with the primary objective of enhancing the safety or effectiveness of patient care [47]. Notably, neither the NHS App Library nor the CMA policy outlines specific rules for symptom checker apps. In the case of the NHS App Library, reviewed apps are organized into 16 categories, none of which refer precisely to the symptom checker designation [45]. Using the App Library search function and the fixed term "symptom checker" returns only 1 result, My Possible Self: The Mental Health App, a self-help app for managing fear, anxiety, and stress [45].

While the nonexistence of regulation and guidance for the clinical use of mHealth apps is a generally pervasive problem [44], it may be that its effects are especially pronounced in the context of apps powered by AI. There are at least 2 reasons for this. First, AISympCheck apps might be significantly more powerful than existing consumer health technologies. Apps using artificially intelligent processing techniques, for example, may be able to deliver diagnostic information with unprecedented levels of accuracy and specificity relative to other mass-market digital tools. This may promote high levels of public adoption while also increasing the clinical use of these systems. But even potentially accurate systems will likely sometimes produce errors, an eventuality that could be heightened by the absence of regulation of AISympCheck apps. The possibility that symptom checker apps will provide

XSL•FO
RenderX

inaccurate information poses a clear potential risk to users. App-mediated misdiagnosis might also cause confusion in medical practice. Care providers engaged in patient triage, for example, might have a difficult time accurately assessing patients presenting in the clinic with erroneous diagnostic information. Second, many AISympCheck apps are likely to be inscrutable in the sense, described above, that programmers and reviewers will be unable to understand how these systems technically operate. Physicians using or interacting with inscrutable decision systems may be unable to satisfactorily explain diagnostic outputs to their patients. Technical inscrutability may likewise impede physicians and regulators from systematically differentiating between high-quality, reliable apps and low-quality, unreliable apps. Inscrutability can also be expected to raise difficult questions about who, ultimately, is responsible should AI-powered tools cause injury. Complex and intersecting systems of medical malpractice, professional responsibility, and manufacturer liability may be ready to address injury caused by AISympCheck apps.

Addressing Entrenched Bias and Professional Responsibility

We contend that AISympCheck apps may work to entrench existing bias and greatly complicate the conventional approach to professional responsibility in the context of novel health technologies. These and related lines of concern are likely to attract the attention of scholars, policy makers, and activists. To be sure, these are far from the only legal and ethical concerns raised by the increasing use of AISympCheck apps. These tools also raise important questions about privacy, data processing, commercialization, premarket review, and the privatization of health care. Indeed, several authors have highlighted these issues, outlining challenges associated with an absence of clarity surrounding data custodianship when users interact with AISympCheck apps [27]. All of this is made more confounding by the tendency of AISympCheck apps to enter the market absent any prior, independent regulatory review. As we outlined above, these systems are effectively unregulated in Canada and elsewhere.

Entrenched bias and professional responsibility are distinct kinds of problems. Though they operate in different ways, both are exacerbated by the unregulated nature of mHealth. There are no clear rules about whether or how AISympCheck apps should be used. This confounds and complicates the question of professional responsibility in a fairly direct way, but so too does it pose additional problems for entrenched bias. As we outlined above, the problem of entrenched bias in AISympCheck decision-making is in large part a problem associated with the quality of data on which these systems are trained. Conceived as a challenge that operates as a function of the characteristics of training data, 3 general kinds of solutions might be helpful.

First, technical safeguards embedded into AISympCheck app programming could operate to mitigate or correct biased outcomes. But technical safeguards alone might be an insufficient remedy, for programmers will need to make assessments about what constitutes bias in the relevant circumstances. While not a problem in itself, often the most pernicious forms of discrimination may be those not initially recognized. Bias may operate on a complex web of interrelated factors; it might not be immediately obvious, for example, that an apparently innocuous decision has a discriminatory effect. These forms of bias could be overlooked in a merely programmatic response.

Second, AISympCheck systems could be trained principally on unbiased or neutral data. In response to the garbage in, garbage out phenomenon, it may be tempting to stress that we simply should not feed garbage into the system. Of course, this would massively oversimplify the entrenched bias challenge, while also again prompting programmers to evaluate what does or does not constitute a biased data set. This does not solve so much as punt the problem. One reaction to the garbage in, garbage out phenomenon could be that it is not a problem unique to AI systems but is a prevailing concern across medicine. We agree but would stress that the unexplainable character of many of our best AI tools makes the phenomenon potentially more pernicious than in other settings. If we do not know how a system processes data to arrive at a conclusion, it may be especially difficult to measure when the data in issue is "garbage."

Third, AISympCheck apps could operate according to a system of regulation and guidance that empowers users to make informed decisions about which apps are safe and reliable, that delineates the responsibilities of clinicians and others using AISympCheck apps in their practice, and that reviews individual apps periodically to ensure their proper function.

While this line of solution does not directly address the core set of issues in biased AI-mediated decision-making, it may be the best medium-term option for reducing discriminatory outcomes in the AISympCheck app context. It has the additional advantage of also working to address the second kind of challenge we identified above, namely that there persists a great deal of uncertainty about professional responsibility for the use of AISympCheck systems. In addressing both challenges, the most urgent unmet need is the absence of regulation or guidance specifically addressing AISympCheck apps. Questions about both bias and responsibility can be expected to fester in the absence of a framework that permits app users to evaluate and make decisions about how AISympCheck apps ought to be used. A reasonable place to begin, particularly in the Canadian context, is for professional medical associations to develop guidance for their members on the use of novel mHealth apps, especially those of the AISympCheck variety. Professional medical associations have played a long and essential role in working to structure the practice of medicine, guide clinicians in the performance of their obligations, and protect patients from risk [46]. In the Canadian context, provincial medical colleges that oversee the professional certification and clinical practice of physicians may have a unique role to play in guiding the use of AISympCheck apps. Medical professionals are, after all, primary reservoirs of expertise on the clinical use of novel health technologies and on the ways such technologies might affect physician-patient relationships. We, therefore, suggest that the dual challenges of entrenched bias and professional responsibility could begin to be addressed through medical college guidance that minimally (1) outlines factors that would assist in determining which apps are safe, reliable, and as best

XSL•FO
RenderX

as possible avoid discriminatory outcomes; (2) provides guidance for physicians on whether and how to recommend AISympCheck apps to their patients; and (3) considers how physicians should manage information provided by an AISympCheck app when a patient presents with a mobile-mediated diagnosis.

Conclusions

This viewpoint introduces the concept of the AISympCheck app, an mHealth app powered by AI and designed to provide users with information on a broadly diagnostic character. We outline 2 legal and policy challenges to which AISympCheck apps are likely to be especially susceptible—entrenched bias and professional responsibility. These issues underscore the critical importance of addressing the regulatory and guidance lacuna existing in this space, as well as the necessity of continued research that monitors the quality, safety, and efficacy of AISympCheck systems. We suggest in this viewpoint that AISympCheck apps will likely have increasing clinical implications and that medical colleges could play a central role in developing guardrails for their use. These guardrails will not only work to provide clarity to physicians and health systems using AISympCheck apps in clinical practice but to the app developers as well. To be sure, the regulation of clinical AI cannot by itself remedy all of the challenges described in his viewpoint. Problems surrounding biased training, for example, have deep structural and sociological roots that likely cannot be solely addressed in AI regulation. We nevertheless propose that forward-looking regulation and guidance in this space will attenuate many of the more serious risks, will help to facilitate the lawful design of these potentially disruptive tools, and ensure that their use serves the interests of patients and the public.

Acknowledgments

MHZ acknowledges the generous support of the Fonds de recherche du Québec—Santé, Junior 1 Research Scholar programme.

Data Availability

Data sharing is not applicable to this paper as no data sets were generated or analyzed during this study.

Conflicts of Interest

None declared.

References

1. Wattanapisit A, Teo CH, Wattanapisit S, Teoh E, Woo WJ, Ng CJ. Can mobile health apps replace GPs? A scoping review of comparisons between mobile apps and GP tasks. BMC Med Inform Decis Mak 2020 Jan 06;20(1):5 [FREE Full text] [doi: 10.1186/s12911-019-1016-4] [Medline: 31906985]
2. Fortuna KL, Aschbrenner KA, Lohman MC, Brooks J, Salzer M, Walker R, et al. Smartphone ownership, use, and willingness to use smartphones to provide peer-delivered services: results from a national online survey. Psychiatr Q 2018 Dec;89(4):947-956 [FREE Full text] [doi: 10.1007/s11126-018-9592-5] [Medline: 30056476]
3. Baxter C, Carroll J, Keogh B, Vandelanotte C. Assessment of mobile health apps using built-in smartphone sensors for diagnosis and treatment: systematic survey of apps listed in international curated health app libraries. JMIR Mhealth Uhealth 2020 Feb 03;8(2):e16741 [FREE Full text] [doi: 10.2196/16741] [Medline: 32012102]
4. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. J Med Internet Res 2020 Jan 30;22(1):e14679 [FREE Full text] [doi: 10.2196/14679] [Medline: 32012052]
5. Salisbury H. Prestidigitation. BMJ 2020 Feb 25;368:m648. [doi: 10.1136/bmj.m648] [Medline: 32098837]
6. McCartney M. AI in medicine must be rigorously tested. BMJ 2018 Apr 24;361:k1752. [doi: 10.1136/bmj.k1752] [Medline: 29691220]
7. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378. [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
8. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18:106820. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
9. Shah N, Martin G, Archer S, Arora S, King D, Darzi A. Exploring mobile working in healthcare: clinical perspectives on transitioning to a mobile first culture of work. Int J Med Inform 2019 May;125:96-101. [doi: 10.1016/j.ijmedinf.2019.03.003] [Medline: 30914187]
10. Reid AJ. The Smartphone Paradox: Our Ruinous Dependency in the Device Age. Cham, Switzerland: Palgrave Publishing; 2018.
11. Olson JA, Sandra DA, Colucci S, Al Bikaii A, Chmoulevitch D, Nahas J, et al. Smartphone addiction is increasing across the world: a meta-analysis of 24 countries. In: Computers in Human Behavior. Netherlands: ELSVIER; 2022:107138.
12. Statista. Number of smartphone users worldwide from 2016 to 2023. Statista. URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/ [accessed 2024-05-15]

XSL•FO
RenderX

13. Davies A, Mueller J. Introduction to mHealth. In: Developing Medical Apps and mHealth Interventions. Switzerland: Springe; 2020:24.
14. Higgins JP. Smartphone applications for patients' health and fitness. Am J Med 2016 Jan;129(1):11-19. [doi: 10.1016/j.amjmed.2015.05.038] [Medline: 26091764]
15. Coppetti T, Brauchlin A, Müggler S, Attinger-Toller A, Templin C, Schönrath F, et al. Accuracy of smartphone apps for heart rate measurement. Eur J Prev Cardiol 2017 Aug;24(12):1287-1293 [FREE Full text] [doi: 10.1177/2047487317702044] [Medline: 28464700]
16. Byambasuren O, Beller E, Glasziou P. Current knowledge and adoption of mobile health apps among Australian general practitioners: survey study. JMIR Mhealth Uhealth 2019 Jun 03;7(6):e13199 [FREE Full text] [doi: 10.2196/13199] [Medline: 31199343]
17. Murnane EL, Huffaker D, Kossinets G. Mobile health apps: adoption, adherence, and abandonment. 2015 Presented at: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers; 7 September 2015; United States. [doi: 10.1145/2800835.2800943]
18. Lupton D, Jutel A. 'It's like having a physician in your pocket!' A critical analysis of self-diagnosis smartphone apps. Soc Sci Med 2015 May;133:128-135. [doi: 10.1016/j.socscimed.2015.04.004] [Medline: 25864149]
19. Lang M, Zawati MH. The app will see you now: mobile health, diagnosis, and the practice of medicine in Quebec and Ontario. J Law Biosci 2018 May;5(1):142-173 [FREE Full text] [doi: 10.1093/jlb/lsy004] [Medline: 29707219]
20. Sullivan H. NHS symptom checker app. Nurs Stand 2013 Oct 23;28(8):33. [doi: 10.7748/ns2013.10.28.8.33.s40] [Medline: 26982424]
21. WebMD app. Apple Apple Store. 2023. URL: https://apps.apple.com/us/app/webmd-trusted-health-and-wellness-information/id295076329 [accessed 2023-11-21]
22. We are eMed UK. eMed. URL: https://www.babylonhealth.com/about [accessed 2023-11-21]
23. Babylon. Learn more about our AI. Babylon. 2023. URL: https://www.babylonhealth.com/ai/learn-more [accessed 2023-11-21]
24. Copestake J. Babylon claims its chatbot beats GPs at medical exam. BBC. 2018. URL: https://www.bbc.com/news/technology-44635134 [accessed 2024-05-10]
25. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018 Nov 24;392(10161):2263-2264. [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]
26. Iacobucci G. London GP clinic sees big jump in patient registrations after Babylon app launch. BMJ 2017 Dec 21;359:j5908. [doi: 10.1136/bmj.j5908] [Medline: 29269519]
27. Hardcastle L, Ogbogu U. Virtual care: enhancing access or harming care? Healthc Manage Forum 2020 Nov;33(6):288-292 [FREE Full text] [doi: 10.1177/0840470420938818] [Medline: 32686506]
28. Hardcastle L, Ogbogu U. Opinion: Alberta's virtual health-care app plagued with problems. Edmonton Journal. URL: https://edmontonjournal.com/opinion/columnists/opinion-albertas-virtual-health-care-app-plagued-with-problems [accessed 2023-11-21]
29. Browne R. Saudi Arabia backs $550 million investment in health start-up Babylon, valuing it at $2 billion. CNBC. 2019. URL: https://www.cnbc.com/2019/08/02/saudi-arabia-backs-550-million-investment-in-health-startup-babylon.html [accessed 2023-11-21]
30. Hi, we're Ada. Ada. URL: https://ada.com/about/ [accessed 2023-11-21]
31. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open 2020 Dec 16;10(12):e040269 [FREE Full text] [doi: 10.1136/bmjopen-2020-040269] [Medline: 33328258]
32. Torous J, Nebeker C. Navigating ethics in the digital age: introducing Connected and Open Research Ethics (CORE), a tool for researchers and institutional review boards. J Med Internet Res 2017 Feb 08;19(2):e38 [FREE Full text] [doi: 10.2196/jmir.6793] [Medline: 28179216]
33. Hirsch MC, Ronicke S, Krusche M, Wagner AD. Rare diseases 2030: how augmented AI will support diagnosis and treatment of rare diseases in the future. Ann Rheum Dis 2020 Jun;79(6):740-743 [FREE Full text] [doi: 10.1136/annrheumdis-2020-217125] [Medline: 32209541]
34. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics 2017 Mar 01;18(1):19 [FREE Full text] [doi: 10.1186/s12910-017-0179-8] [Medline: 28249596]
35. Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, et al. Social determinants of health and diabetes: a scientific review. Diabetes Care 2020 Nov 02;44(1):258-279 [FREE Full text] [doi: 10.2337/dci20-0053] [Medline: 33139407]
36. Karmakar M, Lantz PM, Tipirneni R. Association of social and demographic factors with COVID-19 incidence and death rates in the US. JAMA Netw Open 2021 Jan 04;4(1):e2036462 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.36462] [Medline: 33512520]
37. Ahmed T, Rizvi SJR, Rasheed S, Iqbal M, Bhuiya A, Standing H, et al. Digital health and inequalities in access to health services in Bangladesh: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e16473 [FREE Full text] [doi: 10.2196/16473] [Medline: 32706736]

XSL•FO
RenderX

38. Castelvecchi D. Can we open the black box of AI? Nature 2016 Oct 06;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
40. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
41. Castelvecchi D. Is facial recognition too biased to be let loose? Nature 2020 Nov;587(7834):347-349. [doi: 10.1038/d41586-020-03186-4] [Medline: 33208976]
42. Mobile fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2023-11-21]
43. Health Canada. Guidance Document: Software as a Medical Device (SaMD): Definition and Classification. Canada: Health Canada; 2019.
44. Magrabi F, Habli I, Sujan M, Wong D, Thimbleby H, Baker M, et al. Why is it so difficult to govern mobile apps in healthcare? BMJ Health Care Inform 2019 Nov;26(1):e100006 [FREE Full text] [doi: 10.1136/bmjhci-2019-100006] [Medline: 31744843]
45. National Health Service. NHS apps library. NHS. 2019. URL: https://www.mobihealthnews.com/news/emea/more-100-digital-health-and-care-tools-evaluated-against-key-standards-nhs-apps-library [accessed 2023-11-21]
46. Zawati MH, Lang M. Mind the app: considerations for the future of mobile health in Canada. JMIR Mhealth Uhealth 2019 Nov 04;7(11):e15301 [FREE Full text] [doi: 10.2196/15301] [Medline: 31682580]
47. Canadian Medical Association. Guiding principles for physicians recommending mobile health applications to patients. CMA. 2019. URL: https://policybase.cma.ca/en/permalink/policy11521 [accessed 2021-11-21]

Abbreviations

AI: artificial intelligence
AISympCheck: AI-powered symptom checker
CMA: Canadian Medical Association
mHealth: mobile health
NHS: National Health Service

Edited by T de Azevedo Cardoso; submitted 27.06.23; peer-reviewed by D Pougheon Bertrand, A Gorny; comments to author 06.09.23; revised version received 01.12.23; accepted 03.04.24; published 05.06.24.

Please cite as:
Zawati MH, Lang M
Does an App a Day Keep the Doctor Away? AI Symptom Checker Applications, Entrenched Bias, and Professional Responsibility
J Med Internet Res 2024;26:e50344
URL: https://www.jmir.org/2024/1/e50344
doi:10.2196/50344
PMID:38838309

©Ma'n H Zawati, Michael Lang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Trust but Verify: Lessons Learned for the Application of AI to Case-Based Clinical Decision-Making From Postmarketing Drug Safety Assessment at the US Food and Drug Administration

Robert Ball[1], MPH, ScM, MD; Andrew H Talal[2], MPH, MD; Oanh Dang[1], PharmD; Monica Muñoz[1], PharmD, PhD; Marianthi Markatou[3], PhD

[1]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

[2]Jacobs School of Medicine and Biomedical Sciences, Buffalo, NY, United States

[3]School of Public Health and Health Professions, University at Buffalo, Buffalo, NY, United States

Corresponding Author:
Robert Ball, MPH, ScM, MD
Office of Surveillance and Epidemiology
Center for Drug Evaluation and Research
US Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD, 20993
United States
Phone: 1 301 796 2380
Fax: 1 301 796 9832
Email: robert.ball@fda.hhs.gov

Abstract

Adverse drug reactions are a common cause of morbidity in health care. The US Food and Drug Administration (FDA) evaluates individual case safety reports of adverse events (AEs) after submission to the FDA Adverse Event Reporting System as part of its surveillance activities. Over the past decade, the FDA has explored the application of artificial intelligence (AI) to evaluate these reports to improve the efficiency and scientific rigor of the process. However, a gap remains between AI algorithm development and deployment. This viewpoint aims to describe the lessons learned from our experience and research needed to address both general issues in case-based reasoning using AI and specific needs for individual case safety report assessment. Beginning with the recognition that the trustworthiness of the AI algorithm is the main determinant of its acceptance by human experts, we apply the Diffusion of Innovations theory to help explain why certain algorithms for evaluating AEs at the FDA were accepted by safety reviewers and others were not. This analysis reveals that the process by which clinicians decide from case reports whether a drug is likely to cause an AE is not well defined beyond general principles. This makes the development of high performing, transparent, and explainable AI algorithms challenging, leading to a lack of trust by the safety reviewers. Even accounting for the introduction of large language models, the pharmacovigilance community needs an improved understanding of causal inference and of the cognitive framework for determining the causal relationship between a drug and an AE. We describe specific future research directions that underpin facilitating implementation and trust in AI for drug safety applications, including improved methods for measuring and controlling of algorithmic uncertainty, computational reproducibility, and clear articulation of a cognitive framework for causal inference in case-based reasoning.

(J Med Internet Res 2024;26:e50274) doi:10.2196/50274

KEYWORDS

drug safety; artificial intelligence; machine learning; natural language processing; causal inference; case-based reasoning; clinical decision support

XSL•FO
RenderX

Introduction

A very common task accomplished by medical professionals many times a day is reasoning about an individual case to make medical decisions. Supporting case-based reasoning with automation for most clinical situations remains a challenge despite advances in artificial intelligence (AI) [1]. To be most successful, predictive "AI" (eg, machine learning [ML]) requires large amounts of annotated data and will, therefore, perform best in large-scale clinical situations where such annotation is possible (eg, image interpretation). However, most of clinical medicine (1) consists of situations that, while common, are very complex, making the identification of all the necessary predictive features difficult; (2) is small-scale, where there is incomplete understanding about what the important features are; and (3) can involve genuine uncertainty in the application of consensus guidelines to an individual patient. Developing strategies for how best to apply AI to these situations is essential to fulfill the potential of AI to assist with patient care.

Case-based reasoning about a drug's risks is a key component of the assessment of postmarketing individual case safety reports (ICSRs) at the Food and Drug Administration (FDA) [2]. An ICSR contains a description of an adverse drug experience related to an individual patient prepared in a standard format for submission to the FDA and other regulators. Information about the risks of a drug tend to increase after the drug is approved and used by large numbers of patients. While some of this information comes from postmarketing trials and observational studies required by the FDA and observational studies conducted in the FDA's Sentinel System [3], a critical source of information is case reports of adverse events (AEs) from individual patients and their providers [4-7]. The process of assessing these ICSRs is akin to a clinician's task of considering multiple possibilities to assign a diagnosis and includes the identification of key features of the drug, the clinical events, the temporal relationship between drug exposure and the clinical events, and demographic features of the patients. These factors must be interpreted against a background of past medical history, other medications and exposures, and the natural history of the disease being treated. The goal of this process is a decision as to whether it is likely the drug in question caused the observed clinical events.

To make this process more efficient and improve its scientific rigor, the FDA embarked on a program to develop automation support for ICSR assessment [8]. In this viewpoint, our aim is to describe the lessons learned from this experience and additional research needed to address both general issues in case-based reasoning using AI and specific needs for ICSR assessment. We first describe the FDA's experience applying natural language processing (NLP) and ML to ICSR assessment and the recognition that, for automation support to be successful, it must have the full trust of those it is supporting. Then, we apply the Diffusion of Innovations theory to the FDA's experience to illuminate the sociotechnical reasons for FDA safety reviewers' acceptance of one AI algorithm (ie, the process of deduplication) but not another (ie, assignment of causal relationships) [9]. This analysis leads to the recognition of the importance of a formal inferential framework for ICSR assessment. We conclude with a discussion of the need for a deeper understanding and potential reframing of the cognitive framework used for causal inference and research priorities for AI to be fully applied to case-based reasoning and clinical drug safety assessment.

FDA's Experience in Applying NLP and ML to ICSR Assessment

Analyzing ICSR data is challenging because of the limitations of these data, including the underreporting of AEs and the lack of accurate data on drug use preventing the calculation of accurate AE occurrence rates; the lack of controls; as well as data limitations within the reports themselves, including missing, imprecise, or occasionally inaccurate clinical information. These and other drawbacks limit the usability of current methods to draw a causal link between a reported drug and the AE based on statistical properties of their occurrence in an AE database alone.

The assessment of ICSRs for possible causality still relies primarily on expert judgment and global introspection. Figure 1 provides an overview of the ICSR evaluation workflow (Figure 1 is adapted from the study by Ball and Dal Pan [8]. Case definitions are a set of prespecified criteria for determining whether a patient should be identified as having a particular disease, injury, or other health condition (ie, AE) [2]). Current practice for drawing inferences from these data involves expert review of case series and comparison with external sources of information (eg, product labeling describing known AEs and pharmacological mechanisms) supplemented with summary statistics and disproportionality scores [2,10,11]. Important clinical information (eg, temporality, concomitant medications, comorbidities or past medical history, or alternative explanations for the AE) is typically found in the ICSR narrative, which is generally believed to be the key to making an accurate assessment, provided that the narrative contains the relevant information. This means that the application of AI techniques must focus on methods to extract and organize meaningful information from these narratives.

The initial development of this approach at the FDA was conducted using reports to the FDA and Centers for Disease Control and Prevention's Vaccine Adverse Event Reporting System focusing on automating the identification of 2 rare, but important, vaccine AEs, anaphylaxis and Guillain-Barre syndrome, as test cases [12-14]. The first step was to develop an NLP approach to select and extract clinical manifestations and demographic characteristics of the patients for the clinical condition of interest. Subsequent work involved further development of the text mining system to extract drug or biologic product exposure, temporality, and alternative explanations for the AE [15-17]. Finding duplicate reports of the same case (ie, if the same AE instance for the same individual is reported by multiple reporters) is an important practical step in the assessment of ICSRs. While algorithms based on structured fields in the ICSRs have been available, an important advance occurred with the incorporation of clinical text extracted using NLP from the case narratives into the algorithm [18-20]. In these examples, the features were extracted from narratives primarily using rule-based approaches.

XSL•FO
RenderX

A complementary approach to the aforementioned one is to take a holistic view of the causality assessment process and allow the machine to identify the key features of a report. Using relatively small subsets of ICSRs from the FDA Adverse Event Reporting System classified on a 5-point scale as to the likelihood that the drug caused the AE, supervised ML was used to predict these classifications [21,22]. Another approach used logistic regression to predict the likelihood that an ICSR contained useful information by its inclusion in an FDA case series review supporting a recommendation to modify product labeling [23]. These efforts demonstrate the feasibility of developing models to predict which ICSRs most likely contain information for a human safety reviewer's triage assessment of the report. However, the algorithm's performance is not sufficient to allow for use without human expert review of the predictions.

The relative success of these algorithms led to their incorporation into decision support tools [24,25]. The experience using these tools in operational pilots is described in the following section.

Figure 1. Individual case safety report (ICSR) evaluation workflow. This figure illustrates the ICSR evaluation workflow from the retrieval of ICSRs to the creation of a case series that has been deduplicated and assessed for causality. The assessment for causality uses case-based reasoning and considers information from the ICSR about the drug or biologic product, adverse event, temporal relationship, and other etiologies for the adverse event. The output from the ICSR causality assessment process is a case series of ICSRs that have been evaluated for causal association.

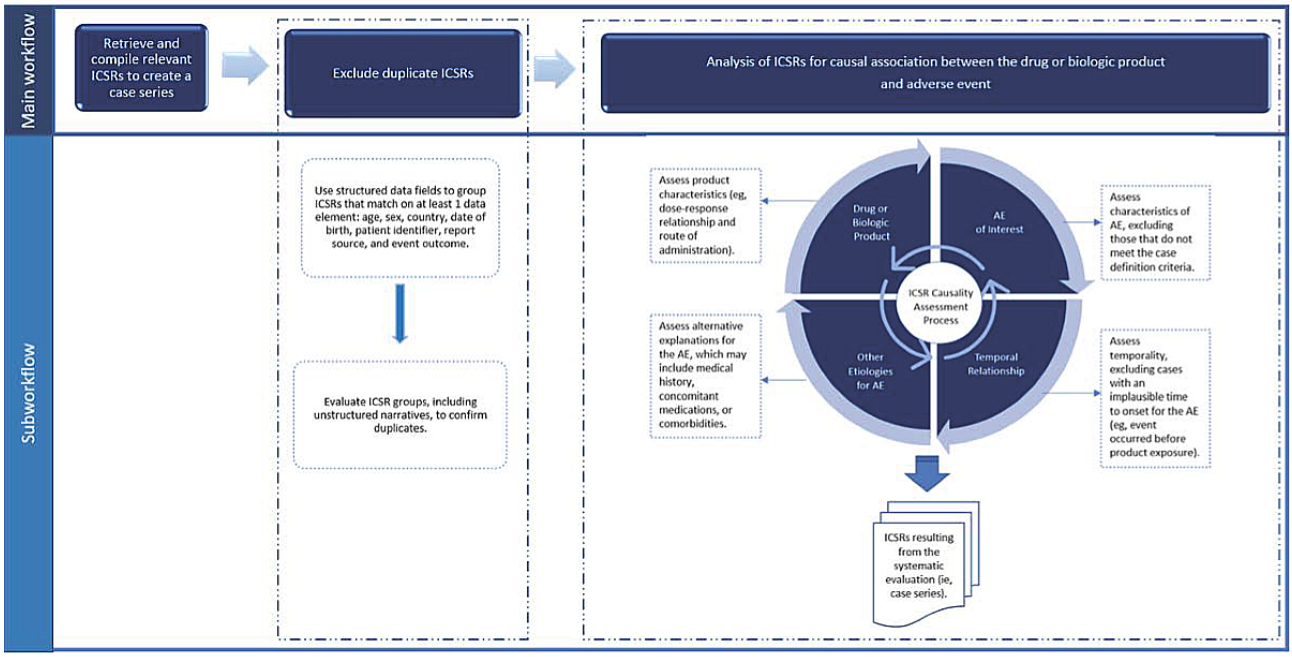

The Importance of Trust in Building AI Support for ICSR Assessment

User trust in AI can be furthered through close collaboration among end users and AI developers in an iterative process [26]. Continuous communication between these stakeholders is needed to understand pharmacovigilance problems and uncover challenges as experienced by safety reviewers, the selection of use cases for AI, the development of required functionalities to support the use case, and appropriate insertion of AI in the pharmacovigilance workflow. Safety reviewers need to have confidence that AI developers sufficiently understand the complexity of the pharmacovigilance problem to propose whether and which AI methods can be optimally applied to address the specific problem. The ICSR evaluation workflow is complicated by many and varied tasks; the consideration of a multitude of features that require interpretation for decision-making; and the application of clinical domain expertise in the context of the use case, such as duplicate detection, clinical classification, or causality assessment. A shared understanding of the intricacies involved in ICSR evaluation supports safety reviewers and AI developers working together to define appropriate use cases and the approach for integration in the pharmacovigilance workflow.

Humans have rich prior knowledge. The engagement of the safety reviewer with the computer algorithm can facilitate addressing the issue of incomplete features and data sparseness, especially when there is a lack of training data, as is the situation with many ICSR assessments and clinical decision-making. The concept of "human-in-the-loop" has been proposed to tackle these challenges, by incorporating human knowledge into the process and addressing the AI algorithm's limitations.

Acceptability of NLP and ML Algorithms by Safety Reviewers and the Diffusion of Innovations Theory

Despite what would appear to be adequate performance to provide an aid for improving ICSR review, the software tools initially were met with limited acceptance by safety reviewers, although each development cycle included their active involvement in the collection of requirements and testing of user interfaces.

Discussion with safety reviewers about AI for causality assessment was met with skepticism because their perception is that this problem is challenging for humans, let alone

XSL•FO
RenderX

machines. In the previously mentioned FDA study [21] classifying reports on a 5-point scale as to the likelihood that the drug caused the AE, at least 2 of the 3 adjudicators tended to agree as to a report's classification. When the reports were regrouped to create a binary classification (ie, Certain, Probable, or Possible vs Unlikely or Unassessable), all 3 adjudicators agreed on 61% of the reports [21].

In contrast to the causality assessment use case, duplicate ICSR detection was perceived by the safety reviewers as a high priority and a potentially solvable problem for AI. Duplicate ICSRs contribute to inefficiencies by increasing workload because they are generally identified via manual processes but can be relatively easily verified if presented by an algorithm. Trust-building efforts, aligning with the safety reviewers' workflow and communication, were applied to explicitly map clinical features used in the safety reviewers' current manual deduplication processes to that of the AI-based algorithm. In addition, the NLP-extracted common clinical features and their effects on the duplicates identified by the algorithm were further visually presented to enhance end users' understanding of the features that were used by the algorithm to identify duplicates.

Feedback from safety reviewers [27] suggested several reasons for their initial limited acceptance for causality assessment when performed using AI. First, the AI approach is outside of their domain of expertise and clinical training. Furthermore, the approach did not offer sufficient explanation or flexibility to accommodate information external to the ICSR (eg, reviewer's knowledge). The software tools were not fully aligned with the reviewer's current business practices and workflows. Because of the lack of trust in the software tools, any efficiency gains, compared to current business practices and workflows, were not perceived as making it easier for the safety reviewers to assess whether the drug caused the outcome described in the report.

Ethical Considerations

FDA employees were asked to provide feedback about their experience with the deduplication algorithm as part of a project conducted in collaboration with Johns Hopkins University under the FDA Centers of Excellence in Regulatory Science and Innovation program. The Johns Hopkins Medicine Institutional Review Board determined that the project did not constitute human subjects research under the US Department Health and Human Services or FDA regulations. As this was not considered human subjects research, informed consent was not obtained. All participation in surveys was voluntary and was part of routine employment at the FDA; hence, participants did not receive additional compensation. Survey participants provided consent to use their anonymized responses in the manuscript.

Diffusion of Innovations Theory

Diffusion of Innovations is a well-known theory that has been applied to explain how new ideas and technologies are adopted [9]. According to the theory, several attributes influence the rate of spread of an innovation and can offer insight into the FDA safety reviewers' acceptance of AI algorithm use (Table 1).

In Table 1, the acceptability of the application of AI algorithms to ICSR deduplication is assessed through the main attributes, including the social system, the communication channels used for dissemination, and the attributes of the innovation. This example illustrates how the AI algorithm for deduplication successfully incorporated the key attributes of an innovation, namely, relative advantage, trialability, and observability. The successful fulfillment of these attributes promoted trust in its use. The key to its acceptance is that the tool transparently implemented an algorithm that closely paralleled existing pharmacovigilance workflows and safety reviewers' cognitive process for deduplication.

In contrast, 2 attributes of innovation, compatibility and complexity, offer a possible explanation for the lack of acceptance of the causality assessment algorithm. Causality assessment is much more complex than duplicate detection, as illustrated by the larger number of attributes required for consideration and their interactions, listed on the right side of Figure 1. The determination of the likelihood that a drug caused an AE in a report is considered the most challenging aspect of a safety reviewer's work. Not only are the drug exposure and the AE outcomes important considerations, but alternative explanations of the phenomenon and temporality must also be considered. The social environment in which safety reviewers work encourages complete and transparent characterization of all evaluated ICSRs following a deterministic logic. Using probabilistic predictions in such an environment requires either perfectly performing algorithms or a validation process that limits the benefits of the algorithms. While the same issue arises in duplicate detection, the validation process is simpler, and the consequences of incorrect duplicate classification are lower than those for causality assessment. The existing technical environment in which safety reviewers work offers no other applications of computing to address a problem of similar complexity. In fact, many simpler problems remain unautomated, for example, automatically flagging an AE as being present in the drug's package insert, leading to the perception that it is unlikely that a machine could successfully help solve the more complex problems.

Furthermore, 2 other "diffusion of innovations" attributes, the ability to readily observe an intervention's functionality (ie, observability) and to readily experiment with it (ie, trialability), are also likely important. The relative difficulty of observing the process by which an ML algorithm assigns a causality assessment may complicate safety reviewer acceptance. While the general parameters of the approach to causal inference in ICSR evaluation have been described [2,10,11], no complete articulation of all the data elements and their interrelationships in the form of an algorithm has been accomplished. This lack of a basis for verification of an algorithm's validity suggests that developing a more complete understanding of how inferences are made in case-based reasoning is an important next step. To increase acceptance of ML approaches to causality assessment, more work is also needed on how to best fit causality assessment algorithms into safety reviewer workflows to allow them to try out the results of the algorithms without total commitment.

XSL•FO
RenderX

Table 1. Diffusion of Innovations theory applied to the Food and Drug Administration's AI[a] algorithms.

Attribute (definition)	Deduplication		Causality assessment	
	Relevance	Evidence	Relevance	Evidence
Social system (interconnected units working collaboratively toward common goal)	Duplicate ICSRs[b] are viewed as a common problem by safety reviewers during their case series evaluation.	Participants viewed themselves as a community of safety reviewers in which their key role involves evaluating ICSRs for drug-related safety issues. During this process, they experienced the challenges of efficiently identifying duplicate ICSRs for exclusion from the case series (Figure 1).	Assessing causality during ICSR evaluation involves evaluating the likelihood and strength of the relationship between a drug and an adverse event at a report level. Causality assessment is a critical component of safety signal management.	Consistent with established practices for safety surveillance, safety reviewers assessed ICSRs before assembling a case series that includes cases assessed as causally associated. Reviewers documented a summary of the considerations or rationale for inclusion of the ICSRs in the case series [2].
Communication channel (method of information spread; users' ability to perceive usefulness)	Safety reviewers were involved in iterative requirements gathering processes. They provided continuous input during the testing and evaluation of the deduplication tool.	Verbal and written communication provided an explanatory description of the deduplication algorithm tool and where it could fit into the safety reviewers' workflow. Multiple rounds of feedback were collected during the testing and evaluation of the deduplication algorithm tool.	As part of a series of research and development efforts to implement a tool to support case series evaluation that include causality assessment, interactive meetings with multiple groups of safety reviewers were held to understand current practices and workflow for conducting causality assessments.	Discussions with safety reviewers revealed concerns about the effectiveness and utility of a one-size-fits all algorithm that classifies ICSRs by level of causality. Safety reviewers viewed causality assessments as a complex task for both humans and AI. As a result, they did not prioritize causality assessment for incorporation in a tool to support case series evaluation.
Attributes of innovation: relative advantage (perception of benefit or improvement over existing technology)	Safety reviewers compared the usefulness and efficiency gained from the automated deduplication algorithm output against that of the current baseline, which is to manually use spreadsheets to find duplicate ICSRs.	"I think the deduplication method could be used instead of the current process, as the algorithm did find some cases that would have been missed. However, the cases still need to be screened to determine if the cases are duplicates or not." "Really a big help to save time." "It is helpful that the algorithm grouped likely matches together; this saved some time."	The existing approach for conducting causality assessments is a complicated manual process that involves many steps. In particular, it requires clinical and pharmacovigilance expertise.	Safety reviewers did not view the automation of drug causality assessment as a relative advantage over the current manual process. Rather, there was skepticism around whether and how well AI could emulate human experts' thinking in terms of applying clinical knowledge and judgment to accurately conduct a causality assessment of ICSRs.
Attributes of innovation: compatibility (consistent with existing technical and social environment)	The deduplication tool is consistent with and supports the current technical, business process, and safety reviewers' workflows.	"Usability was very straightforward, easy to run." "Very quick turnaround time (minutes) allowed me to start working on it right away." "I will continue to use the algorithm." "This tool doesn't replace my deduplication process but is a helpful addition."	Causality assessments to detect and evaluate the relationship between a drug or biologic and adverse event of interest are inherently part of the safety reviewer's workflow and processes.	The use of AI for causality assessments is viewed by safety reviewers as a complicated task that is likely not yet solvable by current technology. Particularly, there are nuances and factors that need to be considered in various use cases to which causality assessments are applied. A question remains whether the AI output for causality assessment could be generalized to any drug and adverse event of interest. In addition, it is not clear how the AI output could be usefully incorporated in the current workflow to support safety reviewers.

XSL•FO
RenderX

Attribute (definition)	Deduplication		Causality assessment	
	Relevance	Evidence	Relevance	Evidence
Attributes of innovation: complexity (perception of the difficulty of implementation, use, or understanding)	During the manual process of detecting duplicate ICSRs, safety reviewers conducted a stepwise comparison of data points from structured fields, followed by those from the narratives between potential duplicate pairs to find actual duplicate reports during case series evaluation.	The deduplication tool considers multiple relevant features from ICSR structured fields and narratives, many of which overlap with those used in the safety reviewers' best practices deduplication processes. Most safety reviewers stated that they were very likely or likely to use the deduplication algorithm for their reviews and had medium to high confidence in the deduplication algorithm tool's output.	The processes involved in causality assessment require more advanced logical reasoning and considerations of interrelationships among various data in the structured and unstructured information from ICSRs. Furthermore, clinical and pharmacovigilance knowledge and expertise is applied during causality assessment (ie, external information that may not be represented within an ICSR).	Safety reviewers participated in successive research focused on applying AI for causality assessments. To train and test the ML[c] classification algorithm for causality, safety reviewers created a reference data set of annotated ICSRs based on the likelihood that the drug caused the adverse event. The ambiguity and complexity of annotating levels of causality was illustrated by the substantially lower interannotator agreement among 3 adjudicators compared to that for 2 adjudicators. The performance characteristics of the ML algorithms for causality were perceived as not being adequate. Moreover, safety reviewers do not have domain expertise in ML, which could affect the perceptions and understanding of the strengths and limitations of technology. When the ML algorithm used certain features not used by human safety reviewers, it was perceived as a limitation. In addition, the ML causality classification algorithm did not use external data sources, such as clinical knowledge of the medical history or concomitant medications, that safety reviewers routinely apply.
Attributes of innovation: trialability (ability to try without total commitment and with minimal investment)	A 6-month study allowed all safety reviewers to test and evaluate the usefulness of the deduplication algorithm within their current workflow and provide additional feedback.	"It's a good backup and second check to my own deduplication." "It's a good second check." "Although the algorithm doesn't replace my own deduplication, I find it helpful in combination with my process."	Safety reviewers were offered a plausible workflow option that would incorporate the output of the ML for causality assessment in their workflow process. The option was to use the ML algorithm's output to prioritize the review of ICSRs that were classified with the highest likelihood of a causal association, followed by those with lower likelihood of causality.	Despite the development of the ML algorithm for ICSR classification of causality, the proposed plausible workflow would still need considerable time and human resource investment. First, from the safety reviewers' perspective, the ML algorithm did not preclude the need to conduct causality assessments, the most resource intensive step, for all the ICSRs within a case series. Second, because ICSR prioritization is a not part of the current workflow, a new workflow that effectively integrates prioritization would need to be developed.

XSL•FO
RenderX

Attribute (definition)	Deduplication		Causality assessment	
	Relevance	Evidence	Relevance	Evidence
Attributes of innovation: observability (visible benefits to potential adopters)	Safety reviewers were able to experience the application of the deduplication algorithm tool to each of their specific case series of interest within the existing workflow. The benefits of the tool were apparent while using the tool.	"Although I don't think it saved any time in this data set, the algorithm identified duplicates I would have missed." "...good for screening large numbers of reports." "I still believe there is beneficial utility to the tool, possibly with searches producing high caseloads."	Safety reviewers were presented with the option to use only the final output from the AI algorithm for causality assessment that automates the classification of ICSRs as assessable or not assessable. Assessable ICSRs contain sufficient information for a safety reviewer to be able to conduct a causality assessment, whereas unassessable reports have insufficient information.	Safety reviewers neither completely trusted nor valued the benefit of solely using only the ML classification output of assessable or not assessable in their current workflow. Concerns were raised about the risk of missing an important ICSR because of misclassification by the ML algorithm. Rather than use the ML algorithm's final classification output, reviewers requested to view and understand components of the algorithm. The potential benefits and how to optimally incorporate this AI output in their workflow remain unclear and thus not observable.

[a]AI: artificial intelligence.

[b]ICSR: individual case safety report.

[c]ML: machine learning.

The Importance of a Formal Inferential Framework for Building Trust in AI Support for ICSR and Clinical Drug Safety Assessment

What are the key components of a plausible inferential framework for building trust in AI systems for drug safety causality assessment? At the heart of this framework is measuring and managing uncertainty. Additional components include understanding and mitigating the impact of data biases as well as understanding how ML algorithms for drug safety operate, their strengths and limitations, and how they can be applied and tuned for a given task. Furthermore, computational reproducibility is a key component for valid inference. Thus, in our context, the term signifies well-described and standardized workflows, computing environments, and the ability to obtain the same results if the same data and algorithm are used by 2 users following the same workflows. In the following paragraphs, we discuss these aspects in detail.

Traditionally, inference from clinical data has been based on a hierarchy of evidence, with randomized, blinded clinical trials considered the gold standard, while individual case reports, such as those discussed in this paper, are considered to have the least evidentiary value. More recently, there has been a movement away from this "hierarchy" toward a recognition that valid causal inferences can be made from a synthesis of different types of data, enabled by advanced computational techniques including AI [28,29]. Traditionally in pharmacovigilance, certain AEs, such as anaphylaxis, have been attributed to drug exposure after only a few case reports if the following two conditions are met: (1) the time between the drug exposure and the onset of the condition is relatively short and consistent with the known mechanism of action and (2) there are no other obvious causal factors. For anaphylaxis, most reactions occur within minutes to hours after exposure [30]. Progressive multifocal leukoencephalopathy (PML) was attributed to the treatment of patients with multiple sclerosis with natalizumab after only a few cases were observed based on the rareness of PML, the plausibility of natalizumab causing immunosuppression, and the presence of the infectious agent that causes PML [31]. What matters more than the traditional hierarchy is the proper application of the "rules of inference" to data that are "fit-for-purpose." In the context of drug safety, this means that better understanding of both elements is necessary to make advancements in creating a computable drug safety cognitive framework.

By a drug safety cognitive framework, we mean the rules of inference applied to fit-for-purpose data in ICSRs by safety reviewers to assess whether there is likely to be a causal relationship between an exposure and an AE. As outlined in Figure 1, the data categories included in the framework include the following:

- The drug exposure
- The timing between the exposure and the onset of the AE of interest
- Concomitant exposures, including other drugs and the timing between their exposure and the onset of the AE of interest
- The natural history of the disease and its relationship with the AE of interest
- Prior medical history including other conditions and their natural history and relationship with the AE of interest
- If available, information about dechallenge or rechallenge, that is, if the AE stops when the drug is discontinued and reoccurs when the drug is restarted, can support a conclusion that the likelihood of a causal relationship is increased

XSL•FO
RenderX

Factors external to the information contained in an ICSR include knowledge of the mechanism of action of the exposures and their relationship with the AE of interest, including whether the timing of exposure and onset of the AE of interest is mechanistically feasible, and knowledge of whether the AE is known to be caused by the exposures.

"Fit-for-purpose" data are data suitable to be used for pharmacovigilance and are characterized by the following properties:

- They are fit-for-purpose from a quality standpoint. Are the data statistically sound to be used for the purpose of identifying AEs?
- They are fit-for-purpose from a timing perspective. Are the data current enough to form the basis for the safety question of interest?
- They are fit-for purpose for regulatory action. Whatever is discovered from the data needs to be understood to take action.
- They are fit-for-purpose for representing the population affected by the safety issue. For example, do the data pertain to certain sex, race, age, ethnicity, or other important categorization of the population affected?

The key to understanding the relationships between "rules of inference" and "fit-for-purpose" data, as it relates to the use of AI, is how different approaches manage uncertainty. The pharmacovigilance setting is that of learning from data and, as such, is inseparably connected to uncertainty. Decision-making involves uncertainty. Some of the uncertainty concerns facts. For example, how long does it take to develop anaphylaxis after exposure to a drug that might initiate this event? Taking full advantage of the scientific research dictates knowing its associated uncertainty. Both extremes, that is, too much confidence or too little confidence, are problematic. The first extreme raises the possibility of facing unexpected problems, such as missing important AEs potentially resulting in increased morbidity and mortality. The second extreme raises the possibility of missing opportunities while wasting time and resources, such as unnecessarily conducting additional studies to address a safety issue. In the context of pharmacovigilance, substantially consequential decisions need to be taken with reference to the modification of the drug product labeling of certain medications or even removal of medicinal products from the market. Thus, improving the understanding of sources of uncertainty and their implications for the consequences of a decision is a fundamental need when using AI.

At a high level, there are 2 types of uncertainty: aleatoric or statistical uncertainty and epistemic uncertainty. Aleatoric uncertainty expresses the inherent randomness associated with an observed process, while epistemic uncertainty refers to the uncertainty associated with limited knowledge, which can be partially reduced by increasing knowledge. These uncertainties are of different nature, with aleatoric being a stochastic component in the data generation process and epistemic being associated with the state of our knowledge about a phenomenon of interest. An example of aleatoric uncertainty is coin flipping; there is a stochastic component to the data generation process that cannot be reduced by adding any type of information, while epistemic uncertainty is reducible as more information becomes available. Human performance in the evaluation of ICSRs is assumed to have the minimum amount of irreducible epistemic uncertainty, but in many situations, no baseline is available to inform the threshold for minimum AI performance characteristics. Senge et al [32] referred to the distinction between the 2 types of uncertainties and proposed a quantification that is illustrated in the context of medical decision-making. Another recent work [33] discussed these concepts in the context of ML. At this point, it is important to emphasize that other sources of uncertainty may exist, for example, measurement error, outliers, model uncertainty, or incorrect model assumptions [34].

In the context of clinical trials, Figure 2 indicates the different components associated with the aleatoric versus epistemic uncertainty framework. Each stage of the trial has a clear role in reducing aleatoric uncertainty. For example, the stage labeled as processes is well defined, in that implementation and the analysis of a clinical trial are executed by well-defined, a priori developed protocols. The data collection follows appropriate designs determined before the collection commences. In addition, the entry criteria in the protocol define the population, and randomization seeks to balance predetermined enrollment criteria to ensure the validity of inference. Therefore, the statistical or the irreducible part of the uncertainty is well controlled. Data collected from clinical trials are experimental, are generally carefully generated, and are subjected to many controls.

In contrast, a parallel framework for the use of the ML algorithms in pharmacovigilance does not currently exist. Recognizing that randomized controlled trials are for causal inference at population levels and ML is used for predicting outcomes for individuals, the lessons learned from randomized controlled trials about how well-defined processes can reduce and control statistical uncertainty can facilitate the development of a framework appropriate for ML that, in turn, contributes toward trusting AI systems. We discuss this in the rest of this section. In the context of ML, data are not collected according to any predesigned experiment; these are observational data. The self-controlled case series method uses only cases to study the association between adverse health outcomes and medical products [35,36]. This method can be applied to population-based databases and has been used in vaccine safety studies. However, the use of these designs for the analysis of data that reside in population-based databases has well-known limitations, including the strong assumptions made for the analysis to be possible [35]. ML methods are used to predict outcomes using different data sources. The uncertainty in predictions can then be assessed to provide a measure of trustworthiness of the results.

ML algorithms are not well understood or easily comprehended by their users, as evidenced by the experiences of the safety reviewers discussed earlier. These algorithms depend on data; to identify the circumstances under which they operate implies that the analyst must know and understand the data with which the algorithm works. In addition, many of these algorithms depend upon their hyperparameter setting (ie, a configuration that is external to the model used; they are often set or tuned

XSL•FO
RenderX

by the users of the model for the specific problem for which they are applied) to have good performance. The definition of "good performance" itself is subject to discussion (eg, whether to emphasize precision or recall) and depends upon the context of use. Returning to the context of hyperparameters, an algorithm's performance depends on the setting of these parameters or tuning, an action that requires the definition of a search space [37,38]. This is indeed a serious limitation of these methods. Currently, the FDA is applying ML algorithms to separate out reports that are highly unlikely to contain information that can be used by a safety reviewer to establish causality (ie, likely not useful data) [22].

A proper representation of algorithmic uncertainty is, therefore, an important prerequisite for the acceptance of AI methods and ML algorithms in pharmacovigilance. Quantifying algorithmic uncertainty is a current active research area, with authors attempting to propose measures that reflect aleatoric and epistemic uncertainty [32]. The field of statistical sciences has contributed to the quantification of aleatoric uncertainty; examples include methods that measure the variance of resampling estimates (ie, cross-validation or bootstrap) of the generalization error of computer algorithms [39-42]. Informally speaking, the generalization error is the error an ML algorithm makes on cases that it has not seen previously and indicates the algorithm's predictive ability. Relatively recent work proposes the construction of CIs for the generalization error [43-45].

An important consideration is the training of the algorithm and the FDA safety reviewers, that is, the individuals who will be using the algorithm. The training of ML algorithms to identify cases that represent AEs requires large amounts of data, which may be a limitation for rare or extremely rare AEs. In the case of training safety reviewers to use ML or rule-based algorithms, trust in the algorithm can be developed by educating end users, such as safety reviewers, about the algorithm. The training should include the features used, how they operate, what their strengths and limitations are, and how the algorithms can be applied to relevant tasks. The more comfortable the reviewers are with the algorithms, the more likely they will be to actually use them.

Algorithms are models and, as such, are imperfect. Box [46] stated that "All models are wrong, some are useful," essentially stating that the real test of knowledge is not truth but utility. How can one then make decisions with imperfect models? What steps should we take to make decisions?

The uncertainty associated with model imperfection has implications for how much trust we put in the model's outcome. Constructing CIs for model uncertainty and observing short lengths of these intervals can provide a measure that facilitates model trust. Short (ie, narrow) CIs necessarily have small magnitude of uncertainty because the length is small. However, it is possible that the magnitude indicated by the length is not acceptable for a given situation, such as, if we need magnitude of <0.5 and we obtain magnitude of 1. Constructing CIs for model uncertainty and observing short lengths of these intervals can provide a measure of understanding the magnitude of the uncertainty and enhancing model trust. Furthermore, computational reproducibility and the existence of protocols that facilitate computational reproducibility adds additional components to model trust. The imperfection of the models used has implications on how to specify, estimate, and evaluate these models as well as for how we interpret the results we obtain and the trust we put in their predictions. Evaluating algorithmic uncertainty via the construction of associated intervals (eg, intervals for assessing predictive uncertainty and model uncertainty) contributes toward better understanding of model performance. The development of a broad collection of models and methods, potentially incorporating varying degrees of uncertainty, provides an approach to decision-making in the presence of imperfect models.

Figure 2. Main aspects of clinical trials and machine learning technologies. Text in parenthesis indicates how uncertainty is controlled. Controlled clinical trials have well-developed protocols that define appropriate processes that aim to reduce statistical uncertainty. Analogous protocols should be developed for machine learning applications.

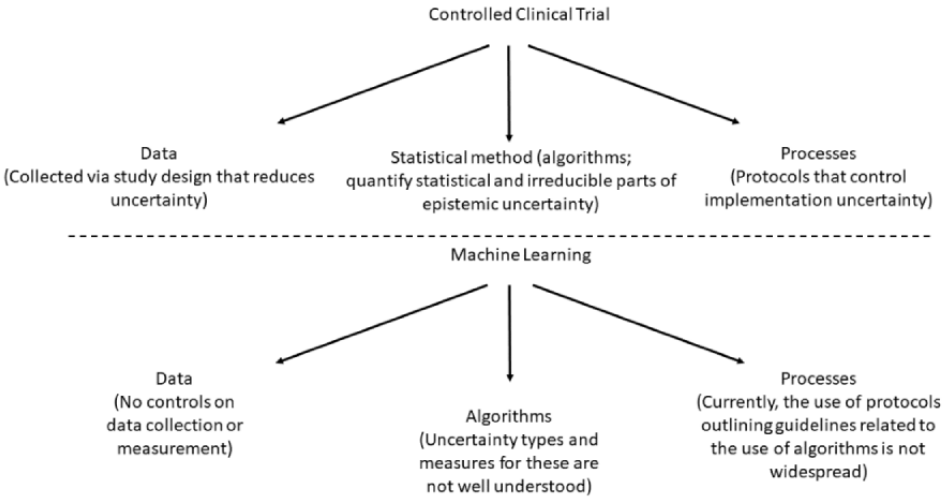

XSL•FO
RenderX

Discussion and Research Priorities

In this viewpoint, our aim is to describe the lessons learned from the FDA's experience of applying NLP ML to ICSR assessment and the research needed to address both general issues in case-based reasoning using AI and specific needs for ICSR assessment. Looking through the lens of the Diffusion of Innovations theory, we found that the AI algorithm for deduplication successfully incorporated the key attributes of an innovation, namely, relative advantage, trialability, and observability. The key to the deduplication algorithm's acceptance is that the tool transparently implemented an algorithm that closely paralleled existing pharmacovigilance workflows and safety reviewers' cognitive process for deduplication. In contrast, 4 attributes of innovation—compatibility, complexity, observability, and trialability—provide possible explanations for the lack of acceptance of the causality assessment algorithm. These attributes are much more difficult to satisfactorily fulfill for case-based reasoning. From this analysis, we conclude that the lack of a basis for verification of the causality assessment algorithm's validity in a detailed exposition of a human safety reviewer workflow suggests that developing a more complete understanding of how inferences are made in case-based reasoning for ICSR causality assessment is needed, including improved methods for measuring and controlling of algorithmic uncertainty and computational reproducibility.

It is widely recognized that there remains a gap between AI algorithm development and deployment [1]. Approaches to narrowing the gap are typically presented in the specific technical context of AI algorithms. For example, according to the FDA document on "Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan," several factors must be considered when deciding whether an AI algorithm might be ready for implementation. These factors include algorithm performance (eg, validity, generalizability, absence of bias, and robustness in real-world settings with changing inputs), documentation, transparency, explainability (ie, the reasons for an algorithm's prediction), quality control with real-world data collection and monitoring, and algorithm change control (ie, a structured approach to updating an algorithm using new data) [26]. While AI algorithms used for ICSR processing submission and evaluation by drug companies generally are not required to be submitted to the FDA for approval, the FDA has seen an increase in regulatory submissions with AI components in drug development more generally, including for postmarketing safety monitoring in epidemiological studies [47]. The FDA recently released a discussion paper on AI in drug development to further engage with stakeholders on defining an approach to regulating such algorithms [48].

On the basis of the FDA's experience of applying AI to ICSR causality assessment, we propose that these factors alone might be insufficient to address the AI development to deployment gap for many case-based reasoning scenarios. Specifically, by applying the Diffusion of Innovations theory, we identify the need for an improved general theory of inference for case-based reasoning as a critical step. In addition, creation of a map of the cognitive framework used by safety reviewers for causality assessment will be necessary for the application of AI to pharmacovigilance.

In moving from these general observations to research priorities to achieve these goals (Textbox 1), we asked ourselves 2 key questions. First, how does the computer "learn" human knowledge? Second, how can we be confident that the knowledge is both correct and accurately captured in the computer algorithms?

Most current approaches answer the first question through a combination of rule-based algorithms, human data annotations, and by using human intervention (ie, the human in the loop) in dialogue with the algorithm to enable iterative learning in machines. Rule-based algorithms are algorithms infused with human knowledge in the form of "if-then" statements for specific rules, and they are not as flexible as ML algorithms. However, rule-based algorithms are used in NLP and are well suited for low data volume and relatively simple rules. A very critical point for developing algorithms of all types is acquiring essential data and annotating them with human intervention. The engagement of the safety reviewer with the computer algorithm can help address the issue of incomplete features and data sparseness, especially when training data are lacking. Thus, the concept "human-in-the-loop" has been proposed to tackle these challenges, by incorporating human knowledge into the process (in this case the algorithmic model).

The answer to the second question is more challenging and currently understudied. New AI technologies using large language models to mimic human reasoning by identifying the most likely sequence of words may become sufficiently robust to superficially meet the Diffusion of Innovation theory's attributes, such as complexity and compatibility. One such algorithm has already been reported to have passed the medical board examinations and as a result is controversially being proposed as potentially supporting clinical decision-making [50]. In the context of application to case-based reasoning in pharmacovigilance, large language models might be used to summarize the narratives of ICSRs as part of a case series evaluation and to provide a narrative description that has all the hallmarks of a careful analysis but which might be riddled with errors. For example, a large language model might construct a sentence that follows the pattern observed in a narrative describing the relationship between a drug and a clinical outcome. The description, however, might be incorrect because it does not incorporate the scientific and clinical knowledge about the relationships among the various factors that a human expert would likely include in their description. Large language models will likely require that additional knowledge models be incorporated into their workflows for a complete analysis of case series and case-based reasoning more generally.

In conclusion, we need an improved understanding of causal inference and the cognitive framework for determining the causal relationship between a drug and an AE. While human expert evaluation is the current gold standard, the cognitive framework remains incompletely articulated [11,12]. Making a computable cognitive framework trustworthy will not just require its full articulation but also the application of a process to measure and quantify uncertainty as well as computational

XSL•FO
RenderX

reproducibility. Implementation should enable transparent comparisons of the data used, the decisions being made by the algorithm to the incorporated data, and decisions made by human experts. Improved understanding of causal inference and the cognitive framework for determining the causal relationship between a drug and an AE will still be important to optimize how such assessments are undertaken and the decision-making process derived from ICSRs regarding the benefits and risks of drugs.

Textbox 1. Research priorities for facilitating the use and trust of artificial intelligence (AI) tools in case-based assessments.

For the application of AI to case-based reasoning generally, future work is needed in the following aspects:

- Evaluate, understand, explain, and ultimately control uncertainty associated with algorithms for them to be useful, using alternative approaches and models.
- Assess computational reproducibility, efficiency, and resource requirements.
- Develop methods that support the evaluation of rule-based algorithms and can measure the extent to which the constructed rules satisfy the end-user's requirements and that can determine whether the rule definitions are accurate.

For the application of AI to individual case safety report (ICSR) evaluation, future work is needed in the following aspects:

- Better understand and document the cognitive framework of safety reviewers, in particular, when and how decisions are made using information external to the report itself, such as a case definition or known adverse effects.
- Empirically derive evidence by capturing the actual steps taken by a safety reviewer in real time and developing consensus on the detailed requirements for a high-quality case.
- Develop "tunable" AI algorithms in which performance characteristics and even features such as gender or race used in the algorithm would be selectable by the safety reviewer, consistent with the notion that computational reproducibility, operationalized as "what you can trust and what you can check," is a key component to the socialization of AI algorithms, which currently has limited formal underlying theory. If the supplied algorithm does not have the flexibility to incorporate these variables and provides results that are different from what the safety reviewer would expect, then the method decreases trust. Qualitative analysis performed by safety reviewers supplies additional variables to incorporate in the revised iteration of the algorithm to improve its performance.
- Explore the relevance of N-of-1 trials in drug development as the closest example of the type of data integration and individualized inferential approach that is needed for case-based reasoning, recognizing that the data limitations of ICSRs make this approach even more challenging. N-of-1 trials are multiple crossover trials, usually randomized and often blinded, and conducted on a single patient. Thus, N-of-1 trials are single-patient trials [49].

Acknowledgments

The authors thank the safety reviewers in the Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, Food and Drug Administration (FDA), who provided valuable feedback on the acceptance of artificial intelligence algorithms for pharmacovigilance. Representatives of the FDA reviewed a draft of the manuscript for presence of confidential information and accuracy regarding presentation of any FDA policy statements. The views expressed are those of the authors and not necessarily those of the US FDA. This work was also partially supported by the Troup Fund of the Kaleida Health Foundation.

Authors' Contributions

RB coconceived the topic, cowrote the original draft, reviewed and edited multiple drafts, interpreted the findings of the Diffusion of Innovation theory application, and coprovided overall supervision to the project. AHT co-developed the topic, interpreted the findings of the Diffusion of Innovation theory application, and reviewed and edited multiple drafts. OD reviewed and edited multiple drafts, applied the Diffusion of Innovation theory methodology to the acceptance of the deduplication and causality assessment algorithms, and administered the project. M Muñoz reviewed and edited multiple drafts and interpreted the findings of the Diffusion of Innovation theory application. M Markatou coconceived the topic, cowrote the original draft and reviewed and edited multiple drafts, interpreted the findings of the Diffusion of Innovation theory application, and coprovided overall supervision as the senior author. All authors approved the final version of the manuscript.

Conflicts of Interest

M Markatou and RB are authors on US Patent 9,075,796, "Text mining for large medical text datasets and corresponding medical text classification using informative feature selection." At present, this patent is not licensed and does not generate royalties. RB, OD, and M Muñoz are employees of the US Food and Drug Administration (FDA) and conducted this work as part of their official duties. M Markatou is the principal investigator and AHT is a coinvestigator on FDA-BAA-20-00123, "Evaluating the LRT Method for Post-Marketing Surveillance for Drug Safety: Comparison and Connection With the Pattern Discovery Framework."

References

XSL•FO
RenderX

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan 20;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
2. Best practices in drug and biological product Postmarket safety surveillance for FDA staff. U.S. Food and Drug Administration. URL: https://www.fda.gov/media/130216/download [accessed 2023-04-28]
3. Brown JS, Mendelsohn AB, Nam YH, Maro JC, Cocoros NM, Rodriguez-Watson C, et al. The US Food and Drug Administration sentinel system: a national resource for a learning health system. J Am Med Inform Assoc 2022 Nov 14;29(12):2191-2200 [FREE Full text] [doi: 10.1093/jamia/ocac153] [Medline: 36094070]
4. Pinnow E, Amr S, Bentzen SM, Brajovic S, Hungerford L, St George DM, et al. Postmarket safety outcomes for new molecular entity (NME) drugs approved by the food and drug administration between 2002 and 2014. Clin Pharmacol Ther 2018 Aug 20;104(2):390-400. [doi: 10.1002/cpt.944] [Medline: 29266187]
5. Bulatao I, Pinnow E, Day B, Cherkaoui S, Kalaria M, Brajovic S, et al. Postmarketing safety-related regulatory actions for new therapeutic biologics approved in the United States 2002-2014: similarities and differences with new molecular entities. Clin Pharmacol Ther 2020 Dec 08;108(6):1243-1253 [FREE Full text] [doi: 10.1002/cpt.1948] [Medline: 32557564]
6. Lester J, Neyarapally GA, Lipowski E, Graham CF, Hall M, Dal Pan G. Evaluation of FDA safety-related drug label changes in 2010. Pharmacoepidemiol Drug Saf 2013 Mar 02;22(3):302-305. [doi: 10.1002/pds.3395] [Medline: 23280652]
7. Croteau D, Pinnow E, Wu E, Muñoz M, Bulatao I, Dal Pan G. Sources of evidence triggering and supporting safety-related labeling changes: a 10-year longitudinal assessment of 22 new molecular entities approved in 2008 by the US Food and Drug Administration. Drug Saf 2022 Feb 03;45(2):169-180. [doi: 10.1007/s40264-021-01142-3] [Medline: 35113347]
8. Ball R, Dal Pan G. "Artificial intelligence" for pharmacovigilance: ready for prime time? Drug Saf 2022 May 17;45(5):429-438 [FREE Full text] [doi: 10.1007/s40264-022-01157-4] [Medline: 35579808]
9. Rogers EM. Diffusions of Innovations. 5th edition. New York, NY: Free Press; 2003.
10. Ralph Edwards I. Causality assessment in pharmacovigilance: still a challenge. Drug Saf 2017 May 28;40(5):365-372. [doi: 10.1007/s40264-017-0509-2] [Medline: 28247280]
11. Agbabiaka TB, Savović J, Ernst E. Methods for causality assessment of adverse drug reactions: a systematic review. Drug Saf 2008;31(1):21-37. [doi: 10.2165/00002018-200831010-00003] [Medline: 18095744]
12. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. J Am Med Inform Assoc 2011 Sep 01;18(5):631-638 [FREE Full text] [doi: 10.1136/amiajnl-2010-000022] [Medline: 21709163]
13. Botsis T, Buttolph T, Nguyen MD, Winiecki S, Woo EJ, Ball R. Vaccine adverse event text mining system for extracting features from vaccine safety reports. J Am Med Inform Assoc 2012 Nov 01;19(6):1011-1018 [FREE Full text] [doi: 10.1136/amiajnl-2012-000881] [Medline: 22922172]
14. Botsis T, Woo EJ, Ball R. The contribution of the vaccine adverse event text mining system to the classification of possible Guillain-Barré Syndrome reports. Appl Clin Inform 2017 Dec 19;04(01):88-99. [doi: 10.4338/aci-2012-11-ra-0049]
15. Botsis T, Woo EJ, Ball R. Application of information retrieval approaches to case classification in the vaccine adverse event reporting system. Drug Saf 2013 Jul 24;36(7):573-582. [doi: 10.1007/s40264-013-0064-4] [Medline: 23703591]
16. Wang W, Kreimeyer K, Woo EJ, Ball R, Foster M, Pandey A, et al. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. J Biomed Inform 2016 Aug;62:78-89 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.006] [Medline: 27327528]
17. Baer B, Nguyen M, Woo EJ, Winiecki S, Scott J, Martin D, et al. Can natural language processing improve the efficiency of vaccine adverse event report review? Methods Inf Med 2016;55(2):144-150. [doi: 10.3414/ME14-01-0066] [Medline: 26394725]
18. Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Saf 2014 Apr 14;37(4):249-258. [doi: 10.1007/s40264-014-0146-y] [Medline: 24627310]
19. Kreimeyer K, Menschik D, Winiecki S, Paul W, Barash F, Woo EJ, et al. Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems. Drug Saf 2017 Jul 14;40(7):571-582. [doi: 10.1007/s40264-017-0523-4] [Medline: 28293864]
20. Norén GN. The power of the case narrative - can it be brought to bear on duplicate detection? Drug Saf 2017 Jul 30;40(7):543-546 [FREE Full text] [doi: 10.1007/s40264-017-0548-8] [Medline: 28560588]
21. Han L, Ball R, Pamer CA, Altman RB, Proestel S. Development of an automated assessment tool for MedWatch reports in the FDA adverse event reporting system. J Am Med Inform Assoc 2017 Sep 01;24(5):913-920 [FREE Full text] [doi: 10.1093/jamia/ocx022] [Medline: 28371826]
22. Kreimeyer K, Dang O, Spiker J, Muñoz MA, Rosner G, Ball R, et al. Feature engineering and machine learning for causality assessment in pharmacovigilance: lessons learned from application to the FDA Adverse Event Reporting System. Comput Biol Med 2021 Aug;135:104517. [doi: 10.1016/j.compbiomed.2021.104517] [Medline: 34130003]
23. Muñoz MA, Dal Pan GJ, Wei YJ, Delcher C, Xiao H, Kortepeter CM, et al. Towards automating adverse event review: a prediction model for case report utility. Drug Saf 2020 Apr 07;43(4):329-338. [doi: 10.1007/s40264-019-00897-0] [Medline: 31912439]

XSL•FO
RenderX

24. Botsis T, Jankosky C, Arya D, Kreimeyer K, Foster M, Pandey A, et al. Decision support environment for medical product safety surveillance. J Biomed Inform 2016 Dec;64:354-362 [FREE Full text] [doi: 10.1016/j.jbi.2016.07.023] [Medline: 27477839]
25. Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, et al. Information visualization platform for post-market surveillance decision support. Drug Saf 2020 Sep 22;43(9):905-915. [doi: 10.1007/s40264-020-00945-0] [Medline: 32445187]
26. Artificial intelligence and machine learning in software as a medical device. Food and Drug Administration, USA. 2021 Jan. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [accessed 2023-04-28]
27. Kreimeyer K, Dang O, Spiker J, Gish P, Weintraub J, Wu E, et al. Increased confidence in deduplication of drug safety reports with natural language processing of narratives at the US Food and Drug Administration. Front Drug Saf Regul 2022 Jun 15;2:918897. [doi: 10.3389/fdsfr.2022.918897]
28. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023 Jan 16;29(1):49-58. [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
29. Liu K, Meng X. There is individualized treatment. why not individualized inference? Annu Rev Stat Appl 2016 Jun 01;3(1):79-111. [doi: 10.1146/annurev-statistics-010814-020310]
30. Montañez MI, Mayorga C, Bogas G, Barrionuevo E, Fernandez-Santamaria R, Martin-Serrano A, et al. Epidemiology, mechanisms, and diagnosis of drug-induced anaphylaxis. Front Immunol 2017 May 29;8:614 [FREE Full text] [doi: 10.3389/fimmu.2017.00614] [Medline: 28611774]
31. Bloomgren G, Richman S, Hotermans C, Subramanyam M, Goelz S, Natarajan A, et al. Risk of natalizumab-associated progressive multifocal leukoencephalopathy. N Engl J Med 2012 May 17;366(20):1870-1880. [doi: 10.1056/NEJMoa1107829] [Medline: 22591293]
32. Senge R, Bösner S, Dembczyński K, Haasenritter J, Hirsch O, Donner-Banzhoff N, et al. Reliable classification: learning classifiers that distinguish aleatoric and epistemic uncertainty. Inf Sci 2014 Jan;255:16-29. [doi: 10.1016/j.ins.2013.07.030]
33. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021 Mar 08;110(3):457-506. [doi: 10.1007/S10994-021-05946-3]
34. Hodges JS. Uncertainty, policy analysis and statistics. Statist Sci 1987 Aug 1;2(3):259-275. [doi: 10.1214/ss/1177013224]
35. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. Stat Med 2006 May 30;25(10):1768-1797. [doi: 10.1002/sim.2302] [Medline: 16220518]
36. Whitaker HJ, Ghebremichael-Weldeselassie Y. Self-controlled case series methodology. Annu Rev Stat Appl 2019 Mar 07;6(1):241-261. [doi: 10.1146/annurev-statistics-030718-105108]
37. Weerts HJ, Mueller AC, Vanschoren J. Importance of tuning hyperparameters of machine learning algorithms. arXiv Preprint posted online July 15, 2020 [FREE Full text]
38. Lavesson N, Davidsson P. Quantifying the impact of learning algorithm parameter tuning. In: Proceedings of the 21st National Conference on Artificial Intelligence. 2006 Presented at: AAAI '06; July 16-20, 2006; Boston, MA p. 395-400 URL: https://dl.acm.org/doi/10.5555/1597538.1597602
39. Markatou M, Tian H, Biswas S, Hripcsak G. Analysis of the variance of cross validation estimators of the generalization error of computer algorithms. J Mach Learn Res 2005;6:1127-1168 [FREE Full text] [doi: 10.7916/D86D5R2X]
40. Wang Q, Lindsay BG. Variance estimation of a general U-statistic with appllication to cross-validation. Stat Sinica 2014;24:1117-1141. [doi: 10.5705/ss.2012.215]
41. Afendras G, Markatou M. The out-of-source error in multi-source cross validation-type procedures. In: Liu A, Chen DG, Li G, Li Y, Zhao Y, Jin Z, editors. New Advances in Statistics and Data Science. New York, NY: Springer; 2017:27-44.
42. Afendras G, Markatou M. Optimality of training/test size and resampling effectiveness in cross-validation. J Stat Plan Inference 2019 Mar;199:286-301. [doi: 10.1016/j.jspi.2018.07.005]
43. Laber EB, Murphy SA. Adaptive confidence intervals for the test error in classification. J Am Stat Assoc 2011 Sep;106(495):904-913. [doi: 10.1198/jasa.2010.tm10053]
44. Mentch L, Hooker G. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. J Mach Learn Res 2016;17:1-41 [FREE Full text]
45. Marques F, Paulo C. Confidence intervals for the random forest generalization error. Pattern Recognit Lett 2022 Jun;158:171-175. [doi: 10.1016/j.patrec.2022.04.031]
46. Box GE. Science and statistics. J Am Stat Assoc 1976 Dec;71(356):791-799. [doi: 10.1080/01621459.1976.10480949]
47. Liu Q, Huang R, Hsieh J, Zhu H, Tiwari M, Liu G, et al. Landscape analysis of the application of artificial intelligence and machine learning in regulatory submissions for drug development from 2016 to 2021. Clin Pharmacol Ther 2023 Apr 16;113(4):771-774. [doi: 10.1002/cpt.2668] [Medline: 35707940]
48. Using artificial intelligence and machine learning in the development of drug and biological products. Food and Drug Administration, USA. URL: https://www.fda.gov/media/167973/download?attachment [accessed 2024-04-29]
49. Kravitz RL, Duan N. Design and implementation of n-of-1 trials: a user's guide. Agency for Health Care Research and Quality. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/n-1-trials_research-2014-5.pdf [accessed 2024-04-29]

XSL•FO
RenderX

50. Kung TH, Cheatham M, Medenilla A, Sillos C, de Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]

Abbreviations

AE: adverse event
AI: artificial intelligence
FDA: Food and Drug Administration
ICSR: individual case safety report
ML: machine learning
NLP: natural language processing
PML: progressive multifocal leukoencephalopathy

Edited by T de Azevedo Cardoso; submitted 25.06.23; peer-reviewed by A Bate, D Chrimes, J Chen; comments to author 26.10.23; revised version received 22.12.23; accepted 26.04.24; published 06.06.24.

Please cite as:
Ball R, Talal AH, Dang O, Muñoz M, Markatou M
Trust but Verify: Lessons Learned for the Application of AI to Case-Based Clinical Decision-Making From Postmarketing Drug Safety Assessment at the US Food and Drug Administration
J Med Internet Res 2024;26:e50274
URL: https://www.jmir.org/2024/1/e50274
doi:10.2196/50274
PMID:38842929

©Robert Ball, Andrew H Talal, Oanh Dang, Monica Muñoz, Marianthi Markatou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Key Considerations for Designing Clinical Studies to Evaluate Digital Health Solutions

Elaina Bolinger[1*], PhD; Benoit Tyl[2*], MD

[1]Integrated Evidence Generation & Business Innovation, Bayer AG, Berlin, Germany

[2]Integrated Evidence Generation & Business Innovation, Bayer HealthCare SAS, La Garenne Colombes, France

[*]all authors contributed equally

Corresponding Author:
Benoit Tyl, MD
Integrated Evidence Generation & Business Innovation
Bayer HealthCare SAS
10 Place de Belgique
La Garenne Colombes, F-92254
France
Phone: 33 6 80 29 07 79
Email: benoit.tyl@bayer.com

Abstract

Evidence of clinical impact is critical to unlock the potential of digital health solutions (DHSs), yet many solutions are failing to deliver positive clinical results. We argue in this viewpoint that this failure is linked to current approaches to DHS evaluation design, which neglect numerous key characteristics (KCs) requiring specific scientific and design considerations. We first delineate the KCs of DHSs: (1) they are implemented at health care system and patient levels; (2) they are "complex" interventions; (3) they can drive multiple clinical outcomes indirectly through a multitude of smaller clinical benefits; (4) their mechanism of action can vary between individuals and change over time based on patient needs; and (5) they develop through short, iterative cycles—optimally within a real-world use context. Following our objective to drive better alignment between clinical evaluation design and the unique traits of DHSs, we then provide methodological suggestions that better address these KCs, including tips on mechanism-of-action mapping, alternative randomization methods, control-arm adaptations, and novel end-point selection, as well as innovative methods utilizing real-world data and platform research.

(J Med Internet Res 2024;26:e54518) doi:10.2196/54518

KEYWORDS

evidence generation; clinical robustness; clinical trials; digital health; solutions; digital health solutions; DHS; health care system; patients; patient; individuals; individual; healthcare system; control arm adaptations; randomization methods; real-world data; platform research

Introduction

Digital health solutions (DHSs) can improve health care access, patient equity, operational efficiency, and cost-effectiveness for health care organizations while delivering clinical outcomes for patients. Despite the rapid proliferation of DHSs, most lack convincing evidence supporting their clinical impact [1], preventing their uptake within the health care system and blocking subsequent development cycles required to realize their full clinical potential. Closing this evidence gap and producing strong evidence in a timely and cost-effective manner is critical for the establishment of trust in DHSs and imperative for their adoption and implementation—prerequisites for unlocking digital health's potential to truly transform health care [2].

Randomized controlled trials (RCTs) using placebo control groups; individual randomization; and strict, arguably artificial clinical settings are the gold standard for evaluating drug efficacy and safety [3]. But contrary to pharmacological treatments, in which efficacy is driven by the modulation of biology at the molecular level, DHSs belong to a specific interventional class known as "complex interventions" [4]. By definition, the impact of this type of intervention is not solely biological but also encompasses psychological, behavioral, and systems-level effects, indicating context-dependency [5]. These attributes may, for example, help explain why effect sizes for pharmaceutical products are generally higher in traditional RCTs

XSL•FO
RenderX

compared to research conducted under conditions reflecting real-world conditions [6,7], whereas DHS effect sizes tend to be higher when using designs that mimic real-world conditions [8]. The unique attributes of DHSs and the "complex" class of interventions they deliver may therefore necessitate a more tailored approach to their evaluation to measure their true impact.

In this viewpoint, we discuss our perspective on what must be considered when designing clinical evaluations of DHSs. We specifically aimed to (1) delineate a series of key characteristics (KCs) of DHSs that make them different from pharmacological agents and should be considered while designing their evaluation, and (2) offer methodological solutions to adapt research to the requirements of DHSs, addressing randomization; control-arm design; end-point selection; and nontraditional, innovative adaptations.

DHS Key Characteristics

KC1: DHSs Are Often Implemented at Health Care System and Patient Levels

While patients may generate most of the data, DHSs frequently involve other individuals such as health care professionals (HCPs) and caregivers, who must integrate the solutions into their workflows for the DHS to reach full effectiveness. The benefits of using DHSs frequently expand beyond patients to caregivers, HCPs, and the whole health care organization. For instance, implementing a remote patient monitoring DHS in a single health care center can have permeating systemic effects, with the product improving outcomes of the patients monitored as well as those of the other patients, either by decreasing readmission and then shifting clinical staff resources to preventive care or by improving health care efficiency (streamlined workflow, optimized patient engagement strategies, etc). Within an evaluation setting, HCPs in the control group may adapt their behavior via learning from those included in the intervention group and change their clinical strategies (ie, leading to contamination). Therefore, it is important to control not just at the patient level but also at the health care site level [9].

KC2: DHSs Are "Complex" Interventions

Owing to their multiuser designs and dependencies on real-world health care systems, the mechanism of action (MoA) of DHSs depends on the historical, situational, environmental, and psychological factors in which treatments are delivered. Contextual dependencies include but are not limited to: HCP characteristics; patient-HCP relationships; perceived intervention credibility; delivery modality; psychological state of the individual; and societal, economic, and cultural factors. Even seemingly small nuances, such as a patient having previously been asked about their health, can have a significant impact on their health-related behavior and consequently also on health outcomes [10-12]. Many of these contextual factors are unintentionally, artificially modified in pharmaceutical RCTs to maximize internal validity. For instance, the Hawthorn effect, which is an increase in engagement when people are observed [13], and other research participation effects [14] are controlled between groups by using placebos. Yet this engagement is in fact a designed benefit of several DHSs; therefore, adding measurement points that would not occur in real-world situations to a DHS control group can render the control noninert. For example, measuring and regularly reporting blood pressure improves engagement and hypertension control [15], and similarly, reporting symptoms improves cancer self-management practices [16]. The artificial contexts and measurement conditions traditionally used in pharmaceutical RCTs can therefore limit the ecological validity of how the DHS is used. Consequently, conclusions drawn from such designs may lack external validity.

KC3: DHSs May Drive Multiple Clinical Outcomes Indirectly Through a Multitude of Smaller Clinical Benefits

DHSs are frequently "holistic" interventions that provide a broad spectrum of solutions to induce diverse changes, mainly behavioral, which then work as levers to deliver clinically meaningful outcomes. For instance, a remote patient monitoring platform for patients with diabetes that reduces time to treatment adjustment and thereby improves pharmaceutical adherence should consequently improve blood glucose regulation. Similarly, any improvement of sleep behavior by a sleep coaching DHS may also improve cardiovascular risk through better exercise, decreased weight, better eating habits, decreased cholesterol, and improved glucose regulation. While the magnitude of effect of the DHS on each of these end points may be limited, the cumulative effects may lead to a clinically significant benefit. Choosing a single primary clinical end point during evaluation therefore may not capture the complexity of the DHS's intended purpose nor its true performance.

KC4: The MoA Driving Clinical Outcomes in DHSs Can Vary Between Individuals and Change Over Time Based on Patient Needs

Many DHS interventions include several separate, situationally activated components or features (HCP-facing interface, symptom monitoring, activity tracking, etc) that can bring unique benefits. For example, one patient might use feature A, which indirectly improves cholesterol levels; another might prefer feature B and experience an improvement in blood pressure, with smaller changes observed in cholesterol; and some patients might benefit from both. The effect size of the intervention for a single end point may therefore depend on how the product is used. Moreover, the second patient may begin to prefer feature A, but only later during use. The effect size of the intervention for a single end point can therefore be weakened when the DHS as a whole rather than specific feature use is considered as the intervention. What is more, many DHSs employ adaptive algorithms that are specifically designed to adjust the intervention over time. Depending on the DHS being considered, researchers might therefore anticipate that due to user-dependent variability in the MoA and emerging synergistic influences (ie, better product engagement following use-based personalization), the clinical impact may require a longer treatment period to be detected.

XSL•FO
RenderX

KC5: DHSs Develop Through Short, Iterative Cycles—Often Within a Real-World Use Context

Contrary to active pharmacological ingredients that cannot change, DHSs continuously evolve in terms of their functionality with each subsequent release. Because DHSs must meet their users' needs in the real world, an environment where many factors remain unknown [17], developers often perform accelerated, iterative tests to both discover product performance factors and adapt them before the next release. Due to the number of factors that must be discovered, how they can influence each other, and how influential certain factors can be (ie, a seemingly small adjustment of written content to an audio format might completely change engagement and therefore clinical outcomes), the solution often must stay moderately malleable, as product design hypotheses are validated over a continuum of nonclinical tests. Clinical researchers often opt to evaluate a "frozen" version of the product to optimize homogeneity in the intervention arm, but these approaches may limit external validity, as the marketed solution may drastically evolve over time and the versions tested by the RCT may be outdated at the time of study completion.

Recalibrating Research Design to Meet the Scientific Canvas of DHSs

Considering the KCs described above, alternative assessment strategies are needed to address the unique challenges of DHS evaluation while maintaining internal validity (unbiased study design), external validity (applicability to different contexts), and ecological validity (generalizability to real-world settings). The solutions discussed below are summarized in Figure 1.

The first step to design a DHS evaluation is to take the time to delineate the MoA of the solution and the context. This includes a map of behavioral changes required from patients and all other individuals who are affected by the product (KC1), as well as the contextual environment of use (KC2). Given the limitations of an active control arm described above (KC2), the use of usual care—usually named treatment as usual (TAU)—may serve as an appropriate control. TAU corresponds to the actual routine care and may differ from the state-of-the-art, guideline-adherent clinical care. Therefore, a prerequisite for this approach would be a detailed understanding of the procedural standard of care, which includes, but is not limited to, *local* (ie, site-specific) procedures, treatment recommendations, information delivered (eg, health literacy counseling), and frequency of contacts with HCPs. This task is nontrivial, as evidenced by standard-of-care descriptions being generally underreported [18], and can pose exceptional challenges due to the number of factors in a real-world environment. Combining patient journeys, DHS user journeys, and HCP patient management protocols is a helpful first step to identify how the product induces change within its context of use. Process evaluation techniques [19] can also be employed to extract MoAs for complex interventions during preliminary implementation research [5]. With the MoA delineated in detail, one can begin designing an appropriate evaluation employing a fit-for-purpose scientific approach.

As many have suggested before [9,20], cluster randomization by health care site, cities, or region is a methodologically sound approach to control for contamination between arms while taking into account DHS benefits, which depend on factors extending beyond the treated patient (KC1). This design, however, can necessitate larger sample sizes, and site matching can be challenging. Pragmatic designs [9], which anchor RCTs in real-world settings and minimize research-related contact outside of TAU, can help maximize the ecological validity of both the experimental (+TAU) and TAU arms (KC2). Keeping the trial decentralized can minimize research-related contact while increasing trial accessibility and make the research population more representative of the target population (KC2). While DHSs, which digitally collect data by nature, are particularly suitable for such designs, limitations secondary to the reduced investigators oversight, including data quality and outcome assessment, persist.

Even with these tools in mind, designing a control arm with high internal validity remains challenging when evaluating DHSs. A reference DHS is an appropriate control but rarely exists. While some have argued that a "sham" app (ie, as defined by the Food and Drug Administration, an app without the "active" features anticipated to drive the MoA [21]) offers optimal internal validity, in that it is a form of "placebo" technology, this design can pose critical threats to external and ecological validity, particularly when the MoA itself is not fully understood. Providing even minimalistic digital support that is nevertheless seemingly endorsed by HCPs (as would be the case in a clinical trial) might still produce a similar, albeit muted, effect as the DHS. For example, a sham app delivering some health literature is not inert, as it can improve health literacy and consequently health-related behaviors. The use of a sham app as a control is therefore challenging unless it is possible to exclude all active ingredients of the MoA, including engagement and interaction with the real world, from the sham app. On the other hand, the use of TAU as a control can potentially inflate effect size with some interventions (KC2). The traditional "waitlist" design used in medical device evaluation may also be appropriate but can inflate effect sizes by decreasing behavioral activation in the control arm (in essence, inducing the opposite of the Hawthorne effect) [22,23]. Researchers considering having a waitlist TAU design as the comparator might use an appropriate *run-in period* before randomization, during which both study arms receive TAU to minimize the influence of treatment expectancy effects [24] within the comparator arm (KC2). Finally, it is important to consider how the nature of the intervention and the control arm may impact the risk of attrition in the study. Beyond minimizing attrition in general, it is crucial to ensure that the attrition rates between the active and control arms are balanced. Significant differences in dropout rates between both arms may lead to variations in the clinical characteristics of the remaining patients, potentially biasing the overall analysis.

DHS researchers may also explore alternate evaluation designs. The use of synthetic control arms built using real-world data sources such as electronic health records, claims databases, and disease registries may address certain limitations inherent to the conventional control group. However, the creation of these

XSL•FO
RenderX

synthetic arms needs to be considered carefully, as they pose their own set of challenges and are not devoid of their own limitations. In case of products already on the market, epidemiological methods could be used to examine causal inference in product registry data—again minimizing research participation effects. These approaches may in some cases both be more cost-effective and provide higher quality data on product performance.

Designing a trial that accounts for the holistic impact of DHSs is also critical for demonstrating performance and efficacy. Contrary to pharmacological agents that usually improve markedly a single biomarker (hemoglobin A_{1c}, C-reactive protein, etc), DHSs frequently improve modestly several biomarkers (physical activity, lipids, weight, etc; KC3). Here, rather than examining the effect on a single end point, researchers might employ a risk score that better reflects the holistic nature of the intervention. Win ratios are also an elegant way to analyze multiple, equally valuable end points [25]. Adaptive designs, such as those incorporating sample size re-estimation, hypothesis-adaptive, or response-adaptive designs [26], could also be both powerful and appropriate given the novel nature of many DHS interventions and the different use patterns and emergent interventions that can arise (KC4). Given the dynamic and personalized nature of DHSs, the evaluation design should also carefully consider anticipated effect size at each time point during planning.

The above can greatly improve the evaluation of DHS clinical performance but do not address the practical issue of evaluating rapidly and continuously evolving DHSs (KC5). While there is no definitive way to address this, a few options exist beyond freezing the product. During evaluation planning, planned future modifications of the product can be scrutinized to determine whether they constitute major changes to the MoA or intended use and would therefore require protocol modification during study execution. By combining strict product design control practices with a Trials of Intervention Principles approach [27], researchers can aggregate groups of users who may have used product versions with minor design variations not anticipated to change the MoA, allowing the product to continue to organically develop (KC5). Platform trials that incorporate key product design and clinical value hypotheses in master protocols can be leveraged to have the successive versions of the DHS tested and the results of the substudies combined to maximize sample sizes, rather than performing several insufficiently powered successive trials, while keeping analyses anchored in clearly documented starting hypotheses and protocols (Figure 2) [20,28]. Likewise, factorial research designs can be used to investigate the efficacy of new features as they are added to the product [29,30].

XSL•FO
RenderX

Figure 1. Overview of key characteristics (KCs) of the digital health solutions (DHSs) and potential methodological solutions to improve their clinical assessment. EHR: electronic health record; MoA: mechanism of action. For a higher-resolution version of this figure, see Multimedia Appendix 1.

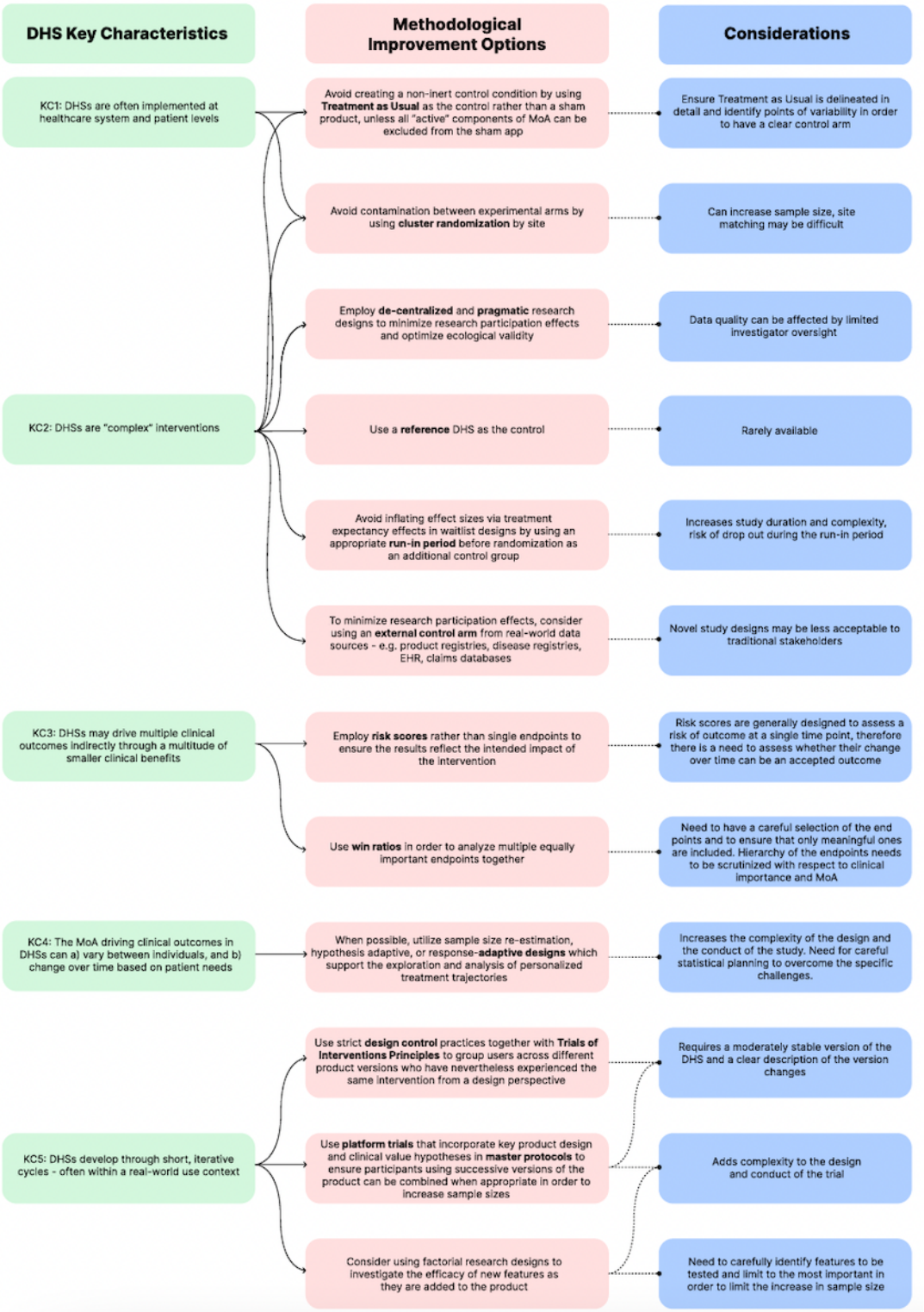

XSL•FO
RenderX

Figure 2. Example of design of a platform trial including 300 patients (150 per arm) enabling the evaluation of multiple versions of a DHS within one study. The figure illustrates the recruitment flow of the patients in the trial. Multiple versions of a DHS can be included in a single study when guidelines for identifying significant changes to study procedures and DHS solutions are delineated in advance. A matched control arm with similar recruitment procedures ensures comparability in the control group. DHS: digital health solution.

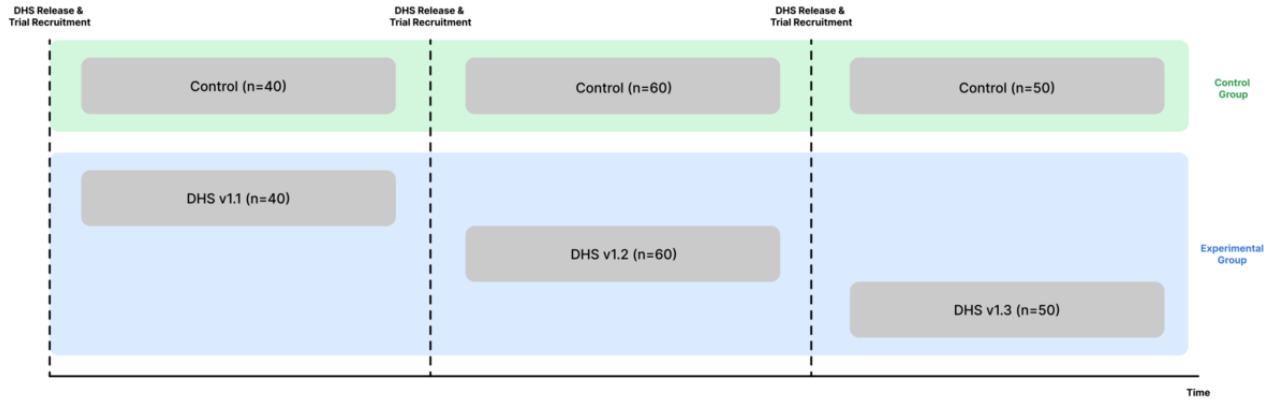

Conclusion

DHSs need to be rigorously evaluated to fulfill the needs of the various stakeholders, including HCPs, patients, regulators, and payers. In this viewpoint, we proposed that their unique characteristics and MoA create a critical need for tailored, innovative approaches that move beyond traditional pharmaceutical RCTs in order to create fit-for-purpose, robust, and scientifically optimized evaluations. While no one-size-fits-all design exists, both researcher and stakeholders should embrace nontraditional methodologies that match the key characteristics of this emerging type of health care intervention.

Acknowledgments

We thank Jason Lott for his feedback and comments.

Authors' Contributions

Both authors equally contributed to the manuscript.

Conflicts of Interest

Both authors are full-time employees of Bayer.

Multimedia Appendix 1
Overview of key characteristics (KCs) of the digital health solutions (DHSs) and potential methodological solutions to improve their clinical assessment. EHR: electronic health record; MoA: mechanism of action.
[PDF File (Adobe PDF File), 1793 KB - jmir_v26i1e54518_app1.pdf]

References

1. Day S, Shah V, Kaganoff S, Powelson S, Mathews SC. Assessing the clinical robustness of digital health startups: cross-sectional observational analysis. J Med Internet Res 2022 Jun 20;24(6):e37677 [FREE Full text] [doi: 10.2196/37677] [Medline: 35723914]
2. Egermark M, Blasiak A, Remus A, Sapanel Y, Ho D. Overcoming pilotitis in digital medicine at the intersection of data, clinical evidence, and adoption. Adv Intell Syst 2022 Sep 21;4(9):2200056. [doi: 10.1002/aisy.202200056]
3. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1986 Feb;89(2 Suppl):2S-3S. [doi: 10.1378/chest.89.2_supplement.2s] [Medline: 3943408]
4. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
5. Minary L, Trompette J, Kivits J, Cambon L, Tarquinio C, Alla F. Which design to evaluate complex interventions? toward a methodological framework through a systematic review. BMC Med Res Methodol 2019 May 07;19(1):92 [FREE Full text] [doi: 10.1186/s12874-019-0736-6] [Medline: 31064323]
6. Vonbank A, Drexel H, Agewall S, Lewis B, Dopheide J, Kjeldsen K, et al. Reasons for disparity in statin adherence rates between clinical trials and real-world observations: a review. Eur Heart J Cardiovasc Pharmacother 2018 Oct 01;4(4):230-236 [FREE Full text] [doi: 10.1093/ehjcvp/pvy028] [Medline: 30099530]

XSL•FO
RenderX

7. Kilcher G, Hummel N, Didden EM, Egger M, Reichenbach S, GetReal Work Package 4. Rheumatoid arthritis patients treated in trial and real world settings: comparison of randomized trials with registries. Rheumatology (Oxford) 2018 Feb 01;57(2):354-369. [doi: 10.1093/rheumatology/kex394] [Medline: 29149289]
8. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct 21;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
9. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020 Aug 27;3(1):110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
10. de Bruin M. Risk of bias in randomised controlled trials of health behaviour change interventions: evidence, practices and challenges. Psychol Health 2015 Jan 09;30(1):1-7. [doi: 10.1080/08870446.2014.960653] [Medline: 25223823]
11. Bishop FL, Fenge-Davies AL, Kirby S, Geraghty AW. Context effects and behaviour change techniques in randomised trials: a systematic review using the example of trials to increase adherence to physical activity in musculoskeletal pain. Psychol Health 2015 Jan 09;30(1):104-121. [doi: 10.1080/08870446.2014.953529] [Medline: 25109300]
12. Conner M, Godin G, Norman P, Sheeran P. Using the question-behavior effect to promote disease prevention behaviors: two randomized controlled trials. Health Psychol 2011 May;30(3):300-309. [doi: 10.1037/a0023036] [Medline: 21553974]
13. Berkhout C, Berbra O, Favre J, Collins C, Calafiore M, Peremans L, et al. Defining and evaluating the Hawthorne effect in primary care, a systematic review and meta-analysis. Front Med (Lausanne) 2022 Nov 8;9:1033486 [FREE Full text] [doi: 10.3389/fmed.2022.1033486] [Medline: 36425097]
14. McCambridge J, Kypri K, Elbourne D. Research participation effects: a skeleton in the methodological cupboard. J Clin Epidemiol 2014 Aug;67(8):845-849 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.03.002] [Medline: 24766858]
15. Milani RV, Lavie CJ, Bober RM, Milani AR, Ventura HO. Improving hypertension control and patient engagement using digital tools. Am J Med 2017 Jan;130(1):14-20. [doi: 10.1016/j.amjmed.2016.07.029] [Medline: 27591179]
16. Darley A, Coughlan B, Maguire R, McCann L, Furlong E. A bridge from uncertainty to understanding: the meaning of symptom management digital health technology during cancer treatment. Digit Health 2023 Jan 22;9:20552076231152163 [FREE Full text] [doi: 10.1177/20552076231152163] [Medline: 36714543]
17. Smits M, Ludden GDS, Verbeek P, van Goor H. How digital therapeutics are urging the need for a paradigm shift: from evidence-based health care to evidence-based well-being. Interact J Med Res 2022 Oct 20;11(2):e39323 [FREE Full text] [doi: 10.2196/39323] [Medline: 36264624]
18. Ayling K, Brierley S, Johnson B, Heller S, Eiser C. How standard is standard care? exploring control group outcomes in behaviour change interventions for young people with type 1 diabetes. Psychol Health 2015 Jan 20;30(1):85-103 [FREE Full text] [doi: 10.1080/08870446.2014.953528] [Medline: 25118842]
19. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015 Mar 19;350(mar19 6):h1258-h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
20. Hemkens LG. Benefit assessment of digital health applications-challenges and opportunities [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct 15;64(10):1269-1277 [FREE Full text] [doi: 10.1007/s00103-021-03413-x] [Medline: 34524477]
21. Lutz J, Offidani E, Taraboanta L, Lakhan SE, Campellone TR. Appropriate controls for digital therapeutic clinical trials: a narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. Front Digit Health 2022 Aug 18;4:823977 [FREE Full text] [doi: 10.3389/fdgth.2022.823977] [Medline: 36060538]
22. Cunningham JA, Kypri K, McCambridge J. Exploratory randomized controlled trial evaluating the impact of a waiting list control design. BMC Med Res Methodol 2013 Dec 06;13(1):150 [FREE Full text] [doi: 10.1186/1471-2288-13-150] [Medline: 24314204]
23. Bendtsen M, Ulfsdotter Gunnarsson K, McCambridge J. Effects of a waiting list control design on alcohol consumption among online help-seekers: protocol for a randomised controlled trial. BMJ Open 2021 Aug 26;11(8):e049810 [FREE Full text] [doi: 10.1136/bmjopen-2021-049810] [Medline: 34446493]
24. Rutherford BR, Wager TD, Roose SP. Expectancy and the treatment of depression: a review of experimental methodology and effects on patient outcome. Curr Psychiatry Rev 2010 Feb 01;6(1):1-10 [FREE Full text] [doi: 10.2174/157340010790596571] [Medline: 24812548]
25. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J 2012 Jan 06;33(2):176-182. [doi: 10.1093/eurheartj/ehr352] [Medline: 21900289]
26. Mahajan R, Gupta K. Adaptive design clinical trials: methodology, challenges and prospect. Indian J Pharmacol 2010 Aug;42(4):201-207 [FREE Full text] [doi: 10.4103/0253-7613.68417] [Medline: 20927243]
27. Mohr DC, Schueller SM, Riley WT, Brown CH, Cuijpers P, Duan N, et al. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. J Med Internet Res 2015 Jul 08;17(7):e166 [FREE Full text] [doi: 10.2196/jmir.4391] [Medline: 26155878]

XSL•FO
RenderX

28. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023 Jan 16;29(1):49-58. [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
29. Tarricone R, Petracca F, Ciani O, Cucciniello M. Distinguishing features in the assessment of mHealth apps. Expert Rev Pharmacoecon Outcomes Res 2021 Aug 04;21(4):521-526. [doi: 10.1080/14737167.2021.1891883] [Medline: 33663324]
30. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]

Abbreviations

DHS: digital health solution
HCP: health care professional
KC: key characteristic
MoA: mechanism of action
RCT: randomized controlled trial
TAU: treatment as usual

Edited by T de Azevedo Cardoso; submitted 13.11.23; peer-reviewed by F Denis, M Savolainen; comments to author 14.02.24; revised version received 05.03.24; accepted 17.05.24; published 17.06.24.

Please cite as:
Bolinger E, Tyl B
Key Considerations for Designing Clinical Studies to Evaluate Digital Health Solutions
J Med Internet Res 2024;26:e54518
URL: https://www.jmir.org/2024/1/e54518
doi:10.2196/54518
PMID:38885020

©Elaina Bolinger, Benoit Tyl. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses

Xufei Luo[1,2,3,4,5]; Fengxian Chen[6], PhD; Di Zhu[7], MPH; Ling Wang[7], MPH; Zijun Wang[1,2,3,4,5]; Hui Liu[1,2,3,4,5]; Meng Lyu[7], MPH; Ye Wang[7], MPH; Qi Wang[8,9], PhD; Yaolong Chen[1,2,3,4,5], MD, PhD

[1]Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China

[2]World Health Organization Collaboration Center for Guideline Implementation and Knowledge Translation, Lanzhou, China

[3]Institute of Health Data Science, Lanzhou University, Lanzhou, China

[4]Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou University, Lanzhou, China

[5]Research Unit of Evidence-Based Evaluation and Guidelines, Chinese Academy of Medical Sciences (2021RU017), School of Basic Medical Sciences, Lanzhou University, Lanzhou, China

[6]School of Information Science & Engineering, Lanzhou University, Lanzhou, China

[7]School of Public Health, Lanzhou University, Lanzhou, China

[8]Department of Health Research Methods, Evidence and Impact, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[9]McMaster Health Forum, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Yaolong Chen, MD, PhD
Evidence-Based Medicine Center
School of Basic Medical Sciences
Lanzhou University
No 199 Donggang West Road
Chengguan District
Lanzhou, 730000
China
Phone: 86 13893104140
Email: chevidence@lzu.edu.cn

Abstract

Large language models (LLMs) such as ChatGPT have become widely applied in the field of medical research. In the process of conducting systematic reviews, similar tools can be used to expedite various steps, including defining clinical questions, performing the literature search, document screening, information extraction, and language refinement, thereby conserving resources and enhancing efficiency. However, when using LLMs, attention should be paid to transparent reporting, distinguishing between genuine and false content, and avoiding academic misconduct. In this viewpoint, we highlight the potential roles of LLMs in the creation of systematic reviews and meta-analyses, elucidating their advantages, limitations, and future research directions, aiming to provide insights and guidance for authors planning systematic reviews and meta-analyses.

(J Med Internet Res 2024;26:e56780) doi:10.2196/56780

KEYWORDS

large language model; ChatGPT; systematic review; chatbot; meta-analysis

Introduction

A systematic review is the result of a systematic and rigorous evaluation of evidence, which may or may not include a meta-analysis [1]. Owing to the strict methodology and comprehensive summary of evidence, high-quality systematic reviews are considered the highest level of evidence, positioned at the top of the evidence pyramid [2]. Additionally, high-quality systematic reviews and meta-analyses are often used to support the development of clinical practice guidelines, aid clinical decision-making, and inform health care policy formulation [3]. Currently, the methods of systematic reviews and meta-analyses are applied in various disciplines in medicine and beyond such as law [4], management [5], and economics [6], and have yielded positive results, contributing to the continuous advancement of these fields [7].

XSL•FO
RenderX

The process of conducting systematic reviews demands a substantial investment in terms of time, resources, human effort, and financial capital [8]. To expedite the development of systematic reviews and meta-analyses, various automated or semiautomated tools such as Covidence have been developed [9,10]. However, the emergence of large language models (LLMs), particularly chatbots such as GPT, presents a set of both challenges and opportunities in the realm of systematic reviews and meta-analyses [11]. Based on the emerging literature in this field, we here provide our perspectives on the potential for harnessing the capabilities of LLMs to accelerate the production of systematic reviews and meta-analyses, while also scrutinizing the potential impacts and delineating the crucial steps involved in this process.

The Process and Challenges of Performing a Systematic Review and Meta-Analysis

The procedures and workflows for conducting systematic reviews and meta-analyses are well-established. Currently, researchers often refer to the Cochrane Handbooks recommended by the Cochrane Library for intervention or diagnostic reviews [12,13]. In addition, some scholars and institutions have developed detailed guidelines on the steps and methodology for performing systematic reviews and meta-analyses [14-17]. Generally speaking, researchers should take the following steps to produce a high-quality systematic review and meta-analysis: determine the clinical question, register and draft a protocol, set inclusion and exclusion criteria, develop and implement a search strategy, screen the literature, extract data from included studies, assess the quality and risk of bias of included studies, analyze and processed data, write up the full text of the manuscript, and submit the manuscript for publication, as illustrated in Figure 1. These different steps contain many subtasks; therefore, conducting a complete systematic review and meta-analysis requires fairly complex and time-consuming work.

Although systematic reviews and meta-analyses have been widely applied and play an important role in developing guidelines and informing clinical decision-making, their production process faces many challenges. One of these challenges is the long production time and large resource requirements. The average estimated time to complete and publish a systematic review is 67.3 weeks, requiring 5 researchers and costing approximately US $140,000 [18,19]. More recently, the development of automated and semiautomated tools using natural language processing and machine learning have accelerated systematic review and meta-analysis production to some extent [20], with studies showing that such tools can help to produce a systematic review and meta-analysis within 2 weeks [21]. However, these tools also have some limitations. First, no single tool can fully accelerate the entire production process of systematic reviews and meta-analyses. Second, these tools cannot process and analyze literature written in different languages. Finally, the reliability of the results generated by these automated and semiautomated tools needs further validation as they are not yet widely adopted for this purpose.

Figure 1. The process of conducting a systematic review and meta-analysis.

XSL•FO
RenderX

Applications of LLMs in Medical Research

Chatbots based on LLMs such as ChatGPT, Google Gemini, and Claude have become widely applied in medical research. These chatbots have proven to be valuable in tasks ranging from knowledge retrieval, language refinement, content generation, and medical exam preparation to literature assessment. ChatGPT has been shown to excel in accuracy, completeness, nuance, and speed when generating responses to clinical inquiries in psychiatry [22]. Moreover, LLMs such as ChatGPT play a pivotal role in automating the evaluation of medical literature, facilitating the identification of accurately reported research findings [23]. Despite their significant contributions, these chatbots are not without limitations. Challenges such as the potential for generating misleading content and susceptibility to academic deception necessitate further scholarly discourse on effective mitigation strategies. Standardized reporting practices may contribute to delineating the applications of ChatGPT and mitigating research biases [24].

ChatGPT has also demonstrated significant application potential and promise in the process of conducting systematic reviews and meta-analyses. Various studies [11,25-32] indicate that ChatGPT can play a pivotal role in formulating clinical questions, determining inclusion and exclusion criteria, screening literature, assessing publications, generating meta-analysis code, and assisting the full-text composition, among other relevant tasks. The details of these capabilities are summarized in Table 1.

Table 1. The possible functions of chatbots in the creation of systematic reviews and meta-analyses encompassing separate stages of the process.

Tasks	Potential roles and application steps of chatbots	References
Determine the research topic/question	• Identify previously published systematic reviews and meta-analyses on the same topic. • Assist in determining the rationale for the research question. • Clarify the PICO (Population, Intervention, Comparison, Outcome) question.	[11,33-35]
Register and write a research proposal	• Generate preliminary, unverified registration information. • Draft an initial research proposal, subject to validation.	[11,36,37]
Define inclusion an exclusion criteria	• Establish inclusion criteria. • Establish exclusion criteria.	[11,38]
Develop a search strategy and conduct searches	• Develop and optimize search strategies. • Implement retrieval. • Collect grey literature.	[11,25,29,33,39-42]
Screen the literature	• Remove duplicate records. • Screen literature titles, abstracts, and keywords. • Screen the full text of the obtained literature. • Download the full text of the literature.	[11,25,27,28,33,34,43-47]
Extract the data	• Extract basic information. • Extract patient information. • Extract outcome information. • Extract table information.	[11,25,26,47-50]
Assess the risk of bias	• Extract relevant information based on the scale. • Evaluate the risk of bias based on the scale. • Present visual results.	[26,51-53]
Analyze the data/meta-analyses	• Extract outcome information. • Generate figures and tables for some results.	[11,26,37,54]
Draft the full manuscript	• Search for relevant references. • Polish language and grammar. • Adjust the reference citation format. • Summarize the abstract.	[25,33,55-58]
Submit and publish	• Assist in selecting a suitable journal. • Adjust the manuscript format. • Compose a cover letter. • Assist in preparing the submission.	[33,59]

XSL•FO
RenderX

Potential Roles of LLMs in Producing Systematic Reviews and Meta-Analyses

Determine the Research Topic/Question

Determining the clinical question of interest represents the initial and paramount step in the process of conducting systematic reviews and meta-analyses. At this juncture, it is crucial to ascertain whether comparable systematic reviews and meta-analyses have already been published and to delineate the scope of the forthcoming review and meta-analysis. Generally, for interventional systematic reviews, the Patient, Intervention, Comparison, and Outcome (PICO) framework is considered for defining the scope and objectives of the research question [60]. In this context, ChatGPT serves a dual role. On the one hand, it expeditiously aids in searching for published systematic reviews and meta-analyses related to the relevant topics (see Multimedia Appendix 1 and Multimedia Appendix 2) [34]. On the other hand, ChatGPT assists in refining the clinical question that needs to be addressed (see Multimedia Appendix 3), facilitating prompt determination of the feasibility of undertaking the proposed study. However, it is important to be cautious of the retrieval of false literature [35].

Register and Write a Research Proposal

The registration and proposal writing process constitutes a pivotal preparatory phase for conducting systematic reviews and meta-analyses. Registration enhances research transparency, fosters collaboration among investigators, and mitigates the redundancy of research endeavors. Drafting a proposal helps in elucidating the research objectives and methods, providing robust support for the smooth execution of the study. For LLMs, generating preliminary registration information and initial proposal content is remarkably convenient and facile (see Multimedia Appendix 4 and Multimedia Appendix 5). For example, ChatGPT can assist researchers in generating the statistical methods for a research proposal [37]. However, considering that LLMs often generate fictitious literature, the content they produce may be inaccurate; thus, discernment and validation of the generated content remain essential considerations.

Define Inclusion and Exclusion Criteria

The inclusion and exclusion criteria for systematic reviews and meta-analyses are instrumental in determining the screening standards for studies. Therefore, strict and detailed inclusion and exclusion criteria contribute to the smooth and high-quality conduct of preparing systematic reviews and meta-analyses. The use of a chatbot based on LLMs can help in establishing the inclusion and exclusion criteria (see Multimedia Appendix 6) [38]; however, the inclusion criteria need to be optimized and adjusted according to the specific research objectives and the exclusion criteria should be based on the foundation of the inclusion criteria. Therefore, manual adjustments and optimizations are also necessary.

Develop a Search Strategy and Conduct Searches

ChatGPT can assist in formulating search strategies, using PubMed as an example [40]. Researchers can simply list their questions using the PICO framework and a search strategy can be quickly generated (Multimedia Appendix 1 and Multimedia Appendix 2). Based on the generated search strategy, one method is to copy the strategy from ChatGPT and paste it into the PubMed search box for direct retrieval [40,41]. Another approach involves using the OpenAI application programming interfaces (APIs) to invoke PubMed APIs with the search strategy generated by ChatGPT. This facilitates searching the PubMed database, obtaining search results, and applying predetermined inclusion and exclusion criteria. Subsequently, ChatGPT can be used to filter the search results, exporting and recording the filtered results in JSON format. This integrated process encompasses search strategy formulation, retrieval, and filtering. However, the direct use of LLMs to generate search strategies and complete the one-stop process of searching and screening may not yet be mature, and this poses a significant challenge for generating the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart. Therefore, we suggest using LLMs to generate search strategies, which should then be optimized and modified by librarians and computer experts (specializing in LLMs) before manually searching the databases. Additionally, to use search strategies transparently and reproducibly, the detailed prompts used should be reported [40,42].

Screen the Literature

Literature screening is one of the most time-consuming steps in the creation of systematic reviews and meta-analyses. Prior to the advent of ChatGPT, there were already many automated and semiautomated tools available for literature screening, such as Covidence, EPPI-Reviewer, DistillerSR, and others [39]. With the emergence of ChatGPT, researchers can now train the model based on predefined inclusion criteria. Subsequently, ChatGPT can be used to automatically screen records retrieved from databases and obtain the filtered results. Previous studies suggested that using ChatGPT in the literature selection process for a meta-analysis substantially diminishes the workload while preserving a recall rate on par with that of manual curation [28,44-47].

Extract the Data

Data extraction involves obtaining information from primary studies and serves as a primary source for systematic reviews and meta-analyses. Generally, when conducting systematic reviews and meta-analyses, basic information must be extracted from the original studies, such as publication date, country of conduct, and the journal of publication. Additionally, characteristics of the population, such as patient samples, age, gender/sex, and outcome data, are also extracted, including event occurrences, mean change values, and total sample size. Currently, tools based on natural language processing and LLMs, such as ChatGPT and Claude, demonstrate high accuracy in extracting information from PDF documents (see Multimedia Appendix 7 for an example) [47-50]. However, it is important to note that despite their promising capabilities, manual verification remains a necessary step in the data extraction process when using these artificial intelligence (AI) tools [61]. Using LLMs to extract data can help avoid random errors;

XSL•FO
RenderX

however, caution is still required when extracting data from figures or tables [47-50].

Assess the Risk of Bias

Assessing the bias of risk involves evaluating the internal validity of studies included in research. For randomized controlled trials, tools such as Risk of Bias (RoB) [62] or its updated version RoB 2 [63] are typically used, with an estimated review time of 10-15 minutes per trial. However, automated tools such as RobotReviewer can streamline the extraction and evaluation process in batches [51-53], thereby improving efficiency, although manual verification is still necessary. Additionally, chatbots based on LLMs can aid in risk of bias assessment (see Multimedia Appendix 8), and their accuracy appears to be comparable to that of human evaluations [23].

Analyze the Data/Meta-Analysis

Data analysis serves as the source of systematic review results, typically encompassing basic information and outcome findings. The meta-analysis may be one outcome, along with potential components such as subgroup analysis, sensitivity analysis, meta-regression, and detection of publication bias. Numerous software options are available to facilitate these data analyses, including Stata, RevMan, Rstudio, and others [43]. Currently, it appears that chatbots based on LLMs may not fully execute data analysis independently, although they can extract the relevant information. Subsequently, one can employ corresponding software for comprehensive data analysis. Alternatively, after extracting information with chatbots, the ChatGPT Code Interpreter can assist in analysis and generating graphical results, although this requires a subscription to ChatGPT Plus. Moreover, an LLM markedly accelerates the data analysis process, empowering researchers to handle larger data sets with greater efficacy [54].

Draft the Full Manuscript

The complete drafting of systematic reviews and meta-analyses should adhere to the PRISMA reporting guidelines [64]. It is not advisable to use chatbots such as ChatGPT for article composition. On the one hand, the accuracy and integrity of content generated by ChatGPT require human verification. On the other hand, various research types and journals have different requirements for full-length articles, making it challenging to achieve uniformity in the generated content. However, using tools such as GPT for language refinement and adjusting the content logic can be considered to enhance the quality and readability of the article [33,55]. It is important to declare the use of GPT-related tools in the methods, acknowledgments, or appendices of the article to ensure transparency [24,65].

Submit and Publish

Submission and publication represent the final steps in the process of conducting systematic reviews and meta-analyses, aside from subsequent updates. At this stage, the potential role of LLM-based tools is to assist authors in recommending suitable journals (see Multimedia Appendix 9). These tools might also aid in crafting components required along with submission of the manuscript such as cover letters and highlights [59]. However, it is imperative to emphasize that the content generated by these tools requires manual verification to ensure accuracy, and all authors should be accountable for the content generated by LLMs.

Benefits and Drawbacks of Using LLMs

Systematic reviews and meta-analyses are crucial evidence types that support the development of guidelines [3]. The benefits of employing LLM-based chatbots in the production of systematic reviews and meta-analyses include increased speed, such as in the stages of evidence searching, data extraction, and assessment of bias risk; these tools can also enhance accuracy by reducing human errors such as those made while extracting essential information and pooling data. However, there are also drawbacks of these applications of LLMs, such as the potential for generating hallucinations, the requirement for human verification owing to the poor reliability of the models, and that the entire systematic review process is not replicable. Moreover, when interacting with LLM chatbots, it is important to manage data privacy. In particular, when using LLMs to analyze data, especially when including personal patient information, ethical approval and management must be properly addressed.

Challenges and Solutions

While LLMs can assist in accelerating the production of systematic reviews and meta-analyses in some steps, enhancing accuracy and transparency, and saving resources, they also face several challenges. For instance, LLMs cannot promptly update their versions and information. For example, ChatGPT 3.5 has been trained on data available in 2021. Thus, limitations such as the length of prompts and token constraints, as well as restrictions related to context associations, may potentially impact the overall results and user experience [25]. Although LLM-based autonomous agents have made strides in tasks related to systematic reviews and meta-analyses, their applications are still associated with various issues related to personalization, updating knowledge, strategic planning, and complex problem-solving. The development of LLM-driven autonomous agents adept at systematic reviews and meta-analyses warrants further exploration [66]. The use of LLMs as centrally controlled intelligent agents encompasses the ability to handle precise literature screening, extract and analyze complex data, and assist in manuscript composition, as highlighted by proof-of-concept demonstrations such as MetaGPT [67]. Moreover, the continuous growth of the use of LLMs can pose a significant challenge in ensuring the accuracy of information provided in systematic reviews, particularly if LLMs are indiscriminately overused.

To better facilitate the use of tools such as ChatGPT in systematic reviews and meta-analyses, we believe that, first and foremost, authors should understand the scope and scenarios for applying ChatGPT, clearly defining which steps can benefit from these tools. Second, for researchers, collaboration with computer scientists and AI engineers is crucial to optimize the prompts and develop integrated tools based on LLMs, such as web applications. These tools can assist in seamless transitions between different tasks in the systematic review process. Lastly, for journal editors, collaboration with authors and reviewers is essential to adhere to reporting and ethical principles associated

XSL•FO
RenderX

with the use of GPT and similar tools [24,68]. This collaboration aims to promote transparency and integrity, while preventing indiscriminate overuse in the application of LLMs in systematic reviews and meta-analyses.

Future Perspectives and Conclusion

The emergence of LLMs could have a significant impact on the production of systematic reviews and meta-analyses. In this process, the application of chatbots such as ChatGPT has the potential to speed up certain steps such as literature screening, data extraction, and risk of bias assessment, which are processes that typically consume a considerable amount of time. However, it is important to note that if AI methods such as GPT are employed in performing systematic reviews, disclosure and declaration of the use of these tools are essential. This includes specifying the AI tools used, their roles, and the areas of application within the review process, among other relevant information for full disclosure [24]. In this context, developing a reporting guideline is warranted to guide the application of LLM tools in systematic reviews and meta-analyses. Although the PRISMA 2020 guideline briefly addresses the use of automation technologies, its coverage is limited to steps such as screening, and there is a lack of comprehensive guidance on the broader spectrum of applications [64].

Acknowledgments

ChatGPT 3.5 designed by OpenAI was used to help with language editing. The authors take the ultimate responsibility for the content of this publication.

Authors' Contributions

XL and YC were responsible for conceptualization of the article. XL, FC, DZ, and LW generated the examples with the large language models and wrote the first draft of the article. XL, ZW, HL, ML, YW, QW, and YC reviewed and edited the manuscript. YC supervised the study, takes full responsibility for the work and conduct of the study, has access to the data, and controlled the decision to publish. All authors read the final manuscript and approved the submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Using ChatGPT 4.0 to assist in generating PubMed search strategies for assessing systematic reviews.
[PNG File , 3022 KB - jmir_v26i1e56780_app1.png]

Multimedia Appendix 2
The results obtained after searching the PubMed database based on the search strategy generated by ChatGPT.
[PNG File , 1409 KB - jmir_v26i1e56780_app2.png]

Multimedia Appendix 3
Using ChatGPT 4.0 to assist in optimizing the clinical question for conducting a systematic review and meta-analysis.
[PNG File , 210 KB - jmir_v26i1e56780_app3.png]

Multimedia Appendix 4
Using ChatGPT 4 to generate PROSPERO (International Prospective Register of Systematic Reviews) registration information.
[PNG File , 482 KB - jmir_v26i1e56780_app4.png]

Multimedia Appendix 5
Proposal of a systematic review and meta-analysis related to exercises for osteoarthritis generated by Claude 3 based on the provided prompts.
[PNG File , 1435 KB - jmir_v26i1e56780_app5.png]

Multimedia Appendix 6
The inclusion and exclusion criteria for a systematic review and meta-analysis on exercise therapy for osteoarthritis based on GPT-4.
[PNG File , 2225 KB - jmir_v26i1e56780_app6.png]

Multimedia Appendix 7
Using Claude 3 for data extraction from PDF documents: an example with three randomized controlled trials.

XSL•FO
RenderX

[PNG File , 3184 KB - jmir_v26i1e56780_app7.png]

Multimedia Appendix 8
Using Claude 3 for risk of bias assessment: an example with two randomized controlled trials.
[PNG File , 4317 KB - jmir_v26i1e56780_app8.png]

Multimedia Appendix 9
Using GPT-4 to assist in selecting target journals for submission of a systematic review and meta-analysis.
[PNG File , 203 KB - jmir_v26i1e56780_app9.png]

References

1. Jahan N, Naveed S, Zeshan M, Tahir MA. How to conduct a systematic review: a narrative literature review. Cureus 2016 Nov 04;8(11):e864 [FREE Full text] [doi: 10.7759/cureus.864] [Medline: 27924252]
2. Wallace SS, Barak G, Truong G, Parker MW. Hierarchy of evidence within the medical literature. Hosp Pediatr 2022 Aug 01;12(8):745-750. [doi: 10.1542/hpeds.2022-006690] [Medline: 35909178]
3. Institute of Medicine, Board on Health Care Services, Committee on Standards for Developing Trustworthy Clinical Practice Guidelines, Graham R, Mancher M, Miller Wolman D, et al. Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Press; 2011.
4. Bystranowski P, Janik B, Próchnicki M, Skórska P. Anchoring effect in legal decision-making: a meta-analysis. Law Hum Behav 2021 Feb;45(1):1-23. [doi: 10.1037/lhb0000438] [Medline: 33734746]
5. Geyskens I, Krishnan R, Steenkamp JEM, Cunha PV. A review and evaluation of meta-analysis practices in management research. J Management 2008 Feb 05;35(2):393-419. [doi: 10.1177/0149206308328501]
6. Bagepally BS, Chaikledkaew U, Chaiyakunapruk N, Attia J, Thakkinstian A. Meta-analysis of economic evaluation studies: data harmonisation and methodological issues. BMC Health Serv Res 2022 Feb 15;22(1):202 [FREE Full text] [doi: 10.1186/s12913-022-07595-1] [Medline: 35168619]
7. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature 2018 Mar 07;555(7695):175-182 [FREE Full text] [doi: 10.1038/nature25753] [Medline: 29517004]
8. Tsertsvadze A, Chen Y, Moher D, Sutcliffe P, McCarthy N. How to conduct systematic reviews more expeditiously? Syst Rev 2015 Nov 12;4:160 [FREE Full text] [doi: 10.1186/s13643-015-0147-7] [Medline: 26563648]
9. Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol 2021 Oct;138:80-94. [doi: 10.1016/j.jclinepi.2021.06.030] [Medline: 34242757]
10. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
11. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev 2023 Apr 29;12(1):72 [FREE Full text] [doi: 10.1186/s13643-023-02243-z] [Medline: 37120563]
12. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.4. 2023 Aug. URL: https://training.cochrane.org/handbook [accessed 2024-06-11]
13. Deeks J, Bossuyt P, Leeflang M, Takwoingi Y. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. 2023 Jul. URL: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current [accessed 2024-06-11]
14. Xiao Y, Watson M. Guidance on conducting a systematic literature review. J Plan Educ Res 2017 Aug 28;39(1):93-112. [doi: 10.1177/0739456x17723971]
15. Muka T, Glisic M, Milic J, Verhoog S, Bohlius J, Bramer W, et al. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. Eur J Epidemiol 2020 Jan;35(1):49-60. [doi: 10.1007/s10654-019-00576-5] [Medline: 31720912]
16. Siddaway AP, Wood AM, Hedges LV. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Annu Rev Psychol 2019 Jan 04;70:747-770. [doi: 10.1146/annurev-psych-010418-102803] [Medline: 30089228]
17. Tawfik GM, Dila KAS, Mohamed MYF, Tam DNH, Kien ND, Ahmed AM, et al. A step by step guide for conducting a systematic review and meta-analysis with simulation data. Trop Med Health 2019;47:46 [FREE Full text] [doi: 10.1186/s41182-019-0165-6] [Medline: 31388330]
18. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545 [FREE Full text] [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
19. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun 2019 Dec;16:100443 [FREE Full text] [doi: 10.1016/j.conctc.2019.100443] [Medline: 31497675]

XSL•FO
RenderX

20. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 2019 Jul 11;8(1):163 [FREE Full text] [doi: 10.1186/s13643-019-1074-9] [Medline: 31296265]
21. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
22. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry 2023 Oct 15;22(3):479-480 [FREE Full text] [doi: 10.1002/wps.21145] [Medline: 37713576]
23. Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. BMJ Health Care Inform 2023 Oct;30(1):e100830 [FREE Full text] [doi: 10.1136/bmjhci-2023-100830] [Medline: 37827724]
24. Luo X, Estill J, Chen Y. The use of ChatGPT in medical research: do we need a reporting guideline? Int J Surg 2023 Dec 01;109(12):3750-3751 [FREE Full text] [doi: 10.1097/JS9.0000000000000737] [Medline: 37707517]
25. Alshami A, Elsayed M, Ali E, Eltoukhy A, Zayed T. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems 2023 Jul 09;11(7):351. [doi: 10.3390/systems11070351]
26. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. Br Dent J 2023 Jul;235(2):90-92. [doi: 10.1038/s41415-023-6132-y] [Medline: 37500847]
27. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 07;13(7):e072254 [FREE Full text] [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
28. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods 2024 Mar 14:online ahead of print. [doi: 10.1002/jrsm.1715] [Medline: 38484744]
29. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation. JMIR Med Inform 2024 May 14;12:e51187-e51187 [FREE Full text] [doi: 10.2196/51187] [Medline: 38771247]
30. Hossain MM. Using ChatGPT and other forms of generative AI in systematic reviews: challenges and opportunities. J Med Imaging Radiat Sci 2024 Mar;55(1):11-12. [doi: 10.1016/j.jmir.2023.11.005] [Medline: 38040497]
31. Giunti G, Doherty CP. Cocreating an automated mHealth apps systematic review process with generative AI: design science research approach. JMIR Med Educ 2024 Feb 12;10:e48949 [FREE Full text] [doi: 10.2196/48949] [Medline: 38345839]
32. Nashwan AJ, Jaradat JH. Streamlining systematic reviews: harnessing large language models for quality assessment and risk-of-bias evaluation. Cureus 2023 Aug;15(8):e43023 [FREE Full text] [doi: 10.7759/cureus.43023] [Medline: 37674957]
33. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res 2023;13(4):1148-1154 [FREE Full text] [Medline: 37168339]
34. Issaiy M, Ghanaati H, Kolahi S, Shakiba M, Jalali AH, Zarei D, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol 2024 Mar 27;24(1):78 [FREE Full text] [doi: 10.1186/s12874-024-02203-8] [Medline: 38539117]
35. Branum C, Schiavenato M. Can ChatGPT accurately answer a PICOT question? Assessing AI response to a clinical question. Nurse Educ 2023;48(5):231-233. [doi: 10.1097/NNE.0000000000001436] [Medline: 37130197]
36. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023 Feb 17;13:01003 [FREE Full text] [doi: 10.7189/jogh.13.01003] [Medline: 36798998]
37. Richard E, Pozzi A. Using ChatGPT to develop the statistical analysis plan for a randomized controlled trial: a case report. Zurich Open Repository and Archive Preprint posted online on October 17, 2023. [doi: 10.21203/rs.3.rs-3433956/v1]
38. Hutson M. How AI is being used to accelerate clinical trials. Nature 2024 Mar;627(8003):S2-S5. [doi: 10.1038/d41586-024-00753-x] [Medline: 38480968]
39. Van der Mierden S, Tsaioun K, Bleich A, Leenaars CHC. Software tools for literature screening in systematic reviews in biomedical research. Altex 2019;36(3):508-517. [doi: 10.14573/altex.1902131] [Medline: 31113000]
40. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? arXiv Preprint posted online on February 9, 2023 [FREE Full text] [doi: 10.1145/3539618.3591703]
41. Alaniz L, Vu C, Pfaff MJ. The utility of artificial intelligence for systematic reviews and Boolean query formulation and translation. Plast Reconstr Surg Glob Open 2023 Oct;11(10):e5339 [FREE Full text] [doi: 10.1097/GOX.0000000000005339] [Medline: 37908326]
42. Guimarães NS, Joviano-Santos JV, Reis MG, Chaves RRM, Observatory of Epidemiology, Nutrition, Health Research (OPENS). Development of search strategies for systematic reviews in health using ChatGPT: a critical analysis. J Transl Med 2024 Jan 02;22(1):1 [FREE Full text] [doi: 10.1186/s12967-023-04371-5] [Medline: 38167166]
43. Tantry TP, Karanth H, Shetty PK, Kadam D. Self-learning software tools for data analysis in meta-analysis. Korean J Anesthesiol 2021 Oct;74(5):459-461 [FREE Full text] [doi: 10.4097/kja.21080] [Medline: 33677944]

XSL•FO
RenderX

44. Cai X, Geng Y, Du Y, Westerman B, Wang D, Ma C, et al. Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. medRxiv Preprint posted online on September 7, 2023. [doi: 10.1101/2023.09.06.23295072]
45. Eugene S, Istvan D, Gauransh K. Assessing the ability of ChatGPT to screen articles for systematic reviews. arXiv Preprint posted online on July 12, 2023. [doi: 10.5860/choice.189890]
46. Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, Samiee R. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. BMJ Evid Based Med 2024 Jan 19;29(1):69-70 [FREE Full text] [doi: 10.1136/bmjebm-2023-112678] [Medline: 37989538]
47. Guo E, Gupta M, Deng J, Park Y, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res 2024 Jan 12;26:e48996 [FREE Full text] [doi: 10.2196/48996] [Medline: 38214966]
48. Polak MP, Morgan D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 2024 Feb 21;15(1):1569. [doi: 10.1038/s41467-024-45914-8] [Medline: 38383556]
49. Mahmoudi H, Chang D, Lee H, Ghaffarzadegan N, Jalali MS. A critical assessment of large language models for systematic reviews: utilizing ChatGPT for complex data extraction. SSRN Preprint posted online on April 19, 2024. [doi: 10.2139/ssrn.4797024]
50. Sun Z, Zhang R, Doi SA, Furuya-Kanamori L, Yu T, Lin L, et al. How good are large language models for automated data extraction from randomized trials? medRXiv Preprint posted online on February 21, 2024. [doi: 10.1101/2024.02.20.24303083]
51. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc 2016 Jan;23(1):193-201 [FREE Full text] [doi: 10.1093/jamia/ocv044] [Medline: 26104742]
52. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the risk of bias in randomized clinical trials with large language models. JAMA Netw Open 2024 May 01;7(5):e2412687 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.12687] [Medline: 38776081]
53. Pitre T, Jassal T, Talukdar JR, Shahab M, Ling M, Zeraatkar D. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: a methods study. medRXiv Preprint posted online on November 22, 2023. [doi: 10.1101/2023.11.19.23298727]
54. Rasheed Z, Waseem M, Ahmad A, Kemell KK, Xiaofeng W, Nguyen Duc A, et al. Can large language models serve as data analysts? A multi-agent assisted approach for qualitative data analysis. arXiv Preprint posted online on February 2, 2024.
55. Kim S. Using ChatGPT for language editing in scientific articles. Maxillofac Plast Reconstr Surg 2023 Mar 08;45(1):13 [FREE Full text] [doi: 10.1186/s40902-023-00381-x] [Medline: 36882591]
56. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
57. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol 2023 Oct;19(5):598-604. [doi: 10.1016/j.jpurol.2023.05.018] [Medline: 37328321]
58. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. J Med Internet Res 2024 Apr 05;26:e52935 [FREE Full text] [doi: 10.2196/52935] [Medline: 38578685]
59. Firdausa Nuzula I, Miftahul Amri M. Will ChatGPT bring a new paradigm to HR World? A critical opinion article. J Manag Stud Devel 2023 Apr 24;2(02):142-161. [doi: 10.56741/jmsd.v2i02.316]
60. Eriksen MB, Frandsen TF. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. J Med Libr Assoc 2018 Oct;106(4):420-431 [FREE Full text] [doi: 10.5195/jmla.2018.345] [Medline: 30271283]
61. Roberts R. I tested how well ChatGPT can pull data out of messy PDFs. Source. 2023 Mar 01. URL: https://source.opennews.org/articles/testing-pdf-data-extraction-chatgpt/ [accessed 2024-06-11]
62. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
63. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
64. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
65. Gaggioli A. Ethics: disclose use of AI in scientific manuscripts. Nature 2023 Feb 14;614(7948):413-413. [doi: 10.1038/d41586-023-00381-x] [Medline: 36788370]

XSL•FO
RenderX

66. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. Front Comput Sci 2024 Mar 22;18(6):186345 [FREE Full text] [doi: 10.1007/s11704-024-40231-1]
67. Hong S, Zhuge M, Chen J, Zheng X, Cheng Y, Zhang C, et al. Metagpt: Meta programming for multi-agent collaborative framework. arXiv Preprint posted online on November 6, 2023.
68. Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting use of AI in research and scholarly publication-JAMA Network Guidance. JAMA 2024 Apr 02;331(13):1096-1098. [doi: 10.1001/jama.2024.3471] [Medline: 38451540]

Abbreviations

AI: artificial intelligence
API: application programming interface
LLM: large language model
PICO: Population, Intervention, Comparison, Outcome
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
RoB: Risk of Bias

Edited by G Eysenbach; submitted 20.02.24; peer-reviewed by A Jafarizadeh, M Chatzimina, AS Van Epps; comments to author 03.05.24; revised version received 21.05.24; accepted 29.05.24; published 25.06.24.

Please cite as:
Luo X, Chen F, Zhu D, Wang L, Wang Z, Liu H, Lyu M, Wang Y, Wang Q, Chen Y
Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses
J Med Internet Res 2024;26:e56780
URL: https://www.jmir.org/2024/1/e56780
doi:10.2196/56780
PMID:38819655

©Xufei Luo, Fengxian Chen, Di Zhu, Ling Wang, Zijun Wang, Hui Liu, Meng Lyu, Ye Wang, Qi Wang, Yaolong Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A Nordic Perspective on Patient Online Record Access and the European Health Data Space

Maria Hägglund[1,2], PhD; Anna Kharko[1,3], PhD; Annika Bärkås[1], PhD; Charlotte Blease[1,4], PhD; Åsa Cajander[5], PhD; Catherine DesRoches[4], DrPH; Asbjørn Johansen Fagerlund[6], PhD; Josefin Hagström[1], MSc; Isto Huvila[7], PhD; Iiris Hörhammer[8], DSC; Bridget Kane[1,9], PhD; Gunnar O Klein[10], MD, PhD; Eli Kristiansen[6], MEc; Jonas Moll[10], PhD; Irene Muli[1], MSc; Hanife Rexhepi[11], PhD; Sara Riggare[1], PhD; Peeter Ross[12,13], PhD, MD; Isabella Scandurra[10], PhD; Saija Simola[8], MSc; Hedvig Soone[12], MSc; Bo Wang[6], BMed, MPH; Maedeh Ghorbanian Zolbin[8], MBA; Rose-Mharie Åhlfeldt[11], PhD; Sari Kujala[8*], PhD; Monika Alise Johansen[6*], PhD

[1]Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
[2]Medtech Science & Innovation Centre, Uppsala University Hospital, Uppsala, Sweden
[3]School of Psychology, Faculty of Health, University of Plymouth, Plymouth, United Kingdom
[4]Division of General Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States
[5]Department of Information Technology, Uppsala University, Uppsala, Sweden
[6]Norwegian Centre for E-Health Research, University Hospital of North Norway, Tromsø, Norway
[7]Department of ALM, Uppsala University, Uppsala, Sweden
[8]Department of Computer Science, Aalto University, Espoo, Finland
[9]Business School, Karlstad University, Karlstad, Sweden
[10]Centre for Empirical Research on Information Systems, School of Business, Örebro University, Örebro, Sweden
[11]School of Informatics, University of Skövde, Skövde, Sweden
[12]E-Medicine Centre, Department of Health Technologies, Tallinn University of Technology, Tallinn, Estonia
[13]Research Department, East Tallinn Central Hospital, Tallinn, Estonia
[*]these authors contributed equally

Corresponding Author:
Maria Hägglund, PhD
Participatory eHealth and Health Data Research Group
Department of Women's and Children's Health
Uppsala University
Dag Hammarskjölds Väg 14b, 1 Tr
Uppsala, 752 37
Sweden
Phone: 46 729999381
Email: maria.hagglund@uu.se

Abstract

The Nordic countries are, together with the United States, forerunners in online record access (ORA), which has now become widespread. The importance of accessible and structured health data has also been highlighted by policy makers internationally. To ensure the full realization of ORA's potential in the short and long term, there is a pressing need to study ORA from a cross-disciplinary, clinical, humanistic, and social sciences perspective that looks beyond strictly technical aspects. In this viewpoint paper, we explore the policy changes in the European Health Data Space (EHDS) proposal to advance ORA across the European Union, informed by our research in a Nordic-led project that carries out the first of its kind, large-scale international investigation of patients' ORA—NORDeHEALTH (Nordic eHealth for Patients: Benchmarking and Developing for the Future). We argue that the EHDS proposal will pave the way for patients to access and control third-party access to their electronic health records. In our analysis of the proposal, we have identified five key principles for ORA: (1) the right to access, (2) proxy access, (3) patient input of their own data, (4) error and omission rectification, and (5) access control. ORA implementation today is fragmented throughout Europe, and the EHDS proposal aims to ensure all European citizens have equal online access to their health data. However, we argue that in order to implement the EHDS, we need more research evidence on the key ORA principles we have

XSL•FO
RenderX

identified in our analysis. Results from the NORDeHEALTH project provide some of that evidence, but we have also identified important knowledge gaps that still need further exploration.

(J Med Internet Res 2024;26:e49084) doi:10.2196/49084

KEYWORDS

patients' online record access; open notes; electronic health records; EHR; patient portals; European Health Data Space; digital health; health care; patient access

Introduction

The digitalization of health care is increasing rapidly, changing the way patients communicate and collaborate with health care providers. The importance of access to and use of digital health data became even more evident during the COVID-19 pandemic [1], including the use of online patient portals and patients' online record access (ORA) [2]. The Nordic countries are, together with the United States, forerunners in providing their residents with online tools that enable interaction not only with health care but also with the patient's own health data [3]. ORA has become a key means to these ends [4-6]. National patient portals have been implemented in all Nordic countries, enabling residents to access different health-related e-services, for example, patient-accessible electronic health records (PAEHRs). In the United States, legislation mandates patients to have ORA [7]. Textbox 1 provides an overview of key concepts related to patients' ORA that will be used in this paper.

Textbox 1. Key terminology.

Electronic health record (EHR)

The World Health Organization defines EHRs as "shared patient records that contain historical data about a patient that are compiled from all local Electronic Medical Records" [8].

Patient-accessible EHR (PAEHR)

PAEHRs are online services providing patients secure access to view and sometimes edit or comment on their electronic health records (EHRs) made available by their health care providers [5], that is, online record access (ORA).

European Health Data Space (EHDS)

The EHDS is a health-specific ecosystem comprising rules, common standards and practices, infrastructure, and a governance framework [9].

Open notes

Online access to the visit note summaries, or the narrative, free-text entries, written by clinicians about patient health.

ORA

ORA has been used as a "solution-neutral" concept to describe the phenomenon of patients' online record access [10]. ORA can be implemented through a PAEHR or any other technical system that gives patients access to their health records online.

Patient portal

Patient portals are online portals that can be provided locally by a specific health care provider or nationally, as is the case in the Nordic countries. Patient portals are increasingly used to provide patients with ORA. In some patient portals, a PAEHR is provided as a specific service [11], whereas others may have more seamlessly integrated ORA through different patient portal functions. In a local patient portal, patients often have ORA to only 1 specific EHR system, whereas national patient portals can provide ORA to several EHR systems.

In many countries, ORA is considered a logical extension of patients' already existing legal rights to request copies of their health records. ORA provides a rapid and convenient method of accessing the information held by clinicians that increases the total number of patients who read their records. Considering the growing body of evidence presenting the benefits of ORA for the individual (in terms of improved health outcomes and self-management) [4-6,12,13], we argue that health organizations in other countries can learn from the Nordic experience and should also consider striving to provide patients' ORA [14].

In parallel with the increased use of digital health services, the importance of accessible and structured health data has also been highlighted by policy makers internationally. In the United States, a federal rule in the 21st Century Cures Act mandates that US health care providers offer patients access, with few exceptions, to all the health information in their electronic medical records without charge [7]. The 21st Century Cures Act is also motivated by the idea of a health app economy, and it is mandated that patients' health information be available in a form that is downloadable to third-party apps. In Europe, the proposal for an EHDS aims to both empower people to control and use their health data in their home country or other member states, as well as offer "a consistent, trustworthy and efficient framework to use health data for research, innovation, policy making, and regulatory activities, while ensuring full compliance with the EU's high data protection standards" [9]. Furthermore, the EU has adopted the NIS directive (EU 2022/255), which sets requirements for security in networks and information systems. The rules cover providers of socially important services and certain digital services where health care is a designated sector [15].

Despite growing international evidence that ORA has the potential to empower patients and yield many health benefits,

XSL•FO
RenderX

its implementation has not always been straightforward [16,17]. Not all patients use online portals [4,5,12,18], and ORA remains controversial among providers [10,19]. Health care professionals have raised concerns regarding patient ORA in several contexts where it has been implemented [10,20-23]. The concerns include that patients might misunderstand what they read and become worried, and that clinicians' workload will increase as patients ask more questions, both during and between appointments. Within mental health care, such concerns have been especially prominent [19,22,24].

Despite clinicians' concerns and the need for a more fine-grained policy concerning, for example, proxy access (when an informal caregiver, such as a family member, has ORA on behalf of a patient) [25], psychiatric care [26], and ethical exemptions from ORA [27], mounting international experience challenges clinicians' skepticism and evinces the benefits to patients of this practice innovation [28]. Nonetheless, as the shift toward giving patients more autonomy over their health data is underway, there is an urgent need to address more contested aspects of ORA. Doing so may simultaneously offer guidance to other countries, where implementation is currently lagging behind. We argue that to ensure the full realization of its potential in the short and long term, there is a pressing need to study patients' ORA from a cross-disciplinary, technical, clinical, humanistic, and social science perspective that looks beyond narrow technical aspects of implementation [29]. A project that aims to do this is the NORDeHEALTH research project, launched in 2021 [3]. NORDeHEALTH focuses on studying novel digital services and innovation, exploring different ways to make national patient portals and patients' ORA more useful to both patients and health care professionals, supporting person-centered care, patient self-management, and empowerment, as well as collaboration.

In this viewpoint paper, we will summarize the key policy changes in the EHDS with relevance to ORA and discuss the proposed changes in the context of the latest research findings from the Nordic Region through the NORDeHEALTH project.

The European Health Data Space and Patient Online Record Access

The EHDS proposal aims to "improve access to, and control by, natural persons over their personal electronic health data in the context of health care (primary use of electronic health data), as well as for other purposes that would benefit the society such as research, innovation, policy-making, patient safety, personalized medicine, official statistics or regulatory activities (secondary use of health data)" [9]. "Natural person" is a legal term used to signify an individual human being, distinguishing them from a "juridical person," which can encompass other entities too. We will use "person" in our text to signify a "natural person," unless in a direct quote from the EHDS proposal. Negotiations among EU member states have been completed, and in April 2024, the European Parliament adopted the EHDS proposal. Once the new EHDS regulation is formally adopted, which is expected to be in autumn 2024, it will become applicable in different stages according to use case and data type, allowing member states time to adapt to the regulation [30].

Primary use of electronic health data is the main focus of patients' ORA and is further defined as the processing of personal electronic health data for the provision of health services to assess, maintain, or restore the state of health of the natural person to whom that data relates, including the prescription, dispensation, and provision of medicinal products and medical devices, as well as for relevant social security, administrative, or reimbursement services.

Embedded within the EHDS Chapter 2, Primary use of electronic health data, Section 1, Article 3 describes the "rights of natural persons in relation to the primary use of their personal electronic health data" [9]. In our analysis of the EHDS proposal, we have identified 5 key principles of high relevance for patient ORA, where results from the NORDeHEALTH project may contribute to the design and implementation of the EHDS across Europe. The five principles we have identified are (1) the right to access, (2) proxy access, (3) patient input of their own data, (4) error and omission rectification, and (5) access control.

Overview of the NORDeHEALTH Research Project

The NORDeHEALTH project, funded by NordForsk (grant 100477), has research partners in Sweden, Norway, Finland, Estonia, and the United States. The goal is to enable further digitalization of the public health sector by providing concrete feedback to the national authorities in the respective countries and providing guidelines and frameworks for the design, implementation, and evaluation of patients' ORA as well as other eHealth services.

The foundation for the NORDeHEALTH research project is a sociotechnical analysis of the context in the respective country using a model proposed previously [29,31]. The model defines eight dimensions that are essential to consider when designing and implementing sociotechnical systems in health care: (1) hardware and software, (2) clinical content, (3) human-computer interface, (4) people, (5) workflow and communication, (6) internal organizational policies, procedures, and culture, (7) external rules, regulations, and pressures, and (8) system measurement and monitoring [31]. Key focus areas in the NORDeHEALTH research project have included policy and regulations for patients' ORA in mental health care [32-34], ORA for adolescents and parents [25,35], and ORA within oncology, specifically focusing on multidisciplinary team conferences. The project also investigates benchmarking for the usability and acceptance of national patient portals and patients' ORA [6,36], which iteratively feeds into the co-design of novel ORA and patient portal functionality.

Empirical data in the project is gathered by exploring the needs of specific patient and demographic groups using the current implementations of PAEHRs as a case; patients with mental health disorders, patients with cancer, and adolescents and their caregivers. Most research into patient ORA to date collects data from one country or region [4-6,12], making it difficult to

XSL•FO
RenderX

compare results across contexts. In the NORDeHEALTH project, we therefore designed an international cross-sectional survey study, and in 2022, data were collected simultaneously in Sweden, Norway, Finland, and Estonia [37]. The survey aimed to study patients' experiences with the PAEHR provided through the national patient portals in the respective countries.

Textbox 2 provides an overview of how the NORDeHEALTH research contributes to the patient ORA principles we have identified in the EHDS proposal: (1) the right to access, (2) proxy access, (3) patient input of their own data, (4) rectifying errors and omissions, and (5) access control. In the next sections, we further deepen the analysis of the 5 principles.

Textbox 2. Online record access (ORA) principles in the context of recent research from the Nordics.

The right to access

- Despite the many similarities, ORA implementation varies in the Nordic region.
- Patients' experiences depend on platform usability.

Proxy access

- Current regulations for parental and adolescent proxy access greatly differ between the Nordic countries.
- Proxy access other than parental has even greater variation, and is not allowed in, eg, Sweden.

Patient input of their own data

- Despite the advanced stage of ORA in the Nordic countries, patient input is not widely available.
- Nordic patients have expressed repeated interest in the ability to contribute to their record.

Error and omission rectification

- A high number of Nordic patients find serious errors in their records through ORA.
- Some groups of patients report more errors than others.
- At present, only a minority of Nordic patients attempt to rectify the errors.

Access control

- A minority of Nordic patients have reported unwanted access to their electronic health records

The Right to Access

The Proposal

The EHDS proposal clearly states that "natural persons shall have the right to access their personal electronic health data" immediately, free of charge, and in an easily readable, consolidated, and accessible form [9]. This is not limited to electronic health records (EHRs) data, but considering the EHR's core role for documentation in health care, patients' ORA must be considered essential for EHDS. Article 3, paragraph 2, continues to declare that:

> *natural persons shall have the right to receive an electronic copy, in the European electronic health record exchange format […] of at least their electronic health data in the priority categories referred to in Article 5 [9].*

These include patient summaries, electronic prescriptions, electronic dispensations, medical images and image reports, laboratory results, and discharge reports.

The Research

Many European countries already have legislation stipulating that patients should have ORA [16]. In Norway, the patient is both the object and the owner of the health record. The Norwegian Patient Rights Act of 2001 states that patients have the right by law to access their health records [38], and in 2013, a White Paper stated that patients should have digital access [39]. Similar legislation is in place in all the Nordic countries [16].

In Germany, the Patient's Rights Act of 2013 stipulates that health care professionals must document diagnosis and treatment promptly and comprehensively. It grants patients the right to fully view their records and attain an electronic copy [40], yet progress in implementing patients' ORA has been slow. In the Netherlands, patients have had the right to a digital copy of all the information in their EHRs since 2020 [17], and different types of incentive programs have been implemented to encourage health care providers to provide such access.

Given the current challenges in implementing patients' full ORA across Europe, the EHDS proposal is ambitious. Mandating patients' ORA is to be broadly encouraged, considering the positive experiences reported by patients with full ORA, but experiences show that regulations are often not enough to ensure implementation.

The NORDeHEALTH project has designed and tested a sociotechnical framework for studying and comparing factors that affect the implementation and adoption of patient ORA. The framework was designed based on the existing Sittig and Singh sociotechnical framework [31]. ORA-specific factors are explored related to, for example, what information patients have access to and when they can access it (immediately or with a delay), what functionality is provided (eg, being able to upload

XSL•FO
RenderX

or edit information, proxy access), rules and regulations for ORA (on national or local levels), the usability of the PAEHR, technical infrastructure, and population characteristics (eg, educational levels, digital literacy, and diversity). An in-depth understanding of the local sociotechnical context is essential for comprehending the impact of ORA and being able to design successful interventions for implementing ORA. A study on the implementation of ORA in Sweden and the Netherlands identified resistance from health care professionals and technical infrastructure challenges as main barriers, whereas existing national infrastructure and program management, strong leadership, and stakeholder engagement (including both patients and health care professionals) were identified as success factors [17].

Viewpoint

We argue that simply enabling ORA is not enough to ensure that all patients can use it. Usability is a key factor in the adoption of ORA [41]. Therefore, the NORDeHEALTH project strives to benchmark PAEHR usability [6] and investigate how it affects the acceptance and adoption of PAEHRs among different patient groups.

Proxy Access

The Proposal

In the EHDS proposal, proxy access is described in Article 3, paragraph 5. Member States shall:

> *establish one or more proxy services enabling a natural person to authorize other natural persons of their choice to access the electronic health data on their behalf [9].*

This includes guardians and other representatives, either automatically or upon request.

The Research

With a growing population of older people (>60 years) seeking health care services, many of whom are likely to have (multiple) chronic conditions, the demands on health and social care services are increasing [42]. However, the time that individuals with chronic conditions seek health and social care represents only a fragment of their 24/7 lived experience of coping with a chronic disease. As we grow older, we often become increasingly dependent on psychosocial and physical support outside of formal health care services, but this is far from only an issue for older people. Patients with cognitive or physical disabilities often rely on family or informal caregivers for support in managing their health. Parents, especially of children with chronic or life-threatening conditions, have an instrumental role in their children's care and report great benefits from ORA [35] when it is available. Having a (strong) social network and informal caregivers (eg, family and friends), especially in times of life-changing illness, could mean the difference between survival and death [43]

Despite informal caregiving being an essential part of health care, it is rarely given a lot of credit. In fact, the vast majority of caregiving is informal, and it is undertaken by family members free of charge and with no support provided for them, often at great burden [44,45]. Thus, health care outcomes highly depend on the competence and ability of informal caregivers. Still, informal caregivers are often left out of the conversation [46], not the least when digitalization is introduced, and informal communication needs to be formalized.

In the NORDeHEALTH project, research into proxy access focuses on how parental ORA differs between the participating countries [25] and what this could mean for streamlining proxy access across Europe. Further studies on parental proxy access and adolescents' ability to deny access in certain situations are in process. For general proxy access, there is even greater diversity across countries and even less research available. Internationally, when patients are given online access to their records, they are often given the option to share their records with a proxy if needed, usually a close family member such as a spouse or adult son or daughter [47]. In a US study, 2 out of 3 surveyed hospitals offered adult patients the option of granting portal access to an informal caregiver, but among hospitals that did, the process for obtaining proxy credentials was often difficult and time-consuming [48]. In the original Swedish implementation of ORA, patients could assign a proxy to be able to access their records. This function was available after secure log-in to the record, and the patient could assign access to any person in Sweden by adding the social security number of the person and choosing what parts of the PAEHR to share [49]. Despite this flexibility of the solution, the Swedish Authority for Privacy Protection requested the function be shut down, and after several appeals from Region Uppsala, the Supreme Administrative Court in Sweden finally decided to prohibit the function where patients can share their information with others, finding it to be in conflict with the Patient Data Act (a part of the Swedish Data Protection Act (2018:218) and the Swedish Data Protection Regulation (2018:219) that entered into force on May 25, 2018 [50]), which refers to allowing only patients themselves direct access to their medical records—not someone else [49].

Viewpoint

In order to implement the EHDS proposal, it will be essential to streamline regulations for, and implementation of, proxy access across Europe. Acknowledging that different types of proxy access exist and come with their own set of challenges will also be important, distinguishing, for example, parental proxy access from other forms of proxy access.

Rectifying Errors and Omissions

The Proposal

The EHDS proposal, Article 3 §7, stresses that Member States shall ensure that:

> *natural persons can easily request rectification online [9]*

The Research

In a US study of 22,000 patients who read their notes, 1 in 5 reported finding an error, and 40% of those perceived the error to be serious [51]. The most common errors were related to diagnoses, medical histories, medications, test results, notes on

XSL•FO
RenderX

the wrong patient, and notes pertaining to the wrong side of the patient's body (left vs right). Erroneous records may contribute to diagnostic errors that are common in health care [52]. Between 1 in 20 and 1 in every 6 medical consultations results in missed, wrong, or delayed diagnoses [53]. Most diagnostic errors relate to common conditions such as congenital heart failure, pneumonia, and urinary tract infections [54]. Research also shows repeated, missed opportunities to detect cancer [55,56].

Patients have so far had a marginal role in diagnostic processes, as acknowledged in the US National Academy of Medicine's report "Improving Diagnosis in Health Care." The report prompts a deeper discussion about the role of patients in closing feedback loops in care and helping to avoid mistakes that can lead to diagnostic errors, and ultimately patient harm [57]. Patient ORA is cited as a mechanism for improving diagnostic accuracy [57] and has been described by medical safety experts as a "transforming concept" in patient safety [58]. Emerging research supports these conclusions [13,59,60]. Patient ORA may help patients avoid delays and missed diagnoses by encouraging timely follow-up of recommended tests, results, and referrals [51]. Patients with ORA who identify and report errors could prevent clinicians from relying on incorrect data that may lead to poor diagnostic or treatment decisions or even legal liability [61]. A meta-analysis of 20 ORA-related randomized clinical trials (involving 17,387 patients) supports the conclusion that ORA could improve patient safety [13]. Most research to date on patients' ORA and documentation errors has been performed in the United States with a remarkably different medico-legal system from the European one. In this respect, the NORDeHEALTH research complements the existing research and provides important evidence for the usefulness of patient ORA in patient safety work in contexts dominated by public health care provision.

In addition, most studies focus on somatic care and exclude mental health care. A recent NORDeHEALTH study made a comparison between patients who had received mental health care (the MHC group) and patients who had not (non-MHC group), regarding their experiences of finding errors or missing information in their online records [62]. MHC respondents (n=3131) experienced errors (MHC 1586/3131, 50.65% and non-MHC 3311/9203, 35.98%) and omissions (MHC 1089/3131, 34.78% and non-MHC 2427/9203, 26.37%) in the EHR at a higher rate compared with non-MHC respondents (n=9203). The statistically significant differences between the MHC and non-MHC groups remained when comparing a stratified subgroup sample adjusted for age and gender [62].

Viewpoint

In the NORDeHEALTH patient survey [37], we explore the extent to which patients in Sweden, Norway, Finland, and Estonia find errors or missing information in their PAEHR and the action they have taken in these situations [62]. The results from these studies will help guide the further implementation of the EHDS with respect to the management of errors and missing information.

Patients' Input of Their Own Data

The Proposal

Article 3, §6 of the EHDS proposal states that:

> *Natural persons may insert their electronic health data into their own EHR or in that of natural persons whose health information they can access, through electronic health data access services or applications linked to these services. That information shall be marked as inserted by the natural person or by his or her representative [9].*

The Research

Although the Nordic countries are advanced in providing ORA, entering health data into the EHR is not widely implemented. Swedish patients could previously comment in their EHRs [63], but this function was removed in 2022 due to technical problems related to the initial implementation of the feature. In Finland, patients can save health data to their personal health record via well-being applications, but the function is in limited use, only certain applications are accepted, and the data are not yet available to health care professionals [64]. Furthermore, patients in both Finland and Sweden have asked for more interactivity in their health records, such as the possibility to comment on the notes or request corrections [4,6].

In the NORDeHEALTH project, we explore how patient input to the EHR might become better designed to adequately meet this function. The EHR has traditionally been available to health care professionals only. Patients report many positive effects from accessing their records, yet to fully achieve the potential benefits of digitalization, we need to further explore how EHRs can shift from being solely a documentation tool for health professionals to a tool for secure collaboration and communication with patients and family caregivers. Here, national patient portals and additional digital services will complement the future development of the EHRs into collaborative, person-centered tools.

Viewpoint

As digitalization is about more than making electronic versions of analog work, we investigate different ways to use the power of digitalization, as provided in Textbox 3.

Textbox 3. Different ways to incorporate patient input to the electronic health record (EHR).

- Patient input to the EHR in narrative form, for example, patients commenting on notes or contributing with descriptions of their symptoms.
- Patient-created structured data, for example, patient-reported outcome or experience measures—patient reported outcome measures and patient reported experience measures, and Integration of data and services from third-party applications, for example, self-tracking data, and decision support.

XSL•FO
RenderX

Access Control

The Proposal

Finally, EHDS proposes increased access control for patients. Patients should be able to request that electronic health data are made accessible to actors in the health or social security sector, and they should also have the right to restrict such access to electronic health data.

> *Natural persons shall have the right to give access to or request a data holder from the health or social security sector to transmit their electronic health data to a data recipient of their choice from the health or social security sector, immediately, free of charge, and without hindrance from the data holder or from the manufacturers of the system used by that holder [Article 3, §8] [9]*
>
> *[...] natural persons shall have the right to restrict access of health professionals to all or part of their electronic health data [Article 3, §9] [9]*

The Research

The Swedish Agency for Health and Care Services Analysis [65] reports that the majority of the population accepts and wants digital data about their own care and health to be used so that it is useful, including for safer care and research. At the same time, it is important that the data be handled securely and protected from unauthorized access.

In the NORDeHEALTH 2022 patient survey [37], questions related to sharing of the respondents' records were asked, as well as requests for unwanted record access. Among the Swedish respondents, 4% (501/12,334) of respondents answered that they had experienced that someone had seen their health records without their consent [66], a finding that requires further analysis. Although 4% can be considered a low number, it stands in stark contrast to the clear message from policy makers that unauthorized access should not occur [9,50].

With the increasing possibilities for secondary use of health data, both by research and industry stakeholders, it will become even more important for patients to be aware of how to manage their health data, and consent to sharing it, in a safe way. Further research is needed both to understand patients' incentives for and experiences of sharing their data for secondary use and to identify interventions to increase digital health literacy regarding secondary use specifically.

Viewpoint

Unauthorized access can reduce both patient safety and trust in health care, but it can also erode opportunities for secondary use of health data. Existing controls for information security and privacy, therefore, need to be improved in line with the EHDS proposal.

Conclusions

We argue that with the realization of the EHDS, patients' opportunities to access and control third-party access to their EHRs are likely to change dramatically. ORA implementation today is fragmented throughout Europe, and the EHDS proposal aims to ensure all European citizens have equal online access to their health data. However, we argue that in order to implement the EHDS, we need more research evidence on the key ORA principles we have identified in our analysis. Results from the NORDeHEALTH project provide some of that evidence, but we have also identified important knowledge gaps that still need further exploration. Research such as that performed in the NORDeHEALTH project offers important firsthand insights and will be essential to inform the design and implementation of ORA to meet the requirements of the EHDS. However, further international collaboration and research, and dedicated funding, are needed to achieve a comprehensive understanding of sociotechnical and contextual factors necessary to consider for ensuring successful, secure, and ethical implementation of EHDS.

Acknowledgments

This work was supported by the Citizen and Health Data Working Group in the European Federation for Medical Informatics (CHD WG, EFMI). This work was supported by NordForsk through the funding to Nordic eHealth for Patients: Benchmarking and Developing for the Future, NORDeHEALTH (project #100477), the Swedish Research Council for Health, Working Life and Welfare (FORTE) through the funding to Beyond Implementation of eHealth (project #2020-01229), the Strategic Research Council at the Academy of Finland (projects #352501 and #352503), and the Norwegian Centre for E-health Research. CB was supported by a Keane Scholar Award. The study funders played no role in the study design, data interpretation, writing of the results, or decision to submit the manuscript for publication.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

MH and AK prepared the initial draft. All authors participated in revising the text and approved the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022;29(3):453-460 [FREE Full text] [doi: 10.1093/jamia/ocab268] [Medline: 34888680]
2. Hägglund M, McMillan B, Whittaker R, Blease C. Patient empowerment through online access to health records. Br Med J 2022;378:e071531 [FREE Full text] [doi: 10.1136/bmj-2022-071531] [Medline: 36175012]
3. Maria Hägglund: Nordic countries lead new initiative on patient access to EHRs. Br Med J. URL: https://blogs.bmj.com/bmj/2021/05/18/maria-hagglund-nordic-countries-lead-new-initiative-on-patient-access-to-ehrs/ [accessed 2022-01-16]
4. Moll J, Rexhepi H, Cajander, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their Electronic Health Records: National Patient Survey in Sweden. J Med Internet Res 2018;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
5. Zanaboni P, Kummervold PE, Sørensen T, Johansen MA. Patient use and experience with online access to Electronic Health Records in Norway: results from an online Survey. J Med Internet Res 2020;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
6. Kujala S, Hörhammer I, Väyrynen A, Holmroos M, Nättiaho-Rönnholm M, Hägglund M, et al. Patients' experiences of web-based access to Electronic Health Records in Finland: cross-sectional Survey. J Med Internet Res 2022;24(6):e37438 [FREE Full text] [doi: 10.2196/37438] [Medline: 35666563]
7. Salmi L, Blease C, Hägglund M, Walker J, DesRoches CM. US policy requires immediate release of records to patients. Br Med J 2021;372:n426. [doi: 10.1136/bmj.n426] [Medline: 33602667]
8. Digital health platform handbook: building a digital information infrastructure (infostructure) for health. World Health Organization, Union IT. URL: https://apps.who.int/iris/handle/10665/337449 [accessed 2023-03-24]
9. European Health Union: a European Health Data Space for people and science. European Commission. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2711 [accessed 2022-05-08]
10. Blease C, Torous J, Dong Z, Davidge G, DesRoches C, Kharko A, et al. Patient online record access in English primary care: qualitative survey study of general practitioners' views. J Med Internet Res 2023;25:e43496 [FREE Full text] [doi: 10.2196/43496] [Medline: 36811939]
11. Hägglund M, Scandurra I. Patients' online access to electronic health records: current status and experiences from the implementation in Sweden. Stud Health Technol Inform 2017;245:723-727. [Medline: 29295193]
12. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' Outpatient Visit Notes. J Med Internet Res 2019;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
13. Neves AL, Freise L, Laranjo L, Carter AW, Darzi A, Mayer E. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. BMJ Qual Saf 2020;29(12):1019-1032 [FREE Full text] [doi: 10.1136/bmjqs-2019-010581] [Medline: 32532814]
14. Kharko A, Blease C, Johansen M, Moen A, Scandurra I, McMillan B, et al. Mapping patients' online record access worldwide: preliminary results from an international survey of healthcare experts. Stud Health Technol Inform 2024;310:114-118. [doi: 10.3233/shti230938] [Medline: 38269776]
15. Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive). URL: https://eur-lex.europa.eu/eli/dir/2022/2555/oj [accessed 2023-03-28]
16. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient access to electronic health records: Differences across ten countries. Health Policy and Technology 2018 Mar;7(1):44-56 [FREE Full text] [doi: 10.1016/j.hlpt.2017.11.003] [Medline: 22925029]
17. Cijvat CD, Cornet R, Hägglund M. Factors influencing development and implementation of patients' access to electronic health records—a comparative study of Sweden and the Netherlands. Front Public Health 2021;9:621210 [FREE Full text] [doi: 10.3389/fpubh.2021.621210] [Medline: 34169054]
18. Kainiemi E, Vehko T, Kyytsönen M, Hörhammer I, Kujala S, Jormanainen V, et al. The factors associated with nonuse of and dissatisfaction with the National Patient Portal in Finland in the era of COVID-19: population-based cross-sectional survey. JMIR Med Inform 2022;10(4):e37500 [FREE Full text] [doi: 10.2196/37500] [Medline: 35404831]
19. Petersson L, Erlingsdóttir G. Open notes in Swedish Psychiatric Care (Part 2): survey among psychiatric care professionals. JMIR Ment Health 2018;5(2):e10521 [FREE Full text] [doi: 10.2196/10521] [Medline: 29929946]
20. Grünloh C, Myreteg G, Cajander A, Rexhepi H. "Why Do They Need to Check Me?" Patient participation through eHealth and the doctor-patient relationship: qualitative study. J Med Internet Res 2018;20(1):e11 [FREE Full text] [doi: 10.2196/jmir.8444] [Medline: 29335237]
21. Grünloh C, Cajander Å, Myreteg G. "The Record is Our Work Tool!"-physicians' framing of a patient portal in Sweden. J Med Internet Res 2016;18(6):e167 [FREE Full text] [doi: 10.2196/jmir.5705] [Medline: 27349531]
22. Kristiansen E, Johansen M, Zanaboni P. Healthcare personnels' experience with patients' online access to electronic health records: differences between professions, regions, and somatic and psychiatric healthcare. : Linköping Electronic Conference

XSL•FO
RenderX

Proceedings; 2019 Presented at: Scandinavian Conference on Health Informatics; 2019; Norway p. 93-98 URL: https://ep.liu.se/en/conference-article.aspx?series=&issue=161&Article_No=16
23. Johansen MA, Kummervold PE, Sørensen T, Zanaboni P. Health professionals' experience with patients accessing their electronic health records: results from an online survey. Stud Health Technol Inform 2019;264:504-508. [doi: 10.3233/SHTI190273] [Medline: 31437974]
24. Petersson L, Erlingsdóttir G. Open notes in Swedish Psychiatric Care (Part 1): survey among psychiatric care professionals. JMIR Ment Health 2018;5(1):e11 [FREE Full text] [doi: 10.2196/mental.9140] [Medline: 29396386]
25. Hagström J, Scandurra I, Moll J, Blease CR, Haage B, Hörhammer I, et al. Minor and parental access to electronic health records: differences across four countries. Stud Health Technol Inform 2022;294:495-499. [doi: 10.3233/SHTI220508] [Medline: 35612129]
26. Blease C, Torous J, Kharko A, DesRoches CM, Harcourt K, O'Neill S, et al. Preparing patients and clinicians for Open Notes in Mental Health: qualitative inquiry of international experts. JMIR Ment Health 2021;8(4):e27397 [FREE Full text] [doi: 10.2196/27397] [Medline: 33861202]
27. Blease CR, O'Neill SF, Torous J, DesRoches CM, Hägglund M. Patient access to mental health notes: motivating evidence-informed ethical guidelines. J Nerv Ment Dis 2021;209(4):265-269. [doi: 10.1097/NMD.0000000000001303] [Medline: 33764954]
28. Blease C, Salmi L, Rexhepi H, Hägglund M, DesRoches CM. Patients, clinicians and open notes: information blocking as a case of epistemic injustice. J Med Ethics 2021;48(10):785-793 [FREE Full text] [doi: 10.1136/medethics-2021-107275] [Medline: 33990427]
29. Hägglund M, Scandurra I. A socio-technical analysis of patient accessible electronic health records. Stud Health Technol Inform 2017;244:3-7. [Medline: 29039366]
30. Commission welcomes European Parliament's adoption of the European Health Data Space and regulation on substances of human origin. European Commission. URL: https://ec.europa.eu/commission/presscorner/detail/en/IP_24_2250
31. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010;19 Suppl 3(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
32. Bärkås A, Scandurra I, Rexhepi H, Blease C, Cajander Å, Hägglund M. Patients' access to their psychiatric notes: current policies and practices in Sweden. Int J Environ Res Public Health 2021;18(17):9140 [FREE Full text] [doi: 10.3390/ijerph18179140] [Medline: 34501730]
33. Bärkås A, Hägglund M, Moll J, Cajander Å, Rexhepi H, Hörhammer I, et al. Patients' access to their Psychiatric Records—a comparison of four countries. Stud Health Technol Inform 2022;294:510-514. [doi: 10.3233/SHTI220511] [Medline: 35612132]
34. Schwarz J, Bärkås A, Blease C, Collins L, Hägglund M, Markham S, et al. Sharing clinical notes and Electronic Health Records with people affected by mental health conditions: scoping review. JMIR Ment Health 2021;8(12):e34170 [FREE Full text] [doi: 10.2196/34170] [Medline: 34904956]
35. Hagström J, Blease C, Haage B, Scandurra I, Hansson S, Hägglund M. Views, use, and experiences of web-based access to pediatric electronic health records for children, adolescents, and parents: scoping review. J Med Internet Res 2022;24(11):e40328 [FREE Full text] [doi: 10.2196/40328] [Medline: 36413382]
36. Simola S, Hörhammer I, Xu Y, Bärkås A, Fagerlund AJ, Hagström J, et al. Patients' experiences of a National Patient Portal and its usability: cross-sectional survey study. J Med Internet Res 2023;25:e45974 [FREE Full text] [doi: 10.2196/45974] [Medline: 37389909]
37. Hägglund M, Kharko A, Hagström J, Bärkås A, Blease C, Cajander Å, et al. The NORDeHEALTH 2022 Patient Survey: cross-sectional study of National Patient Portal users in Norway, Sweden, Finland, and Estonia. J Med Internet Res 2023;25:e47573 [FREE Full text] [doi: 10.2196/47573] [Medline: 37955963]
38. HOD, The Norwegian Patient Right Act. 2001. URL: https://lovdata.no/dokument/NL/lov/1990-07-02- [accessed 2024-05-23]
39. Én inbygger - én journal? [in Norwegian] Meld. St. 9 (2012–2013). 2012. URL: https://www.regjeringen.no/no/dokumenter/meld-st-9-20122013/id708609/ [accessed 2023-05-17]
40. Perlich A, Meinel C. Automatic treatment session summaries in psychotherapy—a step towards therapist-patient cooperation. Procedia Comput Sci 2015;63:276-283. [doi: 10.1016/j.procs.2015.08.344]
41. Kaihlanen AM, Virtanen L, Buchert U, Safarov N, Valkonen P, Hietapakka L, et al. Towards digital health equity—a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022;22(1):188 [FREE Full text] [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
42. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 2012;380(9836):37-43 [FREE Full text] [doi: 10.1016/S0140-6736(12)60240-2] [Medline: 22579043]
43. Chou AF, Stewart SL, Wild RC, Bloom JR. Social support and survival in young women with breast carcinoma. Psychooncology 2012;21(2):125-133 [FREE Full text] [doi: 10.1002/pon.1863] [Medline: 20967848]
44. Goren A, Gilloteau I, Lees M, DaCosta Dibonaventura M. Quantifying the burden of informal caregiving for patients with cancer in Europe. Support Care Cancer 2014;22(6):1637-1646. [doi: 10.1007/s00520-014-2122-6] [Medline: 24496758]

XSL•FO
RenderX

45. Braun M, Mikulincer M, Rydall A, Walsh A, Rodin G. Hidden morbidity in cancer: spouse caregivers. J Clin Oncol 2007;25(30):4829-4834. [doi: 10.1200/JCO.2006.10.0909] [Medline: 17947732]
46. Adelman RD, Tmanova LL, Delgado D, Dion S, Lachs MS. Caregiver burden: a clinical review. J Am Med Assoc 2014;311(10):1052-1060. [doi: 10.1001/jama.2014.304] [Medline: 24618967]
47. Wolff JL, Darer JD, Berger A, Clarke D, Green JA, Stametz RA, et al. Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. J Am Med Inform Assoc 2016;24(e1):e166-e172. [doi: 10.1093/jamia/ocw108] [Medline: 27497795]
48. Latulipe C, Mazumder SF, Wilson RKW, Talton JW, Bertoni AG, Quandt SA, et al. Security and privacy risks associated with adult patient portal accounts in US hospitals. JAMA Intern Med 2020;180(6):845-849 [FREE Full text] [doi: 10.1001/jamainternmed.2020.0515] [Medline: 32364562]
49. Nurgalieva L, Cajander Å, Moll J, Åhlfeldt RM, Huvila I, Marchese M. 'I do not share it with others. No, it's for me, it's my care': on sharing of patient accessible electronic health records. Health Informatics J 2020;26(4):2554-2567 [FREE Full text] [doi: 10.1177/1460458220912559] [Medline: 32264734]
50. Lag (2018:218) med kompletterande bestämmelser till EU:s dataskyddsförordning. URL: https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-2018218-med-kompletterande-bestammelser_sfs-2018-218 [accessed 2023-05-06]
51. Bell SK, Folcarelli P, Fossa A, Gerard M, Harper M, Leveille S, et al. Tackling ambulatory safety risks through patient engagement: what 10,000 patients and families say about safety-related knowledge, behaviors, and attitudes after reading visit notes. J Patient Saf 2021;17(8):e791-e799. [doi: 10.1097/PTS.0000000000000494] [Medline: 29781979]
52. Newman-Toker DE, Pronovost PJ. Diagnostic errors—the next frontier for patient safety. J Am Med Assoc 2009;301(10):1060-1062. [doi: 10.1001/jama.2009.249] [Medline: 19278949]
53. Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 2014;23(9):727-731 [FREE Full text] [doi: 10.1136/bmjqs-2013-002627] [Medline: 24742777]
54. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013;173(6):418-425 [FREE Full text] [doi: 10.1001/jamainternmed.2013.2777] [Medline: 23440149]
55. Singh H, Hirani K, Kadiyala H, Rudomiotov O, Davis T, Khan MM, et al. Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record-based study. J Clin Oncol 2010;28(20):3307-3315 [FREE Full text] [doi: 10.1200/JCO.2009.25.6636] [Medline: 20530272]
56. Singh H, Daci K, Petersen LA, Collins C, Petersen NJ, Shethia A, et al. Missed opportunities to initiate endoscopic evaluation for colorectal cancer diagnosis. Am J Gastroenterol 2009;104(10):2543-2554 [FREE Full text] [doi: 10.1038/ajg.2009.324] [Medline: 19550418]
57. Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care. Washington, DC: The National Academies Press; 2015.
58. Gandhi TK, Kaplan GS, Leape L, Berwick DM, Edgman-Levitan S, Edmondson A, et al. Transforming concepts in patient safety: a progress report. BMJ Qual Saf 2018;27(12):1019-1026 [FREE Full text] [doi: 10.1136/bmjqs-2017-007756] [Medline: 30018115]
59. Blease CR, Bell SK. Patients as diagnostic collaborators: sharing visit notes to promote accuracy and safety. Diagnosis (Berl) 2019;6(3):213-221. [doi: 10.1515/dx-2018-0106] [Medline: 31039128]
60. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care 2021;46(2):192-204. [doi: 10.1080/17538157.2021.1879810] [Medline: 33840342]
61. Blease C, Cohen IG, Hoffman S. Sharing clinical notes: potential medical-legal benefits and risks. J Am Med Assoc 2022;327(8):717-718. [doi: 10.1001/jama.2021.23179] [Medline: 35119468]
62. Bärkås A, Kharko A, Blease C, Cajander Å, Johansen Fagerlund A, Huvila I, et al. Errors, omissions, and offenses in the health record of mental health care patients: results from a nationwide survey in Sweden. J Med Internet Res 2023;25:e47841 [FREE Full text] [doi: 10.2196/47841] [Medline: 37921861]
63. Hägglund M. Electronic health records in Sweden—how can we go from transparency to collaboration? BMJ Opinion. 2017. URL: https://blogs.bmj.com/bmj/2017/06/23/maria-hagglund-electronic-health-records-in-sweden-how-can-we-go-transparency-to-collaboration/ [accessed 2022-03-10]
64. Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. FinJeHeW 2018;10(4):381-395. [doi: 10.23996/fjhw.74511]
65. For Safety's Sake: The population's attitude to the benefits and risks of digital health data [in Swedish]. Swedish Agency for Health and Care Services Analysis. 2017. URL: https://www.vardanalys.se/rapporter/for-sakerhets-skull/#:~:text=Resultaten%20i%20korthet,s%C3%A4kert%20och%20skyddas%20fr%C3%A5n%20obeh%C3%B6riga [accessed 2023-12-01]
66. Bärkås A, Kharko A, Åhlfeldt RM, Hägglund M. Patients' experiences of unwanted access to their online health records. Stud Health Technol Inform 2023;302:356-357. [doi: 10.3233/shti230138] [Medline: 37203682]

XSL•FO
RenderX

Abbreviations

EHDS: European Health Data Space
EHR: electronic health record
NORDeHEALTH: Nordic eHealth for Patients: Benchmarking and Developing for the Future
ORA: online record access
PAEHR: patient accessible electronic health record

Edited by T de Azevedo Cardoso; submitted 17.05.23; peer-reviewed by E Meier-Diedrich, J Sippel, O Petrovskaya, Y Feng; comments to author 06.09.23; revised version received 31.10.23; accepted 25.04.24; published 27.06.24.

Please cite as:
Hägglund M, Kharko A, Bärkås A, Blease C, Cajander Å, DesRoches C, Fagerlund AJ, Hagström J, Huvila I, Hörhammer I, Kane B, Klein GO, Kristiansen E, Moll J, Muli I, Rexhepi H, Riggare S, Ross P, Scandurra I, Simola S, Soone H, Wang B, Ghorbanian Zolbin M, Åhlfeldt RM, Kujala S, Johansen MA
A Nordic Perspective on Patient Online Record Access and the European Health Data Space
J Med Internet Res 2024;26:e49084
URL: https://www.jmir.org/2024/1/e49084
doi:10.2196/49084
PMID:38935430

©Maria Hägglund, Anna Kharko, Annika Bärkås, Charlotte Blease, Åsa Cajander, Catherine DesRoches, Asbjørn Johansen Fagerlund, Josefin Hagström, Isto Huvila, Iiris Hörhammer, Bridget Kane, Gunnar O Klein, Eli Kristiansen, Jonas Moll, Irene Muli, Hanife Rexhepi, Sara Riggare, Peeter Ross, Isabella Scandurra, Saija Simola, Hedvig Soone, Bo Wang, Maedeh Ghorbanian Zolbin, Rose-Mharie Åhlfeldt, Sari Kujala, Monika Alise Johansen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Patient Portals Fail to Collect Structured Information About Who Else is Involved in a Person's Care

Liz Salmi[1], AS; Danielle Peereboom[2], MPH; David A Dorr[3], MS, MD; Leilani R Graham, BA; Jennifer L Wolff[2], PhD; Catherine M DesRoches[4], DrPH

[1]Department of General Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States
[2]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[3]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, United States
[4]Department of General Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Liz Salmi, AS
Department of General Medicine
Beth Israel Deaconess Medical Center
133 Brookline Avenue
HVMA Annex, Suite 2200
Boston, MA, 02215
United States
Phone: 1 617 754 9600
Email: lsalmi@bidmc.harvard.edu

Abstract

The US health care delivery system does not systematically engage or support family or friend care partners. Meanwhile, the uptake and familiarity of portals to personal health information are increasing among patients. Technology innovations, such as shared access to the portal, use separate identity credentials to differentiate between patients and care partners. Although not well-known, or commonly used, shared access allows patients to identify who they do and do not want to be involved in their care. However, the processes for patients to grant shared access to portals are often limited or so onerous that interested patients and care partners often circumvent the process entirely. As a result, the vast majority of care partners resort to accessing portals using a patient's identity credentials—a "do-it-yourself" solution in conflict with a health systems' legal responsibility to protect patient privacy and autonomy. The personal narratives in this viewpoint (shared by permission) elaborate on quantitative studies and provide first-person snapshots of challenges faced by patients and families as they attempt to gain or grant shared access during crucial moments in their lives. As digital modalities increase patient roles in health care interactions, so does the importance of making shared access work for all stakeholders involved—patients, clinicians, and care partners. Electronic health record vendors must recognize that both patients and care partners are important users of their products, and health care organizations must acknowledge and support the critical contributions of care partners as distinct from patients.

(J Med Internet Res 2024;26:e49394) doi:10.2196/49394

KEYWORDS

patient portals; shared access; proxy access; portal delegates; caregivers; care partners; health IT; electronic health records

Introduction

As digital tools continue to expand patient roles in health care interactions, technology integrations such as shared access, which uses separate identity credentials for patients and their portal delegates, allow patients to identify who they do and do not want involved in their care. However, clinical teams rarely communicate with patients about the benefits of shared access, and health systems often make registration for these features so onerous that interested patients and their desired delegates circumvent the steps entirely. Health care organizations must acknowledge and support the critical contributions of "portal delegates" as distinct from patients, and electronic health record (EHR) vendors need to make shared access features in portals easier to use for both.

The personal narratives in this viewpoint, shared by DAD, LS, and LRG with permission, elaborate on quantitative studies and provide first-person snapshots of challenges faced by patients and families as they attempt to gain or grant shared access during crucial moments in their lives.

XSL•FO
RenderX

Shared Access to the Patient Portal

When functioning well, patient portals can place patients in the driver's seat for their care and improve communication with the health care team. However, portals typically fail to collect structured information about who else is involved in the person's care, what these individuals do, and their capacity to handle care demands [1]. Caregivers (or "care partners") play an important role in navigating health system demands, and they commonly provide, oversee, and coordinate medical, nursing, and daily living tasks [1]. Care partners comanage care in myriad ways, such as (1) arranging and scheduling services, (2) joining patients in visits, (3) providing emotional support, and (4) participating in making important care decisions. Care partners are often family members but may also be neighbors, friends, or others having a close relationship with a patient and may or may not be involved in the provision of hands-on assistance with daily activities [2]. Unfortunately, health systems commonly under-engage and lack basic support services for care partners. If services for care partners are available, they are often not covered by insurance. As such, clinicians—facing burnout, staffing shortages, and a growing level of burden with in-basket management—rely on a "shadow workforce" of care partners who are often ill-prepared to assume these roles [3], and reimbursement mechanisms are often insufficient to encourage and support such engagement [4].

According to the Office of the National Coordinator for Health Information Technology (ONC), 86% of US acute care hospitals allow patients to register someone they trust for shared access to a patient portal account with identity credentials (ie, username and password) [5]. Shared access to the portal allows patients to formally identify and involve care partners in their care by designating them as "portal delegates" with their identity credentials. Shared portal access facilitates an information exchange among patients, clinicians, and care partners. It helps clinicians better understand who is involved in health system interactions when it is someone other than a patient, while simultaneously providing care partners with legitimacy and information transparency when engaging in health system activities. However, the uptake of shared access is low, leaving clinicians and other staff guessing about the support available at home or in the community (Textbox 1) [6].

It is important to note that portal delegates are different from health care or medical proxies. A medical proxy is a person who is legally authorized to make decisions on behalf of a person who is physically or mentally unable to make their own medical decisions. Emerging focus group research shows that "shared access" is the preferred term of use for explaining these concepts because it better conveys the intended purpose and functionality of separate login credentials for patients and their portal delegates [7,8].

More structured information about a patient's informal care team, when placed in the clinical record or integrated through clinical decision support tools [9], could help clinical teams understand who may help the patient at home or in the community (such as social organizations, religious groups, or neighbors) and provide context about who a patient trusts with knowledge of their health and information.

The following personal narratives, shared by permission, elaborate on quantitative studies and provide first-person snapshots of challenges faced by patients and portal delegates as they attempt to gain or grant shared access during crucial moments in their lives.

Textbox 1. Defining shared access.

- Shared access is sometimes referred to as proxy access.
- Shared access is different from a health care or medical proxy—which is a person legally authorized to make decisions on behalf of a person who is physically or mentally unable to make their own medical decisions.
- Shared access provides portal delegates with legitimacy and information transparency when engaging in health system activities about a patient's health and their treatments.
- Shared access can help clinicians better understand who is involved in health system interactions when it is someone other than a patient.
- When used as designed, shared access facilitates information exchange between patients, clinicians, and portal delegates.
- Getting shared access to work well for patients, portal delegates, and clinicians is critically important as digital modalities become the mainstream mode of interaction in health care.
- Shared access is especially important for patients who have lower digital health literacy and less experience with health tools, which are increasingly being moved online.

A Care Partner Perspective (DAD)

My mother and aunt both receive excellent care at major medical centers that use the same EHR vendor. The only difference is that my aunt's health system realized that not all portal delegates are patients in their system, and they made the experience of being a delegate easy. In contrast, my mother's health system made a deliberate choice to make shared access the responsibility of individual clinics.

Within a year, my mother and aunt were both diagnosed with complex medical conditions. As a physician and informaticist, I recommend patients and their care partners register for patient portals so they can easily communicate with their medical team, read visit notes, and view lab results. You would think with today's technology patient portals would be a breeze to navigate, but from a technical perspective, my mother's and aunt's portals could not have been more different.

I attempted to talk my mother through the shared access authorization process to register me as her delegate by having

XSL•FO
RenderX

her use a desktop computer, but she could not get it to work. I tried to help by accessing her portal through a mobile app. This time it was me who could not get it to work. We eventually called her doctor's office; they did not know what to do either. In the end, we learned my mother would need to visit a clinic in person to initiate a "nonpatient" request on my behalf. We tried this with the clinic staff but, despite sending reminder emails to me, it never allowed me to fully log in.

Compare this with my aunt who lives quite far away from me. As the doctor in our family, I help remotely whenever I can. A few clicks around her health system's website revealed my aunt could grant shared access to me as her delegate even though I was not a patient at her hospital. The solution was handled electronically. It was elegant and trouble free. I was surprised and impressed.

If medical professionals are truly committed to advancing shared access, we must be open to learning even more efficient ways of doing things and committed to innovation.

A Patient Perspective (LS)

Brett became my care partner at a relatively young age after I was diagnosed with a malignant brain tumor at the age of 29 years. He was witness to 2 brain surgeries before we married in the middle of a 2-year stint of chemotherapy.

Ten years later, my tumor recurred. I was scheduled for a third brain surgery, and we had 2 weeks to prepare. I knew the importance of having Brett officially registered as my portal delegate, but the hospital I trust with operating on my brain is 2 hours away and required a wet signature. On top of getting everything else in my life in order in the days leading to brain surgery, I spent hours calling my doctor's office, sending emails, and sitting on hold in an attempt to grant my husband shared access. I fruitlessly fiddled with my iPhone (Apple Inc) and attempted to send pictures of my signature authorizing my spouse as a portal delegate. According to the hospital, nothing short of a wet signature is acceptable.

Out of necessity and frustration, I circumvented formal channels by manually changing my portal settings to send immediate email and text alerts directly to my husband's email address. This made it possible for Brett to receive real-time notifications for every test result and progress note.

My over-preparedness and advanced digital planning turned out to be a blessing. The surgery left me with unexpected neurological deficits I am still learning to navigate today. It became hard for me to process numbers and understand dates and times. Making sense of medications and dosing, and tracking appointments became a challenge. Rather than turning off text and email alerts as originally intended, we remained subscribed, which permitted us to not worry about missed messages. I asked Brett if the alerts were information overload, but he believes it is worth it so he can make sure nothing falls through the cracks.

A Young Adult Perspective (LRG)

In the beginning, having access to my medical record at my fingertips was just annoying.

"I can't have an outpatient echo today; I'm literally in the ICU," I told my mother, as I silenced yet another appointment reminder notification.

My unscheduled heart transplant at the age of 24 years nullified the need for routine testing of my native, diseased heart. However, my transition from being cared for by the heart failure clinic to the exclusive property of the transplant team was not communicated between electronic medical record systems.

"I will ask someone to let them know," my mother soothed. At this stage in my recovery, I did not yet know how to read lab values or translate procedure jargon. I was not yet able to communicate to my usual degree, and telling people, "I no longer have that heart," was not as straightforward a declaration as I believed it to be.

However, the second I was discharged from the hospital, circumstances forced my hand (rather, all of me) to become a vessel of amateur medical knowledge. Staying alive was a full-time job, and the patient portal became an operating manual. Over time, this access has brought me a great sense of peace, inclusion, and partnership with my care team.

Barely a month after my stay, my dad was admitted to the hospital for a cardiac procedure. Although less severe, my father's heart developed from the same genetic mutation he passed on to me. We do not, however, share the same desire to be medically literate. After repeated unsuccessful attempts to glean a level of detail from his clinic visit recaps, I insisted on conferencing in, often dominating the conversation with my questions.

"Lei," he finally said. "Why don't I just sign something to give you full access to my medical record? You had the disease. You know what needs to be asked."

Using my patient portal made me a better patient and a stronger advocate. Being made a portal delegate to my dad's record has made me a better daughter.

Shared Access Honors Patient Health Information Preferences

As demonstrated in these narratives, adults of all ages commonly rely on and desire care partner involvement in their care, which is facilitated by access to timely and accurate health information [10-12]. Shared access to patient portals can facilitate and improve patient and care partner experiences of care; patient and care partner insight into patient health, activation, and continuity of care; and the ability to engage in health system interactions and the management of tasks [13-16]. However, the ease of shared access registration varies widely even though the functionality exists across health systems [17].

When used as designed, shared access facilitates information exchange between patients, clinicians, and their portal delegates. Shared access helps clinicians better understand with whom they are interacting when it is someone other than the patient. It ensures patient health information preferences are honored while being able to provide care partners with legitimacy for their digital interactions and easy access to their loved one's

XSL•FO
RenderX

health information. Shared access is especially important for patients with lower levels of digital health literacy and less experience with tools that are increasingly being moved online.

As illustrated by DAD and LS, the processes for patients to grant shared access to a portal delegate can be so limited or onerous that interested patients and delegates circumvent the process entirely. Instead, most potential delegates access the portal using the patient's identity credentials by sharing usernames and passwords—a "do-it-yourself" solution in conflict with a health system's legal responsibility to protect patient privacy and autonomy [13,18-21]. For example, a study by Pecina et al [20] looked at messages sent to a single health system through an online portal over a calendar year (n=752,551). Of these messages, the study found that 99.3% of messages were sent using the patient's identity credentials, and 0.7% were sent using a portal delegate's identity credentials [20]. The study team reviewed 3000 randomly selected portal messages and found that 83.8% (n=2512) were sent by the patient; 7.4% (n=221) were sent from someone other than the patient (indicated by the use of a third-person pronoun, eg, "John needs a refill on his medication"); and in 8.9% (n=266) of messages, the identity of the sender was unclear [20] (Table 1). A recent scoping review of the published literature found that in 7 articles, <3% of adult patient portal accounts were identified as having 1 or more formally registered portal delegates [16].

Table 1. Messages sent via the patient portal: from the patient, from someone other than the patient, and unclear of sender (adapted from Pecina et al [20]).

Messages sent via the patient portal	Message sent from a patient, n (%)	Message sent from someone other than the patient, n (%)	Unclear whether the message was sent by a patient or someone else, n (%)
Identity credentials used	747,581 (99.3%)	4970 (0.7%)	__[a]
Text analysis results	2512 (83.8%)	221 (7.4%)	266 (8.9%)

[a]Not available.

Sharing usernames and passwords between patients and care partners is an undesirable practice for several reasons. First, it compromises clinician-patient confidentiality, as clinicians can often identify when the message sender is not the patient, leading to confusion and mistrust. This concern is amplified for older adults who may lack digital proficiency. This practice undermines data integrity and accountability and potentially leads to misinterpretation or misuse of health information. In addition, sharing credentials conflicts with legal and ethical regulations, such as the Health Insurance Portability and Accountability Act (HIPAA), which could jeopardize patient privacy and health care provider obligations [22].

While shared access holds potential as a powerful tool, raising awareness and uptake, and getting shared access to work well for all parties involved, is critically important as digital modalities become the mainstream mode of interaction in health care [23]. In the patient narratives, LRG and LS highlighted the importance of the ability to update portal delegate credentials during crucial moments. Ensuring that all patients and care partners can access these options when they need it will require maintained effort by EHR vendors and health care organizations.

First, the process of requesting and granting shared access must be streamlined, and EHR vendors should work to make the electronic process more intuitive for patients and care partners by incorporating human-centered design principles into the portal interface [24]. As shown through DAD's story, health care organizations should remove barriers such as requiring that patients request shared access in person and recognize that portal delegates may not be patients in the same health care system or even live in the same geographic area as the patient.

Second, health care organizations need to make organizational commitments to strategies that enhance awareness and adoption of shared access, including education of patient communities and health care teams. Because shared access is so rarely used, awareness is quite low among patients, care partners, and clinical teams. The availability of shared access should be prominently described in print and web-based information about the patient portal. The shared access registration process should be clearly detailed with organizational contact information when people have questions (eg, phone number and email) and include working links to the documentation necessary for the patient authorization and delegate identity credentials. Accreditation organizations, such as the Joint Commission, could further spread these practices by recognizing organizations that prioritize shared access. Initiatives such as the Institute for Healthcare Improvement's Age-Friendly Health Systems and "dementia-capable" systems could recommend shared access as a standard to support care partners of all ages [25-27]. Furthermore, conversations around assigning portal delegates are best practices that apply to adults of any age.

Third, EHR vendors should offer granular privacy controls to meet the needs of various patient or care partner dyads. Health care organizations currently make "one size fits all" decisions about what portal delegates can access through the portal. Similar to the way most portals allow patients to set preferences about notifications, patients could decide what their portal delegate can see or do with their records. These granular privacy controls could respect patient autonomy and place decision-making power in their hands.

Finally, as illustrated by these stories, prioritizing care partner access to the patient's electronic health information while supporting patient autonomy will require new educational efforts and technology changes that require culture change at a fundamental level. EHR vendors will need to respect that both patients and portal delegates are important users of their products, and health care organizations will need to acknowledge and support the critical contributions of care partners as distinct from patients. There is a growing recognition that everyone is affected by health and functional setbacks that necessitate care

XSL•FO
RenderX

partner involvement. It is well past time to acknowledge this reality by making it easier for care partners to get the information they need and to communicate with the care team.

Ethical Considerations

According to the Beth Israel Deaconess Medical Center (BIDMC) Committee on Clinical Investigations, viewpoints coauthored with individuals representing patient or family perspectives are not human participants' research. In addition, none of the individuals sharing personal perspectives in this viewpoint are patients within the institutions named in the author list, and informed consent has been waived. No authors representing patient or care partner perspectives were offered compensation for their views.

Acknowledgments

This work was supported by The John A. Hartford Foundation research grant (LS, DP, JLW, and CMD).

Authors' Contributions

LS conceptualized the manuscript. LS, DP, DAD, LRG, JLW, and CMD contributed to writing and editing the manuscript. LS analyzed and interpreted the data. LS, DP, DAD, LRG, JLW, and CMD critically revised the text.

Conflicts of Interest

LS, DP, JLW, and CMD received support from The John A. Hartford Foundation to support the work described in this paper. The funder had no role in the design, conduct, or reporting of this work.

References

1. Committee on Family Caregiving for Older Adults, Board on Health Care Services, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine, Schulz R, Eden J. Families Caring for an Aging America. Washington, DC: National Academies Press; 2016.
2. Coalition for Care Partners. Boston, MA: OpenNotes & Johns Hopkins Bloomberg School of Public Health; 2023. URL: https://coalitionforcarepartners.org/ [accessed 2023-05-22]
3. Tai-Seale M, Dillon EC, Yang Y, Nordgren R, Steinberg RL, Nauenberg T, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019 Jul;38(7):1073-1078. [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
4. Riffin C, Wolff JL, Pillemer KA. Assessing and addressing family caregivers' needs and risks in primary care. J Am Geriatr Soc 2021;69(2):432-440 [FREE Full text] [doi: 10.1111/jgs.16945] [Medline: 33217776]
5. Henry J, Barker W, Kachay L. The Office of the National Coordinator for Health Information Technology. Electronic capabilities for patient engagement among US non-federal acute care hospitals: 2013-2017. 2019. URL: https://www.healthit.gov/sites/default/files/page/2019-04/AHApatientengagement.pdf [accessed 2024-05-31]
6. DesRoches C, Wolff J, Salmi L, Dorr D. Disconnections & dead ends: fixing proxy access for older adults.: OpenNotes; 2022. URL: https://youtu.be/u_kCvzvJqgg [accessed 2023-06-13]
7. Dukhanin V, Wolff JL, Salmi L, Harcourt K, Wachenheim D, Byock I, et al. Co-designing an initiative to increase shared access to older adults' patient portals: stakeholder engagement. J Med Internet Res 2023;25:e46146 [FREE Full text] [doi: 10.2196/46146] [Medline: 37991827]
8. Providence. Providence MyChart Shared Access Instructions. Providence MyChart Shared Access Instructions.: Providence; 2023 Nov 09. URL: https://www.youtube.com/watch?v=0llmL6zLpm4 [accessed 2023-11-22]
9. Patient-Centered Outcomes Research: clinical decision support.: Agency for Healthcare Research and Quality; 2023. URL: https://cdsic.ahrq.gov/cdsic/clinicians [accessed 2023-11-22]
10. Zulman DM, Nazi KM, Turvey CL, Wagner TH, Woods SS, An LC. Patient interest in sharing personal health record information: a web-based survey. Ann Intern Med 2011;155(12):805-810 [FREE Full text] [doi: 10.7326/0003-4819-155-12-201112200-00002] [Medline: 22184687]
11. Crotty BH, Walker J, Dierks M, Lipsitz L, O'Brien J, Fischer S, et al. Information sharing preferences of older patients and their families. JAMA Intern Med 2015;175(9):1492-1497. [doi: 10.1001/jamainternmed.2015.2903] [Medline: 26147401]
12. Latulipe C, Quandt SA, Melius KA, Bertoni A, Miller DP, Smith D, et al. Insights into older adult patient concerns around the caregiver proxy portal use: qualitative interview study. J Med Internet Res 2018;20(11):e10524 [FREE Full text] [doi: 10.2196/10524] [Medline: 30389654]
13. Wolff JL, Berger A, Clarke D, Green JA, Stametz R, Yule C, et al. Patients, care partners, and shared access to the patient portal: online practices at an integrated health system. J Am Med Inform Assoc 2016 Nov;23(6):1150-1158. [doi: 10.1093/jamia/ocw025] [Medline: 27026614]
14. Wolff JL, Aufill J, Echavarria D, Blackford AL, Connolly RM, Fetting JH, et al. A randomized intervention involving family to improve communication in breast cancer care. NPJ Breast Cancer 2021;7(1):14 [FREE Full text] [doi: 10.1038/s41523-021-00217-9] [Medline: 33579966]

XSL•FO
RenderX

15. Wolff JL, Darer JD, Berger A, Clarke D, Green JA, Stametz RA, et al. Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. J Am Med Inform Assoc 2017 Apr 01;24(e1):e166-e172 [FREE Full text] [doi: 10.1093/jamia/ocw108] [Medline: 27497795]
16. Gleason KT, Peereboom D, Wec A, Wolff JL. Patient portals to support care partner engagement in adolescent and adult populations: a scoping review. JAMA Netw Open 2022 Dec 01;5(12):e2248696 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.48696] [Medline: 36576738]
17. Wolff J, Kim V, Mintz S, Stametz R, Griffin J. An environmental scan of shared access to patient portals. J Am Med Inform Assoc 2018 Apr 01;25(4):408-412 [FREE Full text] [doi: 10.1093/jamia/ocx088] [Medline: 29025034]
18. Reed ME, Huang J, Brand R, Ballard D, Yamin C, Hsu J, et al. Communicating through a patient portal to engage family care partners. JAMA Intern Med 2018 Jan 01;178(1):142-144 [FREE Full text] [doi: 10.1001/jamainternmed.2017.6325] [Medline: 29159402]
19. Ramirez-Zohfeld V, Seltzer A, Xiong L, Morse L, Lindquist LA. Use of electronic health records by older adults, 85 years and older, and their caregivers. J Am Geriatr Soc 2020 May;68(5):1078-1082. [doi: 10.1111/jgs.16393] [Medline: 32159860]
20. Pecina J, Duvall MJ, North F. Frequency of and factors associated with care partner proxy interaction with health care teams using patient portal accounts. Telemed J E Health 2020;26(11):1368-1372. [doi: 10.1089/tmj.2019.0208] [Medline: 31971889]
21. Latulipe C, Mazumder SF, Wilson RKW, Talton JW, Bertoni AG, Quandt SA, et al. Security and privacy risks associated with adult patient portal accounts in US hospitals. JAMA Intern Med 2020;180(6):845-849 [FREE Full text] [doi: 10.1001/jamainternmed.2020.0515] [Medline: 32364562]
22. U.S. Department of Health and Human Services. Health information privacy. URL: https://www.hhs.gov/hipaa/index.html [accessed 2023-12-05]
23. Wolff JL, Dukhanin V, Burgdorf JG, DesRoches CM. Shared access to patient portals for older adults: implications for privacy and digital health equity. JMIR Aging 2022;5(2):e34628 [FREE Full text] [doi: 10.2196/34628] [Medline: 35507405]
24. Health+human-centered design.: U.S. Department of Health and Human Services; 2022. URL: https://www.hhs.gov/ash/osm/innovationx/human-centered-design/index.html [accessed 2023-05-07]
25. Adler-Milstein J, Raphael K, Bonner A, Pelton L, Fulmer T. Hospital adoption of electronic health record functions to support age-friendly care: results from a national survey. J Am Med Inform Assoc 2020;27(8):1206-1213 [FREE Full text] [doi: 10.1093/jamia/ocaa129] [Medline: 32772089]
26. Fulmer T, Berman A. Age-friendly health systems: how do we get there?. Washington, DC: Health Affairs; 2016.
27. Fulmer T, Mate KS, Berman A. The age-friendly health system imperative. J Am Geriatr Soc 2018;66(1):22-24. [doi: 10.1111/jgs.15076] [Medline: 28876455]

Abbreviations

BIDMC: Beth Israel Deaconess Medical Center
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
ONC: Office of the National Coordinator for Health Information Technology

Edited by T de Azevedo Cardoso; submitted 03.08.23; peer-reviewed by J Maxwell, J Ross, H Veldandi, A Hill; comments to author 06.10.23; revised version received 15.12.23; accepted 30.04.24; published 27.06.24.

Please cite as:
Salmi L, Peereboom D, Dorr DA, Graham LR, Wolff JL, DesRoches CM
Patient Portals Fail to Collect Structured Information About Who Else is Involved in a Person's Care
J Med Internet Res 2024;26:e49394
URL: https://www.jmir.org/2024/1/e49394
doi:10.2196/49394
PMID:

©Liz Salmi, Danielle Peereboom, David A Dorr, Leilani R Graham, Jennifer L Wolff, Catherine M DesRoches. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support

Carlos Sáez[1], PhD; Pablo Ferri[1], PhD; Juan M García-Gómez[1], PhD

Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Valencia, Spain

Corresponding Author:
Carlos Sáez, PhD
Biomedical Data Science Lab
Instituto Universitario de Tecnologías de la Información y Comunicaciones
Universitat Politècnica de València
Camino de Vera s/n
Valencia, 46022
Spain
Phone: 34 963877000 ext 85247
Email: carsaesi@upv.es

Abstract

Artificial intelligence (AI)–based clinical decision support systems are gaining momentum by relying on a greater volume and variety of secondary use data. However, the uncertainty, variability, and biases in real-world data environments still pose significant challenges to the development of health AI, its routine clinical use, and its regulatory frameworks. Health AI should be resilient against real-world environments throughout its lifecycle, including the training and prediction phases and maintenance during production, and health AI regulations should evolve accordingly. Data quality issues, variability over time or across sites, information uncertainty, human-computer interaction, and fundamental rights assurance are among the most relevant challenges. If health AI is not designed resiliently with regard to these real-world data effects, potentially biased data-driven medical decisions can risk the safety and fundamental rights of millions of people. In this viewpoint, we review the challenges, requirements, and methods for resilient AI in health and provide a research framework to improve the trustworthiness of next-generation AI-based clinical decision support.

(J Med Internet Res 2024;26:e50295) doi:10.2196/50295

KEYWORDS

artificial intelligence; clinical decision support; resilience; clinical medicine; machine learning; data quality; fairness; trustworthy AI; regulation; AI regulation; AI Act; EHDS; European Health Data Space; emergency medical dispatch; clinical decision support systems

Introduction

The advent of electronic health record (EHR) data sharing posed expectations for improved, trustworthy development of artificial intelligence (AI) in health through a larger volume and variety of data. However, the development process of health AI and the generalization and fairness of the resulting models face significant challenges due to the inherent biases, uncertainty, variability, and quality levels of real-world data (RWD). These challenges include variable information across different settings and over time, biases affecting underrepresented groups, uncertainty from lacking or overlapping information, or data quality (DQ) issues such as incomplete or implausible information. These issues can be present in training data feeding the AI learning or manifest de novo in the AI routine clinical use (Textbox 1). If health AI is not designed resiliently with regard to these RWD effects, potentially biased data-driven medical decisions can risk the safety and fundamental rights of millions of people.

XSL•FO
RenderX

Textbox 1. Clinical case with potentially suboptimal clinical decision support outcomes in a high-risk artificial intelligence (AI) system according to the European Union (EU) AI Act—Article 6(2)—Annex III.

A medical emergency dispatch center receives a call from a professor informing about an aged 20 years female student showing apparent respiratory distress. After data input, with no known chronic respiratory disease reported, an AI triage support system confirms that the case is not life-threatening. They send a basic life support ambulance. Eventually, the patient died during transport. The autopsy revealed a pulmonary embolism: she had recently started taking oral contraceptives. Clearly, this lack of information affected the AI outcome. Should AI have reacted by warning about potentially high uncertainty or asked for more information?

Alternatively, without previous embolism, a similar case might have occurred in March 2020 as a then-unknown effect of the SARS-CoV-2 infection. Instead of using a static AI decision support system trained on prepandemic data, should AI have automatically adapted to the very recent data patterns to provide better outcomes on new data?

Many recent papers in the highest-impact medical journals warned about the increasing obstacles imposed by RWD challenges for health AI. Cabitza et al [1] warned about the effects of intrinsic uncertainty in medical observations and interpretations on the reliability and accuracy of machine learning (ML). They claimed to develop and validate ML adaptable to the variable nature and actual DQ of medical information. Chen and Asch [2] argued that just relying on past data can have diminishing effects on AI's usefulness and future generalization, as well as the "butterfly effect" of tiny current variations into the future. Gianfrancesco et al [3] outlined the potential contribution to socioeconomic health disparities of ML based on biased data. Rajkomar et al [4] classified the availability of high-quality data and learning from undesirable past practices among the key challenges for medical ML, especially in RWD and nonimaging data from EHR. Google Health and DeepMind teams identified data variability as among the key challenges in delivering clinical impact with AI [5]. Besides, the COVID-19 pandemic highlighted potential limitations in medical AI from inadequate training and evaluation design and DQ and variability issues [6,7]. Not surprisingly, the European Commission has recognized DQ as a risk for the safety and fundamental rights assurance (FRA) in AI and included it along with other RWD challenges within the recently approved AI regulation (Article 10) [8,9] and among the significant issues for the quality and economy of the European Health Data Space (EHDS) [10-13].

We focus on ML as the methodological driver for current well-established health AI and clinical decision support systems (CDSSs) across diverse clinical fields. In health ML, new knowledge is learned computationally from data in a training phase, generally from dedicated data cohorts or the EHR. This knowledge is then used to assist decision-making for new cases in a prediction phase. At prediction, a CDSS can retrieve its inputs from the EHR or manual input. Potential data-related uncertainties, variability, and biases for health AI can arise both at the training and prediction phases (eg, see Table S1 in Multimedia Appendix 1 for some examples).

The definitions we provide in this work apply to multiple traditional and state-of-the-art ML approaches for CDSS, including predictive analytics and the potential upcoming use of conversational AI [14]. The related techniques include, but are not limited to, deep learning, ensemble models, and statistical methods, with knowledge acquisition schemes based on learning from scratch and using pretrained or foundation models [15,16].

Resilient AI: Definition and Requirements

Definition

We define resilient AI (RAI) as an AI that can automatically or semiautomatically adapt and react to unprecedented, uncertain, or undesired situations in real-world environments during model training and its use.

RAI emerges as a new paradigm over conventional AI procedures, generally aimed at model generalization and stability in controlled environments. The conventional AI approach expects learning on preprocessed, curated datasets and then inferring on equivalently consistent input data. The RAI approach, however, aims to learn and predict raw RWD by relying on resiliency-enabling methodologies and functions while improving its generalization and stability in variable environments. In RAI, adaptation would involve using new information to change or update AI knowledge, and reaction would involve providing appropriate information to support decision-making. Overall, RAI aims to enable trustworthy CDSSs for real-world environments, from adapting and explaining against potential biases and variability in the secondary use of EHR to handling uncertainty in decision-making, covering the whole AI lifecycle (Figure 1).

XSL•FO
RenderX

Figure 1. Comparison of the conventional AI process and the proposed resilient AI paradigm. Conventional AI is generally aimed at generalization and stability in controlled environments. That is, typically learning and validating a static model from a standardized, quality-controlled dataset, and applying the inference on equivalently consistent inputs. Resilient AI, however, aims to handle real-world variable environments during all its lifecycle, from adapting and explaining against potential biases and variability in the secondary use of electronic health records, both at model training and during use, to handling uncertainty in decision-making, enabling next-generation, trustworthy clinical decision support systems. AI: artificial intelligence, MD: medical device, I/O: input/output.

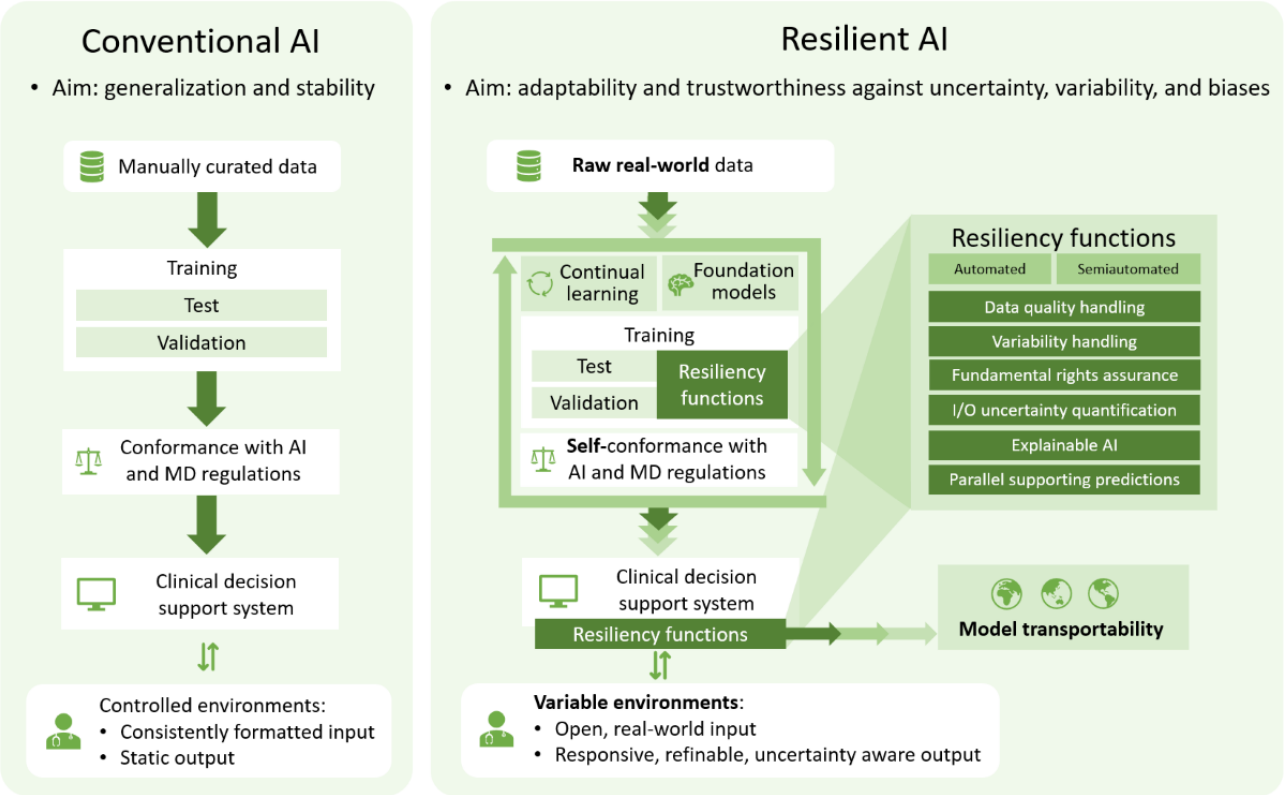

The challenges for RAI include DQ issues, data variability over time or across sources, information uncertainty, human-computer interaction, security and privacy, and FRA, among others. By wrapping the desired behavior of health AI against these challenges, we propose a set of requirements for the behavior and functionality of RAI in health (Figure 2). These requirements are organized in the 4 blocks described next.

XSL•FO
RenderX

Figure 2. A classification for the expected behavior and functionality of resilient AI in health, regarding the AI training phase, the prediction in new cases, the routine maintenance of AI during production, and related regulatory aspects. We focus on supervised ML as the methodological driver for current well-established medical AI and clinical decision support systems. AI: artificial intelligence; FRA: fundamental rights assurance; GDPR: General Data Protection Regulation; ML: machine learning; MNAR: missing not-at-random.

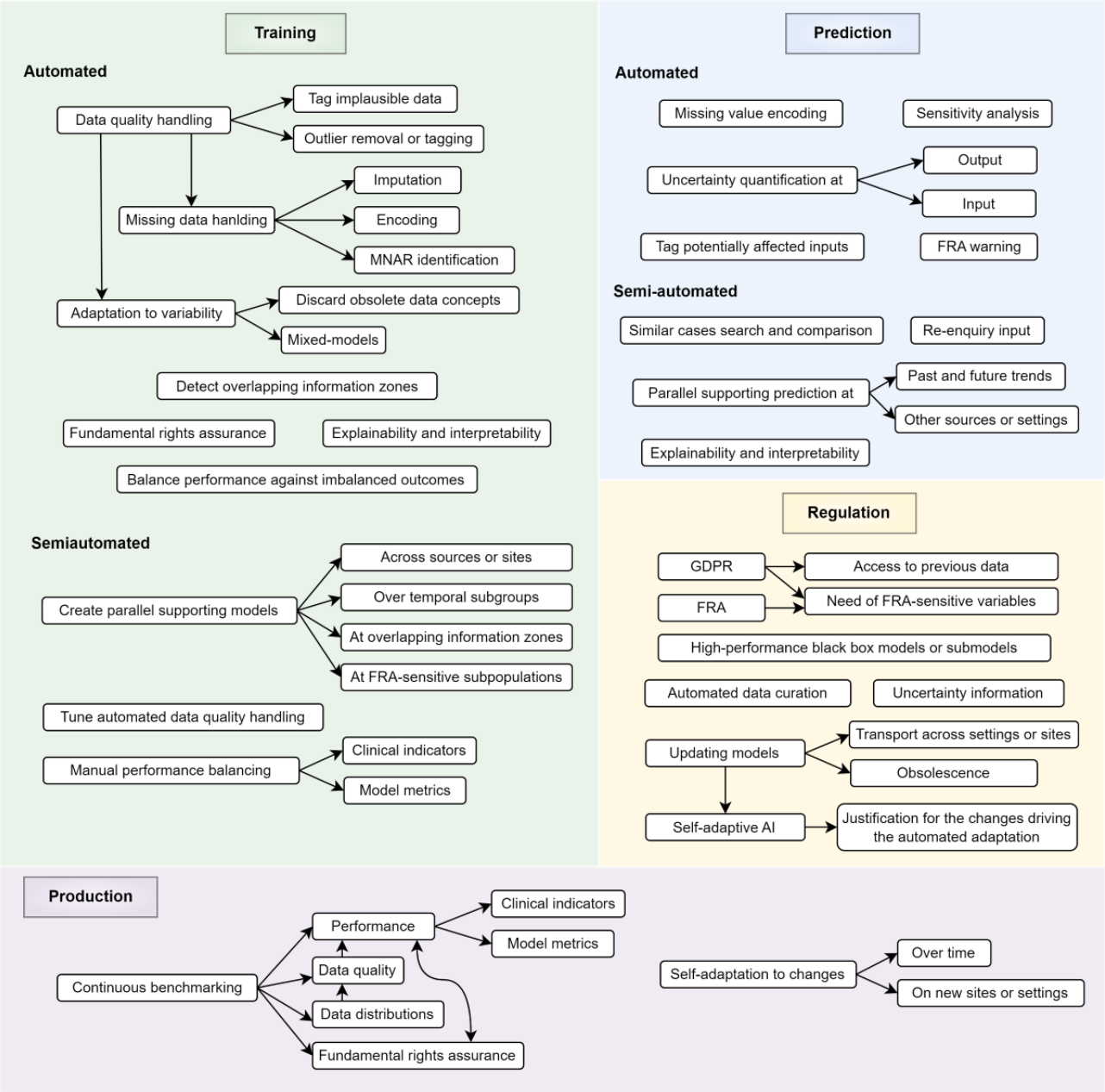

Training

RAI should ensure quality, generalizable knowledge through automated or semiautomated data preparation and lifelong learning processes cost-effectively. Manual data preparation generally consumes 80% of the costs of AI projects, and many data issues neither generally addressed nor detected in apparently error-free data can potentially lead to biased or inaccurate models.

Most data-related barriers in training are categorized as DQ issues [17-19], including missing, implausible, or outlier data. RAI should tag DQ issues and allow models to discard or fix them, infer any useful knowledge associated with faulty inputs, or suggest multiple modelling pipelines. For example, repeated missing laboratory tests at emergency admission in patients with COVID-19 due to rapid intensive care unit admissions might lead to a missing not-at-random situation associated with higher severity in the outcome.

RAI training should be resilient against variability over time and across settings or sites. Changing or unexpected information patterns can present as shifts in the covariates—$p(x)$—prior probability shifts in the outcomes—$p(y)$—or as concept shifts in one conditioned to the other—$p(y|x)$ or $p(x|y)$—leading to dataset shifts [20-22]. For example, would prediction models for acute respiratory distress syndrome trained on the first COVID-19 wave EHRs perform on new cases as initially evaluated? Would a model trained on a global European sample perform equally cross-border? These shifts can even relate to FRA-sensitive variables such as gender or race [10]. To delineate and characterize variability is critical, as it is to

XSL•FO
RenderX

optimize model performance according to the population where the model will be used. Therefore, variability-resilient AI training should apply both at the initial dataset learning and during its prospective production cycle, as described in this paper.

Similarly, ML responds suboptimally at overlapping information zones, that is, where repeated similar cases show distinct outcomes, bounding the Bayes error, the minimum and irreducible error rate achievable by a model. RAI should help delineate these zones during training, enabling reviewing, skipping, or fixing these cases or suggesting complementary models trying to optimize prediction at these zones.

RAI training should account for user requirements regarding the expected performance, including cost-benefits for each miss-correct classification, targets for sensitivity, specificity, or positive predictive value, and their trade-offs. These estimations are key in unbalanced outcomes and against FRA-unbalanced subpopulations. Lastly, training should enable prospective explainability and interpretability in the prediction phase.

Prediction

Directly related to clinical decision-making, resilience becomes critical in prediction. Balancing clinical usefulness with time-efficiency—with minimum intrusiveness—RAI should encompass fully automated support with smart human-AI interaction: a fully automated RAI will act without additional user feedback, while a semiautomated RAI will help improve or investigate the results through additional feedback.

Determinant for the user's adherence to the AI decision processes, RAI predictions should be explainable and interpretable [23,24]. Interpretability means understanding in human terms why a model provides an output and how it could vary from changes in the input—for example, through a sensitivity analysis. Explainability means understanding the internal mechanics of a model driving causality in a specific decision-making process.

Uncertainty is intrinsic to decision-making. Therefore, informing and quantifying uncertainty in health AI would significantly increase the system's trustworthiness and confidence in derived decisions. Particularly relevant are input cases showing unprecedented data patterns. RAI outcomes should count, where relevant, with CIs or levels. Likewise, RAI should handle potential input issues by identifying DQ-affected inputs, smartly encoding missing data, and identifying lacking optional information potentially relevant to the prediction.

Toward CDSS transportability, endurance, and fairness, RAI should enable the comparison of predictions obtained at multisite consensus or local level, identify potential changes in the outcome over time, or check for differences at distinct FRA-related subpopulations. Of note is that RAI should compare our case with similar past cases with validated outcomes, particularly beneficial in rare cases, and warn about potential FRA violations.

Production

Once an AI is in routine clinical use, changes can occur in the data acquisition contexts, the target populations of the model, or in clinical knowledge and user requirements. The dataset shifts described before in the training block can still occur in production, either over time or using the AI in different settings from those we initially trained it in. To avoid unexpected biases and obsolescence, we must continuously benchmark AI performance on target populations and enable automated or semiautomated self-adaptation mechanisms to these changes.

Besides benchmarking model performance in clinical indicators—for example, number of hospital readmissions in readmission prediction—or confusion matrix-derived metrics—for example, sensitivity and specificity—an RAI in production should monitor dataset shifts, user-defined DQ rules, user requirements—for example, deprecated or new International Classificafion of Diseases codes—and continuously benchmarking FRA.

On any significant change in AI performance, or even in advance of them, RAI should be able to self-adapt its knowledge over time or when transporting a model to other settings. Further, changes in requirements and clinical knowledge should be easily incorporated into previously built models without requiring a whole reengineering process.

Regulation

Production health AI and CDSS require regulations for medical device products, such as the European Union (EU) Medical Device Regulation [25], and specific regulations for AI, including the European Regulatory Framework for AI or AI Act [8,9,11]. The US Food and Drug Administration provides the AI and ML in Medical Devices white paper and its AI and ML Software as a Medical Device Action Plan [26,27].

The flexibility of current regulations on putting trustworthiness and patient safety as the top priority accommodates most of the described resilient AI needs. However, the wide RAI casuistry, involving more system self-decisions and higher levels of interaction and data access, might deserve more detailed definitions.

To promote AI while addressing potential risks, the EU AI Act includes a list of prohibited AI practices, rules for high-risk AI systems, transparency obligations, liability rules, and an innovation support legal framework, all consistent with the EU General Data Protection Regulation (GDPR). Specifically, the EU AI Act Article 15—Accuracy, Robustness, and Cybersecurity—Section 4 states that high-risk AI systems shall be as resilient as possible regarding errors, faults, or inconsistencies that may occur within the system or the environment in which the system operates, in particular, due to their interaction with natural persons or other systems. How to implement these resilience features is left to the system designers. At the simplest level, RAI could stop or warn users of unexpected situations, such as potential uncertainties or FRA risks at outputs or inputs. How this information is presented and interacted with, since it potentially affects decision-making, might be the subject of more specific regulations. Further,

XSL•FO
RenderX

Article 15, Section 5, also claims resilience against attempts to exploit system vulnerabilities.

However, in addition to during system operation, issues can occur as well at initial model training and in prospective retraining, where alternative resilience features could also take place. In this regard, Article 10—data and data governance—claims that data governance and management practices can be used to mitigate these data-related biases. However, fully resilient AI should, to some extent, make self-decisions for automated data curation or FAIRification procedures, which could be applied both during training and system operation. These may require stricter regulation and a justification supported by the added value of this process.

An innovative regulatory aspect for RAI focuses on self-adaptive AI, which automatically updates its knowledge to maintain performance and avoid obsolescence during production or, similarly, when transporting the AI to other settings. RAI self-updates should avoid, to some extent, new conformance evaluations. Self-adaptive RAI might also need data access regulations for benchmarking and retraining AI while respecting data security and privacy, such as in compliance with the EU GDPR [28]. Besides, accessing the EHRs for similar cases search and comparison, and the need for sensitive, FRA-related information potentially improving the system performance should also be regulated.

Lastly, for black-box models significantly outperforming other interpretable solutions, regulations should consider mitigation through resilient explainable and FRA features rather than limiting the AI use indications [29,30].

Available Methods and Solutions

Overview

Next, we describe a collection of currently available methods and solutions with a feasible while relevant use to address to some extent the RAI objectives and resiliency functions defined in the previous section.

Data Quality

The assurance of DQ can apply to the whole health data and AI lifecycles. Initially, AI is typically trained on secondary use data, of which quality and utility for AI should be labeled appropriately for trustworthy use. The current proposal for a regulation of the EHDS addresses the necessity of labeling the DQ and utility (Article 56) for data uses including research and personalized medicine, as is the case for AI.

DQ assurance can be specified based on DQ dimensions, which characterize data attributes facilitating or impeding its expected use. Dimension definitions vary according to the field of study [17,18,31,32]. Some dimensions are intrinsic to the data contents, including completeness, correctness, plausibility, or stability. Others relate to data access, including availability, accessibility, or security. Since potentially hindering or impeding AI development and use, we set intrinsic and access dimensions as the initial DQ targets in RAI.

Addressing DQ generally involves costly processes, from quality rules checks and exploratory analysis to data curation. Simple rules can find missing data footprints—blank spaces or, in the worst case, negative numbers such as “–9.” Logic rules defined by experts can help find implausible patterns in and between variables. Principal component analysis plots can help find apparently plausible but unlikely data, such as outliers. Further, information variability methods can help delineate temporal and source variability [22,33,34].

Addressing faulty data in training includes skipping faulty cases or recovering faulty values through data imputation methods. DQ flags can occasionally provide information relevant to AI performance, such as in the missing not-at-random case described in section 2. Further, as described in this paper, continual learning and model transportability methods can help address data variability in AI training.

Besides, DQ issues for AI can also occur when data is inputted into the AI systems, either being created de novo or passed through the EHR. At this stage, DQ handling methods can warn or automatically update missing, implausible, or outlying inputs with the most likely values given a context. However, any fully automated modification should be notified, and its effect on the output quantified through sensitivity analysis. This can be achieved through smart human-computer interaction and explainable AI methods.

Uncertainty Management

In health AI, data-related uncertainty can stem from epistemic or aleatoric factors, including lacking or faulty data or from pure statistical randomness in data or information overlap [35-37].

Quantification of uncertainty in AI focuses mainly on prediction accuracy, that is, informing about variability and confidence of the outputs. Available approaches include deterministic, Bayesian, ensemble, and augmentation methods [35,37]. Deterministic methods aim to predict jointly an output and its variability in single models, thus requiring this information at training. Bayesian methods model the statistical distribution of AI parameters—for example, the coefficients of logistic regression—rather than as single values, capturing the training data variability and enabling an output distribution from which to estimate a mean and CI. Bayesian methods include Monte Carlo dropout or Gaussian mixture models. Ensemble methods estimate multiple models—for example, through boosting or bagging approaches—in which combined outputs conform to a distribution. Lastly, augmentation methods apply artificial but realistic perturbations to inputs enabling multiple predictions which then also conform to a distribution.

Further, uncertainty from lacking information is particularly relevant in open inputs, including free text and dialogue systems such as in large language models (LLMs). It can be managed through deep learning network architectures, such as using recurrent neural networks that treat the input as a sequence with order information, recursive neural networks that treat the input as a hierarchical structure, transformers masked models such as BERT, where the different parts of an input sequence can influence each other through attention mechanisms [38], or using input embedding layers, where partial inputs can be expressed as indexes of dense vector representations [39].

XSL•FO
RenderX

Continual Learning and Model Transportability

Continual learning provides AI with mechanisms to self-adapt to information changes over time [40]. The model knowledge, generally represented as hyperparameters—for example, an artificial neural network layout—and parameters—the network weights—updates as new data batches arrive. Continual learning must balance forward and backward knowledge transfer, that is, rapidly adapting to dataset shifts while avoiding a catastrophic forgetting of important predictive past patterns. When using an AI model at settings or sites different from where it was trained, we face the problem of model transportability.

The continual learning and model transportability fields encompass similar methods from diverse ML paradigms, including online, active, transfer, or multitask learning [41]. Given their adaptative flexibility, these methods generally relate to artificial neural networks and deep learning. However, other simpler methods, including logistic regression or random forests, can also use batch or iterative learning. Of note is that these methods can update their knowledge without needing past data, conforming to data access regulations. Continual learning and model transportability can also benefit from infrastructure and organizational practices, including interoperability and domain knowledge management. Further, recent ML methodologies such as ML operations consider continual benchmarking and model updating [42]. Of note is that the use of AI foundation models, pretrained across various domains, can be of potential benefit to optimize resilience in continual learning and model transportability [43].

Fundamental Rights Assurance

Health AI involves potential risks for FRA regarding equality and nondiscrimination, economic and social rights, access to a fair trial and effective remedies, protection of personal data, or good administration [10]. RAI can address some of these rights by attaching to principles including the EU Charter of Fundamental Rights [44], the Ethics Guidelines for Trustworthy AI [45], or the GDPR [28]. In fact, FRA is at the core of medical AI regulations [46]. A successful implementation demands specific computational methods handling potential algorithmic biases.

We can classify FRA computational methods into pretraining, in-training, and posttraining methods [47,48]. Initially, we should identify available or derived variables potentially defining FRA-sensitive subpopulations, such as gender, race, or socioeconomic status. In pretraining methods, we compensate underrepresented subpopulations or imbalanced outcomes via resampling methods. Reweighting mechanisms can mitigate discrimination in outcomes. Further, we can remove or obfuscate FRA-sensitive variables if they do not affect model performance or usability. Besides, in-training methods focus mainly on defining objective functions or constraints in the AI learning process to compensate for unfair performance metrics in sensitive subpopulations—for example, optimizing the false-negative rate for a discriminated subpopulation. In screening and classification tasks, we can set subpopulation-specific cost benefits in confusion matrices or use specific deep learning loss functions—that is, the functions that measure the difference between the model prediction to the true answers to learn the network weights. Lastly, posttraining methods adapt a model's outcome to the sensitive situations of the tested individual—for example, setting custom receiver operating characteristic curves and decision thresholds or readjusting uncertain outcomes to favor sensitive subpopulations.

DQ and variability methods can also help detect unfair differences in data representations, DQ levels, or model performance across subpopulations. Another FRA aspect is decision-making transparency, where the models' interpretability is highly important.

Human-Computer Interaction

Resilient human-computer interaction is key to trustworthy AI and FRA. The methodological drivers for RAI by human-computer interaction include user experience and interface design methods [49], toward developing resilient user-centered CDSSs from the beginning [50], and AI and CDSS engineering aiming for dynamic and explainable AI [23,24].

Dynamic human-computer interaction in AI can use sequential prediction models that generate and update outcomes on partial inputs as new information is introduced, such as in recurrent or transformer neural networks [24]. Sequential models can establish a smart human-computer dialogue, particularly useful for sequential anamnesis inputs and natural language processing, such as using LLM.

Regarding explainable AI, the simpler the model, the more interpretable it is. For example, we can easily reason about a logistic regression behavior based on the feature's coefficients. Explainable AI post hoc methods provide explanations for less interpretable or "black box" models, which typically perform higher in complex tasks, including natural language processing or medical imaging. Post hoc local methods explain specific predictions, including the model-agnostic Shapley values or neural network gradient attribution techniques such as saliency maps. Post hoc global methods estimate each feature's average importance for any prediction. Surrogate interpretable models in points close to the test case can assist in explaining individual predictions from less-interpretable models [51]. Lastly, explainable AI local techniques can provide a sensitivity analysis for the outcomes.

Regulation

Current regulations supporting the development and maintenance of RAI solutions AI include those described in section 2, namely the EU AI Act, the EU Medical Device Regulation, the AI Action Plans by the Food and Drug Administration, and the proposal for the EHDS.

As a first step to facilitate the transportability of production AI systems, the EHDS provides a set of common specifications (Article 23) and a list of interoperability and security requirements (Annex II). These include structural and semantical requirements, and requirements related to DQ, patient safety, and data protection.

Some of the harmonized rules (HRs) that provide concrete details on how to meet the EU AI Act's goals align with our aim for RAI in health: high DQ and proper statistical properties

XSL•FO
RenderX

in training, testing, and validation are essential for the performance of AI and to avoid biases—HR-67; explainability and transparency to avoid opacity is key to use the system appropriately—HR-72; the system performance should be consistent throughout its lifecycle and monitored in real-life settings—HR-74; and the systems should be resilient against errors, faults, inconsistencies, or unexpected situations, as well as malicious actions, to avoid erroneous, potentially harmful decisions or biased outputs—HR-75.

Regarding self-adaptive AI, once a system is put into service, the EU regulatory framework for AI suggests providing rules establishing that changes in the algorithm and its performance, predetermined by the provider and assessed at conformity assessment, should not constitute a substantial modification—HR-128. Similarly, the American Medical Informatics Association suggests a set of recommendations for the safe, effective use of self-adaptive CDSS and their prospective regulation concerning: the design and development based on transparency metrics; implementation using standards for communication and planning retraining criteria; in-situ evaluations and testing; and on-going monitoring including system maintenance and user training [52].

Others

Further transversal approaches can benefit RAI. Federated learning provides algorithms and interoperability infrastructure for multisite AI learning, enabling secure model transportability and continual learning addressing potential data access regulations. In the prediction phase, fast nearest-neighbor search and phenotyping algorithms can proactively display relevant comparative past cases [53-55]. Lastly, we bring attention to the proposed commandments of ethical medical AI by Muller et al [30], which can help design RAI-enabled CDSS while encompassing human-AI decision roles.

Figure 3 summarizes the main solutions described in this section, linking to potential new research aims, as described in this paper in more detail.

Figure 3. Summary of available solutions and new proposed research for resilient AI in health and CDSS. AI: artificial intelligence; CDSS: clinical decision support systems; DQ: data quality; EHDS: European Health Data Space; EU: European Union; FDA: Food and Drug Administration; GDPR: General Data Protection Regulation; UX/UI: user experience and interface.

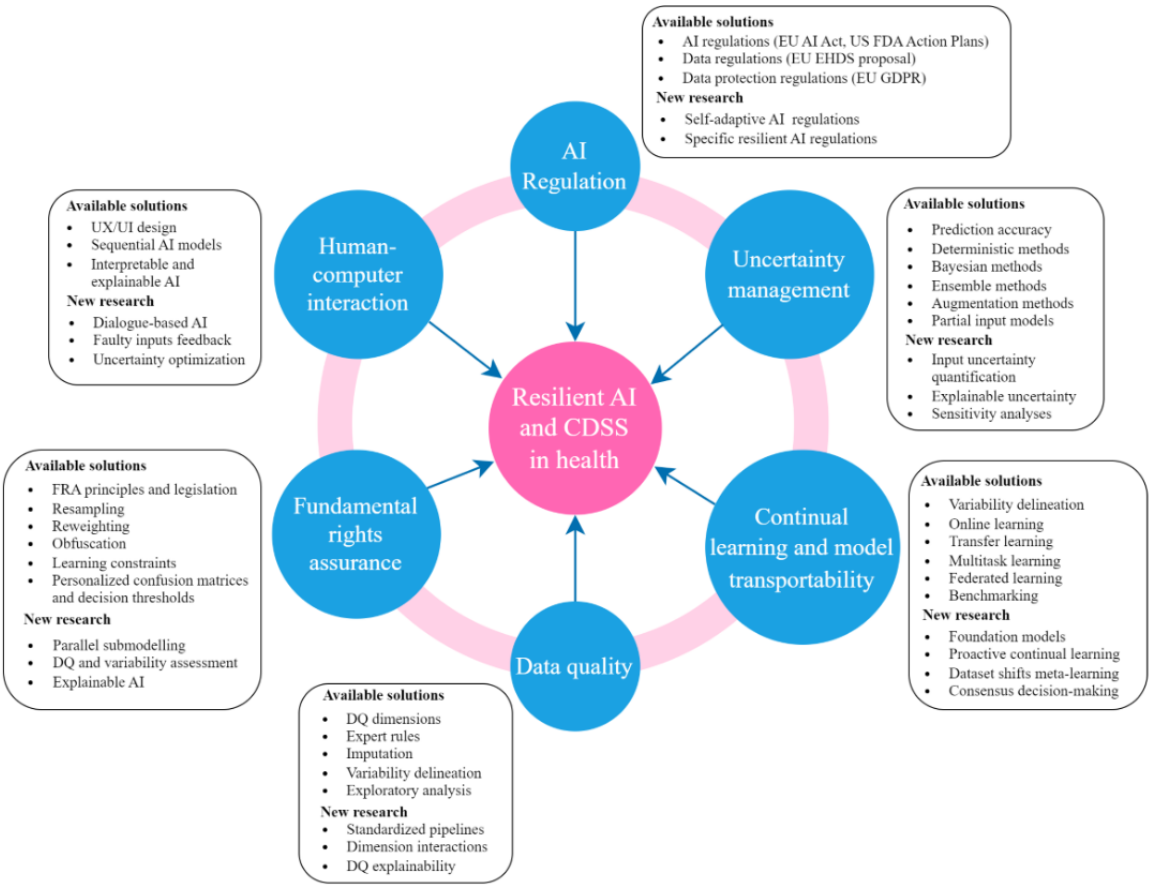

Research Agenda

Putting the methods and solutions described into practice would be already a huge step toward RAI and trustworthy CDSS (see Textbox 2 for an example of a desired use of an RAI-enabled CDSS). However, this is not free of challenges. The development complexity and costs can be considerable, regulations must be met or extended, and additional validation of the methods is needed. Fortunately, user acceptance and organizational culture regarding RAI may benefit from conventional AI, given its increased, user-centered flexibility. Next, we propose several research aims to address current and future challenges in RAI.

XSL•FO
RenderX

Textbox 2. Example desired use of a resilient artificial intelligence–enabled clinical decision support system following the example from emergency medical dispatch in Textbox 1.

Context

- Suppose a European emergency medical dispatch center develops a clinical decision support system based on resilient artificial intelligence (RAI) for triage support. The aim is to predict the life-threatening situation of incidents based on the live transcription of the during-call conversations to support resource allocation.

Training

- The initial RAI learning is fed from a raw database containing transcriptions, in situ maneuvers and diagnoses, transport, and hospitalization—as the final gold standard—data of the last 10 years. The system vendor and its triage tree changed during this period, leading to different conversation types. Further, the International Classification of Diseases codes at hospitalization were updated from version 9 to version 10. The RAI training is based on a continual, deep learning methodology; the neural network is learnt in temporal batches through a continual fine-tuning strategy. Therefore, the final model will retain past knowledge, but recent patterns will be more relevant. The training also relied on resiliency functions that permitted, for example, alerting about outlying operators with training data biased toward overclassification, prospective uncertainty quantification of inputs based on a Monte Carlo dropout function, and suggesting a parallel model for female cases with poor information in cardiac disease incidents with positive outcome to avoid false negatives.

Prediction

- The case from Textbox 1 is inputted into the system, which classifies it as a high-uncertainty case using the uncertainty quantification function. The system automatically asks for potentially relevant information to reduce uncertainty, such as the use of drugs or recent anxiety attack episodes. In light of neither changes in the input nor interoperability with the patient EHR to retrieve this data, the system uses a reject option-based classification function to classify the case as life-threatening to avoid potentially biased decisions.

Production

- A continuous artificial intelligence (AI) benchmarking system is used, counting with secure interoperability with gold-standard feedback data and automated inspection of causes of changes. The system eventually detects reduced performance in older adults and newborns, finding multidrug-resistant bacterial infections associated with these cases, and suggests retraining the model with these new data.

Regulation

- The system is a high-risk AI system in compliance with the European Union (EU) AI Act, ensuring its safe and trustworthy use. It conforms with the European Health Data Space proposal's common specifications, interoperability, and security requirements. An updated regulation framework considered the digital document with clinically explained motivation for the retraining described before as a valid source for self-conformance of the system to continue its use after its self-update.

We must standardize pipelines for fully automated and semiautomated DQ processing during training. The implications of DQ dimensions interactions, for example, across-site missingness patterns, should be studied. Excessive curation might lead to an unreal dataset; therefore, combining curation with using affected informative patterns—for example, informative missingness [56]—should be further studied. In prediction, we must investigate user-centered human-computer interaction approaches against faulty inputs, for example, using automated DQ corrections and DQ-related sensitivity analysis. Legal aspects for data curation, especially if automated, should also be realized technically [57].

Uncertainty quantification can benefit from addressing input uncertainty besides outcome uncertainty. Input uncertainty quantification could combine imputation and data augmentation methods. This approach can also help analyze the effect of DQ on output uncertainty and lead to input-outcome sensitivity analysis. Additionally, explainability might improve from uncertainty quantification at specific decision stages, for example, visualizing uncertainty at neural network branches.

Conversational, dialogue-based RAI interactions, smartly asking for new parameters anticipating the user needs, while minimizing input costs following an uncertainty-reduction targeted dialogue, are key research aims. This could be achieved by combining recurrent AI with dynamic human-computer interaction, for example, through voice or text-based LLMs [58,59], enabling AI to work with partial information, the natural case in medicine. Besides, user-centered design with a focus on explainability deserves specific research toward its acceptance [60].

Human-AI interaction is key to FRA since it influences users' decision-making. Previous methods for uncertainty quantification, DQ assessment, or RAI interactions shall be put in common with AI and health data regulations, for example, through specific HRs and methods enabling FRA as a priority layer in RAI-based CDSS interactions.

The use of foundation models in CDSS development should be further validated for safe and trustworthy clinical use. They can enable rapid development and model transportability, and show promising behavior to fill knowledge gaps in the training data, increasing generalization [15,16,43]. However, risks and uncertainty must still be thoroughly quantified. Besides, foundation models will still need to be updated as data changes through continual learning methods.

Continual learning can evolve from reactive to proactive. Current resilience against dataset shifts is constrained by the need for new labeled data to update the model's parameters, which is always one step behind the actual context. Further than adapting periodically or after significant shifts, proactive

XSL•FO
RenderX

continual learning can aim to anticipate changes. Domain adaptation methods—for example, transductive learning—can help update knowledge from cases before their gold-standard outcomes are available [61,62]. Besides, comprehensive variability characterizations in historical datasets can provide an extensive knowledge base of variability patterns, potentially enabling the forecasting of changes [22,63]. Overall, proper benchmarking for continual learning in health is still challenging, where we motivate the availability of publicly available patient-level datasets including the cases' date of acquisition.

Predictions based on the "consensus" of parallel models may support human decision-making resiliency—not generally based on a single mental process—through cooperative knowledge from different perspectives. Parallel models can apply to FRA-sensitive subpopulations, missingness patterns, multisite and temporal variability, and high uncertainty, overlapping-information zones. However, this can risk the CDSS user understanding; for example, what is the trustworthiness balance between local and multisite outcomes? Transparent, distributed, and continuous AI benchmarking is key. Complementarily, consensus decision-making can benefit from research in context and logic-based argumentation algorithms [64].

Specific resilience and self-adaptiveness features must be included in health AI regulations. The adaptability of AI to different settings and the temporal evolution of medical practice are recognized regulatory challenges [52,65]. Any change in an AI production system currently requires human intervention and conformance evaluation. Therefore, automatically providing a clinically explainable justification for why and how self-adaptive AI is updated without human intervention—besides planned retraining—is significant future work. Further, we must standardize AI documentation and record-keeping practices to define the rules firing a self-adaptation process and keep track of the latter—for example, based on performance changes or dataset shifts—and, similarly, justify automated DQ curation processes both at training and prediction.

Self-adaptiveness and federated learning infrastructures should be validated and regulated for CDSSs' retraining, production use, and benchmarking with data access limitations. In cases when protected information is required—for example, for training or similarity searches—synthetic data fabrication can be a solution.

Lastly, future work in RAI can address unsupervised and reinforcement learning. Unsupervised learning aims to discover natural subgroups in data. Generally applied at the population level—for example, to find potential immune response patterns in blood tests—it can also apply at the patient level—for example, for tissue segmentation in medical imaging. Besides, reinforcement learning aims to learn optimal decision-making choices in nonstationary environments—for example, to optimize intensive care unit procedures [66]. Improving the resilience of both will potentially improve their contribution to actual clinical practice.

Conclusions

Resilience is a significant factor in the success of health AI. Health data is imperfect, incomplete, and permeated with variable representations. However, this is not wrong. Instead of artificially modifying data for AI, AI should be resilient to the data nature. Methods enabling RAI involve disciplines including DQ, uncertainty management, continual learning, model transportability, foundation models, conversational AI, human-computer interaction, and regulatory aspects. Their implementation and specific research in RAI will define new-generation CDSSs and maximize trustworthiness in AI-enabled health care.

Acknowledgments

This work was cofunded (PID2022-138636OA-I00; KINEMAI) by Agencia Estatal de Investigación—Proyectos de Generación de Conocimiento 2022.

Authors' Contributions

CS conceived this work and wrote this paper. PF and JMG-G reviewed this paper and provided critical revision. All authors approved the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Potential data-related uncertainties, variability, and biases for health artificial intelligence at the training and prediction phases.
[PDF File (Adobe PDF File), 211 KB - jmir_v26i1e50295_app1.pdf]

References

1. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]

XSL•FO
RenderX

2. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507-2509 [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
3. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
5. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
6. Wynants L, van Calster B, Collins G, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ 2020;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
7. Sáez C, Romero N, Conejero J, García-Gómez JM. Potential limitations in COVID-19 machine learning due to data source variability: a case study in the nCov2019 dataset. J Am Med Inform Assoc 2021;28(2):360-364 [FREE Full text] [doi: 10.1093/jamia/ocaa258] [Medline: 33027509]
8. European Commission. Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 [accessed 2024-06-12]
9. European Commission. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Outcome of the European Parliament's first reading (Strasbourg, 11 to 14 March 2024). 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CONSIL:ST_7536_2024_INIT [accessed 2024-06-12]
10. European Union Agency for Fundamental Rights. Data quality and artificial intelligence: mitigating bias and error to protect fundamental rights. LU: Publications Office. 2019. URL: https://data.europa.eu/doi/10.2811/546219 [accessed 2022-06-01]
11. European Commission. White paper on artificial intelligence—a european approach to excellence and trust. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065 [accessed 2024-06-12]
12. European Commission. A european strategy for data. Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 [accessed 2024-06-12]
13. European Commission. Regulation of the European Parliament and of the Council on the European Health Data Space. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197 [accessed 2024-06-12]
14. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
15. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
16. Rosenbloom L, Hudson DA, Adeli E. On the opportunities and risks of foundation models. arXiv 2022. [doi: 10.48550/arXiv.2108.07258]
17. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
18. Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform 2012;180:721-725. [Medline: 22874286]
19. Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: organizing data for trustworthy artificial intelligence. Gov Inf Q 2020;37(3):101493. [doi: 10.1016/j.giq.2020.101493]
20. Quiñonero-Candela J. Dataset shift in machine learning. In: Mass: MIT Press. London, England: MIT Press; 2009.
21. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit 2012;45(1):521-530. [doi: 10.1016/j.patcog.2011.06.019]
22. Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. Gigascience 2020;9(8):giaa079 [FREE Full text] [doi: 10.1093/gigascience/giaa079] [Medline: 32729900]
23. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
24. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
25. Regulation (EU) 2017/745 of the European Parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives

XSL•FO
RenderX

90/385/EEC and 93/42/EEC. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745 [accessed 2024-06-12]
26. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food & Drug Administration. US: U.S. Food & Drug; 2021. URL: https://www.fda.gov/media/145022/download?attachment [accessed 2024-06-12]
27. Artificial Intelligence and Machine Learning in Software as a Medical Device. U.S. Food & Drug Administration. US: U.S. Food & Drug; 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [accessed 2024-06-12]
28. European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation). 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2024-06-12]
29. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA 2018;320(11):1107-1108. [doi: 10.1001/jama.2018.11029] [Medline: 30178025]
30. Muller H, Mayrhofer MT, van Veen E, Holzinger A. The ten commandments of ethical medical AI. Computer 2021;54(7):119-123. [doi: 10.1109/mc.2021.3074263]
31. Madnick SE, Wang RY, Lee YW, Zhu H. Overview and framework for data and information quality research. J Data Inf Qual 2009;1(1):1-22. [doi: 10.1145/1515693.1516680]
32. Aerts H, Kalra D, Sáez C, Ramírez-Anguita JM, Mayer M, Garcia-Gomez JM, et al. Quality of hospital electronic health record (EHR) data based on the international consortium for health outcomes measurement (ICHOM) in heart failure: pilot data quality assessment study. JMIR Med Inform 2021;9(8):e27842 [FREE Full text] [doi: 10.2196/27842] [Medline: 34346902]
33. Kahn MG, Raebel M, Glanz J, Riedlinger K, Steiner J. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012;50 Suppl:S21-S29 [FREE Full text] [doi: 10.1097/MLR.0b013e318257dd67] [Medline: 22692254]
34. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. J Am Med Inform Assoc 2016;23(6):1085-1095 [FREE Full text] [doi: 10.1093/jamia/ocw010] [Medline: 27107447]
35. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion 2021;76:243-297. [doi: 10.1016/j.inffus.2021.05.008]
36. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110(3):457-506. [doi: 10.1007/s10994-021-05946-3]
37. Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev 2023;56(S1):1513-1589. [doi: 10.1007/s10462-023-10562-9]
38. Goodfellow I, Bengio Y, Courville A. Deep learning. USA: Cambridge, MA. USA: MIT Press; 2016.
39. Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. Advances in Neural Information Processing Systems. 2000. URL: https://papers.nips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html [accessed 2024-04-16]
40. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. Neural Netw 2019;113:54-71 [FREE Full text] [doi: 10.1016/j.neunet.2019.01.012] [Medline: 30780045]
41. Mundt M, Lang S, Delfosse Q, Kersting K. CLEVA-compass: a continual learning evaluation assessment compass to promote research transparency and comparability. arXiv 2022. [doi: 10.48550/arXiv.2110.03331]
42. Dreyfus-Schmidt L. Introducing MLOps. URL: https://learning.oreilly.com/library/view/~/9781492083283/?ar?orpq&email=^u [accessed 2022-06-01]
43. Guo LL, Steinberg E, Fleming SL, Posada J, Lemmon J, Pfohl SR, et al. EHR foundation models improve robustness in the presence of temporal distribution shift. Sci Rep 2023;13(1):3767 [FREE Full text] [doi: 10.1038/s41598-023-30820-8] [Medline: 36882576]
44. European Commission. Charter of Fundamental Rights of the European Union 2012/C 326/02. 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:C2012/326/02 [accessed 2024-06-12]
45. European Commission. Ethics guidelines for trustworthy AI. Directorate-General for Communications Networks, Content and Technology. 2019. URL: https://data.europa.eu/doi/10.2759/346720 [accessed 2024-06-12]
46. Stöger K, Schneeberger D, Holzinger A. Medical artificial intelligence: the European legal perspective. Commun ACM 2021;64(11):34-36. [doi: 10.1145/3458652]
47. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv 2022;55(3):1-44. [doi: 10.1145/3494672]
48. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI Ethics 2023;3(1):223-240. [doi: 10.1007/s43681-022-00195-z]
49. Margetis G, Ntoa S, Antona M, Stephanidis C. Human-centered design of artificial inteligence. In: Handbook of human factors and ergonomics. USA: John Wiley & Sons; 2021:1085-1086.

XSL•FO
RenderX

50. Melnick ER, Hess EP, Guo G, Breslin M, Lopez K, Pavlo AJ, et al. Patient-centered decision support: formative usability evaluation of integrated clinical decision support with a patient decision aid for minor head injury in the emergency department. J Med Internet Res 2017;19(5):e174 [FREE Full text] [doi: 10.2196/jmir.7846] [Medline: 28526667]
51. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You": explaining the predictions of any classifier. arXiv 2016. [doi: 10.48550/arXiv.1602.04938]
52. Petersen C, Smith J, Freimuth R, Goodman KW, Jackson GP, Kannry J, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc 2021;28(4):677-684 [FREE Full text] [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
53. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
54. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol 2015;33(5):462-463. [doi: 10.1038/nbt.3223] [Medline: 25965751]
55. Li W, Zhang Y, Sun Y, Wang W, Li M, Zhang W, et al. Approximate nearest neighbor search on high dimensional data — experiments, analyses, and improvement. IEEE Trans Knowl Data Eng 2020;32(8):1475-1488. [doi: 10.1109/tkde.2019.2909204]
56. Ferri P, Romero-Garcia N, Badenes R, Lora-Pablos D, Morales TG, Gómez de la Cámara A, et al. Extremely missing numerical data in electronic health records for machine learning can be managed through simple imputation methods considering informative missingness: a comparative of solutions in a COVID-19 mortality case study. Comput Methods Programs Biomed 2023;242:107803 [FREE Full text] [doi: 10.1016/j.cmpb.2023.107803] [Medline: 37703700]
57. Stöger K, Schneeberger D, Kieseberg P, Holzinger A. Legal aspects of data cleansing in medical AI. Comput Law Secur Rev 2021;42:105587. [doi: 10.1016/j.clsr.2021.105587]
58. Moore MR, Arar R. Studies in Conversational UX Design. In: Moore RJ, Szymanski MH, Arar R, Ren GJ, editors. Conversational UX design: an introduction. US: Cham: Springer International Publishing; 2018:1-16.
59. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine 2023;90:104512 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
60. Cabitza F, Campagner A, Malgieri G, Natali C, Schneeberger D, Stoeger K, et al. Quod erat demonstrandum?—towards a typology of the concept of explanation for the design of explainable AI. Expert Syst Appl 2023;213:118888. [doi: 10.1016/j.eswa.2022.118888]
61. Vladimir V. Transductive inference semi-supervised learning. Semi-Supervised Learning. London, UK: The MIT Press; 2006. URL: https://axon.cs.byu.edu/~martinez/classes/778/Papers/transductive.pdf [accessed 2024-06-12]
62. Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B. Covariate shift by kernel mean matching. Dataset Shift in Machine Learning. UK: The MIT Press; 2008. URL: https://www.gatsby.ucl.ac.uk/~gretton/papers/covariateShiftChapter.pdf [accessed 2024-06-12]
63. Sáez C, García-Gómez JM. Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: functional data analysis of data temporal evolution over non-parametric statistical manifolds. Int J Med Inform 2018;119:109-124 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.09.015] [Medline: 30342679]
64. Vassiliades A, Bassiliades N, Patkos T. Argumentation and explainable artificial intelligence: a survey. Knowl Eng Rev 2021;36:e5. [doi: 10.1017/s0269888921000011]
65. Cohen IG, Evgeniou T, Gerke S, Minssen T. The European artificial intelligence strategy: implications and challenges for digital health. Lancet Digit Health 2020;2(7):e376-e379 [FREE Full text] [doi: 10.1016/S2589-7500(20)30112-6] [Medline: 33328096]
66. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med 2019;25(1):16-18 [FREE Full text] [doi: 10.1038/s41591-018-0310-5] [Medline: 30617332]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
DQ: data quality
EHDS: European Health Data Space
EHR: electronic health record
EU: European Union
FRA: fundamental rights assurance
GDPR: General Data Protection Regulation
HR: harmonized rule
LLM: large language model
ML: machine learning
RAI: resilient artificial intelligence

XSL•FO
RenderX

RWD: real-world data

Edited by G Tsafnat; submitted 26.06.23; peer-reviewed by Y Tolias, Z Su, C Manlhiot, A Holzinger; comments to author 27.01.24; revised version received 16.04.24; accepted 18.05.24; published 28.06.24.

Please cite as:
Sáez C, Ferri P, García-Gómez JM
Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support
J Med Internet Res 2024;26:e50295
URL: https://www.jmir.org/2024/1/e50295
doi:10.2196/50295
PMID:38941134

©Carlos Sáez, Pablo Ferri, Juan M García-Gómez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Enhancing Agency in Posttraumatic Stress Disorder Therapies Through Sensorimotor Technologies

Vladimir Adrien[1,2,3], MD, PhD; Nicolas Bosc[3], MD, PhD; Claire Peccia Galletto[4], BSc; Thomas Diot[5], MSc, MD; Damien Claverie[6], MD, PhD; Nicco Reggente[7], PhD; Marion Trousselard[6,8,9], MD, PhD; Eric Bui[10,11,12], MD, PhD; Thierry Baubet[3,13,14], MD, PhD; Félix Schoeller[7,15], PhD

[1]Department of Infectious and Tropical Diseases, Avicenne Hospital, AP-HP, Université Sorbonne Paris Nord, Bobigny, France
[2]Institute of Psychiatry and Neuroscience of Paris, Inserm UMR-S 1266, Université Paris Cité, Paris, France
[3]Department of Psychopathology, Avicenne Hospital, AP-HP, Université Sorbonne Paris Nord, Bobigny, France
[4]Unités René Diatkine, Association de Santé Mentale du 13è arrondissement, Paris, France
[5]Department of Adult Psychiatry, Impact, Mondor Hospital, AP-HP, Université Paris-Est Créteil, Créteil, France
[6]Institut de Recherche Biomédicale des Armées, Brétigny-sur-Orge, France
[7]Institute for Advanced Consciousness Studies, Santa Monica, CA, United States
[8]INSPIIRE, Inserm UMR 1319, Université de Lorraine, Nancy, France
[9]ADES, CNRS UMR 7268, Aix-Marseille Université, Marseille, France
[10]Department of Psychiatry, Caen Normandy University Hospital, Normandie Université, Caen, France
[11]Physiopathology and Imaging of Neurological Disorders, UNICAEN, Inserm UMR-S 1237, Normandie Université, Caen, France
[12]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States
[13]Unité Transversale de Psychogénèse et Psychopathologie, Université Sorbonne Paris Nord, Villetaneuse, France
[14]Centre National de Ressources et de Résilience, Lille, France
[15]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States

Corresponding Author:
Vladimir Adrien, MD, PhD
Department of Infectious and Tropical Diseases
Avicenne Hospital, AP-HP
Université Sorbonne Paris Nord
125 rue de Stalingrad
Bobigny, 93000
France
Phone: 33 148955421
Fax: 33 148955428
Email: vladimir.adrien@aphp.fr

Abstract

Posttraumatic stress disorder (PTSD) is a significant public health concern, with only a third of patients recovering within a year of treatment. While PTSD often disrupts the sense of body ownership and sense of agency (SA), attention to the SA in trauma has been lacking. This perspective paper explores the loss of the SA in PTSD and its relevance in the development of symptoms. Trauma is viewed as a breakdown of the SA, related to a freeze response, with peritraumatic dissociation increasing the risk of PTSD. Drawing from embodied cognition, we propose an enactive perspective of PTSD, suggesting therapies that restore the SA through direct engagement with the body and environment. We discuss the potential of agency-based therapies and innovative technologies such as gesture sonification, which translates body movements into sounds to enhance the SA. Gesture sonification offers a screen-free, noninvasive approach that could complement existing trauma-focused therapies. We emphasize the need for interdisciplinary collaboration and clinical research to further explore these approaches in preventing and treating PTSD.

(J Med Internet Res 2024;26:e58390) doi:10.2196/58390

KEYWORDS

posttraumatic stress disorder; PTSD; agency; proprioception; trauma; self-control; sensorimotor technology; enactive perspective; peritraumatic dissociation; proprioceptive reafferent fibers; gesture sonification devices

XSL•FO
RenderX

Introduction

Posttraumatic stress disorder (PTSD) stands among the 10 major public health issues [1]. Remission is obtained for a mere one-third of patients at 1-year follow-up, while another third continues to grapple with symptoms a decade later [2]. The gold standard treatments are trauma-focused cognitive behavioral therapies (CBTs), including cognitive processing therapy and prolonged exposure [3,4]. These approaches aim at improving self-regulation [5] but lead to remission in only approximately 40% of patients [6], although criteria for treatment nonresponse are not well defined [7].

Virtual reality exposure therapy (VRET) has been used with immersive simulations of trauma-relevant environments [8,9]. However, its superiority over standard prolonged exposure remains controversial [10-12], probably due to some inherent limitations, such as the sensory conflicts virtual reality (VR) induces as well as how its visual interface decouples patients from their body and environment [13]. In this context, other sensorimotor technologies letting patients connect to their environment may be of interest due to their potential to modulate the sense of body ownership (SO) and sense of agency (SA).

Emerging data have in fact reported how the SO and SA may be impacted in PTSD [14-19]. While the loss of the SO and control has been extensively studied in the context of PTSD and is historically among the main targets of trauma-focused therapies, much less attention has been paid to the loss and restoration of the SA, as opposed to emotion regulation and cognitive control. Despite promising theoretical perspectives on this topic under the umbrella of enactive theories of mind [15,20-28], there is still a dearth of agency-based treatment and recovery options available to patients with PTSD and clinicians. In this perspective paper, we explore the SA in the context of PTSD and its importance in understanding symptoms and improving treatments. Our contribution entails providing a road map for the development of agency-based therapies in the future, along with offering an agency-based perspective on psychological trauma. Finally, we explore the potential of using screen-free innovative technologies such as gesture sonification (GS) to complement existing trauma-focused therapies that target internal regulation by directly influencing the SA, thereby paving the way for the creation of new and effective intervention tools.

SO and SA: Definition and Brief Review

The SO refers to the perception of one's own body, feelings, thoughts, or movements, integrating somatosensory signals fundamental to distinguish between self and other, that is, to self-consciousness and control [29,30]. It depends on the interaction of afferent feedback and a top-down contribution of body representations [31,32]. The interplay between afferent internal (interoceptive and proprioceptive) and external multisensory (eg, visual and tactile) stimuli is sufficient for the SO [33], which is viewed as a psychophysiological baseline involving the brain's default mode network (DMN) [34]. This multisensory integration in the premotor cortex enables bodily self-attribution [35,36], with the prioritization of the most relevant sensory domains [37]. Computational (Bayesian) models have been suggested to account comprehensively for this multisensory integration [38]. Conversely, the SA refers to the immediate feeling of initiating and controlling an action [39], that is, to the subjective perception of being an agent effecting changes in the external world with a sensorial reflection (ie, sensorimotor contingencies). That is, the SA refers to the feeling of being "the one who is causing some event in the external world" [40]. It depends on prior intention and prereflective perceptual monitoring of the consequences of self-generated action, achieved through basic efferent motor-control processes and sensory feedback [31,41-43]. The SA involves premotor, motor, and temporoparietal areas, intention, and action monitoring [34,44] and integrates sensory and motor signals into a coherent representation of the self-world system (ie, the sensory expectations deriving from motor output). The SA also relies on proprioception, which is broadly defined as the sense of (self-generated) movements, integrating signals arising from cutaneous, muscular, and joint receptors [45], and on causal attribution [46]. Therefore, the SA is best described as a multisensory process that integrates motor and nonmotor cues [47,48].

All in all, the SA is the sense of originating and controlling our actions, whereas the SO is the sense of being the one to act [29,39,47]. For instance, involuntary movement gives rise to the SO but not to the SA [42]. This phenomenological distinction has neural correlates [44], but recent work seems to show a more interactive model where the SA and SO are strongly related to one another [49-51]. For instance, both share a network in the left middle insula [52], and intereffector regions in the motor cortex display strong functional connectivity, suggesting that body control and action are part of a common circuit [53].

Prereflective SA and SO can be distinguished from attribution judgments [54-56], that is, the ability to attribute an action to its proper agent (self or other), which also integrates sensory information (the visual sense being determinant [57,58]). This integration processes multiple sources indirectly due to action and its congruence to the prediction of one's self-narrative and understanding [41]. These second-order reflective (ie, cognitively processed) attributions of agency and ownership involve *judgments* of agency and ownership [42,47,59] and can be distinguished from first-order prereflective (ie, motor processed) *experience* (or *feeling*) of the SA and SO, although experience can influence judgment [60]. It is worth noting that this separation between the SO and attribution of ownership [61,62], as well as the separation between the SA and attribution of agency [63], has been challenged.

The SO can be measured by experiments, such as the rubber hand illusion (RHI) [33,35,64,65], in which a lifelike left rubber hand is viewed by participants whose real left hand is hidden, and the same afferent signals (eg, tactile stroke by a paintbrush) are being delivered to both hands, generating the transfer of the SO from the real to the rubber hand (ie, the illusion that the rubber hand is the real hand). Participants are then asked to point to their left hand with the right one and point toward a position between both hidden and rubber hands, elucidating the proprioceptive drift.

XSL•FO
RenderX

To measure the SA, both direct and indirect measures have been proposed [66]. Direct measures are rating scales or self-report questionnaires [66], whereas indirect measures focus on intentional binding [60], that is, the perceived time interval between the action and its outcome in comparison to involuntary action [67]. There are ongoing debates about the reliability and validity of these measures, as no correlations were identified between both types that may assess different aspects of the SA [68]. In this context, the Sense of Agency Scale, a new psychometric scale, has been developed and validated to directly assess global SA [69]. This self-report questionnaire includes 2 factors, the sense of positive agency and the sense of negative agency (respectively the control and lack of control over the environment), and is thus interesting for assessing altered SA in psychopathology. The Sense of Agency Scale has been validated in multiple languages [70]. The SA can also be measured in the laboratory with the RHI, which has been adapted recently [71] to a "dynamic" RHI where the index finger of the rubber hand moves when the participants move their own finger, both hands being mechanically connected, and the participants transfer their SA to the rubber hand (ie, feel being the ones tapping on the table with their finger). Interestingly, the illusion stays equally strong for various combinated sensory domains [72]. The SO and SA can be dissociated by varying the mode of movement (passive or active) and the position of the rubber hand (congruent or not). This strongly suggests that the SO and SA still represent distinct cognitive processes. The alteration of both SO and SA has been extensively studied across psychiatric conditions [42,57,58,73-75].

Psychological Trauma as a Major Loss of the SA

Human experience constantly presents us with challenges that either meet our physical and cognitive skills or require that we develop new ones. When none of these are an option, or these options are overwhelmed, because of an unexpected event that goes largely beyond one's representations and ability to adapt, threatens survival and physical or cognitive integrity of oneself or others [76], and induces a response of intense fear, helplessness, or horror, this event is called a *traumatic event* (TE; derived from τρα μα, the Greek word for wound, hurt, or injury). In the Bayesian approach of the mind as a hierarchical predictive model of its reality, the TE does not relate to any predictive models available to make sense of the external world, as no empirical priors exist to account for the incoming sensory signals or the most likely motor response. Significantly, in severe cases of psychological trauma, the lack of a predictive model results in the loss of both vision and hearing [77]. From a physiological perspective, both central and peripheral regulatory systems are ineffective in carrying out their retrocontrol functions. The stress response system, which involves the activation of the amygdala at the central level and the corticotrope and noradrenergic hormonal systems at the peripheral level, is hyperactivated in contrast to the underactivation of brain structures responsible for typical retrocontrol [78,79]. This includes the prefrontal cortex (PFC) and the hippocampus [80-83]. Regarding circuit regulation, this aligns with an overactive salience network (SN; eg, amygdala with insula and the anterior cingulate cortex [ACC]), that is, with increased threat detection and fear learning [84], in contrast to weakly connected DMN and central executive network [85,86]. Along with the disruptions of the hypothalamic-pituitary-adrenal gland axis, the entire system is unable to effectively regulate the adaptive stress response.

Understanding the TE as a potential sensorimotor failure leads to a novel perspective of psychological trauma as *a major breakdown of the SA*. A lot of attention has been directed toward the SO and self-control, which depend largely on interoceptive signals (eg, "my body is unable to mitigate its own state of stress when being mugged"), whereas the clinical focus on the SA, directed toward exteroceptive signals, is insufficient. In the context of psychological trauma, the felt sense of helplessness (eg, "my body is unable to defend against an aggressor"), peritraumatic distress [87]; or tonic immobility, that is, the freeze response [88], translate into the failure of the SA. Hence, the breakdown of the SA automatically triggers a hardwired sensorimotor response, sometimes referred to as the defense cascade [88]. The alteration of both perception and action during the TE leads to dissociation [89], that is, disconnecting from incoming sensory information, generating a psychological distance from the traumatic experience, and allowing the patient to "tolerate" the intolerable [90]. Peritraumatic dissociation—an array of reactions to the TE that includes depersonalization, derealization, and emotional numbness [91]—has been thoroughly investigated as being the result of the absence of adequate sensory and motor representations (ie, "I do not understand what is happening to me and there is nothing I can do"). In derealization, the sense of the world is lost, whereas in depersonalization, the sense of self is lost. Emotional numbness corresponds to the absence of defensive emotions such as fear or anxiety. In certain cases, individuals may enter a state of sideration or extreme surprise, unable to move or plan any action, essentially losing their SA. Peritraumatic dissociation can be seen as a protective response to the intense emotional distress during the TE when there is no possibility of escape or avoidance and no sensory or motor representations of the event [23,92-94]. It provides a sense of safety and physical and psychic analgesia, reducing engagement with the TE [89,95]. Nevertheless and importantly, experiencing severe peritraumatic dissociation raises the risk of developing PTSD [96-99], interfering with trauma memory processing and coherence [89], thus, in this case, being deemed maladaptive, impeding the processing of the TE.

There is indeed a significant memory-related aspect of PTSD [100]: the emotions or sensory memories attached to the TE are either not integrated as long-term declarative memory, resulting in denial or the inability of verbalization of the TE, or are integrated as semantic instead of episodic memory [101,102], that is, as a factual general knowledge that does not seem to belong to the individual [103,104]. In other terms, the traumatic memory will be related to noetic instead of autonoetic consciousness [105,106]. This correlates with reduced activity or dysfunction in the hippocampal structures [107,108]. When the traumatic memory remains unprocessed, it fails to integrate with the individual's conscious perception of reality. Over time, the failure to integrate the traumatic memory leads to a default

XSL•FO
RenderX

in contextual processing [84,85,109,110], resulting in trauma-related cognitions, such as guilt and shame, which in turn induce a wide array of comorbid complications, such as depression [111], anxiety [112], and obsessive-compulsive disorder [113].

Posttraumatic dissociative states can also emerge later on as a protective reaction against the abnormal emotional associations linked to the traumatic memory [114]. Dissociative PTSD (D-PTSD) has been recognized in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) [115], and is supported by a neurophysiological basis that distinguishes it from conventional PTSD. Emotional overmodulation (ie, hypoarousal) in D-PTSD is thought to be linked to an underactivation of the amygdala and an overactivation of the PFC [114], in contrast to the standard PTSD model. The Clinician-Administered PTSD Scale for *DSM-5* [116] allows for differentiation between these 2 types of PTSD, and approximately 30% of individuals with PTSD may have the dissociative subtype [15,114,117]. It has already been demonstrated that patients with D-PTSD exhibit a higher proprioceptive drift in the RHI [15], suggesting that individuals with D-PTSD integrate the illusion more significantly due to their diminished SO. Further research is needed to determine if the SA is altered differently between these clinical subtypes. The latest version of the *International Classification of Diseases* [118-120] has recently introduced a subtype known as complex PTSD (C-PTSD), which is characterized by the symptoms of PTSD and 4 additional groups of symptoms: dissociation; difficulties in regulating emotions; a negative self-concept involving feelings of worthlessness, defeat, shame, and guilt; and challenges in social cognition, such as maintaining relationships and feeling emotionally close to others. This distinction is strongly supported by research [119,121,122]. C-PTSD affects approximately 40% of individuals with PTSD [123]. As for both D-PTSD and C-PTSD subtypes, there are currently no established guidelines for best practices.

Trauma-focused CBT may restore the functionality of the brain structures (PFC, ACC, and hippocampus) that are involved in executive retrocontrol, which is initially ineffective during the TE [124]. Notably, trauma-focused CBT is safe [125] and has been found to be as effective in treating D-PTSD [126,127], despite its reversed neurophysiological model, and C-PTSD [123] than conventional PTSD. By concentrating on the traumatic memory, the goal of this therapy is to alleviate the emotional dysregulation experienced with PTSD symptoms. Alternatively, the enactive perspective asserts that the traumatized brain is engaged, interconnected with the body and environment, and dynamically integrated [128].

PTSD Symptoms Might Be Understood as an Attempt to Restore the SA

Overview

When considering psychological trauma from an enactive perspective and acknowledging that perception and action are intrinsically linked, the mechanisms described in the previous section and their accompanying symptoms can also be understood as an attempt to control oneself, the environment, and the causes of the TE that pervade both perception and action, whether physically or virtually. Thus, understanding PTSD symptoms as an adaptive response to psychological trauma can inform future treatment development, creating a more holistic approach to PTSD treatment that focuses on empowering patients to regain an SA and control over their environment. In this section, we briefly review PTSD symptoms and how they relate to the SA (Table 1).

XSL•FO
RenderX

Table 1. Posttraumatic stress disorder (PTSD) symptoms, their relation to the sense of agency (SA), and how intervening on the SA may target these symptoms.

	How does the symptom relate to the SA?	How agency-based therapies can help?
Intrusion syndrome	• Thoughts: attempt to get another shot at unfolding an optimal motor response to control the TE[a]. • Repetitive behaviors: attempt to attribute to our own action the exposition or occurrence of the TE or the destruction of oneself during the TE.	Exposure therapy aims at self-control during intrusion symptoms, rather than the SA. Agency-based therapy may augment gradual exposure by letting the patient control exposure level and would aim at increasing environmental control and replacing repetitive behaviors by enhancing the SA.
Avoidance behaviors	• Attempt to regain control over the environment and reminders of the TE or protect oneself from further harm.	Using gradual exposure therapy to triggers with body control of these triggers.
Hyperarousal and hypervigilance	• Minimizes the expected loss of the SA (as a failure to avoid or control the TE, the patient anticipates it anytime).	The patient controls sensory triggers themself with the help of the therapist to regain adapted and coherent salience.
Trauma cognitions	• Guilt: attempt to integrate the causes of the TE within a wider model of self, other, and reality in which oneself had some SA over what caused the TE to occur. • Anxiety: signal anxiety, anticipate the expected loss of the SA, and work as a protection. • Depression: loss of omnipotence or frustration of not having SA over the environment.	By controlling the environmental sensorial triggers, the patient can get pleasure in return, reducing anxiety and frustration and regaining some SA over the environment.
Dissociation	• Loss of the SA. • In addition, adaptive response against the passive loss of the SA felt during the TE or PTSD symptoms.	Increasing the sensorial signals produced by one's action in the external environment reduces the disconnection felt between oneself and reality.

[a]TE: traumatic event.

Intrusion Symptoms

Intrusive thoughts and memories as well as vivid re-experiencing, also called flashbacks, and traumatic nightmares are common symptoms in the aftermath of a TE [129,130]. These intrusive symptoms are distressing and make it difficult for patients to function in their daily life. Some patients with PTSD may have a tendency to replay the TE mentally in an obsessive-compulsive or addictive fashion, supposedly in an attempt to make sense of it and understand what had happened [131]. Intrusion symptoms can be regarded as a facet of dissociation [104]. Still, the recent findings on the activation of brain structures and neural circuits pointed out in the previous part, which would make PTSD and D-PTSD different entities, oppose this interpretation [114,132]. Intrusion symptoms have recently been analyzed from a Bayesian perspective, where the TE perceptual hypothesis gains a very high prior due to its life-threatening significance and is reselected independently of the actual sensory input [133]. From an enactive perspective, this can be understood as a way for the individual to try to regain some control over the TE to make it less distressing. Indeed, instructions to freely express intrusive thoughts led to a decrease in intrusive sexual assault thoughts over time, while suppressing them led to an automatic rebound in intrusive thoughts over time [134].

In addition to intrusive cognitions, repetitive behaviors, such as re-exposing oneself to potentially traumatic situations, are often seen in individuals with PTSD and can be a way for them to try to regain their SA. This can be observed in patients with C-PTSD and premorbid personality disorders, who may engage in risky or self-destructive behaviors [135]. Perhaps one of the most striking examples of repetition syndrome would be patterns of hypersexuality in children who have been victims of sexual abuse [136]. Patients are at risk of repeated harm, either self-inflicted or at the hands of others, adopting self-destructive behaviors [136] or repeated exposure to violence (suicide attempts, self-mutilation, military enlistment, development of substance use disorders, risky sexual behavior, and inability to give consent for sexual intercourse leading to exposure to rape situations), which can be seen as a tentative to attribute to our own action the destruction of oneself experienced during the original TE [121]. The climax of this viewpoint is found in the classical example of survivors becoming executioners (children who experienced abuse, child soldiers, etc) after a phase of repetitive behaviors such as engaging in plays or reenactments, often seen in childhood traumas [137]. Freud [138] famously discussed the idea of individuals using objects or activities to regain a sense of control over their environment. He gave the example of a baby playing with a toy that represents his mother to feel like he has control over the situation and soothe his distress. This idea was later developed by Winnicott [139] into the concept of transitional objects, used by children to cope with the loss of the subjective omnipotence over their environment, deriving from the caretaker's attention to the child's need, a frustration that can be related to the one experienced during a TE. Separation (ie, autonomy) from the parents accompanies the rise of the SA in childhood, as the sum of the attempts to compensate for the loss results in the gained ability to be alone. Intrusive cognitions and repetitive behaviors

XSL•FO
RenderX

are thus common symptoms observed in individuals with PTSD, which can be seen as attempts to regain an SA and cope with the psychological trauma. However, these symptoms can also lead to repeated harm or victimization, making it crucial for therapists to develop treatment strategies that focus on restoring the SA while addressing the underlying causes of these symptoms.

Avoidance Behaviors

The second category of PTSD symptoms is avoidance behaviors [140], that is, the active avoidance of thoughts, feelings, or external reminders of the TE. This can include avoiding certain places, people, or activities that may trigger memories of the event [141]. As patients have lost their SA over what triggers their flashbacks, the best way to control their occurrence is simply to avoid the stimulation. Here again, this might be understood as a displacement of the cause of the TE, where patients assume they can take responsibility for generating these sensory signals that largely surpassed their perception and action capabilities. One possible explanation for avoidance symptoms in PTSD is, therefore, that they may be an attempt by patients to regain their SA over their environment. For example, patients with PTSD due to a car accident may avoid driving, as driving may trigger memories of the TE [142].

An alternative explanation for avoidance symptoms in PTSD is that they serve as a means of self-protection, aiming to minimize the risk and vulnerability to risk. Social cognition impairment is commonly observed in PTSD [143]. By avoiding social situations, patients can shield themselves from re-experiencing the TE and encountering additional distress. Another instance is postdisaster PTSD, which is more likely to occur if the disaster originates from human actions as opposed to natural causes [144,145]. This underscores the greater psychological impact of interpersonal trauma, as it is perceived as potentially controllable, thus amplifying the distress. Patients may instinctively seek to regain the SA by avoiding situations associated with interpersonal trauma when addressing PTSD symptoms.

It should be noted that substance use is often seen as a form of avoidance in individuals with PTSD who may use drugs or alcohol as a way to numb or suppress their traumatic memories [146] or to escape from the difficult emotions that can accompany these memories. For example, self-reported PTSD is associated with increased use of 3,4-Methylenedioxymethamphetamine (MDMA) in adolescents with substance use disorders [147], and substance abuse is widespread in victims of childhood sexual abuse [148]. In addition to numbing their emotions, substance use can also be a way for individuals with PTSD to avoid trauma reminders (people, places, or situations).

Negative Alterations in Cognition and Mood and Maladaptive Schemas

Negative cognitions are thoughts and beliefs about oneself, others, and the world that are negatively distorted and not based on reality. They include beliefs that one is to blame for the TE, that the world is dangerous and unpredictable, and that one is not safe. They also include decreased interest in activities, negative emotions, and the feeling of being isolated. They may be a way of coping with the loss of the SA. Patients with PTSD often believe they are to blame (ie, the feeling of guilt) for the TE, perhaps as this belief provides them with a (virtual) sense of control over the situation. By experiencing guilt and self-blaming, they make sense of the event and feel like they have some SA over it—“I am responsible for exposing myself to the environment that generated the TE” [149].

In addition, negative cognitions in PTSD relate to the prediction of an *expected* loss of the SA. A TE can cause a person to feel vulnerable and at risk: negative cognitions may be a way of anticipating and preparing for future loss of control. For example, patients with PTSD may believe that the world is dangerous and unpredictable, as this belief helps them to be prepared for potential threats. This culminates with the concept of signal anxiety developed by Freud [150], one of the main defense mechanisms. By expecting the worst, patients are able to feel like they have some control over their environment and can protect themselves from further harm. Klein [151] related trauma negative cognitions to the “depressive position” in the early phase of existence. This preverbal infantile attitude comes with a loss of subjective omnipotence (ie, control) over external objects that coincides with the emergence of emotional valence in infancy: the child becomes aware that objects not only gratify (“good object”) but can also frustrate (“bad object”), generating feelings of guilt and grief and a desire for reparation. This example of “proto-traumas” or “micro-traumas” experienced repeatedly during early childhood lightens what happens when a real TE occurs later in life, altering the sense of control and leading to greater PTSD susceptibility [99,152,153]. These negative thoughts and beliefs may be a way of coping with the distress and vulnerability caused by the TE, but in a vicious cycle, they can also have negative effects on daily life and overall well-being related to an expected loss of the SA, if nothing is done to remedy it.

Alterations in Arousal and Reactivity: Disturbed Attentional Patterns

Patients with PTSD develop a heightened state of awareness, alertness, and physiological arousal to regain a sense of control and SA [154]. This is referred to as hypervigilance [114], characterized by emotional undermodulation, resulting in an increased threat detection and fear learning, and a state of alertness and readiness to respond to these threats [155]. This also includes irritability, difficulty concentrating, sleeping disorders, and an overall sense of feeling on edge. On a sensorimotor level, hypervigilance may manifest as an increased sensitivity to sensory stimuli, such as changes in light or sound, as well as increased physiological arousal, such as increased muscle tension or an elevated heart rate, which is a risk factor for developing PTSD when it immediately follows the TE [156,157]. PTSD is in return a risk factor and shares genetic risk for cardiovascular diseases [158]. By being constantly on guard and ready to respond to potential threats, patients may feel they are able to anticipate and prevent dangerous situations from occurring. This restores the SA that was lost as a result of the TE. Hypervigilance, as an attentional bias toward potential threats, can lead to anxiety and further hypervigilance [155,159]. This can cause the individual to misinterpret ambiguous cues

XSL•FO
RenderX

as threats and exaggerate minor threats, as shown by the overactive SN, whose function is to recognize and prioritize stimuli while regulating emotional reactivity [160]. These changes in attention and threat perception can also manifest in the individual's eye movements [161], leading to an inability to disengage from potential threats. This is referred to in the literature as "oculomotor reflexes" [162]. From a Bayesian inference perspective, dysfunctional SN amounts to aberrant precision control, where precision denotes the confidence placed in prediction errors (mirroring the reliability of the stimulation that causes them) within the hierarchy of information processing, that is, a high precision will favor bottom-up ascending prediction errors, while a low precision will bias perception toward top-down prior beliefs. Interestingly, the SN implies the ACC and insula, which are strongly related to interoception [163]. The dysregulation of bodily signals in psychiatric illnesses may, therefore, offer an important way forward in terms of phenotyping [164,165].

Dissociative Symptoms

After Winnicott [139], Anzieu [166] developed the concept of "skin-ego," that is, the presence of a symbolic skin allowing for the creation of psychic and bodily boundaries. In PTSD, this symbolic frontier is blurred or constricted and no longer filters or organizes sensory and perceptual signals as easily. It becomes porous and loses its function as a protective structure between interoceptive cues and cues originating from the external world. This sudden immense vulnerability exacerbates the loss of self-integrity and SO and forces the maintenance of a state of psychic survival with various defensive symptoms involving both temporality and space: agitation, withdrawal, and dissociative symptoms.

The view of peritraumatic dissociation as a protective response can be related to the anthropological interpretation of dissociative experiences with their positive affective valence, with the DMN mainly activated in situations such as daydreaming, hypnotic responses, fatigue, anxiety, drug intoxication, and boredom [167]. In this case, dissociation has a social and discursive meaning rather than being seen as a mechanism. It can thus function adaptively, which depends fundamentally on context [94,168].

By contrast, dissociation interferes with the integration of sensory information and the creation of a coherent sense of self. If dissociative symptoms of PTSD are the reminiscences of peritraumatic dissociation and protect patients against the distress provoked by intrusion symptoms or hyperarousal, they are nevertheless severely debilitating. Patients often complain of the distress related to their dissociative states, which can be transitory or permanent and are associated with high disability. Dissociation can indeed be interpreted literally as the loss of the SA: "if I see myself from an external point of view, thus my body's actions do not belong to myself" and "if the external world is not real anymore, then I can no longer be an agent in the real world."

Dissociation can thus at the same time be a loss of the SA and a protective phenomenon against it. There is an ongoing debate on this matter, with the concept of protective dissociation culminating in at-risk professions such as the military or fire service: workers experience deliberate dissociative states [169,170] to gain automatic action and execute tasks more efficiently without being emotionally overwhelmed. Further research is required to thoroughly investigate and clarify this phenomenon. One potential avenue is to distinguish between the *passive* loss of the SA resulting from external TE and the symptoms actively produced (deliberately or not) by patients as a means to combat passivity ("if it is due to my action or my mind that I lose SA, thus the weight of external events in the loss of SA is reduced"). This includes dissociation or dissociative behaviors (eg, drug use).

Sensorimotor Technologies for Agency-Based Therapies

The enactive, agency-based perspective on PTSD suggests that the disruption of the SA results from the TE, overwhelming the patient's ability to adapt and make sense of sensorimotor signals, leading to a breakdown in the hierarchical predictive model of reality. This disrupted SA may contribute to or be further exacerbated by peritraumatic dissociation, explaining why its intensity is a predictive factor of PTSD onset [96-99].

Currently, there is a lack of agency-based treatment options for PTSD. New treatment options to manage nonresponse to PTSD conventional treatments [7] include the promising use of ketamine [171] or psychedelic-assisted psychotherapy [172], whose dissociative effect is now reconsidered as a phenomenological therapeutic tool, as these changes of the self-experience have a subjective meaning and are transitory and necessary for coping. These drugs could help improve prefrontal function and contextual processing, modifying beliefs, refining predictions and thus the SA [173-178]. Ketamine could also induce brain-derived neurotrophic factor increase in the hippocampus [179]. Conversely, endocannabinoids modulators or drugs that reduce the glutamate response (such as D-cycloserine, a partial N-methyl-D-aspartate [NMDA] receptor agonist) could not only treat dissociative PTSD symptoms [180,181] but also enhance extinction training by preventing the hyperglutamatergic state of the stress response responsible for abnormal fear conditioning [182,183] as well as the defaults in memory processing [181]. Indeed, NMDA receptors are highly concentrated in the hippocampus and implied in long-term potentiation, a mechanism for encoding long-term (eg, episodic) memory. Neuromodulatory treatments [7] such as transcranial magnetic stimulation [184-186], transcranial direct current stimulation [187,188], deep brain stimulation [189], and vagus nerve stimulation [190,191] could increase body awareness, SO, and SA. Nevertheless, less invasive or expensive treatments, such as sensorimotor therapies [192-194], may have the potential to enhance and restore the SO and, more importantly, SA during PTSD recovery. A mounting body of evidence [165] suggests that bodily signals play an essential role in driving precision control, hinting toward the relevance of reliable body-based interventions for mental health disorders depending on the patient's life history, conditions, and symptoms. The controlled generation of artificial sensations could, therefore, lead to novel options for the diagnosis, monitoring, intervention, and treatment of disorders of emotional

XSL•FO
RenderX

and interoceptive inference. One such potential solution is "*affordance training*" with a narrative therapy map [195], which involves helping the individual to focus on action possibilities rather than action impossibilities. This can involve training patients to engage in goal-directed behavior and to seek out opportunities for action, rather than focusing solely on potential threats.

Sensorimotor technologies, such as VR, have been used for exposure therapy in immersive simulations of trauma-relevant environments [8,9]. The aim is to allow a precise control of stimulus [196]. VRET can discriminate between patients with PTSD and patients without PTSD on a measure of psychophysiological arousal such as skin conductance reactivity [197,198] as well as between patients with low symptoms of subthreshold PTSD and patients with high symptoms of subthreshold PTSD through heart rate [199].

However, various meta-analyses did not find any difference in clinical efficacy between conventional VRET and other psychotherapies [10-12], although some found moderate effects [200]. However, specific VR-graded exposure therapy seems to be beneficial [11]. All in all, VRET may be beneficial only for specific patient profiles (eg, those who are younger, with greater hyperarousal symptoms, with comorbid depression or suicidal risk, with no antidepressant medication, or who cannot engage in imaginal imagery) compared to prolonged exposure therapy [201-203]. One explanation of the limitation of VRET is that patients may become diverted by the technology, expressing doubts about its authenticity ("this isn't real") and using this divergence or the lack of personalization of virtual environments as a means to evade forming an emotional connection with their distressing memories [196], especially because sensorial integration prioritizes sight, as shown by the RHI [65]. Indeed, VR deprives patients of their gaze and hands due to the screen and interface, creating a disconnection between the patient's body and its environment, a restriction of virtual objects as actionable targets, contrary to tangible real-world objects [204], compromising the transfer of visuomotor skills from virtual to real settings [205]. n addition, the disembodying effect of VR can create interoceptive and visuo-proprioceptive issues that make the experience less immersive [206,207], besides causing discomfort, headaches, nausea, or instability [205,208]. Another issue is the lack of facial or trauma-focused interaction with the therapist, which can be problematic for the therapeutic alliance, limiting a shared experience, impeding regaining a sense of social safety [209-211] as well as the possibility of group therapy, a useful method in PTSD [212]. In this context, other sensorimotor technologies letting patients connect to their environment may be of interest as an alternative or adjunctive intervention. They could provide individuals with an SA over their environment.

The Example of GS Technologies to Treat PTSD

Overview

GS is a technique that involves using natural body gestures to control and generate sound. It uses sensors to detect the movement of an individual's body and then translates this movement into sound (Figure 1).

GS allows overcoming VR's limitations, as it is a screen-free alternative where only the auditory sense is augmented. It enables participant-body-environment coupling, the maintenance of contact with the therapist, and a more individualized virtual environment. It is applicable in case of language barrier, is more acceptable and user-friendly, and provides clinicians with ample opportunities for exploration. However, despite its potential benefits, some challenges might be encountered, such as acceptability. Indeed, patients may refuse to use GS; feel inhibited or startled while using it; and be exposed to various risks, such as allergies to sensors or, in case of improper use, hearing loss due to excessive volume or physical injury from uncontrolled movement. In addition, it is important to exercise caution both in the moment and in selecting sounds according to the way GS is used. For instance, some sounds that might be used for providing a sense of security in some patients could trigger intrusive symptoms in others due to specific traumatic experiences (eg, water sounds for a patient who experienced drowning).

While GS has found applications in various health contexts, such as physical rehabilitation [213] and music therapy [214], its potential in psychiatric illnesses has largely remained untapped. By combining GS with sensorimotor exercises, patients with PTSD could use their bodily movements to generate sounds and manipulate and control virtual objects and environments, creating a more immersive and engaging experience that leaves room for interaction with the therapist. GS could, therefore, emphasize the body and the sensations of movements that are often overlooked in first-line therapies such as prolonged exposure. By directing attention to the body and the physical sensations experienced, GS could assist patients in connecting with their present moment experience and developing a greater SA [193,215]. Indeed, multisensory retroaction can reinforce patients' sense of control over their symptoms [213]. Importantly, GS can be combined with guided imagery (GI) to provide a personalized experience for patients, allowing them to manipulate, through sensorimotor exercises, the traumatic memories and facilitate their integration. GI is a behavioral mind-body intervention using appropriate scripts and imagination to manipulate representations and enable positive affective and body responses [216,217]. GI uses sensory integration to enhance affective and cognitive retrocontrol over hyperarousal through muscular relaxation and positive mental images. GI has been shown to improve depression, anxiety, and stress [218] and change the meaning of pain [219] and is also widely used in prolonged exposure as well as a self-management intervention to alleviate PTSD symptoms [220], especially hyperarousal. Nevertheless, individuals are not equal in their ability to engage in GI, and its efficacy is correlated with absorption abilities [221] and stays controversial [222]. Our team has experimented with GS and found that the technology could potentially help augment such abilities and thus mind-body interventions. We identified 5 potential applications of the system following the common course of existing therapies.

XSL•FO
RenderX

Figure 1. System diagram of gesture sonification illustrating (1) the integration of a wearable band (2) with sensors and a Bluetooth module, (3) which communicates with a computational device (4) running specific signal processing software. The device (5) processes the data through loudspeakers or headphones (6) following the therapist guidelines (7) to provide auditory feedback.

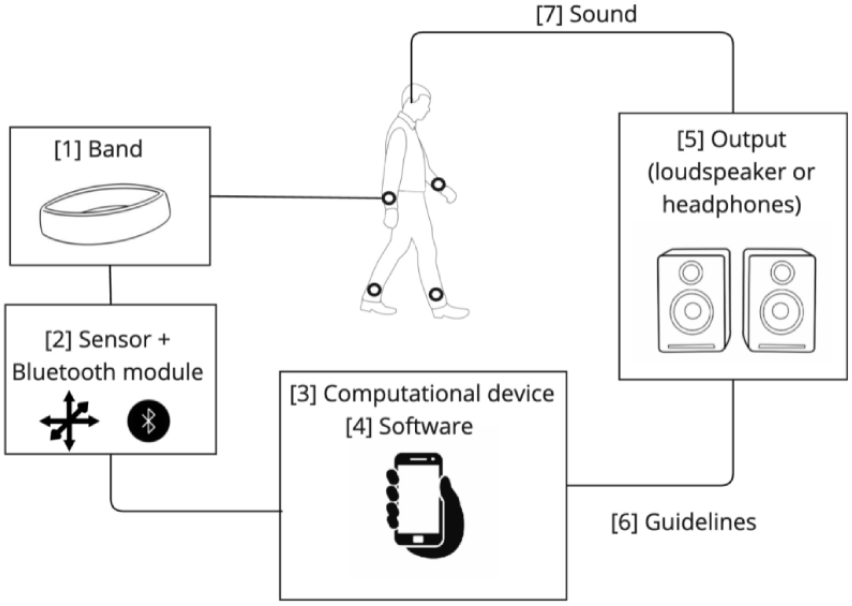

PTSD Primary Prevention

Functional alterations in brain circuitry identified in patients with PTSD could already be present before the TE and constitute risk factors for developing PTSD [83]. Occasionally, they may exist before PTSD and worsen after the onset of PTSD. Indeed, PTSD has been found to be highly heritable due to epigenetic factors [83,223-226], which means that prior trauma and cumulative life adversity may induce alterations in brain circuitry responsible for vulnerability to PTSD [99,152,153].

Preventing PTSD, for instance, among at-risk populations such as women, refugees, and military and rescue workers [227] involves more than just averting TEs. It means identifying cognitive impairments predisposing for PTSD and enhancing or correcting these impairments. This includes improving cortical and hippocampus retrocontrol abilities, activating the central executive network or DMN, and regulating the SN to maintain threat detection at an optimal level. Ultimately, this entails bolstering the SA before the TE occurs, potentially averting SA failure, peritraumatic dissociation, and thus PTSD later on. GS could be used to assess vulnerability and prepare military and rescue workers before field operations.

In addition, GS could serve as an SA enhancer in the immediate and postimmediate phases following the TE. During the immediate phase, where defusing by talking [228] aims at reintegrating individuals in the present moment with the presence of the other, GS could help mitigate peritraumatic dissociation by amplifying environmental sounds and facilitating reorientation. In the postimmediate phase, during early psychological interventions or debriefing, which may not prevent the onset of PTSD but may improve subsequent adherence to necessary care [229-235], GS could help reduce stress activation by providing soothing sensorimotor environments.

Reducing Hyperarousal and Negative Alteration of Mood: Securitization

GS could potentially help patients with PTSD to reduce their hyperarousal symptoms. Patients may indeed engage their senses to focus on the present moment. Using personalized sound environments considered secure by patients (eg, water, beach, and fireplace) can help reduce physiological arousal (eg, elevated heart rate or increased muscle tension). Furthermore, by letting patients control with their body movements the volume of the sounds produced by the device, it may help them acquire the ability to filter the external sensory signals whose input is distorted in PTSD with hypervigilance.

By strengthening proprioception and sensory modality with movement, patients can learn to tame the external world, the one that has betrayed or shattered. GS would thus help construct coherence and new congruences in multisensory processing: sorting, filtering, or hierarchization. Enhancing the patient's self with an auditory signal, a vibrational charge to movement from the environment, offers the patient the opportunity to reverse defensive modes of withdrawal: it awakens curiosity and momentarily diverts anxious rumination about the external world. The environment becomes a facilitator of action, contributing to the reshaping of mental images and the reinforcement of motor planning. The patient actively participates, seeking to recombine the missing pattern in psychological trauma, creating sensory coherence and ideas through action. This synergistic process involving the postural system and the psychomotor feedback allowed by GS nourishes past sensorimotor experiences while updating new ones. Patients are placed in a position to synchronize with the world and derive pleasure from this tuning.

Reducing Dissociation: Reassociation of Patients With Their Body

Importantly, GS could help patients with PTSD to reduce dissociation symptoms. Sensorimotor therapy facilitates the

XSL•FO
RenderX

reconstruction of the body schema to align with functional reality. Patients are guided to inhabit their body and reshape their representations within a somatic reality. Through the pursuit of body wholeness and conscious work on muscle tone, the therapist contributes to this aim. Dance, slow movement in dynamic relaxation, and postural work (eg, Qi gong and yoga), through the antagonistic interplay of tension and relaxation, create a dynamic musculo-psychological synergy. Tonus is initially a function of dialogue with the world. Among the wide variety of sensorimotor or body-oriented therapies, GS could be combined with dance movement therapy (DMT) for instance. DMT has demonstrated its effectiveness in treating PTSD [236,237]. It calls for the awakening of supportive responses: a confident relationship with support generates a repelling force. Attention is focused on coordination and responses to orientation, alignment, and balance in rhythm. DMT is a therapeutic mediation that allows for the revisiting of early tuning connections; engages in the quality of gaze, joint attention, and gesturing; and addresses support, verticalization, and initial connections. Combined with GS, patients may be able to control with their body various elements (eg, rhythm, timbre, melody, harmony, and dynamics) of the music to which they are dancing and position themselves in a realm beneath language: the reality of music interplaying with movement that spatializes and shifts. This dual modality is a prelude to the transformation of psychological trauma through somato-psychological adjustments, when the body becomes “the thought-thinking apparatus” [238]. In essence, GS combined with sensorimotor therapies would work toward the generation of new body images, reinforcing tonic regulation, self-esteem, and primary narcissism and finally restoring not only proprioceptive feelings enabling patients to reintegrate their body, as well as a sense of contact to the ground (ie, reducing depersonalization), but also the SA over incoming sensory signals from external objects (ie, reducing derealization).

Reducing Intrusion Symptoms and Avoidance: Exposure to the TE

GS involving sounds related to triggers of intrusive symptoms or to the TE itself may help patients with PTSD to engage with their environment in a more active, meaningful way than mere passive exposure, controlling intrusive symptoms and overcoming avoidance behaviors. Gradually, they would expose themselves to triggers that they are avoiding, gaining progressively the ability to control them. GS may indeed let the patient control the level of gradual exposure to TE triggers by themself. As in prolonged exposure, the uncontrollability of intrusion symptoms or triggers would then be reduced by the gained ability to control the irruption, intensity, and end of triggers or intrusion symptoms through body-environment interactions.

Social and Cognitive Rehabilitation

Social cognition is disturbed in PTSD [143], especially social perception, affective theory of mind, affective empathy, and social interactions. Of note, social cognition is more altered if the TE is interpersonal (ie, originating from human actions) than not interpersonal [239]. Patients with PTSD exhibit an overactive SN, which has a crucial function in affective empathy, which in turn also relates to a higher sensitivity to stress [240] or the negative social impact [241]. The potential higher affective empathy may probably explain the increasing attention on social rehabilitation in PTSD treatment, such as communication training, group therapies, or community programs [242]. Indeed, social rehabilitation favors the sense of belonging, just as collective commemorations of disruptive events are perceived useful by victims and help rebuild social links [243].

Sensorimotor group therapies have recently been found to be effective in treating C-PTSD [244,245]. Just as GS would help restore localization and a sense of contact with the ground and external objects by increasing the sensorimotor afferences, it could also restore the sense of contact with other individuals and help in the process of mourning. The involvement of the body and the sensorimotor pleasures that reconstruct and reshape body schemas enhance bodily sensations of gathering or synchronization. This, in turn, boosts the effective connectivity of the mirror neuron system, which is crucial for social cognition [246].

Conclusions

In this paper, we offer a perspective of PTSD in the light of the essential role of the SA in functional behavior. Recognizing the TE as a sensorimotor failure, we view psychotrauma as a profound breakdown of the SA. We offer an enactive perspective of PTSD, where symptoms represent efforts to uncover and restore the SA in response to the TE. We advocate for PTSD therapy to develop interventions fostering direct engagement with one’s body and environment, gradually rebuilding the SA. We suggest that agency-based therapies could mitigate PTSD risk and enhance treatment effectiveness. GS, alongside active somatic or movement-based approaches such as sensorimotor therapy, DMT, and somatic experiencing, directs attention to the body and sensory experiences, fostering present moment connection and bolstering the SA. Overall, this paper highlights the significance of the SA in psychotrauma and PTSD, offering insights to enhance treatments and advocating for further research in this critical area.

Conflicts of Interest

In the past years, FS cofounded and received compensation from BeSound SAS and Nested Minds LLC. In the past months, VA also received compensation from BeSound SAS.

References

XSL•FO
RenderX

1. Karam EG, Friedman MJ, Hill ED, Kessler RC, McLaughlin KA, Petukhova M, et al. Cumulative traumas and risk thresholds: 12-month PTSD in the World Mental Health (WMH) surveys. Depress Anxiety 2014 Feb;31(2):130-142 [FREE Full text] [doi: 10.1002/da.22169] [Medline: 23983056]
2. Kessler RC, Sonnega A, Bromet E, Hughes M, Nelson CB. Posttraumatic stress disorder in the National Comorbidity Survey. Arch Gen Psychiatry 1995 Dec;52(12):1048-1060. [doi: 10.1001/archpsyc.1995.03950240066012] [Medline: 7492257]
3. Charney ME, Hellberg SN, Bui E, Simon NM. Evidenced-based treatment of posttraumatic stress disorder: an updated review of validated psychotherapeutic and pharmacological approaches. Harv Rev Psychiatry 2018;26(3):99-115. [doi: 10.1097/HRP.0000000000000186] [Medline: 29734225]
4. Schnurr PP, Chard KM, Ruzek JI, Chow BK, Resick PA, Foa EB, et al. Comparison of prolonged exposure vs cognitive processing therapy for treatment of posttraumatic stress disorder among US veterans: a randomized clinical trial. JAMA Netw Open 2022 Jan 04;5(1):e2136921 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.36921] [Medline: 35044471]
5. Lee AH, Brown E. Examining the effectiveness of trauma-focused cognitive behavioral therapy on children and adolescents' executive function. Child Abuse Negl 2022 Apr;126:105516. [doi: 10.1016/j.chiabu.2022.105516] [Medline: 35093801]
6. Steenkamp MM, Litz BT, Hoge CW, Marmar CR. Psychotherapy for military-related PTSD: a review of randomized clinical trials. JAMA 2015 Aug 04;314(5):489-500. [doi: 10.1001/jama.2015.8370] [Medline: 26241600]
7. Fonzo GA, Federchenco V, Lara A. Predicting and managing treatment non-response in posttraumatic stress disorder. Curr Treat Options Psychiatry 2020 Jun 23;7(2):70-87 [FREE Full text] [doi: 10.1007/s40501-020-00203-1] [Medline: 33344106]
8. Rizzo A, Reger G, Gahm G, Difede J, Rothbaum B. Virtual reality exposure therapy for combat-related PTSD. In: LeDoux JE, Keane T, Shiromani P, editors. Post-Traumatic Stress Disorder: Basic Science and Clinical Practice. Totowa, NJ: Humana Press; 2009:375-399.
9. Beidel DC, Frueh BC, Neer SM, Bowers CA, Trachik B, Uhde TW, et al. Trauma management therapy with virtual-reality augmented exposure therapy for combat-related PTSD: a randomized controlled trial. J Anxiety Disord 2019 Jan;61:64-74. [doi: 10.1016/j.janxdis.2017.08.005] [Medline: 28865911]
10. Eshuis LV, van Gelderen MJ, van Zuiden M, Nijdam MJ, Vermetten E, Olff M, et al. Efficacy of immersive PTSD treatments: a systematic review of virtual and augmented reality exposure therapy and a meta-analysis of virtual reality exposure therapy. J Psychiatr Res 2021 Nov;143:516-527 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.11.030] [Medline: 33248674]
11. Heo S, Park JH. Effects of virtual reality-based graded exposure therapy on PTSD symptoms: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Nov 29;19(23):15911 [FREE Full text] [doi: 10.3390/ijerph192315911] [Medline: 36497989]
12. Kothgassner OD, Goreis A, Kafka JX, Van Eickels RL, Plener PL, Felnhofer A. Virtual reality exposure therapy for posttraumatic stress disorder (PTSD): a meta-analysis. Eur J Psychotraumatol 2019 Aug 19;10(1):1654782 [FREE Full text] [doi: 10.1080/20008198.2019.1654782] [Medline: 31489138]
13. Adrien V, Bui E, Baubet T, Galletto CP, Duhamel J, McNamara A, et al. Beyond virtual reality: towards screen-free interfaces for post-traumatic stress disorder interventions. PsyArXiv Preprints Preprint posted online November 17, 2023. [doi: 10.31234/osf.io/24beh]
14. Rabellino D, Harricharan S, Frewen PA, Burin D, McKinnon MC, Lanius RA. "I can't tell whether it's my hand": a pilot study of the neurophenomenology of body representation during the rubber hand illusion in trauma-related disorders. Eur J Psychotraumatol 2016;7:32918 [FREE Full text] [doi: 10.3402/ejpt.v7.32918] [Medline: 27876453]
15. Rabellino D, Burin D, Harricharan S, Lloyd C, Frewen PA, McKinnon MC, et al. Altered sense of body ownership and agency in posttraumatic stress disorder and its dissociative subtype: a rubber hand illusion study. Front Hum Neurosci 2018 May 1;12:163 [FREE Full text] [doi: 10.3389/fnhum.2018.00163] [Medline: 29765311]
16. van der Hart O, van Dijke A, van Son M, Steele K. Somatoform dissociation in traumatized World War I combat soldiers. J Trauma Dissociation 2001 Mar 28;1(4):33-66. [doi: 10.1300/J229v01n04_03]
17. Ataria Y. Body Disownership in Complex Posttraumatic Stress Disorder. London, United Kingdom: Palgrave Macmillan; 2018.
18. Ataria Y, Gallagher S. Somatic apathy: body disownership in the context of torture. J Phenomenol Psychol 2015 Jun 10;46(1):105-122. [doi: 10.1163/15691624-12341286]
19. Bracha HS. Freeze, flight, fight, fright, faint: adaptationist perspectives on the acute stress response spectrum. CNS Spectr 2004 Sep;9(9):679-685 [FREE Full text] [doi: 10.1017/s1092852900001954] [Medline: 15337864]
20. Arieli A, Ataria Y. Helplessness: the inability to know-that you don't know-how. Philos Psychol 2018 May 23;31(6):948-968. [doi: 10.1080/09515089.2018.1468559]
21. Ataria Y. Sense of ownership and sense of agency during trauma. Phenom Cogn Sci 2013 Aug 30;14(1):199-212. [doi: 10.1007/s11097-013-9334-y]
22. Ataria Y. Trauma from an enactive perspective: the collapse of the knowing-how structure. Adapt Behav 2015 May 27;23(3):143-154. [doi: 10.1177/1059712315578542]
23. Ataria Y. Dissociation during trauma: the ownership-agency tradeoff model. Phenom Cogn Sci 2014 Oct 17;14(4):1037-1053. [doi: 10.1007/s11097-014-9392-9]

XSL•FO
RenderX

24. Linson A, Friston K. Reframing PTSD for computational psychiatry with the active inference framework. Cogn Neuropsychiatry 2019 Sep;24(5):347-368 [FREE Full text] [doi: 10.1080/13546805.2019.1665994] [Medline: 31564212]
25. Linson A, Parr T, Friston KJ. Active inference, stressors, and psychological trauma: a neuroethological model of (mal)adaptive explore-exploit dynamics in ecological context. Behav Brain Res 2020 Feb 17;380:112421 [FREE Full text] [doi: 10.1016/j.bbr.2019.112421] [Medline: 31830495]
26. Webb RE, Widseth JC. Traumas with and without a sense of agency. J Aggress Maltreatment Trauma 2009 Jul 23;18(5):532-546. [doi: 10.1080/10926770903050993]
27. Beste JE. The vulnerable self and loss of agency: trauma theory and the challenge to a Rahnerian theology of freedom and grace. In: Beste JE, editor. God and the Victim: Traumatic Intrusions on Grace and Freedom. Oxford, United Kingdom: Oxford University Press; 2007.
28. Frabetti M, Gayraud F, Auxéméry Y. [Study of agency in the discourse of women suffering from post-traumatic stress disorder in the aftermath of domestic violence]. Encephale 2023 Oct;49(5):516-524. [doi: 10.1016/j.encep.2022.09.002] [Medline: 36257851]
29. Tsakiris M, Schütz-Bosbach S, Gallagher S. On agency and body-ownership: phenomenological and neurocognitive reflections. Conscious Cogn 2007 Sep;16(3):645-660. [doi: 10.1016/j.concog.2007.05.012] [Medline: 17616469]
30. Braun N, Debener S, Spychala N, Bongartz E, Sörös P, Müller HH, et al. The senses of agency and ownership: a review. Front Psychol 2018 Apr 16;9:535 [FREE Full text] [doi: 10.3389/fpsyg.2018.00535] [Medline: 29713301]
31. Tsakiris M, Hesse MD, Boy C, Haggard P, Fink GR. Neural signatures of body ownership: a sensory network for bodily self-consciousness. Cereb Cortex 2007 Oct;17(10):2235-2244. [doi: 10.1093/cercor/bhl131] [Medline: 17138596]
32. Tsakiris M. Looking for myself: current multisensory input alters self-face recognition. PLoS One 2008;3(12):e4040 [FREE Full text] [doi: 10.1371/journal.pone.0004040] [Medline: 19107208]
33. Chancel M, Ehrsson HH. Which hand is mine? Discriminating body ownership perception in a two-alternative forced-choice task. Atten Percept Psychophys 2020 Nov;82(8):4058-4083 [FREE Full text] [doi: 10.3758/s13414-020-02107-x] [Medline: 32856222]
34. Tsakiris M. My body in the brain: a neurocognitive model of body-ownership. Neuropsychologia 2010 Feb;48(3):703-712. [doi: 10.1016/j.neuropsychologia.2009.09.034] [Medline: 19819247]
35. Ehrsson HH, Spence C, Passingham RE. That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. Science 2004 Aug 06;305(5685):875-877. [doi: 10.1126/science.1097011] [Medline: 15232072]
36. Guterstam A, Larsson DE, Szczotka J, Ehrsson HH. Duplication of the bodily self: a perceptual illusion of dual full-body ownership and dual self-location. R Soc Open Sci 2020 Dec;7(12):201911 [FREE Full text] [doi: 10.1098/rsos.201911] [Medline: 33489299]
37. Litwin P, Zybura B, Motyka P. Tactile information counteracts the attenuation of rubber hand illusion attributable to increased visuo-proprioceptive divergence. PLoS One 2020 Dec 30;15(12):e0244594 [FREE Full text] [doi: 10.1371/journal.pone.0244594] [Medline: 33378385]
38. Litwin P. Extending Bayesian models of the rubber hand illusion. Multisens Res 2020 Jan 08;33(2):127-160. [doi: 10.1163/22134808-20191440] [Medline: 31648196]
39. Gallagher S. Philosophical conceptions of the self: implications for cognitive science. Trends Cogn Sci 2000 Jan;4(1):14-21. [doi: 10.1016/s1364-6613(99)01417-5] [Medline: 10637618]
40. Haggard P, Chambon V. Sense of agency. Curr Biol 2012 May 22;22(10):R390-R392 [FREE Full text] [doi: 10.1016/j.cub.2012.02.040] [Medline: 22625851]
41. Blakemore SJ, Wolpert DM, Frith CD. Abnormalities in the awareness of action. Trends Cogn Sci 2002 Jun 01;6(6):237-242. [doi: 10.1016/s1364-6613(02)01907-1] [Medline: 12039604]
42. Gallagher S, Trigg D. Agency and anxiety: delusions of control and loss of control in schizophrenia and agoraphobia. Front Hum Neurosci 2016 Sep 26;10:459 [FREE Full text] [doi: 10.3389/fnhum.2016.00459] [Medline: 27725796]
43. Seghezzi S, Giannini G, Zapparoli L. Neurofunctional correlates of body-ownership and sense of agency: a meta-analytical account of self-consciousness. Cortex 2019 Dec;121:169-178. [doi: 10.1016/j.cortex.2019.08.018] [Medline: 31629195]
44. Tsakiris M, Longo MR, Haggard P. Having a body versus moving your body: neural signatures of agency and body-ownership. Neuropsychologia 2010 Jul;48(9):2740-2749. [doi: 10.1016/j.neuropsychologia.2010.05.021] [Medline: 20510255]
45. Farrer C, Franck N, Paillard J, Jeannerod M. The role of proprioception in action recognition. Conscious Cogn 2003 Dec;12(4):609-619. [doi: 10.1016/s1053-8100(03)00047-3] [Medline: 14656504]
46. Wegner DM. Précis of the illusion of conscious will. Behav Brain Sci 2004 Oct;27(5):649-59; discussion 659. [doi: 10.1017/s0140525x04000159] [Medline: 15895616]
47. Synofzik M, Vosgerau G, Newen A. Beyond the comparator model: a multifactorial two-step account of agency. Conscious Cogn 2008 Mar;17(1):219-239. [doi: 10.1016/j.concog.2007.03.010] [Medline: 17482480]
48. Moore JW, Fletcher PC. Sense of agency in health and disease: a review of cue integration approaches. Conscious Cogn 2012 Mar;21(1):59-68 [FREE Full text] [doi: 10.1016/j.concog.2011.08.010] [Medline: 21920777]
49. Caspar EA, De Beir A, Magalhaes De Saldanha Da Gama PA, Yernaux F, Cleeremans A, Vanderborght B. New frontiers in the rubber hand experiment: when a robotic hand becomes one's own. Behav Res Methods 2015 Sep;47(3):744-755. [doi: 10.3758/s13428-014-0498-3] [Medline: 24942249]

XSL•FO
RenderX

50. Pyasik M, Furlanetto T, Pia L. The role of body-related afferent signals in human sense of agency. J Exp Neurosci 2019 May 16;13:1179069519849907 [FREE Full text] [doi: 10.1177/1179069519849907] [Medline: 31205423]
51. Schütz-Bosbach S, Avenanti A, Aglioti SM, Haggard P. Don't do it! Cortical inhibition and self-attribution during action observation. J Cogn Neurosci 2009 Jun;21(6):1215-1227. [doi: 10.1162/jocn.2009.21068] [Medline: 18702585]
52. Seghezzi S, Zirone E, Paulesu E, Zapparoli L. The brain in (willed) action: a meta-analytical comparison of imaging studies on motor intentionality and sense of agency. Front Psychol 2019 Apr 12;10:804 [FREE Full text] [doi: 10.3389/fpsyg.2019.00804] [Medline: 31031676]
53. Gordon EM, Chauvin RJ, Van AN, Rajesh A, Nielsen A, Newbold DJ, et al. A somato-cognitive action network alternates with effector regions in motor cortex. Nature 2023 May;617(7960):351-359 [FREE Full text] [doi: 10.1038/s41586-023-05964-2] [Medline: 37076628]
54. Farrer C, Frith CD. Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. Neuroimage 2002 Mar;15(3):596-603. [doi: 10.1006/nimg.2001.1009] [Medline: 11848702]
55. Bulot V, Thomas P, Delevoye-Turrell Y. [Sense of agency: experiencing is not judging]. Encephale 2007 Sep;33(4 Pt 1):603-608. [doi: 10.1016/s0013-7006(07)92060-6] [Medline: 18033150]
56. Gallagher S. Self-defense: deflecting deflationary and eliminativist critiques of the sense of ownership. Front Psychol 2017 Sep 21;8:1612 [FREE Full text] [doi: 10.3389/fpsyg.2017.01612] [Medline: 28970813]
57. Daprati E, Franck N, Georgieff N, Proust J, Pacherie E, Dalery J, et al. Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. Cognition 1997 Dec;65(1):71-86. [doi: 10.1016/s0010-0277(97)00039-5] [Medline: 9455171]
58. Franck N, Farrer C, Georgieff N, Marie-Cardine M, Daléry J, d'Amato T, et al. Defective recognition of one's own actions in patients with schizophrenia. Am J Psychiatry 2001 Mar;158(3):454-459. [doi: 10.1176/appi.ajp.158.3.454] [Medline: 11229988]
59. Stephens GL, Graham G. When Self–Consciousness Breaks: Alien Voices and Inserted Thoughts. Cambridge, MA: The MIT Press; 2000.
60. Haggard P, Tsakiris M. The experience of agency: feelings, judgments, and responsibility. Curr Dir Psychol Sci 2009 Aug 01;18(4):242-246. [doi: 10.1111/j.1467-8721.2009.01644.x]
61. Bermúdez JL. Bodily awareness and self-consciousness. In: Gallagher S, editor. The Oxford Handbook of the Self. Oxford, United Kingdom: Oxford University Press; 2011.
62. Gallagher M, Colzi C, Sedda A. Dissociation of proprioceptive drift and feelings of ownership in the somatic rubber hand illusion. Acta Psychol (Amst) 2021 Jan;212:103192 [FREE Full text] [doi: 10.1016/j.actpsy.2020.103192] [Medline: 33137614]
63. Grünbaum T. The feeling of agency hypothesis: a critique. Synthese 2015 Feb 24;192(10):3313-3337. [doi: 10.1007/s11229-015-0704-6]
64. Decety J, Perani D, Jeannerod M, Bettinardi V, Tadary B, Woods R, et al. Mapping motor representations with positron emission tomography. Nature 1994 Oct 13;371(6498):600-602. [doi: 10.1038/371600a0] [Medline: 7935791]
65. Botvinick M, Cohen J. Rubber hands 'feel' touch that eyes see. Nature 1998 Feb 19;391(6669):756. [doi: 10.1038/35784] [Medline: 9486643]
66. Aarts H, Custers R, Wegner DM. On the inference of personal authorship: enhancing experienced agency by priming effect information. Conscious Cogn 2005 Sep;14(3):439-458. [doi: 10.1016/j.concog.2004.11.001] [Medline: 16091264]
67. Moore JW, Obhi SS. Intentional binding and the sense of agency: a review. Conscious Cogn 2012 Mar;21(1):546-561. [doi: 10.1016/j.concog.2011.12.002] [Medline: 22240158]
68. Dewey JA, Knoblich G. Do implicit and explicit measures of the sense of agency measure the same thing? PLoS One 2014 Oct 16;9(10):e110118 [FREE Full text] [doi: 10.1371/journal.pone.0110118] [Medline: 25330184]
69. Tapal A, Oren E, Dar R, Eitam B. The sense of agency scale: a measure of consciously perceived control over one's mind, body, and the immediate environment. Front Psychol 2017 Sep 12;8:1552 [FREE Full text] [doi: 10.3389/fpsyg.2017.01552] [Medline: 28955273]
70. Hurault JC, Broc G, Crône L, Tedesco A, Brunel L. Measuring the sense of agency: a French adaptation and validation of the sense of agency scale (F-SoAS). Front Psychol 2020 Oct 8;11:584145 [FREE Full text] [doi: 10.3389/fpsyg.2020.584145] [Medline: 33132992]
71. Kalckert A, Ehrsson HH. Moving a rubber hand that feels like your own: a dissociation of ownership and agency. Front Hum Neurosci 2012 Mar 14;6:40 [FREE Full text] [doi: 10.3389/fnhum.2012.00040] [Medline: 22435056]
72. Kalckert A, Ehrsson HH. The moving rubber hand illusion revisited: comparing movements and visuotactile stimulation to induce illusory ownership. Conscious Cogn 2014 May;26:117-132 [FREE Full text] [doi: 10.1016/j.concog.2014.02.003] [Medline: 24705182]
73. Gallagher S. Neurocognitive models of schizophrenia: a neurophenomenological critique. Psychopathology 2004;37(1):8-19. [doi: 10.1159/000077014] [Medline: 14988645]
74. Torregrossa LJ, Snodgress MA, Hong SJ, Nichols HS, Glerean E, Nummenmaa L, et al. Anomalous bodily maps of emotions in schizophrenia. Schizophr Bull 2019 Sep 11;45(5):1060-1067 [FREE Full text] [doi: 10.1093/schbul/sby179] [Medline: 30551180]

XSL•FO
RenderX

75. Ardizzi M, Ambrosecchia M, Buratta L, Ferri F, Ferroni F, Palladini B, et al. The motor roots of minimal self disorders in schizophrenia. Schizophr Res 2020 Apr;218:302-303. [doi: 10.1016/j.schres.2020.03.007] [Medline: 32171636]
76. Pitman RK, Rasmusson AM, Koenen KC, Shin LM, Orr SP, Gilbertson MW, et al. Biological studies of post-traumatic stress disorder. Nat Rev Neurosci 2012 Nov;13(11):769-787 [FREE Full text] [doi: 10.1038/nrn3339] [Medline: 23047775]
77. Myers CS. Shell Shock in France, 1914-1918: Based on a War Diary. Cambridge, United Kingdom: Cambridge University Press; 1940.
78. Forster GL, Simons RM, Baugh LA. Revisiting the role of the amygdala in posttraumatic stress disorder. In: Ferry B, editor. The Amygdala - Where Emotions Shape Perception, Learning and Memories. London, United Kingdom: IntechOpen; 2017.
79. Malikowska-Racia N, Salat K. Recent advances in the neurobiology of posttraumatic stress disorder: a review of possible mechanisms underlying an effective pharmacotherapy. Pharmacol Res 2019 Apr;142:30-49. [doi: 10.1016/j.phrs.2019.02.001] [Medline: 30742899]
80. Likhtik E, Stujenske JM, Topiwala MA, Harris AZ, Gordon JA. Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. Nat Neurosci 2014 Jan;17(1):106-113 [FREE Full text] [doi: 10.1038/nn.3582] [Medline: 24241397]
81. Likhtik E, Paz R. Amygdala-prefrontal interactions in (mal)adaptive learning. Trends Neurosci 2015 Mar;38(3):158-166 [FREE Full text] [doi: 10.1016/j.tins.2014.12.007] [Medline: 25583269]
82. Lopresto D, Schipper P, Homberg JR. Neural circuits and mechanisms involved in fear generalization: implications for the pathophysiology and treatment of posttraumatic stress disorder. Neurosci Biobehav Rev 2016 Jan;60:31-42. [doi: 10.1016/j.neubiorev.2015.10.009] [Medline: 26519776]
83. Ross DA, Arbuckle MR, Travis MJ, Dwyer JB, van Schalkwyk GI, Ressler KJ. An integrated neuroscience perspective on formulation and treatment planning for posttraumatic stress disorder: an educational review. JAMA Psychiatry 2017 Apr 01;74(4):407-415 [FREE Full text] [doi: 10.1001/jamapsychiatry.2016.3325] [Medline: 28273291]
84. Liberzon I, Abelson JL. Context processing and the neurobiology of post-traumatic stress disorder. Neuron 2016 Oct 05;92(1):14-30 [FREE Full text] [doi: 10.1016/j.neuron.2016.09.039] [Medline: 27710783]
85. Shalev A, Liberzon I, Marmar C. Post-traumatic stress disorder. N Engl J Med 2017 Jun 22;376(25):2459-2469. [doi: 10.1056/NEJMra1612499] [Medline: 28636846]
86. Akiki TJ, Averill CL, Abdallah CG. A network-based neurobiological model of PTSD: evidence from structural and functional neuroimaging studies. Curr Psychiatry Rep 2017 Sep 19;19(11):81 [FREE Full text] [doi: 10.1007/s11920-017-0840-4] [Medline: 28924828]
87. Brunet A, Weiss DS, Metzler TJ, Best SR, Neylan TC, Rogers C, et al. The Peritraumatic Distress Inventory: a proposed measure of PTSD criterion A2. Am J Psychiatry 2001 Sep;158(9):1480-1485. [doi: 10.1176/appi.ajp.158.9.1480] [Medline: 11532735]
88. Kozlowska K, Walker P, McLean L, Carrive P. Fear and the defense cascade: clinical implications and management. Harv Rev Psychiatry 2015;23(4):263-287 [FREE Full text] [doi: 10.1097/HRP.0000000000000065] [Medline: 26062169]
89. van der Velden PG, Wittmann L. The independent predictive value of peritraumatic dissociation for PTSD symptomatology after type I trauma: a systematic review of prospective studies. Clin Psychol Rev 2008 Jul;28(6):1009-1020 [FREE Full text] [doi: 10.1016/j.cpr.2008.02.006] [Medline: 18406027]
90. Herman JL. Trauma and Recovery: The Aftermath of Violence--From Domestic Abuse to Political Terror. New York, NY: Basic Books; 2015.
91. Thompson-Hollands J, Jun JJ, Sloan DM. The association between peritraumatic dissociation and PTSD symptoms: the mediating role of negative beliefs about the self. J Trauma Stress 2017 Apr;30(2):190-194 [FREE Full text] [doi: 10.1002/jts.22179] [Medline: 28449364]
92. Spiegel D, Cardeña E. Disintegrated experience: the dissociative disorders revisited. J Abnorm Psychol 1991 Aug;100(3):366-378. [doi: 10.1037//0021-843x.100.3.366] [Medline: 1918616]
93. Vermetten E, Spiegel D. Trauma and dissociation: implications for borderline personality disorder. Curr Psychiatry Rep 2014 Feb;16(2):434. [doi: 10.1007/s11920-013-0434-8] [Medline: 24442670]
94. Seligman R, Kirmayer LJ. Dissociative experience and cultural neuroscience: narrative, metaphor and mechanism. Cult Med Psychiatry 2008 Mar;32(1):31-64 [FREE Full text] [doi: 10.1007/s11013-007-9077-8] [Medline: 18213511]
95. Schauer M, Elbert T. Dissociation following traumatic stress. J Psychol 2010 Jan;218(2). [doi: 10.1027/0044-3409/a000018]
96. Marmar CR, Weiss DS, Metzler T. Peritraumatic dissociation and posttraumatic stress disorder. In: Bremner JD, Marmar CR, editors. Trauma, Memory, and Dissociation. Washington, DC: American Psychiatric Association; 1998.
97. Marshall GN, Schell TL. Reappraising the link between peritraumatic dissociation and PTSD symptom severity: evidence from a longitudinal study of community violence survivors. J Abnorm Psychol 2002 Nov;111(4):626-636. [doi: 10.1037//0021-843x.111.4.626] [Medline: 12428776]
98. Birmes P, Brunet A, Carreras D, Ducassé JL, Charlet JP, Lauque D, et al. The predictive power of peritraumatic dissociation and acute stress symptoms for posttraumatic stress symptoms: a three-month prospective study. Am J Psychiatry 2003 Jul;160(7):1337-1339. [doi: 10.1176/appi.ajp.160.7.1337] [Medline: 12832251]
99. Ozer EJ, Best SR, Lipsey TL, Weiss DS. Predictors of posttraumatic stress disorder and symptoms in adults: a meta-analysis. Psychol Bull 2003 Jan;129(1):52-73. [doi: 10.1037/0033-2909.129.1.52] [Medline: 12555794]

XSL•FO
RenderX

100. Samuelson KW, Neylan TC, Lenoci M, Metzler TJ, Cardenas V, Weiner MW, et al. Longitudinal effects of PTSD on memory functioning. J Int Neuropsychol Soc 2009 Nov;15(6):853-861. [doi: 10.1017/S1355617709990282] [Medline: 19703319]
101. Tulving E. Episodic and semantic memory. In: Tulving E, Donaldson W, editors. Organization of Memory. Cambridge, MA: Academic Press; 1972:381-403.
102. Samuelson KW. Post-traumatic stress disorder and declarative memory functioning: a review. Dialogues Clin Neurosci 2011;13(3):346-351 [FREE Full text] [doi: 10.31887/DCNS.2011.13.2/ksamuelson] [Medline: 22033732]
103. van der Hart O, Nijenhuis ER, Steele K. The Haunted Self: Structural Dissociation and the Treatment of Chronic Traumatization (Norton Series on Interpersonal Neurobiology). New York, NY: W. W. Norton & Company; 2006.
104. Nijenhuis E, van der Hart O, Steele K. Trauma-related structural dissociation of the personality. Act Nerv Super 2017 Feb 23;52:1-23. [doi: 10.1007/bf03379560]
105. Tulving E. Memory and consciousness. Can Psychol 1985;26(1):1-12. [doi: 10.1037/h0080017]
106. Wheeler MA, Stuss DT, Tulving E. Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. Psychol Bull 1997 May;121(3):331-354. [doi: 10.1037/0033-2909.121.3.331] [Medline: 9136640]
107. Woodward SH, Kaloupek DG, Grande LJ, Stegman WK, Kutter CJ, Leskin L, et al. Hippocampal volume and declarative memory function in combat-related PTSD. J Int Neuropsychol Soc 2009 Nov;15(6):830-839. [doi: 10.1017/S1355617709990476] [Medline: 19703322]
108. Acheson DT, Gresack JE, Risbrough VB. Hippocampal dysfunction effects on context memory: possible etiology for posttraumatic stress disorder. Neuropharmacology 2012 Feb;62(2):674-685 [FREE Full text] [doi: 10.1016/j.neuropharm.2011.04.029] [Medline: 21596050]
109. Rougemont-Bücking A, Linnman C, Zeffiro TA, Zeidan MA, Lebron-Milad K, Rodriguez-Romaguera J, et al. Altered processing of contextual information during fear extinction in PTSD: an fMRI study. CNS Neurosci Ther 2011 Aug;17(4):227-236. [doi: 10.1111/j.1755-5949.2010.00152.x] [Medline: 20406268]
110. Sheynin J, Liberzon I. Circuit dysregulation and circuit-based treatments in posttraumatic stress disorder. Neurosci Lett 2017 May 10;649:133-138 [FREE Full text] [doi: 10.1016/j.neulet.2016.11.014] [Medline: 27845239]
111. Vitriol V, Cancino A, Weil K, Salgado C, Asenjo MA, Potthoff S. Depression and psychological trauma: an overview integrating current research and specific evidence of studies in the treatment of depression in public mental health services in chile. Depress Res Treat 2014;2014:608671 [FREE Full text] [doi: 10.1155/2014/608671] [Medline: 24695633]
112. Kascakova N, Furstova J, Hasto J, Madarasova Geckova A, Tavel P. The unholy trinity: childhood trauma, adulthood anxiety, and long-term pain. Int J Environ Res Public Health 2020 Jan 08;17(2):414 [FREE Full text] [doi: 10.3390/ijerph17020414] [Medline: 31936285]
113. Semiz UB, Inanc L, Bezgin CH. Are trauma and dissociation related to treatment resistance in patients with obsessive-compulsive disorder? Soc Psychiatry Psychiatr Epidemiol 2014 Aug;49(8):1287-1296. [doi: 10.1007/s00127-013-0787-7] [Medline: 24213522]
114. Yehuda R, Hoge CW, McFarlane AC, Vermetten E, Lanius RA, Nievergelt CM, et al. Post-traumatic stress disorder. Nat Rev Dis Primers 2015 Oct 08;1:15057. [doi: 10.1038/nrdp.2015.57] [Medline: 27189040]
115. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Association; 2013.
116. Weathers FW, Bovin MJ, Lee DJ, Sloan DM, Schnurr PP, Kaloupek DG, et al. The clinician-administered PTSD scale for DSM-5 (CAPS-5): development and initial psychometric evaluation in military veterans. Psychol Assess 2018 Mar;30(3):383-395 [FREE Full text] [doi: 10.1037/pas0000486] [Medline: 28493729]
117. Wolf EJ, Hawn SE, Sullivan DR, Miller MW, Sanborn V, Brown E, et al. Neurobiological and genetic correlates of the dissociative subtype of posttraumatic stress disorder. J Psychopathol Clin Sci 2023 May;132(4):409-427. [doi: 10.1037/abn0000795] [Medline: 37023279]
118. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak 2021 Nov 09;21(Suppl 6):206 [FREE Full text] [doi: 10.1186/s12911-021-01534-6] [Medline: 34753471]
119. Brewin CR, Cloitre M, Hyland P, Shevlin M, Maercker A, Bryant RA, et al. A review of current evidence regarding the ICD-11 proposals for diagnosing PTSD and complex PTSD. Clin Psychol Rev 2017 Dec;58:1-15. [doi: 10.1016/j.cpr.2017.09.001] [Medline: 29029837]
120. International statistical classification of diseases and related health problems (ICD). World Health Organization. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2024-06-05]
121. Herman JL. Complex PTSD: a syndrome in survivors of prolonged and repeated trauma. J Trauma Stress 1992 Jul;5(3):377-391 [FREE Full text] [doi: 10.1002/jts.2490050305]
122. Maercker A, Cloitre M, Bachem R, Schlumpf YR, Khoury B, Hitchcock C, et al. Complex post-traumatic stress disorder. Lancet 2022 Jul 02;400(10345):60-72. [doi: 10.1016/S0140-6736(22)00821-2] [Medline: 35780794]
123. Sachser C, Keller F, Goldbeck L. Complex PTSD as proposed for ICD-11: validation of a new disorder in children and adolescents and their response to trauma-focused cognitive behavioral therapy. J Child Psychol Psychiatry 2017 Feb;58(2):160-168. [doi: 10.1111/jcpp.12640] [Medline: 27677771]

XSL•FO
RenderX

124. Murray H, Grey N, Warnock-Parkes E, Kerr A, Wild J, Clark DM, et al. Ten misconceptions about trauma-focused CBT for PTSD. Cogn Behav Therap 2022 Jul 22;15:s1754470x22000307 [FREE Full text] [doi: 10.1017/S1754470X22000307] [Medline: 36247408]
125. van Minnen A, Harned MS, Zoellner L, Mills K. Examining potential contraindications for prolonged exposure therapy for PTSD. Eur J Psychotraumatol 2012;3 [FREE Full text] [doi: 10.3402/ejpt.v3i0.18805] [Medline: 22893847]
126. Halvorsen J, Stenmark H, Neuner F, Nordahl HM. Does dissociation moderate treatment outcomes of narrative exposure therapy for PTSD? A secondary analysis from a randomized controlled clinical trial. Behav Res Ther 2014 Jun;57:21-28. [doi: 10.1016/j.brat.2014.03.010] [Medline: 24762779]
127. Hoeboer CM, De Kleine RA, Molendijk ML, Schoorl M, Oprel DA, Mouthaan J, et al. Impact of dissociation on the effectiveness of psychotherapy for post-traumatic stress disorder: meta-analysis. BJPsych Open 2020 May 19;6(3):e53 [FREE Full text] [doi: 10.1192/bjo.2020.30] [Medline: 32423501]
128. Di Paolo EA, Buhrmann T, Barandiaran XE. The sense of agency. In: Sensorimotor: Life An Enactive Proposal. Oxford, United Kingdom: Oxford University Press; 2017.
129. Bomyea J, Lang AJ. Accounting for intrusive thoughts in PTSD: contributions of cognitive control and deliberate regulation strategies. J Affect Disord 2016 Mar 01;192:184-190 [FREE Full text] [doi: 10.1016/j.jad.2015.12.021] [Medline: 26741045]
130. Falsetti SA, Monnier J, Davis JL, Resnick HS. Intrusive thoughts in posttraumatic stress disorder. J Cogn Psychother 2002 Jun;16(2):127-143. [doi: 10.1891/jcop.16.2.127.63993]
131. Shipherd JC, Salters-Pedneault K. Attention, memory, intrusive thoughts, and acceptance in PTSD: an update on the empirical literature for clinicians. Cogn Behav Pract 2008 Nov;15(4):349-363 [FREE Full text] [doi: 10.1016/j.cbpra.2008.01.003]
132. Bryant RA, Sackville T, Dang ST, Moulds M, Guthrie R. Treating acute stress disorder: an evaluation of cognitive behavior therapy and supportive counseling techniques. Am J Psychiatry 1999 Nov;156(11):1780-1786. [doi: 10.1176/ajp.156.11.1780] [Medline: 10553743]
133. Kube T, Berg M, Kleim B, Herzog P. Rethinking post-traumatic stress disorder - a predictive processing perspective. Neurosci Biobehav Rev 2020 Jun;113:448-460. [doi: 10.1016/j.neubiorev.2020.04.014] [Medline: 32315695]
134. Shipherd JC, Beck JG. The effects of suppressing trauma-related thoughts on women with rape-related posttraumatic stress disorder. Behav Res Ther 1999 Feb;37(2):99-112. [doi: 10.1016/s0005-7967(98)00136-3] [Medline: 9990742]
135. Tull MT, Trotman A, Duplinsky MS, Reynolds EK, Daughters SB, Potenza MN, et al. The effect of posttraumatic stress disorder on risk-taking propensity among crack/cocaine users in residential substance abuse treatment. Depress Anxiety 2009;26(12):1158-1164 [FREE Full text] [doi: 10.1002/da.20637] [Medline: 19957281]
136. Tyler KA. Social and emotional outcomes of childhood sexual abuse: a review of recent research. Aggress Violent Behav 2002;7(6):567-589. [doi: 10.1016/s1359-1789(01)00047-7]
137. Terr LC. Childhood traumas: an outline and overview. Am J Psychiatry 1991 Jan;148(1):10-20. [doi: 10.1176/ajp.148.1.10] [Medline: 1824611]
138. Freud S, Strachey J. Beyond the Pleasure Principle. Scotts Valley, CA: CreateSpace Independent Publishing Platform; 2010.
139. Winnicott DW. Transitional objects and transitional phenomena. In: Caldwell L, Taylor Robinson H, editors. The Collected Works of D. W. Winnicott: Volume 3. Oxford, United Kingdom: Oxford Academic Press; 2016.
140. Thompson BL, Waltz J. Mindfulness and experiential avoidance as predictors of posttraumatic stress disorder avoidance symptom severity. J Anxiety Disord 2010 May;24(4):409-415. [doi: 10.1016/j.janxdis.2010.02.005] [Medline: 20304602]
141. Schneider A, Gudiño OG. Predicting avoidance symptoms in U.S. Latino youth exposed to community violence: the role of cultural values and behavioral inhibition. J Trauma Stress 2018 Aug;31(4):509-517. [doi: 10.1002/jts.22313] [Medline: 30058738]
142. Măirean C. Fear and avoidance of driving among drivers involved in a road traffic crash. The role of traumatic fear and driving cognitions. Transp Res Part F Traffic Psychol Behav 2020 Oct;74:322-329. [doi: 10.1016/j.trf.2020.08.026]
143. Couette M, Mouchabac S, Bourla A, Nuss P, Ferreri F. Social cognition in post-traumatic stress disorder: a systematic review. Br J Clin Psychol 2020 Jun;59(2):117-138. [doi: 10.1111/bjc.12238] [Medline: 31696974]
144. Galea S, Nandi A, Vlahov D. The epidemiology of post-traumatic stress disorder after disasters. Epidemiol Rev 2005;27:78-91. [doi: 10.1093/epirev/mxi003] [Medline: 15958429]
145. Neria Y, Olfson M, Gameroff MJ, Wickramaratne P, Pilowsky D, Verdeli H, et al. Trauma exposure and posttraumatic stress disorder among primary care patients with bipolar spectrum disorder. Bipolar Disord 2008 Jun;10(4):503-510 [FREE Full text] [doi: 10.1111/j.1399-5618.2008.00589.x] [Medline: 18452446]
146. Flynn AJ, Navarro GY, Basehore HK. PTSD avoidance symptoms associated with alcohol craving in treatment-seeking veteran population. J Dual Diagn 2022;18(3):135-143. [doi: 10.1080/15504263.2022.2089799] [Medline: 35761472]
147. Basedow LA, Kuitunen-Paul S, Wiedmann MF, Roessner V, Golub Y. Self-reported PTSD is associated with increased use of MDMA in adolescents with substance use disorders. Eur J Psychotraumatol 2021 Sep 28;12(1):1968140 [FREE Full text] [doi: 10.1080/20008198.2021.1968140] [Medline: 34603636]

XSL•FO
RenderX

148. Müller M, Vandeleur C, Rodgers S, Rössler W, Castelao E, Preisig M, et al. Posttraumatic stress avoidance symptoms as mediators in the development of alcohol use disorders after exposure to childhood sexual abuse in a Swiss community sample. Child Abuse Negl 2015 Aug;46:8-15 [FREE Full text] [doi: 10.1016/j.chiabu.2015.03.006] [Medline: 25828861]
149. Ciccone A, Ferrant A. Honte, Culpabilité Et Traumatisme. Paris, France: Dunod; 2015.
150. Freud S. Inhibitions, Symptoms and Anxiety. Tempe, AZ: Norton; 1926.
151. Klein M. Notes on some schizoid mechanisms. J Psychother Pract Res 1996;5(2):160-179 [FREE Full text] [Medline: 22700275]
152. Keane TM, Marshall AD, Taft CT. Posttraumatic stress disorder: etiology, epidemiology, and treatment outcome. Annu Rev Clin Psychol 2006;2:161-197. [doi: 10.1146/annurev.clinpsy.2.022305.095305] [Medline: 17716068]
153. Brewin CR, Andrews B, Valentine JD. Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. J Consult Clin Psychol 2000 Oct;68(5):748-766. [doi: 10.1037//0022-006x.68.5.748] [Medline: 11068961]
154. Cameron AY, Mamon D. Towards a better understanding of hypervigilance in combat veterans. Mil Behav Health 2019 Jan 07;7(2):206-217. [doi: 10.1080/21635781.2018.1526144]
155. Kimble M, Boxwala M, Bean W, Maletsky K, Halper J, Spollen K, et al. The impact of hypervigilance: evidence for a forward feedback loop. J Anxiety Disord 2014 Mar;28(2):241-245 [FREE Full text] [doi: 10.1016/j.janxdis.2013.12.006] [Medline: 24507631]
156. Shalev AY, Sahar T, Freedman S, Peri T, Glick N, Brandes D, et al. A prospective study of heart rate response following trauma and the subsequent development of posttraumatic stress disorder. Arch Gen Psychiatry 1998 Jun;55(6):553-559. [doi: 10.1001/archpsyc.55.6.553] [Medline: 9633675]
157. Bryant RA, Salmon K, Sinclair E, Davidson P. Heart rate as a predictor of posttraumatic stress disorder in children. Gen Hosp Psychiatry 2007;29(1):66-68. [doi: 10.1016/j.genhosppsych.2006.10.002] [Medline: 17189749]
158. Seligowski AV, Misganaw B, Duffy LA, Ressler KJ, Guffanti G. Leveraging large-scale genetics of PTSD and cardiovascular disease to demonstrate robust shared risk and improve risk prediction accuracy. Am J Psychiatry 2022 Nov 01;179(11):814-823 [FREE Full text] [doi: 10.1176/appi.ajp.21111113] [Medline: 36069022]
159. Wood J, Mathews A, Dalgleish T. Anxiety and cognitive inhibition. Emotion 2001 Jun;1(2):166-181. [doi: 10.1037/1528-3542.1.2.166] [Medline: 12899195]
160. Menon V, Uddin LQ. Saliency, switching, attention and control: a network model of insula function. Brain Struct Funct 2010 Jun;214(5-6):655-667 [FREE Full text] [doi: 10.1007/s00429-010-0262-0] [Medline: 20512370]
161. Stewart L, Brewin CR, Muggleton NG, Javardi AH, Tcheang L. Hypervigilance without threat: eye-movements reveal vigilant behaviours. In: Proceedings of the International Society for Traumatic Stress Studies 29th Annual Meeting. 2013 Presented at: ISTSS 2013; November 7-9, 2013; Philadelphia, PA.
162. Coll SY, Eustache F, Doidy F, Fraisse F, Peschanski D, Dayan J, et al. Avoidance behaviour generalizes to eye processing in posttraumatic stress disorder. Eur J Psychotraumatol 2022;13(1):2044661 [FREE Full text] [doi: 10.1080/20008198.2022.2044661]
163. Critchley HD, Harrison NA. Visceral influences on brain and behavior. Neuron 2013 Feb 20;77(4):624-638 [FREE Full text] [doi: 10.1016/j.neuron.2013.02.008] [Medline: 23439117]
164. Schoeller F, Haar AJ, Jain A, Maes P. Enhancing human emotions with interoceptive technologies. Phys Life Rev 2019 Dec;31:310-319. [doi: 10.1016/j.plrev.2019.10.008] [Medline: 31757602]
165. Schoeller F, Horowitz AH, Jain A, Maes P, Reggente N, Christov-Moore L, et al. Interoceptive technologies for psychiatric interventions: from diagnosis to clinical applications. Neurosci Biobehav Rev 2024 Jan;156:105478. [doi: 10.1016/j.neubiorev.2023.105478] [Medline: 38007168]
166. Anzieu D. The Skin Ego. London, United Kingdom: Yale University Press; 1989.
167. Schoeller F. Primary states of consciousness: a review of historical and contemporary developments. Conscious Cogn 2023 Aug;113:103536. [doi: 10.1016/j.concog.2023.103536] [Medline: 37321024]
168. Butler LD. Normative dissociation. Psychiatr Clin North Am 2006 Mar;29(1):45-62, viii. [doi: 10.1016/j.psc.2005.10.004] [Medline: 16530586]
169. Taylor MK, Morgan CA. Spontaneous and deliberate dissociative states in military personnel: relationships to objective performance under stress. Mil Med 2014 Sep;179(9):955-958. [doi: 10.7205/milmed-d-14-00081]
170. Morgan CA3, Taylor MK. Spontaneous and deliberate dissociative states in military personnel: are such states helpful? J Trauma Stress 2013 Aug;26(4):492-497. [doi: 10.1002/jts.21834] [Medline: 23893559]
171. Liriano F, Hatten C, Schwartz TL. Ketamine as treatment for post-traumatic stress disorder: a review. Drugs Context 2019 Apr 08;8:212305 [FREE Full text] [doi: 10.7573/dic.212305] [Medline: 31007698]
172. Elsouri KN, Kalhori S, Colunge D, Grabarczyk G, Hanna G, Carrasco C, et al. Psychoactive drugs in the management of post traumatic stress disorder: a promising new horizon. Cureus 2022 May 23;14(5):e25235 [FREE Full text] [doi: 10.7759/cureus.25235] [Medline: 35747039]
173. Feder A, Parides MK, Murrough JW, Perez AM, Morgan JE, Saxena S, et al. Efficacy of intravenous ketamine for treatment of chronic posttraumatic stress disorder: a randomized clinical trial. JAMA Psychiatry 2014 Jun;71(6):681-688. [doi: 10.1001/jamapsychiatry.2014.62] [Medline: 24740528]

XSL•FO
RenderX

174. Yarnell S. The use of medicinal marijuana for posttraumatic stress disorder: a review of the current literature. Prim Care Companion CNS Disord 2015 May 07;17(3):10.4088/PCC.15r01786 [FREE Full text] [doi: 10.4088/PCC.15r01786] [Medline: 26644963]
175. Niciu MJ, Shovestul BJ, Jaso BA, Farmer C, Luckenbaugh DA, Brutsche NE, et al. Features of dissociation differentially predict antidepressant response to ketamine in treatment-resistant depression. J Affect Disord 2018 May;232:310-315 [FREE Full text] [doi: 10.1016/j.jad.2018.02.049] [Medline: 29501990]
176. Breeksema JJ, Niemeijer A, Kuin B, Veraart J, Vermetten E, Kamphuis J, et al. Phenomenology and therapeutic potential of patient experiences during oral esketamine treatment for treatment-resistant depression: an interpretative phenomenological study. Psychopharmacology (Berl) 2023 Jul;240(7):1547-1560 [FREE Full text] [doi: 10.1007/s00213-023-06388-6] [Medline: 37222753]
177. Mathai DS, Nayak SM, Yaden DB, Garcia-Romeu A. Reconsidering "dissociation" as a predictor of antidepressant efficacy for esketamine. Psychopharmacology (Berl) 2023 Apr;240(4):827-836. [doi: 10.1007/s00213-023-06324-8] [Medline: 36729145]
178. Barrett FS, Carbonaro TM, Hurwitz E, Johnson MW, Griffiths RR. Double-blind comparison of the two hallucinogens psilocybin and dextromethorphan: effects on cognition. Psychopharmacology (Berl) 2018 Oct;235(10):2915-2927 [FREE Full text] [doi: 10.1007/s00213-018-4981-x] [Medline: 30062577]
179. Garcia LS, Comim CM, Valvassori SS, Réus GZ, Barbosa LM, Andreazza AC, et al. Acute administration of ketamine induces antidepressant-like effects in the forced swimming test and increases BDNF levels in the rat hippocampus. Prog Neuropsychopharmacol Biol Psychiatry 2008 Jan 01;32(1):140-144. [doi: 10.1016/j.pnpbp.2007.07.027] [Medline: 17884272]
180. Laruelle M, Abi-Dargham A, van Dyck CH, Gil R, D'Souza CD, Erdos J, et al. Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. Proc Natl Acad Sci U S A 1996 Aug 20;93(17):9235-9240 [FREE Full text] [doi: 10.1073/pnas.93.17.9235] [Medline: 8799184]
181. Chambers RA, Bremner JD, Moghaddam B, Southwick SM, Charney DS, Krystal JH. Glutamate and post-traumatic stress disorder: toward a psychobiology of dissociation. Semin Clin Neuropsychiatry 1999 Oct;4(4):274-281. [Medline: 10553033]
182. Maren S. Long-term potentiation in the amygdala: a mechanism for emotional learning and memory. Trends Neurosci 1999 Dec;22(12):561-567 [FREE Full text] [doi: 10.1016/s0166-2236(99)01465-4] [Medline: 10542437]
183. Ori R, Amos T, Bergman H, Soares-Weiser K, Ipser JC, Stein DJ. Augmentation of cognitive and behavioural therapies (CBT) with d-cycloserine for anxiety and related disorders. Cochrane Database Syst Rev 2015 May 10;2015(5):CD007803 [FREE Full text] [doi: 10.1002/14651858.CD007803.pub2] [Medline: 25957940]
184. Furukawa TA, Cipriani A, Cowen PJ, Leucht S, Egger M, Salanti G. Optimal dose of selective serotonin reuptake inhibitors, venlafaxine, and mirtazapine in major depression: a systematic review and dose-response meta-analysis. Lancet Psychiatry 2019 Jul;6(7):601-609 [FREE Full text] [doi: 10.1016/S2215-0366(19)30217-2] [Medline: 31178367]
185. Thierrée S, Raulin-Briot M, Legrand M, Le Gouge A, Vancappel A, Tudorache AC, et al. Combining trauma script exposure with rTMS to reduce symptoms of post-traumatic stress disorder: randomized controlled trial. Neuromodulation 2022 Jun;25(4):549-557. [doi: 10.1111/ner.13505] [Medline: 35667770]
186. Ferreri F, Mouchabac S, Sylvestre V, Millet B, El Hage W, Adrien V, et al. Repetitive transcranial magnetic stimulation (rTMS) in post-traumatic stress disorder: study protocol of a nationwide randomized controlled clinical trial of neuro-enhanced psychotherapy "TraumaStim". Brain Sci 2023 Aug 31;13(9):1274 [FREE Full text] [doi: 10.3390/brainsci13091274] [Medline: 37759875]
187. van 't Wout-Frank M, Shea MT, Larson VC, Greenberg BD, Philip NS. Combined transcranial direct current stimulation with virtual reality exposure for posttraumatic stress disorder: feasibility and pilot results. Brain Stimul 2019;12(1):41-43 [FREE Full text] [doi: 10.1016/j.brs.2018.09.011] [Medline: 30266416]
188. Han J, Choi KM, Yang C, Kim HS, Park SS, Lee SH. Treatment efficacy of tDCS and predictors of treatment response in patients with post-traumatic stress disorder. J Affect Disord 2022 Dec 01;318:357-363. [doi: 10.1016/j.jad.2022.08.111] [Medline: 36055537]
189. Abelson JL, Curtis GC, Sagher O, Albucher RC, Harrigan M, Taylor SF, et al. Deep brain stimulation for refractory obsessive-compulsive disorder. Biol Psychiatry 2005 Mar 01;57(5):510-516. [doi: 10.1016/j.biopsych.2004.11.042] [Medline: 15737666]
190. Noble LJ, Souza RR, McIntyre CK. Vagus nerve stimulation as a tool for enhancing extinction in exposure-based therapies. Psychopharmacology (Berl) 2019 Jan;236(1):355-367 [FREE Full text] [doi: 10.1007/s00213-018-4994-5] [Medline: 30091004]
191. Souza RR, Robertson NM, Pruitt DT, Gonzales PA, Hays SA, Rennaker RL, et al. Vagus nerve stimulation reverses the extinction impairments in a model of PTSD with prolonged and repeated trauma. Stress 2019 Jul;22(4):509-520. [doi: 10.1080/10253890.2019.1602604] [Medline: 31010369]
192. Ogden P, Pain C, Fisher J. A sensorimotor approach to the treatment of trauma and dissociation. Psychiatr Clin North Am 2006 Mar;29(1):263-79, xi. [doi: 10.1016/j.psc.2005.10.012] [Medline: 16530597]
193. Ogden P, Fisher J. Sensorimotor Psychotherapy: Interventions For Trauma And Attachmnent. New York, NY: W. W. Norton & Company; Apr 27, 2015.

XSL•FO
RenderX

194. van Gelderen MJ, Nijdam MJ, Vermetten E. An innovative framework for delivering psychotherapy to patients with treatment-resistant posttraumatic stress disorder: rationale for interactive motion-assisted therapy. Front Psychiatry 2018 May 04;9:176 [FREE Full text] [doi: 10.3389/fpsyt.2018.00176] [Medline: 29780334]
195. Beaudoin MN. Agency and choice in the face of trauma: a narrative therapy map. J Syst Ther 2005 Dec;24(4):32-50 [FREE Full text] [doi: 10.1521/jsyt.2005.24.4.32]
196. Rothbaum BO, Rizzo AS, Difede J. Virtual reality exposure therapy for combat-related posttraumatic stress disorder. Ann N Y Acad Sci 2010 Oct;1208:126-132. [doi: 10.1111/j.1749-6632.2010.05691.x] [Medline: 20955334]
197. Webb AK, Vincent AL, Jin AB, Pollack MH. Physiological reactivity to nonideographic virtual reality stimuli in veterans with and without PTSD. Brain Behav 2015 Feb;5(2):e00304 [FREE Full text] [doi: 10.1002/brb3.304] [Medline: 25642387]
198. van 't Wout M, Spofford CM, Unger WS, Sevin EB, Shea MT. Skin conductance reactivity to standardized virtual reality combat scenes in veterans with PTSD. Appl Psychophysiol Biofeedback 2017 Sep;42(3):209-221. [doi: 10.1007/s10484-017-9366-0] [Medline: 28646388]
199. Roy MJ, Costanzo ME, Jovanovic T, Leaman S, Taylor P, Norrholm SD, et al. Heart rate response to fear conditioning and virtual reality in subthreshold PTSD. Stud Health Technol Inform 2013;191:115-119. [Medline: 23792855]
200. Deng W, Hu D, Xu S, Liu X, Zhao J, Chen Q, et al. The efficacy of virtual reality exposure therapy for PTSD symptoms: a systematic review and meta-analysis. J Affect Disord 2019 Oct 01;257:698-709. [doi: 10.1016/j.jad.2019.07.086] [Medline: 31382122]
201. Difede J, Cukor J, Jayasinghe N, Patt I, Jedel S, Spielman L, et al. Virtual reality exposure therapy for the treatment of posttraumatic stress disorder following September 11, 2001. J Clin Psychiatry 2007 Nov;68(11):1639-1647. [Medline: 18052556]
202. Norr AM, Smolenski DJ, Katz AC, Rizzo AA, Rothbaum BO, Difede J, et al. Virtual reality exposure versus prolonged exposure for PTSD: which treatment for whom? Depress Anxiety 2018 Jun;35(6):523-529. [doi: 10.1002/da.22751] [Medline: 29734488]
203. Difede J, Rothbaum BO, Rizzo AA, Wyka K, Spielman L, Reist C, et al. Enhancing exposure therapy for posttraumatic stress disorder (PTSD): a randomized clinical trial of virtual reality and imaginal exposure with a cognitive enhancer. Transl Psychiatry 2022 Jul 27;12(1):299 [FREE Full text] [doi: 10.1038/s41398-022-02066-x] [Medline: 35896533]
204. Gomez MA, Skiba RM, Snow JC. Graspable objects grab attention more than images do. Psychol Sci 2018 Feb;29(2):206-218 [FREE Full text] [doi: 10.1177/0956797617730599] [Medline: 29215960]
205. Harris DJ, Buckingham G, Wilson MR, Vine SJ. Virtually the same? How impaired sensory information in virtual reality may disrupt vision for action. Exp Brain Res 2019 Nov;237(11):2761-2766 [FREE Full text] [doi: 10.1007/s00221-019-05642-8] [Medline: 31485708]
206. El Ali A, Ney R, van Berlo ZM, Cesar P. Is that my heartbeat? Measuring and understanding modality-dependent cardiac interoception in virtual reality. IEEE Trans Vis Comput Graph 2023 Nov;29(11):4805-4815. [doi: 10.1109/TVCG.2023.3320228] [Medline: 37782606]
207. Kilteni K, Bergstrom I, Slater M. Drumming in immersive virtual reality: the body shapes the way we play. IEEE Trans Vis Comput Graph 2013 Apr;19(4):597-605 [FREE Full text] [doi: 10.1109/TVCG.2013.29] [Medline: 23428444]
208. Davis S, Nesbitt K, Nalivaiko E. Comparing the onset of cybersickness using the Oculus Rift and two virtual roller coasters. In: Proceedings of the 11th Australasian Conference on Interactive Entertainment. 2015 Presented at: IE 2015; January 27-30, 2015; Sydney, Australia.
209. Coulter S. Creating safety for trauma survivors: what can therapists do? Child Care Pract 2008 Jan 17;7(1):45-56. [doi: 10.1080/13575270108413233]
210. Ross N, Brown C, Johnstone M. Beyond medicalized approaches to violence and trauma: empowering social work practice. J Soc Work 2023 Jan 09;23(3):567-585. [doi: 10.1177/14680173221144557]
211. Botella C, Serrano B, Baños RM, Garcia-Palacios A. Virtual reality exposure-based therapy for the treatment of post-traumatic stress disorder: a review of its efficacy, the adequacy of the treatment protocol, and its acceptability. Neuropsychiatr Dis Treat 2015 Oct 3;11:2533-2545 [FREE Full text] [doi: 10.2147/NDT.S89542] [Medline: 26491332]
212. Schleyer W, Zona K, Quigley D, Spottswood M. Group therapy in primary care settings for the treatment of posttraumatic stress disorder: a systematic literature review. Gen Hosp Psychiatry 2022;77:1-10. [doi: 10.1016/j.genhosppsych.2022.03.010] [Medline: 35390567]
213. Hunt A, Wanderley MM, Paradis M. The importance of parameter mapping in electronic instrument design. J New Music Res 2003 Dec 1;32(4):429-440. [doi: 10.1076/jnmr.32.4.429.18853]
214. Thompson WF, Schlaug G. The healing power of music. Sci Am Mind 2015 Mar;26(2):32-41. [doi: 10.1038/scientificamericanmind0315-32]
215. Koch SC, Riege RF, Tisborn K, Biondo J, Martin L, Beelmann A. Effects of dance movement therapy and dance on health-related psychological outcomes. A meta-analysis update. Front Psychol 2019 Aug 20;10:1806 [FREE Full text] [doi: 10.3389/fpsyg.2019.01806] [Medline: 31481910]
216. Rossman ML. Guided Imagery for Self-Healing: An Essential Resource for Anyone Seeking Wellness. Tiburon, CA: HJ Kramer Incorporated; 2000.

XSL•FO
RenderX

217. Özü Ö. Guided imagery as a psychotherapeutic mind-body intervention in health psychology: a brief review of efficacy research. Eur J Psychol 2010 Nov 29;6(4). [doi: 10.5964/ejop.v6i4.232]
218. Apóstolo JL, Kolcaba K. The effects of guided imagery on comfort, depression, anxiety, and stress of psychiatric inpatients with depressive disorders. Arch Psychiatr Nurs 2009 Dec;23(6):403-411. [doi: 10.1016/j.apnu.2008.12.003] [Medline: 19926022]
219. Lewandowski W, Good M, Draucker CB. Changes in the meaning of pain with the use of guided imagery. Pain Manag Nurs 2005 Jun;6(2):58-67. [doi: 10.1016/j.pmn.2005.01.002] [Medline: 15970919]
220. Strauss J, Calhoun P, Marx C. Guided imagery as a therapeutic tool in post-traumatic stress disorder. In: LeDoux J, Keane T, Shiromani P, editors. Post-Traumatic Stress Disorder. Totowa, NJ: Humana Press; 2009.
221. Kwekkeboom K, Huseby-Moore K, Ward S. Imaging ability and effective use of guided imagery. Res Nurs Health 1998 Jun;21(3):189-198. [doi: 10.1002/(sici)1098-240x(199806)21:3<189::aid-nur2>3.0.co;2-d] [Medline: 9609504]
222. Neuner F. Safety first? Trauma exposure in PTSD. In: Neudeck P, Wittchen HU, editors. Exposure Therapy. New York, NY: Springer; 2012.
223. Yehuda R, Bierer LM. The relevance of epigenetics to PTSD: implications for the DSM-V. J Trauma Stress 2009 Oct;22(5):427-434 [FREE Full text] [doi: 10.1002/jts.20448] [Medline: 19813242]
224. Yehuda R, Daskalakis NP, Lehrner A, Desarnaud F, Bader HN, Makotkine I, et al. Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. Am J Psychiatry 2014 Aug;171(8):872-880 [FREE Full text] [doi: 10.1176/appi.ajp.2014.13121571] [Medline: 24832930]
225. Yehuda R, Daskalakis NP, Bierer LM, Bader HN, Klengel T, Holsboer F, et al. Holocaust exposure induced intergenerational effects on FKBP5 methylation. Biol Psychiatry 2016 Sep 01;80(5):372-380. [doi: 10.1016/j.biopsych.2015.08.005] [Medline: 26410355]
226. Yehuda R, Lehrner A. Intergenerational transmission of trauma effects: putative role of epigenetic mechanisms. World Psychiatry 2018 Oct;17(3):243-257. [doi: 10.1002/wps.20568] [Medline: 30192087]
227. Javidi H, Yadollahie M. Post-traumatic stress disorder. Int J Occup Environ Med 2012 Jan;3(1):2-9. [Medline: 23022845]
228. Fillion JS, Clements PT, Averill JB, Vigil GJ. Talking as a primary method of peer defusing for military personnel exposed to combat trauma. J Psychosoc Nurs Ment Health Serv 2002 Aug;40(8):40-49. [doi: 10.3928/0279-3695-20020801-15] [Medline: 12174514]
229. Stileman HM, Jones CA. Revisiting the debriefing debate: does psychological debriefing reduce PTSD symptomology following work-related trauma? A meta-analysis. Front Psychol 2023 Dec 21;14:1248924 [FREE Full text] [doi: 10.3389/fpsyg.2023.1248924] [Medline: 38204890]
230. Wessely S, Rose S, Bisson J. Brief psychological interventions ("debriefing") for trauma-related symptoms and the prevention of post traumatic stress disorder. Cochrane Database Syst Rev 2000;2002(2):CD000560 [FREE Full text] [doi: 10.1002/14651858.CD000560] [Medline: 10796724]
231. Bolwig TG. Debriefing after psychological trauma. Acta Psychiatr Scand 1998 Sep;98(3):169-170. [doi: 10.1111/j.1600-0447.1998.tb10062.x] [Medline: 9761401]
232. Sensky T. The utility of systematic reviews: the case of psychological debriefing after trauma. Psychother Psychosom 2003;72(4):171-175. [doi: 10.1159/000070780] [Medline: 12817571]
233. Rose S, Bisson J, Wessely S. A systematic review of single-session psychological interventions ('debriefing') following trauma. Psychother Psychosom 2003;72(4):176-184. [doi: 10.1159/000070781] [Medline: 12792122]
234. van Emmerik AA, Kamphuis JH, Hulsbosch AM, Emmelkamp PM. Single session debriefing after psychological trauma: a meta-analysis. Lancet 2002 Sep 07;360(9335):766-771. [doi: 10.1016/S0140-6736(02)09897-5] [Medline: 12241834]
235. Richins MT, Gauntlett L, Tehrani N, Hesketh I, Weston D, Carter H, et al. Early post-trauma interventions in organizations: a scoping review. Front Psychol 2020 Jun 25;11:1176 [FREE Full text] [doi: 10.3389/fpsyg.2020.01176] [Medline: 32670143]
236. Levine B, Land HM. A meta-synthesis of qualitative findings about dance/movement therapy for individuals with trauma. Qual Health Res 2016 Feb;26(3):330-344. [doi: 10.1177/1049732315589920] [Medline: 26063604]
237. Grasser LR, Al-Saghir H, Wanna C, Spinei J, Javanbakht A. Moving through the trauma: dance/movement therapy as a somatic-based intervention for addressing trauma and stress among Syrian refugee children. J Am Acad Child Adolesc Psychiatry 2019 Nov;58(11):1124-1126. [doi: 10.1016/j.jaac.2019.07.007] [Medline: 31348987]
238. Bion WR. A theory of thinking. In: Spillius EB, editor. Melanie Klein Today: Developments in Theory and Practice, Volume 1. Milton Park, United Kingdom: Taylor & Francis Group; 1988:178-186.
239. Stevens JS, Jovanovic T. Role of social cognition in post-traumatic stress disorder: a review and meta-analysis. Genes Brain Behav 2019 Jan;18(1):e12518 [FREE Full text] [doi: 10.1111/gbb.12518] [Medline: 30221467]
240. Tibi-Elhanany Y, Shamay-Tsoory SG. Social cognition in social anxiety: first evidence for increased empathic abilities. Isr J Psychiatry Relat Sci 2011;48(2):98-106 [FREE Full text] [Medline: 22120444]
241. Elmi LM, Clapp JD. Interpersonal functioning and trauma: the role of empathy in moderating the association of PTSD and interpersonal functioning. Behav Ther 2021 Sep;52(5):1251-1264. [doi: 10.1016/j.beth.2021.02.004] [Medline: 34452677]
242. Maercker A, Horn AB. A socio-interpersonal perspective on PTSD: the case for environments and interpersonal processes. Clin Psychol Psychother 2013;20(6):465-481. [doi: 10.1002/cpp.1805] [Medline: 22730216]

XSL•FO
RenderX

243. Mitima-Verloop HB, Boelen PA, Mooren TT. Commemoration of disruptive events: a scoping review about posttraumatic stress reactions and related factors. Eur J Psychotraumatol 2020 Jan 13;11(1):1701226 [FREE Full text] [doi: 10.1080/20008198.2019.1701226] [Medline: 32082507]
244. Gene-Cos N, Fisher J, Ogden P, Cantrell A. Sensorimotor psychotherapy group therapy in the treatment of complex PTSD. Ann Psychiatr Ment Health 2016;4(6):1080 [FREE Full text]
245. Classen CC, Hughes L, Clark C, Hill Mohammed B, Woods P, Beckett B. A pilot RCT of a body-oriented group therapy for complex trauma survivors: an adaptation of sensorimotor psychotherapy. J Trauma Dissociation 2021;22(1):52-68. [doi: 10.1080/15299732.2020.1760173] [Medline: 32419670]
246. Sadeghi S, Schmidt SN, Mier D, Hass J. Effective connectivity of the human mirror neuron system during social cognition. Soc Cogn Affect Neurosci 2022 Aug 01;17(8):732-743 [FREE Full text] [doi: 10.1093/scan/nsab138] [Medline: 35086135]

Abbreviations

ACC: anterior cingulate cortex
CBT: cognitive behavioral therapy
C-PTSD: complex posttraumatic stress disorder
DMN: default mode network
DMT: dance movement therapy
D-PTSD: dissociative posttraumatic stress disorder
GI: guided imagery
GS: gesture sonification
MDMA: 3,4-Methylenedioxymethamphetamine
NMDA: N-methyl-D-aspartate
PFC: prefrontal cortex
PTSD: posttraumatic stress disorder
RHI: rubber hand illusion
SA: sense of agency
SN: salience network
SO: sense of body ownership
TE: traumatic event
VR: virtual reality
VRET: virtual reality exposure therapy

Edited by A Mavragani; submitted 14.03.24; peer-reviewed by S Gallagher, J Mistry; comments to author 03.04.24; revised version received 17.04.24; accepted 02.05.24; published 01.07.24.

Please cite as:
Adrien V, Bosc N, Peccia Galletto C, Diot T, Claverie D, Reggente N, Trousselard M, Bui E, Baubet T, Schoeller F
Enhancing Agency in Posttraumatic Stress Disorder Therapies Through Sensorimotor Technologies
J Med Internet Res 2024;26:e58390
URL: https://www.jmir.org/2024/1/e58390
doi:10.2196/58390
PMID:38742989

©Vladimir Adrien, Nicolas Bosc, Claire Peccia Galletto, Thomas Diot, Damien Claverie, Nicco Reggente, Marion Trousselard, Eric Bui, Thierry Baubet, Félix Schoeller. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions

Yu-Qing Cai[1*], BD; Da-Xin Gong[2*], MD, PhD; Li-Ying Tang[1*], BD; Yue Cai[1*], BD; Hui-Jun Li[3], BD; Tian-Ci Jing[2], MD; Mengchun Gong[4], MD; Wei Hu[5], MD; Zhen-Wei Zhang[6], MD; Xingang Zhang[7], PhD; Guang-Wei Zhang[2], MD, PhD

[1]The First Hospital of China Medical University, Shenyang, China
[2]Smart Hospital Management Department, The First Hospital of China Medical University, Shenyang, China
[3]Shenyang Medical & Film Science and Technology Co, Ltd, Shenyang, China
[4]Digital Health China Co, Ltd, Beijing, China
[5]Bayi Orthopedic Hospital, Chengdu, China
[6]China Rongtong Medical & Healthcare Co, Ltd, Chengdu, China
[7]Department of Cardiology, The First Hospital of China Medical University, Shenyang, China
[*]these authors contributed equally

Corresponding Author:
Guang-Wei Zhang, MD, PhD
Smart Hospital Management Department
The First Hospital of China Medical University
Shenyang,
China
Phone: 86 24 88283350
Fax: 86 24 83283350
Email: gwzhang@cmu.edu.cn

Abstract

In recent years, there has been explosive development in artificial intelligence (AI), which has been widely applied in the health care field. As a typical AI technology, machine learning models have emerged with great potential in predicting cardiovascular diseases by leveraging large amounts of medical data for training and optimization, which are expected to play a crucial role in reducing the incidence and mortality rates of cardiovascular diseases. Although the field has become a research hot spot, there are still many pitfalls that researchers need to pay close attention to. These pitfalls may affect the predictive performance, credibility, reliability, and reproducibility of the studied models, ultimately reducing the value of the research and affecting the prospects for clinical application. Therefore, identifying and avoiding these pitfalls is a crucial task before implementing the research. However, there is currently a lack of a comprehensive summary on this topic. This viewpoint aims to analyze the existing problems in terms of data quality, data set characteristics, model design, and statistical methods, as well as clinical implications, and provide possible solutions to these problems, such as gathering objective data, improving training, repeating measurements, increasing sample size, preventing overfitting using statistical methods, using specific AI algorithms to address targeted issues, standardizing outcomes and evaluation criteria, and enhancing fairness and replicability, with the goal of offering reference and assistance to researchers, algorithm developers, policy makers, and clinical practitioners.

(J Med Internet Res 2024;26:e47645) doi:10.2196/47645

KEYWORDS

cardiovascular diseases; risk prediction models; machine learning; problem; solution

Background

Cardiovascular diseases (CVDs) refer to both ischemic and hemorrhagic diseases that affect the heart, brain, and systemic vasculature, such as heart failure (HF), atrial fibrillation, acute coronary syndrome (ACS), myocardial infarction (MI), coronary heart disease (CHD), stroke, and cerebrovascular disease. As the most common noncommunicable diseases worldwide, CVDs remain a major cause of death in both low- and high-income countries. In 2019, there were 18.6 million deaths due to CVDs [1]. As awareness of the serious threat that CVDs pose to human

XSL•FO
RenderX

health has grown, many studies have focused on developing tools and guidelines for predicting these diseases. These prediction models can help identify individuals at high risk of developing CVDs, enabling preventive measures to be taken in a timely manner and potentially reducing both the human and economic costs of the disease. In this context, effective CVD risk prediction and prevention are critical to addressing this global challenge [2,3]. As a result, developing reliable and feasible prediction models for CVDs has become an ongoing area of exploration and study.

Early in 1976, the first CVD risk prediction equations were developed by the Framingham Heart Study, called the Framingham risk score (FRS). As one of the most classic CVD risk models, this score was widely applied all over the world in the following decades to provide important guidance for public health and clinical practice [4]. With the development of CVD preventive research, other important prediction tools have emerged for regional applicability, such as the Systematic Coronary Risk Evaluation model in Europe, the QRISK in the United Kingdom, the pooled cohort equation (PCE) for atherosclerotic CVDs (ASCVD) reported recently by the American College of Cardiology and American Heart Association guideline, and so on [5,6]. These models have a common feature: they all consist of fixed equations and lack scalability and are thereby defined as traditional models.

Although traditional models remain the most popular tools in the field of CVD prevention, they have proven to be inefficient and inflexible in the face of rapidly expanding amounts and types of data and increasing clinical requirements for precise, comprehensive, and continuous CVD risk prediction and treatment recommendations. The ideal risk prediction models are expected to include the broadest possible range of parameters and clinically relevant outcomes and to provide real-time and continuous support for doctors' decision-making, such as suggestions for smoking cessation, physical activity, diet, and medication use [7]. Obviously, the traditional models are no longer competent for these responsibilities.

Machine learning (ML) involves endowing computers with the ability to simulate or replicate human learning behavior, allowing them to acquire new knowledge or skills, reorganize existing knowledge structures, and continually improve their performance. As an important subset of artificial intelligence (AI), it has emerged as a promising research area in recent years [8-10]. ML models have also led to a significant evolution in the field of CVD risk prediction, allowing for the handling of new features of existing variables, such as nonlinearity and temporal dynamics, as well as novel variables such as electrocardiography results, medical images, and even genomics data [11-21]. Our recent systematic review [22], which included 486 AI-CVD prediction models across 79 articles, identified that AI has initiated a promising digital revolution in CVD risk prediction, characterized by an increase in the number and dimensions of predictors, as well as a notable diversity in applied algorithms, encompassing 66 specific algorithms across 13 categories.

Despite the stronger predictive ability and more promising development prospects of ML models compared with traditional ones, the development and clinical application of AI prediction technology are still strictly limited by a series of key problems. In the field of CVD prediction, this situation appears more severe than in other disease prediction or prognosis research [22-24]. As our systematic reviews have found [22], all current published AI-CVD prediction models exhibited a high risk of bias, lacked independent external validation, and had no clinical implementation. Moreover, the onset of CVDs is an exceptionally prolonged process, making prospective clinical validation of a flawed model potentially very costly in terms of research resources and possibly harmful to populations. Therefore, summarizing the potential pitfalls in AI-CVD prediction model research is crucial. Providing researchers with adequate warnings and references before initiating their studies is not only necessary but also of significant importance.

Although there are already some criteria in AI research that can be used as references, such as the Prediction Model Risk of Bias Assessment Tool (PROBAST) [25], Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD) [26], and AI transparent, replicable, ethical, and effective research (AI-TREE) [27], summarizing and analyzing key considerations and strategies during the research process will directly contribute to the practical implementation, replication, dissemination, and clinical application of model studies [28]. This holds significant practical significance not only for researchers but also for algorithm developers, cohort investigators, policy makers, health care providers, and professionals. Russo and Bonassi [29] previously outlined several "pitfalls" encountered in AI studies in nutritional epidemiology, including issues related to measurement methods, confounding factors, nonlinearities, missing data, overfitting, and interpretability, among others. Similarly, Chiarito et al [30] identified a few "pitfalls" in their review of AI and CVD risk prediction [30]. However, there is currently a lack of a comprehensive and systematic summary specific to the AI-CVD prediction field. In this paper [25,27,31], we summarize and analyze these existing problems and their possible solutions, aiming to provide guidance and references during the process of the development of AI models for the prediction of CVDs and other diseases.

Design and Summary of a Framework for Presenting Existing Pitfalls

To develop a framework outlining the key pitfalls, the following process was executed (Figure 1). A literature review of existing assessment guidelines or tools related to AI or ML medical research was performed, with the search strategy of "assessment tool/guideline"+"AI/ML"+"bias" and their synonyms and related terms (details are shown in Multimedia Appendix 1). Full-text articles in English related to human participants were included, while low-quality or irrelevant articles, such as those discussing the current state of ML or AI or guidelines outside the scope of this study, were excluded. Detailed information is provided in Figure 1. The inclusion and exclusion process was conducted by YQC, YC, and GWZ. Ultimately, 31 papers were included in the preliminary summary of candidate items (Multimedia Appendix 2 [25,27,31-60]). Subsequently, incorporating the

XSL•FO
RenderX

risk issues identified in our previous systematic review of AI-CVD prediction models [22], a final framework was further discussed and confirmed by a panel of experts. This panel included AI experts (TCJ and MG), statisticians (ZWZ), clinicians (DXG, XZ, and GWZ), and information technology specialists (WH), among others. As summarized in Multimedia Appendix 3 [15,18,21,28,61-100], the framework encompasses four major categories with 15 subcategories: (1) data quality (data source, subjective factors of researchers, incomplete data, and parameter acquisition methods); (2) data set characteristics (small sample size, low event rate, characteristics of the data distribution, and multimodal data); (3) model design and statistical methods (outcome definitions, incomplete inclusion of covariates, overfitting, and defects in evaluation criteria); and (4) clinical implications (generalization, interpretability, and AI ethics).

Figure 1. The flowchart for developing a framework for this paper based on the assessment guidelines or tools in the field of medical artificial or machine learning research. CVD: cardiovascular disease; ML: machine learning.

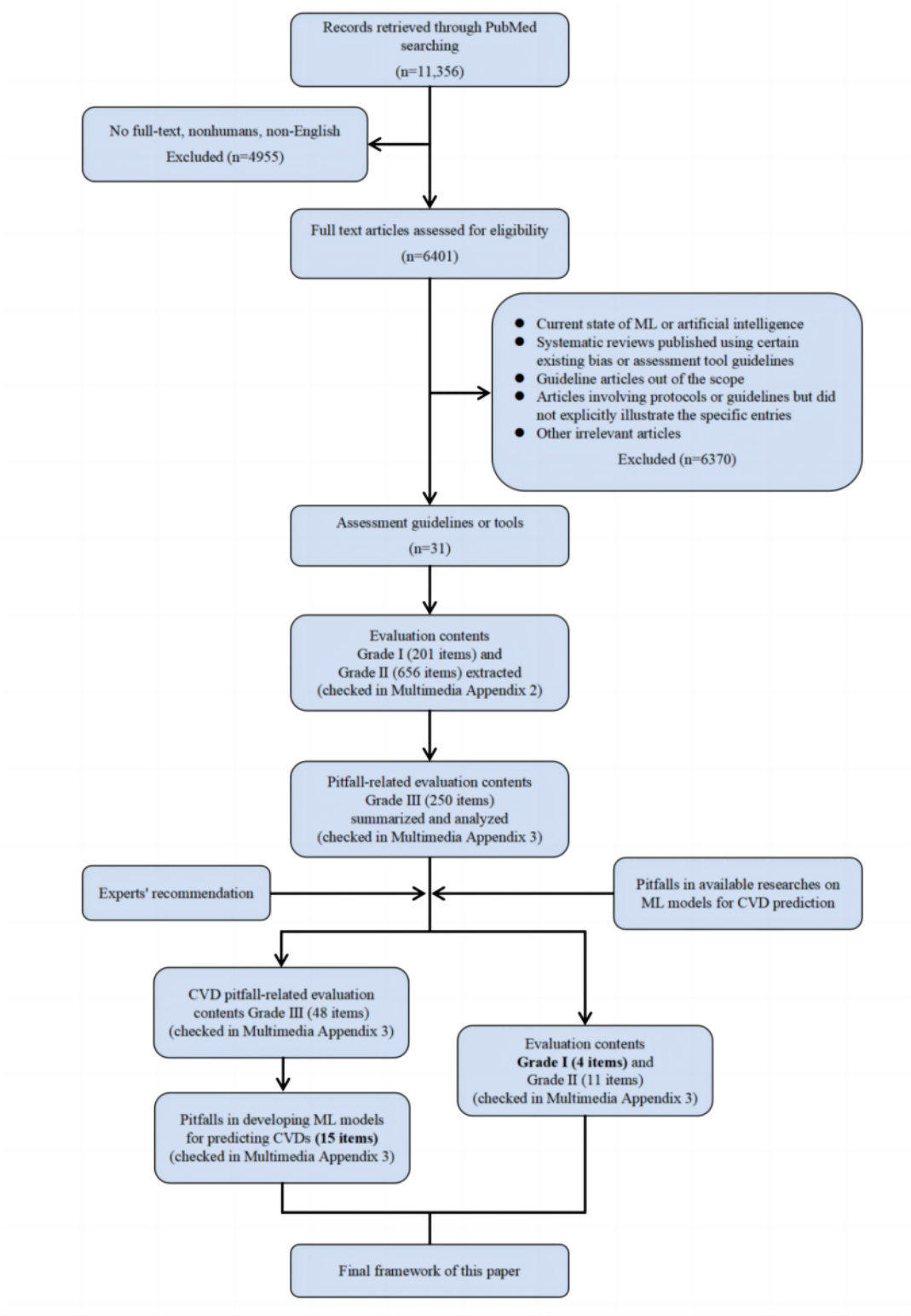

XSL•FO
RenderX

The Effectiveness of the ML Models Still Largely Depends on Data Quality

Overview

Data are undeniably a fundamental element in ML models, which are inherently sensitive to the quality of the data used for training and validation. The adage "garbage in, garbage out" is particularly pertinent here [101]; poor-quality input data can also result in a biased ML model, and even minor errors or biases in training data can lead to unforeseen consequences in a model's predictions. These considerations raise ethical concerns regarding the reliability of decisions based on ML model predictions. Consequently, understanding and mitigating pitfalls in data collection, preprocessing, analysis, and application become critically important. Specific methods involve selecting objective indicators collected through reasonable means based on a thorough evaluation of data authenticity.

Data Recourse Is the Key Factor

It is well-known that data accuracy is a key influencing factor for data-based prediction models, as inaccurate data can significantly impact the predictive ability of subsequent models. One of the main reasons for data inaccuracy is the collection of subjective data, such as self-reported data on blood pressure (BP), obesity, and family history, as noted by Hippisley-Cox et al [102] and Manuel et al [103] in traditional models, based on the Canadian Community Health Survey (104,219 participants) and QResearch database (3,610,918 participants), respectively. In fact, AI models have encountered the same dilemma [104,105]. As reported by Han et al [61] in their ML research based on the Korea Initiatives on Coronary Calcification (86,155 patients), inaccurate data from self-administered questionnaires had a limited contribution to model development; after feature selection based on information gain ranking (such as the history of diabetes and hypertension, this limitation resulted in an exclusion of important predictive factors from 10 most pertinent variables of a future all-cause mortality (ACM) events. Ultimately, this underestimation of important risk factors may lead to insufficient attention being paid to them, thereby misdirecting primary prevention strategies such as aggressive glycemic control.

Similar to traditional models, the most effective solution in AI models is still to collect objective data, instead of relying solely on self-reported data whenever possible. Even electronic health records (EHR) can contain incorrect data, which may be difficult to detect and therefore easily overlooked [62]. For example, in a study conducted by Rodriguez et al [63], race data, which were obtained from a community-based EHR with 231,622 participants, were mainly self-reported, while other data were inferred through validation methods that were based on PCE, leading to possible misclassification, and may affect the risk prediction for ASCVD. By way of illustration, it is advocated by Teoh [64] that the diagnosis or the value of the check index should be applied to further confirm the disease status of the participants, including hypertension, diabetes, and dyslipidemia reported by the participants themselves, and they found that the area under the curve (AUC) dropped from 0.623 to 0.608 when the diabetes examination inputs were removed (including blood glucose, glycated hemoglobin [as standardized by Japan Diabetes Society], and glycated hemoglobin [as standardized by National Glycohemoglobin Standardization Program]). These examples clearly demonstrate the impact of data authenticity on the robustness of the models. Unfortunately, even some classic scoring tools, such as PROBAST, TRIPOD, and CREMLS (Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models), do not specify methods for collecting objective data [32,33,106,107]. Therefore, we suggest the implementation of appropriate validation measures during model development.

The Subjective Factor of Researchers Is a Potential Interference

In some studies, basic background, sociodemographic, and follow-up information may be collected by local doctors and nurses [65,108,109]. Although these interviewers have been strictly trained, the data collection process is likely to be interfered with by their personal thoughts. Benjamins et al [66] reported that in the Glycometabolic Intervention as Adjunct to Primary Percutaneous Intervention in ST-Segment Elevation Myocardial Infarction-III study, which included 222 people and defined ventricular function dysfunction as an outcome, using data created by local observers to train and validate the U-Net model may result in an unstable model quality. Liu et al [67] analyzed the registry data of 53,213 inpatients in the Cardiovascular Department of Xiangya Hospital; they found that student extractors, who manually filled missing data using a hierarchical mean filling method, subjectively adjusted variable parameters of the extracted data based on experience and introduced a potential risk of bias. Although these reports mention the potential impact of researchers' subjective factors on the models, they lack effective proof methods, highlighting that such factors are often overlooked and difficult to evaluate. Therefore, it is essential to identify subjective factors of researchers before the study.

The most common and effective approaches to address this issue are to improve training and supervision of data collectors. While both traditional and ML models require manual data processing, some studies have shown that AI can potentially replace humans in data processing to reduce the impact of subjective factors. For instance, in the Systolic Blood Pressure Intervention Trial (SPRINT) and the Hong Kong eHealth cohort studies with 8133 and 1094 participants, respectively, Tsoi et al [68] used a K-means clustering ML algorithm to identify variations in BP without the need for human adjustment, which resulted in low clustering similarities (Davies-Bouldin Index: 0.653 in SPRINT and 0.680 in the Hong Kong eHealth cohort). This not only further highlights the importance of controlling subjective factors in model construction but also provides an alternative solution beyond simply enhancing training.

Incomplete Data Also Affect the Effectiveness of ML Models

Incomplete data are a common problem in retrospective studies and may significantly impact the final predictive results of ML studies. For instance, in the Melbourne Collaborative Cohort Study with 32,611 participants, Sajeev et al [69] reported that

XSL•FO
RenderX

32% of high-density lipoprotein cholesterol data were missing and that this resulted in a lower AUC of 0.753 (95% CI 0.729-0.777) in a logistic regression (LR) model, while an AUC of 0.874 (95% CI 0.833-0.915) was achieved in another similar LR model that was developed by the North West Adelaide Health Study, which had only 41 missing data among 3654 participants. It is also argued by Han et al [61] that unmeasured factors could introduce bias in their ACM prediction ML model, which used the Korean Initiatives on Coronary Calcification registry, a retrospective cohort. The severity of this issue lies in its potential to also cause underestimation or neglect of the impact of risk factors in actual clinical practice, thereby affecting CVD prevention strategies.

Both traditional and ML models face the issue of missing data. In traditional models, missing data are often directly excluded, which can impact model stability and weaken its significance [70,108,110]. In contrast, ML models use several statistical methods to address the problem of missing data and optimize the models, such as multiple imputation, multiple imputation with chained equations, logitBoost, multiple imputation by fully conditional specification, Markov chain Monte Carlo method, median imputation, and k-nearest neighbor (KNN) imputation [71,72,111,112]. The application of these statistical methods gives ML models an advantage over traditional models in handling missing data. For instance, in a study that included 423,604 participants from the UK Biobank, Alaa et al [73] used the missForest algorithm to impute missing data in an ML model, achieving a significantly higher AUC of 0.774 (95% CI 0.768-0.780) than with FRS (AUC=0.724, 95% CI 0.720-0.728). In the ATTICA prospective study involving 2020 participants, Dimopoulos et al [112] excluded variables with missing values in >70% of the individuals. In order to improve the performance of the CVD risk estimator, the missing data in part of the remaining 22 variables were then imputed by KNN imputation, resulting in relatively high accuracy rates (KNN: 96%, random forest [RF]: 99%, decision tree classifier: 99%). Therefore, ML models can better manage missing data due to the application of the aforementioned statistical methods. Furthermore, ML models also can register patients with missing necessary variables in traditional models [62], providing an advantage in handling incomplete data. In addition, besides the aforementioned statistical methods, Weng et al [71] used a novel approach in the Clinical Practice Research Datalink (CPRD) study that included 378,256 patients, using missing data as an independent variable to predict the risk of CVDs. Specifically, they created dummy variables to indicate missing continuous variables and a separate category of “unknown” to represent missing categorical variables such as the Townsend deprivation index and race, with an acceptable AUC of 0.764 (95% CI 0.759-0.769) and a specificity of 70.7% in the neural networks model eventually. Although this method needs to be further confirmed, exploring such new approaches provides additional options for ML models in handling missing data, thereby enhancing AI-CVD prediction models.

Appropriate Parameter Acquisition Method Is Another Important Factor

In most studies predicting the risk of CVDs, researchers have typically used one-time measurements of risk factors, ignoring the fact that some factors exhibit significant variability over time, such as systolic BP (SBP), plasma glucose levels, low-density lipoprotein, and serum total cholesterol [113-115]. As demonstrated in the SPRINT study mentioned in The Subjective Factor of Researchers Is a Potential Interference section, BP measured at fixed times (visit-to-visit BP variability) can vary greatly from 24-hour ambulatory BP variability (r^2<0.026), for the data that are acquired intermittently with a short follow-up time may introduce bias in the data [68]. These common problems illustrate that variables measured at a certain time point cannot necessarily represent the true level of the variable in the population. In fact, there have already been diagnostic criteria for multiple measurements of some risk factors in guidelines, for single measurements are likely to introduce significant bias within clinical work. For example, Stergiou et al [116] advocated that a minimum of 2 to 3 office visits at 1- to 4-week intervals are frequently necessary for the office BP assessment.

The solving method is to use repeated measurements of risk factors to improve the model prediction [117]. In the National Health Insurance System-National Health Screening Cohort (NHIS-HEALS) study with 361,239 participants selected, Sung et al [74] advocated that by offering discrimination and calibration with repeatedly measured data, the deep learning model takes advantages in CVD risk prediction in the EHR era (female participants: AUC=0.94, 95% CI 0.91-0.97 in; male participants: AUC=0.96, 95% CI 0.95-0.97 in). In another study including 80,964 people, multiple SBP recordings from EHR (provided by UK primary care) were analyzed. It was found that the multiple SBP recordings had a better correlation with CVDs than a single recording, as identified by an increased hazard ratio from 1.22 (95% CI 1.18-1.30) to 1.39 (95% CI 1.31-1.46) with the use of repeated measured factors [75]. Besides, in the NHIS-HEALS study, Cho et al [76] handled continuous variables by using their mean, minimum and maximum values, and SDs in the developing of recurrent neural network models and achieved a high level of discriminative accuracy (female participants: AUC=0.921, 95% CI 0.908-0.934; male participants: AUC=0.896, 95% CI 0.886-0.907), which was highlighted by authors as significant improvement compared with single-measured method [76]. Furthermore, collecting data multiple times can enhance data consistency and to the benefit of exploring the changing trend of data.

Data Set Characteristics Largely Affect the Effectiveness of the ML Models

Overview

The impact of the data set characteristics used for ML on risk prediction is crucial and should not be disregarded. This includes several elements: the impact of the small sample size and low event rates, the alignment of data distribution characteristics with the model, and the challenges associated with the application of multimodal data. These pitfalls can significantly impede ML studies, highlighting the critical need for heightened attention in the various stages such as research design, implementation, data analysis, model construction, paper

XSL•FO
RenderX

preparing, and even in submission and peer review processes. Specific methods include using sufficiently large cohorts, extending follow-up periods, using specific AI algorithms to address targeted issues, and selecting clinical routine examinations as risk factors whenever possible.

Small Sample Size Is a More Serious Problem

It is well known that insufficient sample size will lead to the risk of bias for the study in the traditional model, and the evaluation standard of sample size has been developed by the PROBAST [106]. Therefore, in order to improve the credibility of research conclusions, more and more researchers have used large data sets that are multiethnic cohorts with high quality [77,118-120], although there are still many substandard studies [108,121]. However, in the field of ML prediction models, the problem of insufficient sample size seems to be more serious. In our previous research of 79 CVD risk prediction studies, the number of participants included in the ML model could even be as low as 80 [22]. The lack of a sufficient sample has become the biggest contributor to the high risk of bias [28]. For example, in a retrospective observational study of 420 patients, Ponomartseva et al [78] stated that when the sample size is smaller than the optimal ML method size, the accuracy of the model may change. In another study carried out with 451 consecutive patients from a tertiary hospital, conventional univariate and multivariate analyses were limited because of the small sample size; the median accuracy of artificial neural network (ANN) models in predicting recurrent stroke was only 75%, as reported by Chan et al [79].

Although a larger population size was prone to the better performance and reliability of the models, as revealed by Alaa et al [73] in an experiment with a series of subpopulations of varying sizes and a fixed number of variables, the sample size of the ML algorithms is dependent on various factors, including the types of algorithms, the number of index, the features of the sample, and so on [122-124]. For example, the classical events per variable method, which often follows the “one in ten rule” for the sample size calculation, may not be applicable to certain ML algorithms due to their specific operation mechanism [22]. Therefore, it is strongly recommended that the sample size calculation be performed before initiating ML research, considering the various factors mentioned above, due to dramatic differences in the mechanisms of sample size calculation between ML and traditional models. Meanwhile, the upcoming PROBAST-AI, a special tool for bias evaluation for AI models, has been expected to provide a reference standard [125-127]. Before this, we are inclined to recommend using the “one in ten rule” as the minimum standard to include as many samples as possible.

Application of the ML Models Is Also Troubled by Low Event Rate

The low incidence rate, a common problem for prediction model construction, may restrict the generalizability and predictive ability of the prediction model in practice [74,80,128]. It may lead to the bad performance of the model in general or cause bias in predictions for a particular segment that has a low incidence rate. For example, on the basis of the data from 13,291 participants from 7 epidemiological cohorts, Jdanov et al [77] found that the low incidence rate of CVD deaths (<7% in male participants and <1% in female participants) at a young age (aged 45-54 years) affected the prediction outcomes at young ages, with a substantial underestimation of CVD mortality risk by 40% to 45% for male individuals and 3% to 4% for female individuals. Therefore, based on this reasoning, if applied in clinical practice, the low event rate may result in the underestimation of a substantial number of individuals who are at high risk for CVD, thus reducing the role of primary prevention.

The problem of low event rate can be overcome by the following methods, such as filtering of oversampled data using noncooperative game theory [80], synthetic minority oversampling technique [129], random oversampling [130], random undersampling [19], bootstrapping [81], and the stacking paradigm [131]. However, these methods can only improve the performance of the ML from the statistical method and cannot solve the problems with the data themselves. By using the data solely from the DaVita Inc EHRs (124,097 participants), Goldstein et al [82] reported a direct way to increase the incidence rate by broadening the outcome range and showed good training and test set loss at the same time. However, by using this method, the outcome will become more heterogeneous and will decrease the quality of the prediction model, and thus this method may not be an adequate option. The better approach may be to extend the follow-up duration or to include more different populations as mentioned above to improve the reliability of ML [61].

Characteristics of the Data Distribution Underscore the Critical Importance of Model Selection

The cornerstone of constructing predictive models lies in the analysis, mining, and use of data relationships between covariates (risk factors) and gold standards (outcome events). Traditional statistical methods often rely on the assumption of linearity when building predictive models. However, it is essential to acknowledge that in clinical practice, many risk factors and outcome events exhibit nonlinear relationships, although linear relationships may predominantly characterize some factors, such as glycated hemoglobin or smoking pack-years. For instance, Angeli et al [83] and Lip et al [84] reported a J-curve relationship between BP and CVDs or ACM. Similarly, dietary factors such as salt, carbohydrates, and fats have demonstrated U- or J-shaped relationships with CVD outcomes [132,133]. Unfortunately, certain ML algorithms are not well suited to handle nonlinear relationships, including linear regression, LR, and so on. Consequently, if linear algorithms are used to construct predictive models that incorporate nonlinear elements, the result may lead to inaccurate associations and distorted effect estimates. In a study based on the National Health and Nutrition Examination Survey cohort (37,079 participants), Dutta et al [85] developed 6 AI models with nonlinear factors (eg, BP) for predicting CHD, and expectedly, the LR algorithm achieved the worst performance (AUC=0.713). Similarly, Wang et al [70] included 40,711 participants from the Life Risk Pooling Project cohort to build 4 AI models with nonlinear factors (eg, BP) for predicting CVDs and also found the worst performance in linear naive Bayes algorithm (AUC=0.786, 95% CI 0.726-0.735; *P*=.001) [70]. In addition,

XSL•FO
RenderX

many predictive models aimed to assess long-term risk for adverse events, but commonly used ML models often struggle with handling time-to-event variations and censored patient data, significantly limiting their performance. Li et al [72] demonstrated that LR algorithms that neglect censoring tend to substantially underestimate CVD risk to only 2.2% to 5.8% when compared with Cox models that had a risk of 9.5% to 10.5%, which are better suited for survival data analysis.

Therefore, it is imperative to emphasize that the application of nonlinear AI algorithms plays a pivotal role in the research of constructing predictive models that encompass intricate data relationships. These algorithms have the capability to capture and model intricate patterns, interactions, and dependencies within the data that may not be adequately addressed by linear approaches. As in the 2 examples elaborated in the previous paragraph, support vector machine (SVM) had the best performance with an AUC of 0.776 in the study by Dutta et al [85], and RF had the best performance with an AUC of 0.892 in the study by Wang et al [70]. When it comes to survival data, random survival forest (RSF) appears to exhibit superior performance, owing to their ability to handle complex, time-dependent relationships and censored data more effectively, ultimately resulting in more accurate survival predictions. For instance, in a study based on the Atherosclerosis Risk in Communities cohort (14,842 participants), Zhuang et al [86] developed 4 survival models (Cox proportional hazards model, Akaike information criteria for Cox regression, least absolute shrinkage and selection operator for Cox regression, and RSF), and the RSF algorithm achieved the best performance in predicting CHD (AUC=0.80, 95% CI 0.79-0.81) and ACM (AUC=0.78, 95% CI 0.77-0.78). Another example is meant by Ambale-Venkatesh et al [87], where 6814 participants from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort were used for ACM risk. Eventually, the RSF algorithm (AUC=0.84) performed the best, yielding better results compared to those obtained with other algorithms (eg, Cox least absolute shrinkage and selection operator: AUC=0.80).

Multimodal Data–Based Prediction Model Pitfalls Limit Clinical Applications

The number of studies of multimodal data–based prediction models for CVDs has risen substantially recently, and there have been a variety of data types, such as radiomics [88,134], proteomics [135], and genomics [136]. The multimodal data–based prediction model, which integrates available heterogeneous data into a unified framework, can fully consider the importance of each modality and incorporate information from multiple aspects, thus improving model performance [136]. However, the clinical accessibility of the used multimodal data presents another pitfall. As advocated by Pujadas et al [88], there was a limitation in the multimodal data–based CVD prediction model with cardiovascular magnetic resonance radiomics, because cardiovascular magnetic resonance was not a routine examination. More significantly, according to the AI-TREE criteria, the inclusion of unconventional or difficult-to-obtain examination results is a crucial factor influencing the clinical application of the model [27]. This may be solved by applying more commonly used tests (eg, electrocardiogram [ECG]) or disease-specific imaging examinations. For instance, it is reported by Lou et al [89] that ECG-based patient characteristics from 2 hospitals in the Tri-Service General System were used for CVD risk prediction, and the deep learning models performed well, yielding excellent results (AUC>0.90). Another example of the solution was carried out by Chao et al [21], who included clinical routine 30,286 low-dose computed tomography data from 2085 patients with lung cancer in the National Lung Cancer Screening Trial to construct a deep learning CVD risk prediction model. Therefore, we suggest that multimodal model studies should focus on the practicality of data acquisition to enhance the feasibility of primary prevention of CVDs.

In addition, studies emphasizing multimodal data often disregard the inclusion of classical risk factors, neglecting essential parameters crucial for comprehensive prediction models. The overarching focus on multimodal data variables can lead to decreased model performance due to the absence of integrated incorporation of traditional risk factors. For instance, in the ECG methods, for the prompt identification of coronary events study (499 patients), Al-Zaiti et al [15] constructed ACS risk prediction models by using 554 temporal-spatial features of the 12-lead ECG without clinical examination factors such as total cholesterol or lifestyle variables such as smoking, which led to unsatisfactory validation performance (LR model: AUC=0.67; gradient boosting machine [GBM] model: ACS=0.71). Therefore, it is strongly recommended that the inclusion of classic predictive factors should form the cornerstone for constructing multimodal models. This is evident in the Singapore epidemiology of eye diseases study (comprised of over 70,000 images), where Cheung et al [134] not only applied imaging data but also comprehensively incorporated multiple types of variables in CVD risk estimators, and the best model achieved a perfect AUC of 0.948 (95% CI 0.941-0.954).

Challenges Arising From Model Design and Statistical Methods

Overview

In AI-CVD prediction research, model design and statistical methods play a critical role not only in aspects of model performance such as discrimination and accuracy but also in reproducibility and generalizability for clinical application. Key pitfalls in this area include issues of outcome definitions, incomplete inclusion of covariates, overfitting, and inadequate evaluation metrics. Some of these pitfalls stem from inherent problems associated with AI algorithms, necessitating researchers to thoroughly understand and mitigate their impact. In addition, other pitfalls can arise from design and implementation choices made by researchers, underscoring the need for careful planning and execution to avoid such issues during the research process. Specific methods include designing studies with standardized outcomes and evaluation criteria, incorporating all necessary variables, and using appropriate algorithms and validation techniques.

XSL•FO
RenderX

The Outcome Definitions May Be Detrimental to the Significance of the Study

The definitions of the CVD outcomes show considerable heterogeneity in both traditional and ML models, concluding from the facts that there have always been some differences more or less in the detailed definitions among almost all studies. All the 3 versions of QRISK set the end points as CVDs, which are defined as a composite outcome of CHD, stroke, and transient ischemic attacks [5,111,137]. Systematic Coronary Risk Evaluation in Europe aimed to predict fatal CVDs, so its end point was CVD mortality [138]. PCE and Prediction for ASCVD Risk in China defined ASCVD as nonfatal MI, CHD death, or fatal or nonfatal stroke [139,140]. The outcomes of adding social deprivation and family history to CVD risk assessments were CVD-cause mortality, CHD, cerebrovascular disease, or interventions for coronary artery (coronary artery bypass grafting or percutaneous transluminal coronary angioplasty) [141]. These differences, to some extent, limit the comparability, reliability, and generalizability of the models, thereby affecting their clinical applicability for CVD prediction.

In studies of ML prediction models, heterogeneity is even more significant, with various origins of outcome definitions, such as disease codes (*International Classification of Diseases [ICD], Ninth Revision* [*ICD-9*] or *ICD-10*), self-reports, and other international guidelines. For example, in a study with 31,466 participants from the UK Biobank, You et al [90] defined CVD events as MI or stroke with *ICD-9* and *ICD-10* codes. Conversely, Cho et al [76] in the NHIS-HEALS study, developed risk prediction models that used death from CVDs, MI, coronary arterial intervention, or bypass surgery stroke as the outcome definitions with only *ICD-10* codes. This heterogeneity in outcome definitions can lead to significant issues such as bias in model performance evaluation, reduced generalizability, challenges in data integration and standardization, difficulties in clinical application, and even misguidance in CVD prevention strategies. Therefore, it is crucial to address this problem adequately. To mitigate these issues, we strongly recommend that risk prediction models be developed with standardized use of *ICD* codes. This standardization would help in achieving consistent outcome definitions, thereby improving the reliability, generalizability, and clinical applicability of the prediction models.

Incomplete Inclusion of Covariates May Diminish the Study's Value

In many CVD risk prediction studies, several classical risk factors have been omitted. This phenomenon is not limited to retrospective studies using existing databases but is also prevalent in prospective research. By way of illustration, Chua et al [81] included 638 participants from Sandwell and West Birmingham Hospitals NHS Trust (local data set) to build 2 atrial fibrillation risk calculators. They did not include classical factors such as BP and smoking status and achieved a low accuracy performance with an AUC of 0.684 (95% CI 0.62-0.75) and 0.697 (95% CI 0.63-0.76), respectively. In the China Health and Retirement Longitudinal Study, Chen et al [91] included 9821 participants in the development of a stroke risk calculator. However, this classical public data set lacked classic variables such as total cholesterol, which resulted in a mediocre performance (AUC=0.7388) in the model. Such omissions may potentially lead to inadequate consideration of confounding factors, resulting in model bias and decreased performance or even rendering the model practically inapplicable [142]. From a statistical perspective, these classic risk factors can be considered confounders, and ignoring them may cause confounding, which represents a critical error and is a primary source of systematic errors that can manifest when assessing causality, potentially leading to misinterpretation of the results [143]. In accordance with the AI-TREE criterion [27], a key consideration is whether the available data can effectively address the clinical question at hand. In other words, a data set lacking essential predictors that are known to be relevant to an outcome is unlikely to satisfactorily address related inquiries. Therefore, studies that exclusively focus on model development while disregarding considerations for clinical application may deviate from sound research principles.

One potential approach to mitigate these challenges is to ensure that potentially biased features, such as ethnicity and social determinants of health, are explicitly incorporated into the models [142,144]. As suggested by Suri et al [28] in a study on understanding bias in ML systems for CVD risk assessment, achieving a robust ML-based design for CHD and CVD prediction necessitates the integration of traditional, laboratory, image-based, and medication-based biomarkers. In the realm of clinical research, where there is often a high number of variables and confounding factors, the adoption of high-capacity AI models is warranted [145,146]. For instance, Kakadiaris et al [80] obtained 6459 participants from the MESA cohort to build a CVD risk calculator based on SVM. They included all classic risk factors such as age, sex, smoking status, diabetes, SBP, total cholesterol, and high-density lipoprotein and achieved a rather high AUC of 0.94 (95% CI 0.93-0.95). In addition, in a study conducted with an outpatient health care system, Ward et al [62] enrolled 262,923 participants with risk factors that contained all classic risk factors as mentioned above for ASCVD risk prediction, and the AI model performed well with an AUC of 0.835 (95% CI 0.825-0.846). These also indicated the importance of incorporating complete covariates for accurate modeling, which is more conducive to individualized, precise, systematic, and comprehensive prediction and assessment of CVDs. This, in turn, facilitates targeted prevention and monitoring.

Overfitting Determines the Performance of ML Models

Overfitting, a more severe problem in ML, means that the data can fit relatively perfectly with the derivation queue during training and derivation, but it will yield large biases and unreliable results when applied to other data, consequently leading to difficult generalization of the model [147,148]. The influence factors accounting for overfitting of ML models are summarized as follows: the features of algorithms, the number of variables, cohorts, outcome events, and so on. It is found in the research performed by Weng et al [71] that the more complex the model, the more likely it was prone to overfitting, while Commandeur et al [92] revealed that the few numbers of the derivation or the low event rate will cause the overfitting problem, which can be assessed by comparing the accuracy of

XSL•FO
RenderX

the training set and the test set. These reports suggest that selecting appropriate algorithms and addressing the issue of low event rates have a decisive impact on preventing model overfitting, necessitating thorough consideration by researchers during study design.

There are numerous approaches to reduce the impact of overfitting, and the most commonly used include pretraining, hyperparameter selection, regularization, and cross-validation [149,150]. There are also other approaches, such as data-smoothing techniques, separating training, tuning, testing data, and the filtering of oversampled data using noncooperative game theory algorithm [80,92]. In addition, in the Prospective Cardiovascular Münster Study, which included 5159 participants, Voss et al [93] have proposed 3 solutions to address overfitting for CHD event estimations: cross-validation, stopping training when the errors in validation data sets are at a minimum, and rendering the networks with synthetic data and modifying them until the results are plausible. Eventually, the average negative log likelihood (ANLL) that explains the loss function of the model and the bias value attained comparatively minor numbers in the probabilistic neural networks model for both training (ANLL: 0.1712, bias: 0.0396) and testing sets (ANLL: 0.1807, bias: 0.0396).

Inadequate Evaluation Metrics Still Exist in the Field of ML Model Prediction

The evaluation metrics help researchers assess the performance of models, and thus, defects in them can lead to obstacles in drawing practical conclusions. For example, although the AUC is the most commonly used evaluation criterion for model assessment, it may not accurately reflect the risk of CVDs in the population. This is because a person with a very low risk makes the same contribution to the AUC's value as a person with a very high risk, as noted in a study developed by the CPRD with 3,660,000 participants [72]. Similarly, the Net Reclassification Index has been shown to be insensitive to changes in the model. For example, in a MESA study (5878 participants) reported by Polonsky et al [94], the Net Reclassification Index remained substantially constant even after including or excluding certain participants. These examples indicate that model evaluation cannot be adequately represented by a single metric; rather, it should be a comprehensive, multifaceted, and complex system, especially for AI models, in line with preventive medicine principles.

It is advisable to use a comprehensive evaluation of predictive model capabilities by considering various metrics such as accuracy, sensitivity, precision, specificity, F_1-score, decision curves, and so on [151], which are also required to be reported in ML prediction studies as advocated by several reporting guidelines [32-34]. This is not only crucial for assessing the reliability of the model but also beneficial for its accurate application. Some commercial medical devices, such as ultrasound-assisted ECGs, have already contained criteria with high transparency and high clarity that are not available for predictive models. Thus, marketizing prediction models may be a solution. Moreover, in contrast to the predictive model, the intervention AI evaluation has different criteria, which are more rigorous and systematic. For instance, as the guideline of clinical trial reports for interventions involving AI, CONSORT (Consolidated Standards of Reporting Trials)-AI recommended presenting absolute and relative effect sizes for binary outcomes [35]. Therefore, we recommend that during model development, comprehensive calculation of these evaluation metrics is essential, as it is crucial for subsequent model dissemination and selection for clinical applications.

Clinical Implication Remains a Significant Challenge in This Field

Overview

In fact, the factors affecting the clinical application of AI-CVD models are complex. Beyond the pitfalls discussed above, 3 directly related and most prevalent factors that researchers should particularly consider are the problem of generalization, the lack of interpretability, and the limitations of AI ethics. These factors not only require researchers to minimize their impact during the study process but also necessitate collaborative efforts from clinicians, policy makers, and even the general public to overcome and address them. This includes clinicians integrating AI tools into their practice, policy makers establishing supportive regulatory frameworks, and the general public engaging in discussions about the ethical implications.

The Problem of Generalization Has Largely Limited ML Algorithms Application

Generalization, which represents the ability to adapt to novel situations, is rather important for the applicability of CVD risk prediction models. However, both traditional and ML models have the problem of generalization, as they are developed based on specific populations. For example, many studies have used cohorts predominantly consisting of White individuals, men, or participants from a single center, even during a relatively narrow socioeconomic range with well-educated and low-risk factor burden compared with the general population [152-155]. Thus, risk prediction models were largely restricted to application of cross-cohorts. The Pooled Cohort Equations to Prevent Heart Failure tool, which was developed with the data of 11,771 individuals from 7 community-based cohorts, was unable to offer precise HF risk estimates for individuals who were not classified to the non-Hispanic White and non-Hispanic Black racial or ethnic groups [7]. The Cox regression models developed by Li et al [156] contained data from 44,869 participants from the Evidence for Cardiovascular Prevention from Observational Cohorts in Japan research group. Models were developed to predict death from CHD, stroke, and CVDs, but they may lack accuracy when predicting risk in women, for the corresponding *P* values of the Hosmer-Lemeshow test were .27, .002, and .04 in women and .51, .49, and .25 in men, respectively.

In ML, it is still a challenge to generalize models with specific populations, although this problem has been improved to a certain extent due to the elimination of some restrictions by AI algorithms. It has been documented by Bouzid et al [18] that using patients from only 1 region (even multiple hospitals) would significantly influence the promotion of ML models. In terms of improvement, it has been proven by Ward et al [62]

XSL•FO
RenderX

that ML models developed in the United States or Europe could be expanded to Asian and Hispanic populations, and they achieved AUCs of 0.803 (95% CI 0.743-0.863) and 0.768 (95% CI 0.663-0.874), respectively, without the limitation of race in the PCE. In addition, the application of PCE is limited to the population aged from 40 to 79 years, while ML models have eliminated this limitation [62]. Moreover, the LR and SVM algorithms have been confirmed to overcome the age limitation (<79 years), which caused 26% of all cases in FRS not to be classified [95].

The way to solve the problem of generalization is to include more participants in a large data set to improve the models' applicability as advocated by Chang et al [157] and Chen et al [158]. For example, it has been demonstrated that the model trained on the largest training set yielded the best performance in external validation [11,159]. Zarkogianni et al [96] have introduced larger data sets corresponding to patients with different ethnicities and races to other cohorts of patients and consequently extended the applicability of the model. Therefore, it is advocated to conduct multicohort, multicountry, or multiracial studies with adequate cross-validation. In the real world, however, it is difficult to obtain large data with high quality, which will cost lots of manpower and money and take a long time. In response to this point, it may be a more appropriate solution to fully use open-source databases, such as the National Health Insurance Service-National Sample Cohort, NHIS-HEALS, Qresearch, Early Identification of Subclinical Atherosclerosis by Noninvasive Imaging Research, and so on [76,92].

The Lack of Interpretability in ML Algorithms Poses a Significant Challenge

Prediction models play a crucial role in helping doctors make accurate decisions in clinical practice. However, the "black box" effect in many ML algorithms, such as ANNs and RF, creates a challenge for doctors to interpret mechanisms correctly [160]. This limits the clinical utility of the models, as they become complex for doctors to understand the impact of risk factors on the prevalence of CVDs. In contrast to the fixed regression coefficient in traditional models, the complexity of ML algorithms hinders their interpretability, posing a common problem for doctors [95,160].

To overcome or reduce the negative impact of the "black box," Han et al [61] suggested that information gain ranking methods could be used, while Weng et al [71] advocated for the use of data visualization to facilitate interpretation. Besides, in the ECG methods for the prompt identification of coronary events study with 1244 participants, Al-Zaiti et al [15] revealed that by incorporating feature selection and annotation based on clinical knowledge, LR could achieve comparable performance to complex and expensive nonlinear algorithms such as ANN and GBM. This approach also resolved the "black box" issue of ANN and GBM, improving clinical utility by identifying predominant features contributing to ACS. Furthermore, methods such as discriminant analysis, Naive Bayes, logistic and Cox regression, and classification provide relationships between predictors and outcomes, enabling risk scores to be calculated for each predictor in clinical practice [160]. In addition, there has recently been an increasing focus on explaining AI algorithms, with Xuan et al [97] developing a visual analytics tool for comparing convolutional neural networks (CNNs) to support in-depth inspection and comparative studies of CNN models. Wang et al [98] have reported a CNN explainer, which can show the CNN's structure and provides on-demand dynamic visual explanations. Furthermore, it is exciting that explainable AI techniques have been applied to CVD prediction models and are reasonably expected to promote their application [161,162].

AI-Ethics Pitfalls Also Should Be Considered During the Study Design

Informed model selection is a pivotal decision-making step in AI, particularly when dealing with CVD risk prediction. It is imperative to initiate the process with a rigorous assessment of bias and fairness [144]. While it is acknowledged that most AI algorithms may inherently carry biases, it is equally important to benchmark these biases against those prevalent in the existing systems. AI's propensity to introduce bias often arises from making unequal errors across diverse demographic groups. The degree to which key demographic variables, including sex, age, and ethnicity, are adequately represented in the data set and incorporated during algorithm development significantly impacts the predictive accuracy across subgroups. Consequently, when these AI-derived predictions inform individual decisions, they can either perpetuate or intensify the existing disparities [163]. This issue is exacerbated by data that may not authentically mirror the entire target population, underscoring the ethical imperative of ensuring fairness at every stage of the project life cycle [164]. In the domain of CVD risk prediction, these ethical considerations loom large. For example, Kim et al [99] and Weng et al [71] constructed AI models for CVD risk prediction in the Korea National Health and Nutrition Examination Survey-VI cohort (4244 participants) and the CPRD cohort (378,256 participants), respectively. However, they failed to perform sensitivity analysis for race and gender subgroups, leading to gender and ethnic bias. These issues will significantly impact the application of the models.

There is an undeniable global demand for the development of AI systems that can be relied upon for their trustworthiness [165-168]. This demand is underscored by real-world examples, such as ProPublica's investigation into the Correctional Offender Management Profiling for Alternative Sanctions software, which serves as a stark illustration of how algorithmic performance can exhibit disparities rooted in racial factors. Notably, the field of fairness evaluation has made significant methodological advancements that directly facilitate such analyses [169]. It is imperative for AI developers and health care professionals to actively engage with these tools. Researchers can effectively showcase bias in critical subgroups, such as minority ethnic communities or distinct age groups, through explicit presentation of these findings. This approach not only enhances transparency but also empowers users of the AI algorithm with the knowledge of its strengths and weaknesses in different demographic contexts. This consideration holds particular relevance in the domain of AI-CVD risk prediction. Segar et al [100] included 19,080 participants from Atherosclerosis Risk in Communities, Dallas Heart Study, Jackson Heart Study, and MESA cohort to

XSL•FO
RenderX

build 6 AI models for predicting incident HF. They performed sensitivity analysis for the subgroup of race and sex, and all models performed well.

Conclusions

In conclusion, we highlight that the effectiveness and application of ML models for predicting CVD risk are heavily reliant on data quality, data set characteristics, model design and statistical methods, and clinical implications. To address these challenges, we propose practical solutions, including gathering objective data, improving training, repeating measurements, increasing sample size, preventing overfitting using statistical methods, using specific AI algorithms to address targeted issues, standardizing outcomes and evaluation criteria, and enhancing fairness and replicability. It may provide a helpful reference and assistance to researchers, algorithm developers, policy makers, and clinical practitioners in the ML field of CVD prediction.

Acknowledgments

The authors would like to gratefully acknowledge the assistance of Dr Jesse Li-Ling not only for technical support but also for their helpful discussions on topics related to this viewpoint paper. The work was supported by the National Key Research and Development Program of China (2020YFC2006406 and 2020YFC2006401) and the National College Students' Innovation and Entrepreneurship Training Program (202110159005).

Authors' Contributions

GWZ and XZ contributed to the overarching research goals of the paper as well as the manuscript review and editing. YQC, YC, MG, TCJ, and LYT performed the literature search together. They also collected important background information and wrote the manuscript. HJL and ZWZ helped with manuscript preparation. DXG and WH performed supervision and project administration. All authors have read and approved the content of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies of artificial intelligence or machine learning assessment guidelines or tools.
[DOCX File , 15 KB - jmir_v26i1e47645_app1.docx]

Multimedia Appendix 2
The evaluation content of assessment guidelines or tools in the field of medical artificial intelligence or machine learning research.
[DOCX File , 270 KB - jmir_v26i1e47645_app2.docx]

Multimedia Appendix 3
The framework for pitfalls in developing machine learning models for cardiovascular disease prediction.
[DOCX File , 149 KB - jmir_v26i1e47645_app3.docx]

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
2. Correction to: 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Circulation 2019 Sep 10;140(11):e649-e650. [doi: 10.1161/CIR.0000000000000725] [Medline: 31498691]
3. Pelliccia A, Sharma S, Gati S, Bäck M, Börjesson M, Caselli S, et al. 2020 ESC guidelines on sports cardiology and exercise in patients with cardiovascular disease. Eur Heart J 2021 Jan 01;42(1):17-96 [FREE Full text] [doi: 10.1093/eurheartj/ehaa605] [Medline: 32860412]
4. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet 2014 Mar 15;383(9921):999-1008 [FREE Full text] [doi: 10.1016/S0140-6736(13)61752-3] [Medline: 24084292]
5. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 2007 Jul 21;335(7611):136 [FREE Full text] [doi: 10.1136/bmj.39261.471806.55] [Medline: 17615182]
6. Goff DCJ, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RBS, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force

XSL•FO
RenderX

on Practice Guidelines. J Am Coll Cardiol 2014 Jul 01;63(25 Pt B):2935-2959 [FREE Full text] [doi: 10.1016/j.jacc.2013.11.005] [Medline: 24239921]
7. Khan SS, Ning H, Shah SJ, Yancy CW, Carnethon M, Berry JD, et al. 10-year risk equations for incident heart failure in the general population. J Am Coll Cardiol 2019 May 21;73(19):2388-2397 [FREE Full text] [doi: 10.1016/j.jacc.2019.02.057] [Medline: 31097157]
8. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019 Dec 31;188(12):2222-2239. [doi: 10.1093/aje/kwz189] [Medline: 31509183]
9. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019 Mar 19;19(1):64 [FREE Full text] [doi: 10.1186/s12874-019-0681-4] [Medline: 30890124]
10. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
11. Rasmy L, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform 2018 Aug;84:11-16 [FREE Full text] [doi: 10.1016/j.jbi.2018.06.011] [Medline: 29908902]
12. Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, et al. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation. JMIR Med Inform 2020 Jul 06;8(7):e17257 [FREE Full text] [doi: 10.2196/17257] [Medline: 32628616]
13. Han L, Askari M, Altman RB, Schmitt SK, Fan J, Bentley JP, et al. Atrial fibrillation burden signature and near-term prediction of stroke: a machine learning analysis. Circ Cardiovasc Qual Outcomes 2019 Oct;12(10):e005595 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.118.005595] [Medline: 31610712]
14. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet 2019 Sep;394(10201):861-867. [doi: 10.1016/s0140-6736(19)31721-0]
15. Al-Zaiti S, Besomi L, Bouzid Z, Faramand Z, Frisch S, Martin-Gill C, et al. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. Nat Commun 2020 Aug 07;11(1):3966 [FREE Full text] [doi: 10.1038/s41467-020-17804-2] [Medline: 32769990]
16. Sengupta PP, Kulkarni H, Narula J. Prediction of abnormal myocardial relaxation from signal processed surface ECG. J Am Coll Cardiol 2018 Apr 17;71(15):1650-1660 [FREE Full text] [doi: 10.1016/j.jacc.2018.02.024] [Medline: 29650121]
17. Kagiyama N, Piccirilli M, Yanamala N, Shrestha S, Farjo PD, Casaclang-Verzosa G, et al. Machine learning assessment of left ventricular diastolic function based on electrocardiographic features. J Am Coll Cardiol 2020 Aug 25;76(8):930-941 [FREE Full text] [doi: 10.1016/j.jacc.2020.06.061] [Medline: 32819467]
18. Bouzid Z, Faramand Z, Gregg RE, Frisch SO, Martin-Gill C, Saba S, et al. In search of an optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department. J Am Heart Assoc 2021 Feb 02;10(3):e017871 [FREE Full text] [doi: 10.1161/JAHA.120.017871] [Medline: 33459029]
19. Dogan MV, Beach SR, Simons RL, Lendasse A, Penaluna B, Philibert RA. Blood-based biomarkers for predicting the risk for five-year incident coronary heart disease in the Framingham Heart Study via machine learning. Genes (Basel) 2018 Dec 18;9(12):641 [FREE Full text] [doi: 10.3390/genes9120641] [Medline: 30567402]
20. Khanna NN, Maindarkar M, Saxena A, Ahluwalia P, Paul S, Srivastava SK, et al. Cardiovascular/stroke risk assessment in patients with erectile dysfunction-a role of carotid wall arterial imaging and plaque tissue characterization using artificial intelligence paradigm: a narrative review. Diagnostics (Basel) 2022 May 17;12(5):1249 [FREE Full text] [doi: 10.3390/diagnostics12051249] [Medline: 35626404]
21. Chao H, Shan H, Homayounieh F, Singh R, Khera RD, Guo H, et al. Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. Nat Commun 2021 May 20;12(1):2963 [FREE Full text] [doi: 10.1038/s41467-021-23235-4] [Medline: 34017001]
22. Cai Y, Cai YQ, Tang LY, Wang YH, Gong M, Jing TC, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. BMC Med 2024 Feb 05;22(1):56 [FREE Full text] [doi: 10.1186/s12916-024-03273-7] [Medline: 38317226]
23. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281 [FREE Full text] [doi: 10.1136/bmj.n2281] [Medline: 34670780]
24. Li B, Feridooni T, Cuen-Ojeda C, Kishibe T, de Mestral C, Mamdani M, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. NPJ Digit Med 2022 Jan 19;5(1):7 [FREE Full text] [doi: 10.1038/s41746-021-00552-y] [Medline: 35046493]
25. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51. [doi: 10.7326/m18-1376]
26. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015 Jan 06;162(1):55-63 [FREE Full text] [doi: 10.7326/M14-0697] [Medline: 25560714]

XSL•FO
RenderX

27. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Apr 01;369:m1312 [FREE Full text] [doi: 10.1136/bmj.m1312] [Medline: 32238345]
28. Suri JS, Bhagawati M, Paul S, Protogeron A, Sfikakis PP, Kitas GD, et al. Understanding the bias in machine learning systems for cardiovascular disease risk assessment: the first of its kind review. Comput Biol Med 2022 Mar;142:105204. [doi: 10.1016/j.compbiomed.2021.105204] [Medline: 35033879]
29. Russo S, Bonassi S. Prospects and pitfalls of machine learning in nutritional epidemiology. Nutrients 2022 Apr 20;14(9):1705 [FREE Full text] [doi: 10.3390/nu14091705] [Medline: 35565673]
30. Chiarito M, Luceri L, Oliva A, Stefanini G, Condorelli G. Artificial intelligence and cardiovascular risk prediction: all that glitters is not gold. Eur Cardiol 2022 Feb 20;17:e29 [FREE Full text] [doi: 10.15420/ecr.2022.11] [Medline: 36845218]
31. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
32. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. J Med Internet Res 2023 Aug 31;25:e48763 [FREE Full text] [doi: 10.2196/48763] [Medline: 37651179]
33. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
34. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020 Sep;26(9):1320-1324 [FREE Full text] [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]
35. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ 2020 Sep 09;370:m3164 [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]
36. Tanguay W, Acar P, Fine B, Abdolell M, Gong B, Cadrin-Chênevert A, et al. Assessment of radiology artificial intelligence software: a validation and evaluation framework. Can Assoc Radiol J 2023 May;74(2):326-333 [FREE Full text] [doi: 10.1177/08465371221135760] [Medline: 36341574]
37. Lennerz JK, Salgado R, Kim GE, Sirintrapun SJ, Thierauf JC, Singh A, et al. Diagnostic quality model (DQM): an integrated framework for the assessment of diagnostic quality when using AI/ML. Clin Chem Lab Med 2023 Jan 25;61(4):544-557 [FREE Full text] [doi: 10.1515/cclm-2022-1151] [Medline: 36696602]
38. Cerdá-Alberich L, Solana J, Mallol P, Ribas G, García-Junco M, Alberich-Bayarri A, et al. MAIC-10 brief quality checklist for publications using artificial intelligence and medical images. Insights Imaging 2023 Jan 16;14(1):11 [FREE Full text] [doi: 10.1186/s13244-022-01355-9] [Medline: 36645542]
39. Sujan M, Smith-Frazer C, Malamateniou C, Connor J, Gardner A, Unsworth H, et al. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health Care Inform 2023 Jun;30(1):e100749 [FREE Full text] [doi: 10.1136/bmjhci-2023-100749] [Medline: 37364922]
40. Mylrea M, Robinson N. Artificial intelligence (AI) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical AI. Entropy (Basel) 2023 Oct 09;25(10):1429 [FREE Full text] [doi: 10.3390/e25101429] [Medline: 37895550]
41. Kwong JC, Khondker A, Lajkosz K, McDermott MB, Frigola XB, McCradden MD, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open 2023 Sep 05;6(9):e2335377 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.35377] [Medline: 37747733]
42. Kocak B, Baessler B, Bakas S, Cuocolo R, Fedorov A, Maier-Hein L, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging 2023 May 04;14(1):75 [FREE Full text] [doi: 10.1186/s13244-023-01415-8] [Medline: 37142815]
43. van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J 2022 Aug 14;43(31):2921-2930 [FREE Full text] [doi: 10.1093/eurheartj/ehac238] [Medline: 35639667]
44. Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M, et al. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR Derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group. JAMA Dermatol 2022 Jan 01;158(1):90-96 [FREE Full text] [doi: 10.1001/jamadermatol.2021.4915] [Medline: 34851366]
45. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904 [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
46. Jha AK, Bradshaw TJ, Buvat I, Hatt M, Kc P, Liu C, et al. Nuclear medicine and artificial intelligence: best practices for evaluation (the RELAINCE guidelines). J Nucl Med 2022 Sep 26;63(9):1288-1299 [FREE Full text] [doi: 10.2967/jnumed.121.263239] [Medline: 35618476]

XSL•FO
RenderX

47. Banerjee A, Chen S, Fatemifar G, Zeina M, Lumbers RT, Mielke J, et al. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. BMC Med 2021 Apr 06;19(1):85 [FREE Full text] [doi: 10.1186/s12916-021-01940-7] [Medline: 33820530]
48. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods 2021 Oct 27;18(10):1122-1127 [FREE Full text] [doi: 10.1038/s41592-021-01205-4] [Medline: 34316068]
49. Olczak J, Pavlopoulos J, Prijs J, Ijpma FF, Doornberg JN, Lundström C, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop 2021 Oct 14;92(5):513-525 [FREE Full text] [doi: 10.1080/17453674.2021.1918389] [Medline: 33988081]
50. Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, et al. The AIMe registry for artificial intelligence in biomedical research. Nat Methods 2021 Oct 25;18(10):1128-1131 [FREE Full text] [doi: 10.1038/s41592-021-01241-0] [Medline: 34433960]
51. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent 2021 Apr;107:103610. [doi: 10.1016/j.jdent.2021.103610] [Medline: 33631303]
52. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform 2021 Feb 05;28(1):e100251 [FREE Full text] [doi: 10.1136/bmjhci-2020-100251] [Medline: 33547086]
53. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020 Oct;2(10):e537-e548 [FREE Full text] [doi: 10.1016/S2589-7500(20)30218-1] [Medline: 33328048]
54. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. JACC Cardiovasc Imaging 2020 Sep;13(9):2017-2035 [FREE Full text] [doi: 10.1016/j.jcmg.2020.07.015] [Medline: 32912474]
55. Cruz Rivera S, Liu X, Chan A, Denniston AK, Calvert MJ, SPIRIT-AICONSORT-AI Working Group, SPIRIT-AICONSORT-AI Steering Group, SPIRIT-AICONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020 Sep;26(9):1351-1363 [FREE Full text] [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
56. Kakarmath S, Esteva A, Arnaout R, Harvey H, Kumar S, Muse E, et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. NPJ Digit Med 2020 Oct 16;3(1):134 [FREE Full text] [doi: 10.1038/s41746-020-00336-w] [Medline: 33083569]
57. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10). [doi: 10.1161/circoutcomes.120.006556]
58. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017 Dec 4;14(12):749-762. [doi: 10.1038/nrclinonc.2017.141] [Medline: 28975929]
59. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 06;162(1):W1-73 [FREE Full text] [doi: 10.7326/M14-0698] [Medline: 25560730]
60. van Royen FS, Asselbergs FW, Alfonso F, Vardas P, van Smeden M. Five critical quality criteria for artificial intelligence-based prediction models. Eur Heart J 2023 Dec 07;44(46):4831-4834 [FREE Full text] [doi: 10.1093/eurheartj/ehad727] [Medline: 37897346]
61. Han D, Kolli KK, Gransar H, Lee JH, Choi SY, Chun EJ, et al. Machine learning based risk prediction model for asymptomatic individuals who underwent coronary artery calcium score: comparison with traditional risk prediction approaches. J Cardiovasc Comput Tomogr 2020;14(2):168-176. [doi: 10.1016/j.jcct.2019.09.005] [Medline: 31570323]
62. Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. NPJ Digit Med 2020 Sep 23;3(1):125 [FREE Full text] [doi: 10.1038/s41746-020-00331-1] [Medline: 33043149]
63. Rodriguez F, Chung S, Blum MR, Coulet A, Basu S, Palaniappan LP. Atherosclerotic cardiovascular disease risk prediction in disaggregated Asian and Hispanic subgroups using electronic health records. J Am Heart Assoc 2019 Jul 16;8(14):e011874 [FREE Full text] [doi: 10.1161/JAHA.118.011874] [Medline: 31291803]
64. Teoh D. Towards stroke prediction using electronic health records. BMC Med Inform Decis Mak 2018 Dec 4;18(1):127 [FREE Full text] [doi: 10.1186/s12911-018-0702-y] [Medline: 30509279]
65. Nakanishi R, Slomka PJ, Rios R, Betancur J, Blaha MJ, Nasir K, et al. Machine learning adds to clinical and CAC assessments in predicting 10-year CHD and CVD deaths. JACC Cardiovasc Imaging 2021 Mar;14(3):615-625 [FREE Full text] [doi: 10.1016/j.jcmg.2020.08.024] [Medline: 33129741]

XSL•FO
RenderX

66. Benjamins JW, van Leeuwen K, Hofstra L, Rienstra M, Appelman Y, Nijhof W, et al. Enhancing cardiovascular artificial intelligence (AI) research in the Netherlands: CVON-AI consortium. Neth Heart J 2019 Sep 20;27(9):414-425 [FREE Full text] [doi: 10.1007/s12471-019-1281-y] [Medline: 31111459]
67. Liu L, Tan S, Li Y, Luo J, Zhang W, Li S. An early aortic dissection screening model and applied research based on ensemble learning. Ann Transl Med 2020 Dec;8(23):1578 [FREE Full text] [doi: 10.21037/atm-20-1475] [Medline: 33437777]
68. Tsoi KK, Chan NB, Yiu KK, Poon SK, Lin B, Ho K. Machine learning clustering for blood pressure variability applied to systolic blood pressure intervention trial (SPRINT) and the Hong Kong community cohort. Hypertension 2020 Aug;76(2):569-576. [doi: 10.1161/hypertensionaha.119.14213]
69. Sajeev S, Champion S, Beleigoli A, Chew D, Reed RL, Magliano DJ, et al. Predicting Australian adults at high risk of cardiovascular disease mortality using standard risk factors and machine learning. Int J Environ Res Public Health 2021 Mar 19;18(6):3187 [FREE Full text] [doi: 10.3390/ijerph18063187] [Medline: 33808743]
70. Wang H, Li Y, Ning H, Wilkins J, Lloyd-Jones D, Luo Y. Using machine learning to integrate socio-behavioral factors in predicting cardiovascular-related mortality risk. Stud Health Technol Inform 2019 Aug 21;264:433-437. [doi: 10.3233/SHTI190258] [Medline: 31437960]
71. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017 Apr 4;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
72. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. BMJ 2020 Nov 04;371:m3919 [FREE Full text] [doi: 10.1136/bmj.m3919] [Medline: 33148619]
73. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 2019 May 15;14(5):e0213653 [FREE Full text] [doi: 10.1371/journal.pone.0213653] [Medline: 31091238]
74. Sung JM, Cho IJ, Sung D, Kim S, Kim HC, Chae MH, et al. Development and verification of prediction models for preventing cardiovascular diseases. PLoS One 2019 Sep 19;14(9):e0222809 [FREE Full text] [doi: 10.1371/journal.pone.0222809] [Medline: 31536581]
75. Ayala Solares JR, Canoy D, Raimondi FE, Zhu Y, Hassaine A, Salimi-Khorshidi G, et al. Long-term exposure to elevated systolic blood pressure in predicting incident cardiovascular disease: evidence from large-scale routine electronic health records. J Am Heart Assoc 2019 Jun 18;8(12):e012129 [FREE Full text] [doi: 10.1161/JAHA.119.012129] [Medline: 31164039]
76. Cho IJ, Sung JM, Kim HC, Lee SE, Chae MH, Kavousi M, et al. Development and external validation of a deep learning algorithm for prognostication of cardiovascular outcomes. Korean Circ J 2020 Jan;50(1):72-84 [FREE Full text] [doi: 10.4070/kcj.2019.0105] [Medline: 31456363]
77. Jdanov DA, Deev AD, Jasilionis D, Shalnova SA, Shkolnikova MA, Shkolnikov VM. Recalibration of the SCORE risk chart for the Russian population. Eur J Epidemiol 2014 Sep 2;29(9):621-628 [FREE Full text] [doi: 10.1007/s10654-014-9947-7] [Medline: 25179794]
78. Ponomartseva DA, Derevitskii IV, Kovalchuk SV, Babenko AY. Prediction model for thyrotoxic atrial fibrillation: a retrospective study. BMC Endocr Disord 2021 Jul 11;21(1):150 [FREE Full text] [doi: 10.1186/s12902-021-00809-3] [Medline: 34246271]
79. Chan KL, Leng X, Zhang W, Dong W, Qiu Q, Yang J, et al. Early identification of high-risk TIA or minor stroke using artificial neural network. Front Neurol 2019 Mar 1;10:171 [FREE Full text] [doi: 10.3389/fneur.2019.00171] [Medline: 30881336]
80. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. J Am Heart Assoc 2018 Nov 20;7(22). [doi: 10.1161/jaha.118.009476]
81. Chua W, Purmah Y, Cardoso VR, Gkoutos GV, Tull SP, Neculau G, et al. Data-driven discovery and validation of circulating blood-based biomarkers associated with prevalent atrial fibrillation. Eur Heart J 2019 Apr 21;40(16):1268-1276 [FREE Full text] [doi: 10.1093/eurheartj/ehy815] [Medline: 30615112]
82. Goldstein BA, Chang TI, Mitani AA, Assimes TL, Winkelmayer WC. Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records. Clin J Am Soc Nephrol 2014 Jan;9(1):82-91 [FREE Full text] [doi: 10.2215/CJN.03050313] [Medline: 24178968]
83. Angeli F, Reboldi G, Verdecchia P. Hypertension and the J-curve phenomenon: implications for tight blood pressure control. Hypertens Res 2013 Feb 15;36(2):109-111. [doi: 10.1038/hr.2012.165] [Medline: 23154585]
84. Lip S, Tan LE, Jeemon P, McCallum L, Dominiczak AF, Padmanabhan S. Diastolic blood pressure J-curve phenomenon in a tertiary-care hypertension clinic. Hypertension 2019 Oct;74(4):767-775. [doi: 10.1161/hypertensionaha.119.12787]
85. Dutta A, Batabyal T, Basu M, Acton ST. An efficient convolutional neural network for coronary heart disease prediction. Expert Syst Appl 2020 Nov;159:113408. [doi: 10.1016/j.eswa.2020.113408]

XSL•FO
RenderX

86. Zhuang XD, Tian T, Liao LZ, Dong YH, Zhou HJ, Zhang SZ, et al. Deep phenotyping and prediction of long-term cardiovascular disease: optimized by machine learning. Can J Cardiol 2022 Jun;38(6):774-782. [doi: 10.1016/j.cjca.2022.02.008] [Medline: 35157988]
87. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res 2017 Oct 13;121(9):1092-1101. [doi: 10.1161/CIRCRESAHA.117.311312] [Medline: 28794054]
88. Pujadas ER, Raisi-Estabragh Z, Szabo L, McCracken C, Morcillo CI, Campello VM, et al. Prediction of incident cardiovascular events using machine learning and CMR radiomics. Eur Radiol 2023 May;33(5):3488-3500 [FREE Full text] [doi: 10.1007/s00330-022-09323-z] [Medline: 36512045]
89. Lou YS, Lin CS, Fang WH, Lee CC, Lin C. Extensive deep learning model to enhance electrocardiogram application via latent cardiovascular feature extraction from identity identification. Comput Methods Programs Biomed 2023 Apr;231:107359. [doi: 10.1016/j.cmpb.2023.107359] [Medline: 36738606]
90. You J, Guo Y, Kang JJ, Wang HF, Yang M, Feng JF, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke Vasc Neurol 2023 Dec 29;8(6):475-485 [FREE Full text] [doi: 10.1136/svn-2023-002332] [Medline: 37105576]
91. Chen X, Liu S, Chu J, Hu W, Sun N, Shen Y. Joint effect of elevated-c-reactive protein level and hypertension on new-onset stroke: a nationwide prospective cohort study of CHARLS. Front Public Health 2022 Oct 3;10:919506 [FREE Full text] [doi: 10.3389/fpubh.2022.919506] [Medline: 36262245]
92. Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, et al. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. Cardiovasc Res 2020 Dec 01;116(14):2216-2225 [FREE Full text] [doi: 10.1093/cvr/cvz321] [Medline: 31853543]
93. Voss R, Cullen P, Schulte H, Assmann G. Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks. Int J Epidemiol 2002 Dec;31(6):1253-62; discussion 1262-64. [doi: 10.1093/ije/31.6.1253] [Medline: 12540731]
94. Polonsky TS, McClelland RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. JAMA 2010 Apr 28;303(16):1610-1616 [FREE Full text] [doi: 10.1001/jama.2010.461] [Medline: 20424251]
95. Unnikrishnan P, Kumar DK, Poosapadi Arjunan S, Kumar H, Mitchell P, Kawasaki R. Development of health parameter model for risk prediction of CVD using SVM. Comput Math Methods Med 2016;2016:3016245 [FREE Full text] [doi: 10.1155/2016/3016245] [Medline: 27594895]
96. Zarkogianni K, Athanasiou M, Thanopoulou AC, Nikita KS. Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication. IEEE J Biomed Health Inform 2018 Sep;22(5):1637-1647. [doi: 10.1109/jbhi.2017.2765639]
97. Xuan X, Zhang X, Kwon OH, Ma KL. VAC-CNN: a visual analytics system for comparative studies of deep convolutional neural networks. IEEE Trans Visual Comput Graphics 2022 Jun;28(6):2326-2337. [doi: 10.1109/tvcg.2022.3165347]
98. Wang ZJ, Turko R, Shaikh O, Park H, Das N, Hohman F, et al. CNN explainer: learning convolutional neural networks with interactive visualization. IEEE Trans Visual Comput Graphics 2021 Feb;27(2):1396-1406. [doi: 10.1109/tvcg.2020.3030418]
99. Kim J, Kang U, Lee Y. Statistics and deep belief network-based cardiovascular risk prediction. Healthc Inform Res 2017 Jul;23(3):169-175 [FREE Full text] [doi: 10.4258/hir.2017.23.3.169] [Medline: 28875051]
100. Segar MW, Jaeger BC, Patel KV, Nambi V, Ndumele CE, Correa A, et al. Development and validation of machine learning-based race-specific models to predict 10-year risk of heart failure: a multicohort analysis. Circulation 2021 Jun 15;143(24):2370-2383 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.053134] [Medline: 33845593]
101. Grimes DA. Epidemiologic research using administrative databases: garbage in, garbage out. Obstet Gynecol 2010;116(5):1018-1019. [doi: 10.1097/aog.0b013e3181f98300]
102. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ 2010 Dec 09;341(dec09 1):c6624 [FREE Full text] [doi: 10.1136/bmj.c6624] [Medline: 21148212]
103. Manuel DG, Tuna M, Bennett C, Hennessy D, Rosella L, Sanmartin C, et al. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the Cardiovascular Disease Population Risk Tool (CVDPoRT). CMAJ 2018 Jul 23;190(29):E871-E882 [FREE Full text] [doi: 10.1503/cmaj.170914] [Medline: 30037888]
104. Inoue K, Seeman TE, Horwich T, Budoff MJ, Watson KE. Heterogeneity in the association between the presence of coronary artery calcium and cardiovascular events: a machine-learning approach in the MESA study. Circulation 2023 Jan 10;147(2):132-141. [doi: 10.1161/circulationaha.122.062626]
105. Li JX, Li L, Zhong X, Fan SJ, Cen T, Wang J, et al. Machine learning identifies prominent factors associated with cardiovascular disease: findings from two million adults in the Kashgar Prospective Cohort Study (KPCS). Glob Health Res Policy 2022 Dec 06;7(1):48 [FREE Full text] [doi: 10.1186/s41256-022-00282-y] [Medline: 36474302]

XSL•FO
RenderX

106. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 01;170(1):W1-33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
107. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 07;350:g7594 [FREE Full text] [doi: 10.1136/bmj.g7594] [Medline: 25569120]
108. Johansson JK, Puukka PJ, Niiranen TJ, Varis J, Peltonen M, Salomaa V, et al. Health 2000 score - development and validation of a novel cardiovascular risk score. Ann Med 2016 Sep;48(6):403-409 [FREE Full text] [doi: 10.1080/07853890.2016.1183257] [Medline: 27187608]
109. Wu Y, Liu X, Li X, Li Y, Zhao L, Chen Z, et al. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. Circulation 2006 Nov 21;114(21):2217-2225. [doi: 10.1161/circulationaha.105.607499]
110. Carresi C, Scicchitano M, Scarano F, Macrì R, Bosco F, Nucera S, et al. The potential properties of natural compounds in cardiac stem cell activation: their role in myocardial regeneration. Nutrients 2021 Jan 19;13(1):275 [FREE Full text] [doi: 10.3390/nu13010275] [Medline: 33477916]
111. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ 2017 May 23;357:j2099 [FREE Full text] [doi: 10.1136/bmj.j2099] [Medline: 28536104]
112. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol 2018 Dec 29;18(1):179 [FREE Full text] [doi: 10.1186/s12874-018-0644-1] [Medline: 30594138]
113. Saeed A, Nambi V, Sun W, Virani SS, Taffet GE, Deswal A, et al. Short-term global cardiovascular disease risk prediction in older adults. J Am Coll Cardiol 2018 Jun 05;71(22):2527-2536 [FREE Full text] [doi: 10.1016/j.jacc.2018.02.050] [Medline: 29535064]
114. Xu J, Li L, Huang S, Song H, Gao J, Ni H, et al. Impact of visit-to-visit fasting plasma glucose variability on the development of diabetes: the mediation by insulin resistance. J Diabetes 2022 Mar 16;14(3):205-215 [FREE Full text] [doi: 10.1111/1753-0407.13253] [Medline: 35170854]
115. Hukportie DN, Li FR, Zhou R, Zheng JZ, Wu XX, Zou MC, et al. Lipid variability and risk of microvascular complications in Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial: a post hoc analysis. J Diabetes 2022 Jun 06;14(6):365-376 [FREE Full text] [doi: 10.1111/1753-0407.13273] [Medline: 35668633]
116. Stergiou GS, Palatini P, Parati G, O'Brien E, Januszewicz A, Lurbe E, et al. 2021 European Society of Hypertension practice guidelines for office and out-of-office blood pressure measurement. J Hypertens 2021 Jul 01;39(7):1293-1302. [doi: 10.1097/HJH.0000000000002843] [Medline: 33710173]
117. Paige E, Barrett J, Pennells L, Sweeting M, Willeit P, Di Angelantonio E, et al. Use of repeated blood pressure and cholesterol measurements to improve cardiovascular disease risk prediction: an individual-participant-data meta-analysis. Am J Epidemiol 2017 Oct 15;186(8):899-907 [FREE Full text] [doi: 10.1093/aje/kwx149] [Medline: 28549073]
118. WHO CVD Risk Chart Working Group. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob Health 2019 Oct;7(10):e1332-e1345 [FREE Full text] [doi: 10.1016/S2214-109X(19)30318-3] [Medline: 31488387]
119. Hajifathalian K, Ueda P, Lu Y, Woodward M, Ahmadvand A, Aguilar-Salinas CA, et al. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. Lancet Diabetes Endocrinol 2015 May;3(5):339-355 [FREE Full text] [doi: 10.1016/S2213-8587(15)00081-9] [Medline: 25819778]
120. McClelland RL, Jorgensen NW, Budoff M, Blaha MJ, Post WS, Kronmal RA, et al. 10-Year coronary heart disease risk prediction using coronary artery calcium and traditional risk factors: derivation in the MESA (multi-ethnic study of atherosclerosis) with validation in the HNR (Heinz Nixdorf Recall) study and the DHS (Dallas Heart Study). J Am Coll Cardiol 2015 Oct 13;66(15):1643-1653 [FREE Full text] [doi: 10.1016/j.jacc.2015.08.035] [Medline: 26449133]
121. Parikh NI, Jeppson RP, Berger JS, Eaton CB, Kroenke CH, LeBlanc ES, et al. Reproductive risk factors and coronary heart disease in the women's health initiative observational study. Circulation 2016 May 31;133(22):2149-2158 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.017854] [Medline: 27143682]
122. Cui Z, Gong G. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage 2018 Sep;178:622-637. [doi: 10.1016/j.neuroimage.2018.06.001] [Medline: 29870817]
123. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J 2019 Nov 29;70(4):344-353 [FREE Full text] [doi: 10.1016/j.carj.2019.06.002] [Medline: 31522841]
124. Ma J, Fong SH, Luo Y, Bakkenist CJ, Shen JP, Mourragui S, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. Nat Cancer 2021 Feb 25;2(2):233-244 [FREE Full text] [doi: 10.1038/s43018-020-00169-2] [Medline: 34223192]

XSL•FO
RenderX

125. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 09;11(7):e048008 [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
126. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. The Lancet 2019 Apr;393(10181):1577-1579. [doi: 10.1016/s0140-6736(19)30037-6]
127. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022 Jan 13;22(1):12 [FREE Full text] [doi: 10.1186/s12874-021-01469-6] [Medline: 35026997]
128. Backholer K, Hirakawa Y, Tonkin A, Giles G, Magliano DJ, Colagiuri S, et al. Development of an Australian cardiovascular disease mortality risk score using multiple imputation and recalibration from national statistics. BMC Cardiovasc Disord 2017 Jan 06;17(1):17 [FREE Full text] [doi: 10.1186/s12872-016-0462-5] [Medline: 28061760]
129. Dritsas E, Trigka M. Efficient data-driven machine learning models for cardiovascular diseases risk prediction. Sensors (Basel) 2023 Jan 19;23(3):1161 [FREE Full text] [doi: 10.3390/s23031161] [Medline: 36772201]
130. Mandair D, Tiwari P, Simon S, Colborn KL, Rosenberg MA. Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data. BMC Med Inform Decis Mak 2020 Oct 02;20(1):252 [FREE Full text] [doi: 10.1186/s12911-020-01268-x] [Medline: 33008368]
131. Subramani S, Varshney N, Anand MV, Soudagar ME, Al-Keridis LA, Upadhyay TK, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front Med (Lausanne) 2023 Apr 17;10:1150933. [doi: 10.3389/fmed.2023.1150933] [Medline: 37138750]
132. Kong YW, Baqar S, Jerums G, Ekinci EI. Sodium and its role in cardiovascular disease - the debate continues. Front Endocrinol (Lausanne) 2016 Dec 23;7:164 [FREE Full text] [doi: 10.3389/fendo.2016.00164] [Medline: 28066329]
133. Dehghan M, Mente A, Zhang X, Swaminathan S, Li W, Mohan V, et al. Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents (PURE): a prospective cohort study. Lancet 2017 Nov 04;390(10107):2050-2062. [doi: 10.1016/S0140-6736(17)32252-3] [Medline: 28864332]
134. Cheung CY, Xu D, Cheng CY, Sabanayagam C, Tham YC, Yu M, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. Nat Biomed Eng 2021 Jun;5(6):498-508. [doi: 10.1038/s41551-020-00626-4] [Medline: 33046867]
135. Unterhuber M, Kresoja KP, Rommel KP, Besler C, Baragetti A, Klöting N, et al. Proteomics-enabled deep learning machine algorithms can enhance prediction of mortality. J Am Coll Cardiol 2021 Oct 19;78(16):1621-1631 [FREE Full text] [doi: 10.1016/j.jacc.2021.08.018] [Medline: 34649700]
136. Xia X, Zhu C, Zhong F, Liu L. MDTips: a multimodal-data-based drug-target interaction prediction system fusing knowledge, gene expression profile, and structural data. Bioinformatics 2023 Jul 01;39(7):btad411 [FREE Full text] [doi: 10.1093/bioinformatics/btad411] [Medline: 37379157]
137. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008 Jun 28;336(7659):1475-1482 [FREE Full text] [doi: 10.1136/bmj.39609.449676.25] [Medline: 18573856]
138. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003 Jun;24(11):987-1003. [doi: 10.1016/s0195-668x(03)00114-3] [Medline: 12788299]
139. Goff DCJ, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014 Jun 24;129(25 Suppl 2):S49-S73 [FREE Full text] [doi: 10.1161/01.cir.0000437741.48606.98] [Medline: 24222018]
140. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese population: the China-PAR project (prediction for ASCVD risk in China). Circulation 2016 Nov 08;134(19):1430-1440. [doi: 10.1161/CIRCULATIONAHA.116.022367] [Medline: 27682885]
141. Woodward M, Brindle P, Tunstall-Pedoe H, SIGN group on risk estimation. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart 2007 Feb;93(2):172-176 [FREE Full text] [doi: 10.1136/hrt.2006.108167] [Medline: 17090561]
142. Zeraatkar D, Cheung K, Milio K, Zworth M, Gupta A, Bhasin A, et al. Methods for the selection of covariates in nutritional epidemiology studies: a meta-epidemiological review. Curr Dev Nutr 2019 Sep 17;3(10):nzz104 [FREE Full text] [doi: 10.1093/cdn/nzz104] [Medline: 31598577]
143. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health 2001;22:189-212. [doi: 10.1146/annurev.publhealth.22.1.189] [Medline: 11274518]
144. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]

XSL•FO
RenderX

145. Brisk R, Bond R, Finlay D, McLaughlin J, Piadlo A, Leslie SJ, et al. The effect of confounding data features on a deep learning algorithm to predict complete coronary occlusion in a retrospective observational setting. Eur Heart J Digit Health 2021 Feb 20;2(1):127-134 [FREE Full text] [doi: 10.1093/ehjdh/ztab002] [Medline: 36711180]
146. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2019 Apr 30;2:31 [FREE Full text] [doi: 10.1038/s41746-019-0105-1] [Medline: 31304378]
147. Blackshear JL, McRee CW, Safford RE, Pollak PM, Stark ME, Thomas CS, et al. von Willebrand factor abnormalities and Heyde syndrome in dysfunctional heart valve prostheses. JAMA Cardiol 2016 May 01;1(2):198-204. [doi: 10.1001/jamacardio.2016.0075] [Medline: 27437891]
148. Kwiendacz H, Wijata AM, Nalepa J, Piaśnik J, Kulpa J, Herba M, et al. Machine learning profiles of cardiovascular risk in patients with diabetes mellitus: the Silesia diabetes-heart project. Cardiovasc Diabetol 2023 Aug 24;22(1):218 [FREE Full text] [doi: 10.1186/s12933-023-01938-w] [Medline: 37620935]
149. Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, Kähönen M, et al. Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young Finns study. PLoS Genet 2010 Sep 30;6(9):e1001146 [FREE Full text] [doi: 10.1371/journal.pgen.1001146] [Medline: 20941391]
150. Tesche C, Bauer MJ, Baquet M, Hedels B, Straube F, Hartl S, et al. Improved long-term prognostic value of coronary CT angiography-derived plaque measures and clinical parameters on adverse cardiac outcome using machine learning. Eur Radiol 2021 Jan;31(1):486-493. [doi: 10.1007/s00330-020-07083-2] [Medline: 32725337]
151. Zhang PI, Hsu CC, Kao Y, Chen CJ, Kuo YW, Hsu SL, et al. Real-time AI prediction for major adverse cardiac events in emergency department patients with chest pain. Scand J Trauma Resusc Emerg Med 2020 Sep 11;28(1):93 [FREE Full text] [doi: 10.1186/s13049-020-00786-x] [Medline: 32917261]
152. Pencina MJ, D'Agostino RBS, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. Circulation 2009 Jun 23;119(24):3078-3084 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.108.816694] [Medline: 19506114]
153. Lloyd-Jones DM, Leip EP, Larson MG, D'Agostino RB, Beiser A, Wilson PW, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation 2006 Feb 14;113(6):791-798. [doi: 10.1161/CIRCULATIONAHA.105.548206] [Medline: 16461820]
154. Wickramasinghe CD, Ayers CR, Das S, de Lemos JA, Willis BL, Berry JD. Prediction of 30-year risk for cardiovascular mortality by fitness and risk factor levels: the Cooper Center Longitudinal Study. Circ Cardiovasc Qual Outcomes 2014 Jul;7(4):597-602 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.113.000531] [Medline: 24987054]
155. Nakai M, Watanabe M, Kokubo Y, Nishimura K, Higashiyama A, Takegami M, et al. Development of a cardiovascular disease risk prediction model using the Suita study, a population-based prospective cohort study in Japan. J Atheroscler Thromb 2020 Nov 01;27(11):1160-1175 [FREE Full text] [doi: 10.5551/jat.48843] [Medline: 32023562]
156. Li Y, Yatsuya H, Tanaka S, Iso H, Okayama A, Tsuji I, et al. Estimation of 10-year risk of death from coronary heart disease, stroke, and cardiovascular disease in a pooled analysis of Japanese cohorts: EPOCH-JAPAN. J Atheroscler Thromb 2021 Aug 01;28(8):816-825 [FREE Full text] [doi: 10.5551/jat.58958] [Medline: 33041313]
157. Chang W, Liu Y, Wu X, Xiao Y, Zhou S, Cao W. A new hybrid XGBSVM model: application for hypertensive heart disease. IEEE Access 2019 Dec 03;7:175248-175258. [doi: 10.1109/access.2019.2957367]
158. Chen Y, Mao Y, Pan X, Jin W, Qiu T. Verification and comparison of three prediction models of ischemic stroke in young adults based on the back propagation neural networks. Medicine (Baltimore) 2021 Mar 19;100(11):e25081. [doi: 10.1097/MD.0000000000025081] [Medline: 33725985]
159. Barbieri S, Mehta S, Wu B, Bharat C, Poppe K, Jorm L, et al. Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. Int J Epidemiol 2022 Jun 13;51(3):931-944 [FREE Full text] [doi: 10.1093/ije/dyab258] [Medline: 34910160]
160. Quesada JA, Lopez-Pineda A, Gil-Guillén VF, Durazo-Arvizu R, Orozco-Beltrán D, López-Domenech A, et al. Machine learning to predict cardiovascular risk. Int J Clin Pract 2019 Oct;73(10):e13389. [doi: 10.1111/ijcp.13389] [Medline: 31264310]
161. Westerlund AM, Hawe JS, Heinig M, Schunkert H. Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence. Int J Mol Sci 2021 Sep 24;22(19):10291 [FREE Full text] [doi: 10.3390/ijms221910291] [Medline: 34638627]
162. Rao S, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Cleland J, et al. An explainable transformer-based deep learning model for the prediction of incident heart failure. IEEE J Biomed Health Inform 2022 Jul;26(7):3362-3372. [doi: 10.1109/jbhi.2022.3148820]
163. Curtis D. Polygenic risk score for schizophrenia is not strongly associated with the expression of specific genes or gene sets. Psychiatr Genet 2018 Aug;28(4):59-65. [doi: 10.1097/YPG.0000000000000197] [Medline: 29672343]
164. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol 2018 Feb 15;19(1):21 [FREE Full text] [doi: 10.1186/s13059-018-1396-2] [Medline: 29448949]

XSL•FO
RenderX

165. Advancing governance, innovation, and risk management for agency use of artificial intelligence. Executive Office of the President, United States. 2024 Mar 28. URL: https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf [accessed 2024-07-10]
166. Ethics guidelines for trustworthy AI. Publications Office of the European Union. URL: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en [accessed 2024-07-03]
167. Rees C, Müller B. All that glitters is not gold: trustworthy and ethical AI principles. AI Ethics 2022 Nov 16;3(4):1-14 [FREE Full text] [doi: 10.1007/s43681-022-00232-x] [Medline: 36406882]
168. Radclyffe C, Ribeiro M, Wortham RH. The assessment list for trustworthy artificial intelligence: a review and recommendations. Front Artif Intell 2023 Mar 9;6:1020592 [FREE Full text] [doi: 10.3389/frai.2023.1020592] [Medline: 36967834]
169. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 2019 Feb 01;21(2):E167-E179 [FREE Full text] [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]

Abbreviations

ACM: all-cause mortality
ACS: acute coronary syndrome
AI-TREE: AI transparent, replicable, ethical, and effective research
AI: artificial intelligence
ANLL: average negative log likelihood
ANN: artificial neural network
ASCVD: atherosclerotic cardiovascular diseases
AUC: area under the curve
BP: blood pressure
CHD: coronary heart disease
CNN: convolutional neural network
CONSORT: Consolidated Standards of Reporting Trials
CPRD: Clinical Practice Research Datalink
CVD: cardiovascular disease
ECG: electrocardiogram
EHR: electronic health record
FRS: Framingham risk score
GBM: gradient boosting machine
HF: heart failure
ICD: International Classification of Diseases
KNN: k-nearest neighbor
LR: logistic regression
MESA: Multi-Ethnic Study of Atherosclerosis
MI: myocardial infarction
ML: machine learning
NHIS-HEALS: National Health Insurance System-National Health Screening Cohort
PCE: pooled cohort equation
PROBAST: Prediction Model Risk of Bias Assessment Tool
RF: random forest
RSF: random survival forest
SBP: systolic blood pressure
SPRINT: Systolic Blood Pressure Intervention Trial
SVM: support vector machine
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis

XSL•FO
RenderX

Edited by T Leung, S Ma; submitted 29.03.23; peer-reviewed by K Gupta, N Jiwani, XC Liu; comments to author 17.09.23; revised version received 30.10.23; accepted 12.06.24; published 26.07.24.

Please cite as:
Cai YQ, Gong DX, Tang LY, Cai Y, Li HJ, Jing TC, Gong M, Hu W, Zhang ZW, Zhang X, Zhang GW
Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions
J Med Internet Res 2024;26:e47645
URL: https://www.jmir.org/2024/1/e47645
doi:10.2196/47645
PMID:

©Yu-Qing Cai, Da-Xin Gong, Li-Ying Tang, Yue Cai, Hui-Jun Li, Tian-Ci Jing, Mengchun Gong, Wei Hu, Zhen-Wei Zhang, Xingang Zhang, Guang-Wei Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Expanding a Behavioral View on Digital Health Access: Drivers and Strategies to Promote Equity

Maura M Kepper[1], MPH, PhD; Lauren A Fowler[2], PhD; Isabelle S Kusters[3,4], MPH, PhD; Jean W Davis[5], EdD, DNP, PhD; Manal Baqer[6], PhD; Sara Sagui-Henson[7], PhD; Yunyu Xiao[8], PhD; Adati Tarfa[9], MS, PharmD, PhD; Jean C Yi[10], PhD; Bryan Gibson[11], DPT, PhD; Kristin E Heron[12,13], PhD; Nicole M Alberts[14], PhD; Marissa Burgermaster[15,16], PhD; Veronica PS Njie-Carr[17], RN, ACNS-BC, PhD; Lisa M Klesges[18], PhD

[1]Prevention Research Center, Washington University in St. Louis, St. Louis, MO, United States
[2]Sexuality, Health, and Gender Center, Washington University in St. Louis School of Medicine, Saint Louis, MO, United States
[3]Department of Health, Human, and Biomedical Sciences, University of Houston-Clear Lake, Houston, TX, United States
[4]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX, United States
[5]College of Nursing, University of Central Florida, Orlando, FL, United States
[6]Neamah Health Consulting, Boston, MA, United States
[7]Clinical Strategy and Research Team, Modern Health, San Francisco, CA, United States
[8]Department of Population Health Science, Weill Cornell Medicine, Cornell University, New York, NY, United States
[9]School of Medicine, Yale University, New Haven, CT, United States
[10]Fred Hutchinson Cancer Center, Seattle, WA, United States
[11]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, United States
[12]Psychology Department, Old Dominion University, Norfolk, VA, United States
[13]Virginia Consortium Program in Clinical Psychology, Norfolk, VA, United States
[14]Department of Psychology, Concordia University, Montreal, QC, Canada
[15]Department of Nutritional Sciences, University of Texas at Austin, Austin, TX, United States
[16]Department of Population Health, Dell Medical School, University of Texas at Austin, Austin, TX, United States
[17]Department of Organizational Systems and Adult Health, University of Maryland, Baltimore, MD, United States
[18]Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, MO, United States

Corresponding Author:
Maura M Kepper, MPH, PhD
Prevention Research Center
Washington University in St. Louis
1 Brookings Drive
St. Louis, MO, 63130
United States
Phone: 1 3149350142
Email: kepperm@wustl.edu

Abstract

The potential and threat of digital tools to achieve health equity has been highlighted for over a decade, but the success of achieving equitable access to health technologies remains challenging. Our paper addresses renewed concerns regarding equity in digital health access that were deepened during the COVID-19 pandemic. Our viewpoint is that (1) digital health tools have the potential to improve health equity if equitable access is achieved, and (2) improving access and equity in digital health can be strengthened by considering behavioral science–based strategies embedded in all phases of tool development. Using behavioral, equity, and access frameworks allowed for a unique and comprehensive exploration of current drivers of digital health inequities. This paper aims to present a compilation of strategies that can potentially have an actionable impact on digital health equity. Multilevel factors drive unequal access, so strategies require action from tool developers, individual delivery agents, organizations, and systems to effect change. Strategies were shaped with a behavioral medicine focus as the field has a unique role in improving digital health access; arguably, all digital tools require the user (individual, provider, and health system) to change behavior by engaging with the technology to generate impact. This paper presents a model that emphasizes using multilevel strategies across design, delivery, dissemination, and sustainment stages to advance digital health access and foster health equity.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51355) doi:10.2196/51355

KEYWORDS

digital health; health equity; mobile health; mHealth; health care access; digital divide; behavioral medicine; implementation; mobile phone

Introduction

Behavioral medicine plays a key role in testing and implementing newly developed digital health technologies and evaluating their use in care delivery, including recognizing unique opportunities and challenges to equitable health services [1,2]. Along with many other health-related specialties, behavioral medicine responded to the COVID-19 pandemic with the rapid deployment of digital health and faced concerns that digital health disparities deepened during the pandemic [3-6]. Digital health tools, which include mobile health, health information technology, wearable devices, telehealth, and telemedicine [7], were rapidly implemented in lieu of in-person activities to limit the spread of COVID-19 while trying to ensure that access to care was uninterrupted [8]. While benefits of digital health were seen in some population groups, the pandemic also exposed long-standing inequities in health and health care, especially access to care. Often described as the "digital divide" [9], the gap between those with access to digital technologies and those without [10,11] came into stark relief during the pandemic. Thus, this again highlights the need to prioritize digital health equity while sustaining the momentum of the paradigm shift in health care delivery to digital health technology [12,13].

Behavioral medicine brings unique perspectives to bridging the digital divide. Other health care fields join behavioral medicine in conceptualizing more comprehensive models to identify strategies that address barriers and enhance facilitators to ensure greater equity in digital health. A body of work within behavioral medicine has focused on improving access to care [14] and reducing health disparities [1]. Alcaraz et al [1] developed the ConNECT framework that links behavioral science with health equity and provides strategies, including harnessing technology, focused on marginalized subgroups to inform practice and policy to promote health equity. Harnessing technology can promote equity by, for example, bridging accessibility barriers; using seamless technological adaptations (eg, language, literacy, or cultural tailoring); or scaling treatment approaches. These considerations are critical given that digital technologies have been repeatedly shown to exacerbate health disparities if equity-promoting approaches are not considered at every stage of the design of digital tools [15].

Complementary models such as the Digital Health Equity Framework address multilevel dimensions of population access to health care and build from the socioecological model to demonstrate that multilevel factors such as digital literacy, community infrastructure, and policy must be addressed to mitigate inequities [4]. *Access* is described by Levesque et al [14] in health care access conceptual framework as an individual's opportunity and ease of using appropriate services in proportion to their needs [14]. Although access is related to price, quality, and availability, it is also impacted by individuals' behaviors; individuals must use the service, which depends on many factors, such as an individual's preferences, values, knowledge, age, abilities, and level of clinical need.

Bringing complementary models together can identify unique opportunities for digital health access that are more vital in addressing health inequity [1]. Behavioral scientists have contributed to this growing field of digital health equity in areas such as integration of equity principles into digital interventions [16], inclusion of user-centered or co-design methods for cultural tailoring of digital health interventions [17], development of new ways to engage historically marginalized groups [18], and engagement of the community through formative evaluations when digitizing interventions for dissemination [19]. As the proliferation of digital health technologies for behavior change continues, the potential for wide-scale public impact across disciplines (eg, engineering, psychology, nursing, and public health) and disease and illness foci (eg, diabetes, cancer, obesity, HIV, and mental health) could be further realized if strategies to promote equitable access are used at all levels of the socioecological model. Our viewpoint is that (1) digital health tools have the potential to improve health equity if equitable access is achieved, and (2) improving access and equity in digital health can be strengthened by considering behavioral science–based strategies embedded in all phases of tool development. Our paper is the first to integrate equity (ConNECT) and access frameworks (Levesque et al [14]) to explore how dimensions of access impact digital health equity comprehensively. This paper aims to examine key drivers of inequities at multiple access levels and provides actionable strategies to promote digital health equity across stages of development, delivery, dissemination, and sustainment.

Methods and Guiding Principles

Health equity is an important focus within behavioral medicine, specifically the Society of Behavioral Medicine [20]. This paper aims to articulate a behavioral medicine perspective to improve digital health equity. Coauthors represent a multidisciplinary group of behavioral scientists from the Society of Behavioral Medicine spanning academia, industry, and nonprofit organizations with expertise in informatics, health services research, dissemination and implementation science, and health disparities research. Leveraging our working group's diverse experiences and expertise, we cataloged the key challenges of equitably implementing, disseminating, and sustaining digital health interventions among diverse populations and settings. As we compiled these challenges, six guiding principles emerged to generate a shared understanding critical to clarifying our perspectives that follow.

First, the field of digital health requires the application of overarching ethical principles and several unique considerations as outlined in the Ethics Checklist for Digital Health Research:

XSL•FO
RenderX

engagement of end users; informed consent; equity, diversity, and access; privacy and partnerships; regulation and law; and return of results [21].

Second, beyond ethical considerations of technology, connectivity, monitoring, and data management, it is critical to consider social, cultural, economic, infrastructure, policy, and other intersectionalities to promote digital health equities.

Third, digital health tools can be designed for and disseminated directly to individuals (eg, wearable devices) or can be used or delivered by health care organizations and teams. These distribution pathways have unique implications for implementation, dissemination, and sustainability (eg, funding sources) that result in unique drivers of inequities at multiple socioecological levels (providers, individuals, health care systems, third-party payors, and public health).

Fourth, understanding the groups that are underserved by digital health is a critical step. This paper refers to these groups broadly as underserved populations. By this, we mean people and communities disadvantaged by our systems and policies, including racial and ethnic groups, non–US-born persons, people with lower incomes, rural communities, people with disabilities, lesbian, gay, bisexual, transgender, queer, intersex, and asexual communities, and people who are incarcerated. Our intention is to speak broadly and without stigma to advance digital health for all.

Fifth, behavior change is inherent to the success of digital health interventions. All digital tools require behavior change; in other words, they require the user (individual, provider, and health system) to use the technology to generate impact for positive health outcomes. Digital tools may also have behavior change (eg, blood pressure self-monitoring and promotion of physical activity) as their primary purpose. Sometimes, behavior change must occur at multiple levels to generate the intended outcome. For example, a clinical decision support tool to promote cancer screening requires the clinician to change the way in which they are delivering care and requires the patient to get screened.

Sixth, to achieve equity, multilevel strategies (individual-, developer-, organizational-, and policy-levels) are necessary to consider during the design, dissemination, implementation, and sustainment stages of digital health interventions.

Theoretical Framework: Limited Access Drives Digital Health Inequity

We propose that overlaying a behavioral medicine perspective with multilevel, intersecting strategies that address digital health access inequities can yield greater reach, adoption, and impact of digital health across diverse populations. Applying the access framework by Levesque et al [14] to the digital divide can promote equity by considering the multiple dimensions of access to digital health: approachability, acceptability, availability, affordability, and appropriateness. Our adaptation of this framework specifies a digital access application, rather than health care access more broadly, and integrates intersecting factors influencing digital health accessibility and accounting for the barriers experienced at multiple steps. We found the framework by Levesque et al [14] useful for examining specific digital health inequities and presenting our application within the 5 access dimensions (Figure 1). We adapted specific digital health access barriers, focusing on multilevel factors that may generate differences in one's desire or perception of need and the ability to seek, reach, pay for, and engage with digital health [14,22]. Our framework uses a socioecological perspective to describe multilevel drivers of inequities and strategies to generate equitable access. This is critical, as digital health tools may be used by individuals and health care teams and disseminated through companies and organizations, as noted in principle 3 in the Methods and Guiding Principles section.

Figure 1. A conceptual framework of drivers of inequitable access to digital health. Adapted from the health care access framework by Levesque et al [14].

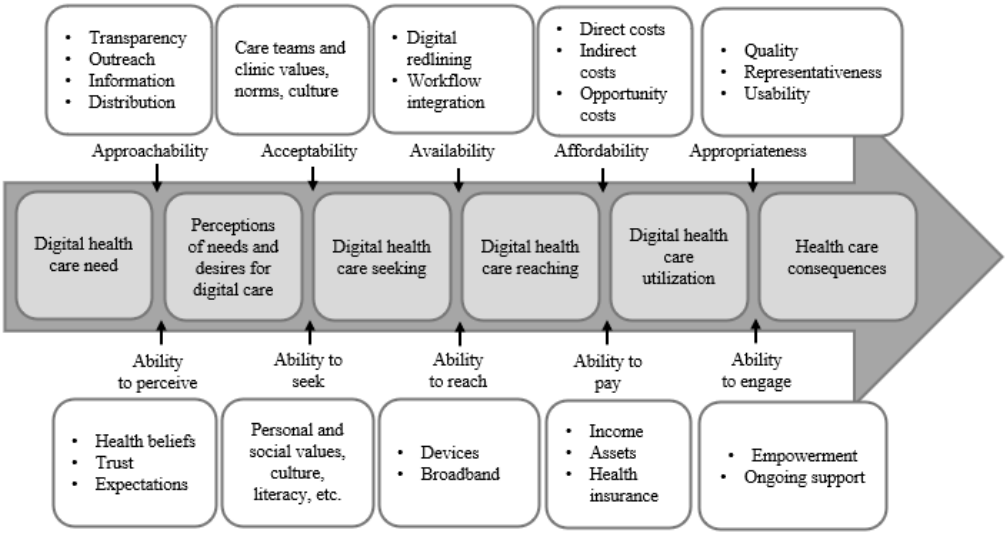

XSL•FO
RenderX

Drivers of Digital Health Inequities Using the Health Care Access Conceptual Framework

Approachability: As a Driver of Digital Health Inequities

A digital health tool is approachable if an individual is aware of the tool and perceives a need for it. However, awareness of digital health tools may vary among social and geographical population groups. Levesque et al [14] propose that transparency of data collection or use, tool content, and outreach activities all contribute to the approachability of health care services. However, outreach activities are often not tailored to diverse subpopulations within the intended audience and, therefore, do not equally increase awareness and produce the intended outcome for all groups, especially underserved populations [23,24]. Additionally, Levesque et al [14] indicate that individuals must perceive a need for care (ie, the digital health intervention), which is impacted by factors such as health knowledge and health-related beliefs. Individuals may lack a perceived need for a digital health tool that has not established credibility, particularly among underserved populations, and faces fundamental challenges with efficacy, validity, and compliance.

Acceptability: As a Driver of Digital Health Access Inequities

The acceptability of a digital health tool influences the decision to initially adopt and continuously engage with the digital tool, which is a precursor to receiving the intended health benefit [1]. Acceptability is interconnected with the approachability of a digital health tool by influencing one's perceptions of needing the tool as well as the willingness to share about the tool with others and influences one's perceptions of the appropriateness. Differences in culture, values, digital health literacy, numeracy, physical ability, visual acuity, hearing, and structural impediments influence an individual's acceptance and interactions with digital health tools [22]. Digital health tools that do not account for unique social and cultural factors, particularly among historically underserved groups, are often less relatable [25] and increase people's hesitance to use and benefit from them. This discrepancy in the acceptance of the digital health tool may be due to the failure of digital developers to incorporate the diverse views of end user perspectives, including the individual user and those who may deliver (eg, health care teams) the tool, in the conceptualization and design phases [26-28]. This results in tools that are unintentionally designed with features that are unappealing or inaccessible for the individuals and communities for which they are intended [29].

People who are visually impaired, color blind, older adults, and people from different races, ethnicities, sexual orientations, or abilities all have unique needs and preferences that should be considered [26]. Beliefs about the efficacy of the digital tool may influence initial acceptance and continued use [30]. For example, beliefs regarding the provider's ability to deliver quality care using telehealth were associated with fewer telehealth visits during the COVID-19 pandemic among African American individuals with diabetes [31], and overall, African American individuals have limited trust in the health benefits of digital health interventions [32]. Concerns with data privacy and security also reduce trust or willingness to use digital tools, especially among African American individuals [33].

A critical consideration is whether the care team also believes in the ability and security of the tools to improve care and patient outcomes, specifically for underserved populations [34]. The care team may be aware and hold beliefs about digital health tools that reduce their acceptability and impact the health care team members' and organizations' readiness and willingness to use a digital tool or recommend one to the patient. If a care team member does not share about a digital health tool, it may reduce a patient's awareness (approachability) of the tool. In addition, the continued lack of resources and support for use (eg, IT support and training) increases perceptions that digital tools disrupt workflows and delivery of care [35,36]. These fundamental challenges may compromise credibility among health care providers and executives and further impact the approachability and integration of digital health into the health care system.

Availability: As a Driver of Digital Health Inequities

Availability is the ability to reach and use digital health tools in a timely manner. Availability is largely dictated by an individual's access to devices (eg, smartphones and computers) and broadband internet. In 2021, 85% of Americans owned a smartphone, and 77% had broadband access, making smartphones and SMS text messaging intervention an appealing approach for widespread reach [37,38]. Access to broadband is so critical to health that it has been advocated for as a social determinant of health [39]. Although broadband access continues to rise, limited access has been associated with specific racial minority groups, lower education, lower income, living in rural areas, and the process of digital redlining (ie, the creation and perpetuation of inequities between marginalized groups through technology) [37,40]. Access to broadband internet also impacts how health interventions are delivered. For example, individuals who lack access to the internet at minimum connection speeds will typically be unable to access digital interventions that depend on features such as streaming and video calls (eg, Zoom).

At the organization and systems levels, enhanced access to services that support digital health is a concern. Health care facilities in underserved rural and urban communities have limited resources for implementing advanced digital functions, such as telehealth services [41]. This has increased disparities in access to mental health services among children in low-resourced areas where they are unable to access internet-based visits with providers [42]. Availability within a clinic or organization may also be limited by poor design and implementation planning, which leads to failure to integrate the digital health tool into the workflow and inadequate training health care teams to deliver digital health interventions routinely [43]. Lack of infrastructure and resources (eg, staff, IT support, internet access, and devices) within clinics has been cited as a continued barrier to using digital health tools despite their rapid

XSL•FO
RenderX

growth [35,36]. Furthermore, academic researchers often rely on time-limited grants to develop digital health tools, and these grants provide limited financing for the long-term operation, hindering the scalability and availability of digital health tools beyond the pilot stage of the intervention.

Affordability: As a Driver of Digital Health Access Inequities

Affordability is the economic capacity for people to spend resources and time to use appropriate services [14]. In digital health, this may be costs for the individual user (eg, patient), staff delivering care, and organization providing the digital health intervention (eg, clinic and hospital system). Costs may include obtaining digital devices or applications (eg, physical activity trackers, blood pressure monitors, or smartphones, and apps); connectivity for internet data transfer; and geolocation, which may be particularly prohibitive for individuals or organizations in underserved communities [29,44-46].

At the clinic level, disparities exist in the type of informatics systems afforded (eg, electronic health record [EHR] systems) and if and how new technologies can be easily adapted. For example, the digital infrastructure in community clinics may be less adaptable than the systems used by academic medical centers (eg, lack of skip logic, display logic in data collection forms, and lack of ability to place a referral from a questionnaire in the EHR). These system design challenges hamper health care by diminishing the process of collecting information, in particular, on patient social and behavioral determinants of health and making referral to appropriate services. Inequities may also stem from disparate costs during the development of digital health tools, as the necessary inclusion of end users in the design process of digital health solutions is time-intensive and costly [47]. In addition, consideration of the costs of maintaining digital health tools, the necessary digital infrastructure, and technical support is necessary for equitable, effective, and efficient digital care.

The unique approach to affordability by Levesque et al [14] provides insight into costs not only as a mechanism of payment for a service or tool but also as indirect costs (eg, time to use the tool) and opportunity costs (eg, time needed to learn how to use a new digital health tool). In particular, populations with low digital literacy (eg, older generations) may require more time and support to learn how to use digital tools, and this increases the opportunity costs and limits their ability to access digital health tools at the same rate as other users [48,49].

Appropriateness: As a Driver of Digital Health Access Inequities

Appropriateness refers to the ability of digital health tools to adequately address an individual's needs. When a tool is appropriate, individuals remain engaged and experience the benefits. Personal factors (eg, digital health literacy, language, stigma, and age) and technological factors (eg, not adapted for different abilities, confusing interface, and lack of structure) drive inequities in engagement and effectiveness that impact the ability of a digital health tool to meet the needs of the user. The rapid digitization of health care has implications for health care providers who are serving these diverse populations, including older adults who are more likely to have low digital literacy and may not be as able to benefit from these technologies [3]. Furthermore, low levels of digital literacy among older adults increases loneliness and reduces health seeking behaviors and quality of life [50,51]. As an indicator of appropriateness, a systematic review on the usability of digital health interventions for anxiety, depression, and somatoform disorders found that participants' initial beliefs and the amount of support and personalization that they received influenced engagement and dropout [52].

Appropriateness can also be conceptualized at a population level by evaluating who is accessing and benefiting and who may not be benefiting [32]. Digital health studies often lack both the measurement of social and digital determinants of health and the representation of people from underserved communities that would enable research evaluations on appropriateness and equitable effectiveness [20]. By being excluded from intervention efficacy and effectiveness studies [53-55], underserved populations experience health data poverty [56] or the inability for some individuals or groups to benefit from innovation due to lack of representation. In addition, data-driven interventions (eg, machine learning techniques) built from homogenous samples risk being ineffective or harmful to groups not included in the initial design process and could increase the digital divide [57]. Without understanding who is accessing and benefiting, developers cannot design digital tools to improve health and well-being for everyone and risk creating biased or racist technology [58]. Further exacerbating these issues are the systems-level issues such as racial disparities in research funding for behavioral medicine [59] and a lack of scholar representation from underserved populations [60,61] that impact the equitability of research evaluation and ultimately impact the appropriateness of digital health interventions.

Proposed Strategies to Improve Equitable Access

Overview

To address these drivers of inequitable access, we propose a strategic model (Figure 2) for advancing digital health access to foster health equity. The digital health model, adapted from the ConNECT framework [1] and health care access conceptual framework by Levesque et al [14], considers health inequities in access (far left), implementation stages and actors (middle), and subsequent influences on health behaviors and more equitable health outcomes (far right). Strategies to improve access are presented and summarized in Multimedia Appendix 1 [26,27,29,35,49,56,62-85]. These include design strategies followed by delivery, dissemination, and sustainability strategies to address digital health access inequities. Consideration of subsequent phases of delivery, dissemination, and sustainability during the design phase (ie, designing with the end in mind) can support sustained impact on health behaviors and health equity [86]. Considerations of evaluation and ethics should be applied across all stages to further ensure equitable access. While many levels of influence exist, we focus primarily on the developers, delivery agents, and systems as key actors for proposed strategies. We consider developers to include

XSL•FO
RenderX

academics, industry, and governmental or nongovernmental organizations that develop digital health solutions in collaboration with end users. Delivery agents include individuals who provide or deliver the solution to the end user (eg, clinicians and health care team members) and their organizations (eg, clinics, hospitals, broadband suppliers, and other organizations). Systems include the policies and funding structures in which users and suppliers are embedded.

Figure 2. A model for advancing digital health access to foster health equity adapted from the ConNECT framework and the health care access conceptual framework by Levesque et al [14]. The model considers health inequities in access, digital health implementation stages and actors, and the subsequent impact on sustained health behaviors and health outcomes and equity.

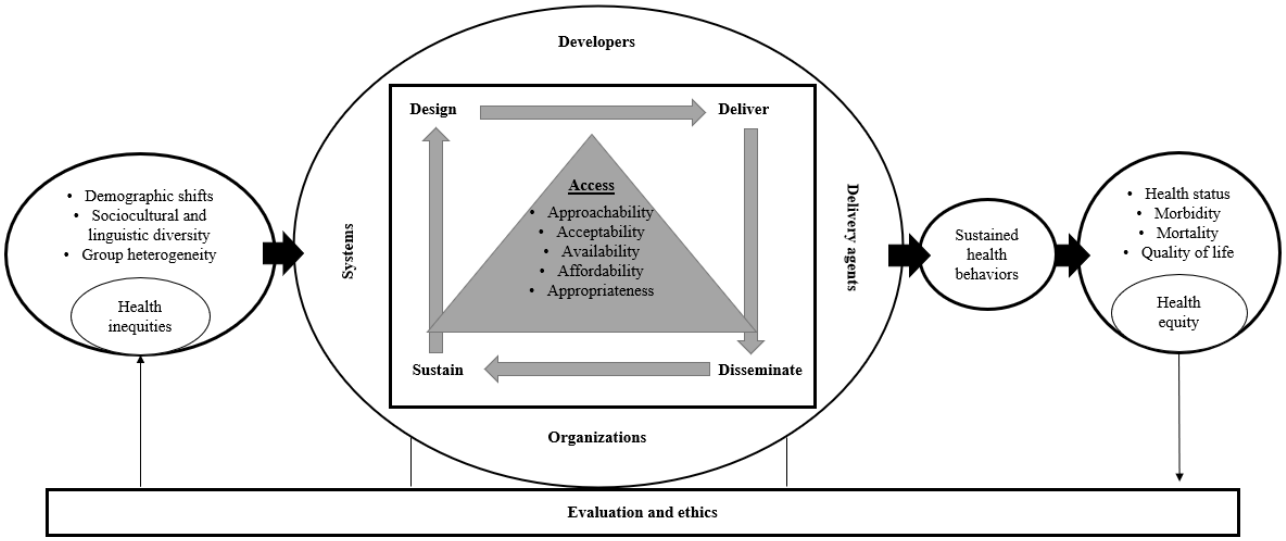

Design Strategies

Digital health tools designed using a theoretical basis, using behavior change techniques and in response to an end user's needs and desires (eg, health issues perceived as critical) and background (eg, social identities, digital and health literacy, language, and stigma) can improve their acceptability, approachability, and appropriateness [87]. Developers improve access by determining and enlisting the input of their customers and target users, whether the tool serves a broad population or offers more personalized support to people with specific social identities (eg, health apps for different races, ethnic, or age groups) [88]. Design approaches that strive for a one-size-fits-all end product may not address the unique challenges faced by different populations, ultimately reducing use and impact of the digital tool.

Customer discovery and value proposition design, a form of stakeholder engagement based on marketing and LeanStartup business methods [89], can be used to understand the problem and articulate the product's hypothesized unique value proposition relative to alternative options available to end users. This approach aims to identify a need or clinical problem requiring a solution, articulate a strong value proposition, and identify outcome measures that are meaningful to users to demonstrate efficacy and achieve buy-in [63,65,90]. Customer discovery has been applied to sustaining health informatics innovation and is foundational to the Innovation-Corps program for academic entrepreneurs developed by the National Science Foundation and adapted for health researchers by the National Center for Advancing Translational Sciences [63,91].

User-centered or participatory design methods may promote equitable access by designing tools that are relevant for different lived experiences and provide the information that individuals want, delivered how they like to receive it, and share it in a way that is secure and appeals to them in a user interface, resulting in a tool that is useful and accessible [92-95]. However, this approach is only successful when the intended users are broadly represented and engaged in the design process, underlying data, and product evaluation. Ensuring the intended audience is well-represented within design sessions and communicating who the tool was designed for can be challenging due to the intersectionality and diversity of individuals. Remaining cognizant of historical drivers of individual- and population-level inequities, including digital and health literacy and numeracy [48]; linguistic barriers [49]; and the uptake of a digital intervention for individuals with physical or cognitive limitations, including those more common among older adults age (eg, visual impairment), throughout design (and subsequent phases of implementation) is critical [96].

Equity of digital health tools is improved by undergoing usability testing by diverse, underserved individuals, including those of different generations, leveraging user experience methods, such as heuristic evaluations (does the tool meet the 10 general principles for user experience design and provide a usable experience?) and unmoderated tests (what is working and not working about the experience?). User testing may reveal technological aspects that can improve engagement from underserved groups, such as a simple interface, succinct content, reminders, feedback on progress, and acknowledgment of achievements [97]. The expansion of funding models to adequately support the described iterative design processes is critical to developing a product with the potential to be effective and equitable for underserved populations.

XSL•FO
RenderX

Delivery Strategies

Intentional implementation and sustainability plans codeveloped with the end users are critical to the integration of these solutions into routine practice to make them an integral, consistent part of patients' experience in health care. Ross et al [35] provide a nice example of using a systematic approach to implementation planning and execution. We urge others to share their implementation plan, including specific strategies, the plan development process, and adaptations made during delivery to generate a shared learning of the best approaches to implementing digital health solutions. The Expert Recommendations for Implementing Change provides discrete implementation strategies that are widely applicable and may be specified for digital health implementation in clinical care [98-100].

Ethnography may be used to observe and generate workflows that clarify the "patient journey" and points in the care visit where and by whom digital health tools should be used [68]. In addition to workflows or process maps, implementation plans may include strategies to address device ownership, broadband access, and digital health literacy screening as part of routine practice [101]. Understanding the ability of individuals to use a digital tool and the number and types of devices available in a household and whether those devices have sufficient bandwidth to support access such as telehealth conferencing is critical. Continued reimbursement for telehealth visits is necessary for successful implementation [102]. Offering digital health interventions in various formats (eg, virtual, real-time web-based help, in-person, and home assistance programs) may improve the ability to engage diverse, underserved populations [102].

Adequate training and ongoing technical support are needed for the end user to engage and benefit from the digital health tool successfully. Some clinics use a community-based navigator who could be trained to explain digital solutions to patients; connect patients to low-cost desktop, tablet, and mobile equipment and broadband or mobile data via local partnerships and federal subsidies; and provide ongoing support. At a systems-level, health care organizations should partner with community colleges, public libraries, and other community-based organizations to develop education and skill-building programs to address digital literacy gaps among underserved populations. Capacity (eg, staffing and time) of health care and community-based organizations and funding may be a challenge to institute and sustain these types of partnerships.

Dissemination Strategies

Active and tailored dissemination of digital health tools to the target audience through determined channels and planned strategies will increase awareness (ie, approachability) and access to digital health solutions [103]. Communicating about digital health using audience segmentation, that is, the process of dividing your audience into segments and matching your communication channels and strategies to meet their preferences and needs, is necessary to increase awareness among historically underserved populations (eg, African American individuals and people living in rural communities) [104].

Forging a partnership between academics and industry is a unique strategy for academics to outsource the commercialization mindset and use strategies to improve widespread access. Using a flexible business and revenue model beyond the organization claiming intellectual property rights may reach more organizations and individuals [105]. One dissemination strategy may be to develop a platform, such as the one developed in Germany [76], where evidence-based digital health tools can be easily found. Criteria for defining what classifies as evidence-based, generating buy-in from health care providers, and maintaining such platform for universal use may present challenges to implementing this strategy [76]. Another crucial strategy is to support the early adopters of these innovations and to make their use of and benefit from the digital health tool observable. Strategies of crowdsourcing innovators and citizen science platforms support the sharing of tools that seek to solve scientific challenges that are important to the user [106-109]. Furthermore, involving health care workers from underserved populations as trusted messengers of digital health solutions may also increase reach and adoption among diverse groups [110]. Finally, and perhaps most importantly, it is necessary to create and foster an organizational culture of tolerance and patience for change and adoption for new innovations within health care institutions because diffusing innovations from early adopters to other segments in society requires significant time and energy from all stakeholders [111,112].

Sustainability Strategies

Addressing disparities in affordability is particularly critical to sustaining access to digital health solutions. When scaling up an effective digital health solution, a sustainable funding plan must be developed from the outset to support long-term growth, which may require private-public partnerships. Government subsidies for broadband subscriptions and data charges for mobile health applications may help reduce individual costs. Additionally, broadband access can be expanded through community infrastructure, such as broadband hot spots in public spaces, including libraries and community centers [102]. McCall et al [40] recently outlined multilevel strategies to combat digital redlining (eg, intentional lack of investment in broadband) and expand broadband access and quality to all communities [113,114].

At the policy-level, regulations are needed that ensure internet service providers build infrastructure that meets the minimum requirements for high-speed internet, as defined by the Federal Communications Commission [40]. Renewed funding of government subsidies and programs such as the Regional Health Information Technology Extension Centers, which successfully promoted the adoption of EHRs in community clinics nationwide up until 2015, should be renewed to support clinics in underserved communities [77,115,116]. As the COVID-19 pandemic has highlighted, offering telehealth services can increase access to care and support underserved populations. Continued reimbursement for telehealth visits by both state Medicaid programs and the Centers for Medicare and Medicaid Services is critical to sustain improved access.

XSL•FO
RenderX

Overarching Considerations: Evaluation and Ethics

Evaluating effectiveness of digital health interventions and sharing the results among underserved target populations is critical to improving approachability of digital health tools among these groups. Process evaluation that includes examination of metrics, such as acceptability, is critical throughout design, implementation, dissemination, and sustainment phases. Examination of acceptability, for example, may provide insights into reasons for low use or effectiveness and implications for the fidelity of both delivery and receipt of the intervention [117]. Evaluation ensures solutions are inclusive, equitable, and appropriate to address a health problem and to solve critical gaps in health care delivery. Implementation science and community-based participatory research literature can provide established approaches to consider in evaluation of digital health tools [118-120]. If digital health equity is to be realized, developers need a strong research focus with expanded funding mechanisms and support for implementing and evaluating innovative designs and analyses (eg, stepped-wedge designs, sequential multiple assignment randomized trial designs, and pragmatic randomized controlled trials) that simultaneously consider effectiveness and implementation (eg, engagement and integration in workflow) [43]. Digital health companies are growing rapidly and should include research teams to use these methodologies to validate the effectiveness of these technologies on health and health equity outcomes. The need for collaboration by researchers and digital health companies has been acknowledged through funding mechanisms, such as the National Institutes for Health (NIH) Small Business Innovation Research and Small Business Technology Transfer, that support research or business collaboration. These NIH funding mechanisms provide great opportunities for digital health solutions created in the private sector to benefit from the research expertise of clinical scientists and evaluate their solutions early in the process.

When considering the appropriateness of digital health tools for achieving the intended impacts across different demographic groups, we must also reduce health data poverty [56] by being advocates for populations who lack data, incorporating deliberative research and citizens' juries (ie, inclusion of the community in decision-making), and communicating with people on how their data are being used and protected. It is critically important to create data sets that are representative of marginalized and underrepresented populations by using community-based participatory research methods and encouraging data sharing initiatives across digital health tools [20,56]. Justifying exclusion criteria should be provided to ensure that we are engaging diverse participants in digital behavioral health interventions and evaluation [46]. In addition, to address structural issues, funders, regulators, and policy makers should require digital health interventions and tools to perform appropriately and be usable for different populations (especially historically underrepresented groups) and settings [56]. In addition, and to act on multiple levers of change within the society, funders may also create separate funding mechanisms to specifically support the development or adaptation of digital health tools tailored for historically marginalized groups. In case successful digital health tools were already created for the mainstream population, these mechanisms can ensure that separate digital health tools with similar missions are available for marginalized groups with specific needs.

All tools should use ethical principles and theories in design, including confidentiality, inclusivity, and transparency [84]. To increase trust in digital health tools, there needs to be greater transparency in transferring and sharing data collected using behavioral digital technology [44] while maintaining privacy and security and returning results to patients or participants [27,29,84]. Core elements of privacy to ensure when using personal and remote individual digital monitoring include safety and security in digital services for data collection, transfer, storage, management, and sharing [21,62,84,121]. The Connected and Open Research Ethics (CORE) tool may support researchers and institutional review boards with data privacy, transfer, and transparency to support digital health research [122]. CORE may be used to find or share informational resources (eg, institutional review board and consent documents) or receive feedback from experts. Using approaches, such as CORE, to improve confidentiality and transparency are particularly important when working with historically marginalized populations who have faced ostracization, discrimination, and even physical danger from privacy violations. For example, during the HIV and AIDs epidemic, lesbian, gay, bisexual, transgender, and queer individuals were fearful that their HIV and AIDs status would be disclosed and that they would face stigma at work or risk losing their job. More recently, individuals who menstruate have expressed concerns regarding data privacy and security when using period-tracking apps and the potential harmful consequences of these data being misused in a post–Roe v. Wade era. Outside health care, data privacy has implications for discriminatory advertising, racially biased policing, and the outing or surveillance of historically marginalized populations. Intellectual property, policies, and data governance structures must be addressed to safeguard personal behavioral data and rights that adhere to policy and law, especially for historically marginalized populations [44,84].

Discussion: Focusing Future Solutions for Equitable Digital Health Access

Overview

The "promises and perils" of digital tools to achieve health equity have been highlighted for over a decade [121,123], but success in achieving equitable access to health technologies remains challenging. Our paper addresses renewed concerns regarding equity in digital health access that were deepened during the COVID-19 pandemic [3-6]. Using behavioral, equity, and access frameworks allows for a unique and comprehensive exploration of current drivers of digital health inequities. It allowed a compilation of strategies that have the potential for actionable impact on digital health equity.

We propose a model that emphasizes using multilevel strategies across stages of design, delivery, dissemination, and sustainment to advance digital health access and foster health equity.

XSL•FO
RenderX

Strategies were shaped with a behavioral medicine focus, as the field has a unique role in improving digital health access; arguably, all digital tools require the user (individual, provider, and health system) to change behavior by engaging with the technology to generate impact. Furthermore, digital health tools targeting health behaviors (eg, physical activity and mental health apps) have recently exploded [2]. Finally, a breadth of behavioral research offers critical evidence to draw upon from individual to population effects, providing a myriad of behavior change tools and strategies that have been shown to be effective, to have broad reach, and to have the potential to be widely disseminated and sustained.

Multilevel factors drive unequal access, so strategies require action from tool developers, individual delivery agents, organizations, and systems to effect change. We outline key strategies at multiple stages, albeit not exhaustive, to promote digital health equity to achieve the intended health outcomes and consequences (Multimedia Appendix 1). Multilevel solutions are highlighted for key producers of digital approaches and recommendations to engage diverse end users to develop representative, acceptable, and appropriate tools for the intended audiences. Once designed, digital health tools demonstrating usability and accessibility require further evaluation to address dissemination and implementation challenges at organizational and systems levels [20]. Successful and sustainable digital health approaches are supported by multifaceted strategies, culturally and linguistically appropriate methods, tailoring, establishment of mutual partnerships, and community engagement.

Implicit in our model is the potential for multidisciplinary and team-based approaches in applying recommended strategies. Given the important role of industry in creating new digital tools, exploring new models of academic, community, and industry partnerships is recommended to work upstream in addressing improved access [124-126]. While academics and industry approach development and evaluation differently [2], synergy could be found in blending their strengths of evaluation and innovation. Academics focus on design, intervention, and efficacy testing, infusing behavioral theory and evidence before disseminating to end users for input. They often start with a targeted product that is personalized to a specific population or social identity and, even more often, is constrained by funding that requires this sequence of development. Conversely, industry develops a product for a broad audience and immediately disseminates it, iterating the functionality with user feedback and eventually examining effectiveness (if at all). The resulting products may be either effective but not widely used or engaging but not meaningfully effective to result in positive health outcomes. While the goal should be effectiveness, there are opportunities to design digital tools to improve the health and well-being for both everyone and for specific groups. This mixed approach is critical as different health providers and systems serve different populations, and the goal of health equity is to meet the specific needs of their population. While a combination of tools for both broad and specific populations is likely needed to achieve health equity, using ethical principles of transparency to state for whom the tool is designed and is proven to benefit and perhaps, more importantly, for whom the tool may not benefit is important to ensure adoption, use, and dissemination to the right individuals to achieve positive health impacts.

Team science practices and experience can shape new ideas about forming digital health partnerships [2]. To begin, academics could consider hiring experienced commercial developers and designers on their team to infuse usability principles and integrate efforts toward sustainability from the outset of product development, and industry companies could consult with academic scientists or build an in-house team of experienced clinical researchers to integrate theory-based interventions and evidence-based practice. Behavioral science expertise could inform considerations of several issues, such as transparency and fairness in application software; how big data and self-monitoring may regulate or constrain behavior; how to obtain user consent for data use, especially by employers or insurance companies; and comparisons of in-person to digitally mediated health care [84]. Developing meaningful collaborations with scholars from diverse underserved communities who are health equity experts is critical in addition to initiatives to transform institutional culture through programs such as NIH Build and the National Research Mentoring Network [127]. Additionally, partnerships with community colleges, health departments, public libraries, and other community-based organizations are necessary for a bidirectional flow of information and resources that may support common goals of digital health access across systems, individuals, and national or state policies [128]. These relationships can bridge multiple levels [129] and consider and account for the interconnectedness and bidirectional nature of movement within and across internal (clinic or organization level) and external factors that shape digital health access.

Limitations

While using behavioral, equity, and access frameworks allowed for a unique and comprehensive exploration of current drivers of digital health inequities and a compilation of strategies, a systematic literature search was not conducted. While the intent was to provide an overview of barriers, not using systematic methods may have limited our ability to generate content that is entirely comprehensive. This paper used a narrative search driven by the discussion and comparison of frameworks and drawn from the expertise of our multidisciplinary team. The authors are mainly behavioral scientists, and some are clinical providers; while they were culturally diverse, they may not reflect all perspectives across all demographic and cultural groups. This may limit or bias the perspectives of our report. To reduce biases, we were intentional about drawing from the literature and do acknowledge that this is a viewpoint paper that reflects some positionality of the authors.

Conclusions

Health equity is an ethical imperative that digital health equity can contribute to achieving. Digital health tools have the potential to reduce health inequities when equitable access is achieved. Applying behavioral science perspectives to better understand the multilevel factors driving unequal access can contribute to solving existing and emerging health disparities. Addressing concerns that threaten the widespread benefit from effective and validated digital health tools to society and further

XSL•FO
RenderX

prevent benefit for historically marginalized communities needs urgent attention to bring these solutions to those most in need. Transdisciplinary teams that use this integrated digital health equity model (Figure 2) and strategies across design, delivery, dissemination, and sustainability stages may significantly improve digital health access and health equity.

Acknowledgments

This group was formed from the Digital Equity and Implementation working group of the Society for Behavioral Medicine. The authors would like to acknowledge the support of the Behavioral Informatics and Technology and Healthy Equity special interests groups. MMK received funding that partially supported this project from the National Heart Lung and Blood Institute of the National Institutes of Health (K01HL167993) and the Centers for Disease Control and Prevention (U48DP006395). MMK and LMK were partially supported by the National Cancer Institute (P50CA244431). LAF was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health (K01MD017630). NMA is the Canada Research Chair (tier 2) in Behavioural Health Intervention, Department of Psychology at Concordia University, and this work was funded, in part, by funds provided via the Canada Research Chair Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

All authors contributed to the writing and editing of the manuscript. MMK and LMK led the conceptualization of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Strategies for developers, delivery agents, and systems to improve access and promote digital health equity.
[DOCX File , 26 KB - jmir_v26i1e51355_app1.docx]

References

1. Alcaraz KI, Sly J, Ashing K, Fleisher L, Gil-Rivas V, Ford S, et al. The ConNECT Framework: a model for advancing behavioral medicine science and practice to foster health equity. J Behav Med 2017 Feb 10;40(1):23-38 [FREE Full text] [doi: 10.1007/s10865-016-9780-4] [Medline: 27509892]
2. Pagoto S, Bennett GG. How behavioral science can advance digital health. Transl Behav Med 2013 Sep 17;3(3):271-276 [FREE Full text] [doi: 10.1007/s13142-013-0234-z] [Medline: 24073178]
3. Chunara R, Zhao Y, Chen J, Lawrence K, Testa PA, Nov O, et al. Telemedicine and healthcare disparities: a cohort study in a large healthcare system in New York City during COVID-19. J Am Med Inform Assoc 2021 Jan 15;28(1):33-41. [doi: 10.1093/jamia/ocaa217] [Medline: 32866264]
4. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020 Jun 2;22(6):e19361. [doi: 10.2196/19361] [Medline: 32452816]
5. Dixit N, Van Sebille Y, Crawford GB, Ginex PK, Ortega PF, Chan RJ. Disparities in telehealth use: how should the supportive care community respond? Support Care Cancer 2022 Feb;30(2):1007-1010 [FREE Full text] [doi: 10.1007/s00520-021-06629-4] [Medline: 34668075]
6. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440 [FREE Full text] [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
7. What is digital health? U.S. Food & Drug Administration. 2020 Sep 22. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2024-07-04]
8. Mohammed HT, Hyseni L, Bui V, Gerritsen B, Fuller K, Sung J, et al. Exploring the use and challenges of implementing virtual visits during COVID-19 in primary care and lessons for sustained use. PLoS One 2021 Jun 24;16(6):e0253665 [FREE Full text] [doi: 10.1371/journal.pone.0253665] [Medline: 34166441]
9. van Dijk J, Hacker K. The digital divide as a complex and dynamic phenomenon. Inf Soc 2003 Sep;19(4):315-326. [doi: 10.1080/01972240309487]
10. Bernhardt JM. Health education and the digital divide: building bridges and filling chasms. Health Educ Res 2000 Oct;15(5):527-531. [doi: 10.1093/her/15.5.527] [Medline: 11184212]
11. Hall AK, Bernhardt JM, Dodd V, Vollrath MW. The digital health divide: evaluating online health information access and use among older adults. Health Educ Behav 2015 Apr 25;42(2):202-209 [FREE Full text] [doi: 10.1177/1090198114547815] [Medline: 25156311]
12. Taylor CB, Fitzsimmons-Craft EE, Graham AK. Digital technology can revolutionize mental health services delivery: the COVID-19 crisis as a catalyst for change. Int J Eat Disord 2020 Jul 25;53(7):1155-1157 [FREE Full text] [doi: 10.1002/eat.23300] [Medline: 32449523]

XSL•FO
RenderX

13. Torous J, Wykes T. Opportunities from the coronavirus disease 2019 pandemic for transforming psychiatric care with telehealth. JAMA Psychiatry 2020 Dec 01;77(12):1205-1206. [doi: 10.1001/jamapsychiatry.2020.1640] [Medline: 32391857]
14. Levesque JF, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int J Equity Health 2013 Mar 11;12(1):18 [FREE Full text] [doi: 10.1186/1475-9276-12-18] [Medline: 23496984]
15. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 01;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
16. Bouchard LC, Yanez B, Dahn JR, Flury SC, Perry KT, Mohr DC, et al. Brief report of a tablet-delivered psychosocial intervention for men with advanced prostate cancer: acceptability and efficacy by race. Transl Behav Med 2019 Jul 16;9(4):629-637 [FREE Full text] [doi: 10.1093/tbm/iby089] [Medline: 30285186]
17. Verbiest M, Corrigan C, Dalhousie S, Firestone R, Funaki T, Goodwin D, et al. Using codesign to develop a culturally tailored, behavior change mHealth intervention for indigenous and other priority communities: a case study in New Zealand. Transl Behav Med 2019 Jul 16;9(4):720-736. [doi: 10.1093/tbm/iby093] [Medline: 30388262]
18. Hurtado-de-Mendoza A, Serrano A, Zhu Q, Graves K, Fernández N, Fernández A, et al. Engaging Latina breast cancer survivors in research: building a social network research registry. Transl Behav Med 2018 Jul 17;8(4):565-574 [FREE Full text] [doi: 10.1093/tbm/ibx053] [Medline: 30016517]
19. Renfro T, Johnson E, Lambert DN, Wingood G, DiClemente RJ. The MEDIA model: an innovative method for digitizing and training community members to facilitate an HIV prevention intervention. Transl Behav Med 2018 Nov 21;8(6):815-823. [doi: 10.1093/tbm/iby012] [Medline: 29471535]
20. Lyles CR, Wachter RM, Sarkar U. Focusing on digital health equity. JAMA 2021 Nov 09;326(18):1795-1796. [doi: 10.1001/jama.2021.18459] [Medline: 34677577]
21. Shen FX, Silverman BC, Monette P, Kimble S, Rauch SL, Baker JT. An ethics checklist for digital health research in psychiatry: viewpoint. J Med Internet Res 2022 Feb 09;24(2):e31146 [FREE Full text] [doi: 10.2196/31146] [Medline: 35138261]
22. Cu A, Meister S, Lefebvre B, Ridde V. Assessing healthcare access using the Levesque's conceptual framework- a scoping review. Int J Equity Health 2021 May 07;20(1):116 [FREE Full text] [doi: 10.1186/s12939-021-01416-3] [Medline: 33962627]
23. Alsan M, Stanford FC, Banerjee A, Breza E, Chandrasekhar AG, Eichmeyer S, et al. Comparison of knowledge and information-seeking behavior after general COVID-19 public health messages and messages tailored for Black and Latinx communities. Ann Intern Med 2021 Apr;174(4):484-492. [doi: 10.7326/m20-6141]
24. Evans WD, Thomas CN, Favatas D, Smyser J, Briggs J. Digital segmentation of priority populations in public health. Health Educ Behav 2019 Dec 19;46(2_suppl):81-89 [FREE Full text] [doi: 10.1177/1090198119871246] [Medline: 31742454]
25. Kwasnicka D, Keller J, Perski O, Potthoff S, Ten Hoor GA, Ainsworth B, et al. White paper: open digital health - accelerating transparent and scalable health promotion and treatment. Health Psychol Rev 2022 Dec 14;16(4):475-491 [FREE Full text] [doi: 10.1080/17437199.2022.2046482] [Medline: 35240931]
26. Bhattacharyya O, Mossman K, Gustafsson L, Schneider EC. Using human-centered design to build a digital health advisor for patients with complex needs: persona and prototype development. J Med Internet Res 2019 May 09;21(5):e10318 [FREE Full text] [doi: 10.2196/10318] [Medline: 31094334]
27. Njie-Carr VP, Jones-Parker H, Massey C, Baker D, Nganga-Good C. Leveraging community engagement to develop a mobile health application for older women with HIV infection. J Obstet Gynecol Neonatal Nurs 2018 Nov;47(6):833-843. [doi: 10.1016/j.jogn.2018.08.005] [Medline: 30278158]
28. Wray TB, Kahler CW, Simpanen EM, Operario D. User-centered, interaction design research approaches to inform the development of health risk behavior intervention technologies. Internet Interv 2019 Mar;15:1-9 [FREE Full text] [doi: 10.1016/j.invent.2018.10.002] [Medline: 30425932]
29. Njie-Carr VP, Zhu S, Williams GC, Corless IB, Himelhoch S. Evaluation of a technology-enhanced intervention for older women with HIV infection: a proof of concept study. AIDS Care 2021 Aug 24;33(8):983-992. [doi: 10.1080/09540121.2020.1810617] [Medline: 32835493]
30. Svendsen MJ, Nicholl BI, Mair FS, Wood K, Rasmussen CD, Stochkendahl MJ. One size does not fit all: participants' experiences of the selfBACK app to support self-management of low back pain-a qualitative interview study. Chiropr Man Therap 2022 Oct 03;30(1):41 [FREE Full text] [doi: 10.1186/s12998-022-00452-2] [Medline: 36192724]
31. Rovner BW, Casten RJ, Chang AM, Hollander JE, Rising K. Mistrust, neighborhood deprivation, and telehealth use in African Americans with diabetes. Popul Health Manag 2021 Dec 01;24(6):699-700 [FREE Full text] [doi: 10.1089/pop.2021.0094] [Medline: 34000211]
32. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]
33. Lai S, Tanner B. Examining the intersection of data privacy and civil rights. Brookings Institution. 2022 Jul 18. URL: https://www.brookings.edu/articles/examining-the-intersection-of-data-privacy-and-civil-rights/ [accessed 2023-05-30]

XSL•FO
RenderX

34. Bondaronek P, Dicken SJ, Singh Jennings S, Mallion V, Stefanidou C. Barriers to and facilitators of the use of digital tools in primary care to deliver physical activity advice: semistructured interviews and thematic analysis. JMIR Hum Factors 2022 Aug 30;9(3):e35070 [FREE Full text] [doi: 10.2196/35070] [Medline: 36040764]
35. Ross J, Stevenson F, Dack C, Pal K, May C, Michie S, et al. Developing an implementation strategy for a digital health intervention: an example in routine healthcare. BMC Health Serv Res 2018 Oct 19;18(1):794 [FREE Full text] [doi: 10.1186/s12913-018-3615-7] [Medline: 30340639]
36. Whitelaw S, Pellegrini DM, Mamas MA, Cowie M, Van Spall HG. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021 Mar;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
37. Internet, broadband fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2024-07-04]
38. Willcox JC, Dobson R, Whittaker R. Old-fashioned technology in the era of "bling": is there a future for text messaging in health care? J Med Internet Res 2019 Dec 20;21(12):e16630 [FREE Full text] [doi: 10.2196/16630] [Medline: 31859678]
39. Benda NC, Veinot TC, Sieck CJ, Ancker JS. Broadband internet access is a social determinant of health!. Am J Public Health 2020 Aug;110(8):1123-1125. [doi: 10.2105/AJPH.2020.305784] [Medline: 32639914]
40. McCall T, Asuzu K, Oladele CR, Leung TI, Wang KH. A socio-ecological approach to addressing digital redlining in the United States: a call to action for health equity. Front Digit Health 2022 Jul 18;4:897250 [FREE Full text] [doi: 10.3389/fdgth.2022.897250] [Medline: 35924138]
41. Eccles MP, Mittman BS. Welcome to implementation science. Implement Sci 2006 Feb 22;1:1. [doi: 10.1186/1748-5908-1-1]
42. Smith-East M, Starks S. COVID-19 and mental health care delivery: a digital divide exists for youth with inadequate access to the internet. J Am Acad Child Adolesc Psychiatry 2021 Jul;60(7):798-800. [doi: 10.1016/j.jaac.2021.04.006] [Medline: 33910040]
43. Palacholla RS, Fischer N, Coleman A, Agboola S, Kirley K, Felsted J, et al. Provider- and patient-related barriers to and facilitators of digital health technology adoption for hypertension management: scoping review. JMIR Cardio 2019 Mar 26;3(1):e11951 [FREE Full text] [doi: 10.2196/11951] [Medline: 31758771]
44. Brall C, Schröder-Bäck P, Maeckelberghe E. Ethical aspects of digital health from a justice point of view. Eur J Public Health 2019 Oct 01;29(Supplement_3):18-22 [FREE Full text] [doi: 10.1093/eurpub/ckz167] [Medline: 31738439]
45. McCall HC, Hadjistavropoulos HD, Loutzenhiser L. Reconsidering the ethics of exclusion criteria in research on digital mental health interventions. Ethics Behav 2019 Oct 31;31(3):171-180. [doi: 10.1080/10508422.2019.1684295]
46. Digital equity in health services: practice brief. National Collaborating Centre for Determinants of Health and Nova Scotia Health Authority. 2022. URL: https://nccdh.ca/resources/entry/digital-equity-in-health-services [accessed 2024-07-04]
47. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
48. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019 Oct 01;29(Supplement_3):13-17 [FREE Full text] [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
49. Rodriguez JA, Casillas A, Cook BL, Marlin RP. The language of equity in digital health: prioritizing the needs of limited English proficient communities in the patient portal 2.0. J Health Care Poor Underserved 2021;32(2S):211-219. [doi: 10.1353/hpu.2021.0059]
50. Li Q, Yang C, Zhao Z, Yang C, Chen Z, Huang D, et al. The relationship between internet use and loneliness of middle-aged and older adult people: the moderating effect of residence. Front Public Health 2024 Jan 31;12:1284180 [FREE Full text] [doi: 10.3389/fpubh.2024.1284180] [Medline: 38356943]
51. Ngiam NH, Yee WQ, Teo N, Yow KS, Soundararajan A, Lim JX, et al. Building digital literacy in older adults of low socioeconomic status in Singapore (project wire up): nonrandomized controlled trial. J Med Internet Res 2022 Dec 02;24(12):e40341 [FREE Full text] [doi: 10.2196/40341] [Medline: 36459398]
52. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 06;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
53. Chastain DB, Osae SP, Henao-Martínez AF, Franco-Paredes C, Chastain JS, Young HN. Racial disproportionality in Covid clinical trials. N Engl J Med 2020 Aug 27;383:e59. [doi: 10.1056/nejmp2021971]
54. Loree JM, Anand S, Dasari A, Unger JM, Gothwal A, Ellis LM, et al. Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. JAMA Oncol 2019 Oct 01;5(10):e191870 [FREE Full text] [doi: 10.1001/jamaoncol.2019.1870] [Medline: 31415071]
55. Polo AJ, Makol BA, Castro AS, Colón-Quintana N, Wagstaff AE, Guo S. Diversity in randomized clinical trials of depression: a 36-year review. Clin Psychol Rev 2019 Feb;67:22-35. [doi: 10.1016/j.cpr.2018.09.004] [Medline: 30292439]
56. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit Health 2021 Apr;3(4):e260-e265. [doi: 10.1016/s2589-7500(20)30317-4]

XSL•FO
RenderX

57. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 04;169(12):866. [doi: 10.7326/m18-1990]
58. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
59. Patel V. Mental health research funding: too little, too inequitable, too skewed. Lancet Psychiatry 2021 Mar;8(3):171-172. [doi: 10.1016/s2215-0366(20)30471-5]
60. Ginther DK, Kahn S, Schaffer WT. Gender, race/ethnicity, and National Institutes of Health R01 research awards: is there evidence of a double bind for women of color? Acad Med 2016 Aug;91(8):1098-1107 [FREE Full text] [doi: 10.1097/ACM.0000000000001278] [Medline: 27306969]
61. The advisory committee to the director working group on diversity: racism in science report. National Institutes of Health. 2021 Feb 14. URL: https://acd.od.nih.gov/documents/presentations/02262021DiversityReport.pdf [accessed 2024-07-04]
62. Chang LW, Njie-Carr V, Kalenge S, Kelly JF, Bollinger RC, Alamo-Talisuna S. Perceptions and acceptability of mHealth interventions for improving patient care at a community-based HIV/AIDS clinic in Uganda: a mixed methods study. AIDS Care 2013 May 08;25(7):874-880 [FREE Full text] [doi: 10.1080/09540121.2013.774315] [Medline: 23452084]
63. Morrato EH, Hamer MK, Sills M, Kwan B, Schilling LM. Applying a commercialization-readiness framework to optimize value for achieving sustainability of an electronic health data research network and its data capabilities: the SAFTINet experience. EGEMS (Wash DC) 2019 Aug 29;7(1):48 [FREE Full text] [doi: 10.5334/egems.295] [Medline: 31523697]
64. Gibson B, Kramer H, Weir C, Fiol G, Borbolla D, Schlechter CR, et al. Workflow analysis for design of an electronic health record-based tobacco cessation intervention in community health centers. JAMIA Open 2021 Jul;4(3):ooaa070 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa070] [Medline: 34514352]
65. Kepper MM, Walsh-Bailey C, Brownson RC, Kwan BM, Morrato EH, Garbutt J, et al. Development of a health information technology tool for behavior change to address obesity and prevent chronic disease among adolescents: designing for dissemination and sustainment using the ORBIT model. Front Digit Health 2021 Mar 10;3:648777 [FREE Full text] [doi: 10.3389/fdgth.2021.648777] [Medline: 34713122]
66. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019 Aug 15;9(8):e029954 [FREE Full text] [doi: 10.1136/bmjopen-2019-029954] [Medline: 31420394]
67. Burgermaster M, Rosenthal M, Tierney WM, Altillo BS, Nordquist E, Enriquez C, et al. Nutri: a behavioral science-based clinical decision support for chronic disease management. AMIA Annu Symp Proc 2022;2022:299-308 [FREE Full text] [Medline: 37128464]
68. Goodson L, Vassar M. An overview of ethnography in healthcare and medical education research. J Educ Eval Health Prof 2011 Apr 25;8:4 [FREE Full text] [doi: 10.3352/jeehp.2011.8.4] [Medline: 21637319]
69. Jin KY, Reichert F, Cagasan LPJ, de la Torre J, Law N. Measuring digital literacy across three age cohorts: exploring test dimensionality and performance differences. Comput Educ 2020 Nov;157:103968. [doi: 10.1016/j.compedu.2020.103968]
70. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021 Feb 03;23(2):e26145 [FREE Full text] [doi: 10.2196/26145] [Medline: 33533727]
71. Blanc J, Hahn K, Oliveira B, Phillips R, Duthely LM, Francois L, et al. Bringing health care equity to diverse and underserved populations in sleep medicine and research through a digital health equity framework. Sleep Med Clin 2023 Sep;18(3):255-267 [FREE Full text] [doi: 10.1016/j.jsmc.2023.05.009] [Medline: 37532367]
72. Vaughan EM, Cardenas VJJ, Chan W, Amspoker AB, Johnston CA, Virani SS, et al. Implementation and evaluation of a mHealth-based community health worker feedback loop for Hispanics with and at risk for diabetes. J Gen Intern Med 2024 Feb 06;39(2):229-238. [doi: 10.1007/s11606-023-08434-7] [Medline: 37803098]
73. Njie-Carr V, Kalengé S, Kelley J, Wilson A, Muliira JK, Nabirye RC, et al. Research capacity-building program for clinicians and staff at a community-based HIV clinic in Uganda: a pre/post evaluation. J Assoc Nurses AIDS Care 2012;23(5):431-441 [FREE Full text] [doi: 10.1016/j.jana.2011.11.002] [Medline: 22265671]
74. Kusters IS, Dean JM, Gutierrez AM, Sommer M, Klyueva A. Assessment of COVID-19 website communication in languages other than English by local health departments in the United States. Health Commun 2023 Jul 29;38(8):1519-1529. [doi: 10.1080/10410236.2021.2017109] [Medline: 34965822]
75. Barbieri M, Andreoni G. mHealth market exploitation through the analysis of the related intellectual property rights. In: Andreoni G, Perego P, Frumento E, editors. m_Health Current and Future Applications. Cham, Switzerland: Springer; 2019.
76. Wangler J, Jansky M. Two years of approved digital health applications in Germany - perspectives and experiences of general practitioners with an affinity for their use. Eur J Gen Pract 2023 Dec 15;29(1):2186396 [FREE Full text] [doi: 10.1080/13814788.2023.2186396] [Medline: 36919621]
77. Connecting Americans to health care. Federal Communications Commission. URL: https://www.fcc.gov/connecting-americans-health-care [accessed 2024-07-04]
78. Phan P, Mitragotri S, Zhao Z. Digital therapeutics in the clinic. Bioeng Transl Med 2023 Jul 03;8(4):e10536 [FREE Full text] [doi: 10.1002/btm2.10536] [Medline: 37476062]

XSL•FO
RenderX

79. Brewer LC, Asiedu GB, Jones C, Richard M, Erickson J, Weis J, et al. Emergency preparedness and risk communication among African American churches: leveraging a community-based participatory research partnership COVID-19 initiative. Prev Chronic Dis 2020 Dec 10;17:E158 [FREE Full text] [doi: 10.5888/pcd17.200408] [Medline: 33301390]
80. Ho E, Jeon M, Lee M, Luo J, Pfammatter AF, Shetty V, et al. Fostering interdisciplinary collaboration: a longitudinal social network analysis of the NIH mHealth Training Institutes. J Clin Transl Sci 2021 Sep 20;5(1):e191 [FREE Full text] [doi: 10.1017/cts.2021.859] [Medline: 34849265]
81. Baptist A, Gibson-Scipio W, Carcone AI, Ghosh S, Jacques-Tiura AJ, Hall A, et al. Asthma and technology in emerging African American adults (the ATHENA project): protocol for a trial using the multiphase optimization strategy framework. JMIR Res Protoc 2022 May 10;11(5):e37946 [FREE Full text] [doi: 10.2196/37946] [Medline: 35536642]
82. Burgermaster M, Desai PM, Heitkemper EM, Juul F, Mitchell EG, Turchioe M, et al. Who needs what (features) when? Personalizing engagement with data-driven self-management to improve health equity. J Biomed Inform 2023 Aug;144:104419. [doi: 10.1016/j.jbi.2023.104419] [Medline: 37301528]
83. Health Data Research UK homepage. Health Data Research UK. URL: https://www.hdruk.ac.uk/ [accessed 2024-05-02]
84. Shaw JA, Donia J. The sociotechnical ethics of digital health: a critique and extension of approaches from bioethics. Front Digit Health 2021 Sep 23;3:725088 [FREE Full text] [doi: 10.3389/fdgth.2021.725088] [Medline: 34713196]
85. Mental health technology advisory committee. American Psychological Association. URL: https://www.apa.org/practice/mental-health-techology-advisory-committee [accessed 2024-05-02]
86. Kwan BM, Brownson RC, Glasgow RE, Morrato EH, Luke DA. Designing for dissemination and sustainability to promote equitable impacts on health. Annu Rev Public Health 2022 Apr 05;43(1):331-353 [FREE Full text] [doi: 10.1146/annurev-publhealth-052220-112457] [Medline: 34982585]
87. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
88. Samarasekera U. The rise of racial minority health apps. Lancet Digit Health 2022 Apr;4(4):e218-e219. [doi: 10.1016/s2589-7500(22)00043-7]
89. Why the lean start-up changes everything. Harvard Business Review. 2013. URL: https://hbr.org/2013/05/why-the-lean-start-up-changes-everything [accessed 2024-07-04]
90. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015 Oct;34(10):971-982 [FREE Full text] [doi: 10.1037/hea0000161] [Medline: 25642841]
91. Morrato EH, Lennox LA, Sendro ER, Schuster AL, Pincus HA, Humensky J, et al. Scale-up of the Accrual to Clinical Trials (ACT) network across the Clinical and Translational Science Award Consortium: a mixed-methods evaluation of the first 18 months. J Clin Transl Sci 2020 Jun 30;4(6):515-528 [FREE Full text] [doi: 10.1017/cts.2020.505] [Medline: 33948228]
92. Labrique A, Vasudevan L, Weiss W, Wilson K. Establishing standards to evaluate the impact of integrating digital health into health systems. Glob Health Sci Pract 2018 Oct 10;6(Supplement 1):S5-17. [doi: 10.9745/ghsp-d-18-00230]
93. Bamgboje-Ayodele A, Smith A, Short CE, Fardell JE, Shaw J, Beatty L, et al. Barriers and facilitators to the availability of efficacious self-directed digital health tools for adults living with cancer and their caregivers: a systematic literature review and author survey study. Patient Educ Couns 2021 Oct;104(10):2480-2489. [doi: 10.1016/j.pec.2021.03.012] [Medline: 33741232]
94. Higgins KS, Tutelman PR, Chambers CT, Witteman HO, Barwick M, Corkum P, et al. Availability of researcher-led eHealth tools for pain assessment and management: barriers, facilitators, costs, and design. Pain Rep 2018 Sep;3(Suppl 1):e686 [FREE Full text] [doi: 10.1097/PR9.0000000000000686] [Medline: 30324177]
95. Young AS. Overcoming the barriers to dissemination and implementation. In: Wulfovich S, Meyers A, editors. Digital Health Entrepreneurship. Cham, Switzerland: Springer; Jun 21, 2019.
96. Henni SH, Maurud S, Fuglerud KS, Moen A. The experiences, needs and barriers of people with impairments related to usability and accessibility of digital health solutions, levels of involvement in the design process and strategies for participatory and universal design: a scoping review. BMC Public Health 2022 Jan 06;22(1):35 [FREE Full text] [doi: 10.1186/s12889-021-12393-1] [Medline: 34991540]
97. van der Vaart R, van Driel D, Pronk K, Paulussen S, Te Boekhorst S, Rosmalen JG, et al. The role of age, education, and digital health literacy in the usability of internet-based cognitive behavioral therapy for chronic pain: mixed methods study. JMIR Form Res 2019 Nov 21;3(4):e12883 [FREE Full text] [doi: 10.2196/12883] [Medline: 31750839]
98. Graham AK, Lattie EG, Powell BJ, Lyon AR, Smith JD, Schueller SM, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020 Nov;75(8):1080-1092 [FREE Full text] [doi: 10.1037/amp0000686] [Medline: 33252946]
99. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the expert recommendations for implementing change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]

XSL•FO
RenderX

100. Perry CK, Damschroder LJ, Hemler JR, Woodson TT, Ono SS, Cohen DJ. Specifying and comparing implementation strategies across seven large implementation interventions: a practical application of theory. Implement Sci 2019 Mar 21;14(1):32 [FREE Full text] [doi: 10.1186/s13012-019-0876-4] [Medline: 30898133]
101. Haggerty JL, Levesque JF. Development of a measure of health care affordability applicable in a publicly funded universal health care system. Can J Public Health 2015 Feb 04;106(2):e66-e71 [FREE Full text] [doi: 10.17269/cjph.106.4562] [Medline: 25955674]
102. Eruchalu CN, Pichardo MS, Bharadwaj M, Rodriguez CB, Rodriguez JA, Bergmark RW, et al. The expanding digital divide: digital health access inequities during the COVID-19 pandemic in New York City. J Urban Health 2021 Apr;98(2):183-186 [FREE Full text] [doi: 10.1007/s11524-020-00508-9] [Medline: 33471281]
103. Proctor EK, Colditz GA, Brownson RC. Dissemination and Implementation Research in Health: Translating Science to Practice. New York, NY: Oxford University Press; 2012.
104. Brownson RC, Eyler AA, Harris JK, Moore JB, Tabak RG. Getting the word out: new approaches for disseminating public health science. J Public Health Manag Pract 2018;24(2):102-111 [FREE Full text] [doi: 10.1097/PHH.0000000000000673] [Medline: 28885319]
105. Hill C, Martin JL, Thomson S, Scott-Ram N, Penfold H, Creswell C. Navigating the challenges of digital health innovation: considerations and solutions in developing online and smartphone-application-based interventions for mental health disorders. Br J Psychiatry 2017 Aug;211(2):65-69. [doi: 10.1192/bjp.bp.115.180372] [Medline: 28522435]
106. Lemmens R, Antoniou V, Hummer P, Potsiou C. Citizen science in the digital world of apps. In: Vohland K, editor. The Science of Citizen Science. Cham, Switzerland: Springer; Jan 12, 2021.
107. Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, et al. Crowdsourcing--harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med 2014 Jan;29(1):187-203 [FREE Full text] [doi: 10.1007/s11606-013-2536-8] [Medline: 23843021]
108. Saez-Rodriguez J, Costello JC, Friend SH, Kellen MR, Mangravite L, Meyer P, et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat Rev Genet 2016 Jul 15;17(8):470-486 [FREE Full text] [doi: 10.1038/nrg.2016.69] [Medline: 27418159]
109. Rumsfeld JS, Brooks SC, Aufderheide TP, Leary M, Bradley SM, Nkonde-Price C, et al. Use of mobile devices, social media, and crowdsourcing as digital strategies to improve emergency cardiovascular care: a scientific statement from the American Heart Association. Circulation 2016 Aug 23;134(8):e87-108. [doi: 10.1161/CIR.0000000000000428] [Medline: 27334603]
110. Yee V, Bajaj SS, Stanford FC. Paradox of telemedicine: building or neglecting trust and equity. Lancet Digit Health 2022 Jul;4(7):e480-e481. [doi: 10.1016/s2589-7500(22)00100-5]
111. Berwick DM. Disseminating innovations in health care. JAMA 2003 Apr 16;289(15):1969-1975. [doi: 10.1001/jama.289.15.1969] [Medline: 12697800]
112. Rogers EM. Diffusion of Innovations, 4th Edition. New York, NY: The Free Press; 2010.
113. Ankawi B, Clark J, Walsh EA, Wesolowicz DM. No patient left behind: ensuring we continue overcoming barriers in the age of telehealth. Society of Behavioral Medicine. 2022. URL: https://www.sbm.org/publications/outlook/issues/fall-2022/no-patient-left-behind-ensuring-we-continue-overcoming-barriers-in-the-age-of-telehealth/full-article [accessed 2024-07-04]
114. Ford S, Buscemi J, Laitner M, Hirko K, Newton RJ, Jonassaint C, et al. Position statement: retain and enhance efforts to increase broadband internet availability for health care access in rural areas. Society of Behavioral Medicine. 2018. URL: https://www.sbm.org/UserFiles/file/RuralBroad_State18_v2.pdf [accessed 2024-07-04]
115. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
116. Apathy NC, Holmgren AJ, Adler-Milstein J. A decade post-HITECH: critical access hospitals have electronic health records but struggle to keep up with other advanced functions. J Am Med Inform Assoc 2021 Aug 13;28(9):1947-1954 [FREE Full text] [doi: 10.1093/jamia/ocab102] [Medline: 34198342]
117. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
118. Chapter 36. Introduction to evaluation. Community Tool Box. URL: https://ctb.ku.edu/en/table-of-contents/evaluate/evaluation [accessed 2024-07-04]
119. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
120. Shelton RC, Chambers DA, Glasgow RE. An extension of RE-AIM to enhance sustainability: addressing dynamic context and promoting health equity over time. Front Public Health 2020 May 12;8:134 [FREE Full text] [doi: 10.3389/fpubh.2020.00134] [Medline: 32478025]
121. Institute of Medicine. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington, DC: The National Academies Press; 2011.

XSL•FO
RenderX

122. Torous J, Nebeker C. Navigating ethics in the digital age: introducing connected and open research ethics (CORE), a tool for researchers and institutional review boards. J Med Internet Res 2017 Feb 08;19(2):e38 [FREE Full text] [doi: 10.2196/jmir.6793] [Medline: 28179216]
123. Anderson KM, Olson S. The Promises and Perils of Digital Strategies in Achieving Health Equity: Workshop Summary. Washington, DC: National Academies Press; 2016.
124. Partnership at the digital health frontier: accelerating access to healthcare through public-private collaboration. U.S. Chamber of Commerce. 2023 Jun 21. URL: https://tinyurl.com/addc9f8h [accessed 2024-07-04]
125. Ford KL, Portz JD, Zhou S, Gornail S, Moore SL, Zhang X, et al. Benefits, facilitators, and recommendations for digital health academic-industry collaboration: a mini review. Front Digit Health 2021 Apr 15;3:616278 [FREE Full text] [doi: 10.3389/fdgth.2021.616278] [Medline: 34713094]
126. MacLeod J, Abdelrahim M, Painter S, Maddula R, Steward A, Hamid A, et al. Ten step academic-industry digital health collaboration methodology: a case-based guide for digital health research teams with the example of cardio-oncology. Front Cardiovasc Med 2022 Sep 30;9:982059 [FREE Full text] [doi: 10.3389/fcvm.2022.982059] [Medline: 36247469]
127. Diversity Program Consortium. URL: https://diversityprogramconsortium.org/pages/ [accessed 2024-07-04]
128. Gustavson AM, Lewinski AA, Fitzsimmons-Craft EE, Coronado GD, Linke SE, O'Malley DM, et al. Strategies to bridge equitable implementation of telehealth. Interact J Med Res 2023 May 15;12:e40358 [FREE Full text] [doi: 10.2196/40358] [Medline: 37184909]
129. Moullin JC, Dickson KS, Stadnick NA, Rabin B, Aarons GA. Systematic review of the exploration, preparation, implementation, sustainment (EPIS) framework. Implement Sci 2019 Jan 05;14(1):1 [FREE Full text] [doi: 10.1186/s13012-018-0842-6] [Medline: 30611302]

Abbreviations

CORE: Connected and Open Research Ethics
EHR: electronic health record
NIH: National Institutes for Health

Edited by T de Azevedo Cardoso; submitted 28.07.23; peer-reviewed by E Vashishtha, H Veldandi, E Blazek; comments to author 26.02.24; revised version received 02.05.24; accepted 27.05.24; published 01.08.24.

Please cite as:
Kepper MM, Fowler LA, Kusters IS, Davis JW, Baqer M, Sagui-Henson S, Xiao Y, Tarfa A, Yi JC, Gibson B, Heron KE, Alberts NM, Burgermaster M, Njie-Carr VPS, Klesges LM
Expanding a Behavioral View on Digital Health Access: Drivers and Strategies to Promote Equity
J Med Internet Res 2024;26:e51355
URL: https://www.jmir.org/2024/1/e51355
doi:10.2196/51355
PMID:39088246

©Maura M Kepper, Lauren A Fowler, Isabelle S Kusters, Jean W Davis, Manal Baqer, Sara Sagui-Henson, Yunyu Xiao, Adati Tarfa, Jean C Yi, Bryan Gibson, Kristin E Heron, Nicole M Alberts, Marissa Burgermaster, Veronica PS Njie-Carr, Lisa M Klesges. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

5G Key Technologies for Helicopter Aviation Medical Rescue

Wei Han Sr[1], MD; Yuanting Li 2nd[2], BA; Changgen Chen 3rd[3], MS; Danni Huang[4], BS; Junchao Wang[5], PhD; Xiang Li[6], BA; Zhongliang Ji[1], PhD; Qin Li[1], PhD; Zhuang Li[1], BA

[1]Emergency Department, Shenzhen University General Hospital, Shenzhen, China
[2]Guangdong Yitong Software, Guangzhou, China
[3]Information and Support Division, Shenzhen Mindray Biomedical Electronics Co, Shenzhen, China
[4]Shenzhen branch of Guangdong Co, Ltd, China Mobile Communications Group, Shenzhen, China
[5]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China
[6]Technical Department, Shenzhen Eastern General Aviation Co, Ltd, Shenzhen, China

Corresponding Author:
Wei Han Sr, MD
Emergency Department
Shenzhen University General Hospital
1098 Xueyuan Avenue, Nanshan District
Shenzhen, 518055
China
Phone: 86 18813980300
Email: sugh_hanwei@szu.edu.cn

Abstract

Rapid global population growth and urbanization have heightened the demand for emergency medical rescue, with helicopter medical rescue emerging as an effective solution. The advent of 5G communication technology, characterized by large bandwidth, low latency, and high reliability, offers substantial promise in enhancing the efficiency and quality of helicopter rescue operations. However, the full integration of 5G technology into helicopter emergency medical services is still in its nascent stages and requires further development. In this viewpoint, we present our experience from the Shenzhen University General Hospital of the application of 5G low-altitude network communication technology, body area network disease sensing technology, and 5G air-ground collaborative rapid diagnosis and treatment technology in aeromedical rescue. We consider that the 5G air-to-ground collaborative rapid diagnosis and treatment technology enables high-quality remote consultation, enhancing emergency medical rescue and providing strong support for future rescue operations.

(J Med Internet Res 2024;26:e50355) doi:10.2196/50355

KEYWORDS

low airspace; helicopters; medical aid; 5G technology; aeronautical engineering

Introduction

Helicopter medical rescue has become an effective option for emergency medical assistance, as populations and urbanization increase [1]. Such rescues must quickly and accurately determine the target location, cooperate closely with ground medical personnel, maintain stability during high-speed flight, and achieve real-time data transmission and efficient medical treatment [2,3]. To meet these needs, effective communication technology is essential. As one of the most advanced mobile communication technologies at present, 5G technology has advantages such as large bandwidth, low latency, and high reliability. For medical rescue, 5G has many advantages [4]. Medical experts can guide on-site medical personnel to provide treatment through remote video and other means, thereby improving the efficiency and quality of medical rescue [5,6].

The development of 5G technology started in 2018, and the first 5G network was put into use in South Korea. Since then, 5G technology has rapidly developed globally. At present, many countries, including the United States, Japan, South Korea, and various countries in Europe, have begun commercial deployment of 5G networks. Applications of 5G technology are also being explored. For example, in the United States, 5G technology has been widely applied in fields such as intelligent transportation and intelligent manufacturing. In Europe, 5G technology is applied in industrial automation and robot control [7]. In Asia, South Korea and Japan are exploring 5G technology in intelligent manufacturing and intelligent health care. China has

XSL•FO
RenderX

also led the world in the research and applications of 5G technology [8,9]. Since the end of 2019, China has built the world's largest 5G network, and at present, 31 provinces and cities across the country have opened 5G commercial services. China has applied 5G technology to fields such as intelligent manufacturing, intelligent logistics, and smart cities; while in health care, it has been used for telemedicine and medical image transmission. Companies such as Huawei and ZTE are leading in the research and development of 5G technology. At the same time, the Chinese government is actively promoting its development and has formulated a series of policy measures to support the promotion and applications of 5G technology.

On August 7, 2022, the aviation rescue drill at Shenzhen University General Hospital successfully adopted 5G private network access [10]. We developed aviation medical rescue programs and services through a cooperation agreement with Shenzhen Eastern Navigation Co, Ltd, which provided essential support and guarantees. The work started with 4 aviation medical rescue drills held from September 2019 to December 2022 in order to lay the technical foundation. The types of diseases considered during these drills included multiple injuries to the whole body and spinal cord, and the scenarios included traffic accidents and tunnel scenes. From November 2019 to March 2023, seven aviation medical rescue operations were conducted covering multiple systemic fractures, spinal injuries, drowning pulmonary edema, multiple organ failure, hemopneumothorax, and other diseases. The types of rescue operations were joint mountain search and rescue, seaside drowning rescue, and traffic accident rescue, with a flight distance of 60-600 km and a flight time of 10-136 minutes. These operations involved the first aerial cardiopulmonary resuscitation (CPR), medical staff delivery, and emergency medical treatment. Among the rescue operations, the joint mountain search and rescue used highly complex technologies such as hovering, winches, and aerial lifting, providing significant support for treating patients [11].

This viewpoint paper explores 5G key technologies for helicopter aviation medical rescue to address the shortcomings of traditional air rescue methods in information transmission, positioning accuracy, real-time data transmission, and medical treatment efficiency. We demonstrate the potential of the applications of 5G technology in helicopter medical rescue in terms of improving rescue efficiency, shortening patient treatment time, and enhancing the collaborative combat ability of rescue teams.

Innovative Technology

Overview

There are 3 major problems for medical rescue operations in the air: difficulty in communication, difficulty in disease monitoring, and difficulty in collaborative diagnosis and treatment. To overcome these problems, our team put forward a series of important technological innovations such as developing airborne 5G network communication technology and achieving efficient and stable communication. The body area network disease perception technology provided precise real-time monitoring and 5G air-to-ground collaborative rapid diagnosis and treatment technology, which strengthened the collaborative capacity and efficiency of emergency response teams. Hence, these innovations brought unprecedented breakthroughs to the field of aviation medical rescue, significantly improving rescue effectiveness.

Low Airspace Medical Rescue 5G Network Technology

Low-Altitude 2.6+4.9-GHz Collaborative Networking

The low-altitude medical rescue helicopter provided side lobe coverage of a 2.6-GHz 5G down-tilt base station as a 5G private network access signal during takeoff and landing, with a true height less than 150 m. The main or side lobes of a 4.9-GHz 5G up-tilt base station provided 5G private network signal access during takeoff or landing and flight at a true height of 150 m or above. The collaborative coverage of 2 frequency bands of base stations provided seamless switching of 5G private network access for helicopter vertical takeoff and landing and horizontal flight. The network switching delay was ≤30 milliseconds (Figure 1).

XSL•FO
RenderX

Figure 1. Schematic diagram of 2.6+4.9-GHz dual-frequency collaborative 5G networking.

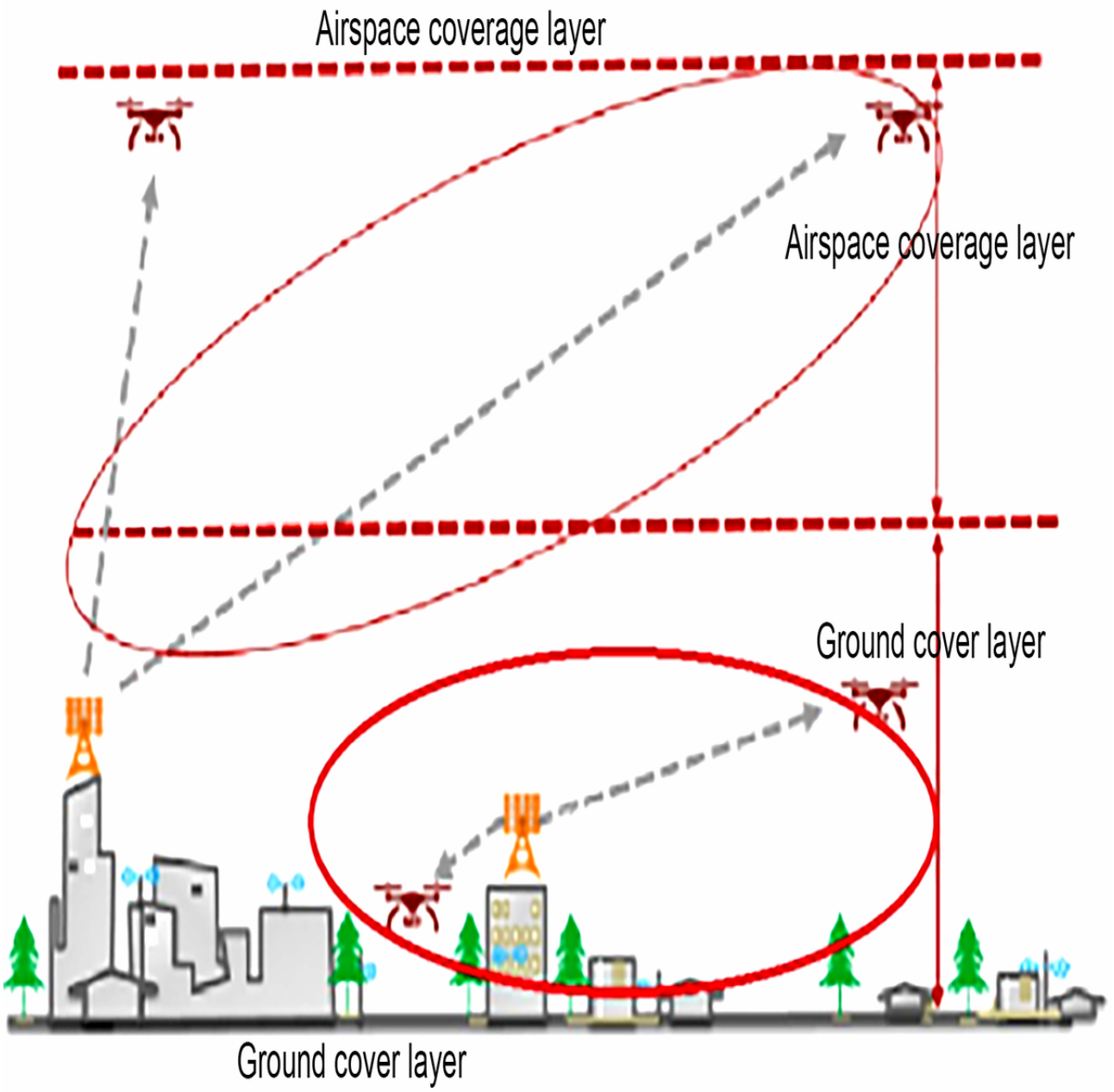

Adopting 5G+Global Navigation Satellite System External Suction Cup Antenna

The aircraft body mainly comprises aluminum, significantly attenuating the signal after blocking radio waves. Therefore, we designed an external pull-away suction cup antenna to dock with the onboard 5G router. The external suction cup antenna could be installed on the glass window side of the helicopter (Figure 2) and directionally receive 5G signals from the external airspace of the aircraft body through glass (window) materials with low signal attenuation.

XSL•FO
RenderX

Figure 2. Deployment image of the external 5G+global navigation satellite system antenna (the antenna was derived from the external suction cup antenna of Guanghetong BGS-910).

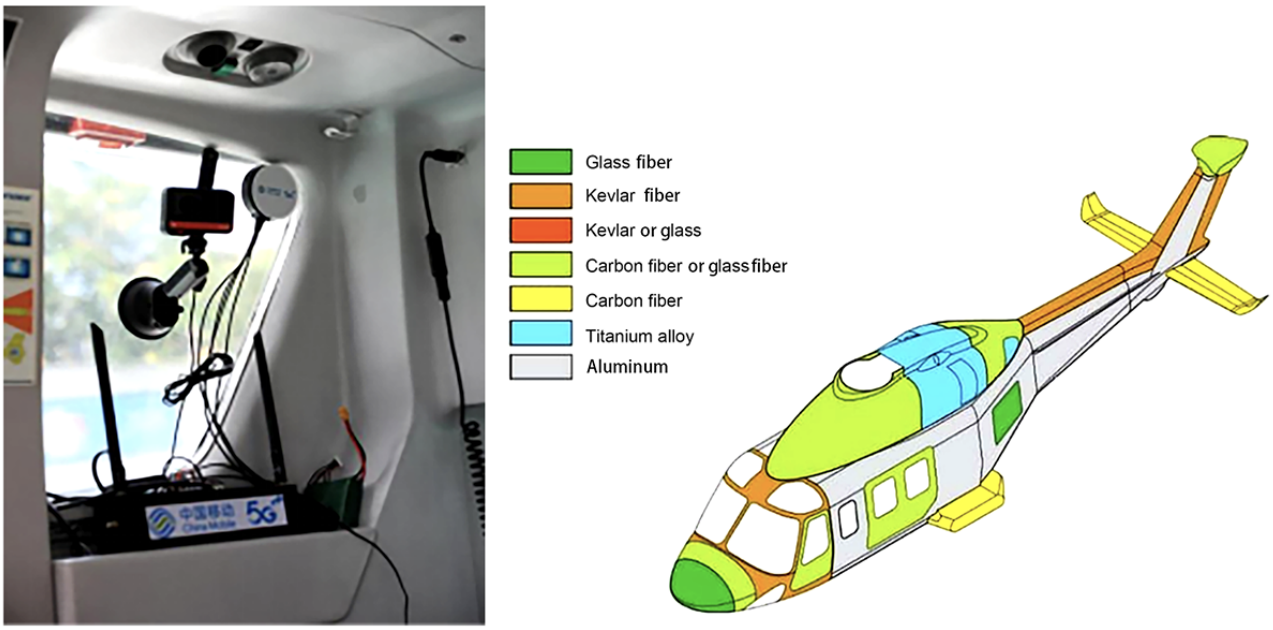

This installation was made on a domestic medical rescue helicopter, with the novel and successful deployment of 5G-Customer Premises Equipment. After testing, under this installation method, the attenuation difference of 5G signal inside and outside the body was only 8 dB, while 5G-CPE's built-in rubber rod antenna deployed inside the cabin, and the attenuation of 5G signal by the metal material of the aircraft body was 20 dB, which is a significant difference.

Collaboration of 5G+Beidou Positioning Technology

Next, we examined the Beidou positioning capability provided by the 5G-Customer Premises Equipment module (Figure 3). The data were transmitted back to the emergency platform deployed by the hospital via the 5G private network to locate the real-time flight position of the rescue helicopter.

Figure 3. Beidou positioning trajectory map obtained from 5G module.

The traditional positioning scheme for airspace helicopters generally uses a GPS for the helicopter's built-in flight-positioning module. It does not output positioning data to external systems. It only interfaces with local airspace management departments (regional civil aviation bureaus or air traffic control bureaus) for airspace flight management. For this scheme, we adopted the Guanghetong FM160 model 5G air rescue equipment (Fibocom; Figure 4) and used the global navigation satellite system capability of the 5G module and the antenna to provide global navigation satellite system reception,

XSL•FO
RenderX

obtaining real-time positioning information for low-altitude medical rescue, and the information output of real-time positioning of low-altitude medical rescue was realized.

Figure 4. 5G module–integrated 5G air rescue equipment with global navigation satellite system capability. The module was derived from Guanghetong FM160.

Air Body Area Network Disease Awareness Technology

Construction of a Body Area Network Transmission Scheme Based on Posture Feature and Channel Quality

Communication energy consumption is the key factor that restricts the wider applications of a network with limited energy resources. Human activities drastically affect near-body communication channels, leading to strong time-varying characteristics. This limits the performance of communication energy consumption optimization methods based on wireless sensor networks [12,13]. Therefore, designing an intelligent communication mechanism for body area networks should consider the strong time-varying characteristics of the channel so that the overall transmission energy efficiency can be improved. Our approach was to design a near-body channel estimation algorithm based on deep learning. Treating the time-frequency response of the pilot point as a low-resolution 2D image, using convolutional neural networks and memory networks to generate high-resolution images, and obtaining complete time-frequency response characteristics helped predict and evaluate the current channel quality (Figure 5). Further, we combined human dynamic channel quality estimation and prediction procedures and studied energy-efficient and adaptive access protocols and resource allocation schemes for improved service quality to achieve stable, flexible, and low-power transmission effects.

XSL•FO
RenderX

Figure 5. Detection of channel characteristics in body area networks. LMMSE: linear minimum mean square error; RCNN: region-based convolutional neural network.

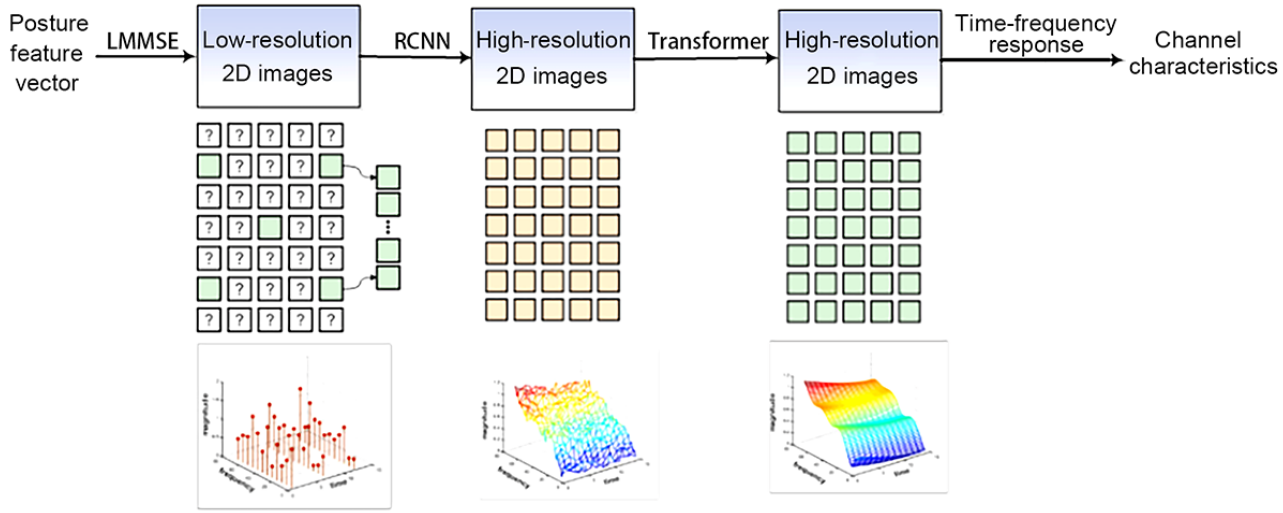

Proposing a Volume Domain Network Data Encryption Method Based on Physical Signal Features and Chaotic Mapping

The body area network involves sensitive data and hence requires a high-level security system. The limited sensor resources limit the applicability of traditional encryption schemes. As shown in Figure 6, we used the characteristics of physical signals as the initial chaotic values to design a high-strength and sensitive encryption scheme, providing asymmetric encryption with 1 key at a time. The main procedures followed in this study were as follows: (1) calculation of feature vectors for sign signals, precise classification of signal characteristics of physical signs using discrete wavelet transform, and design of multidimensional convolutional neural networks; (2) design and quantification of the algorithm for generating heterogeneous combination chaotic sequences, proposing power spectral entropy as the evaluation criterion, and designing a quantified heterogeneous combination chaotic sequence that met the security requirements of the body area network; and (3) implementation and verification of data encryption scheme, proposing a heterogeneous combination chaotic encryption scheme based on sign signal feature vectors and conducting hardware implementation and performance verification based on the field programmable gate array platform (Figure 6).

XSL•FO
RenderX

Figure 6. Body area network data encryption method.

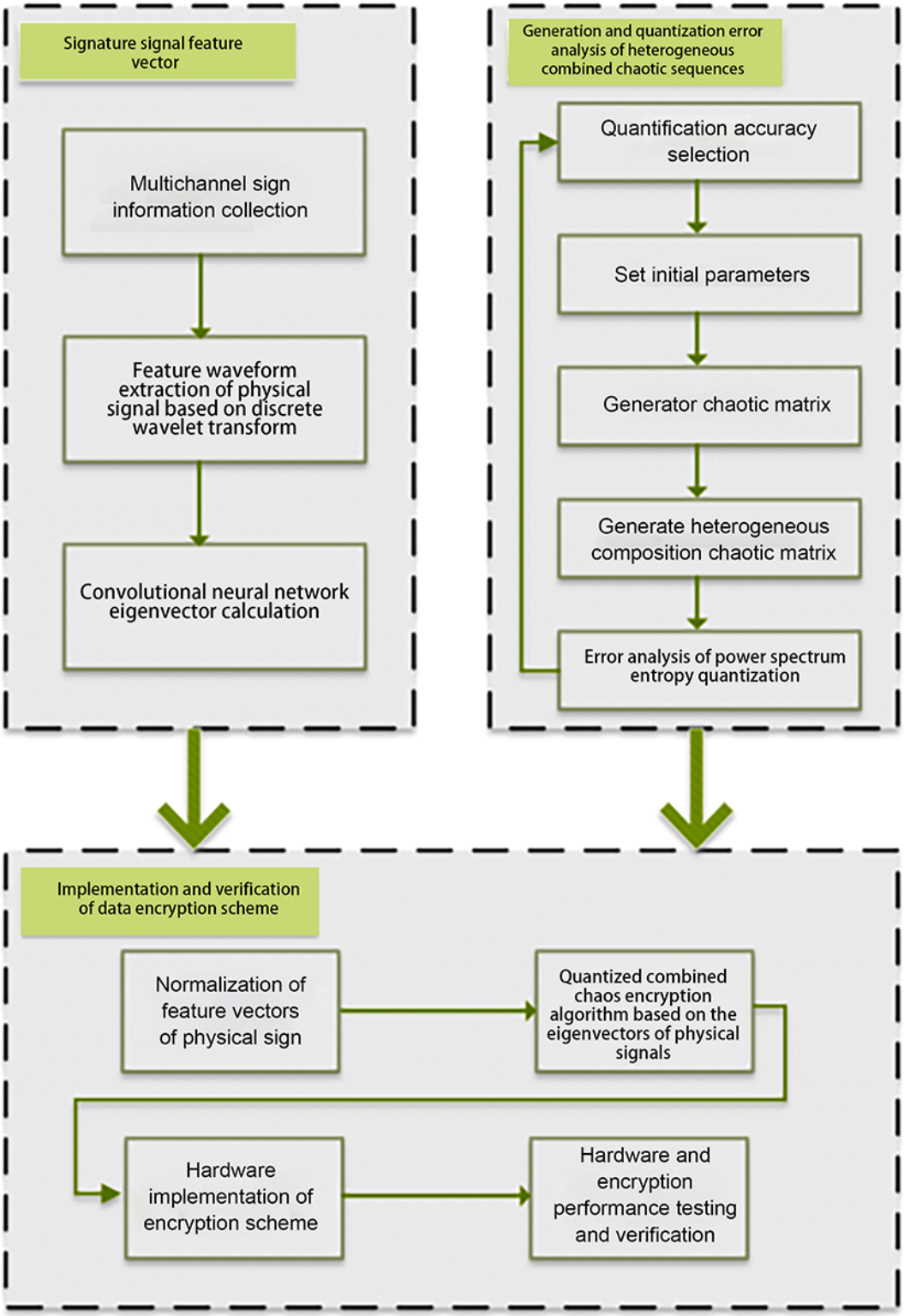

5G Air-to-Ground Collaborative Rapid Diagnosis and Treatment Technology

Development of a Plug-and-Play Highly Reliable 5G Emergency Equipment

The 5G module has high power consumption, insufficient USB power supply capacity for digital medical devices, and unreliable long-term continuous data transmission and communication through a USB module. (1) We developed a 5G single-slot module using the existing 12-V power supply and direct communication between network ports of digital medical equipment. Additionally, we used anti-interference and heat treatment technologies to achieve 5G standardization of devices such as monitors and ventilators. There was no need to change the software and hardware of the digital equipment and special installation. This provided us with “plug-and-play” 5G emergency equipment for use in hospitals (Figure 7). (2) We

XSL•FO
RenderX

established highly reliable 5G medical equipment. Our team developed the antenna technology in the form of multiple antennas; coupled devices, antipower technology, and anti-interference technology between equipment; and designed switching performance improvement, network disconnection alarm optimization, and alarm strategy under high-speed movement with highly stable software design technology and wireless parameter tuning. This antenna technology also included switching performance improvement under high-speed movement, network outage alarm optimization, alarm strategy, and other designs, with high reliability of 5G monitors, 5G transport monitors, 5G ventilators, 5G defibrillators, and other equipment. The packet loss rate of the data was within 0.5% at high speeds and edge locations with a reference signal received power of –95 dBm for extended periods, which was much higher in reliability compared with other wireless communication consumer products (packet loss rate of 3%-5%). The 5G rescue equipment developed in this study completed the first domestic air rescue and remote intensive care unit rescue application, playing an essential role in epidemic prevention and control (Figure 7).

Figure 7. 5G first-aid equipment.

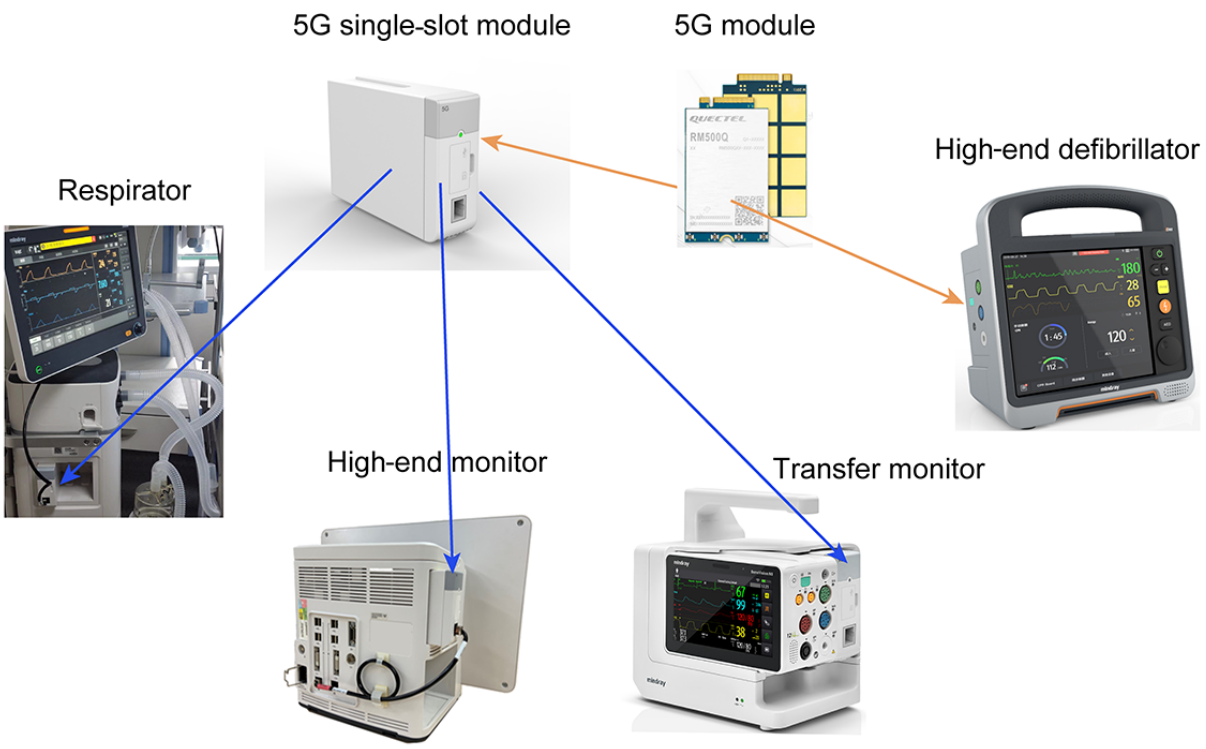

Establishing an Air-to-Ground Collaborative Information-Sharing Platform

A panoramic archive storage and retrieval system was developed for prehospital emergency patients. A panoramic archive of prehospital emergency patients was formed based on the patient health archive, combined with the vital signs collected by 5G emergency equipment, which helped achieve real-time data exchange with the hospital system. We focused on building a standard for emergency rescue data in sky-ground spaces, considering that the panoramic archive of prehospital emergency patients included emergency dispatch and diagnosis and treatment information to support comprehensive and complete access and transmission of cross-institutional rescue information. An air-to-ground emergency rescue information exchange and scheduling system was developed. We used microservice, distributed, and message service technologies to develop an application platform relying on the 5G private network and information security infrastructure, achieving information collaboration covering the entire chain of prehospital first-aid, interhospital transportation, and in-hospital treatment. These technologies supported information exchange and business collaborative applications among cross-regional, cross-institutional, and cross-departmental application systems. By using the Beidou module for satellite positioning and leveraging the high-speed, low-latency characteristics of 5G technology, we can support aviation transportation tracking, injury assessment and management, and remote real-time monitoring and diagnosis of vital signs. It also opened up an air green channel, significantly improving the overall rescue success rate. We developed a remote and multidisciplinary consultation system for sky-ground spaces, established a multichannel collection and transmission of vital signs, and transmitted the archival information and treatment images of patients who are critically ill and outside the hospital to the hospital's emergency expert group or multidisciplinary consultation on time. This provided technical guidance and full monitoring for the diagnosis and treatment of patients who are critically ill and outside the hospital and improved the efficiency of diagnosis and treatment (Figure 8).

XSL•FO
RenderX

Figure 8. Air-to-ground collaborative information-sharing platform. CT: computed tomography.

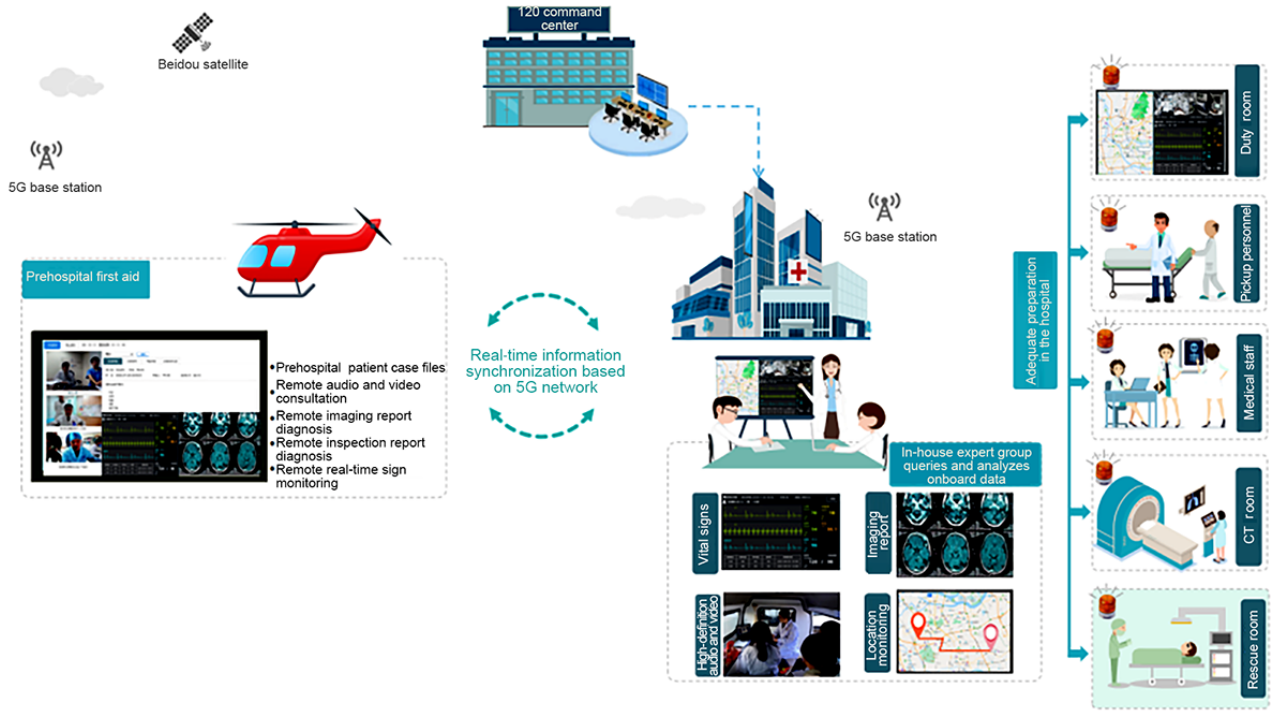

Our Direct Experience of the Helicopter Aviation Medical Rescue System

Applications of 5G Technologies in 2 Cases

On August 7, 2022, the aviation rescue drill successfully adopted 5G private network access, achieving advanced functions such as remote video consultation, vital sign data transmission, and the high-precision positioning of Beidou, which transmits the rescue trajectory of the rescue helicopter to the hospital emergency platform in real time through the 5G private network. This drill was novel in testing domestically produced portable extracorporeal membrane oxygenation on a helicopter, laying the technical foundation for practical rescue [10]. These achievements were successfully applied in the first international cross-regional long-distance aviation medical rescue case on November 9, 2022. This case further optimized the application of advanced technologies such as remote video consultation, innovative 5G module equipment, vital sign data transmission, and Beidou high-precision positioning; improved rescue time and efficiency; and shortened transit time [11]. The case validated the low airspace medical rescue 5G network technology, air body area network disease perception technology, and 5G air-to-ground collaborative rapid diagnosis and treatment technology. The entire rescue process seamlessly connected the accident scene via the applications of 5G technologies, medical ambulances or helicopters, and hospital emergency departments, achieving "admission as soon as boarding" and "integration of prehospital and in-hospital diagnosis and treatment," significantly improving the level of medical rescue.

Ethical Considerations

The study was conducted in accordance with the principles of the declaration of Helsinki. The study involving human participants was approved by the Ethics Committee of Shenzhen University General Hospital (KYLL-2023019A). All participants provided written informed consent before participating in the study. In this study, the principles of patient privacy protection were strictly followed. Identifying information, including patients' names, initials, or hospital numbers, was not disclosed.

Clinical Rescue Effectiveness

Improvement in Response Time

The helicopter had a flight distance of 60-600 km and a flight time of 10-136 minutes to cover this distance. In contrast, ambulances typically traveled at speeds of 40-60 km/hour depending on road conditions and distance and covered a distance of 60-600 km in 1-15 hours. Compared to the traditional ambulance, the flight time was less than a quarter of the traditional ambulance time, significantly reducing the rescue response time.

Innovative 5G Medical Equipment

On August 7, 2022, the aviation rescue trial first adopted the 5G private network access. On November 9, 2022, the 5G technology assisted the first international cross-regional long-distance aviation medical rescue operation using the novel airworthy 5G module vital sign monitoring equipment jointly developed with Mindray. The 5G emergency equipment could improve treatment efficiency [11,14].

Expanding the Rescue Scope

The traditional rescue radius is generally 5-10 km. However, in this study, the helicopter rescue radius was 60-600 km. The helicopter rescue could significantly expand the rescue scope and improve rescue efficiency compared with traditional ambulance vehicles expanding the rescue radius to 600 km.

XSL•FO
RenderX

Improving Rescue Effectiveness

A helicopter using 5G technology can transmit real-time high-definition images and videos through high-speed networks, helping doctors conduct remote diagnosis and treatment in the air. Further, 5G technology can also provide high-precision positioning and map navigation functions, helping rescue personnel locate patients more accurately, shortening search and rescue response time, and improving rescue efficiency.

Technical Application Effect

Previous study showed an improved model of helicopter air-ground cooperative rescue shortened the rescue time by 7.51% and reduced the rescue cost by 4.18% compared with a previous model [15]. Another study of 36 critical cases transferred between hospitals by emergency helicopters showed that the average transfer time was 54.95 (SD 17.89) minutes, and the average transfer distance was 205.74 (SD 74.68) km [16]. Such time for transfer requires airborne medical equipment and rescue techniques. A study used remote mobile medical technology, incorporating Internet of Things technology and a cloud computing system, by deploying information sensing devices on traditional transportation methods, such as cars and helicopters, and connecting these devices to the internet to enable intelligent identification, positioning, tracking, monitoring, and management [17]. This method can realize real-time video monitoring, call and dispatch command of the rescue scene, and remote real-time transmission of information in the car to improve the efficiency of prehospital first aid. The advancement of 5G low-altitude network communication technology was compared and analyzed in domestic and foreign literatures (Table 1) in order to provide a more scientific and effective decision-making basis for future medical rescue.

Table 1. Comparison of similar aviation technological parameters at home and abroad.

Technological parameters	Domestic traditional technology	Similar technologies used abroad	Technology used in this study	Comparison results
Low-altitude 5G network technology				
Communications technology	≤150 m, –108 dB, 95% [18]	≤80 m, –120 dB, 95% [19]	≤300 m, –100 dB, 100% [11]	Internationally leading
Positioning technology	Foreign GPS high-precision positioning; dynamic positioning accuracy ≤0.1 m [20]	Foreign GPS high-precision positioning; dynamic positioning accuracy ≤0.1 m [21]	China Beidou satellite high-precision positioning; dynamic positioning accuracy ≤1 m [11]	100% localization of technology
Air body area network disease awareness technology				
Human channel model	Multipose human model, estimation algorithm based on deep learning [22]	Static human body model [23]	A human model of walking and running states	Internationally leading
Encryption algorithm information entropy	7.9992 [24]	7.99930 [25]	7.9993	Similar effect
Encryption time	0.0834 [26]	1.476551 [25]	0.0385	Internationally leading
5G air-to-ground collaborative rapid diagnosis and treatment technology				
5G emergency equipment	Does not support 5G direct connection; Wi-Fi wireless access; packet loss rate <2% [27]	Does not support 5G direct connection; Wi-Fi wireless access; packet loss rate <2% [28]	Supports 5G direct connection with a packet loss rate of ≤0.2% [11]	Internationally leading
Remote real-time audio and video multidisciplinary consultation technology	None, mainly focused on intercom voice communication [29]	One-on-one audio and video communication [30]	Supports multiparty remote consultation under 1080P high-definition video image transmission [11]	Internationally leading

Future Directions

Due to the delayed development of domestic aviation rescue, the application of 5G technology in low-altitude scenarios is still in its infancy, resulting in limited use of 5G technology in aviation rescue operations. Proposed solutions include enhancing the aviation rescue system, promoting the development of 5G applications in low-altitude scenarios, researching 5G solutions suitable for various situations, strengthening coordination within the industry chain, and engaging in international cooperation. 5G network coverage is insufficient; although this study adopted ground tilt coverage technology in the 4.9-GHz frequency band, low-altitude coverage in some remote areas still needs improvement. The solution includes increasing base stations and optimizing their layout, using high-performance equipment and advanced algorithms, improving low-altitude signal transmission, sharing infrastructure, collaborating with civil aviation and meteorological industries for more information support, and considering satellite communication and ground relay as supplementary methods. Currently, 5G networks still have certain stability issues in high-speed mobile scenarios, which may affect the real-time communication performance of helicopter aviation medical rescue. To enhance the stability and communication quality of 5G networks in high-speed mobile

XSL•FO
RenderX

scenarios, it is important to explore and implement advanced signal processing technologies and network architectures. Although this study made breakthroughs in 5G key technologies, its global popularization and promotion still need strengthening. To strengthen collaboration with governments, enterprises, and research institutions, the focus is on collectively advancing the development of international standards and specifications for 5G technology. We will accelerate the construction and application of 5G networks. At the same time, it is necessary to increase the publicity of the advantages and application scenarios of 5G technology to improve the awareness and acceptance of 5G technology by the public and enterprises.

Improvement of Key Technologies

The key 5G technologies for helicopter aviation medical rescue need the following improvements:

- Improvement in the 5G communication coverage range and signal stability: higher frequency band ground tilt coverage technology can be studied to meet a wider range of low-altitude coverage needs, and multiantenna technology and anti-interference technology can be optimized to improve signal coverage range and stability.
- Improvement in personal positioning technology: besides the existing Beidou positioning terminal, hybrid positioning technology that integrates multiple satellite navigation systems (GPS, GLONASS [Global Navigation Satellite System], etc) can be studied to improve positioning accuracy and reliability.
- Optimizing the transmission scheme of the body area network: the near-body channel estimation algorithm based on deep learning and human body posture recognition can be further optimized to adapt to more complex human motion states and improve data transmission efficiency.
- Enhanced body area network data encryption method: more efficient encryption algorithms and protocols can be studied to improve data transmission security based on signal characteristics of physical signs and chaotic mapping.
- Improving 5G emergency equipment: 5G modular devices with higher transmission rates, lower latency, and lower power consumption can be explored to better adapt to various environments and application scenarios.
- Improvement in the functions of emergency platforms: besides real-time audio and video communications, emergency platforms with more intelligent functions can be developed, such as real-time data analysis, intelligent diagnostic advice, and so forth, to improve the effectiveness and efficiency of aerial medical treatment.

The key technologies of 5G also include millimeter wave communication, multiuser multiple-input multiple-output, network slicing, and edge computing. Millimeter wave communication offers higher data transmission speeds, multiuser multiple-input multiple-output technology supports simultaneous connections for more users, network slicing provides tailored network services for various application scenarios, and edge computing reduces data transmission delays while enhancing real-time performance. The 5G key technologies used in helicopter aviation medical rescue can better meet practical application needs through the aforementioned technological improvements and play a significant role in improving medical rescue efficiency and treatment level.

Conclusions

This viewpoint suggests that 5G technology had significant advantages in terms of improvement in response rescue time, innovative medical equipment, expansion of rescue scope, and improvement in rescue effectiveness. Although existing technologies have shortcomings in terms of network coverage, network stability, and other aspects, the 5G technologies for helicopter aviation medical rescue can better meet practical application needs. We believe that 5G technology can make breakthroughs in terms of rapid response, advanced equipment, and expanding rescue coverage, playing a significant role in improving the efficiency and level of global emergency medical rescue. Furthermore, research and practice on industrial application expansion, intelligent scheduling and management, and data security assurance can be important directions in developing 5G technologies for helicopter aviation medical rescue in the future.

Acknowledgments

The authors would like to express their deep gratitude to the following institutions. Without the support and funding of the “5G+ Medical and Health Application Pilot Project” of the Ministry of Industry and Information and the National Health Commission (Guangdong Province, Direction 1: 5G+Emergency Treatment, 11), this viewpoint could not be completed. Their support enables us to further study the application of 5G technology in emergency treatment. This study was funded by the Shenzhen Medical and Health Three Project Support (project SZSM201911007). Their funding has provided us with the necessary resources and platforms for research. The authors would also like to thank all their colleagues and collaborators who participated for their contributions and efforts on this report. No artificial intelligence tools were used in this study.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Lin X. Current situation and development of helicopter emergency medical rescue. China Sci Technol Inform 2022;4:113-115.
2. Chen Z, Dong B, Luo R. The particularity and existing problems of helicopter emergency medical rescue services. Chin Disaster Rescue Med 2021;9(8):1184-1188. [doi: 10.13919/j.issn.2095-6274.2021.08.011]
3. Yang L, Yang S, Plotnick L. How the internet of things technology enhances emergency response operations. Technol Forecast Soc Change 2013;80(9):1854-1867. [doi: 10.1016/j.techfore.2012.07.011]
4. Berlet M, Vogel T, Gharba M, Eichinger J, Schulz E, Friess H, et al. Emergency telemedicine mobile ultrasounds using a 5G-enabled application: development and usability study. JMIR Form Res 2022;6(5):e36824 [FREE Full text] [doi: 10.2196/36824] [Medline: 35617009]
5. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
6. Zhai Y, Xu X, Chen B, Lu H, Wang Y, Li S, et al. 5G-network-enabled smart ambulance: architecture, application, and evaluation. IEEE Netw 2021;35(1):190-196. [doi: 10.1109/mnet.011.2000014]
7. Attaran M. The impact of 5G on the evolution of intelligent automation and industry digitization. J Ambient Intell Humaniz Comput 2023;14(5):5977-5993 [FREE Full text] [doi: 10.1007/s12652-020-02521-x] [Medline: 33643481]
8. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
9. Naruse K, Yamashita T, Onishi Y, Niitaka Y, Uchida F, Kawahata K, et al. High-quality transmission of cardiotocogram and fetal information using a 5G system: pilot experiment. JMIR Med Inform 2020;8(9):e19744 [FREE Full text] [doi: 10.2196/19744] [Medline: 32897237]
10. Li G. Aeromedical emergency rescue drillG and ECMO join hands in medical rescue helicopter. People's Daily Client Guangdong Channel. URL: https://wap.peopleapp.com/article/6917993/6779744 [accessed 2022-08-10]
11. Li G. Eastern Navigation has completed the first practical application of 5G+aviation medical rescue. People's Daily Client Guangdong Channel. URL: https://szdbth.com/gongsiyaowen/212.html [accessed 2022-11-09]
12. Liu L, Pu H. Efficient on-demand charging planning in large-scale wireless sensor networks. Appl Res Comput 2022;39(1):5.
13. Gulati K, Kumar Boddu RS, Kapila D, Bangare SL, Chandnani N, Saravanan G. A review paper on wireless sensor network techniques in Internet of Things (IoT). Mater Today 2022;51:161-165. [doi: 10.1016/j.matpr.2021.05.067]
14. Zhao H. ECMO team appeared on the rescue scene and 5G+ aeromedical emergency rescue opened the green channel for air life. URL: https://baijiahao.baidu.com/s?id=1740685720448154135&wfr=spider&for=pc [accessed 2022-08-05]
15. Zhang M, Li S, Li B. An air-ground cooperative scheduling model considering traffic environment and helicopter performance. Comput Ind Eng 2021;158:107458. [doi: 10.1016/j.cie.2021.107458]
16. Li Y, Liao X, Zhao H, Zeng G, Ling Z, Wu G, et al. Thirty-six critical cases of emergency helicopter transferring between hospitals. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue 2021;33(8):1003-1006. [doi: 10.3760/cma.j.cn121430-20201110-00708] [Medline: 34590572]
17. He Z, Ren G, Ning L. Work requirements, process and training of primary health workers in seamless rescue system. Med Chinese Armed Police Force 2018;29(10):2.
18. Patzold M. 5G sets sail around the globe [mobile radio]. IEEE Veh Technol Mag 2019;14(2):5-11. [doi: 10.1109/mvt.2019.2902773]
19. Säe J, Wirén R, Kauppi J, Torsner J, Andreev S, Valkama M. Reliability of UAV connectivity in dual-MNO networks: a performance measurement campaign. 2020 Presented at: 2020 IEEE International Conference on Communications Workshops (ICC Workshops); 2020; Dublin, Ireland URL: https://ieeexplore.ieee.org/document/9145477 [doi: 10.1109/ICCWorkshops49005.2020.9145477]
20. Dong Y, Zhang L, Wang D, Li Q, Wu J, Wu M. Low-latency, high-rate, high-precision relative positioning with moving base in real time. GPS Solut 2020;24(2):56. [doi: 10.1007/s10291-020-0969-1]
21. Walko C, Maibach MJ. Flying a helicopter with the HoloLens as head-mounted display. Opt Eng 2021;60(10):103103. [doi: 10.1117/1.oe.60.10.103103]
22. Lei Y, He S, Khamis M, Ye J. An end-to-end review of gaze estimation and its interactive applications on handheld mobile devices. ACM Comput Surveys 2023;56(2):1-38. [doi: 10.1145/3606947]
23. Gupta S, Kakkar V, Ahmed S, Khanday FA, Sharma S, Singh S, et al. Modeling of on-chip biosensor for the in vivo diagnosis of hypertension in wireless body area networks. IEEE Access 2021;9:95072-95082. [doi: 10.1109/access.2021.3094227]
24. Wang J, Huang D, Fan S, Han K, Jeon G, Rodrigues JJPC. PSDCE: physiological signal-based double chaotic encryption for instantaneous E-healthcare services. Future Generation Comput Syst 2023;141:116-128. [doi: 10.1016/j.future.2022.10.034]
25. Zhang Y. A chaotic system based image encryption scheme with identical encryption and decryption algorithm. Chin J Electronics 2017;26(5):1022-1031. [doi: 10.1049/cje.2017.08.022]
26. Wang X, Li Y, Jin J. A new one-dimensional chaotic system with applications in image encryption. Chaos Solitons Fractals 2020;139:110102. [doi: 10.1016/j.chaos.2020.110102]

XSL•FO
RenderX

27. Yuan H, Xiao N, Cheng M, Kaidi W. A teleoperation control method for vascular interventional surgery robot. : IEEE; 2021 Presented at: 2021 IEEE International Conference on Mechatronics and Automation (ICMA); August 08-11, 2021; Takamatsu, Japan p. 1027-1032. [doi: 10.1109/ICMA52036.2021.9512797]
28. Pervez F, Qadir J, Khalil M, Yaqoob T, Ashraf U, Younis S. Wireless technologies for emergency response: a comprehensive review and some guidelines. IEEE Access 2018;6:71814-71838. [doi: 10.1109/access.2018.2878898]
29. Wei S. Effectiveness and feasibility of helicopter rescue in trauma treatment. Fire Protection Today 2023;8(9):24-26.
30. Johnsen AS. Helicopter emergency medical services in sudden-onset major incidents: patterns of use [dissertation]. University of Stavanger. 2021. URL: https://ebooks.uis.no/index.php/USPS/catalog/book/85 [accessed 2024-06-07]

Edited by A Mavragani; submitted 28.06.23; peer-reviewed by H Fan, J Wu, B Qu; comments to author 08.02.24; revised version received 22.02.24; accepted 21.05.24; published 01.08.24.

Please cite as:
Han Sr W, Li 2nd Y, Chen 3rd C, Huang D, Wang J, Li X, Ji Z, Li Q, Li Z
5G Key Technologies for Helicopter Aviation Medical Rescue
J Med Internet Res 2024;26:e50355
URL: https://www.jmir.org/2024/1/e50355
doi:10.2196/50355
PMID:

©Wei Han Sr, Yuanting Li 2nd, Changgen Chen 3rd, Danni Huang, Junchao Wang, Xiang Li, Zhongliang Ji, Qin Li, Zhuang Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint

Li Zhui[1], PhD; Li Fenghe[1], PhD; Wang Xuehu[1], PhD; Fu Qining[1], PhD; Ren Wei[1], PhD

Department of Vascular Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

Corresponding Author:
Ren Wei, PhD
Department of Vascular Surgery
The First Affiliated Hospital of Chongqing Medical University
No. 1 of Youyi Road, Yuzhong District
Chongqing, 400016
China
Phone: 86 13658339771
Fax: 86 89011231
Email: renwei_2301@yeah.net

Abstract

This viewpoint article first explores the ethical challenges associated with the future application of large language models (LLMs) in the context of medical education. These challenges include not only ethical concerns related to the development of LLMs, such as artificial intelligence (AI) hallucinations, information bias, privacy and data risks, and deficiencies in terms of transparency and interpretability but also issues concerning the application of LLMs, including deficiencies in emotional intelligence, educational inequities, problems with academic integrity, and questions of responsibility and copyright ownership. This paper then analyzes existing AI-related legal and ethical frameworks and highlights their limitations with regard to the application of LLMs in the context of medical education. To ensure that LLMs are integrated in a responsible and safe manner, the authors recommend the development of a unified ethical framework that is specifically tailored for LLMs in this field. This framework should be based on 8 fundamental principles: quality control and supervision mechanisms; privacy and data protection; transparency and interpretability; fairness and equal treatment; academic integrity and moral norms; accountability and traceability; protection and respect for intellectual property; and the promotion of educational research and innovation. The authors further discuss specific measures that can be taken to implement these principles, thereby laying a solid foundation for the development of a comprehensive and actionable ethical framework. Such a unified ethical framework based on these 8 fundamental principles can provide clear guidance and support for the application of LLMs in the context of medical education. This approach can help establish a balance between technological advancement and ethical safeguards, thereby ensuring that medical education can progress without compromising the principles of fairness, justice, or patient safety and establishing a more equitable, safer, and more efficient environment for medical education.

(J Med Internet Res 2024;26:e60083) doi:10.2196/60083

KEYWORDS

medical education; artificial intelligence; large language models; medical ethics; AI; LLMs; ethics; academic integrity; privacy and data risks; data security; data protection; intellectual property rights; educational research

Introduction

Large language models (LLMs), such as OpenAI's ChatGPT series, Microsoft's Copilot, and Google's Gemini, are sophisticated artificial intelligence (AI) tools that are trained on vast text datasets via deep learning techniques. These models exhibit the ability to generate language that is highly similar to human speech; thus, their use in medicine is becoming increasingly widespread. From assisting scientific research to informing clinical decisions and supporting medical education, LLMs can be used to optimize the use of information and resources across diverse domains [1].

In the technology-driven transformation of medical education, the introduction of LLMs has the potential to revolutionize traditional learning and teaching methods. These advanced language models can significantly enhance the personalization and practicality of education by generating tailored learning materials and simulating digital experiences that resemble real clinical scenarios [2]. LLMs dynamically generate customized learning materials and assessments based on students' progress

XSL•FO
RenderX

and needs, thereby improving the efficiency and effectiveness of learning [3]. Moreover, by simulating patient interactions, LLMs can help students practice and hone their diagnostic and communication skills in a safe environment, which is a crucial element of attempts to nurture the practical abilities of medical professionals [4].

Biomedical ethics is founded on 4 fundamental principles: autonomy, nonmaleficence, beneficence, and justice [5]. Medical education should not only aim to cultivate these principles among students but also continuously evolve and improve under their guidance [6]. Due to the increasingly widespread use of LLMs in medical education research, some foreseeable ethical issues present challenges to these basic principles. For example, the privacy and data security risks associated with LLMs can infringe upon individuals' autonomy to control their own information, thus potentially causing harm. Biases in training data and algorithms can lead to unfair diagnostic and treatment decisions that can harm patients, and these biases may also risk spreading misinformation, thereby violating the principle of beneficence. Additionally, issues related to academic integrity and concerns regarding responsibility and copyright may risk violating the principle of justice, among others.

While previous research on education-related LLMs has emphasized the tasks of enhancing key competencies or exploring educational LLM programs and related concepts, it has failed to account for ethical principles specifically in the development and application of such technologies [7,8]. Addressing and resolving these issues are vital not only for technological advancement but also for the tasks of ensuring high-quality medical education and protecting the rights of both practitioners and patients. Therefore, this viewpoint article aims to explore and analyze the potential ethical issues that LLMs might raise in the context of future medical education and to identify fundamental principles for developing an ethical framework for these technologies as well as corresponding methods for their implementation.

Ethical Concerns Related to the Development of LLMs

Privacy and Data Risks

In the era of electronic information technology, privacy risks and data security have always been associated with advancements in information technology. LLMs are associated with enormous numbers of parameters and training data sets and may “memorize” sensitive personal information from their training data, thus, enabling these models to generate information that contains specific individual data [9]. Jegorova et al [10] confirmed the risk of privacy leakage during the inference process exhibited by learning models. When teachers or educational institutions use data, including students' personal information, social background information, and health data, to train models to provide personalized feedback regarding students’ learning interests and progress, this process entails the risk of sensitive information leakage. When LLMs are used for the purpose of assisted learning, such as in the contexts of clinical case discussions, tasks involving sensitive patient data such as name, sex, age, medical history, or final diagnoses, and even imaging data such as computed tomography scans and magnetic resonance imaging for assistance in interpretation, patient privacy is put at risk. Previous studies have shown that even if this information is anonymized when it is input, some advanced models can reidentify such personal information from large data sets through so-called linkage attacks, thus, resulting in information exposure [10]. Rocher et al [11] reported that LLMs could use as few as 15 demographic attributes to reidentify 99.98% of personal data in any anonymized data set. Therefore, deidentifying information alone is insufficient to protect patient privacy, a situation which poses tremendous challenges with regard to protecting patient privacy when LLMs are used in the context of medical education. Furthermore, as the training and user data of LLMs are stored on cloud servers, any security vulnerabilities could result in the leakage of such sensitive information [12]. LLM development companies, such as OpenAI, may use users' personal information for service analysis, improvement, or research purposes and retain the right to share users' personal information with third parties without explicit user consent [13]. All of these factors exacerbate the risks pertaining to privacy and data security in medical education processes.

AI Hallucination

Clinical practice guidelines and consensus documents, which are continually updated, are significantly valuable as references in the context of medical education. However, “AI hallucinations” on the part of LLMs represent a major obstacle to the effective use of these documents as educational content. In the field of medicine, AI hallucinations occur when LLMs generate responses that appear to be logical but are actually incorrect, inconsistent, or fabricated, including those that rely on fabricated data and forged references [14]. Amir et al [15] compared ChatGPT 3.5 and Google BARD in terms of their ability to answer questions related to lung cancer, revealing that these LLMs exhibited error rates ranging between 17.5% and 27.5%. Similar issues have been observed in the field of dentistry, in which context LLMs can offer vague, outdated, and inaccurate information [16]. Studies have highlighted accuracy concerns with regard to LLMs’ responses in medical contexts [17,18]. Several factors contribute to AI hallucinations in the field of medicine. First, LLMs' training data, which are sourced mostly from the internet, include outdated and inaccurate information that lacks quality control. Second, access to authoritative databases such as PubMed, UpToDate, and Cochrane requires subscriptions, thus, restricting LLMs from obtaining up-to-date and reliable research data [19,20]. Additionally, limitations regarding LLMs’ reasoning and probabilistic text generation capabilities exacerbate the problem of AI hallucination.

Given the high level of precision required in medicine, even minor errors can have catastrophic consequences for patients. When medical educators use unverified LLM-generated content, the “one-to-many” model facilitates the widespread propagation of erroneous information. Medical students, who are in the process of acquiring foundational knowledge and trust their educators, are especially vulnerable to such misinformation. This issue can severely impact their future clinical decisions. The use of inaccurate LLM content to evaluate students reinforces incorrect information, leading to flawed assessments and continued learning on the basis of faulty knowledge. In

XSL•FO
RenderX

academic research, students may unknowingly cite false references or use incorrect data, resulting in academic misconduct. If erroneous LLM-generated information is widely disseminated in medical education, it can lead to systemic errors in clinical decision-making, thus, endangering patient health and safety. This situation undermines the rigor of medical education, erodes public trust in health care, and raises ethical issues and challenges related to academic integrity, which ultimately impact the professionalism and social responsibility of the medical field.

Training Data and Algorithmic Biases

Zack et al [21] reported that ChatGPT-4 often resorts to stereotypes and exhibits bias toward specific races, ethnicities, and sexes in terms of its diagnostic outputs. Additionally, treatment suggestions have been shown to be linked to demographic factors, to favor costlier procedures and to exhibit variations in terms of patient perceptions [21]. This widespread problem of unfair responses is primarily due to such LLMs' training data and algorithmic biases.

Bias in training data is the most direct cause of bias in LLMs' responses. LLMs learn from vast amounts of unfiltered text; thus, if these data contain biases related to sex, race, culture, or socioeconomic background, the model replicates these biases. A study on a chest x-ray classifier that had been trained mainly on White patients revealed that it lacked recognition capability in cases involving non-White patients [22]. As the amount of software-generated data increases, such information may be used as training data, further reinforcing the original biases [23].

Algorithmic bias on the part of LLMs exacerbates the problem of output bias. During the process of LLM design, the optimization goals and evaluation criteria used often focus on the majority population, neglecting the needs of the minority. For example, if an LLM is trained primarily on Western cases of alcoholic cirrhosis without considering geographic and cultural diversity or including data concerning hepatitis B-induced cirrhosis from other regions, it may provide inaccurate guidance regarding the diagnosis and treatment of hepatitis B-related cirrhosis. Failure to address training data bias and adjust algorithms perpetuates unfairness in terms of clinical practices [24]. Additionally, developers may intentionally introduce biases, which can affect data processing, model training, algorithm selection, and application, ultimately causing specific groups to receive unfair treatment. Algorithm bias impacts LLMs' responses regardless of the presence of training data bias [25].

Biases embedded in data and algorithms are reproduced in LLMs' outputs, resulting in incomplete educational content and depriving students of comprehensive and in-depth learning guidance. This lack of diversity in educational content deprives students of a balanced medical perspective, which impacts the quality and fairness of education. Future doctors may lack the knowledge and ethical sensitivity they need to serve diverse patient populations effectively, potentially leading to misdiagnoses or inappropriate treatments. Biased information based on stereotypes and unfair descriptions of specific groups can distort students' ethical concepts during the process of diagnosis and treatment, inadvertently fostering biases against these groups. Furthermore, this gap not only affects individual students' achievements but is also amplified through medical education, thereby weakening the health care system's ability to meet the needs of diverse patient populations. LLM biases violate core biomedical ethics principles, such as nonmaleficence, beneficence, and justice. Ensuring fair and neutral LLM use in medical education is crucial with regard to efforts to maintain educational quality and uphold professional ethics.

Deficiencies in Transparency and Interpretability

When users ask questions, LLMs usually generate a predefined response. However, users often cannot comprehend how this response is generated or the underlying logic. This issue is known as the "black box" effect, in which context transparency and interpretability are lacking [26]. Transparency refers to whether the model's operation and decision-making process can be presented clearly to the user, while interpretability refers to whether users can understand the logic and reasons that underlie the decisions or text, thus, generated [9]. Such a lack of transparency and interpretability is mainly due to LLMs' large model structures and complex training data, which make their internal operations difficult to understand.

In the context of medical education, the lack of transparency and interpretability in LLMs can prevent teachers and students from assessing the accuracy of information effectively [27]. If teachers convey inaccurate information to students, such information inevitably entails risks to patient safety in clinical practice. The fact that students exhibit a high level of trust in their teachers indicates that these deficiencies can have more widespread and serious impacts than in other fields. Additionally, a lack of understanding of the logic underlying LLMs may lead to rigid thinking on the part of students, who may neglect the importance of logical thinking in medical education, thus weakening efforts to cultivate their critical thinking. Logical reasoning is crucial in clinical practice with regard to analyzing conditions accurately and formulating treatment plans. A lack of logical reasoning on the part of doctors may lead to misdiagnosis, mistreatment, and risks to patient safety. Moreover, when teachers cannot grasp the logic employed by LLMs, they may convey conclusions to students without explaining the corresponding deductive process, thus hampering their ability to guide students and decreasing the quality of teaching.

Ethical Concerns Related to the Application of LLMs

Emotional Intelligence Deficiency

Although LLMs are trained to provide empathetic responses in the context of engaging with patients [28]—recent research has shown that ChatGPT even exhibits higher levels of empathy when consulting on systemic sclerosis than do neurologists [29]—the artificial empathy expressed by AI generally cannot replicate the subtle genuineness of the emotions conveyed by health care professionals [30]. True emotional expression is the result of perceiving and understanding the patient's emotions through real-time communication, especially by sensing changes in tone, facial expressions, and body language, with the goals of comprehending the patient's needs and concerns and

XSL•FO
RenderX

providing personalized feedback. However, existing LLMs cannot perceive these nonverbal emotional changes. Furthermore, the expression of empathy by LLMs is limited to textual descriptions of emotions, whereas variations in tone, facial expressions, and body language on the part of health care professionals, such as handshakes or hugs, often convey empathy more effectively than can words alone [1]. Prof Marc Succi suggested that excessive reliance on artificial emotional communication in medical settings may exacerbate the societal public health issue of "loneliness" [31]. Establishing and developing empathy involves addressing the complexity and uncertainty that characterize real interactions, which are shaped by patients' diverse backgrounds, beliefs, and values. LLMs, as teaching tools, cannot convey a diverse range of emotional expressions beyond the level permitted by text, thus, preventing medical students from gathering materials that can help them improve their communication skills and empathy through such interactions. In addition to direct patient communication, teachers who exhibit empathy and care in clinical settings also serve as good examples who can enhance students' empathy. Sole reliance on LLMs for tasks such as medical history collection, condition disclosure, and obtaining consent can reduce the presence of empathy in clinical environments, thus, conflicting with the goals of medical ethics education. Excessive reliance on artificial emotional intelligence can lead to a lack of empathy among medical students. The development of empathy requires ongoing practice in diverse emotional settings, which static LLMs cannot provide.

Educational Unfairness

Unfairness in the context of traditional medical education is characterized primarily by several aspects: uneven resource distribution, disparities in faculty strength, high tuition fees that limit student access, and geographical disparities in terms of school location [32]. While LLMs offer rich educational resources and opportunities, their application is a double-edged sword with respect to educational fairness.

For institutions that have only limited teaching resources, LLMs can aggregate rich teaching resources at a global level, thus, providing students with high-quality educational experiences. They also offer relatively equal learning opportunities to regions located in remote geographical areas and enable nonnative speakers to access and understand medical knowledge in other languages through cross-language support [33]. However, medical schools that have better technological infrastructure, higher budgets, and access to the most up-to-date technologies may find it easier to take advantage of the benefits of LLMs in the context of education. In contrast, schools and regions with limited resources may experience exacerbated resource allocation inequalities because of their inability to afford these technologies, their lack of the professional knowledge necessary to implement these technologies, and restrictions on accessing LLMs (such as ChatGPT, which is currently unavailable to Chinese users). Moreover, the use of LLMs requires extensive data processing, which may expose data to privacy and security risks. Institutions that lack sufficient resources and robust infrastructure may struggle to safeguard user data, thus, increasing the risk of data leakage and misuse. With respect to academic achievements, individuals who can use these tools may gain access to more recognition and opportunities within the academic community, whereas students and researchers who lack such access may face disadvantages in terms of academic competition [34].

Academic Integrity Concerns

Issues related to academic integrity in the context of medical education notably include plagiarism and cheating. LLMs greatly facilitate medical research by summarizing published articles and assisting with data analysis or literature retrieval. Research has indicated that when abstracts generated by LLMs and those authored by humans are presented simultaneously, university professors struggle to distinguish between them [35]. Another study that examined LLMs' generation of fictitious article abstracts revealed that peer reviewers could identify only 68% of abstracts generated by ChatGPT as fictitious [36]. This outcome indicates that the quality of the text produced by LLMs has approached or reached a professional level, thus, giving rise to opportunities for academic plagiarism; namely, students may use LLMs to generate abstracts, data, or even complete draft papers. Despite the originality of the content generated by LLMs, students who submit these texts as their own work may be suspected of plagiarism. With regard to one recently published article, the direct replication of generated content, including the prompts from LLMs—the regenerated response—resulted in article retraction [37]. Numerous other articles now face similar scrutiny, and the question of whether these will be retracted by the journal or revised remains undecided [38,39]. This excessive reliance and lack of critical use of LLMs have caused serious damage to academic integrity and posed ethical challenges. Furthermore, the integration of LLMs into browsers facilitates direct searches for source texts or other multimedia documents, which can exacerbate issues related to plagiarism and make them more prevalent when teachers or students use these data directly. A well-known test of the accuracy of LLMs in the context of medical questions is ChatGPT's successful passing of the United States Medical Licensing Exam [40]. Many studies have shown that LLMs achieve high scores in various disciplines [41-43]. This high level of accuracy enables students to use LLMs to cheat on their examinations. Excessive reliance on LLMs may decrease students' research ability, critical thinking, innovation, and motivation to engage in proper learning [44]. This situation can perpetuate students' reliance on LLMs and academic misconduct, thus, leading to a vicious cycle. Additionally, if students and teachers plagiarize LLM-generated content without critical evaluation, erroneous or biased information may become widespread.

Responsibility and Copyright Concerns

In the context of traditional clinical education, clinical teachers supervise students' operations, provide students with guidance, and take responsibility for adverse clinical events, while students adhere to medical standards and ethical norms. Medical institutions must ensure a safe learning environment and sufficient supervision. This clear division of responsibilities can help protect patient safety and improve the quality of education.

XSL•FO
RenderX

In the future, the introduction of LLMs into medical education will inevitably add new dimensions to the division of responsibilities. From a legal perspective, AI lacks the legal standing of a human being, thus leaving humans as the ultimate accountable parties [45]. However, in medical settings, the responsibility for patient harm due to LLM biases and inaccuracies is unclear, a situation which affects doctors, patients, and institutions such as medical facilities and developers. LLM developers should ensure that their tools are accurate and safe, but entities such as OpenAI disclaim responsibility for the texts generated by their LLMs [46]. Existing ethical and legal frameworks have not fully adapted to the challenges posed by such emerging tools, thus, leading to ambiguity in terms of responsibility attribution and regulation in the context of actual application.

When teachers use LLMs to generate teaching outlines, lecture notes, or textbooks or when students use LLMs to write academic papers—especially when LLMs access original texts and images from the internet through search engines—and then copy from copyrighted source materials directly or borrow from them excessively, copyright disputes can occur [4]. To protect copyrights and maintain academic integrity more effectively, current guidelines issued by the International Committee of Medical Journal Editors and the Committee on Public Ethics suggest that LLMs such as ChatGPT should not be listed as authors of papers [47]. LLMs cannot bear the same responsibility as human authors or provide substantive explanations for the content they generate. However, this approach is not universally accepted, and the role of LLMs in authorship declarations remains the subject of ongoing discussion by various publishing institutions [48,49].

AI-Related Laws and Ethics

Overview

Due to the advancement of AI technology in various domains, particularly with regard to the development of LLMs, multiple countries and international organizations have actively formulated or updated ethical and legal guidelines. These measures focus on key areas such as privacy protection, transparency, algorithmic fairness, and accountability and offered reliable guidance concerning the reasonable use and development of artificial intelligence technologies. AI legal and ethical principles are in Multimedia Appendix 1.

Laws

The General Data Protection Regulation of the European Union, which was enacted in 2018, offers stringent guidelines for handling sensitive data in the education and health care sectors. These guidelines emphasize the pseudonymization or anonymization of personal data storage, requiring the highest privacy settings, and ensuring transparency in terms of data processing. Additionally, they strengthen users' control over their personal information, thus, empowering individuals to manage their data effectively [50]. In China, the Personal Information Protection Law, which was implemented in 2021, establishes rigorous standards for the protection of personal information, particularly with regard to user awareness and consent rights [51]. Similarly, the Health Insurance Portability and Accountability Act in the United States prioritizes the protection of medical information and respect for patient privacy [52].

Ethical Guidelines and Initiatives

International guiding principles have extensively discussed issues pertaining to transparency, interpretability, algorithmic fairness, and accountability in the context of AI. The Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems, which was founded in 2019, emphasized the transparency of autonomous systems in AI design processes and ethical guidelines for handling private data and algorithmic biases [53]. The European Union's White Paper on Artificial Intelligence, which was released in February 2020, elaborated on ethical frameworks for system transparency, privacy data, and algorithmic biases, particularly with regard to the importance of avoiding bias and discrimination and ensuring the fair use of AI decision support systems across diverse populations [54]. The United Nations Educational, Scientific and Cultural Organization's "Recommendations on the Ethics of Artificial Intelligence," which was issued in November 2021, systematically addressed privacy rights, data protection, transparency, interpretability, responsibility, and accountability in the context of AI applications, emphasizing fairness and nondiscrimination [55]. In January 2024, the World Health Organization issued the "Ethics and Governance of AI in Health: Guidance on Large Multimodal Models," which was the first ethical standard for large multimodal models such as LLMs. This document emphasized transparency, interpretability, and accountability; promoted fair access; and prioritized inclusivity [56]. In March 2024, the United Nations General Assembly adopted a US-led resolution titled "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development," which represented the first global consensus on AI governance. This resolution addressed the needs to bridge regional AI gaps, protect against discrimination and bias, safeguard privacy, and respect intellectual property (IP) rights, among other 13 proposals [57].

Limitations

Existing legal and ethical guidelines ensure the safe, transparent, and fair use of AI systems. However, when LLMs are used in the context of medical education, significant limitations emerge. First, these guidelines, which have been shaped by various national, organizational, and cultural contexts, reflect different social values, historical backgrounds, and priorities. For instance, the General Data Protection Regulation and the Personal Information Protection Law in China emphasize data privacy, whereas the Institute of Electrical and Electronics Engineers guidelines prioritize technological innovation. Such differences can lead to conflicts and increase the complexity and uncertainty of applying LLMs across different countries. Second, despite the fact that these guidelines emphasize transparency and interpretability, the complex algorithms and nonlinear structures of LLMs make these principles difficult to achieve. Existing guidelines lack detailed, actionable recommendations that can help users understand and trust LLM decision-making in the context of medical education.

XSL•FO
RenderX

Furthermore, LLMs may amplify biases in their training data, thus, leading to unfair outcomes. While many guidelines have emphasized the importance of avoiding discrimination and bias, they have not provided specific tools or methods to detect and correct biases within LLMs. Additionally, current ethical oversight and accountability mechanisms are inadequate with respect to the use of LLMs in medical education. When LLMs generate misleading or inaccurate content that could cause harm, existing frameworks lack clear guidelines for determining when human intervention is necessary, how to monitor and adjust the model's output, and how to establish accountability. Most importantly, existing legal and ethical guidelines in the field of medical education are fragmented and lack cohesiveness. They do not comprehensively address unique challenges such as academic misconduct, plagiarism, and copyright issues. Consequently, educators often struggle to find clear, authoritative solutions in practice. Therefore, it is critical to develop an ethical framework that is specifically tailored to the application of LLMs in medical education, thereby ensuring the safety, efficacy, and fairness of such technologies and promoting their widespread and in-depth application in this critical field.

Fundamental Principles of and Ways of Implementing LLM Ethics in Medical Education

Overview

To ensure that LLMs are used within appropriate legal and ethical frameworks, it is essential to establish a unified ethical framework. In light of the broad biomedical principles proposed by Beauchamp and Childress, namely, beneficence, justice, autonomy, and nonmaleficence [5], we propose fundamental principles and implementation methods for the use of LLMs in the context of medical education that are in line with these biomedical ethical principles (Figure 1 and Table 1).

Figure 1. Eight fundamental principles regarding the use of large language models in medical education.

XSL•FO
RenderX

Table 1. Implementation methods aligned with biomedical ethical principles regarding the use of large language models (LLMs) in medical education.

Ethical principles for LLMs in medical education	Implementation methods	Biomedical ethical principles
Privacy and data protection	1. Multilayered privacy protection mechanisms 2. Data encryption and secure storage 3. Strict data usage policies 4. Strict access controls	Nonmaleficence and autonomy
Data quality control and supervision	1. Develop independent LLM databases 2. Link to authoritative databases 3. Maintain a critical approach	Beneficence and justice
Fairness and equal treatment	1. Diverse and comprehensive training data 2. Fairness review and monitoring mechanisms 3. Open and accessible resources and tools 4. Cross-regional and cross-economic background collaboration	Justice
Transparency and interpretability	1. Model documentation and operational guidelines 2. Implement interpretative interfaces or tools 3. Regular public evaluations and feedback	Autonomy
Academic integrity and moral norms	1. Establish policies and guidelines 2. Develop and use detection tools 3. Maintain educators' leadership roles 4. Foster empathy and compassion	Nonmaleficence and beneficence
Accountability and traceability	1. Clear responsibility framework 2. Detailed logging and monitoring systems 3. Independent review and investigation committees 4. Clear complaint and feedback channels	Nonmaleficence and justice
Intellectual property protection	1. Strict IP compliance processes 2. Copyright statements and terms of use 3. Monitoring and detection systems 4. Complaint and feedback mechanisms	Justice and autonomy
Educational research and innovation	1. Establish dedicated research and development funds 2. Creating an open teaching experimentation environment 3. Optimizing teaching content	Beneficence

Privacy and Data Protection

To mitigate the risk of sensitive data breaches effectively and ensure both data security and privacy protection, it is crucial to implement a series of robust data protection measures involving dedicated LLM training databases that are specifically designed for medical education.

Multilayered Privacy Protection Mechanisms: Advanced data deidentification techniques, such as generalizing and suppressing sensitive information, must be used to prevent reidentification. Differential privacy techniques must be employed to enhance protection by adding noise to the data [58].

Data Encryption and Secure Storage: Data encryption and storage must be enhanced based on tamper-proof hardware security modules and distributed storage systems with the goals of preventing attacks and ensuring data security in terms of both storage and transmission [59].

Strict Data Usage Policies: Strict policies must be established and followed to ensure that data providers have control over and transparency regarding their data use. Users must be informed about data usage and allowed to withdraw their consent at any time.

Strict Access Controls: Multifactor authentication and stringent access controls, such as SMS text messaging verification and facial recognition, must be implemented to ensure that only authorized personnel access sensitive data [60].

Data Quality Control and Supervision

To prevent partial and inaccurate responses based on erroneous or outdated training data, effective quality control and supervision mechanisms must be established.

Develop independent LLM databases: all training data should be reviewed by clinical experts before incorporation to ensure authenticity and accuracy. LLM-generated content must be regularly reviewed and evaluated to identify and correct biases and errors promptly.

Link to authoritative databases: authoritative databases such as PubMed and UpToDate must be used to ensure the accuracy and relevance of training data. These resources are to be updated regularly with the latest medical research, thereby reducing errors and biases as well as enhancing data credibility.

XSL•FO
RenderX

Maintain a critical approach: educators and students should assess LLM-generated information critically by verifying it by reference to authoritative databases and other language models. This practice can ensure high-quality standards in medical education involving LLMs.

Fairness and Equal Treatment

Mitigating biases in LLMs can ensure that all students and educators have equal opportunities and resources in the context of medical education, thereby eliminating the health disparities and knowledge inequalities caused by these biases.

Diverse and comprehensive training data: developers must use diverse datasets during model training to avoid biases. This process involves reviewing and filtering data with the goal of eliminating the factors that lead to discrimination or unfair treatment.

Fairness review and monitoring mechanisms: fairness review and monitoring mechanisms must be incorporated into the model's deployment and usage. The biases or unjust tendencies exhibited by the model outputs must be evaluated and detected regularly, and adjustments and optimizations must be made as needed.

Open and accessible resources and tools: open and accessible resources and tools must be provided to students and educators at various economic levels and geographic locations to enable them to use LLMs on an equal basis. For example, developing lightweight or localized versions of a model can support advanced technology adoption in resource-limited areas.

Cross-regional and cross-economic background collaboration: cross-regional and cross-economic collaboration platforms and communities must be established and supported with the goal of promoting the sharing of knowledge and resources, thereby mitigating the educational inequalities caused by regional disparities.

Transparency and Interpretability

Ensuring that the operations and decision-making processes of LLMs are transparent and explainable to students and educators is critical with respect to enhancing students' and educators' understanding and trust in the system.

Model documentation and operational guidelines: developers should provide comprehensive documentation and guidelines to explain the model's mechanisms, input-output processes, and decision-making logic. These guidelines should cover design objectives, the sources of training data and corresponding methods, and potential limitations with the aim of helping users understand the model's construction and functionality.

Interpretative interfaces or tools: in the context of model use, interpretative interfaces or tools must be implemented to allow students and educators to understand the model's reasoning and decisions in real time. For example, visual tools can reveal how the model processes input data and generate outputs, thereby enhancing the model's transparency and interpretability [61].

Regular public evaluations and feedback: involving experts and users in regular evaluations and reviews and establishing effective feedback mechanisms can help identify and correct issues or biases associated with the use of the model, thereby promoting continuous improvement and optimization.

Academic Integrity and Moral Norms

An emphasis on academic integrity and ethical norms can prevent students and educators from gaining undue advantages from LLMs or from engaging in misconduct, such as plagiarism or cheating. Students and educators should use LLMs reasonably, avoid excessive reliance on such technologies, and focus on cultivating empathy, compassion, and critical thinking.

Establish policies and guidelines: educational institutions should establish specific LLM usage policies and guidelines based on a unified set of ethical principles to specify the legal and ethical use of LLMs as well as to prohibit any form of academic misconduct clearly and explain the corresponding consequences.

Develop and use detection tools: detection tools must be developed to identify and prevent inappropriate content or plagiarism with regard to LLM-generated material, thus ensuring that students' submissions and research results are genuine and original [62].

Maintain educators' leadership roles: during the teaching process, educators should continue to play a leadership role by encouraging students to analyze and evaluate LLM-generated content critically and emphasizing that models are auxiliary tools rather than substitutes for human thought. Educators should encourage students to question and verify information from LLMs, thereby enhancing their independent thinking skills.

Foster empathy and compassion: in clinical practice, cultivating students' empathy and compassion through rich interpersonal interactions can enhance the humanistic aspect of patient care. Educators should lead by example, such as by demonstrating material and guiding students in the process of applying these skills in real clinical settings. This approach can help students understand and respect patients' emotions and needs, ultimately improving the overall quality of health care services.

Accountability and Traceability

To ensure that responsibilities and accountability procedures are defined clearly when LLM decision errors negatively affect patient health, clear accountability and traceability mechanisms must be established.

Clear framework for responsibility: developing a clear framework of responsibilities and obligations for model developers, deployers, and users at different stages is crucial. Developers should ensure that the model's design and training meet safety and fairness standards; deployers must ensure compliance with supervision and quality control measures; and users must make and verify professional judgments on the basis of the model's recommendations.

Detailed logging and monitoring systems: implementing detailed logging and monitoring during the use of LLMs involves recording the model's inputs, outputs, and decision paths, which can make it possible to trace the decision-making process when issues arise.

Independent review and investigation committees: independent committees must be established to address and investigate health

XSL•FO
RenderX

issues resulting from LLM decision errors while ensuring fair procedures. These committees should include technical experts, legal advisors, and medical professionals who can assess responsibilities jointly and recommend improvements.

Clear complaint and feedback channels: providing clear complaint and feedback channels for patients and users allows them to report issues and seek help promptly in cases involving harm.

Intellectual Property Protection

Ensuring compliance with IP laws during the use of LLMs and protecting the rights of knowledge creators and holders are critical to the task of preventing unauthorized knowledge use and infringement.

Strict IP compliance processes: strict IP compliance processes must be established to ensure that LLMs use only authorized or publicly licensed data. Developers should document data sources meticulously and review them regularly for legality and compliance.

Copyright statements and terms of use: when LLM-generated content is used, clear copyright statements and terms of use must be provided. Such statements and terms of use can ensure that users understand and comply with IP laws, thus preventing the unauthorized creation or dissemination of protected content.

Monitoring and detection systems: effective monitoring and detection systems can be implemented to identify and prevent unauthorized use and potential infringement. For example, copyright detection tools can automatically flag potentially infringing content in model outputs [63].

Complaint and feedback mechanisms: complaint and feedback mechanisms can be established to allow creators and IP holders to report and address potential infringement issues promptly, thus facilitating efficient remedial actions.

Educational Research and Innovation

Taking advantage of the benefits of LLM technology for educational research and innovation can enhance the quality and effectiveness of medical education, thus promoting the continuous improvement and optimization of teaching methods and content.

Establish dedicated research and development funds: dedicated research and development funds must be established to support the attempts of educational institutions and academic groups to conduct LLM-centered innovative educational projects and research. Encouraging interdisciplinary collaboration and integrating the most recent advancements in AI and medical education can facilitate the exploration of novel teaching methods and content.

Open teaching experimentation environment: establishing an open experimentation and testing environment can allow educators to explore innovative teaching methods safely. The powerful analytical capabilities of LLMs can be used to evaluate the effectiveness of these methods. The development of LLM-based intelligent tutoring tools and digital learning modules can further optimize teaching strategies and enhance the student learning experience, ultimately improving overall teaching quality.

Optimizing teaching content: to ensure that medical education remains aligned with the most up-to-date advancements in medical research and technology, researchers should regularly monitor and evaluate recent findings and clinical practices. These insights must be integrated into the curriculum via LLMs to ensure that the curriculum remains current and promotes continuous improvement.

While we have to identify certain fundamental principles and methods for an ethical framework for LLMs in medical education, several key areas require further refinement. First, this framework should be dynamic and adaptable to facilitate continuous updates as technology and society evolve. Second, a cross-cultural and global perspective is crucial with regard to respecting and accommodating ethical norms associated with different backgrounds. Additionally, involving a wide range of stakeholders—data providers, developers, policymakers, educators, students, and patients—is vital if their needs and concerns are to be reflected. Finally, strengthening ethical education and training is essential for stakeholders' ability to understand and apply these principles. In the future, adhering to a unified ethical framework can establish a safe, transparent, fair, and high-quality environment to support the development of medical education.

Conclusions

This paper explores the complex challenges that LLMs may entail with regard to the future of medical education, and it covers the entire spectrum ranging from the development of such technologies to their practical application. It examines the limitations of current AI-related legal and ethical frameworks with regard to guiding LLM use in the context of medical education and advocates for the development of a unified ethical framework specifically for this purpose. This paper highlights 8 fundamental principles and detailed implementation measures necessary to create such a framework: privacy and data protection, data quality control and supervision, fairness and equal treatment, transparency and interpretability, academic integrity and moral norms, accountability and traceability, IP protection, and educational research and innovation. These principles provide a solid foundation for the development of a comprehensive and actionable ethical framework. Adhering to this unified ethical framework can offer clear guidance concerning the application of LLMs in medical education. Future research should refine and implement this ethical framework in the context of a rapidly evolving technological landscape, including by facilitating dynamic updates in light of new technologies, ensuring global cross-cultural applicability, engaging diverse stakeholders, and strengthening ethical education and training. These efforts aim to balance technological advancement with ethical values, thus, promoting progress in medical education without compromising the principles of fairness, equity, or patient safety with the goal of ensuring higher quality and greater inclusivity.

XSL•FO
RenderX

Acknowledgments

We used the cross-language capabilities of ChatGPT-4 and ChatGPT-4o to enhance the language used in this paper. The revisions were reviewed by all the authors and subsequently refined and polished by American Journal Experts to ensure publication quality. We thank Ms Ding Ding for her assistance in the literature search.

Authors' Contributions

LZ and RW contributed to the conceptual design, data collection, drafting, final revision, and final submission of this paper. LFH, WXH, and FQN contributed to the data collection, language correction, and revision of this paper. All the authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Artificial intelligence–related legal and ethical principles.
[DOCX File , 18 KB - jmir_v26i1e60083_app1.docx]

References

1. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023;9:e50945 [FREE Full text] [doi: 10.2196/50945] [Medline: 37578830]
2. Xu X, Wang X, Zhang Y, Zheng R. Applying ChatGPT to tackle the side effects of personal learning environments from learner and learning perspective: an interview of experts in higher education. PLoS One 2024;19(1):e0295646 [FREE Full text] [doi: 10.1371/journal.pone.0295646] [Medline: 38170691]
3. Kavadella A, Dias da Silva MA, Kaklamanos E, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2023;10:e51344 [FREE Full text] [doi: 10.2196/51344] [Medline: 38111256]
4. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
5. Beauchamp T, Childress J. Principles of biomedical ethics: marking its fortieth anniversary. Am J Bioeth 2019;19(11):9-12. [doi: 10.1080/15265161.2019.1665402] [Medline: 31647760]
6. Busch F, Adams LC, Bressem KK. Biomedical ethical aspects towards the implementation of artificial intelligence in medical education. Med Sci Educ 2023;33(4):1007-1012 [FREE Full text] [doi: 10.1007/s40670-023-01815-x] [Medline: 37546190]
7. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One 2022;17(7):e0271872 [FREE Full text] [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
8. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021;7(4):e31043 [FREE Full text] [doi: 10.2196/31043] [Medline: 34898458]
9. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
10. Jegorova M, Kaul C, Mayor C, O'Neil AQ, Weir A, Murray-Smith R, et al. Survey: leakage and privacy at inference time. IEEE Trans Pattern Anal Mach Intell 2023;45(7):9090-9108 [FREE Full text] [doi: 10.1109/TPAMI.2022.3229593] [Medline: 37015684]
11. Rocher L, Hendrickx JM, de Montjoye Y. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019;10(1):3069 [FREE Full text] [doi: 10.1038/s41467-019-10933-3] [Medline: 31337762]
12. Li J. Security implications of AI Chatbots in health care. J Med Internet Res 2023;25:e47551 [FREE Full text] [doi: 10.2196/47551] [Medline: 38015597]
13. Data Usage for Consumer Services FAQ. URL: https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq [accessed 2024-04-18]
14. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
15. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
16. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based

XSL•FO
RenderX

dentistry: comparative mixed methods study. J Med Internet Res 2023;25:e51580 [FREE Full text] [doi: 10.2196/51580] [Medline: 38009003]
17. Zúñiga Salazar G, Zúñiga D, Vindel CL, Yoong AM, Hincapie S, Zúñiga B, et al. Efficacy of AI chats to determine an emergency: a comparison between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus 2023;15(9):e45473 [FREE Full text] [doi: 10.7759/cureus.45473] [Medline: 37727841]
18. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
19. Xue E, Bracken-Clarke D, Iannantuono GM, Choo-Wosoba H, Gulley JL, Floudas CS. Utility of large language models for health care professionals and patients in navigating hematopoietic stem cell transplantation: comparison of the performance of ChatGPT-3.5, ChatGPT-4, and Bard. J Med Internet Res 2024;26:e54758 [FREE Full text] [doi: 10.2196/54758] [Medline: 38758582]
20. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
21. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6(1):e12-e22 [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
22. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput 2021;26:232-243 [FREE Full text] [Medline: 33691020]
23. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024;30(2):80-90 [FREE Full text] [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
24. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
25. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin A, Park H, et al. AI in Medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023;330(9):818-820. [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
26. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
27. Patino GA, Amiel JM, Brown M, Lypson ML, Chan TM. The promise and perils of artificial intelligence in health professions education practice and scholarship. Acad Med 2024;99(5):477-481. [doi: 10.1097/ACM.0000000000005636] [Medline: 38266214]
28. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
29. Maida E, Moccia M, Palladino R, Borriello G, Affinito G, Clerico M, Digital Technologies, Web, Social Media Study Group of the Italian Society of Neurology (SIN). ChatGPT vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. J Neurol 2024;271(7):4057-4066. [doi: 10.1007/s00415-024-12328-x] [Medline: 38568227]
30. Guidi C, Traversa C. Empathy in patient care: from 'clinical empathy' to 'empathic concern'. Med Health Care Philos 2021;24(4):573-585 [FREE Full text] [doi: 10.1007/s11019-021-10033-4] [Medline: 34196934]
31. Koranteng E, Rao A, Flores E, Lev M, Landman A, Dreyer K, et al. Empathy and equity: key considerations for large language model adoption in health care. JMIR Med Educ 2023;9:e51199 [FREE Full text] [doi: 10.2196/51199] [Medline: 38153778]
32. Tsouroufli M, Malcolm I. Equality, diversity and fairness in medical education: international perspectives. Med Educ 2015;49(1):4-6. [doi: 10.1111/medu.12601] [Medline: 25545562]
33. Biswas S. ChatGPT and the future of medical writing. Radiology 2023;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
34. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
35. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
36. Gao CA, Howard FM, Dyer EC, Ramesh S, Luo Y. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
37. Sibel T, Karmina K, Abdullahi Y. Retracted: exploring new optical solutions for nonlinear Hamiltonian amplitude equation via two integration schemes. Phys. 2023. URL: https://iopscience.iop.org/article/10.1088/1402-4896/aceb40/meta [accessed 2024-04-30]

XSL•FO
RenderX

38. Zhang M, Wu L, Yang T, Zhu B, Liu Y. The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries. Surfaces and Interfaces 2024;46:104081. [doi: 10.1016/j.surfin.2024.104081]
39. Patil M, Sidramappa A, Shetty SK, Hebbale AM. Experimental study of solar PV/T panel to increase the energy conversion efficiency by air cooling. Materials Today: Proceedings 2023;92:309-313. [doi: 10.1016/j.matpr.2023.05.007]
40. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
41. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. Cardiol J 2024;31(3):442-450 [FREE Full text] [doi: 10.5603/cj.97517] [Medline: 37830257]
42. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
43. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, Pérez-Sancristóbal I, Pato-Cour E, Plasencia-Rodríguez C, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 2023;13(1):22129 [FREE Full text] [doi: 10.1038/s41598-023-49483-6] [Medline: 38092821]
44. Jeyaraman M, Jeyaraman N, Nallakumarasamy A, Yadav S, Bondili SK. ChatGPT in medical education and research: a boon or a bane? Cureus 2023;15(8):e44316 [FREE Full text] [doi: 10.7759/cureus.44316] [Medline: 37779749]
45. Zhang J, Zhang Z. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023;23(1):7 [FREE Full text] [doi: 10.1186/s12911-023-02103-9] [Medline: 36639799]
46. Terms of Use. URL: https://openai.com/policies/terms-of-use [accessed 2024-04-30]
47. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023;5(3):e105-e106 [FREE Full text] [doi: 10.1016/S2589-7500(23)00019-5] [Medline: 36754725]
48. ChatGPT Generative Pre-trained Transformer, Zhavoronkov A. Rapamycin in the context of pascal's wager: generative pre-trained transformer perspective. Oncoscience 2022;9:82-84 [FREE Full text] [doi: 10.18632/oncoscience.571] [Medline: 36589923]
49. Marchandot B, Matsushita K, Carmona A, Trimaille A, Morel O. ChatGPT: the next frontier in academic writing for cardiologists or a pandora's box of ethical dilemmas. Eur Heart J Open 2023;3(2):oead007 [FREE Full text] [doi: 10.1093/ehjopen/oead007] [Medline: 36915398]
50. General Data Protection Regulation. URL: https://gdpr-info.eu/ [accessed 2024-04-18]
51. Personal Information Protection Law. URL: https://www.gov.cn/xinwen/2021-08/20/content_5632486.htm [accessed 2024-04-18]
52. Health Insurance Portability and Accountability Act. URL: https://www.hhs.gov/hipaa/index.html [accessed 2024-04-18]
53. Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems. URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e_general_principles.pdf [accessed 2024-04-18]
54. European Union's White Paper on Artificial Intelligence. URL: https://www.ouvrirlascience.fr/wp-content/uploads/2020/03/commission-white-paper-artificial-intelligence-feb2020_en.pdf [accessed 2024-04-18]
55. Recommendations on the Ethics of Artificial Intelligence. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137 [accessed 2024-04-18]
56. Ethics and Governance of AI in Health: Guidance on Large Multi-modal Models. URL: https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf [accessed 2024-04-18]
57. Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development. URL: https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf?token=TozR3GWFrppLwXOHMd&fe=true [accessed 2024-04-18]
58. El Emam K. Methods for the de-identification of electronic health records for genomic research. Genome Med 2011;3(4):25 [FREE Full text] [doi: 10.1186/gm239] [Medline: 21542889]
59. Amanat A, Rizwan M, Maple C, Zikria YB, Almadhor AS, Kim SW. Blockchain and cloud computing-based secure electronic healthcare records storage and sharing. Front Public Health 2022;10:938707 [FREE Full text] [doi: 10.3389/fpubh.2022.938707] [Medline: 35928494]
60. de Beijer IAE, van den Oever SR, Charalambous E, Cangioli G, Balaguer J, Bardi E, PanCareSurPass Consortium. IT-related barriers and facilitators to the implementation of a new European eHealth solution, the digital survivorship passport (SurPass version 2.0): semistructured digital survey. J Med Internet Res 2024;26:e49910 [FREE Full text] [doi: 10.2196/49910] [Medline: 38696248]
61. Cao Y, Shi Q, Shen L, Chen K, Wang Y, Zeng W, et al. NFTracer: tracing NFT impact dynamics in transaction-flow substitutive systems with visual analytics. IEEE Trans Vis Comput Graph 2024. [doi: 10.1109/TVCG.2024.3402834] [Medline: 38768002]

XSL•FO
RenderX

62. King MR, chatGPT. A conversation on artificial intelligence, Chatbots, and plagiarism in higher education. Cell Mol Bioeng 2023;16(1):1-2 [FREE Full text] [doi: 10.1007/s12195-022-00754-8] [Medline: 36660590]
63. Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM, et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. Cell Rep Med 2024;5(1):101356 [FREE Full text] [doi: 10.1016/j.xcrm.2023.101356] [Medline: 38232690]

Abbreviations

AI: artificial intelligence
IP: intellectual property
LLM: large language model

Edited by A Mavragani; submitted 30.04.24; peer-reviewed by MO Khursheed; comments to author 10.06.24; accepted 06.07.24; published 01.08.24.

Please cite as:
Zhui L, Fenghe L, Xuehu W, Qining F, Wei R
Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint
J Med Internet Res 2024;26:e60083
URL: https://www.jmir.org/2024/1/e60083
doi:10.2196/60083
PMID:38971715

©Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, Ren Wei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Evaluating Artificial Intelligence in Clinical Settings—Let Us Not Reinvent the Wheel

Kathrin Cresswell[1], PhD; Nicolette de Keizer[2,3], PhD; Farah Magrabi[4], PhD; Robin Williams[5], PhD; Michael Rigby[6], PhD; Mirela Prgomet[7], PhD; Polina Kukhareva[8], PhD; Zoie Shui-Yee Wong[9], PhD; Philip Scott[10], PhD; Catherine K Craven[11], PhD; Andrew Georgiou[4], PhD; Stephanie Medlock[2,12], PhD; Jytte Brender McNair[13], PhD; Elske Ammenwerth[14], PhD

[1]Usher Institute, The University of Edinburgh, Usher Building, Edinburgh, United Kingdom
[2]Amsterdam UMC, University of Amsterdam, Medical Informatics, Amsterdam, Netherlands
[3]Amsterdam Public Health Research Institute, Digital Health and Quality of Care, Amsterdam, Netherlands
[4]Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[5]Institute for the Study of Science, Technology and Innovation, The University of Edinburgh, Edinburgh, United Kingdom
[6]School of Social, Political and Global Studies and School of Primary, Community and Social Care, Keele University, Keele, United Kingdom
[7]Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia
[8]Department of Biomedical Informatics, University of Utah, Utah, UT, United States
[9]St. Luke's International University, Tokyo, Japan
[10]University of Wales Trinity St David, Swansea, United Kingdom
[11]University of Texas Health Science Center, San Antonio, TX, United States
[12]Amsterdam Public Health, Methodology & Aging & Later Life, Amsterdam, Netherlands
[13]Department of Health Science and Technology, Aalborg University, Aalborg, Denmark
[14]Institute of Medical Informatics, Private University for Health Sciences and Health Technology, UMIT TIROL, Hall in Tirol, Austria

Corresponding Author:
Kathrin Cresswell, PhD
Usher Institute
The University of Edinburgh
Usher Building
5-7 Little France Road
Edinburgh, EH16 4UX
United Kingdom
Phone: 44 131 650 6984
Email: kathrin.cresswell@ed.ac.uk

Abstract

Given the requirement to minimize the risks and maximize the benefits of technology applications in health care provision, there is an urgent need to incorporate theory-informed health IT (HIT) evaluation frameworks into existing and emerging guidelines for the evaluation of artificial intelligence (AI). Such frameworks can help developers, implementers, and strategic decision makers to build on experience and the existing empirical evidence base. We provide a pragmatic conceptual overview of selected concrete examples of how existing theory-informed HIT evaluation frameworks may be used to inform the safe development and implementation of AI in health care settings. The list is not exhaustive and is intended to illustrate applications in line with various stakeholder requirements. Existing HIT evaluation frameworks can help to inform AI-based development and implementation by supporting developers and strategic decision makers in considering relevant technology, user, and organizational dimensions. This can facilitate the design of technologies, their implementation in user and organizational settings, and the sustainability and scalability of technologies.

(J Med Internet Res 2024;26:e46407) doi:10.2196/46407

KEYWORDS

artificial intelligence; evaluation; theory; patient safety; optimisation; health care; optimization

XSL•FO
RenderX

Introduction

The last two decades have seen rapid growth in artificial intelligence (AI) initiatives in health care settings, driven by the promises of improved treatment, quality, safety, and efficiency [1]. AI systems are computer algorithms that are able to mimic human intelligence to perform tasks. They are potentially capable of improving clinical decision-making. However, there is currently a lack of high-quality evidence of effectiveness, and an overoptimism regarding AI-based technologies in health care [2,3]. Many existing algorithms and applications fail to scale and migrate across settings [4], potentially leading to missed benefits or compromised patient safety.

Evidence from other sectors, such as finance and retail, may have limited applicability given the particular social, economic, technical processes, and legal challenges of health and social care settings [5]. Across the digital economy, AI has been successfully applied to historical data, for example, in financial forecasting [6] or retail marketing, where personalized advertisements have transformed consumer behavior [7]. These methods are harder to deploy in the more complex and sensitive settings of health and social care [5]. This is largely because developers and implementers focus on tool development and do not sufficiently draw on existing work to inform the conception and design of technologies, their use and optimization, and organizational strategies to implement them.

Theory-informed approaches to evaluation can help to ensure that technologies are effectively validated, implemented, and adopted. They can also help to ensure that systems do not result in unintended negative consequences, such as inappropriate or suboptimal care, exacerbated inequities, or clinician burnout [8]. Theories seek to explain complex relationships at an abstract level and can help to integrate a particular implementation with the empirical evidence base. As such, theory-informed evaluation frameworks can enable learning from experience, thus guiding developers, implementers, and evaluators through development, implementation, and optimization [9]. Ideally, the real-world experience gathered during this process is then used also to inform the refinement of evaluation frameworks.

Despite significant investments, there are currently only a few examples of the use of AI-based systems in health care and most systems are only beginning to be rolled out and embedded [10-12]. This is in contrast to the finance and retail sectors, where processes and products are standardized. To date, most activity has focused on diagnostic image-based systems and text or language processing, while complex precision medicine efforts are in very early stages of development. We here call for the increasing use of theory-informed approaches to evaluation to help ensure that developed systems can be adopted, scaled, and sustained within settings of use, and are safe and effective. Until now, this has not been done consistently, which has resulted in limited learning and limited ability to transfer learning across settings, as well as limited clinical and patient reassurance. If done appropriately, the implications for clinical settings are significant, as validated new knowledge can be disseminated and shared. This, in turn, obviates the need to learn through experience that can be painful, dangerous, and costly.

Unfortunately, despite increasing attention in research, the current application of theory-informed strategy and evaluation in AI practice is relatively limited in both health care and other sectors [13]. This may be due to a lack of understanding surrounding the theoretical literature (ie, why theories are useful in practice and how they may be used by different stakeholders), and the immediate focus of developers on demonstrating that technology works. Politically and managerially, there may be a drive to show modernization processes rather than making clinical and organizational decisions based on evidence-based outcomes. Where theories have been applied, these have been driven by business approaches to value creation in organizations [14], or by approaches designed to influence consumer behavior [15]. In these contexts, they have been strategically used to help address a particular stakeholder need (eg, how to maximize value through implementing AI in organizations and how to get consumers to accept AI technology). In health care, the range of stakeholders and associated needs however varies significantly from other sectors. While the managers and policymakers may focus on value and efficiency, patients are likely to be concerned about avoidable illness, and practitioners may focus on workloads and potential liability.

It is therefore often difficult to know what needs (and consequently what theory) to focus on and in what context. For example, while developers of technology now increasingly draw on cocreation with users to promote the adoption of AI, these approaches may not consider organizational drivers, workflow integration, multiplicity of stakeholders, or ethical considerations in implementation, thereby limiting the scalability of emerging applications.

Theory-informed approaches to evaluation in health care must be considered within their specific context, recognizing their relative positions and identifying which needs they address at various stages of the technology lifecycle. We aim to begin this journey by providing a conceptual overview of existing theory-informed frameworks that could usefully inform the development and implementation of AI-based technologies in health care. Despite some differences in technological properties and performance between AI- and non–AI-based technologies (Table 1) [16], many existing frameworks are likely to be applicable.

XSL•FO
RenderX

Table 1. Differences between artificial intelligence (AI)–based and non–AI-based health IT.

Applications	AI-based	Evidence	Non–AI-based	Evidence
Health services management	AI can help in optimizing resource allocation, scheduling, and workflow management by analyzing large data sets and identifying patterns and trends. For example, modeling of waiting times and underlying reasons	Limited evidence in relation to impact, mainly in relation to proof-of-concept [17,18]	Non–AI-based approaches typically rely on manual processes and human decision-making for resource management, scheduling, and workflow optimization. For example, patient flow management applications	High potential of data-driven approaches to improve organizational performance [19,20]
Predictive medicine	AI algorithms can analyze patient data, genetic information, and medical records to predict disease risks, treatment outcomes, and responses to therapies. This enables personalized medicine and targeted interventions	Many proof-of-concept studies but limited evidence in relation to how outputs are incorporated into clinical decision-making [21,22]	Non–AI-based approaches rely on statistical analysis and clinical expertise to make predictions about disease risks, treatment outcomes, and responses to therapies	Many proof-of-concept studies but limited evidence in relation to how outputs are incorporated into clinical decision-making [23,24]
Clinical decision support systems	AI to analyze large amounts of medical literature, patient data, and clinical guidelines to support clinical decision-making	Area of most focus, especially in imaging applications, AI has the potential to improve practitioner performance [25,26], but limited evidence surrounding organizational impacts or patient outcomes	Non–AI-based approaches rely on the expertise and experience of health care professionals, along with clinical guidelines and published research, to make clinical decisions	Demonstrated benefits for practitioner performance and patient outcomes in some areas of use (eg, drug-drug interactions) [27,28]
Laboratory and radiology information systems	Use of AI to detect abnormalities and to enhance the accuracy of diagnoses	Most progress has been made in relation to imaging [29,30], but limited attention has been paid to integration with organizational practices as above [26]	Non–AI-based diagnostics typically rely on visual inspection by health care professionals and manual analysis of patient data	This is associated with information overload but does take account of contextual factors
Patient data repositories	AI algorithms can process patient data to identify trends, patterns, and risk factors	Promising proof-of-concept studies, but limited implementation [31,32]	Patient data are stored in a centralized repository	Some evidence that digitized records and repositories can lead to improved quality, safety, and efficiency, but hard to assess and take a long time to materialize [33,34]
Population health management	Precision prevention approaches to identify populations at risk and tailor preventative interventions	Promising approaches to precision prevention in specific cohorts, but limited implementation [35,36]	Understanding factors that influence health outcomes and developing tailored interventions	Significant evidence of population health interventions [37,38]
Patient portals	AI-based symptom checkers and triage tools	Inconsistent evidence in relation to symptom checkers and triage tools, concerns in relation to diagnostic accuracy [39,40]	Access to generic informational resources	Tailored informational resources can improve satisfaction, involvement, and decision-making [41,42]
Telehealth and telecare	Online health assistants and chatbots	Mixed evidence of effectiveness usability and user satisfaction [43,44]	Access to generic informational resources.	Tailored informational resources can improve satisfaction, involvement, and decision-making [41,42]
Health information exchange	Extracting and converting unstructured or semistructured data into a standardized format	The use of free text data is still in its infancy but is promising. There is limited data on integration with existing ways of working and organizational functioning. [45,46]	Coding and transfer into standardized formats are often done by health care staff	Increased workloads for health care staff and coding are often not done accurately [47,48]

We here provide a conceptual overview of existing frameworks, focusing on practical applications of examples of existing theory-informed frameworks and their potential application to AI-based technologies in health care [49]. Frameworks were selected as examples illustrating these extracted categories. This work is not intended to be exhaustive but to provide a pragmatic introduction to the topic for nonspecialists [50,51].

To categorize frameworks in a meaningful way, we focused on their potential area of application and the particular interest or focus of various stakeholder groups who may need to draw on existing experience to inform their current efforts to develop, implement, and optimize AI-based technologies in health care settings.

XSL•FO
RenderX

Health IT Evaluation Frameworks and Their Potential Application to AI

Overview

The 3 distinct dimensions identified are illustrated in Table 2, along with potential applications of AI-based technologies in health care and example use cases. These include frameworks with a technology, user, and organizational focus. We discuss each of these categories, the application of exemplary frameworks, and practical implications for various stakeholders in the paragraphs below.

However, it is important to recognize that the categorization of frameworks provided here is a simplification. Various frameworks have common and, in some instances, overlapping elements. The categories presented are intended to facilitate navigation and application.

Table 2. Examples of the focus of existing health IT evaluation frameworks and their potential application to artificial intelligence.

Focus of the framework	Area of application	Example theoretical lenses	Practical implications	Stakeholders	Examples
Technology focus	• Informing the conception and design of technologies • To help AI[a] system developers design a system that is usable and useful within intended use settings	Human-centered design	Actively and iteratively involve end users in system design and development	End users and developers	A team had developed an algorithm to predict arterial fibrillation from electrocardiograms, but prospective users stated that the information would not change their practice [10]
User focus	• Informing and helping to optimize the use of technologies • To help developers and implementers understand the various contexts of use of AI as well as unintended consequences, and tailor systems to maximize benefits and minimize harms	Sociotechnical systems	Plan with users to effectively integrate the system in their work practices and monitor progress over time	End users, implementers	• IBM Watson encountered adoption-related issues, including usability and perceived usefulness of their oncology software, which eventually led to the abandonment • The system increased the workloads of doctors and made treatment recommendations that were viewed as unsafe by doctors [52]
Organizational focus	• Informing organizational strategies to implement technologies • To help AI system implementers integrate AI safely within existing organizational structures and processes	Institutional theory	Plan and monitor how systems and their outputs are integrated within and across organizational units and existing technological and social structures	End users, organizational stakeholders, and implementers	• Babylon Health UK (an AI-based remote service provider) failed because it did not fit with existing health system financing structures and cultures • Many patients from outside the local area enrolled in the service, which meant that the product was not commercially viable for local organizations [53]

[a]AI: artificial intelligence.

Frameworks With a Technology Focus

Many current AI applications in health care settings have been developed by AI specialists in laboratory settings. Consequently, they have struggled to successfully translate into clinical settings and deliver the performance achieved in research trials [54]. Frameworks with a technology focus can help to inform the "conception and design" of technologies, thereby helping to ensure that AI system developers design a system that is readily implemented and useful within intended use settings. For instance, techniques such as technology assessment and requirements analysis can help to identify use cases, constraints, and requirements that the new technology needs to fulfill.

Frameworks include, for example, design and usability frameworks such as the Health IT Usability Evaluation Model (Health-ITUEM) for evaluating mobile health technology [55]. This includes assessment of subjective properties of the technology from the perspective of users, which have been shown to be crucial to user adoption of technology, but that developers may not necessarily consider as a priority during the development process, including ease of use and perceived usefulness.

Frameworks With a User Focus

While use is crucial for the successful development of AI-based technology, empirical work has shown that systems may be

XSL•FO
RenderX

used in ways other than intended, which may in turn result in unanticipated threats to organizational functioning and patient safety [56]. For example, users may develop workarounds to compensate for usability issues of technologies, but these workarounds may compromise the intended performance of a system [57]. Frameworks that focus on the user of the technology can help to address these issues and facilitate the "optimization of technology use". In doing so, they can help developers and implementers understand the various contexts of the use of AI-based technologies, as well as unintended consequences, and tailor systems to maximize benefits and minimize harms. For instance, a contextual analysis can help to gain a deep understanding of the various contexts in which a technology will be deployed. This includes examining cultural and social factors, as well as user behavior, user expectations, and existing systems or practices.

An example framework in this context is the Health Information Technology Evaluation Framework (HITREF), which includes an assessment of a technology's impact on quality of care as well as an assessment of unintended consequences [58].

Frameworks With an Organizational Focus

AI-based technologies are not adopted in a vacuum but must be integrated within organizational contexts. Previous work has shown that organizational strategies to implement health IT (HIT) and organizational cultures can have significant consequences for adoption and use [59]. For example, lack of integration with existing health information infrastructures can slow down system performance and impede practical use, and hence, impact adversely on safety and user experience [60]. Frameworks with an organizational focus can facilitate the development of "organizational strategies" to implement new technologies. In doing so, they can help AI system implementers integrate AI safely within existing organizational structures and processes. For instance, these can help to inform communication strategies, training programs, and support mechanisms to help users understand the benefits and risks of AI technologies and adapt to new roles and responsibilities.

An example of a framework with an organizational focus is the Safety Assurance Factors for Electronic Health Record Resilience (SAFER) guides, which help implementing organizations identify existing risks and facilitate the development of mitigation strategies to promote the effective integration of technologies within organizational processes [61].

Discussion

A range of theory-informed evaluation frameworks for diverse kinds of HIT already exist [62]. Although not all of these may be relevant for AI-based applications, many aspects of existing frameworks are likely to apply. Exploring the transferability of these dimensions, therefore, needs to be a central component of work going forward [63].

Existing frameworks examine various aspects of technology design, implementation, adoption, and optimization. On the most basic level, they can be distinguished according to their focus, which then influences their application and context of use. A simplified overview of selected HIT evaluation frameworks and their potential application to AI is shown in Table 2. Frameworks with a technology focus can help to inform the conception and design of technologies through actively and iteratively involving end users, bridging the gap between technology development and application. This can, in turn, mitigate risks around nonadoption due to a lack of need or actionable system outputs. Frameworks with a user focus can help to ensure that systems are effectively embedded with adoption contexts and thereby mitigate the risk of systems not being used or not being used as intended. Finally, frameworks with an organizational focus can help to ensure that systems fit with existing organizational structures, and thereby help to ensure sustained use over time and across contexts.

We recommend that researchers, implementers, and strategic decision makers consider the use of existing theory-informed HIT evaluation frameworks before embarking on an AI-related initiative. This can help to mitigate emerging risks and maximize the chances of successful implementation, adoption, and scaling. To achieve this, existing and emerging guidelines for the evaluation of AI must promote the use of theory-informed evaluation frameworks.

Although many of the frameworks are well-known in the academic clinical informatics community, there is an urgent need to incorporate them into general AI design, implementation, and evaluation activities, as they can help to facilitate learning from experience and ensure building on the existing empirical evidence base. Unfortunately, this is currently not routinely done, perhaps reflecting disciplinary silos leading to lessons having to be learned the hard way. This, in turn, potentially compromises the safety, quality, and sustainability of applications. For example, although AI applications in radiology are now getting more established, the existing evidence base focuses on demonstrating effectiveness in proof-of-concept or specific clinical settings (the technology dimension in Table 2) [25]. Wider organizational and user factors are somewhat neglected, potentially threatening the wider sustainability and acceptability of such applications.

Conclusions

We aimed to provide a conceptual overview of existing theory-informed frameworks that could usefully inform the development and implementation of AI-based technologies in health care, and we identified several frameworks with technological, user, and organizational foci. Future research could involve conducting a systematic review based on this pragmatic overview to synthesize existing evidence across evaluation frameworks, spanning the dimensions of technology, user, and organization.

Evaluation of AI-based systems needs to be based on theoretically informed empirical studies in contexts of implementation or use to ensure objectivity and rigor in establishing the benefits and thwarting risks. This will ensure that systems are based on relevant and transferable evidence and can be implemented safely and effectively. Theory-based HIT evaluation frameworks should be integrated into existing and emerging guidelines for the evaluation of AI [64-66]. The examples of frameworks provided could also help to stimulate

XSL•FO
RenderX

the development of other related frameworks that can guide further evaluation efforts.

Drawing effectively on theory-based HIT evaluation frameworks will help to strengthen the evidence-based implementation of AI systems in health care and help to refine and tailor existing theoretical approaches to AI-based HIT. Learning from the wealth of existing HIT evaluation experience will help patients, professionals, and wider health care systems.

Acknowledgments

The authors are members of the International Medical Informatics Association Working Group on Technology Assessment and Quality Development and the European Federation for Medical Informatics Working Group on Evaluation. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

KC led on drafting of the manuscript and all authors (NDK, FM, RW, MR, MP, PK, ZSW, PS, CKC, AG, SM, JBM, and EA) critically commented on various iterations.

Conflicts of Interest

None declared.

References

1. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med 2021;4(1):1-8 [FREE Full text] [doi: 10.1038/s41746-021-00423-6] [Medline: 33742085]
2. Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform 2019;28(01):128-134. [doi: 10.1055/s-0039-1677903]
3. Patel VL, Kannampallil. Cognitive informatics in biomedicine and healthcare. J Biomed Inform 2015;53:3-14. [doi: 10.1016/j.jbi.2014.12.007] [Medline: 25541081]
4. Lobo JL, Del Ser J, Bifet A, Kasabov N. Spiking neural networks and online learning: an overview and perspectives. Neural Netw 2020;121:88-100. [doi: 10.1016/j.neunet.2019.09.004] [Medline: 31536902]
5. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
6. Li X, Sigov A, Ratkin L, Ivanov LA, Li L. Artificial intelligence applications in finance: a survey. Journal of Management Analytics 2023;10(4):676-692. [doi: 10.1080/23270012.2023.2244503]
7. Weber FD, Schütte R. State-of-the-art and adoption of artificial intelligence in retailing. DPRG 2019;21(3):264-279. [doi: 10.1108/dprg-09-2018-0050]
8. Wiegand T, Krishnamurthy R, Kuglitsch M, Lee N, Pujari S, Salathé M, et al. WHO and ITU establish benchmarking process for artificial intelligence in health. Lancet 2019;394(10192):9-11. [doi: 10.1016/S0140-6736(19)30762-7] [Medline: 30935732]
9. Craven CK, Doebbeling B, Furniss D, Holden RJ, Lau F, Novak LL. Evidence-based Health Informatics Frameworks for Applied Use. Stud Health Technol Inform 2016;222:77-89. [Medline: 27198094]
10. Williams R, Anderson S, Cresswell K, Kannelønning MS, Mozaffar H, Yang X. Domesticating AI in medical diagnosis. Technology in Society 2024:102469. [doi: 10.1016/j.techsoc.2024.102469]
11. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG, Right Data Right Now Collaborators. Machine learning in intensive care medicine: ready for take-off? Intensive Care Med 2020;46(7):1486-1488. [doi: 10.1007/s00134-020-06045-y] [Medline: 32399747]
12. Mackenzie SC, Sainsbury CAR, Wake DJ. Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges. Diabetologia 2024;67(2):223-235 [FREE Full text] [doi: 10.1007/s00125-023-06038-8] [Medline: 37979006]
13. Reim W, Åström J, Eriksson O. Implementation of artificial intelligence (AI): a roadmap for business model innovation. AI 2020;1(2):180-191. [doi: 10.3390/ai1020011]
14. Enholm IM, Papagiannidis E, Mikalef P, Krogstie J. Artificial intelligence and business value: a literature review. Inf Syst Front 2021;24:1709-1734 [FREE Full text] [doi: 10.1007/s10796-021-10186-w]

XSL•FO
RenderX

15. Vlačić B, Corbo L, Costa e Silva S, Dabić M. The evolving role of artificial intelligence in marketing: A review and research agenda. Journal of Business Research 2021;128:187-203. [doi: 10.1016/j.jbusres.2021.01.055]
16. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021;21(1):125 [FREE Full text] [doi: 10.1186/s12911-021-01488-9] [Medline: 33836752]
17. Pianykh OS, Guitron S, Parke D, Zhang C, Pandharipande P, Brink J, et al. Improving healthcare operations management with machine learning. Nat Mach Intell 2020;2(5):266-273. [doi: 10.1038/s42256-020-0176-3]
18. Sebla AK. Use of artificial intelligence in health services management in Türkiye. International Journal of Health Services Research and Policy 2023;8(2):139-161. [doi: 10.33457/ijhsrp.1298068]
19. Madsen LB. Data-Driven Healthcare: How Analytics and BI are Transforming the Industry. Hoboken, NJ: John Wiley & Sons; 2014.
20. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Serv Res 2021;21(1):200 [FREE Full text] [doi: 10.1186/s12913-021-06215-8] [Medline: 33663508]
21. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577-1579. [doi: 10.1016/S0140-6736(19)30037-6] [Medline: 31007185]
22. Schork NJ. Artificial intelligence and personalized medicine. Cancer Treat Res 2019;178:265-283 [FREE Full text] [doi: 10.1007/978-3-030-16391-4_11] [Medline: 31209850]
23. Bellazzi R, Ferrazzi F, Sacchi L. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdiscip Rev Data Min Knowl Discov 2011;1(5):416-430. [doi: 10.1002/widm.23]
24. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ 2018;363:k4245 [FREE Full text] [doi: 10.1136/bmj.k4245] [Medline: 30530757]
25. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32(11):7998-8007 [FREE Full text] [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
26. Farič N, Hinder S, Williams R, Ramaesh R, Bernabeu MO, van Beek E, et al. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings: a qualitative study. J Am Med Inform Assoc 2023;31(1):24-34 [FREE Full text] [doi: 10.1093/jamia/ocad191] [Medline: 37748456]
27. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
28. Greenes R, Del FG, editors. Clinical Decision Support and Beyond Progress and Opportunities in Knowledge-Enhanced Health and Healthcare. Cambridge, MA: Academic Press; 2023.
29. Hafizović L, Čaušević A, Deumić A, Spahić Bećirović L, Gurbeta Pokvić L, Badnjević A. The use of artificial intelligence in diagnostic medical imaging: systematic literature review. 2021 Presented at: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE); 15 December 2021; Kragujevac, Serbia p. 1-6. [doi: 10.1109/bibe52308.2021.9635307]
30. Rebitzer JB, Rege M, Shepard C. Influence, information overload, and information technology in health care. Adv Health Econ Health Serv Res 2008;19:43-69. [Medline: 19548513]
31. Chen X, Liu Z, Wei L, Yan J, Hao T, Ding R. A comparative quantitative study of utilizing artificial intelligence on electronic health records in the USA and China during 2008-2017. BMC Med Inform Decis Mak 2018;18(Suppl 5):117 [FREE Full text] [doi: 10.1186/s12911-018-0692-9] [Medline: 30526643]
32. Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, et al. The secondary use of electronic health records for data mining: data characteristics and challenges. ACM Comput. Surv 2022;55(2):1-40. [doi: 10.1145/3490234]
33. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS Med 2011;8(1):e1000387 [FREE Full text] [doi: 10.1371/journal.pmed.1000387] [Medline: 21267058]
34. Bergmo TS. How to measure costs and benefits of eHealth interventions: an overview of methods and frameworks. J Med Internet Res 2015;17(11):e254 [FREE Full text] [doi: 10.2196/jmir.4521] [Medline: 26552360]
35. Loomans-Kropp HA, Umar A. Cancer prevention and screening: the next step in the era of precision medicine. NPJ Precis Oncol 2019;3:3 [FREE Full text] [doi: 10.1038/s41698-018-0075-9] [Medline: 30701196]
36. Herman WH, Ye W. Precision prevention of diabetes. Diabetes Care 2023;46(11):1894-1896. [doi: 10.2337/dci23-0052] [Medline: 37890107]
37. Kaplan RM. Two pathways to prevention. American Psychologist 2000;55(4):382-396. [doi: 10.1037//0003-066x.55.4.382]
38. Jacka FN, Mykletun A, Berk M. Moving towards a population health approach to the primary prevention of common mental disorders. BMC Med 2012;10:149 [FREE Full text] [doi: 10.1186/1741-7015-10-149] [Medline: 23186355]
39. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023;25:e43803 [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]

XSL•FO
RenderX

40. Ilicki J. Challenges in evaluating the accuracy of AI-containing digital triage systems: a systematic review. PLoS One 2022;17(12):e0279636 [FREE Full text] [doi: 10.1371/journal.pone.0279636] [Medline: 36574438]
41. Reynolds TL, Ali N, McGregor E, O'Brien T, Longhurst C, Rosenberg AL, et al. Understanding patient questions about their medical records in an online health forum: opportunity for patient portal design. : American Medical Informatics Association; 2017 Presented at: AMIA annual symposium proceedings; 16 April 2018; Washington, DC p. 1468.
42. Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ 2002;324(7339):710 [FREE Full text] [doi: 10.1136/bmj.324.7339.710] [Medline: 11909789]
43. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
44. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR Hum Factors 2022;9(4):e35882 [FREE Full text] [doi: 10.2196/35882] [Medline: 36197708]
45. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
46. Li D, Simon G, Chute CG, Pathak J. Using association rule mining for phenotype extraction from electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:142-146 [FREE Full text] [Medline: 24303254]
47. Bajaj Y, Crabtree J, Tucker AG. Clinical coding: how accurately is it done? Clinical Governance: An International Journal 2007;12(3):159-169. [doi: 10.1108/14777270710775873]
48. Campbell S, Giadresco K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. Health Inf Manag 2020;49(1):5-18. [doi: 10.1177/1833358319851305] [Medline: 31159578]
49. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
50. Salkind NJ, editor. Encyclopedia of Research Design. North America: Sage; 2010.
51. Glasgow RE. What does it mean to be pragmatic? Pragmatic methods, measures, and models to facilitate research translation. Health Educ Behav 2013;40(3):257-265. [doi: 10.1177/1090198113486805] [Medline: 23709579]
52. Faheem H, Dutta S. Artificial intelligence failure at IBM' Watson for Oncology'. IUP Journal of Knowledge Management 2023;21(3):47-75.
53. Mahase E. Babylon looks to sell GP at hand and other UK business amid financial issues. BMJ 2023;382:1835. [doi: 10.1136/bmj.p1835] [Medline: 37558246]
54. Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: a systematic review. Artif Intell Med 2020;103:101785. [doi: 10.1016/j.artmed.2019.101785] [Medline: 32143792]
55. Brown W, Yen PY, Rojas M, Schnall R. Assessment of the health IT usability evaluation model (Health-ITUEM) for evaluating mobile health (mHealth) technology. J Biomed Inform 2013;46(6):1080-1087 [FREE Full text] [doi: 10.1016/j.jbi.2013.08.001] [Medline: 23973872]
56. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc 2004;11(2):104-112 [FREE Full text] [doi: 10.1197/jamia.M1471] [Medline: 14633936]
57. Cresswell KM, Mozaffar H, Lee L, Williams R, Sheikh A. Workarounds to hospital electronic prescribing systems: a qualitative study in english hospitals. BMJ Qual Saf 2017;26(7):542-551 [FREE Full text] [doi: 10.1136/bmjqs-2015-005149] [Medline: 27129493]
58. Sockolow PS, Bowles KH, Rogers ML. Health information technology evaluation framework (HITREF) comprehensiveness as assessed in electronic point-of-care documentation systems evaluations. Stud Health Technol Inform 2015;216:406-409. [Medline: 26262081]
59. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
60. Cresswell KM, Mozaffar H, Lee L, Williams R, Sheikh A. Safety risks associated with the lack of integration and interfacing of hospital health information technologies: a qualitative study of hospital electronic prescribing systems in England. BMJ Qual Saf 2017;26(7):530-541 [FREE Full text] [doi: 10.1136/bmjqs-2015-004925] [Medline: 27037303]
61. Sittig DF, Ash JS, Singh H. The SAFER guides: empowering organizations to improve the safety and effectiveness of electronic health records. Am J Manag Care 2014;20(5):418-423 [FREE Full text] [Medline: 25181570]
62. Scott P, de KN, Georgiou A, editors. Applied Interdisciplinary Theory in Health Informatics: A Knowledge Base for Practitioners. Netherlands: IOS Press; 2019.
63. Schloemer T, Schröder-Bäck P. Criteria for evaluating transferability of health interventions: a systematic review and thematic synthesis. Implement Sci 2018;13(1):88 [FREE Full text] [doi: 10.1186/s13012-018-0751-8] [Medline: 29941011]

XSL•FO
RenderX

64. Nykänen P, Brender J, Talmon J, de Keizer N, Rigby M, Beuscart-Zephir MC, et al. Guideline for good evaluation practice in health informatics (GEP-HI). Int J Med Inform 2011;80(12):815-827. [doi: 10.1016/j.ijmedinf.2011.08.004] [Medline: 21920809]
65. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI -statement on reporting of evaluation studies in health informatics. Yearb Med Inform 2009;78(1):23-31. [Medline: 19855867]
66. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]

Abbreviations

AI: artificial intelligence
Health-ITUEM: Health IT Usability Evaluation Model
HIT: health IT
HITREF: Health Information Technology Evaluation Framework
SAFER: Safety Assurance Factors for Electronic Health Record Resilience

Edited by T Leung; submitted 10.02.23; peer-reviewed by P Aovare, M Yusof, D Chrimes; comments to author 14.04.23; revised version received 20.04.23; accepted 02.03.24; published 07.08.24.

Please cite as:
Cresswell K, de Keizer N, Magrabi F, Williams R, Rigby M, Prgomet M, Kukhareva P, Wong ZSY, Scott P, Craven CK, Georgiou A, Medlock S, Brender McNair J, Ammenwerth E
Evaluating Artificial Intelligence in Clinical Settings—Let Us Not Reinvent the Wheel
J Med Internet Res 2024;26:e46407
URL: https://www.jmir.org/2024/1/e46407
doi:10.2196/46407
PMID:39110494

©Kathrin Cresswell, Nicolette de Keizer, Farah Magrabi, Robin Williams, Michael Rigby, Mirela Prgomet, Polina Kukhareva, Zoie Shui-Yee Wong, Philip Scott, Catherine K Craven, Andrew Georgiou, Stephanie Medlock, Jytte Brender McNair, Elske Ammenwerth. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies

Tianhui Xu[1,2], MSN, RN; Huiting Weng[1], MSN, RN; Fang Liu[1], MSN, RN, PhD; Li Yang[1], MSN, RN; Yuanyuan Luo[2], MSN, RN; Ziwei Ding[2], MSN, RN; Qin Wang[1,2], MSN, RN

[1]Clinical Nursing Teaching and Research Section, The Second Xiangya Hospital of Central South University, Changsha, China
[2]Xiangya School of Nursing, Central South University, Changsha, China

Corresponding Author:
Qin Wang, MSN, RN
Clinical Nursing Teaching and Research Section
The Second Xiangya Hospital of Central South University
139 Middle Renmin Road
Changsha, 410011
China
Phone: 86 18774806226
Email: wangqin3421@csu.edu.cn

Abstract

ChatGPT, a generative pretrained transformer, has garnered global attention and sparked discussions since its introduction on November 30, 2022. However, it has generated controversy within the realms of medical education and scientific research. This paper examines the potential applications, limitations, and strategies for using ChatGPT. ChatGPT offers personalized learning support to medical students through its robust natural language generation capabilities, enabling it to furnish answers. Moreover, it has demonstrated significant use in simulating clinical scenarios, facilitating teaching and learning processes, and revitalizing medical education. Nonetheless, numerous challenges accompany these advancements. In the context of education, it is of paramount importance to prevent excessive reliance on ChatGPT and combat academic plagiarism. Likewise, in the field of medicine, it is vital to guarantee the timeliness, accuracy, and reliability of content generated by ChatGPT. Concurrently, ethical challenges and concerns regarding information security arise. In light of these challenges, this paper proposes targeted strategies for addressing them. First, the risk of overreliance on ChatGPT and academic plagiarism must be mitigated through ideological education, fostering comprehensive competencies, and implementing diverse evaluation criteria. The integration of contemporary pedagogical methodologies in conjunction with the use of ChatGPT serves to enhance the overall quality of medical education. To enhance the professionalism and reliability of the generated content, it is recommended to implement measures to optimize ChatGPT's training data professionally and enhance the transparency of the generation process. This ensures that the generated content is aligned with the most recent standards of medical practice. Moreover, the enhancement of value alignment and the establishment of pertinent legislation or codes of practice address ethical concerns, including those pertaining to algorithmic discrimination, the allocation of medical responsibility, privacy, and security. In conclusion, while ChatGPT presents significant potential in medical education, it also encounters various challenges. Through comprehensive research and the implementation of suitable strategies, it is anticipated that ChatGPT's positive impact on medical education will be harnessed, laying the groundwork for advancing the discipline and fostering the development of high-caliber medical professionals.

(J Med Internet Res 2024;26:e57896) doi:10.2196/57896

KEYWORDS

chat generative pretrained transformer; ChatGPT; artificial intelligence; medical education; natural language processing; clinical practice

Introduction

Artificial intelligence (AI) is the simulation of human cognitive capacities using computer programming, allowing robots to emulate human thought and behavior [1]. AI generation is the automated creation of various content formats, such as text, photos, video, and audio, using AI technologies. This approach generates content using language, visuals, and multimodal macro models [2]. Among these technologies, ChatGPT stands out as a sophisticated, large-scale language model developed by

XSL•FO
RenderX

OpenAI, which has reached stage 4.0, the most recent iteration of the OpenAI system. ChatGPT, which has been trained on large amounts of textual data, is designed to participate in conversational exchanges with users while responding contextually to their prompts [3]. Particularly, ChatGPT has advanced capabilities in natural language processing, logical reasoning, task execution, information retrieval, picture analysis, content development, and other areas [4]. Moreover, ChatGPT provides a vast range of services accessible through global registration, facilitating its integration across various domains.

In the field of education, the introduction of ChatGPT has generated considerable interest and prompted in-depth discussions. Its exceptional generating capabilities offer new avenues for scholarly research, increased learning, classroom improvement, and knowledge sharing [5]. The incorporation of ChatGPT into medical education has become a major focus, with the goal of combining the advancements of education and medicine. In order to provide a well-informed assessment of the potential and limitations of integrating ChatGPT into medical education, this research aims to examine both the technology's capabilities and the challenges it may present. Furthermore, tactics are put out to support the smooth integration of technology and medicine.

Status of ChatGPT in Medical Education

In this study, we searched the Web of Science database using specific terms related to ChatGPT and education or medical science within the timeframe of January 2022 to December 2024. The search formula is ("Chat generative pretrained transformer" OR "chat GPT" OR ChatGPT OR chat- GPT OR GPT-3.5 OR GPT-4.0") AND (Education OR educate OR educator OR Students, Medical OR medical science OR medicine OR health care OR health science). We used CiteSpace v6.1R6 (64-bit) Basic to analyze the posted keywords [6]. The basic parameters used in CiteSpace were configured as follows: the time partition ranged from January 2022 to December 2024, with a one-time slice; the node selection criterion was set to k=25 for the g-index in each time slice, while the remaining parameters were kept at default settings. The top 7 keywords, ranked by frequency in the keyword network, are natural language processing, impact, academic integrity, AI in medicine, writing, health literacy, and health care professionals (Figure 1), and the top 10 countries by publication volume are United States, India, China, Australia, England, Germany, Canada, Italy, Spain, and United Arab Emirates (Figure 2).

Figure 1. Keywords of ChatGPT in medicine and education.

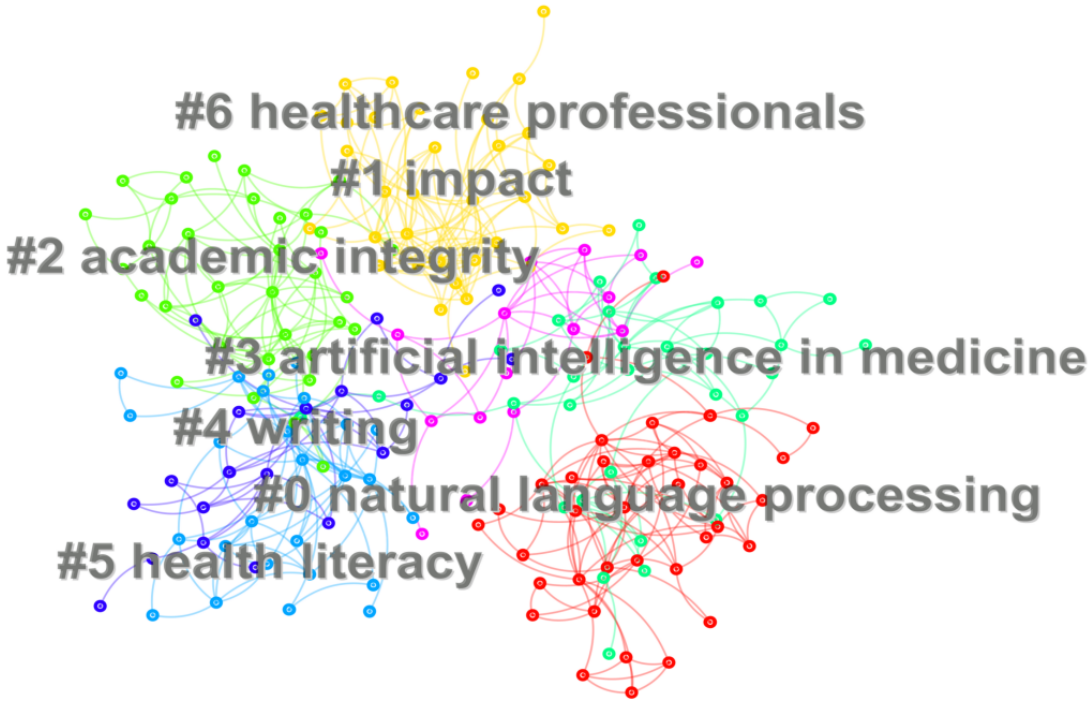

XSL•FO
RenderX

Figure 2. National publications of ChatGPT in medicine and education.

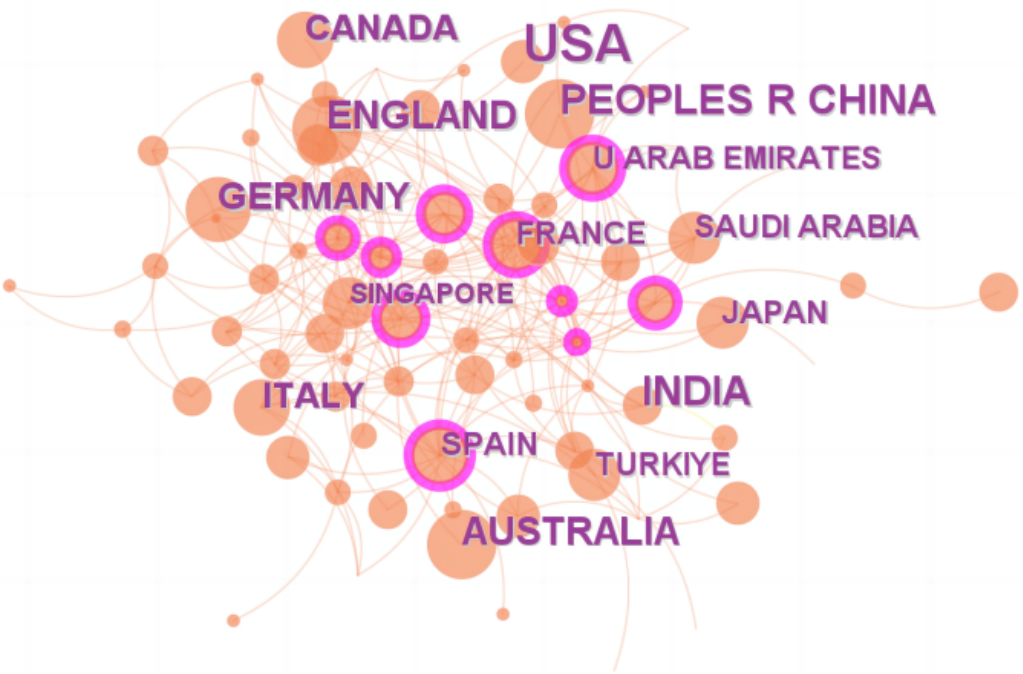

Next, we will analyze the impact of ChatGPT and focus on its advantages, disadvantages, and coping strategies in the areas of education, academics, clinical decision-making, and health education.

Potentials

A summary of the functions of ChatGPT and its potential role in the field of medical education is shown in Figure 3.

Figure 3. Functions of ChatGPT and the potentials of medical education.

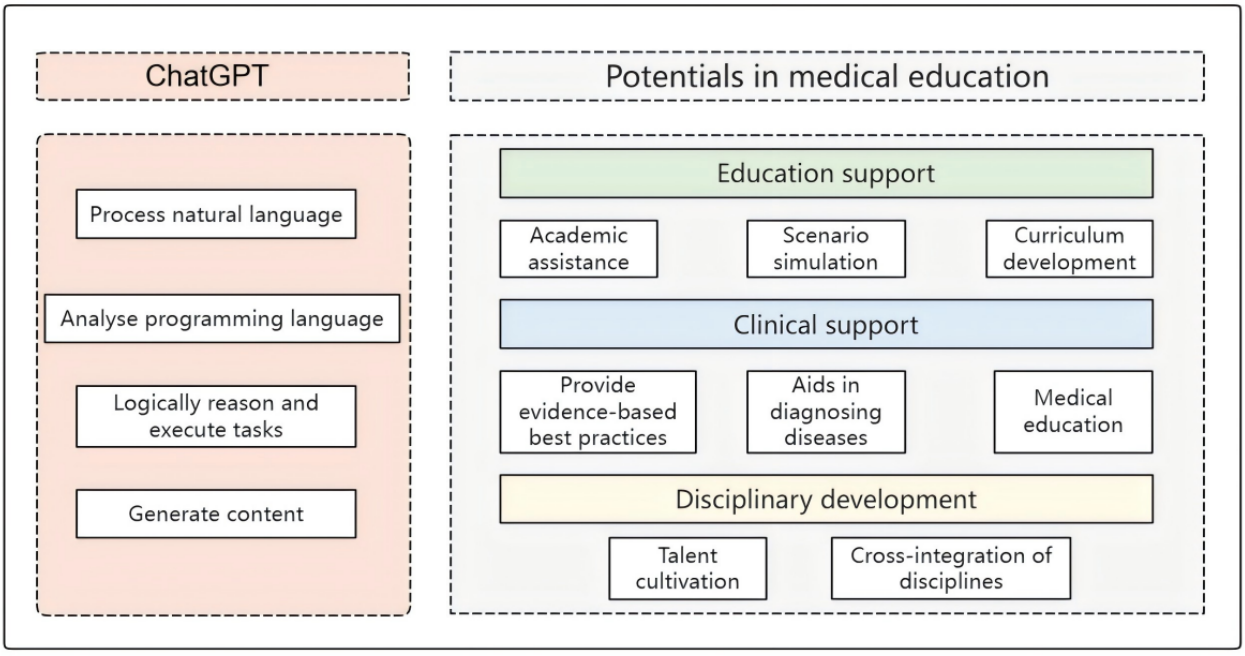

XSL•FO
RenderX

Education Support

Academic Assistance

ChatGPT excels in both language understanding and content generation. It relies on semantic understanding and reasoning to decipher user intent through mutual dialog. In addition, by applying deep learning techniques, ChatGPT efficiently retrieves information from a variety of sources to provide consumers with reliable answers [7]. As a result, ChatGPT emerges as an invaluable resource for medical students, especially in helping them understand complex ideas. ChatGPT improves topic understanding by providing examples and conducting text analyses [8]. Furthermore, ChatGPT serves an important role in alleviating academic difficulty among medical students [9]. With its ability to produce and answer questions as well as assist with revision tasks, ChatGPT can help medical students complete coursework, assess the quality of their coursework, and reinforce previously learned concepts [10]. Serving as a personalized tutor, it develops customized learning programs and time management strategies based on individual interests and learning preferences [11]. In addition, renowned for its prowess in research and writing, ChatGPT contributes significantly to academic endeavors [12]. On one hand, it assists students in comprehensively exploring research literature, gaining a preliminary understanding of current research trends [13]. On the other hand, ChatGPT aids in structuring thesis frameworks and generating writing prompts, while also offering grammar and spelling checks to enhance writing proficiency and quality [14].

Scenario Simulation

ChatGPT has superior social features that allow it to replicate clinical settings for medical students through situational simulation and role-playing [15]. This functionality assists the transition of medical students from a theoretical to a clinical attitude. Furthermore, ChatGPT provides the potential to faithfully mimic clinical circumstances while dynamically adapting to changes in patients' conditions [16]. This function lets students gain practical experience managing unexpected medical problems in simulated contexts, thereby improving their preparedness and psychological fortitude.

Curriculum Development

ChatGPT is crucial in educational curriculum development since it helps teachers enhance their logical thinking and task performance skills [17]. It assists teachers create lesson plans, course handouts, and lesson plan content, which accelerates curriculum development [13]. For example, we asked ChatGPT to create a teaching plan on pressure ulcer care, and the resulting content is presented in Figure 4. Upon assessing the information, it is obvious that ChatGPT could generate core teaching plan content. However, teachers must supplement with more precise and detailed knowledge points, as well as validate the created information.

Furthermore, ChatGPT promotes innovation in teaching approaches by enabling scenario-based learning, role-playing, and the integration of diverse educational resources [18]. Recognizing the value of continuous pedagogical innovation, teachers are encouraged to apply the capabilities of AI, such as ChatGPT, alongside traditional teaching methods [19]. This integration allows for constant advancement in teaching approaches, which improves the effectiveness of teaching practices and enables educators to fulfill their instructional responsibilities more efficiently.

XSL•FO
RenderX

Figure 4. Lesson plan for pressure ulcer care.

ChatGPT

Pressure Ulcer Care Teaching Plan

Teaching Objectives:

1. Understand the definition, classification, and pathogenesis of pressure ulcers.
2. Master the prevention measures and assessment methods of pressure ulcers.
3. Understand the principles of treatment and nursing methods for pressure ulcers.
4. Cultivate students' comprehensive abilities in pressure ulcer care, including teamwork and communication skills.

Teaching Challenges:

1. Understanding the pathogenesis and grading system of pressure ulcers.
2. Details and practical operations of pressure ulcer prevention measures.
3. Understanding treatment methods and nursing techniques for pressure ulcers of different grades.

Teaching Focus:

1. Prevention measures and assessment methods of pressure ulcers.
2. Principles of treatment and nursing methods for pressure ulcers.
3. Case analysis and practical exercises.

Teaching Process:

Part One: Theoretical Explanation

1. **Definition and Classification of Pressure Ulcers**:
   - Explain the definition of pressure ulcers, introduce common grading systems (such as Braden or NPUAP classifications), and describe the characteristics and pathophysiological changes of pressure ulcers at different levels.
   - Provide examples to illustrate the clinical manifestations and prognosis of pressure ulcers of different grades.
2. **Pathogenesis of Pressure Ulcers**:
   - Explain the main mechanisms of pressure ulcer formation, including poor blood circulation caused by sustained pressure, tissue hypoxia, and cellular damage.
   - Discuss the occurrence mechanisms of pressure ulcers in different situations through case analysis.
3. **Prevention Measures for Pressure Ulcers**:
   - Provide a detailed introduction to prevention measures for pressure ulcers, including regular repositioning, maintaining skin cleanliness and dryness, adequate nutrition and fluid intake, and the use of appropriate mattresses and supports.
   - Analyze actual cases to discuss how to develop effective prevention measures for different patient situations.

Part Two: Practical Operations

1. **Assessment Methods for Pressure Ulcers**:
   - Demonstrate how to assess pressure ulcers, including observing the location, size, depth, color, and surrounding tissue status.
   - Students work in groups to practice assessment and recording of simulated pressure ulcers.
2. **Principles of Pressure Ulcer Treatment**:
   - Demonstrate the treatment principles for pressure ulcers of different grades, including local cleansing and moist wound healing, pressure ulcer dressing, surgical treatment, etc.
   - Conduct group exercises to simulate the nursing process of pressure ulcers, including wound cleaning, dressing changes, etc.

Part Three: Case Discussion and Summary

1. **Case Analysis**:
   - Provide real cases for students to analyze and discuss prevention, assessment, and treatment strategies for pressure ulcers based on the knowledge learned.
   - Guide students to think and discuss the advantages, disadvantages, and possible implementation difficulties of different strategies.
2. **Conclusion and Reflection**:
   - Summarize the key points of this class, emphasizing the importance and comprehensiveness of pressure ulcer care.
   - Encourage students to reflect on their learning and provide questions and suggestions.

Teaching Evaluation:

1. Classroom Q&A: Assess students' mastery of pressure ulcer care knowledge through questioning.
2. Practical Operation Assessment: Evaluate students' practical operations, including the accuracy and standardization of pressure ulcer assessment and treatment operations.
3. Case Analysis Report: Require students to submit case analysis reports in groups, evaluating their comprehensive analysis and problem-solving abilities.

Clinical Support

ChatGPT plays a pivotal role in clinical support, standardizing procedures, aiding in disease diagnosis, and delivering health education. In terms of literature search and operational procedures, ChatGPT is capable of accessing the most recent literature and clinical guidelines [20]. This capability enables the provision of evidence-based best practices to health care professionals, facilitating the identification of current operational procedures and the enhancement of operational protocols [21]. Regarding disease diagnosis, ChatGPT can analyze patient data and test results, assisting doctors in diagnosing conditions and offering treatment recommendations [22]. In addition, it supports health care professionals in telemedicine by engaging in real-time communication with patients and providing remote diagnosis and treatment suggestions [23]. Moreover, ChatGPT serves as a valuable tool for patient health education. It can translate health education materials into multiple languages and deliver personalized health education using straightforward language, thereby aiding patients in understanding condition guidance and adopting healthier lifestyles [24].

Overall, ChatGPT enhances medical education by integrating essential functionalities into clinical support through standardized processes, aiding in disease diagnosis, and delivering health education [25]. This not only empowers health care professionals to provide medical services more efficiently but also provides students with a comprehensive and enriching learning experience, fostering the growth of medical professionals with practical skills and professional competence.

Disciplinary Development

The integration of AI and medicine represents a future frontier, characterized by the convergence of technological innovation and advancements in medical care [26]. This cross-disciplinary collaboration not only promotes technical advancement but also drives medical treatment to better intelligence and efficiency. Furthermore, it promotes the development of comprehensive capabilities in academic institutions that can navigate the junction of medicine and technology. In the future, coordinated activities in medical and technical domains have the potential to expedite technological innovation and move the field of medicine forward. This collaboration promises to usher in a new era of health care marked by innovation and efficiency, efficiently meeting society's changing health needs [27]. ChatGPT advances medical education by enabling interdisciplinary collaboration and encouraging innovation at the confluence of medicine and technology.

Challenges

Learning Dependency and Uneven Education

While ChatGPT can serve as a valuable learning aid by providing answers, assisting in understanding complex concepts, and offering personalized tutoring, excessive reliance on it can yield detrimental consequences in the long term [10]. Overdependence on ChatGPT may result in the erosion of critical thinking skills, creativity, and self-directed learning capabilities. The ease of obtaining answers quickly through ChatGPT may foster complacency among students, discouraging

XSL•FO
RenderX

them from engaging in reflective problem-solving [28]. To mitigate this issue, students should proactively disclose which parts of their work were aided by ChatGPT, allowing teachers to assess the overall quality of assignments more accurately [10]. Failure to do so may lead to the perpetuation of an "information cocoon," wherein students are only exposed to solutions that align with their existing preferences, hindering the exploration of diverse perspectives. Furthermore, the widespread adoption of ChatGPT may exacerbate educational inequalities. Developing and underdeveloped countries may lack the necessary technological infrastructure and resources to fully leverage ChatGPT, widening the educational gap between these regions and more developed countries [29]. Hence, it is crucial to address these challenges to ensure equitable access to educational resources and opportunities worldwide.

Copying and Plagiarism

When tackling assignments or final papers, students often seek to leverage technology to address challenges, enhance content, and elevate the overall quality of their work. ChatGPT, with its exceptional social capabilities and vast knowledge base, offers students the opportunity to obtain answers, prepare for exams, outline papers, and even complete them through mutual dialogue [29]. However, such practices are commonly perceived as copying answers and plagiarism, thereby contravening the fundamental principles of scientific research and academic integrity [30]. Moreover, ChatGPT encounters issues related to the fabrication of reference citations, as it may generate citations without verifiable sources, making it difficult for users to locate the original literature on academic platforms [31]. This inability to access and verify the sources provided by ChatGPT poses significant obstacles to maintaining academic integrity and conducting rigorous scholarly research [32].

It is noteworthy that Som Biswas, a radiologist in the United States, has authored 16 papers with ChatGPT, resulting in the publication of 6 articles across 4 journals [33-38]. Nevertheless, a thorough examination of the content by experts in the field has revealed significant inaccuracies, with all references found to be fictitious [39]. The editors of the journal Nature stated that while ChatGPT cannot be held accountable for the content and integrity of scientific studies, its contributions can be recognized [40]. Similarly, scientific journal editors claimed that using ChatGPT-generated content without adequate citations may be considered plagiarism, while contributions to ChatGPT might be acknowledged in the acknowledgments section [40]. Therefore, there is an urgent need to develop clearer criteria for distinguishing between authorized use and plagiarism when using ChatGPT assistance in academic research.

Insufficient Factualness, Timeliness, and Interpretability

ChatGPT's credibility is not absolute, as it grapples with issues, such as illusion, poor timeliness, and interpretability, akin to other large language models [41]. Despite its impressive performance, ChatGPT has been known to generate convincing yet erroneous information, undermining its reliability, particularly in the health care domain. Furthermore, because its training data is only up to January 2022, its conclusions may not always be up to date [12]. Furthermore, ChatGPT lacks specialist medical knowledge and may struggle to understand complicated illness relationships [42]. Its algorithms function as a black box, providing findings without divulging the underlying mechanism, creating ambiguity about their applicability for health care applications [43]. While doctors may leverage ChatGPT for disease diagnosis to enhance clinical accuracy, the absence of evidence supporting diagnosis and treatment may leave patients questioning the reliability of the results.

Ethical Issues

While ChatGPT has the potential to revolutionize medical education, it also raises a number of ethical concerns. It is of the utmost importance to address issues, such as algorithmic discrimination, the allocation of responsibility for medical malpractice involving ChatGPT, and the safeguarding of data privacy and security [44,45].

ChatGPT displays algorithmic biases, including gender stereotypes, racial discrimination, and cultural insensitivity [46]. These biases not only undermine the model's accuracy, fairness, and reliability, but they also perpetuate disparities in clinical health care [29]. Furthermore, ChatGPT's inadequacies, such as inadequate timeliness, interpretability, and accuracy, increase the risk of incorrect clinical diagnoses, treatment protocols, and the transmission of incorrect information, risking patient care [47]. Furthermore, safeguarding patient privacy is critical in the therapeutic arena, mandating strict safeguards for sensitive patient data. Given that ChatGPT may share information, there may be a possibility of privacy breaches, as patient data are temporarily stored on Open AI's servers [48]. This raises concerns regarding the potential leaking of patients' private information, such as personal details, medical conditions, and examination results, while using ChatGPT for assisted diagnosis, treatment, and health education [49]. Hence, there is an urgent need to establish practical ethical norms to harness the value of ChatGPT while ensuring alignment with scientific and technological advancements and societal development goals. These norms should address concerns related to algorithmic biases, data privacy, accuracy, and accountability to foster responsible and ethical use of ChatGPT in health care and other domains [50].

Undermining Communication and Trust

Currently, research on the applications of ChatGPT for clinical communication and health education is limited, and the outcomes are not substantial [51]. Health care professionals' humanistic care qualities, as well as the emotional rapport they build with patients, are critical components of disease treatment [49]. While ChatGPT can aid in accomplishing clinical tasks, it lacks emotion, empathy, and the capacity to perceive patient emotions. Therefore, health care professionals cannot depend on ChatGPT for communication and health education with patients [52]. This necessitates health care professionals integrating emotional value into patient interactions, fostering a more holistic approach to care, and combining the rationality of AI with the empathetic senses of health care providers. By striking a balance between technological assistance and human compassion, health care professionals can cultivate a

XSL•FO
RenderX

patient-centric environment that addresses both medical needs and emotional well-being.

Strategies

Table 1 summarizes the challenges and strategies of ChatGPT.

Table 1. The challenges and strategies of ChatGPT.

Challenges	Strategies	Feasible plans
Learning dependency and uneven education	• Delineate the relationship between individuals and ChatGPT • Apprise the limitations of ChatGPT • Optimize curriculum design	• ChatGPT serves as a useful tool to aid us • Adopt an objective and cautious stance • Integrate ChatGPT into medical curriculum • Optimize methods for assessing assignments and grades
Copying and plagiarism	• Guide proper use • Evaluate and review the content • Develop guidelines and regulatory mechanisms	• Address the basic principles and ethical boundaries and enhance the awareness of academic ethics and integrity • Develop specialized software or use fake text detection technology • Formulate guidelines and management mechanism
Insufficient factualness, timeliness, and interpretability	• Enhance and tailor training methodologies • Improve accuracy • Improve transparency and interpretability	• Large-scale training and use transfer learning • Fine-tuning and reinforcement learning • Establish visual interfaces, output human-machine interaction code, and using XAI[a] techniques
Ethical issues, algorithmic discrimination, responsibility allocation, and safeguarding data privacy and security	• Address algorithmic discrimination • Assign responsibility • Protecting personal privacy and information data	• Large-scale training, XAI, and quality improvement • Review content, involve health care professionals in decision-making, and ensure patients' right • Conceal original motives, implement manual review mechanisms, data encryption, anonymization, and establish guidelines, ethical frameworks, and institutional norms
Undermine communication and trust	• Humanistic care and emotional support from health care professionals • Teacher's quality education	• Provide emotional support and humanistic care • Implement student-centered education and cultivate excellent qualities

[a]XAI: explainable artificial intelligence.

Prevention of Overdependence

Delineate the Relationship Between Individuals and ChatGPT

Humans remain the primary agents in social activities, with ChatGPT serving as a useful tool to aid them [53]. While students may use the answers and suggestions provided by ChatGPT, it is crucial that they engage in critical thinking and judgment to arrive at their conclusions [54].

Apprise the Limitations of ChatGPT

Despite students' optimism toward ChatGPT [55], educators must underscore that it is not a panacea and guide students to adopt an objective and cautious stance [30]. Educators can focus on nurturing students' independent thinking, creative problem-solving abilities, and information literacy skills. This includes cultivating habits of reading and lifelong learning, fostering critical thinking and effective communication, and enhancing independent problem-solving skills [21]. The ultimate goal is to empower students to transition from mere questioners to creators and decision makers [56].

Optimize Curriculum Design

This can include introducing a teacher-student-machine interaction paradigm for instruction, giving courses on AI, and experimenting with new assignment forms and evaluation methodologies [57].

1. Integrating ChatGPT into the medical curriculum: problem-based learning allows students to analyze clinical problems, encouraging proactive thinking, and comparing their own ideas to those generated by ChatGPT [58]. This comparison encourages students to reflect on the strengths and limitations of both human-generated and AI-generated solutions, fostering a deeper understanding of clinical reasoning and decision-making processes. In addition to integrating ChatGPT with problem-based learning, it can also be combined with other instructional methods, such as case-based learning, team-based learning, group meetings, etc [59]. By using diverse teaching modalities, critical thinking and innovative thinking among students can be nurtured.
2. Update the methods for assessing assignments and grades. Educators might replace typical writing tasks with presentation reports, oral debates, group discussions, and peer reviews [17]. Encouraging students to preserve transcripts of their interactions using ChatGPT can also help [31]. Rather than relying on ChatGPT for direct responses, emphasizing knowledge and competency allows students to interact more deeply and think critically.

XSL•FO
RenderX

Academic Integrity

Guide Proper Use

ChatGPT has distinct advantages, and it is critical not to restrict students from using it totally, but rather to guide them to use right [60]. Thus, improving the quality of student learning is critical. In terms of educational orientation, teachers ought to lead students toward developing appropriate values about science and technology. Prioritizing education on academic integrity is crucial, with a focus on reiterating the basic principles and ethical boundaries of scientific research, enhancing awareness of academic ethics and integrity, and deepening reverence for science. In addition, organizations and institutions can conduct academic integrity seminars to educate individuals about the ethical use of ChatGPT [57]. Educators should also educate students about the consequences and repercussions of violating research integrity.

Evaluate and Review the Content

The text produced by ChatGPT should be evaluated and reviewed to ensure academic accuracy and integrity [61]. Developing specialized software or using fake text detection technologies created particularly to recognize text generated by ChatGPT can help detect whether a communication contains faked or nonsensical text [62]. To ensure that the answers are correct, they must be approached with reasonable skepticism and verified for accuracy. It is also vital to clearly identify which pieces came from ChatGPT. For example, using plagiarism detection software allows students to ensure that the information generated by ChatGPT does not infringe on other people's academic work, lowering the danger of plagiarism and ensuring the accuracy and authenticity of cited references [63].

Develop Guidelines and Regulatory Mechanisms

Stakeholders can collaborate to create relevant recommendations for the standardized use of ChatGPT [64]. Simultaneously, implementing a corresponding management system can enhance the management approach by providing training, education, assessment, review, feedback, and improvement activities to ensure the ethical use of ChatGPT [65].

Model Enhancement

Enhance and Tailor Training Methodologies

While ChatGPT boasts significant power, its susceptibility to hallucination poses a challenge to its credibility. However, effective mitigation of this issue and refinement of its specialization could unlock limitless potential in the medical domain [66]. Using regular input of data into the model or using transfer learning can effectively augment a vast, diverse, accurate, and high-quality training dataset, thereby enhancing the performance of ChatGPT [67].

Improve Accuracy

Through fine-tuning or reinforcement learning, the process of continuously incorporating user feedback, collecting and analyzing suggestions, and reintegrating ChatGPT resources is achieved to sustainably improve ChatGPT performance [62]. Collecting feedback on ChatGPT responses allows for iterative adjustments and enhancements to the model's accuracy [68]. Timely updates of data, especially for time-sensitive matters, facilitate a more accurate understanding of queries and the generation of relevant answers. For instance, regular updates on clinical data, research findings, expert consensus, and medical guidelines can effectively inform clinical practice.

Improve Transparency and Interpretability

ChatGPT's transparency can be improved by creating visual interfaces, generating human-machine interaction code, and using explainable AI techniques [8]. Increased openness builds confidence between humans and robots. Not only does it help medical workers understand the model's decision-making process, but it also allows for improved evaluation and interpretation of generated outcomes, lowering the risk of medical errors [69]. Incorporating transparency measures is thus critical for increasing ChatGPT's use in medical settings.

Emphasizing Ethical Issues

Value alignment is a contentious subject in ChatGPT, with the goal of aligning its capabilities and actions with human objectives, ethical standards, and values in order to promote safety and confidence in human-ChatGPT cooperation. The top 3 ethical concerns related to ChatGPT include algorithmic discrimination, medical liability allocation, and privacy and security difficulties [49].

Address Algorithmic Discrimination

To address algorithmic bias, on the one hand, incorporating diverse and balanced samples for large-scale training can enhance ChatGPT's fairness awareness [70]. On the other hand, using explainable AI can help identify biased patterns in ChatGPT while implementing fair models [8]. Furthermore, continuous review and quality improvement can minimize gender and racial discrimination in the health care sector to the greatest extent possible [44].

Assign Responsibility

Clear responsibility allocation is paramount when collaborating with ChatGPT for clinical disease diagnosis and health education. In other words, ensuring maximum safety in the clinical use of ChatGPT requires explicit delineation of responsibilities. Health care professionals using ChatGPT for decision support should actively engage in its decision-making process and critically assess its recommendations. Collaboration between health care professionals and ChatGPT can enhance decision-making accuracy [71]. Patients' right to information should be upheld. Organizations and institutions must ensure ChatGPT's responsible participation in treatment and adherence to ethical standards [72].

Protecting Personal Privacy and Information Data

Measures such as analogously asking ChatGPT questions to conceal original motives, implementing manual review mechanisms for uploaded information, data encryption, and anonymization can mitigate privacy risks [73]. Compliance with privacy laws and regulations is essential to secure patient privacy and medical information, ensuring data handling and storage integrity [43]. Establishing clear guidelines, ethical frameworks, and institutional norms for data collection, storage, and use is crucial [66].

XSL•FO
RenderX

Irreplaceability of Medical Personnel and Educators

Humanistic Care and Emotional Support From Health Care Professionals

While ChatGPT has the potential to enhance access to primary health care in underdeveloped regions and streamline repetitive tasks for medical staff, it cannot be considered a substitute for health care professionals in any capacity [13]. ChatGPT is devoid of the attributes that are typically associated with independent consciousness, ethical standards, emotional empathy, and the capacity to anticipate unforeseen circumstances. The interpersonal communication between health care professionals and patients, as well as the emotional support and humanistic care that are provided in person, cannot be replicated by ChatGPT [74]. Furthermore, medical professionals provide invaluable empirical assistance and support to patients based on their clinical experience, which enhances the quality of clinical services. This is a capability that ChatGPT does not possess [74]. Consequently, ChatGPT should be regarded as a valuable adjunct to the work of medical personnel in clinical settings, rather than as a replacement for human health care providers.

Teacher's Quality Education

Education fosters individual growth, community advancement, and the continuation of human civilization. While ChatGPT has effectively helped to progress education by equalizing and enriching instructional resources, it is still only a tool for teachers, and not a replacement [75]. Teachers have distinct powers that ChatGPT lacks. They practice student-centered teaching and help kids develop moral qualities and abilities like ideals, beliefs, values, critical thinking, emotional intelligence, and creativity. Furthermore, people are naturally social animals that require interpersonal interactions to thrive and find spiritual fulfillment.

In conclusion, despite the benefits of ChatGPT, technology cannot replace humans in health care and education. Our individual features and capacities allow us to maintain social value in an era of rapid developments in AI.

Conclusion

ChatGPT has demonstrated considerable potential in medical education, but it has also introduced a number of thought-provoking difficulties. As ChatGPT advances, it may present new obstacles and opportunities. On a worldwide basis, cultural diversity is under peril. ChatGPT disseminates information based on the training data it gets, making well-known and popular information more easily shared and transferred. However, this has the tendency to marginalize niche or local cultures and languages, hence reducing cultural variety. Roles in the health care sector may change. The increased implementation of automation and intelligent technologies may result in the displacement of positions, such as primary diagnostic personnel and imaging diagnostics. But this might also open new career paths for professionals with expertise in medical AI and information management. Promoting AI technology development in developing nations can help advance fields, such as the creation of diagnostic software.

It is imperative that we approach ChatGPT with caution and subject it to critical evaluation, weighing its benefits against its drawbacks. This paper presents a dialectical examination of the current state of ChatGPT application in medical education, conducting an in-depth analysis of its advantages and the dilemmas it presents. In addition, targeted strategies are proposed to address these challenges effectively. The aim is to standardize and rationalize ChatGPT's maximum potential in the future, paving the way for innovative approaches in medical education and contributing to the advancement of medicine.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021-HLKY-06) and the Clinical Nursing Research Fund Project of the Second Xiangya Hospital, Central South University (2021-HLKY-06), People's Republic of China.

Authors' Contributions

TX performed the conceptualization of topics, analysis and interpretation of the image, and writing of the paper. QW and HW supervised and controlled writing quality. FL and LY contributed to project administration. YL and ZD monitored and regulated the project.

Conflicts of Interest

None declared.

References

1. Holmes JH, Sacchi L, Bellazzi R, Peek N. Artificial intelligence in medicine AIME 2015. Artif Intell Med 2017 Sep;81:1-2. [doi: 10.1016/j.artmed.2017.06.011] [Medline: 28733119]
2. Wiggins WF, Tejani AS. On the opportunities and risks of foundation models for natural language processing in radiology. Radiol Artif Intell 2022 Jul;4(4):e220119 [FREE Full text] [doi: 10.1148/ryai.220119] [Medline: 35923379]
3. Kurian N, Cherian JM, Sudharson NA, Varghese KG, Wadhwa S. AI is now everywhere. Br Dent J 2023 Jan;234(2):72. [doi: 10.1038/s41415-023-5461-1] [Medline: 36707552]

XSL•FO
RenderX

4. Ruiz-Rojas LI, Acosta-Vargas P, De-Moreta-Llovet J, Gonzalez-Rodriguez M. Empowering education with generative artificial intelligence tools: approach with an instructional design matrix. Sustainability 2023 Jul;15(15):11524. [doi: 10.3390/su151511524]
5. Lin X. Exploring the role of ChatGPT as a facilitator for motivating self-directed learning among adult learners. Adult Learning 2023;35(3):156-166. [doi: 10.1177/10451595231184928]
6. Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci 2006;57(3):359-377. [doi: 10.1002/asi.20317]
7. Vitorino LM, Júnior GHY. ChatGPT and the teaching of contemporary nursing: and now professor? J Clin Nurs 2023 Nov;32(21-22):7921-7922. [doi: 10.1111/jocn.16706] [Medline: 37004198]
8. Khennouche F, Elmir Y, Himeur Y, Djebari N, Amira A. Revolutionizing generative pre-traineds: insights and challenges in deploying ChatGPT and generative chatbots for FAQs. Expert Systems with Applications 2024;246:123224. [doi: 10.1016/j.eswa.2024.123224]
9. Chan MMK, Wong ISF, Yau SY, Lam VSF. Critical reflection on using ChatGPT in student learning: benefits or potential risks? Nurse Educ 2023;48(6):E200-E201. [doi: 10.1097/NNE.0000000000001476] [Medline: 37348135]
10. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. Int J Educ Technol High Educ 2023 Oct;20(1):56. [doi: 10.1186/s41239-023-00426-1]
11. Thakur A, Parikh D, Thakur A. ChatGPT in nursing education: is there a role for curriculum development? Teaching and Learning in Nursing 2023 Apr;18(3):450-451. [doi: 10.1016/j.teln.2023.03.011]
12. Gandhi Periaysamy A, Satapathy P, Neyazi A, Padhi BK. ChatGPT: roles and boundaries of the new artificial intelligence tool in medical education and health research - correspondence. Ann Med Surg (Lond) 2023 Mar;85(4):1317-1318 [FREE Full text] [doi: 10.1097/MS9.0000000000000371] [Medline: 37113819]
13. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
14. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? J Med Syst 2023;47(1):61. [doi: 10.1007/s10916-023-01957-w] [Medline: 37160568]
15. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
16. Hobert S. Say hello to coding tutor design and evaluation of a Chatbot-based learning system supporting students to learn to program. 2019 Presented at: Fortieth International Conference on Information Systems; 2019 December 15-18; Munich, Germany p. 1-17.
17. Sun GH, Hoelscher SH. The ChatGPT storm and what faculty can do. Nurse Educ 2023;48(3):119-124. [doi: 10.1097/NNE.0000000000001390] [Medline: 37043716]
18. Memarian B, Doleck T. ChatGPT in education: methods, potentials, and limitations. Computers in Human Behavior: Artificial Humans 2023;1(2):100022. [doi: 10.1016/j.chbah.2023.100022]
19. Irwin P, Jones D, Fealy S. What is ChatGPT and what do we do with it? Implications of the age of AI for nursing and midwifery practice and education: an editorial. Nurse Education Today 2023 Aug;127:105835. [doi: 10.1016/j.nedt.2023.105835] [Medline: 37267643]
20. Odom-Forren J. The role of ChatGPT in perianesthesia nursing. J Perianesth Nurs 2023 Apr;38(2):176-177. [doi: 10.1016/j.jopan.2023.02.006] [Medline: 36965923]
21. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med 2023 Mar;13(3):e1206 [FREE Full text] [doi: 10.1002/ctm2.1206] [Medline: 36854881]
22. Ferdush J, Begum M, Hossain ST. ChatGPT and clinical decision support: scope, application, and limitations. Ann Biomed Eng 2024 Jul;52(5):1119-1124. [doi: 10.1007/s10439-023-03329-4] [Medline: 37516680]
23. Sharma M, Sharma S. A holistic approach to remote patient monitoring, fueled by ChatGPT and metaverse technology: the future of nursing education. Nurse Educ Today 2023 Dec;131:105972. [doi: 10.1016/j.nedt.2023.105972] [Medline: 37757713]
24. Moons P, Van Bulck L. ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. Eur J Cardiovasc Nurs 2023 Oct;22(7):e55-e59. [doi: 10.1093/eurjcn/zvad022] [Medline: 36752788]
25. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
26. Wang Y, Li N, Chen L, Wu M, Meng S, Dai Z, et al. Guidelines, consensus statements, and standards for the use of artificial intelligence in medicine: systematic review. J Med Internet Res 2023 Nov;25:e46089 [FREE Full text] [doi: 10.2196/46089] [Medline: 37991819]
27. Ahmed Y, Ireland S. Utilization of ChatGPT in medical education: applications and implications for curriculum enhancement. Acta Inform Med 2023;31(4):300-305 [FREE Full text] [doi: 10.5455/aim.2023.31.300-305] [Medline: 38379690]
28. Abdulai A, Hung L. Will ChatGPT undermine ethical values in nursing education, research, and practice? Nurs Inq 2023 Jul;30(3):e12556. [doi: 10.1111/nin.12556] [Medline: 37101311]

XSL•FO
RenderX

29. Liu J, Liu F, Fang J, Liu S. The application of chat generative pre-trained transformer in nursing education. Nurs Outlook 2023;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
30. Choi EPH, Lee JJ, Ho M, Kwok JYY, Lok KYW. Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Educ Today 2023 Jun;125:105796. [doi: 10.1016/j.nedt.2023.105796] [Medline: 36934624]
31. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023 Jul;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
32. Dashti M, Londono J, Ghasemi S, Moghaddasi N. How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? The Journal of Prosthetic Dentistry 2023 Jul;S0022-3913(23):00371-00372. [doi: 10.1016/j.prosdent.2023.05.023]
33. Biswas S, Dobaria D, Cohen HL. ChatGPT and the future of journal reviews: a feasibility study. Yale J Biol Med 2023 Sep;96(3):415-420 [FREE Full text] [doi: 10.59249/SKDH9286] [Medline: 37780993]
34. Biswas S. ChatGPT and the future of medical writing. Radiology 2023 Apr;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
35. Biswas S. Passing is great: can ChatGPT conduct USMLE exams? Ann Biomed Eng 2023 Sep;51(9):1885-1886. [doi: 10.1007/s10439-023-03224-y] [Medline: 37154989]
36. Biswas SS. Potential use of chat GPT in global warming. Ann Biomed Eng 2023 Jun;51(6):1126-1127. [doi: 10.1007/s10439-023-03171-8] [Medline: 36856927]
37. Perera Molligoda Arachchige AS, Biswas S. Re: potential applications and impact of ChatGPT in radiology. Acad Radiol 2023 Nov;S1076-S6332(23):00601-00603. [doi: 10.1016/j.acra.2023.10.048] [Medline: 37973516]
38. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng 2023 May;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
39. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles. Skeletal Radiol 2023 Nov;52(9):1755-1758. [doi: 10.1007/s00256-023-04340-5]
40. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023 Jan;613(7945):620-621. [doi: 10.1038/d41586-023-00107-z] [Medline: 36653617]
41. Wu T, He S, Liu J, Sun S, Liu K, Han Q, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J. Autom. Sinica 2023 May;10(5):1122-1136. [doi: 10.1109/jas.2023.123618]
42. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
43. Li J. Security implications of AI Chatbots in health care. J Med Internet Res 2023;25:e47551 [FREE Full text] [doi: 10.2196/47551] [Medline: 38015597]
44. Ray PP, Majumder P. The potential of ChatGPT to transform healthcare and address ethical challenges in artificial intelligence-driven medicine. J Clin Neurol 2023 Nov;19(5):509-511 [FREE Full text] [doi: 10.3988/jcn.2023.0158] [Medline: 37635433]
45. Ali MJ, Djalilian A. Readership awareness series - paper 4: Chatbots and ChatGPT - ethical considerations in scientific publications. Ocul Surf 2023 Apr;28:153-154. [doi: 10.1016/j.jtos.2023.04.001] [Medline: 37028488]
46. Arshad HB, Butt SA, Khan SU, Javed Z, Nasir K. ChatGPT and artificial intelligence in hospital level research: potential, precautions, and prospects. Methodist Debakey Cardiovasc J 2023 Nov;19(5):77-84 [FREE Full text] [doi: 10.14797/mdcvj.1290] [Medline: 38028967]
47. Sridi C, Brigui S. The use of ChatGPT in occupational medicine: opportunities and threats. Ann Occup Environ Med 2023 Oct;35:e42 [FREE Full text] [doi: 10.35371/aoem.2023.35.e42] [Medline: 38029273]
48. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024 Feb;52(2):130-133. [doi: 10.1007/s10439-023-03296-w] [Medline: 37378876]
49. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
50. Liu J, Chen C, Qu Y, Yang S, Xu L. RASS: enabling privacy-preserving and authentication in online AI-driven healthcare applications. ISA Trans 2023;141:20-29. [doi: 10.1016/j.isatra.2023.03.049] [Medline: 37059673]
51. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using Chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul;22(7):e16021 [FREE Full text] [doi: 10.2196/16021] [Medline: 32673216]
52. Scerri A, Morin KH. Using chatbots like ChatGPT to support nursing practice. J Clin Nurs 2023 Aug;32(15-16):4211-4213. [doi: 10.1111/jocn.16677] [Medline: 36880216]
53. Busch F, Adams LC, Bressem KK. Biomedical ethical aspects towards the implementation of artificial intelligence in medical education. Med Sci Educ 2023 Jun;33(4):1007-1012 [FREE Full text] [doi: 10.1007/s40670-023-01815-x] [Medline: 37546190]

XSL•FO
RenderX

54. Shoufan A. Exploring students' perceptions of ChatGPT: thematic analysis and follow-up survey. IEEE Access 2023;11:38805-38818. [doi: 10.1109/access.2023.3268224]
55. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659 [FREE Full text] [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
56. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. Innovations in Education and Teaching International 2023;61(3):460-474. [doi: 10.1080/14703297.2023.2195846]
57. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
58. Divito CB, Katchikian BM, Gruenwald JE, Burgoon JM. The tools of the future are the challenges of today: the use of ChatGPT in problem-based learning medical education. Med Teach 2024;46(3):320-322. [doi: 10.1080/0142159X.2023.2290997] [Medline: 38149617]
59. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024;10:e52483 [FREE Full text] [doi: 10.2196/52483] [Medline: 38598263]
60. Harrison LM, Hurd E, Brinegar KM. Critical race theory, books, and ChatGPT: moving from a ban culture in education to a culture of restoration. Middle School Journal 2023;54(3):2-4. [doi: 10.1080/00940771.2023.2189862]
61. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol 2023 Sep;13(4):170-178. [doi: 10.5662/wjm.v13.i4.170]
62. Sohail SS, Farhat F, Himeur Y, Nadeem M, Madsen D, Singh Y, et al. Decoding ChatGPT: a taxonomy of existing research, current challenges, and possible future directions. Journal of King Saud University - Computer and Information Sciences 2023;35(8):101675. [doi: 10.1016/j.jksuci.2023.101675]
63. Vignesh R, Pradeep P, Balakrishnan P. A tête-à-tête with ChatGPT on the impact of artificial intelligence in medical education. Med J Malaysia 2023 Jul;78(4):547-549 [FREE Full text] [Medline: 37518931]
64. Shoja MM, Van de Ridder JMM, Rajput V. The emerging role of generative artificial intelligence in medical education, research, and practice. Cureus 2023 Jun;15(6):e40883 [FREE Full text] [doi: 10.7759/cureus.40883] [Medline: 37492829]
65. Kleebayoon A, Wiwanitkit V. ChatGPT and the teaching of contemporary nursing: Comment. J Clin Nurs 2023 Oct;32(19-20):7642. [doi: 10.1111/jocn.16762] [Medline: 37194403]
66. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng 2023 Dec;51(12):2647-2651. [doi: 10.1007/s10439-023-03284-0] [Medline: 37328703]
67. Al-Hasan TM, Sayed AN, Bensaali F, Himeur Y, Varlamis I, Dimitrakopoulos G. From traditional recommender systems to GPT-based chatbots: a survey of recent developments and future directions. BDCC 2024 Mar;8(4):36. [doi: 10.3390/bdcc8040036]
68. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences 2023;13(4):410. [doi: 10.3390/educsci13040410]
69. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
70. Stahl BC, Eke D. The ethics of ChatGPT – exploring the ethical issues of an emerging technology. International Journal of Information Management 2024 Feb;74:102700. [doi: 10.1016/j.ijinfomgt.2023.102700]
71. Woodnutt S, Allen C, Snowden J, Flynn M, Hall S, Libberton P, et al. Could artificial intelligence write mental health nursing care plans? J Psychiatr Ment Health Nurs 2024 Feb;31(1):79-86. [doi: 10.1111/jpm.12965] [Medline: 37538021]
72. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
73. Li J. A privacy preservation model for health-related social networking sites. J Med Internet Res 2015;17(7):e168 [FREE Full text] [doi: 10.2196/jmir.3973] [Medline: 26155953]
74. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: what does GPT-4 add to healthcare? The dawn of a new era. Cardiol J 2023 Oct;30(6):1018-1025 [FREE Full text] [doi: 10.5603/cj.97515] [Medline: 37830256]
75. Reiss M. The use of AI in education: practicalities and ethical considerations. London Review of Education 2021 Feb;19(1):15. [doi: 10.14324/lre.19.1.05]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 29.02.24; peer-reviewed by H Zhai, Y Himeur; comments to author 11.04.24; revised version received 05.06.24; accepted 29.06.24; published 28.08.24.

Please cite as:
Xu T, Weng H, Liu F, Yang L, Luo Y, Ding Z, Wang Q
Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies
J Med Internet Res 2024;26:e57896
URL: https://www.jmir.org/2024/1/e57896
doi:10.2196/57896
PMID:39196640

©Tianhui Xu, Huiting Weng, Fang Liu, Li Yang, Yuanyuan Luo, Ziwei Ding, Qin Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Digital Platform and Its Emerging Role in Decentralized Clinical Trials

Rachel R Copland[1], BSc; Sten Hanke[2], MSc, PhD; Amy Rogers[1], MBChB, MD; Lampros Mpaltadoros[3], MSc; Ioulietta Lazarou[3], PhD; Alexandra Zeltsi[3], MSc; Spiros Nikolopoulos[3], PhD; Thomas M MacDonald[1], BSc, MBChB, MD; Isla S Mackenzie[1], MBChB, PhD

[1]MEMO Research, School of Medicine, University of Dundee, Dundee, United Kingdom

[2]eHealth, FH Joanneum, Graz, Austria

[3]Information Technologies Institute, Centre for Research & Technology Hellas, Thessaloniki, Greece

Corresponding Author:
Rachel R Copland, BSc
MEMO Research
School of Medicine
University of Dundee
Ninewells Hospital and Medical School
Dundee, DD1 9SY
United Kingdom
Phone: 44 1382388380
Email: rcopland001@dundee.ac.uk

Abstract

Decentralized clinical trials (DCTs) are becoming increasingly popular. Digital clinical trial platforms are software environments where users complete designated clinical trial tasks, providing investigators and trial participants with efficient tools to support trial activities and streamline trial processes. In particular, digital platforms with a modular architecture lend themselves to DCTs, where individual trial activities can correspond to specific platform modules. While design features can allow users to customize their platform experience, the real strengths of digital platforms for DCTs are enabling centralized data capture and remote monitoring of trial participants and in using digital technologies to streamline workflows and improve trial management. When selecting a platform for use in a DCT, sponsors and investigators must consider the specific trial requirements. All digital platforms are limited in their functionality and technical capabilities. Integrating additional functional modules into a central platform may solve these challenges, but few commercial platforms are open to integrating third-party components. The lack of common data standardization protocols for clinical trials will likely limit the development of one-size-fits-all digital platforms for DCTs. This viewpoint summarizes the current role of digital platforms in supporting decentralized trial activities, including a discussion of the potential benefits and challenges of digital platforms for investigators and participants. We will highlight the role of the digital platform in the development of DCTs and emphasize where existing technology is functionally limiting. Finally, we will discuss the concept of the ideal fully integrated and unified DCT and the obstacles developers must address before it can be realized.

(J Med Internet Res 2024;26:e47882) doi:10.2196/47882

KEYWORDS

decentralized clinical trials; digital platform; digitalization; clinical trials; mobile phone

Introduction

In the last 2 decades, advances in digital health have given rise to clinical trial designs that are expected to disrupt the conventional clinical trial model [1-10]. Conventional clinical trials often depend on frequent in-person site visits, and are therefore limited to individuals who can easily access trial sites. The decentralized clinical trial (DCT), however, allows individuals to take part in research regardless of their geographical proximity to the trial site [11,12]. Moving trial activities away from sites and into participant homes or other local settings can also make trials less burdensome for participants [13].

DCTs are often, but not exclusively, facilitated by technology. Participants can be recruited through web-based advertising campaigns, receive study medications directly to their homes, self-report outcomes through smartphone apps, attend trial visits at home through videoconferencing software, and be monitored

XSL•FO
RenderX

remotely with interconnected wearable devices [14-23]. Sponsors increasingly use these technologies as the pharmaceutical industry shifts toward using more decentralized or hybrid [24] methods [1,2]. To fully harness the benefits of this shift toward decentralization, we recommend the adoption of unified, integrated, and DCT-specific digital platforms.

A widely agreed definition of digital platforms does not yet exist; in this paper, we define them as software environments that facilitate the completion of clinical trial tasks. Digital platforms, hosted locally or on remote (cloud-based) servers, can also facilitate centralized data monitoring and oversight of trial activity, and typically adopt a modular architecture. A unified platform is one where disparate systems and modules centralize (or unify) to create a seamless user interface by linking systems to allow uninterrupted data flow. DCTs often incorporate several modes of data collection (eg, devices, electronic participant reported outcomes) from disparate sources. Unifying these data in a single platform avoids data sprawl while providing the opportunity for richer data insights and streamlined workflows [25,26]. The Eureka digital platform, for example, was developed by the University of California and has been used in several DCTs. The Eureka platform allows participants to consent, upload data, and attend remote trial visits. Meanwhile, trial staff can access real-time study progress data, generate reports, and collect analysis-ready data sets [27]. Eureka permits the customization of user interfaces with a granular role-based permissions system. Platforms like Eureka can be tailored to multiple users, including trial participants, researchers, sponsors, and third-party vendors, and are accessible on any suitable internet-connected device.

Although the use of digital platforms in DCTs is relatively common [27-34], existing platforms are not without limitations. Lack of interoperability and unclear data standardization protocols are known challenges to building unified systems [35-37]. Many proprietary digital platforms offer modules to support common DCT trial activities such as electronic data capture (EDC), drug dispensing, or remote visits. However, not all planned DCT activities may be available in a single proprietary system. To overcome this limitation, 1 option would be to integrate a third-party software solution that provides the necessary functionality often not permitted on proprietary systems. Even where third-party integration is allowed, subsequent software updates may be problematic [38]. Lack of data standardization leads to platforms and modules lacking interoperability, and custom implementations increase build complexity, extending timelines, and adding to costs.

With hybrid and decentralized study designs becoming more popular, the demand for innovative systems to support these trials is likely to grow; this will result in technology partners becoming important stakeholders in the success of DCTs, particularly for those studies requiring custom builds. The trend toward decentralization, however, has highlighted discordance between trialists and technology partners. Where technology partners specialize in full-stack development and user experience design, they are not experts on clinical trials. Likewise, trialists are adept at the intricacies of the clinical trial and DCT landscape, but they may harbor unrealistic expectations of system capabilities. This lack of shared understanding between stakeholders slows progress [39-41], and literature discussing platforms from both technical and clinical perspectives is scarce. We have attempted to bridge this gap in this viewpoint. We examine how digital platforms and DCTs have developed and highlight the advantages of a digital platform approach. We also discuss where user engagement in platform design can strengthen the DCT model. Finally, we demonstrate the key challenges of deploying a unified, modular, and fully-integrated platform in terms of technical limitations and data standardization.

Digital Platforms: Then and Now

Early digital solutions for clinical trials took many forms, including EDC programs and software that helped streamline trial management. The INVEST trial compared treatment strategies for the treatment of hypertension in individuals with coronary artery disease [42,43]. The trial began in 1997 and was conducted using an internet-based software tool akin to a modern randomization and trial supply management system. Hosted on a Netscape Enterprise Server, the tool incorporated a database for real-time collection, verification, and validation of participant data at study visits. In the same year, Schoichet et al [44] described a communications network set up initially for the COMET trial, which investigated the efficacy of a chronic obstructive pulmonary disease treatment. The network included tools for data transfer, randomization and trial supply management activities, communications protocols, and an information access system for monitoring the trial status. The infrastructure was developed from preexisting local area networks, and frame relays were established between the participating organizations. An internal interactive voice response system was deployed to conduct and monitor tasks relating to drug supply management, collecting data in real-time. To monitor study activities, a platform was built on Lotus Notes (Lotus Development Corporation) with separate modules for enrollment of participants, trial monitoring, adverse event reporting, and tracking of case report forms.

Many modern platform interfaces, derived from early EDC systems, include participant-facing electronic diaries and web-based questionnaires. These have been used successfully in several early DCTs [45-47], and remain popular [32,48,49]. EDC platforms later expanded their capabilities beyond remote data capture to include other DCT elements, such as electronic consenting and telemedicine. Various DCTs have used the REDCap (Research Electronic Data Capture; Vanderbilt University) [50,51] electronic data capture platform to allow participants to upload trial data [52-54]. Other DCTs have used ResearchKit (Apple Inc.) [55], an open-source framework, to create participant platforms for users to consent, upload data, and view study results through a study-specific app [16,56,57].

There are several technology vendors offering ready-to-deploy, customizable, and often modular, platforms for DCTs. Companies such as IQVIA [58]; Medable [59]; Castor [60]; THREAD [61]; ObvioHealth [62]; Science37 [63]; Cognizant [64]; and AiCure [65], among others, offer digital platforms. For example, the RADIAL trial conducted by Trials@Home [66] is supported by eClinicalHealth's Clinpal [67] digital

XSL•FO
RenderX

platform. While conducting a recent systematic review of methods used to conduct DCTs, we found that, of 45 included trials, 31 managed their trials using a digital platform of some kind [24].

Why Use Digital Platforms for DCTs?

The benefits of using digital platforms for DCTs extend to both participants and researchers. Digital platforms allow participants to be monitored remotely, which is associated with improved self-management of conditions and better clinical outcomes for participants [68-70]. Meanwhile, remote monitoring gives researchers greater insight into the day-to-day variability of disease activity and its wider impact [71,72]. Platforms with electronic participant reported outcome modules can increase participant satisfaction, improve data quality, reduce costs, and improve response rates [73-75]. There are also benefits associated with participants using their own devices to access study platforms. Directly uploading one's information to the platform (ie, active data collection) offers participants greater flexibility and enables "on the go" or real-time reporting. On the other hand, data collected passively through devices, such as geolocation or screen interactions, provide researchers with greater insight into real-world behavior. The Brighten (Bridging Research Innovations for Greater Health in Technology, Emotion, and Neuroscience) DCTs, for example, actively and passively collected data to better understand the behaviors of participants with depression [76]. This combination of active and passive data collection could allow researchers to better monitor changes in behavior and, in theory, intervene sooner. DCTs with little to no participant-study team interaction may be experienced as disengaging by participants, but platform technologies can address these challenges. Push notifications, alerts, and reminders received through devices, phones, or personal computers have been cited as facilitators of engagement [77], and research has shown how these features have encouraged participants to complete study tasks and promote the use of study apps [78,79]. Research increasingly shows participants value having access to their own data [41,80], and displaying meaningful insights through participant dashboards could improve participant satisfaction. Another strategy for increasing engagement and promoting product longevity that is frequently adopted in commercial settings is gamification. Gamification incorporates game design elements in nongame contexts [81] to encourage reward-seeking behavior through incentives, such as unlocking new levels, receiving badges, or collecting points [82]. This strategy of promoting engagement has proved effective in several trials [83-88] and could be used to enhance both trial and platform engagement.

Digital platforms can be integrated with videoconferencing software, which allows participants to attend trial visits remotely [49,89,90]. Participant satisfaction with video visits is generally high, even when visits are completed on smartphones [91]. For instance, a platform could offer a web-based booking system, giving participants flexibility when scheduling follow-up visits. Further, integrating calendar application programming interfaces with the platform would allow visits to be added to the participant's existing email or device calendars, which could be accompanied by pop-ups or reminders, potentially reducing missed or forgotten appointments. A combination of the tools mentioned above could generate a platform that promotes participant engagement with the trial while also allowing investigators to supervise trial progress and participant safety.

Conducting DCTs through digital platforms can also result in improved data integrity; this can be accomplished through platforms with data visualization dashboards, data tables, activity and audit logs, and real-time error flags, with unauthorized data edits restricted through access controls. Currently, most data visualizations and dashboards are implemented with web access in JavaScript or similar tools. More modern visualizations incorporate business intelligence (BI) approaches where different data sources can be combined and structured for quick visualization. BI tools can further help to find patterns and correlations between data. Data traceability can be achieved with rules-based sanity tests, semiautomated validation, and automated report generation [92]. Research increasingly supports the adoption of blockchain technology in clinical trials, with advocates suggesting that an immutable ledger using smart contracts would improve data integrity, accountability, and traceability [93-96]. For example, Liang et al [97], demonstrated a system whereby users controlled data access and protected privacy through blockchain technology and Intel software guard extensions. Data integrity and accountability were ensured by hashing protected personal health data and data access records and anchoring them to a long-lasting but secure ledger with platform dependencies. Building digital platforms on blockchain technology and other similar technological solutions warrants further exploration but could have significant implications for improving data integrity and security in DCTs.

Anticipating Risks When Designing Platforms

Overview

Maintaining digital security and compliance with local legislation and data standards are paramount to the success of DCTs, and digital platforms are not exempt from potential risks. While an unstable internet connection presents a source of weakness for data sharing, personal devices can also fall victim to malware or expose personal information if lost or stolen. As more trials embrace platform technologies, data security and governance risks will continue to grow [98]. Security and governance teams must, therefore, be established both at the Sponsor site and among all trial partners to scan for malware, apply patches to applications, and keep study personnel and participants informed on how to protect their data. As with conventional clinical trial approaches, DCTs are vulnerable to nonadherence and attrition. Learnings from mHealth studies show that users lose momentum with app use after a goal has been achieved and nothing new can be gained [99], which could have consequences for trials using digital platforms. Following best practice guidelines for DCTs, like choosing a research question of high perceived value [13] or offering flexibility in the DCT approach [100], may ameliorate some of these challenges, but research in this area is limited.

XSL•FO
RenderX

Finally, population demographics and individual preferences can impact how a user interacts with digital platforms, and anticipating these factors in advance can minimize later delays. For example, lack of digital literacy, inexperience with trial technology, insufficient digital infrastructure, and lack of confidence using devices are frequently cited reasons for withdrawal in app-based studies [101,102]. Embedding customization features into platforms may address this and encourage use, as has been demonstrated with mHealth applications [103-108]. Additionally, customization such as enlarged text functionality or text-to-speech embedded software would accommodate those with accessibility needs. Designing straightforward and simple user interfaces can avoid "overwhelm" and encourage regular use [99,109], and being able to access the platform on multiple devices allows participants to prioritize familiar devices. Age-related differences can influence how a user perceives the graphical user interface of the platform. Older adults perform better with classic, skeuomorphic interfaces (mimicking real-world objects), but younger adults may prefer the minimalism of flat design (clean and straightforward) [110]. Busy or cluttered platforms risk excluding individuals with dyslexia or visual impairments [111], potentially discouraging use [41,77], and improper use of color hues risks excluding color-blind individuals [112,113]. Further, research has shown that participants value simple interfaces [41,80] in digitally enabled DCTs. Additionally, users have preexisting expectations of platform functionality [114], (eg, "hamburger icons" denoting navigation menus), which should be considered in platform design. Typography, color, formatting, and layout can also influence user experience [115-118].

The Challenge of Fully Integrated, Unified Digital Platforms

Many commercially available platform solutions are not yet open for integrating third-party components or are limited in their service provision and functionalities. Using existing components and already established solutions typically still results in a lack of interoperability, dependence on functionalities, and vendor lock-in. The future of digital platform technology, therefore, would look toward developing an end-to-end platform applying international standards where, if desired, users can integrate individual modules from different vendors.

Given the relative novelty of fully integrated, unified digital platforms for DCTs, specific interoperability standards for clinical trial platforms are not yet available. Nevertheless, commonly recommended standards are broadly applicable. We have summarized some of these standards in Table 1. The Healthcare Information and Management Systems Society [119] recommends interoperability standards in health care. These standards describe data sharing and offer recommendations regarding privacy and security. They define levels of interoperability, including technical and semantic. Future digital components used in DCTs should adapt and integrate these standards and recommendations for interoperability and modularity. As more data standardization processes become available, better ways to communicate and analyze data in trial platforms will become apparent [120].

Figure 1 provides a reference architecture of components typically comprising a DCT. Architectures provide an abstract view of systems and can be used to manage communications and decisions. Domain, application, and site requirements lead to domain reference architectures which can be further refined to application architectures. These, in turn, define the final implementation architecture. The architecture comprising a digital platform is technically complex, and a further layer of complexity is added when integrating modules from external providers. Developing such a platform would first require defined system requirements (functional and nonfunctional), user requirements (user point of view, user goals, and user input and output), and business requirements (sponsor point of view, scope of trial, and business objectives). Within our reference model, data exists in a common format and the use of preexisting standards guarantees interoperability, but this exhibits a best-case scenario. Lack of interoperability represents a significant hurdle to the development of unified digital platforms, and the lack of specific data standardization processes may prevent innovation in this technology from being explored. Moreover, given the proprietary nature of many of the available commercial platforms, specific literature on integrations, architectures, data standardization, and general deployment strategies is restricted or unavailable. Despite these challenges, the development of a fully-integrated digital platform with in-built customization and tailoring is likely to provide suppliers with a competitive advantage in the growing DCT landscape.

XSL•FO
RenderX

Table 1. Existing data standardization processes and examples illustrating how these standards can be implemented in digital platform solutions.

Publisher	Data standard	Applicable use in digital platforms
HL7[a]	FHIR[b]	Profile released to store and transmit health data in a standardized, easy, and accessible format [121,122]. Digital platforms would benefit from using international standards for health data exchange partially already used in hospital information systems. HL2 version 3 promotes clinical document architecture which can be used to save clinical reports in a standardized form.
CDISC[c]	PRM[d]	Provides a standard for planning and designing a research protocol with focus on study characteristics such as study design, eligibility criteria, and requirements from ClinicalTrials.gov, WHO[e], and EudraCT registries. PRM assists in automating case report form creation and EHR[f] configuration to support clinical research and data sharing.
CDISC	CDASH[g]	Establishes a standardized way to collect data consistently across studies and sponsors so that data collection formats and structures provide clear traceability of submission data into SDTM[h], delivering more transparency to regulators and others conducting data review.
CDISC	AdaM[i]	Defines dataset and metadata standards that support efficient generation, replication, and review of clinical trial statistical analyses, and traceability among analyses results, analyses data, and data represented in the SDTM [123]. Required standard for data submission to the FDA[j] and Japan's PMDA[k].
GS1[l]	GS1 Global Data Model	Process automation in DCTs[m]. GS1 in health care is a global, voluntary group for all participants in the health care supply chain, which includes manufacturers, distributors, health care services, solution providers, regulators, and associations. A clinical trial electronic messaging standard implementation guide is provided [121].

[a]HL7: Health Level 7.
[b]FHIR: Fast Healthcare Interoperability Resources.
[c]CDISC: Clinical Data Interchange Standards Consortium.
[d]PRM: Protocol Representation Model.
[e]WHO: World Health Organization.
[f]EHR: electronic health record.
[g]CDASH: Clinical Data Acquisition Standards Harmonization.
[h]SDTM: Study Data Tabulation Model.
[i]ADaM: Analysis Data Model.
[j]FDA: United States Food and Drug Administration.
[k]PMDA: Pharmaceuticals and Medical Devices Agency.
[l]GS1: Global Standards 1.
[m]DCT: decentralized clinical trial.

XSL•FO
RenderX

Figure 1. A reference architecture for components typically involved in decentralized clinical trials for data sharing and storage between participants' homes and study centers. Data exists in a common format with preexisting standards to guarantee interoperability and modularity. Figure adapted with permission from Trials@Home Deliverable 2.4 [124]. CDISC: Clinical Data Interchange Standards Consortium; FHIR: Fast Healthcare Interoperability Resources; GCP: good clinical practice; GS1: Global Standards 1; HIS: hospital information system; HL7: Health Level 7; IHE: Integrating the Healthcare Enterprise; REST: representational state transfer; SOP: standard operating procedure; WHO: World Health Organization.

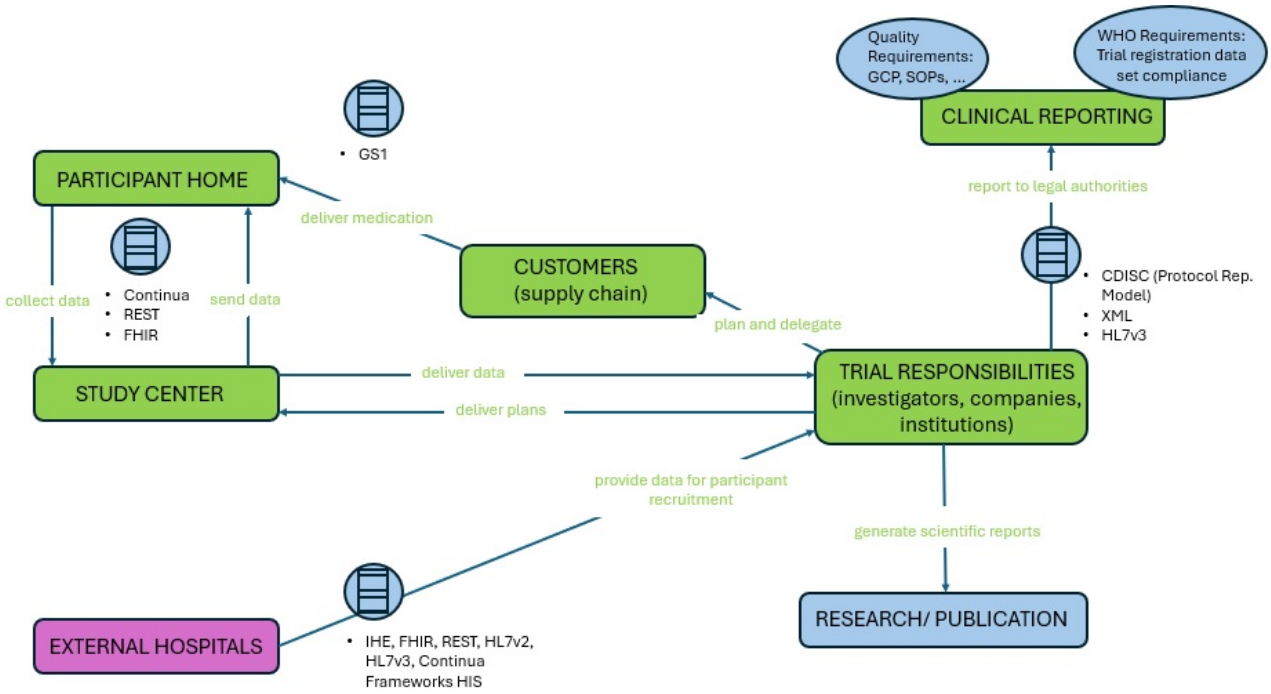

Conclusions

The growing popularity of DCTs necessitates an increasing reliance on technology. We believe the benefits of such technologies will be best realized through unified digital platforms. Platform approaches have historically been used to support DCTs, and the technology underpinning these platforms continues to develop. Tools for promoting trial engagement, maintaining investigator oversight, and improving data integrity can all be facilitated through digital platforms, and careful design planning helps accommodate the various needs of platform users. Digital platforms will likely evolve into singular, unified platforms that integrate external modules to generate a fully customizable system. However, challenges with poor interoperability and lack of data standards must first be addressed. Regardless of the current state of digital platforms for DCTs, the relationship between technology and trial innovation is likely to strengthen. Therefore, developing and implementing new DCT tools and solutions that support all stakeholders will require clinical and technical expertise.

Acknowledgments

The Trials@Home project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (grant agreement 831458). This joint undertaking received support from the European Union's Horizon 2020 research and innovation programme and European Federation of Pharmaceutical Industries and Associations. The research leading to these results was conducted as part of the Trials@Home consortium. This paper only reflects the personal views of the stated authors. No generative artificial intelligence was used in any portion of manuscript writing.

Authors' Contributions

Manuscript conceptualization is credited to RC. RC and SH wrote the original draft of the manuscript. All authors were involved in reviewing and editing the final manuscript.

Conflicts of Interest

ISM reports research grants from Menarini, EMA, Sanofi, HDR UK, National Institute for Health and Care Research (NIHR) Health Technology Assessment (HTA), and Innovative Medicines Initiative; institutional consultancy income from AstraZeneca outside of the submitted work; personal income from AstraZeneca, Amgen, Amarin, Novartis and Novo Nordisk outside of the submitted work; and vice-chairmanship of a trial steering committee for a trial funded by Novo Nordisk. TMM reports grants from the British Heart Foundation, Menarini/Ipsen/Teijin, NIHR HTA, and MSD outside of the submitted work; personal income for consultancy from Novartis and AstraZeneca, and honorarium from Viatris outside of the submitted work; membership of

XSL•FO
RenderX

independent data monitoring committees for Novo Nordisk and an NIHR trial; and is a trustee of the Scottish Heart Arterial Risk Prevention Society. AR reports unpaid membership of an NIHR trial steering committee. The remaining authors declare no competing interests.

References

1. Marquis-Gravel G, Roe MT, Turakhia MP, Boden W, Temple R, Sharma A, et al. Technology-enabled clinical trials: transforming medical evidence generation. Circulation 2019;140(17):1426-1436. [doi: 10.1161/CIRCULATIONAHA.119.040798] [Medline: 31634011]
2. Rosa C, Campbell ANC, Miele GM, Brunner M, Winstanley EL. Using e-technologies in clinical trials. Contemp Clin Trials 2015;45:41-54 [FREE Full text] [doi: 10.1016/j.cct.2015.07.007] [Medline: 26176884]
3. Steinhubl SR, McGovern P, Dylan J, Topol EJ. The digitised clinical trial. Lancet 2017;390(10108):2135. [doi: 10.1016/S0140-6736(17)32741-1] [Medline: 29143747]
4. Intelligent clinical trials: transforming through AI-enabled engagement. Deloitte Centre for Health Solutions. 2020. URL: https://www2.deloitte.com/content/dam/insights/us/articles/22934_intelligent-clinical-trials/DI_Intelligent-clinical-trials.pdf [accessed 2024-05-14]
5. Dorsey ER, Darwin KC, Mohammed S, Donohue S, Tethal A, Achey MA, et al. Virtual research visits and direct-to-consumer genetic testing in Parkinson's disease. Digit Health 2015;1:2055207615592998 [FREE Full text] [doi: 10.1177/2055207615592998] [Medline: 29942542]
6. Dorsey ER, Wagner JD, Bull MT, Rizzieri A, Grischkan J, Achey MA, et al. Feasibility of virtual research visits in Fox Trial Finder. J Parkinsons Dis 2015;5(3):505-515 [FREE Full text] [doi: 10.3233/JPD-150549] [Medline: 26406130]
7. Hirsch IB, Martinez J, Dorsey ER, Finken G, Fleming A, Gropp C, et al. Incorporating site-less clinical trials into drug development: a framework for action. Clin Ther 2017;39(5):1064-1076 [FREE Full text] [doi: 10.1016/j.clinthera.2017.03.018] [Medline: 28413148]
8. Jadhav S. Virtual clinical trials: the future of patient engagement? Applied Clinical Trials. 2016. URL: https://www.appliedclinicaltrialsonline.com/view/virtual-clinical-trials-future-patient-engagement [accessed 2024-05-14]
9. Rutkove SB, Qi K, Shelton K, Liss J, Berisha V, Shefner JM. ALS longitudinal studies with frequent data collection at home: study design and baseline data. Amyotroph Lateral Scler Frontotemporal Degener 2019;20(1-2):61-67 [FREE Full text] [doi: 10.1080/21678421.2018.1541095] [Medline: 30486680]
10. Vogel MME, Combs SE, Kessel KA. mHealth and application technology supporting clinical trials: today's limitations and future perspective of smartRCTs. Front Oncol 2017;7:37 [FREE Full text] [doi: 10.3389/fonc.2017.00037] [Medline: 28348978]
11. Rogers A, Copland R, Houston Pichardo S, Harwood K, Müller-Vössing C, Naster P, et al. Glossary of terms and definitions used. Trials@Home. 2020. URL: https://trialsathome.com/glossary/ [accessed 2024-05-14]
12. Decentralized clinical trials for drugs, biological products, and devices: guidance for industry, investigators, and other stakeholders. Food and Drug Administration. 2023. URL: https://www.fda.gov/media/167696/download [accessed 2024-05-14]
13. Rogers A, Mackenzie I, Hawkins K, Work Package 1 BEST Team of the Trials@Home Project. First set of recommendations for RDCTs (to be implemented in the pan-EU pilot RDCT). Trials@Home. 2020. URL: https://trialsathome.com/first-set-of-recommendations-for-rdcts-d1-1/ [accessed 2023-04-06]
14. Steinhubl SR, Waalen J, Edwards AM, Ariniello LM, Mehta RR, Ebner GS, et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. JAMA 2018;320(2):146-155 [FREE Full text] [doi: 10.1001/jama.2018.8102] [Medline: 29998336]
15. Carpinella I, Cattaneo D, Bonora G, Bowman T, Martina L, Montesano A, et al. Wearable sensor-based biofeedback training for balance and gait in Parkinson disease: a pilot randomized controlled trial. Arch Phys Med Rehabil 2017;98(4):622-630.e3 [FREE Full text] [doi: 10.1016/j.apmr.2016.11.003] [Medline: 27965005]
16. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019;381(20):1909-1917 [FREE Full text] [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
17. Bergenstal RM, Peyrot M, Dreon DM, Aroda VR, Bailey TS, Brazg RL, et al. Implementation of basal-bolus therapy in type 2 diabetes: a randomized controlled trial comparing bolus insulin delivery using an insulin patch with an insulin pen. Diabetes Technol Ther 2019;21(5):273-285 [FREE Full text] [doi: 10.1089/dia.2018.0298] [Medline: 31025878]
18. Wisk LE, Nelson EB, Magane KM, Weitzman ER. Clinical trial recruitment and retention of college students with type 1 diabetes via social media: an implementation case study. J Diabetes Sci Technol 2019;13(3):445-456 [FREE Full text] [doi: 10.1177/1932296819839503] [Medline: 31010315]
19. Bailey JV, Pavlou M, Copas A, McCarthy O, Carswell K, Rait G, et al. The Sexunzipped trial: optimizing the design of online randomized controlled trials. J Med Internet Res 2013;15(12):e278 [FREE Full text] [doi: 10.2196/jmir.2668] [Medline: 24334216]
20. Parker JN, Hunter AS, Bauermeister JA, Bonar EE, Carrico A, Stephenson R. Comparing social media and in-person recruitment: lessons learned from recruiting substance-using, sexual and gender minority adolescents and young adults for

XSL•FO
RenderX

a randomized control trial. JMIR Public Health Surveill 2021;7(12):e31657 [FREE Full text] [doi: 10.2196/31657] [Medline: 34855613]
21. Durkalski V, Zhao W, Dillon C, Kim J. A web-based clinical trial management system for a sham-controlled multicenter clinical trial in depression. Clin Trials 2010;7(2):174-182. [doi: 10.1177/1740774509358748] [Medline: 20083496]
22. Cramon P, Rasmussen Å, Bonnema SJ, Bjorner JB, Feldt-Rasmussen U, Groenvold M, et al. Development and implementation of PROgmatic: a clinical trial management system for pragmatic multi-centre trials, optimised for electronic data capture and patient-reported outcomes. Clin Trials 2014;11(3):344-354. [doi: 10.1177/1740774513517778] [Medline: 24519964]
23. Gupta A, Calfas KJ, Marshall SJ, Robinson TN, Rock CL, Huang JS, et al. Clinical trial management of participant recruitment, enrollment, engagement, and retention in the SMART study using a marketing and information technology (MARKIT) model. Contemp Clin Trials 2015;42:185-195 [FREE Full text] [doi: 10.1016/j.cct.2015.04.002] [Medline: 25866383]
24. Rogers A, De Paoli G, Subbarayan S, Copland R, Harwood K, Coyle J, et al. A systematic review of methods used to conduct decentralised clinical trials. Br J Clin Pharmacol 2022;88(6):2843-2862 [FREE Full text] [doi: 10.1111/bcp.15205] [Medline: 34961991]
25. Vesset D, Olofson CW. Unified data cloud for simplicity and intelligence to drive better business outcomes. International Data Corporation. 2022. URL: https://services.google.com/fh/files/misc/idc_whitepaper_unified_data_cloud_for_simplicity_and_intelligence.pdf [accessed 2024-05-14]
26. Kashyap D. Streamlining clinical trials: the impact of unified management platforms on drug development. Cloudbyz. 2023. URL: https://www.cloudbyz.com/resources/clinical-operations/streamlining-clinical-trials-the-impact-of-unified-management-platforms-on-drug-development/ [accessed 2024-05-14]
27. Peyser ND, Marcus GM, Beatty AL, Olgin JE, Pletcher MJ. Digital platforms for clinical trials: the Eureka experience. Contemp Clin Trials 2022;115:106710 [FREE Full text] [doi: 10.1016/j.cct.2022.106710] [Medline: 35183763]
28. van Dam J, Omondi Onyango K, Midamba B, Groosman N, Hooper N, Spector J, et al. Open-source mobile digital platform for clinical trial data collection in low-resource settings. BMJ Innov 2017;3(1):26-31 [FREE Full text] [doi: 10.1136/bmjinnov-2016-000164] [Medline: 28250964]
29. Dockendorf MF, Hansen BJ, Bateman KP, Moyer M, Shah JK, Shipley LA. Digitally enabled, patient-centric clinical trials: shifting the drug development paradigm. Clin Transl Sci 2021;14(2):445-459 [FREE Full text] [doi: 10.1111/cts.12910] [Medline: 33048475]
30. Doyle J, Murphy E, Gavin S, Pascale A, Deparis S, Tommasi P, et al. A digital platform to support self-management of multiple chronic conditions (ProACT): findings in relation to engagement during a one-year proof-of-concept trial. J Med Internet Res 2021;23(12):e22672. [doi: 10.2196/22672] [Medline: 34914612]
31. Spertus JA, Birmingham MC, Nassif M, Damaraju CV, Abbate A, Butler J, et al. The SGLT2 inhibitor canagliflozin in heart failure: the CHIEF-HF remote, patient-centered randomized trial. Nat Med 2022;28(4):809-813 [FREE Full text] [doi: 10.1038/s41591-022-01703-8] [Medline: 35228753]
32. Tempini N, Teira D. Is the genie out of the bottle? Digital platforms and the future of clinical trials. Econ Soc 2019;48(1):77-106 [FREE Full text] [doi: 10.1080/03085147.2018.1547496]
33. Triebell E. Enhancing decentralized clinical trials through advanced digital platforms. Health Cloud Solutions. 2022. URL: https://healthcloudsolutions.org/enhancing-decentralized-clinical-trials-through-advanced-digital-platforms [accessed 2023-04-06]
34. Harrington M. Identifying integrated digital platforms designed for a decentralized future. Applied Clinical Trials. 2022. URL: https://www.appliedclinicaltrialsonline.com/view/identifying-integrated-digital-platforms-designed-for-a-decentralized-future [accessed 2023-04-06]
35. Sadeghi M, Carenini A, Corcho O, Rossi M, Santoro R, Vogelsang A. Interoperability of heterogeneous systems of systems: from requirements to a reference architecture. J Supercomput 2023;80(7):8954-8987 [FREE Full text] [doi: 10.1007/s11227-023-05774-3]
36. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023;23(1):18. [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
37. Dudzinski MA, Garty D. A guide for establishing and governing data collection standards. Journal for Clinical Studies 2024:20-23 [FREE Full text]
38. Nosheen M, Syed Z. An integration model resolving 3rd party source code/component incorporation issues during open source software development. Int Arab J E Technol 2014;3(3):146-154 [FREE Full text]
39. Steinhubl SR, Wolff-Hughes D, Nilsen W, Iturriaga E, Califf RM. Digital clinical trials: creating a vision for the future. NPJ Digit Med 2019;2:126 [FREE Full text] [doi: 10.1038/s41746-019-0203-0] [Medline: 31872066]
40. Moorthy V, Abubakar I, Qadri F, Ogutu B, Zhang W, Reeder J, et al. The future of the global clinical trial ecosystem: a vision from the first WHO Global Clinical Trials Forum. Lancet 2024;403(10422):124-126. [doi: 10.1016/S0140-6736(23)02798-8] [Medline: 38128557]
41. Coyle J, Rogers A, Copland R, De Paoli G, MacDonald T, Mackenzie IS, Trials@Home Consortium. Learning from remote decentralised clinical trial experiences: a qualitative analysis of interviews with trial personnel, patient representatives and

XSL•FO
RenderX

other stakeholders. Br J Clin Pharmacol 2022;88(3):1031-1042 [FREE Full text] [doi: 10.1111/bcp.15003] [Medline: 34296777]
42. Pepine CJ, Handberg-Thurmond E, Marks RG, Conlon M, Cooper-DeHoff R, Volkers P, et al. Rationale and design of the International Verapamil SR/Trandolapril Study (INVEST): an internet-based randomized trial in coronary artery disease patients with hypertension. J Am Coll Cardiol 1998;32(5):1228-1237 [FREE Full text] [doi: 10.1016/s0735-1097(98)00423-9] [Medline: 9809930]
43. Pepine CJ, Handberg EM, Cooper-DeHoff RM, Marks RG, Kowey P, Messerli FH, et al. A calcium antagonist vs a non-calcium antagonist hypertension treatment strategy for patients with coronary artery disease. The International Verapamil-Trandolapril Study (INVEST): a randomized controlled trial. JAMA 2003;290(21):2805-2816. [doi: 10.1001/jama.290.21.2805] [Medline: 14657064]
44. Schoichet S, Creasey G, Kasay S. A systems infrastructure for sponsor/CRO collaboration on international clinical trials. Drug Inf J 1998;32(1):155-161 [FREE Full text] [doi: 10.1177/009286159803200122]
45. McAlindon T, Formica M, LaValley M, Lehmer M, Kabbara K. Effectiveness of glucosamine for symptoms of knee osteoarthritis: results from an internet-based randomized double-blind controlled trial. Am J Med 2004;117(9):643-649. [doi: 10.1016/j.amjmed.2004.06.023] [Medline: 15501201]
46. Jacobs BP, Bent S, Tice JA, Blackwell T, Cummings SR. An internet-based randomized, placebo-controlled trial of kava and valerian for anxiety and insomnia. Medicine 2005;84(4):197-207 [FREE Full text] [doi: 10.1097/01.md.0000172299.72364.95] [Medline: 16010204]
47. Eilenberg KL, Hoover AM, Rutherford ML, Melfi CA, Segal S. From informed consent through database lock: an interactive clinical trial conducted using the internet. Drug Inf J 2004;38(3):239-251 [FREE Full text] [doi: 10.1177/009286150403800303]
48. Gelfand AA, Qubty W, Patniyot I, Grimes B, Pletcher MJ, Goadsby PJ, et al. Home-based trials in adolescent migraine: a randomized clinical trial. JAMA Neurol 2017;74(6):744-745 [FREE Full text] [doi: 10.1001/jamaneurol.2017.0285] [Medline: 28437521]
49. Olden M, Wyka K, Cukor J, Peskin M, Altemus M, Lee FS, et al. Pilot study of a telehealth-delivered medication-augmented exposure therapy protocol for PTSD. J Nerv Ment Dis 2017;205(2):154-160 [FREE Full text] [doi: 10.1097/NMD.0000000000000563] [Medline: 27441461]
50. REDCap. Vanderbilt University. URL: https://www.project-redcap.org/ [accessed 2023-04-06]
51. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
52. Higgins SB, Jiang K, Plummer WD, Edens TR, Stroud MJ, Swindell BB, et al. Pivot/remote: a distributed database for remote data entry in multi-center clinical trials. Medinfo 1995;8:1097. [Medline: 8591379]
53. Kianersi S, Luetke M, Ludema C, Valenzuela A, Rosenberg M. Use of research electronic data capture (REDCap) in a COVID-19 randomized controlled trial: a practical example. BMC Med Res Methodol 2021;21(1):175 [FREE Full text] [doi: 10.1186/s12874-021-01362-2] [Medline: 34418958]
54. Steger FL, Donnelly JE, Hull HR, Li X, Hu J, Sullivan DK. Intermittent and continuous energy restriction result in similar weight loss, weight loss maintenance, and body composition changes in a 6 month randomized pilot study. Clin Obes 2021;11(2):e12430 [FREE Full text] [doi: 10.1111/cob.12430] [Medline: 33305526]
55. ResearchKit. Apple Inc. URL: https://researchkit.org/ [accessed 2023-04-06]
56. Crouthamel M, Quattrocchi E, Watts S, Wang S, Berry P, Garcia-Gancedo L, et al. Using a ResearchKit smartphone app to collect rheumatoid arthritis symptoms from real-world participants: feasibility study. JMIR Mhealth Uhealth 2018;6(9):e177 [FREE Full text] [doi: 10.2196/mhealth.9656] [Medline: 30213779]
57. Radin JM, Steinhubl SR, Su AI, Bhargava H, Greenberg B, Bot BM, et al. The healthy pregnancy research program: transforming pregnancy research through a ResearchKit app. NPJ Digit Med 2018;1:45 [FREE Full text] [doi: 10.1038/s41746-018-0052-2] [Medline: 31304325]
58. IQVIA. URL: https://www.iqvia.com/ [accessed 2023-04-06]
59. Medable. URL: https://www.medable.com/ [accessed 2023-04-06]
60. Castor. URL: https://www.castoredc.com/ [accessed 2023-04-06]
61. THREAD. URL: https://www.threadresearch.com/ [accessed 2023-04-06]
62. ObvioHealth. URL: https://www.obviohealth.com/ [accessed 2023-04-06]
63. Science 37. URL: https://www.science37.com/ [accessed 2023-04-06]
64. Cognizant. URL: https://www.cognizant.com/uk/en [accessed 2023-04-06]
65. AiCure. URL: https://aicure.com/ [accessed 2023-04-06]
66. Trials@Home. URL: https://trialsathome.com/ [accessed 2023-04-06]
67. Clinpal. URL: https://clinpal.com/ [accessed 2023-04-06]
68. Farias FAC, Dagostini CM, Bicca YA, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]

XSL•FO
RenderX

69. Su D, Michaud TL, Estabrooks P, Schwab RJ, Eiland LA, Hansen G, et al. Diabetes management through remote patient monitoring: the importance of patient activation and engagement with the technology. Telemed J E Health 2019;25(10):952-959. [doi: 10.1089/tmj.2018.0205] [Medline: 30372366]
70. Wintrich J, Pavlicek V, Brachmann J, Bosch R, Butter C, Oswald H, et al. Remote monitoring with appropriate reaction to alerts was associated with improved outcomes in chronic heart failure: results from the OptiLink HF study. Circ Arrhythm Electrophysiol 2021;14(1):e008693. [doi: 10.1161/CIRCEP.120.008693] [Medline: 33301362]
71. Anguera JA, Jordan JT, Castaneda D, Gazzaley A, Areán PA. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. BMJ Innov 2016;2(1):14-21 [FREE Full text] [doi: 10.1136/bmjinnov-2015-000098] [Medline: 27019745]
72. Omberg L, Neto EC, Perumal TM, Pratap A, Tediarjo A, Adams J, et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat Biotechnol 2022;40(4):480-487. [doi: 10.1038/s41587-021-00974-9] [Medline: 34373643]
73. Meirte J, Hellemans N, Anthonissen M, Denteneer L, Maertens K, Moortgat P, et al. Benefits and disadvantages of electronic patient-reported outcome measures: systematic review. JMIR Perioper Med 2020;3(1):e15588 [FREE Full text] [doi: 10.2196/15588] [Medline: 33393920]
74. Zylla DM, Gilmore GE, Steele GL, Eklund JP, Wood CM, Stover AM, et al. Collection of electronic patient-reported symptoms in patients with advanced cancer using Epic MyChart surveys. Support Care Cancer 2020;28(7):3153-3163. [doi: 10.1007/s00520-019-05109-0] [Medline: 31701269]
75. Coons SJ, Eremenco S, Lundy JJ, O'Donohoe P, O'Gorman H, Malizia W. Capturing patient-reported outcome (PRO) data electronically: the past, present, and promise of ePRO measurement in clinical trials. Patient 2015;8(4):301-309 [FREE Full text] [doi: 10.1007/s40271-014-0090-z] [Medline: 25300613]
76. Pratap A, Homiar A, Waninger L, Herd C, Suver C, Volponi J, et al. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. Sci Data 2022;9(1):522 [FREE Full text] [doi: 10.1038/s41597-022-01633-7] [Medline: 36030226]
77. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-Health: a scoping review. Telemed J E Health 2022;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
78. Sandborg J, Henriksson P, Larsen E, Lindqvist AK, Rutberg S, Söderström E, et al. Participants' engagement and satisfaction with a smartphone app intended to support healthy weight gain, diet, and physical activity during pregnancy: qualitative study within the HealthyMoms trial. JMIR Mhealth Uhealth 2021;9(3):e26159 [FREE Full text] [doi: 10.2196/26159] [Medline: 33666554]
79. Eliasen A, Abildtoft MK, Krogh NS, Rechnitzer C, Brok JS, Mathiasen R, et al. Smartphone app to self-monitor nausea during pediatric chemotherapy treatment: user-centered design process. JMIR Mhealth Uhealth 2020;8(7):e18564 [FREE Full text] [doi: 10.2196/18564] [Medline: 32706744]
80. Coyle J, Rogers A, Copland R, De Paoli G, MacDonald TM, Mackenzie IS, Trials@Home Consortium. A secondary qualitative analysis of stakeholder views about participant recruitment, retention, and adherence in decentralised clinical trials (DCTs). Trials 2022;23(1):614 [FREE Full text] [doi: 10.1186/s13063-022-06521-4] [Medline: 35907888]
81. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining "gamification". Tampere, Finland: Association for Computing Machinery; 2011 Presented at: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments; 2011; New York, NY, United States p. 9-15 URL: https://www.researchgate.net/publication/230854710_From_Game_Design_Elements_to_Gamefulness_Defining_Gamification [doi: 10.1145/2181037.2181040]
82. Sailer M, Hense JU, Mayr SK, Mandl H. How gamification motivates: an experimental study of the effects of specific game design elements on psychological need satisfaction. Comput Hum Behav 2017;69:371-380 [FREE Full text] [doi: 10.1016/j.chb.2016.12.033]
83. Pramana G, Parmanto B, Lomas J, Lindhiem O, Kendall PC, Silk J. Using mobile health gamification to facilitate cognitive behavioral therapy skills practice in child anxiety treatment: open clinical trial. JMIR Serious Games 2018;6(2):e9 [FREE Full text] [doi: 10.2196/games.8902] [Medline: 29748165]
84. Waddell KJ, Patel MS, Clark K, Harrington TO, Greysen SR. Effect of gamification with social incentives on daily steps after stroke: a randomized clinical trial. JAMA Neurol 2022;79(5):528-530 [FREE Full text] [doi: 10.1001/jamaneurol.2022.0231] [Medline: 35344027]
85. Patel MS, Small DS, Harrison JD, Fortunato MP, Oon AL, Rareshide CAL, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019;179(12):1624-1632 [FREE Full text] [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]
86. Allam A, Kostova Z, Nakamoto K, Schulz PJ. The effect of social support features and gamification on a web-based intervention for rheumatoid arthritis patients: randomized controlled trial. J Med Internet Res 2015;17(1):e14 [FREE Full text] [doi: 10.2196/jmir.3510] [Medline: 25574939]

XSL•FO
RenderX

87. Lukas CA, Eskofier B, Berking M. A gamified smartphone-based intervention for depression: randomized controlled pilot trial. JMIR Ment Health 2021;8(7):e16643 [FREE Full text] [doi: 10.2196/16643] [Medline: 34283037]
88. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PLoS One 2017;12(3):e0173403 [FREE Full text] [doi: 10.1371/journal.pone.0173403] [Medline: 28362821]
89. Blake K, Holbrook JT, Antal H, Shade D, Bunnell HT, McCahan SM, et al. Use of mobile devices and the internet for multimedia informed consent delivery and data entry in a pediatric asthma trial: study design and rationale. Contemp Clin Trials 2015;42:105-118 [FREE Full text] [doi: 10.1016/j.cct.2015.03.012] [Medline: 25847579]
90. Tarolli C, Andrzejewski K, Bull M, Goldenthal S, O'Brien M, Simuni T, et al. Virtual research visits in individuals with Parkinson disease enrolled in a clinical trial: REACT-PD study interim analysis (P4.005). Neurology 2017;88 [FREE Full text] [doi: 10.1212/wnl.88.16_supplement.p4.005]
91. Sun C, Sun L, Xi S, Zhang H, Wang H, Feng Y, et al. Mobile phone-based telemedicine practice in older Chinese patients with type 2 diabetes mellitus: randomized controlled trial. JMIR Mhealth Uhealth 2019;7(1):e10664 [FREE Full text] [doi: 10.2196/10664] [Medline: 30609983]
92. Why does data traceability matter—and how do platforms help? Aetion. 2019. URL: https://aetion.com/evidence-hub/data-traceability-how-do-platforms-help/ [accessed 2023-04-06]
93. Nugent T, Upton D, Cimpoesu M. Improving data transparency in clinical trials using blockchain smart contracts. F1000Res 2016;5:2541. [doi: 10.12688/f1000research.9756.1] [Medline: 28357041]
94. Ichikawa D, Kashiyama M, Ueno T. Tamper-resistant mobile health using blockchain technology. JMIR Mhealth Uhealth 2017;5(7):e111 [FREE Full text] [doi: 10.2196/mhealth.7938] [Medline: 28747296]
95. Maslove DM, Klein J, Brohman K, Martin P. Using blockchain technology to manage clinical trials data: a proof-of-concept study. JMIR Med Inform 2018;6(4):e11949 [FREE Full text] [doi: 10.2196/11949] [Medline: 30578196]
96. Zhuang Y, Sheets L, Shae Z, Tsai JJP, Shyu CR. Applying blockchain technology for health information exchange and persistent monitoring for clinical trials. AMIA Annu Symp Proc 2018;2018:1167-1175 [FREE Full text] [Medline: 30815159]
97. Liang X, Shetty S, Zhao J, Bowden D, Li D, Liu J. Towards decentralized accountability and self-sovereignty in healthcare systems. : Springer International; 2017 Presented at: International Conference on Information and Communications Security, 19th International Conference; 2017 Dec 08; Beijing, China p. 387-398 URL: https://link.springer.com/content/pdf/10.1007/978-3-319-89500-0_34.pdf [doi: 10.1007/978-3-319-89500-0_34]
98. De Brouwer W, Patel CJ, Manrai AK, Rodriguez-Chavez I, Shah NR. Empowering clinical research in a decentralized world. NPJ Digit Med 2021;4(1):102 [FREE Full text] [doi: 10.1038/s41746-021-00473-w] [Medline: 34211085]
99. Anderson K, Burford O, Emmerton L. Mobile health apps to facilitate self-care: a qualitative study of user experiences. PLoS One 2016;11(5):e0156164 [FREE Full text] [doi: 10.1371/journal.pone.0156164] [Medline: 27214203]
100. CTTI recommendations: decentralized clinical trials. Clinical Trials Transformation Initiative. 2018. URL: https://ctti-clinicaltrials.org/wp-content/uploads/2021/06/CTTI_DCT_Recs.pdf [accessed 2023-04-06]
101. Lau AY, Piper K, Bokor D, Martin P, Lau VS, Coiera E. Challenges during implementation of a patient-facing mobile app for surgical rehabilitation: feasibility study. JMIR Hum Factors 2017;4(4):e31 [FREE Full text] [doi: 10.2196/humanfactors.8096] [Medline: 29217504]
102. Reade S, Spencer K, Sergeant JC, Sperrin M, Schultz DM, Ainsworth J, et al. Cloudy with a chance of pain: engagement and subsequent attrition of daily data entry in a smartphone pilot study tracking weather, disease severity, and physical activity in patients with rheumatoid arthritis. JMIR Mhealth Uhealth 2017;5(3):e37 [FREE Full text] [doi: 10.2196/mhealth.6496] [Medline: 28341616]
103. Shabir H, D'Costa M, Mohiaddin Z, Moti Z, Rashid H, Sadowska D, et al. The barriers and facilitators to the use of lifestyle apps: a systematic review of qualitative studies. Eur J Investig Health Psychol Educ 2022;12(2):144-165 [FREE Full text] [doi: 10.3390/ejihpe12020012] [Medline: 35200235]
104. Adu MD, Malabu UH, Malau-Aduli AEO, Malau-Aduli BS. Users' preferences and design recommendations to promote engagements with mobile apps for diabetes self-management: multi-national perspectives. PLoS One 2018;13(12):e0208942 [FREE Full text] [doi: 10.1371/journal.pone.0208942] [Medline: 30532235]
105. Juarascio AS, Goldstein SP, Manasse SM, Forman EM, Butryn ML. Perceptions of the feasibility and acceptability of a smartphone application for the treatment of binge eating disorders: qualitative feedback from a user population and clinicians. Int J Med Inform 2015;84(10):808-816 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.06.004] [Medline: 26113461]
106. Abreu M, Carmo AS, Franco A, Parreira S, Vidal B, Costa M, et al. Mobile applications for epilepsy: Where are we? Where should we go? A systematic review. Signals 2022;3(1):40-65 [FREE Full text] [doi: 10.3390/signals3010005]
107. Morton E, Nicholas J, Lapadat L, O'Brien H, Barnes SJ, Poh C, et al. Use of smartphone apps in bipolar disorder: an international web-based survey of feature preferences and privacy concerns. J Affect Disord 2021;295:1102-1109. [doi: 10.1016/j.jad.2021.08.132] [Medline: 34706421]
108. Flaherty SJ, McCarthy M, Collins A, McAuliffe F. Can existing mobile apps support healthier food purchasing behaviour? Content analysis of nutrition content, behaviour change theory and user quality integration. Public Health Nutr 2018;21(2):288-298 [FREE Full text] [doi: 10.1017/S1368980017002889] [Medline: 29081322]

XSL•FO
RenderX

109. Peng W, Kanthawala S, Yuan S, Hussain SA. A qualitative study of user perceptions of mobile health apps. BMC Public Health 2016;16(1):1158 [FREE Full text] [doi: 10.1186/s12889-016-3808-0] [Medline: 27842533]
110. Urbano ICVP, Guerreiro JPV, Nicolau HMAA. From skeuomorphism to flat design: age-related differences in performance and aesthetic perceptions. Behav Inf Technol 2020;41(3):452-467 [FREE Full text] [doi: 10.1080/0144929x.2020.1814867]
111. Evett L, Brown D. Text formats and web design for visually impaired and dyslexic readers—clear text for all. Interacting Comput 2005;17(4):453-472 [FREE Full text] [doi: 10.1016/j.intcom.2005.04.001]
112. Sakamoto K, Kato C. Effects on memorized information quantity in web pages using bicolor design—from the perspective of color blind people and non-color blind people. Indian J Public Health Res Dev 2020;11(1):1839-1843 [FREE Full text] [doi: 10.37506/ijphrd.v11i1.1389]
113. Ellfattah A, Tawfik M. Web design for color blind persons. Int Des J 2014;4(4):37-46 [FREE Full text] [doi: 10.21608/idj.2014.109774]
114. Kirwan M, Duncan MJ, Vandelanotte C, Mummery WK. Design, development, and formative evaluation of a smartphone application for recording and monitoring physical activity levels: the 10,000 steps "iStepLog". Health Educ Behav 2013;40(2):140-151. [doi: 10.1177/1090198112449460] [Medline: 22984196]
115. Lazard AJ, Watkins I, Mackert MS, Xie B, Stephens KK, Shalev H. Design simplicity influences patient portal use: the role of aesthetic evaluations for technology acceptance. J Am Med Inform Assoc 2016;23(e1):e157-e161 [FREE Full text] [doi: 10.1093/jamia/ocv174] [Medline: 26635314]
116. Ling J, van Schaik P. The influence of font type and line length on visual search and information retrieval in web pages. Int J Hum Comput Stud 2006;64(5):395-404 [FREE Full text] [doi: 10.1016/j.ijhcs.2005.08.015]
117. Varela M, Mäki T, Skorin-Kapov L, Hoßfeld T. Towards an understanding of visual appeal in website design. : IEEE; 2013 Presented at: Fifth International Workshop on Quality of Multimedia Experience (QoMEX); July 3, 2013; Klagenfurt am Wörthersee, Austria p. 70-75 URL: https://ieeexplore.ieee.org/document/6603213 [doi: 10.1109/qomex.2013.6603213]
118. Ashraf N, Faisal MN, Jabbar S, Habib MA. The role of website design artifacts on consumer attitude and behavioral intentions in online shopping. Technical Journal 2019;24(2):50-60 [FREE Full text]
119. Healthcare Information and Management Systems Society. URL: https://www.himss.org/ [accessed 2023-04-06]
120. Lu Z, Su J. Clinical data management: current status, challenges, and future directions from industry perspectives. Open Access J Clin Trials 2010;2:93-105 [FREE Full text] [doi: 10.2147/oajct.s8172]
121. Fast Healthcare Interoperability Resources. HL7. 2011. URL: https://www.hl7.org/fhir/ [accessed 2023-04-06]
122. Clinical Data Interchange Standards Consortium. URL: https://www.cdisc.org/ [accessed 2023-04-06]
123. GS1 pharmaceutical clinical trial electronic messaging standard implementation guideline. GS1. 2021. URL: https://ref.gs1.org/guidelines/pharma-clinical-trial/ [accessed 2023-04-06]
124. Neumayer B, Schnalzer B, Alcalde B, Trappenburg J. Technology package for the pan-EU pilot study including data flows and connections with external systems where necessary. Trials@Home. 2021. URL: https://trialsathome.com/deliverables/ [accessed 2023-04-06]

Abbreviations

DCT: decentralized clinical trial
EDC: electronic data capture
REDCap: Research Electronic Data Capture

Edited by T Leung; submitted 06.04.23; peer-reviewed by R Yunis, A Blasimme, J Waalen; comments to author 21.08.23; revised version received 11.10.23; accepted 09.07.24; published 03.09.24.

Please cite as:
Copland RR, Hanke S, Rogers A, Mpaltadoros L, Lazarou I, Zeltsi A, Nikolopoulos S, MacDonald TM, Mackenzie IS
The Digital Platform and Its Emerging Role in Decentralized Clinical Trials
J Med Internet Res 2024;26:e47882
URL: https://www.jmir.org/2024/1/e47882
doi:10.2196/47882
PMID:

©Rachel R Copland, Sten Hanke, Amy Rogers, Lampros Mpaltadoros, Ioulietta Lazarou, Alexandra Zeltsi, Spiros Nikolopoulos, Thomas M MacDonald, Isla S Mackenzie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook

Rawan AlSaad[1], PhD; Alaa Abd-alrazaq[1], PhD; Sabri Boughorbel[2], PhD; Arfan Ahmed[1], PhD; Max-Antoine Renault[1], PhD; Rafat Damseh[3], PhD; Javaid Sheikh[1], MD

[1]Weill Cornell Medicine-Qatar, Education City, Doha, Qatar

[2]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

[3]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, United Arab Emirates

Corresponding Author:
Rawan AlSaad, PhD
Weill Cornell Medicine-Qatar, Education City
Street 2700
Doha
Qatar
Phone: 974 44928830
Email: rta4003@qatar-med.cornell.edu

Abstract

In the complex and multidimensional field of medicine, multimodal data are prevalent and crucial for informed clinical decisions. Multimodal data span a broad spectrum of data types, including medical images (eg, MRI and CT scans), time-series data (eg, sensor data from wearable devices and electronic health records), audio recordings (eg, heart and respiratory sounds and patient interviews), text (eg, clinical notes and research articles), videos (eg, surgical procedures), and omics data (eg, genomics and proteomics). While advancements in large language models (LLMs) have enabled new applications for knowledge retrieval and processing in the medical field, most LLMs remain limited to processing unimodal data, typically text-based content, and often overlook the importance of integrating the diverse data modalities encountered in clinical practice. This paper aims to present a detailed, practical, and solution-oriented perspective on the use of multimodal LLMs (M-LLMs) in the medical field. Our investigation spanned M-LLM foundational principles, current and potential applications, technical and ethical challenges, and future research directions. By connecting these elements, we aimed to provide a comprehensive framework that links diverse aspects of M-LLMs, offering a unified vision for their future in health care. This approach aims to guide both future research and practical implementations of M-LLMs in health care, positioning them as a paradigm shift toward integrated, multimodal data–driven medical practice. We anticipate that this work will spark further discussion and inspire the development of innovative approaches in the next generation of medical M-LLM systems.

(J Med Internet Res 2024;26:e59505) doi:10.2196/59505

KEYWORDS

artificial intelligence; large language models; multimodal large language models; multimodality; multimodal generative artificial intelligence; multimodal generative AI; generative artificial intelligence; generative AI; health care

Introduction

Large language models (LLMs) are sophisticated machine learning algorithms designed to process, understand, and generate humanlike language, enabling key developments in applications such as automated conversation, text analysis, creative writing, and complex problem-solving [1]. In health care, LLMs have shown remarkable potential, primarily through their ability to process and analyze textual content [2,3]. These models play a crucial role in assisting with diagnoses as they can efficiently process extensive textual patient histories and vast medical literature, providing clinicians with valuable insights [4-7]. However, most current LLMs are primarily limited to processing and generating textual content. While this unimodal focus on text-based operation has been transformative in the medical field, it does not fully capture the complex and diverse nature of health care practice [8].

In health care, diagnosing and treating a patient often involves a health care professional engaging in a comprehensive approach: listening to the patient, reviewing their health records, examining medical images, and analyzing laboratory test results—and all this over time. This multidimensional process

XSL•FO
RenderX

exceeds the capabilities of current unimodal LLM systems. Moreover, nontextual data types play a crucial role in diagnosis, effective treatment planning, research, and patient care [9-11]. Such data may include medical imaging (eg, x-rays, magnetic resonance imaging [MRI], computed tomography [CT] scans, positron emission tomography scans, and pathology slides), electrophysiological data (eg, electrocardiography, electroencephalography (EEG), and electromyography), sensory data (eg, data from sensors of medical devices, such as pacemakers and continuous glucose monitors), videos (eg, recordings of surgeries, procedures, and patient interactions), omics data (eg, genomics, proteomics, metabolomics, and transcriptomics), and audio data (eg, recordings of patient interviews and heart and respiratory sounds).

The introduction of LLMs has been a key development in the field of artificial intelligence (AI) and natural language processing (NLP). In 2010, the emergence of deep learning revolutionized LLMs. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks [12], allowed models to better capture sequential data and context. However, the major breakthrough occurred in 2017 with the introduction of transformer models [13], which are widely used for NLP tasks. A transformer is a type of neural network architecture that uses a self-attention mechanism to capture long-range dependencies between words in a sentence. While the computation in architectures such as RNNs and LSTM networks is sequential and slow for long sequences [14], self-attention can be parallelized and made highly scalable. Transformers have been widely trained using 2 objectives. The first objective is mask language modeling (MLM), where the objective is to learn text reconstruction by randomly masking several words in text (eg, 10%) and update the transformer weights toward this goal. Encoder transformers such as Bidirectional Encoder Representations From Transformers (BERT) [15] have been trained with the MLM objective. The second widely used objective is the next word prediction or causal language modeling. The self-attention mechanism is masked such that, at each position in the sequence, the model is able to attend only to the left words. This modeling approach mimics how text is read by humans in one direction. The self-attention mechanism allows for the computation of the probability of predicting the next word in a document by attending to the most relevant parts of the input sequence [13,16]. By applying the prediction autoregressively, the transformer model performs a text completion task by generating multiple words. Interestingly, transformers extend beyond just handling natural language data. They can effectively compute representations for various data types provided these can be represented as a sequence of tokens. The letters are elementary entities that constitute the sequences. The unique set of tokens represents the vocabulary. For example, in DNA sequence, each nucleotide could be represented by tokens from the vocabulary of 4 tokens: A, C, G, and T. This capability includes processing elements such as video frames, audio spectrograms, time-series data, code snippets, or protein sequences. BERT [15] is among the first major models to use transformers. Subsequently, a series of medical BERT models were proposed to accelerate medical research [6,17-20].

In 2022, OpenAI released ChatGPT (GPT-3.5), a significant iteration in the generative pretrained transformer (GPT) series [21]. As an LLM, ChatGPT has been trained on a vast collection of text data, which enables it to generate humanlike responses across a broad spectrum of topics and formats. ChatGPT has also shown its potential to become a valuable resource in health care, making significant contributions to various medical applications. It provided opportunities for advancing diagnostic accuracy, personalized treatment planning, and medical research, as well as optimizing health care administration and enhancing communication in patient care [22-28]. In addition, several open-source LLMs such as LLaMA [29], Flan-T5 [30], Vicuna [31], and Alpaca [32] have substantially driven progress and contributed to the field of LLMs. Although these LLM systems have achieved considerable success, they are predominantly limited to single data types. This limitation makes them less effective for the multimodal nature of medicine, where handling multiple data types is often required. Therefore, considerable efforts have been dedicated to creating LLMs that handle multimodal inputs and tasks, ultimately leading to the development of multimodal LLMs (M-LLMs). In 2023, OpenAI released GPT-4, an M-LLM with the dual capability to process and respond to both text and images. Following the release of GPT-4, several medically adapted versions of this model have been developed [33-37]. These specialized versions of GPT-4 have been tailored to interpret medical data, understand patient queries, and assist in diagnostic processes using both text and image modalities. Building on these insights, M-LLMs are increasingly recognized as systems capable of integrating various data types to facilitate comprehensive patient assessments, ensuring accurate diagnoses. In addition, they hold the potential to streamline operations, significantly improving efficiency in both clinical and administrative tasks. Most importantly, with appropriate oversight, M-LLMs could provide personalized care by tailoring treatment plans to meet the individual needs of patients, thereby enhancing the quality of health care services.

Recent studies [38,39] have explored the capabilities of M-LLMs within the health care sector. However, these studies exhibit several limitations. First, the range of data modalities examined is often restricted to text, images, videos, and audio [38], with some studies focusing narrowly on a limited number of clinical applications [39]. Second, the discussion regarding the potential uses of M-LLMs in health care is largely theoretical [38], leading to a significant gap in demonstrating their application in actual health care environments. Third, although the challenges of integrating diverse data types into M-LLMs are recognized, there is limited exploration of possible solutions or ongoing research aimed at overcoming these technical barriers [38,39].

This paper aims to present a detailed, practical, and solution-oriented perspective on the use of M-LLMs in the medical field. We unify the discussion by focusing on how M-LLMs can serve as a transformative tool that integrates various data modalities to enhance health care outcomes. Specifically, we aim to (1) broaden the analysis of M-LLM applications in health care to include additional data modalities, such as time-series data and omics data, alongside conventional

XSL•FO
RenderX

modalities such as images, text, audio, and video; (2) highlight practical examples in which M-LLMs have been or could be effectively applied in health care settings; (3) outline current technological advancements to address the technical and ethical challenges; and (4) propose future research directions to fully exploit the capabilities of M-LLMs. Our unique contribution lies in providing a comprehensive framework that links these diverse aspects, offering a unified vision for the future of M-LLMs in health care.

Background

Multimodal Learning

In the context of M-LLMs, the term *multimodal* encompasses a range of scenarios in data processing and interpretation. First, it refers to LLMs in which the input and output to the system involve different modalities, such as text-to-image or image-to-text conversions. Second, it describes LLM systems capable of handling inputs from multiple modalities, such as those that can process both text and images simultaneously. Finally, multimodality characterizes systems designed to generate outputs in >1 modality, such as systems capable of producing both textual and image-based content [40].

Several previous works have developed basic M-LLMs by aligning the well-trained encoders from different modalities with the textual feature space of LLMs. This approach enables LLMs to process inputs other than text, as seen in various examples [41-44]. For instance, Flamingo [45] uses a cross-attention layer to link a frozen image encoder with LLMs. LLaVA [46] uses a basic projection method to incorporate image features into the word embedding space. Similarly, models such as Video-Chat [47] and Video-LLaMA [48] are designed for video comprehension, whereas SpeechGPT [49] is tailored for audio processing. A notable example is PandaGPT [50], which uniquely manages to interpret 6 different modalities at the same time, achieved through the integration of a multimodal encoder known as ImageBind [51].

Despite numerous efforts focusing on understanding multimodal content at the input side, there is a significant gap in the ability to produce outputs in various modalities beyond textual content. This underscores the importance of developing any-to-any M-LLMs, which are crucial for realizing real artificial general intelligence (AGI) [52,53]. Such models should be capable of receiving inputs in any form and providing responses in the appropriate form of any modality.

From Unimodal Limitations to Multimodal Solutions

Unimodal LLMs generate content in the same modality as that in which they receive inputs, typically text, whereas M-LLMs are capable of processing inputs from various modalities and delivering outputs across multiple modalities, as illustrated in Figure 1. Despite their remarkable abilities, unimodal LLMs in medicine have inherent limitations that can be effectively overcome by shifting toward multimodal systems. In Table 1, we summarize these limitations in the medical field and illustrate how the integration of a multimodal approach can address these challenges.

Figure 1. Unimodal large language model (LLM) and multimodal LLM (M-LLM) examples. ECG: Electrocardiogram.

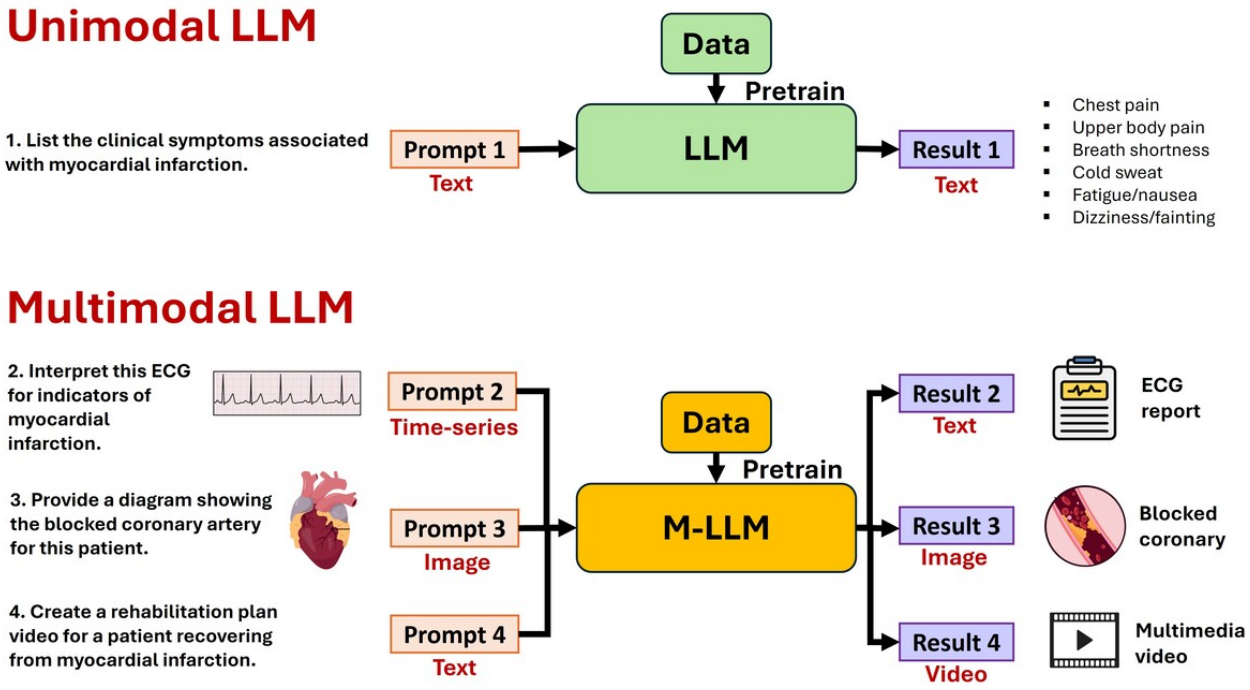

XSL•FO
RenderX

Table 1. Summary of unimodal (text) large language model (LLM) limitations in medicine and potential multimodal LLM solutions.

Unimodal (text) LLM limitation	Description of unimodal limitation	Multimodal LLM solution	Description of multimodal solution
Lack of diagnostic imaging context	Unimodal LLMs in medicine can only process textual patient data and cannot interpret diagnostic images, which are vital in many clinical scenarios.	Integration of diagnostic imaging data	Multimodal models process and integrate diagnostic imaging information (eg, x-rays and MRIs[a]), improving diagnostic accuracy and patient outcomes.
Inability to analyze temporal data	Text LLMs often struggle with interpreting time-series data, such as continuous monitoring data or progression of diseases, which are vital for tracking patient health over time.	Time-series data integration	Multimodal systems incorporate and analyze temporal data, such as ECG[b] readings or continuous monitoring data, enabling dynamic tracking of patient health and disease progression.
Absence of auditory data interpretation	Unimodal LLMs grapple with audio analysis, which limits their effectiveness in health care applications that rely on processing spoken interactions or auditory signals.	Audio data processing	Multimodal systems can process and understand audio signals, such as patient verbal descriptions and heartbeats, enhancing diagnostic precision.
Limited comprehension of complex medical scenarios	Unimodal LLMs struggle with interpreting complex medical conditions that require a multisensory understanding beyond text.	Multisensory data integration	By processing clinical notes, diagnostic images, and patient audio, multimodal systems offer more comprehensive analyses of complex medical conditions.
Overfitting to clinical textual patterns	Sole reliance on clinical texts can lead LLMs to overfit to textual anomalies, potentially overlooking critical patient information.	Diverse clinical data sources	Diversifying input types with clinical imaging and audio data allows multimodal systems to increase the number of training data points and, hence, reduce overfitting, enhancing diagnostic reliability.
Bias and ethical concerns	Unimodal LLMs, especially text-based ones, can inherit biases and misconceptions present in their training data sets, affecting patient care quality.	Richer contextual patient data	Multimodal systems use diverse modalities, including patient interviews and diagnostic images, to provide a broader context that can mitigate biases in clinical decision-making.

[a]MRI: magnetic resonance imaging.

[b]ECG: electrocardiography.

Foundational Principles of M-LLMs

Overview

The field of M-LLMs is evolving rapidly, with new ideas and methodologies being continuously developed. The training of medical M-LLMs is a complex process designed to effectively integrate and interpret the diverse types of data encountered in clinical practice. Typically, the architecture of an M-LLM system encompasses four key stages (Figure 2): (1) modality-specific encoding, (2) embedding alignment and fusion, (3) contextual understanding and cross-modal interactions, and (4) decision-making or output generation. In addition to these stages, pretraining and fine-tuning processes play a crucial role, interacting with and enhancing each of the aforementioned stages.

This section presents the foundational principles that currently guide the development and functioning of medical M-LLMs. Importantly, the specific architecture of an M-LLM might vary significantly to meet particular requirements, such as the types of data it needs to handle, the tasks it is designed to perform, and the level of interpretability and performance required. Therefore, while the stages outlined provide a high-level overview of an M-LLM system's architecture, actual implementations may vary widely to accommodate the unique demands of each application. As this field progresses, we anticipate that the guiding principles of medical M-LLMs will continue to be shaped by emerging ideas and technological advancements.

XSL•FO
RenderX

Figure 2. Main components of multimodal large language model training.

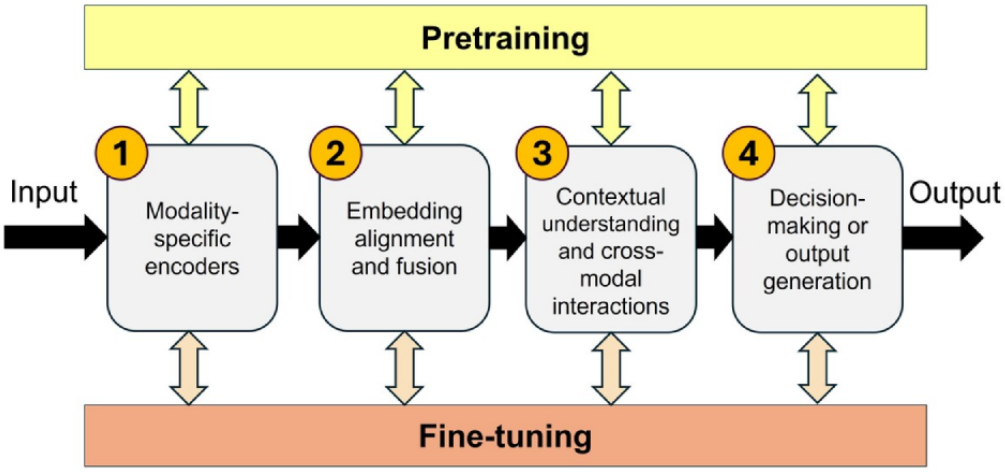

Modality-Specific Encoding

The purpose of this stage is to transform raw data from each modality into a format that the model can understand and process. This involves using modality-specific encoders to encode various data types into rich and informative representations that subsequent components of the M-LLM architecture can effectively leverage. These modality-specific encoders are meticulously trained using extensive data sets of unlabeled information to generate embeddings that accurately encapsulate the data's content. The encoders are trained in an unsupervised manner using a large collection of data sets. Selecting the appropriate encoding architecture and optimizing the training methodology are imperative and often tailored to the specific characteristics of the data and the requirements of the medical task at hand. For example, image encoders (eg, transformers [54] and convolutional neural networks (CNNs) [55,56]) are designed to capture fine-grained patterns or anomalies crucial for diagnosis, whereas text encoders (BERT [15]) aim to comprehend complex medical terminology and patient histories. Similarly, audio encoders (such as WaveNet [57] and DeepSpeech [58]) are optimized to distinguish subtle variations in sounds, such as differentiating between normal and abnormal heart or lung sounds. Time-series encoders (such as transformer-based models [15,59-61] and LSTM [12]) are intended to detect critical changes over time, signaling the need for urgent medical intervention. Finally, omics encoders (eg, DeepVariant [62], Basenji [63], and DeepCpG [64]) focus on identifying genetic markers or patterns associated with specific diseases, aiding in the development of targeted therapies.

Embedding Alignment and Fusion

The purpose of this stage is to harmonize the embeddings from different modality-specific encoders into a combined representation that reflects the combined information from all input modalities. This might involve techniques such as concatenation [65] and attention mechanisms [13] or more sophisticated methods such as cross-modal attention [66,67] and tensor fusion [68]. While modality-specific encoding relies solely on unsupervised data, embedding alignment needs annotated data across modalities. Moreover, the alignment mechanism in medical M-LLMs may require incorporating domain-specific knowledge to enhance its understanding and integration of medical data. For example, it might use known relationships between symptoms and diseases or anatomical correlations to better align and interpret the multimodal data. This results in a more accurate, reliable, and clinically relevant synthesis of information.

Contextual Understanding and Cross-Modal Interactions

The objective of this stage is that the M-LLM not only comprehends the individual modalities but also discerns their interrelations and collective contributions to the overall medical analysis or diagnostic task. This necessitates the deployment of advanced neural network architectures, notably, transformers equipped with cross-modal attention mechanisms [66,67]. These mechanisms enable the M-LLM to dynamically prioritize and integrate features across different modalities, enhancing its ability to make informed medical decisions. In addition, attention-based fusion strategies [68] could be implemented to weigh and integrate information from disparate sources, adjusting the focus of the model according to the contextual relevance of each data point from each modality.

Decision-Making or Output Generation

This component is the actionable end of the model that produces the final output or decision based on the integrated and interpreted multimodal data. This could be a classification layer capable of distinguishing between different medical conditions or a sequence generator for creating detailed medical reports. When encoder architectures are used, the model head layer can be trained for downstream classification tasks. When decoder architectures are used, the model head layer outputs logits of vocabulary tokens that can be applied in an autoregressive manner to synthesize a response. For instance, in diagnostic imaging, the model might analyze combined textual and visual embeddings to identify and categorize pathologies. In treatment recommendation systems, the model could synthesize patient history, current symptoms, and laboratory test results to suggest personalized treatment plans. The effectiveness of this stage depends on the precision of the previous components.

Pretraining and Fine-Tuning

Pretraining and fine-tuning are fundamental processes in the development and optimization of LLMs, including multimodal

XSL•FO
RenderX

ones [69]. They are not just single steps but integral, ongoing processes that influence and enhance all components of an M-LLM system's architecture. They interact with the 4 previous components of the M-LLM architecture in multiple ways.

Pretraining begins with modality-specific encoders, focusing on learning general features and representations for each modality. For instance, encoders are pretrained on large data sets to understand text, images, or audio before they are combined or applied for specific tasks. Within the embedding alignment and fusion component, pretraining enables models to learn preliminary methods for aligning and integrating embeddings from different modalities, especially in unsupervised or self-supervised setups in which the model is exposed to vast amounts of multimodal data. In the context of understanding and cross-modal interactions, pretraining lays the foundation for learning complex relationships between modalities. As the model is exposed to a wide and varied range of multimodal data, it learns to identify common patterns and interactions. Although pretraining does not directly result in final decisions or outputs for the decision-making or output generation component, it establishes essential capabilities and knowledge. This foundational understanding equips the model to later perform specific tasks more effectively.

Fine-tuning adapts a pretrained model to downstream tasks or data sets. It involves adjusting and optimizing the model's parameters and architecture components using a smaller, more task-specific data set. The fine-tuned models are capable of following instructions and responding to questions and queries. In the context of M-LLMs, fine-tuning would adjust how individual modalities are encoded, how they are aligned and fused, and how the model makes decisions based on this refined understanding.

Applications

Overview

M-LLMs hold transformative potential for numerous medical applications, demonstrating unparalleled proficiency in processing and integrating diverse data types, as shown in Figure 3. In this section, we discuss the applications of M-LLMs in clinical practice organizing them according to data type. These categories include medical images, temporal data (encompassing time-series and event data), audio, video, text, omics data, and any-to-any M-LLMs. This structured approach enables a thorough exploration of how these models can revolutionize health care practices based on their ability to synthesize and analyze complex multimodal information.

Figure 3. Applications of multimodal large language models in health care.

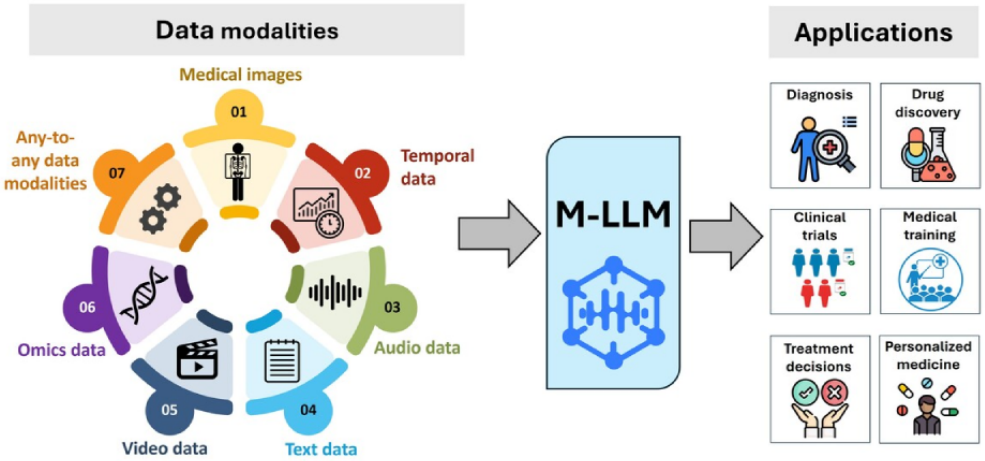

Medical Images

M-LLMs, equipped with advanced capabilities to process and interpret various image modalities, can significantly enhance diagnostic accuracy and efficiency in medical imaging applications. Examples of these image modalities include x-rays, MRI scans, CT scans, positron emission tomography scans, ultrasound images, digital pathology slides, and retinal images. Each modality provides unique insights into the body's internal structures, facilitating comprehensive analysis and aiding in the early detection, diagnosis, and monitoring of diseases. For example, in radiology, M-LLMs are instrumental in analyzing CT and MRI images to offer precise, quantifiable data for identifying and characterizing anomalies such as tumors, fractures, and signs of chronic diseases. In addition, these models support the generation of automated radiological reports that summarize findings and suggest potential diagnoses. It is also possible to use M-LLM embedding to retrieve similar cases based on keyword searching. Conversely, M-LLMs allow for the annotation and tagging of medical images with keywords. This enables additional analytics applications. Similarly, in pathology, M-LLMs interpret tissue sample slides, identifying disease markers that are often subtle and challenging to discern. In dermatology, M-LLMs apply their image analysis processes to assess photos of skin lesions, aiding in the early detection of skin cancers such as melanoma [70].

Significant progress has been made in the field of general-domain image-text M-LLMs through the development of models such as GLaMM [71], Qwen-VL [72], SpatialVLM [73], InternVL [74], Osprey [75], Vary [76], ShareGPT4V [77], OtterHD [78], LION [79], SPHINX [80], BLIVA [81], SVIT [82], LLaVA [46], and CoVLM [83]. These advancements have been made possible by leveraging billions of image-text pairs predominantly sourced from the public web, enabling these models to analyze and integrate visual and textual information

XSL•FO
RenderX

to understand and generate complex and contextually relevant responses. Such M-LLMs with vision capabilities can be adapted for medical imaging applications (eg, LLaVA-Med [84], PMC-VQA [85], Med-Flamingo [86], and PeFoMed [87]). However, an important question arises regarding whether such general-domain models can deeply understand medical images or whether they simply recognize superficial patterns from extensive pretraining. Previous work [88] evaluated the performance of a general-domain M-LLM in biomedical image classification tasks. The study aimed to determine whether such M-LLMs can develop usable internal representations of medical images and whether these representations could effectively distinguish between various medical subclasses. The results showed that generalist models can inherently understand medical images and, in some medical contexts, even outperform specialized, task-specific pretraining methods. Therefore, using representations from generalist models may offer a data-effective solution for developing classification models in the medical imaging domain.

Temporal Data

M-LLMs with the ability to process and interpret time-stamped sequences of data offer significant potential in areas such as real-time patient status tracking in intensive care units, longitudinal studies for chronic disease management, and predictive analytics for patient risk assessment. M-LLMs designed with temporal dimensions acquire predictive capability and skills in extrapolating the understanding of medical conditions over time. Temporal data include time-series, spatiotemporal, and event data. For the purpose of this paper, our focus will be on time-series and event data.

Time-series data are a sequence of data points collected or recorded at regular time intervals, with each data point being time-stamped. Examples include a patient's heart rate recorded over time and continuous glucose monitoring (CGM). In critical care settings, M-LLMs can detect early signs of clinical deterioration, such as sepsis or cardiac events, from continuous monitoring of vital signs. In neurology, M-LLMs process EEG data to detect neurological anomalies, such as seizure patterns.

Event data are a record of discrete actions or occurrences at specific time points or over intervals. Unlike time-series data, they do not have to be regularly timed. Examples include electronic health records (EHRs) detailing various discrete events in a patient's medical history, such as physician visits, hospital admissions, and prescription records, or sensor data recording specific occurrences, such as motion sensors being triggered with movement. Each event is time-stamped but does not occur at regular intervals. M-LLMs are instrumental in extracting meaningful insights from EHRs, which encompass diverse and nonregularly timed medical events [89]. M-LLMs can analyze the sequence and context of these events, providing a comprehensive understanding of a patient's medical history. This analysis can lead to more accurate diagnoses, tailored treatment strategies, and improved management of chronic conditions. In addition, M-LLMs can process sensor data, such as motion sensor activations in older adult care settings, offering real-time insights into patient activity and well-being.

Significant advancements have been made in M-LLMs with temporal analysis capabilities, including models such as Time-LLM [90], LLM4TS [91], TEMPO [92], and PromptCast [93], among others [94,95]. However, there is still a lack of M-LLMs specifically designed for medical temporal data. Some of the existing M-LLMs with temporal capabilities could be adapted for medical applications [89,96], or new models specifically designed and pretrained on medical temporal data can be developed.

Audio

Medical M-LLMs that can process and comprehend audio signals have the potential to significantly enhance health care. These models can analyze vocal patterns and breathing sounds to identify respiratory conditions such as asthma or chronic obstructive pulmonary disease (COPD) early in their development. In addition, M-LLMs can be used in mental health to detect subtle changes in speech patterns, affective tone, and vocal tone that may indicate depression, anxiety, or stress, offering a noninvasive diagnostic tool that complements traditional assessment methods. Moreover, audio-based M-LLMs facilitate continuous monitoring of patients in intensive care unit settings, using sound analysis to alert medical staff to changes in patient condition that might necessitate immediate intervention. Furthermore, these models enhance patient engagement and education by converting medical advice into accessible audio formats tailored to individual patient needs and comprehension levels. They can also aid in the early detection of neurological disorders through speech irregularities, help monitor sleep apnea by analyzing breathing patterns during sleep, and support speech therapy for stroke survivors by tracking progress in speech fluency and pronunciation.

Numerous audio-text M-LLMs, leveraging transformer-based architectures, have integrated text- and audio-based language models, such as AudioPaLM [97], AudioLM [98], Pengi [99], AudioGPT [100], SpeechGPT [49], VioLA [101], and SALMONN [102], into a unified multimodal architecture. This architecture is capable of processing and generating both text and speech, facilitating applications such as speech recognition and speech-to-speech translation. However, there is a gap in the development of large audio models specifically tailored for medical applications [103]. Nonetheless, these existing M-LLMs with audio capabilities may be adapted and refined to address the requirements of medical-related tasks.

Text

Although text-based LLMs are not inherently multimodal, integrating text with other data modalities such as images and audio transforms them into the core of M-LLMs. In clinical practice, these text-based components of M-LLMs can be applied in several ways. For instance, they facilitate the automated generation of patient reports by interpreting and summarizing complex medical language and data, including diagnostic imaging and laboratory test results. M-LLMs with additional skills in understanding tabular and other structured textual data are expected to perform better on EHR data. Furthermore, text M-LLMs play a crucial role in analyzing the large volumes of clinical notes routinely available in EHRs to predict clinical outcomes. In addition, they enhance medical

XSL•FO
RenderX

education and training by providing simulations and interactive learning experiences based on extensive medical literature and case studies.

There is a growing interest in the development of M-LLMs that incorporate text data, demonstrating the vast potential and ongoing innovations in this field. Examples of biomedical text LLMs include BiMediX [104], BioBERT [105], PubMedBERT [106], and ClinicalBERT [20]. BioBERT is a biomedical language representation model designed for text mining tasks such as named entity recognition, relation extraction, and question answering in the biomedical domain. PubMedBERT is specifically pretrained from scratch on PubMed articles, ensuring a highly focused approach to understanding medical literature. ClinicalBERT is a BERT model pretrained on generic EHR clinical documentation and discharge summaries. BiMediX is the first bilingual medical LLM with expertise in both English and Arabic, facilitating several medical interactions, including multiturn conversations, multiple-choice queries, and closed question answering.

Videos

M-LLMs hold significant promise in transforming the analysis and interpretation of various types of video data within medical settings. In surgical training, M-LLMs can analyze and interpret surgical videos, providing real-time feedback and educational insights. In physical therapy, M-LLMs can analyze patient movement videos, aiding in designing targeted rehabilitation programs and monitoring patient progress. They can also be used in psychiatric evaluations to assess behavioral patterns through video assessments. Furthermore, M-LLMs can be used in internal examinations, interpreting recordings from endoscopic and laparoscopic procedures to identify abnormalities and support real-time decision-making during these procedures. Their applications extend to home health care, allowing for remote patient monitoring through video to track well-being. They are also used in sleep studies, where video recordings assist in diagnosing disorders such as sleep apnea. In dermatology, video analysis of skin conditions over time helps in tracking disease progression.

The progress in M-LLMs for video data analysis, demonstrated by models such as Video-Chat [47], Video-ChatGPT [107], Video-LLaMA [48], LLaMA-VID [108], MotionGPT [109], LAVENDER [110], MovieChat [111], Vid2Seq [112], VideoLLM [113], and VTimeLLM [114], shows significant promise for the development of models tailored to medical applications. The success of these models in nonmedical settings lays a foundation for similar advancements in the health care sector. However, a critical aspect in applying these models to medicine is the incorporation of domain-specific medical knowledge. Medical videos require not just technical analysis but also contextual interpretation aligned with patient history, presenting symptoms, and potential diagnoses. Furthermore, the operational demands of these models in clinical environments are stringent. They must function in real time or near real time to offer actionable insights during critical medical procedures, such as providing alerts during surgeries or continuous patient monitoring.

Omics Data

M-LLMs leveraging omics data, encompassing genomics, transcriptomics, proteomics, and other omics technologies, have the potential to significantly enhance personalized medicine and clinical diagnostics. By integrating and interpreting complex omics data sets, M-LLMs can uncover novel biomarkers for diseases, predict patient responses to specific treatments, and facilitate the development of targeted therapies. For example, in oncology, these models can analyze genetic mutations and expression patterns to guide cancer treatment strategies. Similarly, in cardiology, omics data analysis can help identify genetic risk factors for heart diseases, enabling preventative interventions. M-LLMs also support drug discovery processes by predicting the efficacy and side effects of potential drugs based on the omics profiles of diverse patient populations.

Several M-LLMs have been developed using omics data for a wide range of biomedical applications [115]. In genomics, DNA sequence language models are used for a variety of predictive tasks. These tasks include predicting genome-wide variant effects (GPN [116]; DNABERT [117]; and its subsequent evolution, DNABERT-2 [118]), predicting DNA cis-regulatory regions (DNAGPT [119], DNABERT, and DNABERT-2), predicting DNA-protein interactions (TFBert [120] and MoDNA [121]), and determining RNA splice sites from DNA sequences (DNABERT and DNABERT-2). In transcriptomics, RNA sequence language models are used for RNA splicing prediction (SpliceBERT [122]), assessment of long noncoding RNAs' coding potential (LncCat [123]), RNA-binding protein interactions (BERT-RBP [124]), RNA modification identification (BERT-m7G [125]), and predictions related to protein expression and messenger RNA degradation (CodonBERT [126]). In proteomics, protein language models are used for secondary structure and contact prediction (ProtTrans [127]), protein sequence generation (ProGen [128]), protein function prediction (ProtST [129]), major posttranslational modification prediction (ProteinBERT [130]), biophysical property prediction (PromptProtein [131]), and advancing the state of the art in proteomics [132,133].

Any-to-Any M-LLMs

Current M-LLMs are primarily limited to multimodal comprehension on the input side, possessing limited capabilities to generate content across various modalities [134,135]. Given that clinicians frequently interact and communicate using a variety of medical modalities, the potential applications of any-to-any M-LLMs, which can accept input in any modality and produce output in any modality, are numerous. For instance, clinicians can provide a combination of textual patient history, radiographic images, and audio recordings of patient symptoms as input to the M-LLM. The M-LLM could then analyze this multimodal input to diagnose the patient's condition. Subsequently, it could generate a multimodal output that includes a textual report summarizing the diagnosis, annotated images highlighting areas of concern, and an audio explanation that can be easily shared with patients or other medical professionals.

There is an increasing interest in the development of any-to-any M-LLMs, highlighting the significant potential of their

XSL•FO
RenderX

applications across various domains. For instance, NExT-GPT [136] enhances an LLM with multimodal adapters and a range of diffusion decoders, enabling the model to process and generate outputs in any combination of text, images, videos, and audio. Macaw-LLM [137] integrates images, audio, and textual data using 3 primary components: a modality module for encoding multimodal data, a cognitive module for leveraging pretrained LLMs, and an alignment module for synchronizing diverse representations. OneLLM [138] incorporates 8 unique modalities within a single framework using a multimodal alignment pipeline, which can be further expanded to include additional data modalities. These models, among others [139,140], can be tailored and fine-tuned to specifically address the unique demands of tasks related to health care.

Use Case Example

In this section, we present a use case that demonstrates the practical application of M-LLMs in health care using the Contrastive Learning From Captions for Histopathology (CONCH) model [141]. CONCH is a vision-language M-LLM specifically designed for computational histopathology. It is pretrained on the largest histopathology-specific vision-language data set, enabling it to create effective representations for non–H&E (hematoxylin and eosin)-stained images, such as immunohistochemistry and special stains, without relying on large public histology slide collections such as The Cancer Genome Atlas, Pancreatic Cancer AI Platform, and Genotype-Tissue Expression.

For this experiment, we used the pretrained model weights available on Hugging Face [141] and installed the CONCH package from the official repository [142]. The experiment was conducted on a Linux machine equipped with an NVIDIA GeForce GTX 1080 Ti graphics card using a web-based demonstration application developed using the Flask web framework. The application created a ChatGPT-like interface for zero-shot cross-modal retrieval, accepting both pathology-related text prompts and pathological images. It computed cosine similarity and provided retrieval scores based on the input data. Figure 4 illustrates how CONCH was used to analyze 2 histopathology slides, providing confidence scores for various diagnostic questions. The model processes both the images and corresponding text prompts, offering a zero-shot cross-modal retrieval approach to assist in diagnosing conditions such as invasive ductal carcinoma, invasive lobular carcinoma, and ulcerative colitis.

This use case example highlights the potential of M-LLMs such as CONCH to enhance computational pathology by enabling advanced, multimodal data retrieval and analysis even in complex and specialized medical imaging tasks.

Figure 4. Demonstration of the Contrastive Learning From Captions for Histopathology (CONCH) model as a text-vision foundation model for histopathology analysis.

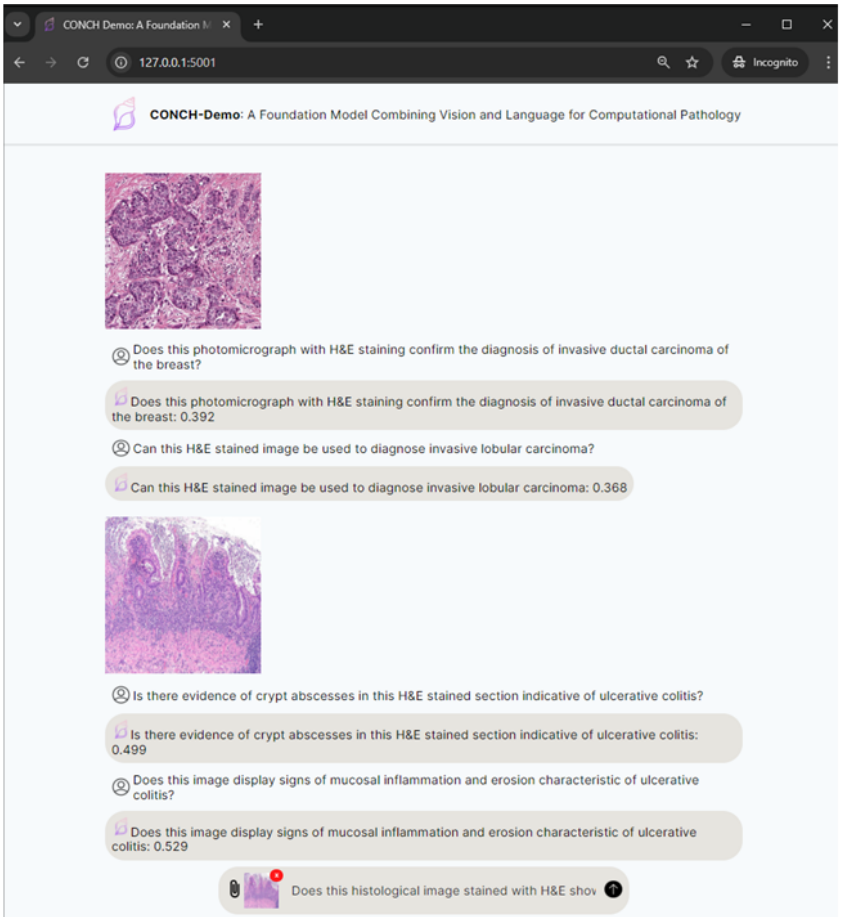

XSL•FO
RenderX

Challenges

Overview

While the potential of M-LLMs is promising, it is crucial to understand the significant technical and ethical challenges and limitations that accompany their development and deployment in health care (Figure 5). From a technical perspective, challenges include integrating diverse data sources (data fusion), meeting extensive data requirements, ensuring scalability and managing computational demands, and improving the interpretability of M-LLMs. Ethically, issues such as bias and fairness, obtaining informed consent, data privacy and security, and the safety and alignment of these models in clinical practice present substantial obstacles. In this section, we discuss these challenges and propose potential solutions to tackle them.

Figure 5. Challenges of multimodal large language models in health care.

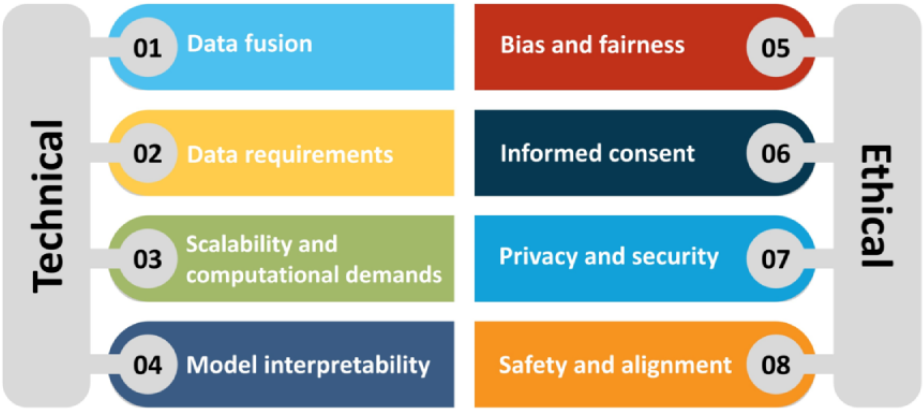

Technical Challenges

Data Fusion

Problem

Data fusion in medical M-LLMs is a sophisticated and complex process that requires the integration of heterogeneous data types to create a comprehensive and multidimensional representation of patient health. This integration process encompasses several technical challenges that must be adeptly managed. The first challenge is the temporal and spatial alignment of different data modalities, where aligning data from diverse sources such as medical images, videos, and text-based records is crucial to ensure that all data points are synchronized and that temporal data (showing changes over time) and spatial data (showing anatomical or physiological details) are correctly correlated. Second, handling data sparsity and missingness is vital as it can significantly impact diagnosis and treatment. For example, missing frames in a medical video could miss critical changes in a patient's condition, incomplete medical images may not fully reveal the extent of a disease, and gaps in EHRs can result in a lack of historical context for patient care, necessitating sophisticated techniques to infer missing information without compromising diagnostic accuracy. Furthermore, normalization and standardization are essential given the varied formats, scales, and resolutions of different data modalities, for example, adjusting the scale of medical images to a standard range, normalizing text data from clinical notes to a uniform format for analysis, and standardizing video data to ensure consistent frame rates and resolutions. These challenges highlight the complexity of integrating diverse data types used in M-LLMs, underscoring the need for advanced computational techniques and algorithms to address these issues effectively.

Potential Solution

Beyond foundational methods for data fusion, a variety of advanced techniques exist that can enable M-LLMs to more effectively integrate different modalities. Prompt-based multimodal fusion [143] is one such framework that enables bidirectional interaction among different modalities through a 2-stream structure, typically involving parallel construction of the multimodal model through pretrained language and image models. Hybrid fusion [144] integrates unstructured and structured data along with other multimodal sources via a pretrained language model, capturing a more comprehensive patient representation. Gated fusion [145,146] uses mechanisms such as neural network gates or attention mechanisms to dynamically emphasize or de-emphasize different aspects or modalities of the data based on the context. Finally, tensor fusion [68] constructs a higher-order tensor representing all feature combinations across modalities, which is then decompressed or factorized to a lower dimension for tractable computation while preserving the depth of multimodal interactions.

Data Requirements

Problem

In the pretraining phase of M-LLMs, large and diverse data sets with extensive labeling in many cases are required to capture a wide range of general knowledge across different modalities (eg, text, images, and audio). The primary goals of pretraining are to develop robust feature representations and ensure that the model can handle the inherent variability in real-world data. However, such multimodal medical data sets are currently limited, and the acquisition of such large-scale labeled data presents logistical, ethical, and privacy challenges [147]. Existing multimodal medical data sets available for public use [84,85,148] are often relatively small in scale and demand the consolidation of numerous resources. For instance, the MIMIC-IV [148] includes a limited range of modalities,

XSL•FO
RenderX

including clinical notes, medical images (chest x-ray Digital Imaging and Communications in Medicine [DICM] images), and time series (diagnostic electrocardiogram and patient records), making it a valuable but constrained resource for training medical M-LLMs. Similarly, PMC-VQA [85] and LLaVA-Med [84] include text and image modalities for medical visual question answering.

It is to be noted that the storage of vast amounts of multimodal data (ie, medical images and scans, videos, and high-resolution audio files) requires substantial storage capacity. Efficient and secure storage solutions are essential to handle these data, ensuring quick access and retrieval while maintaining data integrity and security.

Potential Solution

To address the limited data challenge in training medical M-LLMs, a combination of synthetic data generation and federated learning could be used. Synthetic data generation using generative models can create realistic, diverse data sets that mimic real-world multimodal medical scenarios, thus expanding the training data set without compromising privacy or ethical standards [149-151]. In addition, federated learning presents a viable solution for leveraging multimodal data from multiple health care institutions without the need to share the actual data, thus maintaining patient privacy [152-156]. This decentralized approach enables multimodal M-LLMs to learn from a vast, distributed data set encompassing a wide range of medical modalities without necessitating centralization of the data.

The few-shot learning approach enables models to generalize from a limited number of examples. By leveraging the pretrained knowledge and adapting quickly to new tasks with minimal data, few-shot learning can be particularly useful in medical scenarios in which labeled data are limited. Another approach to reducing computational requirements and addressing the problem of unavailable labeled data is in-context learning. This approach enables models to perform tasks by providing examples in the input context without fine-tuning the model weights. This approach can be effective for tasks such as medical image interpretation or clinical note analysis.

To address data storage demands when building M-LLMs, cloud-based storage solutions offer a flexible and scalable way to store big data and allow organizations to scale their storage capacity as needed without the upfront investment in physical infrastructure. Other benefits include improved accessibility and cost efficiency, whereas providers can implement robust security measures (eg, data encryption and access control). Moreover, the combination of cloud-based storage and distributed storage systems provides a robust and adaptable solution for managing the extensive and complex data sets needed for M-LLMs.

Scalability and Computational Demands

Problem

The development and deployment of M-LLMs in the medical field pose significant scalability and computational challenges. During training, such complex M-LLMs require substantial computational power, often involving parallel processing and sophisticated algorithms to manage and analyze the data effectively. Moreover, M-LLMs face memory limitations due to processing vast amounts of data, and their large size necessitates considerable storage capacity. This can also lead to network latency, slowing down model performance and affecting user experience. The scalability issue is further compounded by the need for continuous model updates to incorporate new medical data and knowledge. These factors translate to high operational costs, making the development of medical M-LLMs feasible mainly for large technology corporations with significant resources. Inference, on the other hand, requires minimizing latency and reducing computational load to ensure real-time or near–real-time responses in clinical settings. Both phases pose unique challenges that need to be addressed to facilitate the practical deployment of M-LLMs in health care.

Potential Solution

To optimize efficiency during both training and inference, several methods can be used. Parameter-efficient fine-tuning methods such as adapter layers help reduce the computational load by fine-tuning only a subset of the model's parameters [157,158]. In addition, quantization approaches can address the scalability and computational demands by shifting toward quantized versions of existing models using curated, domain-specific data rather than pretraining from scratch [159]. This method capitalizes on the foundational strengths of established models, significantly reducing the computational resources needed for initial training [160]. Knowledge distillation is another approach that involves training a smaller "student" model to replicate the behavior of a larger "teacher" model, requiring less computational power while retaining performance [161]. Fine-tuning using targeted medical data sets enhances accuracy and relevance in medical applications while also cutting down development time and costs. Furthermore, developing more efficient transformer architectures tailored for multimodal data, such as Kosmos-1 [162], Muse [163], and PixArt-α [164], presents a viable solution. Optimizing algorithms for parallel processing is another approach that promotes more efficient use of computational resources. During inference, quantization and pruning continue to be beneficial by reducing the computational burden and speeding up model execution. Knowledge distillation allows for the use of smaller, faster models that maintain high performance, ideal for real-time applications. Additional optimization techniques, such as model compression [165] and hardware acceleration using graphics processing units (GPUs) or tensor processing units (TPUs) [166], further enhance efficiency.

Model Interpretability

Problem

In contrast to unimodal LLMs, the scale of M-LLMs in terms of parameters and training data introduces a unique set of interpretability challenges alongside potential opportunities in the field of research on model explainability. First, as these models expand in size, the task of understanding and interpreting their decision-making processes becomes increasingly challenging [167]. This difficulty is amplified by the added

XSL•FO
RenderX

internal complexity of M-LLMs and the extensive variety of their training data sets. Moreover, this complexity necessitates substantial computational resources to facilitate the generation of explanations. Such increased complexity poses significant hurdles for in-depth analysis, thereby hindering the debugging and diagnostic processes essential for understanding and improving M-LLMs.

Potential Solution

Addressing these interpretability challenges in the context of health care is critical as clinicians—accountable to patients and regulators—should have a reasonable ability to explain how a complex model assists and makes medical recommendations. Choosing between model performance and interpretability can be problematic and is often down to trust (in model development methods, data, metrics, and outcome data, among other things). This challenge necessitates the development of advanced methods for explaining transformer-based language models [167,168], particularly methods for local explanations, such as feature attribution explanation, attention-based explanation, example-based explanation, and natural language explanation [169-172], and global explanations, such as probing-based explanation, neuron activation explanation, concept-based explanation, and mechanistic interpretability [168,173,174]. In addition, being able to use these explanations is crucial for debugging and improving M-LLMs. An effective approach is the development of integrated explanation frameworks specifically designed for medical M-LLMs that can integrate both local and global explanations. Such frameworks are essential for handling the multimodal nature of medical data, including the combination of textual and imaging information. In addition, incorporating a human-in-the-loop approach, where clinician feedback on the model's explanations is used for continuous improvement, can significantly enhance the practical utility and trustworthiness of these M-LLM systems in medical settings [167].

Ethical Challenges

Bias and Fairness

Problem

The potential for bias represents one of the primary ethical challenges in using M-LLMs in health care. Specifically, in the health care domain, data often exhibit bias due to the uneven distribution of demographic attributes, preconceptions held by health care professionals involved in data collection and interpretation, and the varied academic and experiential backgrounds that influence their perspectives [175-177]. If M-LLMs are trained on patient data that contain biases related to gender, ethnicity, socioeconomic status, or geographic location, they may inadvertently cause biases in their predictions or recommendations [175,178,179]. For example, a recently developed M-LLM, LLaVA [46], when asked to analyze an image featuring 2 Black men and 2 gorillas, erroneously identified one of the men as a gorilla. This error suggests the existence of racial bias within the algorithmic framework of the model [180]. In health care, biased M-LLMs can lead to differential treatment, misdiagnoses, and unequal access to medical resources. For example, an M-LLM analyzing medical images might miss subtle symptoms in darker-skinned individuals due to biases in the training data. One study showed that CNNs, when trained on publicly available chest x-ray data sets, may show a tendency to underdiagnose specific populations, such as individuals from marginalized communities (eg, Black and Hispanic patients), women, and Medicaid recipients [181].

Potential Solutions

Mitigating bias and improving fairness within medical M-LLMs necessitates a multifaceted approach centered on 3 pillars: data integrity, model refinement, and comprehensive evaluation [181,182]. Essential to this strategy is the curation of diverse and representative data. This involves compiling multimodal medical data sets that encompass a wide array of demographics, languages, and cultures to ensure balanced representation and guide targeted model fine-tuning efforts [183]. Fine-tuning these models through transfer learning and bias reduction techniques, such as counterfactual data augmentation [184], can effectively minimize patterns of gender, racial, or cultural bias. Furthermore, deploying multiple methods and metrics for evaluation is crucial. These may include human, automatic, or hybrid evaluations alongside metrics such as accuracy, sentiment, and fairness, which provide feedback on bias in M-LLM outputs. Through such rigorous evaluation, biases can be detected and continuously addressed, improving the reliability of M-LLMs. Moreover, incorporating logic-aware mechanisms into medical M-LLMs involves integrating clinical reasoning and decision-making processes into the M-LLMs. This approach promotes the generation of more accurate and less biased outputs by applying medical reasoning to the relationships between data tokens. For instance, logic-aware M-LLMs can differentiate between correlational and causal relationships in patient data, recognize the significance of laboratory values within clinical contexts, and apply diagnostic criteria accurately across diverse patient populations. Ultimately, the goal is to reduce bias without compromising the performance of M-LLMs. It is a careful balance of debiasing and enhancing the models, requiring ongoing monitoring and adjustment to align with ethical standards, particularly in the sensitive domain of health care [185].

Informed Consent

Problem

Obtaining informed consent in the context of M-LLMs presents unique challenges. In particular, it remains uncertain whether patient consent is necessary for training M-LLMs using their data if consent was previously obtained for research purposes in general or for AI development specifically [178,186]. Furthermore, given the complexity of M-LLMs, it might be difficult for patients to grasp what they are consenting to, especially in terms of how their data will be used, how these models operate, and the potential risks involved. This raises questions about the validity of consent and the level of detail required to adequately inform patients [177,178]. In addition, it can be argued that traditional institutional review boards (IRBs) and ethical oversight committees may be ill-equipped to deal with AI and M-LLM applications due to the lack of

XSL•FO
RenderX

understanding of such novel technologies in the medical arena [187].

Potential Solutions

Health care providers and developers have a responsibility to empower patients to make informed decisions about the use of their data in developing M-LLMs. This requires providing them with clear, transparent, simplified explanations of how M-LLMs work, how their data will be used, the nature of the data they handle, the steps taken to protect privacy, and the potential risks of using their data (eg, algorithmic bias and privacy issues). These explanations may take various forms, including written text, visual aids, educational videos, or other materials tailored to different levels of understanding. Professional training should be provided to health care professionals on the capabilities, limitations, and ethical considerations of using M-LLMs in practice to effectively communicate these aspects to patients. To this point, it may be necessary for health care and academic medical institutions to adapt their IRBs for a more effective governance and use of AI, first through incorporating a sufficiently diverse set of expert members (eg, experts in machine learning, experts in data science, and experts in previous studies of marginalized or discriminated communities) and, second, through more targeted, ongoing training of board members. In doing so, IRBs are more likely to constructively navigate issues pertaining to informed consent, data privacy and security, and safety.

Data Privacy and Security

Problem

As mentioned previously, M-LLMs require a massive amount of patient data (eg, medical history, clinical notes, medical images, laboratory test results, and prescriptions) that are inherently sensitive. This, in turn, raises substantial privacy and security concerns—how will patient data be collected, stored, and used? Who will have access to them and for what purposes [175-177]? Researchers have demonstrated that bombarding an LLM with specific questions (ie, adversarial attacks) could force it to expose its training data, which contain verbatim personal identifiable information and chat conversations [188]. They have also concluded that larger models seem to be more susceptible to attacks than smaller models [188]. Other studies have shown that, even when sensitive patient data are anonymized, certain algorithms can still identify individual patients [189-191]. Unauthorized access or breaches can have severe consequences, including reputational damage, misuse of personal health information, and compromise of patient confidentiality.

Potential Solutions

It is crucial to implement stringent data protection measures to mitigate data privacy and security concerns when using patient data for developing M-LLMs. One of these measures is the implementation of federated learning techniques [153,155,156] to enable M-LLMs to be trained on decentralized data sources without the need to transfer sensitive or private information to a central location, thereby preserving data privacy and security. Furthermore, robust encryption protocols and anonymization techniques should be applied to the data before transferring or processing them. Secure storage infrastructure should be in place to safeguard patient information. It is important to conduct auditing of M-LLMs using data extraction attacks to understand how well M-LLMs resist unauthorized attempts to extract data and identify areas for improvement in terms of security and privacy. Health care providers and developers must establish strong data governance frameworks and policies and comply with relevant privacy regulations (eg, Health Insurance Portability and Accountability Act [HIPAA]). They also need to adopt a proactive approach to cybersecurity and regularly update security measures to counter-emerging threats.

Safety and Alignment

Problem

Ensuring the safety and alignment of M-LLMs in health care is paramount. These models must not only be effective in processing and analyzing medical data but also align with human ethical standards, particularly those of health care professionals. Similar to text-based models, where fine-tuning, reinforcement learning from human feedback, and dynamic policy optimization (DPO) are used to minimize harm and align outputs with human preferences, M-LLMs could adopt analogous methodologies to ensure that their recommendations are in harmony with the preferences and ethical considerations of medical practitioners. The challenge lies in aligning M-LLMs with the complex, nuanced, and sometimes subjective decision-making processes of human physicians. This involves training models on a diverse array of scenarios, encompassing ethical dilemmas, treatment preferences, and patient-centered care principles. By integrating feedback loops in which health care professionals review and adjust model outputs alongside technical and other professionals, M-LLMs can learn to prioritize patient safety, privacy, and the nuances of human empathy and ethical considerations in their recommendations.

Potential Solutions

Developing a framework for continuous learning and adaptation is crucial. This could involve iterative cycles of feedback and adjustment in which M-LLMs are fine-tuned based on direct input from health care professionals regarding the appropriateness and ethical alignment of their outputs. Incorporating mechanisms for DPO in which models adjust their decision-making strategies in real time based on new information or feedback could further enhance alignment with human values. Moreover, simulating diverse clinical and ethical scenarios during training phases can prepare M-LLMs to handle real-world complexities.

Future Outlook

Overview

In the evolving landscape of medical M-LLMs, anticipating future directions is crucial for advancing their application in health care. In this section, we outline prospective advancements and necessary adaptations that could enhance the functionality, efficacy, and ethical integration of M-LLMs in health care. Specifically, we explore the evolution in generating multimodal outputs, the critical need for establishing performance benchmarks, the shift in explainability paradigms toward

XSL•FO
RenderX

comprehensive explainability, the role of M-LLMs in enhancing interoperability within hospital systems, the formulation of robust regulatory frameworks, and the essential role of multidisciplinary collaboration (Figure 6). We envision that these areas collectively represent key future perspectives where M-LLMs are expected to transform both medical applications and patient care.

Figure 6. Future directions of multimodal large language models in health care.

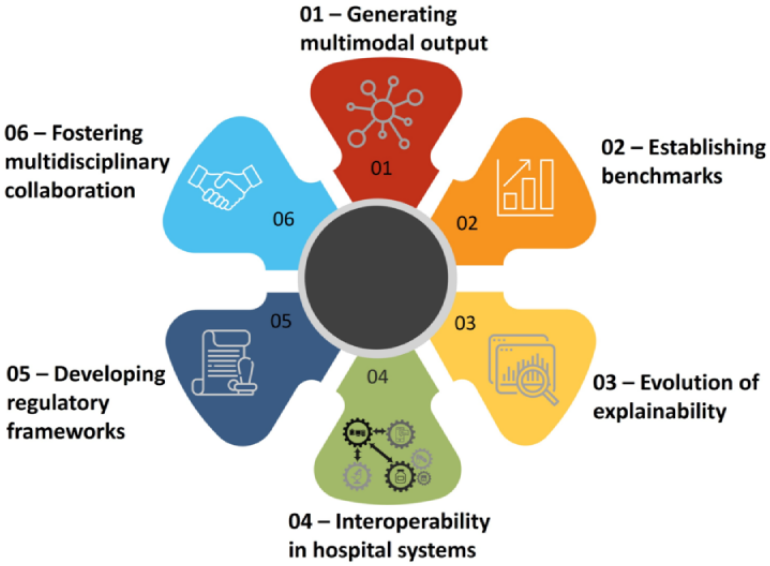

Generating Multimodal Outputs

While medical M-LLMs are rapidly evolving in processing multimodal inputs, the development of multimodal outputs is still trailing behind. The importance of multimodal outputs in medical contexts is significant. For example, when asking ChatGPT to explain complex medical concepts, such as interpreting radiological images or outlining surgical procedures, effective explanations should ideally blend textual descriptions with graphical representations, mathematical equations, audio narratives, or animations for enhanced comprehension. This highlights the need for medical M-LLMs capable of producing such varied outputs. A critical step toward this goal is the creation of a shared intermediate output by the model, which raises the following question: what form should this intermediate output take? A practical method is using text as the intermediate output, serving as a basis for generating additional modalities. For example, the causal masked multimodal (CMM) model [192] produces HTML markup that can be transformed into rich web pages with text, formatting, links, and images. Alternatively, using multimodal tokens where each token is tagged to represent different modalities such as text or image offers another route. Image tokens could feed into an image generation model such as a diffusion model to generate visual content, whereas text tokens are processed by a language model. This dual-token approach paves the way for more sophisticated and contextually appropriate multimodal outputs. Further exploration and development in this field could lead to models that seamlessly integrate a variety of output formats, revolutionizing the way in which medical information is conveyed and understood.

Establishing Benchmarks

Benchmarks are crucial in assessing the performance, accuracy, and effectiveness of generative AI, especially in the context of medical M-LLMs. The expansive scope and complex nature of health care and medicine necessitate continuous advancements in robust evaluation methods and frameworks. This is essential to ensure that medical M-LLMs are effectively aligned with the unique requirements of these domains. These benchmarks enable model comparisons, highlighting efficiencies and creative capabilities in specific tasks and data modalities both individually and collectively. They also play a critical role in detecting biases and limitations. Furthermore, they play a crucial role in establishing industry standards for medical M-LLMs, ensuring their ethical and safe use in sensitive medical contexts. Recent initiatives in M-LLM benchmarks, such as AesBench [193], Mementos [194], MME [195], MM-BigBench [196], MLLM-Bench [197], and VLM-Eval [198], offer a foundational framework that could be adapted to medical M-LLMs. However, there is an urgent need for more comprehensive evaluation methods and frameworks as well as rigorous rubrics for human evaluation of M-LLM performance in real-world clinical workflows and scenarios.

Evolution of Explainability: From Snapshot to Temporal Explainability

Snapshot explainability refers to the ability of M-LLMs to provide explanations for decisions or predictions at a single, specific point in time. In contrast, temporal analysis offers a more comprehensive understanding by tracking and interpreting changes over time. Most current interpretability research on M-LLMs neglects training dynamics, focusing mainly on post hoc explanations of fully trained models [167]. This lack of developmental investigation into the training process can lead to biased explanations. Moreover, examining interpretability based on a single data modality fails to reflect interactions between modalities. Therefore, transitioning from static snapshot explainability to dynamic temporal analysis is essential for medical M-LLMs. This approach is particularly beneficial for

XSL•FO
RenderX

using multimodal data in monitoring patient progress, understanding disease trajectories, and predicting outcomes. By leveraging temporal explainability, M-LLMs can better contextualize data, uncovering patterns and trends that might be overlooked in static analysis. This shift not only enhances the accuracy of diagnoses and treatment plans but also improves the personalization of patient care by taking advantage of rich multimodal data.

Interoperability in Hospital Systems

An M-LLM could act as a central hub in hospitals, integrating various unimodal AI systems such as radiology, insurance, and EHRs. Currently, each department uses different AI tools from various companies, and most of these systems do not intercommunicate, resulting in access being limited to only department-specific systems. For instance, radiologists use radiological AI, whereas cardiologists might not have access to this, and likewise for other specialties. The introduction of M-LLMs can change this landscape significantly. M-LLMs understand the language and format of all these disparate software applications, allowing for seamless interaction. This means that health care practitioners regardless of specialty could easily work with any AI tool in the hospital, breaking down the silos that currently exist. This potential is vital as it enables comprehensive, integrated care, which individual organizations cannot achieve alone due to proprietary restrictions on data.

Developing Regulatory Frameworks

The development of a regulatory framework for medical M-LLMs is essential to ensure their safe, effective, and ethical use. Regulatory bodies need to establish standards and guidelines that define acceptable accuracy for various M-LLM applications, ensuring that these tools are reliable and trustworthy in clinical settings. A critical aspect of this framework also includes algorithmic transparency; therefore, regulatory guidelines must clearly stipulate requirements for explainability. Furthermore, the protection of patient data privacy is essential given that M-LLMs process sensitive health information. Therefore, regulatory frameworks must enforce strict data protection standards and formulate strategies for ethically collecting and processing multimodal data sets. Moreover, regardless of whether regulations are sufficiently developed or comprehensive in any given jurisdiction, medical and research institutions have an obligation to upgrade the knowledge and diversity of their ethics approval boards.

Fostering Multidisciplinary Collaboration and Stakeholder Engagement

AI, and specifically M-LLMs, is so new and complex in the health care domain that the expertise and insights needed extend far beyond the capabilities of any one health care or academic medical organization. Thus, it is imperative for those implementing M-LLM solutions to draw upon the know-how of 4 major external stakeholders. First, because many AI projects are expected to pose ethical concerns, the relevant applicable regulatory bodies and local health authorities should be engaged on a regular basis to ensure compliance with regulations. Indeed, guidelines and laws are rapidly changing; at the time of writing, the European Union has endorsed a world-first AI Act [199]. Second, much of the M-LLM innovation is expected to stem from academic and research contexts, where scientists continually push the boundaries of evidence-based, validated AI projects commonly published and made available for public benefit. Collaborating and partnering with such institutions ensures that the latest approaches and technologies can be incorporated into a health care project. Third, the industry is often a forgotten collaborator due to perceived entry barriers (eg, intellectual property ownership, exclusivity, and so forth). However, large commercial companies have access to far wider resources and technical expertise, particularly in engineering development, than medical institutions and, when negotiated with a win-win perspective, can significantly accelerate AI project deployment in the health care context. The same may apply to vendors who are infrastructure and deployment experts and who may be able to contribute beyond the limited scope of a purchase agreement. Moreover, when applicable, industry partners may offer greater commercialization pathways for projects. Finally, the fourth external stakeholder is the patient advocacy organization. Such groups should be engaged early and continuously and can help ensure that patients' critical perspectives are communicated and included within the requirements of an M-LLM project. This is especially the case in projects that directly impact the patients' needs and preferences, for instance, an M-LLM that interacts by providing clinical insights and recommendations to the physician during a patient consultation. Such advocacy groups can also be an effective way for health care institutions to more naturally engage in awareness and trust building with their communities. Naturally, with external stakeholders, appropriate collaboration and data agreements should be sought to protect the health care institutions' interests as well as those of their patients. In addition, regardless of whether projects require internal or external collaboration, best practices should be used to ensure that roles, responsibilities, and decision-making structures are clarified upfront.

Conclusions

In this paper, we explored the foundational principles, applications, challenges, and future perspectives of M-LLMs in health care practice. While this work suggests a promising direction for the application of M-LLMs in medicine, it also highlights the need for further evaluation and benchmarking of their capabilities and limitations in real-world medical settings. In addition, despite the momentum toward models capable of processing multimodal inputs, the progression toward sophisticated multimodal outputs remains comparatively slow. Furthermore, it is crucial to acknowledge that the emergence of M-LLMs does not render traditional LLMs obsolete. Instead, M-LLMs serve as an extension, building upon the foundational strengths and capabilities of LLMs to enhance health care delivery. This association underscores that the efficiency of M-LLMs is inherently tied to the robustness of the underlying LLMs. As we advance toward more general AI systems, M-LLMs offer a promising path to a comprehensive form of AI in health care practice. The journey has its challenges, but the potential rewards could significantly redefine our interaction with technology in the medical field.

XSL•FO
RenderX

Conflicts of Interest

AAA is an associate editor for *JMIR Nursing*. All other authors declare no coflicts of interest.

References

1. Tamkin A, Brundage M, Clark J, Ganguli D. Understanding the capabilities, limitations, and societal impact of large language models. arXiv Preprint posted online on February 4, 2021. [doi: 10.48550/arXiv.2102.02503 Focus to learn more]
2. Chen R. [Prospects for the application of healthcare big data combined with large language models]. Sichuan Da Xue Xue Bao Yi Xue Ban 2023 Sep;54(5):855-856 [FREE Full text] [doi: 10.12182/20230960301] [Medline: 37866938]
3. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141 [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
4. Botelho F, Tshimula JM, Poenaru D. Leveraging ChatGPT to democratize and decolonize global surgery: large language models for small healthcare budgets. World J Surg 2023 Nov;47(11):2626-2627. [doi: 10.1007/s00268-023-07167-2] [Medline: 37689598]
5. Praveen SV, Deepika R. Exploring the perspective of infection clinicians on the integration of large language models (LLMs) in clinical practice: a deep learning study in healthcare. J Infect 2023 Oct;87(4):e68-e69. [doi: 10.1016/j.jinf.2023.07.011] [Medline: 37487951]
6. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
7. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. Healthcare (Basel) 2023 Oct 20;11(20):2776 [FREE Full text] [doi: 10.3390/healthcare11202776] [Medline: 37893850]
8. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med 2023 May;116(5):181-182 [FREE Full text] [doi: 10.1177/01410768231173123] [Medline: 37199678]
9. Holzinger A. Explainable AI and multi-modal causability in medicine. I Com (Berl) 2021 Jan 26;19(3):171-179 [FREE Full text] [doi: 10.1515/icom-2020-0024] [Medline: 37014363]
10. Liang J, Li ZW, Yue CT, Hu Z, Cheng H, Liu ZX, et al. Multi-modal optimization to identify personalized biomarkers for disease prediction of individual patients with cancer. Brief Bioinform 2022 Sep 20;23(5):bbac254. [doi: 10.1093/bib/bbac254] [Medline: 35858208]
11. Zheng S, Zhu Z, Liu Z, Guo Z, Liu Y, Yang Y, et al. Multi-modal graph learning for disease prediction. IEEE Trans Med Imaging 2022 Sep;41(9):2207-2216. [doi: 10.1109/tmi.2022.3159264]
12. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA.
14. AlSaad R, Malluhi Q, Janahi I, Boughorbel S. Interpreting patient-specific risk prediction using contextual decomposition of BiLSTMs: application to children with asthma. BMC Med Inform Decis Mak 2019 Nov 08;19(1):214 [FREE Full text] [doi: 10.1186/s12911-019-0951-4] [Medline: 31703676]
15. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018.
16. AlSaad R, Malluhi Q, Abd-Alrazaq A, Boughorbel S. Temporal self-attention for risk prediction from electronic health records using non-stationary kernel approximation. Artif Intell Med 2024 Mar;149:102802 [FREE Full text] [doi: 10.1016/j.artmed.2024.102802] [Medline: 38462292]
17. Li Y, Mamouei M, Salimi-Khorshidi G, Rao S, Hassaine A, Canoy D, et al. Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. IEEE J Biomed Health Inform 2023 Feb;27(2):1106-1117 [FREE Full text] [doi: 10.1109/JBHI.2022.3224727] [Medline: 36427286]
18. Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep 2020 Apr 28;10(1):7155 [FREE Full text] [doi: 10.1038/s41598-020-62922-y] [Medline: 32346050]
19. Ho QT, Nguyen TT, Khanh Le NQ, Ou YY. FAD-BERT: improved prediction of FAD binding sites using pre-training of deep bidirectional transformers. Comput Biol Med 2021 Apr;131:104258. [doi: 10.1016/j.compbiomed.2021.104258] [Medline: 33601085]
20. Mulyar A, Uzuner O, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. J Am Med Inform Assoc 2021 Sep 18;28(10):2108-2115 [FREE Full text] [doi: 10.1093/jamia/ocab126] [Medline: 34333635]
21. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/index/chatgpt/ [accessed 2024-09-12]

XSL•FO
RenderX

22. Zhang Q, Liang Y. Comments on "ChatGPT and its role in the decision-making for the diagnosis and treatment of lumbar spinal stenosis: a comparative analysis and narrative review". Global Spine J 2024 May;14(4):1452 [FREE Full text] [doi: 10.1177/21925682231222268] [Medline: 38090761]
23. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol 2024 May 08;34(3):e13207 [FREE Full text] [doi: 10.1111/bpa.13207] [Medline: 37553205]
24. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 2024 Jan 23;66(1):73-79. [doi: 10.1007/s00234-023-03252-4] [Medline: 37994939]
25. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023 Feb;15(2):e35237 [FREE Full text] [doi: 10.7759/cureus.35237] [Medline: 36968864]
26. Mese I, Taslicay CA, Sivrioglu AK. Improving radiology workflow using ChatGPT and artificial intelligence. Clin Imaging 2023 Nov;103:109993. [doi: 10.1016/j.clinimag.2023.109993] [Medline: 37812965]
27. Mallio CA, Sertorio AC, Bernetti C, Beomonte Zobel B. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. Radiol Med 2023 Jul 29;128(7):808-812. [doi: 10.1007/s11547-023-01651-4] [Medline: 37248403]
28. Perera Molligoda Arachchige AS. Empowering radiology: the transformative role of ChatGPT. Clin Radiol 2023 Nov;78(11):851-855. [doi: 10.1016/j.crad.2023.08.006] [Medline: 37657969]
29. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv Preprint posted online on February 27, 2023 [FREE Full text]
30. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. arXiv Preprint posted online on October 20, 2022 [FREE Full text]
31. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. LMSYS. 2023 Mar 30. URL: https://lmsys.org/blog/2023-03-30-vicuna/ [accessed 2024-09-12]
32. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Alpaca: a strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. 2023. URL: https://crfm.stanford.edu/2023/03/13/alpaca.html [accessed 2024-09-12]
33. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CV, Belfort RJ. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. Rev Assoc Med Bras (1992) 2023;69(10):e20230848 [FREE Full text] [doi: 10.1590/1806-9282.20230848] [Medline: 37792871]
34. Lai UH, Wu KS, Hsu TY, Kan JK. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) 2023 Sep 19;10:1240915 [FREE Full text] [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]
35. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for AI-assisted medical education and decision making in radiation oncology. Front Oncol 2023 Sep 14;13:1265024 [FREE Full text] [doi: 10.3389/fonc.2023.1265024] [Medline: 37790756]
36. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach 2023 Oct 15;46(3):1-7. [doi: 10.1080/0142159x.2023.2249588]
37. Kleebayoon A, Wiwanitkit V. ChatGPT-4, medical education, and clinical exposure challenges. Indian J Orthop 2023 Nov 21;57(11):1912. [doi: 10.1007/s43465-023-00997-1] [Medline: 37881293]
38. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023 Nov 02;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]
39. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 2023 Jul 28;15(1):29 [FREE Full text] [doi: 10.1038/s41368-023-00239-y] [Medline: 37507396]
40. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: a scoping review. NPJ Digit Med 2022 Nov 07;5(1):171 [FREE Full text] [doi: 10.1038/s41746-022-00712-8] [Medline: 36344814]
41. Huang W, Tu S, Xu L. PFB-Diff: progressive feature blending diffusion for text-driven image editing. arXiv Preprint posted online on June 28, 2023 [FREE Full text]
42. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision-language understanding with advanced large language models. arXiv Preprint posted online on April 20, 2023 [FREE Full text]
43. Su Y, Lan T, Liu Y, Liu F, Yogatama D, Wang Y, et al. Language models can see: plugging visual controls in text generation. arXiv Preprint posted online on May 5, 2022 [FREE Full text]
44. Koh JY, Fried D, Salakhutdinov R. Generating images with multimodal language models. arXiv Preprint posted online on May 26, 2023 [FREE Full text]
45. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. arXiv Preprint posted online on April 29, 2022 [FREE Full text]

XSL•FO
RenderX

46. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv Preprint posted online on April 17, 2023. [doi: 10.48550/arXiv.2304.08485]
47. Li K, He Y, Wang Y, Li Y, Wang W, Luo P, et al. VideoChat: chat-centric video understanding. arXiv Preprint posted online on May 10, 2023. [doi: 10.48550/arXiv.2305.06355]
48. Zhang H, Li X, Bing L. Video-LLaMA: an instruction-tuned audio-visual language model for video understanding. arXiv Preprint posted online on June 5, 2023. [doi: 10.48550/arXiv.2306.02858]
49. Zhang D, Li S, Zhang X, Zhang J, Wang P, Zhou Y, et al. SpeechGPT: empowering large language models with intrinsic cross-modal conversational abilities. arXiv Preprint posted online on May 18, 2023. [doi: 10.18653/v1/2023.findings-emnlp.1055]
50. Su Y, Lan T, Li H, Xu J, Wang Y, Cai D. PandaGPT: one model to instruction-follow them all. arXiv Preprint posted online on May 25, 2023. [doi: 10.48550/arXiv.2305.16355]
51. Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, et al. ImageBind: one embedding space to bind them all. arXiv Preprint posted online on May 9, 2023. [doi: 10.48550/arXiv.2305.05665]
52. Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. Nat Commun 2022 Jun 02;13(1):3094 [FREE Full text] [doi: 10.1038/s41467-022-30761-2] [Medline: 35655064]
53. Buttazzo G. Rise of artificial general intelligence: risks and opportunities. Front Artif Intell 2023 Aug 25;6:1226990 [FREE Full text] [doi: 10.3389/frai.2023.1226990] [Medline: 37693010]
54. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv Preprint posted online on October 22, 2020. [doi: 10.48550/arXiv.2010.11929]
55. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv Preprint posted online on September 4, 2014. [doi: 10.48550/arXiv.1409.1556]
56. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: CVPR 2016; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
57. Albaqami H, Hassan GM, Datta A. Automatic detection of abnormal EEG signals using WaveNet and LSTM. Sensors (Basel) 2023 Jun 27;23(13):5960 [FREE Full text] [doi: 10.3390/s23135960] [Medline: 37447810]
58. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep speech: scaling up end-to-end speech recognition. arXiv Preprint posted online on December 17, 2014. [doi: 10.48550/arXiv.1412.5567]
59. Zhu S, Zheng J, Ma Q. MR-Transformer: multiresolution transformer for multivariate time series prediction. IEEE Trans Neural Netw Learn Syst 2023 Nov 06;PP (forthcoming). [doi: 10.1109/TNNLS.2023.3327416] [Medline: 37930914]
60. Baidya R, Jeong H. Anomaly detection in time series data using reversible instance normalized anomaly transformer. Sensors (Basel) 2023 Nov 19;23(22):9272 [FREE Full text] [doi: 10.3390/s23229272] [Medline: 38005658]
61. Bai N, Wang X, Han R, Wang Q, Liu Z. PAFormer: anomaly detection of time series with parallel-attention transformer. IEEE Trans Neural Netw Learn Syst 2023 Dec 11;PP (forthcoming). [doi: 10.1109/TNNLS.2023.3337876] [Medline: 38079369]
62. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 2018 Nov;36(10):983-987. [doi: 10.1038/nbt.4235] [Medline: 30247488]
63. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 2016 Jul;26(7):990-999 [FREE Full text] [doi: 10.1101/gr.200535.115] [Medline: 27197224]
64. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol 2017 Apr 11;18(1):67 [FREE Full text] [doi: 10.1186/s13059-017-1189-z] [Medline: 28395661]
65. Wang X, Jiang Y, Bach N, Wang T, Huang Z, Huang F, et al. Automated concatenation of embeddings for structured prediction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021 Presented at: ACL/IJCNLP 2021; August 1-6, 2021; Virtual Event. [doi: 10.18653/v1/2021.acl-long.206]
66. Xi C, Lu G, Yan J. Multimodal sentiment analysis based on multi-head attention mechanism. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing. 2020 Presented at: ICMLSC '20; January 17-19, 2020; Haiphong City, Viet Nam. [doi: 10.1145/3380688.3380693]
67. Kiela D, Bhooshan S, Firooz H, Perez E, Testuggine D. Supervised multimodal bitransformers for classifying images and text. arXiv Preprint posted online on September 6, 2019. [doi: 10.48550/arXiv.1909.02950]
68. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. arXiv Preprint posted online on July 23, 2017. [doi: 10.48550/arXiv.1707.07250]
69. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. arXiv Preprint posted online on November 27, 2023. [doi: 10.48550/arXiv.2311.16079]
70. Akrout M, Cirone KD, Vender R. Evaluation of vision LLMs GTP-4V and LLaVA for the recognition of features characteristic of melanoma. J Cutan Med Surg 2024;28(1):98-99. [doi: 10.1177/12034754231220934] [Medline: 38174854]
71. Rasheed H, Maaz M, Mullappilly SS, Shaker A, Khan S, Cholakkal H, et al. GLaMM: pixel grounding large multimodal model. arXiv Preprint posted online on November 6, 2023.

XSL•FO
RenderX

72. Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond. arXiv Preprint posted online on August 24, 2023. [doi: 10.48550/arXiv.2308.12966]
73. Chen B, Xu Z, Kirmani S, Ichter B, Driess D, Florence P, et al. SpatialVLM: endowing vision-language models with spatial reasoning capabilities. arXiv Preprint posted online on January 22, 2024. [doi: 10.48550/arXiv.2401.12168]
74. Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, et al. InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv Preprint posted online on December 21, 2023. [doi: 10.48550/arXiv.2312.14238]
75. Yuan Y, Li W, Liu J, Tang D, Luo X, Qin C, et al. Osprey: pixel understanding with visual instruction tuning. arXiv Preprint posted online on December 15, 2023. [doi: 10.48550/arXiv.2312.10032]
76. Wei H, Kong L, Chen J, Zhao L, Ge Z, Yang J, et al. Vary: scaling up the vision vocabulary for large vision-language models. arXiv Preprint posted online on December 11, 2023. [doi: 10.48550/arXiv.2312.06109]
77. Chen L, Li J, Dong X, Zhang P, He C, Wang J, et al. ShareGPT4V: improving large multi-modal models with better captions. arXiv Preprint posted online on November 21, 2023. [doi: 10.48550/arXiv.2311.12793]
78. Li B, Zhang P, Yang J, Zhang Y, Pu F, Liu Z. OtterHD: a high-resolution multi-modality model. arXiv Preprint posted online on November 7, 2023. [doi: 10.48550/arXiv.2311.04219]
79. Chen G, Shen L, Shao R, Deng X, Nie L. LION : empowering multimodal large language model with dual-level visual knowledge. arXiv Preprint posted online on November 20, 2023. [doi: 10.48550/arXiv.2311.11860]
80. Lin Z, Liu C, Zhang R, Gao P, Qiu L, Xiao H, et al. SPHINX: the joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv Preprint posted online on November 13, 2023. [doi: 10.48550/arXiv.2311.07575]
81. Hu W, Xu Y, Li Y, Li W, Chen Z, Tu Z. BLIVA: a simple multimodal LLM for better handling of text-rich visual questions. Proc AAAI Conf Artif Intell 2024 Mar 24;38(3):2256-2264. [doi: 10.1609/aaai.v38i3.27999]
82. Zhao B, Wu B, He M, Huang T. SVIT: scaling up visual instruction tuning. arXiv Preprint posted online on July 9, 2023. [doi: 10.48550/arXiv.2307.04087]
83. Li J, Chen D, Hong Y, Chen Z, Chen P, Shen Y, et al. CoVLM: composing visual entities and relationships in large language models via communicative decoding. arXiv Preprint posted online on November 6, 2023. [doi: 10.48550/arXiv.2311.03354]
84. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv Preprint posted online on June 1, 2023. [doi: 10.48550/arXiv.2306.00890]
85. Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, et al. PMC-VQA: visual instruction tuning for medical visual question answering. arXiv Preprint posted online on May 17, 2023. [doi: 10.48550/arXiv.2305.10415]
86. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-Flamingo: a multimodal medical few-shot learner. arXiv Preprint posted online on July 27, 2023. [doi: 10.48550/arXiv.2307.15189]
87. He J, Liu G, Li P, Zhao Z, Zhong S. PeFoMed: parameter efficient fine-tuning on multimodal large language models for medical visual question answering. arXiv Preprint posted online on April 16, 2024. [doi: 10.48550/arXiv.2401.02797]
88. Han T, Adams LC, Nebelung S, Kather JN, Bressem KK, Truhn D. Multimodal large language models are generalist medical image interpreters. medRxiv Preprint posted online on December 22, 2023. [doi: 10.1101/2023.12.21.23300146]
89. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
90. Jin M, Wang S, Ma L, Chu Z, Zhang JY, Shi X, et al. Time-LLM: time series forecasting by reprogramming large language models. arXiv Preprint posted online on October 3, 2023. [doi: 10.48550/arXiv.2310.01728]
91. Chang C, Wang WY, Peng WC, Chen TF. LLM4TS: aligning pre-trained LLMs as data-efficient time-series forecasters. arXiv Preprint posted online on August 16, 2023. [doi: 10.48550/arXiv.2308.08469]
92. Cao D, Jia F, Arik SO, Pfister T, Zheng Y, Ye W, et al. TEMPO: prompt-based generative pre-trained transformer for time series forecasting. arXiv Preprint posted online on October 8, 2023. [doi: 10.48550/arXiv.2310.04948]
93. Xue H, Salim FD. PromptCast: a new prompt-based learning paradigm for time series forecasting. IEEE Trans Knowl Data Eng 2023 Dec 13:1-14. [doi: 10.1109/tkde.2023.3342137]
94. Jin M, Wen Q, Liang Y, Zhang C, Xue S, Wang X, et al. Large models for time series and spatio-temporal data: a survey and outlook. arXiv Preprint posted online on October 16, 2023. [doi: 10.48550/arXiv.2310.10196]
95. Gruver N, Finzi M, Qiu S, Wilson AG. Large language models are zero-shot time series forecasters. arXiv Preprint posted online on October 11, 2023. [doi: 10.48550/arXiv.2310.07820]
96. Liu X, McDuff D, Kovacs G, Galatzer-Levy I, Sunshine J, Zhan J, et al. Large language models are few-shot health learners. arXiv Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.15525]
97. Rubenstein PK, Asawaroengchai C, Nguyen DD, Bapna A, Borsos Z, de Chaumont Quitry F, et al. AudioPaLM: a large language model that can speak and listen. arXiv Preprint posted online on June 22, 2023. [doi: 10.48550/arXiv.2306.12925]
98. Borsos Z, Marinier R, Vincent D, Kharitonov E, Pietquin O, Sharifi M, et al. AudioLM: a language modeling approach to audio generation. IEEE/ACM Trans Audio Speech Lang Process 2023 Jun 21;31:2523-2533. [doi: 10.1109/taslp.2023.3288409]
99. Deshmukh S, Elizalde B, Singh R, Wang H. Pengi: an audio language model for audio tasks. arXiv Preprint posted online on May 19, 2023. [doi: 10.48550/arXiv.2305.11834]

XSL•FO
RenderX

100. Huang R, Li M, Yang D, Shi J, Chang X, Ye Z, et al. AudioGPT: understanding and generating speech, music, sound, and talking head. Proc AAAI Conf Artif Intell 2024 Mar 24;38(21):23802-23804. [doi: 10.1609/aaai.v38i21.30570]
101. Wang T, Zhou L, Zhang Z, Wu Y, Liu S, Gaur Y, et al. VioLA: conditional language models for speech recognition, synthesis, and translation. IEEE/ACM Trans Audio Speech Lang Process 2024 Jul 29;32:3709-3716. [doi: 10.1109/taslp.2024.3434425]
102. Tang C, Yu W, Sun G, Chen X, Tan T, Li W, et al. SALMONN: towards generic hearing abilities for large language models. arXiv Preprint posted online on October 20, 2023. [doi: 10.48550/arXiv.2310.13289]
103. Latif S, Shoukat M, Shamshad F, Usama M, Ren Y, Cuayáhuitl H, et al. Sparks of large audio models: a survey and outlook. arXiv Preprint posted online on August 24, 2023. [doi: 10.48550/arXiv.2308.12792]
104. Pieri S, Mullappilly SS, Khan FS, Anwer RM, Khan S, Baldwin T, et al. BiMediX: bilingual medical mixture of experts LLM. arXiv Preprint posted online on February 20, 2024. [doi: 10.48550/arXiv.2402.13253]
105. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
106. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
107. Maaz M, Rasheed H, Khan S, Khan FS. Video-ChatGPT: towards detailed video understanding via large vision and language models. arXiv Preprint posted online on June 8, 2023. [doi: 10.48550/arXiv.2306.05424]
108. Li Y, Wang C, Jia J. LLaMA-VID: an image is worth 2 tokens in large language models. arXiv Preprint posted online on November 28, 2023. [doi: 10.48550/arXiv.2311.17043]
109. Jiang B, Chen X, Liu W, Yu J, Yu G, Chen T. MotionGPT: human motion as a foreign language. arXiv Preprint posted online on June 26, 2023. [doi: 10.48550/arXiv.2306.14795]
110. Li L, Gan Z, Lin K, Lin CC, Liu Z, Liu C, et al. LAVENDER: unifying video-language understanding as masked language modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 Presented at: CVPR 2023; June 17-24, 2023; Vancouver, BC. [doi: 10.1109/cvpr52729.2023.02214]
111. Song E, Chai W, Wang G, Zhang Y, Zhou H, Wu F, et al. MovieChat: from dense token to sparse memory for long video understanding. arXiv Preprint posted online on July 31, 2023. [doi: 10.48550/arXiv.2307.16449]
112. Yang A, Nagrani A, Seo PH, Miech A, Pont-Tuset J, Laptev I, et al. Vid2Seq: large-scale pretraining of a visual language model for dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 Presented at: CVPR 2023; June 17-24, 2023; Vancouver, BC. [doi: 10.1109/cvpr52729.2023.01032]
113. Chen G, Zheng YD, Wang J, Xu J, Huang Y, Pan J, et al. VideoLLM: modeling video sequence with large language models. arXiv Preprint posted online on May 22, 2023. [doi: 10.48550/arXiv.2305.13292]
114. Huang B, Wang X, Chen H, Song Z, Zhu W. VTimeLLM: empower LLM to grasp video moments. arXiv Preprint posted online on November 30, 2023. [doi: 10.48550/arXiv.2311.18445]
115. Liu J, Yang M, Yu Y, Xu H, Li K, Zhou X. arge language models in bioinformatics: applications and perspectives. arXiv Preprint posted online on January 8, 2024 [FREE Full text] [doi: 10.48550/arXiv.2401.04155]
116. Benegas G, Batra SS, Song YS. DNA language models are powerful predictors of genome-wide variant effects. Proc Natl Acad Sci U S A 2023 Oct 31;120(44):e2311219120 [FREE Full text] [doi: 10.1073/pnas.2311219120] [Medline: 37883436]
117. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 2021 Aug 09;37(15):2112-2120 [FREE Full text] [doi: 10.1093/bioinformatics/btab083] [Medline: 33538820]
118. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. arXiv Preprint posted online on June 26, 2023. [doi: 10.48550/arXiv.2306.15006]
119. Zhang D, Zhang W, Zhao Y, Zhang J, He B, Qin C, et al. DNAGPT: a generalized pre-trained tool for versatile DNA sequence analysis tasks. BioRxiv Preprint posted online on January 04, 2024. [doi: 10.1101/2023.07.11.548628]
120. Luo H, Shan W, Chen C, Ding P, Luo L. Improving language model of human genome for DNA-protein binding prediction based on task-specific pre-training. Interdiscip Sci 2023 Mar;15(1):32-43. [doi: 10.1007/s12539-022-00537-9] [Medline: 36136096]
121. An W, Guo Y, Bian Y, Ma H, Yang J, Li C, et al. MoDNA: motif-oriented pre-training for DNA language model. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2022 Presented at: BCB '22; August 7-10, 2022; Northbrook, Illinois. [doi: 10.1145/3535508.3545512]
122. Chen K, Zhou Y, Ding M, Wang Y, Ren Z, Yang Y. Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. BioRxiv Preprint posted online on February 3, 2023. [doi: 10.1101/2023.01.31.526427]
123. Feng H, Wang S, Wang Y, Ni X, Yang Z, Hu X, et al. LncCat: an ORF attention model to identify LncRNA based on ensemble learning strategy and fused sequence information. Comput Struct Biotechnol J 2023 Feb 08;21:1433-1447 [FREE Full text] [doi: 10.1016/j.csbj.2023.02.012] [Medline: 36824229]
124. Yamada K, Hamada M. Prediction of RNA-protein interactions using a nucleotide language model. Bioinform Adv 2022 Apr 07;2(1):vbac023 [FREE Full text] [doi: 10.1093/bioadv/vbac023] [Medline: 36699410]

XSL•FO
RenderX

125. Zhang L, Qin X, Liu M, Liu G, Ren Y. BERT-m7G: a transformer architecture based on BERT and stacking ensemble to identify RNA N7-methylguanosine sites from sequence information. Comput Math Methods Med 2021 Aug 25;2021:7764764 [FREE Full text] [doi: 10.1155/2021/7764764] [Medline: 34484416]
126. Babjac AN, Lu Z, Emrich SJ. CodonBERT: using BERT for sentiment analysis to better predict genes with low expression. In: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2023 Presented at: BCB '23; September 3-6, 2023; Houston, TX. [doi: 10.1145/3584371.3613013]
127. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022 Oct;44(10):7112-7127. [doi: 10.1109/TPAMI.2021.3095381] [Medline: 34232869]
128. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023 Aug;41(8):1099-1106 [FREE Full text] [doi: 10.1038/s41587-022-01618-2] [Medline: 36702895]
129. Xu M, Yuan X, Miret S, Tang J. ProtST: multi-modality learning of protein sequences and biomedical texts. arXiv Preprint posted online on January 28, 2023. [doi: 10.48550/arXiv.2301.12040]
130. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022 Apr 12;38(8):2102-2110 [FREE Full text] [doi: 10.1093/bioinformatics/btac020] [Medline: 35020807]
131. Wang Z, Zhang Q, Yu H, Hu S, Jin X, Gong Z, et al. Multi-level protein structure pre-training with prompt learning. In: Proceedings of the Eleventh International Conference on Learning Representations. 2023 Presented at: ICLR 2023; May 1-5, 2023; Kigali, Rwanda.
132. Wang S, You R, Liu Y, Xiong Y, Zhu S. NetGO 3.0: protein language model improves large-scale functional annotations. Genomics Proteomics Bioinformatics 2023 Apr;21(2):349-358 [FREE Full text] [doi: 10.1016/j.gpb.2023.04.001] [Medline: 37075830]
133. Flamholz ZN, Biller SJ, Kelly L. Large language models improve annotation of prokaryotic viral proteins. Nat Microbiol 2024 Feb;9(2):537-549. [doi: 10.1038/s41564-023-01584-8] [Medline: 38287147]
134. Driess D, Xia F, Sajjadi MS, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. arXiv Preprint posted online on March 6, 2023. [doi: 10.48550/arXiv.2303.03378]
135. Moon S, Madotto A, Lin Z, Nagarajan T, Smith M, Jain S, et al. AnyMAL: an efficient and scalable any-modality augmented language model. arXiv Preprint posted online on September 27, 2023. [doi: 10.48550/arXiv.2309.16058]
136. Wu S, Fei H, Qu L, Ji W, Chua TS. NExT-GPT: any-to-any multimodal LLM. arXiv Preprint posted online on September 11, 2023. [doi: 10.48550/arXiv.2309.05519]
137. Lyu C, Wu M, Wang L, Huang X, Liu B, Du Z, et al. Macaw-LLM: multi-modal language modeling with image, audio, video, and text integration. arXiv Preprint posted online on June 15, 2023. [doi: 10.48550/arXiv.2306.09093]
138. Han J, Gong K, Zhang Y, Wang J, Zhang K, Lin D, et al. OneLLM: one framework to align all modalities with language. arXiv Preprint posted online on December 6, 2023. [doi: 10.48550/arXiv.2312.03700]
139. Ye Q, Xu H, Ye J, Yan M, Hu A, Liu H, et al. mPLUG-Owl2: revolutionizing multi-modal large language model with modality collaboration. arXiv Preprint posted online on November 7, 2023. [doi: 10.48550/arXiv.2311.04257]
140. Gemini Team Google. Gemini: a family of highly capable multimodal models. arXiv Preprint posted online on December 19, 2023. [doi: 10.48550/arXiv.2312.11805]
141. Lu MY, Chen B, Williamson DF, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nat Med 2024 Mar;30(3):863-874. [doi: 10.1038/s41591-024-02856-4] [Medline: 38504017]
142. CONCH: a vision-language foundation model for computational pathology. GitHub. URL: https://github.com/mahmoodlab/CONCH [accessed 2024-08-02]
143. Li Y, Quan R, Zhu L, Yang Y. Efficient multimodal fusion via interactive prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 Presented at: CVPR 2023; June 17-24, 2023; Vancouver, BC. [doi: 10.1109/cvpr52729.2023.00256]
144. Ye J, Hai J, Song J, Wang Z. Multimodal data hybrid fusion and natural language processing for clinical prediction models. medRxiv Preprint posted online on August 25, 2023. [doi: 10.1101/2023.08.24.23294597]
145. Quan Z, Sun T, Su M, Wei J. Multimodal sentiment analysis based on cross-modal attention and gated cyclic hierarchical fusion networks. Comput Intell Neurosci 2022 Aug 9;2022:4767437 [FREE Full text] [doi: 10.1155/2022/4767437] [Medline: 35983132]
146. Yuan H, Xu H. Deep multi-modal fusion network with gated unit for breast cancer survival prediction. Comput Methods Biomech Biomed Engin 2024 May;27(7):883-896. [doi: 10.1080/10255842.2023.2211188] [Medline: 37166185]
147. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J, et al. A survey of large language models in medicine: progress, application, and challenge. arXiv Preprint posted online on November 9, 2023.
148. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 03;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
149. Wu W, Wang Y, Liu Q, Wang G, Zhang J. Wavelet-improved score-based generative model for medical imaging. IEEE Trans Med Imaging 2024 Mar;43(3):966-979. [doi: 10.1109/TMI.2023.3325824] [Medline: 37856266]

XSL•FO
RenderX

150. Li W, Yang J, Min X. Next-day medical activities recommendation model with double attention mechanism using generative adversarial network. J Healthc Eng 2022 Nov 7;2022:6334435 [FREE Full text] [doi: 10.1155/2022/6334435] [Medline: 36389105]
151. Barreto AG, de Oliveira JM, Gois FN, Cortez PC, de Albuquerque VH. A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks. Bioengineering (Basel) 2023 Sep 19;10(9):1098 [FREE Full text] [doi: 10.3390/bioengineering10091098] [Medline: 37760200]
152. Zhao L, Huang J. A distribution information sharing federated learning approach for medical image data. Complex Intell Systems 2023 Mar 29 (forthcoming) [FREE Full text] [doi: 10.1007/s40747-023-01035-1] [Medline: 37361966]
153. Yue G, Wei P, Zhou T, Song Y, Zhao C, Wang T, et al. Specificity-aware federated learning with dynamic feature fusion network for imbalanced medical image classification. IEEE J Biomed Health Inform 2023 Sep 26;PP (forthcoming). [doi: 10.1109/JBHI.2023.3319516] [Medline: 37751333]
154. Wang R, Lai J, Zhang Z, Li X, Vijayakumar P, Karuppiah M. Privacy-preserving federated learning for internet of medical things under edge computing. IEEE J Biomed Health Inform 2023 Feb;27(2):854-865. [doi: 10.1109/JBHI.2022.3157725] [Medline: 35259124]
155. Ma Y, Wang J, Yang J, Wang L. Model-heterogeneous semi-supervised federated learning for medical image segmentation. IEEE Trans Med Imaging 2024 Jan 01;PP (forthcoming). [doi: 10.1109/TMI.2023.3348982] [Medline: 38163306]
156. Mantey EA, Zhou C, Anajemba JH, Arthur JK, Hamid Y, Chowhan A, et al. Federated learning approach for secured medical recommendation in internet of medical things using homomorphic encryption. IEEE J Biomed Health Inform 2024 Jun;28(6):3329-3340. [doi: 10.1109/JBHI.2024.3350232] [Medline: 38190666]
157. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv Preprint posted online on June 17, 2021. [doi: 10.48550/arXiv.2106.09685]
158. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. arXiv Preprint posted online on May 23, 2023. [doi: 10.48550/arXiv.2305.14314]
159. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv Preprint posted online on Oct 1, 2015. [doi: 10.48550/arXiv.1510.00149]
160. Lu Y, Li C, Lu H, Yang J, Gao J, Shen Y. An empirical study of scaling instruct-tuned large multimodal models. arXiv Preprint posted online on September 18, 2023. [doi: 10.48550/arXiv.2309.09958]
161. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv Preprint posted online on March 9, 2015. [doi: 10.48550/arXiv.1503.02531]
162. Huang S, Dong L, Wang W, Hao Y, Singhal S, Ma S, et al. Language is not all you need: aligning perception with language models. arXiv Preprint posted online on February 27, 2023. [doi: 10.48550/arXiv.2302.14045]
163. Chang H, Zhang H, Barber J, Maschinot A, Lezama J, Jiang L, et al. Muse: text-to-image generation via masked generative transformers. arXiv Preprint posted online on January 2, 2023. [doi: 10.48550/arXiv.2301.00704]
164. Chen J, Yu J, Ge C, Yao L, Xie E, Wu Y, et al. PixArt-α: fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv Preprint posted online on September 30, 2023. [doi: 10.48550/arXiv.2310.00426]
165. Cheng Y, Wang D, Zhou P, Zhang T. A survey of model compression and acceleration for deep neural networks. arXiv Preprint posted online on October 23, 2017. [doi: 10.48550/arXiv.1710.09282]
166. Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, et al. In-datacenter performance analysis of a tensor processing unit. arXiv Preprint posted online on April 16, 2017 [FREE Full text] [doi: 10.1145/3079856.3080246]
167. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey. ACM Trans Intell Syst Technol 2024 Feb 22;15(2):1-38. [doi: 10.1145/3639372]
168. Hoover B, Strobelt H, Gehrmann S. exBERT: a visual analysis tool to explore learned representations in transformers models. arXiv Preprint posted online on October 11, 2019 [FREE Full text] [doi: 10.18653/v1/2020.acl-demos.22]
169. Wu T, Ribeiro MT, Heer J, Weld D. Polyjuice: generating counterfactuals for explaining, evaluating, and improving models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021 Presented at: ACL-IJCNLP 2021; August 1-6, 2021; Online. [doi: 10.18653/v1/2021.acl-long.523]
170. Chen H, Covert IC, Lundberg SM, Lee SI. Algorithms to estimate shapley value feature attributions. Nat Mach Intell 2023 May 22;5:590-601. [doi: 10.1038/s42256-023-00657-x]
171. Yordanov Y, Kocijan V, Lukasiewicz T, Camburu OM. Few-shot out-of-domain transfer learning of natural language explanations in a label-abundant setup. arXiv Preprint posted online on December 12, 2021 [FREE Full text]
172. Luo S, Ivison H, Han SC, Poon J. Local interpretations for explainable natural language processing: a survey. ACM Comput Surv 2024 Apr 25;56(9):1-36. [doi: 10.1145/3649450]
173. Dalvi F, Durrani N, Sajjad H, Belinkov Y, Bau A, Glass J. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. Proc AAAI Conf Artif Intell 2019 Jul 17;33(01):6309-6317. [doi: 10.1609/aaai.v33i01.33016309]
174. Hewitt J, Manning CD. A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL-HLT 2019; June 2-7, 2019; Minneapolis, MN.

XSL•FO
RenderX

175. Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024 Jun;6(6):e428-e432. [doi: 10.1016/s2589-7500(24)00061-x]
176. Price N. Problematic interactions between AI and health privacy. Utah Law Rev 2021;2021(4):925-936. [doi: 10.26054/0d-th4e-sgvq]
177. Gerke S, Minssen T, Cohen G. Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Cambridge, MA: Academic Press; 2020:295-336.
178. Becker J, Gerke S, Cohen IG. The development, implementation, and oversight of artificial intelligence in health care: legal and ethical issues. In: Valdés E, Lecaros JA, editors. Handbook of Bioethical Decisions. Volume I. Cham, Switzerland: Springer; 2023.
179. Ma W, Scheible H, Wang B, Veeramachaneni G, Chowdhary P, Sun A, et al. Deciphering stereotypes in pre-trained language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023 Presented at: EMNLP 2023; December 6-10, 2023; Singapore, Singapore.
180. Christos K. Multimodal LLMs: fairness and transparency concerns. Media Analysis, Verification and Retrieval Group (MeVer). 2023 Nov 2. URL: https://mever.gr/post/fairness-and-transparency-concerns-in-multimodal-llms/ [accessed 2024-09-13]
181. He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. arXiv Preprint posted online on October 9, 2023 [FREE Full text] [doi: 10.2139/ssrn.4809363]
182. Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. ACM J Data Inf Qual 2023 Jun 22;15(2):1-21. [doi: 10.1145/3597307]
183. Lee N, Bang Y, Lovenia H, Cahyawijaya S, Dai W, Fung P. Survey of social bias in vision-language models. arXiv Preprint posted online on September 24, 2023 [FREE Full text]
184. Reddy AG, Bachu S, Dash S, Sharma C, Sharma A, Balasubramanian VN. On counterfactual data augmentation under confounding. arXiv Preprint posted online on May 29, 2023 [FREE Full text]
185. Chen RJ, Wang JJ, Williamson DF, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742 [FREE Full text] [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
186. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? Georgetown Law J 2020 May 1;108:1425-1469 [FREE Full text] [doi: 10.2139/ssrn.3529576]
187. Friesen P, Douglas-Jones R, Marks M, Pierce R, Fletcher K, Mishra A, et al. Governing AI-driven health research: are IRBs up to the task? Ethics Hum Res 2021 Mar;43(2):35-42. [doi: 10.1002/eahr.500085] [Medline: 33683015]
188. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. arXiv Preprint posted online on December 14, 2020 [FREE Full text]
189. Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. Science 2018 Nov 09;362(6415):690-694 [FREE Full text] [doi: 10.1126/science.aau4832] [Medline: 30309907]
190. Na L, Yang C, Lo CC, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. JAMA Netw Open 2018 Dec 07;1(8):e186040 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.6040] [Medline: 30646312]
191. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339(6117):321-324. [doi: 10.1126/science.1229566] [Medline: 23329047]
192. Aghajanyan A, Huang B, Ross C, Karpukhin V, Xu H, Goyal N, et al. CM3: a causal masked multimodal model of the internet. arXiv Preprint posted online on January 19, 2022 [FREE Full text]
193. Huang Y, Yuan Q, Sheng X, Yang Z, Wu H, Chen P, et al. AesBench: an expert benchmark for multimodal large language models on image aesthetics perception. arXiv Preprint posted online on January 16, 2024 [FREE Full text]
194. Wang X, Zhou Y, Liu X, Lu H, Xu Y, He F, et al. Mementos: a comprehensive benchmark for multimodal large language model reasoning over image sequences. arXiv Preprint posted online on January 19, 2024 [FREE Full text]
195. Fu C, Chen P, Shen Y, Qin Y, Zhang M, Lin X, et al. MME: a comprehensive evaluation benchmark for multimodal large language models. arXiv Preprint posted online on June 23, 2023 [FREE Full text]
196. Yang X, Wu W, Feng S, Wang M, Wang D, Li Y, et al. MM-BigBench: evaluating multimodal models on multimodal content comprehension tasks. arXiv Preprint posted online on October 13, 2023 [FREE Full text]
197. Ge W, Chen S, Chen GH, Chen Z, Chen J, Yan S, et al. MLLM-Bench: evaluating multimodal LLMs with per-sample criteria. arXiv Preprint posted online on November 23, 2023 [FREE Full text]
198. Li S, Zhang Y, Zhao Y, Wang Q, Jia F, Liu Y, et al. VLM-Eval: a general evaluation on video large language models. arXiv Preprint posted online on November 20, 2023 [FREE Full text]
199. The act texts. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/the-act/ [accessed 2024-09-13]

Abbreviations

AGI: artificial general intelligence

XSL•FO
RenderX

AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
CGM: continuous glucose monitoring
CMM: causal masked multimodal
CNN: convolutional neural network
CONCH: Contrastive Learning From Captions for Histopathology
COPD: chronic obstructive pulmonary disease
CT: computed tomography
DICM: Digital Imaging and Communications in Medicine
DPO: dynamic policy optimization
EHR: electronic health record
GPT: generative pretrained transformer
GPU: graphics processing unit
H&E: hematoxylin and eosin
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
LLM: large language model
LSTM: long short-term memory
M-LLM: multimodal large language model
MLM: mask language modeling
MRI: magnetic resonance imaging
NLP: natural language processing
RNN: recurrent neural network
TPU: tensor processing unit

Edited by A Coristine, A Mavragani; submitted 13.04.24; peer-reviewed by B Solaiman, G Sun; comments to author 14.05.24; revised version received 07.08.24; accepted 20.08.24; published 25.09.24.

Please cite as:
AlSaad R, Abd-alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, Sheikh J
Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook
J Med Internet Res 2024;26:e59505
URL: https://www.jmir.org/2024/1/e59505
doi:10.2196/59505
PMID:

©Rawan AlSaad, Alaa Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, Javaid Sheikh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Engine of Innovation in Hospital Pharmacy: Applications and Reflections of ChatGPT

Xingang Li[1], PhD; Heng Guo[1], MD; Dandan Li[1], MD; Yingming Zheng[1], MD

Department of Pharmacy, Beijing Friendship Hospital, Capital Medical University, Beijing, China

Corresponding Author:
Xingang Li, PhD
Department of Pharmacy, Beijing Friendship Hospital
Capital Medical University
No 95, Yongan Road, Xicheng District
Beijing, 100050
China
Phone: 86 1081608511
Email: lxg198320022003@163.com

Abstract

Hospital pharmacy plays an important role in ensuring medical care quality and safety, especially in the area of drug information retrieval, therapy guidance, and drug-drug interaction management. ChatGPT is a powerful artificial intelligence language model that can generate natural-language texts. Here, we explored the applications and reflections of ChatGPT in hospital pharmacy, where it may enhance the quality and efficiency of pharmaceutical care. We also explored ChatGPT's prospects in hospital pharmacy and discussed its working principle, diverse applications, and practical cases in daily operations and scientific research. Meanwhile, the challenges and limitations of ChatGPT, such as data privacy, ethical issues, bias and discrimination, and human oversight, are discussed. ChatGPT is a promising tool for hospital pharmacy, but it requires careful evaluation and validation before it can be integrated into clinical practice. Some suggestions for future research and development of ChatGPT in hospital pharmacy are provided.

(J Med Internet Res 2024;26:e51635) doi:10.2196/51635

KEYWORDS

ChatGPT; hospital pharmacy; natural language processing; drug information; drug therapy; drug interaction; scientific research; innovation; pharmacy; quality; safety; pharmaceutical care; tool; medical care quality

Introduction

Hospital pharmacists play an important role in ensuring the safe and effective use of medications in hospitals. However, pharmacists also face many challenges, such as the increasing complexity of patients' medication regimens, the potential risks of medication errors and drug-drug interactions, the constant updating of pharmaceutical knowledge, and the growing volume of pharmaceutical literature [1-4]. Due to the large volume of medical-related literature, pharmacists spend a lot of time searching, analyzing, and applying relevant information in clinics. In addition, it takes a lot of effort to keep up with the latest information on new drugs, clinical guidelines, adverse drug reactions, and drug-drug interactions. Therefore, how to quickly and effectively obtain and interpret medication-related information has emerged as a challenge for pharmacists.

With progress in deep learning and natural language processing technologies, generative large language models like ChatGPT have come to the forefront. Its outstanding performance in understanding context and generating text lays a good foundation for its wide application in various fields [5-9]. If ChatGPT is applied to hospital pharmacy, it may become a new tool for hospital pharmacists to solve the aforementioned problems and challenges, improve work efficiency, ensure safe and rational use of drugs, and even greatly change the status quo of hospital pharmacy operations and research [10].

This article aims to explore the application potential of ChatGPT in daily hospital pharmacy operations and research, and analyze the challenges and problems faced by ChatGPT. Simultaneously, we provide our suggestions for the development of artificial intelligence (AI) in this field.

Overview of ChatGPT

The ChatGPT system is built upon a generative pretrained transformer (GPT) model, which has advantages in natural language processing (NLP). GPT pretrains a large amount of

XSL•FO
RenderX

text data, such as books, articles, website content, etc, from which it acquires the ability to understand natural language deeply. The transformer architecture within GPT captures long-distance dependencies and contextual relationships in text through a self-attention mechanism, which enables ChatGPT to generate coherent and context-relevant responses, outputting content similar to human-written content. The working principle underlying its generation of text can be summarized as follows (Figure 1):

Figure 1. ChatGPT's working principle of generating text.

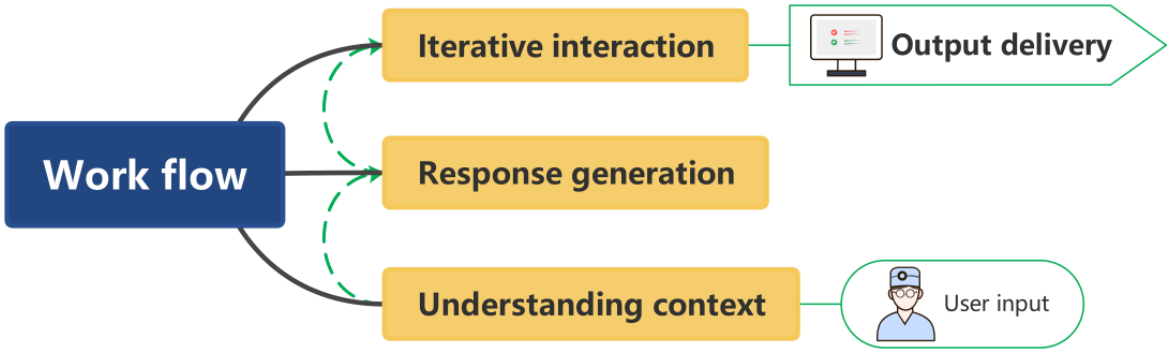

1. User input: ChatGPT receives user input information in text form, including single sentences and longer dialogue contexts.
2. Understanding context: ChatGPT processes the input text and uses pretrained parameters to understand the context and extract relevant information.
3. Response generation: based on its understanding of the context, ChatGPT uses its knowledge base and language patterns to construct replies and generate coherent and context-relevant responses.
4. Iterative interaction: in continuous dialogue, ChatGPT can retain dialogue history and context, enabling it to use previous interactions to generate more appropriate responses and ensure smooth dialogue.
5. Output delivery: ChatGPT presents the generated response to the user in human-readable text to complete the dialogue.

Due to its advanced transformer model, ChatGPT excels at understanding and generating language. This capability expands its application potential across diverse fields, notably enhancing interactions and decision support. In medicine, it promises improvements in patient care, telehealth, information retrieval, and education, etc [11-13].

Application of ChatGPT in Hospital Pharmacy

Daily Operations in Hospital Pharmacy

Drug Information Retrieval and Interpretation

In daily hospital pharmacy operations, pharmacists often need to quickly and accurately retrieve the latest pharmaceutical information. ChatGPT could be used as an auxiliary tool for drug information retrieval and interpretation [14]. Pharmacists can enter specific information for querying, such as drug name, indication, contraindication, medication regimen, adverse reactions, clinical guideline, etc, and provide relevant information [15], thus helping pharmacists save time on information retrieval, improve learning efficiency, and keep their pharmaceutical knowledge up-to-date [16]. In a recent study assessing ChatGPT's proficiency in clinical pharmacy knowledge, the AI language model was presented with 264 multiple-choice questions commonly used to maintain clinical pharmacists' basic knowledge. ChatGPT not only achieved a higher accuracy rate of 79% in answering these questions but also demonstrated excellent reproducibility and quality of substantiation, surpassing the performance of pharmacists by a notable margin [17].

Prescription Review and Medication Guidance

ChatGPT can assist pharmacists in the review of prescriptions and offer guidance for the rational use of medications. When used for prescription review, ChatGPT can interact with prescribers, feedback-inappropriate prescriptions or medication orders, and improve the work efficiency of prescription review. For controversial prescriptions or medication orders, they are then submitted to clinical pharmacists for manual review. In terms of medication guidance, pharmacists can enter relevant information of specific patients, such as medical history, laboratory test results and comorbidities, and seek advice from ChatGPT on the most appropriate drug therapy regimen for the patient. By considering the patient's individual factors, drug information, and evidence-based guidelines and other comprehensive factors, ChatGPT supports pharmacists in making decisions regarding drug selection, medication regimen adjustment, and individualized treatment [18]. Finally, pharmacists integrate the relevant suggestions with their own professional medical knowledge and provide individualized medication guidance and optimized medication regimen for patients [15]. The potential of ChatGPT in proactive polypharmacy management was explored. By processing clinical scenarios, ChatGPT demonstrated an ability to make deprescribing decisions, suggesting it can assist in reviewing prescriptions and guiding medication use. The findings indicate that with specialized training, ChatGPT could offer valuable clinical assistance to primary care physicians in managing complex medication regimens for older patients [19].

XSL•FO
RenderX

Drug Interaction Management

For patients with multiple medications, managing drug interactions is important and could reduce the potential risks. Upon receiving detailed clinical information and medication profiles from patients, an in-depth analysis can be conducted to identify and prevent possible drug interactions. By accessing specialized databases, a comprehensive evaluation of these drug interactions can be carried out, providing professional advice to patients on how to avoid harmful interactions [15]. This helps pharmacists optimize medication regimens, minimize the risk of adverse reactions, and improve the safety of patient medication. There have been reports that ChatGPT has achieved 95% accuracy in predicting and explaining drug-drug interactions. It can generate logical and coherent text output, rather than simply listing the types and mechanisms of interactions [20]. The application of ChatGPT in medication therapy management has the potential to enhance patient safety and involvement, lower health care costs, and assist health care providers in medication management and identifying drug interactions [21]. In a comparative analysis of AI chatbots' ability to predict drug-drug interactions, Bing AI outperformed others, including different versions of ChatGPT. Despite some limitations, these AI tools showed promise, with ChatGPT demonstrating potential in accurately identifying drug-drug interactions. This underscores the budding utility of AI, and specifically ChatGPT, in streamlining drug interaction management to bolster patient safety [22].

Medication Education and Consultation

Effective medication education and consultation are essential for encouraging patients to actively participate in drug therapy, which is important for promoting the safe and effective use of drugs, improving medication adherence, and optimizing treatment outcomes. However, the reality is that clinical pharmacists have limited time and resources, which limited the extent of medication education and consultation they can provide. ChatGPT could be used as a virtual resource for patient medication education [23], providing information on drug indications, usage, dosage, potential adverse effects, precautions, etc [24]. By serving as a digital assistant, ChatGPT can support pharmacists in consulting on information related to drugs and diseases [25]. It can answer common questions raised by patients and provide professional explanations, such as when to take medicine, how many times a day, what dose to take, whether some foods or alcohol have any effect on medicine, etc, to help patients reduce the risk of medication errors.

ChatGPT's language processing ability is outstanding, which can solve the language barrier problem in the process of medication education and consultation. ChatGPT's multilingual support ensures comprehensive medication guidance for patients, regardless of their language preferences [26]. A real-world study found that ChatGPT exhibited a higher appropriateness rate in responding to public medication consultation questions compared to those asked by health care providers in a hospital setting [27]. It is particularly important to note that the medication education and consultation services provided by ChatGPT need to be reviewed by pharmacists to ensure the accuracy of information [28]. Pharmacists support the use of ChatGPT in pharmacy practice but have concerns about its use due to ethical reasons, legal problems, privacy concerns, worries about the accuracy of the data generated, data learning, and bias risk [29]. A summary of the application of ChatGPT in daily work of hospital pharmacy is provided in Figure 2.

Figure 2. Summary of the application of ChatGPT in daily work in hospital pharmacy.

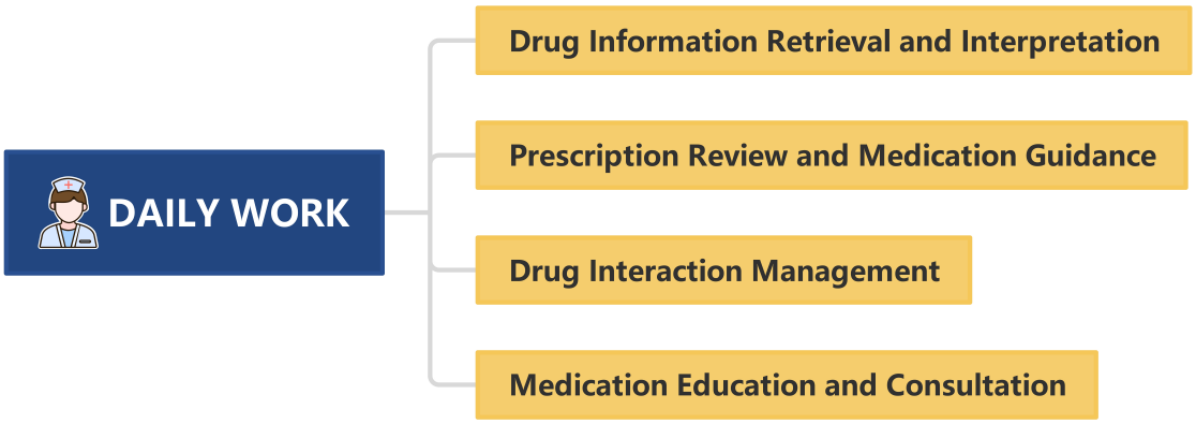

Application of ChatGPT in Research

Literature Review and Analysis

ChatGPT serves as a tool for pharmacists delving into literature, as it aids in the reviews and analyses process [14]. When presented with specific research topics or questions by researchers, it facilitates the retrieval of relevant literature and synthesizes findings, innovations, future trends, and current research gaps. With ChatGPT's NLP ability, researchers can accelerate the screening, learning, and summarization of related literature and relevant knowledge and development trends in specific fields. For nondomain knowledge, it is possible to sort out relevant knowledge points through progressive questioning, quickly understand the field, and build a knowledge system. By summarizing and abstracting in a concise way, it provides a way for researchers to quickly understand the research achievements of predecessors, saving time for consulting a large volume of literature. In a systematic review assessing medication adherence improvement in patients with ischemic stroke via mobile health interventions, ChatGPT was pitted against human

XSL•FO
RenderX

researchers for identifying relevant studies. While human researchers showed a slightly higher precision (0.86) and relevance percentage (9.8%), ChatGPT demonstrated significant efficiency by identifying a substantial number of relevant studies, including a majority of studies that reported adherence improvements, in a fraction of the time required by human researchers. This case illustrates the utility of ChatGPT in literature review and analysis, suggesting that its speed and broad search capabilities can augment traditional research methods, particularly in the initial stages of study identification [30].

Research Topic Selection and Research Design

Researchers can engage in role-playing to consult ChatGPT on a range of questions related to specific field-related research topics, including representative literature, study design, inclusion and exclusion criteria, sample size calculation, research methods, quality control, etc. By offering expert-level insights, ChatGPT serves as a reference for researchers [31]. With ChatGPT's guidance, researchers not only enrich the research process but also enhance the overall integrity and impact of the scholarly work produced. A study explores the utility of AI-based transformers, specifically ChatGPT, in aiding epidemiological research. By converting STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines into prompts, ChatGPT was assessed on its ability to generate coherent and relevant responses for observational studies. With mean scores of 3.6 for coherence and 3.3 for relevance, ChatGPT proved to be a valuable asset, particularly in adhering to recognized research standards. However, the study underscores the need for a critical approach to AI-generated content, highlighting the importance of user expertise and awareness of ethical considerations in scientific research [31].

Data Statistical Analysis and Plotting

ChatGPT offers a robust suite of support for data statistical analysis, enabling researchers to extract meaningful information from datasets. ChatGPT's Code Interpreter plugin is a powerful tool that can translate natural language into executable code for data statistical analysis and plotting. After entering datasets or specific research questions, ChatGPT suggests suitable statistical techniques and interprets the data accordingly. ChatGPT is adept at generating code in programming languages such as Python, SAS, R, etc, and then using Python and other tools for data mining, analysis, and plotting [32]. ChatGPT can work with interactive plotting software to assist researchers in generating various types of figures, including scatter plots, bar charts, line charts, heat maps, and network graphs. By understanding researchers' input or description of data, ChatGPT could offer plotting suggestions based on the research objectives of the data, recommending the most appropriate graphic types, layouts, and color schemes that best represent the information conveyed. One study compared the statistical analysis capabilities of ChatGPT-4 with traditional biostatistical software (SAS, SPSS, and R) using data from the China Health and Nutrition Survey. ChatGPT-4 demonstrated high consistency, efficiency, and user-friendliness in descriptive statistics, and maintained these advantages in intergroup and correlational analyses despite minor discrepancies with the traditional software. ChatGPT-4 is a powerful tool for epidemiological data analysis, particularly for researchers with intermediate data analysis experience. It suggests that AI's integration with data analysis platforms can simplify operations and enhance research focus on result interpretation, potentially revolutionizing epidemiological and medical research [33].

Academic Writing

ChatGPT has a very obvious advantage in assisting paper writing [34]. It offers strategic advice on the structure of the paper, guiding researchers in the logical arrangement of the sections and content. In academic paper writing, the accuracy and fluency of language expression are very important. ChatGPT aids in polishing the manuscript's language and improving the paper's quality and readability [28,35,36]. However, it should be noted that although ChatGPT can provide a lot of help in paper writing, there is the potential for misunderstandings or errors [37]. It should not supplant the professional knowledge and judgment of researchers [38,39]. Therefore, when using ChatGPT, researchers should verify the reliability of its output results [40]. Simultaneously, ChatGPT should not be used to write entire papers, but rather be used as an auxiliary tool to assist researchers in research and writing [41,42]. ChatGPT serves as a versatile technological tool, one whose influence can be profoundly positive or negative, depending on the wisdom and discretion with which it is used. It holds significant promise as an asset to scientific writing; yet, it lacks the ability to supersede the irreplaceable qualities of human composition and the nuanced faculties of critical thought. Hence, the use of AI in academic writing should be conducted with unwavering adherence to the foundational principles of honesty, rigor, and originality [43]. A summary of the application of ChatGPT in research is provided in Figure 3.

XSL•FO
RenderX

Figure 3. Comprehensive artificial intelligence–enhanced scientific research framework incorporating literature review, topic selection, design, data analysis, and result interpretation.

Challenges and Issues Faced by ChatGPT

Although ChatGPT enhances the efficiency of daily work in hospital pharmacy and provides assistance in scientific research, it still faces some technical hurdles and challenges [28,44-46]. ChatGPT's responses rely on the context provided in the input. However, it may sometimes misunderstand or misinterpret vague queries or complex medical scenarios, resulting in potentially inaccurate or irrelevant responses. Repeated questions to ChatGPT do not ensure consistent answers [24]. Its knowledge is rooted in pretraining data, which means that it lacks the ability to update medical guidelines, latest research results, or constantly evolving clinical practices in real time. Therefore, its answers may not be the most up-to-date information [31]. In addition, biases present in its training data may influence its responses, introducing the risk of biased or incorrect content [47,48]. The performance of ChatGPT-4 and Google Bard in providing accurate preventive medicine and primary care recommendations were evaluated, and a considerable portion of responses from both AI models contained inaccuracies or missing details. The findings underscore the necessity for AI tools to be regularly updated, especially in fast-paced medical fields, and to be used as supplementary resources rather than sole authorities on medical information [49].

Incorporating ChatGPT into hospital settings requires a careful approach to patient privacy and ethical considerations [50]. When interacting with ChatGPT, sensitive information should be deidentified to protect patient privacy. In daily work and scientific research, one should be clear about the advantages and disadvantages of using ChatGPT and obtain informed consent from patients before using ChatGPT. ChatGPT, while capable of addressing professional inquiries and providing recommendations, serves as an auxiliary decision support tool rather than a substitute for the professional knowledge of clinical pharmacists [51]. Pharmacists or researchers should critically evaluate and verify the information provided by ChatGPT before making a final decision, and they should be responsible for the final decision (Figure 4).

Figure 4. Key challenges and issues in artificial intelligence applications highlighting inaccurate responses, outdated information, patient privacy, and ethical concerns.

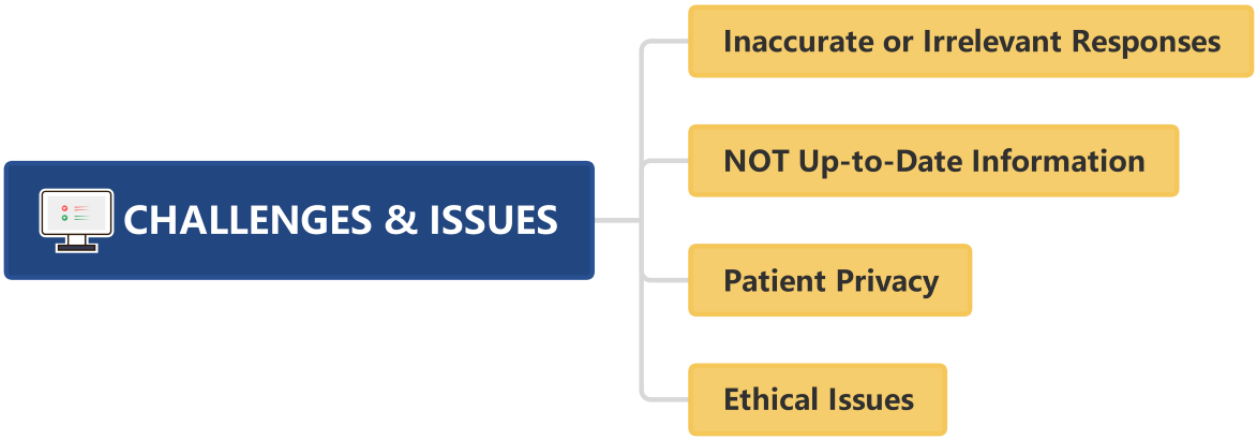

XSL•FO
RenderX

Thoughts on AI Development

While the potential of ChatGPT in enhancing hospital pharmacy operations is significant, the practicalities of its integration into existing workflows require careful consideration. The necessary infrastructure includes robust IT systems capable of handling the computational demands of AI, secure data storage solutions to safeguard patient privacy, and a reliable internet connection to ensure continuous access to ChatGPT's services. Staff training is equally vital, encompassing not only the technical aspects of using the AI tool but also the interpretation of its outputs within a clinical context. Change management strategies must be used to facilitate a smooth transition, addressing potential resistance and fostering a culture of continuous learning and adaptation. Piloting ChatGPT in select areas of the pharmacy, soliciting feedback from frontline staff, and iterating on the implementation plan based on real-world experiences can significantly mitigate challenges and optimize outcomes. Ongoing evaluation and refinement of the integration process are imperative to ensure that the benefits of ChatGPT are fully realized and sustained over time.

To facilitate the development and scientific rational use of AI, it is imperative for governments and industry associations to establish pertinent innovation and regulatory frameworks [37,50]. Innovation policies include formulating technical standards for AI, supporting the research and development of AI, increasing investment in AI technology, strengthening talent cultivation for AI, promoting popularization and promotion of AI, etc. Regulatory policies include strengthening protection of intellectual property rights related to AI, ensuring that AI is only used for reasonable and legal activities, ensuring that relevant policies are conducive to innovation and development of AI, protection of privacy by AI, ethical issues related to AI, etc. The United States has already formulated relevant policies in this regard.

To optimize ChatGPT's role in hospital pharmacy, enhancing its model performance and reliability is important [26]. On the basis of the original data model, it is crucial to strengthen its training on professional medical datasets. This will enhance its understanding of medical terminology, drug information, and clinical background knowledge. Refining its ability to deliver precise responses to medical inquiries is paramount. By bolstering ChatGPT's interpretive skills in grasping nuanced contexts, it can generate more accurate and contextually fitting answers. Advancing its contextual understanding will not only minimize the incidence of incorrect responses but also elevate its efficacy in intricate medical situations. Medical professionals should participate in training and calibration of ChatGPT and regularly evaluate and verify its performance in medical applications. Through rigorous testing, comparison with gold standards for calibration, and incorporation of feedback from medical professionals, its reliability in the medical field can be enhanced.

AI-assisted diagnosis is based on the intelligent analysis of big data. It is necessary to solve the problem of fragmentation of medical health data to achieve a leap from data to knowledge to intelligence. Efforts must be made to bridge the gaps among disparate data sources, creating a comprehensive medical knowledge center that facilitates a connection between individuals and health care providers. To ensure the widespread adoption of AI in health care, developing interfaces that seamlessly integrate AI technologies with existing hospital information systems is essential. One study introduces PMC-LLaMA, an open-source large language model tailored for medical applications. By integrating a vast database of biomedical literature and textbooks, and fine-tuning with domain-specific instructions, PMC-LLaMA has reportedly outperformed models like ChatGPT in medical question–answering benchmarks. The research provides an open-source framework that could drive future developments in medical AI, offering a foundational model that can be further trained and adapted for various medical tasks [52]. This integration will enable AI to play role in enhancing hospitals' information infrastructure (Figure 5).

Figure 5. Strategic recommendations for the future development of artificial intelligence.

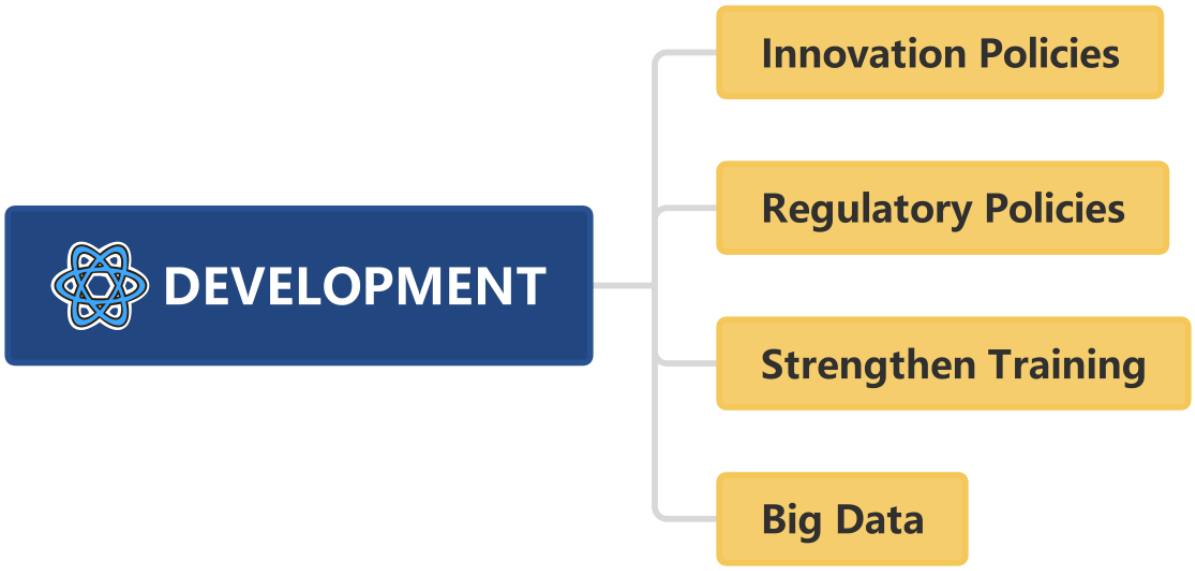

XSL•FO
RenderX

Integrating ChatGPT into the practice of hospital pharmacy requires a detailed strategy that addresses technology, training, data security, and ethical compliance across several key areas. Initially, a needs assessment determines the scope of assistance that ChatGPT can provide, such as drug information retrieval and patient counseling. Subsequently, IT infrastructure is strengthened to ensure that servers, data storage, and network security can support the integration of ChatGPT, with the development of data interfaces for seamless integration with hospital systems like electronic health records. Concurrently, pharmacy personnel are trained in the use of ChatGPT, emphasizing best practices and potential risks. In terms of privacy and compliance, it is imperative that the use of ChatGPT adheres to data protection regulations such as HIPAA (Health Insurance Portability and Accountability Act) and undergoes review by an ethics committee to prevent ethical issues. Moreover, strict data security protocols and oversight mechanisms are implemented to protect patient information and monitor ChatGPT's performance. Intuitive user interfaces have been designed to facilitate interaction between medical staff and ChatGPT, and testing in real environments has been conducted, with feedback collected for iterative improvements. Regular assessments of ChatGPT's performance, including accuracy and user satisfaction, are conducted, and necessary improvements are made based on the evaluation results. Medical personnel are encouraged to engage in research on the application of ChatGPT in pharmacy and to publish their findings to share knowledge. Patients are clearly informed about the supportive role of ChatGPT, its benefits, and the data security measures in place, maintaining full transparency. Additionally, collaboration across various disciplines such as pharmacy, IT, data science, and health care management is fostered to ensure the effective integration of ChatGPT. Finally, clear policies and guidelines for AI use are established to ensure that all relevant personnel understand and comply with these regulations. Through these comprehensive strategies, hospital pharmacy departments can improve work efficiency while ensuring patient safety and data protection.

Conclusion

In summary, with its advanced language generation and understanding capabilities, ChatGPT has great value and broad application prospects in the field of hospital pharmacy. Its potential to streamline daily tasks and spur advancements in scientific research is noteworthy. However, there are still some challenges with ChatGPT's predictive performance, privacy protection, and ethics. In the future, through policy formulation, model optimization, information integration, and other measures are required to further promote the development and application of AI in hospital pharmacy. ChatGPT is expected to contribute to the transformation of hospital pharmacy in the future, benefiting patients, pharmacists, medical practitioners, and the broader health care ecosystem.

Acknowledgments

This study was funded by Research and Application of Clinical Characteristic Diagnosis and Treatment Technology in Beijing (grant Z221100007422032).

Authors' Contributions

All authors contributed to the writing and reviewing of this viewpoint paper.

Conflicts of Interest

None declared.

References

1. Bishop J, Kong D, Schulz T, Thursky K, Buising K. Meeting the challenge for effective antimicrobial stewardship programs in regional, rural and remote hospitals - what can we learn from the published literature? Rural Remote Health 2018 May;18(2):4442-4442 [FREE Full text] [doi: 10.22605/RRH4442] [Medline: 29792036]
2. Kennedy MJ. Personalized medicines - are pharmacists ready for the challenge? Integr Pharm Res Pract 2018 Sep;7:113-123 [FREE Full text] [doi: 10.2147/IPRP.S133083] [Medline: 30310772]
3. Sakeena MHF, Bennett AA, McLachlan AJ. Enhancing pharmacists' role in developing countries to overcome the challenge of antimicrobial resistance: a narrative review. Antimicrob Resist Infect Control 2018 May 02;7:63 [FREE Full text] [doi: 10.1186/s13756-018-0351-z] [Medline: 29744044]
4. Zhang T, Lambert H, Zhao L, Liu R, Shen X, Wang D, et al. Antibiotic stewardship in retail pharmacies and the access-excess challenge in China: a policy review. Antibiotics (Basel) 2022 Jan 21;11(2):141 [FREE Full text] [doi: 10.3390/antibiotics11020141] [Medline: 35203744]
5. Nguyen J, Pepping CA. The application of ChatGPT in healthcare progress notes: a commentary from a clinical and research perspective. Clin Transl Med 2023 Jul 02;13(7):e1324 [FREE Full text] [doi: 10.1002/ctm2.1324] [Medline: 37394880]
6. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci U S A 2023 Jul 25;120(30):e2305016120 [FREE Full text] [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
7. Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese Society of Hypertension guidelines. Circ J 2023;87(7):1030-1033. [doi: 10.1253/circj.cj-23-0308]

XSL•FO
RenderX

8. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
9. Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clin Infect Dis 2023 Nov 11;77(9):1322-1328 [FREE Full text] [doi: 10.1093/cid/ciad407] [Medline: 37399030]
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
11. Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. J Med Internet Res 2023 Jun 22;25:e48392 [FREE Full text] [doi: 10.2196/48392] [Medline: 37347508]
12. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec 12;28(1):2220920 [FREE Full text] [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
13. Cross J, Robinson R, Devaraju S, Vaughans A, Hood R, Kayalackakom T, et al. Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. Cureus 2023 Jul;15(7):e41399 [FREE Full text] [doi: 10.7759/cureus.41399] [Medline: 37426402]
14. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? J Am Soc Nephrol 2023 Aug 01;34(8):1302-1304. [doi: 10.1681/ASN.0000000000000166] [Medline: 37254254]
15. Wang H, Ding YJ, Luo Y. Future of ChatGPT in pharmacovigilance. Drug Saf 2023 Aug 12;46(8):711-713 [FREE Full text] [doi: 10.1007/s40264-023-01315-2] [Medline: 37306853]
16. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol 2024 Jan 13;90(1):232-238. [doi: 10.1111/bcp.15896] [Medline: 37626010]
17. van Nuland M, Erdogan A, Açar C, Contrucci R, Hilbrants S, Maanach L, et al. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. J Clin Pharmacol 2024 Sep 16;64(9):1095-1100. [doi: 10.1002/jcph.2443] [Medline: 38623909]
18. Murugan M, Yuan B, Venner E, Ballantyne C, Robinson K, Coons J, et al. Empowering personalized pharmacogenomics with generative AI solutions. J Am Med Inform Assoc 2024 May 20;31(6):1356-1366. [doi: 10.1093/jamia/ocae039] [Medline: 38447590]
19. Rao A, Kim J, Lie W, Pang M, Fuh L, Dreyer KJ, et al. Proactive polypharmacy management using large language models: opportunities to enhance geriatric care. J Med Syst 2024 Apr 18;48(1):41. [doi: 10.1007/s10916-024-02058-y] [Medline: 38632172]
20. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus 2023 Mar;15(3):e36272 [FREE Full text] [doi: 10.7759/cureus.36272] [Medline: 37073184]
21. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. J Am Pharm Assoc (2003) 2024 Mar;64(2):422-428.e8 [FREE Full text] [doi: 10.1016/j.japh.2023.11.023] [Medline: 38049066]
22. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard against conventional drug-drug interactions clinical tools. Drug Healthc Patient Saf 2023;15:137-147 [FREE Full text] [doi: 10.2147/DHPS.S425858] [Medline: 37750052]
23. Luo Y, Hu N. Can ChatGPT be a new educational tool in medicine? Med Clin (Barc) 2023 Oct 27;161(8):363-364. [doi: 10.1016/j.medcli.2023.05.018] [Medline: 37438191]
24. He N, Yan Y, Wu Z, Cheng Y, Liu F, Li X, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. J Telemed Telecare 2023 Jun 22:1357633X231181922. [doi: 10.1177/1357633X231181922] [Medline: 37350055]
25. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating artificial intelligence responses to public health questions. JAMA Netw Open 2023 Jun 01;6(6):e2317517 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.17517] [Medline: 37285160]
26. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-English-speaking outpatient clinic settings. Digit Health 2023 Jun 26;9:20552076231184091 [FREE Full text] [doi: 10.1177/20552076231184091] [Medline: 37434733]
27. Hsu H, Hsu K, Hou S, Wu C, Hsieh Y, Cheng Y. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med Educ 2023 Aug 21;9:e48433 [FREE Full text] [doi: 10.2196/48433] [Medline: 37561097]
28. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023 Jul 13;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
29. Abu Hammour K, Alhamad H, Al-Ashwal FY, Halboup A, Abu Farha R, Abu Hammour K. ChatGPT in pharmacy practice: a cross-sectional exploration of Jordanian pharmacists' perception, practice, and concerns. J Pharm Policy Pract 2023 Oct 03;16(1):115 [FREE Full text] [doi: 10.1186/s40545-023-00624-2] [Medline: 37789443]
30. Ruksakulpiwat S, Phianhasin L, Benjasirisan C, Ding K, Ajibade A, Kumar A, et al. Assessing the efficacy of ChatGPT versus human researchers in identifying relevant studies on mHealth interventions for improving medication adherence in

XSL•FO
RenderX

patients with ischemic stroke when conducting systematic reviews: comparative analysis. JMIR Mhealth Uhealth 2024 May 06;12:e51526 [FREE Full text] [doi: 10.2196/51526] [Medline: 38710069]
31. Sanmarchi F, Bucci A, Nuzzolese AG, Carullo G, Toscano F, Nante N, et al. A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of ChatGPT using the STROBE checklist for observational studies. Z Gesundh Wiss 2023 May 26:1-36 [FREE Full text] [doi: 10.1007/s10389-023-01936-y] [Medline: 37361298]
32. Cloesmeijer ME, Janssen A, Koopman SF, Cnossen MH, Mathôt RAA, SYMPHONY consortium. ChatGPT in pharmacometrics? Potential opportunities and limitations. Br J Clin Pharmacol 2024 Jan 06;90(1):360-365. [doi: 10.1111/bcp.15895] [Medline: 37621112]
33. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. J Glob Health 2024 Mar 29;14:04070 [FREE Full text] [doi: 10.7189/jogh.14.04070] [Medline: 38547497]
34. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science 2023 Jul 14;381(6654):187-192. [doi: 10.1126/science.adh2586] [Medline: 37440646]
35. Lenharo M. ChatGPT gives an extra productivity boost to weaker writers. Nature 2023 Jul 13. [doi: 10.1038/d41586-023-02270-9] [Medline: 37443308]
36. Darzidehkalani E. ChatGPT in medical publications. Radiology 2023 Jun 01;307(5):e231188. [doi: 10.1148/radiol.231188] [Medline: 37278630]
37. Garcia MB. Using AI tools in writing peer review reports: should academic journals embrace the use of ChatGPT? Ann Biomed Eng 2024 Feb 27;52(2):139-140. [doi: 10.1007/s10439-023-03299-7] [Medline: 37368125]
38. Munoz-Zuluaga C, Zhao Z, Wang F, Greenblatt M, Yang H. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine. Clin Chem 2023 Aug 02;69(8):939-940. [doi: 10.1093/clinchem/hvad058] [Medline: 37231970]
39. Graham F. Daily briefing: the pros and cons of writing a paper with ChatGPT. Nature 2023 Jul 07. [doi: 10.1038/d41586-023-02264-7] [Medline: 37433941]
40. Baker N, Thompson B, Fox D. ChatGPT can write a paper in an hour - but there are downsides. Nature 2023 Jul 12. [doi: 10.1038/d41586-023-02298-x] [Medline: 37438635]
41. Ring J. Artificial intelligence-ChatGPT and scientific publishing. J Eur Acad Dermatol Venereol 2023 Jul 12;37(7):1253-1254. [doi: 10.1111/jdv.19176] [Medline: 37306599]
42. Shafiee A. Matters arising: authors of research papers must cautiously use ChatGPT for scientific writing. Int J Surg 2023 Sep 01;109(9):2853-2854 [FREE Full text] [doi: 10.1097/JS9.0000000000000515] [Medline: 37222674]
43. He S, Yang F, Zuo J, Lin Z. ChatGPT for scientific paper writing-promises and perils. Innovation (Camb) 2023 Nov 13;4(6):100524 [FREE Full text] [doi: 10.1016/j.xinn.2023.100524] [Medline: 38028132]
44. Dashti M, Londono J, Ghasemi S, Moghaddasi N. How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? J Prosthet Dent 2023 Jul 10. [doi: 10.1016/j.prosdent.2023.05.023] [Medline: 37438164]
45. Heng JJY, Teo DB, Tan LF. The impact of Chat Generative Pre-trained Transformer (ChatGPT) on medical education. Postgrad Med J 2023 Sep 21;99(1176):1125-1127. [doi: 10.1093/postmj/qgad058] [Medline: 37466157]
46. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev 2023 Apr 29;12(1):72 [FREE Full text] [doi: 10.1186/s13643-023-02243-z] [Medline: 37120563]
47. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. Eur J Hosp Pharm 2023 Jun 01:ejhpharm-2023-003750. [doi: 10.1136/ejhpharm-2023-003750] [Medline: 37263772]
48. Kao Y, Chuang W, Yang J. Use of ChatGPT on Taiwan's examination for medical doctors. Ann Biomed Eng 2024 Mar 11;52(3):455-457. [doi: 10.1007/s10439-023-03308-9] [Medline: 37432530]
49. Kassab J, Hadi El Hajjar A, Wardrop RM, Brateanu A. Accuracy of online artificial intelligence models in primary care settings. Am J Prev Med 2024 Jun;66(6):1054-1059. [doi: 10.1016/j.amepre.2024.02.006] [Medline: 38354991]
50. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023 Jul 25;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
51. Li S. Exploring the clinical capabilities and limitations of ChatGPT: a cautionary tale for medical applications. Int J Surg 2023 Sep 01;109(9):2865-2867 [FREE Full text] [doi: 10.1097/JS9.0000000000000523] [Medline: 37222684]
52. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc 2024 Sep 01;31(9):1833-1843. [doi: 10.1093/jamia/ocae045] [Medline: 38613821]

Abbreviations

AI: artificial intelligence
GPT: generative pretrained transformer
HIPAA: Health Insurance Portability and Accountability Act

XSL•FO
RenderX

NLP: natural language processing
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 06.08.23; peer-reviewed by Z Wang, X Li, Z Wang, T Ma; comments to author 07.03.24; revised version received 09.04.24; accepted 06.09.24; published 04.10.24.

Please cite as:
Li X, Guo H, Li D, Zheng Y
Engine of Innovation in Hospital Pharmacy: Applications and Reflections of ChatGPT
J Med Internet Res 2024;26:e51635
URL: https://www.jmir.org/2024/1/e51635
doi:10.2196/51635
PMID:39365643

©Xingang Li, Heng Guo, Dandan Li, Yingming Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Applying Implementation Science to Advance Electronic Health Record–Driven Learning Health Systems: Case Studies, Challenges, and Recommendations

Katy E Trinkley[1,2,3,4], PharmD, PhD; Anna M Maw[2,5], MD, MS; Cristina Huebner Torres[6], PhD, MA; Amy G Huebschmann[2,7,8], MD, MSc; Russell E Glasgow[1,2,9], PhD

[1]Department of Family Medicine, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[2]Adult and Child Center for Outcomes Research and Delivery Science Center, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[3]Department of Biomedical Informatics, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[4]Colorado Center for Personalized Medicine, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[5]Division of Hospital Medicine, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[6]Caring Health Center, Springfield, MA, United States
[7]Division of General Internal Medicine, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[8]Ludeman Family Center for Women's Health Research, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[9]VA Eastern Colorado Geriatric Research Education and Clinical Center, Aurora, CO, United States

Corresponding Author:
Katy E Trinkley, PharmD, PhD
Department of Family Medicine
School of Medicine
University of Colorado Anschutz Medical Campus
12631 E 17th Ave
Mail Stop F496
Aurora, CO, 80045
United States
Phone: 1 303 724 3103
Email: katy.trinkley@cuanschutz.edu

Abstract

With the widespread implementation of electronic health records (EHRs), there has been significant progress in developing learning health systems (LHSs) aimed at improving health and health care delivery through rapid and continuous knowledge generation and translation. To support LHSs in achieving these goals, implementation science (IS) and its frameworks are increasingly being leveraged to ensure that LHSs are feasible, rapid, iterative, reliable, reproducible, equitable, and sustainable. However, 6 key challenges limit the application of IS to EHR-driven LHSs: barriers to team science, limited IS experience, data and technology limitations, time and resource constraints, the appropriateness of certain IS approaches, and equity considerations. Using 3 case studies from diverse health settings and 1 IS framework, we illustrate these challenges faced by LHSs and offer solutions to overcome the bottlenecks in applying IS and utilizing EHRs, which often stymie LHS progress. We discuss the lessons learned and provide recommendations for future research and practice, including the need for more guidance on the practical application of IS methods and a renewed emphasis on generating and accessing inclusive data.

(J Med Internet Res 2024;26:e55472) doi:10.2196/55472

KEYWORDS

learning health systems; implementation science; chronic care; electronic health record; evidence-based medicine; information technology; research and technology

XSL•FO
RenderX

Introduction

Nearly all health care settings in the United States use electronic health records (EHRs) to document the continuum of patient care [1,2], a trend that is increasingly evident in other countries as well [3-5]. Driven by federal incentives and regulations, health systems are being progressively encouraged or even mandated to document patient care in structured and actionable ways to facilitate reporting on quality of care metrics. As the breadth and depth of actionable data captured in EHRs expand, health systems are increasingly able to develop learning health systems (LHSs) [6,7]. In these systems, EHR data and other decision support features are utilized to advance the quintuple aim of health care: improving population health, promoting health equity, reducing health care costs, and enhancing both patient and care team experiences [8].

An LHS aims to "align science, informatics, incentives and culture for continuous improvement and innovation, with best practices seamlessly embedded in the health care delivery process in such a way that new knowledge is captured as an integral by-product" [9]. While the concept and the term LHSs are not new, significant progress has recently been made in their development, largely due to the growing capabilities of EHRs that make LHS more feasible. As interest in EHR-driven LHS increases, so does the interest in implementation science (IS), which plays a crucial role in ensuring that LHSs are feasible, rapid, iterative, valid, reliable, reproducible, equitable, and sustainable [10,11]. Both IS and LHSs aim to advance health and health care in ways that are locally relevant and externally valid, with IS providing methods and approaches that can help achieve these goals within LHS [10,12].

"IS is the study of how evidence-based practices are feasibly adopted, implemented, and sustained in real-world settings" [10,12]. A core aspect of IS is the use of its theories, models, and frameworks (TMFs), which are theory-driven approaches for evaluating the context in which LHSs are implemented and for guiding the selection of an implementation evaluation plan for a given LHS learning cycle project. IS and its TMFs draw on methods and theories from various disciplines and evolve over time as they are applied to new perspectives and situations, including different LHSs. A unique strength of IS and its TMFs is their capacity to adapt as new scientific methods emerge, while rigorously translating evidence into practice in ways that are locally relevant, yet generalizable, replicable, sustainable, and equitable [12]. Central to achieving the benefits of IS is the use of its TMFs, which offer a systematic and replicable approach to adapting evidence-based practices to local settings in ways that are both pragmatic and scientifically robust. IS TMFs can be integrated with broader LHS frameworks [13,14] to provide more specific guidance on aligning and evaluating the overarching LHS or learning cycle with the local context in ways that remain generalizable [10]. The adaptability of IS, along with its dual emphasis on locally relevant and externally valid findings, makes it particularly well-suited to support the goals of LHSs in their rapidly changing and inherently complex settings. While the success of LHSs undeniably requires a multidisciplinary approach, with theories and models from other disciplines playing a key role [15], this paper specifically highlights the value of applying IS to advance LHSs [16].

With the increasing use of IS TMFs for LHSs, certain inherent challenges have emerged, partly due to the limitations of EHRs [17,18]. Although EHRs serve as a rich data source, they often lack crucial data needed to effectively apply TMFs, such as key patient-reported outcomes and social determinants of health [19]. Such data are often too complex to collect consistently for most patients but are essential for informing patient-centered care decisions and evaluating the impact of a learning cycle. While EHR constraints challenge the application of IS TMFs, these frameworks can also help address some limitations of EHRs, such as expanding the use of EHR data beyond patient care and billing [19,20]. Nonetheless, there is a lack of guidance and ambiguities in applying IS TMF constructs to learning cycle projects, particularly when using EHR data. Additionally, some perceive TMFs as overly academic, with concerns that operationalizing them is resource-intensive and not practical, intuitive, or sufficiently rapid [18,21].

To advance the goals of LHSs, there is a critical need for guidance and support on pragmatically adapting and applying IS TMFs. This adaptation must align with the varying resources and expectations of different health systems and work within existing constraints, such as competing priorities, resource limitations, and data availability. The purposes of this paper are to (1) discuss the benefits and challenges of applying IS TMFs to EHR-driven LHS learning cycles; (2) outline the key features of a widely used IS TMF, the Practical, Robust Implementation and Sustainability Model (PRISM), as applied to LHS research; (3) provide 3 pragmatic case studies of this application; and (4) explore future directions, challenges, and opportunities for incorporating IS TMFs to support rapid LHSs and address the limitations of EHR data. We offer recommendations and resources for leaders, clinicians, implementers, quality improvement specialists, and researchers involved in developing an LHS or conducting learning cycles. These recommendations focus on how to feasibly adapt IS approaches and methods to various LHS situations.

Methods

Overview of PRISM Applications

In this article, we review 3 case studies illustrating the pragmatic application of an IS TMF and its associated methods, the PRISM, to LHSs. These examples highlight useful and practical applications of PRISM and serve as a basis for discussing challenges and future directions related to the use of IS methods and EHR data.

Although many IS TMFs exist [22,23], we selected PRISM to concretely illustrate the application of a TMF within an LHS. PRISM is an expanded version of the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework, incorporating 4 context domains and the 5 RE-AIM outcome dimensions [17,24-27], as detailed in Tables 1 and 2. PRISM evolved from a blend of frameworks from agriculture (Diffusion of Innovation), engineering (Plan-Do-Study-Act quality improvement cycles), and health care (Chronic Care

XSL•FO
RenderX

Model) [24]. PRISM is designed to be used throughout the life cycle of a study or project to guide the systematic assessment and alignment of the project with its context, aiming to maximize equitable impact on relevant outcomes and sustainability. PRISM underscores the importance of aligning a project with the diverse perspectives and characteristics of various partners, including those who will be directly or indirectly impacted or involved in approving and funding the programs, such as frontline staff, clinicians, clinic or department leaders, system-level leaders, and boards of directors. For effective alignment, it is crucial to ensure the representation of these partners. A key aspect of promoting equity is defining relevant outcomes that are important from diverse perspectives and measuring the representativeness of these outcomes across various demographics (patients) and types of clinics or providers (settings) [27]. PRISM's RE-AIM outcomes facilitate discussions about relevant and meaningful outcome measures at different levels of perspective (eg, leadership, clinician, patient) while emphasizing representativeness and pragmatic issues such as adoption and uptake. PRISM also takes into account the external context (eg, policies, guidelines) and the support or infrastructure available for initial implementation and sustainability (eg, resources, audit, and feedback processes). This consideration enhances the likelihood that the project will continue beyond the study timeline or funding. The systematic approach provided by a TMF like PRISM, combined with the RE-AIM pragmatic outcomes, allows a project to be adapted and scaled up in ways that are locally relevant to different situations and health systems.

To illustrate the challenges and potential solutions of applying IS TMFs to LHS learning cycles and using EHR data, we present retrospective case studies of 3 applications of PRISM to different LHS projects across diverse health care settings. These case studies were selected based on the authors' experiences to represent various types of projects and settings, as well as different approaches to applying PRISM and addressing key issues within LHSs. We focus on identifying challenges and solutions related to (1) operationalizing and applying PRISM within the EHR-embedded LHS context; (2) pragmatically adapting PRISM and its RE-AIM outcomes based on available resources, expectations regarding speed, and data availability; and (3) effectively utilizing EHR data. The solutions provided aim to offer guidance and direction on addressing these challenges and improving the practical application of PRISM and other TMFs and IS methods to EHR-driven LHS.

To facilitate a systematic evaluation of how PRISM was applied in each case study and to minimize recall bias, we first adapted our previously published framework for a fully mature LHS to illustrate where and how PRISM and its associated methods can be applied (Figure 1) [10]. The original framework was designed to be agnostic to any specific IS TMF, so we modified it by overlaying PRISM and highlighting where and how it integrates with the broader LHS framework. In this figure, PRISM guides and informs key aspects and activities of an LHS, including representativeness and equity of perspectives and outcomes; achieving local relevance and external validity or generalizability; rapidity of change and impact; and designing for sustainability. This adapted figure of a fully mature LHS was used to stimulate recall and identify challenges and solutions when applying PRISM and using EHR data. For each case study, we thematically reflected on the challenges and both actual and potential solutions at each phase of the implementation continuum (ie, preimplementation or planning, implementation, and sustainment or evaluation) [28]. This reflection was conducted both inductively to identify new themes and deductively by considering preidentified themes. We also identified aspects of the EHR that pose barriers to applying PRISM and other IS TMFs, which are crucial to address in order to achieve the aspirational goals of a high-functioning LHS.

XSL•FO
RenderX

Table 1. PRISM's[a] contextual domains.

PRISM context domain	Description
Patient and organizational *characteristics*	• The characteristics, priorities, and needs of the setting, including those affected by or involved in the intervention, are crucial to consider when designing the intervention. • LHSs[b] should design learning cycles to align with the priorities and values of the setting.
Patient and organizational *perspectives* of the intervention	• The setting's view of the intervention, including the perspectives of those directly and indirectly affected, influences its uptake and impact. • LHSs should prioritize learning cycles that address local gaps in ways that are both relevant and acceptable to the setting.
Implementation and sustainability infrastructure	• The time, staff, and money required to feasibly implement and maintain the intervention are crucial considerations. This also includes alignment with existing processes, norms, and priorities to promote sustainability. • LHSs must consider the available resources and supporting infrastructure to ensure sustainability and design learning cycles with lasting effects.
External environment	• This includes clinical practice guidelines, policies or regulations, reimbursement mechanisms, and other incentives such as national benchmarking. • LHSs are influenced by contextual factors outside their health setting and consider published literature when prioritizing and designing learning cycles.

[a]PRISM: Practical, Robust Implementation and Sustainability Model.

[b]LHS: learning health system.

Table 2. PRISM's[a] RE-AIM[b] outcome dimensions.

RE-AIM outcome dimension[c]	Description
Reach	• Who was intended to benefit and who participated or was exposed to the intervention? • Representativeness and equity of reach • LHSs[d] should assess what proportion of the target group was impacted and consider the characteristics of those affected to ensure representativeness and equity in reach.
Effectiveness	• What was the most important benefit you were trying to achieve and what were the negative outcomes (eg, safety issues)? • Include quality of life and equity of outcomes. • LHS should evaluate the impact from various perspectives and compare the characteristics of those who were positively and negatively affected.
Adoption	• Where was the intervention applied and who applied it, and who declined? • LHSs should examine how and why uptake or use varies among different users and settings.
Implementation	• How consistently was the intervention delivered? (Fidelity) • How was it adapted? • How much did it cost? • Equity and representativeness (subgroup effects) across implementation outcomes • LHSs should consider how and why adaptations are made based on different contextual influences, as well as the costs associated with initial and ongoing implementation. This information is crucial for making informed decisions about local sustainability and scalability.
Maintenance	• How long was the intervention sustained and how long are the results sustained? • Equity and representativeness of maintenance • LHSs should plan for and assess sustainability and contextual drivers

[a]PRISM: Practical, Robust Implementation and Sustainability Model.

[b]RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance.

[c]Issues of proportion and representativeness of participants compared with all those eligible or those who decline are important and relevant across all RE-AIM dimensions.

[d]LHS: learning health system.

XSL•FO
RenderX

Figure 1. PRISM-adapted conceptual model of a fully mature learning health system NOTES: The orange denotes ways PRISM helps guide and inform key components and activities of a fully mature LHS, including: alignment of a project with the context internal and external to the health system to ensure local relevance and external validity; consideration of inclusive precision health data to promote equity; engagement and integration of diverse partner perspectives to assess and align a project with the context in ways that promote equity and sustainability; iterative assessments and adaptations across all stages of a project (pre-implementation planning, implementation and evaluation/sustainment) to increase speed, effectiveness, and sustainability; use of pragmatic RE-AIM outcomes that consider equity and issues important to partners; and adapting how PRISM is applied based on available timelines and resources.

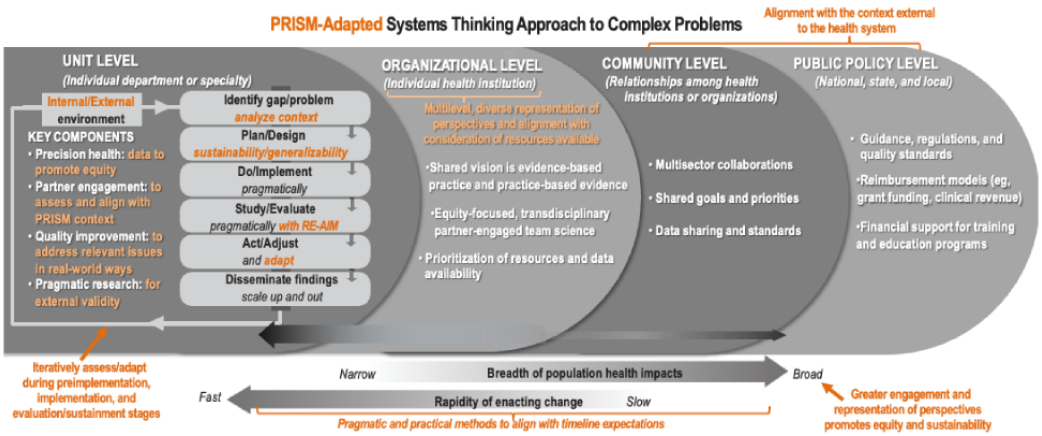

Case Studies

Overview

Below, we provide a brief description of each LHS-based project where PRISM was applied. Table 3 offers an overview of the different features of these case studies.

Table 3. Overview of the key differentiating features of the case studies.

Feature	EHR[a]-embedded CDS[b] for heart failure [29]	EHR-based dashboard for LUS[c] [30]	FQHC[d]-led social needs screening and response equity study
Problem addressed	Gaps in guideline-concordant prescribing of beta-blocker medications for patients with heart failure	Address the heightened need for accurate bedside chest imaging during the COVID-19 pandemic by implementing lung ultrasound among hospitalists.	Define a multipartner implementation plan for social risk screening and response equity to address the CMS[e] mandate and emerging Medicaid ACO[f] equity scores.
Setting	The EHR of 28 primary care clinics in a large integrated health care system	A quaternary care academic medical center	Partnership of multiple FQHCs
Target audience	Primary care clinicians	Hospitalists	Researchers and health center/ACO leaders and staff
Intervention and implementation strategies	A CDS tool was integrated into clinical EHR workflows to alert clinicians and recommend initiating a beta-blocker medication during patient visits.	Hospitalist training in lung ultrasound and the use of iterative PRISM[g] to implement lung ultrasound in the management of patients with COVID-19.	Collaboration across FQHCs to develop a plan using the iPRISM[h] webtool iteratively as a strategy to examine social risk screening and response equity in IS[i] research.
How PRISM was used	For the planning and sustainment/evaluation phases	Iteratively in the planning and implementation phases	Iteratively in the planning phase for cocreation

[a]EHR: electronic health record.

[b]CDS: clinical decision support.

[c]LUS: lung ultrasound.

[d]FQHC: federally qualified health center.

[e]CMS: Centers for Medicare and Medicaid Services.

[f]ACO: accountable care organization.

[g]PRISM: Practical, Robust Implementation and Sustainability Model.

[h]iPRISM: Iterative Practical, Robust Implementation and Sustainability Model.

[i]IS: implementation Science.

XSL•FO
RenderX

Case Study 1: EHR-Embedded Clinical Decision Support for Heart Failure

PRISM was used to design and evaluate a clinical decision support (CDS) tool that recommended prescribing a beta-blocker for primary care clinicians treating patients with heart failure with reduced ejection fraction [29,31]. PRISM guided the systematic assessment of context, which included the following:

- focus groups with patients to understand their treatment preferences and needs;
- iterative, multilevel user-centered design and testing procedures, including focus groups and semistructured interviews with clinicians and clinician leaders; and
- meetings with executive-level informatics and operational leaders, and governance groups.

Based on this iterative partner engagement process, the CDS tool was designed to align with the context while also addressing the technical and data limitations of the EHR. The process also involved identifying pragmatic RE-AIM outcomes that were important and relevant from various perspectives.

A total of 28 primary care clinics were cluster-randomized to either the new, contextually customized CDS tool or an active control group for 6 months. PRISM was used to guide a mixed methods evaluation and to identify adaptations that would promote sustainability and expand the scale of the CDS tool. Specifically, adoption and effectiveness were quantitatively assessed using EHR data. Clinicians were interviewed to determine whether either of the CDS tools should be continued and to understand what changes were needed to optimize adoption and effectiveness within their workflows. Based on the evaluation, the customized CDS tool was found to be more effective, and clinicians expressed a preference for its continued use. As a result, all clinics have since transitioned to the customized CDS tool, and its effectiveness has been sustained [32]. It was also determined that the CDS tool needs to include additional evidence-based medications for heart failure and should be expanded to cardiology clinics. Plans are underway to broaden the tool's scope to include these medications and extend its use to cardiology clinics.

Case Study 2: EHR-Based Dashboard for Lung Ultrasound

In this pilot implementation study, an iterative assessment process was used to ensure high-fidelity and equitable implementation of lung ultrasound for patients hospitalized with COVID-19 [30]. This was achieved through a recurring mixed methods evaluation of prioritized PRISM constructs, including context and progress on outcomes. The evaluation incorporated qualitative interviews and an operational dashboard that displayed prioritized RE-AIM outcomes using real-time EHR data.

Specifically, PRISM was used to assess context by guiding the interview questions posed to multilevel partners during the planning and implementation phases of this study. The questions were designed to identify and characterize barriers to implementation and to uncover potential strategies for overcoming these barriers. The contextual data collected were then used by implementers to inform iterative adaptations to implementation strategies, with the goal of improving PRISM's RE-AIM outcomes. Additionally, prioritized outcomes were iteratively assessed quantitatively using an *operational dashboard* that displayed the representativeness and extent of reach and adoption of lung ultrasound.

The implementation team met every 2 weeks to review outcomes displayed on the dashboard and the qualitative interview data. They assessed barriers to progress in reach and adoption and screened for evidence of emerging disparities in implementation. Based on these assessments, the team collaboratively decided which implementation strategies to deploy, adapt, or discontinue. *This iterative PRISM approach*, combined with an operational dashboard, offered a low-burden method for monitoring progress and disparities in implementation, as well as for making timely, data-driven adjustments to implementation strategies. This approach has since been expanded to support the equitable implementation of additional evidence-based applications of lung ultrasound (eg, management of heart failure) and other point-of-care ultrasound applications within the same health system.

Case Study 3: Social Needs Data Within Federally Qualified Health Centers

In this ongoing project, PRISM is being used to support cross-institutional partnership engagement and the cocreation of an LHS project focused on social risk screening and response equity across multiple federally qualified health centers (FQHCs). Building on long-standing collaborative research by an FQHC in Springfield, Massachusetts, the project brings together partners from a model B Medicaid Accountable Care Organization comprising multiple FQHCs in Massachusetts, collaborators from the Massachusetts Primary Care Association, and research partners from Harvard's Implementation Science Center for Cancer Control Equity. The focus of the project was on mandated universal social risk screening and response requirements for FQHCs [33,34]. The team used PRISM to assess each partner's perspective on anticipated RE-AIM outcomes and to identify contextual issues of importance.

To operationalize PRISM, the team used the Iterative Practical, Robust Implementation and Sustainability Model (iPRISM) webtool [17] during the preimplementation planning phase. The iPRISM webtool includes 21 assessment questions designed to systematically guide individuals or teams through the process of assessing context and anticipated RE-AIM outcomes for projects, while also facilitating shared input or cocreation. The iPRISM webtool was developed to help implementers efficiently apply PRISM to various types of projects and support implementation teams from diverse backgrounds (eg, clinicians/researchers with and without IS expertise). Each partner completed the iPRISM webtool independently (n=6; responses from 1 partner were not linked because they selected a different stage/phase within the webtool form).

During 2 sequential debrief meetings, the partners reviewed their responses, which led to increased clarity on (1) the RE-AIM and PRISM factors with lower scores, highlighting areas for priority focus; and (2) the variation in scoring across different partner perspectives, which contributed to a better understanding of each FQHC and defined multilevel assessment opportunities. Examining and discussing the mean scores and

XSL•FO
RenderX

ranges of responses for each item provided valuable insights into both aspects. The tangible outcomes from this process included the development of a cocreated set of specific aims for the project, a better understanding of each FQHC's perspective and contextual issues, and consensus among the partners to apply for grant funding to support the ongoing work as an innovation in FQHC-led social care research.

Results

Challenges, Solutions, and Future Directions

In each case study, we identified several challenges, potential solutions, and considerations for future research when using EHR data, PRISM, and other IS approaches. The detailed evaluation and findings from each case study are described in Multimedia Appendix 1 and are categorized into 6 overarching themes of challenges with corresponding solutions. These themes and solutions are described below and summarized in Table 4. Across these themes, there are interdependencies, which are expected from a systems science perspective. For instance, resource and time constraints can exacerbate EHR data limitations, and varying levels of IS expertise can complicate the appropriate application or adaptation of PRISM and IS methods. For each challenge, we discuss the issues,potential solutions,and future directions.

XSL•FO
RenderX

Table 4. Challenges, solutions, and future directions.

Theme of challenge	Description of challenge	Potential solutions or ways to mitigate challenges and future directions
Team science issues: *LHS*[a]*and IS*[b]*involve diverse teams with different backgrounds and perspectives*	• Level setting of vocabulary and terminology that is new or has different meanings in different fields • Shared goals and understanding of the problem and project are important • Can be difficult to moderate and understand different perspectives and ensure openness in sharing different perspectives • Consistency in applying TMF[c] to assess context/outcomes across individuals of a team	• Employ team science best practices [35], including the creation of a shared vision and understanding of each partner's perspectives. • Use the iPRISM[d] webtool to facilitate the use of team science best practices; level set vocabulary, issues, and goals; identify and summarize different perspectives; and consistently apply a TMF across partners.
Limited/no IS experience: *LHS teams may have limited IS expertise*	• Makes the application of IS difficult which is a barrier to using these methods that aim to improve relevance, sustainability, scalability, and equity	• Identify external expertise or a consultant • Leverage existing resources and tools that make IS more accessible (eg, iPRISM webtool; see Multimedia Appendix 2 for additional tools/resources) • Create additional resources to increase the accessibility of IS, including guidance on how to (1) feasibly and systematically anticipate, mitigate, and assess for unintended consequences including exacerbation of inequities, and (2) design for sustainability, including identifying and securing resources • Invest in capacity building of IS including training and resources aimed at implementer/practitioner education in addition to implementation scientist training
Data and technology limitations: *LHS and IS methods are limited by the data available*	• LHSs often rely on EHR[e] data and the value of these data is limited by completeness, accuracy, equity, biases of documentation, or is difficult to capture because it is unstructured • Contextual data are often not documented or accessible in structured formats, which often limits contextual assessments to qualitative analyses which can be limited by partial or small samples • Reliance solely on quantitative or qualitative data limits a full picture of the context and impact or outcomes (eg, issues of actual vs stated and depth of understanding) • Accessibility and functionality of software and technology to manage and use data can limit use	• Proactively consider potential data issues and the implications for a given project, and develop workarounds (eg, proxies) or strategies (eg, transparency in reporting) to mitigate the negative impact • Promote better/different data collection practices to ensure high-quality, unbiased, and inclusive data documented in standardized and structured ways • Develop more guidance on how to qualitatively assess context when not accessible via quantitative data sources • Engage diverse partners in the decision of what data are collected and how they are collected • Invest in capacity building of personnel skilled in: a. making timely/relevant data accessible to implementers/system leaders (eg, can build dashboards) b. using advanced analytics such as natural language processing to transform data from unstructured to structured formats • Invest in capacity building of software and technology that is more accessible and tailored to the needs of LHS including: a. transforming unstructured data into structured formats b. analyzing data and images c. intuitively conveying complex data in meaningful visualizations and other ways on demand d. using decision support and other tools that are more precise and able to nimbly embed within existing clinical and EHR workflows e. interoperability or integration across different EHRs and health systems

XSL•FO
RenderX

Theme of challenge	Description of challenge	Potential solutions or ways to mitigate challenges and future directions
Time and resource constraints: *LHS and IS methods must fit within available timelines and resources*	• Health system or other timelines may be fast and constrain LHS projects and IS methods • System-level and implementation team–level resources and time may limit data access; collection of quality, representative, and iterative quantitative and qualitative data to assess context that dynamically changed over time; type, intensity, and frequency of partner engagement and other contextual assessment methodology that can be conducted; and sustained support. • Evaluations of context and outcomes may be limited to a. what is discretely documented unless skilled staff is available to apply advanced analytic approaches (eg, natural language processing) to abstract unstructured data b. availability of skilled staff to query data and generate a report c. staff time to manually collect data • The availability of skilled staff to pragmatically design and evaluate projects can limit internal or external validity	• Avoid the trap of perfect and aim for a minimum viable product when there is no anticipated harm • Collaborate to extend resource availability with: a. trainees who can contribute meaningfully while gaining valuable experience and knowledge b. methodologists who can provide needed skills while gaining authorship opportunities • Apply IS "designing for sustainability" principles to plan for sustainability and develop the supporting infrastructure from the beginning • Leverage advanced analytics (eg, natural language processing, machine learning) and data visualization platforms (eg, dashboards) to automate data collection, analysis, and reporting back to implementers and partners which can: a. make iterative assessments of context and outcomes feasible and sustainable b. increase the speed of positive impact or change when data are acted on c. improve overall efficiency • Invest in capacity building of personnel with the skills that can build low-burden means (eg, dashboards) that allow for evaluation of process and effectiveness outcomes in real time
Appropriateness of certain IS principles, outcomes, and TMF constructs: *IS needs to adapt to each situation*	• Some aspects may not be relevant or perceived as important • Some aspects may not apply or some outcomes may not be addressed because of: a. resource or data constraints (eg, limit iterative assessments of context or adaptations, deter costing analyses) b. expectations around speed (eg, system priorities or patient safety issues require quick action) c. anticipated benefits or needs preclude resource allocation for certain evaluations (eg, cost, unintended consequences) d. difficulty measuring certain outcomes (eg, rare clinical outcomes or time to event, denominators unavailable to assess representativeness) or establishing reasonable causality (eg, unable to control for other influences) • Application of IS can be challenging when the a. intervention is mandated and expected to change but is not evidence based and yet there is a need to (1) design for future and ongoing sustainability and (2) get buy-in from key partners b. evidence for an intervention varies by contextual situation (eg, all CDS[f] interventions are not equal)	• Create best practices and guidance on how/when to adapt IS principles, outcomes, and TMF constructs for diverse projects and situations • Avoid the trap of perfect and work within the constraints of the available timeline, data, and other resources • More guidance is needed on how to apply IS to interventions that are mandated without an established evidence base including how to: a. use an iterative LHS approach to evaluate effectiveness at intervals under different situations and inform intervention adaptations until it is evidence based (ie, effective) b. shift the incentive for partnership and buy-in from the strength of evidence to the requirement to implement c. use a "designing for sustainability" approach to assist in prioritizing resource allocation for a project that is likely to change as the evidence evolves • For interventions with effectiveness that varies by contextual situation, mixed methods evaluations and transparent reporting of contextual factors will provide clarity of the conditions needed for effectiveness
Representation and equity: *LHS and IS need to proactively promote and assess for equity in perspectives and outcomes*	• Partner engagement that does not represent the spectrum of perspectives: a. biases a project toward certain perspectives and priorities b. stymies equity • Incomplete access to accurate data limits the ability to a. evaluate the equitable impact of projects b. assess for unintended consequences such as exacerbation of inequities	

XSL•FO
RenderX

Theme of challenge	Description of challenge	Potential solutions or ways to mitigate challenges and future directions
		• Aim to represent the perspectives of partners engaged at each stage of the project (planning, implementation, and evaluation), rather than just an average perspective across all stages. • Use the iPRISM webtool [36] and other tools to systematically capture different perspectives • Consider less traditional data sources and methods such as crowd-sourcing and social media to expand the representation and inclusion of diverse data types • Proactively consider potential inequities inherent within quantitative or qualitative data sources and any potential unintended consequences • Transparently report perspectives engaged and completeness of data evaluated to guard against lack of diversity in perspectives that inform a project

[a]LHS: learning health system.

[b]IS: implementation science.

[c]TMF: theories, models, and frameworks.

[d]iPRISM: Iterative Practical, Robust Implementation and Sustainability Model.

[e]EHR: electronic health record.

[f]CDS: clinical decision support.

Challenge 1: Team Science

Team science refers to the cross-disciplinary collaboration necessary to address scientific questions and challenges [37]. Both LHS and IS require multilevel engagement from partners with diverse roles, backgrounds, and perspectives, which introduces complexities. Representatives from various partner groups (eg, patients, community members, clinicians, leaders, nonclinical staff, and researchers from different fields) bring unique histories, perspectives, terminologies, biases, assumptions, and knowledge. Sometimes, communication issues or differences in perspectives are recognized, but they can often remain unnoticed for extended periods.

These team science challenges are manageable and should be addressed during the planning phase and throughout the project [35]. Taking the time to understand and respect each partner's perspective and developing a shared vision and vocabulary is essential for team effectiveness and efficiency. It is also important to continually emphasize that different perspectives are not only beneficial but also expected [35]. The use of tools such as the iPRISM webtool [17,36] can systematically capture the diverse perspectives on teams and summarize the mean and distribution of scores, which can be used to focus team discussions on areas with lower mean scores (indicating areas for improvement) or where scores vary and perspectives differ. In the future, greater use and availability of tools such as the iPRISM webtool are recommended to facilitate team science principles within LHS and when using IS, including other TMFs [38,39] beyond PRISM.

> *In the social needs example, the diverse partner team included researchers and executive-level leaders with a range of IS expertise. The iPRISM webtool offered a grounded and shared entry point and opportunity to contextualize each partner's perspective and expertise by providing (1) a guided process assessment, (2) a framework to understand the team's similarities and differences, and (3) a shared language. This resulted in deepening the understanding of the range of potential themes related to social risk screening and response to be addressed and the various contexts in which they need to be considered.*

Challenge 2: Limited or No Implementation Science Experience

In an LHS, and within health systems more generally, there may be variable or no IS expertise, which can preclude the application of IS altogether or lead to inconsistent or incomplete implementation. Inconsistent application may result in replication issues or incomplete assessments of context or project alignment, ultimately affecting project outcomes and sustainability. Understanding how to apply IS principles and TMFs such as PRISM can be challenging without training, and it is often impractical to train all team members.

One potential solution is identifying external IS expertise, which may be feasible through existing consultation services [40-42]. Utilizing resources and tools that make IS more accessible to individuals and teams by simplifying its application can also help. Tools such as the iPRISM webtool [17] and other resources [43-50], which can increase the accessibility of IS, are described in Multimedia Appendix 2. In our case studies, we identified specific aspects of applying PRISM and IS generally that would benefit from greater guidance and tools, such as how to feasibly and systematically anticipate, assess for, and mitigate unintended consequences—including those that can exacerbate or create inequities—and how to design for sustainability, including identifying and securing resources. Capacity-building efforts are also needed to train implementation researchers and practitioners [51].

> *In the social needs example, the iPRISM webtool was used successfully to standardize the application of IS across a team that had variable IS experience, including some members who had no IS experience.*

XSL•FO
RenderX

At the time of the heart failure and lung ultrasound examples, the lead researchers were being mentored through an institutional implementation science K12 program and are now independent implementation scientists supporting others' LHS projects.

Challenge 3: Data and Technology Limitations

Timely and feasible access to complete, accurate, and actionable quantitative and qualitative data is a universal challenge for LHS and IS [6,10,52-54]. These challenges limit outcome evaluations, contextual assessments, and the capacity to rapidly and strategically design interventions and programs that are equitable and optimally fit within workflows [55-58]. Often considered together under the umbrella of "informatics," access to the needed technology is also a challenge. Even when technologies are accessible, they do not always have the functionality needed to seamlessly embed within clinical workflows. Furthermore, there may not be enough skilled staff to use or configure the technology to achieve the goals of delivering the "right" information at the "right" time, to the "right" person, in the "right" format, and through the "right" channel [59].

When faced with data and technology limitations, it is important to first recognize the potential issues and how they might impact a project, and then explore workarounds or strategies to monitor for potential downstream consequences. Workarounds to address data access issues include creating proxies or estimates and conducting sensitivity or subgroup analyses, which may not be precise but can provide valuable insights and aid in understanding. Addressing technology functional limitations often requires adopting a "good enough" mindset, provided the benefits are expected to outweigh potential sacrifices in user experience and no harm is anticipated [60,61]. Other strategies include transparently reporting the data and acknowledging technology limitations, allowing the audience to make informed interpretations of the findings. The current reality may compel LHS projects to operate within existing constraints, but it is crucial to advocate for the goals of truly inclusive precision health [62], which requires comprehensive integration of data to support holistic care decisions. Guidance is needed to help current LHS teams optimize their projects and evaluations within these constraints. However, it is also important to challenge and push against current limitations when necessary. For example, when available data are known to produce biased information or fail to include critically important patient or contextual factors, the health system may need to add essential data elements or utilize novel methods [63,64] to achieve impactful and equitable results [56,62,65,66].

In the lung ultrasound example, reach was trended month to month as a prioritized outcome to monitor the progress of implementation. However, the reach presented as simply the percentage of eligible patients that received lung ultrasounds was somewhat misleading because the denominator, the number of patients hospitalized with COVID-19, fluctuated so greatly with each surge of the pandemic. In this case, it seemed more transparent and less misleading to present reach with the absolute value of the numerator and denominator visible as opposed to just a percentage. The heart failure example illustrates ways to work within the functional constraints of technology that is good enough and proactively considers potential unintended consequences.

Challenge 4: Time and Resource Constraints

In most projects, restrictions on time, availability of skilled personnel, and other resources can slow data access or preclude it entirely. Additionally, the time required from the implementation team and participants can limit the breadth of perspectives engaged and the frequency of qualitative assessments. These limitations can negatively impact the equity and sustainability of projects.

To make progress, LHS teams often need to adapt evaluation plans and partner engagement methods to fit within available time and resource constraints. Finding partners and collaborators who can generate win-win situations and extend resources is also crucial for overcoming these limitations. There is also a growing availability of consultant-type services to support LHS initiatives, offer specific methodological expertise, or connect projects with needed resources [67,68]. Advances in artificial intelligence, such as natural language processing and machine learning, can enhance efficiency and reduce resource usage by automating data collection and analysis. Although these advancements are increasingly present within LHS, improving accessibility to these skills, resources, and software for automation is essential for further enhancing efficiency. Additionally, while the application of artificial intelligence approaches can offer significant benefits, it is important to exercise caution and carefully balance and vet automated processes [69,70]. To further improve efficiency, creating and disseminating "how-to" or implementation guides and recommendations [42,71-73] could help reduce the resources needed to understand, adapt, or apply specific technologies, data, or methodologies.

In the lung ultrasound example, due to the limited time available to the implementers between iterative PRISM cycles, the qualitative interview data collected were not as systematically analyzed as is ideal. In the social needs example, a partner was not able to engage in all iPRISM webtool planning and contextual assessment activities due to time constraints, thus measures were taken to ensure they were able to engage via asynchronous methods, provide summary updates on progress, and gain full team consensus at various points in the process. In the heart failure example, resource availability prevented the ability to iteratively assess for and make adaptations, which may have led to more impactful outcomes.

Challenge 5: Appropriateness of IS Principles, Outcomes, and TMF Constructs

Not all aspects of IS apply to every LHS situation for a variety of reasons. In some cases, a TMF construct or IS method may not align with a project's goals, may not fit within resource or timeline constraints, or may need to be adapted to suit the

XSL•FO
RenderX

situation [74]. Additionally, when implementing interventions without an established evidence base, flexibility is required in applying IS methods and partner engagement strategies. For instance, the LHS may be implementing a clinical guideline recommendation based on poor-quality or low-strength evidence, or following a new regulatory mandate for an intervention that has yet to demonstrate effectiveness [75].

IS is intended to be practical and pragmatic, and its principles, outcomes, and TMF constructs should be considered as a guide and adapted to the situation at hand, focusing on what is feasible and relevant for the context [74]. When changes are made, it is important to document and report the adaptations along with the rationale to facilitate future scalability. For projects without an established evidence base, some common IS strategies, including how and which partners are engaged, may need to be adjusted. Further, when a project's evidence base is uncertain or may vary based on contextual conditions (eg, the effectiveness of a CDS alert varying by clinical situation and design), iterative IS and LHS approaches can be leveraged to develop the evidence base and understand the necessary conditions for success.

To increase the uptake of IS, it is essential to promote awareness that IS should be adapted to fit the specific needs and context of each project [76,77]. Current misconceptions that IS cannot be adapted may inhibit its application. Providing guidance on how to adapt TMFs and IS methods, supported by case examples, can help address these misconceptions and enhance the accessibility and use of IS. Additionally, interactive tools such as the iPRISM webtool, which dynamically guides implementers through the process of adapting IS methods for their specific project, could further facilitate this adaptation.

> *In the heart failure example, the adoption measure was modified from the original definition to be relevant to the situation at hand and to still facilitate the collection of important implementation outcomes.*
>
> *In the social needs example, the intervention was not yet evidence based, which changed aspects of IS partner engagement. Specifically, partner buy-in shifted from a shared understanding of the effectiveness to a shared incentive to meet the mandate with a common interest in contributing to the development of an evidence base. Across all 3 examples, none were able to assess all of PRISM's RE-AIM outcomes or evaluate the cost of implementation due to data and resource constraints as well as a need to focus efforts on those that were mission aligned amidst substantial competing priorities.*

Challenge 6: Representation, Representativeness, and Equity

Limitations in the representativeness of documented data or in the range of partners engaged can impede the ability to design equitable solutions. This may stem from difficulties in assessing equity of outcomes and having a limited number of partners to strategically address existing disparities [27,62,78]. Such limitations could exacerbate or create new inequities without the capability to use data to identify and resolve these issues. IS methods, including PRISM, promote representativeness in data and partner engagement, which may not always be achievable within existing constraints. Therefore, when planning LHS or learning cycles, equity should be clearly defined and prioritized from the outset [79].

To the extent feasible, inclusive use of data and engagement of partners across the spectrum of perspectives—not just the average or majority perspective—is important for promoting equity within LHS [27]. Proactively considering the potential unintended consequences of using different types of data is also key to mitigating inequities. When possible, integrating data sources beyond the EHR (eg, social media, patient and staff satisfaction, community forums, community partner data) and using systems science approaches that include patient-reported outcomes and other social determinants and behavioral data can aid in more comprehensive consideration of the data needed to promote health equity [80]. There is a clear need for health systems to access a more inclusive integration of reliable, structured data.

> *Across all examples, none were able to gain the breadth of partner perspectives that is ideal to sufficiently assess the representativeness of outcomes, but they did what was feasible. For instance, the representativeness (eg, gender, race, age) of clinicians who adopted the CDS or lung ultrasound was not assessed because these data are stored outside of the EHR and inaccessible to those evaluating this type of LHS work.*

Discussion

In many ways, EHRs have enabled the visionary idea of an LHS to become a reality for many health systems. Yet, as highlighted in our case studies, this reliance on EHRs also limits their potential. These case examples demonstrate how IS—when applied in practical, accessible, and adaptable ways—can help LHSs navigate the challenges posed by EHRs while also addressing other crucial factors, such as team science. We highlight aspects of EHR-based LHSs that can complicate the application of IS, notably limitations in the type and completeness of available data. To enable LHS to practically apply IS, our case studies illustrate how an IS framework and its methods (PRISM) can be adapted to drive meaningful change. While IS can sometimes appear overly academic, complex, or inflexible, we emphasize that IS should be tailored to fit specific situations. We encourage others to utilize existing tools and resources to make IS more accessible and practical for their needs.

A cross-cutting key take-home message for LHSs broadly, and particularly when applying IS to EHR-based LHS, is to "do what you can with what you have while proactively anticipating and mitigating unintended consequences and harm" [60,81]. Historically, health care has often pursued perfection, which has led to rigidity in applying IS methods and utilizing EHR data and technology. This mindset can significantly delay or impede progress and is at odds with the visionary goals of LHS, which emphasize practical, relevant, and rapid learning cycles.

XSL•FO
RenderX

Perfection is neither realistic nor attainable, and while striving for a perfect solution, health systems, priorities, and innovations evolve quickly, rendering solutions obsolete before they are even implemented [82,83].

In our case studies, we also identified areas for future development to enhance the accessibility of IS and the utility of EHR data and technology for LHS. First, there is a need for more user-friendly tools and resources to guide the use and adaptation of IS TMFs and methods across various types of projects and situations. Such resources should provide guidance on simplifying the application and adaptation of TMFs, taking into account relevance, data, and resource constraints. They should also address designing for sustainability, equity, and generalizability, including for mandated projects that lack an evidence base. Additionally, these resources should help systematically anticipate and mitigate unanticipated consequences, including those that could potentially misinform future policy. Additionally, we reinforce the decade-long call for more inclusive integration of accessible, high-quality data essential for achieving precision health goals [84,85]. Change is needed to ensure that LHSs have equitable, collaborative, and agreed-upon access to a comprehensive range of data—such as mental, physical, and behavioral health information; social determinants; environmental risks; patient preferences; and genomic data—that drive equitable health care outcomes and are crucial for making informed, patient-centered, and personalized health care decisions [84,85].

Since the original call for LHS in 2007 [7], significant progress has been made, with a growing number of functional LHS [10,86]. EHRs provide foundational infrastructure and data that make LHS possible, and IS methods can help both new and existing LHSs achieve their goals of equitable, sustainable, reproducible, and relevant knowledge generation and translation. However, to foster the growth of new LHSs and support existing LHSs in achieving the aspirational goals of a fully mature, equitable, and sustainable LHS [10], there is a clear need for greater access to inclusive data and more guidance on the practical application of IS methods.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of challenges and solutions to applying PRISM to EHR-based LHS research categorized according to whether the issue primarily stems from PRISM or the EHR.
[DOCX File , 22 KB - jmir_v26i1e55472_app1.docx]

Multimedia Appendix 2
Additional tools and resources for guidance on how to apply IS.
[DOCX File , 16 KB - jmir_v26i1e55472_app2.docx]

References

1. Adler-Milstein J, Jha A. HITECH Act drove large gains in hospital electronic health record adoption. Health Aff (Millwood) 2017 Aug 01;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]
2. National trends in hospital and physician adoption of electronic health records. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records [accessed 2024-09-19]
3. European Commission: Directorate-General for Communications Networks, Content and Technology, Valverde-Albacete J, Folkvord F, Lupiáñez-Villanueva F, Hocking L, Devaux A, et al. Benchmarking deployment of eHealth among general practitioners (2018): final report. Publications Office of the European Union. 2018. URL: https://op.europa.eu/en/publication-detail/-/publication/d1286ce7-5c05-11e9-9c52-01aa75ed71a1/language-en [accessed 2024-09-18]
4. Woldemariam M, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
5. Kim Y, Jung K, Park Y, Shin D, Cho S, Yoon D, et al. Rate of electronic health record adoption in South Korea: a nation-wide survey. Int J Med Inform 2017 May;101:100-107. [doi: 10.1016/j.ijmedinf.2017.02.009] [Medline: 28347440]
6. Etheredge LM. Rapid learning: a breakthrough agenda. Health Aff (Millwood) 2014 Jul;33(7):1155-1162. [doi: 10.1377/hlthaff.2014.0043] [Medline: 25006141]
7. Etheredge LM. A rapid-learning health system. Health Aff (Millwood) 2007 Jan;26(2):w107-w118. [doi: 10.1377/hlthaff.26.2.w107] [Medline: 17259191]
8. Nundy S, Cooper L, Mate K. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 08;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
9. Committee on the Learning Health Care System in America, Institute of Medicine. In: Smith M, Saunders R, Stuckhardt L, McGinnis JM, editors. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC: National Academies Press (US); May 10, 2013.

XSL•FO
RenderX

10. Trinkley K, Ho P, Glasgow R, Huebschmann A. How dissemination and implementation science can contribute to the advancement of learning health systems. Acad Med 2022 Oct 01;97(10):1447-1458 [FREE Full text] [doi: 10.1097/ACM.0000000000004801] [Medline: 35796045]
11. Chambers D, Feero W, Khoury M. Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research. JAMA 2016 May 10;315(18):1941-1942 [FREE Full text] [doi: 10.1001/jama.2016.3867] [Medline: 27163980]
12. Brownson RC, Colditz GA, Proctor EK. Dissemination and Implementation Research in Health: Translating Science to Practice (Second Edition). Oxford: Oxford University Press; Nov 23, 2017.
13. Foley T, Vale L. A framework for understanding, designing, developing and evaluating learning health systems. Learn Health Syst 2023 Jan;7(1):e10315 [FREE Full text] [doi: 10.1002/lrh2.10315] [Medline: 36654802]
14. Allen C, Coleman K, Mettert K, Lewis C, Westbrook E, Lozano P. A roadmap to operationalize and evaluate impact in a learning health system. Learn Health Syst 2021 Oct;5(4):e10258 [FREE Full text] [doi: 10.1002/lrh2.10258] [Medline: 34667878]
15. Nilsen P. In: Nilsen P, editor. Implementation Science: Theory and Application. New York, NY: Routledge; May 13, 2024.
16. Easterling D, Perry AC, Woodside R, Patel T, Gesell SB. Clarifying the concept of a learning health system for healthcare delivery organizations: implications from a qualitative analysis of the scientific literature. Learn Health Syst 2022 Apr;6(2):e10287 [FREE Full text] [doi: 10.1002/lrh2.10287] [Medline: 35434353]
17. Trinkley KE, Glasgow RE, D'Mello S, Fort MP, Ford B, Rabin BA. The iPRISM webtool: an interactive tool to pragmatically guide the iterative use of the Practical, Robust Implementation and Sustainability Model in public health and clinical settings. Implement Sci Commun 2023 Sep 19;4(1):116 [FREE Full text] [doi: 10.1186/s43058-023-00494-4] [Medline: 37726860]
18. Beidas R, Dorsey S, Lewis C, Lyon A, Powell B, Purtle J, et al. Promises and pitfalls in implementation science from the perspective of US-based researchers: learning from a pre-mortem. Implement Sci 2022 Aug 13;17(1):55 [FREE Full text] [doi: 10.1186/s13012-022-01226-3] [Medline: 35964095]
19. Boyd A, Gonzalez-Guarda R, Lawrence K, Patil C, Ezenwa M, O'Brien EC, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. J Am Med Inform Assoc 2023 Aug 18;30(9):1561-1566 [FREE Full text] [doi: 10.1093/jamia/ocad115] [Medline: 37364017]
20. Richesson R, Marsolo K, Douthit B, Staman K, Ho P, Dailey D, et al. Enhancing the use of EHR systems for pragmatic embedded research: lessons from the NIH Health Care Systems Research Collaboratory. J Am Med Inform Assoc 2021 Nov 25;28(12):2626-2640 [FREE Full text] [doi: 10.1093/jamia/ocab202] [Medline: 34597383]
21. Damschroder LJ, Knighton AJ, Griese E, Greene SM, Lozano P, Kilbourne AM, et al. Recommendations for strengthening the role of embedded researchers to accelerate implementation in health systems: findings from a state-of-the-art (SOTA) conference workgroup. Healthc (Amst) 2021 Jun;8 Suppl 1(Suppl 1):100455 [FREE Full text] [doi: 10.1016/j.hjdsi.2020.100455] [Medline: 34175093]
22. Tabak R, Khoong E, Chambers D, Brownson R. Bridging research and practice: models for dissemination and implementation research. Am J Prev Med 2012 Sep;43(3):337-350 [FREE Full text] [doi: 10.1016/j.amepre.2012.05.024] [Medline: 22898128]
23. Rabin B. Dissemination Implementation. URL: https://dissemination-implementation.org/ [accessed 2024-09-19]
24. Feldstein AC, Glasgow RE. A Practical, Robust Implementation and Sustainability Model (PRISM) for integrating research findings into practice. Jt Comm J Qual Patient Saf 2008 Apr;34(4):228-243. [doi: 10.1016/s1553-7250(08)34030-6] [Medline: 18468362]
25. Rabin B, Cakici J, Golden C, Estabrooks P, Glasgow R, Gaglio B. A citation analysis and scoping systematic review of the operationalization of the Practical, Robust Implementation and Sustainability Model (PRISM). Implement Sci 2022 Sep 24;17(1):62 [FREE Full text] [doi: 10.1186/s13012-022-01234-3] [Medline: 36153628]
26. Glasgow R, Harden S, Gaglio B, Rabin B, Smith M, Porter G, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
27. Fort MP, Manson SM, Glasgow RE. Applying an equity lens to assess context and implementation in public health and health services research and practice using the PRISM framework. Front Health Serv 2023;3:1139788 [FREE Full text] [doi: 10.3389/frhs.2023.1139788] [Medline: 37125222]
28. Goodrich D, Miake-Lye I, Braganza M, Wawrin N, Kilbourne A. The QUERI Roadmap for Implementation and Quality Improvement. Washington, DC: Department of Veterans Affairs (US); 2020.
29. Trinkley K, Kroehl M, Kahn M, Allen L, Bennett T, Hale G, et al. Applying clinical decision support design best practices with the practical robust implementation and sustainability model versus reliance on commercially available clinical decision support tools: randomized controlled trial. JMIR Med Inform 2021 Mar 22;9(3):e24359 [FREE Full text] [doi: 10.2196/24359] [Medline: 33749610]
30. Maw A, Morris M, Glasgow R, Barnard J, Ho P, Ortiz-Lopez C, et al. Using Iterative RE-AIM to enhance hospitalist adoption of lung ultrasound in the management of patients with COVID-19: an implementation pilot study. Implement Sci Commun 2022 Aug 12;3(1):89. [doi: 10.1186/s43058-022-00334-x] [Medline: 35962441]

XSL•FO
RenderX

31. Trinkley K, Kahn M, Bennett T, Glasgow R, Haugen H, Kao D, et al. Integrating the Practical Robust Implementation and Sustainability Model with best practices in clinical decision support design: implementation science approach. J Med Internet Res 2020 Oct 29;22(10):e19676 [FREE Full text] [doi: 10.2196/19676] [Medline: 33118943]
32. Trinkley KE, Wright G, Allen LA, Bennett TD, Glasgow RE, Hale G, et al. Sustained effect of clinical decision support for heart failure: a natural experiment using implementation science. Appl Clin Inform 2023 Oct;14(5):822-832. [doi: 10.1055/s-0043-1775566] [Medline: 37852249]
33. Cole M, Nguyen K, Byhoff E, Murray G. Screening for social risk at federally qualified health centers: a national study. Am J Prev Med 2022 May;62(5):670-678 [FREE Full text] [doi: 10.1016/j.amepre.2021.11.008] [Medline: 35459451]
34. Billioux A, Verlander K, Anthony S, Alley D. Standardized screening for health-related social needs in clinical settings: the accountable health communities screening tool. NAM Perspectives 2017 May 30;7(5):1-9. [doi: 10.31478/201705b]
35. Patient-Centered Outcomes Research Institute (PCORI). Best practices in multi-stakeholder team science. PCORI. URL: https://research-teams.pcori.org/best-practices [accessed 2024-09-19]
36. iPRISM Webtool. URL: https://prismtool.org/ [accessed 2024-09-26]
37. Little MM, St Hill CA, Ware KB, Swanoski MT, Chapman SA, Lutfiyya MN, et al. Team science as interprofessional collaborative research practice: a systematic review of the science of team science literature. J Investig Med 2017 Jan;65(1):15-22 [FREE Full text] [doi: 10.1136/jim-2016-000216] [Medline: 27619555]
38. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
39. Moullin J, Dickson K, Stadnick N, Rabin B, Aarons G. Systematic review of the Exploration, Preparation, Implementation, Sustainment (EPIS) framework. Implement Sci 2019 Jan 05;14(1):1 [FREE Full text] [doi: 10.1186/s13012-018-0842-6] [Medline: 30611302]
40. University of Colorado Anschutz Medical Campus. URL: https://medschool.cuanschutz.edu/accords/cores-and-programs/dissemination-implementation-science-program [accessed 2024-09-18]
41. Dissemination and Implementation Science Center (DISC). Altman Clinical and Translational Research Institute UC San Diego. URL: https://actri.ucsd.edu/centers-services/portfolio/disc/index.html [accessed 2024-09-18]
42. Resources: Implementation Science and Practice. The Center for Implementation. URL: https://thecenterforimplementation.com/toolbox [accessed 2024-09-18]
43. Coyle K, Carcone A, Butame S, Pooler-Burgess M, Chang J, Naar S. Adapting the self-assessment of contextual fit scale for implementation of evidence-based practices in adolescent HIV settings. Implement Sci Commun 2022 Oct 22;3(1):115 [FREE Full text] [doi: 10.1186/s43058-022-00349-4] [Medline: 36273221]
44. Robinson C, Damschroder L. A pragmatic context assessment tool (pCAT): using a think aloud method to develop an assessment of contextual barriers to change. Implement Sci Commun 2023 Jan 11;4(1):3 [FREE Full text] [doi: 10.1186/s43058-022-00380-5] [Medline: 36631914]
45. Palinkas L, Spear S, Mendon S, Villamar J, Reynolds C, Green C, et al. Conceptualizing and measuring sustainability of prevention programs, policies, and practices. Transl Behav Med 2020 Feb 03;10(1):136-145 [FREE Full text] [doi: 10.1093/tbm/ibz170] [Medline: 31764968]
46. Malone S, Prewitt K, Hackett R, Lin J, McKay V, Walsh-Bailey C, et al. The Clinical Sustainability Assessment Tool: measuring organizational capacity to promote sustainability in healthcare. Implement Sci Commun 2021 Jul 17;2(1):77 [FREE Full text] [doi: 10.1186/s43058-021-00181-2] [Medline: 34274004]
47. Schell S, Luke D, Schooley M, Elliott M, Herbers S, Mueller N, et al. Public health program capacity for sustainability: a new framework. Implement Sci 2013 Feb 01;8:15 [FREE Full text] [doi: 10.1186/1748-5908-8-15] [Medline: 23375082]
48. The Hexagon – an exploration tool. National Implementation Research Network. URL: https://implementation.fpg.unc.edu/resource/the-hexagon-an-exploration-tool/ [accessed 2024-09-18]
49. Hull L, Goulding L, Khadjesari Z, Davis R, Healey A, Bakolis I, et al. Designing high-quality implementation research: development, application, feasibility and preliminary evaluation of the implementation science research development (ImpRes) tool and guide. Implement Sci 2019 Aug 14;14(1):80 [FREE Full text] [doi: 10.1186/s13012-019-0897-z] [Medline: 31412887]
50. Weiner B, Sherr K, Lewis CC, editors. Practical Implementation Science. New York City, NY: Springer Publishing; Mar 28, 2022.
51. Huebschmann AG, Johnston S, Davis R, Kwan BM, Geng E, Haire-Joshu D, et al. Promoting rigor and sustainment in implementation science capacity building programs: a multi-method study. Implement Res Pract 2022;3:26334895221146261 [FREE Full text] [doi: 10.1177/26334895221146261] [Medline: 37091073]
52. Borsky AE, Savitz LA, Bindman AB, Mossburg S, Thompson L. AHRQ series on improving translation of evidence: perceived value of translational products by the AHRQ EPC learning health systems panel. Jt Comm J Qual Patient Saf 2019 Nov;45(11):772-778 [FREE Full text] [doi: 10.1016/j.jcjq.2019.08.002] [Medline: 31668330]
53. Slutsky JR. Moving closer to a rapid-learning health care system. Health Aff (Millwood) 2007;26(2):w122-w124. [doi: 10.1377/hlthaff.26.2.w122] [Medline: 17259193]

XSL•FO
RenderX

54. Gould M, Sharp A, Nguyen H, Hahn E, Mittman B, Shen E, et al. Embedded research in the learning healthcare system: ongoing challenges and recommendations for researchers, clinicians, and health system leaders. J Gen Intern Med 2020 Dec;35(12):3675-3680 [FREE Full text] [doi: 10.1007/s11606-020-05865-4] [Medline: 32472492]
55. Wu AW, Snyder C, Clancy CM, Steinwachs DM. Adding the patient perspective to comparative effectiveness research. Health Aff (Millwood) 2010 Oct;29(10):1863-1871. [doi: 10.1377/hlthaff.2010.0660] [Medline: 20921487]
56. Glasgow RE, Kaplan RM, Ockene JK, Fisher EB, Emmons KM. Patient-reported measures of psychosocial issues and health behavior should be added to electronic health records. Health Aff (Millwood) 2012 Mar;31(3):497-504. [doi: 10.1377/hlthaff.2010.1295] [Medline: 22392660]
57. Nerenz DR, Austin JM, Deutscher D, Maddox KEJ, Nuccio EJ, Teigland C, et al. Adjusting quality measures for social risk factors can promote equity in health care. Health Aff (Millwood) 2021 Apr;40(4):637-644. [doi: 10.1377/hlthaff.2020.01764] [Medline: 33819097]
58. Lavallee DC, Chenok KE, Love RM, Petersen C, Holve E, Segal CD, et al. Incorporating patient-reported outcomes into health care to engage patients and enhance care. Health Aff (Millwood) 2016 Apr;35(4):575-582. [doi: 10.1377/hlthaff.2015.1362] [Medline: 27044954]
59. Osheroff J, Teich J, Levick D, Saldana L, Velasco F, Sittig D. Improving Outcomes With Clinical Decision Support: An Implementers Guide (Second Edition). Chicago, IL: Healthcare Information Management Systems Society (HIMSS); 2012.
60. Tolf S, Nyström ME, Tishelman C, Brommels M, Hansson J. Agile, a guiding principle for health care improvement? Int J Health Care Qual Assur 2015;28(5):468-493. [doi: 10.1108/IJHCQA-04-2014-0044] [Medline: 26020429]
61. Krebs K, Milani L. Translating pharmacogenomics into clinical decisions: do not let the perfect be the enemy of the good. Hum Genomics 2019 Aug 27;13(1):39 [FREE Full text] [doi: 10.1186/s40246-019-0229-z] [Medline: 31455423]
62. Glasgow R, Kwan B, Matlock D. Realizing the full potential of precision health: the need to include patient-reported health behavior, mental health, social determinants, and patient preferences data. J Clin Transl Sci 2018 Jun;2(3):183-185 [FREE Full text] [doi: 10.1017/cts.2018.31] [Medline: 30370072]
63. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? EGEMS (Wash DC) 2016;4(1):1203 [FREE Full text] [doi: 10.13063/2327-9214.1203] [Medline: 27668265]
64. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Netw Open 2021 Feb 01;4(2):e210184 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0184] [Medline: 33635321]
65. Nelson TA, Anderson B, Bian J, Boyd AD, Burton SV, Davis K, et al. Planning for patient-reported outcome implementation: development of decision tools and practical experience across four clinics. J Clin Transl Sci 2020 Apr 06;4(6):498-507 [FREE Full text] [doi: 10.1017/cts.2020.37] [Medline: 33948226]
66. Glasgow R, Huebschmann A, Krist A, Degruy FV. An adaptive, contextual, technology-aided support (ACTS) system for chronic illness self-management. Milbank Q 2019 Sep;97(3):669-691 [FREE Full text] [doi: 10.1111/1468-0009.12412] [Medline: 31424137]
67. McDonald PL, Van Der Wees P, Weaver GC, Harwood K, Phillips JR, Corcoran M. Learning health systems from an academic perspective: establishing a collaboratory within a school of medicine and health sciences. Med Educ Online 2021 Dec;26(1):1917038. [doi: 10.1080/10872981.2021.1917038] [Medline: 33876701]
68. Learning health sciences. Department of Learning Health Sciences, University of Michigan Medical School. URL: https://medicine.umich.edu/dept/learning-health-sciences [accessed 2024-09-18]
69. Feehan M, Owen L, McKinnon I, DeAngelis MM. Artificial intelligence, heuristic biases, and the optimization of health outcomes: cautionary optimism. J Clin Med 2021 Nov 14;10(22):5284 [FREE Full text] [doi: 10.3390/jcm10225284] [Medline: 34830566]
70. Cabitza F, Rasoini R, Gensini G. Unintended consequences of machine learning in medicine. JAMA 2017 Aug 08;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
71. Soong C, Burry L, Cho HJ, Gathecha E, Kisuule F, Tannenbaum C, et al. An implementation guide to promote sleep and reduce sedative-hypnotic initiation for noncritically ill inpatients. JAMA Intern Med 2019 Jul 01;179(7):965-972. [doi: 10.1001/jamainternmed.2019.1196] [Medline: 31157831]
72. Trecartin KW, Wolfe RE. Emergency department observation implementation guide. J Am Coll Emerg Physicians Open 2023 Aug;4(4):e13013 [FREE Full text] [doi: 10.1002/emp2.13013] [Medline: 37520081]
73. Draffan E, Danger C, Banes D. Reflections on building a multi-country AAC implementation guide. Stud Health Technol Inform 2023 Aug 23;306:181-187. [doi: 10.3233/SHTI230616] [Medline: 37638914]
74. Glasgow R, Estabrooks P. Pragmatic applications of RE-AIM for health care initiatives in community and clinical settings. Prev Chronic Dis 2018 Jan 04;15:E02 [FREE Full text] [doi: 10.5888/pcd15.170271] [Medline: 29300695]
75. Brownson RC, Shelton RC, Geng EH, Glasgow RE. Revisiting concepts of evidence in implementation science. Implement Sci 2022 Apr 12;17(1):26 [FREE Full text] [doi: 10.1186/s13012-022-01201-y] [Medline: 35413917]
76. Quinn AK, Neta G, Sturke R, Olopade CO, Pollard SL, Sherr K, et al. Adapting and operationalizing the RE-AIM framework for implementation science in environmental health: clean fuel cooking programs in low resource countries. Front Public Health 2019;7:389 [FREE Full text] [doi: 10.3389/fpubh.2019.00389] [Medline: 31921753]

XSL•FO
RenderX

77. Soicher RN, Becker-Blease KA, Bostwick KCP. Adapting implementation science for higher education research: the systematic study of implementing evidence-based practices in college classrooms. Cogn Res Princ Implic 2020 Nov 05;5(1):54 [FREE Full text] [doi: 10.1186/s41235-020-00255-0] [Medline: 33152095]
78. Shelton R, Chambers D, Glasgow R. An extension of RE-AIM to enhance sustainability: addressing dynamic context and promoting health equity over time. Front Public Health 2020;8:134 [FREE Full text] [doi: 10.3389/fpubh.2020.00134] [Medline: 32478025]
79. Lee-Foon NK, Reid RJ, Brown A. Fairness for whom? Learning health systems' approach to equity in healthcare. Healthc Policy 2023 Nov;19(2):15-20 [FREE Full text] [doi: 10.12927/hcpol.2023.27237] [Medline: 38105663]
80. Bierman AS, Mistry KB. Commentary: Achieving health equity - the role of learning health systems. Healthc Policy 2023 Nov;19(2):21-27 [FREE Full text] [doi: 10.12927/hcpol.2023.27236] [Medline: 38105664]
81. Palakshappa D, Miller D, Rosenthal G. Advancing the learning health system by incorporating social determinants. Am J Manag Care 2020 Jan 01;26(1):e4-e6 [FREE Full text] [doi: 10.37765/ajmc.2020.42146] [Medline: 31951360]
82. MGowan H, Shipley C. The Adaptation Advantage: Let Go, Learn Fast, and Thrive in the Future of Work. Hoboken, NJ: John Wiley and Sons Inc; 2020.
83. Chambers D, Norton W. The adaptome: advancing the science of intervention adaptation. Am J Prev Med 2016 Oct;51(4 Suppl 2):S124-S131 [FREE Full text] [doi: 10.1016/j.amepre.2016.05.011] [Medline: 27371105]
84. Hood L, Friend S. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011 Mar;8(3):184-187. [doi: 10.1038/nrclinonc.2010.227] [Medline: 21364692]
85. Khoury M, Gwinn M, Glasgow R, Kramer B. A population approach to precision medicine. Am J Prev Med 2012 Jun;42(6):639-645 [FREE Full text] [doi: 10.1016/j.amepre.2012.02.012] [Medline: 22608383]
86. Lannon C, Schuler C, Seid M, Provost L, Fuller S, Purcell D, et al. A maturity grid assessment tool for learning networks. Learn Health Syst 2021 Apr;5(2):e10232 [FREE Full text] [doi: 10.1002/lrh2.10232] [Medline: 33889737]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
FQHC: federally qualified health center
iPRISM: Iterative Practical, Robust Implementation and Sustainability Model
IS: implementation science
LHS: learning health system
PRISM: Practical, Robust Implementation and Sustainability Model
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
TMF: theories, models, and frameworks

Edited by G Tsafnat; submitted 13.12.23; peer-reviewed by S Greene, T Watterson; comments to author 28.02.24; revised version received 17.05.24; accepted 24.08.24; published 07.10.24.

Please cite as:
Trinkley KE, Maw AM, Torres CH, Huebschmann AG, Glasgow RE
Applying Implementation Science to Advance Electronic Health Record–Driven Learning Health Systems: Case Studies, Challenges, and Recommendations
J Med Internet Res 2024;26:e55472
URL: https://www.jmir.org/2024/1/e55472
doi:10.2196/55472
PMID:39374069

©Katy E Trinkley, Anna M Maw, Cristina Huebner Torres, Amy G Huebschmann, Russell E Glasgow. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Using Social Media to Recruit Participants in Health Care Research: Case Study

Amy L Wright[1], MScN, NP-Pediatrics, NCC-BC, CNeoN(C), PhD; Ysabella Jayne Willett[1], BKin; Era Mae Ferron[1], BScN, MN, PhD; Vithusa Kumarasamy[1], BScN, MN; Sarah M Lem[1], BSc, BScN; Ossaid Ahmed[1], BHSc

Lawrence Bloomberg Faculty of Nursing, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Amy L Wright, MScN, NP-Pediatrics, NCC-BC, CNeoN(C), PhD
Lawrence Bloomberg Faculty of Nursing
University of Toronto
155 College Street
Toronto, ON, M5T 1P8
Canada
Phone: 1 4169468183
Email: amyl.wright@utoronto.ca

Abstract

This paper presents a case study describing the use of social media, specifically Facebook and Instagram, as a valuable tool for recruiting participants in community-engaged health care studies. Drawing on the experiences of our team during a qualitative study aiming to understand the needs of Indigenous fathers and Two-Spirit parents as they transition to parenthood, we offer an in-depth exploration of our social media recruitment strategy. This strategy encompasses deliberate content creation and online engagement with local Indigenous community organizations and people. Through the implementation of this recruitment strategy, we successfully recruited 18 Indigenous fathers and 4 Two-Spirit parents to our community-engaged project. We learned that social media can be used to enhance recruitment by building community trust, engagement, tailored content for specific audiences, and adaptive strategies guided by data metrics provided by social media platforms. Our journey included several challenges, such as dealing with fraudulent participants, navigating budget and resource constraints, and facing recruitment limitations, which we also describe in detail. Our paper provides essential insights for researchers considering the use of social media as a recruitment tool but we are unsure of how to begin. Health care researchers may find our experience and recommendations helpful for developing and implementing their own effective social media recruitment strategy. Meanwhile, sharing our experience contributes to the broader understanding of the role of social media in participant recruitment.

(J Med Internet Res 2024;26:e51751) doi:10.2196/51751

KEYWORDS

social media; qualitative methods; recruitment strategies; healthcare research; digital health; internet;

Introduction

Social media refers to online platforms and technologies that facilitate the creation, sharing, and exchange of user-generated content and enables interactions and connections between individuals and groups in an online space [1]. These platforms often support various forms of content, including text, images, videos, and links, allowing users to communicate, share information, and build online communities. Social media, with its current estimate of 3.8 billion active users worldwide [1,2], has transformed the way researchers connect with potential study participants. In the era of digital connectivity, social media stands as a powerful tool, offering an opportunity for researchers to harness this digital landscape for participant recruitment beyond traditional communication methods, such as word-of-mouth, posters, newspaper advertisements, and direct mailing [1,3]. Facebook, Instagram, and X (formerly Twitter) stand as prominent social media platforms, collectively hosting billions of monthly users, and have gained popularity among researchers for participant recruitment [4-6]. Facebook allows users to create individual profiles and organizational pages to foster community building around common interests [7]. In contrast, Instagram offers a visual communication stream, with a primary focus on photos, graphics, and videos [8], catering to younger users with a preference for visual content [9]. Like Facebook and Instagram, X allows users to share and discover short text messages, as well as various multimedia content, including photos, videos, and live streams [10]. In this paper, we focus our discussion on the 2 social media platforms that

XSL•FO
RenderX

we used in our study: Facebook and Instagram. We chose not to use X due to limits on time and the funding necessary to implement our social media recruitment strategy across more than 2 platforms.

Benefits of Using Social Media for Recruitment

The versatility of social media has positioned it as a valuable resource for researchers to disseminate study information to the public, generate interest, and enhance recruitment efforts [11-13]. The role of social media in health care research has grown considerably over the years, serving as a platform to improve research literacy and individuals' understanding of ongoing research studies [9,11]. Using social media for participant recruitment in health care–related studies has several benefits. First, it presents a cost-effective approach to recruitment [13,14], making it particularly appealing to researchers with limited budgets. Second, social media offers extensive accessibility, reaching a diverse global audience using various electronic devices [4,13,15], and allowing for anonymous interaction by users [4,16,17]. It has been found to be highly effective in connecting with adolescents and young adults, who are active users [9]. Third, it aids in engaging individuals who have historically been difficult to recruit due to stigmatization, social inequities, low disease prevalence, or disabilities [4,9,16,18-21]. Facebook, in particular, has been recognized for its success in reaching underrepresented communities, allowing researchers to connect with individuals based on demographics, geographical location, interests, and behaviors [3,5,22]. Lastly, social media platforms can support researchers in reaching their desired sample sizes, a crucial advantage for studies requiring large participant populations [3].

Despite these benefits, there are some challenges worth noting. First, using social media for recruitment requires users to have electronic devices with access to the internet, so recruitment using only social media may exclude those without this access [4,5]. Second, there is an increased risk of fraudulent participants, threatening data integrity [16,17,19-21,23]. This concern will be outlined later in detail.

Our paper presents insights gained from using Facebook and Instagram to recruit participants for phase 1 of a 3-phased community-engaged research study, Fathers of the Next Generation (FOTNG). By using this example as a case study with which to illustrate our use of social media for participant recruitment and its associated risks and benefits, we aim to provide researchers with key insights to consider to effectively develop and implement this type of recruitment strategy. For definitions of common social media terminology used in Facebook and Instagram, please see Multimedia Appendix 1.

Case Study: FOTNG

The FOTNG is a 3-year, 3-phase project, funded by the Canadian Institutes of Health Research. The project focuses on understanding the unique needs of self-identifying First Nations, Métis, and Inuit fathers and Two-Spirit parents during the transition to parenthood to build a community-derived parenting program that addresses their needs. Recognizing the diversity of Indigenous Peoples, we will use the term Indigenous to refer to more than one cultural group, and use the specific culture or Nation whenever possible. This study received ethics approval from the Six Nations Council Research Ethics Committee and the University of Toronto Research Ethics Committee. The project is led by a steering committee of representatives from Indigenous-led organizations within Southern Ontario, Canada, Indigenous service providers, parents, and Indigenous and non-Indigenous researchers of settler-European ancestry. The first phase of the project involves a qualitative study, which commenced participant recruitment in September 2022 and will be the focus of our discussion here. Phase 1 included one-on-one interviews with self-identifying Indigenous fathers, Two-Spirit parents, and service providers. Parent participants were eligible to participate if they identified as an Indigenous father (or father-to-be), or Two-Spirit parent (or parent-to-be), and had a child aged younger than 3 years or were expecting a baby, and lived in Southern Ontario, Canada. Service providers were eligible to participate if they provided parenting programs or services to Indigenous fathers or Two-Spirit parents in the same region. At the time of publication, we have recruited 18 Indigenous fathers, 4 Two-Spirit parents, and 23 service providers, and recruitment for Two-Spirit parents is ongoing. Although this phase also included the recruitment of service providers, we will focus our case study on the recruitment strategy we used for fathers and Two-Spirit parents, as service providers were primarily recruited through word-of-mouth by our community partners and steering committee.

Our recruitment strategy was devised by the lead author and steering committee and supported by a project coordinator and 2 research assistants, and included numerous strategies including traditional approaches: word-of-mouth, distributing digital and hard copies of flyers, posting flyers and other information on the project website, attending community events, in addition to a comprehensive online social media communication strategy. As our project is community-led by Indigenous organizations and community members, we collectively agreed it was important to use our social media platforms to not only recruit participants to participate in our project but to also build awareness among Indigenous and non-Indigenous people about Indigenous cultures and priorities, inform the public about our study and ongoing activities and findings, and, most importantly, facilitate trust and building relationships with Indigenous community members and organizations.

Social Media Communication Strategy: Initial Steps

Any researcher seeking to use social media to recruit participants must first have a social media account. After creating an account for our project, we established our presence on both Facebook and Instagram by following the accounts of Indigenous organizations, relevant health and wellness groups, Indigenous activists, educators, and other established Indigenous community members. Following these accounts helped us to elicit interest in our account through other users by sharing their content on our page. Before we began to create our own content, we developed a logo and short project name (FOTNG) to support our study's *brand recognition.* We developed a project logo with support from a graphic designer and the steering committee. The project logo and its associated color palette provided a standardized and appealing visual that was crucial for

XSL•FO
RenderX

establishing our project's recognition within our social media account profiles and across posts [22].

Content Creation

Our content strategy was guided by three objectives: (1) to create awareness about Indigenous wellness, including information relevant to this study (parenting), family health and wellness, and Indigenous holidays and observances; (2) to recruit Indigenous participants who met our eligibility criteria; and (3) to communicate study activities and findings. We posted our recruitment flyer, which contained the purpose of this study, eligibility requirements, incentive information, and research coordinator contact information (see Figure 1). We used single images (Figure 2 [24]), carousels (several images in a single post; Figure 3 [25]), graphics, and reels or videos to share information. We reshared content multiple times to reach new people [26], and to foster engagement and support among our followers. Initially, we posted content about the project, such as this study's purpose, objectives, and phases, and once we covered relevant project details, we shifted our focus to promoting participant recruitment and celebrating and promoting health and wellness among parents, young children, families, and communities.

To ensure our content was informative and visually appealing, our team carefully conducted thorough research and sourced credible information, and included graphics found using the platform Canva. Captions were used to summarize information not depicted in the visual content and typically included a summary, contact details, a link to the FOTNG website, and relevant hashtags [11,28]. Hashtags are links to other relevant social media topics (eg, #FOTNG), and by adding these to our posts, we hoped to promote the engagement of the followers interested in those topics with our content [29]. Finally, we were deliberate in our choice of language, aiming to maximize end user engagement by keeping content simple, concise, and at an 8th-grade reading level or a Gunning Fog Index score of less than eight [30,31].

Further, 2 research assistants managed content creation and posting to the FOTNG social media accounts. They met monthly to brainstorm ideas and adjust their approach based on social media data metrics provided through the platforms. For instance, we initially posted a 58-second recruitment video which reached 123 user accounts on Facebook. To increase the number of people viewing our recruitment video content, we condensed the video to a 21-second reel, which increased our engagement 12-fold, reaching 1628 user accounts. Our experience is supported by the literature indicating that shorter videos (less than 60 seconds) are more engaging [32,33], also reflecting end user preference for quick access to information in today's fast-paced world [34].

Figure 1. Recruitment flyer for Indigenous fathers. Note: the name and phone number of the research coordinator have been removed due to privacy considerations. Text, images, and graphics were created using Canva.

XSL•FO
RenderX

Figure 2. Single image post commemorating Indigenous Peoples Day [24]. The right side of the figure displays the accompanying caption and indication of the number of likes and comments. Text, images, and graphics were created using Canva.

Figure 3. Carousel post for Treaties Recognition Week [25], an important Indigenous event celebrated every November in the province of Ontario, Canada, and accompanying caption. Text, images, and graphics were created using Canva.

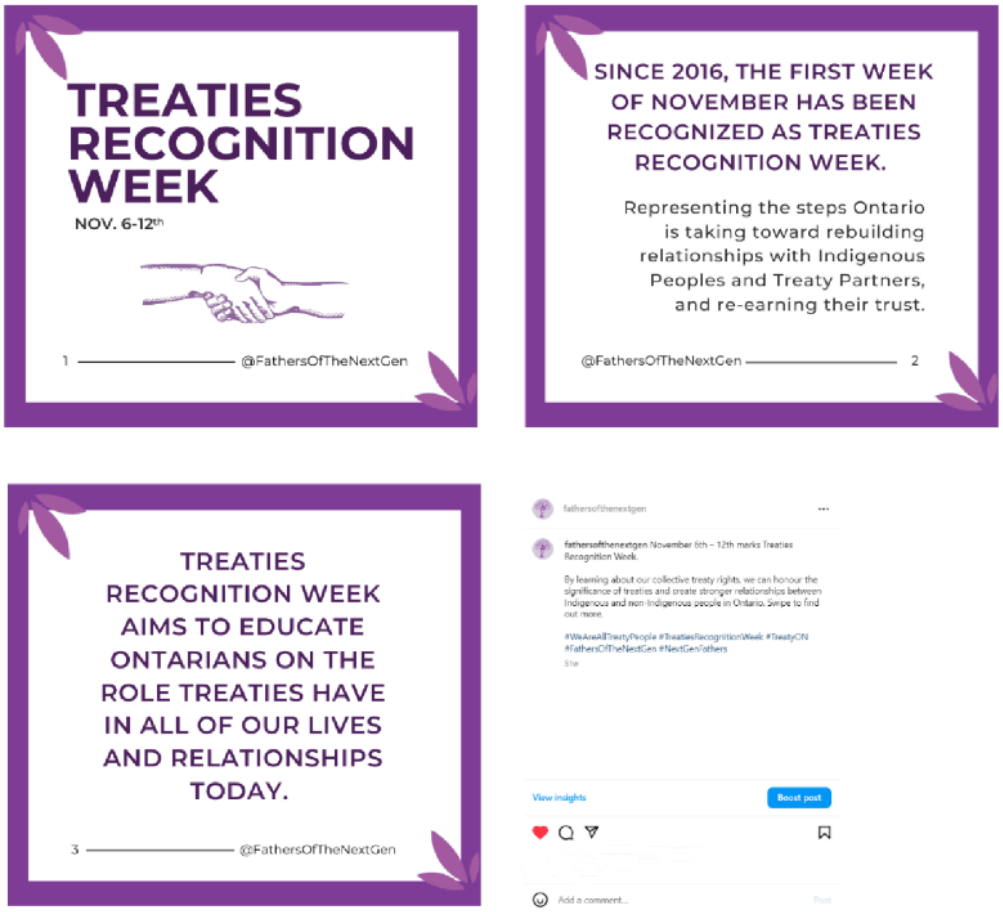

Posting Timing and Frequency

Before we were granted access to special data metrics detailing network growth, top locations and age range of followers, and times of most activity on each platform, we experimented with posting at various times of day and frequencies, to identify audience engagement trends. Once our Instagram page reached over 100 followers, we were provided access to special insights and data metric reports via the platform, which we used to adjust our posting schedule to align with peak user activity on Instagram, primarily between 12 PM and 6 PM EST, on Tuesdays and Fridays. Additional posts occurred every other week as part of our monthly social media plan. For Facebook, our follower count remained below the 100 users required to

XSL•FO
RenderX

gain access to special data metrics, so we were unable to use these data to inform our approach. Instead, we mirrored our Facebook posting schedule to match our Instagram metrics, using the same peak times and preferred days of the week [12,35].

Notably, posts of our recruitment flyer generated the highest engagement concerning views, shares, impressions, and accounts reached compared to other posted content. This aligns with the experiences of Luarn et al [36], who found that posts advertising tangible benefits (such as compensation and honoraria) were more engaging than posts without. Additionally, we observed increased engagement on both Facebook and Instagram when our content was posted more frequently. For example, months with an average of 3 to 5 posts per week garnered greater impressions and engagement than those with only 2 to 3 posts per week. Recognizing that posting too frequently can diminish end user interest and engagement, we sought to balance our posting schedule to maximize interaction [37].

Connecting With Respected Community Members and Organizations

Overview

We prioritized building relationships with other relevant account holders, to support the engagement of our followers with their content by sharing their posts on our platforms, as well as to promote engagement with potential research participants, by their sharing our posts on their platforms. These account holders included Indigenous organizations, parenting and early childhood education groups, and popular leaders within Indigenous communities in Southern Ontario. To establish connections with these accounts, we sent private messages via the platform requesting they share our recruitment flyer with their followers. In turn, we shared their content on our sites. We aimed to foster trust and build online relationships with the account holders and their followers by promoting one another's content. Additionally, we asked our partner organizations and steering committee members to share our content on their personal and organization's social media accounts. As trusted sources of information, their sharing of our content enhanced the credibility of our accounts and instilled a stronger sense of confidence among eligible participants who contacted us to express their interest in participating [12,38].

Using Data Metrics

By analyzing data metrics, we recognized better engagement with our Instagram account than with our Facebook account. This was made evident by the number of likes, comments, and reactions to posts and stories, and followers, as well as the additional special data metrics provided by Instagram. Further, 1 reason for the difference in engagement between the 2 platforms may be the increased popularity of Instagram among adolescents and young adults, compared with Facebook [39]. While Facebook has been identified as the most widely used social media platform among researchers, its use is declining among younger users [2,9]. As a result, Facebook may not have been as appealing to our target audience, which primarily consisted of parents, generally aged younger than 40 years [40]. Despite this, more study participants expressed that they became aware of the FOTNG study through our Facebook page rather than our Instagram site.

How we delivered content also impacted user engagement. Data metrics on both Facebook and Instagram showed that content delivered in the form of a carousel performed better than single-image posts, and visually appealing content garnered more attention from our target audience, regardless of the day of the week or time the content was posted. In addition, we found that posts containing the faces of actual people had more views in comparison to posts with text-only, cartoon images, or other graphic variations. For instance, our World Smile Day post on Facebook, which consisted of actual human faces, reached more than 10 times the number of accounts than our text-only image posted 10 days later which outlined the FOTNG study framework. This is in line with the findings by Cyhlarova et al [38], who found that more personable recruitment methods are more effective. Similarly, posting on and about Indigenous-specific national holidays or observances performed significantly better than posts about Canadian national holidays or observances. As an illustration, during Treaties Recognition Week (Figure 3 [25]), we posted content that detailed the history and importance of acknowledging treaties in our local area. This post reached nearly 5 times the number of accounts on Facebook and nearly double the number of Instagram accounts when compared to our World Kindness Day post, which was published 7 days later. Our experience suggests that content specific to the target audience's cultural, spiritual, and historical needs and interests may generate more engagement and interactions than other content. As explained by Sutherland [41], creating appealing content that is tailored to a target audience's interests can foster an online community with which end users can connect and interact, generating greater engagement.

Time Invested by the Research Team

Gaining attention, credibility, and followers on Facebook and Instagram involved a significant amount of time and effort on behalf of our research team. Specifically, devising and implementing our social media recruitment plan involved: (1) brainstorming, researching, and sourcing appropriate topics; (2) creating clear, concise, engaging, and culturally appropriate content; (3) sourcing appropriate visuals; (4) prescheduling content; and (5) analyzing data metrics [41]. While these efforts were necessary to present relevant, worthwhile content that our followers wanted to engage with [41], it required 2 research assistants to spend a total of approximately 10 hours per week over 3 months to develop and execute our social media recruitment strategy to recruit 22 parent participants. Our experience aligns with Sutherland [41] who found that teams need to dedicate extensive time to researching their targeted audience and creating evidence-based social media strategies that can guide actions and decisions. Failing to allocate an adequate amount of time and funding to content creation can result in posts containing incorrect information or poor use of images, potentially damaging the account's (and study's) reputation.

XSL•FO
RenderX

Discussion: Lessons Learned

Thus far we have described our approach to creating and implementing an online recruitment strategy using Facebook and Instagram social media platforms. This approach has been successful in creating community engagement with local Indigenous organizations, parent and child-focused organizations, and Indigenous community members. Combined with other more traditional recruitment strategies, such as distributing and posting hard copies of flyers, attending events, and word of mouth, we have achieved our recruitment goals. Throughout our use of online recruitment using social media, we have realized many benefits as well as challenges, which we outline next. These are particularly salient for researchers considering using online recruitment strategies, to design a thoughtful and effective, yet informed and cautious approach.

Benefits

Community Trust and Engagement

One of the benefits of using social media for participant recruitment is engaging and building trust with the community from which our target participants are members. Our recruitment strategy was intentionally designed to foster a safe and trustworthy online space for Indigenous families. By not only sharing content directly related to our study, but also content around Indigenous-specific cultural holidays and observances, and information about family health and parenting, our efforts facilitated public engagement and trust in the project and demonstrated our sincere desire to support Indigenous communities [41]. The trust of local organizations in our project became evident as several Indigenous-led organizations shared our recruitment materials with their followers. This collaborative effort resulted in individuals reaching out to participate in this study. While we did not systematically collect data on how participants learned about our study, many individuals voluntarily informed us during the screening process that they learned about our research through recruitment flyers posted on Indigenous community organizations' social media accounts.

Tailored Content

Another benefit that social media has to offer researchers is the distribution of tailored content through diverse formats [13]. Facebook and Instagram allowed us to disseminate our original recruitment flyer as well as iterations of it that were more visually appealing, in addition to standalone posts, carousels, and reels or videos of varying lengths. Data metrics from each platform also showed us that by tailoring our content to cultural, spiritual, and historical interests, and featuring real people's faces (eg, fathers and children), we significantly improved engagement, as they emphasized relatability [38,41].

Adaptability through Data Metrics

Data metric analyses are another benefit of leveraging social media to recruit participants as the data informs researchers of the effectiveness of their content and can be used to quickly revise one's approach to better meet the engagement activities of users. For example, after reflecting on the analytic data, we were able to quickly shorten our recruitment video from 58 to 21 seconds increasing our engagement by 12 times the number of followers.

Challenges

Using Facebook and Instagram for recruitment was not without significant challenges, however. Fraudulent participants, budget constraints, inaccurate data metrics, and user access are real concerns to be thoughtfully considered before initiating a recruitment strategy using social media.

Fraudulent Participants

First, the presence of fraudulent participants, defined as ineligible persons or computer bots designed to pose as real people in a research study [17], is a known pervasive problem in health care studies where online recruitment methods are used [16,42]. As our recruitment materials initially mentioned a CAD $50 (US $37) gift card honorarium as an incentive for participation, some individuals (or bots) were motivated to falsify their eligibility and participate in this study in hopes of receiving the honorarium [20,43,44]. After discovering a fraudulent participant had passed our email screening process undetected, the research team quickly modified the screening procedures to mitigate the enrollment of fraudulent participants. The modified process included screening over the phone instead of by email, allowing the project coordinator to interpret an individual's eligibility by gauging the ease with which they provided answers to screening questions. Realizing that outlining this study's specific inclusion criteria in our social media posts had made it easy for fraudulent participants to answer our screening questions, we removed many of these details from our posts and redesigned our questions to be challenging for ineligible people to answer. For instance, we changed an original screening question from "Do you live in Southern Ontario" to "Where do you live?" If an individual was unable to answer one of our screening questions, they were deemed ineligible to participate. Given that we did not use a more targeted online recruitment approach (eg, paid advertisements) to reach eligible individuals, we inevitably reached a large pool of people, any one of whom might choose to participate fraudulently. This put us at risk of collecting fraudulent data [21] and resulted in a great deal of stress and wasted time for the research team.

Researchers planning to leverage social media for recruitment purposes should be aware of the growing threat of fraudulent participants and incorporate strategies to prevent its occurrence and mitigate its impact. Lawlor et al [45] present a structured approach for researchers to both prevent and detect fraudulent participants in their studies called the REAL framework. This method consists of 4 key stages. First, researchers are prompted to reflect on the vulnerabilities within their recruitment plan, considering the survey design elements in place to minimize fraudulent entries. Second, they establish expectations by identifying typical data patterns and anomalies they might encounter. In the third stage, researchers analyze collected data to check for any deviations from these expectations. Finally, the framework emphasizes transparency through the establishment of criteria to label and exclude potentially fraudulent responses. Creating a recruitment strategy using the REAL framework can support researchers in anticipating and mitigating the threat of fraudulent participants.

XSL•FO
RenderX

Budget and Resource Constraints

A second challenge was the substantial financial investment required to create and implement a social media communication plan and recruitment strategy [2]. Each social media post necessitated several hours of work, from researching topics to finding appealing graphics and writing captions. To maintain a presence in our followers' social media feeds, we needed to create frequent and original content, which is also time-consuming [12]. This can represent a significant challenge for researchers, particularly if time and budget are limited. Researchers must carefully consider whether they have the necessary personnel and funds to use an online recruitment strategy, especially without a guaranteed achievement of the desired sample size [46].

Data Metrics Versus Participant Feedback

A third challenge involved inaccurate interpretation of the data metrics provided by the social media platforms. In our project, we learned that metrics can sometimes misrepresent the actual reach of content. For instance, our metrics on Facebook indicated limited engagement with our target audience compared to Instagram, yet many study participants reported discovering our study through Facebook, despite not following our page. This discrepancy underscores the importance of asking participants directly about how they learned about this study to gain a more accurate understanding of whether the recruitment strategies being used are effective.

Access Limitations

Finally, although social media provided us with access to a diverse and massive set of individuals, our social media recruitment efforts were likely limited to people with access to an electronic device and the internet, which may have introduced sampling biases [12]. To mitigate this limitation, we encouraged study participants to share information about this study within their own social networks [47] and asked our partner organizations and steering committee members to do the same. We also used additional traditional means of recruitment, such as distributing hard copies of the flyer and attending in-person and online events to share our study. Using traditional methods and continuing to prioritize in-person recruitment strategies was important to our steering committee, which prioritizes building relationships with members of the communities we are working with throughout our project to inform the project design and implementation, give back to community members, and support knowledge dissemination efforts. Participants' access to information about study participation is important for all researchers to consider, particularly if recruiting using social media will introduce critical sampling biases that may jeopardize study findings. For instance, using social media is likely not an effective way to recruit participants where internet access is low or absent, such as in some rural areas, or when participants are unlikely to use social media or own a device with internet access, such as young children. Researchers should carefully reflect on how their intended participants access information and make efforts to mitigate the risk of sampling biases in their recruitment strategies.

Recommendations and Future Research

Given our experience developing and implementing an online recruitment strategy using popular social media platforms, we have several recommendations for researchers interested in using this approach, as well as areas that require further research.

Create an Online Recruitment Plan

First, researchers must make well-considered decisions about online recruitment, considering factors such as platform suitability and feasibility given this study's objectives. A thoughtful plan is essential for guiding researchers in making informed choices about which platforms are most likely to yield successful results. For example, Macapagal et al [46] present the CAN-DO-IT model as a guide for researchers to use when using online recruitment methods, irrespective of platform. The model encompasses 7 iterative steps, from conceptualization to implementation and evaluation of a recruitment strategy. Researchers can support the growing body of literature on online recruitment strategies by writing about how they design their approaches, their experiences of effectiveness, and other lessons learned.

Fraudulent Participant Prevention Strategies

Second, due to the threat of fraudulent participants in online research [48], it is imperative for researchers to plan for and implement prevention strategies, recognizing that more than one strategy is often necessary to mitigate this risk. Several prevention and verification actions may be used, such as using CAPTCHA [49] in online surveys, refraining from including the existence of incentives in recruitment advertisements, advertising in specific social media groups, using paid advertisements, checking IP addresses, screening for eligibility in real time and on-camera, requesting that participants show government-issued identification, and asking questions only legitimate participants would be able to answer with ease [19-22,44,49-54]. The REAL framework described earlier, can be used as a structured approach to help researchers prevent and detect fraudulent participants in their studies [45].

Given the growing popularity of online recruitment methods in research, we recommend that future research be conducted to better understand the problem of fraudulent participants. This includes exploring study-related factors that attract fraudulent participants, and the development of mitigation, verification, and prevention strategies that researchers can use in conjunction with social media recruitment strategies to encourage study participation without the added risk of compromising data.

Optimal Resource Allocation

Third, it is crucial to plan for adequate financial resources for recruitment purposes when writing grant funding applications. While the initial set-up costs of using social media for recruitment are often more cost-effective compared to traditional methods [14,19,21], the true costs of a social media approach may manifest in the form of human resource capital and time. Unlike mailing where sending flyers is a 1-time or infrequent effort to reach the target audience, social media requires a sustained and consistent posting strategy to effectively leverage the platform's abilities to engage potential participants [19,55]. This entails dedicating time and effort to curate captivating and

XSL•FO
RenderX

eye-catching content to capture the attention of potential participants and prevent them from simply scrolling past your content [46]. While incentives may help solve this problem, it is imperative to recognize that they may also attract fraudulent participants. As a result, researchers must allocate additional time to verify participant identities as discussed earlier. This may not pose a significant challenge for studies with smaller sample sizes, however, with large-scale studies (eg, Ali et al [47]), the added time and resource requirements can substantially impact the overall costs and project feasibility. Overall, the production of compelling content can translate into an increased workload for research staff, in addition to the time needed to analyze performance metrics, which are essential for optimizing user engagement.

Build Trust and Credibility

Fourth, building trust and credibility is very important when using social media for recruitment, particularly in community-engaged research with Indigenous communities, with whom unethical research practices have understandably resulted in distrust of researchers and academia [4,56-60]. In our experience, the use of social media enabled us to introduce our project to organizations and people we might otherwise have not encountered using traditional methods. However, our efforts to build an authentic online presence with support from Indigenous-led organizations and community members were what significantly enhanced our success. Researchers would do well to build relationships with community members and organizations relevant to their project aims, who can then support participant recruitment by sharing opportunities with their followers. This in turn will increase the reach of a social media campaign to engage participants who are more likely to be eligible to participate, while simultaneously fostering trust and credibility of the project to the public [4].

Contributing to the Literature

Fifth, as online recruitment methods are constantly evolving, collaboration is necessary for researchers to share experiences, challenges, and recommendations that can support one another in developing effective recruitment strategies. Researchers should detail their recruitment approaches and experiences in their publications and presentations, to contribute to the growing body of knowledge regarding social media recruitment. This collective effort can help to improve the efficiency and effectiveness of social media recruitment strategies, and help to address challenges such as representative bias [12,47] and fraudulent participants [20].

Final Thoughts

Our case study offers important insights into the use of social media as a recruitment tool for engaging Indigenous fathers and Two-Spirit parents, using a community-engaged research approach. We have detailed our approach from its initial creation, considerations for creating content, and how one might interpret and respond to user engagement metrics. While we experienced numerous benefits, we also outline several challenges of using social media for study recruitment which researchers should consider before developing their approach. Despite these challenges, our social media accounts were able to foster trust and build relationships with other Indigenous-led organizations and community members in the virtual space, which supported our study recruitment through their sharing about our study on their platforms and promoting our project. Building trusting relationships with community groups, while ethically imperative in conducting research with Indigenous communities, proved indispensable to our recruitment efforts, and cannot be replaced by social media alone. This comprehensive approach, encompassing both social media and community engagement, allowed us to make significant strides in achieving our recruitment goals.

Limitations

Several limitations of our experience are worth discussing. First, our experiences using social media for health care research is limited to Facebook and Instagram. Recruiting via other platforms might have generated more interest in our study. Our approach has also limited our ability to compare the effectiveness of Facebook and Instagram for study recruitment to other social media platforms such as X. Second, we used the special data metrics we were granted access to from Instagram to guide our social media plan for both Facebook and Instagram. Although our followers continue to increase, obtaining access to the special data metrics for Facebook is necessary for devising a tailored plan for this platform. Finally, as we did not intend to test our recruitment strategies for effectiveness at the outset of our project, we did not specifically ask participants how they learned about this study. Researchers should consider collecting this data in all their studies from the outset so they can gauge the effectiveness of their recruitment strategies and adjust their approach as needed.

Conclusion

Our case study provides researchers with a detailed guide on what to consider and how to design and implement a social media recruitment strategy. Our experience is derived from a community-engaged project with Indigenous communities in Southern Ontario and demonstrates our ability to create relationships in an online space, which in turn enhanced the effectiveness of our approach. While there are many benefits to recruiting using social media, researchers should be thoughtful in their approach to devising a communication strategy, using data metrics to enhance user engagement, while planning mitigation tactics for the risk of fraudulent participants. Finally, researchers can contribute to the ongoing discourse on social media recruitment strategies by writing about their experiences in publications, facilitating the advancement of the use of social media for participant recruitment.

Acknowledgments

This work was funded by the Canadian Institutes of Health Research. We would like to acknowledge the entire FOTNG steering committee for their continued guidance on our social media communication and recruitment strategies used for this project.

XSL•FO
RenderX

Steering committee members include: C Bomberry, ML Butt, J Dame, B Freeman, S George, S Hachey, S Harris, P Hill, B Jacobs, K Jacobs, D Johnson, V Miller, D Miller, C Neuert, J Pace, W Spring, R VanEvery, YJ Willet, and A Wright.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Terms used by Facebook and Instagram.
[DOCX File , 21 KB - jmir_v26i1e51751_app1.docx]

References

1. Aichner T, Grünfelder M, Maurer O, Jegeni D. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. Cyberpsychol Behav Soc Netw 2021;24(4):215-222 [FREE Full text] [doi: 10.1089/cyber.2020.0134] [Medline: 33847527]
2. Darko EM, Kleib M, Olson J. Social media use for research participant recruitment: integrative literature review. J Med Internet Res 2022;24(8):e38015 [FREE Full text] [doi: 10.2196/38015] [Medline: 35925655]
3. Reagan L, Nowlin SY, Birdsall SB, Gabbay J, Vorderstrasse A, Johnson C, et al. Integrative review of recruitment of research participants through Facebook. Nurs Res 2019;68(6):423-432. [doi: 10.1097/NNR.0000000000000385] [Medline: 31693547]
4. Garrido-Hernansaiz H. The use of online social media for the recruitment of people living with HIV in Spain and Latin America: lessons from two studies. Health Soc Care Community 2022;30(6):e4065-e4073 [FREE Full text] [doi: 10.1111/hsc.13799] [Medline: 35318765]
5. Shaver LG, Khawer A, Yi Y, Aubrey-Bassler K, Etchegary H, Roebothan B, et al. Using Facebook advertising to recruit representative samples: feasibility assessment of a cross-sectional survey. J Med Internet Res 2019;21(8):e14021 [FREE Full text] [doi: 10.2196/14021] [Medline: 31429409]
6. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020;22(10):e22810 [FREE Full text] [doi: 10.2196/22810] [Medline: 33104015]
7. Company information, culture, and principles | About Meta. Meta. 2023. URL: https://about.meta.com/company-info/ [accessed 2023-11-07]
8. Facebook. Meta. 2023. URL: https://about.meta.com/technologies/facebook-app/ [accessed 2023-11-07]
9. Ford KL, Albritton T, Dunn TA, Crawford K, Neuwirth J, Bull S. Youth study recruitment using paid advertising on Instagram, Snapchat, and Facebook: cross-sectional survey study. JMIR Public Health Surveill 2019;5(4):e14080 [FREE Full text] [doi: 10.2196/14080] [Medline: 31599739]
10. It's what's happening / X. X Corp. Twitter. 2023. URL: https://twitter.com/?lang=en-ca [accessed 2023-07-27]
11. Flood-Grady E, Solberg LB, Baralt C, Meyer M, Stevens J, Krieger JL. Engaging institutional stakeholders to develop and implement guidelines for recruiting participants in research studies using social media: mixed methods, multi-phase process. J Med Internet Res 2021;23(10):e23312 [FREE Full text] [doi: 10.2196/23312] [Medline: 34623319]
12. Marks A, Wilkes L, Blythe S, Griffiths R. A novice researcher's reflection on recruiting participants for qualitative research. Nurse Res 2017;25(2):34-38. [doi: 10.7748/nr.2017.e1510] [Medline: 29115753]
13. Farsi D. Social media and health care, part 1: literature review of social media use by health care providers. J Med Internet Res 2021;23(4):e23205 [FREE Full text] [doi: 10.2196/23205] [Medline: 33664014]
14. Ellington M, Connelly J, Clayton P, Lorenzo CY, Collazo-Velazquez C, Trak-Fellermeier MA, et al. Use of Facebook, Instagram, and Twitter for recruiting healthy participants in nutrition-, physical activity-, or obesity-related studies: a systematic review. Am J Clin Nutr 2022;115(2):514-533 [FREE Full text] [doi: 10.1093/ajcn/nqab352] [Medline: 34669955]
15. Benedict C, Hahn AL, Diefenbach MA, Ford JS. Recruitment via social media: advantages and potential biases. Digit Health 2019;5:2055207619867223 [FREE Full text] [doi: 10.1177/2055207619867223] [Medline: 31431827]
16. Ballard AM, Cardwell T, Young AM. Fraud detection protocol for web-based research among men who have sex with men: development and descriptive evaluation. JMIR Public Health Surveill 2019;5(1):e12344 [FREE Full text] [doi: 10.2196/12344] [Medline: 30714944]
17. Hohn KL, Braswell AA, DeVita JM. Preventing and protecting against Internet research fraud in anonymous web-based research: protocol for the development and implementation of an anonymous web-based data integrity plan. JMIR Res Protoc 2022;11(9):e38550 [FREE Full text] [doi: 10.2196/38550] [Medline: 36094806]
18. Dol J, Tutelman PR, Chambers CT, Barwick M, Drake EK, Parker JA, et al. Health researchers' use of social media: scoping review. J Med Internet Res 2019;21(11):e13687 [FREE Full text] [doi: 10.2196/13687] [Medline: 31719028]
19. Bethel C, Rainbow JG, Dudding KM. Recruiting nurses via social media for survey studies. Nurs Res 2021;70(3):231-235. [doi: 10.1097/NNR.0000000000000482] [Medline: 33060416]

XSL•FO
RenderX

20. Griffin M, Martino RJ, LoSchiavo C, Comer-Carruthers C, Krause KD, Stults CB, et al. Ensuring survey research data integrity in the era of internet bots. Qual Quant 2022;56(4):2841-2852 [FREE Full text] [doi: 10.1007/s11135-021-01252-1] [Medline: 34629553]
21. Pozzar R, Hammer MJ, Underhill-Blazey M, Wright AA, Tulsky JA, Hong F, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020;22(10):e23021 [FREE Full text] [doi: 10.2196/23021] [Medline: 33026360]
22. Sanchez C, Grzenda A, Varias A, Widge AS, Carpenter LL, McDonald WM, et al. Social media recruitment for mental health research: a systematic review. Compr Psychiatry 2020;103:152197 [FREE Full text] [doi: 10.1016/j.comppsych.2020.152197] [Medline: 32992073]
23. Kumarasamy V, Goodfellow N, Ferron EM, Wright AL. Evaluating the problem of fraudulent participants in health care research: multimethod pilot study. J Med Internet Res 2024;8:e51530 [FREE Full text] [doi: 10.2196/51530]
24. National Indigenous Peoples Day. Government of Canada. 2023. URL: https://www.rcaanc-cirnac.gc.ca/eng/1100100013718/1534874583157 [accessed 2023-11-19]
25. Treaties recognition week. Ontario. 2023. URL: https://www.ontario.ca/page/treaties-recognition-week [accessed 2023-11-13]
26. Farid Z. Why repurposing content is beneficial for social media reach? SocialChamp. 2021. URL: https://www.socialchamp.io/blog/repurposing-content-for-social-media-reach/ [accessed 2021-11-12]
27. Canva. 2023. URL: https://www.canva.com/ [accessed 2023-11-12]
28. Tully L, Spyreli E, Allen-Walker V, Matvienko-Sikar K, McHugh S, Woodside J, et al. Recruiting 'hard to reach' parents for health promotion research: experiences from a qualitative study. BMC Res Notes 2021;14(1):276 [FREE Full text] [doi: 10.1186/s13104-021-05653-1] [Medline: 34289873]
29. How do I use hashtags on Facebook? | Facebook help centre. Meta. 2023. URL: https://www.facebook.com/help/587836257914341/ [accessed 2023-11-15]
30. Gkikas DC, Tzafilkou K, Theodoridis PK, Garmpis A, Gkikas MC. How do text characteristics impact user engagement in social media posts: modeling content readability, length, and hashtags number in Facebook. Int J Inf Manage Data Insights 2022;2(1):100067. [doi: 10.1016/j.jjimei.2022.100067]
31. Gunning R. The fog index after twenty years. Journal of Business Communication 1969;6(2):3-13. [doi: 10.1177/002194366900600202]
32. Zhang C, Zheng H, Wang Q. Driving factors and moderating effects behind citizen engagement with mobile short-form videos. IEEE Access 2022;10:40999-41009. [doi: 10.1109/access.2022.3167687]
33. Zhu C, Xu X, Zhang W, Chen J, Evans R. How health communication via Tik Tok makes a difference: a content analysis of Tik Tok accounts run by chinese provincial health committees. Int J Environ Res Public Health 2019;17(1):192 [FREE Full text] [doi: 10.3390/ijerph17010192] [Medline: 31892122]
34. Wang J, Calderon G, Hager ER, Edwards LV, Berry AA, Liu Y, et al. Identifying and preventing fraudulent responses in online public health surveys: lessons learned during the COVID-19 pandemic. PLOS Glob Public Health 2023;3(8):e0001452 [FREE Full text] [doi: 10.1371/journal.pgph.0001452] [Medline: 37610999]
35. Cuevas-Molano E, Sánchez-Cid M, Gordo-Molina V. Brand strategy and content management on Instagram: scheduling and message length as factors to improve engagement. Commun Soc 2022;35(2):71-87. [doi: 10.15581/003.35.2.71-87]
36. Luarn P, Lin YF, Chiu YP. Influence of Facebook brand-page posts on online engagement. Online Inf Rev 2015;39(4):505-519. [doi: 10.1108/oir-01-2015-0029]
37. Aydin S, Koçak O, Shaw TA, Buber B, Akpinar EZ, Younis MZ. Investigation of the effect of social media addiction on adults with depression. Healthcare (Basel) 2021;9(4):450 [FREE Full text] [doi: 10.3390/healthcare9040450] [Medline: 33920478]
38. Cyhlarova E, Clark M, Knapp M. Recruiting Participants for Adult Social Care Studies: Challenges and Mitigating Strategies. 2020. URL: https://core.ac.uk/download/pdf/323304729.pdf [accessed 2024-09-27]
39. Arigo D, Pagoto S, Carter-Harris L, Lillie SE, Nebeker C. Using social media for health research: methodological and ethical considerations for recruitment and intervention delivery. Digit Health 2018;4:2055207618771757 [FREE Full text] [doi: 10.1177/2055207618771757] [Medline: 29942634]
40. Trifiro B. Instagram use and its effect on well-being and self-esteem. Master of Arts in Communication. 2018. URL: https://digitalcommons.bryant.edu/macomm/4 [accessed 2024-08-30]
41. Sutherland KE. Intrategic Social Media Management. Singapore: Springer Nature Singapore; 2021:1-4.
42. Heffner J, Watson N, Dahne J, Croghan I, Kelly MM, McClure JB, et al. Recognizing and preventing participant deception in online nicotine and tobacco research studies: suggested tactics and a call to action. Nicotine Tob Res 2021;23(10):1810-1812 [FREE Full text] [doi: 10.1093/ntr/ntab077] [Medline: 33876244]
43. Mournet AM, Kleiman EM. Internet-based mental health survey research: navigating internet bots on Reddit. Cyberpsychol Behav Soc Netw 2023;26(2):73-79. [doi: 10.1089/cyber.2022.0173] [Medline: 36724303]
44. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133 [FREE Full text] [doi: 10.1111/jlme.12200] [Medline: 25846043]

XSL•FO
RenderX

45. Lawlor J, Thomas C, Guhin AT, Kenyon K, Lerner MD, Drahota A. Suspicious and fraudulent online survey participation: introducing the REAL framework. Method Innov 2021;14(3):1-52. [doi: 10.1177/20597991211050467]
46. Macapagal K, Li DH, Clifford A, Madkins K, Mustanski B. The CAN-DO-IT model: a process for developing and refining online recruitment in HIV/AIDS and sexual health research. Curr HIV/AIDS Rep 2020;17(3):190-202 [FREE Full text] [doi: 10.1007/s11904-020-00491-5] [Medline: 32444929]
47. Ali SH, Foreman J, Capasso A, Jones AM, Tozan Y, DiClemente RJ. Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. BMC Med Res Methodol 2020;20(1):116 [FREE Full text] [doi: 10.1186/s12874-020-01011-0] [Medline: 32404050]
48. Woolfall K. Identifying and preventing fraudulent participation in qualitative research. Arch Dis Child 2023;108(6):421-422. [doi: 10.1136/archdischild-2023-325328] [Medline: 37068922]
49. CAPTCHA: telling humans and computers apart automatically. Carnegie Mellon University. 2010. URL: http://www.captcha.net/ [accessed 2023-07-27]
50. Fox AU. Social Media Analytics Strategy. 2022. URL: https://ideas.repec.org/b/spr/sprbok/978-1-4842-8306-6.html [accessed 2024-08-30]
51. Glazer JV, MacDonnell K, Frederick C, Ingersoll K, Ritterband LM. Liar! Liar! Identifying eligibility fraud by applicants in digital health research. Internet Interv 2021;25:100401. [doi: 10.1016/j.invent.2021.100401]
52. Moss A, Rosenzweig C, Jaffe SN, Gautam R, Robinson J, Litman L. Bots or inattentive humans? Identifying sources of low-quality data in online platforms. PsyArXiv Prepr 2021:1-41. [doi: 10.31234/osf.io/wr8ds]
53. Parks AM, Duffecy J, McCabe JE, Blankstein Breman R, Milgrom J, Hirshler Y, et al. Lessons learned recruiting and retaining pregnant and postpartum individuals in digital trials: viewpoint. JMIR Pediatr Parent 2022;5(2):e35320 [FREE Full text] [doi: 10.2196/35320] [Medline: 35107422]
54. Storozuk A, Ashley M, Delage V, Maloney EA. Got Bots? Practical recommendations to protect online survey data from bot attacks. TQMP 2020;16(5):472-481. [doi: 10.20982/tqmp.16.5.p472]
55. Singh C. Understanding social media algorithm 2023 complete guide. SocialPilot. URL: https://www.socialpilot.co/blog/social-media-algorithm [accessed 2024-07-31]
56. Mashford-Pringle A, Pavagadhi K. Using OCAP and IQ as frameworks to address a history of trauma in indigenous health research. AMA J Ethics. 2020. URL: https://journalofethics.ama-assn.org/sites/journalofethics.ama-assn.org/files/2020-09/stas1-2010_1.pdf [accessed 2023-11-15]
57. Cain T. Decolonizing Methodologies: Research and Indigenous Peoples. New York, NY, USA: Zed Books; 2013.
58. Mashford-Pringle A, Hiscock C, Rice EJ, Scott B. Weaving First Nations, Inuit, and Métis principles and values into health research processes. J Clin Epidemiol 2023;160:54-60. [doi: 10.1016/j.jclinepi.2023.05.012] [Medline: 37217105]
59. Jull J, Queen's University, King A. Indigenous community research partnerships can help address health inequities. The Canadian Press. 2021. URL: https://theconversation.com/indigenous-community-research-partnerships-can-help-address-health-inequities-152705 [accessed 2024-08-30]
60. Nickels S, Shirley J, Laidler G. Negotiating research relationships with inuit communities, a guide for researchers. Inuit Tapiriit Kanatami and Nunavut Research Institute. 2007. URL: https://www.itk.ca/negotiating-research-relationships-guide/ [accessed 2023-11-15]

Abbreviations

FOTNG: Fathers of the Next Generation

Edited by N Cahill; submitted 11.08.23; peer-reviewed by M Eder, E Miyagi, C Bethel, J Martin-Kerry; comments to author 06.10.23; revised version received 21.11.23; accepted 31.07.24; published 11.10.24.

Please cite as:
Wright AL, Willett YJ, Ferron EM, Kumarasamy V, Lem SM, Ahmed O
Using Social Media to Recruit Participants in Health Care Research: Case Study
J Med Internet Res 2024;26:e51751
URL: https://www.jmir.org/2024/1/e51751
doi:10.2196/51751
PMID:

©Amy L Wright, Ysabella Jayne Willett, Era Mae Ferron, Vithusa Kumarasamy, Sarah M Lem, Ossaid Ahmed. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Enhancing Research Involvement of Young People With Lived Expertise: Reflecting on Experiences in Digital Mental Health Research

Josephine Brogden[1*]; Zsofi de Haan[1*], MCP; Carla Gorban[1*]; Samuel J Hockey[1*]; Alexis Hutcheon[1*]; Frank Iorfino[1], PhD; Yun Ju C Song[1], PhD; Elizabeth Scott[1], PhD; Ian B Hickie[1], MD; Sarah McKenna[1], PhD

Brain and Mind Centre, The University of Sydney, Camperdown, NSW, Australia

[*]these authors contributed equally

Corresponding Author:
Sarah McKenna, PhD
Brain and Mind Centre
The University of Sydney
94-100 Mallett St
Camperdown, NSW, 2050
Australia
Phone: 61 (02) 9351 0774
Email: sarah.mckenna@sydney.edu.au

Abstract

Given the rapid development of digital mental health technologies and a focus on connecting with youth, there is an urgent need to enhance the engagement of young people with lived expertise in research. Even so, youth with lived experience of accessing mental health services are particularly affected by power imbalances and may receive limited compensation in academic settings. Therefore, an emphasis on how research engagement not only improves the work but can benefit young people themselves is required. Here, 5 young people with lived expertise report on their experience of being employed as researchers at the University of Sydney's Brain and Mind Centre. As such, this team is uniquely placed to offer reflections from their work across multiple stages of research. This led to four key insights, including (1) creating accepting work cultures, (2) providing diverse opportunities for involvement, (3) giving young people agency and flexibility around sharing lived experiences, and (4) creating accommodating work environments for all researchers. We suggest that these insights can support more diverse ways of engaging young people and maximizing the value of participation for both researchers and young people themselves.

(J Med Internet Res 2024;26:e55441) doi:10.2196/55441

KEYWORDS

youth mental health; digital mental health; lived expertise; young people; youth; adolescence; technologies; university; universities; Sydney; real-world; engagement; work environment

Introduction

In Australia, 26% of young adults (aged 15 to 25 years) have experienced a long-term mental health or behavioral condition, representing a major health crisis [1]. This has prompted the rapid development of digital technology solutions, such as e-therapies and tracking tools, to provide support options that are youth-friendly and accessible. Yet, while clinical trials in controlled settings have frequently demonstrated the potential effectiveness of these technologies, these trials often report high attrition rates and decreasing usage over time [2-4]. For this reason, there are increasing calls to involve young people in digital mental health research so that new tools are optimally designed to suit the needs and preferences of the populations they seek to help [5-7].

Given this is the case, it is important to consider what young people themselves can gain from their involvement. It is concerning to find limited data regarding the value of research involvement for young people and, more particularly, how this value can be enhanced [5,8-11]. For example, a 2023 targeted review of youth involvement in mental health research found 16 articles that had described both youth involvement in research and the impacts of that involvement [12]. In work that does exist, the scope of young people's involvement is limited as compared with other groups. The 2023 review found that youth were engaged at the advisory level in 4 studies, at the

XSL•FO
RenderX

consultation level in 5 studies, at a partnership level in 8 studies, and only 1 study was "youth-led" [12]. In a more targeted review of youth involvement in digital mental health, no studies involved youth in the planning or leadership of research [13]. By contrast, a review of 70 articles detailing partnerships with adult patients in health research found that most research included patients across multiple stages, and 38 had included patients in the whole research cycle, including planning stages [11]. Thus, it is important that researchers consider more diverse types of contributions that young people with lived expertise can make, as well as the potential value of expanding their involvement.

When considering the depth of consumer involvement, Arnsteins' ladder of participation has been hugely influential in both academic and public policy spaces. This ladder conceptualizes "citizen-control" as the highest form of participation, suggesting that projects should ideally be led by young people with lived experience of mental ill-health, including having key roles in decision-making and planning [14]. There is also an acknowledgment that this "ladder" may need to be more flexible and dynamic when working with youth populations, who often prefer to choose and change the depth of their involvement based on their interests [15]. Hence, while there is a need to increase young people's involvement in decision-making and leadership roles, young people should have a say in designing their involvement and may not always want "citizen control." Ultimately, it is incumbent upon research institutions to create optimal conditions for young people to have agency and flexibility regarding their own involvement [5,16]. In addition, much attention has been given to how researchers interact with individuals with lived expertise, not just what activities they are involved in. While frameworks are often project-specific, key principles that are continuously promoted include supportive institutional policies, supportive attitudes that ensure strong communication and shared goals, mutual respect, addressing training needs, providing resources and advanced planning, and recognizing the value of partnership throughout all stages of research [5,16].

Taken together, while increasing engagement between young people and researchers is vital, we should also consider the extent to which this participation is benefitting young people and whether key principles of involvement are being realized, in other words, addressing whether young people do feel respected and recognized by academic researchers. Young people themselves should be supported in providing their own reflections in their own words. There are increasing calls to elevate lived expertise perspectives of being involved in research, given a longstanding history of coercion and damaging power dynamics in academic settings [17,18]. When planning research involvement of individuals with lived expertise, it is important for academic researchers to be empathic and understanding about how this involvement will be experienced, particularly by vulnerable populations. Furthermore, they should seek to enhance the benefits of research participation for these individuals where possible.

The Youth Mental Health and Technology team at The University of Sydney's Brain and Mind Centre collaborates on a range of projects related to improving youth mental health services, partly through novel digital tools. Our team has consistently documented involving young people with lived expertise in our work [19-21]; however, our partnerships with young people were strengthened in 2021 by the establishment of a lived experience working group, coled by permanent lived experience researchers on our team, and documented in a previous publication [22]. In addition, our team has successfully employed 5 young people as lived expertise researchers on an ongoing basis to ensure that young people are given more flexibility and autonomy in their contributions to our work. This has also led to stronger collaborations between academic researchers and young people in our community more broadly. Accordingly, 5 lived expertise researchers on our team wrote independent reflections and participated in reflective group discussions. We hope that this can provide a greater understanding of how to enhance the value of research involvement for young people themselves in line with existing frameworks and guidelines.

Overview of Lived Expertise Research Team and Reflective Process

Ethical Considerations

Our team has published a framework that underpins our team's lived expertise contributions to research [22] and has received ethical approval from the University of Sydney's Human Research and Ethics Committee (Project number: 2020/786) for this work. Given that lived expertise researchers are reflecting on their own experiences, we did not seek informed consent. Even so, quotes and reflections were deidentified so that lived expertise researchers could provide commentary confidentially. All lived expertise researchers were hired as casual, full-time, or part-time professional staff.

Description of Lived Expertise Research Team and Responsibilities

In this paper, the term "lived expertise researcher" refers to young people with lived experiences of seeking treatment for mental ill-health. In line with Killackey's framework, most lived expertise researchers also had other types of expertise in mental health, including laboring (which includes working in mental health services), loving (caring for someone with mental ill-health), and learning (being involved in mental health research) [23]. By contrast, this paper will refer to all academic staff with learned expertise and no declared lived expertise in mental ill-health as "academic researchers." The lived expertise researchers on our team are between 20 and 31 years old and have been employed for between 9 months and 5 years. They have been involved in a range of projects at various stages. Furthermore, they have moved between advisor, partner, and decision-making roles [22,24,25]. Table 1 describes the research activities of this team according to the Smits et al [24] involvement matrix. Many of these roles have been described in more detail in previous publications [22,25,26].

XSL•FO
RenderX

Table 1. Overview of lived expertise researchers' involvement in projects conducted by the Youth Mental Health and Technology team.

Role in project	Nature of lived expertise researchers' involvement in project	Stage of project when this occurred
Listener is given information	• Grant funding needed to be sought before hiring lived expertise researchers, so the major responsibilities of these roles had been designed before lived expertise researchers joined the team.	• Preparation[a]
Cothinker is asked to give opinion	• Lived expertise researchers are included in regular team meetings during which new projects or grant proposals are being discussed, researchers provide updates on ongoing projects and troubleshoot problems, and findings from data analysis are discussed and interpreted.	• Preparation • Execution[b] • Implementation[c]
Advisor gives unsolicited advice	• Lived expertise researchers work alongside academic researchers as part of working groups that are developing new digital tools and helping to design user flow and content.	• Preparation • Execution • Implementation
Partner works as an equal partner.	• A lived expertise researcher has contributed to designing procedures for an ongoing clinical trial, as part of which they perform the role of "Digital Navigator," assisting young people to understand how to use a novel digital tool. This included partnering with the academic lead to design technology implementation protocols, role descriptions, and qualitative data collection tools. This lived expertise researcher has also assisted with recruitment and data collection and is a co–first author on publications about the novel role that they have performed in this trial. • Lived expertise researchers have coled the development and ongoing running of a lived expertise working group and have coled qualitative workshops. This has included creating protocols, assisting with ethics proposals, recruiting members, organizing meetings, collecting evaluative data, and contributing to outputs.	• Preparation • Execution • Implementation
Decision maker takes initiative (final) decisions	• A lived expertise researcher has developed an information workshop to translate the Brain and Mind Centres Model of Care (that underpins the digital tools we are developing) to young people. This will be delivered to health services and evaluated.	• Implementation

[a]Preparation stage includes developing ideas, formulating questions, financing, and ethical approval.

[b]Execution stage includes recruiting study subjects, choosing instruments, data collection, and data analysis.

[c]Implementation stage includes report writing, dissemination in the media, and translation to practice.

Reflective Process and Approach to Data Analysis

To ensure that all members of our team had an equal opportunity to contribute, despite varying experience and training in academia, we used a process of coreflection that was designed by our team and was underpinned by the principles of thematic analysis [17,27] as this approach allows both inductive and theory-driven data analysis. Our approach followed reflective processes that had previously been adopted by lived expertise groups to provide their collective experiences of being involved in research [17]. Based on existing guidelines, a clinical researcher was also involved in providing guidance and support, as this can improve young people's level of involvement in academic projects [15].

Young people wrote personal reflections using the prompt, "How have you experienced being a member of the Lived Expertise team?" This broad prompt allowed us to identify key aspects of the research experience from the perspective of our team rather than being encouraged to focus on key topics. Following this, all team members were given several weeks to read over the reflections, which informed 2 group discussions, both of which were held over 2 hours through Zoom (Zoom Video Communications), the focus of which was to identify common topics. Discussions were facilitated by a clinical researcher, video recorded, and later transcribed. Subsequently, a clinical researcher (SM) who was trained in thematic analysis and a lived expertise researcher (SJH) summarized topics into themes based on these discussions and existing frameworks on participatory practices. All lived expertise researchers also contributed to the drafting of this paper to ensure it reflected their beliefs and wishes. For example, the first draft of this paper was edited so that language use was more inclusive.

Data analysis was informed by thematic analyses, as themes were derived from what was in the data (written reflections and discussions), as well as from existing frameworks on creating supportive partnerships between academic and lived expertise researchers [5,27]. This was done to ensure that we captured themes that were most relevant to the lived expertise researchers themselves while also considering existing literature on creating positive experiences for young people. Data analyses followed the constructivist grounded theory that argues all knowledge is constructed by meanings that individuals bring to data analysis [28]. As this was a group reflective process, researchers' own perspectives and experiences influenced the organization of data into themes and categories.

XSL•FO
RenderX

Insights From Lived Expertise Researchers on a Digital Mental Health Team

Creating an Accepting Culture and Addressing Academic Ableism

Academic institutions are often seen as ableist and exclusionary by those with lived expertise, particularly young people, due to unequal compensation and opportunities for development as compared with academic staff. Previous reviews have found that the risks of hiring researchers with disabilities are overemphasized as compared with the benefits, that lived experience is not valued in the process of fellowship schemes, and that, in academic institutions, there are few individuals with disabilities of any kind in positions of power [29]. As such, there is a need to enable greater inclusion in academic institutions.

> *There was an uncomfortable feeling of being exposed before I had even arrived... I was worried about how I would navigate this awareness with my colleagues, but my apprehensions were allayed almost instantly. The Digital Health team welcomed me as a researcher first, with valuable, unique insights and able to contribute to the research projects, not solely because of my lived experience.*

For this lived expertise researcher, it was rewarding to be valued because of experiences that are generally highly stigmatized, and being so highly valued by colleagues also led to a sense of belonging and acceptance within their workplace. A lived expertise researcher who had been on the team for over 5 years felt that this increased acceptance was linked to a change in culture over time.

> *There has been a necessary encouragement toward an overall culture change in the role and importance of lived experience. I believe this to be from a new understanding of the process of inclusion in academic outputs... One main problem is ableism, one of the core things for me has been the differences in pay, and the fact that you can't go beyond a certain point if you don't have an academic qualifications.*

Relatedly, other lived expertise researchers felt that because of "ableist" policies in academic institutions, it is important for research teams to actively build a culture of inclusivity and acceptance.

> *In a broader research team, there's maybe less of an understanding about where the value comes from as someone with lived experience, and then maybe more of a sense of a need to justify that by sharing some of your lived experience.*

> *Sometimes it is hard to recognise what I bring to the table in a team of very accomplished researchers. So while no one has actively said or done anything to me I think the academic environment sometimes reinforces this insecurity. Being included as a core member of our team has been an important step in recognising that my lived experience adds value and is not something I need to use to justify why I am here.*

Overall, cultural and systemic problems in academic institutions can create power imbalances that negatively impact lived expertise researchers. Yet, when academic researchers can create welcoming and supportive environments for those who have traditionally been excluded from research, this can be highly rewarding for individuals. Furthermore, this can lead to more accepting workplace environments that benefit all researchers.

Providing Diverse Opportunities to Use a Variety of Skills

Given the nature of our teams' work, which is focused on designing and implementing new digital tools, lived expertise researchers were encouraged to contribute to projects in a variety of ways as shown in Table 1. Furthermore, these opportunities were not always linked to their lived expertise but allowed them to build capacities in new areas or to use other existing skill sets. For example, lived expertise researchers who were originally employed to implement a specific digital platform in services have also been involved in the planning and development of further digital tools, partly due to their experience working in health services or their clinical training.

> *The really great part about working in a broader research team is that you can actually be valued for other skill sets, and other parts of who you are… like, you know that you're also intelligent, or you're really good at writing or whatever it may be. There are other things that make you who you are, that are outside of this mental illness that's defined your life in many ways.*

> *[I have been] given opportunities, that I wouldn't necessarily think that I was capable of doing, [and wouldn't have] put my hand up for… [My manager] doesn't make assumptions. She's like, you don't have to but yeah, and it's that exposure therapy thing. That's really helped me build up confidence as hard as it is.*

Importantly, all lived expertise researchers on our team agreed that lived expertise advisory groups were highly valuable and rewarding in and of themselves, yet having the opportunity to contribute more diverse skills significantly enhanced the value of research involvement. As expressed by one lived expertise researcher below:

> *I think something that I'm like cautious of [is that] oh, are [academic researchers] gonna, like see me as someone who's lived experience and now assume that there's a set of things I can and can't do… you're just like this young lived experience person, is there anything else more than that? And the answer is like always, yes.*

Taken together, academic researchers can enhance the value of being involved in research by creating diverse opportunities for young people with lived expertise. Given the rapid progress being made in digital mental health, this field is particularly likely to benefit from more reflexive involvement of individuals with diverse skill sets.

XSL•FO
RenderX

Creating Agency and Flexibility Around Sharing Lived Experiences

Again, it was rewarding for our lived expertise researchers to be in an environment where their experiences of mental ill-health were valued, given that these illnesses continue to be highly stigmatized. [5,30] At the same time, as expressed by one lived expertise researcher below, publicly declaring mental ill-health was a difficult decision and did come with challenges.

> *Not having the luxury of hiding my mental-ill health at work has left me feeling vulnerable at times... Initially, I thought I had to reveal everything about myself so that others would consider my experiences valid enough. Despite having an on-going battle with this, the team's encouragement and inclusive culture minimise my insecurities that I am not accomplished or sick enough to be called a lived experience researcher."*

Thus, we believe it is important for all academic colleagues involved in mental health research to have training and additional support regarding how to openly and safely discuss experiences of mental ill-health. Without this step, lived expertise involvement may be limited, and the value may not be realized because of academic researchers' own attitudes and stigmas. For example, a lived expertise researcher who was originally employed to assist with the implementation of a technology platform in health services has also been involved in the development of further digital tools. This researcher reflected on an interaction they had had with other researchers on the project, who appeared uncomfortable when she discussed her lived experience.

> *My experience so far working with other researchers in the team and other projects that I'm on, is they get a real shock if I do reference my lived experience, and they get really uncomfortable about it… I was curious to observe that my colleagues initially did not know how to respond to me when I shared parts of my lived experience related to our project. Further, that they felt they needed to acknowledge my value as coming from my clinical experience, rather than my lived experience. Almost as though they wanted me to see I was more than my lived experience. In time this team has responded very positively when I have shared lived experience and go out of their way to seek my lived experience input now they know me more.*

It is also important that lived expertise researchers are fully empowered to choose how, when, and what they share of their personal experiences. One individual appreciated that academic researchers supported her in choosing what and when to share.

> *We've all had parts in our lives and different jobs where we've had to hide it… And I think it's, it's unusual for me to actually have a group of people that encourage you to sort of let go a little bit, and I struggle to, to let my walls down… [But] it's also being really respected about what I share and don't share… I've been able to sort of weave my lived experience and mental illness into the conversations in a way that I'm comfortable.*

Therefore, academic researchers may need more support to discuss mental ill-health openly and safely and must understand how to give lived expertise researchers agency and flexibility around sharing their personal stories.

Creating More Accommodating Workplaces for All Researchers

Finally, these discussions caused our lived expertise researchers to reflect on the need for research institutions to become more accommodating workplaces generally. Provisions such as more flexible working arrangements, as well as increased training and support, helped our team to successfully contribute to our research and have meaningful work experiences. Yet, these accommodations should be made more routine across academic workplaces (and all other workplaces) to ensure that research benefits from a diverse range of voices with all types of backgrounds and skills. In addition, 2 lived expertise researchers provided reflections on how workplace accommodations had helped them to succeed.

> *Navigating work responsibilities, medical appointments and keeping on top of my mental health can be overwhelming and has impacted my performance at times. Having a manager who supports me, allows the flexibility to work from home, attend mental health appointments and provides a safe space to talk openly has been crucial for my success in this role.*

> *I currently have a flexible working arrangement, working between 5 – 10 hours per week on a casual basis. I feel incredibly supported with my workload and am able to work within my capacity. This allows me to maintain a strong work life balance, manage my mental and physical health, as well as pursue my studies. I also have frequent supervision meetings, as well as peer support meetings, which allow me to properly manage my workload, and learn from other lived experience team members.*

In conclusion, our experiences demonstrate that it is not only feasible for research institutions to make better workplace accommodations but also an important step to ensuring that the skills and experiences of all individuals from all backgrounds are being used to their full potential.

Summary

Taken together, this paper summarizes our experiences of integrating 5 young people with lived expertise in mental ill-health across a range of digital mental health research projects. These steps include (1) creating accepting work cultures and addressing academic ableism, (2) providing diverse opportunities for involvement, (3) giving young people agency and flexibility around when and how to share lived experiences, and (4) creating accommodating work environments for all researchers. These reflections have important implications for youth mental health research, as they demonstrate key steps that academic researchers can take to enhance the value of lived

XSL•FO
RenderX

expertise research, particularly for the young people involved. Textbox 1 provides a summary of why such reflections are important from the perspective of one of our lived experienced researchers.

Textbox 1. Commentary on findings from a lived expertise researcher (SJH).

Our paper highlights that cultural and systemic changes are necessary for research institutions to improve inclusivity and safety for all individuals. The current definition of a "researcher" is someone from science, academia, or clinical disciplines. Inherently, these disciplines come from a knowledge or expertise rooted in education and training. Comparatively, our understanding of the research subject—the "researched"—is someone having lived and living experience of mental ill-health. In their role, lived, and living experience, individuals are solely there to share their experiences on a consultation or advisory basis.

Equal to knowledge-based expertise, individuals with lived and living experience have lived experience-based expertise that pertains to mental health. Therefore, a new understanding—redefinition—is needed to acknowledge the value that individuals with lived and living experiences hold as researchers and not just the researched, or paradoxically in some cases, a combination of both. To address this change in researcher credentials, clarity of roles, expectations, and role extent is needed.

A key reason for this change is that prejudices, pay inequities, oversight of qualifications, and dismissal of lived expertise researchers ultimately cap career progression for many. The previously mentioned definitions of a researcher have shaped a surreptitious culture of academic bias and confusion to value, expectations of experience remit, and enrichment of research lived-expertise researchers bring [7]. Commonly lived expertise researchers experience dismissal, inadequacy, and lack of qualifications due to this bias and lack of clarity, often oversharing at their own expense to vie for their worth [7,16]. Sharing one's lived and living experience daily is emotionally burdensome; without modification of the current culture, this potential of oversharing can impact a lived expertise researcher's mental wellness [7,31].

Institutionally, this must change. Due to the size of academic institutions (like universities), a combined approach of top-down and bottom-up would be needed for this to happen. An example of implementable, bottom-up approaches is that responsibilities held by supervisors and academic staff foster the current culture and, consequently, are accountable for social and systemic academic ableism. By facilitating opportunities to upskill and build research skill sets, supervisors enable further career progression for lived expertise researchers due to increased abilities and responsibilities.

A key strength of this work is that young people have been able to compare and contrast the benefits and disadvantages of different types of engagement and, therefore, to provide important recommendations about how to enhance the benefits and mitigate the risks of lived expertise involvement in youth mental health research. While much has been written about the impacts of lived expertise involvement, recent meta-analyses found that young people were underrepresented in research as, unlike adult populations, they are generally not involved in planning stages, and their involvement is often restricted to working groups or advisory panels [11,13]. Generally, academic researchers have sole decision-making power when identifying how to enhance the value of involving young people in research. Meanwhile, young people themselves lack opportunities to critically reflect on ways to improve the value of their engagement. By contrast, this paper included perspectives from 5 young people with lived expertise in mental ill-health who have been involved in a range of different projects at different stages and in a variety of ways. As such, a key contribution to our work is demonstrating how young people can be integrated into a research team and drive more diverse and valuable contributions to research in partnership with academic researchers.

Despite these contributions, our paper has important limitations. This piece reflects our teams' personal perspectives of being involved in digital mental health research and is based on the experiences of 5 individuals. We believe that highlighting the perspectives of individuals with lived expertise is an important step toward improving the accessibility of academic research. However, this approach and limited sample size also impact the generalizability of our reflections. Furthermore, it is important to acknowledge that lived expertise researchers are still employed by our team, and there are inherent risks involved in reflecting openly on their experiences, which may also have biased the perspectives they were willing to provide. As such, we took several steps to mitigate these risks. First, we deliberately designed a group reflective process so that individual researchers were able to share their perspectives without needing to identify themselves and their opinions to academic researchers on the team; all quotes used in this piece are anonymous for the same reason. Second, we have included leading academic researchers as coauthors to show that the perspectives of our lived expertise researchers, including critical perspectives, are fully supported and endorsed by our team. Third, existing guidelines suggest that vulnerable populations may need additional support to increase their level of involvement in research [15,32]. In line with these recommendations, a clinical researcher, who was a clinical psychologist and academic researcher, was involved throughout all stages of planning, data collection, analysis, and write-up to provide support as needed. Finally, we have sought and been given ethics approval by the relevant university human ethics committee to ensure that young people's involvement in our research is conducted safely.

In conclusion, the benefits of being involved in research for our team were that they felt valued for their experiences rather than stigmatized, they felt accepted by academic researchers, they were able to contribute a variety of skills and to develop others, and they were given accommodations when needed to maintain a healthy work-life balance. This, in turn, highlights important steps that researchers can take to ensure that academic work and young people themselves are benefitting fully from lived expertise involvement in research.

XSL•FO
RenderX

Acknowledgments

The funding sources of this study have had no input into the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. This work was supported by the Medical Research Future Fund Applied Artificial Intelligence in Health Care grant (MRFAI000097). CG was supported by The Ainsworth 4 Foundation. SM was supported by the Cottle Family Fellowship in Youth Mental Health. IBH was supported by a National Health and Medical Research Council Research Fellowship (51192). FI was supported by the Bill and Patricia Richie Foundation. We thank all the young people who participated in this study and all the staff in the Youth Mental Health Team at the University of Sydney's Brain and Mind Centre, past and present, who contributed to this work.

Authors' Contributions

SJH and SM were involved in conceptualization, design, data collection, analysis, interpretation, and writing. AH, CG, ZDH, and JB were involved in conceptualization, analysis, interpretation, review, and editing. YS, ES, FI, and IBH assisted with conceptualization, review and editing, supervision, and funding acquisition.

Conflicts of Interest

IBH is the codirector, Health and Policy at the Brain and Mind Centre (BMC), University of Sydney. The BMC operates an early-intervention youth service at Camperdown under contract to Headspace. He is the chief scientific advisor to, and a 3.2% equity shareholder in, InnoWell Pty Ltd, which aims to transform mental health services through the use of innovative technologies. He is funded by an NHMRC Senior Principal Research Fellowship. ES is a Principal Research Fellow at the Brain and Mind Centre at the University of Sydney. She is a Discipline Leader of Adult Mental Health at the School of Medicine, University of Notre Dame, and a Consultant Psychiatrist. She was the Medical Director of the Young Adult Mental Health Unit at St Vincent's Hospital Darlinghurst until January 2021. She has received honoraria for educational seminars related to the clinical management of depressive disorders supported by Servier, Janssen, and Eli-Lilly Pharmaceuticals. She has participated in a national advisory board for the antidepressant compound Pristiq, manufactured by Pfizer. She was the National Coordinator of an antidepressant trial sponsored by Servier.

References

1. National Study of Mental Health and Wellbeing. Canberra: Australian Bureau of Statistics; 2020. URL: https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release
2. Marinucci A, Grové C, Allen K. A scoping review and analysis of mental health literacy interventions for children and youth. School Psychology Review 2022 Feb 08;52(2):144-158. [doi: 10.1080/2372966x.2021.2018918]
3. Koh J, Tng GYQ, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. JPM 2022 Aug 25;12(9):1376 [FREE Full text] [doi: 10.3390/jpm12091376] [Medline: 36143161]
4. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: A systematic review of randomised controlled trials. Children and Youth Services Review 2021 Aug;127:106073 [FREE Full text] [doi: 10.1016/j.childyouth.2021.106073]
5. Hawke LD, Sheikhan NY, Roberts S, McKee S. Research evidence and implementation gaps in the engagement of people with lived experience in mental health and substance use research: a scoping review. Res Involv Engagem 2023 May 11;9(1):32 [FREE Full text] [doi: 10.1186/s40900-023-00442-5] [Medline: 37170357]
6. Hawke LD, Sheikhan NY, Jones N, Slade M, Soklaridis S, Wells S, et al. Embedding lived experience into mental health academic research organizations: Critical reflections. Health Expectations 2022 Aug 23;25(5):2299-2305 [FREE Full text] [doi: 10.1111/hex.13586] [Medline: 35999670]
7. Jones N, Atterbury K, Byrne L, Carras M, Brown M, Phalen P. Lived experience, research leadership, and the transformation of mental health services: building a researcher pipeline. Psychiatr Serv 2021 May 01;72(5):591-593. [doi: 10.1176/appi.ps.202000468] [Medline: 33691492]
8. Vanderhout SM, Bhalla M, Van A, Fergusson DA, Potter BK, Karoly A, et al. The impact of patient and family engagement in child health research: a scoping review. J Pediatr 2023 Feb;253:115-128. [doi: 10.1016/j.jpeds.2022.09.030] [Medline: 36179891]
9. Viksveen P, Bjønness SE, Cardenas NE, Game JR, Berg SH, Salamonsen A, et al. User involvement in adolescents' mental healthcare: a systematic review. Eur Child Adolesc Psychiatry 2021 Jun 05;31(11):1765-1788 [FREE Full text] [doi: 10.1007/s00787-021-01818-2] [Medline: 34089383]
10. Greenhalgh T, Hinton L, Finlay T, Macfarlane A, Fahy N, Clyde B, et al. Frameworks for supporting patient and public involvement in research: Systematic review and co‐design pilot. Health Expectations 2019 Apr 22;22(4):785-801 [FREE Full text] [doi: 10.1111/hex.12888] [Medline: 31012259]
11. Karlsson AW, Kragh-Sørensen A, Børgesen K, Behrens KE, Andersen T, Kidholm ML, et al. Roles, outcomes, and enablers within research partnerships: A rapid review of the literature on patient and public involvement and engagement in health

XSL•FO
RenderX

research. Res Involv Engagem 2023 Jun 15;9(1):43 [FREE Full text] [doi: 10.1186/s40900-023-00448-z] [Medline: 37322525]
12. McCabe E, Amarbayan M, Rabi S, Mendoza J, Naqvi SF, Thapa Bajgain K, et al. Youth engagement in mental health research: A systematic review. Health Expectations 2022 Nov 16;26(1):30-50 [FREE Full text] [doi: 10.1111/hex.13650] [Medline: 36385452]
13. Bevan Jones R, Stallard P, Agha SS, Rice S, Werner‐Seidler A, Stasiak K, et al. Practitioner review: Co‐design of digital mental health technologies with children and young people. Child Psychology Psychiatry 2020 Jun 22;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
14. Arnstein SR. A ladder of citizen participation. JAPA 1969 Jul;35(4):216-224. [doi: 10.1080/01944366908977225]
15. Arunkumar K, Bowman D, Coen S, El-Bagdady M, Ergler C, Gilliland J. Time for a rope ladder? In: Lauria M, Schively SC, editors. Learning from Arnstein's Ladder: From Citizen Participation to Public Engagement (1st ed). United Kingdom: Routledge; 2020:304-324.
16. Beames JR, Kikas K, O'Gradey-Lee M, Gale N, Werner-Seidler A, Boydell KM, et al. A new normal: integrating lived experience into scientific data syntheses. Front. Psychiatry 2021 Oct 29;12:763005 [FREE Full text] [doi: 10.3389/fpsyt.2021.763005] [Medline: 34777064]
17. Richmond L, Caton N, Downs J, Newton A, Devereux-Fitzgerald A, Brooks H. Creating positive experiences of involvement in mental health research. The Lancet Psychiatry 2023 Mar;10(3):220-227. [doi: 10.1016/S2215-0366(22)00430-8] [Medline: 36696910]
18. Davis S, Pinfold V, Catchpole J, Lovelock C, Senthi B, Kenny A. Reporting lived experience work. The Lancet Psychiatry 2024 Jan;11(1):8-9. [doi: 10.1016/S2215-0366(23)00402-9] [Medline: 38101876]
19. Hickie IB, Davenport TA, Burns JM, Milton AC, Ospina-Pinillos L, Whittle L, et al. Project Synergy: co-designing technology-enabled solutions for Australian mental health services reform. Med J Aust 2019 Oct;211 Suppl 7:S3-S39. [doi: 10.5694/mja2.50349] [Medline: 31587276]
20. LaMonica HM, Davenport TA, Braunstein K, Ottavio A, Piper S, Martin C, et al. Technology-enabled person-centered mental health services reform: strategy for implementation science. JMIR Ment Health 2019 Sep 19;6(9):e14719 [FREE Full text] [doi: 10.2196/14719] [Medline: 31538938]
21. Crouse JJ, Carpenter JS, Song YJC, Hockey SJ, Naismith SL, Grunstein RR, et al. Circadian rhythm sleep-wake disturbances and depression in young people: implications for prevention and early intervention. Lancet Psychiatry 2021 Sep;8(9):813-823. [doi: 10.1016/S2215-0366(21)00034-1] [Medline: 34419186]
22. Lee GY, McKenna S, Song YJC, Hutcheon A, Hockey SJ, Laidler R, et al. Strengthening mental health research outcomes through genuine partnerships with young people with lived or living experience: A pilot evaluation study. Health Expectations 2023 May 17;26(4):1703-1715 [FREE Full text] [doi: 10.1111/hex.13777] [Medline: 37195575]
23. Killackey E. Lived, loved, laboured, and learned: experience in youth mental health research. The Lancet Psychiatry 2023 Dec;10(12):916-918. [doi: 10.1016/S2215-0366(23)00270-5] [Medline: 37625427]
24. Smits DW, van Meeteren K, Klem M, Alsem M, Ketelaar M. Designing a tool to support patient and public involvement in research projects: the Involvement Matrix. Res Involv Engagem 2020 Jun 16;6(1):30 [FREE Full text] [doi: 10.1186/s40900-020-00188-4] [Medline: 32550002]
25. Hickie IB, Iorfino F, Rohleder C, Song YJC, Nichles A, Zmicerevska N, et al. EMPOWERED trial: protocol for a randomised control trial of digitally supported, highly personalised and measurement-based care to improve functional outcomes in young people with mood disorders. BMJ Open 2023 Oct 11;13(10):e072082 [FREE Full text] [doi: 10.1136/bmjopen-2023-072082] [Medline: 37821139]
26. McKenna S, Hutcheon A, Gorban C, Song Y, Scott E, Hickie I. Youth perspectives on 'highly personalised and measurement‐based care': qualitative co‐design of education materials. Health Expectations 2024 Jul 08;27(4):e14137 [FREE Full text] [doi: 10.1111/hex.14137] [Medline: 38976368]
27. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
28. Mills J, Bonner A, Francis K. The development of Constructivist Grounded Theory. Int J Qual Methods 2006 Mar 01;5(1):25-35. [doi: 10.1177/160940690600500103]
29. Mellifont D, Smith-Merry J, Dickinson H, Llewellyn G, Clifton S, Ragen J, et al. The ableism elephant in the academy: a study examining academia as informed by Australian scholars with lived experience. Disability & Society 2019 Apr 23;34(7-8):1180-1199. [doi: 10.1080/09687599.2019.1602510]
30. Sheikhan NY, Kuluski K, McKee S, Hiebert M, Hawke LD. Exploring the impact of engagement in mental health and substance use research: A scoping review and thematic analysis. Health Expectations 2023 Jun 06;26(5):1806-1819 [FREE Full text] [doi: 10.1111/hex.13779] [Medline: 37282732]
31. Orlowski SK, Lawn S, Venning A, Winsall M, Jones GM, Wyld K, et al. Participatory research as one piece of the puzzle: a systematic review of consumer involvement in design of technology-based youth mental health and well-being interventions. JMIR Human Factors 2015 Jul 09;2(2):e12 [FREE Full text] [doi: 10.2196/humanfactors.4361] [Medline: 27025279]
32. van Ham CR, Burgers VWG, Sleeman SHE, Dickhout A, Harthoorn NCGL, Manten-Horst E, et al. A qualitative study on the involvement of adolescents and young adults (AYAs) with cancer during multiple research phases: "plan, structure,

XSL•FO
RenderX

and discuss". Res Involv Engagem 2022 Jul 08;8(1):30 [FREE Full text] [doi: 10.1186/s40900-022-00362-w] [Medline: 35804443]

Edited by N Cahill; submitted 12.12.23; peer-reviewed by I Bell, J Povey; comments to author 13.02.24; revised version received 03.03.24; accepted 21.08.24; published 18.10.24.

Please cite as:
Brogden J, de Haan Z, Gorban C, Hockey SJ, Hutcheon A, Iorfino F, Song YJC, Scott E, Hickie IB, McKenna S
Enhancing Research Involvement of Young People With Lived Expertise: Reflecting on Experiences in Digital Mental Health Research
J Med Internet Res 2024;26:e55441
URL: https://www.jmir.org/2024/1/e55441
doi:10.2196/55441
PMID:

©Josephine Brogden, Zsofi de Haan, Carla Gorban, Samuel J Hockey, Alexis Hutcheon, Frank Iorfino, Yun Ju C Song, Elizabeth Scott, Ian B Hickie, Sarah McKenna. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Video Abstracts in Research

Sophie Nachman[1,2], MSIS, MPH; Esteban Ortiz-Prado[3], MD; Joseph D Tucker[1,4], MD, PhD

[1]Institute for Global Health and Infectious Diseases, Chapel Hill, NC, United States
[2]UNC Gillings School of Global Public Health, Chapel Hill, NC, United States
[3]One Health Research Group at Universidad de las Américas, Universidad de las Américas, Quito, Ecuador
[4]Clinical Research Department, London School of Hygiene and Tropical Medicine, London, United Kingdom

Corresponding Author:
Joseph D Tucker, MD, PhD
Institute for Global Health and Infectious Diseases
130 Mason Farm Road
2nd Floor
Chapel Hill, NC, 27599
United States
Phone: 1 9199662536
Email: jdtucker@med.unc.edu

Abstract

Video abstracts can be useful in health research. A video abstract provides key messages about a research article and can increase public engagement, spark conversations, and may increase academic attention. A growing number of open source software programs make it easier to develop a video abstract. This viewpoint provides practical tips for creating a video abstract for health research.

(J Med Internet Res 2024;26:e64221) doi:10.2196/64221

KEYWORDS

video abstract; abstract; dissemination; public engagement; online; videos; public audience; communication; infographics; health literacy; patient education; public health

Introduction

Video abstracts can breathe life into health research studies, adding a human dimension to what would otherwise be a dry text. A video abstract is a brief video that provides several key messages about a research article. Video abstracts catalyze public engagement, increase the number of manuscript views, and may increase manuscript citations [1,2]. A growing open science movement provides a strong rationale for developing a video abstract as part of a health research manuscript. Video abstracts provide an opportunity for research participants to codevelop dissemination materials and tailor dissemination for patients, policymakers, and others [3]. Video abstracts in health research are often more appropriate for public audiences compared to text manuscripts [1]. Communicating scientific research in an easy-to-understand way is crucial to improve shared decision-making, trust in health institutions, and health literacy. Many medical journals not only allow video abstracts but encourage them to broaden research impact.

Despite the compelling reasons to develop video abstracts, most health researchers have never developed one. Many researchers are not familiar with video creation, lack open source resources, and have limited training in public engagement. The process of developing a video can be challenging, even among researchers who have communications support. Many online resources for creating videos are paywalled or require software. The purpose of this viewpoint is to review practical, open source strategies for developing video abstracts with a focus on expanding research accessibility for a broad public audience.

Developing a Video Abstract

Several strategies can help researchers develop their first video abstract. Video abstracts should provide the most important points of the research. A strong video abstract uses plain language [4], communicating the message in a way that a broad public audience can easily understand without reading the accompanying text. One study analyzing COVID-19 videos found them easier to understand and more actionable compared to websites and infographics [5]. Consulting health literacy guidelines and evaluation tools such as the Patient Education Materials Assessment Tool [6] or the Centers for Disease Control and Prevention (CDC) Clear Communication Index [7] can help confirm that content is clearly communicated.

XSL•FO
RenderX

Developing Inclusive Videos

Accessibility guidelines provide specific strategies for reaching a wider public audience that includes diverse individuals. The Web Accessibility Initiative provides many practical tips for creating accessible audio and video content (Multimedia Appendix 1) [8]. For example, adding captions and providing a transcript are important not only for people with limited hearing but also for people with limited English proficiency [8]. Open source software provides initial drafts of captions that can be revised. Additionally, ensure color contrast in the visual design is sufficient and include narration and audio descriptions for the visuals to provide useful context for the audience.

Organizing the whole research team is essential for creating a compelling video abstract (Table 1). Researchers should consult their local libraries (university and local) to find out what video creation resources are available. Video abstracts are an opportunity for cocreation with community partners who are often involved in research as participants, funders, or evaluators but excluded from research dissemination [3]. Since patients or consumers of health research may not have access to the full-text publication, engaging the intended audience in video creation can increase comprehension and relevance. Cocreation has been a valuable strategy for developing relevant video messages related to consumer health, health literacy, and public health interventions [3,9]. When planning your video abstract, consider who in the community may have creative, artistic, or technological skills to contribute. Also consider whose voice is missing from the manuscript and who is best suited to deliver the message to the intended audience. Aiming for a brief duration (less than 10 minutes) is important for keeping the viewer's attention. Open source teleprompter software can help to facilitate a smooth delivery.

Table 1. Considerations for creating video abstracts.

Category	Considerations
Setting	• Find a quiet space with good lighting.
Audio	• Use an additional microphone to enhance audio quality.
Filming	• Use a tripod for steady filming.
Teleprompter software	• Use a teleprompter software application to enhance delivery.
Iterate	• Solicit and incorporate audience feedback.
Visuals	• Use clear, concise visuals or infographics.
Accessibility	• Include captions or subtitles; provide a transcript.
Length	• Aim for a brief length (less than 10 minutes)
Open source software	• Video editing: Shotcut (Meltytech LLC; Windows, Mac, Linux), OpenShot (OpenShot Studios LLC; Windows, Linux, Mac) • Graphic design: GIMP (Kimball & Mattis, GNU Image Manipulation Program), Inkscape (Inkscape Project; Windows, Mac, Linux) • Audio recording and editing: Audacity (The Audacity Team) • Screen recording: OBS Studio (Bailey, Open Broadcaster Software) • Animation: Blender (Roosendaal) • Storyboarding: Trelby (Gulecha; Windows, Linux), Storyboarder

Approaches to Video Abstract Creation

There are several approaches to choose from depending on the goals, skills, and resources available to your team (Table 2). A do-it-yourself (DIY) 1-person approach can be executed within a few days by recording yourself (or a team member) using a cell phone, Zoom, or creating a presentation in PowerPoint (Multimedia Appendix 2, point 2a). This approach requires no special equipment and is relatively simple. As a result, this approach can be used to develop a video abstract within a shorter period of time (eg, several hours). However, this method centers on you and requires some basic editing skills. Alternatively, a DIY drawing approach involves making a time-lapse video of drawn images or text, which can effectively tell a story and incorporate creative drawings (Multimedia Appendix 2, point 2b). A time-lapse video requires some basic materials, including a storyboard, whiteboard, markers, video recording device, basic editing software, and an elevated surface [3]. At the same time, creating time-lapse videos requires some sketching ability, in addition to video editing skills. For a more team-based approach, you can use open source tools that are available online to develop a video abstract (Multimedia Appendix 2, point 2c). This approach can easily include subtitles and draw on existing templates but does require familiarity with open source programs. Generally, this requires a person with greater experience in video creation and a longer period of development. Finally, artificial intelligence (AI) video generation can be used to develop a brief video. While there are increasing resources available to support AI video generation (eg, RunwayML for

XSL•FO
RenderX

video generation; Mozilla TTS for voice synthesis), this approach does require more technical expertise [10].

Table 2. Approaches and methods for creating a video abstract.

Approach and method	Strengths	Weaknesses
DIY[a] 1-person approach		
Record yourself (or a team member) using a cell phone, Zoom, or PowerPoint	• No equipment required • Relatively simple • Can create a video within days	• Centers on you (or a team member) • Requires some basic editing skills
DIY drawing where you make a video of drawn images and text [3]	• Easier to tell a story • Can incorporate creative drawings • Can create a video within days	• Some basic equipment needed • Requires sketch drawing skills
Team-based approach		
Animation using online tools	• Easy to integrate subtitles and text • Use of templates and defaults	• Requires open source software[b] for animation • May take a longer period of time
AI[c] video generation	• Easy to develop a brief video of an avatar speaking	• Requires open source software[d] and greater technical expertise

[a]DIY: do-it-yourself.

[b]There are several open source animation programs, including Synfig Studio and Blender.

[c]AI: artificial intelligence.

[d]Open source resources can help with AI video generation (RunwayML) and AI voice synthesis (Mozilla TTS).

Recommendations for Developing a Video Abstract

We have several recommendations for those interested in developing their first video abstract. First, creating a draft storyboard is an essential first step in developing a video. Storyboards are visual representations of how the video will progress, frame by frame. This can help the researcher create a list of necessary materials and identify who else needs to be involved in the video development. Second, the storyboard can be turned into a detailed script that conveys key points. Third, iteratively recording video and audio can help to develop a compelling video.

Conclusion

Video abstracts provide a unique opportunity to broaden research dissemination, meaningfully engage communities, and provide a human touch within research. By concisely explaining research in a plain and accessible way, video abstracts can make the results and implications of health research available to a wider audience, including patients and consumers. Once a video abstract is created, this can be included as a URL in supplemental data, uploaded to a video platform or shared drive, or embedded within an online version of a journal. Case studies of video abstracts have demonstrated the impact of brief videos within the larger process of research dissemination (Multimedia Appendix 2). Given the brief duration, a compelling video abstract could also be part of a presubmission inquiry or cover letter. Open access resources and simple strategies can empower researchers to create videos with little or no budget. While research is needed to understand the impact of video abstracts on key research outcomes, there is already a strong evidence base supporting their use. Video abstracts provide an exciting opportunity to spark innovation and creativity in the research process.

Acknowledgments

We acknowledge the helpful comments from the participants of the Advanced Course on Diagnostics in Annecy, France.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily represent those of any institutions with which they are currently or have been formerly affiliated.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Open source resources for video creation.
[DOCX File , 15 KB - jmir_v26i1e64221_app1.docx]

Multimedia Appendix 2
Case studies of video abstracts.
[DOCX File , 16 KB - jmir_v26i1e64221_app2.docx]

References

1. Bonnevie T, Repel A, Gravier F, Ladner J, Sibert L, Muir J, et al. Video abstracts are associated with an increase in research reports citations, views and social attention: a cross-sectional study. Scientometrics 2023;128(5):3001-3015 [FREE Full text] [doi: 10.1007/s11192-023-04675-9] [Medline: 37101977]
2. Zong Q, Xie Y, Tuo R, Huang J, Yang Y. The impact of video abstract on citation counts: evidence from a retrospective cohort study of New Journal of Physics. Scientometrics 2019 Apr 20;119(3):1715-1727. [doi: 10.1007/s11192-019-03108-w]
3. Finley N, Swartz TH, Cao K, Tucker JD. How to make your research jump off the page: co-creation to broaden public engagement in medical research. PLoS Med 2020 Sep;17(9):e1003246 [FREE Full text] [doi: 10.1371/journal.pmed.1003246] [Medline: 32925970]
4. What is plain language? Plainlanguage.gov. URL: https://www.plainlanguage.gov/about/definitions/ [accessed 2024-10-28]
5. Mani NS, Ottosen T, Fratta M, Yu F. A health literacy analysis of the consumer-oriented COVID-19 information produced by ten state health departments. J Med Libr Assoc 2021 Jul 01;109(3):422-431 [FREE Full text] [doi: 10.5195/jmla.2021.1165] [Medline: 34629971]
6. Shoemaker S, Wolf M, Brach C. The Patient Education Materials Assessment Tool (PEMAT) and User's Guide (version 1.0). Agency for Healthcare Research and Quality. 2013. URL: https://www.ahrq.gov/health-literacy/patient-education/pemat.html [accessed 2024-10-28]
7. The CDC Clear Communication Index. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/ccindex/index.html [accessed 2024-10-28]
8. Making audio and video media accessible. W3C Web Accessibility Initiative (WAI). 2019. URL: https://www.w3.org/WAI/media/av/ [accessed 2024-10-28]
9. Kerrigan V, Park D, Ross C, Davies J, Ralph AP. Co-design effective health-literacy videos. Lancet 2023 Feb 04;401(10374):343. [doi: 10.1016/S0140-6736(23)00010-7] [Medline: 36739131]
10. Ferreira M, Lopes B, Granado A, Siopa C, Gaspar H, Castro H, et al. Video abstract production guide. Front Commun 2023 Mar 28;8. [doi: 10.3389/fcomm.2023.1060567]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
DIY: do-it-yourself

Edited by A Mavragani; submitted 11.07.24; peer-reviewed by C Bruns, F Frielitz, KK Raja; comments to author 06.09.24; revised version received 11.09.24; accepted 22.10.24; published 04.11.24.

Please cite as:
Nachman S, Ortiz-Prado E, Tucker JD
Video Abstracts in Research
J Med Internet Res 2024;26:e64221
URL: https://www.jmir.org/2024/1/e64221
doi:10.2196/64221
PMID:

©Sophie Nachman, Esteban Ortiz-Prado, Joseph D Tucker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Economics and Equity of Large Language Models: Health Care Perspective

Radha Nagarajan[1], PhD; Midori Kondo[2], MHA, PharmD; Franz Salas[3], MBA; Emre Sezgin[4], PhD; Yuan Yao[5], BSE; Vanessa Klotzman[6], MS; Sandip A Godambe[7], MD; Naqi Khan[8], MD, MS; Alfonso Limon[7], PhD; Graham Stephenson[9], MD; Sharief Taraman[10], MD; Nephi Walton[11], MD; Louis Ehwerhemuepha[7], PhD; Jay Pandit[12], MD; Deepti Pandita[9], MD; Michael Weiss[1], MD; Charles Golden[1], MD; Adam Gold[7], MBA; John Henderson[1], MBA; Angela Shippy[13], MD; Leo Anthony Celi[14], MD; William R Hogan[15], MD; Eric K Oermann[16], MD; Terence Sanger[7], MD, PhD; Steven Martel[7,17], MD

[1]Children's Hospital of Orange County, Orange, CA, United States
[2]Fred Hutch Patient Care, Seattle, WA, United States
[3]Amazon Web Services, Detroit, MI, United States
[4]Nationwide Children's Hospital, Columbus, OH, United States
[5]Amazon Web Services, San Francisco, CA, United States
[6]University of California Irvine, Irvine, CA, United States
[7]Children's Hospital of Orange County, Orange, CA, United States
[8]Amazon Web Services, Seattle, WA, United States
[9]University of California Irvine Health, Irvine, CA, United States
[10]Cognoa LLC, Palo Alto, CA, United States
[11]National Institutes of Health, Bethesda, MD, United States
[12]Scripps Research Translational Institute, La Jolla, CA, United States
[13]Amazon Web Services, Houston, TX, United States
[14]Massachusetts Institute of Technology, Cambridge, MA, United States
[15]Medical College of Wisconsin, Milwaukee, WI, United States
[16]NYU Langone Medical Center, New York, NY, United States
[17]Physicians Specialty Faculty, Orange, CA, United States

Corresponding Author:
Radha Nagarajan, PhD
Children's Hospital of Orange County
1201 W. La Veta Ave
Orange, CA, 92868
United States
Phone: 1 714 997 3000
Email: Radha.Nagarajan@choc.org

Abstract

Large language models (LLMs) continue to exhibit noteworthy capabilities across a spectrum of areas, including emerging proficiencies across the health care continuum. Successful LLM implementation and adoption depend on digital readiness, modern infrastructure, a trained workforce, privacy, and an ethical regulatory landscape. These factors can vary significantly across health care ecosystems, dictating the choice of a particular LLM implementation pathway. This perspective discusses 3 LLM implementation pathways—training from scratch pathway (TSP), fine-tuned pathway (FTP), and out-of-the-box pathway (OBP)—as potential onboarding points for health systems while facilitating equitable adoption. The choice of a particular pathway is governed by needs as well as affordability. Therefore, the risks, benefits, and economics of these pathways across 4 major cloud service providers (Amazon, Microsoft, Google, and Oracle) are presented. While cost comparisons, such as on-demand and spot pricing across the cloud service providers for the 3 pathways, are presented for completeness, the usefulness of managed services and cloud enterprise tools is elucidated. Managed services can complement the traditional workforce and expertise, while enterprise tools, such as federated learning, can overcome sample size challenges when implementing LLMs using health care data. Of the

XSL•FO
RenderX

3 pathways, TSP is expected to be the most resource-intensive regarding infrastructure and workforce while providing maximum customization, enhanced transparency, and performance. Because TSP trains the LLM using enterprise health care data, it is expected to harness the digital signatures of the population served by the health care system with the potential to impact outcomes. The use of pretrained models in FTP is a limitation. It may impact its performance because the training data used in the pretrained model may have hidden bias and may not necessarily be health care–related. However, FTP provides a balance between customization, cost, and performance. While OBP can be rapidly deployed, it provides minimal customization and transparency without guaranteeing long-term availability. OBP may also present challenges in interfacing seamlessly with downstream applications in health care settings with variations in pricing and use over time. Lack of customization in OBP can significantly limit its ability to impact outcomes. Finally, potential applications of LLMs in health care, including conversational artificial intelligence, chatbots, summarization, and machine translation, are highlighted. While the 3 implementation pathways discussed in this perspective have the potential to facilitate equitable adoption and democratization of LLMs, transitions between them may be necessary as the needs of health systems evolve. Understanding the economics and trade-offs of these onboarding pathways can guide their strategic adoption and demonstrate value while impacting health care outcomes favorably.

(J Med Internet Res 2024;26:e64226) doi:10.2196/64226

KEYWORDS

large language model; LLM; health care; economics; equity; cloud service providers; cloud; health outcome; implementation; democratization

Introduction

Overview

The past decade has witnessed unprecedented growth and digitization of *multivariate* and *multimodal* health care data from diverse sources (eg, the electronic health record [EHR], claims, registries, Internet of Things, and molecular) [1,2]. While multivariate data represent data of a given type across a set of entities (eg, text), multimodal data represent distinct types of data (eg, text, image, and genomics) across entities of interest and play a critical role in generating comprehensive patient and population profiles. Multimodal health care data fall under 2 broad categories, namely *structured* (eg, diagnosis codes and the *International Classification of Disease*, ninth and tenth revisions) and *unstructured* data (eg, text and image). About 80% of health care data are unstructured, including text from clinical narratives [3,4]. The prevalence of unstructured textual clinical data is perhaps a primary motivating factor behind the continued evolution and adoption of natural language processing (NLP) [5] approaches for gaining novel insights from these datasets [6-8]. More recently, advanced machine learning techniques, such as deep learning (DL) [9-11], large language models (LLMs) [12,13], and foundation models, have accelerated these efforts with enhanced capabilities in deciphering patterns from unstructured data [14-16]. Multimodal health care data are usually extracted, transformed, and loaded (extract, transform, and load [ETL]) from diverse source systems into an enterprise data warehouse (EDW; Figure 1). Several variants (eg, Data Lake house) have also been proposed [17]. Subsequently, textual data from EDW are retrieved in a context-specific manner for downstream analytics and ingestion by LLMs (Figure 1). LLM implementations are governed by needs as well as affordability. This perspective discusses factors that impact LLM implementation and proposes 3 broad LLM onboarding pathways for its equitable distribution and adoption.

Multimodal digital footprints (Figure 1) capture unique characteristics of a given population with the potential to assist in decision-making in an evidence-based and data-driven manner, impacting outcomes and key performance indicators (KPIs). These outcomes typically fall under 3 broad categories (*clinical*, *operational*, and *financial*) that are not necessarily independent. For instance, data-driven approaches that can improve preventive care use can minimize aggressive disease-impacting clinical outcomes. Improved clinical outcomes can enable optimal resource allocation impacting operational outcome, reducing the economic burden on the patient, provider, as well as the payer impacting financial outcome. Therefore, the outcomes are represented by bidirectional arrows in Figure 1. While there is considerable excitement over the transformative potential of LLMs in health care [18], it is accompanied by significant *economic* challenges impacting their equitable distribution across health care organizations, especially those that serve economically disadvantaged communities. The choice of a particular LLM implementation pathway is dictated by the *needs* as well as *affordability*. In this perspective, 3 different LLM implementation pathways (training from scratch pathway [TSP], fine-tuned pathway [FTP], and out-of-the-box pathway [OBP]; Figure 1) across 4 major cloud service providers (CSPs; Amazon Web Services [AWS], Google Cloud Platform [GCP], Azure: Microsoft, and Oracle Cloud Infrastructure [OCI]) are presented as “onboarding points.” The risks, benefits, and economics of these pathways are presented and expected to assist in choosing a pathway and subsequent migration across pathways.

XSL•FO
RenderX

Figure 1. Essential ingredients for equitable distribution of large language models (LLMs) comprising 3 interconnected components: (1) digital readiness, (2) infrastructure workforce and privacy, and (3) LLM onboarding pathways (training from scratch pathway [TSP], fine-tuned pathway [FTP], and out-of-the-box pathway [OBP]) and applications to impact health care outcomes and key performance indicators (KPIs) in health care settings. AI: artificial intelligence; API: application programming interface; EDW: enterprise data warehouse; ETL: extract, transform, and load; MRI: magnetic resonance imaging.

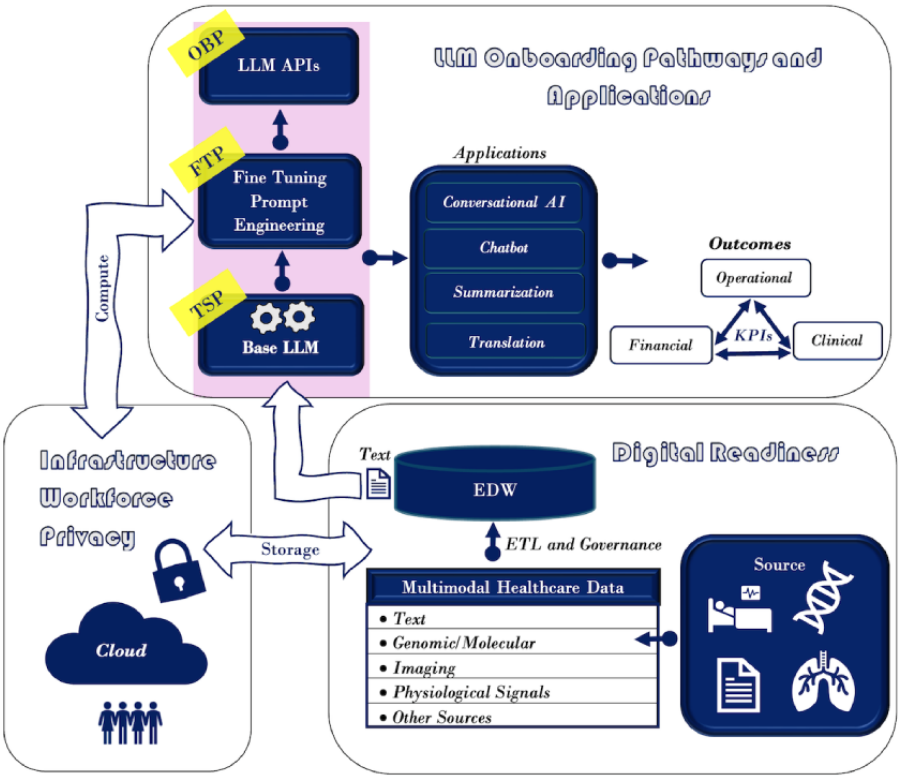

Artificial Intelligence

Operational definition of artificial intelligence (AI) relies on the Turing test, which emphasizes the ability of computers to imitate humans in performing certain tasks [19]. These include automated reasoning, machine learning (ML), and NLP [19]. The hierarchical relationship of AI, ML, DL, and LLMs is shown in Figure 2. In contrast to classical statistical hypothesis testing, AI, ML, or DL assist in discovery, hypothesis generation, and validation in an evidence-based and data-driven manner, with LLMs exhibiting emergent abilities. ML is a branch of AI with a focus on the ability of machines to learn patterns from experience for a given task in an automated manner and draw inferences on previously unseen instances. Popular ML approaches with health care applications include supervised learning, unsupervised learning, reinforcement learning, and association mining [20,21]. DL [22-25] is a subfield of ML that specifically uses neural networks with multiple hidden layers to capture patterns of varying resolution in the given data. Unlike traditional ML, model parameters in deep neural networks (DNNs) can be considerably large with significant storage and computational demands. For example, DL models, such as transformers [26], an essential ingredient of LLMs, have billions of parameters [27]. The layered architecture of DNNs has also been shown to support "transfer learning," where an existing neural network trained on a large dataset can be reused as a base network for predictions on related datasets by retraining only a subset of the layers. This is in stark contrast to traditional, shallow ML models where all model parameters may be altered upon retraining. Recently, proposed foundation models [5,14], exploit the transfer learning ability of DNN, where a "base" model trained on large multimodal data is subsequently adapted for various downstream tasks. LLMs use transfer learning in conjunction with augmented intelligence for enhanced performance, as discussed in the subsequent sections.

XSL•FO
RenderX

Figure 2. Augmented intelligence along with hierarchical representation of artificial intelligence (AI), machine learning (ML), deep learning (DL), and large language model (LLM) is shown in panels (A) and (B), respectively. The primary driver is shown to the right of the inequality in each of the panels.

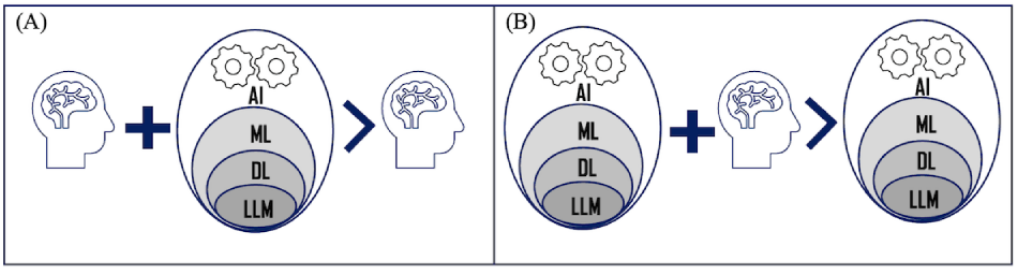

Augmented Intelligence

Augmented intelligence emphasizes the importance of AI and humans working in concert for enhanced performance and generalization ability (Figure 2). The fundamental theorem of biomedical informatics by Friedman [28] emphasized the importance of information resources assisting domain experts by complementing the knowledge of the domain expert. In the present context, it essentially addresses the question, "Can domain experts in partnership with an AI resource lead to better insights than those unassisted?" (Figure 2A). As noted earlier (Figure 1), health care data are typically large, high-dimensional, multimodal, ingested from diverse sources and evolving rapidly, challenging manual interpretation. Therefore, AI models can assist in gaining novel insights from these datasets in an evidence-based manner while validating what is already known. The primary driver in Figure 2A is the domain expert with AI assisting the knowledge discovery process. On a related note, AI models implicitly subscribe to optimizing a chosen objective function and may converge prematurely to a local optimum. Several factors dictate the convergence aspects of these models [29]. This raises the question, "Can AI models with feedback from domain experts perform and generalize better than those unassisted?" (Figure 2B). More specifically, the role of the domain expert can assist in narrowing the set of variables in a context-specific manner, significantly reducing the search space of potential solutions and improving the performance and generalization ability while imposing necessary guardrails for optimal performance. Unlike Figure 2A, the primary driver in Figure 2B is AI, with the domain expert assisting in the knowledge discovery process. LLMs incorporate human feedback (reinforcement learning with human feedback) in minimizing bias, toxicity, hallucinations with improved performance, and generalization ability [30].

LLMs in Health Care

Factors Impacting LLM Implementation

This section discusses three critical factors accompanying successful LLM implementations and deployments in health care settings: (1) *digital readiness*, (2) *infrastructure and workforce*, and (3) *privacy, ethics, and regulatory aspects* (Figure 1). These factors are closely related to the analytics maturity of a health care organization (eg, Healthcare Information and Management Systems Society Adoption Model for Analytics Maturity) [31], and their impact varies across the 3 LLM implementation pathways (TSP, FTP, and OBP).

Digital Readiness

Typically, LLMs (eg, LLaMA [Meta AI], GPT-4 [OpenAI], Med PALM-2, and Claude) are trained on massive amounts of data (eg, TB) integrated from multiple data sources, including those available publicly [32,33]. Large datasets used to train LLMs often lead to sizeable models with billions of parameters with a direct impact on their performance [27], while marking the transition from language models to LLMs with emergent abilities. So, access to large digital health care data (Figure 1) is critical in developing LLMs with superior performance. However, accessing sensitive health care data (eg, protected health information [PHI] and personally identifiable information) to train LLMs poses significant privacy and security challenges. Health care data, such as clinical narratives, are primarily governed by regulations, such as Health Insurance Portability and Accountability Act (HIPAA) [34] in the United States and the General Data Protection Regulation (GDPR) [35] in Europe. HIPAA establishes national standards for the protection of individually identifiable health information by covered entities and their business associates. Similarly, GDPR, although a European Union regulation, is designed to protect the privacy of European Union citizens and residents and applies to all organizations regardless of location. Both HIPAA and GDPR impose clear regulations on the release and sharing of health care data, with civil and criminal penalties for violations. Addressing privacy and security challenges associated with accessing sensitive health care data for training LLMs demands a robust data management strategy. Currently, there are 2 key strategies that stand out in this context: *data deidentification* and *confidential computing* [36,37]. Data deidentification involves removing personally identifiable information, both direct and indirect, from datasets to reduce the risk of patient reidentification. This allows the use of clinical data for model training purposes without compromising patient privacy. The deidentification process involves techniques such as removing names, addresses, social security numbers, and other direct identifiers, as well as managing quasi-identifiers such as dates of birth, gender, and medical diagnoses. However, these identifiers could potentially be used in combination with other information to reidentify individuals [38,39]. Therefore, deidentification singularly is not devoid of reidentification risks. The granularity of the data upon adequate deidentification can also significantly impact what could be effectively inferred. A recent study by de Kok et al [40] elucidated some of the challenges and best practices for sharing health care data compliant with GDPR. Subsequently, 4 approaches were proposed [40] for sharing open health care data (*consent*

XSL•FO
RenderX

pseudonymized, *no consent pseudonymized*, *no consent anonymized*, and *no consent cloud*). However, the sensitive nature of health care data sharing has a direct impact on the sample size for training and fine-tuning LLMs. In general, health care datasets are relatively smaller by several orders of magnitude compared with datasets used to train popular general-purpose LLMs. Recently, Jiang et al [12] pretrained their LLM (NYUTron) primarily on health care data comprising 4,112,249,482, nearly 4 billion words, resulting in a 109-million-parameter model. In a related yet independent study, Yang et al [41] pretrained their LLM (GATORTron) primarily on health care data comprising 82 billion deidentified clinical words, resulting in 3 different LLM models of varying sizes (baseline: 345 million parameters, medium: 3.9 billion parameters, and large: 8.9 billion parameters). It might not be surprising to note that the size of these LLMs was markedly smaller than general-purpose LLMs (eg, LLaMA: 65 billion, GPT-3: 175 billion, and Google Pathways Language Model: 540 billion) [27] by several orders of magnitude. Empirical studies on the scaling behavior of LLMs reported improved performance and emergent behavior with increasing number of parameters (ie, size of the model) [42,43], size of the data, and compute time [44]. However, sample size challenges are likely to persist for health care datasets. Recent efforts in boosting the sample size of the training data from LLMs have explored supplementing health care data from EDW with biomedical text from sources, such as PubMed [41]. Approaches, such as federated learning (FL) [45-48] can also assist in overcoming some of the sample size-related challenges. However, FL could be susceptible to data leakage, breach, and fair and equitable data representation [49].

Multimodal health care data (eg, text, images, and signals) are usually integrated from diverse enterprise source systems (eg, electronic medical record, registries, the Internet of Things, next-generation sequencing, and Picture Archiving and Communication System) and reside in a centralized EDW (Figure 1). EDWs support querying, reporting, and enterprise analytics. Data governance and ETL are essential ingredients of EDW implementation, dictating the quality and granularity of the data ingested by downstream analytics tools, such as LLMs, impacting their performance. Data governance and ETL can also exhibit marked variations across health care organizations attributed to several factors, including variations in source systems and business processes. More importantly, EDWs demand upfront investment, continued support from executive leadership, and an existing, evidence-based culture. EDW economics is governed by several factors, including (1) architecture of EDW (eg, Inmon [50] and Kimball and Ross [51]); (2) source systems; (3) type, velocity, and volume of data; (4) enterprise data governance and ETL processes; (5) infrastructure and workforce supporting secured storage and retrieval; and (6) existing culture of data-driven and evidence-based approaches in impacting outcomes and KPIs in the health care organization. It is important to note that ETL and EDW can exhibit marked variations across organizations, posing challenges in seamless sharing of health care data, deployment of AI and ML models, and demonstrating their generalization ability. Common data models, such as the Observational Medical Outcomes Partnership, have been helpful in data standardization, data sharing, and federated querying [52,53]. However, there are inherent limitations to standardization, including incompleteness in data models and terminologies resulting in data that cannot be mapped [54], errors in mapping [55], and the potential loss of information due to granularity mismatches between the source data and the standard. In addition, data standards require significant expertise and a steep learning curve in their absence. The challenges mentioned earlier can especially be accentuated across health care organizations that primarily serve the economically disadvantaged, underserved, and marginalized communities. Factors contributing to data inequities [56] are multidimensional and include ethnicity (eg, Hispanic), race (eg, African American), disease groups and treatment regimens (eg, rare disease), gender or gender identity (eg, lesbian, gay, bisexual, transgender, and queer [LGBTQ]), age (eg, pediatric population), geographic location (eg, rural areas), language barriers (eg, Spanish), digital divide, and patient literacy [57,58]. It is of interest to note that these dimensions are not mutually exclusive, and their combination can significantly impact the representation of a given population in the data and digital readiness of organizations that serve these communities. Inequity along these dimensions can lead to potential biases [59,60]. These biases broadly fall under *systemic*, *statistical*, *computational*, and *cognitive* bias [61]. Systemic bias can be further categorized into *measurement bias*, *missing validation bias*, *label bias*, and *modeling bias* [61]. Measurement bias [62,63] may be the result of variations in quality and representation of the entities of interest across subpopulations in the data. Missing validation bias may result due to a lack of adequate validation studies across certain populations. It is important to note that measurement bias may accentuate validation bias because validation at small sample sizes can be statistically challenging. Label bias may be a result of surrogate variables substituting the actual health care outcomes of interest. Modeling bias is attributed to the biased results generated by a specific model. Computational and statistical biases can be a result of inadequate representation of select groups and populations in the given data. Human cognitive bias is a bias due to human perception of AI and ML systems. Mitigating human cognitive bias may be critical for successful adoption of AI systems. From this discussion, digital readiness is critical for health care organizations to embark on their LLM journey and could dictate the choice of the LLM implementation pathway.

Infrastructure and Workforce Needs

Infrastructure Needs

LLMs primarily rely on the transformer architecture [26], with the ability to be trained in a massively parallel manner on sequential data. However, parallel processing in turn demands specialized hardware accelerators, such as graphics processing units (GPUs), and seamless interfaces between various hardware components in conjunction with significant network bandwidth for avoiding storage, data transfer, latencies, and bottlenecks (Figure 1). This is especially true when the training involves large amounts of data and pretrained LLMs have billions of parameters. Parallelism falls under 2 broad categories, namely (1) *data parallelism* [64] and (2) *model parallelism* [65]. Data

XSL•FO
RenderX

parallelism addresses challenges with large training datasets that cannot fit within a single GPU by partitioning and distributing them across multiple GPUs. This is especially helpful because optimization techniques, such as stochastic gradient descent, used by LLMs rely on small batches of data that can be distributed across GPUs. Model parallelism addresses challenges with the size of the LLMs (eg, billions of parameters) by distributing them (eg, weights and layers) across multiple GPUs and further subdivided into *pipeline parallelism* and *tensor parallelism*. While pipeline parallelism facilitates the distribution of the layers of a DNN [66], tensor parallelism [67] distributes the tensor computation across hardware accelerators. While partitioning the data and model is useful in overcoming some of the challenges with their size, fast interconnect between the GPUs is especially critical for enhanced communication between them in a cluster (eg, DGX A100 Data Center) and between clusters (eg, Super PODs) through specialized high-bandwidth (eg, 900 GB/s) communication links, such as NVLink/NVLink Switch. Training LLM models using multiple GPUs is also accompanied by significant carbon footprint and heating, demanding specialized cooling systems for optimal performance [68]. While LLM implementations are traditionally dependent on multiple software libraries [69], there has been recent interest in developing graphical user interfaces for LLMs to alleviate some of these challenges for the end users [70]. The size of the GPUs required is dependent on several factors, such as the size of the LLM models so that a model can be loaded into the GPU successfully while permitting necessary computation. For instance, LLM models with 7 billion parameters (13 GiB) may need a 16-GiB GPU, such as NVIDIA T4 Tensor Core for processing, while those with 13 billion (25 GiB) and 70 billion parameters (130 GiB) might require a 32 GiB (eg, NVIDIA-A100 GPU Server) and 160 GiB GPUs (eg, NVIDIA 2 × A100 multi-GPU Server), respectively.

Cloud-Computing Platforms

On-premise infrastructure had supported more traditional AI and ML implementations and analytics dashboards in the past. However, increasing size of the data and compute along with evolving needs of health care organizations, demand scalable infrastructure for storage (horizontal scaling) and computing (vertical scaling), with the latter playing a critical role for LLMs. For instance, data and model parallelism demand scalable infrastructure that can vary with training data size as well as the model size. Cloud-computing environments (Figure 1) provide infrastructure as a service and software as a service (SaaS) with pay-as-you-go payment models to address challenges from scalability, privacy, workforce, and economic standpoints. CSPs (AWS, GCP, Azure, and OCI) offer robust solutions for LLM deployment by leveraging their global infrastructures built around *high availability*, *enterprise security and compliance*, *low latency* access to computing resources, and managed services as needed. CSPs can play a critical role in bridging the chasm between needs, such as growth scale and security of information technology infrastructure, and affordability by leveraging existing technology investments on-premises or on other clouds. More importantly, CSPs can help facilitate enhanced democratization and equitable adoption of LLMs by the broader health care communities and not by a privileged few, a critical aspect for the equitable distribution, widespread adoption, and long-term success of LLMs in health care. CSPs also provide hybrid options, such as “cloud bursting,” that allow organizations to use their private cloud, on-premise infrastructure, for routine operations and burst into a CSP temporarily when additional computing resources are needed to handle peak loads and prevent queuing of computational workloads. A hybrid approach enables health care organizations to handle the variable and intensive computational demands of LLMs at scale. By cloud bursting, health care organizations can maintain a cost-effective private cloud for steady-state workloads and then burst into public clouds during periods of high demand without needing to overprovision their private cloud infrastructure while minimizing latencies. CSPs provide essential tools to help organizations manage security and compliance risks. AWS identity access management and Azure controls are examples of services that integrate into existing organizational security investments and aid in configuring fine-grained access controls and ensuring authorized access to sensitive data. Hybrid cloud offerings from AWS, GCP, and Azure also adhere to regulations and compliance (eg, HIPAA) for handling health care data containing PHI through data encryption and network security between on-premises and on-cloud workloads, creating a single flexible, cost-effective enterprise infrastructure technology solution. These environments can also assist in minimizing on-premises workforce needs and advanced skillset needs using readily available cloud-based tools and managed services. This aspect is especially critical for health care organizations that do not have sufficient up-front investment and a history of supporting analytics teams but would like to experiment with the utility of LLMs for impacting outcomes. In essence, CSPs can assist in equitable adoption by starting small with existing resources and scaling out as needed without large upfront investments. Utility services from CSPs can significantly reduce the operational overhead and technical challenges associated with LLM development, testing, and scale, allowing for health care organizations to focus more on outcome-based applications rather than infrastructure management [71]. Serverless options by CSPs for storage and computing are also expected to minimize carbon footprints [68].

Cloud-Based Hardware Accelerators

Hardware accelerators assist in overcoming data and computational bottlenecks working in concert with base processors (eg, central processing units) [72]. Several types of accelerators, such as application specific integrated circuits, field programmable gate arrays, GPUs, and dedicated chips for AI, have been explored [73-76]. Accelerators, such as GPUs, as noted earlier, have become an integral part of DL models and LLMs in overcoming computational bottlenecks. However, there is renewed interest in developing hardware accelerators for LLM training and inference that are affordable. Hardware accelerators offered by CSPs include Trainium and Inferentia (AWS), tensor processing units (TPUs; GCP), Maia (Azure), and MI300X (OCI). These chips are optimized for specific workloads, cost-effective relative to GPUs, and interface to popular open-source environments. However, multiple factors, such as the choice of open-source environment and

XSL•FO
RenderX

machine-learning libraries (eg, Tensorflow and PyTorch), can impact the benchmarking and performance of hardware accelerators [77].

- AWS—Trainium AI accelerator [78] supports training DL models with >100 billion parameters and up to 50% in cost-to-train savings over comparable elastic compute cloud instances. Inferentia [79] supports inference, delivering up to 2.3× higher throughput and up to 70% lower cost per inference than comparable elastic compute cloud instances. Both are supported by the AWS Neuro software development kit, which integrates natively with open-source DL environments such as PyTorch, TensorFlow, and HuggingFace. Inference on Meta's llama 8B would cost US $0.99/h on Inferentia, a savings of >65% on an NVIDIA A10G GPU instance.
- GCP—TPUs [80] are custom-designed AI accelerators optimized for training and inference of AI models [81]. They scale cost-efficiently for a wide range of AI workloads, spanning training, fine-tuning, and inference. TPUs provide the versatility to accelerate workloads on leading AI frameworks, including PyTorch, JAX, and TensorFlow.
- Azure—Maia 100 AI accelerator [82] is supported by the Maia software development kit and interfaces to open-source frameworks such as PyTorch, ONNX Runtime, and Triton from OpenAI. It can support services such as Microsoft Copilot and Azure OpenAI Service.
- OCI—MI300X accelerators are powered by AMD's CDNA 3 architecture, offering memory bandwidth and compute performance supporting a broad range of precision data that enable OCI to support larger and more complex computations for AI and ML workloads [83-85]. Compared with NVidia's H100GPUs, MI300X could provide a cost-effective alternative while still delivering competitive performance with its 304 compute units, 19,456 stream cores, and 1216 Matrix cores.

Cloud-Based Managed Services

Managed services from CSPs can assist in LLM implementation and management while minimizing on-premises workforce needs. Managed services features can be readily accessed through application programming interfaces (APIs), accelerating implementations with an enhanced focus on impacting health care outcomes. This may include access to multiple LLMs across vendors via a single API, enabling experimentation by end users without managing multiple end points, keys, and payloads. This permits experimentation with new LLM models and automates this process via CSP-managed services. These features are especially helpful in exploring the available LLMs that best suit the current needs of the health care organization. In contrast, the exploration phase could be fairly involved across on-premise implementations, accompanied by multiple tokens, API end points, and access controls. CSP-managed LLM services also facilitate centralized governance structures for access management, billing, auditing, and security scanning. Given the sensitive nature of health care data, using fewer end points allows health care organizations to set up the necessary access controls for data and APIs in a seamless manner.

Workforce Needs

LLM implementation and deployment in health care workflows typically demands a workforce with expertise across a spectrum of areas (Figure 1). However, the workforce needs will be dependent on several factors that include (1) LLM implementation pathway, (2) applications of LLM in health care workflows, (3) commercial or open-source platforms. and (4) on-premise or cloud-based implementations. Workforce needs are especially critical across LLM implementations that involve training and fine-tuning. This would ideally consist of a (1) core team comprising data scientists, architects, engineers, and CSPs with expertise in areas such as warehousing, NLP, LLMs, ML implementation, deployment, and assessment (eg, MLOps) [86] in concert with (2) an infrastructure team that addresses storage and computing needs on-premises and on-cloud, (3) subject matter experts and health care personnel who guide the implementation, reinforcement learning, and prompt engineering aspects of LLMs for optimal performance, (4) regulatory and compliance teams for ensuring ethical use of health care data and establishing guardrails, and (5) information technology that assists in ensuring privacy and security of health care data while deploying the LLM applications or API in enterprise workflows to impact outcomes and KPIs (Figure 1). Agile strategies (eg, DevOps or MLOps) may be critical for implementation, validation, and seamless deployment of LLMs in workflows. Given potential bias and toxicity that may accompany LLM implementations, an inclusive framework that incorporates critical feedback from stakeholders, patients, providers, and subject matter experts across diverse communities can minimize bias and assist in developing the necessary guardrails.

Privacy, Ethics, and Regulatory Aspects

Privacy and Security in the Cloud

Privacy and security are critical for storage, retrieval, and analysis of health care data (Figure 1). There has been an increasing shift in moving health care data and analytics from on-premises to the cloud [71]. CSPs provide confidential computing environments (CCEs) that facilitate computations in hardware-based trusted execution environments (TEEs). These ensure sensitive data (eg, PHI data) to remain encrypted in an isolated environment, preventing modification of data and applications by unauthorized parties, including CSPs, during processing and transmission [87,88]. They facilitate secure collaboration of first- and third-party data with the potential to assist in overcoming sample size constraints for LLM training as discussed in this section. CCE capabilities across the 4 major CSPs (AWS, GCP, Azure, and OCI) are discussed in the subsequent sections.

- AWS—Confidential computing capabilities through the processor agnostic AWS Nitro System and AWS Nitro Enclaves, enabling secure isolation of sensitive workloads [89].
- GCP—Confidential virtual machines (VMs) and Confidential Google Kubernetes Nodes allow customers to process sensitive data while keeping them encrypted in memory [90].

XSL•FO
RenderX

- Azure—Confidential computing VMs with AMD SEV-SNP and Intel SGX support ensuring VM-level confidentiality and protection from cloud operators [91].
- OCI—Confidential computing through Oracle's confidential instances leveraging AMD Secure Encrypted Virtualization for VMs and AMD secure memory encryption for bare metal instances protecting data and application processing the data [92].

Attestation

Attestation is a critical component of CCEs that ensures the trustworthiness of the computing environment. It allows the integrity and authenticity of the hardware, software, and configuration to be verified, effectively establishing trust between parties. Attestation offered by major CSPs (AWS, GCP, Azure, and OCI) is discussed in the subsequent sections.

- AWS—CCE attestation involves the use of an attestation document signed by the Nitro Hypervisor. This document is critical for providing the identity of the enclave to AWS Key Management Service (KMS), which validates the document against the KMS key policy. This allows the enclave to perform cryptographic operations with KMS keys [93].
- GCP—Provides attestation through the Binary Authorization and Certificate Authority Service. Through this service, the confidential workload collects measurements of itself and the TEE and sends an attestation request to the Binary Authorization service, which compares the measurements against an attestation policy. If they match, service returns a signed attestation [94].
- Azure—Provides CCE attestation via the Microsoft Azure Attestation service. The confidential workload includes an attestation client that collects measurements and evidence from the TEE. It then sends an attestation request with its evidence to the Microsoft Azure Attestation service that is verified against policy. If valid, it returns a signed attestation token.
- OCI—Provides attestation using a hardware-based trusted security module that generates an attestation report containing measurements of the hardware and firmware environment and verified by the customer to ensure confidential workload is running in a legitimate TEE.

FL Architecture

CSPs also provide FL architectures [47] for decentralized training of LLM addressing sample size challenges. An FL model is trained locally and refined through shared updates, resulting in an aggregated global model without explicit sharing of health care data. Using federated data for training may leverage collective knowledge, perhaps resulting in models with enhanced generalization ability. It is important to note that because only model updates are shared instead of the actual data, FL implicitly minimizes the amount of data transferred over the network. This can be particularly beneficial in scenarios where data transfer is costly or limited by bandwidth constraints. FL frameworks developed by CSPs include:

- AWS—FedML on AWS is an open-source library that supports several FL models. AWS provides tools, libraries, and algorithms to implement and experiment with FL algorithms in a private and secured environment [95].
- GCP—TensorFlow Federated is an open-source framework for ML on decentralized data. TensorFlow Federated is used to implement FL on Google Cloud, leveraging Google Kubernetes Engine for hosting and managing the FL process [96].
- Azure—the AzureML platform supports Azure FL frameworks NVFlare and Flower for running a FL pipeline. Azure's capabilities are leveraged for provisioning and orchestration of FL algorithms [97].
- OCI—Supports FL through various tools, frameworks, and services, such as Private Federated Learning with Domain Adaptation [98].

Confidential FL

While FL ensures compliance with data protection regulations, such as HIPAA and GDPR, it does have some limitations, as noted earlier [49]. A possible solution is to combine confidential computing and FL, resulting in confidential FL (CFL) [99]. This decentralized approach works well for a hybrid cloud environment that spans on-premise data centers, edge devices, and public clouds from different CSPs. Confidential computing TEEs secure the data during processing, while FL enables collaborative training without explicit sharing of health care data. Incorporating a deidentification process as a part of CFL workflow ensures access to large sensitive data for training LLMs without compromising privacy and security while maintaining the integrity of health care data management. Key characteristics of CFL across CSPs are discussed in the following sections.

- Enhanced privacy and security—while deidentification removes identifiable information from the data, CFL ensures that the data are processed in a secure and private manner.
- Compliance or regulations—Deidentification and CFL can help health care organizations comply with HIPAA and GDPR. CFL provides the necessary security measures to protect data in use. Both technologies address regulations regarding data security.
- Facilitation of data sharing—by combining these technologies, health care organizations can safely engage in collaborative data sharing initiatives and develop LLM models with enhanced performance and generalization ability.
- Intellectual property protection—CFL can protect intellectual property such as health care AI algorithms and research data during collaborative training.
- Building trust—secure handling of health care data builds trust among patients, providers, and payers.

While CFL has the potential to accelerate LLM implementation using sensitive federated health care data, some of the challenges listed here need to be addressed for its successful deployment and adoption.

- Data heterogeneity—CFL assumes that the data across participating organizations are independently and identically distributed. However, in practice, health care data may exhibit significant heterogeneity, which can impact the performance of the models. Techniques, such as transfer

XSL•FO
RenderX

learning and domain adaptation, can be used to address this challenge.

- Communication efficiency—the iterative nature of CFL involves frequent communication between the participating organizations and the central data server. Efficient communication protocols and compression techniques are necessary to minimize the communication overhead and ensure scalability.
- Model interpretability—CFL models may lack interpretability due to the distributed nature of the training process. Techniques, such as model distillation and explainable AI can be used to improve the interpretability of CFL models.
- Incentive mechanisms—encouraging health care organizations to participate in CFL initiatives may require appropriate incentive mechanisms. Developing fair and transparent incentive approaches that align with the interests of all stakeholders is an important consideration for success.
- Human in the loop—integrating human expertise and oversight as a part of the LLM training and decision-making process ensures that the models are accurate, reliable, and aligned with human values. Human in the loop also ensures that the models comply with legal and regulatory requirements, such as data protection laws and medical standards.

Ethics and Regulatory Aspects

The potential of AI tools, such as LLM, to transform health care outcomes does come with various ethical and regulatory challenges [100-102]. US president Joe Biden's October 2023 executive order [103] underscored the necessity of ensuring AI safety and security. It mandated AI-generated content to be clearly identified and called for substantial investments in AI-related education, training, and research. The order emphasized protecting intellectual property, supporting American workers, advancing equity, civil rights, while safeguarding privacy and civil liberties. It also directed the Department of Health and Human Services to establish safety parameters for AI, including frameworks for identifying and tracking clinical errors, generating improvement guidelines, and sharing these among health care organizations. A recent Health and Human Services 2024 ruling (Section 1557, Patient Protection and Affordable Care Act) also emphasized protection to patients against bias and discrimination from AI and ML decision support tools and the importance of mitigating such biases [104,105]. In addition, 29 countries attended the AI Safety Summit in November 2023 and signed the Bletchley Declaration [106] to "cooperate on AI to promote inclusive economic growth, sustainable development, and innovation, to protect human rights and fundamental freedoms, and to foster public trust and confidence in AI systems to completely realize their potential." The Institute for Healthcare Improvement's Lucian Leape Institute (LLI) [107] predicted increased use of AI in clinical documentation support, clinical decision support, and patient-supportive chatbots in the health care setting. They recommended prioritizing patient safety, engaging clinicians, ensuring AI efficacy and bias mitigation, establishing AI governance and oversight, and fostering collaborative learning across health systems. In addition, LLI emphasized the importance of AI systems in reporting confidence levels and rationale and the need for continuous human monitoring to maintain trust and accuracy in AI-generated outputs. LLI also suggested several considerations for regulators and policy makers: establishing clear guidelines for ethical and trustworthy AI use, promoting transparency and accountability, supporting AI literacy, incentivizing AI development that prioritizes safety, and facilitating localized decision-making. The European AI Act [108], the first legal framework on AI, categorized AI risks into 4 levels—unacceptable, high, limited, and minimal—and introduced transparency obligations for all AI models. Accreditation agencies, such as the Joint Commission, will need to advocate to create governance structures and processes for monitoring patient safety issues related to AI. The World Health Organization also recently commented on regulatory considerations for AI in health [109]. It outlined essential guidelines covering documentation and transparency, risk management, intended use validation, data quality, privacy protection, and stakeholder engagement. The ethical development of AI must adhere to principles, such as beneficence, nonmaleficence, autonomy, justice, data quality, transparency, fairness, responsibility, privacy, freedom, trust, sustainability, dignity, and solidarity. Trustworthiness of AI-based clinical decision support is often compromised by the lack of transparency in how AI tool's function and the basis of their decisions. There are concerns about the use of proprietary data, the absence of robust regulation, and the risk of bias from datasets that do not adequately represent marginalized populations. Overall, as AI continues to evolve and integrate into health care, maintaining a balance between innovation and ethical responsibility is crucial. Regulatory frameworks and ethical guidelines at health care organizations must evolve to ensure that AI enhances health care delivery while protecting the interests and rights of patients and providers alike.

LLM Guardrails for Responsible AI

There has been interest in developing guardrails and regulatory frameworks to facilitate responsible AI. These guardrails ensure the behavior of AI tools, such as LLM, falls within expected bounds while being resilient to adversarial attacks. These efforts include recent open-source initiatives, such as NeMo Guardrails [110] for improved trustworthiness [111] of LLM conversational systems. These guardrails assist in customizing the LLM interaction with users using *topical rails* and *execution rails* [110]. While topical rails prevent the LLM from veering off topic, execution rails assist in moderating the LLM output and ensure it is factual. A recent study by Meskó and Topol [102] on regulatory oversight of LLMs identified several associated challenges. The study pointed out the challenges in translating existing Food and Drug Administration's regulatory frameworks for medical devices [112] to contemporary AI-based technologies, such as LLMs, and the need for new regulatory frameworks for LLMs. More specifically, it highlighted 2 unique characteristics of LLMs: (1) the ability to adapt their performance to training data as well as tasks in contrast to traditional AI and ML approaches, and (2) the ability to learn in a self-supervised (autodidactic) manner without the need for explicit guidance and ground truth labels as in a more classical supervised ML paradigm. Subsequently, a list of LLM

XSL•FO
RenderX

regulatory challenges were identified (Table 1 in the study by Meskó and Topol [102]), including privacy, intellectual property, medical malpractice liability, quality control and standardization, informed consent, interpretability and transparency, fairness and bias, data ownership, overreliance, and need for continuous monitoring and validation. Some of these challenges have also been acknowledged in a more recent US Food and Drug Administration release with a broader focus on AI and medical products [113].

XSL•FO
RenderX

Table 1. Summary of risks and benefits of the 3 large language models (LLMs) onboarding pathways (training from scratch pathway [TSP], fine-tuned pathway [FTP], and out-of-the-box pathway [OBP]).

	TSP	FTP	OBP
Digital readiness			
Benefits	• LLMs are trained on health care data and capture characteristics of that target population with the potential to impact outcomes in that population. • Enhanced transparency of the data, implementation, and deployment. • Enhanced quality of training data through enterprise governance, minimizing bias.	• Digital readiness of FTP is much lesser than TSP because FTP focuses on fine-tuning as opposed to training.	• Digital readiness for OBP is minimum.
Risks	• Demands upfront investment in data warehousing, enterprise governance, and dedicated workforce.	• Susceptible to biases in training data used in the pretrained LLMs. • General purpose LLMs are often trained on nonhealth care data. • Prompt engineering demands can be significant.	• General purpose, out-of-the-box models pretrained on nonhealth care data may have limited utility, prone to bias, and hallucinations.
LLM			
Benefits	• Train LLM from scratch using either existing transformer architectures or novel architectures. • Long-term maintenance, customization, with evolving needs.	• Uses off-the-shelf pretrained LLMs without explicitly training from scratch. • The number of open-source pretrained LLMs continues to grow.	• Multiple choices of off-the-shelf LLMs accessed as APIs[a]. • Readily accessible with minimal training.
Risks	• Cost of training LLMs can be significant. • Sharing checkpointed LLMs trained on PHI[b] data is a risk. • Novel architectures demand considerable experimentation for optimal performance. May result in implementation and deployment delays.	• Susceptible to biases in the pretrained LLMs. • Pretrained LLMs on nonhealth care data may have performance limitations. • Dependency on pretrained LLMs is a risk. Limited transparency may be a security risk.	• Use is dependent on the features exposed by the vendors. • Generic nature of the output may have limited utility in addressing the unique needs of health systems. No explicit training of the LLM model.
Workforce			
Benefits	• On-premise workforce can assist in customizing LLMs with enhanced transparency and evolving needs.	• Workforce demand is significantly less than TSP.	• Minimal on-premise workforce needs and training. • Rapid implementation using managed services.
Risks	• Demands a skilled workforce with expertise across a spectrum of areas for implementation and deployment. • Demands recruitment, growth, and retention of skilled workforce. • Continued buy-in from leadership for sustaining workforce.	• Limited customization of the pretrained LLMs.	• Complete dependence on vendor models and services with minimal transparency
Infrastructure and security			
Benefits	• Options for training LLMs across cloud service providers with on-demand and spot instance pay-as-you-go pricing models. • Secured cloud environments are available	• Infrastructure needs for compute are significantly less than TSP.	• Infrastructure is needed primarily for inference.

XSL•FO
RenderX

	TSP	FTP	OBP
Risks	• Training LLMs on GPUs[c] is expensive. • Security and governance for training sensitive data in the cloud.	• Vendor pricing may vary based on adoption. • Prompt engineering on pretrained models can be involved.	• Infrastructure costs increase with the number of users. • Vendor pricing may vary with increasing adoption. Availability of services is not guaranteed. • Interfacing to downstream applications.

[a]API: application programming interface.

[b]PHI: protected health information.

[c]GPU: graphics processing unit.

LLM Implementation Triumvirate

Overview

As noted earlier, LLM implementation is dependent on several factors. In this section, 3 broad LLM implementation pathways (Triumvirate) are discussed along with associated risks, benefits, and economics (Figure 1). These pathways are not necessarily independent and expected to serve as onboarding points for equitable distribution and enhanced adoption of LLMs. A summary of the risk and benefits of these 3 pathways is also enclosed in Table 1 for convenience.

The TSP

In TSP, an LLM is trained from scratch using health care data and subsequently customized for specific needs and tasks of the health care organization. FTP and OBP may follow TSP.

Benefits

The TSP provides enhanced transparency of the data and code, complete ownership of the model parameters, implementation, and the ability to assess the quality of the training data at the most granular level. TSP is expected to facilitate long-term maintenance of the model and customization to the evolving needs of the health care organization, including seamless deployment in workflows and interface to enterprise dashboards. In addition to using an established LLM architecture, TSP may also implement novel architectures or modifications to existing general-purpose transformer architectures. Training LLMs using data from the EDW of a health care organization implicitly adheres to the data governance and ETL ensuring high-quality data and accommodating characteristics of the population served by the health care organization that may not necessarily be captured in generic datasets, such as those used to train general purpose LLMs. This in turn is expected to result in a model with enhanced performance [114] addressing the needs of the organization impacting clinical, operational, and financial outcomes and KPIs (Figure 1). Such a model is also expected to be better used while demonstrating value because the training data can significantly impact its behavior and performance [115]. As noted earlier, it is not uncommon for the organizational data to be supplemented by high-quality external data during training [41]. Unlike pretrained LLMs, access to the training data and enhanced transparency may assist in mitigating biases, minimizing perpetuation and amplification of biases, and reducing toxicity by the model as well as downstream applications and APIs that are dependent on the base model (Figure 1), leading to improved overall performance. This aspect is especially critical when deploying the LLM model in clinical workflows to assist in clinical decision-making.

Risks

TSP implicitly demands digital readiness, infrastructure and workforce, and regulatory compliance. Because sample-size challenges can impact TSP, approaches and FL techniques may be explored. The digital and analytics maturity of an organization, along with an existing culture of data-driven and evidence-based approaches to impacting outcomes, may be critical for successful TSP implementation and continued buy-in from the enterprise leadership. Unlike pretrained LLMs, TSP may demand experimentation to identify the optimal model size, parameters, and checkpointing the model before deployment. Agile implementation strategies across multiple teams, such as data science, information technology, clinical, and support from executive leadership may be critical for the timely progress of TSP. Therefore, timelines for TSP implementation and deployment are expected to be significantly larger than FTP and OBP. In addition to implementation and validation, timelines would also include seamless deployment in health care workflows and providing necessary training for the end users. Delays in demonstrating value are to be expected as with any new AI tool. TSP will also demand access to specialized infrastructure for storage and computing, including distributed frameworks and GPU Clusters or PODS. Enterprise CSPs can be critical partners in this regard. Unlike general purpose LLMs, TSPs using cloud infrastructure should follow strict compliance and security protocols that in turn may incur additional costs. Because TSP demands unique skill sets for implementation and critical evaluation, existing skillsets and investment in the workforce in areas such as data science, DL, NLP, LLM, and IT, as well as protected time for health care personnel for critical assessment of the models, would be important. The quality and performance of TSP will be dependent on the knowledge of the subject matter experts assisting with the reinforcement learning with human feedback process. FTP is likely to follow TSP as a part of customizing the LLMs to end users and downstream applications (Figure 1). Given the large number of parameters of LLM models, there is the possibility of the LLM models memorizing some of the information in the training data [116]. This in turn may discourage sharing checkpointed TSP models due to the risk of information leak.

XSL•FO
RenderX

Economics of TSP

Among the 3 pathways, TSP demands considerable up-front investment with regards to digital readiness, infrastructure, workforce, privacy, and regulatory aspects. TSP implementation will demand (1) an existing EDW for querying and retrieval of unstructured data for ingestion by LLMs; (2) a workforce with competencies across a spectrum of areas, including implementation, integration as well as technical aspects in multiple areas, including DL; (3) concerted working of multiple teams, including subject matter experts for validation and prompt engineering; and (4) regulatory oversight because the training phase in TSP would involve using health care data with the DL algorithm. Data-warehouse implementation and continued management could cost millions of dollars. The cost of hiring and retaining a workforce to support LLM implementation can be substantial, especially given the high demand for such specialized skillsets. For a midsized health care organization, the cost of a workforce capable of handling LLM development, implementation, deployment, and maintenance can range from US $2 million to US $5 million per year, including salaries, benefits, and training. Training and deploying LLMs require significant computational resources, including high-performance storage and computing infrastructure. For example, DGX A100 data centers (80 GB) were priced at approximately US $200,000 in 2020. LLaMA implementation [32] required 2048 A100 GPUs and 21 days for training their 65 billion parameter model, resulting in significant costs in millions of dollars. Therefore, the cost of compute, along with power consumption, physical space requirements, and dedicated personnel, could easily reach into the tens of millions of dollars for TSP. Working in partnership with cloud platforms can address several of these challenges. Major CSPs, such as AWS, Azure, GCP, and OCI, offer on-demand compute instances across GPU clusters and cost profiles based on the user needs and affordability within a secured framework (Table 2) [117-122]. TSP is usually followed by FTP to tune the response of the LLMs.

Because TSP demands building an LLM using the health care organizations data, it may require *hundreds of thousands of hours* of training [123]. Comparable pricing of 320 and 640 GiB of GPU memory using 8 × A100 GPUs in a single instance across the 4 major CSPs is presented in Table 2. On the basis of the comparison table and the time for training an LLM from scratch, it might be economical to purchase a long-term (3-year commitment), which may save around 60% when compared with on-demand costs. Another option would be to use a "spot instance" (Table 3) [124-127]. Spot instances are spare compute capacity offered by CSPs at a reduced cost compared with the on-demand pricing and serve as a suitable alternative. These pricing estimates vary with demand and can change throughout the day, week, or month. However, to use spot instances for LLM training, organizations need to implement strategies to handle instance reclamations and checkpoint management. This is especially critical for TSP, as it takes considerable time to train an LLM from scratch. CSP-managed services offer managed spot training or fine-tuning and resume jobs from the checkpoints. Due to spot interruptions, training or fine-tuning using spot instances may also take longer to complete compared with on-demand or reserved instances. With increasing adoption of LLMs by health care organizations in conjunction with the popularity of AI and ML availing spot instances in general can be challenging and could be prone to interruptions with marked variations in availability as well as pricing across the different geographic regions. Spot instances can also vary across CSPs, with some (eg, AWS, GCP, and OCI) providing more options and higher-memory GPU instances.

Table 2. Comparable per-hour pricing of graphics processing unit (GPU) clusters across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]) for 320 or 640 GiB GPU memory for the training from scratch pathway (TSP). Representative data retrieved on June 2024.

CSP	Instance type	CPU[a] (cores)	Memory (GiB)	GPUs	Per GPU memory	Total GPU memory	On demand (US $)	On demand per GPU (US $)	1 year (US $)	3 years (US $)
AWS	p4d.24xlarge	96	1152	8	40	320	32.77	4.10	19.22	11.57
AWS	p4de.24xlarge	96	1152	8	80	640	40.96	5.12	24.01	14.46
Azure	ND96asr A100 v4	96	900	8	40	320	27.20	3.40	22.62	13.63
Azure	ND96amsr A100 v4	96	1900	8	80	640	32.77	4.10	20.97	14.42
GCP	a2-highgpu-8g	96	680	8	40	320	29.39	3.67	18.52	10.29
GCP	a2-ultragpu-8g	96	1360	8	80	640	40.22	5.03	__[b]	—
OCI	BM.GPU4.8	64	2048	8	40	320	24.40	3.05	—	—
OCI	BM.GPU.A100-v2.8	128	2048	8	80	640	32.00	4	—	—

[a]CPU: central processing unit.

[b]Not applicable.

XSL•FO
RenderX

Table 3. Comparison of per-hour spot-instance and on-demand pricing across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]) large language models (eg, 320 GiB, 8 × A100 graphics processing units [GPUs] single instance) and smaller and medium large language models (eg, 64, 16 GiB V100 GPU). Representative data were retrieved in June 2024.

CSP	Large models			Medium and smaller models		
	Instance type	On-demand cost per hour (US $)	Spot cost per hour (US $)	Instance type	On-demand cost per hour (US $)	Spot cost per hour (US $)
AWS	p4d.24xlarge	32.77	8.37	p3.8xlarge	12.24	3.97
Azure	ND96asr A100 v4	27.20	8.19	NC24rs_v3	13.46	0.91
GCP	a2-highgpu-8g	29.39	11.75	n1-highmem-32	21.73	3.63
OCI	BM.GPU4.8	24.40	12.20	BM.GPU3.4	12.03	6.02

FTP Overview

While TSP focuses on pretraining LLMs, FTP focuses on adapting an existing pretrained LLM with a given architecture and parameters to tasks at hand in a domain-specific manner. This is usually accomplished by (1) adjusting the model parameters of the LLM using context-specific data that are much smaller than the training data and (2) adjusting the LLM performance and behavior by prompt engineering inputs and outputs of the LLM. The pretrained LLMs can be either open-source or proprietary LLMs, with those pretrained on health care data expected to perform better than general-purpose LLMs.

Benefits

The timeline for implementation, budgeting, infrastructure, and workforce needs for FTP is expected to be significantly lower than that of TSP because it does not involve training LLMs from scratch [114]. Typically, FTP uses readily available, pretrained proprietary or open-source LLM with open-source licenses for modifying the source code as per user needs. The number of open-source LLM offerings has continued to increase with time with communities, such as Hugging Face hosting leaderboards comparing their performance. Managed services by CSPs can assist in setting up multistep tasks across systems and data sources, generate knowledge bases from private data sources for FTP, and implement safeguards on inputs and outputs adhering to governance and responsible AI.

Risks

FTP will demand resources, access to quality data, relevant prompts (eg, input-output pairs), protected time for subject matter experts, and agile implementation strategies for adapting the pretrained LLMs for specific tasks. While automated approaches have been proposed for prompt engineering [128], prompt engineering risks for FTP may be relatively higher compared with fine-tuning on TSP because the training data of the pretrained models may not be domain-specific and can have potential biases. Implementation details of proprietary pretrained LLMs may not be readily accessible, limiting innovation and modification with evolving needs. Pricing of proprietary LLMs used by FTP may also increase with enhanced adoption across health care organizations, and their downtime may impact several dependent downstream applications in health care workflows. While several open-source LLMs are readily available from platforms, such as Hugging Face, these are primarily a result of crowdsourcing efforts and voluntary contributions posing challenges in translating them to enterprise tools. Dependency on an existing pretrained open-source or proprietary LLM can be a risk because these models are rapidly evolving and that could challenge active maintenance of legacy models. Open-source implementations traditionally do not support extensive documentation and training materials for onboarding. These in turn may demand workforce and digital capacity on-premises. This includes challenges in interfacing these tools with other systems and health care workflows. Open-source implementations may also be susceptible to vulnerabilities that may not be readily apparent; hence, they could be susceptible to security breaches, malicious content, malware, and ransomware attacks on models and downstream applications compromising patient privacy and leading to liabilities. Pretrained proprietary LLMs may have minimal flexibility, transparency, and interface options to downstream applications and dashboards in health care workflows. Because health care data can contain PHI, open-source and pretrained proprietary LLMs should be HIPAA compliant.

Economics of FTP

The economics of FTP (Table 4) are expected to be markedly lower compared with TSP as it does not involve the computationally intensive task of pretraining an LLM. More specifically, digital readiness and infrastructure and workforce costs are expected to be markedly lesser than those of TSP. In addition, FTP may require access to the annotated health care data that are several magnitudes less than the training data used in TSP. Unlike TSP, fine-tuning is typically accompanied by high GPU consumption for a short burst of time and ideal for availing pay-as-you-go models offered by CSPs in contrast to on-premise systems.

XSL•FO
RenderX

Table 4. Comparable pricing of graphics processing unit [GPU] clusters across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]) for the fine-tuned pathway. Representative data were retrieved in June 2024.

CSP	Instance type	CPU[a] cores	Memory (GiB)	GPUs	Per GPU memory	Total GPU memory	On demand (US $)	1 year (US $)	3 year (US $)
AWS	p3.16xlarge	64	488	8	16	128	24.48	15.91	8.39
AWS	p3dn.24xlarge	96	768	8	32	256	31.22	18.39	9.64
Azure	NC24rs_v3	24	448	4	16	64	12.24	8.98	6.52
GCP	n1-highmem-64	64	416	8	16	128	23.63	14.88	10.63
OCI	BM.GPU3.8	52	768	8	16	128	23.60	__[b]	—

[a]CPU: central processing unit.

[b]Not applicable.

OBP Overview

OBP includes commercial off-the-shelf LLMs typically accessed by end users through Representational State Transfer APIs (SaaS) with minimal or no local customization.

Benefits

Unlike TSP and FTP, OBP is usually enterprise-ready. OBPs do not require digital readiness in access to integrated datasets and data warehouses. LLMs are accessed through a web interface; hence, no budgeting needs to be allocated for the storage and computing infrastructure needs of LLMs or the workforce to support LLM implementation and maintenance, as in TSP and FTP. Because OBP is provided by multiple vendors, there is an option to choose the best-performing LLM for a given price. Transitioning between OBP services can be done with ease because there is no explicit sharing of sensitive health care data or customized LLM architecture. User training needs in OBP are minimal compared with TSP and FTP. Managed services provided by CSPs can assist OBP implementation with minimal workforce needs on-premises. This includes CSP-managed API end points that offer access to 1 or more LLMs, whose cost is proportional to the number of input and output tokens. Services provided by the 4 major CSPs in this regard include (AWS: Claude, Azure: OpenAI Service, GCP: Gemini, and OCI: Cohere). Managed services by CSPs also support OBP LLM deployment on GPUs with minimal ease and a per-hour charge (AWS: Amazon SageMaker Canvas, Azure: AI Studio, and GCP: Vertex AI).

Risks

Unlike TSP and FTP, OBP is completely dependent on the SaaS or infrastructure as a service option provided by the vendor, with no control over training and fine-tuning and limited transparency on potential biases and the details of the LLM model and the training data. Prompt engineering risks of OBP may be larger than those of FTP and TSP because the OBP may use models that are not domain-specific, and the absence of domain-specific knowledge may result in OBP being sensitive to prompting. Output variability can be relatively higher for OBP, leading to inconsistent results and challenging deploying these models in clinical workflows. Copyright protection may limit the extent to which OBP can reveal the implementation details, and OBP services can be black boxes. Therefore, OBP services are expected to provide generic insights that may not necessarily accommodate the digital footprints and characteristics of the population served by a health care organization. This in turn may diminish the value and utility of OBP in addressing the needs that are specific to the organization. It might not be feasible to use PHI data across OBP services that are not HIPAA compliant. Posing questions to OBP services may unintentionally compromise patient privacy, especially when the sample size of the cohort being queried is small (eg, rare disease). OBP services may pose challenges in interfacing other applications, dashboards, and workflows without the explicit involvement of the vendor. Cost is usually incurred per inference and can aggregate over time with increasing dependency and number of users. These in turn may demand user authentication, quota allocation, and auditing. The reliability of the OBP service is dependent on the vendor and not necessarily guaranteed, with the potential for pricing options to increase with increasing adoption.

Economics of OBP

The cost of digital readiness for OBP is expected to be minimal as OBP services are provided as ready-to-use solutions without any need for training or fine-tuning as in TSP and FTP. Ideally, OBP would not require a dedicated team to support other than training materials for end users. However, hosting open-source LLM models, such as those from Hugging Face Hub, may demand infrastructure costs on-premises or in the cloud. Alternatively, LLM models in the OBP can be availed through APIs whose charges vary based on the input and output tokens and use patterns. Use patterns need to be carefully monitored as they could gradually increase to millions of tokens per month, significantly impacting the cost of OBP long-term. Pricing estimates of LLM APIs across major CSPs are shown in Table 5 [129-132]. Due to the per-unit pricing of certain CSPs, an average of 4 characters per token is assumed in generating these estimates.

XSL•FO
RenderX

Table 5. Comparable pricing of large language models through application programming interface across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]). Representative data were retrieved in June 2024.

CSP	Model	Input unit cost (US $)	Input parameters	Output unit cost (US $)	Output parameters	10,000 Input tokens (US $)	500 Output tokens (US $)	Total cost (US $)
AWS	Claude 3 Sonnet	0.003	Per 1000 tokens	0.015	Per 1000 tokens	0.030	0.008	0.038
Azure	GPT-4o global deployment	0.005	Per 1000 tokens	0.015	Per 1000 tokens	0.050	0.008	0.058
GCP	Gemini 1.5 pro	0.00125	Per 1000 characters (approximately 250 tokens)	0.00375	Per 1000 characters (approximately 250 tokens)	0.050	0.015	0.058
OCI	Cohere large	0.022	Per 10,000 transactions (approximately 2500 tokens)	0.022	Per 10,000 transactions (approximately 2500 tokens)	0.088	0.004	0.092

LLM Applications in Health Care

LLMs have the potential to offer health care providers new mechanisms for optimization and automation of documentation, clinical review, and direct patient communication. Its use is expected to reduce provider time in using systems, such as EHR, and improve clinical documentation while minimizing repetitive workflows. LLMs can also benefit patients by serving as an approachable tool to manage their health information assist in health literacy, appointment scheduling, and ambulatory encounters. Typical LLM applications in health care are shown in Figure 3. These applications broadly rely on the principles underlying dialogue systems [133]. They facilitate text-based and voice-based interactions between LLMs and the end user.

Figure 3. Typical health care applications of (A) large language models (LLMs) and conversational artificial intelligence (AI), (B) chatbots, (C) summarization, and (D) translation. Q/A: question and answer.

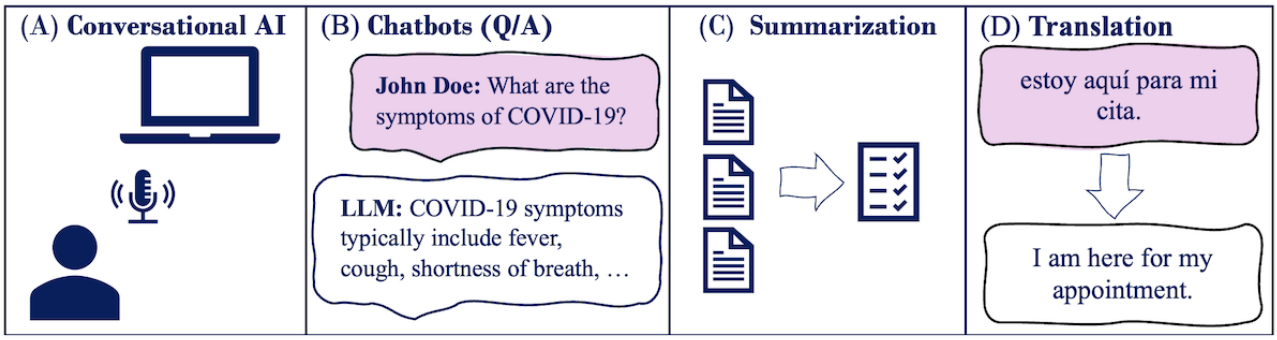

Conversational AI in Clinical Practice

Conversational AI (Figure 3A) focuses on the application of AI-based approaches, such as LLMs, for developing dialogue systems [133]. It is an omnibus term that includes chatbots and virtual assistants as well as text and voice-based interaction with the end user. The ability of these systems to understand user intent and generate relevant and accurate responses has continued to improve over time [134,135]. Conversational AI-enabled systems can significantly enhance information seeking and retrieval capabilities of health care information by patient, provider, and payer. Specifically in clinical practice, conversational AI integrated within EHRs [136] can facilitate increased patient-provider engagement and reduced documentation time, minimizing physician burnout. Moreover, these ambient listening, generative AI solutions have the opportunity to dramatically reduce documentation and cognitive burden, further leading to an improved clinician EHR experience. Depending on the services needed, the integration can be pursued through TSP for a bespoke solution, FTP for modifying existing models, and the OBP pathway with minimal customization. In addition to the risks and benefits of these pathways, their deployment in EHR workflows may demand expertise in user interface design for functional integration. The focus should be on ensuring that AI-supported interactions enhance rather than impede the patient-provider relationship. Operationally, such systems streamline data management processes, leading to increased efficiency and potential financial savings due to reduced time spent on repetitive tasks. Depending on digital and analytics maturity, there could be challenges related to integration and deployment challenges and enhanced adoption. In enhancing the conversational AI interfaces within EHR systems, a more detailed exploration of specific use cases and scenarios would reveal their impact on clinical workflow, such as patient portal communications or clinical decision support. Long-term adoption studies, focusing on the acceptance and resistance among health care professionals, can provide insights into the practical aspects of implementing these technologies. Moreover, a deeper discussion on how these AI interfaces evolve through ML and adaptation to user behaviors and preferences can illustrate the potential for increasingly efficient and user-friendly medical systems. The development of guardrails on AI-moderated operations and data security, particularly in the context of sensitive EHR data, is imperative

XSL•FO
RenderX

to address concerns around privacy and compliance with health care regulations.

Conversational AI economics depends on the quantity, complexity, and scale of integration. For instance, conversational AI tools for EHR demand seamless and secured integration of the tool to the EHR either directly or through third-party interfaces, training the users for enhanced adoption, and customization of the tools through fine-tuning. Training conversational AI tools is critical to maintaining high accuracy on clinical tasks and dialogue understanding impacting its adoption. Successful deployment will demand active involvement of multiple teams, including analytics, information technology, the governance team, compliance, and security. Scalability considerations are crucial if deployed across multiple facilities, necessitating robust infrastructure investments. In addition, ongoing optimization costs are incurred as the AI must continuously update to align with new medical guidelines and treatment protocols, ensuring reliability and accuracy in patient data management.

Chatbots in Health Care

A chatbot is a conversational AI system that enables communication between computers and humans using natural language in completing specific tasks, such as question answering (Figure 3B). LLM-based chatbots are transforming the way patients, providers, and payers interact within health care settings [137]. They have the potential to improve health care outcomes by providing timely health information to patients, assisting in patient education and interventions, improving operational productivity by assisting decision-making processes, and reducing administrative tasks and overhead costs [138]. Chatbots can also assist in individualized services, including symptom assessment through virtual consults, appointment scheduling, and improving health literacy by making health care information accessible in a preferred language [139-141]. In developing chatbots, health care systems may opt for TSP for a completely customized solution, FTP for adapting existing models, or OBP for ready-to-use prebuilt templates. In addition to the risks and benefits of these pathways discussed earlier, chatbot implementations demand a good understanding of patient engagement strategies and health care communication norms. Successful deployment should ensure privacy and security of patients, adhere to regulations, and establish necessary guardrails [110]. Their design should also prioritize empathy and cultural sensitivity to ensure inclusive and respectful interactions with diverse patient populations in addition to minimizing bias and assumptions in conversations [137]. These, in turn, are expected to build trust and harmony with users, leading to enhanced adoption [142]. Successful chatbot design should also accommodate nuances of human interactions (eg, patient emotions and expectations). These in turn may demand well-articulated prompt engineering, fine-tuning, and optimization of the parameters in the underlying model [143]. Incorporating feedback from patients and health care providers will offer a deeper understanding of user experience and areas for improvement. Chatbots can also be integrated with other digital health technologies and workflows, such as telemedicine and electronic medical record. In addition, it can assist in personalization of services, tailoring interactions based on individual patient profiles and needs, significantly impacting patient engagement. Integrating chatbots with predictive analytics can also assist in assessing the usefulness of these tools by incorporating feedback.

The economics of LLM-powered chatbots in health care can be impacted by integration, compliance needs, feature complexity, and deployment scale. OBP is generally the least expensive, potentially costing a few thousand to tens of thousands of dollars annually, based on a subscription model. However, these figures can easily grow with the frequency of use and number of users. Fine-tuning existing models can cost tens to hundreds of thousands of dollars, depending on the extent of customization and licensing fees. Developing a highly customized chatbot from scratch is the most expensive option, with expenses running into hundreds of thousands to millions of dollars. Additional costs arise from ongoing maintenance, server costs, updates, and compliance with regulations such as HIPAA or GDPR, which necessitate robust security measures. The complexity of features, such as multilingual support advanced diagnostic capabilities, and the scale of deployment can also significantly impact the overall cost.

Summarization in Health Care

LLMs excel in summarization (Figure 3C), tasks critical for managing extensive medical documentation and improving clinical workflows [144]. These AI-driven systems can distill complex medical records into concise summaries for improved decision-making and patient management. Summarization also reduces task load while improving documentation quality, operational efficiency in processing documents, and financial savings by reducing the time and resources spent on administrative tasks. Summarization can also succinctly capture details of patient-provider conversations in team settings comprising a large number of clinicians, resulting in enhanced care continuity, coordination, and overall quality. While similar in integration complexity to conversational AI systems, the implementation of LLMs for summarization specifically requires tuning the models to capture critical medical insights accurately. Summarization leverages advanced natural language understanding capabilities, a step beyond general chatbot applications to ensure that summaries are not only succinct but also clinically relevant. These can be developed through TSP for high specificity, FTP for a balanced approach, or OBP for broader applications. The summarization process requires an augmented framework comprising a group of experts in the domain of AI, clinical knowledge and medical terminology, and data-processing infrastructure for critical validation. Successful implementation of LLM summarization is expected to ensure integrity of medical information, preventing any loss of critical details in the summarization process, minimizing the risk of misinterpretation of condensed information, and seamless integration of these tools into existing clinical workflows.

The economics of summarization while overlapping with those of broader AI integrations, such as conversational AI tools, are particularly influenced by the need for high-quality training data and the development of interfaces that clinicians can use effectively within existing digital health frameworks and workflows. The cost and investment may increase for

XSL•FO
RenderX

summarization tools that meet high standards of accuracy and reliability in medical contexts while minimizing risks. This includes rigorous testing and validation in real-world settings to adhere to the data handling and privacy regulations characteristic of the health care institution and industry. In addition, the ongoing maintenance to update the models with new medical information and guidelines further adds to the overall expenditure. These factors combined make the economics of summarization technologies substantial yet crucial for enhancing efficiency and decision-making in health care environments.

Machine Translation in Health Care

Machine translation (Figure 3D) serves diverse linguistic communities by the translation of text or speech from one language to another. Its role is especially helpful in overcoming language barriers in medical communication and documentation, especially across health care organizations that serve non–English-speaking communities. Machine translation can improve patient–health care provider communication, patient understanding of instructions, and discharge summaries, as well as operational benefits by facilitating multilingual documentation and financial advantages by potentially reducing the workload for human interpreters in low-resource settings [145]. It also has the potential to assist in the transmission of critical medical information in a culturally sensitive and empathetic manner, with the potential to minimize adverse events and impact health care outcomes favorable, especially across non–English-speaking communities. Machine translation can be developed through TSP for precise, context-specific translations using data from communities served by a specific health care organization, FTP for adapting existing models to medical language nuances, or OBP for immediate implementation with off-the-shelf translation tools. In addition to the applications mentioned here, machine translation requires collaboration with linguists and cultural sensitivity advisers to ensure translations are accurate and culturally appropriate. Ethical and regulatory considerations revolve around the accuracy and cultural appropriateness of translations. There is a strong emphasis on avoiding miscommunication in critical medical contexts while respecting linguistic diversity. Challenges include the risk of mistranslation, cultural insensitivity, and loss of nuanced medical context. The impact on target populations is usually diverse. It enables payers to offer multilingual services efficiently, aids providers in delivering equitable care to non–English-speaking patients and empowers patients by providing access to medical information in their preferred native languages. Here, underlining cultural competency alongside language translation is crucial. This includes not only translating text but also understanding and conveying cultural nuances, which is critical in medical contexts with potential favorable impact outcomes. Establishing specific metrics or standards to gauge the accuracy and reliability of translations can provide a benchmark for evaluating these tools. Discussing the legal implications and responsibilities in cases of mistranslation or miscommunication is also vital to understanding the potential liabilities involved. Extensive testing may be required before deployment in high-stakes areas, such as emergency medicine, where quick and accurate translation is vital.

The economics of machine translation in health care settings largely overlap with those of conversational AI technologies, as discussed previously. Costs can vary based on the development pathway chosen: OBP solutions may offer a lower upfront cost with general translation tools available for immediate use, while the FTP and the TSP require more substantial investments to adapt or develop models that handle medical language nuances and specific community dialects. However, OBP solutions could have severe limitations in high-stakes applications compared with FTP and TSP. Costs usually include customization, system integration with existing health care IT infrastructures such as EHRs, and ongoing expenses for maintenance and updates. Additional significant expenses are incurred in ensuring accuracy and cultural appropriateness, which involves collaboration with linguists and cultural experts. This collaboration is essential to mitigate risks of mistranslation and to comply with health care communication standards. Therefore, while the base technology may be like those used in conversational AI, the specificity and critical nature of medical translations can lead to higher costs, particularly when ensuring the system meets the stringent requirements of medical accuracy and regulatory compliance.

Conclusions

LLMs have the potential to meaningfully impact health care delivery and health outcomes. However, LLM implementations are impacted by the needs and affordability. This perspective provided 3 LLM implementation pathways (TSP, FTP, and OBP) and a road map for onboarding, enhanced democratization, and equitable adoption by the health care ecosystem. The economics, risks, and benefits of these pathways were also presented across 4 major CSPs (AWS, GCP, Azure, and OCI) to assist in choosing the best pathway for an organization. As LLMs continue to evolve [146], additional onboarding pathways are expected to join the repertoire.

The critical role of cloud-computing frameworks to support onboarding efforts from scalability, privacy, workforce, and economic standpoints was discussed. Pay-as-you-go models offered by CSPs alleviate the need for significant upfront investments while providing the ability to experiment with different pathways with the flexibility to scale and transition between pathways based on the usefulness of these tools in impacting outcomes and KPIs and use with evolving needs of health care organizations. Managed services provided by CSPs can assist in optimal management of resources and infrastructure while streamlining workflows and minimizing the need for considerable expertise across a variety of areas. These aspects are especially suited for organizations that do not have sufficient resources and upfront investment for LLM implementation. CSPs also provide privacy and security features, including confidential computing and TEEs for safeguarding sensitive health care data and maintaining regulatory compliance. The size of health care datasets used to train LLMs is often small compared with datasets used to train general-purpose models. CSPs can facilitate FL in conjunction with TEEs and deidentification in overcoming sample size constraints by

XSL•FO
RenderX

enabling collaborative training strategies across health care organizations without explicit data sharing and ensuring privacy. FL approaches can result in models with enhanced generalization ability in contrast to those trained using data from a single health care organization. Because LLMs may have the potential to memorize sensitive information from training data, hindering the sharing of checkpointed models due to privacy concerns. Techniques such as differential privacy and secure multiparty computation in CSPs can mitigate such risks. CSPs also provide access to specialized hardware accelerators, presenting an opportunity to improve the efficiency and cost-effectiveness of LLM training and inference. However, it is important to consider compatibility and performance trade-offs when integrating these accelerators into existing workflows.

As LLMs are used across health care, it is crucial to consider potential challenges and unintended consequences. Choice of an LLM implementation pathway can be significantly impacted by digital readiness, infrastructure, workforce, and ethical and regulatory landscape. Overreliance on LLMs could also diminish the critical thinking skills of health care professionals. As with all AI and ML tools, optimization is an essential ingredient of LLMs. Therefore, it is essential that LLMs be used within an augmented framework to support human decision-making rather than serving as a replacement. Establishing guardrails, ethical guidelines, and training programs for ensuring the responsible use of LLMs in clinical settings is important. Providing training and support for health care professionals and actively engaging them in the LLM implementation is critical for their successful deployment and long-term adoption. Workforce is a critical ingredient for LLM implementation, deployment, and maintenance. Prioritizing the inclusion of belonging, diversity, equity, and inclusion leaders as a part of LLM development is crucial to ensuring implementation that is inclusive and representative of diverse populations. Engaging policy makers and educating them about LLM limitations and adoption in health care is critical for realistic expectations from these tools and developing the necessary regulatory frameworks. Incentives could be introduced to encourage LLM adoption across health care organizations, and KPIs should be identified to assess its impact on health care outcomes. Short-term incentives can facilitate initial adoption of a particular onboarding pathway, while long-term incentives may assist in shifting across these pathways.

Identifying onboarding pathways for LLM implementation leveraging cloud computing along with metrics to demonstrate value while incorporating the necessary regulations and guardrails of responsible AI is critical for its equitable distribution and enhanced adoption in the health care ecosystem. Widespread adoption is also expected to facilitate feedback from diverse communities served by the health care ecosystem, improving patient outcomes and operational efficiency and addressing the unique challenges and considerations in health care.

Authors' Contributions

RN conceived the presented idea. All authors contributed to the content and development of the manuscript. All authors contributed to the editing, reviewing, and approval of the final manuscript.

Conflicts of Interest

ES serves on the editorial board of JMIR Publications.

References

1. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep 15;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
2. Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science 2023 Sep 15;381(6663):adk6139. [doi: 10.1126/science.adk6139] [Medline: 37708283]
3. Kong HJ. Managing unstructured big data in healthcare system. Healthc Inform Res 2019 Jan;25(1):1-2 [FREE Full text] [doi: 10.4258/hir.2019.25.1.1] [Medline: 30788175]
4. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health 2023 Oct 11;2(10):e0000347 [FREE Full text] [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
5. Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, et al. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. arXiv Preprint posted online on February 18, 2023 [FREE Full text]
6. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med 2019 Dec 17;2(1):130 [FREE Full text] [doi: 10.1038/s41746-019-0208-8] [Medline: 31872069]
7. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017 Sep;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
8. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]

XSL•FO
RenderX

9. Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, et al. Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. JAMA Netw Open 2023 Mar 01;6(3):e231204 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.1204] [Medline: 36862411]
10. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020 Mar 01;27(3):457-470 [FREE Full text] [doi: 10.1093/jamia/ocz200] [Medline: 31794016]
11. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc 2019 Nov 01;26(11):1247-1254 [FREE Full text] [doi: 10.1093/jamia/ocz149] [Medline: 31512729]
12. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023 Jul 07;619(7969):357-362 [FREE Full text] [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
13. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 02;6(10):e2336997 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
14. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv Preprint posted online on August 16, 2021 [FREE Full text]
15. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug 12;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
16. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv Preprint posted online on May 16, 2023 [FREE Full text]
17. Armbrust M, Ghodsi A, Xin R, Zaharia M. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of the 11th Annual Conference on Innovative Data Systems Research. 2021 Presented at: CIDR '21; January 11-15, 2021; Online URL: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf
18. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 05;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
19. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. London, UK: Pearson; 2010.
20. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Cham, Switzerland: Springer International Publishing; 2009.
21. Aggarwal CC. Data Mining: The Textbook. Cham, Switzerland: Springer International Publishing; 2015.
22. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: The MIT Press; 2016.
23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
24. Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun 2022 Apr 01;13(1):1728 [FREE Full text] [doi: 10.1038/s41467-022-29268-7] [Medline: 35365602]
25. Chollet F. Deep Learning with Python. Shelter Island, NY: Manning Publications; Nov 2017.
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv Preprint posted online on June 12, 2017 [FREE Full text]
27. Ananthaswamy A. In AI, is bigger always better? Nature 2023 Mar 08;615(7951):202-205. [doi: 10.1038/d41586-023-00641-w] [Medline: 36890378]
28. Friedman CP. A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc 2009 Mar;16(2):169-170. [doi: 10.1197/jamia.m3092]
29. Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization. arXiv Preprint posted online on November 9, 2018 [FREE Full text]
30. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv Preprint posted online on March 4, 2022 [FREE Full text]
31. Adoption model for analytics maturity (AMAM). Healthcare Information and Management Systems Society. URL: https://www.himss.org/maturity-models/amam/ [accessed 2024-09-30]
32. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv Preprint posted online on February 27, 2023 [FREE Full text]
33. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. PaLM 2 technical report. arXiv Preprint posted online on May 17, 2023. [doi: 10.48550/arXiv.2305.10403]
34. Health information privacy. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/security/index.html [accessed 2024-10-14]
35. General Data Protection Regulation (GDPR) compliance guidelines. General Data Protection Regulation. URL: https://gdpr.eu/ [accessed 2024-10-14]

XSL•FO
RenderX

36. Radhakrishnan L, Schenk G, Muenzen K, Oskotsky B, Ashouri Choshali H, Plunkett T, et al. A certified de-identification system for all clinical text documents for information extraction at scale. JAMIA Open 2023 Oct;6(3):ooad045 [FREE Full text] [doi: 10.1093/jamiaopen/ooad045] [Medline: 37416449]
37. Mulligan DP, Petri G, Spinale N, Stockwell G, Vincent HJ. Confidential computing—a brave new world. In: Proceedings of the International Symposium on Secure and Private Execution Environment Design. 2021 Presented at: SEED 2021; September 20-21, 2021; Washington, DC. [doi: 10.1109/seed51797.2021.00025]
38. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010;17(2):169-177 [FREE Full text] [doi: 10.1136/jamia.2009.000026] [Medline: 20190059]
39. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011 Dec 2;6(12):e28071 [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
40. de Kok JW, de la Hoz M, de Jong Y, Brokke V, Elbers PW, Thoral P, Collaborator Group, et al. A guide to sharing open healthcare data under the General Data Protection Regulation. Sci Data 2023 Jun 24;10(1):404 [FREE Full text] [doi: 10.1038/s41597-023-02256-2] [Medline: 37355751]
41. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
42. Schaeffer R, Miranda B, Koyejo S. Are emergent abilities of large language models a mirage? arXiv Preprint posted online on April 28, 2023 [FREE Full text]
43. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv Preprint posted online on March 31, 2023 [FREE Full text]
44. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv Preprint posted online on January 20, 2020. [doi: 10.48550/arXiv.2001.08361]
45. Chen C, Feng X, Zhou J, Yin J, Zheng X. Federated large language model: a position paper. arXiv Preprint posted online on July 18, 2023. [doi: 10.48550/arXiv.2307.08925]
46. Ye R, Wang W, Chai J, Li D, Li Z, Xu Y, et al. OpenFedLLM: training large language models on decentralized private data via federated learning. arXiv Preprint posted online on February 10, 2024. [doi: 10.48550/arXiv.2402.06954]
47. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021 Sep 07;4(1):132 [FREE Full text] [doi: 10.1038/s41746-021-00489-2] [Medline: 34493770]
48. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. arXiv Preprint posted online on February 17, 2016. [doi: 10.48550/arXiv.1602.05629]
49. Bak M, Madai VI, Celi LA, Kaissis GA, Cornet R, Maris M, et al. Federated learning is not a cure-all for data ethics. Nat Mach Intell 2024 Mar 18;6:370-372. [doi: 10.1038/s42256-024-00813-x]
50. Inmon WH. Building the Data Warehouse. Hoboken, NJ: John Wiley & Sons; 2005.
51. Kimball R, Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Hoboken, NJ: John Wiley & Sons; 2002.
52. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009 Sep;16(5):624-630. [doi: 10.1197/jamia.m3191]
53. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
54. Saitwal H, Qing D, Jones S, Bernstam EV, Chute CG, Johnson TR. Cross-terminology mapping challenges: a demonstration using medication terminological systems. J Biomed Inform 2012 Aug;45(4):613-625 [FREE Full text] [doi: 10.1016/j.jbi.2012.06.005] [Medline: 22750536]
55. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf 2014 Nov;37(11):945-959 [FREE Full text] [doi: 10.1007/s40264-014-0214-3] [Medline: 25187016]
56. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
57. Chang BL, Bakken S, Brown SS, Houston TK, Kreps GL, Kukafka R, et al. Bridging the digital divide: reaching vulnerable populations. J Am Med Inform Assoc 2004;11(6):448-457 [FREE Full text] [doi: 10.1197/jamia.M1535] [Medline: 15299002]
58. Wang M, Gago CM, Rodriguez K. Digital redlining-the invisible structural determinant of health. JAMA 2024 Apr 16;331(15):1267-1268. [doi: 10.1001/jama.2024.1628] [Medline: 38497952]
59. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
60. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
61. Blueprint for trustworthy AI: implementation guidance and assurance for healthcare coalition for health AI. Coalition for Health AI. 2023 Apr 4. URL: https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf [accessed 2024-10-14]

XSL•FO
RenderX

62. Charpignon ML, Byers J, Cabral S, Celi LA, Fernandes C, Gallifant J, et al. Critical bias in critical care devices. Crit Care Clin 2023 Oct;39(4):795-813. [doi: 10.1016/j.ccc.2023.02.005] [Medline: 37704341]
63. Teotia K, Jia Y, Link Woite N, Celi LA, Matos J, Struja T. Variation in monitoring: glucose measurement in the ICU as a case study to preempt spurious correlations. J Biomed Inform 2024 May;153:104643. [doi: 10.1016/j.jbi.2024.104643] [Medline: 38621640]
64. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
65. Dean J, Corrado GS, Monga R, Chen K, Devin M, Le QV, et al. Large scale distributed deep networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. 2012 Presented at: NIPS'12; December 3-6, 2012; Lake Tahoe, Nevada.
66. Huang Y, Cheng Y, Bapna A, Firat O, Chen MX, Chen D, et al. GPipe: efficient training of giant neural networks using pipeline parallelism. arXiv Preprint posted online on November 16, 2018. [doi: 10.48550/arXiv.1811.06965]
67. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-lm: training multi-billion parameter language models using model parallelism. arXiv Preprint posted online on September 17, 2019. [doi: 10.48550/arXiv.1909.08053]
68. Faiz A, Kaneda S, Wang R, Osi R, Sharma P, Chen F, et al. LLMCarbon: modeling the end-to-end carbon footprint of large language models. arXiv Preprint posted online on September 25, 2023. [doi: 10.48550/arXiv.2309.14393]
69. Bui ND, Le H, Wang Y, Li J, Gotmare AD, Hoi SC. CodeTF: one-stop transformer library for state-of-the-art code LLM. arXiv Preprint posted online on May 31, 2023. [doi: 10.48550/arXiv.2306.00029]
70. Candel A, McKinney J, Singer P, Pfeiffer P, Jeblick M, Lee CM, et al. H2O open ecosystem for state-of-the-art large language models. arXiv Preprint posted online on October 17, 2023 [FREE Full text] [doi: 10.18653/v1/2023.emnlp-demo.6]
71. Cresswell K, Domínguez Hernández A, Williams R, Sheikh A. Key challenges and opportunities for cloud technology in health care: semistructured interview study. JMIR Hum Factors 2022 Jan 06;9(1):e31246 [FREE Full text] [doi: 10.2196/31246] [Medline: 34989688]
72. Peccerillo B, Mannino M, Mondelli A, Bartolini S. A survey on hardware accelerators: taxonomy, trends, challenges, and perspectives. J Syst Archit 2022 Aug;129:102561. [doi: 10.1016/j.sysarc.2022.102561]
73. Chen H, Zhang J, Du Y, Xiang S, Yue Z, Zhang N, et al. A comprehensive evaluation of FPGA-based spatial acceleration of LLMs. In: Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays. 2024 Presented at: FPGA '24; March 3-5, 2024; Monterey, CA. [doi: 10.1145/3626202.3637600]
74. Mohaidat T, Khalil K. A survey on neural network hardware accelerators. IEEE Trans Artif Intell 2024 Aug;5(8):3801-3822. [doi: 10.1109/tai.2024.3377147]
75. Machupalli R, Hossain M, Mandal M. Review of ASIC accelerators for deep neural network. Microprocess Microsyst 2022 Mar;89:104441. [doi: 10.1016/j.micpro.2022.104441]
76. Batra G, Jacobson Z, Madhav S, Queirolo A, Santhanam N. Artificial-intelligence hardware: new opportunities for semiconductor companies. McKinsey and Company. 2018 Dec. URL: https://tinyurl.com/22a7vapn [accessed 2024-10-14]
77. Mince F, Dinh D, Kgomo J, Thompson N, Hooker S. The grand illusion: the myth of software portability and implications for ML progress. arXiv Preprint posted online on September 12, 2023. [doi: 10.1016/0950-5849(89)90153-5]
78. Bshara N. AWS Trainium: the journey for designing and optimization full Stack ML hardware. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. 2024 Presented at: ASPLOS '24; April 27-May 1, 2024; La Jolla, CA. [doi: 10.1145/3620666.3655592]
79. Romero F, Li Q, Yadwadkar NJ, Kozyrakis C. INFaaS: automated model-less inference serving. In: Proceedings of the 2021 USENIX Annual Technical Conference. 2021 Presented at: USENIX ATC 2021; July 14-16, 2021; Virtual Event URL: https://www.usenix.org/conference/atc21/presentation/romero
80. Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, et al. In-datacenter performance analysis of a tensor processing unit. In: Proceedings of the 44th Annual International Symposium on Computer Architecture. 2017 Presented at: ISCA '17; June 24-28, 2017; Toronto, ON. [doi: 10.1145/3079856]
81. Wang YE, Wei GY, Brooks D. Benchmarking TPU, GPU, and CPU platforms for deep learning. arXiv Preprint posted online on July 24, 2019 [FREE Full text] [doi: 10.1090/mbk/121/79]
82. Azure Maia for the era of AI: from silicon to software to systems. Azure. 2024 Apr 3. URL: https://azure.microsoft.com/en-us/blog/azure-maia-for-the-era-of-ai-from-silicon-to-software-to-systems/ [accessed 2024-10-14]
83. Smith A, Chapman E, Patel C, Swaminathan R, Wuu J, Huang T. 11.1 AMD InstinctTM MI300 series modular chiplet package – HPC and AI accelerator for exa-class systems. In: Proceedings of the IEEE International Solid-State Circuits Conference. 2024 Presented at: ISSCC 2024; February 18-22, 2024; San Francisco, CA. [doi: 10.1109/isscc49657.2024.10454441]
84. AMD CDNA™ 3 Architecture. Advanced Micro Devices. URL: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf [accessed 2024-10-14]
85. AMD delivers leadership portfolio of data center AI solutions with AMD instinct MI300 series. AMD. 2023 Dec 6. URL: https://ir.amd.com/news-events/press-releases/detail/1173/amd-delivers-leadership-portfolio-of-data-center-ai [accessed 2024-10-14]

XSL•FO
RenderX

86. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access 2023;11:31866-31879. [doi: 10.1109/access.2023.3262138]
87. A technical analysis of confidential computing. The Confidential Computing Consortium. 2022 Nov. URL: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf [accessed 2024-10-14]
88. Confidential computing: hardware-based trusted execution for applications and data. The Confidential Computing Consortium. 2022 Nov. URL: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC_outreach_whitepaper_updated_November_2022.pdf [accessed 2024-09-30]
89. Brown D. Confidential computing: an AWS perspective. Amazon Web Services. 2021 Aug 24. URL: https://aws.amazon.com/blogs/security/confidential-computing-an-aws-perspective/ [accessed 2024-09-30]
90. Confidential computing. Google. URL: https://cloud.google.com/security/products/confidential-computing [accessed 2024-10-14]
91. Azure offerings. Microsoft. 2024 May 21. URL: https://learn.microsoft.com/en-us/azure/confidential-computing/overview-azure-products [accessed 2024-10-14]
92. Confidential computing. Oracle. URL: https://docs.oracle.com/en-us/iaas/Content/Compute/References/confidential_compute.htm [accessed 2024-10-14]
93. Bean JD, Raghu A, Rudzitis A. Protect sensitive data in use with AWS confidential computing. Amazon Web Services. 2023. URL: https://d1.awsstatic.com/events/Summits/reinvent2023/CMP307_Protect-sensitive-data-in-use-with-AWS-confidential-computing.pdf [accessed 2024-10-14]
94. Confidential VM attestation. Google Cloud. URL: https://cloud.google.com/confidential-computing/confidential-vm/docs/attestation [accessed 2024-10-14]
95. Choudhury O, He C, Avestimehr S, Bhargavi D, Ravipati VS, Aziz W, et al. Federated learning on AWS with FedML: health analytics without sharing sensitive data – part 1. Amazon Web Services. 2023 Jan 13. URL: https://aws.amazon.com/blogs/machine-learning/part-1-federated-learning-on-aws-with-fedml-health-analytics-without-sharing-sensitive-data/ [accessed 2024-10-14]
96. Cross-silo and cross-device federated learning on Google Cloud. Google Cloud. 2024. URL: https://cloud.google.com/architecture/cross-silo-cross-device-federated-learning-google-cloud [accessed 2024-10-14]
97. Federated learning in Azure ML. GitHub. URL: https://github.com/Azure-Samples/azure-ml-federated-learning [accessed 2024-10-14]
98. Peterson D, Kanani P, Marathe VJ. Private federated learning with domain adaptation. arXiv Preprint posted online on December 13, 2019. [doi: 10.48550/arXiv.1912.06733]
99. Eichner H, Ramage D, Bonawitz K, Huba D, Santoro T, McLarnon B, et al. Confidential federated computations. arXiv Preprint posted online on April 16, 2024. [doi: 10.48550/arXiv.2404.10764]
100. Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024 Jun;6(6):e428-e432. [doi: 10.1016/s2589-7500(24)00061-x]
101. Goldberg CB, Adams L, Blumenthal D, Brennan PF, Brown N, Butte AJ, et al. To do no harm - and the most good - with AI in health care. Nat Med 2024 Mar;30(3):623-627. [doi: 10.1038/s41591-024-02853-7] [Medline: 38388841]
102. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
103. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023 Oct 30. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [accessed 2024-10-14]
104. Section 1557 of the Patient Protection and Affordable Care Act. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/civil-rights/for-individuals/section-1557/index.html [accessed 2024-10-14]
105. LaRose C, Edwards E. 1557 final rule protects against bias in health care algorithms. National Health Law Program. URL: https://healthlaw.org/1557-final-rule-protects-against-bias-in-health-care-algorithms/ [accessed 2024-10-14]
106. The Bletchley Declaration by Countries attending the AI Safety Summit, 1-2 November 2023. United Kingdom Government. 2023 Nov 1. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [accessed 2024-10-14]
107. Artificial intelligence in health care: implications for patient and workforce safety. Institute for Healthcare Improvement. URL: https://www.ihi.org/resources/publications/artificial-intelligence-health-care-implications-patient-and-workforce#downloads [accessed 2024-10-14]
108. Shaping Europe's digital future: AI Act. European Commission. URL: https://tinyurl.com/455s5hb8 [accessed 2024-10-02]
109. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf [accessed 2024-10-14]
110. Rebedea T, Dinu R, Sreedhar M, Parisien C, Cohen J. NeMo guardrails: a toolkit for controllable and safe LLM applications with programmable rails. arXiv Preprint posted online on October 16, 2023 [FREE Full text] [doi: 10.18653/v1/2023.emnlp-demo.40]

XSL•FO
RenderX

111. Sun L, Huang Y, Wang H, Wu S, Zhang Q, Li Y, et al. Trustllm: trustworthiness in large language models. arXiv Preprint posted online on January 10, 2024. [doi: 10.48550/arXiv.2401.05561]
112. Software as a Medical Device (SAMD): clinical evaluation. U.S. Food & Drug Administration. 2017 Dec 8. URL: https://www.fda.gov/media/100714/download [accessed 2024-10-14]
113. Artificial intelligence and medical products: how CBER, CDER, CDRH, and OCP are working together. U.S. Food and Drug Administration. URL: https://www.fda.gov/media/177030/download?attachment [accessed 2024-10-14]
114. Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, et al. Do we still need clinical language models? arXiv Preprint posted online on February 16, 2023. [doi: 10.48550/arXiv.2302.08091]
115. Yu KH, Healey E, Leong TY, Kohane IS, Manrai AK. Medical artificial intelligence and human values. N Engl J Med 2024 May 30;390(20):1895-1904. [doi: 10.1056/NEJMra2214183] [Medline: 38810186]
116. Hartmann V, Suri A, Bindschaedler V, Evans D, Tople S, West R. SoK: memorization in general-purpose large language models. arXiv Preprint posted online on October 24, 2023. [doi: 10.48550/arXiv.2310.18362]
117. Amazon EC2 P4 instances. Amazon Web Services. URL: https://aws.amazon.com/ec2/instance-types/p4/ [accessed 2024-10-14]
118. Azure machine learning pricing. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/machine-learning/ [accessed 2024-10-14]
119. NDasrA100_v4 sizes series. Microsoft. 2024 Aug 22. URL: https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/ndasra100v4-series?tabs=sizebasic [accessed 2024-10-14]
120. VM instance pricing. Google Cloud. URL: https://cloud.google.com/compute/vm-instance-pricing [accessed 2024-10-14]
121. Compute pricing. Oracle. URL: https://www.oracle.com/cloud/compute/pricing/#compute-gpu [accessed 2024-10-14]
122. Compute and EC2 instance savings plans. Amazon Web Services. URL: https://aws.amazon.com/savingsplans/compute-pricing/ [accessed 2024-10-14]
123. Gurney M. meta-llama / llama. GitHub. URL: https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md [accessed 2024-10-14]
124. Amazon EC2 spot instances pricing. Amazon Web Services. URL: https://aws.amazon.com/ec2/spot/pricing/ [accessed 2024-10-14]
125. Linux virtual machines pricing. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/#pricing [accessed 2024-10-14]
126. Spot VMs pricing. Google Cloud. URL: https://cloud.google.com/spot-vms/pricing [accessed 2024-10-17]
127. Cloud cost estimator. Oracle. URL: https://www.oracle.com/cloud/costestimator.html [accessed 2024-10-17]
128. Hsieh CJ, Si S, Yu FX, Dhillon IS. Automatic engineering of long prompts. arXiv Preprint posted online on November 16, 2023. [doi: 10.48550/arXiv.2311.10117]
129. AWS Bedrock pricing. Amazon Web Services. URL: https://aws.amazon.com/bedrock/pricing/ [accessed 2024-10-14]
130. Azure OpenAI Service pricing. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/ [accessed 2024-10-14]
131. Vertex AI pricing. Google Cloud. URL: https://cloud.google.com/vertex-ai/pricing [accessed 2024-10-14]
132. Generative AI service pricing. Oracle. URL: https://www.oracle.com/artificial-intelligence/generative-ai/generative-ai-service/pricing/ [accessed 2024-10-14]
133. McTear M. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Williston, VT: Morgan & Claypool Publishers; 2020.
134. Goodman RS, Patrinely JR, Stone CAJ, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 02;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
135. Sezgin E. Redefining virtual assistants in health care: the future with large language models. J Med Internet Res 2024 Jan 19;26:e53225 [FREE Full text] [doi: 10.2196/53225] [Medline: 38241074]
136. Haberle T, Cleveland C, Snow GL, Barber C, Stookey N, Thornock C, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc 2024 Apr 03;31(4):975-979. [doi: 10.1093/jamia/ocae022] [Medline: 38345343]
137. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
138. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022 Feb 10;10(2):e32875 [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
139. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]

XSL•FO
RenderX

140. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RT, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. Ann Med 2024 Dec;56(1):2302980 [FREE Full text] [doi: 10.1080/07853890.2024.2302980] [Medline: 38466897]
141. Wollny S, Schneider J, Di Mitri D, Weidlich J, Rittberger M, Drachsler H. Are we there yet? - a systematic literature review on chatbots in education. Front Artif Intell 2021 Jul 15;4:654924 [FREE Full text] [doi: 10.3389/frai.2021.654924] [Medline: 34337392]
142. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. J Med Internet Res 2021 Jan 06;23(1):e19928 [FREE Full text] [doi: 10.2196/19928] [Medline: 33404508]
143. Schulhoff S, Ilie M, Balepur N, Kahadze K, Liu A, Si C, et al. The prompt report: a systematic survey of prompting techniques. arXiv Preprint posted online on June 6, 2024. [doi: 10.48550/arXiv.2406.06608]
144. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Clinical text summarization: adapting large language models can outperform human experts. Research Square Preprint posted online on October 30, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3483777/v1] [Medline: 37961377]
145. Hudelson P, Chappuis F. Using voice-to-voice machine translation to overcome language barriers in clinical communication: an exploratory study. J Gen Intern Med 2024 May;39(7):1095-1102 [FREE Full text] [doi: 10.1007/s11606-024-08641-w] [Medline: 38347346]
146. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey. arXiv Preprint posted online on February 9, 2024. [doi: 10.48550/arXiv.2402.06196]

Abbreviations

AI: artificial intelligence
API: application programming interface
AWS: Amazon Web Services
CCE: confidential computing environment
CFL: confidential federated learning
CSP: cloud service provider
DL: deep learning
DNN: deep neural network
EDW: enterprise data warehouse
EHR: electronic health record
ETL: extract, transform, and load
FL: federated learning
FTP: fine-tuned pathway
GCP: Google Cloud Platform
GDPR: General Data Protection Regulation
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
KMS: Key Management Service
KPI: key performance indicator
LGBTQ: lesbian, gay, bisexual, transgender, and queer
LLI: Lucian Leape Institute
LLM: large language model
ML: machine learning
NLP: natural language processing
OBP: out-of-the-box pathway
OCI: Oracle Cloud Infrastructure
PHI: protected health information
SaaS: software as a service
TEE: trusted execution environment
TPU: tensor processing unit
TSP: training from scratch pathway
VM: virtual machine

XSL•FO
RenderX

Edited by A Coristine; submitted 11.07.24; peer-reviewed by K Srinivasan, A Abreu, Q Shi; comments to author 20.08.24; revised version received 28.08.24; accepted 16.09.24; published 14.11.24.

Please cite as:
Nagarajan R, Kondo M, Salas F, Sezgin E, Yao Y, Klotzman V, Godambe SA, Khan N, Limon A, Stephenson G, Taraman S, Walton N, Ehwerhemuepha L, Pandit J, Pandita D, Weiss M, Golden C, Gold A, Henderson J, Shippy A, Celi LA, Hogan WR, Oermann EK, Sanger T, Martel S
Economics and Equity of Large Language Models: Health Care Perspective
J Med Internet Res 2024;26:e64226
URL: https://www.jmir.org/2024/1/e64226
doi:10.2196/64226
PMID:

©Radha Nagarajan, Midori Kondo, Franz Salas, Emre Sezgin, Yuan Yao, Vanessa Klotzman, Sandip A Godambe, Naqi Khan, Alfonso Limon, Graham Stephenson, Sharief Taraman, Nephi Walton, Louis Ehwerhemuepha, Jay Pandit, Deepti Pandita, Michael Weiss, Charles Golden, Adam Gold, John Henderson, Angela Shippy, Leo Anthony Celi, William R Hogan, Eric K Oermann, Terence Sanger, Steven Martel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Challenges and Lessons Learned Building a New UK Infrastructure for Finding and Accessing Population-Wide COVID-19 Data for Research and Public Health Analysis: The CO-CONNECT Project

Emily Jefferson[1,2], BSc, PhD; Gordon Milligan[3], BSc, MSc; Jenny Johnston[3]; Shahzad Mumtaz[3], BSc, MCS, PhD; Christian Cole[1,3], BSc, PhD; Joseph Best[4]; Thomas Charles Giles[4], BSc, PhD; Samuel Cox[4], MMath, PhD; Erum Masood[3], BE, MSc; Scott Horban[3], BSc; Esmond Urwin[4], BEng, MSc, PhD; Jillian Beggs[5], BSc; Antony Chuter[5]; Gerry Reilly[2], BSc, MSc; Andrew Morris[2], MD, PhD; David Seymour[2], BSc; Susan Hopkins[6], BA, MA, MSc, MD; Aziz Sheikh[7], BSc, MBBS, MSc, MD; Philip Quinlan[4], BSc, PhD

[1]Population Health and Genomics, School of Medicine, University of Dundee, Dundee, United Kingdom
[2]Health Data Research UK, London, United Kingdom
[3]Health Informatics Centre, School of Medicine, University of Dundee, Dundee, United Kingdom
[4]School of Medicine, University of Nottingham, Nottingham, United Kingdom
[5]Patient and Public Representatives, United Kingdom
[6]Public Health England, London, United Kingdom
[7]Centre for Population Health Sciences, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Emily Jefferson, BSc, PhD
Population Health and Genomics
School of Medicine
University of Dundee
The Health Informatics Centre
Ninewells Hospital and Medical School
Dundee, DD2 1FD
United Kingdom
Phone: 44 01382383943
Email: e.r.jefferson@dundee.ac.uk

Abstract

The COVID-19-Curated and Open Analysis and Research Platform (CO-CONNECT) project worked with 22 organizations across the United Kingdom to build a federated platform, enabling researchers to instantaneously and dynamically query federated datasets to find relevant data for their study. Finding relevant data takes time and effort, reducing the efficiency of research. Although data controllers could understand the value of such a system, there were significant challenges and delays in setting up the platform in response to COVID-19. This paper aims to present the challenges and lessons learned from the CO-CONNECT project to support other similar initiatives in the future. The project encountered many challenges, including the impacts of lockdowns on collaboration, understanding the new architecture, competing demands on people's time during a pandemic, data governance approvals, different levels of technical capabilities, data transformation to a common data model, access to granular-level laboratory data, and how to engage public and patient representatives meaningfully on a highly technical project. To overcome these challenges, we developed a range of methods to support data partners such as explainer videos; regular, short, "touch base" videoconference calls; drop-in workshops; live demos; and a standardized technical onboarding documentation pack. A 4-stage data governance process emerged. The patient and public representatives were fully integrated team members. Persistence, patience, and understanding were key. We make 8 recommendations to change the landscape for future similar initiatives. The new architecture and processes developed are being built upon for non–COVID-19–related data, providing an infrastructural legacy.

(J Med Internet Res 2024;26:e50235) doi:10.2196/50235

XSL•FO
RenderX

KEYWORDS

COVID-19; infrastructure; trusted research environments; safe havens; feasibility analysis; cohort discovery; federated analytics; federated discovery; lessons learned; population wide; data; public health; analysis; CO-CONNECT; challenges; data transformation

Introduction

In 2017, the UK Clinical Research Collaboration Tissue Directory and Coordination Centre (TDCC) [1] started a program to understand the processes required to allow federated discovery of biobanks based on the clinical and biomedical data they hold. There were many federated software solutions providing data discovery and data access, for example, TriNetx [2], i2b2 [3]/SHRINE [4], European Health Data & Evidence Network [5], iQVIA [6], and Leaf [7]. We chose the BC Platforms [8] solution for the TDCC work as they were a global leader in the field; had proven security and data confidentiality mechanisms built in; and most importantly, were willing to open their application programming interfaces such that we could implement new tools around their existing system. This program launched a beta version in February 2020 with 4 initial biobanks (Avon Longitudinal Study of Parents and Children [ALSPAC], National Institute for Health and Care Research [NIHR] Bioresource, Generation Scotland, and the Asymptomatic COVID19 in Education [ACE] cohort). The work also piloted the mechanisms in which to convert data to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) [9]. In March 2020, this was pivoted as part of the response to the COVID-19 pandemic, and the tools, know-how, and approaches were pivoted to the response to the pandemic.

The COVID-19-Curated and Open Analysis and Research Platform (CO-CONNECT) project was launched in November 2020, with funding from the Department of Health and Social Care UK Research and Innovation COVID-19 Rapid Response Initiative. At the time, new data were being generated by newly funded research projects and by the National Health Service (NHS) response to COVID-19, such as laboratory data from testing centers, in addition to routinely collected health data [10]. Research using these data was fundamental to the UK's fight against the pandemic; however, it was challenging for researchers and public health agencies to find which data were being captured and by which organization or organizations [11]. This lack of visibility was further complicated by the 4 devolved nations within the United Kingdom each collecting data in different ways [12,13]. As the data were held by many different organizations, researchers had to contact multiple groups to determine if they held the data that could answer the key questions of policy interest. This took data analysis time away from researchers and time responding to the pandemic away from the organizations collecting the data. This was not a new issue [13], but its impact on timely research was made stark with COVID-19 [14-16]. Equally from a technical perspective, the challenges were identical to those sought to be solved by the earlier work of the TDCC.

Once researchers or public health agencies had found relevant data needed for their study, they required data sharing and data governance approvals. This took significant time, especially if data were required from multiple organizations, even considering the streamlined processes put in place during the pandemic (such as Control of Patient Information [COPI] notices) [17,18].

CO-CONNECT was funded to respond to the needs of the research community to find and access data at pace while maintaining patient confidentiality [19]. The new CO-CONNECT infrastructure enables approved researchers and approved individuals from public health organizations to run aggregate-level, dynamic queries across UK-wide datasets in a federated manner. They can instantaneously discover the data that are available and run analyses with data remaining in situ (Figure 1). For example, "How many people have had a polymerase chain reaction (PCR) test that was positive and are younger than 40 years old?" An aggregate-level count of the people who meet the criteria is returned, for example, 102 people across 4 datasets. Queries are initiated from a single common website: the Health Data Research Innovation Gateway (Gateway) from within the Cohort Discovery Tool user interface [20]. Through a streamlined data governance application process and semiautomated pipeline, researchers can then request access to a linked, standardized, pseudonymized, project-specific subset of data to answer specific research questions. This "do once, reuse often" paradigm reduces the challenges and complexity of finding and making data available to answer urgent questions in the national interest.

The CO-CONNECT architecture (described in detail in Jefferson et al [21] and based on the early TDCC work) enables data sources to make their data discoverable and standardized while the data sources still maintain control over who is given access to the data. Each data source (termed CO-CONNECT data partners) hosts a secure virtual machine (VM) within their own network. The VM is separate from the location where identifiable data are stored. The data partners create a pseudonymized and standardized version of their data and deposit these within a database on the VM. Software installed within the VM runs queries received from the Gateway [22] against the database and returns an aggregate count of the number of individuals that meet the search criteria. The results from the queries run across all the data partners and are automatically collated and displayed within the Gateway for analysis by the user who initiated the query.

To configure the federated platform, we worked across 22 different data partners to on-board their data. Data partners included academic groups collecting data from a consented research cohort and national Trusted Research Environments (TREs) collecting population-wide, routinely collected data. We spent significant effort prior to grant submission to explain the benefits of the solution and included a representative from each data partner as a coapplicant on the proposal.

XSL•FO
RenderX

Figure 1. Schematic of federated querying from the Health Data Research Innovation Gateway. PCR: polymerase chain reaction.

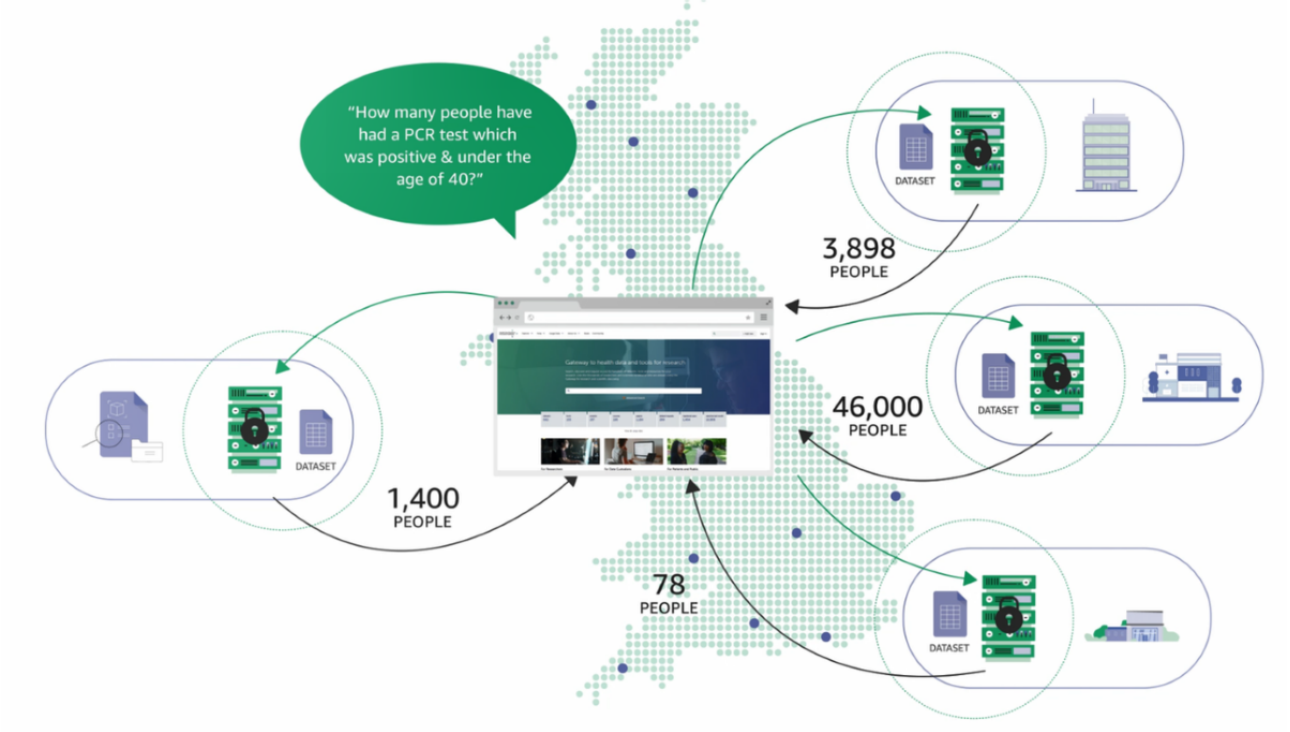

The Cohort Discovery tool is live within the Gateway, supporting authenticated researchers to find and access data—delivering the main output from CO-CONNECT [20]. However, there were significant challenges and delays along the journey. This paper describes the challenges faced and solutions developed and provides a series of recommendations to streamline future similar initiatives and access to UK-wide data for research and innovation.

Challenges

Addressing Cultural Norms During an Emergency Pandemic Response

It was initially challenging for TREs to understand how the infrastructure we were building was different from the "normal" architecture of TREs. Figure 2A illustrates how TREs generally support research projects. TREs have an area within their infrastructure that hosts a data repository of population-wide information from a range of databases. For each research project, a project-specific set of data is copied from the data repository and placed in a separate area of their infrastructure where researchers can analyze the data. The project-specific set of data only includes the minimum amount of data that are needed to answer the specific scientific question and researchers can only access their specific dataset or datasets. There are stringent ingress and egress disclosure controls on the researcher area and there is no access to the internet. Only anonymous aggregate-level data can be exported from the researcher area, which are manually checked by trained TRE staff before release. Therefore, the primary role of the TRE is not to make data discoverable, nor to make it easy to link data across TREs, but to focus on providing a governance wrapper and keeping data secure within their given environments. Understanding the CO-CONNECT architecture required a mindset change for TREs, who are used to working in a set way.

For researchers, they are used to having collaborative projects, in which a group of researchers will come together, form a consortium, and generate and analyze data. Within a consortium, data may be shared between each of the entities involved, and data sharing agreements will be in place. The researchers will most likely have an agreed publication strategy, which also recognizes how involvement will equate to authorship and the research outputs. The reuse of data and the publishing of data are often only seen as something to do at the end of the project, as part of archiving or validating the results as a requirement from funders or journals. Therefore, a project that sought to make data discoverable before publication, even if just basic metadata, was a very different ethos and culture for academic researchers to operate.

Patient and public involvement and engagement (PPIE) have become a cornerstone of medical research to ensure the public voice is heard and indeed where public and patient members help to cocreate the solutions. Those with lived experience may be able to inform on the development of a new medicine or medical device or service, and therefore, PPIE on medically oriented projects are commonplace. CO-CONNECT was a very technical program in parts and we had a strong desire not to exclude our PPIE members from the technical design components so that we could ensure full transparency and scrutiny across all decisions. While much guidance on PPIE does exist, there was very little on how to embed PPIE into a software development team for example.

XSL•FO
RenderX

Figure 2. Comparison of traditional TRE architecture versus the new CO-CONNECT architecture. CO-CONNECT: COVID-19-Curated and Open Analysis and Research Platform; HDR: Health Data Research; OMOP: Observational Medical Outcomes Partnership; TRE: Trusted Research Environment; VM: virtual machine.

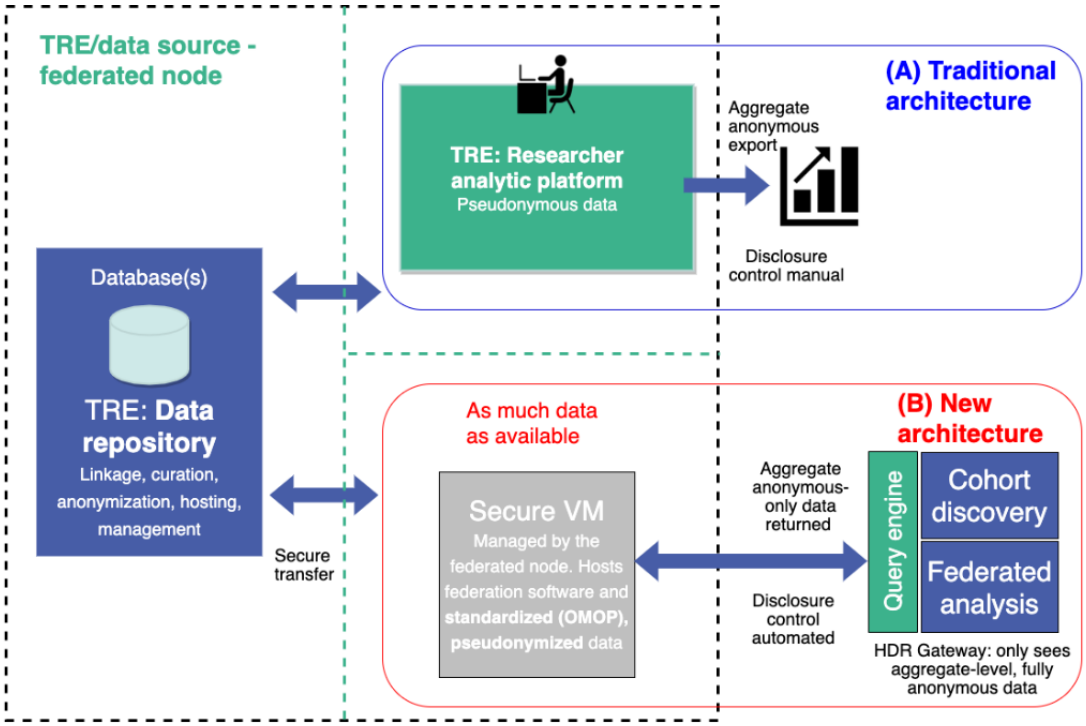

Awareness of Data Discovery Approaches

Although there are many different solutions that provide data discovery and federated analysis, most of the data partners we worked with were unfamiliar with such solutions (Figure 2B). For both the academic and TRE teams, this meant quite a significant amount of effort in helping them to understand the core concepts: for example, for a researcher, how they could make data discoverable without sharing it and protecting their datasets before they had published. For TREs, this was about how the technical approach could be consistent with their existing security models and approaches. Federated and data discovery systems are not new, but it was clear that they were still not understood or known by a significant portion of the community. This was especially true when it came to local configuration and how these approaches could be demonstrated to be consistent with the five safes framework.

Standardized Software to Profile Data Within Trusted Research Environments

CO-CONNECT uses the OMOP CDM [9] to enable federated queries with comparable answers across data partners. We needed metadata describing the datasets held by data partners to use as input to a tool that would convert their datasets into OMOP format [23-25]. Metadata are anonymous data about the data that are captured, for example, field names, ranges of data within the fields, and descriptions of what data are collected in each field. We sought to use the WhiteRabbit data profiling tool from Observational Health Data Science and Informatics to generate the metadata [26]. The academic community was largely happy to install this, but for the TREs that operate within a highly controlled environment, the software had to undergo several layers of security assessment to assure this open-source software to confirm that only metadata were shared (therefore, data sharing agreements are not required) and then to test the solution and confirm the outputs were not disclosive.

Data Governance

Overview

Implementing the technological components of the project was relatively straightforward. However, although the data partners were supportive of the initiative, obtaining data governance approvals, especially for the national TREs, was a significant challenge. Most of the time spent with TREs was for them to understand the process and which groups needed to approve this new infrastructure and process. The actual governance approvals took a long time. In some cases, but not all, governance approval was a relatively short time in comparison to collaborating with the TREs to populate forms with the information they required prior to submission. There were several governance challenges.

No Existing Governance Framework

There was no process that we could simply apply to undertake the work; indeed, it was also not clear whether the CO-CONNECT team should be applying the process or simply helping each data partner to form an opinion. The design of the system meant that at no stage would any third party have access to the data. The questions asked within all data governance approval documents assumed that someone would be accessing the data. What we were doing was to provide support to enable each data partner to connect to a national data discovery framework, via software. Most TREs did not have a documented process to assess new software being installed within their environments for a research project (as it was not normally requested). They did have methods for assessing third-party

XSL•FO
RenderX

provider software for their core systems, but then that process would normally have followed a procurement route, where a strategic decision and funding were made to redesign or develop parts of their infrastructure.

The academic-led studies were also supportive of the initiative but were not sure of the formal approaches to approve a project such as CO-CONNECT. The CO-CONNECT team did mandate that all data partners undertake a data protection impact assessment to ensure that each organization considers its involvement carefully. Overall, once data partners understood how the architecture worked, they could see the benefits of the solution and how security and patient confidentiality were protected, but the national TREs were unsure of how to say "yes."

Bespoke Approaches Across Data Partners

Each data partner subsequently sought approval from the various committees to ensure all bases were covered. Each committee, for each data partner, had different document requirements, different formats, and different levels of detail. For some national TREs, 4 different forms were required to go to 4 different committees or approvers, with each form being >100 pages long. Drafting such information across many data partners took significant time.

Single Points of Failure for Approval of Data Governance Applications

Many data partners only had a single individual within their organization who could formally approve and sign off the data governance applications. The pandemic resulted in a much higher than usual number of applications, especially because of the temporary COPI notice in England [18], which resulted in an increase in the volume of research and data sharing requests. We were told in one instance that there was a 6-month wait for applications to be reviewed due to the backlog and only a single person with the authority to approve requests. The technical nature of the project meant that some institutions had preferred individuals who could review this type of application. This led to further bottlenecks when this individual was time-constrained or unavailable for periods of time due to other priorities or unforeseen circumstances.

Technical Knowledge Gap Within Data Governance Approval Teams

Given there was not a defined process, all TREs felt that CO-CONNECT should still apply to the data access teams for access to the data. The paradox though was that the CO-CONNECT team never required access to the data and the whole architecture was designed such that the only teams accessing the data are the teams within the TREs. It was the software, rather than a person, that was accessing the data to enable questions to be answered, which naturally led the data access teams to request more technical details about the software. However, simultaneously, the data security teams within the TREs were independently assessing the software for compliance. The TREs had processes that were more established to assess a typical research project in which a researcher requires access to data, for a particular project. As the application to set up the infrastructure within each data partner was highly technical, it was challenging to find the right level of detail to be both clear and understandable taking into account the knowledgebase of data access teams. In the case of the national TREs, there were different groups or committees assessing the technical and security aspects of the project; however, it did not appear that the data access groups were content to rely on the approval of the technical groups. The data governance committees asked for more technical details, but in some instances, appeared not to have the required expertise to assess the information.

Different Levels of Technical Capability Across Data Partners

There was a vast difference between the technical capabilities of different data partners. National TREs had different teams of people covering specialist areas, such as data sharing, disclosure control, cyber security, information technology, and data governance, for the CO-CONNECT team to interact with for each technical component. Groups hosting research cohorts of data often only had small teams. The documentation and technical support needed to be tailored to each data partner which took time.

Production of Linked Extracts of Data From Across Different Databases Held by the Same Organization

Once the data governance had been approved, data partners needed to set up a pseudonymized version of their data for federated querying. For research cohorts of data, this was relatively straightforward, but for national TREs, this was much more complex. Population-level data are held within different databases often managed by different teams within the same organization. The infrastructure was not in place for automated linkage and extraction of data from across these databases. For example, the data we required from Scotland were very similar to the EAVE II dataset [27]. However, we were told that we could not request a copy of that data as it was not a "real" dataset; instead, it was extracts of data from many sources, which would take months to link and clean before being useable. It was unfortunate that we could not reuse the linking and cleaning efforts of others.

Laboratory Data Feed

Research using granular-level (such as antibody volume and assay type rather than just positive or negative), serology data was key to the COVID-19 response. In the early stages of the pandemic, it was not clear if individuals who had contracted the disease would be immune, and if so, how long that immunity would last. Vendors developed new assays for COVID-19, but data were not easily accessible to calibrate these assays with evidence-driven cutoff ratios. A workstream of the CO-CONNECT project was to collaborate with health boards or trusts to capture granular-level data and to make this data, linked to longitudinal health care records, available for research. Although we discovered that the laboratory systems could be relatively easily configured to capture the additional granular-level data, we found that sending this data onward to national groups collecting population-wide health data was extremely challenging. In England, laboratory data are not routinely collated at a national level and shared with NHS

XSL•FO
RenderX

England (the organization that manages other English national datasets), although there is a long-term plan to build such a pipeline. In Scotland, although there is a pipeline set up for health boards to share data with Public Health Scotland (the Scottish national organization that manages relevant Scottish national datasets), they could not accept new laboratory data fields without risking the stability of antiquated systems, and therefore, could not approve boards sending this additional data.

PPIE on a Highly Technical Project

It was key for CO-CONNECT to include patient and public representatives in the project in a meaningful way. We wanted to ensure their voice was heard in the way that the system was designed and the controls we were applying to protect patients confidentially. However, this was a technical project, and at times, it was challenging to explain complex solutions to a nontechnical audience.

Lockdown Impacts

CO-CONNECT was initiated during a pandemic when people could not meet in person due to national lockdown restrictions. Outside of a pandemic, such a project would have started with a face-to-face kickoff meeting to get everyone on the same page and to build new relationships. Such a meeting was not possible in person (or even virtually due to the added diary pressures caused by the pandemic). This inability to build relationships and efficiently communicate to everyone at once hampered the timeliness and cohesiveness of the project.

Solutions and Lessons Learned

Support for Data Partners

We adapted our approach to support data partners to progress the project while under the added pressures of the pandemic using the following strategies.

Regular Short Videoconference Calls

Regular emailing with data partners was generally not as successful for progressing issues as regular “touch base” video calls. The video calls provided a rapid way of addressing any outstanding questions, helped build relationships, and developed or maintained momentum.

Standardized Technical Onboarding Pack

Data partners were asking the same questions but in different ways. We initially provided bespoke answers to each data partner based on how they had framed the question. However, over time, we recognized the patterns of the questions and covered the answers in a standardized format within the technical onboarding pack. This was part of the process to scale the onboarding process. We provided the pack as an appendix to all of the different types of governance documentation, and where possible, pointed answers to the questions to the section in the pack rather than bespoke answers [28].

Animated Explainer Videos

We worked with a company (Mode13) to develop a series of short, animated videos explaining the design of the federated platform and how results to queries could be securely set externally while protecting data confidentiality [29-32]. Data partners and their data governance committees found that they grasped the solutions much faster than they did reading about them in the abstract. The videos had active input from our PPIE group.

Demonstrating the Live System

Demonstrating the system once it was live with the first sets of data also helped data partners earlier on in their journey to understand the benefits of the project and how it could be securely configured.

Technical Videos

We developed a series of videos aimed at the data partner technical teams who would be implementing the solution [32]. The technical teams found these videos helped them to grasp the solution quickly and could then follow up with more details within the technical onboarding pack.

Workshops and Drop-In Sessions

Bringing data partners together for shared workshops, drop-in sessions, question and answer sessions, and demo sessions helped to develop a shared understanding and confidence in the solution. We also ran a workshop with representatives from the Information Commissioner’s Office and Health Data Research UK (UK National Institute of Health Data Research with expertise in data governance) [33] to address any concerns from the data governance groups. We brought data partner technical teams together for technical workshops with representatives from BC Platforms and the CO-CONNECT technical team.

Engaging the Academic Community

Engaging academic data partners (those who had collected data from research cohorts) was achieved by inviting academics to present their research in monthly CO-CONNECT all-partners meetings. This improved understanding of the benefits of the project and the details of the datasets held by the other data partners; developed collaborations; and discussed exemplar research projects that could use the infrastructure.

Emergence of a 4-Stage Governance Approval Progress for Unconsented Population-Scale Data

Although each national TRE was initially unsure of how to approve the project, a similar 4-stage governance approval process emerged across TRE organizations: a Data Protection Impact Assessment, a data governance request, a Disclosure Control Assessment, and a Security Risk Assessment. Understanding which groups needed to be consulted before the organizations could say “yes” will help similar new infrastructure projects in the future. However, as the groups all required similar information, it could be argued that a more streamlined approach would be for all groups to meet to review the application together based on a single application form containing all the required information. Adoption of the Health Data Research UK standardized data access request form [34] across data partners might have been helpful, however, it was not mature enough to be adopted at the time of the project.

XSL•FO
RenderX

The Goldilocks Approach to Data Governance Approvals

Different individuals or committees required different levels of detail on application forms. Subsequently, it was difficult to determine the level of detail required at the offset. On many occasions, we were asked to remove details from an application, only to be asked for those same details to be added as part of the reviewer's response. The lesson learned was that there is a Goldilocks level of detail required. It is unfortunate that Goldilocks' taste buds were different for each committee within each organization.

We remain unsure of how to address this challenge directly other than perhaps some more guidance throughout the process in the form of a consistent individual within the organization to guide us. This individual could then attend any governance review meetings and answer any questions that may arise to help the approvals process be more efficient and expedient. This process was used within one national TRE for its data access and data governance applications, and it was the only national TRE to complete all tasks and have data accessible via the tool.

Persistence, Patience, and Understanding

The CO-CONNECT timelines were tight, with the desire to positively impact the COVID-19 response. The delays that we experienced were frustrating, especially for the team directly working with data partners. However, we recognized that all collaborators were under immense pressure from competing demands and still tried their best to support the project. We tried to approach our communications with persistence, patience, and understanding. As with so many things in life, engendering change is hard and takes time, but even more so in the middle of a pandemic.

Development of a National COVID-19 Serology Laboratory Data Standard

To support the capture of granular-level, serology data, we worked across data partners and 3 different NHS boards or trusts to develop a new serology laboratory data standard. The standard included 12 data attributes providing the required fields for serology research using data from multiple different vendors. Details of the standard are presented in [35].

PPIE Group

CO-CONNECT had 2 lay coapplicants on the grant proposal. A PPIE group was set up and was heavily involved in all aspects of the project including attendance at workstream meetings. One of the lay coapplicants chaired our monthly all-partners meetings, demonstrating our commitment to PPIE. Another of our PPIE members was a poet and author and wrote several papers covering various aspects of the project including one paper published in a news outlet [36]. PPIE representatives were highly involved in the main project. They attended our workstream meetings and we scheduled time after the core meeting to answer any specific questions they might have and to summarize the conclusions of the meeting to this nontechnical audience. The PPIE group provided helpful feedback and edits to the narrative of the explainer videos and also reviewed the animations—helping us to pitch these videos to nontechnical audiences.

Recommendations

We recommend 10 key changes to the data landscape so the United Kingdom can respond more quickly to future pandemics and make similar projects more efficient and impactful:

- Single points of failure within organizations that have data and provide access to research could be reduced by training more people in assessing data requests and allowing delegated authority for approvals.
- Having an assigned individual within the organization who is your local "sponsor" proved incredibly useful. Some organizations have it once a project is approved, like the "data buddy" in ALSPAC [37], but having a similar process during the governance approvals could help smooth the process, especially for more complex applications.
- Technical expertise should be embedded within national and regional data governance application panels or committees.
- Data partners should adopt a standard data access request form (such as the one being developed by Health Data Research UK) [34] so that applications in different formats across different data partners are not required. Each application for data to a single data partner should not require multiple forms in different formats for different committees. Targeted information for each group could be provided within appendices rather than requiring the same information to be provided but requested in a slightly different way by each group.
- A light-touch governance process should become standard for the sharing of metadata or profiling data. This data does not contain identifiable information.
- Significant investment should be provided by the government or NHS to enhance data systems across the UK health care system, making them more robust, automated, and scalable. This is not only needed for data for the secondary purposes of research but also to provide timely, high-quality data to respond to clinical and administrative needs. As part of this, data standards adoption should be embedded in the dataset creation process to reduce the requirement for highly specialized, domain expert data engineering which is a limited pool of people who are in high demand in other sectors. If datasets were created to standardized data models, the requirement would be for validation and testing, which can be automated, as opposed to mapping and transformation which requires clinician and engineering input in addition to those steps. The specific standard model is not necessarily an obstacle if its specification is published or available, as mapping from one standard to another is a much less arduous task than mapping from no standard to a standard.
- A national laboratory data feed should be commissioned for NHS boards or trusts to robustly share data with national bodies collecting population-wide data. The infrastructure should be scalable and flexible, supporting additional laboratory tests and data fields as required over time. This

XSL•FO
RenderX

should support new laboratory standards as they are developed.

- To modernize the completion of the various governance forms, it would be beneficial if digital portals could be created in place of "paper" forms. These systems could easily provide features such as tracking differences, providing notification of submissions or changes or review, and any comments that have been made by reviewers or collaborators all in one place. They could also provide more guidance and be transparent in showing how the forms are progressing. More than one TRE has a system like this for 1 or 2 forms making the process simpler as everything was more transparent, and it was easier for multiple people to work on the forms simultaneously. It has added benefits of always being up to date with the latest form, and guidance from the host organization, and not falling into the trap of submitting a governance application only to be told that the Word template had changed since you started writing your document. The final benefit is that it provides a single source of truth for your application instead of multiple different versions of the same document circulating between collaborators and reviewers. Health Data Research UK is developing a standardized form and process that can be used across data custodians supporting such needs of the community [38].
- There should be a national process to approve software solutions for use across Health Boards or Trusts and Research Environments. We had to do a security assessment of the software at every location. This challenge could be reduced through a central list of approved or safe software.
- National funding schemes that support innovation in the national infrastructures are vital. The COVID-19 response placed a strain on the existing infrastructures and funding, such as CO-CONNECT, was seen as vital to the response. However, funding for innovation in research infrastructure has previously been hard to secure. The DARE UK program [39] offers a significant opportunity for the technical infrastructure to be given opportunities to innovate and support emerging technologies.

Long-Term Impacts

Although CO-CONNECT experienced delays, the project was successful in onboarding federated data for cohort discovery. The Cohort Discovery Tool is live within the Gateway and has been queried over 3600 times supporting researchers to carry out feasibility queries across multiple datasets. Data custodians are supported to promote their dataset and researchers can find out about datasets they otherwise would not have known about. At the time of publication, there are 19 research cohorts live within the tool and population-scale datasets from Scotland. NHS England worked with us to test the solution and provide metadata but could not go live with the solutions within the timeframe, due to other competing demands. Data governance has been approved in Wales and the system is ready to go live following final testing.

A new secure architecture has been approved and deployed by >18 organizations hosting both unconsented and consented datasets, which can be enhanced and reused. This new architecture includes new processes for onboarding new datasets, open-source software [21,23-25] to map datasets into the OMOP CDM, open application programming interfaces to enhance and expand the functionality, WhiteRabbit acceptance as software for profiling and producing metadata on sensitive datasets, acceptance of automated disclosure control, federated analytics capability using the same architecture, and datasets that have been converted to OMOP and can be reused by other research projects such as the pan-Europe European Health Data & Evidence Network project [5].

The software and new processes developed by CO-CONNECT will be taken forward by Health Data Research UK, to onboard new datasets covering different diseases; for example, the Alleviate data hub [40] is currently onboarding pain data from across the United Kingdom. The materials developed by CO-CONNECT can be used to onboard new datasets including a range of explainer videos and documentation for data custodians, technical teams, and the general public to understand the solution, benefits, and how we protect patient confidentiality.

The technology has demonstrated the flexibility to be converted to different health care challenges, as it was converted from consented biobanks to be used across large population-level TREs. This flexibility is achieved predominately by the OMOP CDM; therefore, so long as data can be represented in this data model, then the solution can be used for any future pandemic response. CO-CONNECT brought a new federated solution to groups generally unaware of such tools and changed the conversation from *why* we should do cohort discovery or federated analytics to *how* we should. From CO-CONNECT, many other initiatives are now spawning and seeking to embed federation as a core requirement. The work from CO-CONNECT now continues in the Health Data Research UK program under the Federated Analytics Program and has been further accelerated in the DARE UK program, which has accelerated the development of additional federated tooling across different TREs and different domains [41,42]. In England, the NHS Secure Data Environment program is creating regional TREs, and federation across these secure data environments is part of the program.

Acknowledgments

We acknowledge the support from CO-CONNECT funded by UKRI (MRC) and DHSC (NIHR), (MR/V03488X/1) and supported by a consortium of 22 partner organizations whose infrastructure made this research possible. This project was also funded by Health Data Research UK, which receives its funding from Health Data Research UK Ltd (HDR-5012) funded by the UK MRC, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation

XSL•FO
RenderX

(BHF), and the Wellcome Trust. Calum Macdonald developed the software and methodology and was involved in the data curation, investigation, and formal analysis.

Authors' Contributions

EJ and PQ led the funding acquisition and led the project, supervised the team, developed the methodology, and drafted and edited the manuscript. GM was a project manager of the project, supervised the software development, supported data curation, drafted and edited the manuscript, was involved in the investigation and formal analysis, and developed the methodology. JJ was a project manager of the project. SM, J Best, TCG, SC, EU, EM, and SH developed the software and methodology and were involved in the data curation, investigation, and formal analysis. CC drafted and edited the manuscript and supervised some of the team. GR, DS, SH, and AM were involved with the funding acquisition and methodology. J Beggs and AC provided patient and public involvement and engagement (PPIE) expertise and drafted and edited the manuscript. AS was involved with the funding acquisition and methodology, drafted and edited the manuscript, and supervised some of the team. All authors reviewed the final manuscript.

Conflicts of Interest

AS was co-investigator on the CO-CONNECT grant.

References

1. Quinlan PR, Lawrence E, Pourabdolla A, Sims JM, Gander A, Fuller B, et al. The UK clinical research collaboration (UKCRC) tissue directory and coordination centre: the UK's centre for facilitating the usage of human samples for medical research. Open J Bioresour 2017;4(1):6. [doi: 10.5334/ojb.31]
2. TriNetX. URL: https://trinetx.com/ [accessed 2022-11-14]
3. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
4. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009;16(5):624-630 [FREE Full text] [doi: 10.1197/jamia.M3191] [Medline: 19567788]
5. European Health Data Evidence Network. URL: https://www.ehden.eu/ [accessed 2021-08-11]
6. IQVIA. URL: https://www.iqvia.com/ [accessed 2022-11-14]
7. Dobbins NJ, Spital CH, Black RA, Morrison JM, de Veer B, Zampino E, et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. J Am Med Inform Assoc 2020;27(1):109-118 [FREE Full text] [doi: 10.1093/jamia/ocz165] [Medline: 31592524]
8. BC Platforms. URL: https://www.bcplatforms.com [accessed 2021-07-23]
9. OMOP CDM v5.3. URL: https://ohdsi.github.io/CommonDataModel/cdm53.html#Clinical_Data_Tables [accessed 2022-05-13]
10. Dron L, Kalatharan V, Gupta A, Haggstrom J, Zariffa N, Morris AD, et al. Data capture and sharing in the COVID-19 pandemic: a cause for concern. Lancet Digital Health 2022;4(10):e748-e756 [FREE Full text] [doi: 10.1016/S2589-7500(22)00147-9] [Medline: 36150783]
11. Nuthi K. Local data collection and coordinated data sharing can tackle the UK's health data divide. Center for Data Innovation. 2022. URL: https://datainnovation.org/2022/11/local-data-collection-and-coordinated-data-sharing-can-tackle-the-uks-health-data-divide/ [accessed 2022-12-23]
12. Digital health care across the UK: where are we now? The Nuffield Trust. 2022. URL: https://www.nuffieldtrust.org.uk/resource/digital-health-care-across-the-uk-where-are-we-now [accessed 2022-12-23]
13. Lugg-Widger FV, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: managing the morass. Int J Popul Data Sci 2018;3(3):432 [FREE Full text] [doi: 10.23889/ijpds.v3i3.432] [Medline: 34095522]
14. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
15. Cavallaro F, Lugg-Widger F, Cannings-John R, Harron K. Reducing barriers to data access for research in the public interest-lessons from COVID-19. BMJ Opinion. 2020 Jul 6. URL: https://blogs.bmj.com/bmj/2020/07/06/reducing-barriers-to-data-access-for-research-in-the-public-interest-lessons-from-covid-19/ [accessed 2022-05-13]
16. Vickery J, Atkinson P, Lin L, Rubin O, Upshur R, Yeoh E, et al. Challenges to evidence-informed decision-making in the context of pandemics: qualitative study of COVID-19 policy advisor perspectives. BMJ Global Health 2022 Apr;7(4):e008268 [FREE Full text] [doi: 10.1136/bmjgh-2021-008268] [Medline: 35450862]
17. Coronavirus (COVID-19): notification to organisations to share information. GOV.UK. URL: https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information [accessed 2023-01-13]

XSL•FO
RenderX

18. Control of patient information (COPI) notice. NHS Digital. URL: https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice [accessed 2023-04-26]
19. CO-CONNECT. CO-CONNECT | Unleashing the power of data through discovery. 2021. URL: https://co-connect.ac.uk/co-connect-home/ [accessed 2022-05-31]
20. Cohort discovery. HDRUK Innovation Gateway. URL: https://www.healthdatagateway.org/about/cohort-discovery [accessed 2022-03-10]
21. Jefferson E, Cole C, Mumtaz S, Cox S, Giles TC, Adejumo S, et al. A hybrid architecture (CO-CONNECT) to facilitate rapid discovery and access to data across the United Kingdom in response to the COVID-19 pandemic: development study. J Med Internet Res 2022 Dec 27;24(12):e40035 [FREE Full text] [doi: 10.2196/40035] [Medline: 36322788]
22. HDR UK Innovation Gateway. URL: https://www.healthdatagateway.org/ [accessed 2021-07-28]
23. Cox S, Macdonald C, Lea D, Adejumo S, Panagi V, Tarr S, et al. HDRUK/CaRROT-Mapper: 2.0.1. Zenodo. 2022. URL: https://doi.org/10.5281/zenodo.6587677 [accessed 2022-05-25]
24. Macdonald C, Panagi V, Tarr S, Santos R, Schlessinger D, Mumtaz S. HDRUK/CaRROT-CDM: 0.6.0. Zenodo. 2022. URL: https://doi.org/10.5281/zenodo.6593954 [accessed 2022-05-25]
25. Cox S, Macdonald C, Mumtaz S, Panagi V, Tarr S, Quinlan P. HDRUK/CaRROT-Docs. 2022. URL: https://carrot4omop.ac.uk/ [accessed 2022-03-10]
26. WhiteRabbit. URL: http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html [accessed 2021-07-23]
27. Mulholland RH, Vasileiou E, Simpson CR, Robertson C, Ritchie LD, Agrawal U, et al. Cohort profile: early pandemic evaluation and enhanced surveillance of COVID-19 (EAVE II) database. Int J Epidemiol 2021 Aug 30;50(4):1064-1074 [FREE Full text] [doi: 10.1093/ije/dyab028] [Medline: 34089614]
28. Milligan G, Johnston J, Horban S, Masood E, Cox S, Urwin E, et al. COVID-curated and OpeN aNalysis and rEsearCh plaTform (CO-CONNECT) implementation guide. Zenodo. 2022. URL: https://doi.org/10.5281/zenodo.7150536 [accessed 2022-10-05]
29. CO-CONNECT - overview. CO-CONNECT YouTube page. 2022 Oct 3. URL: https://www.youtube.com/watch?v=dqZtWi6VGGo [accessed 2024-10-10]
30. CO-CONNECT - finding data. CO-CONNECT YouTube page. 2022 Oct 3. URL: https://www.youtube.com/watch?v=HAI3NhJsCKE [accessed 2024-10-10]
31. CO-CONNECT- accessing and analysing data. CO-CONNECT YouTube page. 2022 Oct 3. URL: https://www.youtube.com/watch?v=oc25k8WW440 [accessed 2024-10-10]
32. CO-CONNECT data pipeline demo videos. CO-CONNECT YouTube page. URL: https://www.youtube.com/playlist?list=PLHXTHXI_vn3bqTDcr3hZ2DgjF83mWFHLp [accessed 2022-11-14]
33. Health Data Research United Kingdom (HDR UK). HDR UK. URL: https://www.hdruk.ac.uk/ [accessed 2023-03-21]
34. Five Safe Data Access Request application form. Health Data Research UK 2022 Feb 02 [FREE Full text] [doi: 10.5281/zenodo.5946891]
35. Urwin E, Martin J, Sebire N, Harris A, Johnson J, Masood E, et al. A SARS-CoV-2 minimum data standard to support national serology reporting. Ann Clin Biochem 2024 Jun 17:45632241261274 [FREE Full text] [doi: 10.1177/00045632241261274] [Medline: 38806176]
36. Mooney K. Information is power: its loss can be devastating. Central Bylines. URL: https://centralbylines.co.uk/business/technology/information-is-power-its-loss-can-be-devastating/ [accessed 2024-10-05]
37. ALSPAC access policy. University of Bristol. 2024. URL: https://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf [accessed 2024-05-29]
38. Bailey A, Coleman G, Harbinson A, Khodiyar V, Mill N, Morris C, et al. Five safe data access request application form. Zenodo. 2022. URL: https://doi.org/10.5281/zenodo.5946891 [accessed 2023-04-27]
39. DARE UK. URL: https://dareuk.org.uk/ [accessed 2024-05-25]
40. Alleviate—our hub for APDP pain research data. HDR UK. URL: https://www.hdruk.ac.uk/helping-with-health-data/our-hubs-across-the-uk/alleviate/ [accessed 2021-07-29]
41. Giles DT, Soiland-Reyes S, Couldridge J, Wheater S, Thomson B, Beggs J, et al. TRE-FX: delivering a federated network of trusted research environments to enable safe data analytics. Zenodo. 2023. URL: https://doi.org/10.5281/zenodo.10055354 [accessed 2024-10-05]
42. Orton C, Thompson PS, Lee A, Whittle J, Clark L, Healy J, et al. TELEPORT: connecting researchers to big data at light speed. Zenodo. 2023. URL: https://doi.org/10.5281/zenodo.10055358 [accessed 2024-10-05]

Abbreviations

ALSPAC: Avon Longitudinal Study of Parents and Children
CDM: common data model
CO-CONNECT: COVID-19-Curated and Open Analysis and Research Platform
COPI: Control of Patient Information
NIHR: National Institute for Health and Care Research

XSL•FO
RenderX

NHS: National Health Service
OMOP: Observational Medical Outcomes Partnership
PPIE: patient and public involvement and engagement
TDCC: Tissue Directory and Coordination Centre
TRE: Trusted Research Environment
VM: virtual machine

Edited by Y Zhuang; submitted 27.06.23; peer-reviewed by A Ahonen-Bishopp, M Behzadifar; comments to author 29.04.24; revised version received 12.06.24; accepted 21.06.24; published 20.11.24.

Please cite as:
Jefferson E, Milligan G, Johnston J, Mumtaz S, Cole C, Best J, Giles TC, Cox S, Masood E, Horban S, Urwin E, Beggs J, Chuter A, Reilly G, Morris A, Seymour D, Hopkins S, Sheikh A, Quinlan P
The Challenges and Lessons Learned Building a New UK Infrastructure for Finding and Accessing Population-Wide COVID-19 Data for Research and Public Health Analysis: The CO-CONNECT Project
J Med Internet Res 2024;26:e50235
URL: https://www.jmir.org/2024/1/e50235
doi:10.2196/50235
PMID:

©Emily Jefferson, Gordon Milligan, Jenny Johnston, Shahzad Mumtaz, Christian Cole, Joseph Best, Thomas Charles Giles, Samuel Cox, Erum Masood, Scott Horban, Esmond Urwin, Jillian Beggs, Antony Chuter, Gerry Reilly, Andrew Morris, David Seymour, Susan Hopkins, Aziz Sheikh, Philip Quinlan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Exploring Participants' Experiences of Digital Health Interventions With Qualitative Methods: Guidance for Researchers

Kristin Harrison Ginsberg[1*], MA; Katie Babbott[1*], PhD; Anna Serlachius[1*], PhD

Department of Psychological Medicine, University of Auckland, Auckland, New Zealand

[*]all authors contributed equally

Corresponding Author:
Anna Serlachius, PhD
Department of Psychological Medicine
University of Auckland
22-30 Park Avenue
Grafton
Auckland, 1023
New Zealand
Phone: 64 9 923 3073
Email: a.serlachius@auckland.ac.nz

Abstract

Digital health interventions have gained prominence in recent years, offering innovative solutions to improve health care delivery and patient outcomes. Researchers are increasingly using qualitative approaches to explore patient experiences of using digital health interventions. Yet, the qualitative methods used in these studies can vary widely, and some methods are frequently misapplied. We highlight the methods we find most fit for purpose to explore user experiences of digital tools and propose 5 questions for researchers to use to help them select a qualitative method that best suits their research aims.

(J Med Internet Res 2024;26:e62761) doi:10.2196/62761

KEYWORDS

qualitative methods; content analysis; thematic analysis; digital health evaluation; user engagement; user experience; digital health intervention; innovation; patient experience; health care; researcher; technology; mobile health; mHealth; telemedicine; digital health; behavior change; usability; tutorial; research methods; qualitative research; study design

Introduction

Digital health interventions encompass a wide range of technologies, such as mobile apps, websites, wearable devices, and telemedicine platforms. Their use in behavioral change and mental health interventions are common; such interventions are increasingly used as both an adjunct to face-to-face care and as standalone interventions [1-4]. Although digital health interventions have ostensible benefits in terms of scalability and potential to improve equity of access, the high dropout rates of digital health interventions, which can reach up to 80% [5,6], raise questions of acceptability and usability. Understanding participants' experiences in these interventions is therefore crucial for developing effective digital tools, improving and tailoring existing digital tools, and optimizing health outcomes.

Qualitative research exploring user experience of digital health tools has surged alongside the exponential increase in digital health interventions. Researchers frequently use qualitative approaches to explore engagement, usability, and uptake of digital health interventions. In this type of "applied research," researchers often start with predefined research questions, such as how well an intervention works and under what circumstances [7]. Yet, despite these focused research questions, the qualitative methods (ie, the qualitative approaches used) and methodology (ie, the theoretical rationale and perspective that guide the research) used in these studies can vary widely and be misapplied. For example, a systematic review of 16 studies that used qualitative methods to assess user experiences of digital interventions for pediatric patients [8] found an eclectic range of data analysis methods used, including thematic analysis (37%), content analysis (31%), hermeneutic research analysis (6%), and the generically described "deductive" analysis (6%). Nearly 20% of the articles did not describe their qualitative analysis methodology (ie, their theoretically informed approach). These vague, incomplete, or absent descriptions of qualitative methodology are common across the health intervention literature [9].

XSL•FO
RenderX

As health psychology researchers who have conducted a range of studies exploring patient experiences using digital health interventions [10,11], we understand the challenge of selecting a qualitative approach for this type of research, as there are no existing guidelines to inform the selection of the right qualitative method (ie, the approach that is used) and/or methodology (ie, the overarching philosophical framework or lens). Although there exist several guidelines for improving the reporting of qualitative research [12-15], these generally do not offer guidance on selecting appropriate qualitative approaches. Furthermore, qualitative approaches are also more flexible and interpretative than quantitative methods, which can add to selection confusion. Therefore, in this article, we summarize the most common qualitative approaches used in digital health research when exploring user engagement and make recommendations based on our experiences. We also provide 5 questions and a decision tree that digital health intervention researchers can use to select a qualitative approach that best meets their research goals.

First Things First: Grounded in Theory, or Not?

As qualitative research has grown in popularity in behavioral research, so too has confusion over the range of approaches available and the theoretical or methodological frameworks involved. For example, frequent misapplication of thematic analysis, one of the most often used qualitative methods, has led Braun and Clarke [16], authors of the seminal paper on the method, to publish 4 recent papers clarifying its methodology and "flexible" theoretical approach [17-20].

Starting first with theory and its underlying epistemology (ie, underlying philosophy) can help you choose the qualitative approach that best fits your research question. Qualitative researchers exploring participant experiences in health-related research often take either a constructionist or realist approach. A constructionist approach, also referred to as interpretivist, critical, or "artfully interpretive" [21], assumes that reality is socially constructed and shaped through interactions, language, and meaning-making processes. It emphasizes that individuals and groups actively create their own realities through their interpretations and interactions with the world. Social constructionist qualitative approaches in health- and mental health–related research often analyze patterns in language, discourse, and narratives to explore the underlying meaning and the social constructs of reality [22].

In contrast, a realist approach, or a "scientifically descriptive" [21] approach, posits that there is a degree of objectivity that exists independently of human perception. It suggests that social phenomena have inherent structures and properties that exist regardless of human interpretation. Realist approaches are often used in mixed methods research and tend to analyze qualitative data with a focus on identifying a "consistency of meaning," often through triangulation and the use of several co-coders [23].

Researchers also need to consider whether they are taking a deductive (top-down), inductive (data-driven), or abductive (combination of both) approach in their analysis and whether they will focus on semantic (explicit) or latent (implicit) meanings in the data. It is increasingly acknowledged that the coding and analysis process is rarely completely inductive or deductive [24] and is usually a combination of both [25,26]. For example, even if you choose a mainly deductive approach, it is not uncommon that unexpected and largely inductive (data-driven) codes may also become apparent during the qualitative coding process, even if you are coding with a framework in mind or are approaching the data with predetermined deductive codes. This "abductive" approach also allows for a more nuanced and contextual story to be told. For example, even if you are coding participants' responses using a digital usability framework, you can also code inductively when participants raise important concepts, experiences, and thoughts that relate to the overall research question and aims.

Untangling Qualitative Methodologies

Overview

Some qualitative approaches are inherently linked to their broader methodological frameworks (eg, narrative analysis, interpretative phenomenological analysis, and grounded theory) while others are more flexible in terms of belonging to a certain methodological framework (eg, thematic analysis [16] and content analysis [27]). Thematic analysis, qualitative content analysis, grounded theory, and interpretative phenomenological analysis are some of the many forms of pattern-seeking qualitative approaches. In our opinion, grounded theory and interpretative phenomenological analysis, which are highly interpretive and explorative, are less likely to fit the needs of most digital health research exploring participant usability of and engagement with digital tools, and we are not going to discuss them here. (Good guidance on these methods can be found via Strauss and Corbin [28] and McLeod [29].) Instead, we've highlighted the 2 most commonly used approaches in digital health research exploring participant experience and user engagement.

Qualitative Content Analysis

Qualitative content analysis, including conventional, directed, and summative approaches [30], uses systematic methods to analyze patterns in text and explore meaning. Sometimes considered an "intermediary" approach between qualitative and quantitative methods [31], content analysis, particularly some types, such as summative, use methods that borrow from quantitative research, such as counting the frequency of words or phrases [30]. Researchers using content analysis may take a deductive, inductive, or abductive (combined) approach and often use it deductively given its utility for exploring predefined categories and/or incorporating an existing behavioral or theoretical framework [27,29]. Conventional content analysis can also be used more interpretively, and shares some similarities with certain types of thematic analysis (such as the framework method, discussed below) [32].

We find qualitative content analysis particularly useful in digital health intervention evaluations that build on existing frameworks or previous research (as most of our work does). We have also found directed content analysis to be an efficient and effective

XSL•FO
RenderX

method to analyze qualitative data within mixed methods randomized controlled trials of digital health interventions (such as in Brenton-Peters et al [33] and Serlachius et al [34]).

Thematic Analysis

Similar in some aspects to content analysis, thematic analysis is a "family" of related methods [19] that involve labeling ("coding") data and then organizing them into themes (patterns of meaning). Reflexive thematic analysis, which directly addresses the researcher's role and process in the analysis, is Braun and Clarke's [16] updated approach and seeks in part to differentiate itself from older styles of thematic analysis, such as codebook and framework analysis, which take a more structured, and often combined inductive/deductive, approach.

Due in part to Braun and Clarke's [16] detailed, step-by-step description of thematic analysis published in 2006, this approach has become widely used. They have written at length on how this approach is frequently misunderstood and misapplied, as demonstrated by a recent study where they reviewed 20 health-related studies that used thematic analysis and found the most common problem was in the creation of themes [20]. In reflexive thematic analysis, a theme should highlight a broader meaning across the dataset, telling an interpretive story. However, many studies confuse themes with "topic summaries," such as "helpfulness of the intervention" or "ease of use." Braun and Clarke [20] note that if the theme could have been created before conducting the data analysis (ie, it could be mapped directly to an interview question), then it is a topic summary.

We feel that if your research question and aims align best with topic summaries and deductive analysis, qualitative content analysis or codebook or framework thematic analysis may be a better fit than reflexive thematic analysis or other more interpretative approaches.

Selecting a Qualitative Approach: 5 Questions to Consider

To use qualitative approaches more effectively in digital health user engagement research, we believe there needs to be more coherence and consistency in choosing qualitative methods. Therefore, we suggest researchers use the following 5 questions to help select the qualitative approach most appropriate for their research aims (Figure 1 shows a decision tree.)

Figure 1. Questions to help select a qualitative method in digital health research evaluation.

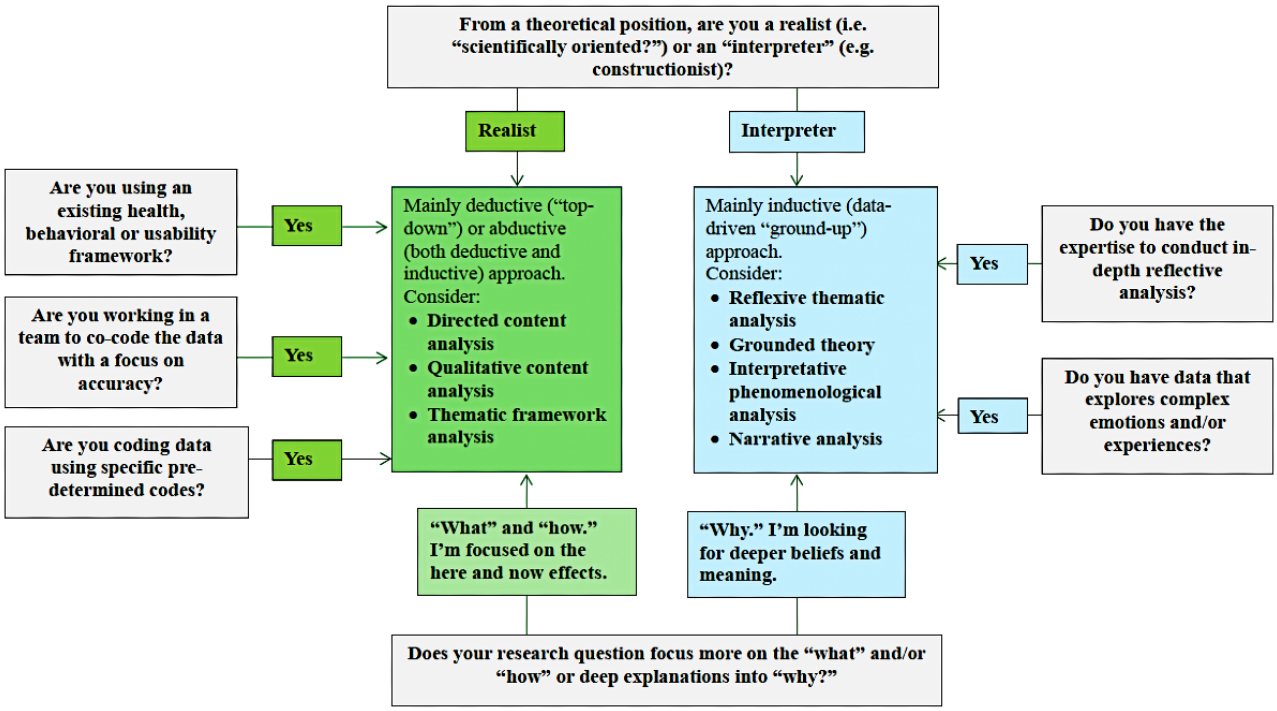

Question 1: What Is Your Theoretical/Methodological Position: Are You a Realist or an Interpreter?

Most digital health intervention evaluation studies have pragmatic goals grounded in a realist (ie, "scientifically oriented") view, building on previous research findings, for example, as part of mixed methods research. If your study fits within this orientation, strongly consider using a method like content analysis or thematic framework analysis, which can flexibly accommodate existing frameworks and build on previous quantitative findings.

While more inductively aligned methods of thematic analysis or other more interpretative approaches such as narrative analysis or grounded theory can be excellent approaches at the development stage of a digital health intervention, when there is more emphasis on exploring in-depth patient experiences, they are frequently less aligned with most research aims around digital health intervention evaluation and exploring questions regarding user experience and user engagement through a realist epistemology. However, there are exceptions; for example, a study by Knox et al [35] explored stakeholder views on virtual pulmonary rehabilitation using a critical epistemology, and a paper by Bleyel et al [36] explored patient perceptions of mental

XSL•FO
RenderX

health care through video consultations from a critical realist perspective [36]. These examples demonstrate the importance of articulating your research aims, methods, and methodological perspective to justify your chosen approach, which is not always easily achievable within the tight word-count limits of medical journals.

Also carefully consider whether you want to take a deductive, inductive, or abductive approach in your analysis. While you can use any of these approaches across the different types of content and thematic analysis methods [27,37], you should select an analysis method that works within your research aims and state this orientation in your reported methodology.

Question 2: Are You Using an Existing Health or Behavioral Change Framework?

Many digital health interventions are built on existing intervention or behavior change frameworks, such as the capability, opportunity, motivation-behavior (COM-B) model [38] and the reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) framework [39]. If you want to use an existing framework in your qualitative analysis, as many digital intervention studies do, choose an approach designed for this, such as directed content analysis or thematic framework analysis. Do not choose a highly interpretive approach, such as reflexive thematic analysis or interpretative phenomenological analysis, if you intend to map your qualitative findings onto an existing framework. A good example by Szinay and colleagues [40] used the framework method to explore engagement with well-being apps informed by the COM-B model. For further details regarding the framework method, refer to Gale and colleagues [32], who provide a comprehensive outline of the 7 stages of analysis using this method in health research.

Question 3: What Type of Research Question Are You Trying to Answer?

Qualitative research is well suited to answer “how” and “why” questions, such as why individuals did (or did not) use a digital health intervention and how they went about it. But different research questions may be better suited to one type of qualitative analysis than another.

For example, let us suppose your digital health intervention for weight loss had a high drop-out rate, and you want to use qualitative methods to explore reasons why. The most suitable qualitative approach depends on your underlying rationale and research strategy (including your methodological/theoretical position) and what you plan to do in response to your findings (ie, changes or iterations to the intervention or exploring a different approach altogether). If you want to explore topics such as participants’ beliefs about weight loss and/or perceived barriers to weight loss, narrative analysis or another more inductive approach may be more appropriate. However, if, as is more common in digital health intervention evaluation research, you have specific questions related to participants’ opinions about the design, content, and activities within the intervention itself that caused them to stop using it, then a more focused and deductive or abductive qualitative approach (ie, content analysis) may be more useful.

Question 4: How Are You Collecting Data?

According to a 2019 systematic review, frequently used qualitative data collection tools in digital health intervention evaluation and testing and exploring users’ perspectives include “think-aloud” protocols, focus groups, and interviews [41]. We often also use open-ended online survey questions in our mixed methods evaluation studies. It is important to consider the type of data you are collecting alongside your analysis plan.

Stemming from cognitive psychology, the think-aloud method asks participants to share what they are thinking while performing a task [42], which can lead to a focused data set. Because of this, the think-aloud method is popular in usability research and has been found to successfully generate intervention improvements [43]. Think-aloud data are most commonly analyzed using a deductive approach and with methods similar to content analysis, such as counting positive or negative sentiment [44], but these data can also be coded and organized into topics or themes [12].

Focus groups and interviews can generate large datasets that may cover a broad range of topics and experiences. While exploratory questions may be useful in digital health intervention development research, a more focused approach is usually necessary to meet the goals of evaluation research or exploring whether a digital tool met participants’ needs. Therefore, it is critical to develop an interview guide that considers your research questions, theoretical framework, and desired result type (ie, themes or frequencies). Otherwise, you may end up with a wide-ranging and meandering dataset that does not answer your research questions and leads to what Braun and Clarke [20] call a “mish-mash” of ineffective qualitative analysis.

Question 5: How Much Time Do You Have?

Digital health interventions often require rapid evolution to stay relevant, and researchers may have limited time, resources, and funding [7]. Therefore, it is important to consider how quickly you need to generate your results, as well as the resources needed to complete the analysis.

In terms of the approaches discussed here, reflexive thematic analysis, interpretive phenomenological analysis, grounded theory, and narrative analysis often require significant time to conduct, with multiple rounds of detailed, in-depth coding and thematic review and ongoing exploration regarding reflexivity and how it influences the process and outcomes of the analysis. Content analysis or framework thematic analysis can be done in a more efficient and timely manner, with codebooks often developed in advance of analysis and the frequent use of co-coders with a focus on accuracy.

Increasingly, so-called rapid approaches have been developed and compared against more traditional qualitative approaches due to this need to reduce time and improve efficiency in health care research [45,46]. For example, Holdsworth et al [47] used a rapid form of framework analysis to summarize evaluation data from intensive care unit site visits while still on site. They found this approach delivered significant savings in analysis and transcription costs.

XSL•FO
RenderX

Conclusion

After more than a decade of conducting qualitative research exploring patient experiences of and user engagement with digital health tools, we too have grappled with choosing appropriate qualitative methods. In this article, we draw on our experience, which has been largely in health psychology and digital health–related research, but acknowledge that similar qualitative design challenges exist in health informatics, health services research, and the broader usability research literature [14,15]. Ultimately, we believe that many of these ongoing interdisciplinary challenges and discussions enhance overall understanding of the benefits of qualitative and mixed methods research and lead to improvements not just in the conducting and reporting of qualitative research but also in the ongoing development and clarification of the methods themselves. This paper did not have the scope to explore questions regarding what constitutes "good" or rigorous qualitative research in digital health, but there are several articles providing useful summaries of key criteria in digital health research, health services research, and digital health assessment [48,49].

Digital health interventions offer innovative solutions to improve health care delivery and patient outcomes. Researchers now frequently use qualitative approaches to explore engagement, usability, and uptake of digital health interventions. However, some of the most popular qualitative approaches are frequently misapplied in this type of research and may not be the most appropriate for this type of research. Therefore, we suggest researchers consider their research aims, theoretical orientation, and the time and resources available before selecting a qualitative approach to use when exploring participants' experiences of digital tools.

Conflicts of Interest

None declared.

References

1. Li J, Theng Y, Foo S. Game-based digital interventions for depression therapy: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2014 Aug;17(8):519-527. [doi: 10.1089/cyber.2013.0481] [Medline: 24810933]
2. McLean G, Band R, Saunderson K, Hanlon P, Murray E, Little P, DIPSS co-investigators. Digital interventions to promote self-management in adults with hypertension systematic review and meta-analysis. J Hypertens 2016 Apr;34(4):600-612 [FREE Full text] [doi: 10.1097/HJH.0000000000000859] [Medline: 26845284]
3. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022 May 12;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
4. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017 Dec;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
5. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
6. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 2019 Sep 04;21(9):e13617 [FREE Full text] [doi: 10.2196/13617] [Medline: 31486409]
7. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
8. Cheng L, Liu F, Mao X, Peng W, Wang Y, Huang H, et al. The pediatric cancer survivors' user experiences with digital health interventions: a systematic review of qualitative data. Cancer Nurs 2022;45(1):E68-E82. [doi: 10.1097/NCC.0000000000000885] [Medline: 34870939]
9. O'Cathain A, Hoddinott P, Lewin S, Thomas KJ, Young B, Adamson J, et al. Maximising the impact of qualitative research in feasibility studies for randomised controlled trials: guidance for researchers. Pilot Feasibility Stud 2015;1(1):32 [FREE Full text] [doi: 10.1186/s40814-015-0026-y] [Medline: 27965810]
10. Garner K, Thabrew H, Lim D, Hofman P, Jefferies C, Serlachius A. Exploring the Usability and acceptability of a well-being app for adolescents living with type 1 diabetes: qualitative study. JMIR Pediatr Parent 2023 Dec 22;6:e52364 [FREE Full text] [doi: 10.2196/52364] [Medline: 38153804]
11. Wallace-Boyd K, Boggiss AL, Ellett S, Booth R, Slykerman R, Serlachius AS. ACT2COPE: A pilot randomised trial of a brief online acceptance and commitment therapy intervention for people living with chronic health conditions during the COVID-19 pandemic. Cogent Psychology 2023 May 11;10(1):2208916. [doi: 10.1080/23311908.2023.2208916]
12. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
13. Ancker J, Benda N, Reddy M, Unertl K, Veinot T. Guidance for publishing qualitative research in informatics. J Am Med Inform Assoc 2021 Nov 25;28(12):2743-2748 [FREE Full text] [doi: 10.1093/jamia/ocab195] [Medline: 34537840]

XSL•FO
RenderX

14. Mann DM, Chokshi SK, Kushniruk A. Bridging the gap between academic research and pragmatic needs in usability: a hybrid approach to usability evaluation of health care information systems. JMIR Hum Factors 2018 Nov 28;5(4):e10721 [FREE Full text] [doi: 10.2196/10721] [Medline: 30487119]
15. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess 1998;2(16):iii-ix, 1 [FREE Full text] [Medline: 9919458]
16. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
17. Braun V, Clarke V, Hayfield N, Terry G. Answers to frequently asked questions about thematic analysis. University of Auckland. URL: https://cdn.auckland.ac.nz/assets/psych/about/our-research/documents/Answers%20to%20frequently%20asked%20questions%20about%20thematic%20analysis%20April%202019.pdf [accessed 2024-11-21]
18. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
19. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern‐based qualitative analytic approaches. Couns Psychother Res 2020 Oct 18;21(1):37-47. [doi: 10.1002/capr.12360]
20. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
21. Finlay L. Thematic analysis: the 'good', the 'bad' and the 'ugly'. Eur J Qual Res Psychother 2021;11:103-116 [FREE Full text] [doi: 10.4337/9781847208552.00018]
22. Harper D, Thompson AR, editors. Qualitative Research Methods in Mental Health and Psychotherapy: A Guide for Students and Practitioners. 1st ed. New York, NY: Wiley; 2011.
23. Madill A, Jordan A, Shirley C. Objectivity and reliability in qualitative analysis: realist, contextualist and radical constructionist epistemologies. Br J Psychol 2000 Feb;91 (Pt 1):1-20. [doi: 10.1348/000712600161646] [Medline: 10717768]
24. Bingham AJ. From data management to actionable findings: a five-phase process of qualitative data analysis. Int J Qual Methods 2023 Aug 10;22:3. [doi: 10.1177/16094069231183620]
25. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 2006 Mar 01;5(1):80-92. [doi: 10.1177/160940690600500107]
26. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2022 Sep 20;17(3):308-326. [doi: 10.1177/15586898221126816]
27. Graneheim UH, Lindgren B, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017 Sep;56:29-34. [doi: 10.1016/j.nedt.2017.06.002] [Medline: 28651100]
28. Strauss A, Corbin JC. Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Thousand Oaks, CA: Sage Publications; 1990.
29. McLeod J. Qualitative Research in Counselling and Psychotherapy. Thousand Oaks, CA: Sage Publications; 2012.
30. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
31. Mayring PAE. Qualitative content analysis. In: International Encyclopedia of Education (Fourth Edition). Amsterdam, Netherlands: Elsevier; 2023:314-322.
32. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
33. Brenton-Peters JM, Consedine NS, Cavadino A, Roy R, Ginsberg KH, Serlachius A. Finding kindness: a randomized controlled trial of an online self-compassion intervention for weight management (SC4WM). Br J Health Psychol 2024 Feb;29(1):37-58. [doi: 10.1111/bjhp.12686] [Medline: 37544883]
34. Serlachius A, Schache K, Kieser A, Arroll B, Petrie K, Dalbeth N. Association between user engagement of a mobile health app for gout and improvements in self-care behaviors: randomized controlled trial. JMIR Mhealth Uhealth 2019 Aug 13;7(8):e15021 [FREE Full text] [doi: 10.2196/15021] [Medline: 31411147]
35. Knox L, Gemine R, Dunning M, Lewis K. Reflexive thematic analysis exploring stakeholder experiences of virtual pulmonary rehabilitation (VIPAR). BMJ Open Respir Res 2021 Jul;8(1):e000800 [FREE Full text] [doi: 10.1136/bmjresp-2020-000800] [Medline: 34301711]
36. Bleyel C, Hoffmann M, Wensing M, Hartmann M, Friederich H, Haun MW. Patients' perspective on mental health specialist video consultations in primary care: qualitative preimplementation study of anticipated benefits and barriers. J Med Internet Res 2020 Apr 20;22(4):e17330 [FREE Full text] [doi: 10.2196/17330] [Medline: 32310139]
37. Roseveare C. Thematic Analysis: A Practical Guide, by Virginia Braun and Victoria Clarke. Can J Program Eval 2023 Jun 01;38(1):143-145. [doi: 10.3138/cjpe.76737]
38. Keyworth C, Epton T, Goldthorpe J, Calam R, Armitage CJ. Acceptability, reliability, and validity of a brief measure of capabilities, opportunities, and motivations ("COM-B"). Br J Health Psychol 2020 Sep;25(3):474-501 [FREE Full text] [doi: 10.1111/bjhp.12417] [Medline: 32314500]

XSL•FO
RenderX

39. Forman J, Heisler M, Damschroder LJ, Kaselitz E, Kerr EA. Development and application of the RE-AIM QuEST mixed methods framework for program evaluation. Prev Med Rep 2017 Jun;6:322-328 [FREE Full text] [doi: 10.1016/j.pmedr.2017.04.002] [Medline: 28451518]
40. Szinay D, Perski O, Jones A, Chadborn T, Brown J, Naughton F. Perceptions of factors influencing engagement with health and well-being apps in the United Kingdom: qualitative interview study. JMIR Mhealth Uhealth 2021 Dec 16;9(12):e29098 [FREE Full text] [doi: 10.2196/29098] [Medline: 34927597]
41. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: A scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
42. Ericsson KA, Simon HA. Verbal reports as data. Psychol Rev 1980 May;87(3):215-251. [doi: 10.1037/0033-295X.87.3.215]
43. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
44. Yardley L, Bradbury K, Morrison L. Using qualitative research for intervention development and evaluation. In: Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design (2nd Ed). Washington, DC: American Psychological Association; 2021.
45. Taylor B, Henshall C, Kenyon S, Litchfield I, Greenfield S. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. BMJ Open 2018 Oct 08;8(10):e019993 [FREE Full text] [doi: 10.1136/bmjopen-2017-019993] [Medline: 30297341]
46. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
47. Holdsworth LM, Safaeinili N, Winget M, Lorenz KA, Lough M, Asch S, et al. Adapting rapid assessment procedures for implementation research using a team-based approach to analysis: a case example of patient quality and safety interventions in the ICU. Implement Sci 2020 Feb 22;15(1):12 [FREE Full text] [doi: 10.1186/s13012-020-0972-5] [Medline: 32087724]
48. Mays N, Pope C. Qualitative research in health care. Assessing quality in qualitative research. BMJ 2000 Jan 01;320(7226):50-52. [doi: 10.1136/bmj.320.7226.50] [Medline: 10617534]
49. Tracy SJ. Qualitative quality: eight “big-tent” criteria for excellent qualitative research. Qual Inq 2010 Oct 01;16(10):837-851. [doi: 10.1177/1077800410383121]

Abbreviations

COM-B: capability, opportunity, motivation-behavior
Re-Aim: reach, effectiveness, adoption, implementation, and maintenance

Edited by T de Azevedo Cardoso; submitted 30.05.24; peer-reviewed by D Singh, Y Sha, L Weinert; comments to author 20.09.24; revised version received 03.10.24; accepted 28.10.24; published 28.11.24.

Please cite as:
Harrison Ginsberg K, Babbott K, Serlachius A
Exploring Participants’ Experiences of Digital Health Interventions With Qualitative Methods: Guidance for Researchers
J Med Internet Res 2024;26:e62761
URL: https://www.jmir.org/2024/1/e62761
doi:10.2196/62761
PMID:

©Kristin Harrison Ginsberg, Katie Babbott, Anna Serlachius. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A Cross-Disciplinary Analysis of the Complexities of Scaling Up eHealth Innovation

Sanne Allers[1], MSc; Chiara Carboni[1], MSc; Frank Eijkenaar[1], PhD; Rik Wehrens[1], PhD

Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, Netherlands

Corresponding Author:
Sanne Allers, MSc
Erasmus School of Health Policy & Management
Erasmus University Rotterdam
Burgemeester van Oudlaan 50
Rotterdam, 3062PA
Netherlands
Phone: 31 104081213
Email: allers@eshpm.eur.nl

Abstract

Innovative eHealth technologies are becoming increasingly common worldwide, with researchers and policy makers advocating their scale-up within and across health care systems. However, examples of successful scale-up remain extremely rare. Although this issue is widely acknowledged, there is still only a limited understanding of why scaling up eHealth technologies is so challenging. This article aims to contribute to a better understanding of the complexities innovators encounter when attempting to scale up eHealth technologies and their strategies for addressing these complexities. We draw on different theoretical perspectives as well as the findings of an interview-based case study of a prominent remote patient monitoring (RPM) innovation in the Netherlands. Specifically, we create a cross-disciplinary theoretical framework bringing together 3 perspectives on scale-up: a structural perspective (focusing on structural barriers and facilitators), an ecological perspective (focusing on local complexities), and a critical perspective (focusing on mutual adaptation between innovation and setting). We then mobilize these perspectives to analyze how various stakeholders (n=14) experienced efforts to scale up RPM technology. We provide 2 key insights: (1) the complexities and strategies associated with local eHealth scale-up are disconnected from those that actors encounter at a broader level scale-up, and this translates into a simultaneous need for stability and malleability, which catches stakeholders in an impasse, and (2) pre-existing circumstances and associated path dependencies shape the complexities of the local context and facilitate or constrain opportunities for the scale-up of eHealth innovation. The 3 theoretical perspectives used in this article, with their diverging assumptions about innovation scale-up, should be viewed as complementary and highlight different aspects of the complexities perceived as playing an important role. Using these perspectives, we conclude that the level at which scale-up is envisaged and the pre-existing local circumstances (2 factors whose importance is often neglected) contribute to an impasse in the scale-up of eHealth innovation at the broader level of scale.

(J Med Internet Res 2024;26:e58007) doi:10.2196/58007

KEYWORDS

innovation; eHealth; remote patient monitoring; scale-up; cross-disciplinary; qualitative case study; health care systems; adaptation; complexity; health care; framework; ecological perspective; barriers and facilitators

Introduction

Innovative eHealth technologies, defined in this paper as tools that support the organization and delivery of health services and information using the internet and related technologies [1], are becoming increasingly common in health care systems worldwide. These technologies have been discussed in the literature with careful enthusiasm, with studies often weighing challenges related to “privacy, liability, and costs” [2] against their potential to support patient-centered care, remote patient monitoring (RPM), and prevention [3-6]. In addition, eHealth technologies have become increasingly central in policy debates, with policy makers consistently articulating high expectations around eHealth’s role in health care’s future sustainability.

The literature often singles out the fragmentation of the eHealth landscape as a potential hindrance [7,8]. Such fragmentation is found to result in various problems, including nondissemination of valuable innovations [9]; inequalities resulting from a failure to reach patients who have the greatest needs [10] or who reside in specific areas [11]; and generally unsustainable

XSL•FO
RenderX

implementation, given that eHealth providers need to achieve a considerable level of coverage to become (economically) viable [12].

Scale-up is often proposed as a remedy in this context [13]. For instance, policy makers have advocated replicating proven eHealth technologies within and across health care systems. In its global strategy 2020-2025, the World Health Organization (WHO) recommends focusing on nationwide scaling of eHealth technologies, proposing that principles such as *scalability* and *replicability* should be at the heart of current efforts around eHealth development and implementation [14]. Many national governments have also attempted to support the scale-up of eHealth technologies, emphasizing that stakeholders should share and adopt best practices rather than reinvent the wheel [15,16].

Despite this dominant rhetoric, examples of successful eHealth technology scale-up are few [17,18]. Widespread adoption is plagued by the "diffusion chasm," with a gap opening up between initial invention and successful market penetration [19]. A recent literature review, for example, found that none of the eHealth technologies implemented in the United Kingdom managed to reach organization-wide or large-scale adoption [20]. Similarly, a publicly funded program in the Netherlands aimed at introducing eHealth technologies nationwide [21] ultimately fell short of its goal, that is, to counteract fragmentation and encourage further scale-up of local initiatives [22].

Although the poor scale-up of eHealth innovations is widely acknowledged [18], we still have only a limited understanding of why scaling up eHealth technologies is so challenging. So far, research has described the stagnation of eHealth implementation without articulating clear strategies to overcome it [23,24]. These studies focus on the local implementation of pilot eHealth technologies and assume that this process has a clear beginning (ie, the introduction of an innovation in a specific organizational setting) and an end (ie, the innovation being structurally embedded in that specific organizational setting). What came before and what comes after are often left out of the picture. Thus, although researchers acknowledge the necessity of scale-up for sustainable eHealth innovation, they typically end up studying implementation issues as though they were separate from scale-up [25]. This narrow approach does not contribute to our knowledge of what scaling up eHealth (and the complexities associated with it) entails or our knowledge of promising strategies to facilitate scale-up.

This article seeks to address this knowledge gap by answering the question of what complexities are encountered when attempting to scale up an eHealth technology and what strategies are applied by stakeholders to deal with these complexities. We start by describing and combining insights from 3 existing theoretical perspectives on technology scale-up in and beyond health care. We then bring these insights to bear on eHealth by examining a case study of a Dutch eHealth technology that, despite being considered a success at the local level, encounters major (and common) challenges when being scaled up. By mobilizing empirical insights stemming from the case analysis, our discussion contributes to existing theorizations of the complexities of scale-up, thus illuminating other dimensions of these difficulties as stakeholders experience them in practice.

Theoretical Perspectives and Insights From a Case Study

Meaning of Scale-Up

Numerous common definitions of scale-up in the literature fall into 2 categories. The first describes scale-up as *the replication* of an existing innovation in "multiple geographic locations and contexts to maximize the number of people that an innovation reaches" [26]. Conversely, the second describes scale-up as *the gradual adaptation* an existing innovation undergoes as it becomes embedded in more and more dimensions of health care practice (ie, covering more patients, involving more providers, or adding to the steps involved in care provision). Although they differ in how they regard the innovation itself, both definitions consider the core of scale-up to be an expansion of an innovation's *coverage*. Consistent with this, Spicer et al [27] defined scale-up as "an increase in the coverage of health interventions that have been tested in pilot and experimental projects in order to benefit more people." This paper defines scale-up, in the broadest sense, as the steps taken to progressively expand the coverage of an existing innovation.

Theoretical Perspectives on Innovation Scale-Up

In this section, we offer a high-level discussion of a number of theoretical perspectives that various disciplines have developed to think through the complexities associated with innovation scale-up. The 3 cross-disciplinary perspectives we have selected are by no means an exhaustive list of the theoretical approaches to implementation and scale-up proposed in the literature. They do, however, mobilize a diversity of viewpoints and arguments that enable us to illuminate crucial aspects of the complexities of technology scale-up, giving us a generative heuristic framework for our analysis. Selected precisely by virtue of their differing approaches, the 3 perspectives synthesize a variety of middle-range theories stemming from conventional management scholarship, complex adaptive system theory, and critical social science approaches, respectively, but for the sake of brevity, we refer to them as structural, ecological, and critical perspectives. Although these labels might appear arbitrary, we argue that they emphasize crucial aspects of each perspective: a structural perspective's focus on high-level mechanisms and a somewhat immobile social world; an ecological perspective's emphasis on interactions of different (human) components of health systems; and a critical perspective's attempt to politicize scale-up as well as subvert its commonsense conceptualizations.

In what follows, we first illustrate these perspectives' conceptualizations of the complexities around and solutions for enabling scale-up. Subsequently, we combine these theoretical perspectives to analyze a real-world eHealth innovation case.

Structural Perspective

The structural perspective, particularly prominent in economic and management theories, foregrounds the role of structural system barriers in hindering scale-up beyond the local level, with the possibility of scale-up resting on the removal of these

XSL•FO
RenderX

barriers. For instance, Wang et al [28], in their study of telehealth adoption in the United States, found that the medical-legal framework of health care delivery impedes the successful scale-up of telehealth. As they argued, "policymakers must rethink and address the economic incentives and payment of telehealth services, the medical-legal issues surrounding virtual care, and the effects of increased competition across geographic areas and jurisdictions." Similarly, the scoping review by Gijsbers et al [29] concluded that "successful upscaling of telemonitoring requires insight into its critical success factors, especially at an overarching national level. […] A wide program on change management, nationally or regionally coordinated, is key." In summary, the structural perspective proposes that system-wide innovation scale-up depends on overcoming all critical barriers, assuming that it is possible to adjust systems to remove such barriers.

Ecological Perspective

The ecological perspective, rooted in organization studies and health systems research, emphasizes the local interrelated factors that must be considered to scale an innovation locally and replicate it in another setting. The analysis by Greenhalgh and Papoutsi [30] of the different logics that challenge the dissemination of innovations exemplifies this perspective: "Complexity can be hard to square with spread strategies that seek to replicate a 'blueprint' innovation in a standardized way across widely different settings. The plan-do-study-act engine might work for small-scale improvement initiatives, but spreading and scaling up major innovations across a health system requires attention to the underlying logic of complex systems, which is ecological rather than mechanical."

In another study, Greenhalgh et al [31] argued that explaining the complexities of innovation in terms of barriers and facilitators does not sufficiently acknowledge the intricacies of the setting in which it is introduced. Studies that we group under the ecological perspective emphasize, for example, the importance of conducting sensemaking work among stakeholders before replicating an innovation [32]. Others emphasize the need to acknowledge the dynamic relationship between different factors in a local setting [31], or argue that an innovation must fit in with the diverging institutional logics of all relevant stakeholders before they can accept it as part of their practice [33]. In short, the human and technical characteristics of the local setting (eg, belief systems of local providers, technical interoperability of adopted and pre-existing systems, and stakeholders' clashing institutional logics) must be acknowledged and worked on before an innovation can be scaled from another context. An example of a widely used theory applying this perspective is the RE-AIM framework, which fits in with a social-ecological perspective [34]. It conceptualizes the impact of innovative programs as depending on the percentage and characteristics of the people who receive or are affected by the intervention. Eventually, its aim is for innovations "to become a relatively stable, enduring part of the behavioral repertoire of an individual, organization or community." In other words, the ecological perspective postulates that a successful innovation scale-up requires work to adapt the local setting so that it will accommodate the innovation.

Critical Perspective

A more critical perspective on scale-up has been formulated by authors active in the field of science and technology studies (STS). For instance, in their article on the politics of scaling, Pfotenhauer et al [35] described scaling as a current "obsession of innovation discourses and, with it, contemporary social, political and economic life at large." Critical reflections on scaling often build on Anna Tsing's definition of scalability as the ability of a system to "expand without changing" [36] and, according to Hanna and Park, without "rethinking its constitutive elements" [37]. These authors shed a different light on the local uniformity that the ecological perspective posits as necessary for scaling up: adapting local systems generates tensions because it disregards the diverse ways in which people define problems and solutions, priorities, and values. As Pfotenhauer et al [35] argued, any narrative presenting scaling up as a smooth process is suspicious because it probably excludes certain perspectives. Hanna and Park [37] argued that the very idea of scalability as replication entails that the work that sustains innovation needs to be something "interchangeable, abstract, and universal." Thus, in scaling-up discourse, the emphasis ends up being on the standardization of infrastructure, for instance, at the expense of more relational views that stress the inherently more unpredictable and therefore flexible work needed to maintain networks of humans and technologies. In other words, the critical perspective describes the need to examine innovations in their specific context of emergence, assuming that during implementation, both the innovation and the local setting are reshaped in a work-intensive process of mutual adaptation, and that some type of local knowledge or practice is inevitably lost in this process. Unlike the ecological perspective, the critical perspective suggests that replicating a blueprint innovation in a new setting requires work to adapt not only the context but also the innovation itself.

Cross-Disciplinary Framework

Table 1 summarizes the characteristics of the 3 perspectives on innovation scale-up. We have combined the 3 perspectives into a cross-disciplinary framework, with each one making a unique contribution.

The structural perspective advocates the removal of systemic barriers and the strengthening of facilitators, assuming that the changes required to stimulate scale-up can be pinpointed. This perspective contributes to our framework by focusing on the structural facilitators and barriers (eg, regulatory and financial) that emerge in systems and that support or hinder innovation up-scaling.

The ecological perspective teaches us that conditions at the local level are complex and diverse, and we cannot expect to know in advance what needs to be done for an innovation to "work" in a specific setting. Local complexity must be considered before introducing an innovation in a new setting or to new users. This perspective assumes that the innovation itself remains largely unchanged in the process of scaling and that it is crucial to convince users of its utility or value and to create a fitting context. This perspective contributes to our cross-disciplinary framework by highlighting aspects of the local context that are

XSL•FO
RenderX

perceived to influence the past and future evolution of the innovation.

The critical perspective assumes that the process of organizational embedding transforms both the context and the innovation to such an extent that the very possibility of scaling a specific innovation needs to be questioned. This perspective does not suggest that scaling up is altogether impossible, but it does emphasize that the innovation itself changes continuously in local scale-up processes as it potentially loses or gains aspects while moving across localities. This perspective contributes to our framework by focusing on the possibility that the innovation itself must adapt during the scale-up process in response to the context to which it is being scaled.

Table 1. Summary of the 3 theoretical perspectives on innovation scale-up.

Perspective	Complexities of scale-up	Type of solution
Structural	Structural factors support or hinder successful scale-up of the innovation	Remove structural barriers and strengthen structural facilitators
Ecological	Local context defines whether the innovation can be successfully scaled up	Prepare local contexts for the replication of an innovation
Critical	Innovations cannot be uniformly applied to different contexts and must adapt in response to changes in the context	Rethink scaling up beyond blueprint innovations; acknowledge the mutual transformation of local context and innovation as necessary

Narrative on the Scale-Up of an eHealth Innovation

Case Description

The eHealth technology considered here consists of several measuring devices for physiological variables (eg, weight, blood pressure, and heart rate) and a smartphone app to enable RPM in a Dutch university medical center. Patients perform the measurements at home according to a predetermined schedule, after which the data are automatically sent to the care professionals at the hospital. This eHealth application is meant to encourage healthy behavior in patients by showing them their progress and to allow care professionals to monitor their treatment more closely by sharing continuous health measures. In addition, it greatly reduces the need for outpatient visits. The cardiology department of a Dutch university medical center initially implemented this innovation to monitor a specific group of patients and then attempted to extend it to other cardiology patient populations and other departments. Research framed the innovation as a local success. First, it showed that the health outcomes of patients using the innovation did not differ substantially from those of patients following standard face-to-face care trajectories. Second, it showed that patient experience improved, that patients became more involved in their own care, and that care professionals had better insight into patients' health (references omitted to preserve anonymity).

Data Collection

We performed 14 semistructured interviews (average duration of 60 minutes) with professionals involved in the development, implementation, and scaling of the innovation. All respondents provided informed consent, and interviews were audio recorded and transcribed verbatim. The respondents included 2 researchers who developed the innovation, 4 health care professionals (nurses and doctors), 3 IT professionals, 2 project managers (cardiologists), 1 department manager, and 2 liaisons from a MedTech company. Potential respondents were identified through snowballing, starting with the current project lead, and recruited via email. Recruitment stopped when respondents started referring us to individuals we had already interviewed, thus indicating that we had consulted all the main actors involved in the case. Moreover, while different respondents may have been better or less well acquainted with particular aspects of the innovation's trajectory, the similarity between their views and arguments around scale-up indicated that we had reached analytical saturation.

Most of the interviews (12 out of 14) were conducted jointly by the first 2 authors (SA and CC) to ensure that the data collection reflected the cross-disciplinary perspectives at the heart of this study's design. The remaining 2 interviews were conducted by one or the other. One of the interviewers has a background in health care economics/management, and the other has a background in sociology/organization science and studies health care systems. The 2 interviewers jointly developed semistructured topic lists and adapted them to the specific expertise of each respondent. Respondents were asked to reflect on such topics as how they experienced developing or using the innovation, difficulties encountered during its development or implementation, changes in care provision and infrastructure, interaction between stakeholders, and views on the future of the innovation.

Ethical Considerations

Given that the data collected involved humans, we applied for ethical approval. Ethical approval for this study was granted by the internal review board of the Erasmus School of Health Policy & Management, Erasmus University Rotterdam (reference: ETH2122-0324). All respondents provided written informed consent prior to the start of the interview and were explicitly told they had the opportunity to withdraw from the study at any time or, without stating reasons, withdraw their permission for the use of their information in the study. Data were deidentified as much as possible as all respondents are only referred to by their function and the specific innovation project is described with the omission of any specific references. No form of compensation was provided to the respondents, as we required only a short amount of their time and the research involved very little to no risk of potential harm for the respondents.

XSL•FO
RenderX

Data Analysis

We performed data analysis abductively, going back and forth between the insights from theory and the empirical data, and following guidelines from Braun and Clarke on thematic analysis [38,39]. This paper's focus on scale-up emerged organically as part of this abductive process. Indeed, we initially aimed to investigate what made the innovation a "successful" eHealth innovation. The respondents questioned this depiction of the innovation, however, as they did not perceive its development as finished. Scale-up emerged inductively from our interview data as a central theme in actors' attempts to make sense of what they were involved in and spurred the development of our cross-disciplinary theoretical framework. Using this framework, we analyzed the narrative generated by the respondents as they reflected on the past and future evolution of this eHealth technology, including their views on scale-up. Our aim in combining and comparing the different perspectives was to gain a richer understanding of the scale-up of eHealth innovation.

After abductively reassessing the theoretical and empirical data, we identified 6 themes, making use of the qualitative analysis software Atlas.ti 23. The first 3 themes were related to the chronological experiences of the respondents, enveloping the initial development and embedding of the innovation, the gradual adaptation of the innovation within the original setting, and the attempts to replicate the innovation in a new setting. The other 3 themes were related to the issues that respondents experienced from diverging (theoretical) perspectives, either the barriers they experienced from a structural perspective, the complex local factors they experienced from an ecological perspective, or the necessity of mutual adaptation of innovation and context they experienced from a critical perspective. Subsequently, using these themes, we created a narrative piecing together the respondents' perspectives.

We noted that the respondents' narrative distinguished between *local* organizational scale-up (instances where the innovation's scope was expanded to include more dimensions of care practice in the same organization) and *broader* cross-organizational scale-up (replicating the innovation in a new organization). We have therefore structured this section according to these experiences. In Part I, we describe how stakeholders made sense of what happened during the initiation and embedding of the innovation. These early experiences demonstrate the importance of pre-existing circumstances for local scale-up. In Part II, we present respondents' reflections on the local scale-up and the complexities encountered during this process. Finally, in Part III, we discuss respondents' current experiences and future ambitions for scaling up their innovation to other organizations.

Part I: Initiation and Embedding of the Original Innovation

Local Aspects Shaping the Facilitating and Complicating Circumstances for Innovation

The innovation project started around 2015 at the cardiology department of a Dutch university medical center. As its initiators reported, the project was driven by the will to innovate the practice of cardiology by moving part of care provision outside the hospital using RPM. However, as one of the respondents described, "it wasn't like this happened suddenly. [...] The undercurrent was already there" (Project Manager 1). Several respondents referred to this "undercurrent" to describe the facilitating circumstances in the department prior to the project's initiation. We discuss 3 of these pre-existing circumstances below: the maturity of the department's IT infrastructure; the highly standardized care pathway for (some) patients; and the facilitating workflow, culture, and resources.

First, in the early 2000s, this department was one of the first in the country to develop and implement an electronic health record (EHR). The in-house EHR gave the department a tailored flexible infrastructure into which RPM devices and data could be integrated. The department also had an internal team of dedicated IT professionals who supported staff in integrating the hardware and software and were available to continuously adapt the EHR structure and its data visualization. As one respondent stated, "that made it easier, let's say, to add more things to our own electronic patient record. That was in fact a reason [for the innovation's success]" (Department Manager 1).

Second, the RPM project was built on another project that had introduced a care pathway protocol, thereby restructuring care provision for a patient group. As reported by one of the innovation's initiators, this standardized care pathway laid the groundwork for RPM: "the idea [of reducing patients' visits to the clinic] came from the project that had already been running in the department for years" (IT Professional 1). Introducing the care pathway protocol brought up 2 important points for consideration: on the one hand, the numerous physical contact moments between medical professionals and patients, and on the other, the lack of data on these patients between hospital visits. Thus, besides providing clarity on disease progression, and a "very well-defined care track" (Department Manager 1), this project also singled out points for improvement that could be addressed through RPM technologies.

Third, the RPM project's initiators also highlighted how the practice and culture of cardiology are highly technology-, data-, and innovation-driven. The department prided itself on its early adoption of earlier innovative technologies, such as pacemakers, and for being "used to problems with patients with home monitoring devices." In turn, this meant that they "already had a very fast technical back office to help patients with their problems" (Department Manager 1). The professional workflow and culture at the department therefore enabled a transition toward technological innovation. In addition, being part of a university medical center allowed the department to invest in several PhD candidates whose research on protocolized care pathways, IT infrastructure, and the innovation itself supported the transformation.

After the innovation's inception, embedding it required a lot of time, effort, and communication. Many issues emerged, ranging from technical (ie, selecting appropriate monitoring devices and finding ways to integrate data from commercial devices into the EHR) to usability (ie, educating patients and professionals about a new form of care provision and discussing it with them). The department needed to undergo further significant changes

XSL•FO
RenderX

to embed the innovation. Below, we discuss examples of changes in professional tasks and roles.

The department's staff, particularly specialized nurses, had to adapt their tasks and knowledge infrastructure. Crucially, nurses became responsible for interpreting the incoming RPM data and for following up with appropriate actions. This new task required them to respecialize to provide care based on data produced remotely by patients. Specifically, as one of them stated, nurses had to become "aware that [a number] is just a number, […] measured by the patient in the home setting, by a device of which I'm not sure how old or how reliable it is" (Nurse 1). This required them to view measurements as not always trustworthy (as numbers that offered guidance but could not be taken at face value). This new orientation stemmed from firsthand experiences on the job and from sharing these experiences with their colleagues. Furthermore, the department created a new role: "eHealth consultants," tasked with distributing the innovation to patients, instructing them in its use, and addressing technical questions. By hiring eHealth consultants, the department acknowledged the work needed to guide patients in their use of the innovation and formally integrated tasks previously conducted informally by PhDs into the organizational structure. These organizational changes resonate with the ecological perspective in our theoretical framework.

Complexity Also Lies in a Process of Mutual Adaptation of Innovation and Context

Embedding the innovation in the department also involved adapting the innovation itself, reflecting the critical perspective articulated above. For example, the IT department, in collaboration with local nurses, developed a new dashboard in the EHR to visualize incoming data. They also installed secured software to enable e-consultations with patients and developed an app to give patients personalized instructions. The hospital's physical infrastructure was also adapted. Specifically, the department transformed a central space in the hospital into an office where patients could meet the eHealth consultants and discuss their questions. As explained by a cardiologist:

> *At this point there also came a different kind of department that had a focus on the innovation only, so they make sure that the patient gets the devices, they make sure that everything is electronically connected the right way, that the data are coming in, they always call the patients within 2-3 weeks to make sure that everything is going well, that the data are coming in. So I think that was crucial.* [Doctor 1]

Concurrently, Structural Barriers Need to be Addressed

While the respondents were busy preparing the local context for the innovation and at the same time adapting the innovation itself, barriers of a structural nature were addressed but not yet solved permanently (ie, negotiating issues of safety, financing, and medical device regulation). For example, all the infrastructural changes were financed through the departmental budget and temporary research grants because there was no regular reimbursement available for these types of care provisions. According to the respondents, this was possible only because the cardiology department had access to more financial resources than other departments in this and other hospitals. Moreover, the department was responsible for the care pathway of the patient target group from beginning to end, giving them the autonomy necessary to transform care provision. An external project manager explained:

> *Because it's only their department, that makes them really quick in making decisions and going forward.* [Project Manager 2]

Analyzing the innovation's organizational embedding highlights the necessity of addressing structural financial and regulatory barriers early on and the need for a flexible and continuously developing local infrastructure. In addition, the innovation itself changed in response to the requirements of the local setting. With regard to embedding, the respondents' narratives resonate with the structural, ecological, and critical perspectives.

Part II: Gradual Adaptation of the Innovation Within the Original Setting

We now turn to an analysis of the innovation's evolution from its initiation (2015) to the period of data collection (2022). This period saw a local scale-up in terms of (1) the patient population covered by the innovation, (2) the technical aspects of the innovation, and (3) the number of departments using the innovation. Each of these expansions of coverage faced several complexities.

Aspects in the Local Setting That Complicate Gradual Scale-Up

Some of these complexities relate to the ecological perspective. To begin with, local scale-up focused on replicating the innovation within the cardiology department to cover new patient groups. This had major consequences for the care professionals involved (again, nurses in particular). Patient numbers and data collection requirements increased, placing a significant burden on the department's care professionals, who perceived the amount of data generated by up to 400 patients daily as overwhelming. Moreover, scaling up to other patient populations increased uncertainty in nurses' daily tasks, leaving them unable to plan and putting them under more stress. One respondent explained:

> *If you have patients who are continuously doing these measurements at times that they find suitable, they contact you at unpredictable times with questions that can be emotional, that can be technical, they can be completely fine, but they can also be extremely ill, and you must adjust your actions as a medical professional accordingly. That type of unpredictability, when you don't know at the beginning of the day where it's going to end, that just introduces some stress.* [Doctor 2]

Several nurses decided to "evaluate critically how often they check those measurements, because it is such a huge investment of time and such a burden" (Nurse 2). Nurses started looking critically at the real benefits of the deluge of data they received and began to wonder whether "it could potentially be better to place the responsibility with the patients" (Nurse 2). As a result, patients were increasingly instructed to keep an eye on their

XSL•FO
RenderX

own data and explicitly made responsible for contacting the hospital if they believed something was wrong.

Gradual Scale-Up Also Requires Continuous Adaptation of the Innovation Itself

Complexities described by the respondents also resonate with the critical perspective, since the innovation had to be adapted based on the needs and possibilities of the local context. For instance, the project leads continuously added hardware (ie, new measuring devices) to the innovation to provide more extensive data. One of them explained:

> *What we're trying to do is check whether there are more non-invasive devices that you can combine to make sure that you can see that the patient is developing heart failure again as soon as possible. So that's why we're continuously monitoring and checking, okay, what could we add that could possibly help us? Because now we only have the step counter and the weight scale and the blood pressure monitor, but maybe, just maybe, it will help us if we can look at the sleep monitor as well. Because there will be some data that we can combine.* [Doctor 1]

Since new hardware and datapoints would further exacerbate the data issues experienced by nurses, the IT department, in collaboration with nurses, started developing an artificial intelligence (AI) model aimed at analyzing the growing amount of patient data and flagging those patients in need of nurses' attention. Although the model had not been implemented yet, many of the respondents agreed that AI was necessary for the innovation to be workable in practice, even more so given the scaling-up ambitions.

All these add-ons made it difficult to draw clear boundaries around the innovation. In addition, the innovation increasingly became an unstable object for local stakeholders, which in turn complicated further scale-up plans. One doctor stated:

> *The project has not been finalized yet. […] So, I would say, in order for it to be implemented in other hospitals, you need to come to at least a sort of 1.0 solution in which you have a proper and clear description of the product service design, [but] we are still designing it as we go along.* [Doctor 2]

Interestingly, despite not considering the innovation as a finished product, local professionals still attempted to scale it to other departments, especially following strategic investments from the hospital. As the context of scale-up started to move away from the original department, the stakeholder strategies that had functioned in the local context began to fall short. Stakeholders attempted to generate a process of mutual adaptation between local context and innovation, replicating the continuous changes that the "original" innovation underwent at the cardiology department. Many respondents acknowledged, however, that it was inherently much more difficult to scale the innovation to other hospital departments.

Structural Barriers Impede the Gradual Scale-Up of the Innovation

The factors impeding the scaling of the innovation within the hospital also resonate with the structural perspective, for example, the limitations of the external IT infrastructure adopted in other departments, their tight budgets, and the lack of eHealth consultants and dedicated IT staff there. Respondents reported being able to work around some of these barriers, for instance, by drawing on temporary grants and budgetary slack to compensate for the lack of structural reimbursement and by postponing discussions around regulatory safety and liability. One respondent explained:

> *Everything we have done so far was a little bit in this department, a little bit in that department. We had different ways of presenting the data [from patients], we had different IT infrastructural routes; […] the hospital picked five different departments and said, "every department gets some money for fifty patients per [RPM innovation], they get some money from a grant". […] For now, the departments, they arrange the technical explanations to the patients themselves. But, if we get new [versions of the innovation addressing different conditions], you know – the workload is already quite high with the nurse practitioners… […] In an ideal world, you would like to have an overall eHealth department that can do the support of all the [innovations in different departments]. I think that would be necessary if we were really scaling up the [innovation] within the [hospital].* [Project Manager 2]

To summarize, local scaling of the innovation within the organization had its challenges, with respondents recounting organizational complexity, mutual adaptation of context and innovation, and structural barriers. Yet despite these complexities, we can conclude that local scale-up successfully extended the innovation's coverage to include more patients, more aspects of the care pathway, and more departments.

Part III: Attempts to Replicate the Innovation in a New Setting

At the time of our interviews, the respondents considered their innovation as one that was still evolving with the gradual addition of new software, hardware, and patients. They also envisaged further replications of the innovation in departments and hospitals nationwide. Generally, they viewed broader scale-up attempts as inevitable. Although many respondents mentioned this as a concrete possibility in the (near) future, they were well aware of the associated complexities.

Aspects of Local Complexity Remain Important Issues for Replication Attempts Across Settings

To begin with, respondents recognized the organizational complexity, arguing that local people and infrastructure cannot be ignored when trying to implement the innovation in another setting, a point clearly reminiscent of the ecological perspective in our framework. In terms of local people, for example, they referred to the attitude of health care professionals toward innovation and to the different ways in which patients interacted

XSL•FO
RenderX

with it. One significant challenge, in their view, lay in convincing medical professionals and local managers to support and adopt RPM, a relatively new type of care provision, because "at this moment […] it's very difficult to prove that it's better than what we used to do" (Doctor 2). One respondent reflected on the attempts to scale up more widely:

> *If you start this, you start with the small groups who are believers, curious people. So, they are motivated to do it. The difficult part comes after that, when you have to scale up, introduce more people to this way of care. And then you find people who say "oh this is extra, I have to look at all the data, I don't have time for that!" We have to explain to them that this is part of our journey.* [Project Manager 1]

In terms of local infrastructure, respondents expected issues to arise around the local IT systems and the funds available to invest in all the necessary adaptations:

> *The other thing is, let's say in terms of the cost structure, if we want to bring this to other hospitals, you also have to think about a lot of logistics. So, the innovation requires centers. [Hospital] has created an office, where you can go and get your stuff as a patient. All that is taken care of. It's not so easy to replicate that in other places [...] with minimal costs. And there are also hidden costs. For example, let's say, the personnel costs are not calculated by [hospital].* [Liaison 1]

The farther away the innovation moves from the epicenter of its origin, the more difficult it seemingly becomes to get the innovation embedded in the organizational setting:

> *The difficulty we have is that the infrastructure of the innovation is now really incorporated in the [hospital] infrastructure. So, if we want to expand to other hospitals, we need to get to a plug-and-play solution, that you get the app with a dashboard that you can connect into as a hospital. To connect them with the [hospital] infrastructure, that is also a step that needs to be made.* [Project Manager 2]

Resolving Structural Barriers Presents One Solution to Stagnating Scale-Up

The respondents emphasized the necessity of a plug-and-play innovation to resolve issues on a broader level, and this is a solution that reflects the structural perspective discussed earlier in this paper. Respondents struggled to articulate viable strategies for overcoming these systemic barriers. This has been illustrated by the following quote:

> *In the short term the business plan behind mobile health technology is bankruptcy. It just means bankruptcy for a classical hospital, so to say. So, there are a lot of hurdles that must either be taken or ignored in order to make this a success.* [Doctor 2]

Unresolved structural barriers mentioned included health care providers not being reimbursed, or not enough, for avoiding unnecessary visits to the hospital; health insurers having to choose from among a growing number of potential eHealth innovations; and a lack of resources for insurers with the largest market share in a region, who would be expected to take the lead in investing in innovations. One respondent summarized:

> *It has nothing to do with the technology, it's purelya cost versus revenue issue. […] that's what makes it difficult to bring this to other organizations.* [Liaison 1]

In addition to financial issues, participants recognized structural barriers to broader scale-up of eHealth innovations in regulatory and quality constraints, in the fragmentation of IT systems used throughout the country, in hospitals' perceived risk aversion when it comes to investing in innovations, in the absence of suitable hospital infrastructure and new professional roles supporting eHealth care provision, and in the environmental unsustainability of data storage and single-use medical devices. Nevertheless, respondents did emphasize that they expected these issues to be resolved over time:

> *Like I said, there is hope, we just need to accept that the way we organize our health insurance system is not going to help us implement digital solutions. There is enough awareness, so I guess at some point in time we will come up with a solution, but it's going to take time.* [Doctor 2]

Moreover, respondents considered several strategies for dealing with the structural complexities:

> *We are working on the general issues, for instance liability, ethics, data ownership. We organize meetings with all the (national) institutes that are responsible, we convene. Because everybody needs the same answers. They have the same questions at least.* [Project Manager 1]

In addition, they recounted that collaborations with private companies had been considered to support broader scale-up, often framed as "commercialization" (Researcher 2). In this scenario, selling the innovation to an independent organization would allow them to outsource legal liability and the development of an independent IT infrastructure. The strategies respondents described to address structural barriers in reimbursement and regulatory systems, however, assumed that the innovation was a finished product. For the insurer to provide a reimbursement code, authorities to provide certification, or a private company to sell the innovation, there needed to be agreement on a stable and finalized innovation.

Acknowledging the Malleability of the Innovation Presents Another Solution to Stagnating Scale-Up

While some respondents described the innovation as always boiling down to "the same object" (IT Professional 2) regardless of setting, others held the view that it could not simply be considered "a thing" that was reproducible across settings. Consistent with the critical perspective identified above, these respondents claimed that every instantiation of the innovation would in fact amount to another entity altogether, because the innovation had to shed some features and acquire new ones in order to work locally. This narrative problematized the depiction of the innovation as a stable technological object:

XSL•FO
RenderX

> *[Although] there is one message and one goal, […] the [innovation] has been expanded from that original one. [...] See, you see an [innovation]. But it is the whole idea behind it, it's the concept. It's not a technology, it's a concept. That can be difficult to explain.* [Project Manager 1]

These "critical" respondents questioned what constituted the core of the innovation. Was it a form of care provision at home with the involvement of remote technology? A preventative intervention to keep patients out of the hospital? Or nothing more than a concept, an idea about the values of contemporary health care? These reflections (ie, which of the core features of the innovation had to be replicated for the replication to count as scale-up) have important implications for how we appreciate the complexities of broader scale-up. One respondent commented:

> *[N]obody cares about the devices. How are you selling something that is a way of working right? How are you selling change management? [...] Maybe the [innovation] is more like a consultancy service that you buy as a hospital. [Liaison 1]*

In summary, in trying to make sense of the complexities of broader scale-up, our respondents were caught up in a paradox. On the one hand, they considered the innovation a clearly demarcated product subject to financial and regulatory issues. On the other hand, they also acknowledged that it needed to be malleable to deal with the complexities of transitioning to another context. In our discussion, we reflect on this tension and draw lessons from it for both theory and practice.

Key Insights and Discussion

Overview

To dissect the complexities involved in scaling up eHealth innovations, we have combined different theoretical perspectives to make sense of the narrative constructed by stakeholders involved in the scale-up of an innovative eHealth technology. Based on insights from different fields of literature, we have (in brief) identified a structural perspective focusing on systemic barriers and facilitators, an ecological perspective focusing on local organizational complexity, and a critical perspective focusing on mutual adaptation of context and innovation. The 3 perspectives provide complementary explanations for the complexities perceived in the scale-up of eHealth innovation. The structural perspective, for example, aligned with respondents' observations regarding flawed reimbursement systems and fragmented IT infrastructure within and between Dutch hospitals. The ecological perspective resonated with respondents' reflections on the importance of convincing medical professionals of the innovation's value, as well as on the reconfiguration of nurses' tasks and the establishment of a central eHealth office to improve workflow. The critical perspective was consistent with respondents acknowledging the ongoing evolution and adaptation of the innovation, for example, with the addition of novel monitoring devices and AI-powered software.

Our analysis thus shows the importance of adopting a cross-disciplinary perspective when examining eHealth scale-up and its associated complexities. While the structural perspective tends to overlook the flexibility and malleability necessary for scaling an innovation to another organizational context, the ecological and critical perspectives fail to connect local experiences with the stability perceived to be required of the innovation to reach upscaling at broader levels. Below, we elaborate on 2 distinct insights that emerged from this combination of perspectives and discuss their implications for researchers, practitioners, and policy makers.

Differing Strategies Used for Scaling Up eHealth at Different Levels of Scale

While the respondents' narratives bore traces of each of the 3 perspectives, the strategies they used to deal with scale-related complexities differed considerably depending on the level (ie, local versus interorganizational). When discussing *local scale-up*, the respondents stated that they worked on changing the context, adapting practices and infrastructure to embed the innovation, and reconfiguring the innovation itself. In this sense, the flexibility of the organization and innovation was crucial. Although respondents mentioned several structural barriers as having affected the progress of the innovation (eg, the lack of reimbursement agreements with health insurers), they did not resolve them formally but rather worked around them informally.

In contrast, when reflecting on the plans for *broader scale-up*, respondents questioned the strategy of informality and flexibility. When it came to scaling the innovation to other organizations, respondents emphasized the need to find formal solutions to systemic barriers, such as national reimbursement arrangements, an integrated IT network, and national regulatory and liability agreements. Moreover, to realize this, the innovation itself had to cease being malleable and become a stable "product." The view of innovation as a formalized stable entity stems from the structural perspective, which assumes that innovation is a thing embedded in formal structures. In contrast, the critical perspective emphasizes how scaling an innovation to another context entails changing what the innovation *is*. The idea that the innovation never becomes a stable entity amenable to replication does not align with the perceived necessity of turning it into a clearly demarcated product. This tension has critical implications for the solutions respondents envisage for dealing with complexities at a broader level of scale-up: proposing concrete solutions to overcome systemic barriers assumes that the innovation can achieve a state of closure, and that is not what the respondents experienced.

To summarize, our analysis revealed a tension between 2 different conceptions of eHealth innovation: as something that is malleable *and* entangled in an organization, and as a stable product that can be replicated across contexts without changing. This tension results in an impasse in developing strategies to overcome the complexities involved in broader scale-up, with respondents being incentivized to keep the innovation malleable and fixed at the same time.

XSL•FO
RenderX

Importance of Path Dependencies Shaping Chances for Scale-Up

In addition to the important role of the level of scale, we identified another aspect that is often overlooked in understanding the complexities of scaling up eHealth innovation. Prior research has often advocated establishing the infrastructure needed for local scale-up without *acknowledging the influence of what already exists*. Although the ecological perspective emphasizes the need to adapt the local setting, what is missing from this view is the path dependencies stemming from previous decisions. Defined by Mahoney [40] as "those historical sequences in which contingent events set into motion institutional patterns or even chains that have deterministic properties," path dependencies refer to how current events depend at least in part on a chain of prior events. When translating this to innovation scale-up, it becomes clear that success is predicated not only on present and future actions but also on how past choices have shaped organizational structures. It is unlikely that all local circumstances can be adapted to accommodate an existing innovation; some are likely to complicate eHealth scaling. An example can be found in the discussion surrounding the pre-existing IT infrastructure of the cardiology department where the RPM innovation originated. The presence of an in-house EHR allowed for a tailored flexible infrastructure into which the innovation's devices and data could be integrated, something that would have been much more difficult in a commercial platform. However, the use of such platforms is the prevailing reality in most hospitals in the Netherlands.

In summary, pre-existing local circumstances can significantly shape eHealth innovations' initiation and scale-up, and the associated path dependencies should be taken into account when attempting eHealth scale-up both locally and more broadly.

Recommendations for Practice and Policy

This article offers several lessons for practice and policy. First, the challenges faced by actors at the local and broader levels are likely to differ greatly. Actors at the local level are likely to face complexities related to the continuous adaptation of both the organizational infrastructure and the innovation itself. Even so, there is little sense of urgency at this level about addressing structural barriers formally and systematically. In contrast, actors at the broader level need to deal with the tension between the limitations imposed by structural systems and the requirement of local flexibility. To do this, decision makers must, first and foremost, acknowledge this tension and the confusion that is likely to result regarding the strategies that actors pursue in their attempts to scale up broadly.

The second lesson, related to the previous point, is that flexibility should be incorporated into system-wide structures that govern eHealth innovation (including IT infrastructure, regulations, and reimbursement mechanisms). Specifically, this would mean creating opportunities for local adaptation work, facilitating ingenuity in local contexts. While fostering such "facilitating space" will not remove all complexity, it will help actors deal with complexity by giving them a certain amount of leeway to tailor and adapt to the requirements of specific local contexts. This complexity includes the path dependencies imposed by pre-existing circumstances that define the opportunities for scaling up specific innovations.

That is why actors should be aware that, when adopting a scaled-up innovation, they will likely need to adapt the innovation itself as well as the local context. At the policy level, such awareness should be made part of reimbursement and regulatory structures. For decision makers, this means not expecting a "plug-and-play" model of scale-up where innovation is simply replicated; rather, they must set aside the time and financial and human resources necessary for adaptation at the local level.

Recommendations for Research

Our analysis leads us to make 3 recommendations for future research. First, despite discussing scale-up at length, existing literature lacks a conceptual vocabulary for studying the concept of scale itself, including its implications for the complexities stakeholders face and the strategies they use. There is almost no research addressing the phenomenon of scale in innovation scale-up, the notable exception being a study of strategies addressing scale-up complexities at different levels of scale in the energy sector [41]. Future research on health care innovation scale-up should thus acknowledge the concept of the scale itself (ie, the level at which scale-up is attempted) as a crucial factor in determining actors' strategies.

Second, research should focus on whether and how actors involved in different types of innovations may experience complexities at different levels of scale. The RPM innovation studied in this paper is only one of many types of eHealth innovations, each with specific characteristics that may affect the process of scaling. Others could, for example, have more stable or more malleable aspects and come up against more or fewer unresolved systemic barriers. More research is needed to further our understanding of these complexities and identify a broader set of action repertoires to deal with them.

Third, this article adopted a cross-disciplinary perspective, applying a generative heuristic framework to highlight the various complexities stakeholders encounter in the scale-up of eHealth innovations through a high-level discussion. Consequently, our analysis leaves considerable scope for research on more specific aspects of these complexities. One potentially relevant direction for research could be the interaction of humans with technologies, the novel challenges emerging from this interaction in care processes [42], and the ways in which this interaction and the related challenges may differ based on the scale attempted.

Limitations Related to the Case Study

There are several limitations associated with our case study. First, for reasons of unavailability and privacy protection, we were unable to include the perspectives of all stakeholders currently involved in the RPM innovation project. Specifically, we did not include the perspectives of eHealth consultants, patients, and stakeholders in other departments outside of the original cardiology setting. Stakeholders in departments attempting to scale the innovation could have provided additional insights into the complexities of scale-up. We were also unable to speak to the nurses directly involved in the

XSL•FO
RenderX

original version of the innovation owing to the work pressure they were experiencing. We did, however, talk to their colleague nurses, who witnessed how the workload associated with the innovation proved demotivating for several nurses.

Second, several biases could have emerged from the interview-based nature of the case study. Specifically, both the retrospective (recall bias, potentially misremembering events in the past) and prospective (declinism, the tendency to perceive the future more negatively than the past) nature of the respondents' reflections might have shaped our findings. We cannot preclude the possibility that we missed relevant aspects, as we were unable to observe the innovation's development as it happened. However, the interviews did allow us to piece together a longitudinal narrative spanning a much longer time period than any direct observations would have allowed. Moreover, we did reach analytical saturation based on our interviews, insofar as respondents presented similar views on scale-up across interviews, strengthening our confidence in the validity of our conclusions.

Conclusion

This article addressed the question of why scaling up innovative eHealth technologies is so challenging in practice. For this purpose, we brought together 3 theoretical perspectives on the complexities of innovation scale-up from different fields of literature. We used these perspectives to make sense of the narrative produced by stakeholders involved in the scale-up of an RPM-based eHealth technology, which was presented as a local success but came up against challenges in broader scale-up attempts. Each of these perspectives highlights different but equally pertinent aspects of scale-up complexities and strategies for addressing them. Two key insights emerged from this cross-disciplinary analysis. First, we found that the level at which scale-up is pursued plays an important, yet so far neglected, role. Contextual complexities were overcome at the local level and systemic barriers were informally worked around. By contrast, at a broader level, tension emerged between the need for stability on the one hand and malleability on the other, leading to an impasse in the scale-up of the eHealth innovation. Second, our study has emphasized the role of path dependencies, namely in terms of pre-existing organizational structures and technological infrastructure, in facilitating and constraining scale-up processes. The path dependencies in local contexts play an important role in shaping the complexities that actors face.

Researchers, policy makers, and stakeholder practitioners need to acknowledge and account for the crucial roles that the level of scale and path dependencies play in shaping the complexities involved in scaling up eHealth innovations. Such projects might enjoy greater success by rethinking structural systems to allow for malleability in the innovation, giving it the necessary leeway to align with the requirements of specific local contexts.

Acknowledgments

We thank all the participants who contributed to our research at various stages, especially the anonymous respondents for generously sharing their time and experiences with us. We are thankful to our colleagues at Erasmus School of Health Policy & Management for their valuable feedback on earlier versions of the paper. We also want to acknowledge the program Medical Delta's Journey from Prototype to Payment for funding the PhD trajectories as part of which this article was written. The funder was not involved in the conduct of the research or the preparation of the article.

Conflicts of Interest

None declared.

References

1. Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan JP, et al. What is eHealth (4): a scoping exercise to map the field. J Med Internet Res 2005 Mar 31;7(1):e9 [FREE Full text] [doi: 10.2196/jmir.7.1.e9] [Medline: 15829481]
2. van den Heuvel JF, Groenhof TK, Veerbeek JH, van Solinge WW, Lely AT, Franx A, et al. eHealth as the next-generation perinatal care: an overview of the literature. J Med Internet Res 2018 Jun 05;20(6):e202 [FREE Full text] [doi: 10.2196/jmir.9262] [Medline: 29871855]
3. Duan Y, Shang B, Liang W, Du G, Yang M, Rhodes RE. Effects of eHealth-based multiple health behavior change interventions on physical activity, healthy diet, and weight in people with noncommunicable diseases: systematic review and meta-analysis. J Med Internet Res 2021 Feb 22;23(2):e23786 [FREE Full text] [doi: 10.2196/23786] [Medline: 33616534]
4. Bashi N, Karunanithi M, Fatehi F, Ding H, Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews. J Med Internet Res 2017 Jan 20;19(1):e18 [FREE Full text] [doi: 10.2196/jmir.6571] [Medline: 28108430]
5. Heckemann B, Wolf A, Ali L, Sonntag SM, Ekman I. Discovering untapped relationship potential with patients in telehealth: a qualitative interview study. BMJ Open 2016 Mar 02;6(3):e009750 [FREE Full text] [doi: 10.1136/bmjopen-2015-009750] [Medline: 26936904]
6. Wildevuur SE, Simonse LWL. Information and communication technology-enabled person-centered care for the "big five" chronic conditions: scoping review. J Med Internet Res 2015 Mar 27;17(3):e77 [FREE Full text] [doi: 10.2196/jmir.3687] [Medline: 25831199]

XSL•FO
RenderX

7. Kierkegaard P. eHealth in Denmark: a case study. J Med Syst 2013 Dec;37(6):9991. [doi: 10.1007/s10916-013-9991-y] [Medline: 24166019]
8. Asthana S, Jones R, Sheaff R. Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors. BMC Health Serv Res 2019 Dec 21;19(1):984 [FREE Full text] [doi: 10.1186/s12913-019-4790-x] [Medline: 31864370]
9. Thoumi A, Udayakumar K, Drobnick E, Taylor A, McClellan M. Innovations in diabetes care around the world: case studies of care transformation through accountable care reforms. Health Aff (Millwood) 2015 Sep;34(9):1489-1497. [doi: 10.1377/hlthaff.2015.0403] [Medline: 26355050]
10. Baur C. An analysis of factors underlying e-health disparities. Camb Q Healthc Ethics 2008 Oct 01;17(4):417-428. [doi: 10.1017/s0963180108080547]
11. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: A systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
12. Oderanti FO, Li F, Cubric M, Shi X. Business models for sustainable commercialisation of digital healthcare (eHealth) innovations for an increasingly ageing population. Technological Forecasting and Social Change 2021 Oct;171:120969. [doi: 10.1016/j.techfore.2021.120969]
13. Schlieter H, Marsch LA, Whitehouse D, Otto L, Londral AR, Teepe GW, et al. Scale-up of digital innovations in health care: expert commentary on enablers and barriers. J Med Internet Res 2022 Mar 11;24(3):e24582 [FREE Full text] [doi: 10.2196/24582] [Medline: 35275065]
14. Global strategy on digital health 2020-2025. World Health Organization. Geneva; 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-09-15]
15. Integraal Zorgakkoord: 'Samen werken aan gezonde zorg'. Rijksoverheid. The Hague; 2022. URL: https://www.rijksoverheid.nl/documenten/rapporten/2022/09/16/integraal-zorgakkoord-samen-werken-aan-gezonde-zorg [accessed 2024-09-15]
16. Kamerbrief voortgangsrapportage eHealth en zorgverbetering. Rijksoverheid. The Hague; 2015 Oct. URL: https://www.rijksoverheid.nl/documenten/kamerstukken/2015/10/08/kamerbrief-voortgangsrapportage-ehealth-en-zorgverbetering [accessed 2024-09-15]
17. Richter P, Harst L. Tackling the scaling-up problem of digital health applications. Z Gesundh Wiss 2022 Jun 05;30(1):1-3 [FREE Full text] [doi: 10.1007/s10389-021-01599-7] [Medline: 34109100]
18. Coiera E. Why e-health is so hard. Med J Aust 2013 Mar 04;198(4):178-179. [doi: 10.5694/mja13.10101] [Medline: 23451947]
19. Cho S, Mathiassen L, Gallivan M. Crossing the diffusion chasm: from invention to penetration of a telehealth innovation. Information Technology & People 2009 Nov 13;22(4):351-366. [doi: 10.1108/09593840911002450]
20. Oderanti FO, Li F. Commercialization of eHealth innovations in the market of the UK healthcare sector: A framework for a sustainable business model. Psychology and Marketing 2018 Jan 10;35(2):120-137. [doi: 10.1002/mar.21074]
21. Rauwerdink A, Kasteleyn MJ, Chavannes NH, Schijven MP. Successes of and lessons from the first joint eHealth program of the Dutch University hospitals: evaluation study. J Med Internet Res 2021 Nov 25;23(11):e25170 [FREE Full text] [doi: 10.2196/25170] [Medline: 34842536]
22. Doorwerking van de Citrienprogramma's. Citrienfonds. 2023 Jul. URL: https://www.citrienfonds.nl/actueel/nieuws/doorwerking-van-de-citrienprogrammas/ [accessed 2024-09-15]
23. Corôa R, Gogovor A, Ben Charif A, Hassine A, Zomahoun H, McLean R, et al. Evidence on scaling in health and social care: An umbrella review. Milbank Q 2023 Sep 25;101(3):881-921. [doi: 10.1111/1468-0009.12649] [Medline: 37186312]
24. Standing C, Standing S, McDermott M, Gururajan R, Kiani Mavi R. The paradoxes of telehealth: a review of the literature 2000–2015. Syst. Res 2016 Dec 01;35(1):90-101. [doi: 10.1002/sres.2442]
25. Ossebaard HC, Van Gemert-Pijnen L. eHealth and quality in health care: implementation time. Int J Qual Health Care 2016 Jun;28(3):415-419. [doi: 10.1093/intqhc/mzw032] [Medline: 27029590]
26. Moore M, Riddell D, Vocisano D. Scaling out, scaling up, scaling deep: strategies of non-profits in advancing systemic social innovation. J Corporate Citizenship 2015 Jun 01;2015(58):67-84. [doi: 10.9774/gleaf.4700.2015.ju.00009]
27. Spicer N, Bhattacharya D, Dimka R, Fanta F, Mangham-Jefferies L, Schellenberg J, et al. 'Scaling-up is a craft not a science': catalysing scale-up of health innovations in Ethiopia, India and Nigeria. Soc Sci Med 2014 Nov;121:30-38 [FREE Full text] [doi: 10.1016/j.socscimed.2014.09.046] [Medline: 25306407]
28. Wang CJ, Liu TT, Car J, Zuckerman B. Design, adoption, implementation, scalability, and sustainability of telehealth programs. Pediatr Clin North Am 2020 Aug;67(4):675-682. [doi: 10.1016/j.pcl.2020.04.011] [Medline: 32650866]
29. Gijsbers H, Feenstra TM, Eminovic N, van Dam D, Nurmohamed SA, van de Belt T, et al. Enablers and barriers in upscaling telemonitoring across geographic boundaries: a scoping review. BMJ Open 2022 Apr 20;12(4):e057494 [FREE Full text] [doi: 10.1136/bmjopen-2021-057494] [Medline: 35443957]
30. Greenhalgh T, Papoutsi C. Spreading and scaling up innovation and improvement. BMJ 2019 May 10;365:l2068 [FREE Full text] [doi: 10.1136/bmj.l2068] [Medline: 31076440]

XSL•FO
RenderX

31. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
32. May CR, Mair F, Finch T, MacFarlane A, Dowrick C, Treweek S, et al. Development of a theory of implementation and integration: Normalization Process Theory. Implement Sci 2009 May 21;4:29 [FREE Full text] [doi: 10.1186/1748-5908-4-29] [Medline: 19460163]
33. Hidefjäll P, Laurell H, Johansson J, Barlow J. Institutional logics and the adoption and implementation of remote patient monitoring. Innovation 2023 Jan 03:1-20. [doi: 10.1080/14479338.2022.2162907]
34. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
35. Pfotenhauer S, Laurent B, Papageorgiou K, Stilgoe AJ. The politics of scaling. Soc Stud Sci 2022 Feb 08;52(1):3-34 [FREE Full text] [doi: 10.1177/03063127211048945] [Medline: 34625011]
36. Tsing A. On nonscalability: the living world is not amenable to precision-nested scales. Common Knowledge 2012;18(3):505-524. [doi: 10.1215/0961754X-1630424]
37. Hanna A, Park T. Against Scale: Provocations and Resistances to Scale Thinking. arXiv preprint. 2020 Oct. URL: https://arxiv.org/abs/2010.08850 [accessed 2024-11-15]
38. Timmermans S, Tavory I. Theory construction in qualitative research. Sociological Theory 2012 Sep 10;30(3):167-186. [doi: 10.1177/0735275112457914]
39. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
40. Mahoney J. Path dependence in historical sociology. Theory and Society 2000;29(4):507-548 [FREE Full text]
41. Raven R, Schot J, Berkhout F. Space and scale in socio-technical transitions. Environmental Innovation Societal Transitions 2012 Sep;4:63-78. [doi: 10.1016/j.eist.2012.08.001]
42. Rahman Jabin MS, Pan D. Software-related challenges in Swedish healthcare through the lens of incident reports: a desktop study. Digit Health 2023;9:20552076231203600 [FREE Full text] [doi: 10.1177/20552076231203600] [Medline: 37744748]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
RPM: remote patient monitoring

Edited by A Schwartz; submitted 02.03.24; peer-reviewed by T Greenhalgh, K Lawrence, M Rahman Jabin; comments to author 17.05.24; revised version received 12.07.24; accepted 02.10.24; published 02.12.24.

Please cite as:
Allers S, Carboni C, Eijkenaar F, Wehrens R
A Cross-Disciplinary Analysis of the Complexities of Scaling Up eHealth Innovation
J Med Internet Res 2024;26:e58007
URL: https://www.jmir.org/2024/1/e58007
doi:10.2196/58007
PMID:

©Sanne Allers, Chiara Carboni, Frank Eijkenaar, Rik Wehrens. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Digital Information Ecosystems in Modern Care Coordination and Patient Care Pathways and the Challenges and Opportunities for AI Solutions

You Chen[1,2], PhD; Christoph U Lehmann[3,4], MD; Bradley Malin[1,2,5], PhD

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[2]Department of Computer Science, Vanderbilt University, Nashville, TN, United States
[3]Clinical Informatics Center, University of Texas Southwestern Medical Center, Dallas, TX, United States
[4]Institut für Medizinische Informatik, Universitäts Klinikum Heidelberg, Heidelberg, Germany
[5]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
You Chen, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave
Nashville, TN, 37203
United States
Phone: 1 6153431939
Email: you.chen@vanderbilt.edu

Abstract

The integration of digital technologies into health care has significantly enhanced the efficiency and effectiveness of care coordination. Our perspective paper explores the digital information ecosystems in modern care coordination, focusing on the processes of information generation, updating, transmission, and exchange along a patient's care pathway. We identify several challenges within this ecosystem, including interoperability issues, information silos, hard-to-map patient care journeys, increased workload on health care professionals, coordination and communication gaps, and compliance with privacy regulations. These challenges are often associated with inefficiencies and diminished care quality. We also examine how emerging artificial intelligence (AI) tools have the potential to enhance the management of patient information flow. Specifically, AI can boost interoperability across diverse health systems; optimize and monitor patient care pathways; improve information retrieval and care transitions; humanize health care by integrating patients' desired outcomes and patient-reported outcome measures; and optimize clinical workflows, resource allocation, and digital tool usability and user experiences. By strategically leveraging AI, health care systems can establish a more robust and responsive digital information ecosystem, improving care coordination and patient outcomes. This perspective underscores the importance of continued research and investment in AI technologies in patient care pathways. We advocate for a thoughtful integration of AI into health care practices to fully realize its potential in revolutionizing care coordination.

(J Med Internet Res 2024;26:e60258) doi:10.2196/60258

KEYWORDS

patient care pathway; care journey; care coordination; digital information ecosystem; digital technologies; artificial intelligence; information interoperability; information silos; workload; information retrieval; care transitions; patient-reported outcome measures; clinical workflow; usability; user experience workflow; health care information systems; networks of health care professionals; patient information flow

Introduction

Care coordination, as defined by the US National Academy of Medicine, involves the organization of patient care activities and the deliberate distribution of information among stakeholders—including clinicians and patients—to ensure the appropriate delivery of health care services [1,2]. These services include diagnostics, treatments, care transitions, and the monitoring of patient self-reported measures and clinical outcomes. Efficient exchange and communication of information across a patient's care pathway is fundamental to the care

XSL•FO
RenderX

coordination process—a continuum that spans all activities from diagnosis through treatment and follow-up [3]. Today's health care landscape relies heavily on digital technologies, such that it would benefit greatly from formally characterizing and optimizing the "digital information ecosystem" [4]. This ecosystem encompasses the complex network of digital technologies and protocols responsible for the generation, updating, transmission, and exchange of information among stakeholders along a patient care pathway. This perspective paper positions digital information ecosystems as the heart of modern care coordination, outlines the current challenges within this ecosystem, and explores how specific applications of artificial intelligence (AI) tools may enhance processes along the patient's care pathway (Figure 1). By focusing on targeted AI solutions, we believe there are major opportunities to reduce the digital burden on stakeholders and enhance the efficiency of digital coordination, thereby improving the overall quality of health care delivery.

Figure 1. Comprehensive overview of the digital information ecosystem in modern care coordination and patient care pathways. The figure provides a detailed visualization of the digital information ecosystem, illustrating the flow of information generation, updating, transmission, and exchange along a patient's care pathway. It encompasses all stages of patient care, including initial consultations, diagnostic tests, imaging, treatment interventions, ongoing monitoring, and referral to follow-up care. The figure highlights several key challenges that hinder the efficient management of the digital information ecosystem, such as interoperability issues, information silos, increased workload and clinician burden, coordination and communication gaps, patient care journey mapping, and the complexities of complying with privacy regulations. Potential artificial intelligence (AI) solutions are depicted as integral to overcoming these challenges, enhancing the efficiency of care coordination, and reducing the burden on clinicians. The figure underscores how AI can streamline processes, improve data flow and accessibility, and foster more effective coordination and collaboration across stakeholders. By integrating AI-driven tools and technologies, this ecosystem can achieve higher efficiency and more responsive patient care.

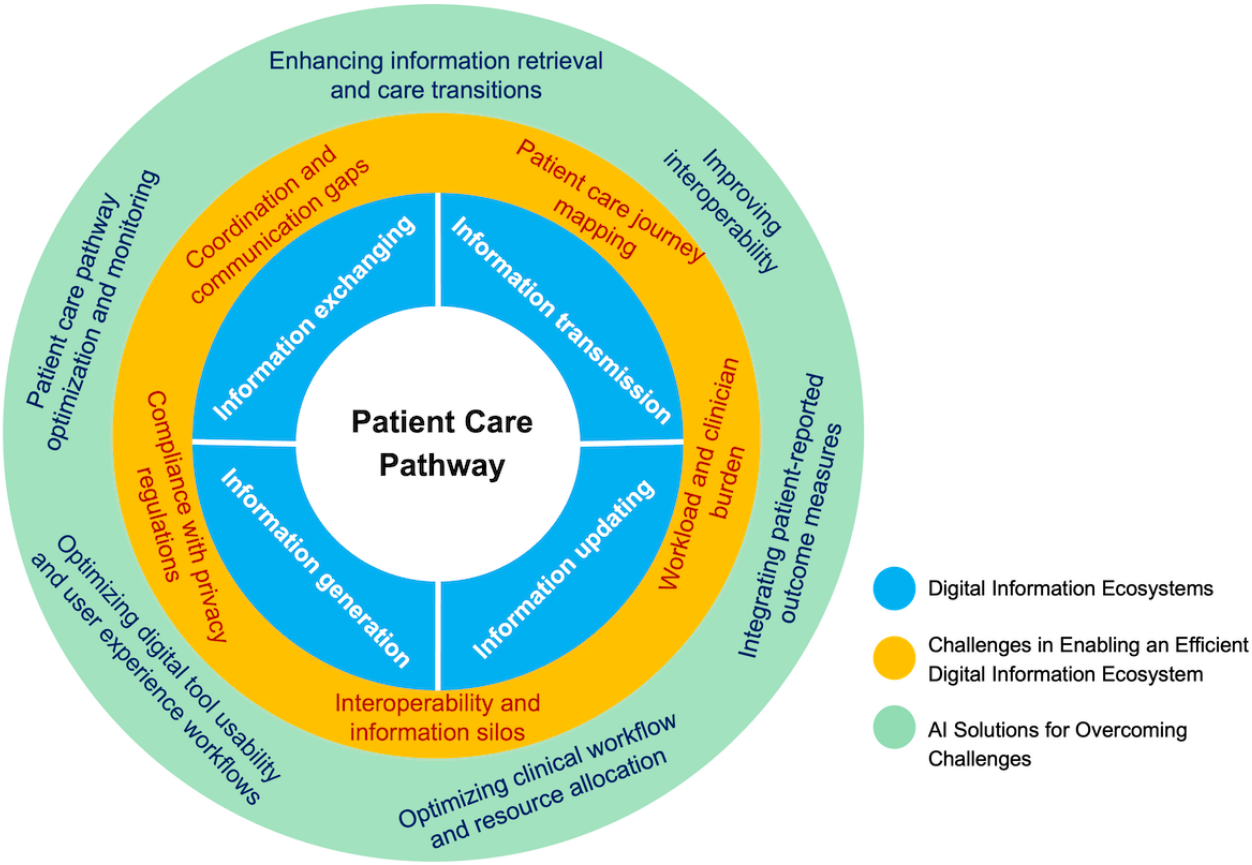

Digital Information Ecosystem: Generation, Updating, Transmission, and Exchange

Overview

The essence of care coordination lies in the seamless generation, update, transmission, and exchange of digital information, which supports both synchronous and asynchronous communications among stakeholders. Information generation corresponds to the creation of new data entries during the provision of health care services. There are numerous examples of information generation, such as the laboratory test results produced during diagnostic testing, or the diagnoses documented by health care professionals following patient evaluation. After generation, information updates occur, which involve modifications to existing health records to reflect new developments or changes in a patient's condition. Examples of these activities include updates to progress reports, medication management, or any alterations made to previously recorded health information. The information transmission process pertains to the movement of information along specific pathways within the health care setting, such as prescriptions being transferred from physicians to pharmacists or care instructions moving from specialists to nurses. Finally, information exchange refers to the dynamic process where health information is not only shared but also

XSL•FO
RenderX

appropriately interpreted and used across different health care systems and stakeholders. Exchange emphasizes the adherence to standards, protocols, or formats to ensure that the information is communicated accurately and meaningfully, supporting coherent and coordinated health care delivery.

An effective digital information ecosystem facilitates a comprehensive and timely understanding of a patient's health, thereby informing clinical decisions and ensuring timely interventions. Given the complexity and extensive scope of the digital information ecosystem, we will simplify our introduction of this concept by focusing on electronic health records (EHRs) and their patient portal extensions. These platforms serve as prime examples of how digital tools facilitate the flow of information among stakeholders. To illustrate this further, we will explore specific scenarios involving the management of chronic diseases, demonstrating how information is exchanged and used within these platforms.

Information Generation Along a Patient's Care Pathway

Information about a patient's health or their health care is generated at every key interaction point along a patient's care pathway, from registration, initial consultations, and diagnostic tests to treatment interventions and ongoing monitoring. Throughout the care pathway for chronic conditions like diabetes, the generated information continuously informs care strategies. The flow of such information, as facilitated by digital platforms such as EHRs or patient portals, evolves with the patient's condition, treatments, and responses, making it a dynamic and integral part of the care coordination process. This flow supports a range of care activities, ensuring that the right stakeholder has the right information at the right time to make informed decisions and provide effective care.

EHRs and patient portals play pivotal roles in generating health care information [5]. EHRs document a comprehensive range of data including diagnoses, procedures, laboratory test results, medications, orders, flowsheets, clinical notes, medical and surgical histories, allergies, and summaries of care during transitions [6,7]. These data can be composed to form the Continuity of Care Record [7], which supports clinical decision-making and ensures continuity of care. Integration of Continuity of Care Record into care coordination was associated, for example, with improvements in glycemic and lipid control for patients with diabetes [7]. Evidence suggests that EHR designs that store patients' values, health goals, and action plans may strengthen the continuity and quality of care between patients and primary care team members [8].

Building on EHRs, patient portals can enhance patient engagement by allowing individuals to actively contribute to their care process and provide self-reported measures and feedback. These platforms can improve communication between clinicians and patients, increase medication adherence, reduce appointment delays, and enhance patient satisfaction [9,10].

Information Updates and Transmission Along a Patient's Care Pathway

Passing information from one care provider to another is an important process that involves accessing and updating patient records. Consider the case of a patient undergoing nephrolithotomy: important stages that rely on shared information include the initial diagnosis, presurgery assessments, the surgery itself, postoperative care, and routine follow-ups. When a diagnosis of kidney stone is confirmed, it is entered into the EHR along with a time stamp and user identification. Every subsequent access or modification of this diagnosis by health care professionals is similarly recorded with a new time stamp. This fine-grained tracking ensures that any changes in the patient's condition or treatment strategy are attributable, illustrating how the diagnosis has been reviewed and adjusted by different health care providers.

The pathway that information takes generates detailed logs regarding which stakeholders accessed or modified the data [11]. Continuing with our example, after the nephrolithotomy, the surgical team may update the patient's EHR with the details of the surgery, any medication administrations, and follow-up needs. These time-stamped logs identify the stakeholder making the modification and indicate the direction of the information flow—such as from a urologist to a primary care physician. This level of detailed tracking enables a clear audit trail, but also is useful in characterizing how medical decisions were made and can be refined [11].

Information Exchanges Along a Patient's Care Pathway

Information flows continuously through care pathways, frequently being exchanged among various nodes within the health system, including hospital departments and units, and diverse stakeholders. Each stakeholder involved in care, whether they are physicians, nurses, administrative staff, or patients, relies on the continuous exchange of data essential for their respective roles. For instance, physicians and nurses require immediate updates to make informed medical decisions, while administrative staff need accurate information for billing and compliance purposes. Patients are increasingly accessing their health data through patient portals to actively participate in their health care decisions.

Stakeholders engaging with EHRs or patient portals participate in asynchronous and synchronous information exchange. Ensuring that the correct information is delivered at the right time is crucial for the efficiency and safety of health care delivery. Delayed or inaccurate exchanges can compromise patient outcomes and increase the administrative burden on health care providers [12,13].

To address these challenges, electronic health information exchange (HIE) standards such as Health Level Seven (HL7) Fast Healthcare Interoperability Resources are crucial to implement and to be refined. These standards have the potential to ensure that every node in the health system involved in a patient's care has access to up-to-date accurate and relevant data, enhancing the overall efficiency and quality of health care delivery [14,15].

XSL•FO
RenderX

Challenges in Enabling an Efficient Digital Information Ecosystem and Effective Care Coordination

Overview

Despite significant advancements in digital technologies and interoperability, including the adoption of standards like Fast Healthcare Interoperability Resources and participation in national networks such as Care Everywhere, CommonWell, and Trusted Exchange Framework and Common Agreement [16], and the use of Direct Secure Messaging for communication between different EHR systems [17], challenges persist in managing the digital information ecosystem and care coordination efficiently. These challenges arise from the complex interactions that occur throughout the patient-centered care journey involving health care information systems and networks of health care professionals, as illustrated in Figure 2. Interoperability issues and information silos emerge when disconnected or incompatible health care information systems hinder seamless data exchange, leading to fragmentation of information. Difficulties in information retrieval, processing, and mapping of key care events impede the modeling of patient care pathways necessary for optimization and monitoring. Coordination and communication gaps can occur due to synchronous and asynchronous interactions through health care information systems, resulting in miscommunication and gaps in coordination among health care professionals. The extensive interaction with multiple health care systems increases the digital workload and contributes to clinician burnout, exacerbating the digital workload and clinician burden. Additionally, navigating regulatory requirements adds complexity to information sharing and coordination efforts, posing challenges in compliance with regulations.

Figure 2. Challenges arising from interactions among the patient-centered care journey, health care information systems, and a network of health care professionals. The figure illustrates 3 interconnected components crucial to modern health care: the patient-centered care journey, health care information systems, and a network of health care professionals. The patient-centered care journey represents a sequence of care events along a patient's health care trajectory. Health care information systems consist of various platforms containing diverse types of health and health care data about the patient. The network of health care professionals includes all the health care providers involved in the patient's care journey, connected through professional collaboration. The interactions among these components give rise to several challenges in establishing an efficient information ecosystem and effective care coordination. EHR: electronic health record.

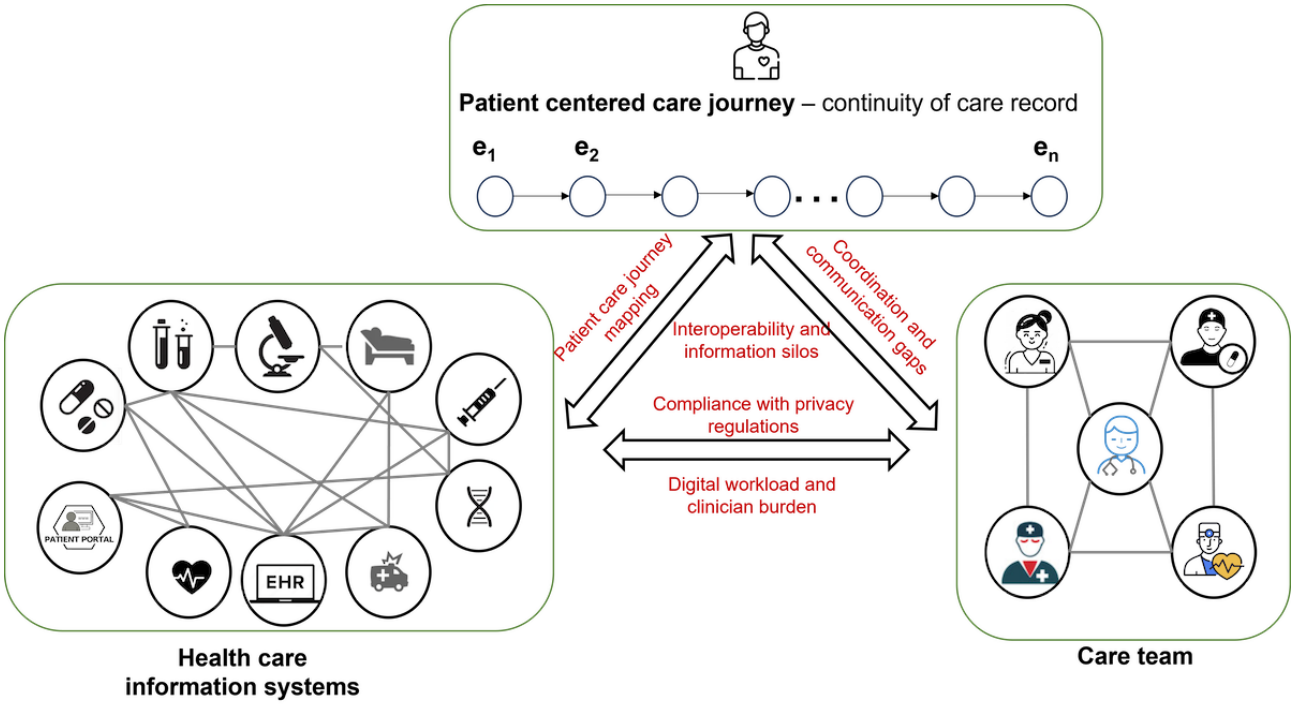

Interoperability and Information Silos

Interoperability in health care is crucial for the seamless sharing and use of information across various systems and platforms. Despite its importance, interoperability remains a significant challenge in digital coordination among health care providers. The landscape of formal HIE platforms in the United States is notably fragmented, characterized by a patchwork of efforts at federal, state, community, enterprise, and EHR vendor levels (eg, Epic Care Everywhere) [18]. As a result, only certain provider organizations are equipped to engage in HIE with others with whom they routinely share patients [19,20].

The use of EHR systems from different vendors, along with other medical devices, often exacerbates challenges, leading to ineffective communication and the creation of information silos within and across organizations [19,21]. These silos significantly hinder the process of coordinated care, as they prevent the efficient flow of crucial health data among the involved parties. This lack of streamlined data exchange continues to be a major barrier to achieving ideal interoperability that enables optimized health care delivery.

XSL•FO
RenderX

Patient Care Journey Mapping

Patient care journey mapping refers to the process of creating a systematic representation of the sequence of care activities experienced by patients, thereby reflecting the continuity of patient care. For instance, Meyer [22] describes a patient with multiple chronic conditions who underwent posterior hip replacement surgery due to advanced avascular necrosis. Over the course of 21 months following the initial surgery, the patient received care across a wide range of settings, including 2 major surgical procedures and 9 care transitions. Without patient care journey mapping and monitoring, this patient and his family were unable to manage or plan medical appointments or self-care tasks. Mapping and monitoring the longitudinal patient journey could identify problems and gaps along the patient's continuum of care, such as prolonged wait times, misalignment between medical needs and care teams, duplicated tests, adverse drug reactions, failure to follow-up, and fragmentation. Fung-Kee-Fung et al [23] showed that a lack of understanding of the patient care journey and the need for dynamic alignment of providers is associated with delays in timely lung cancer care.

Various approaches have been developed for mapping patient care journeys. The Customer Journey Mapping Language is a framework for customer journey analysis [24]. Additionally, data-driven pathway modeling, such as Process Mining, has been relied upon to build pathways by using existing EHR log data to display the actual patient pathway [25]. Conformance analysis in the context of process mining is one of the main forms of process mining [26-28], using a pre-existing process model and an event log. The pre-existing process model is often a protocol, guideline, or formally defined care pathway, and EHRs generally serve as the event logs.

The Unified Modeling Language provides a more system-centric modeling method, connecting stakeholders, tasks, and system intended functionality [29]. However, it is nontrivial to capture the entirety of the patient care journey due to the fact that not all care activities are documented in health information systems. Moreover, even when such information is documented, there are challenges with (1) information siloing, (2) differences in use behaviors and terminologies, and (3) a lack of integrated technologies that can perform information retrieval, processing, and mapping of key care events impede the modeling of patient care pathways, and generate a common language easily understood by involved stakeholders.

Coordination and Communication Gaps

While EHR systems enable real-time access to data, retrieving relevant information efficiently remains a substantial challenge. EHR systems frequently lack direct communication tools that would allow stakeholders to discuss specific patient information. This deficiency can nontrivially impede effective interdisciplinary collaboration, as it complicates team members' ability to access and discuss comprehensive patient data seamlessly, resulting in delay and inaccurate information. In one recent study, it was shown that the application of an EHR system could negatively affect team function, noting that while they support task-oriented and efficient communication, the system may shift attention away from the human needs of the care team [30]. Such observations underscore the double-edged nature of EHR utility in clinical settings.

Introducing features that allow direct annotation or discussion within patient records in the EHR could significantly enhance the clarity and speed of health care delivery. Such functionalities would ensure that all team members are well-informed and aligned in their clinical strategies, fostering a collaborative environment and improving health care outcomes.

Digital Workload and Clinician Burden

EHRs and patient portals are designed to enhance information exchange and care coordination, yet often these tools end up increasing clinician burden due to their complex interfaces and inefficient workflows. For instance, physician-rated EHR usability was independently linked to the odds of burnout [31]. Specifically, each 1-point increase in the System Usability Scale score was associated with a 3% decrease in the odds of experiencing burnout. Thus, despite their intended purpose, EHRs frequently require extensive data entry and navigation, which diverts attention away from patient engagement and care and can contribute to physician fatigue and burnout [32].

This problem becomes particularly pronounced in settings that manage complex care pathways, such as treatment for patients with chronic kidney disease. In these scenarios, timely and accurate information exchange is critical, and the demands placed on clinicians by digital tools can significantly impede the efficiency and effectiveness of care delivery.

Compliance With Regulations

Research shows that while patients are generally open to sharing their protected health information (PHI) to support collaborative care, they continue to have concerns about the privacy of their personal health information during the sharing process [33]. Addressing these concerns is essential. By enhancing security measures and increasing transparency on PHI management, health care providers can build greater trust among patients. This improves the flow of information and ensures that patients understand and feel reassured about the protections in place to safeguard their privacy, which is crucial for maintaining compliance and integrity within the digital information ecosystem.

As digital information traverses the care pathway, ensuring adherence to regulations like the Health Insurance Portability and Accountability Act (HIPAA) becomes increasingly complex. In one survey of surgeons, a non-trivial proportion used text messaging to communicate PHI [34]. Yet HIPAA stipulates that text messaging of PHI is only permissible if the patient has initiated communication via SMS or has requested confidential communications through this medium. It is crucial for stakeholders to implement robust security measures and establish clear protocols to prevent any breach of privacy during electronic communications, ensuring compliance with HIPAA and other federal or state-specific laws.

XSL•FO
RenderX

AI Solutions for Overcoming Challenges in the Digital Information Ecosystem

AI in Improving Interoperability

As the health care industry continues to enhance its health information technology connectivity infrastructure, AI has contributed to integrating and harmonizing data across multiple systems [35-37]. AI significantly advances semantic interoperability—the ability of different systems to exchange data accurately with a common format and meaning [38,39]. This is particularly evident in AI's capacity to extract meaningful information or infer actionable insights from structured and unstructured data sources, thereby improving data exchange across the care continuum.

A key aspect of AI's role is the conversion of free-text information into structured data, which is invaluable for creating comprehensive patient records that are compatible across different systems. This enhances data fluidity and accessibility, making it easier to share information across systems. Natural language processing (NLP)–driven AI tools are essential in transforming unstructured clinical text into a structured format. A recent study using Stanford Open Information Extraction demonstrated how AI can structure diagnostic information from free-text pathology reports [40]. The AI model effectively captured complex diagnostic entities and their relationships, converting this information into knowledge graphs. The resulting structured data showed a high semantic similarity to the original reports (mean weighted overlap of 0.83), illustrating the model's ability to preserve the integrity of the original clinical information.

Moreover, AI can integrate structured and unstructured data to generate actionable insights. An AI algorithm—an ensemble model combining latent Dirichlet allocation with gradient-boosted trees—was developed to predict and diagnose sepsis [41]. By leveraging structured data and unstructured clinical notes, the AI model achieved high predictive accuracy (area under the receiver-operating characteristic curve of 0.94, sensitivity of 0.87, and specificity of 0.87) up to 12 hours before the onset of sepsis. This model demonstrated the potential to increase early detection of sepsis by up to 32% and reduce false positives by up to 17% compared with physician predictions.

In addition to its clinical applications, AI-powered tools can streamline administrative tasks in health care, such as scheduling appointments, managing billing, and processing insurance claims [42-44]. Improved interoperability between different administrative systems leads to more efficient health care operations, reducing the administrative burden, and allowing health care professionals to focus on patient care rather than paperwork.

AI in Patient Care Pathway Optimization and Monitoring

AI has been increasingly applied to optimize patient care pathways by improving data integration, standardization, and predictive modeling. These AI-driven solutions facilitate better coordination among health care professionals, reduce delays in treatment, and enhance overall patient outcomes. One notable example is the implementation of a learning health system by The Ottawa Hospital, which connected data across silos to redesign regional lung cancer diagnostic pathways [23]. By integrating 12 major processes—including referral, review, diagnostics, assessment, triage, and consultation—the hospital significantly streamlined the patient care journey resulting in a reduction in the median time from referral to initial treatment by 48%, dropping from 92 to 47 days. This demonstrates how data integration and process optimization can substantially improve the efficiency of patient care. Enhancing data quality and transforming unstructured data into actionable insights are also helpful for effective care pathway management. In a study by Levine et al [45] who developed NLP annotators to transform unstructured EHR text into structured attributes for patients with stage III breast cancer, the researchers enabled a clearer understanding of the clinical course by compiling events in the patients' journeys to create timelines. The NLP system showed high agreement with manual chart reviews, differing in only 6 out of 171 data elements, thus validating its accuracy and utility in clinical settings.

Building upon data integration and standardization, AI can automate the generation of care pathways aligned with clinical guidelines. In pediatric oncology, one study demonstrated how knowledge engineering methods alongside AI planning and scheduling techniques could automatically generate care pathways from computer-interpretable clinical practice guidelines [46]. This approach coordinated the activities of health care professionals involved in patient treatment, adapting care plans based on patients' current health conditions and deviations from expected progress. The system effectively handled exceptions, producing suitable care plans in 71.88% (23 out of 32) of the executed care pathways during simulation. Advancing further, AI models can predict patient health trajectories and outcomes, enabling proactive care planning. The development of a long short-term memory–based tool called DeepCare exemplifies this capability [47]. DeepCare represents care episodes as vectors and models patient health state trajectories by aggregating historical and current health data through multiscale temporal pooling. This information is used in a deep dynamic neural network that estimates future outcomes, handles irregularly timed events, and models medical interventions. When applied to datasets related to diabetes and mental health, DeepCare demonstrated improved prediction accuracy in outcomes, showcasing its potential in predictive analytics. Additionally, simulation models can aid in understanding complex patient care pathways and the potential outcomes of interventions. Researchers created a digital simulation to model the clinical pathways of critically ill patients by transforming causal and associative relationships into structured expert rules [48]. Depicted in a directed acyclic graph format and stored in a graph database (Neo4j), these rules drove a simulation app enabling users to simulate patients' state trajectories over time and test the effect of different interventions on outcomes, facilitating decision-making in critical care settings.

Collectively, these examples illustrate the great potentiality of AI in patient care pathway optimization and monitoring. By automating the generation of care pathways, adapting to patients'

XSL•FO
RenderX

changing health statuses, and predicting future outcomes, AI can bridge and integrate information from various health information systems. This integration facilitates the creation of a common language easily understood by the stakeholders involved. Moreover, AI enables the prediction of steps in the patient journey and the simulation of care pathways with different parameters, identifying the most effective ones to improve patient outcomes. However, the development and implementation of such AI technologies are still in their early stages, indicating a need for continued research and refinement in this area.

AI in Enhancing Information Retrieval and Care Transition in Patient Care Pathways

In modern health care, despite the prevalence of digital technologies, stakeholders—including health care professionals and patients—face challenges in effectively retrieving key health information to improve responsiveness along a patient-centered care journey [49-53]. The heavy workload associated with managing vast amounts of heterogeneous information can lead to gaps or inaccuracies in understanding patient health conditions complicating clinical decision-making and affecting patient outcomes. AI applications that successfully retrieve, code, and summarize key health information along a patient care journey to make predictions or decisions demonstrate the transformative potential of these technologies.

Care transitions are particularly vulnerable phases where inconsistencies in handoff records can lead to adverse events. AI has the potential to automate and verify handoff records, such as medication lists across care transitions, ensuring accuracy and continuity of care [54]. Inefficient care summaries can also create gaps in care, especially during transitions between different services. AI-generated summaries that automatically highlight critical care points can improve the effectiveness of these transitions [55,56]. A recent simulation study assessed the effectiveness of NLP combined with a support vector machine to automatically code handoff communication content and behaviors during transitions from the operating room to the intensive care unit [57]. The coded communication content included patient information, anesthesia details, surgical data, nursing information, equipment and technology used, patient transfer, and the professional environment. The communication behaviors analyzed encompassed information giving, seeking, and verification, as well as assessment, planning, and decision-making. The support vector machine model achieved coding accuracy scores of 74.2% for communication content and 54.8% for communication behaviors. A qualitative study involving semistructured interviews with 11 nurses explored the potential uses of AI in perioperative nursing handoffs, which heavily rely on structured checklists [58]. The findings suggest that AI highlighted critical care points and identified issues that could be incorporated into handoff discussions and team communications. AI-estimated elevated risks might prompt patient re-evaluation, and AI-highlighted important data could add value to nursing assessments.

The AI-Pathway Companion is another example of the potential for using AI to manage vast amounts of data, including radiology, pathology, genetics, and laboratory results from hospital information systems and picture archiving and communication systems [59]. The AI-Pathway Companion Prostate Cancer VA10B, approved for use in Europe as a medical device, leverages NLP to match the data available for an individual patient with prostate cancer guidelines from the European Association of Urology and the National Comprehensive Cancer Network. It identifies the recommended treatment approach that suits the patient's current treatment status [59]. The companion automatically shows the patient's location in the pathway and recommends the next steps, including any missing information required. The implementation of this software significantly reduced consultation preparation times in prostate cancer management and effectively improved the decision-making process and customer satisfaction. Another example is an e-pathway embedded into the EHR [60], such as a Head and Neck Oncology e-pathway that was developed to categorize by tumor localization and tumor stage. For every phase of the care process, agreements among involved health care professionals from different specialties were made regarding when and who should perform specific tasks, record specific information, determine required orders for providing optimal care, and schedule when these orders should be placed. The e-pathway aimed to automatically capture all relevant discrete data during the care process, eliminating the need for manual data collection. In this project, the team used smart phrases, autotexts, and content-importing technology to automate and standardize documentation of every phase in the pathway, including the initial oncology consultation, multidisciplinary tumor board meeting, diagnostic result consultation, follow-up consultation, and prefilled, standardized order sets. The study showed that total EHR time in initial oncology consultations was significantly reduced by 3.7 minutes—a 27% decrease compared with before using the e-pathways.

AI in Humanizing Health Care by Integrating Patient-Reported Outcome Measures

The integration of patient-reported outcome measures (PROMs) into AI-enabled processes can further help AI to personalize and potentially humanize health care [61-63]. AI algorithms are increasingly capable of processing PROM data in real-time and suggesting possible adjustments to care plans. In a review of 152 trials conducted before September 2022, it was shown that AI-enabled tools, including machine learning, reinforcement learning, neural networks, and NLP incorporating PROMs were frequently evaluated in areas like musculoskeletal health, mental health, and metabolic conditions such as diabetes [64]. Notably, 24 of these trials specifically evaluated AI models that used PROMs as input, highlighting the growing trend of integrating patient perspectives and personal outcome desires into AI-driven health care. In one significant study, researchers developed a shared decision-making tool using machine learning to generate personalized predictions of risks and benefits for total joint replacement surgery based on PROMs and clinical data [65]. This tool was evaluated in a randomized controlled trial, where its effect on decision quality and patient outcomes was assessed. Similarly, in the field of rehabilitation, machine learning models relying on commonly used algorithms like linear regression,

XSL•FO
RenderX

Random Forest regressor, Extra Trees regressor, Linear Support Vector Regression, and Kernel Ridge with a polynomial kernel were trained on clinical and PROM data from over a thousand patients. This model successfully predicted the likelihood of rehabilitation success, allowing clinicians to tailor treatment plans from the start of therapy [66].

While AI-enabled smart devices, clinical decision support systems, and chatbots have shown potential in retrieving, summarizing, coding, and analyzing clinical and PROM data to personalize and enhance treatment effectiveness [62,63], their adoption is not without challenges. These include fragmented data collection processes, the scarcity of comprehensive clinical and PROM datasets, and the rigorous monitoring and validation requirements for AI systems. Moreover, many studies integrating clinical, PROMs, and AI often exhibit suboptimal design [62,67].

AI in Optimizing Clinical Workflows and Resource Allocation Along the Patient Care Journey

There is a growing demand for AI models that can enhance clinical workflow and resource allocation across various stages of the patient care journey—such as registration, triage, diagnosis, and treatment. A qualitative study conducted in 2 German hospitals confirmed the potential of AI-based processes, specifically AI-driven text recognition and monitoring, to alleviate overcrowding in emergency departments [68]. By automating data entry and monitoring patient flow, AI can manage patient influx more effectively, reduce wait times, and improve patient satisfaction. Another study developed machine learning models to automate scalp electroencephalogram seizure tracking, supporting a digital care pathway for epilepsy [69]. The results showed that automating the diagnostic process within the digital care pathway could significantly reduce the time needed to diagnose epilepsy. This acceleration in diagnosis not only improves patient outcomes but also optimizes the use of health care resources by reducing the burden on medical professionals.

Recent advancements in AI, such as the release of ChatGPT-4, demonstrate its potential to optimize clinical workflows and reduce clinician burden. For example, integrating large language models (LLMs) into clinical workflows to draft responses to patient inbox messages has been shown to improve clinician well-being [70]. Additionally, in one study focusing on OpenAI's ChatGPT models, it was shown that AI assistance not only reduced physician workload but also enhanced the consistency, informativeness, and educational value of responses to patient messages [71].

These benefits highlight AI's potential to transcend beyond traditional tools, such as dashboards, by adding an adaptive, real-time optimization layer. Unlike static dashboards that visualize a patient's journey through the health care system, LLM-driven AI can continuously refine and update patient journey maps, ensuring that health care providers have the most accurate and timely information. This ability to actively process and synthesize complex information sets AI apart, making it an asset for improving clinical workflows and patient outcomes.

Another significant challenge in clinical workflow is the complexity of accessing and managing patient care-related digital information, which contributes to EHR workload [30]. LLMs have the potential to proactively deliver the required information to the correct clinicians at the right time, eliminating the need for extensive searches or queries [72]. By providing customized information displays tailored to the needs of each stakeholder group, AI has the potential to enhance efficiency and improve user satisfaction. For example, AI-generated tools could include context-aware messaging systems that automatically link discussions to relevant patient data, or intelligent summarization algorithms that quickly provide overviews of patient status changes [73]. This could be particularly useful in environments like intensive care units, where real-time monitoring and customized insights can significantly affect strategic planning and patient outcomes.

In the realm of medical imaging where many Food and Drug Administration (FDA)–approved AI technologies exist [74], AI, particularly deep learning, is already assuming a pivotal role in optimizing clinical workflows. The advanced computational capabilities of deep learning enable it to analyze complex features in imaging data, thereby improving image acquisition, providing real-time assessments of image quality, and aiding in objective diagnoses [75]. For instance, convolutional neural network-powered ultrasound systems can evaluate multimodal data, guide sonographers, and offer objective qualifications and diagnoses [76]. This not only supports clinical decision-making but also enhances workflow efficiency and reduces costs by minimizing dependency on operator expertise. A practical example is the health care enterprise radiology AI workflow for imaging, which streamlines the process from imaging examinations to report generation, significantly reducing radiologists' workload [77]. The deployment of an AI system with integrated components—ranging from image delivery and quality control to error correction and performance monitoring—demonstrates how AI can improve the efficiency of radiology workflows.

AI in Optimizing Digital Tool Usability and User Experience Workflows

Beyond clinical workflows, AI's applications extend to improving user experience (UX) in digital health tools. AI-based tools that monitor user behavior, such as mouse tracking and click models, can infer a user's needs and optimize interface designs accordingly. For instance, a deep neural network model called the "click model" was developed to simulate user interactions within a mobile app, leading to an optimized design that improved efficiency, reduced errors, and increased user satisfaction [78].

These AI-driven enhancements in UX are directly linked to the broader context of health care operations, where digital tools play a role in coordinating asynchronous and synchronous communication along a patient's care journey. By monitoring the use of these tools and identifying inefficiencies, AI can enable health care providers to optimize prompt use, address barriers to information exchange, and mitigate potential clinician burdens and adverse events [79]. Studies have shown that workload, as measured by EHR audit logs, is linked to physician

XSL•FO
RenderX

burnout and responsiveness to patient inquiries, underscoring how AI-driven insights can identify these issues and address them [80]. AI's ability to analyze large-scale, granular, and continuously updated data allows it to identify workflow bottlenecks and inefficiencies that might be overlooked in traditional usability studies [81-83]. For example, AI can identify features within the EHR interface that are frequently accessed but inefficiently positioned, suggesting redesigns that reduce the number of clicks and screen transitions required for routine tasks [84,85]. While traditional usability studies can highlight areas for improvement based on observed user behavior, AI can go further by tracking every interaction across thousands of user sessions, revealing inefficiencies that may occur in specific contexts or evolve over time.

Furthermore, AI has the potential to uncover non-intuitive patterns that might not be apparent even to experienced usability experts. For example, an AI might discover that certain combinations of actions consistently lead to longer task completion times, identifying areas for redesign that were not previously considered [86].

Last, the growing interest in applying AI to support UX workflows highlights AI's versatility in health care. UX design heavily depends on human creativity and input, aiming to meet end user requirements and expectations. It is a user-centered, iterative process where piloting, testing, and refining are key elements. AI offers new opportunities for human-AI collaboration in creative tasks, enabling cooperation in different phases of the UX design process. Various industry surveys indicate that UX professionals view AI-driven features in design tools positively, seeing AI as a creative partner that can iterate and innovate alongside human designers [87]. A critical step for AI is automating the transformation of a mock-up of a graphical user interface (GUI) into code. Such automation can facilitate the creation of adaptive interfaces that dynamically evolve based on changing user requirements. One study developed a machine learning-based tool called ReDraw [88], which integrates software repository mining, automated dynamic analysis, deep convolutional neural networks, and k-nearest neighbors to automate the transformation of GUI mock-ups into code. This enables accurate prototyping of GUIs. Interviews with industry practitioners illustrate ReDraw's potential to improve real-world UX workflows.

While most AI-enabled systems supporting UX work focus mainly on GUI elements, design activities that involve more "design thinking," such as user interviews and user testing, are particularly helpful for designers. LLMs have shown remarkable capabilities in creative tasks like idea generation, providing exceptionally human-like responses. A study [89] examined human-LLM interaction in UX designers' practices of incorporating ChatGPT into their professional design activities from a communication perspective, demonstrating the potential of ChatGPT in creating more productive UX tools. Another study [90] showcased a proof-of-concept implementation where users submit textual descriptions of their desired website features, prompting the LLM to produce corresponding HTML and CSS code and even provide recommendations for UX design. This approach allows users to update specific parts of the code or add new components generated by the LLM, rather than rewriting the entire document and including unnecessary code revisions, thereby improving the creativity and efficiency of UX design.

AI-human collaboration in enhanced user interface design has the potential to improve operational efficiency and reduce the learning curve for stakeholders, including health care professionals and patients.

AI-Based Tools: From Evidence to Practice

Transitioning from evidence to practical application involves multiple steps. Before fully implementing new tools or features developed through AI and big data analyses, it is essential to conduct pilot and clinical testing in controlled environments. This preliminary phase helps to evaluate the effect of these innovations on coordination efficiency and clinician workload.

Engaging the stakeholders—not just clinicians but also patients, IT professionals, and administrative staff—in the redesign process is crucial to ensure that the changes effectively address real-world needs and increase the likelihood of their adoption [36,91]. For example, AI tools designed to summarize and maintain key information along a patient care journey need to be intuitive and responsive. The stakeholder can be involved in shaping these tools to include various anticipated features, such as alerts for updates to critical information, ensuring that these enhancements are useful and user-friendly.

Establishing mechanisms to continuously collect feedback from all stakeholders and perform iterative updates based on actual use is vital. An ongoing feedback loop will allow for the refinement of AI tools, so they better meet the evolving needs of health care professionals and patients. By adapting AI tools based on direct input from end users, health care systems can optimize the functionality and influence of these technologies, ensuring they effectively support clinical decision-making and patient care.

Limitations and Considerations in the Design of AI-Based Tools

One of the significant limitations of AI-based tools is the potential for inherent biases, which can arise from the data on which these systems are trained. If the training data are not representative of diverse patient populations, or the specific population they will be applied to, AI algorithms can develop biases that lead to unequal treatment outcomes. These biases could potentially affect diagnostic accuracy, treatment recommendations, and patient management, disproportionately influencing certain demographic groups and leading to inequitable care [92-95].

While AI has already demonstrated its potential to enhance accuracy in some areas of health care, such as medical imaging, addressing the challenges in other fields—like mental health—requires ongoing research and development. Their accuracy remains a critical concern, particularly for multimodal data, where a lack of comprehensive frameworks can affect the performance, reproducibility, and transparency of AI algorithms.

XSL•FO
RenderX

Continuous monitoring and updating of the algorithms are essential to maintaining accuracy as new data become available. By focusing on improving data quality, model transparency, and rigorous validation, AI can become a more trustworthy and effective tool across all areas of health care [94].

As AI systems rely on large volumes of data to function optimally, this dependence raises additional concerns about privacy and security, particularly regarding sensitive patient information [96]. Ensuring compliance with regulations like HIPAA in the United States or the emerging AI Act in the European Union will be critical. AI tools must be designed with robust security measures to protect patient data and prevent unauthorized access.

Conclusions

Addressing the challenges in managing information flow—including data generation, updating, passing, and exchange along the patient care pathway—is critical to enhancing digital coordination. AI offers promising solutions by boosting the efficiency of the digital information ecosystem, improving interoperability and clinical workflow, optimizing patient care pathways, enhancing digital tool usability and UX workflows, reducing clinician burden, and ensuring privacy compliance. Moreover, initial practical implementations in organizations equipped to effectively implement and maintain these technologies are showing significant promise. As more health care systems begin to incorporate these technologies into their operational and clinical workflows, a richer picture will emerge, providing clearer demonstrations of their effectiveness. We anticipate that this growing body of evidence will inform continuous improvements in health care delivery systems. As such, we advocate for strategic investments in these technologies, along with ongoing evaluation and adaptation, to advance digital collaboration in health care and ultimately enhance the quality and outcomes of patient care.

Acknowledgments

This research was supported, in part, by the National Human Genome Research Institute of the National Institutes of Health under award U54HG012510, the National Library of Medicine of the National Institutes of Health under award R01LM012854, and the National Institutes of Health's Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity under award OT2OD032581. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

The idea for this paper was initiated by YC, who also carried out the literature review and composed and drafted the manuscript. Feedback on the organization, content flow, and language was provided by CUL and BM, and the manuscript underwent critical revisions for important intellectual content with their input. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Institute of Medicine (US) Committee on Identifying Priority Areas for Quality Improvement. In: Adams K, Corrigan JM, editors. Priority Areas for National Action: Transforming Health Care Quality. Washington, DC: The National Academies Press; 2003.
2. McDonald KM, Sundaram V, Bravata DM, Lewis R, Lin N, Kraft SA, et al. Closing the quality gap: a critical analysis of quality improvement strategies (Vol. 7: Care Coordination). AHRQ Tech Rev 2007:441. [Medline: 20734531]
3. Rigby M, Georgiou A, Hyppönen H, Ammenwerth E, de Keizer N, Magrabi F, et al. Patient portals as a means of information and communication technology support to patient- centric care coordination - the missing evidence and the challenges of evaluation. a joint contribution of IMIA WG EVAL and EFMI WG EVAL. Yearb Med Inform 2015;10(1):148-159 [FREE Full text] [doi: 10.15265/IY-2015-007] [Medline: 26123909]
4. Lupton D. Digital health technologies and digital data: new ways of monitoring, measuring and commodifying human bodies. In: In Research Handbook on Digital Transformations. Cheltenham, UK: Edward Elgar Publishing; 2016:85-102.
5. Fennelly O, Moroney D, Doyle M, Eustace-Cook J, Hughes M. Key interoperability factors for patient portals and electronic health records: a scoping review. Int J Med Inform 2024;183:105335. [doi: 10.1016/j.ijmedinf.2023.105335] [Medline: 38266425]
6. Burton LC, Anderson GF, Kues IW. Using electronic health records to help coordinate care. Milbank Q 2004;82(3):457-481, table of contents [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00318.x] [Medline: 15330973]
7. Graetz I, Harding B. EHR use and care coordination. Agency for Healthcare Research and Quality. 2012. URL: https://digital.ahrq.gov/sites/default/files/docs/publication/r36hs021082-graetz-final-report-2012.pdf [accessed 2024-09-09]
8. Chunchu K, Mauksch L, Charles C, Ross V, Pauwels J. A patient centered care plan in the EHR: improving collaboration and engagement. Fam Syst Health 2012;30(3):199-209. [doi: 10.1037/a0029100] [Medline: 22866953]

XSL•FO
RenderX

9. Osborn CY, Mayberry LS, Wallston KA, Johnson KB, Elasy TA. Understanding patient portal use: implications for medication management. J Med Internet Res 2013;15(7):e133 [FREE Full text] [doi: 10.2196/jmir.2589] [Medline: 23823974]
10. Short-Russell M, Thompson J, Waldrop J. Secure messaging: demonstration and enrollment patient portal program: patient portal use in vulnerable populations. Comput Inform Nurs 2024;42(2):104-108. [doi: 10.1097/CIN.0000000000001098] [Medline: 38206326]
11. Rule A, Chiang M, Hribar M. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. J Am Med Inform Assoc 2020;27(3):480-490 [FREE Full text] [doi: 10.1093/jamia/ocz196] [Medline: 31750912]
12. Dymek C, Kim B, Melton GB, Payne TH, Singh H, Hsiao CJ. Building the evidence-base to reduce electronic health record-related clinician burden. J Am Med Inform Assoc 2021;28(5):1057-1061 [FREE Full text] [doi: 10.1093/jamia/ocaa238] [Medline: 33340326]
13. Mazur LM, Mosaly PR, Moore C, Marks L. Association of the usability of electronic health records with cognitive workload and performance levels among physicians. JAMA Netw Open 2019;2(4):e191709 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.1709] [Medline: 30951160]
14. Ranade-Kharkar P, Narus SP, Anderson GL, Conway T, Del Fiol G. Data standards for interoperability of care team information to support care coordination of complex pediatric patients. J Biomed Inform 2018;85:1-9 [FREE Full text] [doi: 10.1016/j.jbi.2018.07.009] [Medline: 30017975]
15. Imler T, Vreeman D, Kannry J. Healthcare data standards and exchange. In: Clinical Informatics Study Guide: Text and Review. Cham, Switzerland: Springer; 2016:233-253.
16. Adler-Milstein J, Garg A, Zhao W, Patel V. A survey of health information exchange organizations in advance of a nationwide connectivity framework. Health Aff (Millwood) 2021;40(5):736-744. [doi: 10.1377/hlthaff.2020.01497] [Medline: 33939510]
17. McCormack J, Lehmann C, Lusk K, Ayers Wickenhauser K, Arvisais-Anhalt S, Feterik K. Direct secure messaging in practice: addressing workflow challenges. Stud Health Technol Inform 2024;310:189-193. [doi: 10.3233/SHTI230953] [Medline: 38269791]
18. Tarabichi Y, Frees A, Honeywell S, Huang C, Naidech AM, Moore JH, et al. The Cosmos collaborative: a vendor-facilitated electronic health record data aggregation platform. ACI Open 2021;5(1):e36-e46 [FREE Full text] [doi: 10.1055/s-0041-1731004] [Medline: 35071993]
19. Adler-Milstein J, Lin SC, Jha AK. The number of health information exchange efforts is declining, leaving the viability of broad clinical data exchange uncertain. Health Aff (Millwood) 2016;35(7):1278-1285. [doi: 10.1377/hlthaff.2015.1439] [Medline: 27385245]
20. Holmgren AJ, Adler-Milstein J. Health information exchange in US hospitals: the current landscape and a path to improved information sharing. J Hosp Med 2017;12(3):193-198. [doi: 10.12788/jhm.2704] [Medline: 28272599]
21. Verstraeten LMG, van Wijngaarden JP, Kim DY, Meskers CGM, Maier AB. Feasibility of bioelectrical impedance analysis in routine clinical care to assess body composition in geriatric rehabilitation inpatients: RESORT. Aging Clin Exp Res 2023;35(2):293-302. [doi: 10.1007/s40520-022-02320-8] [Medline: 36609845]
22. Meyer MA. Mapping the patient journey across the continuum: lessons learned from one patient's experience. J Patient Exp 2019;6(2):103-107 [FREE Full text] [doi: 10.1177/2374373518783763] [Medline: 31218254]
23. Fung-Kee-Fung M, Maziak D, Pantarotto J, Smylie J, Taylor L, Timlin T, et al. Regional process redesign of lung cancer care: a learning health system pilot project. Curr Oncol 2018;25(1):59-66 [FREE Full text] [doi: 10.3747/co.25.3719] [Medline: 29507485]
24. Halvorsrud R, Kvale K, Følstad A. Improving service quality through customer journey analysis. JSTP 2016;26(6):840-867. [doi: 10.1108/jstp-05-2015-0111]
25. Kempa-Liehr AW, Lin CY, Britten R, Armstrong D, Wallace J, Mordaunt D, et al. Healthcare pathway discovery and probabilistic machine learning. Int J Med Inform 2020;137:104087. [doi: 10.1016/j.ijmedinf.2020.104087] [Medline: 32126509]
26. van der Aalst W, Adriansyah A, van Dongen B. Replaying history on process models for conformance checking and performance analysis. WIREs Data Min Knowl 2012;2(2):182-192. [doi: 10.1002/widm.1045]
27. van der Aalst W. Process mining: overview and opportunities. ACM Trans Manage Inf Syst 2012;3(2):1-17. [doi: 10.1145/2229156.2229157]
28. Rinderle-Ma S, Stertz F, Mangler J, Pauker F. Process mining—discovery, conformance, and enhancement of manufacturing processes. In: Digital Transformation: Core Technologies and Emerging Topics from a Computer Science Perspective. Berlin, Germany: Springer Vieweg; 2023:363-383.
29. Askari M, Westerhof R, Eslami S, Medlock S, de Rooij SE, Abu-Hanna A. A combined disease management and process modeling approach for assessing and improving care processes: a fall management case-study. Int J Med Inform 2013;82(10):1022-1033. [doi: 10.1016/j.ijmedinf.2013.06.011] [Medline: 23891085]
30. Amano A, Brown-Johnson CG, Winget M, Sinha A, Shah S, Sinsky CA, et al. Perspectives on the intersection of electronic health records and health care team communication, function, and well-being. JAMA Netw Open 2023;6(5):e2313178 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.13178] [Medline: 37171816]

XSL•FO
RenderX

31. Melnick ER, Dyrbye LN, Sinsky CA, Trockel M, West CP, Nedelec L, et al. The association between perceived electronic health record usability and professional burnout among US physicians. Mayo Clin Proc 2020;95(3):476-487 [FREE Full text] [doi: 10.1016/j.mayocp.2019.09.024] [Medline: 31735343]
32. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
33. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021;23(8):e26162. [doi: 10.2196/26162] [Medline: 34236994]
34. Abdelhamid M, Gaia J, Sanders GL. Putting the focus back on the patient: how privacy concerns affect personal health information sharing intentions. J Med Internet Res 2017;19(9):e169 [FREE Full text] [doi: 10.2196/jmir.6877] [Medline: 28903895]
35. Tangi L, Combetto M, Martin BJ, Rodriguez MP, Joint Research Centre (Seville site). Artificial intelligence for interoperability in the European public sector. European Commission. 2023. URL: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC134713/JRC134713_01.pdf [accessed 2024-09-09]
36. Lopez-Perez L, Gaeta E, Alonso I, Esteban A, Dominguez VG, Mercalli F, et al. Clinical natural language processing and health interoperability to support knowledge management and governance in rare cancers. 2023 Presented at: Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON) and International Conference on Medical and Biological Engineering (CMBEBIH); September 14-16, 2023; Sarajevo, Bosnia and Herzegovina. [doi: 10.1007/978-3-031-49062-0_58]
37. Gandhi TK, Classen D, Sinsky CA, Rhew DC, Vande Garde N, Roberts A, et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? JAMIA Open 2023;6(3):ooad079 [FREE Full text] [doi: 10.1093/jamiaopen/ooad079] [Medline: 37655124]
38. Razzaque A, Hamdan A. Artificial intelligence based multinational corporate model for EHR interoperability on an e-health platform. In: Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications. Cham, Switzerland: Springer; 2021:71-81.
39. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med 2023;155:106649. [doi: 10.1016/j.compbiomed.2023.106649] [Medline: 36805219]
40. Giannaris PS, Al-Taie Z, Kovalenko M, Thanintorn N, Kholod O, Innokenteva Y, et al. Artificial intelligence-driven structurization of diagnostic information in free-text pathology reports. J Pathol Inform 2020;11:4 [FREE Full text] [doi: 10.4103/jpi.jpi_30_19] [Medline: 32166042]
41. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021;12(1):711. [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
42. Zheng Y, Wang L, Feng B, Zhao A, Wu Y. Innovating healthcare: the role of ChatGPT in streamlining hospital workflow in the future. Ann Biomed Eng 2024;52(4):750-753. [doi: 10.1007/s10439-023-03323-w] [Medline: 37464178]
43. Knight DR, Aakre CA, Anstine CV, Munipalli B, Biazar P, Mitri G, et al. Artificial intelligence for patient scheduling in the real-world health care setting: a metanarrative review. Health Policy Technol 2023;12(4):100824. [doi: 10.1016/j.hlpt.2023.100824]
44. Zhang W, Shi J, Wang X, Wynn H. AI-powered decision-making in facilitating insurance claim dispute resolution. Ann Oper Res 2023:1-30. [doi: 10.1007/s10479-023-05631-9]
45. Levine MN, Alexander G, Sathiyapalan A, Agrawal A, Pond G. Learning health system for breast cancer: pilot project experience. JCO Clin Cancer Inform 2019;3:1-11 [FREE Full text] [doi: 10.1200/CCI.19.00032] [Medline: 31369338]
46. Sánchez-Garzón I, González-Ferrer A, Fernández-Olivares J. A knowledge-based architecture for the management of patient-focused care pathways. Appl Intell 2013;40(3):497-524. [doi: 10.1007/s10489-013-0466-0]
47. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: a deep learning approach. J Biomed Inform 2017;69:218-229 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.001] [Medline: 28410981]
48. Trevena W, Lal A, Zec S, Cubro E, Zhong X, Dong Y, et al. Modeling of critically ill patient pathways to support intensive care delivery. IEEE Robot Autom Lett 2022;7(3):7287-7294. [doi: 10.1109/lra.2022.3183253]
49. Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. JAMA Netw Open 2020;3(1):e1919396 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.19396] [Medline: 31951272]
50. Bradwell K, Wooldridge J, Amor B, Bennett T, Anand A, Bremer C, et al. Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (EHR) dataset. J Am Med Inform Assoc 2022;29(7):1172-1182 [FREE Full text] [doi: 10.1093/jamia/ocac054] [Medline: 35435957]
51. Song Y, Wu P, Hu C, Zhang K, Dai D, Chang H, et al. Research on the structuring of electronic medical records based on joint extraction using BART. 2023 Presented at: In China Health Information Processing Conference; Singapore; Springer Nature Singapore. [doi: 10.1007/978-981-99-9864-7_14]

XSL•FO
RenderX

52. Ge J, Li M, Delk MB, Lai JC. A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. Gastroenterology 2024;166(4):707-709.e3. [doi: 10.1053/j.gastro.2023.12.019] [Medline: 38151192]
53. Huang J, Yang D, Chi Z, Rong R, Wang S, Nezafati K, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. SSRN 2023:1-21. [doi: 10.2139/ssrn.4488945]
54. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy 2018;38(8):822-841. [doi: 10.1002/phar.2151] [Medline: 29884988]
55. Davis R, Eppler M, Ayo-Ajibola O, Loh-Doyle JC, Nabhani J, Samplaski M, et al. Evaluating the effectiveness of artificial intelligence-powered large language models application in disseminating appropriate and readable health information in urology. J Urol 2023;210(4):688-694. [doi: 10.1097/JU.0000000000003615] [Medline: 37428117]
56. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
57. Mayes E, Gehlbach JA, Jeziorczak PM, Wooldridge AR. Machine learning to operationalize team cognition: a case study of patient handoffs. Hum Factors Healthcare 2023;3:100036. [doi: 10.1016/j.hfh.2023.100036]
58. King C, Shambe A, Abraham J. Potential uses of AI for perioperative nursing handoffs: a qualitative study. JAMIA Open 2023;6(1):ooad015 [FREE Full text] [doi: 10.1093/jamiaopen/ooad015] [Medline: 36935899]
59. Henkel M, Horn T, Leboutte F, Trotsenko P, Dugas SG, Sutter SU, et al. Initial experience with AI pathway companion: evaluation of dashboard-enhanced clinical decision making in prostate cancer screening. PLoS One 2022;17(7):e0271183 [FREE Full text] [doi: 10.1371/journal.pone.0271183] [Medline: 35857753]
60. Ebbers T, Takes RP, Smeele LE, Kool RB, van den Broek GB, Dirven R. The implementation of a multidisciplinary, electronic health record embedded care pathway to improve structured data recording and decrease electronic health record burden. Int J Med Inform 2024;184:105344 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105344] [Medline: 38310755]
61. Allam S, Nadikattu A. AI economical wearable smart device to alert real time health reports to doctors. Int J Creat Res Thoughts 2020:2320-2882 [FREE Full text]
62. Cruz Rivera S, Liu X, Hughes SE, Dunster H, Manna E, Denniston AK, et al. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. Lancet Digit Health 2023;5(3):e168-e173 [FREE Full text] [doi: 10.1016/S2589-7500(22)00252-7] [Medline: 36828609]
63. Hassan AM, Biaggi-Ondina A, Rajesh A, Asaad M, Nelson JA, Coert JH, et al. Predicting patient-reported outcomes following surgery using machine learning. Am Surg 2023;89(1):31-35 [FREE Full text] [doi: 10.1177/00031348221109478] [Medline: 35722685]
64. Pearce FJ, Cruz Rivera S, Liu X, Manna E, Denniston AK, Calvert MJ. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997-2022). Lancet Digit Health 2023;5(3):e160-e167 [FREE Full text] [doi: 10.1016/S2589-7500(22)00249-7] [Medline: 36828608]
65. Jayakumar P, Bozic KJ. Advanced decision-making using patient-reported outcome measures in total joint replacement. J Orthop Res 2020;38(7):1414-1422 [FREE Full text] [doi: 10.1002/jor.24614] [Medline: 31994752]
66. Tschuggnall M, Grote V, Pirchl M, Holzner B, Rumpold G, Fischer MJ. Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. Inform Med Unlocked 2021;24:100598. [doi: 10.1016/j.imu.2021.100598]
67. Capriulo M, Pizzolla I, Briganti G. On the use of patient-reported measures in digital medicine to increase healthcare resilience. In: Ordóñez de Pablos P, Zhang X, editors. Artificial Intelligence, Big Data, Blockchain and 5G for the Digital Transformation of the Healthcare Industry. Cambridge, MA: Academic Press; 2024:41-66.
68. Born C, Wildmoser J, Schwarz R, Böttcher T, Hein A, Krcmar H. Identifying potentials for artificial intelligence-based process support along the emergency department care pathway to alleviate overcrowding. Procedia Comput Sci 2024;239:1705-1712. [doi: 10.1016/j.procs.2024.06.348]
69. Keikhosrokiani P, Isomursu M, Uusimaa J, Kortelainen J. A sustainable artificial-intelligence-augmented digital care pathway for epilepsy: automating seizure tracking based on electroencephalogram data using artificial intelligence. Digit Health 2024;10:20552076241287356 [FREE Full text] [doi: 10.1177/20552076241287356] [Medline: 39381810]
70. Garcia P, Ma SP, Shah S, Smith M, Jeong Y, Devon-Sand A, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024;7(3):e243201 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
71. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024;6(6):e379-e381 [FREE Full text] [doi: 10.1016/S2589-7500(24)00060-8] [Medline: 38664108]
72. Peng C, Yang X, Smith KE, Yu Z, Chen A, Bian J, et al. Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. J Biomed Inform 2024;153:104630 [FREE Full text] [doi: 10.1016/j.jbi.2024.104630] [Medline: 38548007]
73. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]

XSL•FO
RenderX

74. Joshi G, Jain A, Araveeti SR, Adhikari S, Garg H, Bhandari M. FDA-Approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. Electronics 2024;13(3):498. [doi: 10.3390/electronics13030498]
75. Pierre K, Haneberg AG, Kwak S, Peters KR, Hochhegger B, Sananmuang T, et al. Applications of artificial intelligence in the radiology roundtrip: process streamlining, workflow optimization, and beyond. Semin Roentgenol 2023;58(2):158-169 [FREE Full text] [doi: 10.1053/j.ro.2023.02.003] [Medline: 37087136]
76. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, et al. A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow. J Am Coll Radiol 2019;16(9 Pt B):1318-1328. [doi: 10.1016/j.jacr.2019.06.004] [Medline: 31492410]
77. Juluru K, Shih H, Keshava Murthy KN, Elnajjar P, El-Rowmeim A, Roth C, et al. Integrating Al algorithms into the clinical workflow. Radiol Artif Intell 2021;3(6):e210013 [FREE Full text] [doi: 10.1148/ryai.2021210013] [Medline: 34870216]
78. Souza KES, Seruffo MCR, de Mello HD, Souza DDS, Vellasco MMBR. User experience evaluation using mouse tracking and artificial intelligence. IEEE Access 2019;7:96506-96515. [doi: 10.1109/access.2019.2927860]
79. Yang B, Wei L, Pu Z. Measuring and improving user experience through artificial intelligence-aided design. Front Psychol 2020;11:595374 [FREE Full text] [doi: 10.3389/fpsyg.2020.595374] [Medline: 33329260]
80. Dyrbye LN, Gordon J, O'Horo J, Belford SM, Wright M, Satele DV, et al. Relationships between EHR-based audit log data and physician burnout and clinical practice process measures. Mayo Clin Proc 2023;98(3):398-409. [doi: 10.1016/j.mayocp.2022.10.027] [Medline: 36868747]
81. King AJ, Cooper GF, Hochheiser H, Clermont G, Hauskrecht M, Visweswaran S. Using machine learning to predict the information seeking behavior of clinicians using an electronic medical record system. AMIA Annu Symp Proc 2018 Dec 5;2018:673-682 [FREE Full text] [Medline: 30815109]
82. Greenbaum NR, Jernite Y, Halpern Y, Calder S, Nathanson LA, Sontag DA, et al. Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces. Int J Med Inform 2019;132:103981. [doi: 10.1016/j.ijmedinf.2019.103981] [Medline: 31605881]
83. Ratwani RM, Savage E, Will A, Arnold R, Khairat S, Miller K, et al. A usability and safety analysis of electronic health records: a multi-center study. J Am Med Inform Assoc 2018;25(9):1197-1201 [FREE Full text] [doi: 10.1093/jamia/ocy088] [Medline: 29982549]
84. McGreevey JD, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing alert burden in electronic health records: state of the art recommendations from four health systems. Appl Clin Inform 2020;11(1):1-12 [FREE Full text] [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
85. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. JAMA Netw Open 2021;4(7):e2117391 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17391] [Medline: 34297075]
86. Chen B, Alrifai W, Gao C, Jones B, Novak L, Lorenzi N, et al. Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning. J Am Med Inform Assoc 2021;28(6):1168-1177 [FREE Full text] [doi: 10.1093/jamia/ocaa338] [Medline: 33576432]
87. Knearem T, Khwaja M, Gao Y, Bentley F, Kliman-Silver C. Exploring the future of design tooling: the role of artificial intelligence in tools for user experience professionals. 2023 Presented at: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany p. 1-6. [doi: 10.1145/3544549.3573874]
88. Moran K, Bernal-Cardenas C, Curcio M, Bonett R, Poshyvanyk D. Machine learning-based prototyping of graphical user interfaces for mobile apps. IIEEE Trans. Software Eng 2020;46(2):196-221. [doi: 10.1109/tse.2018.2844788]
89. Zhou Z, Gao W, Li Y, Yu J. Developing an interaction framework for human-large language models collaboration in creative tasks: insights from Ux professionals‘ communication with Chatgpt. SSRN 2024:3257. [doi: 10.2139/ssrn.4853257]
90. Calò T, de Russis L. Leveraging large language models for end-user website generation. 2023 Presented at: 9th International Symposium, IS-EUD 2023; June 6-8, 2023; Cagliari, Italy p. 52-61. [doi: 10.1007/978-3-031-34433-6_4]
91. Zhang G, Jin Q, Jered McInerney D, Chen Y, Wang F, Cole CL, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform 2024;153:104640. [doi: 10.1016/j.jbi.2024.104640] [Medline: 38608915]
92. Gichoya J, Thomas K, Celi L, Safdar N, Banerjee I, Banja J, et al. AI pitfalls and what not to do: mitigating bias in AI. Br J Radiol 2023;96(1150):20230023 [FREE Full text] [doi: 10.1259/bjr.20230023] [Medline: 37698583]
93. Handley JL, Lehmann CU, Ratwani RM. Safe and equitable pediatric clinical use of AI. JAMA Pediatr 2024;178(7):637-638. [doi: 10.1001/jamapediatrics.2024.0897] [Medline: 38739385]
94. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. J Med Internet Res 2023;25:e43251 [FREE Full text] [doi: 10.2196/43251] [Medline: 36961506]
95. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc 2022;29(4):585-591 [FREE Full text] [doi: 10.1093/jamia/ocac006] [Medline: 35190824]
96. Golda A, Mekonen K, Pandey A, Singh A, Hassija V, Chamola V, et al. Privacy and security concerns in generative AI: a comprehensive survey. IEEE Access 2024;12:48126-48144. [doi: 10.1109/access.2024.3381611]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
EHR: electronic health record
FDA: Food and Drug Administration
GUI: graphical user interface
HIE: health information exchange
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level Seven
LLM: large language model
NLP: natural language processing
PHI: protected health information
PROM: patient-reported outcome measure
UX: user experience

Edited by A Mavragani; submitted 06.05.24; peer-reviewed by FA Causio, D Chrimes; comments to author 14.08.24; revised version received 26.08.24; accepted 28.10.24; published 02.12.24.

Please cite as:
Chen Y, Lehmann CU, Malin B
Digital Information Ecosystems in Modern Care Coordination and Patient Care Pathways and the Challenges and Opportunities for AI Solutions
J Med Internet Res 2024;26:e60258
URL: https://www.jmir.org/2024/1/e60258
doi:10.2196/60258
PMID:

©You Chen, Christoph U Lehmann, Bradley Malin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Reflections From the Pandemic: Is Connectivism the Panacea for Clinicians?

Jennifer Benjamin[1,2*], MD, MS; Tyson Pillow[3], MD, MEd; Heather MacNeill[4], MD; Ken Masters[5], PhD; Anoop Agrawal[6], MD; Neil Mehta[7*], MBBS, MS

[1]Texas Children's Hospital, Baylor College of Medicine, Houston, TX, United States
[2]Huffington Department of Education Innovation and Technology, Baylor College of Medicine, Houston, TX, United States
[3]Department of Emergency Medicine, Baylor College of Medicine, Houston, TX, United States
[4]Department of Medicine, Continuing Professional Development, Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[5]Medical Informatics, Sultan Qaboos University, Alkhod, Oman
[6]Department of Internal Medicine Pediatrics, Baylor College of Medicine, Houston, TX, United States
[7]Department of Internal Medicine, Lerner College of Medicine, Cleveland Clinic, Cleveland, OH, United States
[*]these authors contributed equally

Corresponding Author:
Jennifer Benjamin, MD, MS
Texas Children's Hospital
Baylor College of Medicine
6701 Fannin Street
Houston, TX, 77030
United States
Phone: 1 832 824 1000 ext 3436
Fax: 1 832 825 3435
Email: jennifer.benjamin@bcm.edu

Abstract

The COVID-19 pandemic and the recent increased interest in generative artificial intelligence (GenAI) highlight the need for interprofessional communities' collaboration to find solutions to complex problems. A personal narrative experience of one of the authors compels us to reflect on current approaches to learning and knowledge acquisition and use solutions to the challenges posed by GenAI through social learning contexts using connectivism. We recognize the need for constructivism and experiential learning for knowledge acquisition to establish foundational understanding. We explore how connectivist approaches can enhance traditional constructivist paradigms amid rapidly changing learning environments and online communities. Learning in connectivism includes interacting with experts from other disciplines and creating nodes of accurate and accessible information while distinguishing between misinformation and accurate facts. Autonomy, connectedness, diversity, and openness are foundational for learners to thrive in this learning environment. Learning in this environment is not just acquiring new knowledge as individuals but being connected to networks of knowledge, enabling health professionals to stay current and up-to-date. Existing online communities with accessible GenAI solutions allow for the application of connectivist principles for learning and knowledge acquisition.

(J Med Internet Res 2024;26:e53344) doi:10.2196/53344

KEYWORDS

learning theory; learning framework; connectivism; panacea; COVID-19; generative artificial intelligence; GAI; health care community; clinician; health care; airborne disease; learning; information; misinformation; autonomy; diversity

Introduction

The COVID-19 pandemic was unprecedented in exposing the importance of staying current as health professional educators, the dangers of misinformation, and the need for a new approach to learning. In our commentary, we aim to highlight these 3 different aspects.

The first case of COVID-19 in the United States was confirmed in January 2020, and stay-at-home orders were not issued till March 2020 [1]. We narrate the personal experience of one of

XSL•FO
RenderX

the authors during this time. The viruses' mode of transmission and strategies to reduce spread were every health professional's priority, and this author, a physician at a large health care system, observed that several preprint articles highlighted that transmission of infection could be potentially from airborne infection [1-5].

In March 2020, just before the COVID-19 pandemic became a reality in the United States, this author was precepting a student in an ambulatory clinic. After some discussion on the topic, the physician and the student both wore masks at the clinic. Coincidentally, another member of the team who was a member of the hospital's COVID-19 task force walked by, commenting, "You know masks don't work for COVID-19, right? You will more likely get infected by transferring droplets to your face as you repeatedly adjust the mask." Due to the information in the preprint articles, the physician and student continued masking, and the physician shared the reasoning behind masking and why several respected health care organizations and experts were advocating against it.

At this hospital, the emphasis was still on droplet precaution with hand washing, thereby aligning with health organizations' statements prioritizing the need for handwashing for droplet precautions [6,7] based on evidence-based recommendations. However, the pandemic exposed years of misunderstanding of the nature of airborne infection by the scientific community [2,8-10]. The transmission was primarily believed to be droplet and fomite-related, with airborne transmission falsely thought to occur only during aerosol-generating procedures, such as nebulizer treatments and intubations. The conventional wisdom was that particles larger than 5 microns would be transmitted as droplets rather than aerosols. This led to the assumption that most respiratory infections spread as large droplets and could be mitigated through droplet precautions.

The confusion of airborne versus droplet transmission of infection existed several decades prior with other infections, such as tuberculosis, when it was mistakenly believed that infections spread primarily as droplets, with reluctance to accept the airborne route as the cause of transmission [5]. Physicists and aerosol scientists, on the other hand, were aware of aerosol particles as large as 100 microns [11], which contradicted the 5-micron particle size distinction between aerosol and droplets [4,5,12,13]. However, the medical community was either unaware or slow to learn these facts due to the predominantly siloed approach of relying solely on previously established studies and beliefs in medical academia [3-6,12,14-17]. Health care experts in environmental sciences voiced concern about how the pandemic was being handled [3]. This physician was intrigued and, rather than dismiss the new differing opinion from environmental scientists, decided to investigate it further. As we know now, although the virus responsible for the pandemic is spread by airborne transmission, the recommendations for masking and the warning that COVID-19 was being transmitted as an aerosol were not communicated until almost 2 years later by institutions [18].

Questions to be considered at this stage are: What convinced this physician about the mode of transmission and to practice masking contrary to recommendations by their hospital or public health organizations during the initial days of the pandemic? How can a health care provider stay informed when traditional peer-reviewed literature is not yet available and recommendations from health care organizations differ from expert opinions?

As physicians, we are expected to be fully informed of the latest advances in medicine, stay connected with accurate sources of information, and, thus, make the best clinical care decisions. While the challenges posed by the pandemic may be behind us, our aim is to reflect and learn from this experience and derive solutions when there are differing opinions. Overall, the question to address is: How can one use critical appraisal to decipher signals from noise when bombarded with information?

While peer-reviewed evidence is still being generated, people often fall back on institutional guidance and policies, and one needs to recognize the need to explore differing opinions with a critical approach. With this personal experience as our background, we pivot our attention to addressing the realistic challenge of today: the role of GenAI in education. We propose the need for a different approach to learning and knowledge building to solve challenges facing health professionals with GenAI, using connectivism as a complementary approach to the tried and tested methods of constructivism. In doing this, we apply the foundational principles of learning from this personal experience to derive solutions to challenges posed by GenAI and provide concrete examples and suggestions to implement this new framework of learning.

Let us consider this physician's approach to knowledge generation.

Building a Knowledge Network

As the news broke of a serious respiratory illness in China, the physician created their personalized stream of information from firsthand observers, including journalists and reporters from reputable newspapers and websites, opinions of a diverse group of experts in infectious disease, virology, immunology, epidemiology, and environmental engineering by creating a list on Twitter (subsequently rebranded X). They checked the profile of experts in each field by their institutional affiliations and qualifications, ignoring follower counts and, instead, focusing on previous work using PubMed, Google Scholar, and organic web searches. Intentionally seeking diverse viewpoints, they expanded their network by adding experts in their field of interest or followed by their existing contacts. They also created an RSS feed of newly published preprint articles on the topic. By regularly perusing the knowledge network and the latest studies on the feed, they stayed current on the latest evidence-based knowledge and multidisciplinary expert opinions by being connected to new knowledge and information in real-time, thus overcoming traditional health care siloes. Their knowledge network expanded by adding authors of key studies from the RSS feed to the Twitter list. Their expanding knowledge allowed them to openly discuss evidence-based facts and highlight scientific knowledge about transmission by airborne spread in earlier studies [16]. In these discussion forums, they learned of additional articles that treated the first few US patients, suggesting multiple modalities of spread,

XSL•FO
RenderX

droplet, and airborne [2]. On further inquiry, they recognized a group of environmental engineers who were elaborating on the aerosolization of large particles, traveling distances larger than the recommended 6 feet for social distancing while staying suspended in the air for long periods of time [3,4,11-13].

Constructivism: Tried and Tested Approach to Learning

Traditionally, the dissemination of evidence-based knowledge is through information taught at medical schools, academic conferences, and peer-reviewed literature, including systematic reviews. Experts from other fields rarely teach in medical schools, and we seldom venture beyond health care literature for evidence-based information. There are few opportunities for the exchange of ideas between basic scientists and clinical researchers, which further delays the application of translational research to clinical practice. Medical training relies on years of education and learning from clinical experience. This process of building upon previous knowledge while constructing new knowledge is known as "constructivism." Constructivism works well for established paradigms of knowledge generation at a controlled pace. However, in settings such as the pandemic and currently with GenAI, evidence is generated rapidly and often sourced from disciplines outside health care. In the example above, it was the interdisciplinary work of aerosol scientists that helped us acknowledge airborne transmission [3,12].

Constructivism takes time and is a gradual process. As individuals, we are unable to process numerous sources of information simultaneously, and technology can help address this challenge by personalizing the information we receive. When faced with new ideas that do not align with previous knowledge, we either ignore new information or work harder to reframe our understanding. Our psychological drive for consensus suppresses disagreement, sometimes even in the face of evidence-based facts [8,19]. Thus, when exposed to information contrary to previous knowledge, we fail to recognize our limitations and misunderstandings. When inundated with a flood of new research from untraditional sources, the challenge is even greater. With the acceleration of knowledge generation, it is impossible to stay informed with advances in literature. Hence, there is a need to belong to communities that amplify accurate information. Learning to distinguish misinformation from tried and tested evidence-based facts becomes an additional role for physicians. As a scientific community, we are at risk of overlooking errors and missing crucial discoveries, especially if we fail to listen to experts from other fields. We further distinguish between constructivism and connectivism in Table 1.

Table 1. Differentiation between constructivism and connectivism.

	Constructivism	Connectivism
Learning focus	• Focuses on an individual building internal understanding	• Distributed within a network through connections between information sources
Factors	• Influenced by engagement, participation, and social and cultural factors	• Influenced by the diversity of the network, strength of ties, reliability, and accuracy in becoming "nodes"
Process of learning	• Learner builds schema based on previous understanding	• Learner builds knowledge by connecting to nodes and growing network
Knowledge	• Exists within an individual • Connecting new information to existing knowledge	• Exists outside the individual • Network's collective intelligence
Information flow	• Sequential builds from a foundational understanding of simple to complex concepts	• Nonlinear, influenced by strength and diversity of connections
Knowledge dissemination	• Source of new information: systematic reviews and conferences, slower pace of new information	• Source of new information: preprints; rapidly changing, diverse pool of resources
Technology	• Supportive, enhances learning experience	• Integral to learning for connections and to access information
Educator's role	• Experiential learning, helps learners construct knowledge	• Learner-driven, facilitates connections, guides learners to reliable sources

Why Consider a New Framework?

Public health institutions' announcements from the World Health Organization and the Centers for Disease Control and Prevention reach billions of people who trust and follow their recommendations as they are based on well-founded evidence. Until fully verified, informing the public of aerosolized spread could cause panic and chaos, leading to a shortage of personal protective equipment with a huge socioeconomic impact [10]. They rely on data from multiple well-designed studies before conclusively deciding on airborne spread. Although the gold standard, this method is not completely foolproof and has its drawbacks. Many randomized controlled trials are refuted within a few years of publication, with a considerable number not published [20,21]. There are biases leading to higher publication rates of studies with positive results. Biomedical science has been in a silo, failing to learn from other disciplines and often discarding information that is obvious to other disciplines [13].

XSL•FO
RenderX

In fact, if the need for unanimity overrides motivation and authority, it can be detrimental, with negative consequences for true learning [22], and one way to combat this as a scientific community is to test and retest a hypothesis to maintain scientific rigor [23]. Overworked clinicians, stretched for time, often rely on online resources and opinions of local experts, practicing eminence-based medicine if there is insufficient evidence-based medicine. Over time, as more evidence emerges from multiple sources, the science becomes clearer, the truth emerges, and the findings are more universally adopted by physicians in practice. This process can take several years to get from the bench to the bedside, sometimes up to 5 years in surgical specialties [20].

A New Approach to Learning: Connectivism

Connectivism was first proposed by Siemens [24] as a learning theory for a digitally connected world. In connectivism, new knowledge resides in the diversity of opinion and is distributed across networks, and learning occurs when individuals form connections across these networks [24,25]. Knowledge in a digital world where information is changing rapidly resides in several sources, experts, and networks, serving as “nodes” of information [26]. New knowledge resides outside of individuals, in digital networks and online communities [25,27]. Learning is not about acquiring new information but connecting to networks of knowledge [26]. Thus, knowing “where” and “who” instead of “what” and “how” becomes the new paradigm for information acquisition and learning. Health care experts need to be active participants in this dynamic network, serving as “nodes” that can recognize accurate information while discarding misinformation. The internet has allowed for the creation of communities and knowledge networks; learning how to navigate the diversity of opinions, differentiate reliable sources of information, and distinguish accurate facts from misinformation is, therefore, a valuable skill for future physicians. The theoretical approach to connectivism relies on the diversity of opinions, the strength of connections, the ability to recognize patterns, and the notion that the capacity to know is more important than what is already known. [25,26,28]. Coherence provides synchronization of different elements in this virtual community and provides organization and structure in the diverse networks of knowledge [29].

We consider the following implications for using connectivism as clinicians: First, the starting point of connectivism are “nodes” of accurate information; these can be individual experts, devices, or communities. Knowledge is created through interaction between the nodes and due to the diversity of opinions [26,28,30,31] by maintaining and nurturing connections and forming alliances with experts across disciplines. In these knowledge networks, Connectivism supports lifelong learning, collaboration, and critical thinking and fosters the ability to synthesize information from multiple sources [28,32]. The diversity of nodes makes it possible for networks to adopt and evolve over time [26]. The 4 foundational aspects of successful networks in connectivism are autonomy, connectedness, diversity, and openness [33].

Second, autonomy allows learners to choose their learning communities through self-directed learning in massive online open courses. Individuals have the freedom to express their ideas and explore the knowledge network to identify experts and acquire knowledge based on their level of motivation and not be restricted by a hierarchical approach to learning. Exploring the diversity of opinions allows the learner to experience the breadth of opinions. A physician with the skills to identify reliable and accurate sources of information with a diversity of opinions has a better understanding of the challenges faced by patients and can provide practical advice by directly addressing misinformation. These dialogues allow for interactivity and problem-solving [29].

Connectedness enables collaboration, and learning occurs through new connections between nodes and being connected to the right sources allows knowledge propagation in networks [29]. Maintaining, nurturing, and contributing to this network with coherence allows for continual learning, with new knowledge residing in the community rather than the individual. Connection with the right people in the right context contributes to learning. The ability to find and analyze new information quickly is crucial. With rapid changes in scientific understanding, staying current by collaborating with experts outside one’s field is essential. Currency and staying current with accurate information is the intent of connectivist learning [34].

Connectivism and the Future Implications

Learning in connectivism is based on the production of knowledge rather than the consumption of knowledge by prioritizing relevant data while filtering out irrelevant information. In the digital environment, learning is enhanced by engagement with virtual communities of practice through blogs, posts on X, podcasts, and meaningful dialogue [35]. Interactions are central in the online educational experience, and for deep learning to occur, these interactions need to be purposeful with systematic discourse [36]. With the recent interest in GenAI, we can recognize that knowledge resides outside of humans and in this increasingly connected world, solutions to problems are provided by teams from multiple disciplines. GenAI can be used to provide personalized learning through communities, fostering a greater sense of autonomy and motivation [28,37].

If academia fails to embrace a connectivist framework, it risks losing its position as an arbiter of information while becoming less relevant.

The application of connectivism helps expand our approach to constructing scientific knowledge. It can move us from traditional classrooms to virtual communities by implementing solutions that are responsive, agile, networked, and dynamic [26]. When faced with emergent challenges with the rapid growth of information from multiple disciplines, Connectivism can bolster a swift response. We envision several applications of connectivism through social learning platforms that already exist for students to solve complex problems using online open-access tutoring through self-motivation [38]. Community-driven GenAI solutions already exist connecting 5 million users to computing resources for learning with open access to the latest technologies [39,40] with the potential to offer personalized learning and cloud-based integrations of

XSL•FO
RenderX

GenAI for smart classrooms [41,42]. These social learning approaches demonstrate that connectivism directs the learner to a socially shared source of knowledge. The idea of distributed cognitive theory proposed by Hutchins [43] is relevant to connectivism in that it provides the framework and approach to lifelong and self-determined knowledge acquisition in social learning contexts

A limitation of connectivism is the overwhelming amount of information and diversity of opinions. Educators should address this by organizing the content by highlighting the common themes through coherence [29]. Another solution is teaching learners to identify the credibility and cognitive authority of web-based resources [44,45]. Teachers can tailor the lesson to the individual needs of learners, and consistent use of internet-based tools across lessons improves the learner's confidence and learning ability to use these tools [46]. We provide a practical guide for the application of connectivism in Table 2. We highlight the need for understanding before applying its principles in real-world contexts by creating feeds and recognizing sources of information. Another solution to information overload is the use of effective filters for feeds and autogenerated information. Education around using social media and digital scholarship is still being defined and will need refinement and guidance from educators, although connectivism, at its core, should be self-directed by the learner.

Table 2. Learning activities for teaching connectivism and its application.

Course objectives	Learning objectives	Learning activities	How can it be incorporated into the curriculum?
Connectivism framework	• Summarize principles of connectivism • Compare the relevance of connectivism to other frameworks for learning about GenAIa • Identify valid resources from misinformation	• Create a session on GenAI; as prework, provide the seminal paper by Siemens (Connectivism: A Learning Theory for the Digital Age) [24] • At the session, review the paper in a large group discussion • Small group discussion on how connectivism is relevant for learning about and staying up to date on GenAI	• Identify a learning community • Identify experts on social media platforms • Discuss a preprint article on GenAI [47] • Describe the cognitive authority of communities or experts
GenAI concerns, limitations, and potential benefits.	• List limitations, concerns, and benefits of GenAI • Recognize settings when it is most applicable	• Discussion board on the dangers of misinformation and how to avoid it • Small group discussion on the dangers of not exploring diverse opinions	• Provide an article on droplet transmission [9] • Provide resources on scientific community sites that show
Practicing connectivism	• Review a strategy for building, curating, and navigating a knowledge network • Practice tools for implementing this strategy • Discuss bias in GenAI	• Provide tasks for identifying reliable sources of information. • Identify tools for knowledge curation	• Identify a list of reputable experts and reliable media articles [13] • Identify expertise by looking at their track record for accuracy: explore institutional affiliations • Track Google Scholar and PubMed citations of experts; explore other experts they follow • Populate the list with people of multiple perspectives and disciplines • Identify resources to address misinformation [48]

[a]GenAI: generative artificial intelligence.

Discussion and Conclusion

In this commentary, we use a personal narrative to make the case for rethinking our approach to learning and acquiring knowledge in a connected world. We recognize the challenges posed by GenAI and the need for innovative solutions using connectivist approaches. Some of the challenges with GenAI are the concerns about biases, inaccuracies with hallucinations, and the accessibility of these tools. Being connected to learning communities, subscribing to credible and accurate sources of information, and being able to critically analyze information are potential solutions. Using freemium GenAI tools addresses some of the problems with accessibility and unfamiliarity with these tools. Embracing digital technologies and internet tools as health professional educators enables us to direct our learners through cohesive approaches to trainees to build, curate, and navigate knowledge networks that will help them address challenges with misinformation. As the amount of information and misinformation grows, one of the essential skills for health care practitioners will be the need to decipher accurate information from misinformation, which is possible by belonging to learning communities that amplify accurate information and expose the dangers of misinformation. GenAI poses an exciting opportunity to democratize resources through open-source collaborative efforts.

XSL•FO
RenderX

GenAI can be used across different learning paradigms, enabling learner-directed learning and empowering the teacher in a connectivist learning environment [49]. This approach of using GenAI as part of a new paradigm of learning allows educators to use these tools to provide powerful tutoring and adaptive testing based on analytics and to deliver personalized learning, improving assessment and feedback [50].

Summary

Practical Steps for Implementation in Health Care Education and Practice

First, creating a dynamic learning network encourages health care professionals and students to build and maintain professional networks through platforms like LinkedIn, X, and specialty-specific forums. Individuals can access real-time updates and exchange information. Use the feed generators from online communities and interested studies to build a knowledge network. Demonstrate the use of desired feed from journals on topics of interest. Identify experts from various disciplines.

Second, integrate technology into learning and practice; demonstrate the use of available resources such as podcasts, online webinars, blogs, and community forums to help with knowledge-building.

Third, develop critical information literacy skills, including training on how to critically assess the reliability and credibility of information sources, especially those found online. It is essential for learners to be able to distinguish between valid and misleading information and address misinformation.

Fourth, by leveraging artificial intelligence capabilities, clinicians can filter out unnecessary information, and GenAI tools can identify relevant scientific articles through feeds from reliable sources of information and amplify their search using freemium tools such as Research Rabbit. In addition, these tools allow for adaptive personalized instruction to the learner.

Fifth, the clinician can incorporate these tools as part of a synchronous classroom activity and model critical inquiry skills when using GenAI. By analyzing the accuracy of outputs, clinicians can model iterative prompting to refine GenAI outputs, refine learning, and expose the gaps in outputs.

Acknowledgments

The authors acknowledge the Huffington Department of Education Innovation and Technology at Baylor College of Medicine and the Lerner College of Medicine Cleveland Clinic for their support. The open access publication fee was funded in part by a nonrestrictive donor gift through the Hennick Bridgepoint Foundation.

Conflicts of Interest

HM declares a paid role as interim assistant dean of clinical faculty development at Toronto Metropolitan University School of Medicine, paid honorariums for invited academic talks for various organizations and conferences on academic technologies., and a nonrestricted educational grant from Novo Nordisk for online faculty development on teaching social determinants of health.

References

1. Gladiss Merlin NR, Jacobsen KH. Statewide COVID-19 stay-at-home orders and population mobility in the United States. World Med Health Policy 2020 Dec;12(4):347-356 [FREE Full text] [doi: 10.1002/wmh3.350] [Medline: 32837774]
2. Santarpia JL, Rivera DN, Herrera VL, Morwitzer MJ, Creager HM, Santarpia GW, et al. Transmission potential of SARS-CoV-2 in viral shedding observed at the University of Nebraska Medical Center. medRxiv. Posted online June 3, 2020 . [doi: 10.1101/2020.03.23.20039446]
3. @linseycmarr. Let's talk about #airborne transmission of #SARSCOV2 and other viruses. A discussion is needed to improve accuracy and reduce fear associated with the term. Mar 5. URL: https://twitter.com/linseymarr/status/1235640400054046724 [accessed 2024-02-22]
4. Morawska L, Milton DK. It is time to address airborne transmission of coronavirus disease 2019 (COVID-19). Clin Infect Dis 2020 Dec 03;71(9):2311-2313 [FREE Full text] [doi: 10.1093/cid/ciaa939] [Medline: 32628269]
5. Randall K, Ewing ET, Marr LC, Jimenez JL, Bourouiba L. How did we get here: what are droplets and aerosols and how far do they go? A historical perspective on the transmission of respiratory infectious diseases. Interface Focus 2021 Dec 06;11(6):20210049 [FREE Full text] [doi: 10.1098/rsfs.2021.0049] [Medline: 34956601]
6. CDC statement for healthcare personnel on hand hygiene during the response to the international emergence of COVID-19. US Centers for Disease Control and Prevention. URL: https://stacks.cdc.gov/view/cdc/85944 [accessed 2024-02-22]
7. @WHO. FACT: #COVID19 is NOT airborne. The #coronavirus is mainly transmitted through droplets generated when an infected person coughs, sneezes or speaks. 2020 Mar 28. URL: https://twitter.com/WHO/status/1243972193169616898 [accessed 2024-02-22]
8. Jimenez JL, Marr LC, Randall K, Ewing ET, Tufekci Z, Greenhalgh T, et al. What were the historical reasons for the resistance to recognizing airborne transmission during the COVID-19 pandemic? Indoor Air 2022 Aug;32(8):e13070 [FREE Full text] [doi: 10.1111/ina.13070] [Medline: 36040283]

XSL•FO
RenderX

9. Greenhalgh T, Jimenez JL, Prather KA, Tufekci Z, Fisman D, Schooley R. Ten scientific reasons in support of airborne transmission of SARS-CoV-2. Lancet 2021 May 01;397(10285):1603-1605 [FREE Full text] [doi: 10.1016/S0140-6736(21)00869-2] [Medline: 33865497]
10. Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations. World Health Organization. URL: https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations [accessed 2024-03-11]
11. Xie X, Li Y, Chwang ATY, Ho PL, Seto WH. How far droplets can move in indoor environments--revisiting the Wells evaporation-falling curve. Indoor Air 2007 Jun;17(3):211-225. [doi: 10.1111/j.1600-0668.2007.00469.x] [Medline: 17542834]
12. Morawska L, Bahnfleth W, Bluyssen P, Boerstra A, Buonanno G, Dancer S, et al. Coronavirus disease 2019 and airborne transmission: science rejected, lives lost. Can society do better? Clin Infect Dis 2023 May 24;76(10):1854-1859 [FREE Full text] [doi: 10.1093/cid/ciad068] [Medline: 36763042]
13. The 60-year-old scientific screwup that helped covid kill. WIRED. URL: https://www.wired.com/story/the-teeny-tiny-scientific-screwup-that-helped-covid-kill/ [accessed 2021-10-24]
14. Chapin CV. The Sources and Modes of Infection. Hoboken, New Jersey: J. Wiley; 1912.
15. Woo PCY, Lau SKP, Chu C, Chan K, Tsoi H, Huang Y, et al. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol 2005 Jan;79(2):884-895 [FREE Full text] [doi: 10.1128/JVI.79.2.884-895.2005] [Medline: 15613317]
16. Yu ITS, Li Y, Wong TW, Tam W, Chan AT, Lee JHW, et al. Evidence of airborne transmission of the severe acute respiratory syndrome virus. N Engl J Med 2004 Apr 22;350(17):1731-1739. [doi: 10.1056/NEJMoa032867] [Medline: 15102999]
17. Hare R. The transmission of respiratory infections. Proc R Soc Med 1964 Mar;57(3):221-230 [FREE Full text] [Medline: 14130877]
18. Lewis D. Why the WHO took two years to say COVID is airborne. Nature 2022 Apr;604(7904):26-31. [doi: 10.1038/d41586-022-00925-7] [Medline: 35388203]
19. Janis IL. Groupthink: Psychological Studies of Policy Decisions and Fiascoes, 2nd Ed. Boston, MA: Houghton Mifflin; 1982:978.
20. Chapman SJ, Shelton B, Mahmood H, Fitzgerald JE, Harrison EM, Bhangu A. Discontinuation and non-publication of surgical randomised controlled trials: observational study. BMJ 2014 Dec 09;349:g6870 [FREE Full text] [doi: 10.1136/bmj.g6870] [Medline: 25491195]
21. Kasenda B, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, et al. Prevalence, characteristics, and publication of discontinued randomized trials. JAMA 2014 Mar 12;311(10):1045-1051 [FREE Full text] [doi: 10.1001/jama.2014.1361] [Medline: 24618966]
22. Janis IL. Victims of Groupthink: A Psychological Study of Foreign Policy Decisions and Fiascoes. Oxford, UK: Houghton Mifflin; 1972.
23. Horgan D. Building eminence through evidence. Biomed Hub 2017;2(Suppl 1):230-238 [FREE Full text] [doi: 10.1159/000481615] [Medline: 31988953]
24. Siemens G. Connectivism. Pressbooks. 2004. URL: https://lidtfoundations.pressbooks.com/chapter/connectivism-a-learning-theory-for-the-digital-age/ [accessed 2021-10-24]
25. Brandao P, Algarvio D. Connectivism, information technologies and distance learning. Kriativ.tech. 2020. URL: http://www.kriativ-tech.com/wp-content/uploads/2018/02/ConnectivismInformationTechnologiesDistanceLearning.pdf [accessed 2024-03-19]
26. Omodan BI, Mtsi N, Mpiti PT. Enhancing virtual teaching and learning through connectivism in university classrooms. J Curric Teach 2023;12(4):116. [doi: 10.5430/jct.v12n4p116]
27. Goldie JGS. Connectivism: a knowledge learning theory for the digital age? Med Teach 2016 Oct;38(10):1064-1069 [FREE Full text] [doi: 10.3109/0142159X.2016.1173661] [Medline: 27128290]
28. Wongwatkit C, Thongsibsong N, Chomngern T, Thavorn S. The future of connectivist learning with the potential of emerging technologies and AI in Thailand: trends, applications, and challenges in shaping education. J Learn Sci Educ 122;2(1):122-154 [FREE Full text]
29. Downes S. Recent work in connectivism. Eur J Open Distance E-Learning 2020;22(2):113-132. [doi: 10.2478/eurodl-2019-0014]
30. Omodan BI. Analysis of connectivism as a tool for posthuman university classrooms. J Curric Stud Res 2023;5(1):1-12. [doi: 10.46303/jcsr.2023.2]
31. AlDahdouh A, Osorio A, Caires S. Understanding knowledge network, learning and connectivism. Int J Instruct Tech Distance Learn 2015;12(10):1-19 [FREE Full text]
32. Lacković N. Thinking with digital images in the post-truth era: a method in critical media literacy. Postdigit Sci Educ 2020;2(2):442-462. [doi: 10.1007/s42438-019-00099-y]
33. Corbett F, Spinello E. Connectivism and leadership: harnessing a learning theory for the digital age to redefine leadership in the twenty-first century. Heliyon 2020 Jan;6(1):e03250 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03250] [Medline: 31993523]

XSL•FO
RenderX

34. Utecht J, Keller D. Becoming relevant again: applying connectivism learning theory to today's classrooms. Crit Quest Educ Acad Educ Stud 2019;10(2):107-119 [FREE Full text]
35. Hendricks GP. Connectivism as a learning theory and its relation to open distance education. Progressio 2019 Dec 20;41(1). [doi: 10.25159/2663-5895/4773]
36. Cleary Y. Fostering communities of inquiry and connectivism in online technical communication programs and courses. J Tech Writ Commun 2020 Dec 11;51(1):11-30. [doi: 10.1177/0047281620977138]
37. Fancsali SE, Holstein K, Sandbothe M, Ritter S, McLaren BM, Aleven V. Towards practical detection of unproductive struggle. In: Bittencourt II, Cukurova M, Muldner K, Luckin R, Millán E, editors. Artificial Intelligence in Education. Cham, Switzerland: Springer International Publishing; 2020:92-97.
38. Dautov R, Husom EJ, Sen S, Song H. Towards community-driven generative AI. Ann Comp Sci Info Sys 2023;36:43-50 [FREE Full text] [doi: 10.15439/2023f5494]
39. Anderson DP, Cobb J, Korpela E, Lebofsky M, Werthimer D. SETI@home. Commun ACM 2002 Nov;45(11):56-61. [doi: 10.1145/581571.581573]
40. Faritha Banu J, Revathi R, Suganya M, Gladiss Merlin NR. IoT based cloud integrated smart classroom for smart and a sustainable campus. Procedia Comp Sci 2020;172:77-81. [doi: 10.1016/j.procs.2020.05.012]
41. Shemshack A, Spector JM. A systematic literature review of personalized learning terms. Smart Learn Environ 2020 Oct 23;7(1):33. [doi: 10.1186/s40561-020-00140-9]
42. Hutchins E. Distributed Cognition. Cornell University. URL: https://arl.human.cornell.edu/linked%20docs/Hutchins_Distributed_Cognition.pdf [accessed 2024-03-30]
43. Fritch JW, Cromwell RL. Evaluating Internet resources: Identity, affiliation, and cognitive authority in a networked world. J Am Soc Inf Sci 2001;52(6):499-507. [doi: 10.1002/asi.1081]
44. Kapoun J. Teaching undergrads WEB evaluation: a guide for library instruction. C&RL News 2020 Feb 11;59(7):522-523 [FREE Full text] [doi: 10.5860/crln.59.7.522]
45. Jeong H, Hmelo-Silver C. Productive use of learning resources in an online problem-based learning environment. Comput Hum Behav 2010 Jan;26(1):84-99 [FREE Full text] [doi: 10.1016/j.chb.2009.08.001]
46. Wang T, Lund BD, Marengo A, Pagano A, Mannuru NR, Teel ZA, et al. Exploring the potential impact of artificial intelligence (AI) on international students in higher education: generative AI, chatbots, analytics, and international student success. Appl Sci 2023;13(11):6716. [doi: 10.3390/app13116716]
47. Confronting medical misinformation: Tips from the trenches. Association of American Medical Colleges. URL: https://www.aamc.org/news/confronting-medical-misinformation-tips-trenches [accessed 2024-04-16]
48. Ouyang F, Jiao P. Artificial intelligence in education: the three paradigms. Comput Educ Artif Intell 2021;2:100020. [doi: 10.1016/j.caeai.2021.100020]
49. Choi E, Borkowski M, Zakoian J, Sagan K, Scholla K, Ponti C, et al. Proc Assoc Info Sci Tech 2016 Feb 24;52(1):1-4. [doi: 10.1002/pra2.2015.145052010069]
50. Al-Zahrani A, Alasmari TM. Exploring the impact of artificial intelligence on higher education: the dynamics of ethical, social, and educational implications. Humanit Soc Sci Commun 2024 Jul 13;11(1):1-12. [doi: 10.1057/s41599-024-03432-4]

Abbreviations

GenAI: generative artificial intelligence

Edited by A Mavragani; submitted 04.10.23; peer-reviewed by Q Lu, S Purkayastha, AR Ashraf; comments to author 06.02.24; revised version received 16.04.24; accepted 14.10.24; published 03.12.24.

Please cite as:
Benjamin J, Pillow T, MacNeill H, Masters K, Agrawal A, Mehta N
Reflections From the Pandemic: Is Connectivism the Panacea for Clinicians?
J Med Internet Res 2024;26:e53344
URL: https://www.jmir.org/2024/1/e53344
doi:10.2196/53344
PMID:39625749

©Jennifer Benjamin, Tyson Pillow, Heather MacNeill, Ken Masters, Anoop Agrawal, Neil Mehta. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Unlocking the Potential for Implementation of Equitable, Digitally Enabled Citizen Science: Multidisciplinary Digital Health Perspective

Lucio Naccarella[1*], BSc, PhD; Jonathan Charles Rawstorn[2*], BPhEd, BSc, PhD; Jaimon Kelly[3*], PhD; Eleanor Quested[4*], PhD; Stuart Jenkinson[5*], BSc; Dominika Kwasnicka[1*], MA, MSc, PhD

[1]Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia
[2]Institute for Physical Activity and Nutrition, School of Exercise and Nutrition Sciences, Deakin University, Geelong, Australia
[3]Centre for Online Health, The University of Queensland, Brisbane, Australia
[4]Curtin School of Population Health, Faculty of Health Sciences, Curtin University, Perth, Australia
[5]Systemic Policy and Strategy Team, Carers Western Australia, Perth, Australia
[*]all authors contributed equally

Corresponding Author:
Dominika Kwasnicka, MA, MSc, PhD
Melbourne School of Population and Global Health
The University of Melbourne
207 Bouverie St Carlton
Victoria
Melbourne, 3010
Australia
Phone: 61 390355511
Fax: 61 390355511
Email: dom.kwasnicka@unimelb.edu.au

Abstract

Citizen science is a community-based participatory research approach with an emphasis on addressing health disparities that is increasingly advocated by the community, researchers, and research funders. Digitally enabled methods can extend the potential of citizen science by enabling citizens to engage in real-time research processes, such as data collection, information sharing, interpreting, acting on data, and informing decision-making. However, the power of any citizen science lies in promoting health equity by providing equal opportunity for citizen engagement. Without appropriate attention to recognize and address equity, digital enablement of citizen science may exacerbate rather than ameliorate health inequalities. In this Viewpoint, we draw on our digital health research experience and perspectives to outline the practice of citizen science in the context of digital health—how it is operationalized, key advocated principles, and challenges. We also discuss citizen science in relation to health equity and implementation science, including emphasizing the importance of integrating health equity principles and frameworks, health equity implementation determinants, and digital determinants of health. We demonstrate how equity could be achieved by providing a working example in the context of a digitally enabled approach to improving social, physical, and mental well-being among people with disability and caregivers.

(J Med Internet Res 2024;26:e50491) doi:10.2196/50491

KEYWORDS

citizen science; digital health; equity; implementation science; community; research; health inequality; health equity; health integration; mental well-being; well-being

Introduction

Citizen science is a community-based participatory research approach that specifically involves the participation of community members, who may not have formal scientific training, in the scientific process—often by collecting, analyzing, and interpreting data. Participatory research methodologies like citizen science are increasingly advocated by consumers, the community, researchers, and research funders [1]. While guiding principles exist [2,3] and many virtues are

XSL•FO
RenderX

claimed [4], implementation into practice remains a challenge for most researchers, practitioners, and policy makers [5,6]. Digital technologies have enabled the global growth of citizen science and led to the development of the field which here we coined as digitally enabled citizen science, where digital technologies support citizen scientists in engaging in real-time research processes.

The successful integration of citizen science, digital health equity, and health equity implementation into digital health care could be achieved by using existing principles, frameworks, and known health equity implementation determinants and digital determinants of health. In this interdisciplinary viewpoint, we outline key principles and challenges of citizen science. We highlight the intersection of citizen science and digital health, and emphasize the importance of health equity in implementation. We present a digitally enabled citizen science project (ConnectUp) as a working example of how relevant principles, frameworks, and determinants can be integrated (Textbox 1).

Textbox 1. Case study: ConnectUp.

ConnectUp is a prototype web-based platform that connects caregivers and people with disability with geographically proximal peers to engage in physical activities. Users create profiles describing personal characteristics (age, physical activities they enjoy engaging in, how far they are willing to travel to meet with someone who they matched with), and ConnectUp matches people with similar profiles to facilitate connections and engagement.

The platform was co-designed by a range of stakeholders including a state-level leading not-for-profit organization supporting caregivers (Carers WA), a digital health agency (Eduka Pty Ltd), and people with disability and their caregivers. A total of 17 participants took part in a co-design process, led by a person with disability, which included 4 web-based or 2 face-to-face workshops. Participants represented a broad range of ages (range 20-73 years), cultural backgrounds (Australian, Cambodian, Singaporean, Romanian, and English), and disabilities (intellectual, physical, and mental health conditions).

The co-designed prototype is currently being extended to support digitally enabled citizen science and enhance the user experience. ConnectUp users will be able to upload new physical activity opportunities into the platform (including defining characteristics such as location, activity type, and accessibility considerations), update characteristics of existing opportunities (eg, changing accessibility, new characteristics relevant to different types of disability), and review existing opportunities to provide lived-experience perspectives that can help other ConnectUp users to determine suitability for their needs. ConnectUp will empower citizens to create, curate, and disseminate information that can strengthen physical, social, and psychological well-being among people with disability and caregivers, as well as generate large-scale evidence about the accessibility and inclusivity of community amenities to support advocacy for policy and business reform.

Citizen science is typically operationalized on a spectrum with varying levels of citizen involvement. Multiple taxonomies reflect this variation [7-9]. Table 1 provides 2 examples.

Higher levels of involvement suggest greater commitments to creating equal opportunity for citizens to engage in the research process, key to promoting health equity. In ConnectUp, we support higher levels of citizen involvement, empowering citizens to be involved at all research stages, including them in Steering Committee and Advisory Committee, and using data to inform and activate community action. While citizen science is increasingly advocated to address long-standing health inequity, the issue of "power dynamics" between researchers and citizens is debated by the community, researchers, and research funders, especially in the context of the COVID-19 pandemic [9] For example, lower involvement may widen power imbalances and potentially exploit vulnerable community groups because it provides citizens with less control and voice than higher involvement.

Table 1. Examples of citizen science taxonomies.

Level of involvement	Example taxonomy 1 (English et al [7])	Example taxonomy 2 (King et al [8] and Tan et al [9])
Low	Crowdsourcing: involves active or passive citizen participation in data collection	For the people: citizens are involved in informing and contributing to research 1-way via meetings
Low	Limited participatory research: involves citizens in problem definition and data collection	With the people: citizens are involved in consulting and collaborating in research processes via consultations, interviews, or surveys
High	Extreme citizen science: involves citizens in analysis and interpretation, study dissemination, and public health action	By the people: citizens are empowered to be involved in all research processes, via advisory committees or community forums and can include applying data to inform and activate community action

Citizen science has been applied and adapted within diverse research contexts, population groups, and content areas, and guiding principles have been developed by regulatory bodies to promote consistent and high-quality practices. For example, the Australian and European Citizen Science Associations both recommend 10 community citizen science engagement principles to ensure research quality, equity, inclusion, and governance [10,11] (Table 2). Table 2 demonstrates how these principles have been actioned in the ConnectUp case study.

XSL•FO
RenderX

Table 2. Citizen science engagement principles [10,11], exemplified with ConnectUp.

Engagement principles	ConnectUp: case study principles in action
1. Project actively involves citizens and generates new knowledge or understanding.	People with disability and caregivers co-defined the problem (lack of physical activity and opportunities to be active and socialize) and solution (web-based platform creating connections with like-minded people and physical activity opportunities in nearby locations).
2. Project has a genuine scientific outcome.	The platform that was proposed by users will expand the initial scope of the project to facilitate social connection. It will include elements of data collection (listing and reviewing places to be active), data summaries and feedback (rating of facilities and activities on offer), and data will be collated, analyzed, and co-summarized with and by consumers (people using the platform). The project will include a robust evaluation of effects on meaningful health/social outcomes, the evaluation will be developed and refined with all stakeholders.
3. Project provides benefits to both science and society.	With the active participation of consumers in gathering data, researchers will be able to better understand several issues including barriers and facilitators to access physical activity opportunities. The benefits for society will include service provision—the users will be able to access and use up-to-date crowdsourced information that is relevant to them and in the area that they live in. They will be able to actively participate in the research process if required.
4. Citizen scientists may participate in various stages of the scientific process.	Stages that citizen scientists participate in include study and platform design (partially completed); data collection, data indexing and categorizing, data processing, analyzing, and summarizing (together with the researchers). Followed by involvement in outcomes interpretation, dissemination and local advocacy, participation in publications and presentations as well as informing local policy.
5. Citizen scientists receive feedback from the project team.	The academic researchers seek to identify citizen scientists' capacity development needs and provide appropriate feedback, training, and resources to support development.
6. Project has limitations and biases that should be considered and controlled for.	Not all citizen scientists may see themselves as equal participants in the research process. Through extensive training, conversations, responsibility sharing, and project goal setting, the researchers will continue to strive to ensure equity of involvement between citizen scientists and researchers.
7. Project data and meta-data are publicly available and results are published in accessible format.	Project data and meta-data will be publicly available and results will be published on the ConnectUp website, the summary of research results will be published in accessible reports. Data will also be available on the Open Science Framework to increase data discoverability in the scientific community.
8. Citizen scientists are suitably acknowledged by project team.	Citizen scientists will continue to colead the academic researchers, for instance, coauthor SJ received training in qualitative research methods and led co-design workshops/focus groups. Lead citizen scientists will be named on the study website and will coauthor scientific publications (eg, SJ and others).
9. Project offers benefits and outcomes are considered in project evaluation.	Citizen scientists will continue to be involved in defining key project benefits and setting outcomes and they will take an active role in project evaluation during the full duration of the project.
10. Project takes into consideration relevant legal and ethical issues.	Legal and ethical aspects of the project are considered by the full team throughout the research project by discussing and providing training in legal and ethical principles of citizen science–led research and codeveloping ethical standards to follow the citizen science best practices.

XSL•FO
RenderX

Several insights from the community members pointed to the need for the proposed work:

> *We've done some research on Google for something and then you really excited or keen and then you look at it and go okay, that's not that program, it's not running or it's finished. And we do believe it would be a full-time job, or potentially a team of people updating the information but the importance of that to us as carers, I think is invaluable. It would be worth having because we are so time poor we want something that is a one stop shop.... You know, there's a lot of "come and try" days for people. But how many people know about it? Unless you're on the right mailing list, you don't know about them or what they offer. I think it's bringing all those together in the one umbrella, and everybody promotes the one thing. At the moment, I get all my resources from different newsletters, and I think you need something that's geographically located.* [caregiver, ConnectUp user co-design workshop 1, hosted by Carers WA]

People with disability echoed this, raising similar issues:

> *I think if you start this database. You will have people then saying, hey, how about us, did you know about this? You can have others that can then put forward ideas. I think that is really important, because, you know, we all have different things that we are knowledgeable about, and it's just bringing it together.* [person with disability, ConnectUp user co-design workshop 3, hosted via web]

While the promises, virtues, and benefits of citizen science are increasingly being espoused [12,13], debates about operational and conceptual challenges in practice are also increasing (Textbox 2).

Textbox 2. Citizen science operational and conceptual challenges.

Operational challenges

- Ensuring equal representation of citizens in all research processes.
- Balancing scientific and social value of citizen-generated knowledge and citizen capabilities.
- Balancing connections between citizens and researchers and research quality.
- Managing the complex link between citizen science and participatory policy development.

Conceptual challenges

- Diverse motivations of citizens for participating in research.
- Dynamic public trust and distrust in scientific knowledge and endeavors.
- Issues of power, exploitation, and commitment to engagement in citizen science projects/programs.

Challenges related to citizen representation and power signal that health equity and implementation (ie, the use of strategies to adopt and integrate evidence-informed interventions and to change practice patterns) requires focused consideration, reflection, and action to ensure inequities are not perpetuated or increased.

Citizen Science in Digital Health

Digital health refers to the development and use of digital technologies to improve health, for example, virtual health, mobile health apps, wearable devices, the internet, and artificial intelligence tools that enable the storage, exchange, advanced analysis, and visualization of data [14]. Digital technologies have contributed to the rise of citizen science and led to the field which we coined here digitally enabled citizen science, where digital methods are used to engage citizens in real-time research processes [15]. However, the integration of "digital health" with "Citizen Science" is multi-faceted and complex. Digital health tools could facilitate equitable citizen participation in health intervention design, development, adoption, implementation, evaluation, and sustainability. This is achieved by providing a voice to citizens who experience health inequities and having the reach and power to source big data to inform health policy decision-making. However, digital health can also contribute to health inequities when designed, developed, implemented, and evaluated without adequate consideration of the digital determinants of health, such as digital health literacy [16]. The rapid digital transformation of health care may contribute to increased inequity as uneven adoption of health interventions can widen existing gaps between advantaged and disadvantaged populations [17].

Equitable Implementation of Digitally Enabled Citizen Science Initiatives: Guiding Frameworks

Health equity means that no one is denied access to optimal health and well-being because they are economically or socially disadvantaged [18]. Health interventions can perpetuate health disparities when implementation leads to inequitable inputs (eg, funding), outputs (eg, uptake and quality), and outcomes (eg, access to care). Race, ethnicity, sexual orientation, gender identity, socioeconomic status, functional limitations, and other characteristics can contribute to disparities in the implementation of health interventions. As implementation science [19] progressed, there was increased use of implementation determinant frameworks to understand why these disparities occur [20]. For instance, the health equity implementation framework [21] recommends the integration of three health equity domains into existing implementation science

XSL•FO
RenderX

frameworks, namely: (1) health care intervention recipient cultural factors (eg, socioeconomic status, race or ethnicity, and language); (2) interaction between clinicians and patients (eg, interactions can predict satisfaction, perceived trust, and health outcomes); and (3) societal context (eg, economic, demographic, or geographical factors).

These health equity domains can guide the use of existing implementation science frameworks and efforts to identify and understand barriers to equitable implementation of Digitally enabled citizen science initiatives. The promise and power of citizen science to address health inequities lie in the opportunity for broader, more equitable citizen participation and engagement in research processes—to identify, systematically collect, analyze, and use data that are meaningful and relevant to citizens, researchers, and policymakers. The known disparities in access to digital technology and opportunities within and between individuals, communities, and nations [22], reinforce the necessity of equitable digitally enabled citizen science, if the use of the paradigm is to result in large-scale health benefits.

A recent review [11] that aimed to summarize existing efforts to use citizen science to address health equity recommended expanding the focus on topics important for health equity (eg, equitable access). The suggestions included increasing the diversity of people serving as citizen scientists and their involvement and integration in research process phases (eg, applying the data to inform and activate community action), continuing to leverage emerging technologies that enable citizen scientists to collect data relevant to health equity, and strengthening the rigor of methods to evaluate impacts on health equity. Application of these recommendations has the potential to unlock the equitable implementation of citizen science projects, in particular where digital health tools are used.

Citizen Science and Digital Health Equity: Principles, Frameworks, and Determinants

Digital health equity refers to having an equal opportunity for individuals to benefit from the knowledge and practices related to the development and use of digital technologies to improve health [22]. Digital health technologies interact with social, cultural, and economic realities and with social determinants of health to indirectly contribute to health equity or inequity. Given that digital health tools are being used to engage citizens in real-time research processes [8], digital health equity principles, frameworks, and determinants need consideration to facilitate equitable citizen participation. While a universal set of digital health equity principles does not exist, principles relevant to digital health interventions have been suggested [23]. For example, equity principles have been derived from a digital equity assessment tool [23]. The digital health equity framework (DHEF) builds upon the health equity measurement framework [24] designed to measure the effects of social determinants of health to support improved statistical modeling and the measurement of health equity. Textbox 3 synthesizes equity principles relevant to digital health and illustrates how they have been actioned in the ConnectUp case study.

We advocate that these principles should also guide the design, development, implementation, and evaluation of digitally enabled citizen science initiatives. To date, several DHEFs [25-27] have been developed as calls to action that seek to guide the planning, implementation, and evaluation of equity-informed digital health interventions. The DHEFs are informed by and build upon integrative literature reviews and syntheses of existing health equity frameworks. These frameworks also signal the importance of digital determinants of health, the unique elements of people's experiences with the digital health environment (eg, digital health literacy, access to digital resources, and infrastructure) [17].

The DHEF [27] reinforces how the digital determinants of health interact with a person's current health state and needs, and other intermediate health factors such as psychosocial stressors, pre-existing health conditions, health-related beliefs and behaviors, and the environment. For example, access to digital health resources and digital health literacy interact with the degree and kind of psychosocial stress a person is currently experiencing; job loss or poverty, level of education, and previous exposure to digital media can all impact access. The framework also highlights the importance of approaching digital health technologies from an ecological perspective, considering the ways an individual's use of technologies extends out into (and is shaped by) their social, cultural, and economic position.

The framework for digital health equity [25] also highlights the ecological perspective to encourage multiple approaches to digital health by identifying individual-level, interpersonal-level, community-level, and societal-level determinants. For example, interventions targeting "upstream" determinants (eg, digital infrastructure) at the community and societal levels have the potential for the greatest impact on more populations. Collectively DHEFs can provide a comprehensive guide to planning, implementing, and evaluating digitally enabled citizen science initiatives to maximize equity.

XSL•FO
RenderX

Textbox 3. Digital health equity principles.

Digital health equity principles

The digital health initiative should

- facilitate equal community participation;
- improve the status of populations who are disadvantaged;
- narrow the health divide between population groups;
- reduce social inequalities throughout the whole population;
- not be driven by or exploit users for profit;
- be provided according to need, not ability to pay;
- tackle fundamental social determinants of health;
- provide a voice to the voiceless;
- facilitate equal access to services;
- evaluate the impacts by sex, gender, race/ethnicity, geographic and socioeconomic communities.

ConnectUp reflections on principles in action

ConnectUp is designed to

- create a community that is inviting to all. Equity challenges include digital health literacy, access to internet-enabled devices, and the impacts of some on user interactions (eg, blindness, neurological impairments);
- improve social connection and physical activity;
- reduce excess health burden and inequity associated with disability and caring;
- reduce social inequalities throughout the target population, making the platform user-friendly and accessible, and promoting it at the population level;
- protect users by not using data for advertising or sale to third parties;
- be freely available;
- improve social inclusion and nondiscrimination, and increase access to affordable, quality health services;
- enable marginalized people to build a community that exchanges knowledge and experiences of social connection and physical activity;
- identify and promote dissemination of social and physical activity opportunities that are suitable for people with disability;
- include a comprehensive process evaluation, designed jointly with people with disability and caregivers, to evaluate the impacts separately for differing sex, gender, race/ethnicity, and geographic and socioeconomic communities.

Conclusion

Digitally enabled citizen science is increasingly being advocated by the community, researchers, and research funders, particularly to address health disparities, and to unlock its equitable implementation potential. We acknowledge that challenges exist, particularly when an action requires collaboration between multiple fields that may have differing understandings and use of key terms and concepts (ie, citizen science, equity, and digital health). Our goal is to discuss citizen science and to create the opportunity by encouraging this coming together of different paradigms to encourage collaboration, and ultimately improve and widen the equitable application of digitally enabled citizen science. We advocate that future Digitally enabled citizen science requires the integration of digital health equity and health equity implementation. That is, to ensure initiatives are designed, developed, implemented, evaluated, and sustained using (1) principles (existing citizen science engagement principles and digital health equity principles provide a critical platform), (2) frameworks (existing health equity implementation and DHEFs provide a comprehensive guide for planning, implementation, and evaluation activities), and (3) determinants (the known health equity implementation determinants and digital determinants of health will optimize the identification of strategies to improve equitable implementation of digitally enabled citizen science initiatives).

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Consumer led research grant opportunity. National Health and Medical Research Council. URL: https://www.nhmrc.gov.au/funding/find-funding/mrff-2021-consumer-led-research-grant-opportunity [accessed 2023-05-15]
2. 10 principles of citizen science. Australian Citizen Science Association. URL: https://citizenscience.org.au/wp-content/uploads/2018/09/10-Principles-of-Citizen-Science.pdf [accessed 2023-05-15]
3. ECSA 10 principles of citizen science. EU Citizen Science Association. URL: https://eu-citizen.science/resource/88 [accessed 2023-05-15]
4. Bonn A, Hecker S, Bowser A, Makuch Z, Vogel J, Haklay M. Citizen science to foster innovation in open science, society and policy. Citizen Sci Innovation Open Sci Soc Policy 2018:465-484. [doi: 10.2307/j.ctv550cf2.38]
5. Vohland K, Land-Zandstra A, Ceccaroni L, Lemmens R, Perelló J, Ponti M, et al. Editorial: the science of citizen science evolves. In: Vohland K, editor. The Science of Citizen Science. Cham, Switzerland: Springer; 2021:1-12.
6. Den Broeder L, Devilee J, Van Oers H, Schuit AJ, Wagemakers A. Citizen Science for public health. Health Promot Int 2018;33(3):505-514 [FREE Full text] [doi: 10.1093/heapro/daw086] [Medline: 28011657]
7. English PB, Richardson MJ, Garzón-Galvis C. From crowdsourcing to extreme citizen science: participatory research for environmental health. Annu Rev Public Health 2018;39:335-350 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-013702] [Medline: 29608871]
8. King AC, Winter SJ, Chrisinger BW, Hua J, Banchoff AW. Maximizing the promise of citizen science to advance health and prevent disease. Prev Med 2019;119:44-47 [FREE Full text] [doi: 10.1016/j.ypmed.2018.12.016] [Medline: 30593793]
9. Tan YR, Agrawal A, Matsoso MP, Katz R, Davis SLM, Winkler AS, et al. A call for citizen science in pandemic preparedness and response: beyond data collection. BMJ Glob Health 2022;7(6):e009389 [FREE Full text] [doi: 10.1136/bmjgh-2022-009389] [Medline: 35760438]
10. Kimura A, Kinchy A. Citizen science: probing the virtues and contexts of participatory research. Engaging Sci Tech Soc 2016;2:331-361. [doi: 10.17351/ests2016.99]
11. Rosas LG, Rodriguez Espinosa P, Montes Jimenez F, King AC. The role of citizen science in promoting health equity. Annu Rev Public Health 2022;43:215-234 [FREE Full text] [doi: 10.1146/annurev-publhealth-090419-102856] [Medline: 34724389]
12. Bröder J, Okan O, Bauer U, Bruland D, Schlupp S, Bollweg TM, et al. Erratum to: health literacy in childhood and youth: a systematic review of definitions and models. BMC Public Health 2017;17(1):419 [FREE Full text] [doi: 10.1186/s12889-017-4365-x] [Medline: 28486939]
13. Hecker S, Bonney R, Haklay M, Hölker F, Hofer H, Goebel C, et al. Innovation in citizen science - perspectives on science-policy advances. Citizen Sci Theor Pract 2018;3(1):4. [doi: 10.5334/cstp.114]
14. Global strategy on digital health 2020-2025. World Health Organization. URL: https://apps.who.int/iris/handle/10665/344249 [accessed 2023-11-16]
15. Katapally TR. The SMART framework: integration of citizen science, community-based participatory research, and systems science for population health science in the digital age. JMIR Mhealth Uhealth 2019;7(8):e14056 [FREE Full text] [doi: 10.2196/14056] [Medline: 31471963]
16. Gallegos-Rejas VM, Thomas EE, Kelly JT, Smith AC. A multi-stakeholder approach is needed to reduce the digital divide and encourage equitable access to telehealth. J Telemed Telecare 2023;29(1):73-78. [doi: 10.1177/1357633X221107995] [Medline: 35733379]
17. Veinot T, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
18. Braveman P. What are health disparities and health equity? we need to be clear. Public Health Rep 2014;129 Suppl 2(Suppl 2):5-8 [FREE Full text] [doi: 10.1177/00333549141291S203] [Medline: 24385658]
19. Eccles MP, Mittman BS. Welcome to implementation science. Implementation Sci 2006;1(1):1-3. [doi: 10.1186/1748-5908-1-1]
20. Baumann AA, Cabassa LJ. Reframing implementation science to address inequities in healthcare delivery. BMC Health Serv Res 2020;20(1):190 [FREE Full text] [doi: 10.1186/s12913-020-4975-3] [Medline: 32164706]
21. Woodward EN, Singh RS, Ndebele-Ngwenya P, Melgar Castillo A, Dickson KS, Kirchner JE. A more practical guide to incorporating health equity domains in implementation determinant frameworks. Implement Sci Commun 2021;2(1):61 [FREE Full text] [doi: 10.1186/s43058-021-00146-5] [Medline: 34090524]
22. Astley CM, Clarke RA, Cartledge S, Beleigoli A, Du H, Gallagher C, et al. Remote cardiac rehabilitation services and the digital divide: implications for elderly populations during the COVID19 pandemic. Eur J Cardiovasc Nurs 2021;20(6):521-523 [FREE Full text] [doi: 10.1093/eurjcn/zvab034] [Medline: 34056650]
23. Were M, Sinha C, Catalani C. A systematic approach to equity assessment for digital health interventions: case example of mobile personal health records. J Am Med Inform Assoc 2019;26(8-9):884-890 [FREE Full text] [doi: 10.1093/jamia/ocz071] [Medline: 31188438]
24. Dover DC, Belon AP. The health equity measurement framework: a comprehensive model to measure social inequities in health. Int J Equity Health 2019;18(1):36 [FREE Full text] [doi: 10.1186/s12939-019-0935-0] [Medline: 30782161]
25. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]

XSL•FO
RenderX

26. Jahnel T, Dassow H, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093 [FREE Full text] [doi: 10.1177/20552076221129093] [Medline: 36204706]
27. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]

Abbreviations

DHEF: digital health equity framework

Edited by N Cahill; submitted 03.07.23; peer-reviewed by D Verran, J Scheibner; comments to author 07.10.23; revised version received 17.04.24; accepted 04.09.24; published 10.12.24.

Please cite as:
Naccarella L, Rawstorn JC, Kelly J, Quested E, Jenkinson S, Kwasnicka D
Unlocking the Potential for Implementation of Equitable, Digitally Enabled Citizen Science: Multidisciplinary Digital Health Perspective
J Med Internet Res 2024;26:e50491
URL: https://www.jmir.org/2024/1/e50491
doi:10.2196/50491
PMID:

©Lucio Naccarella, Jonathan Charles Rawstorn, Jaimon Kelly, Eleanor Quested, Stuart Jenkinson, Dominika Kwasnicka. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

From Doubt to Confidence—Overcoming Fraudulent Submissions by Bots and Other Takers of a Web-Based Survey

Jeffrey J Hardesty[1], MPH; Elizabeth Crespi[1], MPH; Joshua K Sinamo[1], BS; Qinghua Nian[1], PhD; Alison Breland[2], PhD; Thomas Eissenberg[2], PhD; Ryan David Kennedy[1], PhD; Joanna E Cohen[1], PhD

[1]Institute for Global Tobacco Control, Department of Health, Behavior and Society, Johns Hopkins University, Baltimore, MD, United States
[2]Center for the Study of Tobacco Products, Department of Psychology, Virginia Commonwealth University, Richmond, VA, United States

Corresponding Author:
Jeffrey J Hardesty, MPH
Institute for Global Tobacco Control
Department of Health, Behavior and Society
Johns Hopkins University
Fourth Floor
2213 McElderry St
Baltimore, MD, 21205
United States
Phone: 1 410 641 4537
Email: jhardesty@jhu.edu

Abstract

In 2019, we launched a web-based longitudinal survey of adults who frequently use e-cigarettes, called the Vaping and Patterns of E-cigarette Use Research (VAPER) Study. The initial attempt to collect survey data failed due to fraudulent survey submissions, likely submitted by survey bots and other survey takers. This paper chronicles the journey from that setback to the successful completion of 5 waves of data collection. The section "Naïve Beginnings" examines the study preparation phase, identifying the events, decisions, and assumptions that contributed to the failure (eg, allowing anonymous survey takers to submit surveys and overreliance on a third-party's proprietary fraud detection tool to identify participants attempting to submit multiple surveys). "A 5-Alarm Fire and Subsequent Investigation" summarizes the warning signs that suggested fraudulent survey submissions had compromised the data integrity after the initial survey launched (eg, an unanticipated acceleration in recruitment and a voicemail alleging fraudulent receipt of multiple gift codes). This section also covers the investigation process, along with conclusions regarding how the methodology was exploited (eg, clearing cookies and using virtual private networks) and the extent of the issue (ie, only 363/1624, 22.4% of the survey completions were likely valid). "Building More Resilient Methodology" details the vulnerabilities and threats that likely compromised the initial survey attempt (eg, anonymity and survey bots); the corresponding mitigation strategies and their benefits and limitations (eg, personal record verification platforms, IP address matching, virtual private network detection services, and CAPTCHA [Completely Automated Public Turing test to tell Computers and Humans Apart]); and the array of strategies that were implemented in future survey attempts. "Staying Vigilant" recounts the identification and management of an additional threat that emerged despite the implementation of an array of mitigation strategies, underscoring the need for ongoing vigilance and adaptability. While the precise nature of the threat remains unknown, the evidence suggested multiple fraudulent surveys were submitted by a single or connected entities, who likely did not possess e-cigarettes. To mitigate the chance of reoccurrence, participants were required to submit an authentic photo of their most used e-cigarette. Finally, in "Reflection 4 Years Later," we share insights after completing 5 waves of data collection without additional threats or vulnerabilities uncovered that necessitated the application of further mitigation strategies. Reflections include reasons for confidence in the data's integrity, the scalability and cost-effectiveness of the study protocols, and the potential introduction of sampling bias through recruitment and mitigation strategies. By sharing our journey, we aim to provide valuable insights for researchers facing similar challenges with web-based surveys and those seeking to minimize such challenges a priori. Our experiences highlight the importance of proactive measures, continuous monitoring, and adaptive problem-solving to ensure the integrity of data collected from participants recruited from web-based platforms.

(J Med Internet Res 2024;26:e60184) doi:10.2196/60184

XSL•FO
RenderX

KEYWORDS

fake data; recruitment; online survey; internet survey; challenges; data integrity; data quality; e-cigs; tobacco control; longitudinal survey; web-based survey; e-cigarette; vaping; smoking; smoke; cessation; prevalence; data collection; United States; US; adult; VAPER; Vaping and Patterns of E-cigarette Use Research

Introduction

Researchers increasingly use web-based surveys for data collection, offering them multiple options to source participants [1]. Companies, such as Qualtrics, offer web-based panels of possible survey participants, who are preidentified persons willing to respond to web questionnaires. Other companies, such as Amazon Mechanical Turk, offer crowdsourcing platforms that pair "workers" with paid web-based tasks that require effort from multiple people, including web-based surveys. In addition, researchers can create an original sample of web-based survey participants through recruitment strategies such as ads on social media platforms.

Each of the 3 options for sourcing participants has relative strengths and trade-offs in terms of the required effort from the research team, cost, speed of recruitment, data quality, maximum sample size, and representativeness of the target sample. For example, those building a web-based sample using a company-provided web panel or a crowdsource platform may find these ready-to-use options for sourcing participants result in studies achieving their desired sample size more quickly and at a lower cost as compared with an original sample recruited via social media ads or similar web-based recruitment approach. However, web-based panels and crowdsourcing platforms are not all created equally. Some companies provide researchers access to their web panel and oversee sampling and survey administration for a cost, while crowdsourcing platforms are more self-service and only provide access to their "workers." Furthermore, Peer et al [2] found that data quality may vary among web panels and crowdsourcing platforms, and others have reported poor data integrity due to participant misrepresentation of credentials after using a well-known company's subcontracted web panels [3]. Researchers expecting to recruit lower-prevalence populations may also find that the attainable sample size, pace of recruitment, or representativeness from existing web-based panels and crowdsourcing platforms are inadequate for their study aims [4].

In contrast to those using a company-provided web panel or crowdsource platform, researchers who are interested in building a web-based original sample must plan, build, test, and implement recruitment and data collection methods and related strategies, which may result in slower recruitment and higher costs. However, these bespoke solutions may allow researchers to recruit lower-prevalence populations more efficiently once established, to achieve a more representative sample, or to improve data quality. Achieving these benefits requires not only expertise in survey sampling techniques but also proficiency in mitigating risk to data integrity [5-7]. Importantly, the latter includes minimizing fraudulent survey submissions from individuals who aim to deceive researchers, typically to receive incentives [5,8-10]. Pozzar et al [11] conducted a data integrity experiment and found that as much as 94.5% of their sample, which was recruited by social media and intended to be English-speaking US adults diagnosed with ovarian cancer within the last 12 months, was fraudulent. This occurred despite implementing practices consistent with lessons learned from previously published research and seeking advice from their institution's Research Electronic Data Capture (REDCap) administrators and survey research core (eg, eligibility criteria were not apparent on promotions, CAPTCHA [Completely Automated Public Turing test to tell Computers and Humans Apart], and a hidden questionnaire item visible only to survey bots) [11]. Others such as Pratt-Chapman et al [6] aimed to recruit cancer survivors through social media and removed three-fourths of their initial sample, despite using CAPTCHA, a hidden questionnaire item, instructions to type a specific word, and other strategies to mitigate the impact of survey bots that automate the completion of multiple surveys. Griffin et al [12] had a similar experience with social media and professional listserve recruitment of people who identify as LGBTQ+ (lesbian, gay, bisexual, transgender, queer) and removed 61.8% of their initial survey responses, despite using Qualtrics' "Prevent Ballot Box Stuffing" feature (prevents multiple submissions from one individual by placing a cookie in the browser) and survey bot detection features [13]. These recent examples suggest that recommendations from the literature and other sources for addressing fraudulent survey submissions may be insufficient, improperly applied, or both, necessitating that researchers continue to share their failures and successes so that others may learn from their experiences.

In 2019, we launched the Vaping and Patterns of E-cigarette Use Research (VAPER) Study, which has been described in detail elsewhere [14]. Participants were surveyed to understand better how their product use may change in response to regulations. We recruited a web-based original sample of a lower-prevalence population, adults who frequently use e-cigarettes (5+ days per week), to join a longitudinal web-based survey. Over the next 4 years, we experienced a failed survey attempt, learned about the threats and vulnerabilities that likely led to the fraudulent survey submissions that compromised the data integrity, learned about the benefits and limitations of risk mitigation strategies, implemented an effective array of mitigation strategies, and, ultimately, achieved the desired sample size and ended with reasonable confidence in the data integrity. This paper has 5 aims. First, we aim to summarize the events, decisions, and assumptions during the study preparation phase that likely led to the failed initial survey attempt ("Naïve Beginnings"). Second, we present the warning signs that suggested fraudulent survey submissions may have compromised the data integrity after the initial survey was launched, and the results from the subsequent investigation into the extent of the issue and how the methodology was exploited ("A 5-Alarm Fire and Subsequent Investigation"). Third, we describe the threats and vulnerabilities that likely led to the fraudulent survey submissions during the initial survey attempt (with a focus on anonymity and survey bots), the corresponding

XSL•FO
RenderX

mitigation strategies and their benefits and limitations, and the array of strategies that were implemented in future survey attempts ("Building More Resilient Methodology"). Fourth, we detail our experience identifying and addressing another threat and vulnerability that led to suspicious survey submissions despite various mitigation strategies in place after restarting wave 1 ("Staying Vigilant"). Last, we share reflections after completing 5 waves of data collection with no additional threats or vulnerabilities uncovered that warranted the application of additional mitigation strategies ("Reflections 4 Years Later").

Ethical Considerations

Ethical approval was obtained from the IRB of the Virginia Commonwealth University (approval HM20015004; date: May 14, 2019), with the Johns Hopkins Bloomberg School of Public Health IRB relying on the IRB of Virginia Commonwealth University as the IRB of record (IRB9277).

Participants were directed to read the written consent form on a web page and click "Agree" or "Do not agree" in response to a prompt asking, "Do you consent to answering the screening and survey questions?" Participants were also provided with the opportunity to re-read the consent form prior to each follow-up questionnaire. A waiver of documentation of signatures was received from the IRB. As part of the informed consent, participants consented to allow the use of their nonidentifying data in future research studies and distribution of their nonidentifying data to the study sponsor or another researcher for future studies without additional informed consent.

During the study, risk for loss of confidentiality was mitigated by securely storing documentation containing participant-identifying information on university servers with restricted access. A Certificate of Confidentiality was also obtained from the National Institutes of Health to help keep participant data private (eg, to prevent a court from obtaining participant data in a court case). Upon study completion, all documentation containing identifying information was deidentified.

Participants who completed the survey during the failed initial survey attempt received a $10 gift code from a company of their choice as compensation. After implementing risk mitigation strategies, participants in waves 1-5 received Amazon gift codes worth US $10-$30, with the amount varying based on the survey wave, whether it was a baseline or follow-up survey, and the subset of questions completed.

Naïve Beginnings

The VAPER Study recruited adults aged 21 years and older who used e-cigarettes at least 5 days per week. This was a lower prevalence population, with 2.3% of US adults using e-cigarettes daily and 5.1% of adults using in the past 30 days in 2020 [15]. A sample of at least 1200 participants was required to adequately power the VAPER Study's initial hypotheses, after adjusting for an anticipated loss to follow-up rate of 25% and assuming an effect size of 10% (for *t* tests to detect differences between 2 dependent means using a 2-tailed test), Cronbach $\alpha<0.05$, and power of 0.85 [14]. Additionally, a subset of the sample was intended to attend an in-person laboratory visit for puff topography measurements (measurements of the duration, volume, and frequency of puffs taken during the use of an e-cigarette), which required limiting recruitment to Richmond, Virginia; Columbus, Ohio; and Los Angeles, California, where collaborating laboratory partners were located.

We found that no existing web-based panel or crowdsourcing platform available during 2019 could support the study due to sample size considerations and the limited catchment areas, necessitating that we build a web-based original sample. We collaborated with a university-based survey research center in the United States with experience in web-based recruitment and survey data collection to build the survey using a questionnaire provided by our team, implement a Facebook- and Instagram-based recruitment strategy (specific platforms were recommended by the research center), and distribute incentives.

The survey was custom-built on a proprietary platform. Data storage protections primarily included network security measures (eg, partitioned network storage to mitigate the potential for data loss and to administer appropriate permissions) and integrity verification (eg, Varonis DatAdvantage to report on authorized and unauthorized changes to file server data). The research center also used a third-party's proprietary fraud detection tool to identify participants attempting to take the survey multiple times. At the time of its inclusion, the tool functioned by monitoring and modeling metadata and behaviors to generate a fraud profile score. In addition, another third-party tool was used to streamline the delivery of incentives to participants in the form of US $10 gift codes from a company of their choosing. Each survey was intended to last about 20 minutes.

To help increase the pace of recruitment, anonymous survey taking was permitted, allowing participants to register for the survey by signing in with a Gmail or Facebook account or by creating an account with any email address. These joint decisions were informed by the research center's previous experiences with web-based recruitment and web-based survey data collection and the assumption that the third-party detection tool intended to prevent multiple survey attempts would minimize fraudulent survey submissions and preserve data integrity.

A 5-Alarm Fire and Subsequent Investigation

The target population's lower prevalence combined with a limited catchment area resulted in modest expectations for the pace of recruitment. During the first 4 weeks of recruitment, each city yielded 1-10 participants per day, which was consistent with these expectations. However, approximately 1 month after recruitment began, the pace of recruitment accelerated from 1-10 to 11-124 submissions per day per city. Upon receiving a participant's voicemail message alleging fraudulent receipt of multiple gift codes, concerns arose about the data integrity. Consequently, recruitment, data collection, and incentive delivery were halted, and a 2-month investigation of 1624 survey submissions commenced to determine how participants had

XSL•FO
RenderX

exploited the methodology and the approximate number of fraudulent survey submissions.

The investigation revealed that participants likely had circumvented the mitigation strategies by clearing cookies, using new browsers or devices, using virtual private networks (VPNs), discerning a pattern in survey URL generation, and interfering with the third-party fraud detection tool. As identifiable data were not collected, determining the precise extent of the issues was not feasible. As a result, a liberal approach was adopted to identify fraudulent survey submissions using a set of criteria. These criteria included, but were not limited to, short completion times, surveys taken without first clicking a social media ad, duplicate account information (eg, same or similar email addresses), and the other indicators, as detailed in Textbox 1.

Based on these criteria, it was determined that only 363 of the 1624 survey completions were likely valid, representing 22.4% of the total. Due to the imprecision of labeling participants as "likely valid" and the possibility of inviting participants who submitted fraudulent surveys to another survey, these data were not used for analysis and the participants were not invited to participate in future survey attempts.

Textbox 1. Key criteria for identifying fraudulent survey submissions used during the investigation of the failed initial survey attempt.

Criteria

- Short completion times
- Surveys taken without first clicking a social media ad
- Duplicate account information (eg, same or similar email addresses)
- Unusually high amount of "don't know" and "prefer not to answer" responses (2 SDs above the mean)
- Third-party fraud detection tool scores suggesting multiple surveys were submitted by a participant
- Duplicate photos submitted across survey submissions (optional photos of their e-cigarette device were requested during the survey)
- Use of non-English alphabet (eg, Cyrillic alphabet)
- Ad-location mismatches (eg, an ad was for Los Angeles, California, and the survey was taken in Boston, Massachusetts)
- Unusually high agreement between responses across survey submissions suggesting multiple surveys were submitted by a participant (2 SDs above the mean)
- Inconsistent responses to questionnaire items (eg, inconsistent device type and brand)

Building More Resilient Methodology

Overview

Prior to resuming data collection, we reviewed the literature and sought guidance from independent data authentication experts. The experts consistently indicated that while fraudulent survey submission and data integrity issues cannot be completely prevented, they can be mitigated through a combination of strategies. Their recommendations included addressing anonymity, using CAPTCHA to mitigate the impact of survey bots, shortening data collection windows, and randomizing survey URLs generated. In this section, we describe the threats and vulnerabilities that likely led to the fraudulent survey submissions during the initial survey attempt (with a focus on anonymity and survey bots), the corresponding mitigation strategies and their benefits and limitations, and the array of strategies that were implemented in future survey attempts.

Anonymity

Anonymity may be appropriate in certain study designs [16], but surveys that recruit anonymous participants and collect data on the internet are vulnerable [7,17] and can be compromised by a single or group of individuals with the intent to deceive. Anonymity allows for the unfettered creation and use of survey bots and multiple email addresses, phone numbers, IP addresses (eg, via VPNs), browsers, and devices to complete one or multiple fraudulent survey submissions. This situation arises because anonymous fraudulent survey submissions are more difficult to detect with certainty and anonymous participants can submit additional fraudulent surveys using the same or different strategies even after detection and removal from the dataset.

Anonymity: Mitigation Strategies and Their Benefits and Limitations

Using a personal record verification platform can mitigate anonymity concerns significantly. To use these services, researchers must collect each participant's name, date of birth, and residential mailing address and cross-reference these data against information about the participant on the platform. These 3 forms of verified personally identifiable data are a powerful tool for determining whether a survey likely was submitted by a unique and real participant, considering that their unverified equivalents can be fabricated and functioning email addresses and phone numbers can be generated and authenticated by individuals who aim to deceive researchers. Nevertheless, this strategy is not foolproof. Limitations include the following: (1) post office (PO) box addresses may not be tracked by the personal record verification platform; (2) personally identifiable data can be obtained illegally (eg, on the dark web, an encrypted subsection on the internet that is not indexed by a typical search engine and frequently used to conduct illegal activities) [18]; (3) participants can own or live at multiple addresses; (4) participants can use the address of a family member, friend, or neighbor; (5) personal record verification platforms have imperfect datasets (eg, participants move and the dataset may

XSL•FO
RenderX

not have updated); and (6) participants can make honest mistakes (eg, typos or provide nicknames), resulting in mismatched data at no fault of the participant.

In addition, IP address matching can be used to identify instances where an individual has submitted multiple surveys, and IP address lookup tools can be used to verify participants' approximate geographic location. However, the utility of these 2 strategies is limited because IP addresses were not invented to identify internet users [19]. Rather, they were created to identify devices on a network that transmit data [19]. This key distinction is important as there are tools available that can be used to manipulate IP addresses. Such tools include VPNs, which are web-based applications that allow users to access the internet through remote servers that may be located anywhere in the world, providing them with access to many IP addresses that can give researchers the false impression they are physically located in a specific country or subnational location [20]. VPNs pose a particular challenge to data integrity given their increasingly common use; a nationally representative survey conducted by a VPN provider suggests that 33.0% of Americans in 2023 use a VPN compared with 24.3% in 2022 [21]. Beyond VPNs, networks of proxy servers, such as Tor via Tor Browser, can be used to route internet traffic through more than one intermediary server [22], effectively hiding its users' real IP addresses. Other approaches to changing a device's IP address include resetting a router [23], using the same device on a different Wi-Fi network [24], and switching a mobile phone from a Wi-Fi network to a cellular network [25].

VPN detection services are an available tool that can assess whether IP addresses are likely associated with VPNs [26]. These services use various detection methods, with the simplest and most effective method being IP address blocklists, which match user IP addresses against known VPN-associated IP addresses. Other less effective methods include deep packet inspection, DNS leak testing, browser fingerprinting, and identifying mismatches in time zones [27]. To circumvent these detection methods, VPN providers offer features aimed at further obfuscating VPN use, and they are also known to change IP addresses when streaming services like Netflix add them to blocklists [28]. Furthermore, some VPN providers allow individual users to create unique and dedicated IP addresses, making it less likely that these IP addresses will be flagged as being associated with a VPN [29]. Taken together, while VPN detection services are a potentially valuable tool for researchers, more skilled and motivated users can still evade detection.

Anonymity: Strategies Applied in Future Survey Attempts

To mitigate concerns about anonymity and limitations involving the related strategies, several protocol changes were implemented in our future survey attempts. Each participant's name, date of birth, and residential mailing address were collected and cross-referenced with corresponding data on LexisNexis' personal record verification platform. After taking into account the limitations of LexisNexis' platform, 3 additional protocol decisions were made. First, PO box addresses were not permitted because LexisNexis did not track this information, and someone may own multiple PO boxes. Next, to delay gratification, the first incentive was mailed to the residential address provided rather than sending it electronically. This procedure also ensured individuals who used another person's identifiable information without their knowledge did not receive the incentive. Third, when self-reported identifiable information did not match the personal record verification platform's dataset, a picture of a utility bill or driver's license, or similar identification, was requested, depending on the information that was unable to be verified. This procedure provided participants with a pathway to enroll in the study if they would have been otherwise excluded due to LexisNexis' imperfect dataset or an honest mistake. IP address–related strategies, including VPN detection services, were not deemed essential given the aforementioned protocol changes. Nevertheless, IP address matching was used to identify multiple survey submissions as a first pass because it was more efficient than conducting a personal record search. Last, participants' cell phone numbers were requested for authentication purposes and to encourage participants to take the survey from their phones, which provided a more streamlined experience for uploading authentic photos of their devices and liquids. Email addresses were requested as well but not for authentication purposes. Rather, email addresses were collected as a means of contacting participants beyond their cell phone numbers and disseminating gift codes to those who had completed 2 or more surveys.

Survey Bots

Bots are computer programs that operate on the internet and perform repetitive tasks [30], and survey bots are a type of bot that can submit multiple web-based surveys [31]. Limited information was available on how developers of survey bots design the bots to accomplish their goals; however, a tutorial on YouTube suggested that survey bots are bespoke and adjustable programming scripts [32]. The number of surveys a survey bot submits and the amount of time it takes to complete each survey can be predetermined by the programming script [32]. While the developers can complete surveys quickly and may do so to maximize the number of incentives received, this quick completion rate is not a rule. If the survey bot developer is motivated by the incentive and believes a short survey completion time will result in the survey bot's detection, they can slow the pace of the survey bot [32]. Furthermore, survey bots can answer different question types, including multiple-choice and open-ended response questionnaire items [32]. For the latter, survey bots can access another service that generates random sentences that are used as a bank of answers for open-ended questionnaire items [32]; however, developers could elect to provide the survey bot with words and sentences that are better aligned with the survey topic to evade detection. Survey bots appear to be an increasingly relevant threat [33]; however, there is a lack of clarity on whether the increasing threat is related to improved reporting, increased prevalence of the issue, or both. Surveys that offer financial incentives are more likely to be targeted by survey bots, although surveys without incentives can also be targeted for practice [33].

XSL•FO
RenderX

Survey Bots: Mitigation Strategies and Their Benefits and Limitations

CAPTCHA is a web-based tool meant to prevent bots from accessing web pages, web services, and more [34-36]. This tool acts as a gatekeeper by requiring the completion of a task that is easy for humans but difficult for bots [34-36]. The earliest versions involved identifying distorted alphanumeric characters and submitting them in a text field [35,36]. By 2014, machine learning (advanced mathematical) algorithms had become advanced enough that they could solve alphanumeric CAPTCHAs better than humans, which led to image-based CAPTCHAs (eg, selecting all the pictures that contain traffic lights) [35]. However, bots have begun to solve image-based CAPTCHAs at a high rate as well. Sivakorn et al [37] conducted an experiment and found an approach that successfully solved image-based CAPTCHAs more than 70.8% of the time. Furthermore, marketing for CAPTCHA solving services in 2023 suggests such services can use a mix of automated processes (ie, optical character recognition) and humans to solve any CAPTCHA at a 90% to 100% success rate for about US $0.50-US $5.00 per 1000 CAPTCHAs [38-40]. CAPTCHA-solving services also offer application programming interfaces that can be integrated into a survey bot's script, which enables communications between the survey bot and the CAPTCHA-solving service [38-41]. These findings and available services suggest CAPTCHA is a solvable hurdle for survey bot developers and is not sufficient to prevent survey bots from completing multiple survey submissions, yet CAPTCHA plausibly can deter some developers of survey bots due to the added planning and programming time and modest costs required to circumvent it.

Survey Bots: Anonymity Strategies Applied in Future Survey Attempts

Despite CAPTCHA's limitations and the previously discussed strategies for reducing anonymity, we used REDCap's CAPTCHA task in conjunction with additional checks, such as manually reviewing open-ended responses for repetitious words and phrases and requiring the submission of authentic photos of participants' most used e-cigarette device. Additional discussion about these strategic choices can be found below in the "Staying Vigilant" section.

Other Vulnerabilities and Challenges

Other vulnerabilities include lengthy data collection periods that may allow participants more time to devise a strategy for submitting fraudulent surveys, the ability to learn the screening questions and questionnaire easily, nonrandom survey URL text generation, and payment of survey incentives without adequate review of survey submission data quality. Although not directly a vulnerability, one additional challenge is institutional review board (IRB) application language that requires amendments and reports each time a new threat or vulnerability is uncovered.

Other Strategies Applied in Future Survey Attempts

To mitigate concerns about the duration of the data collection period, we shortened the data collection period as much as possible by increasing the pace of recruitment. Specifically, the catchment area expanded from 3 cities to as many as 404 areas across the United States using Craigslist "jobs" and "gigs" boards. However, a by-product of this decision meant that the puff topography component of the study became a separate and more limited study. To increase the difficulty of learning the screening questions and questionnaire, the "back button" was disabled. Anyone wishing to communicate honest mistakes could notify the research team at the end of the survey in a comment box. Additional mitigation strategies included the randomization of survey URL text, flexible IRB application language that enabled the removal of any participant suspected of misusing the survey, written warnings to participants against survey misuse in the consent form and reminders before the screening questions each wave, and an initial data quality review before incentives were sent to participants. The latter review included monitoring for the use of non-Latin alphabet characters; invalid mailing addresses; multiple submissions from the same individual or household from current or previous waves of data collection; inconsistent responses (eg, age provided in the survey differs from the age provided during registration); minimal number of questionnaire items answered, which was highly unlikely given the skip logic; short completion time; missing attention-checking questions; unusable photos submitted; and remarkably similar open-ended responses across multiple submissions.

Staying Vigilant

Beyond learning a lesson in the importance of planning mitigation strategies for building a web-based original sample, we also learned the value of staying vigilant as a new threat and vulnerability emerged after commencing with wave 1 data collection (ie, the first wave of data collection after the initial survey attempt that failed).

At the start of wave 1, participants still had the optional ability to submit photos of their most used e-cigarette device to allow for the accurate collection of technical details, which can be difficult to self-report. They were intended to be authentic photos of their device rather than images downloaded from the internet. Initially, this worked as expected, with only a handful of survey submissions including images downloaded from the internet, which we attributed to participants not having their devices readily available. However, approximately 2 months into the relaunch of wave 1, there was a sudden increase in device images that appeared to be downloaded from the internet (34 images over 3 days, of which many were identical images). Out of an abundance of caution, we temporarily suspended the delivery of gift codes to investigate the cluster of survey submissions more thoroughly.

Upon closer examination, other issues with these survey submissions were discovered. Notably, open-response answers were remarkably similar, with all the suspicious survey submissions providing nearly identical responses to the inquiry for feedback to improve the survey. These included responses such as "good," "GOOD," "very good," "VERY GOOD," "well," and "very well." In addition, self-report device brand and model were limited to only a few responses (eg, "JUUL," "Vaporesso," "Luxe," "Voopoo," and "Drag"), which aligned

XSL•FO
RenderX

with the examples provided in the questionnaire item. Of further concern, these submissions passed data quality and personal record checks and there were no obvious connections between the individuals who appeared to submit the suspicious surveys (ie, there was no geographic pattern observed from the addresses provided, and reported ages also widely varied).

Discerning precisely how these suspicious survey submissions were submitted and evaded detection is difficult. Plausibly, someone developed a survey bot capable of circumventing the CAPTCHA and submitted multiple surveys, although, equally plausibly, a single individual or a group of individuals who are somehow connected could have repeatedly attempted the survey without the aid of a survey bot [10]. Presuming that the culprits were not a group of individuals operating from a written script, the idea that a survey bot developer or other individuals involved passed the personal record check is perplexing. One possibility is that they read the consent form, recognized we were authenticating participant identities, and used a database with authentic identifying information. While unlikely, we also cannot dismiss the possibility that these data and circumstances were a coincidence.

The motivating factors for the individuals responsible for the suspicious survey submissions remain unclear as well, although the motivation could have been to create chaos or, more likely, financial gain [5,8-10]. Assuming a database with authentic identifying information was used, the individuals responsible would not have received the initial gift code for the baseline survey, as the first gift code was delivered by mail to the address provided. This detail may have been overlooked by the individuals; however, they may have recognized from the consent form that future gift codes would be delivered electronically and could potentially profit from these efforts in subsequent waves. A similarly patient approach to survey fraud has been taught by at least 1 website dedicated to training individuals to fraudulently submit a large volume of surveys on their own: "...just create enough accounts and wait for the right opportunity. You can in just 1 day, make enough money to last you for several months" [10].

Ultimately, we decided that the cluster of submissions was suspicious enough to warrant being dropped from the dataset, and 3 actions were taken to mitigate the chances of recurrence. First, we required photo submissions of each participant's most used device to ensure that all were in possession of an e-cigarette. Second, all photo submissions were reviewed as part of the initial data quality review to ensure they were not professional, downloaded, or screenshot images; duplicates; remarkably similar to photos already submitted (eg, distinctive marks, lighting, and background objects); an object other than an e-cigarette; taken from a store; or inconsistent with the self-reported brand (for brands that were listed as examples in the corresponding questionnaire item only). Finally, all open-ended responses were reviewed once a week as part of the initial data quality review to ensure that repetitive open-ended answers were not submitted. These strategic choices further raised the bar required to deceive our team. For example, developers of survey bots or others attempting to deceive our team in the future would be required to purchase or access e-cigarettes, take authentic photos of the e-cigarettes in different lighting environments, and submit nonrepetitive and logical open-ended responses.

Reflections 4 Years Later

Four years after our unsuccessful attempt to recruit participants for a web-based survey, we successfully created a cohort and completed 5 survey waves of data collection between May 2020 and April 2023 (wave 1: n=1179, wave 2: n=1187, wave 3: n=1219, wave 4: n=1224, and wave 5: n=1290). After implementing additional mitigation strategies in response to the suspicious survey submissions in wave 1, no new threats or vulnerabilities were uncovered (Textbox 2). In addition, although we cannot dismiss the possibility that the final dataset contains a few fraudulent survey submissions, there is little to no evidence that the dataset's integrity has been compromised to a degree that it would bias the sample and analyses. For example, the sample of those who frequently use e-cigarettes primarily consists of people who self-reported daily use (wave 1: 1081/1179, 91.7%; wave 2: 1104/1187, 93.0%; wave 3: 1154/1219, 94.7%; wave 4: 1170/1224, 95.6%; and wave 5: 1234/1290, 95.7%), which is demographically similar to respondents using e-cigarettes daily from the 2019 Tobacco Use Supplement to the Current Population Survey [42], with the exceptions that participants in the VAPER Study sample were relatively younger and had lower incomes. Moreover, often the VAPER Study findings have been consistent with the broader literature and policy context. For example, participants' most used flavor for their most used device included a wide range of flavors for all device types except disposable pod devices (eg, JUUL). Specifically, 93.4% (156/167) of those mostly using disposable pod devices during wave 5 indicated that their most used flavor was tobacco or menthol, which largely is consistent with expectations given a US Food and Drug Administration regulation that only allows the sale of tobacco- and menthol-flavored pods [43,44].

XSL•FO
RenderX

Textbox 2. Risk mitigation strategies used in the VAPER (Vaping and Patterns of E-cigarette Use Research) Study, waves 1-5.

Presurvey strategies:

- Requested flexible institutional review board protocols for easy removal of participants suspected of misuse

In-survey strategies:

- Authenticated cell phone numbers
- Collected IP addresses
- Used CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart)
- Shortened the data collection period
- Generated random URLs for accessing the survey
- No back button
- Required photo submissions of their most used e-cigarette
- Used written warnings to warn against misuse in the consent and before the screening questions

Postsurvey strategies:

- Authenticated names, dates of birth, and residential mailing addresses
- Requested identification cards and utility bills, as needed
- Mailed incentives to physical addresses
- Completed an initial data quality review prior to sending incentives

The study protocols, inclusive of the mitigation strategies, were adequately scalable and cost-effective given the project resources available [14]. One of the more significant expenses unique to studies with an original sample was the Craigslist postings to advertise the survey. These costs peaked at US $53 per valid participant in wave 2 (covering 150 Craigslist areas over 4 months to replace those lost to follow-up). Costs decreased to US $6-US $10 per valid participant in waves 3-5 (covering 404 Craigslist areas over 2-2.5 months to replace those lost to follow-up). These costs decreased in response to the number of Craigslist areas increasing (individual Craigslist areas can become more depleted over a period of a few months), a larger proportion of the posts being on the “gigs” boards (these posts are cheaper than those on the “jobs” boards), and improved reposting schedules (more populated areas require more frequent reposting). In terms of mitigation strategies, several were implemented at minimal or no cost, as follows: (1) authenticating phone numbers, (2) using CAPTCHA, (3) generating random URLs for survey access, (4) prohibiting the use of the back button, (5) requesting flexible IRB protocols to easily remove participants suspected of misuse, (6) including written warnings against misuse in the consent and before the screening questions, and (7) requiring photo submissions of participants’ most used e-cigarettes. Other mitigation strategies entailed significantly higher costs in terms of subscription fees. For example, an annual subscription to LexisNexis’ personal record verification platform costs US $4876 for access by 3 users, as of March 2020 (prices may vary and could be outdated). The most substantial and difficult-to-quantify cost was the staff effort required to (1) post and repost Craigslist postings, (2) use the LexisNexis’ personal record verification platform, (3) follow up with participants who could not be authenticated using LexisNexis’ platform, (4) conduct an initial data quality review prior to sending incentives, (5) mail incentives to physical addresses, and (6) conduct data cleaning and analysis. Moreover, staff effort was essential in planning mitigation strategies ahead of data collection and swiftly implementing postsubmission mitigation strategies to ensure a positive participant experience. For example, after our investigation into the initial survey attempt that failed in 2019, preparation and ethics approval for the relaunched wave 1 took approximately 9 months, and starting with wave 1, all incentives were sent within 2 weeks of survey submission to participants who had their identities authenticated and passed the initial data quality checks.

As previously highlighted by our experiences and the learnings from the literature, the specific tactics used by individuals attempting to submit fraudulent surveys can be dependent upon mitigation strategies in place, but, likely, study design and the intended population (eg, lower prevalence vs higher prevalence) play important roles as well. The methods and array of related strategies highlighted in this paper were successful for the VAPER Study’s study design and population, but a similar approach may be insufficient for other study designs and populations and may be excessive for others. For instance, we initially paid participants US $10 to complete a 20-minute survey and included screening questions intended to only include a lower prevalence population. For the financially motivated, this payment schedule may have provided a financial incentive to learn the screening questions through multiple attempts in order to complete one or multiple fraudulent survey submissions. This possibility is consistent with the work of Bowen et al [8], who found that participants eligible for an incentive were 6 times more likely to submit more than one survey. Furthermore, general population studies likely will recruit participants more quickly than studies recruiting a lower prevalence population, resulting in less time to develop bespoke survey bots or to

XSL•FO
RenderX

implement other strategies. Thus, general population studies may encounter fewer threats, although this is an empirical question that awaits further investigation.

Nearly all the mitigation strategies implemented in the VAPER Study can be used by researchers in other medical and public health fields seeking to strengthen their web-based original samples, resources permitting. However, requiring photo submissions is not an appropriate strategy for all studies. For the VAPER Study, the required submission of device photos provided us with added confidence that the survey submissions came from the target population. A similar strategy can be used for other web-based surveys that recruit populations likely in possession of an item, such as people who frequently use combustible tobacco products or alcohol (eg, photos of product packaging), who take a specific medication (eg, photos of the bottle containing the medication), or who work in a specific occupation (eg, photos of an identification badge from their employer). The strategy is less appropriate in instances where the target population is less likely to be able to submit photos of the desired item. For example, people who infrequently use combustible tobacco products or alcohol may not be in possession of a product package at the time of the survey. Additionally, this strategy is not applicable when there is no physical item uniquely associated with all individuals in the target population (eg, members of a religious group). In such cases, open-ended questions that solicit information likely to be commonly known by the target population—but not by others—such as colloquial neighborhood names in city-specific studies, may be a useful alternative [45]. However, the efficacy of this approach remains uncertain, as specific information commonly known to a target population can still be searched for by those attempting to submit fraudulent surveys.

While we are satisfied with the integrity of the VAPER Study dataset, we recognize that the recruitment and mitigation strategies used likely introduced at least some sampling bias. Regarding recruitment strategies, the sample was drawn from the internet, where use is lower among individuals aged 65 years and older (88%) compared with younger age groups (96%-98%) [46]. Additionally, our web-based recruitment predominantly targeted the “jobs” and “gigs” boards on Craigslist, where individuals are often seeking income-generating opportunities. We suspect that these factors likely contributed to a younger, lower-income sample compared with the Tobacco Use Supplement to the Current Population Survey. Recruitment from other websites or social media platforms may have yielded a sample with a different level of representativeness. Address-based sampling using mailed survey invitations with an access code may have resulted in a more representative sample and should be considered for future studies [47]. In terms of mitigation strategies, privacy concerns may have deterred some individuals from participating due to potential data security risks, which are inherent to both in-person and web-based studies that require identifiable information for enrollment. Furthermore, American adults who could not write in languages using the Latin alphabet were excluded given our experience with the failed survey attempt in 2019.

Overall, web-based recruitment and web-based survey data collection proved to be more challenging than we anticipated, but much has been learned from our initial survey attempt that failed and subsequent successes. As a result, we have reasonable confidence in the integrity of the data and a stronger understanding of their strengths and limitations. We encourage other researchers conducting web-based surveys to consider the threats and vulnerabilities that could impact their data integrity, to incorporate appropriate mitigation strategies, and to develop effective monitoring procedures. In addition, we encourage more researchers to report their challenges and successes and to report additional details about their web-based survey methodologies for mitigating the submission of fraudulent surveys and preserving data integrity in their manuscripts.

Acknowledgments

The authors thank Kevin Welding for his contributions to the investigation of the 2019 survey data. Research reported in this manuscript was supported by the National Institute on Drug Abuse and the Food and Drug Administration’s Center for Tobacco Products under award U54DA036105. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Food and Drug Administration. This manuscript has not been reviewed by the National Institute on Drug Abuse or the Food and Drug Administration’s Center for Tobacco Products.

Data Availability

The datasets for waves 1-5 generated during this study are available from the corresponding author on reasonable request and after a review of the proposed research and qualifications of the applicant.

Authors' Contributions

JJH wrote the first draft of the manuscript. Subsequent drafts included contributions from EC, JKS, QN, AB, TE, KW, RDK, and JEC. The final text was approved by all authors.

Conflicts of Interest

TE is a paid consultant in litigation against the tobacco industry; has been a paid consultant in litigation against the electronic cigarette industry; and is named on one patent for a device that measures the puffing behavior of electronic cigarette users, on

XSL•FO
RenderX

another patent application for a smartphone app that determines electronic cigarette device and liquid characteristics, and on a third patent application for a smoking cessation intervention. JEC was a paid consultant in litigation against a tobacco company.

References

1. Evans JR, Mathur A. The value of online surveys: a look back and a look ahead. Internet Res 2018;28(4):854-887. [doi: 10.1108/intr-03-2018-0089]
2. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
3. Bell AM, Gift T. Fraud in online surveys: evidence from a nonprobability, subpopulation sample. J Exp Polit Sci 2023;10(1):148-153. [doi: 10.1017/xps.2022.8]
4. Guillory J, Kim A, Murphy J, Bradfield B, Nonnemaker J, Hsieh Y. Comparing twitter and online panels for survey recruitment of e-cigarette users and smokers. J Med Internet Res 2016;18(11):e288 [FREE Full text] [doi: 10.2196/jmir.6326] [Medline: 27847353]
5. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133 [FREE Full text] [doi: 10.1111/jlme.12200] [Medline: 25846043]
6. Pratt-Chapman M, Moses J, Arem H. Strategies for the identification and prevention of survey fraud: data analysis of a web-based survey. JMIR Cancer 2021 Jul 16;7(3):e30730-e30133 [FREE Full text] [doi: 10.2196/30730] [Medline: 34269685]
7. Sterzing PR, Gartner RE, McGeough BL. Conducting anonymous, incentivized, online surveys with sexual and gender minority adolescents: lessons learned from a national polyvictimization study. J Interpers Violence 2018;33(5):740-761. [doi: 10.1177/0886260517744845] [Medline: 29295005]
8. Bowen AM, Daniel CM, Williams ML, Baird GL. Identifying multiple submissions in internet research: preserving data integrity. AIDS Behav 2008;12(6):964-973 [FREE Full text] [doi: 10.1007/s10461-007-9352-2] [Medline: 18240015]
9. Reips UD. Standards for internet-based experimenting. Exp Psychol 2002;49(4):243-256. [doi: 10.1026//1618-3169.49.4.243] [Medline: 12455331]
10. Market research fraud: distributed survey farms exposed. Greenbook. 2019. URL: https://www.greenbook.org/insights/insights-technology/market-research-fraud-distributed-survey-farms-exposed [accessed 2023-11-10]
11. Pozzar R, Hammer MJ, Underhill-Blazey M, Wright AA, Tulsky JA, Hong F, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020;22(10):e23021 [FREE Full text] [doi: 10.2196/23021] [Medline: 33026360]
12. Griffin M, Martino RJ, LoSchiavo C, Comer-Carruthers C, Krause KD, Stults CB, et al. Ensuring survey research data integrity in the era of internet bots. Qual Quant 2022;56(4):2841-2852 [FREE Full text] [doi: 10.1007/s11135-021-01252-1] [Medline: 34629553]
13. Security survey options. Qualtrics. URL: https://www.qualtrics.com/support/survey-platform/survey-module/survey-options/survey-protection/ [accessed 2024-04-24]
14. Hardesty JJ, Crespi E, Nian Q, Sinamo JK, Breland AB, Eissenberg T, et al. The vaping and patterns of e-cigarette use research study: protocol for a web-based cohort study. JMIR Res Protoc 2023;12:e38732 [FREE Full text] [doi: 10.2196/38732] [Medline: 36862467]
15. Boakye E, Osuji N, Erhabor J, Obisesan O, Osei AD, Mirbolouk M, et al. Assessment of patterns in e-cigarette use among adults in the US, 2017-2020. JAMA Netw Open 2022;5(7):e2223266 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.23266] [Medline: 35867060]
16. Hoft J. Anonymity and confidentiality. In: The Encyclopedia of Research Methods in Criminology and Criminal Justice. Hoboken, NJ: John Wiley & Sons, Inc; 2021:223-227.
17. Goodrich B, Fenton M, Penn J, Bovay J, Mountain T. Battling bots: experiences and strategies to mitigate fraudulent responses in online surveys. Appl Econ Perspect Pol 2023;45(2):762-784. [doi: 10.1002/aepp.13353]
18. ElBahrawy A, Alessandretti L, Rusnac L, Goldsmith D, Teytelboym A, Baronchelli A. Collective dynamics of dark web marketplaces. Sci Rep 2020;10(1):18827 [FREE Full text] [doi: 10.1038/s41598-020-74416-y] [Medline: 33139743]
19. Dennis SA, Goodson BM, Pearson CA. Online worker fraud and evolving threats to the integrity of MTurk data: a discussion of virtual private servers and the limitations of IP-based screening procedures. Behav Res Account 2020;32(1):119-134. [doi: 10.2308/bria-18-044]
20. Hohn KL, Braswell AA, DeVita JM. Preventing and protecting against internet research fraud in anonymous web-based research: protocol for the development and implementation of an anonymous web-based data integrity plan. JMIR Res Protoc 2022;11(9):e38550 [FREE Full text] [doi: 10.2196/38550] [Medline: 36094806]
21. Globyt E. NordVPN survey shows: a third of Americans use a VPN. NordVPN. URL: https://nordvpn.com/blog/nordvpn-usage-survey-us/ [accessed 2023-11-07]
22. About Tor Browser. Tor Project. URL: https://tb-manual.torproject.org/about/ [accessed 2023-11-07]
23. Baluch A. How to change your IP address (4 different ways in 2024). Forbes. URL: https://www.forbes.com/advisor/business/software/how-to-change-ip-address/ [accessed 2023-11-02]

XSL•FO
RenderX

24. Why does your IP address change now and then. WhatIsMyIPAddress. URL: https://whatismyipaddress.com/keeps-changing [accessed 2023-11-07]
25. Spector L. Your mobile IP address: its safety is one thing, its privacy is another. PCworld. 2015. URL: https://www.pcworld.com/article/422808/your-mobile-ip-address-its-safety-is-one-thing-its-privacy-is-another.html [accessed 2023-11-14]
26. Effective protection. Proxycheck.io. URL: https://proxycheck.io/ [accessed 2024-10-28]
27. Detecting VPN usage: what you need to know. Forest VPN. 2024. URL: https://forestvpn.com/blog/internet-privacy/how-to-tell-if-a-vpn-is-being-used/ [accessed 2024-10-28]
28. Hauk C. How to spoof your location and get around geo-blocking with a VPN. Pixelprivacy. 2023. URL: https://pixelprivacy.com/vpn/spoof-location-geo-blocking/ [accessed 2024-10-28]
29. Get your dedicated IP. NordVPN. URL: https://nordvpn.com/offer/dedicated-ip/?vpn=generic&nc=G-SR-US-EN-VPN-SRC-Device-Product-17/48-D-(WW)-USD-[Broad]-&ns=google&nm=cpc&nt=do%20i%20need%20dedicated%20ip%20vpn&gad_source=1 [accessed 2024-10-28]
30. What is a bot? | bot definition. Cloudflare. URL: https://www.cloudflare.com/learning/bots/what-is-a-bot/ [accessed 2023-11-14]
31. Hallberg L. Understanding survey bots and tools for data validation: strategies for identifying possibly fraudulent responses. The University of Kansas. 2022. URL: https://lifespan.ku.edu/online-surveys-and-data-collection-tools [accessed 2023-11-14]
32. Jeremy H. Automated bot to take CollegePulse surveys. Youtube. 2019. URL: https://www.youtube.com/watch?v=QbEZDLqxWPc [accessed 2023-11-14]
33. Xu Y, Pace S, Kim J, Iachini A, King LB, Harrison T, et al. Threats to online surveys: recognizing, detecting, and preventing survey bots. Soc Work Res 2022;46(4):343-350. [doi: 10.1093/swr/svac023]
34. Noury Z, Rezaei M. Deep-CAPTCHA: a deep learning based CAPTCHA solver for vulnerability assessment. arXiv. Preprint posted online on June 15, 2020 [FREE Full text] [doi: 10.48550/arXiv.2006.08296]
35. Why captchas are getting harder. Vox YouTube page. 2021 May 14. URL: https://www.youtube.com/watch?v=lUTvB1O8eEg [accessed 2023-11-14]
36. von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: human-based character recognition via web security measures. Science 2008;321(5895):1465-1468. [doi: 10.1126/science.1160379] [Medline: 18703711]
37. Sivakorn S, Polakis J, Keromytis A. I'm not a human: breaking the Google reCAPTCHA. Black Hat. 2016. URL: https://www.blackhat.com/docs/asia-16/materials/asia-16-Sivakorn-Im-Not-a-Human-Breaking-the-Google-reCAPTCHA-wp.pdf [accessed 2023-11-14]
38. Captcha solving service. 2captcha. URL: https://2captcha.com/ [accessed 2023-11-14]
39. Best CAPTCHA solver bypass service. Death by Captcha. URL: https://deathbycaptcha.com/ [accessed 2023-11-14]
40. CAPTCHA solving service. Anti Captcha. URL: https://anti-captcha.com/ [accessed 2023-11-14]
41. How to bypass reCAPTCHA with puppeteer and headless chrome. Jarrod Overson Youtube page. 2019 Nov 6. URL: https://www.youtube.com/watch?v=wsDRkAD6lPs [accessed 2023-11-14]
42. Tobacco use supplement to the current population survey harmonized data, 1992-2019. National Cancer Institute. 2019. URL: https://cancercontrol.cancer.gov/brp/tcrb/tus-cps [accessed 2023-11-14]
43. Ali FRM, Seidenberg AB, Crane E, Seaman E, Tynan MA, Marynak K. E-cigarette unit sales by product and flavor type, and top-selling brands, United States, 2020-2022. MMWR Morb Mortal Wkly Rep 2023;72(25):672-677 [FREE Full text] [doi: 10.15585/mmwr.mm7225a1] [Medline: 37347717]
44. FDA finalizes enforcement policy on unauthorized flavored cartridge-based e-cigarettes that appeal to children, including fruit and mint. Food and Drug Administration. 2020. URL: https://www.fda.gov/news-events/press-announcements/fda-finalizes-enforcement-policy-unauthorized-flavored-cartridge-based-e-cigarettes-appeal-children [accessed 2023-11-06]
45. Bonett S, Lin W, Sexton Topper P, Wolfe J, Golinkoff J, Deshpande A, et al. Assessing and improving data integrity in web-based surveys: comparison of fraud detection systems in a COVID-19 study. JMIR Form Res 2024;8:e47091 [FREE Full text] [doi: 10.2196/47091] [Medline: 38214962]
46. Petrosyan A. Share of adults in the United States who use the internet as of september 2023, by age group. Statista. 2024. URL: https://www.statista.com/statistics/266587/percentage-of-internet-users-by-age-groups-in-the-us/#:~:text=As%20of%202023%2C%2098%20percent,18%20and%2029%20years%20old [accessed 2024-10-28]
47. Understanding America study. USC Dornsife: Center for Economic and Social Research. URL: https://uasdata.usc.edu/index.php [accessed 2024-10-28]

Abbreviations

CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart
IRB: institutional review board
LGBTQ+: lesbian, gay, bisexual, transgender, queer
PO: post office
REDCap: Research Electronic Data Capture
VAPER: Vaping and Patterns of E-cigarette Use Research

XSL•FO
RenderX

VPN: virtual private network

Edited by A Coristine; submitted 03.05.24; peer-reviewed by Y Duan, ER Park; comments to author 18.06.24; revised version received 31.07.24; accepted 28.10.24; published 16.12.24.

Please cite as:
Hardesty JJ, Crespi E, Sinamo JK, Nian Q, Breland A, Eissenberg T, Kennedy RD, Cohen JE
From Doubt to Confidence—Overcoming Fraudulent Submissions by Bots and Other Takers of a Web-Based Survey
J Med Internet Res 2024;26:e60184
URL: https://www.jmir.org/2024/1/e60184
doi:10.2196/60184
PMID:

©Jeffrey J Hardesty, Elizabeth Crespi, Joshua K Sinamo, Qinghua Nian, Alison Breland, Thomas Eissenberg, Ryan David Kennedy, Joanna E Cohen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Selecting an Ecological Momentary Assessment Platform: Tutorial for Researchers

Lauren M Henry[1], PhD; Eleanor Hansen[1], BA; Justin Chimoff[2], BS; Kimberly Pokstis[2], BS; Miryam Kiderman[1], PsyD; Reut Naim[3], PhD; Joe Kossowsky[2], MMSc, PhD; Meghan E Byrne[1], PhD; Silvia Lopez-Guzman[1], MD, PhD; Katharina Kircanski[1], PhD; Daniel S Pine[1], MD; Melissa A Brotman[1], PhD

[1]Emotion and Development Branch, National Institute of Mental Health, Bethesda, MD, United States

[2]Department of Anesthesiology, Critical Care and Pain Medicine Research, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States

[3]The School of Psychological Sciences, Tel-Aviv University, Tel-Aviv, Israel

Corresponding Author:
Lauren M Henry, PhD
Emotion and Development Branch, National Institute of Mental Health
9000 Rockville Pike, Building 15K
Bethesda, MD, 20892
United States
Phone: 1 301 480 3895
Email: lauren.henry@nih.gov

Abstract

Background: Although ecological momentary assessment (EMA) has been applied in psychological research for decades, delivery methods have evolved with the proliferation of digital technology. Technological advances have engendered opportunities for enhanced accessibility, convenience, measurement precision, and integration with wearable sensors. Notwithstanding, researchers must navigate novel complexities in EMA research design and implementation.

Objective: In this paper, we aimed to provide guidance on platform selection for clinical scientists launching EMA studies.

Methods: Our team includes diverse specialties in child and adolescent behavioral and mental health with varying expertise on EMA platforms (eg, users and developers). We (2 research sites) evaluated EMA platforms with the goal of identifying the platform or platforms with the best fit for our research. We created a list of extant EMA platforms; conducted a web-based review; considered institutional security, privacy, and data management requirements; met with developers; and evaluated each of the candidate EMA platforms for 1 week.

Results: We selected 2 different EMA platforms, rather than a single platform, for use at our 2 research sites. Our results underscore the importance of platform selection driven by individualized and prioritized laboratory needs; there is no single, ideal platform for EMA researchers. In addition, our project generated 11 considerations for researchers in selecting an EMA platform: (1) location; (2) developer involvement; (3) sample characteristics; (4) onboarding; (5) survey design features; (6) sampling scheme and scheduling; (7) viewing results; (8) dashboards; (9) security, privacy, and data management; (10) pricing and cost structure; and (11) future directions. Furthermore, our project yielded a suggested timeline for the EMA platform selection process.

Conclusions: This study will guide scientists initiating studies using EMA, an in vivo, real-time research tool with tremendous promise for facilitating advances in psychological assessment and intervention.

(J Med Internet Res 2024;26:e51125) doi:10.2196/51125

KEYWORDS

ecological momentary assessment; methodology; psychology and psychiatry; child and adolescent; in vivo and real time

XSL•FO
RenderX

Introduction

Overview

Ecological momentary assessment (EMA), a tool for collecting naturalistic data in real time, has been used in psychological science for decades [1]. In the past 10 years, the use of EMA in published scientific research has increased by 168% (Multimedia Appendix 1 [2]). Surging interest is justifiable; for example, EMA can enhance ecological validity and support the examination of within-person dynamics [1]. EMA delivery methods have evolved with technological advances [3], and leveraging mobile devices garners additional benefits, including greater accessibility and convenience, increased measurement precision, and an opportunity for the integration of passive sensing [3-5]. A proliferation of developers and platforms must be navigated to select suitable equipment [5-8]. Here, we provide guidance on platform selection for researchers launching EMA studies.

What Is EMA?

The term EMA was coined in the 1990s to describe "...monitoring and sampling strategies to assess phenomena at the moment they occur in natural settings..." [9]. EMA is an idiographic approach. In contrast to nomothetic research, which focuses on principles that may generalize more broadly, ideographic research examines individual-, event-, and time-based idiosyncrasies [10]. EMA is characterized by (1) repeated and (2) real-time assessments in (3) naturalistic settings [1]. Phenomena of interest are measured dynamically (in contrast to statically) as they unfold (in contrast to retrospectively) and in the contexts in which they occur naturally (in contrast to the clinic, laboratory, or another location). Long-standing traditions of repeated, random sampling [11] and self-report diaries capturing naturalistic data [12] are among the bases for EMA methods.

Historically, EMA was first administered through paper-and-pencil diaries with technology available before the 21st century [1]. Currently, smartphones and tablets are the primary method for EMA delivery. Burgeoning advances have facilitated the integration of EMA with wearable devices and passive sensors, generating further naturalistic data [5]. EMA is an especially promising method for mental and behavioral health research [1]. Accordingly, EMA has been implemented across disciplines, including addiction [13], depression and anxiety [14,15], neurodevelopmental disorders [16], personality disorders [17], disruptive behavior disorders [18], eating disorders [19], schizophrenia [20], suicide [21], and chronic pain [22].

Why Use EMA?

Self-report and other-report (eg, parent and teacher) questionnaires and laboratory-based methods are useful but limited. Questionnaires rely on retrospection, which is prone to error and may be systematically impacted by recall biases [23], and laboratory-based methods are high in internal validity but may not approximate real life [24]. Comparatively, EMA facilitates the collection of data with high ecological validity, allowing for generalization to authentic and natural individual experiences [25]. By tracing change across time relative to each individual, EMA reveals temporal associations that are inaccessible with less granular methods [24,26]. EMA has also been shown to be useful in examining dynamic clinical phenomena that are otherwise difficult to track [25,27,28].

EMA facilitates the transition from assessment to intervention. Ecological momentary interventions (EMIs), and relatedly just-in-time adaptive interventions (JITAIs), capitalize on EMA, wearable devices, and mobile device–based passive sensors to provide in-situ personalized support by detecting and adapting to individual internal and contextual states [29,30]. Identifying specific factors (eg, negative affect) that temporally precede intervention targets (eg, self-harm) via EMA can lead to novel interventions deployed in the moment that they are most needed (eg, prompting use of coping skills) via EMI or JITAI [1].

Mobile device–delivered EMA and EMI or JITAI may help increase access to mental health assessments and interventions at scale, respectively. Geographic, racial, ethnic, and socioeconomic disparities, as well as a dearth of providers, severely limit access to in-person psychological assessment and intervention [31]. Importantly, 97% of Americans report cellular phone ownership and 85% report smartphone ownership [32], which is nearly universal across gender, race, ethnicity, socioeconomic status, and aspects of psychological health [33-35]. Furthermore, internet access is expanding rapidly in low- and middle-income countries [36], providing promise for a global health impact. Accordingly, technology-facilitated assessment, intervention, and monitoring may ultimately prove to be more cost-effective and scalable than in-person alternatives [5].

This Research

As mentioned earlier, research leveraging EMA has rapidly increased since its inception (Multimedia Appendix 1). At the time of writing, at least 63 EMA platforms were available for research use, and at least 12 research groups have developed their own platforms (Multimedia Appendix 2 [37-48]). Researchers may benefit from direct guidance on selecting an EMA platform, given the quantity of platforms available and researchers who have dedicated resources to developing their own platforms. At the National Institute of Mental Health (NIMH) and Boston Children's Hospital (BCH), we are engaged in individual and overlapping research programs using EMA. Recognizing the unmet needs of our laboratories by our existing EMA platforms, we collaborated to select our next EMA platform.

Methods

Purpose and Approach

Team Experience With EMA Platforms

This project was the product of a collaboration between the NIMH and BCH. We represent diverse specialties in child and adolescent behavioral and mental health (irritability, chronic pain, and substance use) and a range of experiences with EMA platforms (users and developers).

XSL•FO
RenderX

Team at the NIMH

The focus of our research is the identification of treatment targets for psychiatric disorders and the development and evaluation of innovative, scalable, and cost-effective interventions for children and adolescents (ages 8-17 years). We have experience as users of multiple text-based and app-based EMA platforms, as well as experience as developers of "RATE-IT," an app-based EMA platform for use in our own research studies [44]. We use EMA in isolation, as well as EMA paired with wearable devices (eg, heart rate monitors and accelerometers) [49-51]. We also have experience developing and evaluating technology-based solutions more broadly, including through assessment (eg, smartphone app–based inhibitory control assessment [52]) and intervention (ie, computer-based interpretation bias training targeting face emotion processing [53]).

Team at BCH

We focus on the development, validation, and implementation of mobile health solutions to assess outcomes reliably and consistently in individuals aged 7 to 24 years with chronic and acute pain, mental health, and substance use disorders. We use EMA [54,55], wearables [56,57], digital phenotyping [58], and the internet [59] to assess behavioral patterns, sleep, pain, physical mobility, motor activity, energy expenditure, and cognitive functioning. In turn, we use these constructs to evaluate and predict clinical and treatment progression, as well as to provide smartphone-based treatments. Methodologically, we have tested methods for improving adherence and retention in longitudinal mobile health studies [60].

Procedure

We used a 5-step process to select an EMA platform (Figure 1). First, we developed a list of platform features (Textbox 1) and individualized and prioritized lists of laboratory-based needs (Textbox 2). Second, we created a list of extant platforms (Multimedia Appendix 2), conducted a web-based review to determine platform capabilities (ie, to the best of our ability, we answered questions about platform features from Textbox 1 and noted platform capabilities relative to our individualized and prioritized lists of laboratory-based needs in Textbox 2), and selected 7 EMA platforms. Third, we met with our data security officer and corresponded with our institutional review board (IRB) to determine the institutional security, privacy, and data management requirements at each of our sites (Multimedia Appendix 2). Fourth, we met with 7 platform developers to confirm our understanding of platform features; inquire about features not explicitly or clearly described on developer websites; and answer our respective institutional security, privacy, and data management questions. Fifth, we evaluated each of the 7 candidate platforms for 1 week (mean 5.00, SD 3.61 laboratory members per platform) using a combined subset of items from our laboratories.

Of note, initially, we created our list of extant EMA platforms through our team's preexisting research-based experiences with EMA platforms; correspondence with collaborators and colleagues, including word of mouth and social media; and multiple web-based searches. For the purpose of this paper, we subsequently supplemented our initial search with data extracted by an independent rater using a search strategy implemented in a forthcoming systematic review on adherence in pediatric EMA studies (LM Henry, unpublished data, June 2023). See Multimedia Appendix 2 for the search strategy and terms.

XSL•FO
RenderX

Figure 1. Timeline and recommendations for researchers beginning the ecological momentary assessment (EMA) platform section process. IRB: institutional review board.

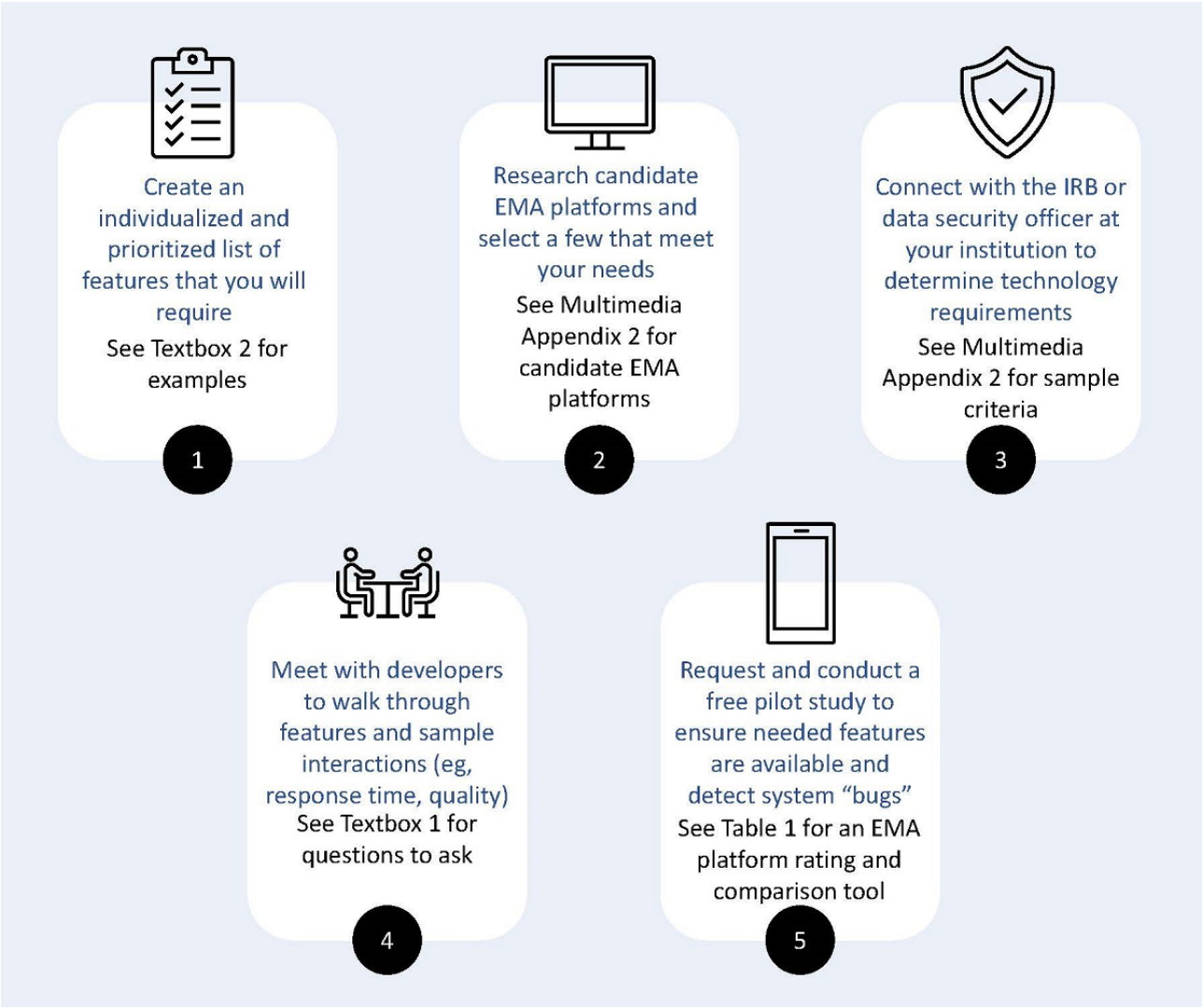

XSL•FO
RenderX

Textbox 1. Ecological momentary assessment platform features and developer question examples.

Alarms

- How are alarms delivered? (eg, phone calls, SMS text messages, and push notifications)
- Is the alarm design customizable? (eg, text [content, font, emphasis, color, size] and sound [type or vibration])
- Are the alarm duration and frequency programable?
- Are there options for users to delay or dismiss an alarm? Can that feature be disabled?

Compatibility

- With which operating systems are the platform compatible? (eg, Android and iOS)

Consent

- Is there an option or requirement for e-consent? Are e-consent forms fully or partially customizable?

Cost structure

- How is cost calculated? (eg, sample size; number of users, surveys, or items; length of study; scope of deployment [academic vs government; multisite vs single site]; extra features [integration with wearables]; type of data collected [text, audio, video]; or some combination of these factors)
- Is there an annual subscription, and how is it renewed? Are there setup fees?
- What costs are associated with developer support and consultation?
- Are discounts available? (eg, for multiyear subscriptions and student projects)

Customization

- Can nonessential features be removed (eg, from the menu or navigation)?
- Can you build out features that are not presently included in the platform?
- Do white label opportunities exist?

Dashboards

- Is a research dashboard available? If so, what are its capabilities, and is it customizable? (eg, create surveys, develop and implement a sampling scheme, message participants, track compliance, visualize data, search data)
- Is a clinician dashboard available? If so, what are its capabilities, and is it customizable? (eg, track responses, flag concerning responses, and message participants)

Downloading results

- When and how are user results uploaded? (eg, manually, upon Wi-Fi connection, or immediately)
- In what file format are user results downloaded?
- What information is provided in data exports? (eg, response time, missing data, compliance, and summary statistics) What structure are the data exports? (eg, long vs wide format) Is the output customizable?
- Can data be exported on a user-by-user basis? Can data be exported for the full sample?

Dyadic interactions

- Is there a capability for dyadic or family-level data to be linked? (eg, survey triggers based on dyadic responding or joint location)

Gamification

- Are there opportunities for goal setting, generic or individualized compliance feedback, rewards, badges, “streaks”?

Item display

- What options are available for item display? (eg, text, image, audio, video, and stopwatch)

Item response

- What options are available for item response? (eg, select one, select many, visual analogue, free text, image, audio, video, and geotag)

Language capabilities

XSL•FO
RenderX

- With which languages are the platform compatible, including infrastructure, interface, surveys, and responses?
- Can the platform automatically recognize language?

Level of support from developers

- Are user guides provided? (eg, written manuals and videos)
- Is developer support required or offered to facilitate setup of, and modifications to, the study?
- How are developers contacted for support? (eg, email, phone call, messaging, and website support ticket)
- Is there a direct point of contact for the study?
- In what timeframe are support requests fulfilled? (eg, in the moment, 1 business day, or 1 week)

Messaging

- Is there a feature that enables real-time or scheduled messaging with users during the study? (eg, general study updates, check-ins for technical difficulties)
- Are there capabilities for triggered notifications to principal investigators or clinicians for concerning responses to critical items?

Onboarding

- Can onboarding occur remotely or virtually, or is onboarding optimized for in-person setup?
- What method(s) are available for onboarding? (eg, email, SMS text messages, QR code, and website link)

Passive sensing and wearable integration

- Is wearable integration offered? If so, with which wearables can the platform pair?
- Is passive sensing offered? (eg, accelerometer, geolocation, and screen time)

Sampling scheme and scheduling

- What sampling schemes are supported? (eg, interval contingent, signal contingent, and event contingent)
- Can multiple forms be created?
- What are the branching logic capabilities?
- Can item responses or passive sensing trigger new surveys?
- Can surveys be deployed and completed while offline? (ie, are data and Wi-Fi required for functionality)

XSL•FO
RenderX

Textbox 2. Individualized and prioritized laboratory-based needs for an ecological momentary assessment (EMA) platform for teams at the National Institute of Mental Health and Boston Children's Hospital.

The National Institute of Mental Health (NIMH)

1. Meets NIMH security requirements and high likelihood of institutional review board approval
2. Child and adolescent friendly
3. Clinical and research dashboards
4. Language: compatibility with English and Hebrew
5. Self-programming option
6. Thermometer item display option and visual analog scale with midpoint item response option
7. Wearable integration
8. Passive sensing
9. Dyadic interactions

Boston Children's Hospital

1. High likelihood of institutional review board approval
2. Child and adolescent friendly
3. Direct messaging with participants
4. Wearable integration
5. Research dashboard
6. Price
7. Support from developer

Results

Overview

Our teams at the NIMH and BCH selected 2 different platforms as the product of our evaluation. We believe that our personalized choices reflect the importance of platform selection driven by individualized and prioritized laboratory needs rather than a single platform being the ideal system for EMA researchers. Our selection processes are described in detail below. Furthermore, Table 1 depicts a 3-point rating system used by our team at the NIMH as one tool for comparing the 7 platforms on each of our individualized and prioritized needs.

XSL•FO
RenderX

Table 1. 3-point rating system for comparing ecological momentary assessment platforms on individualized and prioritized laboratory-based needs; example from the National Institute of Mental Health.

Individualized and prioritized need	Platform 1	Platform 2	Platform 3	Platform 4	Platform 5	Platform 6	Platform 7
Meets the National Institute of Mental Health security requirements and high likelihood of IRB[a] approval	☒ b	☒ c	☒	☒ d	☒	☒	☒
Child and adolescent friendly	☒	☒	☒	☒	☒	☒	☒
Clinical and research dashboards	☒	☒	?[e]	☒	☒	☒	☒
Language: compatibility with English and Hebrew	☒	☒	☒	☒	☒	☒	☒
Self-programming option	☒	☒	☒	☒	☒	☒	☒
Thermometer item display option and visual analog scale with midpoint item response option	☒	☒	?	☒	☒	☒	☒
Wearable integration	☒	☒	☒	☒	☒	☒	☒
Passive sensing	☒	☒	☒	☒	☒	☒	☒
Dyadic interactions	☒	☒	☒	☒	☒	☒	☒

[a]IRB: institutional review board.

[b]☒: full capability.

[c]☒: no capability.

[d]☒: some capability or requires modifications to achieve full capability.

[e]?: no information.

Our team at the NIMH ultimately selected platform 5.

Platform Selection

Our teams selected platforms that were, for the most part, comprehensive in standard features (eg, alarms, device compatibility, downloading results, item display and response, onboarding, sampling schemes, and scheduling; Textbox 1). Unavailable features were anticipated to be added within the next year (ie, audio item response) or could be added for an additional cost (ie, image and video item response) and importantly were not necessary for our ongoing or upcoming laboratory studies.

A subset of our individualized and prioritized laboratory-based needs (Textbox 2) were shared across teams at the NIMH and BCH. Compliance with institutional data security requirements and a high likelihood of IRB approval were paramount (priority 1). Through meetings and correspondence with developers, we obtained answers to the relevant data security questions. At the NIMH, our selected platform did not initially meet security requirements because of data storage outside the United States. However, the developers proposed a viable technical solution for an additional cost. In addition, both teams valued child- and adolescent-friendly user interfaces (priority 2). Our team at BCH recruits a wider age range (7-24 years) than our team at the NIMH (8-17 years). Therefore, BCH selected a platform with customizable fonts, colors, and graphics that would allow for tailoring, depending on the study and sample. The NIMH selected a platform with a fixed, simple, and brightly colored user interface, child-friendly item displays (eg, smiley faces), and gamified incentives. Although both teams were interested in wearable integration, we prioritized this feature differently (priority 4 at BCH and priority 7 at the NIMH). At BCH, a majority of our studies use actigraphy devices to collect activity and sleep data. Thus, the seamless integration of EMA and wearable device data was highly valued. At the NIMH, no ongoing studies actively integrate EMA and wearable devices. In fact, wearable integration was the only item on the NIMH's list of individualized and prioritized laboratory-based needs not offered by our selected platform. Although several other platforms offered wearable integration, we decided to forgo it given the low prioritization. In addition, both teams valued research dashboards that would facilitate the tracking of individual- and sample-level EMA survey compliance (priority 3 at the NIMH; priority 5 at BCH). At the NIMH, we additionally required a dashboard for clinician use (eg, visualizing participant responses to critical clinical items, priority 3). The platforms that we selected conformed to our clinical and research dashboard needs. Finally, we considered interactions with developers, including response time and quality. Both teams had troubleshot countless technical errors on researcher and participant platform interfaces in previous research. At BCH, it was important for our team to select a platform with a strong developer support system (priority 7). Although not originally prioritized at the NIMH, developer

XSL•FO
RenderX

interactions contributed to our decision-making. Across teams, the developers of our selected platforms were quick and comprehensive in correspondence and were willing to collaborate with our teams to address concerns as they developed. In contrast, for example, developers from one platform did not respond to our email inquiries regarding their features (Table 1).

A subset of individualized and prioritized laboratory-based needs were unique to BCH. EMA adherence was a challenge in our previous study. Therefore, we prioritized direct messaging capabilities to send participants survey reminders, troubleshoot issues with wearable devices, and inquire about adherence (priority 3). Platform pricing was also a consideration (priority 6). We compared platform pricing and packages (eg, the maximum number of participants and studies allowed) relative to our needs and the corresponding platform features.

A subset of individualized and prioritized laboratory-based needs were unique to the NIMH. All evaluated platforms were fully or partially (eg, item and response options but not app infrastructure, such as menu options) compatible with Hebrew with additional customization (priority 4). Five platforms, including our selected platform, were entirely self-programmable (priority 5); 2 platforms required partnering with developers for the study setup and allowed subsequent self-programmed modifications. Our selected platform did not initially offer our desired thermometer item display or visual analog scale with midpoint item response options (priority 6), but the developers built these options free of charge in time for our evaluation. Platforms ranged in passive sensing capabilities (priority 8), with some offering full, continuous passive sensing for multiple metrics (eg, text transcripts, step count, browser history, and social media posts); others offering passive sensing during (or close in time to) survey completion (our selected platform); and still others without passive sensing capabilities. Finally, our selected platform was 1 of 3 evaluated platforms with dyadic interaction capabilities (eg, triggering surveys based on the responses of a designated partner; priority 9).

Discussion

Overview

On the basis of our experiences as EMA platform users and developers, along with lessons learned from the current research, here, we discuss 11 considerations in selecting an EMA platform: (1) location; (2) developer involvement; (3) sample characteristics; (4) onboarding; (5) survey design features; (6) sampling scheme and scheduling; (7) viewing results; (8) dashboards; (9) security, privacy, and data management; (10) pricing and cost structure; and (11) future directions. For each consideration, we provide example questions to ask platform developers, which vary based on individualized and prioritized lab-based needs (Textbox 1).

Consideration 1: Location

Overview

Think about the location of the developer relative to your study and institution. The developer's location may influence the language or languages of the participant and researcher platform interfaces and instructional materials. In addition, developer location may impact the timing of support from developers and dictate the location of the data storage.

Are Platform Features and Instructions Available in (My Language or Languages)?

Determine whether each platform is available in the language or languages of your sample and study team. It is not only necessary to confirm that survey items and response options are programmable in the language or languages of your study, but it is also important that participant- and researcher-facing features are available in the appropriate language or languages. Can the platform accommodate languages with non-Roman alphabets (eg, Arabic and Greek) or languages that read from right to left (eg, Hebrew and Farsi)? Some developers are willing to build new languages into their platforms (potentially for a fee). Furthermore, developers may offer knowledge bases, manuals, or other informational materials that will help you use the platform, and it is important that these resources are accessible to your team.

Will You Be in Office While (My Study) Is Live, so That I Can Receive Support, if Needed?

If developer support with quick turnaround time is a priority, determine when your study will be live, when developers will be available, and how quickly developers will respond to service requests. Consider the time zone differences between the developer and study sites. Additional coordination may be necessary if there are gaps in developer coverage. For example, in previous research, our team at BCH scheduled a weekly check-in and troubleshooting meeting with developers outside the United States to circumvent response delays due to time zone differences.

Where Will the Data Be Stored?

Determine whether your country and institution impose restrictions on the location of your data. For example, the General Data Protection Regulation stipulates how the data of European Union citizens and residents must be collected, stored, processed, and disseminated (eg, individuals must have the ability to erase their personal data), and platforms that do not comply may not be used [61]. Regarding institutional policies on data storage, early conversations with your data security officer or IRB are critical (see the section *Consideration 9: Security, Privacy, and Data Management*). If data storage outside of your country is prohibited by your institution, some developers may collaborate on customizing data storage locations.

Consideration 2: Developer Involvement

Overview

Although some studies are straightforward and require low maintenance, others are more complex and require more frequent communication with developers. Relative to the technical skill levels of your team members, consider the support (type and quantity) you will need from developers at the beginning and throughout the life span of your study.

XSL•FO
RenderX

How Much Support Will You Provide at the Beginning of (My Study)?

Developers offer varying levels of support when initiating a new study, and accordingly, our teams encountered a range of support at the outset of our evaluation. Some developers programmed the study on our behalf, in part or in whole, after receiving our protocol. Other developers provided access to knowledge bases, manuals, and other informational materials to facilitate the setup led by our teams. The cost of developer involvement in the study setup varies. Some platforms offer unlimited support or build a fixed number of training sessions into their pricing. Other platforms support users on a fee-per-session basis.

How Much Support Will You Provide Throughout the Life Span of (My Study)?

Developers offer varying levels of support throughout the life span of a study. In our evaluation, developer response times spanned hours to weeks. Typically, we requested support by emailing a prespecified point of contact or submitting "help tickets" on the web. If support is provided via email, work with developers to determine an alternate point of contact for vacation and sick days. If support is provided via help tickets, inquire about the average time from receipt to resolution. Regardless, be informed and develop a backup plan. Plan ahead, as most technical malfunctions tend to occur when studies are launched.

Consideration 3: Sample Characteristics

Overview

The characteristics of your sample may inform the optimal participant user experience (UX) for your platform.

Is Your Participant UX Optimized for (My Sample Characteristics)?

Intuitive and engaging UX is a goal for developers, regardless of sample characteristics. Still, depending on factors, including participant age range and level of cognitive functioning, it may be critical to implement an easy-to-navigate and visually appealing UX. For example, the NIMH prioritized a child-friendly UX. Of the platforms we evaluated, a few were characterized by child-friendly UX (eg, large fonts and bright colors), and some (but not all) allowed for user customization. Gamification and rewards may increase survey compliance and minimize dropout. Of the platforms that we evaluated, one offered awards for survey completion and another offered a virtual "shop" to unlock incentives such as fun facts. Creative response options may make items clearer and more engaging. In our evaluation, one platform offered emoji response options and another platform enabled researchers to upload images as response options.

What Features Support Platform Use With High-Risk Samples?

If recruiting a high-risk sample (eg, individuals at risk for suicidal ideation), you may consider platforms that allow for responses (eg, thoughts of death or dying) to be "flagged" and forwarded to a point of contact on your team (eg, clinician) for follow-up. Some platforms leverage natural language processing to identify flagged terms in free-response items. In addition, some platforms can be programmed to automatically generate follow-up questions or messages for a flagged response.

Consideration 4: Onboarding

Overview

Determine whether you will enroll participants in person or virtually, factoring in the technical skills of your team members and sample, and assess the relative fit of platform onboarding options.

How Can Participants Enroll in (My Study)?

Consider how participants will enroll in your study. Although participants may benefit from in vivo technical support, virtual onboarding may be less resource-intensive (eg, lower cost, easier to schedule). Both the technical skills of your sample (considering age, cognitive functioning, and clinical status) and research team are factors to consider. Some platforms autogenerate onboarding invitations that are sent to participants via email or SMS text message. Notably, email invitations may be sent to spam folders, emphasizing the importance of detailed study protocols, including guidance on troubleshooting technical issues during onboarding. Other platforms direct participants to scan a QR code for study enrolment. The level of automaticity and researcher time and effort in onboarding varies by platform. For example, when evaluating one platform, our team was responsible for generating mobile codes for enrollment.

Can or Must Participants Provide Consent Through the Platform?

Some platforms allow for the administration of a researcher-developed e-consent form as part of onboarding. Understand whether this feature is available if needed and can be disabled if you decide to consent participants through another method. In addition, given passive sensing integrations to some platforms, participants may be prompted to provide permission to access smartphone features and sensors (eg, microphones). In our evaluation, multiple participants expressed concerns about privacy upon receiving smartphone sensor permission requests from the platform, even though we instructed participants to refuse those requests and did not collect passive data.

Consideration 5: Survey Design

Overview

Determine which design features are required for your surveys to be maximally effective, user-friendly, and tailored to your sample.

Does the Platform Offer (My Prioritized Survey Display Options)?

Determine whether you require your surveys to display text, images, audio, or video to the user. Depending on your sample, visual or auditory cues may be beneficial (see the section *Consideration 3: Sample Characteristics*). Although text display is ubiquitous across platforms, options to present images, audio, and video vary by platform. In addition, consider how you would like users to move through your survey; some platforms display

XSL•FO
RenderX

multiple items on a single page (ie, via scrolling), other platforms display one item per page, and still others allow for customization. Think about whether participants should be able to skip items and return to previous items and determine if those features are modifiable in the platform.

Does the Platform Offer (My Prioritized Survey Response Options)?

Determine whether your response options will be a single-choice, multiple-choice, Likert-scale, or free-text response or something else. For example, you might ask participants to respond through captured images, audio, video, time stamps, or geotags. In addition, consider your need for "branching logic" or custom paths based on user responses. Many platforms allow participant responses to dictate the items to follow, but fewer platforms allow responses to trigger future surveys (eg, later in the day).

Consideration 6: Sampling Scheme and Scheduling

Overview

Determine the sampling scheme for your study and whether the platform supports your sampling scheme. Consider the other platform features you may need for successful survey deployment and completion.

Which Sampling Schemes Are Supported by the Platform?

Decide whether your surveys will be interval contingent, signal contingent, event contingent, or some combination thereof. Some platforms also allow for pairing with wearable devices and passive sensors, such that geolocation or another target might trigger a survey. Inquire as to whether multiple surveys can be programmed and deployed or if the platform can only manage a single survey. In addition, understand the method for developing the sampling scheme structure within the platform; our team found the evaluation to be critical in testing the researcher-facing platform design.

What Aspects of the Alarm or Notification System Are Modifiable?

Alarms signaling users to complete a survey might use sound, vibration, a banner, or a push notification. Some platforms allow for reminder notifications if a survey is not completed. Understand the level of customization offered by your platform. Greater flexibility will allow you to develop a system tailored to your sample such that participants remain engaged but not overwhelmed with notifications.

Can Survey Scheduling Vary Based on Participant Personal Preferences, or Is Scheduling Fixed Across Participants?

Some degree of personalization in survey scheduling may increase survey compliance. For example, the morning survey for participant A (an early riser) may be deployed randomly between 7 AM and 8 AM, and the morning survey for participant B (a late sleeper) may be deployed randomly between 8:30 AM and 9:30 AM. Determine whether the platform allows for personalization or if survey scheduling must be held constant across your sample. In addition, when enrolling participants across time zones (eg, multisite studies), understand whether the survey will be deployed at the local time of the participant or researcher.

Does the Platform Require Wi-Fi or Cellular Data to Function?

Understand the impact of participant phones being "offline" (eg, survey deployment, response, and upload to the server). For example, some platforms are native apps that allow users to access and respond to surveys even offline, and other platforms store responses locally until Wi-Fi or cellular data are available. In addition, consider the impact of the platform on data use. Significant data use may be prohibitive if participants use their own phones. Passive sensing may rely heavily on data.

Consideration 7: Dashboards

Overview

Dashboards can be used for data monitoring. Given that dashboards are not universal to platforms, consider whether a dashboard is a priority.

Do You Offer Dashboards, What Are Their Capabilities, and to What Extent Are They Customizable?

Dashboards can be programmed to report high-level information about your sample, specific information about individual participants, or both. When determining your dashboard needs, consider which team members will need to view your data (eg, research coordinator, clinician, or principal investigator). Then, consider which variables each team member will need to view (eg, compliance and symptoms), including level (eg, full sample and individual participants) and interval (eg, daily and weekly). In addition, consider how the data should be visualized (eg, graphs and tables). For example, clinicians may want to track depression symptom scores weekly over the course of a clinical trial using a time-series graph. Some groups may be interested in data collection on dyads, making platforms with dyadic linking and dashboards with side-by-side data visualization optimal. If multiple users require dashboards for distinct purposes (eg, clinicians and researchers), inquire whether the platform supports multiple dashboards and users in a single study.

Consideration 8: Viewing Results

Overview

Consider how and how often data should be uploaded to the server and downloaded by your research team. In addition, consider your preferences for structuring your data export.

How and How Often Are Data Uploaded?

Consider how (eg, Wi-Fi and cellular data) and how often (eg, immediately and daily) data are uploaded to the server. Some platforms require Wi-Fi or cellular data to deliver surveys and subsequently upload data. Other platforms are accessible offline and do not require connections to Wi-Fi or cellular data for functionality. Still other platforms deliver surveys offline, and data are stored on the platform and uploaded when connected to Wi-Fi or cellular data. Data upload considerations may vary

XSL•FO
RenderX

in importance depending on your study design (eg, use of participant phones compared with study phones).

In What Formats Are Data Exported?

In many cases, platforms export data as comma-separated values files. Some platforms offer additional formats (eg, JavaScript Object Notation) and application programming interfaces (eg, SPSS and R). In addition, some platforms export individual participant data (which may be less feasible for data collection on large samples) and others export data for the full sample.

How Will the Data Export Be Structured?

Platforms vary in how their data exports are structured and whether (and to what extent) their data exports are customizable. Consider whether you will need your data in a long or wide format, how missing data should be reported, and which fields are necessary (eg, skipped items, response time, and summary statistics). Inconsistency between data exports and your needs will yield extra time and effort in preprocessing.

Consideration 9: Security, Privacy, and Data Management

Overview

Consider security, privacy, and data management requirements as they relate to your country and institution. The presence or absence of certain platform features (eg, 2-factor authentication; 2FA) may render a platform incompatible with those requirements. Depending on your institution and familiarity with their policies, consider meeting or corresponding with your data security officer or IRB before consulting with developers. Our team at the NIMH met with our data security officer to understand the requirements of IT applications and services (Multimedia Appendix 2). We presented the corresponding questions to developers when we met to discuss their platforms. On the basis of developer responses, our data security officer assigned platforms to varying levels of security clearance for NIMH-specific standards, which informed the selection of our platform. Our team at BCH was not required to have an initial conversation with data security. Instead, our IRB application included questions about our candidate platform (Multimedia Appendix 2), and we met with the data security team following submission.

How Are the Platform and Data Secured?

Inquire about whether and how the platform secures access, including requiring a personal identification number, a password, or 2FA. Although 2FA by some means (eg, email) may complicate, slow, and impede survey responses, other means may be simpler, quicker, and more feasible (eg, Touch ID, Face ID, and text). Some of the platforms we evaluated offered (and some required) 2FA. Other platforms did not offer 2FA, stating that study accounts do not store user phone numbers to enable 2FA. In addition, inquire about whether and how data are stored on the platform and mobile device and how data are encrypted. All the platforms evaluated encrypted data as they were transported between phones and servers, and most platforms used Amazon Web Services for cloud storage.

How Does the Platform Maintain User Privacy?

Ask developers what kind of information is stored about the participant (eg, log-in information, survey data, and location), where (eg, on the platform and in a server), and for how long (eg, days, months, duration of the contract, or in perpetuity). Follow up to determine how data are deidentified, if at all, and where the key to the data is stored. Inquire as to whether the platform is compliant with the laws and regulations of your country that protect an individual's data or information (eg, Health Insurance Portability and Accountability Act in the United States and General Data Protection Regulation in the European Union). In our evaluation, platforms typically temporarily store survey data on phones in an encrypted format not visible in phone storage or accessible to other apps. These data were deleted from the phone once they were uploaded to the server. Some platforms avoid storing user information by not creating user accounts. Other platforms create user accounts but do not require personal information. Still others use information such as email addresses to create accounts.

How Are Data Managed?

Determine whether the data will be stored in the country of research or abroad (see the section *Consideration 1: Location*) and on a physical server or the cloud. Determine how long the platform will retain the data, along with whether and when long-term data storage will incur an additional cost. During our evaluation, one platform maintained data storage for the license period and another platform maintained data storage for 6 months following the license period to allow researchers time for data analysis.

Consideration 10: Pricing and Cost Structure

Overview

Before meeting with developers, know your budget. The platforms we evaluated offered complimentary project consultations and estimates, which were key in selecting among candidates.

What Is Your Pricing and Cost Structure?

Platforms vary widely in their pricing and cost structures. Platforms may charge a flat fee for a user license with varying levels (basic and premium), or platforms may provide quotes that are team and project specific, varying depending on your estimated number of participants, the size of your team, the roles of your team members (eg, researchers and clinicians), your institution, and required features (eg, basic EMA vs EMA plus passive sensing). Some platforms offer free, basic versions of their platform. For example, one platform in our evaluation offered a basic package for clinicians to use in private practice with their clients while researchers were asked to purchase a premium package. Another platform offered discounted costs with the purchase of a multiple-year license. Costs may also vary based on whether you are part of an academic institution or a government institution and whether your project is single or multisite. Academic institutions are typically lower in cost than private entities, and single-site projects are lower in cost than multisite projects.

XSL•FO
RenderX

What Additional Features Do You Offer and Are They Included in the Price?

Consider how necessary, desired, and unnecessary features may impact platform costs. Sophisticated tools such as passive sensing, pairing with wearable devices, dyadic linking of participants, or compatibility with health care systems are costly to develop and maintain. Furthermore, images and videos may be expensive for developers to store. One of the platforms that we evaluated offered advanced features, including the ability to access electronic health records and facilitate biospecimen collection, which resulted in high overall costs for the platform. If your project requires only EMA, choose a platform with basic offerings or ask if you can opt out of extra features.

Consideration 11: Future Directions of Your Study and Research Program

Overview

The next steps for the platform (eg, financial viability and developments) and your study and research program (eg, the need for additional features) should be considered before you partner with developers.

How Long Do You Expect the Platform to Be Operational?

Consider the length of your study and explicitly ask developers their expectation for being operational for that duration and longer. A single, reliable platform is critical for stable data collection. Funding sources may inform the long-term stability of a platform. For example, if a platform is grant funded, inquire about plans for financing the platform when the funding period elapses. In addition, ask whether and when developers plan to launch upgrades to, or new versions of, the platform and how those developments might interface with an ongoing study. These considerations are especially critical for longitudinal studies.

How Could the Platform Adapt to the Evolving Needs of (My Laboratory)?

As you plan your research program (eg, integration of additional measures delivered via EMA, passive sensing, and wearable devices), consider whether the platform has or can develop the capacity to support these plans. At the same time, it may be appropriate to select a particular platform for one phase of research with the understanding that the platform may be reevaluated in the future.

Conclusions

EMA is a tool for the repeated, real-time assessment of constructs of interest in natural settings, which offers great opportunity for advancing psychological science. Accordingly, interest in incorporating EMA into individual studies and research programs has burgeoned since EMA emerged in the 1990s. However, numerous platforms are available to researchers, and choosing between options is an arduous task to the point that some researchers elect to develop their own platforms. Taken together, we found that the optimal EMA platform is the one that best meets the study and laboratory needs. Reconciling needs with opportunities to optimize outcomes requires information, understanding, and planning. Here, we offer specific guidance on platform selection for researchers interested in integrating EMA methods into their research programs.

Acknowledgments

This research was supported by the National Institute of Mental Health (MEB: ZIAMH002969) and the National Institute on Drug Abuse (JK: K01DA057374).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Number of publications by year from 1994 through 2023 including the term ecological momentary assessment (EMA) and related terms. Overall, 1032 publications in 2012 and 2767 publications in 2022 used the term EMA and related terms, yielding an increase of over 168% across time. These data reflect a search of Embase, PsycINFO, Web of Science, Scopus, and PubMed from 1994 (when the term EMA was coined by Stone and Shiffman) through 2023, which yielded 25,353 articles. An additional 1061 articles were published from 1829 through 1993 using the term “daily diary,” a technique preceding EMA consisting of a single end-of-day assessment typically administered via paper and pencil or a telephone call.

[PNG File , 18 KB - jmir_v26i1e51125_app1.png]

Multimedia Appendix 2

A list of app-based, text-based, web-based, and user-design ecological momentary assessment (EMA) platforms; a list of EMA platforms developed by researchers for individual use; the PubMed search strategy; and sample EMA platform data security questions.

[DOCX File , 49 KB - jmir_v26i1e51125_app2.docx]

References

XSL•FO
RenderX

1. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
2. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16(3):199-202 [FREE Full text] [doi: 10.1093/abm/16.3.199]
3. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9:151-176 [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
4. Kim J, Marcusson-Clavertz D, Yoshiuchi K, Smyth JM. Potential benefits of integrating ecological momentary assessment data into mHealth care systems. Biopsychosoc Med 2019 Aug 09;13:19 [FREE Full text] [doi: 10.1186/s13030-019-0160-5] [Medline: 31413726]
5. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
6. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017 Mar 15;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]
7. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based Ecological Momentary Assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017 Nov 01;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
8. Stone A, Shiffman S, Atienza A, Nebeling L. The Science of Real-Time Data Capture: Self-Reports in Health Research. Oxford, UK: Oxford University Press; 2007.
9. Stone AA, Shiffman SS. Ecological Momentary Assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16(3):199-202 [FREE Full text] [doi: 10.1093/abm/16.3.199]
10. Allport GW. Personality: A Psychological Interpretation. In: Sociometry. London, UK: Constable and Co Limited; Jan 1938.
11. Brunswik E. Organismic achievement and environmental probability. Psychological Review 1943 May;50(3):255-272. [doi: 10.1037/h0060889]
12. Cattell RB, Cattell AK, Rhymer RM. P-technique demonstrated in determining psychophysiological source traits in a normal individual. Psychometrika 1947 Dec;12(4):267-288. [doi: 10.1007/bf02288941] [Medline: 18921433]
13. Serre F, Fatseas M, Swendsen J, Auriacombe M. Ecological momentary assessment in the investigation of craving and substance use in daily life: a systematic review. Drug Alcohol Depend 2015 Mar 01;148:1-20. [doi: 10.1016/j.drugalcdep.2014.12.024] [Medline: 25637078]
14. Colombo D, Fernández-Álvarez J, Patané A, Semonella M, Kwiatkowska M, García-Palacios A, et al. Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: a systematic review. J Clin Med 2019 Apr 05;8(4):233 [FREE Full text] [doi: 10.3390/jcm8040465] [Medline: 30959828]
15. Walz LC, Nauta MH, Aan Het Rot M. Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: a systematic review. J Anxiety Disord 2014 Dec;28(8):925-937. [doi: 10.1016/j.janxdis.2014.09.022] [Medline: 25445083]
16. Miguelez-Fernandez C, de Leon SJ, Baltasar-Tello I, Peñuelas-Calvo I, Barrigon ML, Capdevila AS, et al. Evaluating attention-deficit/hyperactivity disorder using ecological momentary assessment: a systematic review. Atten Defic Hyperact Disord 2018 Dec;10(4):247-265. [doi: 10.1007/s12402-018-0261-1] [Medline: 30132248]
17. Santangelo P, Bohus M, Ebner-Priemer UW. Ecological momentary assessment in borderline personality disorder: a review of recent findings and methodological challenges. J Pers Disord 2014 Aug;28(4):555-576. [doi: 10.1521/pedi_2012_26_067] [Medline: 22984853]
18. Naim R, Smith A, Chue A, Grassie H, Linke J, Dombek K, et al. Using ecological momentary assessment to enhance irritability phenotyping in a transdiagnostic sample of youth. Dev Psychopathol 2021 Dec 07;33(5):1734-1746 [FREE Full text] [doi: 10.1017/s0954579421000717]
19. Schaefer LM, Engel SG, Wonderlich SA. Ecological momentary assessment in eating disorders research: recent findings and promising new directions. Curr Opin Psychiatry 2020 Nov;33(6):528-533 [FREE Full text] [doi: 10.1097/YCO.0000000000000639] [Medline: 32740204]
20. Mote J, Fulford D. Ecological momentary assessment of everyday social experiences of people with schizophrenia: a systematic review. Schizophr Res 2020 Feb;216:56-68. [doi: 10.1016/j.schres.2019.10.021] [Medline: 31874743]
21. Sedano-Capdevila A, Porras-Segovia A, Bello HJ, Baca-García E, Barrigon ML. Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Curr Psychiatry Rep 2021 May 18;23(7):41. [doi: 10.1007/s11920-021-01255-7] [Medline: 34003405]
22. May M, Junghaenel DU, Ono M, Stone AA, Schneider S. Ecological momentary assessment methodology in chronic pain research: a systematic review. J Pain 2018 Jul;19(7):699-716 [FREE Full text] [doi: 10.1016/j.jpain.2018.01.006] [Medline: 29371113]
23. Ebner-Priemer UW, Trull TJ. Ambulatory assessment. Eur Psychol 2009 Jan;14(2):109-119 [FREE Full text] [doi: 10.1027/1016-9040.14.2.109]

XSL•FO
RenderX

24. Mestdagh M, Dejonckheere E. Ambulatory assessment in psychopathology research: current achievements and future ambitions. Curr Opin Psychol 2021 Oct;41:1-8 [FREE Full text] [doi: 10.1016/j.copsyc.2021.01.004] [Medline: 33550191]
25. Thunnissen MR, Aan Het Rot M, van den Hoofdakker BJ, Nauta MH. Youth psychopathology in daily life: systematically reviewed characteristics and potentials of ecological momentary assessment applications. Child Psychiatry Hum Dev 2022 Dec 01;53(6):1129-1147 [FREE Full text] [doi: 10.1007/s10578-021-01177-8] [Medline: 34075490]
26. Russell M, Gajos J. Annual research review: ecological momentary assessment studies in child psychology and psychiatry. J Child Psychol Psychiatry 2020 Mar;61(3):376-394 [FREE Full text] [doi: 10.1111/jcpp.13204] [Medline: 31997358]
27. Czyz E, Horwitz A, Arango A, King C. Short-term change and prediction of suicidal ideation among adolescents: a daily diary study following psychiatric hospitalization. J Child Psychol Psychiatry 2019 Jul;60(7):732-741 [FREE Full text] [doi: 10.1111/jcpp.12974] [Medline: 30246870]
28. Smith KE, Juarascio A. From Ecological Momentary Assessment (EMA) to Ecological Momentary Intervention (EMI): past and future directions for ambulatory assessment and interventions in eating disorders. Curr Psychiatry Rep 2019 Jun 04;21(7):53. [doi: 10.1007/s11920-019-1046-8] [Medline: 31161276]
29. Schueller SM, Aguilera A, Mohr DC. Ecological momentary interventions for depression and anxiety. Depress Anxiety 2017 Jun;34(6):540-545 [FREE Full text] [doi: 10.1002/da.22649] [Medline: 28494123]
30. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
31. Henry LM, Compas BE. Review: preventing psychopathology in the digital age: leveraging technology to target coping and emotion regulation in adolescents. JAACAP Open 2023 Oct (forthcoming) [FREE Full text] [doi: 10.1016/j.jaacop.2023.09.006]
32. Mobile fact sheet. Pew Research Center. 2021 Apr. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2023-04-28]
33. Rice E, Lee A, Taitt S. Cell phone use among homeless youth: potential for new health interventions and research. J Urban Health 2011 Dec;88(6):1175-1182 [FREE Full text] [doi: 10.1007/s11524-011-9624-z] [Medline: 22076445]
34. Anderson M, Jiang J. Teens, social media and technology 2018. Pew Research Center. 2018. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2023-04-23]
35. George MJ, Jensen MR, Russell MA, Gassman-Pines A, Copeland WE, Hoyle RH, et al. Young adolescents' digital technology use, perceived impairments, and well-being in a representative sample. J Pediatr 2020 Apr;219:180-187 [FREE Full text] [doi: 10.1016/j.jpeds.2019.12.002] [Medline: 32057438]
36. West DM. Digital divide: Improving Internet access in the developing world through affordable services and diverse content. Center for Technology Innovation at Brookings. 2015. URL: https://www.brookings.edu/articles/digital-divide-improving-internet-access-in-the-developing-world-through-affordable-services-and-diverse-content/#:~:text=From%20this%20research%2C%20it%20is,for%20local%20content%20and%20services [accessed 2023-04-28]
37. Rachuri K, Musolesi M, Mascolo C, Rentfrow P, Longworth C, Aucinas A. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM international conference on Ubiquitous computing. 2010 Sep Presented at: UbiComp '10; September 26-29, 2010; Copenhagen, Denmark p. 281-290 URL: https://dl.acm.org/doi/10.1145/1864349.1864393 [doi: 10.1145/1864349.1864393]
38. Barchitta M, Maugeri A, Favara G, Magnano San Lio R, Riela PM, Guarnera L, et al. Development of a web-app for the ecological momentary assessment of dietary habits among college students: the HEALTHY-UNICT project. Nutrients 2022 Jan 13;14(2):330 [FREE Full text] [doi: 10.3390/nu14020330] [Medline: 35057511]
39. Slofstra C, Klein NS, Nauta MH, Wichers M, Batalas N, Bockting CL. Imagine your mood: study design and protocol of a randomized controlled micro-trial using app-based experience sampling methodology to explore processes of change during relapse prevention interventions for recurrent depression. Contemp Clin Trials Commun 2017 Sep;7:172-178 [FREE Full text] [doi: 10.1016/j.conctc.2017.07.003] [Medline: 29696182]
40. Torous J, Staples P, Shanahan M, Lin C, Peck P, Keshavan M, et al. Utilizing a personal smartphone custom app to assess the Patient Health Questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR Ment Health 2015 Mar 24;2(1):e8 [FREE Full text] [doi: 10.2196/mental.3889] [Medline: 26543914]
41. Reid SC, Kauer SD, Dudgeon P, Sanci LA, Shrier LA, Patton GC. A mobile phone program to track young people's experiences of mood, stress and coping. Development and testing of the mobiletype program. Soc Psychiatry Psychiatr Epidemiol 2009 Jun;44(6):501-507. [doi: 10.1007/s00127-008-0455-5] [Medline: 19011723]
42. Brannon EE, Cushing CC, Crick CJ, Mitchell TB. The promise of wearable sensors and ecological momentary assessment measures for dynamical systems modeling in adolescents: a feasibility and acceptability study. Transl Behav Med 2016 Dec;6(4):558-565 [FREE Full text] [doi: 10.1007/s13142-016-0442-4] [Medline: 27678501]
43. González-Barato LJ, Rubio VJ, Hernández JM, Sánchez-Iglesias I. PSIXPORT: mobile app for ecological momentary assessment of psychological dimensions in sport injury. Front Psychol 2021 Jul 27;12:697293 [FREE Full text] [doi: 10.3389/fpsyg.2021.697293] [Medline: 34385962]

XSL•FO
RenderX

44. Meigs J, Naim R, Pandya U, Shaughnessy S, Dombek K, Kircanski K, et al. The initial validation of RATE-IT, a novel mobile ecological momentary assessment application to probe irritability. Society for Digital Mental Health Conference. 2022 Jun 13. URL: https://www.societydmh.org/2022-posters/ [accessed 2023-04-28]
45. Rabbi M, Philyaw Kotov M, Cunningham R, Bonar EE, Nahum-Shani I, Klasnja P, et al. Toward increasing engagement in substance use data collection: development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults. JMIR Res Protoc 2018 Jul;7(7):e166 [FREE Full text] [doi: 10.2196/resprot.9850] [Medline: 30021714]
46. Blaauw FJ, van der Gaag MA, Snell NR, Emerencia AC, Kunnen ES, de Jonge P. The u-can-act platform: a tool to study intra-individual processes of early school leaving and its prevention using multiple informants. Front Psychol 2019;10:1808 [FREE Full text] [doi: 10.3389/fpsyg.2019.01808] [Medline: 31616330]
47. Fellendorf FT, Hamm C, Dalkner N, Platzer M, Sattler MC, Bengesser SA, et al. Monitoring sleep changes via a smartphone app in bipolar disorder: practical issues and validation of a potential diagnostic tool. Front Psychiatry 2021 Mar 24;12:641241 [FREE Full text] [doi: 10.3389/fpsyt.2021.641241] [Medline: 33841209]
48. Bakolis I, Hammoud R, Smythe M, Gibbons J, Davidson N, Tognin S, et al. Urban mind: using smartphone technologies to investigate the impact of nature on mental well-being in real time. Bioscience 2018 Feb 01;68(2):134-145 [FREE Full text] [doi: 10.1093/biosci/bix149] [Medline: 29599549]
49. Grasser LR, Erjo T, Goodwin MS, Naim R, German RE, White J, et al. Can peripheral psychophysiological markers predict response to exposure-based cognitive behavioral therapy in youth with severely impairing irritability? A study protocol. Research Square. Preprint posted online October 02, 2023 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3396443/v1]
50. Naim R, Shaughnessy S, Smith A, Karalunas SL, Kircanski K, Brotman MA. Real-time assessment of positive and negative affective fluctuations and mood lability in a transdiagnostic sample of youth. Depress Anxiety 2022 Dec;39(12):870-880 [FREE Full text] [doi: 10.1002/da.23293] [Medline: 36325887]
51. Naim R, Goodwin MS, Dombek K, Revzina O, Agorsor C, Lee K, et al. Cardiovascular reactivity as a measure of irritability in a transdiagnostic sample of youth: preliminary associations. Int J Methods Psychiatr Res 2021 Dec;30(4):e1890 [FREE Full text] [doi: 10.1002/mpr.1890] [Medline: 34390050]
52. Cardinale EM, Naim R, Haller SP, German R, Botz-Zapp C, Bezek J, et al. Rationale and validation of a novel mobile application probing motor inhibition: proof of concept of CALM-IT. PLoS One 2021 Jun 4;16(6):e0252245 [FREE Full text] [doi: 10.1371/journal.pone.0252245] [Medline: 34086728]
53. Haller SP, Stoddard J, Botz-Zapp C, Clayton M, MacGillivray C, Perhamus G, et al. A randomized controlled trial of computerized interpretation bias training for disruptive mood dysregulation disorder: a fast-fail study. J Am Acad Child Adolesc Psychiatry 2022 Jan;61(1):37-45 [FREE Full text] [doi: 10.1016/j.jaac.2021.05.022] [Medline: 34147585]
54. Koechlin H, Beeckman M, Meier AH, Locher C, Goubert L, Kossowsky J, et al. Association of parental and adolescent emotion-related factors with adolescent chronic pain behaviors. Pain 2022 Jul 01;163(7):e888-e898 [FREE Full text] [doi: 10.1097/j.pain.0000000000002508] [Medline: 35050962]
55. Kaczynski KJ, Chang CY, Chimoff J, Koike C, Berde CB, Logan DE, et al. Initial adjustment to the COVID-19 pandemic and the associated shutdown in children and adolescents with chronic pain and their families. Front Pain Res (Lausanne) 2021 Sep 30;2:713430 [FREE Full text] [doi: 10.3389/fpain.2021.713430] [Medline: 35295442]
56. Kelleher S, Cravero J, Baumer A, Bertisch S, Kossowsky J. Monitoring sleep, activity and function using wearable devices in post-surgical patients: a pilot study. Society for Technology in Anesthesia. URL: https://www.stahq.org/userfiles/files/033_STA_20AM_Abstract_Kelleher%28579%29.pdf [accessed 2023-04-28]
57. Lee-Archer PF, Kossowsky J, Cravero JP. If the watch fits-wear it. Paediatr Anaesth 2020 May;30(5):518-519. [doi: 10.1111/pan.13835] [Medline: 32500966]
58. Nestor B, Chimoff J, Koike C, Weitzman E, Riley B, Uhl K, et al. Adolescent and Parent Perspectives on Digital Phenotyping in Youth with Chronic Pain: A Cross-Sectional Mixed-Methods Survey Study. J Med Internet Res (forthcoming) [FREE Full text] [doi: 10.2196/47781]
59. Aeberhard A, Gschwind L, Kossowsky J, Luksys G, Papassotiropoulos A, de Quervain D, et al. Introducing COSMOS: a web platform for multimodal game-based psychological assessment geared towards open science practice. J Technol Behav Sci 2019;4:234-244 [FREE Full text] [doi: 10.31234/osf.io/cb4f9]
60. Lalloo C, Boerner K, Kossowsky J. Mobile health data for pediatric pain: inter-institutional experiences and best practices for patient-centred data capture, analysis, and visualization. Pediatr Pain Lett 2023 May 10;25(1):1-5 [FREE Full text]
61. 2016/679 (General Data Protection Regulation). The European Parliament and of the Council. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 [accessed 2023-04-23]

Abbreviations

2FA: 2-factor authentication
BCH: Boston Children's Hospital
EMA: ecological momentary assessment
EMI: ecological momentary intervention

XSL•FO
RenderX

IRB: institutional review board
JITAI: just-in-time adaptive intervention
NIMH: National Institute of Mental Health
UX: user experience

Edited by A Mavragani; submitted 21.07.23; peer-reviewed by K Yoshiuchi, M Ono; comments to author 05.09.23; revised version received 25.09.23; accepted 18.10.23; published 04.01.24.

Please cite as:
Henry LM, Hansen E, Chimoff J, Pokstis K, Kiderman M, Naim R, Kossowsky J, Byrne ME, Lopez-Guzman S, Kircanski K, Pine DS, Brotman MA
Selecting an Ecological Momentary Assessment Platform: Tutorial for Researchers
J Med Internet Res 2024;26:e51125
URL: https://www.jmir.org/2024/1/e51125
doi:10.2196/51125
PMID:38175682

©Lauren M Henry, Eleanor Hansen, Justin Chimoff, Kimberly Pokstis, Miryam Kiderman, Reut Naim, Joe Kossowsky, Meghan E Byrne, Silvia Lopez-Guzman, Katharina Kircanski, Daniel S Pine, Melissa A Brotman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Developing a Chatbot to Support Individuals With Neurodevelopmental Disorders: Tutorial

Ashwani Singla[1], MSc; Ritvik Khanna[1], BSc; Manpreet Kaur[1], MEng; Karen Kelm[1], MA; Osmar Zaiane[1], MSc, PhD; Cory Scott Rosenfelt[1], BSc; Truong An Bui[1], PhD; Navid Rezaei[1], PhD; David Nicholas[1], PhD; Marek Z Reformat[1], MSc, PhD; Annette Majnemer[2], OT, PhD; Tatiana Ogourtsova[2], OT, PhD; Francois Bolduc[1], MD, PhD

[1]Department of Pediatrics, University of Alberta, Edmonton, AB, Canada

[2]School of Physical & Occupational Therapy, McGill University, Montreal, QC, Canada

Corresponding Author:
Francois Bolduc, MD, PhD
Department of Pediatrics
University of Alberta
11315 87th Avenue
Edmonton, AB, T6G 2E1
Canada
Phone: 1 780 492 9713
Email: fbolduc@ualberta.ca

Abstract

Families of individuals with neurodevelopmental disabilities or differences (NDDs) often struggle to find reliable health information on the web. NDDs encompass various conditions affecting up to 14% of children in high-income countries, and most individuals present with complex phenotypes and related conditions. It is challenging for their families to develop literacy solely by searching information on the internet. While in-person coaching can enhance care, it is only available to a minority of those with NDDs. Chatbots, or computer programs that simulate conversation, have emerged in the commercial sector as useful tools for answering questions, but their use in health care remains limited. To address this challenge, the researchers developed a chatbot named CAMI (Coaching Assistant for Medical/Health Information) that can provide information about trusted resources covering core knowledge and services relevant to families of individuals with NDDs. The chatbot was developed, in collaboration with individuals with lived experience, to provide information about trusted resources covering core knowledge and services that may be of interest. The developers used the Django framework (Django Software Foundation) for the development and used a knowledge graph to depict the key entities in NDDs and their relationships to allow the chatbot to suggest web resources that may be related to the user queries. To identify NDD domain–specific entities from user input, a combination of standard sources (the Unified Medical Language System) and other entities were used which were identified by health professionals as well as collaborators. Although most entities were identified in the text, some were not captured in the system and therefore went undetected. Nonetheless, the chatbot was able to provide resources addressing most user queries related to NDDs. The researchers found that enriching the vocabulary with synonyms and lay language terms for specific subdomains enhanced entity detection. By using a data set of numerous individuals with NDDs, the researchers developed a knowledge graph that established meaningful connections between entities, allowing the chatbot to present related symptoms, diagnoses, and resources. To the researchers' knowledge, CAMI is the first chatbot to provide resources related to NDDs. Our work highlighted the importance of engaging end users to supplement standard generic ontologies to named entities for language recognition. It also demonstrates that complex medical and health-related information can be integrated using knowledge graphs and leveraging existing large datasets. This has multiple implications: generalizability to other health domains as well as reducing the need for experts and optimizing their input while keeping health care professionals in the loop. The researchers' work also shows how health and computer science domains need to collaborate to achieve the granularity needed to make chatbots truly useful and impactful.

(J Med Internet Res 2024;26:e50182) doi:10.2196/50182

KEYWORDS

chatbot; user interface; knowledge graph; neurodevelopmental disability; autism; intellectual disability; attention-deficit/hyperactivity disorder

XSL•FO
RenderX

Introduction

Knowledge exchange in the medical domain presents multiple challenges, including accessibility, readability, and accuracy. Chatbots, or computer programs that simulate conversations, can help answer users' or caregivers' questions [1,2]. Moreover, a chatbot offers several advantageous features for medical needs: flexibility (service providers available 24/7 and from any location [3]), speed (rapid delivery of a large number of resources), privacy (confidential access for users), engagement (an appealing user interface that fosters interaction), and trustworthiness (information developed by professionals, ensuring reliability). Chatbots have already been developed for diagnosing heart disease [4,5], providing counseling for mental health [6], improving patient monitoring and medical services [7], and preventing eating disorders [8]. Some chatbots have integrated coaching elements to support youth with weight management and prediabetes symptoms [9,10], young adults with depression and anxiety [3,11], people with obesity and emotional eating issues, adults wishing to improve wellness [4,5], and young adults with a high level of stress [6]. User trust remains a key challenge faced by medical and health-related chatbots [12].

Chatbots powered by advances in natural language processing (NLP) such as large language models (LLMs; eg, ChatGPT [7]) have shown how chatbots can revolutionize the way information is shared and accessed.

Nonetheless, developing a chatbot for the medical domain is challenging, especially when targeting complex medical conditions such as neurodevelopmental disorders (NDDs). NDDs represent a diverse group of conditions affecting development, including conditions such as attention-deficit/hyperactivity disorder (ADHD), intellectual disability, autism spectrum disorder, cerebral palsy, and learning difficulty. Together, NDDs affect up to 14% of children [8] and have major implications not only for the individuals themselves but also for their families and society [13]. It is increasingly recognized that individuals with 1 NDD diagnosis (eg, autism spectrum disorder) often also present with features of ADHD or learning difficulty. Moreover, several associated conditions (often referred to as comorbidities) [14,15] can also significantly affect the clinical presentation of individuals with NDDs and their needs in terms of health, social participation, and education; for instance, sleep disorders, gastrointestinal symptoms, anxiety, depression, or even seizures are more commonly found in individuals with NDDs than in the general population. NDDs are also increasingly linked to genes. While this has paved the way for personalized medicine, it has also led to silos where parents have access to information only through associations created for their genes of interest, resulting in very limited information in terms of management for many rare conditions. Furthermore, NDDs are chronic conditions with changing manifestations and needs over the course of an individual's lifespan.

While LLM technology is evolving quickly, here, we discuss key steps to be considered when developing a chatbot for the medical domain and present how our team applied these steps to our use case of developing a chatbot for medical information related to NDDs named CAMI (Coaching Assistant for Medical/Health Information).

Methods

Consulting End Users About Their Needs to Target the Needs of Potential Users

It is important to consult directly with individuals with lived experience. While this is now common practice in industry, it remains challenging to reach out to individuals with medical issues.

Step 1: We Recommend Developing an Advisory Group That Includes Representative Individuals

In our case, we developed a national advisory group that included 9 caregivers of individuals with NDDs (we advertised for the advisory group position through associations and partners involved in NDD research in Canada). The caregivers were predominantly female (8/9, 89%), and their ages ranged from 32 to 51 years. In terms of marital status, of the 9 participants, 7 (78%) were married or had common law spouses, and 2 (22%) were divorced or separated. They had diverse levels of education: bachelor's degree (7/9, 78%), master's degree (1/9, 11%), and PhD (1/9, 11%). Of the 9 participants, 6 (67%) were employed full time, while 3 (33%) worked part time. The participants had various occupations: special education teacher (1/9, 11%), planning and programming officer (1/9, 11%), senior IT consultant (1/9, 11%), community development coordinator (1/9, 11%), academic professor (2/9, 11%), research assistant (1/9, 11%), psychologist (1/9, 11%), and manufacturer (1/9, 11%). Of the 9 individuals, 8 (89%) were White, and 1 (11%) identified as a First Nations person. All participants were biological parents.

The advisory group will not only provide direct feedback on the project but will also, through their personal and professional networks, recruit participants for testing the ideas or chatbot prototype.

Step 2: Ethical Considerations

It is important to prepare in advance a complete set of questions to be used in semistructured interviews with individuals with lived experience. The advisory group can provide key insights into the development of the consent forms, study protocol, and interview material. For developers from industry, it is key to collaborate with clinicians, association leaders, or other allied health professionals who are not only able to provide feedback but can also help engage participants in the project.

In our case, the project was approved by the research ethics board at the University of Alberta (Pro00081113). Informed consent was obtained from participants in interviews. The participants were compensated for their work as per ethics regulation.

Step 3: Consultations

We used semistructured interviews (Multimedia Appendix 1) adapted from user interface evaluation resources [12,16-18] to ask advisory group members about their current needs, their

XSL•FO
RenderX

use of web-based platforms to gather information, and the current barriers to access to information. All interviews were video recorded after receiving consent from the participants.

Our interviews allowed us to identify overlapping themes using thematic analysis [19]. We were able to identify the patterns and themes among different user groups and build a plan on how to represent the data. The advisory group members (who were individuals with lived experience) suggested that the chatbot should provide rapid access to information:

> *I like the way that the resources just pop up in the side window as the user is answering the questions.*
>
> *You could be asking me questions all day long but I'm not sure how close it can come to identifying the biggest stressors and challenges for our family. According to those responses they could target the info resources accordingly. If the chatbot doesn't get to the bottom of our challenges fast enough, as a parent, I would likely find myself turning to Google more quickly to find what I'm looking for.*

Step 4: Knowledge Mapping

It is important to identify the potential differences in the conceptual framework associated with medical conditions. These differences in mental representations, known as mental models, can lead to blind spots that may prevent the chatbot from being widely useful [20].

Developing a Database of Resources for the Chatbot

Step 1: Gathering Information

Providing trusted, exhaustive, and actionable information is key for a chatbot meant to provide medical information. It is important to partner with individuals with lived experience, associations (patients, parents, and professionals) relevant to the condition, and health authorities to obtain such information. The information about the relevant web pages, books, or other formats needs to be stored centrally and visible to others so that it can be peer reviewed. This can be done using shared sheets or websites.

In our case, we developed a database of NDD resources by leveraging existing databases (Alberta Children's Hospital, AIDE Canada, and InformAlberta [21,22]) as well as by developing a nationwide consultation with individuals with lived experience from across Canada who submitted 1422 resources. AIDE Canada and InformAlberta shared their resource database with the research team, which has been incorporated in the main CAMI database. The same data set or a superset of the data is used on the general websites of the aforementioned organizations. The individuals with lived experience used Google Sheets to add the known or found resources with the appropriate data labels. These data included the web page URL, type of resource, language, age group, location, eligibility criteria, and some important keywords. All resources were collected and annotated manually by group members. Submitted resources covered several topics such as health, education, and support programs and were grouped under the categories *core knowledge*, educational, *financial support*, and *services*.

A key feature brought up by the families of individuals we consulted was the need for an in-depth, trusted, and diverse set of resources. Therefore, we collaborated with several organizations involved in providing NDD treatment as well as individuals with lived experience. We encountered challenges in obtaining resources for diverse NDD subtypes and covering all regions of Canada at the start because we connected only with organizations that tend to have better coverage in urban areas. We found that forming a network of individuals with lived experience was very helpful and allowed us to achieve broader coverage in identifying resources (Figure 1; Textbox 1).

Figure 1. Representation of the geographic distribution of neurodevelopmental difference resources. (A) The number of resources per province and territory is represented with a circle (blue) on the map of Canada. (B) The distribution of resources by type of resource (services, core knowledge, financial support, or education) is shown.

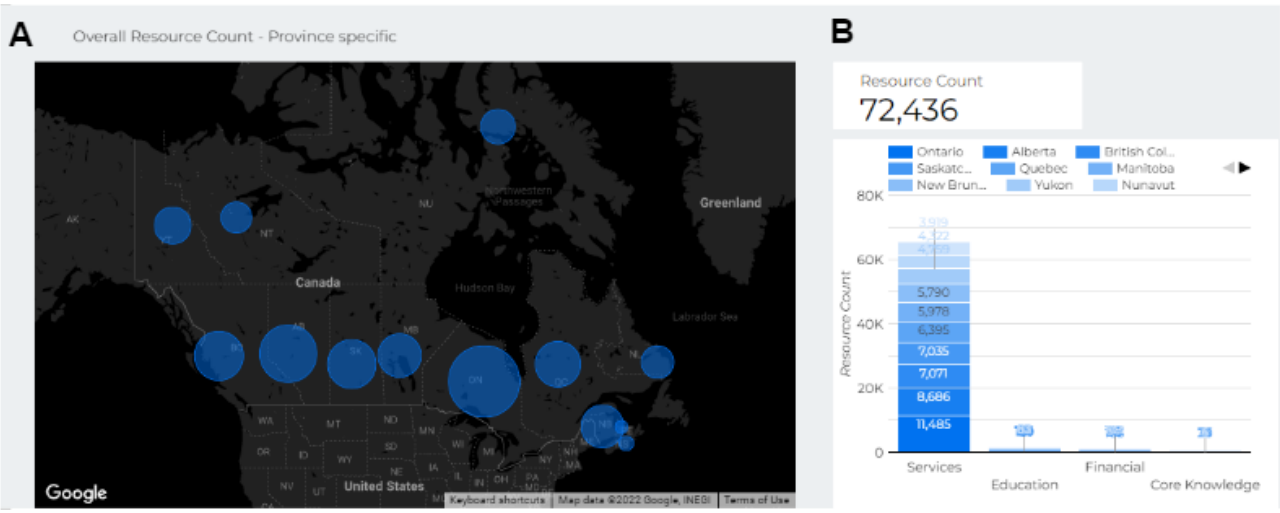

XSL•FO
RenderX

Textbox 1. Number of web resources per province and territory in the CAMI (Coaching Assistant for Medical/Health Information) database. The textbox shows the number of resources available in each province in descending order. The resource count includes all categories available for each province.

Province and number of resources

- Ontario: 11,819
- Alberta: 8986
- British Columbia: 7330
- Saskatchewan: 7285
- Quebec: 6640
- Manitoba: 6218
- New Brunswick: 6095
- Yukon: 4951
- Nunavut: 4508
- Newfoundland and Labrador: 4085

Step 2: Data Annotation

It is important to annotate the resources included in the data set in a meaningful way so that the information can be retrieved in response to related queries by future chatbot users. Again, discussions with potential users are key in capturing topics of interest and making sure that they are properly covered in the database.

We developed an automated annotation tool [23] using an NLP pipeline that uses a combination of named entity recognition, topic modeling, and text classification model to annotate the resources. All resource annotations along with their weight were stored in the Neo4j graph database (Neo4j, Inc) in the form of a weighted knowledge graph as shown in Figure 2 in the paper authored by Costello et al [23]. Ordered weighted aggregation operators are used to rank the resources that will be printed out in response to a user query.

Alternatively, one could use LLMs to extract keywords from the individual web pages. This would include selecting a suitable prompt and providing the web page content extracted from the URL. By leveraging LLMs such as ChatGPT and Google Gemini, among others, the content can be analyzed to identify and filter out important keywords from the web page.

Figure 2. Knowledge graph of co-occurring entities in neurodevelopmental disabilities or differences. Co-occurring phenotypes for autism and attention-deficit/hyperactivity disorder in the Deciphering Developmental Disorders (DDD) database are represented as nodes. Arrows or connections between nodes represent the is_associated_with relationship. Numerical values represent the weight of the connections, which refers to the number of co-occurrences, and the thickness of the arrow is proportional to this weight. The Deciphering Developmental Disorders database contains 715 and 280 individuals with autism and attention-deficit/hyperactivity disorder, respectively.

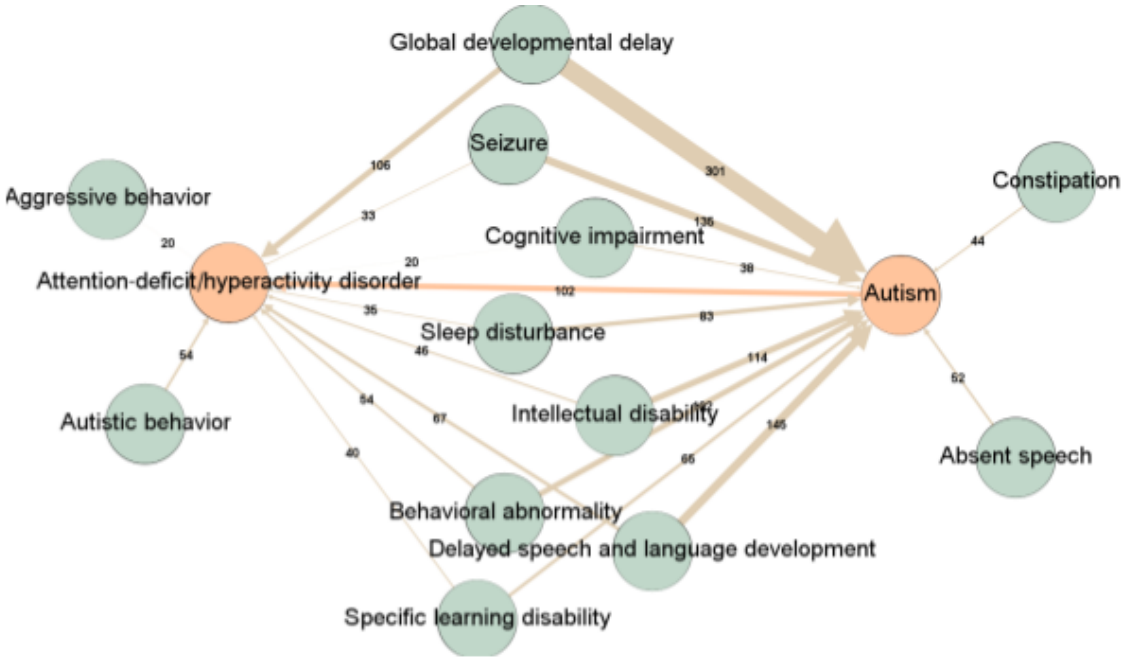

XSL•FO
RenderX

Step 3: Database Formatting

The information is stored in the database in a format that retains the dependencies and relations among different data. The data are divided into multiple tables or classes such that the data are modular and can be used independently.

We used the Neo4j database to store the database models, and the database includes several nodes, relationships, and properties. The web page content was initially stored in the MongoDB database (MongoDB, Inc) and was further annotated into multiple properties and relationships. These properties are *is_about*, *is_associated_with*, *is_located_in*, *occurred_together*, and so on. When querying the database, these properties and corresponding relationships are extracted to filter out the resources from the database and rank them in accordance with the defined parameters.

Designing and Evaluating the Chatbot User Interface Design

Overview

A crucial aspect of the chatbot relates to its user interface, which requires it to be functional, easy to use, and attractive so that it appeals to users [6-8,13]. Of particular importance is the landing page because it represents the first impression of the application and should connect with the target audience [7] by reflecting the user's [6] as well as the chatbot's goals [14]. Important components of the page design include a logo, design, fonts, colors, and layout [15] (Multimedia Appendices 2-7).

Step 1: Consultation Regarding the Interface

To streamline the process of multiple iterations, mock-up representations of the chatbot can be used in interviews with individuals with lived experience. This is important to increase the cost-effectiveness and reduce the time required for coding each version. In our case, the mock-ups, or the design wireframes of the chatbot, were shared with the participants (individuals with lived experience) by research team members on a Zoom call (Zoom Video Communications, Inc), and their feedback was recorded on Google Docs during these calls. We usually included both content experts and computer scientists to uncover gaps. This cycle should be repeated until a consensus is reached among the target user base, and the design is finalized [24]. In the design process, we used Figma (Figma Inc) [25], a collaborative design tool that allows users to work on various designs while also allowing others to review them and provide comments. Optimal user interface design uses an agile methodology that includes iterative design, implementation, and testing [26].

We started with the home page of the chatbot (Figure 3). We found that families of individuals with NDDs favored a simple design with a clear indication of the purpose of the chatbot. We also included a tutorial video about the purpose and intended use of the chatbot. The families also recommended the inclusion of the names of the institutions involved in the development of the chatbot as a mark of trusted information.

The home page was designed using Gestalt principles [27], and all closely matching content was kept on the same page. During the initial interviews, participants indicated that the home page content was not self-explanatory, clear, and trustworthy. The feedback was considered and the required changes made to showcase the authenticity of the chatbot.

Figure 3. Evolution of the chatbot home page design. (A) The initial design of the home page by our team before consultation with caregivers, which displays text indicating that the resources are trusted and diverse but does not explain what the chatbot can do. (B) The updated home page, revised after several consultations with families. It has a neutral color scheme and a graphical visualization that represents a robot conversing with a human. Adding the names of partners or affiliated organizations on the home page establishes trust in the website and proves that the product was developed by reputed organizations.

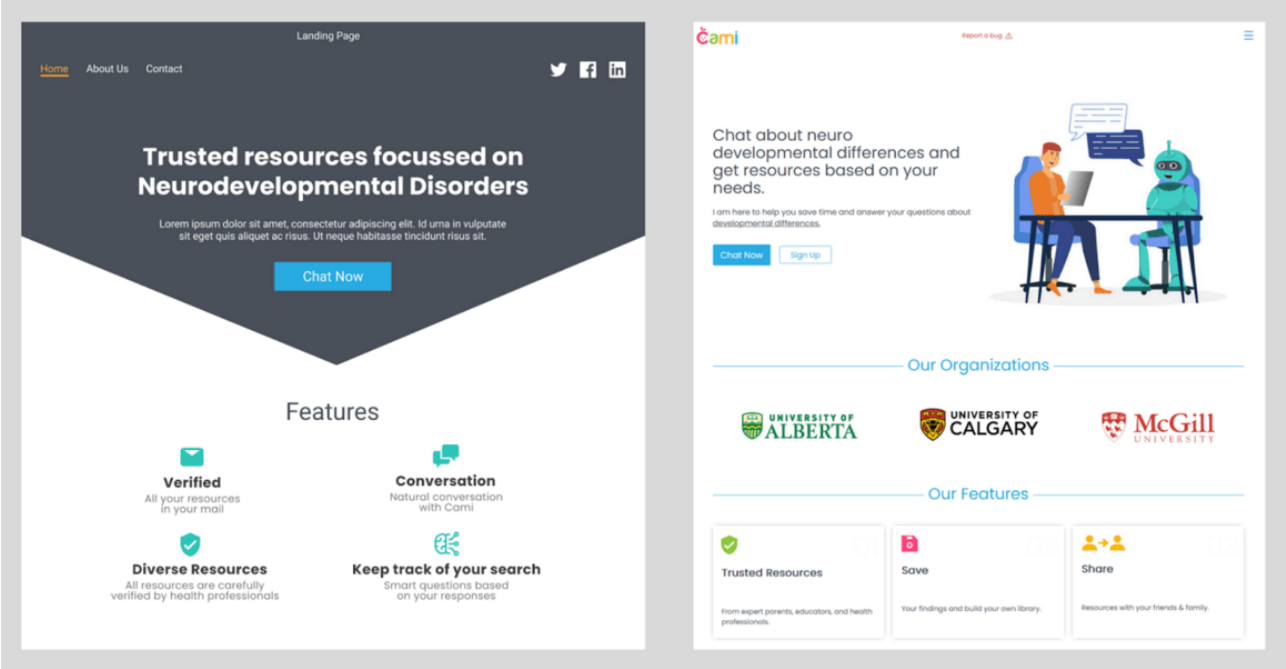

XSL•FO
RenderX

Step 2: Conversation Design

The conversation (or generic flow), which represents the exchange between the user and the chatbot, is a very important aspect in usability. Most chatbots will include an introductory text explaining the purpose of the chatbot. Subsequently, the conversation can be more or less prestructured.

An introductory video can help the user understand rapidly the aim of the chatbot as well as how best to ask questions. Importantly, the video needs to be short enough for the user to stay engaged with the chatbot.

In our case, the conversation (chat) design went through several iterations based on input from caregivers of individuals with NDDs (Figure 4). In the initial design, the users were confused when the chatbot was not able to understand them. Either they had to start the chat again or follow along with the conversation. In the final design, users have the ability to inform the chatbot if its understanding is incorrect, and they can enter their query again. Users have more flexibility in responding to certain questions by clicking "I am not sure" if they are not comfortable with answering them. In certain cases, we also provided radio buttons based on participant feedback:

> *Instead of having the user type yes or no every time, could a selection not appear using radio buttons, one for yes and one for no? It would be easier for the user.*

We developed a video explaining the aim and scope of the chatbot to avoid users asking questions outside of our domain of expertise. The video is available on the web [28].

Figure 4. Representation of the conversation (chat). (A) In the initial design, before consultation with caregivers, the user had to always answer the question, and there was no way to correct the chatbot if it misunderstood. (B) The updated design, revised after several consultations with individuals with lived experience, is much more conversational and flexible for users, allowing them to skip, make corrections, or end the chat when they want to. The design also makes use of radio buttons (yellowish-orange) that can help the user to quickly acknowledge what the chatbot is saying or correct course if needed.

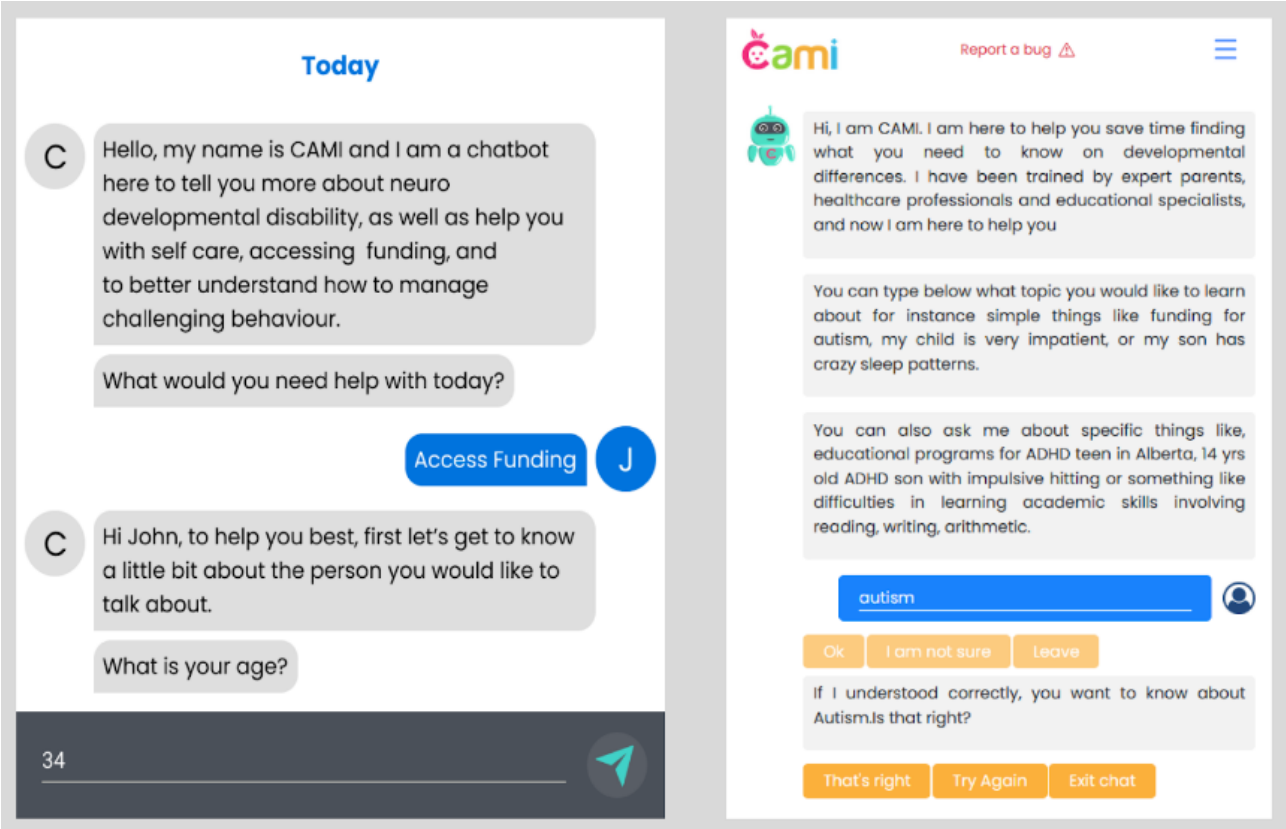

Step 3: Output Representation

Another major focus was on how to present the results of the query.

In our case, web resources identified by CAMI are presented to the user (Figure 5). This part underwent significant modification to provide information to users that would influence their decision to continue browsing or not. Users have major time constraints, and they would lose interest quickly if they are unclear about the quality of the resource. The resource card evolved from showing a title and rating to showing a static image of the website, the type of resource, tags describing the resource, and some key functions such as sharing or saving the resource (which was identified by users as key in being able to access the chatbot from multiple environments, eg, work, commute, and home). All static images are captured as screenshots and saved in the database. The families reported that in addition to being more appealing, the images would help to build trust and identify the authenticity of a site if they could see its home page. Moreover, users mentioned that they would be able to recognize and remember the site more effectively if it displayed an image. We also took advantage of the tags to allow the user to *conversewith* or *direct* the chatbot dynamically. Indeed, when the user clicks on a radio button, say, "parent" or "autism," the chatbot would respond by providing resources

XSL•FO
RenderX

related to this tag. This not only allows the user to start with a query but also enables them to navigate to other topics of interest that they may not have initially considered.

Next, we investigated how many resources to display during a single interaction. The families' feedback showed consensus at 3. Resources are presented in sets of 3 with page numbers at the bottom showing the user that there are more resources for them to look at in the future (Figure 6).

Figure 5. Representation of the resources identified by CAMI (Coaching Assistant for Medical/Health Information). (A) Early version of the resource display, before consultation with caregivers: only resource names and tags are shown to the user. (B) Updated resource card, revised after several consultations with families: each resource is represented with a static picture of the website, the website name, tags describing the types of resources (core knowledge and resources) as well as key concepts (age and location when appropriate). Radio buttons at the bottom allow users to either share the content with a friend (using SMS text messaging) or save a resource to their profile. They can also report any issue with the resource (eg, not trusted or inaccurate) to our team.

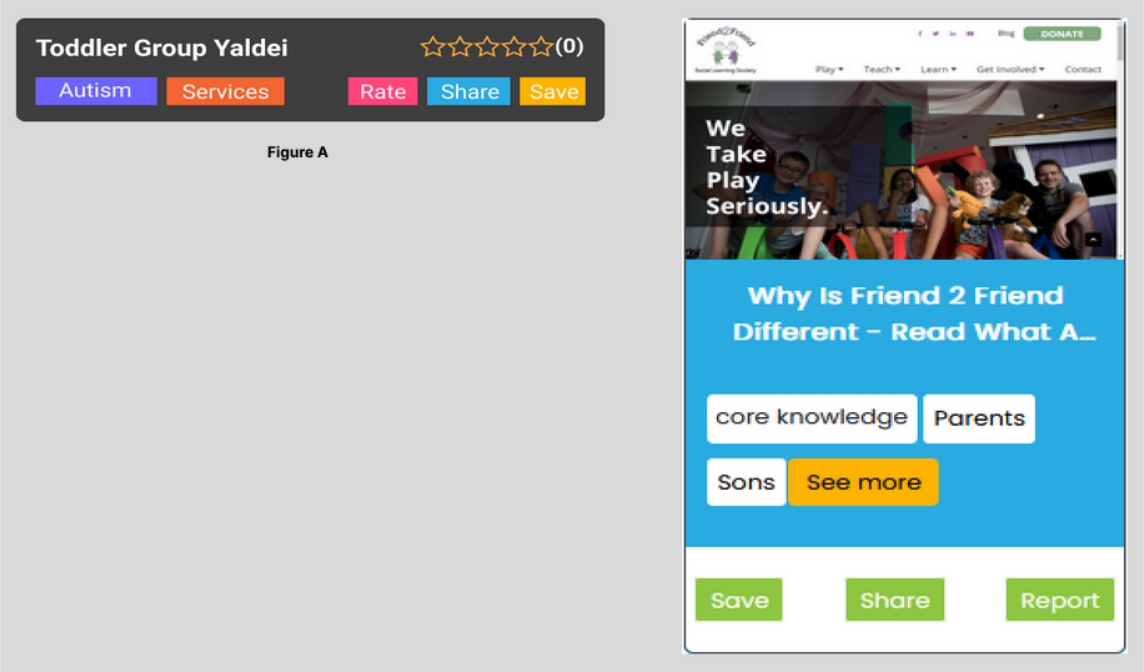

Figure 6. Resource presentation format. After receiving a query, the chatbot will provide a set of 3 resources to the user. The page numbers at the bottom indicate to the user that there are more resources available (when applicable). The user can click on Prev and Next to go to the other pages. They can also click on See More (yellowish-orange radio button) to check out more tags on the resources and decide whether they want to save, share, or report the resource.

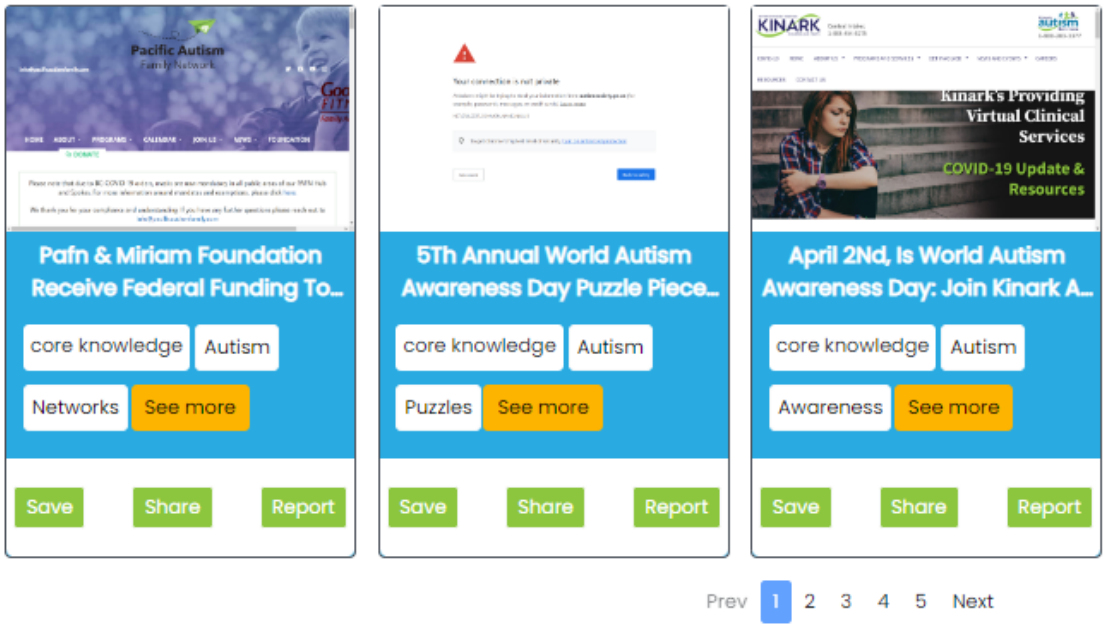

XSL•FO
RenderX

Step 4: User Engagement and Coaching Strategies

While websites have used several ways to encourage users to remain on their platform, it is important to consider the vulnerable nature of patients or their families as well as the potential detrimental effect of prolonged use. At the same time, encouraging the user to reconnect with the platform will allow repeated use and coaching.

Health coaching and care coordination [29,30] have been shown to improve health outcomes for individuals with NDDs, but they are not offered to most families because of cost, a lack of specialized health care professionals [31], access barriers due to geography, or a lack of integration into the health system (eg, insurance and social factors).

In our case, we consulted with health experts in NDDs and identified key coaching aspects that could be provided to users. We also asked users whether they were interested in being sent tips or related resources via SMS text messaging or email (Figure 7).

Figure 7. Tips provided by experts in neurodevelopmental disabilities or differences used by the chatbot. Conversation flow for sending users tips is shown, along with related information about their query. Users can also input their telephone number and receive other tips in the future.

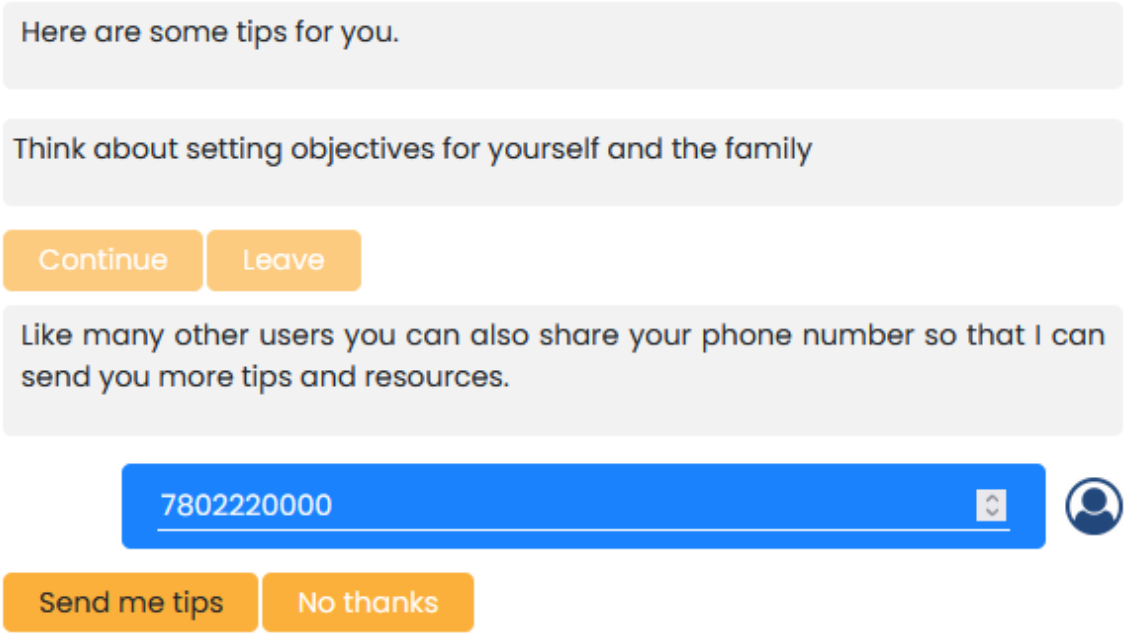

Optimizing User Input Understanding in Domain-Specific Terms

Understanding user input is an important factor in providing responses to user queries. NLP is a domain of artificial intelligence focused on developing computer programs that are able to *read* textual data, analyze the content, and extract meaningful information [32]. Advanced deep network NLP algorithms, such as named entity recognition and relation extraction, have been repurposed to identify diseases, symptoms, and drugs from user input [33-35]. Chatbots use NLP to process user text and respond to the text. While the underlying processes are different, chatbots aim to behave like people who listen and respond to their conversational partners. Day-to-day language is difficult for a computer to understand, but, using NLP, the computer is able to break up user text into a set of attribute-value pairs; for example, if the user types “I want to know if there are any services available for Global Developmental Delay in Edmonton,” the NLP pipeline would break up the text into subparts by extracting keywords: {*Services*, *Global Developmental Delay*, *Edmonton*}.

We leveraged an annotation NLP pipeline [23] to extract structured information from the user free-text query. Every query submitted by the user is divided into multiple medical-related domains such as Human Phenotype Ontology (HPO) for medical terms and Alliance of Information & Referral Systems (now known as Inform USA) for service-related terms, as well as symptom-specific terms (eg, challenging behavior), geographic location, and the age of the individual.

We found that formatting the query into multiple domains helped us identify the resources with more accuracy compared to performing a search using keywords. We experimented with certain queries from the test group and identified the related conditions for common queries. Textbox 2 shows the extraction of a single query to a list of multiple medical database keywords that relate to each other when filtering the resources.

In CAMI, certain keywords are hardcoded so that the recommendations can be domain-specific and adhere closely to user requirements. If keywords such as “knowledge” or “information” are included in the initial user query, the user will not be asked about their geographic location because location is not relevant; for instance, general knowledge regarding autism is likely consistent in Canada, the United States, or Germany, but services or policies will differ based on location. By contrast, if the keyword “services” is detected in the user query, and the location is not extracted from the query itself, the user will be prompted to provide their location (city or province) such that the recommended services can be actionable.

XSL•FO
RenderX

Textbox 2. Query analysis by domain. A query (eg, "My son is suffering from ADHD") is analyzed by using multiple annotation sources: Human Phenotype Ontology (HPO) from the Deciphering Developmental Disorders (DDD) database, the Unified Medical Language System (UMLS), a service-based annotation using the Education Resources Information Center (ERIC) thesaurus, and a tool for annotating challenging behaviors at home. Ngrams represent topic keywords from topic modeling.

Term and value

- HPO-DDD: "attention-deficit/hyperactivity disorder"
- UMLS: "mental suffering"
- ERICterm: "sons"
- cb_category: "behavioral concerns"
- Ngrams: "son, suffering, adhd"
- searchedCategory: "core knowledge"
- Location: "not found/not required"
- Age: "not found"

Chatbot Framework Coding

There are several tools available to create chatbots with diverse functionalities and their use varies with each use case. Wit.ai [36], Rasa [37], and Dialogflow [38] are some of the popular frameworks used to design rule-based chatbots.

We used the Django library in Python to create the backend for CAMI. It follows the model-view-template framework and app-based architecture, in which all classes are called apps and can work independently as well as in conjunction with other apps in the system ecosystem [36]. Each app includes a model file, a views file, a test file, and an init file. The schema was defined for the database table in the *models.py* file, where the column name is declared with its definition, which includes data type, constraints, and default values. All changes in the database model generate a migration file, which is an auto-created SQL code that creates or updates the database's internal schema [37,38]. View files include a set of functions with decorators that define the type of http request served [39]. Django (Django Software Foundation) was selected over comparable frameworks such as Flask because using Django's authentication module streamlined the development process for registration and log-in functionalities through auto-encryption, password verification, and authorization [34,35].

Once the information is extracted from the user query, it is sent to the matching engine. The matching engine incorporate the logic to find the similarity between the given input data dictionary and the list of resources from the database. The resources are annotated in the same way as the query to maintain the uniformity between the query and the list of resources to be considered for matching after the resources are filtered with the type of resource requested (Figure 8).

We optimized the system to obtain faster throughput to display the results faster. This was done by incrementing the heap size and page cache in the Neo4j configuration file. The average response time for query analysis is approximately 2 seconds, but this depends on the length of the sentence. For smaller sentence sizes, the output will be returned in a shorter time period (Figure 9).

Similarly, the time required to obtain the recommendations for the queries (Figure 9) changes with the size of the payload. The engine checks all web pages and ranks them on the basis of the topic, entity, and location (Figure 10).

In developing the chatbot, it is important to follow the CONSORT (Consolidated Standards of Reporting Trials) guidelines [39]. CONSORT guidelines are usually meant for reporting clinical trials, ensuring transparency and accuracy in the reporting of trial methods and results, but the same principles can be applied to software development as well. The guidelines include the following: (1) mention the names, credentials, and affiliations of the developers, sponsors, and owners; (2) describe the history and development process of the application and previous formative evaluations; (3) report revisions and updates; (4) provide information on quality assurance methods to ensure the accuracy and quality of the information provided; (5) ensure replicability by publishing the source code; (6) provide the URL of the application; however, because the intervention is likely to change or disappear over the years, make sure that the intervention is archived; and (7) describe how participants accessed the application, in what setting or context, and if they had to pay for access.

In our case, we applied the CONSORT guidelines as outlined in Textbox 3.

XSL•FO
RenderX

Figure 8. Chatbot workflow for query analysis. The inputs from the user interface are converted to the user query, which is passed to the annotator to procure health-related keyword pairs. The extracted data are sent to a matching engine for filtering. The matching engine checks the database to find the matching keywords and returns filtered resources.

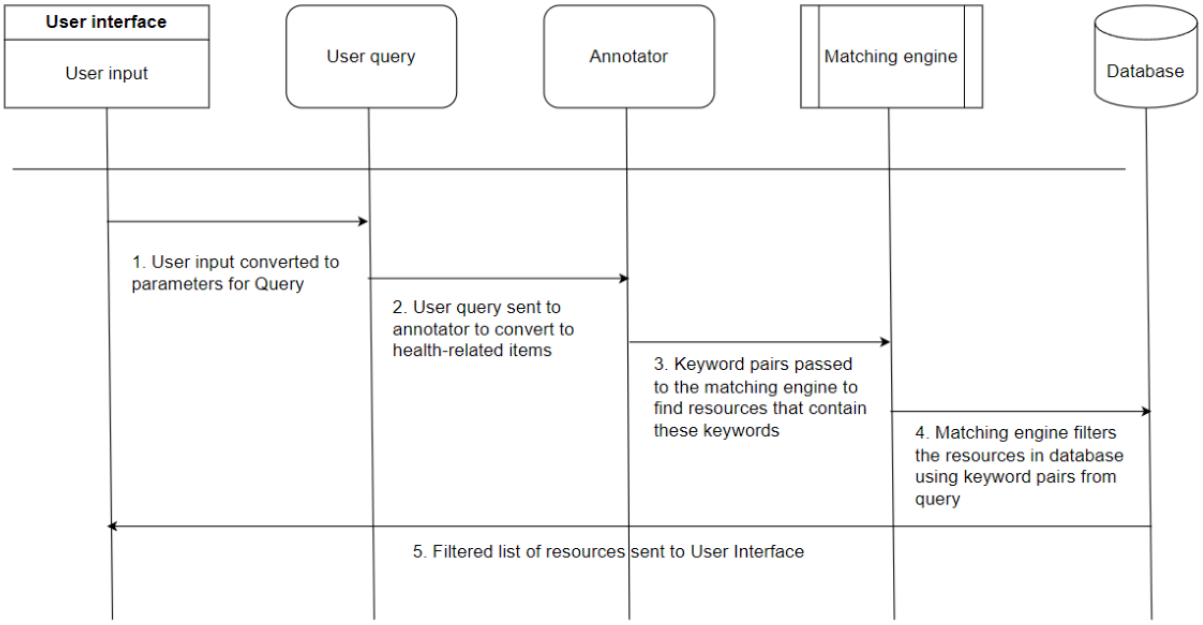

Figure 9. Query annotation time. Representation of examples of estimated time for the chatbot to analyze a user query. Two different queries are shown: (A) test 1 ("My child has sleep issues") and (B) test 2 ("My child has sleep issues and he is suffering from ADHD and cries in the night"). (A) The application programming interface timing, which includes queuing, connectivity, and response return or throughput time, is displayed. (B) For a longer query (test 2), the same parameters as the test 1 query are displayed. SSL: secure sockets layer.

XSL•FO
RenderX

Figure 10. Chatbot recommendation time. “Socket Initialization” is an event that enables the client (browser and third-party application programming interface tool) to connect to the internet. “DNS [domain name system] Lookup” retrieves the IP address from the web by deciphering the domain name, and the “TCP [transmission control protocol] Handshake” establishes a reliable connection between client and server. The “SSL [secure sockets layer] Handshake” establishes a secure communication channel between client and server. “Transfer Start” refers to the retrieval of the complete data from the server. “Download” refers to the time required to obtain all the packets at the client destination. The response times for the recommendations for both (A) test 1 (“My child have sleep issues”) and (B) test 2 (“My child have sleep issues and he is suffering from ADHD and cries in the night”) are displayed.

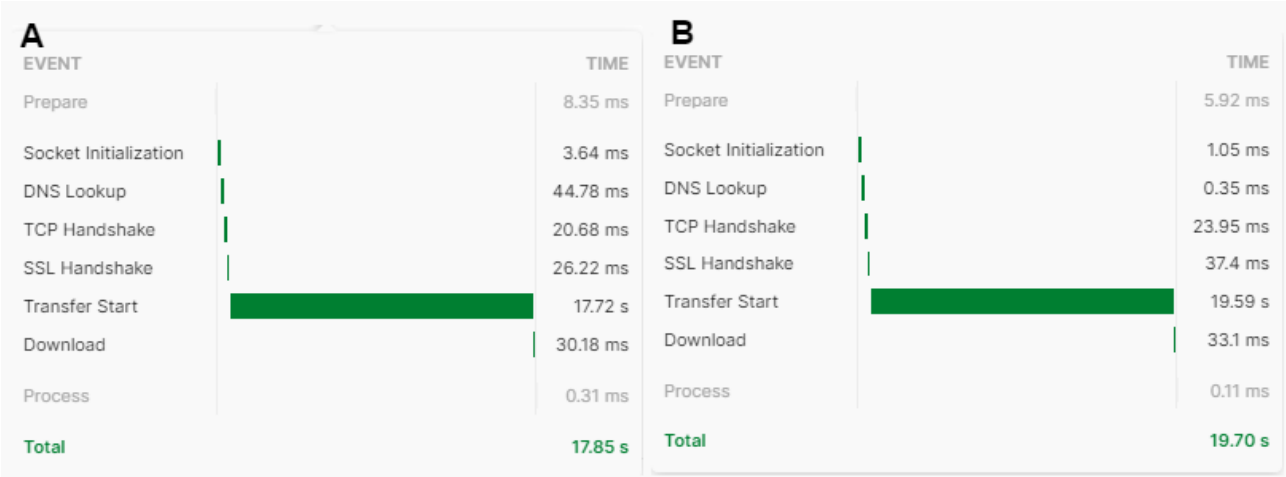

Textbox 3. Application of the CONSORT (Consolidated Standards of Reporting Trials) guidelines.

Adhering to the CONSORT guidelines

1. We provided the following information about the developers and authors: Ashwani Singla (backend), Ritvik Khanna (front-end), Manpreet Kaur (NLP [natural language processing]), and Francois Bolduc (conceptual and content expert).
2. CAMI (Coaching Assistant for Medical/Health Information) went through multiple major changes throughout the project timeline as well as numerous ongoing minor changes that were tackled on the way to building the Epic (project milestone) or the goal of the system. All code changes are committed to the GitHub repository with separate branches for front-end, backend, and NLP. Any modification to the code will go to the respective branch, and the integration developer will perform integration testing to merge the code. All modifications to the chatbot are subject to feedback from the testing group and the principal investigators of the project.
3. CAMI went through 3 major releases and 403 commits on the GitHub repository [40]. The most recent stable version for the backend functionality was released on April 16, 2023, and some changes were made on the front-end progressively. Development was frozen for all versions during the testing phase so that the same system could be tested by multiple people. There were significant changes to version 3.0 with regard to using the knowledge graph to find the connection between the entities and share the information with users.
4. During testing, the chatbot's quality was evaluated based on 3 main components: user understanding of the chatbot's functionality, user satisfaction with the overall quality of resources, and user confidence in sharing information on the platform. These metrics helped the development team to enhance the quality of the resources by placing more relevant resources on the first page and others on the subsequent pages. Similarly, the question flow changes are a result of the confidence shown by users, which was evaluated based on the user being willing to answer at least 80% of the questions asked by the chatbot.
5. CAMI is deployed on the University of Alberta’s MedIT server with the URL [41]. The default page corresponds to the home page or landing page. The backend end points can be accessed using the same URL but with different subroutings.
6. The testing method varied based on the version of the chatbot. In the first and second test phases, the participants were on a video call on Zoom (Zoom Video Communications, Inc), and the screen view was shared with them to enable them to view the website. Participants were granted remote control access to the screen to test the system, although some participants guided the proctor to click certain options. In the third version of testing, the URL was shared with the participants so that they could use a web browser to test the system. Participants could report any bugs they found. Finally, the testing group completed a feedback form.
7. The chatbot is currently freely accessible.

Establishing a Conversation Flow for the Chatbot

In the case of medical conditions, especially complex ones such as NDDs, patients present with variable or multiple issues that can all be of interest. Therefore, it is important to include a knowledge of such associated concepts to provide comprehensive information. In some cases, the core disease, say, autism or advanced stage cancer, may not be treatable, but associated conditions (eg, anxiety or constipation) may have interventions available to the patients. Concepts (or entities) and their relationships are often stored in knowledge graphs. A knowledge graph consists of a graphical and structured representation of information: terms or concepts can be stored as nodes that are connected to one another through edges that define different relationships [42]. Knowledge graphs are already being used to store information about disorders [8,13,14] and could therefore be leveraged by chatbots.

XSL•FO
RenderX

Using a knowledge graph in a health assistant or chatbot allows the question-answering system to provide (1) information regarding, for instance, disease symptoms [43,44]; (2) internet-based diagnosis and risk assessment [45]; (3) personal lifestyle interventions for serious conditions [46-48]; and (4) prediagnosis and triaging using the patient's symptoms and medical history [49,50].

We developed a knowledge graph representing key entities observed in individuals with NDDs as well as their interrelation by leveraging the largest data set of individuals with NDDs, the Deciphering Developmental Disorders (DDD) database [51]. The DDD database includes phenotypic and genotypic information along with the ages of 13,424 patients with severe and undiagnosed NDDs [30]. The DDD database labels phenotypes using the HPO. The knowledge graph contains 4181 HPO phenotypes for which we calculated the co-occurrence counts among all patient profiles and stored them in the knowledge graph using the "is_associated_with" relationship. The knowledge graph contains 357,514 "is_associated_with" relations. As all resources are already annotated with HPO phenotypes, the chatbot suggests resources to the user for other phenotypes associated with the queried phenotype.

Leveraging the DDD database allowed us to identify the concepts that co-occurred in individuals with NDDs (Figure 2). Using CAMI, we are successfully able to identify the relationship between different nodes (medical entities related to NDDs). This is possible, based on the large number of individuals in the DDD database, which makes a salient difference between the rates of co-occurrence.

The tips or suggestions are based currently on established concepts from the coaching literature developed for families or caregivers of individuals with NDDs (Figure 11); for instance, suggesting that the user set up objectives, which is a key step and a common coaching strategy when dealing with individuals with complex conditions such as NDDs. The chatbot will provide a tip, which, if the user wishes, can be sent to them via email or SMS text message.

This allowed CAMI to provide the user with the names of conditions related to their initial query. It then offered the user the opportunity to see resources for these related entities (Figure 11).

The entities extracted from the query will be passed to the knowledge graph by calling the Neo4j database's object instance, and the resources are matched [26]; for example, a query such as "My child has sleep issues" will be parsed and subcategorized into multiple key-value pairs. The aforementioned sample will return the result in the following format:

> *{HPO-DDD: [], UMLS: [Problem], EricTerm: [Sleep], AIRS: [], cb_category: [Sleep issues], location: AGE: Child, ngrams: [child, sleep, issue], relatedConditions: [], searchedCategory: [core knowledge]}*

The keywords in the query, including "sleep," "child," and "issues," are mapped with other medical data sets and categorized in the format shown.

Figure 11. Using a knowledge graph to present the user with related conditions. The user query is subcategorized into condition and age group as parameters. These parameters are passed to the knowledge graph, and it returns the closest condition within the same or similar subgroup. The condition is passed to the resource extraction engine, and new resources are displayed to the user.

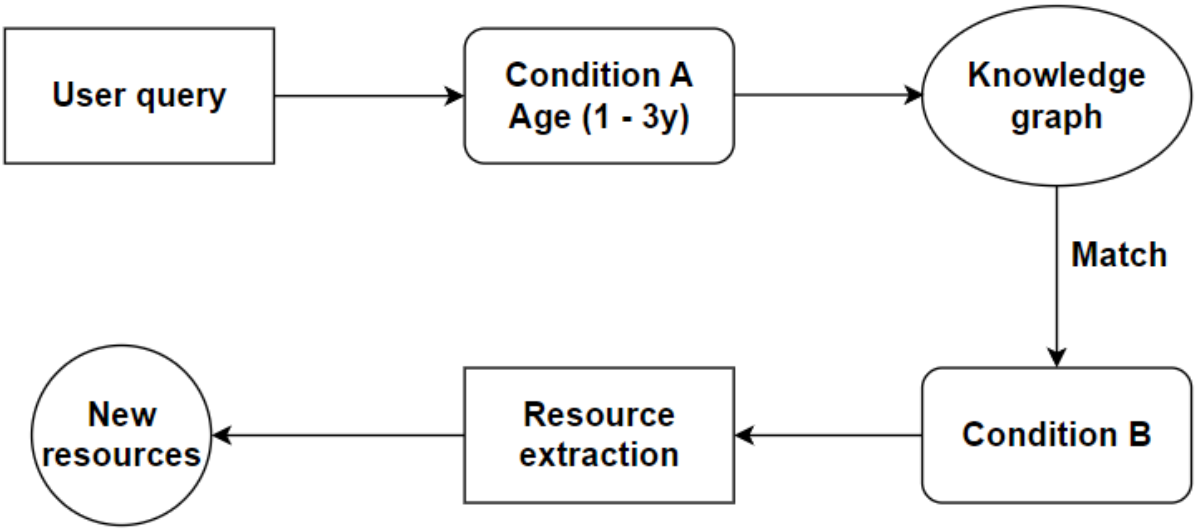

Evaluating Chatbot Outputs

Step 1: Internal Testing

It is important, especially for the medical domain, to assess the output of the chatbot internally with a content expert before conducting further testing. This process, known as *red teaming*, is carried out to make sure that the chatbot does not provide inappropriate results.

We started by reviewing the chatbot output with 3 caregivers of individuals with NDDs as well as the resources identified by the chatbot to assess their quality. Overall, these individuals with lived experience appreciated the resource quality; most of

XSL•FO
RenderX

the suggestions were from trusted websites and did not include any advertisements or false information. The chatbot was also able to recognize, for instance, the location and age and provide different recommendations integrating this information (Table 1).

Although query 1 and query 2 in Table 1 are similar, the top resource recommendation differs. When the age is included in the query, the system goes through a different pipeline and categorizes or ranks the resources on the basis of age as well. More specific resources will be ranked higher, and very general resources will be ranked lower. Similarly, if the location is added, then another ranking filter will be appended to the recommendations, and the list is sorted with 2 filters.

Table 1. Recommendations made by CAMI (Coaching Assistant for Medical and Health Information) based on user query. Top recommendations for the input query types are shown.

Query and title of web page	Reference
"My child is suffering from ADHD [attention-deficit/hyperactivity disorder]"	
"Siblings of ADHD Children: Family Dynamics"	[52]
"ADHD Behavior: Expert Discipline Skills"	[53]
"Expert Answers to Common Questions About ADHD"	[54]
"My child is suffering from ADHD and he is 7 years old"	
"Being Strength-Minded: An Introduction To Growth Mindset - Foothills Academy"	[55]
"Baby Registry Tips: Baby Clothes | Cando Kiddo"	[56]
"Global Developmental Delay Services in Edmonton for 10 year old kid"	
"Alberta Child Health Benefit | Alberta.ca"	[57]
"Services — Qi Creative Inc."	[58]

Step 2: Large-Scale Testing

Next, it is important to conduct a broader assessment of the chatbot output with a larger number of queries and evaluators. This is more challenging because patients or families have limited time. Similarly, pilot testing by clinicians is difficult to carry out.

We collaborated with undergraduate students who received training about NDDs and subsequently were asked to evaluate the resources provided by the chatbot. A total of 17 students signed up for the evaluation, but 3 (18%) dropped out, leaving 14 (82%) to evaluate the recommendations. The 14 students were divided into 3 subgroups, and each subgroup evaluated the recommendations individually for the full testing set. The evaluation sheet was divided into 4 tabs: *behavioral concerns*, *support groups or programs*, *cognitive development*, and *other topics*. Multiple user queries (from the interviews or feedback) were grouped together into each tab, and up to 50 recommendations were provided for each query. In a single subgroup, each member checked 1 query and all recommendations and answered the questions in the evaluation form (Multimedia Appendix 8). Thus, each query and its recommendations were reviewed by 3 individuals, and a majority decision (2 out of 3) was considered the final label for the evaluation.

After the evaluation, the queries were subcategorized into NDD topics such as autism, ADHD, sleep concerns, and so on, and were analyzed further (Figure 12). The analysis included the relevance scores for the top 10, top 15, and top 50 recommendations for each topic. The results varied for each topic; for ADHD, the relevance scores decreased from the top 10 recommendations to the top 15 recommendations but subsequently increased when considering the top 50 results. For autism, the relevance score gradually decreased as the considered recommendations increased, which shows that the initial recommendations are considered better and more relevant than the later recommendations. Upon analysis, we also found that the data in the database and the mapping keywords play an important role in the data ranking.

The ranking relies on multiple factors, such as location, category, Unified Medical Language System (UMLS) terms, Education Resources Information Center (ERIC) terms, type of resource required, and so on. Although we had approximately 11,000 resources that include the term *ADHD* or *attention-deficit/hyperactivity disorder*, the filters and ranking algorithms need a weighted mechanism to decide which resource to rank first. Regarding *autism*, 4169 resources in the CAMI database include the term, and CAMI performed better in terms of ranking, which shows that having more resources does not necessarily equate to better recommendations. We kept all weights at the same scale for our research, but varying weights have a better chance of improving the ranking.

XSL•FO
RenderX

Figure 12. The relevance scores for top 10, top 15, and top 50 results for each topic. (A-C) Relevant resources for attention-deficit/hyperactivity disorder (ADHD). The relevant portions are almost the same for the 3 graphs, but the not relevant and partially relevant portions vary with the increase in the number of results. (D-F) Relevant resources for autism. The top 10 results show that 93.8% of the resources are relevant or partially relevant to the user query, but this percentage reduces with the increase in the number of recommendations. (G-I) Relevant resources for behavior. The performance is similar to that for ADHD, with the relevant portion increasing with the increase in the number of recommendations from top 10 to top 15 but significantly decreasing subsequently. Dark green=relevant, light green=partially relevant, yellow=not relevant, and orange=not sure.

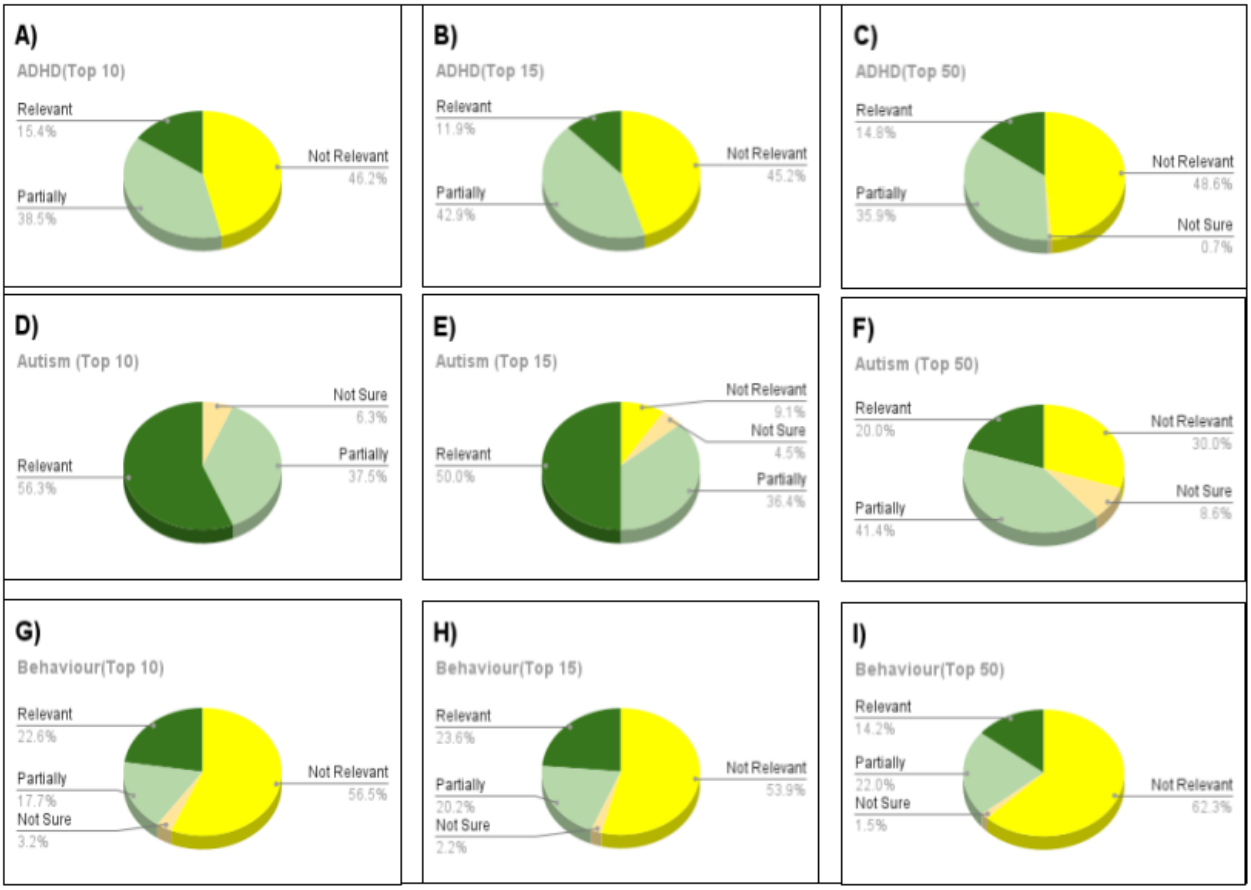

Discussion

Summary

While chatbots are emerging as a key aspect in customer service workflow for several businesses, their use in the medical domain remains limited. Although several features of chatbots, such as 24/7 availability, accessibility from remote areas, and privacy, promise a bright future for the use of chatbots in health care, significant challenges remain; for instance, users expect chatbots to be able to process their query. This relies on NLP, and while there are extensive ontologies for general domains, an understanding of the medical domain, especially when specific to a subfield such as NDDs, remains limited. The vocabulary used by users who are lay people requires the development of synonyms for most medical terms used usually by web resources that provide either core knowledge or specialized services. We found that, in addition to using the UMLS, involving medical experts as well as individuals with lived experience could greatly improve on this. This proved to be engaging for families of individuals with NDDs because they felt involved in the development of the chatbot. We would therefore encourage developers to form, as we did, advisory groups that include individuals with lived experience.

Another important aspect to consider when developing a medical domain chatbot is the level of complexity of most medical conditions. This may be related to sex differences, age-dependent differential manifestations, or associated conditions (comorbid conditions) that may influence the clinical presentation and the needs of the user; for example, an individual with an NDD is more at risk of developing seizures. This is important information for the user to be aware of when asking about, for instance, change in behavior. However, this information may not be easily available to computer scientists developing the chatbot. Even general practitioners may not consider this information shared knowledge. Interviewing domain-specific medical experts is challenging due to access or time constraints. Therefore, we developed an alternative approach by using a large data set of information and integrating the data into a knowledge graph. This allowed us to identify coassociated conditions and provide the chatbot with knowledge of the topic.

In addition to these more conceptual aspects, several practical points need to be considered when developing a chatbot in the medical domain. Possibly the most important concerns how to manage a highly multidisciplinary team (medical, social science, and computer science professionals as well as individuals with lived experience) with very limited overlap in background

XSL•FO
RenderX

knowledge. First, we noted the importance of involving potential users (caregivers) in identifying the needs of users, the features to be included, and the user interface best suited for efficient and useful mobilization of knowledge. We found that it was very important to bring together individuals who share a common interest in the topic to surmount the challenges related to differences in language and background. Indeed, we noted that an agile-inspired approach was extremely important because caregivers could better identify needs and refine approaches as they were presented with multiple iterations of the chatbot [59]. The agile approach has gained popularity because it focuses more on the potential users and the results, which enables projects to swiftly adapt to new requirements or changes as they arise.

Second, we identified several technical points that made knowledge mobilization particularly challenging for a chatbot when using automation, as opposed to manual human curation. We observed from all our user experience testing that, to be useful, the chatbot needed a wide array of resources from which to make recommendations (as was pointed out during the initial steps with the advisory group). However, we found it difficult to identify a labeled set of such resources that could be used by the chatbot. Furthermore, we noted that, for the most part, resources were labeled manually by each individual group focused on a specific health-related entity. This made it difficult for the chatbot to identify the relevant resources using a common set of vocabulary. Related to this is the difficulty for the chatbot to provide resources according to a rank most useful to users; for instance, some websites (eg, services) may have more content and therefore might include more keywords of interest than other sites with less content (which includes a list of telephone numbers for a service). At the moment, it is still unclear how simpler sites, which may be more impactful to the user, could be added.

Finally, developing a chatbot that uses a coaching approach to engage with the user and build a long-term relationship proved to be difficult. We incorporated several elements to foster engagement (customization, allowing resources to be shared, and providing expert tips), but recreating the relationship between a caregiver and a human coach remains a complex task that will require further research in mood analysis, personalization based on deeper understanding of a user's inner mental model (which can vary significantly between users), and an awareness of the course of the disease.

Limitations

Individuals with NDDs can present with different clinical features based on age and sex. Unfortunately, there is currently no database containing longitudinal data. In the future, user interactions, such as resource ratings of the resources combined with data on individuals with NDDs and cluster analysis approaches, could be used to recommend age-, sex-, or phenotypic profile–specific resources [60-63].

The conversation between chatbot and the user is still *scripted* to some degree, although the user can decide to branch out by either navigating to the related topics suggested by the chatbot or simply clicking on the tags presented with the resources provided for their query. In the future, we would like to expand on this and allow a question-generation system that will use the knowledge graph of entities related to NDDs to ask users questions to provide more specific recommendations [64]. Developing a question-generation system could allow parents to access required information without answering too many queries.

From a technical point of view, querying the database imposes significant load, resulting in prolonged wait times for results, which may impact the user experience. Using the correct database or adding indexing to the database queries can expedite data processing, but this will need many changes in database modeling and extraction queries [65]. In addition, regularly checking for annotation updates on websites is crucial to ensure their continued operation and content consistency. Sometimes, websites do not follow coding standards, and it is very difficult to spot abandoned or corrupted web pages and make sure that they do not show up in the suggestions. We could potentially develop a script that runs automatically regularly to find and remove pages that do not exist or have broken links. The website's status can be confirmed by either receiving the response using the Python request module or using the Linux *ping* command [66,67]. However, the authenticity of the website needs to be reverified after the data update to ensure that false information is not disseminated.

Finally, in its current version, CAMI is only available in English, but newcomers to the country, immigrants, and people who are comfortable in a different language should be able to access the chatbot so that it can reach, and provide resources to, a larger audience [68,69].

Future Plan

We would like to highlight the importance for our project of creating a large data set of longitudinal data to track patients' data over a period of time using the chatbot. This will allow the system to monitor patient behavior and can provide proactive solutions or resources in the early stages itself. This type of information will also be crucial for other artificial intelligence tools aiming at predicting outcomes and assessing intervention impact.

Another key aspect of chatbots operating within the medical domain will be a precise NLP system, necessitating the creation of a lexicon of domain-specific medically relevant terms alongside layperson language that can be recognized by the language models. Along the same lines, a truly dynamic system with automatic question generation will make the chatbot smarter in terms of its conversational manner. In addition to the user asking general questions, the specificity of interactions with the chatbot will be determined by the responses or queries typed by the user. Better questions will lead to better and more specific responses and better recommendations [70-72].

XSL•FO
RenderX

Authors' Contributions

FB conceptualized the study, obtained funding, and managed the project. OZ, AM, and DN assisted with the conceptualization, navigation, and development of the chatbot. MZR and OZ assisted with the technical validation of the project. AS developed and designed the architecture and backend of the chatbot and cowrote the paper. RK developed the user interface of the chatbot and cowrote the paper. MK assisted with the knowledge graph and matching engine of the chatbot and cowrote the paper. MZR assisted with the validation of the knowledge graph. CR assisted with interviews with individuals with lived experience as well as managed the resource identification team, user data validation, and interviews for requirement gathering and analysis. KK assisted with the parent advisory group and provided feedback about the chatbot from the point of view of an individual with lived experience. NR assisted with the development of the text classification model that was used in natural language processing. AM also assisted with feedback on the paper. TAB assisted with the development of the CAMI (Coaching Assistant for Medical/Health Information) logo as well as other visual aspects of the chatbot. TO assisted with the development of the parent advisory group and initial interviews.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview consent and questions.
[DOCX File , 16 KB - jmir_v26i1e50182_app1.docx]

Multimedia Appendix 2
Coaching Assistant for Medical/Health Information home page (top portion).
[PNG File , 193 KB - jmir_v26i1e50182_app2.png]

Multimedia Appendix 3
Coaching Assistant for Medical/Health Information home page (bottom portion).
[PNG File , 90 KB - jmir_v26i1e50182_app3.png]

Multimedia Appendix 4
Coaching Assistant for Medical/Health Information sign-up page.
[PNG File , 61 KB - jmir_v26i1e50182_app4.png]

Multimedia Appendix 5
Coaching Assistant for Medical/Health Information chat interface.
[PNG File , 111 KB - jmir_v26i1e50182_app5.png]

Multimedia Appendix 6
Coaching Assistant for Medical/Health Information resource card.
[PNG File , 246 KB - jmir_v26i1e50182_app6.png]

Multimedia Appendix 7
Coaching Assistant for Medical/Health Information chat radio buttons that enable users to explore related conditions.
[PNG File , 21 KB - jmir_v26i1e50182_app7.png]

Multimedia Appendix 8
Coaching Assistant for Medical/Health Information questions for recommendation evaluation.
[DOCX File , 15 KB - jmir_v26i1e50182_app8.docx]

References

1. Følstad A, Araujo T, Papadopoulos S, Law EL, Granmo OC, Luger E, et al. Chatbot Research and Design. Cham, Switzerland: Springer; 2020.
2. Grassi L, Recchiuto CT, Sgorbissa A. Knowledge-grounded dialogue flow management for social robots and conversational agents. Int J Soc Robot 2022;14(5):1273-1293 [FREE Full text] [doi: 10.1007/s12369-022-00868-z] [Medline: 35341063]
3. Youper homepage. Youper. URL: https://www.youper.ai/ [accessed 2023-03-07]
4. Welltory homepage. Welltory. URL: https://app.welltory.com/landing/main/ [accessed 2023-03-07]
5. Meditation and sleep made simple - Headspace. Headspace. URL: https://www.headspace.com/ [accessed 2023-03-07]

XSL•FO
RenderX

6. MindShift® CBT App. Anxiety Canada. URL: https://www.anxietycanada.com/resources/mindshift-cbt/ [accessed 2023-03-07]
7. ChatGPT 3.5. ChatGPT. URL: https://chat.openai.com/ [accessed 2024-05-21]
8. Child functional characteristics explain child and family outcomes better than diagnosis: population-based study of children with autism or other neurodevelopmental disorders/disabilities. Statistics Canada. 2016 Jun 15. URL: https://www150.statcan.gc.ca/n1/pub/82-003-x/2016006/article/14635-eng.htm [accessed 2023-03-06]
9. mySugr Global - Make diabetes suck less. mySugr. URL: https://www.mysugr.com/en/ [accessed 2023-03-07]
10. GlucoseZone® homepage. GlucoseZone. URL: https://glucosezone.com/home [accessed 2023-03-07]
11. Woebot health homepage. Woebot Health. URL: https://woebothealth.com/ [accessed 2023-03-07]
12. Shevat A. Designing Bots: Creating Conversational Experiences. Sebastopol, CA: O'Reilly Media; 2017.
13. Ansari NJ, Dhongade RK, Lad PS, Borade A, Yg S, Yadav V, et al. Study of parental perceptions on health and social needs of children with neuro-developmental disability and it's impact on the family. J Clin Diagn Res 2016 Dec;10(12):SC16-SC20 [FREE Full text] [doi: 10.7860/JCDR/2016/22538.9039] [Medline: 28208966]
14. Wang CH, Lee TY, Hui KC, Chung MH. Mental disorders and medical comorbidities: association rule mining approach. Perspect Psychiatr Care 2019 Jul;55(3):517-526. [doi: 10.1111/ppc.12362] [Medline: 30734309]
15. Dewey D. What is comorbidity and why does it matter in neurodevelopmental disorders? Curr Dev Disord Rep 2018 Sep 22;5(4):235-242. [doi: 10.1007/s40474-018-0152-3]
16. Levy J. UX Strategy: How to Devise Innovative Digital Products that People Want. Sebastopol, CA: O'Reilly Media; 2015.
17. Krug S. Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability. Indianapolis, IN: New Riders; 2014.
18. Unger R, Chandler C. A Project Guide to UX Design: For User Experience Designers in the Field or in the Making. London, UK: Pearson Education; 2009.
19. Delve HL, Limpaecher A. How to do thematic analysis. Delve. 2020. URL: https://delvetool.com/blog/thematicanalysis [accessed 2024-03-01]
20. Holtrop JS, Scherer LD, Matlock DD, Glasgow RE, Green LA. The importance of mental models in implementation science. Front Public Health 2021;9:680316 [FREE Full text] [doi: 10.3389/fpubh.2021.680316] [Medline: 34295871]
21. Default. AIDE Canada. URL: https://library.aidecanada.ca/ [accessed 2023-05-10]
22. ca. InformAlberta. URL: https://informalberta.ca/public/common/index_Search.do [accessed 2023-03-09]
23. Costello J, Kaur M, Reformat MZ, Bolduc FV. Leveraging knowledge graphs and natural language processing for automated web resource labeling and knowledge mobilization in neurodevelopmental disorders: development and usability study. J Med Internet Res 2023 Apr 17;25:e45268 [FREE Full text] [doi: 10.2196/45268] [Medline: 37067865]
24. Rohrer C. When to use which user-experience research methods. Nielsen Norman Group. 2022 Jul 17. URL: https://www.nngroup.com/articles/which-ux-research-methods/ [accessed 2022-10-03]
25. Figma: the collaborative interface design tool. Figma. URL: https://www.figma.com/ [accessed 2024-02-27]
26. Sharma S, Sarkar D, Gupta D. Agile processes and methodologies: a conceptual study. Int J Comput Sci Eng 2012 May;4(5) [FREE Full text]
27. Todorovic D. Gestalt principles. Scholarpedia 2008;3(12):5345. [doi: 10.4249/scholarpedia.5345]
28. cami3.0. GitHub. URL: https://github.com/ashwani227/cami3.0/blob/master/frontend/src/videos/CAMI.mp4 [accessed 2024-05-21]
29. Majnemer A, O'Donnell M, Ogourtsova T, Kasaai B, Ballantyne M, Cohen E, et al. BRIGHT coaching: a randomized controlled trial on the effectiveness of a developmental coach system to empower families of children with emerging developmental delay. Front Pediatr 2019;7:332 [FREE Full text] [doi: 10.3389/fped.2019.00332] [Medline: 31440489]
30. Ogourtsova T, O'Donnell M, De Souza Silva W, Majnemer A. Health coaching for parents of children with developmental disabilities: a systematic review. Dev Med Child Neurol 2019 Nov;61(11):1259-1265 [FREE Full text] [doi: 10.1111/dmcn.14206] [Medline: 30883717]
31. Ghanouni P, Quirke S, Blok J, Casey A. Independent living in adults with autism spectrum disorder: stakeholders' perspectives and experiences. Res Dev Disabil 2021 Dec;119:104085. [doi: 10.1016/j.ridd.2021.104085] [Medline: 34600352]
32. Chowdhary KR. Natural language processing. In: Fundamentals of Artificial Intelligence. New Delhi, India: Springer; 2020.
33. Lai T, Ji H, Zhai C, Tran QH. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021 Presented at: ACL-IJCNLP 2021; August 1-6, 2021; Online. [doi: 10.18653/v1/2021.acl-long.488]
34. El-allaly ED, Sarrouti M, En-Nahnahi N, Ouatik El Alaoui S. MTTLADE: a multi-task transfer learning-based method for adverse drug events extraction. Inf Process Manage 2021 May;58(3):102473. [doi: 10.1016/j.ipm.2020.102473]
35. Zhang Y, Lin H, Yang Z, Wang J, Zhang S, Sun Y, et al. A hybrid model based on neural networks for biomedical relation extraction. J Biomed Inform 2018 May;81:83-92 [FREE Full text] [doi: 10.1016/j.jbi.2018.03.011] [Medline: 29601989]
36. Wit.ai homepage. Wit.ai. URL: https://wit.ai/ [accessed 2024-02-27]
37. Conversational AI platform. Rasa. URL: https://rasa.com/ [accessed 2024-02-27]

XSL•FO
RenderX

38. Dialogflow. Google Cloud. URL: https://cloud.google.com/dialogflow [accessed 2024-02-27]
39. Falci SG, Marques LS. CONSORT: when and how to use it. Dental Press J Orthod 2015;20(3):13-15 [FREE Full text] [doi: 10.1590/2176-9451.20.3.013-015.ebo] [Medline: 26154451]
40. cami3.0. GitHub. URL: https://github.com/ashwani227/cami3.0 [accessed 2024-05-21]
41. Canadian Network for Personalized Interventions in Intellectual Disability. University of Alberta. URL: https://cami.med.ualberta.ca [accessed 2024-05-21]
42. Fensel D, Şimşek U, Angele K, Huaman E, Kärle E, Panasiuk O, et al. Introduction: what is a knowledge graph? In: Knowledge Graphs: Methodology, Tools and Selected Use Cases. Cham, Switzerland: Springer; 2020.
43. Zou X. A survey on application of knowledge graph. J Phys Conf Ser 2020 Mar;1487(1):012016. [doi: 10.1088/1742-6596/1487/1/012016]
44. Wang X, Luo Z, He R, Shao Y. Novel medical question and answer system: graph convolutional neural network based with knowledge graph optimization. Expert Syst Appl 2023 Oct;227:120211. [doi: 10.1016/j.eswa.2023.120211]
45. Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med 2020 Mar;103:101817. [doi: 10.1016/j.artmed.2020.101817] [Medline: 32143785]
46. Chen Y, Guo Y, Fan Q, Zhang Q, Dong Y. Health-aware food recommendation based on knowledge graph and multi-task learning. Foods 2023 May 22;12(10):2079 [FREE Full text] [doi: 10.3390/foods12102079] [Medline: 37238897]
47. Seneviratne O, Harris J, Chen CH, McGuinness DL. Personal health knowledge graph for clinically relevant diet recommendations. arXiv Preprint posted online October 19, 2021 [FREE Full text]
48. Chen Y, Sinha B, Ye F, Tang T, Wu R, He M, et al. Prostate cancer management with lifestyle intervention: from knowledge graph to Chatbot. Clin Transl Dis 2022 Feb 20;2(1):e29. [doi: 10.1002/ctd2.29]
49. Liu Z, Li X, You Z, Yang T, Fan W, Yu P. Medical triage chatbot diagnosis improvement via multi-relational hyperbolic graph neural network. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21). 2021 Presented at: SIGIR '21; July 11-15, 2021; Virtual Event, Canada. [doi: 10.26420/austinhematol.2021.1037]
50. Ma Y, Chen G, Yan W, Xu B, Qi J. Artificial intelligence chronic disease management system based on medical resource perception. In: Sun X, Zhang X, Xia Z, Bertino E, editors. Artificial Intelligence and Security. Cham, Switzerland: Springer; 2021.
51. Deciphering Developmental Disorders (DDD) project - Homepage. Deciphering Developmental Disorders. URL: https://www.ddduk.org/ [accessed 2023-05-29]
52. ADDitude asked: parent-to-parent. ADDitude. URL: https://www.additudemag.com/additude-asked-parent-to-parent/ [accessed 2024-05-23]
53. Loftin A. How to encourage good behavior in children with ADHD. ADDitude. 2021 May 10. URL: https://www.additudemag.com/adhd-behavior-expert-discipline-skills/ [accessed 2024-05-23]
54. Jones C, Brady C, Hirschfeld RM, Barkley R. Expert answers to common questions about ADHD. ADDitude. 2018 Nov 7. URL: https://www.additudemag.com/expert-answers-adhd-myths-facts/ [accessed 2024-05-23]
55. Yee A. Being strength-minded: an introduction to growth mindset. Foothills Academy. 2018 Nov 7. URL: https://www.foothillsacademy.org/community/articles/growth-mindset [accessed 2024-05-23]
56. Baby registry tips: baby clothes. CanDo Kiddo. URL: https://www.candokiddo.com/news/baby-registry-tips-baby-clothes [accessed 2024-05-23]
57. Alberta child health benefit. Government of Alberta. URL: https://www.alberta.ca/alberta-child-health-benefit [accessed 2024-05-23]
58. Qi Creative homepage. Qi Creative. URL: https://www.qicreative.com/services [accessed 2024-05-23]
59. Keita B. What are the reasons behind agile popularity? Invensis. URL: https://www.invensislearning.com/blog/reasons-behind-agile-popularity/ [accessed 2022-10-26]
60. Testa D, Jourde-Chiche N, Mancini J, Varriale P, Radoszycki L, Chiche L. Unsupervised clustering analysis of data from an online community to identify lupus patient profiles with regards to treatment preferences. Lupus 2021 Oct;30(11):1837-1843. [doi: 10.1177/09612033211033977] [Medline: 34313509]
61. Najafabadi MK, Mahrin MN, Chuprat S, Sarkan HM. Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. Comput Hum Behav 2017 Feb;67:113-128. [doi: 10.1016/j.chb.2016.11.010]
62. Vitolo M, Proietti M, Shantsila A, Boriani G, Lip GY. Clinical phenotype classification of atrial fibrillation patients using cluster analysis and associations with trial-adjudicated outcomes. Biomedicines 2021 Jul 20;9(7):843 [FREE Full text] [doi: 10.3390/biomedicines9070843] [Medline: 34356907]
63. Zou J, Luo JF, Shen Y, Cai JF, Guan JL. Cluster analysis of phenotypes of patients with Behçet's syndrome: a large cohort study from a referral center in China. Arthritis Res Ther 2021 Jan 30;23(1):45 [FREE Full text] [doi: 10.1186/s13075-021-02429-7] [Medline: 33514418]
64. Zeng J, Nakano YI. Exploiting a large-scale knowledge graph for question generation in food preference interview systems. In: Proceedings of the Companion Proceedings of the 25th International Conference on Intelligent User Interfaces. 2020 Presented at: IUI '20 Companion; March 17-20, 2020; Cagliari, Italy. [doi: 10.1145/3379336.3381504]

XSL•FO
RenderX

65. Fortis AE. Indexing research papers in open access databases. arXiv Preprint posted online May 28, 2009 [FREE Full text]
66. Matotek D, Turnbull J, Lieverdink P. Pro Linux System Administration: Learn to Build Systems for Your Business Using Free and Open Source Software. New York, NY: Apress; 2017.
67. Chandra RV, Varanasi BS. Python Requests Essentials. Mumbai, India: Packt Publishing Pvt Ltd; 2015.
68. O'Hare S. 7 reasons why having a multi-language site benefits your business. Weglot. URL: https://weglot.com/blog/reasons-why-a-multi-language-site-benefits-your-business/ [accessed 2022-10-31]
69. Fowler W. 9 reasons a multi-language website benefits your business. TechSpective. 2018 Sep 19. URL: https://techspective.net/2018/09/19/9-reasons-a-multi-language-website-benefits-your-business/ [accessed 2022-10-31]
70. Hapke H, Howard C, Lane H. Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python. New York, NY: Manning Publications; 2019.
71. Ni P, Okhrati R, Guan S, Chang V. Knowledge graph and deep learning-based text-to-GQL model for intelligent medical consultation chatbot. Inf Syst Front 2022 Jul 06:1-19 [FREE Full text] [doi: 10.1007/s10796-022-10295-0] [Medline: 35815295]
72. Weeks R, Cooper L, Sangha P, Sedoc J, White S, Toledo A, et al. Chatbot-delivered COVID-19 vaccine communication message preferences of young adults and public health workers in urban American communities: qualitative study. J Med Internet Res 2022 Jul 06;24(7):e38418 [FREE Full text] [doi: 10.2196/38418] [Medline: 35737898]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
CAMI: Coaching Assistant for Medical/Health Information
CONSORT: Consolidated Standards of Reporting Trials
DDD: Deciphering Developmental Disorders
ERIC: Education Resources Information Center
HPO: Human Phenotype Ontology
LLM: large language model
NDD: neurodevelopmental disability/difference
NLP: natural language processing
UMLS: Unified Medical Language System

Edited by A Mavragani; submitted 22.06.23; peer-reviewed by M Chatzimina; comments to author 21.07.23; revised version received 27.07.23; accepted 19.04.24; published 18.06.24.

Please cite as:
Singla A, Khanna R, Kaur M, Kelm K, Zaiane O, Rosenfelt CS, Bui TA, Rezaei N, Nicholas D, Reformat MZ, Majnemer A, Ogourtsova T, Bolduc F
Developing a Chatbot to Support Individuals With Neurodevelopmental Disorders: Tutorial
J Med Internet Res 2024;26:e50182
URL: https://www.jmir.org/2024/1/e50182
doi:10.2196/50182
PMID:38888947

©Ashwani Singla, Ritvik Khanna, Manpreet Kaur, Karen Kelm, Osmar Zaiane, Cory Scott Rosenfelt, Truong An Bui, Navid Rezaei, David Nicholas, Marek Z Reformat, Annette Majnemer, Tatiana Ogourtsova, Francois Bolduc. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Machine Learning and Health Science Research: Tutorial

Hunyong Cho[1*], PhD; Jane She[1*], BA; Daniel De Marchi[1], BSc; Helal El-Zaatari[1], BSc; Edward L Barnes[2,3], MPH, MD; Anna R Kahkoska[4,5,6], MD, PhD; Michael R Kosorok[1], MM, PhD; Arti V Virkud[7], PhD

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[2]Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[3]Center for Gastrointestinal Biology and Diseases, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[4]Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[5]Division of Endocrinology and Metabolism, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[6]Center for Aging and Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[7]Kidney Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[*]these authors contributed equally

Corresponding Author:
Jane She, BA
Department of Biostatistics
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB #7420
Chapel Hill, NC, 27599-7420
United States
Phone: 1 (919) 966 7250
Email: jane.she@unc.edu

Abstract

Machine learning (ML) has seen impressive growth in health science research due to its capacity for handling complex data to perform a range of tasks, including unsupervised learning, supervised learning, and reinforcement learning. To aid health science researchers in understanding the strengths and limitations of ML and to facilitate its integration into their studies, we present here a guideline for integrating ML into an analysis through a structured framework, covering steps from framing a research question to study design and analysis techniques for specialized data types.

(J Med Internet Res 2024;26:e50890) doi:10.2196/50890

KEYWORDS

health science researcher; machine learning pipeline; machine learning; medical machine learning; precision medicine; reproducibility; unsupervised learning

Introduction

As a brief overview, machine learning (ML) is generally characterized by model complexity and capacity for processing high-dimensional or complicated data forms and is often mentioned as an antonym to traditional statistical learning algorithms. However, this division is not clear, and ML algorithms range from traditional statistical analysis tools such as simple linear regression to cutting-edge deep neural network algorithms. While often used interchangeably with artificial intelligence (AI), ML is a subset of AI and seeks to use data-driven methods to identify patterns and make decisions. This can then be used in the field of AI to allow problem-solving and decision-making.

ML is becoming increasingly popular in the research community due to the proliferation of complex or unstructured data sets and the increased capacity and access to computing power needed to run these models. ML models can often discover sophisticated and surprising patterns in these data sets that would be difficult to discover using classical methods [1,2]. The health science research domain has been no exception to this paradigm, as the health science fields have an abundance of data well suited for these models, such as genomics sequencing data and electronic health records (EHR) data [3-6]. Applications of ML to the health field can lead to targeted interventions to provide support for health care professionals [7]. ML has also become almost indispensable to the fast-growing field of PM, which uses rich patient information to precisely target interventions [8].

XSL•FO
RenderX

This paper provides a sequential framework for health scientists intending to use ML in a research proposal and discusses types of analyses that can be done and factors to consider. It will also include a special introduction to the field of PM, which has become a popular research area with the development of new ML methodologies. Finally, we discuss some unique data types and analysis techniques specific to those application areas. In general, throughout the study design process, documentation and preplanning are highly recommended for the sake of reproducibility of the work carried out. For a visual illustration of the research pipeline flowchart, see Figure 1. There are also existing pipelines such as MLOps and CRISP used in business and industry settings that may be adapted to health science research fields; however, this paper will follow a framework more commonly seen in health science research. We relegate some technical topics, such as general sample size calculations, model training, and model tuning and validation to Multimedia Appendix 1 [9-21]. Readers are also encouraged to reference other ML primers, such as one for epidemiologists [22] and 1 for biologists [23].

Figure 1. Machine learning workflow for a health science research question, from research question refinement to results reporting, with additional considerations. The cyclic nature of the process is reflected in the arrows, as several different iterations may be considered before narrowing down to a decisive pipeline, leading to result reporting.

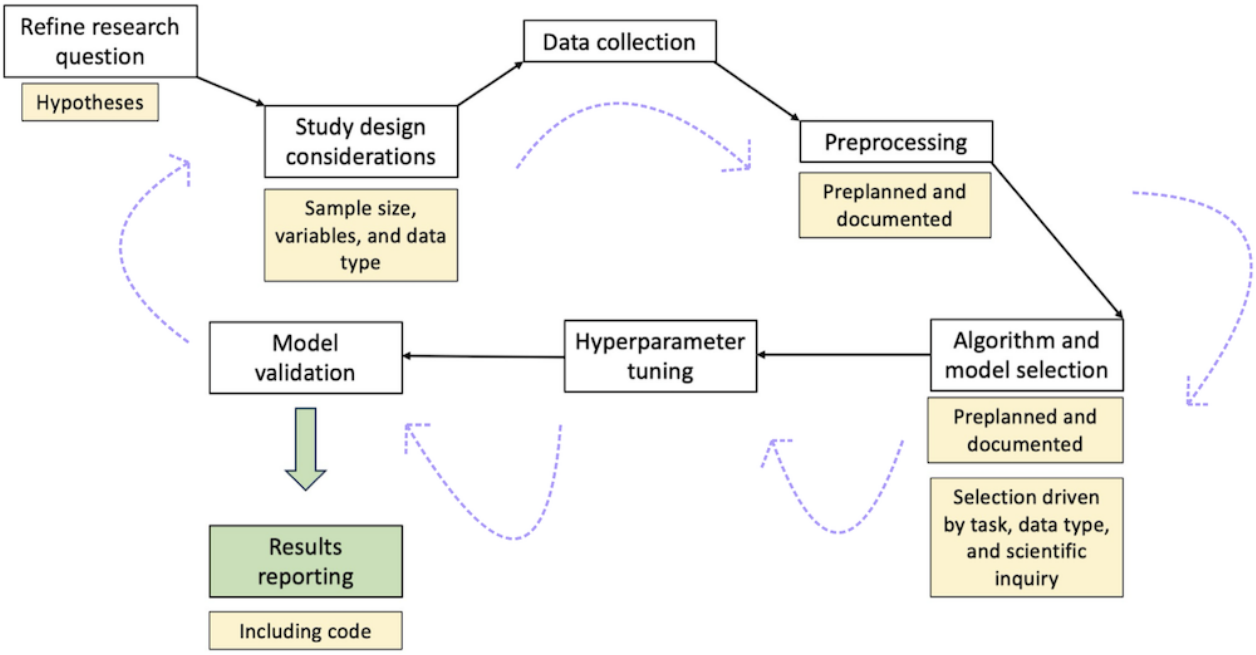

Experimentation

This section introduces step-by-step core considerations for designing an ML-involved research project.

Refining Research Questions: What Can Machine Learning Do?

ML methods can be used to answer questions for studies that may fall within the following categories: prediction, estimation, understanding causal associations, and decision support. ML can also help support main analyses as an auxiliary tool through missing data imputation, inverse propensity score weighting, dimensionality reduction, and variable selection. This last point will be covered further in the Data Collection and Preprocessing section.

Inquiries in traditional studies are often limited to the discovery and measurement of the size of certain effects or to establishing the causal relationship between variables. These are called estimation and causal inference, respectively. The recent realm of research has also expanded to prediction [24], where algorithms can predict an outcome for a patient when given a set of input variables. Functionalities such as estimation, establishing causal relationships, and prediction are meant to, in an indirect manner, support clinical decision-making. However, there are decision-support frameworks that explicitly provide recommendations through reinforcement learning (RL; defined in the section Advanced Concepts). One of the most important applications in the medical domain is PM, where the goal is to provide an optimal treatment for individual patients each with unique characteristics [8].

Prediction

As a concrete example, suppose a researcher is interested in investigating the health effects of electronic screen time use over several months [25]. This is an example of a study where the research question hinges on accurate prediction of the use of a screen. Since it is impractical and unethical to monitor an individual for several months and self-reported measures can be unreliable, the use of ML algorithms for predicting screen time is useful [26]. Prediction has also been used in the identification of early cancer diagnoses using image data analysis [27]. Classification of patients for disease screening, a prediction task, can be performed with high accuracy using ML.

Although the accuracy of a predictive algorithm is considered one of the most important virtues, interpretability [28] is another important aspect to consider, especially in health science research. Interpretability often comes at the price of reduced

XSL•FO
RenderX

accuracy, which is sometimes framed as the "interpretability-accuracy trade-off." More complex models, which may improve prediction accuracy, maybe less interpretable, as it can be difficult to trace why the model arrived at such a decision, how the predictors relate to the outcome, and how to interpret the results. Interpretability generally is used to mean being able to understand the inner workings of a model, but as evidenced by the previous sentence, it can encompass several different aspects. These can range from overall model structure, ability to explain individual predictors, transparency of decision-making processes, and more. Measuring interpretability is a challenge, as it can be context-dependent for the problem you are working on; more information on interpretability can be found at [29]. In the screen time prediction example above, interpretability is not of concern but rather maximizing the prediction accuracy, as it may not be of interest how the algorithm predicted the values, but rather the predicted values themselves.

Estimation

ML algorithms can also be used to estimate associations between exposures and health outcomes [30]. Examples include calculating the odds ratio of obesity while comparing 2 socioeconomic statuses, measuring the association between physical activity and mortality [31], and estimating the association between sociodemographic traits and diabetes prevalence [32]. However, the estimation procedures of ML algorithms are often limited to point estimation and usually lack inferential abilities such as *P* values and Bayes factors. Meaning, estimation procedures can usually find an approximate value of a parameter (like an average) through point estimation but are not usually able to output other quantities such as CIs or hypothesis tests which provide information on the population as a whole. This is because models that are nonparametric or complex may not make certain distributional assumptions, which makes quantifying CIs for a point estimate not easily doable. For an investigator wanting to confirm the positive effects of a medical treatment on patient health outcomes, ML often cannot discern whether the estimated effect size is statistically significant or not. Rather, this can be done through classical statistical tests, which possess inferential capabilities. However, this limitation is not the same as generating CIs for model performance, which is a separate procedure and generally more straightforward as model evaluation may involve data splitting or repeated sampling.

That being said, certain ML algorithms still have the potential for inferential capacity. Recently, a random forest-based framework for judging the statistical significance of heterogeneous treatment effects for individuals with specific covariate values has been developed [33]. Additionally, many other algorithms, such as support vector machine (SVM) and *k*-nearest neighbors (*k*-NN), can output CIs and *P* values for estimated effects [34,35]. However, these approaches are, in general, much less efficient than classical statistical tests and thus should be used after carefully considering the trade-off between flexibility (model specification) and efficiency (power of the test).

Causal Inference

Understanding causal associations is the activity of investigating the cause of an outcome, such as the occurrence of disease. In the statistical literature, it is known as causal inference, which provides a foundation for establishing causality [36,37]. Research questions related to understanding causal associations include estimation of average or individual treatment effects (ATE and ITE, respectively) and identification of important risk factors or subgroups for a health outcome. A rich literature for causal inference methods has been developed in statistics. For example, when estimating average treatment effects [38] from observational data, propensity score matching [39] is frequently used, which is often done using flexible models such as random forests [40]. However, its use should be carefully considered due to potential small sample size issues and covariate imbalance.

Study Design Considerations

Quality of data is a key design consideration for the successful use of ML. Given the complexity of ML, which often involves managing a vast range of input variables coming in various formats, it is crucial to plan the identification, collection, and management of these variables. Data from multiple sources—for example, clinical information, genomics data, and medical images have different dimensions—eventually needs to be aligned for downstream analyses using techniques such as feature concatenation, feature extraction, and tree and metric-based learning, so planning the process ahead of time is essential to consider any feasibility issues [41,42].

In ML studies, missing data are one of the most frequently observed issues that can harm the quality of data and can lead to bias. Thus, planning the data collection process to minimize missing data and setting up quality control checks on data entry errors is essential. Approaches to data missingness will be described in the Data Collection and Preprocessing section.

Sample Size and Strategies for Sample Size Determination

In general, ML models with tunable parameters require much larger sample sizes than traditional statistical models to achieve the same level of estimation or prediction accuracy. Since ML models usually have much weaker model assumptions than traditional parametric models, when the dimension of a parameter is much larger, more data are needed for the estimator to determine the model structure on top of estimating the mean outcome or the parameters of interest. The phenomenon when the required sample size grows exponentially with the dimension of the parameter is called the "curse of dimensionality," which is attributable to the nonparametric nature of ML models.

This relatively large sample size requirement is not the only issue, but precisely calibrating the required size is another challenge. Unlike traditional clinical trials, where the sample size of a study is planned to achieve a certain amount of power to detect a certain effect size [43,44], the sample size determination for ML has a different meaning, and there is no generic framework for it [45]. In ML, where the model performance is often measured in terms of prediction accuracy; measures, such as mean squared error and classification error

XSL•FO
RenderX

rate, are meant to be controlled under a predefined level, and the sample size that meets such prediction accuracy is to be derived.

Popular choices of evaluation metrics include mean squared error and R^2 for continuous outcomes, Brier scores for survival outcomes, classification error rate (accuracy rate) for categorical outcomes, and area under the receiver operating characteristics curve (AUC) for binary outcomes. However, these evaluation metrics should be chosen after consideration of the cost of wrong predictions and the benefits of correct predictions. For example, a model for predicting cancer may have to impose a higher cost for false negative than for false positive. Thus, a true negative rate (TNR) or a partial AUC could be considered for its evaluation measure after considering threshold selection and other possible reporting metrics. There are no "best" evaluation metrics, as this is highly dependent on the problem itself beyond the characterization of a classification or regression framework; differences in metrics can emerge when there are outliers in the data set, model comparison, and differential penalties for errors.

Although there is no deterministic sample size formula for predictive models, one can fit a learning curve on the training data for a given ML algorithm based on some evaluation measures such as the prediction error rate and AUC, which quantifies the overall accuracy of a binary classification model [45]. Essentially, the researcher is required to run the ML algorithm for the pilot data using training data and project the evaluation measure based on the fitted learning curve through the evaluation of the testing data. This evaluation measure is then used to inform the sample size or amount of data needed for the specific accuracy or statistical power desired [46].

To accurately estimate this curve, at least 2 or 3 points are required [47]. This means that the researcher is required to take at least 2 subsets of the available data and calculate 2 respective error rates. However, the pilot data might not capture all the biases present in the larger data set, as the sample may not be fully representative of the population or phenomena of interest. The researcher must be wary of generalizations using this pilot data. It is therefore recommended that a statistician trained in ML be present to assist with these technical sample size estimation procedures.

More details are included in section A of Multimedia Appendix 1 [9-21], which also includes information on how to mitigate the large sample requirement in neural networks through augmentation techniques and transfer learning (defined in the section Advanced Concepts).

Data Collection and Data Preprocessing

As previously mentioned, EHR, administrative claims, clinical trials, and longitudinal cohort data are major data types in the ML world. However, there are also "specialized data types," which require their own distinct methods of analysis due to their unique qualities. These include textual or language data, imaging, and genomics, and will be discussed in the Applications section. Due to the highly complex nature of the data being used, ML analyses often involve heavy data preprocessing. This step often requires more time than the main analysis itself and not only includes screening for erratic values, detecting and understanding outliers, and handling of missing values, but also transformation of the data into a software-friendly format, feature scaling, feature selection, dimensionality reduction, and sample splitting for validation, among others [48].

These procedures, while seemingly not important, may bring significant changes to the conclusion. For example, data preprocessing is an essential step for categorical features when using certain gradient boosting algorithms such as XGBoost, as the algorithm requires the categorical variables to be coded through mean coding or one hot coding before use in the model. Additionally, feature scaling would change the results of any methods involving Euclidean distance metrics such as principal component analysis (PCA), *k*-means, and *k*-NN.

As mentioned in the section Refining Research Questions: What Can Machine Learning Do? ML can be used as an auxiliary tool for missing data imputation [49], dimensionality reduction [50] before regression analysis, and variable selection, all of which can make an analysis more manageable.

Missing data, which typically arises in survival analysis, longitudinal studies, among other scientific studies, has great potential to create statistical bias if not accounted for in an auxiliary analysis [51]. Simply discarding observations with missing data may lead to selection bias and reduced sample size, resulting in incorrect estimation of relationships. Instead, the mechanism behind the missing data can be accounted for through an auxiliary analysis to mitigate the effects of the bias using tools such as imputation and maximum likelihood estimation. See, for example, "missForest" for imputation based on random forests [52]. As an example, [49] provides a real-use case of how ML methods can be used to impute missing data in a breast cancer problem.

Algorithm and Model Selection

The choice of an ML method largely depends on the type of task and data type. For example, linear discriminant analysis (LDA) and *k*-means clustering can only be used with continuous predictors; SVM and support vector regression can be used for classification and regression problems, respectively; and random forests and neural networks are capable of both classification and regression. Table 1 lists commonly available algorithms in each category and summarizes their benefits and drawbacks.

Once the candidate algorithms are identified, the choice of the algorithm may be driven by the scientific inquiry, as discussed in the section Refining Research Questions: What Can Machine Learning Do? Additional factors for algorithm choice may include computing resources, data limitations, and data assumptions. Figure 2 gives a list of common ML algorithms and the purposes they may be used for. The nature of the scientific study will determine the importance of interpretability in the prediction of particular phenomena. A "black box" predictive model may not clearly explain why such predictions were made, only what the predictions are [53]. For clinicians who want to attribute a specific cause of an output, these methods may be less suited for their research question, and it is suggested to use a more interpretable model. As an extreme example, consider using an ML algorithm to support the decision

XSL•FO
RenderX

of no amputation, minor amputation, or major amputation for a patient with diabetic foot ulcer. One can imagine that an interpretable ML algorithm must be preferred, as was proposed in [54], as the decision-making process needs to be clear before an amputation is carried out.

Table 1. Benefits and drawbacks of common machine learning methods in supervised and unsupervised settings. This list is not exhaustive and includes popular machine-learning algorithms in each category.

Methods	Strengths	Limitations
Benefits and drawbacks of some supervised methods		
Logistic regression and linear regression	• Interpretable • Easy implementation	• Overfitting with highly correlated data (use variable selection or shrinkage methods) • Poor performance for nonlinearly separable data
Naive Bayes	• Performs well even without conditional independence • Easy implementation	• Simple; outperformed by well-tuned, more complex models
k-nearest neighbors	• Nonparametric (no model assumptions needed) • High level of flexibility; performs well for nonlinear boundaries	• Low interpretability • Poor performance for high-dimensional data • Difficulty dealing with missing values
Support vector machine	• Performs well with high-dimensional data and nonlinear boundaries	• Low interpretability • Poor performance for imbalanced data • Usually outperformed by newer methods
Decision trees	• Interpretable for trivial data sets • Nonparametric (no model assumptions needed) • Works with nonlinear relationships • Classification for more than 2 classes	• Prone to overfitting • Difficult to interpret for nontrivial data sets
Random forest	• Handles high-dimensional data well • Reduces overfitting from decision trees • Reduces variance	• Low interpretability • Poor performance for sparse data
Gradient boosted trees	• Increased accuracy over random forests	• More difficult to implement due to tuning parameter selection
Artificial neural networks	• Works well with many data types (images, text, audio, etc) • Adaptable architecture	• Low interpretability • Overfitting if trained too long • Requires a great deal of data
Benefits and drawbacks of some unsupervised methods		
k-means clustering	• Fast, easy implementation	• Only quantitative data • No clear best way to choose k • Poor performance for noncircular cluster
Hierarchical cluster	• Reproducible • Visually interpretable by dendrograms • Cluster shape not assumed to be globular	• Poor performance on high-dimensional data • Hierarchy level must be selected
Gaussian mixture models	• Flexibility since clusters can have irregular shapes • No assumption of cluster number or level • Accommodates mixed cluster membership	• Poor performance on high-dimensional data
Linear discriminant analysis	• Interpretable • Can lower model variance over logistic regression if model assumptions are met	• Can only be used with continuous predictors • Poor performance for nonlinearly separable data (try quadratic discriminant analysis)

XSL•FO
RenderX

Figure 2. Commonly used algorithms in the supervised setting by algorithm type distinguished between classification and regression problems, as well as methods used in unsupervised learning.

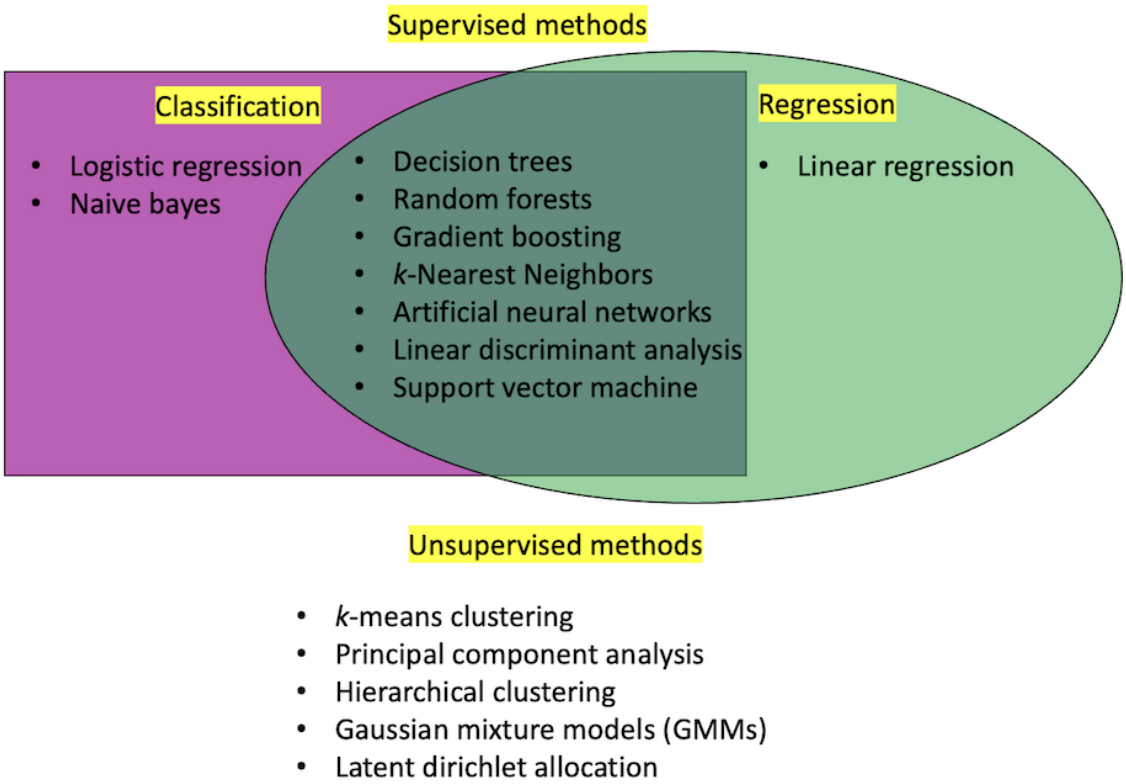

Being too open-ended about model possibilities may lead to nonreproducibility or phenomena analogous to *P*-hacking, where researchers may choose the model that leads to the highest accuracy for the data at hand after trials of multiple approaches. This highlights the importance of having a held-out test set, which is used only at the end of model development to report model performance results, as well as having an appropriate justification for model selection. More details about a held-out test set can be found in the Hyperparameter and Model Validation section. It is important to be specific enough about the goals of the analysis to justify the use of different algorithms. At the same time, being too specific can put too many undesired constraints on research, unduly limiting the use of adequate algorithms and models [55].

Hyperparameter Tuning and Model Validation

Assuming an appropriate model has been chosen, hyperparameter tuning and model validation become the next steps for an ML practitioner. In section B of Multimedia Appendix 1 [9-21], we provide guidelines for tuning hyperparameters for 2 popular ML methods—tree-based methods and neural networks. The relatively high performance of these models is achieved by adequately tuning the hyperparameters.

Model performance assessment metrics are used to determine how well a trained model performs on new, unseen data. Popular model assessment metrics include in regression: R^2 values, mean absolute error, mean square error, and in classification: recall, F_1-score, and AUC. Beyond these metrics, aspects of model performance can also include ease of use and deployment feasibility. In many health care cases, understanding how a model reached the conclusion as well as interpreting the results of the conclusion may be preferred over blackbox models, as medical decisions are made based on the results of the model. The model deployment aspect focuses on its practical use; a complicated model may not be used in resource-constrained settings, so its use may not be feasible.

Typically, the preprocessed data are split into separate training and testing sets. The term “validation set” is often used interchangeably with the term “test set” and usually refers to a portion of the data that are not used in training the model. The model is evaluated on the test set to give an unbiased estimate of model performance on unseen data. This test set cannot be touched before model fitting and is not used for training the model or tuning model parameters.

In some literature, the training data themselves may also be split into 2 separate data sets: one, dubbed the “training set,” is used to train the models to get parameter estimates, and the other, the “validation set,” is used to help tune parameters. Therefore, rather than a split into a training and testing set as previously mentioned, we have a training, validation, and testing split, where the test set is held out until model performance evaluation. The interchangeable use of “test set” and “validation set” in this case may be confusing, as they do not refer to the same thing when data are portioned this way—one must be careful in reading to understand which scenario is occurring.

When partitioning data, it is important for the test set to be representative of the data rather than having different characteristics than the training set. There are various factors

XSL•FO
RenderX

to consider when forming a test set, which can depend on the use case. For example, one may want the training and test sets to contain records from different individuals for diagnosis purposes, or for the training and test sets to contain observations from different time points on the same individuals for prognosis. Saeb et al [56] discuss examples of using different types of splitting in cross-validation and how results can differ based on the partitioning.

Section B of Multimedia Appendix 1 [9-21] continues to discuss how to validate these models using *k*-fold cross-validation, which is a validation method that uses a given sample, assuming collecting additional data is difficult. For high-dimensional data where *k*-fold cross-validation is infeasible to implement due to computational costs, an alternative approach is introduced. It should be noted that cross-validation methods still require a held-out test set to evaluate model performance at the very end; *k*-fold cross-validation is used on the training data set to tune the parameters, but it does not replace the need for a separate testing set. An introduction to how ML models are trained is also discussed in section C of Multimedia Appendix 1 [9-21] for additional information and completion in the understanding of model training.

Results and Reproducibility

The final step of any project is to report the results. Luo and colleagues [57] set up reporting standards of ML predictive model-based research for biomedical researchers, which include a list of reporting items to be included. Reporting of such items is essential to promoting reproducibility in research. Among the items are details including the nature of the study along with a background, objectives, clinical rationale, data sources, type of modeling, inclusion and exclusion criteria, time span, model validation strategies, handling of missing values, cleaning and transformation, candidate modeling techniques with justification, model selection criteria, clinical implications, and model limitations [57].

Several of these reporting items have been discussed in the paper and fall under the outlined categories of research question refinement, study design considerations, data collection and preprocessing, algorithm and model selection, hyperparameter tuning, and model validation. The goal of the list of reporting standards for predictive modeling is to encourage transparency and reproducibility to ensure credibility in the scientific community and the methodological soundness of research.

The need for transparent reporting is even more apparent when considering the nuances of ML. While studies involving ML have experimental design steps that overlap with general study considerations such as refining a research question, study design, and data collection, the use of ML in health science studies requires ML-specific considerations in terms of quality and size of data, adequacy of methods, and reproducibility. These issues are inherent in ML-involved research owing to both the complexity of data and ML models and a wide spectrum of ML methods. Readers are encouraged to read the literature [58-60] for guidelines on general study considerations. Therefore, there is a necessity for conscientious approaches to reporting.

Within each ML method, there are usually one or more hyperparameters, such as the depth and node size in tree methods and the penalty term in kernel regression [61]. Cherry-picking a hyperparameter after looking at the data multiple times may result in irreproducibility. This is why the held-out test set can only be used for result reporting and should not be used for further model development. As previously mentioned, for the sake of reproducibility, keeping documentation of the research pipeline from start to finish as outlined is also necessary.

Application

This section includes specialized data types in ML.

Natural Language Processing

Medical notes of physicians may contain important information beyond quantitative clinical records. They, however, are not readily analyzable without processing such as transcription and topic extraction. Natural language processing (NLP) does what was previously considered impossible by processing such nonstandard form of massive data into a readily analyzable format, opening huge opportunities for health science research.

NLP as a field has undergone a revolution since 2018. The seminal paper, “BERT” [62], delivered unprecedented performance on almost every major language task. NLP models using transformers (defined in the section Advanced Concepts), such as BERT and GPT [63,64], can be used for various language tasks, including classification, summarization, imputation, and prediction. The most common and useful tasks in medical NLP usually deal with hospital documents and patient interactions. For instance, NLP models can be used to automatically transcribe patient conversations, predict disease from medical notes, or impute missing values in medical forms [65]. There are many high-quality models trained on massive text corpora that can, out of the box, deliver state-of-the-art performance on almost any task.

Therefore, the first step in any NLP project is to select a pretrained model closest to the language domain being used and perform transfer learning. Transfer learning is where a model pretrained on 1 task is then trained on a related but different task. For example, for medical tasks, “Med-BERT: pretrained contextualized embeddings on large-scale structured EHR for disease prediction” and “SciBERT: a pretrained language model for scientific text” may be of use, as they are trained on similar language as is used in medical contexts [60,61,66]. The original BERT model will also work well for most purposes. The common structure of language and the size of most of these training data sets (terabytes for some models) mean that a general model will have almost certainly been exposed to any sort of text problem a researcher may be interested in due to the sheer breadth of data.

The next step is to preprocess the data so that text is converted into simple numeric tokens that can be used as inputs and process those tokens into small sets for the model to interpret. Once this is done, the model can be fine-tuned on the new data. This is done by taking the previously selected pretrained model and carefully training it with a low learning rate on the new data. Once this is done, the model should be ready for use.

XSL•FO
RenderX

Extreme care needs to be taken to derive an optimal train and test split.

Imaging

Imaging research has long been the most high-profile ML task. High-quality benchmark data sets such as the ImageNet challenge have provided robust methods for model assessment with useful pretrained networks for transfer learning [67,68]. The uses of imaging in biomedical applications are myriad. Diagnostics, such as automatic reading and classification of radiological scans or tissue biopsies, are an active area of research. Computer-assisted decision support, where ML algorithms mark anomalous areas for clinicians to investigate, is also relatively well developed. Each use case often requires highly specific knowledge and training data; we leave the specifics to clinical experts.

Convolutional neural networks are a type of deep learning model heavily used in image analyses, such as medical imaging. Convolutional neural networks can extract patterns from image pixels and are thus widely used in abnormality detection, segmentation, and classification [69].

For imaging, as with language, it is strongly encouraged to take advantage of transfer learning. High-quality models are available on many tasks. In addition, vision has demonstrated similar properties as language, and seemingly unconnected tasks often turn out to be very similar, such as categorizing pastries and segmenting tumors. Even if the images in question are very different sizes, it can still be effective to simply resize them to fit the network in question. As in all things, the best strategy is simply to experiment [70].

Genomics

With its high-dimensional nature and the growing availability of large-scale data, genomics has become one of the largest research areas where ML is used [71]. The capacity of traditional statistics is often limited without the support of ML, especially in “multi-omics,” where multiple modes of genomics data, such as DNA-seq, RNA-seq, ChIP-seq, proteomics, and metagenomics, are analyzed together.

ML is used in genomics in multiple ways. For example, ML can be used to predict a certain gene’s expression level given the corresponding DNA-seq information. Genome-wide association studies (GWAS) that aim to identify genetic variants associated with a medical condition of interest, frequently involve ML algorithms such as neural networks and random forests [72]. Zou et al [73] provide more examples of deep learning applications to genomics.

The usefulness of ML as an auxiliary tool should not be underestimated. The overwhelming number of genes is often screened using the variable importance of random forests before downstream analyses. High-dimensional features can be reduced using autoencoders [74,75] to lower-resolution data, which can then be analyzed using traditional statistical analysis tools. Graphical illustrations of the data can also be created through 2D or 3D summaries using algorithms such as tSNE and PCoA [76,77].

Despite its broad capacity in genomics research and ability to handle high-dimensional data, ML has limitations. In genomics data, the number of features outnumbering the sample size, or high dimensionality, is a commonly seen attribute; even with the use of ML, the relatively small sample size can cause reliability and reproducibility issues.

Precision Medicine

As previously mentioned, the goal of the research question of interest may be to indirectly support clinical decision-making. RL is a subset of ML that explicitly provides recommendations for decision-making at sequential time points. PM is a field where such algorithms can be applied to make treatment recommendations for individuals according to their unique characteristics.

PM starts from patient heterogeneity, where reactions to treatment vary from patient to patient [78]. For many illnesses, no panacea exists. PM seeks to recommend different treatment options for unique individuals based on their characteristics; this is formally called individualized treatment rules (ITRs) [79,80]. ITR forms the basis of PM by providing the best treatment recommendations tailored for each patient, as treatment effects can be heterogeneous among individuals. These rules are best identified with rich information about patients such as sociodemographic, clinical, and genomic data. Recently, a wealth of ITR methods have been developed [79-82].

Health care professionals and clinicians are often faced with treating a patient multiple times based on changes in response. For example, researchers can plan adaptive intervention programs for weight loss where later interventions are adjusted depending on responses to the previous treatment [83]. Such dynamic strategies are called dynamic treatment regimes (DTRs) [84,85]. DTRs aim to provide tailored decisions over more than 1 time point based on subject characteristics and their evolving contexts so that a long-term outcome of interest is optimized. The literature on this subtopic is fast-growing [86-88].

For example, patients with cancer may be given frontline chemotherapy followed by a salvage treatment if the response to the initial treatment is not successful [89]. A DTR can then be used to account for potential changes in a patient so that optimal recommendations can be made for each patient for unique stages in their disease to optimize a long-term outcome of interest, such as patient survival. An estimation of such DTR may require a large sample size. To address this issue, investigators can design a multistage randomized trial in an adaptive way through a sequential multiple assignment randomized trial (SMART) [83,90]. For instance, if a patient responded well to the first treatment, increasing the dose may not be particularly effective, and assigning a continued or decreased dose would be worth exploring. ITR and DTR are decision-support tools that provide the best treatment recommendations for patients.

SMARTs are an adaptive study design approach to finding a DTR. While SMARTs are used only for a fixed number of time points, the number of decision time points could be arbitrarily large for some problems. This is formally called an infinite-horizon setting. For example, artificial pancreas

XSL•FO
RenderX

programs decide the amount of insulin infused every minute, so that numerous actions are taken even during a day [91]. A class of DTRs that provide essentially continuous recommendations is called just-in-time adaptive intervention [92]. A Markov decision process (MDP) is often used for these problems. MDP is a class of dynamic decision rules that base their decision only on the current state information, not necessarily depending on the history of the change. V-learning [93] is an example of such an infinite-horizon DTR that uses MDP structure.

PM has a strong connection with ML. Treatment effects are often dictated by thousands of patient characteristics, such as sociodemographic, genetic, clinical, and behavioral factors. Genetic factors alone are high-dimensional and can contain millions of traits. The goal of PM is to recommend the best treatment for a patient given their unique characteristics by providing them with an ITR. For example, a clinician may be interested in delineating an optimal ITR for each patient that best achieves cancer remission [94].

Limitations and Optimizations

ML models trained on data that inaccurately represent the population cause fundamental issues such as biased prediction and suboptimality of decisions. Data, where the healthy population is poorly represented, may not be used to make a conclusion for the general population without adjustment. As ML can be used to support decision-making processes, it is also crucial that these decisions arising from the data are not discriminatory toward certain populations [95]. In the ML world, this term is called fairness. When the underlying data are biased, the ML algorithms that are trained on such data may produce biased results, which can lead to inaccurate predictions or withholding of resources. Bias in data may result from measurement bias, representation bias, and sampling bias, among others [96,97]. As an example, unbalanced gender data in the medical imaging field has led to algorithmic underperformance [98]. This discussion should be considered in the data quality assessment when planning data collection.

Advanced Concepts

Transformers

Transformers in the context of NLP are a type of deep learning model architecture that was first introduced by Vaswani in 2017 [63] and have outperformed other model types such as neural networks in both language generation and language interpretation [99]. Transformer models have a unique self-attention mechanism where the model can weigh the importance of different pieces of input data.

Transfer Learning

Transfer learning is the process of taking knowledge from one task and applying it to a different task. This can be useful in several scenarios, such as when training data for models can be difficult to collect or it is computationally expensive to train a new model. In such cases, a pretrained model used in one task can then be trained using data from a different but related task, and the information learned from the previous task can be useful in the new task. This can have reduced data requirements and improved performance as opposed to training a brand-new model, especially when using large, pretrained, publicly available models [100].

Reinforcement Learning

RL is a type of ML focused on training an algorithm to make sequential decisions in potentially changing environments to maximize a cumulative reward [101]. An agent, or decision maker, will receive some quantification of the current environment, also known as the state; the agent will then take an action that will change the state of the environment. The value associated with taking the action and transitioning to the next state is quantified by a “reward”; the agent should choose actions to maximize long-term reward. The goal may be to find the optimal sequence of actions to take to maximize the long-term reward. An application of this is decision support, which has previously been discussed. Other applications of RL range from PM to the development of self-driving cars to financial trading to the creation of ChatGPT. Those interested in the applications of RL in health care should see reference [102] for examples of use and [103] for guidelines of use.

Outlook

The field of ML has a vast trove of tools and resources for use. Its potential, though impressive and exciting, can also be a drawback, as the inherent flexibility in the analysis process gives room for the researcher to arbitrarily or questionably choose methods that result in overfitting and false discovery. This paper has provided a framework for steps involved in using ML in research, discusses analyses for specialized data formats, reviews decision support and bias in ML, and introduces PM, a popular field in the health research domain. Consulting ML experts throughout the process will not only streamline the analysis but also play a large role in legitimizing justifications for choice selections.

Furthermore, the paper has highlighted the importance of preplanned documentation to ensure transparency and foster credibility within the scientific community. The incorporation of concrete examples within the health care domain, in addition to the provision of techniques involved in specialized data types, illustrates the vast applications of ML methodologies and their potential impact in the health science field.

Overall, the specifics of different data types and the wide variety of research goals make it difficult to make more specific recommendation guidelines. However, major ML considerations and how to approach them are discussed, with specific examples. While this paper provides a general recommended research framework and major considerations for the use of ML, it is not comprehensive, as the aim was to provide a general overview of potential methods and considerations.

While ML has strong performance potential in a variety of situations, its use needs to be carefully planned through the aforementioned steps and justified to obtain the best results, as ML cannot overcome poor study design or data quality despite all its virtues. By acknowledging these limitations, the research

XSL•FO
RenderX

community can better strive for high-quality data and reproducible results to continue driving innovation in society.

Acknowledgments

JS, DDM, HE, and MRK were supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health (grant UL1TR002489).

ARK is supported by the National Center for Advancing Translational Sciences, National Institutes of Health (grant KL2TR002490). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ARK also reports receiving research grants from the Diabetes Research Connection and the American Diabetes Association and a prize from the National Academy of Medicine outside the submitted work.

AVV is supported by the National Institutes of Health T32 grant for the UNC Kidney Center.

ELB is supported by the National Institutes of Health (grant K23DK127157).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

HC, JS, DDM, HE, and MRK were responsible for paper concept and design. HC, JS, DDM, and HE contributed to the drafting of the manuscript. AVV and MRK are joint senior authors. All authors were involved in the critical revision of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary information on general sample size calculations, model training, and model tuning and validation.
[DOCX File , 34 KB - jmir_v26i1e50890_app1.docx]

References

1. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
2. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
3. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
4. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
5. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet 2007;8:343-364. [doi: 10.1146/annurev.genom.7.080505.115721] [Medline: 17550341]
6. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res 2014;42(Database issue):D975-d979 [FREE Full text] [doi: 10.1093/nar/gkt1211] [Medline: 24297256]
7. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis 2018;66(1):149-153 [FREE Full text] [doi: 10.1093/cid/cix731] [Medline: 29020316]
8. Kosorok MR, Laber EB. Precision medicine. Annu Rev Stat Appl 2019;6(1):263-286 [FREE Full text] [doi: 10.1146/annurev-statistics-030718-105251]
9. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv Preprint posted online on November 19, 2015. [doi: 10.48550/arXiv.1511.06348]
10. Incorporating nesterov momentum into adam. International Conference of Learning Representations. URL: https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ [accessed 2024-01-16]
11. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv Preprint posted online on June 12, 2017. [doi: 10.48550/arXiv.1706.03825]
12. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. 2019 Presented at: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining; August 4-8, 2019; Anchorage, AK. [doi: 10.1145/3292500.3330701]
13. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6(1):60. [doi: 10.1186/s40537-019-0197-0]

XSL•FO
RenderX

14. Xiao D, Huang Y, Qin C, Liu Z, Li Y, Liu C. Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 2019;233(14):5131-5143. [doi: 10.1177/0954406219840381]
15. Tsinganos P, Cornelis B, Cornelis J, Jansen B, Skodras A. Data augmentation of surface electromyography for hand gesture recognition. Sensors (Basel) 2020;20(17) [FREE Full text] [doi: 10.3390/s20174892] [Medline: 32872508]
16. Stone CJ. Optimal rates of convergence for nonparametric estimators. Ann Stat 1980:1348-1360 [FREE Full text]
17. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006;7:91 [FREE Full text] [doi: 10.1186/1471-2105-7-91] [Medline: 16504092]
18. Qi Y. Random forest for bioinformatics. Ensemble Mach Learn 2012:307-323 [FREE Full text]
19. Zeiler M. ADADELTA: An adaptive learning rate method. arXiv Preprint posted online on December 22, 2012. [FREE Full text]
20. Bergstra DY, James, Cox D. HYPEROPT: A python library for optimizing the hyperparameters of machine learning algorithms. 2013 Presented at: Proceedings of the 12th Python in Science Conference; July 29-August 2, 2013; Austin, TX. [doi: 10.25080/Majora-8b375195-003]
21. Kingma D, Ba J. Adam: a method for stochastic optimization. arXiv Preprint posted online on December 22, 2014. [FREE Full text]
22. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019;188(12):2222-2239 [FREE Full text] [doi: 10.1093/aje/kwz189] [Medline: 31509183]
23. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23(1):40-55. [doi: 10.1038/s41580-021-00407-0] [Medline: 34518686]
24. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507-2509 [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
25. Stiglic N, Viner RM. Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. BMJ Open 2019;9(1):e023191 [FREE Full text] [doi: 10.1136/bmjopen-2018-023191] [Medline: 30606703]
26. Fletcher RR, Chamberlain D, Richman D, Oreskovic N, Taveras E. Wearable sensor and algorithm for automated measurement of screen time. 2016 Presented at: 2016 IEEE Wireless Health (WH); October 25-27, 2016; Bethesda, MD p. 1-8. [doi: 10.1109/WH.2016.7764564]
27. Debelee TG, Schwenker F, Ibenthal A, Yohannes D. Survey of deep learning in breast cancer image analysis. Evol Syst 2019;11(1):143-163. [doi: 10.1007/s12530-019-09297-2]
28. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019;267:1-38 [FREE Full text] [doi: 10.1016/j.artint.2018.07.007]
29. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv Preprint posted online on February 28, 2017. [FREE Full text] [doi: 10.48550/arXiv.1702.08608]
30. Levin KA. Study design v. case-control studies. Evid Based Dent 2006;7(3):83-84 [FREE Full text] [doi: 10.1038/sj.ebd.6400436] [Medline: 17003803]
31. Rockhill B, Willett WC, Manson JE, Leitzmann MF, Stampfer MJ, Hunter DJ, et al. Physical activity and mortality: a prospective study among women. Am J Public Health 2001;91(4):578-583 [FREE Full text] [doi: 10.2105/ajph.91.4.578] [Medline: 11291369]
32. Cowie CC, Eberhardt MS. Sociodemographic characteristics of persons with diabetes. In: Diabetes in America. 2nd Edition. Bethesda, MD: National Institute of Diabetes and Digestive and Kidney Diseases; 1995:85-116.
33. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018;113(523):1228-1242. [doi: 10.1080/01621459.2017.1319839]
34. Jiang B, Zhang X, Cai T. Estimating the confidence interval for prediction errors of support vector machine classifiers. J Mach Learn Res 2008;9:521-540 [FREE Full text]
35. Arbach L, Reinhardt J, Bennett D, Fallouh G. Mammographic masses classification: comparison between Backpropagation Neural Network (BNN), K Nearest Neighbors (KNN), and human readers. 2003 Presented at: CCECE 2003—Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436); May 4-7, 2003; Montreal, QC p. 1441-1444. [doi: 10.1109/ccece.2003.1226174]
36. Hernan MA, Robins JM. Causal inference. GRASS (Grup de Recerca en Anàlisi eStadística de la Supervivència). 2010. URL: https://grass.upc.edu/en/seminar/presentation-files/causal-inference/chapters-1-i-2/@@download/file/BookHernanRobinsCap1_2.pdf [accessed 2024-01-10]
37. Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY: Cambridge University Press; 2015.
38. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A 2019;116(10):4156-4165 [FREE Full text] [doi: 10.1073/pnas.1804597116] [Medline: 30770453]
39. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010;29(3):337-346 [FREE Full text] [doi: 10.1002/sim.3782] [Medline: 19960510]
40. Zhao P, Su X, Ge T, Fan J. Propensity score and proximity matching using random forest. Contemp Clin Trials 2016;47:85-92 [FREE Full text] [doi: 10.1016/j.cct.2015.12.012] [Medline: 26706666]

XSL•FO
RenderX

41. Li Y, Ngom A. Data integration in machine learning. 2015 Presented at: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 9-12, 2015; Washington, DC p. 1665-1671. [doi: 10.1109/bibm.2015.7359925]
42. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform 2018;19(2):325-340 [FREE Full text] [doi: 10.1093/bib/bbw113] [Medline: 28011753]
43. Zhong B. How to calculate sample size in randomized controlled trial? J Thorac Dis 2009;1(1):51-54 [FREE Full text] [Medline: 22263004]
44. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20(5):453-458 [FREE Full text] [doi: 10.1136/emj.20.5.453] [Medline: 12954688]
45. Vapnik VN. The Nature of Statistical Learning Theory. New York, NY: Springer; 1999.
46. Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: a learning-curve analysis. J Mach Learn Res 2003;4:211-255 [FREE Full text]
47. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta 2013;760:25-33. [doi: 10.1016/j.aca.2012.11.007] [Medline: 23265730]
48. Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Riche NH, et al. Research directions in data wrangling: visualizations and transformations for usable and credible data. Inf Vis 2011;10(4):271-288. [doi: 10.1177/1473871611415994]
49. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 2010;50(2):105-115. [doi: 10.1016/j.artmed.2010.05.002] [Medline: 20638252]
50. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 2015;16:241 [FREE Full text] [doi: 10.1186/s13059-015-0805-z] [Medline: 26527291]
51. Guo CY, Yang YC, Chen YH. The optimal machine learning-based missing data imputation for the cox proportional hazard model. Front Public Health 2021;9:680054 [FREE Full text] [doi: 10.3389/fpubh.2021.680054] [Medline: 34291028]
52. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112-118 [FREE Full text] [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
53. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. 2018 Presented at: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); October 01-03, 2018; Turin, Italy p. 80-89. [doi: 10.1109/dsaa.2018.00018]
54. Xie P, Li Y, Deng B, Du C, Rui S, Deng W, et al. An explainable machine learning model for predicting in-hospital amputation rate of patients with diabetic foot ulcer. Int Wound J 2022;19(4):910-918 [FREE Full text] [doi: 10.1111/iwj.13691] [Medline: 34520110]
55. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv Preprint posted online on November 13, 2018. [FREE Full text]
56. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience 2017;6(5):1-9 [FREE Full text] [doi: 10.1093/gigascience/gix019] [Medline: 28327985]
57. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
58. Peat JK, Mellis C, Williams K, Xuan W. Health Science Research: A Handbook of Quantitative Methods. London, UK: Routledge; 2020.
59. Woodward M. Epidemiology: Study Design and Data Analysis, 3rd Edition. Boca Raton, FL: CRC press; 2013.
60. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 2018;16(1):29 [FREE Full text] [doi: 10.1186/s12916-018-1017-7] [Medline: 29490655]
61. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 2020;415:295-316. [doi: 10.1016/j.neucom.2020.07.061]
62. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018. [FREE Full text]
63. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017). New York, NY: Curran Associates; 2017.
64. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Papers With Code. 2018. URL: https://paperswithcode.com/paper/improving-language-understanding-by [accessed 2024-01-19]
65. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4(1):86 [FREE Full text] [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
66. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

XSL•FO
RenderX

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3–7, 2019; Hong Kong, China p. 3615-3620 URL: https://aclanthology.org/D19-1371.pdf [doi: 10.18653/v1/d19-1371]
67. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;6(6):84-90. [doi: 10.1145/3065386]
68. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 Presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL p. 248-255. [doi: 10.1109/cvpr.2009.5206848]
69. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42(11):226. [doi: 10.1007/s10916-018-1088-1] [Medline: 30298337]
70. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016;3:9 [FREE Full text] [doi: 10.1186/s40537-016-0043-6]
71. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16(6):321-332 [FREE Full text] [doi: 10.1038/nrg3920] [Medline: 25948244]
72. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet 2020;11:350 [FREE Full text] [doi: 10.3389/fgene.2020.00350] [Medline: 32351543]
73. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019;51(1):12-18. [doi: 10.1038/s41588-018-0295-5] [Medline: 30478442]
74. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 2019;20(7):389-403. [doi: 10.1038/s41576-019-0122-6] [Medline: 30971806]
75. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac Symp Biocomput 2015;20:132-143 [FREE Full text] [Medline: 25592575]
76. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605 [FREE Full text]
77. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform 2016;17(4):628-641 [FREE Full text] [doi: 10.1093/bib/bbv108] [Medline: 26969681]
78. Grutters JPC, Sculpher M, Briggs AH, Severens JL, Candel MJ, Stahl JE, et al. Acknowledging patient heterogeneity in economic evaluation: a systematic literature review. Pharmacoeconomics 2013;31(2):111-123. [doi: 10.1007/s40273-012-0015-4] [Medline: 23329430]
79. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc 2012;107(449):1106-1118 [FREE Full text] [doi: 10.1080/01621459.2012.695674] [Medline: 23630406]
80. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics 2012;68(4):1010-1018 [FREE Full text] [doi: 10.1111/j.1541-0420.2012.01763.x] [Medline: 22550953]
81. Zhang Y, Laber EB, Tsiatis A, Davidian M. Using decision lists to construct interpretable and parsimonious treatment regimes. Biometrics 2015;71(4):895-904 [FREE Full text] [doi: 10.1111/biom.12354] [Medline: 26193819]
82. Athey S, Wager S. Efficient policy learning. Stanford Institute for Economic Policy Research (SIEPR). 2017. URL: https://siepr.stanford.edu/publications/working-paper/efficient-policy-learning [accessed 2024-01-10]
83. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. Transl Behav Med 2014;4(3):260-274 [FREE Full text] [doi: 10.1007/s13142-014-0265-0] [Medline: 25264466]
84. Kosorok MR, Moodie EE. Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2015.
85. Chakraborty B, Murphy SA. Dynamic treatment regimes. Annu Rev Stat Appl 2014;1:447-464 [FREE Full text] [doi: 10.1146/annurev-statistics-022513-115553] [Medline: 25401119]
86. Zhao YQ, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Stat Assoc 2015;110(510):583-598 [FREE Full text] [doi: 10.1080/01621459.2014.937488] [Medline: 26236062]
87. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika 2013;100(3):681-684 [FREE Full text] [doi: 10.1093/biomet/ast014] [Medline: 24302771]
88. Zhang Y, Laber EB, Davidian M, Tsiatis AA. Interpretable dynamic treatment regimes. J Am Stat Assoc 2018;113(524):1541-1549 [FREE Full text] [doi: 10.1080/01621459.2017.1345743] [Medline: 30774169]
89. Song Z, Yu X, Lou G, Shi X, Zhang Y. Salvage treatment with apatinib for advanced non-small-cell lung cancer. Onco Targets Ther 2017;10:1821-1825 [FREE Full text] [doi: 10.2147/OTT.S113435] [Medline: 28367065]
90. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol 2012;8:21-48 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032511-143152] [Medline: 22224838]
91. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: a systematic review. Artif Intell Med 2020;104:101836. [doi: 10.1016/j.artmed.2020.101836] [Medline: 32499004]

XSL•FO
RenderX

92. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
93. Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR. Estimating dynamic treatment regimes in mobile health using V-learning. J Am Stat Assoc 2020;115(530):692-706 [FREE Full text] [doi: 10.1080/01621459.2018.1537919] [Medline: 32952236]
94. Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer 2015;15(12):747-756 [FREE Full text] [doi: 10.1038/nrc4015] [Medline: 26536825]
95. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. Nature 2018;559(7714):324-326. [doi: 10.1038/d41586-018-05707-8] [Medline: 30018439]
96. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54(6):1-35. [doi: 10.1145/3457607]
97. Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv Preprint posted online on July 31, 2018. [FREE Full text] [doi: 10.48550/arXiv.1808.00023]
98. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020;117(23):12592-12594 [FREE Full text] [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
99. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. : Association for Computational Linguistics; 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Virtual p. 38-45 URL: https://aclanthology.org/2020.emnlp-demos.6/ [doi: 10.18653/v1/2020.emnlp-demos.6]
100. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. Proc IEEE 2021;109(1):43-76. [doi: 10.1109/jproc.2020.3004555]
101. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res 1996;4:237-285 [FREE Full text] [doi: 10.1613/jair.301]
102. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv 2021;55(1):1-36. [doi: 10.1145/3477600]
103. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med 2019;25(1):16-18 [FREE Full text] [doi: 10.1038/s41591-018-0310-5] [Medline: 30617332]

Abbreviations

AI: artificial intelligence
ATE: average treatment effects
AUC: area under the receiver operating characteristics curve
DTR: dynamic treatment regime
EHR: electronic health record
GWAS: genome-wide association studies
ITE: individual treatment effects
ITR: individualized treatment rule
k-NN: k-nearest neighbors
LDA: linear discriminant analysis
MDP: Markov decision process
ML: machine learning
NLP: natural language processing
PCA: principal component analysis
PM: precision medicine
RL: reinforcement learning
SMART: sequential multiple assignment randomized trial
SVM: support vector machine
TNR: true negative rate

XSL•FO
RenderX

Edited by A Mavragani; submitted 15.07.23; peer-reviewed by C Sáez, Z Ren, M Aczon; comments to author 11.10.23; revised version received 30.11.23; accepted 21.12.23; published 30.01.24.

Please cite as:
Cho H, She J, De Marchi D, El-Zaatari H, Barnes EL, Kahkoska AR, Kosorok MR, Virkud AV
Machine Learning and Health Science Research: Tutorial
J Med Internet Res 2024;26:e50890
URL: https://www.jmir.org/2024/1/e50890
doi:10.2196/50890
PMID:38289657

©Hunyong Cho, Jane She, Daniel De Marchi, Helal El-Zaatari, Edward L Barnes, Anna R Kahkoska, Michael R Kosorok, Arti V Virkud. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Mental Wellness Self-Care in Singapore With mindline.sg: A Tutorial on the Development of a Digital Mental Health Platform for Behavior Change

Janice Huiqin Weng[1], MPP; Yanyan Hu[1], PhD; Creighton Heaukulani[1], PhD; Clarence Tan[1], BSc; Julian Kuiyu Chang[1], PhD; Ye Sheng Phang[1], BEng; Priyanka Rajendram[1], MBBS, MS; Weng Mooi Tan[1], PharmD, EMBA; Wai Chiong Loke[1], MBBS, MMed, MBA; Robert J T Morris[1,2], PhD

[1]MOH Office for Healthcare Transformation, Singapore, Singapore

[2]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Creighton Heaukulani, PhD
MOH Office for Healthcare Transformation
1 Maritime Square
Harbourfront Centre
Singapore, 099253
Singapore
Phone: 65 66793209
Email: creighton.heaukulani@moht.com.sg

Abstract

Background: Singapore, like the rest of Asia, faces persistent challenges to mental health promotion, including stigma around unwellness and seeking treatment and a lack of trained mental health personnel. The COVID-19 pandemic, which created a surge in mental health care needs and simultaneously accelerated the adoption of digital health solutions, revealed a new opportunity to quickly scale innovative solutions in the region.

Objective: In June 2020, the Singaporean government launched mindline.sg, an anonymous digital mental health resource website that has grown to include >500 curated local mental health resources, a clinically validated self-assessment tool for depression and anxiety, an artificial intelligence (AI) chatbot from Wysa designed to deliver digital therapeutic exercises, and a tailored version of the website for working adults called mindline at work. The goal of the platform is to empower Singapore residents to take charge of their own mental health and to be able to offer basic support to those around them through the ease and convenience of a barrier-free digital solution.

Methods: Website use is measured through click-level data analytics captured via Google Analytics and custom application programming interfaces, which in turn drive a customized analytics infrastructure based on the open-source platforms Titanium Database and Metabase. Unique, nonbounced (users that do not immediately navigate away from the site), engaged, and return users are reported.

Results: In the 2 years following launch (July 1, 2020, through June 30, 2022), the website received >447,000 visitors (approximately 15% of the target population of 3 million), 62.02% (277,727/447,783) of whom explored the site or engaged with resources (referred to as nonbounced visitors); 10.54% (29,271/277,727) of those nonbounced visitors returned. The most popular features on the platform were the dialogue-based therapeutic exercises delivered by the chatbot and the self-assessment tool, which were used by 25.54% (67,626/264,758) and 11.69% (32,469/277,727) of nonbounced visitors. On mindline at work, the rates of nonbounced visitors who engaged extensively (ie, spent ≥40 seconds exploring resources) and who returned were 51.56% (22,474/43,588) and 13.43% (5,853/43,588) over a year, respectively, compared to 30.9% (42,829/138,626) and 9.97% (13,822/138,626), respectively, on the generic mindline.sg site in the same year.

Conclusions: The site has achieved desired reach and has seen a strong growth rate in the number of visitors, which required substantial and sustained digital marketing campaigns and strategic outreach partnerships. The site was careful to preserve anonymity, limiting the detail of analytics. The good levels of overall adoption encourage us to believe that mild to moderate mental health conditions and the social factors that underly them are amenable to digital interventions. While mindline.sg was primarily used in Singapore, we believe that similar solutions with local customization are widely and globally applicable.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e44443) doi:10.2196/44443

KEYWORDS

digital mental health; artificial intelligence; AI; AI chatbot; digital therapeutics; mental health; mental wellness; mobile phone

Introduction

Mental Health and Wellness in Singapore

Mental health conditions are now responsible for 1 in 5 years lived with disability [1], and depression is the leading cause of disability worldwide, affecting >300 million people [2]. In Singapore, 13.9% of the adult population experienced at least 1 mood, anxiety, or alcohol use disorder in their lifetime [3]. As in other Asian countries, the treatment gap is large; 78.6% of individuals met the criteria for a 12-month mental health condition and were in need of mental health care but did not receive help or treatment [4]. Asia faces unique challenges in improving mental health. The region experiences a shortage of mental health professionals [5,6]. For example, in 2016, Singapore and South Korea reported 4.19 and 5.79 psychiatrists, respectively, working in the mental health sector per 100,000 population [7]; most Southeast Asian countries have <1 psychiatrist per 100,000 population [6]. In comparison, Australia reported 13.53 psychiatrists per 100,000 population in 2015, and the United Kingdom reported 17.98 psychiatrists per 100,000 population in 2019 [7,8].

Digital solutions for mental health are poised to overcome these challenges. Wearable devices and smartphones provide increased accessibility and less-stigmatizing avenues for mental health care [9], and there is evidence to support the use of internet-based interventions for mental health treatment [10,11] and self-help [12,13]. The COVID-19 pandemic has accelerated the adoption of these tools; there was an increase in the use of digital mental health technology during the pandemic in the United States [14], and in Singapore, there was a rise in demand for digital mental health services and mobile apps [15]. The market today is large; there are an estimated 10,000 to 20,000 mental health mobile apps on the market today [16], and mental health start-ups garnered >US $1 billion in funding in the first half of 2020 [17].

In June 2020, driven by the COVID-19 pandemic, the MOH Office for Healthcare Transformation (under the Ministry of Health in Singapore) launched a digital mental health resource website called mindline.sg. Initially intended to serve emergent mental health needs following lockdowns and major societal disruption, the initiative has evolved to address the broader gap in digital mental wellness provision in the country. In particular, there is a need to recognize that mental health exists on a spectrum and maintaining a healthy state of mental well-being involves being equipped with the necessary skills and knowledge for coping with regular life stresses [18]. To this end, *mental wellness* (or *mental well-being*) can be defined as a positive state of mental health beyond the absence of a mental health condition; being mentally well means you are able to think, feel, and act in ways that create a positive impact on your physical and social well-being [18]. Where the health care system primarily serves the needs of those diagnosed with a mental health condition, digital mental wellness self-help resources may provide a scalable complement that enables people to maintain their mental wellness.

To the best of our knowledge, mindline.sg is the first of its kind in Asia. The platform provides >500 curated local resources, a clinically validated self-assessment tool for depression and anxiety, and an emotionally intelligent artificial intelligence (AI) chatbot from Wysa delivering a suite of interactive digital therapies (the chatbot will be described in detail in the Methods section). Data-driven digital marketing strategies, the addition of a tailored product targeting working adults, and strategic outreach partnerships have all expanded the platform's reach and ability to engage, with 447,783 unique visitors (447,783/3,000,000, 14.93% of the target population, which includes those who are able to understand and access a website such as mindline.sg independently, including youth, adults, digitally literate older adults; a baseline of 3 million is taken) in the 2 years following launch. Of these visitors, 62.02% (277,727/447,783) of these visitors engage with the site by exploring self-care content, completing exercises, or conversing with the AI chatbot. Furthermore, 10.54% (29,271/277,727) of those users return. The framework around the platform's development is driven by behavior change goals and unique requirements for mental wellness in Asia.

In this descriptive paper, we outline this framework, which focuses on anonymity, trustworthiness, multiple levels of content engagement, customizability, and tight integration into the local mental health care ecosystem, and we present data and analyses accumulated throughout implementation, culminating in a set of recommendations to guide the development of similar platforms in the region and around the world.

The mindline.sg Platform

The goals of the platform and the theory of change underlying it can be summarized as described herein. The designers of the platform believe that providing individuals with localized and trusted digital mental health resources that have low barriers to use will lead them to adopt healthy behaviors that maintain their mental wellness. As such, mindline.sg is a stress management and coping website that can be accessed and used anonymously from anywhere for free (the Wysa chatbot component, however, is only licensed for free use within Singapore). The site first exposes individuals to mental health education (when they view content and read articles) and subsequently aims to train them in mental wellness self-management skills (through practicing therapeutic and self-care exercises). The platform recognizes that a wide variety of social, economic, and personal lifestyle factors can cause stress or emotional distress, and therefore, feeling down, anxious, or depressed is something that anyone in the society can encounter at any point in their lives. Therefore, the site includes resources for financial stress, work support, caregiver support, keeping fit and healthy, etc. The goal of the platform is to empower all residents in Singapore to take charge of their mental health and to be able to offer support to those

XSL•FO
RenderX

around them; this is accomplished when individuals develop self-care habits and coping skills as a lifestyle and possess the know-how to self-refer to professional services when needed. This theory is summarized in Figure 1. Broader goals to which the initiative may contribute include destigmatizing mental health and increasing mental health literacy. At the time of creating mindline.sg, there were no comparable services available in Singapore or Asia and few global resources offered similar depth and breadth.

Figure 1. Depiction of the theory of change and goals of the mindline.sg initiative. eCBT: electronic cognitive behavioral therapy; EAP: employee assistance program.

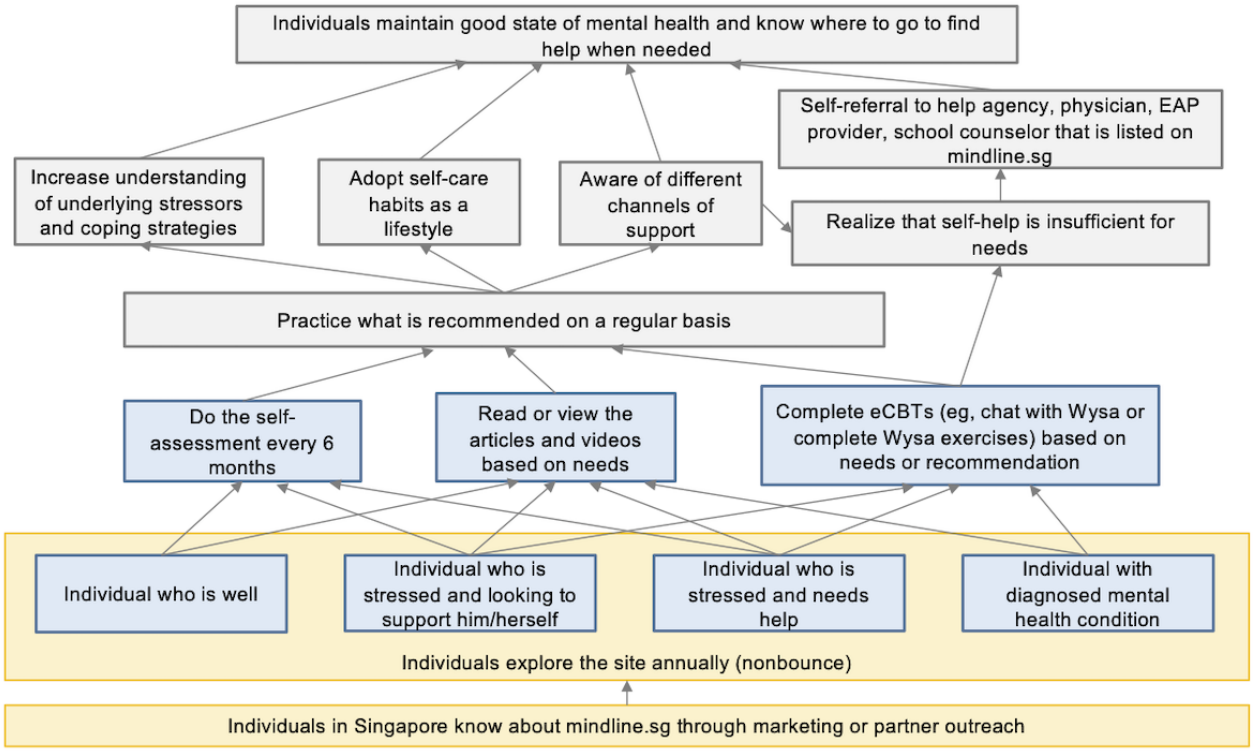

Methods

A Framework for the Development of a Digital Mental Health Platform

A New Framework

We describe a framework for the development of mindline.sg that is centered around anonymity, building trust, progressive engagement levels, and customizability to different population segments. The framework was developed referencing other recognized digital mental health frameworks, such as the American Psychiatric Association's App Evaluation Framework [19]. Most if not all such frameworks, however, are designed to develop and evaluate mobile apps. As we will discuss, the mindline.sg platform was implemented as a web app to satisfy requirements to reduce barriers to use and increase perceived anonymity. Our framework nonetheless satisfies or fits within the components of existing frameworks for digital mental health apps.

Trust and Anonymity

An advisory and editorial board (AEB) consisting of mental health and health care leaders and experts from among the health care ecosystem in Singapore was formed to drive decision-making around platform development and to bring in delivery partnerships and value-adding stakeholders. The AEB also included a clinical review panel consisting of qualified mental health clinicians. Together, the AEB and clinical review panel provided detailed advice; oversight; support; and validation of the clinical effectiveness and safety of the content, referral and escalation methodology, etc, providing assurances to users that the quality of resources and methodologies used are trustworthy.

Singaporeans have a high level of confidence in the government, with 81.5% of respondents in a World Values Survey study indicating *a great deal of confidence* or *quite a lot of confidence* in the local government [20]. In addition, >70% of survey respondents in Singapore indicated a high confidence in state institutions such as the civil service, as compared to the media or other nongovernment organizations [20]. Therefore, the government agencies behind mindline.sg are highlighted on the website landing page (Figure 2A). While government origins on the landing page will give some users confidence, overt associations with the government could deter others. A generic name and website address mindline.sg was therefore chosen, as opposed to a website address with a .gov.sg suffix.

XSL•FO
RenderX

Figure 2. The mindline.sg landing page on (A) a mobile device and (B) the home page on a desktop computer.

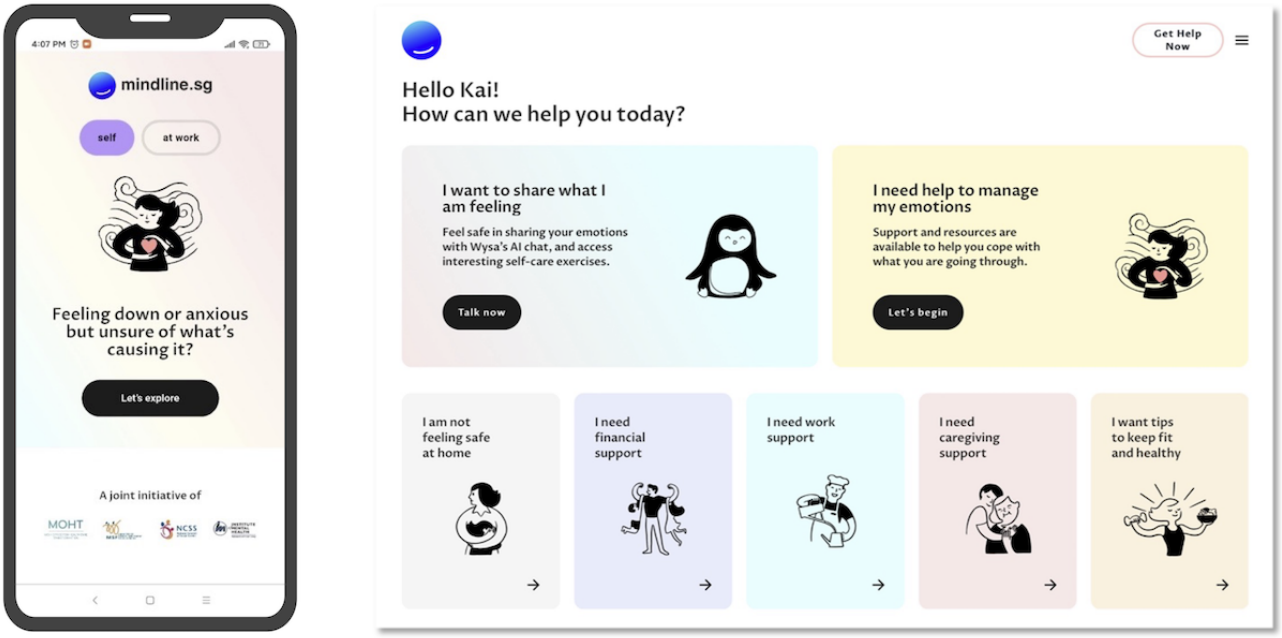

(A) (B)

Stigma remains a challenge for mental health care in Singapore [21-23], perhaps partially explaining the 78.6% treatment gap described in the Introduction section. Moreover, there is a general fear (especially in Asia) of being "labeled" with a mental health condition. Users must therefore be able to access the platform anonymously to minimize barriers to use. A substantial emphasis is therefore placed on anonymity; no personally identifiable data are collected by mindline.sg. Only an anonymous user ID associated with the browser and device that accessed mindline.sg (ie, a cookie ID) is generated and collected by the site, along with an optionally provided nickname (to personally address the user; Figure 2B) or an age range (to personalize the promotion of content). To further emphasize anonymity, the website is accessible through a web app, instead of a mobile app, and requests no registration for use. (We note that this feature may merely improve *perceived* anonymity, as it may be possible to offer an equivalently anonymous experience with mobile apps.)

Progressive Engagement for Different Stages of Learning and Readiness for Change: Incorporating a Self-Assessment Tool and Digital Therapeutics Delivered via an AI Chatbot

The platform strategically offers multiple levels of progressively deeper engagement driving behavior change goals. The first step is making users aware of mindline.sg through marketing and outreach efforts or visiting the site. Raising awareness of mindline.sg is a necessary outcome for the initiative. Reach is achieved through search engine optimization, marketing (through both traditional and digital means such as advertisements on social media), and building strategic partnerships (eg, with educational institutes and organizations supporting a profession) to leverage influence with targeted audiences (such as youth and working adults).

The second step is for users to view and read articles, increasing their mental health literacy and awareness of local resources, including channels of support. Curated and trustworthy resources are provided on the site.

The final step is to interactively practice a self-care skill offered on the site. One such skill is self-assessment, for which the site offers a self-assessment tool, developed to measure one's own mental health state easily and accurately so that it may be regularly monitored and acted upon when needed. The tool is composed of the Generalized Anxiety Disorder-7 (GAD-7), which assess anxiety symptoms, and the Patient Health Questionnaire (PHQ; PHQ-9), which assesses depression symptoms and self-harm ideation; these tools are regarded to be well-clinically validated. To be as brief as possible, the tool starts with the ultrabrief instrument PHQ-4 (consisting of the first 2 items of GAD-7 and PHQ-9 each) and only continues onto the rest of the instruments if the user's responses indicate more than minimal to mild symptoms, according to the instruments' defined thresholds. The tool ultimately triages the user into 1 of 4 wellness protocols: well, mild, moderate, and in crisis, based on symptom severity, which provides the user with an actionable insight and allows the platform to personalize resource recommendation. The self-assessment tool is described in detail in Multimedia Appendix 1. In Figure 3, we show a screenshot of a question in the self-assessment tool. In Figure 4, we show an example user journey on mindline.sg involving the recommendation of resources according to assigned wellness protocol by the self-assessment tool.

XSL•FO
RenderX

Figure 3. User journey on mindline.sg. The self-assessment wellness protocol determines the resources recommended to the user, including exercises with the artificial intelligence chatbot (for well and mild protocols) and direction to support services (for moderate and severe protocols). AI: artificial intelligence.

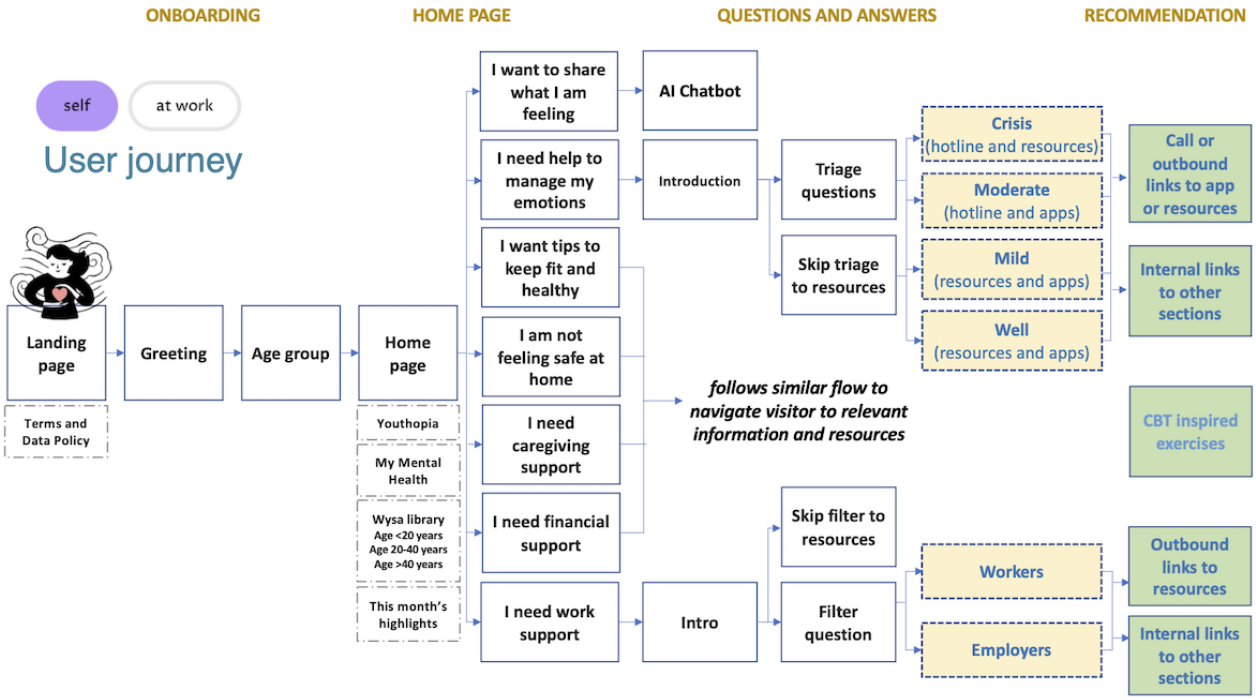

Another important source of training offered through the site are digital therapeutic exercises delivered by an AI chatbot from Wysa [24]. This emotionally intelligent chatbot is embedded into the platform and deploys a suite of dialogue-based interactive exercises inspired by cognitive behavioral therapy. These exercises teach skills such as mindfulness, managing stressors, meditation, reframing thoughts, and sleep techniques. Moreover, the chatbot can engage in free-form conversations, providing a pocket therapist for the user to share their emotions in a safe and anonymous manner, anytime and anywhere. The efficacy, safety, and impact of the Wysa chatbot have been evaluated in multiple studies [25]. A customized version of the Wysa web app was used, though could be comparable to Android app versions 0.7.2.8 (September 8, 2020) through 3.0.4 (May 30, 2022), available throughout the duration under analysis. A screenshot of a dialogue session with the Wysa AI chatbot is shown in Figure 4.

XSL•FO
RenderX

Figure 4. Interventions designed to deepen engagement. (A) A view of the self-assessment tool. (B) The landing page for the mindline at work product customized for working adults. (C) An interactive dialogue-based exercise inspired by cognitive behavioral therapy with the Wysa artificial intelligence chatbot.

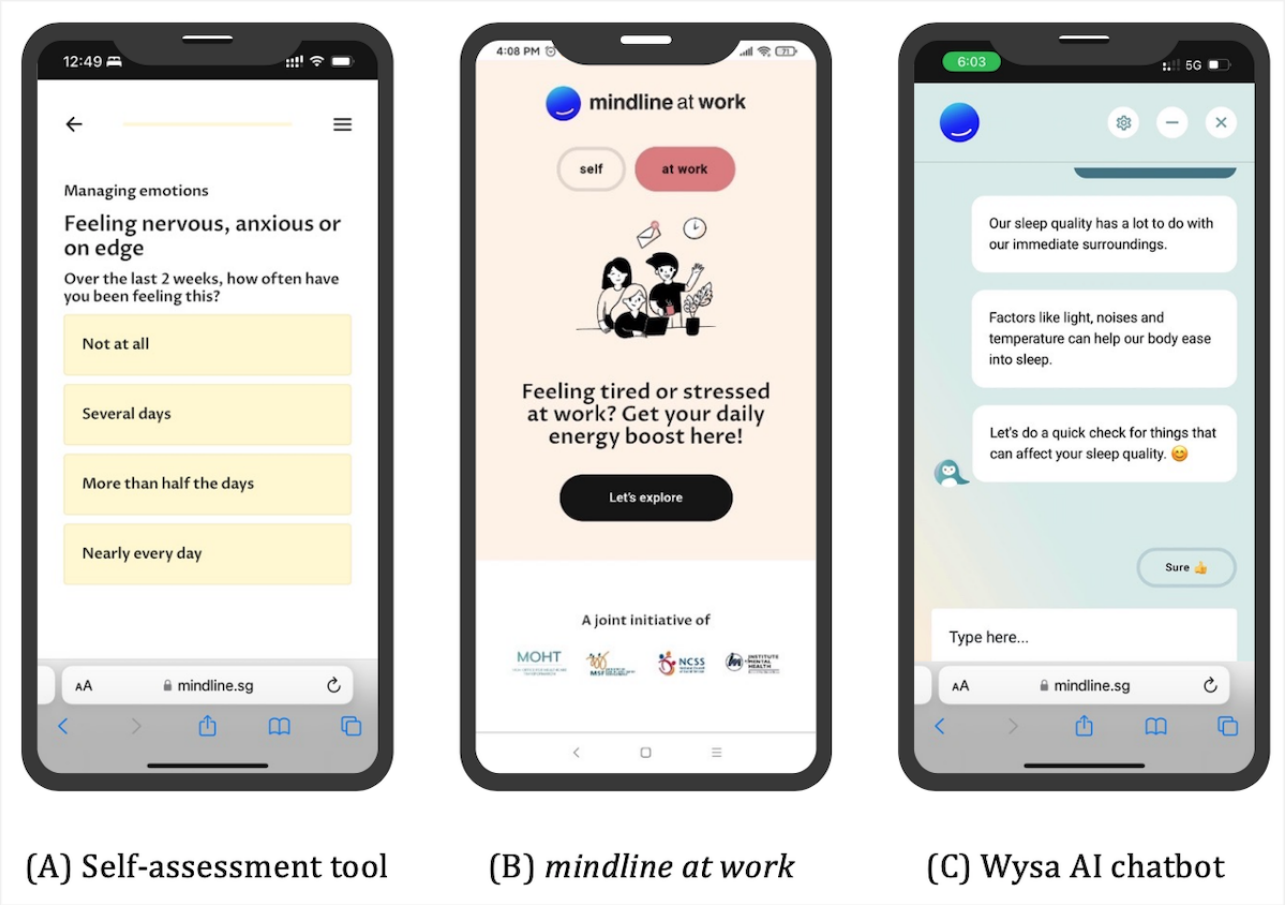

Directing Users With Moderate Symptom Severity or Who Are in a Crisis to Professional and Emergency Care

The designers of the platform are acutely aware that digitally delivered services have limits and appropriate referral to crisis services must be offered when indicated. Indeed, a key component of mindline.sg was the identification, either through the self-assessment tool or the chatbot, of users who are at risk or in a crisis and directing them to human-based hotlines and emergency services. We take care to clearly publicize these limitations in the following ways: first, at the footer of all mindline.sg platform pages, users are advised to contact emergency services if they are at risk of immediate harm. Second, when the navigation (aka hamburger) menu is opened from any page, a Get Help Now link is immediately visible, which directs individuals to emergency services (including hotlines) and provides information on how to help others in crisis. Third, if a user is self-assessed into the moderate protocol, they are directed to counseling centers and hotlines, with options to launch calls directly from the site. Fourth, if a user is self-assessed into the in-crisis protocol, they are directed to emergency hotlines and a suicide prevention help page. Fifth, if the AI chatbot detects a user in an emergency (such as one expressing self-harm ideation), they are immediately directed to local emergency services.

Ensuring Customizability of the Platform: Population Segmentation and mindline at work

mindline.sg was created during the COVID-19 pandemic to address the acute increase in stress and anxiety levels, which were particularly borne by working adults because of job insecurities, new working arrangements, and the lack of usual supporting services (eg, childcare or older adults care). A version of the platform tailored to this subpopulation was therefore added in July 2021 (approximately 1 year following launch), branded as mindline at work. Key differentiating features of the product include mood check-ins (which is an ecological momentary assessment) and user journeys developed around common workplace stressors for both employees (eg, on navigating a toxic work environment) and employers (eg, on nurturing positive workplace cultures). The landing page of the mindline at work product is shown in Figure 4.

This completes the description of the components of the development framework. In Table 1, we provide a summary of the principle mental wellness interventions deployed on the site, the mental wellness goals they intend to serve, and the evidence supporting their use.

XSL•FO
RenderX

Table 1. Summary of specific evidence-based interventions deployed on the site along with details on the goals they attempt to serve.

Mental wellness intervention	Goal	Evidence
Curated local mental health resources	Increase mental health literacy and knowledge of stressors and coping strategies	• Content from authoritative (evidence-backed) sources curated with help from the clinical review panel and advisory and editorial board
Self-assessment tool (comprising the PHQ-9[a] and GAD-7[b] instruments)	Enable users to self-monitor and take action when needed	• PHQ-9 and GAD-7 among the most well-validated mental health screening tools
Digital therapeutic exercises delivered through the Wysa chatbot	Provide evidence-based tools to practice self-help skills	• Smartphones and internet-based interventions are accessible and effective [9-13] • Wysa chatbot is well-validated [25]
Direction to professional services	Ensure users know where to seek help when needed	• Site not designed to serve the needs of users with moderate to severe mental health symptoms • Digital apps are more effective when used in conjunction with a clinician [26]

[a]PHQ-9: Patient Health Questionnaire-9.

[b]GAD-7: Generalized Anxiety Disorder-7.

Analytical Methods

While the anonymity features of the site facilitated users' trust in the platform, they severely limit evaluation efforts to what can be determined from event logs associated with anonymous user IDs. In particular, all usage is captured in a database that records every event, that is, interaction with the site, by the user. These data are further supplemented by Urchin Tracking Module (aka UTM) tags (which are special codes segmenting site traffic) that record the origin of an incoming link (eg, from a targeted advertisement or QR code) and by Google Analytics, which collects inbound activity from other sites. The activity from crawlers (a common method of operating search engines or searching for web resources) is not included in our data.

Unique anonymous user IDs are interpreted as *unique users*, which has the limitation that a single user could be accessing the site from multiple devices or a new anonymous user ID could be assigned when a user clears their browser cache or switches to a different browser. A *bounceduser* visits the site and performs at most 1 page load. Because the platform serves as content directory rather than repository (users are merely directed to external content, including when engaging with the chatbot), nonbounced users are the most accurate metric available for engagement measurement. An *engageduser* is defined as a nonbounced user that spends ≥40 seconds on the site, which indicates a higher level of engagement most likely involving an exploration of multiple resources. A threshold of 40 seconds was chosen during initial site design as an estimate by its designers on how long a user interested in ≥1 resources would likely spend navigating the site or completing an exercise. Note that this is likely to be an underestimate of engagement again because mindline.sg does not host its own content; users could therefore be redirected from the site but engage for far longer than 40 seconds on the linked content. Recent research also estimates the average attention time spent on a digital device screen to be 47 seconds, which is a comparable number [27]. A *return user* is identified by a returning anonymous user ID on another day after the first visit, which again could be inaccurate if users return from a different device or clear their web browser cache. A *session* is defined as contiguous segments of clicks containing pauses (periods of no web activities) of <30 minutes.

We only count users who complete the self-assessment tool and do not exit the assessment early. We can also accurately count users who use the AI chatbot by only counting those who successfully handover to the Wysa platform and engage in a chat.

Statistical Analyses

We compute conversion rates of users that are nonbounced, that are engaged, and that return and we test whether the distribution of those who engage and return differ significantly between years and product using the Fisher exact test.

Ethical Considerations

Ethics review was not required because this research involved analysis of data from an anonymous, nonclinical, and global population (though most users are assumed to have been in Singapore). All users of the mindline.sg platform accept the terms of use and data protection policy (copies of which are provided in Multimedia Appendix 2), which states that statistical data on use is collected and may be used for "research and service enhancement." These policies highlight that these statistical data are anonymous, that the site stores no personally identifiable information, and that the data are transmitted and stored according to industry standards. No individuals were compensated.

Results

Site Use and Growth

The mindline.sg website was launched in June 2020, and mindline at work was launched a year later in June 2021. Visitor and engagement metrics across the platform and for the individual mindline.sg and mindline at work products during these years are summarized as user funnels, depicted in Figure 5.

XSL•FO
RenderX

Figure 5. Engagement funnels for mindline.sg and mindline at work during the first year of each product and across the platforms in the 2 years following launch.

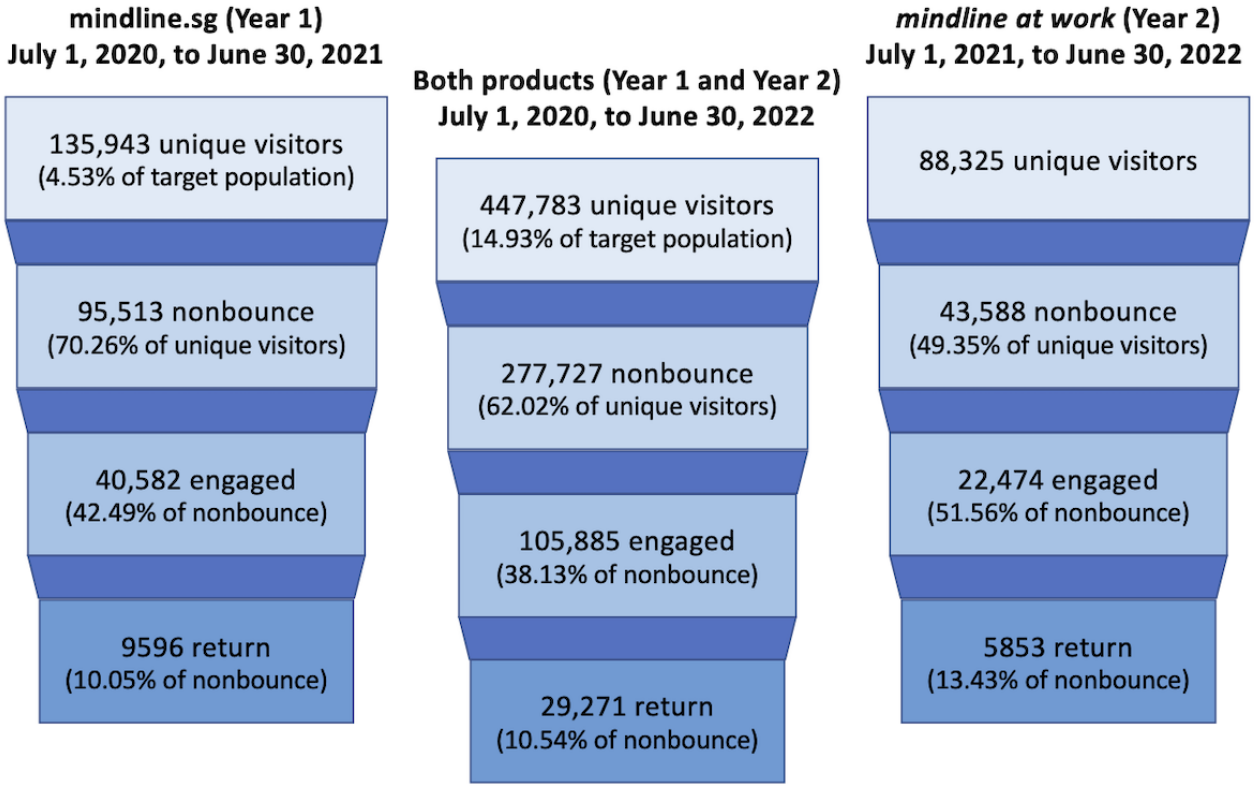

The following year-on-year growth metrics describe the growth of the platform use from the period July 1, 2020, to June 30, 2021, to the period July 1, 2021, to June 30, 2022, across both sites:

- There was a 126% growth year-on-year in total unique visitors (from 135,943 to 447,783).
- There was a 39.7% growth year-on-year in nonbounced users (from 95,513 to 277,727).
- There was a 59% growth year-on-year in engaged users (from 40,582 to 105,885).
- There was a 110% growth year-on-year in return users (from 9596 to 29,271).

The platform's 126% year-on-year growth rate is high. For reference, a marketing firm reported that nonprofit and health care websites studied achieved year-on-year growth rates of 70% and 113%, respectively [28].

The median duration spent on either the mindline.sg or mindline at work sites by nonbounced users during the first 2 years was approximately 137 seconds. As noted in the Analytical Methods section, this is not a good measure of engagement, as the time spent off the site on externally hosted resources is not known. This duration is merely a measure of time spent navigating around different resources on the site.

Development of the Analytics Dashboard

During implementation, an analytics dashboard for click-data analytics was implemented using Metabase [29], a commercial and open-source platform for data visualization and business intelligence. One such dashboard is shown in Figure 6. These dashboards were essential in driving digital marketing strategy and providing customized analytics to strategic partners. The underlying datastore was Titanium Database [30], an open-source hybrid transactional and analytical processing database, running on a 10-node Kubernetes cluster on the cloud.

XSL•FO
RenderX

Figure 6. Customizable analytics dashboards used to continually drive process evaluation, digital marketing strategy, and site redesign.

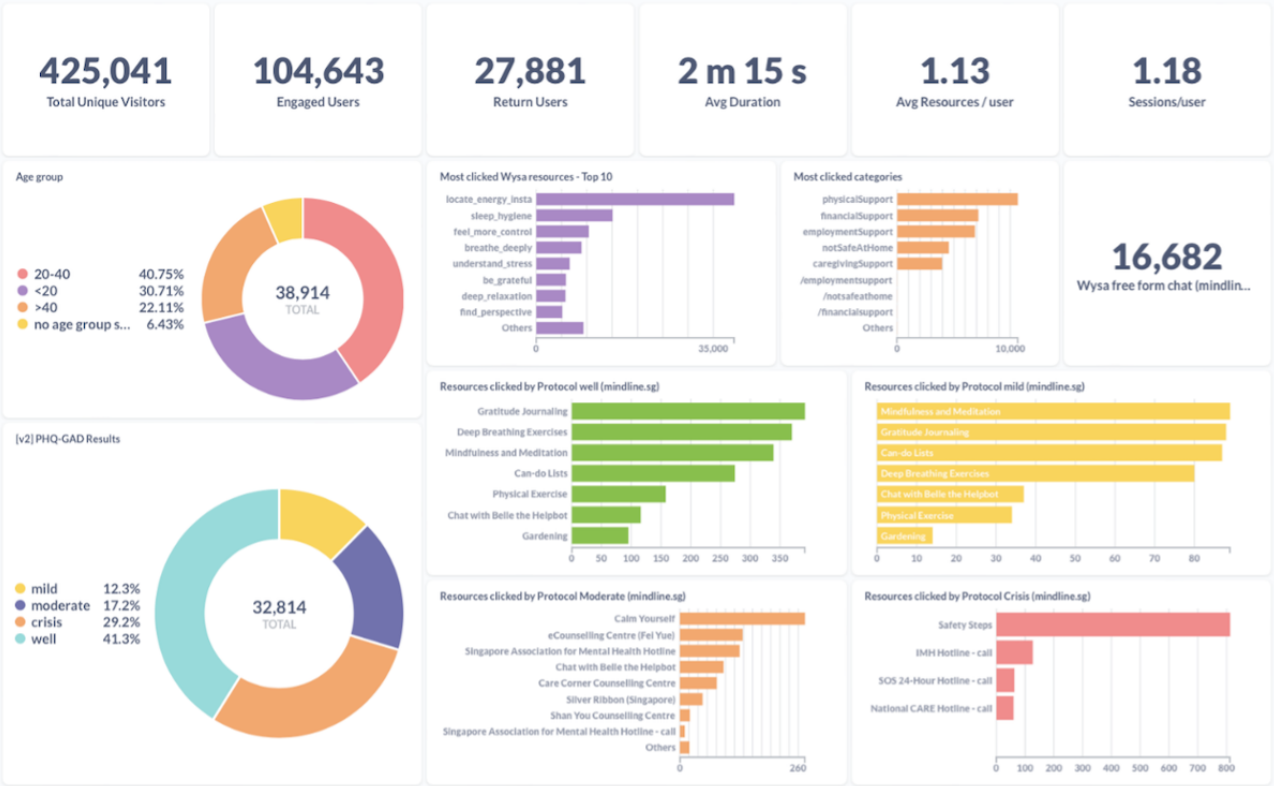

Digital Marketing

Data analytics showed a strong relationship between publicity efforts (ie, newspaper advertisement, radio advertisement, or Facebook advertisement) and visitors to mindline.sg, as seen in Figure 7. Digital marketing, that is, advertising to potential users through digital channels ranging from search platforms to social media platforms, effectively reaches out to population groups who are more likely to use and explore a site such as mindline.sg and to use products such as the AI chatbot. The Analytics Dashboard was analyzed daily to dynamically drive the placement of advertisement campaigns.

Figure 7. Number of weekly users to mindline.sg before and after digital marketing.

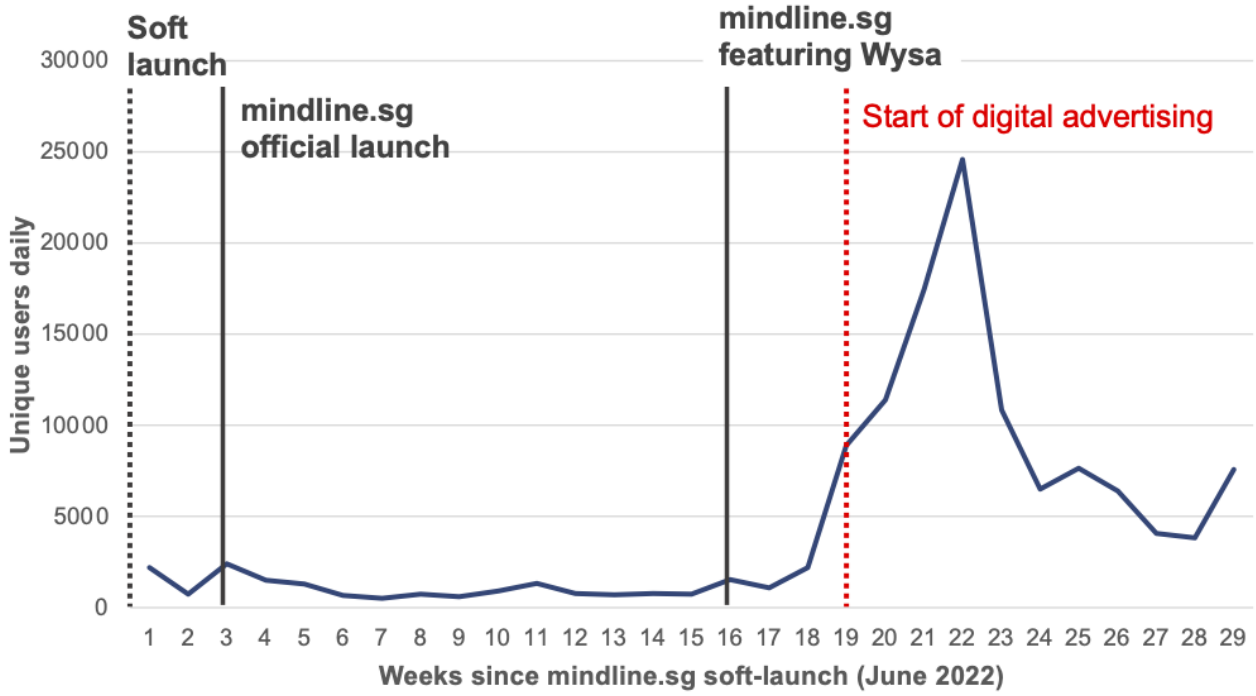

XSL•FO
RenderX

Digital marketing campaigns focused on the placement of advertisements onto social media platforms, including Facebook, Instagram, Google, Reddit, and YouTube. Advertisements usually took the form of static graphics, though occasional video advertisements were used. Advertisements were placed using a combination of placements through advertising platforms, including Taboola and Outbrain, and custom advertisement placements. These advertisements often targeted a common stressor, for example, with invitations to users to "improve sleep" and "locate energy," and often linked directly into a resource on the site such as a Wysa therapeutic exercise. Micro influencers with strategically targeted audiences were also engaged to promote the site and advertisement content.

Use of the AI Chatbot and Therapeutic Exercises

The Wysa AI chatbot was added to the site in October 2020. Through June 30, 2022, a total of 67,626 unique users (which is 25.54% of the 264,758 nonbounced visitors to the platform over this same period) engaged with the chatbot through a cognitive behavioral therapy–inspired exercise or a free-form chat. As a measure of meaningful engagement, around 75.6% (51,125/67,626) of these users exchanged >5 messages with the chatbot. A completion rate of 74.5% (50,381/67,626) was achieved for all therapeutic exercises initiated.

By comparison, 11.69% (32,469/277,727) and 9.52% (26,434/277,727) of the nonbounced visitors completed a self-assessment exercise or viewed some other resource (such as an article or video) through the first 2 years on the site, respectively. The therapeutic exercises delivered by the AI chatbot are therefore the most popular feature of the site. The most popular exercises were energy maintenance, feeling in control, understanding stress, sleep hygiene, breathing, and relaxation.

Self-Assessment Use and Behaviors by Wellness Protocols

Taking the self-assessment questionnaire has consistently proven to be one of the most popular activities on the site. During the 2-year period from July 1, 2020, through June 30, 2022, a total of 41,366 self-assessments have been completed from among 32,469 unique individuals (11.69% of the 277,727 nonbounced visitors over this period). This makes it the second most popular feature on the site after the suite of exercises offered by the chatbot.

Among the set of final self-assessments completed across the users, 44.82% (13,403/32,469) were assigned to the well protocol, 12.28% (3986/32,469) to the mild protocol, 17.26% (5604/32,469) to the moderate protocol, and 29.18% (9476/32,469) to the crisis protocol.

Of the users assigned to the well and mild protocols, 13.01% (1744/13,403) and 10.76% (429/3,986), respectively, explored the resources recommended to them. The most popular of these were gratitude journaling, mindfulness, meditation, and deep breathing exercises. Users assigned to the moderate protocol were slightly more engaged, with 14.72% (825/5604) exploring the resources recommended to them. These included 52.5% (433/825) of those who further engaged, navigating to professional support services, including counseling centers and a mental health hotline. Finally, 11.21% (1062/9476) of the users assigned to the crisis protocol engaged with a recommended resource. Moreover, 23.54% (250/1062) of these individuals elected to launch a call immediately to one of the linked emergency care hotlines. However, most of the individuals (812/1062, 76.46%) in this protocol that further engaged elected to instead read an emergency support page for those experiencing suicidal ideation.

Use Behaviors in mindline at work

Feedback from surveys, focus group discussions, and the AEB drove cocreation of the mindline at work product and its curated resources. The visitor and engagement metrics can be seen in Figure 5.

Conversion rates (ie, rates of successful engagement) used to compare engagement between the 2 sites are shown in Table 2. Digital marketing efforts were substantially increased during year 2, resulting in a large influx of new visitors and a subsequently lower nonbounced user rate. As can be seen in Table 2, the engaged and return user rates of 51.56% (22,474/43,588) and 13.43% (5853/43,588) of the nonbounced users, respectively, were notably higher on mindline at work than they were on mindline.sg during its comparable first year period, which were 42.49% (40,582/95,513) and 10.05% (9596/95,513), respectively. By comparison, the mindline.sg site saw 30.9% (42,829/138,626) and 9.97% (13,822/138,626) engaged and return user rates in the second year, respectively. The Fisher exact test (with test statistic H and P value) confirmed that both the increases in engaged and return users on the mindline.sg site from year 1 to year 2 were significant (engaged: H=0.73; P<.001 and return: H=1.14; P<.001). This could be explained by the increased strategic outreach initiatives to target user groups in the second year, which brought users that were more likely to engage with the platform. Similarly, the Fisher exact test confirmed that both engaged and return user rates on the mindline at work platform were significantly higher than those on the mindline.sg platform during that same year (engaged: H=1.96; P<.001 and return: H=1.21; P<.001). We conclude that the mindline at work site achieved better engagement among its target population.

XSL•FO
RenderX

Table 2. User conversion rates in the first and second years of mindline.sg and in the first year of mindline at worka.

Conversion rates	mindline.sg year 1, n (%)	mindline.*sg* year 2, n (%)	mindline at work year 2, n (%)
Nonbounced rate (of all visitors)	95,513 (70.26)	138,626 (62.02)	43,588 (49.35)
Engaged user rate (of nonbounced users)	40,582 (42.49)	42,829 (30.9)	22,474 (51.56)
Return rate (of nonbounced users)	9596 (10.05)	13,822 (9.97)	5853 (13.43)

[a]Year 1 denotes the period July 1, 2020, to June 30, 2021, and year 2 denotes the period July 1, 2021, to June 30, 2022. The nonbounced rates seen in year 2 are lower than those in year 1 due to substantially larger numbers of visitors because of increased digital marketing efforts. We see that the mindline at work site has higher engaged user and return user rates than the mindline.sg site.

Discussion

Limitations of This Study

The site only collects anonymous user IDs to maintain the level of anonymity required. A major limitation of this feature, however, is that no demographic or other user data are collected, substantially limiting any ability to understand user profiles and to evaluate the impact of the platform. These limitations have already been detailed when defining important metrics in the Analytical Methods section.

Another limitation is that some users could be adopting exploratory behaviors on the site out of curiosity, for example, artificially triggering different wellness protocols during the self-assessment. While this might result in some statistical noise, we do view exploration of the site and the range of resources available for different situations to be a desirable and healthy behavior. To mitigate any impact of this effect on our data analyses, we have taken appropriate data processing steps; for example, we discard all repeated self-assessments in a single session.

Trust and Anonymity

To serve the goal of building a site that users trust in the local Singaporean context, both in terms of content reliability and on anonymity associated with use, the site designers curated content and tools that were endorsed by both local experts in the field and the government. While Singaporeans generally have a high level of public trust in the government (as detailed in the Methods section), this may not be so for all communities.

While the promise of anonymity aims to build trust with users, the anonymity requirements restricted use metrics to anonymous user ID–based data. Evaluation and personalization were therefore difficult without detailed tracking of user-level data. Future iterations of the platform should include a framework for anonymous user profiling.

Marketing and Outreach for Awareness

A key learning from implementation is that digital well-being solutions aimed at population-level intervention require substantial and systematic marketing and outreach efforts to achieve desired reach and economies of scale.

From early process evaluation efforts, it was learned that an extensive and customizable, yet still user-friendly, analytics infrastructure was needed to support digital marketing and outreach partnerships, and the analytics dashboard was therefore implemented. Future iterations of the platform must therefore aim to drive user growth with data-driven digital marketing and ecosystem partnerships, which must be supported by such an appropriate analytics infrastructure.

Progressive Engagement

The progressive engagement approach on mindline.sg encourages users to initially explore the site and read articles, followed by deeper engagement practicing self-care skills through tailored products, nudges, and personalization. Therapeutic exercises delivered through the AI chatbot are the most used feature on the site, indicating that digital therapeutics that enable users to practice skillsets adds immense value. However, we cannot confidently rule out that the AI chatbot appeared to be the most used feature as people tried it more out of curiosity than actual need and so did not engage meaningfully with it. We note that AI chatbots are generally targeted toward those who are well or have only mild symptoms.

The self-assessment tool was the second most used feature on the platform, with 11.69% (32,469/277,727) of the nonbounced users completing at least 1 self-assessment. This is comparable to the uptake seen on another digital mental health platform, Beyond Blue (run by the Australian government), which saw approximately 6.5% and 17.4% of the nonbounced users complete the GAD-7 instrument and a custom 10-item anxiety and depression checklist on their site in 2017 (we have computed these estimates based on data made available in an evaluation by Siggins Miller Consultants [31] and from a report by Beyond Blue [32]). We have learned that users like to assess themselves; there could be an implicit interest in benchmarking oneself against others. Future iterations could perhaps reflect individual assessments against aggregate results from the assessments to further nudge those who may be in the moderate-to-crisis protocols to seek help. Because this self-assessment tool is derived from existing, validated instruments and is low cost to deploy and scale digitally, it has the potential to be a cost-effective intervention for the platform.

Customization and Personalization to Increase Engagement

While most users were self-assessed to be well or have only mild symptoms, a substantial proportion were self-assessed to have moderate symptoms or were in crisis, with 29.18% (9476/32,469) falling under the crisis protocol. Although we surmise that some of these users may be triggering the crisis protocols out of curiosity, we have no independent assessment to determine the actual rate, other than numerous press and scholarly reports that have suggested that mental anguish is greater than generally believed. We can conclude that mindline.sg is attractive to people at all levels of symptom

XSL•FO
RenderX

severity, and this is supported in part by the fact that those in the moderate protocol were found to be slightly more committed to help seeking in their follow-up actions after self-assessment. The platform therefore has the potential to further drive help-seeking behavior in users with moderate to severe symptoms and should link such users to human-based support within the platform itself (eg, community forums and therapists) as well as outside the platform (eg, telemedicine and clinical consults).

On the basis of our customization experience for a more specific population segment with mindline at work, we found that the tailored product for working adults deepened the level of user engagement. The hypothesis that customization could help create more targeted, meaningful impact to the mental well-being of specific populations could be tested further with other populations that have been identified to have greater mental health needs, especially since the COVID-19 pandemic, such as youth and health care workers.

Scaling Digital Mental Health Initiatives and Summary of Recommendations

An estimated 12 full-time employment staff were involved in the development and maintenance of the initiative over the first 2 years.

A summary of our design principles, key features, learning points, and key recommendations from the initial implementation of mindline.sg can be found in Table 3.

Table 3. Summary of the design principles, key features, learning points, and key recommendations for mindline.sg.

Design principles and key features	Learning points	Key recommendations
Design 1 (trust): built as a trusted source for high-quality mental health information and resources through cocreation methodologies		
• Participatory approach with feedback from service providers and intended users • Expert input and oversight through an advisory and expert board • Government endorsed	• Cocreative methodologies, along with endorsement from multiple authoritative entities, build trust and facilitate acceptance of the digital intervention by a diverse group of people, which would be unlikely if purely government endorsed	• Maintain participatory approach to keep enhancing the anonymous user experience • Maintain credibility of content that is overseen by expert advisors • Obtain endorsement by multitude of authoritative figures or organizations with high public trust.
Design 2 (anonymity): maintaining anonymity to encourage convenient help seeking		
• No collection of personally identifiable data • Developed as a web app rather than a mobile app	• The promise of anonymity contributed to building trust • A web app proved to be easy to use • Anonymity severely limits evaluation	• Develop a framework for anonymous user profiling • Remain as a web app to maintain convenience and anonymity
Design 3 (progressive engagement): progressive engagement to facilitate users' different stages of learning and readiness for change		
• A robust branding and marketing strategy to raise awareness about the existence of mindline.sg	• Substantial digital marketing and strategic outreach partnerships are required to achieve desired reach and economies of scale.	• Drive user growth with data-driven digital marketing and ecosystem partnerships, which must be supported by an extensive and customizable analytics infrastructure.

• Tools to improving mental health literacy
  • Links to curated, expert-endorsed articles and videos for self-learning from local and international sources
  • Self-assessment tool with personalized resource recommendations through expert-developed algorithms
  • Links to local resources and helplines
  • AI[a] chatbot to guide therapeutic exercises
	• The substantial proportion of resource use attributable to the AI chatbot indicates that therapeutic content on practicing skillsets adds immense value
	• Users like to assess themselves; there could be an implicit interest in benchmarking oneself against others
	• mindline.sg was attractive to people with varying levels of symptom severity, including those with moderate to severe needs.
		• Existing features should be retained for their usefulness to users with different needs, though some expansion of features should be considered, for example:
		  • Reflect individual assessments against aggregate results for contextualization.
		  • Provide deeper links to human-based services from within the platform for those who need further support, for example, linkage to therapy, telemedicine, and community forums

Design 4 (customizability): ensuring customizability of the platform for a targeted approach to improving mental health outcomes		
• Adapting features on the platform that can be tailored to different population segments, for example, mindline at work for working adults	• The customized product for working adults deepened the level of engagement	• Explore customizations for a variety of population segments with mental health needs to deepen impact with a targeted approach, for example, with youth and health care workers

[a]AI: artificial intelligence.

XSL•FO
RenderX

Relevance of the Findings

The theory of change for the platform (described in the Introduction section) requires reach and engagement to be achieved so that users are aware of the resources available on mindline.sg and they gain experience practicing self-care skills. This paper provides evidence that such necessary processes have been accomplished. This supports the subsequent study on whether the desired outcomes (that users practice self-care skills as a habit, access resources when in need, and self-refer to professional mental health services if required, among others) are being achieved. We leave this work to ongoing evaluation studies.

Conclusions

The mindline.sg digital mental wellness initiative aims to empower all residents in Singapore to maintain their own state of mental wellness, to be able to support others around them, and to self-refer to professional mental health services when needed. This study provides evidence that the initiative has successfully achieved substantial reach and engagement in the 2 years since its launch. Achieving these necessary outcomes required substantial and sustained investment into digital marketing as well as multiple strategic outreach partnerships. The development of an extensive and customizable data analytics infrastructure was required to support these efforts. A self-assessment tool designed for self-monitoring and content personalization, as well as a suite of digital therapeutic exercises delivered by the Wysa AI chatbot, successfully engaged users and were the 2 most popular features of the site. The creation of a customized version of the site for working adults, mindline at work, also achieved higher engaged in its first year compared to the generic mindline.sg site. We have provided recommendations for revision, and we believe this report may help to guide the development of similar platforms around the world for communities with similar needs. With this successful process evaluation, the initiative should now proceed to conduct impact evaluation studies around its desired outcomes.

Acknowledgments

Funding to create the mindline.sg platform and to support the writing of this descriptive paper was provided by the MOH Office for Healthcare Transformation (Singapore).

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to the mindline.sg data protection policy but are available from the corresponding author on reasonable request.

Authors' Contributions

JHW, WCL, and RJTM contributed to conceptualization, formal analysis, funding acquisition, project administration, resources, supervision, and reviewing and editing the draft. YH contributed to project administration, resources, software, and supervision. CH contributed to data curation, formal analysis, methodology, resources, supervision, validation, visualization, preparing the original draft, and reviewing and editing the final draft. CT contributed to conceptualization, formal analysis, and resources. JKC and YSP contributed to data curation, formal analysis, software, validation, and visualization. WMT contributed to conceptualization and resources.

Conflicts of Interest

All authors were employees of the MOH Office for Healthcare Transformation (Singapore), which created the mindline.sg platform, at the time of writing.

Multimedia Appendix 1
Description of the self-assessment and triaging tool.
[DOCX File , 159 KB - jmir_v26i1e44443_app1.docx]

Multimedia Appendix 2
Terms of use and data protection policy for the mindline.sg platform.
[DOCX File , 33 KB - jmir_v26i1e44443_app2.docx]

References

1. World mental health report: transforming mental health for all. World Health Organization. URL: https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report [accessed 2024-05-24]
2. Some 300 million people suffer from depression, UN warns ahead of World Health Day. United Nations News. 2017 Mar 31. URL: https://news.un.org/en/story/2017/03/554462 [accessed 2024-05-24]

XSL•FO
RenderX

3. Subramaniam M, Abdin E, Vaingankar JA, Shafie S, Chua BY, Sambasivam R, et al. Tracking the mental health of a nation: prevalence and correlates of mental disorders in the second Singapore mental health study. Epidemiol Psychiatr Sci 2019 Apr 05;29:e29 [FREE Full text] [doi: 10.1017/S2045796019000179] [Medline: 30947763]
4. Subramaniam M, Abdin E, Vaingankar JA, Shafie S, Chua HC, Tan WM, et al. Minding the treatment gap: results of the Singapore Mental Health Study. Soc Psychiatry Psychiatr Epidemiol 2020 Nov;55(11):1415-1424 [FREE Full text] [doi: 10.1007/s00127-019-01748-0] [Medline: 31317246]
5. Maddock A, Blair C, Ean N, Best P. Psychological and social interventions for mental health issues and disorders in Southeast Asia: a systematic review. Int J Ment Health Syst 2021 Jun 05;15(1):56 [FREE Full text] [doi: 10.1186/s13033-021-00482-y] [Medline: 34090491]
6. Sharan P, Sagar R, Kumar S. Mental health policies in South-East Asia and the public health role of screening instruments for depression. WHO South East Asia J Public Health 2017 Apr;6(1):5-11. [doi: 10.4103/2224-3151.206165] [Medline: 28597852]
7. Psychiatrists working in mental health sector (per 100,000). World Health Organization. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/psychiatrists-working-in-mental-health-sector-(per-100-000) [accessed 2024-05-24]
8. Number of psychiatrists: how do countries compare? Eurostat. 2020 May 6. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20200506-1 [accessed 2024-05-24]
9. Hollis C, Sampson S, Simons L, Davies EB, Churchill R, Betton V, et al. Identifying research priorities for digital technology in mental health care: results of the James Lind Alliance Priority Setting Partnership. Lancet Psychiatry 2018 Oct;5(10):845-854 [FREE Full text] [doi: 10.1016/S2215-0366(18)30296-7] [Medline: 30170964]
10. Andersson G, Carlbring P, Titov N, Lindefors N. Internet interventions for adults with anxiety and mood disorders: a narrative umbrella review of recent meta-analyses. Can J Psychiatry 2019 Jul 16;64(7):465-470 [FREE Full text] [doi: 10.1177/0706743719839381] [Medline: 31096757]
11. Fu Z, Burger H, Arjadi R, Bockting CL. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. The Lancet Psychiatry 2020 Oct;7(10):851-864. [doi: 10.1016/s2215-0366(20)30256-x]
12. Heckendorf H, Lehr D, Boß L. Effectiveness of an internet-based self-help intervention versus public mental health advice to reduce worry during the COVID-19 pandemic: a pragmatic, parallel-group, randomized controlled trial. Psychother Psychosom 2022 Jan 20;91(6):398-410 [FREE Full text] [doi: 10.1159/000521302] [Medline: 35051939]
13. Mühlmann C, Madsen T, Hjorthøj C, Forman JL, Kerkhof AJ, Nordentoft M, et al. Effectiveness of an internet-based self-help therapy program for suicidal ideation with follow-up at 6 months: results of a randomized controlled trial. J Clin Psychiatry 2021 Aug 31;82(5). [doi: 10.4088/jcp.20m13803]
14. Sorkin DH, Janio EA, Eikey EV, Schneider M, Davis K, Schueller SM, et al. Rise in use of digital mental health tools and technologies in the United States during the COVID-19 pandemic: survey study. J Med Internet Res 2021 Apr 16;23(4):e26994 [FREE Full text] [doi: 10.2196/26994] [Medline: 33822737]
15. Kit TS. Pandemic blues drive demand for mental health apps in Singapore. Channel News Asia. 2021 Oct 7. URL: https://www.channelnewsasia.com/singapore/covid-19-mental-health-startups-singapore-mindfi-intellect-2203336 [accessed 2024-05-24]
16. Clay RA. Mental health apps are gaining traction. American Psychological Association. 2021 Jan 1. URL: https://www.apa.org/monitor/2021/01/trends-mental-health-apps [accessed 2024-05-24]
17. State of wellness H1'20 report: investment and sector trends to watch. CB Insights. 2020 Aug 25. URL: https://www.cbinsights.com/research/report/wellness-trends-h1-2020/#:~:text=Report%20Highlights%3A&text=Mental%20health%20startups%20saw%20over,taking%20place%20in%20H1'20 [accessed 2024-05-24]
18. Mental health. World Health Organization. 2022 Jun 17. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response [accessed 2024-05-24]
19. The app evaluation model. American Psychiatric Association. URL: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/the-app-evaluation-model [accessed 2024-05-24]
20. Attitudes towards institutions, politics, and policies: key findings from the world value survey. Institute of Policy Studies, Lee Kuan Yew School of Public Policy, National University of Singapore. 2021 Mar 23. URL: https://lkyspp.nus.edu.sg/ips/news/details/attitudes-towards-institutions-politics-and-policies-key-findings-from-the-world-value-survey [accessed 2024-05-24]
21. Pang S, Liu J, Mahesh M, Chua BY, Shahwan S, Lee SP, et al. Stigma among Singaporean youth: a cross-sectional study on adolescent attitudes towards serious mental illness and social tolerance in a multiethnic population. BMJ Open 2017 Oct 16;7(10):e016432 [FREE Full text] [doi: 10.1136/bmjopen-2017-016432] [Medline: 29042379]
22. Subramaniam M, Abdin E, Picco L, Pang S, Shafie S, Vaingankar JA, et al. Stigma towards people with mental disorders and its components - a perspective from multi-ethnic Singapore. Epidemiol Psychiatr Sci 2017 Aug 28;26(4):371-382 [FREE Full text] [doi: 10.1017/S2045796016000159] [Medline: 27018715]

XSL•FO
RenderX

23. Yuan Q, Abdin E, Picco L, Vaingankar JA, Shahwan S, Jeyagurunathan A, et al. Attitudes to mental illness and its demographic correlates among general population in Singapore. PLoS One 2016 Nov 28;11(11):e0167297 [FREE Full text] [doi: 10.1371/journal.pone.0167297] [Medline: 27893796]
24. Wysa homepage. Wysa. URL: https://www.wysa.com/ [accessed 2024-05-24]
25. Clinical evidence - Wysa. Wysa. URL: https://www.wysa.io/clinical-validation [accessed 2024-05-24]
26. Lagan S, Aquino P, Emerson MR, Fortuna K, Walker R, Torous J. Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ Digit Med 2020;3:100 [FREE Full text] [doi: 10.1038/s41746-020-00312-4] [Medline: 32821855]
27. Mark G. Attention Span: A Groundbreaking Way to Restore Balance, Happiness and Productivity. New York, NY: Hanover Square Press; 2023.
28. Website growth: annual visitor increases and what that means for your business. Kerigan Marketing Associates. 2020 Oct 13. URL: https://keriganmarketing.com/website-growth-annual-visitor-increases-what-that-means-for-your-business/ [accessed 2024-05-24]
29. Metabase. URL: https://www.metabase.com/ [accessed 2024-05-24]
30. PingCAP. URL: https://www.pingcap.com/ [accessed 2024-05-24]
31. Beyond blue evaluation 2015–2018. Siggins Miller. URL: https://cdn.intelligencebank.com/au/share/zKG/7MpVv/LVXLy/original/independent-evaluation_final-report-for-web_sept-2019 [accessed 2024-05-24]
32. Annual highlights 2017–18. Beyond Blue. URL: https://edge.sitecorecloud.io/beyondblue1-beyondblueltd-p69c-fe1e/media/Project/Sites/beyondblue/PDF/Annual-Highlights/beyond-blue-annual-highlights-1017-18.pdf [accessed 2024-05-24]

Abbreviations

AEB: advisory and editorial board
CRP: clinical review panel
GAD-7: Generalized Anxiety Disorder-7
PHQ: Patient Health Questionnaire

Edited by T Leung; submitted 20.11.22; peer-reviewed by J Rose, M Brown, A Salamanca-Sanabria; comments to author 03.03.23; revised version received 12.04.23; accepted 31.03.24; published 04.06.24.

Please cite as:
Weng JH, Hu Y, Heaukulani C, Tan C, Chang JK, Phang YS, Rajendram P, Tan WM, Loke WC, Morris RJT
Mental Wellness Self-Care in Singapore With mindline.sg: A Tutorial on the Development of a Digital Mental Health Platform for Behavior Change
J Med Internet Res 2024;26:e44443
URL: https://www.jmir.org/2024/1/e44443
doi:10.2196/44443
PMID:38833294

©Janice Huiqin Weng, Yanyan Hu, Creighton Heaukulani, Clarence Tan, Julian Kuiyu Chang, Ye Sheng Phang, Priyanka Rajendram, Weng Mooi Tan, Wai Chiong Loke, Robert J T Morris. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Transforming Digital Phenotyping Raw Data Into Actionable Biomarkers, Quality Metrics, and Data Visualizations Using Cortex Software Package: Tutorial

James Burns[1], BS; Kelly Chen[1], MS; Matthew Flathers[1], BA; Danielle Currey[1,2], BS; Natalia Macrynikola[1], PhD; Aditya Vaidyam[1,3], MS; Carsten Langholm[1], BS; Ian Barnett[4], PhD; Andrew (Jin Soo) Byun[1], MS; Erlend Lane[1], MPA; John Torous[1], MD

[1]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

[2]Case Western Reserve University School of Medicine,, Cleveland, OH, United States

[3]Carle Illinois College of Medicine, Urbana, IL, United States

[4]Department of Biostatistics, Epidemiology and Informatics Perelman School of Medicine at the University of Pennsylvania, Philadephia, PA, United States

Corresponding Author:
John Torous, MD
Division of Digital Psychiatry
Beth Israel Deaconess Medical Center
Harvard Medical School
330 Brookline Ave
Boston, MA, 02215
United States
Phone: 1 6176676700
Email: jtorous@bidmc.harvard.edu

Abstract

As digital phenotyping, the capture of active and passive data from consumer devices such as smartphones, becomes more common, the need to properly process the data and derive replicable features from it has become paramount. Cortex is an open-source data processing pipeline for digital phenotyping data, optimized for use with the mindLAMP apps, which is used by nearly 100 research teams across the world. Cortex is designed to help teams (1) assess digital phenotyping data quality in real time, (2) derive replicable clinical features from the data, and (3) enable easy-to-share data visualizations. Cortex offers many options to work with digital phenotyping data, although some common approaches are likely of value to all teams using it. This paper highlights the reasoning, code, and example steps necessary to fully work with digital phenotyping data in a streamlined manner. Covering how to work with the data, assess its quality, derive features, and visualize findings, this paper is designed to offer the reader the knowledge and skills to apply toward analyzing any digital phenotyping data set. More specifically, the paper will teach the reader the ins and outs of the Cortex Python package. This includes background information on its interaction with the mindLAMP platform, some basic commands to learn what data can be pulled and how, and more advanced use of the package mixed with basic Python with the goal of creating a correlation matrix. After the tutorial, different use cases of Cortex are discussed, along with limitations. Toward highlighting clinical applications, this paper also provides 3 easy ways to implement examples of Cortex use in real-world settings. By understanding how to work with digital phenotyping data and providing ready-to-deploy code with Cortex, the paper aims to show how the new field of digital phenotyping can be both accessible to all and rigorous in methodology.

(J Med Internet Res 2024;26:e58502) doi:10.2196/58502

KEYWORDS

digital phenotyping; mental health; data visualization; data analysis; smartphones; smartphone; Cortex; open-source; data processing; mindLAMP; app; apps; data set; clinical; real world; methodology; mobile phone

XSL•FO
RenderX

Introduction

The popularity of smartphone-based digital phenotyping for advancing health research has resulted in a plethora of platforms and tools. Many have lauded the potential to capture real-time behavioral data from patients' smartphones in a scalable manner [1,2]. However, recent reviews and commentaries consistently highlight the need for improved methodological rigor, more return of results to participants, and better data sharing [3-5]. Solving these challenges is multifactorial but can be framed around issues of (1) data collection, (2) data processing, and (3) return of results.

The evolving landscape of consumer technology, and thus digital phenotyping apps, presents numerous challenges to their scientific study. These apps require constant efforts to ensure they remain secure and able to collect the data they claim to capture. The result of this dynamic environment is that many digital phenotyping apps are no longer supported, and many past apps are already abandoned. Keeping pace with new updates from Apple and Google, security vulnerabilities or patches, user interface or experience improvements, and new sensors accessible from the phone requires constant effort and financial resources. Furthermore, the shifting tools and libraries used in this research has also created challenges for transparent data processing and return of results. Few digital phenotyping studies have ever been replicated because differences in feature processing procedures make the resulting outputs from these studies incomparable. Of note, in the rare cases in which replication has been attempted, the results have been negative [6,7].

Beyond supporting replicable science, better data collection and data processing can enable better data visualization, better sharing of results with participants, and meaningful use outside of research. A 2022 paper proposed the return of results as a necessary aspect of ethical digital psychiatry research [8], and a 2023 ethical framework was proposed that specially identified the need to return results in digital phenotyping in psychiatry [9]. However, neither proposal identified how this could be accomplished or provided resources to ensure data sharing is possible. The need for data visualizations extends beyond just research. A December 2023 systematic review and meta-analysis on the utility of mental health apps to augment care for psychiatric illnesses found that, while nearly two-thirds of those apps reviewed captured digital phenotyping–related data, less than half of these apps were even able to share data with clinicians [10]. A 2022 review on data visualization in mental health highlighted the numerous benefits of return of results for patients including enabling proactive self-management, more effective communication with clinicians, and better engagement with remote tools like apps [11].

Toward supporting more transparent and ethical research, the need for better data collection, data processing, and return of results is clear. Data processing pipelines offer a productive focus, as they can support data from multiple apps and transform that data to support myriad visualizations. There are several pipelines specific to smartphone digital phenotyping data including Rapids [12], Forest [13], and Cortex [14]. While each of the platforms offers value, they also feature differences, especially around natively supporting data visualizations. Toward the goal of expanding the utility of digital phenotyping methods and enabling the return of results, we review the Cortex data processing pipeline and provide a tutorial to encourage more replicable, ethical, and clinically relevant research. Toward building the use cases of data visualization, this paper (1) reviews digital phenotyping data types, (2) explores preprocessing steps, (3) discusses the basic use of Cortex for data quality assessments, (4) presents advanced Cortex features, and (5) provides several real-world examples of using all the steps together. Figure 1 below provides a visual overview of the paper and details of the topics covered.

The goal of this paper is to facilitate an understanding of the steps, assumptions, and limitations around processing digital phenotyping data and transforming it into data visualizations. The steps outlined here correspond to interactive examples, and we highlight in each step the ability to customize, expand, or alter the procedures to ensure the process can serve a range of outcomes, from simple reporting of survey scores to more complex time series results. To facilitate these examples, we draw from the open-source mindLAMP application, which is used by nearly 100 research teams across the world [15-18].

Figure 1. Paper schematic.

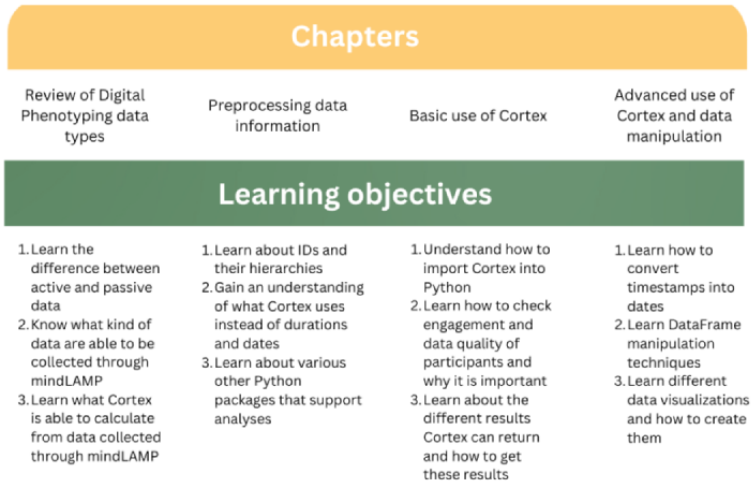

XSL•FO
RenderX

Review of Digital Phenotyping Data Types

Different types of digital phenotyping data require different approaches toward labeling, storing, and processing. In digital phenotyping and with mindLAMP, data are classified as either active or passive. Active data include surveys and cognitive assessments. These data streams are customizable by the researchers or clinician and require the patient or user to actively engage (ie, take a survey) for data to be collected. mindLAMP can be deployed solely as a survey tool, making it possible to use the platform for ecological momentary assessment research alone. All survey and cognitive data are stored within a secure server with no personal identifiable information; only the participant's user ID is attached for identification [19].

By contrast, passive data are collected using the phone's native sensors enabled by the participant and also specified by the researcher. These sensors commonly include an accelerometer and a GPS, both of which are assigned to collect at a frequency of up to 5 and 1 Hz, respectively. The passive data collected through sensors can be customized by a study to use as few or as many sensors as researchers desire, with options ranging from screen time to Bluetooth connections. It is also important to note that these sensor outputs are also referred to as raw data. They are what Cortex uses to calculate more actionable metrics. Table 1 offers a more comprehensive list of options at the time of this writing. Table 1 lists all the sensors LAMP currently has the ability to collect and their name within mindLAMP. Sampling type represents how much data are collected within a certain time frame. If the sampling type is continuous, all data are collected the moment the sensor is activated. If the sampling type is discrete, the data are recorded every *n* seconds, in which *n* is defined by a frequency parameter. If the sampling type is interval, which only relates to step count, the data describe how many steps were taken within *n* minutes, which is described by a frequency parameter.

Unlike active data, passive data are captured at a high volume and velocity that are not conducive to most ecological momentary assessment data analysis methods. It must be processed into actionable metrics through a pipeline such as Cortex. For example, Cortex can process thousands of coordinates from the GPS phone sensor and transform them into a metric that reflects how much time a participant spent at home, appropriately named "hometime." Table 2 offers a more comprehensive list of calculated data features offered by Cortex at the time of this writing. Table 2 lists all the secondary features currently offered by Cortex within the "Secondary feature name" column. It also lists the reference citations that feature links to the documentations, which offer a description and an example (per link) of how to pull values using cortex.secondary.feature_name.feature name(). More information regarding the command can be found in the *Passive Data Processing* section. The final column, "Requires additional parameters for cortex?" has a Boolean (True or False) outcome that represents if Cortex needs any additional parameters beyond start, end, and resolution to run either cortex.secondary.feature_name.feature_name() or cortex.run().

Table 1. Sensor types.

Name	SensorSpec	Sampling type	Requires watch or other devices
Analytics	lamp.analytics	Continuous	No
Location	lamp.gps	Discrete	No
Accelerometer	lamp.accelerometer	Discrete	No
Device Motion	lamp.device_motion	Discrete	No
Screen	lamp.device_state	Continuous	No
Pedometer	lamp.steps	Interval	No
Bluetooth & WiFi	lamp.nearby_device	Discrete	No
Calls & Texts	lamp.telephony	Continuous	No
Sleep	lamp.sleep	Continuous	Yes
Workouts	lamp.segment	Continuous	Yes
Activity Recognition	lamp.activity_recognition	Continuous	Yes
Nutrition	lamp.nutrition	Continuous	Yes
Blood Glucose	lamp.blood_glucose	Continuous	Yes
Oxygen Saturation	lamp.oxygen_saturation	Continuous	Yes
Body Temperature	lamp.body_temperature	Continuous	Yes
Blood Pressure	lamp.blood_pressure	Continuous	Yes
Heart Rate	lamp.heart_rate	Continuous	Yes
Heart Rate Variability	lamp.heartratevariability_sdnn	Continuous	Yes
Respiratory Rate	lamp.respiratory_rate	Continuous	Yes

XSL•FO
RenderX

Table 2. Secondary features.

Secondary feature name	References featuring links to documentation	Requires additional parameters for Cortex?
Call Duration	[20]	Yes
Call Number	[21]	Yes
Data Quality	[22]	Yes
Entropy	[23]	No
Cognitive Assessment Results	[24]	Yes
HealthKit Sleep Duration	[25]	Yes
Hometime	[26]	No
Inactive Duration	[27]	No
Nearby Device Count	[28]	No
Screen Duration	[29]	No
Step Count	[30]	Yes
Survey Results	[31]	Yes
Trip Distance	[32]	No
Trip Duration	[33]	No
Call Degree	[34]	No

Pre–Data Processing Information

Overview

The design and function of Cortex are motivated by the volume of data, particularly passive data, that a single user can generate in 2 weeks. Assuming a user captures GPS and accelerometer data from the smartphone, each at a rate of 5 Hz, there will be just >1.2 million data points gathered within 14 days. While lower sampling frequencies are possible, the ability to capture additional sensors and data over months or even years illustrates the size of the data set in question. Cortex is designed to work flexibly with these data. This section reviews its foundations and assumptions to provide key background information to enable easy and effective use of Cortex for broad applications including data processing and data visualization.

IDs and Their Hierarchy

Working with the data requires being able to identify users and studies. For every participant that downloads the mindLAMP app, a researcher needs to create a user profile for them in the mindLAMP dashboard. Upon creation of a user profile, mindLAMP will automatically generate a unique ID specific to the patient and the investigator account (researcher ID). The user ID is given in the format of a U followed by 10 digits (eg, U0123456789). These IDs are not to be confused with the researcher ID, which is provided in a 20-character string format of numbers and letters (eg, 123456abcdef78g912hi). Within the investigator mindLAMP account, multiple studies (groups) can also exist. The studies themselves also have a unique ID, following the same format as the researcher ID.

The utility of having multiple studies under a singular researcher ID is to allow for multiple versions (eg, A/B testing, all sensors active or no sensors active, feedback shared in the app or no feedback shared in the app) of a larger investigation (Figure 2). This can be useful when assessing the impact of features or when seeking to offer participants multiple ways to partake, especially around how much passive data they wish to share. It can also enable different schedules of assessments or app reminders for participants as set by the research team.

Figure 2. LAMP hierarchy: This figure shows the hierarchy within LAMP. Researchers, also known as investigators, are at the top. Each researcher/investigator can have multiple different studies, and within each of those studies are individual participants.

Epoch Time and Milliseconds

Instead of dates, which can have multiple formats, Cortex uses a modified version of Epoch time. Epoch time is the number of seconds that have passed since January 1, 1970, 12 AM Coordinated Universal Time (UTC) to a specified date and time. The modification Cortex uses is that it uses the number of milliseconds instead of seconds. To provide an example, December 1, 2023, 12 AM is 1,701,450,000,000 in the modified Epoch time. In addition to representing date values in modified Epoch time, all other time-related fields (eg, time duration) will also use this format. For the remainder of the paper, when Epoch time is mentioned, it will be the modified version.

XSL•FO
RenderX

For the sake of convenience to the user, this paper will provide resources to turn dates into Epoch time and helpful variables to change milliseconds to more comprehensive units, such as seconds or days [35]. In addition, many different Epoch time converters on the internet can assist with conversions.

Dates can be converted into Epoch time using a Python function combined with various Python packages, such as “pytz” and “datetime” (Figure 3).

This function takes a date in the form of MM/DD/YYYY in a string and converts it into Epoch time. An example of this function in use can be seen in Figure 4.

Cortex has a useful command that gives the epoch time when the command is run (Figure 5). This command will print the Epoch time if it is not saved into a variable.

Finally, it is good practice to keep variable names that allow for easy conversions of milliseconds such as the variables given in Figure 6.

Figure 3. Codebook 1.

```
import pytz
from datetime import datetime, timezone

def timestamp(dt):
    local = pytz.timezone("America/New_York")
    date = datetime.strptime(dt, '%m/%d/%Y')
    local_dt = local.localize(date, is_dst=None)
    utc_dt = local_dt.astimezone(pytz.utc)
    utc_dt.replace(tzinfo=timezone.utc).timestamp() * 1000
    return int(utc_dt.replace(tzinfo=timezone.utc).timestamp() * 1000)
```

Figure 4. Codebook 2.

```
timestamp('01/01/2001')
978325200000
```

Figure 5. Codebook 3.

```
cortex.now()
```

Figure 6. Codebook 4.

```
MS_IN_DAY = 24 * 3600 * 1000
MS_IN_HOUR = 3600 * 1000
```

Imports

Cortex is designed to provide users with meaningful and understandable metrics. When combined with various Python packages, it is even more powerful for data processing. Table 3 shows common Python packages that are often useful to run with Cortex. These are the packages that will be used in the rest of the paper and are useful to explore if planning to run Cortex. There are many more Python packages that can be used, and Table 3 should not be viewed as a limiting factor on what can be done with the data processed by Cortex.

XSL•FO
RenderX

Table 3. Common Python Packages used to help manage and support data manipulation and analysis.

Package name	Package usability	References featuring the documentation link
Pandas	Allows for data manipulation and simple analysis within DataFrames	[36]
Numpy	Allows for scientific computations in Python	[37]
Datetime	Allows for manipulation of dates and times	[38]
Pytz	Allows cross-platform time zone calculations	[39]
Seaborn	Allows for data visualizations	[40]
Matplotlib	Library for creating visualizations in Python	[41]
Calplot	Used to create heat maps per calendar year by plotting time series data	[42]
Altair	Visualization Library for Python	[43]
Plotly	Python Graphing Library that makes interactive visualizations	[44]

Basic Use of Cortex

Building on the previous sections, this section explores how to import Cortex, score active data surveys, pull active data, pull passive data, and use several miscellaneous Cortex commands. This section discusses issues related to poor data quality (ie, low engagement from the user and errors in data collection) and how to use these commands to easily assess and troubleshoot both active and passive digital phenotyping data quality issues.

Importing Cortex

Cortex is a Python package that uses the mindLAMP application programming interface (API) to collect data from various users. In order to use its API, the user must have an access key provided by the mindLAMP consortium (Figure 7).

Figure 7. Codebook 5.

```
import LAMP
api = 'exampleapi.api'
email = 'exampleemail@gmail.com'
password = 'password'
lamp.connect(email, password, api)
import cortex
```

Checking Engagement and Data Quality

Participant engagement is critical to any study, especially for digital phenotyping, in which the goal is to collect active and passive data from their smartphones. A lack of active participant engagement limits the amount of data that can be pulled from LAMP, and missing data will reduce the quality of analysis that can be performed [6]. Engagement thus encapsulates (1) a participant's active participation in taking surveys and cognitive assessments and (2) a participant's passive data quality, a ratio of actual versus expected smartphone signal (passive) data captured each day. Due to this importance, it is best practice to check engagement and data quality metrics before proceeding with other analyses. Prior research suggests that active engagement is linked to passive data quality; if the app is not actively opened and used for several days, the phone may ignore the app's request to capture high-frequency passive data [6].

To assess participant's engagement with active data, it is useful to evaluate how many surveys or cognitive assessments have been completed in a certain time frame. Cortex supports such assessments with the command shown in Figure 8.

This command displays a graph of the activities that the participant completed, with the date of completion as the x-axis (Figure 9). Activities seen can vary from mindfulness exercises to the surveys that the participants complete.

Given that a study can be customized to capture different active and passive data streams at different frequencies, the parameters within the command to check data quality can be changed to suit different purposes. For the scope of the paper, every parameter for every command will not be explained in detail. Instead, the most important parameters will be explained, and a link to the documentation will be provided. The documentation will contain the other parameters not covered in the paper [45]. If a documentation link is not provided, all parameters are covered in detail.

The important parameters of cortex.visualizations.participant.active are explained as follows:

1. id_list: this parameter represents the user for which the researcher is generating the graph. To generate the graph for a single participant, one should make this parameter a Python string with the user ID. This parameter also accepts a list of participant IDs, study IDs, and researcher IDs. If >1 participant ID is provided or represented within a study or researcher ID, multiple graphs will be printed.
2. sample_length: this parameter represents the maximum number of days that can be represented in the generated

XSL•FO
RenderX

graph. It is possible to use an arbitrarily large number to pull every single day that a participant has completed a task. The graph will start with the first event completed.

3. attach_graphs: this parameter is a Boolean value (True or False) that, if true, will attach the graphs to the participant's Portal Tab in the mindLAMP app for which the graph is generated. This enables real-time sharing of results with participants directly via mindLAMP and provides a simple means for return of results.

An example of this command in use can be seen in Figure 10.

Figure 9 shows the activities of a participant who was asked to complete 2 of the same daily surveys per day. These surveys were named "All Daily Surveys" in mindLAMP and are represented as blue bars as seen in the legend to the right. Looking at the graph, the participant has been active in terms of their engagement with surveys and has completed at least 1 survey between December 19 and December 27. The participant completed the task of taking the survey twice on December 21, 22, 23, 24, and 26.

The command used to assess passive data quality is similar to that used to assess active data as shown in Figure 11.

This command generates 4 graphs per participant. Two of the graphs represent the amount of data coming from the GPS sensor. The other 2 graphs will represent the amount of data coming from the accelerometer sensor. The graphs generated will be either a scatterplot or a heat map. Each of these 2 graphs will have a representation of the "ideal" amount of data being received. For the scatterplot, a horizontal red line is placed. For the ideal amount of data collected, the scatter plot should be above the line. In terms of the heat map, a blue star will represent the ideal amount of data collected. These "ideal" amounts should only be used as a reference point. The true ideal amount of data will always be up to the researchers and set by the study.

The important parameters of cortex.visualizations.participant.passive are explained as follows [45]:

1. id_list: this parameter represents the user for which the researcher is generating the graph. To generate the graph for a single participant, make this parameter a Python string with the user ID. This parameter also accepts a list of participant IDs, study IDs, and researcher IDs. If >1 participant ID is provided or represented within a study or researcher ID, multiple graphs will be printed.
2. sample_length: this parameter represents how many days the user wants to query for. Unlike the sample_length parameter in cortex.visualizations.participant.active, the user cannot use an arbitrarily large number to query for all the data from a participant. If a user does take this approach, they will end up querying for days that a participant was not collecting data, causing the scale of the axes to be off and creating unnecessary blank space on the heat map.
3. days_ago: this parameter represents where the user wants the end of the x-axis to be.
4. reset: this parameter is a Boolean value (True or False) that will reset any previous graphs generated for the participant. If a graph has already been generated, any new information being queried for will be added to the existing graph. If the user does not want this, use reset=True.

An example of this command in use can be found in Figure 12. In this example, the researcher wanted to query for the first week of the participant's time during a study. The researcher knows that the participant consented to data collection 1 month ago (31 days). The user also had generated a previous graph and wanted to reset the existing graph.

The examples shown in Figures 13-16 highlight that LAMP is collecting passive data of accelerometer and GPS for this single user, although on some days, the data quality captured is below the target. While beyond the immediate scope of this paper, possible reasons include the phone being in low-power or airplane mode, lack of active engagement (which was assessed at the beginning of this section), or technical errors (one common error being the participant's device being on low-power mode). Monitoring passive data quality is easy and can be done on a weekly or even daily basis to rapidly spot issues and help a study or patient troubleshoot.

Figure 8. Codebook 6.

```
cortex.visualizations.participant.active(id)
```


XSL•FO
RenderX

Figure 9. Cortex active data visualization. This figure shows the activities that a participant has completed over time using mindLAMP. The x-axis shows the date of the activities completed, while the y-axis shows how many activities were completed. The color of the bar represents which activity was completed. The blue bars represent a created survey that the participant was supposed to take twice a day. The orange bar represents a survey that the participant completed in regard to the app itself and its ease of use. The red bar is another survey that represents responses to questions on the digital working alliance. D-WAI: Digital Working Alliance Inventory.

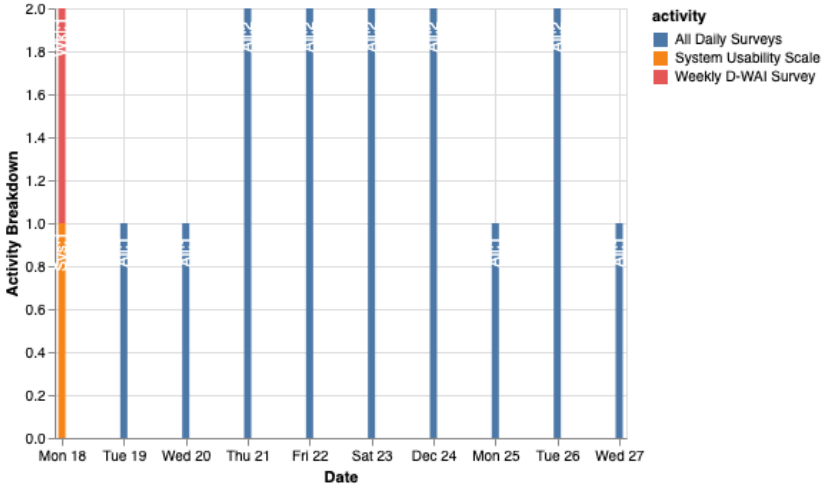

Figure 10. Codebook 7.

```
cortex.visualizations.participant.active("U0929038949", sample_length=100,
attach_graphs = True)
```

Figure 11. Codebook 8.

```
cortex.visualizations.participant.passive(id)
```

Figure 12. Codebook 9.

```
cortex.visualizations.participant.passive("U0929038949", sample_length =
7,
days_ago=27, reset=True)
```

Figure 13. This figure shows how many GPS data points are being collected over time. The x-axis represents the time of data point collection, and the y-axis represents the amount. Ideally, the user would want to see the blue line never drop below the red.

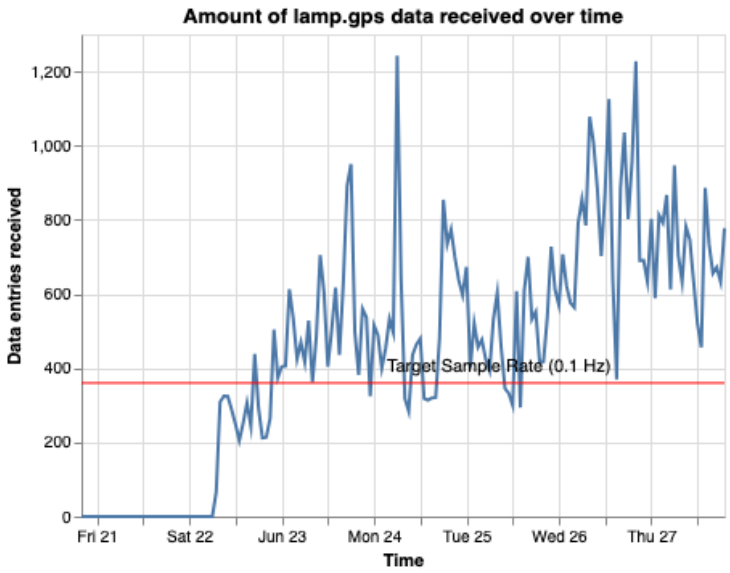

XSL•FO
RenderX

Figure 14. This figure shows how many GPS data points are being collected over time. The beginning of the x-axis represents the very start of the day, 12 AM or 00:00 in 24-hour time. Each hour is represented as a square and given a gradient of blue, with the deeper blues representing more data points collected. Each row represents 1 day and starts at the top. Ideally, the user would want each square to be blue with a little star within.

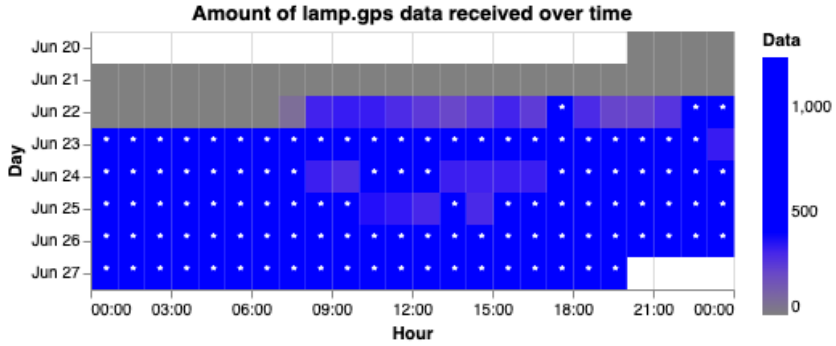

Figure 15. This figure shows how many accelerometer data points are being collected over time. The x-axis represents the time of data point collection, and the y-axis represents the amount. Ideally, the user would want to see the blue line never drop below the red.

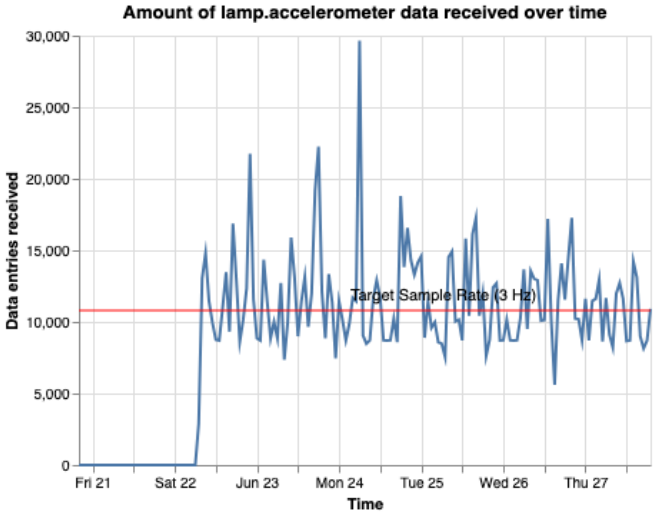

Figure 16. This figure shows how many accelerometer data points are being collected over time. The beginning of the x-axis represents the very start of the day, 12 AM or 00:00 in 24-hour time. Each hour is represented as a square and given a gradient of blue, with the deeper blues representing more data points collected. Each row represents 1 day and starts at the top. Ideally, the user would want each square to be blue with a little star within.

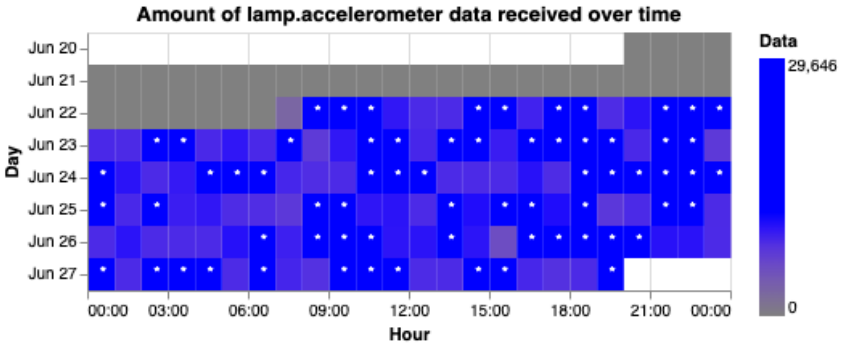

Activity Events

Surveys and sensors are not the only types of data that can be captured. mindLAMP also offers functionality beyond active and passive data collection to enable users to engage with psychoeducation (the L in LAMP) and complete therapeutic exercises (the M in LAMP). For some studies, this can be important or even the focus of the research question. For example, “Does the efficacy of completing mindfulness on mindLAMP vary by location or does cognitive therapy work better on nights with more sleep duration?” To accomplish this from a data perspective, the user can access information regarding everything a participant engages in within mindLAMP. Every time a participant completes an activity within mindLAMP, details about the event are recorded. These details include when the activity was started, how long it took, and specific details relating to the activity. The specific details can include the answer to a question within a survey, the results of a cognitive assessment, or the time stamp when a mindfulness exercise was completed.

XSL•FO
RenderX

A researcher or clinician can create their own activities, including custom surveys, custom psychoeducation, and custom therapeutic activities. To ensure that these custom activities can be easily scored and work well with Cortex, there is a standard data format that is important to understand when seeking data from any activity event. In order to view these activity events for any study, this Cortex command, which can be seen in Figure 17, can be used.

The command will print a Python dictionary that contains every activity a participant has completed. The dictionary will also contain information related to the activity being completed, which will be contained within "temporal_slices," a key within the dictionary (or column within a DataFrame). For example, if a participant took a survey, every question within the survey would be contained within "temporal_slices."

An example of the command in use can be found in Figure 18.

Since this command will print every activity completed, it can have a very long output. In the example (Figure 19), only the most recent activity completed will be printed.

Going through this output, it can be seen that there are 6 keys within the Python dictionary, each describing something specific about the activity that was completed. Starting at the top in the "temporal_slices" key, the information contained within is specific details about the activity event and actions within. In this example (Figure 19), since the activity event was a survey, each answer to the questions is recorded, along with how long it took to answer the question. The next key in the dictionary is "duration." This is simply how long the entire activity event took. The "static_data" key is empty in this activity. Static data refer to the details about certain activity events, specifically the games offered within mindLAMP. These details include points scored, correct answers, attempts, and so on. Moving to the next key, "activity" represents the activity event that was completed. Each activity event is given a string for identification regardless of activity type (eg, 1 survey will have a different identification string from another). Within this example, the activity completed was "All Daily Surveys." The final 2 keys are "timestamp" and "_parent," which are the time the activity was completed and who completed the activity, respectively.

Figure 17. Codebook 10.

```
LAMP.ActivityEvent.all_by_participant(id)["data"]
```

Figure 18. Codebook 11.

```
LAMP.ActivityEvent.all_by_participant(part)["data"][0]
```

Figure 19. Codebook 12.


```
{'temporal_slices': [{'item': 'Overall, how would I rate my anxiety
today?',
   'value': '2',
   'duration': 3750,
   'type': None,
   'level': None},
  {'item': 'I am able to manage my day-to-day life',
   'value': '3',
   'duration': 7116,
   'type': None,
   'level': None},
  {'item': 'Overall, how would you rate your mood today?',
   'value': '7',
   'duration': 11154,
   'type': None,
   'level': None},
  {'item': 'Approximately what time did you fall asleep last night?',
   'value': '04:00AM',
   'duration': 56773,
   'type': None,
   'level': None},
  {'item': 'Approximately what time did you wake up this morning?',
   'value': '07:00AM',
   'duration': 75563,
   'type': None,
   'level': None},
  {'item': 'How much physical exercise did you engage in today? ',
   'value': '30-45 minutes',
   'duration': 23957,
   'type': None,
   'level': None}],
 'duration': 178313,
 'static_data': {},
 'activity': 'ywse81hs411cdwvgcr5n',
 'timestamp': 1703646117283,
 '_parent': 'U0929038949'}
```


XSL•FO
RenderX

Survey Grading

Surveys, which are a common form of active data collection, can be completed on the LAMP platform and scored with Cortex. Scoring with Cortex is an efficient way of analyzing the results of the surveys and prevents mistakes caused by calculating scores by hand.

To score a survey using Cortex, a Python dictionary needs to be created. These dictionaries provide Cortex with instructions on how the user wants the survey to be scored. For example, a participant may have completed a daily survey that asked them 3 questions. The 3 questions being "Overall, how would you rate your mood today?" "Overall, how would I rate my anxiety today?" and "How much physical exercise did you engage in today?" relating to the participant's mood, anxiety, and exercise level, respectively.

This scoring dictionary can be seen in Figure 20, with scoring parameters easily adjusted by the researcher.

This dictionary can be updated with any survey that has been taken, but the dictionary can only take numeric outputs, whether that be a Boolean (True or False) or an integer.

With a start time, end time, and scoring dictionary, the user can query for scores. The scores will be labeled to match how categories are named, in this case they will be labeled "Daily Mood Survey," "Daily Anxiety Survey," and "Daily Exercise Survey."

The command to grade the survey can be seen in Figure 21.

The important parameters of cortex.primary.survey_scores.survey_scores are explained as follows:

1. id: this parameter represents the participant's survey that is being graded. Unlike the Cortex visualizations, ID is limited to 1 participant at a time. If the user wants to grade multiple participants at once, consider a loop. In addition, only putting the user ID without the parameter title (id=) will cause an error.
2. start: this parameter represents the start of the time period the user wants to query for (epoch time).
3. end: this parameter represents the end of the time period the user wants to query for (epoch time).
4. scoring_dict: this parameter represents the scoring dictionary that was created in Figure 20.
5. return_ind_ques (optional): this parameter is a Boolean value (True or False) that, if True, will return the individual question along with the results of the survey. Since this parameter is optional, if it is not included within the command, Cortex assumes it to be false.

An example of this command can be found in Figure 22.

In this example, the researcher wanted to query every survey taken from December 19, 2023, to December 20, 2023. The researcher also did not want Cortex to return the individual questions within the surveys (Figure 23).

Figure 20. Codebook 13.

```
score_dict = {'category_list': ['Daily Mood Survey', 'Daily Anxiety
Survey', 'Daily Exercise Survey'],
              'questions': {
                  'Overall, how would you rate your mood today?':
{'category': 'Daily Mood Survey', 'scoring': 'M_map'},
                  'Overall, how would I rate my anxiety today?':
{'category': 'Daily Anxiety Survey', 'scoring': 'A_map'},
                  'How much physical exercise did you engage in today? ':
{'category': 'Daily Exercise Survey', 'scoring': 'Exercise_map'},
              },
              'A_map': {'10': 10, '9': 9, '8': 8, '7': 7, '6': 6, '5': 5,
'4': 4, '3': 3, '2': 2, '1': 1, '0': 0},
              'M_map': {'10': 10, '9': 9, '8': 8, '7': 7, '6': 6, '5': 5,
'4': 4, '3': 3, '2': 2, '1': 1, '0': 0},
              'Exercise_map': {'None': 0, '< 15 minutes': 1, '15-30
minutes': 2, '30-45 minutes': 3, '45-60 minutes': 4, '60+': 5}}
```

Figure 21. Codebook 14.

```
cortex.primary.survey_scores.survey_scores(id=, start=, end=,
                                           scoring_dict=score_dict)['data']
```


XSL•FO
RenderX

Figure 22. Codebook 15.

```
cortex.primary.survey_scores.survey_scores(id="U0929038949",
                                           start=1703005200000,
                                           end=1703091600000,
                                           scoring_dict=score_dict)['data']
```

Figure 23. Codebook 16.

```
[{'start': 1703021202227,
  'end': 1703022012667,
  'category': 'Daily Anxiety Survey',
  'question': 'Daily Anxiety Survey',
  'score': 1},
 {'start': 1703021202227,
  'end': 1703022012667,
  'category': 'Daily Mood Survey',
  'question': 'Daily Mood Survey',
  'score': 10},
 {'start': 1703021202227,
  'end': 1703022012667,
  'category': 'Daily Exercise Survey',
  'question': 'Daily Exercise Survey',
  'score': 5}]
```

Passive Data Processing

Passive data streams (eg, accelerometers) are often the most useful when that data can be transformed into behavioral features. Cortex has 2 ways of processing the raw passive data collected from the LAMP platform. The first way a user can process passive data is individually, through the format of the command shown in Figure 24.

The important parameters of cortex.secondary.feature_name.feature_name are explained as follows:

1. feature_name: this parameter is in the name of the command. It represents what passive data feature the user wants to query for (refer to Table 2 for the list of feature_names that are currently available).
2. id: this parameter represents the participant that the user wants to query for.
3. start: this parameter represents the start of the time period the user wants to query for (epoch time).
4. end: this parameter represents the end of the time period the user wants to query for (epoch time).
5. resolution: this parameter represents the time frame of which the passive data feature is calculated in milliseconds. For example, if the user wanted to calculate hometime in A DAY, the user would have a resolution equal to the number of milliseconds in a day.
6. The parameters above are required for ALL features, but some features require more parameters. Cortex will produce an error if the required parameters are not included while executing the command. The features that require more than the parameters above can be seen in Table 2.

An example of this command can be found in Figure 25.

With this command, the user is looking for how much time the participant spent at home from December 19, 2023, to December 20, 2023, specifically 12 AM, for both dates. Since the resolution is binned by the number of milliseconds in a day, Cortex will return the amount of time at home per day. If the user changed the resolution to the number of milliseconds in a week, Cortex would return the amount of time at home that week. It is important to note that the difference between the start and end time needs to be greater than the resolution by at least 1 millisecond. This is why, in the example, the end time ends with a 1 (Figure 26).

Looking at the results, it can be seen that this participant spent 85,599,870 milliseconds at home, but for the sake of understanding, this value should be changed into something more digestible. This can be seen in the example continuation in Figure 27.

This command uses the pandas package to take the “value,” which is the amount of home time in milliseconds and converts it into hours (Figure 28).

Within this example, it can be seen that this participant has spent nearly the entire day at home.

The second way that Cortex can process sensor data is with the cortex.run() command. The functionality of the parameters is largely the same as that of the cortex.secondary (Figure 29).

XSL•FO
RenderX

Cortex.run is meant to be customizable. Every parameter, from ID to the start and end times, is meant to give all the information one needs in one go. Going through the parameters, examples of what can be customized are given as follows (rest of the parameters are given elsewhere) [46]:

1. id_or_set: this parameter is who the user is querying for. The user can run this command with 1 participant, many participants in a contained list, or a study or researcher ID.
2. features: this parameter describes what to query for (refer to Table 2 for the features that are available to be queried).
3. start: this parameter represents the start time in epoch time.
4. end: this parameter represents the end time in epoch time.
5. resolution: this parameter represents the time frame of which the passive data feature is calculated in milliseconds. For example, if the user wanted to calculate hometime in A DAY, the user would have a resolution equal to the number of milliseconds in a day.

An example of this command in use can be found in Figure 30.

In this example, the researcher wanted to query for both screen_duration and data_quality for 1 week. Within the feature_params parameter, it is specified that data quality will be calculated using the GPS sensor and not the accelerometer. The bin size is set at 3,600,000, which means that Cortex will split the time between the start and end time into “chunks” of 3,600,000 milliseconds, which is the equivalent to 1 hour. The value calculated under data_quality is the percentage of bins or “chunks” that have a point of either GPS or accelerometer data. In this case it is GPS data. The results below are cleaned up from the normal result using the “tabulate” Python package and provide a visually appealing table (Figure 31).

The actual code chunk will print a data dictionary, which can be confusing to read; thus, it is recommended that the user save it as a variable when using cortex.run. Once it is saved as a variable, the user can call individual keys from the dictionary. An example of this can be seen in Figure 32.

Researchers also can pull raw sensor data. The functionality of this command is much like pulling activity events, which can be referred to in the *Activity Events* section. The command can be run without parameters but will return an incomprehensible amount of data (Figure 33).

Parameters can be used within this command to narrow how much data are received and what sensor spec to return:

1. id: this parameter represents the participant that the user wants to query for.
2. origin: this parameter represents what sensor spec the user wants to query for (refer
   Table 1
   for the features that are available to be queried).
3. _from: this parameter represents the start time in epoch time.
4. to: this parameter represents the end time in epoch time.
5. _limit: this parameter represents the maximum amount of data points to pull.

An example of this command in use can be seen in Figure 34.

In this example, the researcher wanted to query the first 5 accelerometer data points for the participant.

Looking at the example (Figure 35), it can be seen what raw data looks like, and it highlights the necessity for data pipelines such as Cortex.

With the information present thus far, the reader is able to pull both active and passive data. With this background, the paper will now shift into data manipulation and how to create data visualizations that can give the user insight into participant’s behavior in relation to data gathered.

Figure 24. Codebook 17.

```
cortex.secondary.feature_name.feature_name(id=, start=, end=, resolution=)
```

Figure 25. Codebook 18.

```
hometime = cortex.secondary.hometime.hometime(id=part,
                                              start=1702962000000,
                                              end=1703048400001,
                                              resolution=MS_IN_DAY)
```

Figure 26. Codebook 19.

```
{'timestamp': 1702962000000,
 'duration': 86400001,
 'resolution': 86400000,
 'data': [{'timestamp': 1702962000000, 'value': 85599870}]}
```


XSL•FO
RenderX

Figure 27. Codebook 20.

```
pd.DataFrame(hometime["data"])["value"] / (3600 * 1000)
```

Figure 28. Codebook 21.

```
0    23.777742
Name: value, dtype: float64
```

Figure 29. Codebook 22.

```
cortex.run(id_or_set, features=[], feature_params={}, start=None,
end=None,
                                                resolution=MS_PER_DAY)
```

Figure 30. Codebook 23.

```
cortex.run(part,
          ['screen_duration', 'data_quality'],
          feature_params={'data_quality': {"feature":"gps",
"bin_size":3600000}},
          start=1702962000000,
          end=1702962000000 + 7 * MS_IN_DAY)
```

Figure 31. Codebook 24.

```
screen_duration:
+-------------+---------------+------------+---------------------+
|     id      |   timestamp   |   value    |      datetime       |
+-------------+---------------+------------+---------------------+
| U0929038949 | 1702962000000 | 36803989.0 | 2023-12-19 05:00:00 |
| U0929038949 | 1703048400000 | 16234234.0 | 2023-12-20 05:00:00 |
| U0929038949 | 1703134800000 | 26260277.0 | 2023-12-21 05:00:00 |
| U0929038949 | 1703221200000 | 48899441.0 | 2023-12-22 05:00:00 |
| U0929038949 | 1703307600000 | 38569785.0 | 2023-12-23 05:00:00 |
| U0929038949 | 1703394000000 | 11286251.0 | 2023-12-24 05:00:00 |
| U0929038949 | 1703480400000 |    0.0     | 2023-12-25 05:00:00 |
+-------------+---------------+------------+---------------------+
data_quality:
+-------------+---------------+---------------------+---------------------
+
|     id      |   timestamp   |        value        |      datetime
|
+-------------+---------------+---------------------+---------------------
+
| U0929038949 | 1702962000000 | 0.9166666666666666  | 2023-12-19 05:00:00
|
| U0929038949 | 1703048400000 |        0.625        | 2023-12-20 05:00:00
|
| U0929038949 | 1703134800000 | 0.5416666666666666  | 2023-12-21 05:00:00
|
| U0929038949 | 1703221200000 |         1.0         | 2023-12-22 05:00:00
|
| U0929038949 | 1703307600000 |         1.0         | 2023-12-23 05:00:00
|
| U0929038949 | 1703394000000 | 0.16666666666666666 | 2023-12-24 05:00:00
|
| U0929038949 | 1703480400000 |         0.0         | 2023-12-25 05:00:00
|
+-------------+---------------+---------------------+---------------------
+
```


XSL•FO
RenderX

Figure 32. Codebook 25.

```
variable = cortex.run(part,
          ['screen_duration', 'data_quality'],
          feature_params={'data_quality': {"feature":"gps", "bin_size":3600000}},
          start=1702962000000,
          end=1702962000001 + 7 * MS_IN_DAY)

variable['screen_duration']
```

Figure 33. Codebook 26.

```
LAMP.SensorEvent.all_by_participant(id)
```

Figure 34. Codebook 27.

```
LAMP.SensorEvent.all_by_participant(part, _limit=5,
                                    origin="lamp.accelerometer")["data"]
```

Figure 35. Codebook 28.


```
[{'data': {'x': 0.31465616822242737,
   'y': 4.24367094039917,
   'z': 8.520841598510742},
  'sensor': 'lamp.accelerometer',
  'timestamp': 1703406991186},
 {'data': {'x': 0.19621145725250244,
   'y': 4.146761417388916,
   'z': 9.060422897338867},
  'sensor': 'lamp.accelerometer',
  'timestamp': 1703406990985},
 {'data': {'x': 0.3026920557022095,
   'y': 4.328616142272949,
   'z': 9.351150512695312},
  'sensor': 'lamp.accelerometer',
  'timestamp': 1703406990786},
 {'data': {'x': 0.2069791555404663,
   'y': 3.957728624343872,
   'z': 8.622536659240723},
  'sensor': 'lamp.accelerometer',
  'timestamp': 1703406990584},
 {'data': {'x': 0.33379876613616943,
   'y': 4.285545349121094,
   'z': 9.06401252746582},
  'sensor': 'lamp.accelerometer',
  'timestamp': 1703406990381}]
```


Advanced Participant Level Example and Data Manipulation

Overview

Cortex makes pulling data simple, but most of the time, the data format given is not compatible with various Python packages for creating charts, tables, and different machine learning techniques. Due to this difficulty, it is worthwhile to consider an example where various levels of data manipulation are used to create a pandas DataFrame that will allow for Python packages to be used. The subsequent subheadings (or sections) will elaborate (1) grading surveys and pulling passive data for a participant, (2) how to convert time stamps in epoch time into a date format, (3) splitting DataFrames, (4) cleaning DataFrames, (5) merging of DataFrames and final cleanup using

XSL•FO
RenderX

pandas, (6) creating the correlation matrix and making it visually appealing, and (7) verifying and gathering more information.

While many types of data visualization are possible, 1 compelling data visualization that is often used is a correlation matrix. These matrices can point the user in the right direction in terms of understanding the relationships between active and passive data. However, it is important to understand that although variables may correlate with each other, there may be confounding variables that are not represented. To combat this, extra examples will be given to implement scientific processes and give weight to what the data are telling the user.

This example will go through each step to create the correlation matrix. These steps will also be labeled in reference to the enumerated list above for easy access.

Within the correlation matrix, the variables that will be included are (1) anxiety (measured daily through surveys on a scale of 1-10), (2) mood (measured daily through surveys on a scale of 1-10), (3) exercise length (measured daily through surveys; scale is a range of durations up to 60 min), (4) screen duration (collected from passive data sensors), (5) step count (collected from passive data sensors), and (6) data quality (calculated using passive data sensors).

Grading Surveys and Pulling Passive Data for a Participant

After importing packages and connecting them to the API, the user should specify the user ID and save it within a variable. in this example, “Newpart.” The user should then pull the active and passive data using the respective Cortex commands and save the results within properly named variables (Figure 36).

Figure 36. Codebook 29.

```
Newpart = 'U3971360469'

active = pd.DataFrame(cortex.primary.survey_scores.survey_scores(id=Newpart,
                start=1702314000000, end=1702314000000 + 14 * MS_IN_DAY,
scoring_dict=score_dict)['data'])

passive = cortex.run(Newpart,
  ['screen_duration', 'data_quality', 'step_count'],
  feature_params={'data_quality': {"feature":"gps", "bin_size":3600000}},
  start=1702314000000,
  end=1702314000000 + 14 * MS_IN_DAY)
```

The dictionary used to grade the surveys is the same dictionary used within the *Survey Grading* subsection.

Converting Time Stamps

Taking a look at the “active” variable, the start time stamp needs to be converted from epoch time, which is in UTC, to a date in EST. Doing this will allow the user to merge the passive and active DataFrames on the formatted date. Knowing that the surveys started and ended on the same date, only the start time stamp needs to be converted. The end time stamp will be deleted. Figure 37 shows the current state of the active variable before time stamp conversion.

The passive data do not need as many steps because the cortex.run produces a datetime value. The user does need to be wary that this datetime value is in the UTC time zone and needs to be converted (Figure 38).

Screen_duration is currently contained within the “passive” dictionary. Though only screen_duration is shown, both data_quality and step_count are also contained within “passive” and look identical to screen_duration in terms of format. Figure 38 shows the first 9 rows of screen_duration within the passive dictionary.

XSL•FO
RenderX

Figure 37. Codebook 30.

start	end	category	question	score
1703377320153	1703379135290	Daily Anxiety Survey	Daily Anxiety Survey	5
1703377320153	1703379135290	Daily Mood Survey	Daily Mood Survey	3
1703377320153	1703379135290	Daily Exercise Survey	Daily Exercise Survey	2
1703206840167	1703256334979	Daily Anxiety Survey	Daily Anxiety Survey	6
1703206840167	1703256334979	Daily Mood Survey	Daily Mood Survey	5
1703206840167	1703256334979	Daily Exercise Survey	Daily Exercise Survey	3
1703189284252	1703191154192	Daily Anxiety Survey	Daily Anxiety Survey	7
1703189284252	1703191154192	Daily Mood Survey	Daily Mood Survey	3
1703189284252	1703191154192	Daily Exercise Survey	Daily Exercise Survey	2

```
# Convert epoch time in milliseconds to datetime
active['start'] = pd.to_datetime(active['start'], unit='ms')

# Set the timezone to NYC
nyc_timezone = pytz.timezone('America/New_York')
active['start'] =
active['start'].dt.tz_localize(pytz.utc).dt.tz_convert(nyc_timezone)

# Extract date without time
active['start'] = active['start'].dt.date

#Dropping end timestamp
active = active.drop('end', axis=1)
```

start	category	question	score
2023-12-23	Daily Anxiety Survey	Daily Anxiety Survey	5
2023-12-23	Daily Mood Survey	Daily Mood Survey	3
2023-12-23	Daily Exercise Survey	Daily Exercise Survey	2
2023-12-21	Daily Anxiety Survey	Daily Anxiety Survey	6
2023-12-21	Daily Mood Survey	Daily Mood Survey	5
2023-12-21	Daily Exercise Survey	Daily Exercise Survey	3
2023-12-21	Daily Anxiety Survey	Daily Anxiety Survey	7
2023-12-21	Daily Mood Survey	Daily Mood Survey	3
2023-12-21	Daily Exercise Survey	Daily Exercise Survey	2

XSL•FO
RenderX

Figure 38. Codebook 31.

```
#Changing screen_duration to date instead of datetime
nyc_timezone = pytz.timezone('America/New_York')
passive['screen_duration']['datetime'] =
passive['screen_duration']['datetime'].dt.tz_localize(pytz.utc).dt.tz_convert(nyc_
timezone)
passive['screen_duration']['datetime'] =
passive['screen_duration']['datetime'].dt.date
#Changing data_quality to date instead of datetime
nyc_timezone = pytz.timezone('America/New_York')
passive['data_quality']['datetime'] =
passive['data_quality']['datetime'].dt.tz_localize(pytz.utc).dt.tz_convert(nyc_tim
ezone)
passive['data_quality']['datetime'] = passive['data_quality']['datetime'].dt.date
#Changing step_count to date instead of datetime
nyc_timezone = pytz.timezone('America/New_York')
passive['step_count']['datetime'] =
passive['step_count']['datetime'].dt.tz_localize(pytz.utc).dt.tz_convert(nyc_timez
one)
passive['step_count']['datetime'] = passive['step_count']['datetime'].dt.date
```

id	timestamp	value	datetime
U3971360469	1702314000000	18776209.0	2023-12-11
U3971360469	1702400400000	18280763.0	2023-12-12
U3971360469	1702486800000	33523357.0	2023-12-13
U3971360469	1702573200000	26536546.0	2023-12-14
U3971360469	1702659600000	18876016.0	2023-12-15
U3971360469	1702746000000	17029980.0	2023-12-16
U3971360469	1702832400000	26569784.0	2023-12-17
U3971360469	1702918800000	15253011.0	2023-12-18
U3971360469	1703005200000	0.0	2023-12-19

Splitting DataFrames

For the sake of simplicity in merging and reducing redundancy, the separation of each variable allows for precise changes to be made (Figure 39).

Figure 39. Codebook 32.

```
screen_duration = passive['screen_duration']
step_count = passive['step_count']
data_quality = passive['data_quality']
mood = active[active['category'] == 'Daily Mood Survey']
anxiety = active[active['category'] == 'Daily Anxiety Survey']
exercise = active[active['category'] == 'Daily Exercise Survey']
```

Cleaning DataFrames Using Pandas

When cleaning data and preparing it for merging, it makes sense to have a plan regarding what the user wants the DataFrame to look like. This can help to prevent unnecessary commands. In this example, the final DataFrame will have a column for the UserID, a date that will be merged, and 1 column per data feature. Currently, the passive DataFrames have the most information in regard to the columns, but only 1 of the 3 needs more than just the date and its respective data. Knowing this, the data_quality DataFrame will be established as the base and will be the only DataFrame that keeps its UserID column. It is also worth noting that each of the “value” columns, which are within each of the DataFrames, will need to be renamed into what the value represents. For instance, for the step_count DataFrame, the “value” column needs to be renamed into “stepcount” or some variation of that (Figure 40).

Notice how in the DataFrames shown in Figure 40, there are multiple instances of the same date that have a score. This is because this participant took >1 survey in a day. There are many

XSL•FO
RenderX

different solutions to remedy this, but for this example, the average between the scores will be taken (Figure 41).

Figure 40. Codebook 33.

```
#Renaming value columns to what the value represents for the passive features
screen_duration = screen_duration.rename(columns = {'value':'Screen Duration'})
data_quality = data_quality.rename(columns = {'value':'Data Quality'})
step_count = step_count.rename(columns = {'value':'Step Count'})
#Dropping unnecessary columns for the passive features
data_quality = data_quality.drop(columns=['timestamp'])
screen_duration = screen_duration.drop(columns=['timestamp', 'id'])
step_count = step_count.drop(columns=['timestamp', 'id'])
+-------------+--------------------+------------+
|     id      |    Data Quality    |  datetime  |
+-------------+--------------------+------------+
| U3971360469 | 0.8333333333333334 | 2023-12-11 |
| U3971360469 | 0.9166666666666666 | 2023-12-12 |
| U3971360469 |        1.0         | 2023-12-13 |
| U3971360469 | 0.9166666666666666 | 2023-12-14 |
| U3971360469 |       0.75         | 2023-12-15 |
| U3971360469 | 0.5833333333333334 | 2023-12-16 |
| U3971360469 | 0.9166666666666666 | 2023-12-17 |
| U3971360469 |      0.625         | 2023-12-18 |
| U3971360469 |       0.0          | 2023-12-19 |
| U3971360469 |       0.0          | 2023-12-20 |
+-------------+--------------------+------------+
+------------+------------+
| Step Count |  datetime  |
+------------+------------+
|  747631.0  | 2023-12-11 |
|  594845.0  | 2023-12-12 |
| 3718114.0  | 2023-12-13 |
|  431599.0  | 2023-12-14 |
|  206163.0  | 2023-12-15 |
| 1043189.0  | 2023-12-16 |
|  726736.0  | 2023-12-17 |
| 1205445.0  | 2023-12-18 |
|    nan     | 2023-12-19 |
|    nan     | 2023-12-20 |
+------------+------------+
+----------------+------------+
| Screen Duration |  datetime  |
+----------------+------------+
|   18776209.0   | 2023-12-11 |
|   18280763.0   | 2023-12-12 |
|   33523357.0   | 2023-12-13 |
|   26536546.0   | 2023-12-14 |
|   18876016.0   | 2023-12-15 |
|   17029980.0   | 2023-12-16 |
|   26569784.0   | 2023-12-17 |
|   15253011.0   | 2023-12-18 |
|       0.0      | 2023-12-19 |
+----------------+------------+
#Dropping unnecessary columns for the active features
anxiety = anxiety.drop(columns=['category', 'question'])
exercise = exercise.drop(columns=['category', 'question'])
mood = mood.drop(columns=['category', 'question'])
#Renaming value columns to what the value represents for the active features
anxiety = anxiety.rename(columns = {'score':'Anxiety Score'})
exercise = exercise.rename(columns = {'score':'Exercise Score'})
mood = mood.rename(columns = {'score':'Mood Score'})
#Renaming date columns from start to datetime in for the active features
anxiety = anxiety.rename(columns = {'start':'datetime'})
exercise = exercise.rename(columns = {'start':'datetime'})
mood = mood.rename(columns = {'start':'datetime'})
+------------+------------+
|  datetime  | Mood Score |
+------------+------------+
| 2023-12-23 |     3      |
| 2023-12-21 |     5      |
| 2023-12-21 |     3      |
| 2023-12-19 |     3      |
| 2023-12-18 |     2      |
| 2023-12-18 |     2      |
| 2023-12-17 |     3      |
| 2023-12-17 |     3      |
+------------+------------+
```

Figure 41. Codebook 34.

```
#Taking the average of two survey scores if scores share a date
mood = mood.groupby('datetime')['Mood Score'].mean().reset_index()
exercise = exercise.groupby('datetime')['Exercise Score'].mean().reset_index()
anxiety = anxiety.groupby('datetime')['Anxiety Score'].mean().reset_index()
```


XSL•FO
RenderX

Merging of DataFrames and Final Cleanup

Now that every data point is attached to its formatted date (Figure 42), each of the DataFrames is ready to be merged.

Now that the DataFrames are merged (Figure 42), the user can look for discrepancies. Two glaring edits that need to be implemented into this DataFrame are switching the locations of "Data Quality" and "datetime" and changing the units of "Screen Duration" from milliseconds to hours.

A keen eye will also notice strange results within the "Step Count" column. It is known that the values within "Step Count" are supposed to represent how many steps a participant has taken within a day, but the values obtained are way too high to be accurate. It is unknown what went wrong while collecting this participant's data, but it goes to show that researchers should keep a sharp eye out for discrepancies. For this example, the step count correlations will be taken with a grain of salt.

Finally, on December 19 and 20, the screen duration is 0 for both. Since 0 can and will be used within correlations, the result can be skewed by outliers. Due to this fact, it is reasonable to turn these *0*'s into NaN values using the Numpy Python package (Figure 43).

Figure 42. Codebook 35.

```
#Merging all DataFrames into one
final_df = pd.merge(data_quality, screen_duration, how = 'left', on = 'datetime')
final_df = pd.merge(final_df, step_count, how = 'left', on = 'datetime')
final_df = pd.merge(final_df, anxiety, how = 'left', on = 'datetime')
final_df = pd.merge(final_df, mood, how = 'left', on = 'datetime')
final_df = pd.merge(final_df, exercise, how = 'left', on = 'datetime')
```

Figure 43. Codebook 36.

```
#Switching the column locations of datetime and data_quality
final_df['Data Quality'], final_df['datetime'] = final_df['datetime'].copy(),
final_df['Data Quality'].copy()
final_df = final_df.rename(columns = {'datetime': 'Data Quality', 'Data Quality':
'datetime'})

#Converting Screen Duration in MS to hours
final_df['Screen Duration'] = final_df['Screen Duration']/3600000

final_df['Screen Duration'][8] = np.nan
final_df['Screen Duration'][9] = np.nan
```

Creating the Correlation Matrix and Making It Visually Appealing

The correlation matrix (Figure 44) is created with a simple command using the "pandas" Python package. To make it visually appealing, the "seaborn" package will be used. The result is shown in Figure 45.

Looking at the graph, it is possible to see an overview of how the variables may correlate with each other. It is important to note not only that correlation does not equal causation but also how this graph can help share back both active and passive data with participants. Scatterplots are also helpful in obtaining an alternative image of what these data may represent.

Figure 44. Codebook 37.

```
#Creation of correlation matrix
correlation_matrix = final_df.corr()

#Eliminating top half of Correlation matrix including correlations of 1
mask = np.triu(np.ones_like(correlation_matrix, dtype=bool))
correlation_matrix = correlation_matrix.mask(mask)
#Dropping rows with no values
correlation_matrix = correlation_matrix.drop(index='Data Quality',
columns='Exercise Score')
#Figure creation
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", vmin=-1,
vmax=1);
```


XSL•FO
RenderX

Figure 45. Correlation matrix. This graph shows each variable and its relationship to the other variables within the final DataFrame. Correlations closer to 1 represent a positive linear correlation, whereas correlations close to negative 1 represent a negative linear correlation. Positive in this context means that both variables are going in the same direction. Negative means that the variables are heading in opposite directions.

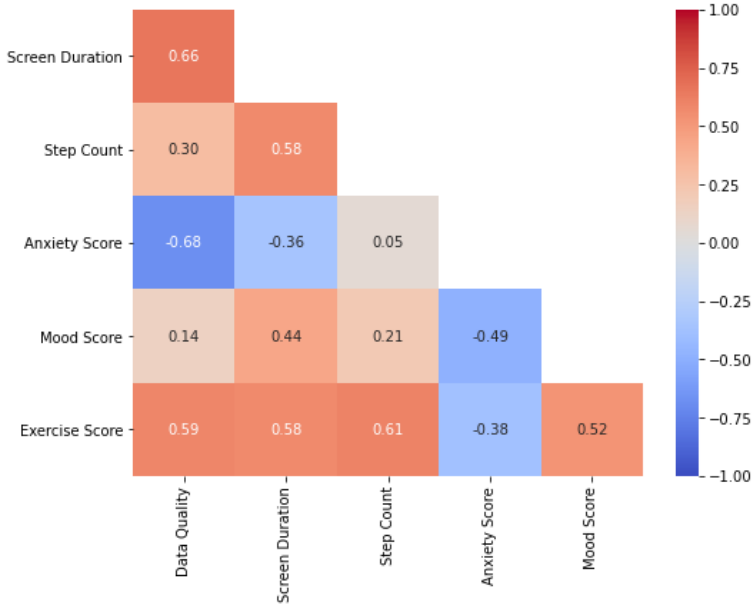

Figure 46. Codebook 38.

```
plt.scatter(final_df['Exercise Score'], final_df['Mood Score'])
plt.xlabel('Exercise Score')
plt.ylabel('Mood Score')
plt.title('Basic Scatterplot')
```

Verifying and Gathering More Information

In the correlation matrix shown in Figure 45, there are some interesting correlations to look at. Perhaps the most interesting is the 0.52 correlation between exercise score and mood. The assumption that most people would make is that if the exercise score is higher, the participant's mood would also be higher, but to ensure this is the case, a scatter plot should be made (Figure 46). The scatter plot results are shown in Figure 47.

Looking at the graph, it can be seen that the initial assumption that mood improves as the exercise score increases looks to be correct. For the truly diligent user, more information can be gathered by running a simple linear regression, especially due to the linear look of the data points. To do this, a Python package named "statsmodels.api" will be used.

Looking at the *P* value in red in Figure 48, it can be seen that this relationship is approaching the point where it would be considered significant ($P<.05$) This is not to say that this is not a valuable result. There absolutely is a case to be made that this participant's mood can benefit from increased exercise.

XSL•FO
RenderX

Figure 47. Scatterplot of exercise score versus mood score. This graph shows the mood score with respect to the exercise score. The x-axis represents the exercise score, with higher scores meaning the participant completed more exercise that day. The y-axis represents the mood score, with higher scores representing a more positive mood. Each dot represents 1 day.

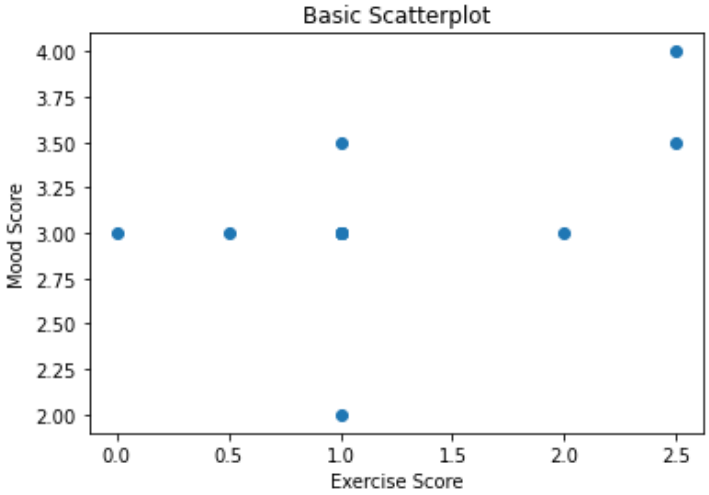

XSL•FO
RenderX

Figure 48. Codebook 39.

```
#Linear Regression
import statsmodels.api as sm
X = final_df[['Exercise Score']]
Y = final_df['Mood Score']
X = sm.add_constant(X)
X = X.dropna()
Y = Y.dropna()

model = sm.OLS(Y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Mood Score   R-squared:                       0.275
Model:                            OLS   Adj. R-squared:                  0.194
Method:                 Least Squares   F-statistic:                     3.409
Date:                Thu, 29 Feb 2024   Prob (F-statistic):             0.0979
Time:                        16:59:16   Log-Likelihood:                -5.4891
No. Observations:                  11   AIC:                             14.98
Df Residuals:                       9   BIC:                             15.77
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              2.6893      0.255     10.552      0.000       2.113       3.266
Exercise Score     0.3272      0.177      1.846      0.098      -0.074       0.728
==============================================================================
Omnibus:                        7.947   Durbin-Watson:                   2.315
Prob(Omnibus):                  0.019   Jarque-Bera (JB):                3.401
Skew:                          -1.185   Prob(JB):                        0.183
Kurtosis:                       4.342   Cond. No.                         3.83
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Discussion

Overview

Thus far, this paper has illustrated the technical process of using Cortex to process, manipulate, visualize, and analyze digital phenotyping data. To more clearly illustrate the utility of these functionalities, we will discuss below the significance of the steps outlined in this paper, as well as use cases. Figure 49 highlights a common workflow with Cortex that is applicable to many types of studies and use cases.

The basic features of Cortex can be immediately translated into analysis of digital phenotyping data. For example, Cortex collects time stamps not only of the start and end time of active data activities (eg, a survey) but also of the time of specific item responses for a given survey and of any discrete passive data points. Prior research has used analytical methods such as this to predict self-reported thoughts of harm based on latency [47]

XSL•FO
RenderX

in responding to items assessing such thoughts. Furthermore, epoch time provides foundational support for pivotal features that can be derived from Cortex, such as sleep duration and screen use data, which allow the transformation of passive data into clinically relevant variables. While discrete data points can be informative, even without incorporating time (eg, an individual's home may be approximated for most by the GPS point most frequently visited), the formatting of data points with time stamps allows for more nuanced analysis and deeper insight. By combining basic Cortex features, it is also possible to derive novel digital biomarkers as well as replicate those created by other teams.

The nature of digital phenotyping data, when combined with Cortex, is the flexibility in using derived features or even creating novel digital biomarkers. For example, Cortex-derived features, such as sleep duration, time spent at home, screen time, and data quality, have been demonstrated when used in research applying anomaly detection. Specifically, anomalies within personal baselines of active data and Cortex-derived features of data for individuals with schizophrenia have been used to predict clinical relapse, an example of advanced data processing [48]. Using the methods outlined above in this paper, any digital phenotyping data can be processed in features and DataFrames that will enable easy use of the code [49] for anomaly detection. Related applications may include using the digital phenotyping data features and anomaly detection to detect early warning of risk of self-harm, relapse in bipolar disorder, or medication adherence as 3 broad examples.

The nature of digital phenotyping is also well suited when combined with Cortex for machine learning research. Linear regression models are a common tool for analyzing digital phenotyping data, and the readily accessible Cortex features, particularly those derived from passive data, make Cortex a productive API for analyzing the relationships between digital phenotyping–derived features. In addition, the abundance of time and duration columns in the data set makes Cortex a useful tool to prepare data for time series analysis. Recent research has leveraged Cortex to pull data, split it using a windowing approach, and train a variety of Recurrent Neural Network models using a diverse architecture including Long Short-Term Memory. Cortex's unique strengths in overtime data allow researchers to use the library for a variety of predictive tasks ranging from prognosis prediction to quantitative score prediction.

The utility features such as visualizations for monitoring data quality and providing data back to users have been illustrated in our Beth Israel Deaconess Medical Center–based clinic using mindLAMP [50]. As illustrated above, patients can be offered heat maps of passive data volume generated daily (see the *Checking Engagement and Data Quality* section) in addition to visualizations of active data. With staff allocated to assess patient data quality multiple times per week, they can reach out to patients about their data quality, with specific strategies for ameliorating poor data quality [6]. For example, low passive data quality across sensors may be a result of low weekly app use, as many devices limit app data collection based on use. However, poor GPS quality with high accelerometer quality may simply indicate a misconfiguration of phone settings. The root cause is easy to detect with data visualization. Similarly, patients provided with visualizations of their own active data overlaid with passive data streams can gain insight by observing and even interacting with their own data, as shown in Figure 50.

Figure 49. A schematic of a common deployment of mindLAMP, hosted by the Beth Israel Deaconess Medical Center team and Cortex. API: application programming interface. HIPAA: Health Insurance Portability and Accountability Act.

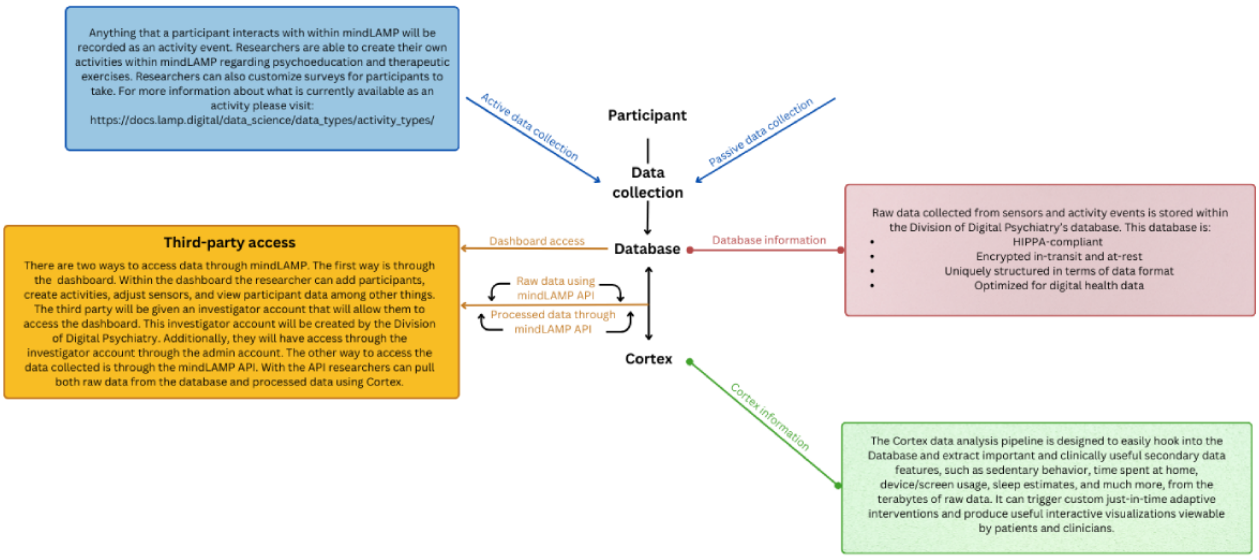

XSL•FO
RenderX

Figure 50. An example of test account data used to illustrate how 1 passive data feature screen time can be visualized in light of different symptoms and functioning metrics.

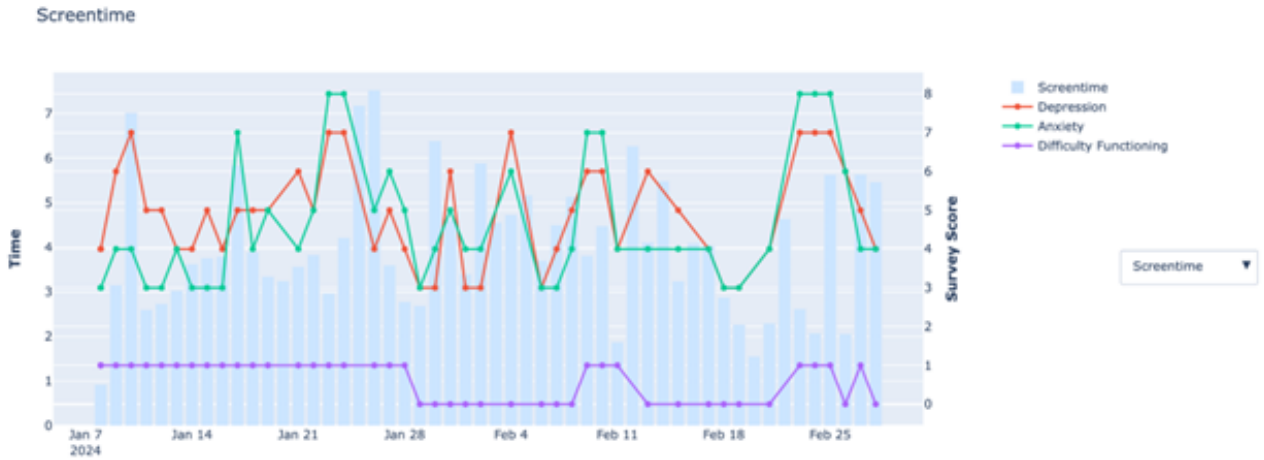

Conclusions

Digital phenotyping presents researchers, clinicians, and patients with a wealth of new data. Packages such as Cortex ensure that these data and resulting insights are accessible to all. By reviewing how to pull active and passive data, check the participants' data quality, manipulate DataFrames, and create basic data visualizations, this paper serves as an interactive guide for advancing reproducible and replicable digital phenotyping research.

Similar to many other tools, Cortex has its limits. While Cortex has been optimized for mindLAMP, it will still work on any data set that meets the formatting requirements outlined in the *Review of Digital Phenotyping Data Types* section. Cortex will attempt to provide meaningful insights and features derived from low data quality, but the quality of such features is questionable. Fortunately, low data quality can be prevented with regular checkups with participants using Cortex for quality checks, as discussed in the *Checking Engagement and Data Quality* section. Clinical judgment is always necessary when interpreting any output, and at the time of this writing, for psychiatry, there are no digital biomarkers recognized by the Food and Drug Administration or other regulatory bodies.

Cortex is an open-source package. As such, it is always evolving, both internally within the team and externally from users worldwide. Due to this evolving nature, this paper is not meant to be rigid but more of a living document to be updated periodically. Updates are posted on the web [51].

Acknowledgments

We would like to thank Jenn Sabbagh, MBA, for her ongoing support of digital psychiatry research over the years and advocacy for this work through her research administration role.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

JT is the editor-in-chief of *JMIR Mental Health* at the time of this publication.

References

1. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 06;2:88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
2. Montag C, Quintana DS. Digital phenotyping in molecular psychiatry-a missed opportunity? Mol Psychiatry 2023 Jan;28(1):6-9 [FREE Full text] [doi: 10.1038/s41380-022-01795-1] [Medline: 36171355]
3. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: a systematic literature review. J Biomed Inform 2023 Feb;138:104278 [FREE Full text] [doi: 10.1016/j.jbi.2022.104278] [Medline: 36586498]
4. Ebner-Priemer U, Santangelo P. Digital phenotyping: hype or hope? Lancet Psychiatry 2020 Apr;7(4):297-299. [doi: 10.1016/S2215-0366(19)30380-3] [Medline: 31635970]
5. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan;46(1):45-54 [FREE Full text] [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]

XSL•FO
RenderX

6. Currey D, Torous J. Increasing the value of digital phenotyping through reducing missingness: a retrospective review and analysis of prior studies. BMJ Ment Health 2023 Feb;26(1):e300718 [FREE Full text] [doi: 10.1136/bmjment-2023-300718] [Medline: 37197799]
7. Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. J Med Internet Res 2016 Mar 29;18(3):e72 [FREE Full text] [doi: 10.2196/jmir.5505] [Medline: 27025287]
8. Shen FX, Silverman BC, Monette P, Kimble S, Rauch SL, Baker JT. An ethics checklist for digital health research in psychiatry: viewpoint. J Med Internet Res 2022 Feb 09;24(2):e31146 [FREE Full text] [doi: 10.2196/31146] [Medline: 35138261]
9. Shen FX, Baum ML, Martinez-Martin N, Miner AS, Abraham M, Brownstein CA, et al. Returning individual research results from digital phenotyping in psychiatry. Am J Bioeth 2024 Feb;24(2):69-90 [FREE Full text] [doi: 10.1080/15265161.2023.2180109] [Medline: 37155651]
10. Fuhrmann LM, Weisel KK, Harrer M, Kulke JK, Baumeister H, Cuijpers P, et al. Additive effects of adjunctive app-based interventions for mental disorders - a systematic review and meta-analysis of randomised controlled trials. Internet Interv 2023 Dec 18;35:100703 [FREE Full text] [doi: 10.1016/j.invent.2023.100703] [Medline: 38225971]
11. Polhemus A, Novak J, Majid S, Simblett S, Morris D, Bruce S, et al. Data visualization for chronic neurological and mental health condition self-management: systematic review of user perspectives. JMIR Ment Health 2022 Apr 28;9(4):e25249 [FREE Full text] [doi: 10.2196/25249] [Medline: 35482368]
12. carissalow / rapids. GitHub. URL: https://github.com/carissalow/rapids [accessed 2024-07-25]
13. onnela-lab / forest. GitHub. URL: https://github.com/onnela-lab/forest [accessed 2024-07-25]
14. BIDMCDigitalPsychiatry / LAMP-cortex. GitHub. URL: https://github.com/BIDMCDigitalPsychiatry/LAMP-cortex [accessed 2024-07-25]
15. Breitinger S, Gardea-Resendez M, Langholm C, Xiong A, Laivell J, Stoppel C, et al. Digital phenotyping for mood disorders: methodology-oriented pilot feasibility study. J Med Internet Res 2023 Dec 29;25:e47006 [FREE Full text] [doi: 10.2196/47006] [Medline: 38157233]
16. Gansner M, Nisenson M, Carson N, Torous J. A pilot study using ecological momentary assessment via smartphone application to identify adolescent problematic internet use. Psychiatry Res 2020 Nov;293:113428. [doi: 10.1016/j.psychres.2020.113428] [Medline: 32889344]
17. Bilden R, Torous J. Global collaboration around digital mental health: the LAMP consortium. J Technol Behav Sci 2022;7(2):227-233 [FREE Full text] [doi: 10.1007/s41347-022-00240-y] [Medline: 35071742]
18. Rodriguez-Villa E, Rozatkar AR, Kumar M, Patel V, Bondre A, Naik SS, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021 Aug 24;8:e30 [FREE Full text] [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
19. Security and privacy policy. LAMP Platform. URL: https://docs.lamp.digital/privacy/ [accessed 2024-06-17]
20. Call duration. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/call_duration [accessed 2024-06-17]
21. Call number. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/call_number [accessed 2024-06-17]
22. Data quality. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/data_quality [accessed 2024-06-17]
23. Entropy. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/entropy [accessed 2024-06-17]
24. Game results. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/game_results [accessed 2024-06-17]
25. HealthKit sleep duration. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/healthkit_sleep_duration [accessed 2024-06-17]
26. Hometime. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/hometime [accessed 2024-06-17]
27. Inactive duration. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/inactive_duration [accessed 2024-06-17]
28. Nearby device count. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/nearby_device_count [accessed 2024-06-17]
29. Screen duration. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/screen_duration [accessed 2024-06-17]
30. Step count. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/step_count [accessed 2024-06-17]
31. Survey results. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/survey_results [accessed 2024-06-17]

XSL•FO
RenderX

32. Trip distance. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/trip_distance [accessed 2024-06-17]
33. Trip duration. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/trip_duration [accessed 2024-06-17]
34. Call degree. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/features/secondary/call_degree [accessed 2024-06-17]
35. Epoch and Unix timestamp conversion tools. Epoch Coverter. URL: https://www.epochconverter.com/ [accessed 2024-07-25]
36. About pandas. Pandas. URL: https://pandas.pydata.org/about/ [accessed 2024-07-25]
37. What is NumPy? NumPy. URL: https://numpy.org/doc/stable/user/whatisnumpy.html [accessed 2024-07-25]
38. datetime — basic date and time types. Python. URL: https://docs.python.org/3/library/datetime.html [accessed 2024-07-25]
39. pytz 2024.1. Python. URL: https://pypi.org/project/pytz/ [accessed 2024-07-25]
40. seaborn: statistical data visualization. Seaborn. URL: https://seaborn.pydata.org/ [accessed 2024-07-25]
41. matplotlib 3.9.1. Python. URL: https://pypi.org/project/matplotlib/ [accessed 2024-07-25]
42. calplot 0.1.7.5. Python. URL: https://pypi.org/project/calplot/ [accessed 2024-07-25]
43. Vega-Altair: declarative visualization in Python. Vega-Altair. URL: https://altair-viz.github.io/ [accessed 2024-07-25]
44. Plotly open source graphing library for Python. Plotly. URL: https://plotly.com/python/ [accessed 2024-07-25]
45. Participant-level visualizations. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/visualizations/participant_level [accessed 2024-06-17]
46. Running cortex. LAMP Platform. URL: https://docs.lamp.digital/data_science/cortex/running_cortex [accessed 2024-06-17]
47. Henson P, Torous J. Feasibility and correlations of smartphone meta-data toward dynamic understanding of depression and suicide risk in schizophrenia. Int J Methods Psychiatr Res 2020 Jun;29(2):e1825 [FREE Full text] [doi: 10.1002/mpr.1825] [Medline: 32333471]
48. Cohen A, Naslund JA, Chang S, Nagendra S, Bhan A, Rozatkar A, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. Schizophrenia (Heidelb) 2023 Jan 27;9(1):6 [FREE Full text] [doi: 10.1038/s41537-023-00332-5] [Medline: 36707524]
49. Anomaly detection in R. Division of Digital Psychiatry. URL: https://www.notion.so/digitalpsychiatry/Anomaly-Detection-in-R-177a40b5120343fdad1bff6db7632118 [accessed 2024-06-17]
50. Macrynikola N, Nguyen N, Lane E, Yen S, Torous J. The digital clinic: an innovative mental health care delivery model utilizing hybrid synchronous and asynchronous treatment. NEJM Catal Innov Care Deliv 2023 Aug 16;4(9). [doi: 10.1056/cat.23.0100]
51. LAMP Platform homepage. LAMP Platform. URL: https://docs.lamp.digital/ [accessed 2024-07-26]

Abbreviations

API: application programming interface
UTC: Coordinated Universal Time

Edited by A Mavragani; submitted 17.03.24; peer-reviewed by J Mistry, D Ghosh; comments to author 17.04.24; revised version received 22.04.24; accepted 19.06.24; published 23.08.24.

Please cite as:
Burns J, Chen K, Flathers M, Currey D, Macrynikola N, Vaidyam A, Langholm C, Barnett I, Byun A(S, Lane E, Torous J Transforming Digital Phenotyping Raw Data Into Actionable Biomarkers, Quality Metrics, and Data Visualizations Using Cortex Software Package: Tutorial
J Med Internet Res 2024;26:e58502
URL: https://www.jmir.org/2024/1/e58502
doi:10.2196/58502
PMID:39178032

©James Burns, Kelly Chen, Matthew Flathers, Danielle Currey, Natalia Macrynikola, Aditya Vaidyam, Carsten Langholm, Ian Barnett, Andrew (Jin Soo) Byun, Erlend Lane, John Torous. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Adapting a Telehealth Physical Activity and Diet Intervention to a Co-Designed Website for Self-Management After Stroke: Tutorial

Dina Pogrebnoy[1,2], BA, GDip; Lee Ashton[3], PhD; Brian A Beh[4]; Meredith Burke[5,6]; Richard Cullen[7], BA; Jude Czerenkowski[7], BA, MSW; Julie Davey[8], BA, BSW; Amy M Dennett[9,10], PhD; Kevin English[11]; Erin Godecke[12], PhD; Nicole Harper[5]; Elizabeth Lynch[13], PhD; Lesley MacDonald-Wicks[1,3], PhD; Amanda Patterson[1,3], PhD; Emily Ramage[1,14,15], PhD; Ben Schelfhaut[16]; Dawn B Simpson[1], PhD; Karly Zacharia[1,3], PhD; Coralie English[1,6,17], PhD

[1]School of Health Sciences, University of Newcastle, Newcastle, Australia
[2]Department of Physiotherapy, Western Health, St Albans, Australia
[3]Food and Nutrition Program, Hunter Medical Research Institute, Newcastle, Australia
[4]Consumer Advisory Group Member, Sydney, Australia
[5]Consumer Advisory Group Member, Newcastle, Australia
[6]Heart and Stroke Program, Hunter Medical Research Institute, Newcastle, Australia
[7]Stroke Foundation, Melbourne, Australia
[8]Consumer Advisory Group Member, Wangaratta, Australia
[9]Allied Health Clinical Research Office, Eastern Health, Melbourne, Australia
[10]School of Allied Health, Human Services and Sport, Latrobe University, Melbourne, Australia
[11]Consumer Advisory Group Member, Melbourne, Australia
[12]Department of Speech Pathology, School of Medical and Health Sciences, Edith Cowan University, Sir Charles Gairdne Park Health Care Group, Perth, Australia
[13]Caring Futures Institute, College of Nursing and Health Sciences, Flinders University, Adelaide, Australia
[14]ASPIRE Unit, Western Health, Melbourne, Australia
[15]Institute of Neurosciences and Mental Health, Florey, Melburne, Australia
[16]Consumer Advisory Group Member, Kent, United Kingdom
[17]Centre of Research Excellence to Accelerate Stroke Trial Innovation and Translation, University of Sydney, Sydney, Australia

Corresponding Author:
Dina Pogrebnoy, BA, GDip
Department of Physiotherapy
Western Health
176 Furlong Rd
St Albans, 3021
Australia
Phone: 61 0403432415
Email: dina.pogrebnoy@uon.edu.au

Abstract

People who experience a stroke are at a higher risk of recurrent stroke when compared with people who have not had a stroke. Addressing modifiable risk factors like physical inactivity and poor diet has been shown to improve blood pressure, a leading contributor to stroke. However, survivors of stroke often experience challenges with accessing risk reduction services including long wait lists, difficulty with transportation, fatigue, impaired function, and diminished exercise capacity. Providing health interventions via a website can extend the reach when compared with programs that are only offered face to face or via real-time telehealth. Given global challenges of accessing secondary prevention programs, it is important to consider alternative ways that this information can be made available to survivors of stroke worldwide. Using the “design thinking” framework and drawing on principles of the integrated knowledge translation approach, we adapted 2 co-designed telehealth programs called i-REBOUND – *Let’s get moving* (physical activity intervention) and i-REBOUND – *Eat for health* (diet Intervention) to create the i-REBOUND *after stroke* website. The aim of this paper is to describe the systematic process undertaken to adapt resources from the telehealth delivered i-REBOUND – *Let’s get moving* and i-REBOUND – *Eat for health* programs to a website prototype with a focus on navigation requirements and accessibility for survivors of stroke. We engaged a variety of key stakeholders with diverse skills

XSL•FO
RenderX

and expertise in areas of stroke recovery, research, and digital health. We established a governance structure, formed a consumer advisory group, appointed a diverse project team, and agreed on scope of the project. Our process of adaptation had the following 3 phases: (1) understand, (2) explore, (3) materialize. Our approach considered the survivor of stroke at the center of all decisions, which helped establish guiding principles related to our prototype design. Careful and iterative engagement with survivors of stroke together with the application of design thinking principles allowed us to establish the functional requirements for our website prototype. Through user testing, we were able to confirm the technical requirements needed to build an accessible and easy-to-navigate website catering to the unique needs of survivors of stroke. We describe the process of adapting existing content and co-creating new digital content in partnership with, and featuring, people who have lived experience of stroke. In this paper, we provide a road map for the steps taken to adapt resources from 2 telehealth-delivered programs to a website format that meets specific navigation and accessibility needs of survivors of stroke.

(J Med Internet Res 2024;26:e58419) doi:10.2196/58419

KEYWORDS

stroke; secondary prevention; co-design; how-to guide, website development; accessibility; navigation; self-management

Introduction

Background

A stroke happens when the blood supply to the brain is interrupted, causing oxygen deprivation and tissue death [1]. Stroke is the second leading cause of death worldwide [2]. With 1 in 4 people experiencing a stroke in their lifetime [3], it is estimated that, by the year 2030, there will be 12 million stroke-related deaths and an estimated 70 million stroke survivors globally [4]. In Australia, there are currently close to half a million people living with the effects of stroke [5], and without action, this number is predicted to exceed 800,000 by 2050 [6]. Survivors of stroke have an increased risk of stroke recurrence, with an estimated 40% of stroke survivors going on to have another stroke within 10 years [7].

Hypertension is one of the biggest risk factors for first stroke [8] and, together with physical activity and diet quality, is one of the most modifiable risk factors for recurrent stroke [8]. Adequate levels of physical activity improve blood pressure control [9], and a reduction of 10 mm Hg in systolic blood pressure has been associated with as much as a 40% reduction in stroke risk [10]. There are no trials of dietary interventions for secondary stroke prevention, but the Mediterranean-style diet holds the most promise [11]. Therefore, together with good medical management, lifestyle interventions that improve dietary behavior and physical activity levels can help reduce the risk of recurrent stroke [12,13]. Despite this, however, few programs exist to support survivors of stroke to achieve and maintain these beneficial lifestyle changes after their stroke [14].

Many survivors of stroke have difficulty with equitable access to follow-up services once discharged from hospital [15]. Common barriers include transportation [16,17], access [16], economic constraints [16], and a lack of suitable programs [16]. An additional challenge facing survivors of stroke is their diminished exercise capacity, with increased oxygen requirements during routine tasks making it hard to engage in secondary prevention programs [18]. Similarly, problems with memory or thinking as well as physical limitations, fatigue, and planning challenges can further impact a stroke survivor's ability to participate in meaningful activity after stroke [18].

Many survivors of stroke will have lifelong effects from their stroke; therefore, resources to improve well-being after stroke need to cater for a variety of abilities and be relevant for survivors at all stages of their recovery. Unlike other chronic health conditions (eg, cardiac rehabilitation), there are no structured self-management programs beyond the initial rehabilitation phase after an acute stroke event. International clinical guidelines recommend that all survivors of stroke should receive support for self-management after their stroke [19]. A recently published Cochrane review provides recommendations on the best ways to develop effective self-management programs for survivors of stroke [19], by emphasizing that programs need to be meaningful and acceptable to people with stroke and that survivors need to be involved in program design [20].

Previous Work

Our team developed 2 co-designed telehealth-delivered interventions to (1) increase physical activity [21] and (2) improve diet quality [22] in people after stroke. These interventions were pilot-tested in a 4-arm randomized control trial and demonstrated a clinically relevant difference in support of the interventions across a range of physical activity, diet quality, and blood pressure outcomes [23]. Building on this work, our research partners with lived experience of stroke felt that adapting our telehealth resources to be publicly available on a website was an important next step to reach more people impacted by stroke.

Websites have been used in recent years to improve accessibility to evidence-based resources for people living with complex health conditions [24,25]. Our recent review found that people living with chronic health conditions could improve their physical activity levels and aspects of diet quality if they accessed websites hosting health information [26]. Despite the apparent utility and effectiveness of websites, to the best of our knowledge, prior to the launch of the i-REBOUND *after stroke* website, there were no publicly available websites aimed at secondary stroke prevention. We hypothesized that a key barrier to creating meaningful websites for survivors of stroke is the absence of published guiding principles or functional requirements to do this well. We hoped that publishing the guiding principles used in our project will help others to develop accessible websites for survivors of stroke.

XSL•FO
RenderX

A key guiding principle of our work, as identified by our research partners with lived experience, was that the i-REBOUND *after stroke* website needed to feature survivors of stroke in the content. It was important for survivors of stroke to be able to relate to the website content, feeling that they "could only be what they could see." Therefore, we created a fit-for-purpose website designed for and featuring content with survivors of stroke to address this gap.

Objective

The purpose of this paper was to describe the systematic approach undertaken to adapt existing co-designed telehealth resources for a website prototype that meets the unique needs and priorities of people living with stroke. In this paper, we provide a road map for clinicians and researchers wanting to adapt nondigital interventions from their original form into digital-driven solutions for survivors of stroke and other similar health conditions.

Methods

Overview

We used the integrated knowledge translation (iKT) approach to adapt our telehealth-delivered interventions and accompanying resources into a website. The iKT approach is a coproduction approach in which "knowledge users" are included in the research team in shared partnership [27-29]. Knowledge users are described as people who administer, influence, or use the research [28,30]. Knowledge users can include health care workers, policymakers who implement research findings, and people with lived experience of the health condition of interest. This project was a partnership between the Stroke Foundation and the University of Newcastle in Australia.

Ethical Considerations

The project commenced in April 2021 (ethical approval 2021/ETH00360 from the Hunter New England research ethics committee) and was completed in November 2022, when the i-REBOUND *after stroke* website was launched to the public. Informed consent was obtained from all research participants, and all data were deidentified prior to analysis.

Project Plan

Our project plan included the following 3 distinct phases, which are consistent with a design thinking framework: (1) understand, (2) explore, and (3) materialize [31]. These areas of focus are summarized in the project road map presented in Figure 1 and described in further detail in the following sections.

XSL•FO
RenderX

Figure 1. Project road map. *Findings from this phase of the project will be reported in a future paper.

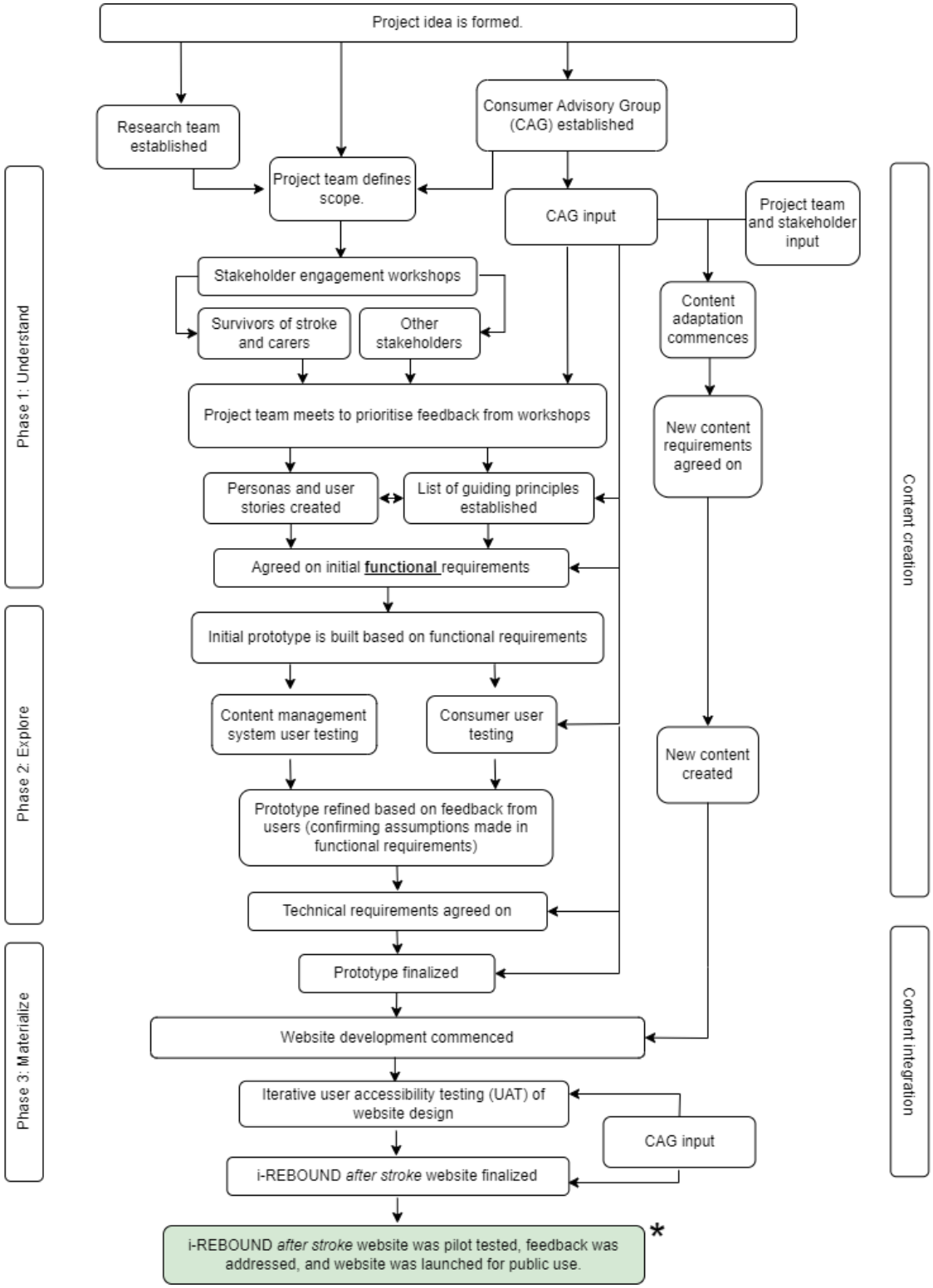

Consumer Advisory Group

A Consumer Advisory Group (CAG) was convened to guide all project aspects from inception to completion. Members for the CAG were chosen to ensure representation of a range of views, experiences, and elements of lived experience. The CAG for the project included 6 people with lived experience of stroke (MB, JD, BS, BAB, KE, NH). The CAG created and agreed to a "Terms of Reference" for their engagement for the duration of this project. A copy of the "Terms of Reference" for the i-REBOUND *after stroke* website CAG can be seen in Multimedia Appendix 1. The CAG met quarterly to discuss project progress. Research leads for the project were present at the beginning of each CAG meeting and would then leave the meeting to allow for free sharing of opinions without researchers present. Cochairs of the CAG (JD and MB) were invited to all project and steering committee meetings to represent views of the CAG and to provide input from the lived experience

XSL•FO
RenderX

perspective at each stage of project development. All CAG members were consulted on their preferences for ways in which they would like to offer their input to meet their communication needs.

Members of the CAG felt that the key ingredient to a successful project was to establish and maintain an authentic partnership and spirit of collaboration between members of the CAG and the broader project team. Members of the CAG identified 5 key "tips" that helped establish and maintain an authentic collaborative partnership including (1) working together from the start, (2) respecting the process as much as the outcome, (3) recognizing survivors of stroke as whole human beings, (4) being authentic and willing to learn, and (5) adopting the concept of forming an alliance with survivors of stroke. A detailed explanation of these tips can be found in Multimedia Appendix 2.

Governance Structure

A steering committee of senior stakeholders was established to provide overall governance of the project with key deliverables being (1) progress and (2) monitoring for risks. Members included the principal investigator (CE), project lead (DP), national manager for stroke services (JC), and digital services (RC) representing the Stroke Foundation of Australia and CAG cochairs (JD and MB). The steering committee was responsible for developing and endorsing a collaboration agreement between the University of Newcastle and the Stroke Foundation and planning for sustainability of the website following public launch.

The Project Team

Members of the project team included clinical experts in delivering physical activity and diet interventions (ER, KZ, LMW, AP, LA, AD, DP, EL, CE) as well as clinical experts in cognitive and communication difficulties (EG) for people after stroke. Other project team members included experts in co-design methodology (ER, KZ, CE), website designers (RC and LA), policymakers (JC), as well as implementation and research translation experts (EL). Most importantly, 2 research partners with lived experience of stroke (JD and MB) were members of the research team and had an equal voice at the decision-making table. The team agreed to respect everyone's views and use a common language and strove to empower all members to share their knowledge and experience freely.

Project Outline

Design Thinking Methodology

We used principles from the "design thinking" methodology throughout the project as outlined in Figure 2 and explained in more detail in the following sections.

Design thinking is a methodology used to tackle complex end user "problems" in a person-centric way [32]. The design thinking methodology was chosen for this project as it allowed us to identify user requirements specific to survivors of stroke.

Figure 2. Design thinking process [31].

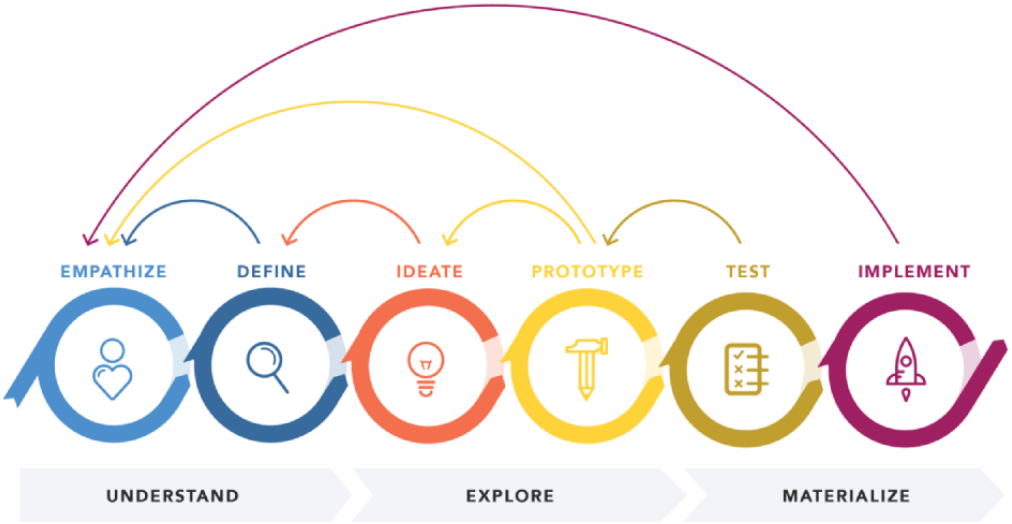

Phase 1: Understand

The first key stage of phase 1, and in keeping with the iKT approach, was to engage with intended users of the i-REBOUND *after stroke* website. The "understand" phase helped to establish the functional requirements for survivors of stroke using websites with health information.

Stakeholder Engagement

We held workshops separately with survivors of stroke (and carers) and with other stakeholders including allied health clinicians, medical professionals, policymakers, and health care managers. We included people living with aphasia who were supported by a member of the research team who was a speech pathologist with 25 years' experience in language disorders after stroke (EG). Workshops were facilitated by a member of the research team (DP and CE or AD) and recorded for transcribing purposes. One member of the research team facilitated the workshop with a second member of the research team there to take note of factors that might aid analysis such as passionate comments, body language, or nonverbal activity.

XSL•FO
RenderX

Workshops—Survivors of Stroke and Carers

The main objectives of the workshop(s) with stroke survivors and carers were to elicit perceived *barriers and facilitators* to use of websites targeting physical activity and diet quality for people with stroke and key *features* for survivors of stroke that help promote engagement and improve usability and uptake when interacting with websites.

Workshop participants were provided with an outline of the workshop (Multimedia Appendix 3) by email 1 week prior to the workshop. The outline allowed participants to formulate their thoughts on the topic and to encourage active participation in facilitated discussion at the time of the workshop. Workshops were held in small groups through an online meeting platform with breaks offered as needed. A welcome video, recorded by a CAG member (Multimedia Appendix 4) with aphasia, was played at the commencement of each workshop to encourage all participants to share their views openly. We asked participants to reflect on websites they currently use and what features help them enjoy their experience and make them want to return to the website on a future occasion. Participants were encouraged to share any difficulties they experience when using websites to help optimize the user experience of our website prototype. Participants were encouraged to comment on experiences shared by others (in relation to themselves) so that recurring themes could be captured. After the workshops, we emailed participants a written summary of discussed themes for cross-checking and invited feedback or any additional comments.

Workshops—"Other Stakeholders"

Stakeholder workshops were held to capture views of health professionals working in the field of stroke rehabilitation. The same process was followed with the "other stakeholders" as with survivors of stroke prior to the workshops. A copy of the workshop outline in relation to "other stakeholders" can be seen in Multimedia Appendix 5, and the welcome video can be viewed in Multimedia Appendix 6. We asked clinicians if they currently provided secondary prevention advice to their patients and in what format. Attendees discussed key considerations that they felt would help clinicians support survivors of stroke to engage with websites aimed at secondary stroke prevention. After the workshops, participants were sent a summary of discussed themes for cross-checking and were invited to provide feedback or any additional comments.

A total of 22 diverse key stakeholders (9 survivors of stroke, 1 carer, and 12 health care professionals representing dietetics, physiotherapy, speech pathology, occupational therapy, and neuropsychology, as well as medical rehabilitation specialists) contributed their views to initial workshops. A written summary of all stakeholder workshops was provided to all members of the CAG via email. The CAG met to review the content together and agree on priority focus areas to guide the project team in creating meaningful user stories and personas [33].

User Stories and Personas

User stories and personas are integral components of the design thinking process, aiding in the creation of more effective and user-friendly interfaces. User stories are succinct narratives describing specific user interactions with a website, structured as:

"As a *[stroke survivor]*, I want *[an action]* so that *[benefit or goal]*" [34].

User stories and user acceptance criteria for the i-REBOUND *after stroke* project were developed after stakeholder engagement workshops and are described in Multimedia Appendix 7.

Personas help website designers understand the target user and therefore humanize design decisions [35]. Personas can be useful to pick up on areas that may not have been identified in other discovery activities. Testing assumptions as part of the iterative user testing process ensured that the persona activity in our project was robust and allowed for the checking of both internal and external hypotheses. Undertaking this process ensured a higher degree of user acceptance and user-centered product design. By generating a list of meaningful personas (Multimedia Appendix 8) from the priority focus areas identified by members of the CAG, the design team could be confident that the i-REBOUND *after stroke* website would resonate with a diverse group of users [36].

Guiding Principles

Once we had a range of personas specific to this project, we could create a list of "guiding principles" underpinning the design and development of our website prototype. The guiding principles (Table 1) included the broad categories of design, navigation, access, function, community, language, and content display with a clear distinction between what the website design "Must be" and what it "Won't be" from a user perspective. Members of the CAG reviewed and endorsed the list of guiding principles for the i-REBOUND *after stroke* website.

XSL•FO
RenderX

Table 1. Guiding principles as identified by participants in stakeholder workshops.

Guiding principle	Must be	Won't be
General	• Ability-specific • Short instructions • Short videos • Research supported	• One size fits all • Long instructions • Long videos • Not trusted
Design	• Minimal links • Clean design • Not a lot of steps to follow • Everyday people doing everyday things	• Lots of links • Colorful and busy • Lots of steps • Unrelatable
Navigation	• Clear to understand • Intuitive	• Confusing • Complex
Access	• Quick to join • Accessible on multiple devices • Optimized for small screens	• Too long = fatigue • Not optimized • Not optimized
Function	• Personalizable • Goal-focused	• Unguided • Without a purpose/aim
Community	• Feeling connected • Encouragement	• Alone • Disengaged
Language	• Simple • Short and direct • Respectful • Motivational • Engaging • Fun • Other languages	• Complex • Long and ambiguous • Patronizing • Discouraging • Won't persist • Boring/dull • Just English
Content displayed	• Big font size • Optimized for print • Ad free • Practical • Free to use • Only show information I need • Easy to interpret content • Sharable • Stroke-related content • Multimedia (images, video, and audio)	• Small font • Not print ready • Ads or promotions • Not relatable/usable • Cost • Oversupply of information • Difficult to follow • Can't share • Generic information • Just text

Functional Requirements

Using the guiding principles, our development team was able to derive a comprehensive and user-focused list of functional requirements. Functional requirements outlined features that the i-REBOUND *after stroke* website must perform to meet the needs of our target users. These requirements specifically described what the system "should do" and how it "should behave" [37]. Our generated list of functional requirements served as the basis for designing, developing, and testing the initial i-REBOUND *after stroke* website prototype. A full list of functional requirements can be found in Multimedia Appendix 9.

Phase 2: Explore

Prototype Development

We used an iterative approach to develop the prototype to uncover design flaws, enable a process to refine features, and offer an opportunity to align the final product with user needs [38].

The initial prototype of the i-REBOUND *after stroke* website was presented to the CAG by members of the research and design team (RC, DP, CE) to seek first impressions. The CAG was then left to meet without members of the research and design team to discuss their impressions freely. The CAG provided a summary of recommendations (Multimedia Appendix 10) to the broader research team based on their review and discussion. Positive feedback was that sections were generally set up well and the "comments" section at the bottom of each page would encourage user engagement. Suggestions for improvement included creating an introduction video for each website landing page to set the scene for the purpose of each section. Another suggestion was to remove the "time I have" to cook button from each recipe and replace it with the number of steps for the recipe, as the time a recipe will take to cook is likely to vary between users with different abilities.

XSL•FO
RenderX

User Testing

Once the initial website prototype was ready, we proceeded to test it with a broader range of users. User testing during prototype development is a crucial step to validate design decisions and identify usability issues [39]. User testing was an important step in helping to uncover user preferences, behavior patterns, and potential pain points. To understand user experience, in relation to navigation of the i-REBOUND *after stroke* website, a member of the research team (DP) conducted 1:1 user testing sessions. User testing was completed with stroke survivors (5 members of the CAG and 2 other survivors of stroke) to help identify further development opportunities for the i-REBOUND *after stroke* website.

User testing sessions (the session guide is in Multimedia Appendix 11) were conducted using an online meeting platform and were recorded (audio and video) with consent. At the beginning of each testing session, users shared their screen and were asked to share their feelings when browsing the site, both positive and negative, to enable further analysis from the development team. During the sessions, users were asked to complete several tasks to see how easy the site was to use. If something did not feel right, users were asked to share this. Similarly, if something worked well, users were encouraged to share what was working well for them. Duration of the user testing sessions ranged from 20 minutes to 30 minutes, and breaks were offered as needed. After all user testing sessions were completed, recordings of the sessions were shared with a member of the design team. These recordings were analyzed for subtle navigation challenges that may not have been articulated through verbal feedback (eg, when it took a user an extended period to navigate somewhere but they did not say anything to indicate this was an issue).

Content Management System User Testing

Having a user-friendly content management system (CMS) for content managers is important for maintaining website content. CMS platforms allow users to create, edit, organize, and publish content such as text, images, videos, and documents on websites [40]. CMS user testing ensured that the content managers and administrators of the i-REBOUND *after stroke* website were able to perform key functions in the day-to-day management of the website. User testing a CMS involves evaluating the usability, functionality, and user experience of the system from the perspective of its intended user. Four members of the Stroke Foundation design and development team tested the CMS and provided feedback to website developers on areas that needed to be addressed to improve their experience.

Refining the Prototype

Analyzing user testing feedback, identifying technical limitations, prioritizing opportunities for enhancement, and mapping enhancements to technical requirements were important steps that needed to occur before website development could commence. Technical requirements refer to the specific capabilities, features, and resources that are necessary for a technology solution to function properly and effectively [41]. This phase of the project was completed in an iterative manner with extensive input from the CAG to agree on the minimally viable product. A minimally viable product in the context of building a website [42] needs to capture essential features and functionality to address the core needs of the target audience taking into consideration project budget and agreed timelines. CAG members guided decision-making around prioritization of enhancements as the initial prototype was refined and finalized [43]. A wish list for future enhancement opportunities was created to reflect that not all functional requirements could be addressed with the limited funding for this project.

Content Creation

In parallel to the process of developing the website prototype, our team worked with members of the CAG and a broader group of stroke survivors to adapt existing content from the i-REBOUND – *Let's get moving* and i-REBOUND – *Eat for health* telehealth programs to be accessible in a digital format. An experienced physiotherapist (DP) was the lead for this phase of the project and worked closely with members of the project team to adapt existing content and create new content for the i-REBOUND *after stroke* website. The i-REBOUND *after stroke* website has 3 sections: Eat Well, Move More, and Hints and Hacks. Members of the CAG felt it was important to convey to website users that all resources on the i-REBOUND *after stroke* website were co-designed with survivors of stroke and were evidence-based to increase user confidence in the credibility of the information provided. To address this, an introductory video was created with project team members and clinical experts in eating well (LMW and KZ) and moving more (CE and ER) after stroke. A member of the CAG with lived experience of stroke (JD) introduced the Hints and Hacks section to connect with users who would most benefit from this content.

Content Adaptation

For the adaptation of i-REBOUND – *Eat for health* content, we used written documents created for the i- REBOUND – *Eat for health* program and adapted them to a digital format in the *Practical tips for eating well after stroke* section of the i-REBOUND *after stroke* website. Key considerations were font size and spacing, clarity of images, and keeping the language simple and consistent. Adapting existing recipes from the i-REBOUND – *Eat for health* recipe book to a digital version involved transferring all written content into templates. Reviewing the number of steps in each recipe was important to make sure all instructions were simple and aligned with feedback received in stakeholder engagement workshops. Users wanted each instruction to include an image; therefore, a key part of adapting each recipe to be accessible and user-focused was to create step-by-step images for each recipe. Images needed to be clear and appropriately reflect the written instruction.

For the adaptation of i-REBOUND – *Let's get moving* content, we adapted the "safety instructions" and "when not to exercise" guides from the i-REBOUND – *Let's get moving* telehealth program by making sure they were presented in an easy-to-read digital format with appropriate font size, spacing, and clear presentation. We added advice about consulting a medical professional when exercising after stroke, as users of the i-REBOUND *after stroke* website would not be supervised during exercise, which differed from the supervised telehealth programs of i-REBOUND – *Let's get moving* and i-REBOUND – *Eat for health*. A written explanation and video about the

XSL•FO
RenderX

evidence relating to recommended dose of exercise after stroke demonstrated that content on the i-REBOUND *after stroke* website was evidence-based and met clinical practice guidelines. [44]

New Content

The i-REBOUND *after stroke* Eat Well section contains 24 Mediterranean diet–friendly recipes as well as "Practical tips for eating well after stroke." Members of the CAG and other stroke survivors contributed their tips about ways to simplify cooking after stroke. New content created in this section included guidance on tools when cooking one-handed, gadgets in the kitchen to conserve energy, and tips on cooking when living alone.

New content was required for the i-REBOUND *after stroke* Move More section because the exercise intervention in the telehealth program was delivered via real-time telehealth supervision by an experienced neurological physiotherapist and tailored to each participant. Our stakeholder engagement early in this project highlighted that exercise videos for the i-REBOUND *after stroke* website needed to feature stroke survivors to be relatable. In keeping with the iKT framework, we involved one of our CAG members who is a qualified personal trainer to develop a list of potential exercises (see Multimedia Appendix 12) that would be suitable for a range of abilities after stroke. Exercise choices included seated options, standing supported exercises, as well as more advanced upright options. We incorporated selections for users with an affected upper limb as well as users who used walking aids. A key criterion for each chosen exercise was for it to help the user work at a moderate level of intensity as they moved.

Members of the research team and staff at the Stroke Foundation of Australia assisted with seeking expressions of interest from survivors of stroke to help create content for the i-REBOUND *after stroke* website. We sought to represent a diverse group of stroke survivors including survivors of stroke of varied abilities, age, and the much harder to achieve aspiration of representing cultural diversity. We were able to identify 6 survivors of stroke with varied levels of functional ability and ages (39 years to 78 years old) but found it more difficult to reflect cultural diversity in our content. Therefore, people from culturally and linguistically diverse groups were underrepresented. The process of onboarding a stroke survivor to create these exercise videos had 4 stages. First, an experienced physiotherapist (DP) met with the stroke survivor (+/- carer) and discussed the key objectives of the i-REBOUND *after stroke* website and the type of content we wanted to create. Individual preferences for exercise options were discussed. Second, the physiotherapist developed a proposed exercise program and sent this for review and feedback to the stroke survivor via email. Third, a second face-to-face session was arranged in which the agreed program was practiced, adjustments were made as needed, and the proposed exercise routine was finalized. The final phase of creating the digital exercise content was the day of filming when the practiced routine was recorded at a mutually agreed location. Throughout the process of filming, stroke survivors were encouraged to share their tips for adapting exercises to suit their unique needs. The instructing physiotherapist also verbalized options to make the exercise more or less challenging to appeal to more users.

After filming, the content was edited, and a total of 34 videos showing exercise routines of different lengths (4 minutes to 35 minutes) were created. All recorded content was shared with the stroke survivor (+/- carer) for their review and approval for use prior to integrating it onto the website. Each exercise package presented on the website described the exercise routine and the functional ability for which it is suitable for in written form. Users could listen to this information (by pressing a "listen" button) if that was their preference. Each routine was presented as a full-length video with closed captions. All exercise routines on the website have a step-by-step written guide with still images describing each key component of the exercise routine. The step-by-step guide can be printed for users who prefer to have a paper guide. In addition to structured exercise routines, we filmed a variety of videos in which stroke survivors were engaged in incidental exercise in and around their home. It was important for stroke survivors to showcase these opportunities to demonstrate that any amount of moving is beneficial and there are a variety of ways to increase activity levels after stroke.

Content for the i-REBOUND *after stroke* Hints and Hacks section evolved organically as part of creating new content for the Eat Well and Move More sections. Survivors of stroke openly shared many tips that they said would have been helpful for them in the early stages of their recovery. There are 117 videos featuring a variety of tips in relation to physical activity and eating well after stroke. A key focus in this section is on what survivors of stroke "can do" rather than what they "cannot do." This section of the website has been the most visited section and continues to receive the most positive feedback since the website was launched.

Phase 3: Materialize

Once the prototype was finalized, the technical requirements were handed over to the web design team to "materialize" and build the i-REBOUND *after stroke* website. This phase, like the rest of the project, relied heavily on input from people with lived experience of stroke to complete user accessibility testing (UAT) during the "build" process. This phase was important to make sure functional requirements as identified in phase 1 (understand phase) were appropriately reflected in the design of the i-REBOUND *after stroke* website. The process for UAT in this phase was the same as the process undertaken during prototype user testing in phase 2 (explore phase). Feedback from iterative UAT (2 cycles of improvement) of website navigation and accessibility by all 6 members of the CAG and 2 additional volunteers from the Stroke Foundation allowed further refinement to improve the user experience. The 2 additional volunteers from the Stroke Foundation had not been involved in prior prototype user testing to allow for a fresh perspective. Further refinement as identified through UAT included creating videos for key recommendations of the Mediterranean diet to visually demonstrate what was being recommended. We also added a specific filter for "practical tips for eating well after stroke" to the Eat Well section of the i-REBOUND *after stroke* website to make this information

XSL•FO
RenderX

more accessible and easier to find. The final version of the i-REBOUND *after stroke* website was agreed on by members of the CAG and was then considered ready for pilot testing for feasibility and usability with a broader group of stroke survivors before being launched to the public.

Discussion

Principal Findings

The steps outlined in this paper demonstrate the systematic approach used to create the first evidence-based website co-designed with survivors of stroke targeting the reduction of risk factors associated with recurrent stroke. We appointed a multidisciplinary project team, which included the establishment of a CAG and was further enhanced by engaging a wide key stakeholder group through the life of the project. This paper provides a road map for partnering with consumers and using the design thinking methodology when adapting telehealth resources to be accessible for survivors of stroke in a website format.

Adapting existing evidence-based programs to online formats has the potential to reach more people who would otherwise not have access to this support. Incorporating stakeholder input into both website prototype design and content creation promotes sustained stakeholder engagement [45]. Several methodologies can be used when designing websites including the waterfall method [46], design thinking [32], and agile [46]. Design thinking is the preferred approach when the goal is to create an innovative solution where user experience is at the core of all decisions made [31]. By following the design thinking framework and drawing on the principles of the iKT approach, we were able to successfully adapt existing telehealth resources from the i-REBOUND – *Let's get moving* and i-REBOUND – *Eat for health* programs to create a website that met the unique needs and priorities of people living with stroke.

Endorsement and ongoing support to promote visibility of the i-REBOUND *after stroke* website from the Australian Stroke Foundation have resulted in 89,242 page views from 22,088 unique users in over 90 countries (data from September 25, 2024) since the website was launched in November 2022 [47].

Strengths

A strength of this work is that both the i-REBOUND – *Let's get moving* and i-REBOUND – *Eat for health* programs were already co-designed with survivors of stroke and were effective at promoting behavior change to reduce the risk of recurrent stroke [23]. Content hosted on the i-REBOUND *after stroke* website was co-designed and evidence-based, giving users confidence in the information made available to them. Partnering with consumers throughout this project has been a key strength of this work. Taking the time to understand consumer needs through the discovery phase enabled the creation of a website prototype that met the accessibility and navigation needs of our end users. Cultivating a safe space for survivors of stroke to share freely enabled the creation of authentic and meaningful content, increasing the likelihood that the website will continue to be a well-used resource by survivors of stroke worldwide and ensuring ongoing sustainability.

Another key strength to this body of work is the partnership with Australia's peak body responsible for initiatives to promote recovery after stroke. The i-REBOUND *after stroke* website is hosted on the main Stroke Foundation of Australia website, ensuring sustainable support for content. Links to the i-REBOUND *after stroke* website are routinely shared with newsletter subscribers of the Stroke Foundation of Australia with a clear increase to website traffic when this is done, highlighting that the content on the i-REBOUND *after stroke* website continues to be of interest to people in the stroke community. In addition to this, a member of the CAG (BAB) actively promotes the website through his weekly blogs on a variety of social media platforms, which are well followed by survivors of stroke and clinicians working in stroke recovery.

Limitations

Although we followed a systematic approach by applying principles from the design thinking framework, there are still some limitations to the design of the i-REBOUND *after stroke* website that must be acknowledged. First, it cannot be assumed that an effective intervention via telehealth will be as effective if adapted to a self-directed website with health information. Although it appears that websites are effective for risk factor management for people with chronic illness [26], it remains unclear whether using self-directed websites with survivors of stroke will be as effective as programs like i-REBOUND – *Let's get moving* and i-REBOUND –*Eat for health* in relation to meaningful and long-lasting behavior change [22].

Second, cost is another important limitation to producing a website that meets all user requirements and has relatable content. The relatively small budget (Aus $100,000 [US $68,610.90]) for this project meant that we had to prioritize what could be achieved in the initial phase of the project and create a wish list of future functional enhancements if further funding becomes available. Discussing these priorities with consumer partners and agreeing on a minimal viable product was an important step in making sure we delivered this project on time and on budget. Despite not being able to deliver all identified functions, we believe that the i-REBOUND *after stroke* website is the first of its kind where many key aspects of accessibility and navigation needs, as identified by people living with stroke, have been met. Equally, if additional funds are available, it would be appropriate to create content with greater representation of a culturally diverse group of stroke survivors so that content on the i-REBOUND *after stroke* website can be more relatable.

Finally, the i-REBOUND *after stroke* website, in its current state, is a static information portal where content is not customized to the user. Future enhancements to the i-REBOUND *after stroke* website should focus on the user being able to set goals in relation to being active and eating well after stroke and for website content to be customized to help them achieve these goals.

Key Take Home Messages

Navigation and accessibility should be considered when adapting nondigital interventions to websites with information for survivors of stroke. When designing a website for survivors of

XSL•FO
RenderX

stroke, members of the project team should be diverse and represent multiple levels of influence for a broad perspective. Establishing and maintaining an authentic relationship with consumer partners is critical to a successful project when utilizing principles from the design thinking methodology. The success of any project depends on how well it is set up at the beginning. Establishing a link with a peak governing body that can support the work in the long term will influence the sustainability of initial success. Engaging a wide stakeholder group early in the project enables the establishment of well-informed guiding principles that underpin future decisions. Prototyping and user-testing a website allow the confirmation of assumptions made about technical requirements from functional requirements. User-centered content, where users are not only featured in the content but are also content creators, is wanted by users who access health-related information on websites.

Future Direction

The i-REBOUND *after stroke* website was pilot tested with stroke survivors (n=42) prior to being launched to the public. Pilot testing looked at usability via the System Usability Scale [48] and feasibility (domains of interest were "acceptability," "demand," and "limited efficacy") as per the feasibility framework by Bowen et al [49]. Findings from the pilot study of the i-REBOUND *after stroke* website will be reported in a future paper. The next step in progressing this work would be to test the real-world effectiveness of the i-REBOUND *after stroke* website for patient outcomes.

Conclusion

This paper provides a road map for the adaptation of diet and physical activity telehealth programs to an evidence-based and co-designed website with health information to promote self-management after stroke. We believe that our systematic approach, as outlined in this paper, can be used as an exemplar for any clinicians or researchers wanting to adapt any nondigital resources to a digital interface suitable to be used by survivors of stroke. Further to this, we are confident that the methods we describe in this paper could also be applied to the development of websites with health information for other consumer groups.

Acknowledgments

We acknowledge the generous support of the Nancy and Vic Allen Memorial fund through the Stroke Foundation of Australia for funding the web development of the i-REBOUND *after stroke* website. We thank the Hunter Medical Research Institute for funding the professional filming of website content. We express our gratitude to Richard Cullen and Kakes Kamata from the web development team at the Stroke Foundation for their "in-kind" support, and we thank all workshop participants for sharing their lived and professional experience with us.

Last, and most importantly, we would like to acknowledge and thank the survivors of stroke and their partners who agreed to be filmed and gave their permission for this content to be shared on the i-REBOUND *after stroke* website. Thank you to Meredith Burke, Colin and Alison Cunningham, Julie Davey, Kevin and Karen English, Paul Fink, Emma Gee, Bev Hopper, Clive Kempson, Shannon Nelson, Brooke Parsons, Ben Schelfhaut, and Nick Waters. Your generosity of time and giving spirit are wholeheartedly appreciated.

Data Availability

All data from this project are reported in this paper as multimedia appendices apart from interview transcripts. DP is happy to be contacted with any questions about this work.

Authors' Contributions

DP led the drafting and revisions of this manuscript with all the other authors making a significant contribution to the iteration of the draft manuscript. All authors signed off on the final version of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Terms of reference.
[DOCX File , 45 KB - jmir_v26i1e58419_app1.docx]

Multimedia Appendix 2
Tips for building an authentic partnership between consumer partners and researchers.
[DOCX File , 15 KB - jmir_v26i1e58419_app2.docx]

Multimedia Appendix 3
Co-design workshop outline for survivors of stroke.

XSL•FO
RenderX

[DOCX File , 27 KB - jmir_v26i1e58419_app3.docx]

Multimedia Appendix 4
Stroke survivor welcome video.
[AVI File , 22162 KB - jmir_v26i1e58419_app4.avi]

Multimedia Appendix 5
Workshop outline for other stakeholders.
[DOCX File , 27 KB - jmir_v26i1e58419_app5.docx]

Multimedia Appendix 6
Welcome video for other stakeholders.
[MP4 File (MP4 Video), 7653 KB - jmir_v26i1e58419_app6.mp4]

Multimedia Appendix 7
User Stories and User Acceptability Criteria.
[DOCX File , 25 KB - jmir_v26i1e58419_app7.docx]

Multimedia Appendix 8
Personas.
[DOCX File , 20 KB - jmir_v26i1e58419_app8.docx]

Multimedia Appendix 9
Functional Requirements.
[DOCX File , 28 KB - jmir_v26i1e58419_app9.docx]

Multimedia Appendix 10
CAG feedback on prototype.
[DOCX File , 18 KB - jmir_v26i1e58419_app10.docx]

Multimedia Appendix 11
User testing guide.
[DOCX File , 18 KB - jmir_v26i1e58419_app11.docx]

Multimedia Appendix 12
Proposed exercises.
[DOCX File , 20 KB - jmir_v26i1e58419_app12.docx]

References

1. Stroke, Cerebrovascular accident. World Health Organization. URL: https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html [accessed 2024-09-23]
2. Australia's Health 2018. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/getmedia/7c42913d-295f-4bc9-9c24-4e44eff4a04a/aihw-aus-221.pdf [accessed 2024-09-23]
3. Global Health Estimates. World Health Organization. 2018. URL: https://www.who.int/data/global-health-estimates [accessed 2018-09-19]
4. GBD 2016 Lifetime Risk of Stroke Collaborators. Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. N Engl J Med 2018 Dec 20;379(25):2429-2437. [doi: 10.1056/nejmoa1804492]
5. Feigin V, Forouzanfar M, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010)the GBD Stroke Experts Group. Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet 2014 Jan 18;383(9913):245-254 [FREE Full text] [doi: 10.1016/s0140-6736(13)61953-4] [Medline: 24449944]
6. Deloitte Access Economics. No postcode untouched: Stroke in Australia 2020. Stroke Foundation. 2020 Nov. URL: https://strokefoundation.org.au/media/juuba3qm/no-postcode-untouched-30-october-final-report.pdf [accessed 2024-09-23]
7. Mohan KM, Wolfe CD, Rudd AG, Heuschmann PU, Kolominsky-Rabas PL, Grieve AP. Risk and cumulative risk of stroke recurrence. Stroke 2011 May;42(5):1489-1494. [doi: 10.1161/strokeaha.110.602615]

XSL•FO
RenderX

8. O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. The Lancet 2016 Aug;388(10046):761-775. [doi: 10.1016/s0140-6736(16)30506-2]
9. Deijle IA, Van Schaik SM, Van Wegen EE, Weinstein HC, Kwakkel G, Van den Berg-Vos RM. Lifestyle interventions to prevent cardiovascular events after stroke and transient ischemic attack. Stroke 2017 Jan;48(1):174-179. [doi: 10.1161/strokeaha.116.013794]
10. D'Isabella NT, Shkredova DA, Richardson JA, Tang A. Effects of exercise on cardiovascular risk factors following stroke or transient ischemic attack: a systematic review and meta-analysis. Clin Rehabil 2017 Dec 19;31(12):1561-1572. [doi: 10.1177/0269215517709051] [Medline: 28523989]
11. English C, MacDonald-Wicks L, Patterson A, Attia J, Hankey GJ. The role of diet in secondary stroke prevention. The Lancet Neurology 2021 Feb;20(2):150-160. [doi: 10.1016/s1474-4422(20)30433-6]
12. Mora S, Cook N, Buring JE, Ridker PM, Lee I. Physical activity and reduced risk of cardiovascular events. Circulation 2007 Nov 06;116(19):2110-2118. [doi: 10.1161/circulationaha.107.729939]
13. Boehme AK, Esenwa C, Elkind MS. Stroke risk factors, genetics, and prevention. Circulation Research 2017 Feb 03;120(3):472-495. [doi: 10.1161/circresaha.116.308398]
14. Razmara A, Ovbiagele B, Markovic D, Towfighi A. Patterns and predictors of blood pressure treatment, control, and outcomes among stroke survivors in the United States. J Stroke Cerebrovasc Dis 2016 Apr;25(4):857-865. [doi: 10.1016/j.jstrokecerebrovasdis.2015.12.027] [Medline: 26778599]
15. Kjörk EK, Gunnel C, Lundgren-Nilsson Å, Sunnerhagen KS. Experiences, needs, and preferences for follow-up after stroke perceived by people with stroke and healthcare professionals: A focus group study. PLoS One 2019 Oct 1;14(10):e0223338 [FREE Full text] [doi: 10.1371/journal.pone.0223338] [Medline: 31574135]
16. Thilarajah S, Mentiplay BF, Bower KJ, Tan D, Pua YH, Williams G, et al. Factors associated with post-stroke physical activity: a systematic review and meta-analysis. Arch Phys Med Rehabil 2018 Sep;99(9):1876-1889. [doi: 10.1016/j.apmr.2017.09.117] [Medline: 29056502]
17. Nicholson SL, Donaghy M, Johnston M, Sniehotta FF, van Wijck F, Johnston D, et al. A qualitative theory guided analysis of stroke survivors' perceived barriers and facilitators to physical activity. Disabil Rehabil 2014 Dec 31;36(22):1857-1868. [doi: 10.3109/09638288.2013.874506] [Medline: 24377329]
18. Bailey RR. Lifestyle modification for secondary stroke prevention. Am J Lifestyle Med 2018 Feb 21;12(2):140-147 [FREE Full text] [doi: 10.1177/1559827616633683] [Medline: 30202386]
19. Winstein CJ, Stein J, Arena R, Bates B, Cherney LR, Cramer SC, et al. Guidelines for adult stroke rehabilitation and recovery. Stroke 2016 Jun;47(6):1. [doi: 10.1161/str.0000000000000098]
20. Fryer C, Luker JA, McDonnell MN, Hillier SL. Self management programmes for quality of life in people with stroke. Cochrane Database Syst Rev 2016 Aug 22;2016(8):CD010442 [FREE Full text] [doi: 10.1002/14651858.CD010442.pub2] [Medline: 27545611]
21. Ramage ER, Burke M, Galloway M, Graham ID, Janssen H, Marsden DL, et al. Fit for purpose. Co-production of complex behavioural interventions. A practical guide and exemplar of co-producing a telehealth-delivered exercise intervention for people with stroke. Health Res Policy Syst 2022 Jan 03;20(1):2 [FREE Full text] [doi: 10.1186/s12961-021-00790-2] [Medline: 34980156]
22. Zacharia K, Patterson AJ, English C, Ramage E, Galloway M, Burke M, et al. i-Rebound after Stroke-Eat for Health: Mediterranean dietary intervention co-design using an integrated knowledge translation approach and the TIDieR checklist. Nutrients 2021 Mar 24;13(4):1 [FREE Full text] [doi: 10.3390/nu13041058] [Medline: 33805076]
23. English C, Ramage ER, Attia J, Bernhardt J, Bonevski B, Burke M, et al. Secondary prevention of stroke. A telehealth-delivered physical activity and diet pilot randomized trial (ENAbLE-pilot). Int J Stroke 2024 Feb 29;19(2):199-208 [FREE Full text] [doi: 10.1177/17474930231201360] [Medline: 37658738]
24. Communications and media in Australia series: How we use the internet. Australian Communications and Media Authority. 2022 Dec. URL: https://www.acma.gov.au/sites/default/files/2023-03/HOWWEU~1.PDF [accessed 2024-09-23]
25. Ramadas A, Quek KF, Chan CK, Oldenburg B, Hussein Z. Randomised-controlled trial of a web-based dietary intervention for patients with type 2 diabetes mellitus: study protocol of myDIDeA. BMC Public Health 2011 May 21;11(1):359 [FREE Full text] [doi: 10.1186/1471-2458-11-359] [Medline: 21599990]
26. Pogrebnoy D, Dennett AM, Simpson DB, MacDonald-Wicks L, Patterson AJ, English C. Effects of using websites on physical activity and diet quality for adults living with chronic health conditions: systematic review and meta-analysis. J Med Internet Res 2023 Oct 19;25:e49357 [FREE Full text] [doi: 10.2196/49357] [Medline: 37856187]
27. Nguyen T, Graham ID, Mrklas KJ, Bowen S, Cargo M, Estabrooks CA, et al. How does integrated knowledge translation (IKT) compare to other collaborative research approaches to generating and translating knowledge? Learning from experts in the field. Health Res Policy Syst 2020 Mar 30;18(1):35 [FREE Full text] [doi: 10.1186/s12961-020-0539-6] [Medline: 32228692]
28. Kothari A, McCutcheon C, Graham ID. Defining integrated knowledge translation and moving forward: a response to recent commentaries. Int J Health Policy Manag 2017 May 01;6(5):299-300 [FREE Full text] [doi: 10.15171/ijhpm.2017.15] [Medline: 28812820]

XSL•FO
RenderX

29. Boland L, Kothari A, McCutcheon C, Graham ID, Integrated Knowledge Translation Research Network. Building an integrated knowledge translation (IKT) evidence base: colloquium proceedings and research direction. Health Res Policy Syst 2020 Jan 20;18(1):8 [FREE Full text] [doi: 10.1186/s12961-019-0521-3] [Medline: 31959184]
30. Knowledge translation framework for ageing and health. World Health Organization. 2012. URL: https://www.who.int/publications/m/item/knowledge-translation-framework-for-ageing-and-health [accessed 2024-09-23]
31. Hunter Miller B. What is Design Thinking? (And What Are The 5 Stages Associated With it?). Medium. 2017 Sep 04. URL: https://medium.com/@bhmiller0712/what-is-design-thinking-and-what-are-the-5-stages-associated-with-it-d628152cf220 [accessed 2024-09-23]
32. Razzouk R, Shute V. What Is design thinking and why is it important? Review of Educational Research 2012 Sep 01;82(3):330-348 [FREE Full text] [doi: 10.3102/0034654312457429]
33. Van Duyne D, Landay JA, Hong JI. The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience. Reading, MA: Addison-Wesley; 2002.
34. Cohn M. User Stories Applied: For Agile Software Development. Reading, MA: Addison-Wesley; 2004.
35. Cooper A. The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity. Carmel, IN: Sams - Pearson Education; 1999.
36. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. Burlington, MA: Morgan Kaufmann; 2010.
37. Li W, Badr Y, Biennier F. Improving Web Service Composition with User Requirement Transformation and Capability Model. In: Meersman R, Panetto H, Dillon T, Eder J, Bellahsene Z, Ritter N, et al, editors. On the Move to Meaningful Internet Systems: OTM 2013 Conferences. OTM 2013. Lecture Notes in Computer Science, vol 8185. Berlin, Germany: Springer; 2013.
38. Lancaster A. Paper prototyping: the fast and easy way to design and refine user interfaces. IEEE Trans. Profess. Commun 2004 Dec;47(4):335-336. [doi: 10.1109/tpc.2004.837973]
39. Dumas JS, Redish JC. A Practical Guide to Usability Testing. Bristol, England: Intellect Ltd; 1999.
40. Short C. Web content management: CMS for competitive advantage. J Direct Data Digit Mark Pract 2010 Jan 28;11(3):198-206 [FREE Full text] [doi: 10.1057/dddmp.2009.32]
41. Garett R, Chiu J, Zhang L, Young SD. A literature review: website design and user engagement. Online J Commun Media Technol 2016 Jul;6(3):1-14 [FREE Full text] [Medline: 27499833]
42. Ries E. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. Surfers Paradise, Queensland: Crown Currency; 2011.
43. Welcome to i-Rebound Prototype. URL: https://irebound-nsf.webflow.io/ [accessed 2024-09-23]
44. Physical activity. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/physical-activity/physical-activity [accessed 2024-09-23]
45. Byrne M, McSharry J, Meade O, Lavoie KL, Bacon SL. An international, Delphi consensus study to identify priorities for methodological research in behavioral trials in health research. Trials 2020 Mar 23;21(1):292 [FREE Full text] [doi: 10.1186/s13063-020-04235-z] [Medline: 32293510]
46. Thesing T, Feldmann C, Burchardt M. Agile versus waterfall project management: decision model for selecting the appropriate approach to a project. Procedia Computer Science 2021;181:746-756 [FREE Full text] [doi: 10.1016/j.procs.2021.01.227]
47. i-REBOUND after Stroke. URL: https://irebound.enableme.org.au/ [accessed 2024-09-23]
48. Brooke J. SUS: A quick and dirty usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1995:189.
49. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]

Abbreviations

CAG: Consumer Advisory Group
CMS: content management system
iKT: integrated knowledge translation
UAT: user accessibility testing

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 15.03.24; peer-reviewed by H Gooch, A Haynes; comments to author 02.07.24; revised version received 08.07.24; accepted 31.07.24; published 22.10.24.

Please cite as:
Pogrebnoy D, Ashton L, Beh BA, Burke M, Cullen R, Czerenkowski J, Davey J, Dennett AM, English K, Godecke E, Harper N, Lynch E, MacDonald-Wicks L, Patterson A, Ramage E, Schelfhaut B, Simpson DB, Zacharia K, English C
Adapting a Telehealth Physical Activity and Diet Intervention to a Co-Designed Website for Self-Management After Stroke: Tutorial
J Med Internet Res 2024;26:e58419
URL: https://www.jmir.org/2024/1/e58419
doi:10.2196/58419
PMID:

©Dina Pogrebnoy, Lee Ashton, Brian A Beh, Meredith Burke, Richard Cullen, Jude Czerenkowski, Julie Davey, Amy M Dennett, Kevin English, Erin Godecke, Nicole Harper, Elizabeth Lynch, Lesley MacDonald-Wicks, Amanda Patterson, Emily Ramage, Ben Schelfhaut, Dawn B Simpson, Karly Zacharia, Coralie English. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Evaluation of a Smartphone-Based Chatbot Coach to Facilitate a Balanced Lifestyle in Individuals With Headaches (BalanceUP App): Randomized Controlled Trial

Sandra Ulrich[1], MSc; Andreas R Gantenbein[2,3], MD; Viktor Zuber[1], BSc; Agnes Von Wyl[1], PhD; Tobias Kowatsch[4,5,6*], PhD; Hansjörg Künzli[1*], LicPhil

[1]School of Applied Psychology, Zurich University of Applied Sciences, Zurich, Switzerland
[2]Pain and Research Unit, ZURZACH Care, Bad Zurzach, Switzerland
[3]Department of Neurology, University Hospital Zurich, Zurich, Switzerland
[4]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland
[5]School of Medicine, University of St.Gallen, St. Gallen, Switzerland
[6]Centre for Digital Health Interventions, Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland
[*]these authors contributed equally

Corresponding Author:
Sandra Ulrich, MSc
School of Applied Psychology
Zurich University of Applied Sciences
Pfingstweidstrasse 96
2
Zurich, 8005
Switzerland
Phone: 41 58 934 ext 8451
Email: sandra.ulrich@zhaw.ch

Abstract

Background: Primary headaches, including migraine and tension-type headaches, are widespread and have a social, physical, mental, and economic impact. Among the key components of treatment are behavior interventions such as lifestyle modification. Scalable conversational agents (CAs) have the potential to deliver behavior interventions at a low threshold. To our knowledge, there is no evidence of behavioral interventions delivered by CAs for the treatment of headaches.

Objective: This study has 2 aims. The first aim was to develop and test a smartphone-based coaching intervention (BalanceUP) for people experiencing frequent headaches, delivered by a CA and designed to improve mental well-being using various behavior change techniques. The second aim was to evaluate the effectiveness of BalanceUP by comparing the intervention and waitlist control groups and assess the engagement and acceptance of participants using BalanceUP.

Methods: In an unblinded randomized controlled trial, adults with frequent headaches were recruited on the web and in collaboration with experts and allocated to either a CA intervention (BalanceUP) or a control condition. The effects of the treatment on changes in the primary outcome of the study, that is, mental well-being (as measured by the Patient Health Questionnaire Anxiety and Depression Scale), and secondary outcomes (eg, psychosomatic symptoms, stress, headache-related self-efficacy, intention to change behavior, presenteeism and absenteeism, and pain coping) were analyzed using linear mixed models and Cohen *d*. Primary and secondary outcomes were self-assessed before and after the intervention, and acceptance was assessed after the intervention. Engagement was measured during the intervention using self-reports and usage data.

Results: A total of 198 participants (mean age 38.7, SD 12.14 y; n=172, 86.9% women) participated in the study (intervention group: n=110; waitlist control group: n=88). After the intervention, the intention-to-treat analysis revealed evidence for improved well-being (treatment: β estimate=–3.28, 95% CI –5.07 to –1.48) with moderate between-group effects (Cohen *d*=–0.66, 95% CI –0.99 to –0.33) in favor of the intervention group. We also found evidence of reduced somatic symptoms, perceived stress, and absenteeism and presenteeism, as well as improved headache management self-efficacy, application of behavior change techniques, and pain coping skills, with effects ranging from medium to large (Cohen *d*=0.43-1.05). Overall, 64.8% (118/182) of the participants used coaching as intended by engaging throughout the coaching and completing the outro.

XSL•FO
RenderX

Conclusions: BalanceUP was well accepted, and the results suggest that coaching delivered by a CA can be effective in reducing the burden of people who experience headaches by improving their well-being.

Trial Registration: German Clinical Trials Register DRKS00017422; https://trialsearch.who.int/Trial2.aspx?TrialID=DRKS00017422

(J Med Internet Res 2024;26:e50132) doi:10.2196/50132

KEYWORDS

chatbot; mobile health; mHealth; smartphone; headache management; psychoeducation; behavior change; stress management; mental well-being; lifestyle; mindfulness; relaxation; mobile phone

Introduction

Background

Primary headaches, including tension-type headaches (TTHs) and migraine, are among the most prevalent neurological illnesses [1]. TTH and migraine are ranked as the third and sixth most common diseases, respectively, worldwide in both women and men [2]. The physical, social, and mental burden of headaches, defined as the summation of all negative consequences [3], is substantial. In addition, it is important to consider the quality of life impacted by headaches, defined as the subjective assessment of general well-being, position, and prospects in life. Individuals who have recurrent headaches are often afraid of the next headache attack, which can lead to avoidance behaviors, such as the cancelation of social activities or not even planning them [4].

The economic costs of headaches are substantial, primarily manifesting as indirect and intangible costs. This includes work absences because of headaches (absenteeism), reduced on-the-job performance while experiencing headaches (presenteeism), reduced quality of life, and increased pain outside the workplace [5-7]. In the European Union, the total annual costs associated with headaches among adults are estimated at €173 billion (US $189 billion; 64% migraine, 12% TTH, and 24% other types of headaches) [8]. A more recent study conducted in Canada estimated the total annual cost associated with migraine to be CAD $23,756.04 (US $17,750.50) per patient [9]. Given the high personal and financial costs, there is an urgent need for effective management of headaches.

Headaches are multifactorial, and besides physiological factors, lifestyle factors, such as stress or sleep, play a significant role in the development and retention of a headache [10-12]. Insufficient perception of stress reactions, individual attitudes toward stress (eg, high-performance orientation and anxiety), or coping strategies, such as avoidance versus endurance, are seen as dysfunctional stress-coping in relation to headaches [13,14]. The belief that pain-related factors are outside one's control and the perceived inability to control these factors (ie, low self-efficacy) are further dysfunctional coping mechanisms associated with poor adjustment to headache and psychological functioning [15,16]. Among people who have headache, stress, stress regulation, and mental tension are perceived as critical triggers of a headache [14,17-19], and headache itself serves as a stressor that negatively affects well-being [20]. However, there are controversial findings regarding the association between lifestyle and headache, and individuals differ in the extent to which these factors interact with headache [21].

Guidelines recommend pharmacological and nonpharmacological interventions as standard therapy [22], including behavioral treatment, which has been shown to be effective in both face-to-face and web-based settings [22-25]. These treatments incorporate psychoeducation, relaxation techniques, physical activity, triggers management, and cognitive behavioral therapy elements, focusing on stress management and coping strategies to modify negative and dysfunctional cognitions, emotions, and behavior related to headaches [26,27]. A person-centered approach that integrates various intervention components is more effective [28,29] and may enhance personal control and efficacy in headache management [14]. However, challenges such as cost, access, motivation [18], and stigma hinder engagement [30]. Despite this, consistent care using nonmedical options has proven beneficial [31].

In addition to traditional evidence-based treatment modalities, mindfulness-based interventions have gained research interest, showing enhanced well-being in various settings [32-34], particularly for coping with chronic conditions, such as chronic pain [35-37]. Studies have demonstrated that mindfulness interventions benefit individuals with headaches by improving psychological functioning [38-40]. Furthermore, they help improve affective conditions, such as anxiety and depression, which are often related to poorer treatment outcomes [41,42]. Improvements in these conditions, even if subclinical, may improve coping skills for headaches and increase treatment adherence. In addition, individuals without a psychiatric diagnosis may experience headache-attributed disability, defined as physical, cognitive, and mental incapacities imposed by headaches [3], such as disabling anxiety related to fear of headaches and perceived triggers [43].

Over the past few years, the adoption of app-based interventions for headache management has increased [44,45]. These interventions primarily take the form of electronic headache diary apps, which offer practical solutions for data monitoring [46]. Furthermore, app-based interventions have improved our understanding of the relationship between lifestyle factors and headaches [21]. Apps are also suitable for providing guideline-compliant therapeutic options [23], such as psychoeducation, relaxation techniques, endurance sports, and other elements of behavioral therapy (eg, stress reduction). However, evidence of the effectiveness of app-based behavioral interventions for managing headaches remains weak [44,47].

XSL•FO
RenderX

Conversational agents (CAs), also known as chatbots or digital assistants, are increasingly being applied in both clinical [48-50] and nonclinical [51-53] health care settings to support disease management and behavioral lifestyle interventions. CAs engage users in humanlike conversations [48,54-56], enabling factual, relational, and emotional communication. This interactive style enhances engagement by establishing a working alliance between the users and the CAs [57,58], reflecting the collaborative relationship between shared treatment goals and tasks [59], which is crucial for treatment success in psychotherapy and counseling [60]. In contrast, conventional mobile health (mHealth) interventions may lack a therapist relationship and suffer from noncommittal timing [61], may not be used as intended [62], and engagement may often be low [63-65]. Nevertheless, CA-based coaching offers the potential to deliver personalized, accessible, and scalable content via web-based or mobile-based apps [56,66].

Objectives

To our knowledge, there is no evidence of mHealth coaching interventions delivered by CAs for the treatment of headaches. Building on a successful pilot study [67], we designed BalanceUP, a smartphone-based and CA-delivered intervention aimed at supporting a healthy lifestyle in people with headache. BalanceUP aims to improve mental well-being by promoting behavior change techniques (BCTs) in behaviors, emotions, thoughts, and beliefs related to headaches while ensuring low-threshold access and scalability. Using smartphone apps' technical potential, scalable interventions can be beneficial in supporting individuals [68]. Consequently, this study had the following objectives: (1) to develop a smartphone-based and CA-delivered intervention for people with headache and (2) to evaluate its effectiveness, engagement, and acceptance.

Methods

App Development

BalanceUP, developed for iOS (Apple Inc) and Android (Google LLC) platforms using MobileCoach [54,69,70], provides a chat-based interface for communicating with the CA (Figure 1B). The communication between users and the server is encrypted. The chat feature offers predefined answer options and free-text input, guiding conversations along dynamic paths to individualized tasks. The CA also shares videos and pictures elaborating on the psychoeducational content. Using the sidebar, users can access the (1) chat channel; (2) audio library (eg, relaxation, mindfulness, and imagination exercises); (3) illustrations; (4) working materials (eg, energy balance and coping circle); (5) video library (eg, animated psychoeducational videos); and (6) frequently asked questions about the study and BalanceUP app (Figure 1D).

Figure 1. Screenshots of the BalanceUP app: (A) conversational agent (CA) selection screen, (B) screen displaying a chat with the CA, (C) screen displaying push reminder from the CA, and (D) sidebar of the app.

Coaching Intervention

Overview

Drawing on the best practice from behavior therapy, BalanceUP is based on the cognitive behavior change migraine therapy manual (Cognitive-Behavioral Therapy for Migraine Management [Kognitiv-verhaltenstherapeutisches Migränemanagement]; MIMA) [13], which has been demonstrated to be feasible [71] and effective, showing results similar to those of the active control group at a 12-month follow-up [72]. BalanceUP comprises 7 consecutive modules: (1) headaches, (2) relaxation, (3) balance, (4) fear, (5) coping, (6) trigger, and (7) stress (refer to Multimedia Appendix 1 for an overview of the coaching), with a procedure similar to MIMA: (1) feedback on tasks, (2) psychoeducation, (3) reflection of behavior, (4) behavioral intention, (5) action planning, and (6) relaxation and imagination (refer to Figure 2 for further details). Each module contains 3 to 4 units (Figure 3), and users can work through these units during sessions with the CA according to their preferences, allowing completion within 24 to 60 days. Examples of the session flow can be found in Figure 4. After completion, users can still access the materials. A diagram showing the intervention flow is presented in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 2. Elements of coaching and engagement.

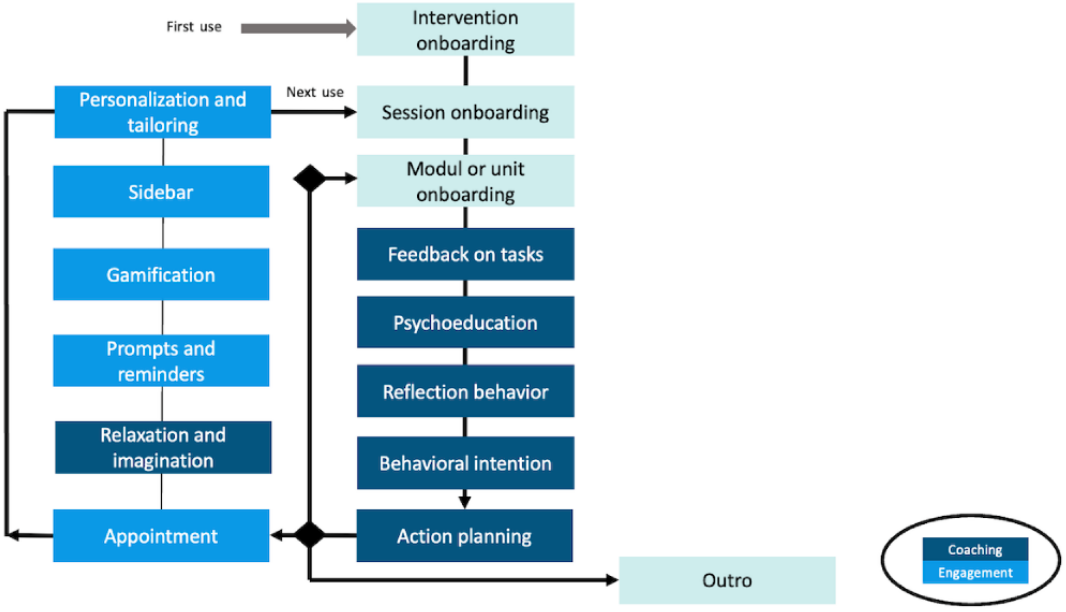

Figure 3. Structure of modules and units.

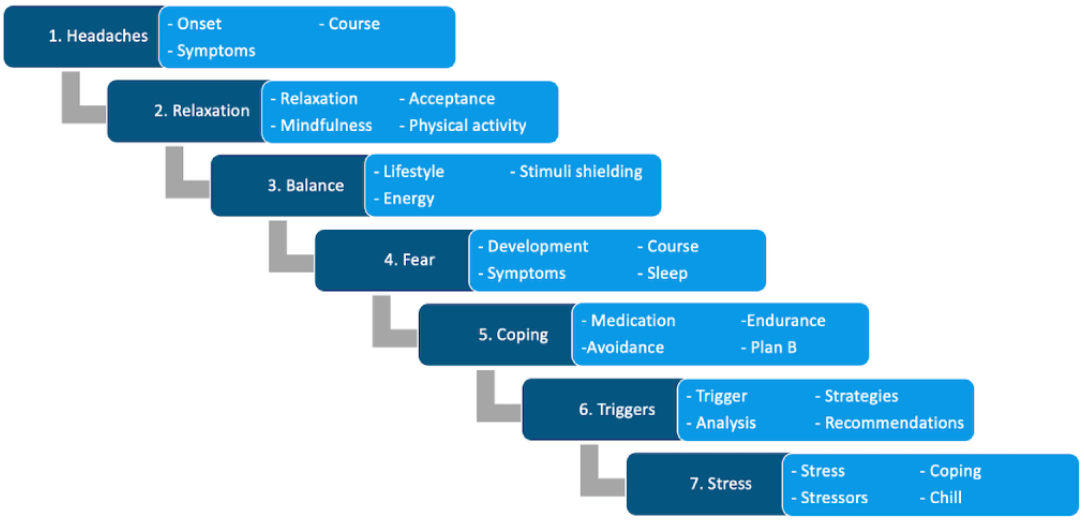

XSL•FO
RenderX

Figure 4. Examples of session flow.

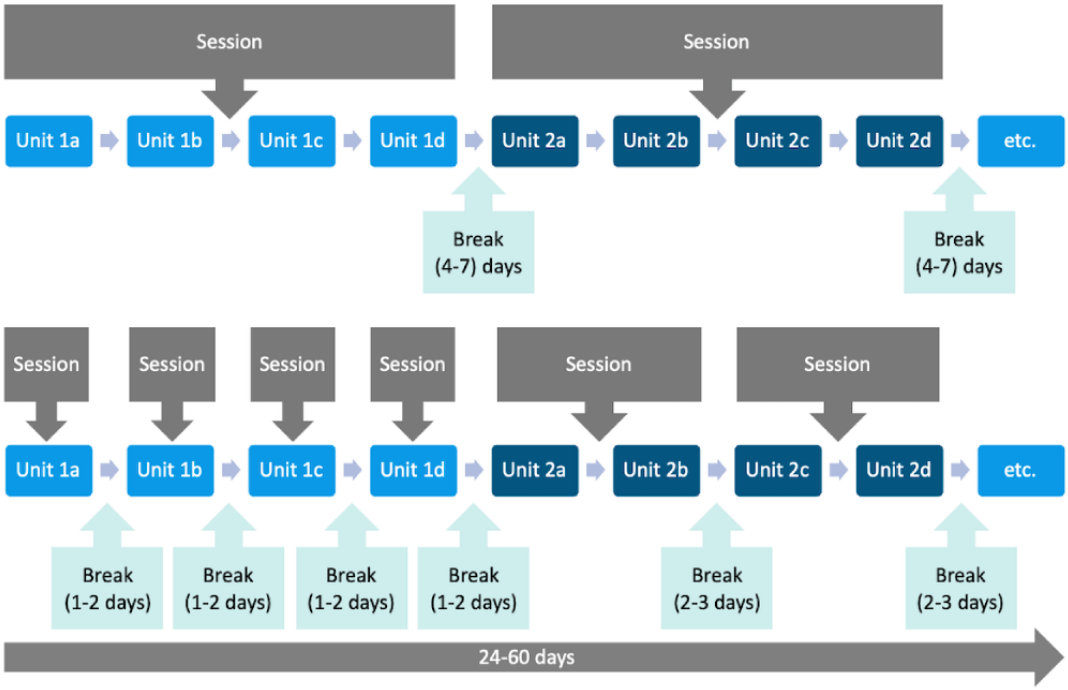

Behavior Change

To foster health-promoting behaviors, various BCTs [73] have been considered (eg, gamification for reward, action planning, and prompts to perform an exercise). A complete list of BCTs and their specific applications is outlined in Multimedia Appendix 3 [73]. BCTs are intervention ingredients designed to alter or redirect causal processes that regulate behavior [74]. Their specific implementation enables accurate replication, precise specification of the intervention content, and investigation of possible mechanisms of action [74].

Engagement

Tailoring (at a subgroup level) and personalization (at an individual level) are essential for promoting trust, engagement, adherence, and effectiveness in mHealth interventions [63,75]. Both tailoring (eg, psychoeducation material based on headache types) and personalization (eg, coach selection [Figure 1A] personal greeting, personalized goals, and individual appointments with the CA) were applied in BalanceUP. Reminders (Figure 1C) are an effective way of improving engagement, especially when they address specific needs [76,77]; refer to Multimedia Appendix 4 [49,56,68,78-81] for the detailed aspects of the engagements implemented.

Intended Use

The intended use of BalanceUP, which is the extent to which an individual needs to experience the content to derive maximum benefit from the intervention [81], is determined by reaching the outro and completing the postsurvey. Users can skip a module or elements based on their diagnosis (migraine vs TTH) and preferences. For instance, participants can skip psychoeducation within a unit and proceed to behavior reflection, whereas relaxation exercises, videos, and worksheets are optional. This procedure offers flexibility in line with the recommended mHealth intervention strategies [82,83] and is similar to the dose-response rate in pharmaceutical research, where medication dosage can vary depending on the patient's condition and characteristics. Furthermore, we base this approach on the self-determination theory [84], which emphasizes autonomy, competence, and engagement.

Study Design and Procedure

We conducted an unblinded, 2-arm, randomized controlled trial (RCT). After onboarding, eligible participants were asked to provide electronic informed consent. Participants who provided consent proceeded to complete the baseline survey (T1), while those who were ineligible were directed to a farewell conversation with the CA. Those meeting the inclusion criteria and consenting were randomly assigned with a 1:1 allocation ratio to either the intervention or waitlist control group using random numbers (0-1) generated by the BalanceUP app, with numbers below 0.5 assigned to the intervention group. The intervention group immediately began with the coaching. The postintervention survey (T2) was conducted 24 to 60 days after randomization. The waitlist control group received weekly reminders from the CA during the 42-day waiting period. After this, they completed the postintervention survey (T2) and were given the option to access coaching or proceed to the outro (Figure 5). Throughout the study, self-reported primary and secondary outcomes were collected within the BalanceUP app and via the in-app survey tool, LimeSurvey (version 3.4). Refer to Multimedia Appendix 5 for outcome details and data collection times.

XSL•FO
RenderX

Figure 5. Study procedure.

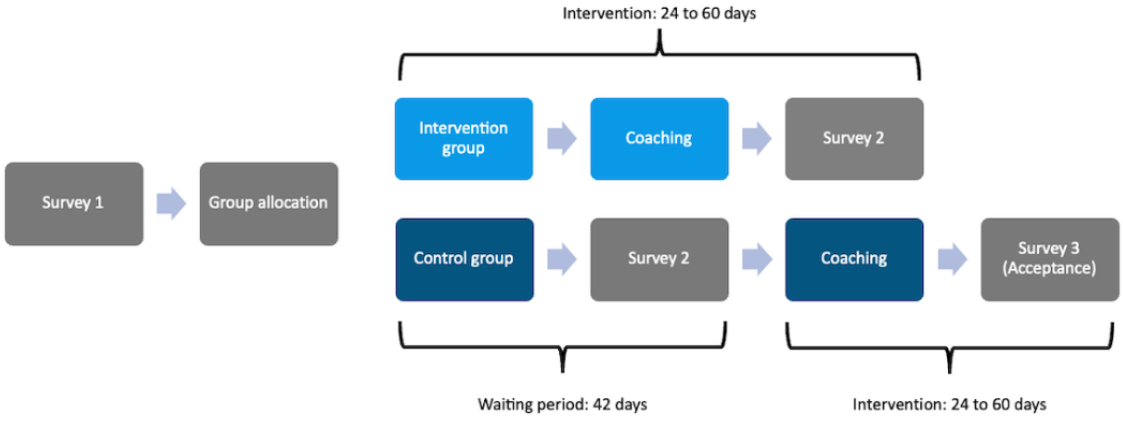

Study Participants

Participants were self-recruited in Switzerland, Germany, and Austria between April and November 2022 (German-speaking parts) via the study website. The link to the website was shared via social media by headache organizations, insurance companies, and health care institutions. This recruitment strategy allowed participation from all German-speaking countries. The study website provided study information and app download links. Inclusion criteria were being an adult (aged ≥18 y) with a smartphone (eg, iOS [Apple Inc] or Android [Google LLC]), fluent German-speaking skills, and experiencing regular headaches for at least 3 months with a minimum of 4 incidents per month. These criteria were assessed using the BalanceUP app during onboarding.

Outcomes

Mental Well-Being

To measure effectiveness, we defined mental well-being as the primary outcome, measured by the Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS) [85]. The PHQ-ADS is a composite of the Patient Health Questionnaire-9 (PHQ-9) [86] and General Anxiety Disorder Scale-7 (GAD-7) [87]. Scores can range from 0 to 48, with higher scores indicating more severe depression and anxiety; 3 to 4 points were considered the minimum clinically important difference. Cutoff scores of 10, 20, and 30 denoted mild, moderate, and severe degrees of depression and anxiety, respectively.

To assess psychological functioning, multiple secondary outcome measures were used in accordance with the established guidelines [88].

Secondary Outcomes

Depression

The PHQ-9 [86] consists of 9 items for evaluating depressive symptoms, rated on a 4-point Likert scale ranging from 0 (not at all) to 3 (nearly every day). Higher scores indicate higher symptom severity, with scores ranging from 0 to 4 indicating no symptoms of depression and scores from 5 to 9, 10 to 14, 15 to 19, and 20 to 27 indicating mild, moderate, moderately severe, and severe depression, respectively.

Anxiety

The GAD-7 [87] is used for evaluating symptoms of generalized anxiety disorder. It comprises 7 items. Similar to PHQ-9, answers are rated on a 4-item Likert scale ranging from 0 (not at all) to 3 (nearly every day). Higher scores indicate higher levels of anxiety, and the total score ranges from 0 to 21. Scores from 0 to 4, 5 to 9, 10 to 14, and 15 to 21 denoted minimal, mild, moderate, and severe anxiety, respectively.

Somatic Symptoms

We measured somatic symptoms using the Patient Health Questionnaire-15 (PHQ-15) [89]. It is a 15-item self-report questionnaire that can be scored on a scale from 0 (not impaired) to 2 (severely impaired). A total score of ≥15 on the PHQ-15 indicates a high level of impairment owing to somatic symptoms [90]. For this study, we adopted 2 items from the PHQ-9 because the items were similar. However, the answer scale of the 2 items differed between the PHQ-9 and PHQ-15 and had to be converted according to the manual.

Stress

We measured stress with the German version of the Perceived Stress Scale-10 [91]. The 10-item questionnaire can be rated on a scale from 0 (never) to 4 (very often); higher scores reflect a higher level of perceived stress, and scores can range from 0 to 40.

Self-Efficacy

To assess headache-related self-efficacy, we used the German short form of the Headache Management Self-Efficacy (HMSE-G-SF) Scale [92]. The measurement consists of 6 items assessing self-efficacy beliefs related to headaches. Answer scales range from 1 (do not agree) to 7 (agree). Higher scores implied higher self-efficacy expectations, and summed scores <19 indicated below-average self-efficacy expectations compared with other people experiencing headache.

Intention to Change Behavior

To assess participants' intention to change behavior, we used the health action process approach model [93], which categorizes behavior change into 3 stages: nonintenders, intenders, and actors. In this study, participants indicated their use of psychological techniques for headache treatment by choosing 1 of the 5 possible answers: (1) no, and I do not intend to do so (nonintender); (2) no, but I am considering it (nonintender); (3)

XSL•FO
RenderX

no, but I have the intention to do so (intender); (4) yes, but it is not easy (actor); and (5) yes, and it is easy (actor).

Absenteeism and Presenteeism

To measure work-related impairment due to headaches, we applied 4 out of 5 questions from the Migraine Disability Assessment [94]. These questions assessed days with complete loss and days with at least 50% reduced productivity (eg, work, household, and school) for the past 3 months. Given the study's runtime and potential recall challenge [94], participants reported headache days for a 1-month period instead of 3 months.

Pain Processing

We applied the Questionnaire for the Assessment of Pain Processing (questionnaire to assess pain management; Fragebogen zur Erfassung der Schmerzverarbeitung) to measure pain coping strategies [95]. This tool assesses coping strategies in individuals with persistent pain and is comprised of 2 parts. In this study, we used the first part to evaluate cognitive and behavioral coping using 24 items. The cognitive coping subscale included the dimensions "action planning skills," "cognitive restructuring," and "experience of competence." The behavioral coping subscale included "mental distraction," "counteracting activities," and "rest and relaxation techniques." Answers were scored from 1 (not at all true) to 6 (always true), with higher scores indicating better pain processing.

Sociodemographics

We collected data on age, sex, level of education, parallel app use for headache tracking, concurrent psychotherapy, commitment to the program, and headache diagnosis at baseline to describe the study population. We further assessed participants' sensitivity to triggers and tendency to avoid triggers optionally in module 6 (Trigger) using the German short version of the Headache Triggers Sensitivity and Avoidance Questionnaire [96]. Mean scores <2.03 indicate below-average trigger sensitivity and scores >3.19 indicate above-average trigger sensitivity, respectively. For avoidance, mean scores <2.09 indicate below-average trigger avoidance and scores >3.23 indicate above-average trigger avoidance, respectively.

Engagement

Overview

According to a systematic review conceptualizing engagement with digital behavior change interventions [64], engagement is both a multidimensional concept and a dynamic process. Engagement consists of 2 parts: (1) the extent of use (eg, amount, frequency, duration, and depth) and (2) a subjective experience characterized by attention, interest, and affect. By using this multidimensional approach, we aimed to capture the various aspects of engagement, as defined by Perski et al [64].

Extent of Use

The following use data were recorded during the coaching intervention: total minutes spent on in-app relaxation and imagination exercises, total reminders sent to participants in cases of inactivity, and the average number of days taken to complete 1 coaching module. In addition, the percentage of answered conversational turns between the participant and the CA coach was calculated, where a higher number indicated higher engagement with the intervention. We also assessed intended use, that is, the number of participants who completed the outro.

Subjective Experience

To gather the subjective experiences of the participants, we used 4 items from the German Group Therapy Session Evaluation by Patients [97] with statements about personal involvement, active participation, perceived comprehensibility, and perceived benefit on a 5-point Likert scale ranging from 1 (disagree) to 5 (agree). Furthermore, we measured perceived enjoyment by applying a single-item measure from technology acceptance research [98,99] ("Did you enjoy the last unit?"), ranging from 1 (not at all) to 5 (very much).

We assessed participants' commitment to changing their behavior with 1 question ("How committed are you towards changing your behavior?") on a scale of 1 to 10.

We used a modified version of the Session Alliance Inventory—Patient Version [100,101] to repeatedly measure the working alliance between the participant and the CA. It consists of 3 items for the Bond Scale and 3 items for the Task and Goal Scale. In this study, we used the items of the validated German version of the Working Alliance Inventory-Short Revised [102], which features a 5-point answer scale ranging from 1 (seldom) to 5 (always). In addition, we contextualized the Session Alliance inventory by replacing the term "therapist" with the CA's name.

Acceptance

We assessed the acceptance of BalanceUP with a slightly modified and translated version of the Mobile App Rating Scale (uMARS) [103] assessing engagement (eg, entertainment, interest, customization, interactivity, and target group of the app); information (eg, quality of information, quantity of information, visual information, and credibility of source); perceived quality (eg, recommendation, use, payment, and overall rating); and perceived impact (eg, awareness, knowledge, attitudes, behavior change, seeking help, and intention to change). All subscales used a 5-point Likert scale ranging from 1 to 5, with higher scores indicating a more favorable judgment.

In addition to the uMARS, the questions "What did you like most about the BalanceUP app?" and "What would you like to see improved about the BalanceUP app?" could be answered in free text.

Impression of Change and Adverse Events

We assessed adverse events with the Patients Global Impression of Change Scale [104]. It is a 7-point scale depicting the perceived overall improvement in general health, rated from 1 (much improved) to 7 (very much worse). In this study, for scores ≥5, we also assessed whether participants believed these changes occurred because of coaching or whether other circumstances (eg, professional situation and conflicts in the social sphere) caused these changes. Participants were asked to note the adverse changes if coaching was given as a reason.

XSL•FO
RenderX

Sample Size Calculation

We estimated the sample size based on the primary outcome (mental well-being) measured by the PHQ-ADS for a linear mixed model (LMM) and a repeated measure ANOVA (within-between interaction). Consistent with previous headache and chronic pain research [105-107], we assumed a small-to-medium effect size for the primary outcome. Statistical power calculation using G*Power[3] software revealed that a sample size of 90 (45 for each group) would be sufficient, with a power of 0.80 to detect a small-to-medium time×group interaction effect size (Cohen *d*=0.35) with an α of .05 and based on 2 measurements. According to our pilot study, we estimated a dropout rate of 40% and aimed to recruit approximately 150 participants.

Data Analysis

The analysis was performed using SPSS (version 28.0; IBM Corp) and R software (version 4.2.2; R Foundation for Statistical Computing) including the *lme4*. For the primary outcome (PHQ-ADS) from before the intervention (T1) to after the intervention (T2), the LMMs were used, considering time (T1 and T2), group (intervene and wait), and their interaction as fixed effects, with participants as a random factor. Secondary outcomes (ie, PHQ-9, GAD-7, PHQ-15, Perceived Stress Scale-10, Questionnaire for the Assessment of Pain Processing, HMSE-G-SF, Migraine Disability Assessment, and health action process approach) were analyzed accordingly. Missing data were managed using LMM, which is based on all observed data and accounts for data missing at random [108-110]. According to the CONSORT (Consolidated Standards of Reporting Trials) guidelines, we reported the LMM analysis for the (1) intention-to-treat (ITT) analysis, in which all randomized participants were included, regardless of whether they used the coaching and (2) per-protocol (PP) analysis of complete cases. Calculations of within- and between-group effect sizes (Cohen *d*) were based on the pooled SD of complete cases only and labeled as small (Cohen *d*=0.2), medium (Cohen *d*=0.5), and large (Cohen *d*=0.8). The influence of predictors on outcomes was explored using the LMM, including the primary outcome (PHQ-ADS), with a focus on the 3-way interaction. Change in engagement over time was analyzed via repeated-measures ANOVA, and the effect of early engagement [64] on treatment outcomes was analyzed using linear regression. Descriptive statistics were used to summarize participant characteristics at baseline, and 2-tailed *t* tests were used to assess baseline differences. We applied qualitative content analysis [111,112] to answer the open-ended questions.

Ethical Considerations

The Swiss Ethics Committee Zurich reviewed the research project and confirmed (Swiss Ethics BASEC-Nr. Req-2021-01365) that it does not fall within the scope of the Human Research Act. This research project was registered in the World Health Organization–accredited German Clinical Trials Register (DRKS00017422). We performed this trial based on the CONSORT-EHEALTH guidelines.

Results

Participant Flow and Baseline Characteristics

During the recruitment phase, from April to November 2020, 405 individuals downloaded the BalanceUP app. Of these, 223 (55.1%) were assessed for eligibility and 7 (1.7%) were excluded from the study. Of those eligible, 198 (48.9%) individuals completed the baseline survey and were randomized into the intervention (n=110) and control (n=88) groups. The full participant flow is presented in Figure 6. The dropout rate for randomized participants after the treatment was 29.1% (32/110) for the intervention group and 18% (16/88) for the waitlist control group and thus can be considered low, particularly for a fully unguided mHealth intervention [113] and in comparison with our pilot study.

XSL•FO
RenderX

Figure 6. Participant flowchart. ITT: intention-to-treat; PP: per-protocol.

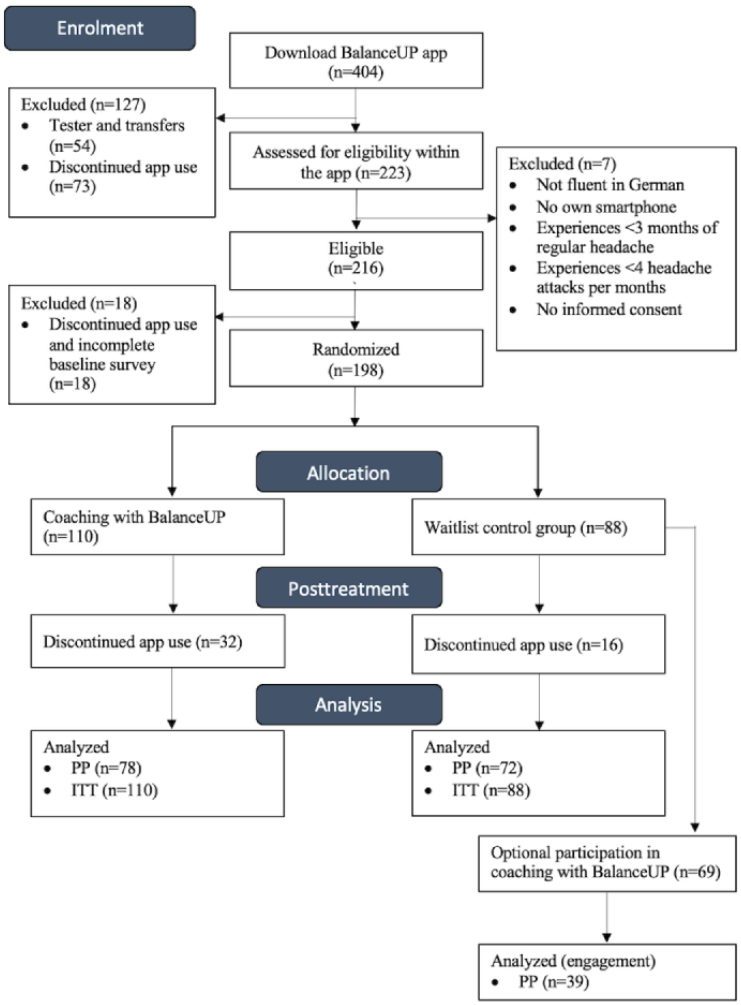

As presented in Table 1, most of the participants were women (180/198, 90.9%) with a mean age of 38.7 (SD 12.14) years, and more than half had a university degree (104/198, 52.5%). Migraine was the most prevalent diagnosis, accounting for 72.2% (143/198) of the sample. Approximately half of the participants (95/198, 48%) reported using headache diaries to track their symptoms. Few participants (32/198, 16.2%) reported attending psychotherapy and using the coaching app. In general, participants reported an average of 6.66 (SD 7.34) days of work per month missed because of headaches and 11.82 (SD 9.77) days per month when their performance was reduced by half or more (including work, school, and household). Compared with other individuals who had headaches [92], the study participants reported average levels of headache-related self-efficacy. On average, participants were classified as “intenders,” indicating that they had the intention to change their behavior (as opposed to “nonintenders” or “actors”). On average, the participants had mild depression (mean 9.06, SD 4.24), mild anxiety (mean 6.76, SD 3.87), and moderate psychosomatic symptoms (mean 10.92, SD 4.34). There was no difference between the groups in any of the outcomes.

XSL•FO
RenderX

Table 1. Demographic, app-related, headache-related, and mental well-being–related characteristics at baseline (N=198).

	Control group (n=88)	Intervention group (n=110)	*P* value[a]
Demographic characteristics			
Age (y), mean (SD)	38.28 (12.82)	39.03 (11.46)	.67
Gender, n (%)			.64
Man	8 (9.1)	12 (10.9)	
Woman	76 (86.4)	96 (87.3)	
Nonbinary	1 (1.1)	1 (0.9)	
No information	3 (3.4)	1 (0.9)	
Education, n (%)			.22
No education	2 (2.3)	0 (0)	
Obligatory or high school	1 (1.1)	6 (5.5)	
Vocational training and high school	26 (29.5)	34 (30.9)	
Higher vocational training	13 (14.8)	12 (10.9)	
University or University of Applied Sciences	46 (52.3)	58 (52.7)	
App related, n (%)			
Platform			.09
iOS (Apple Inc)	53 (60.2)	53 (48.2)	
Android (Google LLC)	35 (39.8)	57 (51.8)	
Chatbot coach			.62
Sophie (woman)	76 (86.4)	94 (85.6)	
David (man)	12 (13.6)	16 (14.5)	
Headache-related characteristics			
Diagnosis, n (%)			.82
Migraine	65 (73.9)	78 (70.9)	
TTH[b]	10 (11.4)	12 (10.9)	
No diagnosis	13 (14.8)	20 (18.2)	
Tracking app in parallel, n (%)			.82
Yes	43 (48.9)	52 (47.3)	
No	45 (51.1)	58 (52.7)	
Psychotherapy, n (%)			.49
Yes	16 (18.2)	16 (14.5)	
No	72 (81.8)	94 (85.5)	
Absenteeism: MIDAS[c] (d), mean (SD)[d]	6.95 (8.3)	6.37 (6.49)	.60
Presenteeism: MIDAS[c] (d), mean (SD)[d]	12.1 (9.99)	11.54 (9.54)	.70
Pain coping: FESV[e] score, mean (SD)			
Cognitive coping	41.06 (9.1)	39.52 (8.51)	.22
Behavioral coping	31.58 (7.55)	29.86 (7.73)	.19
Self-efficacy: HMSE-G-SF[f] Scale score, mean (SD)	22.77 (6.78)	24.37 (6.94)	.11
Application of behavior change techniques[g], mean (SD)	3.36 (0.97)	3.45 (1.04)	.09
Mental well-being, mean (SD)			
Depression: PHQ-9[h] score	9.43 (4.13)	8.69 (4.34)	.22

XSL•FO
RenderX

	Control group (n=88)	Intervention group (n=110)	*P* value[a]
Anxiety: GAD-7[i] score	7.25 (3.86)	6.25 (3.87)	.07
Somatic symptoms: PHQ-15[j] score	11.26 (4.28)	10.57 (4.4)	.27
Stress: PSS[k] score	30.32 (5.75)	29.8 (6.33)	.55

[a]Baseline group comparison between intervention and waitlist control groups using the 2-tailed *t* test and chi-square test.

[b]TTH: tension-type headache.

[c]MIDAS: Migraine Disability Assessment conducted for a period of 1 month.

[d]Waitlist control group: n=83 and intervention group: n=102 (outliers removed).

[e]FESV: questionnaire to assess pain management (Fragebogen zur Erfassung der Schmerzverarbeitung).

[f]HMSE-G-SF: German short form of the Headache Management Self-Efficacy.

[g]Application of behavior change techniques based on the health action process approach stages of change: 1 to 2=nonintenders, 3=intenders, and 4 to 5=actors.

[h]PHQ-9: Patient Health Questionnaire-9.

[i]GAD-7: General Anxiety Disorder Scale-7.

[j]PHQ-15: Patient Health Questionnaire-15.

[k]PSS: Perceived Stress Scale.

Effectiveness

Primary Outcome

Table 2 presents the results of the LMM analyses for the ITT and PP analyses. For both the ITT and PP analyses, we found evidence of a treatment effect (group by time interaction) for the PHQ-ADS after the intervention (t_{342}=–3.6; *P*<.001, 95% CI –5.06 to –1.47 and t_{294}=–3.58; *P*<.001, 95% CI –5.11 to –1.49, respectively). BalanceUP significantly affected mental well-being, as shown by the change in the PHQ-ADS. However, the waitlist control group did not improve with time (Cohen *d*=–0.07, 95% CI –0.23 to 0.08), the intervention group improved from before the intervention to after the intervention with a medium effect (Cohen *d*=0.62, 95% CI –0.84 to –10.39).

XSL•FO
RenderX

Table 2. Results of per-protocol (PP) and intention-to-treat (ITT) outcome analysis using linear mixed models[a,b].

Outcome	PP		ITT	
	β estimate (SE; 95% CI)	*P* value	β estimate (SE; 95% CI)	*P* value
Mental well-being (PHQ-ADS)[c]				
Intercept	17.01 (N/A[d])	N/A	16.68 (N/A)	N/A
Time[e]	–0.60 (0.67; –1.91 to 0.72)	.37	–0.50 (0.66; –1.80 to 0.80)	.45
Group[f]	–0.6 (0.38; –4.00 to 0.74)	.18	–1.75 (1.07; –3.84 to 0.35)	.10
Treatment[g]	–3.3 (0.93; –5.12 to –1.48)	<.001	–3.28 (0.91; –5.07 to –1.48)	<.001
Depression (PHQ-9[h])				
Intercept	9.68 (N/A)	N/A	9.43 (N/A)	N/A
Time	–0.33 (0.37; –1.06 to 0.40)	.37	–0.27 (0.36; –0.99 to 0.45)	.47
Group	–0.80 (0.70; –2.19 to 0.59)	.26	–0.74 (0.61; –1.95 to 0.47)	.23
Treatment	–1.78 (0.51; –2.79 to –0.77)	<.001	–1.80 (0.51; –2.79 to –0.80)	<.001
Anxiety (GAD-7[i])				
Intercept	7.33 (N/A)	N/A	7.25 (N/A)	N/A
Time	–0.26 (0.39; –1.04 to 0.51)	.50	–0.23 (0.39; –0.99 to 0.53)	.55
Group	–0.83 (0.60; –2.07 to 0.06)	.17	–1.01 (0.54; –2.06 to 0.06)	.06
Treatment	–1.52 (0.54; –2.59 to –0.45)	.006	–1.45 (0.53; –2.50 to –0.40)	.007
Somatic symptoms (PHQ-15[j])				
Intercept	11.42 (N/A)	N/A	11.26 (N/A)	N/A
Time	–4.44 (0.38; –0.75 to 0.75)	>.99	0.04 (0.38; –0.70 to 0.78)	.92
Group	–0.66 (0.72; –2.07 to 0.75)	.36	–0.69 (0.64; –1.95 to 0.58)	.29
Treatment	–2.33 (0.53; –3.37 to –1.30)	<.001	–2.33 (0.52; –3.35 to –1.30)	<.001
Stress (PSS-10[k])				
Intercept	30.74 (N/A)	N/A	30.32 (N/A)	N/A
Time	–0.90 (0.62; –2.13 to 0.33)	.15	–0.76 (0.62; –1.96 to 0.45)	.22
Group	–0.25 (1.03; –2.28 to 1.79)	.81	–0.52 (0.91; –2.31 to 1.27)	.57
Treatment	–2.69 (0.86; –4.39 to –0.99)	.002	–2.60 (0.85; –4.26 to –0.93)	.003
Headache management self-efficacy (HMSE-G-SF[l]Scale)				
Intercept	23.14 (N/A)	N/A	22.77 (N/A)	N/A
Time	–0.07 (0.73; –1.51 to 1.37)	.92	0.08 (0.72; –1.34 to 1.50)	.91
Group	1.37 (1.20; –0.80 to 3.52)	.26	1.60 (0.98; –0.34 to 3.54)	.11
Treatment	4.15 (1.01; 2.15 to 6.14)	<.001	4.05 (0.99; 2.10 to 6.00)	<.001
Application of behavior change techniques[m] (HAPA[n])				
Intercept	3.47 (N/A)	N/A	3.36 (N/A)	N/A
Time	0.01 (0.11; –0.22 to 0.23)	.93	–0.07 (0.12; –0.16 to 0.29)	.55
Group	0.21 (1.14; –0.08 to 0.49)	.15	0.09 (0.14; –0.18 to 0.36)	.51
Treatment	0.70 (0.16; 0.39 to 1.01)	<.001	0.76 (0.16; 0.45 to 1.07)	<.001
Absenteeism and presenteeism[o] (MIDAS[p])				
Intercept	20.10 (N/A)	N/A	19.05 (N/A)	N/A
Time	–0.90 (1.17; –3.21 to 1.41)	.45	–0.68 (1.15; –2.95 to 1.59)	.56

XSL•FO
RenderX

Outcome	PP		ITT	
	β estimate (SE; 95% CI)	*P* value	β estimate (SE; 95% CI)	*P* value
Group	–2.12 (2.58; –7.20 to 2.96)	.41	–1.14 (2.20; –5.46 to 3.18)	.61
Treatment	–4.61 (1.63; –7.83 to –1.39)	.005	–4.81 (1.60; –7.97 to –1.66)	.003
Cognitive pain coping (FESV[q])				
Intercept	40.69 (N/A)	N/A	41.06 (N/A)	N/A
Time	1.40 (0.83; –0.24 to 3.03)	.93	1.28 (0.82; –0.33 to 2.89)	.12
Group	–1.48 (1.42; –4.23 to 1.32)	.30	–1.54 (1.26; –4.02 to 0.94)	.22
Treatment	5.56 (1.14; 3.30 to 7.83)	<.001	5.58 (1.13; 3.37 to 7.80)	<.001
Behavioral pain coping (FESV)				
Intercept	31.72 (N/A)	N/A	31.58 (N/A)	N/A
Time	0.64 (0.71; –2.03 to 0.76)	.37	–0.59 (0.70; –1.97 to 0.79)	.40
Group	–1.40 (1.23; –3.84 to 1.03)	.26	–1.72 (1.10; –3.89 to 0.45)	.12
Treatment	4.91 (0.98; 2.87 to 6.84)	<.001	5.00 (0.97; 3.10 to 6.90)	<.001

[a]Outcomes of the PP analyses: only participants who completed the outro (ie, intended use).

[b]Outcomes of the ITT analyses: all participants who were randomized into groups.

[c]PHQ-ADS: Patient Health Questionnaire Anxiety and Depression Scale.

[d]N/A: not applicable.

[e]Time effect represents the rate of improvement for both intervention and waitlist control groups.

[f]Group effect represents intervention or waitlist control group.

[g]Treatment effect is represented by group and time interaction.

[h]PHQ-9: Patient Health Questionnaire-9.

[i]GAD-7: General Anxiety Disorder Scale-7.

[j]PHQ-15: Patient Health Questionnaire-15.

[k]PSS-10: Perceived Stress Scale-10.

[l]HMSE-G-SF: German short form of the Headache Management Self-Efficacy.

[m]Application of behavior change techniques, based on the health action process approach stages of change: 1 to 2=nonintenders, 3=intenders, 4 to 5=actors.

[n]HAPA: health action process approach.

[o]Assessment of absenteeism and presenteeism based on Migraine Disability Assessment, conducted for a period of 1 month.

[p]MIDAS: Migraine Disability Assessment.

[q]FESV: questionnaire to assess pain management (Fragebogen zur Erfassung der Schmerzverarbeitung).

Changes in the PHQ-ADS score differed significantly between groups with a medium effect (Cohen *d*=0.66, 95% CI –0.99 to –0.33); refer to Table 3 for observed means and effect sizes (Cohen *d*) for participants who completed the coaching intervention as intended (PP).

XSL•FO
RenderX

Table 3. Results of per-protocol outcome measures: means and effect sizes (Cohen *d*).

Measure	Before the intervention, mean (SD)	After the intervention, mean (SD)	Within-group effect (before the intervention vs after the intervention), Cohen *d*[a] (95% CI)	Between-group effect (intervention vs waitlist control group), Cohen *d* (95% CI)
Primary outcome				
Mental well-being (PHQ-ADS[b])				
Intervention (n=78)	15.38 (7.09)	11.49 (5.48)	–0.62 (–0.84 to –0.39)	–0.66 (–0.99 to –0.33)
Control (n=72)	17.01 (7.36)	16.42 (9.18)	–0.07 (–0.23 to 0.08)	N/A[c]
Secondary outcome				
Depression (PHQ-9[d])				
Intervention (n=78)	8.88 (4.19)	6.77 (3.55)	–0.54 (–0.74 to –0.35)	–0.59 (–0.91 to –0.26)
Control (n=72)	9.68 (4.27)	9.35 (5.15)	–0.07 (–0.23 to 0.09)	N/A
Anxiety (GAD-7[e])				
Intervention (n=78)	6.50 (3.64)	4.72 (2.41)	–0.58 (–0.83 to –0.32)	–0.66 (–0.98 to –0.33)
Control (n=72)	7.33 (3.91)	7.07 (4.53)	–0.06 (–0.24 to 0.12)	N/A
Somatic symptoms (PHQ-15[f])				
Intervention (n=78)	10.76 (4.03)	8.42 (3.66)	–0.61 (–0.82 to –0.39)	–0.65 (–0.98 to –0.32)
Control (n=72)	11.42 (4.26)	11.42 (5.46)	0.00 (–0.15 to 0.15)	N/A
Stress (PSS-10[g])				
Intervention (n=78)	30.49 (6.12)	26.9 (6.06)	–0.59 (–0.80 to –0.37)	–0.43 (–0.76 to –0.12)
Control (n=71)	30.66 (5.68)	29.79 (7.34)	–0.13 (–0.32 to 0.05)	N/A
Headache management self-efficacy (HMSE-G-SF[h] Scale)				
Intervention (n=78)	24.5 (9.79)	28.58 (7.21)	0.58 (0.34 to 0.82)	0.81 (0.48 to 1.14)
Control (n=71)	23.14 (6.37)	23.07 (6.34)	–0.01 (–0.20 to 0.18)	N/A
Application of behavior change techniques[i] (HAPA[j])				
Intervention (n=78)	3.45 (0.89)	4.38 (0.52)	1.28 (0.92 to 1.63)	1.05 (0.71 to 1.39)
Control (n=71)	3.46 (0.94)	3.48 (1.12)	0.02 (–0.20 to 0.24)	N/A
Absenteeism and presenteeism[k] (MIDAS[l])				
Intervention (n=71)	17.99 (14.18)	12.48 (12.05)	–0.42 (–0.60 to –0.23)	–0.45 (–0.79 to –0.12)
Control (n=67)	20.10 (16.48)	19.21 (17.36)	–0.05 (–0.18 to 0.07)	N/A
Cognitive pain coping (FESV[m])				
Intervention (n=78)	39.22 (7.98)	46.18 (8.58)	0.84 (0.59 to 1.09)	0.46 (0.14 to 0.79)
Control (n=71)	40.70 (9.29)	42.1 (9.02)	0.15 (0.00 to –0.30)	N/A
Behavioral pain coping (FESV)				
Intervention (n=78)	30.32 (7.4)	34.59 (7.58)	0.57 (0.36 to 0.78)	0.45 (0.15 to 0.78)
Control (n=71)	31.76 (7.45)	31.11 (7.83)	–0.08 (–0.25 to 0.08)	N/A

[a]Effect size according to Cohen *d*.

[b]PHQ-ADS: Patient Health Questionnaire Anxiety and Depression Scale.

[c]N/A: not applicable.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]GAD-7: General Anxiety Disorder Scale-7.

[f]PHQ-15: Patient Health Questionnaire-15.

[g]PSS-10: Perceived Stress Scale-10.

[h]HMSE-G-SF: German short form of the Headache Management Self-Efficacy.

XSL•FO
RenderX

[i]Application of behavior change techniques based on the health action process approach stages of change: 1 to 2=nonintenders, 3=intenders, 4 to 5=actors.

[j]HAPA: health action process approach.

[k]Assessment of absenteeism and presenteeism based on Migraine Disability Assessment conducted for a period of 1 month.

[l]MIDAS: Migraine Disability Assessment.

[m]FESV: questionnaire to assess pain management (Fragebogen zur Erfassung der Schmerzverarbeitung).

Secondary Outcomes

Regarding secondary outcomes, the ITT LMM analyses demonstrated evidence of treatment effects for depression (t_{342}=–3.56; *P*<.001, 95% CI –2.79 to –0.80), somatic symptoms (t_{348}=–4.48; *P*<.001, 95% CI –3.35 to –1.31), stress (t_{341}=–3.07; *P*=.003, 95% CI –4.25 to –0.94), headache-related self-efficacy (t_{342}=4.08; *P*<.001, 95% CI 2.10-5.99), application of BCTs (t_{342}=4.82; *P*<.001, 95% CI 0.45-1.07), presenteeism and absenteeism (t_{317}=–3.00; *P*=.003, 95% CI –7.96 to –1.68), cognitive pain coping (t_{341}=4.96; *P*<.001, 95% CI 3.38-7.79), behavioral coping (t_{341}=5.18; *P*<.001, 95% CI 3.11-6.89), and suggestive evidence for anxiety (t_{342}=–2.73; *P*=.007, 95% CI –2.50 to –0.40). The PP analyses showed similar results (Table 2). The effect sizes of secondary outcomes between groups after the intervention were medium (eg, depression, anxiety, somatic symptoms, stress, absenteeism and presenteeism, and pain coping) and large (eg, headache-related self-efficacy and application of BCTs). Refer to Table 3 for further details.

Predictors

We also explored whether diagnostic status (participants diagnosed with migraine vs participants with other or no headache-related diagnosis), concurrent psychotherapy, concurrent tracking of headaches, and headache-related self-efficacy influenced the pre- and postintervention effects. We did not find evidence of a 3-way interaction among group, time, and predictors. There was no evidence of a difference in the decrease of the PHQ-ADS score between the intervention and waitlist control groups for participants with a diagnosis of migraine (t_{338}=–0.81; *P*=.42, 95% CI –5.93 to 2.48), concurrent psychotherapy (t_{338}=–0.45; *P*=.65, 95% CI –5.69 to 3.57), concurrent headache tracking (t_{338}=–1.92; *P*=.06, 95% CI –6.99 to 0.09), and self-efficacy at baseline (t_{338}=–1.50; *P*=.14, 95% CI –0.48 to 0.06).

Engagement

Table 4 shows the rate of the intended use of the BalanceUP coaching app among participants who started module 1 (n=182), that is, participants from the intervention group (n=110) and participants who started coaching after the waiting time (n=72). As anticipated, the highest dropout rates occurred during module 1 (34/182, 18.7%), with a subsequent decrease in dropout rates during the subsequent modules. Of the 182 participants who began the coaching program with module 1, 118 (64.8%) completed the coaching and thus used it as intended. A visual inspection of the engagement data related to subjective experience revealed that participants who discontinued using the app did not show differences in active participation, internal engagement, perceived benefit, or comprehensibility compared with those who continued using BalanceUP (refer to Multimedia Appendix 6).

XSL•FO
RenderX

Table 4. Indicators of engagement: intended use, extent of use, and subjective experience of the BalanceUP app (n=182).

Indicators of engagement and acceptance	All participants who started with module 1[a]
Intended use of BalanceUP and use data, n (%)	
Started coaching (ie, intervention group and waitlist control group after a waiting time of 43 d)	182 (100)
Ceased interacting during module 1 (dropout)	34 (18.7)
Ceased interacting during module 2 (dropout)	11 (6)
Ceased interacting during module 3 (dropout)	6 (3.3)
Ceased interacting during module 4 (dropout)	4 (2.2)
Ceased interacting during module 5 (dropout)	2 (1.1)
Ceased interacting during module 6 (dropout)	1 (0.5)
Ceased interacting during module 7 (dropout)	6 (3.3)
Completed outro (intended use)	118 (64.8)
Extent of use among participants with intended use (n=117-112)	
Number of participants who completed all 7 modules, n (%)	101 (86.3)
Number of days to complete a module, mean (SD)	6.91 (1.50)
Ratio of reply in conversational turns, mean (SD)	77.84 (3.73)
Push reminders after 1 h of inactivity, mean (SD)	11.70 (5.99)
Push reminders after 2 h of inactivity, mean (SD)	8.95 (5.87)
Push reminders after 1 d of inactivity, mean (SD)	4.14 (3.62)
Push reminders after 3 d of inactivity, mean (SD)	1.46 (1.66)
Push reminders after 5 d of inactivity, mean (SD)	0.74 (1.12)
Email reminders after 10 d of inactivity, mean (SD)	0.34 (0.65)
Push reminders during survey, mean (SD)	0.75 (0.62)
Relaxation (audio listened in min), mean (SD)	113.44 (182.52)
Subjective experience among participants with intended use (n=117-112), mean (SD)	
Personal involvement during sessions[b]	4.30 (0.55)
Active participation during sessions[b]	4.27 (0.63)
Perceived comprehensibility of units[c]	3.87 (0.68)
Perceived benefit of units[c]	3.87 (0.68)
Commitment to change[d]	8.15 (5.46)
Perceived enjoyment[e]	4.13 (0.54)
Perceived alliance with the chatbot coach[f]	3.94 (0.81)
Acceptance of the BalanceUP app (n=117-114), mean (SD)	
Self-reported data	
Engagement (eg, entertainment and personalization)[g]	3.72 (0.65)
Information[g]	4.47 (0.47)
Perceived app quality[g]	3.56 (0.77)
Perceived impact of the app[g]	4.00 (0.62)

[a]Participants from the intervention group and waitlist control group who optionally participated after a waiting period.
[b]Measured after every session using 1 item of the Patient Session Evaluation Questionnaire.
[c]Measured randomly at the end of a unit using 1 items of the Patient Session Evaluation Questionnaire.
[d]Measured 3 times during coaching using the question “How committed are you to changing your behavior?”

XSL•FO
RenderX

[e]Measured randomly during coaching by a single item ("Did you enjoy the last unit?") from technology acceptance research.

[f]Measured 3 times during coaching using the Session Alliance Inventory Scale.

[g]Measured postcoaching intervention via Mobile Application Rating Scale.

We measured engagement among the participants with intended use based on the extent of use (use data) and subjective experience (self-reported data). The use data analysis showed that 86.3% (101/117) of the participants completed all 7 modules, taking an average of 6.9 (SD 1.5) days to work through a module. Participants replied in an average of 77.8% (SD 3.73%) conversational turns. In the event of inactivity, across the entire study period, participants were sent an average of 11.70 (SD 5.99) push notifications for no activity for a 1-hour period and 8.95 (SD 5.87) notifications for no activity for a 2-hour period, with a subsequent decrease in reminders sent for 1, 3, 5, and 10 days of no activity. Participants spent an average of 113.44 (SD 182.51) minutes on relaxation exercises. Refer to Table 4 for further details.

Participants self-reported a mean commitment to change their behavior of 8.15 (SD 5.46). The reported commitment to change behavior significantly increased with time ($F_{2,192}$=8.17; P<.001) with a medium effect (Cohen f=0.29), and evidence of higher commitment toward the end of the coaching (mean 8.17, SD 1.49) than in the middle (mean 7.57, SD 1.55, mean difference 0.60, 95% CI 0.31-0.89; P<.001) or beginning of the intervention (mean 7.61, SD 1.57, mean difference 0.56, 95% CI 0.21-0.91; P=.002). A linear regression (β=.05, SE 0.42; P=.67) with pre- to posttreatment PHQ-ADS changes showed that early reported commitment did not predict before improvement to after improvement of mental well-being. The mean perceived alliance with the CA was 3.94 (SD 0.82). The alliance significantly increased with time ($F_{2,200}$=10.66; P<.001) with a medium-large effect (Cohen f=0.32), with the evidence of alliance becoming higher toward the end of the coaching (mean 4.07, SD 0.83; P<.001) than at the beginning (mean 3.83, SD 0.83, mean difference 0.23, 95% CI 0.12-0.34; P<.001); refer to Table 5. A linear regression (β=.11, SE 0.80; P=.36) with pre- to posttreatment PHQ-ADS change showed that early reported alliance with the CA did not predict pre- to posttreatment improvement of mental well-being.

Table 5. Results of a repeated measure ANOVA for commitment to change and perceived alliance with the conversational agent (CA).

Outcome	Start of the intervention[a], mean (SD)	Midintervention[b], mean (SD)	End of the intervention[c], mean (SD)	*P* value	*F* test (*df*)
Commitment to change[d]	7.61 (1.57)	7.57 (1.55)	8.16 (1.49)	<.001	8.17 (2,192)
Perceived alliance with the CA[e]	3.84 (0.85)	4.02 (0.81)	4.07 (0.83)	<.001	10.66 (2,200)

[a]Measured during module 1 (commitment to change) and module 2 (alliance).

[b]Measured during module 3 (commitment to change) and module 4 (alliance).

[c]Measured during module 6.

[d]Measured by the question "How committed are you to changing your behavior" (n=97).

[e]Measured using Session Alliance Inventory Scale (n=101).

Participants with intended use rated the quality of BalanceUP in terms of engagement (eg, personalization and interactivity), information quality (eg, source and credibility), perceived subjective quality (eg, recommend app and pay for app), and perceived impact (eg, knowledge and awareness). The highest-rated subscale was information quality, with a mean of 4.47 (SD 0.47), followed by perceived impact (mean 4.00, SD 0.62), engagement (mean 3.72, SD 0.65), and subjective quality (mean 3.56, SD 0.77).

Impression of Change and Adverse Events

We found evidence of an improved perceived global impression of change in the Patient Global Impression of Change Scale from module 1 to module 7 (t_{101}=8.45; P<.001; Cohen d=–0.62, 95% CI –0.91 to –0.33), indicating an improvement in perceived general health. Of the participants who reported worsening of their general condition in module 1 (6/142, 4.2%), only 1 participant reported that this was because of the current coaching intervention and provided the following explanation:

> *Dealing with migraine triggers constant self-monitoring, which may cause slight discomfort and prevents me from feeling completely free from them. However, in the long term, it is worth pursuing education and behavior change.*

The remaining 5 participants reported that the worsening was because of other reasons, such as job and social conflicts. In module 7, only 2% (2/102) of the participants reported worsening of their general condition; however, all cited other reasons as the cause.

Triggers

Participants rated their sensitivity to trigger with a mean of 2.51 (SD 1.09) and avoidance with a mean of 2.32 (SD 1.07), indicating average sensitivity and avoidance compared with a sample of people who have headache [96].

Qualitative Evaluation

We further assessed the BalanceUP app's positive (Figure 7) and negative (Figure 8) aspects using qualitative content analysis. We used a deductive approach to identify themes based on insights from our pilot study and the existing literature. When examining the positive aspects, participants expressed a strong appreciation for the extensive and comprehensive information provided, which included various types of exercises and the

XSL•FO
RenderX

delivery mode of the intervention. When evaluating negative aspects, participants expressed a desire for more flexibility in their interactions with the CA.

Figure 7. The thematic map illustrates the positive aspects of BalanceUP. Numbers in parentheses indicate the frequency of mentions for each topic.

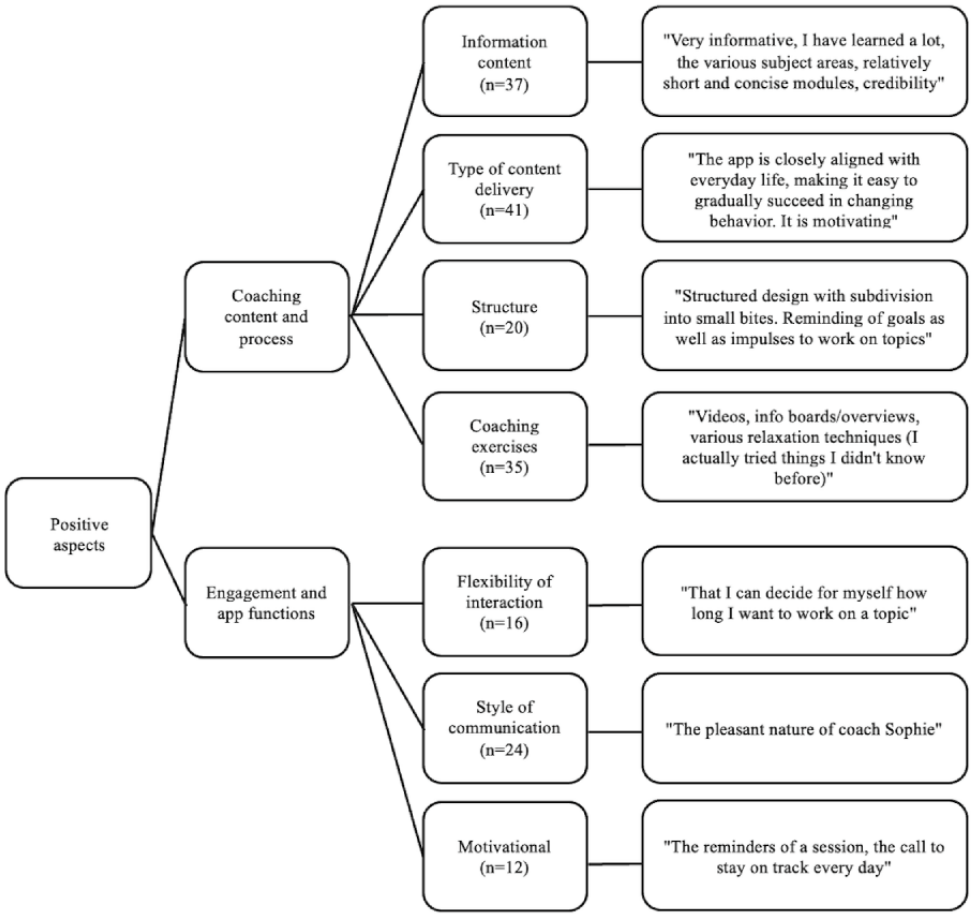

XSL•FO
RenderX

Figure 8. The thematic map illustrates suggestions to improve BalanceUP. Numbers in parentheses indicate the frequency of mentions for each topic.

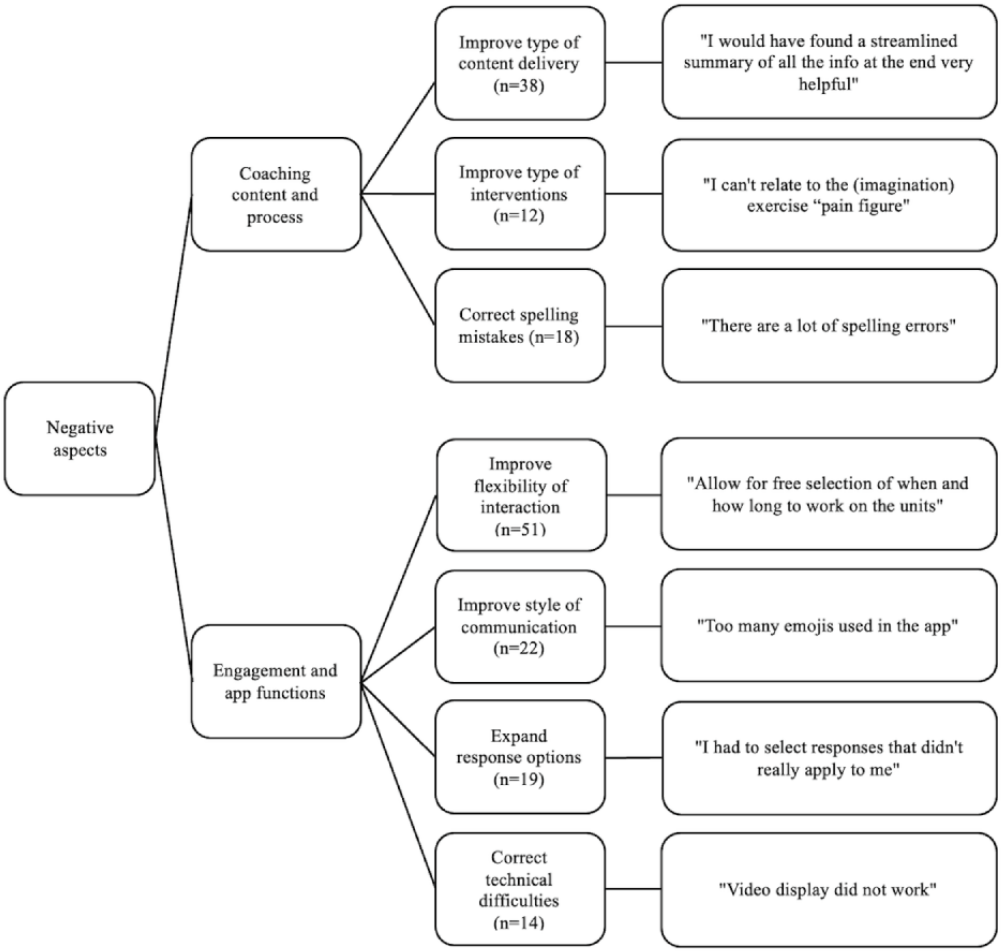

Discussion

Principal Findings

This study aimed to describe the development and evaluation of the effectiveness of the BalanceUP app. This is the first RCT of a fully unguided coaching intervention delivered by a rule-based CA to facilitate mental well-being in individuals with headaches. We described the BalanceUP app's evidence-based design and systematic evaluation.

With regard to effectiveness, we found evidence of improved mental well-being, as measured by the PHQ-ADS, in individuals with frequent headaches who received BalanceUP, with a medium to large between-group effect size (Cohen d=–0.66). Participants who interacted with BalanceUP experienced a clinically important improvement, reporting, on average, a 3.9 (SD 5.59) point reduction [85] of perceived depression and anxiety symptoms after the intervention. In contrast, participants in the waitlist control group did not show substantial changes in their mental well-being (0.6-point reduction on the PHQ-ADS score). Moreover, we found evidence of reduced anxiety, somatic symptoms, perceived stress, and absenteeism and presenteeism, as well as improved HMSE-G-SF, application of BCTs, and pain coping skills, with effects ranging from medium (Cohen d=0.43) to large (Cohen d=1.05). Diagnostic status, concurrent psychotherapy, concurrent tracking of headaches, and headache-related self-efficacy did not influence the effects of chatbot coaching.

No notable adverse effects were observed owing to the use of BalanceUP. Among the participants who initiated the coaching by starting the first module, a substantial portion (118/182, 64.8%) successfully adhered to the program. Participants who used BalanceUP as intended established a pronounced working alliance with the CA, which significantly improved with time.

In terms of acceptance, the program's information content received the highest rating, followed by the perceived impact, engagement, and subjective quality of BalanceUP. Its overall average rating, on a scale of 1 to 5 stars, was 3.91 (SD 0.67), indicating a high level of acceptance. Participants expressed their willingness to recommend the app to individuals who might benefit, with a mean rating of 4.02 (SD 1.05), reflected by the following statement in the uMARS questionnaire: "There are many people I would recommend this app to."

Comparison With Prior Work

Evidence on the effectiveness of mobile interventions for headaches is limited, despite their wide availability [47]. Minen et al [44] found that, despite a trend toward mHealth, most studies using electronic behavioral interventions to treat headaches did not use mobile devices. Only 1 single-arm study explored mHealth migraine behavioral therapy but with potential bias owing to missing diary entries. A recent study by Grazzi

XSL•FO
RenderX

et al [114] demonstrated the feasibility and effectiveness of an mHealth mindfulness program for chronic migraineurs (headaches for ≥15 d in a month), resulting in a 50% retrospective reduction in migraine days and medication intake.

With regard to internet-based headache treatments, 1 pilot study by Day et al [115] showed improvement in self-efficacy (Cohen *d*=0.82) and pain acceptance (Cohen *d*=0.82), although there was no evidence of headache reduction. An RCT by Bromberg et al [105] assessed a web-based intervention for migraine self-management and coping, revealing improvements in multiple outcomes, including depression, stress, headache self-efficacy, pain catastrophizing, and coping strategies, although without changes in anxiety and disability. Another study on web-based behavioral training for migraine self-management found no difference in headache attack frequency but found evidence for improved migraine self-efficacy and locus of control [116].

This study's results align with face-to-face behavior change interventions for headaches. Cognitive behavioral therapy, relaxation, or mindfulness interventions have effectively improved the cognitive, behavioral, anxiety, and stress-related aspects associated with headaches [30,117]. Notably, the evaluation of MIMA [72], the basis of our coaching, found limited improvement in headache-related outcomes, except for headache self-efficacy. Furthermore, BalanceUP was extended to address TTH (eg, onset of headache, course of headache, and medication). Although migraine was predominant among participants, we found no evidence that diagnostic status affected treatment outcomes. Tailoring allowed BalanceUP to effectively address both types of headaches.

There was no evidence of the influence of concurrent psychotherapy on coaching effectiveness in this study, which is noteworthy given prior suggestions of potential benefits of combining face-to-face and digital interventions [118,119]. The hypothesized role of headache-related self-efficacy in treatment outcome [120] remained unconfirmed in this study.

The BalanceUP app's evidence-based design and systematic evaluation contribute to the growing body of evidence on the acceptance and effectiveness of interventions involving CAs in health care. A recent meta-analysis examined 32 RCTs focusing on mental health and CA use [121]. Short-term effects on outcomes such as depressive and generalized anxiety symptoms, quality of life, well-being, and stress were found, with effect sizes ranging from Hedges *g*=0.24 to Hedges *g*=0.62. However, the long-term effects remained unclear. Personalization and empathetic responses emerged as crucial effectiveness facilitators, with longer CA interactions linked to larger effect sizes. These findings align with those of another meta-analysis [122], showing the effectiveness of CAs in outcomes related to lifestyle changes, smoking cessation, substance misuse, and medication adherence. Notably, <50% of the participants reported overall satisfaction with the CA, content likeability, and future use. However, many studies had a pre- and postintervention design or were feasibility trials, indicating the need for further RCTs in this field. Abd-Alrazaq et al [123] evaluated the effectiveness of CAs in 12 clinical and nonclinical RCTs. Weak evidence of reduced depression, stress, or agoraphobia rates, but not improved mental well-being, was found. However, bias, low-quality evidence, small sample size, or contradictory results limited the conclusions, necessitating further high-quality RCTs following the guidelines.

Contextual factors, such as psychological traits, motivation, personal relevance, and attributes of digital behavior change interventions themselves (eg, content, reminders, delivery, support, and personalization) influencing engagement [64]. BalanceUP integrates personalization and tailoring (eg, relevant topic selection and adjustment of interaction length) to empower participants and promote a sense of control and ownership over the coaching process. Emotional support, encouragement, or validation from the CA also expresses empathy, reinforcing participants' feelings (eg, "This is excellent news, well done, keep on track" and "I am sorry to hear, but setbacks are also part of a change process"). Qualitative feedback confirms program flexibility and suggests that addressing individual preferences can enhance satisfaction and further improve the coaching experience.

In BalanceUP, we observed that 65% of the participants used coaching as intended by completing the outro. In behavioral headache treatment, engagement has yet to be thoroughly assessed in terms of dose and duration; however, earlier studies reported high dropout rates [124].

Consistent with previous research [67,125], participants using BalanceUP established a strong alliance with the CA, which improved significantly with time. This finding aligns with studies conducted in in-person, digital, and group settings [126-128] and with individuals with recurrent headaches [129]. However, contradictory to findings in internet-based therapy studies [130,131], the alliance was unrelated to improvements in participants' mental well-being. These results suggest that, although the alliance between participants and the CA was established and strengthened throughout the coaching program, other factors may significantly influence treatment effectiveness.

Limitations and Future Work

This study had several limitations. First, guidelines for trials of behavioral headache treatments [88] recommend using headache frequency as the primary outcome. However, they also urge investigators to use standardized disability, functional status, or quality of life measures. A recent Delphi study by Leudtke et al [132] emphasized the need for outcome measures that truly reflect patients' experiences. Therefore, the inclusion of functional disability, quality of life, and associated symptoms should be considered in nonpharmacological interventions. BalanceUP aimed to capture the biopsychosocial impact of headaches by addressing various lifestyle factors.

Second, the self-selection of the participants limits the generalizability of our findings and introduces potential self-selection bias. Participants' particular interest in the subject matter might make them nonrepresentative of the broader population. Caution is needed when extending these findings to a broader context because of the possibility of differing preexisting characteristics.

Third, it is important to acknowledge potential improvements for enhanced interactions in BalanceUP. The current rule-based

XSL•FO
RenderX

nature of the CA allowed for the implementation of an evidence-based program. However, participants' desire for more flexible interactions with the ability to input personal responses was evident. Previous research [133] indicates user preference for a combination of predefined answer options and text input to enhance perceived interactivity. To enhance text processing in BalanceUP, integrating artificial intelligence (AI)–based technology, such as large language models or natural language processing, could be considered. Natural language processing and large language models enable the CA to interpret user inputs more dynamically, yet existing AI-based chatbots struggle with unforeseen user responses [133,134]. Furthermore, ethical aspects related to AI technology should be considered, as they could lead to misjudgments and potential risks. Research has highlighted a lack of transparency in describing the handling of input data and algorithms, affecting the reliability and validity of findings [122,135]. One approach might involve using AI technology for specific tasks (eg, providing content that humans will select or tailor) while maintaining a rule-based approach for other tasks (eg, handling sensitive information and following guidelines), ensuring predictability and preventing harm.

Fourth, it is worth noting that most participants in this study were women (86%), in line with the higher prevalence of headache in women [136]. Approximately half of them had university degrees, consistent with our pilot study [67]. However, the trial's sociographic distribution may not fully represent the general population. Future studies should seek greater diversity using a more representative sample.

Fifth, a potential limitation is the digital placebo effect [137]. In this unblinded trial, participants might attribute improvements to using an mHealth intervention rather than interventionist ingredients. Expectations and engagement could bias outcomes positively. Future research should carefully design control conditions, considering active control groups or sham interventions [138].

Sixth, participants needing a smartphone introduced a limitation [139]. Although mHealth usability is favored by a substantial proportion of middle-aged individuals with headache in German-speaking countries, the exclusion of smartphone-less individuals may impact generalizability [4]. Ownership rates are high, for example, 92% in Switzerland and 91% in Europe [140].

Seventh, a small number of participants (28/198, 14.1%) in this study encountered technical issues, such as missing audio tones in videos or loss of internet connection. Although technical support was provided to address these problems, it is likely that these technical issues may have had a negative impact on participants' engagement.

Finally, the primary objective of developing this app was to enhance the mental well-being outcomes of individuals with headache. The app was specifically designed for this research project. Currently, it is freely available for public use. However, the future availability of the app is uncertain because of ongoing support costs and a lack of collaboration with potential providers. We recognize the potential value of making such interventions accessible in the future, and discussions regarding their availability are ongoing. It is essential to note that the university's role is not to provide the app as a service but to contribute through accompanying research. The sustainability of digital health interventions, particularly with limited public health funding, remains challenging [141,142]. Clear cost-benefit outcomes and accountability strategies should be addressed in future research.

Conclusions

This study represents the first empirical evaluation of an evidence-based and CA-delivered coaching intervention (BalanceUP) designed to promote a healthy lifestyle among individuals with headaches. The findings provide evidence of the positive impact of BalanceUP on participants' mental well-being, as indicated by improvements in depression and anxiety symptoms. The establishment of a strong alliance between participants and the CA, along with their high commitment to the program, further reinforces the potential effectiveness of this intervention. The scalability and accessibility of automated coaching through a CA highlight its potential as an engaging and effective tool for behavior change. Further research is needed to explore the long-term effects, intensity, and duration of delivery of coaching interventions for lifestyle modifications in the health sector. Therefore, the applicability of mHealth interventions in diverse populations should be investigated. Understanding these aspects will contribute to the development of effective and inclusive interventions that promote positive health outcomes.

Acknowledgments

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

SU designed the study and was supervised by HK and ARG. SU drafted the ethics application and registered with the trial. SU oversaw the design of the coaching program, with assistance from VZ. Technical implementation was a collaborative effort supported from Zurich University of Applied Sciences (ZHAW) and Swiss Federal Institute of Technology Zurich (ETH). Statistical analyses were performed by SU and supervised by HK, including consultation with a statistician from ZHAW.

XSL•FO
RenderX

Interpretation and discussion of the results were led by SU in consultation with TK and ARG. SU wrote the manuscript, and TK critically reviewed the entire manuscript. All the authors approved the final manuscript.

Conflicts of Interest

TK is affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics at Swiss Federal Institute of Technology Zurich; and the Institute of Technology Management and School of Medicine at the University of St. Gallen. The Centre for Digital Health Interventions is funded in part by CSS, a Swiss health insurer; Mavie Next, an Austrian health insurer; and MTIP, a Swiss digital health investor. TK is also a cofounder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, neither CSS nor Pathmate Technologies were involved in this research. ARG has no conflicts of interest to report to this publication. He is a board member of the Swiss Headache Society. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Overview of the coaching intervention.
[PDF File (Adobe PDF File), 151 KB - jmir_v26i1e50132_app1.pdf]

Multimedia Appendix 2
Intervention flow.
[PDF File (Adobe PDF File), 112 KB - jmir_v26i1e50132_app2.pdf]

Multimedia Appendix 3
Behavior change techniques implemented.
[PDF File (Adobe PDF File), 148 KB - jmir_v26i1e50132_app3.pdf]

Multimedia Appendix 4
Aspects of engagement.
[PDF File (Adobe PDF File), 228 KB - jmir_v26i1e50132_app4.pdf]

Multimedia Appendix 5
Outcomes and time points.
[PDF File (Adobe PDF File), 123 KB - jmir_v26i1e50132_app5.pdf]

Multimedia Appendix 6
Engagement with subjective experience dropout versus intended use.
[PDF File (Adobe PDF File), 96 KB - jmir_v26i1e50132_app6.pdf]

Multimedia Appendix 7
CONSORT-EHEALTH Checklist.
[PDF File (Adobe PDF File), 763 KB - jmir_v26i1e50132_app7.pdf]

References

1. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017 Sep 16;390(10100):1211-1259 [FREE Full text] [doi: 10.1016/S0140-6736(17)32154-2] [Medline: 28919117]
2. Steiner TJ, Stovner LJ, Birbeck GL. Migraine: the seventh disabler. J Headache Pain 2013 Jan 10;14(1):1 [FREE Full text] [doi: 10.1186/1129-2377-14-1] [Medline: 23566305]
3. Steiner TJ, Terwindt GM, Katsarava Z, Pozo-Rosich P, Gantenbein AR, Roche SL, et al. Migraine-attributed burden, impact and disability, and migraine-impacted quality of life: expert consensus on definitions from a Delphi process. Cephalalgia 2022 Nov;42(13):1387-1396 [FREE Full text] [doi: 10.1177/03331024221110102] [Medline: 35791285]
4. Lampl C, Thomas H, Stovner LJ, Tassorelli C, Katsarava Z, Laínez JM, et al. Interictal burden attributable to episodic headache: findings from the Eurolight project. J Headache Pain 2016;17:9 [FREE Full text] [doi: 10.1186/s10194-016-0599-8] [Medline: 26879832]
5. Berg J, Stovner LJ. Cost of migraine and other headaches in Europe. Eur J Neurol 2005 Jun;12 Suppl 1(s1):59-62. [doi: 10.1111/j.1468-1331.2005.01192.x] [Medline: 15877781]

XSL•FO
RenderX

6. Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, CDBE2010Study Group. Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 2011 Oct;21(10):718-779. [doi: 10.1016/j.euroneuro.2011.08.008] [Medline: 21924589]
7. Sokolovic E, Riederer F, Szucs T, Agosti R, Sándor PS. Self-reported headache among the employees of a Swiss university hospital: prevalence, disability, current treatment, and economic impact. J Headache Pain 2013 Mar 26;14(1):29 [FREE Full text] [doi: 10.1186/1129-2377-14-29] [Medline: 23565579]
8. Linde M, Gustavsson A, Stovner LJ, Steiner TJ, Barré J, Katsarava Z, et al. The cost of headache disorders in Europe: the Eurolight project. Eur J Neurol 2012 May;19(5):703-711. [doi: 10.1111/j.1468-1331.2011.03612.x] [Medline: 22136117]
9. Amoozegar F, Khan Z, Oviedo-Ovando M, Sauriol S, Rochdi D. The burden of illness of migraine in Canada: new insights on humanistic and economic cost. Can J Neurol Sci 2022 Mar;49(2):249-262. [doi: 10.1017/cjn.2021.75] [Medline: 33875042]
10. Hasenbring MI, Korb J, Pfingsten M. Psychologische Mechanismen der Chronifizierung ? Konsequenzen für die Prävention. In: Kröner-Herwig B, Frettlöh J, Klinger R, Nilges P, editors. Schmerzpsychotherapie: Grundlagen – Diagnostik – Krankheitsbilder – Behandlung. Berlin, Germany: Springer; 2017:115-131.
11. Turk DC, Okifuji A. Psychological factors in chronic pain: evolution and revolution. J Consult Clin Psychol 2002 Jun;70(3):678-690. [doi: 10.1037//0022-006x.70.3.678] [Medline: 12090376]
12. Martin PR. Stress and primary headache: review of the research and clinical management. Curr Pain Headache Rep 2016 Jul 23;20(7):45. [doi: 10.1007/s11916-016-0576-6] [Medline: 27215628]
13. Klan T, Liesering-Latta E. Kognitiv-verhaltenstherapeutisches Migränemanagement (MIMA): Ein Behandlungsmanual zur Krankheitsbewältigung und Attackenprophylaxe bei Migräne. Göttingen, Germany: Hogrefe Verlag; 2020.
14. Fritsche G, Gaul C. Migräne. In: Kröner-Herwig B, Frettlöh J, Klinger R, Nilges P, editors. Schmerzpsychotherapie: Grundlagen - Diagnostik - Krankheitsbilder - Behandlung. Berlin, Germany: Springer; 2017:475-502.
15. French DJ, Holroyd KA, Pinell C, Malinoski PT, O'Donnell F, Hill KR. Perceived self-efficacy and headache-related disability. Headache 2000 Sep;40(8):647-656. [doi: 10.1046/j.1526-4610.2000.040008647.x] [Medline: 10971661]
16. Martin NJ, Holroyd KA, Penzien DB. The headache-specific locus of control scale: adaptation to recurrent headaches. Headache 1990 Nov;30(11):729-734. [doi: 10.1111/j.1526-4610.1990.hed3011729.x] [Medline: 2074167]
17. Pellegrino AB, Davis-Martin RE, Houle TT, Turner DP, Smitherman TA. Perceived triggers of primary headache disorders: a meta-analysis. Cephalalgia 2018 May;38(6):1188-1198 [FREE Full text] [doi: 10.1177/0333102417727535] [Medline: 28825314]
18. Limbrecht-Ecklundt K, Bischoff C, Traue HC. Kopfschmerz vom Spannungstyp. In: Kröner-Herwig B, Frettlöh J, Klinger R, Nilges P, editors. Schmerzpsychotherapie: Grundlagen – Diagnostik – Krankheitsbilder – Behandlung. Berlin, Germany: Springer; 2017:451-474.
19. Rasmussen BK. Migraine and tension-type headache in a general population: precipitating factors, female hormones, sleep pattern and relation to lifestyle. Pain 1993 Apr;53(1):65-72. [doi: 10.1016/0304-3959(93)90057-V] [Medline: 8316392]
20. Nash JM, Thebarge RW. Understanding psychological stress, its biological processes, and impact on primary headache. Headache 2006 Oct;46(9):1377-1386. [doi: 10.1111/j.1526-4610.2006.00580.x] [Medline: 17040334]
21. Seng EK, Martin PR, Houle TT. Lifestyle factors and migraine. Lancet Neurol 2022 Oct;21(10):911-921. [doi: 10.1016/S1474-4422(22)00211-3] [Medline: 36115363]
22. Diener HC, Gaul C, Kropp P. Therapie der migräneattacke und prophylaxe der migräne. AWMF Online. 2018. URL: https://register.awmf.org/assets/guidelines/030-057l_S1_Migraene-Therapie_2020-12.pdf [accessed 2023-12-28]
23. Kropp P, Meyer B, Dresler T, Fritsche G, Gaul C, Niederberger U, et al. Entspannungsverfahren und verhaltenstherapeutische interventionen zur behandlung der migräne: leitlinie der Deutschen migräne- und kopfschmerzgesellschaft. Schmerz 2017 Oct 31;31(5):433-447. [doi: 10.1007/s00482-017-0214-1] [Medline: 28364171]
24. Fritsche G, Kröner-Herwig B, Kropp P, Niederberger U, Haag G. Psychologische therapie der migräne: systematische übersicht. Schmerz 2013 Jun 19;27(3):263-274. [doi: 10.1007/s00482-013-1319-9] [Medline: 23685993]
25. Lüking M, Martin A. Entspannung, Im agination, Biofeedback und Meditation. In: Kröner-Herwig B, Frettlöh J, Klinger R, Nilges P, editors. Schmerzpsychotherapie: Grundlagen - Diagnostik - Krankheitsbilder - Behandlung. Berlin, Germany: Springer; 2017:303-324.
26. Roditi D, Robinson ME. The role of psychological interventions in the management of patients with chronic pain. Psychol Res Behav Manag 2011;4:41-49 [FREE Full text] [doi: 10.2147/PRBM.S15375] [Medline: 22114534]
27. Frettlöh J, Hermann C. Kognitiv-behaviorale therapie. In: Kröner-Herwig B, Frettlöh J, Klinger R, Nilgesditors P, editors. Schmerzpsychotherapie: Grundlagen - Diagnostik - Krankheitsbilder - Behandlung. Berlin, Germany: Springer; 2017:349-371.
28. Gaul C, van Doorn C, Webering N, Dlugaj M, Katsarava Z, Diener HC, et al. Clinical outcome of a headache-specific multidisciplinary treatment program and adherence to treatment recommendations in a tertiary headache center: an observational study. J Headache Pain 2011 Aug 5;12(4):475-483 [FREE Full text] [doi: 10.1007/s10194-011-0348-y] [Medline: 21544647]
29. Wallasch TM, Kropp P. Multidisciplinary integrated headache care: a prospective 12-month follow-up observational study. J Headache Pain 2012 Oct;13(7):521-529 [FREE Full text] [doi: 10.1007/s10194-012-0469-y] [Medline: 22790281]

XSL•FO
RenderX

30. Pérez-Muñoz A, Buse DC, Andrasik F. Behavioral interventions for migraine. Neurol Clin 2019 Nov;37(4):789-813. [doi: 10.1016/j.ncl.2019.07.003] [Medline: 31563233]
31. Schaetz L, Rimner T, Pathak P, Fang J, Chandrasekhar D, Mueller J, et al. Employee and employer benefits from a migraine management program: disease outcomes and cost analysis. Headache 2020 Oct;60(9):1947-1960 [FREE Full text] [doi: 10.1111/head.13933] [Medline: 32799346]
32. Cavanagh K, Strauss C, Forder L, Jones F. Can mindfulness and acceptance be learnt by self-help?: a systematic review and meta-analysis of mindfulness and acceptance-based self-help interventions. Clin Psychol Rev 2014 Mar;34(2):118-129 [FREE Full text] [doi: 10.1016/j.cpr.2014.01.001] [Medline: 24487343]
33. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
34. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279:131-142. [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
35. Kabat-Zinn J, Lipworth L, Burney R. The clinical use of mindfulness meditation for the self-regulation of chronic pain. J Behav Med 1985 Jun;8(2):163-190. [doi: 10.1007/BF00845519] [Medline: 3897551]
36. Wong SY, Chan FW, Wong RL, Chu MC, Kitty Lam YY, Mercer SW, et al. Comparing the effectiveness of mindfulness-based stress reduction and multidisciplinary intervention programs for chronic pain: a randomized comparative trial. Clin J Pain 2011 Oct;27(8):724-734. [doi: 10.1097/AJP.0b013e3182183c6e] [Medline: 21753729]
37. Rosenzweig S, Greeson JM, Reibel DK, Green JS, Jasser SA, Beasley D. Mindfulness-based stress reduction for chronic pain conditions: variation in treatment outcomes and role of home meditation practice. J Psychosom Res 2010 Jan;68(1):29-36. [doi: 10.1016/j.jpsychores.2009.03.010] [Medline: 20004298]
38. Day MA, Thorn BE, Rubin NJ. Mindfulness-based cognitive therapy for the treatment of headache pain: a mixed-methods analysis comparing treatment responders and treatment non-responders. Complement Ther Med 2014 Apr;22(2):278-285. [doi: 10.1016/j.ctim.2013.12.018] [Medline: 24731899]
39. Simshäuser K, Lüking M, Kaube H, Schultz C, Schmidt S. Is mindfulness-based stress reduction a promising and feasible intervention for patients suffering from migraine? A randomized controlled pilot trial. Complement Med Res 2020;27(1):19-30. [doi: 10.1159/000501425] [Medline: 31390617]
40. Wells RE, O'Connell N, Pierce CR, Estave P, Penzien DB, Loder E, et al. Effectiveness of mindfulness meditation vs headache education for adults with migraine: a randomized clinical trial. JAMA Intern Med 2021 Mar 01;181(3):317-328 [FREE Full text] [doi: 10.1001/jamainternmed.2020.7090] [Medline: 33315046]
41. Lake AE, Rains JC, Penzien DB, Lipchik GL. Headache and psychiatric comorbidity: historical context, clinical implications, and research relevance. Headache 2005 May 09;45(5):493-506. [doi: 10.1111/j.1526-4610.2005.05101.x] [Medline: 15953266]
42. Andrasik F, Wittrock DA, Passchier J. PsychologicalMechanisms of tension-type headaches. International Headache Society. 2005. URL: https://ihs-headache.org/wp-content/uploads/2020/06/Andrasik.pdf [accessed 2023-12-14]
43. Wells RE, Smitherman TA, Seng EK, Houle TT, Loder EW. Behavioral and mind/body interventions in headache: unanswered questions and future research directions. Headache 2014 Jun 15;54(6):1107-1113. [doi: 10.1111/head.12362] [Medline: 24735261]
44. Minen MT, Torous J, Raynowska J, Piazza A, Grudzen C, Powers S, et al. Electronic behavioral interventions for headache: a systematic review. J Headache Pain 2016;17:51 [FREE Full text] [doi: 10.1186/s10194-016-0608-y] [Medline: 27160107]
45. Stubberud A, Linde M. Digital technology and mobile health in behavioral migraine therapy: a narrative review. Curr Pain Headache Rep 2018 Jul 31;22(10):66. [doi: 10.1007/s11916-018-0718-0] [Medline: 30066141]
46. Göbel H, Frank B, Heinze A, Zimmermann W, Göbel C, Göbel A, et al. Gesundheitsverhalten von Migräne- und kopfschmerzpatienten bei digitaler therapiebegleitung mit der Migräne-app. Schmerz 2019 Apr;33(2):147-155. [doi: 10.1007/s00482-018-0355-x] [Medline: 30649625]
47. Mosadeghi-Nik M, Askari MS, Fatehi F. Mobile health (mHealth) for headache disorders: a review of the evidence base. J Telemed Telecare 2016 Dec;22(8):472-477. [doi: 10.1177/1357633X16673275] [Medline: 27799450]
48. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
49. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
50. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 09;19(5):e151 [FREE Full text] [doi: 10.2196/jmir.6553] [Medline: 28487267]
51. Stieger M, Flückiger C, Rüegger D, Kowatsch T, Roberts BW, Allemand M. Changing personality traits with the help of a digital personality change intervention. Proc Natl Acad Sci U S A 2021 Feb 23;118(8):e2017548118 [FREE Full text] [doi: 10.1073/pnas.2017548118] [Medline: 33558417]

XSL•FO
RenderX

52. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
53. Gardiner PM, McCue KD, Negash LM, Cheng T, White LF, Yinusa-Nyahkoon L, et al. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Educ Couns 2017 Sep;100(9):1720-1729 [FREE Full text] [doi: 10.1016/j.pec.2017.04.015] [Medline: 28495391]
54. Kowatsch T, Volland D, Shih I, Rüegger D, Künzler F, Barata F, et al. Design and evaluation of a mobile chat app for the open source behavioral health intervention platform MobileCoach. In: Proceedings of the 12th International Conference on Designing the Digital Transformation. 2017 Presented at: DESRIST '17; May 30-June 1, 2017; Karlsruhe, Germany p. 485-489 URL: https://link.springer.com/chapter/10.1007/978-3-319-59144-5_36 [doi: 10.1007/978-3-319-59144-5_36]
55. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
56. Pereira J, Díaz Ó. Using health chatbots for behavior change: a mapping study. J Med Syst 2019 Apr 04;43(5):135. [doi: 10.1007/s10916-019-1237-1] [Medline: 30949846]
57. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005 Oct;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
58. Bickmore T, Schulman D, Yin L. Maintaining engagement in long-term interventions with relational agents. Appl Artif Intell 2010 Jul 01;24(6):648-666 [FREE Full text] [doi: 10.1080/08839514.2010.492259] [Medline: 21318052]
59. Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol 1989 Apr;36(2):223-233 [FREE Full text] [doi: 10.1037/0022-0167.36.2.223]
60. Flückiger C, Del Re AC, Wampold BE, Symonds D, Horvath AO. How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. J Couns Psychol 2012 Jan;59(1):10-17. [doi: 10.1037/a0025749] [Medline: 21988681]
61. Laux G. Online-/internet-programme zur psychotherapie bei depression – eine zwischenbilanz // online-/internet-based psychological therapies for depression – a summarizing report. J Neurol Neurosurg Psychiatry 2017;18(1):16-24 [FREE Full text]
62. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
63. Karekla M, Kasinopoulos O, Neto DD, Ebert DD, Van Daele T, Nordgreen T, et al. Best practices and recommendations for digital interventions to improve engagement and adherence in chronic illness sufferers. Eur Psychol 2019 Jan;24(1):49-67 [FREE Full text] [doi: 10.1027/1016-9040/a000349]
64. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
65. Baumel A, Birnbaum ML, Sucala M. A systematic review and taxonomy of published quality criteria related to the evaluation of user-facing eHealth programs. J Med Syst 2017 Aug 22;41(8):128. [doi: 10.1007/s10916-017-0776-6] [Medline: 28735372]
66. Bickmore TW, Kimani EN, Trinh H, Pusateri A, Paasche-Orlow MK, Magnani JW. Managing chronic conditions with a smartphone-based conversational virtual agent. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. 2018 Presented at: IVA '18; November 5-8, 2018; Sydney, Australia p. 119-124 URL: https://dl.acm.org/doi/10.1145/3267851.3267908 [doi: 10.1145/3267851.3267908]
67. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 03;8(4):e15806 [FREE Full text] [doi: 10.2196/15806] [Medline: 32242820]
68. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019 Dec 2;2(1):118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
69. Home page. MobileCoach. URL: https://www.mobile-coach.eu/ [accessed 2023-12-14]
70. Filler A, Kowatsch T, Haug S, Wahle F, Staake T, Fleisch E. MobileCoach: a novel open source platform for the design of evidence-based, scalable and low-cost behavioral health interventions: overview and preliminary evaluation in the public health context. In: Proceedings of the 2015 Wireless Telecommunications Symposium. 2015 Presented at: WTS '15; October 15-17, 2015; New York, NY p. 1-6 URL: https://ieeexplore.ieee.org/document/7117255 [doi: 10.1109/wts.2015.7117255]
71. Klan T, Liesering-Latta E, Gaul C, Martin PR, Witthöft M. An integrative cognitive behavioral therapy program for adults with migraine: a feasibility study. Headache 2019 May 10;59(5):741-755. [doi: 10.1111/head.13532] [Medline: 30970172]
72. Klan T, Gaul C, Liesering-Latta E, Both B, Held I, Hennemann S, et al. Efficacy of cognitive-behavioral therapy for the prophylaxis of migraine in adults: a three-armed randomized controlled trial. Front Neurol 2022 Apr 28;13:852616 [FREE Full text] [doi: 10.3389/fneur.2022.852616] [Medline: 35572937]
73. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change

XSL•FO
RenderX

interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
74. Michie S, Wood CE, Johnston M, Abraham C, Francis JJ, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technol Assess 2015 Nov;19(99):1-188 [FREE Full text] [doi: 10.3310/hta19990] [Medline: 26616119]
75. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
76. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
77. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
78. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
79. Kowatsch T, Schachner T, Harperink S, Barata F, Dittler U, Xiao G, et al. Conversational agents as mediating social actors in chronic disease management involving health care professionals, patients, and family members: multisite single-arm feasibility study. J Med Internet Res 2021 Feb 17;23(2):e25060 [FREE Full text] [doi: 10.2196/25060] [Medline: 33484114]
80. Stieger M, Nißen M, Rüegger D, Kowatsch T, Flückiger C, Allemand M. PEACH, a smartphone- and conversational agent-based coaching intervention for intentional personality change: study protocol of a randomized, wait-list controlled trial. BMC Psychol 2018 Sep 04;6(1):43 [FREE Full text] [doi: 10.1186/s40359-018-0257-9] [Medline: 30180880]
81. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
82. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017 Dec 06;19(12):e402 [FREE Full text] [doi: 10.2196/jmir.8578] [Medline: 29212630]
83. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
84. von Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY: The Guilford Press; 2017.
85. Kroenke K, Wu J, Yu Z, Bair MJ, Kean J, Stump T, et al. Patient health questionnaire anxiety and depression scale (PHQ-ADS): initial validation in three clinical trials. Psychosom Med 2016;78(6):716-727 [FREE Full text] [doi: 10.1097/PSY.0000000000000322] [Medline: 27187854]
86. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
87. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
88. Penzien DB. Guidelines for trials of behavioral treatments for recurrent headache: purpose, process, and product. Headache 2005 May;45 Suppl 2:S87-S89. [doi: 10.1111/j.1526-4610.2005.4502001.x] [Medline: 15921504]
89. Kroenke K, Spitzer RL, Williams JB. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. Psychosom Med 2002;64(2):258-266. [doi: 10.1097/00006842-200203000-00008] [Medline: 11914441]
90. Toussaint A, Hüsing P, Kohlmann S, Löwe B. Detecting DSM-5 somatic symptom disorder: criterion validity of the Patient Health Questionnaire-15 (PHQ-15) and the Somatic Symptom Scale-8 (SSS-8) in combination with the Somatic Symptom Disorder - B Criteria Scale (SSD-12). Psychol Med 2020 Jan;50(2):324-333. [doi: 10.1017/S003329171900014X] [Medline: 30729902]
91. Klein EM, Brähler E, Dreier M, Reinecke L, Müller KW, Schmutzer G, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16:159 [FREE Full text] [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
92. Graef JE, Rief W, French DJ, Nilges P, Nestoriuc Y. German language adaptation of the headache management self-efficacy scale (HMSE-G) and development of a new short form (HMSE-G-SF). Headache 2015 Apr 22;55(7):958-972. [doi: 10.1111/head.12564] [Medline: 25904007]
93. Schwarzer R. Modeling health behavior change: how to predict and modify the adoption and maintenance of health behaviors. Appl Psychol 2008 Jan 30;57(1):1-29 [FREE Full text] [doi: 10.1111/j.1464-0597.2007.00325.x]

XSL•FO
RenderX

94. Benz T, Lehmann S, Gantenbein AR, Sandor PS, Stewart WF, Elfering A, et al. Translation, cross-cultural adaptation and reliability of the German version of the migraine disability assessment (MIDAS) questionnaire. Health Qual Life Outcomes 2018 Mar 09;16(1):42 [FREE Full text] [doi: 10.1186/s12955-018-0871-5] [Medline: 29523138]
95. Geisser E. Fragebogen zur Erfassung der Schmerzverarbeitung. Göttingen, Germany: Hogrefe; 2001.
96. Caroli A, Klan T, Gaul C, Liesering-Latta E, Martin PR, Witthöft M. Die erfassung von triggerempfindlichkeit und -vermeidung bei primären kopfschmerzen: entwicklung und psychometrische erprobung einer deutschen version des Headache Triggers Sensitivity and Avoidance Questionnaire (HTSAQ-G). Verhaltenstherapie 2019 Jul 2;30(3):234-245. [doi: 10.1159/000501218]
97. Zoubek K. Prozessevaluation einer kognitiv-verhaltenstherapeutischen gruppentherapie bei diabetes und depression : entwicklung und validierung der patienten- und therapeuten- gruppentherapiestundenbögen (GTS-P, GTS-T, GTS-TP) zur vorhersage des therapieerfolgs. Johannes Gutenberg-Universität Mainz. 2013. URL: https://openscience.ub.uni-mainz.de/handle/20.500.12030/1758 [accessed 2023-12-28]
98. van der Heijden H. User acceptance of hedonic information systems. MIS Q 2004;28(4):695-704 [FREE Full text] [doi: 10.2307/25148660]
99. Kamis A, Koufaris M, Stern T. Using an attribute-based decision support system for user-customized products online: an experimental investigation. MIS Q 2008 Mar;32(1):159-177 [FREE Full text] [doi: 10.2307/25148832]
100. Falkenström F, Hatcher RL, Skjulsvik T, Larsson MH, Holmqvist R. Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. Psychol Assess 2015 Mar;27(1):169-183 [FREE Full text] [doi: 10.1037/pas0000038] [Medline: 25346997]
101. Falkenström F, Hatcher RL, Skjulsvik T, Larsson MH, Holmqvist R. Session alliance inventory—patient version. APA PsycTests. 2015. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft39594-000 [accessed 2023-12-14]
102. Wilmers F, Munder T, Leonhart R, Herzog T, Plassmann R, Barth J, et al. Die deutschsprachige Version des Working Alliance Inventory: short revised (WAI-SR): ein schulenübergreifendes, ökonomisches und empirisch validiertes Instrument zur erfassung der therapeutischen allianz. Klin Diagnostik u Evaluation 2018;1(3):343-358.
103. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
104. Ferguson L, Scheman J. Patient global impression of change scores within the context of a chronic pain rehabilitation program. J Pain 2009 Apr;10(4):S73. [doi: 10.1016/j.jpain.2009.01.258]
105. Bromberg J, Wood ME, Black RA, Surette DA, Zacharoff KL, Chiauzzi EJ. A randomized trial of a web-based intervention to improve migraine self-management and coping. Headache 2012 Feb;52(2):244-261 [FREE Full text] [doi: 10.1111/j.1526-4610.2011.02031.x] [Medline: 22413151]
106. Wells RE, Burch R, Paulsen RH, Wayne PM, Houle TT, Loder E. Meditation for migraines: a pilot randomized controlled trial. Headache 2014 Oct;54(9):1484-1495. [doi: 10.1111/head.12420] [Medline: 25041058]
107. Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting group medical visits with conversational agents for stress management behavior change. In: Proceedings of the 12th International Conference on Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors. 2017 Presented at: PERSUASIVE '17; April 4-6, 2017; Amsterdam, The Netherlands p. 55-67 URL: https://link.springer.com/chapter/10.1007/978-3-319-55134-0_5 [doi: 10.1007/978-3-319-55134-0_5]
108. Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. J Biopharm Stat 2001;11(1-2):9-21. [doi: 10.1081/BIP-100104194] [Medline: 11459446]
109. Siddiqui O, Hung HM, O'Neill R. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. J Biopharm Stat 2009 Feb 12;19(2):227-246. [doi: 10.1080/10543400802609797] [Medline: 19212876]
110. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol 2014 Nov 19;14:118 [FREE Full text] [doi: 10.1186/1471-2288-14-118] [Medline: 25407057]
111. Mayring P. Qualitative content analysis. Forum Qual Soc Res 2000;1(2):251-257. [doi: 10.17169/fqs-1.2.1089]
112. Mayring P. Qualitative inhaltsanalyse. In: Flick U, von Kardoff E, Keupp H, Rosenstiel L, Wolff S, editors. Handbuch Qualitative Sozialforschung: Grundlagen, Konzepte, Methoden und Anwendungen. Weinheim, Germany: Beltz; 1991:209-213.
113. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
114. Grazzi L, Rizzoli P, Andrasik F. Effectiveness of mindfulness by smartphone, for patients with chronic migraine and medication overuse during the COVID-19 emergency. Neurol Sci 2020 Dec 13;41(Suppl 2):461-462 [FREE Full text] [doi: 10.1007/s10072-020-04659-0] [Medline: 32794128]
115. Day MA, Thorn BE, Ward LC, Rubin N, Hickman SD, Scogin F, et al. Mindfulness-based cognitive therapy for the treatment of headache pain: a pilot study. Clin J Pain 2014 Feb;30(2):152-161. [doi: 10.1097/AJP.0b013e318287a1dc] [Medline: 23446085]

XSL•FO
RenderX

116. Kleiboer A, Sorbi M, van Silfhout M, Kooistra L, Passchier J. Short-term effectiveness of an online behavioral training in migraine self-management: a randomized controlled trial. Behav Res Ther 2014 Oct;61:61-69. [doi: 10.1016/j.brat.2014.07.009] [Medline: 25133856]
117. Raggi A, Grignani E, Leonardi M, Andrasik F, Sansone E, Grazzi L, et al. Behavioral approaches for primary headaches: recent advances. Headache 2018 Jun;58(6):913-925. [doi: 10.1111/head.13337] [Medline: 29802634]
118. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017 Sep 15;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
119. Berger T, Bur O, Krieger T. Internet-interventionen in der psychotherapie. Psychother Psychosom Med Psychol 2019 Oct 10;69(9-10):413-426. [doi: 10.1055/a-0963-9055] [Medline: 31600812]
120. Nicholson RA, Hursey KG, Nash JM. Moderators and mediators of behavioral treatment for headache. Headache 2005 May;45(5):513-519. [doi: 10.1111/j.1526-4610.2005.05103.x] [Medline: 15953268]
121. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862 [FREE Full text] [doi: 10.2196/43862] [Medline: 37115595]
122. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
123. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021 [FREE Full text] [doi: 10.2196/16021] [Medline: 32673216]
124. Gewirtz A, Minen M. Adherence to behavioral therapy for migraine: knowledge to date, mechanisms for assessing adherence, and methods for improving adherence. Curr Pain Headache Rep 2019 Jan 19;23(1):3 [FREE Full text] [doi: 10.1007/s11916-019-0739-3] [Medline: 30661135]
125. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Form Res 2021 May 11;5(5):e27868 [FREE Full text] [doi: 10.2196/27868] [Medline: 33973854]
126. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022 Apr 11;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
127. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working alliance inventory-short revised (WAI-SR): psychometric properties in outpatients and inpatients. Clin Psychol Psychother 2010 May;17(3):231-239. [doi: 10.1002/cpp.658] [Medline: 20013760]
128. Jasper K, Weise C, Conrad I, Andersson G, Hiller W, Kleinstäuber M. The working alliance in a randomized controlled trial comparing internet-based self-help and face-to-face cognitive behavior therapy for chronic tinnitus. Internet Interv 2014 Apr;1(2):49-57. [doi: 10.1016/j.invent.2014.04.002]
129. Trautmann E, Kröner-Herwig B. A randomized controlled trial of Internet-based self-help training for recurrent headache in childhood and adolescence. Behav Res Ther 2010 Jan;48(1):28-37. [doi: 10.1016/j.brat.2009.09.004] [Medline: 19782343]
130. Bergman Nordgren L, Carlbring P, Linna E, Andersson G. Role of the working alliance on treatment outcome in tailored internet-based cognitive behavioural therapy for anxiety disorders: randomized controlled pilot trial. JMIR Res Protoc 2013 Jan 18;2(1):e4 [FREE Full text] [doi: 10.2196/resprot.2292] [Medline: 23612437]
131. Meyer B, Bierbrodt J, Schröder J, Berger T, Beevers CG, Weiss M, et al. Effects of an internet intervention (Deprexis) on severe depression symptoms: randomized controlled trial. Internet Interv 2015 Mar;2(1):48-59 [FREE Full text] [doi: 10.1016/j.invent.2014.12.003]
132. Luedtke K, Basener A, Bedei S, Castien R, Chaibi A, Falla D, et al. Outcome measures for assessing the effectiveness of non-pharmacological interventions in frequent episodic or chronic migraine: a Delphi study. BMJ Open 2020 Feb 12;10(2):e029855 [FREE Full text] [doi: 10.1136/bmjopen-2019-029855] [Medline: 32051295]
133. Lim SM, Shiau CW, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
134. Chan WW, Fitzsimmons-Craft EE, Smith AC, Firebaugh ML, Fowler LA, DePietro B, et al. The challenges in designing a prevention chatbot for eating disorders: observational study. JMIR Form Res 2022 Jan 19;6(1):e28003 [FREE Full text] [doi: 10.2196/28003] [Medline: 35044314]
135. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
136. Stovner LJ, Andree C. Prevalence of headache in Europe: a review for the Eurolight project. J Headache Pain 2010 Aug;11(4):289-299 [FREE Full text] [doi: 10.1007/s10194-010-0217-0] [Medline: 20473702]
137. Torous J, Firth J. The digital placebo effect: mobile mental health meets clinical psychiatry. Lancet Psychiatry 2016 Feb;3(2):100-102. [doi: 10.1016/S2215-0366(15)00565-9] [Medline: 26851322]

XSL•FO
RenderX

138. Lutz J, Offidani E, Taraboanta L, Lakhan SE, Campellone TR. Appropriate controls for digital therapeutic clinical trials: a narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. Front Digit Health 2022;4(6):823977-823950 [FREE Full text] [doi: 10.3389/fdgth.2022.823977] [Medline: 36060538]
139. Vogel MM, Combs SE, Kessel KA. mHealth and application technology supporting clinical trials: today's limitations and future perspective of smartRCTs. Front Oncol 2017 Mar 13;7:37 [FREE Full text] [doi: 10.3389/fonc.2017.00037] [Medline: 28348978]
140. Grampp M, Brandes D. Smartphones - bane or boon? Opportunities and risks for customers, consumers and companies. Deloitte. 2019. URL: https://www2.deloitte.com/ch/en/pages/technology-media-and-telecommunications/articles/global-mobile-consumer-survey.html [accessed 2023-12-15]
141. Iyamu I, Gómez-Ramírez O, Xu AX, Chang HJ, Watt S, Mckee G, et al. Challenges in the development of digital public health interventions and mapped solutions: findings from a scoping review. Digit Health 2022;8:20552076221102255 [FREE Full text] [doi: 10.1177/20552076221102255] [Medline: 35656283]
142. Wienert J, Zeeb H. Implementing health apps for digital public health - an implementation science approach adopting the consolidated framework for implementation research. Front Public Health 2021;9:610237 [FREE Full text] [doi: 10.3389/fpubh.2021.610237] [Medline: 34026702]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique
CA: conversational agent
CONSORT: Consolidated Standards of Reporting Trials
GAD-7: General Anxiety Disorder Scale-7
HMSE-G-SF: German short form of the Headache Management Self-Efficacy
ITT: intention-to-treat
LMM: linear mixed model
mHealth: mobile health
MIMA: Cognitive-Behavioral Therapy for Migraine Management (Kognitiv-verhaltenstherapeutisches Migränemanagement)
PHQ-15: Patient Health Questionnaire-15
PHQ-9: Patient Health Questionnaire-9
PHQ-ADS: Patient Health Questionnaire Anxiety and Depression Scale
PP: per-protocol
RCT: randomized controlled trial
TTH: tension-type headache
uMARS: translated version of the Mobile App Rating Scale

Edited by T de Azevedo Cardoso; submitted 27.06.23; peer-reviewed by C Gaul, R Davis-Martin, P Wicks; comments to author 16.08.23; revised version received 20.09.23; accepted 12.12.23; published 24.01.24.

Please cite as:
Ulrich S, Gantenbein AR, Zuber V, Von Wyl A, Kowatsch T, Künzli H
Development and Evaluation of a Smartphone-Based Chatbot Coach to Facilitate a Balanced Lifestyle in Individuals With Headaches (BalanceUP App): Randomized Controlled Trial
J Med Internet Res 2024;26:e50132
URL: https://www.jmir.org/2024/1/e50132
doi:10.2196/50132
PMID:38265863

©Sandra Ulrich, Andreas R Gantenbein, Viktor Zuber, Agnes Von Wyl, Tobias Kowatsch, Hansjörg Künzli. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Group–Based Intervention for Physical Activity Promotion Among Thai Adults During the COVID-19 Lockdown: Randomized Controlled Trial

Nanthawan Pomkai[1], MA; Piyawat Katewongsa[1], PhD; Aphichat Chamratrithirong[1], PhD; Kanokwan Tharawan[1], PhD; Teeranong Sakulsri[1], PhD; Bhubate Samutachak[1], PhD; Dyah Anantalia Widyastari[1], PhD; Niramon Rasri[2], MBA; Boonyanuch Wijarn[3], MEng; Yodchanan Wongsawat[4], PhD

[1]Institute for Population and Social Research, Mahidol University, Nakhon Pathom, Thailand

[2]Thai Health Promotion Foundation, Bangkok, Thailand

[3]Popmed Co, Ltd, Bangkok, Thailand

[4]Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand

Corresponding Author:
Piyawat Katewongsa, PhD
Institute for Population and Social Research
Mahidol University
Salaya Campus
999 Phutthamonthon 4 Road
Nakhon Pathom, 73170
Thailand
Phone: 66 875459998
Email: piyawat.kat@mahidol.edu

Abstract

Background: The COVID-19 pandemic significantly diminished the physical activity (PA) level of Thai adults belonging to Generation Y (Gen Y). As a response to the global crisis, many individuals worldwide have turned to social community platforms, recognizing their potential in promoting PA during the pandemic. Gen Y, in particular, demonstrates exceptional proficiency in using social media platforms, showcasing a remarkable aptitude for swiftly accessing new information and knowledge. However, their proclivity for reckless behavior exposes them to various health risks, potentially leading to enduring adverse health consequences. Consequently, there arises a pressing need to develop a comprehensive model aimed at elevating the PA levels among individuals belonging to Gen Y.

Objective: This research aimed to examine the effectiveness of a digital group–based activity in promoting PA among Gen Y in Thailand.

Methods: This was a parallel 2-arm randomized controlled trial with single-blind allocation to experimental and control groups and pre- and posttest measurements. Measurements were administered on the web and were designed for respondents to complete by themselves. The sample comprised 100 Gen Y individuals who met the inclusion criteria. Both groups were matched for background characteristics. The two 8-week intervention activities were (1) two weeks of education and (2) six weeks of motivation by target groups that set goals for PA together (using the Zoom meeting application), with a time limit and group consensus as to when the goal was to be achieved. The intervention activities were implemented one by one at specified intervals and delivered daily through health apps and the official LINE account.

Results: The intervention starts from August 22 to October 16, 2021. Of the 100 participants, 20 (20%) left the study, and the remaining 80 (80%) participated in the study (40 individuals each in the experimental and control groups). After participating in the experiment, a statistically significant difference in PA was found between the 2 groups (moderate to vigorous PA; 25/40, 63%; P=.03). Participants in the intervention group collected a higher cumulative minute of moderate to vigorous PA weekly (283 minutes) than those in the control group (164 minutes), and this was statistically significant (P=.03). For the transition to the fourth stage of behavior (ie, action), the improvement in the experimental group, after participating in the trial, was statistically significant compared to that of the control group (P=.01).

XSL•FO
RenderX

Conclusions: Digital group–based activity showed its effectiveness in improving the PA of Gen Y individuals in the intervention group. It created a process-based intervention activity that corresponds to the stages of behavior changes, from contemplation to action. The digital community can also connect individuals to comparable groups locally and globally.

Trial Registration: Thai Clinical Trials Registry TCTR20211101005; https://www.thaiclinicaltrials.org/show/TCTR20211101005

(J Med Internet Res 2024;26:e43366) doi:10.2196/43366

KEYWORDS

physical activity; digital group–based activity; collective action; Thai adults; COVID-19 pandemic; COVID-19; design; model; effectiveness; digital; Thailand; behavior

Introduction

The importance of physical activity (PA) for health has been documented worldwide. For adults, regular and adequate PA has been associated with a lower risk of premature deaths due to noncommunicable diseases, lower incidence of hypertension, type 2 diabetes, and cancer; improved cognitive function; and better mental health outcomes [1,2]. Nevertheless, although these benefits have been acknowledged and public health promotion efforts have been undertaken, physical inactivity remains a global public health problem [3].

The COVID-19 pandemic has forced nationwide lockdowns, and most of the countries use social distancing measures and policies to encourage people to stay at home to reduce the spread of the disease [4]. These measures have restricted people's movement and led to a reduction in PA globally [5,6]. With limited opportunity to engage in regular outdoor PA during containment periods, lifestyle changes are required. Health promotion strategies have then focused on the message of promoting home-based PA and encouraging population of all ages to stay active while at home [7]. The World Health Organization (WHO) initiated the "Healthy at Home" campaign, and it was followed by many countries with a similar focus: encouraging home-based PA to regenerate the PA of the population [7-10].

The promotion of home-based PA was purposefully executed through social media platforms and group-based activities. This involved recommending home-based exercises; providing sample training videos; and offering free web-based exercise classes on platforms such as YouTube, Facebook, and Twitter. Access was facilitated through social media platforms or applications, with the primary goal of helping individuals maintain a healthy weight, reduce the risk of chronic diseases, and foster social interaction [2,8,10-12]. A few studies have documented the positive effect of home-based PA promotion (eg, Fit From Home in improving the PA of the population) [8,12].

While the actual community interaction was limited during the pandemic, the digital community has been increasingly demanding and becoming an everyday part of every individual's life [13]. This community is a place for people to gather on the web, who may or may not meet one another face-to-face, to stay in touch with or perform certain activities in the groups they belong to [14]. The digital engagement creates and maintains relationships between group members in the digital group–based activity, affecting both psychological and social motivation, and this can lead to behavior modification of members within the group [8,13].

With more than half of the people globally now using social community [15], digital group–based activity (a digital community platform) has the potential to promote PA during the pandemic [8,10-12,16]. Unlike the mass campaigns that are addressed to the general population, information delivered in the digital group–based activity targeted specific groups that share similar interest. Good engagement within the group increases the effectiveness of the messages in driving the behavior changes of its members [12,17,18].

Adults who belong to the Generation Y (Gen Y) are a group of population with high literation in technology or digital native generation and are those who use IT as part of their daily life [19-21]. Gen Y in Thailand (those born between 1981 and 2000) [22-26] who have access to the internet, comprising a total of 17,644,290 (26.6%) people [16], are the most active in using the social media platform compared to other working-age groups [16]. They are exceptionally quick to access new information and knowledge and share this at lightning speed among their social networks. They are creative and have high self-esteem, and they prefer working in teams or working together in groups [27]. However, it is also the group that has the opportunity to contract various diseases due to reckless behavior, and that hazard may have lifelong adverse health consequences [28]. Therefore, it is necessary to find specific measures to encourage PA, especially for those who are undecided about changing their behavior.

Previous studies reported a high probability of behavior changes if appropriate intervention and health promotion efforts are undertaken [18]. Collective action theory has identified mechanisms to increase motivation through the creation of a common goal and by setting conditions for achieving that goal with others [29]. By providing education and motivation through a social media platform, the use of group processes should be able to induce behavioral changes and help promote the adoption of healthier lifestyles. While the collective action theory was useful in developing the model intervention, the transtheoretical model (TTM) is relevant in describing the stages of behavioral change [30,31]. This theory argues that an individual's health behavior change can be observed by following the stages of "precontemplation, contemplation, preparation, action, maintenance, and termination" [31]. By understanding which stage an individual is in, an appropriate intervention could be tailored. Guided by the 2 theories, this study aimed to examine the effectiveness of digital group–based activity to improve

XSL•FO
RenderX

Thai adults' PA with a focus on Gen Y by considering the context of their lifestyle and habits. The findings of this study should be useful in providing information for decision-making in promoting PA of adults in the future.

Methods

Study Design

A parallel 2-arm randomized controlled trial design with single-blind assignment of participants to experimental and control groups was used. Outcomes were measured by pre- and posttest assessments (pre- and posttest design with the control group). Matching between the experimental and control groups was performed by aligning the individual characteristics, for example, sex, age group, and primary occupation, by using probability random sampling. While the intervention group received a set of interventions, the control group performed their routines without any intervention.

Population and Sample

The population universe in this research is Thai adults who belong to Gen Y and who have access to the internet, or a total of 17,644,290 people. Gen Y was defined as a group of people born between 1981 and 2000 [22-26]. This segment of the population is critical for the future development of the country, both in the labor market and in terms of the economy (insofar as the Gen Y group dominates the spending power of the country) [32]. This digital population was randomly sampled (with probability known) to provide information on Thailand's PA surveillance system. The prescribed sample size was calculated by considering the value α=.05, effect size=0.5 (a medium effect size from the previous study [33]), power=0.95, and a 2-tailed *t* test. To reduce attrition due to withdrawal from the study or refusal to participate, the sample size was increased from the minimum prescribed number (54 persons), resulting in a total sample of 100 people, evenly divided into experimental and control groups.

Eligible participants were those who met the inclusion criteria of having insufficient PA as recommended by the WHO [34] and were in the second stage of behavior change (ie, contemplation) because 1 in 3 Gen Y individual is in this stage [35]. Simple random sampling with replacement was used to randomly assign the participants to the experimental group or control group. In case of sample rejection, a suitable replacement was selected.

Randomization

Once the sample size was known, the researcher conducted random sampling to obtain the participants in the research by concealment of allocation sequences, and participants did not know that they were assigned to the experimental or control group. The steps are as follows:

Step 1 involved screening samples to participate in the experiment. The population universe was Thai individuals of all genders and ages who had access to the internet at the time of the survey. The sample was selected through multistage random sampling. For this study, the population of interest was members of Gen Y. Eligible participants were screened according to the control variables, namely, sex, age group, and primary occupation. The inclusion criteria were those who had insufficient PA as recommended by the WHO and were in the second stage of behavior change (ie, contemplation).

In step 2, simple random sampling with replacement was used to randomly select between the experimental group and control group. At this stage, the prior random individual was assigned to the experimental group, and then, another of its pair would be assigned to the control group. This loop proceeded until all required samples were matched as planned.

In step 3, after the sampling process was completed, the consent and voluntary participation were inquired. In case of sample rejection, the new random sample with the same eligible participants and method was used for replacement.

The Intervention: Digital Group–Based Activity

Overview

For the prototype model in this study, the intervention covered 8 weeks of activity (from August 22 to October 16, 2021). The designed interventions consisted of education and motivation (as per theory), with the intervention activities implemented one by one at specified intervals, delivered daily through social media platforms (health apps and the official LINE account [LY Corporation]). The "Light-hearted" app (Popmed Co Ltd) was designed as a mobile-based app considering that Gen Y individuals are the heavy users of digital media. This app recorded the number of steps and other PAs including biking, swimming, walking, boxing, volleyball, and badminton. Participants were also requested to record their daily calorie intake. Additional features including health information and feedback mechanism (ie, suggestion box) were also made available to improve participants' engagement apart from the chat box (Table 1).

XSL•FO
RenderX

Table 1. The intervention program for the experimental group was sent through the health apps "Light-hearted" and LINE.

Week and topic	Activity
Intervention 1: Providing knowledge to increase awareness	
Week 1: Exposed to PA[a] education messages	• PA self-assessment • The importance of PA • Knowledge of the PA • Advantages and disadvantages of the original behavior • Cobenefits of sufficient PA • Dramatic relief
Week 2: The guideline practice of various PAs	• Information about alternatives to PA and PA for the working-age population (at home or around the house) • Information about alternatives to PA at work and in public areas
Intervention 2: Creating motivation	
Week 3-6: Collective motivation of the group by encouraging the determination to reach the set goals	• Setting group goals • Target groups set goals for PA together with a time limit and group consensus • Start accumulating steps through the health app • The group's cumulative steps were reported every 3 days including sending motivation messages to groups

[a]PA: physical activity.

Providing Knowledge to Increase Awareness

This part of the intervention lasted 2 weeks. This activity aimed to build knowledge and understand the guidelines on PA for the intervention group. In the first week, participants were exposed to health education messages (infographic and video clips) sent through the health apps "Light-hearted " and LINE. The health apps "Light-hearted" and LINE were installed after the intervention group consented to participate in the intervention. The messages covered PA self-assessment, the importance of PA, advantages and disadvantages of the original behavior, cobenefits of sufficient PA, and dramatic relief. In the second week, the health education message described the practice of various PAs. This included information about alternatives to PA and PA for the working-age population at home or around the house, at work, and in public areas. The intention was to add a variety of options for PA. The video was produced by Social Marketing Team of ThaiHealth Promotion Foundation, was in the public domain (SocialMarketingTH YouTube channel), and was free to be used. The videos contained messages on the causes of noncommunicable diseases, shared experiences, and motivation for regular exercise.

Creating Motivation

A total of 6 weeks was spent to apply the collective action theory, which was the main process hypothesized to drive behavior change in this study. Target groups set goals for PA together, with a time limit and group consensus as to when the goal was to be achieved. The collective action theory prescribes both internal and external motivation. Both types of motivation were directly targeted by the intervention activity and indirectly targeted through discussion and encouragement via group chat as an additional impetus to behavior change. The process started with setting group goals by using Zoom (Zoom Video Communications) meetings to discuss PA goals in the entire experimental group. Ultimately, it was agreed that all participants would accumulate steps through the health app "Light-hearted." The group's cumulative steps were reported every 3 days. At the end of the intervention, if the goal of the activity was achieved, then the research project would make a donation to charity as promised.

Measurement

A structured questionnaire was used as the data collection tool. PA and sedentary behavior were assessed by using the Global Physical Activity Questionnaire (version 2), while behavior assessment used the standardized and validated TTM questionnaire [35]. Sufficient PA was calculated following the guideline that includes (1) work-related activity, (2) traveling from place to place, and (3) leisure time activity [34]. The questions were self-administered on the web in the LimeSurvey web application (LimeSurvey) and were designed appropriately for the respondents to complete by themselves (ie, PA).

Data Analysis

Univariate analysis was used to describe the characteristics and distribution of the sample, such as percentage, mean, median, and range. Bivariate analysis was used to test for significant association of variables and group-level differences. To test the effectiveness of the model, the mean before-and-after values in the same group were compared by the paired 2-tailed *t* test, while the means between the 2 groups were compared by the independent samples 2-tailed *t* test.

Ethical Considerations

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study protocol was reviewed and approved by the Research Ethics Committee of the Institutional Review Board, Institute for Population and Social

XSL•FO
RenderX

Research, and Mahidol University on April 28, 2021 (project code COA 2021/03-048) and registered on the Thai Clinical Trials Registry (TCTR20211101005). All participants agreed in advance to provide data to the researchers and were assured that all their information would be kept confidential. Participants provided their consent to be included in the study by clicking on the agreement box available in the LimeSurvey web application. No personal information of the participant was collected or used in a way that would identify the data provider. As compensation, participants in the intervention group who completed two rounds received a T-shirt valued 250 Thai Baht (US $7). If the intervention was successful in the experimental group, the intervention would also be offered to the control group participants (Figure 1).

Figure 1. Process of the intervention of this study.

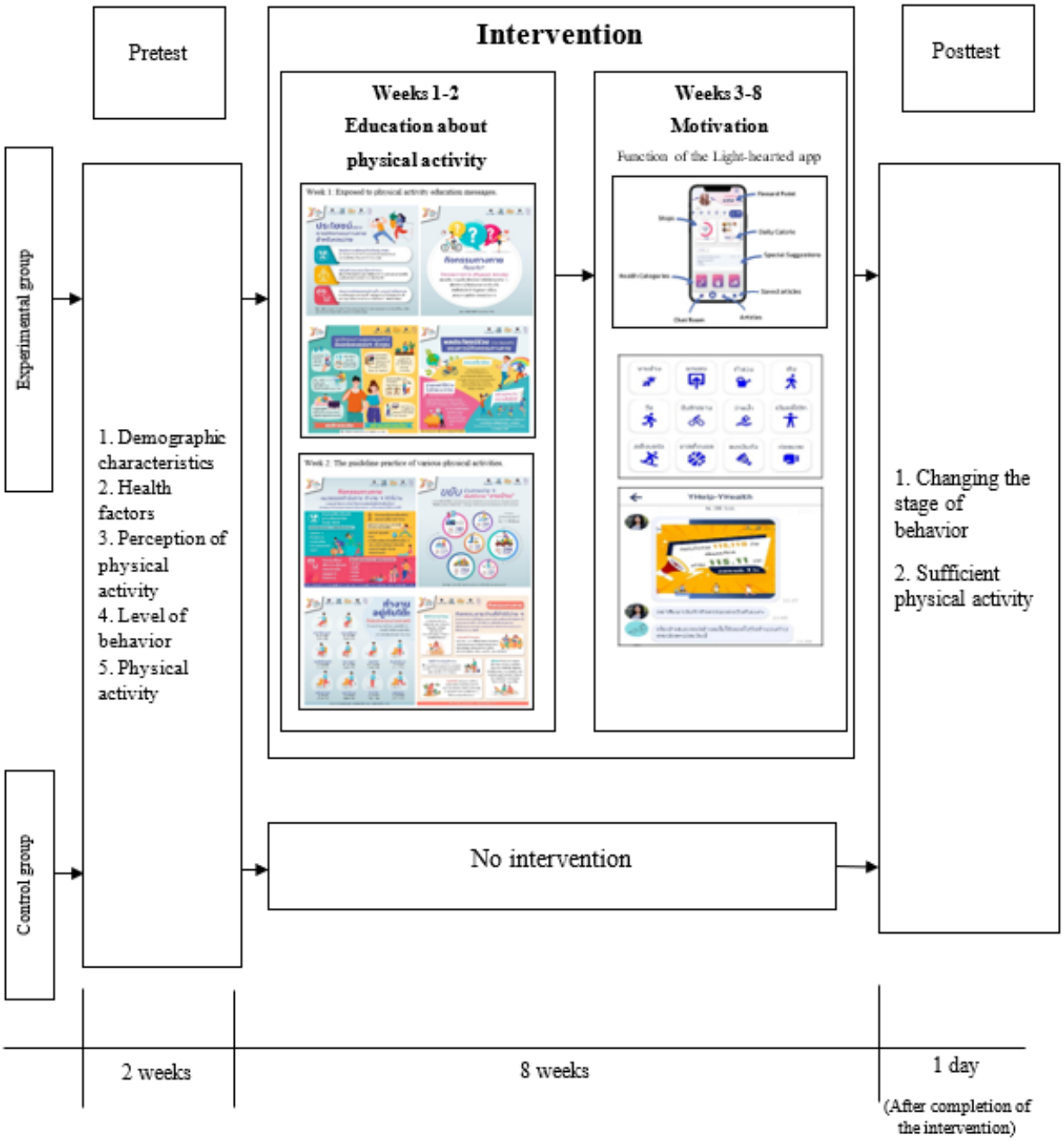

Results

Overview

Of the 100 participants, 8 left the study during the trial period and 12 were excluded because they declined to provide posttest data, yielding a 20% (n=20) attrition rate. The remaining 80 (80%) participants in the study comprised 40 individuals each in the experimental and control groups (Figure 2; Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

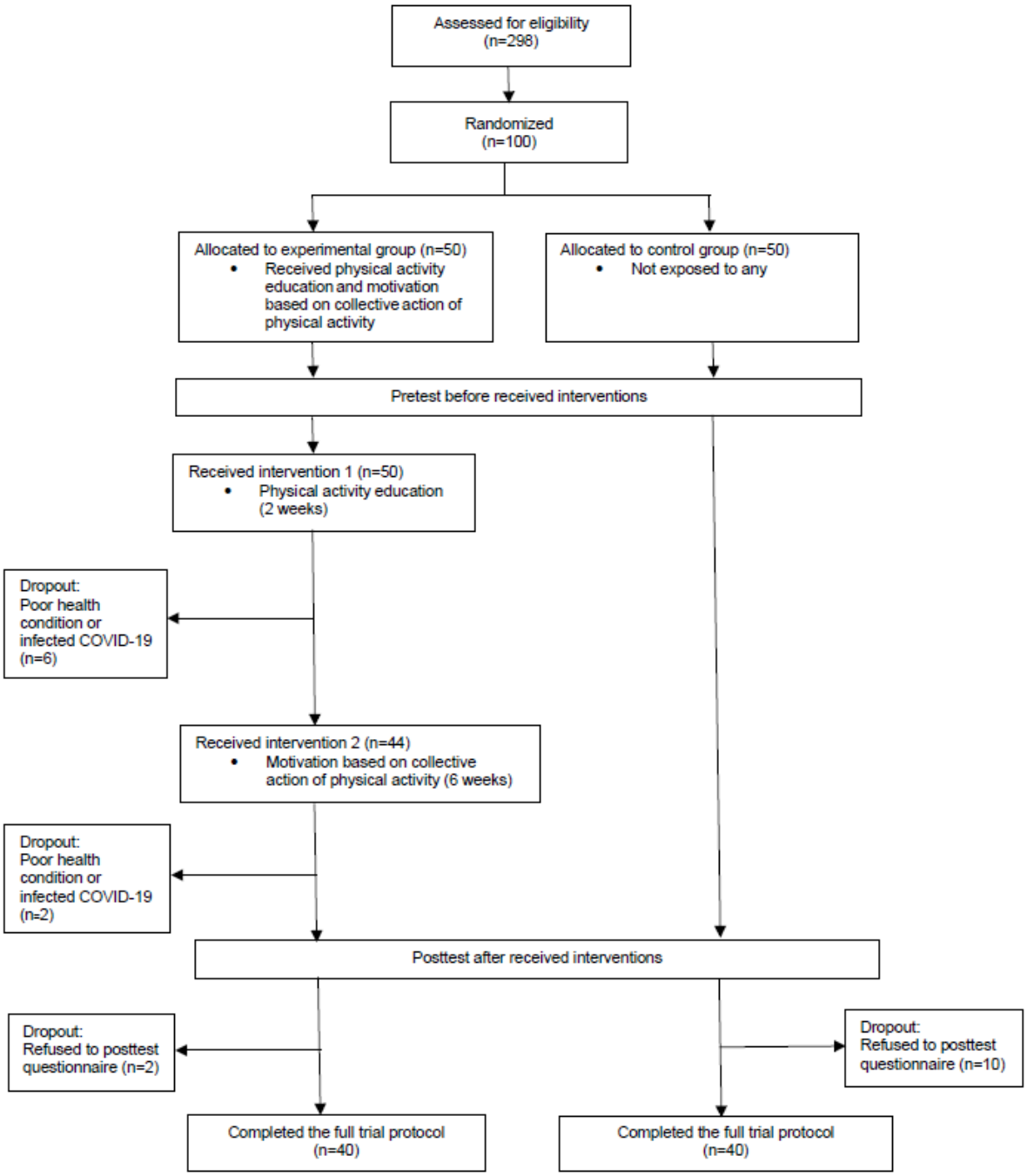

Baseline Characteristics of the Participants

The characteristics and distribution of the sample were described using univariate analysis variables such as percentage, mean, median, and range. The sample of the study comprised Gen Y individuals with similar ages in both experimental (n=40; mean age 29.6, SD 5.52 years) and control (n=40; mean age 29.2, SD 6.48 years) groups. In terms of sex, the proportion of female participants in both intervention and control groups was slightly higher (24/40, 60% and 22/40, 55%, respectively) than male participants (16/40, 40% and 18/40, 45%, respectively). The proportion of participants in the experiment group who were employed as office workers was slightly higher (18/40, 45%) than their counterparts in the control group (14/40, 35%; Table 2). There was no statistically significant difference for general characteristics, health factors, perceptions about PA, and level of moderate to vigorous physical activity (MVPA) at baseline (all $P>.05$). Prior to the intervention, individuals in the experimental group collected a slightly higher average cumulative MVPA (62.15, SD 89.38 minutes per week) than those in the control group (43.55, SD 62.15 minutes per week). Nevertheless, there was no significant difference in the PA level between the 2 groups ($P=.29$; Table 2).

XSL•FO
RenderX

Table 2. General characteristics of the sample by group.

Characteristic	Control (n=40)	Experimental (n=40)	*P* value
Sex, n (%)			.65[a]
Male	18 (45)	16 (40)	
Female	22 (55)	24 (60)	
Age group (years) , n (%)			.63[a]
20-29	21 (52)	21 (52)	
30-39, n (%)	19 (47)	19 (47)	
Age (years), mean (SD)	29.20 (6.48)	29.98 (4.41)	
Occupation, n (%)			.33[a]
Office worker	14 (35)	18 (45)	
Not an office worker	16 (40)	17 (42)	
Unemployed	10 (25)	5 (12)	
Currently have a chronic illness or condition, n (%)			.80[a]
Yes	10 (25)	11 (27)	
No	30 (75)	29 (72)	
Received education about physical activity, n (%)			.27[a]
Yes	29 (72)	31 (77)	
No	11 (27)	9 (22)	
MVPA[b] (minutes per week), mean (SD)	43.55 (62.15)	62.15 (89.38)	.29[c]

[a]*P* values are based on the chi-square test

[b]MVPA: moderate to vigorous physical activity.

[c]*P* values are based on the 2-tailed *t* test.

Comparison in the Stages of Behavioral Change

For bivariate analysis, the significance of association between variables and group-level differences was examined. Before participating in the trial, both sample groups were at the second stage of behavior change (ie, contemplation). According to TTM of behavioral change, contemplation means the individuals intend to change their behavior (into a positive one) within the next 6 months [31]. Upon completing an 8-week intervention, 63% (25/40) of participants in the experimental group and 35% (14/40) participants in the control group improved their level of behavior to the fourth stage (ie, action) by reporting an improvement in their PA for less than 6 months (Table 3). While most of the participants in the experimental group have progressed to the action stage, 50% (20/40) of participants in the control group remained in the contemplation stage. The findings suggest that the intervention helped those in the experimental group to advance in their stage of behavioral readiness, that is, to take action (χ^2_2=6.1; *P*=.01; Table 3).

Table 3. Comparison in the stages of behavioral change of the contemplation stage, preparation stage, and action stage by group and round of data collection. The chi-square test was used[a].

Round of data collection and group	Stage of change (n=40), n (%)		
	Contemplation	Preparation	Action
Baseline			
Control	40 (100)	0 (0)	0 (0)
Experimental	40 (100)	0 (0)	0 (0)
Posttest			
Control	20 (50)	6 (15)	14 (35)
Experimental	6 (15)	9 (22)	25 (62)

[a]Considerations on (1) perceived negative effects of insufficient physical activity (PA), (2) sufficient PA based on the opinions and Global Physical Activity Questionnaire, and (3) intention to increase PA.

XSL•FO
RenderX

Effectiveness of the Intervention in Increasing Cumulative Minutes of MVPA

To examine the effectiveness of the model, the mean before-and-after values within the same group were compared using the paired 2-tailed *t* test, whereas the means between the 2 groups were compared using the independent samples 2-tailed *t* test. While participants in the control group performed their regular routines, participants in the experimental group were exposed to an 8-week intervention. Health education messages were delivered through selected channels (ie, "Light-hearted" and LINE apps) to provide PA-related knowledge and create awareness of the importance of PA. In the first 2 weeks, various topics were delivered through LINE app and YouTube videos to improve participants' knowledge in PA, including the importance, cobenefit, and choices of PA, as well as guideline practices for various PA. In the third to sixth week, participants were invited to join Zoom application to collectively set up their goals. Motivational messages and reminders were sent throughout the 4 weeks through the "Light-hearted" app to ensure participants' compliance.

To test the effectiveness of the intervention, the independent samples 2-tailed *t* test was performed to compare the duration of MVPA (minutes) between participants in the experimental and control groups. After being exposed to the intervention, participants in the intervention group collected a higher cumulative minute of MVPA weekly (283 minutes) than those in the control group (164 minutes), and this was statistically significant (t_{78}=2.19; *P*=.03; Table 4).

Table 4. Differences in the duration of level of MVPA[a] in minutes per week by group and round of data collection. Independent samples 2-tailed *t* test and paired 2-tailed t test were used.

Variable	MVPA (minutes per week), mean (SD)		Difference (within group), mean (SD)	*t* test (*df*=78)
	Baseline	Posttest		
Group				
Control	43.55 (62.15)	164.05 (202.84)	__[b]	—
Experimental	61.75 (89.38)	283.13 (279.22)	221.38 (296.25)	4.73[c]
Difference (within round of data collection)	—	119.08 (54.57)	—	2.19[d]

[a]MVPA: moderate to vigorous physical activity.

[b]Not available.

[c]Paired 2-tailed *t* test was conducted. Difference indicates a difference of outcome measures between baseline and posttest in each group. Significant difference in comparison with the control group (*P*<.001).

[d]Independent samples 2-tailed *t* test was conducted. Difference indicates the difference of outcome measures between the experimental and control groups after interventions. Significant difference in comparison with the control group (*P*=.03).

The effectiveness of the intervention could also be observed from the increase in the cumulative minutes of MVPA collected by participants in the experimental group. The analysis considered the difference in the duration of PA before and after participating in the study intervention among members of the experimental group using the paired samples 2-tailed *t* test statistic. The analysis found that the mean duration of PA before and after participating in the intervention among the experimental group was significantly different at the *P*<.001 level (t_{78}=4.73). PA increased by 221.38 minutes, suggesting that the model can significantly increase the PA of Gen Y persons exposed to the intervention (Table 4).

Effectiveness of the Intervention in Improving the Proportion of Participants With Sufficient MVPA

In the baseline, 10% (4/40) of participants in the experimental group and 8% (3/40) in the control group met the recommended level of MVPA for adults. After being exposed to the study intervention, the experimental group had more sufficient PA than their counterparts in the control group, and that difference was statistically significant (*P*=.03). Over two-fifths (25/40, 63%) of members in the experimental group had adequate PA compared to only slightly over one-third of the control group (14/40, 35%). With an 80% (80/100) general compliance rate and 63% (25/40) success rate yielding the effect size of the intervention at 28% (effect size=27.5 and number needed to treat=3.64), those who were exposed to the full package of interventions and complied with the intervention (24/25, 96%) recorded positive change and had sufficient PA, and that improvement was statistically significant (χ^2_1=31.9; *P*<.001). Male adults aged 30-39 years and employed in a nonformal sector are more likely to comply with the intervention compared to female younger adults (20-29 years) who are unemployed or office workers (Figure 3).

XSL•FO
RenderX

Figure 3. Ability of the intervention to create sufficient physical activity and comparison of behavior change by group and round of data collection sent through the health apps "Light-hearted" and LINE.

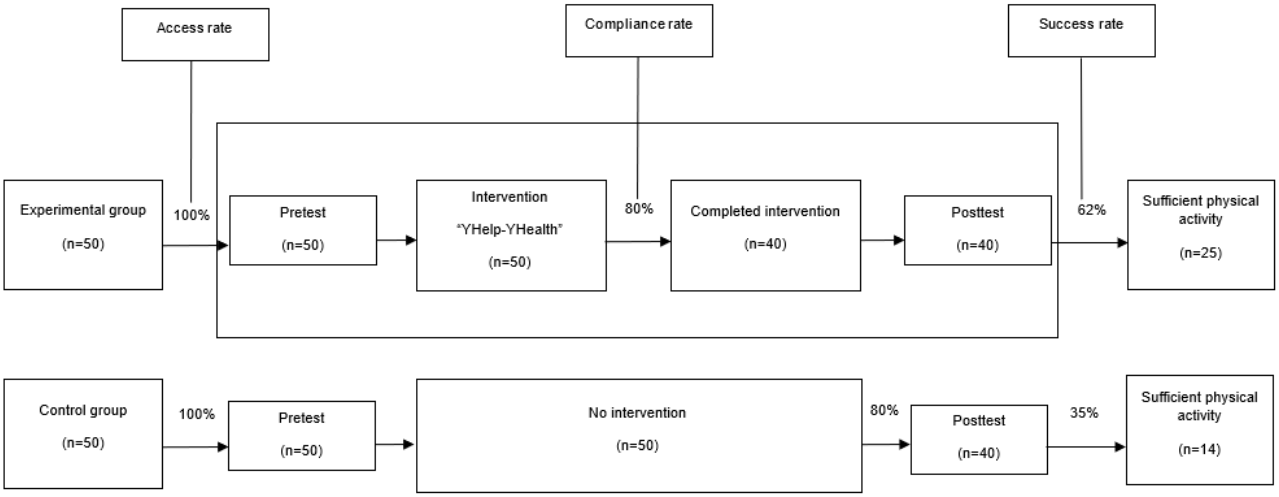

Discussion

Principal Findings

The effectiveness of a digital group–based activity has been shown to enhance the PA of Gen Y individuals in the intervention group. The COVID-19 pandemic has affected the PA of the population globally. This study found that applying digital group–based activity intervention can help in regenerating PA of Gen Y by considering the context of their lifestyle and habits. The digital group–based activity was able to significantly boost the PA level of behavior of members of the experimental group and shift from the contemplation stage to action. Over 3 in 5 (25/40, 63%) members of the experimental group have successfully changed their behavior by taking action and improving their PA. The findings of this study indicated that the prototype model of digital group–based activity was appropriate for Gen Y as it fits their characteristics and lifestyle. The health information packages in the intervention were beneficial for Gen Y in accelerating their behavior changes, as they contained a series of practical knowledge and examples on PA during COVID-19 pandemic containment periods. The incorporation of the TTM further contributed to the alignment of the intervention with the specific stage of behavioral readiness that varied among individual participants [30,31].

The experimental group showed a greater mean increase in PA than the control group, and that difference is statistically significant. The findings strongly suggest that the prototype of the digital group–based activity to promote PA for Gen Y Thai individuals was successful in building motivation through setting common goals [12,36]. As in-person peer interaction was limited during the containment measures, those Gen Y members who wanted to adopt a healthier lifestyle participated more actively in setting behavior change goals through the digital group–based activity. When engaged with a group, the drive or motivation of each person propels the goals of the group, and that phenomenon can be exploited to increase positive behavior at both the level of behavior and the duration of PA [37].

Strengths and Limitations

Olson's collective action theory (1974) [38,39], which focuses on the process of behavior change (ie, motivation) through the creation of collective goals, was applied in this study by leading the participants to believe that a group of their peers had adopted common goals. At least among this sample of Gen Y Thai individuals, this kind of motivation can lead to behavior modification and the ability to achieve the goals as set. The change occurs through direct motivation from intervention activities and indirect motivation from peer group expectations. This process is consistent with collective action theory [38,40], which, according to data in this study, can clearly enhance the PA of Gen Y.

It is noteworthy that members in the control group also reported an increase in PA during the trial period, but not to the same level as those in the experimental group. It is noteworthy that while the control group participants also exhibited an increase in PA during the trial, it was not as significant as the experimental group. This suggests external factors in the societal context may have influenced the behavior of Gen Y participants, such as societal trends emphasizing health and wellness during the pandemic and individual circumstances like remote work or altered routines. In addition, perhaps the mere act of inviting those members of the control group to participate in the study may have been motivational in and of itself [41]. The members of the control group may have become more self-aware of their PA level during the pretest phase and took steps to improve their PA on their own initiative. Or perhaps, the changes in the posttest PA of both groups were due to fewer restriction measures with limited facility opening.

After the completion of the intervention, the experimental group provided valuable feedback, suggesting improvements such as enhancing the frequency of transmission notifications, addressing issues with group chat notifications, and integrating a diverse range of PA selection functions. These suggestions aimed to enhance the overall user experience and make the app more engaging for participants.

All the individuals in the experimental group who dropped out from the study did so because of COVID-19 infection (and were

XSL•FO
RenderX

admitted to a clinical facility). Attrition in the control group was due to the inability to participate in the posttest survey. The sample attrition was assumed to be random and, thus, was assumed not to introduce statistical bias in the analysis.

That said, there are certain limitations of this study. For example, the package of interventions was not assessed for efficacy over time after the trial period was complete. This makes it impossible to track the persistence of the recorded PA gains in the intervention group. In addition, this model was designed to be consistent with the Gen Y lifestyle. Even if the intervention improves PA in Gen Y, applying it to other adult populations may have a different effect. In terms of chronic diseases as a variable, the proportion of the sample group who have the disease is not a national proportion, but it is the nature of the sample group who consented to participate and had no effect on the intervention.

Future Directions

To expand the intervention in other population age groups, it would be necessary to take into account the prevailing stage of behavior and way of life of that population segment. In addition, the findings from this study are short term. Creating profound behavior change to the fourth stage (action) and then on to the fifth stage (maintenance) takes extensive time and a greater range and intensity of interventions than were possible in this study. As a result, there is no guarantee that the observed improvements in the behavior of the experimental group were sustained for any length of time after the end of the trial. However, the fact that the intervention could be delivered on the web means that it would be possible to provide follow-up prompts and ongoing motivation (ie, pandemic or postpandemic) for the target population to sustain or even improve the gains they have made and at very little cost to the program. Future research should compare the collective action of people in a digital community to that of people in a physical or actual community to determine whether or not change occurs. Promoting PA in the digital community can be a very cost-effective way to increase the likelihood that younger individuals will become more physically active. Since the digital community can be implemented on numerous internet platforms, its theoretical reach is essentially limitless. The digital community can also connect individuals to comparable groups locally and globally. Finally, the digital community is substantially less expensive to implement than a comparable physical community intervention.

Conclusions

There was a significant change in the PA level of participants in the intervention group who were exposed to digital group–based activity for 8 weeks. The model was effective in promoting PA for members of Gen Y seeking to change their health behavior through the setting of common goals and the provision of adequate health-related information. This study showed that the prototype model was tailored to the prevailing lifestyle of Gen Y and, thus, could be integrated in their daily routine with least disruption.

Acknowledgments

The authors would like to thank the participation of Generation Y individuals in this study. The authors gratefully acknowledge the funding support from Thai Health Promotion Foundation (grant 64-00-0120). In addition, the authors acknowledge the support from the Biomed laboratory, the health app “Light-hearted” from Popmed Co Ltd, and the Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Thailand.

Authors' Contributions

NP and PK conceived the study, designed the research methodology, performed the formal data analysis, and conceptualized the paper. NP drafted the paper. NP, AC, KT, TS, BS, and PK interpreted the results. NP, PK, and DAW wrote the final paper. All authors have read and approved the final version of the paper and agreed with the order of presentation of the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 3123 KB - jmir_v26i1e43366_app1.pdf]

References

1. Saqib ZA, Dai J, Menhas R, Mahmood S, Karim M, Sang X, et al. Physical activity is a medicine for non-communicable diseases: a survey study regarding the perception of physical activity impact on health wellbeing. Risk Manag Healthc Policy 2020;13:2949-2962 [FREE Full text] [doi: 10.2147/RMHP.S280339] [Medline: 33335436]
2. Noncommunicable diseases. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2023-12-30]
3. Piercy KL, Dorn JM, Fulton JE, Janz KF, Lee SM, McKinnon RA, et al. Opportunities for public health to increase physical activity among youths. Am J Public Health 2015;105(3):421-426. [doi: 10.2105/AJPH.2014.302325] [Medline: 25602864]

XSL•FO
RenderX

4. Singh B, Jain S, Rastogi A. Effects of nationwide COVID-19 lockdown on lifestyle and diet: an Indian survey. J Family Med Prim Care 2021;10(3):1246-1250 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_2046_20] [Medline: 34041159]
5. Institute for Population and Social Research. Thailand's Surveillance on Physical Activity. 2020. URL: https://tpak.or.th/th [accessed 2022-01-09]
6. Stockwell S, Trott M, Tully M, Shin J, Barnett Y, Butler L, et al. Changes in physical activity and sedentary behaviours from before to during the COVID-19 pandemic lockdown: a systematic review. BMJ Open Sport Exerc Med 2021;7(1):e000960 [FREE Full text] [doi: 10.1136/bmjsem-2020-000960] [Medline: 34192010]
7. Hammami A, Harrabi B, Mohr M, Krustrup P. Physical activity and coronavirus disease 2019 (COVID-19): specific recommendations for home-based physical training. Manag Sport Leis 2020;27(1-2):26-31. [doi: 10.1080/23750472.2020.1757494]
8. Katewongsa P, Widyastaria DA, Saonuam P, Haematulin N, Wongsingha N. The effects of the COVID-19 pandemic on the physical activity of the Thai population: evidence from Thailand's Surveillance on Physical Activity 2020. J Sport Health Sci 2021;10(3):341-348 [FREE Full text] [doi: 10.1016/j.jshs.2020.10.001] [Medline: 33039655]
9. Katewongsa P, Potharin D, Rasri N, Palakai R, Widyastari DA. The effect of containment measures during the Covid-19 pandemic to sedentary behavior of Thai adults: evidence from Thailand's Surveillance on Physical Activity 2019-2020. Int J Environ Res Public Health 2021;18(9):4467 [FREE Full text] [doi: 10.3390/ijerph18094467] [Medline: 33922340]
10. Schwendinger F, Pocecco E. Counteracting physical inactivity during the COVID-19 pandemic: evidence-based recommendations for home-based exercise. Int J Environ Res Public Health 2020;17(11):3909 [FREE Full text] [doi: 10.3390/ijerph17113909] [Medline: 32492778]
11. Dwyer MJ, Pasini M, De Dominicis S, Righi E. Physical activity: benefits and challenges during the COVID-19 pandemic. Scand J Med Sci Sports 2020;30(7):1291-1294 [FREE Full text] [doi: 10.1111/sms.13710] [Medline: 32542719]
12. Parker K, Uddin R, Ridgers ND, Brown H, Veitch J, Salmon J, et al. The use of digital platforms for adults' and adolescents' physical activity during the COVID-19 pandemic (Our Life at Home): survey study. J Med Internet Res 2021;23(2):e23389 [FREE Full text] [doi: 10.2196/23389] [Medline: 33481759]
13. Shahbaznezhad H, Dolan R, Rashidirad M. The role of social media content format and platform in users' engagement behavior. J Interact Mark 2021;53:47-65. [doi: 10.1016/j.intmar.2020.05.001]
14. Rheingold H. Virtual community. Britannica. 2021. URL: https://www.britannica.com/topic/virtual-community [accessed 2023-12-30]
15. Digital 2021 april global statshot report. Global Statshot Report. 2021. URL: https://datareportal.com/reports/digital-2021-april-global-statshot [accessed 2023-12-30]
16. National Statistical Office. The 2021 Household Survey on the Use of Information and Communication Technology (Annually). Bangkok: National Statistical Office; 2021.
17. DiPietro L, Al-Ansari SS, Biddle SJH, Borodulin K, Bull FC, Buman MP, et al. Advancing the global physical activity agenda: recommendations for future research by the 2020 WHO physical activity and sedentary behavior guidelines development group. Int J Behav Nutr Phys Act 2020;17(1):143 [FREE Full text] [doi: 10.1186/s12966-020-01042-2] [Medline: 33239105]
18. Physical Activity Promotion Development Committee. Thailand National Physical Activity Plan 2018-2030. Bangkok: Thai Health Promotion Foundation; 2018.
19. Bolton R, Parasuraman A, Hoefnagels A. Understanding Generation Y and their use of social media: a review and research agenda. J Serv Manag 2013;24(3):245-267 [FREE Full text] [doi: 10.1108/09564231311326987]
20. Coombes B. Generation Y: are they really digital natives or more like digital refugees? Synergy 2009;7(1):31-40 [FREE Full text]
21. Liu Q, Yuxuan X, Sijia C. Y-Generation digital natives' impulsive buying behavior. : IEEE; 2018 Presented at: 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON); December 12-14, 2018; Bangkok, Thailand. [doi: 10.1109/times-icon.2018.8621840]
22. Bussin MHR, Mohamed-Padayachee K, Serumaga-Zake P. A total rewards framework for the attraction of Generation Y employees born 1981–2000 in South Africa. SA J Hum Resour Manag 2019;17:a1066 [FREE Full text] [doi: 10.4102/sajhrm.v17i0.1066]
23. Gregg P, Gardiner L. Study, work, progress, repeat? how and why pay and progression outcomes have differed across cohorts. Resolution Foundation. 2017. URL: https://www.resolutionfoundation.org/app/uploads/2017/02/IC-labour-market.pdf [accessed 2024-01-03]
24. Lancaster LC, Stillman D. The M-Factor: How the Millennial Generation Is Rocking the Workplace. New York: HarperCollins; 2010.
25. Russo RA, Dallaghan GB, Balon R, Blazek MC, Goldenberg MN, Spollen JJ, et al. Millennials in psychiatry: exploring career choice factors in Generation Y psychiatry interns. Acad Psychiatry 2020;44(6):727-733. [doi: 10.1007/s40596-020-01272-3] [Medline: 32661946]
26. Strauss W, Howe N. Generations: The History of America's Future. New York: William Morrow and Company Inc; 1991.
27. Dechawatthanapaisan D, Nualuk J, Numpraya K, Pluemboon C. A study of viewpoints and expectations of generational characteristics between Generation X and Generation Y. CBSJ 2014;36(3):1-17 [FREE Full text]

XSL•FO
RenderX

28. Strategy and Planning Division MoPH. Public Health Statistics A. D. Nonthaburi, Thailand: Ministry of Public Health; 2019.
29. van Zomeren M, Postmes T, Spears R. Toward an integrative social identity model of collective action: a quantitative research synthesis of three socio-psychological perspectives. Psychol Bull 2008;134(4):504-535 [FREE Full text] [doi: 10.1037/0033-2909.134.4.504] [Medline: 18605818]
30. Prochaska JO, Norcross JC, DiClemente CC. Changing for Good: A Revolutionary Six-Stage Program for Overcoming Bad Habits and Moving Your Life Positively Forward. New York: Quill; 2014.
31. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
32. Department of Mental Health. Gen Y/Gen Me, the group that rules the world. Ministry of Public Health. 2020. URL: https://www.dmh.go.th/news/view.asp?id=2293 [accessed 2023-12-30]
33. Wang F, Boros S. The effect of daily walking exercise on sleep quality in healthy young adults. Sport Sci Health 2020;17(2):393-401 [FREE Full text] [doi: 10.1007/s11332-020-00702-x]
34. WHO guidelines on physical activity and sedentary behaviour: at a glance. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240014886 [accessed 2024-01-03]
35. Sriklad J. Development of Policy Options to Reduce Sedentary Lifestyle of Working-Age Population. Nakhon Pathom: Institute for Population and Social Research, Mahidol University; 2020.
36. de Buisonjé DR, Reijnders T, Cohen Rodrigues TR, Prabhakaran S, Kowatsch T, Lipman SA, et al. Investigating rewards and deposit contract financial incentives for physical activity behavior change using a smartphone app: randomized controlled trial. J Med Internet Res 2022;24(10):e38339 [FREE Full text] [doi: 10.2196/38339] [Medline: 36201384]
37. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
38. Mancur O. The Logic of Collective Action. Cambridge, MA: Harvard University Press; 1971.
39. Chamberlain NW, Olson M. Reviewed work: the logic of collective action: public goods and the theory of groups. Am Econ Rev 1966;56(3):603-605. [doi: 10.2307/2521228]
40. Benford RD, Snow DA. Framing processes and social movements: an overview and assessment. Annu Rev Sociol 2000;26(1):611-639. [doi: 10.1146/annurev.soc.26.1.611]
41. Legault L. Intrinsic and extrinsic motivation. In: Zeigler-Hill V, Shackelford TK, editors. Encyclopedia of Personality and Individual Differences. Cham: Springer International Publishing; 2016:1-4.

Abbreviations

Gen Y: Generation Y
MVPA: moderate to vigorous physical activity
PA: physical activity
TTM: transtheoretical model
WHO: World Health Organization

Edited by A Mavragani; submitted 11.10.22; peer-reviewed by W Wongpipit, J Xu; comments to author 11.01.23; revised version received 17.05.23; accepted 24.11.23; published 31.01.24.

Please cite as:
Pomkai N, Katewongsa P, Chamratrithirong A, Tharawan K, Sakulsri T, Samutachak B, Widyastari DA, Rasri N, Wijarn B, Wongsawat Y
Digital Group–Based Intervention for Physical Activity Promotion Among Thai Adults During the COVID-19 Lockdown: Randomized Controlled Trial
J Med Internet Res 2024;26:e43366
URL: https://www.jmir.org/2024/1/e43366
doi:10.2196/43366
PMID:38294853

©Nanthawan Pomkai, Piyawat Katewongsa, Aphichat Chamratrithirong, Kanokwan Tharawan, Teeranong Sakulsri, Bhubate Samutachak, Dyah Anantalia Widyastari, Niramon Rasri, Boonyanuch Wijarn, Yodchanan Wongsawat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Evaluation of the GET.ON Nationwide Web-Only Treatment Service for Depression- and Stress-Related Symptoms: Naturalistic Trial

Anne Etzelmueller[1,2,3], Dipl-Psych; Elena Heber[2], PhD; Hanne Horvath[2], PhD; Anna Radkovsky[4], PhD; Dirk Lehr[5], PhD; David Daniel Ebert[1,2], PhD

[1]Department of Sports and Health Sciences, Professorship Psychology and Digital Mental Health Care, Technical University of Munich, Munich, Germany

[2]HelloBetter, GET.ON Institut für Online Gesundheitstrainings GmbH, Hamburg / Berlin, Germany

[3]Department of Clinical Psychology and Psychotherapy, Institute of Psychology, Friedrich-Alexander-University of Erlangen-Nürnberg, Erlangen, Germany

[4]Department of Psychology, Philipps University Marburg, Marburg, Germany

[5]Department of Health Psychology, Institute of Psychology, Leuphana University Lüneburg, Lüneburg, Germany

Corresponding Author:
Anne Etzelmueller, Dipl-Psych
Department of Sports and Health Sciences, Professorship Psychology and Digital Mental Health Care, Technical University of Munich
Georg-Brauchle-Ring 60
Munich, 80992
Germany
Phone: 49 621345167
Email: anneetzelmueller@gmail.com

Abstract

Background: GET.ON (HelloBetter) treatment interventions have been shown to be efficacious in multiple randomized controlled trials.

Objective: This study evaluated the effectiveness of 2 GET.ON interventions, GET.ON Mood Enhancer and GET.ON Stress, in a national digital mental health service implemented across Germany.

Methods: Following an initial web-based questionnaire, participants were allocated to either intervention based on their baseline symptom severity and personal choice and received a semistandardized guided, feedback-on-demand guided, or self-guided version of the treatment. Uncontrolled routine care data from 851 participants were analyzed using a pretest-posttest design. Half of the participants (461/851, 54.2%) were allocated to the stress intervention (189/461, 41% semistandardized; 240/461, 52% feedback on demand; and 32/461, 6.9% self-guided), and almost all participants in the mood intervention (349/352, 99.2%) received semistandardized guidance.

Results: Results on depression-related symptom severity indicated a reduction in reported symptoms, with a large effect size of d=−0.92 (95% CI −1.21 to −0.63). Results on perceived stress and insomnia indicated a reduction in symptom severity, with large effect sizes of d=1.02 (95% CI −1.46 to −0.58) and d=−0.75 (95% CI −1.10 to −0.40), respectively. A small percentage of participants experienced deterioration in depression-related symptoms (11/289, 3.8%), perceived stress (6/296, 2%), and insomnia (5/252, 2%). After completing treatment, 51.9% (150/289) of participants showed a clinically reliable change in depression-related symptoms, whereas 20.4% (59/289) achieved a close to symptom-free status. Similar improvements were observed in perceived stress and insomnia severity. Guidance moderated the effectiveness of and adherence to the interventions in reducing depressive symptom severity. Effect sizes on depression-related symptom severity were d=−1.20 (95% CI −1.45 to −0.93) for the semistandardized group, d=−0.36 (95% CI −0.68 to −0.04) for the feedback-on-demand group, and d=−0.83 (95% CI −1.03 to −0.63) for the self-guided group. Furthermore, 47.6% (405/851) of the participants completed all modules of the intervention. Participant satisfaction was high across all patient groups and both interventions; 89.3% (242/271) of participants would recommend it to a friend in need of similar help. Limitations include the assignment to treatments and guidance formats based on symptom severity. Furthermore, part of the differences in symptom change between groups must be assumed to be due to this baseline difference in the measures.

XSL•FO
RenderX

Conclusions: Future digital health implementation and routine care research should focus on monitoring symptom deterioration and other negative effects, as well as possible predictors of deterioration and the investigation of individual patient trajectories. In conclusion, this study supports the effectiveness of tailored digital mental health services in routine care for depression- and stress-related symptoms in Germany. The results highlight the importance of guidance in delivering internet-based cognitive behavioral therapy interventions and provide further evidence for its potential delivered as web-only solutions for increasing access to and use of psychological treatments.

(J Med Internet Res 2024;26:e42976) doi:10.2196/42976

KEYWORDS

depression; stress; digital; internet; effectiveness; routine care

Introduction

Background

Major depressive disorder is not only highly prevalent [1] but also associated with substantial impairment [2] and economic costs [3]. Although psychological interventions for depression have been shown to be effective [4], most individuals with depression fail to receive minimally adequate intervention [5]. In a large German sample (N=1186) reporting a 12-month diagnosis of a mental disorder, Mack et al [6] found that most individuals (81.1%) had not used mental health offers in the previous 12 months. Furthermore, they found a substantial time lapse of approximately 7 years between the onset of a mood disorder and service use.

Internet-based cognitive behavioral therapy (iCBT) may help overcome some of the limitations of traditional intervention services [7]. iCBT interventions have a broad reach and accessibility and might especially attract people who might not make use of traditional mental health services [8,9], whereas acceptability is reported to be lower in self-guided iCBT interventions than in other interventions [10]. With regard to the efficacy of iCBT, a recent meta-analysis has provided evidence that iCBT for depression can have positive effects on the mean symptom improvement for subthreshold depression [11] or major depressive disorder [12], resulting in clinically relevant changes in terms of response and remission [13], with effects found to be comparable with face-to-face psychotherapy [8]. In addition, recent research has shown that self-guided iCBT interventions result in significant effects on depression outcomes [14] as well, whereas guided interventions have yielded better outcomes in comparative reviews [15,16].

Recently, evidence for the effectiveness of guided iCBT interventions in routine mental health care has accumulated. Within a systematic literature review and meta-analysis, iCBT for the treatment of adult depression and anxiety has been shown to be effective when implemented in routine mental health care, with a within-group effect size of g=0.42-1.88 and a pooled effect of 1.18 for depression studies and 0.94 for anxiety studies [17]. In their review of computer-based interventions in randomized trials, Andrews et al [18] identified 8 studies on iCBT in routine clinical practice with an average within-group effect size of g=1.07, indicating symptom reduction across the treatment of depression, panic disorder, generalized anxiety disorder, and social phobia [18].

The GET.ON (today HelloBetter) Mood Enhancer is a German iCBT intervention for depression that has been evaluated in several randomized controlled trials (RCTs) with demonstrated efficacy in different target populations [19-23]. A study by Reins et al [21] found between-group effects for guided versions of the intervention of d=0.36 (95% CI 0.01-0.70; P=.03) in major depression when compared with web-based psychoeducation. In this study, both groups showed significant reductions in depression severity from pretreatment to posttreatment time points—8.31 points (d=1.09, 95% CI 0.72-1.46; P<.001) and 5.42 points (d=0.59, 95% CI 0.24-0.94; P<.001) on the Hamilton Rating Scale for Depression, respectively. Additional studies on treating depression-related symptoms in adults with type 1 and type 2 diabetes indicate the effectiveness of the adaptation of this intervention, with effects of d=0.58 at posttreatment measurement and d=0.83 (95% CI 0.57-1.08) after 6 months for depression and diabetes compared with web-based psychoeducation [19,20]. Moderator analyses also indicated the effectiveness in nonsuicidal participants with severe depression, with d=1.05 (95% CI 0.11-1.98) at posttreatment measurement and d=0.71 (95% CI 0.19-1.61) at the 6-month follow-up [24]. In addition, studies have found between-group effects for guided versions of the intervention of d=0.69 (95% CI 0.49-0.89) in subclinical depression compared with psychoeducation [22] and treatment as usual and d=0.37 (95% CI 0.09-0.64) when delivered with feedback-on-demand guidance only [23]. The intervention has also been found to be effective in reducing the risk of onset or delaying the onset of a major depressive disorder for 12 months [25]. The results of health economic studies showed that the guided intervention for depression and diabetes had a 97% probability of being cost-effective (at a willingness-to-pay ceiling of €5000 [US $5619.25] for a treatment response) compared with an active control group [26] and that the guided stress intervention indicated a 67% likelihood of being more cost-effective than no immediate intervention, whereas it showed net savings of €181 [US $203.42] on average per participant already in the first 6 months following the intervention [27].

Targeting individuals experiencing mild to moderate depression- and stress-related symptoms with interventions directed at stress might be a possibility to reach those individuals who are waiting for a specialized mental health treatment or might usually not seek help [9,28]. As individuals with mental health problems wait on average 3 months on an outpatient psychotherapeutic treatment [29], offering low-threshold interventions can help bridge the time until another evidence-based treatment is available. Reasons for the limited health care use also include

XSL•FO
RenderX

the preference for solving problems on one's own [30]. iCBT interventions might attract this population as they allow for independent and self-reliant work processes. In addition, stigmatizing attitudes toward mental illness are associated with less active help seeking for mental health problems [31]. Targeting mental health issues with a stress intervention might appeal to individuals potentially not seeking help otherwise, for example, providing interventions specifically labeled as targeting depression. Studies have shown that internet-based cognitive behavioral stress and occupational health interventions targeting depression can be effective [32,33]. Such interventions include techniques based on cognitive behavioral therapy principles, which have been shown to be effective in treating depression [34,35]. Weisel et al [36] showed that the GET.ON (today HelloBetter) internet-based stress management intervention is also effective for overcoming depressive symptoms and that it poses an adequate way for even highly affected participants to make use of psychotherapeutic interventions [36]. Effect sizes were moderate to large compared with the waitlist control condition both at the postassessment time point (d=0.67, 95% CI 0.32-1.02) and 6-month follow-up (d=0.79, 95% CI 0.44-1.15). In addition, in this study, 86.5% of participants reported having no experience with health-related digital applications, and 60.7% reported not having had experiences with face-to-face psychotherapy. These results indicate that such interventions might be a good entry point for individuals already showing high levels of depression who otherwise might not seek treatment.

The iCBT intervention GET.ON Stress, a web-based iCBT intervention including problem-solving and emotion regulation techniques, has been evaluated in 7 RCTs so far, demonstrating efficacies within different target groups [32,37,38]. In samples of employees with elevated symptoms of perceived stress, these studies found between-group effects for the self-guided intervention of d=0.83 (95% CI 0.58-1.08) at posttest measurement and d=1.02 (95% CI 0.76-1.27 [37]) at follow-up compared with waitlist controls and d=0.79 (95% CI 0.54-1.04) at posttest measurement and d=0.85 (95% CI 0.59-1.10 [32]) at follow-up for the intervention including feedback-on-demand guidance. Similarly, in college students reporting elevated symptoms of perceived stress, Harrer et al [39] found between-group effects for the feedback-on-demand guided intervention of d=0.69 (95% CI 0.36-1.02) at posttreatment measurement and d=0.57 (95% CI 0.24-0.89) at the 6-month follow-up compared with a waitlist control.

Taken together, there is evidence suggesting that guided and unguided self-help interventions for the treatment of depression are effective in a research setting within RCTs. However, the effectiveness outside of a highly structured research setting is less documented. In comparison with reports on RCTs, these often uncontrolled studies might present higher generalizability [40], whereas RCTs trade external validity for maximal internal validity [41,42]. Furthermore, RCTs often apply stricter inclusion and exclusion criteria than studies administered under routine care conditions and provide a highly structured setting accompanying the intervention setting [43]. Rothwell [41] found that the proportion of individuals with a specific disorder in a specific area recruiting for a trial would often be <1%. In addition, RCTs are assumed to have a potential adherence-fostering effect because of their highly structured nature [43,44]. Thus, the efficacy derived from RCTs of internet-based interventions might be overestimated for what can be expected when implementing them in routine care, limiting the knowledge base for routine practice [42]. Reporting on the effectiveness of interventions after establishing their efficacy in a controlled setting is important as efficacy trials may or may not yield similar effects in routine care conditions [45], and establishing the effectiveness of an intervention to evaluate its safety and scalability is crucial. Nonrandomized trials and open cohort studies on pretest-posttest (within-group) effects investigate events in a natural setting without the involvement of experimental interference.

In several recent effectiveness studies, a greater focus has been put on factors such as the quality of iCBT interventions and their clinical outcomes when delivered under naturalistic conditions [17]. These studies reported the within-group effect sizes of guided iCBT interventions (Hedges g=0.42−1.88, with a pooled effect of g=1.78 for depression treatments [17]).

Following this evidence, in 2015, the GET.ON Mood Enhancer and GET.ON Stress were implemented in German routine health care. Interventions provided by GET.ON are aimed at adults experiencing depression- and stress-related symptoms within a nationwide, web-only service for the prevention of and early intervention on depression- and stress-related symptoms. Clients were provided with a brief online assessment followed by semistandardized, feedback-on-demand, or self-guided iCBT.

Objectives

The aim of this study was to investigate the use of, adherence to, effectiveness of, and patient satisfaction with this semistandardized, feedback-on-demand, and self-guided iCBT intervention for the treatment of depression- and stress-related symptoms when initially implemented in routine care. Participants were recommended either a stress- or depression-related intervention in a guided or self-guided format based on their baseline depression- and stress-related symptom severity. Using this tailored approach allowed for a broader implementation of the service and a higher reach in comparison with a *one-size-fits-some* approach.

Methods

Design

This effectiveness study used data routinely collected between January 2015 and June 2017 within a service provided in partnership with one of Germany's major health insurance companies (HICs). All service users were invited to participate in this open trial using a naturalistic within-group design. We included all participants in the study who reported a Patient Health Questionnaire–8 (PHQ-8) [46,47] score of >5 at baseline and started the intervention. This study reports on GET.ON clients treated during the course of the MasterMind project [48], a European project set up to foster iCBT uptake in Europe. Today, the service is offered by HelloBetter within the German health care system and internationally.

XSL•FO
RenderX

Procedure

The GET.ON (now HelloBetter) service comprised a fully web-based service without face-to-face client-coach interaction. Participants became aware of the intervention via the cooperating HIC's personal recommendation or their health insurance online platform (members' portal) or members' magazine. Furthermore, clients were informed on the GET.ON website and the general media. Participants interested in the service registered either via the HIC's or GET.ON's system and could then access the intervention platform. There, the client filled out the screening questionnaire. The sole inclusion criteria for participation were being insured with the cooperating HIC and being aged >18 years. No further screening took place, and no clients were excluded with regard to any other criteria. Clients participated in the intervention anonymously. Upon entering the intervention platform, clients provided information on the intake questionnaires, namely, PHQ-8 [46,47,49], Perceived Stress Scale (PSS-10) [50,51], and Insomnia Severity Index (ISI) [52]. On the basis of the results, a psychologist provided the client with their individual intervention recommendation within 24 hours. The recommendation was based on the combined consideration of depression-related symptoms and perceived stress—a high PHQ-8 indicating higher levels of depression-related symptoms led to the recommendation of GET.ON Mood Enhancer even if perceived stress was elevated as well. Lower levels of depression-related symptoms combined with higher levels of perceived stress resulted in the recommendation of GET.ON Stress. Clients with high ISI values were recommended additional content focusing on sleep (sleep hygiene, stimulus control to enhance sleep quality, and the reduction of insomnia symptoms) in addition to the regular iCBT intervention. The client then received access to their training modules.

Ethical Considerations

Data were routinely collected within the treatment (mental health check) as part of the provision of the intervention within a partnership with one of Germany's major HICs. The study was a retrospective participant data analysis using anonymized routine data for which it is not possible to trace the data back to individual participants. In accordance with German legal regulations (§ 15 MBO-Ä), which state that "physicians who participate in a research project which invades the mental or physical integrity of a human being, or uses human body material or data which can be traced to a particular individual, must ensure that advice on questions of professional ethics and professional conduct associated with the project is obtained from an Ethics Committee established at the responsible Chamber of Physicians, or from another independent, interdisciplinary Ethics Committee set up according to state law, before conducting the research" [53], ethics approval was not regarded as a requirement at the time the study was conducted, as the study solely included anonymized routine data. There was no compensation to the participants for either receiving the treatment or filling out the mental health checks.

Intervention

The service consisted of the GET.ON Mood Enhancer and GET.ON Stress interventions, with the level of guidance tailored to both individual symptom severity and the participants' preference.

GET.ON Mood Enhancer was an evidence-based internet-based intervention consisting of 6 modules (online lessons) and was mainly based on problem-solving and behavioral activation. The modules relied on evidence-based face-to-face manuals that have been shown to be effective at reducing depressive symptomatology, including psychoeducation, and exercises for behavioral activation, problem-solving, and relapse prevention, with 6 optional modules (sleep problems, time management, better sleep, antidepressive medication, relaxation techniques, and worrying) that could be chosen depending on the individual user needs or preferences. A strong emphasis was placed on homework assignments designed to integrate acquired coping skills into daily life. Relative to the standard version of the intervention, which was originally developed to target subclinical depressive symptoms, the current version was shortened, updated with regard to design, and simplified to also account for potentially reduced ability to concentrate among individuals with more severe depressive symptoms, including reducing the length of explanatory text.

The GET.ON Stress intervention was based on the transactional model of stress by Lazarus et al [54]. This intervention included both problem-solving and emotion regulation strategies. Important health behavior change principles such as goal setting, action planning, and coping planning were followed. GET.ON Stress consisted of 7 sessions and a booster session provided 4 weeks after training completion. Following psychoeducation (session 1), the participants learned a 6-step procedure to systematically solve problems (sessions 2-3). In sessions 4 to 6, the participants were introduced to emotion regulation techniques (muscle and breathing relaxation, acceptance of negative emotions, and self-support in difficult situations). Session 7 included planning for the future, in which participants set goals to maintain the achieved results. Moreover, the same 6 aforementioned optional modules were included in this version as well. The application of exercises in daily life was strongly recommended. The participants were advised to complete 1 to 2 sessions per week. The program included exercises, audio and video files, and downloadable material and was presented on a secured web-based platform. A more detailed description of the overall intervention can be found in the protocol of the accompanying efficacy trial [55], whereas the implemented version of the GET.ON Stress intervention did not include mobile coaching via SMS text messaging.

Standard operating procedures in case of a crisis, such as suicidal ideation and suicidality, included coaches contacting their supervisor. In the case of suicidal ideation, coaches sent a message to the participant taking up what the participant had been writing. In this message, information on suicidal ideation was provided, and further support options were laid out. In an attachment, a detailed description of further support options was sent to the participant, detailing intermediate-term (therapeutic options such as a general practitioner and face-to-face psychotherapy and how to schedule an appointment, social consulting options, and other consulting centers) and instant offers (crisis and emergency numbers). Furthermore,

XSL•FO
RenderX

participants showing symptom deterioration were referred to more intensive care.

Guidance and Professional Training

On the basis of the participants' baseline depressive symptom severity (measured using the PHQ-8) and perceived stress (measured using the PSS-10), they were recommended a semistandardized, feedback-on-demand, or self-guided version of the intervention. Following their personal recommendation, participants could decide to enter the version with less guidance (from semistandardized to feedback on demand or from feedback on demand to self-guided), but they could not enter a more intensive guidance format (from self-guided to feedback on demand or from feedback on demand to semistandardized guidance).

The self-guided version was completely unguided and self-administered by the participant. For technical questions, participants were able to contact the IT team. The feedback-on-demand guidance included standardized reminders from the platform once a week. In addition, participants were able to ask questions or request feedback from their online coach at any time during the first 3 months. Within the semistandardized guidance, the coaches provided personalized written feedback based on templates after the completed treatment modules as a patient safety measure. Online coaches were licensed clinical psychologists or clinical psychologists in training for the license; in exceptional cases, they were psychologists under the supervision of experienced licensed clinical psychologists. Professional training included the provision of a detailed intervention manual and close supervision by an experienced licensed psychologist.

Measures

All questionnaires were administered at baseline before starting the intervention as part of the intervention; afterward, the participants could fill in the questionnaires every 2 weeks. The analysis used postmeasurement data that were collected closest in time to the completion of the last treatment session that the participant engaged in.

Primary Outcome Measure

The PHQ-8 is a self-report measure for the assessment of depressive symptoms. It administers the first 8 items of the Patient Health Questionnaire–9 (PHQ-9) [46,47], omitting the item on thoughts of death or self-harm. The PHQ-8 has been used in clinical or research settings where the follow-up to positive responses to the ninth item of the PHQ-9 may be delayed, for example, in a web-based screening. Erbe et al [56] found that delivering the PHQ-9 in a digital format does not affect the psychometric properties in a clinically meaningful way.

Secondary Outcome Measures

As a secondary outcome, the level of perceived stress was measured using the PSS-10 (5-point Likert scale; range 0-40; Cronbach α=.78-.91 [50,51,57]). Furthermore, insomnia severity was measured using the ISI (7 items; range 0-28; Cronbach α=.90; [52]). The participants' satisfaction was assessed using the Client Satisfaction Questionnaire–8 (CSQ-8; 8 items; range 0-32 [58,59]). The questionnaire has been translated into various languages and is used to measure global participant satisfaction.

Statistical Analyses

Participant characteristics were analyzed using descriptive statistics. Linear mixed-effects models were applied to estimate intervention effects. We used Satterthwaite approximations [60] to derive *P* values for the fixed effects and calculated the effect sizes by dividing the estimated mean difference by the SD of the postmeasure [61]. We included the variables symptom severity (mild, moderate, and severe), guidance (semistandardized guided, feedback-on-demand guided, and self-guided), and intervention (mood and stress) into the linear models and compared this model with the model including the interaction of the variable with the measurement time point using chi-square tests. If a significant interaction was found, we investigated the effects of a linear mixed model for each level of the variable. As we applied 3 group comparisons, we adjusted the *P* value indicating significance to .05/3=.02 [62].

Intervention completers were defined as participants who started all the intended modules. A minimal adequate dose of the intervention was defined as starting at least 5 modules of any intervention. As the aforementioned statistical model assumes missing-at-random data, we also included the analysis using only completer data. To determine the number of participants achieving a reliable, positive outcome, we coded participants as responders or nonresponders according to the widely used Reliable Change Index [63]. To determine the potential negative effects of the intervention, the number of participants showing a reliable symptom deterioration was assessed regarding depressive symptom severity and perceived stress, defined as a negative change in symptom severity based on a negative Reliable Change Index. According to Jacobson and Truax [63], a cutoff point indicating symptom-free status was calculated and defined as scoring >2 SDs below the baseline mean.

Logistic regression analysis was conducted to explore the influence of depressive symptom severity (PHQ-8) at baseline, before the start of the intervention, and at completion. We also tested the influence of guidance (semistandardized, feedback on demand, and self-guided) on adherence (defined as the number of completed modules).

Participant satisfaction measured using the CSQ-8 was reported as means and SDs at the item level and as the "percentage of agreement" operationalized as the positive answer on the 4-point Likert scale on the CSQ-8. We analyzed the data of participants who filled out the complete questionnaire.

Information on participants' recruitment pathways was provided in open-text answers. These text answers were interpreted and assigned to one of six groups: (1) information from the HIC, (2) via the internet (through the GET.ON, HIC website, membership portal, or Google), (3) from the HIC's membership magazine, (4) through direct HIC consultation, (5) recommendations by health care professionals, or (6) recommendations by others (family and friends).

R (version 3.5.2; R Foundation for Statistical Computing) [64] was used for all analyses.

XSL•FO
RenderX

Results

Sample Selection and Baseline Characteristics

Figure 1 shows the participant flowchart. Of the 1195 cases reported in the data set, 1096 (91.72%) reported a PHQ-8 score of >5 at baseline. Of those 1096 participants, 851 (77.65%) also started the intervention (opened at least one session). Of those 851 participants, 327 (38.4%) were identified as having severe depressive symptoms according to the PHQ-8 and 310 (36.4%) were identified as having moderately severe symptoms. On average, participants were aged 41.7 (SD 11.31) years, and 72.2% (612/848) were female. Approximately half (405/833, 48.7%) were working full time, whereas 24.6% (205/833) were employed part time, 10.7% (89/833) were employed but on sick leave, and 7.2% (60/833) were unemployed. In total, 59.7% (503/843) had received a higher education, and 3.8% (32/843) had received a lower education. Of the included participants, 40.4% (337/834) reported having received psychotherapeutic treatment before, 18.3% (153/834) were in psychotherapeutic treatment while receiving the iCBT intervention, and 41.3% (345/834) had not received any form of psychotherapy before. Of the included participants, 38.4% (327/851) reported severe depressive symptoms, 36.4% (310/851) reported moderate depressive symptoms, and 25.1% (214/851) reported mild depressive symptoms at baseline. In addition, 98.7% (840/851) of participants followed the recommendation regarding the guidance format based on their symptom severity. Furthermore, 41.4% (352/851) of participants (*mild symptoms*: 1/352, 0.3%; *moderate symptoms*: 36/352, 10.2%; *severe symptoms*: 315/352, 89.5%) entered the mood intervention, 54.2% (461/851) of participants (*mild symptoms*: 196/461, 42.5%; *moderate symptoms*: 236/461, 51.2%; *severe symptoms*: 2/461, 0.4%) entered the stress management intervention, 2.6% (22/851) of participants (*mild symptoms*: 8/22, 36%; *moderate symptoms*: 8/22, 36%; *severe symptoms*: 6/22, 27%) entered both, and 1.9% (16/851) of participants started optional online modules only. Furthermore, 66.2% (563/851) of clients participated in the semistandardized guided intervention (*mood*: 349/563, 62%; *stress*: 191/563, 33.9%; *only optional modules*: 7/563, 1.2%; *both stress and depression*: 16/563, 2.8%), 29.7% (253/851) of clients participated in the feedback-on-demand guided intervention (*mood*: 1/253, 0.4%; *stress*: 241/253, 95.3%; *only optional modules*: 6/253, 2.4%; *bothstress and depression*: 5/253, 2%), and 4.1% (35/851) of clients participated in the self-guided intervention (*mood*: 2/35, 6%; *stress*: 29/35, 83%; *only optional modules*: 3/35, 9%; *both stress and depression*: 1/35, 3%). Of all participants, 32.9% (279/851) were recruited via the internet, 23.3% (198/851) were recruited via direct contact with the HIC, and 19.9% (169/851) were recruited via the HIC's membership magazine. In addition, 2.7% (23/851) were recruited via a direct HIC consultation, the treatment was recommended by their general practitioner to 0.5% (4/851), and the treatment was recommended by a nonprofessional (ie, friend or family) to 1.1% (9/851). Participants' demographic characteristics are presented in Table 1.

Figure 1. Participant flowchart. PHQ-8: Patient Health Questionnaire–8.

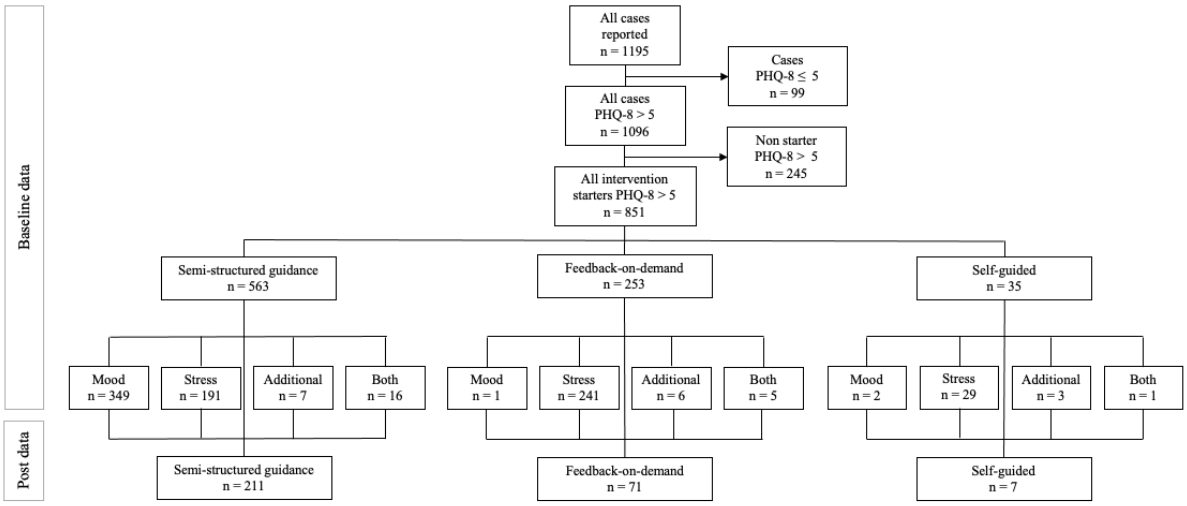

XSL•FO
RenderX

Table 1. Demographics for all starters reporting a Patient Health Questionnaire–8 score of >5 (N=851).

Characteristics	Values
Age (years), mean (SD)	41.68 (11.31)
Sex (female), n (%)	612 (72.2)
Employment, n (%)	
Employed but on sick leave	89 (10.7)
Employed (part time)	205 (24.6)
Employed (full time)	405 (48.7)
Unemployed	60 (7.2)
Nonworking (eg, pensioner, housewife, or househusband)	73 (8.8)
Education, n (%)	
High	503 (59.7)
Middle	306 (36.3)
Low	32 (3.8)
None	1 (0.1)
Symptom severity, n (%)	
Mild	214 (25.1)
Moderate	310 (36.4)
Severe	327 (38.4)
Experience with PT[a], n (%)	
"I have received PT before"	337 (40.4)
"I am currently in PT"	153 (18.3)
"I have never received PT before"	345 (41.3)
Intervention, n (%)	
Mood	352 (41.4)
Stress	461 (54.2)
Only optional modules	16 (1.9)
Both mood and stress main modules	22 (2.6)
Number of sessions, mean (SD)	6.64 (4.85)
Number of sessions from the main intervention, mean (SD)	3.91 (2.64)
Guidance modality, n (%)	
Feedback-on-demand guided	253 (29.7)
Semistandardized guided	563 (66.2)
Self-guided	35 (4.1)

[a]PT: psychotherapy.

Primary Analysis: Effects on the Depression Measure

Statistical Significance and Effect Size

Table 2 shows the means and SDs of participants' depressive symptom severity measured using the PHQ-8 at both measurement time points. The primary analysis showed a significant reduction in depressive symptoms on the PHQ-8 over time (β=−4.61; SE 0.27; P<.001), reflecting a large effect size of d=−0.92 (95% CI −1.21 to −0.63). Multimedia Appendix 1 depicts the sensitivity analyses for these outcomes and effect sizes for depressive symptom severity within the subsets of study and treatment completers.

XSL•FO
RenderX

Table 2. Means and SDs of participants' Patient Health Questionnaire–8 (PHQ-8), Perceived Stress Scale (PSS-10), and Insomnia Severity Index (ISI) scores at both measurement time points.

	Participants, n (%)[a]	PHQ-8, mean (SD)	PSS-10, mean (SD)	ISI, mean (SD)
All				
Before	851 (100)	13.02 (4.57)	25.77 (5.63)	13.65 (5.86)
After	307 (36.1)	8.65 (5.03)	18.23 (7.53)	9.37 (6.10)
Mood intervention				
Before	352 (41.3)	17.22 (2.87)	28.19 (5.43)	16.15 (5.64)
After	127 (14.9)	10.43 (5.46)	19.89 (7.49)	11.38 (6.44)
Stress intervention				
Before	461 (54.2)	9.94 (2.69)	24.06 (5.03)	11.79 (5.30)
After	148 (17.4)	7.07 (3.92)	16.68 (7.19)	7.63 (5.22)
Both interventions				
Before	22 (2.6)	12.05 (4.57)	25.52 (5.78)	12.68 (6.66)
After	13 (1.5)	9.46 (6.24)	19.92 (8.79)	9.90 (6.44)
Only optional modules				
Before	16 (1.8)	10.75 (5.08)	22.31 (6.59)	13.56 (5.14)
After	1 (0.1)	4.00 (N/A[b])	17.00 (N/A)	6.00 (N/A)
Semistandardized guidance				
Before	563 (66.1)	15.51 (3.36)	27.87 (4.80)	14.98 (5.82)
After	211 (24.8)	9.36 (5.18)	19.00 (7.35)	9.80 (6.23)
Feedback-on-demand guidance				
Before	253 (29.8)	8.43 (1.41)	21.93 (4.77)	11.43 (4.90)
After	71 (8.3)	6.97 (4.14)	16.35 (7.91)	8.41 (5.72)
Self-guided				
Before	35 (4.1)	6.23 (3.61)	19.76 (4.99)	8.17 (4.99)
After	7 (0.8)	4.29 (0.95)	14.29 (4.82)	6.57 (4.35)
Mild symptoms				
Before	214 (25.1)	7.35 (1.42)	20.88 (4.73)	10.32 (4.88)
After	59 (6.9)	6.61 (4.08)	16.54 (8.16)	8.02 (5.36)
Moderate symptoms				
Before	310 (36.4)	11.87 (1.48)	25.44 (4.73)	12.88 (5.29)
After	106 (12.5)	7.78 (4.24)	17.17 (6.68)	7.72 (5.20)
Severe symptoms				
Before	327 (38.4)	17.83 (2.35)	29.30 (4.32)	16.55 (5.59)
After	124 (14.6)	10.35 (5.51)	19.96 (7.62)	11.39 (6.55)

[a]Sample size in the subset was based on answers on the PHQ-8.

[b]N/A: not applicable.

Influence of Guidance on Treatment Outcome

Guidance was associated with depressive symptom severity over time ($F_{2,755.96}$=45.53; P<.001). Planned comparisons of the guidance groups revealed that semistandardized guidance was associated with a greater symptom reduction compared with feedback-on-demand guidance (β=4.68; SD 0.51; t=9.23; P<.001) and the self-guided format (β=4.54; SD 1.38; t=3.28; P=.001). The effect of "self-guided" was not significant over time. Participants in the feedback-on-demand guided group did not show a significantly higher symptom reduction than those in the self-guided group (β=−.14; SD 1.43; t=−0.1; P=.92). The group sizes differed in this analysis as only 4.1% (35/851) of participants were included in the self-guided condition, 66.2% (563/851) were included in the semistandardized condition, and 29.7% (253/851) were included in the feedback-on-demand

XSL•FO
RenderX

guided condition. The corresponding effect sizes in the guidance groups were d=−1.20 (95% CI −1.45 to −0.93) for the semistandardized guided group, d=−0.36 (95% CI −0.68 to −0.04) for the feedback-on-demand guided group, and d=−0.83 (95% CI −1.03 to −0.63) for the self-guided group.

Influence of the Intervention Type on Treatment Outcome

The intervention type (*mood* or *stress*) was associated with depression-related symptom severity measured using the PHQ-8 over time ($F_{1,680.77}$=79.59; P<.001), with β=3.9 (SD 0.44; t=8.92; P<.001) indicating a higher symptom change in the mood treatment. The corresponding within-group effect sizes were d=−1.26 (95% CI −1.6 to −0.92) for the mood intervention and d=−0.75 (95% CI −0.97 to −0.53) for the stress intervention.

Influence of Baseline Symptom Severity on Treatment Outcome

Depression-related symptom severity was associated with intervention outcomes ($F_{2,858.05}$=94.49; P<.001). Planned comparisons of depression-related symptom severity revealed that severe depression-related symptoms at baseline were associated with a greater symptom reduction compared with mild depression-related symptom severity (β=−6.64; SD 0.5; t=−13.42; P<.001), moderate depression-related symptoms were associated with a greater symptom reduction compared with mild depression-related symptom severity (β=−3.27; SD 0.51; t=−6.44; P<.001), and severe depression-related symptoms were associated with a greater symptom reduction compared with moderate depression-related symptom severity (β=3.37; SD 0.42; t=8.05; P<.001). The corresponding effect sizes in the symptom severity groups were d=−1.36 (95% CI −1.72 to −1.00) for severe depression-related symptoms, d=−1.00 (95% CI −1.26 to −0.68) for moderate depression-related symptoms, and d=−0.21 (95% CI −0.55 to 0.13) for mild depression-related symptoms. The results of the analyses of the pre- and postintervention changes in symptom severity are shown in Table 3.

Table 3. Linear mixed-effects model outcomes and effect sizes for depressive symptom severity (measured using the Patient Health Questionnaire–8).

	Estimate (SD)	*t* test	*F* test (*df*)	*P* value	Effect size (*d*) (95% CI)
All starters (n=851)			288.42 (504.90)	<.001	−0.92 (−1.21 to −0.63)
Time	−4.61 (0.27)	−16.98			
Guidance			45.53 (755.96)	<.001	
SSF[a]	−6.17 (0.28)	−21.70		<.001	−1.19 (−1.45 to −0.93)
FoD[b]	−1.47 (0.30)	−4.91		<.001	−0.36 (−0.68 to −0.04)
SG[c]	−0.76 (0.36)	−2.10		.08	−0.83 (−1.03 to −0.63)
SSF × SG	4.54 (1.38)	3.28		.001	N/A[d]
SSF × FoD	4.68 (0.51)	9.23		<.001	N/A
FoD × SG	−0.14 (1.43)	−0.10		.92	N/A
Intervention			79.59 (680.77)	<.001	
Mood	−6.87 (0.35)	−19.40		<.001	−1.26 (−1.60 to −0.92)
Stress	−2.95 (0.27)	−10.87		<.001	−0.75 (−0.97 to −0.53)
Symptom severity			94.49 (858.05)	<.001	
Severe	−7.48 (0.34)	−21.92		<.001	−1.36 (−1.72 to −1.00)
Medium	−4.09 (0.28)	−14.77		<.001	−0.97 (−1.26 to −0.68)
Mild	−0.85 (0.30)	−2.81		.005	−0.21 (−0.55 to −0.13)
Mild × severe	−6.64 (0.50)	−13.42		<.001	N/A
Mild × medium	−3.27 (0.51)	−6.44		<.001	N/A
Medium × severe	3.37 (0.42)	8.05		<.001	N/A

[a]SSF: semistandardized feedback.

[b]FoD: feedback on demand.

[c]SG: self-guided.

[d]N/A not applicable.

Reliable Change

Table 2 shows the means and SDs of the participants' PHQ-8 scores at both measurement time points. Of all participants reporting postmeasure data, 3.8% (11/289) experienced a deterioration in symptoms. In this study, the cutoff score for reliable change was 9.14, indicating a value of 2 SDs below the mean of the population at the premeasurement time point. At

XSL•FO
RenderX

the postmeasurement time point, 51.9% (150/289) and 20.4% (59/289) of participants showed a clinically reliable change in depressive symptoms and close to symptom-free status, respectively. The reliable change, remission, and deterioration rates are shown in Table 4.

Table 4. Reliable change, deterioration, and remission[a].

	Reliable change, n (%)	Deterioration, n (%)	Remission, n (%)
PHQ-8[b]	150 (51.9)	11 (3.80)	59 (20.4)
PSS-10[c]	152 (51.4)	6 (2)	109 (36.8)
ISI[d]	128 (50.8)	5 (2)	61 (24.2)

[a]Reliable Change Index was used following Jacobson and Truax [63]; a cutoff point indicating symptom-free status was calculated and defined as scoring >2 SDs below the mean (preintervention measurement).

[b]PHQ-8: Patient Health Questionnaire–8.

[c]PSS-10: Perceived Stress Scale.

[d]ISI: Insomnia Severity Index.

Secondary Outcomes

Table 2 shows the means and SDs of participants' perceived stress measured using the PSS-10 and insomnia severity measured using the ISI at both measurement time points.

Effects on Perceived Stress

Statistical Significance of Perceived Stress

There was a significant reduction in perceived stress on the PSS-10 over time ($F_{1,614.39}$=404.73; P<.001), with β=−7.7 (SE 0.38; P<.001), reflecting a large effect size of d=−1.02 (95% CI −1.46 to −0.58). Results are depicted in Table 5. Multimedia Appendix 1 shows the sensitivity analyses for these outcomes and the effect sizes for perceived stress within the subsets of study and treatment completers.

XSL•FO
RenderX

Table 5. Linear mixed-effects model outcomes and effect sizes for perceived stress (measured using the Perceived Stress Scale).

	Estimate (SD)	*t* test	*F* test (*df*)	*P* value	Effect size (*d*; 95% CI)
All starters (n=847)			404.73 (614.39)	<.001	−1.02 (−1.46 to −0.58)
Time	−7.70 (0.38)	−20.12			
Guidance			8.38 (762.22)	<.001	
SSF[a]	−8.88 (0.43)	−20.86		<.001	−1.21 (−1.58 to −0.84)
FoD[b]	−5.66 (0.70)	−8.06		<.001	−0.72 (−1.33 to −0.11)
SG[c]	−5.45 (1.67)	−3.26		.012	−1.18 (−2.18 to −0.18)
SSF × SG	3.42 (2.24)	1.52		.13	N/A[d]
SSF × FoD	3.24 (0.82)	3.93		<.001	N/A
FoD × SG	0.18 (2.31)	0.08		.94	N/A
Intervention			1.15 (622.21)	.28	
Mood	−8.35 (0.56)	−14.87		<.001	−1.12 (−1.59 to −0.65)
Stress	−7.49 (0.50)	−14.85		<.001	−1.04 (−1.44 to −1.44)
Symptom severity			94.49 (858.05)	<.001	
Severe	−9.30 (0.53)	−17.58		<.001	−1.22 (−1.72 to −0.72)
Medium	−8.30 (0.57)	−14.55		<.001	−1.25 (−1.70 to −0.80)
Mild	−4.57 (0.78)	−5.85		<.001	−0.56 (−1.24 to 0.12)
Mild × severe	−4.74 (0.93)	−5.12		<.001	N/A
Mild × medium	−3.75 (0.95)	−3.95		<.001	N/A
Medium × severe	0.99 (0.79)	1.26		.21	N/A

[a]SSF: semistandardized feedback.

[b]FoD: feedback on demand.

[c]SG: self-guided.

[d]N/A: not applicable.

Influence of Guidance on Stress Treatment Outcome

Guidance was associated with the level of perceived stress ($F_{2,762,22}$=8.38; *P*<.001). Planned comparisons of the guidance groups revealed that semistandardized guidance was associated with a greater symptom reduction compared with feedback-on-demand guidance (β=3.24; SD 0.82; *t*=3.93; *P*<.001) but not with self-guidance (β=3.42; SD 2.24; *t*=1.52; *P*=.13). Participants in the feedback-on-demand guided group did not show a significantly higher symptom reduction than those in the self-guided group (β=.18; SD 2.31; *t*=0.08; *P*=.94). The corresponding within-group effect sizes in the guidance groups were *d*=−1.21 (95% CI −1.58 to −0.84) for the semistandardized guided group, *d*=−0.72 (95% CI −1.33 to −0.11) for the feedback-on-demand guided group, and *d*=−1.18 (95% CI −2.18 to −0.18) for the self-guided group.

Influence of Treatment on Stress Treatment Outcome

There was no influence of the intervention (*mood* or *stress*) on the change in perceived stress over time ($F_{1,622.21}$=1.15; *P*=.28), with β=.81 (SD 0.75; *t*=1.07; *P*=.28). The corresponding effect sizes were *d*=−1.12 (95% CI −1.59 to −0.65) for the mood intervention and *d*=−1.04 (95% CI −1.44 to −0.64) for the stress intervention.

Influence of Baseline Symptom Severity on Stress Treatment Outcome

Depression-related symptom severity was associated with perceived stress ($F_{2,745.27}$=13.44; *P*<.001). Planned comparisons of the depression-related symptom severity groups revealed that severe depression-related symptoms at baseline were associated with a greater symptom reduction compared with mild symptom severity (β=−4.74; SD 0.93; *t*=−5.12; *P*<.001), and moderate depression-related symptoms led to a greater symptom reduction compared with mild depression-related symptom severity (β=−3.75; SD 0.95; *t*=−3.95; *P*<.001), whereas there was no significant difference between severe and moderate depression-related symptoms (β=.99; SD 0.79; *t*=1.26; *P*=.21). The corresponding effect sizes were *d*=−1.22 (95% CI −1.72 to −0.72) for severe depression-related symptoms, *d*=−1.25 (95% CI −1.7 to −0.8) for moderate depression-related symptoms, and *d*=−0.56 (95% CI −1.24 to 0.12) for mild depression-related symptoms.

Clinically Reliable Change in Perceived Stress

Furthermore, 2% (6/296) of the participants experienced a deterioration in symptoms. The cutoff score for clinically reliable change was 10.80, indicating a value of 2 SDs below the mean PSS-10 score of the population at the premeasurement time point. In total, 51.4% (152/296) of participants showed a

XSL•FO
RenderX

clinically reliable change, and 36.8% (109/296) had a close to symptom-free status. The reliable change, remission, and deterioration rates are shown in Table 4.

Effects on Insomnia

Statistical Significance of Insomnia

There was a significant reduction in insomnia severity on the ISI over time ($F_{1,456.25}$=192.32; P<.001), with β=−4.57 (SE 0.33; P<.001), reflecting a large effect size of d=−0.75 (95% CI −1.10 to −0.40). Results are depicted in Table 6.

Table 6. Linear mixed-effects model outcomes and effect sizes for insomnia severity (measured using the Insomnia Severity Index).

	Estimate (SD)	*t* test	*F* test (*df*)	*P* value	Effect size (*d*; 95% CI)
All starters (n=849)			192.32 (456.25)	<.001	−0.75 (−1.10 to −0.40)
Time	−4.57 (0.33)	−13.87			
Guidance			3.81 (523.37)	.023	
SSF[a]	−5.21 (0.38)	−13.59		<.001	−0.84 (−1.15 to −0.53)
FoD[b]	−3.36 (0.63)	−5.33		<.001	−0.59 (−1.03 to −0.15)
SG[c]	−0.94 (0.87)	−1.08		.32	−0.22 (−1.11 to 0.67)
SSF × SG	3.93 (2.04)	1.93		.05	N/A[d]
SSF × FoD	1.59 (0.74)	2.15		.03	N/A
FoD × SG	2.34 (2.10)	1.12		.27	N/A
Intervention			0.26 (461.31)	.61	
Mood	−4.82 (0.49)	−9.81		<.001	−0.75 (−1.16 to −0.34)
Stress	−4.45 (0.43)	−10.39		<.001	−0.85 (−1.14 to −0.56)
Symptom severity			94.49 (858.05)	<.001	
Severe	−5.23 (0.50)	−10.36		<.001	−0.80 (−1.23 to −0.37)
Medium	−5.22 (0.53)	−9.83		<.001	−1.01 (−1.36 to −0.66)
Mild	−2.65 (0.64)	−4.15		<.001	−0.50 (−0.95 to −0.05)
Mild × severe	−2.59 (0.86)	−3.02		.003	N/A
Mild × medium	−2.61 (0.88)	−2.97		.003	N/A
Medium × severe	−0.02 (0.72)	−0.03		.97	N/A

[a]SSF: semistandardized feedback.

[b]FoD: feedback on demand.

[c]SG: self-guided.

[d]N/A: not applicable.

Influence of Guidance on Insomnia Treatment Outcome

Guidance was associated with insomnia severity ($F_{2,523,37}$=3.81; P=.02). Planned comparisons of the guidance groups revealed that there was a significant difference between semistandardized guidance and feedback-on-demand guidance (β=1.59; SD 0.74; t=2.15; P=.03) but no difference between semistandardized guidance and self-guidance (β=3.93; SD 2.04; t=1.93; P=.05) or between self-guidance and feedback-on-demand guidance (β=2.34; SD 2.1; t=1.12; P=.27). The corresponding effect sizes were d=−0.84 (95% CI −1.15 to −0.53) for semistandardized guidance, d=−0.59 (95% CI −1.03 to −0.15) for feedback-on-demand guidance, and d=−0.22 (95%-CI −1.11 to 0.67) for self-guidance.

Influence of the Type of Treatment on Insomnia Treatment Outcome

We did not find that the type of intervention (*mood* or *stress*) was associated with change in insomnia severity (F_1=0.26; P=.61), with β=.33 (SD 0.65; t=0.51; P=.61). The corresponding effect sizes were d=−0.75 (95% CI −1.16 to −0.34) for the mood intervention and d=−0.85 (95% CI −1.14 to −0.56) for the stress intervention.

Influence of Baseline Symptom Severity on Insomnia Treatment Outcome

Depression-related symptom severity was associated with insomnia severity ($F_{2,517.42}$=5.4; P<.001). Planned comparisons of the symptom severity groups revealed that severe depression-related symptoms at baseline were associated with a greater symptom reduction in insomnia compared with mild symptom severity (β=−2.59; SD 0.86; t=−3.02; P=.003), and

XSL•FO
RenderX

moderate depression-related symptoms were associated with a greater symptom reduction compared with mild depression-related symptoms (β=−2.61; SD 0.88; *t*=−2.97; *P*=.003), whereas there was no significant difference between severe and moderate depression-related symptoms (β=−.02; SD 0.72; *t*=−0.03; *P*=.97). The corresponding effect sizes in the symptom severity groups were *d*=−0.80 (95% CI −1.23 to −0.37) for severe depression-related symptoms, *d*=−1.01 (95% CI −1.36 to −0.66) for moderate depression-related symptoms, and *d*=−0.50 (95% CI −0.95 to −0.05) for mild depression-related symptoms.

Clinically Reliable Change in Insomnia

Furthermore, 50.8% (128/252) of participants reporting insomnia severity scores at the postintervention measure reported a clinically reliable change in the insomnia measure (measured using the ISI) from baseline to postintervention measure, and 2% (5/252) experienced a deterioration in symptoms. In this study, the cutoff score was 11.73, indicating a value of 2 SDs below the mean ISI score of the population at the premeasurement time point. At the posttreatment time point, 24.2% (61/252) of participants showed a close to symptom-free status on the insomnia severity measure as defined by an ISI score of <8. The reliable change, remission, and deterioration rates are shown in Table 4. In total, 50.8% (128/252) of participants showed a clinically reliable change, and 24.2% (61/252) had a close to symptom-free status. The reliable change, remission, and deterioration rates are shown in Table 4.

Adherence

Table 7 reports the number and percentage of participants completing all intervention modules as intended as well as those receiving a minimal adequate dose of the intervention (at least 5 modules).

Table 7. Participant adherence.[a]

	Participants, n (%)	Completed all modules as intended, n (%)	Received a minimal adequate dose of the intervention (at least 5 modules), n (%)
All participants	851 (100)	405 (47.6)	474 (55.7)
Per intervention			
Mood	352 (41.4)	184 (52.3)	210 (59.7)
Stress	461 (54.2)	202 (43.8)	243 (52.7)
Both	22 (2.6)	18 (81.8)	20 (90.9)
Per guidance format			
Semistandardized	563 (66.2)	298 (52.9)	344 (61.1)
Feedback on demand	253 (29.7)	98 (38.7)	120 (47.4)
Self-guided	35 (4.1)	9 (25.7)	10 (28.6)
Stress intervention and guidance format			
Semistandardized stress	191 (22.4)	99 (51.8)	118 (61.8)
Feedback-on-demand stress	241 (28.3)	95 (39.4)	116 (48.1)
Self-guided stress	29 (3.4)	8 (27.6)	9 (31.0)

[a]In the mood training, only 1 participant received the feedback-on-demand and 2 self-guided interventions; therefore, adherence was not reported by guidance format for the mood intervention.

There was a significant effect of guidance on treatment adherence (χ^2_2=49.6; *P*<.001). Planned comparisons of the guidance groups revealed that participants receiving semistandardized guidance (563/851, 66.2%; mean 7, SD 4.8; range 1-25) completed significantly more sessions than participants in the self-guided group (35/851, 4.1%; mean 4.8, SD 4.7; range 1-16; β=−.39; SE 0.08; *z*=−4.93; *P*<.001), participants in the semistandardized guided group completed significantly more sessions than participants in the feedback-on-demand guided group (253/851, 29.7%; mean 6, SD 4.8; range 1-20; β=.16; SE 0.03; *z*=5.26; *P*<.001), and participants in the feedback-on-demand guided group completed significantly more sessions than participants in the self-guided group (β=−.23; SE 0.08; *z*=−2.83; *P*=.005).

There was a significant effect of the baseline depressive symptom severity on treatment adherence (χ^2_2=20.9, *P*<.001). Planned comparisons of the guidance groups revealed that participants reporting severe depressive symptoms (327/851, 38.4%; mean 6.9, SD 4.6; range 1-25) completed significantly more sessions than participants reporting mild depressive symptoms (214/851, 25.1%; mean 6, SD 5.1; range 1-24; β=−.15; SE 0.04; *z*=−4.26; *P*<.001), and participants reporting moderate depressive symptoms (310/851, 36.4%; mean 6.8, SD 4.9; range 1-20) completed significantly more sessions than participants reporting mild depressive symptoms (β=.14; SE 0.04; *z*=3.85; *P*<.001). There was no difference between participants reporting moderate and severe symptoms with regard to the number of sessions completed (β=−.01; SE 0.03; *z*=−0.43; *P*=.67).

XSL•FO
RenderX

Additional analyses showed that neither the baseline depressive symptom severity (χ^2_1=0.8; *P*=.39), perceived stress (χ^2_1=2.3; *P*=.13), nor insomnia severity (χ^2_1=0.0; *P*=.83) predicted whether clients started the treatment or not.

Participant Satisfaction

Of all participants, 31.8% (271/851) provided data on the CSQ-8. Overall, satisfaction with the intervention, measured using the CSQ-8, was high. Most participants (227/271, 83.8%) indicated that they were satisfied in an overall sense ("very satisfied" or "mostly satisfied"). They rated the training as being of high quality (249/271, 91.9%) and the type of intervention they wanted to receive (223/271, 82.3%). They indicated that the intervention met their needs (264/271, 97.4% "almost all" and "most of them") and helped them deal effectively with problems (231/271, 85.2%). Moreover, 81.9% (222/271 were satisfied with the amount of help they received and would use the intervention again if they needed to (224/271, 82.7%), and 89.3% (242/271) would recommend it to a friend in need of similar help.

Discussion

Principal Findings

This study investigated the use of, effectiveness of, adherence to, and participant satisfaction with 2 iCBT interventions addressing depression- and stress-related symptoms within routinely collected data. A total of 851 participants who reported a PHQ-8 score of >5 and who started the GET.ON Mood Enhancer and GET.ON Stress interventions (today HelloBetter) were included in the reported analysis. Approximately half (461/851, 54.2%) of the participants followed the stress intervention, and of those participants following the stress intervention, 41% (189/461) received semistandardized guidance, 52% (240/461) received feedback-on-demand guidance, and 7% (32/461) followed the self-guided intervention. Of those following the mood intervention, all but 3 patients (349/352, 99.2%) received semistandardized guidance. The results confirm the effect of the service in reducing depressive symptom severity, perceived stress, and insomnia in routine care. Guidance significantly moderated both the effectiveness of and adherence to the interventions regarding reducing depressive symptom severity. Approximately half (150/289, 51.9%) of the participants showed a reliable change in depressive symptom severity after treatment, and 3.8% (11/289) showed reliable symptom deterioration. In such cases, participants were referred to routine health care services. Participant satisfaction was high, and across all patient groups and both interventions, approximately half (405/851, 47.6%) of the participants completed all modules of the intervention provided.

The results indicate the effectiveness of the service in reducing depressive symptom severity, with a large within-group effect size of *d*=−0.92 (ranging from *d*=−1.36 in the group of participants with severe depression to *d*=−0.21 in the group of participants with mild depression). Furthermore, the results indicate the interventions to be effective in reducing perceived stress, with a large effect size of *d*=−1.02 (ranging from *d*=−1.25 in the group of participants with moderate depression to *d*=−0.56 in the group of participants with mild depression), and insomnia severity, with a medium effect size of *d*=−0.75 (ranging from *d*=−1.01 in the group of participants with severe depression to *d*=−0.22 in the group of self-guided participants). The within-group effect sizes for the guided intervention correspond to the effect sizes reported for similar interventions in routine care [65-68]. Furthermore, these results indicate the successful implementation of iCBT interventions in (German) mental health care, where the interventions show similar effects in routine care as in previous RCTs reporting within-group effects of *d*=1.54 [37].

Interestingly, participants receiving the self-guided intervention showed similar symptom reduction as the participants in the semistandardized guidance group. This result might be due to the recommendation procedures based on the screening questions and the resulting differences between the groups. Participants received a recommendation on the treatment and guidance format. Although they could choose to opt out of the semistandardized guidance after it was recommended, most participants (840/851, 98.7%) followed the recommendations regarding treatment and guidance format. People in the self-guided group might have believed that they truly did not need more guidance, and therefore, this guidance format sufficed for them. Participants experiencing high depression-related symptoms, whether combined with high stress-related symptoms, received the semistandardized guided intervention. Less affected patients received the self-guided intervention, and this group could benefit from this low-threshold intervention. This also indicates the usefulness of the GET.ON recommendation scheme. Participants in the feedback-on-demand group showed a lower symptom reduction. Participants receiving feedback-on-demand interventions who were recommended this type showed mild to moderate depressive symptoms. A potential explanation for the lower but still significant symptom reduction of *d*=−0.72 might be that participants in need of more support did not make use of the option to ask questions or request feedback from an online coach. Future studies should analyze this participant group, potentially looking at individual symptom trajectories and routine outcome monitoring, and evaluate the need for additional support of potential subgroups.

In this study, 59.6% (163/273) of participants were first-time help seekers reporting not having received previous psychotherapeutic treatment. This result underlines the great potential of iCBT interventions in reaching populations that otherwise would not seek or find adequate help. In addition, participants in this study were highly educated (503/843, 59.7% had the test level of education). Higher education is associated with a positive view of psychotherapy in general and higher service use in routine care [69,70], and the distribution of educational levels in our sample is comparable with distributions in trials on iCBT for the treatment of depression [13]. The higher proportion of female participants is in line with the use rates of psychotherapy found in epidemiological studies in Germany and elsewhere [71,72]. This effect might be explained by gender differences in help-seeking behavior rather than being related to iCBT service–related factors [73]. Future studies should focus

XSL•FO
RenderX

on ways to encourage men to use iCBT interventions. In addition, the mean age of 41.7 years was comparable with the results found in epidemiological studies on the use of face-to-face therapy [74] and with the results of the previous RCT on the efficacy of GET.ON Stress [37].

The high participant satisfaction reported by those who answered the CSQ-8 (ranging between 82% and 97% agreement) is comparable with participant satisfaction rates in similar interventions [75,76]. Although this result must be interpreted with caution as only a third of the participants (271/851, 31.8%) provided information on the CSQ-8, the result is an indicator of the acceptability of iCBT once people have chosen to use such an intervention.

The deterioration rate of 3.8% is distributed across all guidance formats as well as both the stress and mood interventions. Deterioration rates of this size have been identified in other studies [67,75,77] as well as in the results of the previous RCT on the efficacy of GET.ON Stress [37] and other comparable RCTs [78]. Symptom deterioration and other negative effects should be monitored, and appropriate care for participants should be provided. GET.ON has a protocolized system in place where participants showing symptom deterioration are referred to more intensive care. Further phase-IV trials should focus on possible predictors of deterioration, and the investigation of individual patient trajectories should be used to prevent potential deterioration.

Adherence rates in this study were similar to those observed in comparable studies [66,67,79] but lower than those yielded in a randomized, controlled setting [80,81]. Differences between adherence to iCBT interventions in routine care as compared with in RCTs might be due to the absence of the SMS text message coaching part in the stress management intervention, a format that has been proven to be effective in a similar setting [82]. Moreover, an assumed adherence-fostering effect of randomized controlled settings versus routine care might contribute to this difference [83]. Participants might be more likely to stop treatment once they have recovered, even in agreement with their assigned professional.

The effect of guidance on internet-based interventions' efficacy in reducing depressive symptom severity has also been investigated in other studies, and the first results indicate that guidance has a significant influence on adherence [15,16,80]. Zarski et al [80] found that adherence to GET.ON Stress delivered with feedback-on-demand guidance was equivalent to the intervention including semistandardized guidance within pooled RCT data [80]. In contrast, our data suggest that there was a difference between participants receiving feedback-on-demand or semistandardized guidance. This difference might be due to the recommendation into a specific guidance format and the subsequent fact that guidance was only offered to a specific group of participants. Another influence might be the different research settings—RCT versus routine care. Recently, Baumel et al [84] found that indications for trial settings have an impact on user engagement in self-guided interventions. A possible explanation might be that the feedback-on-demand guidance condition in RCTs still provides additional contacts because of routine study administration. This might have an additional adherence-fostering effect, adding to a difference between feedback-on-demand and semistandardized guidance not present in the routine care application of the service.

Limitations

Although we observed a significantly greater change in depressive symptom severity in the group of participants receiving semistandardized guidance than in the group receiving no guidance, it must be noted that participants with lower baseline symptom severity were assigned to partake in the feedback-on-demand guided or self-guided intervention, whereas participants with higher baseline symptom severity were assigned the semistandardized guided intervention. This resulted in baseline differences in symptom severity in these groups, and a greater change can be expected in a population with a higher burden. This also accounts for the significant difference between the mood and stress interventions—participants reporting a lower depressive symptom severity were mainly recommended the stress intervention, and participants with higher depressive symptom severity were recommended the mood intervention. Again, part of this difference must be assumed to be due to this baseline difference in the measure. Moreover, the group comparison of mild, moderate, and severe symptom severity should be interpreted with caution as the outcome and predictor are defined by the same measure. In this research setup, a causal relationship between guidance and effectiveness cannot be assumed or interpreted. Moreover, the limited total sample size of this implementation project, the small sample sizes in the different analysis groups, and the implementation in the German public health care context might limit the generalizability of results to people with depression overall. In addition, the absence of postdata poses a constraining impact on the generalizability and reliability of the results of this study. As an additional limitation, it needs to be mentioned that dividing the sample into subsets with regard to guidance and mood resulted in group effect sizes with large CIs. Furthermore, the report of effect sizes might be biased as they do not take the random variance of participants into account.

Conclusions

This study adds to the body of literature on the effectiveness of iCBT under routine care conditions. Its main strength is that the interventions described have been thoroughly researched and their clinical efficacy has been established. They are now implemented in German mental health care as tested. Data reported are routinely collected and, therefore, depict routine outcomes and adherence without the potential confounding of additional research. Such results are crucial as depression disorders are highly prevalent and costly and are substantially undertreated, and the presented results support the hypothesis that iCBT may well be able to help bridge this gap in reaching people in need of actual routine care.

To overcome some of the limitations of this study, future research should focus on the analysis of individual participant trajectories to identify characteristics of those participants not benefiting from the intervention provided. Furthermore, these findings must be replicated in different settings and samples to gain further knowledge of the influence of such factors on iCBT

XSL•FO
RenderX

in routine care. In addition, future studies could investigate further the recommendation processes, such as those applied in the GET.ON (HelloBetter) service, to ensure that the maximum number of participants receive the intervention best suited to their needs.

The results of this early implementation study indicate the effectiveness of semistandardized guided, feedback-on-demand guided, and self-guided iCBT interventions for depression- and stress-related symptoms under routine care conditions. These findings highlight that the provision of semistandardized, feedback-on-demand, and self-guided iCBT interventions is possible in the German mental health context and that such interventions can help reach more people in need of treatment for depressive or stress-related symptoms.

Acknowledgments

The health insurance companies as funders of this service intervention had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The MasterMind project was partially funded under the Information and Communication Technologies Policy Support Programme as part of the Competitiveness and Innovation Framework Programme of the European Community (grant agreement 621000). This funding body had no influence on the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to acknowledge the statistical consultation of Professor Johannes Zimmermann during the writing of this manuscript.

Data Availability

The data used in this research will be made available by the authors upon reasonable request.

Authors' Contributions

AE and DDE initiated and conceptualized the study and drafted the manuscript with initial feedback from all authors. EH and HH led the project and collected the data, with supervision from DDE and DL. AE and AR contributed to data preparation. AE analyzed the data with contribution from AR. All authors contributed to the review and editing process of this manuscript.

Conflicts of Interest

DDE reports to have received consultancy fees or served on the scientific advisory board of several companies such as Minddistrict, Sanofi, Lantern, Schön Kliniken, German health insurance companies (BARMER and Techniker Krankenkasse), and chambers of psychotherapists. DDE, HH, and EH are the founders of HelloBetter and the GET.ON Institut für Online Gesundheitstrainings GmbH (GET.ON), which aims to implement scientific findings related to digital health interventions in routine care. AE is employed by HelloBetter, GET.ON Institut für Online Gesundheitstrainings GmbH (GET.ON), as team lead of clinical research.

Multimedia Appendix 1
Sensitivity analysis.
[DOCX File , 35 KB - jmir_v26i1e42976_app1.docx]

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1&isAllowed=y [accessed 2023-08-23]
2. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 2013 Nov 09;382(9904):1575-1586. [doi: 10.1016/S0140-6736(13)61611-6] [Medline: 23993280]
3. Kessler RC. The costs of depression. Psychiatr Clin North Am 2012 Mar;35(1):1-14 [FREE Full text] [doi: 10.1016/j.psc.2011.11.005] [Medline: 22370487]
4. Cuijpers P, Karyotaki E, Reijnders M, Ebert DD. Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. Epidemiol Psychiatr Sci 2019 Feb;28(1):21-30 [FREE Full text] [doi: 10.1017/S2045796018000057] [Medline: 29486804]
5. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017 Feb;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
6. Mack S, Jacobi F, Gerschler A, Strehle J, Höfler M, Busch MA, et al. Self-reported utilization of mental health services in the adult German population--evidence for unmet needs? Results of the DEGS1-Mental Health module (DEGS1-MH). Int J Methods Psychiatr Res 2014 Sep;23(3):289-303 [FREE Full text] [doi: 10.1002/mpr.1438] [Medline: 24687693]

XSL•FO
RenderX

7. Ebert DD, Van Daele T, Nordgreen T, Karekla M, Compare A, Zarbo C, et al. Internet- and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. Eur Psychol 2018;23(2):167-187 [FREE Full text] [doi: 10.1027/1016-9040/a000318]
8. Andersson G, Carlbring P, Titov N, Lindefors N. Internet interventions for adults with anxiety and mood disorders: a narrative umbrella review of recent meta-analyses. Can J Psychiatry 2019 Jul;64(7):465-470 [FREE Full text] [doi: 10.1177/0706743719839381] [Medline: 31096757]
9. Ebert DD, Berking M, Cuijpers P, Lehr D, Pörtner M, Baumeister H. Increasing the acceptance of internet-based mental health interventions in primary care patients with depressive symptoms. A randomized controlled trial. J Affect Disord 2015 May 01;176:9-17. [doi: 10.1016/j.jad.2015.01.056] [Medline: 25682378]
10. Apolinário-Hagen J, Fritsche L, Bierhals C, Salewski C. Improving attitudes toward e-mental health services in the general population via psychoeducational information material: a randomized controlled trial. Internet Interv 2018 Jan 05;12:141-149 [FREE Full text] [doi: 10.1016/j.invent.2017.12.002] [Medline: 30135778]
11. Reins JA, Buntrock C, Zimmermann J, Grund S, Harrer M, Lehr D, et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 2021;90(2):94-106 [FREE Full text] [doi: 10.1159/000507819] [Medline: 32544912]
12. Königbauer J, Letsch J, Doebler P, Ebert DD, Baumeister H. Internet- and mobile-based depression interventions for people with diagnosed depression: a systematic review and meta-analysis. J Affect Disord 2017 Dec 01;223:28-40. [doi: 10.1016/j.jad.2017.07.021] [Medline: 28715726]
13. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, et al. Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. Clin Psychol Rev 2018 Jul;63:80-92. [doi: 10.1016/j.cpr.2018.06.007] [Medline: 29940401]
14. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
15. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215 [FREE Full text] [doi: 10.1016/j.invent.2014.08.003]
16. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
17. Etzelmueller A, Vis C, Karyotaki E, Baumeister H, Titov N, Berking M, et al. Effects of internet-based cognitive behavioral therapy in routine care for adults in treatment for depression and anxiety: systematic review and meta-analysis. J Med Internet Res 2020 Aug 31;22(8):e18100 [FREE Full text] [doi: 10.2196/18100] [Medline: 32865497]
18. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
19. Nobis S, Lehr D, Ebert DD, Baumeister H, Snoek F, Riper H, et al. Efficacy of a web-based intervention with mobile phone support in treating depressive symptoms in adults with type 1 and type 2 diabetes: a randomized controlled trial. Diabetes Care 2015 May;38(5):776-783 [FREE Full text] [doi: 10.2337/dc14-1728] [Medline: 25710923]
20. Ebert DD, Nobis S, Lehr D, Baumeister H, Riper H, Auerbach RP, et al. The 6-month effectiveness of Internet-based guided self-help for depression in adults with Type 1 and 2 diabetes mellitus. Diabet Med 2017 Jan;34(1):99-107. [doi: 10.1111/dme.13173] [Medline: 27334444]
21. Reins JA, Boß L, Lehr D, Berking M, Ebert DD. The more I got, the less I need? Efficacy of internet-based guided self-help compared to online psychoeducation for major depressive disorder. J Affect Disord 2019 Mar 01;246:695-705. [doi: 10.1016/j.jad.2018.12.065] [Medline: 30611913]
22. Buntrock C, Ebert D, Lehr D, Riper H, Smit F, Cuijpers P, et al. Effectiveness of a web-based cognitive behavioural intervention for subthreshold depression: pragmatic randomised controlled trial. Psychother Psychosom 2015;84(6):348-358. [doi: 10.1159/000438673] [Medline: 26398885]
23. Ebert DD, Buntrock C, Lehr D, Smit F, Riper H, Baumeister H, et al. Effectiveness of web- and mobile-based treatment of subthreshold depression with adherence-focused guidance: a single-blind randomized controlled trial. Behav Ther 2018 Jan;49(1):71-83. [doi: 10.1016/j.beth.2017.05.004] [Medline: 29405923]
24. Schlicker S, Weisel KK, Buntrock C, Berking M, Nobis S, Lehr D, et al. Do nonsuicidal severely depressed individuals with diabetes profit from internet-based guided self-help? Secondary analyses of a pragmatic randomized trial. J Diabetes Res 2019 May 14;2019:2634094 [FREE Full text] [doi: 10.1155/2019/2634094] [Medline: 31218230]
25. Buntrock C, Ebert DD, Lehr D, Smit F, Riper H, Berking M, et al. Effect of a web-based guided self-help intervention for prevention of major depression in adults with subthreshold depression: a randomized clinical trial. JAMA 2016 May 03;315(17):1854-1863. [doi: 10.1001/jama.2016.4326] [Medline: 27139058]

XSL•FO
RenderX

26. Nobis S, Ebert DD, Lehr D, Smit F, Buntrock C, Berking M, et al. Web-based intervention for depressive symptoms in adults with types 1 and 2 diabetes mellitus: a health economic evaluation. Br J Psychiatry 2018 Apr;212(4):199-206. [doi: 10.1192/bjp.2018.10] [Medline: 30071909]
27. Ebert DD, Kählke F, Buntrock C, Berking M, Smit F, Heber E, et al. A health economic outcome evaluation of an internet-based mobile-supported stress management intervention for employees. Scand J Work Environ Health 2018 Mar 01;44(2):171-182 [FREE Full text] [doi: 10.5271/sjweh.3691] [Medline: 29144535]
28. Andersson G, Titov N. Advantages and limitations of Internet-based interventions for common mental disorders. World Psychiatry 2014 Feb;13(1):4-11 [FREE Full text] [doi: 10.1002/wps.20083] [Medline: 24497236]
29. BPtK-Studie Zu Wartezeiten in Der Ambulanten Psychotherapeutischen Versorgung: Umfrage Der Landespsycho-Therapeutenkammern Und Der BPtK. Bundes Psychotherapeuten Kammer. 2011. URL: https://www.bptk.de/wp-content/uploads/2019/01/20180411_bptk_studie_wartezeiten_2018.pdf [accessed 2023-08-23]
30. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014 Apr;44(6):1303-1317 [FREE Full text] [doi: 10.1017/S0033291713001943] [Medline: 23931656]
31. Schnyder N, Panczak R, Groth N, Schultze-Lutter F. Association between mental health-related stigma and active help-seeking: systematic review and meta-analysis. Br J Psychiatry 2017 Apr;210(4):261-268. [doi: 10.1192/bjp.bp.116.189464] [Medline: 28153928]
32. Ebert DD, Heber E, Berking M, Riper H, Cuijpers P, Funk B, et al. Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 2016 May;73(5):315-323. [doi: 10.1136/oemed-2015-103269] [Medline: 26884049]
33. Ebert DD, Lehr D, Boß L, Riper H, Cuijpers P, Andersson G, et al. Efficacy of an internet-based problem-solving training for teachers: results of a randomized controlled trial. Scand J Work Environ Health 2014 Nov;40(6):582-596 [FREE Full text] [doi: 10.5271/sjweh.3449] [Medline: 25121986]
34. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJ. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry 2016 Oct 22;15(3):245-258 [FREE Full text] [doi: 10.1002/wps.20346] [Medline: 27717254]
35. Butler AC, Chapman JE, Forman EM, Beck AT. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev 2006 Jan;26(1):17-31. [doi: 10.1016/j.cpr.2005.07.003] [Medline: 16199119]
36. Weisel KK, Lehr D, Heber E, Zarski AC, Berking M, Riper H, et al. Severely burdened individuals do not need to be excluded from internet-based and mobile-based stress management: effect modifiers of treatment outcomes from three randomized controlled trials. J Med Internet Res 2018 Jun 19;20(6):e211 [FREE Full text] [doi: 10.2196/jmir.9387] [Medline: 29921562]
37. Heber E, Lehr D, Ebert DD, Berking M, Riper H. Web-based and mobile stress management intervention for employees: a randomized controlled trial. J Med Internet Res 2016 Jan 27;18(1):e21 [FREE Full text] [doi: 10.2196/jmir.5112] [Medline: 26818683]
38. Ebert DD, Lehr D, Heber E, Riper H, Cuijpers P, Berking M. Internet- and mobile-based stress management for employees with adherence-focused guidance: efficacy and mechanism of change. Scand J Work Environ Health 2016 Sep 01;42(5):382-394 [FREE Full text] [doi: 10.5271/sjweh.3573] [Medline: 27249161]
39. Harrer M, Adam SH, Fleischmann RJ, Baumeister H, Auerbach R, Bruffaerts R, et al. Effectiveness of an internet- and app-based intervention for college students with elevated stress: randomized controlled trial. J Med Internet Res 2018 Apr 23;20(4):e136 [FREE Full text] [doi: 10.2196/jmir.9293] [Medline: 29685870]
40. Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients: results from the medical outcomes study. RAND Corporation. 1989. URL: http://www.rand.org/content/dam/rand/pubs/notes/2009/N3037.pdf [accessed 2023-08-23]
41. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005 Jan;365(9453):82-93. [doi: 10.1016/S0140-6736(04)17670-8] [Medline: 15639683]
42. Singal AG, Higgins PD, Waljee AK. A primer on effectiveness and efficacy trials. Clin Transl Gastroenterol 2014 Jan 02;5(1):e45 [FREE Full text] [doi: 10.1038/ctg.2013.13] [Medline: 24384867]
43. Ebert DD, Baumeister H. Internet-based self-help interventions for depression in routine care. JAMA Psychiatry 2017 Aug 01;74(8):852-853. [doi: 10.1001/jamapsychiatry.2017.1394] [Medline: 28636717]
44. van der Lem R, van der Wee NJ, van Veen T, Zitman FG. Efficacy versus effectiveness: a direct comparison of the outcome of treatment for mild to moderate depression in randomized controlled trials and daily practice. Psychother Psychosom 2012;81(4):226-234. [doi: 10.1159/000330890] [Medline: 22584117]
45. Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, et al. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prev Sci 2005 Sep;6(3):151-175. [doi: 10.1007/s11121-005-5553-y] [Medline: 16365954]
46. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002 Sep 01;32(9):509-515 [FREE Full text] [doi: 10.3928/0048-5713-20020901-06]
47. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

48. Vis C, Kleiboer A, Prior R, Bønes E, Cavallo M, Clark SA, et al. Implementing and up-scaling evidence-based eMental health in Europe: the study protocol for the MasterMind project. Internet Interv 2015 Nov;2(4):399-409 [FREE Full text] [doi: 10.1016/j.invent.2015.10.002]
49. Corson K, Gerrity MS, Dobscha SK. Screening for depression and suicidality in a VA primary care setting: 2 items are better than 1 item. Am J Manag Care 2004 Nov;10(11 Pt 2):839-845 [FREE Full text] [Medline: 15609737]
50. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396 [FREE Full text] [doi: 10.2307/2136404]
51. Reis D, Lehr D, Heber E, Ebert DD. The German version of the Perceived Stress Scale (PSS-10): evaluation of dimensionality, validity, and measurement invariance with exploratory and confirmatory bifactor modeling. Assessment 2019 Oct;26(7):1246-1259. [doi: 10.1177/1073191117715731] [Medline: 28627220]
52. Bastien CH, Vallières A, Morin CM. Validation of the insomnia severity index as an outcome measure for insomnia research. Sleep Med 2001 Jul;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]
53. (Model) Professional Code for Physicians in Germany - MBO-Ä 1997 - The Resolutions of the 121st German Medical Assembly 2018 in Erfurt as amended by a Resolution of the Executive Board of the German Medical Association (English version). German Medical Association. URL: https://www.bundesaerztekammer.de/fileadmin/user_upload/_old-files/downloads/pdf-Ordner/MBO/MBO-AE_EN_2018.pdf
54. Lazarus RS, Folkman S. Stress, Appraisal, and Coping. New York, NY: Springer; 1984.
55. Heber E, Ebert DD, Lehr D, Nobis S, Berking M, Riper H. Efficacy and cost-effectiveness of a web-based and mobile stress-management intervention for employees: design of a randomized controlled trial. BMC Public Health 2013 Jul 15;13:655 [FREE Full text] [doi: 10.1186/1471-2458-13-655] [Medline: 23855376]
56. Erbe D, Eichert HC, Rietz C, Ebert D. Interformat reliability of the patient health questionnaire: validation of the computerized version of the PHQ-9. Internet Interv 2016 Jun 27;5:1-4 [FREE Full text] [doi: 10.1016/j.invent.2016.06.006] [Medline: 30135800]
57. Cohen S, Janicki-Deverts D. Who's stressed? Distributions of psychological stress in the United States in probability samples from 1983, 2006, and 2009. J Appl Soc Psychol 2012;42(6):1320-1334 [FREE Full text] [doi: 10.1111/j.1559-1816.2012.00900.x]
58. Nguyen TD, Attkisson CC, Stegner BL. Assessment of patient satisfaction: development and refinement of a service evaluation questionnaire. Eval Program Plann 1983;6(3-4):299-313. [doi: 10.1016/0149-7189(83)90010-1] [Medline: 10267258]
59. Boß L, Lehr D, Reis D, Vis C, Riper H, Berking M, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234 [FREE Full text] [doi: 10.2196/jmir.5952] [Medline: 27582341]
60. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics 1946 Dec;2(6):110 [FREE Full text] [doi: 10.2307/3002019]
61. Feingold A. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. Psychol Methods 2009 Mar;14(1):43-53 [FREE Full text] [doi: 10.1037/a0014699] [Medline: 19271847]
62. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014 Sep 02;34(5):502-508. [doi: 10.1111/opo.12131] [Medline: 24697967]
63. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991 Feb;59(1):12-19. [doi: 10.1037//0022-006x.59.1.12] [Medline: 2002127]
64. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: http://www.R-project.org [accessed 2023-08-23]
65. Hedman E, Ljótsson B, Rück C, Bergström J, Andersson G, Kaldo V, et al. Effectiveness of internet-based cognitive behaviour therapy for panic disorder in routine psychiatric care. Acta Psychiatr Scand 2013 Dec;128(6):457-467. [doi: 10.1111/acps.12079] [Medline: 23406572]
66. Mathiasen K, Riper H, Andersen TE, Roessler KK. Guided internet-based cognitive behavioral therapy for adult depression and anxiety in routine secondary care: observational study. J Med Internet Res 2018 Nov 28;20(11):e10927 [FREE Full text] [doi: 10.2196/10927] [Medline: 30487118]
67. Ruwaard J, Lange A, Schrieken B, Dolan CV, Emmelkamp P. The effectiveness of online cognitive behavioral treatment in routine clinical practice. PLoS One 2012;7(7):e40089 [FREE Full text] [doi: 10.1371/journal.pone.0040089] [Medline: 22792217]
68. Morrison C, Walker G, Ruggeri K, Hacker Hughes JH. An implementation pilot of the MindBalance web-based intervention for depression in three IAPT services. Cogn Behav Ther 2014 Nov 14;7:e15 [FREE Full text] [doi: 10.1017/s1754470x14000221]
69. Petrowski K, Hessel A, Körner A, Weidner K, Brähler E, Hinz A. Die Einstellung zur Psychotherapie in der Allgemeinbevölkerung. Psychother Psych Med 2014 Feb;64(02):82-85. [doi: 10.1055/s-0033-1361155]
70. Gallas C, Kächele H, Kraft S, Kordy H, Puschner B. Inanspruchnahme, verlauf und ergebnis ambulanter psychotherapie: Befunde der TRANS-OP-studie und deren implikationen für die richtlinienpsychotherapie. Psychotherapeut 2008 Nov 9;53(6):414-423 [FREE Full text] [doi: 10.1007/s00278-008-0641-7]

XSL•FO
RenderX

71. Albani C, Blaser G, Geyer M, Schmutzer G, Goldschmidt S, Brähler E. Wer nimmt in Deutschland ambulante Psychotherapie in Anspruch? Psychother Psychosom Med Psychol 2009 Jul;59(7):281-283. [doi: 10.1055/s-0028-1103267] [Medline: 19274608]
72. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]
73. Thompson AE, Anisimowicz Y, Miedema B, Hogg W, Wodchis WP, Aubrey-Bassler K. The influence of gender and other patient characteristics on health care-seeking behaviour: a QUALICOPC study. BMC Fam Pract 2016 Mar 31;17:38 [FREE Full text] [doi: 10.1186/s12875-016-0440-0] [Medline: 27036116]
74. Epping J, Muschik D, Geyer S. Social inequalities in the utilization of outpatient psychotherapy: analyses of registry data from German statutory health insurance. Int J Equity Health 2017 Aug 16;16(1):147 [FREE Full text] [doi: 10.1186/s12939-017-0644-5] [Medline: 28814311]
75. Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. The first 30 months of the MindSpot clinic: evaluation of a national e-mental health service against project objectives. Aust N Z J Psychiatry 2017 Dec;51(12):1227-1239. [doi: 10.1177/0004867416671598] [Medline: 27733709]
76. Etzelmueller A, Radkovsky A, Hannig W, Berking M, Ebert DD. Patient's experience with blended video- and internet based cognitive behavioural therapy service in routine care. Internet Interv 2018 Jun;12:165-175 [FREE Full text] [doi: 10.1016/j.invent.2018.01.003] [Medline: 30135780]
77. Hadjistavropoulos HD, Nugent MM, Alberts NM, Staples L, Dear BF, Titov N. Transdiagnostic internet-delivered cognitive behaviour therapy in Canada: an open trial comparing results of a specialized online clinic and nonspecialized community clinics. J Anxiety Disord 2016 Aug;42:19-29 [FREE Full text] [doi: 10.1016/j.janxdis.2016.05.006] [Medline: 27244278]
78. Rozental A, Magnusson K, Boettcher J, Andersson G, Carlbring P. For better or worse: an individual patient data meta-analysis of deterioration among participants receiving internet-based cognitive behavior therapy. J Consult Clin Psychol 2017 Feb;85(2):160-177. [doi: 10.1037/ccp0000158] [Medline: 27775414]
79. Hedman E, Ljótsson B, Kaldo V, Hesser H, El Alaoui S, Kraepelien M, et al. Effectiveness of internet-based cognitive behaviour therapy for depression in routine psychiatric care. J Affect Disord 2014 Feb;155:49-58. [doi: 10.1016/j.jad.2013.10.023] [Medline: 24238951]
80. Zarski AC, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146 [FREE Full text] [doi: 10.2196/jmir.4493] [Medline: 27357528]
81. van Ballegooijen W, Cuijpers P, van Straten A, Karyotaki E, Andersson G, Smit JH, et al. Adherence to Internet-based and face-to-face cognitive behavioural therapy for depression: a meta-analysis. PLoS One 2014 Jul 16;9(7):e100674 [FREE Full text] [doi: 10.1371/journal.pone.0100674] [Medline: 25029507]
82. Eckert M, Ebert DD, Lehr D, Sieland B, Berking M. Does SMS-support make a difference? Effectiveness of a two-week online-training to overcome procrastination. A randomized controlled trial. Front Psychol 2018 Jul 5;9:1103 [FREE Full text] [doi: 10.3389/fpsyg.2018.01103] [Medline: 30026713]
83. Ebert DD, Baumeister H. nternet- und mobil-basierte Interventionen in der Psychotherapie: Ein Überblick. Psychotherapeutenjournal 2016;1:22-31 [FREE Full text] [doi: 10.1515/pubhef-2015-0063]
84. Baumel A, Edan S, Kane JM. Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs. Transl Behav Med 2019 Nov 25;9(6):1020-1033. [doi: 10.1093/tbm/ibz147] [Medline: 31689344]

Abbreviations

CSQ-8: Client Satisfaction Questionnaire–8
HIC: health insurance company
iCBT: internet-based cognitive behavioral therapy
ISI: Insomnia Severity Index
PHQ-8: Patient Health Questionnaire–8
PHQ-9: Patient Health Questionnaire–9
PSS-10: Perceived Stress Scale
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by G Eysenbach; submitted 26.09.22; peer-reviewed by T Tillmann, M Löbner; comments to author 15.11.22; revised version received 27.04.23; accepted 03.05.23; published 01.02.24.

Please cite as:
Etzelmueller A, Heber E, Horvath H, Radkovsky A, Lehr D, Ebert DD
The Evaluation of the GET.ON Nationwide Web-Only Treatment Service for Depression- and Stress-Related Symptoms: Naturalistic Trial
J Med Internet Res 2024;26:e42976
URL: https://www.jmir.org/2024/1/e42976
doi:10.2196/42976
PMID:38300701

©Anne Etzelmueller, Elena Heber, Hanne Horvath, Anna Radkovsky, Dirk Lehr, David Daniel Ebert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Mobile Game–Based Intervention to Enhance Child Safety: Randomized Controlled Trial

Rosa S Wong[1], PhD; Keith T S Tung[2], PhD; Frederick K W Ho[3], PhD; Wilfred H S Wong[2], PhD; Chun Bong Chow[2], MD; Ko Ling Chan[4], PhD; King Wa Fu[5], PhD; Patrick Ip[2], MD

[1]Department of Special Education and Counselling, The Education University of Hong Kong, Hong Kong, China (Hong Kong)
[2]Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong, China (Hong Kong)
[3]Institute of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom
[4]Department of Applied Social Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[5]Journalism and Media Studies Centre, The University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Patrick Ip, MD
Department of Paediatrics and Adolescent Medicine
The University of Hong Kong
Pok Fu Lam
Hong Kong
China (Hong Kong)
Phone: 852 22554090
Email: patricip@hku.hk

Abstract

Background: Evidence supports the effectiveness of serious games in health education, but little is known about their effects on the psychosocial well-being of children in the general population.

Objective: This study aimed to investigate the potential of a mobile game–based safety education program in improving children's safety and psychosocial outcomes.

Methods: Safe City is a mobile roleplaying game specifically designed to educate children in Hong Kong about safety. This randomized controlled trial included 340 children in grades 4 through 6. Intervention arm participants (n=170) were instructed to play the Safe City mobile game for 4 weeks, whereas control arm participants (n=170) received a safety booklet. All participants completed a survey on safety knowledge and behaviors and psychosocial problems at baseline (T1), 1 month postintervention (T2), and 3 months postintervention (T3). Cumulative game scores and mini-game performance were analyzed as a proxy for the extent of exposure to the game. Outcome data were analyzed using 2-sample 2-tailed *t* tests to compare mean change from T1 to T2 and to T3 for intervention versus control arm participants. The association of game use with outcome changes postintervention was analyzed using generalized additive models.

Results: No significant differences were found in mean changes between the intervention and control arms. However, use analyses showed that higher game scores were associated with improvements in safe behavior (P=.03) and internalizing problems (P=.01) at T3. Matching and Spot the Danger mini-game performance significantly predicted improvements in safety knowledge at T2 and T3.

Conclusions: Analysis of use has shown that playing the Safe City mobile game can result in significant improvements in safety knowledge and reductions in unsafe behavior and internalizing problems. These findings provide evidence for the positive impact of serious games on psychological and social well-being, highlighting the potential of technology-driven interventions to assist children in learning about safety and preventing injuries.

Trial Registration: ClinicalTrials.org NCT04096196; https://clinicaltrials.gov/show/NCT04096196

International Registered Report Identifier (IRRID): RR2-10.2196/17756

(J Med Internet Res 2024;26:e51908) doi:10.2196/51908

XSL•FO
RenderX

KEYWORDS

mobile game; injury prevention; safety education; unintentional injuries; gamification; game; games; gaming; safety; injury; injuries; gamify; gamified; child; children; youth; pediatric; pediatrics; danger; hazard; hazards; RCT; randomized; controlled trial; controlled trials; mobile phone

Introduction

Injuries remain a top contributor to mortality and disability worldwide, despite a decline on a global level [1]. Notably, children's natural curiosity, energy, and desire for exploration can make them more susceptible to unintentional injuries [2]. Therefore, it is crucial to provide safety education to protect them from harm. Traditional teaching methods such as lectures, on-site demonstrations, and paper-based materials have shown success in preventing injuries, but they have limitations that could hinder their effectiveness in modern learning environments. For example, these methods may not provide enough stimulation or interaction with the material being taught, which can make it challenging for young students to feel connected with the subject matter and retain critical safety information. Recognizing the drawbacks of these conventional teaching methods, online technology is increasingly being used in health promotion because it offers a more captivating, personalized, and accessible approach to providing health education and information [3].

In recent years, there has been a growing trend of using games as an educational tool. Research has shown that this approach can improve students' critical thinking skills, as well as increase their motivation and engagement in the learning process. [4]. Game-based learning can also stimulate curiosity and imagination, making the learning experience more enjoyable and interesting [5]. Serious games, which are computerized games developed for educational or training purposes, are more effective and efficient in achieving learning objectives compared to traditional teaching methods [6]. Meta-analyses have indicated that video game features such as engaging storylines and immediate feedback are beneficial for the promotion of health-related behaviors such as symptom management, physical activity, and diet [7,8]. On the other hand, only a limited number of research studies have used online game technology to enhance safety awareness and education among children. To our knowledge, only 1 trial study has used online game technology to enhance safety awareness and education among children. In the study, a mobile game app named iBsafe was created to reduce the risk of injuries associated with bicycles and dogs among kindergarten children in the United States [9]. Despite positive trial results [10], the intervention only lasted for 1 week and the follow-up assessment was done a week after the randomization. Therefore, the study only revealed short-term effects of the intervention. It remains unclear whether mobile game–based safety education programs can have a sustained impact on the safety knowledge and skills of young students.

Serious games serve not only as a means to enhance learning but are also frequently used as a tool in mental health interventions. Psychosocial problems are typically categorized into internalizing (eg, anxiety, depression, and social withdrawal) and externalizing problems (eg, hyperactivity, impulsivity, and rule violation) [11]. While the pooled evidence suggests a small yet significant association between increased screen time and a higher occurrence of internalizing and externalizing problems [12], trials provide evidence of the effectiveness of serious games, which are often based on cognitive behavioral therapy principles or incorporate elements of both exercise and gaming, in reducing symptoms of depression and anxiety [13,14]. A meta-analysis also reported that digital games created for therapeutic purposes were consistently associated with a reduction in negative emotional experiences among children and early adolescents [15]. Commercial video games, which are primarily designed for entertainment rather than therapy, were found to be as advantageous as clinical video games in producing feelings of joy, happiness, and positivity among older adults and patients [16]. Another meta-analysis found that the type of game (whether educational or therapeutic) did not moderate the impact of serious games on the mental health of children and adolescents, but the duration of the game and the length of each gaming session can have an impact [17]. Although the results yielded positive outcomes, notable discrepancies were observed, and evidence on externalizing problems remains limited, highlighting the necessity for further research to affirm the mental health consequences of serious games in the younger generation.

In Hong Kong, injuries pose a significant risk to childhood mortality and can result in long-term disabilities. A previous study revealed that between 2001 and 2012, 742,552 children and adolescents sought treatment for injuries at local public hospital emergency departments in Hong Kong. The majority of these injuries were unintentional, with home accidents (39%), sports-related incidents (18%), and road traffic accidents (4%) being the most common causes. The annual direct medical cost of these injuries amounted to US $29.4 million, with indirect costs reaching up to 2-3 times higher [18]. Among different age groups, domestic and sports injuries were among the top 3 leading causes of injury for children aged 5-9, 10-14, and 15-19 years, respectively. On the other hand, traffic injuries were more prevalent among younger children [18]. During the same study period, over 753 children lost their lives due to injuries, with traffic accidents accounting for the highest rate of unintentional injury-related deaths (0.69 deaths per 100,000 people) [18]. A study comparing childhood injuries in Hong Kong with 4 other countries (United States, Nigeria, Canada, and Greece) found that Hong Kong had the highest rate of falls (44%), similar to Greece, while transport injuries in Hong Kong (12%) were lower than in Nigeria (27%), but higher than in the other 3 countries (United States [6%], Canada [8%], and Greece [4%]) [19,20]. These findings underscore the importance of creating better tactics to educate children in Hong Kong about safety, particularly in the domains of home safety, sports safety, and traffic safety.

Accordingly, this study used a randomized controlled trial (RCT) design to examine the effectiveness of digital gaming methods

XSL•FO
RenderX

compared to traditional booklet techniques in enhancing safety in the 3 specific domains of home safety, sports safety, and traffic safety, as well as in improving psychosocial outcomes among young students. In contrast to the traditional booklet method, the digital gaming methods included various mini-games such as Spot the Danger, Matching, multiple choice questions (MCQs), and Whac-A-Mole. These mini-games required players to engage in a learning loop where they applied, tested, and updated their safety knowledge. We hypothesized that students who received a mobile game–based safety education program would show significantly greater improvements in safety knowledge, safe behaviors, and internalizing problems compared to those who received a safety booklet. Additionally, we aimed to explore whether the level of improvement in postintervention outcomes was influenced by the total scores accumulated during gameplay. We hypothesized that higher cumulative scores would be associated with larger improvements in postintervention outcomes.

Methods

Study Design

We conducted an RCT to compare the effectiveness of a mobile game–based strategy and a traditional strategy to improve safety and psychosocial outcomes in a sample of 340 children aged 8-13 years in Hong Kong. The intervention was carried out over 30 days from enrollment to the trial end. Assessment was performed via a survey at baseline (T1), at 4 weeks postintervention (T2), and at 12 weeks postintervention (T3; Figure 1). A detailed study protocol has been published elsewhere [21]. The registered protocol for this study initially stated a larger sample size, but it was later reduced to 340 participants. The reason for the reduction was that the effect size cited in the registered protocol was based on preliminary results from a formative research study conducted with children and adolescents in the United Kingdom who likely had different learning preferences and experiences related to road safety compared to Hong Kong children [22].

XSL•FO
RenderX

Figure 1. Study flowchart. ITT: intention-to-treat; PP: per-protocol.

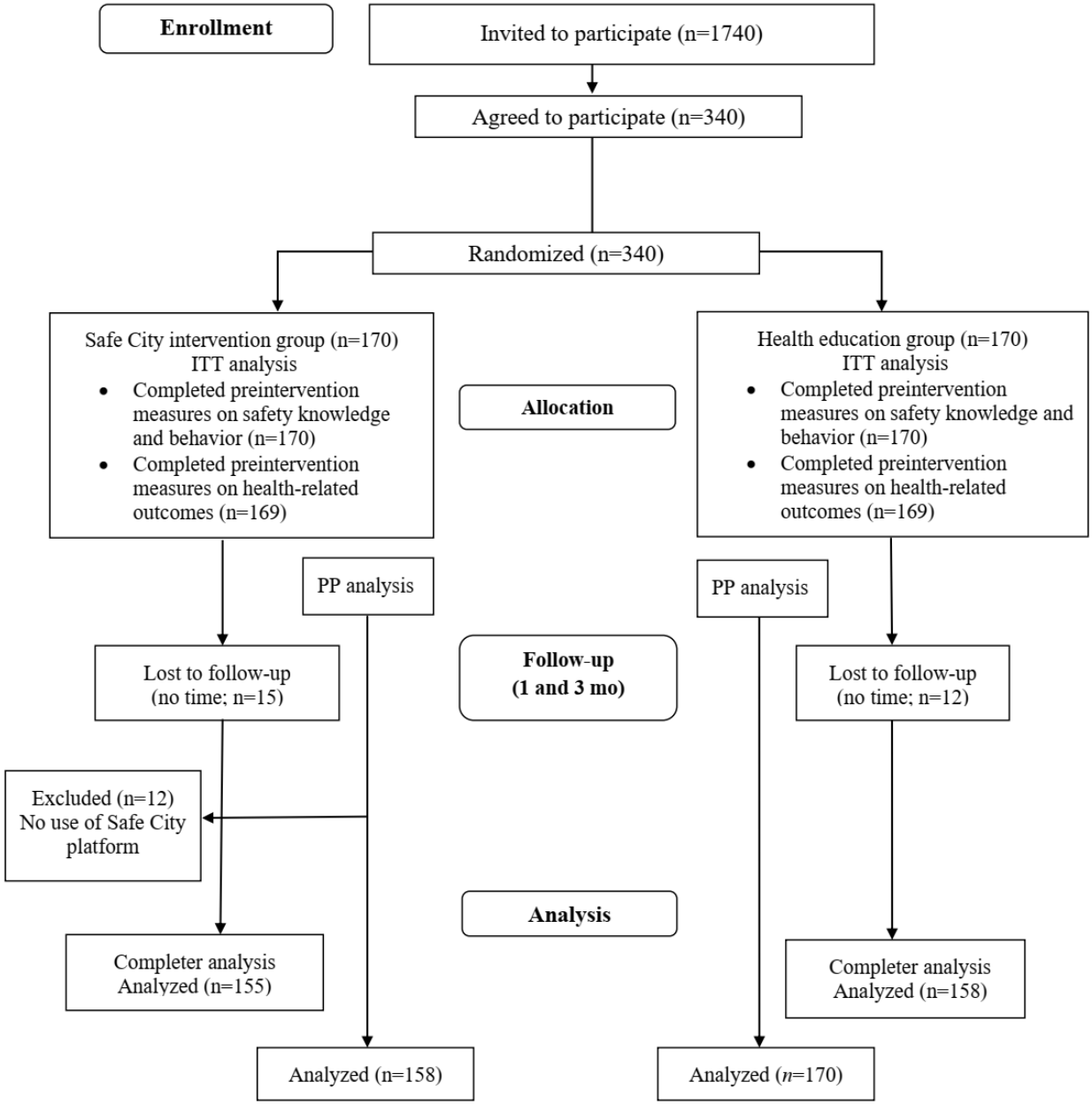

Sample Size Estimation

In 2020, the initial research in Hong Kong that used mobile apps and gamification methods to administer a physical activity intervention was published [23]. This study demonstrated a medium-to-large intervention effect on hyperactivity or inattention symptoms among healthy Hong Kong children aged 9-12 years, with an effect size of d=0.40 at a 6-month follow-up. Since hyperactivity or inattention symptoms are closely related to injuries and safety actions, and the results were derived from the local child population, we considered this result more suitable for estimating the sample size for this project. Based on the effect size of d=0.40, a minimum sample size of 200 (100 in each group) would be sufficient to achieve 80% statistical power at a 5% significance level.

Recruitment and Participants

We recruited participants by distributing promotional leaflets through local mainstream primary schools across 7 districts in Hong Kong that had a significant prevalence of traffic, domestic, and sports injuries. These leaflets emphasized the appealing features of our newly developed educational mobile game. To be eligible for participation in this study, participants had to be in grade 4 to grade 6 and proficient in reading and writing traditional Chinese without the aid of an interpreter, as well as have access to a smartphone with internet connectivity. Exclusion criteria involved having neurodevelopmental disorders or suspected learning difficulties, as reported by teachers, parents, or caregivers.

Procedure

Potential participants and their parents were directed to a study website with information about the project and web-based registration. After obtaining written consent for participation

XSL•FO
RenderX

in this study, participants were granted access to a web-based screening survey that included questions about their demographic information and medical history. If participants did not meet any of the exclusion criteria, they were included in this study and invited to complete the T1 survey electronically. The included participants were then grouped into blocks, with each block consisting of 20 participants, based on the sequence of completion times for their T1 survey. Within each block, a data analyst who was not involved in other aspects of this study randomly assigned participants to either the intervention or control condition, using the R package *blockrand*, in a 1:1 ratio. Assignments were revealed to participants after they had completed the T1 survey. Participants were reassessed at T2 and T3 using the same survey that was administered at T1.

Intervention

Participants assigned to the intervention arm played Safe City, a theoretically grounded roleplaying game that can work on both iOS and Android smartphones. The game's design draws on social cognitive theory [24] and experiential gaming model [25], with input from experts, educators in child safety and injury prevention, as well as local children and their parents. Safe City is designed to promote safety knowledge and behavior by gamifying traditional safety education materials with digital and attractive features such as city scenes, character customization, and a scoring system. The game also uses story missions to enhance motivation for learning by instructing players to create and manage a fictional character (safety inspector) whose mission is to explore and eliminate potential injury hazards in the game's world. The game world is divided into 5 zones, each dedicated to addressing safety risks in the 3 specific domains of home safety (2 zones), road safety (2 zones), and sport safety (1 zone). During the game process, learning happens in a continuous cycle. First, players identify potential injury risks and test their safety knowledge. Then, they apply this knowledge to ensure safety. Afterward, players reflect on the feedback received from the game and update their safety knowledge base accordingly [25].

Within each zone, there are 3 levels, each comprising of 5 components. In the first component, players are required to roleplay the correct way to cross a street by navigating through various checkpoints. In particular, the safety inspector (ie, the player) must adhere to standard street-crossing guidelines, such as avoiding jaywalking, while crossing a street in the game (Figure S1 in Multimedia Appendix 1). The remaining 4 components consist of 3 educational mini-games and 1 leisure-based mini-game, all of which are presented as checkpoints. The first educational mini-game (Spot the Danger) requires the safety inspector to perform a safety check and spot hazards in different areas within 35 seconds (Figure S2 in Multimedia Appendix 1); the second educational mini-game (Matching) requires the safety inspector to match safety signs with their corresponding description in 35 seconds (Figure S3 in Multimedia Appendix 1); and the third education mini-game is carried out in the format of MCQs, where the safety inspector is required to select a correct answer to the safety-related question within 35 seconds (Figure S4 in Multimedia Appendix 1). The educational mini-games award game points for correct answers and provide immediate feedback to the players by indicating which answer is correct after each question or gameplay. Lastly, the leisure-based mini-game adopts the Whac-A-Mole game design (Figure S5 in Multimedia Appendix 1). At the beginning of the gameplay, the safety inspector is presented with a list of safe and hazardous items. During the 35-second gameplay, the inspector is prompted to hit the items that randomly pop up from 9 holes. Whacking safe items earns points, but whacking dangerous items results in a loss of points. This experience of being rewarded for hitting safe items and penalized for hitting dangerous ones not only generates exhilaration but also motivates the player to continuously review and update their knowledge on safety. After the initial 35-second playtime, players have the option to replay any of the mini-games. However, to mitigate the risk of game addiction, every safety inspector begins with the same work energy level. A full energy level is indicated by 10, while no energy is represented by 0. When the energy runs out, players must wait for at least 2 hours to gain 1 energy step. In addition to the game zones, there is a designated rest area within the game known as the safety park. This area provides access to 3 other leisure-based mini-games that follow similar gameplay and rules as the Whac-A-Mole game but are presented in various styles. The points earned from participating in these mini-games determine the player's position on the global scoreboard, which is another game feature designed to encourage use.

Intervention arm participants were instructed to play the Safe City mobile game with unlimited access for 30 days and asked not to share their own account and gameplay information with others until after the T3 assessment. Control participants had no access to the Safe City mobile game platform but received safety education in a traditional format, namely a booklet containing information about childhood injury hazards and safety measures, during this study period.

Ethical Considerations

This study was approved by the ethics committee of the Institutional Review Board of the University of Hong Kong and Hospital Authority Hong Kong West Cluster (UW 19-028) and was registered with ClinicalTrials.gov (NCT04096196). Since the participants were aged younger than 18 years, their parents were required to provide separate written consent both for themselves and for their children to participate in this study. The data were deidentified before analysis and were securely stored in a password-protected file and on a password-protected computer. As an incentive, intervention players ranked in the top 20 were rewarded with a book coupon worth US $30.

Survey Measures

Sociodemographic Questionnaire

A brief questionnaire was used to collect sociodemographic information including the participant's gender and age. Parental educational attainment and household income were combined based on the factor loadings generated by the principal components analysis to create a composite family socioeconomic status index.

XSL•FO
RenderX

Safety Knowledge and Safe Behaviors

Safety knowledge and safe behaviors were measured using items adopted from the existing literature and questionnaires under the 3 contexts: home safety [26], road safety [22], and sports safety [27]. The wording of the items was modified and validated by local injury specialists to ensure their compliance with the safety regulations specific to the region. Participants provided ratings for their involvement in safe behavior on a 5-point Likert scale ranging from 0=never to 4=very often. For the safety knowledge items, a 4-point Likert scale was used, ranging from 1=strongly disagree to 4=strongly agree. The total scores, 1 for safety knowledge (Cronbach α=.61) and 1 for safe behavior (Cronbach α=.65) were calculated by averaging the relevant item scores and then transformed into a scale ranging from 0 to 100. Higher scores indicated a higher level of safety knowledge and safe behaviors.

Psychosocial Problems

Participants were asked to complete the psychosocial problem scales of the Strength and Difficulties Questionnaire, which included emotional symptoms, conduct problems, hyperactivity or inattention, and peer relationship problems [28]. The emotional symptoms and peer relationship problems scale scores were added up into an internalizing problem score (Cronbach α=.70), whereas the hyperactivity or inattention and conduct problems scale scores were added up into an externalizing problem score (Cronbach α=.79). Higher scores indicated more problems. The self-reported Strength and Difficulties Questionnaire version has been used on children between the ages of 8 and 13 years, and has demonstrated satisfactory psychometric properties [29].

Data Analysis

Before the analysis, we assessed the distribution, outliers, and missing data of all variables. We examined the skewness and kurtosis to evaluate the assumption of a normal distribution. Any potential outliers were eliminated if their values were ≥ 3 SDs from the mean of the group. Descriptive statistics were computed to characterize this study's groups. Group differences were examined using chi-square tests for categorical variables or independent 2-tailed *t* tests for continuous variables. Changes in safety and psychosocial outcomes from T1 were compared between the 2 study groups at T2 and T3 using independent *t* tests. Sensitivity analyses were conducted to assess robustness. Intention-to-treat (ITT), per-protocol (PP), and completer analyses were conducted. The PP group included all control participants and those intervention participants who logged into the Safe City platform and accumulated a game score above zero. The completer group included all participants who completed the primary measures at T1-T3. In addition, generalized additive models (GAMs) were applied to explore the linear and nonlinear relationships between Safe City game use and participant outcomes adjusting for gender, age, baseline parameter value, family socioeconomic status index, random assignment, and school cluster. The effective degrees of freedom (edf) estimated from the GAM were used to reflect the degree of nonlinearity in the use-outcome relationship, with an edf of 1 indicating a linear relationship and a larger edf indicating a more complex, nonlinear relationship. Missing outcome data at T2 and T3 were imputed using the last value carried forward. R software (version 4.2.1; R Foundation for Statistical Computing) was used to perform all data analyses. The significance level was set at $P<.05$ (2-sided).

Results

Participant Characteristics

We recruited and randomly assigned 340 participants to either control (n=170, 50%) or intervention (n=170, 50%). Demographics are reported in Table 1. The sample included 166 (48.8%) males and 174 (51.2%) females, with a mean age of 10.71 years (SD 0.91 years). Among the participants, 52.6% (179/340) had a monthly household income below the median household income of US $3835 in Hong Kong. Additionally, approximately 34% (232/680) of parents had completed tertiary education, which was comparable to the population statistics in Hong Kong. Baseline demographics were similar between intervention and control groups, with the exception that fathers and mothers in the control group were slightly more likely to have completed tertiary education than those in the intervention group, although these differences were not statistically significant ($P>.05$).

XSL•FO
RenderX

Table 1. Baseline participant characteristics.

Characteristic	Intervention participants (n=170), n (%)	Control participants (n=170), n (%)	*P* value
Sex			.19
Male	89 (52.4)	77 (45.3)	
Female	81 (47.6)	93 (54.7)	
Age (y), mean (SD)	10.70 (0.93)	10.72 (0.88)	.85
Monthly household income (US $)			.33
<3835	85 (50)	94 (53.3)	
≥3835	85 (50)	76 (44.7)	
Maternal education level			.30
Primary education or below	9 (5.3)	6 (3.5)	
Secondary education	107 (62.9)	105 (61.8)	
Tertiary education or above	54 (31.8)	56 (32.9)	
Missing	0 (0)	3 (1.8)	
Paternal education level			.06
Primary education or below	11 (6.5)	3 (1.8)	
Secondary education	98 (57.6)	97 (57.1)	
Tertiary education or above	59 (34.7)	63 (37.1)	
Missing	2 (1.2)	7 (4.1)	

Use

Of those randomized to the intervention condition, 100% (170/170) downloaded the Safe City app. In total, 96.4% (164/170) of participants logged into the platform at least once. Further, 92.9% (158/170) played at least 1 mini-game. The highest score accumulated over the 4-week intervention period was 14,598,100 and the lowest score was 4000, reflecting the wide-ranging use rates among the participants. On a scale of 1 to 30, the highest game level achieved was level 1 among 29.2% (38/158) of participants and level 15 or above among 8.2% (13/158) of participants.

Effects of Intervention Participation

Table 2 presents the effects of participating in the Safe City game-based intervention on safety and psychosocial outcomes for ITT (n=340), PP (n=328), and the completer (n=325) analyses. At T1, both groups scored relatively high on safety domains (the average score was above 78 when the maximum score was 100) and relatively low on psychosocial problem domains (the average score was below 6 when the maximum score was 20). Overall, results showed that there were no meaningful variations between the groups in terms of mean change at T2 or T3, regardless of the method of analysis used. Additionally, the group × time interaction did not yield statistically significant results for any of the outcome measures.

XSL•FO
RenderX

Table 2. Baseline (T1) values and improvement of safety and psychosocial outcomes at 1 month (T2) and 3 months postintervention (T3).

Characteristic	Control (n=170)	ITT[a] analysis		PP[b] analysis		Completer analysis		
		Intervention (n=170)	*P* value	Intervention (n=158)	*P* value	Control (n=158)	Intervention (n=155)	*P* value
Safe knowledge, mean (SD)								
Baseline score[c]	78.11 (6.11)	78.11 (6.89)	N/A[d]	78.43 (6.65)	N/A	78.48 (6.04)	78.16 (6.83)	N/A
T1-T2[e]	0.12 (6.36)	–0.21 (6.17)	.62	–0.08 (6.14)	.77	0.13 (6.60)	–0.23 (6.47)	.62
T1-T3[e]	0.08 (6.13)	0.57 (6.46)	.47	0.67 (6.50)	.40	0.09 (6.36)	0.63 (6.76)	.47
Safe behavior, mean (SD)								
Baseline score[c]	79.72 (10.23)	78.60 (11.00)	N/A	79.25 (10.72)	N/A	79.94 (10.26)	78.80 (10.90)	N/A
T1-T2[e]	0.75 (10.12)	0.95 (9.59)	.85	0.98 (9.52)	.83	0.80 (10.49)	1.04 (10.04)	.84
T1-T3[e]	0.80 (10.27)	1.43 (10.20)	.57	1.19 (9.87)	.73	0.86 (10.65)	1.57 (10.67)	.56
Internalizing problems[f], mean (SD)								
Baseline score[c]	4.80 (3.08)	4.72 (3.28)	N/A	4.64 (3.27)	N/A	4.80 (3.09)	4.75 (3.36)	N/A
T1-T2[g]	0.15 (2.71)	0.27 (2.98)	.69	0.23 (2.97)	.80	0.16 (2.82)	0.30 (3.12)	.68
T1-T3[g]	0.17 (3.00)	–0.03 (2.79)	.54	–0.01 (2.83)	.58	0.18 (3.12)	–0.03 (2.92)	.54
Externalizing problems[f], mean (SD)								
Baseline score[c]	5.95 (3.59)	5.79 (3.78)	N/A	5.65 (3.75)	N/A	5.85 (3.50)	5.89 (3.83)	N/A
T1-T2[g]	0.07 (2.71)	–0.11 (2.53)	.55	–0.08 (2.56)	.63	0.07 (2.81)	–0.12 (2.65)	.55
T1-T3[g]	–0.22 (3.03)	–0.36 (2.96)	.68	–0.29 (2.96)	.84	–0.24 (3.14)	–0.40 (3.10)	.66

[a]ITT: intention-to-treat.

[b]PP: per-protocol.

[c]Mean domain score at baseline for participants in each study arm.

[d]N/A: not applicable.

[e]Positive values indicate a desirable change in scores.

[f]Missing data from 1 participant.

[g]Negative values indicate a desirable change in scores.

Effects of Intervention Dosage

Table 3 presents the results of GAMs of safety and psychosocial outcome changes at T2 and T3 as a function of the Safe City game use. We found a significant and linear relationship between cumulative game scores and improvements in safe behaviors at T3 (*P*=.03), with higher game scores predicting greater improvements in safe behaviors (Figure S6B in Multimedia Appendix 1). The linear association can also be seen between cumulative game scores and improvements in internalizing problems at T3 (*P*=.01). The level of improvements in internalizing problems at T3 increased monotonically with an increase in game scores during the intervention period (Figure S6C in Multimedia Appendix 1). Regarding mini-games, we observed a significant and monotonic increase in safety knowledge improvements at T2 and T3 as the number of correct answers in Spot the Danger (Figure S8A in Multimedia Appendix 1) and Matching mini-games (Figure S9A in Multimedia Appendix 1) increased. Curvilinear associations were found between MCQ performance and improvements in safety knowledge at T2 (*P*=.008; Figure S7A in Multimedia Appendix 1). At T3, the relationship between MCQ performance and improvements in safety knowledge became linear and marginally significant (*P*=.053; Figure S7A in Multimedia Appendix 1). As the number of correct answers in MCQs increased, the improvements in internalizing problems at T3 became more apparent (*P*=.005; Figure S7C in Multimedia Appendix 1). For the Matching mini-games, the improvements in internalizing problems at T3 reached a plateau when more correct answers were achieved (*P*=.06; Figure S4C in Multimedia Appendix 1). No relationship was found between these game use parameters and changes in externalizing problems at T2 and T3 following the intervention.

XSL•FO
RenderX

Table 3. Effect of game use on improvement in safety and psychosocial outcomes at the 1 month (T2) and 3 months postintervention (T3).

Outcome	Cumulative game score		Multiple choice questions		Spot the Danger		Matching	
	edf[a]	*P* value[b]	edf	*P* value[b]	edf	*P* value[a]	edf	*P* value[a]
Safe knowledge								
Improvement, T2	4.27	.35	5.11	.008	1	.008	1	.002
Improvement, T3	1	.27	1	.05	1	.02	1	.03
Safe behavior								
Improvement, T2	1	.07	1	.67	1	.58	1.12	.45
Improvement, T3	1	.03	1	.48	1.41	.68	1	.53
Internalizing problems								
Improvement, T2	3.07	.09	1	.09	1.17	.27	1	.12
Improvement, T3	1	.01	1.12	.005	1	.20	1.66	.06
Externalizing problems								
Improvement, T2	1	.72	1	>.99	1	.85	1	.42
Improvement, T3	1	.51	1.17	.70	1	.83	1	.64

[a]edf: effective degrees of freedom.

[b]Adjusted for gender, age, baseline level, family socioeconomic status, random assignment, and school cluster.

Discussion

Principal Results

This RCT aimed to investigate the effectiveness of a mobile game–based safety education program in improving children's safety and psychosocial outcomes. The results showed that participating in a 4-week Safe City game-based intervention did not lead to significant improvements in safety and psychosocial well-being at T2 and T3. However, we found evidence of significant effects of game use on changes in safety and psychosocial outcomes after the intervention. The notable enhancement in safe behavior and decrease in internalizing problems at T3 with increasing game scores (intervention dosage), indicates that the game is an effective and enjoyable tool for learning about safety. This, in turn, contributes to the improvement of safety and psychosocial well-being among the participants. Indeed, the analysis of use showed significant effects of the question types used in mini-games on the postintervention changes in safety knowledge and internalizing problems. Specifically, the Matching and Spot the Danger mini-games proved to be effective in promoting safety knowledge. On the other hand, the MCQ mini-game was found to be effective in reducing internalizing problems, although its effect on knowledge gains was not as strong as the other 2 types of mini-game.

Comparison to Prior Work

In this study, it was observed that higher game scores during the intervention period were linked to a greater improvement in safe behavior at T3. The widely used Knowledge, Attitude, and Practices model in health research suggests that repeated acquisition of proper health-related knowledge leads to the formation and maintenance of healthy behavior [30]. While cumulative game scores did not have an impact on the improvement of safety knowledge postintervention, performance in educational mini-games was associated with changes in safety knowledge after the intervention. This finding may be attributed to the fact that the Safe City mobile game includes both educational mini-games and leisure mini-games, all of which award game points for successful attempts. While overall game scores reflect the level of engagement with the intervention, the educational mini-games specifically aim to directly modify or reinforce existing knowledge. By improving safety knowledge, individuals are more likely to adopt and maintain safe behavior, leading to an overall improvement in their safety practices at T3.

Although the results of the use-outcome analysis provide support for the efficacy of the Safe City mobile game as an educational tool to enhance safety and psychosocial well-being in children, it is important to note that the observed differences in mean changes between intervention and control participants were not statistically significant. Further, one possible explanation for this finding is that the participants, regardless of their assignment to the intervention or control group, already had a considerable amount of safety knowledge and compliance, as well as a low level of psychosocial problems, at the beginning of this study. The preexisting level of positive attributes may have limited their potential for improvement through the intervention. A previous study has suggested that demonstrating the effectiveness of an intervention can be challenging if participants' scores on outcome assessments at the start of the study (ie, ceiling effects) are already within the desirable range [31]. Given the promising results regarding the use of the game, it is likely that the Safe City mobile game–based intervention will yield more favorable participation outcomes in scenarios where most participants have insufficient safety awareness and behaviors, or among younger participants who have not yet acquired adequate safety knowledge.

Another encouraging finding is that the intervention showed no effect on externalizing problems, such as conduct problems and

XSL•FO
RenderX

symptoms of hyperactivity or inattention, regardless of participation in the intervention or the dosage received. On the other hand, we noticed that the positive effects of game use on internalizing problems were delayed. Specifically, we found that higher cumulative game scores were only associated with greater reductions in internalizing problems at T3. These findings are consistent with emerging research showing that playing video games may not be entirely detrimental to children's mental health [32]. Previous reviews have highlighted the potential benefits of video games in promoting student engagement and motivation for learning. When used appropriately, video games can even serve as a powerful tool for enhancing cognition in children with mental disorders [33]. Furthermore, studies have demonstrated that various types of video games, including roleplaying games and even short gameplay sessions as brief as 1 or 5 minutes, can help reduce stress and anxiety [34]. Nevertheless, it is important to note that children with self-regulation difficulties, such as attention-deficit/hyperactivity disorder, may be more prone to problematic gaming behavior [35]. Hence, it is crucial for future research to explore the potential benefits of mobile game–based education interventions for both typically developing children and those with special education needs. Understanding how these interventions can effectively support the learning and development of all children, regardless of their individual circumstances, is of utmost importance.

Another notable finding is the differential effects of the types of questions used in mini-games on the postintervention changes in safety and psychosocial outcomes. In this study, the Matching and Spot the Danger performance showed a linear relationship with the postintervention increase in safety knowledge. On the other hand, the MCQ performance exhibited a nonlinear relationship with the improvements in safety knowledge and internalizing problems after the intervention. This distinction could be attributed to the higher involvement of guessing in MCQs compared to the other 2 types of questions. Previous studies have shown that matching items are better at differentiating well-prepared students from marginal students and are less susceptible to cueing issues compared to MCQs [36]. Therefore, it is plausible to speculate that participants who had a limited number of correct answers in MCQs may have relied more on guessing. This could have led to a wider range of knowledge improvement among participants at T2, which subsequently declined at T3.

Limitations and Strengths

One of the major strengths of this study was collecting game use data as a proxy for intervention dosage. However, some limitations need to be taken into consideration when interpreting the findings. Firstly, 93% (158/170 participants) engaged in playing the Safe City mini-games at least once, but there was a significant variation in use rates, with game scores ranging from 4000 to 14,598,100. This variability poses challenges in obtaining an unbiased estimation of the intervention's effects. Secondly, the sample had a minimal number of participants with insufficient safety knowledge or inadequate safety practices. Due to the limited statistical power, we were unable to investigate whether this specific subgroup would show more positive responses to the Safe City mobile game–based intervention. Lastly, the outcome measures relied on self-reports provided by the participants. This introduces the potential for social desirability bias, as individuals may provide responses that align with societal expectations rather than their true experiences or behaviors.

Conclusions

This study contributes to the growing body of research suggesting that serious games, especially when engaged with at high levels, can effectively yield immediate and long-term improvements in knowledge and psychosocial outcomes among typically developing children. Additionally, the findings support the investigation of gamification as a viable approach to health education. Furthermore, this study underscores the importance of integrating traditional and digital teaching methods in education to maximize the benefits of digital tools. It may be necessary to develop question items in various styles, such as matching or MCQs, to attain different outcome objectives, which would require further evaluation.

Acknowledgments

We thank the schools, children, and parents for participating in this study. This study was supported by the Health and Medical Research Fund, Health Bureau, the Hong Kong Special Administrative Region, China (02180768).

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Authors' Contributions

RSW conceptualized the paper, participated in game design, conducted data analysis, and wrote the initial draft. KTST participated in game design and data collection and critically revised this paper. FKWH commented on data analysis and critically revised this paper. WHSW, CBC, KLC, and KWF contributed to data interpretation and critically revised this paper. PI contributed to the overall study design and data interpretation and critically revised this paper. All authors have approved the final paper as submitted.

XSL•FO
RenderX

Conflicts of Interest

RSW is an associate editor of the *Journal of Medical Internet Research*. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Additional figures and tables.
[PDF File (Adobe PDF File), 1158 KB - jmir_v26i1e51908_app1.pdf]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1183 KB - jmir_v26i1e51908_app2.pdf]

References

1. GBD 2019 Adolescent Transport and Unintentional Injuries Collaborators. Adolescent transport and unintentional injuries: a systematic analysis using the Global Burden of Disease Study 2019. Lancet Public Health 2022;7(8):e657-e669 [FREE Full text] [doi: 10.1016/S2468-2667(22)00134-7] [Medline: 35779567]
2. Morrongiello B. Preventing unintentional injuries to young children in the home: understanding and influencing parents' safety practices. Child Dev Perspectives 2018;12(4):217-222 [FREE Full text] [doi: 10.1111/cdep.12287]
3. Bernacki ML, Greene JA, Crompton H. Mobile technology, learning, and achievement: advances in understanding and measuring the role of mobile technology in education. Contemp Educ Psychol 2020;60:101827 [FREE Full text] [doi: 10.1016/j.cedpsych.2019.101827]
4. Chen HL, Wu CT. A digital role-playing game for learning: effects on critical thinking and motivation. Interact Learn Environ 2023;31(5):3018-3030. [doi: 10.1080/10494820.2021.1916765]
5. Behnamnia N, Kamsin A, Ismail MAB, Hayati A. The effective components of creativity in digital game-based learning among young children: a case study. Child Youth Serv Rev 2020;116:105227 [FREE Full text] [doi: 10.1016/j.childyouth.2020.105227]
6. Iten N, Petko D. Learning with serious games: is fun playing the game a predictor of learning success? Brit J Educational Tech 2016;47(1):151-163 [FREE Full text] [doi: 10.1111/bjet.12226]
7. Bossen D, Broekema A, Visser B, Brons A, Timmerman A, van Etten-Jamaludin F, et al. Effectiveness of serious games to increase physical activity in children with a chronic disease: systematic review with meta-analysis. J Med Internet Res 2020;22(4):e14549 [FREE Full text] [doi: 10.2196/14549] [Medline: 32234697]
8. Sajeev M, Kelada L, Yahya Nur AB, Wakefield CE, Wewege MA, Karpelowsky J, et al. Interactive video games to reduce paediatric procedural pain and anxiety: a systematic review and meta-analysis. Br J Anaesth 2021;127(4):608-619 [FREE Full text] [doi: 10.1016/j.bja.2021.06.039] [Medline: 34340838]
9. Dixon CA, Ammerman RT, Dexheimer JW, Meyer B, Jung H, Johnson BL, et al. Development of iBsafe: a collaborative, theory-based approach to creating a mobile game application for child safety. AMIA Annu Symp Proc 2014;2014:477-485 [FREE Full text] [Medline: 25954352]
10. Dixon CA, Ammerman RT, Johnson BL, Lampe C, Hart KW, Lindsell CJ, et al. A randomized controlled field trial of iBsafe-a novel child safety game app. mHealth 2019;5:3 [FREE Full text] [doi: 10.21037/mhealth.2019.01.02] [Medline: 30842951]
11. Vergunst F, Commisso M, Geoffroy MC, Temcheff C, Poirier M, Park J, et al. Association of childhood externalizing, internalizing, and comorbid symptoms with long-term economic and social outcomes. JAMA Netw Open 2023;6(1):e2249568 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.49568] [Medline: 36622675]
12. Eirich R, McArthur BA, Anhorn C, McGuinness C, Christakis DA, Madigan S. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. JAMA Psychiatry 2022;79(5):393-405 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.0155] [Medline: 35293954]
13. Abd-Alrazaq A, Al-Jafar E, Alajlani M, Toro C, Alhuwail D, Ahmed A, et al. The effectiveness of serious games for alleviating depression: systematic review and meta-analysis. JMIR Serious Games 2022;10(1):e32331 [FREE Full text] [doi: 10.2196/32331] [Medline: 35029530]
14. Abd-Alrazaq A, Alajlani M, Alhuwail D, Schneider J, Akhu-Zaheya L, Ahmed A, et al. The effectiveness of serious games in alleviating anxiety: systematic review and meta-analysis. JMIR Serious Games 2022;10(1):e29137 [FREE Full text] [doi: 10.2196/29137] [Medline: 35156932]
15. Reynard S, Dias J, Mitic M, Schrank B, Woodcock KA. Digital interventions for emotion regulation in children and early adolescents: systematic review and meta-analysis. JMIR Serious Games 2022;10(3):e31456 [FREE Full text] [doi: 10.2196/31456] [Medline: 35984681]
16. Kowal M, Conroy E, Ramsbottom N, Smithies T, Toth A, Campbell M. Gaming your mental health: a narrative review on mitigating symptoms of depression and anxiety using commercial video games. JMIR Serious Games 2021;9(2):e26575 [FREE Full text] [doi: 10.2196/26575] [Medline: 34132648]

XSL•FO
RenderX

17. David OA, Costescu C, Cardos R, Mogoaşe C. How effective are serious games for promoting mental health and health behavioral change in children and adolescents? A systematic review and meta-analysis. Child Youth Care Forum 2020;49(6):817-838. [doi: 10.1007/s10566-020-09566-1]
18. Chow CB. Childhood injuries in Hong Kong: from epidemiology to community intervention. The University of Hong Kong. 2015. URL: https://hub.hku.hk/handle/10722/230619 [accessed 2024-01-30]
19. Chan CC, Cheng JC, Wong TW, Chow CB, Luis BP, Cheung WL, et al. An international comparison of childhood injuries in Hong Kong. Inj Prev 2000;6(1):20-23 [FREE Full text] [doi: 10.1136/ip.6.1.20] [Medline: 10728536]
20. Injury survey. Department of Health, Hong Kong. 2008. URL: https://www.chp.gov.hk/files/pdf/injury_survey_eng.pdf [accessed 2024-01-30]
21. Wong RS, Tung KT, Wong HT, Ho FK, Wong HS, Fu KW, et al. A mobile game (Safe City) designed to promote children's safety knowledge and behaviors: protocol for a randomized controlled trial. JMIR Res Protoc 2020;9(6):e17756 [FREE Full text] [doi: 10.2196/17756] [Medline: 32530436]
22. Dunwell I, de Freitas S, Petridis P, Hendrix M, Arnab S, Lameras P, et al. A game-based learning approach to road safety: the code of everand. 2014 Presented at: SIGCHI Conference on Human Factors in Computing Systems; April 26-May 1, 2014; Toronto, ON p. 3389-3398. [doi: 10.1145/2556288.2557281]
23. Wong RSM, Yu EYT, Wong TWL, Fung CSC, Choi CSY, Or CKL, et al. Development and pilot evaluation of a mobile app on parent-child exercises to improve physical activity and psychosocial outcomes of Hong Kong Chinese children. BMC Public Health 2020;20(1):1544 [FREE Full text] [doi: 10.1186/s12889-020-09655-9] [Medline: 33054753]
24. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
25. Kiili K. Digital game-based learning: towards an experiential gaming model. Internet High Educ 2005;8(1):13-24. [doi: 10.1016/j.iheduc.2004.12.001]
26. Home Safety. Safe Kids Worldwide. 2023. URL: https://www.safekids.org/safetytips/field_venues/home [accessed 2024-01-30]
27. Wang KM, Lin YH, Huang YC. The knowledge and attitude of sports injury prevention and management of senior high school athletes in Taiwan. Int J Sport Health Sci 2012;10:12-22. [doi: 10.5432/ijshs.201122]
28. Lai KYC, Luk ESL, Leung PWL, Wong ASY, Law L, Ho K. Validation of the Chinese version of the strengths and difficulties questionnaire in Hong Kong. Soc Psychiatry Psychiatr Epidemiol 2010;45(12):1179-1186. [doi: 10.1007/s00127-009-0152-z] [Medline: 19820885]
29. Muris P, Meesters C, Eijkelenboom A, Vincken M. The self-report version of the Strengths and Difficulties Questionnaire: its psychometric properties in 8- to 13-year-old non-clinical children. Br J Clin Psychol 2004;43(Pt 4):437-448. [doi: 10.1348/0144665042388982] [Medline: 15530213]
30. Xu N, Zhang Y, Zhang X, Zhang G, Guo Z, Zhao N, et al. Knowledge, attitudes, and practices of urban residents toward COVID-19 in Shaanxi during the post-lockdown period. Front Public Health 2021;9:659797 [FREE Full text] [doi: 10.3389/fpubh.2021.659797] [Medline: 34095063]
31. Jaarsma T, Strömberg A, Dunbar SB, Fitzsimons D, Lee C, Middleton S, et al. Self-care research: how to grow the evidence base? Int J Nurs Stud 2020;105:103555 [FREE Full text] [doi: 10.1016/j.ijnurstu.2020.103555] [Medline: 32199150]
32. Stenseng F, Hygen BW, Wichstrøm L. Time spent gaming and psychiatric symptoms in childhood: cross-sectional associations and longitudinal effects. Eur Child Adolesc Psychiatry 2020;29(6):839-847. [doi: 10.1007/s00787-019-01398-2] [Medline: 31492978]
33. Peñuelas-Calvo I, Jiang-Lin LK, Girela-Serrano B, Delgado-Gomez D, Navarro-Jimenez R, Baca-Garcia E, et al. Video games for the assessment and treatment of attention-deficit/hyperactivity disorder: a systematic review. Eur Child Adolesc Psychiatry 2022;31(1):5-20. [doi: 10.1007/s00787-020-01557-w] [Medline: 32424511]
34. Pallavicini F, Pepe A, Mantovani F. Commercial off-the-shelf video games for reducing stress and anxiety: systematic review. JMIR Ment Health 2021;8(8):e28150 [FREE Full text] [doi: 10.2196/28150] [Medline: 34398795]
35. Wang BQ, Yao NQ, Zhou X, Liu J, Lv ZT. The association between attention deficit/hyperactivity disorder and internet addiction: a systematic review and meta-analysis. BMC Psychiatry 2017;17(1):260 [FREE Full text] [doi: 10.1186/s12888-017-1408-x] [Medline: 28724403]
36. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. J Family Community Med 2006;13(3):125-133 [FREE Full text] [Medline: 23012132]

Abbreviations

edf: effective degrees of freedom
GAM: generalized additive model
ITT: intention-to-treat
MCQ: multiple choice question
PP: per-protocol
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 16.08.23; peer-reviewed by T Baranowski, Y Chen; comments to author 20.12.23; revised version received 28.12.23; accepted 22.01.24; published 14.02.24.

Please cite as:
Wong RS, Tung KTS, Ho FKW, Wong WHS, Chow CB, Chan KL, Fu KW, Ip P
Effect of a Mobile Game–Based Intervention to Enhance Child Safety: Randomized Controlled Trial
J Med Internet Res 2024;26:e51908
URL: https://www.jmir.org/2024/1/e51908
doi:10.2196/51908
PMID:38354042

©Rosa S Wong, Keith T S Tung, Frederick K W Ho, Wilfred H S Wong, Chun Bong Chow, Ko Ling Chan, King Wa Fu, Patrick Ip. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial

Oh Young Kwon[1,2], PhD; Mi Kyung Lee[3], PhD; Hye Won Lee[4,5], MD, PhD; Hyerang Kim[6], PhD; Jae Seung Lee[4,5], MD, PhD; Yeonsoo Jang[2], PhD

[1]College of Nursing, Brain Korea 21 FOUR Project, Yonsei University, Seoul, Republic of Korea
[2]College of Nursing, Mo-Im Kim Nursing Research Institute, Yonsei University, Seoul, Republic of Korea
[3]Frontier Research Institute of Convergence Sports Science, Yonsei University, Seoul, Republic of Korea
[4]Department of Internal Medicine, College of Medicine, Yonsei University, Seoul, Republic of Korea
[5]Yonsei Liver Center, Severance Hospital, Seoul, Republic of Korea
[6]Department of Nursing Science, VISION College of Jeonju, Jeollabuk-Do, Republic of Korea

Corresponding Author:
Yeonsoo Jang, PhD
College of Nursing, Mo-Im Kim Nursing Research Institute
Yonsei University
50-1 Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 3343
Email: ysjang517@yuhs.ac

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e57499

Abstract

Background: Lifestyle modification in patients with nonalcoholic fatty liver disease (NAFLD) is key to improving health outcomes. Mobile health technologies may offer potential effective and efficient health care support to facilitate self-management.

Objective: This study aims to develop a lifestyle coaching intervention using a mobile app for patients with NAFLD and evaluate physiological and psychological health outcomes for 6 months.

Methods: This study was a randomized controlled trial. The personalized lifestyle coaching intervention using a mobile app was developed through established guidelines and literature reviews. This intervention consisted of information on NAFLD management, diet and physical activity self-monitoring, and coaching sessions based on patient records and SMS text messages. A total of 102 individuals were enrolled in the study and randomly assigned to the intervention group (n=48) or the control group (n=54). The outcomes were improvements in physiological (weight, liver fat score, aspartate aminotransferase, alanine transferase, and gamma-glutamyl transferase) and clinical outcomes (self-management, NAFLD self-management knowledge, self-efficacy, fatigue, depression, and quality of life). Data were analyzed using descriptive analysis and a linear mixed model to test the effects of the intervention.

Results: All participants completed the study. The mean age of the participants was 48.9 (SD 13.74) years, 38.2% (39/102) were female participants, and 65.7% (67/102) were married. There were no differences in baseline demographic and clinical data between the intervention and control groups. Changes from baseline to 6 months were significant only within the intervention group for weight (P<.001), liver fat score (P=.01), aspartate aminotransferase (P=.03), alanine transferase (P=.002), gamma-glutamyl transferase (P=.04), self-management (P<.001), fatigue (P=.005), depression (P=.003), and quality of life (P<.001). The differences between the 2 groups for the changes over the 6 months were significant in self-management (P=.004), self-management knowledge

XSL•FO
RenderX

(*P*=.04), fatigue (*P*=.004), depression (*P*=.04), and quality of life (*P*=.01). However, the intervention-by-time interaction was significantly effective only for self-management (*P*=.006) and fatigue (*P*=.02).

Conclusions: Nonpharmacological interventions using a mobile app may be effective in improving the physiological and psychological health outcomes of patients with NAFLD.

Trial Registration: Clinical Research Information Service KCT0005549; http://tinyurl.com/y2zb6usy

(J Med Internet Res 2024;26:e49839) doi:10.2196/49839

KEYWORDS

lifestyle modification; mobile health; nonalcoholic fatty liver disease; self-management; randomized controlled trial

Introduction

Nonalcoholic fatty liver disease (NAFLD) is the most common liver disease globally, and its incidence continues to increase rapidly. Recently, the prevalence of NAFLD worldwide was estimated to be around 32.4%, and its reported overall incidence was estimated to be 46.9 cases per 1000 persons every year [1]. Its prevalence in Asian countries ranges from 5% to 30% and is reported to have increased by 30.3% over the last 10 years in South Korea [2].

Most patients with NAFLD are asymptomatic, but more than half may develop nonalcoholic steatohepatitis, which can progress to liver fibrosis, cirrhosis, and liver cancer [1,3]. Moreover, early NAFLD is often associated with metabolic syndromes, obesity, and diabetes [4]. The progressive nature of NAFLD and the association of these factors may put a strain on health care systems in the future if not addressed in advance. Therefore, the early diagnosis and management of NAFLD have the potential to improve the health outcomes of the patients.

Lifestyle modification is key to improving the histologic condition of NAFLD following weight loss. Clinical practice guidelines recommend a weight loss of at least 5% in patients with NAFLD who are overweight or obese [5]. Previous studies have reported the effects of regular exercise practice, associated with various types of diets or not, on weight loss [6-8]. However, some factors related to long-term weight loss maintenance and motivation for weight loss remain unclear in clinical practice [9-11].

Mobile health technology has contributed to improving people's health conditions. Apps on mobile phones can deliver health care information beyond clinical settings and enable tailored interventions to facilitate healthy lifestyles [12,13]. Furthermore, because of the possibility of real-time interactions between people without restrictions on time and place, such applications are effective in helping patients achieve specific health goals through continuous and comprehensive information delivery through SMS text messages, informative texts, illustrations, and instructional videos [14,15]. Given these benefits, mobile apps have been used to manage various diseases daily and are expected to help patients overcome obstacles in the management of NAFLD. Lifestyle modification interventions using mobile apps are increasingly being implemented for patients with NAFLD. A recent meta-analysis reported that interventions using eHealth technologies were effective for BMI, aspartate aminotransferase (AST), and alanine transferase (ALT) in these patients [16]. However, previous studies have focused on physiological outcomes related to weight, BMI, and clinical findings rather than psychological aspects. Some studies have found associations between NAFLD (with or without fibrosis) and mental health status, such as depressive symptoms, anxiety, and general quality of life decline [17-19]. However, studies that focus specifically on the psychological and physical health aspects of patients with NAFLD are lacking.

This study aimed to evaluate the effects of a lifestyle coaching intervention on self-management behaviors and the treatment of patients with NAFLD. We developed a long-term lifestyle coaching intervention routine through a mobile app aiming to improve the health of patients with NAFLD. We hypothesized that the use of this app for 6 months would be associated with improvement of NAFLD self-management compared to traditional follow-up care and would be effective in improving physiological and psychological health parameters.

Methods

Study Design

A randomized controlled trial was conducted using parallel treatment groups for 6 months. Participants were randomly allocated to either the self-management app for repeat tutoring (SMART-Liver, intervention group) or the standard care group (control group) in a 1:1 ratio.

Study Participants and Setting

The participants were recruited from the outpatient department of gastroenterology at a tertiary university hospital in Seoul, South Korea. In total, 138 patients with NAFLD were assessed for eligibility and participated in a study between November 2020 and January 2022. Inclusion criteria were patients who were aged 19 years or older, diagnosed with NAFLD without secondary causes for liver fat accumulation, able to read and write in English, and owned a smartphone using an Android system. The exclusion criteria were a history of alcohol intake of more than 20 g per day for women and 30 g per day for men based on the guidelines of the Korean Association for the Study of the Liver [20], those with hepatitis B or C virus infections, pregnant women who received hepatotoxic medication, patients who had cirrhosis, those with end-stage disease, those with concomitant liver disease, those diagnosed with psychiatric conditions, and those with clinically or biochemically recognized systemic diseases. The patients were screened for eligibility by clinicians and researchers.

XSL•FO
RenderX

Sample Size

Power analysis was performed using G*Power software (version 3.1; Heinrich Heine University) to determine the sample size. The sample size for each group was calculated to be at least 64 participants to obtain a 0.5 effect size, α=.05, and 80% statistical power. Accounting for a 10% dropout rate, the total sample size was planned to be 140 participants.

Randomization and Allocation

The recruited study participants were patients whose physicians recommended screening for eligibility when they visited the clinic. Participants who agreed to participate in the study and provided written informed consent were randomly assigned to either the control or intervention group in a 1:1 ratio. The allocation sequence was concealed using lots marked with the letter I (intervention group) or C (control group) in sealed, impenetrable envelopes prepared by a research assistant who was not involved in the intervention. The lots were stored in cabinets with locks before and after recruitment.

Participants in the intervention group received a unique code number for logging in to the SMART-Liver app after learning how to use the app and registering it on the spot. We provided a written manual for the SMART-Liver app to each participant in the intervention group. The research assistant surveyed participants in the intervention and control group using a questionnaire to obtain their general characteristics, self-management levels, and other psychosocial behaviors. The study period of 24 weeks began after a week in both groups. Participants in the control group received routine care based on treatment protocol in the hospital.

Intervention

Development of SMART-Liver Program

The SMART-Liver (self-management app for repeat tutoring) program is a lifestyle coaching intervention designed to improve the self-management of patients with NAFLD. The intervention was developed following the process of the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) model [21]. The needs of patients with NAFLD and existing studies using mobile apps for self-management behaviors were analyzed in the analysis step, which included literature reviews and the guidelines for NAFLD treatment. Aiming to improve the level of self-management of patients with NAFLD, the intervention consisted of modules for self-monitoring, information, coaching, and SMS text messages related to diet and physical activities. An app with artificial intelligence technology was selected to operate the functions of self-monitoring of diet, which enabled participants to log into the app their diet at every meal. The app was created by modifying the Diet Camera app developed by Doing Lab in South Korea and is commercially available in app stores with basic features that are accessible for free. Full features of the app were made available to participants for the SMART-Liver intervention in collaboration with Doing Lab during the intervention period. A representation of the Smart-Liver app interface is shown in Figure 1.

The information provided in the app include (1) educational slides for NAFLD treatment based on guidelines and diet tips, (2) diet news for 1 portion size of food, (3) exercise video clips, and (4) summary quizzes to facilitate self-management behaviors.

Self-monitoring instructions include (1) diets for every meal and snack; (2) checking of the total caloric intake based on their records; and (3) lifestyle health habits such as exercise, sleep, alcohol intake, and smoking.

The participants were coached to (1) meet with the research team and physical activity specialist to set personal weight and physical activity goals based on their weight and diet information logged in the first week after registering in the app, (2) identify barriers and facilitators in their performances of health-related behavioral change, (3) evaluate or reset the goal in the 12th week, and (4) provide feedback through reports that displayed analyses of their diet logs every 4 weeks for 6 months.

The participants received (1) SMS text messages automatically every morning at 8 AM to remind participants about daily missions, including diet logs; (2) SMS text messages when the missions were completed to encourage the participants' performances; (3) tailored feedback messages for diet and physical activity records once a week; and (4) real-time messages using a chat channel when they have any queries about their disease or obstacles encountered.

XSL•FO
RenderX

Figure 1. SMART-Liver app.

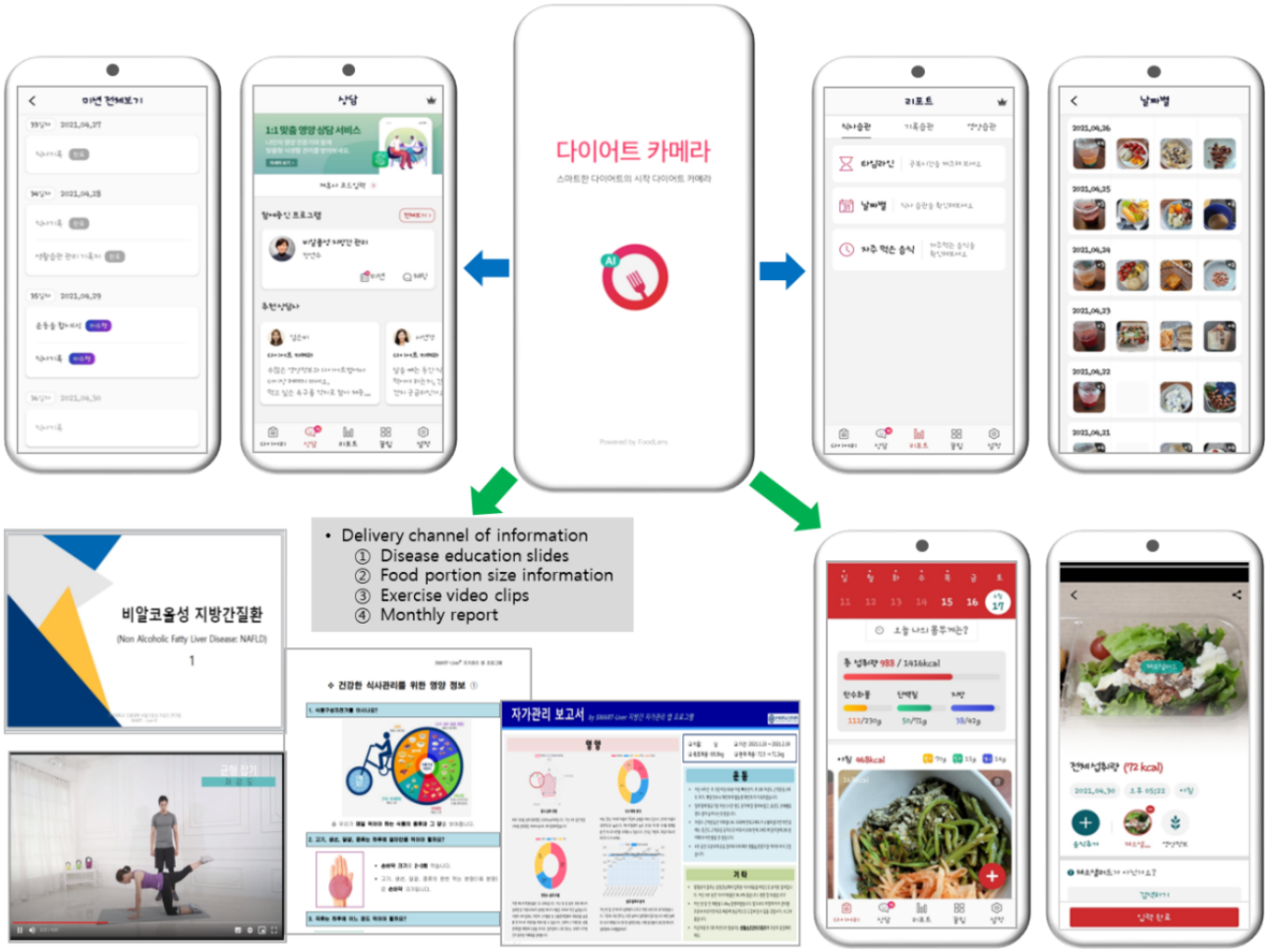

SMART-Liver Intervention

The participants input their body weight and height in the app and received information on the proper amount of calories for their weight through a chat channel every week in the first month. During the rest of the intervention period, the researchers continuously reminded the participants to incorporate this information into their daily routines. From the day of registration onward, participants logged their diets, such as the type of food and the amount they ate, and answered questions about lifestyle management related to exercise, smoking habits, alcohol consumption, and sleep patterns. Participants could check or edit their logs in the app. After using the app for 1 week to practice, they set a weight loss goal based on their records of diet, exercise, and lifestyle logged into the app, with the assistance of a dietitian and a physical activity specialist via remote meeting. After this meeting, the actual intervention program was initiated. Daily missions such as diet logs or weight measurements were sent every morning at 8 AM. During the first month, individually catered educational exercise video clips and information materials on liver disease and diet were sent weekly. After the first month, participants continued to have access to these educational materials on the app at any time.

In addition, the research team sent feedback messages based on the performance rate of diet records every week to encourage compliance and provided monthly feedback reports that included comments from dietitians and physical activity specialists via the app's chat channel. If there were any questions or obstacles to participating in the intervention, participants could ask the research team using a chat channel, and the research team responded in real time during working hours announced in advance. In the third month, participants reset their goals for body weight or lifestyle health habits for the remaining 3-month period through counseling with a dietitian and physical activity specialist.

In the control group, participants received phone calls from the research team a week after enrollment. The research team took note of the participants' body weights and provided information on the importance of weight loss, diet, and exercise for NAFLD management via phone calls. The participants received the standard care for NAFLD for 6 months and were reached by phone calls in the third and sixth months for weight measurement and survey.

Measurements

Overview

We measured general characteristics, physiological, and clinical outcomes over 6 months. Clinical outcomes were measured at baseline and 3 and 6 months. Clinical data were obtained at baseline and after 6 months.

XSL•FO
RenderX

General and Clinical Characteristics

Demographic and clinical information including age, sex, marital status, education level, family income, health-related factors, comorbidities, and clinical findings (liver fat score, AST, ALT, and gamma-glutamyl transferase [-GT]) were collected at the beginning of the study via survey.

Self-Management

Self-management was measured using the NAFLD Self-Management Questionnaire developed based on the Individual and Family Self-Management Theory in patients with NAFLD in South Korea [22]. This 22-item self-report questionnaire is scored on a 5-point Likert scale ranging from 1 (never) to 5 (always). It consists of 6 subdomains: medical treatment compliance, management of medications and dietary supplements, alcohol consumption management, sleep management, family support, and lifestyle management. The mean scores were calculated, with a higher score indicating a higher level of self-management. The Cronbach α value was .87 in the original study and .84 in this study.

Self-Efficacy

Self-efficacy was assessed using the 22-item Chronic Disease Self-Efficacy Scale–Korean Version [23], which was modified from the Chronic Disease Self-Efficacy Scale developed by Lorig et al [24]. Responses were provided on a 10-point Likert scale ranging from 1 (not at all confident) to 10 (totally confident), with higher scores indicating higher confidence. In the original study, the Cronbach α of the scale was .91 and .93 in the Korean version. In this study, the Cronbach α was .95.

Fatigue

Fatigue was assessed using the Korean version of the Revised Piper Fatigue Scale [25], which was translated from the Revised Piper Fatigue Scale [26]. This scale consists of 22 items, each rated on a 10-point numeric rating scale ranging from 0 (none) to 10 (maximum symptom intensity). The scores for all items were added, and the total scores were divided by the number of items. This scale provides an estimate of fatigue, with 0 representing none, 1-3 as mild, 4-6 as moderate, and 7-10 as severe fatigue. Higher mean scores indicated higher perceived fatigue levels. The Cronbach α of the original scale was .97, and that of the Korean version was .96. In this study, the Cronbach α was .97.

Depression

Depression was assessed using the Beck Depression Inventory I-Korean version [27], which is composed of 21 items. Each item is rated on a 4-point Likert scale ranging from 0 to 3. The total score ranges from 0 to 63. Individuals with a total score between 0 and 13 were considered to have minimal depressive symptoms, 14-19 mild depressive symptoms, 20-28 moderate depressive symptoms, and 29-63 severe depressive symptoms. The Cronbach α of the scale was reported as .85 in a previous study and was .91 in this study.

Quality of Life

Quality of life was assessed using the World Health Organization Quality of Life–Brief Version [28], a self-reported questionnaire containing 26 items. The responses were rated on a 5-point Likert scale, where 1 indicated high satisfaction and 5 indicated low satisfaction. The mean score of the items was used in the analyses, and higher scores indicated a better quality of life. The Cronbach α of the scale was reported as .90 in the original study and .89 in this study.

Self-Management Knowledge

The participants' self-management knowledge was evaluated using 10 items of essential knowledge in the NAFLD self-management questionnaire developed by the researchers. The items were based on the NAFLD management guidelines, and the participants answered "yes=1" or "no=0." The total scores were added, and a higher score indicated a higher level of NAFLD self-management knowledge.

SMART-Liver App Compliance

Patient compliance was defined as the participants' diet-log record rate in this study. All participants in the intervention group were expected to log their diet information 3 times using the SMART-Liver app, and their compliance with the intervention was evaluated based on diet-log rates during the 6 months. The diet logs were collected from the app and calculated by the researchers. The levels of compliance were analyzed as >70%, 70%-90%, and more than 90%.

Data Analysis

Data analyses were performed using SPSS Statistics for Windows (version 25.0; IBM). A per-protocol analysis was conducted to evaluate the results of the intervention among 102 participants. General characteristic data are described using mean and SD or frequencies. Differences between the intervention and control groups in sociodemographic, physiological, and psychosocial variables at baseline were analyzed using a 2-tailed *t* test. This process allowed us to confirm the homogeneity between the intervention and control groups and ensured that we did not require further adjustments for any imbalance.

We conducted a paired 2-tailed *t* test for continuous variables within each group and an independent 2-tailed *t* test between groups to compare the baseline and 3- and 6-month changes. Subsequently, a linear mixed model was used to evaluate any statistically significant differences over time in repeated measurements between the study groups. This approach assessed the intervention effect by simultaneously adjusting for the correlation between the intervention and control groups and between repeated measurements of the same participants.

Ethical Considerations

This study was approved by the institutional review board of Yonsei University Health System (IRB 4-2020-0955). All participants were provided written informed consent prior to participating. This study only recruited participants who were willing to participate, and they were fully informed of the rights and obligations associated with the intervention. They retained the freedom to withdraw from the study at any time. All study-related and participant data were securely stored in a password-protected file cabinet with restricted access. To ensure the confidentiality of participants, all identifiable data were

XSL•FO
RenderX

de-identified using coding. Participants were compensated for their participation and completion of the study survey on three occasions (Baseline, 3 months, and 6 months). At the 3-month survey, participants received US $15 upon completion of both the baseline and 3-month surveys. Additionally, participants who completed the 6-month intervention were provided with a free transient elastography test for liver fat valued at US $150.

The trial was prospectively registered in the World Health Organization–accredited Clinical Research Information Service on October 29, 2020 (registration number KCT0005549). We performed this trial based on the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Application and online TeleHealth) guidelines (Multimedia Appendix 1).

Results

Characteristics of Participants

In total, 111 participants were enrolled in this study (Figure 2). During the first month, 2 patients withdrew from the study for personal reasons. Five patients in the control group dropped out of the study because they did not respond to the several contact attempts made by the researchers. Two patients in the intervention group dropped out due to poor compliance and reluctance to use the app. Thus, a total of 102 (91.9%) participants were included in the final analysis (48 in the intervention group and 54 in the control group). Overall, the mean age of the participants was 48.9 (SD 13.74) years.

Table 1 shows the demographic data and baseline characteristics of the participants in each group. There were no significant differences between the groups in terms of age, sex, socioeconomic status, and the presence of relevant comorbidities (all $P>.05$). The physiological and psychosocial outcome variables showed no differences between the participants in either group at the beginning of the intervention (Table 2).

Figure 2. Summary of participation flow diagram.

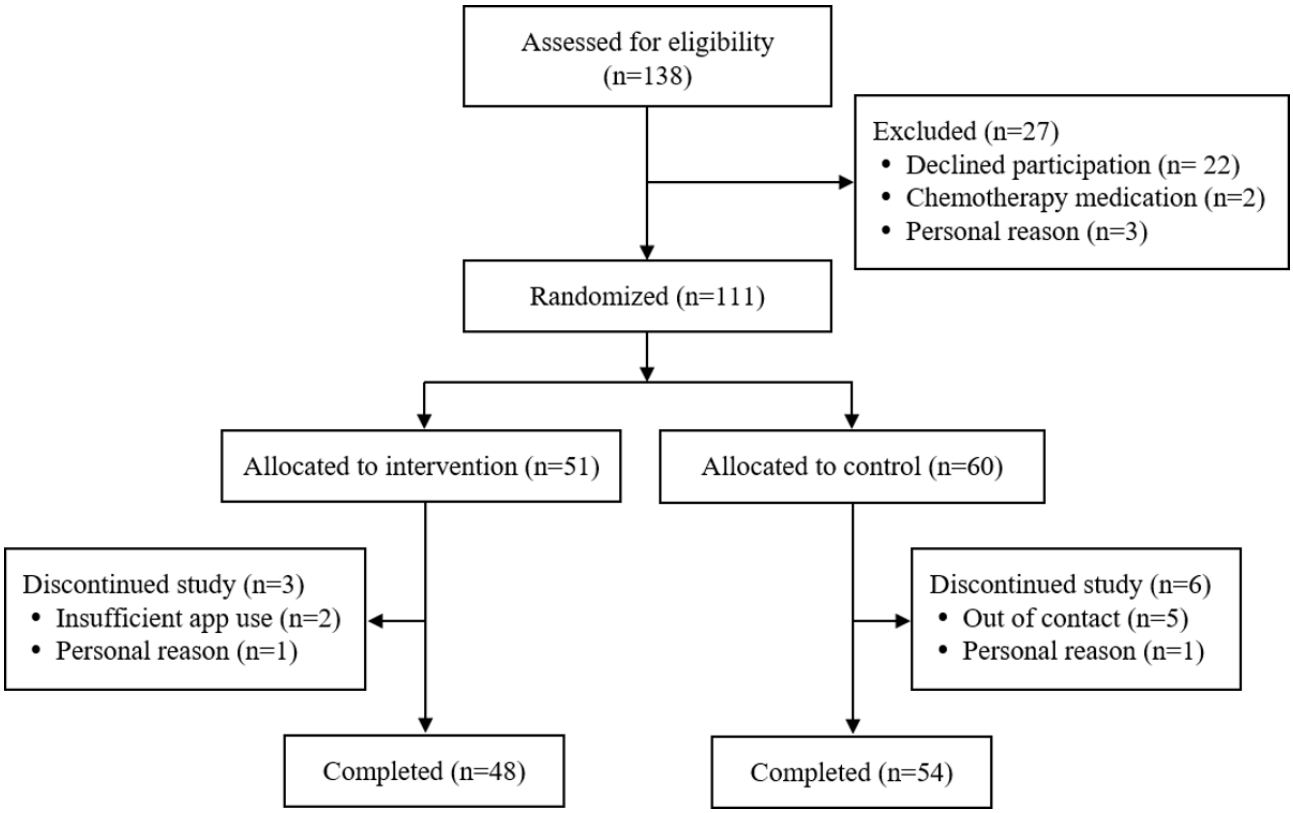

XSL•FO
RenderX

Table 1. Baseline characteristics of the intervention and control groups (N=102).

Characteristics	Control (n=54)	Intervention (n=48)	*P* value
Age (years), mean (SD)	47.1 (13.91)	51.0 (13.37)	.17
Sex (male), n (%)	29 (54)	34 (71)	.10
Occupation (yes), n (%)	41 (76)	36 (75)	≥.99
Marital status, n (%)			.18
Married	36 (67)	31 (65)	
Single	17 (32)	12 (25)	
Widowed, divorce, or separation	1 (2)	5 (10)	
Education, n (%)			.12
Elementary school	0 (0)	1 (2)	
Middle school	0 (0)	0 (0)	
High	22 (41)	11 (23)	
College	23 (28)	29 (60)	
Graduate school or over	9 (17)	7 (15)	
Family income (US $ per month), n (%)			.95
≤1750	7 (13)	6 (13)	
1751-3500	13 (24)	13 (27)	
3501-5250	19 (36)	14 (29)	
≥5251	14 (26)	15 (31)	
Smoking, n (%)			.88
Yes	6 (11)	7 (15)	
No	34 (63)	28 (58)	
Quit	14 (26)	13 (27)	
Drinking (yes), n (%)	20 (37)	15 (31)	.68
Sleep sufficiency (yes), n (%)	26 (48)	20 (42)	.55
Comorbidity[a] (yes), n (%)	41 (76)	40 (83)	.46
Hyperlipidemia	26 (48)	22 (46)	.85
Hypertension	22 (41)	18 (38)	.84
Diabetes	14 (26)	17 (35)	.39
Osteoporosis	1 (2)	1 (2)	≥.99
Others	20 (37)	20 (42)	.69
Medication[a] (yes), n (%)	40 (74)	42 (88)	.13
Hyperlipidemia	25 (46)	23 (48)	≥.99
Hypertension	19 (40)	19 (35)	.69
Diabetes	14 (26)	16 (33)	.52
Others	22 (41)	21 (44)	.84

[a]Respondents are allowed to select more than 1 answer for a single question.

XSL•FO
RenderX

Table 2. Differences of outcome variables between 2 groups at baseline (N=102).

Variable	Control (n=54), mean (SD)	Intervention (n=48), mean (SD)	*P* value
Physiological			
Weight (kg)	84.3 (15.93)	83.4 (19.21)	.81
BMI (kg/m^2)	29.9 (4.13)	29.6 (4.57)	.70
Liver fat score	333.6 (97.23)	317.0 (35.15)	.27
AST[a] (IU/L)	36.4 (24.40)	33.5 (14.23)	.47
ALT[b] (IU/L)	43.3 (31.50)	42.8 (29.80)	.93
-GT[c]	46.9 (34.86)	51.2 (40.49)	.57
Psychosocial			
Self-management	3.5 (0.65)	3.5 (0.52)	.75
Knowledge	7.6 (1.50)	7.3 (1.90)	.37
Self-efficacy	7.2 (1.42)	7.2 (1.14)	.81
Fatigue	4.3 (2.09)	4.3 (1.97)	.98
Depression	8.5 (8.73)	8.6 (7.65)	.96
Quality of life	92.2 (16.51)	91.3 (16.49)	.78

[a]AST: aspartate aminotransferase.

[b]ALT: alanine transaminase.

[c] -GT: gamma-glutamyl transferase.

Within-Group Effects

Changes in Physiological Outcomes

There were significant decreases in weight and BMI from baseline to 6 months after using the app (Table 3). Weight decreased significantly in the intervention group (*P*<.001), and BMI was improved over 6 months (*P*<.001). Furthermore, there were significant reductions in the liver fat score (*P*=.01), AST (*P*=.03), ALT (*P*=.002), and -GT (*P*=.04) in participants of the intervention group. However, there were no significant improvements in the physiological factors of the control group (all *P*>.05).

XSL•FO
RenderX

Table 3. The changes from baseline to 6 months within each group (N=102).

Variable	Control group (n=54)			Intervention group (n=48)		
	Baseline, mean (SD)	6 months, mean (SD)	*P* value	Baseline, mean (SD)	6 months, mean (SD)	*P* value
Physiological						
Weight (kg)	84.3 (15.93)	82.2 (16.10)	.13	83.4 (19.21)	80.9 (19.62)	<.001
BMI (kg/m^2)	29.9 (4.13)	29.6 (4.30)	.08	29.6 (4.57)	29.0 (4.81)	<.001
Liver fat score	333.6 (97.23)	318.7 (35.52)	.26	317.0 (35.5)	305.4 (45.10)	.01
AST[a,b] (IU/L)	37.1 (25.49)	36.2 (30.84)	.79	33.6 (14.13)	29.5 (12.52)	.03
ALT[c,d] (IU/L)	42.1 (31.06)	36.5 (32.53)	.15	39.1 (24.50)	31.0 (23.34)	.002
-GT[e,f]	47.9 (37.64)	43.9 (33.48)	.30	50.6 (38.71)	38.4 (18.34)	.04
Psychosocial						
Self-management	3.5 (0.65)	3.6 (0.62)	.66	3.5 (0.52)	3.9 (0.66)	<.001
Knowledge	7.6 (1.50)	7.2 (1.28)	.11	7.3 (1.90)	7.6 (1.60)	.18
Self-efficacy	7.2 (1.42)	7.3 (1.58)	.80	7.2 (1.14)	7.5 (1.32)	.07
Fatigue	4.3 (2.09)	4.5 (2.04)	.25	4.3 (1.97)	3.7 (1.80)	.005
Depression	8.5 (8.73)	8.5 (8.59)	.97	8.6 (7.65)	6.8 (7.57)	.003
Quality of life	92.2 (16.51)	93.4 (17.56)	.35	91.3 (16.49)	98.0 (17.15)	<.001

[a]AST: aspartate aminotransferase.

[b]Samples for the analysis in the control and intervention group were 49 and 43, respectively.

[c]ALT: alanine transaminase.

[d]Samples for the analysis in the control and intervention group were 49 and 43, respectively.

[e] -GT: gamma-glutamyl transferase.

[f]Samples for the analysis in the control and intervention group were 45 and 38, respectively.

Changes in Psychosocial Outcomes

In the intervention group, there were significant improvements in self-management (*P*<.001), fatigue (*P*=.005), depression (*P*=.003), and quality of life (*P*<.001) scores. However, no significant changes were observed in the control group (Table 3).

Between-Group Effects

There were changes in the physical and physiological outcomes in both groups, but no significant differences were observed between the intervention and control groups at 3 or 6 months (Table 4). In psychosocial outcomes, self-management was significantly improved in the intervention group at 3 months and 6 months compared to the control group (mean 0.30, SD 0.47 vs mean 0.09, SD 0.47; *P*=.02; and mean 0.37, SD 0.61 vs mean 0.03, SD 0.55; *P*=.004). Self-management knowledge also was significantly improved in the intervention group at 3 months and 6 months compared to the control group (mean 0.50, SD 1.38 vs mean –0.31, SD 1.68; *P*=.009; and mean 0.37, SD 0.61 vs mean –0.35, SD 1.59; *P*=.04). Fatigue and depression were significantly reduced in the intervention group at 6 months compared to the control group (mean –0.60, SD 1.42 vs mean 0.22, SD 1.39; *P*=.004; and mean –1.75, SD 3.84 vs mean –0.08, SD 4.05; *P*=.04). Quality of life in the intervention group had significantly improved at 6 months than that in the control group (mean 6.71, SD 12.21 vs mean 1.24, SD 9.70; *P*=.01). The treatment-by-time interaction was significant in self-management and fatigue, indicating that the intervention effects over time in these variables differed significantly between the intervention group and control group across the 6 months.

XSL•FO
RenderX

Table 4. The effects of intervention between the control and intervention groups (N=102).

Variable	Change from baseline, mean (SD)		Between-groups difference, mean (95% CI)	*P* value	
	Control (n=54)	Intervention (n=48)		Intervention vs control	Treatment×time
Weight (kg)					
3 months	–1.57 (2.98)	–2.07 (3.11)	0.50 (–0.70 to 1.69)	.41	.65
6 months	–0.79 (3.76)	–1.54 (2.74)	0.74 (–0.56 to 2.05)	.26	.59
Weight loss (%)					
3 months	–1.82 (3.40)	–2.52 (3.51)	0.70 (–0.66 to 2.06)	.31	.32
6 months	–1.07 (3.34)	–2.08 (3.44)	1.01 (–0.54 to 2.55)	.20	.15
BMI (kg/m^2)					
3 months	–0.57 (1.05)	–0.75 (1.09)	0.18 (–0.24 to 0.60)	.40	.47
6 months	–0.32 (1.32)	–0.58 (0.98)	0.26 (–0.20 to 0.72)	.27	.41
Liver fat score					
6 months	–14.91 (96.31)	–11.60 (30.34)	–3.30 (–32.09 to 25.48)	.82	N/A[a]
AST[b,c,d] (IU/L) (n=92)					
6 months	–0.90 (23.52)	–4.09 (12.21)	3.20 (–4.73 to 11.12)	.43	N/A
ALT[c,e,f] (IU/L) (n=92)					
6 months	–5.73 (27.04)	–8.05 (15.96)	2.32 (–6.84 to 11.48)	.62	N/A
-GT[c,g,h] (n=83)					
6 months	–3.96 (25.07)	–12.16 (34.36)	8.20 (–4.81 to 21.21)	.21	N/A
Self-management					
3 months	0.09 (0.47)	0.30 (0.47)	–0.21 (–0.40 to –0.03)	.025	.12
6 months	0.03 (0.55)	0.37 (0.61)	–0.34 0.57 to –0.11)	.004	.006
Knowledge					
3 months	–0.31 (1.68)	0.50 (1.38)	–0.81 (–1.42 to –0.21)	.009	.07
6 months	–0.35 (1.59)	0.29 (1.47)	–0.64 (–1.25 to –0.04)	.04	.23
Self-efficacy					
3 months	–0.10 (0.93)	–0.03 (1.12)	–0.07 (–0.47 to 0.33)	.73	.97
6 months	0.04 (1.03)	0.29 (1.10)	–0.26 (–0.67 to 0.16)	.230	.45
Fatigue					
3 months	0.13 (1.24)	–0.13 (1.36)	0.27 (–0.24 to 0.78)	.30	.46
6 months	0.22 (1.39)	–0.60 (1.42)	0.82 (0.27 to 1.37)	.004	.02
Depression					
3 months	–0.64 (4.14)	–1.00 (3.73)	0.36 (–1.19 to 1.92)	.65	.85
6 months	–0.08 (4.05)	–1.75 (3.84)	1.67 (0.11 to 3.22)	.04	.26
Quality of life					
3 months	1.17 (8.01)	3.71 (8.11)	–2.54 (–5.71 to 0.63)	.12	.59
6 months	1.24 (9.70)	6.71 (12.21)	–5.47 (–9.78 to –1.16)	.01	.13

[a]N/A: not applicable.

[b]AST: aspartate aminotransferase.

[c]Total numbers were not 102 because of missing data.

[d]Samples for the analysis in the control and intervention group were 49 and 43, respectively.

[e]ALT: alanine transaminase. Samples for the analysis in the control and intervention group were 49 and 43, respectively.

XSL•FO
RenderX

[f]Samples for the analysis in the control and intervention group were 49 and 43, respectively.

[g] -GT: gamma-glutamyl transferase.

[h]Samples for the analysis in the control and intervention group were 45 and 38, respectively.

SMART-Liver App Compliance

Mean compliance was 82.6% (SD 12.9%) at 3 months and 79.8% (SD 13.9%) at 6 months. Table 5 shows whether there were significant differences in study outcomes among 3 compliance groups in the intervention group, which were classified as 70% or less, 70% to 90%, and 90% or over. In the group with 90% or more, there were significantly greater improvements in weight loss (*P*=.01), BMI (*P*=.005), liver fat score (*P*=.002), AST (*P*<.001), and -GT (*P*=.001) at 6 months than in the groups with 70% or less and 70% to 90%. Self-management and knowledge at 3 and 6 months showed significantly greater improvement in the group with higher compliance as well (*P*<.001 and *P*<.001; *P*=.008 and *P*=.001, respectively). Fatigue in the group with 70% to 90% at 6 months was significantly lower compared with the group with 70% or less compliance (*P*=.03), and depression in the groups with 70% to 90% and with 90% or more at 3 months was higher than in the group with 70% or less (*P*=.008). The patterns of mean changes in outcomes from baseline to 6 months in each compliance group are shown in Figure 3.

XSL•FO
RenderX

Table 5. Changes of outcomes from baseline by the compliance level in the intervention group.

Outcome	Compliance (n=48), mean (SD)			*P* value
	A: < 70% (n=15)	B: 70%-90% (n=18)	C: ≥90% (n=15)	
Weight (kg)				
3 months	–1.5 (3.97)	–2.2 (3.27)	–2.4 (1.80)	.37
6 months	–0.6 (2.22)	–1.8 (2.91)	–2.2 (2.92)	.01 (A<C)
BMI (kg/m^2)				
3 months	–0.5 (1.37)	–0.8 (1.13)	–0.9 (0.69)	.26
6 months	–0.2 (0.77)	–0.7 (1.05)	–0.8 (1.02)	.005 (A<C)
Liver fat score				
6 months	–4.4 (23.90)	–6.8 (30.09)	–24.5 (34.00)	.002 (A=B<C)
AST[a] (IU/L)				
6 months	–1.8 (6.35)	–0.3 (12.95)	–12.1 (13.81)	<.001 (A=B<C)
ALT[b] (IU/L)				
6 months	–12.2 (12.36)	–3.6 (15.75)	–8.8 (19.68)	.03 (A>B)
-GT[c]				
6 months	–11.3 (21.89)	–1.4 (15.47)	–28.6 (55.21)	.001 (B<C)
Self-management				
3 months	<0.1 (0.31)	0.3 (0.48)	0.6 (0.38)	<.001 (A<B<C)
6 months	<0.1 (0.58)	0.4 (0.63)	0.6 (0.53)	<.001 (A<B<C)
Knowledge				
3 months	0.4 (1.30)	0.2 (1.42)	1.0 (1.36)	.008 (B<C)
6 months	0.2 (1.70)	–0.2 (1.50)	0.9 (0.96)	.001 (A=B<C)
Self-efficacy				
3 months	<0.1 (0.79)	–0.2 (1.28)	0.1 (1.25)	.35
6 months	0.4 (0.83)	0.3 (1.06)	0.2 (1.41)	.73
Fatigue				
3 months	0.2 (1.57)	–0.2 (1.20)	–0.4 (1.35)	.08
6 months	–0.1 (1.33)	–0.8 (1.39)	–0.8 (1.51)	.03 (A<B=C)
Depression				
3 months	–0.2 (3.88)	–1.3 (3.53)	–1.4 (3.94)	.22
6 months	–0.9 (2.59)	–1.8 (5.01)	–2.6 (3.27)	.10
Quality of life				
3 months	0.7 (6.81)	4.8 (8.77)	5.5 (8.15)	.008 (A<B=C)
6 months	3.8 (7.63)	8.7 (15.12)	7.3 (12.26)	.13

[a]AST: aspartate aminotransferase.

[b]ALT: alanine transaminase.

[c] -GT: gamma-glutamyl transferase.

XSL•FO
RenderX

Figure 3. Pattern of mean changes in outcomes for 6 months in the intervention group. For a higher-resolution version of this figure, see Multimedia Appendix 2.

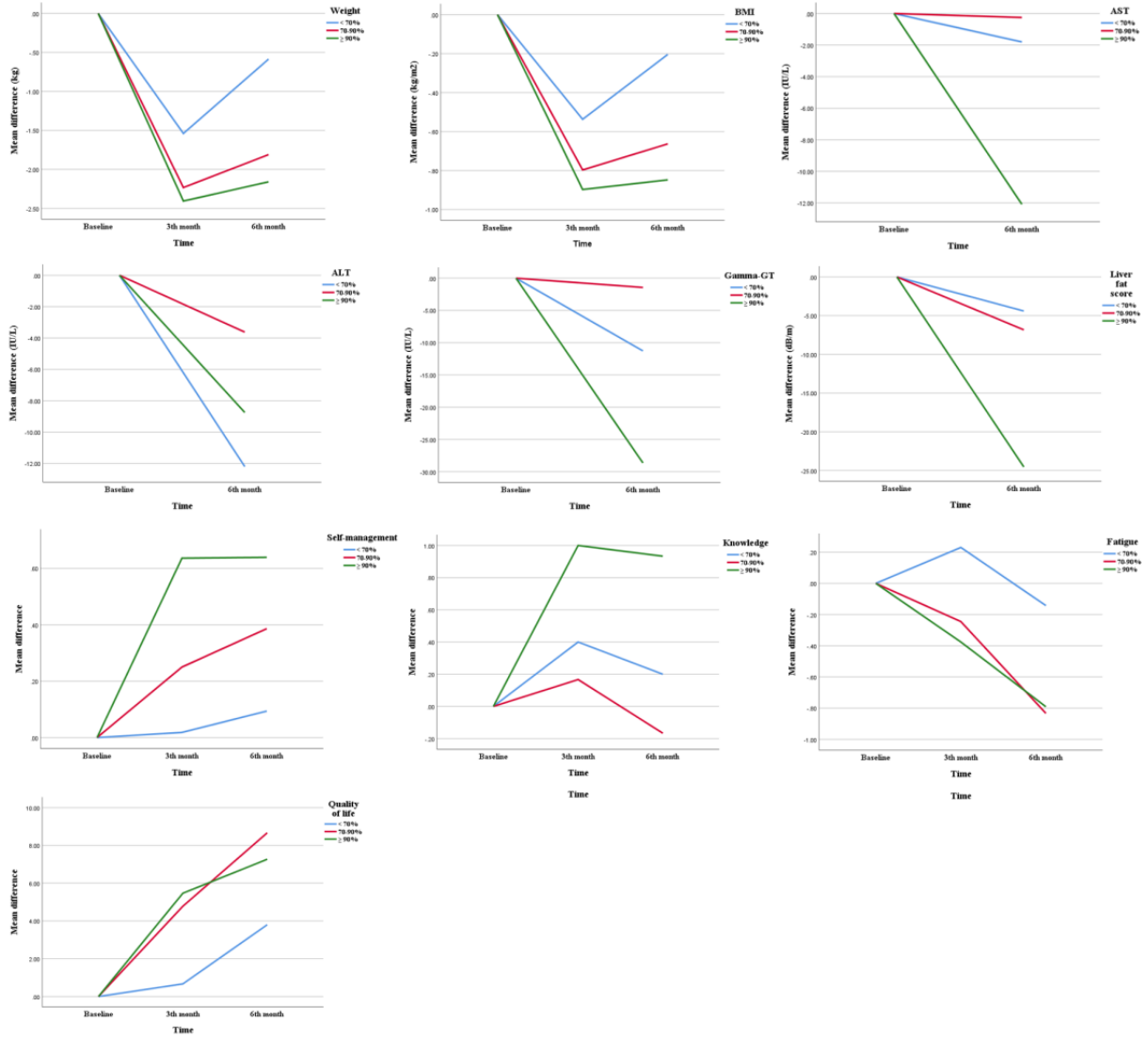

Discussion

Principal Findings

The purposes of this study were to develop and evaluate the physiological and psychosocial effects of the lifestyle coaching intervention for patients with NAFLD. The results showed that there were significant changes in physiological outcomes (body weight, BMI, liver fat score, AST, ALT, and -GT) and psychosocial outcomes (self-management, fatigue, depression, and quality of life) in the intervention group for 6 months, and the intervention-by-time interaction was effective for self-management and fatigue improvement.

There was a significant reduction in weight within the intervention group from baseline at 6 months in this study. Similarly, there was a significant change in BMI in the intervention group than in the control group, which meant a more accurate weight loss representation. This result is consistent with previous studies that reported effects on weight loss in patients with NAFLD [17,29]. Weight loss is recommended for the treatment of NAFLD in the absence of pharmacological treatment. Previous intervention studies on patients with NAFLD have mainly focused on improving physiological outcomes, including body weight and clinical aspects [17,29,30].

In this study, other physiological outcomes, such as liver fat score, AST, ALT, and -GT levels, significantly changed in the intervention group at 6 months. The previous studies presented that weight loss through lifestyle modification is known to be associated with a reduction in liver fat and recommend a 5% minimum weight loss [5,31,32]. Lifestyle modifications, such as diet changes and exercise routines, have been reported to be effective in reducing body weight [33-35] but did not directly result in the reduction of participants' liver fat scores in this study. The reason for this ineffectiveness in improving liver fat scores in this intervention may be related to the fact that participants did not present a significant body weight loss that could result in the improvement of NAFLD-related outcomes. Despite the lack of liver fat reduction at 6 months, there were significant changes in liver enzymes (AST, ALT, and -GT). Generally, liver enzyme levels in simple steatosis cases are mostly not high because their abnormalities are related to inflammatory processes [36]. Nevertheless, our results showed

XSL•FO
RenderX

a reduction in the mean levels of liver enzymes at 6 months. Furthermore, we found that changes in liver enzymes were significantly different only among participants with a high level of intervention compliance. Although AST and ALT have a limitation of moderate accuracy in evaluating the fatty liver [37], this result suggests that further intervention studies need to be developed to identify the effects of changes in liver enzymes and to determine the significant level of weight loss and adherence on these changes.

In the interest of improving the health conditions of patients with NAFLD, psychological factors should also be taken into consideration, as many studies have reported a relationship between physiological biomarkers and psychosocial factors [38,39]. In this respect, our study identified that participants' level of self-management improved in the intervention group compared to the control group and that the SMART-Liver app was significantly effective in improving self-management among patients with NAFLD at 6 months. Self-management behaviors that modify unhealthy lifestyles related to diet and lack of physical activity are a cornerstone of NAFLD treatment [35]. Several studies have conducted lifestyle modification interventions using mobile phones or other electronic devices and have identified the effects of physiological factors in patients with NAFLD [17,29,40,41]. However, other interventions targeting patients with chronic illnesses have evaluated their effects on psychosocial variables, such as self-management, self-efficacy, and quality of life, as well as on physical factors [40-43]. Our findings support the idea that lifestyle modification interventions can significantly improve psychosocial and physical parameters among patients with NAFLD.

Even though NAFLD is an asymptomatic disease, the SMART-Liver program was also effective in managing NAFLD's nonspecific symptoms, such as fatigue. Recently, Golubeva et al [44] reported that fatigue was almost twice more common in patients with NAFLD than in those without it. However, chronic liver disease fatigue management remains a challenge because the pathophysiology of fatigue is complex. From this perspective, health outcomes following lifestyle modifications such as weight loss or reduction in liver fat may positively affect fatigue.

Additionally, participants with high compliance (90% and above) showed higher reductions in body weight and BMI at 6 months. Clinical factors (liver fat, AST, and -GT) and psychosocial factors (self-management and self-management knowledge) were also significantly improved by higher participant compliance (90% and above) compared to other groups. Similarly, there were significant differences in the level of fatigue and quality of life among compliance groups, with higher changes in the 70% or more compliance groups than in the group of 70% or less. Generally, higher compliance can improve health outcomes [45-47]. However, evidence on the standard level of compliance in lifestyle modification interventions using mobile apps is lacking, and many studies that included patients with various chronic diseases have mainly reported medication adherence [48,49]. Our study was conducted to derive changes in health outcomes according to lifestyle behavior modification compliance. These findings may be helpful in setting goals to improve the outcomes of the intervention in this population.

In this study, the attrition rate in the intervention group was low (9/111, 8.1%). The attrition of mobile-using interventions for chronic diseases was previously reported to be 40% [49]. In prior interventions in patients with NAFLD, attrition rates ranged from 5% to 27%, and interventions using the coaching method had especially lower rates [15,16]. Health coaching is an individual education method that can improve health behaviors and the effectiveness of interventions for chronic disease management [50]. Our study applied the coaching method to guide tailored goals by considering the participants' circumstances. The participants were provided counseling via a remote meeting system halfway through the intervention period, which reset the goal based on their behavior and compliance and discussed their obstacles in changing unhealthy habits. In addition, the number of male participants with NAFLD was higher than female participants in this study even though we did not control for the sex ratio. Generally, in research, the proportion of female participants is higher than that of male participants [51]. As this study did not analyze the difference between male and female participants, further research is needed to compare this aspect.

In addition, compared to other liver diseases, most patients with NAFLD have less severe physical symptoms, and they can continue to engage in social activities, including work, relatively without constraints. In this respect, patients could participate in this program using a mobile app without time or space constraints, and if they had obstacles in sustaining their healthy lifestyles, they could contact and receive help from our research team. This might contribute to lowering the attrition rate. Further studies in this population may be successful if conducted considering these individual approaches.

Strengths and Limitations

We found that this lifestyle coaching intervention using a mobile app was successful in improving the level of self-management, which could facilitate healthy behaviors for weight loss and reduction of liver fat. Therefore, the results of this study can help future studies to develop lifestyle modification interventions and apply them to interventions for patients with other chronic diseases.

This study had some limitations. First, the intervention was conducted in a single hospital with only mobile phone users using the Android system. Although the use of mobile phones has increased, it can be affected by age, educational level, and household income. Therefore, future studies could include various types of mobile phones to diversify the target population. Second, the intervention involved counseling for goal setting via a remote meeting system that was not embedded in the app. The research team separately provided a meeting link to each participant, even though the meeting was informed through a chat channel within the app. Future studies may benefit from the development of an app with all possible functions needed for intervention. Finally, it would be beneficial to identify the effects of self-management on participant groups classified by NAFLD progression or clinical findings. Additionally, we did not collect data on app usage of frequency in this study.

XSL•FO
RenderX

However, this frequency may be associated with the intervention effect [52]. Further study may be conducted on the relationship between intervention effect and app usage frequency in this population.

Conclusions

We developed a mobile app–based intervention for patients with NAFLD that provides personalized information, counseling, and interactions to facilitate self-management behaviors. This study showed that mobile app–based lifestyle modification interventions for patients with NAFLD effectively improved their physiological and psychosocial health outcomes. Our results suggest that future studies should include a more diverse and wider population and evaluate the effects of psychosocial aspects on self-management, as well as on clinical health outcomes in lifestyle coaching interventions using a mobile app. To confirm the effectiveness of the only coaching intervention, further studies that consider a mobile app–based lifestyle intervention without coaching as a group should be conducted.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2017R1D1A1B04032264) and the Brain Korea 21 FOUR project funded by the National Research Foundation (NRF) of Korea, Yonsei University College of Nursing. We would like to thank Doing Lap for their cooperation and help throughout this research with their special technology. During the preparation of this, work the authors did not use generative artificial intelligence in any portion of the manuscript writing.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

OYK contributed to developing the study concept, data collection and analysis, and manuscript writing—original draft and editing. MKL and HK provided methodological guidance and data analysis. HWL and JSL contributed to the data collection and analysis. YJ led the development of the study concept, data analysis and interpretations, and manuscript review and editing. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist.
[PDF File (Adobe PDF File), 8397 KB - jmir_v26i1e49839_app1.pdf]

Multimedia Appendix 2
Pattern of mean changes in outcomes for 6 months in the intervention group.
[PNG File , 361 KB - jmir_v26i1e49839_app2.png]

References

1. Riazi K, Azhari H, Charette JH, Underwood FE, King JA, Afshar EE, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022;7(9):851-861. [doi: 10.1016/S2468-1253(22)00165-0] [Medline: 35798021]
2. Im HJ, Ahn YC, Wang JH, Lee MM, Son CG. Systematic review on the prevalence of nonalcoholic fatty liver disease in South Korea. Clin Res Hepatol Gastroenterol 2021;45(4):101526 [FREE Full text] [doi: 10.1016/j.clinre.2020.06.022] [Medline: 32919911]
3. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016;64(1):73-84. [doi: 10.1002/hep.28431] [Medline: 26707365]
4. Leoni S, Tovoli F, Napoli L, Serio I, Ferri S, Bolondi L. Current guidelines for the management of non-alcoholic fatty liver disease: a systematic review with comparative analysis. World J Gastroenterol 2018;24(30):3361-3373 [FREE Full text] [doi: 10.3748/wjg.v24.i30.3361] [Medline: 30122876]
5. Vilar-Gomez E, Martinez-Perez Y, Calzadilla-Bertot L, Torres-Gonzalez A, Gra-Oramas B, Gonzalez-Fabian L, et al. Weight loss through lifestyle modification significantly reduces features of nonalcoholic steatohepatitis. Gastroenterology 2015;149(2):367-378.e5; quiz e14 [FREE Full text] [doi: 10.1053/j.gastro.2015.04.005] [Medline: 25865049]

XSL•FO
RenderX

6. Nakanishi N, Hashimoto Y, Okamura T, Ohbora A, Kojima T, Hamaguchi M, et al. A weight regain of 1.5 kg or more and lack of exercise are associated with nonalcoholic fatty liver disease recurrence in men. Sci Rep 2021;11(1):19992 [FREE Full text] [doi: 10.1038/s41598-021-99036-y] [Medline: 34620897]
7. Oh S, Tsujimoto T, Kim B, Uchida F, Suzuki H, Iizumi S, et al. Weight-loss-independent benefits of exercise on liver steatosis and stiffness in Japanese men with NAFLD. JHEP Rep 2021;3(3):100253 [FREE Full text] [doi: 10.1016/j.jhepr.2021.100253] [Medline: 33898958]
8. Yki-Järvinen H, Luukkonen PK, Hodson L, Moore JB. Dietary carbohydrates and fats in nonalcoholic fatty liver disease. Nat Rev Gastroenterol Hepatol 2021;18(11):770-786 [FREE Full text] [doi: 10.1038/s41575-021-00472-y] [Medline: 34257427]
9. Gu Y, Zhou R, Kong T, Zhang W, Chen Y, Wang C, et al. Barriers and enabling factors in weight management of patients with nonalcoholic fatty liver disease: a qualitative study using the COM-B model of behaviour. Health Expect 2023;26(1):355-365 [FREE Full text] [doi: 10.1111/hex.13665] [Medline: 36385729]
10. Nguyen VH, Yeo YH, Zou B, Le MH, Henry L, Cheung RC, et al. Discrepancies between actual weight, weight perception and weight loss intention amongst persons with NAFLD. J Intern Med 2021;289(6):840-850 [FREE Full text] [doi: 10.1111/joim.13203] [Medline: 33340186]
11. Schnall R, Higgins T, Brown W, Carballo-Dieguez A, Bakken S. Trust, perceived risk, perceived ease of use and perceived usefulness as factors related to mHealth technology use. Stud Health Technol Inform 2015;216:467-471 [FREE Full text] [Medline: 26262094]
12. Smahel D, Elavsky S, Machackova H. Functions of mHealth applications: a user's perspective. Health Informatics J 2019;25(3):1065-1075 [FREE Full text] [doi: 10.1177/1460458217740725] [Medline: 29121831]
13. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, et al. Guidelines for reporting of health interventions using mobile phones: mobile Health (mHealth) Evidence Reporting and Assessment (mERA) checklist. BMJ 2016;352:i1174 [FREE Full text] [doi: 10.1136/bmj.i1174] [Medline: 26988021]
14. Nasi G, Cucciniello M, Guerrazzi C. The role of mobile technologies in health care processes: the case of cancer supportive care. J Med Internet Res 2015;17(2):e26 [FREE Full text] [doi: 10.2196/jmir.3757] [Medline: 25679446]
15. Axley P, Kodali S, Kuo YF, Ravi S, Seay T, Parikh NM, et al. Text messaging approach improves weight loss in patients with nonalcoholic fatty liver disease: a randomized study. Liver Int 2018;38(5):924-931. [doi: 10.1111/liv.13622] [Medline: 29117472]
16. Kwon OY, Choi JY, Jang Y. The effectiveness of eHealth interventions on lifestyle modification in patients with nonalcoholic fatty liver disease: systematic review and meta-analysis. J Med Internet Res 2023;25:e37487 [FREE Full text] [doi: 10.2196/37487] [Medline: 36689264]
17. Tomeno W, Kawashima K, Yoneda M, Saito S, Ogawa Y, Honda Y, et al. Non-alcoholic fatty liver disease comorbid with major depressive disorder: the pathological features and poor therapeutic efficacy. J Gastroenterol Hepatol 2015;30(6):1009-1014. [doi: 10.1111/jgh.12897] [Medline: 25619308]
18. Choi JM, Chung GE, Kang SJ, Kwak MS, Yang JI, Park B, et al. Association between anxiety and depression and nonalcoholic fatty liver disease. Front Med (Lausanne) 2020;7:585618 [FREE Full text] [doi: 10.3389/fmed.2020.585618] [Medline: 33537324]
19. Tapper EB, Lai M. Weight loss results in significant improvements in quality of life for patients with nonalcoholic fatty liver disease: a prospective cohort study. Hepatology 2016;63(4):1184-1189 [FREE Full text] [doi: 10.1002/hep.28416] [Medline: 26698379]
20. Kang SH, Lee HW, Yoo JJ, Cho Y, Kim SU, Lee TH, et al. KASL clinical practice guidelines: management of nonalcoholic fatty liver disease. Clin Mol Hepatol 2021;27(3):363-401 [FREE Full text] [doi: 10.3350/cmh.2021.0178] [Medline: 34154309]
21. Kurt S. ADDIE model: instructional design. Educational Technology. 2017. URL: https://educationaltechnology.net/the-addie-model-instructional-design/ [accessed 2022-09-07]
22. Kwon OY, Jang Y, Lee JY, Kim SU, Ahn SH. Development and initial validation of the nonalcoholic fatty liver disease self-management questionnaire. Res Nurs Health 2021;44(5):844-853. [doi: 10.1002/nur.22164] [Medline: 34120361]
23. Kim CJ, Chae SM, Yoo H. Psychometric testing of the Chronic Disease Self-Efficacy Scale-Korean version (CDSES-K). J Transcult Nurs 2012;23(2):173-180. [doi: 10.1177/1043659611434059] [Medline: 22294333]
24. Lorig K, Stewart A, Ritter P, Gonzalez V, Lynch J, Laurent D. Outcome Measures for Health Education and Other Health Care Interventions. Thousand Oaks, CA: Sage Publications; 1996.
25. Jang Y, Kim JH, Lee K. Validation of the revised piper fatigue scale in Koreans with chronic hepatitis B. PLoS One 2017;12(5):e0177690 [FREE Full text] [doi: 10.1371/journal.pone.0177690] [Medline: 28542250]
26. Piper BF, Dibble SL, Dodd MJ, Weiss MC, Slaughter RE, Paul SM. The revised piper fatigue scale: psychometric evaluation in women with breast cancer. Oncol Nurs Forum 1998;25(4):677-684. [Medline: 9599351]
27. Rhee M, Lee Y, Park S, Sohn C, Chung WY, Hong JS, et al. A standardization study of Beck Depression Inventory I-Korean version (K-BDI): reliability and factor analysis. Korean J Psychopathol 1995;4:77-95.
28. Min SK, Lee CI, Kim KI, Suh SY, Kim DK. Development of Korean version of WHO Quality of Life scale abbreviated version (WHOQOL-BREF). J Korean Neuropsychiatr Assoc 2000:571-579.

XSL•FO
RenderX

29. Koutoukidis DA, Astbury NM, Tudor KE, Morris E, Henry JA, Noreik M, et al. Association of weight loss interventions with changes in biomarkers of nonalcoholic fatty liver disease: a systematic review and meta-analysis. JAMA Intern Med 2019;179(9):1262-1271 [FREE Full text] [doi: 10.1001/jamainternmed.2019.2248] [Medline: 31260026]
30. Golabi P, Locklear CT, Austin P, Afdhal S, Byrns M, Gerber L, et al. Effectiveness of exercise in hepatic fat mobilization in non-alcoholic fatty liver disease: systematic review. World J Gastroenterol 2016;22(27):6318-6327 [FREE Full text] [doi: 10.3748/wjg.v22.i27.6318] [Medline: 27468220]
31. Hallsworth K, Adams LA. Lifestyle modification in NAFLD/NASH: facts and figures. JHEP Rep 2019;1(6):468-479 [FREE Full text] [doi: 10.1016/j.jhepr.2019.10.008] [Medline: 32039399]
32. El-Agroudy NN, Kurzbach A, Rodionov RN, O'Sullivan J, Roden M, Birkenfeld AL, et al. Are lifestyle therapies effective for NAFLD treatment? Trends Endocrinol Metab 2019;30(10):701-709 [FREE Full text] [doi: 10.1016/j.tem.2019.07.013] [Medline: 31422872]
33. Younossi ZM, Corey KE, Lim JK. AGA clinical practice update on lifestyle modification using diet and exercise to achieve weight loss in the management of nonalcoholic fatty liver disease: expert review. Gastroenterology 2021;160(3):912-918 [FREE Full text] [doi: 10.1053/j.gastro.2020.11.051] [Medline: 33307021]
34. Ma J, Hennein R, Liu C, Long MT, Hoffmann U, Jacques PF, et al. Improved diet quality associates with reduction in liver fat, particularly in individuals with high genetic risk scores for nonalcoholic fatty liver disease. Gastroenterology 2018;155(1):107-117 [FREE Full text] [doi: 10.1053/j.gastro.2018.03.038] [Medline: 29604292]
35. Funuyet-Salas J, Martín-Rodríguez A, Conrad R, Pérez-San-Gregorio M. Psychological biomarker profile in NAFLD/NASH with advanced fibrosis. In: Romero-Gomez M, editor. NAFLD and NASH: Biomarkers in Detection, Diagnosis and Monitoring. Cham: Springer; 2020:205-223.
36. Sharma P, Arora A. Clinical presentation of alcoholic liver disease and non-alcoholic fatty liver disease: spectrum and diagnosis. Transl Gastroenterol Hepatol 2020;5:19 [FREE Full text] [doi: 10.21037/tgh.2019.10.02] [Medline: 32258523]
37. Wong VWS, Adams LA, de Lédinghen V, Wong GLH, Sookoian S. Noninvasive biomarkers in NAFLD and NASH—current progress and future promise. Nat Rev Gastroenterol Hepatol 2018;15(8):461-478. [doi: 10.1038/s41575-018-0014-9] [Medline: 29844588]
38. Macavei B, Baban A, Dumitrascu DL. Psychological factors associated with NAFLD/NASH: a systematic review. Eur Rev Med Pharmacol Sci 2016;20(24):5081-5097 [FREE Full text] [Medline: 28051263]
39. Fard SJ, Ghodsbin F, Kaviani MJ, Jahanbin I, Bagheri Z. The effect of follow up (telenursing) on liver enzymes in patients with nonalcoholic fatty liver disease: a randomized controlled clinical trial. Int J Community Based Nurs Midwifery 2016;4(3):239-246 [FREE Full text] [Medline: 27382590]
40. Pfirrmann D, Huber Y, Schattenberg JM, Simon P. Web-based exercise as an effective complementary treatment for patients with nonalcoholic fatty liver disease: intervention study. J Med Internet Res 2019;21(1):e11250 [FREE Full text] [doi: 10.2196/11250] [Medline: 30602434]
41. Chao DY, Lin TM, Ma WY. Enhanced self-efficacy and behavioral changes among patients with diabetes: cloud-based mobile health platform and mobile app service. JMIR Diabetes 2019;4(2):e11017 [FREE Full text] [doi: 10.2196/11017] [Medline: 31094324]
42. Robinson SA, Shimada SL, Quigley KS, Moy ML. A web-based physical activity intervention benefits persons with low self-efficacy in COPD: results from a randomized controlled trial. J Behav Med 2019;42(6):1082-1090 [FREE Full text] [doi: 10.1007/s10865-019-00042-3] [Medline: 30980223]
43. Yamamura S, Nakano D, Hashida R, Tsutsumi T, Kawaguchi T, Okada M, et al. Patient-reported outcomes in patients with non-alcoholic fatty liver disease: a narrative review of chronic liver disease questionnaire-non-alcoholic fatty liver disease/non-alcoholic steatohepatitis. J Gastroenterol Hepatol 2021;36(3):629-636 [FREE Full text] [doi: 10.1111/jgh.15172] [Medline: 32627871]
44. Golubeva JA, Sheptulina AF, Yafarova AA, Mamutova EM, Kiselev AR, Drapkina OM. Reduced quality of life in patients with non-alcoholic fatty liver disease may be associated with depression and fatigue. Healthcare (Basel) 2022;10(9):1699 [FREE Full text] [doi: 10.3390/healthcare10091699] [Medline: 36141310]
45. Rossouw F, Du Plessis R, Reynecke F. Exercise compliance and health outcome in a chronic disease management programme. Afr J Phys Health Educ Recreat Dance 2012;18(Suppl 1):77-89.
46. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
47. Hayes E, McCahon C, Panahi MR, Hamre T, Pohlman K. Alliance not compliance: coaching strategies to improve type 2 diabetes outcomes. J Am Acad Nurse Pract 2008;20(3):155-162. [doi: 10.1111/j.1745-7599.2007.00297.x] [Medline: 18336692]
48. Fan K, Zhao Y. Mobile health technology: a novel tool in chronic disease management. Intell Med 2022;2(1):41-47 [FREE Full text] [doi: 10.1016/j.imed.2021.06.003]
49. Yasmin F, Banu B, Zakir SM, Sauerborn R, Ali L, Souares A. Positive influence of short message service and voice call interventions on adherence and health outcomes in case of chronic disease care: a systematic review. BMC Med Inform Decis Mak 2016;16:46 [FREE Full text] [doi: 10.1186/s12911-016-0286-3] [Medline: 27106263]

XSL•FO
RenderX

50. Olsen JM, Nesbitt BJ. Health coaching to improve healthy lifestyle behaviors: an integrative review. Am J Health Promot 2010;25(1):e1-e12. [doi: 10.4278/ajhp.090313-LIT-101] [Medline: 20809820]
51. Steinberg JR, Turner BE, Weeks BT, Magnani CJ, Wong BO, Rodriguez F, et al. Analysis of female enrollment and participant sex by burden of disease in US clinical trials between 2000 and 2020. JAMA Netw Open 2021;4(6):e2113749 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.13749] [Medline: 34143192]
52. Schoeppe S, Alley S, Van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016;13(1):127 [FREE Full text] [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]

Abbreviations

ADDIE: Analysis, Design, Development, Implementation, and Evaluation
ALT: alanine transferase
AST: aspartate aminotransferase
NAFLD: nonalcoholic fatty liver disease
-GT: gamma-glutamyl transferase

Edited by T de Azevedo Cardoso; submitted 13.06.23; peer-reviewed by M Bonacini, Y Qiu; comments to author 06.07.23; revised version received 22.08.23; accepted 17.01.24; published 15.02.24.

Please cite as:
Kwon OY, Lee MK, Lee HW, Kim H, Lee JS, Jang Y
Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial
J Med Internet Res 2024;26:e49839
URL: https://www.jmir.org/2024/1/e49839
doi:10.2196/49839
PMID:38358794

©Oh Young Kwon, Mi Kyung Lee, Hye Won Lee, Hyerang Kim, Jae Seung Lee, Yeonsoo Jang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing the Effectiveness of the Blended Delivery Mode With the Face-to-Face Delivery Mode of Smoking Cessation Treatment: Noninferiority Randomized Controlled Trial

Lutz Siemer[1,2,3], PhD; Marcel E Pieterse[2], PhD; Somaya Ben Allouch[4,5], PhD; Marloes G Postel[3], PhD; Marjolein G J Brusse-Keizer[6,7], PhD

[1]School of Social Work, Saxion University of Applied Sciences, Enschede, Netherlands

[2]Department of Psychology, Health and Technology, Centre for eHealth & Well-being Research - Behavioural, Management and Social Sciences, University of Twente, Enschede, Netherlands

[3]Research Group Technology, Health & Care, Saxion University of Applied Sciences, Enschede, Netherlands

[4]Digital Life Research Group, Amsterdam University of Applied Science, Amsterdam, Netherlands

[5]Digital Interactions Lab (DIL), Informatics Institute, University of Amsterdam, Amsterdam, Netherlands

[6]Medical School Twente, Medisch Spectrum Twente, Enschede, Netherlands

[7]Health Technology & Services Research, Technical Medical (TechMed) Centre, University of Twente, Enschede, Netherlands

Corresponding Author:
Lutz Siemer, PhD
School of Social Work
Saxion University of Applied Sciences
M. H. Tromplaan 28
Enschede, 7513 AB
Netherlands
Phone: 31 657459469
Email: l.siemer@saxion.nl

Abstract

Background: Tobacco consumption is a leading cause of death and disease, killing >8 million people each year. Smoking cessation significantly reduces the risk of developing smoking-related diseases. Although combined treatment for addiction is promising, evidence of its effectiveness is still emerging. Currently, there is no published research comparing the effectiveness of blended smoking cessation treatments (BSCTs) with face-to-face (F2F) treatments, where web-based components replace 50% of the F2F components in blended treatment.

Objective: The primary objective of this 2-arm noninferiority randomized controlled trial was to determine whether a BSCT is noninferior to an F2F treatment with identical ingredients in achieving abstinence rates.

Methods: This study included 344 individuals who smoke (at least 1 cigarette per day) attending an outpatient smoking cessation clinic in the Netherlands. The participants received either a blended 50% F2F and 50% web-based BSCT or only F2F treatment with similar content and intensity. The primary outcome measure was cotinine-validated abstinence rates from all smoking products at 3 and 15 months after treatment initiation. Additional measures included carbon monoxide–validated point prevalence abstinence; self-reported point prevalence abstinence; and self-reported continuous abstinence rates at 3, 6, 9, and 15 months after treatment initiation.

Results: None of the 13 outcomes showed statistically confirmed noninferiority of the BSCT, whereas 4 outcomes showed significantly (P<.001) inferior abstinence rates of the BSCT: cotinine-validated point prevalence abstinence rate at 3 months (difference 12.7, 95% CI 6.2-19.4), self-reported point prevalence abstinence rate at 6 months (difference 19.3, 95% CI 11.5-27.0) and at 15 months (difference 11.7, 95% CI 5.8-17.9), and self-reported continuous abstinence rate at 6 months (difference 13.8, 95% CI 6.8-20.8). The remaining 9 outcomes, including the cotinine-validated point prevalence abstinence rate at 15 months, were inconclusive.

Conclusions: In this high-intensity outpatient smoking cessation trial, the blended mode was predominantly less effective than the traditional F2F mode. The results contradict the widely assumed potential benefits of blended treatment and suggest that further research is needed to identify the critical factors in the design of blended interventions.

XSL•FO
RenderX

Trial Registration: Netherlands Trial Register 27150; https://onderzoekmetmensen.nl/nl/trial/27150
International Registered Report Identifier (IRRID): RR2-doi.org/10.1186/s12889-016-3851-x

(J Med Internet Res 2024;26:e47040) doi:10.2196/47040

KEYWORDS
tobacco; blended treatment; smoking cessation; randomized controlled trial; effectiveness; noninferiority; evaluation; mobile phone

Introduction

Tobacco's Global Impact

According to the World Health Organization [1], the tobacco epidemic is one of the biggest public health threats the world has ever faced: tobacco kills up to half of its consumers, which means >8 million people every year. Of these, >7 million deaths are because of direct tobacco consumption, whereas approximately 1.2 million are because of the exposure of nonsmokers to passive smoking [1]. The economic costs of tobacco consumption are considerable and include significant health costs of treating the disease caused by tobacco consumption and the loss of human capital through the morbidity and mortality attributable to tobacco consumption [1]. Smoking addiction is more prevalent in specific, often susceptible subpopulations, such as individuals in lower education or socioeconomic groups [2]. Approximately 80% of the world's 1.1 billion smokers live in low- and middle-income countries, where the burden of tobacco-related diseases and deaths is the highest [1].

Smoking Cessation Progress

People who stop smoking greatly reduce their risk of disease and early death [3] and will have major immediate and long-term health benefits [4,5]. Among smokers who are aware of the dangers of tobacco and the benefits of quitting, most want to quit [1]. Compared with quitting without professional support, smoking cessation treatment can more than double the success rates of quitting attempts [1]; this ultimately results—for treatments comparable with those in this study—in estimated point prevalence abstinence rates of 28.4% (95% CI 21.3-35.5) for treatments with a total amount of contact time of 91 to 300 minutes and of 24.7% (95% CI 21.0-28.4) for treatments with >8 person-to-person treatment sessions (both intention-to-treat [ITT]; 6 months after the quit date [6]). Previous research in the hospital smoking cessation clinic where this study was conducted showed a 19% abstinence rate for comparable treatment in the target group of patients with chronic obstructive pulmonary disease at the 12-month follow-up [7].

Blended Treatment Evolution

In the past decades, a variety of effective interventions for smoking cessation have become available [8,9], including, more recently, eHealth services such as web-based interventions [10,11] or mobile phone–based interventions [12,13]. At present, traditional face-to-face (F2F) interventions, on the one hand, and both web-based and mobile phone–based interventions, on the other hand, are increasingly being developed into blended treatments. This development is consistent with the idea that blended treatment combines the *best of both worlds* [14,15], as the strengths of one type of treatment should compensate for the weaknesses of the other [14-21]. For example, personal attention from a professional in F2F treatment could compensate for the lack of personal contact in web-based treatment. In turn, one of the main features of web-based treatment is the possibility of being available anytime and anywhere, which could bridge the interval between sessions in F2F treatment.

A systematic review of randomized controlled trials (RCTs) by Erbe et al [17] on blended F2F and web-based interventions suggests that compared with stand-alone F2F therapy, blended therapy may save clinician time and lead to lower dropout rates and higher abstinence rates in patients with substance abuse. The authors concluded that for common mental health disorders, blended interventions are feasible and can be more effective than no-treatment controls, but more RCTs on the effectiveness of blended treatments compared with nonblended treatments are necessary.

Objectives

Although promising, evidence available for blended deaddiction treatment is still emerging, addressing abuse of substances such as alcohol [22], cocaine, marijuana [23], or opioids [24]. For smoking cessation, we only found studies on the promising adjunctive use of smartphone apps with F2F contact [25-27]. To the best of our knowledge, there is no published research on the effectiveness of blended smoking cessation treatments (BSCTs) compared with F2F treatments, where in the blended treatment, the web-based components are not an adjunct but a substitute for specific F2F treatment components. Therefore, in this study, we present the results of an RCT comparing a blended 50% F2F and 50% web-based smoking cessation treatment with a traditional F2F treatment that was similar in content and intensity. The primary objective was to determine if a BSCT resulted in noninferior abstinence rates compared with an F2F treatment with identical ingredients. The rationale for choosing a noninferiority design was that we expected secondary benefits for the BSCT, such as lower costs, lower dropouts, and higher patient satisfaction, even if the BSCT only led to comparable abstinence rates.

Methods

Design

This study reports the results of an unblinded 2-arm, parallel group, noninferiority RCT with 1:1 allocation using stratified randomization (nicotine dependency, internet skills, and quitting strategy).

XSL•FO
RenderX

Setting

The study was conducted at the outpatient smoking cessation clinic (Dutch: *Stoppen met Roken Poli*) of the Medical Spectrum Twente Hospital in Enschede, the Netherlands. Enschede is a municipality and city in the east of the Netherlands with a population of 150,000 inhabitants. The estimated daily smoking prevalence in Enschede was 17.2% in 2017, which is approximately the same as the average (17.4%) in the Netherlands, which is one of the countries with the least number of smokers in Europe [28].

Trial Registration

The trial was registered in the Netherlands Trial Register on March 24, 2015 (Acronym: LiveSmokefree-Study; Title: Blended Smoking Cessation Treatment; new ID: NL5975; old ID: NTR5113), and a detailed protocol has been published previously [29].

Participants

We recruited participants between March 2015 and March 2019. The participants were self-referred to the treatment or were referred to the clinic by their general practitioner or hospital physician and were called by members of the research department to check for eligibility. Eligible participants were current daily smokers (eg, at least 1 cigarette, cigar, pipe, or e-cigarette per day [30]), those who were aged ≥18 years, those who had access to the internet (eg, email and websites), and those who were able to read and write Dutch. Eligible patients completed a questionnaire at the beginning of the study before being randomized.

Ethical Considerations

Consistent with the Dutch Medical Research Ethics Committee guidelines, this study was approved by the accredited Medical Research Ethics Committee Twente (P14-37/NL50944.044.14) and subsequently by the Board of Directors of Medisch Spectrum Twente Hospital. Before initiation, the trial was registered and a detailed protocol has been published previously [29].

A patient information letter outlining the burden of participation was distributed to all patients, and eligible patients attended an intake interview and signed a consent form.

We processed participants' personal data in accordance with the Dutch Personal Data Protection Act. The data were collected in 2 ways as follows:

1. The data of the personal contacts were recorded on data collection forms and collected centrally at Medisch Spectrum Twente Hospital. The data manager of the study recorded all collected data in an Access 2007 (Microsoft Corp) database.
2. Most data were collected by Tactus Addiction Treatment, a regional addiction care organization with expertise in web-based treatment, using web-based questionnaires offered to both treatment groups.

Individual patients and caregivers had a log-in with a username and password secured by the Secure Sockets Layer. All data transferred between the patient's PC and the application were encrypted and sent using the https protocol. All data were encrypted and stored on servers in secure data centers in the Netherlands. To further ensure data security, daily backups of the server were performed.

The participants did not receive any compensation for their participation in the study.

Interventions

The study interventions to be compared were a blended F2F treatment and web-based BSCT and an F2F treatment. Except for the differences in the mode of delivery (ie, F2F mode and web mode), both treatments had the same features as follows:

1. High-intensity treatment that comprised 10 sessions (20-minute contact time for each session, except for the first session, which lasts 50 minutes) and supportive pharmacotherapy, if needed, within a 6-month period with an expected quit date after about 3 months
2. Delivered by health care professionals in an outpatient cessation clinic
3. Concordant with the Dutch guidelines for tobacco addiction [31], fulfilling the requirements of the Dutch care module for smoking cessation [32]
4. Executed by counselors registered in the Dutch quality register of qualified smoking cessation counselors
5. Supporting 3 quitting strategies that patients could choose at the start of the treatment: (1) stop at once, (2) change gradually by increasing the number of daily activities that are performed smoke free, or (3) decrease smoking at regular intervals (eg, scheduled smoking reduction by 100%->75% and 75%->50%). The chosen quitting strategy did not generally influence the course of the treatment. The order, pace, duration, and intensity were the same for all strategies.

Both BSCT and F2F treatment covered 52 behavior change techniques (using behavior change technique taxonomy v1 of 93 hierarchically clustered techniques by Michie et al [33]) as shown in Table 1.

XSL•FO
RenderX

Table 1. Behavior change techniques included in the BSCT[a] and F2F[b] treatment.

Grouping	Behavior change techniques
1. Goals and planning	1.1. Goal setting (behavior)
	1.2. Problem-solving
	1.3. Goal setting (outcome)
	1.4. Action planning
	1.5. Review behavior goal or goals
	1.6. Discrepancy between current behavior and goal
	1.8. Behavioral contract
	1.9. Commitment
2. Feedback and monitoring	2.3 Self-monitoring of behavior
	2.4 Self-monitoring of outcome or outcomes of behavior
	2.6 Biofeedback
	2.7 Feedback on outcome or outcomes of behavior
3. Social support	3.1 Social support (unspecified)
4. Shaping knowledge	4.2 Information about antecedents
	4.3 Reattribution
5. Natural consequences	5.1 Information about health consequence
	5.2 Salience of consequences
	5.3 Information about social and environmental consequences
	5.4 Monitoring of emotional consequences
	5.5 Anticipated regret
	5.6 Information about emotional consequences
6. Comparison of behavior	6.2 Social comparison
	6.3 Information about others' approval
7. Associations	7.4 Remove access to the reward
8. Repetition and substitution	8.1 Behavioral practice and rehearsal
	8.2 Behavior substitution
	8.3 Habit formation
	8.4 Habit reversal
	8.6 Generalization of a target behavior
	8.7 Graded tasks
9. Comparison of outcomes	9.1 Credible source
	9.2 Pros and cons
	9.3 Comparative imagining of future outcomes
10. Reward and threat	10.7 Self-incentive
	10.9 Self-reward
11. Regulation	11.1 Pharmacological support
	11.2 Reduce negative emotions
12. Antecedents	12.1 Restructuring the physical environment
	12.2 Restructuring the social environment
	12.3 Avoidance or reducing exposure to cues for the behavior
	12.4 Distraction
13. Identity	13.1 Identification of self as role model
	13.2 Framing and reframing
	13.5 Identity associated with changed behavior

XSL•FO
RenderX

Grouping	Behavior change techniques
14. Scheduled consequences	14.4 Reward approximation
	14.5 Rewarding completion
	14.6 Situation-specific reward
	14.7 Reward incompatible behavior
	14.8 Reward alternative behavior
15. Self-belief	15.1 Verbal persuasion about capability
	15.3 Focus on past success
16. Covert learning	16.3 Vicarious consequences

[a]BSCT: blended smoking cessation treatment.

[b]F2F: face-to-face.

F2F treatment consisted of 10 F2F sessions delivered at an outpatient smoking cessation clinic. BSCT consisted of 5 F2F sessions at the outpatient clinic and 5 web-based sessions delivered via the web-based treatment platform Roken De Baas (which translates loosely as “in control of smoking”). During the RCT, the software had to be revised once, as the European General Data Protection Regulation became enforceable from May 25, 2018, which changed the appearance and handling but not the content of the interventions.

Both F2F treatment and BSCT consisted of counselor-dependent and counselor-independent components. The counselor-dependent web-based components of BSCT were interactive and relied on asynchronous communication (eg, email and SMS text messaging) between the counselor and patient. The counselor-independent components such as psychoeducational content or the smoking diary were used by the patients on their own and at their own time. In F2F treatment, these components were provided in a paper manual that clients took home. In BSCT, these components were accessible over the web. As such, both treatments were equivalent in terms of content and intensity. However, an additional benefit of BSCT was that the content of previous counselor-dependent components remained accessible as email and SMS text messaging correspondence saved on the web.

The characteristic feature of BSCT is an equal balance between F2F and web-based sessions, and the focus of the treatment was not supposed to be on the F2F mode or the web mode; in addition, there was a constantly alternating and interacting use of the F2F mode and web mode. Table 2 presents the order, timing, main features, duration, and modes of delivery of the treatment sessions for F2F treatment and BSCT. Although an even distribution was planned for BSCT with regard to the number of sessions, there was an uneven distribution for the duration of treatment because the first session (50 minutes of F2F mode) was longer than the remaining sessions (20 minutes of F2F mode or 20 minutes of web mode); therefore, BSCT patients spent 130 minutes in the F2F mode and 100 minutes in the web mode.

XSL•FO
RenderX

Table 2. Characteristics of the treatment sessions in F2F[a] treatment and BSCT[b] according to treatment protocol.

Session	Week	Main features	Duration (N=230 minutes)	Mode of delivery of treatment sessions	
				BSCT (130 minutes of F2F mode[c] and 100 minutes of web mode[d])	F2F treatment (230 minutes of F2F mode)
1	1	Goal setting, prompt smoking diary, and measure CO[e] level	50	F2F	F2F
2	3	Measures for self-control	20	Web	F2F
3	5	Dealing with withdrawal	20	F2F	F2F
4	7	Breaking habits	20	Web	F2F
5	9	Dealing with triggers	20	F2F	F2F
6	11	Food for thought	20	Web	F2F
7	14	Think differently and measure CO level and cotinine[f] level	20	F2F	F2F
8	18	Do differently	20	Web	F2F
9	22	Action plan and measure CO level	20	F2F	F2F
10	26	Closure	20	Web	F2F

[a]F2F: face-to-face.

[b]BSCT: blended smoking cessation treatment.

[c]F2F mode: F2F sessions of BSCT.

[d]Web mode: web-based sessions of BSCT.

[e]CO: carbon monoxide.

[f]Cotinine measurement was only performed in patients who reported quitting smoking either in the 3-month follow-up questionnaire or during treatment to the counselor.

More information about both treatments can be found in the study protocol of the RCT [29] and in the description of the user experiences of BSCT [21]. The treatment fidelity of the counselors was not recorded. The adherence to the treatments was described elsewhere [34,35]; but, in brief, levels of adherence were comparable for BSCT and F2F treatment sessions. To provide an impression of the look and feel of the web interventions of BSCT, Multimedia Appendix 1 displays screenshots of the web-based sessions of BSCT.

Outcomes

For the primary objective (ie, effectiveness) of the analysis, the primary outcome for the ITT analysis of the treatments' effectiveness in smoking cessation was the proportions of biochemically (ie, cotinine) validated point prevalence abstinence from all combustible tobacco products (eg, cigarettes, bags, cigars, and pipes) at 3 and 15 months after the start of the treatment. Additional outcomes were the proportions of carbon monoxide (CO)–validated point prevalence abstinence; self-reported point prevalence abstinence; and self-reported continuous abstinence at 3 (ie, shortly after the expected quit date), 6 (ie, end of treatment), and 9 and 15 (follow-up measurements) months. Applying the noninferiority margin justified in our protocol paper [29], BSCT was considered as noninferior if it resulted in abstinence rates that were <5% points lower than those of F2F treatment [29].

Measurements

Effectiveness

Cotinine-validated and CO-validated abstinence measurements were used to measure biochemically validated point prevalence abstinence rates [36-38].

Cotinine measurement was performed at approximately the 3-month and 15-month follow-up (ie, shortly after the expected quit day, week 14; refer to Table 2) and at the 15-month follow-up only in patients who reported quitting either during the treatment to the counselor or in the 3-month or 15-month follow-up questionnaire. A 0.5 mL to 1 mL salivary sample was collected using a Salivette (Sarstedt AG and Co). Under supervision, patients chew on a cotton swab for 1 minute to stimulate the saliva flow rate. All saliva specimens were frozen until assayed and transported to the laboratory for the determination of cotinine levels using a gas chromatography technique. Abstinence was defined as having a salivary cotinine level <20 ng/mL [39].

The CO level was measured in all patients (independent of reporting quitting) at 3 months, at the last F2F treatment session at the hospital (for the BSCT group, the last F2F treatment session was at 5 months after the start of the treatment [week 22], and for the F2F treatment group, at 6 months after start of the treatment [week 26]; refer to Table 2), and in patients who reported quitting at 15 months together with the cotinine level. A breath CO level of 5 ppm was taken as the cutoff value between smokers and nonsmokers (≥5 ppm in smokers and <5

XSL•FO
RenderX

ppm in nonsmokers [40]). Breath CO levels were monitored using a piCO Smokerlyzer (Bedfont Instruments), a portable CO monitor.

Furthermore, self-reported point prevalence abstinence and self-reported continuous abstinence rates were measured at 3, 6, 9, and 15 months after treatment initiation. The measurement tool was a standardized questionnaire for Dutch tobacco research [30], which patients in both BSCT and F2F treatment completed over the web. Self-reported point prevalence abstinence rate was assessed by asking patients whether they had smoked ≥1 cigarette (eg, bags, cigars, and pipe) in the last 7 days, and the self-reported continuous abstinence rate was assessed by asking whether they had smoked since the current stop.

For each measurement during and after treatment, the participants were prompted twice via email and, in the absence of measurements, were additionally notified twice via telephone. If no measurement was available after 2 emails and 2 telephone calls, the participants were classified as lost to follow-up for the respective measurement and notified again for the next measurement.

Sample Size

For the RCT, we calculated the abstinence rates for 344 participants, assuming a long-term abstinence rate of 10% for those receiving F2F treatment [6,7,41] and—based on its expected benefits—15% for those receiving BSCT. If BSCT would lead to an abstinence rate not <5%, it would be considered as noninferior compared with F2F treatment. Therefore, 172 patients per group with a power of 80% and a Cronbach α of .025 were needed for this RCT (calculated using Power Analysis & Sample Size [NCSS Statistical Software]).

Randomization

We randomly allocated patients to either BSCT or F2F treatment using computerized randomization (Qminim Online Minimization). Randomization was performed at the individual level (allocation ratio 1:1). The minimization was stratified according to (1) the level of internet skills [42], (2) the level of nicotine dependence [30], and (3) the quitting strategy favored by the patient (eg, stop at once, gradual change, and scheduled reduced smoking; for details refer to the description in the *Interventions* section). The data used for minimization were collected using the baseline questionnaire, which was completed over the web by the patient after providing consent.

Blinding

Owing to the nature of the treatment conditions, it was self-evidently impossible to blind the staff and patients involved in the study.

Statistical Methods

For both the BSCT group and the F2F treatment group, the patients' demographic, smoking-related, and health-related characteristics at baseline were reported as means with SD for normally distributed continuous variables and as medians with IQR for nonnormally distributed continuous variables. Categorical variables were reported as numbers with corresponding percentages. To identify between-group differences, an independent 1-sided (1-tailed) *t* test or Mann-Whitney *U* test was performed as appropriate for continuous variables, and Pearson chi-square or Fisher exact test was performed for categorical variables.

As this was an ITT analysis, participants with missing data on smoking status were considered as smokers. The absolute and proportional abstinence rates in the treatment group were reported.

The noninferiority was analyzed by calculating the difference and the 95% CI of the observed difference in the abstinence rates and by comparing that to the previously defined noninferiority margin of 5% points [29]. In addition, the noninferiority analysis is illustrated in a forest chart.

To be able to compare the results of this study with those of other studies conducted using a more traditional RCT design, additional repeated measures analyses were conducted using generalized estimating equation to test for group, time, and group×time differences in abstinence rates.

All analyses were performed using the SPSS software (version 26.0; IBM Corp), except for the calculation of the CIs of the difference between abstinence rates, for which we used the web tool by VassarStats [43] for "The Confidence Interval For The Difference Between Two Independent Proportions."

Results

Participant Flow

Figure 1 shows the flow of participants throughout the study. A total of 344 patients were eligible for the study, provided written consent, and were randomized (smoking cessation treatment: BSCT, n=177; F2F treatment, n=177). Of 177 patients each in both groups, 167 (94.3%) patients of the BSCT group and all 177 (100%) of the F2F treatment group started treatment (ie, they received at least 1 session). Before the start of treatment, 151 (85.3%) of the 177 patients in the BSCT group and 175 (98.8%) of the 177 patients in the F2F treatment group completed the baseline questionnaire. Three months after starting treatment (ie, shortly after the expected quit date), of the 177 patients in the BSCT group, 14 (7.9%) who self-reported quitting were available for cotinine measurement, 68 (38.4%) were available for CO measurement, and 26 (14.6%) completed the follow-up questionnaire. Of the 177 patients in the F2F treatment group, 47 (26.5%) who reported quitting were available for cotinine measurement, 77 (43.5%) were available for CO measurement, and 47 (26.5%) completed the 3-month follow-up questionnaire. Of 177 patients in the BSCT group, 53 (29.9%) were available for the 5-month CO measurement and 18 (10.1%) completed the 6-month follow-up questionnaire. Of 177 patients in the F2F treatment group, 61 (34.4%) were available for the 6-month CO measurement and 53 (29.9%) completed the 6-month follow-up questionnaire. The 9-month follow-up questionnaire was completed by 20 (11.2%) patients of the BSCT group and 42 (23.7%) patients of the F2F treatment group. After 15 months of starting treatment, 9 (5.1%) of the 177 patients in the BSCT group who self-reported quitting were available for cotinine level measurement. A total of 16 (9%) patients in the BSCT group were available for CO level measurement and 7 (4%) completed the follow-up questionnaire.

XSL•FO
RenderX

Of 177 patients in the F2F treatment group, 12 (6.8%) patients who reported quitting were available for cotinine level measurement, 15 (8.5%) were available for CO level measurement, and 31 (17.5%) completed the 15-month follow-up questionnaire.

Figure 1. Flow of participants through the study. Withdrawals are noncumulative. BSCT: blended smoking cessation treatment; CO: carbon monoxide; F2F: face-to-face.

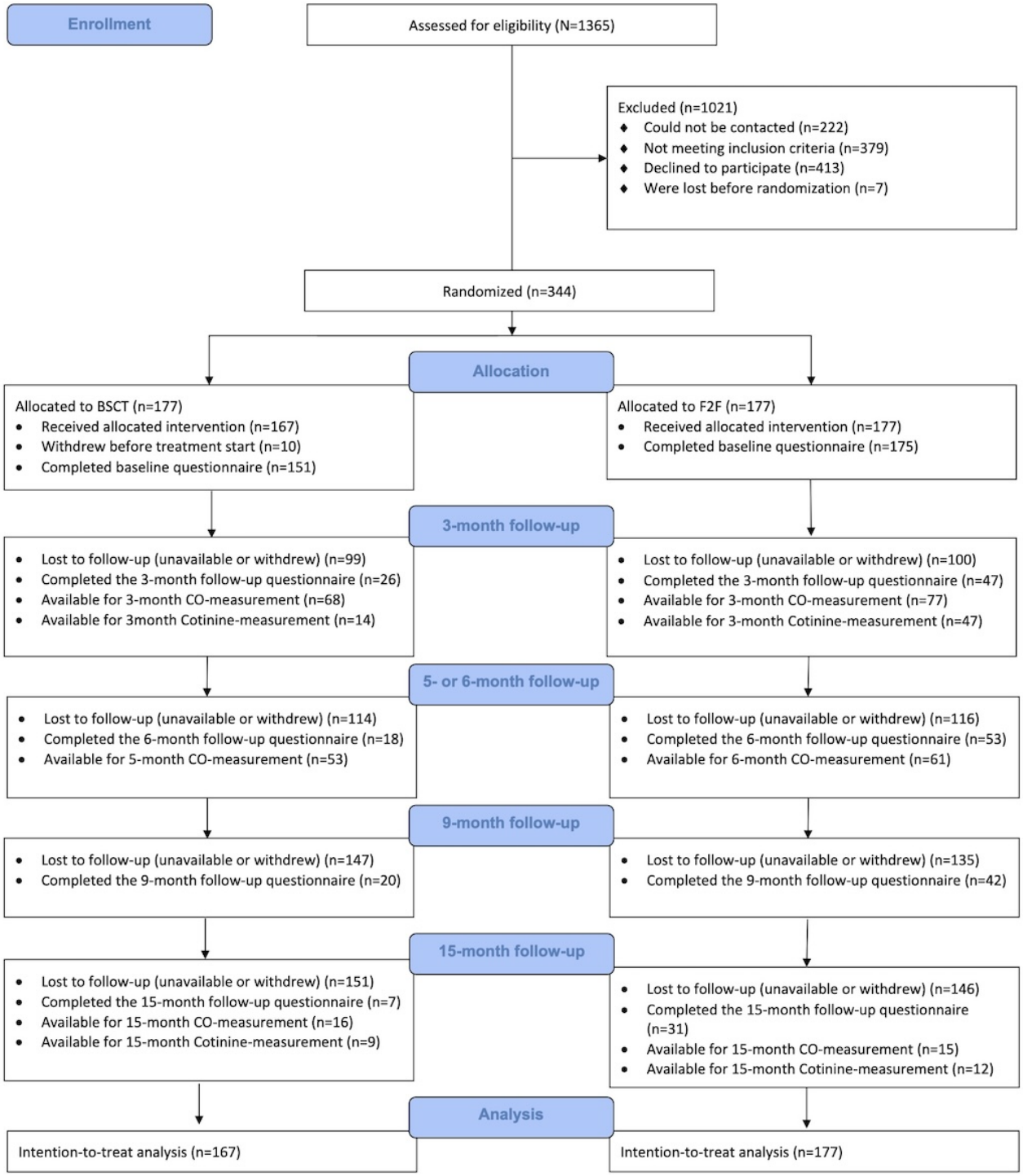

Baseline Characteristics

Table 3 shows that the baseline characteristics, including demographic, smoking-related, and health-related characteristics, were comparable between the participants in both groups.

XSL•FO
RenderX

Table 3. Patients' characteristics of the BSCT[a] group and the F2F[b] treatment group.

Characteristics	BSCT	F2F treatment
Demographics		
Sex (BSCT: n=152; F2F: n=175), n (%)		
Female	77 (50.7)	87 (49.7)
Male	75 (49.3)	88 (50.3)
Nationality (BSCT: n=152; F2F: n=175), n (%)		
Dutch	147 (96.7)	173 (98.9)
Other	5 (3.3)	2 (1.1)
Cultural background (BSCT: n=152; F2F: n=175), n (%)		
Dutch	137 (90.1)	161 (92)
Other	15 (9.9)	14 (8)
Age (y; BSCT: n=152; F2F: n=175), mean (SD)	47.5 (12.4)	46.4 (13.2)
Marital status (BSCT: n=151; F2F: n=175), n (%)		
With partner	99 (65.6)	107 (61.1)
Alone	52 (34.4)	68 (38.9)
Housing situation (BSCT: n=152; F2F: n=172), n (%)		
With children	63 (41.5)	73 (42.4)
Without children	89 (58.6)	99 (57.6)
Education (BSCT: n=152; F2F: n=171), n (%)		
VET[c] or higher	96 (63.1)	109 (63.7)
Lower than VET	56 (36.8)	62 (36.3)
Main income (BSCT: n=152; F2F: n=175), n (%)		
Wage or own company	79 (52)	89 (50.9)
Income support	73 (48)	86 (49.1)
Main day activity (BSCT: n=152; F2F: n=175), n (%)		
Paid work	77 (50.7)	85 (48.6)
Other	75 (49.3)	90 (51.4)
Internet skills[d] (BSCT: n=151; F2F: n=175), mean (SD)	38.4 (6.5)	39.4 (5.8)
Smoking-related characteristics		
Quitting strategy (BSCT: n=167; F2F: n=176), n (%)		
Stop at once	68 (40.7)	72 (40.9)
Change gradually	44 (26.4)	44 (25)
Scheduled reduction	55 (32.9)	60 (34.1)
Reason to start treatment (BSCT: n=152; F2F: n=175), n (%)		
Intrinsic	104 (68.4)	117 (66.9)
Extrinsic	48 (31.6)	58 (33.1)
Nicotine dependency (Fagerström[e]; BSCT: n=151; F2F: n=175), mean (SD)	5.3 (2.1)	5.2 (2.1)
Negative attitude toward quitting[f] (BSCT: n=151; F2F: n=175), mean (SD)	−5.5 (3.1)	−5.1 (2.9)
Positive attitude toward quitting[g] (BSCT: n=151; F2F: n=175), median (IQR)	10 (8-11)	10 (9-11)
Self-efficacy[h] (BSCT: n=151; F2F: n=175), mean (SD)	−0.2 (5.2)	−0.4 (4.9)
Readiness to quit[i] (BSCT: n=151; F2F: n=175), median (IQR)	2 (1-3)	2 (1-3)

XSL•FO
RenderX

Characteristics	BSCT	F2F treatment
Earlier quit attempts (BSCT: n=151; F2F: n=175), n (%)	128 (84.8)	154 (88)
Social support[j] (BSCT: n=151; F2F: n=175), median (IQR)	4 (3-4)	4 (3-5)
Social modeling[k] (BSCT: n=151; F2F: n=175), median (IQR)	3 (1-6)	3 (1-5)
Use of alcohol[l] (BSCT: n=151; F2F: n=175), median (IQR)	2 (1-3)	2 (1-3)
Use of (recreational) drugs (BSCT: n=151; F2F: n=175), n (%)	11 (7.3)	15 (8.6)
Health-related characteristics		
Use of medication in general (BSCT: n=151; F2F: n=175), n (%)	102 (67.6)	130 (74.3)
Use of medication for addiction treatment (BSCT: n=151; F2F: n=175), n (%)	0 (0)	0 (0)
Use of medication for psychiatric treatment (BSCT: n=151; F2F: n=159), n (%)	30 (19.9)	26 (16.4)
Use of medication for physical treatment (BSCT: n=151; F2F: n=159), n (%)	78 (51.7)	93 (58.5)
Use of other medication (BSCT: n=151; F2F: n=159), n (%)	25 (16.6)	31 (19.5)
Health-related complaints[m] (MAP HSS[n]; BSCT: n=151; F2F: n=175), mean (SD)	12.6 (6.2)	12.6 (6.6)
Smoking-related complaints[o] (BSCT: n=151; F2F: n=175), mean (SD)	21.0 (13.6)	20.8 (9.2)
Health- and smoking-related complaints[p] (BSCT: n=151; F2F: n=175), mean (SD)	33.6 (13.6)	33.4 (14.2)
Depression[q] (BSCT: n=151; F2F: n=175), median (IQR)	4 (0-10)	4 (2-12)
Anxiety[q] (BSCT: n=151; F2F: n=175), median (IQR)	4 (2-8)	6 (2-10)
Stress[q] (BSCT: n=151; F2F: n=175), median (IQR)	8 (4-16)	10 (4-16)
DASS[r] (BSCT: n=151; F2F: n=175), median (IQR)	18 (8-32)	20 (8-36)
EQ-5D-3L[s] (BSCT: n=151; F2F: n=175), median (IQR)	0.8 (0.7-1.0)	−0.8 (0.7-0.9)
EQ VAS[t] (BSCT: n=151; F2F: n=175), mean (SD)	65.5 (18.7)	64.6 (18.7)

[a]BSCT: blended smoking cessation treatment.
[b]F2F: face-to-face.
[c]VET: vocational education and training.
[d]Internet skills: range 10-60; higher numbers indicate better skills.
[e]Nicotine dependency (Fagerström): range 0-10; higher numbers indicate higher nicotine dependency.
[f]Negative attitude toward quitting: range −12 to 0; lower numbers indicate a more negative attitude toward quitting smoking.
[g]Positive attitude toward quitting: range 0-12; higher numbers indicate a more positive attitude toward quitting smoking.
[h]Self-efficacy: range −12 to 12; higher numbers indicate higher self-efficacy related to smoking cessation.
[i]Readiness to quit: range 0-4; higher numbers indicate higher readiness to quit.
[j]Social support: range 0-5; higher numbers indicate more social support in smoking cessation.
[k]Social modeling: range 0-8; higher numbers indicate more smokers in the social environment.
[l]Use of alcohol: range 0-4; 0=Never, 1=1 time per month, 2=2-4 times per month, 3=2-3 times per week, and 4=≥4 times per week.
[m]Health-related complaints: range 0-40; higher numbers indicate poorer health status.
[n]MAP HSS: Maudsley Addiction Profile Health Symptoms Scale.
[o]Smoking-related complaints: range 0-64; higher numbers indicate more smoking-related complaints.
[p]Health- and smoking-related complaints: range 0-104; higher numbers indicate poorer health status and more smoking-related complaints.
[q]Depression, anxiety and stress: range 0-42; higher numbers indicate a higher level of depression, anxiety and stress.
[r]DASS: sum score of depression, anxiety and stress (range 0-126; higher numbers indicate a more negative emotional status).
[s]EQ-5D-3L: societal-based quantification of the patients' health status (range 0-1; higher numbers indicate better health status).
[t]EQ VAS: visual analog scale for quality of life (range 0-100, higher numbers indicate better state of health).

Effectiveness

Table 4 shows the results of effectiveness measurements at 3, 5 or 6, 9, and 15 months after the start of treatment. The cotinine-validated point prevalence abstinence shortly after the expected stop day (ie, 3 months after the treatment initiation) showed a significantly lower and inferior abstinence rate in the BSCT group (4.8%) than in the F2F treatment group (17.5%; difference of 12.7, 95% CI 6.2-19.4; $P<.001$). The differences found in the 15-month cotinine level measurement (difference

XSL•FO
RenderX

of 1.5, 95% CI −3.5 to 6.4) and in all CO level measurements at 3 months (difference of 2.5, 95% CI −6.9 to 11.8), 5 or 6 months (difference 3.7, 95% CI −4.0 to 11.4), and 15 months (difference 0.7, 95% CI −4.9 to 6.7) were not substantial and inconclusive in terms of inferiority.

Table 4. Treatment effects on the participants in both BSCT[a] (n=167) and F2F[b] treatment (n=177) groups at 3, 5 or 6, 9, and 15 months after the start of treatment.

Outcome	3 months	5 or 6 months	9 months	15 months
Cotinine-validated point prevalence abstinence				
F2F treatment, n (%)	31 (17.5)	__[c]	—	10 (5.7)
BSCT, n (%)	8 (4.8)	—	—	7 (4.2)
Percentage points, difference (95% CI)	12.7 (6.2 to 19.4)	—	—	1.5 (−3.5 to 6.4)
CO[d]-validated point prevalence abstinence				
F2F treatment, n (%)	50 (28.2)	31 (17.5)	—	13 (7.3)
BSCT, n (%)	43 (25.7)	23 (13.8)	—	11 (6.6)
Percentage points, difference (95% CI)	2.5 (−6.9 to 11.8)	3.7 (−4.0 to 11.4)	—	0.7 (−4.9 to 6.7)
Self-reported point prevalence abstinence[e]				
F2F treatment, n (%)	35 (19.8)	48 (27.1)	39 (22)	26 (14.7)
BSCT, n (%)	22 (13.2)	13 (7.8)	19 (11.4)	5 (3)
Percentage points, difference (95% CI)	6.6 (−1.3 to 14.4)	19.3 (11.5 to 27.0)	10.7 (2.8 to 18.4)	11.7 (5.8 to 17.9)
Self-reported continuous abstinence[f]				
F2F treatment, n (%)	30 (16.9)	35 (19.8)	23 (13)	20 (11.3)
BSCT, n (%)	19 (11.4)	10 (6)	13 (7.8)	3 (1.8)
Percentage points, difference (95% CI)	5.6 (−1.9 to 13.0)	13.8 (6.8 to 20.8)	5.2 (−1.4 to 11.8)	9.5 (4.4 to 15.1)

[a]BSCT: blended smoking cessation treatment.

[b]F2F: face-to-face.

[c]Data not available.

[d]CO: carbon monoxide.

[e]Answer “no” to the questionnaire question “Have you smoked one or more cigarettes (bags, cigars, pipe) in the last 7 days?”

[f]Answer “no” to the questionnaire question “Have you smoked since the stop?”

Furthermore, we observed significantly lower and inferior abstinence rates in the BSCT group for self-reported point prevalence abstinence at 5 or 6 months (BSCT 7.8% vs F2F treatment 27.1%; difference 19.3, 95% CI 11.5-27.0; *P*<.001), for self-reported point prevalence abstinence at 15 months (BSCT 3% vs F2F treatment 14.7%; difference 11.7, 95% CI 5.8-17.9; *P*<.001), and for self-reported continuous abstinence at 5 or 6 months (BSCT 6% vs F2F treatment 19.8%; difference 13.8, 95% CI 6.8-20.8; *P*<.001). Significantly lower—but in terms of inferiority, inconclusive—abstinence rates in the BSCT group were found for self-reported point prevalence abstinence at 9 months (BSCT 11.4% vs F2F treatment 22%; difference 10.7, 95% CI 2.8-18.4; *P*=.009) and for self-reported continuous abstinence at 15 months (BSCT 1.8% vs F2F treatment 11.3%; difference 9.5, 95% CI 4.4-15.1; *P*<.001).

Figure 2 presents the 95% CIs of the differences between BSCT and F2F treatment groups for all abstinence outcome measures by applying the 5% points noninferiority margin. The forest plot illustrates the inferiority of BSCT with cotinine-validated point prevalence abstinence at 3 months, self-reported point prevalence abstinence at 6 and 15 months, and self-reported continuous abstinence at 6 months. For the remaining outcomes, the forest plot shows inconclusive results.

XSL•FO
RenderX

Figure 2. Forest plot of the risk differences between the blended smoking cessation treatment (BSCT; 95% CI) and face-to-face (F2F) treatment. CO: carbon monoxide.

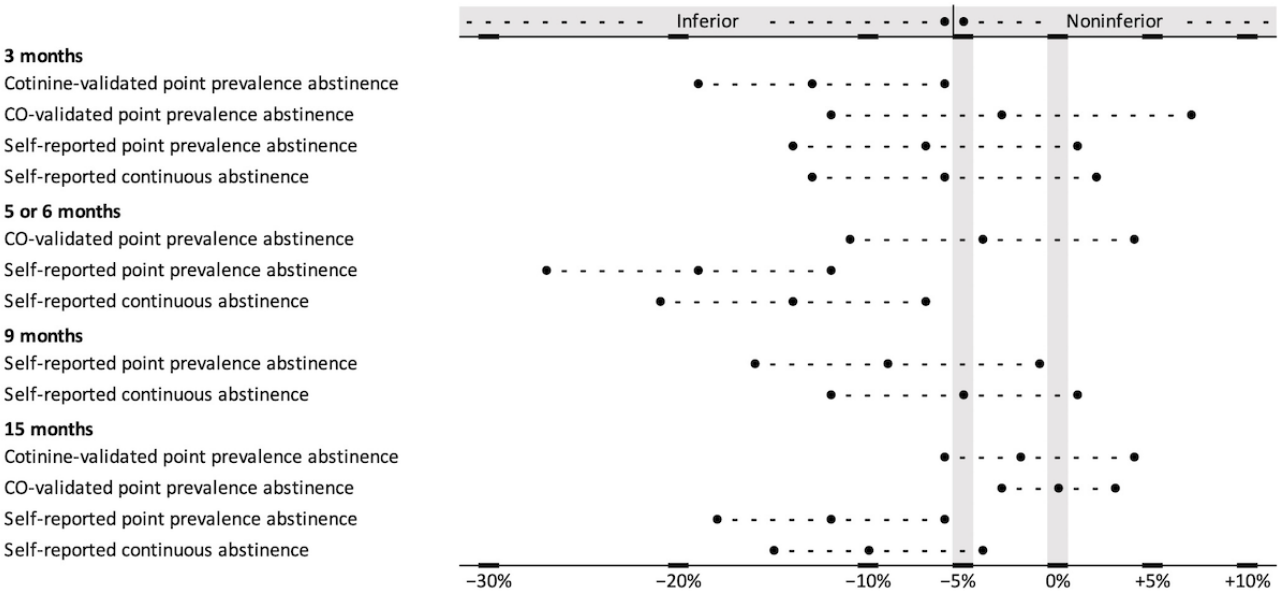

The generalized estimating equation analysis showed significant differences ($P<.05$) between both the groups with time and the group×time interaction for cotinine-validated point prevalence abstinence, self-reported point prevalence abstinence, and self-reported continuous abstinence rates. For the CO-validated point prevalence abstinence, a significant difference was found with time, but neither was there a difference between the groups nor a time×group interaction.

Discussion

Principal Findings

This paper presents the results of an RCT comparing the effectiveness of a blended 50% F2F treatment and 50% web-based BSCT to F2F-only treatment with similar ingredients and intensity. Contrary to our expectations, the abstinence rates of the BSCT group were lower than those of the F2F group. For the primary outcome (ie, cotinine-validated point prevalence abstinence rate), applying the 5%-point noninferiority margin indicated inferiority of BSCT at 3 months, whereas the outcome at 15 months was inconclusive. Both results should be considered with caution as the statistical power to detect differences was limited owing to nonresponse. Furthermore, BSCT was found to be inferior in 3 of the secondary outcomes (ie, self-reported point prevalence abstinence rate, self-reported continuous abstinence rate at 6 months, and self-reported point prevalence abstinence rate at 15 months), whereas the remaining outcomes were inconclusive. Although most outcomes from the repeated measures analyses showed significantly lower abstinence rates for the blended treatment, all remaining outcomes were nonsignificant, further corroborating the inferiority of BSCT against F2F treatment.

Given that our results suggest that it is more likely that BSCT is inferior to F2F treatment, our study is not consistent with the higher abstinence rates reported in the literature [17] for blended treatments compared with F2F treatment. Explanations for this likely inferiority of BSCT require further study. As the patients' demographic, smoking-related, and health-related characteristics were comparable in both treatment groups, these factors did not seem to play a role in this context. This also applies to adherence; as we found in previous analyses [35], adherence was comparable for both the groups. However, we know from qualitative analyses conducted as part of this RCT [21] that participants found the web-based components of BSCT to be rather unmotivating and not enjoyable, which may have resulted in BSCT patients making less use of the web-based components both during and after the treatment, and thus, may be a factor in the lower abstinence rates. The experience of patients in the BSCT suggests that the highly protocolized, equally balanced mix for blended treatment chosen in this study, with a fixed sequence of alternating F2F and web-based sessions, was too restrictive for blended treatment in practice [35], thus limiting tailoring to individual patient needs. Which intervention components should be offered when and in what form to achieve optimal treatment outcomes requires further investigation.

Furthermore, even if this cannot be supported by systematic observations and analyses, we believe that provider-related factors at the microlevel (eg, the treatment fidelity of counselors and therapist drift) and at the mesolevel (eg, the organization's preexisting knowledge, routines, and leadership) should be considered more closely. A relevant factor could be that, in the development of BSCT, half of the counseling sessions of the F2F treatment established in the outpatient smoking cessation clinic were replaced by web-based tools from a web platform unfamiliar to the clinic and counselors. Therefore, counselors had half of their F2F intervention replaced and had to integrate the new web-based components into a new blended workflow. The preexisting routine and familiarity with the F2F treatment among counselors might have disadvantaged the quality of execution of the blended treatment. The normalization process theory [44] could provide valuable perspectives in this context. It posits that the unclear definition of BSCT's meaningfulness (coherence) for counselors may have diminished their motivation and engagement (cognitive participation). Limited collective agency in BSCT's implementation, owing to rigid protocols

XSL•FO
RenderX

and insufficient reflective monitoring, may have further impeded its establishment in clinical practice. Understanding these barriers through targeted investigations could enhance the integration and efficacy of the BSCT.

Although not the focus of this analysis, we noticed that both treatments mostly showed lower abstinence rates than those reported in the literature for comparable treatments (ie, point prevalence abstinence rates of 28.4% for treatments with a total contact time of 91 to 300 minutes 6 months after the quit date [6]). At 9 months (ie, 6 months after the quit date), for F2F treatment, we found a self-reported point prevalence abstinence rate of 22% and a self-reported continuous abstinence rate of 13%. For BSCT, this rate was even lower with 11.4% for self-reported point prevalence abstinence and 7.8% for self-reported continuous abstinence. These relatively low abstinence rates could be because of the population characteristics (ie, patients in an outpatient smoking cessation clinic in a hospital context). Further analysis should investigate whether the sample differs from the general population in terms of known effectiveness predictors [45], such as, in this context, age, socioeconomic status, alcohol and drug use, health status, nicotine dependence, motivation to quit, or family status. However, a previous study by Christenhusz [7] in the same clinic with a comparable treatment for the specific target group of patients with chronic obstructive pulmonary disease showed much higher cotinine-validated abstinence rates (19%) compared with F2F treatment (5.7%) and BSCT (4.2%) at the 12-month follow-up. A more plausible explanation for this is the high dropout rate and missing data in this study. As we applied the common penalized imputation procedure (assuming missing=smoking [46,47]) to deal with missing data in our analyses, imputed quit rates will decrease proportionally to dropout rates.

A final point to consider is that toward the end of the RCT, the software of the web platform had to be updated because of legal changes, which temporarily caused accessibility problems. However, as only a few patients were affected by this and only toward the end of the RCT, these had no relevant influence on the results of this study.

Although blended treatment appears promising and reflects today's digitalization of the lifestyle of patients and health care professionals [14,15], not every realization of blended treatment is automatically an improvement. This also underscores the need to answer the question Greenhalgh et al [48] raised earlier: "What explains the success of a blended treatment in one context and the failure of a comparable blended treatment in another context?" The likely inferiority of BSCT in this study indicates that the current realization of BSCT will have to be reconsidered, which may involve aspects such as the optimal balance and mix of F2F and web-based components or the use of synchronous versus asynchronous counseling within web-based components. Such a redesign process can be supported by an analysis using the normalization process theory [44] and guided by an eHealth development model such as the Center for eHealth Research Road map [49]; the nonadoption, abandonment, scale-up, spread, and sustainability framework [48]; or more practically by the "Fit for Blended Care" instrument [15], which is intended to support therapists and patients in deciding whether and how blended care can be established.

For the generalization of the results, it should be noted that this analysis referred to a hospital context and a blended treatment with a strict 50:50 ratio of web-based and F2F interventions. For example, hospital patients could be expected to have a higher disease burden and, possibly, owing to age, a lower eHealth literacy than the general population. The question arises whether the results would have been different in a healthier, younger population. In addition, as mentioned above, a fixed 50:50 ratio of web-based and F2F interventions was defined for BSCT, which did not consider the individual needs of patients or counselors. A blended treatment that is better tailored to the needs, characteristics, and skills of both the patients and the counselors could have led to better results [15]. We know from an earlier study by Siemer et al [21] that patients would have preferred to use a smartphone app instead of a web platform, for example, or that they would have liked to be free to choose the ratio and sequence of F2F and web-based interventions.

Limitations

A major limitation of this study was the high dropout rate at several follow-up time points, resulting in many missing values for both self-reported and biochemical measures, which had a major impact on the ITT analysis. According to the ITT procedure, all missing values for the outcomes were coded as smoking. As a result, both the biochemically validated outcomes and the self-reported outcomes of this study are likely to be overly conservative, which largely explains the relatively low abstinence rates found in both study groups compared with the existing literature.

Furthermore, because of the high dropout rate, we conducted an analysis of the factors associated with dropout. We identified 2 main predictors of dropout: having a smoking partner at baseline and lower mental health scores as indicated by the Depression Anxiety Stress Scale [50]. Both predictors are known to be associated with poorer treatment outcomes [51]. This finding suggests that neither of the interventions used in this study sufficiently reduced the barriers to successful intervention completion. Such nonrandom patterns of dropout pose a threat to the external validity of our findings as they suggest that our sample may not be fully representative of the wider population. However, it is notable that these attrition factors alone are unlikely to fully explain the relatively low quit rates observed in this trial compared with other similarly intensive interventions [7] as similar reasons for dropout are likely to occur in any smoking cessation trial sample. Nevertheless, the underrepresentation of participants with these risk factors at later follow-ups may have led to an overestimation of the effectiveness of our interventions, although it remains difficult to assess whether this occurred to a greater extent in this study than in other smoking cessation trials.

However, the most critical aspect is whether these predictors of attrition varied between the 2 treatment conditions [52]. Such differential attrition could compromise the internal validity of our findings, particularly in the noninferiority test comparing blended treatment with F2F treatment delivery. Owing to low cell counts of the two above-mentioned dropout predictors, a

XSL•FO
RenderX

detailed analysis to examine the interaction effects of attrition predictors by treatment condition was not feasible. In addition, there were no consistent differences in the attrition rates between the study groups at any follow-up time points.

Another limitation to consider is the risk of bias in patient-reported outcome measures (PROMs), such as social desirability or recall bias, particularly given that the study relies in part on self-reported smoking cessation for an extended period of 15 months. The participants' ability to accurately recall their smoking behavior could be impaired, especially in the context of continuous abstinence, which could bias the study results. In general, the lower quit rates in our study compared with the existing literature argue against a significant self-report bias, as PROMs tend to overestimate quit rates compared with biochemical validation. Furthermore, biased PROMs will only have affected the internal validity of this study if they occur differently in the 2 study groups. We have no indications of this, but we lack data to verify this statistically. Because the determination of noninferiority is ultimately based on applying the 5% margin, it can also be considered a weakness that this 5% margin is based only on our considerations, as stated in the protocol paper of this study [29]. However, a slightly higher or lower margin would have led to slight changes in the results but not fundamental changes in the conclusions.

In addition to the cotinine measurements, this study collected CO measurements at the last 3 follow-up points. However, the second of these CO measurements showed a 4-week difference between the groups: 5 months after baseline for the blended treatment group and 6 months for the F2F group. Assuming that relapse rates would be expected to increase with time, this difference should have favored the effectiveness of the blended treatment. However, our results at this time point show the opposite, further supporting our claim of inferiority of the blended treatment compared with the F2F treatment.

Another limitation of this study is that our data, which were designed to compare the 2 approaches of blended and F2F treatment, did not allow analyses at the level of treatment components within the 2 delivery modes. Nevertheless, studies comparing different blended protocols are warranted to enable the design of improved BSCT in the future.

A final limitation of this study is that, as is often the case in clinical studies [53], we have not recorded the treatment fidelity and therefore deviations from the treatment protocol favoring one of both modes of delivery may have biased our findings. Although we cannot rely on systematic observations, we have some reason to believe that the implementation and adoption of the innovative BSCT may have had a negative impact on the effectiveness of the BSCT compared with the usual F2F treatment. However, based on our data on adherence from previous papers [34,35] and the findings on satisfaction (not reported in this study), we found no indication of a fidelity issue.

Conclusions

In this analysis of an RCT comparing a BSCT with a comparable F2F treatment, we found predominant results indicating inferiority of the blended mode compared with the traditional F2F mode, exceeding a 5% margin in abstinence rate. This could not be explained by lower adherence. Further research is required on the critical factors involved in the design of blended interventions.

Acknowledgments

The authors would like to thank the directors, the staff of both Medisch Spectrum Twente Hospital and Tactus Addiction Treatment, and the patients of the outpatient smoking cessation clinic of the Medical Spectrum Twente Hospital.

During the revision process, the authors assessed the efficacy of artificial intelligence (AI) tools such as ChatGPT 4.0 (OpenAI) [54], DeepL Translator (DeepL SE) [55], and DeepL Write (DeepL SE) [56] in enhancing language use. The objective of the authors was to determine whether this approach could optimize the academic writing quality for nonnative English speakers. Where the AI did not alter the content, the authors refined the language by altering the phrasing based on AI suggestions.

Data Availability

The data underlying this paper are available in Data Archiving and Networked Services [57].

Authors' Contributions

LS, MGP, MGJB-K, MEP, and SBA initiated collaboration with the data provider, designed the study, and wrote the study protocol. LS conducted the literature search, monitored data collection, drafted the paper, and is the guarantor of the paper. LS, MGP, and MGJB-K conducted the trial. LS and MGJB-K wrote the statistical analysis plan. LS and MGP designed the data collection tools. LS, MGJB-K, and MEP analyzed the data. LS, MGP, MGJB-K, MEP, and SBA revised the draft paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of web sessions of the blended smoking cessation treatment.
[PDF File (Adobe PDF File), 1447 KB - jmir_v26i1e47040_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
CONSORT eHEALTH checklist.
[PDF File (Adobe PDF File), 322 KB - jmir_v26i1e47040_app2.pdf]

References

1. Tobacco key facts. World Health Organization. URL: https://www.who.int/en/news-room/fact-sheets/detail/tobacco [accessed 2019-07-26]
2. Drope J, Liber AC, Cahn Z, Stoklosa M, Kennedy R, Douglas CE, et al. Who's still smoking? Disparities in adult cigarette smoking prevalence in the United States. CA Cancer J Clin 2018 Mar;68(2):106-115 [FREE Full text] [doi: 10.3322/caac.21444] [Medline: 29384589]
3. The health consequences of smoking—50 years of progress: a report of the surgeon general. Centers for Disease Control and Prevention. 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK179276/ [accessed 2024-01-23]
4. WHO report on the global tobacco epidemic 2019: offer help to quit tobacco use. World Health Organization. 2019 Jul 25. URL: https://www.who.int/publications/i/item/9789241516204 [accessed 2024-01-23]
5. Tobacco: health benefits of smoking cessation. World Health Organization. 2020 Feb 25. URL: https://www.who.int/tobacco/quitting/benefits/en/ [accessed 2024-01-23]
6. U. S. Department Human Services, Public Service. Treating Tobacco Use and Dependence 2008 Update - Clinical Practice Guideline. Scotts Valley, CA: CreateSpace Independent Publishing Platform; 2013.
7. Christenhusz LC. Smoking cessation in COPD patients: (cost-)effectiveness of the smokestoptherapy and validation of abstinence. University of Twente. 2006 Mar 24. URL: https://research.utwente.nl/en/publications/smoking-cessation-in-copd-patients-cost-effectiveness-of-the-smok [accessed 2024-01-23]
8. Stead LF, Koilpillai P, Fanshawe TR, Lancaster T. Combined pharmacotherapy and behavioural interventions for smoking cessation. Cochrane Database Syst Rev 2016 Mar 24;3(3):CD008286 [FREE Full text] [doi: 10.1002/14651858.CD008286.pub3] [Medline: 27009521]
9. Lancaster T, Stead LF. Individual behavioural counselling for smoking cessation. Cochrane Database Syst Rev 2005 Apr 18;3(2):CD001292. [doi: 10.1002/14651858.CD001292.pub2] [Medline: 15846616]
10. Civljak M, Stead LF, Hartmann-Boyce J, Sheikh A, Car J. Internet-based interventions for smoking cessation. Cochrane Database Syst Rev 2013 Jul 10;7:CD007078. [doi: 10.1002/14651858.CD007078.pub4] [Medline: 23839868]
11. Taylor GM, Dalili MN, Semwal M, Civljak M, Sheikh A, Car J. Internet-based interventions for smoking cessation. Cochrane Database Syst Rev 2017 Sep 04;9(9):CD007078 [FREE Full text] [doi: 10.1002/14651858.CD007078.pub5] [Medline: 28869775]
12. Whittaker R, Borland R, Bullen C, Lin RB, McRobbie H, Rodgers A. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2009 Oct 07;4:CD006611. [doi: 10.1002/14651858.CD006611.pub2] [Medline: 19821377]
13. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
14. van der Vaart R, Witting M, Riper H, Kooistra L, Bohlmeijer ET, van Gemert-Pijnen LJ. Blending online therapy into regular face-to-face therapy for depression: content, ratio and preconditions according to patients and therapists using a Delphi study. BMC Psychiatry 2014 Dec 14;14:355 [FREE Full text] [doi: 10.1186/s12888-014-0355-z] [Medline: 25496393]
15. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JE. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. JMIR Ment Health 2016 Feb 09;3(1):e9 [FREE Full text] [doi: 10.2196/mental.4534] [Medline: 26860537]
16. Barak A, Hen L, Boniel-Nissim M, Shapira N. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. J Technol Hum Serv 2008 Sep 08;26(2-4):109-160. [doi: 10.1080/15228830802094429]
17. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017 Sep 15;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
18. Kemmeren LL, van Schaik DJ, Riper H, Kleiboer AM, Bosmans JE, Smit JH. Effectiveness of blended depression treatment for adults in specialised mental healthcare: study protocol for a randomised controlled trial. BMC Psychiatry 2016 Apr 21;16:113 [FREE Full text] [doi: 10.1186/s12888-016-0818-5] [Medline: 27102812]
19. Postel MG, Witting M, van Gemert-Pijnen JE. Blended behandeling in de geestelijke gezondheidszorg. Dth kwartaaltijdschrift voor directieve therapie en hypnose 2013;33(3):210-221 [FREE Full text]
20. Siemer L, Allouch SB, Pieterse ME, Brusse-Keizer M, Sanderman R, Postel MG. Blended face-to-face and web-based smoking cessation treatment (BSCT): a description of patients' user experience (UX). Eur Respir J 2019;54(suppl 63):PA1690. [doi: 10.1183/13993003.congress-2019.PA1690]

XSL•FO
RenderX

21. Siemer L, Ben Allouch S, Pieterse ME, Brusse-Keizer M, Sanderman R, Postel MG. Patients' user experience of a blended face-to-face and web-based smoking cessation treatment: qualitative study. JMIR Form Res 2020 Jun 03;4(6):e14550 [FREE Full text] [doi: 10.2196/14550] [Medline: 32343245]
22. Blankers M. Commentary on Sundström et al. (2019): digital interventions for alcohol problems-time for more research on blended therapy. Addiction 2020 May;115(5):875-876. [doi: 10.1111/add.14962] [Medline: 32022373]
23. Carroll KM, Ball SA, Martino S, Nich C, Babuscio TA, Nuro KF, et al. Computer-assisted delivery of cognitive-behavioral therapy for addiction: a randomized trial of CBT4CBT. Am J Psychiatry 2008 Jul;165(7):881-888 [FREE Full text] [doi: 10.1176/appi.ajp.2008.07111835] [Medline: 18450927]
24. Christensen DR, Landes RD, Jackson L, Marsch LA, Mancino MJ, Chopra MP, et al. Adding an internet-delivered treatment to an efficacious treatment package for opioid dependence. J Consult Clin Psychol 2014 Dec;82(6):964-972 [FREE Full text] [doi: 10.1037/a0037496] [Medline: 25090043]
25. Barroso-Hurtado M, Suárez-Castro D, Martínez-Vispo C, Becoña E, López-Durán A. Smoking cessation apps: a systematic review of format, outcomes, and features. Int J Environ Res Public Health 2021 Nov 06;18(21):11664 [FREE Full text] [doi: 10.3390/ijerph182111664] [Medline: 34770178]
26. O'Connor M, Whelan R, Bricker J, McHugh L. Randomized controlled trial of a smartphone application as an adjunct to acceptance and commitment therapy for smoking cessation. Behav Ther 2020 Jan;51(1):162-177. [doi: 10.1016/j.beth.2019.06.003] [Medline: 32005334]
27. Nomura A, Tanigawa T, Muto T, Oga T, Fukushima Y, Kiyosue A, et al. Clinical efficacy of telemedicine compared to face-to-face clinic visits for smoking cessation: multicenter open-label randomized controlled noninferiority trial. J Med Internet Res 2019 Apr 26;21(4):e13520 [FREE Full text] [doi: 10.2196/13520] [Medline: 30982776]
28. van Laar M, van Gestel B, Cruts G, van der Pol P, Ketelaars T, Beenakkers E, et al. Jaarbericht nationale drug monitor 2018. Trimbos-instituut. 2018. URL: https://www.trimbos.nl/aanbod/webwinkel/af1643-jaarbericht-nationale-drug-monitor-2018/ [accessed 2024-01-23]
29. Siemer L, Pieterse ME, Brusse-Keizer MG, Postel MG, Ben Allouch S, Sanderman R. Study protocol for a non-inferiority trial of a blended smoking cessation treatment versus face-to-face treatment (LiveSmokefree-Study). BMC Public Health 2016 Nov 24;16(1):1187 [FREE Full text] [doi: 10.1186/s12889-016-3851-x] [Medline: 27881108]
30. Mudde AN, Willemsen MC, Kremers SP, de Vries H. Meetinstrumenten Voor Onderzoek Naar Stoppen Met Roken. The Hague, The Netherlands: Stivoro; Jan 1, 2000.
31. Richtlijn behandeling van tabaksverslaving herziening 2009. Partnership Stop met Roken. 2009. URL: https://www.nvvp.net/stream/richtlijn-behandeling-van-tabaksverslaving-2009.pdf [accessed 2024-01-23]
32. Willems E. Zorgmodule stoppen met roken. Partnership Stop met Roken. 2009. URL: https://www.zorginzicht.nl/binaries/content/assets/zorginzicht/kwaliteitsinstrumenten/ZorgmoduleStoppenmetRoken.pdf [accessed 2024-01-23]
33. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
34. Siemer L, Brusse-Keizer MG, Postel MG, Ben Allouch S, Patrinopoulos Bougioukas A, Sanderman R, et al. Blended smoking cessation treatment: exploring measurement, levels, and predictors of adherence. J Med Internet Res 2018 Aug 01;20(8):e246 [FREE Full text] [doi: 10.2196/jmir.9969] [Medline: 30068503]
35. Siemer L, Brusse-Keizer MG, Postel MG, Ben Allouch S, Sanderman R, Pieterse ME. Adherence to blended or face-to-face smoking cessation treatment and predictors of adherence: randomized controlled trial. J Med Internet Res 2020 Jul 23;22(7):e17207 [FREE Full text] [doi: 10.2196/17207] [Medline: 32459643]
36. Benowitz NL, Bernert JT, Foulds J, Hecht SS, Jacob P, Jarvis MJ, et al. Biochemical verification of tobacco use and abstinence: 2019 update. Nicotine Tob Res 2020 Jun 12;22(7):1086-1097 [FREE Full text] [doi: 10.1093/ntr/ntz132] [Medline: 31570931]
37. SRNT Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. Nicotine Tob Res 2002 May;4(2):149-159. [doi: 10.1080/14622200210123581] [Medline: 12028847]
38. Hughes JR, Keely JP, Niaura RS, Ossip-Klein DJ, Richmond RL, Swan GE. Measures of abstinence in clinical trials: issues and recommendations. Nicotine Tob Res 2003 Feb;5(1):13-25. [Medline: 12745503]
39. Jarvis MJ, Fidler J, Mindell J, Feyerabend C, West R. Assessing smoking status in children, adolescents and adults: cotinine cut-points revisited. Addiction 2008 Sep;103(9):1553-1561 [FREE Full text] [doi: 10.1111/j.1360-0443.2008.02297.x] [Medline: 18783507]
40. Cropsey KL, Trent LR, Clark CB, Stevens EN, Lahti AC, Hendricks PS. How low should you go? Determining the optimal cutoff for exhaled carbon monoxide to confirm smoking abstinence when using cotinine as reference. Nicotine Tob Res 2014 Oct;16(10):1348-1355 [FREE Full text] [doi: 10.1093/ntr/ntu085] [Medline: 24891552]
41. Christenhusz LC, Prenger R, Pieterse ME, Seydel ER, van der Palen J. Cost-effectiveness of an intensive smoking cessation intervention for COPD outpatients. Nicotine Tob Res 2012 Jun;14(6):657-663. [doi: 10.1093/ntr/ntr263] [Medline: 22180589]
42. van Deursen AJ, Courtois C, van Dijk JA. Internet skills, sources of support, and benefiting from internet use. Int J Hum Comput Interact 2014 Mar 07;30(4):278-290. [doi: 10.1080/10447318.2013.858458]

XSL•FO
RenderX

43. Lowry R. The confidence interval for the difference between two independent proportions 2023. VassarStats. URL: http://vassarstats.net/prop2_ind.html [accessed 2024-01-23]
44. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554. [doi: 10.1177/0038038509103208]
45. Holm M, Schiöler L, Andersson E, Forsberg B, Gislason T, Janson C, et al. Predictors of smoking cessation: a longitudinal study in a large cohort of smokers. Respir Med 2017 Nov;132:164-169 [FREE Full text] [doi: 10.1016/j.rmed.2017.10.013] [Medline: 29229092]
46. West R, Hajek P, Stead L, Stapleton J. Outcome criteria in smoking cessation trials: proposal for a common standard. Addiction 2005 Mar;100(3):299-303. [doi: 10.1111/j.1360-0443.2004.00995.x] [Medline: 15733243]
47. Blankers M, Smit ES, van der Pol P, de Vries H, Hoving C, van Laar M. The missing=smoking assumption: a fallacy in internet-based smoking cessation trials? Nicotine Tob Res 2016 Jan;18(1):25-33. [doi: 10.1093/ntr/ntv055] [Medline: 25744969]
48. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
49. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
50. Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the depression anxiety stress scales in clinical groups and a community sample. Psychol Assess 1998;10(2):176-181. [doi: 10.1037/1040-3590.10.2.176]
51. Metse AP, Stockings E, Bailey J, Regan T, Bartlem K, Wolfenden L, et al. Rates of retention of persons with a mental health disorder in outpatient smoking cessation and reduction trials, and associated factors: protocol for a systematic review and meta-analysis. BMJ Open 2019 Sep 04;9(9):e030646 [FREE Full text] [doi: 10.1136/bmjopen-2019-030646] [Medline: 31488491]
52. Crutzen R, Viechtbauer W, Spigt M, Kotz D. Differential attrition in health behaviour change trials: a systematic review and meta-analysis. Psychol Health 2015 Jan;30(1):122-134. [doi: 10.1080/08870446.2014.953526] [Medline: 25109224]
53. Waller G. Evidence-based treatment and therapist drift. Behav Res Ther 2009 Feb;47(2):119-127. [doi: 10.1016/j.brat.2008.10.018] [Medline: 19036354]
54. ChatGPT-4. OpenAI. URL: https://openai.com/gpt-4 [accessed 2024-01-26]
55. DeepL Translator. DeepL. URL: https://www.deepl.com/ [accessed 2024-01-26]
56. DeepL Write. DeepL. URL: https://www.deepl.com/write/ [accessed 2024-01-26]
57. Blended smoking cessation smoking database (all data). DANS Data Station Social Sciences and Humanities. 2022. URL: https://ssh.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/dans-xht-kn55 [accessed 2024-01-25]

Abbreviations

BSCT: blended smoking cessation treatment
CO: carbon monoxide
F2F: face-to-face
ITT: intention-to-treat
PROM: patient-reported outcome measure
RCT: randomized controlled trial

Edited by A Mavragani; submitted 06.03.23; peer-reviewed by L Harst, Z Ehtesham; comments to author 24.07.23; revised version received 04.11.23; accepted 29.12.23; published 20.02.24.

Please cite as:
Siemer L, Pieterse ME, Ben Allouch S, Postel MG, Brusse-Keizer MGJ
Comparing the Effectiveness of the Blended Delivery Mode With the Face-to-Face Delivery Mode of Smoking Cessation Treatment: Noninferiority Randomized Controlled Trial
J Med Internet Res 2024;26:e47040
URL: https://www.jmir.org/2024/1/e47040
doi:10.2196/47040
PMID:38376901

XSL•FO
RenderX

©Lutz Siemer, Marcel E Pieterse, Somaya Ben Allouch, Marloes G Postel, Marjolein G J Brusse-Keizer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Social Media–Based Mindfulness Psycho-Behavioral Intervention (MCARE) for Patients With Acute Coronary Syndrome: Randomized Controlled Trial

Huijing Zou[1], PhD; Sek Ying Chair[2], PhD; Bilong Feng[3], BSN; Qian Liu[1], PhD; Yu Jia Liu[1], BSN; Yu Xin Cheng[1], BSN; Dan Luo[1], PhD; Xiao Qin Wang[1], PhD; Wei Chen[1], BSN; Leiqing Huang[1], BSN; Yunyan Xianyu[4], PhD; Bing Xiang Yang[1], PhD

[1]School of Nursing, Wuhan University, Wuhan, China

[2]The Nethersole School of Nursing, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Zhongnan Hospital of Wuhan University, Wuhan, China

[4]Renmin Hospital of Wuhan University, Wuhan, China

Corresponding Author:
Bing Xiang Yang, PhD
School of Nursing
Wuhan University
No. 115 Donghu Road, Wuchang District
Wuhan, 430071
China
Phone: 86 02768788685
Email: yangbx@whu.edu.cn

Abstract

Background: Psychological distress is common among patients with acute coronary syndrome (ACS) and has considerable adverse impacts on disease progression and health outcomes. Mindfulness-based intervention is a promising complementary approach to address patients' psychological needs and promote holistic well-being.

Objective: This study aims to examine the effects of a social media–based mindfulness psycho-behavioral intervention (MCARE) on psychological distress, psychological stress, health-related quality of life (HRQoL), and cardiovascular risk factors among patients with ACS.

Methods: This study was a 2-arm, parallel-group randomized controlled trial. We recruited 178 patients (mean age 58.7, SD 8.9 years; 122/178, 68.5% male) with ACS at 2 tertiary hospitals in Jinan, China. Participants were randomly assigned to the MCARE group (n=89) or control group (n=89). The 6-week intervention consisted of 1 face-to-face session (phase I) and 5 weekly WeChat (Tencent Holdings Ltd)–delivered sessions (phase II) on mindfulness training and health education and lifestyle modification. The primary outcomes were depression and anxiety. Secondary outcomes included psychological stress, HRQoL, and cardiovascular risk factors (ie, smoking status, physical activity, dietary behavior, BMI, blood pressure, blood lipids, and blood glucose). Outcomes were measured at baseline (T0), immediately after the intervention (T1), and 12 weeks after the commencement of the intervention (T2).

Results: The MCARE group showed significantly greater reductions in depression (T1: β=–2.016, 95% CI –2.584 to –1.449, Cohen *d*=–1.28, *P*<.001; T2: β=–2.089, 95% CI –2.777 to –1.402, Cohen *d*=–1.12, *P*<.001) and anxiety (T1: β=–1.024, 95% CI –1.551 to –0.497, Cohen *d*=–0.83, *P*<.001; T2: β=–0.932, 95% CI –1.519 to –0.346, Cohen *d*=–0.70, *P*=.002). Significantly greater improvements were also observed in psychological stress (β=–1.186, 95% CI –1.678 to –0.694, Cohen *d*=–1.41, *P*<.001), physical HRQoL (β=0.088, 95% CI 0.008-0.167, Cohen *d*=0.72, *P*=.03), emotional HRQoL (β=0.294, 95% CI 0.169-0.419, Cohen *d*=0.81, *P*<.001), and general HRQoL (β=0.147, 95% CI 0.070-0.224, Cohen *d*=1.07) at T1, as well as dietary behavior (β=0.069, 95% CI 0.003-0.136, Cohen *d*=0.75, *P*=.04), physical activity level (β=177.542, 95% CI –39.073 to 316.011, Cohen *d*=0.51, *P*=.01), and systolic blood pressure (β=–3.326, 95% CI –5.928 to –0.725, Cohen *d*=–1.32, *P*=.01) at T2. The overall completion rate of the intervention (completing ≥5 sessions) was 76% (68/89). Positive responses to the questions of the acceptability questionnaire ranged from 93% (76/82) to 100% (82/82).

XSL•FO
RenderX

Conclusions: The MCARE program generated favorable effects on psychological distress, psychological stress, HRQoL, and several aspects of cardiovascular risk factors in patients with ACS. This study provides clues for guiding clinical practice in the recognition and management of psychological distress and integrating the intervention into routine rehabilitation practice.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000033526; https://www.chictr.org.cn/showprojEN.html?proj=54693

(J Med Internet Res 2024;26:e48557) doi:10.2196/48557

KEYWORDS

acute coronary syndrome; psychological distress; depression; anxiety; mindfulness; mindfulness-based intervention; quality of life; risk factors; cardiac rehabilitation; social media

Introduction

Acute coronary syndrome (ACS), an acute manifestation of ischemic heart disease, has become a major public health problem worldwide [1]. In China, ischemic heart disease affects approximately 11.4 million people [2], and there remains a rising trend in the morbidity and mortality of ACS [3], thus posing a huge challenge for the health care system.

Psychological distress, such as depression and anxiety, is highly prevalent in patients with ACS [4] and has considerable adverse impacts on disease progression and health outcomes. Clear evidence supports that psychological distress is associated with functional disability, reduced health-related quality of life (HRQoL), and increased risks of cardiac events [5,6]. Nonetheless, current health care practice has paid inadequate attention to the recognition and management of psychological distress. A growing consensus advocates that psychological distress is a crucial risk factor of ACS that should be addressed in disease management [5,7,8]. The American Heart Association has recommended mindfulness-based intervention as a promising complementary approach to promoting psychological health and well-being for patients with cardiovascular disease [8]. Emerging evidence has proven its benefits in improving a wide range of psychological and physical outcomes [9,10], including among patients with ischemic heart disease [11], indicating its potential as an additional supplement to conventional cardiac care.

Another practice gap in China is that optimal rehabilitation and effective control of cardiovascular risk factors are rarely achieved due to insufficient awareness and competency of health care professionals, inadequate workforce, and limited resources [12]. The health care system primarily focuses on in-hospital treatments of acute attacks of ACS and neglects the posthospital management of risk factors. An investigation of 991 hospitals in China showed that only 228 (23%) hospitals provided center-based cardiac rehabilitation services, which were mainly distributed in urban areas (89.1%) [2].

In recent years, mobile health or eHealth technologies are increasingly being used to improve the availability, feasibility, and affordability of posthospital care with inspiring results in promoting medication adherence, lifestyle changes, and health outcomes [13,14]. With the popularization and widespread coverage of mobile internet access, WeChat (Tencent Holdings Ltd), a free smartphone app, has become one of the most popular and widely used social media in China. It provides various services, including instant messaging, voice and video calls, microblogging and subscription services, and web-based banking. WeChat may provide an unprecedented approach to addressing the shortage of posthospital care, considering its wide population reach and powerful peripheral functions.

This study proposed a WeChat-based mindfulness psycho-behavioral intervention (MCARE), which integrated mindfulness training with health education and lifestyle modification to assist patients in managing risk factors. This randomized controlled trial (RCT) aimed to examine the effects of the MCARE program on psychological distress (primary outcomes), psychological stress, HRQoL, and cardiovascular risk factors (secondary outcomes) among patients with ACS.

Methods

Study Design and Participants

This study was a 2-arm, 1:1 parallel-group RCT. Participants were recruited using convenience sampling from June to September 2020 at the wards of the cardiology department of 2 public tertiary hospitals in Jinan, China. Inclusion criteria were as follows: (1) age 18 to 75 years; (2) clinical diagnosis of ACS (including unstable angina and acute myocardial infarction); (3) ability to read, understand, communicate, and complete questionnaires in Chinese; and (4) possession of an operational smartphone and an active WeChat account. Participants were excluded if they (1) were in the active state of myocardial infarction or receiving open-heart surgical treatment; (2) had a clinical diagnosis of serious physical comorbidities, for example, cancer and renal failure; (3) had psychiatric disorders; (4) had cognitive impairments as documented in the health records; or (5) were currently participating in other interventions.

The sample size was calculated following a power analysis approach using G*Power 3.1 (Universität Düsseldorf). Considering a significance level of 0.05; a statistical power of 80%; an effect size of Cohen d=0.50 for primary outcomes, namely depression and anxiety [15]; and a potential attrition rate of 20%, in all,160 participants (80 per group) were required.

Ethical Considerations

Ethics approvals were obtained from the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee (2019.323) and the Ethics Committee of the School of Nursing, Shandong University (2019-R-017). Participants gave informed consent to participate in the study before taking part. All information and data from participants were anonymous and confidentially was guaranteed

XSL•FO
RenderX

by coding participants with unique identification numbers (eg, 001). All participants received a small red envelope sent via WeChat upon completion of each intervention session and after each data collection to acknowledge the time and effort they dedicated to participating in the study.

Procedure

All procedures of the RCT were completed in the same manner in 2 research hospitals. Two research assistants (registered nurses) were hired and received training on participant recruitment; data collection; and the rationale, principles, and procedure of the study. The research assistants approached eligible participants, introduced the study details, and invited them to participate in this study. Then they obtained written informed consent and performed a baseline assessment.

Participants were randomly allocated into either the intervention or control groups at a 1:1 ratio using a sequence of block randomized numbers generated by an independent statistician using a computer procedure with a block size of 4. The group allocation was concealed by placing the random sequence in a sequentially numbered, opaque, and sealed envelope by an independent person. The research assistants were blinded to group allocation.

Interventions

The MCARE program was developed on the basis of the Common-Sense Model of Self-Regulation theory [16], clinical guideline recommendations [7,17], and experimental studies [8,10,11]. It aimed to promote psychological and physical well-being by targeting emotional and cognitive procedures for managing health threats. Figure 1 illustrates the conceptual framework underpinning the MCARE program.

Figure 1. The conceptual framework underpinning the mindfulness psycho-behavioral intervention program. ACS: acute coronary syndrome.

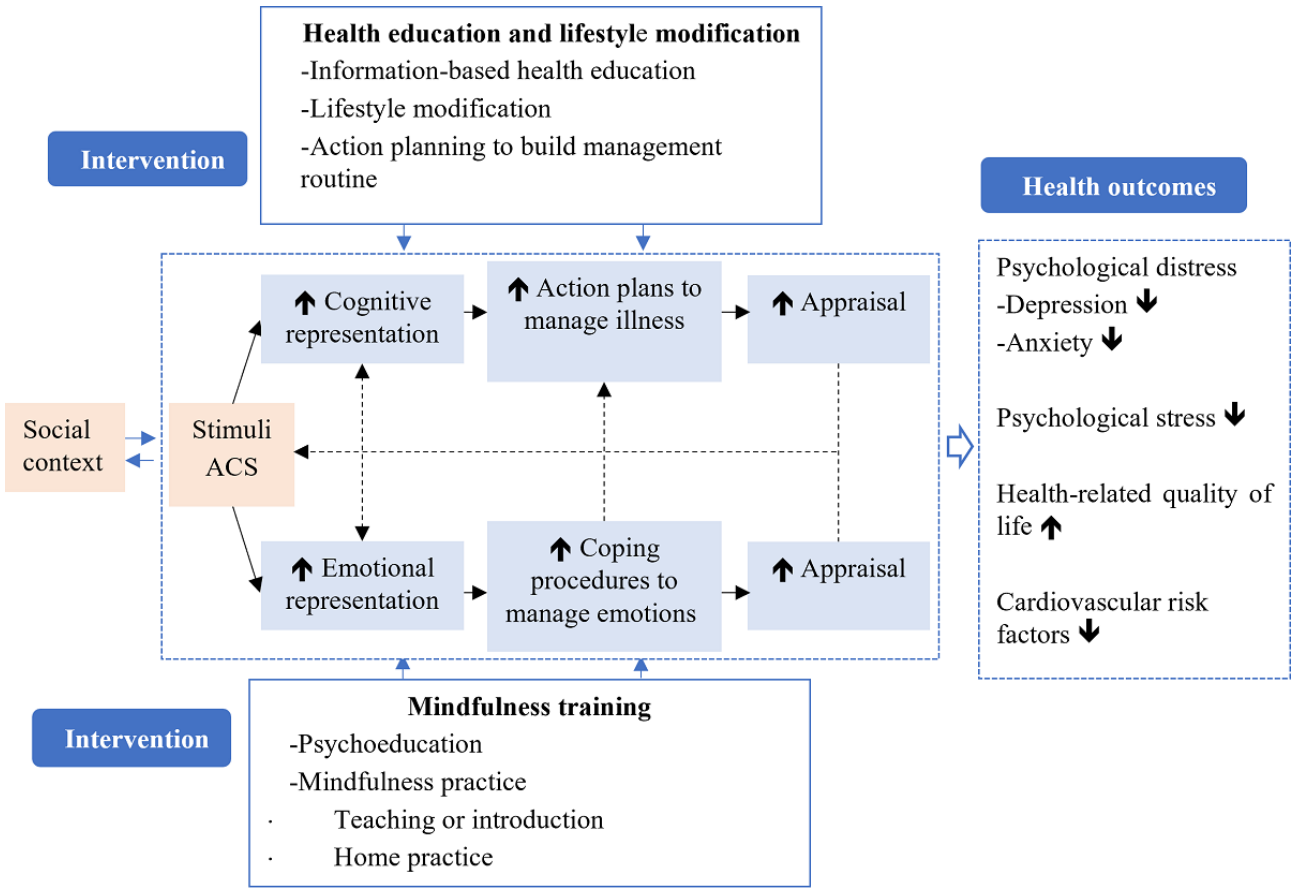

The MCARE program comprised 6 weekly sessions and each session focused on a thematic topic in mindfulness training and disease management (Table S1 in Multimedia Appendix 1). Mindfulness training topics included simple awareness, mindfulness of breath, mindfulness of the body, dealing with thoughts, dealing with difficulties, and maintenance. Disease management topics were basic disease information, healthy dietary behavior, physical activity, body weight control, smoking cessation, and management of metabolic risk factors. Additionally, participants were required to perform home-based mindfulness practice for 10 to 20 minutes per day for 6 days per week. Figure 2 summarizes the detailed delivery plan. The first session was delivered face-to-face during hospitalization (phase I) and the following 5 sessions were delivered using WeChat after discharge (phase II). Following the third and sixth sessions, participants received a voice call to monitor and review their performance, address barriers or problems, and provide support and encouragement. Participants also received an information handbook and audio and video guides on mindfulness practice.

XSL•FO
RenderX

Figure 2. The delivery plan of the mindfulness psycho-behavioral intervention program.

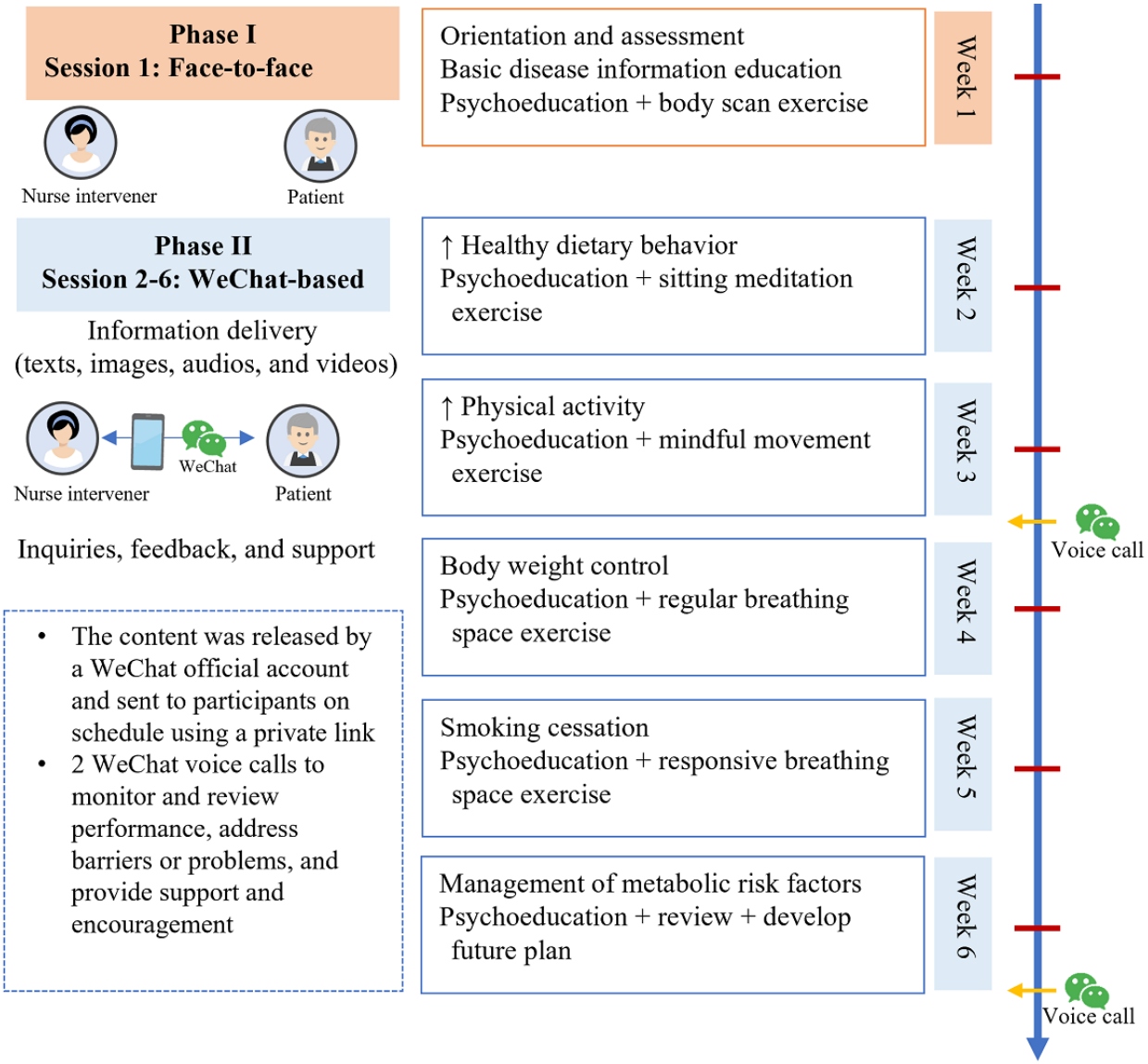

Participants in the control group received routine medical treatment and care before hospital discharge. To control the nonspecific effect of attention, the control group also received WeChat contacts at the same frequency as the intervention group. The contents of the contacts were limited to general information about the importance of managing stress and risk factors without providing specific advice and strategies.

Measures

The research assistants collected data at baseline (T0), immediately after the intervention (T1), and 12 weeks following the commencement of the intervention (T2). A self-designed data collection sheet was used to assess baseline sociodemographic and clinical characteristics through patient interviews and medical record reviews.

Primary outcomes were psychological distress, including depression and anxiety, which were assessed using the 9-item Patient Health Questionnaire (PHQ-9) [18] and the 7-item Generalized Anxiety Disorder (GAD-7) [19], respectively. Secondary outcomes were psychological stress, HRQoL, and cardiovascular risk factors. Psychological stress was assessed using the 10-item Perceived Stress Scale [20]. HRQoL was measured using the HeartQoL questionnaire [21], which comprises 14 items capturing disease-specific HRQoL in physical (10 items) and emotional (4 items) dimensions. Cardiovascular risk factors included (1) smoking status, as measured by self-reported 7-day smoking history [22]; (2) physical activity, as measured using the International Physical Activity Questionnaire-Short Form [23]; (3) dietary behavior, as assessed using the nutrition subscale of the Health-Promoting Lifestyle Profile-II [24]; (4) BMI, as calculated using the formula: *BMI = body weight (in kilograms) / height (in meters) squared*; (5) blood pressure (BP); (6) blood lipid profiles; and (7) blood glucose. Body weight, height, and BP were obtained by anthropometric measures, and blood lipids and blood glucose were measured via laboratory tests of fasting blood samples.

The acceptability of the MCARE program among participants in the MCARE group was also measured at T1 using a self-developed dichotomous questionnaire (positive ratings ≥80% are considered acceptable). Additionally, the completion of the intervention, performance of home mindfulness practice,

XSL•FO
RenderX

difficulties or problems encountered by participants, and adverse events were collected.

Baseline (T0) assessment was conducted at the wards, and follow-up assessments (T1 and T2) were completed via telephone interviews. After the first face-to-face session, the assessors would schedule the telephone interview for each participant. Participants were reminded to return to the research hospitals or go to a nearby accredited hospital to complete fasting blood tests at T2. If the participants did not answer the telephone call, they would contact the participant 3 times at different periods of a day within 1 week. If none of these telephone calls reached the participant, the participant would be considered lost to follow-up.

Statistical Analysis

Statistical analyses were performed using IBM SPSS (version 25.0). All statistical tests were 2-tailed tests and statistical significance was set at 0.05. Appropriate descriptive statistics were calculated to summarize the participant characteristics and outcomes. The intention-to-treat principle was applied in outcome analysis. The generalized estimating equation (GEE) analyses were performed to examine the differential changes of each outcome variable across 3 data collection time points between intervention and control groups. Baseline characteristics and outcomes between intervention and control groups were compared and no significant differences were observed, therefore only the crude GEE models were performed without adjustment of confounding variables. A dummy variable (group) was set to represent the MCARE group with the control group as the reference. To represent time differences, the baseline (T0) was set as the reference, and another 2 dummy variables, T1 and T2, were assigned to correspond to immediate postintervention and 12 weeks. The interaction terms of the group-by-time dummy variables, group×T1 and group×T2, were included in the GEE models to assess the overall differences in the outcomes between the 2 groups at T1 and T2. Effect sizes were estimated using Cohen *d* statistic for continuous outcomes and odds ratio for binary outcomes.

We calculated the percentage of missing data for each outcome (9.0% to 9.9%) and compared baseline sociodemographic and clinical characteristics between participants who had completed all observations and those who had at least 1 missing observation. Additionally, to examine the effects of the missing data on the estimation of intervention effects we conducted a sensitivity analysis using completed case analysis. In the completed case analysis, only participants who had completed all assessments at T0, T1, and T2 were included in the estimation of intervention effects. We also examined the correlation between participants' completion of intervention and dosage of home mindfulness practice on changes in outcome variables between T1 and T0 by performing Pearson correlation analysis.

Results

Participant Recruitment and Retention

Of the 275 patients with ACS screened for eligibility, 52 patients did not meet the eligibility criteria and 45 patients declined to participate. Finally, 178 participants were enrolled (Figure 3). At follow-up, 157 (88.2%) participants completed T1 assessment, and 146 (82.0%) participants completed T2 assessment. Cardiovascular risk factors were measured only at T0 and T2 with a total of 356 observations, and 32 (9.0%) were missing. All the other outcome variables had a total of 534 observations across 3 study time points with 53 (9.9%) missing observations. There was no significant difference in baseline characteristics between participants who completed all observations (n=146) and who had at least 1 missing observation (n=32; Table S2 in Multimedia Appendix 1). No adverse events related to the intervention were reported.

XSL•FO
RenderX

Figure 3. Flow chart of participant recruitment, allocation, intervention delivery, follow-up, and data analysis of the study.

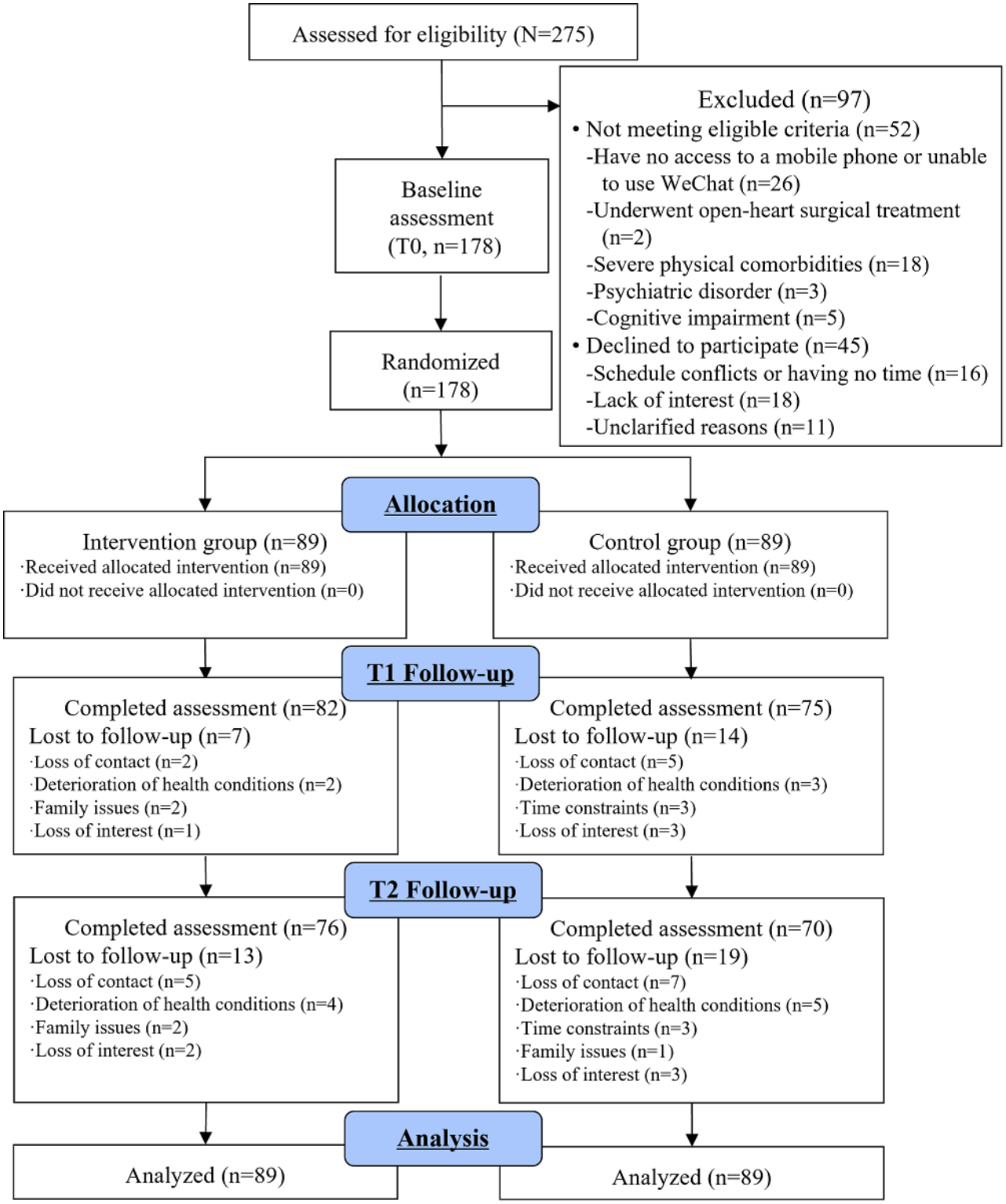

Baseline Characteristics

Table 1 presents the detailed baseline characteristics of participants and the intervention and control groups were well-matched. Mean age of the participants was 58.7 (SD 8.9) years, ranging from 28 to 75 years. The majority of them were male (122/178, 68.5%), married (172/178, 96.6%), received a secondary education or less (147/178, 82.6%), and had a New York Heart Association class of I or II (148/178, 83.1%). Over half of the participants were experiencing ACS for the first time (104/178, 58.4%) and did not receive percutaneous transluminal coronary intervention (99/178, 55.6%). The mean scores for PHQ-9 and GAD-7 were 5.66 (SD 3.30) and 5.38 (SD 2.97), respectively. Over half of the participants had depressive symptoms (PHQ-9 score ≥5; 107/178, 60.1%) and anxiety symptoms (GAD-7 score ≥5; 95/178, 53.4%), indicating psychological distress is highly prevalent.

XSL•FO
RenderX

Table 1. Baseline characteristics and outcomes of all participants and by the mindfulness psycho-behavioral intervention.

Characteristics	All (n=178)	Intervention group (n=89)	Control group (n=89)	*P* value
Age (years), mean (SD)	58.73 (8.90)	59.48 (8.95)	58.02 (8.90)	.28
Sex, n (%)				.52
Male	122 (68.5)	59 (66.3)	63 (70.8)	
Female	56 (31.5)	30 (33.7)	26 (29.2)	
Marital status, n (%)				.41
Married	172 (96.6)	87 (97.8)	85 (95.5)	
Single, separated, divorced, or widowed	6 (3.4)	2 (2.2)	4 (4.5)	
Educational level, n (%)				.49
Primary education	34 (19.1)	20 (22.5)	14 (15.7)	
Secondary education	113 (63.5)	55 (61.8)	58 (65.2)	
Tertiary education	31 (17.4)	14 (15.7)	17 (19.1)	
Employment status, n (%)				.35
Employed	66 (37.1)	30 (33.7)	36 (40.4)	
Unemployed, farmer, retired, or freelance	112 (62.9)	59 (66.3)	53 (59.6)	
Monthly income (CNY ¥[a]), n (%)				.75
<1500	27 (15.2)	12 (13.5)	15 (16.8)	
1500-3000	68 (38.2)	36 (40.4)	32 (36)	
≥3000	83 (46.6)	41 (46.1)	42 (47.2)	
Episode of ACS[b], n (%)				.13
First	104 (58.4)	57 (64)	47 (52.8)	
Recurrent	74 (41.6)	32 (36)	42 (47.2)	
Percutaneous transluminal coronary intervention, n (%)				.05
Yes	79 (44.4)	33 (37.1)	46 (51.7)	
No	99 (55.6)	56 (62.9)	43 (48.3)	
NYHA[c] functional classification, n (%)				.37
I	83 (46.6)	45 (50.6)	38 (42.7)	
II	65 (36.5)	30 (33.7)	35 (39.3)	
III	21 (11.8)	8 (9.0)	13 (14.6)	
IV	9 (5.1)	6 (6.7)	3 (3.4)	
LVEF[d] (%), mean (SD)	56.61 (10.69)	57.00 (8.66)	56.21 (10.57)	.59
Number of comorbidities, mean (SD)	1.13 (0.87)	1.06 (0.82)	1.20 (0.92)	.26
Depression (PHQ-9[e]), mean (SD)	5.66 (3.30)	5.37 (3.26)	5.96 (3.34)	.24
Anxiety (GAD-7[f]), mean (SD)	38 (2.97)	5.25 (2.96)	5.48 (2.93)	.59
Psychological stress (PSS-10[g]), mean (SD)	16.24 (4.54)	16.35 (4.66)	16.12 (4.45)	.74
HRQoL[h] (HeartQoL), mean (SD)				
Physical HRQoL	1.89 (0.61)	1.91 (0.63)	1.87 (0.58)	.67
Emotional HRQoL	2.18 (0.70)	2.19 (0.74)	2.17 (0.65)	.92
General HRQoL	1.98 (0.56)	1.99 (0.58)	1.96 (0.54)	.70
Smoking status, n (%)				.86
Nonsmoker	133 (74.7)	66 (74.2)	67 (75.3)	

XSL•FO
RenderX

Characteristics	All (n=178)	Intervention group (n=89)	Control group (n=89)	*P* value
Current smoker	45 (25.3)	23 (25.8)	22 (24.7)	
Physical activity (IPAQ-SF[i]) (MET·min/wk), mean (SD)	712.78 (665.45)	727.85 (686.95)	697.70 (646.78)	.76
Dietary behavior (HPLP-II[j] nutrition subscale), mean (SD)	2.43 (0.37)	2.46 (0.38)	2.40 (0.35)	.26
BMI (kg/m^2), mean (SD)	25.10 (3.72)	24.98 (3.66)	25.10 (3.20)	.81
Systolic BP[k] (mm Hg), mean (SD)	126.06 (17.29)	126.64 (18.08)	125.48 (16.55)	.66
Diastolic BP (mm Hg), mean (SD)	75.92 (10.24)	75.47 (10.03)	76.36 (10.48)	.56
LDL-C[l] (mmol/L), mean (SD)	2.24 (0.82)	2.22 (0.82)	2.26 (0.82)	.76
HDL-C[m] (mmol/L), median (IQR)	1.09 (0.91, 1.30)	1.00 (0.90, 1.25)	1.12 (0.92, 1.37)	.20
TG[n] (mmol/L), median (IQR)	1.36 (1.00-1.99)	1.36 (0.98-2.00)	1.36 (1.02-2.04)	.75
TC[o] (mmol/L), mean (SD)	4.00 (1.07)	3.90 (1.02)	4.10 (1.12)	.22
FBG[p] (mmol/L), median (IQR)	5.47 (4.74-6.51)	5.56 (4.89-6.93)	5.22 (4.69-6.28)	.31

[a]CNY ¥: Chinese yuan, US $1=CNY ¥6.90 at the time of the study (2020).
[b]ACS: acute coronary syndrome.
[c]NYHA: New York Heart Association.
[d]LVEF: left ventricular ejection fraction.
[e]PHQ-9: 9-item Patient Health Questionnaire.
[f]GAD-7: 7-item Generalized Anxiety Disorder.
[g]PSS-10: 10-item Perceived Stress Scale.
[h]HRQoL: health-related quality of life.
[i]IPAQ-SF: International Physical Activity Questionnaire-Short Form.
[j]HPLP-II: Health-Promoting Lifestyle Profile-II.
[k]BP: blood pressure.
[l]LDL-C: low-density lipoprotein cholesterol.
[m]HDL-C: high-density lipoprotein cholesterol.
[n]TG: triglyceride.
[o]TC: total cholesterol.
[p]FBG: fasting blood glucose.

Intervention Effects

There were significant time-by-group interaction effects on psychological distress with moderate to large effects at both T1 and T2 (Table 2). Participants in the intervention group demonstrated significantly greater reductions in depression (T1: β=–2.016, 95% CI –2.584 to –1.449, Cohen *d*=–1.28, *P*<.001; T2: β=–2.089, 95% CI –2.777 to –1.402, Cohen *d*=–1.12, *P*<.001) and anxiety (T1: β=–1.024, 95% CI –1.551 to –0.497, Cohen *d*=–0.83, *P*<.001; T2: β=–0.932, 95% CI –1.519 to –0.346, Cohen *d*=–0.70, *P*=.002) than those in the control group.

XSL•FO
RenderX

Table 2. Generalized estimating equation analyses for the comparison of depression and anxiety of the mindfulness psycho-behavioral intervention program[a].

Outcome and time point	Intervention group	Control group	Group effect[b]		Time effect[c]		Group×time effect[d]		
			β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	Effect size	*P* value
Depression (PHQ-9[e], total score range 0-27)[f]									
T0	5.37 (3.26)	5.96 (3.34)	–0.584 (–1.548 to 0.379)	.24	N/A[g]	N/A	N/A	N/A	N/A
T1	3.72 (3.03)	6.32 (3.19)	N/A	N/A	0.365 (–0.060 to 0.790)	.09	–2.016 (–2.584 to –1.449)	–1.28	<.001
T2	2.53 (2.39)	5.20 (3.02)	N/A	N/A	–0.755 (–1.255 to –0.256)	.003	–2.089 (–2.777 to –1.402)	–1.12	<.001
Anxiety (GAD-7[h], total score range 0-21)[f]									
T0	5.25 (2.96)	5.48 (2.93)	–0.236 (–1.096 to 0.625)	.59	N/A	N/A	N/A	N/A	N/A
T1	4.07 (2.38)	5.33 (2.71)	N/A	N/A	–0.150 (–0.555 to 0.255)	.47	–1.024 (–1.551 to –0.497)	–0.83	<.001
T2	1.63 (2.03)	2.80 (2.45)	N/A	N/A	–2.683 (–3.121 to –2.245)	<.001	–0.932 (–1.519 to –0.346)	–0.70	.002

[a]The control group (group =0) and the baseline measurement (time =0) were set as the reference categories in the generalized estimating equation model and its corresponding null variables.

[b]Group effect was defined as group differences at baseline between intervention and control groups.

[c]Time effect at T1 is defined as change of scores for the control group at T1 compared with T0; T2 is defined as change of scores for the control group at T2 compared with T0.

[d]Group×time effect at T1 defined as additional change of scores for the intervention group compared with the control group at T1; T2 defined as additional change of scores for the intervention group compared with the control group at T2. Effect sizes were estimated using Cohen *d* statistic for continuous outcomes and odds ratio for binary outcomes.

[e]PHQ-9: 9-item Patient Health Questionnaire.

[f]Intervention and control group data are presented as mean (SD).

[g]N/A: not applicable.

[h]GAD-7: 7-item General Anxiety Disorder.

Compared with control group, the intervention group demonstrated significantly greater improvements in psychological stress (β=–1.186, 95% CI –1.678 to –0.694, Cohen *d*=–1.41, *P*<.001), physical HRQoL (β=0.088, 95% CI 0.008-0.167, Cohen *d*=0.72, *P*=.03), emotional HRQoL (β=0.294, 95% CI 0.169-0.419, Cohen *d*=0.81, *P*<.001), and general HRQoL (β=0.147, 95% CI 0.070-0.224, Cohen *d*=1.07, *P*<.001) at T1 (Table 3). However, the significant effects were only sustained for psychological stress (β=–1.268, 95% CI –1.992 to –0.544, Cohen *d*=–1.17, *P*=.001) and emotional HRQoL (β=0.249, 95% CI 0.102-0.395, Cohen *d*=0.62, *P*=.001) but not for physical HRQoL and general HRQoL at T2. For cardiovascular risk factors, the intervention group showed significantly greater improvements in dietary behavior (β=0.069, 95% CI 0.003-0.136, Cohen *d*=0.75, *P*=.04), physical activity level (β=177.542, 95% CI –39.073 to 316.011, Cohen *d*=0.51, *P*=.01), and systolic BP (β=–3.326, 95% CI –5.928 to –0.725, Cohen *d*=–1.32, *P*=.01) at T2 (Table 4). No significant group-by-time interaction effects were observed for other outcomes (all *P*>.05). The sensitivity analysis showed consistent results in the directions of the GEE regression coefficients (Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 3. Generalized estimating equation analyses for the comparison of psychological stress and health-related quality of life of the mindfulness psycho-behavioral intervention program[a].

Outcome and time point	Intervention group	Control group	Group effect[b]		Time effect[c]		Group×time effect[d]		
			β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	Effect size	*P* value
Psychological stress (PSS-10[e], total score range 0-40)[f]									
T0	16.35 (4.66)	16.12 (4.45)	0.225 (–1.106 to 1.555)	.74	N/A[g]	N/A	N/A	N/A	N/A
T1	14.59 (4.71)	15.55 (4.59)	N/A	N/A	–0.577 (–0.950 to –0.204)	.002	–1.186 (–1.678 to –0.694)	–1.41	<.001
T2	14.00 (4.52)	15.04 (4.60)	N/A	N/A	–1.081 (–1.538 to –0.624)	<.001	–1.268 (–1.992 to –0.544)	–1.17	.001
Physical HRQoL (Physical HeartQoL[h], subscale score range 0-30)[f]									
T0	1.91 (0.63)	1.87 (0.58)	0.039 (–0.139 to 0.217)	.66	N/A	N/A	N/A	N/A	N/A
T1	2.11 (0.62)	1.98 (0.58)	N/A	N/A	0.110 (0.048 to 0.172)	<.001	0.088 (0.008 to 0.167)	0.72	.03
T2	2.13 (0.61)	2.09 (0.58)	N/A	N/A	0.213 (0.131 to 0.295)	<.001	0.004 (–0.099 to 0.107)	0.12	.94
Emotional HRQoL (Emotional HeartQoL, subscale score range 0-12)[f]									
T0	2.19 (0.74)	2.17 (0.65)	0.011 (–0.193 to 0.215)	.91	N/A	N/A	N/A	N/A	N/A
T1	2.52 (0.63)	2.21 (0.61)	N/A	N/A	0.036 (–0.062 to 0.134)	.47	0.294 (0.169 to 0.419)	0.81	<.001
T2	2.55 (0.63)	2.29 (0.55)	N/A	N/A	0.115 (–0.001 to 0.232)	.05	0.249 (0.102 to 0.395)	0.62	.001
General HRQoL (General HeartQoL, total score range 0-42)[f]									
T0	1.99 (0.58)	1.96 (0.54)	0.031 (–0.133 to 0.195)	.71	N/A	N/A	N/A	N/A	N/A
T1	2.23 (0.54)	2.05 (0.51)	N/A	N/A	0.089 (0.029 to 0.148)	.003	0.147 (0.070 to 0.224)	1.07	<.001
T2	2.25 (0.53)	2.15 (0.48)	N/A	N/A	0.185 (0.111 to 0.260)	<.001	0.073 (–0.021 to 0.167)	0.47	.13

[a]The control group (group =0) and the baseline measurement (time =0) were set as the reference categories in the generalized estimating equation model and its corresponding null variables.

[b]Group effect was defined as group differences at baseline between intervention and control groups.

[c]Time effect at T1 is defined as change of scores for the control group at T1 compared with T0; T2 is defined as change of scores for the control group at T2 compared with T0.

[d]Group×time effect at T1 defined as additional change of scores for the intervention group compared with the control group at T1; T2 defined as additional change of scores for the intervention group compared with the control group at T2. Effect sizes were estimated using Cohen *d* statistic for continuous outcomes and odds ratio for binary outcomes.

[e]PSS-10: 10-item Perceived Stress Scale.

[f]Intervention and control group data are presented as mean (SD).

[g]N/A: not applicable.

[h]HRQoL: health-related quality of life.

XSL•FO
RenderX

Table 4. Generalized estimating equation analyses for the comparison of cardiovascular risk factors of the mindfulness psycho-behavioral intervention program[a].

Outcome and time point	Intervention group	Control group	Group effect[b]		Time effect[c]		Group×time effect[d]		
			β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	Effect size	*P* value
Smoking status (current smoker)[e]									
T0	23 (25.8)	22 (24.7)	0.059 (–0.617 to 0.736)	.86	N/A[f]	N/A	N/A	N/A	N/A
T2	6 (7.9)	9 (12.9)	N/A	N/A	–0.800 (–1.358 to –0.242)	.005	–0.603 (–1.532 to 0.327)	0.55	.20
Physical activity (IPAQ-SF[g], MET•min/week)[h]									
T0	727.85 (686.95)	697.70 (646.78)	30.152 (–164.766 to 225.069)	.76	N/A	N/A	N/A	N/A	N/A
T2	881.41 (605.83)	673.71 (541.13)	N/A	N/A	–23.988 (–127.430 to 79.454)	.65	177.542 (39.073 to 316.011)	0.51	.01
Dietary behavior (HPLP-II[i] nutrition subscale, total score range 9-36)[h]									
T0	2.46 (0.38)	2.40 (0.35)	0.061 (–0.045 to 0.168)	.26	N/A	N/A	N/A	N/A	N/A
T2	2.88 (0.36)	2.75 (0.34)	N/A	N/A	0.353 (0.309 to 0.397)	<.001	0.069 (0.003 to 0.136)	0.75	.04
BMI (kg/m^2)[h]									
T0	24.98 (3.66)	25.10 (3.20)	–0.121 (–1.126 to 0.883)	.81	N/A	N/A	N/A	N/A	N/A
T2	24.38 (3.59)	24.90 (3.11)	N/A	N/A	–0.203 (–0.557 to 0.150)	.26	–0.403 (–0.835 to 0.030)	–0.46	.07
Systolic BP[j] (mmHg)[h]									
T0	126.64 (18.08)	125.48 (16.55)	1.157 (–3.906 to 6.221)	.65	N/A	N/A	N/A	N/A	N/A
T2	121.97 (14.44)	124.14 (16.37)	N/A	N/A	–1.340 (–3.073 to 0.393)	.13	–3.326 (–5.928 to –0.725)	–1.32	.01
Diastolic BP (mmHg)[h]									
T0	75.47 (10.03)	76.36 (10.48)	–0.888 (–3.884 to 2.108)	.56	N/A	N/A	N/A	N/A	N/A
T2	73.12 (9.29)	75.10 (9.24)	N/A	N/A	–1.260 (–2.546 to 0.027)	.06	–1.094 (–2.772 to 0.584)	–0.67	.20
LDL-C[k] (mmol/L)[h]									
T0	2.22 (0.82)	2.26 (0.82)	–0.037 (–0.276 to 0.202)	.76	N/A	N/A	N/A	N/A	N/A
T2	2.14 (0.75)	2.16 (0.79)	N/A	N/A	–0.100 (–0.199 to –0.001)	.05	0.025 (–0.112 to 0.162)	0.07	.72

XSL•FO
RenderX

Outcome and time point	Intervention group	Control group	Group effect[b]		Time effect[c]		Group×time effect[d]		
			β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	Effect size	*P* value
HDL-C[l] (mmol/L)[m]									
T0	1.00 (0.90-1.25)	1.12 (0.92-1.37)	–0.121(–0.247 to 0.006)	.06	N/A	N/A	N/A	N/A	N/A
T2	1.18 (0.95-1.37)	1.15 (1.00-1.45)	N/A	N/A	0.043 (–0.025 to 0.112)	.21	0.043 (–0.034 to 0.121)	–0.04	.27
TG[n] (mmol/L)[m]									
T0	1.36 (0.98-2.00)	1.36 (1.02, 2.04)	–0.051(–0.411 to 0.309)	.78	N/A	N/A	N/A	N/A	N/A
T2	1.26 (0.87-1.96)	1.25 (0.99-1.90)	N/A	N/A	–0.104 (–0.245 to 0.037)	.15	–0.015 (–0.185 to 0.155)	–0.27	.86
TC[o] (mmol/L)[h]									
T0	3.90 (1.02)	4.10 (1.12)	–0.195 (–0.508 to 0.117)	.22	N/A	N/A	N/A	N/A	N/A
T2	3.96 (0.92)	4.07 (0.93)	N/A	N/A	–0.025 (–0.157 to 0.107)	.71	0.080 (–0.096 to 0.257)	0.15	.37
FBG[p] (mmol/L)[m]									
T0	5.56 (4.89-6.93)	5.22 (4.69-6.28)	0.246 (–0.292 to 0.784)	.37	N/A	N/A	N/A	N/A	N/A
T2	5.35 (4.80-6.49)	5.21 (4.61-6.27)	N/A	N/A	–0.169 (–0.327 to –0.012)	.04	–0.101 (–0.363 to 0.160)	–0.19	.45

[a]The control group (group =0) and the baseline measurement (time =0) were set as the reference categories in the generalized estimating equation model and its corresponding null variables.

[b]Group effect was defined as group differences at baseline between intervention and control groups.

[c]Time effect at T1 is defined as change of scores for the control group at T1 compared with T0; T2 is defined as change of scores for the control group at T2 compared with T0.

[d]Group×time effect at T1 defined as additional change of scores for the intervention group compared with the control group at T1; T2 defined as additional change of scores for the intervention group compared with the control group at T2. Effect sizes were estimated using Cohen *d* statistic for continuous outcomes and odds ratio for binary outcomes.

[e]Intervention and control group data are presented as n (%).

[f]N/A: not applicable.

[g]IPAQ-SF: International Physical Activity Questionnaire-Short Form.

[h]Intervention and control group data are presented as mean (SD).

[i]HPLP-II: health-promoting lifestyle profile-II.

[j]BP: blood pressure.

[k]LDL-C: low-density lipoprotein cholesterol.

[l]HDL-C: high-density lipoprotein cholesterol.

[m]Intervention and control group data are presented as median (IQR).

[n]TG: triglyceride.

[o]TC: total cholesterol.

[p]FBG: fasting blood glucose.

Intervention Adherence

The overall attrition rate of this study was 18% (32/178; intervention group: 13/89, 15% and control group: 19/89, 21%). The completion rate for each intervention session ranged from 68.5% (61/89) to 100% (89/89) and the overall completion rate (defined as completing at least 5 of the 6 intervention sessions) was 76% (68/89). Further analysis showed the number of completed sessions was significantly and positively correlated with changes in depression (r=0.324, P=.003), psychological

XSL•FO
RenderX

stress (r=0.224, P=.04), emotional HRQoL (r=0.224, P=.04), and general HRQoL (r=0.279, P=.01) at T1 (Table S4 in Multimedia Appendix 1), suggesting that adherence to the intervention may influence the intervention effects. The average frequency of home mindfulness practice ranged from 2.2 (SD 2.0) to 3.7 (SD 1.7) times per week and the total average amount was 19.0 (SD 8.9; range 1 to 38) times during the 6 weeks, which was much less than the designed dosage. In addition, the frequency of home mindfulness practice was significantly and positively correlated with changes in depression (r=0.865, P<.001), anxiety (r=0.626, P<.001), psychological stress (r=0.353, P=.001), emotional HRQoL (r=0.497, P<.001), and general HRQoL (r=0.399, P<.001) at T1 (Table S4 in Multimedia Appendix 1).

Acceptability of the Intervention

At T1, a total of 82 (92%) of 89 participants in the intervention group completed the acceptability questionnaire. Positive responses to the questions ranged from 93% (76/82) to 100% (82/82; Table S5 in Multimedia Appendix 1), indicating high satisfaction with the intervention. Furthermore, 13 (15%) participants reported difficulties or problems in applying intervention skills or strategies in daily life over the 6-week intervention, including lack of a suitable environment to apply the skills and strategies (n=6), lack of a quiet environment to concentrate for home mindfulness practice (n=5), lack of time for home mindfulness practice (n=3), and physical discomfort (n=3).

Discussion

Principal Findings

This study provided evidence for the effects of a social media–based intervention for patients with ACS. The MCARE program significantly improved psychological distress in terms of depression and anxiety at immediate postintervention and 12-week follow-up. Furthermore, the MCARE program has significant effects on psychological stress, HRQoL, dietary behavior, physical activity, and systolic BP.

The findings were supported by previous reports that mindfulness-based interventions [11,25] and health education [26] had significant effects on reducing depression and anxiety for patients with cardiovascular disease. Mindfulness training together with health education mainly targeted promoting awareness of and response to the feelings, emotions, and bodily sensations caused by the physical and psychological distress and increasing disease management knowledge and skills. Therefore, the MCARE program was assumed to have meaningful effects on psychological distress for patients with ACS. Moreover, the benefits of the MCARE program on psychological distress were likely to be sustained for a short-term period from immediate postintervention to 12-week follow-up. This might be explained by the residual gains of mindfulness skills and disease management knowledge and skills from the MCARE program. The present-focused mindfulness practice may generate a short-term, sustainable, beneficial effect on improving emotional regulation skills to facilitate participants to cope with difficult situations and lead to reduced psychological distress.

The MCARE program also significantly improved psychological stress and HRQoL, which is consistent with previous systematic reviews of mindfulness-based interventions [27] and educational interventions [28] for patients with ischemic heart disease. The MCARE program could help participants cope with their condition and thus reduce mental and emotional distress, which in turn contributed to the improvement of HRQoL, particularly in the emotional dimension.

Previous research rarely reported the effects of mindfulness-based interventions on cardiovascular risk factors. This study demonstrated the promising effects of the MCARE program on dietary behavior, physical activity, and systolic BP for patients with ACS, providing a valuable reference for future research. Health education would help patients understand the etiology, development, duration, and prognosis of their illness and learn how to manage the risk factors [29], and mindfulness training would increase their awareness of managing health threats, which together, in turn, might have empowered them to eat healthier and be more physically active. The enhanced knowledge and awareness about the illness, together with regular BP surveillance and behavioral change might have contributed to improved BP control.

The findings suggested the MCARE program had nonsignificant effects on smoking status, BMI, blood lipid profiles, and blood glucose. This might be due to that the MCARE program only provided general education and support without more effective strategies such as targeted and direct medical cessation therapy for smokers [30] and targeted weight loss strategy for patients who are overweight or obese. Blood lipid and glucose control are important targets of ACS management, which may mostly rely on adherence to core cardioprotective medications that lower blood lipid and glucose. Complementary interventions via the promotion of a healthy lifestyle, providing education on disease knowledge, and regular monitoring may facilitate blood lipid and glucose control; however, it may take a long time to observe a significant improvement in the concentration of blood lipid and glucose.

Limitations

This study has several limitations. First, this study was carried out in the Chinese population and used a convenience sampling method, which may lead to selection bias, thus limiting the representativeness of the study population and the generalizability of findings. In order to improve the representativeness of the study population, multicenter studies conducted in various regions and diverse settings are warranted. Second, the measure of the majority of outcomes relied on participants' self-reporting, although the instruments adopted in the study were reliable and validated. The subjective data may be subject to self-reporting and recall bias. Third, this study only evaluated the intervention effects at immediate postintervention and 12-week follow-up, which was unable to explore the sustainability of the intervention effects over the long term. Furthermore, we did not evaluate the consistent use of the intervention and analyze its association with health outcomes including cardiometabolic status. Thus, the patterns of use and the relationships between the use of intervention and health outcomes require further exploration in studies with a

XSL•FO
RenderX

larger sample size and a longer follow-up period. Fourth, due to the scope of this study, we did not examine whether the MCARE program supported the conceptual framework or explore whether the key concepts related to the framework would explain the possible mechanism of the intervention effects. Last, this study did not assess all potential confounding factors that may impact the psychological distress and secondary outcomes, such as the baseline cardiovascular risk of participants by objective ergometry test, which may introduce bias in the estimation of intervention effects.

Conclusions

This study pioneered a social media–based intervention for patients with ACS. The findings demonstrated that the MCARE program was an effective approach to improving psychological distress, psychological stress, HRQoL, and several cardiovascular risk factors.

Acknowledgments

The authors thank the head nurses and doctors in the research hospitals for their assistance and coordination for data collection and the patients who participated in the study.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structure and content of the mindfulness psycho-behavioral intervention (MCARE) program and supplemental results.
[DOCX File , 25 KB - jmir_v26i1e48557_app1.docx]

Multimedia Appendix 2
CONSORT eHEALTH checklist (V 1.6.1)
[PDF File (Adobe PDF File), 1087 KB - jmir_v26i1e48557_app2.pdf]

References

1. Bergmark BA, Mathenge N, Merlini PA, Lawrence-Wright MB, Giugliano RP. Acute coronary syndromes. Lancet 2022;399(10332):1347-1358 [FREE Full text] [doi: 10.1016/S0140-6736(21)02391-6] [Medline: 35367005]
2. National Center for Cardiovascular Diseases, China. Annual Report on Cardiovascular Disease in China 2022. Beijing: Peking Union Medical College Press; 2023.
3. Li Y, Cao GY, Jing WZ, Liu J, Liu M. Global trends and regional differences in incidence and mortality of cardiovascular disease, 1990-2019: findings from 2019 global burden of disease study. Eur J Prev Cardiol 2023;30(3):276-286. [doi: 10.1093/eurjpc/zwac285] [Medline: 36458973]
4. Helmark C, Harrison A, Pedersen SS, Doherty P. Systematic screening for anxiety and depression in cardiac rehabilitation—are we there yet? Int J Cardiol 2022;352:65-71 [FREE Full text] [doi: 10.1016/j.ijcard.2022.02.004] [Medline: 35143875]
5. Vaccarino V, Badimon L, Bremner JD, Cenko E, Cubedo J, Dorobantu M, et al. Depression and coronary heart disease: 2018 position paper of the ESC working group on coronary pathophysiology and microcirculation. Eur Heart J 2020;41(17):1687-1696 [FREE Full text] [doi: 10.1093/eurheartj/ehy913] [Medline: 30698764]
6. Goldberg RJ, Saczynski JS, McManus DD, Waring ME, McManus R, Allison J, et al. Characteristics of contemporary patients discharged from the hospital after an acute coronary syndrome. Am J Med 2015;128(10):1087-1093 [FREE Full text] [doi: 10.1016/j.amjmed.2015.05.002] [Medline: 26007672]
7. Collet JP, Thiele H, Barbato E, Barthélémy O, Bauersachs J, Bhatt DL, et al. 2020 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Eur Heart J 2021;42(14):1289-1367 [FREE Full text] [doi: 10.1093/eurheartj/ehaa575] [Medline: 32860058]
8. Levine GN, Cohen BE, Commodore-Mensah Y, Fleury J, Huffman JC, Khalid U, et al. Psychological health, well-being, and the mind-heart-body connection: a scientific statement from the American Heart Association. Circulation 2021;143(10):e763-e783 [FREE Full text] [doi: 10.1161/CIR.0000000000000947] [Medline: 33486973]
9. Saban KL, Tell D, De La Pena P. Nursing implications of mindfulness-informed interventions for stroke survivors and their families. Stroke 2022;53(11):3485-3493 [FREE Full text] [doi: 10.1161/STROKEAHA.122.038457] [Medline: 35904017]

XSL•FO
RenderX

10. Taylor H, Strauss C, Cavanagh K. Can a little bit of mindfulness do you good? a systematic review and meta-analyses of unguided mindfulness-based self-help interventions. Clin Psychol Rev 2021;89:102078. [doi: 10.1016/j.cpr.2021.102078] [Medline: 34537665]
11. Zou H, Cao X, Chair SY. A systematic review and meta-analysis of mindfulness-based interventions for patients with coronary heart disease. J Adv Nurs 2021;77(5):2197-2213. [doi: 10.1111/jan.14738] [Medline: 33433036]
12. Zhang Z, Pack Q, Squires RW, Lopez-Jimenez F, Yu L, Thomas RJ. Availability and characteristics of cardiac rehabilitation programmes in China. Heart Asia 2016;8(2):9-12 [FREE Full text] [doi: 10.1136/heartasia-2016-010758] [Medline: 27326243]
13. Kuan PX, Chan WK, Ying DKF, Rahman MAA, Peariasamy KM, Lai NM, et al. Efficacy of telemedicine for the management of cardiovascular disease: a systematic review and meta-analysis. Lancet Digit Health 2022;4(9):e676-e691 [FREE Full text] [doi: 10.1016/S2589-7500(22)00124-8] [Medline: 36028290]
14. Dorje T, Zhao G, Tso K, Wang J, Chen Y, Tsokey L, et al. Smartphone and Social Media-Based Cardiac Rehabilitation and Secondary Prevention in China (SMART-CR/SP): a parallel-group, single-blind, randomised controlled trial. Lancet Digit Health 2019;1(7):e363-e374 [FREE Full text] [doi: 10.1016/S2589-7500(19)30151-7] [Medline: 33323210]
15. Hou Y, Zhao X, Lu M, Lei X, Wu Q, Wang X. Brief, one-on-one, telephone-adapted mindfulness-based stress reduction for patients undergoing percutaneous coronary intervention: a randomized controlled trial. Transl Behav Med 2019;9(6):1216-1223. [doi: 10.1093/tbm/ibz130] [Medline: 31504974]
16. Leventhal H, Phillips LA, Burns E. The Common-Sense Model of self-regulation (CSM): a dynamic framework for understanding illness self-management. J Behav Med 2016;39(6):935-946. [doi: 10.1007/s10865-016-9782-2] [Medline: 27515801]
17. Ibanez B, James S, Agewall S, Antunes MJ, Bucciarelli-Ducci C, Bueno H, et al. 2017 ESC guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: the task force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2018;39(2):119-177 [FREE Full text] [doi: 10.1093/eurheartj/ehx393] [Medline: 28886621]
18. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
19. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
20. Cohen S, Williamson GM. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of Health: Claremont Symposium on Applied Social Psychology. Newbury Park, CA: Sage Publications, Inc; 1988:31-67.
21. Oldridge N, Höfer S, McGee H, Conroy R, Doyle F, Saner H, (for the HeartQoL Project Investigators). The HeartQoL: part II. validation of a new core health-related quality of life questionnaire for patients with ischemic heart disease. Eur J Prev Cardiol 2014;21(1):98-106 [FREE Full text] [doi: 10.1177/2047487312450545] [Medline: 22822180]
22. Hennrikus D, Joseph AM, Lando HA, Duval S, Ukestad L, Kodl M, et al. Effectiveness of a smoking cessation program for peripheral artery disease patients: a randomized controlled trial. J Am Coll Cardiol 2010;56(25):2105-2112 [FREE Full text] [doi: 10.1016/j.jacc.2010.07.031] [Medline: 21144971]
23. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35(8):1381-1395 [FREE Full text] [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
24. Walker SN, Sechrist KR, Pender NJ. The health-promoting lifestyle profile: development and psychometric characteristics. Nurs Res 1987;36(2):76-81. [Medline: 3644262]
25. Scott-Sheldon LAJ, Gathright EC, Donahue ML, Balletto B, Feulner MM, DeCosta J, et al. Mindfulness-based interventions for adults with cardiovascular disease: a systematic review and meta-analysis. Ann Behav Med 2020;54(1):67-73 [FREE Full text] [doi: 10.1093/abm/kaz020] [Medline: 31167026]
26. Shi W, Ghisi GLM, Zhang L, Hyun K, Pakosh M, Gallagher R. A systematic review, meta-analysis, and meta-regression of patient education for secondary prevention in patients with coronary heart disease: impact on psychological outcomes. Eur J Cardiovasc Nurs 2022;21(7):643-654 [FREE Full text] [doi: 10.1093/eurjcn/zvac001] [Medline: 35134883]
27. Zou H, Cao X, Geng J, Chair SY. Effects of mindfulness-based interventions on health-related outcomes for patients with heart failure: a systematic review. Eur J Cardiovasc Nurs 2020;19(1):44-54 [FREE Full text] [doi: 10.1177/1474515119881947] [Medline: 31635481]
28. Anderson L, Brown JP, Clark AM, Dalal H, Rossau HK, Bridges C, et al. Patient education in the management of coronary heart disease. Cochrane Database Syst Rev 2017;6(6):CD008895 [FREE Full text] [doi: 10.1002/14651858.CD008895.pub3] [Medline: 28658719]
29. Kris-Etherton PM, Petersen KS, Després JP, Anderson CAM, Deedwania P, Furie KL, et al. Strategies for promotion of a healthy lifestyle in clinical settings: pillars of ideal cardiovascular health: a science advisory from the American Heart Association. Circulation 2021;144(24):e495-e514 [FREE Full text] [doi: 10.1161/CIR.0000000000001018] [Medline: 34689589]

XSL•FO
RenderX

30. Hartmann-Boyce J, Chepkin SC, Ye W, Bullen C, Lancaster T. Nicotine replacement therapy versus control for smoking cessation. Cochrane Database Syst Rev 2018;5(5):CD000146 [FREE Full text] [doi: 10.1002/14651858.CD000146.pub5] [Medline: 29852054]

Abbreviations

ACS: acute coronary syndrome
BP: blood pressure
GAD-7: 7-item Generalized Anxiety Disorder
GEE: generalized estimating equation
HRQoL: health-related quality of life
MCARE: mindfulness psycho-behavioral intervention
PHQ-9: 9-item Patient Health Questionnaire
RCT: randomized controlled trial

Edited by A Mavragani; submitted 30.04.23; peer-reviewed by J Chen, D Fletcher; comments to author 24.10.23; revised version received 30.10.23; accepted 03.01.24; published 20.02.24.

Please cite as:
Zou H, Chair SY, Feng B, Liu Q, Liu YJ, Cheng YX, Luo D, Wang XQ, Chen W, Huang L, Xianyu Y, Yang BX
A Social Media–Based Mindfulness Psycho-Behavioral Intervention (MCARE) for Patients With Acute Coronary Syndrome: Randomized Controlled Trial
J Med Internet Res 2024;26:e48557
URL: https://www.jmir.org/2024/1/e48557
doi:10.2196/48557
PMID:38376899

©Huijing Zou, Sek Ying Chair, Bilong Feng, Qian Liu, Yu Jia Liu, Yu Xin Cheng, Dan Luo, Xiao Qin Wang, Wei Chen, Leiqing Huang, Yunyan Xianyu, Bing Xiang Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Internet-Based Cognitive Behavioral Therapy for Children and Adolescents With Dental or Injection Phobia: Randomized Controlled Trial

Robert Schibbye[1,2], MSci; Erik Hedman-Lagerlöf[3], PhD; Viktor Kaldo[4,5], PhD; Göran Dahllöf[1,2,6], PhD; Shervin Shahnavaz[4], PhD

[1]Department of Dental Medicine, Karolinska Institutet, Huddinge, Sweden

[2]Center of Pediatric Oral Health, Stockholm, Sweden

[3]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

[4]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm Health Care Services, Region Stockholm, Sweden

[5]Department of Psychology, Faculty of Health and Life Sciences, Linnaeus University, Växjö, Sweden

[6]Center for Oral Health Services and Research, Mid-Norway (TkMidt), Trondheim, Norway

Corresponding Author:
Robert Schibbye, MSci
Department of Dental Medicine
Karolinska Institutet
Box 4064
Huddinge, 141 04
Sweden
Phone: 46 8 524 800 0
Email: robert.schibbye@ki.se

Abstract

Background: Dental phobia (DP) and injection phobia (IP) are common in children and adolescents and are considered some of the biggest obstacles to successful treatment in pediatric dentistry. Cognitive behavioral therapy (CBT) is an evidence-based treatment for anxiety and phobias. As the availability of CBT in dentistry is low, internet-based CBT (ICBT) was developed. Open trials have shown that ICBT is a promising intervention, but randomized trials are lacking.

Objective: This randomized controlled trial tests whether therapist-guided ICBT supported by a parent could reduce fear, allowing children and adolescents with DP or IP to receive dental treatment.

Methods: We enrolled 33 participants (mean age 11.2, SD 1.9 y) whom a clinical psychologist had diagnosed with DP, IP, or both. After inclusion, participants were randomized to either ICBT (17/33, 52%) or a control group of children on a waitlist (16/33, 48%). ICBT was based on exposure therapy and comprised a 12-week at-home program combined with visits to their regular dental clinic. Participants corresponded weekly with their therapist after completing each module, and 1 parent was designated as a coach to support the child in the assignments during treatment. All participants completed measurements of the outcome variables before treatment start and after 12 weeks (at treatment completion). The measurements included a structured diagnostic interview with a clinical psychologist. Our primary outcome measure was the Picture-Guided Behavioral Avoidance Test (PG-BAT), which assesses the ability to approach 17 dental clinical procedures, and a positive clinical diagnosis. Secondary outcome measures included self-report questionnaires that measured self-efficacy and levels of dental and injection anxiety. The children and their parents completed the questionnaires.

Results: All participants underwent the 12-week follow-up. After treatment, 41% (7/17) of the participants in the ICBT group no longer met the diagnostic criteria for DP or IP, whereas all participants in the control group did (P=.004). Repeated-measure ANOVAs showed that ICBT led to greater improvements on the PG-BAT compared with the control group; between-group effect sizes for the Cohen d were 1.6 (P<.001) for the child-rated PG-BAT and 1.0 (P=.009) for the parent-rated PG-BAT. Reductions in our secondary outcomes—dental fear and anxiety (P<.001), negative cognitions (P=.001), and injection fear (P=.011)—as well as improvements in self-efficacy (P<.001), were all significantly greater among children in the ICBT group than in the controls. No participants reported adverse events.

XSL•FO
RenderX

Conclusions: ICBT seems to be an effective treatment for DP and IP in children and adolescents. It reduced fear and anxiety and enabled participants to willingly receive dental treatment. ICBT should be seriously considered in clinical practice to increase accessibility; this therapy may reduce the need for sedation and restraint and lead to better dental health in children and adolescents.

Trial Registration: ClinicalTrials.gov NCT02588079; https://clinicaltrials.gov/study/NCT02588079

(J Med Internet Res 2024;26:e42322) doi:10.2196/42322

KEYWORDS

adolescents; cognitive behavioral therapy; CBT; children; dental anxiety; dental fear; internet; specific phobia

Introduction

Background

Dental fear and anxiety (DFA) refers to strong negative emotions associated with dental treatment [1]. In children and adolescents, the prevalence of DFA is 24%, making it a common clinical problem in pediatric dentistry [2]. Dental phobia (DP) and injection phobia (IP) are specific phobias classified as psychiatric disorders in *the Diagnostic and Statistical Manual of Mental Disorders* [3]. A specific phobia is defined as a marked and persistent fear that is excessive or unreasonable, in this case concerning the dentist, dental care in general, or a specific part of the dental treatment. The phobic object or situation is either avoided or endured with intense anxiety or distress. The fear is also persistent and typically lasts for at least 6 months, interfering significantly with the person's normal functioning [3]. In dental care, if the phobia only pertains to receiving an injection but not to any other situation or object, it should be referred to as an IP if the fear of injections is general (ie, the fear is triggered by all types of injections), and it should be referred to as intraoral IP (a subtype of IP) if the fear only pertains to injections in the mouth. Both DP and IP are classified as subtypes of blood-injury-injection phobia, that is, a specific phobia related to seeing blood, injections, injuries, and disability or exposure to these or similar medical procedures [4]. It has been shown that 20% of children meet the diagnostic criteria for blood-injury-injection phobia at least once between the ages of 4 and 14 years [5]. DFA and DP or IP often lead to avoidance of dental care, which can cause decreased dental and general health [6].

Systematic reviews [1,7] have observed that traditional methods of treating pediatric patients with DFA, DP, or IP (ie, tell-show-do, restraint of some sort, sedation with midazolam or nitrous oxide, and general anesthesia) show little scientific evidence of treating the fear itself. These methods may be effective for preventing the development of DFA, DP, or IP before the first dental treatment or temporarily enabling dental treatment in groups of patients with anxiety and fear. However, they are ineffective in patients with severe DFA or previous negative treatment experiences, such as those with DP or IP. One reason may be that they do not incorporate exposure-based cognitive behavioral therapy (CBT), which is arguably the gold-standard method for treating specific phobias. Another reason is that they were not developed using a multidisciplinary approach that combines findings from the fields of both dental care and psychological science on reduction of anxiety and fear [1].

Stronger forms of DFA, DP, and IP are primarily learned through adverse events (ie, painful treatment procedures) in dental or general health care contexts, which results in fear being associated with dental care or specific objects or procedures, a process called respondent or classic conditioning [8,9]. CBT treatments are widely recognized as the most effective treatment for specific phobias and anxieties, but they are rarely implemented in dental practice [10]. The focus of CBT in the pediatric dental care setting is to reduce patients' anxiety so that they can willingly receive treatment without the need for sedation or restraint. CBT, which is normally administered face-to-face by a trained psychotherapist, has been shown to be effective for both adults [11,12] and children and adolescents [13-16] with DFA, DP, and IP. However, accessibility to and use of CBT are generally low in dentistry. Common barriers include a lack of trained CBT therapists, high costs for the family if the child is treated in private psychotherapeutic services, and time constraints (ie, it might be difficult for parents to take time off from work for an extended period for weekly visits to the psychologist) [17]. In addition, there might be dentistry-specific barriers, such as a lack of knowledge about CBT methods and problems integrating workflows and new personnel (ie, CBT therapists) into the clinic.

Internet-Based CBT

Internet-based CBT (ICBT) was developed to make CBT more accessible. ICBT is delivered over the internet through text, video clips, animations, and audio files instead of via a therapist in face-to-face sessions. ICBT uses the same mechanisms and principles as face-to-face CBT, and they have been shown to have comparable effect sizes [18]. Previous research has shown that ICBT, in which a therapist guides the patient, is associated with greater effects than unguided treatments [19]. Therapist-guided ICBT for fears and anxieties consists of a base of web-based material that conveys information and concrete exercises to reduce the patient's fears, and the patient receives support from a psychologist through a web-based chat or email function. ICBT has been found to be effective for children with specific phobias [20,21] and children and adolescents with DP [22]. These studies have used parents as coaches guiding the child through the treatments.

The aim of this study was to test, in a randomized controlled trial (RCT), whether ICBT can reduce fear and increase willingness to receive dental treatment in children and adolescents with DP or IP.

XSL•FO
RenderX

Methods

Recruitment

Participants were recruited through dental clinics and social media advertisements that referred interested families to a website of the Department of Dental Medicine at Karolinska Institute. The study website contained brief information about the study and the targeted population. Interested parents (or caregivers) applied directly through the website and were then assigned a log-in for web-based screening. We informed dental clinics throughout Sweden and encouraged them to advertise for participants in the waiting rooms. We also asked dentists, especially specialists in pediatric dentistry, to recommend patients with DFA that was so severe that it interfered with dental treatment to apply to the study. In Sweden, both general and specialist pediatric dentistry are publicly funded and offered free of charge to all children. However, waiting times and the availability of specialist care differ between regions. Participants were recruited from October 2015 to December 2019. We planned to recruit 50 participants. However, owing to the slower-than-expected recruitment pace, we extended the originally planned recruitment time and ended the recruitment with 33 participants included in the study. Of the 58 interested applicants who provided informed consent and filled out the background information, 45% (26/58) were currently under treatment or on the waiting list of a specialist pediatric dentist. Furthermore, 76% (44/58) lived in the Stockholm region, whereas the rest were spread out throughout other regions of Sweden. Finally, 66% (38/58) were advised directly by their dentist to apply to the study, 2% (1/58) saw an advertisement on a notice board at their dentist's office, 10% (6/58) found the study by searching on the internet or through social media advertisements, 3% (2/58) had heard of it through a friend or relative, and 19% (11/58) were patients in general dental care who were advised to apply to the study when consulting a specialist.

Screening

The first step in web-based screening provided information on the study and details about informed consent from the caregivers (henceforth referred to as the parent or parents). After providing informed consent, the parent (parents had to designate one of them to fill out the questionnaires and conduct the telephone interview) and child provided basic background information on a questionnaire and then filled out 4 standardized questionnaires: the Picture-Guided Behavioral Avoidance Test (PG-BAT) [16], the Children's Fear Survey Schedule–Dental Subscale (CFSS-DS) [23], the Children's Negative Cognitions in Dentistry (CNCD) scale [22], and the Injection Phobia Scale for Children (IPSC) [24].

After assessment of the first screening step, we asked the parents of eligible participants to fill out the Development and Well-Being Assessment [25] and gave them new log-in credentials for accessing the survey on a different platform.

A clinical psychologist then interviewed the parent over the telephone. The semistructured diagnostic interview used the specific phobia section of the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime (K-SADS-PL) [26] to establish whether the child had a specific phobia diagnosis of DP or IP. The interviewer also followed up the results of the Development and Well-Being Assessment with the parent when they indicated that the child might meet the potential exclusion criteria.

Inclusion and Exclusion Criteria

Eligible participants who met the inclusion criteria (Textbox 1) were asked to join the study; they all agreed to participate.

Participants who met any of the exclusion criteria (Textbox 2) were excluded.

Textbox 1. Inclusion criteria.

- The child's age is between 8 and 15 years.
- Child *and* parents sign informed consent forms.
- A psychologist establishes a diagnosis of dental phobia or injection phobia according to the *Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition)* from the results of the web-based parental version of the Development and Well-Being Assessment in the initial screening and of the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime in the semistructured diagnostic interview.
- The Swedish language skills of the child *and* parents are sufficient to manage treatment and the questionnaires.
- Access to a computer and the internet is readily available.
- Child *and* parents have sufficient time and motivation to work with internet-based cognitive behavioral therapy 3 hours each week for 12 weeks.
- Parents agree to book at least 3 visits to the dental clinic during the 12 weeks of treatment.
- If the child is diagnosed with injection phobia, the parents agree to exposure training for intraoral injection phobia at the dental clinic even if the child does not need dental treatment.

XSL•FO
RenderX

Textbox 2. Exclusion criteria.

- Full points on the child and parent versions of the Picture-Guided Behavioral Approach Test (the maximum score of 17 means that the child and the parent assess that the child is already able to manage the most challenging dental situations)
- A score of ≤31 on the child and parent versions of the Children's Fear Survey Schedule–Dental Subscale and no diagnosis of injection phobia by the psychologist during screening
- A previously established diagnosis of a neurodevelopmental disorder or a likely diagnosis of a neurodevelopmental disorder according to the results of the Development and Well-Being Assessment or the psychologist during screening
- Other psychiatric disorders, such as, severe depression, eating disorders, or self-harm behavior, that require treatment first before dental-related specific phobias
- Current or planned psychiatric or psychological examination
- Current or planned psychological treatment
- Stressful life experiences in the previous 12 months, such as parental divorce or somatic illnesses, that the parent or psychologist consider an obstacle in treatment
- Cognitive behavioral treatment for dental fear and anxiety or dental phobia or injection phobia in the previous 3 years

Randomization and Outcome Measures

The study began by establishing a baseline for the outcome measures. These were the PG-BAT and CFSS-DS, which the child and parent rated separately. The CNCD scale, IPSC, and Self-Efficacy Questionnaire for Phobic Situations [27] were rated by the child only. Finally, the Parental Self-Efficacy Questionnaire for Dental Anxiety [22] was rated by the parent. One person uninvolved with the study and blinded to the identities of the participants then randomized the participants consecutively in a 1:1 ratio to active treatment or to a waiting list (the control group) using the list randomization tool [28] and the participants' study ID number. New participants were continuously assigned using a randomized block design; the block size varied depending on the number of participants available at the time of randomization. In case there was 1 participant or an uneven number of participants, a dummy participant was added to the block to maintain a 1:1 ratio for randomization. A follow-up at 12 weeks, upon treatment completion, included all outcome measures in the baseline measurement in addition to a semistructured diagnostic interview with a psychologist over the telephone using the specific phobia section of the K-SADS-PL. Furthermore, the follow-up included a questionnaire concerning qualitative aspects of the child's current dental anxiety and the ICBT treatment; response options included free text, multiple choice, and visual analog scales (VASs). The questionnaires also contained clear questions about adverse events or unwanted treatment effects, which the psychologists covered in their interviews. After the follow-up, the controls were offered the same treatment that the ICBT group had received. Figure 1 shows a flowchart of the participants in this study.

XSL•FO
RenderX

Figure 1. Flowchart of participant recruitment and allocation to the 2 study groups: internet-based cognitive behavioral therapy (ICBT) and control group. CONSORT: Consolidated Standards of Reporting Trials.

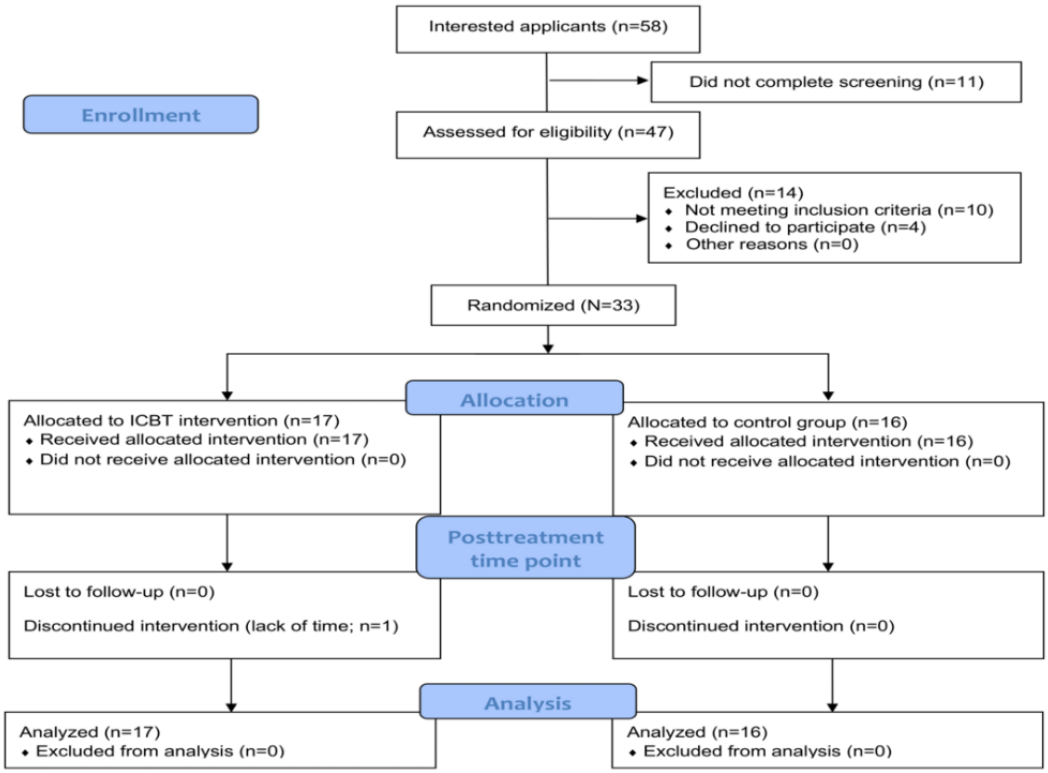

Primary Outcome Measures

This study had 2 primary outcome measures. The first was the specific phobia section of the K-SADS-PL [26], which we used to diagnose DP and IP among the participants. A clinical psychologist included this in a semistructured interview over the telephone with the parent. The K-SADS-PL has been shown to be a reliable and valid instrument for assessing psychiatric diagnoses.

The second primary outcome measure was the PG-BAT [17,22], which the child and parent rated separately. The recommended outcome measure of CBT treatment for specific phobias is the Behavioral Approach Test [9,13]. The PG-BAT is similar and was used in a previous study of ICBT [22] for dental anxiety. The test has been shown to have good psychometric properties and measures the number of dental procedures (n=17; response options: yes=1 and no=0; Textbox 3) that the participants or parents felt that the participants could manage on their own. Pictures of 17 dental clinical procedures are organized according to increasing ability to provoke anxiety. Each image includes a written description of the situation and procedure. Scores range from 0 (not even able to enter the dentist's room) to 17 (able to manage all treatment steps).

XSL•FO
RenderX

Textbox 3. The dental procedures in the Picture-Guided Behavioral Avoidance Test.

- Enter the treatment room.
- Sit in the treatment chair with a paper bib around the neck.
- Sit in the treatment chair while the chair is lowered.
- Lie in the dental chair with the lamp turned on and the dental tools close by on a tray.
- Open the mouth and let the dentist look into the mouth.
- Let the dentist use a small saliva ejector in the mouth.
- Let the dentist blow air and water into the mouth.
- Undergo a clinical exam with a dental mirror.
- Undergo a clinical exam with a mirror and a dental probe.
- Let the dentist take an x-ray in the back of the mouth.
- Receive topical anesthesia.
- Receive an injection of a local anesthetic.
- Let the dentist use a large saliva ejector in the mouth.
- Let the dentist attach composite filling to the surface of a tooth.
- Let the dentist drill with a high-speed drill.
- Let the dentist polish a filling with a low-speed drill.
- Let the dentist extract a tooth.

Secondary Outcome Measures

The CFSS-DS [22] measures DFA and comprises 15 items rated on a 5-point scale from *no fear* (1) to *high fear* (5). The items describe situations in dental and medical care and were rated separately by the child and the parent. The Swedish version of the CFSS-DS has been shown to have good psychometric properties [23].

The children rated the 5-item CNCD scale [22] on a 10-point VAS. The scale end points are 0=*no negative thoughts at all* and 1=*some negative thoughts*; happy and sad figures illustrate these concepts [10,14]. The questions regard the presence and strength of 5 negative thoughts that are common in dentistry: uncontrollability, distrust of dentists, unpredictability, dangerousness, and pain related to dentistry.

The IPSC [24] was rated by the child and consists of 18 items rated on a 5-point scale ranging from *no fear* (1) to *high fear* (5). The questions regard the fear of different situations and procedures related to injections. This test has been shown to have good psychometric properties [24].

Participants rated the Swedish translation of the 14-item Self-Efficacy Questionnaire for Phobic Situations [27] on a 5-point scale with the following end points: 1=*low self-efficacy* and 5=*high self-efficacy*.

Parents rated a Swedish version of the 12-item Parental Self-Efficacy Questionnaire for Dental Anxiety [22] on a 10-point scale with the following end points: 0=*no parental self-efficacy* and 10=*very high parental self-efficacy*. The instrument was evaluated in our open trial study on ICBT for dental anxiety in children and adolescents [22].

The ICBT Intervention Platform

The form of ICBT used in this study was based on a previously published manual [17] that was used in a pilot study [22]. The central component of the treatment was exposure through video and audio recordings, a toolkit for use at home, and in vivo exposure at the dental clinic. Other components of the treatment were psychoeducation, behavioral analyses, controlled breathing, and parental education (Textbox 4). Parents were asked to decide among themselves who would be responsible for coaching and supporting the children and adolescents, thereby becoming the designated coach throughout the treatment. The second treatment module targeted the coach with information on how to support, motivate, and assist the children and adolescents with their assignments.

XSL•FO
RenderX

Textbox 4. Outline of the 12 internet-based cognitive behavioral therapy (ICBT) modules (one module each week).

- Introduction to cognitive behavioral therapy and the web-based treatment
- Psychoeducation, practical arrangements, home assignment, how to guide a child to elicit and reinforce behavior change, reward strategies, and enhancing the child's self-efficacy (for the coach only)
- Behavioral analyses, child psychoeducation and treatment rationale, and goal setting
- Constructing an exposure list and beginning exposure
- Continued exposure (films and training package) and controlled breathing
- Dentistry-related communication training and preparation for the first dental visit
- Evaluation of the dental visit and cognitive restructuring
- Midterm evaluation of ICBT and of exposure at the dentist and relaxation techniques
- Pain and pain management education; fear, thoughts, and pain; and focus shift and acceptance training
- Problem-solving and mindfulness training
- Repetition, strategies for maintaining change and relapse prevention, and letter to yourself
- Relapse prevention plan, enhance your self-efficacy, and diploma

The intervention and all data collection occurred on a secure web-based platform hosted by the Internetpsykiatri (internet psychiatry) unit run by Stockholm Health Care Services, Region Stockholm, Sweden. The coaches and children and adolescents shared the same log-in. Each module indicated who was responsible for completing the various measurements, treatment modules, and assignments. One of the assignments for the coach, for example, was to call the participant's local dental clinic and book at least 3 visits for in vivo exposure, which the dental staff would perform. Participants downloaded information and instructions on exposure from the platform and mailed printed or digital copies of these to the clinic. During week 2, the participants received a toolkit with dental instruments and a VAS that the children and adolescents and coach could use for the exposure assignments at home (Multimedia Appendix 1).

The ICBT intervention comprised 12 web-based treatment modules that were made accessible to the participants, with 1 module per week. Each participant was assigned a psychologist who supported them throughout the treatment. In total, 3 psychologists were involved in this study; all had a 5-year degree in clinical psychology at a minimum. In addition, all 3 had face-to-face clinical dental experience in CBT with children and adolescents who had DP or IP. All modules had some kind of text; most modules also had pictures, animations, and videos and audio of various dental procedures for exposure purposes. Multimedia Appendices 2-4 show example screenshots of the interface and treatment modules.

Each week, the children and adolescents and coaches completed a new module that concluded with an assignment containing knowledge questions about the module and practical exercises. Each week, participants sent their responses to the questions and a log of their assignments to their psychologist, who would send feedback and grant access to the next module within 2 days. Participants could also message their psychologist directly through their account on the platform and expect a reply, on average, within 2 working days. The psychologists would send reminders to inactive participants about continuing work with the modules via SMS text message and email. If participants were inactive for >10 days, the psychologist would try to reach them by telephone.

A total of 12 weeks after treatment had begun, and regardless of whether the participants had finished the treatment modules, all treatment modules were made available on the platform. The account of the children and adolescents and coaches with its log-in credentials remained active for the next 12 months; however, participants were no longer able to communicate with their psychologist.

Statistical Analysis

All statistical analyses were conducted using SPSS (version 27; IBM Corp). We accepted a 5% type-I error in all analyses. To compare the means between the 2 study conditions, we used 2-tailed *t* tests. Repeated-measure ANOVAs were conducted to evaluate possible differences in changes over the 12 weeks (baseline to follow-up) between the ICBT and control groups. We estimated the effect size using the Cohen *d*, that is, the standardized mean difference [29]. Chi-square tests of independence were conducted to evaluate possible between-group differences in meeting the diagnostic criteria. Before conducting our analyses, we checked the data for normality. We expected the effects to be in line with our open trial study using the same procedures and treatment [22]. Thus, the power calculation was based on an estimated effect size of Cohen d=1.0 and showed that, to obtain 80% power (Cronbach α=.05), 17 participants in each arm (N=34) were required. We used an intention-to-treat design; that is, participants were included in the analyses irrespective of the extent to which they had completed the treatment.

Ethical Considerations

Before the participants could complete screening and enter the study, informed consent was obtained. Information about the study and informed consent was provided in Swedish and included permission for secondary analysis of the data without additional consent. Informed consent was provided by both caregivers separately if there were 2. There was no compensation given to the participants, and similar to all dental

XSL•FO
RenderX

care for children and adolescents in Sweden, they received the intervention for free. For added security, the secure web-based platform in which data were collected required a 2-step authentication via SMS text message for logging in, both for participants and the study staff. When extracting data from the platform for analysis, they were deidentified, and an anonymous study ID was used as an identifier for each participant. The regional ethics review board of Stockholm approved this study (ID 2014/633-31/5).

Results

Overview

All participants had a diagnosis of DP or IP (or both) at baseline. Of the 33 children included in this study, 21 (64%) were female. The mean age of the study sample was 11.2 (SD 1.9) years. During previous dental treatment, sedation or general anesthesia had been administered or restraint had been necessary for all but 2 participants (31/33, 94%; 1 participant in each group). Table 1 presents the participant characteristics.

Table 2 presents the baseline scores of the children and their parents in the 2 study groups. The 2-tailed *t* tests revealed no significant between-group differences in these measures (all *P*>.05). Chi-square tests of independence showed no significant between-group associations in meeting the diagnostic criteria for DP, IP, or DP and IP (all *P*>.05).

Table 1. Participant characteristics of the internet-based cognitive behavioral therapy (ICBT) and control groups (N=33).

Variable	ICBT (n=17)	Control (n=16)	Total (N-=33)
Age (y), mean (SD)	11.3 (1.8)	11.1 (1.9)	11.2 (1.9)
Female participants, n (%)	10 (59)	11 (69)	21 (64)
Both parents born in Sweden, n (%)	15 (88)	14 (88)	29 (88)
Living with both parents, n (%)	16 (94)	12 (75)	28 (85)
Duration (mo), mean (SD)			
Fear of the dentist	38 (30)	47 (50)	45 (41.5)
Fear of injections	60 (41)	72 (55)	69.8 (48.4)
Diagnosis, n (%)			
Dental phobia	14 (82)	14 (88)	28 (85)
Injection phobia	15 (88)	14 (88)	29 (88)
Dental and injection phobia	12 (71)	12 (75)	24 (73)
Another specific phobia[a]	5 (29)	5 (31)	10 (30)
The following occurred during previous dental treatment, n (%)			
Sedation was administered	16 (94)	15 (94)	31 (94)
Restraint was necessary	10 (59)	8 (50)	18 (55)
General anesthesia was administered	4 (24)	2 (12)	6 (18)

[a]Diagnosis of a specific phobia other than dental or injection phobia.

XSL•FO
RenderX

Table 2. Baseline self-report scores of the children and their parents in the internet-based cognitive behavioral therapy (ICBT) and control groups.

Outcome variable	ICBT (n=17), mean (SD; 95% CI)	Control (n=16), mean (SD; 95% CI)	*P* value[a]
Child			
Dental procedures managed[b]	11.1 (4.2; 9.0-13.3)	11.7 (3.2; 10.0-13.4)	.67
Dental fear and anxiety[c]	37.1 (10.5; 31.6-42.5)	35.3 (12.3; 28.7-41.8)	.65
Negative cognitions[d]	25.5 (13.0; 18.8-32.2)	22.0 (11.6; 15.8-28.2)	.43
Injection fear[e]	41.7 (10.5; 36.2-47.1)	43.3 (12.1; 36.9-49.7)	.68
Self-efficacy[f]	25.6 (6.6; 22.2-29.0)	29.8 (7.3; 26.0-33.7)	.09
Parent			
Dental procedures managed[b]	10.6 (4.6; 8.2-13.0)	11.2 (4.1; 9.0-13.4)	.70
Dental fear and anxiety[c]	35.5 (9.5; 30.6-40.4)	36.8 (9.7; 31.7-42.0)	.69
Parental self-efficacy[g]	111.2 (15.2; 103.4-119.0)	105.6 (18.3; 95.9-115.4)	.34

[a]*P* values are based on the 2-tailed *t* test.

[b]Picture-Guided Behavioral Avoidance Test; score range: 0 to 17.

[c]Children's Fear and Survey Schedule–Dental Subscale; score range: 15 to 75.

[d]Children's Negative Cognitions in Dentistry scale; score range: 0 to 50.

[e]Injection Phobia Scale for Children; score range: 18 to 90.

[f]Self-Efficacy Questionnaire for Phobic Situations; score range: 14 to 70.

[g]Parental Self-Efficacy Questionnaire for Dental Anxiety; score range: 0 to 120.

Treatment Adherence and Dropout Frequency

All participants underwent the 12-week follow-up; thus, no data were missing. Of the 17 participants in the ICBT group, 1 (6%) chose not to continue the intervention after week 2 because of other priorities and lack of time; we categorized this participant as a dropout. The other 94% (16/17) of the participants completed at least 5 modules and started the first steps of exposure training. A total of 81% (13/16) of the participants were considered treatment completers as they finished the most important steps of the treatment (module 8) and were conducting in vivo exposure sessions at their local dental clinic at the time of the follow-up. The mean number of completed modules after 12 weeks was 8.4 (SD 3.4). In total, 12% (2/16) of the participants in the control group chose not to enroll in ICBT after the study was concluded.

Primary Outcome Measures

In the ICBT group, 41% (7/17) of the participants no longer met the diagnostic criteria for DP, IP, or DP and IP at the posttreatment clinical interview compared with 0% in the control group. A chi-square test of independence showed that the difference was significant (χ^2_1=8.4; *P*=.004). A total of 65% (11/17) of the participants lost at least 1 of their earlier diagnoses of either DP or IP (N=33, χ^2_1=15.5; *P*<.001) compared with 0% in the control group. Table 3 presents the results of the intervention stratified by diagnosis.

Furthermore, children in the ICBT group who rated the PG-BAT noted being able to manage a significantly larger number of dental procedures at the follow-up (mean 3.6, SD 3.1) compared with children in the control group, who reported a decrease over the same period (mean −0.1, SD 1.4; *P*<.001; Figure 2). Parents in the ICBT group also noted that their children were able to manage a significantly larger number of dental procedures (mean 3.7, SD 3.7) compared with parents in the control group (mean 0.6, SD 2.5; *P*=.009; Figure 3). The between-group effect sizes (Cohen *d*) at the follow-up, calculated from the group-by-time interaction effects resulting from a repeated-measure ANOVA analysis of the child- and parent-rated PG-BAT, were 1.6 (95% CI 0.8-2.3) and 1.0 (95% CI 0.3-1.7), respectively; this indicated large treatment effects in the primary outcome measuring willingness and capability to manage dental treatment.

Table 3. Participants meeting the criteria for diagnoses of dental phobia or injection phobia at baseline and at the follow-up in the internet-based cognitive behavioral therapy (ICBT) and control groups.

Diagnosis	Treatment period, n (%)				Chi-square (*df*)	*P* value
	ICBT group (n=17)		Control group (n=16)			
	Before	After	Before	After		
Dental phobia	14 (82)	7 (41)	14 (88)	14 (88)	9.4 (1)	.002
Injection phobia	15 (88)	10 (59)	14 (88)	14 (88)	5.6 (1)	.02

XSL•FO
RenderX

Figure 2. Number of dental procedures managed according to the Picture-Guided Behavioral Avoidance Test (PG-BAT). Child ratings at 2 time points (baseline and posttreatment time point) in the 2 study groups: internet-based cognitive behavioral therapy (ICBT) and control group.

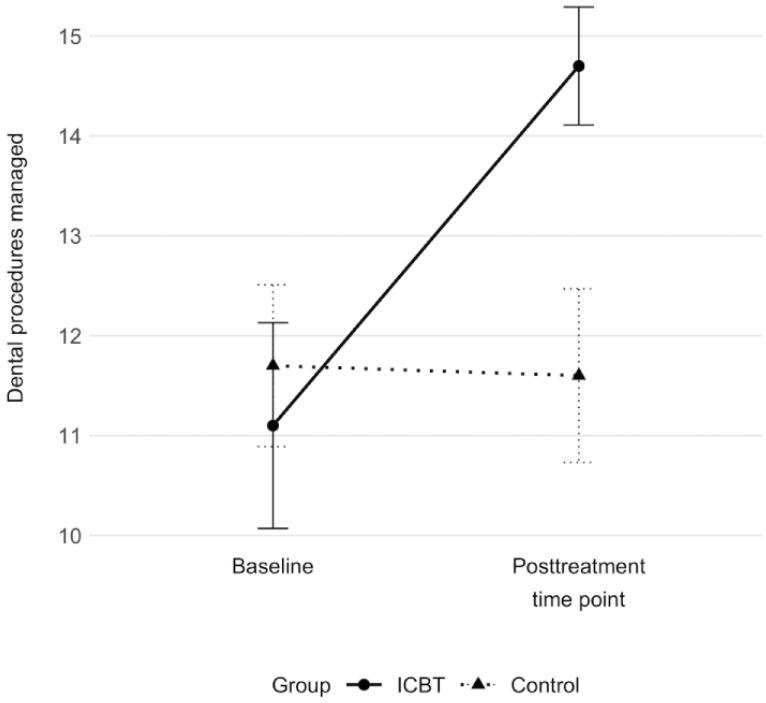

Figure 3. Number of dental procedures managed according to the Picture-Guided Behavioral Avoidance Test (PG-BAT). Parent ratings at 2 time points (baseline and posttreatment time point) in the 2 study groups: internet-based cognitive behavioral therapy (ICBT) and control group.

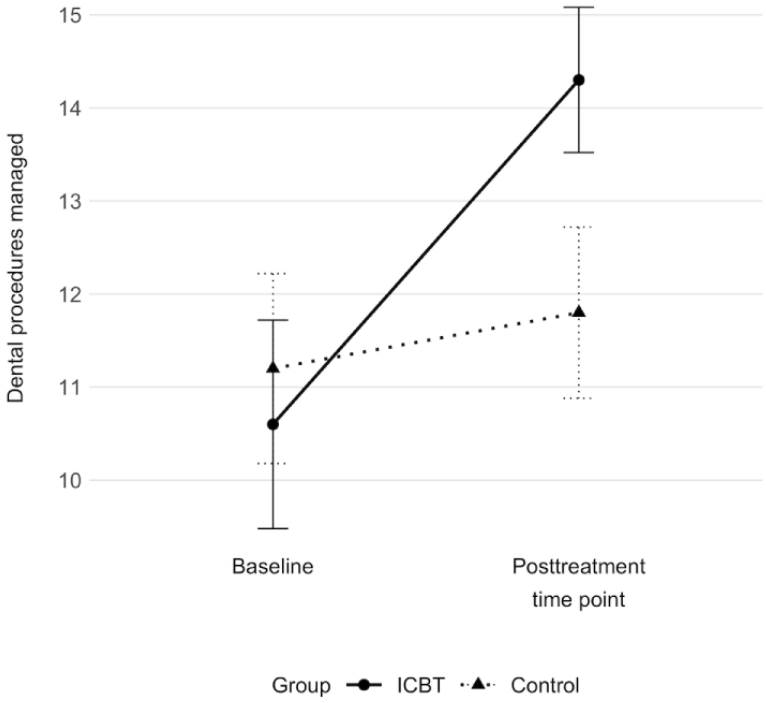

Secondary Outcome Measures

During the 12 weeks between baseline and follow-up, significantly larger improvements were observed in the ICBT group than among the controls in all secondary outcome measures except parental self-efficacy (Table 4) for mean changes and the group-by-time interaction effects calculated in the repeated-measure ANOVA. Multimedia Appendix 5 presents a complete table of means and SDs at follow-up for all outcome measurements stratified by allocation group. Multimedia Appendix 6 presents within-group effects for the ICBT group; no within-group effects were observed in the control group. Participants reported no adverse events or unintended effects during the intervention.

XSL•FO
RenderX

Table 4. Child- and parent-rated changes between baseline and follow-up in secondary outcomes and between-group effect sizes in the internet-based cognitive behavioral therapy (ICBT) and control groups.

Participant and change	ICBT group (n=17)		Control group (n=16)		Repeated-measure ANOVA[a]	
	Participants, n (%)	Values, mean (SD)	Participants, n (%)	Values, mean (SD)	*P* value[b]	Cohen *d* (95% CI)
Child						
Δ Dental fear and anxiety[c]	17 (100)	−11.7 (8.7)	16 (100)	2.6 (9.2)	*<.001* [d]	−1.6 (−2.4 to −0.8)
Δ Negative cognitions[e]	15 (88)	−13.3 (9.7)	16 (100)	1.7 (13.1)	*.001*	−1.3 (−2.1 to −0.5)
Δ Injection fear[f]	16 (94)	−10.6 (10.0)	16 (100)	0.7 (13.4)	*.01*	−1.0 (−1.7 to −0.2)
Δ Self-efficacy[g]	16 (94)	18.3 (12.2)	16 (100)	−0.3 (5.8)	*<.001*	2.0 (1.1 to 2.8)
Parent						
Δ Dental fear and anxiety[c]	17 (100)	−10.88 (9.38)	16 (100)	0.9 (8.9)	*.001*	−1.3 (−2.0 to −0.5)
Δ Parental self-efficacy[h]	17 (100)	3.71 (13.68)	16 (100)	−3.8 (8.2)	.07	0.7 (−0.1 to 1.4)

[a]Group-by-time interaction effects.

[b]*P* values are based on group-by-time interaction effects of repeated-measure ANOVA.

[c]Children's Fear and Survey Schedule–Dental Subscale; score range: 15 to 75.

[d]Italicized values indicate significance.

[e]Children's Negative Cognitions in Dentistry scale; score range: 0 to 50.

[f]Injection Phobia Scale for Children; score range: 18 to 90.

[g]Self-Efficacy Questionnaire for Phobic Situations; score range: 14 to 70.

[h]Parental Self-Efficacy Questionnaire for Dental Anxiety; score range: 0 to 120.

Discussion

Principal Findings

After treatment, 41% (7/17) of the participants in the ICBT group no longer met the diagnostic criteria for either DP or IP, whereas all participants in the control group did. Furthermore, an additional 24% (4/17) of children in the ICBT group no longer met the diagnostic criteria for one of their 2 diagnoses at baseline, totaling 65% (11/17) of children in the ICBT group in whom at least one diagnosis had remitted. Compared with the control group, participants in the ICBT group could also willingly receive significantly more treatment steps at the dentist, and their fear, anxiety, and negative cognitions toward dentistry and injections significantly decreased after 12 weeks. As rated by both the children and parents, the effect size calculated from the primary outcome (PG-BAT) was large. In addition, children in the ICBT group reported higher self-efficacy. In summary, this RCT indicates that ICBT is an effective treatment for children with DP and IP.

Comparison With Prior Work

This is the first RCT of ICBT for DP and IP in pediatric dentistry. The effect sizes are in line with those of earlier studies on in-person exposure-based CBT methods in pediatric dentistry. Our research group previously conducted a nonrandomized pilot study using the same interventions as in this trial [22]. That study found within-group effect sizes between baseline and follow-up of 1.5 (95% CI 0.7-2.3) for the child-rated PG-BAT and 1.0 (95% CI 0.5-1.6) for the child-rated CFSS-DS; these are comparable with the within-group effect sizes in this study, which were 1.2 (95% CI 0.5-1.8) and 1.3 (95% CI 0.7-2.0), respectively. In the previous pilot study [22], ICBT participants were further improved at the 1-year follow-up, suggesting that the results of CBT for DP and IP are stable and can even increase over time.

An RCT of face-to-face CBT for dental anxiety in children and adolescents found strong between-group effects of Cohen *d*=1.4 after treatment and 1.9 at the 1-year follow-up for the main outcome of a clinician-administered Behavioral Approach Test [15]. After treatment, 64% of the patients in the CBT group no longer met the diagnostic criteria of their initial diagnosis, and at the 1-year follow-up, the proportion was 91%. This study used an active control group receiving care from specialist pediatric dentists, indicating that exposure-based CBT can yield improvements above and beyond those achieved with traditional care at specialist dental clinics for pediatric care. When calculated from the child-rated CFSS-DS, the within-group effect size of that study was 1.8 (95% CI 0.9-2.8) compared with a within-group effect size of 1.3 (95% CI 0.7-2.0) for the ICBT group in this study.

Finally, a research group in Norway conducted a 5-session CBT series for children and adolescents diagnosed with intraoral IP [16]. In that study, dentists specially trained in CBT carried out treatment, showing that exposure-based CBT can also be effectively implemented by dentists without the need for psychologists. During CBT treatment, 70% (47/67) of the participants were able to tolerate an injection; at the 1-year follow-up, 69% (34/49) were able to manage the required intraoral injections by their regular dentist according to dental records. As no effect sizes or correlations were reported, we calculated the Cohen *d* of the CFSS-DS from the pre- and posttreatment scores—33.8 (SD 9.7) and 23.9 (SD 6.7)—of the sample (N=57). The resulting effect size of 1.2 (95% CI 0.6-1.8)

XSL•FO
RenderX

is once again comparable with the within-group effects of 1.3 (95% CI 0.7-2.0) for the ICBT group in this study.

Taken together, there is an emerging base of evidence for exposure-based CBT in pediatric dentistry, and it seems that this type of treatment can be delivered effectively in several ways. An important avenue for future research is to investigate whether ICBT also produces effects superior to those of active control conditions and noninferiority compared with traditional face-to-face CBT.

Clinical Implications

One of the criteria for a phobic diagnosis is that it interferes significantly with a person's normal functioning. In the case of DP and IP, this usually means not being able to receive needed dental care, negatively affecting the person's dental and general health [17]. This study shows that ICBT can be of major clinical value in dentistry. ICBT lowers avoidance and fear, enabling children and adolescents to willingly receive dental care without the need for sedation or restraint methods; thus, the dental and general health of children and adolescents who previously avoided dental care because of DP or IP will improve over the long term.

An important clinical aspect of this study was that dental treatment and in vivo exposure were provided by general dental personnel, many times dental assistants, with most not having any association with the research team and having no previous education in the method or specialist training. Earlier studies have also used exposure-based CBT provided by trained dental personnel [11,12,16] face-to-face. This requires a shift in roles from their traditional dental practices to providing exposure therapy, with communication style and time viewed as the most important factors for the treatment to be successful [30]. However, our results suggest that, when ICBT is implemented, it is clinically feasible to only briefly instruct the dentist on how to conduct exposure with patients and their parents, and this information can then be conveyed to the dental assistant and implemented sufficiently well from a therapeutic perspective. We consider this important as it indicates that exposure-based ICBT can be integrated into routine dental care practices and, thus, has the potential to be effectively disseminated while also requiring less time invested by the dental personnel as important parts of the treatment are provided on the web-based platform. In addition, participants reported no adverse events, indicating that ICBT is safe even when in vivo exposure is conducted by general dental personnel with no training in CBT and at home by the patients and their parents themselves.

Previous research has indicated a need for greater use of evidence-based CBT methods with children and adolescents in dentistry [13]. ICBT has been shown to be as effective as face-to-face treatment in many settings and for different conditions [18]. More specifically, ICBT has been shown to be effective for specific phobias [20]. This study showed that ICBT is also an effective treatment that can be used for children with DP or IP in dentistry. As no other evidence-based treatment for children or adolescents with DP or IP in dentistry currently exists, ICBT could fill a gap by providing an evidence-based treatment for the field of dentistry if disseminated correctly. ICBT can potentially be administered to more patients at a lower cost, thus overcoming the hurdle of too few CBT-trained clinicians. Furthermore, ICBT is more accessible than face-to-face CBT for patients as ICBT allows them to freely choose treatment times. This also extends to adult populations. Currently, no trials of ICBT for adults have been conducted, but we find it highly probable that ICBT for adults will be as effective as it is for children and adolescents.

Dentists may hesitate to recommend CBT treatment to their patients, which the slow recruitment in this study suggests. Previous studies have also shown that recruitment for CBT trials in dentistry is difficult [31]. This might be partly owing to a lack of interest by general practicing dentists to refer patients for CBT, which demonstrates the need to make CBT treatment attractive to dental clinicians. Consequently, one aspect that we believe is key for successful dissemination of ICBT is to increase interest in general dentistry for this type of treatment; in this way, families can be referred to and recommended exposure-based treatments when motivation is likely to be high (ie, shortly after a dental visit in which the child expressed marked fear).

Finally, this study excluded participants with neurodevelopmental disorders, a group of patients that is common in specialist pediatric dentistry. Future research needs to explore how exposure-based treatments can be successfully adapted for these patients.

Strengths and Limitations

Important strengths of this study were the randomized controlled design, absence of missing data, low treatment dropout rate and generally high treatment adherence, and use of both validated self-rating scales and clinician assessments, which addressed methodological problems observed in previous research [1]. This study also had high external validity as the dental treatment and in vivo exposure were administered at general dental clinics by personnel with no formal training in CBT. Finally, the sample was clinically relevant, having had fears of visiting the dentist for an extended time and previous experience of unsuccessful treatment to reduce their DFA or DP and IP through sedation, restraint, or general anesthesia.

Regarding limitations, the control group in this study was a waiting list. We initially hoped to establish the efficacy of ICBT treatment before testing it against active control conditions or establish noninferiority by comparing it with face-to-face CBT. The blinding of the clinicians who performed the follow-up interview was not assessed in this study. In addition, the study had a fairly small sample size, which precludes, for example, meaningful subgroup analyses. Furthermore, we used the IP diagnosis instead of intraoral IP in this study. Although it was required that the IP affect the dental treatment of participants when entering the study, it would have been preferable to use the intraoral IP diagnosis. This is because, in some cases, after the treatment, we had patients who no longer met the criteria for intraoral IP diagnosis but still met the criteria for IP diagnosis (ie, being able to receive intraoral injections in dentistry but not intramuscular vaccine injections). However, this only attenuated our results, which would have been even stronger if an intraoral IP diagnosis had been used.

XSL•FO
RenderX

Another limitation of this study is that recruitment through advertisements might not reach the general population, creating a nonrepresentative group. This could be another factor behind the slow recruitment, suggesting treatment barriers for certain groups that warrant further investigation. Speculatively, some families may have too few resources or be reluctant to commit to a program as long as 12 weeks. This is further exemplified by one participant in the study who dropped out because of time constraints and 3 participants who did not complete the modules during the treatment time. The extra burden of investing time in a treatment of this type makes it unsuitable for some families and individuals. This limits the use of both CBT and ICBT in its current form, and more research is needed on how to provide a feasible, evidence-based alternative for this group. Perhaps one alternative that should be tested is an even shorter course in exposure-based CBT that is less text dependent.

Finally, this study was conducted in Sweden, which has publicly funded, free-of-charge dental services for children. The generalizability of the findings to other dental health care contexts needs to be further investigated. More specifically, how the cost of this type of treatment might influence its acceptability for the patient and dentist in different contexts needs to be explored.

Conclusions

ICBT for DP or IP seems to be an effective treatment for children and adolescents. This therapy reduces fear and enables the child to willingly receive dental treatment, thereby leading to improved dental and general health. ICBT should be considered a method that increases accessibility to effective psychological treatment in pediatric dentistry.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Content and photographs of the exposure toolkit and visual analog scale.
[DOCX File , 56 KB - jmir_v26i1e42322_app1.docx]

Multimedia Appendix 2
Screenshot from internet-based cognitive behavioral therapy (module with text).
[PNG File , 442 KB - jmir_v26i1e42322_app2.png]

Multimedia Appendix 3
Screenshot from internet-based cognitive behavioral therapy (module with text and picture).
[PNG File , 401 KB - jmir_v26i1e42322_app3.png]

Multimedia Appendix 4
Screenshot from internet-based cognitive behavioral therapy (home assignment at end of module).
[PNG File , 277 KB - jmir_v26i1e42322_app4.png]

Multimedia Appendix 5
Scores of the children and their parents in the internet-based cognitive behavioral therapy (n=17) and control (n=16) groups at the follow-up.
[DOCX File , 16 KB - jmir_v26i1e42322_app5.docx]

Multimedia Appendix 6
Within-group changes and effect sizes between baseline and treatment follow-up for the internet-based cognitive behavioral therapy group.
[DOCX File , 16 KB - jmir_v26i1e42322_app6.docx]

Multimedia Appendix 7
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 150 KB - jmir_v26i1e42322_app7.pdf]

References

XSL•FO
RenderX

1. Seligman LD, Hovey JD, Chacon K, Ollendick TH. Dental anxiety: an understudied problem in youth. Clin Psychol Rev 2017 Jul;55:25-40. [doi: 10.1016/j.cpr.2017.04.004] [Medline: 28478271]
2. Grisolia BM, Dos Santos AP, Dhyppolito IM, Buchanan H, Hill K, Oliveira BH. Prevalence of dental anxiety in children and adolescents globally: a systematic review with meta-analyses. Int J Paediatr Dent 2021 Mar;31(2):168-183. [doi: 10.1111/ipd.12712] [Medline: 33245591]
3. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th Edition. Arlington, VA: American Psychiatric Publishing; 2013.
4. Wani AL, Ara A, Bhat SA. Blood injury and injection phobia: the neglected one. Behav Neurol 2014;2014:471340-471347 [FREE Full text] [doi: 10.1155/2014/471340] [Medline: 25049451]
5. Berg Johnsen I, Wichstrøm L, Dahllöf G. Prevalence and stability of blood-injection-injury phobia in childhood-a prospective community study in Norway. Acta Paediatr 2024 Jan;113(1):105-112. [doi: 10.1111/apa.17003] [Medline: 37850719]
6. Armfield JM. Towards a better understanding of dental anxiety and fear: cognitions vs. experiences. Eur J Oral Sci 2010 Jun;118(3):259-264. [doi: 10.1111/j.1600-0722.2010.00740.x] [Medline: 20572859]
7. Mejàre IA, Klingberg G, Mowafi FK, Stecksén-Blicks C, Twetman SH, Tranæus SH. A systematic map of systematic reviews in pediatric dentistry--what do we really know? PLoS One 2015;10(2):e0117537 [FREE Full text] [doi: 10.1371/journal.pone.0117537] [Medline: 25706629]
8. Davey GC. Dental phobias and anxieties: evidence for conditioning processes in the acquisition and modulation of a learned fear. Behav Res Ther 1989;27(1):51-58. [doi: 10.1016/0005-7967(89)90119-8] [Medline: 2914005]
9. Berggren U, Carlsson SG, Hägglin C, Hakeberg M, Samsonowitz V. Assessment of patients with direct conditioned and indirect cognitive reported origin of dental fear. Eur J Oral Sci 1997 Jun;105(3):213-220 [FREE Full text] [doi: 10.1111/j.1600-0722.1997.tb00203.x] [Medline: 9249187]
10. Öst L, Skaret E. Cognitive Behavioral Therapy for Dental Phobia and Anxiety. Chichester, UK: John Wiley & Sons; 2013.
11. Hauge MS, Stora B, Willumsen T. Dental anxiety treatment by a dentist in primary care: a 1-year follow-up study. Eur J Oral Sci 2022 Aug;130(4):e12872 [FREE Full text] [doi: 10.1111/eos.12872] [Medline: 35569119]
12. Haukebø K, Skaret E, Ost LG, Raadal M, Berg E, Sundberg H, et al. One- vs. five-session treatment of dental phobia: a randomized controlled study. J Behav Ther Exp Psychiatry 2008 Sep;39(3):381-390. [doi: 10.1016/j.jbtep.2007.09.006] [Medline: 18005932]
13. Gomes H, Viana KA, Batista AC, Costa LR, Hosey MT, Newton T. Cognitive behaviour therapy for anxious paediatric dental patients: a systematic review. Int J Paediatr Dent (Forthcoming) 2018 Jul 08 [FREE Full text] [doi: 10.1111/ipd.12405] [Medline: 29984460]
14. Davis TE, Ollendick TH, Ost LG. Intensive treatment of specific phobias in children and adolescents. Cogn Behav Pract 2009 Aug 01;16(3):294-303 [FREE Full text] [doi: 10.1016/j.cbpra.2008.12.008] [Medline: 20161063]
15. Shahnavaz S, Hedman E, Grindefjord M, Reuterskiöld L, Dahllöf G. Cognitive behavioral therapy for children with dental anxiety: a randomized controlled trial. JDR Clin Trans Res 2016 Oct;1(3):234-243 [FREE Full text] [doi: 10.1177/2380084416661473] [Medline: 29417092]
16. Berge KG, Agdal ML, Vika M, Skeie MS. Treatment of intra-oral injection phobia: a randomized delayed intervention controlled trial among Norwegian 10- to 16-year-olds. Acta Odontol Scand 2017 May;75(4):294-301. [doi: 10.1080/00016357.2017.1297849] [Medline: 28270029]
17. Shahnavaz S. Cognitive behavioral therapy for children and adolescents with dental phobia. Karolinska Institute. 2016. URL: https://openarchive.ki.se/xmlui/handle/10616/45319 [accessed 2024-01-11]
18. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18 [FREE Full text] [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
19. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215 [FREE Full text] [doi: 10.1016/j.invent.2014.08.003]
20. Mor S, Grimaldos J, Tur C, Miguel C, Cuijpers P, Botella C, et al. Internet- and mobile-based interventions for the treatment of specific phobia: a systematic review and preliminary meta-analysis. Internet Interv 2021 Dec;26:100462 [FREE Full text] [doi: 10.1016/j.invent.2021.100462] [Medline: 34646752]
21. Vigerland S, Ljótsson B, Thulin U, Öst LG, Andersson G, Serlachius E. Internet-delivered cognitive behavioural therapy for children with anxiety disorders: a randomised controlled trial. Behav Res Ther 2016 Jan;76:47-56 [FREE Full text] [doi: 10.1016/j.brat.2015.11.006] [Medline: 26649465]
22. Shahnavaz S, Hedman-Lagerlöf E, Hasselblad T, Reuterskiöld L, Kaldo V, Dahllöf G. Internet-based cognitive behavioral therapy for children and adolescents with dental anxiety: open trial. J Med Internet Res 2018 Jan 22;20(1):e12 [FREE Full text] [doi: 10.2196/jmir.7803] [Medline: 29358158]
23. Klingberg G. Reliability and validity of the Swedish version of the dental subscale of the children's fear survey schedule, CFSS-DS. Acta Odontol Scand 1994 Aug;52(4):255-256. [doi: 10.3109/00016359409029055] [Medline: 7985512]
24. Wahlund T. Injection phobi scale for children. Stockholms Universitet Psykologiska Institutionen. 2008. URL: https://www.diva-portal.org/smash/get/diva2:199342/FULLTEXT01.pdf [accessed 2024-01-11]

XSL•FO
RenderX

25. Goodman R, Ford T, Richards H, Gatward R, Meltzer H. The development and well-being assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. J Child Psychol Psychiatry 2000 Jul;41(5):645-655. [Medline: 10946756]
26. Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 1997 Jul;36(7):980-988 [FREE Full text] [doi: 10.1097/00004583-199707000-00021] [Medline: 9204677]
27. Flatt N, King N. The Self-Efficacy Questionnaire for Phobic Situations (SEQ-SP): development and psychometric evaluation. Behav Change 2012 Feb 22;26(2):141-152 [FREE Full text] [doi: 10.1375/bech.26.2.141]
28. List randomizer. random.org. URL: https://www.random.org/lists/ [accessed 2024-01-29]
29. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Mahwah, NJ: Lawrence Erlbaum Associates; 1988.
30. Bryne E, Hean S, Evensen K, Bull V. More than just a dental practitioner: a realist evaluation of a dental anxiety service in Norway. Eur J Oral Sci 2021 Dec;129(6):e12820. [doi: 10.1111/eos.12820] [Medline: 34448277]
31. Wide Boman U, Broberg AG, Krekmanova L, Staberg M, Svensson C, Robertson A. An explorative analysis of the recruitment of patients to a randomised controlled trial in adolescents with dental anxiety. Swed Dent J 2014;38(1):47-54. [Medline: 26995811]

Abbreviations

CBT: cognitive behavioral therapy
CFSS-DS: Children's Fear Survey Schedule–Dental Subscale
CNCD: Children's Negative Cognitions in Dentistry
DFA: dental fear and anxiety
DP: dental phobia
ICBT: internet-based cognitive behavioral therapy
IP: injection phobia
IPSC: Injection Phobia Scale for Children
K-SADS-PL: Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime
PG-BAT: Picture-Guided Behavioral Avoidance Test
RCT: randomized controlled trial
VAS: visual analog scale

Edited by A Mavragani; submitted 12.02.23; peer-reviewed by JA Johnsen, S Hughes; comments to author 06.10.23; revised version received 30.10.23; accepted 19.12.23; published 21.02.24.

Please cite as:
Schibbye R, Hedman-Lagerlöf E, Kaldo V, Dahllöf G, Shahnavaz S
Internet-Based Cognitive Behavioral Therapy for Children and Adolescents With Dental or Injection Phobia: Randomized Controlled Trial
J Med Internet Res 2024;26:e42322
URL: https://www.jmir.org/2024/1/e42322
doi:10.2196/42322
PMID:38381476

©Robert Schibbye, Erik Hedman-Lagerlöf, Viktor Kaldo, Göran Dahllöf, Shervin Shahnavaz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of Internet-Based Cognitive Behavioral Therapy for Subthreshold Depression Among Older Adults in Institutional Long-Term Care Settings: Pragmatic Randomized Controlled Trial

Fanqian Kong[1], MSc; Libo Yu[2], PhD; Yanbin Hou[3], MD; Lijie Zhu[4], MSc; Jing Zhou[2], MSc; Lingjie Huang[2], MPH; Yueer Lv[2], MPH; Li Wang[2], MSc; Li Zhang[2], MSc; Yiling Yang[2], MSc; Yuchen Ying[2], MPH

[1]Ningbo Medical Center Lihuili Hospital, Ningbo, China
[2]Ningbo College of Health Sciences, Ningbo, Zhejiang, China
[3]Ningbo First Hospital, Ningbo, China
[4]Pfizer Research and Development Co, Ltd, Shanghai, China

Corresponding Author:
Yuchen Ying, MPH
Ningbo College of Health Sciences
51 Xuefu Street
Yinzhou District
Ningbo, Zhejiang, 315100
China
Phone: 86 13685733802
Email: 1246401741@qq.com

Abstract

Background: Subthreshold depression (sD) is prevalent in older populations in long-term care (LTC) settings, but psychological therapy in LTC settings in China is not readily available. Thus, internet-based cognitive behavioral therapy (ICBT) may be suitable for this population, but research on the efficacy of ICBT for older adults with sD, especially those living in LTC settings, is limited.

Objective: This study aimed to evaluate the efficacy and acceptability of ICBT treatment for sD among LTC residents in China. We also examined whether ICBT is as effective as group-based cognitive behavioral therapy (CBT) for treating sD in this population.

Methods: We conducted a pragmatic randomized controlled trial, which included 18 LTC institutions. A total of 354 participants were randomized to ICBT, group-based CBT, or a waiting list and were followed up for 12 months. The primary outcome was self-reported depressive symptoms on the Center for Epidemiological Studies Depression Scale (CES-D). Secondary outcomes were the scores of the Patient Health Questionnaire–9 (PHQ-9), Generalized Anxiety Disorder 7-Item (GAD-7), and Geriatric Depression Scale (GDS). A mixed-effects model was used to assess the efficacy of ICBT.

Results: The ICBT group showed a significant improvement in self-reported depressive symptoms, which was maintained at the 12-month follow-up (all $P<.001$). The ICBT group exhibited a significantly larger reduction in the scores of the CES-D (Cohen $d=0.07$, 95% CI 0.04-0.09; $P=.01$), PHQ-9 ($d=0.30$, 95% CI 0.28-0.33; $P<.001$), GDS ($d=0.10$, 95% CI 0.08-0.13; $P<.001$), and GAD-7 ($d=0.19$, 95% CI 0.17-0.22; $P<.001$) compared with a waiting list at postintervention. ICBT had significantly stronger effects than CBT on the PHQ-9 and GAD-7 at postintervention ($d=0.15$, 95% CI 0.13-0.17; $P<.001$ and $d=0.21$, 95% CI 0.19-0.23; $P<.001$, respectively), 6-month follow-up ($d=0.18$, 95% CI 0.16-0.21; $P<.001$ and $d=0.18$, 95% CI 0.15-0.21; $P<.001$, respectively), and 12-month follow-up ($d=0.15$, 95% CI 0.11-0.19; $P<.001$ and $d=0.18$, 95% CI 0.14-0.21; $P<.001$, respectively).

Conclusions: ICBT is a relatively effective and acceptable intervention for reducing depressive symptoms among Chinese LTC residents with sD. These findings indicate the usefulness of ICBT application for sD in LTC settings.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000030697; https://www.chictr.org.cn/showproj.aspx?proj=50781

(J Med Internet Res 2024;26:e40187) doi:10.2196/40187

KEYWORDS

internet-based cognitive behavioral therapy; subthreshold depression; randomized controlled trial; older adults; LTC setting

XSL•FO
RenderX

Introduction

Subthreshold depression (sD) is a significant health burden for older adults [1]. It is even more prevalent than diagnosed clinical depression among older adults [2] and adversely affects their social functions [3], activities of daily living [4], and quality of life [5]. In addition, it is related to increased risk of major depression [6], cognitive impairment and dementia [7], physical ailments [8], and increased mortality and medical costs in older adults [9]. Furthermore, epidemiological research on older adults with sD has revealed a pattern: increasing rates of moving from community to primary care to long-term care (LTC) settings [10]. Therefore, it is necessary to develop appropriate interventions to treat sD among older adults, especially those living in LTC settings.

With respect to the treatment of depression, older adults prefer psychotherapy to pharmacotherapy [11]; cognitive behavioral therapy (CBT) has proven effective [12]. However, the treatment rate for older adults is low due to numerous barriers to face-to-face CBT, including mobility limitations, relatively high medical costs, limited number of trained therapists, and stigma [13]. Fortunately, internet-based CBT (ICBT), an intervention that can be conducted anytime and anywhere over the internet, can overcome these barriers [14]. Nevertheless, although several studies have demonstrated that ICBT may significantly reduce depressive symptoms in older adults with moderate to severe depression [15], the impact of ICBT on older adults with sD is unclear. In addition, although preliminary evidence demonstrated that guided ICBT and face-to-face CBT might be equally effective [16], only 1 prior study has demonstrated that ICBT is more effective than face-to-face CBT in treating sD among older adults living in the community [17]. However, this study did not provide a between-group effect size on the outcomes of ICBT versus CBT, and, therefore, whether ICBT is as effective as face-to-face CBT among LTC residents with sD remains unknown.

Many older Chinese adults are unlikely to seek face-to-face psychotherapy due to stigma that discourages disclosure influenced by Chinese cultural values; thus, undertreatment is common among older adults in China [18]. Of note, the overall detection rate of mild depressive symptoms among Chinese LTC residents is only 30% [19]. Considering China's 2.15 million LTC residents and the extreme shortage of its mental health workforce, especially psychological therapists in LTC settings [20], many cases of sD might go undetected and untreated. To address this issue, the Chinese government released an action plan for depression prevention and treatment to improve access to web interventions (eg, ICBT) for depression among older adults [21]. However, the efficacy and acceptability of ICBT among the older adult Chinese population, especially those in LTC settings, have not been investigated. Therefore, evidence-based research is needed to promote the use of ICBT in LTC settings in China. LTC settings in China are broadly divided into two categories: (1) institutional settings and (2) home- and community-based settings [22]. This study focused on institutional LTC facilities because they have a somewhat similar definition to that of LTC settings in previous studies [10].

The purpose of this study was to test the following hypotheses: (1) ICBT is more effective than a waiting list (WL) in improving depressive symptoms and other secondary outcomes, (2) ICBT is equally effective as group CBT in improving depressive symptoms and other secondary outcomes, and (3) ICBT is effective and acceptable for Chinese LTC residents.

Methods

Study Design

This study was a pragmatic, multicenter randomized controlled trial (RCT) with 3 arms comparing clinician-guided ICBT with group CBT and a WL control. The patients were assessed at baseline, immediately after the intervention, and at 6- and 12-month follow-ups.

The interventions and WL control were not compared at 6- and 12-month follow-ups because the WL group had only 2 assessment points (ie, baseline and postintervention).

Ethical Considerations

This study was approved by the Medical Ethics Committee of Ningbo University (approval number: NBU-2020-139) and registered in the Chinese Clinical Trials Registry (registration ChiCTR2000030697).

Participants and Procedures

A total of 618 participants were recruited via 18 LTC settings across 6 cities in Zhejiang Province. The choice of the cities and LTC settings ensured that the sample included participants from locations with different population compositions.

To protect the residents' privacy, initial contact was made by the designated research liaison staff member at each LTC setting. This person asked about the residents' willingness to participate in the study, examined their health records, and identified eligible participants based on our screening criteria. The inclusion criteria were as follows: (1) be a resident of China, (2) be 60 years or older, (3) have access to the internet, (4) have sD (defined as having a score of ≥16 on the Centre for Epidemiological Studies Depression Scale [CES-D], according to the widely used criteria in comparable studies [6,23]), (5) no suicidal ideation (indicated by a score of <2 on question 9 on the Patient Health Questionnaire–9 [PHQ-9]) [24], (6) no cognitive impairment with a score on a Mini-Mental State Examination >24 [25], (7) not currently undergoing psychotherapy, (8) no alcohol dependency, and (9) no terminal illness on clinical grounds.

The staff invited potential eligible participants to join the study after obtaining written informed consent. These potential participants were further invited for the telephone-administered Semistructured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) Axis Disorders (SCID) to assess final eligibility. The exclusion criteria were as follows: meeting the DSM-IV criteria for (1) major depressive disorder (MDD), (2) bipolar disorder, or (3) psychotic disorder. Participants on antidepressants remained eligible if they were on a stable dose for the past 4 weeks and had no change in medication during the study. New residents who met the inclusion criteria could also be invited.

XSL•FO
RenderX

Standardized questionnaires including all measures and sociodemographic data were administered to eligible participants. Of the recruited participants, 354 met the inclusion criteria. The participants were stratified by age, sex, study site, use of antidepressants, and depression symptom severity (determined using CES-D scores) and then randomized to ICBT (n=118), CBT (n=119), or WL (n=117). Figure 1 depicts the flow of participants through the trial.

Figure 1. Flowchart of participants. CBT: cognitive behavioral therapy; CES-D: Center for Epidemiological Studies Depression Scale; ICBT: internet-based cognitive behavioral therapy.

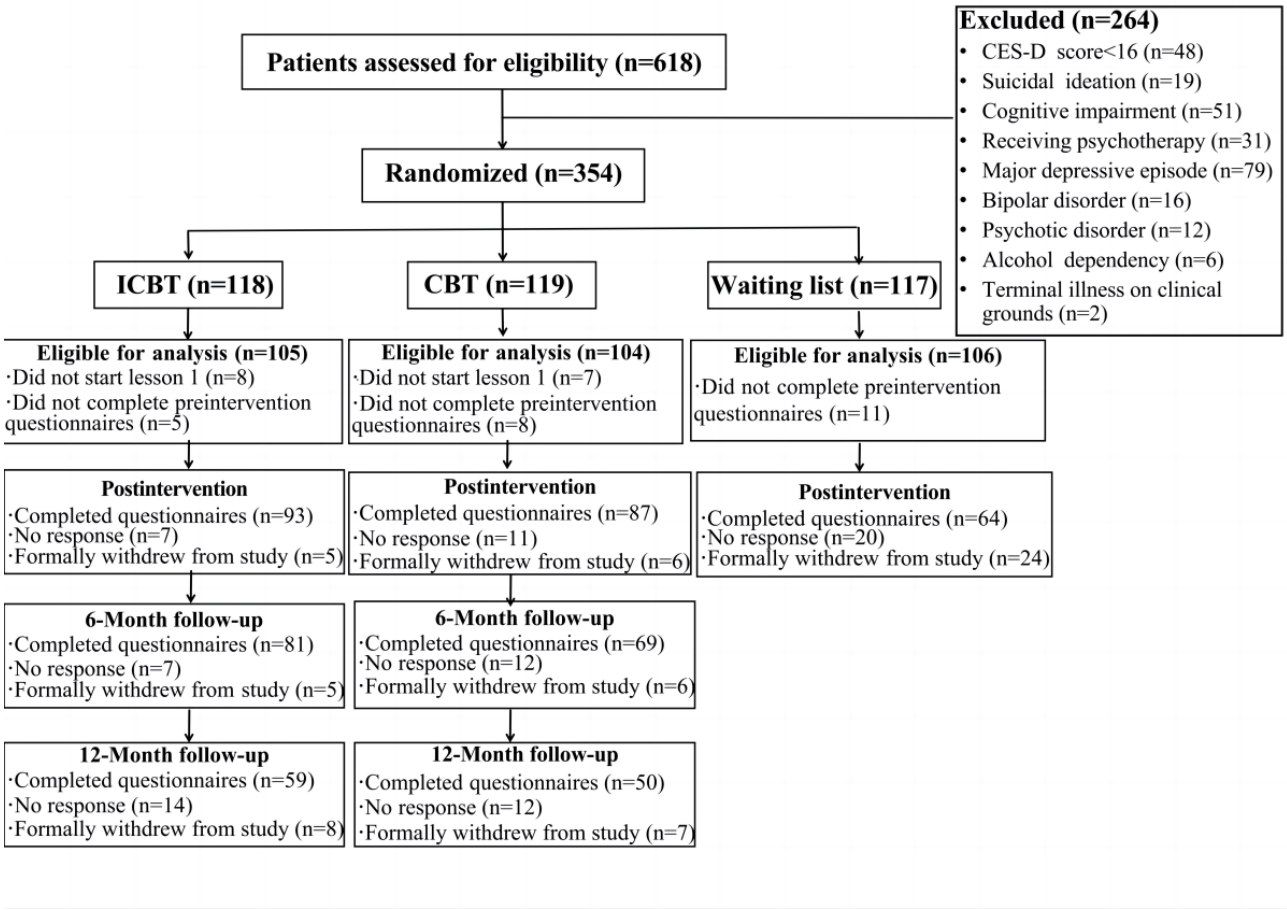

Randomization was performed at a ratio of 1:1:1 in blocks of 6 using a web randomizer by an external researcher from another institution. Each allocation was kept in sequentially numbered, sealed opaque envelopes. Clinical psychologists conducting SCID were blinded to the group assignments. Except for the WL group as a control, none of the participants were blinded to group allocation. Participants who completed postintervention and follow-up questionnaires received CN ¥75 (approximately US $11.80).

Interventions

Clinician-Guided ICBT

This clinician-guided ICBT is based on the principle of the Chinese version of the *Handbook of Cognitive-Behavioral Therapies* [26]. In this trial, the depression module was used. This module comprises 5 lessons with corresponding case stories, which were developed from real cases of Chinese patients who recovered from sD or MDD after receiving CBT, and was designed to educate the general Chinese population about practical CBT skills for depression [27]. Textbox 1 depicts each lesson's descriptions. Although the duration of lessons could vary among participants, they were advised to undertake at least 1 lesson per week in an individual format. The role of clinicians, who were blinded to the hypotheses, was to provide technical assistance and encouragement to participants during the study.

XSL•FO
RenderX

Textbox 1. Lesson content of the depression module.

Lesson 1

- Lesson content

Getting support, building relationship, and signing contracts

Introduction to listening, empathy, active attention, depression, cognitive behavioral therapy, psychological education, etc

Understanding, relationship building, and cognitive assessment (preliminary establishment of case conceptualization and treatment plan)

- Case story
  - Instructions and examples of building a relationship and signing contracts with clinicians
  - Instructions and examples of assessing own symptoms of depression
  - Instructions and examples of establishing cognitive case conceptualization and treatment plan

Lesson 2

- Lesson content
  - Introduction to the psychological meaning of symptoms of depression
  - Introduction to the role of cognition (cognitive recognition, behavioral function analysis, cognitive behavior recording, reality test, etc)
- Case story
  - Instructions and examples of cognitive recognition, behavioral function analysis, cognitive behavior recording, and reality test

Lesson 3

- Lesson content
  - Introduction to the relationship between cognition and emotion
  - Introduction to cognitive reconstruction and behavioral training (behavioral activation, daily activity arrangement, etc)
- Case story
  - Instructions and examples of cognitive reconstruction and behavioral training

Lesson 4

- Lesson content
  - Introduction to the cognition, discussion, and reconstruction of basic beliefs
  - Further introduction to behavioral training (further case conceptualization of cognition, vertical analysis of behavior, problem-solving strategies, further reconstruction of cognition and behavioral training, cognitive flash cards and cognitive motto, testing basic beliefs, etc)
- Case story
  - Instructions and examples of the reconstruction of basic beliefs and behavioral training

Lesson 5

- Lesson content
  - Further introduction to the cognition, reconstruction of basic beliefs (consolidation of positive cognition, confirming and modifying basic beliefs, etc)
  - Further introduction to behavioral training (continued behavior training, etc)
  - Introduction to relapse prevention and constructing relapse prevention plans
  - A summary of the 5 lessons and future behavior plan, etc
- Case story
  - Instructions and examples of the reconstruction of basic beliefs and behavior training
  - Instructions and examples of relapse prevention and relapse prevention plans
  - Instructions and examples of the summary and future plan

XSL•FO
RenderX

Group-Based Face-to-Face CBT

CBT participants received similar treatment modules in the same chronological order as the ICBT participants to ensure that both groups were provided with the same treatment components. There were 5 weekly sessions, each lasting 1-1.5 hours. Two experienced licensed psychologists in each LTC setting, who were blinded to the hypotheses, delivered the interventions. They were rigorously trained on the entire CBT procedure by the research clinicians before the study began to ensure that participants from different settings were treated consistently. All the group CBT sessions were conducted in a private room within the LTC setting.

WL Condition

Participants randomized to the WL condition were informed that they would receive an intervention within a few weeks, while no further contact or information was provided. They would choose the CBT or ICBT intervention independently after the test.

Outcomes

The primary outcome was depressive symptoms measured by the Chinese version of the CES-D (Cronbach α=.88 in this study) [28]. Secondary measures included the Chinese version of the Geriatric Depression Scale (GDS; Cronbach α=.92 in this study) [29] and PHQ-9 (Cronbach α=.81 in this study) [24], which assess depressive symptoms. Another secondary outcome was anxiety symptoms on the Chinese version of the Generalized Anxiety Disorder 7-Item (GAD-7; Cronbach α=.83 in this study) [30]. These measurement tools have shown good reliability and validity among Chinese populations [31-34]. To assess the participants' satisfaction with and acceptability of the ICBT, they were asked to respond with either "yes" or "no" to the following two questions: (1) "are you confident to recommend this program?" and (2) "was this program worth your time?"

Statistical Analysis

All tests were 2-sided, and a *P* value of <.05 was considered statistically significant. Differences between the groups on baseline characteristics were compared using either 1-way ANOVA or Kruskal-Wallis tests for continuous variables and χ^2 tests for categorical variables. All analyses were performed using SAS version 9.4 (SAS Institute).

Participants in the 2 intervention groups who did not start lesson 1 were excluded from the analyses. The least-squares means of outcomes were estimated based on an intention-to-treat basis, performing a mixed-effects model for repeated measures [35], which contained the time, treatment group, and time×group interaction as fixed effects. Effects were further modeled using the restricted maximum likelihood and an unstructured covariance matrix, including demographic and baseline characteristics as covariates. Mean between-group differences and mean changes for each group at each time point were analyzed using appropriate contrasts in the mixed-effects model for repeated measures.

We evaluated the effect sizes (Cohen *d*) using the criteria proposed for mixed model analysis [36-38] and calculated CIs based on the method suggested by Kelley [39]. A Cohen *d* of 0.2-0.49 was considered a small effect size, 0.5-0.79 was considered a medium effect size, and 0.8 or more was considered a large effect size [40].

We also calculated reliable change and clinically significant improvement on the CSE-D from baseline to postintervention, as well as 6- and 12-month follow-ups, according to Jacobson and Truax [41]. Reliable change was defined as a decrease of at least 3 points (using the criterion that the reliable change index must be >1.96 and using retest reliability reported for the CES-D [0.88] in the reliable change index formula) on the CES-D score, plus a CES-D postintervention score below the clinical cutoff of 16 to achieve clinically significant improvement. Next, we used a χ^2 test to assess the frequency differences in reliable change and clinically significant improvements between the 3 groups. We also calculated the number needed to treat with 95% CIs to achieve 1 additional reliable change and clinically significant improvement at postintervention, relative to the WL condition [42].

Results

Participants' Demographic Characteristics

Table 1 shows the demographic characteristics of the participants. Demographic characteristics and outcomes were similar for the 3 groups at baseline. Overall, the mean (SD) age was 78.60 (7.80) years. Most participants were female (215/315, 68.3%), were married (270/315, 85.7%), were graduates of junior high school or lower education (281/315, 89.2%), had low-to-medium income (298/315, 94.6%), and had no histories of MDD (257/315, 81.6%) or therapy for MDD (202/315, 64.1%). Participants were evenly recruited from public social welfare facilities (104/315, 33%), nursing homes (99/315, 31.4%), and other residential care facilities evenly (112/315, 35.6%).

XSL•FO
RenderX

Table 1. Demographic characteristics of the participants.

Participant characteristics	Total (N=315)	CBT[a] (n=104)	ICBT[b] (n=105)	WL[c] (n=106)	*P* value
Sex, n (%)					.08
Male	100 (31.7)	30 (28.8)	42 (40)	28 (26.4)	
Female	215 (68.3)	74 (71.2)	63 (60)	78 (73.6)	
Age (years), mean (SD)	71.6 (7.80)	71.8 (7.97)	71.2 (8.31)	71.9 (6.79)	.92
Study site, n (%)					.20
Public social welfare facilities	104 (33)	38 (36.5)	30 (28.6)	36 (34)	
Nursing homes	99 (31.4)	33 (31.7)	30 (28.6)	36 (34)	
Other residential care facilities	112 (35.6)	33 (31.7)	45 (42.9)	34 (32.1)	
Marital status, n (%)					.98
Unmarried	8 (2.5)	3 (2.9)	2 (1.9)	3 (2.8)	
Married	270 (85.7)	89 (85.6)	91 (86.7)	90 (84.9)	
Divorced or widowed	37 (11.7)	12 (11.5)	12 (11.4)	13 (12.3)	
Monthly income, n (%)					.98
Low	148 (47)	51 (49)	49 (46.7)	48 (45.3)	
Medium	150 (47.6)	46 (44.2)	51 (48.6)	53 (50)	
High	17 (5.4)	7 (6.7)	5 (4.8)	5 (4.7)	
Education, n (%)					.53
Primary school or below	155 (49.2)	54 (51.9)	51 (48.6)	50 (47.2)	
Junior high school	126 (40)	39 (37.5)	46 (43.8)	41 (38.7)	
Senior high school	21 (6.7)	7 (6.7)	5 (4.8)	9 (8.5)	
College or above	13 (4.1)	4 (3.8)	3 (2.9)	6 (5.7)	
History of MDD[d], n (%)					.44
No	257 (81.6)	85 (81.7)	82 (78.1)	90 (84.9)	
Yes	58 (18.4)	19 (18.3)	23 (21.9)	16 (15.1)	
Previous therapy for MDD, n (%)					.53
None	202 (64.1)	73 (70.2)	60 (57.1)	69 (65.1)	
Antidepressant	49 (15.6)	11 (10.6)	23 (21.9)	15 (14.2)	
Psychotherapy	44 (14)	14 (13.5)	16 (15.2)	14 (13.2)	
Both antidepressant and psychotherapy	20 (6.3)	6 (5.8)	6 (5.7)	8 (7.5)	

[a]CBT: cognitive behavioral therapy.

[b]ICBT: internet-based cognitive behavioral therapy.

[c]WL: waiting list.

[d]MDD: major depressive disorder.

Main Outcomes

Table 2 depicts the baseline, postintervention, 6-month follow-up, and 12-month follow-up least-squares mean and SD of outcomes for each group. The mixed model for repeated measures analyses revealed significant time×group interactions across the primary and secondary outcomes (CES-D: $F_{2,205}$=3.37, P=.04; PHQ-9: $F_{2,220}$=24.33, P<.001; GDS: $F_{2,199}$=13.80, P<.001; GAD-7: $F_{2,219}$=25.86, P<.001).

XSL•FO
RenderX

Table 2. Mean, SD, and effect sizes (Cohen *d*) of outcomes for study conditions.

Outcome and time point	CBT[a], mean (SD)	ICBT[b], mean (SD)	WL[c], mean (SD)	Time × group, *F* test (*df*)	Significant contrasts between 2 groups	Within-group effect size (95% CI)			Between-group effect size (95% CI)		
						CBT	ICBT	WL	ICBT vs CBT	CBT vs WL	ICBT vs WL
CES-D[d]				3.37 (2, 205)[h]							
Baseline	18.8 (4.15)	19.1 (4.44)	19.4 (4.37)		__[j]	—	—	—	—	—	—
Postintervention	15.6 (3.87)	14.4 (4.26)	15.8 (3.76)		WL<ICBT[h]	0.16 (0.14 to 0.19)	0.42 (0.40 to 0.44)	0.15 (0.13 to 0.18)	0.00 (–0.02 to 0.02)	0.05 (0.02 to 0.07)	0.07 (0.04 to 0.09)
6-month follow-up	15.8 (3.63)	14.4 (4.05)	—		—	0.17 (0.14 to 0.19)	0.43 (0.41 to 0.45)	—	0.03 (0.00-0.05)	—	—
12-month follow-up	16.3 (3.40)	14.6 (3.71)	—		—	0.13 (0.11 to 0.16)	0.30 (0.28 to 0.33)	—	0.05 (0.01 to 0.08)	—	—
PHQ-9[e]				24.33 (2, 220)[i]							
Baseline	8.0 (5.10)	8.4 (5.48)	8.1 (5.36)		—	—	—	—	—	—	—
Postintervention	6.4 (4.74)	5.0 (2.66)	7.2 (4.62)		WL<ICBT[i], WL<CBT[k], CBT<ICBT[i]	0.20 (0.18 to 0.22)	0.71 (0.69 to 0.73)	0.17 (0.14 to 0.19)	0.15 (0.13 to 0.17)	0.12 (0.10 to 0.15)	0.30 (0.28 to 0.33)
6-month follow-up	6.2 (4.46)	4.7 (2.53)	—		CBT<ICBT[i]	0.19 (0.17 to 0.21)	0.77 (0.75 to 0.80)	—	0.18 (0.16 to 0.21)	—	—
12-month follow-up	6.5 (4.20)	5.7 (2.32)	—		CBT<ICBT[i]	0.13 (0.10 to 0.15)	0.51 (0.49 to 0.54)	—	0.15 (0.11 to 0.19)	—	—
GDS[f]				13.80 (2, 199)[i]							
Baseline	16.3 (4.85)	17.2 (5.15)	16.8 (5.06)		—	—	—	—	—	—	—
Postintervention	12.8 (4.50)	12.2 (4.94)	15.4 (4.36)		WL<ICBT[i], WL<CBT[i]	0.06 (0.04 to 0.08)	0.18 (0.16 to 0.20)	0.04 (0.01 to 0.06)	0.05 (0.02 to 0.07)	0.06 (0.04 to 0.09)	0.10 (0.08 to 0.13)
6-month follow-up	11.8 (4.22)	11.5 (4.70)	—		—	0.06 (0.04 to 0.08)	0.23 (0.21 to 0.25)	—	0.04 (0.02 to 0.07)	—	—
12-month follow-up	11.5 (3.95)	12.6 (4.30)	—		—	0.06 (0.03 to 0.08)	0.17 (0.14 to 0.19)	—	0.05 (0.02 to 0.09)	—	—
GAD-7[g]				25.86 (2, 219)[i]							
Baseline	7.7 (2.47)	7.9 (2.67)	7.5 (2.59)		—	—	—	—	—	—	—
Postintervention	6.4 (2.30)	4.7 (2.56)	7.0 (2.23)		CBT<ICBT[i], WL<ICBT[i]	0.19 (0.17 to 0.21)	0.47 (0.45 to 0.49)	0.09 (0.06 to 0.11)	0.21 (0.19 to 0.23)	0.01 (0.02 to 0.03)	0.19 (0.17 to 0.22)
6-month follow-up	6.4 (2.16)	4.8 (2.43)	—		CBT<ICBT[i]	0.21 (0.19 to 0.23)	0.50 (0.48 to 0.52)	—	0.18 (0.15 to 0.21)	—	—

XSL•FO
RenderX

Outcome and time point	CBT[a], mean (SD)	ICBT[b], mean (SD)	WL[c], mean (SD)	Time × group, *F* test (*df*)	Significant contrasts between 2 groups	Within-group effect size (95% CI)			Between-group effect size (95% CI)		
						CBT	ICBT	WL	ICBT vs CBT	CBT vs WL	ICBT vs WL
12-month follow-up	6.7 (2.03)	4.8 (2.23)	—		CBT<ICBT[i]	0.15 (0.12 to 0.17)	0.51 (0.48 to 0.53)	—	0.18 (0.14 to 0.21)	—	—

[a]CBT: cognitive behavioral therapy.
[b]ICBT: internet-based cognitive behavioral therapy.
[c]WL: waiting list.
[d]CES-D: Center for Epidemiologic Studies Depression Scale.
[e]PHQ-9: Patient Health Questionnaire-9 Item.
[f]GDS: Geriatric Depression Scale.
[g]GAD-7: Generalized Anxiety Disorder 7-Item.
[h]*P*<.05.
[i]*P*<.001.
[j]Not available.
[k]*P*<.005.

The intervention groups showed significant improvements in almost all outcomes (*P*<.001) at postintervention and follow-up, with no significant change in the GAD-7 for the CBT group at 12-month follow-up (*P*=.07). Furthermore, the WL group showed significant improvements in the CES-D (*P*<.001) and PHQ-9 (*P*=.01). At postintervention, there were statistically significant between-group differences in the GDS and PHQ-9 scores in favor of the intervention group (all *P*<.001), and the ICBT participants had a larger reduction of the CES-D (*P*=.01) and GAD-7 scores (*P*<.001) compared with the WL participants. The ICBT participants exhibited significantly larger improvements in the PHQ-9 and GAD-7 scores than the CBT participants did over the study period (all *P*<.001). No other significant differences in the effects of ICBT and CBT were observed.

Effect Size

Primary Outcome

Table 2 shows effect sizes (Cohen *d*). We observed small within-group effect sizes on the CES-D among the 3 groups at postintervention (ICBT *d*=0.42, 95% CI 0.40-0.44; CBT *d*=0.16, 95% CI 0.14-0.19; WL *d*=0.15, 95% CI 0.13-0.18), 6-month follow-up (ICBT *d*=0.43, 95% CI 0.41-0.45; CBT *d*=0.17, 95% CI 0.14-0.19), and 12-month follow-up (ICBT *d*=0.30, 95% CI 0.28-0.33; CBT *d*=0.13, 95% CI 0.11-0.16).

The between-group effect size for CES-D among the 3 groups was trivial during the study period (postintervention: ICBT vs CBT *d*=0.00, 95% CI 0.02-0.02; CBT vs WL *d*=0.05, 95% CI 0.02-0.07; ICBT vs WL *d*=0.07, 95% CI 0.04-0.09; 6-month follow-up: ICBT vs CBT *d*=0.03, 95% CI 0.00-0.05; 12-month follow-up: ICBT vs CBT *d*=0.05, 95% CI 0.01-0.08).

Secondary Outcome

The within-group effect sizes for secondary outcomes ranged from trivial to medium (postintervention: ICBT *d* range=0.18-0.71, CBT *d* range=0.06-0.20, WL *d* range=0.04-0.17; 6-month follow-up: ICBT *d* range=0.23-0.77, CBT *d* range=0.06-0.21; 12-month follow-up: ICBT *d* range=0.17-0.51, CBT *d* range=0.06-0.15). The between-group effect sizes for secondary outcomes were trivial (postintervention: ICBT vs CBT *d* range=0.05-0.21, ICBT vs WL *d* range=0.10-0.30, CBT vs WL *d* range=0.01-0.12; 6-month follow-up: ICBT vs CBT *d* range=0.04-0.18; 12-month follow-up: ICBT vs CBT *d* range=0.05-0.18).

Reliable Change and Clinically Significant Improvement

The proportions of participants reporting reliable change and clinically significant improvement are shown in Table 3. We observed significantly larger rates of reliable change and clinically significant improvement in the ICBT group compared with the CBT group at postintervention (all *P*<.001) and 6-month follow-up (reliable change: *P*<.001; clinically significant improvement: *P*=.007), but these significant differences were not maintained at 12-month follow-up. The numbers needed to treat for ICBT to achieve 1 additional reliable change and clinically significant improvement were 2.89 (95% CI 2.12-4.56) and 3.16 (95% CI 2.28-5.15), respectively, and those for CBT were 6.26 (95% CI 3.45-33.84) and 5.98 (95% CI 3.48-21.01), respectively.

XSL•FO
RenderX

Table 3. Proportion of participants showing clinically significant improvement or reliable change based on the Center for Epidemiologic Studies Depression Scale (N=315).

Outcome and time point	CBT[a] (n=104), n (%)	ICBT[b] (n=105), n (%)	WL[c] (n=106), n (%)	Chi-square (*df*)	*P* value	NNT[d] (95% CI)	
						CBT vs WL	ICBT vs WL
Clinically significant improvement							
Postintervention	38 (36.5)	54 (51.4)	21 (19.8)	22.952 (2)	<.001	5.98 (3.48-21.01)	3.16 (2.28-5.15)
6-month follow-up	28 (26.9)	47 (44.8)	—[e]	7.227 (1)	.007	—	—
12-month follow-up	15 (14.4)	26 (24.8)	—	3.542 (1)	.06	—	—
Reliable change							
Postintervention	49 (47.1)	69 (65.7)	33 (31.1)	25.319 (2)	<.001	6.26 (3.45-33.84)	2.89 (2.12-4.56)
6-month follow-up	41 (39.4)	66 (62.9)	—	11.484 (1)	<.001	—	—
12-month follow-up	33 (31.7)	39 (37.1)	—	0.678 (1)	.41	—	—

[a]CBT: cognitive behavioral therapy.

[b]ICBT: internet-based cognitive behavioral therapy.

[c]WL: waiting list.

[d]NNT: number needed to treat.

[e]Not available.

Adherence, Attrition, and Response Rates and Treatment Satisfaction

Of the ICBT and CBT participants, 79.1% (83/105) and 62.5% (65/104) completed all the lessons, respectively. On average, a mean (SD) of 4.09 (10.41) and 3.80 (7.85) lessons were completed for the ICBT and CBT groups, respectively. The mean (SD) total clinical psychologist time per participant in the ICBT group was 35.26 (29.07) minutes. In total, 88.6% (93/105), 77.1% (81/105), and 56.2% (59/105) of ICBT participants completed the postintervention, 6-month follow-up, and 12-month follow-up assessments, respectively; 83.6% (87/104), 66.4% (69/104), and 48.1% (50/104) of CBT participants completed these respective assessments. Among the participants who completed the postintervention satisfaction questionnaires, 72% of ICBT participants (67/93) and 68% of CBT participants (60/87) indicated that they would confidently recommend this program to others. Furthermore, 68% of ICBT participants (64/93) and 62% of CBT participants (54/87) found it worthwhile to participate in this program.

Discussion

Principal Findings

The main finding of this RCT is that clinician-guided ICBT was relatively effective in reducing depressive symptoms for older adults with sD recruited from institutional LTC settings. ICBT participants showed greater improvement in the CES-D, PHQ-9, GDS, and GAD-7 scores compared with the WL participants at postintervention. In addition, 37.10% and 24.80% of ICBT participants reported reliable change and clinically significant improvement in CES-D from pretreatment to 12-month follow-up, respectively. Furthermore, 79.05% of ICBT participants completed 5 lessons, and more than 60% reported a high degree of satisfaction.

Comparison With Previous Work

The findings of this study are encouraging and are consistent with those of similar studies concerning ICBT for adults with sD [43]. However, both the within-group and between-group effect sizes for ICBT versus WL regarding reduction in depressive symptom severity at postintervention were lower than those in a previous trial on older adults with sD [17]. We surmise that the inferior effect might be explained by the relatively high mean age of participants in this study (71.60 years), which was substantially higher than that in the previous study (55.00 years). A recent meta-analysis showed that the therapeutic effects of ICBT for depression might diminish slightly with age due to technological challenges and limited age-appropriate components [15]. In addition, participants in this study were less educated than those in the prior study, which might have led to limited technology competency and therefore fewer ICBT benefits [44,45].

Compared with CBT participants, ICBT participants showed a significantly larger improvement in the PHQ-9 and GAD-7 at postintervention and follow-up. In addition, a significantly higher proportion of ICBT participants reported clinically significant improvement and reliable change for CES-D at postintervention and 6-month follow-up, which are consistent with the findings by Spek et al [17]. Although compelling meta-analysis evidence shows that ICBT and CBT may have equal overall benefits for depressive and anxiety symptoms [16,46], this study findings may be explained by the higher adherence among ICBT participants than among CBT

XSL•FO
RenderX

participants. Hilvert-Bruce et al [47] revealed that improved adherence is an important determinant of effectiveness of both ICBT and face-to-face CBT for depression and anxiety. For other outcomes, we observed a nonsignificant trend toward improvement for ICBT compared with CBT, possibly because our program was developed based on conventional CBT. Further, ICBT participants received relatively intensive guidance, which mirrors the therapeutic alliance in face-to-face CBT, and as a result, the treatment effect of these 2 interventions might be similar. Nevertheless, most of the prior studies contrasting ICBT and CBT were based on the general population instead of older adults, and no study has reported between-group effect sizes on the outcomes of ICBT versus CBT among older adults with sD, not to mention LTC residents with sD. Thus, more research with a larger sample should be conducted to further compare ICBT with CBT for sD among LTC residents and explore adherence as a moderator of treatment effects.

To the best of our knowledge, this study is the first to demonstrate the feasibility and acceptability of ICBT for Chinese older adults with sD. Our findings are comparable to those of other trials that investigated the use of ICBT for older adults with sD or MDD in Western countries [15,17]. Chinese older adults tend to have limited technology competency and thus had difficulties in using ICBT compared with older adults from developed countries with prior experience in technology use [48]. Therefore, it is encouraging to find that the adherence, response rates, and treatment satisfaction in this study were as high as those observed in the aforementioned studies. This finding might be explained by the relatively intensive therapist-guided ICBT in this study, and given the culture of obedience to the authority of doctors among older Chinese individuals, clinician-guided ICBT may be more acceptable for elderly Chinese compared with older adults from other racial and ethnic backgrounds. In addition, ICBT participants who did not attend scheduled lessons were traced through active outreach via email or telephone. These human support strategies can decrease participant attrition rate [49] and are particularly important for older adults with limited technology competency to improve their adherence to ICBT [15].

Nevertheless, concerns remain as to whether ICBT for LTC residents with sD is as effective as that for the general older population because the effect sizes observed in this study were substantially smaller than those in previous studies on ICBT for depression among the general older population [15]. There are 4 possible reasons for these findings. First, our ICBT program was not tailored for older adults, especially those living in LTC settings who tend to be older, have lower education, and have lower levels of digital competency compared with those living at home [50]. In addition, LTC residents were more likely to have a history of MDD or a history of psychotropic medication [51], which are possible predictive factors for the lack of treatment response and slower improvement [52]. Second, our ICBT program was based on conventional CBT. However, 2 prior systematic reviews have demonstrated that although CBT for LTC residents may alleviate depression, the evidence for effectiveness was mixed [53,54]. The possible reason is that the goal of CBT is to distinguish unrealistic or exaggerated thoughts from more realistic ones [55], but LTC residents are less likely to benefit from it due to institutional policies and practices. For example, this study was conducted during the COVID-19 pandemic, a time when closed management and suspension of family visits were irregularly implemented in all LTC settings in China. This lack of physical interaction might have generated more unrealistic or exaggerated thoughts and therefore reduced the effectiveness of ICBT. Third, compared with older adults living in the community, LTC residents may feel deprived of autonomy and lack of privacy by the institutional environment [55,56], which might be detrimental to their mental health [57] and consequently reduce the effectiveness of ICBT. Fourth, our ICBT program has only 5 modules, which seem to be fewer than the usual ICBT sessions for general older populations [15]. Thus, participants in this study may have limited access to more CBT skills to alleviate depressive symptoms. In this regard, researchers should develop a tailored ICBT with additional lesson content for LTC residents, considering the unique characteristics of LTC settings, and explore its effectiveness in the future.

Notably, more than 60% of the ICBT participants failed to achieve a reliable change at 12-month follow-up; this may lower the motivation of older adults with sD for more intensive treatment, such as face-to-face CBT or pharmacotherapy [58]. Consequently, untreated LTC residents with sD may have higher risk of MDD and impaired physical and mental health [10] and may therefore become less likely to discharge to the community. Thus, future studies are needed to explore the predictive factors for nonresponders in LTC settings.

Implications

First, our findings highlight the feasibility and potential effectiveness of ICBT for improving the mental health of older adults with sD, especially those living in LTC settings. In effect, they are likely to be discharged to the community, which may alleviate the burden on the LTC system. Considering the relatively undeveloped LTC system [22], limited resources for mental health services, and many undetected and untreated cases of sD in LTC settings in China, this study underscores the importance of developing a tailored ICBT for LTC residents with sD and promoting its use in China. Second, guided ICBT can produce equivalent overall effects as face-to-face CBT for LTC residents with sD. Nevertheless, there is a shortage of psychotherapists in Chinese LTC settings, and most of them have low professionalism [20]. In addition, because of the advantages of ICBT, including flexibility, scalability, privacy, and cost-effectiveness, ICBT can be a realistic alternative to CBT in LTC settings in China. Third, our findings provide preliminary evidence for the potential usefulness of ICBT to reduce depressive symptoms among LTC residents in China and other countries during the COVID-19 pandemic, given the closed management and visitor restrictions.

Strengths and Limitations

The core strengths of this trial were its RCT design with direct comparison of 2 interventions, validated measures with good reliability and validity, large sample size, and a high level of adherence. The inclusion of participants with a history of MDD and on antidepressants is both a strength and a weakness; it increases the external validity of the findings but may obfuscate

XSL•FO
RenderX

the effects of the intervention [52,59]. Nevertheless, several limitations should be considered. First, the choice of group CBT intervention, which has a lower degree of privacy than individual CBT, as a comparator group, may adversely affect participants' engagement. In future studies, individual CBT may be preferable as a comparator. Second, this study included participants over 60 years old, which is a group of people in whom depression and other age-related disorders are perceived as normal in most cases. Furthermore, this study relied on self-reported measures rather than objective depression diagnoses at postintervention and follow-up. This may cause a high risk of selection bias, and results should therefore be interpreted carefully. Third, these findings may have lower generalizability because all the LTC facilities in this study were located in Zhejiang Province, one of the most economically vibrant and developed provinces in China. Thus, a further RCT should be conducted using a more diverse population from different Chinese provinces.

Conclusions

Guided ICBT has been shown to be equally effective and acceptable as CBT for improving depressive symptoms among Chinese LTC residents with sD, and the improvements are maintained over 12 months after the intervention. Given the insufficient mental health services in LTC settings in China, ICBT has the potential to reach millions of Chinese LTC residents exposed to sD, who are at risk of being undetected and untreated.

Acknowledgments

We thank all the clinicians for providing clinician assistance, programmers for constructing the internet-based cognitive behavioral therapy platform, and Editage for English language editing. The study was approved the Medical and Health Science and Technology Plan Project of Zhejiang Province (grant 2021KY335), Ningbo Philosophy and Social Planning Project (grant G21-3-ZX94), Zhejiang Province Public Welfare Technology Application Research Project (grant LGF21H090004), General Scientific Research Project of Department of Education of Zhejiang Province (grant Y201942547), and Science and Technology Innovation Activity Plan of Zhejiang University Student & XinMiao Talents Program (grants 2021R464003 and 2020R479003). None of the funding organizations had any role in the conception of the study design or in the collection, analysis, or interpretation of the data; in the writing of the report; or in the decision to submit the paper.

Authors' Contributions

FK was involved in study design, participant management, data analysis, data interpretation, and manuscript preparation. LY was involved in study design. YH was involved in study design, participant management, and manuscript preparation. L Zhu was involved in data analysis and data interpretation. JZ was involved in project management. LH was involved in participant management. YL was involved in participant management. LW was involved in participant management. L Zhang was involved in participant management. Y Yang was involved in participant management. Y Ying was involved in study design, data interpretation, and manuscript preparation. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V.1.6.1)
[PDF File (Adobe PDF File), 19181 KB - jmir_v26i1e40187_app1.pdf]

References

1. Newmann JP. Aging and depression. Psychol Aging 1989;4(2):150-165. [doi: 10.1037//0882-7974.4.2.150] [Medline: 2675927]
2. Oh DJ, Han JW, Kim TH, Kwak KP, Kim BJ, Kim SG, et al. Epidemiological characteristics of subsyndromal depression in late life. Aust N Z J Psychiatry 2020;54(2):150-158. [doi: 10.1177/0004867419879242] [Medline: 31595770]
3. Bosman RC, Have MT, de Graaf R, Muntingh AD, van Balkom AJ, Batelaan NM. Prevalence and course of subthreshold anxiety disorder in the general population: a three-year follow-up study. J Affect Disord 2019;247:105-113. [doi: 10.1016/j.jad.2019.01.018] [Medline: 30660019]
4. Hybels CF, Pieper CF, Blazer DG. The complex relationship between depressive symptoms and functional limitations in community-dwelling older adults: the impact of subthreshold depression. Psychol Med 2009;39(10):1677-1688 [FREE Full text] [doi: 10.1017/S0033291709005650] [Medline: 19356260]
5. Nierenberg AA, Rapaport MH, Schettler PJ, Howland RH, Smith JA, Edwards D, et al. Deficits in psychological well-being and quality-of-life in minor depression: implications for DSM-V. CNS Neurosci Ther 2010;16(4):208-216 [FREE Full text] [doi: 10.1111/j.1755-5949.2009.00108.x] [Medline: 20236139]

XSL•FO
RenderX

6. Lee YY, Stockings EA, Harris MG, Doi SAR, Page IS, Davidson SK, et al. The risk of developing major depression among individuals with subthreshold depression: a systematic review and meta-analysis of longitudinal cohort studies. Psychol Med 2019;49(1):92-102. [doi: 10.1017/S0033291718000557] [Medline: 29530112]
7. Polyakova M, Sonnabend N, Sander C, Mergl R, Schroeter ML, Schroeder J, et al. Prevalence of minor depression in elderly persons with and without mild cognitive impairment: a systematic review. J Affect Disord 2014;152-154:28-38. [doi: 10.1016/j.jad.2013.09.016] [Medline: 24103852]
8. Bremmer MA, Hoogendijk WJG, Deeg DJH, Schoevers RA, Schalk BWM, Beekman ATF. Depression in older age is a risk factor for first ischemic cardiac events. Am J Geriatr Psychiatry 2006;14(6):523-530. [doi: 10.1097/01.JGP.0000216172.31735.d5] [Medline: 16731721]
9. Ludvigsson M, Bernfort L, Marcusson J, Wressle E, Milberg A. Direct costs of very old persons with subsyndromal depression: a 5-year prospective study. Am J Geriatr Psychiatry 2018;26(7):741-751 [FREE Full text] [doi: 10.1016/j.jagp.2018.03.007] [Medline: 29673895]
10. Meeks TW, Vahia IV, Lavretsky H, Kulkarni G, Jeste DV. A tune in "a minor" can "b major": a review of epidemiology, illness course, and public health implications of subthreshold depression in older adults. J Affect Disord 2011;129(1-3):126-142 [FREE Full text] [doi: 10.1016/j.jad.2010.09.015] [Medline: 20926139]
11. Hindi F, Dew MA, Albert SM, Lotrich FE, Reynolds CF. Preventing depression in later life: state of the art and science circa 2011. Psychiatr Clin North Am 2011;34(1):67-78 [FREE Full text] [doi: 10.1016/j.psc.2010.11.008] [Medline: 21333840]
12. Cuijpers P, Karyotaki E, Pot AM, Park M, Reynolds CF. Managing depression in older age: psychological interventions. Maturitas 2014;79(2):160-169 [FREE Full text] [doi: 10.1016/j.maturitas.2014.05.027] [Medline: 24973043]
13. Unützer J. Diagnosis and treatment of older adults with depression in primary care. Biol Psychiatry 2002;52(3):285-292. [doi: 10.1016/s0006-3223(02)01338-0] [Medline: 12182933]
14. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019;18(1):20-28 [FREE Full text] [doi: 10.1002/wps.20610] [Medline: 30600624]
15. Xiang X, Wu S, Zuverink A, Tomasino KN, An R, Himle JA. Internet-delivered cognitive behavioral therapies for late-life depressive symptoms: a systematic review and meta-analysis. Aging Ment Health 2020;24(8):1196-1206 [FREE Full text] [doi: 10.1080/13607863.2019.1590309] [Medline: 30913898]
16. Andersson G, Topooco N, Havik O, Nordgreen T. Internet-supported versus face-to-face cognitive behavior therapy for depression. Expert Rev Neurother 2016;16(1):55-60. [doi: 10.1586/14737175.2015.1125783] [Medline: 26610160]
17. Spek V, Cuijpers P, Nyklícek I, Smits N, Riper H, Keyzer J, et al. One-year follow-up results of a randomized controlled clinical trial on internet-based cognitive behavioural therapy for subthreshold depression in people over 50 years. Psychol Med 2008;38(5):635-639. [doi: 10.1017/S0033291707002590] [Medline: 18205965]
18. Chen S, Conwell Y, He J, Lu N, Wu J. Depression care management for adults older than 60 years in primary care clinics in urban China: a cluster-randomised trial. Lancet Psychiatry 2015;2(4):332-339. [doi: 10.1016/S2215-0366(15)00002-4] [Medline: 26360086]
19. Zhong Y, Gao J, Bai D, Liu R, Chen S, Liao Q, et al. A meta-analysis of detection rate of depressive symptoms among the elderly in pension institutions in China. Chin Ment Health J 2020;34(12):1006-1051. [doi: 10.3969/j.issn.1000-6729.2020.12.007]
20. Wang T. Research on the psychological and mental service system of pension institutions under the background of combination of medical care and nursing care. China Dev 2020;20(5):26-30. [doi: 10.15885/j.cnki.cn11-4683/z.2020.05.005]
21. China pledges to scale up efforts in depression prevention. Xinhua News Agency. 2020. URL: http://english.www.gov.cn/statecouncil/ministries/202009/11/content_WS5f5b1d76c6d0f7257693bd5b.html [accessed 2023-12-22]
22. Feng Z, Glinskaya E, Chen H, Gong S, Qiu Y, Xu J, et al. Long-term care system for older adults in China: policy landscape, challenges, and future prospects. Lancet 2020;396(10259):1362-1372. [doi: 10.1016/S0140-6736(20)32136-X] [Medline: 34338215]
23. Buntrock C, Ebert D, Lehr D, Riper H, Smit F, Cuijpers P, et al. Effectiveness of a web-based cognitive behavioural intervention for subthreshold depression: pragmatic randomised controlled trial. Psychother Psychosom 2015;84(6):348-358. [doi: 10.1159/000438673] [Medline: 26398885]
24. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
25. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
26. Dobson KS, Li Z. Chinese Version of Handbook of Cognitive-Behavioral Therapies, 3rd Edition. Beijing: People's Medical Publishing House; 2015.
27. Ying Y, Ji Y, Kong F, Chen Q, Lv Y, Hou Y, et al. Internet-based cognitive behavioral therapy for psychological distress in older adults without cognitive impairment living in nursing homes during the COVID-19 pandemic: a feasibility study. Internet Interv 2021;26:100461 [FREE Full text] [doi: 10.1016/j.invent.2021.100461] [Medline: 34631432]
28. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 2016;1(3):385-401. [doi: 10.1177/014662167700100306]

XSL•FO
RenderX

29. Yesavage JA, Sheikh JI. 9/Geriatric Depression Scale (GDS). Clin Gerontol 2008;5(1-2):165-173. [doi: 10.1300/J018v05n01_09]
30. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
31. Lin N. Measuring depressive symptomatology in China. J Nerv Ment Dis 1989;177(3):121-131. [doi: 10.1097/00005053-198903000-00001] [Medline: 2918295]
32. Chan AC. Clinical validation of the Geriatric Depression Scale (GDS): Chinese version. J Aging Health 1996;8(2):238-253. [doi: 10.1177/089826439600800205] [Medline: 10160560]
33. Wang W, Bian Q, Zhao Y, Li X, Wang W, Du J, et al. Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry 2014;36(5):539-544. [doi: 10.1016/j.genhosppsych.2014.05.021] [Medline: 25023953]
34. Yu W, Singh SS, Calhoun S, Zhang H, Zhao X, Yang F. Generalized anxiety disorder in urban China: prevalence, awareness, and disease burden. J Affect Disord 2018;234:89-96. [doi: 10.1016/j.jad.2018.02.012] [Medline: 29524751]
35. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. Arch Gen Psychiatry 2004;61(3):310-317. [doi: 10.1001/archpsyc.61.3.310] [Medline: 14993119]
36. Feingold A. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. Psychol Methods 2009;14(1):43-53 [FREE Full text] [doi: 10.1037/a0014699] [Medline: 19271847]
37. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods 2002;7(1):105-125. [doi: 10.1037/1082-989x.7.1.105] [Medline: 11928886]
38. Thorsell J, Finnes A, Dahl J, Lundgren T, Gybrant M, Gordh T, et al. A comparative study of 2 manual-based self-help interventions, acceptance and commitment therapy and applied relaxation, for persons with chronic pain. Clin J Pain 2011;27(8):716-723. [doi: 10.1097/AJP.0b013e318219a933] [Medline: 21540740]
39. Kelley K. Confidence intervals for standardized effect sizes: theory, application, and implementation. J Stat Soft 2007;20(8):1-24 [FREE Full text] [doi: 10.18637/jss.v020.i08]
40. Cohen J. Statistical Power Analysis for the Behavioural Sciences 2nd Edition. Mahwah, NJ: Lawrence Erlbaum Associates; 1988.
41. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037//0022-006x.59.1.12] [Medline: 2002127]
42. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310(6977):452-454 [FREE Full text] [doi: 10.1136/bmj.310.6977.452] [Medline: 7873954]
43. Reins JA, Buntrock C, Zimmermann J, Grund S, Harrer M, Lehr D, et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 2021;90(2):94-106 [FREE Full text] [doi: 10.1159/000507819] [Medline: 32544912]
44. Imamura K, Kawakami N, Furukawa TA, Matsuyama Y, Shimazu A, Umanodan R, et al. Effects of an internet-based cognitive behavioral therapy (iCBT) program in manga format on improving subthreshold depressive symptoms among healthy workers: a randomized controlled trial. PLoS One 2014;9(5):e97167 [FREE Full text] [doi: 10.1371/journal.pone.0097167] [Medline: 24844530]
45. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
46. Esfandiari N, Mazaheri MA, Akbari-Zardkhaneh S, Sadeghi-Firoozabadi V, Cheraghi M. Internet-delivered versus face-to-face cognitive behavior therapy for anxiety disorders: systematic review and meta-analysis. Int J Prev Med 2021;12:153 [FREE Full text] [doi: 10.4103/ijpvm.ijpvm_208_21] [Medline: 35070186]
47. Hilvert-Bruce Z, Rossouw PJ, Wong N, Sunderland M, Andrews G. Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders. Behav Res Ther 2012;50(7-8):463-468. [doi: 10.1016/j.brat.2012.04.001] [Medline: 22659155]
48. Hong YA, Zhou Z, Fang Y, Shi L. The digital divide and health disparities in China: evidence from a national survey and policy implications. J Med Internet Res 2017;19(9):e317 [FREE Full text] [doi: 10.2196/jmir.7786] [Medline: 28893724]
49. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
50. Abrahamson K, Hass Z, Arling G. Shall I stay or shall I go? The choice to remain in the nursing home among residents with high potential for discharge. J Appl Gerontol 2020;39(8):863-870. [doi: 10.1177/0733464818807818] [Medline: 30366510]
51. Levin CA, Wei W, Akincigil A, Lucas JA, Bilder S, Crystal S. Prevalence and treatment of diagnosed depression among elderly nursing home residents in Ohio. J Am Med Dir Assoc 2007;8(9):585-594 [FREE Full text] [doi: 10.1016/j.jamda.2007.07.010] [Medline: 17998115]

XSL•FO
RenderX

52. Andersson G, Bergström J, Holländare F, Ekselius L, Carlbring P. Delivering cognitive behavioural therapy for mild to moderate depression via the internet: predicting outcome at 6-month follow-up. Verhaltenstherapie 2004;14(3):185-189. [doi: 10.1159/000080914]
53. Simning A, Simons KV. Treatment of depression in nursing home residents without significant cognitive impairment: a systematic review. Int Psychogeriatr 2017;29(2):209-226 [FREE Full text] [doi: 10.1017/S1041610216001733] [Medline: 27758728]
54. Bharucha AJ, Dew MA, Miller MD, Borson S, Reynolds C. Psychotherapy in long-term care: a review. J Am Med Dir Assoc 2006;7(9):568-580. [doi: 10.1016/j.jamda.2006.08.003] [Medline: 17095422]
55. Konnert C, Dobson K, Stelmach L. The prevention of depression in nursing home residents: a randomized clinical trial of cognitive-behavioral therapy. Aging Ment Health 2009;13(2):288-299. [doi: 10.1080/13607860802380672] [Medline: 19347696]
56. Choi NG, Ransom S, Wyllie RJ. Depression in older nursing home residents: the influence of nursing home environmental stressors, coping, and acceptance of group and individual therapy. Aging Ment Health 2008;12(5):536-547. [doi: 10.1080/13607860802343001] [Medline: 18855169]
57. Queirós AM, von Gunten A, Martins M, Wellens NIH, Verloo H. The forgotten psychopathology of depressed long-term care facility residents: a call for evidence-based practice. Dement Geriatr Cogn Dis Extra 2021;11(1):38-44 [FREE Full text] [doi: 10.1159/000514118] [Medline: 33790939]
58. Zhou T, Li X, Pei Y, Gao J, Kong J. Internet-based cognitive behavioural therapy for subthreshold depression: a systematic review and meta-analysis. BMC Psychiatry 2016;16(1):356 [FREE Full text] [doi: 10.1186/s12888-016-1061-9] [Medline: 27769266]
59. Hadjistavropoulos HD, Pugh NE, Hesser H, Andersson G. Predicting response to therapist-assisted internet-delivered cognitive behavior therapy for depression or anxiety within an open dissemination trial. Behav Ther 2016;47(2):155-165. [doi: 10.1016/j.beth.2015.10.006] [Medline: 26956649]

Abbreviations

CBT: cognitive behavioral therapy
CES-D: Epidemiological Studies Depression Scale
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
GAD-7: Generalized Anxiety Disorder 7-Item
GDS: Geriatric Depression Scale
ICBT: internet-based cognitive behavioral therapy
LTC: long-term care
MDD: major depressive disorder
PHQ-9: Patient Health Questionnaire–9
RCT: randomized controlled trial
SCID: Semistructured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Axis Disorders
sD: subthreshold depression
WL: waiting list

Edited by A Mavragani; submitted 09.06.22; peer-reviewed by AA Seid, N Esfandiari, Z Lei; comments to author 12.08.22; revised version received 03.01.23; accepted 31.01.23; published 01.03.24.

Please cite as:
Kong F, Yu L, Hou Y, Zhu L, Zhou J, Huang L, Lv Y, Wang L, Zhang L, Yang Y, Ying Y
Efficacy of Internet-Based Cognitive Behavioral Therapy for Subthreshold Depression Among Older Adults in Institutional Long-Term Care Settings: Pragmatic Randomized Controlled Trial
J Med Internet Res 2024;26:e40187
URL: https://www.jmir.org/2024/1/e40187
doi:10.2196/40187
PMID:38427424

©Fanqian Kong, Libo Yu, Yanbin Hou, Lijie Zhu, Jing Zhou, Lingjie Huang, Yueer Lv, Li Wang, Li Zhang, Yiling Yang, Yuchen Ying. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Layperson-Supported, Web-Delivered Cognitive Behavioral Therapy for Depression in Older Adults: Randomized Controlled Trial

Xiaoling Xiang[1], MSW, PhD; Jay Kayser[1], MSW; Skyla Turner[1], MSW; Samson Ash[1], MSW; Joseph A Himle[1], MSW, PhD

School of Social Work, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Xiaoling Xiang, MSW, PhD
School of Social Work
University of Michigan
1080 S University Ave
Ann Arbor, MI, 48109
United States
Phone: 1 (734) 763 6581
Email: xiangxi@umich.edu

Abstract

Background: Depression is the most prevalent mental health condition in older adults. However, not all evidence-based treatments are easily accessible. Web-delivered cognitive behavioral therapy (wCBT) facilitated by laypersons is a viable treatment alternative.

Objective: This randomized controlled trial aims to evaluate the efficacy of a novel wCBT program, Empower@Home, supported by trained lay coaches, against a waitlist attention control. Empower@Home is among the very few existing wCBT programs specifically designed for older adults. The primary objective was to assess the efficacy of the intervention compared with attention control. The secondary objective was to evaluate the program's impact on secondary psychosocial outcomes and explore potential change mechanisms.

Methods: Older adults (N=70) were recruited via web-based research registries, social media advertisements, and community agency referrals and randomly assigned to either the intervention or control group in a 1:1 allocation ratio. The intervention group received access to Empower@Home, which included 9 web-delivered self-help lessons and weekly telephone coaching sessions by a trained layperson over 10 weeks. The control group received weekly friendly phone calls and depressive symptom monitoring. The primary clinical outcome was the severity of depressive symptoms assessed using the Patient Health Questionnaire-9. The secondary clinical outcomes included anxiety, anger, social isolation, insomnia, pain intensity, and quality of life. Linear mixed modeling was used to determine the treatment effects on depression, and 2-tailed *t* tests were used to assess within-group changes and between-group differences.

Results: Most participants in the intervention group completed all 9 sessions (31/35, 89%). The usability and acceptability ratings were excellent. The intervention group had a large within-group change in depressive symptoms (Cohen d=1.22; P<.001), whereas the attention control group experienced a medium change (Cohen d=0.57; P<.001). The between-group effect size was significant, favoring the intervention group over the control group (Cohen d=0.72; P<.001). In the linear mixed model, the group-by-time interaction was statistically significant (b=–0.68, 95% CI –1.00 to –0.35; P<.001). The treatment effects were mediated by improvements in cognitive behavioral therapy skills acquisition; behavioral activation; and satisfaction with the basic psychological needs of autonomy, competence, and relatedness. Furthermore, the intervention group showed significant within-group improvements in secondary psychosocial outcomes, including anxiety (P=.001), anger (P<.001), social isolation (P=.02), insomnia (P=.007), and pain (P=.03). By contrast, the control group did not experience significant changes in these outcome domains. However, the between-group differences in secondary outcomes were not statistically significant, owing to the small sample size.

Conclusions: Empower@Home, a wCBT program supported by lay coaches, was more efficacious in reducing depressive symptoms than friendly telephone calls and depression symptom monitoring. Future studies should examine the effectiveness of the intervention in community and practice settings using nonclinician staff already present in these real-world settings as coaches.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT05593276; https://clinicaltrials.gov/ct2/show/NCT05593276
International Registered Report Identifier (IRRID): RR2-10.2196/44210

(J Med Internet Res 2024;26:e53001) doi:10.2196/53001

KEYWORDS
internet-based cognitive behavioral therapy; iCBT; digital mental health interventions; older adults; depression

Introduction

Background

Depression is the most prevalent mental health condition among older adults and is particularly common among those with chronic physical health conditions and functional limitations [1]. Although evidence-based pharmacological and nonpharmacological treatments exist, not every option is readily accessible to all populations of older adults [2,3]. For example, older adults with mobility difficulties face significant logistical barriers to accessing office-based treatments in addition to common access barriers such as cost, provider shortages, and stigma [2].

Web-delivered cognitive behavioral therapy (wCBT), also known as internet-based cognitive behavioral therapy or computerized cognitive behavioral therapy, is a promising option for addressing the unmet mental health needs of older adults. wCBT, initially introduced in the 90s via CD-ROM manuals, now typically presents cognitive behavioral therapy (CBT) principles through a combination of audio, video, and text elements hosted on dedicated websites or apps. The advantages of wCBT include its low cost, efficiency, convenience, and accessibility [4]. Moreover, preprogrammed components in wCBT can minimize variability between trials and dissemination, thus maintaining greater fidelity to treatment protocols compared with the potential variability or *drift* seen in face-to-face CBT sessions [5].

Although wCBT is promising, its real-world impact can be hampered by low user engagement [6]. Common issues with wCBT programs include dense text, academic oriented content, and a tendency to adopt a *one size fits all* approach [7]. Furthermore, there is a scarcity of wCBT programs specifically designed for older adults, with only a handful having undergone development and testing [8-10]. Although some generic wCBT programs have demonstrated effective depression reduction among older adults [11-13], our assessments of underserved older adult populations, encompassing individuals with low income, limited literacy, technological challenges, and disabilities, have revealed many usability and engagement problems [14,15]. For wCBT to truly benefit older adults, there is a pressing need to address these issues with more inclusive, user-friendly, and engaging programs.

To our knowledge, wCBT programs tailored to the needs of older adults and readily available for US consumers are lacking. We addressed this shortage and developed Empower@Home in response to the scarcity of wCBT programs tailored to the needs of older adults in underserved communities. The design process of Empower@Home, detailed elsewhere [16], was grounded in a user-centered and community-engaged approach. The final program included 9 web-delivered sessions grounded in CBT principles, presented via a custom-made website with a user-friendly interface. Along with the web sessions, we provided a large print user workbook featuring session summaries, home practices, and wellness resources. A unique aspect of our approach is the infusion of entertainment through a character-driven storyline. This story centers on Jackie, a homebound older adult character presented as a 74-year-old woman who navigates health challenges typical of the target demographic. Her experiences are depicted in an animated series shaped by stakeholder input. The web program was enhanced with the inclusion of “Empower Coaches,” laypersons trained to support users. Research has shown that wCBT interventions with human support are more efficacious than those without human assistance [17-19]. Layperson coaches offer an alternative to relying on scarce gerontological mental health professionals.

During usability and field tests, Empower@Home showed superior usability compared with 2 evidence-supported and commercially available wCBT programs [16]. Findings from an uncontrolled trial demonstrated a medium effect size for depression reduction in older adults with mild depression and a large effect size in those with moderate depression [14]. However, the intervention is yet to be assessed against a control condition.

Objectives

This study aims to evaluate the efficacy of layperson-supported Empower@Home for depression in older adults through a randomized controlled trial (RCT). The primary hypothesis was that the treatment group would experience a significant reduction in depressive symptoms compared with the attention control group. In addition, this study aims to investigate the impact of the intervention on secondary clinical outcomes, including anxiety, loneliness, pain severity, and overall disability burden, with the hypothesis that the treatment group would show greater improvements in these areas than the control group. Furthermore, this study intends to explore potential mediators to gain a deeper understanding of the mechanisms responsible for the expected reductions in depressive symptoms.

Methods

Ethical Considerations

The study protocol has been published elsewhere [3]. An abbreviated description of the study methods has been provided in this study to inform the readers. The study was approved by the University of Michigan Institutional Review Board–Health Sciences and Behavioral Science (HUM00212950) and was registered on ClinicalTrials.gov (NCT05593276) on October

XSL•FO
RenderX

24, 2022. Participants were eligible to receive up to US $100 for participation in the study. Figure 1 illustrates the flow and allocation of the participants. After screening, eligible participants were invited to provide their informed consent. They completed the consent document by phone and were provided with a mailed copy upon providing verbal consent. Before beginning the first session, the participants provided their consent on the website hosting the program.

Figure 1. Participant flow through the study.

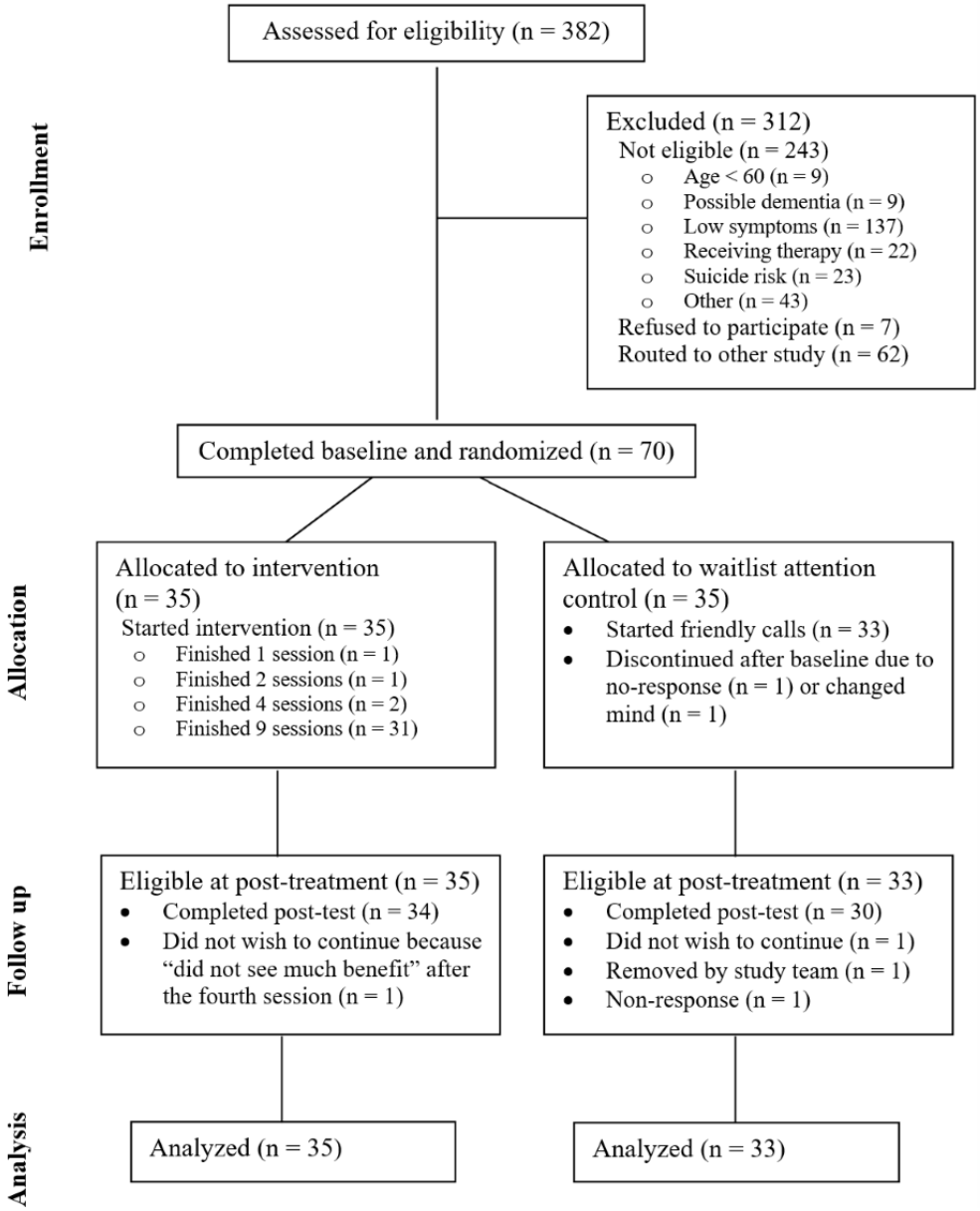

Study Design

The study adopted a parallel RCT design, with participants randomly assigned to either Empower@Home supported by trained lay coaches or a waitlist attention control group, maintaining a 1:1 allocation ratio. The allocation followed a computer-generated random sequence. The principal investigator (XX) generated the sequence, whereas the project coordinator (ST) managed the actual allocation. Other research staff, who were uninformed of the sequence and allocation decision, conducted screening assessments. Although it was not feasible to blind participants to their conditions owing to the intervention's content and study design, research staff conducting follow-up assessments were blinded.

Participant Recruitment

Participants were recruited from research volunteer registries, social media advertisements, and referrals from community agencies. To qualify, participants were required to be able to read and speak English, reside in Michigan, be aged at least 60 years, and have elevated depressive symptoms at screening (as indicated by a score of ≥8 on the Patient Health Questionnaire-9 [PHQ-9]). Participants were excluded if they had probable dementia, a psychotic disorder, moderate to high risk of suicide, a terminal illness, a current substance use disorder, or uncorrected severe vision impairment (eg, blindness) or if they were receiving or planning to receive psychotherapy during the trial. Lack of device ownership or internet access was not an exclusion criterion. Participants without a computer or internet

XSL•FO
RenderX

access were provided with a cellular tablet at no cost during the trial.

Procedures

Potential participants identified from volunteer registries and social media (Facebook) advertisements were directed to a web survey to complete a general screening survey used across multiple studies. Facebook advertisements were directed across all areas of Michigan to users aged ≥60 years. The research team contacted individuals who met the preliminary eligibility criteria up to 3 times for further phone screening. To accommodate potential internet access issues, referrals from community agencies skipped the web screener and were called directly by the research team. Recruitment was completed between January and June 2023.

Assessments were conducted over the phone, with a baseline assessment occurring within 2 weeks preceding the start of the 10-week intervention and a posttest assessment within 2 weeks following the intervention or waiting period. Participants were eligible to receive a total compensation of up to US $100 for completing research assessments, comprising US $30 for the baseline assessment, US $50 for the posttest assessment, and US $20 for a 10-week follow-up (not discussed in this paper). We emphasized that this compensation was solely contingent on completing the research assessments and had no connection to participation in the intervention or coaching. This distinction was repeated multiple times by our consent staff and coaches to ensure clarity regarding payment structure.

Participants in the waitlist control group were contacted by trained research staff with educational backgrounds similar to those of the Empower Coaches who supported the participants in the treatment group. These staff members (N=3) conducted weekly friendly phone calls, averaging 17 minutes each, and administered depression assessments using the PHQ-9 every 2 weeks. These “friendly callers” followed a structured call guide with a bank of questions to guide each call, such as “How has your week been?” or “What was the highlight of your week?” Callers were asked to prioritize empathetic listening during the calls. After their posttest assessment, the control participants were offered the Empower@Home program.

We used a waitlist group with attention control to differentiate between the effects of human interaction and the wCBT program itself. This approach was crucial for assessing whether improvements were because of the program or merely the result of coach-participant interactions. Ethically, we offered the intervention to the waitlist control group after the posttest assessment, as many participants needed treatment and were not service connected. A comparison with another wCBT program was not ideal, as our goal was to evaluate the efficacy of our novel program and not to prove its superiority. In addition, the efficacy of other wCBT programs in the target population was uncertain. In summary, this design enabled a clear assessment of our intervention’s efficacy in improving mental health outcomes in older adults.

All 35 participants in the treatment group started the wCBT program and were included in the analysis. In the control group, 2 participants withdrew from the study after the baseline assessment and before the initiation of friendly calls, and they were excluded from the analysis. This resulted in 33 participants in the control group included in the analysis. Notably, 1 participant from the control group was removed from the study because of a recent significant loss and pronounced grief concerns, as assessed by our clinical social worker. It was determined that this participant required more immediate assistance than the program was designed to provide, and a removal decision was made according to our protocol.

Intervention

Empower@Home contains 9 sequenced sessions, drawing from the CBT manual for working with older people from the study by Gallagher-Thompson and Thompson [20] and the Behavioral Activation Manual from the study by Lejuez et al [21], with adaptations to address the needs of older adults. Specific adaptations, development, and details of the program content, including the coaching component, have been documented elsewhere [14,16]. Essentially, users accessed self-help therapeutic tools and lessons through a dedicated website using a pregenerated username and password found in their user workbook for easy reference. Once logged onto the website, they navigated the web-delivered sessions through a blend of brief videos, narrated text pages, short exercises, and offline home practices. Each session was designed to be completed independently in approximately 20 to 30 minutes. Users completed sessions in their homes and had the option to complete sessions during their weekly coaching calls, based on their preferences. The participants were encouraged to complete 1 session weekly. Apart from the coaching calls, participants did not receive additional prompts to use the web program.

Coaching

Each participant received support from a trained coach for up to 10 weeks. Coaches adhered to a structured coaching guide developed and refined based on insights from our previous studies [14,16]. This guide is designed to be adaptable, ensuring an optimal balance between the benefits realized and the sources used. For instance, coaches have the flexibility to work through web-delivered sessions with participants who may be less motivated or may frequently face technical issues. By contrast, for those who are highly motivated and rarely encounter technical difficulties, coaches can encourage the independent completion of web sessions and then discuss the content during weekly coaching calls. This tailored approach, grounded in the principle of self-determination [22], seeks to cater to participants’ individual needs while maximizing staff efficiency. Coaches contacted users via SMS text messaging or phone calls to schedule weekly sessions, and reminders were sent to users before the coaching sessions, based on coach discretion. All the coaching sessions were conducted via telephone calls.

The coaches participating in this RCT included undergraduate students in psychology (n=1), students pursuing a master’s degree in social work (n=5), and a master’s-level social worker without licensure or prior psychotherapy experience. Hence, they were all considered lay coaches without specialized mental health expertise. The coaches received group supervision and individualized support from the study’s management team, with

XSL•FO
RenderX

a licensed clinical social worker (JK) serving as a clinical backup.

Measures

Primary Clinical Outcome

Depressive symptoms were assessed using the PHQ-9, which is the most commonly used outcome measure in wCBT studies [23] and has been validated for use in remote computerized delivery [24]. Scores of 5, 10, 15, and 20 on the PHQ-9 represent thresholds for mild, moderate, moderately severe, and severe depressive symptoms, respectively. A change of 5 points is deemed clinically significant; a score <10 indicates a partial response, and a score <5 denotes remission [25]. Participants completed the PHQ-9 on up to 7 occasions: at baseline, during 5 biweekly assessments throughout the trial, and at posttest assessment. For the intervention group, the PHQ-9 was completed on the web portal during sessions 1, 3, 5, 7, and 9. The control group was assessed via phone calls during weeks 1, 3, 5, 7, and 9 of the waiting period.

Usability, Acceptability, and Engagement Outcomes

Across all measures and participants, data were collected between January and August 2023. Usability was assessed during posttest assessments using the System Usability Scale (SUS), a 10-item scale commonly used to evaluate the usability of websites, software, and other human-machine systems [26]. SUS scores can range from 0 to 100, with scores ≥68 considered above average. The SUS had good internal consistency in the study sample (Cronbach α=.81).

For acceptability assessment following the intervention, we used the Treatment Evaluation Inventory (TEI), specifically designed to evaluate acceptability of depression treatments for older adults [27]. To alleviate respondents' burden during phone-based assessments, we adapted the original TEI's 7-point Likert scale responses to a 5-point scale. The modified TEI yields total scores ranging from 11 to 55, with a score ≥32 suggesting positive attitudes toward the treatment. Among the Empower@Home participants, TEI scores exhibited excellent reliability (Cronbach α=.89).

Program engagement was measured by the number of sessions completed, which were logged into the web portal and confirmed by the coaches. These measures were only administered to the intervention group participants, as they were not applicable to the control group.

Coaching

The coaches were required to document each coaching call using a structured electronic form. This form captured details such as the call duration, the type of support provided, and a narrative summary of the coaching session. Information collected in the coaching forms was used to calculate the average number and duration of coaching calls, to calculate the percentage of sessions completed independently by participants, and to describe the types of support provided during the coaching calls.

Secondary Clinical Outcomes

Secondary outcomes, including anxiety, anger, and social isolation, were measured in baseline and posttest assessments using validated instruments. Generalized Anxiety Disorder 7-item is a validated population-based survey instrument for measuring anxiety symptoms [28] (Cronbach α=.81 in this study sample). Scores of 5, 10, and 15 on the Generalized Anxiety Disorder 7-item correspond to mild, moderate, and severe anxiety, respectively. The Patient-Reported Outcome Measurement Information System (PROMIS) Anger 5a short form [29] assesses self-reported angry mood (eg, irritability; Cronbach α=.82). The PROMIS–Social Isolation 8a [30] contains 8 items that evaluate feelings of being avoided, excluded, detached, or disconnected (Cronbach α=.89). For both PROMIS measures, raw scores were converted into *t* scores, which had a mean of 50 and an SD of 10, using the conversion tables provided [31].

Physical health outcomes, including insomnia and pain, were assessed using 2 measures. Designed as a brief screening tool for insomnia, the Insomnia Severity Index has 7 questions asking about the nature and symptoms of sleep problems [32] (Cronbach α=.78). Scores of 8, 15, and 22 on the Insomnia Severity Index represent thresholds for subthreshold, moderate clinical, and severe clinical insomnia, respectively. The Pain, Enjoyment, General Activity scale includes 3 items measuring the severity of pain and its interference with enjoyment of life and general activity (Cronbach α=.93) [33].

In addition, the EQ-5D-5L is a self-report survey of global health and health-related quality of life, containing questions across 5 domains: mobility, self-care, usual activities, pain or discomfort, and anxiety or depression [34] (Cronbach α=.76). Responses to the EQ-5D-5L were converted into utility values using an algorithm for the US population [35], with a higher score corresponding to better health-related quality of life.

Potential Mediators

Potential mediators include the CBT-related mechanisms of change. A total of 14 items from the Cognitive Behavioral Therapy Skills Questionnaire [36] were used for a general assessment of CBT skill acquisition and the use of the 2 core CBT skills: cognitive restructuring and behavioral activation (Cronbach α=.83). The 9-item Behavioral Activation for Depression Scale–Short Form, which includes subscales for activation and avoidance, is often used to track changes in the behaviors hypothesized to underlie depression and specifically targeted for change by behavioral activation [37] (Cronbach α=.80). In addition, the 16-item Basic Needs Satisfaction in General Scale [38] was designed to assess satisfaction with 3 basic psychological needs: autonomy, competence, and relatedness (Cronbach α=.77). Autonomy pertains to an individual's desire to perceive their actions and subsequent consequences as self-determined, rather than being subject to external influences or control [39]. Competence refers to the inherent desire to experience a sense of efficacy and proficiency in task execution across a range of complexities [40]. Relatedness refers to the inherent human desire for social connection, support, and caring from others [40]. According to

XSL•FO
RenderX

the self-determination theory, all 3 needs must be fulfilled to achieve psychological well-being [39].

Statistical Analysis

Descriptive statistics were computed for the demographic characteristics, psychosocial outcome measures, program usability, acceptability, and engagement measures. To compare baseline characteristics between the groups, chi-squared tests were used for categorical variables, whereas 2-sample 2-tailed *t* tests were used for continuous variables. Paired 2-tailed *t* tests were used to assess within-group differences from baseline to posttest. In addition, 2-sample 2-tailed *t* tests were used to evaluate between-group differences in outcomes at the posttest stage. Linear mixed modeling was further applied to analyze within-group changes over time and compare changes in depressive symptoms between the treatment and control groups.

Causal mediation analysis [41,42] was conducted to explore the mechanisms driving the intervention effects, specifically focusing on the mediators described in the *Potential Mediators* section. Using the *mediate* command in Stata, individual models were estimated separately for each mediator [43]. Potential confounders that might influence both the mediator and the outcome were adjusted in the model, including age, gender, education, income, living arrangement, and count of chronic physical conditions. In addition, to obtain more robust estimates of indirect effects, the baseline PHQ-9 score was adjusted in the outcome equation, and the baseline score of each corresponding mediator was included in the mediator equation.

For analyses related to depressive symptoms, an intention-to-treat design was used by imputing missing values using the last observation carried forward method. Owing to the small sample size, the low incidence of missing values, and the exploratory nature of the secondary aims, analyses related to secondary outcomes and potential mediators were conducted using complete data, without imputation of missing values. Unless otherwise specified, 2-sided tests (α=.05) were used to determine statistical significance. All analyses were performed using Stata SE (version 18; StataCorp LLC).

Results

Participants

In total, 70 participants were enrolled and assigned to either the treatment or waitlist attention control group, as depicted in Figure 1. The baseline characteristics of the participants are presented in Table 1. Bivariate analyses revealed no significant differences between the treatment and control groups at baseline (all *P* values were >.05 for 2-tailed *t* tests). The participants mainly consisted of women (57/70, 81%) and were primarily non-Hispanic White (56/70, 80%). More than half held at least a college degree (40/70, 57%) and were unmarried (37/70, 53%). In addition, over half had an annual household income <US $50,000 (40/70, 58%). Most participants (60/70, 86%) owned a laptop or PC. The majority had been diagnosed with depression by a health care professional (44/70, 63%), but only around a third (25/70, 36%) reported taking antidepressants. Notably, nearly all participants (68/70, 97%) indicated some lower body mobility difficulties, with an average of approximately 4 items (mean 4.46, SD 2.53) endorsed on the mobility limitation index.

XSL•FO
RenderX

Table 1. Baseline characteristics of the study sample (N=70).

	Control group	Treatment group	Control vs treatment		
			t test (*df*)	Chi-square (*df*)	*P* value
Age (years), mean (SD; range)	67.91 (5.83; 60-85)	69.80 (5.36; 61-86)	–1.41 (68)	N/A[a]	.16
Gender, n (%)			N/A	0.9 (1)	.36
Female	30 (85)	27 (77)			
Male	5 (14)	8 (22)			
Race or ethnicity, n (%)			N/A	5.4 (4)	.25
African American or Black	4 (11)	5 (14)			
American Indian	1 (3)	0 (0)			
Asian or Pacific Islander	0 (0)	2 (6)			
White, non-Hispanic	30 (86)	26 (74)			
Mixed race or other	0 (0)	2 (5)			
Education, n (%)			N/A	2.9 (3)	.41
High school or GED[b]	4 (11)	6 (17)			
Associate's or some college	8 (23)	12 (34)			
4-year college degree	13 (37)	12 (34)			
Master's degree or above	10 (29)	5 (14)			
Household income (US $), n (%)			N/A	2.5 (3)	.47
<20,000	2 (6)	5 (15)			
20,000-49,999	17 (49)	16 (47)			
50,000-74,999	9 (26)	5 (15)			
≥75,000	7 (20)	8 (23)			
Living status, n (%)			N/A	0.1 (1)	.81
Lives with others	21 (60)	20 (57)			
Lives alone	14 (40)	15 (43)			
Marital status, n (%)			N/A	4.7 (4)	.30
Married	18 (51)	13 (37)			
Living with a partner	1 (3)	1 (3)			
Divorced or separated	6 (17)	14 (40)			
Widowed	8 (23)	5 (14)			
Never married	2 (6)	2 (6)			
Count of chronic conditions[c], mean (SD; range)	2.31 (0.99; 0-5)	2.25 (1.27; 0-5)	0.2 (68)	N/A	.83
Mobility limitation index[d], mean (SD; range)	4.20 (2.29; 1-9)	4.71 (2.76; 0-10)	–0.9 (68)	N/A	.40
Device ownership, n (%)			N/A	0.7 (2)	.69
No device ownership	2 (6)	2 (6)			
Has tablet or iPad	4 (11)	2 (6)			
Has laptop or PC	29 (83)	31 (89)			
Ever received a depression diagnosis[e], n (%)			N/A	0 (1)	.99
No	13 (37)	13 (37)			
Yes	22 (63)	22 (63)			
Taking an antidepressant, n (%)			N/A	0.6 (1)	.45

XSL•FO
RenderX

	Control group	Treatment group	Control vs treatment		
			t test (*df*)	Chi-square (*df*)	*P* value
No	21 (60)	24 (69)			
Yes	14 (40)	11 (31)			
PHQ-9[f] score, mean (SD; range)	11.29 (2.65; 8-18)	11.49 (2.63; 8-19)	–0.3 (68)	N/A	.75

[a]N/A: not applicable.

[b]GED: General Educational Development Test.

[c]The chronic disease count was the sum of self-reported conditions including hypertension, diabetes, chronic lung disease, chronic kidney disease, heart disease, stroke, arthritis, and cancer.

[d]The mobility limitation index asks if a participant has difficulty with 11 tasks: running or jogging for a mile, walking several blocks, walking one block, sitting for 2 hours, standing for 30 minutes, getting up from a chair after sitting for a long period, climbing several flights of stairs, climbing 1 flight of stairs, walking across a room, using equipment or devices when crossing a room, and getting in or out of bed. "Yeses" were summed to create a total score, ranging from 0 to 11.

[e]Participants were asked, "has a doctor, therapist, psychologist, or other health care provider ever told you that you have depression?"

[f]PHQ-9: Patient Health Questionnaire-9.

Primary Clinical Outcome

Table 2 shows the results of the linear mixed effects regression analysis of depressive symptoms. The main effect of time was significant (b=–0.49, 95% CI –0.73 to –0.26; *P*<.001), showing a decrease of nearly 0.5 points on the PHQ-9 for every biweekly assessment time point in the control group. The group-by-time interaction was also significant (b=–0.68, 95% CI –1.00 to –0.35; *P*<.001), which indicates that the treatment group experienced a faster decline in their PHQ-9 score by 0.68 points per assessment time point compared with the control group. Considering both the main and interaction effects, the PHQ-9 score in the treatment group was projected to decrease by 1.17 points every 2 weeks.

Table 2. Linear mixed modeling predicting PHQ-9[a] scores[b].

	Estimate (95% CI; SE)	*t* test	*P* value
Intercept	10.48 (9.50 to 11.45;0.50)	21.14	<.001
Group allocation			
Control group	Reference		
Treatment group	1.06 (–0.31 to 2.43; 0.70)	1.52	.13
Time	–0.49 (–0.73 to –0.26; 0.12)	–4.09	<.001
Group×Time	–0.68 (–1.00 to –0.35; 0.17)	–4.09	<.001

[a]PHQ-9: Patient Health Questionnaire-9.

[b]Time codes represent the sequential order of assessments: 0 for the baseline, 1 for the first in-app assessment, and so on, culminating with 6 for the posttest assessment. Each assessment was performed approximately 2 weeks apart.

Effect sizes were calculated using pretest scores to subtract posttest assessment scores so that a positive effect size indicates reduced depressive symptoms as measured by the PHQ-9. The treatment group had a large within-group effect size from baseline to posttest assessment (Cohen *d*=1.22; *P*<.001). By contrast, the waitlist attention control group experienced a medium within-group effect (Cohen *d*=0.57; *P*=.002). Taking everything into account, Empower@Home had a medium between-group effect size compared with waitlist attention control at the posttest (Cohen *d*=0.72; *P*=.004). Descriptive statistics were used to describe clinically significant improvements in the treatment group. Most participants in Empower@Home (21/35, 60%) experienced a clinically meaningful change in depressive symptoms, defined as a reduction of ≥5 points on the PHQ-9. Of those who initially scored ≥10 on the PHQ-9 before treatment, two-thirds (19/ 25, 76%) exhibited a partial response by scoring <10 at posttreatment. Furthermore, almost half of the Empower@Home participants (16/35, 46%) achieved a score <5 on the PHQ-9 at posttest assessment, suggesting remission.

Coaching

In total, 7 coaches provided 286 coaching calls to 35 participants, with an average of 8.2 (SD 2.1) calls per participant. Calls varied in length from 11 to 92 minutes, with an average of 49 (SD 16.9) minutes each. On the basis of the coaching notes, most participants (168/286, 58.7%) completed web-delivered lessons on their own before their coaching session. The remainder either partially completed or went through an entire web-delivered lesson along with their coach during the session. Notably, calls were shorter for those who completed sessions independently, averaging 41 (SD 16.6) minutes, compared with an average of 59 (SD 10.9) minutes when a coach guided them through the web-delivered sessions.

XSL•FO
RenderX

On the basis of the coach self-reports, the primary form of assistance provided was simple feedback, such as words of encouragement, which was given in 88.5% (253/286) of the coaching calls. This was followed by facilitating the understanding of the lesson content (147/286, 51.4%), assisting with the application of program tools (143/286, 50%), reviewing or aiding with home practices (120/286, 41.6%), and offering technical assistance (30/286, 10.5%).

Usability, Acceptability, and Engagement Outcomes

Usability, acceptability, and engagement outcomes were only pertinent to participants in the treatment group. Most participants completed all 9 program sessions (31/35, 89%), averaging 8.3 (SD 2.1) sessions.

Regarding usability, the average SUS score was 82 (SD 12.7), exceeding the benchmark score of 68, indicating that participants perceived the program as usable. A vast majority (30/35, 88%) had an SUS score of ≥68, and all participants (35/35, 100%) agreed or strongly agreed that the program was easy to use.

For acceptability, the average TEI score was 45.9 (SD 6.6), surpassing the benchmark score of 32, indicating favorable attitudes toward the treatment. Most participants (30/35, 88%) stated that they would recommend this program to others who experience depressed moods. However, about a quarter of the participants (9/35, 26%) mentioned that they "experienced discomfort during this program." Further analysis of their feedback from open-ended follow-up questions revealed that most of these responses stemmed from the program evoking distressing emotions or recalling memories from their past. Interestingly, 2 participants clarified that they perceived this discomfort as positive ("good discomfort"), describing it as a constructive force encouraging them to change their behaviors. Furthermore, 3 participants either agreed or strongly agreed that the program had undesirable side effects; analyzing their open-ended feedback mirrored sentiments from the discomfort-related feedback, indicating that the program helped them confront their issues more directly.

Secondary Clinical Outcomes

Table 3 displays the within- and between-group effect sizes for the secondary outcomes.

XSL•FO
RenderX

Table 3. Within-group and between-group differences for secondary outcomes and mediators[a].

Measure and condition	Baseline, mean (SD)	Posttest assessment, mean (SD)	Within-group difference			Between-group difference at posttest assessment[b]		
			Cohen *d*	*t* test (*df*)	*P* value	Cohen *d* (95% CI)	*t* test (*df*)	*P* value
GAD-7[c]								
Treatment	7.85 (3.73)	5.06 (4.71)	0.60	3.48 (33)	.001	0.04 (–0.45 to 0.54)	0.17 (62)	.87
Control	6.17 (5.25)	5.23 (3.38)	0.27	1.48 (29)	.15	N/A[d]	N/A	N/A
PROMIS-anger[e]								
Treatment	55.20 (6.52)	50.06 (9.08)	0.65	3.79 (33)	<.001	0.27 (–0.22 to 0.77)	1.48 (62)	.16
Control	52.67 (7.06)	52.32 (7.14)	0.06	0.34 (29)	.74	N/A	N/A	N/A
PROMIS-social isolation								
Treatment	54.81 (8.62)	52.33 (7.98)	0.41	2.37 (33)	.02	0.31 (–0.18 to 0.80)	1.24 (62)	.22
Control	53.75 (6.85)	54.48 (5.52)	–0.14	–0.75 (29)	.46	N/A	N/A	N/A
Insomnia Severity Index								
Treatment	13.68 (5.95)	10.74 (6.81)	–0.49	2.88 (33)	.007	0.21 (–0.28 to 0.70)	0.83 (62)	.41
Control	12.33 (4.58)	12.00 (4.68)	0.09	0.50 (29)	.62	N/A	N/A	N/A
PEG[f]								
Treatment	4.44 (2.79)	3.54 (2.50)	0.39	2.30 (33)	.03	0.35 (–0.14 to 0.84)	1.40 (62)	.17
Control	4.36 (2.37)	4.38 (2.33)	–0.02	–0.10 (29)	.92	N/A	N/A	N/A
EQ-5D-5L								
Treatment	0.72 (0.21)	0.74 (0.20)	0.12	0.68 (33)	.50	0.33 (–0.15 to 0.82)	1.31 (62)	.20
Control	0.72 (0.25)	0.66 (0.27)	–0.26	–1.14 (29)	.17	N/A	N/A	N/A
CBTSQ[g]								
Treatment	40.85 (8.73)	50.26 (8.11)	1.11	6.45 (33)	<.001	0.67 (0.17 to 1.17)	2.69 (62)	.009
Control	44.97 (8.08)	44.90 (7.81)	–0.01	–0.05 (29)	.96	N/A	N/A	N/A
BADS-SF[h]								
Treatment	26.53 (12.22)	34.21 (9.98)	0.64	3.73 (33)	<.001	0.43 (–0.07 to 0.92)	1.70 (62)	.09
Control	27.50 (10.31)	29.77 (10.93)	0.19	1.05 (29)	.30	N/A	N/A	N/A
BNSG-S[i]								
Treatment	62.18 (9.44)	66.91 (8.11)	0.70	4.11 (33)	<.001	0.52 (0.02 to 1.02)	2.08 (62)	.04
Control	61.0 (8.75)	62.3 (9.64)	0.22	1.23 (29)	.23	N/A	N/A	N/A

[a]There were 34 participants in the treatment group and 30 in the control group. Within-group differences were evaluated using paired 2-tailed *t* tests for each measure or group separately. Between-group differences were evaluated using 2-sample 2-tailed *t* tests for each measure separately. Analyses were conducted using complete data without imputation. Effect sizes were calculated such that a positive effect size indicated changes in the desired direction (ie, increased quality of life, cognitive behavioral therapy skill acquisition, behavioral activation, and basic needs satisfaction and decreased anxiety, pain, insomnia severity, anger, and social isolation).

[b]Between-group difference at posttest assessment is calculated by comparing the treatment and control condition, cells with a dash correspond with nonapplicable values.

[c]GAD-7: Generalized Anxiety Disorder scale, 7-item.

[d]N/A: not applicable.

[e]PROMIS: Patient-Reported Outcome Measurement Information System.

XSL•FO
RenderX

[f]PEG: Pain, Enjoyment, General Activity.

[g]CBTSQ: Cognitive Behavioral Therapy Skills Questionnaire.

[h]BADS-SF: Behavioral Activation for Depression Scale–Short Form.

[i]BNSG-S: Basic Needs Satisfaction in General Scale.

In the treatment group, significant reductions were noted from baseline to posttest in multiple domains: anxiety (Cohen d=0.60; t_{33}=3.48; P=.001), anger (Cohen d=0.65; t_{33}=3.79; P<.001), social isolation (Cohen d=0.41; t_{33}=2.37; P=.02), insomnia severity (Cohen d=0.49; t_{33}=2.88; P=.007), and pain intensity and interference (Cohen d=0.39; t_{33}=2.30; P=.03). By contrast, the control group showed only minor absolute within-group changes, and none of these changes reached statistical significance. Although there were no statistically significant between-group differences in these outcomes owing to the small sample size, the findings were consistent with our expectations that the treatment group would exhibit greater improvements in the secondary outcome measures from baseline to posttest.

The within-group change in health-related quality of life, as measured by EQ-5D-5L, was not statistically significant for either the treatment or the control group. Notably, the treatment group experienced a slight increase in the quality of life, which aligns with our expectations, whereas a decrease was observed in the control group. Although the differences were not statistically significant, the between-group effect was more pronounced than the within-group effect owing to the divergent trends observed in each group.

Mediation Analysis

As shown in Table 3, within the treatment group, significant increases were observed at posttest assessment from baseline in all 3 mediators: CBT skills acquisition (Cohen d=1.11; t_{33}=6.45; P<.001), behavioral activation (Cohen d=0.64; t_{33}=3.73; P<.001), and satisfaction with basic psychological needs (Cohen d=0.70; t_{33}=4.11; P<.001). By contrast, the control group showed only minor absolute changes within group, and none of these changes reached statistical significance.

At posttest assessment, significant between-group differences were observed in CBT skills acquisition (Cohen d=0.67; t_{33}=2.69; P=.009) and basic needs satisfaction (Cohen d=0.52; t_{33}=2.08; P=.04). Moreover, the between-group difference in behavioral activation approached the critical value (Cohen d=0.43; t_{33}=1.70; P=.09), reaching significance in a 1-tailed test.

As illustrated in Table 4, the indirect effects associated with all 3 mediators were statistically significant. This suggests that the effect of the intervention on depression was partially mediated by an increase in CBT skills acquisition (b=–2.29, 95% CI –3.58 to –1.01; P<.001), behavioral activation (b=–1.27, 95% CI –2.39 to –0.15; P=.03), and satisfaction with basic psychological needs (b=–1.38, 95% CI –2.45 to –0.31; P=.01). The largest proportion of the mediation effect was seen in CBT skills acquisition (62.9%), followed by satisfaction with basic psychological needs (46.8%), and behavioral activation (34.5%). This proportion was calculated as the ratio of the natural indirect effect to the total effect.

Table 4. Indirect and direct effects from causal mediation analysis on depressive symptoms at posttreatment (n=64)[a].

Mediator	Nature indirect effect		Nature direct effect		Proportion mediated	
	b (95% CI)	P value	b (95% CI)	P value	Percentage, %	P value
CBTSQ[b]	–2.29 (–3.58 to –1.01)	<.001	–1.35 (–3.38 to 0.68)	.19	62.9	.005
BADS-SF[c]	–1.27 (–2.39 to –.15)	.03	–2.41 (–4.09 to –0.72)	.005	34.5	.02
BNSG-S[d]	–1.38 (–2.45 to –.31)	.01	–1.57 (–3.36 to 0.23)	.09	46.8	.02

[a]The outcome was depressive symptoms, as measured by the Patient Health Questionnaire-9, with a higher score indicating greater symptoms. The control group was used as the reference in the treatment model. Each mediator was evaluated individually in separate models. Each model included potential confounders, such as age, gender, education, income, living arrangement, and count of chronic physical conditions, in both the mediator and outcome models. Baseline depression was included in the outcome equation, and the baseline score of each mediator was included in the mediator equation. In addition, the outcome equation includes the treatment-mediator interaction. Analyses were conducted using complete data without imputation).

[b]CBTSQ: Cognitive Behavioral Therapy Skills Questionnaire.

[c]BADS-SF: Behavioral Activation for Depression Scale–Short Form.

[d]BNSG-S: Basic Needs Satisfaction in General Scale.

Discussion

Principal Findings

This RCT assessed the efficacy of layperson-supported Empower@Home, a wCBT program specifically designed to alleviate depression among older adults, compared with a waitlist attention control group that received weekly friendly calls. The retention rate in the study was excellent, with 91% (64/70) of the participants engaging in the posttest interview, surpassing the 80% benchmark. Moreover, the intervention engagement rate was high, with 89% (31/35) of individuals in the treatment group completing all 9 program sessions. To put this into perspective, the average completion rate of wCBT programs, which is typically defined as completing 80% of treatment lessons, is only 17% for self-administered interventions and 65% for supported interventions [44,45]. In addition, Empower@Home demonstrated a large within-group

XSL•FO
RenderX

effect in the treatment group (Cohen *d*=1.22) and a medium between-group effect accounting for the attention control group at posttest assessment (Cohen *d*=0.72). These findings suggest that the novel intervention is acceptable to older adults and more efficacious in alleviating depression in older adults than weekly friendly calls and symptom monitoring.

Several factors may have contributed to the notably higher engagement rate observed with Empower@Home. Supported wCBT programs tend to exhibit better completion rates compared with self-guided programs, as evidenced by previous studies [44,45]. The presence of coaches likely played a crucial role in this regard by enhancing accountability, fostering interactivity within the program, and establishing a routine for session participation. Another potential factor is the increased global demand for mental health treatments, which has gained momentum during the COVID-19 pandemic [46]. Reports from our community partners who serve older adults in Michigan align with these global observations. Participants may have been motivated to complete the program in their pursuit of improved well-being, especially when faced with barriers in accessing alternative treatment options. Furthermore, although a monetary incentive of up to US $100 was available to the participants, it is improbable that this had a substantial impact on program engagement. This is because our research staff consistently emphasized that payments were not contingent on program completion. In addition, it is worth noting that studies offering higher incentive amounts have sometimes seen lower engagement rates compared with our study [9].

The findings concerning the program's usability, acceptability, and effects observed in this RCT align with those found in a previous uncontrolled study in which an initial version of the program was examined [14]. In comparison with the uncontrolled study, individuals in this RCT were older by an average of 5 years (68.9 vs 63.7), and this RCT had a larger percentage of individuals with at least 3 physical health conditions (31/70, 44% vs 29/103, 28.2%). However, the educational background and income distribution of the participants were similar across studies. More than half of the participants (45/70, 64%) in the RCT were reached through social media advertisements, a method not used in the uncontrolled study. Furthermore, most of the coaches (n=4) involved in the RCT were new and did not participate in the preceding uncontrolled study. Considering all factors, these consistent findings enhance the likelihood that the study results can be replicated under varied conditions.

The layperson-supported Empower@Home program showed a large within-group effect (Cohen *d*=1.22), aligning with previous wCBT trials with older adults, which had a pooled effect size of 1.27 as reported in a meta-analysis evaluating wCBT programs with varying levels of human support [23]. Meanwhile, the between-group effect was medium (Cohen *d*=0.72), which appears smaller than the average of 1.18 reported in the meta-analysis [23]. Notably, none of the controlled trials in the meta-analysis featured a potent attention control condition such as the one used in this study; instead, they used usual care or waitlist control. In this study, the attention control group also showed a significant reduction in depressive symptoms, as indicated by a medium within-group change (Cohen *d*=0.57). Given that empathy-oriented telephone programs administered by lay callers have been shown to improve depression and mental health among older adults [47], the diminished between-group effect observed here, which results from significant improvements in the control group, is understandable. This observation is consistent with a recent RCT involving homebound older adults with depression, in which a tele-delivered behavioral activation treatment conducted by lay counselors demonstrated a medium effect (Cohen *d*=0.62, 95% CI 0.35-0.89) compared with an attention control that received weekly support telephone calls [48].

The clinical efficacy of the intervention was bolstered by the change mechanisms identified through the causal mediation analysis. In CBT, depressive symptoms are believed to result from maladaptive thoughts and behavioral patterns [49]. Consistent with the theory underlying CBT, we found that reductions in depressive symptoms among participants were partially explained by cognitive restructuring and behavioral activation, as measured by the Cognitive Behavioral Therapy Skills Questionnaire and Behavioral Activation for Depression Scale–Short Form. In addition to these well-known change mechanisms inherent in CBT treatment, we found that the treatment also improved satisfaction with basic psychological needs through an enhanced sense of autonomy, competence, and relatedness, which, in turn, contributed to the reduction of depressive symptoms. This finding resonates with the self-determination theory, which underscores the importance of autonomy, competence, and relatedness in fostering intrinsic motivation [50] and posits that fulfilling these needs is essential for attaining psychological well-being.

This study provides evidence that laypersons lacking specialized mental health skills can effectively offer human support in digital mental health interventions such as wCBT programs. Previous research has shown that supported interventions, in which individuals receive human assistance, tend to enhance engagement and adherence, resulting in more significant positive outcomes compared with unsupported interventions [17,18,51]. However, most existing wCBT programs typically involve therapists with specialized mental health training. Given the scarcity of mental health professionals specialized in working with older adults, particularly in underserved areas, a pragmatic approach may involve training laypersons to assist users of wCBT programs. As the core therapeutic components are preprogrammed, the training requirements for wCBT supporters are significantly reduced. This approach makes layperson-supported wCBT a potentially cost-effective and scalable alternative to therapist-supported interventions. Notably, Titov et al [52] showed that layperson-supported wCBT for depression was as effective as a clinician-supported intervention in an RCT involving adults (mean age 44, SD 12.3 years). Although research specific to older adults remains limited, Tomasino et al [9] showed that a wCBT program supported by peers was well received and associated with a significant reduction in depressive symptoms in a small sample of older adults with depression. In another study focusing on low-income homebound older adults, Choi et al [48] reported that behavioral activation treatment administered via telecommunication and led by lay counselors with bachelor's degrees was more effective

XSL•FO
RenderX

than the control condition of weekly support calls. Although the study by Choi et al [48] also found that the lay counselor-led intervention was not as effective as tele-delivered problem-solving therapy conducted by clinicians in alleviating depression, it yielded comparable results in terms of secondary outcomes. These emerging findings provide support for the feasibility and potential cost-effectiveness of having laypersons as human supporters of wCBT programs.

The findings of this study have important practical implications. wCBT programs, such as Empower@Home, can serve as valuable resources for health care professionals working with older adults, including case managers, care coordinators, and primary care physicians. These tools can be especially beneficial for individuals facing persistent treatment barriers such as transportation issues and financial constraints. Although challenges related to technology access and digital literacy may exist, this study shows that with guidance and support, older adults are generally receptive to using wCBT programs. However, it is important to note that wCBT programs should be considered as complementary tools rather than replacements for mental health professionals. They can supplement the traditional one-on-one therapy provided by mental health professionals by offering additional, easily accessible resources for clients to use outside of therapy sessions. wCBT programs can also serve as intermediate support for clients waiting for traditional treatments or as a means to enhance the continuum of care. By recognizing the role of wCBT in augmenting mental health care, practitioners across various health care disciplines can better address the mental health needs of their clients, especially in underserved or marginalized communities.

Limitations

This study was limited by its small sample size, rendering it underpowered for the detection of small effect sizes. Although our study was powered to detect a medium effect in the primary clinical outcome of depression based on a linear mixed effects model with up to 7 assessments [3], it was not powered to detect a small effect in the secondary clinical outcomes based on 2-sample 2-tailed *t* tests. We performed an ad hoc power analysis and found that a sample size of 204 is necessary to identify an effect size of 0.35 with a power of 80%, using a 2-sample 2-tailed *t* test, as per G*Power 3.1.9.7 [53]. Moreover, without long-term follow-up, it is unclear whether the treatment effect observed at posttest assessment would be sustained in the long term. In addition, when considering the generalizations of the study findings, it is important to take into account the sources of participant recruitment and their characteristics. Although participants came from across Michigan, including metropolitan and rural areas, their education levels and technology device ownership exceeded the national and state averages among older adults [54,55]. The overrepresentation of college-educated and technology-savvy participants was not unexpected because most participants were recruited from social media advertisements and a research volunteer registry that required internet access.

Conclusions

wCBT, tailored specifically for older adults and augmented with support from trained laypersons, is efficacious in reducing depressive symptoms compared with friendly telephone calls and symptom monitoring. The intervention works by enhancing several change mechanisms, including facilitating the acquisition of CBT skills, promoting behavioral activation, and fostering self-determination, which manifests as a heightened sense of autonomy, competence, and relatedness. Future well-designed trials are needed to evaluate the effectiveness of the intervention in community and practice settings, leveraging nonclinician staff already present in these real-world settings as wCBT coaches.

Acknowledgments

This study received grant funding from the Michigan Health Endowment Fund. Participant recruitment used the resources provided by the Michigan Center for Urban African American Aging Research, which was funded by the National Institutes of Health (5P30 AG015281). Data collection and management used Research Electronic Data Capture (REDCap; Vanderbilt University) electronic data capture tools hosted at the University of Michigan, which was supported by the Michigan Institute for Clinical and Health Research and funded by the National Institutes of Health (UM1TR004404). Many individuals and entities that were not listed as coauthors contributed to the study's implementation. Alex Halavanau and Robert LaRoe provided technical support to ensure the seamless operation of the program's web portal. The following students contributed as coaches for participants in the program: Asha Dimoji, Charlotte Drumm, Hanuel Chun, Natalie Francis, Elena McLellan, Sarah Minnis, Nicole Kodkany, Sarai Koster-Stetson, Sofia Ruiz-Sierra, and Glory Kubicek. The following students conducted friendly visit calls for waitlist attention control participants: Kayla Drifka, Carly Gosdzinski, and Madeline Poupard. Sari Bircoll conducted assessments. The following students provided administrative and recruitment support: Luke Estey, Ella Baranski, Emily Clausen, Nahush Keshav, and Zilin Zhu. The authors thank the social service providers who provided client referrals: A&D Home Health Care, Ann Arbor Meals on Wheels, EightCAP Inc, Grand Traverse Area Agency on Aging, Healthier Black Elders Center, Huron Valley Pace, Jewish Family Service of Metro Detroit, Leelanau County Senior Services, Northern Health Care Management, Region VII Area Agency on Aging, Shiawassee Council on Aging, National Church Residences, Related Property Management, and Ypsilanti Meals on Wheels.

XSL•FO
RenderX

Authors' Contributions

XX led the design of Empower@Home, conceptualized the study, conducted data analysis, and drafted the manuscript. JK, SA, and JAH contributed to the design of Empower@Home and reviewed the manuscript. ST coordinated subject recruitment and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V 1.6.1)
[PDF File (Adobe PDF File), 16138 KB - jmir_v26i1e53001_app1.pdf]

References

1. Zenebe Y, Akele B, W/Selassie M, Necho M. Prevalence and determinants of depression among old age: a systematic review and meta-analysis. Ann Gen Psychiatry 2021 Dec 18;20(1):55 [FREE Full text] [doi: 10.1186/s12991-021-00375-x] [Medline: 34922595]
2. Pepin R, Segal DL, Coolidge FL. Intrinsic and extrinsic barriers to mental health care among community-dwelling younger and older adults. Aging Ment Health 2009 Sep;13(5):769-777. [doi: 10.1080/13607860902918231] [Medline: 19882416]
3. Kayser J, Wang X, Wu Z, Dimoji A, Xiang X. Layperson-facilitated internet-delivered cognitive behavioral therapy for homebound older adults with depression: protocol for a randomized controlled trial. JMIR Res Protoc 2023 Feb 22;12:e44210 [FREE Full text] [doi: 10.2196/44210] [Medline: 36811937]
4. Muñoz RF. Using evidence-based internet interventions to reduce health disparities worldwide. J Med Internet Res 2010 Dec 17;12(5):e60 [FREE Full text] [doi: 10.2196/jmir.1463] [Medline: 21169162]
5. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
6. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
7. Weaver A, Zhang A, Xiang X, Felsman P, Fischer DJ, Himle JA. Entertain me well: an entertaining, tailorable, online platform delivering CBT for depression. Cogn Behav Pract 2023 Feb;30(1):96-115 [FREE Full text] [doi: 10.1016/j.cbpra.2021.09.003] [Medline: 36741472]
8. Dear BF, Zou J, Titov N, Lorian C, Johnston L, Spence J, et al. Internet-delivered cognitive behavioural therapy for depression: a feasibility open trial for older adults. Aust N Z J Psychiatry 2013 Feb;47(2):169-176. [doi: 10.1177/0004867412466154] [Medline: 23152358]
9. Tomasino KN, Lattie EG, Ho J, Palac HL, Kaiser SM, Mohr DC. Harnessing peer support in an online intervention for older adults with depression. Am J Geriatr Psychiatry 2017 Oct;25(10):1109-1119 [FREE Full text] [doi: 10.1016/j.jagp.2017.04.015] [Medline: 28571785]
10. Titov N, Dear BF, Ali S, Zou JB, Lorian CN, Johnston L, et al. Clinical and cost-effectiveness of therapist-guided internet-delivered cognitive behavior therapy for older adults with symptoms of depression: a randomized controlled trial. Behav Ther 2015 Mar;46(2):193-205. [doi: 10.1016/j.beth.2014.09.008] [Medline: 25645168]
11. Hobbs MJ, Joubert AE, Mahoney AE, Andrews G. Treating late-life depression: comparing the effects of internet-delivered cognitive behavior therapy across the adult lifespan. J Affect Disord 2018 Jan 15;226:58-65. [doi: 10.1016/j.jad.2017.09.026] [Medline: 28963865]
12. McMurchie W, Macleod F, Power K, Laidlaw K, Prentice N. Computerised cognitive behavioural therapy for depression and anxiety with older people: a pilot study to examine patient acceptability and treatment outcome. Int J Geriatr Psychiatry 2013 Nov 14;28(11):1147-1156. [doi: 10.1002/gps.3935] [Medline: 23495124]
13. O'Moore KA, Newby JM, Andrews G, Hunter DJ, Bennell K, Smith J, et al. Internet cognitive-behavioral therapy for depression in older adults with knee osteoarthritis: a randomized controlled trial. Arthritis Care Res (Hoboken) 2018 Jan 15;70(1):61-70. [doi: 10.1002/acr.23257] [Medline: 28426917]
14. Xiang X, Kayser J, Turner S, Zheng C. Layperson-supported internet-delivered cognitive behavioral therapy for depression among older adults. Res Soc Work Pract 2023 Jun 26 [FREE Full text] [doi: 10.1177/10497315231184143]
15. Xiang X, Kayser J, Sun Y, Himle J. Internet-based psychotherapy intervention for depression among older adults receiving home care: qualitative study of participants' experiences. JMIR Aging 2021 Nov 22;4(4):e27630 [FREE Full text] [doi: 10.2196/27630] [Medline: 34813491]
16. Xiang X, Kayser J, Ash S, Zheng C, Sun Y, Weaver A, et al. Web-based cognitive behavioral therapy for depression among homebound older adults: development and usability study. JMIR Aging 2023 Oct 19;6:e47691 [FREE Full text] [doi: 10.2196/47691] [Medline: 37725423]

XSL•FO
RenderX

17. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
18. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
19. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205. [doi: 10.1080/16506070903318960] [Medline: 20183695]
20. Gallagher-Thompson D, Thompson L. Applying cognitive-behavioral therapy to the psychological problems of later life. In: Zarit SH, Knight BG, editors. A Guide to Psychotherapy and Aging: Effective Clinical Interventions in a Life-Stage Context. Washington, DC: American Psychological Association; 1996:61-82.
21. Lejuez CW, Hopko DR, Hopko SD. A brief behavioral activation treatment for depression: treatment manual. Behav Modif 2001 Apr;25(2):255-286. [doi: 10.1177/0145445501252005] [Medline: 11317637]
22. Dwyer LA, Hornsey MJ, Smith LG, Oei TP, Dingle GA. Participant autonomy in cognitive behavioral group therapy: an integration of self-determination and cognitive behavioral theories. J Soc Clin Psychol 2011 Jan;30(1):24-46. [doi: 10.1521/jscp.2011.30.1.24]
23. Xiang X, Wu S, Zuverink A, Tomasino KN, An R, Himle JA. Internet-delivered cognitive behavioral therapies for late-life depressive symptoms: a systematic review and meta-analysis. Aging Ment Health 2020 Aug;24(8):1196-1206 [FREE Full text] [doi: 10.1080/13607863.2019.1590309] [Medline: 30913898]
24. Erbe D, Eichert HC, Rietz C, Ebert D. Interformat reliability of the patient health questionnaire: validation of the computerized version of the PHQ-9. Internet Interv 2016 Sep;5(4):1-4 [FREE Full text] [doi: 10.1016/j.invent.2016.06.006] [Medline: 30135800]
25. Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
26. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Int 2018 Mar 30;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
27. Landreville P, Guérette A. Psychometric properties of a modified version of the treatment evaluation inventory for assessing the acceptability of treatments for geriatric depression. Can J Aging 2010 Nov 29;17(4):414-424 [FREE Full text] [doi: 10.1017/s071498080001268x]
28. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
29. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, PROMIS Cooperative Group. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS®): depression, anxiety, and anger. Assessment 2011 Sep;18(3):263-283 [FREE Full text] [doi: 10.1177/1073191111411667] [Medline: 21697139]
30. Carlozzi NE, Ianni PA, Lange RT, Brickell TA, Kallen MA, Hahn EA, et al. Understanding health-related quality of life of caregivers of civilians and service members/veterans with traumatic brain injury: establishing the reliability and validity of PROMIS social health measures. Arch Phys Med Rehabil 2019 Apr;100(4S):S110-S118 [FREE Full text] [doi: 10.1016/j.apmr.2018.06.026] [Medline: 30075148]
31. Scoring instructions. Health Measures. URL: https://www.healthmeasures.net/score-and-interpret/calculate-scores/scoring-instructions [accessed 2024-01-18]
32. Bastien CH, Vallières A, Morin CM. Validation of the insomnia severity index as an outcome measure for insomnia research. Sleep Med 2001 Jul;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]
33. Krebs EE, Lorenz KA, Bair MJ, Damush TM, Wu J, Sutherland JM, et al. Development and initial validation of the PEG, a three-item scale assessing pain intensity and interference. J Gen Intern Med 2009 Jun;24(6):733-738 [FREE Full text] [doi: 10.1007/s11606-009-0981-1] [Medline: 19418100]
34. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990 Dec;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9] [Medline: 10109801]
35. Pickard AS, Law EH, Jiang R, Pullenayegum E, Shaw JW, Xie F, et al. United States valuation of EQ-5D-5L health states using an international protocol. Value Health 2019 Aug;22(8):931-941 [FREE Full text] [doi: 10.1016/j.jval.2019.02.009] [Medline: 31426935]
36. Jacob KL, Christopher MS, Neuhaus EC. Development and validation of the cognitive-behavioral therapy skills questionnaire. Behav Modif 2011 Nov 05;35(6):595-618. [doi: 10.1177/0145445511419254] [Medline: 21893554]
37. Manos RC, Kanter JW, Luo W. The behavioral activation for depression scale-short form: development and validation. Behav Ther 2011 Dec;42(4):726-739. [doi: 10.1016/j.beth.2011.04.004] [Medline: 22036000]
38. Johnston MM, Finney SJ. Measuring basic needs satisfaction: evaluating previous research and conducting new psychometric evaluations of the basic needs satisfaction in general scale. Contemp Educ Psychol 2010 Oct;35(4):280-296. [doi: 10.1016/j.cedpsych.2010.04.003]
39. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]

XSL•FO
RenderX

40. Ryan RM, Deci EL. Overview of self-determination theory: an organismic-dialectical perspective. In: Deci EL, Ryan RM, editors. Handbook of Self-Determination Research. New York, NY: University of Rochester Press; 2002:3-33.
41. Imai K, Jo B, Stuart EA. Using potential outcomes to understand causal mediation analysis: comment on. Multivariate Behav Res 2011 Sep 30;46(5):861-873 [FREE Full text] [doi: 10.1080/00273171.2011.606743] [Medline: 23788819]
42. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. Statist Sci 2010 Feb 1;25(1):51-71. [doi: 10.1214/10-sts321]
43. Stata Causal Inference and Treatment-Effects Estimation Reference Manual. College Station, TX: Stata Press; 2023.
44. van Ballegooijen W, Cuijpers P, van Straten A, Karyotaki E, Andersson G, Smit JH, et al. Adherence to internet-based and face-to-face cognitive behavioural therapy for depression: a meta-analysis. PLoS One 2014;9(7):e100674 [FREE Full text] [doi: 10.1371/journal.pone.0100674] [Medline: 25029507]
45. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
46. Mental health and COVID-19: early evidence of the pandemic's impact. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [accessed 2024-01-18]
47. Kahlon MK, Aksan N, Aubrey R, Clark N, Cowley-Morillo M, Jacobs EA, et al. Effect of layperson-delivered, empathy-focused program of telephone calls on loneliness, depression, and anxiety among adults during the COVID-19 pandemic: a randomized clinical trial. JAMA Psychiatry 2021 Jun 01;78(6):616-622 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0113] [Medline: 33620417]
48. Choi NG, Marti CN, Wilson NL, Chen GJ, Sirrianni L, Hegel MT, et al. Effect of telehealth treatment by lay counselors vs by clinicians on depressive symptoms among older adults who are homebound: a randomized clinical trial. JAMA Netw Open 2020 Aug 03;3(8):e2015648 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15648] [Medline: 32865577]
49. McGinn LK. Cognitive behavioral therapy of depression: theory, treatment, and empirical status. Am J Psychother 2000 Apr;54(2):257-262 [FREE Full text] [doi: 10.1176/appi.psychotherapy.2000.54.2.257] [Medline: 10928248]
50. Ryan RM, Deci EL. A self-determination theory approach to psychotherapy: the motivational basis for effective change. Can Psychol 2008 Aug;49(3):186-193. [doi: 10.1037/a0012753]
51. Andersson G, Waara J, Jonsson U, Malmaeus F, Carlbring P, Ost LG. Internet-based self-help versus one-session exposure in the treatment of spider phobia: a randomized controlled trial. Cogn Behav Ther 2009;38(2):114-120. [doi: 10.1080/16506070902931326] [Medline: 20183690]
52. Titov N, Andrews G, Davies M, McIntyre K, Robinson E, Solley K. Internet treatment for depression: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010 Jun 08;5(6):e10939 [FREE Full text] [doi: 10.1371/journal.pone.0010939] [Medline: 20544030]
53. Kang H. Sample size determination and power analysis using the G*Power software. J Educ Eval Health Prof 2021 Jul 30;18:17 [FREE Full text] [doi: 10.3352/jeehp.2021.18.17] [Medline: 34325496]
54. Perrin A. Mobile technology and home broadband. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/ [accessed 2024-01-18]
55. Educational attainment in the United States. United States Census Bureau. 2022. URL: https://www.census.gov/newsroom/press-releases/2023/educational-attainment-data.html [accessed 2024-01-18]

Abbreviations

CBT: cognitive behavioral therapy
PHQ-9: Patient Health Questionnaire-9
PROMIS: Patient-Reported Outcome Measurement Information System
RCT: randomized controlled trial
SUS: System Usability Scale
TEI: Treatment Evaluation Inventory
wCBT: web-delivered cognitive behavioral therapy

Edited by G Eysenbach; submitted 21.09.23; peer-reviewed by M Zhou, K Kast; comments to author 16.10.23; revised version received 27.10.23; accepted 16.01.24; published 04.03.24.

Please cite as:
Xiang X, Kayser J, Turner S, Ash S, Himle JA
Layperson-Supported, Web-Delivered Cognitive Behavioral Therapy for Depression in Older Adults: Randomized Controlled Trial
J Med Internet Res 2024;26:e53001
URL: https://www.jmir.org/2024/1/e53001
doi:10.2196/53001
PMID:38437013

XSL•FO
RenderX

©Xiaoling Xiang, Jay Kayser, Skyla Turner, Samson Ash, Joseph A Himle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of a Social Media Intervention on Vaping Intentions: Randomized Dose-Response Experiment

William Douglas Evans[1], PhD; Jeffrey Bingenheimer[1], PhD; Jennifer Cantrell[2], PhD; Jennifer Kreslake[3], PhD; Shreya Tulsiani[3], MPH; Megumi Ichimiya[1], MPH; Alexander P D'Esterre[3], MPH; Raquel Gerard[1], MPH; Madeline Martin[1], MPH; Elizabeth C Hair[3], PhD

[1]Milken Institute School of Public Health, The George Washington University, Washington, DC, United States

[2]New York University School of Global Public Health, New York, NY, United States

[3]Truth Initiative, Washington, DC, United States

Corresponding Author:
William Douglas Evans, PhD
Milken Institute School of Public Health
The George Washington University
950 New Hampshire Ave NW
Washington, DC, 20052
United States
Phone: 1 2023519546
Email: wdevans@gwu.edu

Abstract

Background: e-Cigarette use, especially by young adults, is at unacceptably high levels and represents a public health risk factor. Digital media are increasingly being used to deliver antivaping campaigns, but little is known about their effectiveness or the dose-response effects of content delivery.

Objective: The objectives of this study were to evaluate (1) the effectiveness of a 60-day antivaping social media intervention in changing vaping use intentions and beliefs related to the stimulus content and (2) the dose-response effects of varying levels of exposure to the intervention on vaping outcomes, including anti-industry beliefs, vaping intentions, and other attitudes and beliefs related to vaping.

Methods: Participants were adults aged 18 to 24 years in the United States. They were recruited into the study through Facebook (Meta Platforms) and Instagram (Meta Platforms), completed a baseline survey, and then randomized to 1 of the 5 conditions: 0 (control), 4, 8, 16, and 32 exposures over a 15-day period between each survey wave. Follow-up data were collected 30 and 60 days after randomization. We conducted stratified analyses of the full sample and in subsamples defined by the baseline vaping status (never, former, and current). Stimulus was delivered through Facebook and Instagram in four 15-second social media videos focused on anti-industry beliefs about vaping. The main outcome measures reported in this study were self-reported exposure to social media intervention content, attitudes and beliefs about vaping, and vaping intentions. We estimated a series of multivariate linear regressions in Stata 17 (StataCorp). To capture the dose-response effect, we assigned each study arm a numerical value corresponding to the number of advertisements (exposures) delivered to participants in each arm and used this number as our focal independent variable. In each model, the predictor was the treatment arm to which each participant was assigned.

Results: The baseline sample consisted of 1491 participants, and the final analysis sample consisted of 57.28% (854/1491) of the participants retained at the 60-day follow-up. We compared the retained participants with those lost to follow-up and found no statistically significant differences across demographic variables. We found a significant effect of the social media treatment on vaping intentions ($\beta=-0.138$, 95% CI −0.266 to −0.010; $P=.04$) and anti-industry beliefs ($\beta=-0.122$, 95% CI 0.008-0.237; $P=.04$) targeted by the intervention content among current vapers but not among the full sample or other strata. We found no significant effects of self-reported exposure to the stimulus.

Conclusions: Social media interventions are a promising approach to preventing vaping among young adults. More research is needed on how to optimize the dosage of such interventions and the extent to which long-term exposure may affect vaping use over time.

Trial Registration: ClinicalTrials.gov NCT04867668; https://clinicaltrials.gov/study/NCT04867668

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50741) doi:10.2196/50741

KEYWORDS

randomized controlled trial; e-cigarettes; vaping; nicotine; tobacco control; social media; dose-response effects

Introduction

Background

e-Cigarettes were the most commonly used tobacco product among young adults in the United States between 2014 and 2019 [1]. In 2019, the current use of cigarettes and e-cigarettes was 4.5% and 14% [2], respectively. Although e-cigarette use among this population has decreased in recent years, use prevalence still remains at concerning levels. Moreover, the most popular e-cigarette brands contain high levels of nicotine, an addictive substance that may harm the developing brain of young adults [3-6]. The use of e-cigarettes has also been associated with worsened lung health and mental well-being [7-11].

Digital media, including social media platforms, have become a part of our daily lives, particularly among young adults. The use of any social media site by adults aged 18 to 29 years has been consistently >80% since 2011, and many people spend several hours a day on these sites (Pew Research Center, 2021) [12]. Because of its ubiquity and potential for influence, digital media can be a valuable or harmful tool for population-level behavior change. Thus, there is a great need for more research on the relationship between digital media and health behaviors, social norms, and social networks. Although research is being conducted to determine what digital media as an intervention tool would look like, how it works, and how effective it is [13,14], these studies have only scratched the surface [15].

The importance of digital media interventions is growing in many health behavior subject areas, including nicotine and tobacco use research. Mass media campaigns have been proven to be effective in creating positive changes in smoking-related attitudes, intentions, and behaviors [16]. More recent research also supports the use of media campaigns to address the aforementioned rapid increase in and continued use of e-cigarettes among youth and young adults [17]. Digital strategies will be central to future campaigns. A recent systematic review of digital behavior change interventions by Ichimiya et al [18] identified 298 relevant articles; 19 of those were for nicotine and tobacco interventions.

Prior Work

Digital media intervention research is currently a small, growing, and highly important field, given the shift in nicotine behavior change campaigns from traditional mass media such as television to digital platforms [19]. These strategies have the potential to change social norms (ie, beliefs among a population about what is widespread behavior and what is socially sanctioned or required) [20] about behaviors such as vaping. Social norms may be influenced by, for example, e-cigarette companies' social media platforms [21], which normalize and effectively promote use among a peer group such as young adults [19]. At the same time, antivaping social media may create a new social norm that vaping is uncommon and less socially accepted among the peer group. Theoretically, the effect of such social media campaigns may be to promote a social norm such as the avoidance of nicotine and tobacco products [20].

Specifically, given the relatively small number of studies found in the systematic review of digital tobacco behavior change interventions by Ichimiya et al [18], there is a need to rigorously test the effects of antivaping social media content on outcomes. Many large-scale campaigns, such as those run by the Food and Drug Administration (eg, Real Cost) and the Truth Initiative (the Truth campaign), are currently using digital content as part of their overall behavior change strategies, but little is known about their mechanisms of change, and the published research does not include randomized trials [18]. Research is needed to build and test theories of change for such campaigns using randomized experimental methods.

Furthermore, this study builds on recent studies using a social media–based data collection platform for random assignment studies [14]. The use of social media recruitment, chatbots for survey delivery, and retargeting technology for intervention delivery and follow-up (FU) have been proven feasible and produce short-term effects on content (eg, advertising) exposure. This study aims to test these methods in a randomized controlled dose-response experiment.

Study Aims

In this study, our goal was to determine whether a social media intervention delivered through an experimental design would have a positive effect on young adult vaping outcomes. We aimed to disseminate the intervention on participants' Facebook and Instagram news feeds in the form of an antivaping campaign consisting of 4 videos, each 15 seconds in length, drawn from previous Truth Initiative content and aimed at young adults aged 18 to 24 years. Participants answered 1 preintervention survey and 2 postintervention surveys on the same platform, Facebook Messenger (FM; Meta Platforms). A chatbot was used to execute the surveys and keep participants engaged over the course of the 60-day study period.

Hypothesis

We tested the hypotheses that exposure to antivaping social media content measured through a social media–based survey would reduce vaping use intentions at the 60-day FU (FU2; hypothesis 1) and increase antivaping industry beliefs at the FU2 (hypothesis 2). We also examined 1 research question: would treatment assignment be associated with lower use intentions and higher antivaping beliefs (ie, a dose-response effect)?

XSL•FO
RenderX

Methods

Study Design

The study design was a randomized controlled experiment with 4 treatment arms and a no-exposure control arm. Using the Virtual Lab platform, participants were recruited into the study (details under *Data Collection and Measures* section), delivered a baseline survey, and then randomized to 1 of the 5 conditions. The design was to achieve a specific number of impressions per arm as follows: 0 (control), 4, 8, 16, and 32 over a 15-day period between each survey wave. Impressions are defined as the number of views of a social media post by a study participant [22].

There were 3 survey waves: baseline, 30-day FU (FU1), and FU2. The aim was to collect sufficient participants within each wave to have sufficient power to detect a treatment effect of the intervention video content exposure on vaping intentions at FU2. The final baseline sample consisted of 1491 participants divided into 5 study conditions: FU1 consisted of 70.28% (1048/1491) of the participants, and FU2 consisted of 57.28% (854/1491) of the participants.

Intervention Content

The intervention content consisted of four 15-second videos drawn from a previous web-based Truth Initiative campaign called *Tested on Humans*. The main themes of the videos were that vaping companies do not know the health and other impacts of using e-cigarettes and that they are "testing" their products on human beings. This is consistent with an "anti-industry" countermarking approach to nicotine and tobacco campaigns, which has been used successfully in the past [23,24]. The campaign was not publicly active during this study. We chose this content because it was designed for social media distribution, focused on preventing vaping, and was not currently active.

Following baseline, videos were promoted in the live Facebook and Instagram feeds of treatment arm participants in a randomized order and combinations to achieve the targeted impressions for each arm (ie, an average number of impressions per condition). For example, the "low" exposure arm was designed to obtain 4 impressions that would receive a randomly ordered assignment of each video 1 time, the next highest exposure arm (8 impressions) was designed to obtain the videos in random order 2 times, and so on. The actual number of impressions per group varied because of the time of the intervention delivery and was measured at the group level because of the confidentiality restrictions Facebook and Instagram place on publicly available user data (ie, the exact number of impressions by an individual user is not available, only by the study condition). This resulted in the use of a 5-level variable corresponding to the 5 treatment arms of increasing intended impressions (arm 1=0 impressions, arm 2=4 impressions, arm 3=8 impressions, arm 4=16 impressions, and arm 5=32 impressions).

The study was implemented by Virtual Lab, a social media–based data collection and intervention content delivery platform [25]. Participants were recruited via Facebook and Instagram advertisements. When a potential participant clicked on a study advertisement, they were asked a series of screening questions using a FM chatbot. Eligible participants were US residents aged 18 to 24 years within the stratified subgroups, with 49.97% (911/1823) of them being current vapers. The participants were asked to provide informed consent and participate in the study through an FM survey delivered by the chatbot. After completing the baseline questionnaire, the participants were randomized to the study condition, received any relevant content over time, and were invited to complete the FU questionnaires.

We used the CONSORT (Consolidated Standards of Reporting Trials) checklist when writing our report [26]. The study design and recruitment procedures are summarized in Figure 1. Note that the total retained sample at FU was 57.28% (854/1491), but the sample sizes for some analyses varied because of participant response patterns.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. FU1: 30-day follow-up; FU2: 60-day follow-up.

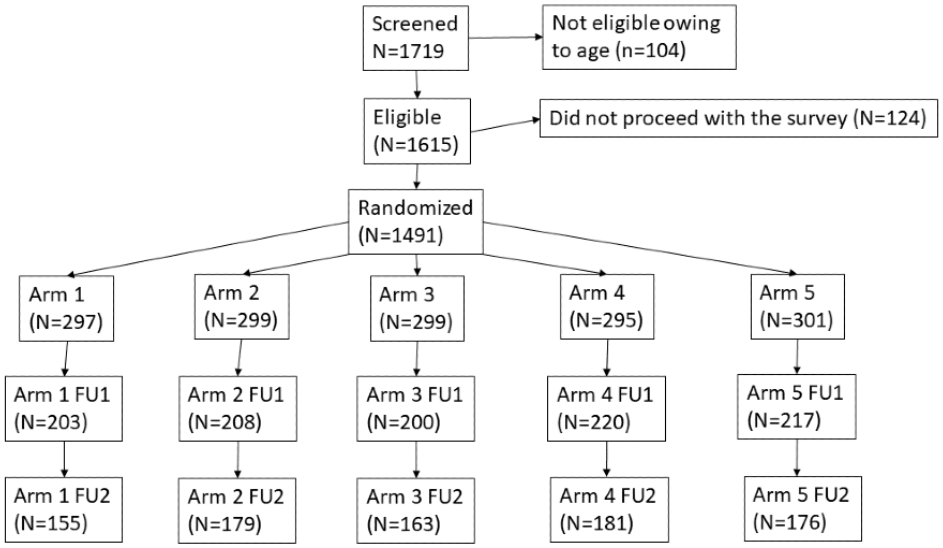

XSL•FO
RenderX

Ethical Considerations

This study was reviewed and approved as not greater than minimal risk human participants research by the George Washington University Institutional Review Board (IRB) on August 5, 2020(IRB number NCR202837). Through the FM chatbot, participants read an IRB-approved statement informing them about the purposes and nature of the research. By clicking on a button to proceed with the survey, they provided consent to participate. All data used in this study have been deidentified and stored following the IRB-approved procedure to ensure confidentiality. The participants received a US $10 e-gift card as compensation for each survey completed.

Data Collection and Measures

Similar to a pilot study reported by Tulsiani et al [14], we worked with Virtual Lab to implement the study and collect data [25]. The study team created a Facebook business account called "Digital Health Research" to recruit participants and manage data collection and a second account, "Consumer Consciousness," to run the target advertisements on the enrolled participants' Facebook and Instagram news feeds. The recruitment advertisements were served to people aged 18 to 24 years who were located in the United States. The advertisements used the text, "Take a 15 minute survey, get paid $10." After the participants clicked on the study's advertisement, they were sent a message via FM inviting them to participate in the study.

The survey was delivered as a series of individual chats through FM using a chatbot. The survey consisted of 40 items drawn from the tobacco control and campaign evaluation literature [27]. For this study, we used a subset of the items contained in the survey, following our study objectives. All items were measured on a 5-point agreement scale, except where noted (strongly agree to strongly disagree).

Our primary end point was future vape intentions, operationalized as the average of responses to 2 items, which were each answered on a 5-point agreement scale: "Thinking about the future, if one of your best friends offered you an e-cigarette/vape (even one or two puffs) in the coming year, would you smoke it?" and "Do you think you will use an e-cigarette/vape (even one or two puffs) in the next year?" Our secondary end point was anti-industry sentiment, measured as the average of responses to 2 items, also on a 5-point agreement scale: "Vape companies make me angry" and "I am willing to stand up with others against vape companies." Our measures of vaping intentions and anti-industry sentiment are both taken from the second FU survey. Finally, we examined self-reported advertisement exposure. For each of the 4 advertisements, the participants were asked, "Overall, about how many times to do you think you've seen this ad? 1-2 times; 3-5 times; more than 5 times." Responses were recoded to approximate the average value for each category ("Never"=0, "1-2 times"=1.5, "3-5 times"=4, and ">5 times"=6), and an average value across each of these 4 advertisements was calculated to generate an average value of reported advertisement exposure. As we were interested in cumulative exposure, the value for both time points was averaged.

Data Analysis

To investigate our hypotheses, a series of multivariate linear regressions were conducted in Stata 17 (StataCorp). To capture the dose-response effect, rather than treating the 5 study arms as 5 independent and nominal groups, we assigned each arm a numerical value corresponding to the number of advertisements delivered to participants in each arm and used this number as our focal independent variable. In each model, the predictor was the treatment arm to which each participant was assigned. The outcome variables for these regressions were self-reported advertisement exposure, anti-industry attitudes and beliefs, and vape use intentions. For each of the multivariate linear regressions, the following covariates were included: race and ethnicity (dummy coded for "non-Hispanic White," "non-Hispanic Black," "Hispanic," and "non-Hispanic other"—"non-Hispanic White" was used as the reference category), gender (dummy coded for "female," "male," and "another identity/nonbinary/transgender"—"female" was used as the reference category), age in years, and baseline use of e-cigarettes (dummy coded for "Never User," "Former User," and "Current User"—"Never User" was used as the reference category).

We included 57.28% (854/1491) of the participants retained at FU2 in our analysis. We compared the retained participants with those lost to FU (LTFU) and found no statistically significant differences across demographic variables. These results are presented in Multimedia Appendix 1.

In addition, different efficacy levels of the treatment were hypothesized for the participants based on their use status at baseline. Therefore, each of the aforementioned model was also applied to subsamples based on e-cigarette use at the final wave of data collection ("Never users," "Former Users," and "Current Users"). We note that these subgroup analyses were not among the original hypotheses of the study and were investigated post hoc. All analyses included the full sample, including those LTFU, following the intention-to-treat principles.

Power Analysis

We conducted a statistical power analysis to determine an appropriate sample size. Because of our planned dose-response analysis, we used correlation analysis as the basis for our calculations. Specifically, we assumed that the correlation between advertisements delivered in each study arm and vaping intentions at FU2 would be small, as low as 0.1, and used Stata 18 (StataCorp) to calculate that a sample of 783 participants would be needed to provide 80% power for rejecting the null hypothesis at the conventional Cronbach α=0.05 level. Second, based on the results of some pilot studies, we assumed that up to 45% (671/1491 based on our actual sample) of the baseline participants would be LTFU before FU2. Thus, we concluded that the baseline sample should include a minimum of 1423 participants, or approximately 19.11% (285/1423) of participants in each of the 5 study arms. We exceeded our recruitment target in the final sample.

XSL•FO
RenderX

Results

Table 1 provides the descriptive statistics for the baseline sample categorized by the treatment arm.

Table 1. Descriptive statistics by treatment arm (n=854).

	Arm 1 (n=155)	Arm 2 (n=179)	Arm 3 (n=163)	Arm 4 (n=181)	Arm 5 (n=176)	Chi-square (*df*)	*P* value
Age (y), n (%)						19.8 (24)	.71
18	22 (14.2)	20 (11.2)	20 (12.3)	27 (14.9)	20 (11.4)		
19	21 (13.6)	24 (13.4)	25 (15.3)	26 (14.4)	19 (10.8)		
20	18 (11.6)	24 (13.4)	22 (13.5)	27 (14.9)	28 (15.9)		
21	26 (16.8)	34 (19)	17 (10.4)	23 (12.7)	22 (12.5)		
22	24 (15.5)	26 (14.5)	22 (13.5)	22 (12.2)	32 (18.2)		
23	16 (10.3)	28 (15.6)	32 (19.6)	27 (14.9)	23 (13.1)		
24	28 (18.1)	23 (12.9)	25 (15.3)	29 (16)	32 (18.2)		
Race and ethnicity, n (%)						11.48 (12)	.49
Black, non-Hispanic	8 (5.2)	12 (6.7)	10 (6.1)	18 (9.9)	9 (5.1)		
Hispanic	36 (23.2)	28 (15.6)	27 (16.6)	35 (19.3)	29 (16.5)		
Other, non-Hispanic	33 (21.3)	52 (29.1)	49 (30.1)	43 (23.8)	50 (28.4)		
White, non-Hispanic	78 (50.3)	87 (48.6)	77 (47.2)	85 (47)	88 (50)		
Sex, n (%)						4.5 (8)	.81
Female	109 (70.3)	125 (69.8)	117 (71.8)	127 (70.2)	121 (68.8)		
Male	45 (29)	47 (26.3)	41 (25.2)	48 (26.5)	48 (27.3)		
Another identity, nonbinary, or transgender	1 (0.7)	7 (3.9)	5 (3.1)	6 (3.3)	7 (4)		
Perceived financial situation, n (%)						11.7	.47
Lives comfortably	51 (32.9)	63 (35.2)	56 (34.4)	55 (30.4)	58 (33)		
Meets needs with a little left over	42 (27.1)	64 (35.8)	51 (31.3)	69 (38.1)	54 (30.7)		
Meets basic expenses	54 (34.8)	39 (21.8)	47 (28.8)	45 (24.9)	54 (30.7)		
Does not meet basic expenses	8 (5.2)	13 (7.3)	9 (5.5)	12 (6.6)	10 (5.7)		
e-Cigarette use at baseline, n (%)						7.8 (12)	.80
Never users	81 (52.3)	108 (60.3)	91 (55.8)	105 (58)	102 (58)		
Former users	40 (25.8)	35 (19.6)	42 (25.8)	41 (22.7)	33 (18.8)		
Current users	33 (21.3)	35 (19.6)	30 (18.4)	34 (18.8)	41 (23.3)		
Missing	1 (0.7)	1 (0.6)	0 (0)	1 (0.6)	0 (0)		
Sexual orientation, n (%)						26.8 (20)	.14
Heterosexual	96 (61.9)	105 (58.7)	106 (65)	124 (68.5)	105 (59.1)		
Bisexual	25 (16.1)	38 (21.2)	25 (15.3)	18 (9.9)	34 (19.3)		
Homosexual	11 (7.1)	12 (6.7)	6 (3.7)	10 (5.5)	12 (6.8)		
Asexual	3 (1.9)	3 (1.7)	8 (4.9)	3 (1.7)	1 (0.6)		
Another sexual orientation	7 (4.5)	6 (3.4)	1 (0.6)	5 (2.8)	6 (3.4)		
Missing	13 (8.4)	15 (8.4)	17 (10.4)	21 (11.6)	19 (10.8)		

We used a chi-square test to examine any potential differences in demographics between the arms and found no statistically significant differences. Overall, the sample is relatively evenly distributed, with ages ranging from 18 to 24 years, and just <47.95% (715/1491) of the sample is non-Hispanic White. Approximately 7.98% (119/1491) and 16.97% (253/1491) of the sample were non-Hispanic Black and of Hispanic ethnicity, respectively. Approximately 69.01% (1029/1491) of the sample was female, and approximately 65% (969/1491) reported having more than enough income to support themselves. Just >70.09% (1045/1491) reported being heterosexual, with the next largest group reporting being bisexual at just <17.91% (267/1491). At

XSL•FO
RenderX

baseline, approximately 20.99% (313/1491) reported current e-cigarette use (meaning within the past 30 d), and 22% (328/1491) reported former (>30 d ago) use.

Table 2 provides a summary of the models for treatment effects with covariates on vape use intentions by final use status (full sample, never, former, and current users).

Overall, there is a significant treatment effect among current vapers on lower vaping intentions (β=−0.138; *P*=.04) but not in the full sample or other subgroups. We also see effects on lower vaping intentions among Black participants in the full sample and among baseline former and current vapers in the full sample.

Table 3 provides a summary of the models for treatment effects with covariates on anti-industry attitudes and beliefs, which were the main beliefs targeted by the intervention content (ie, beliefs that the e-cigarette industry harms its customers).

Table 2. Treatment effects on vape use intentions (N=836)[a].

	Full sample		Stratified analyses					
	Analysis (n=836)		Never users (n=478)		Former users (n=197)		Current users (n=157)	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Treatment	−0.009 (−0.059 to 0.040)	.71	0.024 (−0.026 to 0.074)	.35	0.059 (−0.051 to 0.169)	.29	−0.138 (−0.266 to −0.010)	.04
Race and ethnicity								
Black, non-Hispanic	0.321 (0.031 to 0.611)	.03	0.072 (−0.221 to 0.364)	.63	0.236 (−0.603 to 1.07)	.58	0.265 (−0.410 to 0.940)	.44
Hispanic	0.009 (−0.180 to 0.198)	.93	0.090 (−0.099 to 0.280)	.35	−0.272 (−0.711 to 0.167)	.22	0.081 (−0.407 to 0.568)	.75
Other, non-Hispanic	0.078 (−0.092 to 0.249)	.37	0.042 (−0.124 to 0.209)	.62	0.158 (−0.225 to 0.541)	.42	−0.198 (−0.690 to 0.294)	.43
White, non-Hispanic	REF[b]	REF	REF	REF	REF	REF	REF	REF
Sex								
Female	REF	REF	REF	REF	REF	REF	REF	REF
Male	0.099 (−0.058 to 0.256)	.22	0.072 (−0.085 to 0.228)	.37	0.096 (−0.250 to 0.442)	.59	−0.065 (−0.511 to 0.381)	.78
Another identity, nonbinary, or transgender	0.264 (−0.144 to 0.672)	.20	0.216 (−0.230 to 0.663)	.34	0.321 (−0.414 to 1.06)	.39	0.401 (−0.951 to 1.75)	.56
Age (y)	0.002 (−0.033 to 0.037)	.90	−0.023 (−0.058 to 0.012)	.20	−0.010 (−0.089 to 0.069)	.80	0.073 (−0.020 to 0.165)	.12
Baseline use status								
Never	REF	REF	N/A[c]	N/A	N/A	N/A	N/A	N/A
Former	0.737 (0.564 to 0.910)	<.001	N/A	N/A	N/A	N/A	N/A	N/A
Current	1.70 (1.52 to 1.88)	<.001	N/A	N/A	N/A	N/A	N/A	N/A

[a]For the full-sample analyses, N was 836 because of item nonresponse, and N was 832 for the stratified analyses.

[b]REF: reference.

[c]N/A: not applicable.

XSL•FO
RenderX

Table 3. Treatment effects on anti-industry attitudes and beliefs (N=838)[a].

	Full sample		Stratified analyses					
	Analysis (n=838)		Never users (n=479)		Former users (n=198)		Current users (n=157)	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Treatment	0.020 (−0.027 to 0.067)	.41	0.007 (−0.052 to 0.067)	.81	−0.065 (−0.164 to 0.033)	.19	0.122 (0.008 to 0.237)	.04
Race and ethnicity								
Black, non-Hispanic	−0.409 (−0.681 to −0.137)	.003	−0.203 (−0.553 to 0.147)	.26	−0.988 (−1.673 to −0.303)	.005	−0.186 (−0.796 to 0.423)	.55
Hispanic	0.025 (−0.153 to 0.204)	.78	−0.064 (−0.290 to 0.162)	.58	0.120 (−0.273 to 0.512)	.55	0.164 (−0.269 to 0.598)	.46
Other, non-Hispanic	−0.245 (−0.406 to −0.084)	.003	−0.246 (−0.444 to −0.047)	.02	−0.385 (−0.725 to −0.045)	.03	0.139 (−0.298 to 0.577)	.53
White, non-Hispanic	REF[b]	REF	REF	REF	REF	REF	REF	REF
Sex								
Female	REF	REF	REF	REF	REF	REF	REF	REF
Male	−0.140 (−0.288 to 0.009)	.07	−0.107 (−0.294 to 0.080)	.26	−0.016 (−0.325 to 0.292)	.92	−0.260 (−0.658 to 0.137)	.20
Another identity, nonbinary, or transgender	0.003 (−0.383 to 0.022)	.99	−0.080 (−0.613 to 0.454)	.77	0.345 (−0.309 to 0.999)	.30	−0.591 (−1.81 to 0.630)	.34
Age (y)	−0.011 (−0.044 to 0.022)	.51	−0.009 (−0.051 to 0.032)	.67	−0.031 (−0.100 to 0.039)	.39	0.027 (−0.056 to 0.110)	.52
Baseline use status								
Never	REF	REF	REF	REF	REF	REF	REF	REF
Former	−0.357 (−0.520 to −0.193)	<.001	N/A[c]	N/A	N/A	N/A	N/A	N/A
Current	−0.622 (−0.792 to −0.451)	<.001	N/A	N/A	N/A	N/A	N/A	N/A

[a]For the full-sample analyses, N was 838 because of item nonresponse, and N was 834 for the stratified analyses.

[b]REF: reference.

[c]N/A: not applicable.

Overall, there is a significant treatment effect among current vapers on anti-industry beliefs (β=0.120; *P*=.046) but not in the full sample or other subgroups. In addition, Black participants were more likely to report an intention to vape than non-Hispanic White participants, and current and former vapers were more likely to report an intention to vape than participants who had never vaped.

Table 4 provides a summary of the models for treatment effects with covariates on self-reported advertisement exposure.

There is no main effect of treatment on the reported advertisement exposure among any of the population groups of interest. In addition, we observe higher self-reported advertisement exposure among Black participants in the full sample, and exposure to intervention content was higher among baseline never vapers and baseline current 30-day vapers. It is possible that the lack of relationship between the treatment group and reported advertisement exposure is because of insufficient elapsed time during the intervention period to achieve the intended number of impressions per group, which resulted in participants in the fourth and fifth groups receiving similar levels of impressions. The average impressions delivered per user per treatment arm were as follows: arm 1=0, arm 2=2.293, arm 3=7.708, arm 4=12.718, and arm 5=14.218.

XSL•FO
RenderX

Table 4. Treatment effects on self-reported advertisement exposure (N=730)[a].

	Full sample		Stratified analyses					
	Analysis (n=730)		Never users (n=407)		Former users (n=180)		Current users (n=139)	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Treatment	−0.004 (−0.042 to 0.034)	.83	−0.020 (−0.069 to 0.030)	.44	−0.006 (−0.068 to 0.055)	.85	0.066 (−0.035 to 0.167)	.20
Race and ethnicity								
Black, non-Hispanic	0.312 (0.087 to 0.536)	.007	0.681 (0.392 to 0.969)	<.001	0.143 (−0.263 to 0.549)	.49	−0.430 (−1.04 to 0.183)	.17
Hispanic	−0.071 (−0.216 to 0.075)	.34	−0.160 (−0.350 to 0.030)	.10	0.008 (−0.236 to 0.254)	.94	−0.024 (−0.349 to 0.397)	.90
Other, non-Hispanic	−0.187 (−0.317 to −0.057)	.005	−0.260 (−0.426 to −0.094)	.002	−0.155 (−0.364 to 0.055)	.15	−0.126 (−0.501 to 0.249)	.51
White, non-Hispanic	REF[b]	REF	REF	REF	REF	REF	REF	REF
Sex								
Female	REF	REF	REF	REF	REF	REF	REF	REF
Male	0.088 (−0.030 to 0.207)	.14	−0.036 (−0.189 to 0.117)	.64	0.106 (−0.084 to 0.296)	.27	0.378 (0.044 to 0.712)	.03
Another identity, nonbinary, or transgender	−0.064 (−0.376 to 0.028)	.69	0.167 (−0.308 to 0.643)	.49	−0.124 (−0.512 to 0.265)	.53	−0.257 (−1.27 to 0.753)	.62
Age (y)	0.001 (−0.026 to 0.028)	.93	−0.007 (−0.042 to 0.027)	.67	−0.012 (−0.055 to 0.031)	.58	0.055 (−0.017 to 0.128)	.13
Baseline use status								
Never	REF	REF	N/A[c]	N/A	N/A	N/A	N/A	N/A
Former	−0.080 (−0.211 to 0.052)	.24	N/A	N/A	N/A	N/A	N/A	N/A
Current	0.124 (−0.014 to 0.262)	.08	N/A	N/A	N/A	N/A	N/A	N/A

[a]For the full-sample analyses, N was 730 because of item nonresponse, and for the stratified analyses, N was 726.

[b]REF: reference.

[c]N/A: not applicable.

Discussion

Overview

e-Cigarette use among young adults is a significant public health threat. Use rates dropped early in the COVID-19 pandemic but have seen a resurgence in the later stages of this public health emergency [28]. Innovative strategies to deliver antivaping messages and reduce use intentions and behavior are needed. Given the high levels of social media use among adolescents and young adults, and especially engagement with provaping content [29], interventions using social media are an important intervention channel for experimentation and population-level campaigns.

Principal Findings

Overall, this study found significant effects in the direction expected for intentions and anti-industry sentiment. Our hypothesis 1 was partially confirmed: there was a significant treatment effect on both anti-industry beliefs and lower vaping intentions, but these effects were limited to current e-cigarette users and were not observed among never or past users or in the full-sample analyses.

This study partially confirmed hypothesis 2: higher levels of treatment (ie, from arms 1 to 5) were associated with the anti-industry beliefs and vaping intentions outcomes of interest. However, we did not observe a direct or dose-response effect of the intervention on content exposure outcomes (ie, awareness of the specific social media posts used as stimulus in the study). This is typically the most proximal outcome resulting from a campaign, and the absence of these anticipated effects deserves further investigation.

One possible explanation is that the experiment did not completely achieve the intended levels of impressions for each study arm. In particular, the level of impression achieved at the highest exposure arms (4 to 5) was quite similar, whereas the intent was to double the number of impressions in each arm. This may be an artifact of the length of time our intervention was in the field, which was only 60 days. Total social media impressions are typically a function of the length of a campaign,

XSL•FO
RenderX

and longer study durations may result in more closely matching the intended exposure levels by study condition [30,31].

The observed effects of treatment on anti-industry beliefs are consistent with the content of the intervention, which focused on messages exposing the misinformation and disinformation that e-cigarette companies use in their marketing and the idea that their products and practices are harmful to consumers. This study provides evidence to support the idea that targeted campaign messages can directly impact attitudes and beliefs focused on the content of those messages. Future studies should examine approaches to optimize these observed effects.

The observed effects on lower vape use intentions suggest that there may be a connection between anti-industry beliefs and future use among current users. If young adults believe that the industry is using misinformation and disinformation and selling a harmful product, they may reconsider their current use [32]. This suggests a potential mediation effect of anti-industry beliefs on intentions and potentially on e-cigarette use. The hypothetical pathways of effects should be formally evaluated in future studies.

Future Directions

One question raised by this study's findings for vaping intentions and anti-industry beliefs is why we did not see effects among the former and never vapers. However, it is simply difficult to shift beliefs and intentions in those other groups than among current vapers. Alternatively, messages focused on topics and persuasive content other than anti-industry sentiment may be needed for those groups. In addition, selective attention bias (ie, the personal relevance of vaping-related content) suggests that vapers may be more responsive to the antivaping social media content used in this study [33]. Future research should examine these questions.

To fully examine the dose-response effects of social media interventions, longer time durations may be needed, and larger small sample sizes per study arm may be needed, especially given the attrition at FU. Cell sizes between the study conditions were reduced at the second FU, which may have reduced the statistical power to below the levels needed to detect some dose-response outcomes of interest. Previous studies have shown the dose-response effects of anti-industry messaging on vaping-related content exposure and attitudinal outcomes [14].

This study also contributes to the growing literature on public health, social media interventions, and theories of change [34,35]. This study further demonstrates the potential of a social media–based research and intervention delivery platform to build evidence for tobacco control. Future studies should expand on this research with longer-term longitudinal studies capable of potentially detecting treatment effects on vaping use behavior; examine diverse subgroups of interest, including high-risk groups for e-cigarette use; and examine multiple types of social media content. Finally, the demonstrated effects of social media on intentions and other outcomes related to vaping should be considered when formulating a tobacco control policy, including recommendations for effective comprehensive prevention and cessation interventions [36].

Limitations

Finally, this study has some limitations. First, it was conducted over a relatively short period (60 d), and thus, only intermediate outcomes were evaluated. In addition, the time duration may have limited our ability to fully generate the intended differences in the objective impressions created within each study arm. We observed a substantial LTFU (>40%) at the second FU survey. Future social media studies should make extensive efforts to limit LTFU, especially when following participants over longer periods. Second, our original data analysis plan did not include an explicit plan to stratify by vaping status, and relatedly, the size of our sample was based on power calculations that assumed a whole sample rather than a stratified analysis. The post hoc nature of those stratified analyses should be considered when interpreting the findings of this study. In addition, the Truth Initiative content used for the stimulus had previously aired, and prior exposure may have limited its treatment potential. Finally, although the observed effects of the intervention occurred among current vapers, the study was not powered by subgroups. Finally, we performed multiple comparisons in our analyses, which raises the possibility of false-positive findings. Future studies should use previously unaired content, where possible, and ensure a sufficient sample size among specific subgroups of interest, where feasible (ie, power at the subgroup level).

Conclusions

Social media interventions are a promising approach to preventing vaping among young adults [13]. More research is needed on how to optimize the dosage of such interventions and the long-term effects on vaping use over time. Social media–based research platforms are a promising methodology to conduct experimental public health research among specific priority populations.

Acknowledgments

This research was sponsored by the National Cancer Institute (grant CA253013).

Data Availability

A data sharing agreement is required for the use of all data. Our research team does not share data with tobacco industry representatives or affiliated researchers. Investigators seeking access to the data used in the study should make a written request to the corresponding author and submit a detailed research plan including the purpose of the proposed research, required variables, duration of the analysis phase, institutional review board approval, Federalwide Assurance information, and documentation of investigator training in human participants.

XSL•FO
RenderX

Authors' Contributions

WDE conceptualized the study, acquired the funding, led the investigation and methodology, administered and supervised the project, designed data visualization, and wrote and edited the paper. JB led data curation, formal analysis, contributed to methodology, and edited the paper. APD conducted formal analysis, wrote and edited the paper. ECH contributed to investigation, methodology, and edited the paper. JC contributed to methodology and editing. MM, ST, JK, and MI contributed to editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of recruited and retained samples.
[PDF File (Adobe PDF File), 86 KB - jmir_v26i1e50741_app1.pdf]

Multimedia Appendix 2
CONSORT eHealth Checklist (V 1.6.1).
[PDF File (Adobe PDF File), 4277 KB - jmir_v26i1e50741_app2.pdf]

References

1. Surgeon general's advisory on e-cigarette use among youth. Centers for Disease Control and Prevention. 2018 Dec. URL: https://www.cdc.gov/tobacco/basic_information/e-cigarettes/surgeon-general-advisory/index.html [accessed 2023-04-20]
2. Cornelius ME, Wang TW, Jamal A, Loretan CG, Neff LJ. Tobacco product use among adults - United States, 2019. MMWR Morb Mortal Wkly Rep 2020 Nov 20;69(46):1736-1742 [FREE Full text] [doi: 10.15585/mmwr.mm6946a4] [Medline: 33211681]
3. 2016 surgeon general's report: e-cigarette use among youth and young adults. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/tobacco/sgr/e-cigarettes/index.htm [accessed 2024-02-15]
4. Cooper M, Park-Lee E, Ren C, Cornelius M, Jamal A, Cullen KA. Notes from the field: e-cigarette use among middle and high school students - United States, 2022. MMWR Morb Mortal Wkly Rep 2022 Oct 07;71(40):1283-1285 [FREE Full text] [doi: 10.15585/mmwr.mm7140a3] [Medline: 36201370]
5. Wang TW, Gentzke AS, Neff LJ, Glidden EV, Jamal A, Park-Lee E, et al. Characteristics of e-cigarette use behaviors among US youth, 2020. JAMA Netw Open 2021 Jun 01;4(6):e2111336 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11336] [Medline: 34097049]
6. Williams R. The rise of disposable JUUL-type e-cigarette devices. Tob Control 2020 Dec;29(e1):e134-e135 [FREE Full text] [doi: 10.1136/tobaccocontrol-2019-055379] [Medline: 31806858]
7. Alqahtani MM, Pavela G, Lein DHJ, Vilcassim R, Hendricks PS. The influence of mental health and respiratory symptoms on the association between chronic lung disease and e-cigarette use in adults in the United States. Respir Care 2022 Jul;67(7):814-822 [FREE Full text] [doi: 10.4187/respcare.09579] [Medline: 35440495]
8. Riehm KE, Young AS, Feder KA, Krawczyk N, Tormohlen KN, Pacek LR, et al. Mental health problems and initiation of e-cigarette and combustible cigarette use. Pediatrics 2019 Jul;144(1):e20182935. [doi: 10.1542/peds.2018-2935] [Medline: 31160343]
9. Becker TD, Arnold MK, Ro V, Martin L, Rice TR. Systematic review of electronic cigarette use (vaping) and mental health comorbidity among adolescents and young adults. Nicotine Tob Res 2021 Feb 16;23(3):415-425. [doi: 10.1093/ntr/ntaa171] [Medline: 32905589]
10. Lechner WV, Janssen T, Kahler CW, Audrain-McGovern J, Leventhal AM. Bi-directional associations of electronic and combustible cigarette use onset patterns with depressive symptoms in adolescents. Prev Med 2017 Mar;96:73-78 [FREE Full text] [doi: 10.1016/j.ypmed.2016.12.034] [Medline: 28024859]
11. Obisesan OH, Mirbolouk M, Osei AD, Orimoloye OA, Uddin SM, Dzaye O, et al. Association between e-cigarette use and depression in the behavioral risk factor surveillance system, 2016-2017. JAMA Netw Open 2019 Dec 02;2(12):e1916800 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.16800] [Medline: 31800073]
12. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 7. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2023-04-20]
13. Evans WD, Abroms LC, Broniatowski D, Napolitano MA, Arnold J, Ichimiya M, et al. Digital media for behavior change: review of an emerging field of study. Int J Environ Res Public Health 2022 Jul 26;19(15):9129 [FREE Full text] [doi: 10.3390/ijerph19159129] [Medline: 35897494]
14. Tulsiani S, Ichimiya M, Gerard R, Mills S, Bingenheimer JB, Hair EC, et al. Assessing the feasibility of studying awareness of a digital health campaign on Facebook: pilot study comparing young adult subsamples. JMIR Form Res 2022 Aug 29;6(8):e37856 [FREE Full text] [doi: 10.2196/37856] [Medline: 36036974]

XSL•FO
RenderX

15. Hair EC, Kreslake JM, Rath JM, Pitzer L, Bennett M, Vallone D. Early evidence of the associations between an anti-e-cigarette mass media campaign and e-cigarette knowledge and attitudes: results from a cross-sectional study of youth and young adults. Tob Control 2023 Mar 21;32(2):179-187. [doi: 10.1136/tobaccocontrol-2020-056047] [Medline: 34290134]
16. Vallone D, Cantrell J, Bennett M, Smith A, Rath JM, Xiao H, et al. Evidence of the impact of the truth FinishIt campaign. Nicotine Tob Res 2018 Apr 02;20(5):543-551. [doi: 10.1093/ntr/ntx119] [Medline: 28575421]
17. MacMonegle AJ, Smith AA, Duke J, Bennett M, Siegel-Reamer LR, Pitzer L, et al. Effects of a national campaign on youth beliefs and perceptions about electronic cigarettes and smoking. Prev Chronic Dis 2022 Apr 07;19:E16 [FREE Full text] [doi: 10.5888/pcd19.210332] [Medline: 35389831]
18. Ichimiya M, Gerard R, Mills S, Brodsky A, Cantrell J, Evans WD. The measurement of dose and response for smoking behavior change interventions in the digital age: systematic review. J Med Internet Res 2022 Aug 25;24(8):e38470 [FREE Full text] [doi: 10.2196/38470] [Medline: 36006682]
19. Noar SM, Rohde JA, Prentice-Dunn H, Kresovich A, Hall MG, Brewer NT. Evaluating the actual and perceived effectiveness of E-cigarette prevention advertisements among adolescents. Addict Behav 2020 Oct;109:106473 [FREE Full text] [doi: 10.1016/j.addbeh.2020.106473] [Medline: 32521287]
20. Cislaghi B, Heise L. Theory and practice of social norms interventions: eight common pitfalls. Global Health 2018 Aug 17;14(1):83 [FREE Full text] [doi: 10.1186/s12992-018-0398-x] [Medline: 30119638]
21. Czaplicki L, Kostygina G, Kim Y, Perks SN, Szczypka G, Emery SL, et al. Characterising JUUL-related posts on Instagram. Tob Control 2020 Nov 02;29(6):612-617. [doi: 10.1136/tobaccocontrol-2018-054824] [Medline: 31266903]
22. Heyman JL, Kerr LG, Human LJ. Social media use predicts greater liking in in-person initial interactions. Pers Soc Psychol Bull 2022 Sep 09;48(9):1393-1405 [FREE Full text] [doi: 10.1177/01461672211040964] [Medline: 34496679]
23. Duke JC, Vallone DM, Allen JA, Cullen J, Mowery PD, Xiao H, et al. Increasing youths' exposure to a tobacco prevention media campaign in rural and low-population-density communities. Am J Public Health 2009 Dec;99(12):2210-2216. [doi: 10.2105/ajph.2008.155127]
24. Farrelly MC, Davis KC, Haviland ML, Messeri P, Healton CG. Evidence of a dose-response relationship between "truth" antismoking ads and youth smoking prevalence. Am J Public Health 2005 Mar;95(3):425-431. [doi: 10.2105/AJPH.2004.049692] [Medline: 15727971]
25. Rao NM, Donati D, Orozco Olvera VH. Conducting surveys and interventions entirely online : a virtual lab practitioner's manual (English. World Bank Group. 2021. URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/819191635501106862/conducting-surveys-and-interventions-entirely-online-a-virtual-lab-practitioner-s-manual [accessed 2024-02-13]
26. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Trials 2010 Mar 24;11(1):32 [FREE Full text] [doi: 10.1186/1745-6215-11-32] [Medline: 20334632]
27. Chan L, O'Hara B, Phongsavan P, Bauman A, Freeman B. Review of evaluation metrics used in digital and traditional tobacco control campaigns. J Med Internet Res 2020 Aug 11;22(8):e17432 [FREE Full text] [doi: 10.2196/17432] [Medline: 32348272]
28. Kreslake JM, O'Connor KM, Liu M, Vallone DM, Hair E. A resurgence of e-cigarette use among adolescents and young adults late in the COVID-19 pandemic. PLoS One 2023 Mar 29;18(3):e0282894 [FREE Full text] [doi: 10.1371/journal.pone.0282894] [Medline: 36989261]
29. Alpert JM, Chen H, Riddell H, Chung YJ, Mu YA. Vaping and Instagram: a content analysis of e-cigarette posts using the content appealing to youth (CAY) index. Subst Use Misuse 2021;56(6):879-887. [doi: 10.1080/10826084.2021.1899233] [Medline: 33749515]
30. Liu R, Greene KT, Liu R, Mandic M, Valentino BA, Vosoughi S, et al. Using impression data to improve models of online social influence. Sci Rep 2021 Aug 16;11(1):16613 [FREE Full text] [doi: 10.1038/s41598-021-96021-3] [Medline: 34400698]
31. Trunfio M, Rossi S. Conceptualising and measuring social media engagement: a systematic literature review. Ital J Mark 2021 Aug 11;2021(3):267-292. [doi: 10.1007/s43039-021-00035-8]
32. Tan AS, Bigman CA. Misinformation about commercial tobacco products on social media—implications and research opportunities for reducing tobacco-related health disparities. Am J Public Health 2020 Oct;110(S3):S281-S283. [doi: 10.2105/ajph.2020.305910]
33. Hwang Y. Selective exposure and selective perception of anti-tobacco campaign messages: the impacts of campaign exposure on selective perception. Health Commun 2010 Mar 26;25(2):182-190. [doi: 10.1080/10410230903474027] [Medline: 20390684]
34. Stellefson M, Paige SR, Chaney BH, Chaney JD. Social media and health promotion. Int J Environ Res Public Health 2020 May 11;17(9):3323 [FREE Full text] [doi: 10.3390/ijerph17093323] [Medline: 32403215]
35. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
36. Farrelly MC, Chaloupka FJ, Berg CJ, Emery SL, Henriksen L, Ling P, et al. Taking stock of tobacco control program and policy science and impact in the United States. J Addict Behav Ther 2017;1(2):8 [FREE Full text] [Medline: 30198028]

XSL•FO
RenderX

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FM: Facebook Messenger
FU: follow-up
IRB: Institutional Review Board
LTFU: lost to follow-up

Edited by G Eysenbach, T Leung; submitted 11.07.23; peer-reviewed by K Feder, J Alpert; comments to author 01.09.23; revised version received 12.09.23; accepted 08.02.24; published 12.03.24.

Please cite as:
Evans WD, Bingenheimer J, Cantrell J, Kreslake J, Tulsiani S, Ichimiya M, D'Esterre AP, Gerard R, Martin M, Hair EC
Effects of a Social Media Intervention on Vaping Intentions: Randomized Dose-Response Experiment
J Med Internet Res 2024;26:e50741
URL: https://www.jmir.org/2024/1/e50741
doi:10.2196/50741
PMID:38470468

©William Douglas Evans, Jeffrey Bingenheimer, Jennifer Cantrell, Jennifer Kreslake, Shreya Tulsiani, Megumi Ichimiya, Alexander P D'Esterre, Raquel Gerard, Madeline Martin, Elizabeth C Hair. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Long-Term Effectiveness of a Multi-Strategy Choice Architecture Intervention in Increasing Healthy Food Choices of High-School Students From Online Canteens (Click & Crunch High Schools): Cluster Randomized Controlled Trial

Tessa Delaney[1,2,3], PhD; Jacklyn Jackson[1,2,3], PhD; Christophe Lecathelinais[1,2,3], DESS; Tara Clinton-McHarg[4], PhD; Hannah Lamont[1,2,3], BNutrDiet (Hons); Sze Lin Yoong[5], PhD; Luke Wolfenden[1,2,3], PhD; Rachel Sutherland[1,2,3], PhD; Rebecca Wyse[1,2,3], PhD

[1]School of Medicine and Public Health, University of Newcastle, Wallsend, Australia

[2]Hunter New England Population Health, Hunter New England Local Health District, Wallsend, Australia

[3]Hunter Medical Research Institute, New Lambton Heights, Australia

[4]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia

[5]Faculty of Health, School of Health and Social Development, Global Centre for Preventive Health and Nutrition, Institute for Health Transformation, Deakin University, Melbourne, Australia

Corresponding Author:
Tessa Delaney, PhD
School of Medicine and Public Health
University of Newcastle
Locked Bag 10
Wallsend, 2287
Australia
Phone: 61 294617441
Email: Tessa.Delaney@health.nsw.gov.au

Abstract

Background: School canteens are a recommended setting to influence adolescent nutrition due to their scope to improve student food choices. Online lunch ordering systems ("online canteens") are increasingly used and represent attractive infrastructure to implement choice architecture interventions that nudge users toward healthier food choices. A recent cluster randomized controlled trial demonstrated the short-term effectiveness (2-month follow-up) of a choice architecture intervention to increase the healthiness of foods purchased by high school students from online canteens. However, there is little evidence regarding the long-term effectiveness of choice architecture interventions targeting adolescent food purchases, particularly those delivered online.

Objective: This study aimed to determine the long-term effectiveness of a multi-strategy choice architecture intervention embedded within online canteen infrastructure in high schools at a 15-month follow-up.

Methods: A cluster randomized controlled trial was undertaken with 1331 students (from 9 high schools) in New South Wales, Australia. Schools were randomized to receive the automated choice architecture intervention (including menu labeling, positioning, feedback, and prompting strategies) or the control (standard online ordering). The foods purchased were classified according to the New South Wales Healthy Canteen strategy as either "everyday," "occasional," or "should not be sold." Primary outcomes were the average proportion of "everyday," "occasional," and "should not be sold" items purchased per student. Secondary outcomes were the mean energy, saturated fat, sugar, and sodium content of purchases. Outcomes were assessed using routine data collected by the online canteen.

Results: From baseline to 15-month follow-up, on average, students in the intervention group ordered significantly more "everyday" items (+11.5%, 95% CI 7.3% to 15.6%; $P<.001$), and significantly fewer "occasional" (–5.4%, 95% CI –9.4% to –1.5%; $P=.007$) and "should not be sold" items (–6%, 95% CI –9.1% to –2.9%; $P<.001$), relative to controls. There were no between-group differences over time in the mean energy, saturated fat, sugar, or sodium content of lunch orders.

Conclusions: Given their longer-term effectiveness, choice architecture interventions delivered via online canteens may represent a promising option for policy makers to support healthy eating among high school students.

XSL•FO
RenderX

Trial Registration: Australian Clinical Trials ACTRN12620001338954, https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380546 ; Open Science Framework osf.io/h8zfr, https://osf.io/h8zfr/

(J Med Internet Res 2024;26:e51108) doi:10.2196/51108

KEYWORDS

randomized controlled trial; web-based ordering systems; ordering; menu; menus; prompt; prompts; prompting; nudge; nudges; behavior change; behaviour change; digital intervention; lunch; school; menu labelling; behavioral economics; secondary school; meal delivery apps; public health nutrition; meal; meals; nutrition; nutritional; diet; eating; food; schools; student; students; RCT; randomized; controlled trial; controlled trials; purchase; purchasing; canteen; canteens; choice; choices; architecture

Introduction

Background

Adolescents internationally are prone to having poor quality diets [1-3], which are associated with a higher risk of obesity, poor mental health and well-being, and an increased risk of chronic diseases during adulthood [4]. In particular, data from the most recent national survey of Australian high school children (aged 12-17 years) found that on average 5.7 (SE 0.2) to 6.6 (SE 0.7) serves of discretionary food choices are consumed per day, contributing to 38%-41% of total daily energy intake [5]. Adolescence represents a transitional life stage, which often coincides with increased autonomy regarding food choices and eating behaviors. Healthy eating interventions that can reach the adolescent population during this key stage are required [6], as dietary behaviors during adolescence have been shown to track throughout the life span [7].

High schools are an ideal setting to deliver interventions to improve adolescent nutrition, as they offer ongoing and widespread access to this traditionally hard-to-reach population [8]. Students have also been shown to consume up to 40% of their daily food intake during school hours, and in Australia over 60% of high school students purchase food at least once per week from their school canteen. However, the foods most commonly purchased from this setting are “less healthy,” discretionary foods high in energy, fat, salt, and sugar [9].

Interventions that incorporate choice architecture strategies (eg, provision of information, changing default options, and using incentives) [10] are effective in improving adolescent diet-related outcomes. A recent systematic review found that out of 137 included choice architecture interventions that aimed to modify child or adolescent diet-related outcomes, 74% were effective [10]. Despite this, of the 137 studies, only 9 were conducted in high schools and while 6 of the 9 studies (67%) were shown to be effective, all of the interventions were short in duration (average 10 weeks) and none assessed long-term effectiveness [11].

Online lunch ordering systems (henceforth referred to as “online canteens”), where students select and preorder their lunch using the web or mobile apps, are common in Australian schools [12]. Online canteens represent the optimal infrastructure to implement choice architecture strategies that support students in selecting healthier foods. The research team recently conducted the “Click & Crunch High Schools” cluster randomized controlled trial (RCT). The trial assessed the short-term (2-month) effectiveness of a multi-strategy choice architecture intervention embedded into an online canteen in increasing the relative healthiness of foods purchased at lunch by high school students. At a 2-month follow-up, relative to controls, intervention students purchased significantly more items classified as “everyday” (healthy +5.5%, $P<.001$) and significantly fewer items classified as “should not be sold” (unhealthy –4.4%, $P<.001$) [13]. Although these initial results are promising, evidence suggests that the effects of behavioral interventions can attenuate over time [14,15]. As such, an assessment of the longer-term impact of the intervention on high school students’ lunch purchases is required to better understand how it contributes to long-term behavior change.

Given digital intervention for public health nutrition is still an emerging field, limited studies have been conducted to assess the sustainability of effective interventions. For example, a 2021 umbrella review of 11 systematic reviews of digital interventions to promote healthy eating in children reported that the effectiveness of such interventions in the medium-term and long-term was not well studied [16]. The research targeting adolescents and high school students is sparser still. As this trial [13] was the first to investigate the effectiveness of embedding choice architecture strategies into online canteen ordering systems in high school students, this longer-term follow-up represents a novel contribution to the public health nutrition literature regarding the sustainability of digital health interventions for this underresearched group.

Objectives

Therefore, this study aims to assess the long-term effectiveness (baseline to 15 months) of the “Click & Crunch High Schools” intervention on increasing the relative healthiness of school canteen lunch purchases by high school students.

Methods

Overview

A description of the trial methods has been previously published [13]. The original trial methods and 2-month follow-up were prospectively deposited on the Open Science Framework on October 23, 2020 [17]. The 15-month follow-up was not preregistered however it was conducted per procedures and outcomes as previously registered.

Study Design

This cohort study was conducted as a parallel-group, cluster RCT. Consenting high schools that were using an existing online canteen hosted by Flexischools (InLoop Pty Ltd; a commercial online canteen provider and partner on this research) and located

XSL•FO
RenderX

in NSW Australia were randomized to receive either a multi-strategy choice architecture intervention delivered via the online canteen infrastructure or a usual practice control (ie, standard online canteen). Outcome data were collected over 8 weeks at baseline (October-December 2020), 2 months (the period immediately following intervention commencement, February-April 2021; results previously published) [13], and again at 15-months postintervention commencement (February-April 2022). This paper reports the 15-month findings.

Sample and Recruitment

Schools

School canteen managers from eligible schools were contacted by mail and telephone to invite study participation. A total of 9 (4 intervention and 5 control) government and nongovernment (ie, independent or catholic) schools located in NSW Australia that enrolled high school students (aged ~12-18 years), and used Flexischools as their online canteen provider were eligible to participate in the 15-month follow-up. Schools were ineligible for the trial if they had participated in another unrelated "online canteen" research program conducted by the team or were a catholic school located within a diocese in which ethical approval had not been obtained.

Students

As per prespecified eligibility criteria, students were ineligible for inclusion if; they were in grade 12 at baseline data collection as they were unlikely to be still attending school at the follow-up data collection period; or if they had placed recurring lunch orders set before the intervention period as these orders would not have been exposed to the intervention.

Randomization and Blinding

Following recruitment, an independent statistician block randomized schools (in blocks of 2 and 4) using a random number function in Microsoft Excel. Randomization was stratified by school sector (eg, government vs nongovernment), as evidence suggests there are differences in the availability of healthy food between the school sectors [18]. Schools were unable to be blinded to their group allocation. However, the intervention was applied centrally, and only students at intervention schools could access the intervention strategies via the online ordering system. All student purchasing data was centrally collected by the online provider, reducing any risk of intervention contamination between the groups.

Intervention

Overview

The "Click & Crunch High Schools" intervention is described in full elsewhere [13]. The intervention was underpinned by the principals of choice architecture and sought to encourage the purchase of healthier (ie, "everyday") items from the school's online canteen menu. All intervention strategies were integrated into the schools' existing online canteen and were displayed to students at the point of purchase. All student users of the online canteen at eligible high schools had access to the intervention strategies. The intervention was in place for approximately 15 months (February 2021 to April 2022) until after the 15-month follow-up data collection period. Intervention strategies (see Figure 1) are described in the following sections.

XSL•FO
RenderX

Figure 1. Screenshots from the online canteen showing the following intervention strategies: (A) menu labeling and positioning, (B) feedback, and (C) prompts.

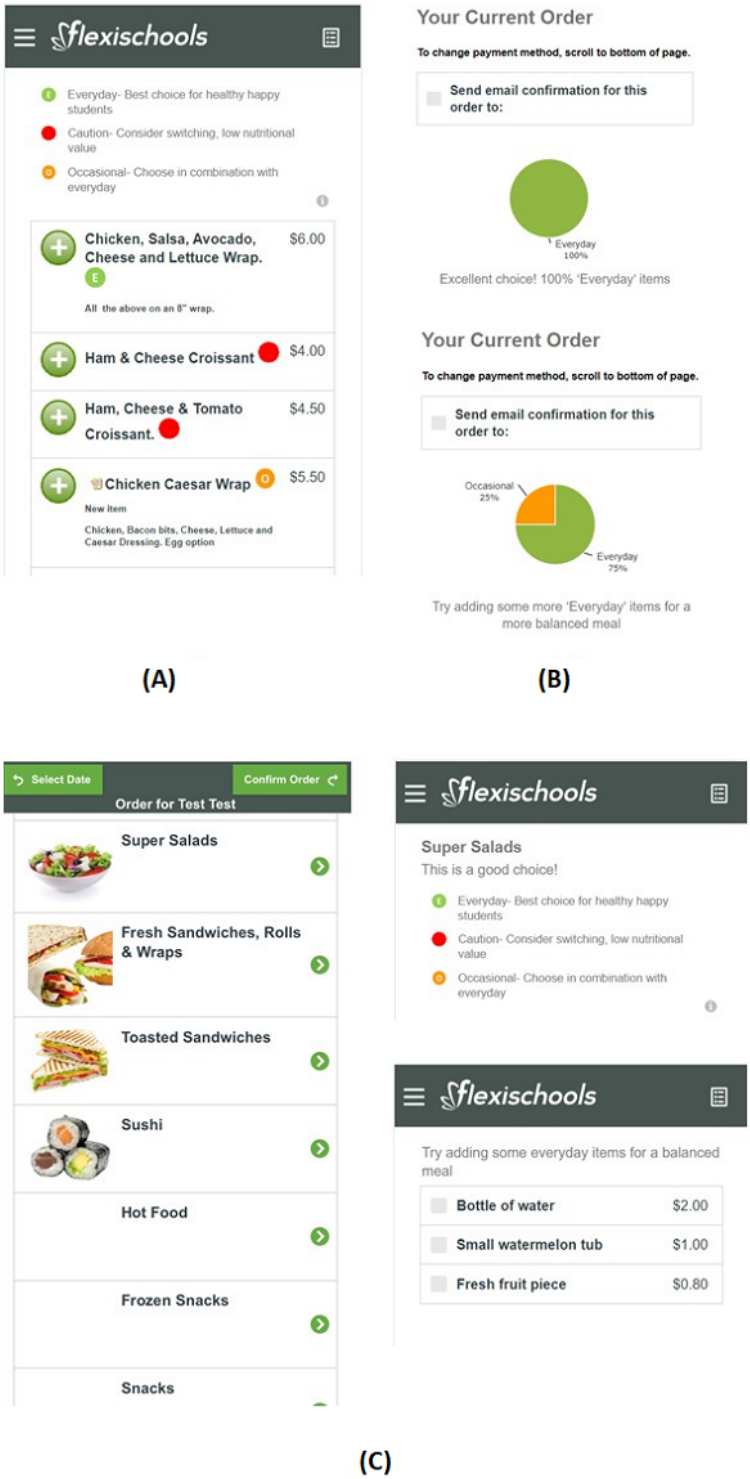

Menu Labeling

All menu items were classified as either "everyday," "occasional," or "should not be sold" based on the criteria outlined in the NSW Healthy School Canteen Strategy. Menu items were labeled with a small colored symbol: a green circle was added next to "everyday" foods, an amber circle was added next to "occasional" foods, and a red circle was added next to "should not be sold" ("caution") foods. A "menu label key" appeared at the top of the page (eg, "Everyday- best choice for healthy happy students"; "Occasional- choose in combination with Everyday"; "Caution- consider switching, low nutritional value").

XSL•FO
RenderX

Positioning

"Everyday" menu items and healthier food categories (eg, fruit, salad, and sandwiches) were positioned prominently (ie, first) in the online menu. Research suggests items placed in the middle of menu lists are two times less likely to be purchased than those at the beginning or end [19]. Therefore, the least healthy "should not be sold" items were placed in the middle and "occasional" items were placed last in menu category lists, respectively. Further, "occasional" or "should not be sold" items with multiple flavors (eg, potato crisps) required the user to first "click" on the item before the full list of flavors appeared (eg, plain, salt and vinegar, and chicken).

Feedback

Before each lunch order was finalized within the online ordering system, users were shown a personalized summary of the healthiness of their lunch order. The summary included a pie graph displaying the proportion of items in their order that was "everyday," "occasional," and "should not be sold," and a tailored message based on the proportion of "everyday" items in the order (eg, if <99% of items were "everyday": "Try adding some 'Everyday' items for a more balanced meal." If 100% of items were "everyday": "Excellent choice! 100% 'Everyday' items").

Prompts

When "occasional" or "should not be sold" hot food items were chosen they included a prompt to add a fruit or vegetable snack and water. Healthier menu categories (eg, fruit, salad, and sandwiches) included an appealing image and positive purchase prompt (eg, "This is a good choice").

To support canteen managers' understanding of the NSW Healthy School Canteen classification system which underpinned the menu labeling, each canteen manager in the intervention group received a "menu feedback report." The report included feedback comparing the online canteen menu to the recommendations of the NSW Healthy School Canteen Strategy and provided suggestions on how to improve the relative availability of "everyday" items on the menu.

Intervention Fidelity

Once every term during the intervention period (approximately every 10 weeks), a member of the research team monitored each school's online canteen menu via the Flexischools website. They checked that all menu items, including any new items, were correctly classified according to the NSW Healthy School Canteen Strategy, and that the intervention strategies were applied accordingly. If any menu items were found to be unlabeled or incorrectly labeled, the research team would notify Flexischools and provide instructions for how to apply the intervention strategy correctly.

Control

Control schools did not receive any of the intervention strategies, and were only provided access to the standard online ordering system.

Data Collection and Outcomes

Overview

Student purchasing data were automatically collected and stored by Flexischools. Data were collected over 3 distinct 8-week periods, with baseline occurring from October to December 2020 and long-term follow-up occurring 15 months after the intervention commenced (February-April 2022). The 2-month follow-up was the primary trial end point (data collected immediately following intervention commencement, February-April 2021), and has been previously published [13].

Primary Trial Outcomes

The primary trial outcomes at the 15-month follow-up were identical to those at the 2-month follow-up and included the mean percentage of all online lunch items purchased per student that were classified according to the NSW Healthy School Canteen Strategy as (1) "everyday," (2) "occasional," and (3) "should not be sold." The NSW Healthy Canteen Strategy classifies foods as "everyday" based on their alignment with the core foods groups within the Australian Dietary Guidelines (eg, fruit, vegetables, dairy and alternatives, lean meat and alternatives, and grains) [20]. Menu items classified as "occasional" or "should not be sold" are considered "noncore" or discretionary foods that are mostly high in energy, saturated fat, sugar, and salt. Further information on the NSW Healthy School Canteen strategy including the nutrition criteria underpinning the strategy are reported elsewhere [20].

Each canteen menu item was classified against the strategy by a research dietitian using detailed item information (ie, brand, product name, service size, flavor, or recipe) obtained from the canteen manager via telephone or email. Following this, a statistician was able to apply the menu item classification (eg, "everyday") to the automatically collected purchase data supplied by Flexischools (eg, fresh fruit equaled "everyday").

Secondary Trial Outcomes

Energy, Saturated Fat, Sugar, and Sodium Content of Online Lunch Orders

Secondary outcomes included the mean total energy (kJ), saturated fat (g), sugar (g), and sodium (mg) content of online lunch orders. Using previously established procedures, the dietitian generated the nutrition profile for each menu item by using data from food product databases (for commercially packaged menu items [13,21,22]) or FoodWorks (version 9; Xyris Software; for menu items requiring a recipe). The statistician then applied the nutritional profile of each menu item to the student purchasing data provided by Flexischools.

Weekly Canteen Revenue From Online Orders

Purchasing data that were automatically collected by Flexischools were used to calculate the mean weekly revenue from all student online lunch orders for the weeks that the canteen was operational at baseline and long-term follow-up. This outcome was assessed to explore any potential adverse effect of the intervention (eg, a reduction in canteen revenue due to the application of the intervention strategies).

XSL•FO
RenderX

Other Data

School Characteristics

At baseline, school characteristics including the number of student enrollments, year range, sector (eg, government vs nongovernment), school type (combined primary and high school students' vs high school only), and postcode were obtained from the government "MySchool" website. As the number of high school student enrollments for combined schools was not available on the "MySchool" website, this data was collected directly from the school.

Canteen Characteristics

Canteen characteristics including operating days per week, frequency of use, and student grade data were obtained from the student purchasing data supplied by Flexischools.

Menu Composition (Pre-Post Intervention)

Using the methods outlined above, a research dietitian assessed the proportion of items on each school's online menus that were classified as "everyday," "occasional," and "should not be sold." This is reported by intervention and control groups at baseline and 15-month follow-up.

Statistical Analysis

Overview

All outcome data were analyzed in SAS (version 9.3; SAS Institute) under an intention-to-treat (ITT) approach whereby all student lunch orders and schools were analyzed based on the groups they were originally allocated. All nutrition outcomes included data from the student cohort (grades 7-11) that had placed at least one order during the baseline period.

Primary and secondary outcomes were assessed using separate linear mixed models by comparing differences between intervention and control groups over time (baseline to 15 months) through the inclusion of a group-by-time interaction fixed effect. All models included a random intercept for schools (to account for potential school-level clustering), a nested random intercept and random time effect for students (to account for repeated measurements between time points), and fixed effects for the school sector and SEIFA (Socio-Economic Indexes for Australia). All available data (baseline, 2 months, and 15 months) were incorporated into the model.

Consistent with previous publications, the denominator for the unit of analysis for primary trial outcomes was the total number of individual items purchased for each student over the three 8-week data collection periods (baseline: October-December 2020; 2 months: February-April 2021; 15 months: February-April 2022).

Differences in the average weekly revenue (a school-level outcome) were assessed using linear mixed models and included data from all students who had placed any order during any of the data collection periods. School and canteen characteristics were previously reported in the 2-month outcome paper and are included here for context.

Given no differences were observed by subgroups (student grade, frequency of canteen use, or school sector) at the primary trial end point (2 months), no subgroup analyses were conducted at the 15-month follow-up.

Sample Size

No sample size calculation was performed for long-term follow-up, sample size estimates were calculated a priori based on the primary trial end point of 2 months [13]. The original sample size required the participation of 10 schools (222 students per school) to ensure a mean detectable difference of 13% of everyday items with 80% power, an intraclass correlation coefficient of 0.05, and an α of .05 at 2-month follow-up.

Ethical Considerations

The ethical approval for the conduct of this study was provided by the Human Research Ethics Committee of the University of Newcastle (H-2017-0402), and State Education Research Approval Process (SERAP 2018065), as well as relevant Catholic School Dioceses.

Results

Overview

The baseline characteristics of the sample are presented in Table 1. At baseline, on average, control schools had higher student enrollments compared with intervention schools (mean enrollments 800, SD 318 vs 496, SD 226). All other baseline characteristics were similar between groups (no significance testing was performed). For example, all school canteens operated 5 days per week, the majority of schools were located in areas of most socioeconomic advantage, and the majority of students were in grades 7 to 9. The number of participants and orders at baseline and 15-month follow-up can be seen in Figure 2. While 1331 students from 9 schools provided data at baseline, 332 (25%) students did not place an online order at the 2-month follow-up, and an additional 268 (20%) students did not place an online order at the 15-month follow-up. Of these 268 students, 70 had completed high school (ie, students that were in grade 11 at baseline were no longer at school 15 months later). There were no statistically significant differences between intervention and control participants being lost to follow-up (P=.08).

XSL•FO
RenderX

Table 1. Characteristics of participating NSW[a] schools (N=9) and students (N=1331) at baseline.

	Intervention	Control
School and canteen characteristics (intervention: n=4 schools; control: n=5 schools)		
School sector, n (%)		
Government	1 (25)	2 (40)
Nongovernment[b]	3 (75)	3 (60)
School type, n (%)		
Combined school (students aged 5-19 years)	3 (75)	2 (40)
High school (students aged 12-19 years)	1 (25)	3 (60)
Number of enrollments[c], mean (SD)	496 (226)	800 (318)
Socioeconomic status of school[d], n (%)		
Least advantaged	1 (25)	2 (40)
Most advantaged	3 (75)	3 (60)
Canteen days of operation[e], n (%)		
Five days a week	4 (100)	5 (100)
Canteen menu characteristics[f], n (%)		
In total, ≥75% “Everyday” items on menu	2 (50)	2 (40)
No “Should not be sold” items on menu	0 (0)	0 (0)
Number of weekly online lunch orders per school[e], mean (SD)	141 (62)	135 (77)
User characteristics[e] (intervention: n=656 participants; control: n=675 participants)		
Grade of student at baseline, n (%)		
Grade 7-9	503 (77)	541 (80)
Grade 10-11	153 (23)	134 (20)
Frequency of use, n (%)		
High users (≥1 order per week on average)	166 (25)	219 (32)
Low users (<1 order per week on average)	490 (75)	456 (68)

[a]NSW: New South Wales.

[b]Nongovernment schools were Catholic and independent schools.

[c]Based on publicly available school statistics (MySchool 2020) or verbally from schools (combined schools only).

[d]Socio-Economic Indexes for Australia 2016, based on the postcode of the school locality and dichotomized at the NSW median.

[e]Based on Flexischools purchasing data.

[f]As classified by a dietitian according to the New South Wales Healthy School Canteen Strategy.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram.

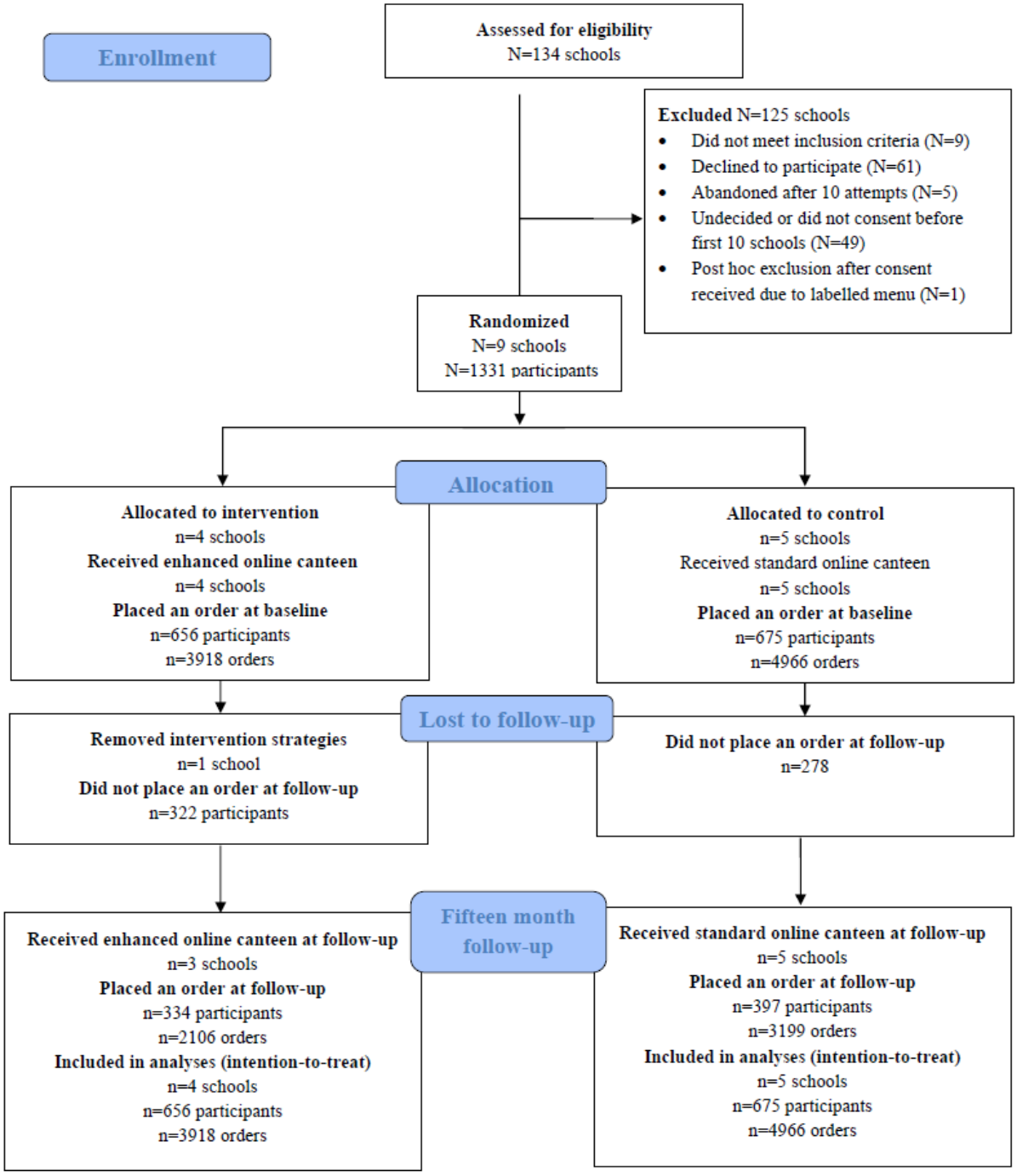

Primary Outcomes

The primary outcomes were the average proportion, per student, of "everyday," "occasional," and "should not be sold" online lunch items purchased. Relative to controls, over time from baseline to 15-month follow-up, students in the intervention group ordered on average significantly more "everyday" items (+11.5%, 95% CI 7.3% to 15.6%; $P<.001$), and significantly fewer "occasional" (–5.4%, 95% CI –9.4% to –1.5%; $P=.007$) and "should not be sold" items (–6%, 95% CI –9.1% to –2.9%; $P<.001$) in an intention-to-treat (ITT) analysis (Table 2).

XSL•FO
RenderX

Table 2. Differences in intervention and control groups over time for primary and secondary outcomes[a].

	Intervention		Control		Intervention versus control[a]	
	Baseline (n=656), mean (SD)	15-Month follow-up (n=334), mean (SD)	Baseline (n=675), mean (SD)	15-Month follow-up (n=397), mean (SD)	Main analysis (15-month versus baseline)	
					Group by time differential effect (95% CI)	*P* value
Primary outcomes						
Percentage per student of lunch items that are "everyday"	44.3 (34.3)	51.0 (34.0)	43.2 (36.3)	39.4 (36.0)	11.5 (7.3 to 15.6)	<.001[b]
Percentage per student of lunch items that are "occasional"	30.7 (30.8)	30.4 (29.7)	40.9 (35.7)	45.7 (37.7)	–5.4 (–9.4 to –1.5)	.007[b]
Percentage per student of lunch items that are "should not be sold"	25.0 (30.9)	18.6 (28.3)	16.0 (24.7)	15.0 (23.2)	–6.0 (–9.1 to –2.9)	<.001[b]
Secondary outcomes						
Energy (kJ) per student lunch order	2172.5 (976.9)	2211.7 (922.3)	1992.8 (793.1)	1943.4 (796.6)	48.8 (–34.6 to 132.2)	.25
Saturated fat (g) per student lunch order	7.3 (5.7)	7.0 (5.0)	5.7 (4.4)	5.5 (4.4)	–0.0 (–0.5 to 0.5)	.99
Sugar (g) per student lunch order	22.4 (22.8)	22.5 (21.0)	15.3 (16.2)	12.8 (15.0)	1.7 (–0.1 to 3.5)	.07
Sodium (mg) per student lunch order	778.9 (354.1)	795.9 (368.8)	808.3 (398.2)	789.3 (384.1)	0.35 (–36.2 to 36.9)	.99
Weekly revenue (Aus $)[c] per school	896.1 (449.9)	1243.8 (485.9)	769.6 (372.8)	1798.5 (653.9)	–673.4 (–1252.6 to –94.2)	.03[b]

[a]All models included a random intercept for school, a nested random intercept and random time effect for students, and fixed effects for the school sector and Socio-Economic Indexes for Australia. All available data were incorporated into the model (baseline, 2-months, and 15-months) to describe purchasing patterns over time.

[b]*P*<.05.

[c]All $ amounts are in Aus $. A currency exchange rate of Aus $1 = US $0.65 was applicable as of February 2024.

Secondary Outcomes

Average Energy, Saturated Fat, Sugar, and Sodium Content of Online Lunch Orders

There were no between-group differences over time (baseline to 15-month follow-up) in the average energy (+48.8 kJ, 95% CI –34.6 to 132.2; *P*=.25), saturated fat (–0.0 g, 95% CI –0.5 to 0.5; *P*=.99), sugar (+1.7 g, 95% CI –0.1 to 3.5; *P*=.07), or sodium (+0.35, 95% CI –36.2 to 36.9; *P*=.99) content of student lunch orders.

Weekly Online Canteen Revenue (Potential Adverse Effect)

While both intervention and control groups increased in revenue (a currency exchange rate of Aus $1=US $0.65 applies) over time (intervention-group baseline: Aus $896.10; intervention-group 15-month follow-up: Aus $1243.80; control-group baseline: Aus $769.60; control-group 15-month follow-up: Aus $1798.50), the increase in the intervention group was significantly lower than the increase in the control group (differential effect –Aus $673.40, 95% CI –Aus $1252.60 to –Aus $94.20; *P*=.03). To further qualify this effect, a post hoc exploratory analysis was undertaken to explore if students spent more money per order between intervention and control groups over time. The exploratory analysis found no difference in the average spend per student order by intervention and control groups over time (difference Aus $0.07, 95% CI –Aus $0.14 to Aus $0.28; *P*=.48).

Other Data

Menu Composition

While no significance testing was performed, the proportion of "everyday," "occasional," and "should not be sold" items available on menus at baseline and 15-month follow-up were similar for the intervention and control schools (Table 3).

XSL•FO
RenderX

Table 3. Menu composition at baseline and 15-month follow-up.

Menu item classification	Intervention		Control	
	Baseline availability, mean %	15-month follow-up availability, mean %	Baseline availability, mean %	15-month follow-up availability, mean %
Everyday	69.4	69.7	68	69.1
Occasional	14.9	15.5	20	21.7
Should not be sold	15.7	14.8	12	9.2

Intervention Fidelity

Of the 4 intervention schools, 3 had 99% (1256/1269 items) of their menu labeled correctly during the 15-month intervention period. The remaining school removed all of their labels in the last 12 weeks of the intervention, resulting in 81% fidelity across the 15-month intervention period.

Discussion

Principal Results

This is the first study to assess the long-term effectiveness of an intervention embedded within an online lunch ordering system for high school students and is one of few studies to assess the long-term effectiveness of food choice architecture interventions more broadly [10,23,24]. The Click & Crunch High Schools cluster RCT found that intervention students, relative to control, ordered significantly more healthy "everyday" items and significantly fewer "less healthy" items from baseline to 15-month follow-up. There were no between-group differences over time in the average energy, saturated fat, sugar, and sodium content of high school student online lunch orders. This study found that the online canteen revenue for both groups increased over 15 months, however, the revenue in the intervention group grew more slowly than the control group. These findings were surprising, given other trials in the school food setting have found no differences between groups in revenue [25-27].

Comparison With Prior Work

While there is limited research to draw direct comparisons of this study, systematic reviews of the school setting have found that very few studies have assessed the long-term effectiveness of nutrition interventions in high schools [24,28]. A systematic review by Mingay and colleagues [24] found that only 6 of 35 studies assessed the long-term effect (≥12 months) of school meal interventions on the selection or purchase of healthier foods by high school students. Similar to our study, the review found mixed evidence for studies that included multiple dietary outcomes (eg, nutrients vs food groups) in their assessment of long-term effectiveness. In contrast to our study, the review found that shorter interventions (<3 months) had a greater effect on dietary outcomes for high school students [24]. Contrary to these review findings, our study found that there was a greater magnitude of effect at 15-months compared to the 2-month follow-up (previously reported) [13]. For example, at 2 months the Click & Crunch High School intervention was effective in increasing "everyday" items (+5.5%, $P<.001$) and decreasing "should not be sold" (–4.4%, $P<.001$) items purchased by students, with no difference in the purchase of "occasional" items (–1.2%, $P=.47$) [13]. At the 15-month follow-up, the magnitude of effect was greater than that observed at 2 months and the decrease in "occasional" items purchased was now significant (15-months: everyday +11.5%, $P<.001$; occasional –5.4%, $P=.007$; should not be sold –6%, $P<.001$). The increase in effect size over time may in part be explained by the high intervention fidelity, the intervention type (choice architecture vs food provision), and the number of strategies employed in this trial. Furthermore, the greater length of time that students were exposed to the intervention may have increased the likelihood of habitual patterns in the purchasing of more healthy foods. The sustained intervention effectiveness may also be attributable to the precommitment involved with "preordering," which may prevent impulse purchasing of "less healthy" foods due to hunger-based cues [10].

Although this is the first RCT to describe the long-term effectiveness of an online choice architecture intervention in the high school setting (enrolling students aged ~12-18 years) a similar pattern of results has been found in related food service settings [29,30]. For example, a longitudinal study undertaken with adults in a large hospital cafeteria found that a 2-year choice architecture intervention involving traffic light labeling, product placement, and promotion increased the sale of "healthy" items by 5% and decreased the sale of unhealthy items by 3% ($P<.001$). In the primary school setting (aged 5-12 years), the same Click & Crunch intervention was found to be effective at improving healthy food purchases by primary school students at 18 months (+3.8% "everyday" and –2.6% "less healthy" items purchased) [29]. Such findings demonstrate the potential merit of the Click & Crunch intervention on improving the nutritional quality of both primary and high school student online lunch purchases over both the short and longer term and challenge the previously held notion that choice architecture interventions may attenuate over time due to their "novelty effect" or "label fatigue" experienced by end users [10].

Broader Implications of This Research

The findings of this trial may have broader relevance to the online food ordering systems more generally. The World Health Organization has identified the need to leverage online food delivery systems for public health benefits [31]. This is the first trial to embed public health nutrition strategies within online food ordering systems for adolescents. With the exponential rise in related meal delivery app use particularly by adolescents and young adults (aged >15 years) [32], these research findings are likely to be of interest to policy makers investigating how to leverage such systems for public health benefit.

XSL•FO
RenderX

Strengths and Limitations

This cluster RCT had several strengths, including the robust trial design, objectively collected purchase data, and the use of a real-world online lunch ordering system to deliver simple choice architecture strategies. Importantly, it is one of few studies assessing the long-term effects on food purchase or consumption of a choice architecture intervention and the first to do so in the high school setting. Despite this, this study had several limitations. In addition to those already discussed in the 2-month follow-up [13], this study did not assess intervention costs or acceptability which are key determinants of intervention scalability [33]. Therefore, to support public health decision-making regarding the scalability of these interventions, future research that explores the acceptability of the intervention to end users (high school canteen managers and students) and intervention costs including cost-effectiveness may be warranted. Furthermore, as this study did not find differences in nutrient outcomes (energy, saturated fat, sugar, and sodium), future research may be required to understand the differential effect of alternate menu labeling systems (eg, kJ labeling) on nutrient-based outcomes. Finally, as outlined in the 2-month follow-up [13], to achieve population-wide improvements in adolescent nutrition this intervention should be considered in addition to broader public health nutrition strategies that reach both users and nonusers of online canteens in the high school setting.

Conclusions

Despite the limitations, this is the first RCT to explore the long-term effectiveness of a choice architecture intervention embedded within an online canteen targeting high-school students and one of only a few choice architecture interventions delivered in the high-school setting. The findings suggest that there are long-term effects of up to 15 months after intervention commencement, including a significant increase in healthy “everyday” items and a significant reduction in less healthy “occasional” and “should not be sold” items. This provides valuable evidence about the potential long-term effect of choice architecture interventions delivered via online canteens on adolescent school lunch ordering and may be useful to policy makers interested in improving adolescent diet within the high school setting. Further research is required to determine the feasibility of disseminating such interventions to schools at scale, and if these effects transfer to other online food environments targeting different end users (ie, adults and health care workers) such as workplaces, hospital settings, and the fast food sector.

Acknowledgments

We wish to thank Flexischools, the research advisory group, and the participating schools, canteen managers, and canteen users. This research was funded by the National Heart Foundation of Australia (102809). RW is supported by a Heart Foundation Postdoctoral Fellowship (102156) and Cancer Institute NSW Early Career Fellowship (2021/ECF1355). LW receives salary support from a National Health and Medical Research Council Investigator Grant Fellowship (APP1197022). RS is supported by a Medical Research Future Fund Investigator Grant (APP1194768). SLY is supported by a National Heart Foundation Future Leader Fellowship (106654). The funders had no role in the conduct of the trial or the analysis or interpretation of findings. The provider (Flexischools) was selected through a competitive tender process. Flexischools is a commercial organization that provided the online canteen ordering infrastructure to schools that were included in this study. Flexischools had no role in this study’s design, data analysis, data interpretation, or writing of this paper.

Data Availability

The data sets generated during or analyzed for this study are available from the corresponding author on reasonable request, pending ethics approval.

Authors' Contributions

RW conceived this study. TD and RW developed the methodology. CL conducted the formal analysis. TD, TC-M, RW, and HL conducted research. CL, TC-M, and TD curated data. TD and JJ led the writing of this paper with all authors contributing to paper revisions. LW, SLY, RS, and RW provided supervision. RW acquired funding. All authors read and approved the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH checklist.
[PDF File (Adobe PDF File), 619 KB - jmir_v26i1e51108_app1.pdf]

References

1. Champion KE, Chapman C, Gardner LA, Sunderland M, Newton NC, Smout S, et al. Lifestyle risks for chronic disease among Australian adolescents: a cross-sectional survey. Med J Aust 2022;216(3):156-157 [FREE Full text] [doi: 10.5694/mja2.51333] [Medline: 34747039]

XSL•FO
RenderX

2. Mayne SL, Virudachalam S, Fiks AG. Clustering of unhealthy behaviors in a nationally representative sample of U.S. children and adolescents. Prev Med 2020;130:105892. [doi: 10.1016/j.ypmed.2019.105892] [Medline: 31715216]
3. Benetou V, Kanellopoulou A, Kanavou E, Fotiou A, Stavrou M, Richardson C, et al. Diet-related behaviors and diet quality among school-aged adolescents living in Greece. Nutrients 2020;12(12):3804 [FREE Full text] [doi: 10.3390/nu12123804] [Medline: 33322459]
4. Khalid S, Williams CM, Reynolds SA. Is there an association between diet and depression in children and adolescents? A systematic review. Br J Nutr 2016;116(12):2097-2108 [FREE Full text] [doi: 10.1017/S0007114516004359] [Medline: 28093091]
5. Dutch DC, Golley RK, Johnson BJ. Diet quality of Australian children and adolescents on weekdays versus weekend days: a secondary analysis of the national nutrition and physical activity survey 2011-2012. Nutrients 2021;13(11):4128 [FREE Full text] [doi: 10.3390/nu13114128] [Medline: 34836383]
6. Todd AS, Street SJ, Ziviani J, Byrne NM, Hills AP. Overweight and obese adolescent girls: the importance of promoting sensible eating and activity behaviors from the start of the adolescent period. Int J Environ Res Public Health 2015;12(2):2306-2329 [FREE Full text] [doi: 10.3390/ijerph120202306] [Medline: 25690003]
7. Craigie AM, Lake AA, Kelly SA, Adamson AJ, Mathers JC. Tracking of obesity-related behaviours from childhood to adulthood: a systematic review. Maturitas 2011;70(3):266-284. [doi: 10.1016/j.maturitas.2011.08.005] [Medline: 21920682]
8. Hunter D, Giyose B, PoloGalante A, Tartanac F, Bundy D, Mitchell A, et al. Schools as a system to improve nutrition: a new statement for school-based food and nutrition interventions. United Nations System Standing Committee on Nutrition. 2017. URL: https://www.unscn.org/uploads/web/news/document/School-Paper-EN-WEB-8oct.pdf [accessed 2024-02-21]
9. Clinton-McHarg T, Delaney T, Lamont H, Lecathelinais C, Yoong SL, Wolfenden L, et al. A cross-sectional study of the nutritional quality of New South Wales high school student food and drink purchases made via an online canteen ordering system. Nutrients 2021;13(12):4327 [FREE Full text] [doi: 10.3390/nu13124327] [Medline: 34959881]
10. Chambers T, Segal A, Sassi F. Interventions using behavioural insights to influence children's diet-related outcomes: a systematic review. Obes Rev 2021;22(2):e13152 [FREE Full text] [doi: 10.1111/obr.13152] [Medline: 33462932]
11. Ensaff H, Homer M, Sahota P, Braybrook D, Coan S, McLeod H. Food choice architecture: an intervention in a secondary school and its impact on students' plant-based food choices. Nutrients 2015;7(6):4426-4437 [FREE Full text] [doi: 10.3390/nu7064426] [Medline: 26043039]
12. Delaney T, Wolfenden L, Wyse R. Online food delivery systems and their potential to improve public health nutrition: a response to 'a narrative review of online food delivery in Australia'. Public Health Nutr 2023;26(1):273-274 [FREE Full text] [doi: 10.1017/S1368980021000926] [Medline: 33706848]
13. Delaney T, Yoong SL, Lamont H, Lecathelinais C, Wolfenden L, Clinton-McHarg T, et al. The efficacy of a multi-strategy choice architecture intervention on improving the nutritional quality of high school students' lunch purchases from online canteens (click and crunch high schools): a cluster randomized controlled trial. Int J Behav Nutr Phys Act 2022;19(1):120 [FREE Full text] [doi: 10.1186/s12966-022-01362-5] [Medline: 36104738]
14. Dave JM. Assessing long-term impact of dietary interventions: important but challenging. Am J Clin Nutr 2019;109(1):5-6 [FREE Full text] [doi: 10.1093/ajcn/nqy339] [Medline: 30624574]
15. Yavuz HM, van Ijzendoorn MH, Mesman J, van der Veek S. Interventions aimed at reducing obesity in early childhood: a meta-analysis of programs that involve parents. J Child Psychol Psychiatry 2015;56(6):677-692. [doi: 10.1111/jcpp.12330] [Medline: 25292319]
16. Prowse R, Carsley S. Digital interventions to promote healthy eating in children: umbrella review. JMIR Pediatr Parent 2021;4(4):e30160 [FREE Full text] [doi: 10.2196/30160] [Medline: 34842561]
17. A cluster randomised controlled trial of the impact of a multi-strategy behavioural intervention on the nutritional quality of secondary school student food purchases from online canteens. Open Science Framework. 2020. URL: https://osf.io/h8zfr/ [accessed 2024-02-21]
18. Reilly KL, Nathan N, Wiggers J, Yoong SL, Wolfenden L. Scale up of a multi-strategic intervention to increase implementation of a school healthy canteen policy: findings of an intervention trial. BMC Public Health 2018;18(1):860 [FREE Full text] [doi: 10.1186/s12889-018-5786-x] [Medline: 29996817]
19. Dayan E, Bar-Hillel M. Nudge to nobesity II: menu positions influence food orders. Judgm Decis Mak 2011;6(4):333-342 [FREE Full text] [doi: 10.1017/s1930297500001947]
20. The NSW Healthy School Canteen Strategy: Food and Drink Criteria (4th Edition). St Leonards, Australia: NSW Ministry of Health; 2020.
21. FoodFinder. Healthy food finder. NSW Government. URL: https://www.foodfinder.health.nsw.gov.au [accessed 2024-02-21]
22. FoodSwitch. The George Institute for Global Health. URL: https://www.foodswitch.com.au/ [accessed 2024-02-21]
23. Metcalfe JJ, Ellison B, Hamdi N, Richardson R, Prescott MP. A systematic review of school meal nudge interventions to improve youth food behaviors. Int J Behav Nutr Phys Act 2020;17(1):77 [FREE Full text] [doi: 10.1186/s12966-020-00983-y] [Medline: 32560731]
24. Mingay E, Hart M, Yoong S, Palazzi K, D'Arcy E, Pursey KM, et al. The impact of modifying food service practices in secondary schools providing a routine meal service on student's food behaviours, health and dining experience: a systematic review and meta-analysis. Nutrients 2022;14(17):3640 [FREE Full text] [doi: 10.3390/nu14173640] [Medline: 36079897]

XSL•FO
RenderX

25. Delaney T, Wyse R, Yoong SL, Sutherland R, Wiggers J, Ball K, et al. Cluster randomized controlled trial of a consumer behavior intervention to improve healthy food purchases from online canteens. Am J Clin Nutr 2017;106(5):1311-1320 [FREE Full text] [doi: 10.3945/ajcn.117.158329] [Medline: 28971849]
26. French SA, Story M, Fulkerson JA, Hannan P. An environmental intervention to promote lower-fat food choices in secondary schools: outcomes of the TACOS study. Am J Public Health 2004;94(9):1507-1512 [FREE Full text] [doi: 10.2105/ajph.94.9.1507] [Medline: 15333303]
27. Wolfenden L, Nathan N, Janssen LM, Wiggers J, Reilly K, Delaney T, et al. Multi-strategic intervention to enhance implementation of healthy canteen policy: a randomised controlled trial. Implement Sci 2017;12(1):6 [FREE Full text] [doi: 10.1186/s13012-016-0537-9] [Medline: 28077151]
28. Durão S, Wilkinson M, Davids EL, Gerritsen A, Kredo T. Effects of policies or interventions that influence the school food environment on children's health and nonhealth outcomes: a systematic review. Nutr Rev 2024;82(3):332-360 [FREE Full text] [doi: 10.1093/nutrit/nuad059] [Medline: 37253393]
29. Wyse R, Delaney T, Stacey F, Lecathelinais C, Ball K, Zoetemeyer R, et al. Long-term Effectiveness of a multistrategy behavioral intervention to increase the nutritional quality of primary school students' online lunch orders: 18-month follow-up of the Click and Crunch cluster randomized controlled trial. J Med Internet Res 2021;23(11):e31734 [FREE Full text] [doi: 10.2196/31734] [Medline: 34847063]
30. Thorndike AN, Riis J, Sonnenberg LM, Levy DE. Traffic-light labels and choice architecture: promoting healthy food choices. Am J Prev Med 2014;46(2):143-149 [FREE Full text] [doi: 10.1016/j.amepre.2013.10.002] [Medline: 24439347]
31. Slide to order: a food systems approach to meals delivery apps: WHO European Office for the Prevention and Control of Noncommunicable Diseases. World Health Organization. 2021. URL: https://www.who.int/europe/publications/i/item/WHO-EURO-2021-4360-44123-62247 [accessed 2024-02-21]
32. Partridge SR, Gibson AA, Roy R, Malloy JA, Raeside R, Jia SS, et al. Junk food on demand: a cross-sectional analysis of the nutritional quality of popular online food delivery outlets in Australia and New Zealand. Nutrients 2020;12(10):3107 [FREE Full text] [doi: 10.3390/nu12103107] [Medline: 33053705]
33. Milat AJ, King L, Bauman AE, Redman S. The concept of scalability: increasing the scale and potential adoption of health promotion interventions into policy and practice. Health Promot Int 2013;28(3):285-298 [FREE Full text] [doi: 10.1093/heapro/dar097] [Medline: 22241853]

Abbreviations

ITT: intention-to-treat
NSW: New South Wales
RCT: randomized controlled trial
SEIFA: Socio-Economic Indexes for Australia

Edited by A Mavragani; submitted 21.07.23; peer-reviewed by S Beresford; comments to author 25.09.23; revised version received 01.10.23; accepted 30.01.24; published 19.03.24.

Please cite as:
Delaney T, Jackson J, Lecathelinais C, Clinton-McHarg T, Lamont H, Yoong SL, Wolfenden L, Sutherland R, Wyse R
Long-Term Effectiveness of a Multi-Strategy Choice Architecture Intervention in Increasing Healthy Food Choices of High-School Students From Online Canteens (Click & Crunch High Schools): Cluster Randomized Controlled Trial
J Med Internet Res 2024;26:e51108
URL: https://www.jmir.org/2024/1/e51108
doi:10.2196/51108
PMID:38502177

©Tessa Delaney, Jacklyn Jackson, Christophe Lecathelinais, Tara Clinton-McHarg, Hannah Lamont, Sze Lin Yoong, Luke Wolfenden, Rachel Sutherland, Rebecca Wyse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial

Hailey N Miller[1], RN, PhD; John A Gallis[2,3], ScM; Miriam B Berger[4], MPH; Sandy Askew[4], MPH; Joseph R Egger[3], PhD; Melissa C Kay[5], RD, MS, MPH, PhD; Eric Andrew Finkelstein[6], PhD; Mia de Leon[7], BSE; Abigail DeVries[8], MD; Ashley Brewer[9], RD; Marni Gwyther Holder[10], RN, MSN; Gary G Bennett[4,11], PhD

[1]School of Nursing, Johns Hopkins University, Baltimore, MD, United States
[2]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, United States
[3]Duke Global Health Institute, Duke University, Durham, NC, United States
[4]Duke Digital Health Science Center, Duke University, Durham, NC, United States
[5]Department of Pediatrics, Duke University, Durham, NC, United States
[6]Duke-NUS Medical School Singapore, Duke Global Health Institute, Duke University, Durham, NC, United States
[7]Caraway, New York, NY, United States
[8]Medical Home Network, Chicago, IL, United States
[9]Piedmont Health Services, Inc, Chapel Hill, NC, United States
[10]Department of Family Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[11]Trinity College of Arts & Sciences, Duke University, Durham, NC, United States

Corresponding Author:
Miriam B Berger, MPH
Duke Digital Health Science Center
Duke University
417 Chapel Drive, Room 048
Campus Box 90086
Durham, NC, 27708
United States
Phone: 1 919 613 8398
Email: miriam.berger@duke.edu

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e60137

Abstract

Background: The prevalence of obesity and its associated comorbidities continue to rise in the United States. Populations who are uninsured and from racial and ethnic minority groups continue to be disproportionately affected. These populations also experience fewer clinically meaningful outcomes in most weight loss trials. Weight gain prevention presents a useful strategy for individuals who experience barriers to weight loss. Given the often-limited weight management resources available to patients in primary care settings serving vulnerable patients, evaluating interventions with pragmatic designs may help inform the design of comprehensive obesity care delivered in primary care.

Objective: This study aims to evaluate the effectiveness of Balance, a 2-arm, 12-month pragmatic randomized controlled trial of a digital weight gain prevention intervention, delivered to patients receiving primary care within federally qualified community health centers.

Methods: Balance was a 2-arm, 12-month pragmatic randomized controlled trial of a digital weight gain prevention intervention delivered to individuals who had a BMI of 25-40 kg/m^2, spoke English or Spanish, and were receiving primary care within a network of federally qualified community health centers in North Carolina. The Balance intervention was designed to encourage behavioral changes that result in a slight energy deficit. Intervention participants received tailored goal setting and tracking, skills

XSL•FO
RenderX

training, self-monitoring, and responsive health coaching from registered dietitians. Weight was measured at regular primary care visits and documented in the electronic health record. We compared the percentage of ≤3% weight gain in each arm at 24 months after randomization—our primary outcome—using individual empirical best linear unbiased predictors from the linear mixed-effects model. We used individual empirical best linear unbiased predictors from participants with at least 1 electronic health record weight documented within a 6-month window centered on the 24-month time point.

Results: We randomized 443 participants, of which 223 (50.3%) participants were allocated to the intervention arm. At baseline, participants had a mean BMI of 32.6 kg/m^2. Most participants were Latino or Hispanic (n=200, 45.1%) or non–Latino or Hispanic White (n=115, 26%). In total, 53% (n=235) of participants had at least 1 visit with weight measured in the primary time window. The intervention group had a higher proportion with ≤3% weight gain at 6 months (risk ratio=1.12, 95% CI 0.94-1.28; risk difference=9.5, 95% CI –4.5 to 16.4 percentage points). This difference attenuated to the null by 24 months (risk ratio=1.00, 95% CI 0.82-1.20; risk difference=0.2, 95% CI –12.1 to 11.0 percentage points).

Conclusions: In adults with overweight or obesity receiving primary care at a community health center, we did not find long-term evidence to support the dissemination of a digital health intervention for weight gain prevention.

Trial Registration: ClinicalTrials.gov NCT03003403; https://clinicaltrials.gov/study/NCT03003403

International Registered Report Identifier (IRRID): RR2-10.1186/s12889-019-6926-7

(J Med Internet Res 2024;26:e50330) doi:10.2196/50330

KEYWORDS

weight gain prevention; digital health; pragmatic clinical trial; primary care; health disparities; obesity; obese; prevalence; weight management; overweight; intervention

Introduction

The prevalence of obesity and its associated comorbidities continue to rise in the United States, and populations who are uninsured and from racial and ethnic minority groups continue to be disproportionately affected [1-8]. These populations also experience fewer clinically meaningful outcomes in most weight loss trials, suggesting a need for alternative intervention approaches to contend with obesity and its associated clinical outcomes [9-11]. Weight gain prevention presents a useful—and guidelines-adherent—strategy for individuals who experience barriers to, or have less success with, weight loss. Compared with weight loss, weight gain prevention can be achieved at lower treatment intensity and is well suited for delivery using electronic health technologies, yielding the potential to reach large, high-risk populations at low costs [12].

We previously demonstrated the efficacy of weight gain prevention interventions in a population of Black female individuals in the community health center (CHC) setting [12,13]. The intervention focused on creating a slight energy deficit sufficient to offset weight gain using medium-intensity digital health strategies. While efficacious, there were concerns about the intervention's suitability for implementation, as it required significant staff effort and used labor-intensive technologies for participant engagement. Given the often-limited weight management resources available to patients in primary care settings serving vulnerable patients, more frequently evaluating intervention with pragmatic designs may help inform the design of comprehensive obesity care delivered in primary care [14].

We report outcomes from Balance, a 2-arm, 12-month pragmatic randomized controlled trial of a digital weight gain prevention intervention delivered to patients receiving primary care within Piedmont Health Services Inc (Piedmont), a network of Federally Qualified Community Health Centers located in central North Carolina. We hypothesized that, compared with usual care, the intervention arm would have a greater proportion of participants with ≤3% weight gain at 24 months after randomization.

Methods

Study Design

Full details of the trial protocol, study population, Balance intervention, and recruitment processes have previously been published [14] and are summarized here. Balance participants were adults with overweight or obesity who received care at participating Piedmont CHCs. Participants were randomized to receive either (1) the 12-month Balance weight gain prevention intervention or (2) a healthy living usual care arm. There were no trial-specific follow-up visits conducted following randomization. Participant outcomes were collected from the Piedmont electronic health record (EHR) at 6, 12, 18, and 24 months after randomization.

Ethics Approval

Approvals for the trial were obtained by the Duke University institutional review board (2017-0738/D0479) and the Piedmont board in 2016. Verbal informed consent was provided from all participants prior to enrolling in the trial. The trial was registered on ClinicalTrials.gov (NCT03003403) on December 16, 2016. All data presented in this manuscript are deidentified. As Balance was a pragmatic trial with outcomes extracted from the Piedmont EHR and thus did not require participant study visits or follow-up surveys for evaluation, monetary compensation was not provided.

Study Setting and Sample

Congruent with the trial's pragmatic design, we sought to limit eligibility criteria to those that were fundamental to intervention implementation or evaluation or represented a significant threat

XSL•FO
RenderX

to patient safety. To be included, participants needed to be fluent in English or Spanish and have a BMI between 25 and 40 kg/m^2 (inclusive) and a weight of 380 pounds or less, as recorded by a provider in the Piedmont EHR within the previous 14 days. The weight eligibility criterion was based on the weight capacity of the connected scales used for intervention delivery. Participants also needed a mobile phone and service plan that could receive weekly trial-related text messages. Exclusion criteria included pregnancy within 12 months, giving birth within 6 months, or breastfeeding within 2 months; having prior weight loss or bariatric surgery; or planning to relocate outside the Piedmont service area during the trial period.

Trial Recruitment, Screening, and Randomization

Trial recruitment launched in February 2017. In-person recruitment was conducted through a joint effort by Piedmont staff and a Balance research staff member, with Piedmont staff identifying possibly eligible participants and referring them to research staff for questions and additional information. Due to slower-than-expected study recruitment, 4 months into the trial, a new recruitment method was added: an EHR query was used to identify patients who met basic criteria (eg, age and BMI) for Balance and had an upcoming appointment at Piedmont. These patients identified via the EHR query were mailed materials ahead of their appointments and presented trial information during visits to the CHC, if a research team member was present at their appointment time. Interested patients from either recruitment method were prompted to sign an authorization form to allow research staff to assess their medical record for eligibility—and later for evaluation purposes, if enrolled. Following completion of the authorization form, patients underwent additional phone screening to assess criteria not available in their medical records, such as willingness to receive text messages. After eligibility was established, research staff continued to conduct verbal informed consent, enrollment, randomization, and program orientation procedures. Additional details regarding recruitment and screening procedures were previously described [14,15]. Randomization took place concurrently with enrollment between February 2017 and December 2018 using block randomization with stratification by the patient's CHC within Piedmont Health, with equal allocation to each treatment group.

Healthy Living Usual Care Arm

Participants randomly assigned to the healthy living usual care arm received enhanced usual care. They received their standard primary care at Piedmont as well as 6 months of automated weekly text messages with healthy living information and printed materials adapted from the National Heart, Lung, and Blood Institute's *Aim for a Healthy Weight* [16].

Balance Weight Gain Prevention Arm

The goal of the Balance intervention was to prevent weight gain by encouraging behavioral changes that result in a slight energy deficit. Intervention participants received (1) tailored behavior change goals, (2) skills training materials and videos, (3) weekly self-monitoring prompts for behavioral goal tracking, (4) connected scales for weight self-monitoring, and (5) responsive health and weight coaching from trained registered dietitians.

Behavioral goals were individually assigned and tailored to participants using the interactive obesity treatment approach. The processes for interactive obesity treatment approach goal assignment for Balance have been described elsewhere [14] and have been used in previous primary care obesity treatment trials [13,17,18]. Briefly, participants were administered a simple survey at the start of the trial with a built-in algorithm that creates a net caloric deficit on the backend. For example, participants were asked to rate the frequency of consuming sugary drinks or walking 10,000 steps per day and their perceived confidence to change this behavior. Based on the results of the frequency of each health behavior and their self-efficacy to change it, participants received a set of 3 goals which changed every 8 weeks. They monitored their goal progress by responding to weekly interactive voice response calls and text messages. Participants then received automated, tailored feedback that described their progress, reinforced successes, and offered motivational strategies or short skills training tips.

Throughout the intervention, participants were also asked to weigh themselves daily on a connected scale, as previously described [14,19,20]. Data were transmitted through cellular networks to a database accessible to the research team. A rolling 7-day average of these weights was used to activate the responsive coaching from a Balance dietitian when a participant reached a designated threshold, or zone, of weight gain. If no weights were received across 1 intervention week, study staff would attempt to contact the participant via text or phone to re-engage or troubleshoot.

Data Collection

Data collection procedures were designed to maximize the use of data collected during routine primary care [21]. Participant sociodemographic characteristics were collected by a brief survey administered as part of the enrollment phone call. All other participant baseline and follow-up data were pulled directly from the EHR (GE Centricity CPS, version 12). Weight measurements in the EHR were recorded in pounds to the nearest 0.1 pound and converted to kilograms for analysis. Height was recorded in inches to the nearest 0.01 inch and converted to centimeters for analysis. BMI was recorded as weight/height (kg/m^2). EHR data were obtained on all participants between August 6, 2015, and April 30, 2021, including up to 18 months before enrollment through 30.5 months after enrollment.

For participants who became pregnant during the study, all clinical values were censored as of the beginning of their pregnancy. For pregnancies that occurred before enrollment, we censored clinical values from the beginning of pregnancy to 6 months after the end of pregnancy.

Statistical Methods

Baseline variables were summarized by the intervention arm as means and SDs for continuous variables and as counts and percentages for categorical variables. We used a 2-stage modeling strategy to answer our primary hypothesis. All analyses were intent-to-treat. For the first stage, we estimated the intervention effect on weight using a constrained longitudinal

XSL•FO
RenderX

linear mixed-effects model [22], using all available data from between 18 months before enrollment through 30.5 months after enrollment. We explored nonlinearities for the fixed effects of time and random effects of time in the models, guided by the Bayesian information criterion and likelihood ratio tests. The model included fixed effects for time and interactions between time segments and intervention, with a random intercept and random slopes for time. The covariance between random effects was modeled using an unstructured covariance matrix. Randomization was stratified by CHC; thus, we adjusted for CHC in the analyses but did not adjust for other variables. Using this linear mixed-effects model, we estimated mean weight change by intervention arm, and the difference between intervention arms in mean weight change, at 6, 12, 18, and 24 months.

Second, we compared the percentage of ≤3% weight gain—our primary outcome—in each arm at 24 months after randomization using individual empirical best linear unbiased predictors (EBLUPs) from the mixed model [23]. To ensure good predictions, we only used individual EBLUPs from participants with at least 1 EHR weight documented within a 6-month window centered on the 24-month time point (ie, between 21 and 27 months after enrollment). We compared the percentage of ≤3% weight gain in intervention and usual care arms using both a log-binomial model and linear risk model on the EBLUP output, in order to obtain both risk ratios and risk differences, which provide estimates of relative and absolute efficacy, respectively [24]. The extra variability induced by the EBLUPs being predicted from the model was taken into account using a resampling procedure, explained further in the Multimedia Appendices 1 and 2. Sensitivity to the 21- to 27-month window was assessed by obtaining predictions after expanding the window. We additionally used EBLUPs to compare the percentage of ≤3% weight gain in each arm at 6, 12, and 18 months after randomization.

Additional sensitivity analyses included adjusting the model for baseline variables imbalanced across arms, by weight measured in the primary time window ($P<.10$) and removing telehealth visits from the analysis. We also explored effect modification by adding interactions to the primary model with age tertiles; gender; BMI class; race and ethnicity; and preferred language (English or Spanish), each in a separate model. Additionally, in post hoc analyses we added the completion date of the study (before or after COVID-19) as an effect modifier to examine the potential impact of COVID-19 on the intervention effect.

Results

We randomized 443 participants, of which 223 (50.3%) participants were allocated to the intervention arm and 220 (49.7%) to the usual care arm. Participant demographics are summarized by intervention arm (Table 1). At baseline, participants had a mean BMI of 32.6 (SD 4.0) kg/m^2 and most participants were Latino or Hispanic (200/443, 45.1%) or non–Latino or Hispanic White (115/443, 26%).

XSL•FO
RenderX

Table 1. Balance baseline demographics.

Characteristics	Control (n=220)	Intervention (n=223)	Total (N=443)
Estimated baseline weight (in kg)[a], mean (SD)	86.8 (14.3)	85.2 (13.1)	86.0 (13.7)
Age (y), mean (SD)	46.8 (13.0)	48.5 (13.6)	47.6 (13.3)
BMI value (kg/m^2; EHR[b]), mean (SD)	32.9 (3.9)	32.4 (4.0)	32.6 (4.0)
BMI class (kg/m^2; EHR), n (%)			
25 to <30: Overweight	58 (26.4)	70 (31.4)	128 (28.9)
30 to <35: Class I obese	97 (44.1)	87 (39)	184 (41.5)
35 to <40: Class II obese	57 (25.9)	64 (28.7)	121 (27.3)
40+: Class III obese	8 (3.6)	2 (0.9)	10 (2.3)
Hours of sleep per 24 hours, mean (SD)	7.1 (1.4)	6.7 (1.3)	6.9 (1.3)
Patient Health Questionnaire-2 score, mean (SD)	0.9 (1.2)	1.0 (1.3)	1.0 (1.3)
Sex, n (%)			
Male	46 (20.9)	43 (19.3)	89 (20.1)
Female	173 (78.6)	179 (80.3)	352 (79.5)
Male to female transgender	1 (0.5)	1 (0.4)	2 (0.5)
Race or ethnicity, n (%)			
Hispanic (all races)	98 (44.5)	102 (45.7)	200 (45.1)
Non-Hispanic Black	53 (24.1)	53 (23.8)	106 (23.9)
Non-Hispanic other or unreported	15 (6.8)	7 (3.1)	22 (5)
Non-Hispanic White	54 (24.5)	61 (27.4)	115 (26)
Preferred language, n (%)			
English	145 (65.9)	141 (63.2)	286 (64.6)
Spanish	75 (34.1)	82 (36.8)	157 (35.4)
Education, n (%)			
Less than high school education	37 (16.8)	45 (20.2)	82 (18.5)
High school graduate	65 (29.5)	84 (37.7)	149 (33.6)
Some college, vocational degree, or associate's degree	88 (40)	68 (30.5)	156 (35.2)
College graduate or beyond	30 (13.6)	26 (11.7)	56 (12.6)
Community health center, n (%)			
Carrboro	111 (50.5)	113 (50.7)	224 (50.6)
Chapel Hill	28 (12.7)	27 (12.1)	55 (12.4)
Moncure	66 (30)	67 (30)	133 (30)
Prospect Hill	10 (4.5)	11 (4.9)	21 (4.7)
Siler City	5 (2.3)	5 (2.2)	10 (2.3)
Recruitment method, n (%)			
Via on site efforts only	123 (55.9)	127 (57)	250 (56.4)
On site after mailing	43 (19.5)	42 (18.8)	85 (19.2)
Via mail or off-site efforts only	54 (24.5)	54 (24.2)	108 (24.4)
Leisure-time physical activity, n (%)			
Refused	2 (0.9)	0 (0)	2 (0.5)
Do not know or not sure	2 (0.9)	1 (0.4)	3 (0.7)
No	91 (41.4)	86 (38.6)	177 (40)

XSL•FO
RenderX

Characteristics	Control (n=220)	Intervention (n=223)	Total (N=443)
Yes	125 (56.8)	136 (61)	261 (58.9)

[a]Estimated from primary linear mixed-effects model.

[b]EHR: electronic health record.

Among our sample, 53% (n=235) of participants had at least 1 visit with weight measured within the primary time window (21 to 27 months; see Figure 1). Participants with visits with weights measured dropped off dramatically during the COVID-19 pandemic (281/443, 63% in April-September 2019 vs 94/443, 21% in April-September 2020; 214/443, 48% in October 2019-February 2020 vs 50/443, 11% in October 2020-February 2021; see Table S1 in Multimedia Appendix 1). This reduced the number of visits in the primary outcome time window since 50% (n=220) of participants had not yet completed all 27 months of follow-up by the start of the pandemic (March 15, 2020). However, there was no significant difference between the arms in the numbers of visits with weights measured in the 21- to 27-month window (119/220, 54% in control vs 116/220, 52% in intervention with a weight measurement in this window; see Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Balance CONSORT (Consolidated Standards of Reporting Trials) flow diagram. For a higher-resolution version of this figure, see Multimedia Appendix 3.

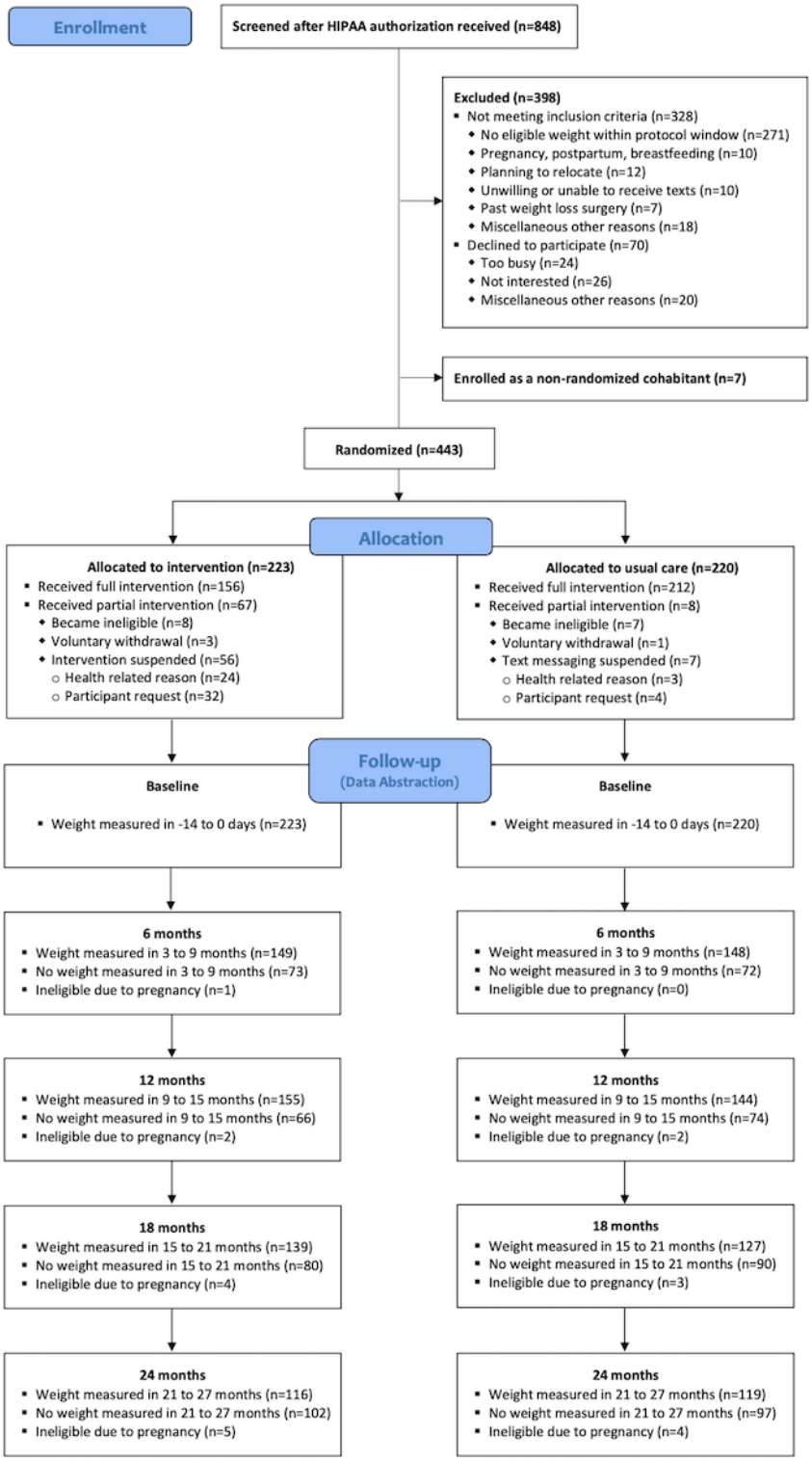

The linear spline model with knots at 6 months before enrollment; enrollment; and 6, 12, and 18 months after enrollment had the lowest Bayesian information criterion. Compared with the usual care arm, the intervention arm had greater mean weight change at 6 months (–1.2, 95% CI –2.1 to –0.2 kg); however, this estimate was attenuated by 24-months (–0.1, 95% CI –1.2 to 1.0 kg; see Figure 2). The intervention arm had a higher proportion with ≤3% weight gain at 6 months (risk ratio=1.12, 95% CI 0.94-1.28; risk difference=9.5, 95% CI –4.5 to 16.4 percentage points), which also attenuated to the null by 24 months (risk ratio=1.00, 95% CI 0.82-1.20; risk difference=0.2, 95% CI –12.1 to 11.0 percentage points; see Figure 3). Changing the window size around the 24-month time point did not appreciably affect these results (see Figures S1 and S2 in Multimedia Appendix 1). Further descriptive statistics

XSL•FO
RenderX

regarding ≤3% weight gain are presented in Table S3 in Multimedia Appendix 1.

Figure 2. Predicted mean weight change (in kilograms) from baseline in intervention and control separately (panel A), and difference in weight change after baseline comparing intervention to control (panel B). Predictions are from the linear mixed-effects model. In panel B, negative values indicate that the intervention group lost more (or gained less) weight on average than the control group.

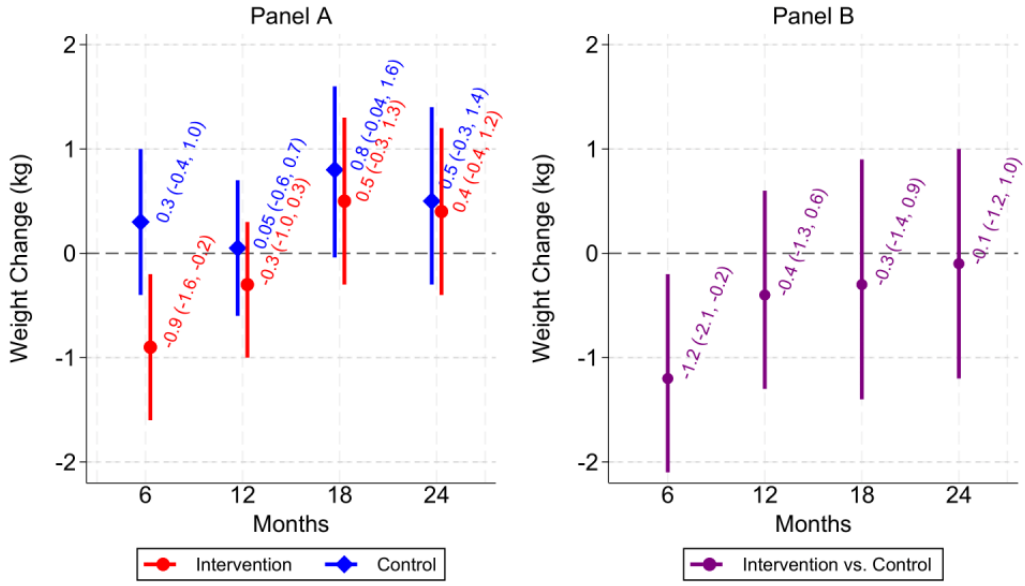

Figure 3. Risk ratio (panel A) and risk difference (panel B) comparing intervention and control on ≤3% weight gain. In panel A, values above 1 indicate that the intervention group had a greater relative probability of ≤3% weight gain versus control. In panel A, values above 0 indicate that the intervention group had a greater absolute probability of ≤3% weight gain versus control.

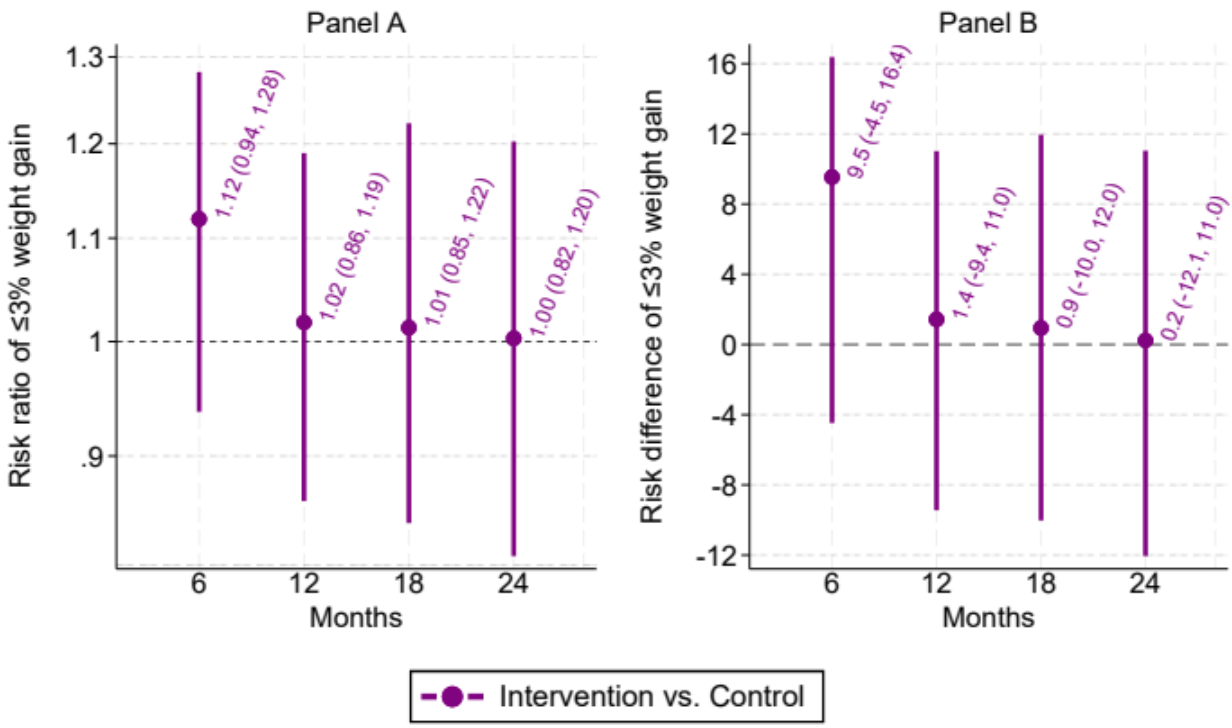

Weight measurements were collected at 75 total telehealth visits during the pandemic. After we removed telehealth visits from the model, there were no significant changes to the effect estimates. However, at 18 and 24 months, the effect attenuated toward the null (see Figure S3 in Multimedia Appendix 1).

Table S4 in Multimedia Appendix 1 displays baseline demographics for those with weight measured in the 21- to 27-month window. No variables were imbalanced between those with weight measured in this window versus those without weight measured in this window (P<.10); thus, no additional analyses were performed.

Figures S4-S7 in Multimedia Appendix 1 display the results of effect modification analysis. At 6 months, there is evidence of a stronger intervention effect in the 44-57 years age range

XSL•FO
RenderX

relative to the ≤43 years age range. No other subgroups across time were significantly different with respect to the intervention effect, including post hoc subgroups based on 24-month completion relative to COVID-19 pandemic onset.

Discussion

Principal Findings

We and others have demonstrated the efficacy of weight gain prevention interventions in adults affected by obesity [13,25]. However, in Balance, we did not find long-term evidence to support the dissemination of a digital weight gain prevention intervention to patients receiving primary care at CHCs. Compared with the usual care arm, the intervention arm experienced greater weight loss at 6 months, yet this difference was not clinically meaningful and had attenuated by 24 months. Further, we did not observe a difference between the 2 arms in the proportion of individuals who stayed within 3% of their baseline weight.

Comparison to Previous Work

Unlike our previous trials using similar approaches resulting in significant weight gain prevention or weight loss [13,18], Balance was designed to use a high level of pragmatism to deliver and evaluate real-world intervention effectiveness. This aligned with the key domains of the Pragmatic-Explanatory Continuum Indicator Summary index, including limited exclusion criteria, nonrestrictive control group, no formal follow-up data collection, and extraction of outcome data directly from the EHR [21]. Pragmatic trial designs vary considerably in the aforementioned domains and these characteristics can assist with contextualizing study findings. For instance, we did not restrict the usual care arm in this trial. As a result, some of these patients may have received medical nutrition therapy or other weight management services from Piedmont registered dietitians or community partners. These efforts may have resulted in the usual care arm experiencing better weight outcomes than expected among a restricted control group, ultimately diminishing trial outcomes.

As described in the methods section and more thoroughly in our protocol paper, the Balance study had a responsive weight and goal coaching protocol. There were 3 zones a participant would be categorized in, depending on their weight change, in which they may have received minimal to no coaching. This differs from our previous weight gain prevention trial that successfully prevented weight gain, in which participants received a standardized coaching protocol that included 20-minute monthly phone calls for 12 months, regardless of weight status [12,13]. It is possible that participants of the Balance intervention would have benefited from this standardized coaching protocol or a higher-intensity intervention.

To our knowledge, no examples exist of other pragmatic weight gain prevention trials in primary care settings among medically vulnerable populations with which to compare our findings. However, there have been several pragmatic weight loss trials implemented in the primary care setting. For example, the PROPEL and REPOWER trials were large, pragmatic randomized weight loss trials implemented in primary care clinics working with medically vulnerable populations and tested a total of 5 weight loss interventions with varying levels of pragmatism [26]. Both trials' intervention arms observed significant weight loss at 6-months. Yet, not all intervention arms were as successful at producing weight loss, and notably, the arm designed with the highest level of pragmatism observed the least sustained weight loss at 24 months. Along with our study, these findings highlight the challenges of long-term weight management, particularly in pragmatic trials. Future studies might benefit from including additional support for changing motivation, group support, and targeted problem-solving strategies [27,28] to overcome external challenges.

Limitations and Considerations for Future Research

This study had limitations that should be noted. A potential challenge facing pragmatic weight management trials is the quality of clinic-measured weights. We previously assessed concordance between EHR and research-collected weight data in our earlier weight management interventions and did not find differences in baseline weight, changes in weight from baseline, or differences between intervention and controls in weight change [29]. However, estimates based on EHR weights were generally more variable than weights directly collected by the study team. Another more recent trial demonstrated that weight loss differences observed at 24 months were smaller when using EHR data [30]. The authors offered several explanations for this finding, including the larger variability in follow-up time and the less restrictive procedures for in-clinic weight measurement, such as those observed in Balance. Because we did not collect a secondary measure of weight, we are unable to assess if more controlled weight measurement procedures would have improved our ability to detect differences in weight gain between groups.

We expect that the timeframe of implementation posed challenges to identifying the effects of the Balance intervention on weight outcomes. Namely, the COVID-19 pandemic began during the implementation of the Balance intervention and our 24-month follow-up period. This decreased the number of in-person visits being conducted at the CHCs and, ultimately, impacted follow-up data collection on nearly half of our participants. COVID-19 could have also impacted other weight-related behaviors and risk factors among participants, including physical activity [31,32], food access and dietary quality [33-35], and mental health stressors [36,37]. In addition to COVID-19, there were several other events during the intervention and follow-up timeframe that could have impacted data collection, including 2 major North Carolina hurricanes and immigration reform resulting in increased deportation of undocumented individuals. These factors likely impacted patient appointment attendance and thus the availability of baseline and follow-up EHR data. As such, these events may have affected the precision of the intervention effect estimates for outcomes at 24 months. However, it is unlikely the intervention effect estimates themselves were biased by these events, as there was no difference between the number of visits with weights measured by treatment arms.

XSL•FO
RenderX

Additionally, some components of trial implementation did not align with a pragmatic approach, such as using research staff to conduct recruitment at the health centers. There were also institutional requirements for research compliance that would not be necessary if implemented within a real-world setting. For instance, a Health Insurance Portability and Accountability Act (HIPAA) authorization form was required from each participant before screening could occur, which posed challenges to recruitment ease and timing, as highlighted in our previous article [15]. However, if Balance were disseminated further within Piedmont or other similar CHCs without the inclusion of external research partners, these specific requirements would likely have been lessened or eliminated, thereby decreasing the amount of time required to enroll in the program. Finally, as discussed above, several events, such as COVID-19, occurred during study implementation that impacted data collection. However, as described above, we expect this did not bias effect estimates.

Strengths

Despite these limitations, there are many notable strengths. The Balance trial was designed with a high level of pragmatism, with limited criteria for eligibility and no study-specific data collection or follow-up procedures, and enrolled a diverse sample of participants. For these reasons, the Balance trial also provides results that can inform decisions in real-world care settings for the management of obesity. Although the intervention did not have a significant impact on weight gain prevention, it is possible that participants adopted healthy dietary and physical activity behaviors that positively impact health and chronic disease management [38]; however, these outcomes were not assessed in this study. Future assessment is warranted given long-term healthy lifestyle changes are possible independent of weight change after participation in a weight gain prevention program [39].

To the best of our knowledge, Balance was also the first weight gain prevention trial to be implemented with a pragmatic design in a primary care setting serving a medically vulnerable population, serving as a strong example of successful partnerships between research teams and CHCs. This sentiment was echoed by Piedmont providers and staff in a qualitative evaluation of Balance [40], in which they stated that early buy-in from providers and staff, respect for the patients and health care setting by the research partnership, and filling a treatment gap for patients all contributed to successful trial implementation. Combined with the trial outcomes, these findings can be used to inform future programs and interventions to be tested and implemented in the CHC setting.

Conclusions

Balance did not exhibit a positive impact on long-term weight maintenance, yet it provides important information for future pragmatic trials in the primary care setting. Future trials might benefit from including an assessment of the comparator group's engagement in weight management behaviors and collecting a secondary weight measurement. Researchers may also consider measuring other outcomes associated with chronic disease prevention that are independent of weight, such as changes in diet quality, physical activity, and psychological health.

Acknowledgments

This trial was funded by the National Institute of Diabetes and Digestive and Kidney Diseases (R18DK109518). The funder had no role in study design, data collection, data analysis, and interpretation of data, in the writing of the report, and in the decision to submit this article for publication. The authors wish to express our deepest gratitude to the administration and staff at Piedmont Health for their continued collaboration and support. In particular, the authors would like to thank Brian Toomey, Jennifer Cunningham, Erica Boshnack, Kristen Norton, Caitlyn Faul, Harriett Burns, Tom Hadley, and all the providers and staff at the participating community health centers. The authors are also grateful to Dori Steinberg, Brian C Batch, and all other Duke staff and research assistants who provided project assistance. Last, the authors would like to especially thank the Piedmont Health patients who participated in Balance.

Conflicts of Interest

GGB is on the scientific advisory board for WeightWatchers and Wondr Health. EAF has past relationships with WeightWatchers, Jenny Craig, and several pharmaceutical companies in the obesity space but has no current industry relationships or financial interests in companies providing weight-loss products or services. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Estimating variance for the empirical best linear unbiased predictions (EBLUPs).
[PDF File (Adobe PDF File), 1700 KB - jmir_v26i1e50330_app1.pdf]

Multimedia Appendix 2
CONSORT E-HEALTH checklist.
[PDF File (Adobe PDF File), 3511 KB - jmir_v26i1e50330_app2.pdf]

Multimedia Appendix 3
Balance CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

XSL•FO
RenderX

[PNG File , 326 KB - jmir_v26i1e50330_app3.png]

References

1. Ard J. Obesity in the US: what is the best role for primary care? BMJ 2015;350:g7846. [doi: 10.1136/bmj.g7846] [Medline: 25656059]
2. Ogden CL, Carroll MD, Flegal KM. Prevalence of obesity in the United States. JAMA 2014;312(2):189-190. [doi: 10.1001/jama.2014.6228] [Medline: 25005661]
3. Flegal KM, Carroll MD, Kit BK, Ogden CL. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. JAMA 2012;307(5):491-497 [FREE Full text] [doi: 10.1001/jama.2012.39] [Medline: 22253363]
4. Yang L, Colditz GA. Prevalence of overweight and obesity in the United States, 2007-2012. JAMA Intern Med 2015;175(8):1412-1413 [FREE Full text] [doi: 10.1001/jamainternmed.2015.2405] [Medline: 26098405]
5. Ogden CL, Lamb MM, Carroll MD, Flegal KM. Obesity and socioeconomic status in adults: United States, 2005-2008. NCHS Data Brief 2010(50):1-8 [FREE Full text] [Medline: 21211165]
6. Chow EA, Foster H, Gonzalez V, McIver L. The disparate impact of diabetes on racial/ethnic minority populations. Clin Diabetes 2012;30(3):130-133 [FREE Full text] [doi: 10.2337/diaclin.30.3.130]
7. Rodriguez F, Ferdinand KC. Hypertension in minority populations: new guidelines and emerging concepts. Adv Chronic Kidney Dis 2015;22(2):145-153 [FREE Full text] [doi: 10.1053/j.ackd.2014.08.004] [Medline: 25704352]
8. Befort CA, Nazir N, Perri MG. Prevalence of obesity among adults from rural and urban areas of the United States: findings from NHANES (2005-2008). J Rural Health 2012;28(4):392-397 [FREE Full text] [doi: 10.1111/j.1748-0361.2012.00411.x] [Medline: 23083085]
9. Osei-Assibey G, Kyrou I, Adi Y, Kumar S, Matyka K. Dietary and lifestyle interventions for weight management in adults from minority ethnic/non-white groups: a systematic review. Obes Rev 2010;11(11):769-776. [doi: 10.1111/j.1467-789X.2009.00695.x] [Medline: 20059708]
10. Yancey AK, Kumanyika SK, Ponce NA, McCarthy WJ, Fielding JE, Leslie JP, et al. Population-based interventions engaging communities of color in healthy eating and active living: a review. Prev Chronic Dis 2004;1(1):A09 [FREE Full text] [Medline: 15634371]
11. Wadden TA, Butryn ML, Hong PS, Tsai AG. Behavioral treatment of obesity in patients encountered in primary care settings: a systematic review. JAMA 2014;312(17):1779-1791 [FREE Full text] [doi: 10.1001/jama.2014.14173] [Medline: 25369490]
12. Foley P, Levine E, Askew S, Puleo E, Whiteley J, Batch B, et al. Weight gain prevention among Black women in the rural community health center setting: the shape program. BMC Public Health 2012;12:305 [FREE Full text] [doi: 10.1186/1471-2458-12-305] [Medline: 22537222]
13. Bennett GG, Foley P, Levine E, Whiteley J, Askew S, Steinberg DM, et al. Behavioral treatment for weight gain prevention among Black women in primary care practice: a randomized clinical trial. JAMA Intern Med 2013;173(19):1770-1777 [FREE Full text] [doi: 10.1001/jamainternmed.2013.9263] [Medline: 23979005]
14. Berger MB, Steinberg DM, Askew S, Gallis JA, Treadway CC, Egger JR, et al. The balance protocol: a pragmatic weight gain prevention randomized controlled trial for medically vulnerable patients within primary care. BMC Public Health 2019;19(1):596 [FREE Full text] [doi: 10.1186/s12889-019-6926-7] [Medline: 31101037]
15. Miller HN, Berger MB, Askew S, Kay MC, Chisholm M, Sirdeshmukh G, et al. Recruitment of diverse community health center patients in a pragmatic weight gain prevention trial. J Clin Transl Sci 2023;7(1):e22 [FREE Full text] [doi: 10.1017/cts.2022.475] [Medline: 36755547]
16. Aim for a healthy weight. NIH 05-5213. National Institutes of Health. 2005. URL: https://www.nhlbi.nih.gov/resources/aim-healthy-weight-patient-booklet [accessed 2024-03-05]
17. Bennett GG, Warner ET, Glasgow RE, Askew S, Goldman J, Ritzwoller DP, et al. Obesity treatment for socioeconomically disadvantaged patients in primary care practice. Arch Intern Med 2012;172(7):565-574 [FREE Full text] [doi: 10.1001/archinternmed.2012.1] [Medline: 22412073]
18. Bennett GG, Steinberg D, Askew S, Levine E, Foley P, Batch BC, et al. Effectiveness of an app and provider counseling for obesity treatment in primary care. Am J Prev Med 2018;55(6):777-786 [FREE Full text] [doi: 10.1016/j.amepre.2018.07.005] [Medline: 30361140]
19. Foley P, Steinberg D, Levine E, Askew S, Batch BC, Puleo EM, et al. Track: a randomized controlled trial of a digital health obesity treatment intervention for medically vulnerable primary care patients. Contemp Clin Trials 2016;48:12-20 [FREE Full text] [doi: 10.1016/j.cct.2016.03.006] [Medline: 26995281]
20. Steinberg DM, Tate DF, Bennett GG, Ennett S, Samuel-Hodge C, Ward DS. The efficacy of a daily self-weighing weight loss intervention using smart scales and e-mail. Obesity (Silver Spring) 2013;21(9):1789-1797 [FREE Full text] [doi: 10.1002/oby.20396] [Medline: 23512320]
21. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A Pragmatic-Explanatory Continuum Indicator Summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009;62(5):464-475 [FREE Full text] [doi: 10.1016/j.jclinepi.2008.12.011] [Medline: 19348971]

XSL•FO
RenderX

22. Lu K. On efficiency of constrained longitudinal data analysis versus longitudinal analysis of covariance. Biometrics 2010;66(3):891-896 [FREE Full text] [doi: 10.1111/j.1541-0420.2009.01332.x] [Medline: 19764951]
23. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis, 2nd Edition. Hoboken, NJ: John Wiley & Sons; 2012.
24. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010;63(8):e1-e37 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.03.004] [Medline: 20346624]
25. Martin JC, Awoke MA, Misso ML, Moran LJ, Harrison CL. Preventing weight gain in adults: a systematic review and meta-analysis of randomized controlled trials. Obes Rev 2021;22(10):e13280. [doi: 10.1111/obr.13280] [Medline: 34028958]
26. Katzmarzyk PT, Apolzan JW, Gajewski B, Johnson WD, Martin CK, Newton RL, et al. Weight loss in primary care: a pooled analysis of two pragmatic cluster-randomized trials. Obesity (Silver Spring) 2021;29(12):2044-2054 [FREE Full text] [doi: 10.1002/oby.23292] [Medline: 34714976]
27. Greaves C, Poltawski L, Garside R, Briscoe S. Understanding the challenge of weight loss maintenance: a systematic review and synthesis of qualitative research on weight loss maintenance. Health Psychol Rev 2017;11(2):145-163 [FREE Full text] [doi: 10.1080/17437199.2017.1299583] [Medline: 28281891]
28. Spreckley M, Seidell J, Halberstadt J. Perspectives into the experience of successful, substantial long-term weight-loss maintenance: a systematic review. Int J Qual Stud Health Well-being 2021;16(1):1862481 [FREE Full text] [doi: 10.1080/17482631.2020.1862481] [Medline: 33455563]
29. Gallis JA, Kusibab K, Egger JR, Olsen MK, Askew S, Steinberg DM, et al. Can electronic health records validly estimate the effects of health system interventions aimed at controlling body weight? Obesity (Silver Spring) 2020;28(11):2107-2115 [FREE Full text] [doi: 10.1002/oby.22958] [Medline: 32985131]
30. Katzmarzyk PT, Mire EF, Martin CK, Newton RL, Apolzan JW, Price-Haywood EG, et al. Comparison of weight loss data collected by research technicians versus electronic medical records: the PROPEL trial. Int J Obes (Lond) 2022;46(8):1456-1462 [FREE Full text] [doi: 10.1038/s41366-022-01129-9] [Medline: 35523955]
31. Park AH, Zhong S, Yang H, Jeong J, Lee C. Impact of COVID-19 on physical activity: a rapid review. J Glob Health 2022;12:05003 [FREE Full text] [doi: 10.7189/jogh.12.05003] [Medline: 35493780]
32. Harrison E, Monroe-Lord L, Carson AD, Jean-Baptiste AM, Phoenix J, Jackson P, et al. COVID-19 pandemic-related changes in wellness behavior among older Americans. BMC Public Health 2021;21(1):755 [FREE Full text] [doi: 10.1186/s12889-021-10825-6] [Medline: 33874931]
33. Picchioni F, Goulao LF, Roberfroid D. The impact of COVID-19 on diet quality, food security and nutrition in low and middle income countries: a systematic review of the evidence. Clin Nutr 2022;41(12):2955-2964 [FREE Full text] [doi: 10.1016/j.clnu.2021.08.015] [Medline: 34535329]
34. Lee MM, Poole MK, Zack RM, Fiechtner L, Rimm EB, Kenney EL. Food insecurity and the role of food assistance programs in supporting diet quality during the COVID-19 pandemic in Massachusetts. Front Nutr 2022;9:1007177 [FREE Full text] [doi: 10.3389/fnut.2022.1007177] [Medline: 36687676]
35. Kent K, Alston L, Murray S, Honeychurch B, Visentin D. The impact of the COVID-19 pandemic on rural food security in high income countries: a systematic literature review. Int J Environ Res Public Health 2022;19(6):3235 [FREE Full text] [doi: 10.3390/ijerph19063235] [Medline: 35328924]
36. Xiong J, Lipsitz O, Nasri F, Lui LMW, Gill H, Phan L, et al. Impact of COVID-19 pandemic on mental health in the general population: a systematic review. J Affect Disord 2020;277:55-64 [FREE Full text] [doi: 10.1016/j.jad.2020.08.001] [Medline: 32799105]
37. Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 march 2022. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [accessed 2024-03-05]
38. Gaesser GA, Angadi SS, Sawyer BJ. Exercise and diet, independent of weight loss, improve cardiometabolic risk profile in overweight and obese individuals. Phys Sportsmed 2011;39(2):87-97. [doi: 10.3810/psm.2011.05.1898] [Medline: 21673488]
39. Driehuis F, Barte JCM, Ter Bogt NCW, Beltman FW, Smit AJ, van der Meer K, et al. Maintenance of lifestyle changes: 3-year results of the Groningen overweight and lifestyle study. Patient Educ Couns 2012;88(2):249-255 [FREE Full text] [doi: 10.1016/j.pec.2012.03.017] [Medline: 22560253]
40. Berger MB, Chisholm M, Miller HN, Askew S, Kay MC, Bennett GG. "We bleed for our community:" a qualitative exploration of the implementation of a pragmatic weight gain prevention trial from the perspectives of community health center professionals. BMC Public Health 2023;23(1):695 [FREE Full text] [doi: 10.1186/s12889-023-15574-2] [Medline: 37060053]

Abbreviations

CHC: community health center
EBLUP: empirical best linear unbiased predictor

XSL•FO
RenderX

EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act

Edited by G Greco; submitted 17.07.23; peer-reviewed by R Newton, A Bucher; comments to author 07.12.23; revised version received 07.02.24; accepted 26.02.24; published 28.03.24.

Please cite as:
Miller HN, Gallis JA, Berger MB, Askew S, Egger JR, Kay MC, Finkelstein EA, de Leon M, DeVries A, Brewer A, Holder MG, Bennett GG
Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial
J Med Internet Res 2024;26:e50330
URL: https://www.jmir.org/2024/1/e50330
doi:10.2196/50330
PMID:38416574

©Hailey N Miller, John A Gallis, Miriam B Berger, Sandy Askew, Joseph R Egger, Melissa C Kay, Eric Andrew Finkelstein, Mia de Leon, Abigail DeVries, Ashley Brewer, Marni Gwyther Holder, Gary G Bennett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of a Web-Based Individual Coping and Alcohol Intervention Program for Children of Parents With Alcohol Use Problems: Randomized Controlled Trial

Håkan Wall[1], PhD; Helena Hansson[2], PhD; Ulla Zetterlind[3], PhD; Pia Kvillemo[1], PhD; Tobias H Elgán[1], PhD

[1]Stockholm Prevents Alcohol and Drug Problems, Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm Health Care Services, Stockholm, Sweden

[2]School of Social Work, Faculty of Social Sciences, Lund University, Lund, Sweden

[3]Clinical Health Promotion Centre, Department of Health Sciences, Lund University, Lund, Sweden

Corresponding Author:
Tobias H Elgán, PhD
Stockholm Prevents Alcohol and Drug Problems, Centre for Psychiatry Research
Department of Clinical Neuroscience
Karolinska Institutet, & Stockholm Health Care Services
Norra Stationsgatan 69
Stockholm, 11364
Sweden
Phone: 46 700011003
Email: tobias.elgan@ki.se

Abstract

Background: Children whose parents have alcohol use problems are at an increased risk of several negative consequences, such as poor school performance, an earlier onset of substance use, and poor mental health. Many would benefit from support programs, but the figures reveal that only a small proportion is reached by existing support. Digital interventions can provide readily accessible support and potentially reach a large number of children. Research on digital interventions aimed at this target group is scarce. We have developed a novel digital therapist-assisted self-management intervention targeting adolescents whose parents had alcohol use problems. This program aims to strengthen coping behaviors, improve mental health, and decrease alcohol consumption in adolescents.

Objective: This study aims to examine the effectiveness of a novel web-based therapist-assisted self-management intervention for adolescents whose parents have alcohol use problems.

Methods: Participants were recruited on the internet from social media and websites containing health-related information about adolescents. Possible participants were screened using the short version of the Children of Alcoholics Screening Test-6. Eligible participants were randomly allocated to either the intervention group (n=101) or the waitlist control group (n=103), and they were unblinded to the condition. The assessments, all self-assessed, consisted of a baseline and 2 follow-ups after 2 and 6 months. The primary outcome was the Coping With Parents Abuse Questionnaire (CPAQ), and secondary outcomes were the Center for Epidemiological Studies Depression Scale, Alcohol Use Disorders Identification Test (AUDIT-C), and Ladder of Life (LoL).

Results: For the primary outcome, CPAQ, a small but inconclusive treatment effect was observed (Cohen d=–0.05 at both follow-up time points). The intervention group scored 38% and 46% lower than the control group on the continuous part of the AUDIT-C at the 2- and 6-month follow-up, respectively. All other between-group comparisons were inconclusive at either follow-up time point. Adherence was low, as only 24% (24/101) of the participants in the intervention group completed the intervention.

Conclusions: The findings were inconclusive for the primary outcome but demonstrate that a digital therapist-assisted self-management intervention may contribute to a reduction in alcohol consumption. These results highlight the potential for digital interventions to reach a vulnerable, hard-to-reach group of adolescents but underscore the need to develop more engaging support interventions to increase adherence.

Trial Registration: ISRCTN Registry ISRCTN41545712; https://www.isrctn.com/ISRCTN41545712?q=ISRCTN41545712

International Registered Report Identifier (IRRID): RR2-10.1186/1471-2458-12-35

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52118) doi:10.2196/52118

KEYWORDS

adolescent; adolescents; alcoholic; alcoholics; CBT; children of impaired parents; cognitive behavioral therapy; coping; digital intervention; mental health; randomized controlled trial; RCT; self-management; substance abuse; substance use; therapist; web-based intervention; youth

Introduction

Children who grow up with parents who have substance use problems or disorders face extraordinary challenges. Approximately 20% of all children have parents with alcohol problems [1-5], while approximately 5% have parents with alcohol use disorders [4,6,7]. Children growing up with parental substance abuse are at an increased risk of several negative outcomes, such as psychiatric morbidity [8-12]; poor intellectual, cognitive, and academic achievement [13-15]; domestic physical abuse [16]; and early drinking onset and the development of substance use problems [9,17,18]. Thus, children exposed to parental substance abuse comprise a target group for selective interventions and prevention strategies [19-22].

In Sweden, municipalities account for most of the support offered to these children. An annual survey by the junior association of the Swedish branch of Movendi International (ie, an international temperance movement) reported that 97% of all municipalities provided support resources [23]. However, estimates from the same survey showed that approximately 2% of the children in the target group received support. Hence, an overwhelming majority never receives support, mainly because of difficulties in identifying and attracting them to intervention programs [22,24].

The internet has become an appealing way to reach and support a large number of people [25,26]. Web-based interventions seem particularly attractive to adolescents, as they generally use digital technology and social media. Furthermore, research has shown that adolescents regard the internet as inviting because it is a readily accessible, anonymous way of seeking help [27]. Web-based interventions can reduce the stigma associated with face-to-face consultations in health care settings [28], and young people appreciate the flexibility of completing web-based sessions to fit their own schedules [29]. The positive effects of web-based interventions have been detected across a broad range of conditions. A recent review by Hedman-Lagerlöf et al [30] concluded that therapist-supported internet-based cognitive behavioral therapy for adults yielded similar effects as face-to-face therapy. To date, most web-based interventions have been designed for adults. Although the number of web-based interventions targeting children or adolescents is increasing [25,31-33], the number of digital interventions aimed at children of substance-abusing parents is still scarce [22,34-38]. Those described in the literature, however, all have in common that they are quite extensive, with a duration over several weeks, and a brief digital intervention could complement these more extended interventions. For instance, our research group initiated a study on a web-based group chat for 15- to 25-year-old individuals who have parents with mental illness or substance use problems [35]. The duration of the program is 8 weeks, and it is a translated version of a program from the Netherlands [34], which has been shown to have inconclusive treatment effects [39]. In Sweden, 2 other programs with inconclusive treatment effects have been tested that target significant others and their children [37,38]. Finally, a digital intervention developed in Australia for 18- to 25-year-old individuals with parents with mental illness or substance use disorder [36] was tested in a pilot study demonstrating positive findings [40].

To meet the need for a brief, web-based intervention that targets adolescents having parents with alcohol problems and build on the evidence base of digital interventions targeting this vulnerable group, we developed a novel internet-delivered therapist-assisted self-management intervention called “Alcohol and Coping.” Our program originated from a manual-based face-to-face intervention called the “Individual Coping and Alcohol Intervention Program” (ICAIP) [41,42]. Previous studies on both the ICAIP, which aimed at college students having parents with alcohol problems, and a coping skills intervention program, which aimed at spouses of partners with alcohol dependency [43], have demonstrated positive effects regarding decreased alcohol consumption and improved mental health and coping behaviors [41-44]. Furthermore, the results from these studies underscore the importance of improving coping skills [42,44]. Among college students, those who received a combination of coping skills and an alcohol intervention program had better long-term outcomes [42].

The aim of this study was to test the effectiveness of Alcohol and Coping among a sample of adolescents aged 15-19 years with at least 1 parent with alcohol use problems. We hypothesized that the intervention group would be superior to the control group in improving coping skills. Secondary research questions concerned the participants’ improvement in (1) depression, (2) alcohol consumption, and (3) quality of life.

Methods

Overview

This study was a parallel-group randomized controlled trial in which participants were randomized to either the intervention or waitlist control group in a 1:1 allocation ratio. The trial design is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram depicting study design and participant flow. ITT: intention-to-treat.

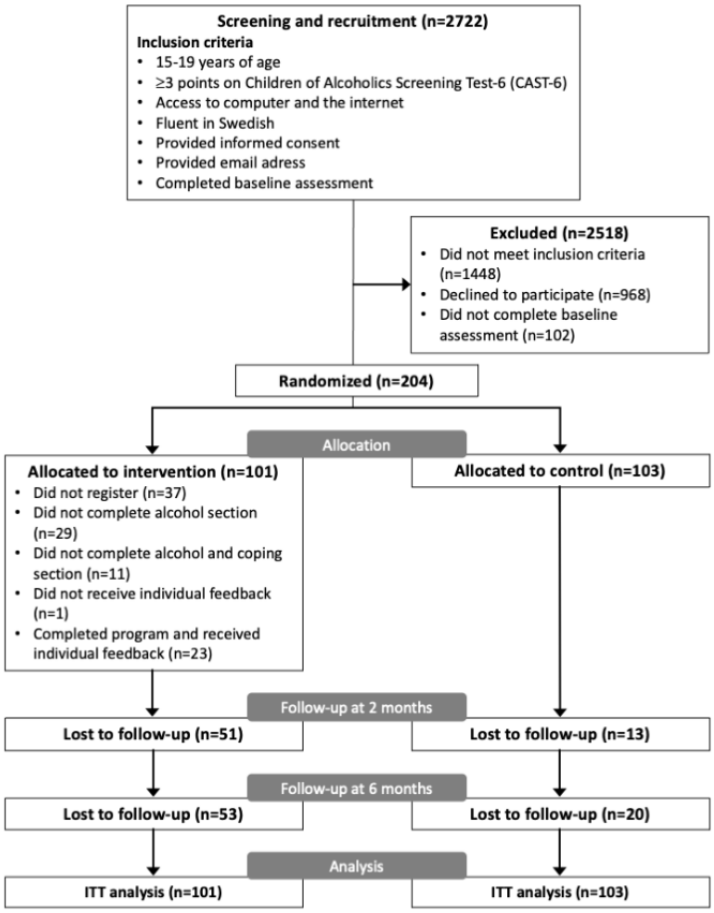

Recruitment and Screening

The participants were recruited from August 2012 to December 2013 through advertisements on social media (Facebook). The advertisements targeted individuals aged 15-19 years with Facebook accounts. Participants were recruited on the internet through advertisements on websites containing health-related information about adolescents. The advertisements included the text, “Do your parents drink too much? Participate in a study.” The advertisement contained an invitation to perform a web-based, self-assessed screening procedure. In addition to questions about age and sex, participants were screened for having parents with alcohol problems using the short version of the Children of Alcoholics Screening Test-6 (CAST-6), developed from a 30-item original version [45]. The CAST-6 is a 6-item true-false measure designed to assess whether participants perceive their parents’ alcohol consumption to be problematic. The CAST-6 has demonstrated high internal consistency (r=0.92-0.94), test-retest reliability (r=0.94), and high validity as compared to the 30-item version (r=0.93) using the recommended threshold score of 3 or higher [45,46]. We previously translated the CAST-6 into Swedish and validated the translated version among 1450 adolescents, showing good internal consistency (α=.88), excellent test-retest reliability (intraclass correlation coefficient=0.93), and loading into 1 latent factor [47]. Additional inclusion criteria included having access to a computer and the internet and being sufficiently fluent in Swedish. Participants were excluded from the study and were referred to appropriate care if there were indications of either suicidal or self-inflicted harmful behaviors. Individuals eligible for inclusion received further information about the study and were asked to provide consent to participate by providing an email address.

Data Collection and Measures

All assessments were administered through email invitations containing a hyperlink to the web-based self-reported assessments. Up to 3 reminders were sent through email at 5, 10, and 15 days after the first invitation. A baseline assessment (t_0) was collected before randomization, and follow-up assessments were conducted at 2 and 6 months (t_1 and t_2, respectively) after the initial assessment.

Participants were asked for age, sex, whether they lived with a parent (mother and father, mother or father, mother or father and stepparent, or alternate between mother and father), where their parents were born (Sweden or a Nordic country excluding Sweden or outside of the Nordic countries), parental status (employed, student, on parental leave, or unemployed), and any previous or present participation in support activities for children having parents with alcohol use problems. The primary outcome was coping, measured using the Coping With Parents Abuse Questionnaire (CPAQ) based on the Coping Behavior Scale developed by Orford et al [48]. Secondary outcomes were the Center for Epidemiological Studies Depression Scale (CES-DC) [49], the 3-question Alcohol Use Disorders Identification Test (AUDIT-C) [50], and the Ladder of Life (LoL), which measures the overall quality of life by asking about the participants’ past,

XSL•FO
RenderX

present, and future ratings of their overall life satisfaction [50]. CPAQ has been shown to be reliable [41,42]. For this study, this scale was factor-analyzed to reduce the number of questions from 37 to 20. The resulting scale measures 6 coping typologies (discord, emotion, control, relationship, avoidance, and taking specific action) using a 4-point Likert scale, with a threshold score above 50 points (out of 80) indicating dysfunctional coping behavior. The CES-DC measures depressive symptoms during the past week using a 4-point Likert scale, where a higher total score indicates more depressive symptoms [49]. A cutoff score of ≥16 indicates symptoms of moderate depression, while a score of ≥30 indicates symptoms of severe depression [51,52]. The scale measures 4 dimensions of depression: depressed mood, tiredness, inability to concentrate, and feelings of being outside and lonely, and has positively stated items [52]. Additionally, this scale is a general measure of childhood psychopathology [53] and has been demonstrated to be reliable and valid among Swedish adolescents [52]. Alcohol consumption was measured using a modified AUDIT-C, which assesses the frequency of drinking, quantity consumed on a typical occasion, and frequency of heavy episodic drinking (ie, binge drinking) [50] using a 30-day perspective (as opposed to the original 12-month perspective). These questions have previously been translated into Swedish [54], and a score of ≥4 and ≥5 points for women and men, respectively, was used as a cutoff for risky drinking. This scale has been demonstrated to be reliable and valid for Swedish adolescents [55]. Furthermore, 2 questions were added concerning whether the participants had ever consumed alcohol to the point of intoxication and their age at the onset of drinking and intoxication. The original version of the LoL was designed for adults and asked the respondents to reflect on their, present, and future life status from a 5-year perspective on a 10-point Visual Analogue Scale representing life status from “worst” to “best” possible life imaginable [56]. A modified version for children, using a time frame of 1 year, has been used previously in Sweden [57] and was used in this study.

Randomization

After completing the baseline assessment, each participant was allocated to either the intervention or the control group. An external researcher generated an unrestricted random allocation sequence using random allocation software [58]. Neither the participants nor the researchers involved in the study were blinded to group allocation.

Procedure

Based on the order in which participants were included in the study, they were allocated to 1 of the 2 study groups and informed of their allocation by email. Additionally, those who were randomized to the intervention group received a hyperlink to the Alcohol and Coping program, whereas the control group participants received information that they would gain access to Alcohol and Coping after the last follow-up assessment (ie, the waitlist control group). All participants were informed about other information and support available through web pages, notably drugsmart [59], which contains general information and facts about alcohol and drugs, in addition to more specific information about having substance-abusing parents. Telephone numbers and contact information for other organizations and primary health care facilities were also provided.

The Intervention

As noted previously, Alcohol and Coping is derived from the aforementioned manual-based face-to-face ICAIP intervention program [41,42]. The ICAIP consists of a combination of an alcohol intervention program, which is based on the short version of the Brief Alcohol Screening and Intervention for College Students program [60], and a coping intervention program developed for the purpose of the ICAIP [41,42]. Like the original ICAIP intervention, Alcohol and Coping builds on psychoeducational principles and includes components such as film-based lectures, various exercises, and both automated and therapist-assisted feedback. Briefly, once the participants logged into the Alcohol and Coping platform, they were introduced to the program, which followed the pattern of a board game (Figure 2). Following the introduction, participants took part in 3 film-based lectures (between 8 and 15 minutes each, Figure 3) concerning alcohol problems within the family. The respective lectures included information about (1) dependency in general as well as the genetic and environmental risks for developing dependency, (2) family patterns and how the family adapts to the one having alcohol problems, and (3) attitudes toward alcohol and how they influence drinking and the physiological effects of alcohol. After completing the lectures, the participants were asked to answer 2 questions about their own alcohol consumption (ie, how often they drink and how often they drink to intoxication), followed by an automatic feedback message that depended on their answers. It was then suggested that the participants log out of the intervention for a 1- to 2-day break. The reason for this break was to give the participants a chance to digest all information and impressions. When they logged back into the intervention, they were asked to answer 20 questions about their coping strategies, which were also followed by automatic feedback. This feedback comprised a library covering all the prewritten feedback messages, each of which was tailored to the participants’ specific answers. The participants then participated in a 5-minute–long film-based lecture on emotion and problem-focused coping in relation to family alcohol problems (Figure 3). This was followed by 4 exercises where the participants read through vignette-like stories from 4 fictional persons describing their everyday lives related to coping and alcohol problems in the family. The stories are presented by film-based introductions that are each 1-2 minutes long. Participants were then requested to respond to each story by describing how the fictive person could have coped with their situation. As a final exercise, participants were asked to reflect on their own family situation and how they cope with situations. The participants then had to take a break for a few days.

During the break, a therapist composed individual feedback that covered reflections and confirmation of the participant’s exercises and answers to questions and included suggestions on well-suited coping strategies. Additionally, the therapist encouraged the participants to talk to others in their surroundings, such as friends, teachers, or coaches, and seek further support elsewhere, such as from municipal social services, youth health care centers, or other organizations.

XSL•FO
RenderX

Finally, the therapist reflected on the participants' alcohol consumption patterns and reminded them of increased genetic and environmental risks. Those who revealed patterns of risky alcohol use were encouraged to look at 2 additional film-based lectures with more information about alcohol and intoxication (4 minutes) and alcohol use and dependency (5 minutes). Participants received this feedback once they logged back into the program, but they also had the opportunity to receive feedback through email. The total estimated effective time for completing the program was about 1 hour, but as described above, there was 1 required break when the individualized feedback was written. To keep track of the dose each participant received, each of the 15 components in the program (Figure 1) is equal to completing 6.7% (1/15) of the program in total.

Figure 2. The procedure of Alcohol and Coping, which followed the pattern of a board game where participants needed to go through all parts of the program from start to end (Swedish, "avslutning").

Figure 3. A total of 4 snapshots from the 4 film-based lectures on alcohol and coping concerning, for instance, alcohol problems and family patterns, emotion and problem-focused coping, and physiological effects of alcohol, intoxication, and addiction.

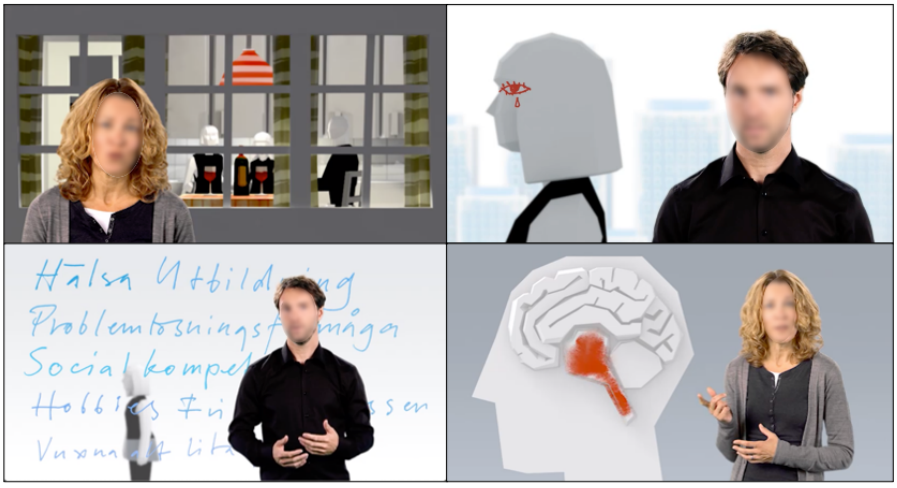

Sample Size

The trial was designed to detect a medium or large effect size corresponding to a standardized mean difference (Cohen d >0.5) [61]. An a priori calculation of the estimated sample size, using the software G*Power (G*Power Team) [62], revealed that a total of 128 participants (64 in each group) were required to enroll in the trial (power=0.80; α=.05; 2-tailed). However, to account for an estimated attrition rate of approximately 30% [34], it was necessary to enroll a minimum of 128/(1 – 0.3) = 183 participants in the trial. After a total of 204 individuals had been recruited and randomized into 2 study arms, recruitment was ended.

Statistical Analysis

Data were analyzed according to the intention-to-treat (ITT) principle, and all randomized participants were included, irrespective of whether they participated in the trial. The 4 research variables were depression (CES-DC), coping (CPAQ), alcohol use (AUDIT-C), and life status (LoL).

Data analysis consisted of comparing outcome measurements at t_1 and t_2. The baseline measurement t_0 value was added as an

XSL•FO
RenderX

adjustment variable in all models. The resulting data from CPAQ, CES-DC, and LoL were normally distributed and analyzed using linear mixed models. The resulting AUDIT-C scores were nonnormally distributed, with an excess of 0 values, and were analyzed using a 2-part model for longitudinal data. This model is sufficiently flexible to account for numerous 0 reports. This was achieved by combining a logistic generalized linear mixed model (GLMM) for the 0 parts and a skewed continuous GLMM for the non-0 alcohol consumption parts. R-package *brms* (Bayesian regression models using Stan; R Foundation for Statistical Computing) [63], a higher-level interface for the probabilistic programming language Stan [64], and a custom *brms* family for a marginalized 2-part lognormal distribution were used to fit the model [65]. The logistic part of the model represents the subject-specific effects on the odds of reporting no drinking. The continuous part was modeled using a gamma GLMM with a log link. The exponentiated treatment effect represents the subject-specific ratio of the total AUDIT-C scores between the treatment and waitlist control groups for those who reported drinking during the specific follow-up period.

Handling of Missing Data

GLMMs include all available data and provide unbiased ITT estimates under the assumption that data are missing at random, meaning that the missing data can be explained by existing data. However, it is impossible to determine whether the data are missing at random or whether the missing data are due to unobserved factors [66]. Therefore, we also assumed that data were not missing at random, and subsequent sensitivity analyses were performed [66]. We used the pattern mixture method, which assumes not missing at random, to compare those who completed the follow-up at 6 months (t_2) with those who did not (but completed the 2-month follow-up). The overall effect of this model is a combination of the effects of each subgroup. We also tested the robustness of the results by performing ANCOVAs at the 2-month follow-up, both using complete cases and with missing values imputed using multilevel multiple imputation.

The effect of the program was estimated using Cohen *d*, where a value of approximately 0.2 indicates a small effect size and values of approximately 0.5 and 0.8 indicate medium and large effect sizes, respectively [61].

Ethical Considerations

All procedures were performed in accordance with the ethical standards of the institutional or national research committees, the 1964 Helsinki Declaration and its later amendments, and comparable ethical standards. Informed consent was obtained from all the participants included in the study. This study was approved by the Swedish Ethical Review Authority (formerly the Regional Ethical Review Board in Stockholm, No. 2011/1648-31/5).

To enhance the response rates, participants received a cinema gift certificate corresponding to approximately EUR 11 (US $12) as compensation for completing each assessment. If a participant completed all assessments, an additional gift certificate was provided. The participants could subsequently receive 4 cinema gift certificates totaling EUR 44 (US $48).

Results

Overview

The trial profile is depicted in Figure 1 and reveals that 2722 individuals who were aged between 15 and 19 years performed the screening procedure. A total of 1448 individuals did not fulfill the inclusion criteria and were excluded, leaving 1274 eligible participants. Another 1070 individuals were excluded because they did not provide informed consent or complete the baseline assessment, leaving 204 participants who were allocated to 1 of the 2 study groups. A total of 140 (69%) and 131 (64%) participants completed t_1 and t_2 assessments, respectively. Of the participants in the intervention group (n=101), 63% (n=64) registered an account on the Alcohol and Coping website, 35% (n=35) completed the alcohol intervention section, and 24% (n=24) completed both the alcohol and coping intervention sections.

Sample Characteristics

The mean age of the sample was 17.0 (SD 1.23) years, and the vast majority were female, with both parents born in Sweden and currently working (Table 1). Approximately one-third of the participants reported living with both parents. The mean score on the CAST-6 was 5.33 (SD 0.87) out of a total of 6, and the majority of the sample (147/204, 72.1%) perceived their father to have alcohol problems. Approximately 12% (25/204) had never consumed alcohol, whereas approximately 70% (144/204) had consumed alcohol at a level of intoxication. The mean age at onset was 13.7 (SD 2.07) years and the age at first intoxication was 14.8 (SD 1.56) years. The proportion of participants with symptoms of at least moderate depression was 77.5% (158/204), of whom 55.1% (87/158) had symptoms of severe depression and 42.6% (87/204) had symptoms of dysfunctional coping behaviors. The percentage of participants who consumed alcohol at a risky level was 39.7% (81/204). Table 1 provides complete information regarding the study sample.

XSL•FO
RenderX

Table 1. Participant demographics at baseline.

	Intervention (n=101)	Control (n=103)	*P* value[a]
Age (years), mean (SD)	17.0 (1.29)	16.9 (1.18)	.74
Sex (female), n (%)	87 (86.1)	92 (89.3)	.49
Children of Alcoholics Screening Test-6 points, mean (SD)	5.45 (.82)	5.23 (.91)	.08
Family member with perceived alcohol problems, n (%)			
Mother	28 (27.7)	35 (34)	.33
Father	75 (74.3)	72 (69.9)	.49
Stepmother	8 (7.9)	7 (6.8)	.76
Stepfather	10 (9.9)	14 (13.6)	.41
Other	5 (5)	6 (5.8)	.78
I live with, n (%)			.71
Mother and father	29 (28.7)	29 (28.2)	
Mother or father	20 (19.8)	25 (24.3)	
Mother, father and stepmother, or father	16 (15.8)	20 (19.4)	
Part-time with mother and father	13 (12.9)	13 (12.6)	
I do not live with any of my parents	16 (15.8)	9 (8.7)	
Mother was born in, n (%)			
Sweden	90 (90.1)	83 (80.6)	.06
Other Nordic country	4 (4)	9 (8.7)	.16
Outside Nordic country	6 (5.9)	12 (11.7)	.15
Father was born in, n (%)			
Sweden	84 (83.2)	84 (81.6)	.76
Other Nordic country	6 (5.9)	9 (8.7)	.44
Outside Nordic country	11 (10.9)	11 (10.7)	.96
Mother's occupation, n (%)			
Employed	85 (84.2)	81 (78.6)	.31
Student	1 (1)	2 (1.9)	.58
On leave or maternal leave	4 (4)	8 (7.8)	.25
Unemployed	9 (8.9)	10 (9.7)	.85
Do not know	6 (5.9)	7 (6.8)	.80
Father's occupation, n (%)			
Employed	72 (71.3)	72 (69.9)	.83
Student	1 (1)	2 (1.9)	.57
On leave or paternal leave	4 (4)	7 (6.8)	.37
Unemployed	9 (8.9)	10 (9.7)	.85
Do not know	16 (15.8)	16 (15.5)	.95
Never consumed alcohol, n (%)	14 (14)	11 (10.7)	.49
Consumed alcohol to the level of intoxication, n (%)	65 (64.4)	79 (76.7)	.05
Age of onset (years), mean (SD)	13.7 (2.01)	13.7 (2.14)	.98
Age of first intoxication (years), mean (SD)	14.8 (1.36)	14.8 (1.71)	.82
Center for Epidemiological Studies Depression Scale scores, mean (SD)	27.0 (12.82)	25.8 (12.61)	.51
Coping With Parents Abuse Questionnaire scores, mean (SD)	48.57 (9.46)	48.70 (10.11)	.94
Alcohol Use Disorders Identification Test scores, mean (SD)	2.49 (2.90)	2.69 (2.75)	.61

XSL•FO
RenderX

	Intervention (n=101)	Control (n=103)	*P* value[a]
Ladder of Life scores, mean (SD)			
Past	4.51 (2.43)	4.40 (2.37)	.73
Present	5.36 (2.02)	5.20 (1.85)	.58
Future	6.80 (2.21)	6.85 (2.19)	.87

[a]Significance levels calculated by Pearson chi-square statistics for categorical variables and 2-tailed *t* tests for continuous variables.

Treatment Effects

For the primary outcome, coping behavior (CPAQ), we found a small but inconclusive treatment effect in favor of treatment at both 2 (t_1) and 6 (t_2) months (Cohen *d*=–0.05 at both t_1 and t_2). For the secondary outcome, alcohol use (AUDIT-C), we found a treatment effect in that the intervention group scored 38% less than the control group on the continuous part (ie, drinking when it occurred) at t_1 and 46% less at t_2. Regarding depression (CES-DC) and life status (LoL), all between-group comparisons of treatment effects were inconclusive at both follow-up time points (Table 2).

Table 2. Estimated treatment effects at 2-month (t_1) and 6-month (t_2) follow-ups.

Outcome	2-month follow-up (t_1)			6-month follow-up (t_2)		
	Treatment effect	Cohen *d* (95% CI)	*P* value	Treatment effect	Cohen *d* (95% CI)	*P* value
CPAQ[a]	–0.44	–0.05 (–0.32 to 0.23)	.74	–0.44	–0.05 (–0.32 to 0.23)	.75
CES-DC[b]	–0.54	–0.04 (–0.31 to 0.23)	.76	–2.61	–0.21 (–0.48 to 0.07)	.11
LoL[c]						
Past	–0.37	–0.16 (–0.44 to 0.12)	.27	0.11	0.05 (–0.23 to 0.33)	.75
Present	–0.24	–0.12 (–0.43 to 0.18)	.43	0.12	0.06 (–0.25 to 0.37)	.69
Future	–0.49	–0.22 (–0.50 to 0.06)	.12	–0.18	–0.08 (–0.36 to 0.20)	.56
AUDIT-C[d]		Ratio	N/A[e]		Ratio	N/A
Drinking (continuous)	–0.47	0.62 (0.44 to 0.88)		–0.62	0.54 (0.38 to 0.76)	
No drinking (binary)	0.71	1.97 (0.25 to 19.52)		1.19	3.07 (0.40 to 40.79)	

[a]CPAQ: Coping With Parents Abuse Questionnaire.

[b]CES-DC: Center for Epidemiological Studies Depression Scale.

[c]LoL: Ladder of Life.

[d]AUDIT-C: Alcohol Use Disorders Identification Test.

[e]N/A: not applicable.

Missing Data

In contrast to the ITT analyses, the sensitivity analyses showed that the treatment group, averaged over the levels of dropout, scored higher (ie, a negative effect) on the main outcome, coping behavior (CPAQ), at t_1 (2.44; *P*=.20). However, the results remain inconclusive.

Dose-Response Effects

We did not find any evidence for greater involvement in the program being linked to improved outcomes with regard to coping behavior.

Discussion

Overview

We did not find any support for the primary hypothesis: the intervention was not superior to the control condition with regard to coping behavior. Inconclusive results with small effect sizes were observed at both follow-up time points. However, for the secondary outcomes, we found that those in the intervention group who drank alcohol drank approximately 40%-50% less than those in the control group at both follow-ups. These results corroborate previous findings on the precursor face-to-face ICAIP intervention program, demonstrating that participants who received a combined alcohol and coping intervention reported superior outcomes with regard to alcohol-related outcomes compared to participants in the other 2 study arms, who received only a coping or alcohol intervention [41,42]. In contrast to this study, Hansson et al [42] found that all groups improved their coping skills, although the between-group comparisons were inconclusive and the improvements were maintained over time. These differences could be explained by the different settings in which the precursor program was provided (ie, face-to-face to young adults in a university setting), whereas this study targeted young people (15-19 years of age) through a web-based digital intervention. Additionally, the poor adherence in this study may explain the absence of primary

XSL•FO
RenderX

results favoring the intervention group. In a recent study, parents without alcohol problems were recruited to participate in a randomized trial evaluating the web-based SPARE (Supportive Parenting and Reinforcement) program to improve children's mental health and reduce coparents' alcohol use. In line with our study, the authors did not find the primary outcome of the SPARE program to be superior to that of the active control group (which received written psychoeducation); however, both groups reported decreased coparental alcohol consumption [38].

Considering that approximately 3600 children in 2022 participated in various forms of support provided by Swedish municipalities [23], our recruitment activities reached a large number of eligible individuals, pointing to the potential of finding these children on these platforms. There were unexpectedly high levels of depression among the participants in this study. Although the intervention did not target depressive symptoms per se, there was a trend for the intervention group to have decreased depression levels compared to the control group. A large proportion of participants had symptoms of severe depression, which may have aggravated their capacity for improvement at follow-up [28,67]. Targeting dysfunctional coping patterns could affect an individual's perceived mental health, and studies have shown that healthy coping strategies positively affect depression and anxiety in a positive way [68]. Using dysfunctional coping strategies, such as negative self-talk and alcohol consumption, can lead to depressive symptoms [69]. Targeting these symptoms in the context of healthy and unhealthy coping strategies may be a viable route to fostering appropriate coping strategies that work in the long run. Given that the young people who were reached by the intervention in this study displayed high levels of depression, future interventions for this group should include programs targeting depressive symptoms.

Almost 37% (37/101) of the intervention group did not log into the intervention at all, and only 24% (24/101) of the intervention group participants completed all parts of the program. The fact that a high proportion of the participants had symptoms of severe depression could explain the low adherence. Another reason could be that the initial film-based lectures were too long to maintain the participants' attention, as the lectures ranged from 8-15 minutes. Yet a final reason could be that we had a 1- to 2-day break built into the intervention, and for unknown reasons, some participants did not log back into the intervention. However, we did not find a dose-response relationship indicating favorable outcomes for those who completed more of the program content. High levels of attrition are not uncommon in self-directed programs such as the one in this study; for example, in a study on a smoking cessation intervention, 37% of the participants never logged into the platform [70], and in a self-directed intervention for problem gamblers, a majority dropped out after 1 week and none completed the entire program [71]. Increased intervention adherence is a priority when developing new digital interventions, particularly for young people. One method is to use more persuasive technologies, such as primary tasks, dialogue, and social support [72]. Considering children whose parents have mental disorders, Grové and Reupert [73] suggested that digital interventions should include components such as providing information about parental mental illness, access to health care, genetic risk, and suggestions for how children might initiate conversations with parents who have the illness. These suggestions should be considered in future studies on interventions for youths whose parents have substance use problems. Representatives of the target group and other relevant stakeholders should also be involved in coproducing new interventions to increase the probability of developing more engaging programs [74]. Moreover, one cannot expect study participants to return to the program more than once, and for the sake of adherence, briefer interventions should not encourage participants to log-out for a break. To keep adherence at an acceptable level, similar future interventions for this target group should also consider having symptoms of severe depression as an exclusion criterion [28,67]. Further, to improve adherence, strategies of coproduction could be used where all stakeholders, including the target group, are involved in intervention development [75]. Other important factors identified to improve adherence to digital interventions are to make the content relatable, useful, and even more interactive [76]. Those participants who have symptoms of severe depression should be referred to other appropriate health care. Finally, it is probably beneficial to develop shorter psychoeducative film-based lectures than ours, lasting up to 15 minutes. Future self-directed digital interventions targeting this population should, therefore, focus on a very brief and focused intervention, which, based on theory, has the potential to foster healthy coping behaviors that can lead to an increased quality of life and improved mental health for this group of young people.

Another concern for future projects would be to use a data-driven approach during the program development phase, where A/B testing can be used to test different setups of the program to highlight which setup works best. Another aspect that must be considered is the fast-changing world of technology, where young people are exposed to an infinite number of different apps that grab their attention, which also calls for interventions to be short and to the point. Furthermore, if the program is to spread and become generally available, one must consider that keeping the program alive for a longer period will require funding and staffing for both product management and technical support.

Strengths and Limitations

This study had several strengths. First, Alcohol and Coping is a web-based intervention program, and it appears as if the internet is a particularly promising way to provide support to adolescents growing up with parents with alcohol problems because it offers an anonymous means of communicating and makes intervention programs readily accessible [25]. Our recruitment strategies reached a considerable number of interested and eligible individuals, demonstrating the potential for recruiting through social media and other web platforms. Additionally, this program is one of the first brief web-based interventions aimed at adolescents with parents with alcohol-related problems. We used the CAST-6, which has been validated among Swedish adolescents [47], to screen eligible participants. Another strength is that the intervention program involved personalized, tailored feedback in the form of prewritten automatic messages and therapist-written

XSL•FO
RenderX

personalized feedback, both of which have proven to be important components of web-based interventions aimed at adolescents [77,78]. Finally, this study evaluated the effectiveness of the Alcohol and Coping program using a randomized controlled trial design, which is considered the strongest experimental design with regard to allocation bias.

This study had some limitations. First, the design with a passive waitlist control group and an active intervention group, both unblinded to study allocation, may have resulted in biased estimates of treatment effects. Intervention adherence was low, and most of the study participants had symptoms of depression, where 55% (87/158) had symptoms of severe depression. This may have contributed to the small and overall inconclusive effects on the primary outcomes of this study. Many digital interventions have problems with low adherence, and in a review by Välimäki et al [79], some studies reported adherence rates as low as 10%. A vast proportion of the study participants were women, making the findings difficult to generalize to men. However, another limitation concerns selection bias and external validity. We recruited study participants through social media and other relevant websites containing health-related information, including information about parents with alcohol-related problems. It is, therefore, possible that the study population can be classified as "information-seeking" adolescents, who may have different personality traits relative to other adolescents in the same home situation. Additionally, as an inclusion criterion was having ready access to computers and the internet, it is possible that participants belonging to a lower socioeconomic class were underrepresented in the study. It should also be noted that the data presented here were collected approximately 10 years ago. However, we believe our findings make an important contribution to the field since, like our intervention, many recent web-based interventions use strategies of psychoeducation, films, exercises, questions, and feedback. Further, the number of web-based interventions for this target group remains scarce in the literature, which underscores the need for future research. Finally, the study was powered to detect a medium effect size. However, given the small effect sizes detected in this study, it is plausible that too few participants were recruited to detect differences between the groups.

Implications for Practice

Although growing up with parents who have alcohol problems per se is not sufficient for developing psychosocial disorders, many children need support to manage their situation. Therefore, it is difficult to recruit children to support these groups. In Sweden, not even 2% of all children growing up with parental alcohol problems attend face-to-face support groups provided by municipalities.

Offering support through web-based intervention programs seems particularly attractive to adolescents whose parents have alcohol-related problems. To date, evidence for such programs is scarce, and there is an urgent need to develop and evaluate digital interventions targeting this group of adolescents. This study makes important contributions to this novel field of research. The results provide insight into effective strategies for delivering intervention programs to children of parents with substance abuse issues, highlighting the potential for digital interventions to reach a vulnerable, hard-to-reach group of adolescents. Our findings underscore the need to develop more engaging interventions in coproduction with the target group.

Conclusions

We found that a digital therapist-assisted self-management intervention for adolescents whose parents have alcohol use problems contributed to a reduction in the adolescents' own alcohol consumption. This result highlights the potential for digital interventions to reach a large, vulnerable, and hard-to-reach group of adolescents with support efforts. Findings were inconclusive for all other outcomes, which may be attributable to low adherence. This points to the need for future research on developing more engaging digital interventions to increase adherence among adolescents.

Acknowledgments

This work was undertaken on behalf of the Swedish Council for Information on Alcohol and Other Drugs (CAN) and was supported by grants from the Swedish National Institute of Public Health and the Swedish Council for Working Life and Social Research.

Conflicts of Interest

HH and UZ developed the study interventions. However, the parties did not derive direct financial income from these interventions. HW, PK, and THE declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 895 KB - jmir_v26i1e52118_app1.pdf]

References

1. Haugland SH, Elgán TH. Prevalence of parental alcohol problems among a general population sample of 28,047 Norwegian adults: evidence for a socioeconomic gradient. Int J Environ Res Public Health 2021;18(10):5412 [FREE Full text] [doi: 10.3390/ijerph18105412] [Medline: 34069384]

XSL•FO
RenderX

2. Elgán TH, Leifman H. Prevalence of adolescents who perceive their parents to have alcohol problems: a Swedish national survey using a web panel. Scand J Public Health 2013;41(7):680-683. [doi: 10.1177/1403494813491859] [Medline: 23761932]
3. Laslett AM, Ferris J, Dietze P, Room R. Social demography of alcohol-related harm to children in Australia. Addiction 2012;107(6):1082-1089. [doi: 10.1111/j.1360-0443.2012.03789.x] [Medline: 22229839]
4. Manning V, Best DW, Faulkner N, Titherington E. New estimates of the number of children living with substance misusing parents: results from UK national household surveys. BMC Public Health 2009;9:377 [FREE Full text] [doi: 10.1186/1471-2458-9-377] [Medline: 19814787]
5. Grant BF. Estimates of US children exposed to alcohol abuse and dependence in the family. Am J Public Health 2000;90(1):112-115 [FREE Full text] [doi: 10.2105/ajph.90.1.112] [Medline: 10630147]
6. Raninen J, Elgán TH, Sundin E, Ramstedt M. Prevalence of children whose parents have a substance use disorder: findings from a Swedish general population survey. Scand J Public Health 2016;44(1):14-17. [doi: 10.1177/1403494815601525] [Medline: 26287020]
7. Christoffersen MN, Soothill K. The long-term consequences of parental alcohol abuse: a cohort study of children in Denmark. J Subst Abuse Treat 2003;25(2):107-116 [FREE Full text] [doi: 10.1016/s0740-5472(03)00116-8] [Medline: 14629993]
8. Martikainen PN, Korhonen K, Moustgaard H, Aaltonen M, Remes H. Substance abuse in parents and subsequent risk of offspring psychiatric morbidity in late adolescence and early adulthood: a longitudinal analysis of siblings and their parents. Soc Sci Med 2018;217:106-111. [doi: 10.1016/j.socscimed.2018.09.060] [Medline: 30300760]
9. Jääskeläinen M, Holmila M, Notkola IL, Raitasalo K. Mental disorders and harmful substance use in children of substance abusing parents: a longitudinal register-based study on a complete birth cohort born in 1991. Drug Alcohol Rev 2016;35(6):728-740. [doi: 10.1111/dar.12417] [Medline: 27246821]
10. Velleman R, Templeton LJ. Impact of parents' substance misuse on children: an update. BJPsych Adv 2018 Apr 11;22(2):108-117 [FREE Full text] [doi: 10.1192/apt.bp.114.014449]
11. Ohannessian CM, Hesselbrock VM, Kramer J, Bucholz KK, Schuckit MA, Kuperman S, et al. Parental substance use consequences and adolescent psychopathology. J Stud Alcohol 2004;65(6):725-730. [doi: 10.15288/jsa.2004.65.725] [Medline: 15700509]
12. Johnson JL, Leff M. Children of substance abusers: overview of research findings. Pediatrics 1999;103(5 Pt 2):1085-1099. [doi: 10.1542/peds.103.S2.1085] [Medline: 10224196]
13. Berg L, Bäck K, Vinnerljung B, Hjern A. Parental alcohol-related disorders and school performance in 16-year-olds-a Swedish national cohort study. Addiction 2016;111(10):1795-1803 [FREE Full text] [doi: 10.1111/add.13454] [Medline: 27178010]
14. Casas-Gil MJ, Navarro-Guzman JI. School characteristics among children of alcoholic parents. Psychol Rep 2002;90(1):341-348. [doi: 10.2466/pr0.2002.90.1.341] [Medline: 11899007]
15. McGrath CE, Watson AL, Chassin L. Academic achievement in adolescent children of alcoholics. J Stud Alcohol 1999;60(1):18-26. [doi: 10.15288/jsa.1999.60.18] [Medline: 10096305]
16. Velleman R, Templeton L, Reuber D, Klein M, Moesgen D. Domestic abuse experienced by young people living in families with alcohol problems: results from a cross - european study. Child Abuse Rev 2008 Nov 24;17(6):387-409. [doi: 10.1002/car.1047]
17. Rothman EF, Edwards EM, Heeren T, Hingson RW. Adverse childhood experiences predict earlier age of drinking onset: results from a representative US sample of current or former drinkers. Pediatrics 2008;122(2):e298-e304. [doi: 10.1542/peds.2007-3412] [Medline: 18676515]
18. Anda RF, Whitfield CL, Felitti VJ, Chapman D, Edwards VJ, Dube SR, et al. Adverse childhood experiences, alcoholic parents, and later risk of alcoholism and depression. Psychiatr Serv 2002;53(8):1001-1009 [FREE Full text] [doi: 10.1176/appi.ps.53.8.1001] [Medline: 12161676]
19. Straussner SLA, Fewell CH. A review of recent literature on the impact of parental substance use disorders on children and the provision of effective services. Curr Opin Psychiatry 2018;31(4):363-367. [doi: 10.1097/YCO.0000000000000421] [Medline: 29794556]
20. Calhoun S, Conner E, Miller M, Messina N. Improving the outcomes of children affected by parental substance abuse: a review of randomized controlled trials. Subst Abuse Rehabil 2015;6:15-24 [FREE Full text] [doi: 10.2147/SAR.S46439] [Medline: 25670915]
21. Emshoff JG, Price AW. Prevention and intervention strategies with children of alcoholics. Pediatrics 1999;103(5 Pt 2):1112-1121. [doi: 10.1542/peds.103.S2.1112] [Medline: 10224198]
22. Cuijpers P. Prevention programmes for children of problem drinkers: a review. Drugs Educ Prev Policy 2009;12(6):465-475. [doi: 10.1080/09687630500337162]
23. Wannberg H. Plats för barnen—Om kommunernas stöd till barn som växer upp med missbrukande föräldrar [Make room for the children—municipalities and their support for children who grow up with parents with substance abuse]. Stockholm, Sweden: Junis, IOGT-NTO's ungdomsförbund; 2023.
24. Elgán TH, Leifman H. Children of substance abusing parents: a national survey on policy and practice in Swedish schools. Health Policy 2011;101(1):29-36. [doi: 10.1016/j.healthpol.2010.10.009] [Medline: 21055839]

XSL•FO
RenderX

25. de Sousa D, Fogel A, Azevedo J, Padrão P. The effectiveness of web-based interventions to promote health behaviour change in adolescents: a systematic review. Nutrients 2022;14(6):1258 [FREE Full text] [doi: 10.3390/nu14061258] [Medline: 35334915]
26. Andersson G. Internet-delivered psychological treatments. Annu Rev Clin Psychol 2016;12:157-179 [FREE Full text] [doi: 10.1146/annurev-clinpsy-021815-093006] [Medline: 26652054]
27. King R, Bambling M, Lloyd C, Gomurra R, Smith S, Reid W, et al. Online counselling: the motives and experiences of young people who choose the internet instead of face to face or telephone counselling. Couns Psychother Res 2007;6(3):169-174. [doi: 10.1080/14733140600848179]
28. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
29. Boggs JM, Beck A, Felder JN, Dimidjian S, Metcalf CA, Segal ZV. Web-based intervention in mindfulness meditation for reducing residual depressive symptoms and relapse prophylaxis: a qualitative study. J Med Internet Res 2014;16(3):e87 [FREE Full text] [doi: 10.2196/jmir.3129] [Medline: 24662625]
30. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023;22(2):305-314 [FREE Full text] [doi: 10.1002/wps.21088] [Medline: 37159350]
31. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
32. Li J, Theng YL, Foo S. Game-based digital interventions for depression therapy: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2014;17(8):519-527 [FREE Full text] [doi: 10.1089/cyber.2013.0481] [Medline: 24810933]
33. Pennant ME, Loucas CE, Whittington C, Creswell C, Fonagy P, Fuggle P, et al. Computerised therapies for anxiety and depression in children and young people: a systematic review and meta-analysis. Behav Res Ther 2015;67:1-18. [doi: 10.1016/j.brat.2015.01.009] [Medline: 25727678]
34. Woolderink M, Smit F, van der Zanden R, Beecham J, Knapp M, Paulus A, et al. Design of an internet-based health economic evaluation of a preventive group-intervention for children of parents with mental illness or substance use disorders. BMC Public Health 2010;10:470 [FREE Full text] [doi: 10.1186/1471-2458-10-470] [Medline: 20698962]
35. Elgán TH, Kartengren N, Strandberg AK, Ingemarson M, Hansson H, Zetterlind U, et al. A web-based group course intervention for 15-25-year-olds whose parents have substance use problems or mental illness: study protocol for a randomized controlled trial. BMC Public Health 2016;16(1):1011 [FREE Full text] [doi: 10.1186/s12889-016-3691-8] [Medline: 27663995]
36. Maybery D, Reupert A, Bartholomew C, Cuff R, Duncan Z, Foster K, et al. A web-based intervention for young adults whose parents have a mental illness or substance use concern: protocol for a randomized controlled trial. JMIR Res Protoc 2020;9(6):e15626 [FREE Full text] [doi: 10.2196/15626] [Medline: 32554368]
37. EÉk N, Romberg K, Siljeholm O, Johansson M, Andreasson S, Lundgren T, et al. Efficacy of an internet-based community reinforcement and family training program to increase treatment engagement for AUD and to improve psychiatric health for CSOs: a randomized controlled trial. Alcohol Alcohol 2020;55(2):187-195 [FREE Full text] [doi: 10.1093/alcalc/agz095] [Medline: 31912156]
38. Siljeholm O, Lindner P, Johansson M, Hammarberg A. An online self-directed program combining Community Reinforcement Approach and Family Training and parenting training for concerned significant others sharing a child with a person with problematic alcohol consumption: a randomized controlled trial. Addict Sci Clin Pract 2022;17(1):49 [FREE Full text] [doi: 10.1186/s13722-022-00332-3] [Medline: 36064466]
39. Woolderink M. Mind the gap: evaluation of an online preventive programme for adolescents with mentally ill or addicted parents. Maastricht University. 2016. URL: https://cris.maastrichtuniversity.nl/ws/portalfiles/portal/7281149/c5531.pdf [accessed 2024-03-14]
40. Maybery D, Reupert A, Bartholomew C, Cuff R, Duncan Z, McAuliffe C, et al. An online intervention for 18-25-year-old youth whose parents have a mental illness and/or substance use disorder: a pilot randomized controlled trial. Early Interv Psychiatry 2022;16(11):1249-1258 [FREE Full text] [doi: 10.1111/eip.13274] [Medline: 35118795]
41. Hansson H, Rundberg J, Zetterlind U, Johnsson KO, Berglund M. An intervention program for university students who have parents with alcohol problems: a randomized controlled trial. Alcohol Alcohol 2006;41(6):655-663 [FREE Full text] [doi: 10.1093/alcalc/agl057] [Medline: 16926171]
42. Hansson H, Rundberg J, Zetterlind U, Johnsson KO, Berglund M. Two-year outcome of an intervention program for university students who have parents with alcohol problems: a randomized controlled trial. Alcohol Clin Exp Res 2007;31(11):1927-1933. [doi: 10.1111/j.1530-0277.2007.00516.x] [Medline: 17908264]
43. Zetterlind U, Hansson H, Aberg-Orbeck K, Berglund M. Effects of coping skills training, group support, and information for spouses of alcoholics: a controlled randomized study. Nord J Psychiatry 2001;55(4):257-262. [doi: 10.1080/080394801681019110] [Medline: 11839116]

XSL•FO
RenderX

44. Hansson H, Zetterlind U, Aberg-Orbeck K, Berglund M. Two-year outcome of coping skills training, group support and information for spouses of alcoholics: a randomized controlled trial. Alcohol Alcohol 2004;39(2):135-140 [FREE Full text] [doi: 10.1093/alcalc/agh024] [Medline: 14998831]
45. Hodgins DC, Maticka-Tyndale E, El-Guebaly N, West M. The CAST-6: development of a short-form of the Children of Alcoholics Screening Test. Addict Behav 1993;18(3):337-345. [doi: 10.1037/t08356-000] [Medline: 8342446]
46. Hodgins DC, Shimp L. Identifying adult children of alcoholics: methodological review and a comparison of the CAST-6 with other methods. Addiction 1995;90(2):255-267. [Medline: 7703819]
47. Elgán TH, Berman AH, Jayaram-Lindström N, Hammarberg A, Jalling C, Källmén H. Psychometric properties of the short version of the children of alcoholics screening test (CAST-6) among Swedish adolescents. Nord J Psychiatry 2021;75(2):155-158 [FREE Full text] [doi: 10.1080/08039488.2020.1812000] [Medline: 32852242]
48. Orford J, Guthrie S, Nicholls P, Oppenheimer E, Egert S, Hensman C. Self-reported coping behavior of wives of alcoholics and its association with drinking outcome. J Stud Alcohol 1975;36(9):1254-1267. [doi: 10.15288/jsa.1975.36.1254] [Medline: 240978]
49. Schoenbach VJ, Kaplan BH, Grimson RC, Wagner EH. Use of a symptom scale to study the prevalence of a depressive syndrome in young adolescents. Am J Epidemiol 1982;116(5):791-800. [doi: 10.1093/oxfordjournals.aje.a113469] [Medline: 7148805]
50. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Alcohol Use Disorders Identification Test. Arch Intern Med 1998;158(16):1789-1795 [FREE Full text] [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
51. Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J Psychiatry 1980;137(9):1081-1084. [doi: 10.1176/ajp.137.9.1081] [Medline: 7425160]
52. Olsson G, von Knorring AL. Depression among Swedish adolescents measured by the self-rating scale Center for Epidemiology Studies-Depression Child (CES-DC). Eur Child Adolesc Psychiatry 1997;6(2):81-87. [doi: 10.1007/BF00566670] [Medline: 9257089]
53. Fendrich M, Weissman MM, Warner V. Screening for depressive disorder in children and adolescents: validating the Center for Epidemiologic Studies Depression Scale for Children. Am J Epidemiol 1990;131(3):538-551. [doi: 10.1093/oxfordjournals.aje.a115529] [Medline: 2301363]
54. Bergman H, Källmen H, Rydberg U, Sandahl C. Tio frågor om alkohol identifierar beroendeproblem. Psykometrisk prövning på psykiatrisk akutmottagning [Ten questions about alcohol as identifier of addiction problems. Psychometric tests at an emergency psychiatric department]. Läkartidningen 1998;95(43):4731-4735 [FREE Full text]
55. Källmén H, Berman AH, Jayaram-Lindström N, Hammarberg A, Elgán TH. Psychometric properties of the AUDIT, AUDIT-C, CRAFFT and ASSIST-Y among Swedish adolescents. Eur Addict Res 2019;25(2):68-77. [doi: 10.1159/000496741] [Medline: 30726842]
56. Andrews FM, Withey SB. Developing measures of perceived life quality: results from several national surveys. Soc Indic Res 1974;1(1):1-26. [doi: 10.1007/bf00286419]
57. Nagy E. Barns känsla av sammanhang—En valideringsstudie av BarnKASAM i årskurserna 1-6 (ålder 7-12 år) [Children's sense of coherence—a study validating SOC for children in grades 1-6 (7-12 years old)]. Lunds Universitet [Lund University]. 2004. URL: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1358959&fileOId=1358960 [accessed 2024-03-14]
58. Saghaei M. Random allocation software for parallel group randomized trials. BMC Med Res Methodol 2004;4:26 [FREE Full text] [doi: 10.1186/1471-2288-4-26] [Medline: 15535880]
59. drugsmart. URL: https://www.drugsmart.se/ [accessed 2024-03-19]
60. Dimeff LA, Baer JS, Kivlahan DR, Marlatt GA. Brief Alcohol Screening and Intervention for College Students: A Harm Reduction Approach. New York: The Guilford Press; 1999.
61. Cohen J. A power primer. Psychol Bull 1992;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
62. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
63. Bürkner PC. brms: an R package for bayesian multilevel models using stan. J Stat Soft 2017;80(1):1-28 [FREE Full text] [doi: 10.18637/jss.v080.i01]
64. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw 2017;76(1):1-32 [FREE Full text] [doi: 10.18637/jss.v076.i01] [Medline: 36568334]
65. Magnusson K, Nilsson A, Carlbring P. Modeling longitudinal gambling data: challenges and opportunities. PsyArxiv Preprint posted online on September 12, 2019 [FREE Full text] [doi: 10.31234/osf.io/uvxk2]
66. Coertjens L, Donche V, De Maeyer S, Vanthournout G, Van Petegem P. To what degree does the missing-data technique influence the estimated growth in learning strategies over time? A tutorial example of sensitivity analysis for longitudinal data. PLoS One 2017;12(9):e0182615 [FREE Full text] [doi: 10.1371/journal.pone.0182615] [Medline: 28902849]

XSL•FO
RenderX

67. Kunas SL, Lautenbacher LM, Lueken PU, Hilbert K. Psychological predictors of cognitive-behavioral therapy outcomes for anxiety and depressive disorders in children and adolescents: a systematic review and meta-analysis. J Affect Disord 2021;278:614-626. [doi: 10.1016/j.jad.2020.09.092] [Medline: 33035949]
68. Stallman HM, Lipson SK, Zhou S, Eisenberg D. How do university students cope? An exploration of the health theory of coping in a US sample. J Am Coll Health 2022;70(4):1179-1185. [doi: 10.1080/07448481.2020.1789149] [Medline: 32672507]
69. Stallman HM, Beaudequin D, Hermens DF, Eisenberg D. Modelling the relationship between healthy and unhealthy coping strategies to understand overwhelming distress: a Bayesian network approach. J Affect Disord Rep 2021;3:100054 [FREE Full text] [doi: 10.1016/j.jadr.2020.100054]
70. McClure JB, Shortreed SM, Bogart A, Derry H, Riggs K, St John J, et al. The effect of program design on engagement with an internet-based smoking intervention: randomized factorial trial. J Med Internet Res 2013;15(3):e69 [FREE Full text] [doi: 10.2196/jmir.2508] [Medline: 23529377]
71. Humphrey G, Bullen C. Smartphone-based problem gambling evaluation and technology testing initiative ('SPGeTTI'): final report reference 354913/00 for the Ministry of Health. National Institute for Health Innovation (NIHI), Auckland UniServices Ltd, The Univerisy of Auckland. 2019. URL: https://www.health.govt.nz/system/files/documents/publications/20190424-spgetti-354913-00-final-report.pdf [accessed 2024-03-14]
72. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
73. Grové C, Reupert A. Moving the field forward: developing online interventions for children of parents with a mental illness. Child Youth Serv Rev 2017;82:354-358. [doi: 10.1016/j.childyouth.2017.10.003]
74. Moffat BM, Haines-Saah RJ, Johnson JL. From didactic to dialogue: assessing the use of an innovative classroom resource to support decision-making about cannabis use. Drugs Educ Prev Policy 2016;24(1):85-95 [FREE Full text] [doi: 10.1080/09687637.2016.1206846]
75. Bevan Jones R, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
76. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
77. Crutzen R, de Nooijer J, Brouwer W, Oenema A, Brug J, de Vries NK. Strategies to facilitate exposure to internet-delivered health behavior change interventions aimed at adolescents or young adults: a systematic review. Health Educ Behav 2011;38(1):49-62. [doi: 10.1177/1090198110372878] [Medline: 21189422]
78. Milward J, Drummond C, Fincham-Campbell S, Deluca P. What makes online substance-use interventions engaging? A systematic review and narrative synthesis. Digit Health 2018;4:2055207617743354 [FREE Full text] [doi: 10.1177/2055207617743354] [Medline: 29942622]
79. Välimäki M, Anttila K, Anttila M, Lahti M. Web-based interventions supporting adolescents and young people with depressive symptoms: systematic review and meta-analysis. JMIR Mhealth Uhealth 2017;5(12):e180 [FREE Full text] [doi: 10.2196/mhealth.8624] [Medline: 29222079]

Abbreviations

AUDIT-C: Alcohol Use Disorders Identification Test
brms: R-package Bayesian regression models using Stan
CAST-6: Children of Alcoholics Screening Test
CES-DC: Center for Epidemiological Studies Depression Scale
CPAQ: Coping With Parents Abuse Questionnaire
GLMM: generalized linear mixed model
ICAIP: Individual Coping and Alcohol Intervention Program
ITT: intention-to-treat
LMM: linear mixed models
LoL: Ladder of Life
SPARE: Supportive Parenting and Reinforcement program

XSL•FO
RenderX

Edited by YH Lin; submitted 24.08.23; peer-reviewed by X Zhang, C Asuzu, D Liu; comments to author 28.01.24; revised version received 08.02.24; accepted 27.02.24; published 10.04.24.

Please cite as:
Wall H, Hansson H, Zetterlind U, Kvillemo P, Elgán TH
Effectiveness of a Web-Based Individual Coping and Alcohol Intervention Program for Children of Parents With Alcohol Use Problems: Randomized Controlled Trial
J Med Internet Res 2024;26:e52118
URL: https://www.jmir.org/2024/1/e52118
doi:10.2196/52118
PMID:38598286

©Håkan Wall, Helena Hansson, Ulla Zetterlind, Pia Kvillemo, Tobias H Elgán. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Therapeutic (Mika) Targeting Distress in Patients With Cancer: Results From a Nationwide Waitlist Randomized Controlled Trial

Franziska Springer[1*], MSc; Ayline Maier[2*], PhD; Michael Friedrich[1], PhD; Jan Simon Raue[2], PhD; Gandolf Finke[2], PhD; Florian Lordick[3,4], Prof Dr; Guy Montgomery[5], Prof Dr; Peter Esser[1], PhD; Hannah Brock[1], MSc; Anja Mehnert-Theuerkauf[1], Prof Dr

[1]Department of Medical Psychology and Medical Sociology, Comprehensive Cancer Center Central Germany, University Medical Center Leipzig, Leipzig, Germany

[2]Fosanis GmbH, Berlin, Germany

[3]Department of Medicine II, University Medical Center Leipzig, Leipzig, Germany

[4]University Cancer Center Leipzig, Comprehensive Cancer Center Central Germany, Leipzig, Germany

[5]Center for Behavioral Oncology, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
Anja Mehnert-Theuerkauf, Prof Dr
Department of Medical Psychology and Medical Sociology, Comprehensive Cancer Center Central Germany, University Medical Center Leipzig
Philipp-Rosenthal-Str. 55, Haus W
Leipzig, 04103
Germany
Phone: 49 341 97 18800
Email: Anja.Mehnert@medizin.uni-leipzig.de

Abstract

Background: Distress is highly prevalent among patients with cancer, but supportive care needs often go unmet. Digital therapeutics hold the potential to overcome barriers in cancer care and improve health outcomes.

Objective: This study conducted a randomized controlled trial to investigate the efficacy of Mika, an app-based digital therapeutic designed to reduce distress across the cancer trajectory.

Methods: This nationwide waitlist randomized controlled trial in Germany enrolled patients with cancer across all tumor entities diagnosed within the last 5 years. Participants were randomized into the intervention (Mika plus usual care) and control (usual care alone) groups. The participants completed web-based assessments at baseline and at 2, 6, and 12 weeks. The primary outcome was the change in distress from baseline to week 12, as measured by the National Comprehensive Cancer Network Distress Thermometer. Secondary outcomes included depression, anxiety (Hospital Anxiety and Depression Scale), fatigue (Functional Assessment of Chronic Illness Therapy-Fatigue), and quality of life (Clinical Global Impression-Improvement Scale). Intention-to-treat and per-protocol analyses were performed. Analyses of covariance were used to test for outcome changes over time between the groups, controlling for baseline.

Results: A total of 218 patients (intervention: n=99 and control: n=119) were included in the intention-to-treat analysis. Compared with the control group, the intervention group reported greater reductions in distress (P=.03; ηp^2=0.02), depression (P<.001; ηp^2=0.07), anxiety (P=.03; ηp^2=0.02), and fatigue (P=.04; ηp^2=0.02). Per-protocol analyses revealed more pronounced treatment effects, with the exception of fatigue. No group difference was found for quality of life.

Conclusions: Mika effectively diminished distress in patients with cancer. As a digital therapeutic solution, Mika offers accessible, tailored psychosocial and self-management support to address the unmet needs in cancer care.

Trial Registration: German Clinical Trials Register (DRKS) DRKS00026038; https://drks.de/search/en/trial/DRKS00026038

(J Med Internet Res 2024;26:e51949) doi:10.2196/51949

XSL•FO
RenderX

KEYWORDS

digital therapeutic; digital health; mobile health; app; cancer; randomized controlled trial; supportive care; oncology; access to care; distress; depression; anxiety; fatigue; mobile phone

Introduction

Background

In addition to somatic symptoms such as pain [1], patients with cancer report elevated levels of distress, anxiety, and depression [2,3]. Epidemiological data show that the prevalence of clinically substantial psychological distress typically ranges from 30% to 60% among patients with cancer [2,4]. Psychological distress can persist long after the end of treatment and is associated with reduced quality of life (QoL), lower cancer treatment adherence, and lower survival rates [5].

Supportive care interventions to prevent and manage the adverse psychological and physical effects of cancer across the cancer trajectory effectively improve outcomes such as emotional distress, QoL, and fatigue [6]. Optimal supportive care is holistic and patient centered, that is, based on the needs of each individual patient [7]. However, access to supportive care is often limited by a lack of specialist staff, organizational deficiencies, and barriers that cause patients to avoid or delay their treatment [8-10]. Thus, emerging or persistent supportive care needs across the cancer trajectory often go unmet, with detrimental psychosocial and emotional impacts on patients with cancer [11]. Moreover, the number of patients living with cancer has increased rapidly in recent years [12] due to improved early detection, diagnosis, and oncological treatments, posing a growing challenge to health systems worldwide to ensure adequate and long-term care for all patients with cancer [13].

The increasing use of digital health has ushered in a new era of patient-centered cancer care due to its potential for cancer care delivery [14]. Digital health interventions provide multiple benefits: they facilitate easy and low-threshold access to care, can overcome barriers to care (eg, location, time, and health status), may enhance symptom management through real-time symptom assessment, are scalable, and provide cost-effective and efficient information sharing [14]. Growing literature suggests that digital therapeutics, a subset of digital health interventions providing evidence-based treatments driven by software, play a useful role in addressing the unmet needs of patients with cancer [15]. For instance, various mobile apps have proven to be effective in catering to specific needs of patients with cancer, such as pain, anxiety, or QoL, by using different types of interventions, such as psychoeducation, physical exercises, or coping skills training (eg, [16-19]). Moreover, large analyses such as systematic reviews and meta-analyses evaluating the efficacy of app-based interventions for patients with cancer show positive effects on patient-relevant outcomes, such as distress, QoL, anxiety, depression, pain, and fatigue [20-23].

Existing app-based supportive care interventions provide various intervention modules, such as symptom monitoring, psychoeducation, mindfulness exercises, physical exercises, and cognitive behavioral therapy (CBT) techniques [24]. However, most of these apps are limited in their scope, targeting only specific symptoms (eg, fatigue) [25] and health behaviors (eg, physical activity) [26], or provide only a single function (eg, mindfulness training or symptom tracking) [27-29]. Furthermore, some of these apps were originally developed for non–oncology patient populations and have only been slightly adapted for patients with cancer [30]. Only a few apps offer a broader range of intervention modules [25,31], but they target specific subgroups of patients with cancer (eg, patients with 1 tumor entity or with specific symptoms).

Despite the evident need, there is yet no digital therapeutic that comprehensively addresses the problems faced by all patients with cancer and simultaneously offers tailored support for each individual patient. Therefore, we investigated the efficacy of Mika (developed by Fosanis GmbH), an app-based digital therapeutic that addresses all patients with cancer transdiagnostically and provides a holistic supportive care intervention. The app incorporates evidence-based supportive care elements, such as distress and symptom monitoring [32], CBT-based coping skills training [33], mindfulness-based stress reduction (MBSR) [34,35], strength and flexibility training [36], and patient education [37], thus targeting different aspects of psychological distress. An artificial intelligence algorithm individually tailors the content of the app to patients' needs, considering cancer type, cancer treatment stage, and use behavior. A previously conducted pilot study of 70 patients with gynecological cancer indicated Mika's feasibility and potential efficacy [38]. Considering the significant prevalence and impact of psychological distress among patients with cancer, this condition was selected as the primary end point of our study. This is underscored by the app's integrated features for distress tracking and management alongside the widespread recommendation for distress screening in routine clinical care. Distress is recognized as a crucial clinical marker for assessing the efficacy of interventions across various tumor types and catering to the immediate and long-term supportive care needs of this patient group.

Objectives

The primary aim of this waitlist randomized controlled trial (RCT) was to examine the efficacy of the Mika app for general distress in patients with cancer. The secondary aim was to assess the efficacy of the Mika app on anxiety, depression, fatigue, and QoL. We hypothesized that participants receiving access to the Mika app plus usual care (UC) for 12 weeks would report greater reductions in distress, anxiety, depression, and fatigue and greater improvements in QoL compared to participants receiving UC only.

Methods

Study Design

This nationwide unblinded 2-arm waitlist RCT evaluated the efficacy of the app-based digital therapeutic Mika in reducing distress in patients with cancer and was conducted fully

XSL•FO
RenderX

decentralized in Germany, that is, participant recruitment, delivery of the study intervention, and outcome data collection were conducted without involving in-person contact between the study team and the participants. In this RCT, participants were assigned to either (1) access to the Mika app plus UC (intervention group [IG]), or (2) UC alone (control group [CG]). Participants were assessed at baseline (t0), 2 weeks (t1), 6 weeks (t2), and 12 weeks (t3) using self-report questionnaires. Once the participants in the CG completed the 12-week questionnaire, they also received access to the Mika app.

Ethical Considerations

The trial was approved by the Ethics Committee of the Medical Faculty of Leipzig University (404/21-ek) and was registered at the German Clinical Trials Register (DRKS00026038) in October 2021. All participants provided written informed consent prior to their participation in the study and retained the autonomy to withdraw from the study at any time. All personal data collected and used for this study underwent deidentification to safeguard the anonymity of participants. Monetary compensation was not provided to participants for their involvement in the study.

Participants

Textbox 1 shows the inclusion and exclusion criteria for this study. We only included patients who had been diagnosed with cancer or relapse within the last 5 years as they are likely to feel burdened by the physical and psychological effects of the disease and its treatment and therefore require supportive care. Epidemiological data indicate that supportive care needs typically decline in the years of long-term survivorship (cancer or relapse diagnosis ≥5 years ago) [39]. Participants were required to confirm their cancer diagnosis during the course of the study by submitting a letter from their treating physician. The study team enrolled patients after they had provided written informed consent, which had to be completed at home and submitted by email or mail.

Textbox 1. Inclusion and exclusion criteria for this study.

Inclusion criteria

- Age≥18 years
- Cancer diagnosis or relapse diagnosis within the last 5 years (10th revision of the International Statistical Classification of Diseases and Related Health Problems: C00-C97)
- Access to a smartphone or tablet
- Ability to provide informed consent

Exclusion criteria

- Insufficient German language skills
- Inability to use a smartphone or tablet
- Prior use of the investigated digital therapeutic

Random Assignment

Participants were randomly assigned (1:1) to either the IG or CG using permuted block randomization with blocks of 4 based on an a priori created randomization list. The allocation sequence was concealed from the study investigators until assignment. Due to the nature of the intervention, it was not feasible to blind participants or the study team to the group assignment.

Recruitment and Procedure

Between September and November 2021, patients were recruited via social media advertising campaigns (Facebook and Instagram, Meta Inc) and informational emails to cancer support groups that directed patients to the trial website with a contact form for study registration. In addition, patients were recruited from a participant pool consisting of participants from previous independent studies at the University Medical Center Leipzig. Patients from the participant pool were approached directly by the study team via phone.

All interested patients were screened by phone to determine eligibility. To identify patients who were already users of the digital therapeutic, the study team asked participants about their use of digital support, however, without referring to the publicly available digital therapeutic by name to prevent CG patients from accessing the digital therapeutic before their enrollment in the study. Eligible patients received study information in the form of a video and text via email. Patients were informed that they were required to submit a physician's letter confirming their cancer diagnosis via a secure cloud data-sharing service (TeamDrive, Crunchbase) during the course of their study participation. After providing informed consent, the participants were randomized into the IG or CG and completed the baseline questionnaires. Participants were informed about their group assignment following a completed baseline assessment. IG participants received a study access code to activate the app after downloading it from the app stores for either Android or iOS smartphones, allowing free use. The questionnaire battery was administered electronically using LimeSurvey (LimeSurvey GmbH). All participants received email invitations and reminders at 2, 6, and 12 weeks to complete the questionnaire. This RCT focused on changes in outcomes from baseline (t0) to week 12 (t3). The 2 assessments in between (t1 and t2) were not part of the analysis; an analysis of the trajectory of the symptoms is planned for the future. Once the CG participants completed the 12-week questionnaire, they also received a study

XSL•FO
RenderX

access code that could be used to activate the app. All the participants received information about the app's content and technical application via a standardized telephone introduction to the app. All participants were contacted for an exploratively structured telephone interview after completing the 12-week questionnaire. During this interview, the use of psychotherapeutic support during study participation was assessed. Data collection ended in March 2022.

Monitoring

Data monitoring was performed via standardized phone calls following questionnaire completion of each participant across all measurement time points to ensure data validity. These phone calls served to ask participants to provide missing questionnaire data, to allow participants to clarify difficulties in understanding single questionnaire items, and to provide assistance with limited app functionality. Missing questionnaire data were entered directly into the database by the study team, with a study team member reading the unanswered questions and associated response options to participants verbatim, prompting them to select their response option.

Self-reported adverse reactions and side effects of the investigated digital therapeutic were assessed at each measurement time point as part of the web-based questionnaire battery.

Intervention

Mika is an app-based digital therapeutic that provides a personalized supportive intervention aiming to reduce distress associated with cancer and its medical treatment, thus improving patients' QoL. Mika comprises 3 modules: *Check-Up*, *Discover*, and *Journeys*. The *Check-Up* module allows for the monitoring of distress and symptom monitoring with electronic patient-reported outcomes that can be shared and discussed with the attending physician. The *Discover* module delivers coaching via articles and videos on cancer types and medical treatments, psychological well-being, physical activity, diet, and social and financial issues, which are based on scientific evidence and presented in a clear and understandable manner for patients. The *Journeys* module provides users with evidence-based, resource-activating training courses combining psychoeducation and exercises to help patients cope with the mental and physical effects of cancer, for example, coping with stress and fatigue, making decisions, or living with immunotherapy (for more details on the app modules, refer to Table 1 and Figure 1). An artificial intelligence algorithm within the app customizes the content for each patient. This includes personalized recommendations based on cancer type, cancer treatment stage, and crucially; the nature and severity of reported symptoms; and ensuring personalized support for each individual. This customization process not only accounts for general patient information but also actively incorporates real-time symptom tracking data and user reading behavior using an attentional factorization machine that predicts a patient's likelihood of engaging with specific content. This approach focuses on important feature interactions related to content consumption [40], ensuring that recommendations are dynamically adjusted as patients report changes in symptoms and interact with the content. In addition, the algorithm uses a Dirichlet loss function to estimate the uncertainty in predictions [41], allowing the content to be ranked and presented based on the estimated read probability. The model undergoes monthly updates using historical data, optimizing through hyperparameter tuning evaluated by 7-fold time series cross-validation.

It is hypothesized that the digital therapeutic empowers patients with cancer by improving their health literacy and self-management along the cancer trajectory using evidence-based methods, such as symptom monitoring, patient education, MBSR, strength and flexibility training, acceptance and commitment therapy, and CBT-based coping skills training.

The Mika app was developed by Fosanis GmbH in collaboration with leading research institutions, such as the Charité University Hospital Berlin, University Hospital Leipzig, and the National Center for Tumor Diseases Heidelberg. All content of the app was carefully reviewed by experts (eg, oncologists, psychotherapists, nutritionists, and physiotherapists) before publication. The feasibility and preliminary efficacy of Mika were investigated in a previously conducted randomized pilot study involving 70 patients with gynecological cancer [38]. Mika is available for download free of charge in German and the United Kingdom app stores for Android and iOS smartphones.

IG participants could freely choose the modules to work on. While regular app use was recommended, participants were instructed to use the app at least 3 times a week.

XSL•FO
RenderX

Table 1. Mika app modules and description.

Distress and symptom monitoring (Check-Up)	Resource-activating training courses (Journeys)	Patient education (Discover)
Monitoring tools (electronic PROs[a]) allow for continuous monitoring of distress levels and symptoms using diary and calendar features. Wearables can be connected to complement PROs. Tracked data can be exported and shared with the treatment team (format: human-readable PDF, machine-readable Fast Healthcare Interoperability Resources). **Monitoring of distress levels** • Distress levels or areas can be tracked using the National Comprehensive Cancer Network Distress Thermometer (scale range 0-10) and problem list (practical problems, family problems, emotional problems, spiritual or religious concerns, and physical problems). A value exceeding the threshold of ≥5 prompts a warning that psychosocial support might be indicated. **Monitoring of symptoms** • Patients can choose from a list of 58 common symptoms along the cancer continuum and can add additional individual symptoms. Severity of symptoms are documented using PRO Common Terminology Criteria for Adverse Events items. If tracked symptom severity values exceed the predefined scale thresholds, a warning is triggered advising the patient to seek medical consultation.	Treatment and recovery programs combining patient education and exercises (journaling, mindfulness, relaxation, physical activity, and infotainment) to help patients cope with the mental and physical effects of cancer. The several days courses are designed to strengthen patients' resilience and trigger behavioral change. **List of available Journeys** • Stress relief • Gain control • Find your way • Sources of strength • Control your feelings • Accept your body • Alleviate fatigue • Making decisions • Living with immunotherapy • Maintenance therapy with Zejula • Ovarian cancer treatment • Living with breast cancer • Yoga and cancer • Nutrition and cancer	Magazine-like content hub consisting of educational articles and videos addressing psychological well-being, physical activity, diet, and social and financial issues. Topics are structured into different categories: • Cancer and treatment types • Symptoms and side effects • Nutrition in cancer • Healthy lifestyle • Partnership and family • Relaxation • Exercise training • Law and finances • COVID-19 and cancer • Law and finances • Recipes • Survivor stories

[a]PRO: patient-reported outcome.

Figure 1. Screenshots of the Mika app modules.

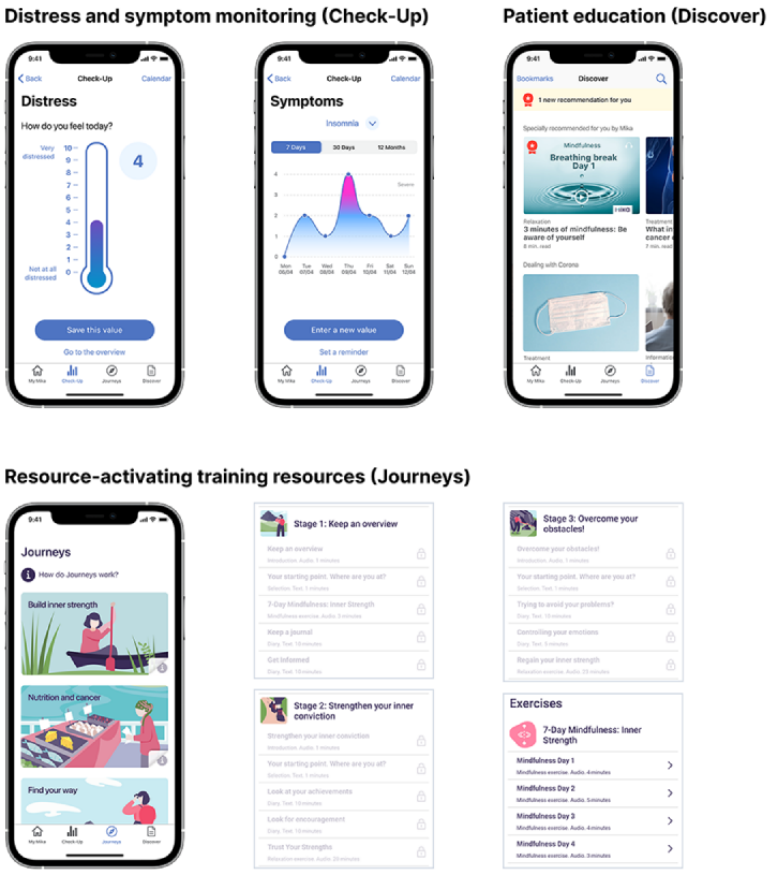

XSL•FO
RenderX

UC Condition

UC consisted of all health care that patients in Germany usually receive. There were no restrictions on health care use.

Outcome Assessment

Primary Outcome

The primary outcome was the change in psychological distress from baseline to 12 weeks, measured using the validated German version of the National Comprehensive Cancer Network Distress Thermometer [42]. Distress Thermometer is a well-established single-item self-report measure that assesses the global level of distress on a 0 (no distress) to 10 (extreme distress)-point Likert scale. It shows excellent psychometric properties across various cancer populations worldwide and is recommended as a clinical tool for routine clinical care [43]. A score ≥5 indicates clinically significant levels of distress.

Secondary Outcomes

The secondary outcomes included changes in anxiety and depression symptoms, fatigue from baseline to 12 weeks, and QoL at 12 weeks. Anxiety and depression symptoms were measured using the Hospital Anxiety and Depression Scale [44], a 14-item self-report measure of anxiety and depression, with 7 items measuring each subscale. Scores for each subscale range from 0 to 21, with a higher score indicating higher levels of anxiety or depression and a cutoff score of ≥8 for each subscale. Fatigue was assessed using the Functional Assessment of Chronic Illness Therapy-Fatigue [45], a 13-item measure that assesses self-reported tiredness, weakness, and difficulty in performing usual activities due to fatigue. The Functional Assessment of Chronic Illness Therapy-Fatigue score ranges from 0 to 52, with higher scores representing less fatigue. Self-reported QoL was measured using an adapted version of the Clinical Global Impression-Improvement Scale [46], a single-item 7-point measure that assesses the overall improvement of a patient's disease relative to a baseline state at the beginning of the intervention. In this trial, the Clinical Global Impression-Improvement Scale measured improvement in QoL relative to the beginning of the study, with a value of 4 indicating no change, <4 indicating improvement, and >4 indicating deterioration in QoL.

Intervention Safety

The safety of the digital therapeutic was assessed by the number and type of self-reported adverse reactions and side effects during the trial duration.

Intervention Adherence and Engagement

Adherence to the intervention was assessed by tracking app activities. IG participants were considered active once they activated the app using the study access code and consented to the Mika app's privacy terms. Subsequently, their pseudonymized in-app activities were automatically recorded as log data. These log data facilitated the evaluation of intervention adherence, defined as the number of days with ≥1 app activity during each of the three 4-week periods (0-4, 5-8, and 9-12 weeks) within the 12-week intervention. Such an approach enabled us to capture the frequency and diversity of app engagement, thus embodying a comprehensive definition of adherence. In addition, engagement across the app's 3 modules—Check-Up, Discover, and Journeys—was analyzed.

Statistical Analysis

Given an estimated dropout rate of 20% (50/250), a priori sample calculations showed that a sample of 2×125 (N=250) at baseline was needed to detect a change of 1 scale point (SD 2; α=.05; 1−β=.8) in the primary outcome.

Primary analyses were performed using the intention-to-treat (ITT) principle, which included all randomized participants with a confirmed cancer diagnosis by a physician's letter. Analyses were also performed per-protocol (PP), which was restricted to participants who (1) completed the self-report questionnaire at all measurement time points, (2) did not receive psychotherapeutic support during study participation, (3) did not use the investigated digital therapeutic before receiving access during study participation, and (4) used the investigated digital therapeutic at least 1 time per period up to the 5- to 8-week period of the 12-week intervention period (only IG).

Analysis of covariance was used to examine changes in distress, depression, anxiety, and fatigue outcomes between the trial arms from baseline to 12 weeks, controlling for baseline scores. Exploratory regression analyses were conducted to investigate potential variables influencing the primary outcome. These analyses focused exclusively on sociodemographic and clinical factors that showed differences between the IG and CG in the initial group comparison. Partial eta–squared was reported as the effect size for all analyses of covariance, with effect sizes interpreted as small, medium, and large at ≥0.01, ≥0.06, and ≥0.14 [47], respectively. Differences in QoL between trial arms at follow-up (12 weeks) were analyzed with a 2-tailed 2-sample *t* test, using Hedges *g*' as a measure of effect size (≥0.2=small effect, ≥0.5=medium effect, and ≥0.8=large effect [47]).

Missing outcome data at random were imputed using the expectation-maximization algorithm. For dropouts, the last observation carried forward was used. For deceased participants, the worst possible values were assumed. Dropouts were participants who failed to complete the baseline or follow-up questionnaires or failed to provide a physician's letter confirming their cancer diagnosis. A dropout analysis was performed to compare the variables of age, sex, and baseline distress between study noncompleters (dropouts) and study completers using chi-square and *t* tests. Furthermore, to model the robustness of the primary efficacy analysis under different assumptions for missing data mechanisms, an explorative sensitivity analysis using reference-based multiple imputation (jump-to-reference) [48] was performed in the extended ITT population (all randomized participants). For this purpose, monotone missing values were replaced using the jump-to-reference approach, whereas sporadic missing values were replaced under the assumption of missing at random. For jump-to-control and jump-to-reference imputation, 50 data sets were generated to minimize the loss of statistical power. The results were then aggregated across the imputed data sets [49].

All statistical tests were 2-tailed, with a significance level of 5%. Analyses were performed using R (version 4.1.0; R Foundation for Statistical Computing) [50].

XSL•FO
RenderX

Results

Study Sample

Over the 3-month recruitment period, 517 persons were screened for eligibility and 321 were determined eligible. Of the 321 participants, 248 (77.3%) gave informed consent and were randomly assigned to the IG and the CG (Figure 2). Of the 248 participants, 37 (14.9%) were considered dropouts because they did not complete baseline or follow-up assessments (n=7), failed to confirm their cancer diagnosis by submission of a physician's letter (n=7), or both (n=23). Age and sex of study dropouts and study completers did not differ (P_{age}=.89 and P_{sex}=.23), but participants who dropped out showed higher distress levels at baseline compared to study completers (*P*=.02). Participants without a verified cancer diagnosis (30/248, 12.1%) were excluded from the ITT analysis, resulting in an ITT population of 218 participants (n=99, 45.4% IG and n=119, 54.6% CG). Of the 218 participants, 173 (79%) were recruited via social media advertisements and cancer support groups and 45 (21%) were recruited using the participant pool of prior studies.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram. CG: control group; IG: intervention group; ITT: intention-to-treat; PP: per-protocol.

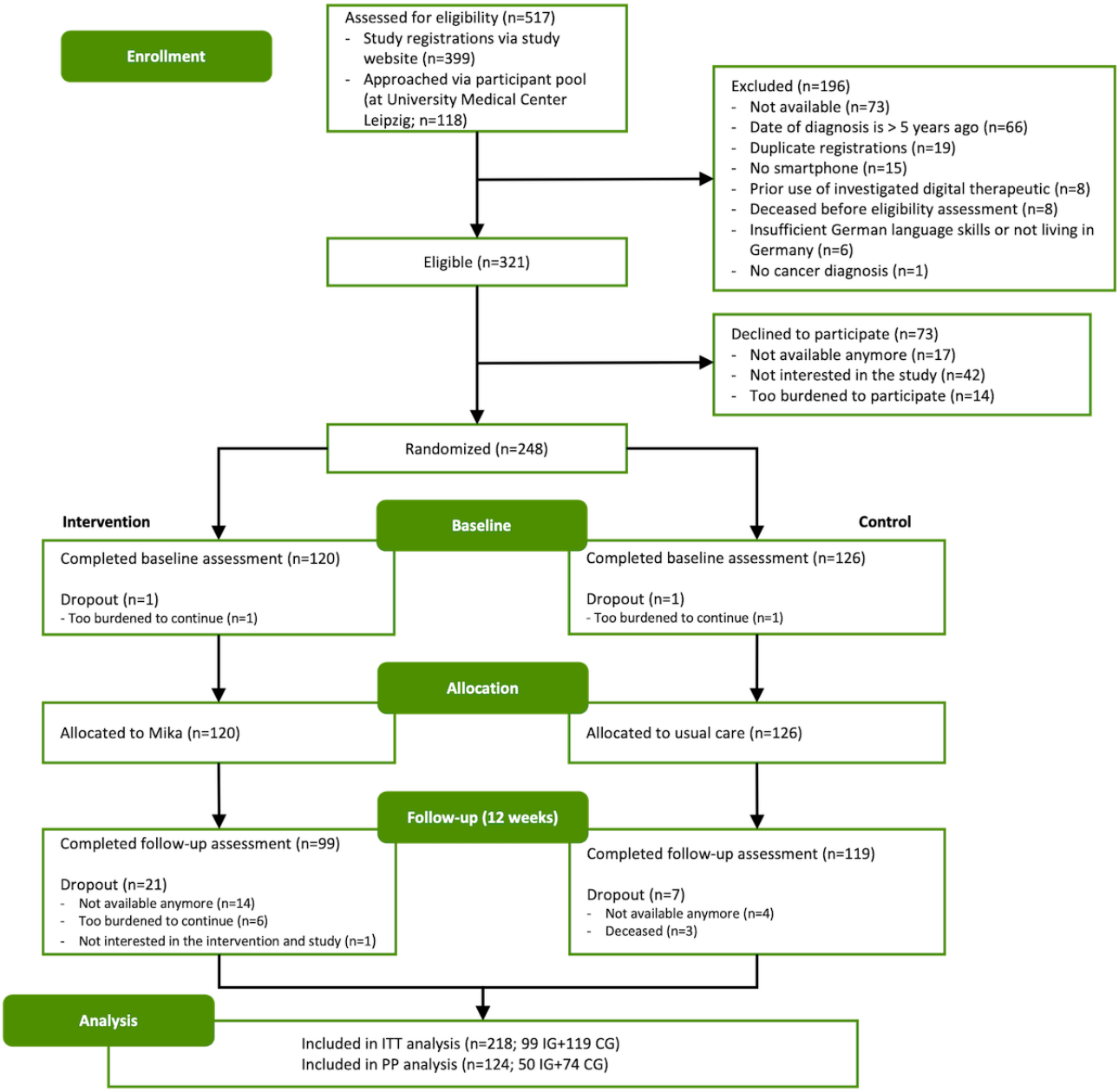

Baseline characteristics were balanced between the groups (Table 2), but participants in the IG were younger compared with those in the CG (*P*=.02). No baseline differences in the primary and secondary outcome parameters were observed between the groups, with *P* values as follows: *P*=.99 (distress), *P*=.25 (depression), *P*=.47 (anxiety), and *P*=.21 (fatigue). On average, participants were 56 (SD 11) years old, and 60.6% (132/218) of the participants were female and had been diagnosed with cancer 25 (SD 17) months earlier. The most frequently reported cancer types were breast cancer (74/218, 33.9%) and hematological cancer (61/218, 28%), with 8.7% (19/218) of participants reporting a diagnosis of relapsed cancer.

XSL•FO
RenderX

The PP population comprised 124 participants, following the exclusion of 94 participants. The primary reasons for exclusion were psychotherapeutic support during study participation and prior use of the investigated digital therapeutic.

Table 2. Baseline sociodemographic and medical sample characteristics (n=218).

Variable	Intervention[a] (n=99)	Control[b] (n=119)
Age (years), mean (SD; range)	55 (11; 20-77)	58 (10; 35-76)
Sex, n (%)		
Female	65 (65.7)	67 (56.3)
Male	34 (34.3)	52 (43.7)
Cancer type, n (%)		
Breast	38 (38.4)	36 (30.3)
Hematological	27 (27.3)	34 (28.6)
Lung	9 (9.1)	12 (10.1)
Urogenital	6 (6.1)	14 (11.8)
Gastrointestinal	7 (7.1)	11 (9.2)
Head-neck	3 (3)	4 (3.4)
Gynecologic	3 (3.0)	3 (2.5)
Other	6 (6.1)	5 (4.2)
Relapse, n (%)	7 (7.1)	12 (10.1)
Time (months) since cancer diagnosis, mean (SD)	25 (18)	24 (16)
Psychotherapeutic support during study participation, n (%)	15 (15.2)	17 (14.3)
Digital therapeutic use before study, n (%)	15 (15.2)	14 (11.8)
NCCN Distress Thermometer≥5[c], n (%)	71 (71.8)	86 (72.3)
Anxiety (HADS-A)≥8[d], n (%)	56 (56.7)	67 (56.3)
Depression (HADS-D)≥8[e], n (%)	42 (42.4)	41 (34.5)
Exclusion in per-protocol analysis[f], n (%)	49 (49.5)	45 (40.9)
Psychotherapeutic support during study	15 (15.2)	17 (14.3)
Prior use of investigated digital therapeutic	15 (15.2)	28 (23.5)
Missing questionnaire	5 (5.1)	5 (4.2)
Digital therapeutic was not used up until the 5- to 9-week period	17 (17.2)	N/A

[a]Intervention=12-week access to digital therapeutic app intervention+usual care.

[b]Control=usual care.

[c]NCCN Distress Thermometer: National Comprehensive Cancer Network Distress Thermometer (at baseline, clinically significant level of distress≥5).

[d]HADS-A: Hospital Anxiety and Depression Scale, anxiety subscale (German version, at baseline, cutoff score ≥8).

[e]HADS-D: Hospital Anxiety and Depression Scale, depression subscale (German version, at baseline, cutoff score ≥8).

[f]Multiple reasons are possible within 1 patient, and cases do not add up to the total number.

Primary Outcome

After 12 weeks, participants in the IG reported a reduced level of distress compared to participants in the CG in the ITT population ($F_{1,215}$=4.7; P=.03; ηp^2=0.02; Table 3). The observed treatment effect was more pronounced in the PP population ($F_{1,121}$=6.9; P=.01; ηp^2=0.05). The analysis revealed that higher levels of baseline distress predicted a greater change in distress after 12 weeks in the IG. An exploratory regression analysis yielded no predictive effect of age on the change in distress. The explorative sensitivity analysis among all randomized participants (n=248) showed comparable treatment effects (jump-to-control: $F_{1,19375.1}$=5.3; P=.02; ηp^2=0.02 and jump-to-intervention: $F_{1,15314.8}$=5.9; P=.02; ηp^2=0.02).

XSL•FO
RenderX

Table 3. Primary outcome: change in distress between baseline and follow-up (12 weeks)[a].

	Intervention[b]		Control[c]		Analysis of covariance		
	Values, n	Values, mean (SD)	Values, n	Values, mean (SD)	*F* test (*df*)	*P* value	ηp²
Intention-to-treat							
Baseline	99	6.0 (2.1)	119	6.0 (2.4)	N/A[d]	N/A	N/A
Follow-up	99	5.1 (2.5)	119	5.9 (2.5)	N/A	N/A	N/A
Change	99	–0.8 (2.8)	119	–0.1 (2.9)	4.7 (1, 215)	.03	0.02
Per-protocol							
Baseline	50	6.0 (2.2)	74	5.7 (2.2)	N/A	N/A	N/A
Follow-up	50	4.8 (2.8)	74	5.9 (2.4)	N/A	N/A	N/A
Change	50	–1.2 (3.2)	74	0.2 (2.8)	6.9 (1, 121)	.01	0.05

[a]An analysis of covariance was used to test for differences in change in distress levels between groups from baseline to follow-up (12 weeks), controlling for baseline. The partial eta–squared is the reported standardized effect size for the mean difference. The effect sizes can be interpreted as small, medium, or large at ≥0.01, ≥0.06, and ≥0.14, respectively. The results of the intention-to-treat and per-protocol analysis are reported.

[b]Intervention=12-week access to digital therapeutic app intervention+usual care.

[c]Control=usual care.

[d]N/A: not applicable.

Secondary Outcomes

In the ITT population, symptoms of anxiety ($F_{1,215}$=4.8; P=.03; ηp²=0.02), depression ($F_{1,215}$=15.5; P<.001; ηp²=0.07), and fatigue ($F_{1,215}$=4.4; P=.04; ηp²=0.02) improved in participants in the IG from baseline to 12 weeks compared to participants in the CG (Table 4). The observed treatment effects on anxiety and depression were more pronounced in the PP population (anxiety: $F_{1,121}$=7.2; P=.01; ηp²=0.06 and depression: $F_{1,121}$=14.9; P<.001; ηp²=0.11). A trend-to-significant treatment effect was observed for fatigue symptoms in the PP population ($F_{1,121}$=3.8; P=.05; ηp²=0.03). QoL did not differ significantly between the groups at 12 weeks (ITT: t_{216}=0.88; P=.38; g=0.12 and PP: t_{122}=1.63; P=.11; g=0.30).

XSL•FO
RenderX

Table 4. Secondary outcomes: changes in anxiety and depression symptoms and fatigue between baseline and follow-up (12 weeks) and quality of life (QoL) at follow-up (12 weeks).

	Intervention[a]		Control[b]		Analysis of covariance				
	Values, n	Values, mean (SD)	Values, n	Values, mean (SD)	*F* test (*df*)	*t* test (*df*)	*P* value	ηp^2	Hedges *g*
Anxiety (HADS-A[c])									
ITT[d]									
Baseline	99	8.6 (4.4)	119	8.2 (3.9)	N/A[e]	N/A	N/A	N/A	N/A
Follow-up	99	7.7 (4.3)	119	8.2 (4.1)	N/A	N/A	N/A	N/A	N/A
Change	99	–0.9 (2.6)	119	0.0 (3.0)	4.8 (1, 215)	N/A	*.03* [f]	0.02	N/A
PP[g]									
Baseline	50	8.0 (4.4)	74	7.6 (3.6)	N/A	N/A	N/A	N/A	N/A
Follow-up	50	7.0 (4.3)	74	7.8 (4.0)	N/A	N/A	N/A	N/A	N/A
Change	50	–1.0 (2.6)	74	0.2 (2.3)	7.2 (1, 121)	N/A	*.01*	0.06	N/A
Depression (HADS-D[h])									
ITT									
Baseline	99	7.2 (4.4)	119	6.6 (4.1)	N/A	N/A	N/A	N/A	N/A
Follow-up	99	6.4 (4.5)	119	7.5 (4.9)	N/A	N/A	N/A	N/A	N/A
Change	99	–0.8 (2.8)	119	0.9 (3.5)	15.5 (1, 215)	N/A	*<.001*	0.07	N/A
PP									
Baseline	50	6.7 (4.3)	74	6.1 (3.7)	N/A	N/A	N/A	N/A	N/A
Follow-up	50	5.6 (4.4)	74	7.1 (4.7)	N/A	N/A	N/A	N/A	N/A
Change	50	–1.1 (3.0)	74	1.1 (3.0)	14.9 (1, 121)	N/A	*<.001*	0.11	N/A
Fatigue (FACIT-F[i])									
ITT									
Baseline	99	29.2 (10.6)	119	31.1 (11.4)	N/A	N/A	N/A	N/A	N/A
Follow-up	99	32.2 (10.8)	119	31.2 (13.0)	N/A	N/A	N/A	N/A	N/A
Change	99	3.0 (8.6)	119	0.1 (9.3)	4.4 (1, 215)	N/A	*.04*	0.02	N/A
PP									
Baseline	50	30.1 (11.5)	74	32.1 (11.2)	N/A	N/A	N/A	N/A	N/A
Follow-up	50	34.2 (11.4)	74	32.9 (12.8)	N/A	N/A	N/A	N/A	N/A
Change	50	4.2 (8.8)	74	0.8 (8.3)	3.8 (1, 121)	N/A	.05	0.03	N/A
Improvement in QoL (CGI-I[j])									
ITT									
Follow-up	99	3.8 (1.1)	119	3.9 (1.4)	N/A	0.88 (216)	.38	N/A	0.12
PP									
Follow-up	50	3.7 (1.2)	74	4.1 (1.3)	N/A	1.63 (122)	.11	N/A	0.30

[a]Intervention=12-week access to digital therapeutic app intervention+usual care.

[b]Control=usual care.

[c]HADS-A: Hospital Anxiety and Depression Scale, anxiety subscale (German version, at baseline, cutoff score ≥8).

[d]ITT: intention-to-treat.

[e]N/A: not applicable.

[f]Italicized values are significant at $P<.05$.

[g]PP: per-protocol.

XSL•FO
RenderX

[h]HADS-D: Hospital Anxiety and Depression Scale, depression subscale (German version, at baseline, cutoff score ≥8).

[i]FACIT-F Functional Assessment of Chronic Illness Therapy–Fatigue.

[j]CGI-I: Clinical Global Impression Improvement.

Safety Outcomes

IG participants reported no adverse reactions or side effects of digital therapeutic during the study.

Intervention Adherence and Engagement

Of the 99 participants in the IG (ITT), 98 (99%), 78 (79%), and 67 (68%) used the digital therapeutic intervention at 0- to 4-, 5- to 8-, and 9- to 12-week periods of the 12-week intervention, respectively, demonstrating good initial adherence to the intervention, which decreased moderately over time. App use (module use and days spent on the app) decreased over time (Table 5). IG participants accessed content from various categories at different frequencies. The most accessed categories were cancer therapy, symptoms and side effects, and nutrition in cancer, with 80% (79/99), 83% (82/99), and 80% (79/99) of users accessing the content in these categories, respectively. Conversely, partnership and family, relaxation, and recipes were accessed less, with 29% (28/99), 34% (33/99), and 32% (31/99) of users, respectively.

Table 5. App use in the intervention group over time (n=99).

	Week 0-4 (n=98)	Week 5-8 (n=78)	Week 9-12 (n=67)
Days of app use, median (IQR)	9 (4-15)	3 (2-14)	4 (2-10)
Tracking of distress in days, median (IQR)	5 (1-12)	2 (0-14)	1 (0-6)
Articles read, median (IQR)	6 (2-16)	2 (0-9)	1 (0-5)
Patients using a Journey, n (%)	64 (65.3)	18 (23.1)	12 (17.9)

Discussion

Principal Findings

This nationwide waitlist RCT examined the efficacy of Mika, an app-based digital therapeutic that provides a personalized supportive intervention for patients with cancer. Participants who had access to the Mika app for 12 weeks showed significant improvements in perceived distress (ie, the primary outcome) and symptoms of anxiety, depression, and fatigue (ie, the secondary outcomes) compared to participants who received UC. The observed treatment effects were similar in the ITT and PP populations but more pronounced in the PP population, indicating the overall robustness of the findings. We observed no group difference in the QoL after 12 weeks. Intervention adherence was good, and no adverse reactions or side effects of the investigated digital therapeutic were reported.

Comparison With Prior Work

While a growing body of research shows evidence of the efficacy of app-based interventions for oncological populations on distress, fatigue, anxiety, and depression [20,25,27,31,51], this is the first study to examine the efficacy of a single holistic app-based digital therapeutic based on multiple intervention modules on these patient-relevant outcomes. Although the improvement in the primary outcome was modest, it reflects the nuanced nature of psycho-oncological interventions, where even modest changes can have significant clinical relevance. Furthermore, we conducted comprehensive testing of the effects of the investigated digital therapeutic on patients with cancer across all tumor entities, using a larger sample size compared to most previous studies [25,27,31,51].

In contrast to the findings of this study, however, other studies found an effect of app-based supportive interventions on QoL [23,27,31]. This difference in findings could be due to differences in the operationalization and measurement of QoL. In this study, participants' global QoL was assessed using a single-item questionnaire after a 12-week intervention period. However, global QoL has been shown to be less affected in patients with cancer compared to specific components of QoL, such as social or cognitive functioning and symptom burden from fatigue or insomnia [52]. Further research using different QoL assessment tools could provide more insights into the efficacy of the investigated digital therapeutic on specific aspects of QoL.

A significant level of intervention adherence and engagement with the digital therapeutic, with varying degrees of interaction across the different app modules, indicates good acceptability and perceived subjective benefit of the investigated digital therapeutic and allows for reliable conclusions about its efficacy in oncological settings. The broad range of engagement, as illustrated by the IQRs, underscores the personalized nature of app use, catering to diverse participant needs and preferences. The variability in engagement levels across different app modules highlights the importance of personalizing digital therapeutics to increase adherence and maximize therapeutic effects.

As we evaluated the app intervention holistically, future studies should examine the impact of the app's individual components.

While the dropout rate in the IG was slightly higher than that in the CG, the dropout rate in the IG as well as the overall dropout rate was low compared to other app-based supportive interventions [25,30]. Considering that patients with cancer have been found to have a positive attitude toward digital health [53,54], the findings of this study add to the notion that digital health interventions have the potential to overcome barriers associated with access to supportive care in oncological populations [55].

We found a positive effect of the investigated digital therapeutic on general psychological distress and a broad range of specific

XSL•FO
RenderX

distress-associated parameters. Importantly, improvements in psychological symptoms, that is, depression and anxiety, can also have a positive tertiary preventive effect on cancer progression [5]. The effect sizes in this study ranged from small (ηp^2=0.02) to medium (ηp^2=0.07) in the ITT population and were more pronounced in the PP population (ηp^2=0.05-0.11). The primary outcome improvement, while subtle, aligns with the expected outcomes in psycho-oncological interventions, highlighting the importance of considering the broad spectrum of therapeutic impacts. The medium to large effects observed in secondary end points, together with the primary outcome, illustrate the broad therapeutic impact and highlight the digital therapeutic's capacity to significantly improve key aspects of psychological well-being in patients with cancer. Small-to-medium effect sizes are common in in-person supportive care interventions [6]. Our results also compare well with other app-based supportive care interventions, such as small effect sizes reported for a CBT and psychoeducation self-management apps on fatigue [25] or small to medium effects of a web-based mindfulness-based intervention on anxiety and depression [56]. This is further supported by the results of several systematic reviews [20,21]. The fact that such effect sizes can be achieved with minimal cost and personnel effort via a digital approach further supports the significant potential for accessibility, reach, and impact of digital therapeutics.

Clinical Implications

The multifaceted intervention modules of the investigated digital therapeutic aim to support patients holistically. The investigated digital therapeutic hereby translates widely used evidence-based intervention methods within supportive care, such as symptom monitoring; patient education; modules of CBT, MBSR, and acceptance and commitment therapy; and strength and flexibility training, into a digital format. The intervention modules of the app are designed to help patients learn about their disease and prepare for discussions with clinicians in an informed decision-making process. This may reduce anxiety and insecurities across the cancer trajectory, while empowering patients and strengthening their self-efficacy.

While it is acknowledged that digital therapeutic interventions might not fully replicate the "in-person" experience, the scope and utility of these tools in the realm of oncology are substantial. For instance, a study evaluating a mobile app designed for tracking patient-reported daily activities found that when supervised by a physician, the data collected were more accurate than when used without guidance [57]. Conversely, a music app was equally effective in alleviating pain and anxiety in emergency department patients irrespective of supervision [58]. This suggests that certain interventions, such as symptom tracking, might be more prone to inaccuracies without proper guidance than passive activities, such as listening to music. In addition, CBT, which is traditionally the most effective in face-to-face settings, has generated interest in the digital domain. A study on the digital adaptation of mindfulness-based cognitive therapy for patients with cancer experiencing distress found the therapeutic connection between therapist and patient to be as potent as in in-person sessions [59]. This underlines the evolving role of digital therapeutics and its potential to reshape therapeutic avenues in oncology, thus paving the way for enhanced patient care.

Furthermore, considering the increasing number of patients with cancer experiencing psychosocial distress and the limited availability of health care professionals, digital therapeutics could present scalable and cost-effective solutions. These solutions can address symptoms and bolster the quality and accessibility of supportive care [55,60,61]. Recognizing patients' diverse needs, tools such as the Mika app leverage artificial intelligence to deliver real-time, tailored support. This has the potential to benefit a broad spectrum of patients with cancer globally while also reducing the pressure on health care infrastructure and professionals. Therefore, digital therapeutics offer a patient-focused approach that is adaptable to specific clinical and lifestyle challenges such as disease management, emotional support, and health-related determinants. They might also further enhance medication adherence, tolerance to chemotherapy, and overall survival rate in the cancer care continuum [15]. Incorporating these digital tools into routine oncological supportive care can augment patient-centric care and enrich patient experience, safety, and interactions with clinicians [15,61]. However, while there is a consensus among medical professionals and stakeholders regarding the revolutionary potential of digital health in addressing cancer treatment challenges, the path to universal adoption remains intricate. Future studies should delve into the assimilation of digital therapeutics, such as Mika, into standard care across varied clinical environments and evaluate hurdles such as digital literacy and the acceptance of digital tools by both patients and health care professionals [62-64].

Strengths and Limitations

The main strength of this study was the app itself. It addresses the overreaching problem areas faced by all patients with cancer while providing tailored support for population-specific areas of burden (ie, cancer type, treatment status, and use behavior). Its flexible and easily accessible use allows for seamless integration into patients' daily lives and continuity of supportive treatment. In addition, the low overall dropout rate and data monitoring led to very little missing data. Similar findings in the ITT, PP, and extended ITT populations suggest overall robustness of the results.

This study has several limitations. First, the web-based recruitment procedure may have led to study registration from patients with cancer who were particularly motivated, digitally literate, and highly functioning in seeking support during their cancer journey, which may limit the generalizability of the study. However, the use of additional recruitment pathways (support groups and participant pool) likely resulted in the recruitment of a more heterogeneous sample, possibly compensating for potential selection bias. Future studies might investigate the impact of various recruitment channels on the efficacy of digital therapeutics, and thus, which population may be particularly responsive to digital interventions. Second, the higher number of dropouts in the IG compared to the CG may reflect treatment dissatisfaction or lost interest in the treatment of some participants, potentially confounding the study's results. Dropouts, who are more likely to show elevated levels of

XSL•FO
RenderX

distress, may have been made aware of the increased need for support through the intervention modules. Patients with clinically significant levels of distress or mental disorders might have accessed support services with more guidance from a health care professional, such as psychotherapy or psycho-oncological counseling. However, no side effects or adverse events were reported in the IG, and the overall robust pattern of results in the ITT, PP, and extended ITT populations suggests a low risk of attrition bias. The fact that participants who dropped out of the study showed higher baseline distress levels may have led to an underestimation of the intervention effect as higher baseline distress levels predicted a greater change in outcome after treatment. Third, due to the nature of the intervention, the group allocation could not be blinded. While experimenter bias was reduced due to a predefined, standardized monitoring procedure and statistical analysis plan, IG participants may have anticipated potential effects. Fourth, the intervention, along with its adherence, was assessed as a whole, which requires the evaluation of specific modules and any potential dose-response relationship in the future. In addition, there was no specific measure to evaluate the subjective usefulness or satisfaction with the digital therapeutic under investigation. Incorporating such a measure could have provided targeted insights into the participants' perceptions and experiences with the app. However, the observed use behavior, characterized by participants repeatedly accessing the app and actively engaging with its content, may serve as an indirect indicator of the app's value to the participants. Future studies should aim to validate this interpretation. Finally, the study sample included participants with a wide variety of cancer diagnoses, which did not allow for the examination of diagnosis-specific intervention effects. However, the sample composition is consistent with the target population of the investigated digital therapeutic, which includes patients with cancer of all entities, and strengthens the study's generalizability and clinical utility. Moreover, a large body of data shows that while variables such as cancer type, treatment status, disease progression, and sex may influence the magnitude of treatment response to supportive therapy, the beneficial effects of supportive therapy are present across various cancer subpopulations [65-67]. In addition, there is a consensus that psychosocial support needs to be integrated into routine cancer care for all cancer types [68,69].

Conclusions

In summary, this RCT demonstrated that Mika, an app-based digital therapeutic that provides a personalized supportive care intervention, can effectively reduce psychological distress and further alleviate symptoms of anxiety, depression, and fatigue in patients with cancer. Digital therapeutics, such as Mika, deliver easily accessible, patient-centered, and effective psychosocial and self-management support for patients with cancer across the course of the disease. Digital therapeutics may present scalable solutions to support patients with cancer worldwide and thus help fill the supportive care gap. Further research is needed to explore the integration of Mika into routine cancer care and its efficacy in diverse clinical settings.

Acknowledgments

The clinical trial was funded by Fosanis GmbH, Berlin.

Data Availability

The data set generated during and analyzed during this study, including individual participant data that underlie the results reported in this article after deidentification (text, tables, and figures), clinical study report, informed consent form, and analytic code, are available from AMT beginning 3 months and ending 5 years following article publication. Access to the data will be granted to investigators whose proposed use of the data has been approved by an independent review committee identified for this purpose, for individual patient data meta-analysis. Proposals for accessing the data may be submitted up to 36 months following article publication.

Authors' Contributions

JSR and GF provided financial support. FS, AM, and HB provided administrative support, with FS also contributing to the collection and assembly of data. MF contributed to data curation. FS, AM, and MF contributed to data analysis and interpretation. FS and AM equally contributed to writing the original draft. All authors contributed to reviewing and editing the draft and provided final approval of the manuscript. AMT, FS, and JSR contributed to the conception and design.

Conflicts of Interest

FS, MF, and HB received research funding for this trial from Fosanis GmbH, which was paid to their institution. AM is an employee at the company Fosanis GmbH. JSR and GF work for the company Fosanis GmbH. They are the managing directors and board members of Fosanis GmbH and own shares of Fosanis GmbH. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1263 KB - jmir_v26i1e51949_app1.pdf]

References

XSL•FO
RenderX

1. Pachman DR, Barton DL, Swetz KM, Loprinzi CL. Troublesome symptoms in cancer survivors: fatigue, insomnia, neuropathy, and pain. J Clin Oncol 2012 Oct 20;30(30):3687-3696. [doi: 10.1200/JCO.2012.41.7238] [Medline: 23008320]
2. Mehnert A, Hartung TJ, Friedrich M, Vehling S, Brähler E, Härter M, et al. One in two cancer patients is significantly distressed: prevalence and indicators of distress. Psychooncology 2018 Jan 16;27(1):75-82. [doi: 10.1002/pon.4464] [Medline: 28568377]
3. Mitchell AJ, Chan M, Bhatti H, Halton M, Grassi L, Johansen C, et al. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. Lancet Oncol 2011 Feb;12(2):160-174. [doi: 10.1016/S1470-2045(11)70002-X] [Medline: 21251875]
4. Meggiolaro E, Berardi MA, Andritsch E, Nanni MG, Sirgo A, Samorì E, et al. Cancer patients' emotional distress, coping styles and perception of doctor-patient interaction in European cancer settings. Pall Supp Care 2015 Jul 09;14(3):204-211. [doi: 10.1017/s1478951515000760] [Medline: 26155817]
5. Brown KW, Levy AR, Rosberger Z, Edgar L. Psychological distress and cancer survival: a follow-up 10 years after diagnosis. Psychosom Med 2003;65(4):636-643. [doi: 10.1097/01.psy.0000077503.96903.a6] [Medline: 12883115]
6. Faller H, Schuler M, Richard M, Heckl U, Weis J, Küffner R. Effects of psycho-oncologic interventions on emotional distress and quality of life in adult patients with cancer: systematic review and meta-analysis. J Clin Oncol 2013 Feb 20;31(6):782-793. [doi: 10.1200/JCO.2011.40.8922] [Medline: 23319686]
7. Epstein RM, Street RL. Patient-centered communication in cancer care: promoting healing and reducing suffering. National Cancer Institute. 2007. URL: https://cancercontrol.cancer.gov/sites/default/files/2020-06/pcc_monograph.pdf [accessed 2024-03-05]
8. Alcalde Castro M, Chavarri Guerra Y, Ramos-Lopez WA, Covarrubias-Gómez A, Sanchez S, Quiroz P, et al. Patient-reported barriers for accessing supportive care among patients with metastatic cancer treated at a public cancer center in Mexico. J Clin Oncol 2018 Dec 01;36(34_suppl):124. [doi: 10.1200/jco.2018.36.34_suppl.124]
9. Carrieri D, Peccatori FA, Boniolo G. Supporting supportive care in cancer: the ethical importance of promoting a holistic conception of quality of life. Crit Rev Oncol Hematol 2018 Nov;131:90-95. [doi: 10.1016/j.critrevonc.2018.09.002] [Medline: 30293711]
10. Kumar P, Casarett D, Corcoran A, Desai K, Li Q, Chen J, et al. Utilization of supportive and palliative care services among oncology outpatients at one academic cancer center: determinants of use and barriers to access. J Palliat Med 2012 Aug;15(8):923-930 [FREE Full text] [doi: 10.1089/jpm.2011.0217] [Medline: 22731514]
11. Bellas O, Kemp E, Edney L, Oster C, Roseleur J. The impacts of unmet supportive care needs of cancer survivors in Australia: a qualitative systematic review. Eur J Cancer Care (Engl) 2022 Nov 12;31(6):e13726. [doi: 10.1111/ecc.13726] [Medline: 36226338]
12. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019 Jan 08;69(1):7-34 [FREE Full text] [doi: 10.3322/caac.21551] [Medline: 30620402]
13. Atun R, Cavalli F. The global fight against cancer: challenges and opportunities. Lancet 2018 Feb;391(10119):412-413. [doi: 10.1016/s0140-6736(18)30156-9]
14. Penedo FJ, Oswald LB, Kronenfeld JP, Garcia SF, Cella D, Yanez B. The increasing value of eHealth in the delivery of patient-centred cancer care. Lancet Oncol 2020 May;21(5):e240-e251 [FREE Full text] [doi: 10.1016/S1470-2045(20)30021-8] [Medline: 32359500]
15. Gussoni G, Ravot E, Zecchina M, Recchia G, Santoro E, Ascione R, et al. Digital therapeutics in oncology: findings, barriers and prospects. A narrative review. Ann Res Oncol 2022 Feb;02(01):55 [FREE Full text] [doi: 10.48286/aro.2022.39]
16. Yang J, Weng L, Chen Z, Cai H, Lin X, Hu Z, et al. Development and testing of a mobile app for pain management among cancer patients discharged from hospital treatment: randomized controlled trial. JMIR Mhealth Uhealth 2019 May 29;7(5):e12542 [FREE Full text] [doi: 10.2196/12542] [Medline: 31144672]
17. Ghanbari E, Yektatalab S, Mehrabi M. Effects of psychoeducational interventions using mobile apps and mobile-based online group discussions on anxiety and self-esteem in women with breast cancer: randomized controlled trial. JMIR Mhealth Uhealth 2021 May 18;9(5):e19262 [FREE Full text] [doi: 10.2196/19262] [Medline: 34003138]
18. Hou IC, Lin HY, Shen SH, Chang KJ, Tai HC, Tsai AJ, et al. Quality of life of women after a first diagnosis of breast cancer using a self-management support mHealth app in Taiwan: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 04;8(3):e17084 [FREE Full text] [doi: 10.2196/17084] [Medline: 32130181]
19. Keum J, Chung MJ, Kim Y, Ko H, Sung MJ, Jo JH, et al. Usefulness of smartphone apps for improving nutritional status of pancreatic cancer patients: randomized controlled trial. JMIR Mhealth Uhealth 2021 Aug 31;9(8):e21088 [FREE Full text] [doi: 10.2196/21088] [Medline: 34463630]
20. Hernandez Silva E, Lawler S, Langbecker D. The effectiveness of mHealth for self-management in improving pain, psychological distress, fatigue, and sleep in cancer survivors: a systematic review. J Cancer Surviv 2019 Feb;13(1):97-107. [doi: 10.1007/s11764-018-0730-8] [Medline: 30635865]
21. Matis J, Svetlak M, Slezackova A, Svoboda M, Šumec R. Mindfulness-based programs for patients with cancer via eHealth and mobile health: systematic review and synthesis of quantitative research. J Med Internet Res 2020 Nov 16;22(11):e20709 [FREE Full text] [doi: 10.2196/20709] [Medline: 33196452]

XSL•FO
RenderX

22. Adriaans DJ, Dierick-van Daele AT, van Bakel MJ, Nieuwenhuijzen GA, Teijink JA, Heesakkers FF, et al. Digital self-management support tools in the care plan of patients with cancer: review of randomized controlled trials. J Med Internet Res 2021 Jun 29;23(6):e20861 [FREE Full text] [doi: 10.2196/20861] [Medline: 34184997]
23. Qin M, Chen B, Sun S, Liu X. Effect of mobile phone app-based interventions on quality of life and psychological symptoms among adult cancer survivors: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Dec 19;24(12):e39799 [FREE Full text] [doi: 10.2196/39799] [Medline: 36534460]
24. Springer F, Mehnert-Theuerkauf A. Content features and its implementation in novel app-based psycho-oncological interventions for cancer survivors: a narrative review. Curr Opin Oncol 2022 Jul 01;34(4):313-319. [doi: 10.1097/CCO.0000000000000836] [Medline: 35837701]
25. Spahrkäs SS, Looijmans A, Sanderman R, Hagedoorn M. Beating cancer-related fatigue with the Untire mobile app: results from a waiting-list randomized controlled trial. Psychooncology 2020 Nov;29(11):1823-1834 [FREE Full text] [doi: 10.1002/pon.5492] [Medline: 33393199]
26. Mayer DK, Landucci G, Awoyinka L, Atwood AK, Carmack CL, Demark-Wahnefried W, et al. SurvivorCHESS to increase physical activity in colon cancer survivors: can we get them moving? J Cancer Surviv 2018 Feb 9;12(1):82-94 [FREE Full text] [doi: 10.1007/s11764-017-0647-7] [Medline: 28994035]
27. Lengacher CA, Reich RR, Ramesar S, Alinat CB, Moscoso M, Cousin L, et al. Feasibility of the mobile mindfulness-based stress reduction for breast cancer (mMBSR(BC)) program for symptom improvement among breast cancer survivors. Psychooncology 2018 Feb;27(2):524-531 [FREE Full text] [doi: 10.1002/pon.4491] [Medline: 28665541]
28. Mikolasek M, Witt CM, Barth J. Effects and implementation of a mindfulness and relaxation app for patients with cancer: mixed methods feasibility study. JMIR Cancer 2021 Jan 13;7(1):e16785 [FREE Full text] [doi: 10.2196/16785] [Medline: 33439132]
29. Gustafson DH, DuBenske LL, Atwood AK, Chih MY, Johnson RA, McTavish F, et al. Reducing symptom distress in patients with advanced cancer using an e-alert system for caregivers: pooled analysis of two randomized clinical trials. J Med Internet Res 2017 Nov 14;19(11):e354 [FREE Full text] [doi: 10.2196/jmir.7466] [Medline: 29138131]
30. Chung IY, Jung M, Park YR, Cho D, Chung H, Min YH, et al. Exercise promotion and distress reduction using a mobile app-based community in breast cancer survivors. Front Oncol 2019;9:1505 [FREE Full text] [doi: 10.3389/fonc.2019.01505] [Medline: 31998651]
31. Greer JA, Jacobs J, Pensak N, MacDonald JJ, Fuh C, Perez GK, et al. Randomized trial of a tailored cognitive-behavioral therapy mobile application for anxiety in patients with incurable cancer. Oncologist 2019 Aug 25;24(8):1111-1120 [FREE Full text] [doi: 10.1634/theoncologist.2018-0536] [Medline: 30683710]
32. Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol 2016 Feb 20;34(6):557-565 [FREE Full text] [doi: 10.1200/JCO.2015.63.0830] [Medline: 26644527]
33. Greer JA, Park ER, Prigerson HG, Safren SA. Tailoring cognitive-behavioral therapy to treat anxiety comorbid with advanced cancer. J Cogn Psychother 2010 Jan 01;24(4):294-313 [FREE Full text] [doi: 10.1891/0889-8391.24.4.294] [Medline: 21234281]
34. Carlson LE, Speca M, Faris P, Patel KD. One year pre-post intervention follow-up of psychological, immune, endocrine and blood pressure outcomes of mindfulness-based stress reduction (MBSR) in breast and prostate cancer outpatients. Brain Behav Immun 2007 Nov;21(8):1038-1049. [doi: 10.1016/j.bbi.2007.04.002] [Medline: 17521871]
35. Speca M, Carlson LE, Goodey E, Angen M. A randomized, wait-list controlled clinical trial: the effect of a mindfulness meditation-based stress reduction program on mood and symptoms of stress in cancer outpatients. Psychosom Med 2000;62(5):613-622. [doi: 10.1097/00006842-200009000-00004] [Medline: 11020090]
36. Winters-Stone KM, Dobek J, Nail L, Bennett JA, Leo MC, Naik A, et al. Strength training stops bone loss and builds muscle in postmenopausal breast cancer survivors: a randomized, controlled trial. Breast Cancer Res Treat 2011 Jun 19;127(2):447-456 [FREE Full text] [doi: 10.1007/s10549-011-1444-z] [Medline: 21424279]
37. Adam R, Bond C, Murchie P. Educational interventions for cancer pain. A systematic review of systematic reviews with nested narrative review of randomized controlled trials. Patient Educ Couns 2015 Mar;98(3):269-282. [doi: 10.1016/j.pec.2014.11.003] [Medline: 25483575]
38. Wolff J, Stupin J, Olschewski J, Pirmorady Sehouli A, Maier A, Fofana M, et al. Digital therapeutic to improve cancer-related well-being: a pilot randomized controlled trial. Int J Gynecol Cancer 2023 Jul 03;33(7):1118-1124. [doi: 10.1136/ijgc-2023-004304] [Medline: 37321673]
39. Harrison SE, Watson EK, Ward AM, Khan NF, Turner D, Adams E, et al. Primary health and supportive care needs of long-term cancer survivors: a questionnaire survey. J Clin Oncol 2011 May 20;29(15):2091-2098. [doi: 10.1200/JCO.2010.32.5167] [Medline: 21519023]
40. Xiao J, Ye H, He X, Zhang H, Wu F, Chua TS. Attentional factorization machines: learning the weight of feature interactions via attention networks. arXiv. Preprint posted online August 15, 2017 2017 [FREE Full text] [doi: 10.48550/arXiv.1708.04617]

XSL•FO
RenderX

41. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Proceedings of the 32nd Conference on Neural Information Processing Systems. 2018 Presented at: NeurIPS '18; December 3-8, 2018; Montreal, QC URL: https://www.proceedings.com/48413.html
42. Mehnert A, Müller D, Lehmann C, Koch U. Die Deutsche version des NCCN distress-thermometers. Z Für Psychiatr Psychol Psychother 2006 Jan;54(3):213-223. [doi: 10.1024/1661-4747.54.3.213]
43. National Comprehensive Cancer Network. Distress management. Clinical practice guidelines. J Natl Compr Canc Netw 2003 Jul 01;1(3):344-374. [doi: 10.6004/jnccn.2003.0031] [Medline: 19761069]
44. Petermann F. Hospital anxiety and depression scale, Deutsche version (HADS-D). Z Für Psychiatr Psychol Psychother 2011 Jul;59(3):251-253. [doi: 10.1024/1661-4747/a000077]
45. Lai J, Cella D, Chang C, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003 Aug;12(5):485-501. [doi: 10.1023/a:1025014509626] [Medline: 13677494]
46. Guy W. Clinical global impressions scale. Psychiatry 1976. [doi: 10.1037/t48216-000]
47. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Oxfordshire, UK: Routledge; 2013.
48. Carpenter JR, Roger JH, Kenward MG. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. J Biopharm Stat 2013 Oct 18;23(6):1352-1371. [doi: 10.1080/10543406.2013.834911] [Medline: 24138436]
49. Enders CK. Applied Missing Data Analysis. New York, NY: The Guilford Press; 2022.
50. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.R-project.org/ [accessed 2024-03-05]
51. Ham K, Chin S, Suh YJ, Rhee M, Yu ES, Lee HJ, et al. Preliminary results from a randomized controlled study for an app-based cognitive behavioral therapy program for depression and anxiety in cancer patients. Front Psychol 2019 Jul 25;10:1592 [FREE Full text] [doi: 10.3389/fpsyg.2019.01592] [Medline: 31402881]
52. Hinz A, Mehnert A, Dégi C, Reissmann DR, Schotte D, Schulte T. The relationship between global and specific components of quality of life, assessed with the EORTC QLQ-C30 in a sample of 2019 cancer patients. Eur J Cancer Care (Engl) 2017 Mar 16;26(2):e12416. [doi: 10.1111/ecc.12416] [Medline: 26568527]
53. Jansen F, van Uden-Kraan CF, van Zwieten V, Witte BI, Verdonck-de Leeuw IM. Cancer survivors' perceived need for supportive care and their attitude towards self-management and eHealth. Support Care Cancer 2015 Jun 26;23(6):1679-1688. [doi: 10.1007/s00520-014-2514-7] [Medline: 25424520]
54. Kessel KA, Vogel MM, Kessel C, Bier H, Biedermann T, Friess H, et al. Mobile health in oncology: a patient survey about app-assisted cancer care. JMIR Mhealth Uhealth 2017 Jun 14;5(6):e81 [FREE Full text] [doi: 10.2196/mhealth.7689] [Medline: 28615159]
55. Marthick M, McGregor D, Alison J, Cheema B, Dhillon H, Shaw T. Supportive care interventions for people with cancer assisted by digital technology: systematic review. J Med Internet Res 2021 Oct 29;23(10):e24722 [FREE Full text] [doi: 10.2196/24722] [Medline: 34714246]
56. Nissen ER, O'Connor M, Kaldo V, Højris I, Borre M, Zachariae R, et al. Internet-delivered mindfulness-based cognitive therapy for anxiety and depression in cancer survivors: a randomized controlled trial. Psychooncology 2020 Jan 18;29(1):68-75 [FREE Full text] [doi: 10.1002/pon.5237] [Medline: 31600414]
57. Egbring M, Far E, Roos M, Dietrich M, Brauchbar M, Kullak-Ublick GA, et al. A mobile app to stabilize daily functional activity of breast cancer patients in collaboration with the physician: a randomized controlled clinical trial. J Med Internet Res 2016 Sep 06;18(9):e238 [FREE Full text] [doi: 10.2196/jmir.6414] [Medline: 27601354]
58. Chai PR, Schwartz E, Hasdianda MA, Azizoddin DR, Kikut A, Jambaulikar GD, et al. A brief music app to address pain in the emergency department: prospective study. J Med Internet Res 2020 May 20;22(5):e18537 [FREE Full text] [doi: 10.2196/18537] [Medline: 32432550]
59. Bisseling E, Cillessen L, Spinhoven P, Schellekens M, Compen F, van der Lee M, et al. Development of the therapeutic alliance and its association with internet-based mindfulness-based cognitive therapy for distressed cancer patients: secondary analysis of a multicenter randomized controlled trial. J Med Internet Res 2019 Oct 18;21(10):e14065 [FREE Full text] [doi: 10.2196/14065] [Medline: 31628791]
60. Parikh RB, Basen-Enquist KM, Bradley C, Estrin D, Levy M, Lichtenfeld JL, et al. Digital health applications in oncology: an opportunity to seize. J Natl Cancer Inst 2022 Oct 06;114(10):1338-1339 [FREE Full text] [doi: 10.1093/jnci/djac108] [Medline: 35640986]
61. Aapro M, Bossi P, Dasari A, Fallowfield L, Gascón P, Geller M, et al. Digital health for optimal supportive care in oncology: benefits, limits, and future perspectives. Support Care Cancer 2020 Oct 12;28(10):4589-4612 [FREE Full text] [doi: 10.1007/s00520-020-05539-1] [Medline: 32533435]
62. Rankin NM, Butow PN, Thein T, Robinson T, Shaw JM, Price MA, et al. Everybody wants it done but nobody wants to do it: an exploration of the barrier and enablers of critical components towards creating a clinical pathway for anxiety and depression in cancer. BMC Health Serv Res 2015 Jan 22;15(1):28 [FREE Full text] [doi: 10.1186/s12913-015-0691-9] [Medline: 25608947]

XSL•FO
RenderX

63. Leader AE, Capparella LM, Waldman LB, Cammy RB, Petok AR, Dean R, et al. Digital literacy at an urban cancer center: implications for technology use and vulnerable patients. JCO Clinical Cancer Informatics 2021 Dec(5):872-880. [doi: 10.1200/cci.21.00039]
64. den Bakker CM, Schaafsma FG, Huirne JA, Consten EC, Stockmann HB, Rodenburg CJ, et al. Cancer survivors' needs during various treatment phases after multimodal treatment for colon cancer - is there a role for eHealth? BMC Cancer 2018 Dec 04;18(1):1207 [FREE Full text] [doi: 10.1186/s12885-018-5105-z] [Medline: 30514325]
65. Cillessen L, Johannsen M, Speckens AE, Zachariae R. Mindfulness-based interventions for psychological and physical health outcomes in cancer patients and survivors: a systematic review and meta-analysis of randomized controlled trials. Psychooncology 2019 Dec 11;28(12):2257-2269 [FREE Full text] [doi: 10.1002/pon.5214] [Medline: 31464026]
66. Tauber NM, O'Toole MS, Dinkel A, Galica J, Humphris G, Lebel S, et al. Effect of psychological intervention on fear of cancer recurrence: a systematic review and meta-analysis. J Clin Oncol 2019 Nov 01;37(31):2899-2915 [FREE Full text] [doi: 10.1200/JCO.19.00572] [Medline: 31532725]
67. Merluzzi TV, Pustejovsky JE, Philip EJ, Sohl SJ, Berendsen M, Salsman JM. Interventions to enhance self-efficacy in cancer patients: a meta-analysis of randomized controlled trials. Psychooncology 2019 Sep 09;28(9):1781-1790 [FREE Full text] [doi: 10.1002/pon.5148] [Medline: 31206917]
68. Jacobsen PB, Wagner LI. A new quality standard: the integration of psychosocial care into routine cancer care. J Clin Oncol 2012 Apr 10;30(11):1154-1159. [doi: 10.1200/JCO.2011.39.5046] [Medline: 22412134]
69. Fann JR, Ell K, Sharpe M. Integrating psychosocial care into cancer services. J Clin Oncol 2012 Apr 10;30(11):1178-1186. [doi: 10.1200/JCO.2011.39.7398] [Medline: 22412139]

Abbreviations

CBT: cognitive behavioral therapy
CG: control group
IG: intervention group
ITT: intention-to-treat
MBSR: mindfulness-based stress reduction
PP: per-protocol
QoL: quality of life
RCT: randomized controlled trial
UC: usual care

Edited by YH Lin; submitted 21.08.23; peer-reviewed by F Denis, A Haussmann, N Schaeffeler, P Chow; comments to author 24.01.24; accepted 23.02.24; published 25.04.24.

Please cite as:
Springer F, Maier A, Friedrich M, Raue JS, Finke G, Lordick F, Montgomery G, Esser P, Brock H, Mehnert-Theuerkauf A
Digital Therapeutic (Mika) Targeting Distress in Patients With Cancer: Results From a Nationwide Waitlist Randomized Controlled Trial
J Med Internet Res 2024;26:e51949
URL: https://www.jmir.org/2024/1/e51949
doi:10.2196/51949
PMID:38663007

©Franziska Springer, Ayline Maier, Michael Friedrich, Jan Simon Raue, Gandolf Finke, Florian Lordick, Guy Montgomery, Peter Esser, Hannah Brock, Anja Mehnert-Theuerkauf. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Cognitive-Physical Intervention for Attention-Deficit/Hyperactivity Disorder: Randomized Controlled Trial

Licong Zhao[1], MD; Heather Agazzi[2], MS, PhD; Yasong Du[3], MD, PhD; Hongdao Meng[4], MD, PhD; Renya Maku[5], BS; Ke Li[1], MD; Peter Aspinall[6], PhD; Cynthia Wilson Garvan[7], PhD; Shuanfeng Fang[1], MD

[1]Department of Child Healthcare, Children's Hospital Affiliated to Zhengzhou University, Zhengzhou, China

[2]Department of Pediatrics & Department of Psychiatry and Behavioral Neurosciences, College of Medicine, University of South Florida, Tampa, FL, United States

[3]Department of Child & Adolescent Psychiatry, Shanghai Mental Health Center, Shanghai Jiaotong University, Shanghai, China

[4]College of Behavioral & Community Sciences, University of South Florida, Tampa, FL, United States

[5]College of Public Health, University of South Florida, Tampa, FL, United States

[6]Recovery Plus, Chengdu, China

[7]Department of Anesthesiology, College of Medicine, University of Florida, Tampa, FL, United States

Corresponding Author:
Shuanfeng Fang, MD
Department of Child Healthcare
Children's Hospital Affiliated to Zhengzhou University
33 Longhu Waihuan East Road
Zhengdong New District
Zhengzhou, 450018
China
Phone: 86 0371 85515939
Email: fangshuanfeng@126.com

Abstract

Background: Attention-deficit/hyperactivity disorder (ADHD) is one of the most common neurodevelopmental disorders among children. Pharmacotherapy has been the primary treatment for ADHD, supplemented by behavioral interventions. Digital and exercise interventions are promising nonpharmacologic approaches for enhancing the physical and psychological health of children with ADHD. However, the combined impact of digital and exercise therapies remains unclear.

Objective: The aim of this study was to determine whether BrainFit, a novel digital intervention combining gamified cognitive and exercise training, is efficacious in reducing ADHD symptoms and executive function (EF) among school-aged children with ADHD.

Methods: This 4-week prospective randomized controlled trial included 90 children (6-12 years old) who visited the ADHD outpatient clinic and met the diagnostic criteria for ADHD. The participants were randomized (1:1) to the BrainFit intervention (n=44) or a waitlist control (n=46) between March and August 2022. The intervention consisted of 12 30-minute sessions delivered on an iPad over 4 weeks with 3 sessions per week (Monday, Wednesday, and Friday after school) under the supervision of trained staff. The primary outcomes were parent-rated symptoms of attention and hyperactivity assessed according to the Swanson, Nolan, and Pelham questionnaire (SNAP-IV) rating scale and EF skills assessed by the Behavior Rating Inventory of Executive Function (BRIEF) scale, evaluated pre and post intervention. Intention-to-treat analysis was performed on 80 children after attrition. A nonparametric resampling-based permutation test was used for hypothesis testing of intervention effects.

Results: Among the 145 children who met the inclusion criteria, 90 consented and were randomized; ultimately, 80 (88.9%) children completed the study and were included in the analysis. The participants' average age was 8.4 (SD 1.3) years, including 63 (78.8%) male participants. The most common ADHD subtype was hyperactive/impulsive (54/80, 68%) and 23 (29%) children had severe symptoms. At the endpoint of the study, the BrainFit intervention group had a significantly larger improvement in total ADHD symptoms (SNAP-IV total score) as compared to those in the control group (β=–12.203, 95% CI –17.882 to –6.523; P<.001), owing to lower scores on the subscales Inattention (β=–3.966, 95% CI –6.285 to –1.647; P<.001), Hyperactivity/Impulsivity (β=–5.735, 95% CI –8.334 to –3.137; P<.001), and Oppositional Defiant Disorder (β=–2.995, 95%

XSL•FO
RenderX

CI –4.857 to –1.132; *P*=.002). The intervention was associated with significant reduction in the Metacognition Index (β=–6.312, 95% CI –10.973 to –1.650; *P*=.006) and Global Executive Composite (β=–5.952, 95% CI –10.214 to –1.690; *P*=.003) on the BRIEF. No severe intervention-related adverse events were reported.

Conclusions: This novel digital cognitive-physical intervention was efficacious in school-age children with ADHD. A larger multicenter effectiveness trial with longer follow-up is warranted to confirm these findings and to assess the durability of treatment effects.

Trial Registration: Chinese Clinical Trial Register ChiCTR2300070521; https://www.chictr.org.cn/showproj.html?proj=177806

(J Med Internet Res 2024;26:e55569) doi:10.2196/55569

KEYWORDS

school-age children; cognitive training; exercise therapy; gamification; ADHD; attention deficit; attention-deficit/hyperactivity disorder; RCT; randomized controlled trial; executive function; digital intervention; AR; augmented reality

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is the most common neurodevelopmental disorder among school-age children [1], with an estimated prevalence of 8.9% in the United States [2] and 6.5% in China [3,4]. The core symptoms of ADHD are characterized by inattention, impulsivity, and/or hyperactivity, and these symptoms lead to functional impairment across multiple settings [5]. These symptoms often co-occur with psychological disorders (eg, disruptive behavior, conduct disorder, anxiety, and depression). Despite the heavy disease burden of ADHD and long-term negative outcomes (eg, substance use disorders, reduced employment and income, and poor social outcomes) [6], many youth with ADHD do not access treatment to mediate the substantial impact of these disorders on daily functioning.

The first-line treatment for ADHD traditionally includes pharmacotherapy, often complemented by nonpharmacological behavioral interventions [7]. In addition, most children with ADHD benefit from school-based interventions that target on-task behavior, academic skills, and social interactions with peers [8]. While pharmacotherapy has proven to be effective for the short-term treatment of ADHD symptoms and approximately two-thirds of children with a current ADHD diagnosis have taken prescribed medication [3], adherence is affected by patient and caregiver attitudes about medication effectiveness and practical barriers. A study on medication adherence in children, adolescents, and adults with ADHD found that the main factors contributing to poor medication adherence included fear of addiction and doubts about the effectiveness of the medication [9]. In addition, some young people with ADHD stop taking their medication due to the challenges associated with accessing the prescription, including committing to medical appointments every 3 months for refills, paying for the cost of the prescription, the time and cost of missing work and attending doctor visits, and finding new providers when transitioning from pediatric to adult care [10]. Thus, even though medication can improve ADHD symptoms in the short term, many people with ADHD perceive the costs of pharmacotherapy to outweigh the benefits.

Behavioral interventions are recommended as the first-line treatment for ADHD in preschool-aged children (4-5 years) [11]. Indeed, behavioral interventions play an important role in the development of organizational skills and executive function (EF) [12]. These interventions include self-monitoring, positive reinforcement for on-task behavior, a token economy, and chunking tasks while providing frequent feedback on performance [13]. These interventions are often taught in the context of parent management training, which has long been considered a gold-standard intervention to address disruptive behaviors in children [14]. Behavioral interventions for ADHD are important because youth learn coping strategies to improve their developmental trajectory across academic, social, and employment settings, potentially reducing overall functional impairment. In addition, behavioral interventions do not carry the same side effects of medication intervention. While behavioral interventions are very effective for treating symptoms of ADHD, less than half of children with a current diagnosis were reported to have received such treatment before 2019 in the United States [15]. The treatment rates could be much lower in less-developed countries and in rural regions of developed countries. In addition, behavioral and psychosocial interventions require extensive time and dedication of caregivers, often resulting in low adherence rates [16,17]. As with medication intervention, there are barriers to accessing behavioral interventions, including a limited number of providers, limited providers who accept Medicaid insurance in the United States, limited availability of treatment in rural settings, long wait lists, and cash-only services [18].

Digital treatment for ADHD may provide an alternative to address some of the challenges in accessing ADHD services with fewer side effects and a lower likelihood of addiction [19]. Digital interventions that can be delivered via mobile phones, tablets, or web-based platforms may provide a mechanism for behavioral interventions to be more accessible, better integrated across multiple settings (eg, home, school, and health services), and enable individuals (and families) to take care of themselves. Some research has shown that cognitive training targeting ADHD symptoms such as attention, response inhibition, and working memory can be designed to engage users through a computer or mobile phone portal [20,21]. A meta-analysis showed that although cognitive training does improve performance related to some neurophysiological functions, the effect on ADHD symptoms is not yet known [22]. Notably, interventions that included multiple cognitive processes had higher effect sizes than those of training that targeted only a single cognitive function [23].

XSL•FO
RenderX

Emerging evidence indicates that digital therapies can affect brain functions such as increased working memory, processing speed, and concentration, and these treatments were proven to have therapeutic effects in a variety of mental and behavioral disorders [24,25]. Kollins et al [19] developed a digital therapy named AKL-T01, which had a beneficial effect on the average attention measure of children with ADHD compared to that of the control group. For its first-in-class breakthrough, the resulting digital intervention has received approval for clinical use by the US Food and Drug Administration (FDA). These findings suggest that digital therapy may be effective in the nonpharmacological management of ADHD symptoms. Digital therapies have the potential to be conducted at any time or place, which could alleviate the burden on caregivers to organize interventions for their children with ADHD. In turn, this may improve patient adherence to the intervention and lead to lasting therapeutic effects. Importantly, digital therapies do not require substantial professional human resources, which could benefit many less-developed countries and regions.

As a nonpharmacologic therapy, exercise has been shown to improve the physical and psychological health of children with ADHD. In recent years, exercise has been linked to positive neuropsychiatric outcomes such as neurocognitive and executive functions [17,26,27]. For example, a meta-analysis showed that exercise therapy can improve core symptoms and EF in children and adolescents with ADHD, especially with closed-skill training [28]. Based on the advantages of digital therapy and exercise therapy in improving ADHD, Benzing et al [25] developed a digital exercise game and conducted an 8-week intervention test on children with ADHD. The results suggested improvements in EF (eg, reaction time in inhibition and switching) when compared to the control group, as well general improvements in psychopathology and motor abilities. However, evidence regarding whether digital behavioral interventions in conjunction with exercise therapy would confer clinical benefits above and beyond those obtained when each intervention is used separately remains limited.

BrainFit is an interactive augmented reality (AR)–based exercise therapy designed and executed in a digital form that uses a pioneering engine to evaluate, develop, and dynamically optimize skills and abilities in a hierarchical manner. BrainFit was created to improve concentration and other EF skills that are typically impaired in youth with ADHD. A variety of training paradigms integrate digital intervention modules through the medium of touchscreen interaction and AR motion interaction. Figure 1 shows the information on BrainFit example modules. The adaptive training algorithm modifies the levels of task difficulty based on the task performance of the user and thus optimizes engagement.

The purpose of this randomized controlled trial was to examine whether a novel digital intervention combining gamified cognitive training and exercise (BrainFit) would be effective in treating ADHD symptoms among children with ADHD aged 6 to 12 years, as well as to evaluate whether any beneficial effects on the core ADHD symptoms and EF could be observed.

XSL•FO
RenderX

Figure 1. Illustrative example of Brainfit setup and modules. AR: augmented reality.

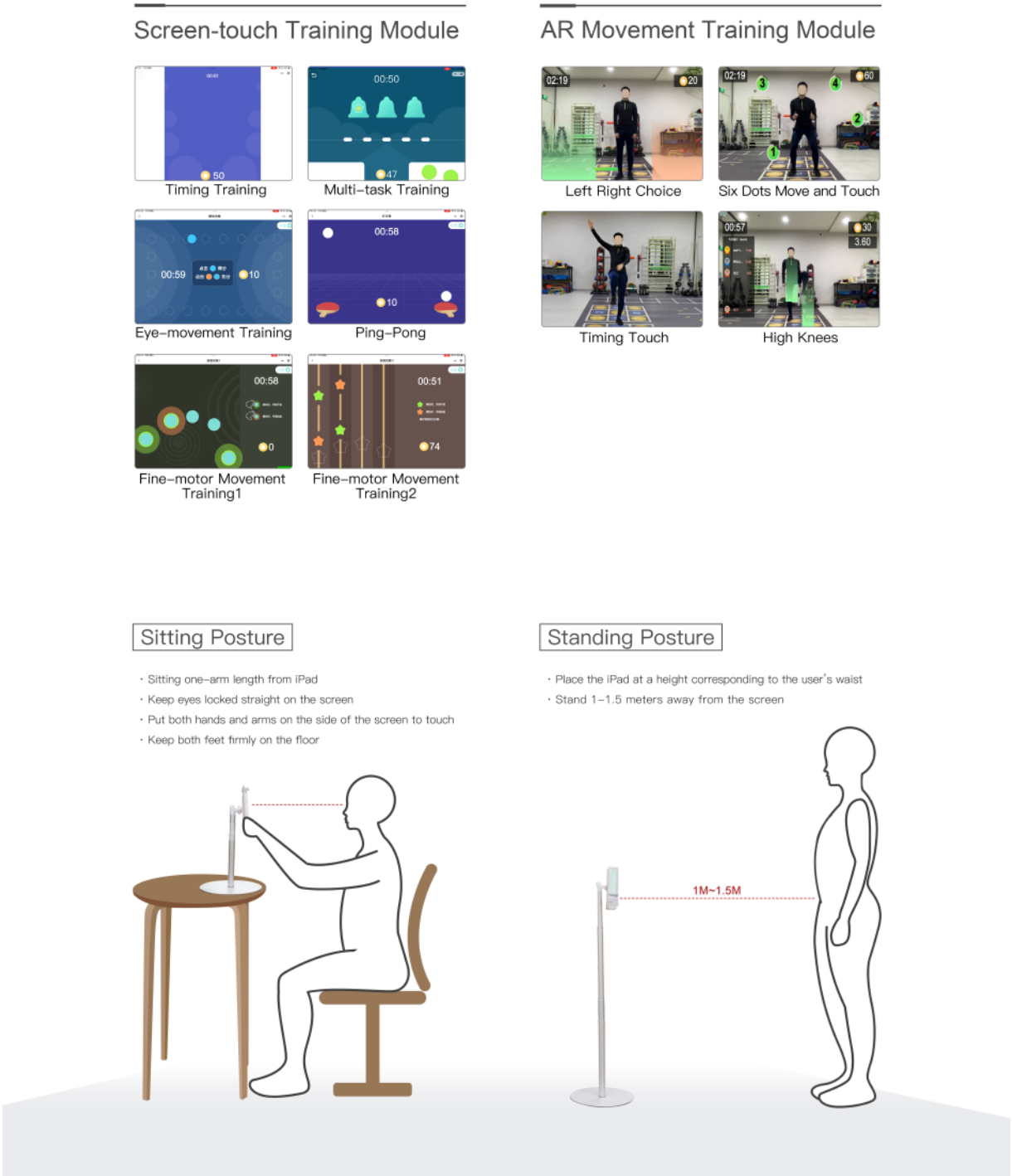

Methods

Study Design

This study was an open-label, randomized, parallel-group trial. The trial was conducted at the Children's Hospital Affiliated to Zhengzhou University (Zhengzhou, Henan, China) from March to August of 2022. All participants were screened for eligibility before being randomly assigned to the BrainFit intervention or control group.

Participants

The participants were recruited from Children's Hospital Affiliated to Zhengzhou University. The inclusion criteria were: (1) aged 6-12 years; (2) ADHD diagnosed by clinicians based on *The Diagnostic and Statistical Manual of Mental Disorders, 5th edition* (DSM-5) [29]; and (3) an IQ score >80, measured by the validated Chinese version of the Wechsler Intelligence Scale for Children, Fourth Edition. Exclusion criteria were: (1) presence of comorbid severe psychiatric or physical disorders that may affect attention (eg, autism spectrum disorder, learning disabilities, abnormal blood glucose, anemia, hyperthyroidism, and hyperadrenocorticism) and (2) current or recent (within the

XSL•FO
RenderX

last 6 months) medication treatment (eg, methylphenidate, atomoxetine, risperidone, modafinil) or (3) behavioral treatment for ADHD.

Randomization

Prior to randomization, participants were informed that they would either be assigned to the 4-week intervention group or the waitlist control group. The randomization schedule was generated by an independent statistician not involved in the clinical trial (using the PROC PLAN procedure in SAS 9.4) and the randomization schema was reproducible. Participants were not blinded to group assignment due to the nature of the digital intervention.

Equipment and Staff Training

The intervention was delivered via iPads preinstalled with the BrainFit software.

Two researchers received training on intervention delivery using BrainFit, which included reviewing the intervention protocol and standardized training manual, practicing intervention delivery steps, and understanding how to respond to potential adverse events. The training material also included a tutorial for BrainFit and a written overview of the study protocol.

Intervention Overview

At the time of diagnosis, participants' parents were provided psychoeducation on ADHD according to Chinese consensus standards for ADHD treatment [30]. The intervention was facilitated by trained researchers and consisted of 12 sessions (30 minutes each) delivered via iPad technology over 4 weeks. Each participant received three sessions per week (Monday, Wednesday, and Friday after school) with appropriate training levels tailored to their individual conditions. The intervention included six modules, delivered in the same sequence, with adaptive difficulty levels to increase engagement. Participants earned scores based on completing tasks correctly and on time, and games were upgraded or downgraded based on their performance. Participants and their caregivers received study session reminders regularly to reduce missing data, missed sessions, and attrition. Participants in the control group received no intervention during the active study period other than the previously mentioned psychoeducation.

Measures

Demographic data were collected from caregivers, including the caregiver report on the child's social skills. ADHD symptoms and EF were assessed as primary outcomes at baseline and at 4 weeks post randomization by experienced trainers. Demographic data included sex (female or male), age (years), school report card–based academic performance (quartiles), living arrangement (living with parents or living with parents and grandparents), parent-rated sociability (low, medium, or high), and referral source (ie, teacher, parent, physician, or other). ADHD subtype and severity were defined according to the DSM-5. ADHD symptoms were assessed using the Chinese version of the Swanson, Nolan, and Pelham Questionnaire (SNAP-IV) 26-item parent rating scale [31], which has satisfactory measurement properties regarding test-retest reliability and validity. The SNAP-IV measures ADHD inattention symptoms (items 1-9) and symptoms of hyperactivity/impulsivity (items 10-18). In addition, items 19-26 assess symptoms of oppositional defiant disorder (ODD). Each item is ranked on a 4-point scale (0="not at all," 2="just a little," 3="quite a bit," and 4="very much"). SNAP-IV is one of the most common methods measuring ADHD symptom severity, in which a higher score indicates more severe symptoms. The Chinese version of the Behavior Rating Inventory of Executive Function (BRIEF) parent form was used to assess EF [32]. The BRIEF has 86 items with each item rated on a 3-point scale (1="never," 2="sometimes," and 3="often"). The BRIEF yields an overall Global Executive Composite (GEC) score, which is composed of two indices: the Behavioral Regulation Index (BRI) and Metacognition Index (MI). The BRI has three clinical scales (Inhibit, Shift, and Emotional Control) and the MI is comprised of five clinical scales (Initiate, Shift, Memory, Plan/Organize, Monitor, and Organization of Materials). The BRIEF has good psychometric properties, including test-retest reliability (intraclass correlation=0.68-0.89), internal consistency (Cronbach α=0.74-0.96), and validity (Pearson correlation coefficient=0.41-0.64) [33]; higher scores indicate worse EF.

Sample Size

In a review of 21 intervention studies and their effect on EF in ADHD, a medium effect size was noted (standard mean difference [SMD] 0.611) [34]. We hypothesized a similar effect in the current study. Power analyses determined that a sample size of at least 30 participants per arm would be sufficient to achieve power of 80% with a Type I error rate of .05. Assuming a 20% dropout rate, a total sample of 80 participants was needed (40 per group).

Adverse Events Reporting

Adverse events were assessed at each session. Participants could also report adverse events by calling a study telephone line. All researchers were trained to record details of any adverse events and a participant would be followed until the event resolved. All potential adverse events were reported and reviewed by the principal investigator. One adverse event was reported wherein a legal guardian withdrew a participant from the study after three sessions due to a fever; this event is unlikely to be related to the intervention.

Statistical Analysis

We used IBM SPSS Statistics for Windows version 23.0 and Stata BE version 17 for all data processing and statistical analyses. We present continuous data as mean (SD) and categorical data as numbers and percentages. To assess any potential differences in participant demographic variables between the intervention and control groups, we used the independent-sample *t* test and Mann-Whitney *U* test for continuous variables and ordinal variables; χ^2 tests were used for comparisons of categorical data. For the multivariable regression analysis of the primary outcomes, SNAP-IV and BRIEF scores were regressed onto the independent variables (treated as dummy variables), including age, sex, whether the participant was the first child, ADHD subtype, academic performance, and sociability ratings. These independent variables were selected based on clinical judgment and the

XSL•FO
RenderX

literature. Controlling for these baseline characteristics in the regression model can generally increase the precision of the estimator for the average treatment effect [35]. All independent variables were entered into the model simultaneously. To obtain unbiased intervention effect estimates and quantify the uncertainty surrounding the point estimates, we used the Fisher-Pitman permutation test for two independent samples. This permutation test is a powerful alternative to the Wilcoxon-Mann-Whitney rank-sum test [36,37]. This approach has the benefit of making no assumptions about the distributional properties of the joint distributions of variables as well as minimizing the false-positive results due to multiple testing with traditional regression-based hypothesis testing [38,39]. Because the exact calculation of the test can be extremely computationally intensive, a Monte Carlo–based algorithm (with 10,000 repetitions) was used to increase efficiency with minimal loss of accuracy. We used an α level of .05 for determining statistical significance. The effect size (Cohen *d*) for each outcome measure was calculated with 95% CIs based on the nonparametric bootstrap method (with 5000 repetitions) to quantify the uncertainty surrounding the point estimates [40].

Ethical Considerations

The study protocol was approved by the Medical Ethics Committee of Henan Children's Hospital (2021-H-H09) before study initiation. The legal guardians of all participants provided written informed consent and children provided assent to participate before baseline assessment. All families were informed that participation was voluntary, that data collected during the study will be kept confidential, and that they can withdraw their consent at any time without negative consequences to their care. Each participant received ¥600 (US $82.80) as compensation for completing all sessions of the intervention. If a participant completed all assessments, the various scale assessments were offered at no cost to the participants. They also receive an explanatory document and consent form from the researchers that included contact information for any inquiries. The personal information of the participants and their guardians was anonymized, and a unique study ID number was assigned to each participant for data entry, management, and analysis. The individual pictured in Figure 1 provided written informed consent to allow their image to be published.

Results

Participants

A total of 199 children were assessed for eligibility and 145 were deemed to be eligible. Among the eligible candidates, 90 children were randomized into either the intervention (n=44) or control (n=46) group. Eighty children completed the study and their data were analyzed, as described in the CONSORT (Consolidated Standards for Reporting Trials) diagram in Figure 2.

Table 1 summarizes the demographic characteristics of participants. The participants' mean age was 8.4 years, including 63 (78.8%) male participants. The participants' mean IQ was 98.7 (SD 10.8). The academic performance of more than 65% of the participants with ADHD ranked in the bottom half of the class. Nearly 90% of the participants were rated as having poor social skills by their caregivers and nearly 80% were the first child in the family. More than 60% had a nuclear family model (ie, living with parents only) and were referred to the hospital by a family member. Age, sex, IQ, achievement, social skills, ADHD severity, and prevalence of each ADHD subtype were not significantly different between the two groups of participants (all $P>.05$).

The primary outcome was ADHD symptoms as measured by the SNAP-IV total and subscale scores. Table 2 shows the results of the parents' rating of SNAP-IV scores at baseline and at the study endpoint in the intervention and control groups. Participants in the intervention group had a significantly larger reduction in SNAP-IV total scores than those in the control group. For subscale scores, the overall reduction was attributable to significantly lower scores on the Inattention, Hyperactivity/Impulsivity, and ODD subscale scores.

In terms of EF as rated by the BRIEF, participants in the intervention group had a significantly larger reduction on the MI and the GEC, with a smaller and nonsignificant reduction in the BRI (Table 2). Except for the Inhibit clinical scale, children in the intervention group showed a larger reduction across all clinical scales as compared to those in the control group, with five clinical scales showing significant decreases: Initiate, Working Memory, Plan/Organize, Organization of Materials, and Monitor. The Initiate clinical scale decreased from a mean of 63.10 to 54.25 for participants in the intervention group, whereas the control group scores decreased from 59.90 to 59.03 ($\beta=-5.80$, $P=.005$).

Figure 3 depicts the improvements in the context of the wide range of baseline scores for all participants. The effect sizes were generally small to moderate for the outcomes that showed a statistically significant difference in improvements between the intervention and control groups.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards for Reporting Trials) diagram of participation flow in the study.

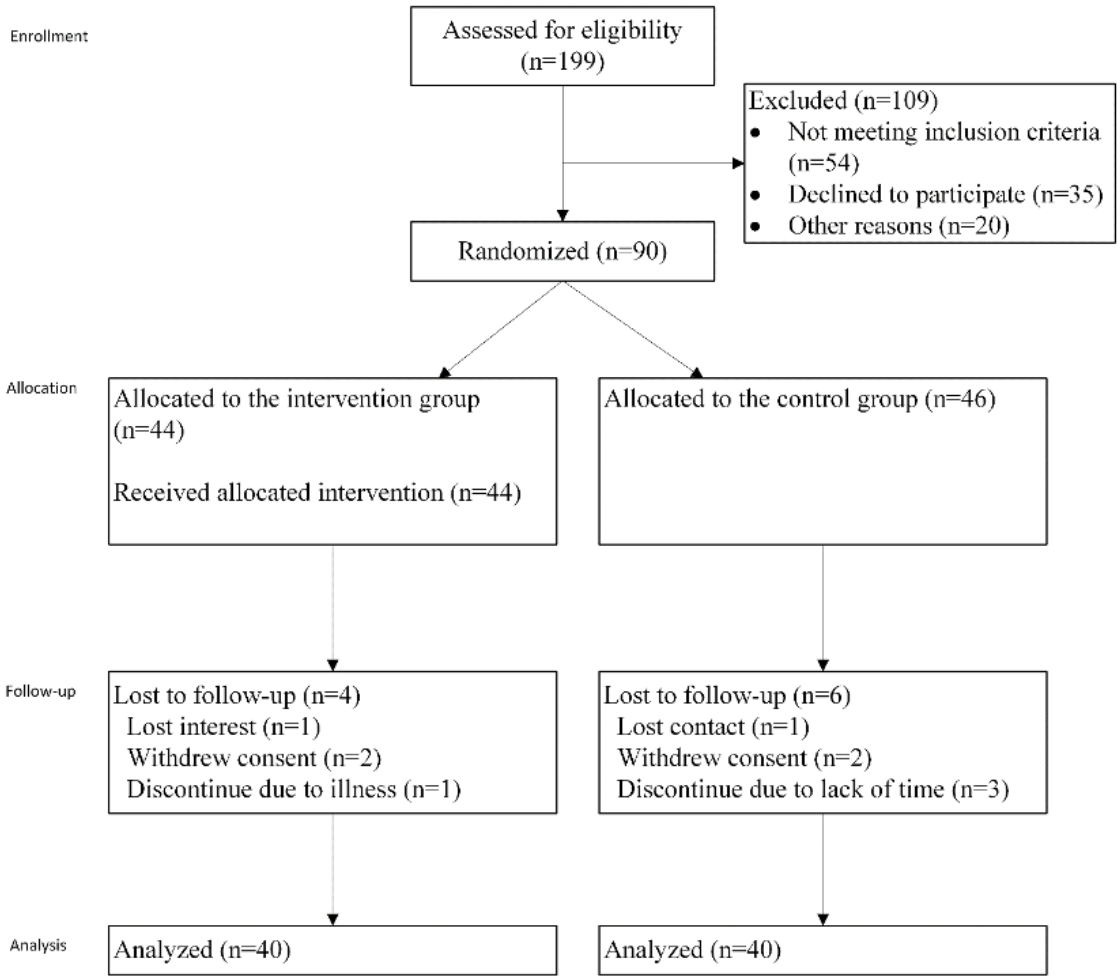

XSL•FO
RenderX

Table 1. Baseline characteristics of the study sample overall and in the study groups allocated by random assignment.

Variables	Overall (N=80)	Intervention group (n=40)	Control group (n=40)	*P* value
Age (years), mean (SD)	8.4 (1.3)	8.5 (1.5)	8.3 (1.1)	.43[a]
Boys, n (%)	63 (79)	34 (85)	29 (73)	.27[b]
IQ, mean (SD)	98.7 (10.8)	98.4 (10.6)	99.0 (10.6)	.81[a]
Academic performance, n (%)				.43[a]
Top quartile	9 (11)	6 (15)	3 (8)	
2nd quartile	16 (20)	5 (13)	11 (28)	
3rd quartile	34 (43)	22 (55)	12 (30)	
Bottom quartile	21 (26)	7 (18)	14 (35)	
Parent-rated sociability, n (%)				.65[a]
Low	29 (36)	16 (40)	13 (33)	
Medium	44 (55)	19 (48)	25 (63)	
High	7 (9)	5 (13)	2 (5)	
First child, n (%)	66 (83)	35 (88)	31 (78)	.24[b]
Living with parents only, n (%)	48 (60)	25 (63)	23 (58)	.65[b]
Family income, n (%)				.88[a]
Below average	12 (15)	6 (15)	6 (15)	
Average	53 (66)	27 (68)	26 (65)	
Above average	15 (19)	7 (18)	8 (20)	
ADHD[c] subtype, n (%)				.14[a]
Inattentive	5 (6)	4 (10)	1 (3)	
Hyperactive/impulsive	54 (68)	23 (58)	31 (78)	
Combined presentation	21 (26)	13 (33)	8 (20)	
ADHD severity, n (%)				.82[a]
Mild	25 (31)	12 (30)	13 (33)	
Moderate	32 (40)	16 (40)	12 (40)	
Severe	23 (29)	12 (30)	11 (28)	
Referral source, n (%)				.63[a]
Teacher	7 (9)	4 (10)	3 (8)	
Physician	10 (13)	4 (10)	5 (15)	
Family member	56 (70)	27 (68)	29 (73)	
Other	7 (9)	5 (13)	2 (5)	

[a]Mann-Whitney *U* test.
[b]χ^2 test.
[c]ADHD: attention-deficit/hyperactivity disorder.

XSL•FO
RenderX

Table 2. BrainFit app intervention effect estimates as compared to the control group.

Measure	Intervention group, mean (SD)		Control group, mean (SD)		Coefficient estimate[a]	Effect size (Cohen *d*) (95% CI)[b]	*P* value[c]
	Baseline	Endpoint	Baseline	Endpoint			
SNAP-IV[d]							
Total	45.83 (13.32)	29.38 (11.87)	36.93 (12.44)	36.65 (12.75)	–12.203	1.258 (0.804-1.712)	<.001
Inattention	18.30 (4.83)	12.73 (4.77)	16.43 (2.18)	16.13 (5.10)	–3.966	0.986 (0.542-1.430)	<.001
Hyperactivity/Impulsivity	16.00 (6.29)	10.00 (5.49)	11.93 (5.74)	12.60 (5.89)	–5.735	1.107 (0.601-1.612)	<.001
ODD[e]	11.55 (5.10)	7.40 (5.10)	8.58 (4.96)	7.93 (4.81)	–2.995	0.894 (0.475-1.314)	.002[c]
BRIEF[f] domain scores							
BRI[g]	57.63 (11.94)	52.15 (10.79)	51.88 (9.48)	51.98 (10.43)	–2.905	0.638 (0.230-1.047)	.13
MI[h]	65.63 (9.30)	57.55 (11.52)	64.40 (11.60)	63.68 (10.93)	–6.312	0.688 (0.251-1.126)	.006
GEC[i]	63.23 (10.03)	55.73 (11.42)	58.50 (10.00)	58.90 (10.16)	–5.952	0.856 (0.476-1.235)	.003

[a]Multivariable linear regression model controlling for age, sex, being the first child, attention-deficit/hyperactivity disorder subtypes, academic performance, and social abilities.

[b]95% CIs were calculated based on the nonparametric bootstrap method (with 1000 repetitions).

[c]*P* values were obtained from Monte Carlo permutation tests with 10,000 repetitions.

[d]SNAP-IV: Swanson, Nolan, and Pelham questionnaire.

[e]ODD: oppositional defiant disorder.

[f]BRIEF: Behavior Rating Inventory of Executive Function.

[g]BRI: Behavior Regulation Index.

[h]MI: Metacognition Index.

[i]GEC: Global Executive Composite.

XSL•FO
RenderX

Figure 3. Box plots of changes in SNAP-IV and BRIEF T scores over 4 weeks according to treatment group. BRIEF: Behavior Rating Inventory of Executive Function; BRI: Behavioral Regulation Index; GEC: Global Executive Composite; MI: Metacognition Index; ODD: oppositional defiant disorder; SNAP-IV: Swanson, Nolan, and Pelham questionnaire.

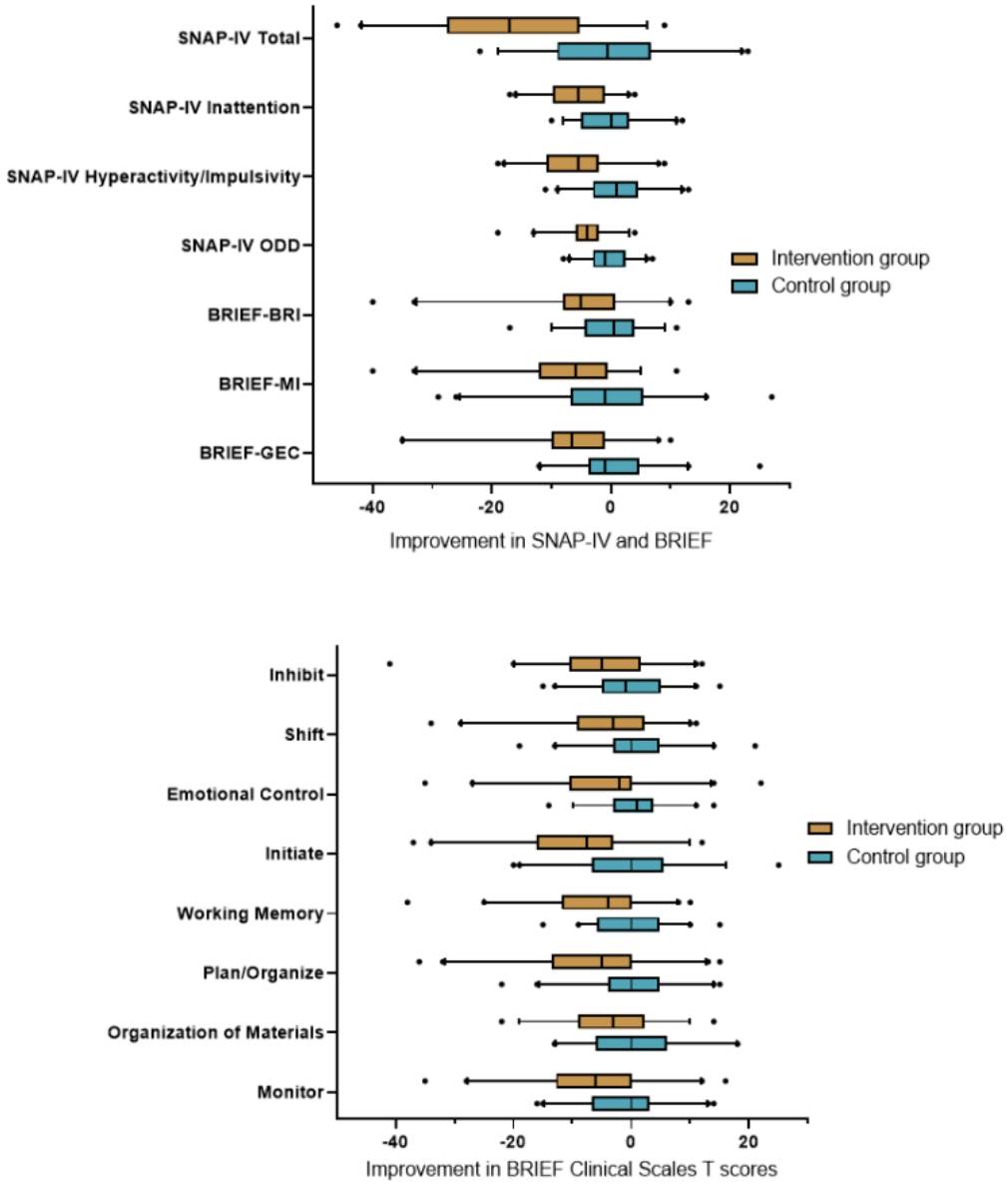

Discussion

Principal Findings

This study investigated one of the first digital interventions combining both cognitive and physical (exercise) training to address the core symptoms of ADHD and EF deficits. Compared to those in the control group, children randomized to the BrainFit intervention showed statistically significant improvement in reducing symptoms of ADHD and enhancing EF, as rated by their caregivers.

To our knowledge, this is the first study combining cognitive training and exercise therapy, two effective interventions for elementary and middle school–aged children with ADHD. Findings from this study provide evidence that a digital cognitive-physical intervention could be a promising nonpharmacologic treatment for ADHD.

The therapeutic effects of BrainFit are comparable to those reported in previous studies. A recent systematic review and meta-analysis of 14 trials reported that digital interventions significantly improved attention of children with ADHD (SMD 0.28, 95% CI 0.14-0.41). Our study demonstrated an improved inattention score with a comparable effect (SMD 0.986, 95% CI 0.581-01.447). This evidence suggested a comparable or even greater improvement of inattention than previous digital interventions, which could be related to the application of the exercise intervention. However, our intervention showed disadvantages in certain subdomains of ADHD symptoms. Davis et al [41] reported that the digital intervention “EVO” could improve attention, working memory, and inhibition in children with ADHD. However, in this study, the scores on the Inhibit, Shift, and Emotional Control subscales did not differ significantly between the intervention and control groups, which could be related to the design of the intervention or subjectivity of caregiver ratings.

XSL•FO
RenderX

With the development of digital technologies [42], increasing studies have demonstrated their promising role in the management of various diseases, especially mental and behavioral disorders [43]. Compared with traditional pharmacotherapy and behavioral interventions, digital interventions have the advantages of a minimal requirement of human resources and less adverse effects, which could reach more patients worldwide, especially for patients in less-developed countries and regions. With greater accessibility and compliance, digital therapies could provide proper and timely intervention for children with ADHD, reducing the number of children who do not receive any intervention to manage their symptoms. The US FDA approved a game-based digital therapy called "EndeavorRx" for the treatment of ADHD on June 15, 2020, which demonstrated therapeutic effects on attention among children aged 8-12 years with ADHD [44]. Emerging digital therapies such as BrainFit have the potential to be translated to clinical settings, thereby offering clinicians and families more choices for ADHD intervention.

Cognitive training has been shown to reduce deficits in EF in children with ADHD by strengthening neural networks, while physical exercise increases the levels of catecholamines such as dopamine and norepinephrine [17,27,45,46]. Exercise that is cognitively engaging can increase cognitive abilities in children with ADHD by engaging the regions of the brain involved in EF [16,27]. An "exergame" developed by Benzing and Schmidt [25] was tested in 51 children (aged 8-12 years) in a randomized controlled study that spanned 8 weeks. The game employed strength, coordination, endurance, attention, inhibition, and task-shifting. The treatment group showed improved attention, inhibition, task-shifting, general mental health, and motor abilities compared with those of the control group [25].

In another study, 35 adolescents were assigned to a treatment arm of ADHD medication and 6 weeks of exercise consisting of a sports program or to a control group of ADHD medication and education, and the results showed that the addition of exercise had a strong influence on task-shifting and improving the speed of attention processing [47]. The exercise program consisted of approximately 90-minute sessions three times per week in which there was a 10-minute warm-up period; a 60-minute aerobic exercise that engaged EF such as zigzag running, jumping rope individually and in groups, and basketball; followed by a 10-minute cool-down period [47]. It is thus recommended that game content should be included in any intervention program prescribing exercise, and the exercises should be planned and combined [27,46]. Intensity should be emphasized, while perceptual motor exercises should be combined with cognitive training such as motor planning [16,27]. This strategy has been shown to improve social, perceptual, and cognitive functions in children with ADHD [27,46].

Limitations

Our study showed promising results for a novel digital cognitive-physical intervention targeting symptoms of ADHD and EF deficits. However, several limitations should be considered in interpreting the findings. First, this was a single-center study with a limited sample size; thus, the study was not powered for subscale analyses across measures. Second, the generalizability of our study findings is limited by our lack of control for ADHD-medication use, ADHD subtype, and age distribution in patients with ADHD. Third, this study measured the outcomes immediately after the conclusion of the 4-week intervention; thus, we are unable to assess the durability of the treatment effects. Based on the encouraging findings of this study, a multicenter efficacy trial controlling for medication use and ADHD subtype with a broader age distribution and a longer follow-up period is planned. Fourth, a dose-finding study with different frequencies of sessions may be necessary to optimize the intervention dose and assess treatment heterogeneity. Finally, the lack of blinding and a sham control may have led to expectancy effects from the parents in completing the assessments.

Conclusion

The findings of this randomized controlled trial conducted in China suggest that BrainFit, a novel digital intervention combining cognitive training and exercise, could improve the symptoms and EF in children with ADHD. The intervention has the potential to increase access to ADHD treatment and improve ADHD symptoms in school-aged children.

Acknowledgments

The authors express their sincere gratitude to all of the children with attention-deficit/hyperactivity disorder and their family members who participated in this study, as well as the staff at Children's Hospital Affiliated to Zhengzhou University and Recovery Plus China, for their contributions to this research project. This research was supported by Recovery Plus China. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data sets generated and used in the current study are available from the corresponding author on reasonable request.

Authors' Contributions

LZ contributed to conceptualization, data curation, funding acquisition, project administration, writing of the original draft, and review and editing of the manuscript. HA contributed to writing of the original draft and review and editing of the manuscript. YD contributed to conceptualization and review and editing of the manuscript. HM contributed to formal analysis and review

XSL•FO
RenderX

and editing of the manuscript. RM contributed to editing of the manuscript. KL contributed to conceptualization and project administration. PA contributed to intervention design and software maintenance. CWG served as the statistical expert for this research and edited the manuscript. SF contributed to conceptualization, funding acquisition, and review and editing of the manuscript.

Conflicts of Interest

PA is employed by Recovery Plus China. SF received research funding from Recovery Plus China. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2586 KB - jmir_v26i1e55569_app1.pdf]

References

1. Rajaprakash M, Leppert M. Attention-deficit/hyperactivity disorder. Pediatr Rev 2022 Mar 01;43(3):135-147. [doi: 10.1542/pir.2020-000612] [Medline: 35229109]
2. Li Y, Yan X, Li Q, Li Q, Xu G, Lu J, et al. Prevalence and trends in diagnosed ADHD among US children and adolescents, 2017-2022. JAMA Netw Open 2023 Oct 02;6(10):e2336872 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36872] [Medline: 37792379]
3. Liu A, Xu Y, Yan Q, Tong L. The prevalence of attention deficit/hyperactivity disorder among Chinese children and adolescents. Sci Rep 2018 Aug 16;8(1):11169. [doi: 10.1038/s41598-018-29488-2] [Medline: 30115972]
4. Danielson ML, Bitsko RH, Ghandour RM, Holbrook JR, Kogan MD, Blumberg SJ. Prevalence of parent-reported ADHD diagnosis and associated treatment among U.S. children and adolescents, 2016. J Clin Child Adolesc Psychol 2018 Jan 24;47(2):199-212 [FREE Full text] [doi: 10.1080/15374416.2017.1417860] [Medline: 29363986]
5. Rubia K, Westwood S, Aggensteiner P, Brandeis D. Neurotherapeutics for attention deficit/hyperactivity disorder (ADHD): a review. Cells 2021 Aug 21;10(8):2156 [FREE Full text] [doi: 10.3390/cells10082156] [Medline: 34440925]
6. Thomas R, Sanders S, Doust J, Beller E, Glasziou P. Prevalence of attention-deficit/hyperactivity disorder: a systematic review and meta-analysis. Pediatrics 2015 Apr 02;135(4):e994-e1001. [doi: 10.1542/peds.2014-3482] [Medline: 25733754]
7. Enriquez-Geppert S, Smit D, Pimenta MG, Arns M. Neurofeedback as a treatment intervention in ADHD: current evidence and practice. Curr Psychiatry Rep 2019 May 28;21(6):46 [FREE Full text] [doi: 10.1007/s11920-019-1021-4] [Medline: 31139966]
8. Daly BP, Creed T, Xanthopoulos M, Brown RT. Psychosocial treatments for children with attention deficit/hyperactivity disorder. Neuropsychol Rev 2007 Mar 27;17(1):73-89. [doi: 10.1007/s11065-006-9018-2] [Medline: 17260167]
9. Khan MU, Aslani P. A review of factors influencing the three phases of medication adherence in people with attention-deficit/hyperactivity disorder. J Child Adolesc Psychopharmacol 2019 Aug;29(6):398-418. [doi: 10.1089/cap.2018.0153] [Medline: 31120328]
10. Titheradge D, Godfrey J, Eke H, Price A, Ford T, Janssens A. Why young people stop taking their attention deficit hyperactivity disorder medication: a thematic analysis of interviews with young people. Child Care Health Dev 2022 Sep 20;48(5):724-735 [FREE Full text] [doi: 10.1111/cch.12978] [Medline: 35102579]
11. Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management, Wolraich M, Brown L, Brown RT, DuPaul G, et al. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics 2011 Nov;128(5):1007-1022 [FREE Full text] [doi: 10.1542/peds.2011-2654] [Medline: 22003063]
12. Rajeh A, Amanullah S, Shivakumar K, Cole J. Interventions in ADHD: a comparative review of stimulant medications and behavioral therapies. Asian J Psychiatr 2017 Feb;25:131-135. [doi: 10.1016/j.ajp.2016.09.005] [Medline: 28262134]
13. Abramowitz AJ, O'Leary SG. Behavioral interventions for the classroom: implications for students with ADHD. School Psychol Rev 2019 Dec 22;20(2):220-234. [doi: 10.1080/02796015.1991.12085547]
14. Zwi M, Jones H, Thorgaard C, York A, Dennis JA. Parent training interventions for attention deficit hyperactivity disorder (ADHD) in children aged 5 to 18 years. Cochrane Database Syst Rev 2011 Dec 07;2011(12):CD003018 [FREE Full text] [doi: 10.1002/14651858.CD003018.pub3] [Medline: 22161373]
15. Weisenmuller C, Hilton D. Barriers to access, implementation, and utilization of parenting interventions: considerations for research and clinical applications. Am Psychol 2021 Jan;76(1):104-115. [doi: 10.1037/amp0000613] [Medline: 32134281]
16. Ng QX, Ho CYX, Chan HW, Yong BZJ, Yeo W. Managing childhood and adolescent attention-deficit/hyperactivity disorder (ADHD) with exercise: a systematic review. Complement Ther Med 2017 Oct;34:123-128. [doi: 10.1016/j.ctim.2017.08.018] [Medline: 28917364]
17. Bustamante E, Davis C, Frazier S, Rusch D, Fogg LF, Atkins MS, et al. Randomized controlled trial of exercise for ADHD and disruptive behavior disorders. Med Sci Sports Exerc 2016 Jul;48(7):1397-1407 [FREE Full text] [doi: 10.1249/MSS.0000000000000891] [Medline: 26829000]

XSL•FO
RenderX

18. Baweja R, Soutullo CA, Waxmonsky JG. Review of barriers and interventions to promote treatment engagement for pediatric attention deficit hyperactivity disorder care. World J Psychiatry 2021 Dec 19;11(12):1206-1227 [FREE Full text] [doi: 10.5498/wjp.v11.i12.1206] [Medline: 35070771]
19. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. Lancet Digit Health 2020 Apr;2(4):e168-e178 [FREE Full text] [doi: 10.1016/S2589-7500(20)30017-0] [Medline: 33334505]
20. Klingberg T, Fernell E, Olesen PJ, Johnson M, Gustafsson P, Dahlström K, et al. Computerized training of working memory in children with ADHD--a randomized, controlled trial. J Am Acad Child Adolesc Psychiatry 2005 Feb;44(2):177-186 [FREE Full text] [doi: 10.1097/00004583-200502000-00010] [Medline: 15689731]
21. Sonuga-Barke EJS, Coghill D. The foundations of next generation attention-deficit/hyperactivity disorder neuropsychology: building on progress during the last 30 years. J Child Psychol Psychiatry 2014 Dec 17;55(12):e1-e5. [doi: 10.1111/jcpp.12360] [Medline: 25399637]
22. Cortese S, Ferrin M, Brandeis D, Buitelaar J, Daley D, Dittmann RW, European ADHD Guidelines Group (EAGG). Cognitive training for attention-deficit/hyperactivity disorder: meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials. J Am Acad Child Adolesc Psychiatry 2015 Mar;54(3):164-174 [FREE Full text] [doi: 10.1016/j.jaac.2014.12.010] [Medline: 25721181]
23. Rubia K. Functional brain imaging across development. Eur Child Adolesc Psychiatry 2013 Dec 24;22(12):719-731 [FREE Full text] [doi: 10.1007/s00787-012-0291-8] [Medline: 22729957]
24. Sha L, Yang X, Deng R, Wang W, Tao Y, Cao H, et al. Automated digital interventions and smoking cessation: systematic review and meta-analysis relating efficiency to a psychological theory of intervention perspective. J Med Internet Res 2022 Nov 16;24(11):e38206 [FREE Full text] [doi: 10.2196/38206] [Medline: 36383408]
25. Benzing V, Schmidt M. The effect of exergaming on executive functions in children with ADHD: a randomized clinical trial. Scand J Med Sci Sports 2019 Aug 23;29(8):1243-1253 [FREE Full text] [doi: 10.1111/sms.13446] [Medline: 31050851]
26. Jiang K, Xu Y, Li Y, Li L, Yang M, Xue P. How aerobic exercise improves executive function in ADHD children: a resting-state fMRI study. Int J Dev Neurosci 2022 Jun 30;82(4):295-302. [doi: 10.1002/jdn.10177] [Medline: 35274372]
27. Chan Y, Jang J, Ho C. Effects of physical exercise on children with attention deficit hyperactivity disorder. Biomed J 2022 Apr;45(2):265-270 [FREE Full text] [doi: 10.1016/j.bj.2021.11.011] [Medline: 34856393]
28. Huang H, Jin Z, He C, Guo S, Zhang Y, Quan M. Chronic exercise for core symptoms and executive functions in ADHD: a meta-analysis. Pediatrics 2023 Jan 01;151(1):e2022057745. [doi: 10.1542/peds.2022-057745] [Medline: 36510746]
29. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington, DC: American Psychiatric Publishing; 2013.
30. Subspecialty Group of Developmental and Behavioral Pediatrics, the Society of Pediatrics, Chinese Medical Association. Consensus on pediatric clinical practice of early identification, standardized diagnosis and treatment of attention deficit hyperactivity disorder. Zhonghua Er Ke Za Zhi 2020 Mar 02;58(3):188-193. [doi: 10.3760/cma.j.issn.0578-1310.2020.03.006] [Medline: 32135589]
31. Swanson JM, Kraemer HC, Hinshaw SP, Arnold LE, Conners CK, Abikoff HB, et al. Clinical relevance of the primary findings of the MTA: success rates based on severity of ADHD and ODD symptoms at the end of treatment. J Am Acad Child Adolesc Psychiatry 2001 Feb;40(2):168-179. [doi: 10.1097/00004583-200102000-00011] [Medline: 11211365]
32. Qian Y, Wang Y. Reliability and validity of behavior rating scale of executive function parent form for school age children in China. Beijing Da Xue Xue Bao Yi Xue Ban 2007 Jun 18;39(3):277-283 [FREE Full text] [Medline: 17572784]
33. Gioia G, Isquith P, Guy S, Kenworthy L. BRIEF. Behavior Rating Inventory of Executive Function Professional Manual. Lutz, Florida: Psychological Assessment Resources, Inc; 2000.
34. Liang X, Li R, Wong SHS, Sum RKW, Sit CHP. The impact of exercise interventions concerning executive functions of children and adolescents with attention-deficit/hyperactive disorder: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 May 22;18(1):68 [FREE Full text] [doi: 10.1186/s12966-021-01135-6] [Medline: 34022908]
35. Imbens G, Rubin D. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge, UK: Cambridge University Press; 2015:A.
36. Fisher R. The Design of Experiments. Edinburgh: Oliver & Boyd; 1935.
37. Pitman E. Significance tests which may be applied to samples from any populations. Suppl JR Stat Soc 1937;4(1):119. [doi: 10.2307/2984124]
38. Heß S. Randomization inference with Stata: a guide and software. The Stata Journal 2017 Sep 01;17(3):630-651. [doi: 10.1177/1536867x1701700306]
39. Manly B. F. In: Randomization, Bootstrap, and Monte Carlo Methods in Biology. 3th edition. Boca Raton, FL: Chapman and Hall/CRC; 2006.
40. Efron B, Tibshirani R. An Introduction to the Bootstrap. Milton Park, UK: Taylor & Francis; 1994.
41. Davis NO, Bower J, Kollins SH. Proof-of-concept study of an at-home, engaging, digital intervention for pediatric ADHD. PLoS One 2018 Jan 11;13(1):e0189749 [FREE Full text] [doi: 10.1371/journal.pone.0189749] [Medline: 29324745]

XSL•FO
RenderX

42. Fleming TM, Bavin L, Stasiak K, Hermansson-Webb E, Merry SN, Cheek C, et al. Serious games and gamification for mental health: current status and promising directions. Front Psychiatry 2016;7:215 [FREE Full text] [doi: 10.3389/fpsyt.2016.00215] [Medline: 28119636]
43. Christie GI, Shepherd M, Merry SN, Hopkins S, Knightly S, Stasiak K. Gamifying CBT to deliver emotional health treatment to young people on smartphones. Internet Interv 2019 Dec;18:100286 [FREE Full text] [doi: 10.1016/j.invent.2019.100286] [Medline: 31890633]
44. FDA permits marketing of first game-based digital therapeutic to improve attention function in children with ADHD. 2020 Jun. URL: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-first-game-based-digital-therapeutic-improve-attention-function-children-adhd [accessed 2020-08-14]
45. Kamp CF, Sperlich B, Holmberg H. Exercise reduces the symptoms of attention-deficit/hyperactivity disorder and improves social behaviour, motor skills, strength and neuropsychological parameters. Acta Paediatr 2014 Jul 29;103(7):709-714. [doi: 10.1111/apa.12628] [Medline: 24612421]
46. Zang Y. Impact of physical exercise on children with attention deficit hyperactivity disorders: evidence through a meta-analysis. Medicine 2019 Nov;98(46):e17980 [FREE Full text] [doi: 10.1097/MD.0000000000017980] [Medline: 31725664]
47. Choi J, Han D, Kang K, Jung H, Renshaw P. Aerobic exercise and attention deficit hyperactivity disorder: brain research. Med Sci Sports Exerc 2015 Jan;47(1):33-39 [FREE Full text] [doi: 10.1249/MSS.0000000000000373] [Medline: 24824770]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
AR: augmented reality
BRI: Behavioral Regulation Index
BRIEF: Behavior Rating Inventory of Executive Function
CONSORT: Consolidated Standards for Reporting Trials
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
EF: executive function
FDA: Food and Drug Administration
GEC: Global Executive Composite
MI: Metacognition Index
ODD: oppositional defiant disorder
SMD: standardized mean difference
SNAP-IV: Swanson, Nolan, and Pelham questionnaire

Edited by T de Azevedo Cardoso; submitted 22.12.23; peer-reviewed by YF Deng; comments to author 31.01.24; revised version received 15.02.24; accepted 25.04.24; published 10.05.24.

Please cite as:
Zhao L, Agazzi H, Du Y, Meng H, Maku R, Li K, Aspinall P, Garvan CW, Fang S
A Digital Cognitive-Physical Intervention for Attention-Deficit/Hyperactivity Disorder: Randomized Controlled Trial
J Med Internet Res 2024;26:e55569
URL: https://www.jmir.org/2024/1/e55569
doi:10.2196/55569
PMID:38728075

©Licong Zhao, Heather Agazzi, Yasong Du, Hongdao Meng, Renya Maku, Ke Li, Peter Aspinall, Cynthia Wilson Garvan, Shuanfeng Fang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of Adding Personalized Instant Messaging Apps to a Brief Smoking Cessation Model in Community Smokers in Hong Kong: Pragmatic Randomized Clinical Trial

Yongda Socrates Wu[1,2], PhD; Yee Tak Derek Cheung[1], PhD; Jay Jung Jae Lee[1], PhD; Carlos King Ho Wong[3], PhD; Sai Yin Ho[4], PhD; William Ho Cheung Li[1,5], PhD; Ying Yao[1], MSc; Tai Hing Lam[4], MD; Man Ping Wang[1], PhD

[1]School of Nursing, The University of Hong Kong, Hong Kong, China (Hong Kong)
[2]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
[3]Department of Family Medicine and Primary Care, The University of Hong Kong, Hong Kong, China (Hong Kong)
[4]School of Public Health, The University of Hong Kong, Hong Kong, China (Hong Kong)
[5]Nethersole School of Nursing, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Man Ping Wang, PhD
School of Nursing
The University of Hong Kong
5/F, Academic Building
3 Sassoon Road, Pokfulam
Hong Kong
China (Hong Kong)
Phone: 852 39176636
Fax: 852 28726079
Email: mpwang@hku.hk

Abstract

Background: While text messaging has proven effective for smoking cessation (SC), engagement in the intervention remains suboptimal.

Objective: This study aims to evaluate whether using more interactive and adaptive instant messaging (IM) apps on smartphones, which enable personalization and chatting with SC advisors, can enhance SC outcomes beyond the provision of brief SC advice and active referral (AR) to SC services.

Methods: From December 2018 to November 2019, we proactively recruited 700 adult Chinese daily cigarette users in Hong Kong. Participants were randomized in a 1:1 ratio. At baseline, all participants received face-to-face brief advice on SC. Additionally, they were introduced to local SC services and assisted in selecting one. The intervention group received an additional 26 personalized regular messages and access to interactive chatting through IM apps for 3 months. The regular messages aimed to enhance self-efficacy, social support, and behavioral capacity for quitting, as well as to clarify outcome expectations related to cessation. We developed 3 sets of messages tailored to the planned quit date (within 30 days, 60 days, and undecided). Participants in the intervention group could initiate chatting with SC advisors on IM themselves or through prompts from regular messages or proactive inquiries from SC advisors. The control group received 26 SMS text messages focusing on general health. The primary outcomes were smoking abstinence validated by carbon monoxide levels of <4 parts per million at 6 and 12 months after the start of the intervention.

Results: Of the participants, 505/700 (72.1%) were male, and 450/648 (69.4%) were aged 40 or above. Planning to quit within 30 days was reported by 500/648 (77.2%) participants, with fewer intervention group members (124/332, 37.3%) reporting previous quit attempts compared with the control group (152/335, 45.4%; *P*=.04). At the 6- and 12-month follow-ups (with retention rates of 456/700, 65.1%, and 446/700, 63.7%, respectively), validated abstinence rates were comparable between the intervention (14/350, 4.0%, and 19/350, 5.4%) and control (11/350, 3.1% and 21/350, 6.0%) groups. Compared with the control group, the intervention group reported greater utilization of SC services at 12 months (RR 1.26, 95% CI 1.01-1.56). Within the intervention group, engaging in chat sessions with SC advisors predicted better validated abstinence at 6 months (RR 3.29, 95% CI 1.13-9.63) and any use of SC services (RR 1.66, 95% CI 1.14-2.43 at 6 months; RR 1.67, 95% CI 1.26-2.23 at 12 months).

XSL•FO
RenderX

Conclusions: An IM-based intervention, providing support and assistance alongside brief SC advice and AR, did not yield further increases in quitting rates but did encourage the utilization of SC services. Future research could explore whether enhanced SC service utilization leads to improved long-term SC outcomes.

Trial Registration: ClinicalTrials.gov NCT03800719; https://clinicaltrials.gov/ct2/show/NCT03800719

(J Med Internet Res 2024;26:e44973) doi:10.2196/44973

KEYWORDS

instant messaging; text messaging; chatting; smoking cessation; COVID-19; community smoker

Introduction

Smoking cessation (SC) can effectively alleviate the burden of diseases caused by tobacco use [1]. SMS text messaging has proven to be effective for SC [2,3]. The effectiveness of SMS text messaging for SC primarily lies in providing psychosocial support [4]. Exploring more interactive and adaptable platforms, such as instant messaging (IM) apps on smartphones that offer personalization and the ability to chat with SC advisors, might offer enhanced support, potentially leading to increased success in SC efforts. The World Health Organization has also embraced platforms such as WhatsApp (Meta Platforms, Inc.) and Facebook Messenger (Meta Platforms, Inc.) to encourage quitting [5], although the evidence regarding their effectiveness remains uncertain. Our literature search (see Multimedia Appendix 1) yielded only 6 relevant randomized controlled trials (RCTs), with full results provided in Multimedia Appendix 2, highlighting the lack of conclusive evidence regarding the effectiveness of IM-based interventions for SC [6-11].

Smartphone usage is widespread in Hong Kong [12], the most Westernized city in China, where WhatsApp and WeChat (Tencent Holdings Ltd.) are the predominant IM apps. In our previous RCT involving community smokers, we offered chat-based support through IM alongside brief SC advice delivered face-to-face [13]. The combined intervention resulted in a 70% increase in the odds of validated abstinence compared with receiving only brief SC advice. However, the level of engagement in chatting was low (16.8%), a trend commonly observed in trials using digital interventions for diverse health concerns [14,15]. Previous studies have indicated that engagement levels can predict successful quitting outcomes [13,15], echoing similar findings in digital interventions aimed at addressing various health issues [14]. These results emphasize the importance of investigating strategies to enhance engagement, including the delivery of multimedia messages and tailoring interventions to individual quitting-related characteristics [16].

SC services provide effective interventions, including nicotine replacement therapies (NRTs), medications such as varenicline and bupropion, and behavioral support [17]. However, these services are underutilized worldwide, as highlighted by the Global Adult Tobacco Survey, which revealed that only a quarter of smokers who attempted to quit in the past 12 months accessed SC services [18]. Similarly, only 4.9% of smokers in Hong Kong had previously used SC services or medications, despite the fact that these services are largely available free of charge or at a nominal fee (HK $50 [US $6.40]). Among those who had not previously used these services or medications, a staggering 94.0% expressed reluctance to consider trying them [19,20]. Over a quarter of local smokers in Hong Kong were classified as hardcore smokers, having never attempted to quit and expressing no desire to do so [21]. Consequently, the prevalence of smoking in Hong Kong has reached a standstill in recent years, with rates at 9.8% (daily) and 0.6% (nondaily) as of 2021 [20]. The majority of smokers were male (83.1%) and aged 40 years or older, with 25.1% in the 40-49-year age group, 23.0% in the 50-59-year age group, and 27.6% aged 60 years or older.

Proactive strategies have proven effective in boosting the utilization of SC services among smokers recruited from clinical or primary care settings, as demonstrated in numerous studies across the United States [22-25], the United Kingdom [26,27], and Australia [28]. Despite its effectiveness, active referral (AR), which involves passing smokers' contact information to SC service providers, resulted in only 29.1% of community smokers using SC services during the study period, albeit leading to increased uptake of SC services and higher rates of smoking abstinence [29]. The primary reasons cited for not using SC services were a busy schedule, timing conflicts, and lack of interest [29]. Subsequent interventions aimed at enhancing AR strategies, such as on-site referral (where service providers are contacted on behalf of smokers at baseline) [30] and AR coupled with a small financial incentive for SC service utilization [31], demonstrated increased rates of quitting compared with brief SC advice alone. However, only the on-site referral approach showed a slight improvement in SC service utilization, with no comparison made to a trial arm offering AR alone.

This RCT examined a novel IM-based intervention that incorporated personalized psychosocial support and guidance for utilizing SC services, in addition to brief SC advice and AR, among smokers proactively recruited from communities across Hong Kong. Our hypothesis posited that tailored messages could enhance engagement with the IM-based intervention, thereby offering increased support for SC efforts. By enhancing assistance for utilizing SC services alongside personalized support through IM, the intervention has the potential to bolster both SC service utilization and smoking abstinence rates. The findings from this study will offer valuable insights into strategies for enhancing SC support via IM platforms and encouraging the uptake of SC services. If proven effective, this intervention model could serve as a valuable reference for promoting SC in countries or regions where IM is widely used and SC services are readily accessible.

XSL•FO
RenderX

Methods

Study Design

This study was a parallel 2-arm, pragmatic RCT conducted within communities across Hong Kong. The protocol has been registered with ClinicalTrials.gov (NCT03800719) and is accessible as Multimedia Appendix 3. We adhered to the CONSORT (Consolidated Standards of Reporting Trials) guidelines for pragmatic trials (Multimedia Appendix 4) [32].

Ethics Approval

Ethical approval was obtained from the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (UW 18-172). Written informed consent was obtained from all participants, and all data and results presented in this study were deidentified.

Recruitment and Participants

Recruitment occurred at community smoking hotspots, areas frequented by many smokers for smoking purposes, including locations near shopping centers, housing estates, and metro stations [33]. Adult daily cigarette users were recruited between December 15, 2018, and November 7, 2019. SC advisors, who were university students trained in a half-day workshop and supervised by research staff on-site, proactively approached smokers using a "foot-in-the-door" technique [33,34]. Smokers were initially queried about their smoking history and habits, including their daily cigarette consumption, age at initiation of smoking, and any previous attempts to quit, including the utilization of SC services. Smokers who expressed a willingness to engage were briefly advised to consider quitting or reducing smoking, and those showing interest were extended an invitation to participate in this RCT. To be eligible for participation, individuals had to be daily cigarette users within the past 3 months, as verified by an exhaled carbon monoxide (CO) level of 4 parts per million (ppm) or above, measured at smoking hotspots during recruitment. Additionally, they needed to be Hong Kong residents aged 18 or older, proficient in Chinese, and in possession of a smartphone equipped with an IM app for receiving intervention messages and follow-up. Exclusion criteria encompassed individuals who (1) had psychiatric or psychological disorders diagnosed by a physician or were under treatment for such conditions, (2) were on regular psychotropic medications, or (3) were using SC medications or services.

Randomization and Blinding

Participants were randomly assigned in a 1:1 ratio using sequentially numbered, opaque, sealed envelopes (SNOSEs). An SC advisor would open 1 SNOSE promptly after receiving written consent from participants. The SNOSEs were prepared by an investigator (YSW) who was not involved in recruitment. They were generated via computerized allocation sequences with block sizes of 4, 8, and 12 in random order. The intervention was administered to 1 smoker at a time to prevent contamination. Following the conclusion of the current intervention, a 5-minute interval was observed to allow the recruited smoker to depart before proceeding to recruit a new smoker at the same smoking hotspot. Allocation concealment was maintained; however, blinding of the SC advisors and participants was not feasible. Nevertheless, all outcome assessors and statistical analysts remained blinded until the primary analyses were concluded.

Interventions

Upon completing the baseline questionnaire, all participants were provided with brief SC advice, typically lasting 5-10 minutes, delivered face-to-face. This advice was guided by the AWARD (Ask, Warn, Advise, Refer, and Do-it-again) model, which was designed and tested in our previous trials [13,29,35,36]. The AWARD model consists of the following components: *Ask* about smoking history; *Warn* about the high risk of smoking using a health warning leaflet; *Advise* smokers to quit as soon as possible and adhere to the chosen quit date; *Refer* smokers to SC services; and *Do* it again. The AWARD model was adapted and simplified from the 5As (Ask, Advise, Assess, Assist, and Arrange) specifically for implementation in the community setting [37]. Additionally, participants were proactively directed to local SC services. They were briefed about the available SC services in Hong Kong and aided in selecting one based on their preferences regarding service offerings, such as available treatments, operational hours, and locations. The contact details of participants who provided written consent for referral were promptly forwarded to their selected service providers within the following day. Subsequently, the service providers reached out to the referred participants to arrange phone or face-to-face counseling sessions or consultations, typically scheduled within 1-2 weeks.

In addition to the standard intervention, participants in the intervention group received an IM-based intervention using platforms such as WhatsApp or WeChat. This intervention involved receiving personalized messages at regular intervals (refer to Multimedia Appendix 5) and engaging in interactive chatting for 3 months following the initiation of the intervention. The content of the regular messages was guided by the principles of the social cognitive theory, which has been used in previous SC interventions utilizing text messages [38,39]. The IM-based intervention aimed to enhance participants' self-efficacy, social support, and behavioral capacity for quitting smoking, while also clarifying the expectations associated with quitting. Practical skills and emotional support were provided to assist participants in managing cravings, withdrawal symptoms, situational triggers, lapses, and relapses. These components were designed based on standard information from the SC kit provided by the Department of Health, Hong Kong SAR [40], and informed by evidence from our previous smoking relapse prevention trial [6]. We devised 3 sets of messages (sent through IM, with some also including pictures and videos) tailored to the planned quit date, categorized as within 30 days, within 60 days, and undecided, in alignment with the transtheoretical model (TTM) of change [41]. Following the steps recommended by Abroms et al [42] for message design, we consulted experienced SC counselors from local service providers to refine the messages before finalization. Additionally, messages were personalized based on participants' sociodemographic characteristics, baseline smoking habits, and updated smoking status obtained during interactive messaging conversations. The primary objective was to boost engagement, defined as active participation in conversations with SC advisors on IM platforms.

XSL•FO
RenderX

A total of 26 messages were sent, comprising various formats such as texts, pictures, and videos, aimed at enhancing engagement. The message schedule involved sending messages once daily during the week encompassing the quit date, followed by 3 times a week for 4 weeks (2 weeks before and after the quit date), and once a week for 7 weeks. In instances where a quit date was not established, daily messages were initiated immediately. Furthermore, the frequency and schedule of messages were adjusted based on participants' preferences as conveyed during IM conversations. Participants had the autonomy to initiate text or voice chats with SC advisors on IM platforms, or these interactions could be prompted by regular messages or proactive inquiries from SC advisors. For instance, advisors might inquire about the current progress of quitting or the utilization of SC services. Specifically, SC advisors reminded participants to utilize the chosen service and provided assistance in addressing any issues encountered in utilizing these services, such as missed appointments or inquiries regarding the location and contact information of service providers. All conversations were meticulously recorded, with 5% randomly selected for review by experienced SC counselors from the research team.

Participants assigned to the control group received 26 messages (refer to Multimedia Appendix 6) focusing on general health information and reminders for follow-up appointments over the same 3-month period. These were delivered as SMS text messages at a frequency comparable to that of the intervention group.

Outcomes

All participants underwent telephone follow-ups at 3, 6, and 12 months after the initiation of the intervention to evaluate their smoking status. A HK $50 (US $6.40) coupon was provided as compensation for each successful follow-up to acknowledge their time and participation. The primary outcome measure was validated abstinence, defined as an exhaled CO level of less than 4 parts per million (ppm), assessed at 6 and 12 months after the initiation of the intervention. Participants who reported complete abstinence (not even a single puff) in the past 7 days at the 6- and 12-month follow-ups were invited to undergo an exhaled CO test using a piCO Smokerlyzer (Bedfont Scientific), facilitated by research staff. These tests were conducted at outdoor community sites that were convenient for the participants. To acknowledge their time and cover travel expenses, participants received a small cash incentive of HK $300 (US $38.40). It is worth noting that previous trials involving community smokers found that this incentive had no impact on quitting outcomes [43]. The main secondary outcomes were assessed at 6 and 12 months, including self-reported 7-day point prevalence abstinence (PPA), continuous abstinence for 24 weeks, reduction of smoking by at least 50% compared with baseline daily cigarette consumption, any quit attempt (abstinence for 24 hours or longer), and utilization of SC services after intervention initiation. The self-reported utilization of SC services, including even a single session, was validated by cross-referencing with records from service providers. Additionally, participants in the intervention group reported whether they had engaged in chat sessions with SC advisors through IM during the 3-month intervention period, which was verified using chat records. To evaluate the quality of life, the validated EQ-5D-5L was administered at baseline and at the 12-month follow-up [44]. As part of the process evaluation for the IM-based intervention, all participants were requested to provide ratings at the 6-month mark regarding the helpfulness of messages, which also included chatting with participants in the intervention group. Ratings were provided on a scale from 0 (not at all helpful) to 10 (most helpful) for the following aspects: (1) increasing quitting motivation, (2) increasing quitting confidence, and (3) overall helpfulness.

Sample Size

The sample size was determined using G-Power (Universität Düsseldorf). As no previous studies had reported the effect of IM messages and chatting in addition to brief advice and AR, calculations were based on the following quit rates: (1) a 10.7% biochemically validated quit rate at 6 months, as reported in a previous trial evaluating SMS text messaging–based interventions [45]; and (2) our previous trials, where only brief advice was provided in the control group, observed a 5.0% validated quit rate at 6 months [29,35]. Assuming a type I error of 0.05 and a power of 0.8, as observed in our previous trials involving community smokers, a total of 696 participants would be required (with 348 participants in each group) to detect a significant difference between the intervention and control groups.

Statistical Analysis

We applied the intention-to-treat approach in the primary analysis, where missing outcome values were imputed with the baseline values. Poisson working models with a log-link were used to assess the intervention effect on primary outcomes, yielding risk ratios (RRs) [46]. Similar analytical methods were applied to conduct prespecified secondary analyses. In our first step, we replicated the primary analysis while adjusting for imbalanced baseline characteristics. Subsequently, subgroup analyses were conducted based on intention to quit, specifically focusing on participants with a planned quit date within 30 days at baseline [41]. Two post hoc analyses have been performed: one was to investigate whether the intervention effects on outcomes at 3 months were consistent with those observed at the subsequent 2 follow-ups and the other was to examine the association between engagement, defined as chatting with SC advisors, and cessation outcomes among participants in the intervention group.

Two prespecified sensitivity analyses were conducted: (1) complete case analyses were conducted, excluding participants with missing outcome data and (2) missing data were handled using multiple imputations by chained equations [47]. The imputation model incorporated outcomes, study group assignment, sociodemographic characteristics (eg, sex, age, highest educational attainment, and monthly household income), and baseline smoking-related characteristics (including daily cigarette consumption, time to the first cigarette after waking, previous quit attempts, and intention to quit). Fifty imputed data sets were generated, and regression results were aggregated according to Rubin's rule [48]. Responses to the EQ-5D-5L questionnaire were transformed using the standard Hong Kong value set, which ranged from –0.865 (indicating the worst

XSL•FO
RenderX

quality of life) to 1 (indicating the best quality of life). The participants' quality of life was assessed using the area under the receiver operating characteristic curve approach [44]. A 2-sided *P* value of less than .05 was considered statistically significant. All statistical analyses were performed using Stata version 15.1 (StataCorp). Additionally, individual interviews were conducted with participants to gain insights into their perceptions and experiences regarding the IM-based intervention. The findings from these interviews will be presented separately.

Results

Characteristics of Participants

Of the 731 smokers initially screened, 700 were found eligible and provided consent to participate in the RCT. A CONSORT flow diagram illustrating the participant flow is provided in Figure 1. Half of the participants (n=350) were randomly assigned to the intervention group, while the remaining half were allocated to the control group. Table 1 presents the baseline sociodemographic and smoking-related characteristics of the participants. It indicates that these characteristics were generally similar between the 2 groups. In total, 505/700 (72.1%) participants were male, with 450/648 (69.4%) aged 40 years or older, and 425/615 (69.1%) having attained secondary education. Additionally, 288/660 (43.6%) reported smoking their first cigarette within 5 minutes after waking, while 385/695 (55.4%) consumed more than 10 cigarettes daily. Notably, 500/648 (77.2%) expressed intentions to quit smoking within 30 days. A comparison between the intervention and control groups revealed that a lower proportion of participants in the intervention group had made previous quit attempts compared with those in the control group (124/332, 37.3%, vs 152/335, 45.4%; *P*=.04). Nearly all participants in both the intervention (331/350, 94.6%) and control (333/350, 95.1%) groups opted for a SC service at baseline. The overall retention rates at follow-up were 69.3% (485/700) at 3 months, 65.1% (456/700) at 6 months, and 63.7% (446/700) at 12 months. Importantly, these retention rates were comparable between the 2 groups at all follow-up time points (*P*=.46-.75).

XSL•FO
RenderX

Figure 1. The CONSORT (Consolidated Standards of Reporting Trials) flow diagram. AWARD: Ask, Warn, Advise, Refer, and Do-it-again; CO: carbon monoxide; IM: instant messaging; SC: smoking cessation.

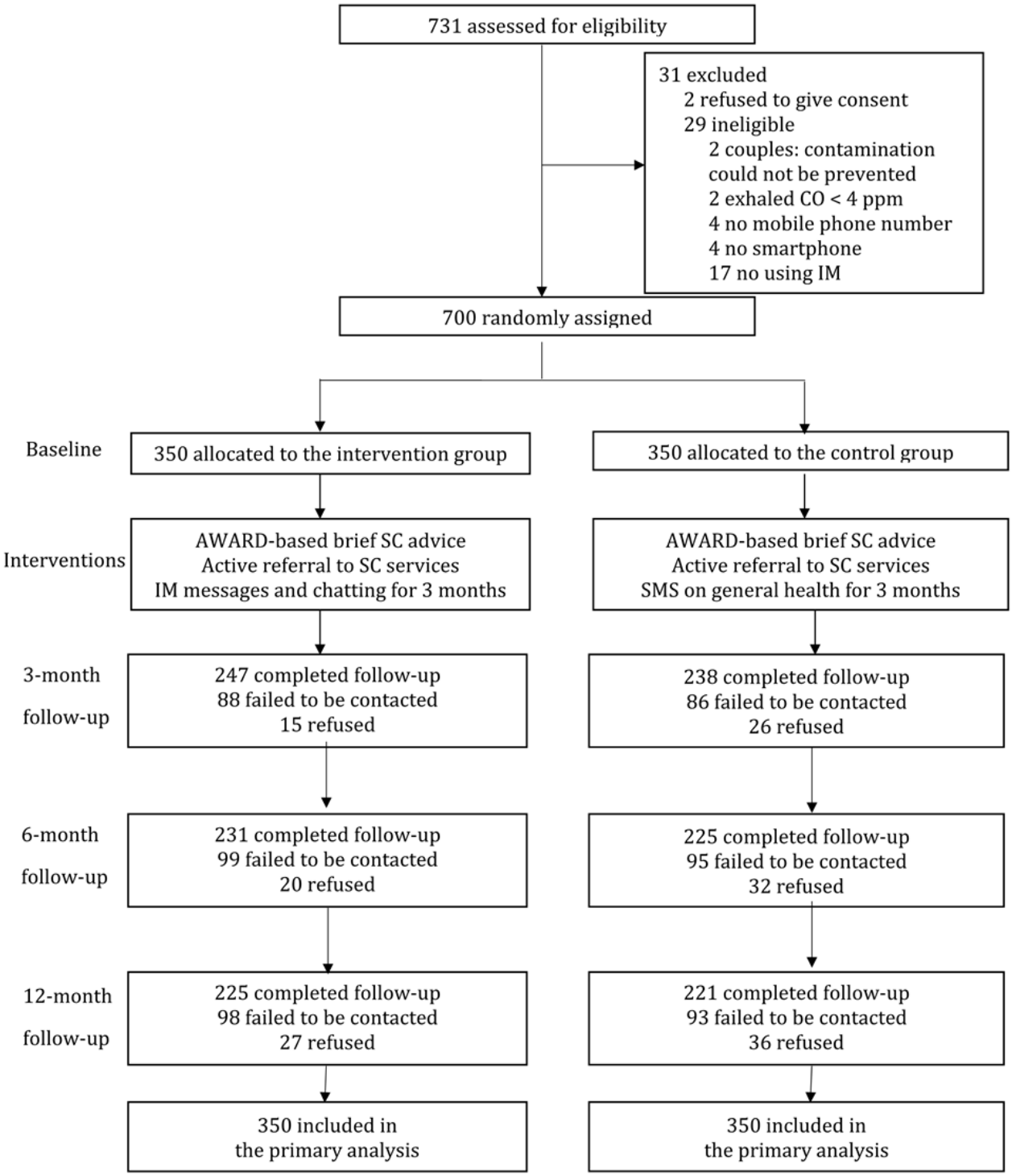

XSL•FO
RenderX

Table 1. Sociodemographic and smoking-related characteristics, and referral to smoking cessation services at baseline (N=700).

Characteristics	Intervention[a] (n=350)	Control[a] (n=350)	*P* value[b]
Sex, n/N (%)			.35
Male	247/350 (70.6)	258/350 (73.7)	
Female	103/350 (29.4)	92/350 (26.3)	
Age (years)			.80
18-29, n/N (%)	40/322 (12.4)	44/326 (13.5)	
30-39, n/N (%)	52/322 (16.1)	62/326 (19.0)	
40-49, n/N (%)	80/322 (24.8)	77/326 (23.6)	
50-59, n/N (%)	82/322 (25.5)	73/326 (22.4)	
60 or above, n/N (%)	68/322 (21.1)	70/326 (21.5)	
Missing, n	28	24	
Highest educational attainment			.80
Primary or below, n/N (%)	41/305 (13.4)	42/310 (13.5)	
Secondary, n/N (%)	214/305 (70.2)	211/310 (68.1)	
Tertiary, n/N (%)	50/305 (16.4)	57/310 (18.4)	
Missing, n	45	40	
Marital status			.18
Single, n/N (%)	85/315 (27.0)	85/314 (27.1)	
Married/cohabited, n/N (%)	216/315 (68.6)	204/314 (65.0)	
Divorced/widow, n/N (%)	14/315 (4.4)	25/314 (8.0)	
Missing, n	35	36	
Children living together			.75
No, n/N (%)	204/297 (68.7)	218/312 (69.9)	
Yes, n/N (%)	93/297 (31.3)	94/312 (30.1)	
Missing, n	53	38	
Housing			.09
Rent, n/N (%)	168/310 (54.2)	196/316 (62.0)	
Owned, n/N (%)	137/310 (44.2)	113/316 (35.8)	
Others, n/N (%)	5/310 (1.6)	7/316 (2.2)	
Missing, n	40	34	
Employment status[c]			.38
Economically active, n/N (%)	213/319 (66.8)	217/310 (70.0)	
Economically inactive, n/N (%)	106/319 (33.2)	93/310 (30.0)	
Missing, n	31	40	
Monthly household income (HK $[d])			.78
Unstable/<19,999, n/N (%)	125/295 (42.4)	137/303 (45.2)	
20,000-29,999, n/N (%)	88/295 (29.8)	85/303 (28.1)	
30,000 or above, n/N (%)	82/295 (27.8)	81/303 (26.7)	
Missing, n	55	47	
Time to the first cigarette after waking			.72
After 60 minutes, n/N (%)	51/330 (15.5)	51/330 (15.5)	
31-60 minutes, n/N (%)	42/330 (12.7)	33/330 (10.0)	

XSL•FO
RenderX

Characteristics	Intervention[a] (n=350)	Control[a] (n=350)	*P* value[b]
6-30 minutes, n/N (%)	94/330 (28.5)	101/330 (30.6)	
Within 5 minutes, n/N (%)	143/330 (43.3)	145/330 (43.9)	
Missing, n	20	20	
Daily cigarette consumption			.70
1-10, n/N (%)	161/349 (46.1)	149/346 (43.1)	
11-20, n/N (%)	147/349 (42.1)	150/346 (43.4)	
21-30, n/N (%)	24/349 (6.9)	31/346 (9.0)	
31 or above, n/N (%)	17/349 (4.9)	16/346 (4.6)	
Missing, n	1	4	
The Heaviness of Smoking Index, median (IQR)[e]	3 (2-4)	3 (2-4)	.48
Previous quit attempts			.04
No attempt, n/N (%)	208/332 (62.7)	183/335 (54.6)	
Had attempts, n/N (%)	124/332 (37.3)	152/335 (45.4)	
Missing, n	18	15	
Intention to quit			.41
Within 7 days, n/N (%)	145/325 (44.6)	135/323 (41.8)	
Within 30 days, n/N (%)	113/325 (34.8)	107/323 (33.1)	
Within 60 days, n/N (%)	21/325 (6.5)	32/323 (9.9)	
Undecided, n/N (%)	46/325 (14.2)	49/323 (15.2)	
Missing, n	25	27	
Perceptions of quitting (score: 0-10)[f]			
Perceived importance of quitting (n=621), median (IQR)	7 (5-10)	7 (5-10)	.71
Perceived difficulty of quitting (n=621), median (IQR)	7 (5-9)	7 (5-9)	.82
Perceived confidence of quitting (n=620), median (IQR)	5 (5-6)	5 (5-6)	.78
Referred to smoking cessation services, n/N (%)	331/350 (94.6)	333/350 (95.1)	.73

[a]Number of missing values excluded from the calculations of the column percentages

[b]*P* values from chi-square tests and Wilcoxon rank-sum tests excluding missing values. Between-group differences were due to randomization (chance); *P* values are for references only.

[c]Being an employer, employee, or self-employed was regarded as economically active; being a student, housekeeper, retired, or unemployed was regarded as economically inactive.

[d]US $1=HK $7.8.

[e]Scored 0-6, with the higher score indicating a higher level of dependence.

[f]Scored 0-10, with the higher score indicating stronger perceptions.

Primary and Sensitivity Analyses

Table 2 indicates that the validated abstinence rates between the intervention and control groups at 6 months (14/350, 4.0%, vs 11/350, 3.1%) and 12 months (19/350, 5.4%, vs 21/350, 6.0%) did not exhibit a significant difference (6-month RR 1.27, 95% CI 0.59-2.77 and 12-month RR 0.90, 95% CI 0.50-1.65). Both groups demonstrated similar rates of self-reported 7-day PPA, smoking reduction by 50% (including self-reported quitters), and any quit attempt at 6 and 12 months. In the intervention group, self-reported 24-week continuous abstinence was notably higher at 6 months (11/350, 3.1%, vs 3/350, 0.9%; RR 3.67, 95% CI 1.03-13.04), while any use of SC services was significantly higher at 12 months (123/350, 35.1%, vs 98/350, 28.0%; RR 1.26, 95% CI 1.01-1.56). Analyses adjusting for previous quit attempts at baseline yielded similar results to those presented earlier. Sensitivity analyses conducted based on complete cases and multiple imputation methods also produced consistent findings regarding abstinence outcomes, as demonstrated in Table S2 in Multimedia Appendix 7. Moreover, outcomes did not exhibit significant differences between the 2 groups at the 3-month mark, as illustrated in Table S3 in Multimedia Appendix 7. For instance, self-reported 7-day PPA was comparable between the intervention and control groups (35/350, 10.0%, vs 25/350, 7.1%; RR 1.51, 95% CI 0.93-2.45).

XSL•FO
RenderX

Table 2. Primary and secondary outcomes in the intervention and control groups (N=700).

Outcomes	Intervention (n=350), n (%)	Control (n=350), n (%)	Crude risk ratio (95% CI)	Adjusted risk ratio[a] (95% CI)
Primary outcomes				
Validated abstinence				
6 months	14 (4.0)	11 (3.1)	1.27 (0.59-2.77)	1.31 (0.60-2.85)
12 months	19 (5.4)	21 (6.0)	0.90 (0.50-1.65)	0.95 (0.52-1.75)
Secondary outcomes				
Self-reported 7-day point prevalence abstinence				
6 months	32 (9.1)	29 (8.3)	1.10 (0.68-1.78)	1.17 (0.73-1.88)
12 months	46 (13.1)	50 (14.3)	0.92 (0.63-1.33)	0.97 (0.67-1.41)
Self-reported 24-week continuous abstinence				
6 months	11 (3.1)	3 (0.9)	3.67 (1.03-13.04)[b]	3.74 (1.07-13.12)[b]
12 months	22 (6.3)	21 (6.0)	1.05 (0.59-1.87)	1.15 (0.66-2.02)
Smoking reduction by at least 50% of baseline[c]				
6 months	100 (28.6)	81 (23.1)	1.23 (0.96-1.59)	1.26 (0.97-1.62)
12 months	111 (31.7)	122 (34.9)	0.91 (0.74-1.12)	0.92 (0.74-1.13)
Any quit attempt since intervention initiation				
6 months	49 (14.0)	36 (10.3)	1.36 (0.91-2.04)	1.45 (0.97-2.16)
12 months	72 (20.6)	63 (18.0)	1.14 (0.84-1.55)	1.21 (0.89-1.63)
Any use of smoking cessation service since intervention initiation				
6 months	90 (25.7)	84 (24.0)	1.07 (0.83-1.39)	1.06 (0.82-1.38)
12 months	123 (35.1)	98 (28.0)	1.26 (1.01-1.56)[b]	1.26 (1.01-1.57)[b]

[a]Adjusted for previous quit attempts.

[b]$P<.05$.

[c]Self-reported quitters included.

Subgroup and Post Hoc Analyses

Table 3 illustrates that among smokers without an intention to quit (ie, those with a planned quit date beyond 30 days or undecided), the intervention group tended to exhibit worse quitting outcomes, although these differences were not statistically significant. Conversely, among participants who had expressed an intention to quit, the intervention group generally demonstrated better quitting outcomes. Specifically, significant results were observed for any quit attempt at 6 months (adjusted RR [ARR] 1.74, 95% CI 1.10-2.76) and any use of SC services at 12 months (ARR 1.31, 95% CI 1.03-1.67). Table S1 in Multimedia Appendix 7 reveals that 74/350 (21.1%) participants in the intervention group engaged in chats with SC advisors during the intervention period. Interestingly, a higher proportion of these participants reported having children living with them and expressed an intention to quit smoking (P=.05 in all cases). Table S4 in Multimedia Appendix 7 demonstrates that among the intervention group participants who engaged in chats with SC advisors, there were notable improvements in various quitting outcomes. Specifically, these participants exhibited significantly higher rates of validated abstinence at 6 months (ARR 3.29, 95% CI 1.13-9.63), smoking reduction by at least 50% (ARR 1.75, 95% CI 1.21-2.55 at 6 months; ARR 1.92, 95% CI 1.36-2.71 at 12 months), and any use of SC services (ARR 1.66, 95% CI 1.14-2.43 at 6 months; ARR 1.67, 95% CI 1.26-2.23 at 12 months). Table S5 in Multimedia Appendix 7 indicates that participants in the intervention group who engaged in chats with SC advisors rated the intervention as more helpful compared with those who did not engage in chats or participants in the control group.

XSL•FO
RenderX

Table 3. Subgroup analyses by intention to quit (planned to quit in 30 days) at baseline (N=648, after excluding 52 participants not reporting intention to quit at baseline).

	No intention to quit (n=148)				Had intention to quit (n=500)			
	Intervention (n=67), n (%)	Control (n=81), n (%)	Crude risk ratio (95% CI)	Adjusted risk ratio[a] (95% CI)	Intervention (n=258), n (%)	Control (n=242), n (%)	Crude risk ratio (95% CI)	Adjusted risk ratio[a] (95% CI)
Validated abstinence								
6 months	1 (1.5)	3 (3.7)	0.40 (0.04-3.81)	0.40 (0.05-3.61)	12 (4.7)	8 (3.3)	1.41 (0.58-3.39)	1.45 (0.60-3.51)
12 months	4 (6.0)	4 (4.9)	1.21 (0.31-4.67)	1.29 (0.35-4.77)	15 (5.8)	17 (7.0)	0.83 (0.42-1.62)	0.87 (0.44-1.71)
Self-reported 7-day point prevalence abstinence								
6 months	5 (7.5)	9 (11.1)	0.67 (0.24-1.92)	0.66 (0.24-1.86)	26 (10.1)	19 (7.9)	1.28 (0.73-2.26)	1.39 (0.79-2.42)
12 months	8 (11.9)	8 (9.9)	1.21 (0.48-3.06)	1.22 (0.48-3.05)	36 (14.0)	39 (16.1)	0.87 (0.57-1.32)	0.93 (0.61-1.41)
Self-reported 24-week continuous abstinence[b]								
6 months	2 (3.0)	0 (0.0)	NA	NA	9 (3.5)	3 (1.2)	2.81 (0.77-10.29)	2.97 (0.82-10.76)
12 months	3 (4.5)	4 (4.9)	0.91 (0.31-3.93)	0.99 (0.24-4.12)	18 (7.0)	15 (6.2)	1.13 (0.58-2.18)	1.23 (0.64-2.34)
Smoking reduction by at least 50% of baseline[c]								
6 months	18 (26.9)	20 (24.7)	1.09 (0.63-1.89)	1.07 (0.62-1.84)	76 (29.5)	57 (23.6)	1.25 (0.93-1.68)	1.29 (0.96-1.73)
12 months	15 (22.4)	24 (29.6)	0.76 (0.43-1.32)	0.76 (0.43-1.33)	86 (33.3)	89 (36.8)	0.91 (0.71-1.15)	0.92 (0.72-1.17)
Any quit attempt since intervention initiation								
6 months	7 (10.4)	11 (13.6)	0.77 (0.31-1.88)	0.77 (0.32-1.89)	41 (15.9)	24 (9.9)	1.60 (1.00-2.57)	1.74 (1.10-2.76)[d]
12 months	11 (16.4)	15 (18.5)	0.89 (0.44-1.80)	0.89 (0.44-1.81)	59 (22.9)	45 (18.6)	1.23 (0.87-1.74)	1.32 (0.94-1.85)
Any use of smoking cessation service since intervention initiation								
6 months	7 (10.4)	12 (14.8)	0.71 (0.29-1.70)	0.70 (0.29-1.69)	79 (30.6)	66 (27.3)	1.12 (0.85-1.48)	1.12 (0.85-1.47)
12 months	11 (16.4)	15 (18.5)	0.89 (0.44-1.80)	0.88 (0.43-1.80)	106 (41.1)	76 (31.4)	1.31 (1.03-1.66)[d]	1.31 (1.03-1.67)[d]

[a]Adjusted for previous quit attempts.

[b]No risk ratio could be calculated as there were no control group participants who had no intention to quit at baseline and achieved 24 weeks of abstinence.

[c]Self-reported quitters included.

[d]$P<.05$.

Quality of Life

Participants reported only minor health issues at baseline, with an overall score of 0.973, which slightly decreased to 0.964 at 12 months. However, there were no statistically significant differences between the intervention and control groups (P=.17 and P=.91, respectively). Furthermore, the area under the receiver operating characteristic curve remained similar between the intervention (0.971) and control (0.967) groups.

Discussion

Principal Findings

The supplementary intervention, comprising personalized regular messages and interactive chatting, did not lead to a significant increase in validated abstinence, self-reported 7-day PPA, smoking reduction (including quitters) by 50%, or quit attempts at both 6 and 12 months, when compared with receiving SMS text messages on general health in addition to brief SC advice and AR. While the intervention group exhibited a significantly higher 24-week continuous abstinence rate at 6 months compared with the control group, this difference was attenuated at the 12-month mark. However, more participants in the intervention group reported using SC services, with a larger effect size observed among those who expressed an intention to quit at baseline.

This study represents one of the few RCTs that provided SC support through a promising yet underutilized mobile health (mHealth) channel—mobile IM apps. Unlike SMS text messaging–based interventions that used only text and can only respond to predefined keywords, this IM-based intervention, guided by the social cognitive theory and the transtheoretical model of change, was integrated into a multicomponent, multimedia, proactive, interactive treatment model. The successful implementation of this intervention in community smokers, many of whom were not actively seeking treatments, suggests that existing SC services could consider integrating IM as an alternative channel to provide SC support. However, while the multicomponent treatment model demonstrated

XSL•FO
RenderX

feasibility, the complex design of this parallel 2-group pragmatic trial limited our ability to isolate the contribution of individual components to cessation outcomes and to assess potential synergies among different components. Factorial designs, wherein participants are randomly assigned to receive either control treatment, IM-based support, AR to SC services, or a combination of these, are necessary to evaluate the additive and interactive effects of the individual components.

Compared with previous studies involving IM-based interventions [7-10,13], we have enhanced our approach by developing 3 sets of messages tailored to the planned quit date. Additionally, we have personalized the messages based on participants' sociodemographic characteristics and smoking behaviors. In contrast to previous studies where the control group typically received minimal treatment [7,10,13], we took a proactive approach by referring all participants in this study to free and effective local SC services. This design was chosen because AR is locally recommended, such as through an online training course funded by the Department of Health for general practitioners, backed by evidence from our prior trials in community smokers [29-31]. By offering AR to both groups, it is possible that the quit rate was boosted, which might have diminished the effectiveness of the additional IM-based intervention.

Efforts to enhance the IM-based intervention included promoting the use of SC services through regular messages and chatting with SC advisors. However, these efforts faced challenges due to social unrest and the COVID-19 pandemic, leading to a delayed increase in SC service utilization. The social unrest in Hong Kong erupted in June 2019, disrupting SC services and leading to ongoing constraints in their operations until the conclusion of this RCT in October 2020. Additionally, 3 waves of COVID-19 outbreaks further compounded the situation, with related social distancing measures such as work-from-home arrangements, bans on public gatherings of more than 4 people, and the closure of many businesses [49]. The provision of SC services in Hong Kong has been further strained by limited resources, exacerbated by factors such as the limited number of centers/clinics serving all 18 districts of Hong Kong. Additionally, most of these facilities operate only during daytime working hours on weekdays, further constraining access to SC support. The prolonged disruption caused by the social unrest and the COVID-19 pandemic could have significantly impacted up to 396 participants recruited in or after May 2019. This is evident in the lower rate of SC service usage at 3 months, which reached a similar level at 6 months [30,31].

The collaboration between one of the local SC providers and the Department of Health, Hong Kong SAR, to implement a pilot fully remote SC treatment is a proactive response to the service disruptions caused by the COVID-19 pandemic. This initiative, which involves phone/videoconferencing counseling and e-mailing NRT since mid-2020, aims to address the constraints posed by limited resources and social distancing measures. More intervention group participants used SC services between the 6- and 12-month marks, a contrast to another RCT that did not use IM to enhance AR, which found minimal difference in service utilization between the 6- and 18-month periods [30]. An uptick in service usage between the 6- and 12-month marks implies that the intervention might have yielded delayed effects on the adoption of SC services. Conducting an additional follow-up at 15- or 18-month intervals could help corroborate this observation.

The post hoc analysis uncovered that engagement, particularly through chatting with SC advisors, seemed to have positively influenced cessation outcomes. This finding aligns with observations from our previous RCT, which offered chat-based support, indicating that chatting could serve as a moderately effective strategy for promoting SC. Although not directly comparable, we did not witness a significant increase in engagement, with nearly 20% engagement observed in both RCTs (74/350, 21.1%, in this study), despite implementing 3 sets of messages tailored to the planned quit date and further personalization involving multimedia. This could be attributed to the fact that many of the recruited smokers had low motivation to quit. A lower motivation to quit has been found to be associated with less desire to use mHealth interventions for quitting [13,50]. We have conducted in-depth qualitative interviews with participants, and the related results, especially from those who did not chat with SC advisors, will be presented elsewhere.

As chatting with SC advisors remained uncommon, regular IM messages appeared to provide no substantially better support than SMS text messages on general health, in addition to brief SC advice and AR. While the mechanisms underlying how messages improve quitting remained understudied, 2 qualitative studies suggested the "feeling of being cared for," that is, the perception of social support, as a potential mechanism [51,52]. The messages also served as a reminder that the individual was trying to quit. Such perceptions of messages suggest that even if the messages are not about SC, they may still increase the quit rate to a certain extent, which appears to be consistent with our observation that control group participants rated the SMS text messages on general health as similarly helpful as did the intervention group participants who did not chat with SC advisors.

Limitations

Our study had some limitations. First, only about two-fifths of participants were not affected by social unrest and the subsequent COVID-19 pandemic in Hong Kong. We conducted post hoc analyses on quitting outcomes in these participants (results not shown) but found no significant between-group differences given a much smaller sample size than the target. We observed some insignificant increases in quit attempts and smoking reduction rates, but there was insufficient statistical power to confirm these trends given the current sample size. Third, we were unable to cross-check the records of using certain SC services due to administrative reasons, but it is worth noting that we referred fewer than 10 participants to those services. We only had a basic record of whether a participant used their chosen service without more detailed information, such as the number of sessions attended and the medication/NRT received. This limitation prevented further in-depth analysis. Although we achieved an acceptable retention rate of over 60% at 12 months (446/700, 63.7%), we cannot rule out nonresponse bias. We conducted multiple imputation analyses, which yielded

XSL•FO
RenderX

results similar to those from the intention-to-treat analyses. Only about two-fifths of the self-reported quitters underwent biochemical validations, a proportion similar to that observed in our recent RCTs of community smokers [13,31], albeit still suboptimal. Because of concerns about potential COVID-19 transmission from mask removal during the exhaled CO test, we implemented stringent infection control measures. These measures included disinfecting the piCO Smokerlyzer immediately before and after each test and maintaining a distance of at least 1.5 m in outdoor venues where the tests were conducted. Additionally, we excluded nondaily smokers from the study, despite the low prevalence of nondaily smoking in Hong Kong (0.6%).

Conclusions

Our RCT demonstrated that an IM-based intervention, featuring personalized psychosocial support and assistance in utilizing SC services alongside brief advice and AR, did not yield additional improvements in quitting rates amidst the challenges posed by social unrest and the COVID-19 pandemic in Hong Kong.

Acknowledgments

We are grateful for the support from the Research Grants Council. We thank the participants, smoking cessation advisors, and research staff who were involved in the study for their contributions. This study is funded by the General Research Fund (No. 17111518) of the Research Grants Council (RGC), University Grants Committee, Hong Kong Special Administrative Region, China. The RGC had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

MPW and THL conceived the study. MPW, YSW, and THL designed the study. YSW coordinated the fieldwork. YSW and MPW had directly accessed and verified the underlying data reported in the manuscript. YSW performed the statistical analysis. YSW wrote the first draft of the manuscript. All authors had full access to and interpreted the data, participated in the critical review of the report, and provided final approval for publication submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Terms for the literature search.
[DOCX File , 14 KB - jmir_v26i1e44973_app1.docx]

Multimedia Appendix 2
Full results of the literature search.
[XLSX File (Microsoft Excel File), 445 KB - jmir_v26i1e44973_app2.xlsx]

Multimedia Appendix 3
Study protocol.
[DOCX File , 49 KB - jmir_v26i1e44973_app3.docx]

Multimedia Appendix 4
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 66 KB - jmir_v26i1e44973_app4.pdf]

Multimedia Appendix 5
Messages for the intervention group.
[XLSX File (Microsoft Excel File), 758 KB - jmir_v26i1e44973_app5.xlsx]

Multimedia Appendix 6
Messages for the control group.
[XLSX File (Microsoft Excel File), 19 KB - jmir_v26i1e44973_app6.xlsx]

XSL•FO
RenderX

Multimedia Appendix 7
Sociodemographic characteristics at baseline, sensitivity and post hoc analyses results, and perceived helpfulness of messages.
[DOCX File , 28 KB - jmir_v26i1e44973_app7.docx]

References

1. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
2. Scott-Sheldon LAJ, Lantini R, Jennings EG, Thind H, Rosen RK, Salmoirago-Blotcher E, et al. Text messaging-based interventions for smoking cessation: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016 May 20;4(2):e49. [doi: 10.2196/mhealth.5436] [Medline: 27207211]
3. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
4. Hoeppner BB, Hoeppner SS, Abroms LC. How do text-messaging smoking cessation interventions confer benefit? A multiple mediation analysis of Text2Quit. Addiction 2017 Apr;112(4):673-682 [FREE Full text] [doi: 10.1111/add.13685] [Medline: 27943511]
5. World Health Organization (WHO). Quitting Toolkit. WHO. URL: https://www.who.int/campaigns/world-no-tobacco-day/2021/quitting-toolkit [accessed 2021-12-04]
6. Cheung YTD, Chan CHH, Lai CJ, Chan WFV, Wang MP, Li HCW, et al. Using WhatsApp and Facebook online social groups for smoking relapse prevention for recent quitters: a pilot pragmatic cluster randomized controlled trial. J Med Internet Res 2015 Oct 22;17(10):e238 [FREE Full text] [doi: 10.2196/jmir.4829] [Medline: 26494159]
7. Xia W, Li H, Cai W, Song P, Zhou X, Lam K, et al. Effectiveness of a video-based smoking cessation intervention focusing on maternal and child health in promoting quitting among expectant fathers in China: a randomized controlled trial. PLoS Med 2020 Sep;17(9):e1003355 [FREE Full text] [doi: 10.1371/journal.pmed.1003355] [Medline: 32991589]
8. Zhao S, Wu Y, Chau S, Fong D, Lam T, Wang M. Mobile chat-based support plus nicotine replacement therapy sampling to promote smoking cessation for community smokers: a randomized controlled trial. Tob Induc Dis 2021;19:32 [FREE Full text] [doi: 10.18332/tid/133373] [Medline: 33927586]
9. Durmaz S, Ergin I, Durusoy R, Hassoy H, Caliskan A, Okyay P. WhatsApp embedded in routine service delivery for smoking cessation: effects on abstinence rates in a randomized controlled study. BMC Public Health 2019 Apr 8;19(1):387. [doi: 10.1186/s12889-019-6727-z]
10. Tang J, Yang J, Liu Y, Liu X, Li L, Sun Y, et al. Efficacy of WeChat-based online smoking cessation intervention ('WeChat WeQuit') in China: a randomised controlled trial. EClinicalMedicine 2023 Jun;60:102009 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102009] [Medline: 37251625]
11. Alsubhi S. The effectiveness of using WhatsApp for smoking relapse prevention among recent quitters. TCTR : Thai Clinical Trials Registry. 2018 Apr 19. URL: https://www.thaiclinicaltrials.org/show/TCTR20180420001 [accessed 2021-12-28]
12. Hong Kong as an information society. Government of Hong Kong/Census and Statistics Department. Hong Kong: Census and Statistics Department, Hong Kong Special Administrative Region; 2020. URL: https://www.statistics.gov.hk/pub/B11100062020AN20B0100.pdf [accessed 2020-10-06]
13. Wang MP, Luk TT, Wu Y, Li WH, Cheung DY, Kwong AC, et al. Chat-based instant messaging support integrated with brief interventions for smoking cessation: a community-based, pragmatic, cluster-randomised controlled trial. Lancet Digit Health 2019 Aug;1(4):e183-e192 [FREE Full text] [doi: 10.1016/S2589-7500(19)30082-2] [Medline: 33323188]
14. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
15. Coa K, Wiseman K, Higgins B, Augustson E. Associations between engagement and outcomes in the SmokefreeTXT program: a growth mixture modeling analysis. Nicotine Tob Res 2019 Apr 17;21(5):663-669 [FREE Full text] [doi: 10.1093/ntr/nty073] [Medline: 29668984]
16. Graham AL, Jacobs MA, Cohn AM, Cha S, Abroms LC, Papandonatos GD, et al. Optimising text messaging to improve adherence to web-based smoking cessation treatment: a randomised control trial protocol. BMJ Open 2016 Mar 30;6(3):e010687 [FREE Full text] [doi: 10.1136/bmjopen-2015-010687] [Medline: 27029775]
17. Patnode CD, Henderson JT, Coppola EL, Melnikow J, Durbin S, Thomas RG. Interventions for tobacco cessation in adults, including pregnant persons: updated evidence report and systematic review for the US preventive services task force. JAMA 2021 Jan 19;325(3):280-298. [doi: 10.1001/jama.2020.23541] [Medline: 33464342]
18. Ahluwalia IB, Tripp AL, Dean AK, Mbulo L, Arrazola RA, Twentyman E, et al. Tobacco smoking cessation and quitline use among adults aged ≥15 years in 31 countries: findings from the Global Adult Tobacco Survey. Am J Prev Med 2021 Mar;60(3 Suppl 2):S128-S135 [FREE Full text] [doi: 10.1016/j.amepre.2020.04.029] [Medline: 33663700]
19. World Health Organization (WHO). WHO Global Report on Trends in Prevalence of Tobacco Use 2000-2025 (4th ed). Geneva, Switzerland: World Health Organization; 2021.

XSL•FO
RenderX

20. Thematic Household Survey Report No. 75. Census and Statistics Department, Hong Kong Special Administrative Region. Hong Kong: Census and Statistics Department, Hong Kong Special Administrative Region; 2022 May. URL: https://www.censtatd.gov.hk/en/data/stat_report/product/B1130201/att/B11302752022XXXXB0100.pdf [accessed 2022-05-26]
21. Leung D, Chan S, Chan V, Lam T. Hardcore smoking after comprehensive smoke-free legislation and health warnings on cigarette packets in Hong Kong. Public Health 2016 Mar;132:50-56 [FREE Full text] [doi: 10.1016/j.puhe.2015.10.007] [Medline: 26612457]
22. Fu SS, van Ryn M, Sherman SE, Burgess DJ, Noorbaloochi S, Clothier B, et al. Proactive tobacco treatment and population-level cessation: a pragmatic randomized clinical trial. JAMA Intern Med 2014 May;174(5):671-677. [doi: 10.1001/jamainternmed.2014.177] [Medline: 24615217]
23. Rigotti NA, Bitton A, Kelley JK, Hoeppner BB, Levy DE, Mort E. Offering population-based tobacco treatment in a healthcare setting: a randomized controlled trial. Am J Prev Med 2011 Nov;41(5):498-503 [FREE Full text] [doi: 10.1016/j.amepre.2011.07.022] [Medline: 22011421]
24. Shelley D, Cantrell J. The effect of linking community health centers to a state-level smoker's quitline on rates of cessation assistance. BMC Health Serv Res 2010 Jan 25;10:25 [FREE Full text] [doi: 10.1186/1472-6963-10-25] [Medline: 20100348]
25. Vidrine JI, Shete S, Cao Y, Greisinger A, Harmonson P, Sharp B, et al. Ask-Advise-Connect: a new approach to smoking treatment delivery in health care settings. JAMA Intern Med 2013 Mar 25;173(6):458-464 [FREE Full text] [doi: 10.1001/jamainternmed.2013.3751] [Medline: 23440173]
26. Gilbert H, Sutton S, Morris R, Petersen I, Galton S, Wu Q, et al. Effectiveness of personalised risk information and taster sessions to increase the uptake of smoking cessation services (Start2quit): a randomised controlled trial. Lancet 2017 Feb 25;389(10071):823-833 [FREE Full text] [doi: 10.1016/S0140-6736(16)32379-0] [Medline: 28129989]
27. Murray RL, Coleman T, Antoniak M, Stocks J, Fergus A, Britton J, et al. The effect of proactively identifying smokers and offering smoking cessation support in primary care populations: a cluster-randomized trial. Addiction 2008 Jun;103(6):998-1006; discussion 1007. [doi: 10.1111/j.1360-0443.2008.02206.x] [Medline: 18422823]
28. Borland R, Balmford J, Bishop N, Segan C, Piterman L, McKay-Brown L, et al. In-practice management versus quitline referral for enhancing smoking cessation in general practice: a cluster randomized trial. Fam Pract 2008 Oct;25(5):382-389. [doi: 10.1093/fampra/cmn046] [Medline: 18689856]
29. Wang MP, Suen YN, Li WH, Lam CO, Wu SY, Kwong AC, et al. Intervention with brief cessation advice plus active referral for proactively recruited community smokers: a pragmatic cluster randomized clinical trial. JAMA Intern Med 2017 Dec 01;177(12):1790-1797 [FREE Full text] [doi: 10.1001/jamainternmed.2017.5793] [Medline: 29059277]
30. Weng X, Luk TT, Suen YN, Wu Y, Li HCW, Cheung YTD, et al. Effects of simple active referrals of different intensities on smoking abstinence and smoking cessation services attendance: a cluster-randomized clinical trial. Addiction 2020 Oct;115(10):1902-1912. [doi: 10.1111/add.15029] [Medline: 32149425]
31. Weng X, Wu Y, Luk TT, Li WHC, Cheung DYT, Tong HSC, et al. Active referral plus a small financial incentive upon cessation services use on smoking abstinence: a community-based, cluster-randomised controlled trial. Lancet Reg Health West Pac 2021 Aug;13:100189 [FREE Full text] [doi: 10.1016/j.lanwpc.2021.100189] [Medline: 34527982]
32. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, CONSORT group, Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008 Nov 11;337:a2390 [FREE Full text] [doi: 10.1136/bmj.a2390] [Medline: 19001484]
33. Cheung Y, Lam T, Li W, Wang M, Chan S. Feasibility, efficacy, and cost analysis of promoting smoking cessation at outdoor smoking "hotspots": a pre-post study. Nicotine Tob Res 2018 Nov 15;20(12):1519-1524. [doi: 10.1093/ntr/ntx147] [Medline: 28655173]
34. Freedman JL, Fraser SC. Compliance without pressure: the foot-in-the-door technique. J Pers Soc Psychol 1966;4(2):195-202. [doi: 10.1037/h0023552]
35. Chan SSC, Wong DCN, Cheung YTD, Leung DYP, Lau L, Lai V, et al. A block randomized controlled trial of a brief smoking cessation counselling and advice through short message service on participants who joined the Quit to Win Contest in Hong Kong. Health Educ Res 2015 Aug;30(4):609-621 [FREE Full text] [doi: 10.1093/her/cyv023] [Medline: 26116584]
36. Chan SSC, Cheung YTD, Wan Z, Wang MP, Lam T. Proactive and brief smoking cessation intervention for smokers at outdoor smoking "hotspots" in Hong Kong. J Cancer Educ 2018 Apr;33(2):365-370. [doi: 10.1007/s13187-016-1085-3] [Medline: 27474113]
37. World Health Organization (WHO). Toolkit for delivering the 5A's and 5R's brief tobacco interventions in primary care. WHO. Geneva, Switzerland: World Health Organization; 2014. URL: https://apps.who.int/iris/handle/10665/112835 [accessed 2022-01-30]
38. Abroms LC, Boal AL, Simmens SJ, Mendel JA, Windsor RA. A randomized trial of Text2Quit: a text messaging program for smoking cessation. Am J Prev Med 2014 Sep;47(3):242-250 [FREE Full text] [doi: 10.1016/j.amepre.2014.04.010] [Medline: 24913220]
39. Naughton F, Cooper S, Foster K, Emery J, Leonardi-Bee J, Sutton S, et al. Large multi-centre pilot randomized controlled trial testing a low-cost, tailored, self-help smoking cessation text message intervention for pregnant smokers (MiQuit). Addiction 2017 Jul;112(7):1238-1249 [FREE Full text] [doi: 10.1111/add.13802] [Medline: 28239919]

XSL•FO
RenderX

40. Smoking cessation information kit 2015 ed. Tobacco and Alcohol Control Office, Department of Health. Hong Kong: Department of Health, HKSAR; 2015. URL: https://www.taco.gov.hk/t/english/quitting/files/kit15.pdf [accessed 2021-12-09]
41. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
42. Abroms LC, Whittaker R, Free C, Mendel Van Alstyne J, Schindler-Ruwisch JM. Developing and pretesting a text messaging program for health behavior change: recommended steps. JMIR Mhealth Uhealth 2015 Dec 21;3(4):e107 [FREE Full text] [doi: 10.2196/mhealth.4917] [Medline: 26690917]
43. Cheung YTD, Wang MP, Li HCW, Kwong A, Lai V, Chan SSC, et al. Effectiveness of a small cash incentive on abstinence and use of cessation aids for adult smokers: a randomized controlled trial. Addict Behav 2017 Mar;66:17-25. [doi: 10.1016/j.addbeh.2016.11.006] [Medline: 27863323]
44. Wong ELY, Ramos-Goñi JM, Cheung AWL, Wong AYK, Rivero-Arias O. Assessing the use of a feedback module to model EQ-5D-5L health states values in Hong Kong. Patient 2018 Apr;11(2):235-247 [FREE Full text] [doi: 10.1007/s40271-017-0278-0] [Medline: 29019161]
45. Free C, Knight R, Robertson S, Whittaker R, Edwards P, Zhou W, et al. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. Lancet 2011 Jul 02;378(9785):49-55 [FREE Full text] [doi: 10.1016/S0140-6736(11)60701-0] [Medline: 21722952]
46. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004 Apr 01;159(7):702-706. [doi: 10.1093/aje/kwh090] [Medline: 15033648]
47. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
48. Rubin D. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ: John Wiley & Sons; 2004.
49. Sun Y, Lam T, Cheung Y, Wang M, Wu Y, Chen J, et al. First report on smoking and infection control behaviours at outdoor hotspots during the COVID-19 pandemic: an unobtrusive observational study. Int J Environ Res Public Health 2021 Jan 25;18(3):1031 [FREE Full text] [doi: 10.3390/ijerph18031031] [Medline: 33503837]
50. Borrelli B, Bartlett YK, Tooley E, Armitage CJ, Wearden A. Prevalence and frequency of mHealth and eHealth use among US and uk smokers and differences by motivation to quit. J Med Internet Res 2015 Jul 04;17(7):e164 [FREE Full text] [doi: 10.2196/jmir.4420] [Medline: 26149323]
51. Douglas N, Free C. 'Someone batting in my corner': experiences of smoking-cessation support via text message. Br J Gen Pract 2013 Nov;63(616):e768-e776 [FREE Full text] [doi: 10.3399/bjgp13X674459] [Medline: 24267860]
52. Grau LE, Pham T, O'Leary T, Weiss J, Toll B, Bernstein SL. Smokers' perspectives on texting for tobacco dependence treatment: a qualitative analysis. Nicotine Tob Res 2017 Mar 01;19(3):307-313. [doi: 10.1093/ntr/ntw184] [Medline: 27613935]

Abbreviations

AR: active referral
ARR: adjusted risk ratio
AWARD: Ask, Warn, Advise, Refer, and Do-it-again
CO: carbon monoxide
CONSORT: Consolidated Standards of Reporting Trials
mHealth: mobile health
NRT: nicotine replacement therapy
PPA: point prevalence abstinence
RCT: randomized controlled trial
RR: risk ratio
SC: smoking cessation
SNOSE: sequentially numbered, opaque, sealed envelope

Edited by T de Azevedo Cardoso; submitted 11.12.22; peer-reviewed by N Jiang, O Byambasuren; comments to author 09.06.23; revised version received 28.09.23; accepted 26.03.24; published 13.05.24.

Please cite as:
Wu YS, Cheung YTD, Lee JJJ, Wong CKH, Ho SY, Li WHC, Yao Y, Lam TH, Wang MP
Effect of Adding Personalized Instant Messaging Apps to a Brief Smoking Cessation Model in Community Smokers in Hong Kong: Pragmatic Randomized Clinical Trial
J Med Internet Res 2024;26:e44973
URL: https://www.jmir.org/2024/1/e44973
doi:10.2196/44973
PMID:38739429

XSL•FO
RenderX

©Yongda Socrates Wu, Yee Tak Derek Cheung, Jay Jung Jae Lee, Carlos King Ho Wong, Sai Yin Ho, William Ho Cheung Li, Ying Yao, Tai Hing Lam, Man Ping Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Assessing the Content and Effect of Web-Based Decision Aids for Postmastectomy Breast Reconstruction: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Lin Yu[1], MD; Jianmei Gong[1], PhD; Xiaoting Sun[1], MD; Min Zang[1], MD; Lei Liu[1*], PhD; Shengmiao Yu[2*], BS

[1]School of Nursing, Liaoning University of Chinese Traditional Medicine, Shenyang, China

[2]Outpatient Department, The Fourth Affiliated Hospital of China Medical University, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Lei Liu, PhD
School of Nursing, Liaoning University of Chinese Traditional Medicine
No.79 Chongshan Dong Road
Shenyang, 110000
China
Phone: 86 17824909908
Email: liulei0428@sina.com

Abstract

Background: Web-based decision aids have been shown to have a positive effect when used to improve the quality of decision-making for women facing postmastectomy breast reconstruction (PMBR). However, the existing findings regarding these interventions are still incongruent, and the overall effect is unclear.

Objective: We aimed to assess the content of web-based decision aids and its impact on decision-related outcomes (ie, decision conflict, decision regret, informed choice, and knowledge), psychological-related outcomes (ie, satisfaction and anxiety), and surgical decision-making in women facing PMBR.

Methods: This systematic review and meta-analysis followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. A total of 6 databases, PubMed, Embase, Cochrane Library, CINAHL, PsycINFO, and Web of Science Core Collection, were searched starting at the time of establishment of the databases to May 2023, and an updated search was conducted on April 1, 2024. MeSH (Medical Subject Headings) terms and text words were used. The Cochrane Risk of Bias Tool for randomized controlled trials was used to assess the risk of bias. The certainty of evidence was assessed using the Grading of Recommendations, Assessment, Development, and Evaluation approach.

Results: In total, 7 studies included 579 women and were published between 2008 and 2023, and the sample size in each study ranged from 26 to 222. The results showed that web-based decision aids used audio and video to present the pros and cons of PMBR versus no PMBR, implants versus flaps, and immediate versus delayed PMBR and the appearance and feel of the PMBR results and the expected recovery time with photographs of actual patients. Web-based decision aids help improve PMBR knowledge, decisional conflict (mean difference [MD]=–5.43, 95% CI –8.87 to –1.99; P=.002), and satisfaction (standardized MD=0.48, 95% CI 0.00 to 0.95; P=.05) but have no effect on informed choice (MD=–2.80, 95% CI –8.54 to 2.94; P=.34), decision regret (MD=–1.55, 95% CI –6.00 to 2.90 P=.49), or anxiety (standardized MD=0.04, 95% CI –0.50 to 0.58; P=.88). The overall Grading of Recommendations, Assessment, Development, and Evaluation quality of the evidence was low.

Conclusions: The findings suggest that the web-based decision aids provide a modern, low-cost, and high dissemination rate effective method to promote the improved quality of decision-making in women undergoing PMBR.

Trial Registration: PROSPERO CRD42023450496; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=450496

(J Med Internet Res 2024;26:e53872) doi:10.2196/53872

KEYWORDS

decision aids; internet; postmastectomy breast reconstruction; decision conflicts; mobile phone

XSL•FO
RenderX

Introduction

Background

Breast cancer (BC) is a major global health problem. In 2020, more than 2.3 million newly diagnosed cases and 685,000 deaths were associated with BC [1]. There has been a gradual increase in the incidence of BC globally over the past few decades, which has been attributed to lifestyle changes (eg, increase in BMI and decrease in birth rate), as well as an increase in screening detection as BC becomes more recognized [2-4]. Although BC has the highest incidence rate among all types of cancer, its mortality rate declined by 43% between 1989 and 2020, and it is concentrated in larger areas [2,5]. Advances in the early detection and treatment of BC have improved patient survival rates, which, in turn, have led to an increased focus on improving the quality of life of the survivors of BC. The surgical approach to BC is complex and usually involves the decision to undergo breast-conserving surgery and mastectomy. For women undergoing mastectomy, the change in appearance due to the missing breast can lead to various types of psychological problems including physical imagery discomfort, psychological distress, anxiety, and depression [6]. Postmastectomy breast reconstruction (PMBR) is now an option for women to restore their appearance [7].

However, when women face a PMBR decision, they must decide whether to use PMBR, and if they choose to do so, they should further decide on the timing and type of PMBR (ie, implant, autologous tissue, or combination) [8,9]. Delayed autologous PMBR results in a localized or regional recurrence rate similar to immediate PMBR [10]. A BC diagnosis can leave patients feeling anxious and uncertain, which is often exacerbated by presenting multiple, complex treatment options for women to choose from in a short period [11]. Most patients with BC who are considering PMBR immediately have clinically substantially decision conflict [12,13]. Patients experience postoperative complications leading to decision regret [14]. These issues can lead to poorer health outcomes, negative perceptions of the health care system, and lower quality of life [14,15]. Therefore, more preoperative patient education about possible complications includes the patient's anatomy, which PMBR to choose, the associated pros and cons, and previous surgical and medication history. Women should be fully informed of their options and given the tools to weigh the pros and cons of each option, which may reduce the incidence of these adverse effects [16]. At the same time, personalized medicine is increasingly becoming the standard of care for patients with BC [17], and based on the current evidence, patients should have equal access to all eligible PMBR options [10]. In a sample of 126 patients who underwent mastectomy, a minority of patients made high-quality decisions about PMBR. Specifically, 43.3% of patients were adequately informed and accepted treatment decisions that were consistent with their preferences [11]. Therefore, patients and providers must work together through dialogue to optimize treatment options and engage in shared decision-making. However, it is not easy for inexperienced physicians to perform shared decision-making in an orderly and correct manner in a limited amount of time [18]. Decision aids may be helpful before a patient decides to undergo PMBR. Some studies [19] also suggest that decision aids may be helpful for some women even after undergoing a PMBR, as some women exhibit decision conflicts after the consultation. Decision aids are powerful tools to support patients in making informed choices based on their own values and are available via the internet, DVDs, and printed materials [20]. With the increasing popularity of the internet worldwide, web-based dissemination of information has been recognized as one of the most promising of all available formats (eg, leaflets, brochures, audio, and video) for delivering decision aids to patients. Web-based decision aids are characterized as being interactive, dynamic, and customizable [21]. On the one hand, web-based decision aids have a greater advantage in facilitating patient access than face-to-face interaction with physicians. On the other hand, decision aids on the internet can store and disseminate information over a longer period than traditional, static decision aids and can personalize the visit according to the patient's values and preferences [21-23].

Prior Research

Paraskeva et al [24] conducted a systematic review exploring the effectiveness of interventions to assist women in making decisions about PMBR, which consisted of 6 studies with mixed results in terms of knowledge, decision-making, overall satisfaction, and quality of life. Berlin et al [25] assessed PMBR decision aids in a systematic review and meta-analysis, concluding that PMBR reduces decision conflict, improves information satisfaction, promotes participation in the decision-making process, and enhances the awareness of participation in the decision-making process. However, the authors included all types of trials (ie, quantitative and qualitative) and only meta-analyzed decision conflict. This review also did not include the effects of decision aids on outcome indicators such as psychologically relevant outcomes. Yang et al [26] conducted a meta-analysis exploring the effects of decision aids on decision-making in PMBR; however, the authors did not compare whether different forms of decision aids would have different effects. Zhao et al [27] conducted a scoping review with the aim of reviewing, comparing, and discussing the current incorporation of the adverse effects of BC treatments into decision aids and examined how web-based decision aids personalized BC treatment decision-making tools in patient–health care provider communication, clinician decision-making processes, and shared decision-making, as yet unassessed patient outcomes (eg, knowledge and anxiety). In summary, there is a lack of descriptions of the impact of web-based decision aids on the decision-making of women facing PMBR. Overall, existing systematic evaluations on related topics have produced mixed results, and more importantly, many primary trials [28-31], following these reviews, have produced conflicting results, which may provide new evidence. Therefore, there is a need for a new systematic evaluation to provide a comprehensive overview of the effectiveness of web-based decision aids on the quality of decision-making for women faced with PMBR, drawn from all available evidence from randomized controlled trials (RCTs) that meet high standards for evidence-based research.

XSL•FO
RenderX

Objectives

The aim of this systematic review and meta-analysis was to assess the content of the web-based decision aids and evaluate their effectiveness on decision-related outcomes (ie, decision conflict, decision regret, informed choice, and knowledge), psychological-related outcomes (ie, satisfaction and anxiety), and surgical decision-making in women facing PMBR.

Methods

This is a systematic review and meta-analysis reported in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 1) guidelines [32]. The protocol was registered in PROSPERO (CRD42023450496).

Eligibility and Exclusion Criteria

An overview of the inclusion and exclusion criteria can be found in Textbox 1.

Textbox 1. Eligibility and exclusion criteria.

Population

- The population included in the study was aged ≥18 years and women who had been diagnosed with breast cancer (BC) and were considering postmastectomy breast reconstruction (PMBR) but had not yet had the surgery and had internet access. If the patient, at the time of enrollment, had attempted PMBR; did not have BC (ie, were considering prophylactic mastectomy); and had an active psychiatric, cognitive, or visual impairment, they were not eligible.

Intervention

- Studies focusing on web-based decision aids (including websites and apps)

Comparison

- Controls for usual care, counseling, health education pamphlets, and non–web-based decision aids

Outcome

- The primary outcomes were decision-related outcomes (ie, informed choice, knowledge, decision conflict, and decision regret); psychological outcomes (ie, satisfaction and anxiety); and PMBR options and tool usability (ie, women's feedback on use)

Study

- Randomized controlled trials

Search Strategy

A systematic search of studies was carried out using English databases such as PubMed, Embase, Cochrane Library, CINAHL, PsycINFO, and Web of Science Core Collection from the date of inception of each database to May 2023, and an updated search was conducted on April 1, 2024, to cover new research. Medical Subject Headings terms and text words were used. The keywords used included "Mastectomy," "mammaplasty," "mastectomy," "informed choice*," "shared decision making," "computer," "web based," and "Internet," which are English search terms. These index terms and keywords were explored and modified according to the different grammatical rules of the database. Specific details of the search algorithm are available in Multimedia Appendix 2. The reference lists of the included studies and relevant articles were hand-searched to identify other potentially eligible articles. The search was limited to articles in English and had no limitations with regard to publication year.

Screening

The results were input into EndNote X9, and duplicates were removed automatically. After removing duplicates, 2 reviewers independently screened the titles and abstracts of identified articles and removed irrelevant citations in accordance with the selection criteria. After the removal of irrelevant studies, the full texts of potentially relevant studies were retrieved. Next, both reviewers independently assessed the full texts. Any disagreements were settled by discussion with a third reviewer.

Data Extraction

Characteristics of the included RCTs (eg, author, year of publication, country, sample size, subject characteristics, form, content, development method and team, theoretical basis, duration of use, reading level, a brief description of the intervention in the control group, outcome measurements, follow-up, and results) were extracted into tables. We wrote to the authors to obtain more information about the results. Two reviewers compared the findings independently.

Risk of Bias Assessment

The quality of RCTs was evaluated using the Cochrane Handbook for RCTs [33]. The tool consists of 7 items: randomized sequence generation, allocation concealment, participant and personnel blinding, blinding for outcome assessment, incomplete outcome, data selective reporting, and other bias. The risk of bias for each domain was judged as low risk of bias, high risk of bias, or unclear risk of bias. The evaluation of study quality was performed independently by 2 reviewers, and a third reviewer was consulted if necessary.

XSL•FO
RenderX

Statistical Analysis

Statistical analysis was performed using Review Manager (version 5.3; Cochrane), illustrated using a forest plot when at least 2 studies were measured for the same outcomes for a PMBR decision at the longest follow-up time point [34,35]. We used mean differences (MDs) for continuous variables that were measured with the same instrument, standardized MDs (SMDs) when a similar outcome was assessed with different instruments, and relative risks for dichotomous variables. We calculated possible missing values such as SD and 95% CI [33]. In the study, heterogeneity was assessed via the Higgins I^2 statistic with I^2 values of ≤25%, 50%, and ≥75% deemed to represent low, medium, and high heterogeneity, respectively [33]. When there was no significant heterogeneity, the fixed effects model (I^2≤50%) was used; otherwise, the random effects model was used, resulting in a more conservative summary effect estimate [33]. To identify potential sources of clinical heterogeneity, we also conducted a post hoc sensitivity analysis to determine the stability of the results by omitting each test [36].

Results

Study Selection

Figure 1 shows the research selection process and results based on the PRISMA 2020 guidelines. A total of 844 studies were identified. A total of 129 of these studies were excluded because they were repetitive. After selecting titles and abstracts, 21 studies were included for the next stage. Consequently, 7 studies met the inclusion criteria.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

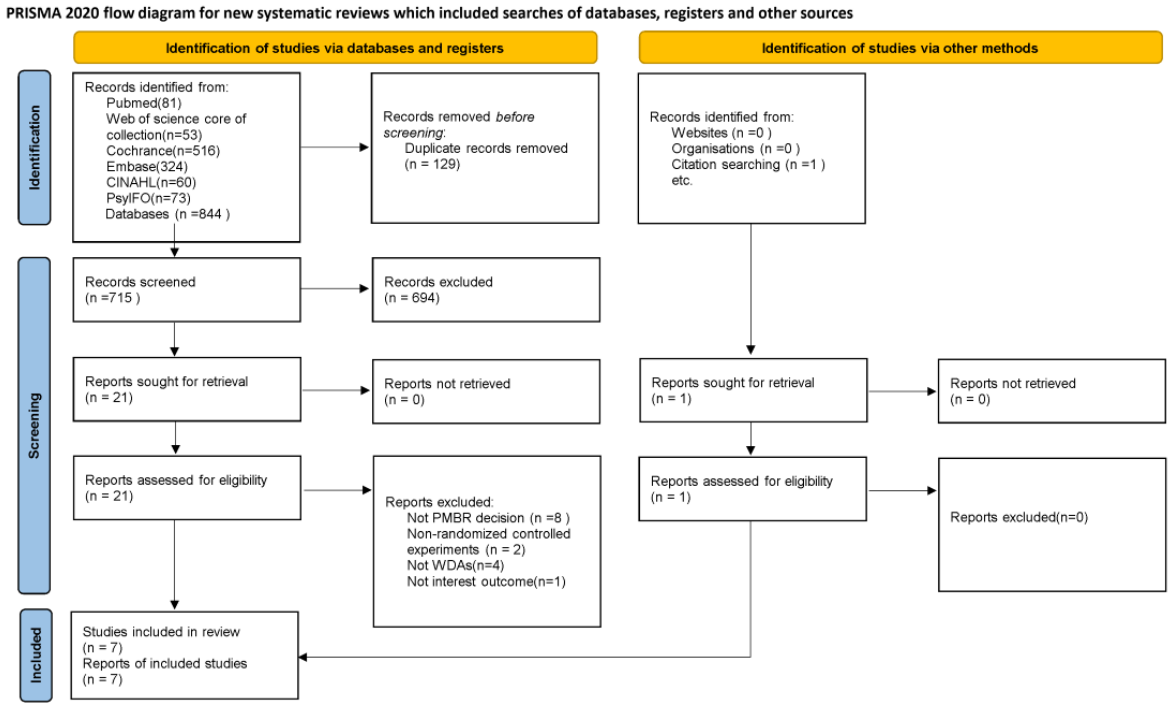

Study Characteristics

The 7 studies included 579 women and were published between 2008 and 2023, and the sample size in each study ranged from 26 to 222. The average age of the women was approximately 50 years; they were in the early stages of BC and facing the PMBR decision. The studies were conducted in 3 countries; 6 studies were conducted in high-income countries—4 in the United States [30,31,37,38] and 2 in Australia [29,39]—and 1 in an upper–middle-income country, China [28]. Detailed characteristics of the included studies are shown in Table 1.

XSL•FO
RenderX

Table 1. Basic characteristics of the included studies.

Study and country	Population	Sample size, n		Outcome measures	Timing	Outcome
		Experimental	Control			
Fang et al [28], China	Aged ≥20 years, recently diagnosed with BC[a], candidates for mastectomy, and able to read and speak Taiwanese or Mandarin	48	48	DCS[b], BR-DMPS[c], DRS[d], BIS[e], and HADS[f]	1 week, 1.8 months, and 12 months	Understanding of medical information, DC[g], anxiety, depression, DR[h], and body image distress
Heller et al [37], United States	Unable to read, write, and speak English, had previously viewed the interactive digital education aid, and the initial consultation was for the purpose of obtaining a second opinion	66	67	Knowledge and satisfaction (self-administered scales) and STAI[i]	1 month after surgery	Knowledge, satisfaction, acquiring information about BR[j], and anxiety
Manne et al [38], United States	Women, aged >18 years, English speaking, diagnosed with ductal carcinoma in situ or stage 1, 2, or 3a BC, considering mastectomy but had not yet had the surgery and had home internet access or willingness to use the patient education center computer to access the web-based decision aids, if assigned to this condition	31	24	Knowledge and satisfaction (self-administered scales), DCS, and STAI	2 weeks	BR knowledge, preparation to make a decision, DC, anxiety, and BR intentions
Mardinger et al [29], Australia	Planning to undergo or having already undergone a therapeutic or prophylactic mastectomy, considering BR, aged >18 years, a proficient in English, and with internet access	30	30	DCS, DRS, and the Satisfaction with Decision Scale	6 weeks and 6 months	Satisfaction with decision, DC, and DR
Politi et al [30], United States	Adult (≥18 years) English-speaking women with stages 0-III BC, who were considering a referral or were referred to 1 of 4 plastic or reconstructive surgeons at a single academic medical center to discuss BR	60	60	DQI[k] the 4-item SURE DCS, the BREAST-Q reconstruction module (preoperative; version 1.0), and 3 items from the patient activation measure	2 months	Knowledge, DC, decision process quality, preferences and preference concordance, quality of life, patient activation, and shared decision-making
Sherman et al [39], Australia	Women diagnosed with BC or ductal carcinoma in situ, who had been advised to undergo and had already undergone a mastectomy, were aged >18 years, were English language competent for reading and writing, had no prior breast surgery (eg, reconstruction or augmentation), and had internet access	116	106	DCS, satisfaction with reconstruction-related information assessed by a 5-item scale, DRS, DASS-21[l], and SSQ-6[m]	1.6 months	DC, satisfaction with information, and DR
Varelas et al [31], United States	English-speaking adult women aged >18 years who had been diagnosed with BC (stage I or II only) and had been advised to undergo or had already undergone a mastectomy	13	13	A short knowledge assessment test, STAI, DCS, and BREAST-Q reconstruction module	12 to 18 weeks, 28 weeks, and 36 weeks of pregnancy, and again 6-8 weeks postnatally	Knowledge, psychological status, DC, satisfaction with information provided to a patient by their surgeon, and time of consultation

[a]BC: breast cancer.

XSL•FO
RenderX

[b]DCS: Decision Conflict Scale.
[c]BR-DMPS: Breast Reconstruction-Decision-Making Process Scale.
[d]DRS: Decision Regret Scale.
[e]BIS: Body Image Scale.
[f]HADS: Hospital Anxiety and Depression Scale.
[g]DC: decisional conflict.
[h]DR: decisional regret.
[i]STAI: State-Trait Anxiety Inventory.
[j]BR: breast reconstruction.
[k]DQI: Decision Quality Index.
[l]DASS-21: Depression Anxiety Stress Scale.
[m]SSQ-6: Social Support Questionnaire.

Characteristics of the Interventions and Controls

The characteristics of the interventions and controls are shown in Multimedia Appendix 3 [28-31,37-39].

Characteristics of the Interventions

In total, 5 of the studies [28-30,38,39] explained that the web-based decision aids development team includes survivors of BC who have undergone mastectomy, plastic or reconstructive surgeons who perform PMBR, and software engineers. The methodology used to develop web-based decision aids includes qualitative research, evidence review and mentoring, and pilot study group meetings. The theoretical basis for the development of web-based decision aids is usually the International Patient Decision Aid Standards [29,30,39] or the Ottawa Decision Support Framework [28]. Except for 2 studies [28,37] that did not report the time of use, most web-based decision aids took between 20 and 74 minutes. Two web-based decision aids [29,30,39] were developed at a reading level written at a seventh- and eighth-grade reading level. The web-based decision aids content specifically includes the patient population and reconstruction options, including implant reconstruction (ie, tissue expanders and implant types), autologous flap reconstruction (ie, latissimus dorsi, rectus abdominis, and free flaps and deep epithelial perforator flaps in the lower abdomen), and skin-sparing and preserving mastectomies (ie, 1-phase and 2-phase procedures). There are also contraindications and general eligibility criteria. Timing of reconstruction includes immediate versus delayed reconstruction, as well as factors that influence the type and timing of reconstruction. It also includes information about the pros and cons of reconstruction versus no reconstruction, implants versus flaps, immediate versus delayed reconstruction, the look and feel of PMBR, and the expected recovery time. The probability of possible implant (eg, wrinkled breast appearance, periosteal contracture after radiation therapy, and possible need for implant replacement over time) and flap (eg, muscle weakness and flap failure) are clearly described in a balanced format with quotes of real patients' opinions. The web-based decision aids show photographs, high-quality 3D animated images, pre- and postoperative photographs, audio, and video of actual patients of different skin colors and body types, A list of frequently asked questions from clinicians is also included. Elements in the tool include patient-tailored risk assessments, patient value clarification exercises, techniques for managing emotions, and strategies for communicating with family members about PMBR decisions. Women's stories explaining their reasons for choosing particular methods and their impact on their lives are also included. Users enter their questions and the system prompts them to print a summary to use in a consultation with their physician. This customized printable page also helps patients discuss their concerns and options with their families.

Characteristics of the Controls

The control for the study by Politi et al [30] was the enhanced urgent care and American Society of Plastic Surgeons pamphlet on PMBR. Varelas et al [31] used traditional counseling. The control for the study by Fang et al [28] was the provider-provided urgent care+pamphlet, which describes the types of surgery, including mastectomy, implant-based PMBR, and autologous PMBR, as well as the advantages and disadvantages of the different types of surgeries. The control for the study by Manne et al [38] was the 56-page pamphlet available at no cost from the Cancer Support Community focusing on PMBR. For the study by Sherman et al [39], the control was the web-based access to excerpts of the public brochure, including basic information on breast surgery and reconstruction, but excluding content unique to the intervention group (ie, video interviews with patients or surgeons, and values clarification exercises). In the study by Mardinger et al [29], the control was the decision aids, which is unvalidated that contains 6 text-based pages that can be accessed in both interactive and noninteractive formats. The control for the study by Heller et al [37] was the group that received the standard patient education, including printed materials in books and pamphlets as well as personal instruction from the attending physician, physician-in-training, physician assistant, and nurse practitioner.

Outcome Measure

A total of 5 studies [28,29,31,38,39] measured decision conflict using the Decision Conflict Scale (DCS), and 1 study [30] measured decision conflict using the 4-item SURE scale. Three studies [28,29,39] measured decision regret using the Decision Regret Scale (DRS) and 2 studies [28,29] measured informed choice using the subdimension of the DCS—feeling informed. Knowledge was measured primarily by the percentage of correct answers to self-administered multiple-choice questions about specific plastic surgery procedures in 6 studies [28-31,37,38]. Satisfaction was measured using the Satisfaction with Decision Scale [29] and some scales adapted from those used in previous

XSL•FO
RenderX

studies [28,37-39]. Anxiety was primarily measured using the Hospital Anxiety and Depression Scale [28] and the State-Trait Anxiety Inventory [31,38].

Decision-Related Outcomes

Decision Conflict

In total, 6 studies [28-31,38,39] investigated the impact of decision conflicts in PMBR. The 5 studies [28,29,31,38,39] that used DCS included in the meta-analysis showed a statistically significant positive impact of web-based decision aids interventions on decision conflict (MD=–5.43, 95% CI –8.87 to –1.99; *P*=.002). Heterogeneity experiments indicated that there was evidence of statistical heterogeneity in the expected summary results (I^2=63%; Figure 2). Politi et al [30] used the 4-item SURE DCS and reported that there was no difference between the 2 groups in terms of decisional conflict (*P*>.05).

Figure 2. Forest plot comparing decision conflict [28,29,31,38,39], decision regret [28,29,39], and informed choice [28,29] in the web-based decision aids group versus control groups.

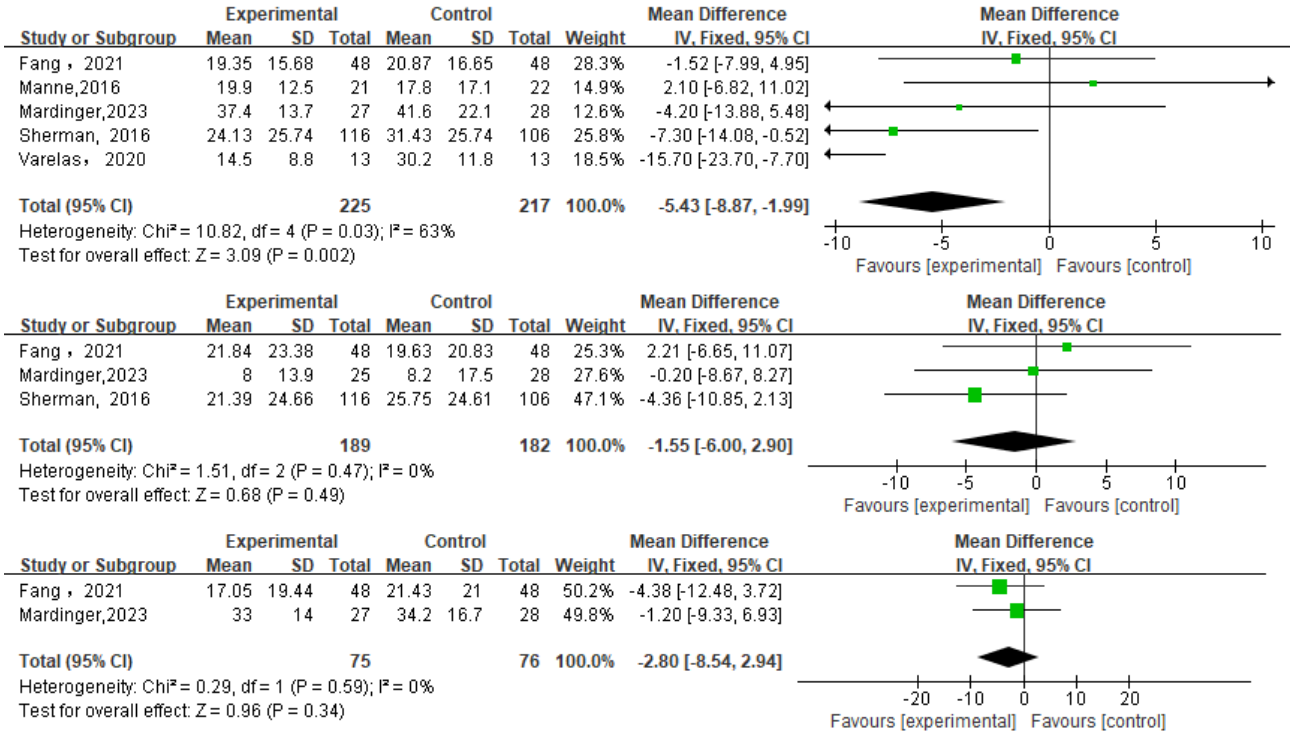

Decision Regret

In total, 3 studies [28,29,39] used DRS to investigate the impact of decision regret in PMBR. The meta-analysis showed that the difference in decision regret after the intervention was not statistically significant compared with the control group (MD=–1.55, 95% CI –6.00 to 2.90; *P*=.49). Heterogeneity experiments indicated that there was evidence of no statistical heterogeneity in the expected summary results (I^2=0%; Figure 2).

Informed Choice

In total, 2 studies [28,29] investigated the impact of informed choice by DCS in PMBR surgery. The meta-analysis showed that the difference in informed choice after the intervention was not statistically significant compared to the control group (MD=–2.80, 95% CI –8.54 to 2.94; *P*=.34). Heterogeneity experiments indicate that there was evidence of no statistical heterogeneity in the expected summary results (I^2=0%; Figure 2).

Knowledge

We did not conduct a meta-analysis of knowledge as an outcome because most of the instruments measuring knowledge were self-administered. The study by Heller et al [37] found significantly higher levels of knowledge in the web-based decision aids group, with a mean increase in correctly answered questions of 14% compared to 8% in the control group (*P*=.02). Politi et al [30] found that participants using web-based decision aids had higher objective knowledge, answering an average of 85% (9.35/11) of the questions correctly compared to 58% (6.35/11) in the control group (*P*<.001). Similarly, Varlas et al [31] showed improved knowledge assessment scores in both groups but significantly higher knowledge assessment scores in the intervention group (control=70.8%, SD 15.5%; intervention=83.1%, SD 13.8%; *P*=.02). However, Manne et al [38] reported similar effects of web-based decision aids on PMBR knowledge versus the booklet, and Fang et al [28] also reported no difference in the amount of PMBR-related medical information between web-based decision aids and the control group at 1 week after consultation (*P*=.13), suggesting that women in both groups had a similar level of comprehension of medical information, whether using the booklet alone or in combination with the web-based decision aids. Mardinger et al [29] also reported that both groups had similar scores on the true or false PMBR knowledge questionnaire over time (*P*>.05).

XSL•FO
RenderX

Psychological Outcomes

Satisfaction

In total, 5 studies [28,29,31,38,39] used different scales to investigate the impact of satisfaction. The meta-analyses indicated that web-based decision aids may improve current form:satisfaction compared to controls, but the results were not statistically significant (SMD=0.48, 95% CI 0.00 to 0.95; *P*=.05). Heterogeneity experiments indicated that there was evidence of statistical heterogeneity in the expected summary results (I^2=79%; Figure 3). Similarly, Heller et al [37] reported a higher level of satisfaction with the way in which information about PMBR was obtained in the web-based decision aids group than in the control group (*P*=.03).

Figure 3. Forest plot comparing satisfaction [28,29,31,38,39] and anxiety [28,31,38] in the web-based decision aids group versus control groups.

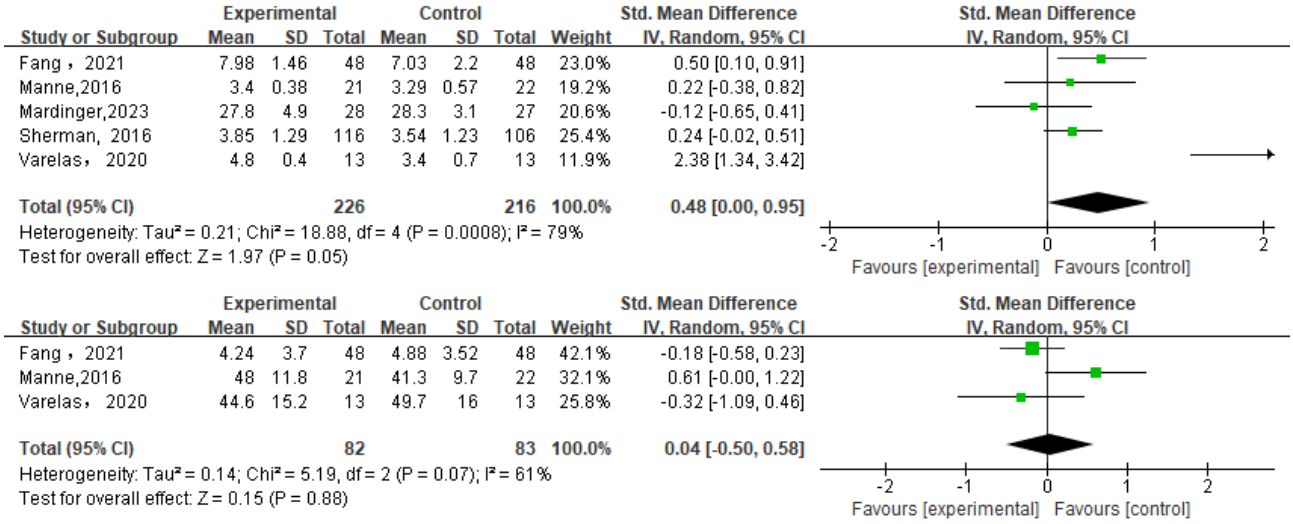

Anxiety

A total of 3 studies used Hospital Anxiety and Depression Scale and State-Trait Anxiety Inventory [28,31,38] to investigate the impact of anxiety in PMBR. The meta-analysis showed that there was no statistically significant difference in the combination of SMD after intervention (SMD=0.04, 95% CI –0.50 to 0.58; *P*=.88). Heterogeneity experiments indicated that there was evidence of statistical heterogeneity in the expected summary results (I^2=61%; Figure 3). Heller et al [37] reported that in the web-based decision aids group, there was a trend toward lower levels of anxiety between the preoperative and postoperative visits, but the difference between the groups was not significant as determined by generalized estimating equation modeling.

Choice of Surgery

The surgical choices differed between the two groups in the study by Fang et al [28]: 56% (27/48) in the web-based decision aids group and 46% (22/48) in the control group opted for immediate PMBR (*P*=.05). In addition, most patients chose implantable PMBR, with no difference between groups. Notably, the web-based decision aids group in the study by Mardinger et al [29] was unbalanced in terms of the choice of type of PMBR, with 10 (36%) women in the web-based decision aids group refusing PMBR compared with 6 (21%) women in the control group (*P*=.20). The results of the study by Politi et al [30] showed that 95 (79.2%) women underwent reconstruction; among them, nearly all (92/95, 97%) underwent immediate PMBR, and there were no differences between groups in median preference scores for reconstruction, type, or time.

Evaluation of the Intervention

In total, 3 studies reported different benefits of web-based decision aids compared to controls. Heller et al [37] reported an upward trend in the number of patients in the web-based decision aids group who reported that they received all the necessary information and improved their ability to choose a PMBR plan, but the difference between the groups was not significant. Manne et al [38] reported that 81% of participants in the web-based decision aids found logging in and navigating easy and the length of time was rated as “just right,” and that the web-based decision aids were more helpful, interesting, and valuable than the brochures. Sherman et al [39] found that women in the intervention group found the web-based decision aids to be 2.94 (SD 0.76) informative, very useful, easy to use, contained enough information, and helped them to clarify their reconstruction ideas. However, Varelas et al [31] reported that surgeon satisfaction was also significantly higher in the intervention group than in the control group. Meanwhile, consultation time was shorter in the intervention group, but the difference was not statistically significant (*P*=.46). Similarly, Politi et al [30] reported no difference between the web-based decision aids group and the control group in terms of mean counseling time after the intervention (29.7 vs 30.0 minutes; *P*>.05). Mardinger et al [29] showed that although women used both decision aids with comparable frequency, the total time spent counseling and the time spent per counseling session was significantly greater for women in the intervention than that for the control group (*P*<.05). Women in the study by Fang et al [28] indicated no difference between the 2 groups in terms of perceived impact and utility of web-based decision aids on PMBR decisions.

XSL•FO
RenderX

Sensitivity Analysis

We conducted sensitivity analyses of decision conflict, satisfaction, and anxiety by removing each study. Sensitivity analysis showed that for decision conflict and satisfaction, after removing 1 study [31], contrary to the previous results, web-based decision aids did improve satisfaction (the I^2 range was 79%-12%) but did not improve decision conflict (the I^2 range was 63%-2%). We found that by removing the study by Manne et al [38], the stability of anxiety did not change but the heterogeneity was reduced from 62% to 0% (Figure 4).

Figure 4. Sensitivity analysis of decision conflict [28,29,38,39], satisfaction [28,29,38,39], and anxiety [28,31].

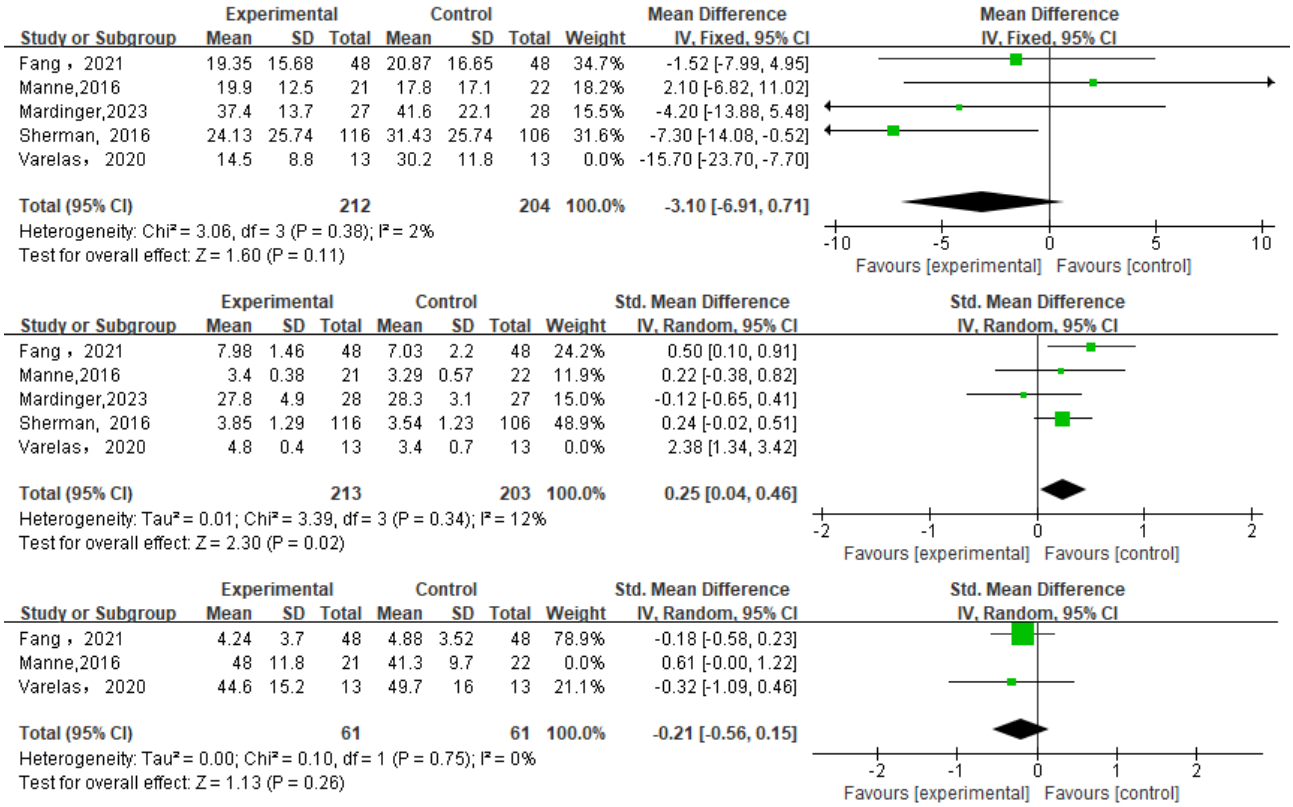

Risk of Bias

Figure 5 [28-31,37-39] presents the summary of the risk of deviation for the included studies. In 6 [28-31,37,39] of the 7 studies, the description of the method used in random assignment was clearly stated (ie, web-based automated randomization software and random number generator), and in the remaining study [38], the information obtained about random assignment was insufficient to make a definitive judgment. Of the 7 studies, 5 [30,31,37-39] were unable to make definitive judgments in this area because of underreporting, whereas in the remaining 2 trials [28,29] sufficient information was obtained about allocation concealment (individually sealed envelopes to conceal allocation). Furthermore, 6 studies [28-30,37-39] were judged to be at unclear risk of bias because the effect of unblinding was unknown, and 1 study [31] described the blinding of participants. Seven studies [28-31,37-39] achieved blinding of outcome evaluators (ie, clinic and surgical staff were blinded to condition assignment) or the blinding was unclear, but the outcome was objectively measured and not subjective to interpretation. Incomplete outcome data appeared to be adequately addressed in 7 studies [28-31,37-39] (ie, incomplete data were fairly evenly balanced across intervention groups or intention-to-treat analyses were reported). In addition, 3 studies [28,30,39] underwent clinical registration or reported relevant protocols, showing that outcomes were reported in full. The impact of selective reporting in the remaining 4 studies [29,31,37,38] was unclear, and this area was judged to be at unclear risk of bias. Information on other potential sources of bias was sufficient. Therefore, this area was judged to be at low risk of bias for all studies [28-31,37-39].

XSL•FO
RenderX

Figure 5. Risk of bias assessment of the included studies [28-31,37-39].

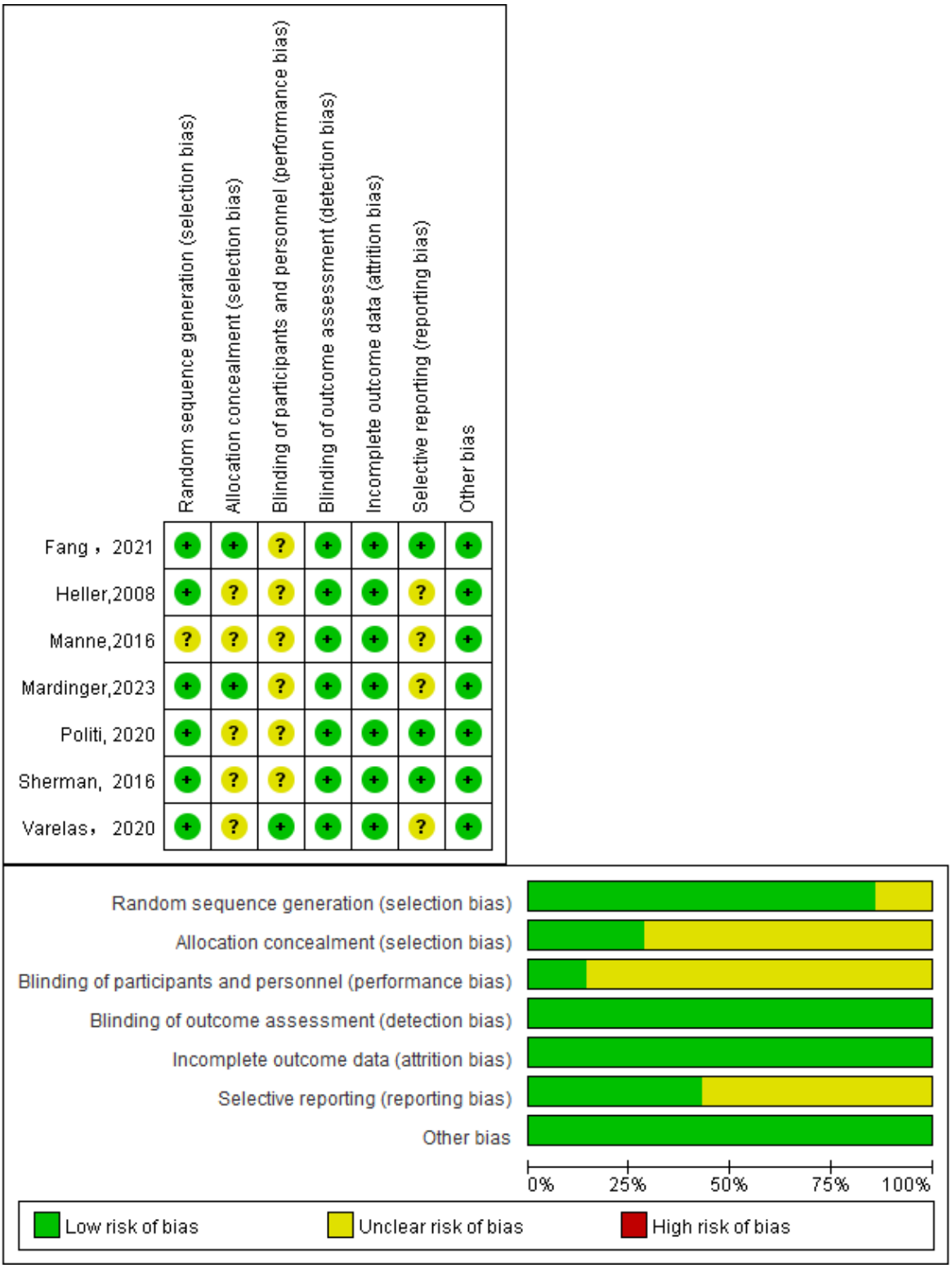

Certainty of Evidence

We assessed the certainty of evidence for the included RCTs using Grades of Recommendation Assessment, Development, and Evaluation (GRADE; Multimedia Appendix 4) except for decision regret, for which the certainty of evidence was low. The certainty of evidence was very low for the rest, that is, decision conflict, satisfaction, anxiety, and informed choice.

Discussion

Principal Findings

Our systematic review and meta-analysis showed that the modules of web-based decision aids include basic information on PMBR, patient stories, risk assessment, value clarification, and emotion management and that patients can be directed to seek information and obtain personalized decision support based

XSL•FO
RenderX

on their individual needs. Therefore, these web-based decision aids are helpful and recommended for women. Regarding the effectiveness of web-based decision aids, the results showed that they may improve PMBR knowledge, decision conflict, and satisfaction but have no effect on informed choice, decision regret, or anxiety. The overall GRADE quality of evidence for decision regret was low, and the overall GRADE quality of evidence for informed choice, decision conflict, and anxiety was very low.

The Content of Web-Based Decision Aids

First, regarding the content of web-based decision aids, few of the studies included in our systematic review and meta-analyses reported comprehensive development of their web-based decision aids. The types of decisions on which most web-based decision aids primarily focused were PMBR decision types and reconstruction times. In addition, some of the studies reported that the development of the tool was obtained through a decisional needs assessment. Research suggests that people tend to have decisional needs when confronted with known outcomes with multiple choices, uncertain outcomes, or valuing people differently and that unmet needs lead to poor quality decisions, which adversely affect health outcomes [40]. Research has shown that some patients have difficulty imagining plastic surgery without photos of women of different body types and skin colors when faced with a decision. Therefore, the use of 3D images during the counseling process is an acceptable web-based decision aid, and the results of our review suggest that web-based decision aids on PMBR decision-making show real photographs of patients by incorporating high-quality, 3D animated images and that viewing 3D images may increase presurgical preparation by giving patients a more realistic understanding of what is actually achievable after PMBR [41]. There are web-based decision aids that are designed with the goal of making patients more comfortable receiving information in a less-stressful environment outside of the hospital, and it also allows family members and friends who are members of the patient support group, but who may not necessarily be able to participate in the counseling, to receive specific information about the procedure and participate in the decision-making process. Women and their families are allowed to express their views about breast surgery because family members act as advocates and care coordinators in the decision-making process [42]. In this era of increasing emphasis on evidence-based medicine, the PMBR risk assessment calculator can help individualize and quantify risk to better inform surgical decisions and better manage patient expectations [43]. The purpose of the values clarification exercise is to help women assess, explore, and identify their personal values and to encourage them to think about how their values affect their decision-making. Using the values clarification exercise can help women increase their satisfaction with their appearance. Patient stories are also important to web-based decision aids, and research has shown that women express a need to learn about other women's experiences to gain a deeper understanding of the impact of PMBR on their daily lives. Web-based decision aids have achieved this by telling the stories of patients who have had previous mastectomies, with or without PMBR. These stories illustrate the decision-making experiences of these patients and the impact of their decisions on their daily lives [44]. Another advantage of web-based decision aids is that they allow patients to absorb the information without being overwhelmed by other information or distracted by other issues. Research has shown that some people feel prepared and emotionally supported for PMBR decision-making, while others feel that the elements of supportive care are missing, making the inclusion of an emotion management module in web-based decision aids essential for women's psychosocial support [45]. However, although the internet has become an easily accessible tool, there is still a persistent digital divide. Therefore, special attention should be given to the sociodemographic characteristics of the population, building more resources for health care infrastructure in underserved communities and providing free or discounted Wi-Fi connections and mobile devices in low-income areas [46]. These actions, combined with the popularity of smartphone users, are measures that may narrow the digital divide [21].

Effectiveness of Web-Based Decision Aids

In line with the results of a previous meta-analysis [26], web-based decision aids reduced decision conflict. Decision conflicts were as high as 45.68 (SD 23.40) among women who were newly diagnosed with early-stage BC in China [47]. Decision conflict was significantly higher among women who chose mastectomy with or without combined reconstruction compared to women who chose conservative breast surgery. Greater decision conflict is associated with less information, higher uncertainty in weighing choices based on personal values, and inadequate social support [40]. Women may second-guess their decisions after the fact, even if those decisions have already been made. Women who face PMBR decision-making need support in making this complex decision, especially those who do not have a strong preference for PMBR. Decision conflict can be reduced by addressing factors of uncertainty, such as providing information about the benefits and risks of each option and helping patients understand their own values [48]. Web-based decision aids can improve the quality of PMBR decision-making by enhancing patient knowledge and providing personalized risk assessments, reducing decision conflict [18].

Uncertainty about whether they are making the best decision can trigger emotional turmoil, and decision regret occurs when women compare the unfavorable outcome of a decision with alternative choices they may have [11,47]. The results of our meta-analysis showed that there was no effect of web-based decision aids on decision regret in the intervention group compared to the control. Women who choose decisions that result in unexpected clinical outcomes or lower-than-expected outcomes will inevitably experience decision regret, a very common but negative emotion, even though the patient's preferences and needs are honored and considered in their treatment [49]. Decision regret can be used as an indicator of decision-making quality, which can contribute to performance improvement in the health care system. Other studies from a psychological perspective have shown that if a decision is regretted, the following "preference reversal" may cause patients to favor another unselected option, which may completely offset their health outcomes, with the degree of decision regret varying widely. However, Becerra Pérez et al [50] reported that most

XSL•FO
RenderX

studies reported a low mean DRS, resulting in an overall mean score of 16.5 out of 100 across studies. It is important to note that there is no consensus on specific thresholds for clinically important decision regret based on DRS, and authors have rarely justified their choice of thresholds; therefore, minimum and maximum efficiency may limit our ability to perform statistical analyses [51].

Previous research has shown that women with BC who use decision aids receive more information that helps them make informed and values-based decisions [26]. Our results, in contrast, showed no effect of web-based decision aids on informed choice in the intervention group compared to the control group possibly because, compared to other forms, web-based decision aids require more effort. Therefore, some women in the web-based decision aids group may have been less inclined to seek more information and consider it carefully. This may explain why women in the web-based decision aids group did not feel less uninformed about their decisions [52]. The results of previous meta-analyses [25,26] suggest that web-based decision aids are promising interventions for improving knowledge related to PMBR decision-making and that web-based decision aids can help patients' knowledge of PMBR and treatment options and can identify patients' PMBR preferences and goals for quality decision-making with their health care providers; however, it is important to note that in this review, the impact of web-based decision aids on PMBR knowledge was mixed, which may be because most of the current instruments on PMBR decision-making knowledge measurement are self-administered scales. We found that web-based decision aids improved PMBR knowledge compared to a control group of some conventional education [37], traditional counseling [31], or conventional pamphlets [30]. When the control group was using pamphlets [19,28] or noninteractive decision aids [29] that contained similar information, web-based decision aids did not have a statistically significant effect on PMBR knowledge. Therefore, to elucidate the impact of web-based decision aids on knowledge, measurement studies using validated and sensitive instruments are needed.

Because the initial anxiety experienced by women may be related to the new diagnosis and anticipated surgery, this anxiety lessened once the surgery was over. There was no difference in the level of anxiety experienced after surgery between the 2 groups. Given the severity of a BC diagnosis, it is very reassuring that web-based decision aids did not exacerbate anxiety while providing benefits in terms of patient satisfaction and knowledge as well as surgeon satisfaction. Several studies have shown that patient satisfaction is higher when receiving PMBR information digitally [53]. Our study also suggests that web-based decision aids improve patient satisfaction with decision-making. Although most of the studies included in our systematic review reported that the use of web-based decision aids increased women's satisfaction with PMBR, most of the measurement tools used to assess the outcomes used self-administered scales. Therefore, more high-quality evidence, including studies using validated and sensitive instruments, is needed to elucidate the impact of web-based decision aids on satisfaction [26].

Sensitivity Analysis

Some of the outcome indicators in this review (ie, decision conflict, satisfaction, and anxiety) showed significant heterogeneity, which may be related to factors such as the fact that the measurement tools were different and the web-based decision aids were delivered in an inconsistent form and content. We conducted sensitivity analyses for decision conflict, satisfaction, and anxiety, and the adjusted total estimates of anxiety did not significantly change these results when studies were progressively omitted, excluding the study by Manne et al [38]. With respect to decision conflict and satisfaction, the adjusted total estimates changed significantly, a result that excludes the study by Varelas et al [31]. Contrary to the original results, the effect of web-based decision aids on improving satisfaction was statistically significant, and the effect of web-based decision aids on improving decision conflict was similar to the control group effect; therefore, the effect of web-based decision aids on decision conflict and satisfaction should be carefully interpreted. Regarding the heterogeneity of this meta-analysis, sensitivity analyses showed that the heterogeneity of all outcomes was also reduced by excluding 1 study.

Limitations

Some limitations of this review must be recognized. First, we did not perform an assessment of publication bias because only 7 studies were ultimately included in the analysis, which may cause publication bias. In addition, the included studies had no follow-up surveys and lacked evidence of the long-term impact of the interventions. Our findings serve as a reminder that even when statistical information is effectively communicated, participants may not make estimates of the same order of magnitude after a period. Finally, the number of included studies was small. Some studies had inconsistent outcome indicators and were therefore not included.

Conclusions

This review shows that web-based decision aids can increase knowledge and satisfaction, and reduce levels of decision conflict among women facing PMBR decision-making; however, there is no effect on informed choice, decision regret, or anxiety. Currently, web-based decision aids for women's PMBR decision-making are relatively easy to implement in terms of content and form. Due to limitations in the number of included studies in our meta-analysis, well-designed studies, including multicenter RCTs using high-quality decision aids, are necessary in the future to further validate our conclusion that web-based decision aids play a role in the quality of decision-making for women facing PMBR.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the authors of studies in this review who took the time to become involved in guiding the development of this review.

Authors' Contributions

LY, LL, SY, JG, XS, and MZ made substantial contributions to the conception and design, acquisition of data, or analysis and interpretation of data and were involved in drafting the manuscript or revising it critically for important intellectual content. All authors provided final approval for the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement.
[DOCX File , 27 KB - jmir_v26i1e53872_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 18 KB - jmir_v26i1e53872_app2.docx]

Multimedia Appendix 3
Characteristics of the interventions and controls.
[DOCX File , 16 KB - jmir_v26i1e53872_app3.docx]

Multimedia Appendix 4
Certainty of evidence.
[DOCX File , 19 KB - jmir_v26i1e53872_app4.docx]

References

1. Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast 2022 Dec;66:15-23 [FREE Full text] [doi: 10.1016/j.breast.2022.08.010] [Medline: 36084384]
2. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast cancer statistics, 2022. CA Cancer J Clin 2022 Nov 03;72(6):524-541 [FREE Full text] [doi: 10.3322/caac.21754] [Medline: 36190501]
3. Pfeiffer RM, Webb-Vargas Y, Wheeler W, Gail MH. Proportion of U.S. trends in breast cancer incidence attributable to long-term changes in risk factor distributions. Cancer Epidemiol Biomarkers Prev 2018 Oct;27(10):1214-1222 [FREE Full text] [doi: 10.1158/1055-9965.EPI-18-0098] [Medline: 30068516]
4. Michaels E, Worthington RO, Rusiecki J. Breast cancer: risk assessment, screening, and primary prevention. Med Clin North Am 2023 Mar;107(2):271-284. [doi: 10.1016/j.mcna.2022.10.007] [Medline: 36759097]
5. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan 12;71(1):7-33 [FREE Full text] [doi: 10.3322/caac.21654] [Medline: 33433946]
6. Sun L, Ang E, Ang WH, Lopez V. Losing the breast: a meta-synthesis of the impact in women breast cancer survivors. Psychooncology 2018 Feb 16;27(2):376-385. [doi: 10.1002/pon.4460] [Medline: 28544436]
7. Kummerow KL, Du L, Penson DF, Shyr Y, Hooks MA. Nationwide trends in mastectomy for early-stage breast cancer. JAMA Surg 2015 Jan 01;150(1):9-16. [doi: 10.1001/jamasurg.2014.2895] [Medline: 25408966]
8. Li X, Meng M, Zhao J, Zhang X, Yang D, Fang J, et al. Shared decision-making in breast reconstruction for breast cancer patients: a scoping review. Patient Prefer Adherence 2021 Dec;15:2763-2781 [FREE Full text] [doi: 10.2147/PPA.S335080] [Medline: 34916786]
9. Santosa KB, Qi J, Kim HM, Hamill JB, Wilkins EG, Pusic AL. Long-term patient-reported outcomes in postmastectomy breast reconstruction. JAMA Surg 2018 Oct 01;153(10):891-899 [FREE Full text] [doi: 10.1001/jamasurg.2018.1677] [Medline: 29926096]
10. Bargon CA, Young-Afat DA, Ikinci M, Braakenburg A, Rakhorst HA, Mureau MA, et al. Breast cancer recurrence after immediate and delayed postmastectomy breast reconstruction-A systematic review and meta-analysis. Cancer 2022 Oct 01;128(19):3449-3469 [FREE Full text] [doi: 10.1002/cncr.34393] [Medline: 35894936]

XSL•FO
RenderX

11. Lee CN, Deal AM, Huh R, Ubel PA, Liu YJ, Blizard L, et al. Quality of patient decisions about breast reconstruction after mastectomy. JAMA Surg 2017 Aug 01;152(8):741-748 [FREE Full text] [doi: 10.1001/jamasurg.2017.0977] [Medline: 28467530]
12. Hoefel L, O'Connor AM, Lewis KB, Boland L, Sikora L, Hu J, et al. 20th anniversary update of the Ottawa decision support framework part 1: a systematic review of the decisional needs of people making health or social decisions. Med Decis Making 2020 Jul 13;40(5):555-581. [doi: 10.1177/0272989x20936209]
13. Ter Stege JA, Oldenburg HS, Woerdeman LA, Witkamp AJ, Kieffer JM, van Huizum MA, et al. Decisional conflict in breast cancer patients considering immediate breast reconstruction. Breast 2021 Feb;55:91-97 [FREE Full text] [doi: 10.1016/j.breast.2020.12.001] [Medline: 33387811]
14. Brehaut JC, O'Connor AM, Wood TJ, Hack TF, Siminoff L, Gordon E, et al. Validation of a decision regret scale. Med Decis Making 2016 Jul 02;23(4):281-292. [doi: 10.1177/0272989x03256005]
15. Fortunato L, Loreti A, Cortese G, Spallone D, Toto V, Cavaliere F, et al. Regret and quality of life after mastectomy with or without reconstruction. Clin Breast Cancer 2021 Jun;21(3):162-169. [doi: 10.1016/j.clbc.2019.11.005] [Medline: 33744100]
16. Cai L, Momeni A. The impact of reconstructive modality and postoperative complications on decision regret and patient-reported outcomes following breast reconstruction. Aesthetic Plast Surg 2022 Apr 29;46(2):655-660. [doi: 10.1007/s00266-021-02660-2] [Medline: 34845512]
17. Pinker K, Chin J, Melsaether AN, Morris EA, Moy L. Precision medicine and radiogenomics in breast cancer: new approaches toward diagnosis and treatment. Radiology 2018 Jun;287(3):732-747. [doi: 10.1148/radiol.2018172171] [Medline: 29782246]
18. Sowa Y, Inafuku N, Tsuge I, Yamanaka H, Katsube M, Sakamoto M, et al. Development and implementation of a decision aid for post-mastectomy breast reconstruction for Japanese women with breast cancer: a field-testing study. Breast Cancer 2023 Jul 18;30(4):570-576. [doi: 10.1007/s12282-023-01447-4] [Medline: 36934211]
19. Manne SL, Topham N, Kirstein L, Virtue SM, Brill K, Devine KA, et al. Attitudes and decisional conflict regarding breast reconstruction among breast cancer patients. Cancer Nurs 2016;39(6):427-436 [FREE Full text] [doi: 10.1097/NCC.0000000000000320] [Medline: 26780376]
20. Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2011 Oct 05;4(10):CD001431. [doi: 10.1002/14651858.CD001431.pub3] [Medline: 21975733]
21. Yu L, Li P, Yang S, Guo P, Zhang X, Liu N, et al. Web-based decision aids to support breast cancer screening decisions: systematic review and meta-analysis. J Comp Eff Res 2020 Oct;9(14):985-1002 [FREE Full text] [doi: 10.2217/cer-2020-0052] [Medline: 33025800]
22. Baptista S, Teles Sampaio E, Heleno B, Azevedo LF, Martins C. Web-based versus usual care and other formats of decision aids to support prostate cancer screening decisions: systematic review and meta-analysis. J Med Internet Res 2018 Jun 26;20(6):e228 [FREE Full text] [doi: 10.2196/jmir.9070] [Medline: 29945857]
23. Tong G, Geng Q, Wang D, Liu T. Web-based decision aids for cancer clinical decisions: a systematic review and meta-analysis. Support Care Cancer 2021 Nov 08;29(11):6929-6941. [doi: 10.1007/s00520-021-06184-y] [Medline: 33834302]
24. Paraskeva N, Guest E, Lewis-Smith H, Harcourt D. Assessing the effectiveness of interventions to support patient decision making about breast reconstruction: a systematic review. Breast 2018 Aug;40:97-105 [FREE Full text] [doi: 10.1016/j.breast.2018.04.020] [Medline: 29730304]
25. Berlin NL, Tandon VJ, Hawley ST, Hamill JB, MacEachern MP, Lee CN, et al. Feasibility and efficacy of decision aids to improve decision making for postmastectomy breast reconstruction: a systematic review and meta-analysis. Med Decis Making 2018 Dec 27;39(1):5-20. [doi: 10.1177/0272989x18803879]
26. Yang S, Yu L, Zhang C, Xu M, Tian Q, Cui X, et al. Effects of decision aids on breast reconstruction: a systematic review and meta-analysis of randomised controlled trials. J Clin Nurs 2023 Apr 22;32(7-8):1025-1044. [doi: 10.1111/jocn.16328] [Medline: 35460127]
27. Zhao A, Larbi M, Miller K, O'Neill S, Jayasekera J. A scoping review of interactive and personalized web-based clinical tools to support treatment decision making in breast cancer. Breast 2022 Feb;61:43-57 [FREE Full text] [doi: 10.1016/j.breast.2021.12.003] [Medline: 34896693]
28. Fang SY, Lin PJ, Kuo YL. Long-term effectiveness of a decision support app (Pink Journey) for women considering breast reconstruction surgery: pilot randomized controlled trial. JMIR Mhealth Uhealth 2021 Dec 10;9(12):e31092 [FREE Full text] [doi: 10.2196/31092] [Medline: 34890354]
29. Mardinger C, Steve AK, Webb C, Sherman KA, Temple-Oberle C. Breast reconstruction decision aids decrease decisional conflict and improve decisional satisfaction: a randomized controlled trial. Plast Reconstr Surg 2023 Feb 01;151(2):278-288. [doi: 10.1097/PRS.0000000000009830] [Medline: 36696307]
30. Politi MC, Lee CN, Philpott-Streiff SE, Foraker RE, Olsen MA, Merrill C, et al. A randomized controlled trial evaluating the BREASTChoice tool for personalized decision support about breast reconstruction after mastectomy. Ann Surg 2020 Feb;271(2):230-237. [doi: 10.1097/SLA.0000000000003444] [Medline: 31305282]

XSL•FO
RenderX

31. Varelas L, Egro FM, Evankovich N, Nguyen V. A randomized controlled trial to assess the use of a virtual decisional aid to improve knowledge and patient satisfaction in women considering breast reconstruction following mastectomy. Cureus 2020 Dec 10;12(12):e12018 [FREE Full text] [doi: 10.7759/cureus.12018] [Medline: 33457123]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol 2021 Jun;134:178-189 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.03.001] [Medline: 33789819]
33. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 03;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
34. Lee AY, Wong AK, Hung TT, Yan J, Yang S. Nurse-led telehealth intervention for rehabilitation (telerehabilitation) among community-dwelling patients with chronic diseases: systematic review and meta-analysis. J Med Internet Res 2022 Nov 02;24(11):e40364 [FREE Full text] [doi: 10.2196/40364] [Medline: 36322107]
35. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2008.
36. Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. Biostatistics 2000 Sep;1(3):247-262. [doi: 10.1093/biostatistics/1.3.247] [Medline: 12933507]
37. Heller L, Parker PA, Youssef A, Miller MJ. Interactive digital education aid in breast reconstruction. Plast Reconstr Surg 2008;122(3):717-724. [doi: 10.1097/prs.0b013e318180ed06]
38. Manne SL, Topham N, D'Agostino TA, Myers Virtue S, Kirstein L, Brill K, et al. Acceptability and pilot efficacy trial of a web-based breast reconstruction decision support aid for women considering mastectomy. Psychooncology 2016 Dec 18;25(12):1424-1433 [FREE Full text] [doi: 10.1002/pon.3984] [Medline: 26383833]
39. Sherman KA, Shaw LK, Winch CJ, Harcourt D, Boyages J, Cameron LD, et al. Reducing decisional conflict and enhancing satisfaction with information among women considering breast reconstruction following mastectomy: results from the BRECONDA randomized controlled trial. Plast Reconstr Surg 2016 Oct;138(4):592e-602e. [doi: 10.1097/PRS.0000000000002538] [Medline: 27673530]
40. O'Connor AM, Drake ER, Wells GA, Tugwell P, Laupacis A, Elmslie T. A survey of the decision-making needs of Canadians faced with complex health decisions. Health Expect 2003 Jun 14;6(2):97-109 [FREE Full text] [doi: 10.1046/j.1369-6513.2003.00215.x] [Medline: 12752738]
41. McCrorie AD, Begley AM, Chen JJ, McCorry NK, Paget G, McIntosh SA. Improving preparedness prior to reconstructive breast surgery via inclusion of 3D images during pre-operative counselling: a qualitative analysis. BMC Womens Health 2021 Aug 31;21(1):323 [FREE Full text] [doi: 10.1186/s12905-021-01463-6] [Medline: 34465323]
42. Flitcroft K, Brennan M, Spillane A. Making decisions about breast reconstruction: a systematic review of patient-reported factors influencing choice. Qual Life Res 2017 Sep 10;26(9):2287-2319. [doi: 10.1007/s11136-017-1555-z] [Medline: 28397191]
43. Kim JY, Mlodinow AS, Khavanin N, Hume KM, Simmons CJ, Weiss MJ, et al. Individualized risk of surgical complications: an application of the breast reconstruction risk assessment score. Plast Reconstr Surg Glob Open 2015;3(5):e405. [doi: 10.1097/gox.0000000000000351]
44. Hung YT, Wu CF, Liang TH, Chou SS, Chen GL, Wu PN, et al. Developing a decision-aid website for breast cancer surgery: an action research approach. J Med Internet Res 2019 Feb 04;21(2):e10404 [FREE Full text] [doi: 10.2196/10404] [Medline: 30714941]
45. Hansen ST, Willemoes Rasmussen LA. 'At least there is something in my bra': a qualitative study of women's experiences with oncoplastic breast surgery. J Adv Nurs 2022 Oct 07;78(10):3304-3319 [FREE Full text] [doi: 10.1111/jan.15309] [Medline: 35799478]
46. Sohn M, Yang J, Sohn J, Lee JH. Digital healthcare for dementia and cognitive impairment: a scoping review. Int J Nurs Stud 2023 Apr;140:104413. [doi: 10.1016/j.ijnurstu.2022.104413] [Medline: 36821951]
47. Zhuang HZ, Wang L, Yu XF, Chan SW, Gao YX, Li XQ, et al. Effects of decisional conflict, decision regret and self-stigma on quality of life for breast cancer survivors: a cross-sectional, multisite study in China. J Adv Nurs 2022 Oct;78(10):3261-3272. [doi: 10.1111/jan.15250] [Medline: 35396872]
48. Subramanian L, Zhao J, Zee J, Knaus M, Fagerlin A, Perry E, et al. Use of a decision aid for patients considering peritoneal dialysis and in-center hemodialysis: a randomized controlled trial. Am J Kidney Dis 2019 Sep;74(3):351-360 [FREE Full text] [doi: 10.1053/j.ajkd.2019.01.030] [Medline: 30954312]
49. Xu RH, Zhou LM, Wong EL, Wang D, Chang JH. Psychometric evaluation of the Chinese version of the decision regret scale. Front Psychol 2020 Dec 3;11:583574 [FREE Full text] [doi: 10.3389/fpsyg.2020.583574] [Medline: 33424697]
50. Becerra Pérez MM, Menear M, Brehaut JC, Légaré F. Extent and predictors of decision regret about health care decisions. Med Decis Making 2016 Mar 14;36(6):777-790. [doi: 10.1177/0272989x16636113]
51. Bruce L, Khouri AN, Bolze A, Ibarra M, Richards B, Khalatbari S, et al. Long-term regret and satisfaction with decision following gender-affirming mastectomy. JAMA Surg 2023 Oct 01;158(10):1070-1077. [doi: 10.1001/jamasurg.2023.3352] [Medline: 37556147]

XSL•FO
RenderX

52. Osaka W, Nakayama K. Effect of a decision aid with patient narratives in reducing decisional conflict in choice for surgery among early-stage breast cancer patients: a three-arm randomized controlled trial. Patient Educ Couns 2017 Mar;100(3):550-562. [doi: 10.1016/j.pec.2016.09.011] [Medline: 28277290]
53. Hoffman AS, Cantor SB, Fingeret MC, Housten AJ, Hanson SE, McGee JH, et al. Considering breast reconstruction after mastectomy: a patient decision aid video and workbook. Plast Reconstr Surg Glob Open 2019 Nov;7(11):e2500 [FREE Full text] [doi: 10.1097/GOX.0000000000002500] [Medline: 31942296]

Abbreviations

BC: breast cancer
DCS: Decision Conflict Scale
DRS: Decision Regret Scale
GRADE: Grades of Recommendation Assessment, Development, and Evaluation
MD: mean difference
PMBR: postmastectomy breast reconstruction
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by S Ma, S Gardezi; submitted 22.10.23; peer-reviewed by A Trojan, T Lund-Jacobsen; comments to author 13.03.24; revised version received 16.04.24; accepted 16.04.24; published 27.05.24.

Please cite as:
Yu L, Gong J, Sun X, Zang M, Liu L, Yu S
Assessing the Content and Effect of Web-Based Decision Aids for Postmastectomy Breast Reconstruction: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2024;26:e53872
URL: https://www.jmir.org/2024/1/e53872
doi:10.2196/53872
PMID:38801766

©Lin Yu, Jianmei Gong, Xiaoting Sun, Min Zang, Lei Liu, Shengmiao Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of the Working Alliance in Blended Cognitive Behavioral Therapy and Treatment as Usual for Depression in Europe: Secondary Data Analysis of the E-COMPARED Randomized Controlled Trial

Asmae Doukani[1], MSc; Matteo Quartagno[2], PhD; Francesco Sera[3], PhD; Caroline Free[1], PhD; Ritsuko Kakuma[1], PhD; Heleen Riper[4], PhD; Annet Kleiboer[5,6], PhD; Arlinda Cerga-Pashoja[1], PhD; Anneke van Schaik[4,7], PhD; Cristina Botella[8,9], PhD; Thomas Berger[10], PhD; Karine Chevreul[11,12], PhD; Maria Matynia[13], PhD; Tobias Krieger[10], DPhil; Jean-Baptiste Hazo[11,12], PhD; Stasja Draisma[14], PhD; Ingrid Titzler[15], PhD; Naira Topooco[16], PhD; Kim Mathiasen[17,18], PhD; Kristofer Vernmark[16], PhD; Antoine Urech[10,19], PhD; Anna Maj[13], PhD; Gerhard Andersson[20,21,22], PhD; Matthias Berking[23], MD, PhD; Rosa María Baños[9,24], PhD; Ricardo Araya[25], PhD

[1]Department of Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom
[2]Medical Research Council Clinical Trials Unit, University College London, London, United Kingdom
[3]Department of Statistics, Computer Science and Applications “G. Parenti”, University of Florence, Florance, Italy
[4]Department of Psychiatry, Amsterdam University Medial Centre, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[5]Department Clinical, Neuro, and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[6]Amsterdam Public Health Institute, Amsterdam, Netherlands
[7]Academic Department for Depressive Disorders, Dutch Mental Health Care, Amsterdam, Netherlands
[8]Department of Basic Psychology, Clinical and Psychobiology, Universitat Jaume I, Castellón de la Plana, Spain
[9]Centro de Investigación Biomédica en Red Fisiopatología Obesidad y Nutrición, Instituto Carlos III, Madrid, Spain
[10]Department of Clinical Psychology and Psychotherapy, University of Bern, Bern, Switzerland
[11]Unité de Recherche Clinique in Health Economics, Assistance Publique–Hôpitaux de Paris, Paris, France
[12]Health Economics Research Unit, Inserm, University of Paris, Paris, France
[13]Faculty of Psychology, SWPS University, Warsaw, Poland
[14]Department on Aging, Netherlands Institute of Mental Health and Addiction (Trimbos Institute), Utrecht, Netherlands
[15]Department of Clinical Psychology and Psychotherapy, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[16]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden
[17]Department of Clinical Medicine, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark
[18]Centre for Digital Psychiatry, Mental Health Services of Southern Denmark, Odense, Denmark
[19]Department of Neurology, Inselspital Bern, Bern University Hospital, Bern, Switzerland
[20]Department of Behavioral Sciences and Learning, Linköping University, Linköping, Sweden
[21]Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden
[22]Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden
[23]Department of Clinical Psychology and Psychotherapy, Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany
[24]Department of Personality, Evaluation and Psychological Treatments, Universidad de Valencia, Valencia, Spain
[25]Department of Health Service and Population Research, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, London, United Kingdom

Corresponding Author:
Asmae Doukani, MSc
Department of Population Health
London School of Hygiene & Tropical Medicine
Keppel Street
London, WC1E 7HT
United Kingdom
Phone: 44 020 7636 8636 ext 2463
Email: asmae.doukani@lshtm.ac.uk

XSL•FO
RenderX

Abstract

Background: Increasing interest has centered on the psychotherapeutic working alliance as a means of understanding clinical change in digital mental health interventions in recent years. However, little is understood about how and to what extent a digital mental health program can have an impact on the working alliance and clinical outcomes in a blended (therapist plus digital program) cognitive behavioral therapy (bCBT) intervention for depression.

Objective: This study aimed to test the difference in working alliance scores between bCBT and treatment as usual (TAU), examine the association between working alliance and depression severity scores in both arms, and test for an interaction between system usability and working alliance with regard to the association between working alliance and depression scores in bCBT at 3-month assessments.

Methods: We conducted a secondary data analysis of the E-COMPARED (European Comparative Effectiveness Research on Blended Depression Treatment versus Treatment-as-usual) trial, which compared bCBT with TAU across 9 European countries. Data were collected in primary care and specialized services between April 2015 and December 2017. Eligible participants aged 18 years or older and diagnosed with major depressive disorder were randomized to either bCBT (n=476) or TAU (n=467). bCBT consisted of 6-20 sessions of bCBT (involving face-to-face sessions with a therapist and an internet-based program). TAU consisted of usual care for depression. The main outcomes were scores of the working alliance (Working Alliance Inventory-Short Revised–Client [WAI-SR-C]) and depressive symptoms (Patient Health Questionnaire-9 [PHQ-9]) at 3 months after randomization. Other variables included system usability scores (System Usability Scale-Client [SUS-C]) at 3 months and baseline demographic information. Data from baseline and 3-month assessments were analyzed using linear regression models that adjusted for a set of baseline variables.

Results: Of the 945 included participants, 644 (68.2%) were female, and the mean age was 38.96 years (IQR 38). bCBT was associated with higher composite WAI-SR-C scores compared to TAU (B=5.67, 95% CI 4.48-6.86). There was an inverse association between WAI-SR-C and PHQ-9 in bCBT (B=−0.12, 95% CI −0.17 to −0.06) and TAU (B=−0.06, 95% CI −0.11 to −0.02), in which as WAI-SR-C scores increased, PHQ-9 scores decreased. Finally, there was a significant interaction between SUS-C and WAI-SR-C with regard to an inverse association between higher WAI-SR-C scores and lower PHQ-9 scores in bCBT (b=−0.030, 95% CI −0.05 to −0.01; P=.005).

Conclusions: To our knowledge, this is the first study to show that bCBT may enhance the client working alliance when compared to evidence-based routine care for depression that services reported offering. The working alliance in bCBT was also associated with clinical improvements that appear to be enhanced by good program usability. Our findings add further weight to the view that the addition of internet-delivered CBT to face-to-face CBT may positively augment experiences of the working alliance.

Trial Registration: ClinicalTrials.gov NCT02542891, https://clinicaltrials.gov/study/NCT02542891; German Clinical Trials Register DRKS00006866, https://drks.de/search/en/trial/DRKS00006866; Netherlands Trials Register NTR4962, https://www.onderzoekmetmensen.nl/en/trial/25452; ClinicalTrials.Gov NCT02389660, https://clinicaltrials.gov/study/NCT02389660; ClinicalTrials.gov NCT02361684, https://clinicaltrials.gov/study/NCT02361684; ClinicalTrials.gov NCT02449447, https://clinicaltrials.gov/study/NCT02449447; ClinicalTrials.gov NCT02410616, https://clinicaltrials.gov/study/NCT02410616; ISRCTN Registry ISRCTN12388725, https://www.isrctn.com/ISRCTN12388725?q=ISRCTN12388725&filters=&sort=&offset=1&totalResults=1&page=1&pageSize=10; ClinicalTrials.gov NCT02796573, https://classic.clinicaltrials.gov/ct2/show/NCT02796573

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-016-1511-1

(J Med Internet Res 2024;26:e47515) doi:10.2196/47515

KEYWORDS

blended psychotherapy; cognitive behavioral therapy; depression; digital mental health interventions; psychotherapy; mental health; program usability; therapeutic alliance; usability heuristics; working alliance

Introduction

Background

Depression is one of the most significant contributors to the global disease burden, affecting an estimated 264 million people globally [1,2]. Depression accounts for 7.2% of the overall disease burden in Europe, costing an estimated €113,405 billion (US $123,038 billion) per year. However, 45% of people with major depression will go untreated [3]. High costs and suboptimal access to mental health care are among the many reasons to foster digital mental health interventions (DMHIs), which promise greater quality of care and lower costs of delivery [4,5].

Evidence concerning the effectiveness of DMHIs has increased substantially over the past decade. Growing evidence indicates that internet-delivered cognitive behavioral therapy (iCBT) might be just as effective as face-to-face cognitive behavioral therapy (CBT) for a range of mental health conditions, particularly depression [6-13]. iCBT is delivered with varying degrees of support ranging from a stand-alone self-administered

XSL•FO
RenderX

digital program to a blended treatment with the active involvement of a therapist through regular face-to-face meetings. Blended psychotherapies provide higher levels of therapist support compared to guided approaches that provide minimal or some guidance from a mental health practitioner [4]. Blended delivery has gained interest, with emerging evidence suggesting that such interventions can lead to improved adherence and treatment outcomes [14].

As interest in DMHIs increases, considerable attention has centered around the concept of the client-therapist alliance, of which there are many variations (therapeutic, working, helping, etc). While different therapeutic approaches have historically failed to agree on a definition of the alliance, Edward Bordin [15-17] proposed a pan-theoretical tripartite conceptualization called the working alliance that is characterized by 3 key dimensions, including the emotional "bond" between the client and the therapist, the agreement on the therapeutic "goals," and the "task" needed to advance the client's goals toward clinical improvement. This concept is particularly important because it has consistently predicted positive treatment outcomes for a range of psychological approaches, including CBT for depression [18-20].

The client-therapist alliance was identified as a key research priority for research policy and funding in digital technologies in mental health care, in a large consensus study involving people with lived experiences of mental health problems and service use, their carers, and mental health practitioners [21]. The integration of digital technologies in psychotherapy has led to changes in the way the alliance is conceptualized and assessed [19], with variability depending on the type of DMHI (digital program [22], avatar [23], or mobile app [24]).

Prior Work

The literature investigating the client-therapist alliance has largely focused on addressing 2 key questions. The first question is "Do *alliance* scores predict changes in clinical outcomes?" [21,25-29], and the second question, which has been focused on to a lesser extent, is "Does the alliance vary depending on how psychotherapy is delivered?" Systematic reviews that have addressed these questions specifically in relation to interventions that are guided, adopt CBT [21], or target the treatment of depression [27] found that the working alliance can be established in guided DMHIs at a comparable level to face-to-face therapy [21]; however, the literature on the outcome-alliance relationship is mixed [21,26,27].

To this end, only 3 studies have examined the working alliance in blended CBT (bCBT). The first was an uncontrolled study in Sweden, which offered 4 face-to-face and 10 iCBT sessions to a total of 73 participants in primary care services and which was part of the E-COMPARED (European Comparative Effectiveness Research on Blended Depression Treatment versus Treatment-as-usual) study [30]. The findings showed that the alliance was rated highly by both clients and therapists. However, only therapist alliance ratings were associated with client score changes in depression, while client ratings were not.

The second study was conducted in the Netherlands and recruited 102 participants from specialist care services. Participants were either randomized to bCBT (n=47), which consisted of a 20-week intervention (10 face-to-face and 10 online sessions), or a control condition (n=45), which consisted of 15-20 face-to-face CBT sessions [31]. Similar to the findings from the study conducted in Sweden [30], the working alliance was rated highly by both clients and therapists, and no differences were observed between scores. Client ratings of the working alliance were associated with lower depression scores over time in face-to-face CBT but not in bCBT. Therapist working alliance ratings were not significantly associated with depression scores over time in both treatment conditions [31].

The third and most recent study was conducted in Denmark. The study recruited a total of 76 participants who were either randomized to bCBT (n=38), which consisted of 6 face-to-face sessions that were alternated with 6-8 online modules of an internet-based program, or a control condition (n=38), which consisted of 12 face-to-face CBT sessions [32]. The findings showed a significant difference in client and therapist working alliance scores, in which clients rated their working alliance higher than therapists. However, only the therapist ratings across conditions were significantly associated with outcomes in depression. Working alliance ratings across face-to-face CBT and bCBT were comparable. Working alliance ratings in both face-to-face CBT and bCBT did not significantly predict treatment outcomes. It is not clear why an in-group effect was found for therapists across the pooled data and not within treatment conditions [32]. These findings might indicate that the study was not powered enough to detect an association for client ratings in each treatment condition.

While research has mainly focused on measuring the alliance between the client and therapist, emerging qualitative research suggests that DMHIs may offer additional relational alliance benefits [29,33,34]. An example comes from a qualitative study that examined the working alliance demands in a bCBT intervention for people with mild-to-moderate depression in the United Kingdom, as part of the E-COMPARED trial [35]. Qualitative data indicated a potential fourth dimension called "usability heuristics," which appeared to uniquely promote the working alliance in bCBT. Usability heuristics defines the digital program's role in promoting active engagement, self-discovery, and autonomous problem-solving, with higher levels expected to enhance the quality of the working alliance. Features that enable "usability heuristics" include digital technologies that increase access and immediacy to the therapeutic task (availability), appropriately respond to the client's input (interactivity), are easy to use, have esthetic appeal, and promote self-directed therapy [36]. Findings regarding usability heuristics and the respective subfeatures were also reported in another qualitative study that tested this framework in a Spanish sample of participants who experienced self-guided or low-intensity supported iCBT [37]. It is therefore possible that experiences of digital program features may influence the way that the working alliance is experienced in blended formats of CBT [36].

XSL•FO
RenderX

Aims and Objectives

To our knowledge, we report the largest investigation of the working alliance in bCBT for depression, using pooled data from 9 country sites involved in a pragmatic noninferiority randomized controlled trial investigating the effectiveness of bCBT for depression when compared with treatment as usual (TAU) [35]. Further to this, our study will explore if system usability, a newly conceptualized feature of the working alliance, in bCBT interacts with the working alliance and treatment outcome association [36]. Our primary objectives are to test the difference in working alliance scores between bCBT and TAU (objective 1), and determine if working alliance scores are associated with depression scores (objective 2). Our secondary objective is to test for an interaction between system usability and the working alliance with regard to an association between the working alliance and depression scores in bCBT (objective 3).

Methods

Study Design and Settings

We conducted a nonprespecified secondary analysis of data collected in the E-COMPARED study, a large European 2-arm, noninferiority randomized controlled trial investigating the effectiveness of bCBT compared with TAU across 9 European countries (France: ClinicalTrials.gov NCT02542891, September 4, 2015; Germany: German Clinical Trials Register DRKS00006866, December 2, 2014; The Netherlands: Netherlands Trials Register NTR4962, January 5, 2015; Poland: ClinicalTrials.gov NCT02389660, February 18, 2015; Spain: ClinicalTrials.gov NCT02361684, January 8, 2015; Sweden: ClinicalTrials.gov NCT02449447, March 30, 2015; Switzerland: ClinicalTrials.gov NCT02410616, April 2, 2015; United Kingdom: ISRCTN Registry ISRCTN12388725, March 20, 2015; Denmark: ClinicalTrials.gov NCT02796573, June 1, 2016) [35,38]. Data were collected between April 2015 and December 2017. Clients seeking treatment for depression were recruited, assessed, and treated across routine primary care in Germany, Poland, Spain, Sweden, and the United Kingdom, and specialized mental health services in France, the Netherlands, Switzerland, and Denmark [35]. Following the start of recruitment, an additional satellite site was added in Denmark to boost recruitment [38]. The E-COMPARED trial was funded by the European Commission FP7-Health-2013-Innovation-1 program (grant agreement number: 603098).

Participants

Recruitment procedures differed in each country, but all sites screened new clients seeking help for depression, who scored 5 or higher on the Patient Health Questionnaire-9 (PHQ-9) [39]. The study was explained to potential participants either face-to-face or over a telephone call. Clients who agreed to take part in the study were invited to an initial appointment to assess eligibility. The inclusion criteria applied at all sites were as follows: age ≥18 years and meeting the diagnostic criteria for major depressive disorder as confirmed by the MINI International Neuropsychiatric Interview (M.I.N.I) version 5.0 [40]. The exclusion criteria were as follows: high risk of suicide and psychiatric comorbidity (ie, substance dependence, bipolar affective disorder, psychotic illness, or obsessive compulsive disorder) assessed during the M.I.N.I. interview; receiving psychological treatment for depression in primary or specialized mental health care at the point of recruitment; inability to comprehend the spoken and written language of the country site; lacking access to a computer or a fast internet connection (ie, broadband or comparable); and lacking a smartphone or being unwilling to carry a smartphone if one was provided by the research team [35].

After baseline assessments, participants were randomized to 1 of 2 treatment arms (bCBT or TAU) using block randomization, with stratification by country [35]. All participants provided written informed consent before taking part in the trial [35].

Ethical Considerations

The trial was conducted in accordance with the Declaration of Helsinki and was approved by all local ethics committees. Ethics approval to conduct a secondary analysis was obtained from the London School of Hygiene and Tropical Medicine Research Ethics Committee on October 7, 2019 (ethics reference number: 17852). For further information on the trial, including local ethics approvals and the randomization process, see the trial protocol [35].

Interventions: bCBT and TAU

bCBT for depression consisted of integrating a digital program (iCBT plus mobile app) with face-to-face CBT in a single treatment protocol [35,41]. iCBT programs included 4 mandatory core modules of CBT (ie, psychoeducation, behavioral activation, cognitive restructuring, and relapse prevention) plus optional modules (eg, physical exercise and problem solving) typically completed at home, while face-to-face CBT was delivered in the clinic [35]. Clients worked through treatment modules, completed exercises, and monitored their symptoms on the digital program, while face-to-face sessions were used by the therapist to set up modules, monitor client progress, and address client-specific needs. Sequencing and time spent on each module were flexibly applied; however, the 4 mandatory modules on the digital program had to be completed. Data on treatment and dosage were not collected for TAU in the trial. See Table 1 for a breakdown of recruitment, bCBT format and dosage, and treatments offered in TAU across all country sites [30,35,42]. It was not possible to blind therapists to treatment allocation; however, assessors were blinded [35].

XSL•FO
RenderX

Table 1. Overview of recruitment in the trial, blended cognitive behavioral therapy format and dosage, and treatment offered in the treatment as usual arm by country site.

Country	Recruitment		bCBT[a] format and dosage				TAU[b] allocation
	Treatment setting	Recruitment procedure	Platform	Duration (weeks)	Online/face-to face, n	Sequencing[c]	
France	Specialized mental health care	New or regular patients recruited by CBT[d] therapists from 11 expert centers throughout France.	Moodbuster	16	8/8	Alternate	Face-to-face CBT
Germany	Primary care	Recruitment in the waiting room of GP[e] clinics or during GP consultations.	Moodbuster	11-13	10/6	Alternate	GP care (eg, watchful waiting, medication prescription, referral to medical specialists, or face-to-face CBT)
Netherlands	Specialized mental health care	Recruitment through mood disorder departments of 3 outpatient clinics in Amsterdam and Leiden.	Moodbuster	20	10/10	Alternate	Evidence-based face-to-face psychotherapy (mainly CBT, interpersonal psychotherapy, problem-solving therapy, antidepressant medication, or a combination of these).
Poland	Primary care	Recruitment through primary care centers by CBT therapists (licensed and in training) in 5 major cities in Poland (Warsaw, Sopot, Poznan, Katowice, and Wroclaw).	Moodbuster	6-10	6/7	Alternate	Face-to-face CBT
Spain	Primary care	Recruitment through routine primary care from the Spanish National Health System in several cites (Valencia, Castellón, and Zamora).	Smiling is fun	10	8/3	1-4-1-4-2	Prescribed medication by the GP or received face-to-face CBT and interpersonal psychotherapy or supportive therapy once a month
Sweden	Primary care	Recruitment through collaborating primary care clinics in 3 Swedish counties (Stockholm, Linköping, and Västerås). Posters and leaflets were distributed in the waiting areas or were provided to GPs in clinics, who in turn referred potentially eligible participants.	Iterapi	10	6/4	Alternate	Usual care paths in Sweden, including general practitioner care; eg, watchful waiting, medication prescription, referral to medical specialist, or face-to-face CBT
Switzerland	Specialized mental health care	Recruitment through 2 outpatient clinics (Bern and Zurich) and individual therapists.	Deprexis	18	9/9	Alternate	Face-to-face CBT
United Kingdom	Primary care	Recruitment through the IAPT[f] NHS[g] program across London, Norfolk, Suffolk, and Berkshire. The IAPT program primarily provides evidence-based psychological therapies to people with depression and anxiety disorders.	Moodbuster	11	5/6	Alternate	Face-to-face CBT

XSL•FO
RenderX

Country	Recruitment		bCBT[a] format and dosage				TAU[b] allocation
	Treatment setting	Recruitment procedure	Platform	Duration (weeks)	Online/face-to face, n	Sequencing[c]	
Denmark[h]	Specialized mental health care	Recruitment through the Center for Telepsychiatry in specialized mental health care at the Mental Health Services of the Region of Southern Denmark, where patients are referred to the study by clinicians. Initially, patients are self-referred to the Center for Telepsychiatry.	NoDep	12	6-8/6	Alternate	Face-to-face CBT

[a]bCBT: blended cognitive behavioral therapy.

[b]TAU: treatment as usual.

[c]Sequencing of face-to-face and online sessions can include more than one session per week for either component.

[d]CBT: cognitive behavioral therapy.

[e]GP: general practitioner.

[f]IAPT: improving access to psychological therapy.

[g]NHS: National Health Service.

[h]Denmark was added as a satellite recruitment site [38] after the commencement of the project.

Based on the registered data, 194 therapists delivered trial interventions. In Germany, therapists only delivered bCBT in the treatment arm, whereas therapists from the remaining 8 country sites delivered interventions across both treatment arms. The risk of contamination was not perceived as a concern, as CBT was also offered in TAU, with the focus of the trial on investigating the blending of an internet-based intervention with face-to-face CBT when compared to routine care. Data on therapist ratings of the working alliance will be published in a separate paper to enable comprehensive reporting and discussion of the findings.

Measures

Diagnostic Assessment

In the E-COMPARED study [35], a diagnosis of major depression according to the Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV) was established at baseline using the M.I.N.I [40], a structured diagnostic interview that has been translated into 65 languages and is used for both clinical and research practice. The interview compares well with the Structured Clinical Interview for DSM-IV disorders [43] and the Composite International Diagnostic Interview [40,43]. The M.I.N.I. was also used to assess the following comorbid disorders that were part of the exclusion criteria: substance dependence, bipolar affective disorder, psychotic illness, and obsessive-compulsive disorder. The M.I.N.I was administered face-to-face or via telephone at baseline and 12-month follow-up assessments. Telephone administration of diagnostic interviews has shown good validity and reliability [44,45].

Primary Measures

The study outcomes were the working alliance and depression severity, which were measured using the Working Alliance Inventory-Short Revised–Client (WAI-SR-C) [46] and the PHQ-9 [39], respectively. The WAI-SR-C scale is based on the theory of working alliance containing 3-item subscales assessing bond, task, and goals by Bordin [15,16]. The 12 items are rated on a 5-point scale from 1 (seldom) to 5 (always), with total scores ranging between 12 and 60. Higher scores on the scale indicate better working alliance. The WAI-SR-C scale has demonstrated good reliability (internal consistency) for all 3 factors, including the bond, task, and goals subscales (Cronbach α=0.92, 0.92, and 0.89, respectively) [47]. The scale has been shown to be correlated with other therapeutic alliance scales such as the California Therapeutic Alliance Rating System [19,48] and the Helping Alliance Questionnaire-II [19,49]. The WAI-SR-C scale was only administered at 3-month assessments. Data for the WAI-SR-C scale were not collected in the TAU arm of the Swedish country site.

The PHQ-9 [39] was used to assess depression as the trial's primary clinical outcome. The PHQ-9 is a 9-item scale that can be used to screen and diagnose people for depressive disorders. Each of the 9 items is scored on a 4-point scale from 0 (not at all) to 3 (nearly every day). The total score ranges between 0 and 27, with higher scores indicating greater symptom severity. Depression severity can be grouped into the following: mild (score 0-5), moderate (6-10), moderately severe (11-15), and severe (≥16). The PHQ-9 has been shown to have good psychometric properties [39] and has demonstrated its utility as a valid diagnostic tool [50]. The PHQ-9 was administered at the baseline and 3-, 6-, and 12-month assessments; however, this study only used baseline and 3-month assessment data as the study was interested in investigating depression scores that generally corresponded to before and after treatment.

Other Measures

System Usability Scale-Client (SUS-C) [51,52] was used to assess the usability of the digital programs. The SUS-C is a 10-item self-reported questionnaire. Items are measured on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly

XSL•FO
RenderX

agree). The total SUS-C score ranges between 10 and 50 to produce a global score. Higher scores indicate better system usability. The total sum score has been found to be a valid and interpretable measure to assess the usability of internet-based interventions by professionals in mental health care settings [53]. The SUS has shown high internal reliability (eg, coefficient Ω=0.91) and good concurrent validity and sensitivity [52,53]. The SUS-C was administered at the 3-month follow-up assessment.

Demographic data on the participant's gender, age, educational attainment, marital status, and country site were collected at baseline. Baseline variables entered as covariates in the regression models included age, gender (male, female, and other), marital status (single, divorced, widowed, living together, and married), and educational level (low, middle, and high, corresponding to secondary school education or equivalent [low], college or equivalent [middle], and university degree or higher [high]).

Baseline data were completed online, face-to-face, via telephone, or a combination of these approaches. The 3-month follow-up assessments were largely completed online, with the exception of the PHQ-9 that was collected via telephone to maximize data collection of the trial's primary outcome. Data that were directly collected by researchers (ie, either in person or via telephone) were double entered to increase the accuracy of the data entry process.

Statistical Analysis

The study used an intention-to-treat (ITT) population for the data analysis [54]. While the ITT approach is standard for RCTs, some methodologists advise that a per-protocol population is more suitable for pragmatic noninferiority trials owing to concerns that a "flawed trial" is likely to incorrectly demonstrate noninferiority (eg, a trial that loses the ability to distinguish any true differences between treatment groups that are present). However, contrary to the primary analysis in the E-COMPARED trial, noninferiority tests were not performed in our analyses. A decision was made to use a pure ITT population in order to maintain the original treatment group composition achieved after the random allocation of trial participants, therefore minimizing the confounding between the treatment groups and providing unbiased estimates of the treatment effects on the working alliance [54].

Data of the E-COMPARED trial were downloaded from a data repository. All analyses employed an ITT population. All models were adjusted for baseline PHQ-9 scores, age, gender, marital status, educational attainment, and country site. Analyses were performed on SPSS (version 26 or above) [55], STATA (version 16 or above) [56], and PROCESS Macro plug-in for SPSS (version 3.5 or above) [57]. Reported *P* values are 2-tailed, with significance levels at $P \leq .05$.

Treatment Assignment as a Predictor for WAI-SR-C Scores at 3-Month Assessments

In order to test if treatment assignment predicted WAI-SR-C scores at 3-month assessments (objective 1), a fixed effects linear regression model [58] was fitted separately for WAI-SR-C composite and subscale scores (goals, task, and bond). Four models were fitted altogether.

Association Between PHQ-9 Scores and WAI-SR-C Scores at 3-Month Assessments

To determine if WAI-SR-C scores were associated with PHQ-9 scores at 3-month assessments (objective 2), a fixed effects linear regression model was fitted to investigate this association separately for the bCBT and TAU arms in order to understand the alliance-outcome association within different treatment conditions in the trial. The model was also fitted separately for WAI-SR-C composite and subscale scores. Eight models were fitted altogether.

Testing the Interaction Between WAI-SR-C and SUS-C Scores With Regard to the Relationship Between WAI-SR-C and PHQ-9 Scores

To test the interaction between 3-month SUS-C and 3-month WAI-SR-C scores in a model examining the relationship between 3-month WAI-SR-C and 3-month PHQ-9 scores, a multiple regression model was fitted separately for WAI-SR-C composite and subscale scores in order to estimate the size of the interaction. Four models were fitted altogether.

Missing Data

Multiple imputation was used to handle high levels of missing data, under the missing at random (MAR) assumption. In particular, 36.6% (345/943) of data were missing for the PHQ-9, 20.7% (195/943) were missing for the WAI-SR-C, and 27.9% (133/476) were missing for the SUS-C at 3-month assessments. We imputed data sets using the chained equation approach [59]. Tabulations of missing data across treatment conditions and country sites are presented in Tables S1-S3 in Multimedia Appendix 1. Chi-square results of differences in missing and complete data between E-COMPARED country sites are presented in Tables S4 and S5 in Multimedia Appendix 1. In the imputation model, we included all variables that were part of the analyses, including observations from the PHQ-9 at baseline and demographic variables. To account for the interaction term in the regression model, data were imputed using the just another variable (JAV) approach [60]. Multiple imputation was performed separately for bCBT and TAU to allow for condition-specific variables to be considered. For example, the SUS-C variable was only entered in the bCBT arm, as those in the TAU arm did not use a digital program.

Post hoc Analysis

Post hoc sensitivity analyses were conducted to examine if the multiple imputation approach that was used to handle missing data would lead to different conclusions when compared to a complete case analysis. Under the MAR assumption, consistent findings between the primary analysis and sensitivity analysis can strengthen the reliability of the findings [61-64], at least in situations where both the primary and sensitivity analyses are expected to be valid under similar assumptions (eg, multiple imputation and complete case analysis under the MAR assumption in the outcome variable only).

Owing to the heterogeneity of interventions offered in the TAU arm within the current pragmatic trial, a subgroup analysis was

XSL•FO
RenderX

conducted to explore the magnitude of treatment effects on the working alliance when using a subset of the sample, which compared bCBT with face-to-face CBT offered in the TAU arm in Denmark, France, Poland, Switzerland, and the United Kingdom country sites of the E-COMPARED trial [35]. The subanalysis replicated the main analysis in just 5 country sites. This enabled the working alliance in bCBT to be directly compared with a defined comparator. Results between the primary analysis and the subanalysis were compared to understand if results vary when there are multiple interventions in TAU and when there is a defined comparator (ie, face-to-face CBT) [65-67].

Results

Clinical and Demographic Characteristics

Table 2 summarizes the baseline characteristics. Among the 943 participants who consented and were randomized in the trial (bCBT=476; TAU=467) (See Figure S1 in Multimedia Appendix 1 for the trial's profile), most were female (644/943, 68.3%), were middle-aged, and had a university degree or higher (447/943, 47.4). The PHQ-9 scores (median 15, IQR 7) reflected depression of moderate severity. PHQ-9 scores at 3 months will be reported in the main trial paper, which is being prepared. The median WAI-SR-C score was 47.42 (IQR 6) in the bCBT arm and 42 (IQR 8) in the TAU arm. The median SUS-C score was 42 (IQR 9) in the bCBT arm. See Table 3 for the median (IQR) values of the WAI-SR-C and SUS-C scores across treatment groups, and see Tables S6-S8 in Multimedia Appendix 1 for the median (IQR) values of the WAI-SR-C and SUS-C scores by country site.

XSL•FO
RenderX

Table 2. Baseline characteristics of participants.

Characteristic	bCBT[a] (n=476)	TAU[b] (n=467)	Total (N=943)
Age (years)			
Median (IQR)	38 (22)	37 (22)	38 (23)
Range (minimum-maximum)	18-74	18-78	18-78
Gender (female), n (%)	318 (67)	326 (70)	644 (68)
Marital status, n (%)			
Single	159 (33)	155 (33)	314 (33)
Divorced	60 (13)	43 (9)	103 (11)
Widowed	3 (1)	6 (1)	9 (1)
Living together	95 (20)	111 (24)	206 (22)
Married	159 (33)	152 (33)	311 (33)
Level of education[c], n (%)			
Secondary school, equivalent	72 (15)	74 (16)	146 (16)
College, equivalent	179 (38)	170 (36)	349 (37)
University degree or higher	225 (47)	222 (48)	447 (47)
Country site (n=943)[d], n (%)			
Germany	86 (18)	87 (19)	173 (18)
Sweden	73 (15)	68 (15)	141 (15)
Netherlands	53 (11)	49 (11)	102 (11)
United Kingdom	49 (10)	52 (11)	101 (11)
Spain	64 (13)	63 (14)	127 (14)
France	51 (11)	54 (12)	105 (11)
Switzerland	26 (6)	24 (5)	50 (5)
Poland	42 (9)	42 (9)	84 (9)
Denmark	32 (7)	28 (6)	60 (6)
Baseline PHQ-9[e] scores[f]			
Median (IQR)	15 (7)	16 (6)	15 (7)
Range (minimum-maximum)	4-27	5-26	4-27

[a]bCBT: blended cognitive behavioral therapy.

[b]TAU: treatment as usual.

[c]Data collected were in respect to what would be considered low, middle, and high levels of education in each setting. Data were missing for 1 of 943 (0.2%) individuals in the bCBT arm.

[d]Self-reported country of birth can be found in Table S9 in Multimedia Appendix 1.

[e]PHQ-9: Patient Health Questionnaire-9.

[f]PHQ-9 severity cutoff points are as follows: 5-9, mild depression; 1-14, moderate depression; 15-19, moderately severe depression; and ≥20, severe depression [39].

XSL•FO
RenderX

Table 3. Data of Working Alliance Inventory-Short Revised–Client composite and subscale scores and System Usability Scale-Client scores collected at 3-month follow-up assessments.

Scale	bCBT[a] (n=476)	TAU[b] (n=467)	Total (N=943)
WAI-SR-C[c], median (IQR)			
Composite	47.42 (6)	42 (8)	46 (9.2)
Goals	16.08 (3.4)	14 (3.9)	15.50 (4.7)
Task	14.45 (3)	12.83 (4)	14 (4)
Bond	17 (4)	15.43 (3)	16 (3.7)
SUS-C[d], median (IQR)	42 (9)	N/A[e]	N/A

[a]bCBT: blended cognitive behavioral therapy.

[b]TAU: treatment as usual.

[c]WAI-SR-C: Working Alliance Inventory-Short Revised–Client.

[d]SUS-C: System Usability Scale-Client.

[e]N/A: not applicable.

Treatment Assignment as a Predictor for WAI-SR-C Scores

Treatment assignment significantly predicted WAI-SR-C composite, goals, task, and bond scores (See Table 4 for model summaries). Being allocated to bCBT predicted higher WAI-SR-C composite and subscale scores at 3-month assessments when compared to TAU.

Table 4. Adjusted linear regression models of treatment assignment as a predictor for Working Alliance Inventory-Short Revised–Client composite and subscale (goals, task, and bond) scores.

WAI-SR-C[a] (outcome)[b]	*B*[c] (95% CI)	*P* value
Composite	5.67 (4.48-6.86)	<.001
Goals	2.32 (1.87-2.78)	<.001
Task	1.99 (1.53-2.44)	<.001
Bond	1.36 (0.91-1.81)	<.001

[a]WAI-SR-C: Working Alliance Inventory-Short Revised–Client.

[b]Separate models were generated for WAI-SR-C composite and subscale scores (ie, goals, task, and bond).

[c]Unstandardized beta.

Association Between PHQ-9 Scores and WAI-SR-C Scores at 3-Month Assessments

Across both treatment arms, WAI-SR-C composite scores and goals and task subscale scores were significantly associated with PHQ-9 scores, in which lower PHQ-9 scores were associated with higher WAI-SR-C composite scores and goals and task subscale scores. WAI-SR-C bond scores were not significantly associated with PHQ-9 scores in both treatment arms (see Table 5 for model summaries).

Table 5. Adjusted linear regression models of associations between Patient Health Questionnaire-9 and Working Alliance Inventory-Short Revised–Client composite and subscale (goals, task, and bond) scores at 3-month assessments.

WAI-SR-C[a] (outcome)[b]	bCBT[c]		TAU[d]	
	B[e] (95% CI)	*P* value	*B*[e] (95% CI)	*P* value
Composite	−0.12 (−0.17 to −0.06)	<.001	−0.06 (−0.11 to −0.02)	.01
Goals	−0.26 (−0.41 to −0.11)	.001	−0.13 (−0.25 to −0.00)	.04
Task	−0.38 (−0.52 to −0.24)	<.001	−0.18 (−0.32 to −0.05)	.008
Bond	−0.13 (−0.27 to 0.02)	.10	−0.12 (−0.25 to 0.01)	.07

[a]WAI-SR-C: Working Alliance Inventory-Short Revised–Client.

[b]Separate models were generated for WAI-SR-C composite and subscale scores (ie, goals, task, and bond).

[c]bCBT: blended cognitive behavioral therapy.

[d]TAU: treatment as usual.

[e]Unstandardized beta.

XSL•FO
RenderX

Testing the Interaction Between WAI-SR-C and SUS-C Scores With Regard to the Relationship Between WAI-SR-C and PHQ-9 Scores

There was a significant interaction between WAI-SR-C and SUS-C scores with regard to the association between WAI-SR-C composite scores and PHQ-9 scores at 3 months (b=−0.008, 95% CI −0.01 to −0.00; P=.03). Similar findings were noted for the goals (b=−0.021, 95% CI −0.04 to −0.00; P=.03) and task (b=−0.028, 95% CI −0.05 to −0.01; P=.003) subscales but not for the bond subscale (b=−0.010, 95% CI −0.03 to 0.01; P=.30). Figure 1 shows the presence of an inverse association between composite WAI-SR-C (for composite, and the goals and task subscales but not the bond subscale) and PHQ-9 scores among those with high SUS-C scores.

Figure 1. Multiple line graphs of the interaction between SUS-C and WAI-SR-C with regard to the association between WAI-SR-C (composite scores and goals, task, and bond subscale scores) and PHQ-9 scores at 3-month assessments in the cognitive behavioral therapy arm. PHQ-9: Patient Health Questionnaire-9; SUS-C: System Usability Scale-Client; WAI-SR-C: Working Alliance Inventory-Short Revised–Client.

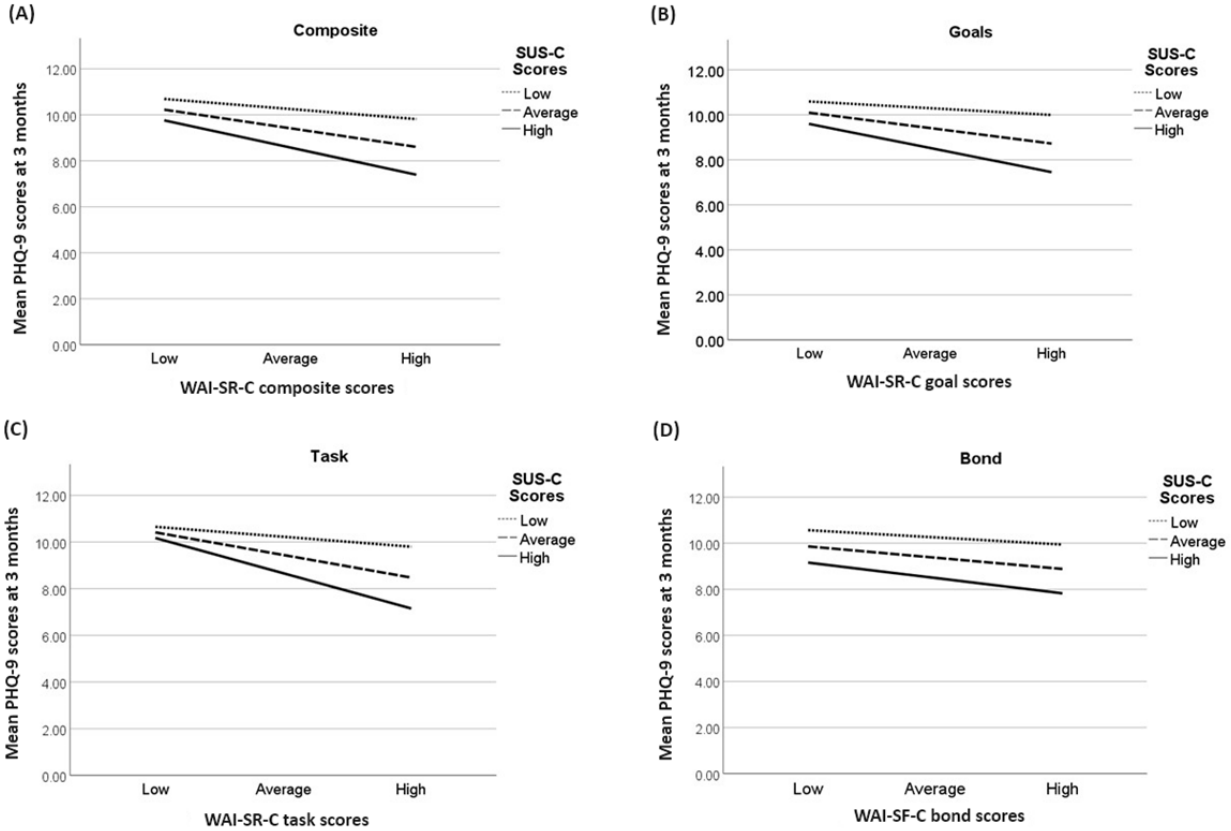

Sensitivity and Subgroup Results

The sensitivity analysis with the complete case data set and subgroup analysis of 5 country sites that only offered face-to-face CBT in the TAU arm produced results that were comparable to those reported in the main paper. However, the interaction between SUS-C (and all subscales) and WAI-SR-C scores with regard to the association between WAI-SR-C and PHQ-9 scores was not significant in terms of sensitivity. Other differences are summarized in Results S2 in Multimedia Appendix 1, while the full results of the sensitivity and subgroup analyses can be found in Results S3 and S4 in Multimedia Appendix 1.

Discussion

Principal Findings

This study investigated the client-rated working alliance in a bCBT intervention for depression when compared to TAU [35]. Overall, our study found that treatment allocation (bCBT versus TAU) was a significant predictor of working alliance scores, in which ratings of the working alliance (composite scale and goals, task, and bond subscales) were higher in bCBT than in TAU. The working alliance was significantly associated with treatment outcomes. Across both bCBT and TAU groups, as working alliance scores increased, PHQ-9 scores decreased for composite, goals, and task scores but not for bond scores. Finally, there was a significant interaction between average and above-average system usability and higher working alliance (composite scale and goals and task subscales, but not bond subscale) scores when examining the relationship between the working alliance and PHQ-9 scores at 3-month assessments.

To our knowledge, our study is the first to report that working alliance composite scores and all subscale scores were higher in bCBT than in TAU. A post hoc analysis using data from country sites that only offered face-to-face CBT in the TAU arm found that the working alliance was significantly higher in the bCBT arm compared to face-to-face CBT. These findings indicate that a blended approach may offer additional alliance-building benefits when compared to face-to-face CBT and other types of usual care for depression offered in TAU

XSL•FO
RenderX

such as talking therapies and psychopharmacological interventions. A possible explanation for our findings is that the digital elements of the intervention may enable better definition and coverage of the goals and the task than what might be possible in face-to-face sessions alone [68]. A study exploring program usage across 4 country sites of the E-COMPARED study found that clients received an average of 10 messages from their therapists online [69]. Features of the digital program that enabled the client to receive contact from the therapist away from the clinic may therefore play a role in increasing the availability of the therapist and enhancing opportunities to further strengthen the working alliance [69].

Further support for our findings comes from a qualitative study that examined the working alliance in bCBT in the United Kingdom country site of the E-COMPARED trial [36], which found that participants preferred bCBT compared to face-to-face CBT alone. The "immediacy" of access to the therapeutic task was reported to enhance engagement with the intervention and provide a higher sense of control and independence. The digital program was also described as a "secure base" that allowed participants to progressively explore self-directed treatment [36]. Similarly, a qualitative study from the German country site of the E-COMPARED trial found that bCBT was perceived to strengthen patient self-management and autonomy in relation to place and location [70].

Our study appears to be the first to identify a significant association between lower depression scores and higher working alliance composite scores and goals and task subscale scores but not bond subscale scores. In alignment with our findings, a narrative review of the working alliance in online therapy found that most guided iCBT studies included in the review reported significant associations between outcomes and the task and goals subscale scores but not the bond subscale scores [26]. A possible explanation could be that the bond is experienced differently in bCBT compared to traditional formats of CBT [26]. Bordin's [15,16] conceptualization of the working alliance suggests that while the pan-theoretical theory allows for the basic measurement of the goals, task, and bond to produce beneficial therapeutic change, the ideal alliance profile is likely to be different across therapeutic approaches and interventions [15,16,18]. The findings may therefore indicate that the working alliance profile might differ in b-CBT. However, further research is needed to investigate this.

Finally, our finding that average and higher system usability ratings may strengthen the working alliance (especially the task subscale) may point to the digital programs' influence on how the working alliance is experienced. This is not surprising given that CBT activities (eg, content and exercises) were primarily completed in the iCBT program and may indicate its relevance in the building of the working alliance and in supporting the task within a capacity that is potentially parallel to the bond. These findings partially test and support a conceptual framework of the working alliance that incorporates features that are derived from the digital program within a blended setting called "digital heuristics" (the promotion of active engagement and autonomous problem solving) in which "ease of use" and "interactivity" were identified as key features for optimizing "active engagement" with the task in the iCBT program [36]. These qualitative findings were mirrored in another study that tested the abovementioned framework, in which digital heuristics emerged as a fourth dimension when examining the working alliance in self-guided and low-intensity supported iCBT for depression [37]. High and low iCBT program functionalities were also identified by therapists as facilitators and barriers, in building the working alliance in bCBT in the German and UK country sites of the E-COMPARED trial [36,70-72]. Although our findings remain preliminary and do not show a causal effect, further investigation concerning the effect of the digital program on the working alliance may be a fruitful direction for future research.

Collectively, our findings suggest that blending face-to-face CBT with an iCBT program may enhance the working alliance and treatment outcomes for depression. These findings hold important implications for clinical practice, especially following the COVID-19 pandemic that resulted in major shifts from in-person care to blended health care provision. The findings of this study suggest that a blended approach may enhance rather than worsen mental health care. Our study's findings regarding the interaction between system usability and the working alliance in terms of treatment outcomes represent a preliminary step to quantitively understand the influence of the digital program and its role in how the working alliance is experienced. While further research is required to explore digital taxonomies that contribute toward fostering the working alliance in bCBT, our findings build on previous qualitative research [29,34,36,68] to explore a conceptualization of the working alliance that goes beyond the client and the therapist in order to consider the role of the digital program. The impact of the digital program on the working alliance may support the case of employing digital navigators who can help clients to use the intervention and troubleshoot technology and program usability issues, and remove the added burden of managing program-related problems that would otherwise fall on the therapist [70,72,73].

We propose 4 directions for future research. First, future research is required to build a comprehensive understanding of what, how, and when digital features (eg, usage, interface, interactivity, and accessibility) influence the working alliance [36]. Second, psychometric scales measuring the working alliance in bCBT should be adapted or developed to conceptually reflect a construct that also incorporates the client-program working alliance [42]. Third, the working alliance should be investigated early in the intervention and across multiple stages of treatment [74]. Fourth, future research should investigate if our results can be replicated across different DMHIs and treatment dosages.

Limitations

Several study limitations should be noted. First, working alliance data were collected at a single point that corresponded with 3-month assessments. While this is common in clinical trials [25,58], the measurement of the alliance is recommended early in treatment within the first 5 sessions and at different points across treatment [74-77]. However, the number of face-to-face sessions varied between the 9 country sites (eg, 5 to 10 sessions), which would have posed significant challenges for the systematic data collection required in a clinical trial [54].

XSL•FO
RenderX

Second, the study engaged in multiple comparisons, which may have increased the risk of type 1 error (a positive result may be due to chance). However, given the exploratory nature of this analysis and the fact that different outcomes are likely to be highly correlated, a multiple adjustment comparison was not deemed necessary [78]. Third, the results of the analysis are valid under the MAR assumption, which we believe to be plausible because the effect of country sites appears to influence the missingness of the main outcome variables, stemming from country-specific data collection procedures and experiences. This is supported by chi-square analyses that indicate significantly higher rates of missing data for the PHQ-9 and WAI-SR-C across some countries compared to others. Nevertheless, it should be noted that this paper cannot rule out that data are missing not at random. Future research can explore this further using a sensitivity analysis. Fourth, the heterogeneity of interventions offered in the TAU group limits the study from conclusively tying causation to a specific comparator intervention. However, it should be noted that interventions offered by services in TAU were regarded as evidence-based, largely consisting of CBT and psychopharmacological interventions [35]. This may reduce the limitations associated with the multiple treatments offered in TAU [66,79] and adhering to the pragmatic trial's ancillary objective to not impose specific constraints on clients and clinicians concerning data collection [79]. However, additional steps were also taken to address this limitation by conducting a subanalysis with a subset of trial country sites that only offered face-to-face CBT in TAU. The findings showed comparable results to those of the main analysis, highlighting that the addition of iCBT to face-to-face CBT may improve the quality of the working alliance. Fifth, another potential limitation is related to the variation in how bCBT was delivered across the trial's country sites, concerning the number of sessions and the types of iCBT programs delivered, across different country sites. However, it should be noted that the study was focused on investigating the noninferiority of blending CBT given that there is a sufficient level of evidence concerning key treatment components, such as the CBT approach, and different delivery formats, including in-person and internet-based delivery of CBT for depression [80,81]. Although the number of treatment sessions varied between settings, to our knowledge, there is no evidence to suggest that the number of sessions of CBT effect the client-therapist alliance as the alliance is typically developed early in treatment and within the first 5 sessions [74-77]. Moreover, another study exploring the usage of different components of bCBT and treatment engagement when compared to intended use in the E-COMPARED study concluded that personalized blended care was more suitable compared to attempting to achieve a standardized optimal blend [69]. Variations in the number of treatment sessions described may enable a pragmatic understanding of the working alliance in bCBT interventions in real-world clinical settings [66].

Conclusions

To our knowledge, this is the first study to show that bCBT may enhance the working alliance when compared to routine care for depression and when compared to face-to-face CBT. The working alliance in bCBT was also associated with clinical improvements in depression, which appear to be enhanced by good program usability. Collectively, our findings appear to add further weight to the view that the addition of iCBT to face-to-face CBT may positively augment experiences of the working alliance.

Authors' Contributions

AD had full access to all of the data and takes full responsibility for the integrity of the data and the accuracy of the data analysis. AD, MQ, FS, RA, and CF contributed to the design concept. AD drafted the manuscript. AD, RA, MQ, FS, CF, RK, HR, AK, ACP, AvS, CB, TB, KC, MM, TK, JBH, SD, IT, NT, KM, KV, AU, GA, MB, and RMB critically revised the manuscript for important intellectual content. AD, MQ, RA, HR, FS, AK, RK, CF, ACP, and SD contributed to data acquisition, analysis, and interpretation. AD contributed to statistical analysis. HR, SD, AK, MB, and ACP provided administrative, technical, or material support. AD, RA, MQ, FS, RK, and CF supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods and results that include information on: Participants' country of birth, information on missing data, medians and IQR for working alliance, participant characteristics, depression and system usability scores, trial profile, and results of the sensitivity analysis and subgroup analysis.
[DOCX File , 91 KB - jmir_v26i1e47515_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V.1.6.1).
[PDF File (Adobe PDF File), 1098 KB - jmir_v26i1e47515_app2.pdf]

References

1. GBD 2017 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis

XSL•FO
RenderX

for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Fact Sheet: Suicide. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/suicide [accessed 2024-03-24]
3. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866 [FREE Full text] [Medline: 15640922]
4. Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. Behav Res Ther 2017 Jan;88:19-25 [FREE Full text] [doi: 10.1016/j.brat.2016.08.012] [Medline: 28110672]
5. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital Mental Health and COVID-19: Using Technology Today to Accelerate the Curve on Access and Quality Tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
6. Kaltenthaler E, Parry G, Beverley C. Computerized Cognitive Behaviour Therapy: A Systematic Review. Behav. Cogn. Psychother 2004 Feb 18;32(1):31-55. [doi: 10.1017/s135246580400102x]
7. Ruwaard J, Lange A, Schrieken B, Emmelkamp P. Efficacy and effectiveness of online cognitive behavioral treatment: a decade of interapy research. Stud Health Technol Inform 2011;167:9-14. [Medline: 21685634]
8. Foroushani P, Schneider J, Assareh N. Meta-review of the effectiveness of computerised CBT in treating depression. BMC Psychiatry 2011 Aug 12;11(1):131 [FREE Full text] [doi: 10.1186/1471-244X-11-131] [Medline: 21838902]
9. Ivarsson D, Blom M, Hesser H, Carlbring P, Enderby P, Nordberg R, et al. Guided internet-delivered cognitive behavior therapy for post-traumatic stress disorder: A randomized controlled trial. Internet Interventions 2014 Mar;1(1):33-40 [FREE Full text] [doi: 10.1016/j.invent.2014.03.002]
10. Cuijpers P, Donker T, Johansson R, Mohr DC, van Straten A, Andersson G. Self-guided psychological treatment for depressive symptoms: a meta-analysis. PLoS One 2011 Jun 21;6(6):e21274 [FREE Full text] [doi: 10.1371/journal.pone.0021274] [Medline: 21712998]
11. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, et al. Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. Clin Psychol Rev 2018 Jul;63:80-92. [doi: 10.1016/j.cpr.2018.06.007] [Medline: 29940401]
12. Andrews G, Basu A, Cuijpers P, Craske M, McEvoy P, English C, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
13. Josephine K, Josefine L, Philipp D, David E, Harald B. Internet- and mobile-based depression interventions for people with diagnosed depression: A systematic review and meta-analysis. J Affect Disord 2017 Dec 01;223:28-40. [doi: 10.1016/j.jad.2017.07.021] [Medline: 28715726]
14. Erbe D, Eichert H, Riper H, Ebert D. Blending Face-to-Face and Internet-Based Interventions for the Treatment of Mental Disorders in Adults: Systematic Review. J Med Internet Res 2017 Sep 15;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
15. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research & Practice 1979;16(3):252-260. [doi: 10.1037/h0085885]
16. Bordin ES. Theory and research on the therapeutic working alliance: New directions. In: Horvath AO, Greenberg LS, editors. The working alliance: Theory, research, and practice. New York, NY: John Wiley & Sons; 1994:13-37.
17. Raue P, Goldfried M. The therapeutic alliance in cognitive-behavior therapy. In: Horvath AO, Greenberg LS, editors. The working alliance: Theory, research, and practice. New York, NY: John Wiley & Sons; 1994:131-152.
18. Lambert MJ. Psychotherapy outcome research: Implications for integrative and eclectical therapists. In: Norcross JC, Goldfried MR, editors. Handbook of psychotherapy integration. New York, NY: Basic Books; 1992:94-129.
19. Norcross JC, Lambert MJ. Psychotherapy relationships that work II. Psychotherapy (Chic) 2011 Mar;48(1):4-8. [doi: 10.1037/a0022180] [Medline: 21401268]
20. Cameron S, Rodgers J, Dagnan D. The relationship between the therapeutic alliance and clinical outcomes in cognitive behaviour therapy for adults with depression: A meta-analytic review. Clin Psychol Psychother 2018 May;25(3):446-456. [doi: 10.1002/cpp.2180] [Medline: 29484770]
21. Pihlaja S, Stenberg J, Joutsenniemi K, Mehik H, Ritola V, Joffe G. Therapeutic alliance in guided internet therapy programs for depression and anxiety disorders - A systematic review. Internet Interv 2018 Mar;11:1-10 [FREE Full text] [doi: 10.1016/j.invent.2017.11.005] [Medline: 30135754]
22. Gómez Penedo J, Berger T, Grosse Holtforth M, Krieger T, Schröder J, Hohagen F, et al. The Working Alliance Inventory for guided Internet interventions (WAI-I). J Clin Psychol 2020 Jun;76(6):973-986. [doi: 10.1002/jclp.22823] [Medline: 31240727]
23. Heim E, Rötger A, Lorenz N, Maercker A. Working alliance with an avatar: How far can we go with internet interventions? Internet Interv 2018 Mar 1;11:41-46 [FREE Full text] [doi: 10.1016/j.invent.2018.01.005] [Medline: 30135758]
24. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych Open 2019 Jan;5(1):e15 [FREE Full text] [doi: 10.1192/bjo.2018.86] [Medline: 30762511]

XSL•FO
RenderX

25. Sucala M, Schnur JB, Constantino MJ, Miller SJ, Brackman EH, Montgomery GH. The therapeutic relationship in e-therapy for mental health: a systematic review. J Med Internet Res 2012 Aug 02;14(4):e110 [FREE Full text] [doi: 10.2196/jmir.2084] [Medline: 22858538]
26. Berger T. The therapeutic alliance in internet interventions: A narrative review and suggestions for future research. Psychother Res 2017 Sep;27(5):511-524. [doi: 10.1080/10503307.2015.1119908] [Medline: 26732852]
27. Wehmann E, Köhnen M, Härter M, Liebherz S. Therapeutic Alliance in Technology-Based Interventions for the Treatment of Depression: Systematic Review. J Med Internet Res 2020 Jun 11;22(6):e17195 [FREE Full text] [doi: 10.2196/17195] [Medline: 32525484]
28. Hayati R, Bastani P, Kabir M, Kavosi Z, Sobhani G. Scoping literature review on the basic health benefit package and its determinant criteria. Global Health 2018 Mar 02;14(1):26 [FREE Full text] [doi: 10.1186/s12992-018-0345-x] [Medline: 29499708]
29. Tremain H, McEnery C, Fletcher K, Murray G. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
30. Vernmark K, Hesser H, Topooco N, Berger T, Riper H, Luuk L, et al. Working alliance as a predictor of change in depression during blended cognitive behaviour therapy. Cogn Behav Ther 2019 Jul;48(4):285-299 [FREE Full text] [doi: 10.1080/16506073.2018.1533577] [Medline: 30372653]
31. Kooistra L, Ruwaard J, Wiersma J, van Oppen P, Riper H. Working Alliance in Blended Versus Face-to-Face Cognitive Behavioral Treatment for Patients with Depression in Specialized Mental Health Care. J Clin Med 2020 Jan 27;9(2):347 [FREE Full text] [doi: 10.3390/jcm9020347] [Medline: 32012722]
32. Askjer S, Mathiasen K. The working alliance in blended versus face-to-face cognitive therapy for depression: A secondary analysis of a randomized controlled trial. Internet Interv 2021 Sep;25:100404 [FREE Full text] [doi: 10.1016/j.invent.2021.100404] [Medline: 34401364]
33. Barazzone N, Cavanagh K, Richards D. Computerized cognitive behavioural therapy and the therapeutic alliance: a qualitative enquiry. Br J Clin Psychol 2012 Nov;51(4):396-417. [doi: 10.1111/j.2044-8260.2012.02035.x] [Medline: 23078210]
34. Clarke J, Proudfoot J, Whitton A, Birch M, Boyd M, Parker G, et al. Therapeutic Alliance With a Fully Automated Mobile Phone and Web-Based Intervention: Secondary Analysis of a Randomized Controlled Trial. JMIR Ment Health 2016 Feb 25;3(1):e10 [FREE Full text] [doi: 10.2196/mental.4656] [Medline: 26917096]
35. Kleiboer A, Smit J, Bosmans J, Ruwaard J, Andersson G, Topooco N, et al. European COMPARative Effectiveness research on blended Depression treatment versus treatment-as-usual (E-COMPARED): study protocol for a randomized controlled, non-inferiority trial in eight European countries. Trials 2016 Aug 03;17(1):387 [FREE Full text] [doi: 10.1186/s13063-016-1511-1] [Medline: 27488181]
36. Doukani A, Free C, Michelson D, Araya R, Montero-Marin J, Smith S, et al. Towards a conceptual framework of the working alliance in a blended low-intensity cognitive behavioural therapy intervention for depression in primary mental health care: a qualitative study. BMJ Open 2020 Sep 23;10(9):e036299 [FREE Full text] [doi: 10.1136/bmjopen-2019-036299] [Medline: 32967872]
37. Barceló-Soler A, García-Campayo J, Araya R, Doukani A, Gili M, García-Palacios A, et al. Working alliance in low-intensity internet-based cognitive behavioral therapy for depression in primary care in Spain: A qualitative study. Front Psychol 2023;14:1024966 [FREE Full text] [doi: 10.3389/fpsyg.2023.1024966] [Medline: 37063543]
38. Mathiasen K, Andersen TE, Riper H, Kleiboer AAM, Roessler KK. Blended CBT versus face-to-face CBT: a randomised non-inferiority trial. BMC Psychiatry 2016 Dec 05;16(1):432 [FREE Full text] [doi: 10.1186/s12888-016-1140-y] [Medline: 27919234]
39. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
40. Lecrubier Y, Sheehan D, Weiller E, Amorim P, Bonora I, Sheehan KH, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur. psychiatr 2020 Apr 16;12(5):224-231. [doi: 10.1016/s0924-9338(97)83296-8]
41. van der Vaart R, Witting M, Riper H, Kooistra L, Bohlmeijer E, van Gemert-Pijnen L. Blending online therapy into regular face-to-face therapy for depression: content, ratio and preconditions according to patients and therapists using a Delphi study. BMC Psychiatry 2014 Dec 14;14:355 [FREE Full text] [doi: 10.1186/s12888-014-0355-z] [Medline: 25496393]
42. Herrero R, Vara M, Miragall M, Botella C, García-Palacios A, Riper H, et al. Working Alliance Inventory for Online Interventions-Short Form (WAI-TECH-SF): The Role of the Therapeutic Alliance between Patient and Online Program in Therapeutic Outcomes. Int J Environ Res Public Health 2020 Aug 25;17(17):6169 [FREE Full text] [doi: 10.3390/ijerph17176169] [Medline: 32854381]
43. Sheehan D, Lecrubier Y, Harnett Sheehan K, Janavs J, Weiller E, Keskiner A, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. European Psychiatry 1997;12(5):232-241 [FREE Full text] [doi: 10.1016/S0924-9338(97)83297-X]
44. Rohde P, Lewinsohn PM, Seeley JR. Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. Am J Psychiatry 1997 Nov;154(11):1593-1598. [doi: 10.1176/ajp.154.11.1593] [Medline: 9356570]

XSL•FO
RenderX

45. Ruskin PE, Reed S, Kumar R, Kling MA, Siegel E, Rosen M, et al. Reliability and acceptability of psychiatric diagnosis via telecommunication and audiovisual technology. Psychiatr Serv 1998 Aug;49(8):1086-1088. [doi: 10.1176/ps.49.8.1086] [Medline: 9712219]
46. Horvath AO, Greenberg LS. Development and validation of the Working Alliance Inventory. Journal of Counseling Psychology 1989;36(2):223-233. [doi: 10.1037//0022-0167.36.2.223]
47. Cahill J, Barkham M, Hardy G, Gilbody S, Richards D, Bower P, et al. A review and critical appraisal of measures of therapist-patient interactions in mental health settings. Health Technol Assess 2008 Jun;12(24):iii, ix-iii, 47 [FREE Full text] [doi: 10.3310/hta12240] [Medline: 18510875]
48. Fenton L, Cecero J, Nich C, Frankforter T, Carroll K. Perspective is everything: the predictive validity of six working alliance instruments. J Psychother Pract Res 2001;10(4):262-268 [FREE Full text] [Medline: 11696653]
49. Luborsky L, Barber J, Siqueland L, Johnson S, Najavits L, Frank A, et al. The Revised Helping Alliance Questionnaire (HAq-II) : Psychometric Properties. J Psychother Pract Res 1996;5(3):260-271 [FREE Full text] [Medline: 22700294]
50. Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. Gen Hosp Psychiatry 2007 Sep;29(5):388-395. [doi: 10.1016/j.genhosppsych.2007.06.004] [Medline: 17888804]
51. Brooke J. SUS: A 'Quick and Dirty' Usability Scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. London, UK: CRC Press; 1996.
52. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]
53. Mol M, van Schaik A, Dozeman E, Ruwaard J, Vis C, Ebert D, et al. Dimensionality of the system usability scale among professionals using internet-based interventions for depression: a confirmatory factor analysis. BMC Psychiatry 2020 May 12;20(1):218 [FREE Full text] [doi: 10.1186/s12888-020-02627-8] [Medline: 32398111]
54. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. Perspect Clin Res 2016;7(3):144-146 [FREE Full text] [doi: 10.4103/2229-3485.184823] [Medline: 27453832]
55. IBM SPSS Statistics 26. IBM Corp. URL: https://www.ibm.com/support/pages/ibm-spss-statistics-26-documentation [accessed 2024-03-24]
56. Stata Statistical Software: Release 16. Stata Corp. URL: https://www.scirp.org/reference/referencespapers?referenceid=2757660 [accessed 2024-03-24]
57. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis (Second Edition): A Regression-Based Approach. New York, NY: Guilford Press; 2017.
58. Kirkwood BR, Stern JAC. Essential Medical Statistics, 2nd Edition. Hoboken, NJ: Wiley; 2003.
59. Carpenter JR, Kenward MG. Multiple Imputation and its Application. New York, NY: John Wiley & Sons; 2013.
60. von Hippel PT. 8. How to Impute Interactions, Squares, and other Transformed Variables. Sociological Methodology 2009 Aug 01;39(1):265-291 [FREE Full text] [doi: 10.1111/j.1467-9531.2009.01215.x]
61. de Souza R, Eisen R, Perera S, Bantoto B, Bawor M, Dennis B, et al. Best (but oft-forgotten) practices: sensitivity analyses in randomized controlled trials. Am J Clin Nutr 2016 Jan;103(1):5-17 [FREE Full text] [doi: 10.3945/ajcn.115.121848] [Medline: 26675766]
62. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 2013 Jul 16;13:92 [FREE Full text] [doi: 10.1186/1471-2288-13-92] [Medline: 23855337]
63. Parpia S, Morris T, Phillips M, Wykoff C, Steel D, Thabane L, Retina Evidence Trials InterNational Alliance (R.E.T.I.N.A.) Study Group. Sensitivity analysis in clinical trials: three criteria for a valid sensitivity analysis. Eye (Lond) 2022 Nov;36(11):2073-2074 [FREE Full text] [doi: 10.1038/s41433-022-02108-0] [Medline: 35585134]
64. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. BMC Med Res Methodol 2014 Jan 24;14(1):11 [FREE Full text] [doi: 10.1186/1471-2288-14-11] [Medline: 24456267]
65. Burke J, Sussman J, Kent D, Hayward R. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015 Nov 04;351:h5651 [FREE Full text] [doi: 10.1136/bmj.h5651] [Medline: 26537915]
66. Dawson L, Zarin D, Emanuel E, Friedman L, Chaudhari B, Goodman S. Considering usual medical care in clinical trial design. PLoS Med 2009 Sep;6(9):e1000111 [FREE Full text] [doi: 10.1371/journal.pmed.1000111] [Medline: 19787044]
67. Farrokhyar F, Skorzewski P, Phillips M, Garg S, Sarraf D, Thabane L, Retina Evidence Trials InterNational Alliance (R.E.T.I.N.A.) Study Group. When to believe a subgroup analysis: revisiting the 11 criteria. Eye (Lond) 2022 Nov;36(11):2075-2077 [FREE Full text] [doi: 10.1038/s41433-022-01948-0] [Medline: 35102244]
68. Doukani A, Free C, Araya R, Michelson D, Cerga-Pashoja A, Kakuma BR. Practitioners' experience of the working alliance in a blended cognitive-behavioural therapy intervention for depression: qualitative study of barriers and facilitators. BJPsych Open 2022 Jul 25;8(4):e142 [FREE Full text] [doi: 10.1192/bjo.2022.546] [Medline: 35876079]
69. Kemmeren LL, van Schaik A, Smit JH, Ruwaard J, Rocha A, Henriques M, et al. Unraveling the Black Box: Exploring Usage Patterns of a Blended Treatment for Depression in a Multicenter Study. JMIR Ment Health 2019 Jul 25;6(7):e12707 [FREE Full text] [doi: 10.2196/12707] [Medline: 31344670]

XSL•FO
RenderX

70. Titzler I, Saruhanjan K, Berking M, Riper H, Ebert D. Barriers and facilitators for the implementation of blended psychotherapy for depression: A qualitative pilot study of therapists' perspective. Internet Interv 2018 Jun;12:150-164 [FREE Full text] [doi: 10.1016/j.invent.2018.01.002] [Medline: 30135779]
71. Titzler I, Berking M, Schlicker S, Riper H, Ebert D. Barriers and Facilitators for Referrals of Primary Care Patients to Blended Internet-Based Psychotherapy for Depression: Mixed Methods Study of General Practitioners' Views. JMIR Ment Health 2020 Aug 18;7(8):e18642 [FREE Full text] [doi: 10.2196/18642] [Medline: 32673213]
72. Cerga-Pashoja A, Doukani A, Gega L, Walke J, Araya R. Added value or added burden? A qualitative investigation of blending internet self-help with face-to-face cognitive behaviour therapy for depression. Psychother Res 2020 Nov 05;30(8):998-1010. [doi: 10.1080/10503307.2020.1720932] [Medline: 32024447]
73. Wisniewski H, Torous J. Digital navigators to implement smartphone and digital tools in care. Acta Psychiatr Scand 2020 Apr;141(4):350-355 [FREE Full text] [doi: 10.1111/acps.13149] [Medline: 31930477]
74. Crits-Christoph P, Gibbons MBC, Hamilton J, Ring-Kurtz S, Gallop R. The dependability of alliance assessments: the alliance-outcome correlation is larger than you might think. J Consult Clin Psychol 2011 Jun;79(3):267-278 [FREE Full text] [doi: 10.1037/a0023668] [Medline: 21639607]
75. Piper W, Azim H, Joyce A, McCallum M. Transference interpretations, therapeutic alliance, and outcome in short-term individual psychotherapy. Arch Gen Psychiatry 1991 Oct;48(10):946-953 [FREE Full text] [doi: 10.1001/archpsyc.1991.01810340078010] [Medline: 1929765]
76. Eames V, Roth A. Patient attachment orientation and the early working alliance-a study of patient and therapist reports of alliance quality and ruptures. Psychother Res 2000 Dec 23;10(4):421-434. [doi: 10.1093/ptr/10.4.421] [Medline: 21756114]
77. Norcross JC, Wampold BE. Evidence-based therapy relationships: research conclusions and clinical practices. Psychotherapy (Chic) 2011 Mar;48(1):98-102. [doi: 10.1037/a0022161] [Medline: 21401280]
78. Rothman K. No Adjustments Are Needed for Multiple Comparisons. Epidemiology 1990;1(1):43-46 [FREE Full text] [doi: 10.1097/00001648-199001000-00010]
79. Giraudeau B, Caille A, Eldridge S, Weijer C, Zwarenstein M, Taljaard M. Heterogeneity in pragmatic randomised trials: sources and management. BMC Med 2022 Oct 28;20(1):372 [FREE Full text] [doi: 10.1186/s12916-022-02569-w] [Medline: 36303153]
80. Karyotaki E, Efthimiou O, Miguel C, Bermpohl F, Furukawa T, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
81. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022 Sep 15;5(1):144 [FREE Full text] [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]

Abbreviations

bCBT: blended cognitive behavioral therapy
CBT: cognitive behavioral therapy
DMHI: digital mental health intervention
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders IV
E-COMPARED: European Comparative Effectiveness Research on Blended Depression Treatment versus Treatment-as-usual
iCBT: internet-delivered cognitive behavioral therapy
ITT: intention to treat
MAR: missing at random
M.I.N.I: MINI International Neuropsychiatric Interview
PHQ-9: Patient Health Questionnaire-9
SUS-C: System Usability Scale-Client
TAU: treatment as usual
WAI-SR-C: Working Alliance Inventory-Short Revised–Client

XSL•FO
RenderX

Edited by Y Hong; submitted 29.03.23; peer-reviewed by A AL-Asadi, A González-Robles; comments to author 26.01.24; revised version received 09.02.24; accepted 11.02.24; published 31.05.24.

Please cite as:
Doukani A, Quartagno M, Sera F, Free C, Kakuma R, Riper H, Kleiboer A, Cerga-Pashoja A, van Schaik A, Botella C, Berger T, Chevreul K, Matynia M, Krieger T, Hazo JB, Draisma S, Titzler I, Topooco N, Mathiasen K, Vernmark K, Urech A, Maj A, Andersson G, Berking M, Baños RM, Araya R
Comparison of the Working Alliance in Blended Cognitive Behavioral Therapy and Treatment as Usual for Depression in Europe: Secondary Data Analysis of the E-COMPARED Randomized Controlled Trial
J Med Internet Res 2024;26:e47515
URL: https://www.jmir.org/2024/1/e47515
doi:10.2196/47515
PMID:38819882

©Asmae Doukani, Matteo Quartagno, Francesco Sera, Caroline Free, Ritsuko Kakuma, Heleen Riper, Annet Kleiboer, Arlinda Cerga-Pashoja, Anneke van Schaik, Cristina Botella, Thomas Berger, Karine Chevreul, Maria Matynia, Tobias Krieger, Jean-Baptiste Hazo, Stasja Draisma, Ingrid Titzler, Naira Topooco, Kim Mathiasen, Kristofer Vernmark, Antoine Urech, Anna Maj, Gerhard Andersson, Matthias Berking, Rosa María Baños, Ricardo Araya. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of a Web-Based Outpatient Asynchronous Consultation Service: Mixed Methods Study

Magdalena Rzewuska Díaz[1], BSc, PhD; Louise Locock[1], MA, MPhil, Prof Dr; Andrew Keen[2], BSc, PhD; Mike Melvin[3], BA; Anthony Myhill[3]; Craig Ramsay[1], BSc, Prof Dr

[1]Health Services Research Unit, University of Aberdeen, Aberdeen, United Kingdom

[2]NHS Grampian, Grampian, Aberdeen, United Kingdom

[3]Public Research Partner, University of Aberdeen, Aberdeen, United Kingdom

Corresponding Author:
Magdalena Rzewuska Díaz, BSc, PhD
Health Services Research Unit
University of Aberdeen
2nd floor
Health Sciences Building
Aberdeen, AB252ZD
United Kingdom
Phone: 44 1224438148
Email: magdalena.rzewuska@abdn.ac.uk

Abstract

Background: Asynchronous outpatient patient-to-provider communication is expanding in UK health care, requiring evaluation. During the pandemic, Aberdeen Royal Infirmary in Scotland expanded its outpatient asynchronous consultation service from dermatology (deployed in May 2020) to gastroenterology and pain management clinics.

Objective: We conducted a mixed methods study using staff, patient, and public perspectives and National Health Service (NHS) numerical data to obtain a rounded picture of innovation as it happened.

Methods: Focus groups (3 web-based and 1 face-to-face; n=22) assessed public readiness for this service, and 14 interviews with staff focused on service design and delivery. The service's effects were examined using NHS Grampian service use data, a patient satisfaction survey (n=66), and 6 follow-up patient interviews. Survey responses were descriptively analyzed. Demographics, acceptability, nonattendance rates, and appointment outcomes of users were compared across levels of area deprivation in which they live and medical specialties. Interviews and focus groups underwent theory-informed thematic analysis.

Results: Staff anticipated a simple technical system transfer from dermatology to other receptive medical specialties, but despite a favorable setting and organizational assistance, it was complicated. Key implementation difficulties included pandemic-induced technical integration delays, misalignment with existing administrative processes, and discontinuity in project management. The pain management clinic began asynchronous consultations (*digital appointments*) in December 2021, followed by the gastroenterology clinic in February 2022. Staff quickly learned how to explain and use this service. It was thought to function better for pain management as it fitted preexisting practices. From May to September 2022, the dermatology (adult and pediatric), gastroenterology, and pain management clinics offered 1709 appointments to a range of patients (n=1417). Digital appointments reduced travel by an estimated 44,712 miles (~71,956.81 km) compared to the face-to-face mode. The deprivation profile of people who chose to use this service closely mirrored that of NHS Grampian's population overall. There was no evidence that deprivation impacted whether digital appointment users subsequently received treatment. Only 18% (12/66) of survey respondents were unhappy or very unhappy with being offered a digital appointment. The benefits mentioned included better access, convenience, decreased travel and waiting time, information sharing, and clinical flexibility. Overall, patients, the public, and staff recognized its potential as an NHS service but highlighted informed choice and flexibility. Better communication—including the use of the term *assessment* instead of *appointment*—may increase patient acceptance.

Conclusions: Asynchronous pain management and gastroenterology consultations are viable and acceptable. Implementing this service is easiest when existing administrative processes face minimal disruption, although continuous support is needed. This study can inform practical strategies for supporting staff in adopting asynchronous consultations (eg, preparing for nonlinearity

XSL•FO
RenderX

and addressing task issues). Patients need clear explanations and access to technical support, along with varied consultation options, to ensure digital inclusion.

(J Med Internet Res 2024;26:e48092) doi:10.2196/48092

KEYWORDS

outpatient care; teleconsultation; asynchronous communication; implementation; mixed methods research; qualitative research; hospital data

Introduction

Background

In recent decades, health care organizations have adopted innovative remote patient-to-provider communication methods such as video, instant messaging, and email. The nature of the COVID-19 pandemic caused extraordinary growth in their use, which relieved health system strains by allowing patients to obtain primary and specialized care without an in-person appointment [1,2]. As these new forms of patient-to-provider communication are expected to become a part of delivering health care, valuable lessons can be learned from this accelerated uptake under emergency measures [3].

Asynchronous consultations, where health care staff and patients do not need to be available at the same time, are a potential alternative to synchronous methods (in real time), such as traditional in-person meetings, phone calls, and videoconferences. They have been widely tested in primary care [4-8] but are increasingly used in secondary care outpatient services, particularly for highly visible symptoms such as dermatological problems [9,10] or epileptic seizures [11]. Typically, outpatient asynchronous approaches involve patients answering a number of specialty-specific questions about their health over the web at times that suit them, usually within a window of approximately a week (responses may include photographs or video footage). This completed form is then reviewed by the outpatient clinic, and they can request further information over a similar period before deciding on the most suitable outcome (eg, examinations, treatments, face-to-face consultations, or outpatient discharge).

Asynchronous consultations provide a new health care model, unlike videoconferencing. A systematic review found that outpatient asynchronous consultations may produce equivalent outcomes to in-person treatment and lower health care expenses for certain conditions [12]. The pandemic has increased the acceptability of remote consultations, but technical, educational, infrastructure, legal, and economic challenges must be solved for them to be sustainable and scalable beyond the pandemic [13]. Because of this, emphasis has been placed on determining the resources and approaches required to effectively implement the technology required to deliver asynchronous consultations (by assessing how they are adopted) and understanding their effects on staff and patients [3,12,14,15].

Objectives

A nationwide case study by Wherton et al [8] indicated that National Health Service (NHS) boards in Scotland expanded video consultations before and after the epidemic. The NHS Grampian board serves a population of nearly 600,000 in urban, remote, and rural locations across a large geographic area. It also provides some specialist services to other boards in the north of Scotland, including the island communities of Shetland and Orkney. Similar to other boards in Scotland but to a greater extent, NHS Grampian substantially increased its use of video consultations [8]. The Department of Dermatology at Aberdeen Royal Infirmary introduced asynchronous consultations as part of a multiboard, nationally led innovation project in May 2020. Due to the pandemic, the organization expanded this new service to the gastroenterology and pain management clinics. The dermatology system provided a template and contextual insights for rolling out this new service. This study examined staff, patient, and public perspectives and NHS quantitative data to provide a complete picture of the outpatient asynchronous consultation service adoption.

Methods

Study Context

The list of parties involved in developing and implementing new pathways is shown in Textbox 1. The NHS Grampian team approached the Health Services Research Unit for help with assessing this implementation's effects. An implementation researcher (MRD), an expert qualitative social science researcher (LL), a service improvement researcher (CR), NHS Grampian's clinical lead for innovation (AK), a service evaluation lead (KM), and 2 public research partners knowledgeable in digital skills assistance (MM and AM) formed the study team. The lay partners participated in project conceptualization and all research stages as coinvestigators (from design to delivery and dissemination).

XSL•FO
RenderX

Textbox 1. A list of adopters of the newly implemented pathways (gastroenterology and pain management) of asynchronous consultations.

Stakeholders and responsibilities

- Asynchronous coordinator: supporting the administrative process surrounding the onboarding of patients and logging issues with the third-party supplier; assisting with troubleshooting
- Clinical leads: defining pathway question sets associated with their specialty
- Clinic coordinators: booking patients into asynchronous consultations and answering some onboarding queries (although this was most likely done by the asynchronous coordinator)
- Clinicians: reviewing the information provided by the patients within the asynchronous consultations and issuing an outcome via the software to the patient
- eHealth: involved in the technical configuration of appointment messaging and outgoing, and onward delivery to appropriate National Health Service (NHS) Grampian electronic patient records; calls may also be logged on the IT service desk, but these are mainly resolved by the third-party supplier, National Services Scotland (NSS), or administrative support teams involved in supporting the system within their departments
- Information governance: ensuring that the project provides the appropriate assurances to enable an NHS Grampian data protection officer Caldicott Guardian to approve required paperwork associated with the system
- IT security: ensuring that the project provides the appropriate assurances to enable the IT security officer to approve required paperwork associated with the system
- NSS and NHS Education for Scotland: involved in technical configuration and delivery of messaging between NHS Grampian and third-party supplier systems and technical troubleshooting
- Outpatient services management: overall responsibility for the patient information leaflets that give patients information about the asynchronous consultations
- Project team: managing the product rollout within the pain management and gastroenterology clinics, coordinating all staff involved to ensure that all prerequisites are met, facilitating a smooth transition to the deployment phase; managing handover to business-as-usual teams
- Service management (gastroenterology only): helping ensure that the process is embedded within gastroenterology
- Storm ID: supplier of the software; enhancing the software and troubleshooting issues

The Asynchronous Consultation System

This service is for routine appointments only. An outpatient team vets returning or new patients referred by primary care physicians to assess suitability for an asynchronous consultation (that NHS Grampian calls *digital appointments*). After reaching the top of the waiting list, a patient receives an information pamphlet explaining a digital appointment and a letter encouraging them to call the operational team to book. If the patient declines, the booking staff offer a synchronous alternative. An appointment is generated in TrakCare, a patient management system, which then triggers the invitation to the patient to register for the digital system. After registration, the patient receives a link to a web-based inquiry form (a web-based questionnaire) with 5 days to complete before the scheduled date (Figure 1). Sample areas of questioning in the pain management pathway include experience of pain, symptom descriptions and improvement level, adverse effects, pain medication, mood, ability to undertake normal work, sleep, enjoyment of life, and any other relevant experiences. Sample areas of questioning in the gastroenterology pathway include description of symptoms, bowel movements and stool, experience of abdominal pain and bloating, weight loss, medication taken, ability to perform normal work and daily tasks, enjoyment of life, and concerns. Patients register, answer questions, check, and submit. After analyzing the information and maybe asking further questions through real-time messaging, a clinician determines whether to provide remote advice and treatment, schedule a face-to-face visit, or discharge. The patient, outpatient clinical team, and general practitioner receive a PDF appointment summary (via the NHS email address of a clinician system user) stored in a document repository (Skystore). Delivery is via Storm ID (third party), and Lenus is the platform used.

XSL•FO
RenderX

Figure 1. Digital appointments: an asynchronous consultation process within the National Health Service Grampian.

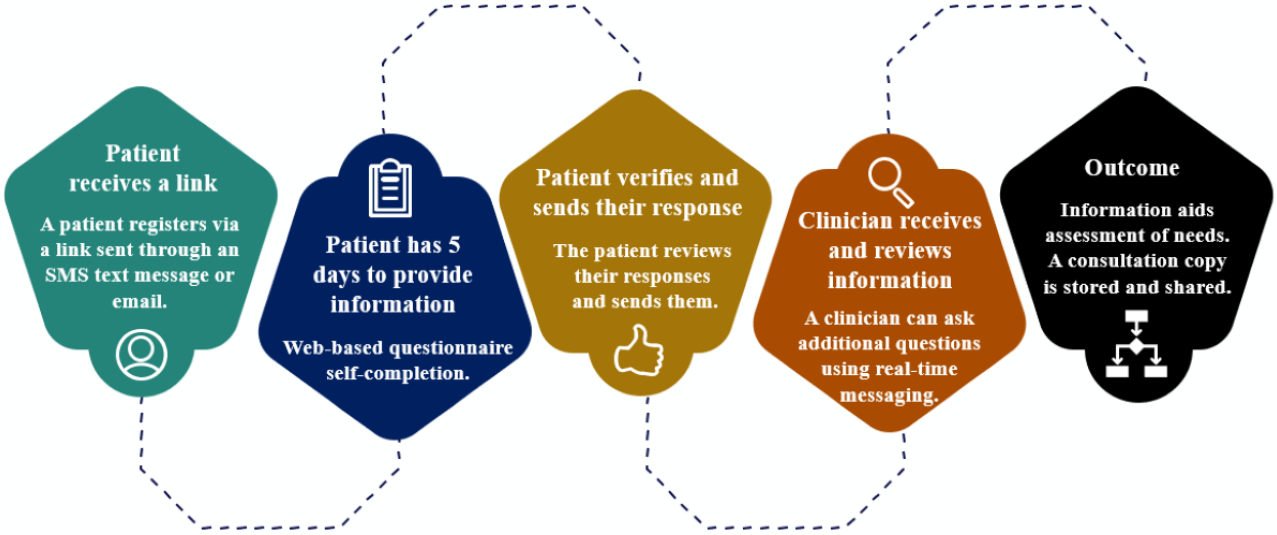

Study Design

We deployed a mixed method study design with a multistage purposeful sampling strategy [16]. It included public focus groups about public readiness for the new service (3 web-based and 1 face-to-face), interviews with staff about their experiences with designing and implementing the service; and an assessment of service uptake conducted through analysis of the service use data and patient experience (including a patient satisfaction survey and follow-up patient interviews about their experiences with receiving the service). Qualitative study, including interviews and focus groups, ran from April 2021 to July 2022. The NHS service evaluation covered a satisfaction survey from January 2022 to September 2022 and analysis of NHS Grampian service usage data between May 2020 and September 2022.

The primary focus was on the newly implemented pathways—gastroenterology and pain management—whereas adult dermatology served effectively to provide context, helping illustrate the transferability from one medical specialty to another in a real-world scenario. Focus groups with members of the public and staff interviews covered all 3 medical specialties. The NHS Grampian data on service use covered 4 medical specialties: pain management, gastroenterology, and adult and pediatric dermatology. The patient satisfaction survey, conducted by NHS Grampian, was confined by our ethical clearance to prospective evaluations of the newly implemented pathways as dermatology was assessed independently. All participants in the satisfaction survey who were happy to be contacted by a researcher were eligible for the interviews.

Overall, the study was designed to answer six research questions:

1. How accepting and prepared is the public for asynchronous consultations?
2. What are clinical and administrative staff experiences of and attitudes toward asynchronous consultation in secondary care?
3. Is asynchronous consultation sustainable beyond the immediate context of COVID-19?
4. How do patients feel about the quality of asynchronous consultations and their effect on their relationship with health care professionals?
5. How do asynchronous consultations affect health inequalities in terms of access?
6. What is the impact of an asynchronous consultation model on NHS performance?

Ethical Considerations

The NHS service evaluation component (satisfaction survey and NHS Grampian service use data analysis) led by the NHS Grampian did not require ethical review. Ethics approval for the qualitative study was granted by the London–Bloomsbury Research Ethics Committee (reference 21/PR/0051). Active informed verbal or written consent was obtained from each qualitative study participant individually. Potential participants had at least 24 hours to consider whether or not they would like to participate in the research and were free to withdraw at any time and without giving a reason. As compensation for their time and effort, the public focus group participants were offered a £15 (approximately US $19.5) retail-gift voucher, but staff and patient interview participants were not compensated for their time. Qualitative data were deidentified before analyzing.

Focus Groups About Public Readiness

Recruitment and Data Collection

We recruited members of the public for participation in focus groups, targeting individuals at risk of digital exclusion, including low-income people, older people, people with disabilities, and those who assist them. Our 2 public coinvestigators (MM and AM) who work with community voluntary sector organizations contacted members of the public to invite them to participate. In total, 3 web-based (using Microsoft Teams platform) focus groups and 1 face-to-face focus group of 4 to 7 participants were held, with 22 adult participants in total (n=8, 36% men and n=14, 64% women; n=6, 27% aged 20-40 years; n=5, 23% aged 41-64 years; and n=11, 50% aged 65 years). The asynchronous consultation service used by the dermatology clinic was explained and demonstrated to participants at the start of the focus groups. In

XSL•FO
RenderX

addition, the group was informed about what new medical specialties this service would be expanded to, highlighting its potential applicability across various populations. The facilitator (MRD) used a semistructured topic guide (Multimedia Appendix 1, item A).

Data Analysis

Focus groups were video or audio recorded and professionally transcribed, checked for accuracy, deidentified, and coded using the NVivo software (version 10; Lumivero). The technology acceptance model [17]—one of the most influential technology acceptance models, which proposes that perceived ease of use and usefulness of a technological tool determine the extent of user acceptance—informed the analysis. Analysis began alongside data collection, with ideas from early analysis informing later data collection in an iterative process. Analysis of transcripts commenced with familiarization conducted by 2 experienced qualitative researchers (MRD and LL) followed by open coding using a mixture of inductive and theory-informed codes. This generated a coding framework, which then one researcher (MRD) systemically applied to all data. The themes identified from the focus groups helped inform later qualitative stages (see the following sections).

Staff Interviews About Design and Implementation

Recruitment and Data Collection

NHS Grampian partners contacted staff members who expressed interest by contacting the project researcher. Staff were interviewed over the web (Microsoft Teams platform) using topic guides (Multimedia Appendix 1, item B) to discuss dermatology routes retrospectively and pain management and gastrointestinal routes prospectively. From the service adopters described in Textbox 1 (gastroenterology and pain management pathways), 8 people, including clinical leads, service or administration leads, and technological leads or experts (including a developer, technical integration lead, and digital health lead), were interviewed at the preimplementation project stage (June to November 2021), of whom 5 (63%) had experience with either the dermatology system alone or dermatology and the newly implemented pathways. At this stage, we additionally interviewed a dermatology clinician, totaling 9 individuals interviewed at the preimplementation stage. A total of 5 adopters of the newly implemented pathways (n=2, 40% interviewed at the preimplementation stage) were interviewed at the postimplementation stage (March to June 2022), including 2 (40%) clinical leads, 1 (20%) service or administration lead, and project management leads or experts (including an evaluation lead and a project manager).

Data Analysis

The approach was the same as for focus group analysis, but this time the coding framework was informed by the evidence-based conceptual model by Greenhalgh et al [18] for evaluating factors affecting diffusion, dissemination, and implementation of innovations in health care delivery and organization.

Assessment of Service Uptake by Patients

Assessment of Routinely Collected Data

The NHS Grampian team examined service use data, including demographics of new system users (age, postcode, and sex), acceptance of service throughout routes (numbers completed, expressly declined, and not accepted for other reasons), nonattendance rates, and appointment results. To assess health inequality, we evaluated rate distribution across the Scottish Index of Multiple Deprivation (SIMD) [19], which ranks every small area in NHS Grampian from *most deprived* (lowest quintile of the SIMD) to *least deprived* (highest quintile of the SIMD). The data were based on hospital data collected between May 2020 and September 2022 (in this, new pathways data spanned the period from January to September 2022) and an audit carried out by administrative staff between February and July 2022. When feasible, a chi-square test was conducted to calculate percentage differences, setting the significance level at $P>.05$.

Patient Experience

All gastroenterology and pain management patients who were offered the new service were given the chance to complete a brief web-based satisfaction survey (Multimedia Appendix 1, items D-F). Collected data were analyzed descriptively. Patients who completed a survey and checked a box to be contacted by the project researcher were eligible for a follow-up interview. The NHS Grampian team research partners gave the details of consenting, eligible participants to the study researcher. The researcher (MRD) contacted participants via email or phone call and sent reminders at 2 weeks. Semistructured interviews were conducted over the web (Microsoft Teams platform) or by phone to better understand their experiences (see Multimedia Appendix 1, item C, for the patient interview topic guide) and analyzed in the same way as the focus group transcripts. From March to July 2022, a total of 6 pain management patients were interviewed, of whom 4 (67%) completed their appointments and 2 (33%) did not. Table S1 in Multimedia Appendix 2 describes the interview participants. Only 2 gastroenterology patients consented to be contacted by the researchers, but neither responded.

Results

Focus Groups About Public Readiness

COVID-19's Impact on the Digital Revolution

Participants described how the COVID-19 pandemic significantly affected the population's digital technology habits, including remote consultations, and has sped up many valuable previously hard-to-imagine digital innovations. They aired concerns about how this may change the fabric of society, adding to the problem of technology addiction and diminishing human interactions.

Attitudes Toward Asynchronous Consultations

Some participants described having personal, satisfactory experiences using remote communication (emails and phone and video calls) with health care providers. Most participants were very positive about the new asynchronous consultation

XSL•FO
RenderX

system. Even those who expressed a clear personal preference for remote synchronous and face-to-face consultations could see how the service could be useful for some aspects of care, such as triage, monitoring the health care needs of people with long-term conditions, or reaching people who have difficulty traveling to hospital.

Relative Advantages and Disadvantages of Asynchronous Consultations

Many possible benefits were cited regarding access, efficiency, time and travel saved, and more flexibility for clinicians, as well as reminders that we should not make assumptions about who would find asynchronous consultations hard to use. Downsides were also aired, particularly for those for whom English is a second language, people without a good internet connection or adequate devices, people with conditions that need physical examination or symptoms that are difficult to describe in words, and people with cognitive or developmental disabilities; the downsides included risks of urgent issues being missed, loss of human contact and empathy, fragmentation of care, and the potential that the length of time one must wait for a reply could create anxiety (although accepting it might still be quicker than waiting for a face-to-face appointment).

Practical Barriers to and Facilitators of Using the Asynchronous Service

Participants noted that the practical barriers, such as the lack of equipment and skills, could be mitigated with the help of community-funded programs or by setting up adequate and safe spaces in the community or providing an alternative method of communication (such as telephone). Patients' fears regarding data privacy and confidentiality would need to be allayed, the group explained, as these issues would otherwise serve as a barrier.

Compatibility With Needs

As the new service might not work for everyone, participants stressed the importance of being mindful of personal preferences (eg, for communication format), tailoring the system to end users' needs, and the service remaining optional. Most participants wondered whether being routed to an asynchronous system might make people feel fobbed off or less valued; another person noted that those who could not use it might feel less valued or like second-class citizens. The underlying concern was that digital exclusion and opting out could exacerbate health inequalities.

Further details on findings under the developed themes can be found in Multimedia Appendix 2.

Staff Interviews About Design and Implementation

Findings from staff interviews are presented drawing on components of the conceptual model by Greenhalgh et al [18] for the diffusion and implementation of innovations. Further details on findings under the developed themes and on implementation setbacks specifically can be found in Tables S2 and S3 in Multimedia Appendix 2, respectively.

Characteristics of the Innovation Itself

Staff involved in the implementation of asynchronous consultations were motivated to roll out the innovation by their assessment that the new system could give clinicians more flexibility, give patients more convenience, enable access to the right care for the right people at the right time (triage or assessment), and improve information sharing (among a patient, an outpatient clinic, and general practice). Over time, those benefits began to manifest. For example, staff noted that, at first, going over a digital appointment response was not much faster than a phone call but, later, it became quicker, taking 15 minutes instead of 30 minutes. Digital exclusion and adding an extra step to the care pathway were among the early concerns that adopters wished to evaluate. The most mentioned concern was a loss of human contact, the risk of which staff felt was reduced by using real-time messaging functions with new patients or creating a pathway where patients are already engaged with care providers. Overall, there was alignment with public opinions, particularly regarding the use for triage and main concerns about digital exclusion and the loss of human contact:

> *You know, we've got a better understanding of the impact of pain and their mood and their function, and how much it's interrupting daily activities, and that's not information that we would have had before we would have had to gained all the information during the first appointment.* [Clinical lead 3]

> *[in pain pathways] Patients are already engaged following a face-to-face appointment with the nursing teams prior to the async appointments. This provides a human touch and reassurance that patients' issues are known and will be considered, listened to.* [Project management lead or expert 1]

It was generally accepted that the innovation might be somewhat changed to satisfy different needs except when the product design and deployment are controlled centrally as opposed to locally (as a part of the national pilot or scale-up project). The new pathways' clinical leads determined the type of information gathered, the questions to be asked during the consultation, and the place of the digital appointment in the patient experience, which mirrors the public's expectations for the service to be tailored to the needs of the end users. Clinical leads described 2 examples of how the system's function evolved according to needs after it went live (eg, the pain management clinic started to use the system to manage medication use) and continued proposing new modifications (eg, a clinical lead in gastroenterology described plans for being able to share educational materials with patients).

The developers of the newly implemented pathways viewed practical demonstrations of the dermatology system as helpful (especially pathway design and patient booking process). For instance, adopters of the new pathway tried to draw lessons from the fact that the dermatology pathway was initially too disjointed and complex for patients. Still, there was a perceived lack of space and time to try out the new systems before they went live, but they recognized that the new process could only be tested once technically integrated with the system.

XSL•FO
RenderX

After launching the service, it became evident that the new service created additional tasks for staff. Staff described additional work tasks related to appointment booking and management process, such as manual record keeping, chasing up dermatology patients, and having to phone patients to offer an appointment. In general, clinicians' manual selection of patients from waiting lists or administrative staff's selection of patients out of order and cold calls to patients were seen as ineffective. A clinical lead explained how, initially, at their end, there was also a lot of paperwork involved. However, later, staff's familiarization with the technology and growing confidence promoted more efficient use of the system (eg, using a filter function to search waiting lists).

The availability of tailored technical support was viewed as needed and important, and there was an appreciation for the support provided through eHealth colleagues. The newly implemented pathways, shortly after being launched, experienced technical issues from the service side that prevented patients from accessing links. Staff worked around this problem by sending new appointments and, ultimately, offering an in-person appointment. Eventually, the eHealth team resolved the problem's source, but a few patients could not join despite trying multiple times. Initially, clinical leads also reported a problem with creating user accounts with the system using their NHS email addresses, which was promptly resolved. When at work, for improved efficiency, clinicians used one monitor for viewing digital appointments and another for the patient appointment booking system. However, this dual-monitor setup was not accessible when working from home:

> *So, when we're trying to test something, we've got to get everyone involved in the test process, we have to get everyone to change everything at the same time, and then if the test doesn't work then we need to go back up through the chain to find out whereabouts it failed.* [Technical expert or lead 2]

Outer Environmental Context

Staff believed that the pandemic created a favorable environment for telemedicine in general, a perspective that was in line with the experiences of the public. The pandemic acted as the catalyst for thinking about it on a larger scale. However, it instigated an increased demand and competition for resources. Participants mentioned the video consultation service used in Grampian called *Near Me*. Even before the pandemic, they said, it was viewed as an effective new method of teleconsultation, which paved the way for teleconsultation innovation and served as a benchmark with which adopters compared the asynchronous consultation system:

> *Although we had been looking at it prior to Covid because we thought this was a way of working anyway, but Covid was the catalyst that got us thinking about it on a larger scale.* [Technical expert or lead 1]

> *In the backdrop of dealing with Covid, it's difficult in terms of the pressure within the system to try and deliver something which was new and maybe something that people were just trying to learn and implement.* [Project management lead or expert 1]

System Antecedents and Organizational Readiness

NHS Grampian was generally viewed as a context receptive to telemedicine and innovations, characterized by a recognized need to introduce a new care model and having technical expert or leadership and innovation high on its agenda. Staff can and do exchange knowledge internally and externally, which is conducive to innovating. However, the receptiveness of the environment for innovation would be further strengthened by a more clearly articulated strategic vision for priorities, more clearly defined roles and responsibilities of innovation and transformation teams, and streamlined processes (such as information governance):

> *I mean I think the main things I would say are that I think NHS Grampian has been very supportive of this.* [Technical expert or lead 2]

> *The reality is there's a lot of things that are urgent and need to be done, so I think that it's more from an innovative perspective creating a kind of culture of innovation and getting agreement that this is how we work for innovative things and understanding that process and really encouraging.* [Project management lead or expert 2]

At the time of implementation, pain management and gastroenterology systems were perceived as more likely to succeed than the dermatology system due to the belief that there were more staff supporters of the innovation in those clinics.

The Importance of Staff and Patients as Adopters of the New Service

The innovation was developed and implemented by motivated and keen innovative clinicians and facilitated by the staff's ability to work across traditional professional and disciplinary boundaries, working around setbacks and task issues. Good uptake of the innovation by patients was viewed as critical to maintaining staff motivation:

> *The clinicians themselves are very motivated to changing how they operate and engage with new digital tools, so they're highly motivated to do that and they know their domain very, very well.* [Technical expert or lead 1]

Efforts were made to select and target the right group of patients considering condition type, demographic characteristics, psychological factors, and lifestyle. They initially expected the system to work better for some patients (returning patients, working people, and those needing continuous low-level input) and less for others (those digitally excluded—no internet or email—and with preference or need for human contact). This belief persisted over time, although subsequently, the notion that the system functions best for medical conditions for which treatment is highly standardized arose:

> *[A clinical lead] felt was suitable for, as the first try at this and she also thought she'd have a high uptake because of the demographic, because of the age range and the working nature of the folk with the condition*

XSL•FO
RenderX

she felt that they were ones who were likely to not want to come into clinic. [Technical expert or lead 1]

I do not see why in a properly selected group of patients or conditions it could not be adopted in other areas. Like, it would be really good for patients with endocrine problems because the majority of their management is pretty standardised. [Clinical lead 1]

Patients' understanding was considered critical to managing patients' expectations and, subsequently, a good uptake. For example, patients were observed to refuse digital appointments when contacted ad hoc due to a lack of previous knowledge of the purpose of the call. The term "appointment" was viewed as a misnomer causing a lack of understanding among the patients, who viewed the term "appointment" as an arrangement to meet someone at a particular time and date. Concerns were also expressed that patients may rely only on information from an invitation letter (rather than an information booklet) or may not have understood all the details (eg, what information exchanged during the consultation is saved).

Staff quickly learned how to effectively communicate the new service to patients to improve their acceptance of it. At the time of conducting the interviews, administrative leads expressed their plans to further revise the information leaflets and update the external website. Patients' preferences affected initial product specification through administrative staff's experiences with patients' preferences (indirectly) and exploring patient opinions earlier in the new pathways' specification process (directly), although some staff members felt that greater patient input might have been beneficial:

We've done a lot of work in what the patient information leaflets were like, letters were like, when we're phoning the patients how we're explaining it to them. I think feedback from when patients did onboard, feedback on the reasons why they didn't onboard. We took that into consideration and changed our information leaflets to change...simple things such as changing the wording. [Service or administrative lead 3]

Assimilation by the System

The adult dermatology system went live in May 2020 relatively smoothly, attributable to the project being driven centrally as a part of the national pilot project, thus having more momentum and being ready to go live by the time the pandemic hit. A major issue was observed during the initial implementation phase when adopters noted and worked around an initially significantly high *did not attend* (DNA) rate. It was not until February to March 2022 that uptake rates increased, which was attributed to appointing an administrative operational support that "pushed" and monitored the system to ensure that patients identified for digital appointments had been offered them and were supported where required, improved clinical buy-in (with more dermatologists offering the digital appointments), and service management supporting and driving digital appointment use within the service:

I think they could all relate to this story of it being given a task and trying to implement it by a date and not hitting that date, and then feeling that you're constantly trying to push something up a hill just to try and get it done. I don't think this is necessarily exclusive to this particular bit of work; it just happened to be the focus of the evaluation for us. [Project management lead or expert 2]

There were several setbacks to the implementation of the gastroenterology and pain management pathways, leading to delays in the system going live: first, a 4-month delay in procurement (attributed to suboptimal coordination of efforts between the innovation and transformation teams); then, a 6-month delay in technical integration caused by capacity issues (loss of technical personnel by the NHS Grampian eHealth team and availability of the external technical resource [National Services Scotland (NSS)] to undertake the development); and, finally, a 1-month delay due to technical issues (eg, information governance):

I think NHS Grampian didn't act upon the emergency procurement process rapidly enough for us to get the window with NSS [National Service Scotland] when they had the capacity to do it, as I look back. [Technical expert or lead 1]

It's really been the bottleneck at NSS that's caused a lot of delays, and then also some personnel changes at NSS which have resulted in the people who've got knowledge of how that integration works moving on to new positions. [Technical expert or lead 2]

Over time, the reality of these setbacks caused staff to grow tired and frustrated, with some losing their confidence in this specific innovation. Contrary to expectations, the groundwork made through the adoption of the dermatology asynchronous consultation system was not easily transferable to the pain management and gastroenterology systems; despite involving the same developer and product, "the reality turned into that it was an absolute uphill battle to even get it live" (project management lead or expert 2):

I think the area that frustrated us more was the integration side of things because we already had integration up and running with dermatology and from our perspective it should've been a relatively simple process. [Technical expert or lead 2]

Overall, almost all adopters described feeling surprised by the complexity and nonlinear nature of the process and implementation effort involved. Initially, this provoked reflections on the timing of technical integration, sequencing of steps, timing of stakeholder involvement, and lack of guidance on core and optional steps. After the implementation, this prompted recognition of the "need to have those champions and people that are pushing for it and really taking it through" (project management lead or expert 2):

If you integrate it too early, you might end up with something that won't scale very well. If you integrate it too late, you might end up with lots of messy ways of working with lots of different workarounds that people don't want to change. I think there is definitely a sweet point of when's the right sweet spot to

XSL•FO
RenderX

integrate and I don't think we got that right with this one. [Technical expert or lead 1]

Implementation and Routinization

While decision-making was devolved locally to pain management and gastroenterology clinical leads (whereas, in practice, dermatology was centrally led as part of a national pilot project), its adopters expressed regrets over missed opportunities to make implementation smoother by taking administrative staff's concerns more seriously early on, especially in terms of system readiness and how to improve patient uptake:

> *Involvement of the stakeholders, as I say, the secretarial staff would have been absolutely brilliant at pointing out where the omissions might happen, or kind of documentation issues might happen that we would never foresee.* [Clinical lead 4]

The organization and delivery of the innovation were human resource intensive, requiring a high level of staff buy-in and administrative input. Allocated time was perceived as needed for clinicians, operational staff, clinic coordinators, project managers, evaluation leads, and eHealth teams. The appointment of administrative operational support was perceived as directly responsible for an increase in uptake. A need for continued funding and dedicated resources was stressed, but this was against extreme pressure on NHS resources generally:

> *If it was funded appropriately that would have maybe improved things as in funding for NHS staff and for NSS involvement. There was very little funding available within the NHS.* [Technical expert or lead 2]

> *There was a new appointment made a number of months ago, which has...actually, you can see it's paid dividends now in terms of the take-up of the figures in terms of the digital appointments as well, but that's kind of an admin support.* [Project management lead or expert 1]

Delays with technical integration prompted reflections on the role of the NSS integration hub as a national-level NHS stakeholder. The NSS appeared to have limited capacity for this project. The innovation was intended as a proof of concept rather than a fully operational system, so it needed more time commitment and a clear description of roles and contractual agreements on escalation routes. The delays in technical integration caused the technical leads to wonder whether an IT system at the proof-of-concept stage should be fully integrated with the host organization's IT system immediately or only after it has undergone functional testing.

Incompatibilities with the current ways of working did not surface until after implementation, leaving adopters unsure of whether it "translates practically into the real world" (service or administrative lead 1). The main issue was the so-called "lifetime of an appointment" in the patient booking system. While in-person appointments are acted upon the same day of the appointment, asynchronous consultations would appear still booked, which was "kind of alien to [admin staff]" (service or administrative lead 2). Another issue concerned how clinicians plan their time, imposing limits on how flexibly the new system could be used (ie, managed by allocating time slots 2 afternoons a week or ad hoc but ultimately depending on patients' timing). At the time the interviews were conducted, a solution for this issue had not been established:

> *I think the notion behind the asynchronous appointments or assessments, or the process is good. I don't know that it translates practically into the real world if that makes sense. So, I think the idea behind it around you don't need to have dedicated clinical sessions, you can spread these appointments out throughout the week and the clinicians can deal with them at any time, that doesn't fit with the world we live in in terms of job planning and service planning and things.* [Service or administrative lead 1]

From early on, there was a perceived need for better 2-way communication between the project manager and adopters, including relevant people from the start (importantly, administrative staff) and keeping them informed throughout, keeping records of communication, and clearly outlining roles (knowing who to contact). Concerns about a perceived lack of a formal project board were aired—essential for acting on staff feedback brought up by a project manager. Transition or loss of personnel, especially project manager changes, significantly disturbed stakeholders' involvement and communication continuity. Later, a new acting project manager addressed this by ensuring a 2-way communication line with the operational staff, which they felt helped restore trust.

Perceived Consequences

Initial low confidence in the success of the dermatology system improved with time following "quite a dramatic increase in Dermatology take-up in terms of the number of appointments that have gone through" (project management lead or expert 1).

As the new pathways went live, it was felt that some of the new pathways were better accepted by patients than others and that the new system would be more successful overall if it fit better with the existing process flow. For example, it was considered to work well for pain management because patients were scheduled in a sequence (when they reached the top of the list as opposed to being chosen by hand) and clinicians introduced patients to the service during their initial appointment:

> *Pain Management it works very well with them because they already used a similar system. The patients are being onboarded by the clinicians at the time, or the care providers at the time of their initial appointment so it's been explained to them what it is and how it works.* [Service or administrative lead 3]

Overall, the new system was viewed as useful to many patients, but "there needs to be an alternative option because it's not for everybody" (clinical lead 3). As for the implementation success, participants seemed to agree that "in the face of all the adversity and challenges that the work has faced, the fact that that still is getting used and patients are still buying into shows that, I guess, as an attempt to introduce this and upscale it, it can be done" (project management lead or expert 2)

XSL•FO
RenderX

Assessment of Service Uptake by Patients

Assessment of Routinely Collected Data

The adult dermatology service was the first in NHS Grampian to offer digital asynchronous appointments to their patients in May 2020, followed by the pediatric dermatology service in January 2021. The gastroenterology and pain management departments started offering digital appointments in December 2021 (10 months later than originally planned).

The responses of patients offered digital appointments were based on an audit carried out by administrative staff between February and July 2022 (n=832). A total of 41% (341/832) of patients accepted and received digital appointments, only 16.9% (141/832) explicitly declined, and 31% (258/832) did not respond to the clinic letter offers or believed that they no longer required a consultation. The remaining 11.1% (92/832) comprised people whom services had already seen, people who had already booked a different type of consultation, and a small number of people classified as *other*.

There were a total of 1709 digital appointments offered to 1417 patients across the 4 medical specialties, between May 2020 and September 2022. The total outcomes (n=2416) were greater than the number of appointments (N=1709) because there could be more than one outcome per consultation. The overall mean age of service users (N=1709) was 38.7 (SD 22.7) years, ranging from 1 to 90 years. A total of 55.59% (950/1709) of patients were women, and 79.4% (1357/1709) were White. Moreover, 4.62% (79/1709) of people who chose to use this service were from our most deprived communities, whereas the figure in the Grampian region overall is 5.36% (31,433/586,530; Figure S1 in Multimedia Appendix 2). This discovery is important as both public and staff groups were concerned with assessing whether the system could potentially worsen health inequalities, a consideration that would be a major disadvantage.

DNA rates were high in the first 2 months that digital appointments were used in the pain management (7/21, 33%) and gastroenterology (17/38, 45%) services. On the other hand, once the service was established the overall DNA rate (between January and September 2022) was 13.7% (125/909), which is approximately 5% higher than the national face-to-face appointment DNA rates for gastroenterology and adult dermatology. There were apparent differences across medical specialties and time (Table 1), yet there was no apparent impact based on the length of time the specialty had been offering the new service. This would be evident if DNA rates in the dermatology pathways were generally lower.

Table 1. Did not attend rates across medical specialties using digital appointments in the first 3 quarters of 2022.

Medical specialties	Total appointments, n	January to March, n/N (%)	April to June, n/N (%)	July to September, n/N (%)
Gastroenterology	234	7/45 (15.6)	3/70 (4.3)	17/119 (14.3)
Pain management	83	3/16 (18.8)	4/23 (17.4)	8/44 (18.2)
Adult dermatology	385	16/132 (12.1)	24/131 (18.3)	23/122 (18.9)
Pediatric dermatology	207	5/36 (13.9)	13/135 (9.6)	2/36 (5.6)

Looking at the 4 medical specialties overall, there was no significant difference in DNA rates across the 5 SIMD deprivation quintiles (χ^2_4=7.8; *P*=.10). The actual DNA rates were as follows: 25.5% (79/310) for SIMD 1 (most deprived); 23.3% (82/352) for SIMD 2; 23.9% (78/327) for SIMD 3; 20.7% (78/377) for SIMD 4; and 17.3% (58/335) for SIMD 5 (least deprived). SIMD could not be calculated for 8 cases.

The most common outcome from appointments overall was open returns (ie, that a patient can request another appointment if and when they feel they require one; 1208/2416, 50%), followed by treatment (894/2416, 37%); discharged (242/2416, 10.02%); and, finally, referred on (72/2416, 2.98%). There was no evidence that deprivation levels impacted on whether people were offered treatment following their appointments, with proportions similar across quintiles (χ^2_4=2.7; *P*=.61). The numbers offered treatment were as follows: SMID 1 (177/894, 19.8%), SIMD 2 (166/894, 18.6%), SIMD 3 (182/894, 20.3%), SIMD 4 (193/894, 21.6%), and SIMD 5 (176/894 19.7%).

In terms of the estimated environmental impact, assuming people would travel from their home address (area spread across 3000 square miles) to where the Aberdeen Royal Infirmary is located, digital appointments resulted in 44,712 fewer miles (approximately 71,956.81 km) traveled than traditional face-to-face approaches.

Patient Experience

Of the 317 patients who attended digital appointments at the gastroenterology and pain management clinics, all of whom were offered to complete the web-based satisfaction survey, only 66 (20.8%) patients completed it. The vast majority of respondents (57/66, 86%) were patients attending the pain management service.

Approximately 1 in 5 of patients were unhappy or very unhappy to be offered a digital appointment (12/66, 19%) and a similar proportion continued to feel the same way about using this type of appointment after they had experienced it (14/66, 21%; Table 2).

XSL•FO
RenderX

Table 2. Results of the satisfaction survey with a digital appointment from those who completed their appointment (n=66).

Survey questions	Very happy, n (%)	Happy, n (%)	Neutral, n (%)	Unhappy, n (%)	Very unhappy, n (%)
How did you feel about being offered a digital appointment?	11 (17)	22 (33)	21 (32)	7 (11)	5 (8)
How do you feel about using digital appointments for this type of health care service now that you have tried it?	8 (12)	26 (39)	18 (27)	8 (12)	6 (9)

The vast majority (59/66, 89%) of patients believed to a large or very large extent it is important to be involved in decisions about their care; however, much fewer felt to a large or very large extent that they were involved in their digital appointments (23/66, 35%; Table 3). Two-thirds (44/66, 67%) rated the quality of their care as excellent or good.

Table 3. Results of the satisfaction survey with a digital appointment from those who completed their appointment (n=66).

Survey questions	To a very large extent, n (%)	To a large extent, n (%)	Moderately, n (%)	To some extent, n (%)	Not at all, n (%)
How important is it to you to be involved in decisions about your health care?	47 (71)	12 (18)	5 (8)	1 (2)	1 (2)
How involved did you feel in the outcome of your digital appointments?	6 (9)	17 (26)	19 (29)	10 (15)	14 (21)

A total of 10% (6/66) of the survey respondents agreed to follow-up interviews. Follow-up interviews suggested that people saw a role for it in the NHS, viewing it as an additional, optional method alongside other forms of consultation, consistent with public opinions. Key specific benefits experienced included the time saved on travel and improved information collection and sharing, which both the public and staff anticipated:

> *I think that there are huge amounts of benefits in having this kind of system, but somewhere in it you have to include the ability to have a face-to-face.* [P3; male]

> *It saves you having appointments with Tom, Dick, and Harry where you can have it with Harry, and he can pass the information on to Tom and Dick.* [P5; female]

Some felt apprehensive about the offered service, fearing that it might eventually replace human connections, and overall preferred in-person communication. This sentiment is in line with the distinct preference for this type of contact expressed by some public focus group participants. They believed that teleconsultations distanced patients and health care professionals, emphasizing the significance of being physically present in creating a trusting physician-patient connection. Losing meaning when writing or even suppressing their answers were also mentioned. However, they acknowledged that the written format offers advantages, such as an opportunity to review a response and writing in one's own words in a free-text box:

> *You're being offered primarily video sessions and telephone sessions from your surgery. There has been a barrier created between the medical profession and Joe punter [a user].* [P3; male]

> *I think if you've got to write things, yeah, that looks bad if I'm saying, "No, this isn't what I want."* [P2; female]

> *It's like a survey, you can ask certain questions but there may not actually be a response available that matches what you want to say. There are benefits to it because you have time to think through...so if for example, it's a free text box, you have the opportunity to construct what you're saying and then review your response.* [P2; female]

Those who were happy to be offered the service, on the other hand, stated that rushed in-person consultations lack the human factor anyway and saw digital appointments as an opportunity to improve relationships because physicians have a better understanding of their needs. They liked having time to think about and reconsider their comments; they thought that they covered things quite well and described the questions set as straightforward, straight to the point, and more personal; and saw it as an unavoidable change:

> *I had time to sit down and actually read the question and properly kind of, not prepare answers, but kind of rethink what my answers would be. Honestly, I think that would probably better the relationship because the doctors have a better understanding of the answers for the questions that they gave.* [P6; female]

> *I think these people have got to get up-to-date with the times...if that's how it's going to be, you're going to have to change with the times or get someone to help you to change with the times.* [P4; male]

A total of 4 types of influences on the ease of use of the new service were mentioned: difficulty receiving timely care, condition, ability, and technical issues. More specifically, some patients had previously had difficulty receiving timely care. For some, this resulted in a preference for in-person contact, whereas for others, it enhanced appreciation for having faster access:

XSL•FO
RenderX

> *You've got to jump through so many hoops...By the time you get to say something, you've become incoherent as well.* [P2; female]
>
> *I was really looking forward to it, and I believed it would be useful.* [P1; female]

In total, 25% (1/4) of the interviewed participants who completed their appointments did so without assistance and felt the direction provided (an email with instructions) to be extremely clear. The remaining 75% (3/4) of the participants expressed doubt in their ability to complete the appointment without assistance, so "to be sure that [they] knew what to do" (P2; female), they sought assistance from a family member, carer, or employee. Consistent with the public views, patients thought that those with limited digital literacy and certain cognitive impairments and those who did not speak English as their first language would be least likely to use this service.

Some people's ability to go to in-person appointments, remain seated, and even respond to the questionnaire was reported to be affected by their condition. In line with the public's perception of possible disadvantages, some people found it difficult to convey their health status, particularly pain, in words, necessitating physical contact with a clinician:

> *Because I'm in pain, I mean, I can't sit in the same place. I'm afraid you just stop listening.* [P1; female]
>
> *When I filled in that, I could quite well be contradicting myself by what I'm saying now. You're not well, things change.* [P2; female]

In total, 2 participants reported experiencing technical issues (the link to the appointment not working, poor broadband connection, and a "device that refused to work" [P3; male]). One participant thought that was normal and resulted from "a lack of experience on both sides" (P3; male). Another participant felt confused and "not being taken seriously" (P2; female). Those patients sought assistance from the clinic, which eventually provided options for in-person and telephone consultations. This approach is in line with public opinions on the necessity of providing alternative forms of consultation and aligns with staff reports on actions taken to sidestep technical issues.

Consistent with staff views, we noted that the primary area for improvement was participants' comprehension of the service. All but 1 patient reported having trouble memorizing and naming specific phases in their treatment pathway, indicating difficulties with eHealth terminology and the complexity of the pathway. Patients did not consider the *digital appointment* as an appointment but rather as an exchange of information or assessment to determine "whether [they] actually need an appointment" (P2; female). Most respondents who completed the consultations did not recall receiving a consultation copy, which indicates that this information was not effectively communicated. In addition, there was no clear connection perceived between the outcome of the consultation and the responses made, indicating a lack of knowledge of how information obtained from consultations influenced treatment decisions.

Discussion

Principal Findings

A range of patients decided to use the asynchronous consultation service of the dermatology, gastroenterology, and pain management departments, with a total of 1709 appointments offered to 1417 patients from May 2020 to September 2022 across 4 medical specialties. The public was overall receptive to the service, and 82% (54/66) of real users felt neutral to very happy about being offered the service. Improved access, ease, information collection and sharing, and more physician flexibility were among the benefits highlighted. The main concerns expressed by both the public and staff was the fear of digital exclusion, potentially worsening health inequalities, and the loss of human contact. However, NHS data showed no evidence that people from the most deprived areas were less likely to accept digital appointments, receive treatment, or be given open return appointments. Regarding human contact, some patients preferred direct interaction, whereas others noted that face-to-face encounters did not always enhance connection and found written communication to be more effective for clinician understanding. Recognizing that it may not work for certain individuals, especially those with limited digital resources or writing abilities, and that other people would prefer in-person interaction, informed choice and flexibility are required. The administrative processes could be improved, starting with better communication to promote patient acceptability by changing the term *digital appointment* to *assessment*.

The adopters (staff) expected a relatively straightforward technical system transfer from dermatology to other receptive specialties, with generally positive support from the clinical and administrative staff tasked with implementing it. Despite a favorable context and organizational support, the reality proved immensely complex. Nonetheless, staff understanding of the asynchronous system, how to describe it to patients, and how to use it to the best effect evolved rapidly. They perceived variances in adoption among specialties, with pain management being regarded to function better (easier to use) as the new approach suited present practices, but numerical data did not show that the approach functioned any differently for pain management clinics. The sustainability of this new system was seen as linked to its transferability across many specialties and the availability of funds to afford allocated continuous time for clinicians, operational staff, clinic coordinators, project managers, evaluation leads, and eHealth teams.

Limitations

There was a substantial unanticipated delay in implementing the system, which is a finding in itself. The NHS is a system under constant pressure with limited support for innovation available, so the risk of delays seems difficult to eliminate. We used this chance to capture and report on implementation challenges, and the NHS Grampian partner's commitment to transparency in reporting the implementation process made this feasible. However, the implementation delay meant that there was less time than we had hoped to recruit patients to interview who had completed the appointment process, and as a result,

XSL•FO
RenderX

our sample of patient interviews was smaller than we had planned. The number of people who were able to complete the NHS satisfaction survey was also small and limited to the gastroenterology and pain management clinics (due to permission restrictions). The NHS Grampian data have limitations as well—we were unable to collect any data on patient outcomes; however, our focus was on assessing the implementation approach across specialties rather than on specific patient outcomes within individual pathways.

Comparison With Prior Work

In accordance with a global agenda [20], the UK Department of Health and Social Care [21] and Scotland's Digital Health and Care Strategy [22] aspire to expand the range of digital clinical and care services and ensure that staff can work remotely and flexibly. Asynchronous patient-to-physician communication offers that. Previous studies have revealed that teleconsultations are generally acceptable to NHS patients [23]. We found that the public is receptive to NHS outpatient asynchronous consultations if individual preferences are respected, the system is suited to end users' needs, and the service is optional. Most outpatient asynchronous consultations involve submitting a response to a set of key questions about a condition and uploading photos or videos (eg, for dermatology [9] or epileptic seizures [11]), but overall, asynchronous consultations have a variety of uses across countries [6]. An asynchronous consultation service for pain management and gastroenterology care involving a set of key questions only is a viable and acceptable option.

Consistent with what we know already from the wider implementation literature [18], future implementation efforts regarding digital innovations should consider assessing ahead of implementation key organization-wide factors identified by our staff participants as desirable, including a degree of consensus on broader innovation priorities and protocols, clearly defined roles and responsibilities of innovation and transformation teams, adequate resources and involvement of stakeholders, and strong and continuous project management. Further research is needed to help organizations build and implement evidence-informed strategies to prioritize and manage innovation.

The nonlinear and complex nature of the implementation process has been previously reported [18]. This study resonates with that and further suggests that it can be expected even when the same product is scaled up in the same setting. However, the reality of staff's frustrations with the complexity of that process is less documented. It is widely recognized that scaling innovation takes space, time, and resources, so organizations need to consciously and strategically drive scaling efforts [24]. We noted the critical importance of numerous influences on clinicians' motivation to adopt and continue engaging with the innovation (eg, uptake by patients, competing priorities and time pressures, exposure to the innovation work, and continuous 2-way communication with a project manager). In the primary care setting, the importance of compatibility with the existing workflow and patient satisfaction with asynchronous consultations was also noted [6]. It is crucial to include administrative staff issues early on as they may anticipate patient ease of use and compatibility with the existing workflow.

The importance of the involvement of end users is reported in the eHealth literature in general [25] and in chronic disease management [26] specifically, emphasizing the need for user-centered design. The importance of providing user-friendly, role-appropriate information and resources to support the individual being cared for [27] and patient involvement in digital service design from the outset have been noted [6,28]. This can be done with the help of one of many frameworks for gathering insights from patients and the public [29]. This study echoes this and emphasizes the need to understand and tackle digital exclusion and provide informed digital choice via system cospecification and effective communication regarding outpatient asynchronous consultations.

The current body of research points to a rising interest in patient safety within telemedicine, although there is a lack of detailed studies on this topic [30]. Consistent with this trend, this perspective warrants attention in this research. In this study, the team responsible for implementation assessed and monitored the new service, worked on their communication with patients, ensured data security and system resilience, and maintained follow-up protocols, which are all key elements of patient safety [31]. We reported issues concerning whether patients fully comprehend the nature and content of the communication process, which is vital for patient safety [31]. Overall, this study supports the relevance of establishing technical standards and guidelines to guarantee safety and quality in asynchronous consultations as a form of telemedicine.

Conclusions

This research shows the viability and great potential of asynchronous consultations and wider digital solutions in the NHS to be a key part of meeting increased demands on the NHS. Recognizing that it may not work for everyone, flexibility and informed choice are key. For potential patients, careful technical support and explanation are needed, as well as a choice of consultation routes, to ensure digital inclusion. These findings also highlight how essential effective patient and public involvement is for the success of any digital technology developments in the NHS, especially as the use of digital technology in health continues to rapidly advance. Our results on implementation effort complexity, the delays that occurred, the characteristics of the innovation, and its reception by hospital staff may help staff deploy and sustain asynchronous consultations.

Acknowledgments

Funding was provided by the Health Foundation (project AIMS 2238252). The funders had no role in the study conduct or interpretation of the results. The authors would like to thank Kyle MacKenzie for his help with participant recruitment and data

XSL•FO
RenderX

interpretation. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author accepted manuscript version arising from this submission.

Data Availability

All relevant data can be found within the paper and Multimedia Appendix 2.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study tools (topic guides for public focus groups, staff and patient interviews, along with the patient satisfaction survey).
[PDF File (Adobe PDF File), 150 KB - jmir_v26i1e48092_app1.pdf]

Multimedia Appendix 2
Supplementary data.
[PDF File (Adobe PDF File), 409 KB - jmir_v26i1e48092_app2.pdf]

References

1. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
2. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021;7:20552076211009396 [FREE Full text] [doi: 10.1177/20552076211009396] [Medline: 33959378]
3. Horton T, Hardie T, Mahadeva S, Warburton W. Securing a positive health care technology legacy from COVID-19. The Health Foundation. 2021 Mar 16. URL: https://www.health.org.uk/publications/long-reads/securing-a-positive-health-care-technology-legacy-from-covid-19 [accessed 2022-08-28]
4. Dixon RF. Enhancing primary care through online communication. Health Aff (Millwood) 2010 Jul;29(7):1364-1369. [doi: 10.1377/hlthaff.2010.0110] [Medline: 20606189]
5. Bishop TF, Press MJ, Mendelsohn JL, Casalino LP. Electronic communication improves access, but barriers to its widespread adoption remain. Health Aff (Millwood) 2013 Aug;32(8):1361-1367 [FREE Full text] [doi: 10.1377/hlthaff.2012.1151] [Medline: 23918479]
6. Fuster-Casanovas A, Vidal-Alaball J. Asynchronous remote communication as a tool for care management in primary care: a rapid review of the literature. Int J Integr Care 2022;22(3):7 [FREE Full text] [doi: 10.5334/ijic.6489] [Medline: 36043028]
7. Mold F, Hendy J, Lai YL, de Lusignan S. Electronic consultation in primary care between providers and patients: systematic review. JMIR Med Inform 2019 Dec 03;7(4):e13042 [FREE Full text] [doi: 10.2196/13042] [Medline: 31793888]
8. Wherton J, Greenhalgh T, Shaw SE. Expanding video consultation services at pace and scale in Scotland during the COVID-19 pandemic: national mixed methods case study. J Med Internet Res 2021 Oct 07;23(10):e31374 [FREE Full text] [doi: 10.2196/31374] [Medline: 34516389]
9. Jiang SW, Flynn MS, Kwock JT, Nicholas MW. Store-and-forward images in teledermatology: narrative literature review. JMIR Dermatol 2022;5(3):e37517 [FREE Full text] [doi: 10.2196/37517] [Medline: 35891983]
10. Wherton J, Greenhalgh T. Evaluation of the Near Me video consulting service in Scotland during COVID-19, 2020. The Scottish Government. URL: https://www.gov.scot/binaries/content/documents/govscot/publications/research-and-analysis/2021/03/evaluation-near-video-consulting-service-scotland-during-covid-19-2020-main-report/documents/evaluation-near-video-consulting-service-scotland-during-covid-19-2020/evaluation-near-video-consulting-service-scotland-during-covid-19-2020/govscot%3Adocument/evaluation-near-video-consulting-service-scotland-during-covid-19-2020.pdf [accessed 2022-09-03]
11. SHTG assessment. Scottish Health Technologies. URL: https://tinyurl.com/329v573d [accessed 2022-05-15]
12. Nguyen OT, Alishahi Tabriz A, Huo J, Hanna K, Shea CM, Turner K. Impact of asynchronous electronic communication-based visits on clinical outcomes and health care delivery: systematic review. J Med Internet Res 2021 May 05;23(5):e27531 [FREE Full text] [doi: 10.2196/27531] [Medline: 33843592]
13. Nittari G, Savva D, Tomassoni D, Tayebati SK, Amenta F. Telemedicine in the COVID-19 era: a narrative review based on current evidence. Int J Environ Res Public Health 2022 Apr 22;19(9):5101 [FREE Full text] [doi: 10.3390/ijerph19095101] [Medline: 35564494]
14. Atherton H, Ziebland S. What do we need to consider when planning, implementing and researching the use of alternatives to face-to-face consultations in primary healthcare? Digit Health 2016;2:2055207616675559 [FREE Full text] [doi: 10.1177/2055207616675559] [Medline: 29942570]

XSL•FO
RenderX

15. Turner A, Morris R, Rakhra D, Stevenson F, McDonagh L, Hamilton F, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2022 Feb;72(715):e128-e137 [FREE Full text] [doi: 10.3399/BJGP.2021.0426] [Medline: 34903520]
16. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
17. Chan S, Li L, Torous J, Gratzer D, Yellowlees PM. Review of use of asynchronous technologies incorporated in mental health care. Curr Psychiatry Rep 2018 Aug 28;20(10):85. [doi: 10.1007/s11920-018-0954-3] [Medline: 30155593]
18. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629 [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
19. Population estimates by Scottish Index of Multiple Deprivation (SIMD). National Records of Scotland. URL: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/2011-based-special-area-population-estimates/population-estimates-by-simd-2016 [accessed 2022-04-08]
20. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2023-01-17]
21. A plan for digital health and social care. Department of Health and Social Care, United Kingdom Government. 2022 Jun 29. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2023-02-01]
22. Digital health and care strategy. Scottish Government. 2021 Oct 27. URL: https://www.gov.scot/publications/scotlands-digital-health-care-strategy/ [accessed 2022-09-02]
23. O'Cathail M, Sivanandan MA, Diver C, Patel P, Christian J. The use of patient-facing teleconsultations in the national health service: scoping review. JMIR Med Inform 2020 Mar 16;8(3):e15380 [FREE Full text] [doi: 10.2196/15380] [Medline: 32175911]
24. Albury D, Beresford T, Dew T, Horton T, Illingworth J, Langford K. Against the odds: successfully scaling innovation in the NHS. The Health Foundation. 2018 Jan. URL: https://www.health.org.uk/publications/against-the-odds-successfully-scaling-innovation-in-the-nhs [accessed 2022-12-08]
25. Ahmed B, Dannhauser T, Philip N. A systematic review of reviews to identify key research opportunities within the field of eHealth implementation. J Telemed Telecare 2019 Jun;25(5):276-285. [doi: 10.1177/1357633X18768601] [Medline: 29703127]
26. Taylor ML, Thomas EE, Vitangcol K, Marx W, Campbell KL, Caffery LJ, et al. Digital health experiences reported in chronic disease management: an umbrella review of qualitative studies. J Telemed Telecare 2022 Dec;28(10):705-717. [doi: 10.1177/1357633X221119620] [Medline: 36346938]
27. Enabling, connecting and empowering: care in the digital age. Scottish Government and COSLA. URL: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2021/10/scotlands-digital-health-care-strategy/documents/enabling-connecting-empowering-care-digital-age/enabling-connecting-empowering-care-digital-age/govscot%3Adocument/enabling-connecting-empowering-care-digital-age.pdf [accessed 2023-01-12]
28. Baines R, Bradwell H, Edwards K, Stevens S, Prime S, Tredinnick-Rowe J, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022 Aug;25(4):1232-1245 [FREE Full text] [doi: 10.1111/hex.13506] [Medline: 35526274]
29. Bird M, McGillion M, Chambers EM, Dix J, Fajardo CJ, Gilmour M, et al. A generative co-design framework for healthcare innovation: development and application of an end-user engagement framework. Res Involv Engagem 2021 Mar 01;7(1):12 [FREE Full text] [doi: 10.1186/s40900-021-00252-7] [Medline: 33648588]
30. Guise V, Anderson J, Wiig S. Patient safety risks associated with telecare: a systematic review and narrative synthesis of the literature. BMC Health Serv Res 2014 Nov 25;14:588 [FREE Full text] [doi: 10.1186/s12913-014-0588-z] [Medline: 25421823]
31. de Micco F, Fineschi V, Banfi G, Frati P, Oliva A, Travaini GV, et al. From COVID-19 pandemic to patient safety: a new "spring" for telemedicine or a boomerang effect? Front Med (Lausanne) 2022;9:901788 [FREE Full text] [doi: 10.3389/fmed.2022.901788] [Medline: 35783642]

Abbreviations

DNA: did not attend
NHS: National Health Service
NSS: National Services Scotland
SIMD: Scottish Index of Multiple Deprivation

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.04.23; peer-reviewed by S Hinder, H Leach; comments to author 24.10.23; revised version received 15.12.23; accepted 12.03.24; published 04.06.24.

Please cite as:
Rzewuska Díaz M, Locock L, Keen A, Melvin M, Myhill A, Ramsay C
Implementation of a Web-Based Outpatient Asynchronous Consultation Service: Mixed Methods Study
J Med Internet Res 2024;26:e48092
URL: https://www.jmir.org/2024/1/e48092
doi:10.2196/48092
PMID:38833695

©Magdalena Rzewuska Díaz, Louise Locock, Andrew Keen, Mike Melvin, Anthony Myhill, Craig Ramsay. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Six-Month Outcomes of a Theory- and Technology-Enhanced Physical Activity Intervention for Latina Women (Pasos Hacia La Salud II): Randomized Controlled Trial

Lauren Connell Bohlen[1], PhD; Shira I Dunsiger[1], PhD; Tayla von Ash[1], ScD; Britta A Larsen[2], PhD; Dori Pekmezi[3], PhD; Becky Marquez[2], MPH, PhD; Tanya J Benitez[1], MSW, PhD; Andrea Mendoza-Vasconez[1], PhD; Sheri J Hartman[4], PhD; David M Williams[1], PhD; Bess H Marcus[1], PhD

[1]Center for Health Promotion and Health Equity, Department of Behavioral and Social Sciences, Brown University School of Public Health, Providence, RI, United States

[2]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, United States

[3]Department of Health Behavior, The University of Alabama at Birmingham School of Public Health, Birmingham, AL, United States

[4]Herbert Wertheim School of Public Health and Human Longevity Science, Moores Cancer Center, University of California San Diego, La Jolla, CA, United States

Corresponding Author:
Lauren Connell Bohlen, PhD
Center for Health Promotion and Health Equity
Department of Behavioral and Social Sciences
Brown University School of Public Health
121 South Main Street
Box G-S121-8
Providence, RI, 02912
United States
Phone: 1 4018636559
Email: lauren_bohlen@brown.edu

Abstract

Background: More than half (55%) of Latina women do not meet aerobic physical activity (PA) guidelines, and frequently cite time, childcare, and transportation as barriers to PA. In addition to linguistic adaptations for this population, successful PA interventions for Latina women addressed these barriers through remote intervention delivery approaches (eg, mail, phone, or web delivery).

Objective: We aimed to evaluate 6-month outcomes of a randomized trial comparing a Spanish-language, individually tailored, web-delivered PA intervention (original) to an enhanced version with text messages and additional features (enhanced). Further, we evaluated if increases in PA at 6 months were moderated by baseline activity status.

Methods: In total, 195 Latina women aged 18-65 years participated in a trial comparing the efficacy of the enhanced versus original interventions at initiating PA behavior change. We examined minutes per week of accelerometer-measured PA in the enhanced versus original arms, and the proportion of each arm meeting aerobic PA guidelines (150 min/wk at 6 mo). For moderator analyses, participants were classified as inactive (0 min/wk) or low active (1-90 min/wk) at baseline, measured via the 7 Day Physical Activity Recall interview.

Results: PA increased from 19.7 (SD 47.9) minutes per week at baseline to 46.9 (SD 66.2) minutes per week at 6 months in the enhanced arm versus 20.6 (SD 42.7) minutes per week to 42.9 (SD 78.2) minutes per week in the original arm (P=.78). Overall, 30% (31/103) of the enhanced group met aerobic PA guidelines at 6 months, compared to 21% (19/92) of the original group (odds ratio [OR] 1.75, 95% CI 0.87-3.55). Baseline PA (inactive vs low active) moderated treatment effects on PA. For inactive participants, there were no group differences at 6 months (b=7.1; SE 22.8; P=.75), while low-active participants increased more in enhanced than original (b=72.5; SE 27.9; P=.01). For low-active participants, 45% (46/103) of the enhanced group met PA guidelines at 6 months, versus 20% (18/92) of the original arm (OR 3.29, 95% CI 1.05-11.31). For inactive participants, there were no group differences (25/103, 24% vs n=19/92, 21% for enhanced vs original, respectively; OR 1.28, 95% CI 0.54-3.06).

XSL•FO
RenderX

Conclusions: Intervention effects were conditional on baseline PA. For low-active Latina women, the enhanced intervention was more effective at increasing PA. Additional tailored intervention enhancements may be necessary to increase PA for inactive Latina women.

Trial Registration: ClinicalTrials.gov NCT03491592; https://www.clinicaltrials.gov/study/NCT03491592

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-022-06575-4

(J Med Internet Res 2024;26:e51708) doi:10.2196/51708

KEYWORDS

digital health; web-based intervention; exercise; social support; behavior change intervention; support; Latina women; women; Latina; physical activity; barrier; aerobic; remote intervention; text message; behavior change; behavior

Introduction

Rates of inactivity-related conditions (overweight, obesity, or type 2 diabetes) are both higher and growing faster among Latina women than in non-Latina White women [1,2], and it is projected that more than half of all Latina women will eventually develop type 2 diabetes or hypertension, or both [3,4]. There is robust evidence that regular physical activity (PA) can help prevent obesity, type 2 diabetes, and hypertension [5-7]. Thus, effective PA interventions that have the potential for widespread dissemination are needed for this at-risk population.

More than half (55%) of Latina women do not meet aerobic PA guidelines [8], and frequently cite barriers to PA such as childcare, transportation, and available time [9-11]. Moreover, Latina women report numerous social, environmental, economic, linguistic, and cultural factors that limit their access to PA programs and resources [12]. Successful PA interventions for Latina women have made linguistic adaptations for this population and further sought to address these barriers through remote intervention delivery approaches (eg, mail, phone, or web delivery) [13,14].

The use of technology-based tools and resources to improve access to health-related interventions shows great promise as a low-cost solution for targeting disproportionately affected populations [15]. There is near universal use of the internet among Latina women (approximately 92%), with access to the internet primarily mediated through smartphones [16]. Yet, few PA interventions use technology-based tools to conduct interventions in Latina populations. Noting the rapid growth of internet use among the Latino population [17], and building upon the success of our research team's evidence-based print intervention [18], we previously conducted formative research with Latina women to develop a theory-based and web-based PA intervention [17]. The results of our initial web-based PA intervention showed that participants achieved significant PA improvements at 6 and 12 months [14,19]; however, participants were on average still well below the national PA guidelines for aerobic activity, suggesting that a more intensive internet-based approach may be necessary to help Latina women reach and sustain health-enhancing PA levels.

As such, a series of technology and theory-supported adaptations were made to the original intervention, including optimization of the website for mobile viewing, text messages to increase engagement with the website, and further targeting the intervention toward changing key social cognitive theory [20] constructs (eg, self-efficacy, enjoyment, and social support).

This study builds upon our previous intervention and formative research with Latina women to test the original individually tailored, Spanish-language, web-based Pasos Hacia La Salud Intervention (original intervention) to a text-message and theory-enhanced version (enhanced intervention) [14,19]. This paper aims to evaluate changes in moderate to vigorous physical activity (MVPA) in the original versus enhanced conditions from baseline to 6 months, as described in this study's protocol paper [21]. A secondary aim of this paper is to explore differences in activity increases in individuals in the contemplation stage versus the action transtheoretical model (TTM) [22] stage of change at baseline; that is, to compare individuals who were doing no activity (0 min/wk) to those doing some amount of activity (>0 min/wk), to better understand who is best aided by the enhancements.

Methods

Study Design

This 6-month randomized controlled trial tested the relative efficacy of an enhanced versus original evidence- and theory-based, Spanish-language PA intervention website on changes in MVPA at the end of the active intervention period (6 mo). The study was conducted at Brown University and registered as a clinical trial (NCT03491592).

Ethical Considerations

All study procedures were approved by the Brown University Institutional Review Board (IRB#: 00000556) and described in detail in a prior report [21]. All participants provided written informed consent.

Participants

Participants included those who self-identified as female and Latina. For inclusion, participants were required to be aged 18-65 years, and engaging in less than 60 minutes per week of self-reported leisure-time MVPA. To be eligible, participants needed to score more than 16 on the Short Test of Functional Health Literacy in Adults, to determine whether participants have adequate literacy levels to read study materials in Spanish [23-25]. This measure has high internal validity (>0.95 in either language) and strong positive associations with education. Additional eligibility criteria included regular access to text-message compatible and internet-connected devices through

XSL•FO
RenderX

home, work, or their community. Exclusion criteria included any serious medical condition that could make unsupervised PA unsafe (based on the modified Physical Activity Readiness Questionnaire) [26], including a history of coronary heart disease (history of myocardial infarction or symptoms of angina), stroke, or orthopedic conditions which limit mobility; other exclusion criteria were hospitalization due to a psychiatric disorder in the past 3 years, BMI above 45 kg/m^2, current or planned pregnancy, or moving from the area within this study's time frame.

Interventions

Original Intervention: Pasos Hacia La Salud I

In the original web-based intervention program, participants attend an in-person or Zoom-based (Zoom Video Communications, Qumu Corporation) baseline session led by trained bilingual or bicultural interventionists to be oriented to the intervention website and set incremental goals. The interventionist guides participants in goal setting (eg, setting a personalized weekly MVPA goal), and instructs them on how to progressively increase their goals to build toward meeting national PA guidelines (ie, graded tasks). The goal-setting session is based on principles of motivational interviewing and also teaches problem-solving skills so participants can adjust plans when necessary. In addition to this baseline goal-setting session, participants complete goal-setting sessions at month 1, and during assessment follow-up visits. Participants were encouraged to accumulate moderate-intensity PA on a majority of the days of the week and recommended to achieve a goal of 150 minutes of MVPA by the 6-month visit. Participants were not required to accumulate MVPA in 10-minute bouts.

Web-based tools to help participants increase their MVPA include (1) self-monitoring of behavior (eg, calendars to log weekly activity), (2) discrepancy between current behavior and goal (eg, calendars display weekly goals alongside weekly activity), (3) social support (eg, a discussion board where participants can interact), (4) instruction about how to be physically active (eg, information about local PA resources and places to be active), (5) problem solving (eg, tips for overcoming common PA barriers; neighborhood safety, and caregiving responsibilities), (6) motivational stage-matched PA manuals based on the TTM [17,22], and (7) individually tailored feedback based on their responses to online questionnaires assessing self-efficacy for PA, processes of change for PA, and social support. Further details about the tailored intervention content are described elsewhere [21]. Participants are encouraged by staff to regularly access the website, log their PA on the website, and receive a tip of the week. Participants are also prompted to complete online surveys to tailor their intervention content every month.

Enhanced Intervention: Pasos Hacia La Salud II

In addition to all of the intervention components in the original intervention described above, Pasos II includes the following enhancements: (1) text messages, (2) additional phone calls, and (3) additional data-driven content and website features that require user participation. In previous trials, the extent to which the Pasos I intervention increased PA behavior was associated with greater use of the website (ie, logging in more frequently, spending more time on the website, and use of specific dynamic features). Previous trials also demonstrated that the social cognitive theory [20] constructs, particularly self-efficacy, enjoyment, and social support, were predictive of PA behavior in this population. Thus, the enhancements to the original intervention were designed to increase participant engagement with the website and further target these constructs.

For 6 months after randomization, participants received daily text messages (in Spanish), which were personalized, and targeted self-efficacy, enjoyment, and social support (eg, "You can combine fun and social time with exercise." "Do you enjoy dancing? Take a weekly dance class of flamenco, salsa or tango. Be adventurous!"), along with reminders to use the tracking features on the website (eg, What would you like your exercise goal to be this week? Login to the website to enter it on the calendar! Login to the website to enter it on the calendar!). Participants also received additional phone calls at 2 and 3 months to check on their progress and discuss any barriers to them achieving their PA goals.

To further promote website engagement, participants were provided with nonmonetary, material incentives for their website use (Multimedia Appendix 1). Specifically, participants could earn points to receive items like water bottles and phone cases. Participants earned points through logging on to this study's website, and using features on the website such as posting a PA goal, completing a questionnaire, and other features which were associated with the largest change in PA in the original study. In weeks that participants self-recorded at least 150 minutes of PA, they were awarded a web-based "medal." These medals could be collected cumulatively and were displayed for participants to view in their goal-setting tab on the website. Earning this medal meant that participants were part of the "150 Minute Club" and they got to see their name posted on the "150 Minute Club" board. All participants in the enhanced arm could view the names of members currently, and recurrently in the "150 Minute Club." To enable additional opportunities for engaging with the intervention content, the motivational stage-matched manuals were made available in both audio and written formats.

To further enhance social support for PA, a discussion board was used to facilitate community meetups among participants. Research staff would post details about free and low-cost PA events in the area so that participants could meet up, discuss coordinating attendance with others, overcome barriers to PA, and motivate each other to attend these events.

Procedures

Recruitment

Community-based strategies were used to recruit participants including a Facebook (Meta) page, internet-based advertising, in-person recruitment at community events and organizations, advertisements in Spanish language newspapers, radio stations and television advertising, and flyers posted in the community. Participants were primarily recruited from in and around Providence County in Rhode Island.

XSL•FO
RenderX

Baseline Sessions

Following recruitment, eligible participants met with bilingual, bicultural staff for an orientation session, who outlined study requirements, potential risks, and benefits, and completed the informed consent process with participants. Participants were given an accelerometer-based device (ActiGraph wGT3X-BT; ActiGraph, LLC) to wear for 7 continuous days before returning for their randomization visit. After participants completed all baseline measures, they were randomly allocated to 1 of the 2 (original or enhanced) PA interventions.

Randomization

Participants were randomized to 1 of 2 conditions using a permuted block randomization scheme designed by this study's biostatistician, which was stratified by baseline stage of change for PA Behavior according to the TTM; with most participants being in contemplation (intending to increase PA but not currently doing any activity) or action (doing some activity, but not meeting national activity guidelines). The study biostatistician created the random allocation sequence, and all staff were blinded to the allocation sequence. Outcome assessment staff are blind to the intervention condition, and intervention staff and participants are not blind to the participant intervention condition.

Measures

Overview

Following baseline assessments, participants' MVPA was assessed again at 6 months. Participants received a questionnaire package and an accelerometer-based device approximately 1 week before the assessment visit. All measures listed below were given at baseline and month 6, available in Spanish, and have been used in our previous studies [14,17-19,27].

Primary Outcome

The primary outcome was the total weekly minutes of MVPA measured via an accelerometer (ActiGraph wGT3X-BT) [28], which measures the movement and intensity of PA. Participants were instructed to wear the ActiGraph wGT3X-BT accelerometer-based device on their hip during waking hours for 7 consecutive days. Participants wore the device on weekdays and weekends, and for example, if a participant started wearing the device on Tuesday they would be asked to wear the device during waking hours until the end of the day on the following Monday. They also were instructed to write down the dates and times they put on and took off the accelerometer-based device. Per standard procedures, the minimum acceptable wear time was 5 days with at least 600 minutes daily or 4 days with at least 3000 minutes total. Daily and weekly minutes of MVPA were calculated with ActiLife (ActiGraph, LLC) software, using a minimum cutoff point of 1952 [29] to define moderate-intensity PA and a minimum activity bout of 10 minutes, for consistency with the original Pasos Hacia La Salud study [14,19].

Secondary Outcome

To provide additional detail about PA, and to capture more information about the types of activities, participants completed an interview to self-report their total weekly MVPA as a secondary outcome measure. The 7-day Physical Activity Recall (PAR) interview [30,31] provides an estimate of the total weekly minutes of PA. The 7-day PAR has repeatedly shown acceptable internal consistency, reliability, and concurrent validity with objective measures of MVPA [32-34], along with sensitivity to changes [33,34] in both moderate and intensive levels of PA [35,36]. Additionally, the 7-day PAR has demonstrated test-retest reliability among Latino participants [37]. The 7-day PAR collects data on any physical activity that is completed during the week. The interviewer would ask the participant to categorize the purpose of the activity into (1) leisure, (2) occupational, (3) housework, and (4) transport. All activities were collected regardless of their purpose, however, activities designated as "housework," "occupational," or "transport" were removed from the total number of minutes when determining eligibility for participation in this study.

Baseline PA

We evaluated baseline PA, as measured via self-report, as a potential moderator of intervention effects on outcomes at 6 months. To evaluate baseline PA as a moderator we dichotomized baseline PA levels based on whether participants were doing any activity at all at baseline (low-active) versus those doing no activity at all at baseline (inactive or 0 minutes of baseline MVPA).

Data Analyses

We assessed potential between-group differences in baseline characteristics (demographics and activity level) using graphical methods, t tests for continuous variables, and χ^2 tests for categorical variables. We examined the effects of the enhanced versus the original intervention at the end of the active intervention period on our primary outcome: minutes per week of objectively measured MVPA at 6 months using a generalized linear model with the identity link function. We controlled for the baseline value of the outcome, accelerometer wear time as well as a time-varying indicator of before vs during pandemic time). A similar series of regression models assessed between-group differences in the secondary outcome (self-reported MVPA). For all analyses involving self-reported MVPA, the total amount of weekly PA reported was based on all types of activity, thus including leisure, occupational, transport, and housework categories of MVPA.

As a subsequent step, given the high proportion of participants who were inactive at baseline (0 min/wk of self-reported MVPA), we examined the potential conditional effects of the intervention as a function of baseline activity level (dichotomized as inactive vs not). Effects were examined using similar models to those described previously: a generalized linear model that included the main effects of intervention (enhanced vs original), the potential moderator (indicator of being inactive at baseline), as well as the interaction between the two. Evidence of moderation exists if the coefficient of the interaction term is statistically different than zero. To assess whether the differences in the effects of treatment were a function of the amount of activity being reported, we ran a similar set of models with a continuous moderator (min/wk of MVPA at baseline), as opposed to the dichotomous version (no activity vs at least some). Analyses were performed with

XSL•FO
RenderX

self-reported MVPA as the moderating variable and with accelerometer-measured MVPA as the moderating variable (continuous and dichotomous).

Next, using a generalized linear model, we examined the effects of the intervention on the proportion of participants meeting national PA guidelines for aerobic activity (defined as at least 150 min/wk of MVPA) at 6 months, measured by an accelerometer-based device. Models specified a logit link function and estimated odds ratios (OR and corresponding 95% CIs) as measures of effect. Models were run first, only specifying the main effect of treatment assignment (original vs enhanced intervention) and then conditional on baseline PA (inactive vs not).

All analyses were based on the intent to treat the sample, with all randomized participants included in the analysis. Details about power analyses conducted to determine the sample size are reported elsewhere [21]. Generalized linear models use a quasi-likelihood-based approach to estimation thus making use of all available data (intent-to-treat sample) without directly imputing missing values. Analyses were run in SAS (version 9.4; SAS Institute Inc) with a significance level set at 5% a priori.

Results

Sample Characteristics

A full description of this study sample is presented in Table 1. The final sample included 195 participants randomized to the original PA intervention (n=92) or the enhanced intervention at baseline (n=103). Overall, participants were aged 43.31 (SD 10.29) years on average, predominantly Dominican (n=80 or 41%), 97.5% (n=190) of participants were first generation, with 62.4% (n=122) participants reporting at least a high school level education. Of note, 68.7% (n=134) of participants were inactive at baseline. There were no differences between intervention conditions at baseline for sociodemographic variables or PA level (P>.05). The flow of participants who were assessed for eligibility, randomized, and who attended follow-up visits are depicted in the CONSORT diagram (Figure 1). In the diagram, “completed” indicates those who attended the 6-month measurement visit and have data for the 7-day PAR. At 6 months, 133 participants completed the accelerometer-based device measurements, with no differences between conditions.

XSL•FO
RenderX

Table 1. Baseline characteristics of study sample by study arm (N=195)[a].

Characteristics	Enhanced (n=103)	Original (n=92)	All (N=195)
Age (years), mean (SD)	42.93 (10.17)	43.73 (10.47)	43.31 (10.29)
Country of origin[b], n (%)			
Puerto Rico	11 (10.7)	16 (17)	27 (13.8)
Mexico (including Mexican American and Chicana women)	6 (5.8)	5 (5)	11 (5.6)
Cuba	0	0	0
Guatemala	14 (13.6)	13 (14)	27 (13.8)
Colombia	22 (21.4)	11 (12)	33 (16.9)
Dominica	39 (37.9)	41 (45)	80 (41)
Other	16 (15.5)	15 (16)	31 (15.9)
Race[c], n (%)			
American Indian or Alaskan Native	5 (4.8)	1 (1)	6 (3.1)
Asian	1 (1)	0	1 (0.5)
Black or African American	14 (13.6)	12 (13)	26 (13.3)
Mixed	4 (3.9)	2 (2)	6 (3.1)
Native Hawaiian or other Pacific Islander	1 (1)	1 (1)	2 (1)
Other	26 (25.2)	16 (17)	42 (21.5)
White	31 (30.1)	39 (42)	70 (35.9)
Education[c], n (%)			
<High school	36 (34.9)	34 (37)	70 (35.9)
High school graduate	14 (13.6)	8 (9)	22 (11.3)
Vocational or technical school	7 (7.1)	11 (12)	18 (9.2)
Some college	18 (17.5)	13 (14)	31 (15.9)
College graduate	22 (21.3)	18 (20)	40 (20.5)
Postgraduate	2 (1.9)	3 (3)	5 (2.6)
Employment status[c], n (%)			
Unemployed	35 (34.7)	19 (21)	54 (27.7)
Employed full time	41 (40.6)	42 (46)	83 (42.6)
Employed part-time	25 (24.8)	25 (27)	255 (28.6)
Marital status[c], n (%)			
Never married or lived with partner	19 (18.4)	15 (16)	34 (17.4)
Divorced	13 (12.6)	9 (10)	22 (11.3)
Separated	11 (10.7)	11 (12)	22 (11.3)
Widowed	3 (2.9)	4 (4)	7 (3.6)
Married	45 (43.7)	44 (48)	89 (45.6)
Living with partner	9 (8.7)	9 (10)	18 (9.2)
BMI, mean (SD)	29.33 (6.59)	30.89 (6.30)	30.09 (6.48)
Baseline MVPA[d] (min/wk), mean (SD)			
Accelerometer-measured	19.70 (47.93)	20.57 (42.69)	20.11 (45.40)
Self-reported[e]	55.28 (262.07)	52.43 (150.27)	53.94 (216.11)
Acculturation, STOFHLA[f] score, mean (SD)	33.50 (3.70)	33.30 (3.34)	33.41 (3.53)

XSL•FO
RenderX

[a]Not all categories sum to 100% due to rounding of percentages.

[b]Participants could select more than one option in response to the question about their country of origin; thus, the numbers sum to more than the total for the respective treatment arm.

[c]Numbers for these categories do not sum to 100 due to missing data for some questions.

[d]MVPA: moderate to vigorous physical activity.

[e]Self-reported moderate to vigorous physical activity represents all types of physical activity in this table; eligibility was based on only leisure time physical activity.

[f]STOFHLA: Short Test of Functional Health Literacy in Adults.

Figure 1. PASOS II month 6 CONSORT diagram. CONSORT: Consolidated Standards of Reporting Trials; PA: physical activity; STOFHLA: Short Test of Functional Health Literacy in Adults.

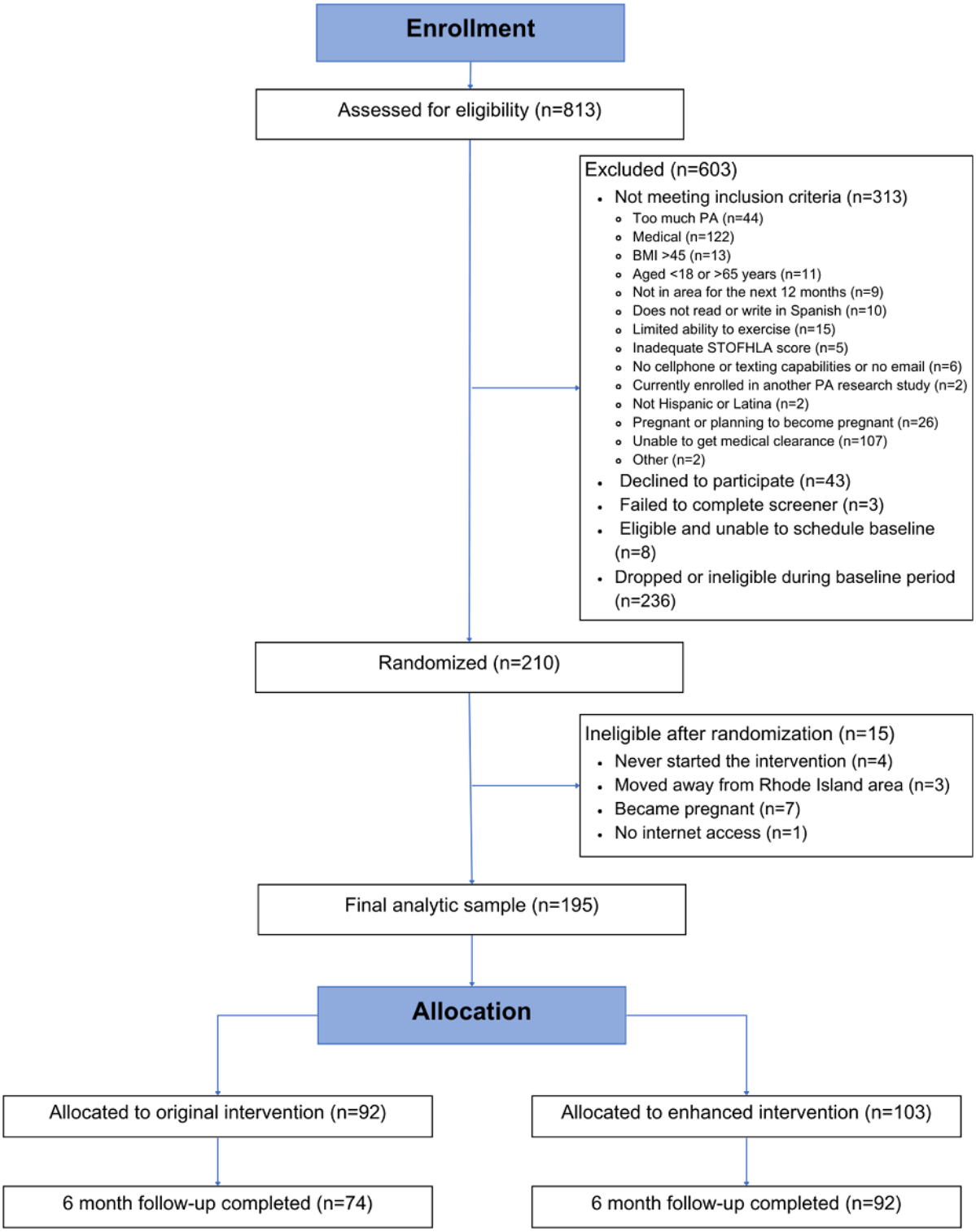

XSL•FO
RenderX

Changes in Physical Activity Over Time

Mean minutes per week of accelerometer-measured MVPA increased significantly in both groups, from 19.70 (SD 47.93, median 0) at baseline to 46.90 (SD 66.24, median 19.00) at 6 months among enhanced participants compared to 20.57 (SD 42.69, median 0) to 42.87 (SD 78.18, median 0) among those randomized to the original intervention. An adjusted model (adjusting for baseline MVPA and accelerometer wear time) did not indicate significant between-group differences at 6 months (*P*=.78).

The 7-day PAR interview findings were similar. Participants in the enhanced intervention condition increased self-reported MVPA from 57.88 (SD 260.25, median 0) minutes per week at baseline to 114.97 (SD 119.63, median 75) compared to the original intervention who reported an increase from 55.19 (SD 149.12, median 0) to 87.88 (SD 115.22, median 52). There were no differences between intervention conditions by self-reported MVPA at 6 months (*P*=.13). Thus, there was a significant increase from baseline to 6 months in MVPA in both groups, but this increase was not significantly different between treatment arms for either self-report or accelerometer-measured PA.

We found a moderating effect of self-reported baseline PA (inactive vs not) on both accelerometer-measured and self-reported MVPA at 6 months (controlling for baseline). For inactive participants, all participants reported 0 total minutes of self-reported MVPA at baseline.

For low-active participants, the median baseline level of MVPA was 48 minutes per week, with 25% (n=49) of the sample reporting 20 minutes per week or less. Among inactive participants, there was no significant effect of the enhanced intervention compared to the original intervention on self-reported MVPA (*b*=7.10; SE 22.79; *P*=.75). For those reporting some activity at baseline (low-active), significant effects were detected which favored the enhanced intervention (*b*=72.51; SE 27.91; *P*=.01). Similar findings were seen with accelerometer-measured MVPA, with significant effects of the enhanced intervention among those who reported at least some activity at baseline (*b*=21.51; SE 8.32; *P*=.02). Moderating effects were not detected based on the *amount* of activity being done at baseline (that is, the number of minutes), only on whether participants did some versus none (see Figure 2). There were no moderating effects of minutes per week of MVPA at baseline, suggesting intervention effects were not conditional on the quantity of PA at baseline, but just whether or not *any* was being completed. The findings were nearly identical when we examined the moderating variable of inactive versus low-active with accelerometer-measured MVPA, and the 6-month MVPA outcomes were self-reported or accelerometer-measured.

Figure 2. Path diagram depicting the results of the moderator analysis of unadjusted minutes per week of MVPA (self-reported MVPA) stratified by baseline self-reported MVPA. MVPA: moderate to vigorous physical activity.

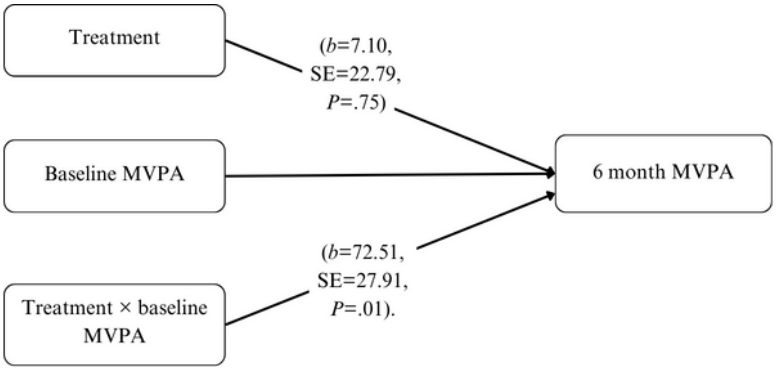

Looking across all participants, 30% (n=31) of participants in the enhanced intervention condition met national PA guidelines for aerobic activity at 6 months based on self-reported data, compared to 21% (n=19) of original intervention (OR 1.75, 95% CI 0.87-3.55). No participants met aerobic PA guidelines at baseline in either condition. Much like the continuous outcome for self-reported MVPA, for low-active participants, 45% (n=46) of participants in the enhanced intervention met national PA guidelines at 6 months compared to 20% (n=18) of original intervention participants (OR 3.29, 95% CI 1.05-11.31). For those who were inactive at baseline, there was no difference between arms in meeting national guidelines (n=25 vs n=19, 24% vs 21%) for enhanced versus original intervention (OR 1.28, 95% CI 0.54-3.06).

Discussion

Principal Findings

The objective of this paper was to evaluate whether the enhanced Pasos Hacia La Salud II intervention significantly increased PA, and helped more Latina women meet PA guidelines within 6 months, compared to the original intervention (Pasos I). The primary goal of this study was to evaluate whether there was a difference between the intervention conditions at reaching our primary end point: meeting national PA guidelines at 6 months. Overall, the results of this study indicate that both the enhanced and original intervention conditions helped Latina women increase their PA during the 6-month intervention. Both arms of the intervention appeared equally effective, with a significant increase in PA from baseline, yet there were no significant differences between groups in overall changes in minutes of activity or in meeting national guidelines.

XSL•FO
RenderX

However, the findings indicated that meeting aerobic PA guidelines and increasing PA differed for those participants who started the program engaging in some amount of MVPA at baseline, compared to those who started with no MVPA. For participants doing even a small amount of MVPA at baseline, the enhanced intervention was more effective at increasing the mean number of minutes per week of MVPA from baseline through 6 months, compared to the original condition. For those doing no activity at baseline, however, there were no differences in the intervention effects by treatment condition. Furthermore, the enhanced intervention resulted in more participants reaching the national aerobic PA guidelines within 6 months than the original intervention, only among low-active participants (ie, those who were doing some activity at baseline). Notably, this moderator effect was seen only for the dichotomous outcome (some vs no activity), not for baseline activity as a continuous outcome, suggesting an important difference between those who were doing some amount of MVPA albeit small (ie, 25-50 min/wk)—compared to those who had not yet begun to engage in any MVPA.

For the overall main effects, the findings are comparable to the original intervention study, which was conducted among a geographically and culturally different population of Latina women. In Pasos I, the participants were predominantly Mexican American (84.4%), and first-generation Americans (81.9%) recruited from the San Diego, California, area [14], whereas in Pasos II the participants were predominantly Dominican recruited from the Providence, Rhode Island, area and 97.5% were first generation. Despite differences in this study's settings and study participants, in both settings, the original intervention increased minutes of PA at 6 months. In Pasos I, intervention participants increased their self-reported MVPA from 8 minutes per week to over 100 minutes per week [14], and in Pasos II, participants receiving the same intervention as those in Pasos I increased their self-reported PA from an average of 55 (SD 262.07) minutes at baseline to more than 85 minutes per week, and the enhanced intervention participants increased from 57 minutes at baseline to more than 114 minutes at 6 months. Future research can examine statistical comparisons of these 2 samples through integrative data analysis, which is beyond the scope of this paper. Of note, the intervention evaluated in both of these trials indicates generalizable efficacy for increasing MVPA in Latina women, in 2 different samples of Latina women, in 2 geographically different regions.

Few other studies have been conducted that evaluate the effects of a culturally tailored, technology-delivered PA intervention for Latina women. In a systematic review of culturally tailored, technology-delivered interventions for changing health behaviors, there is mixed support across health behaviors to recommend culturally adapting technology-delivered interventions [38]. Notably, for PA behavior, there is a small but significant effect that favors culturally tailored PA interventions from 5 previous randomized controlled trials. This review included the efficacy trial which tested the original Pasos intervention, and the addition of this study's findings support the evidence in favor of a culturally tailored PA intervention for Latina women [38]. Additional studies evaluating PA interventions tailored to the Latino population detect similar results. For example, a previous study examined the effects of a PA intervention for Latina women in faith-based organizations. This study identified a marginal increase in self-reported MVPA and nonsignificant findings for accelerometer-measured PA behavior [39]. Although the findings are comparable, the key difference between this study and the faith-based culturally tailored PA intervention is that the latter requires substantial in-person involvement, and thus a large amount of staff time is required to implement an intensive in-person PA intervention. This difference in intervention delivery has important implications for the future scalability and dissemination of PA interventions for Latina women in real-world contexts, with constraints around staff- and participant time for in-person interventions.

The results of this study expand upon the generalizability of the findings for the original intervention by testing the effects among a geographically and culturally different population. However, the conditional effects of the enhanced intervention may not generalize to other Latina subgroups, Latino men, or other ethnic groups. For example, features specific to the enhanced arm such as the discussion board with opportunities for group meetups, text messages, and other content may not be as engaging or as motivating as it was for this group. Additionally, although the enhanced intervention added features to improve engagement with the intervention (by prompting the use of the website, and offering both audio and written versions of some materials), future research could further expand engagement with the program through the addition of infographics, video, or additional audio-based content.

While it is possible that the improvements in PA in this study could be seen as modest, with one-third of participants reaching the recommended PA levels in 6 months, it is important to note the distinction in the effects of the intervention for those who were doing any activity at all, versus those doing none. Nearly two-thirds of the sample were doing no PA at all at baseline. To motivate completely inactive individuals to become active may be the most difficult part of the behavior change process, compared to those who are low-active, which was reflected in our results. While both conditions significantly helped participants increase their minutes of MVPA per week, the intervention enhancements were most helpful for participants doing some activity at baseline, not for completely inactive participants. Specifically, the moderator findings in particular highlight that there are possibly differences in the effects of the enhanced versus original intervention for those who are contemplating change, versus those who have already started to change their MVPA behavior [22]. Previous research suggests that the underlying barrier to the stage of change progression is different for those who are in the contemplation stage (deciding whether to change) versus those who are in the preparation stage (insufficient beliefs about their abilities to successfully change) [40]. Thus, while the original intervention is known to increase self-efficacy [14], the intervention enhancements were designed to further support increases in self-efficacy, and promote engagement with the original intervention which may be 1 explanation for the moderating treatment effects. However, these enhancements may not be providing enough support in the decision-making process for

XSL•FO
RenderX

inactive Latina women to become physically active. This could potentially be facilitated by using additional forms of technology.

For example, delivering the intervention via a smartphone app, wearables, or other technology-based tools may increase convenience and would enable the use of push notifications to nudge participants, which might be particularly beneficial for inactive individuals. To date, few other technology-delivered interventions have been tested that are culturally tailored to the Latina and Latino populations. Four other published studies used similar technology to Pasos I and Pasos II, using the internet, or text messages to increase MVPA [41]. Two other studies evaluated MVPA interventions specifically designed for the Latino population and used fitness wearables, or fitness-tracking mobile apps as part of the intervention [42,43]. These interventions resulted in significant increases in MVPA, and given that both were feasibility studies using smaller sample sizes, with less than 6 months follow-up [42,43], future research might evaluate whether more advanced fitness technology may help to augment the effects of the Pasos intervention. However, such interventions would likely exclude individuals without access to certain devices, including smartphones, having implications for addressing health disparities.

Lastly, this study identified changes in MVPA after 6 months, and previous MVPA interventions among Latina women have either evaluated shorter follow-up durations [41-43] or have failed to identify longer-term maintenance of intervention effects [19]. Thus, while this study identified that the enhanced intervention arm may lead to significantly greater increases in MVPA after an initial period for Latina women doing any MVPA at baseline, future research should evaluate whether (1) these findings persist beyond 6 months and (2) whether those who were inactive at baseline need either more support or more time (>6 mo) for the original or enhanced Pasos intervention to increase MVPA behavior.

The findings from this study indicate that an enhanced version of the web-delivered, individually tailored intervention for Spanish-speaking Latina women resulted in larger increases in PA compared to the original (Pasos I) intervention, but only among those who were low-active at baseline. Overall, the enhanced version of the intervention did not outperform the original version, and both versions of the intervention significantly increased MVPA in Latina women. The enhanced version of the website incorporated text messaging, contained dynamic features to encourage website use, and featured additional targeting of key theoretical constructs identified as important mechanisms of action in the prior study [14,44]. Future research should seek to better understand the extent to which additional tailoring, time, or intervention techniques may be needed for those who report doing no PA at all and seek to understand methods for the dissemination and implementation of the enhanced intervention for low-active Latina women.

Acknowledgments

We would like to thank Susan Pinheiro, Rachelle Edgar, Esther Solis-Becerra, Viveka Ayala-Heredia, and the dedicated study staff for their valuable assistance and contributions to this work. This work was supported by the National Cancer Institute (NCI) of the National Institutes of Health (2R01CA159954). The NCI was not involved in study design, data collection or management, data analysis and interpretation, report writing, or the decision to submit the report for publication. The results of this study do not constitute an endorsement by the American College of Sports Medicine. The results of this study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of material incentives participants in the enhanced intervention condition could earn.
[DOCX File , 16 KB - jmir_v26i1e51708_app1.docx]

Multimedia Appendix 2
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 4363 KB - jmir_v26i1e51708_app2.pdf]

References

1. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity and severe obesity among adults: United States, 2017-2018. NCHS Data Brief 2020(360):1-8 [FREE Full text] [Medline: 32487284]
2. Hales CM, Carroll MD, Fryar CD, Ogden CL. Publications and information products from the National Center for Health Statistics. Centers for Disease Control and Prevention. 2017. URL: https://www.cdc.gov/nchs/products/index.htm [accessed 2022-11-15]
3. Narayan KV, Boyle JP, Geiss LS, Saaddine JB, Thompson TJ. Impact of recent increase in incidence on future diabetes burden: U.S., 2005-2050. Diabetes Care 2006;29(9):2114-2146 [FREE Full text] [doi: 10.2337/dc06-1136] [Medline: 16936162]

XSL•FO
RenderX

4. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. Circulation 2015;131(4):e29-e322 [FREE Full text] [doi: 10.1161/CIR.0000000000000152] [Medline: 25520374]
5. Liu X, Zhang D, Liu Y, Sun X, Han C, Wang B, et al. Dose-response association between physical activity and incident hypertension: a systematic review and meta-analysis of cohort studies. Hypertension 2017;69(5):813-820 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.116.08994] [Medline: 28348016]
6. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Himmelfarb CD, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Hypertension 2018;71(6):e13-e115 [FREE Full text] [doi: 10.1161/HYP.0000000000000065] [Medline: 29133356]
7. Kontis V, Mathers CD, Rehm J, Stevens GA, Shield KD, Bonita R, et al. Contribution of six risk factors to achieving the 25×25 non-communicable disease mortality reduction target: a modelling study. Lancet 2014;384(9941):427-437 [FREE Full text] [doi: 10.1016/S0140-6736(14)60616-4] [Medline: 24797573]
8. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA 2018;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
9. Martinez SM, Arredondo EM, Perez G, Baquero B. Individual, social, and environmental barriers to and facilitators of physical activity among Latinas living in San Diego County: focus group results. Fam Community Health 2009;32(1):22-33 [FREE Full text] [doi: 10.1097/01.FCH.0000342814.42025.6d] [Medline: 19092432]
10. Evenson KR, Sarmiento OL, Tawney KW, Macon ML, Ammerman AS. Personal, social, and environmental correlates of physical activity in North Carolina Latina immigrants. Am J Prev Med 2003;25(3 Suppl 1):77-85. [doi: 10.1016/s0749-3797(03)00168-5] [Medline: 14499813]
11. Larsen BA, Noble ML, Murray KE, Marcus BH. Physical activity in Latino men and women. Am J Lifestyle Med 2015;9(1):4-30. [doi: 10.1177/1559827614521758]
12. Aguayo-Mazzucato C, Diaque P, Hernandez S, Rosas S, Kostic A, Caballero AE. Understanding the growing epidemic of type 2 diabetes in the Hispanic population living in the United States. Diabetes Metab Res Rev 2019;35(2):e3097 [FREE Full text] [doi: 10.1002/dmrr.3097] [Medline: 30445663]
13. Benitez TJ, Dunsiger SI, Pekmezi DJ, Larsen BA, Mendoza-Vasconez AS, Linke SE, et al. Design and rationale for a randomized trial of a theory- and technology- enhanced physical activity intervention for Latinas: the Seamos Activas II study. Contemp Clin Trials 2020;96:106081 [FREE Full text] [doi: 10.1016/j.cct.2020.106081] [Medline: 32687974]
14. Marcus BH, Hartman SJ, Larsen BA, Pekmezi D, Dunsiger SI, Linke S, et al. Pasos Hacia La Salud: a randomized controlled trial of an internet-delivered physical activity intervention for Latinas. Int J Behav Nutr Phys Act 2016;13:62 [FREE Full text] [doi: 10.1186/s12966-016-0385-7] [Medline: 27234302]
15. Larsen B, Marcus B, Pekmezi D, Hartman S, Gilmer T. A web-based physical activity intervention for Spanish-speaking Latinas: a costs and cost-effectiveness analysis. J Med Internet Res 2017;19(2):e43 [FREE Full text] [doi: 10.2196/jmir.6257] [Medline: 28228368]
16. Descubrimiento digital: the online lives of Latinx consumers. Nielsen. 2018. URL: https://www.nielsen.com/insights/2018/descubrimiento-digital-the-online-lives-of-latinx-consumers/ [accessed 2024-04-16]
17. Marcus BH, Hartman SJ, Pekmezi D, Dunsiger SI, Linke SE, Marquez B, et al. Using interactive internet technology to promote physical activity in Latinas: rationale, design, and baseline findings of Pasos Hacia La Salud. Contemp Clin Trials 2015;44:149-158 [FREE Full text] [doi: 10.1016/j.cct.2015.08.004] [Medline: 26255237]
18. Marcus BH, Dunsiger SI, Pekmezi DW, Larsen BA, Bock BC, Gans KM, et al. The Seamos saludables study: a randomized controlled physical activity trial of Latinas. Am J Prev Med 2013;45(5):598-605 [FREE Full text] [doi: 10.1016/j.amepre.2013.07.006] [Medline: 24139773]
19. Hartman SJ, Dunsiger SI, Bock BC, Larsen BA, Linke S, Pekmezi D, et al. Physical activity maintenance among Spanish-speaking Latinas in a randomized controlled trial of an internet-based intervention. J Behav Med 2017;40(3):392-402 [FREE Full text] [doi: 10.1007/s10865-016-9800-4] [Medline: 27752866]
20. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs: Prentice-Hall; 1986.
21. Mendoza-Vasconez AS, Benitez T, Dunsiger S, Gans KM, Hartman SJ, Linke SE, et al. Pasos Hacia La Salud II: study protocol for a randomized controlled trial of a theory- and technology-enhanced physical activity intervention for Latina women, compared to the original intervention. Trials 2022;23(1):621 [FREE Full text] [doi: 10.1186/s13063-022-06575-4] [Medline: 35915473]
22. Prochaska J, Redding C, Evers K. The transtheoretical model and stages of change. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education: Theory, Research, and Practice, 4th Edition. San Francisco: Jossey-Bass; 2008:97-122.
23. Baker DW, Williams MV, Parker RM, Gazmararian JA, Nurss J. Development of a brief test to measure functional health literacy. Patient Educ Couns 1999;38(1):33-42. [doi: 10.1016/s0738-3991(98)00116-5] [Medline: 14528569]
24. Aguirre AC, Ebrahim N, Shea JA. Performance of the English and Spanish S-TOFHLA among publicly insured Medicaid and Medicare patients. Patient Educ Couns 2005;56(3):332-339. [doi: 10.1016/j.pec.2004.03.007] [Medline: 15721976]

XSL•FO
RenderX

25. Parker RM, Baker DW, Williams MV, Nurss JR. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. J Gen Intern Med 1995;10(10):537-541. [doi: 10.1007/BF02640361] [Medline: 8576769]
26. Thomas S, Reading J, Shephard RJ. Revision of the Physical Activity Readiness Questionnaire (PAR-Q). Can J Sport Sci 1992;17(4):338-345. [Medline: 1330274]
27. Pekmezi DW, Neighbors CJ, Lee CS, Gans KM, Bock BC, Morrow KM, et al. A culturally adapted physical activity intervention for Latinas: a randomized controlled trial. Am J Prev Med 2009;37(6):495-500 [FREE Full text] [doi: 10.1016/j.amepre.2009.08.023] [Medline: 19944914]
28. ActiGraph wGT3X-BT. ActiGraph Corporation. URL: https://theactigraph.com/actigraph-wgt3x-bt [accessed 2024-04-16]
29. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc 1998;30(5):777-781 [FREE Full text] [doi: 10.1097/00005768-199805000-00021] [Medline: 9588623]
30. Blair SN, Haskell WL, Ho P, Paffenbarger RS, Vranizan KM, Farquhar JW, et al. Assessment of habitual physical activity by a seven-day recall in a community survey and controlled experiments. Am J Epidemiol 1985;122(5):794-804. [doi: 10.1093/oxfordjournals.aje.a114163] [Medline: 3876763]
31. Sallis JF, Haskell WL, Wood PD, Fortmann SP, Rogers T, Blair SN, et al. Physical activity assessment methodology in the five-city project. Am J Epidemiol 1985;121(1):91-106. [doi: 10.1093/oxfordjournals.aje.a113987] [Medline: 3964995]
32. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
33. Hayden-Wade HA, Coleman KJ, Sallis JF, Armstrong C. Validation of the telephone and in-person interview versions of the 7-day PAR. Med Sci Sports Exerc 2003;35(5):801-809 [FREE Full text] [doi: 10.1249/01.MSS.0000064941.43869.4E] [Medline: 12750590]
34. Leenders NY, Sherman WM, Nagaraja HN, Kien CL. Evaluation of methods to assess physical activity in free-living conditions. Med Sci Sports Exerc 2001;33(7):1233-1240 [FREE Full text] [doi: 10.1097/00005768-200107000-00024] [Medline: 11445774]
35. Dunn AL, Garcia ME, Marcus BH, Kampert JB, Kohl HW, Blair SN. Six-month physical activity and fitness changes in Project Active, a randomized trial. Med Sci Sports Exerc 1998;30(7):1076-1083 [FREE Full text] [doi: 10.1097/00005768-199807000-00009] [Medline: 9662676]
36. Dunn AL, Marcus BH, Kampert JB, Garcia ME, Kohl HW, Blair SN. Comparison of lifestyle and structured interventions to increase physical activity and cardiorespiratory fitness: a randomized trial. JAMA 1999;281(4):327-334. [doi: 10.1001/jama.281.4.327] [Medline: 9929085]
37. Rauh MJ, Hovell MF, Hofstetter CR, Sallis JF, Gleghorn A. Reliability and validity of self-reported physical activity in Latinos. Int J Epidemiol 1992;21(5):966-971. [doi: 10.1093/ije/21.5.966] [Medline: 1468861]
38. Balci S, Spanhel K, Sander LB, Baumeister H. Culturally adapting internet- and mobile-based health promotion interventions might not be worth the effort: a systematic review and meta-analysis. NPJ Digit Med 2022;5(1):34 [FREE Full text] [doi: 10.1038/s41746-022-00569-x] [Medline: 35322172]
39. Arredondo EM, Haughton J, Ayala GX, Slymen D, Sallis JF, Perez LG, et al. Two-year outcomes of faith in action/Fe en Acción: a randomized controlled trial of physical activity promotion in Latinas. Int J Behav Nutr Phys Act 2022;19(1):97 [FREE Full text] [doi: 10.1186/s12966-022-01329-6] [Medline: 35907867]
40. Prochaska JO, Prochaska JM. Why don't continents move? Why don't people change? J Psychother Integr 1999;9:83-102. [doi: 10.1023/A:1023210911909]
41. Joseph RP, Royse KE, Benitez TJ. A systematic review of electronic and mobile health (e- and mHealth) physical activity interventions for African American and Hispanic women. J Phys Act Health 2019;16(3):230-239. [doi: 10.1123/jpah.2018-0103] [Medline: 30782040]
42. Rowland S, Ramos AK, Trinidad N, Quintero S, Beller RJ, Struwe L, et al. Feasibility, usability and acceptability of a mHealth intervention to reduce cardiovascular risk in rural Hispanic adults: descriptive study. JMIR Form Res 2022;6(12):e40379 [FREE Full text] [doi: 10.2196/40379] [Medline: 36563025]
43. Aguiñaga S, Marques IG, Kitsiou S, Balbim GM, Gerber BS, Buchholz SW, et al. BAILAMOS with mHealth technology! improving physical activity and well-being in middle-aged and older Latinxs: a pre-post feasibility study. Health Educ Behav 2021;48(5):575-583. [doi: 10.1177/10901981211027517] [Medline: 34521228]
44. Larsen B, Dunsiger SI, Pekmezi D, Linke S, Hartman SJ, Marcus BH. Psychosocial mediators of physical activity change in a web-based intervention for Latinas. Health Psychol 2021;40(1):21-29 [FREE Full text] [doi: 10.1037/hea0001041] [Medline: 33370154]

Abbreviations

MVPA: moderate to vigorous physical activity
OR: odds ratio
PA: physical activity
PAR: Physical Activity Recall

XSL•FO
RenderX

TTM: transtheoretical model

Edited by G Greco; submitted 09.08.23; peer-reviewed by S Cruz, V Stavric; comments to author 14.10.23; revised version received 10.11.23; accepted 01.03.24; published 06.06.24.

Please cite as:
Connell Bohlen L, Dunsiger SI, von Ash T, Larsen BA, Pekmezi D, Marquez B, Benitez TJ, Mendoza-Vasconez A, Hartman SJ, Williams DM, Marcus BH
Six-Month Outcomes of a Theory- and Technology-Enhanced Physical Activity Intervention for Latina Women (Pasos Hacia La Salud II): Randomized Controlled Trial
J Med Internet Res 2024;26:e51708
URL: https://www.jmir.org/2024/1/e51708
doi:10.2196/51708
PMID:38842930

©Lauren Connell Bohlen, Shira I Dunsiger, Tayla von Ash, Britta A Larsen, Dori Pekmezi, Becky Marquez, Tanya J Benitez, Andrea Mendoza-Vasconez, Sheri J Hartman, David M Williams, Bess H Marcus. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

App Engagement as a Predictor of Weight Loss in Blended-Care Interventions: Retrospective Observational Study Using Large-Scale Real-World Data

Marco Lehmann[1], PhD; Lucy Jones[2], MRES; Felix Schirmann[1], PhD

[1]Oviva AG, Potsdam, Germany
[2]Oviva UK Limited, London, United Kingdom

Corresponding Author:
Marco Lehmann, PhD
Oviva AG
Dortustraße 48
Potsdam, 14467
Germany
Phone: 49 3055572034
Email: marco.lehmann@oviva.com

Abstract

Background: Early weight loss is an established predictor for treatment outcomes in weight management interventions for people with obesity. However, there is a paucity of additional, reliable, and clinically actionable early predictors in weight management interventions. Novel blended-care weight management interventions combine coach and app support and afford new means of structured, continuous data collection, informing research on treatment adherence and outcome prediction.

Objective: Against this backdrop, this study analyzes app engagement as a predictor for weight loss in large-scale, real-world, blended-care interventions. We hypothesize that patients who engage more frequently in app usage in blended-care treatment (eg, higher logging activity) lose more weight than patients who engage comparably less frequently at 3 and 6 months of intervention.

Methods: Real-world data from 19,211 patients in obesity treatment were analyzed retrospectively. Patients were treated with 3 different blended-care weight management interventions, offered in Switzerland, the United Kingdom, and Germany by a digital behavior change provider. The principal component analysis identified an overarching metric for app engagement based on app usage. A median split informed a distinction in higher and lower engagers among the patients. Both groups were matched through optimal propensity score matching for relevant characteristics (eg, gender, age, and start weight). A linear regression model, combining patient characteristics and app-derived data, was applied to identify predictors for weight loss outcomes.

Results: For the entire sample (N=19,211), mean weight loss was –3.24% (SD 4.58%) at 3 months and –5.22% (SD 6.29%) at 6 months. Across countries, higher app engagement yielded more weight loss than lower engagement after 3 but not after 6 months of intervention ($P_{3\ \text{months}}$<.001 and $P_{6\ \text{months}}$=.59). Early app engagement within the first 3 months predicted percentage weight loss in Switzerland and Germany, but not in the United Kingdom ($P_{\text{Switzerland}}$<.001, $P_{\text{United Kingdom}}$=.12, and P_{Germany}=.005). Higher age was associated with stronger weight loss in the 3-month period ($P_{\text{Switzerland}}$=.001, $P_{\text{United Kingdom}}$=.002, and P_{Germany}<.001) and, for Germany, also in the 6-month period ($P_{\text{Switzerland}}$=.09, $P_{\text{United Kingdom}}$=.46, and P_{Germany}=.03). In Switzerland, higher numbers of patients' messages to coaches were associated with higher weight loss ($P_{3\ \text{months}}$<.001 and $P_{6\ \text{months}}$<.001). Messages from coaches were not significantly associated with weight loss (all P>.05).

Conclusions: Early app engagement is a predictor of weight loss, with higher engagement yielding more weight loss than lower engagement in this analysis. This new predictor lends itself to automated monitoring and as a digital indicator for needed or adapted clinical action. Further research needs to establish the reliability of early app engagement as a predictor for treatment adherence and outcomes. In general, the obtained results testify to the potential of app-derived data to inform clinical monitoring practices and intervention design.

(J Med Internet Res 2024;26:e45469) doi:10.2196/45469

XSL•FO
RenderX

KEYWORDS

obesity; weight loss; blended-care; digital health; real-world data; app engagement; mHealth; mobile health; technology engagement; weight management; mobile phone

Introduction

The rise of digital technology enables new treatments for people with overweight or obesity. Digital care elements increasingly augment weight management interventions, resulting in a variety of different digitally supported modes of care [1]. Some modes use digital technology as a communication tool between patients and clinicians; other variants provide stand-alone digital care. A third variant uses digital elements to enhance clinicians' care, blending digital and human care. This so-called blended care has been shown to deliver beneficial health effects [2] and holds the promise of scalability due to its partially digital character and remote mode of delivery [3].

Apps are frequently used for blended-care interventions, mainly due to the ubiquity of smartphones. Accordingly, a host of app-based blended-care interventions for weight management has emerged, spurring research on their efficacy. Several studies indicate their effectiveness in supporting people with overweight or obesity in weight management [1,4,5]. Furthermore, these interventions have additional benefits. Remote delivery simplifies access to care, which is especially relevant for underserved populations [6]. Also, they can increase the uptake of referred patients (to start the intervention) [7] and adherence to the intervention [8].

Moreover, blended care enables the collection of additional, otherwise inaccessible, data originating from patients' interactions with health apps. The generated data are continuous and dense, potentially representing the entire length of an intervention with a plethora of data points. In addition, the data can be passively collected, that is, it is a by-product of patients' interaction with a given app that reduces the burden of data collection on patients (eg, through questionnaires) and offers an extra source of information (eg, on top of patient reports). Thus, digitally derived data could support remedying the dearth of research data relating to nutrition [9].

Connected to this is the growing availability of real-world evidence from app-based blended-care interventions, affording large-scale analyses of weight management interventions [10]. As the reproducibility of real-world evidence has recently been documented, there is growing trust in its reliability [11]. Accordingly, using real-world evidence side-by-side with evidence from randomized controlled trials for the evaluation of interventions is becoming feasible [12]. Research on weight management interventions is profiting from real-world evidence due to the great magnitude of the data sets as well as the density, continuity, and high ecological validity of the data. Moreover, real-world evidence enables in-depth analyses (post hoc, but also live) on how patients respond to an intervention. For example, a large-scale, real-world data analysis using a digital app without face-to-face support recently revealed significant effects of usage frequency on weight loss in the short- and midterm for up to 4 months of treatment [10].

The mentioned usage frequency or, more broadly, patients' engagement with an app is a key indicator of their treatment adherence [8]. Adherence to digital interventions has been succinctly defined as "[a] composite measure encompassing time online, activity completion, and active engagements with the intervention" [13]. However, various definitions of adherence (and its relation to engagement) coexist [14] and, in light of this complexity, standardization of the concept has been called for [15].

Nonetheless, app engagement is considered to be a valid indicator of adherence as well as a predictor of intervention outcomes [8], and app usage data assumes a pivotal role in terms of the measurability of engagement [16]. As high dropout rates (often measured as complete cessation of engagement) impact app-based interventions [17], researching engagement as a predictor of prospective adherence and anticipated treatment success is critical to inform intervention monitoring and to, potentially, trigger countermeasures preventing dropout. Regarding weight management interventions, several engagement metrics have been identified as predictors of overall weight loss. For example, self-weighing, persistent food logging, and activity were found to be significant predictors of weight loss during a 6-month intervention [18]. Using a point-based incentive system in a single-arm study, overall app engagement was shown to be associated with weight loss in a 3-month period [19]. Coaching, self-monitoring, and self-management were identified to be positively correlated with weight loss at 3 and 6 months [20]. Notably, the predictors seem to be time-sensitive, that is, their predictive value varies as per the time period in the treatment trajectory. Accordingly, time-dependent engagement, for example, in the first month, can be applied to inform the modeling of usage trajectories [21].

Against this backdrop, this retrospective study analyzes time-dependent app engagement as a predictor for weight loss in large-scale, real-world, blended-care settings. We hypothesized that patients who frequently used the app (higher engagers, ie, showing comparably higher engagement with the educational content, logging meals, and completing tasks) lose more weight than patients who use the app less frequently (lower engagers, ie, showing comparably lower engagement) at 3 and 6 months of a weight management intervention. Our analysis of real-world data aims to aggregate known indicators of app engagement into a standardized metric to predict weight loss. Furthermore, propensity score matching will allow us to isolate the causal effect of this metric of app engagement on weight loss in the 3- and 6-month periods. Finally, we provide separate analyses of structured weight management programs reimbursed in 3 different European health systems, enabling a first cross-country comparison of blended-care weight management interventions.

XSL•FO
RenderX

Methods

Study Design

This observational study used longitudinal real-world patient data from obesity treatment in weight management programs and obesity therapy in Switzerland, the United Kingdom, and Germany. Patients living with obesity used a mobile app (Oviva app [Oviva AG]) along with their treatment by dietitians and health coaches, known as the blended-care approach [22]. The study aimed to demonstrate the significant impact of using the mobile app (ie, engagement) on weight loss after 3 and 6 months of treatment. The data used for the analysis were automatically collected through the use of the app in the course of the treatment. Demographic data were collected in face-to-face, introductory, internet-based consultations and were recorded in a patient management system. There was no experimental manipulation involved, so the design yielded real-world data. The timeframe of enrollment was from January 3, 2019, to August 25, 2022. This report complies with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement for transparent reporting of observational studies (STROBE checklist in Multimedia Appendix 1).

Weight Management and Obesity Treatment Programs

The blended-care approach in Switzerland, the United Kingdom, and Germany similarly offers personalized support by dietitians and health coaches and combines this with digital care using the Oviva app. The app logs patient-reported weight, food, and activity data and keeps this information as a personal diary for the patient. It further provides learning content to empower patients to engage in a healthier lifestyle. The app also offers messaging between the patient and the dietitian or health coach. In all countries, treatment starts with an introductory consultation with a dietitian or health coach. However, there are also notable differences between countries. In Switzerland, patients are enrolled in a weight management program lasting 36 weeks. This consists of alternating coaching sessions and app-based messaging. Coaching sessions are structured interactions consisting of anamnesis, knowledge exchange, and providing patients with different options for weight reduction (eg, focus on meal size, snacks, and unhealthy beverages). Compliance with the intervention is monitored by the coaches and success is evaluated in the subsequent consultations. Patients from the United Kingdom are enrolled in the Tier 3 weight management program Way to Wellness (patients in obesity class 2 and above [BMI>35 kg/m^2]). The program consists of 12 months of personalized support by the dietitians and health coaches. Patients can choose between app-based chat or telephone consultations weekly or biweekly for the first 16 weeks, with 30 minutes of monthly follow-up over 12 months. The topics and curriculum covered align with the international obesity guidelines, incorporating goal setting, action planning, healthy eating and weight loss strategies, problem-solving, lapse and relapse management, and maintaining behavior change. German patients are enrolled in an app-supported individual dietetic therapy for obesity lasting 3 to 6 months. The program consists of an introductory consultation of 30 minutes to identify the patient's needs. Then, 4 further appointments of 15 to 30 minutes are scheduled within 6 months. Coaching sessions are personalized to cater to the individual patients' therapeutic needs (eg, fostering nutritional literacy). Contact with the coach is further possible through app-based messaging. If patients are willing to continue after the first 5 appointments, a represcription to extend the treatment is possible.

Participants

All patients were enrolled in the blended-care weight management program of their respective countries with the treatment aim of weight loss. Upon enrollment to the treatment program, patients provided written informed consent that their data stored in the Oviva app could be used for scientific purposes. Generally, patient access to the program was possible through a general practitioner referral. However, patients could sign up online with Oviva, declare their interest in the therapy, and then provide the referral later. In the United Kingdom, patients with BMI≥35 kg/m^2 were eligible for participation in the Tier 3 weight management program. Besides referral, access to the program was also possible through self-referral using a prescription prepayment certificate maintained by the National Health Service. In Switzerland, patients with BMI≥25 kg/m^2 were eligible, thus allowing patients with overweight and obesity in the program. In Germany, patients may enroll in the obesity therapy program after referral from their general practitioner. Eligibility here is not bound to a BMI threshold but depends on the respective health insurers' policy for reimbursement of the treatment. After completing the program, patients may receive another referral to continue with the program.

Measures

Primary Outcome: Percentage Weight Loss

Patients used their own devices to measure their body weight and then entered the measurements in the Oviva app as weight logs (patient-reported outcomes; Figure 1). There was no limit regarding the frequency of weight logging in a given time frame. To derive percentage weight loss as the primary outcome, body weight was collected at baseline and at 3 and 6 months after baseline.

XSL•FO
RenderX

Figure 1. Oviva app showing development of body weight logs.

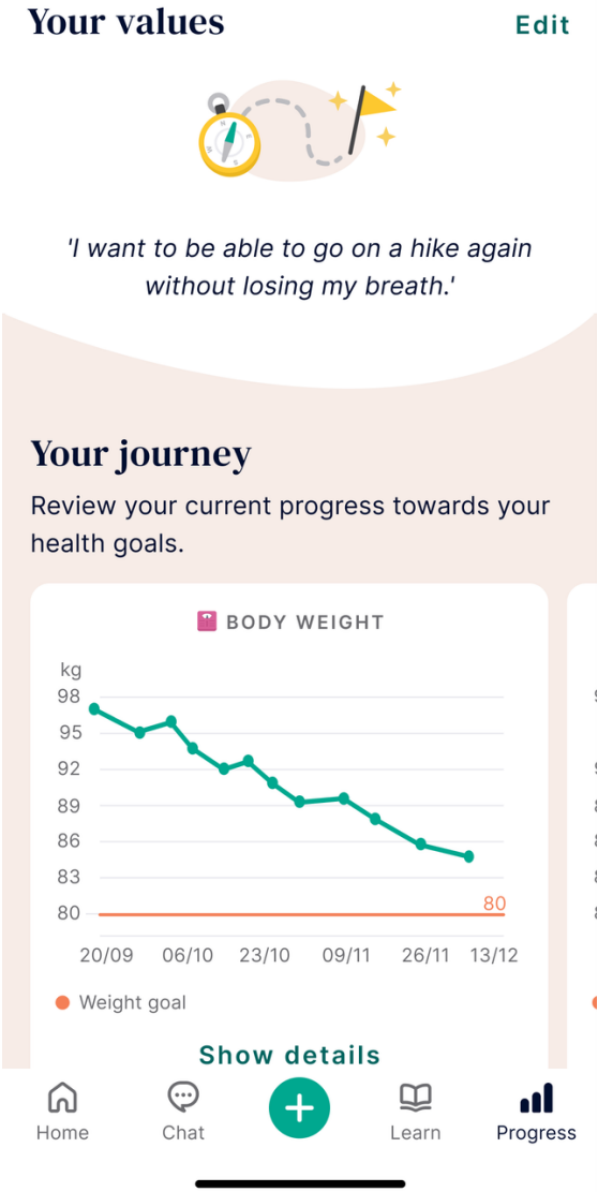

App Engagement

App engagement was measured using 5 indicators: meal logs, activity logs, weight logs, the number of completed goal-oriented tasks, and the number of access pages of learn content (Table 1). As part of the Oviva therapy, patients are asked to log meals and activities; complete tasks; and use learning modules that provide important dietary, behavior change, and health information. Logging takes minimal patient effort, focusing on a 1-click principle, which is always visible in the user interface (Figure 1). The app then offers several categories for logging, including meals and activities. There are different specifiers for meals and activities, such as meal ingredients or the type of activity. However, to derive a general measure of app engagement, we only counted the number of entries and not the content of entries. Task completion was handled through the diary function of the app. Patients could set and commit themselves to fulfill tasks. Task completion was recorded in the diary. Finally, patients could access several courses of learning content, each comprising several pages of the educational material. We used the count of accessed pages in the analysis. Counts in the period from baseline to 3 months of participation were captured as 3-month app engagement, and counts in the period between 3 months and 6 months of participation were captured as 6-month app engagement. All indicators were count variables.

XSL•FO
RenderX

Table 1. Indicators of app engagement used in multiple regression.

Intervention element and indicator of app engagement	Description	Unit
Self-monitoring		
Meal logs	Number of meals tracked through text or photo within the Oviva app	Count
Activity logs	Number of logged activities, symptoms, or measurements (eg, weight or blood glucose) tracked within the Oviva app	Count
Weight logs	Number of logged weights within the Oviva app	Count
Self-management		
Completed tasks	Number of completed tasks assigned by the coach or the patient, for example, track your meal today and make 5000 steps	Count
Education		
Learn content	Accessed pages of learn content	Count

Other Covariates

Additional patient characteristics included in the analysis were gender, age, and the presence of an obesity diagnosis (*International Classification of Diseases, Tenth Revision* [*ICD-10*] obesity [E66]). These characteristics were gathered by the dietitians or health coaches as part of the program enrollment. Furthermore, messages sent to coaches and messages received from coaches were considered in the analysis. These covariates are automatically recorded in the patient management system.

Statistical Analysis

Sample Description

Patient data were obtained from the Oviva database. We used all available data within the described time period from the 3 programs in the United Kingdom, Switzerland, and Germany. All patient data were anonymized for the analysis. We split the available data according to the 3 treatment pathways in Switzerland, Germany, and the United Kingdom. For each, we have shown patient characteristics and weight loss after 3 and 6 months of treatment. We further reported linear regression models with percentage weight loss after 3 and 6 months as outcomes. App engagement, message exchange, and further patient characteristics served as predictors. App engagement was used as a binary predictor distinguishing between lower and higher engagement.

Characteristics of Lower and Higher App Engagement

The 5 indicators of app engagement (ie, meal count, accessed learn pages, weight logs, activity count, and task completion) were $\log_{10}$ transformed to enhance the impact of differences in the lower end of the scale and to remove the impact of extreme counts. For example, the distinction between 10 and 50 accessed pages of learn content should have more impact on the analysis than the difference between 510 and 550 accessed pages. All indicators entered a principal component analysis (PCA), which yielded 1 engagement metric as the first principal component (PC). A median split was then applied to this metric to yield 2 levels of app engagement, identifying lower and higher engagers. This split enabled optimal propensity score matching for gender, age, obesity diagnosis, start weight, and message exchange [23,24].

Weight Loss Models

Multiple regression of weight loss after 3 and 6 months regarding lower and higher engagement and matched covariates will be reported. As main predictors, app engagement after 3 months and after 6 months were entered into the models. As covariates, we entered basic patient characteristics: gender, age, *ICD-10* E66 diagnosis (yes or no), and start weight. Furthermore, because the blended-care approaches in each of the 3 countries comprise digital patient-coach interaction, we further added the number of messages sent from or to the coach, respectively, as covariates. To make the 2 levels of higher and lower app engagement comparable, we used propensity score matching to achieve balanced data regarding the used covariates. So, in the regression models, there will be no confounding between app engagement and covariates. Matching was done separately for the 3- and 6-month periods and separately for the pooled data of all the countries and each of the 3 countries. This yielded 8 prespecified regression models to be reported. There was no covariate selection or removal using statistical criteria. Regression coefficients were tested using a 2-tailed *t* test at a level of significance α=.05. Statistical analysis used the R statistical software (R Core Team) [25]. Demographic information and descriptive statistics were aggregated using the *CreateTableOne* function from the *tableone* package [26], PCA used the *princomp* function from the *stats* package [25], propensity score matching used the *matchit* function (method parameter set to full matching or optimal matching) from the *MatchIt* package [23], and regression analyses used the *lm* function from the *stats* package [25]. All statistical methods used data from completed cases (Figure 2). We used no data imputation methods.

XSL•FO
RenderX

Figure 2. Flow diagram of available cases in the analysis. PCA: principal component analysis.

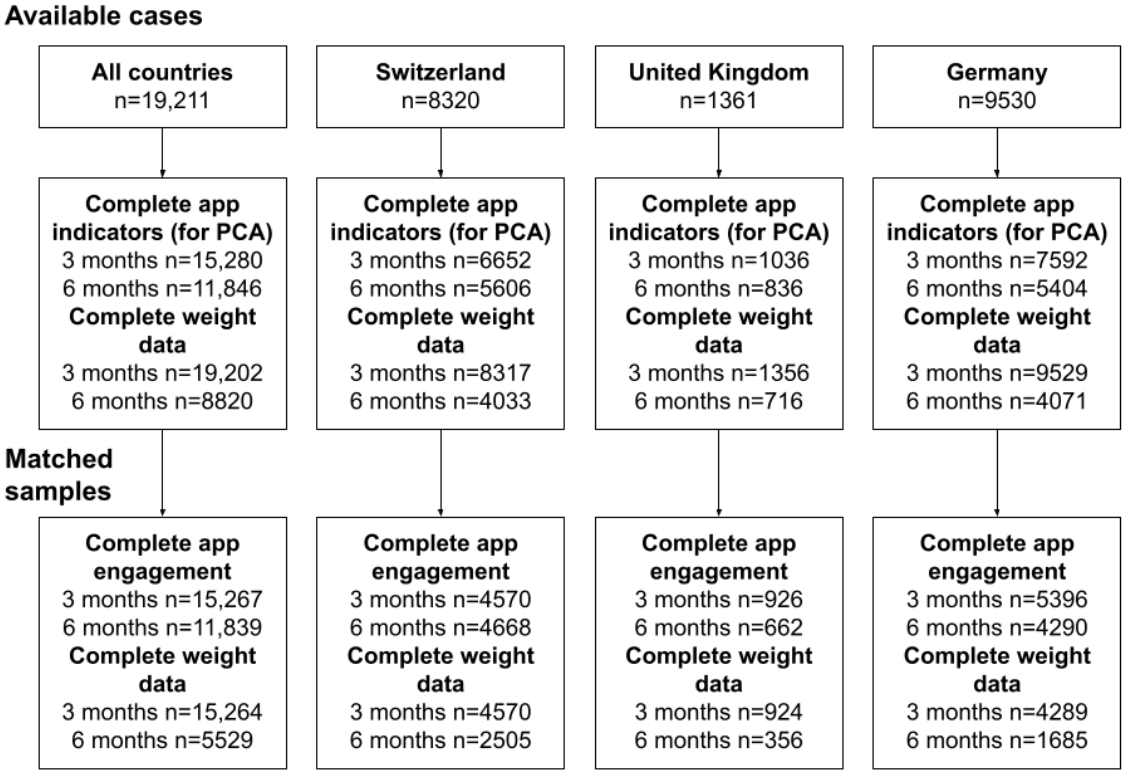

Ethical Considerations

This study is completely based on anonymized observational data using no experimental manipulation or randomization. The Oviva app used by the patients is a registered Conformité Européenne–marked medical device according to the European Union Medical Device Regulation 2017/745 and is certified according to ISO 27001. As such, data collection was performed within the postmarket surveillance activities [27] and the app was used within its intended purpose by all patients. This study is not a clinical investigation according to Medical Device Regulation (MDR) 2017/745 [27] and there were no invasive or other burdensome methods applied. To use the app, patients gave informed consent that enables data analysis. Participants received no compensation. This study obtained no external funding.

Results

Sample Description

Data points from 19,211 patients (age range: 18-96 years) were available and were split according to the blended-care weight management interventions in 3 countries: individual dietetic therapy (Germany), app-supported dietary advice (Switzerland), and Tier 3 weight management (United Kingdom). A fraction of patients had no Oviva app available during their participation; these 2314 (12%) patients were excluded from further analyses. Table 2 shows the summary statistics of patient characteristics, weight loss, and indicators of app engagement. As can be seen, in all 3 countries, the majority (14,763/19,211, 76.9%) of patients were female and the mean age was 49.87 (SD 13.38) years. In Germany and Switzerland, the majority of patients had an *ICD-10* E66 diagnosis of obesity (Germany: 7281/9530, 76.4%; Switzerland: 5820/8320, 70%), whereas only a minority of patients in the United Kingdom had this diagnosis (354/1361, 26%). However, in the United Kingdom, patients were required to have >class 2 obesity (BMI≥35 kg/m^2) to be eligible for the program, and this is also reflected in the much higher average start weight of these patients compared with the patients in Switzerland and Germany. For the entire sample (N=19,211), mean weight loss was –3.24% (SD 4.58%) at 3 months and –5.22% (SD 6.29%) at 6 months. Across countries, percentage weight loss for 3 and 6 months followed an expected pattern of –2.93% (–2.99 kg) to –4.44% (–5.6 kg) after 3 months of intervention and –4.68% (–4.78 kg) to –5.82% (–7.35 kg) after 6 months. Counts for messages and engagement indicators are not transformed in this table to allow for the appraisal of the average magnitude of these indicators. As can be seen, average meal counts are higher than counts of the 3 other indicators in both time periods, indicating the high importance of meal logging in obesity therapy.

XSL•FO
RenderX

Table 2. Patient characteristics.

Factor levels	All countries (N=19,211)	Switzerland (n=8320)	United Kingdom (n=1361)	Germany (n=9530)
Gender, n (%)				
Male	4433 (23.1)	2691 (32.3)	288 (21.2)	1454 (15.3)
Female	14,763 (76.9)	5629 (67.7)	1073 (78.8)	8061 (84.7)
Age (years), mean (SD)	49.87 (13.38)	50.19 (14.38)	48.63 (12.68)	49.77 (12.53)
E66[a] diagnosis, n (%)				
E66	13,455 (70.0)	5820 (70.0)	354 (26.0)	7281 (76.4)
No diagnosis	1893 (9.9)	173 (2.1)	996 (73.2)	724 (7.6)
Other diagnosis	3863 (20.1)	2327 (28.0)	11 (0.8)	1525 (16.0)
Start weight (kg), mean (SD)	100.53 (23.95)	94.48 (20.17)	126.23 (28.14)	102.13 (23.64)
Percent weight loss, mean (SD)				
3 months	–3.24 (4.58)	–3.40 (4.93)	–4.44 (5.93)	–2.93 (3.98)
6 months	–5.22 (6.29)	–5.65 (6.22)	–5.82 (7.37)	–4.68 (6.11)
Messages from coach, mean (SD)				
3 months)	16.94 (22.65)	22.36 (29.00)	7.13 (16.58)	13.60 (14.50)
6 months	8.35 (14.58)	11.40 (18.77)	1.44 (4.21)	6.67 (10.03)
Messages to coach, mean (SD)				
3 months	2.34 (15.94)	2.44 (9.62)	11.35 (29.44)	0.97 (17.15)
6 months	1.13 (15.10)	1.29 (6.46)	3.13 (8.19)	0.71 (20.32)
Meal count, mean (SD)				
3 months	130.23 (117.51)	113.47 (90.53)	116.28 (80.28)	144.05 (135.54)
6 months	130.17 (164.09)	123.03 (124.20)	110.02 (109.01)	137.54 (191.16)
Activity count, mean (SD)				
3 months	48.35 (117.72)	34.86 (77.35)	33.66 (68.37)	57.01 (137.06)
6 months	63.68 (166.85)	46.84 (109.12)	51.04 (121.31)	73.68 (192.72)
Number of weight logs, mean (SD)				
3 months	13.57 (27.24)	10.60 (23.34)	9.64 (9.70)	16.72 (31.42)
6 months	10.51 (23.18)	9.17 (23.54)	7.81 (12.78)	12.12 (23.93)
Number of completed tasks, mean (SD)				
3 months	71.98 (96.53)	47.64 (57.23)	58.10 (69.70)	89.44 (114.27)
6 months	76.17 (129.23)	54.66 (84.61)	65.00 (103.50)	90.87 (151.06)
Pages of learn content, mean (SD)				
3 months	24.05 (31.13)	9.12 (15.59)	30.10 (35.89)	30.86 (33.59)
6 months	18.04 (27.62)	6.61 (12.60)	19.41 (25.41)	24.26 (31.77)

[a]E66: obesity.

Characteristics of Lower and Higher App Engagement

The 5 indicators of app engagement—meal count, activity count, weight logs, accessed pages of learn content, and the number of completed tasks—were aggregated with PCA to a single metric of app engagement, that is, the first PC. For app engagement in the 3-month period, the first PC captures 58% of the variance comprising the 5 engagement indicators, whereas the other 4 components capture at most 16% of the variance. For app engagement in the 6-month period, the first PC captures 65% of the variance, whereas the other 4 components capture at most 14%. According to these numbers, we treat the first PCs as valid metrics of app engagement in the 3- and 6-month periods, respectively. The coefficients for the linear combinations of indicators of app engagement that yield the PCs are (1) $PC_{3\ months} = 0.32$ *accessed pages of learn content* $+ 0.53$ *meal count* $+ 0.22$ *weight count* $+ 0.47$ *activity count* $+ 0.59$ *completed tasks* and (2) $PC_{6\ months} = 0.18$ *accessed pages*

XSL•FO
RenderX

of learn content + 0.65 meal count + 0.24 weight count + 0.39 activity count + 0.58 completed tasks (all counts were log10 transformed before PCA). A median split was then applied to this metric to distinguish between the patients showing lower app engagement and the patients showing higher app engagement. The subsamples of patients with higher and lower app engagement were then matched according to the patient characteristics (ie, gender, age, diagnosis, and start weight) and message exchange between the patient and the coach. Multimedia Appendix 2 shows the statistics for weight outcome, for the generic predictors, and for the indicators of app engagement in the matched subsamples of lower and higher app engagement. The table has an upper part for the 3-month period and a lower part for the 6-month period. This division is necessary because the patients could engage differently with the app in the 2 periods and were categorized differently accordingly. As a consequence, the composition of the matched samples varied between the 2 periods.

Descriptively, weight outcome shows stronger average weight loss for higher app engagement than for lower engagement across countries (Figure 3). Furthermore, the same pattern was observed for the data from Switzerland and the United Kingdom, but not for the data from Germany (Figure 4). In fact, stronger weight loss with lower app engagement was observed on average in Germany after 6 months. Furthermore, the generic predictors and the indicators of app engagement show the required pattern. Because of the matching, the averages for the generic predictors vary little between lower and higher engagement, indicating successful matching. Conversely, because of the categorization into lower and higher engagement, the averages of the engagement indicators show higher variation.

Figure 3. Mean percent weight loss across countries for lower and higher app engagement.

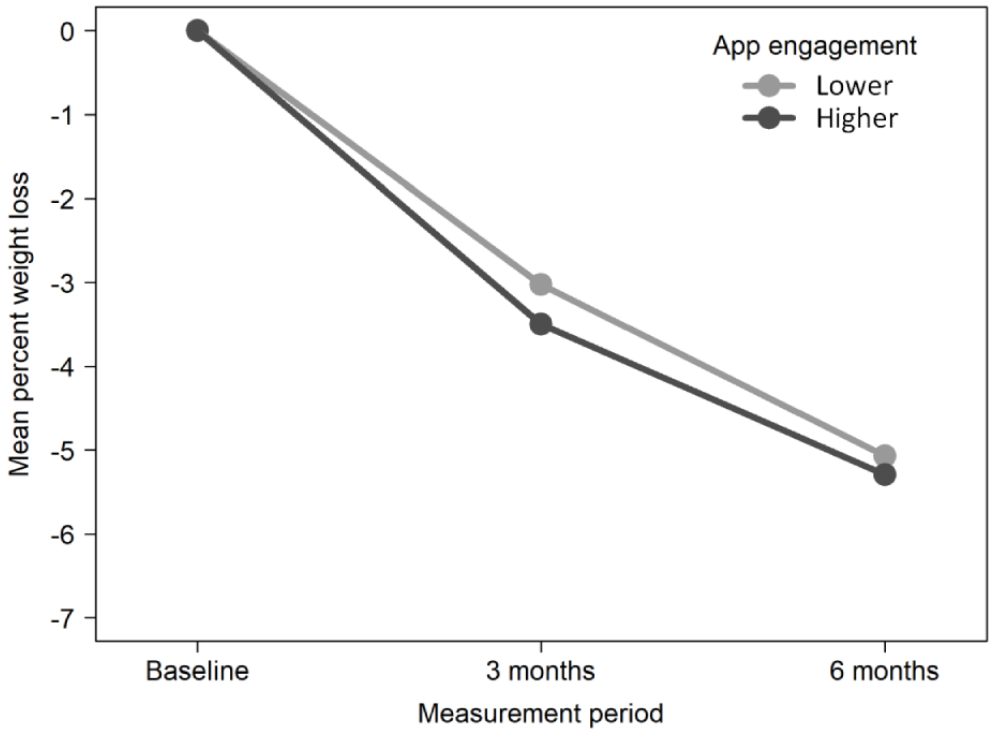

XSL•FO
RenderX

Figure 4. Mean percent weight loss for lower and higher app engagement in 3 countries. Error bars represent 95% CIs.

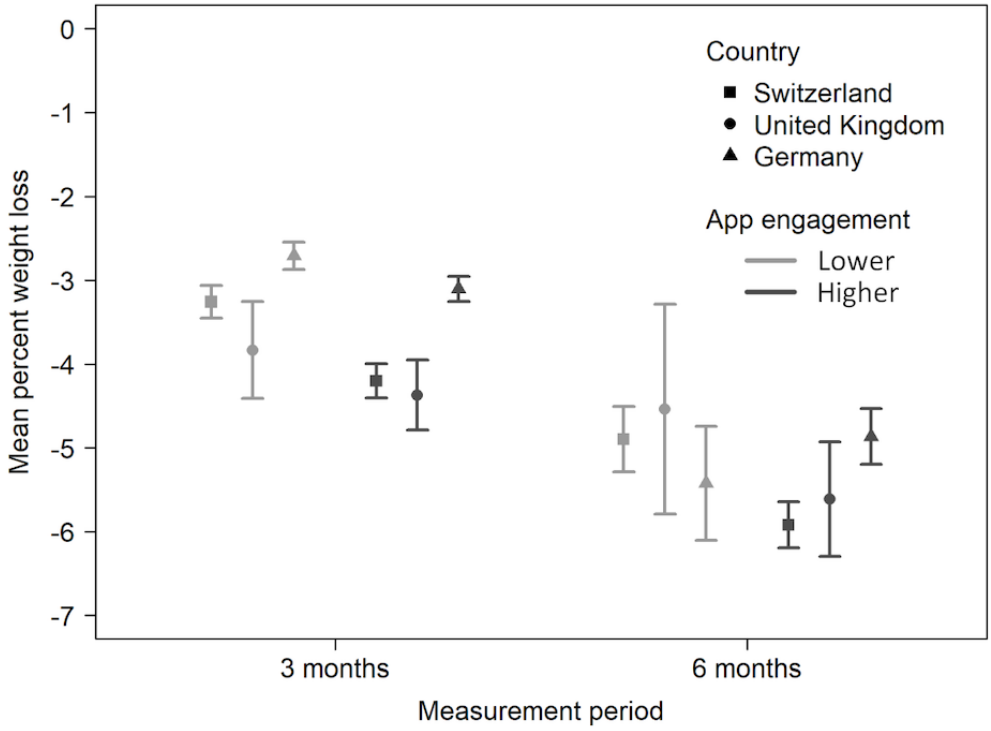

Weight Loss Models

Multimedia Appendix 3 relates to the hypothesized association between percentage weight loss and app engagement after 3 and 6 months of intervention. This association was confirmed in the 3-month period for all countries pooled, for Switzerland, and for Germany. Here, patients with higher app engagement lost significantly more weight than patients with lower engagement (pooled: $b_{3\ months}$=–0.34; *P*<.001; Switzerland: $b_{3\ months}$=–0.9; *P*<.001; and Germany: $b_{3\ months}$=–.34; *P*=.005). In the 6-month period, no significant association between higher app engagement and higher weight loss was found (pooled: $b_{6\ months}$=–0.10; *P*=.59), except for 1 coefficient. That is, in the 6-month period, patients with higher app engagement lost significantly more weight only in Switzerland ($b_{6\ months}$=–0.9; *P*<.001). For the data from the United Kingdom, the association could not be established (United Kingdom: $b_{6\ months}$=–1.06; *P*=.12) and, conflicting with our hypothesis, in individual dietetic therapy in Germany, a positive regression coefficient indicated more weight loss with lower app engagement (Germany: $b_{6\ months}$=0.42; *P*=.21). However, a consistently significant predictor across the countries was start weight, indicating that people with higher start weights achieved more weight loss (pooled: $b_{3\ months}$=–.03; *P*<.001 and $b_{6\ months}$=–.04; *P*<.001; Switzerland: $b_{3\ months}$=–0.04; *P*<.001 and $b_{6\ months}$=–0.04; *P*<.001; United Kingdom: $b_{3\ months}$=–0.04; *P*<.001 and $b_{6\ months}$=–0.04; *P*=.006; and Germany: $b_{3\ months}$=–0.03; *P*<.001 and $b_{6\ months}$=–0.06; *P*<.001). Similarly, in the 3-month period, persons of a higher age achieved more weight loss (pooled: $b_{3\ months}$=–0.02; *P*<.001; Switzerland: $b_{3\ months}$=–0.02; *P*<.001; United Kingdom: $b_{3\ months}$=–0.05; *P*=.002; and Germany: $b_{3\ months}$=–0.02; *P*<.001). Interestingly, the number of messages received from the coach yielded no significant effect on weight loss (pooled: $b_{3\ months}$=–0.05; *P*=.45 and $b_{6\ months}$=0.12; *P*=.37; Switzerland: $b_{3\ months}$=–0.07; *P*=.51 and $b_{6\ months}$=0.2; *P*=.26; United Kingdom: $b_{3\ months}$=0.26; *P*=.37 and $b_{6\ months}$=0.8; *P*=.39; and Germany: $b_{3\ months}$=–0.02; *P*=.91 and $b_{6\ months}$=0.31; *P*=.39), contrasting with significant effects of the number of messages to the coach, at least for the pooled data across countries ($b_{3\ months}$=–0.72; *P*<.001 and $b_{6\ months}$=–1.12; *P*<.001) and for Switzerland ($b_{3\ months}$=–0.57; *P*<.001 and $b_{6\ months}$=–1.09; *P*<.001). This may indicate a kind of asymmetrical impact of communication between the patient and the coach on weight loss.

Discussion

Principal Findings

App engagement was a consistent predictor of percent weight loss across countries in the 3-month period (except for the United Kingdom) but not in the 6-month period. In the 6-month period, only in the data from Switzerland was higher app engagement associated with higher weight loss. In the data from the United Kingdom and Germany, this association was not statistically significant. In fact, higher app engagement appears to be associated with less weight loss after 6 months in Germany, potentially linked to the focus on individual follow-up consultations, rather than app-based care in that program. Furthermore, in the German real-world setting, the program in some cases extends beyond 6 months, so that not all parts of it may have affected weight outcomes measured at 6 months. Therefore, our hypothesis that app engagement yields weight loss was confirmed in the 3-month period, but not in the 6-months period. Notably, whereas higher numbers of patient messages to the coach were associated with more weight loss

XSL•FO
RenderX

in the Swiss data and in the overall data set, this association was not found with the number of messages sent from the coach to the patient. The most consistent predictors for weight loss were start weight and age. Patients with higher start weight achieved significantly stronger weight loss after 3 and 6 months and those of higher age achieved stronger loss after 3 months of intervention.

Comparison With Previous Work

Our results confirm that early app engagement in blended-care weight management interventions (ie, time period of the first 3 months) is associated with additional weight loss as an outcome. Our contribution shows similarities to a retrospective real-world data analysis of 2113 patients with overweight from 2011 to 2015 [18]. This research revealed the number of weight measurements, activity times, and the number of food logs to be predictors of weight loss after 6 months. Using a multiple of the earlier sample size and separating the analysis for 3 different countries, our study revealed that such behavioral indicators truly predict weight loss, however, only for the 3-month period using aggregated indicators of app engagement.

More similar results originate from a large sample of 251,718 patients from China recorded from 2016 and 2017 [10]. This study showed that the frequency of app usage predicts weight loss after 4 months. A comparable effect of frequency of weighing on weight loss was observed by the same research group in patients who are severely obese with BMI≥35 kg/m² [28]. The obtained weight loss outcomes (overall: –3.24%,–3.26 kg at 3 months and –5.22%,–5.25 kg at 6 months) exceeded the weight loss identified in recent systematic reviews on the efficacy of smartphone-based weight loss interventions. For example, absolute weight loss of –1.99 kg [29] and –2.18 kg [30] for the 3-month period were documented. Moreover, the observed overall weight loss at 6 months is clinically meaningful and in line with the weight loss goals for this time period stipulated in the pertinent national and international guidelines [31,32].

To further improve weight loss as well as the intervention design, it is key to analyze the factors that drive engagement, as per the cohort and at the individual level [8]. Moreover, as earlier research indicated the time dependency of digital intervention elements that contribute to app engagement, the dimension of time needs to be considered in addition to quantitative and qualitative factors of engagement [20]. An in-depth understanding of these factors and how patients respond to them is a precondition for tailored engagement, and thus, personalized interventions.

Strengths and Limitations

In our study, app engagement was a particularly strong metric, based on recorded app data, which was not prone to data shortage or data loss. Therefore, we consider these data to yield a valid distinction between lower and higher app engagement. In fact, we recommend future use of this metric because it covers 3 behavioral areas known to be crucial in required lifestyle changes in patients with obesity—self-monitoring, self-management, and education. Furthermore, the propensity score matching used to equalize the groups of patients with higher and lower app engagement allowed for successful isolation of the effect of engagement on weight loss. This supports the idea that the observed differences in average weight loss after 3 or 6 months of intervention can be traced back to the differences in app engagement between the 2 groups of higher and lower app engagement in the respective time period. However, there may still be unobserved factors that may account for both quantities, app engagement and weight loss. Only a randomized controlled trial could rule out this possibility. This trial would have to manipulate (not only measure) app engagement and control for confounding variables using randomization [33].

Data collection for this study covered a long period from January 3, 2019, to August 25, 2022. Although we may be confident that this duration leveled out any potential seasonal effects on the study outcome, it was, nevertheless, possible that the 3 programs in Switzerland, Germany, and the United Kingdom underwent structural changes and staff rotation. Moreover, because the 3 programs differed in more than one respect, causal agents within the programs that contribute to app engagement and eventually yield weight loss are difficult to identify. However, the used programs can be considered as standard blended care in the 3 countries, which emphasizes the transferability of results across persons, settings, and time.

Conclusions

App engagement predicts weight loss in blended-care interventions, with higher engagers losing more weight than lower engagers. Our data consolidate the association between early app engagement and successful weight loss. The results stress that patients may be most receptive in the early phase of a weight loss intervention, opening opportunities for innovative treatment concepts such as just-in-time adaptive interventions supported by chatbot-based digital coaching [34]. Fitting in with this is our secondary result that messages sent by the patient have a stronger effect than messages sent by the coach. Possibly, human-written coach messages could be gradually replaced by automatic messages to maintain the messaging activity of the patient. However, future research should focus on the patient-intervention-fit for these options to be practicable. Furthermore, future research should focus on the reasons for lower app engagement. The potential fear of data exploitation as a known impediment to trust in digital technology [35] may be a tenable explanation. Future research should further scrutinize the differential impact of the components of app engagement [14].

In general, blended-care weight management interventions have surged in popularity in recent years for various reasons. App use for the treatment entails low barriers for most patients, increasing uptake. Digital care furthermore offers clinicians and health coaches real-time insights into the trajectories of important clinical outcomes. Finally, digital care produces comprehensive data sets needed for pivotal insights in health care research. While following the currents of digitization in health care to facilitate clinical delivery and data acquisition in routine care is beneficial, the nascent digital treatment revolution lies in the introduction of digital devices with a manifest effect on health outcomes [36].

XSL•FO
RenderX

Acknowledgments

We thank Tobias Larsen for aggregating the data set for this analysis and Mandana Rezwan for her counsel regarding European Union Medical Device Regulation (EU MDR) 2017/745. This research received no external funding.

Data Availability

This study used real-world data from actual therapeutic care. The data sets generated and analyzed during this study are not publicly available because patient consent only allows for the usage of anonymized data by Oviva for the purpose of this study but precludes making the data publicly available.

Conflicts of Interest

ML and FS are employed at Oviva AG, Potsdam, Germany. LJ is employed at Oviva UK Limited, London, United Kingdom.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist cross-sectional studies.
[PDF File (Adobe PDF File), 106 KB - jmir_v26i1e45469_app1.pdf]

Multimedia Appendix 2
Characteristics of patients with lower and higher app engagement.
[PDF File (Adobe PDF File), 45 KB - jmir_v26i1e45469_app2.pdf]

Multimedia Appendix 3
Weight loss prediction models using patient characteristics and app engagement as predictors.
[PDF File (Adobe PDF File), 50 KB - jmir_v26i1e45469_app3.pdf]

References

1. Cavero-Redondo I, Martinez-Vizcaino V, Fernandez-Rodriguez R, Saz-Lara A, Pascual-Morena C, Álvarez-Bueno C. Effect of behavioral weight management interventions using lifestyle mHealth self-monitoring on weight loss: a systematic review and meta-analysis. Nutrients 2020;12(7):1977 [FREE Full text] [doi: 10.3390/nu12071977] [Medline: 32635174]
2. Berry MP, Sala M, Abber SR, Forman EM. Incorporating automated digital interventions into coach-delivered weight loss treatment: a meta-analysis. Health Psychol 2021;40(8):534-545. [doi: 10.1037/hea0001106] [Medline: 34618500]
3. Ashrafzadeh S, Hamdy O. Patient-driven diabetes care of the future in the technology era. Cell Metab 2019;29(3):564-575 [FREE Full text] [doi: 10.1016/j.cmet.2018.09.005] [Medline: 30269984]
4. Dounavi K, Tsoumani O. Mobile health applications in weight management: a systematic literature review. Am J Prev Med 2019;56(6):894-903 [FREE Full text] [doi: 10.1016/j.amepre.2018.12.005] [Medline: 31003801]
5. Wang Y, Min J, Khuri J, Xue H, Xie B, Kaminsky LA, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. JMIR Mhealth Uhealth 2020;8(4):e15400 [FREE Full text] [doi: 10.2196/15400] [Medline: 32343253]
6. Lee EW, McCloud RF, Viswanath K. Designing effective eHealth interventions for underserved groups: five lessons from a decade of eHealth intervention design and deployment. J Med Internet Res 2022;24(1):e25419 [FREE Full text] [doi: 10.2196/25419] [Medline: 34994700]
7. Ross JAD, Barron E, McGough B, Valabhji J, Daff K, Irwin J, et al. Uptake and impact of the English National Health Service digital diabetes prevention programme: observational study. BMJ Open Diabetes Res Care 2022;10(3):e002736 [FREE Full text] [doi: 10.1136/bmjdrc-2021-002736] [Medline: 35504697]
8. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
9. Fanzo J. The need for nuance with dietary data. Am J Clin Nutr 2022;116(2):297-298 [FREE Full text] [doi: 10.1093/ajcn/nqac120] [Medline: 35679433]
10. Senecal C, Widmer RJ, Larrabee BR, de Andrade M, Lerman LO, Lerman A, et al. A digital health weight loss program in 250,000 individuals. J Obes 2020;2020:9497164 [FREE Full text] [doi: 10.1155/2020/9497164] [Medline: 32300485]
11. Wang SV, Sreedhara SK, Schneeweiss S, REPEAT Initiative. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. Nat Commun 2022;13(1):5126 [FREE Full text] [doi: 10.1038/s41467-022-32310-3] [Medline: 36045130]

XSL•FO
RenderX

12. Stern AD, Brönneke J, Debatin JF, Hagen J, Matthies H, Patel S, et al. Advancing digital health applications: priorities for innovation in real-world evidence generation. Lancet Digit Health 2022;4(3):e200-e206 [FREE Full text] [doi: 10.1016/S2589-7500(21)00292-2] [Medline: 35216754]
13. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
14. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
15. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017;19(12):e402 [FREE Full text] [doi: 10.2196/jmir.8578] [Medline: 29212630]
16. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
17. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
18. Painter SL, Ahmed R, Hill JO, Kushner RF, Lindquist R, Brunning S, et al. What matters in weight loss? an in-depth analysis of self-monitoring. J Med Internet Res 2017;19(5):e160 [FREE Full text] [doi: 10.2196/jmir.7457] [Medline: 28500022]
19. Hales S, Turner-McGrievy GM, Wilcox S, Davis RE, Fahim A, Huhns M, et al. Trading pounds for points: engagement and weight loss in a mobile health intervention. Digit Health 2017;3:2055207617702252 [FREE Full text] [doi: 10.1177/2055207617702252] [Medline: 29942590]
20. Schirmann F, Kanehl P, Jones L. What intervention elements drive weight loss in blended-care behavior change interventions? a real-world data analysis with 25,706 patients. Nutrients 2022;14(14):2999 [FREE Full text] [doi: 10.3390/nu14142999] [Medline: 35889956]
21. Lavikainen P, Mattila E, Absetz P, Harjumaa M, Lindström J, Järvelä-Reijonen E, et al. Digitally supported lifestyle intervention to prevent type 2 diabetes through healthy habits: secondary analysis of long-term user engagement trajectories in a randomized controlled trial. J Med Internet Res 2022;24(2):e31530 [FREE Full text] [doi: 10.2196/31530] [Medline: 35200147]
22. Lungu A, Jun JJ, Azarmanesh O, Leykin Y, Chen CEJ. Blended care-cognitive behavioral therapy for depression and anxiety in real-world settings: pragmatic retrospective study. J Med Internet Res 2020;22(7):e18723 [FREE Full text] [doi: 10.2196/18723] [Medline: 32628120]
23. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42(8):1-28 [FREE Full text] [doi: 10.18637/jss.v042.i08]
24. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2017;15(3):199-236 [FREE Full text] [doi: 10.1093/pan/mpl013]
25. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. URL: https://www.R-project.org/ [accessed 2024-04-25]
26. Yoshida K, Bartel A. Tableone: create 'table 1' to describe baseline characteristics with or without propensity score weights. CRAN. 2022. URL: https://CRAN.R-project.org/package=tableone [accessed 2024-04-25]
27. Regulation (EU) 2017/745 of the European parliament and of the council. The European Parliament and the Council of the European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745&from=DE#d1e5060-1-1 [accessed 2023-08-23]
28. Senecal C, Collazo-Clavell M, Larrabee BR, de Andrade M, Lin W, Chen B, et al. A digital health weight-loss intervention in severe obesity. Digit Health 2020;6:2055207620910279 [FREE Full text] [doi: 10.1177/2055207620910279] [Medline: 32180992]
29. Antoun J, Itani H, Alarab N, Elsehmawy A. The effectiveness of combining nonmobile interventions with the use of smartphone apps with various features for weight loss: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022;10(4):e35479 [FREE Full text] [doi: 10.2196/35479] [Medline: 35394443]
30. Chew HSJ, Koh WL, Ng JSHY, Tan KK. Sustainability of weight loss through smartphone apps: systematic review and meta-analysis on anthropometric, metabolic, and dietary outcomes. J Med Internet Res 2022;24(9):e40141 [FREE Full text] [doi: 10.2196/40141] [Medline: 36129739]
31. Hauner H, Moss A, Berg A, Bischoff SC, Colombo-Benkmann M, Ellrott T, et al. Interdisziplinäre Leitlinie der Qualität S3 zur "Prävention und Therapie der Adipositas": der Deutschen Adipositas-Gesellschaft e.V.; der Deutschen Diabetes Gesellschaft; der Deutschen Gesellschaft für Ernährung e.V.; der Deutschen Gesellschaft für Ernährungsmedizin e.V. Version 2.0 (April 2014); AWMF-Register Nr. 050-001. Adipositas Ursachen Folgeerkrankungen Ther 2017;08(04):179-221. [doi: 10.1055/s-0037-1618857]

XSL•FO
RenderX

32. Yumuk V, Tsigos C, Fried M, Schindler K, Busetto L, Micic D, et al. European guidelines for obesity management in adults. Obes Facts 2015;8(6):402-424 [FREE Full text] [doi: 10.1159/000442721] [Medline: 26641646]
33. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences, 4th Edition. Thousand Oaks, California, US: SAGE Publications, Inc; 2013.
34. Mishra V, Künzler F, Kramer JN, Fleisch E, Kowatsch T, Kotz D. Detecting receptivity for mHealth interventions in the natural environment. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021;5(2):74 [FREE Full text] [doi: 10.1145/3463492] [Medline: 34926979]
35. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
36. Burnham JP, Lu C, Yaeger LH, Bailey TC, Kollef MH. Using wearable technology to predict health outcomes: a literature review. J Am Med Inform Assoc 2018;25(9):1221-1227 [FREE Full text] [doi: 10.1093/jamia/ocy082] [Medline: 29982520]

Abbreviations

ICD-10*: *International Classification of Diseases, Tenth Revision
MDR: Medical Device Regulation
PC: principal component
PCA: principal component analysis
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by G Eysenbach, T Leung, T de Azevedo Cardoso; submitted 03.01.23; peer-reviewed by A Parks, J Antoun; comments to author 14.04.23; revised version received 02.10.23; accepted 23.03.24; published 07.06.24.

Please cite as:
Lehmann M, Jones L, Schirmann F
App Engagement as a Predictor of Weight Loss in Blended-Care Interventions: Retrospective Observational Study Using Large-Scale Real-World Data
J Med Internet Res 2024;26:e45469
URL: https://www.jmir.org/2024/1/e45469
doi:10.2196/45469
PMID:38848556

©Marco Lehmann, Lucy Jones, Felix Schirmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of an Artificial Intelligence–Based Tailored Mobile Intervention for Nurse Burnout: Single-Arm Trial

Aram Cho[1], MSN, RN; Chiyoung Cha[1], PhD, RN; Gumhee Baek[1], MSN, RN

College of Nursing & Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, Republic of Korea

Corresponding Author:
Chiyoung Cha, PhD, RN
College of Nursing & Graduate Program in System Health Science and Engineering
Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu, Seoul
Hellen #202
Seoul, 03760
Republic of Korea
Phone: 82 0232772883
Email: chiyoung@ewha.ac.kr

Abstract

Background: Nurse burnout leads to an increase in turnover, which is a serious problem in the health care system. Although there is ample evidence of nurse burnout, interventions developed in previous studies were general and did not consider specific burnout dimensions and individual characteristics.

Objective: The objectives of this study were to develop and optimize the first tailored mobile intervention for nurse burnout, which recommends programs based on artificial intelligence (AI) algorithms, and to test its usability, effectiveness, and satisfaction.

Methods: In this study, an AI-based mobile intervention, Nurse Healing Space, was developed to provide tailored programs for nurse burnout. The 4-week program included mindfulness meditation, laughter therapy, storytelling, reflective writing, and acceptance and commitment therapy. The AI algorithm recommended one of these programs to participants by calculating similarity through a pretest consisting of participants' demographics, research variables, and burnout dimension scores measured with the Copenhagen Burnout Inventory. After completing a 4-week program, burnout, job stress, stress response using the Stress Response Inventory Modified Form, the usability of the app, coping strategy by the coping strategy indicator, and program satisfaction (1: very dissatisfied; 5: very satisfied) were measured. The AI recognized the recommended program as effective if the user's burnout score reduced after the 2-week program and updated the algorithm accordingly. After a pilot test (n=10), AI optimization was performed (n=300). A paired 2-tailed *t* test, ANOVA, and the Spearman correlation were used to test the effect of the intervention and algorithm optimization.

Results: Nurse Healing Space was implemented as a mobile app equipped with a system that recommended 1 program out of 4 based on similarity between users through AI. The AI algorithm worked well in matching the program recommended to participants who were most similar using valid data. Users were satisfied with the convenience and visual quality but were dissatisfied with the absence of notifications and inability to customize the program. The overall usability score of the app was 3.4 out of 5 points. Nurses' burnout scores decreased significantly after the completion of the first 2-week program (t=7.012; P<.001) and reduced further after the second 2-week program (t=2.811; P=.01). After completing the Nurse Healing Space program, job stress (t=6.765; P<.001) and stress responses (t=5.864; P<.001) decreased significantly. During the second 2-week program, the burnout level reduced in the order of participation (r=–0.138; P=.04). User satisfaction increased for both the first (F=3.493; P=.03) and second programs (F=3.911; P=.02).

Conclusions: This program effectively reduced burnout, job stress, and stress responses. Nurse managers were able to prevent nurses from resigning and maintain the quality of medical services using this AI-based program to provide tailored interventions for nurse burnout. Thus, this app could improve qualitative health care, increase employee satisfaction, reduce costs, and ultimately improve the efficiency of the health care system.

(J Med Internet Res 2024;26:e54029) doi:10.2196/54029

XSL•FO
RenderX

KEYWORDS

artificial intelligence; burnout; mobile app; nurses; nurse; mHealth; mobile health; app; apps; applications; usability; satisfaction; effectiveness; tailored; mind-body; meditation; mindfulness; ACT; algorithm; algorithms; occupational health; digital health; recommender; optimization; acceptance and commitment therapy; job; worker; workers; stress; employee; employees

Introduction

Background

Burnout is a physical and mental exhaustion syndrome that includes phenomena such as a negative self-concept and work attitude and loss of interest in patients. This is a form of stress response that occurs when stress can no longer be tolerated. Burnout occurs simultaneously in 3 dimensions [1], that is, personal, work-related, and client-related dimensions [2]. The World Health Organization has defined burnout as a work-related risk factor that affects health and has included it in the 11th edition of the *International Classification of Diseases* in 2019 [3].

Nurses are known to experience high levels of burnout. The prevalence of nurse burnout has been approximately 30% worldwide in the last 10 years [4]. With the COVID-19 pandemic, nurses' workload has increased leading to increased burnout. Nurse burnout is a serious issue in the health care system as it leads to an increase in the nurse turnover rate [5]. With increased turnover, the nursing workforce has to be replaced with new nurses, resulting in reduced work productivity and an increased burden on hospital finances [6]. In addition, burnout negatively influences the quality of nursing services and leads to frequent absences [7,8].

Acknowledging the significance of nurse burnout, various programs have been developed and tested. In a study that reviewed burnout reduction interventions for nurses and doctors, interventions such as yoga, meditation, and mindfulness reportedly alleviated burnout [9]. However, the program was conducted in person, and interventions for medical personnel can be burdensome due to time and location constraints [10]. Moreover, as nurses work in shifts, there are time and space restrictions for interventions. However, web-based interventions do not require a facilitator and are, therefore, better than in-person interventions in terms of efficiency [11].

In previous studies, interventions through mobile apps to confirm the effect of burnout were based on voice programs [12,13]. Previous studies on burnout interventions introduced one-size-fits-all programs. However, nurses experience varying dimensions of burnout. Artificial intelligence (AI)–based recommendation systems based on similarity are widely used in various fields. Netflix and YouTube recommend similar media based on individual content preferences [14,15]. In addition, in the marketing field, similar products are recommended by analyzing the commonalities of products that consumers have purchased [16]. As the accuracy and diversity of recommendations increase, user satisfaction increases [17]. Recommendation systems commonly use a user-based collaborative filtering system, which recommends items or content based on similarities between users' interest profiles [18]. Digital Me, an AI-based product service system, measures a user's state through user behavior analysis, predicts it, and makes recommendations. The recommendation system maximizes a specific score or finds similar but superior existing users, thus attaining its goal [19]. Therefore, a similar system could be used to provide a tailored program for nurse burnout.

Nurses' burnout is influenced by various factors, such as their demographic and job-related characteristics [20-22]. Burnout patterns differ depending on coping strategies, whether working night shifts, working at a tertiary hospital, age, and department [23-25]. Therefore, when recommending burnout reduction programs for nurses, we could consider that similar burnout experiences are shared among individuals with similar general and job-related characteristics, rather than simply relying on user preferences or uniform delivery. The system we have developed measures the similarity between users based on demographics, research variables (job stress, stress response, and coping strategy), and burnout dimension scores.

Literature Review

Cognitive behavioral therapy (CBT)–based interventions are evidence-based interventions to reduce burnout and stress in health care professionals and improve mental health, healthy lifestyle beliefs, and job satisfaction [26]. Several programs have been suggested to reduce and prevent nurse burnout. One of these, mindfulness meditation, is a well-known technique for managing stress and burnout. Mindfulness is the attitude of paying attention to present moment events and accepting feelings, ideas, and sensations without passing judgment [27]. Mindfulness-based cognitive therapy is a group psychotherapy program that combines mindfulness and CBT and has been proven to effectively improve depression and prevent its recurrence [28]. Mindfulness-based training is effective in reducing burnout not only in nurses [29] but also in other health care professionals [30,31]. In addition, mindfulness meditation has been reported to enhance positive cognitive retraining by increasing positive mood and brain activity [32]. A recent study demonstrated that a website and a smartphone app for mindfulness were effective in reducing burnout among health care professionals. [33].

Another therapeutic intervention is laughter therapy. Laughter therapy improves healing and coping abilities in physical, psychological, and cognitive aspects by inducing laughter, smiles, and pleasant emotions through physical and intellectual interactions with the body and the mind [34]. It is well known that laughter therapy can reduce psychosocial conditions, such as stress, anxiety, and depression, in people with health issues, such as patients with cancer, middle-aged women, and mothers [35-37]; it has also recently been reported as an effective way to improve life satisfaction and reduce burnout in nurses [38]. Using humor to deal with stressful and burnout situations can increase happiness and life satisfaction in individuals by fostering more positive evaluations [39].

Storytelling, one of the most fundamental and traditional forms of communication, can influence behavior by allowing people

XSL•FO
RenderX

to express human identity and self-identify, share similar life experiences, and explain situations [40]. A form of CBT called "narrative CBT" uses storytelling [41]. The use of storytelling among nurses provides a space for nurses to talk about, listen to, perceive, and evaluate their experiences [42]. In previous studies, nurses who took part in storytelling workshops gave storytelling a positive evaluation [43]. In addition, expressive writing has been shown to improve mood, cognitive function, and mental health problems [44,45]. In a recent study that used storytelling and reflective writing as an intervention for oncology nurses, self-awareness and self-compassion reportedly increased, and insomnia and loneliness decreased [46]. Therefore, storytelling and reflective writing can improve personal well-being by providing skills to recognize one's emotions and cope with work-related emotions.

Acceptance and commitment therapy (ACT) is part of the third wave of CBT [47]. It is a comprehensive psychological therapeutic strategy that can be used to treat a variety of problems and disorders. It incorporates acceptance and mindfulness processes, active engagement, and behavioral change processes [48]. Recently, it was provided to a group of employees in the workplace [49] and was verified as an intervention for stress and burnout management for various medical staff, including nurses [50]. It has been suggested that ACT may have the potential to reduce burnout among various occupational groups [51].

Objective

The purpose of this study was to develop and optimize an AI-based tailored mobile intervention, Nurse Healing Space, to reduce nurse burnout. Specifically, we designed an AI algorithm that provides tailored burnout reduction programs using user similarity scores and assessed the usability of the intervention. After optimizing the AI algorithm, we tested the effectiveness of the program for burnout, job stress, stress response, coping strategies, and user satisfaction.

Methods

Research Overview

We developed the mobile app, Nurse Healing Space, from December 2021 to July 2022 and then conducted a pilot test in August 2022 to evaluate the usability and test the operation of the app. From September 2022 to February 2023, 300 people were recruited to accumulate data and optimize algorithms. Before starting the program, we instructed participants on how to participate. This was necessary to prevent dropout as the program's duration was 4 weeks (pretest, program 1 first and second weeks, posttest 1, program 2 first and second weeks, and posttest 2), with a total of 7 stages. A total of 325 potential participants showed interest in the study, of whom, 300 completed the intervention.

Nurse Healing Space

Overview

Nurse Healing Space is a 4-week intervention that consists of two 2-week programs to reduce burnout (Figure 1). Users fill out questionnaires (pretest), and the AI recommends 1 of 4 programs (program 1) based on the similarity of users. After the completion of program 1, users are tested for burnout and program satisfaction (posttest 1). The AI recommends a second program out of the 3 remaining programs (program 2) based on the similarity of users from posttest 1. After completion of program 2, users fill out another questionnaire (posttest 2).

Each program consists of 6 sessions of 10-15 minutes each. The animated female character in the video guides participants (Figure 2). All the programs were developed by the research team in consultation with experts in each field. Below are the descriptions of the four 2-week programs.

Figure 1. Nurse Healing Space operation process. ACT: acceptance and commitment therapy.

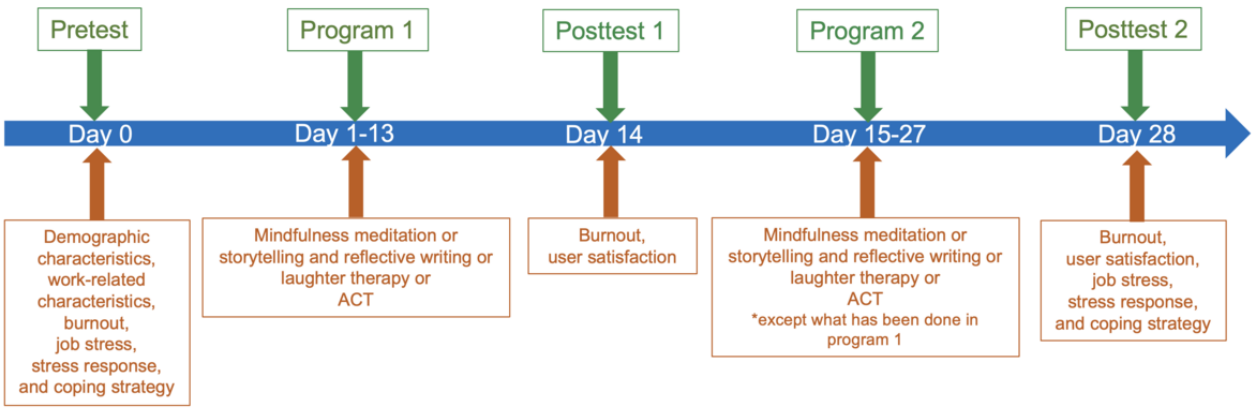

XSL•FO
RenderX

Figure 2. Scenes from the program.

Mindfulness Meditation Program

Mindfulness meditation is conducted by the fictional nurse character guiding the program through narration and demonstration of movements. The contents for mindfulness meditation were organized based on Kabat-Zinn's [27] mindfulness meditation by a researcher (CC) who had previously developed mindfulness programs and reviewed them by an expert. Each session consists of an intro, mindfulness meditation, and wrap-up. The intro explains the mindset before starting the mindfulness meditation. Mindfulness meditation is composed of a mixture of the following 2 in each session: body scan, sitting up straight meditation, relaxation meditation, yoga meditation, and self-compassion meditation. The wrap-up explains ways to apply mindfulness meditation to everyday life.

Storytelling and Reflective Writing Program

In storytelling, a fictional nurse narrates her story of burnout and coping strategies during each session. The topics in the sessions consist of "Work is piling up," "I am a person too," "There aren't enough hours in the day," and "Am I an emotional outlet?" The work environment is poor, with not enough nurses to take care of patients. After the storytelling, the fictional nurse asks participants to reflect on their experiences and write about them. Participants can freely share their reflective writings on the bulletin board of the app.

Laughter Therapy Program

For laughter therapy, a fictional nurse guides the laughter exercise with background music. Each session consists of an intro, laughter exercise, and wrap-up. The contents were organized based on the laughter therapy program of the Korean Society of Laughter Clinic and were reviewed by an expert. The introductory part consists of a brief explanation of laughter therapy and prelaughter exercises. Each laughter exercise consists of 2 exercises with exciting music and the laughter follows the rhythm. The wrap-up consists of self-expression and laughter meditation.

ACT Program

In ACT, a fictional nurse guides the ACT by narration and showing images for cognitive change by using various techniques, such as the breathing technique and music meditation. The contents were organized based on the ACT founder's book [52] and reviewed by an expert. The ACT consists of "Hello my heart" (introduction), "Thinking is just thinking" (cognitive diffusion), "This is the moment" (being present in the moment), "Look back at me" (self as context), "Worthy life" (the best 3 moments of my life), and "Let's do it together" (committed action and willingness).

AI-Based Program Recommendation System

When users fill out the pretest, their responses are calculated and stored as scores. One of 4 programs that were effective for a previous user who is most similar to the new user, is recommended to the new user. We define similar users as having similar demographic and work-related characteristics and burnout dimensions. After the completion of program 1, users are tested for burnout dimensions and program satisfaction. Thereafter, users are matched with program 2, which is 1 of 3 remaining programs that were evaluated as effective for burnout by the most similar users based on similarity scores for demographic and work-related characteristics and burnout dimensions. After the completion of program 2, users fill out posttest 2. The operation of the Nurse Healing Space app is shown in Figure 3.

XSL•FO
RenderX

Figure 3. Nurse Healing Space operation.

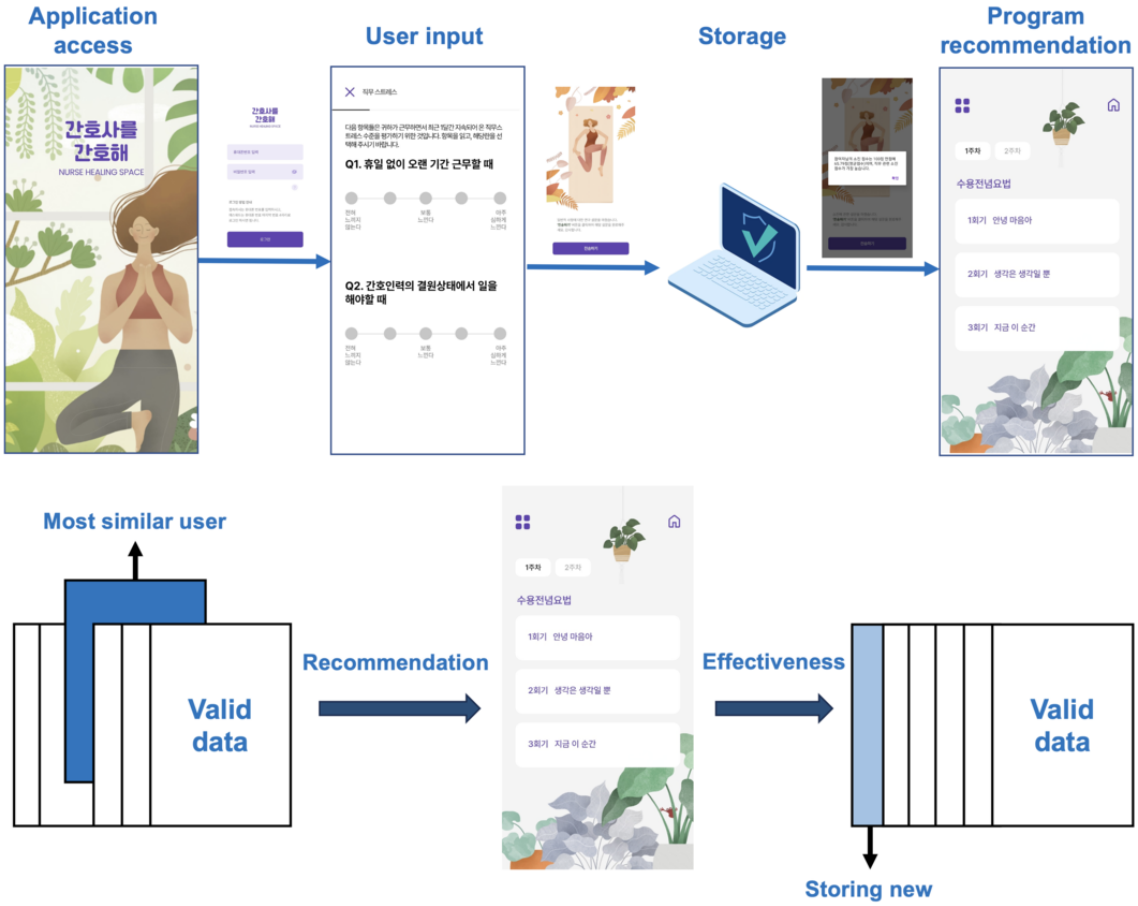

Measuring Similarity Between Users

We used a pilot test with 10 nurses to collect initial data using the preallocated program according to the highest subdimension recommended. With 10 sets of existing data stored through the pilot test, when new users are entered, the algorithm compares demographic and work-related characteristics, job stress, stress response, and coping strategies with the stored data to find the most similar existing users with a matching subdimension of burnout. The search method is as follows.

In the recommendation system, the existing user is denoted as i and the new user as j. Zero is assigned to the element if the value of the k variable of the user i and the k variable of the user j matches, and 1 is assigned to the inconsistent element. Specifically, the similarity of each element of the categorical variable can be determined.

(1)

d_{ijk_x} refers to the similarity of each element of the categorical variable; i is an existing user, j is a new user, and x is an element included in the categorical information. Thereafter, the recommendation system calculates the similarity of each categorical variable by adding the similarity of each element.

The recommendation system calculates the similarity between the users and the existing data by applying the weight of each element to the numerical variable. The numerical type is calculated by dividing the absolute value of the difference between the value of the k variable of user i and the value of the k variable of user j by the range of the k variable. Specifically, the similarity for each element in the numerical type of information is determined.

(2)

indicates the similarity of each element in a numerical variable, i is an existing user, j is a new user, x is an element included in the numerical information, and R may indicate the range of the variable. Thereafter, the recommendation system calculates the similarity of the numerical variable by adding the similarity for each element.

Table 1 shows that when a new user is entered, a similarity between the new user and the existing users A, B, and C is calculated. The recommendation system determines the similarity between the new user and the existing users by calculating the similarity scores from 5 categorical variables and 17 numerical variables. Similarity is based on the sum of each value. The smaller the sum of the similarity level, the higher the similarity with the new user. The possible ranges of

XSL•FO
RenderX

similarity scores are between 0 and 22. For example, in Table 1, user B has the lowest similarity scores and is determined as the user that is most similar to the new user.

Table 1. Measuring similarity between users.

Variable	Raw score			New user	Similarity calculation		
	User A	User B	User C		User A	User B	User C
Categorical data							
Sex	1	1	1	1	1	1	1
Marital status	1	2	1	3	1	1	1
Continuous data							
Age	30	45	25	45	0.39	0.00	0.53
Hospital size	5	3	4	5	0.00	0.50	0.25
Clinical experiences	1	1	2	3	0.40	0.40	0.20
Similarity	__[a]	—	—	—	8.03	5.96	7.64

[a]Not applicable.

Measurement

Burnout was measured using Copenhagen Burnout Inventory (CBI) [2], which was translated into the Korean version. Cronbach α at the time of translation was 0.93 [53]. The CBI consists of 19 items that measure 3 subdomains: personal burnout (6 items, ranging between 0 and 31.5), work-related burnout (7 items, ranging between 0 and 37), and client-related burnout (6 items, ranging between 0 and 31.5). Possible scores range between 0 and 100, with higher scores indicating higher burnout. In this study, Cronbach α was 0.94.

Job stress was measured with an instrument developed for Korean nurses [54]. This instrument consists of 23 items. The scores range from 23 to 115, with a higher score indicating greater job stress. In this study, Cronbach α was 0.92. Stress response was measured using the Stress Response Inventory Modified Form [55], which consists of 22 items. Possible scores ranged between 0 and 88, with higher scores indicating higher levels of stress responses. In this study, Cronbach α was 0.95.

Coping strategy was measured using the Coping strategy indicator [56], which was translated into the Korean version. Cronbach α at the time of translation was 0.84 [57]. The instrument consists of 33 items. The possible range of scores was 33-99. In this study, Cronbach α was 0.89.

A rating scale was used to evaluate the usability of apps [58]. This measure consisted of 23 items in 5 areas—engagement, functionality, aesthetics, information, and subjective mobile app quality. The possible range of scores was 26-130, with higher scores indicating higher usability and quality. App satisfaction was measured with 1 item using a 5-point Likert-type rating scale (1: very dissatisfied; 5: very satisfied).

Valid Data

Data are evaluated as valid when burnout dimension scores after program completion decrease by at least 5 points on a 0-100 scale based on the criterion suggested in a previous study [2]. Having valid data implies that the program effectively lowers burnout and data are stored in the AI algorithm.

Table 2 illustrates an example of the process of determining valid data and the program provided to each user with the change in burnout scores after program completion. For example, user A achieved a 5-point reduction in personal burnout after completion of program 1; this was classified as valid data and included in the algorithm. However, after the completion of program 2, the burnout scores of user A reduced for all burnout dimensions by less than 5 points; this was considered invalid and not stored in the algorithm. None of user B's data were stored because user B did not have valid data. For user C, burnout scores for personal and work-related burnout decreased by 5 points after completion of program 1. For 2 burnout subdimensions the score reduced by more than 5 points. In this case, it was considered that burnout had decreased as a result of other causes, and the data were regarded as invalid. For user D, the personal burnout score reduced by 5 points after completion of program 1 and work-related burnout score by 6 points after completion of program 2; thus, their data were stored as valid data.

XSL•FO
RenderX

Table 2. Valid data selection criteria.

	Personal burnout score	Work-related burnout score	Client-related burnout score	Program number	Data type
User A					
Pretest	45	40	39	__[a]	—
Reduction after program 1	5	3	4	Program 1	Valid
Posttest 1	40	37	35	—	—
Reduction after program 2	4	2	2	Program 2	Invalid
Posttest 2	36	35	33	—	—
User B					
Pretest	45	40	39	—	—
Reduction after program 1	3	3	4	Program 1	Invalid
Posttest 1	42	37	35	—	—
Reduction after program 2	2	0	2	Program 2	Invalid
Posttest 2	40	37	33	—	—
User C					
Pretest	45	40	39	—	—
Reduction after program 1	5	5	4	Program 1	Invalid
Posttest 1	40	35	35	—	—
Reduction after program 2	0	2	2	Program 2	Invalid
Posttest 2	40	33	33	—	—
User D					
Pretest	45	40	39	—	—
Reduction after program 1	5	3	4	Program 1	Valid
Posttest 1	40	37	35	—	—
Reduction after program 2	4	6	2	Program 2	Valid
Posttest 2	36	31	33	—	—

[a]Not applicable.

AI Algorithm

As valid data are continuously accumulated, the recommendation system learns and updates the artificial neural network. The recommendation system determines the most similar user from the user information using the learned artificial neural network. The artificial neural network determines similar users using all the similarity scores of the valid data, and programs that have been effective in reducing burnout for similar users are provided to new users. Thereafter, the artificial neural network compares the degree of change in the pre- and posttest burnout scores of the new user to determine whether the program provided was effective for the new user and stores data about the user if it is determined to be valid. The recommendation system updates the artificial neural network by continuously updating valid data. Thus, the artificial neural network can acquire valid data for more users over time, and by determining users that are similar to new users, it provides more effective programs to new users over time. To ensure that the algorithm operates as intended, we extracted valid data from 10 pilot tests and 299 optimizations. Subsequently, we manually calculated the similarity score of the 300th participant using Excel and checked the suggested program and compared it with what was suggested to the user.

Data Collection and Setting

We used 2 large web-based nursing communities with 94,000, and 170,000 members for data collection. We posted an advertisement on 2 large web-based nursing communities in August 2022 to recruit participants for a pilot test. For the pilot test, we adopted a minimum of 12 participants recommended by Whitehead et al [59]; however, only 10 participants were able to complete the pilot test. After the pilot test, we posted advertisements on 2 large web-based nursing communities between September 2022 and February 2023 to recruit participants for AI algorithm optimization. The number of participants required for measuring the intervention effect was calculated based on the paired 2-tailed *t* test for measuring the effect of burnout, a research variable of this study. Using G-power (version 3.1.9.4) with an effect size of 0.2, significance level of .05, and power of 0.95, a total of 272 people were calculated. Due to the insufficiency of previous studies on

XSL•FO
RenderX

developing mobile interventions for nursing burnout, the effect size was conservatively considered at a low threshold to achieve treatment effects. A total of 325 potential participants showed interest in the study, of whom, 300 completed the study. Their data were used for AI algorithm optimization. The inclusion criteria for the participants were nurses who had been providing care independently at hospitals for the past month at the time of data collection.

Ethical Considerations

This study was reviewed and approved by the institutional review board (ewha-202208-0005-01) at the principal investigator's institution. Informed consent was obtained from all participants. Participants received information about the study's goals, anonymity, confidentiality, and their right to withdraw from the study at any time before signing the online informed consent form. As compensation for their participation, participants were provided with mobile coupons valued at US $3.60. The collected data were encrypted and stored on a computer accessible only to the researchers.

Statistical Analysis

We used software SPSS (version 23.0; IBM Corp) to facilitate data analysis. Descriptive statistics were used to report sociodemographic and work-related characteristics and study variables. All variables met the assumption for normal distribution: skewness of 1 or less and kurtosis of 5 or less [60]. A paired *t* test was conducted on burnout, job stress, stress response, and coping strategies before and after the program. Spearman correlation was performed to examine changes in research variables and burnout reduction based on data accumulation. Nonparametric analysis (Spearman correlation) was used because data accumulation is considered an ordinal variable in that it follows the order of participation. ANOVA was conducted on satisfaction change by grouping participants into 3 groups: the first 100 participants, the second 100 participants, and the third 100 participants; the Scheffe test was performed for post hoc analysis.

Results

Algorithm Operation

The 300th participant had the highest work-related burnout and client-related burnout among the burnout dimensions, and laughter therapy was recommended. The similarity scores between the 300th participant and all the accumulated valid data ranged between 1.17 and 8.14. It was confirmed that the program completed by the participant with the lowest similarity was laughter therapy, which was recommended for the 300th participant. The highest burnout subtype for the most similar participant was work-related burnout, and the participant's data were classified as valid data because of the effect of laughter therapy. The calculated similarity with the participant was 1.17. To make it more indisputable, we checked whether the similarity system worked in the same way for the 299th participant, and it was confirmed that the program recommended to the most similar participant among the valid data was the same.

Valid Data

Of the total data, 45.7% (274/600) were selected as valid data. Personal burnout was the highest subdimension with 57.4% of valid data and work-related burnout was the next highest subdimension with 25.9% of valid data, followed by client burnout with 40% of valid data including duplicates. In terms of programs, laughter therapy accounted for the largest proportion of valid data, with 32.8%, followed by 28.1% for mindfulness meditation, 20.4% for ACT, and 18.6% for storytelling and reflective writing.

Usability Assessment

Users scored ease of use, good visual design, and high quality of content graphics more than 4 out of 5. However, users felt most uncomfortable with the absence of reminders and notification functions and the inability to interact with the app. Moreover, they were not satisfied with the fact that the app could not be set up to suit their preferences. The overall score for this app was 3.4 out of 5 on a Likert scale. Furthermore, the total mean score for each question was 3.4 out of 5 on a Likert scale.

Intervention Effects

The mean age of the participants was 32.8 (SD 5.75) years. Of these, 97.7% (293/300) were female, 87% (261/300) were staff nurses, and 62.3% (187/300) worked in hospitals with more than 500 beds (Table 3).

The decrease in burnout was statistically significant, from 62.36 (SD 18.0) at pretest to 55.22 (SD 18.79) at posttest 1 (t=7.012; P<.001) and 52.70 (SD 20.15) at posttest 2 (t=8.692; P<.001). Reduction in burnout scores from posttest 1 to posttest 2 was also statistically significant (t=2.811; P=.01). Moreover, burnout scores decreased significantly before the program and after the first and second programs (t=8.692; P<.001). Personal burnout decreased statistically from 20.40 (SD 6.39) at pretest to 18.15 (SD 6.38) at posttest 1 (t=6.134; P<.001) and 17.30 (SD 7.33) at posttest 2 (t=7.535; P<.001). In addition, the reduction in personal burnout from posttest 1 to posttest 2 was statistically significant (t=2.597; P=.01). Work-related burnout decreased statistically from 22.22 (SD 6.39) at pretest to 19.83 (SD 6.49) at posttest 1 (t=6.816; P<.001) and 19.06 (SD 6.68) at posttest 2 (t=8.286; P<.001). The reduction in work-related burnout from posttest 1 to posttest 2 was also statistically significant (t=2.306; P=.02). Client-related burnout decreased statistically from 19.72 (SD 6.63) at pretest to 17.22 (SD 6.91) at posttest 1 (t=6.439; P<.001) and 16.32 (SD 7.15) at posttest 2 (t=8.329; P<.001). Reduction in client-related burnout from posttest 1 to posttest 2 was also statistically significant (t=2.704; P=.01).

Job stress, stress response, and coping strategy were measured twice before the program and after the first and second programs. After the completion of the two 2-week programs, the participants reported a statistically significant reduction in job stress (t=6.765; P<.001) and stress response (t=5.864; P<.001) (Table 4).

XSL•FO
RenderX

Table 3. General characteristics of participants (N=300).

Characteristic	Values
Age (years), mean (SD)	32.81 (5.754)
Gender (female), n (%)	293 (97.7)
Marital status, n (%)	
Married	178 (59.3)
Single	118 (39.3)
Divorced or separated	4 (1.3)
Position, n (%)	
Staff nurse	261 (87)
Charge nurse	35 (11.7)
Head nurse or team leader	4 (1.3)
Hospital size (beds), n (%)	
<30	7 (2.3)
31-100	25 (8.3)
101-300	63 (21)
301-500	18 (6)
≥500	187 (62.3)
Clinical experience (months), mean (SD)	95.83 (69.66)
Working department, n (%)	
General ward	172 (57.3)
ICU[a]	16 (5.3)
Special units[b]	30 (10)
Outpatient or clinic or administration	63 (21)
Others	19 (6.3)
Overtime during the past month (hours), n (%)	
None	41 (13.7)
≤0.5	78 (26)
0.5-1	77 (25.7)
1-1.5	50 (16.7)
1.5-2	17 (5.7)
>2	37 (12.3)
Type of shift, n (%)	
8-hour shift	180 (60)
12-hour shift	18 (6)
Fixed	101 (33.7)
Others	1 (0.3)
Turnover intention (range of score: 0-10), mean (SD)	4.88 (2.53)

[a]ICU: intensive care unit.

[b]Special units include emergency rooms, operating rooms, recovery rooms, and delivery rooms.

XSL•FO
RenderX

Table 4. Effectiveness of intervention by variables (N=300).

Variables	Pretest score, mean (SD)	Posttest 1 score, mean (SD)	Posttest 2 score, mean (SD)	Pretest vs posttest 1		Posttest 1 vs posttest 2		Pretest vs posttest 2	
				t (df=299)	*P* value	*t (df=299)*	*P* value	*t (df=299)*	*P* value
Burnout[a]	62.36 (18.00)	55.22 (18.79)	52.70 (20.15)	7.012	<.001	2.811	.01	8.692	<.001
Personal[b]	20.40 (6.39)	18.15 (6.38)	17.30 (7.33)	6.134	<.001	2.597	.01	7.535	<.001
Work-related[c]	22.22 (6.39)	19.83 (6.49)	19.06 (6.68)	6.816	<.001	2.306	.02	8.286	<.001
Client-related[d]	19.72 (6.63)	17.22 (6.91)	16.32 (7.15)	6.439	<.001	2.704	.01	8.329	<.001
Job stress[e]	88.39 (12.65)	__[f]	81.85 (17.34)	—	—	—	—	6.765	<.001
Stress response[g]	42.00 (17.95)	—	35.56 (18.39)	—	—	—	—	5.864	<.001
Coping strategy[h]	70.25 (9.66)	—	70.97 (10.05)	—	—	—	—	–1.367	.17

[a]Burnout total score range is between 0 and 100.

[b]Personal burnout score range is between 0 and 31.5.

[c]Work-related burnout score range is between 0 and 37.

[d]Client-related burnout score range is between 0 and 31.5.

[e]Job stress score range is between 23 and 115.

[f]Not applicable.

[g]Stress response score range is between 0 and 88.

[h]Coping strategy score range is between 33 and 99.

For program 1, the most frequently recommended program was laughter therapy (46.3%), followed by storytelling and reflective writing (20%), ACT (18.3%), and mindfulness meditation (15.3%). For program 2, the most frequently recommended program was mindfulness meditation (41.3%), followed by laughter therapy (23.7%), storytelling and reflective writing (21.3%), and ACT (13.7%). The effect of each program in reducing burnout was as follows: laughter therapy (8.7%), storytelling and reflective writing (8.7%), ACT (12.3%), and mindfulness meditation (4.5%). Thus, among the 4 programs, ACT reduced burnout the most. Client-related burnout was decreased the most by ACT (18.2% decrease). Laughter therapy reduced personal burnout by 10.4%, which was the highest reduction. Work-related burnout was reduced the most through storytelling and reflective writing and ACT (both resulted in a reduction of 10.5%) (Table 5).

Table 6 shows the correlation with the order of participation and burnout reduction. The correlation of burnout differences between posttest 1 and pretest and the order of participation was 0.060 (*P*=.30). The correlation of burnout differences between posttest 2 and posttest 1 and the order of participation was 0.138 (*P*=.04). The correlation of job stress differences between pretest and posttest 2 and the order of participation was 0.134 (*P*=.03). The correlation of stress response differences between pretest and posttest 2 and the order of participation was 0.094 (*P*=.13). The correlation of coping strategy differences between pretest and posttest 2 and the order of participation was 0.140 (*P*=.08).

XSL•FO
RenderX

Table 5. Details for each program (N=300).

	Laughter therapy	Storytelling and reflective writing	ACT[a]	Mindfulness meditation
Program 1 (participants), n (%)	139 (46.3)	60 (20)	55 (18.3)	46 (15.3)
Program 2 (participants), n (%)	71 (23.7)	64 (21.3)	41 (13.7)	124 (41.3)
Burnout reduction				
Personal				
Score in points (%)	–6.9 (10.4)	–4.2 (8.0)	–4.9 (7.9)	–3.0 (4.9)
t (df)	5.403 (209)	2.020 (123)	2.635 (95)	2.266 (169)
P value	<.001	.046	.01	.03
Work-related				
Score in points (%)	–4.3 (7.2)	–5.6 (10.5)	–6.4 (10.5)	–2.2 (4.0)
t (df)	4.272 (209)	3.283 (123)	3.702 (95)	1.820 (169)
P value	<.001	.00	<.001	.07
Client-related				
Score in points (%)	–5.2 (8.7)	–3.8 (7.4)	–12.7 (18.2)	–2.7 (4.8)
t (df)	3.718 (209)	1.870 (123)	6.897 (95)	1.977 (169)
P value	<.001	.06	<.001	.05
Total				
Score in points (%)	–5.4 (8.7)	–4.6 (8.7)	–7.9 (12.3)	–2.6 (4.5)
t (df)	4.833 (209)	2.546 (123)	4.886 (95)	2.241 (169)
P value	<.001	.01	<.001	.03
Satisfaction, mean (SD)	4.58 (0.60)	4.69 (0.54)	4.52 (0.68)	4.61 (0.57)

[a]ACT: acceptance and commitment therapy.

Table 6. Correlation with the order of participation and research variables.

Stages	Order of participation
Burnout differences between posttest 1 and pretest	
R	0.060
P value	.30
Burnout differences between posttest 2 and posttest 1	
R	–0.138
P value	.04
Job stress differences between posttest 2 and pretest	
R	–0.134
P value	.03
Stress response differences between posttest 2 and pretest	
R	–0.094
P value	.13
Coping strategy differences between posttest 2 and pretest	
R	–0.140
P value	.08

XSL•FO
RenderX

Participant Satisfaction

When programs were compared, the satisfaction score for storytelling and reflective writing (4.69 points) was the highest, followed by mindfulness meditation (4.61 points), laughter therapy (4.58 points), and ACT (4.52 points) (Table 5). The satisfaction scores increased over time from participant 1 in October 2022 to participant 300 in January 2023 by optimization (Figure 4). By grouping participants into groups of 100, we were able to calculate the average satisfaction of programs 1 and 2 to understand how satisfaction changed. For program 1, satisfaction scores for the 3 groups increased from 4.46 to 4.60 and 4.69 (F=3.493; P=.03). The post hoc test showed statistically significant differences between the first 100 and the third 100 (P=.03). For program 2, the satisfaction score was 4.49 points for the initial 100 participants, it was 4.66 points for the next 100 participants, and the score was 4.70 for the last 100 participants; thus, the satisfaction for program 2 also gradually increased (F=3.911; P=.02). In addition, the difference in satisfaction with the second program was statistically significant as a result of the post hoc analysis between the first 100 and the third 100 participants (P=.03).

Figure 4. Changes in participant satisfaction. Group 1 included the first to the 100th participants, group 2 the 101st to the 200th, and group 3 the 201st to the 300th.

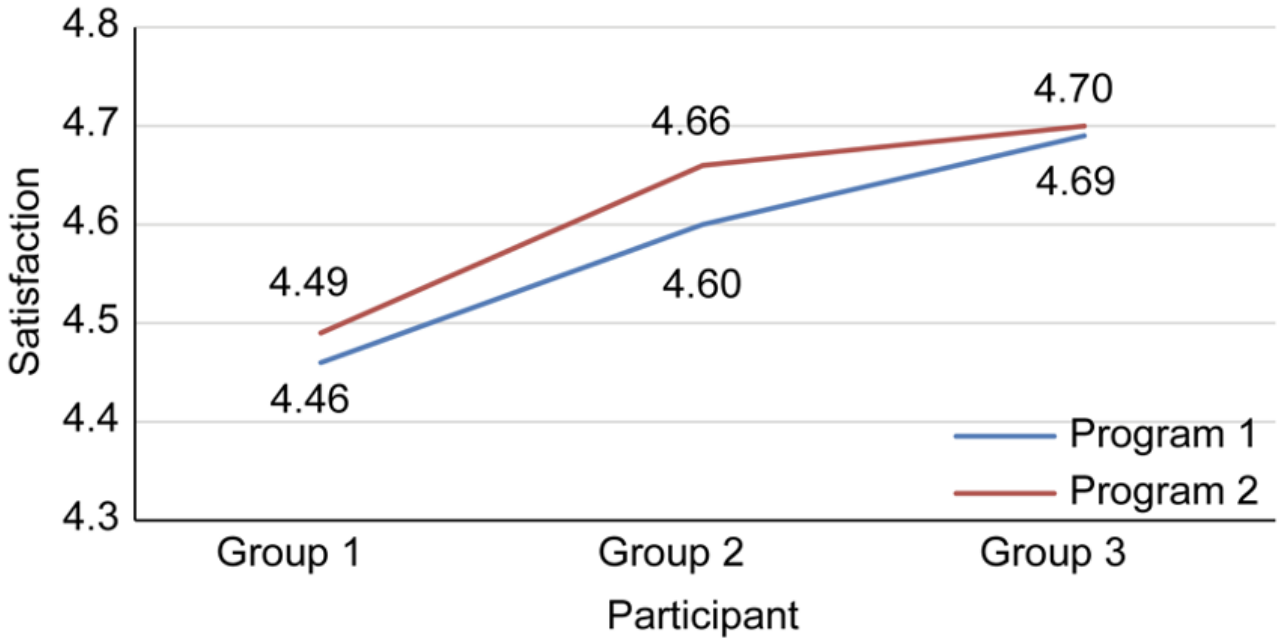

Discussion

Principal Findings

In this study, we developed an AI-based tailored mobile app and optimized the algorithm for nurse burnout. As a result of organizing and optimizing the system with an algorithm that recommends programs that have been effective for highly similar users, burnout decreased and satisfaction scores increased as data accumulated. Furthermore, when we examined whether the algorithm was being applied correctly, we verified that the system recommends the most similar valid data to new users, proving its accuracy.

The unique aspect of this study is the introduction of AI into a program recommendation system. Thus, our system is distinct from other such programs and has unique features in terms of its development. Several studies have used AI to predict or manage burnout based on physiological indicators or to predict the degree of burnout [61-64]. Although predictions and measurements were developed, no suitable strategy for reducing burnout was presented. However, 1 study used AI in stress management programs [65]. AI algorithms were used only to guide programs and programs were conducted in group settings, thus overlooking the aspect of diversity. To the best of our knowledge, no previous study has incorporated AI to provide tailored programs to users to reduce burnout.

Our AI-based algorithm was designed to effectively reduce burnout in each user by referring to similarity. As the algorithm was optimized with an accumulation of valid data, satisfaction increased significantly. In terms of burnout reduction, it was found that the change in the burnout value from posttest 1 to posttest 2 gradually increased as the algorithm was optimized. Thus, the recommendation for the second program was more optimal for burnout reduction.

The strength of our intervention is that it is a tailored program based on user input on their burnout dimensions and demographic data. Previous studies provided general burnout programs even though dimensions of nurse burnout have long been studied. Providing tailored intervention is effective in reducing the emotional exhaustion of nurses [66]. A tailored program is an effective intervention that provides motivation to continue the program, reduces stress by increasing participant engagement, and increases participant satisfaction [67]. Therefore, traditional and conventional care is less effective while personalized care is considered the best [68]. However, interventions that provide tailored programs suited to individual characteristics are scarce in the literature. The program that we developed proved to be an effective intervention by increasing

XSL•FO
RenderX

satisfaction and reducing burnout, job stress, and stress response. This program could become a pioneer and serve as a guide for the development of burnout reduction programs in the future.

Nurse Healing Space consists of visual programs in which nurse characters guide users by narrating, demonstrating the movement of body parts, showing facial expressions, and singing which might overcome the limitations of previously developed audio programs. Recent interventions that have proven to be effective for nurse burnout are commercially available audio-based programs, such as Headspace and UCLA Mindful [12,13]. However, audio-based programs through mobile apps could result in increased dropout rates due to a lack of participant concentration [69]. Moreover, a previous study demonstrated that video-based exercise programs were effective [70] and had a low dropout rate [71].

The users were highly satisfied with the app's convenience and high-quality visuals. However, they were dissatisfied with the limitations of alarm functions and individual settings that need improvement.

CBT has long been used for burnout. In this study, among 4 CBT programs, ACT showed the greatest burnout reduction rate. This is interesting considering that the most researched program for nurse burnout is mindfulness meditation [29]. In recent years, ACT has been increasingly adopted to prevent and reduce professional staff burnout in medical and other workplace settings [50]. In addition, ACT was reported as an effective stress management intervention for medical staff, including nurses [49]. Moreover, in our study, storytelling and reflective writing and laughter therapy reduced burnout scores more effectively than mindfulness meditation. Although many studies have verified the effectiveness of a single program, further research is needed to compare the effects of different burnout programs.

Although participants experienced reduced job stress and stress responses after the intervention, no statistically significant change was reported in the coping strategy level. This could be because the coping strategies introduced in the interventions were not put into action. Conducting debriefing sessions or forming online communities to enhance behavior changes could assist users in coping.

Limitations

We recruited 300 users to optimize the AI algorithm. The resulting sample is not truly representative of the entire population and hence the results are not generalizable [72]. Moreover, the absence of a control group restricts our capacity to determine the causal links underlying the effects of the intervention. Although the satisfaction scores for the program increased as the number of users increased, algorithm optimization is an ongoing process.

Conclusions

In this study, we developed an AI-based tailored mobile intervention for burnout and optimized the algorithm to suggest programs to users based on user similarities. Our mobile app's usability was acceptable, as was discovered through postchecking and satisfaction measurement , which showed that the algorithm functions normally. In addition, based on the CBI scores before and after the program, valid data were accumulated for 45.7% (274/600) of the participants. Through the development of interventions using an AI program, we showed that not only did nurse burnout decrease, job stress and stress responses also decreased. Nurses may be motivated to participate in programs to reduce burnout through interventions tailored to various forms of burnout. This program is the first customized burnout reduction program for nurses, and nurse managers can make use of AI-based systems to offer specialized programs for nurse burnout. This could reduce nurse retirement rates, improve patient safety and qualitative health care, increase employee satisfaction, reduce costs, and ultimately improve the efficiency of the health care system.

Acknowledgments

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and Information & Communications Technology; 2021R1A2C2008166) and the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (Korea) and the NRF (NRF-5199990614253).

Authors' Contributions

Conceptualization was done by CC, AC, and GB; data curation and formal analysis by AC and GB; funding acquisition by CC; methodology by CC; project administration by CC; visualization by AC and GB; writing (original draft) by CC, AC, and GB; and writing (review and editing) by CC, AC, and GB.

Conflicts of Interest

None declared.

References

1. Maslach C, Jackson SE. The measurement of experienced burnout. J Organ Behav 2007;2(2):99-113. [doi: 10.1002/job.4030020205]
2. Kristensen TS, Borritz M, Villadsen E, Christensen KB. The Copenhagen Burnout Inventory: a new tool for the assessment of burnout. Work Stress 2005;19(3):192-207. [doi: 10.1080/02678370500297720]

XSL•FO
RenderX

3. International Classification of Diseases 11th Revision: The global standard for diagnostic health information. World Health Organization. 2021. URL: https://icd.who.int/en [accessed 2024-05-22]
4. Ge M, Hu F, Jia Y, Tang W, Zhang W, Chen H. Global prevalence of nursing burnout syndrome and temporal trends for the last 10 years: a meta-analysis of 94 studies covering over 30 countries. J Clin Nurs 2023;32(17-18):5836-5854. [doi: 10.1111/jocn.16708] [Medline: 37194138]
5. Sheng H, Tian D, Sun L, Hou Y, Liu X. Nurse practice environment, perceived organizational support, general well-being, occupational burnout and turnover intention: a moderated multi-mediation model. Nurs Open 2023;10(6):3828-3839 [FREE Full text] [doi: 10.1002/nop2.1641] [Medline: 36916419]
6. Dewanto A, Wardhani V. Nurse turnover and perceived causes and consequences: a preliminary study at private hospitals in Indonesia. BMC Nurs 2018;17(Suppl 2):52 [FREE Full text] [doi: 10.1186/s12912-018-0317-8] [Medline: 30574016]
7. Kim S, Yang Y. A meta analysis of variables related to burnout of nurse in Korea. J Digit Convergence 2015;13(8):387-400. [doi: 10.14400/jdc.2015.13.8.387]
8. Molina-Praena J, Ramirez-Baena L, Gómez-Urquiza JL, Cañadas GR, De la Fuente EI, Cañadas-De la Fuente GA. Levels of burnout and risk factors in medical area nurses: a meta-analytic study. Int J Environ Res Public Health 2018;15(12):2800 [FREE Full text] [doi: 10.3390/ijerph15122800] [Medline: 30544672]
9. Aryankhesal A, Mohammadibakhsh R, Hamidi Y, Alidoost S, Behzadifar M, Sohrabi R, et al. Interventions on reducing burnout in physicians and nurses: a systematic review. Med J Islam Repub Iran 2019;33:77 [FREE Full text] [doi: 10.34171/mjiri.33.77] [Medline: 31696071]
10. Coster S, Gould RL, Coulson M, Norman IJ. An online mindfulness intervention to enhance compassion in nursing practice: a feasibility and acceptability study with nursing students. Int J Nurs Stud Adv 2020;2:100004 [FREE Full text] [doi: 10.1016/j.ijnsa.2020.100004] [Medline: 38745902]
11. Fish J, Brimson J, Lynch S. Mindfulness interventions delivered by technology without facilitator involvement: what research exists and what are the clinical outcomes? Mindfulness (N Y) 2016;7(5):1011-1023 [FREE Full text] [doi: 10.1007/s12671-016-0548-2] [Medline: 27642370]
12. Taylor H, Cavanagh K, Field AP, Strauss C. Health care workers' need for headspace: findings from a multisite definitive randomized controlled trial of an unguided digital mindfulness-based self-help app to reduce healthcare worker stress. JMIR Mhealth Uhealth 2022;10(8):e31744 [FREE Full text] [doi: 10.2196/31744] [Medline: 36006668]
13. Qualls BW, Payton EM, Aikens LG, Carey MG. Mindfulness for outpatient oncology nurses: a pilot study. Holist Nurs Pract 2022;36(1):28-36. [doi: 10.1097/HNP.0000000000000487] [Medline: 34783698]
14. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer 2009;42(8):30-37. [doi: 10.1109/mc.2009.263]
15. Covington P, Jay A, Emre S. Deep neural networks for YouTube recommendations. 2016 Sep 07 Presented at: Proceedings of the 10th ACM conference on recommender systems; 2016; New York, NY p. 191-198 URL: https://dl.acm.org/doi/pdf/10.1145/2959100.2959190
16. Schafer JB, Konstan J, John R. Recommender systems in e-commerce. 1999 Presented at: Proceedings of the 1st ACM Conference on Electronic Commerce; November 3-5, 1999; Denver, CO p. 158-166 URL: https://dl.acm.org/doi/pdf/10.1145/336992.337035
17. Kim J, Choi I, Li Q. Customer satisfaction of recommender system: examining accuracy and diversity in several types of recommendation approaches. Sustainability 2021;13(11):6165. [doi: 10.3390/su13116165]
18. Shardanand U, Pattie M. Social information filtering: Algorithms for automating "word of mouth". 1995 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 1995; United States p. 210-217 URL: https://dl.acm.org/doi/pdf/10.1145/223904.223931
19. Lee KJ, Baek J, Hwangbo YJ, Kim YC, Bae SW, Baek TH. AMPER (Aim-Measure-Predict-Evaluate-Recommend): The paradigm of digital me. In: 23rd International Conference on Electronic Commerce. South Korea: DBpia; 2022 Presented at: 23rd International Conference on Electronic Commerce 2022; June 22-24, 2022; Daegu, South Korea p. 182-184 URL: https://www-dbpia-co-kr-ssl.access.ewha.ac.kr/journal/articleDetail?nodeId=NODE11150700
20. Lewis HS, Cunningham CJL. Linking nurse leadership and work characteristics to nurse burnout and engagement. Nurs Res 2016;65(1):13-23. [doi: 10.1097/NNR.0000000000000130] [Medline: 26657477]
21. Jansen PG, Kerkstra A, Abu-Saad HH, van der Zee J. The effects of job characteristics and individual characteristics on job satisfaction and burnout in community nursing. Int J Nurs Stud 1996;33(4):407-421. [doi: 10.1016/0020-7489(95)00060-7] [Medline: 8836765]
22. Potard C, Landais C. The use of cluster analysis to identify different burnout profiles among nurses and care assistants for older adults. Geriatr Nurs 2021;42(5):1135-1142. [doi: 10.1016/j.gerinurse.2021.07.005] [Medline: 34352680]
23. Park CS. Factors Influencing on Burnout Experience in Working Nurses at Hospital. Korean J Women Health Nurs 2002;8(4):550-558. [doi: 10.4069/kjwhn.2002.8.4.550]
24. Lee H, Song R, Cho YS, Lee GZ, Daly B. A comprehensive model for predicting burnout in Korean nurses. J Adv Nurs 2003;44(5):534-545. [doi: 10.1046/j.0309-2402.2003.02837.x] [Medline: 14651702]
25. Harrad R, Sulla F. Factors associated with and impact of burnout in nursing and residential home care workers for the elderly. Acta Biomed 2018;89(7-S):60-69 [FREE Full text] [doi: 10.23750/abm.v89i7-S.7830] [Medline: 30539935]

XSL•FO
RenderX

26. Barrett K, Stewart I. A preliminary comparison of the efficacy of online Acceptance and Commitment Therapy (ACT) and Cognitive Behavioural Therapy (CBT) stress management interventions for social and healthcare workers. Health Soc Care Community 2021;29(1):113-126. [doi: 10.1111/hsc.13074] [Medline: 32613644]
27. Kabat-Zinn J. Mindfulness-based interventions in context: Past, present, and future. Clin Psychol Sci Pract 2003;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
28. Yeonhee C, Sanghae B. Therapeutic mechanism of MBCT and clinical application of MBCT (Mindfulness-Based Cognitive Therapy) program on chronic depression. J Korea Academia Ind Cooperation Soc 2017;18(7):237-248. [doi: 10.5762/KAIS.2017.18.7.237]
29. Suleiman-Martos N, Gomez-Urquiza JL, Aguayo-Estremera R, Cañadas-De La Fuente GA, De La Fuente-Solana EI, Albendín-García L. The effect of mindfulness training on burnout syndrome in nursing: a systematic review and meta-analysis. J Adv Nurs 2020;76(5):1124-1140. [doi: 10.1111/jan.14318] [Medline: 32026484]
30. Fendel JC, Bürkle JJ, Göritz AS. Mindfulness-based interventions to reduce burnout and stress in physicians: a systematic review and meta-analysis. Acad Med 2021;96(5):751-764. [doi: 10.1097/ACM.0000000000003936] [Medline: 33496433]
31. Salvado M, Marques DL, Pires IM, Silva NM. Mindfulness-based interventions to reduce burnout in primary healthcare professionals: a systematic review and meta-analysis. Healthcare (Basel) 2021;9(10):1342 [FREE Full text] [doi: 10.3390/healthcare9101342] [Medline: 34683022]
32. Huang F, Hsu A, Hsu L, Tsai J, Huang C, Chao Y, et al. Mindfulness improves emotion regulation and executive control on bereaved individuals: an FMRI study. Front Hum Neurosci 2018;12:541 [FREE Full text] [doi: 10.3389/fnhum.2018.00541] [Medline: 30745865]
33. Keng S, Chin JWE, Mammadova M, Teo I. Effects of mobile app-based mindfulness practice on healthcare workers: a randomized active controlled trial. Mindfulness (N Y) 2022;13(11):2691-2704 [FREE Full text] [doi: 10.1007/s12671-022-01975-8] [Medline: 36160038]
34. Pan P, Yeh C. Laughter therapy in a geriatric patient with treatment-resistant depression and comorbid tardive dyskinesia. Aust N Z J Psychiatry 2016;50(12):1206-1207. [doi: 10.1177/0004867416646859] [Medline: 27164922]
35. Morishima T, Miyashiro I, Inoue N, Kitasaka M, Akazawa T, Higeno A, et al. Effects of laughter therapy on quality of life in patients with cancer: an open-label, randomized controlled trial. PLoS One 2019;14(6):e0219065 [FREE Full text] [doi: 10.1371/journal.pone.0219065] [Medline: 31247017]
36. Cha MY, Hong HS. [Effect and path analysis of laughter therapy on serotonin, depression and quality of life in middle-aged women]. J Korean Acad Nurs 2015;45(2):221-230. [doi: 10.4040/jkan.2015.45.2.221] [Medline: 25947184]
37. Ryu KH, Shin HS, Yang EY. Effects of laughter therapy on immune responses in postpartum women. J Altern Complement Med 2015;21(12):781-788. [doi: 10.1089/acm.2015.0053] [Medline: 26496141]
38. Si S Çeli K, Kılınç T. The effect of laughter yoga on perceived stress, burnout, and life satisfaction in nurses during the pandemic: a randomized controlled trial. Complement Ther Clin Pract 2022;49:101637 [FREE Full text] [doi: 10.1016/j.ctcp.2022.101637] [Medline: 35810525]
39. Cann A, Stilwell K, Taku K. Humor styles, positive personality and health. EJOP 2010;6(3):213-235. [doi: 10.5964/ejop.v6i3.214]
40. Murphy R, Higgins A. The complex terrain of peer support in mental health: What does it all mean? J Psychiatr Ment Health Nurs 2018;25(7):441-448. [doi: 10.1111/jpm.12474] [Medline: 29797631]
41. Rhodes J. Narrative CBT. London: Routledge Taylor & Francis; 2013:ISBN:9780415533973.
42. Fitzpatrick JJ. Narrative nursing: applications in practice, education, and research. Appl Nurs Res 2017;37:67. [doi: 10.1016/j.apnr.2017.08.005] [Medline: 28985924]
43. Jun J, Siegrist K, Weinshenker D. Evaluation of nurses' experiences with digital storytelling workshop: new way to engage, connect, and empower. J Nurs Manag 2022;30(5):1317-1323 [FREE Full text] [doi: 10.1111/jonm.13619] [Medline: 35403291]
44. Bittman B, Bruhn KT, Stevens C, Westengard J, Umbach PO. Recreational music-making: a cost-effective group interdisciplinary strategy for reducing burnout and improving mood states in long-term care workers. Adv Mind Body Med 2003;19(3-4):4-15. [Medline: 14686266]
45. Barclay LJ, Skarlicki DP. Healing the wounds of organizational injustice: examining the benefits of expressive writing. J Appl Psychol 2009;94(2):511-523. [doi: 10.1037/a0013451] [Medline: 19271805]
46. Phillips CS, Volker DL, Davidson KL, Becker H. Storytelling through music: a multidimensional expressive arts intervention to improve emotional well-being of oncology nurses. JCO Oncol Pract 2020;16(4):e405-e414 [FREE Full text] [doi: 10.1200/JOP.19.00748] [Medline: 32275852]
47. Salari N, Khazaie H, Hosseinian-Far A, Khaledi-Paveh B, Ghasemi H, Mohammadi M, et al. The effect of acceptance and commitment therapy on insomnia and sleep quality: a systematic review. BMC Neurol 2020;20(1):300 [FREE Full text] [doi: 10.1186/s12883-020-01883-1] [Medline: 32791960]
48. Hayes S, Strosahl K. A Practical Guide to Acceptance and Commitment Therapy. New York, NY: Springer Science & Business Media; 2004:ISBN:9780387233673.
49. Moran DJ. Acceptance and commitment training in the workplace. Curr Opin Psychol 2015;2:26-31. [doi: 10.1016/j.copsyc.2014.12.031]

XSL•FO
RenderX

50. Prudenzi A, Graham CD, Flaxman PE, Wilding S, Day F, O'Connor DB. A workplace Acceptance and Commitment Therapy (ACT) intervention for improving healthcare staff psychological distress: a randomised controlled trial. PLoS One 2022;17(4):e0266357 [FREE Full text] [doi: 10.1371/journal.pone.0266357] [Medline: 35442963]
51. Towey-Swift KD, Lauvrud C, Whittington R. Acceptance and commitment therapy (ACT) for professional staff burnout: a systematic review and narrative synthesis of controlled trials. J Ment Health 2023;32(2):452-464. [doi: 10.1080/09638237.2021.2022628] [Medline: 35105253]
52. Hayes S, Smith SX. Get Out of Your Mind and Into Your Life: The New Acceptance and Commitment Therapy. Oakland, CA: New Harbinger Publications; 2005.
53. Ham M. Path analysis of emotional labor and burnout of nurses [thesis]. Gyeongsang National University. 2011. URL: https://dcollection.gnu.ac.kr/jsp/common/DcLoOrgPer.jsp?sItemId=000000014045 [accessed 2024-06-17]
54. Gu MO, Kim MJ. An analytical study of work stress among clinical nurses. J Nurs Acad Soc 1985;15(3):39. [doi: 10.4040/jnas.1985.15.3.39]
55. Choi S, Kang T, Woo J. Development and validation of a modified form of the stress response inventory for workers. J Korean Neuropsychiatr Assoc 2006;45(6):541-553 [FREE Full text]
56. Amirkhan JH. A factor analytically derived measure of coping: The Coping Strategy Indicator. J Pers Soc Psychol 1990;59(5):1066-1074. [doi: 10.1037/0022-3514.59.5.1066]
57. Shin H, Kim C. A validation study of coping strategy indicator (CSI). Korean J Couns Psychother 2002;14(4):919-935.
58. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
59. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 2016;25(3):1057-1073 [FREE Full text] [doi: 10.1177/0962280215588241] [Medline: 26092476]
60. West S, Finch J, Curran P. Structural equation models with nonnormal variables: problemsremedies. In: Structural Equation Modeling: Concepts, Issues and Applications. Thousand Oaks, CA: Sage Publications; 1995:56-75.
61. Mentis AA, Lee D, Roussos P. Applications of artificial intelligence-machine learning for detection of stress: a critical overview. Mol Psychiatry 2023;28:1-13. [doi: 10.1038/s41380-023-02047-6] [Medline: 37020048]
62. Martínez-Ramón JP, Morales-Rodríguez FM, Pérez-López S. Burnout, resilience, and COVID-19 among teachers: predictive capacity of an artificial neural network. Appl Sci 2021;11(17):8206. [doi: 10.3390/app11178206]
63. Grządzielewska M. Using machine learning in burnout prediction: a survey. Child Adolesc Soc Work J 2021;38(2):175-180. [doi: 10.1007/s10560-020-00733-w]
64. Gupta MD, Bansal A, Sarkar PG, Girish M, Jha M, Yusuf J, et al. Design and rationale of an intelligent algorithm to detect BuRnoUt in HeaLthcare workers in COVID era using ECG and artificiaL intelligence: the BRUCEE-LI study. Indian Heart J 2021;73(1):109-113 [FREE Full text] [doi: 10.1016/j.ihj.2020.11.145] [Medline: 33714394]
65. Kawakami N, Imamura K, Watanabe K, Sekiya Y, Sasaki N, Sato N, SMART-CBT Project Team. Effectiveness of an internet-based machine-guided stress management program based on cognitive behavioral therapy for improving depression among workers: protocol for a randomized controlled trial. JMIR Res Protoc 2021;10(9):e30305 [FREE Full text] [doi: 10.2196/30305] [Medline: 34460414]
66. Matthew J, Mike L, Huang H, Wang C, Shih C, Chen Y, et al. Effects of personalized music intervention on nurse burnout: a feasibility randomized controlled trial. Nurs Health Sci 2022;24(4):836-844. [doi: 10.1111/nhs.12984] [Medline: 36089738]
67. Mistretta EG, Davis MC, Temkit M, Lorenz C, Darby B, Stonnington CM. Resilience training for work-related stress among health care workers: results of a randomized clinical trial comparing in-person and smartphone-delivered interventions. J Occup Environ Med 2018;60(6):559-568. [doi: 10.1097/JOM.0000000000001285] [Medline: 29370014]
68. Golubnitschaja O, Baban B, Boniolo G, Wang W, Bubnov R, Kapalla M, et al. Medicine in the early twenty-first century: paradigm and anticipation—EPMA position paper 2016. EPMA J 2016;7(1):23 [FREE Full text] [doi: 10.1186/s13167-016-0072-4] [Medline: 27800037]
69. Ma J, Kyu M, Gim W, Lee J. A study on UI/UX design of mobile application of mindfulness. J Korea Convergence Soc 2018;9(11):179-192. [doi: 10.15207/JKCS.2018.9.11.179]
70. Bruschwein H, Gettle LS. Multipronged intervention for reducing burnout and increasing resiliency in an interdisciplinary care team. BMJ Open Qual 2020;9(4):e001015 [FREE Full text] [doi: 10.1136/bmjoq-2020-001015] [Medline: 33208306]
71. Liu F, Cai H, Tang Q, Zou Y, Wang H, Xu Z, et al. Effects of an animated diagram and video-based online breathing program for dyspnea in patients with stable COPD. Patient Prefer Adherence 2013;7:905-913 [FREE Full text] [doi: 10.2147/PPA.S43305] [Medline: 24049441]
72. Elston DM. Participation bias, self-selection bias, and response bias. J Am Acad Dermatol 2021;85:1-2. [doi: 10.1016/j.jaad.2021.06.025] [Medline: 34153389]

Abbreviations

ACT: acceptance and commitment therapy

XSL•FO
RenderX

AI: artificial intelligence
CBI: Copenhagen Burnout Inventory
CBT: cognitive behavioral therapy

Edited by T de Azevedo Cardoso; submitted 02.11.23; peer-reviewed by J Brewer, H Cho, G Humphreys; comments to author 30.11.23; revised version received 21.12.23; accepted 07.05.24; published 21.06.24.

Please cite as:
Cho A, Cha C, Baek G
Development of an Artificial Intelligence–Based Tailored Mobile Intervention for Nurse Burnout: Single-Arm Trial
J Med Internet Res 2024;26:e54029
URL: https://www.jmir.org/2024/1/e54029
doi:10.2196/54029
PMID:

©Aram Cho, Chiyoung Cha, Gumhee Baek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Symptom-Checker for Adult Patients Visiting an Interdisciplinary Emergency Care Center and the Safety of Patient Self-Triage: Real-Life Prospective Evaluation

Andreas Meer[1], MSc, MD; Philipp Rahm[2], MD; Markus Schwendinger[2], MD; Michael Vock[3], PhD; Bettina Grunder[1], MD; Jacopo Demurtas[1], MD, PhD; Jonas Rutishauser[4], MD

[1]In4medicine Inc, Bern, Switzerland

[2]Cantonal Hospital Baden, Baden, Switzerland

[3]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

[4]Clinical Trial Unit, Cantonal Hospital Baden and Medical Faculty, University of Basel, Baden, Switzerland

Corresponding Author:
Andreas Meer, MSc, MD
In4medicine Inc
Monbijoustrasse 23
Bern, 3011
Switzerland
Phone: 41 313701330
Email: a.meer@in4medicine.ch

Abstract

Background: Symptom-checkers have become important tools for self-triage, assisting patients to determine the urgency of medical care. To be safe and effective, these tools must be validated, particularly to avoid potentially hazardous undertriage without leading to inefficient overtriage. Only limited safety data from studies including small sample sizes have been available so far.

Objective: The objective of our study was to prospectively investigate the safety of patients' self-triage in a large patient sample. We used SMASS (Swiss Medical Assessment System; in4medicine, Inc) pathfinder, a symptom-checker based on a computerized transparent neural network.

Methods: We recruited 2543 patients into this single-center, prospective clinical trial conducted at the cantonal hospital of Baden, Switzerland. Patients with an Emergency Severity Index of 1-2 were treated by the team of the emergency department, while those with an index of 3-5 were seen at the walk-in clinic by general physicians. We compared the triage recommendation obtained by the patients' self-triage with the assessment of clinical urgency made by 3 successive interdisciplinary panels of physicians (panels A, B, and C). Using the Clopper-Pearson CI, we assumed that to confirm the symptom-checkers' safety, the upper confidence bound for the probability of a potentially hazardous undertriage should lie below 1%. A potentially hazardous undertriage was defined as a triage in which either all (consensus criterion) or the majority (majority criterion) of the experts of the last panel (panel C) rated the triage of the symptom-checker to be "rather likely" or "likely" life-threatening or harmful.

Results: Of the 2543 patients, 1227 (48.25%) were female and 1316 (51.75%) male. None of the patients reached the prespecified consensus criterion for a potentially hazardous undertriage. This resulted in an upper 95% confidence bound of 0.1184%. Further, 4 cases met the majority criterion. This resulted in an upper 95% confidence bound for the probability of a potentially hazardous undertriage of 0.3616%. The 2-sided 95% Clopper-Pearson CI for the probability of overtriage (n=450 cases,17.69%) was 16.23% to 19.24%, which is considerably lower than the figures reported in the literature.

Conclusions: The symptom-checker proved to be a safe triage tool, avoiding potentially hazardous undertriage in a real-life clinical setting of emergency consultations at a walk-in clinic or emergency department without causing undesirable overtriage. Our data suggest the symptom-checker may be safely used in clinical routine.

Trial Registration: ClinicalTrials.gov NCT04055298; https://clinicaltrials.gov/study/NCT04055298

(J Med Internet Res 2024;26:e58157) doi:10.2196/58157

XSL•FO
RenderX

KEYWORDS

safety; telemedicine; teletriage; symptom-checker; self-triage; self-assessment; triage; triaging; symptom; symptoms; validation; validity; telehealth; mHealth; mobile health; app; apps; application; applications; diagnosis; diagnoses; diagnostic; diagnostics; checker; checkers; check; web; neural network; neural networks

Introduction

In potentially critical situations, clinical warning signs and symptoms may be considered too late by patients due to a lack of professional triage [1]. In this context, various initiatives have been launched to improve outpatient emergency care and the population's access to a low-threshold initial medical assessment [2]. Symptom-checkers, which enable medical self-triage, have recently been introduced for this purpose. Such tools could assist the increasing number of persons without ready access to a primary care physician, for example, migrants or young persons who had been previously healthy. If implemented in settings outside the hospital, that is, at home or work, tools for efficient and safe self-triage could help avoid unnecessary emergency hospital visits, thus contributing to reducing overcrowding and costs.

To fulfill the regulatory requirements [3] and to be used as a part of standard care, the appropriateness and safety of these instruments must be evaluated in concrete clinical settings with real patients [4-6].

Appropriate care results from adequate triage and treatment, while inappropriate care may lead to unsuitable or even dangerous health care delivery. The concept of appropriateness hence includes a widespread range of quality aspects, of which safety is only one. The difficulty of assessing appropriateness in health care and of gaining agreement between clinicians on acceptable and safe care is highlighted by different authors [7,8]. When assessing the appropriateness of medical triage, the question "Was the decision right" suggests that there is merely one single correct triage decision. This question does not appropriately reflect the complex interaction of clinical, social, and environmental factors in medical decision-making. Rather, physicians should consider a range of appropriate triage decisions to guide their actions. Safety is an essential quality attribute of a medical service. In contrast to the idea of appropriateness, the concept of safety focuses on the risk of a specific conduct. When asking about the safety of a symptom-checker, a risk-based approach should be taken, and safety should encompass possible risks to a patient's health and life [3].

An evaluation of 23 symptom-checkers using 45 patient vignettes concluded that most symptom-checkers were deficient in both appropriate triage and correct diagnosis [9]. However, the study did not comment on the safety of the tested devices. A review paper including 14 studies found inconsistent evidence regarding the triage and diagnostic appropriateness of symptom-checkers for common health problems. The average appropriateness of triage ranged from 27% to 92%. This paper did not specifically evaluate the safety of symptom-checkers [5]. Another review paper cited only 6 studies that analyzed the safety of symptom-checkers [10]. These studies were mostly short-term and included samples that were too small and heterogeneous to make reliable statements about safety.

Given the present shortage of data on self-triage, we aimed to investigate the safety of a newly developed symptom-checker (SMASS; Swiss Medical Assessment System) in a concrete clinical setting with patients seeking emergency care.

Methods

Study Design

Before the inclusion of the first patient, this study was registered (ClinicalTrials.gov identifier: NCT04055298).

This study was performed between November 25, 2019, and May 1, 2020, at the walk-in clinic and interdisciplinary emergency department (WIC/ED) of the cantonal hospital of Baden, Switzerland. The WIC/ED is open 24 hours a day, 365 days a year and treats about 55,000 patients annually. Patients are routinely triaged by a nurse using the Emergency Severity Index (ESI) [11]. ESI 1-2 patients are treated in the ED, while ESI 3-5 patients are treated in the WIC.

The symptom-checker used in this study (SMASS in the pathfinder version, release 4.1.12) was developed by in4medicine, Inc. The first author (author AM) is the chief executive officer and founder of this company. To minimize bias, a majority of independent researchers were involved in this study, including establishing the protocol and all practical aspects of the trial. No employee of in4medicine took part in the actual conduct of the trial. Data analysis and statistical calculations were performed by an independent biostatistician. This study was independently monitored by the clinical trial unit of the Medical Faculty of the University of Bern.

The SMASS pathfinder symptom-checker is a medical device class I under the Medical Device Directive and medical device class IIb under the Medical Device Regulation. The Conformité Européenne declaration of conformity to the Swiss Agency for Therapeutic Products (Swissmedic) was made on June 4, 2018. The symptom-checker is a web-based software that aims to support health professionals and laypersons in the structured documentation and assessment of health problems and to advise users about possible medical assessment steps and treatment measures. It is based on a computerized neural network that incorporates extensive data from scientific studies, guidelines, and expertise from various professional boards of specialists in the field of prehospital medical triage. The symptom-checker provides digitalized questionnaires of 125 frequent reasons for consultations (eg, fever, cough, and abdominal pain) and their associated red flags. Based on the triage result, a report including patient gender, age group, symptoms, medical history, and recommendations as to the appropriate time-to-treat and point-of-care is provided. Depending on the presence of red flags, the symptom-checker assigns the clinical condition of the patient to a triage level (Tables 1 and 2). If 5 or more assessment

XSL•FO
RenderX

questions are answered as "unclear," the user is notified that the software cannot provide targeted triage advice and that the patient should seek immediate consultation with a physician concerning his or her medical complaints.

Table 1. Triage levels as recommended to the patient by the symptom-checker. Recommendations are given regarding time-to-treat (emergency, immediately, today, later, and unclear) and point-of-care (ambulance, hospital, physician, call center, pharmacy, self-care, and unclear).

	Ambulance	Hospital	Doctor	Call center	Pharmacy	Self-care	Unclear
Emergency	16	15	__[a]	—	—	—	—
Immediately	14	13	12	—	—	—	—
Today	—	11	10	8	6	4	—
Later	—	—	9	7	5	3	1
Unclear	—	—	—	—	—	2	0

[a]Not applicable.

XSL•FO
RenderX

Table 2. Recommended actions to be taken by the patient, as defined by triage levels (left column). Levels range from 0 (lowest level) to 16 (highest level). Interpretations of the triage level and measures to be taken are specified in the right column.

Triage level	Name	Recommended action
Level 16	Emergency ambulance	CPR[a]/CPR readiness. There is a potentially life-threatening condition. Medical treatment must be given now. Alert the emergency services via the number 144.
Level 15	Emergency hospital	CPR/CPR readiness. There is a potentially life-threatening condition. Medical treatment must be given now. Alert the emergency services via the number 144. Medical treatment should be provided at a hospital.
Level 14	Immediately ambulance	Medical treatment does not allow any delay. Treatment should be given immediately. Alert the emergency services via the number 144.
Level 13	Immediately hospital	Medical treatment does not allow any delay. Treatment should be given immediately. Medical treatment should be provided at a hospital.
Level 12	Immediately doctor	Medical treatment does not allow any delay. Treatment should be given immediately. Medical treatment should be provided by a registered doctor [b].
Level 11	Today hospital	Medical treatment does not have to take place immediately, but should not be delayed until tomorrow or over the weekend. Medical treatment should take place within the next 24 hours. Medical treatment should be provided at a hospital.
Level 10	Today doctor	Medical treatment does not have to take place immediately, but should not be delayed until tomorrow or over the weekend. Medical treatment should take place within the next 24 hours. Medical treatment should be provided by a registered doctor [b].
Level 9	Later doctor	Medical treatment is not urgent. If the symptoms do not subside in the next 2 days, treatment by a doctor is indicated. Medical treatment should be provided by a registered doctor [b].
Level 8	Today call center	Medical treatment does not have to take place immediately, but should not be delayed until tomorrow or over the weekend. Medical treatment should take place within the next 24 hours. The affected person should be advised by a telemedicine center on how to proceed.
Level 7	Later call center	Medical treatment is not urgent. If the symptoms do not subside in the next 2 days, treatment by a doctor is indicated. The affected person should be advised by a telemedicine center on how to proceed.
Level 6	Today pharmacy	Medical treatment does not have to take place immediately, but should not be delayed until tomorrow or over the weekend. Medical treatment should take place within the next 24 hours. The affected person should be advised at a pharmacy on how to proceed.
Level 5	Later pharmacy	Medical treatment is not urgent. If the symptoms do not subside in the next 2 days, treatment by a doctor is indicated. The affected person should be advised at a pharmacy on how to proceed.
Level 4	Today self-care	Medical treatment does not have to take place immediately, but should not be delayed until tomorrow or over the weekend. Medical treatment should take place within the next 24 hours. The complaints can be treated independently by simple measures.
Level 3	Later self-care	Medical treatment is not urgent. If the symptoms do not subside in the next 2 days, treatment by a doctor is indicated. The complaints can be treated independently by simple measures.
Level 0-2	Unclear	The survey contains too many ambiguities. A targeted initial assessment is not possible.

[a]CPR: cardiopulmonary resuscitation.

[b]For example, family doctor, family doctor substitute, family doctor emergency service, or suitable specialist.

All patients aged ≥18 years attending the WIC/ED between 8 AM and 5 PM were eligible. Exclusion criteria included aged <18 years; ESI 1 patients requiring immediate, life-saving intervention; inability to use a tablet PC; inability to communicate in German, French, Italian, or English; inability or unwillingness to give written informed consent and follow the procedures of this study; known or suspected noncompliance; known drug or alcohol abuse; the presence of symptoms or complaints not encompassed by the symptom-checker database (eg, long-lasting hiccups, hair loss).

After instruction by the study's staff and providing written informed consent, the participants independently assessed their health status and complaints as instructed by the symptom-checker on a tablet PC. They were subsequently evaluated and treated by routine medical staff.

In primary care, medical triage decisions usually have to be based solely on the patient's symptoms. We have chosen experts independent of the treatment (panels A, B, and C) as evaluators to ensure that the triage decision is based purely on the symptoms of this study's patients. Including treating physicians as comparators in this study could have influenced the triage decision by additional information (physical examination and diagnostic test results).

Our evaluation of the symptom-checker focused on safety, as this is an essential quality attribute of a medical device [3]. To reflect the highly individual nature of medical decision-making,

XSL•FO
RenderX

which usually results in low interrater reliability [12-14], an independent team of experienced physicians engaged in a stepwise evaluation procedure in which each case that was classified as undertriaged by panel A experts was assessed by several experts.

A research assistant and 3 external interdisciplinary panels of board-certified physicians were involved in the evaluation process (panel A, 5 experts; panel B, 2 experts; panel C, 5 experts). Except for one of the 12 panelists (author BG), they were not affiliated with in4medicine, Inc. None of them took part in the conduct of this study. For every patient, the symptom-checker issued a report summarizing the clinical information. Patients and panelists were unaware of the triage-level recommendations (time-to-treat and point-of-care) made by the symptom-checker. All reports were first assessed by members of panel A, who adjudicated an appropriate range of triage levels to every case. The research assistant then compared the adjudication of the panel A experts with the recommendation issued by the symptom-checker. If the comparison showed that the recommendation of the symptom-checker was below the appropriate range of triage levels determined by the rater of panel A, hence was undertriaged, the case was assigned to panel B. In 80 instances, panelists erroneously examined the same cases twice and concluded on diverse triage recommendations. In these cases, the first of the 2 recommendations was used for the analysis.

The evaluation procedure was repeated by panel B. Each of the 2 panelists evaluated all diverging cases. If the case was undertriaged according to 3 experts (1 expert from panel A and 2 experts from panel B), the case was subsequently analyzed by panel C.

Each member of panel C individually assessed the clinical safety of the triage decision based on the complete structured reports generated by the symptom-checker as well as the WIC/ED's redacted discharge reports. Each of the 5 panelists decided individually on potentially hazardous undertriage. In a modified Delphi process, the panelists first individually adjudicated potentially hazardous undertriage on a 4-point Likert scale. Possible ratings were "unlikely," "rather unlikely," "rather likely," and "likely" that the patient was exposed to a risk to life or health. If the panelists subsequently reached a consensus that the triage of the symptom-checker was "rather likely" or "likely" exposing a patient to a risk to life or health, the case was considered a potentially hazardous undertriage (consensus criterion). As a complement to the original analysis plan, a modified criterion for potentially hazardous undertriage was evaluated, defined as a majority of panel C members judging a risk to life or health as "rather likely" or "likely" (majority criterion).

The primary analysis consisted of the calculation of the 95% upper Clopper-Pearson confidence bound for the probability of undertriage resulting in a risk to life or health (potentially hazardous undertriage). To confirm the safety of the symptom-checker, this upper confidence bound should lie below 1%. For the sample size calculation, we assumed that a 20% probability of failure to meet this criterion is acceptable for a true probability of potentially hazardous undertriage of no more than 0.5%. This is equivalent to requiring a 1-sided test at level 5% to show that the probability of potentially hazardous undertriage is below 1% with a power of 80%, assuming that the true probability is 0.5%. This resulted in a minimal sample size of 2185 patients. Accounting for an estimated rate of 2% "unclear" responses, at least 2230 patients were planned to be included. Secondary analyses included central 95% Clopper-Pearson CIs for the further probabilities, based on corresponding empirical proportions. The software R (version 4.2.0; R Foundation for Statistical Computing) was used for the statistical evaluations.

Ethical Considerations

This study was approved by the competent ethics committee (Ethikkommission Nordwest- und Zentralschweiz EKNZ, project ID 01784) and was conducted per the most recent version of the Declaration of Helsinki, complying with International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use—good clinical practice and International Organization for Standardization European Norm 14155 (clinical investigation of medical devices for human subjects—good clinical practice) as well as with applying national legal and regulatory requirements. All patients gave written informed consent to participate in this study. They did not receive any financial or other compensation. Patients were anonymized upon data collection. Discharge notes studied by panel C were redacted.

Generative artificial intelligence was not used in any portion of this paper's writing.

Results

The baseline characteristics of the participants are shown in Table 3 and the recommendations obtained by the symptom-checker in Table 4. Figure 1 shows the flow of analyses by panels A-C.

XSL•FO
RenderX

Table 3. Characteristics of the study population (N=2543).

Characteristics	Participants, n (%)
Age (years)	
18-49	1397 (54.94)
50-65	668 (26.27)
66-80	360 (14.16)
>80	118 (4.64)
Gender	
Female	1227 (48.25)
Male	1316 (51.75)
Reason for encounter (15 most frequent)	
Stomach pain	287 (11.26)
Chest pain	168 (6.61)
Lumbar back pain	144 (5.66)
Urinary tract problems	124 (4.88)
Trauma or fall	121 (4.76)
Headache	90 (3.54)
Dizziness	87 (3.42)
Wound or skin injury	82 (3.22)
Foot injury (caused by an accident)	81 (3.19)
Leg problems	74 (2.91)
Breathlessness	69 (2.71)
Cold or influenza infection	64 (2.52)
Finger injury (caused by an accident)	55 (2.16)
Knee injury (caused by an accident)	51 (2.01)
Hand injury (caused by an accident)	43 (1.69)

Table 4. Distribution of cases according to the various triage levels, as defined in (N=2543).

Triage level	Participants, n (%)
Emergency ambulance	57 (0.02)
Emergency hospital	142 (0.06)
Immediately ambulance	2 (0)
Immediately hospital	685 (0.27)
Immediately doctor	844 (0.33)
Today hospital	3 (0)
Today doctor	579 (0.23)
Later doctor	36 (0.01)
Today call center	26 (0.01)
Later call center	60 (0.02)
Today pharmacy	0 (0)
Later pharmacy	30 (0.01)
Today self-care	0 (0)
Later self-care	77 (0.03)
Unclear	2 (0)

XSL•FO
RenderX

Figure 1. Flow of patients through this study and triage assessment steps by expert panels A, B, and C. ED: emergency department; ESI: Emergency Severity Index; WIC: walk-in clinic.

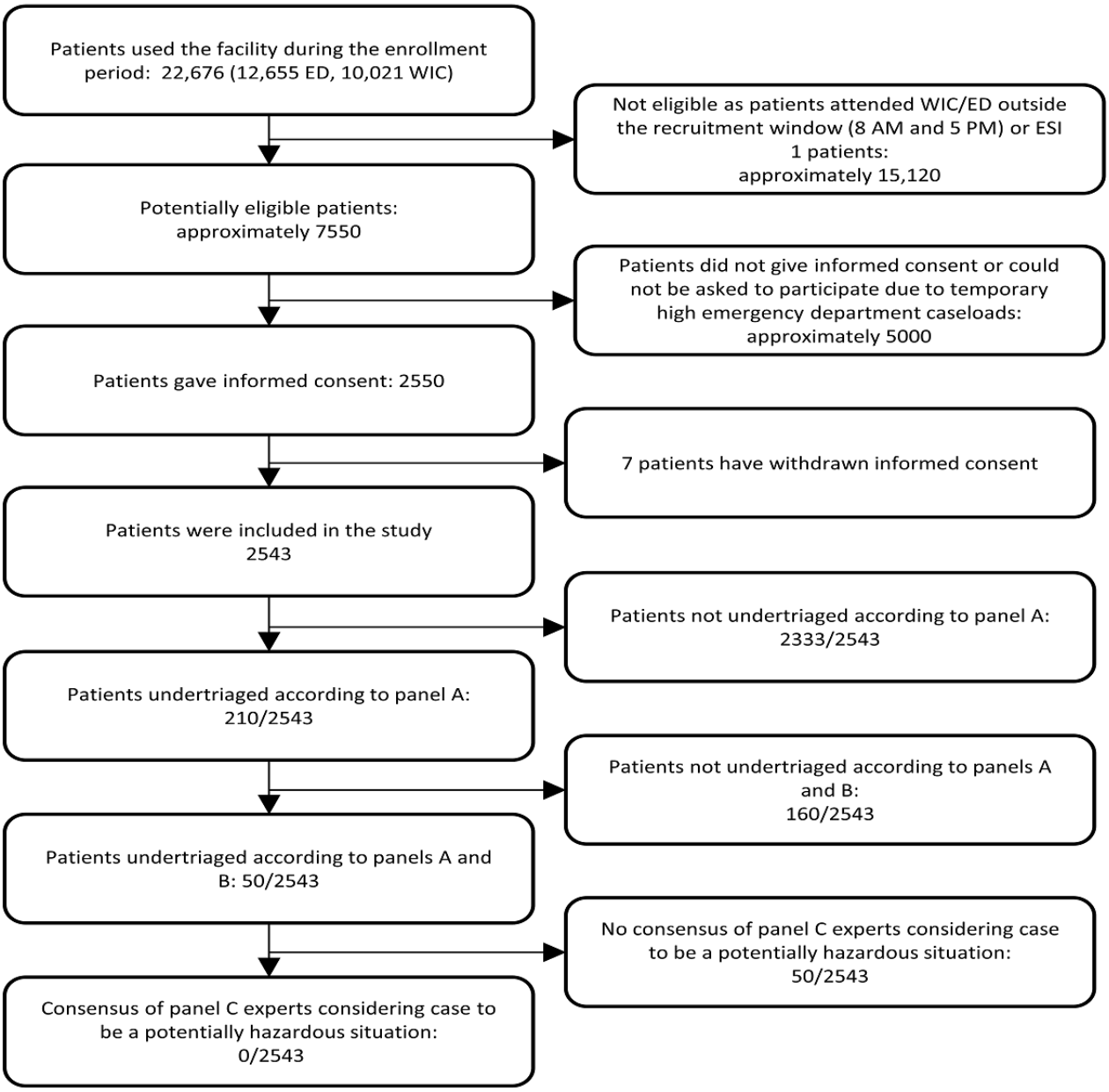

In 210 (8.26%) of the 2543 cases, the recommendation issued by the symptom-checker was below the range of appropriate triage levels defined by the panel A experts and therefore undertriaged. Further, 50 (1.96%) of these 210 patients were equally undertriaged according to panel B. However, for none of these 50 patients did panel C reach a consensus that the undertriage was potentially hazardous. This resulted in an upper 95% confidence bound for the probability of a potentially hazardous undertriage of 0.1184%. If the criterion for potentially hazardous undertriage was defined as a majority of panel C members considering life-threatening or harmful self-triage "rather likely" or "likely," 4 of the 50 cases fulfilled this criterion. This resulted in an upper 95% confidence bound for the probability of a potentially hazardous undertriage of 0.3616%.

Table 5 shows the adjudication of potentially hazardous undertriage for all 50 cases evaluated by the experts of panel C.

XSL•FO
RenderX

Table 5. Distribution of assessment for potentially hazardous undertriage for all 50 cases, as adjudicated by each of the 5 members of panel C.

Case number	Unlikely	Rather unlikely	Rather likely	Likely
Case 1	2	3	0	0
Case 2	5	0	0	0
Case 3	3	2	0	0
Case 4	3	2	0	0
Case 5	4	0	1	0
Case 6	5	0	0	0
Case 7	2	2	1	0
Case 8	4	1	0	0
Case 9	0	2	2	1
Case 10	4	0	1	0
Case 11	5	0	0	0
Case 12	5	0	0	0
Case 13	1	1	2	1
Case 14	4	1	0	0
Case 15	2	2	1	0
Case 16	5	0	0	0
Case 17	4	1	0	0
Case 18	5	0	0	0
Case 19	5	0	0	0
Case 20	5	0	0	0
Case 21	3	1	1	0
Case 22	2	1	1	1
Case 23	3	1	1	0
Case 24	5	0	0	0
Case 25	5	0	0	0
Case 26	5	0	0	0
Case 27	5	0	0	0
Case 28	5	0	0	0
Case 29	2	1	2	0
Case 30	4	0	1	0
Case 31	5	0	0	0
Case 32	2	1	1	1
Case 33	4	1	0	0
Case 34	4	1	0	0
Case 35	5	0	0	0
Case 36	0	2	3	0
Case 37	4	1	0	0
Case 38	5	0	0	0
Case 39	5	0	0	0
Case 40	2	1	1	1
Case 41	1	3	1	0
Case 42	5	0	0	0

XSL•FO
RenderX

Case number	Unlikely	Rather unlikely	Rather likely	Likely
Case 43	5	0	0	0
Case 44	5	0	0	0
Case 45	5	0	0	0
Case 46	5	0	0	0
Case 47	5	0	0	0
Case 48	1	0	4	0
Case 49	4	1	0	0
Case 50	5	0	0	0

The central (2-sided) 95% Clopper-Pearson CI for the probability of undertriage according to panel A is 7.22% to 9.40%. The central (2-sided) 95% Clopper-Pearson CI for the probability of overtriage according to panel A (450 cases, 17.69%) is 16.23% to 19.24%.

For the 50 out of 2543 cases that were undertriaged according to the judgments of panels A and B, the central (2-sided) 95% Clopper-Pearson CI for the corresponding probability is 1.539% to 2.688%.

The central (2-sided) 95% Clopper-Pearson CI for the probability of a potentially hazardous undertriage for the consensus criterion (0 out of 2543 cases) is 0% to 0.1458% and 0.0431% to 0.4045%, according to the majority criterion (4 out of 2543 cases).

Discussion

Principal Findings

Our study corroborates the safety of the SMASS pathfinder symptom-checker for medical self-assessment of acute complaints in a real-life clinical setting. A stepwise evaluation of 2543 consecutive patients by 3 independent expert panels yielded no cases of potentially hazardous undertriage when the consensus criterion was applied and 4 cases when the majority criterion was applied.

In a systematic literature search, we found insufficient evidence from comparatively small studies for the safe use of symptom-checkers in clinical routine (Demurtas et al, unpublished data, 2021). Further, 1 study with 825 patients showing “exactly matched” triage in 52.6% has been published in abstract form only [15]. Another study yielded correct triage in only 50%-74% of cases [16]. A third study, from Germany, evaluated the safety of urgency advice provided to 378 patients at an interdisciplinary ED by a symptom-checker [17], showing undertriage in 34 (8.9%) and overtriage in 216 (57.1%) cases. A potentially hazardous situation was identified in 20 (5.3%) cases. This figure appears considerably higher than our finding, although an interrater variability was not taken into account in the German study. Another study aimed to analyze the performance of a clinical decision support system that allowed patients to self-triage in the ED of a university hospital. The authors concluded that the self-triage device was safe, as the assessments by the system and the physicians were congruent concerning the classification as an emergency. However, in contrast to our study, the risk to life or health was not assessed [18].

In the absence of a broad study base, we cannot compare our results with previous, similarly designed studies for symptom-checkers. In contrast, medical telephone triage has been extensively evaluated during the last 25 years [19-24] and has gained broad clinical support, despite ambivalent conclusions regarding safety.

In a systematic review analyzing 13 observational studies and 10 studies that simulated high-risk patients, safe triage was found to be 46% to 97% [25]. Another systematic review involving computer-assisted telephone triage in urgent care [26] pointed out 4 studies that indicated potential undertriage errors [27-30]. Notably, hospitalization rates of patients who were advised to seek nonurgent care ranged from 9.2% to 48%. Potentially life-threatening situations emerged in 0.84% of cases according to 1 study [29].

We have previously investigated the safety of computer-assisted telephone triage in 208 patients with non–life threatening conditions consulting the ED at a university hospital [31]. We found poor agreement between the assessments by the call center, the emergency physician, and the general practitioners who later cared for the patients. In 1 case, a risk to health or life was found.

The Cochrane Collaboration in their 2004 systematic review on telephone triage concluded that insufficient data existed regarding safety [32]. In light of the available information, the results of our study compare favorably to the published data on telephone triage.

Our study has several strengths and weaknesses. We included a large number of patients in a real-world clinical setting. In addition, this study’s design enabled us to eliminate the low interrater reliability of medical triage decisions by having 3 independent expert panels. This allows robust conclusions about the safety of the evaluated symptom-checker.

For reasons of feasibility, we performed our study in a hospital setting, where patients were triaged to the WIC or ED according to ESI criteria. Thus, a wide variety of cases could be assessed. On the other hand, the symptom-checker was not used in a setting outside the hospital, limiting generalizability. However, presenting symptoms largely overlap with those encountered in primary care, and a potential selection bias toward more

XSL•FO
RenderX

severe cases would support the conclusion on the device's safety if it were used in primary care.

A potential limitation of our study is its single-center design. However, the Cantonal Hospital Baden serves a mixed urban and rural population of approximately 300,000 people and offers all medical services except cardiac surgery and neurosurgery. We therefore believe that the patient sample in our study is fairly representative of the general population.

The total number of patients frequenting the WIC and ED during the time of recruitment was 22,676; thus, only approximately 11% of them participated in this study, potentially resulting in selection bias. Due to limited resources, inclusions were possible only during the daytime, leaving approximately 7550 potential participants. Further, 1.5% (340/22,676) were ESI 1 patients, who were not eligible for this study. It could be speculated that patients visiting an ED at night time might be more seriously ill than those during the daytime. This potential bias would make our cohort more comparable to a setting in primary care.

In our study, we have focused on the safety of the symptom-checker. A possible limitation may have resulted from the fact that each case was initially assessed by a single member of panel A. This could have precluded passing a potentially hazardous case to panel B. While maximum patient safety may theoretically be desirable, it should be weighed against the disadvantages of overtriage, notably inefficiency, unnecessary referrals, and a higher risk of overmedicalization, all of which increase costs. In our study, the overtriage rate after assessment by panel A was 17.69% (450 cases). This figure is comparable to published rates of overtriage by teleconsultation and teletriage, which range from 12% to 57% [33-37]. In a further round of data analysis, we will also have the overtriaged cases assessed by panel B to include the low interrater reliability in the analysis. As with undertriaged cases, this is likely to reduce the overtriaged cases.

From the end of February, the COVID-19 pandemic required special hygiene measures for the tablet computers used, making patient recruitment more difficult as the first wave of the pandemic peaked in March 2020. The pandemic is also likely to have affected the case mix, which may have shifted slightly toward COVID-19–positive patients.

The urgency grading used in the 2D matrix for the triage levels (Table 1) was defined at the discretion of this study's team, implicating a certain degree of subjectiveness. While the range of appropriate triage levels was defined based on this order, the experts did not always explicitly mark all of the intermediate triage levels as appropriate.

Conclusions

The SMASS pathfinder symptom-checker proved to be a safe triage tool, avoiding undertriage in a real-life clinical setting of emergency consultations at a walk-in clinic and ED. Although for practical reasons the symptom-checker was not evaluated outside the hospital environment, our data do not suggest that its safety may have been compromised if used for self-triage by patients in a domestic setting.

Acknowledgments

We thank this study's nurses of the clinical trial unit, Simone Fontana, Stefanie Leuenberger, and Franziska Rutz, PhD, for their excellent work as well as Elena Righi, PhD, and Lee Smith, PhD, for proofreading this paper. This study was funded by the Health Innovation Hub of the Cantonal Hospital Baden, Switzerland. The symptom-checker was provided by in4medicine, Inc, at no charge.

Data Availability

The raw data of our study is available from in4medicine, Inc [38].

Authors' Contributions

AM, JR, MS, and MV were responsible for the concept and design. PR, MS, and BG acquired the data, which were analyzed and interpreted by MV, AM, JD, and JR. The first draft was written by AM and then critically revised in conjunction with JR, JD, and MV. Data analysis was done by MV, with oversight by AM and JR.

Conflicts of Interest

AM is the founder and chief executive officer of in4medicine, Inc. BG is a part-time employee of in4medicine, Inc, and JD received a scientific grant from in4medicine, Inc. MV received an honorarium from in4medicine, Inc. The other authors report no conflicts of interest.

References

1. Thierrin C, Augsburger A, Dami F, Monney C, Staeger P, Clair C. Impact of a telephone triage service for non-critical emergencies in Switzerland: a cross-sectional study. PLoS One 2021;16(4):1-13 [FREE Full text] [doi: 10.1371/journal.pone.0249287] [Medline: 33798216]
2. Steeman L, Uijen M, Plat E, Huibers L, Smits M, Giesen P. Out-of-hours primary care in 26 European countries: an overview of organizational models. Fam Pract 2020;37(6):744-750 [FREE Full text] [doi: 10.1093/fampra/cmaa064] [Medline: 32597962]

XSL•FO
RenderX

3. Regulation (EU) 2017/745 of The European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745 [accessed 2024-05-31]
4. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018;392(10161):2263-2264. [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]
5. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023;25:e43803 [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]
6. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Taghavi Azar Sharabiani M, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. BMJ Open 2022;12(4):e053566 [FREE Full text] [doi: 10.1136/bmjopen-2021-053566] [Medline: 35477872]
7. Munro J, Clancy M, Knowles E, Sampson F, Nicholl J. Evaluation of NHS direct: impact and appropriateness: final report of the phase 1 research. NHS Direct. 2001. URL: https://www.researchgate.net/publication/246112002_Evaluation_of_NHS_Direct_first_wave_sites_Final_report_of_the_phase_1_research [accessed 2024-05-02]
8. Snooks H, Peconi J, Munro J, Cheung WY, Rance J, Williams A. An evaluation of the appropriateness of advice and healthcare contacts made following calls to NHS direct Wales. BMC Health Serv Res 2009;9:178 [FREE Full text] [doi: 10.1186/1472-6963-9-178] [Medline: 19793398]
9. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
10. Chambers D, Cantrell A, Johnson M, Preston L, Baxter S, Booth A, et al. Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review. Southampton (UK): NIHR Journals Library 2019;7(29):1-87. [doi: 10.3310/hsdr07290] [Medline: 31433612]
11. Christ M, Bingisser R, Nickel CH. Emergency Triage. An Overview. Dtsch Med Wochenschr 2016;141(5):329-335. [doi: 10.1055/s-0041-109126] [Medline: 26939102]
12. St George I, Cullen M, Branney M. Healthline: do primary care doctors agree with the advice? N Z Med J 2005;118(1224):U1693. [Medline: 16258573]
13. Gribben B. General practitioners' assessments of the primary care caseload in Middlemore Hospital Emergency Department. N Z Med J 2003;116(1169):U329. [Medline: 12601406]
14. Durand AC, Gentile S, Gerbeaux P, Alazia M, Kiegel P, Luigi S, et al. Be careful with triage in emergency departments: interobserver agreement on 1,578 patients in France. BMC Emerg Med 2011;11:19 [FREE Full text] [doi: 10.1186/1471-227X-11-19] [Medline: 22040017]
15. Koskela T, Liu V, Kaila M. How does triage by an electronic symptom checker match with triage by a nurse? Stud Health Technol Inform 2022;294:571-572. [doi: 10.3233/SHTI220528] [Medline: 35612149]
16. Yu SWY, Ma A, Tsang VHM, Chung LSW, Leung SC, Leung LP. Triage accuracy of online symptom checkers for accident and emergency department patients. Hong Kong J Emerg Med 2020;27(4):217-222. [doi: 10.1177/1024907919842486]
17. Cotte F, Mueller T, Gilbert S, Blümke B, Multmeier J, Hirsch MC, et al. Safety of triage self-assessment using a symptom assessment app for walk-in patients in the emergency care setting: observational prospective cross-sectional study. JMIR mHealth uHealth 2022;10(3):e32340 [FREE Full text] [doi: 10.2196/32340] [Medline: 35343909]
18. Määttä J, Lindell R, Hayward N, Martikainen S, Honkanen K, Inkala M, et al. Diagnostic performance, triage safety, and usability of a clinical decision support system within a university Hospital emergency department: algorithm performance and usability study. JMIR Med Inform 2023;11:e46760 [FREE Full text] [doi: 10.2196/46760] [Medline: 37656018]
19. Derkx HP, Rethans JJE, Muijtjens AM, Maiburg BH, Winkens R, van Rooij HG, et al. Quality of clinical aspects of call handling at Dutch out of hours centres: cross sectional national study. BMJ 2008;337:a1264 [FREE Full text] [doi: 10.1136/bmj.a1264] [Medline: 18790814]
20. Derkx HP, Rethans JJE, Maiburg BH, Winkens RA, Muijtjens AM, van Rooij HG, et al. Quality of communication during telephone triage at Dutch out-of-hours centres. Patient Educ Couns 2009;74(2):174-178. [doi: 10.1016/j.pec.2008.08.002] [Medline: 18845413]
21. Niemann S, Meer A, Simonin C, Abel T. Medical telephone triage and patient behaviour: how do they compare? Swiss Med Wkly 2004;134(9-10):126-131. [doi: 10.4414/smw.2004.10276] [Medline: 15106022]
22. Lattimer V, George S, Thompson F, Thomas E, Mullee M, Turnbull J, et al. Safety and effectiveness of nurse telephone consultation in out of hours primary care: randomised controlled trial. The South Wiltshire Out of Hours Project (SWOOP) Group. BMJ 1998;317(7165):1054-1059. [doi: 10.1136/bmj.317.7165.1054] [Medline: 9774295]
23. Campbell JL, Fletcher E, Britten N, Green C, Holt TA, Lattimer V, et al. Telephone triage for management of same-day consultation requests in general practice (the ESTEEM trial): a cluster-randomised controlled trial and cost-consequence analysis. Lancet 2014;384(9957):1859-1868 [FREE Full text] [doi: 10.1016/S0140-6736(14)61058-8] [Medline: 25098487]
24. Murdoch J, Varley A, Fletcher E, Britten N, Price L, Calitri R, et al. Implementing telephone triage in general practice: a process evaluation of a cluster randomised controlled trial. BMC Fam Pract 2015;16:47 [FREE Full text] [doi: 10.1186/s12875-015-0263-4] [Medline: 25887747]

XSL•FO
RenderX

25. Huibers L, Smits M, Renaud V, Giesen P, Wensing M. Safety of telephone triage in out-of-hours care: a systematic review. Scand J Prim Health Care 2011;29(4):198-209 [FREE Full text] [doi: 10.3109/02813432.2011.629150] [Medline: 22126218]
26. Sexton V, Dale J, Bryce C, Barry J, Sellers E, Atherton H. Service use, clinical outcomes and user experience associated with urgent care services that use telephone-based digital triage: a systematic review. BMJ Open 2022;12(1):e051569 [FREE Full text] [doi: 10.1136/bmjopen-2021-051569] [Medline: 34980613]
27. Foster J, Jessopp L, Chakraborti S. Do callers to NHS direct follow the advice to attend an accident and emergency department? Emerg Med J 2003;20(3):285-288 [FREE Full text] [doi: 10.1136/emj.20.3.285] [Medline: 12748156]
28. Sprivulis P, Carey M, Rouse I. Compliance with advice and appropriateness of emergency presentation following contact with the HealthDirect telephone triage service. Emerg Med Australas 2004;16(1):35-40. [doi: 10.1111/j.1742-6723.2004.00538.x] [Medline: 15239753]
29. Dale J, Higgins J, Williams S, Foster T, Snooks H, Crouch R, et al. Computer assisted assessment and advice for "non-serious" 999 ambulance service callers: the potential impact on ambulance despatch. Emerg Med J 2003;20(2):178-183. [doi: 10.1136/emj.20.2.178] [Medline: 12642540]
30. Stewart B, Fairhurst R, Markland J, Marzouk O. Review of calls to NHS direct related to attendance in the paediatric emergency department. Emerg Med J 2006;23(12):911-914 [FREE Full text] [doi: 10.1136/emj.2006.039339] [Medline: 17130596]
31. Meer A, Gwerder T, Duembgen L, Zumbrunnen N, Zimmermann H. Is computer-assisted telephone triage safe? A prospective surveillance study in walk-in patients with non-life-threatening medical conditions. Emerg Med J 2012;29(2):124-128. [doi: 10.1136/emj.2009.080614] [Medline: 20961939]
32. Bunn F, Byrne G, Kendall S. Telephone consultation and triage: effects on health care use and patient satisfaction. Cochrane Database Syst Rev 2004(4):CD004180. [doi: 10.1002/14651858.CD004180.pub2] [Medline: 15495083]
33. Brasseur E, Gilbert A, Donneau AF, Monseur J, Ghuysen A, D'Orio V. Reliability and validity of an original nurse telephone triage tool for out-of-hours primary care calls: the SALOMON algorithm. Acta Clin Belg 2022;77(3):640-646. [doi: 10.1080/17843286.2021.1936353] [Medline: 34081571]
34. Nørøxe KB, Huibers L, Moth G, Vedsted P. Medical appropriateness of adult calls to Danish out-of-hours primary care: a questionnaire-based survey. BMC Fam Pract 2017;18(1):34 [FREE Full text] [doi: 10.1186/s12875-017-0617-1] [Medline: 28292257]
35. Cook R, Thakore S, Morrison W, Meikle J. To ED or not to ED: NHS 24 referrals to the emergency department. Emerg Med J 2010;27(3):213-215. [doi: 10.1136/emj.2008.064261] [Medline: 20304891]
36. Giesen P, Ferwerda R, Tijssen R, Mokkink H, Drijver R, van den Bosch W, et al. Safety of telephone triage in general practitioner cooperatives: do triage nurses correctly estimate urgency? Qual Saf Health Care 2007;16(3):181-184 [FREE Full text] [doi: 10.1136/qshc.2006.018846] [Medline: 17545343]
37. Scarfone RJ, Luberti AA, Mistry RD. Outcomes of children referred to an emergency department by an after-hours call center. Pediatr Emerg Care 2004;20(6):367-372. [doi: 10.1097/01.pec.0000133610.42699.41] [Medline: 15179144]
38. 2024_05_01_Data_ms_58157. in4medicine. URL: https://pub.in4medicine.ch/fileadmin/pub/2024_05_01_Data_ms_58157.pdf [accessed 2024-06-01]

Abbreviations

ED: emergency department
ESI: Emergency Severity Index
SMASS: Swiss Medical Assessment System
WIC: walk-in clinic

Edited by A Mavragani; submitted 07.03.24; peer-reviewed by C Lowe, R Payne; comments to author 30.03.24; revised version received 15.04.24; accepted 29.05.24; published 27.06.24.

Please cite as:
Meer A, Rahm P, Schwendinger M, Vock M, Grunder B, Demurtas J, Rutishauser J
A Symptom-Checker for Adult Patients Visiting an Interdisciplinary Emergency Care Center and the Safety of Patient Self-Triage: Real-Life Prospective Evaluation
J Med Internet Res 2024;26:e58157
URL: https://www.jmir.org/2024/1/e58157
doi:10.2196/58157
PMID:38809606

XSL•FO
RenderX

©Andreas Meer, Philipp Rahm, Markus Schwendinger, Michael Vock, Bettina Grunder, Jacopo Demurtas, Jonas Rutishauser. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Interventions to Modify Skin Cancer Risk Behaviors in a National Sample of Young Adults: Randomized Controlled Trial

Carolyn J Heckman[1], PhD; Anna Mitarotondo[1], BS; Yong Lin[2], PhD; Olga Khavjou[3], MA; Mary Riley[4], MPH; Sharon L Manne[1], PhD; Amy L Yaroch[5], PhD; Zhaomeng Niu[6], PhD; Karen Glanz[7], MPH, PhD

[1]Rutgers Cancer Institute, New Brunswick, NJ, United States
[2]Rutgers School of Public Health, Piscataway, NJ, United States
[3]RTI International, Research Triangle Park, NC, United States
[4]Medtronic, Denver, CO, United States
[5]Gretchen Swanson Center for Nutrition, Omaha, NE, United States
[6]Rutgers School of Health Professions, Piscataway, NJ, United States
[7]University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Carolyn J Heckman, PhD
Rutgers Cancer Institute
120 Albany St
New Brunswick, NJ, 08901
United States
Phone: 1 732 235 8830
Email: ch842@rutgers.edu

Abstract

Background: Young adults engage in behaviors that place them at risk for skin cancer. Dissemination of digital health promotion interventions via social media is a potentially promising strategy to modify skin cancer risk behaviors by increasing UV radiation (UVR) protection and skin cancer examinations.

Objective: This study aimed to compare 3 digital interventions designed to modify UVR exposure, sun protection, and skin cancer detection behaviors among young adults at moderate to high risk of skin cancer.

Methods: This study was a hybrid type II effectiveness-implementation randomized controlled trial of 2 active interventions, a digital skin cancer risk reduction intervention (UV4.me [basic]) compared with an enhanced version (UV4.me2 [enhanced]), and an electronic pamphlet (e-pamphlet). Intervention effects were assessed over the course of a year among 1369 US young adults recruited primarily via Facebook and Instagram. Enhancements to encourage intervention engagement and behavior change included more comprehensive goal-setting activities, ongoing proactive messaging related to previously established mediators (eg, self-efficacy) of UVR exposure and protection, embedded incentives for module completion, and ongoing news and video updates. Primary outcome effects assessed via linear regression were UVR exposure and sun protection and protection habits. Secondary outcome effects assessed via logistic regression were skin self-exams, physician skin exams, sunscreen use, indoor tanning, and sunburn.

Results: The active interventions increased sun protection (basic: P=.02; enhanced: P<.001) and habitual sun protection (basic: P=.04; enhanced P=.01) compared with the e-pamphlet. The enhanced intervention increased sun protection more than the basic one. Each active intervention increased sunscreen use at the 3-month follow-up (basic: P=.03; enhanced: P=.01) and skin self-exam at 1 year (basic: P=.04; enhanced: P=.004), compared with the e-pamphlet. Other intervention effects and differences between the Basic and Enhanced Intervention effects were nonsignificant.

Conclusions: The active interventions were effective in improving several skin cancer risk and skin cancer prevention behaviors. Compared with the basic intervention, the enhanced intervention added to the improvement in sun protection but not other behaviors. Future analyses will explore intervention engagement (eg, proportion of content reviewed).

Trial Registration: ClinicalTrials.gov NCT03313492; http://clinicaltrials.gov/ct2/show/NCT03313492

(J Med Internet Res 2024;26:e55831) doi:10.2196/55831

XSL•FO
RenderX

KEYWORDS

skin cancer prevention; sun protection; sun exposure; digital intervention; social media; young adults; sun; skin cancer; prevention; cancer; randomized controlled trial; Facebook; Instagram; engagement; behavior; protection; skin; sunscreen

Introduction

Skin Cancer Risk

Almost 5 million Americans are treated for skin cancer annually, and its incidence is rising [1-5]. Additionally, keratinocyte carcinomas can be a chronic disease, requiring ongoing costly treatments and resulting in compromised quality of life similar to some other cancers [6]. Although largely preventable, skin cancers, particularly melanomas, can be deadly, debilitating, damaging to tissues and organs, and disfiguring. Risk factors for skin cancers include personal or family history of melanoma or keratinocyte carcinomas, certain phenotypic (eg, fair skin) and other physical characteristics (eg, numerous moles) [7-19], as well as excessive UV radiation (UVR) exposure from the sun or tanning devices [20-25]. Most skin cancers are preventable with sun protection such as minimizing UVR exposure and wearing protective clothing and sunscreen [26,27]. Early detection via skin examination by an individual, partner, or health care provider may also help improve treatment outcomes and reduce morbidity and mortality [28].

Young adulthood, defined here as ages 18 to 25 years, is an important window for skin cancer risk reduction interventions. Melanoma is one of the most common cancers in people younger than 30 years [29]. Adolescents in the United States have the lowest sun protection rates of all age groups [30] and engage in increased UVR exposure as they move into adulthood [31]. This time period represents a critical period in forming sun protection habits because only 25% of lifetime UVR is accumulated by age 18 years [32]. Skin cancer risk behaviors, including lack of sun protection, sunburns, and indoor tanning, peak around age 25 years [33,34]. Unfortunately, young adults tend to be resistant to preventive health recommendations because, as a group, they perceive themselves as having more immediate priorities than disease prevention and believe that the consequences of their current health behaviors are in the distant future [35-37]. In addition, young adults are more likely to engage in risky behaviors and can be more influenced by their peers than older adults [35-37]. Prior studies of behavioral interventions to increase sun protection or decrease UVR exposure among healthy young adults in the United States [38-45] have been limited in their reach or duration of assessment.

Social Media and Digital Intervention

Social media has been almost universally adopted by young adults. In 2021, almost all 18-29 year olds used the internet [46], approximately 96% owned a smartphone [47], and approximately 70% used Facebook and Instagram [48]. Social media recruitment can be efficient [49-52], and digital behavioral interventions can be disseminated widely and be cost-effective to maintain [53,54]. Digital interventions designed to improve health behaviors (eg, exercise and weight loss) in various populations have been found to produce medium effect sizes and consistently outperform similar nondigital interventions [54], but longer-term outcomes have been minimally evaluated for skin cancer prevention digital interventions [55].

Given the potential of digital behavioral health approaches, we developed an individually tailored, interactive, multimedia, and theoretically grounded (Integrative Model of Behavioral Prediction [IM]) web-based intervention targeted to young adults (basic) [56,57]. We evaluated the intervention in a large sample of 18-25 year olds (n=964) at moderate to high risk of developing skin cancer, recruited from a consumer research panel hosted by a market research company, in a randomized controlled efficacy trial. Behavioral risk factors for melanoma were significantly improved. Intervention effect sizes (Cohen *d*) were 0.53 for sun protection and 0.43 for UVR exposure behaviors at the 3-month follow-up compared with an assessment-only condition [58]. Sunburns were also significantly reduced. This intervention is the basic intervention for the study described in this paper.

Study Purpose

Although the results of the efficacy trial were promising, we proposed that the intervention and outcomes could be further improved. Our goal was to improve engagement with and impact of the Basic Intervention on sun protection, UVR exposure, and engagement in skin cancer examination. To improve effects of the basic intervention, we added several key interactive features and strategies suggested by prior participants, our data, and supported by the literature (ie, by creating a mobile version, adding incentives embedded in the intervention, an enhanced goal-oriented feature, ongoing email and SMS text messages related to previously identified mediators of behavior change, and ongoing news and video updates).

To assess the potential for future dissemination, we conducted a hybrid type II effectiveness-implementation trial [59], which has the goal of increasing efficiency of moving novel evidence-based interventions to the community through mass access and reach. The purpose of this study was to implement the enhanced intervention (enhanced) with young adults aged 18-25 years at moderate to high risk of developing skin cancer and evaluate the intervention's effectiveness in a sample recruited via the web through national dissemination to a general population. We hypothesized that effectiveness and longer-term maintenance effects on UVR and sun protection outcomes would be best for enhanced, second for the original basic condition, and worst for an electronic pamphlet (e-pamphlet) control group. To our knowledge, these are the first entirely digital behavioral skin cancer prevention interventions targeted to healthy young adults recruited from the community that have been evaluated in a national longitudinal randomized controlled trial (RCT) in the United States. If found to be effective, these interventions could be feasibly scaled up for widespread delivery, with the potential to reduce skin cancer risk in a large population of US young adults.

XSL•FO
RenderX

Methods

Procedures

This study assessed 2 digital skin cancer risk reduction programs compared with a control condition for young adults in the United States between the ages of 18 and 25 years (ClinicalTrials.gov NCT03313492). Recruitment efforts are described in detail elsewhere [60]. Briefly, recruitment of study participants was conducted through paid advertising on Facebook, Instagram, and Twitter and unpaid social media posts between September 2018 and April 2019. Audience demographic and behavioral characteristics including age range, location (United States), and interests such as outdoor activities and physical fitness were identified to efficiently target the intended audience. Advertisement content focused on potential heath, appearance, and financial benefits of participation (eg, "Healthy skin is beautiful skin"). The objective of the advertisements was to encourage potential participants to click the call-to-action buttons such as "Sign Up" or "Learn More" that directed them to a study-specific landing web page or sign-up web page. The landing page included relevant images and brief information about the study and reasons why individuals might want to participate, including brief testimonials from prior participants. Individuals were then instructed to create an account on the sign-up web page with a phone number, email, and password. Once an individual indicated their interest and created a study account, they were automatically directed to complete a brief eligibility screener. If eligible, individuals were invited to complete the online informed consent form. After consent, participants were directed to the 10-minute online baseline survey. If enrollees completed the baseline survey, they were automatically directed to their intervention condition and received a US $5 electronic gift card. A computerized program was created to randomize participants on a 2:2:1 basis in blocks of 10 to either basic, enhanced (described below), or a noninteractive skin cancer prevention educational web page modeled after an e-pamphlet from the American Cancer Society [61]. Participants were invited by email, SMS text message, and telephone if necessary to complete online 10- to 15-minute follow-up surveys at 4, 12, 24, and 52 weeks.

Ethical Considerations

This research adhered to appropriate ethical review and approvals as per institutional guidelines (Institutional Review Board #Pro2018001543). Participants provided informed consent electronically and were informed of their ability to opt out. Identifiable data were stored in password-protected files accessible only to approved research personnel. Data were deidentified for analysis. Participants were informed that the total study incentives would be up to US $120 for completing all 5 online surveys over the course of 12 months, plus periodic gift card raffles.

Active Interventions

UV4.me (Basic)

The Integrated Model of Behavioral Prediction (IM) informed the development of the Basic Intervention, for example, by focusing on modification of attitudes, beliefs, perceived norms, and self-efficacy, as well as intentions and behaviors. The basic intervention, described elsewhere in more detail [56], is targeted to young adults, individually tailored, and includes interactive, multimedia, and goal-setting components. It includes 12 modules with content related to a specific topic important in terms of risk or protective behaviors: Why do people tan? To tan or not to tan? Indoor tanning, UVR & looks, UVR & health, Skin cancer, Skin damage, Sunscreen, Shade, Clothes, Skin exams, and Sunless tanning. Several more general sections (eg, My Stuff and Resources) are also included. Tailoring algorithms were created to direct participants to focus on certain modules based on their responses to a few initial questions (eg, the indoor tanning module was recommended for indoor tanners). Throughout the program, participants were asked questions and provided with tailored feedback (eg, "Do you know people who tan? If so, how likely are they to affect your choice to tan or not?"). A number of interactive elements (eg, videos and quizzes) were created to encourage participant engagement with the intervention. For example, at the end of each module, participants could choose to set a goal or not (eg, "I want to put sunscreen on most mornings").

UV4.me2 (Enhanced)

New features and strategies were chosen based on participant feedback from the original UV4.me (eg, strategies suggested to make basic more interactive), our data (eg, number of mobile users who tried to access basic), and evidence-based literature and reviews and models of effective eHealth interventions [62,63] and implementation strategies [64-66] (Table 1). Several enhancements were expected to improve recruitment and representativeness, engagement, and behavioral outcomes:

1. Mobile site: In the original UV4.me study, the intervention was not mobile-friendly [56,58]. For the study described in this paper, both active interventions and the control condition were created using a responsive design, so that they were accessible via a variety of devices including mobile Android and Apple devices. It was thought that a mobile version could improve participant reach directly because more people would have access and may improve effectiveness indirectly by facilitating ongoing engagement.
2. Goal setting: Goal setting addresses several aspects of the IM, for example, intentions, environmental constraints, self-efficacy, and behavior. Users were given the option to commit to a behavioral goal relevant to each module (eg, "cut back on indoor tanning"). They were given tips on goal setting (eg, "set realistic goals"), asked to select from a list of potential reasons why the goal might be important to them (eg, "to avoid getting skin cancer"), strategies to assist them (eg, "check out sunless tanners"), note their progress ("great," "not great," "ok") and challenges (open-ended comments from participants) over time, see their goals summarized on a page called MyGoals, and receive motivating email or SMS text messages (eg, "If you chose a goal, keep working on it") and online feedback (eg, "Keep trying!"). Goal setting is a well-established empirically supported behavior change technique including for internet interventions [67,68]. Both basic and enhanced participants also received 4 automated goal-related messages with decreasing frequency over the year.

XSL•FO
RenderX

3. IM mediator messages: Email or SMS text messages were sent to participants based on IM constructs that were found to mediate UVR exposure and sun protection outcomes in the prior study, for example, knowledge and self-efficacy. Enhanced participants were sent 6 automated mediator messages with decreasing frequency over the course of the year, for example, "Plan ahead, and bring protection with you. Make it part of your routine. Encourage your friends to protect their skin too."
4. Incentives: The enhanced site offered incentives in the form of clickable discount codes (eg, 5%-50% off) and links to sales websites of companies that sell sun protection products (eg, sunscreen, hats, and sunglasses) at the end of each of the 12 main content modules. Incentives can reinforce behavioral implementation, retention, and health behavior change, including in online trials, especially among those initially least motivated or for behaviors that are not intrinsically enjoyable (eg, applying sunscreen) [69-76].
5. Ongoing news updates and video library. Ongoing news and video updates were provided by adding new material (eg, videos, news and media stories, and new research summaries) to the enhanced intervention throughout the study (approximately once per month for 12 months). Updated information was a factor identified in a review or model of effective eHealth interventions [63], is important for dissemination of online interventions [64,65], and has been shown to increase intervention engagement [77,78].

Table 1. Basic and enhanced intervention features.

Intervention features	Basic	Enhanced
Major content		
12 interactive, multimedia modules on topics such as skin cancer risk factors, sun protection, and indoor tanning	✓	✓
Some content tailored to participant demographics and skin cancer risk factors and behaviors		
My Stuff repository of personalized goals and recommendations		
Clickable URLs for skin cancer–related resource websites (eg, Skin Cancer Foundation and American Academy of Dermatology)		
Goal setting		
Basic: prompts users to set behavior change goals (eg, related to sunbathing, indoor tanning, sunburns, and sunscreen use)	✓	✓
Enhanced: prompts users to set behavior change goals, identify barriers, problem-solve, note progress, and receive motivating feedback (eg, "Great job, your tanning is going down over time!"), and reminders via email or SMS text messages		
Video library		
Educational, news, and entertaining videos related to skin cancer prevention (updated at least monthly)		✓
News updates		
Timely news and media articles, new research, charity events, etc related to skin cancer prevention (updated at least monthly)		✓
Product discounts		
Unique coupon discount codes (5%-50% off) for sun protection products (eg, sunscreen, protective clothing, hats, and sunglasses) became available at the completion of each module		✓
Mediator messages		
Six email and SMS text messages were intermittently sent to participants related to theory-informed intervention effect mediators identified in the prior efficacy trial (eg, knowledge and self-efficacy)		✓

Control Condition: E-Pamphlet

A free online American Cancer Society pamphlet on skin cancer prevention and early detection [61] was used as the control condition and was distributed on a noninteractive web page.

Measures

Eligibility Screener and Enrollment

Participants eligible for the study were 18-25 years old, or 19-25 years old if in Alabama or Nebraska, for which 19 years is the age of legal majority. Additional criteria were English speakers, living in a US state or the District of Columbia, had at least weekly internet access, did not have a personal history of skin cancer, and did not report "always" protecting their skin from the sun when outdoors during warm weather. Participants also responded to 11 items related to phenotypic and behavioral risk factors (eg, fair skin and history of sunburns) that indicated moderate to high risk of developing skin cancer (ie, score of ≥27 on the Brief Risk Assessment Tool [BRAT]) [79]. The numbers of participants who were screened, eligible, and consented to participate in the study were assessed (Figure 1). Baseline completion date was noted.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

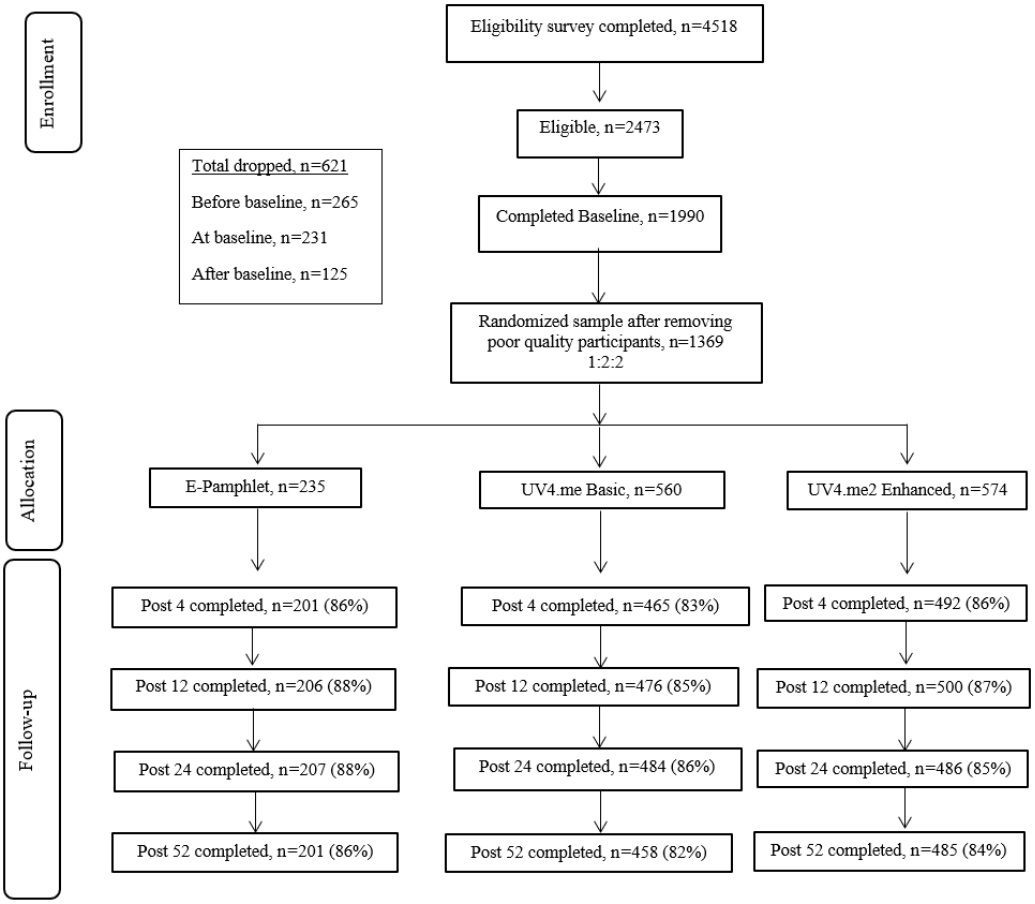

Demographics

Demographic characteristics assessed were age, gender identity, sexual orientation, cohabitation status, race, ethnicity, student status, educational level, employment status, annual income, current climate, and zip code. Weekday time spent on the internet and personal time spent on a mobile device were also assessed using items adapted from the Health Information National Trends Survey [80].

Social Desirability

Social desirability was assessed using the 4-item Brief Social Desirability Scale (eg, “I would never lie to people”), which has demonstrated adequate internal reliability and validity among young adults [81]. Endorsing 3 or 4 of the items indicates high social desirability.

Attention Check

Attention to survey questions was assessed in baseline and all follow-up surveys using the attention check question: “How often in the LAST MONTH did you breathe when you were outside? Please choose ‘not applicable.’ This will verify your careful reading of this item” [82]. This was used as part of an overall data quality score (DQS) described below.

Skin Cancer–Related History

Additional items asked about history of skin cancer among first-degree relatives, the degree of tanning and burning that occurs after 1 week in the sun, and whether participants had received a full body skin cancer examination by a health care provider in the last 12 months [58].

Behavioral Outcomes

Primary Outcomes

Skin cancer–related behavioral outcomes were assessed at baseline and 4 follow-up time points: 4, 12, 24, and 52 weeks. We assessed *sun protection* frequency in the last month (eg, face and body sunscreen use, wide-brimmed hats, long-sleeve shirts, long-sleeve pants, sunglasses, and shade). Response options were 0=never to 4=almost always. Participants were asked how many days in the past month they experienced various types of *UVR exposure* (eg, sunburns, sunbathing, intentional and incidental sun exposure, indoor tanning, clothes that exposes skin to the sun, and tanning products). Summary indexes for sun protection and UVR exposure were created by averaging the items. Items were adapted from Glanz et al [83] and Ingledew et al [84] that our team has cognitively tested and assessed psychometrically with young adults [22].

We adapted the 4 items from the Habit Automaticity subscale of the Self-Report Habit Index to measure behavior patterns associated with the formation of a habit (eg, “Protecting my skin from UVR is something I do without thinking”) [85].

Secondary Outcomes

Specific items including sunburns and indoor tanning were considered secondary outcomes because sunburns and indoor tanning are unique risk factors for skin cancers, and these occur

XSL•FO
RenderX

less frequently than sun exposure and protection in general. Participants were also asked about whether they had been screened for skin cancer by a health care provider or by themselves or a partner. The secondary outcome items were adapted from Glanz et al [83] and Ingledew et al [84] and cognitively tested and assessed psychometrically among young adults [22]. In terms of psychometric evidence for the outcome measures, several studies have demonstrated the reliability and validity of self-report questionnaires of UVR exposure, protection, and detection compared with observation and objective measures with no systematic bias identified among various populations including young adults [86-90].

Data Quality Score

As occurs in other internet-based and survey research, some participants may attempt to enroll more than once or provide inaccurate responses to be deemed eligible and earn study incentives [91-93]. Participants were excluded from this study for giving responses that were very likely of poor quality for issues such as providing a nonunique or nonworking e-mail address or phone number [91]. In addition to removing participants who appeared to attempt to enroll more than once or provided clearly poor-quality data, we used several additional items and responses to assess the quality of the remaining participants. These included responses to the attention check item, social desirability item, 2 items with potentially discrepant responses (ie, work hours and income), responses that were extremely similar to another participant submitted close in time, nonsensical open-ended responses, extremely short survey completion times, reporting on features only included in another condition, reporting that they had seen social media postings related to the study or program since enrolling, or reporting that they knew someone else in the study. Responses to each item were coded as “suspicious” or not. These codes were summed to create a DQS for each participant. Scores could range from 0 to 17, with higher scores indicating worse quality. Participants with a score of 9 or higher were excluded from data analyses. Scores from 0 to 8 were used to weight the outcome data so that higher quality data were given higher weights as described below. DQS scores and procedures were determined before outcome analyses.

Analyses

Effectiveness

We compared demographics across randomization arms using χ^2 tests for categorical variables and heterogeneous variance models for continuous variables to account for potential differences in variance across intervention arms at baseline. The primary outcomes were continuous and reassessed at weeks 4, 12, 26, and 52: sun protection, habitual sun protection, and UVR exposure, all in the last month. The secondary outcomes were binary and reassessed at the same time points as the primary outcomes: sunburn in the last month, indoor tanned in the last month, skin self-exam in the last 3 months, skin exam by a health care provider in the last 12 months, and sunscreen worn on body in the last month. Weighted multivariable mixed effect linear regression together with a robust error variance estimation method was used for analysis of the primary outcomes. The weight was a linear function of the DQS with a value of 1 if DQS=0 and a value of 1/2 if DQS=8. Post hoc analysis based on the model was used to compare the effectiveness of enhanced with basic and enhanced with the e-pamphlet. For analyses of the primary outcomes, the covariates included baseline value, recruitment season (eg, summer, fall, winter, and spring) since some of the behaviors are associated with changes in temperature, residential climate (northern vs southern and tropical), study arm, follow-up time point, and interaction of arm and time point. When the interaction was not significant, it was not included in the final model. For analyses of the secondary outcomes at each time point, covariate-adjusted risk difference estimation based on the logistic regression model proposed by Ge et al [94] was used. The covariates included baseline value, recruitment season (eg, summer, fall, winter, and spring), residential climate (northern vs southern and tropical), and study arm. All the tests were 2-sided with a significance level of 5%.

Missing Data

We also used multiple imputation (MI) methods for missing data to estimate the effectiveness of intervention as sensitivity analyses. MI was based on outcomes of interests and baseline variables: age, gender, race and ethnicity, sexual orientation, family history of skin cancer, level of education, income, climate, cohabitation, BRAT skin cancer total risk score, and number of hours in a day on the internet for personal reasons. A total of 15 imputed data sets were generated by the multivariate sequential regression approach, called the chained equations method [95], using IVEware software [96], and Rubin’s rule [97] was applied to combine results using SAS PROC MIANALYZE [98]. Because the results without MI and with MI were consistent, only results without MI are reported.

Results

Participant Characteristics

Online eligibility surveys were completed by 4518 individuals (Figure 1). Of these, 2473 (54.7%) were deemed eligible. The main reason for being ineligible was a BRAT skin cancer risk score denoting lower risk (1056/4518, 23.4%, of individuals screened not meeting criteria), followed by age other than 18-25 years (767/4518, 17%, not meeting criteria) or self-reporting “always” engaging in sun protection behaviors outside (279/4518, 6.2%, not meeting criteria). Of these, 80.5% (1990/2473) submitted online baseline surveys. The study team dropped 13.7% (621/4518) of potentially eligible participants for attempting to enroll more than once or for poor data quality. This figure is lower than our prior study (22%) [58] but higher than the goal of 5% being careless responders. DQS ranged from 0 to 17 (worse quality), with a median (IQR) score of 3.00 (1.00-5.00) and a mean (SD) score of 3.63 (3.01). Of the 1369 who were randomized to 1 of the 3 intervention conditions, the proportion of completed surveys at follow-up was 1158 (84.6%) at 4 weeks, 1182 (86.3%) at 12 weeks, 1177 (86.0%) at 26 weeks, and 1144 (83.6%) at 52 weeks. These completion rates were similar across the 3 intervention conditions.

At baseline, participants’ average age was 22.3 (SD 2.31) years (Table 2). Approximately 69% (949/1369) of the sample was female, 83% (1141/1369) White non-Hispanic, 73% (999/1369)

XSL•FO
RenderX

heterosexual, 34.3% (470/1369) had at least a college education, 47% (649/1369) earned ≤US $15,000 in the last year, 75% (1021/1369) lived in a state with a northern climate (consistent with classifications of states by climate), and over 87% (1185/1369) lived with someone part or all of the last year. There were no significant differences in these demographic variables across the intervention conditions (*P* values ≥.13), suggesting successful randomization.

Table 2. Participant demographic characteristics at baseline by intervention condition (N=1369).[a]

Characteristics	Total sample (N=1369)	Basic (n=560)	Enhanced (n=574)	E-pamphlet (n=235)
Age (years), mean (SD)	22.30 (2.32)	22.24 (2.32)	22.37 (2.34)	22.26 (2.29)
Gender identity, n (%)				
Male	382 (27.9)	74 (31.5)	155 (27.7)	153 (26.7)
Female	949 (69.3)	156 (66.4)	390 (69.6)	403 (70.2)
Transgender male	35 (2.6)	4 (1.7)	15 (2.7)	16 (2.8)
Transgender female	3 (0.2)	1 (0.4)	0 (0.0)	2 (0.3)
Race and ethnicity, n (%)				
White non-Hispanic	1141 (83.3)	471 (84.7)	476 (83.5)	194 (83.3)
Other	218 (15.9)	85 (15.3)	94 (16.5)	39 (16.7)
Employment status, n (%)				
0 hours (does not work)	123 (9.0)	45 (8.0)	58 (10.1)	20 (8.5)
1-29 hours (part time)	631 (46.1)	254 (45.4)	265 (46.2)	112 (47.7)
>30-39 hours (approximately full time or more)	615 (44.9)	261 (46.6)	251 (43.7)	103 (43.8)
Sexual orientation, n (%)				
Heterosexual	999 (73.0)	404 (72.1)	426 (74.2)	169 (71.9)
Gay, lesbian, bisexual, other	343 (25.1)	144 (25.7)	138 (24.0)	61 (26.0)
Education level, n (%)				
Less than high school or not yet graduated	35 (2.6)	15 (2.7)	12 (2.1)	8 (3.4)
High school graduate or general educational diploma	107 (7.8)	43 (7.7)	42 (7.3)	22 (9.4)
Some college	757 (55.3)	328 (58.6)	311 (54.2)	118 (50.2)
Graduated college	397 (29.0)	146 (26.1)	171 (29.8)	80 (34.0)
Graduate degree or higher	73 (5.3)	28 (5.0)	38 (6.6)	7 (3.0)
Annual income (US $), n (%)				
0-15,000	649 (47.4)	254 (47.2)	288 (51.0)	107 (46.3)
>15,000	685 (51.3)	284 (52.8)	277 (49.0)	124 (53.7)
Current climate, n (%)				
Northern	1021 (74.6)	422 (75.4)	425 (74.3)	174 (74.4)
Southern or Hawaii	345 (25.2)	138 (24.6)	147 (25.7)	60 (25.6)
Cohabitation status, n (%)				
Lived alone	123 (9.0)	47 (8.4)	54 (9.4)	22 (9.4)
Lived with someone else	1185 (86.6)	485 (86.8)	493 (86.0)	207 (88.1)
Lived alone and with someone else (lived with most)	59 (4.3)	27 (4.8)	26 (4.5)	6 (2.6)

[a]Missing data were excluded from Table 2. The greatest percentage of missing data for any variable was 2.6% (income).

Primary and Secondary Outcomes

Mean scores at baseline on primary outcomes were low, with an average score of 1.72 (SD 0.49) for overall sun protection, an average score of 1.48 (SD 0.84) for habitual sun protection on scales from 0 to 4, and an average of 0.72 (SD 1.35) for overall UVR exposure on a scale from 0 to 30 in the last month (Table 3). Baseline values on secondary outcomes were also low, with 11.2% (n=154) reporting at least 1 skin self-exam in the last 3 months, 10.4% (n=143) reporting a physician skin exam in the last 12 months, 31.6% (n=416) reporting wearing

XSL•FO
RenderX

sunscreen on their body at least once in the last month, 6.9% (n=94) reporting indoor tanning at least once in the last month, yet 23.4% (n=321) reporting at least 1 sunburn in the last month (Table 4).

Table 3. Summary descriptive statistics for primary outcomes by intervention and time point (unadjusted).

Primary outcomes (for the last month) and time point (weeks)	Total, mean (SD)	Basic (n=560), mean (SD)	Enhanced (n=574), mean (SD)	E-pamphlet (n=235), mean (SD)	*P* value
Sun protection (0=never to 4=almost always)					
Baseline	1.72 (0.49)	1.73 (0.46)	1.70 (0.51)	1.74 (0.50)	.51
Week 04	1.82 (0.55)	1.80 (0.54)	1.86 (0.58)	1.77 (0.49)	.09
Week 12	1.96 (0.57)	1.96 (0.58)	2.01 (0.58)	1.85 (0.53)	.004
Week 24	2.09 (0.58)	2.10 (0.59)	2.12 (0.59)	2.00 (0.54)	.04
Week 52	1.98 (0.59)	1.98 (0.62)	2.02 (0.58)	1.89 (0.54)	.02
Habitual protection (0=never to 4=almost always)					
Baseline	1.48 (0.84)	1.47 (0.86)	1.49 (0.83)	1.46 (0.84)	.90
Week 04	1.70 (0.90)	1.69 (0.91)	1.75 (0.91)	1.60 (0.84)	.10
Week 12	1.86 (0.95)	1.88 (0.92)	1.90 (0.98)	1.73 (0.92)	.08
Week 24	1.95 (0.94)	1.99 (0.91)	2.00 (0.94)	1.78 (0.99)	.02
Week 52	2.13 (0.97)	2.16 (0.98)	2.14 (0.97)	2.04 (0.97)	.38
UVR exposure (number of days)					
Baseline	0.72 (1.35)	0.74 (1.34)	0.75 (1.45)	0.60 (1.09)	.20
Week 04	0.44 (1.00)	0.43 (0.99)	0.43 (1.01)	0.46 (1.00)	.92
Week 12	0.70 (1.20)	0.66 (1.12)	0.71 (1.20)	0.79 (1.36)	.43
Week 24	1.26 (1.57)	1.31 (1.67)	1.19 (1.47)	1.33 (1.56)	.36
Week 52	0.38 (0.86)	0.39 (0.95)	0.40 (0.80)	0.34 (0.77)	.68

XSL•FO
RenderX

Table 4. Summary descriptive statistics for secondary outcomes by intervention and time point (unadjusted).

Secondary outcomes and visit (weeks)	Total, n (%)	Basic (n=560), n (%)	Enhanced (n=574), n (%)	E-pamphlet (n=235), n (%)	*P* value
Skin self-exam (≥1 time in last 3 months)					
Baseline	154 (11.2)	65 (11.6)	70 (12.2)	19 (8.1)	.23
Week 12	206 (17.4)	91 (19.1)	91 (18.2)	24 (11.7)	.052
Week 24	254 (21.6)	111 (22.9)	107 (22.0)	36 (17.4)	.26
Week 52	272 (23.8)	109 (23.8)	130 (26.8)	33 (16.4)	.01
Physician skin exam (≥1 time in last 12 months)					
Baseline	143 (10.4)	61 (10.9)	64 (11.1)	18 (7.7)	.31
Week 04	137 (11.6)	60 (12.6)	56 (11.1)	21 (10.1)	.61
Week 12	171 (15.5)	79 (16.6)	71 (14.2)	21 (10.2)	.09
Week 24	185 (15.7)	75 (15.5)	80 (16.5)	30 (14.5)	.80
Week 52	229 (20.0)	101 (22.1)	92 (19.0)	36 (17.9)	.35
Sunscreen on body (≥1 time in last month)					
Baseline	416 (31.6)	165 (30.4)	181 (33.3)	70 (30.7)	.56
Week 04	375 (33.1)	143 (31.3)	168 (35.3)	64 (32.3)	.41
Week 12	517 (45.6)	211 (45.6)	233 (49.2)	73 (36.9)	.01
Week 24	857 (73.6)	356 (74.6)	355 (74.0)	146 (70.5)	.52
Week 52	456 (41.6)	179 (40.8)	199 (43.1)	78 (40.0)	.69
Indoor tanning (≥1 time in last month)					
Baseline	94 (6.9)	39 (7.0)	45 (7.8)	10 (4.3)	.19
Week 04	55 (4.6)	23 (4.8)	25 (5.0)	7 (3.4)	.64
Week 12	54 (4.6)	21 (4.4)	23 (4.6)	10 (4.9)	.97
Week 24	39 (3.3)	17 (3.5)	17 (3.5)	5 (2.4)	.73
Week 52	34 (3.0)	17 (3.7)	15 (3.1)	2 (1.0)	.16
Sunburn (≥1 time in last month)					
Baseline	321 (23.4)	145 (25.9)	124 (21.6)	52 (22.1)	.20
Week 04	202 (17.0)	81 (17.0)	79 (15.7)	42 (20.3)	.34
Week 12	357 (30.2)	141 (29.6)	157 (31.4)	59 (28.6)	.72
Week 24	567 (48.2)	238 (49.2)	224 (46.1)	105 (50.7)	.45
Week 52	167 (14.6)	68 (14.8)	74 (15.3)	25 (12.4)	.62

Intervention Effects on Primary Outcomes

Figure 2 shows the error bar plots of the changes from Baseline of the primary outcomes. As the interactions of intervention and time were not significant for any primary outcome ($P \geq .23$), they were not included in the final models. Overall sun protection in the e-pamphlet group was statistically significantly lower than that in the basic or the enhanced group ($P=.007$ and $<.001$ with Cohen d of 0.209 and 0.376, respectively). In addition, sun protection in the enhanced group was statistically significantly higher than that in the basic group ($P=.01$, $d=0.149$). Habitual sun protection in the e-pamphlet group was statistically significantly lower than that in the basic or the enhanced group ($P=.02$ and .004 with d of 0.186 and 0.224, respectively); there was no difference between the basic and the enhanced groups ($P=.46$). All groups reported increased UVR exposure during the summer, and UVR exposure in the e-pamphlet group was greater than that in the basic or the enhanced group, but these differences did not reach statistical significance ($P=.07$ and .09, respectively) In addition, there was no significant difference in UVR between the basic and enhanced groups ($P=.89$).

XSL•FO
RenderX

Figure 2. Intervention effects on primary outcomes. UVR: Ultraviolet radiation.

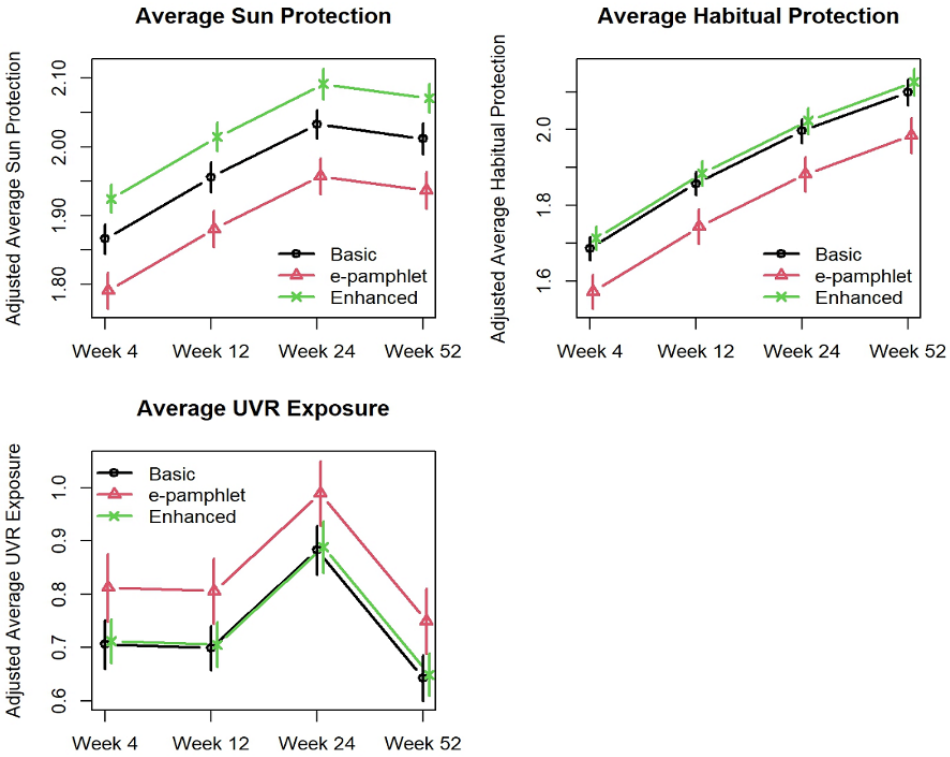

Intervention Effects on Secondary Outcomes

Figure 3 shows the error bar plots of the adjusted proportions for secondary outcomes. For the adjusted proportions of skin self-exam in the past 3 months, the basic group had a significantly higher proportion of individuals who reported doing at least 1 skin self-exam than those in the e-pamphlet group at week 52 (P=.04). The proportion for the enhanced group was also significantly higher than that for the e-pamphlet group at week 52 (P=.004). For the adjusted proportion of individuals who use sunscreen on their body, which increased for all groups during the summer, the proportions for both the basic and enhanced groups were significantly higher than that for the e-pamphlet group at week 12 (P=.03 and .01, respectively). For the adjusted percent of indoor tanning in the past month, the basic group had a significantly higher proportion of individuals who reported indoor tanning than the e-pamphlet group at week 52 (P=.048). All other group comparisons from week 4 to week 52 were not statistically significant.

XSL•FO
RenderX

Figure 3. Intervention effects on secondary outcomes.

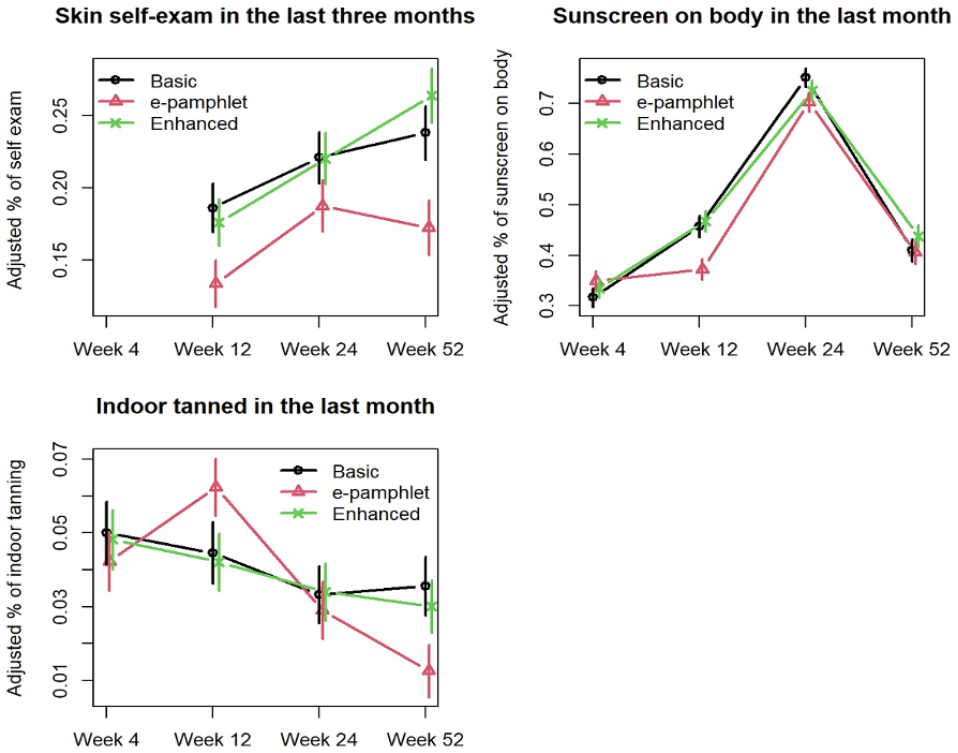

Discussion

This study was a relatively large national hybrid type II effectiveness-implementation RCT of the first, if not only, existing fully digital behavioral intervention that has been demonstrated to modify skin cancer risk behaviors among healthy young adults in the United States at risk of skin cancer recruited from the community with longitudinal follow-ups over the course of a year. Compared with the e-pamphlet, the active interventions increased overall sun protection and habitual sun protection and decreased overall UVR exposure as expected. Similarly, the active interventions also increased sunscreen use worn on the body at the 3-month follow-up (end of summer or early fall) and skin self-examination at 1 year compared with the e-pamphlet. However, the enhanced intervention only differentially improved overall sun protection but did not improve other behaviors more than the basic intervention did. Other outcome effects (eg, for sunburn) were nonsignificant.

Our results suggest that all the automated digital interventions that can be delivered via a smartphone had a significant impact on UVR exposure and protection behaviors as well as skin self-examination. Overall sun protection increased over time for all groups, suggesting that a simple digital pamphlet can produce some behavioral change. However, participants in the e-pamphlet group did engage in (nonstatistically significantly) more UVR exposure at all follow-ups compared with the other groups, so content on UVR exposure may need to be intensified in future e-pamphlets. Additionally, the content on skin self-exam in the active interventions was relatively minimal, which is promising for future interventions addressing this particular behavior. However, the basic intervention increased the physician skin exam significantly in the prior study [58] but not this one.

UVR exposure and sun protection patterns and intervention effects differed by outcome and across time as well as compared with our prior study of the original basic intervention, which demonstrated stronger early effects [58]. Overall sun protection and skin self-exam behavior showed more linear trends over time. UVR exposure, sunscreen use, and sunburns appeared to be more closely associated with changes in seasons. Sunburn patterns were most similar to sunscreen use on the body as opposed to UVR exposure, suggesting that interventions addressing long-term sunscreen use (vs UVR exposure or tanning) may be beneficial to reduce sunburns. Although intervention effects on overall sun protection were fairly stable through 1 year, the differences in the impact of the active interventions on wearing sunscreen compared with e-pamphlets were only significant at 12 weeks and not beyond. Current controversies about potential advantages and disadvantages of mineral, chemical, and hybrid sunscreens for humans and the environment add to the challenges of refining content for such behavioral interventions [99]. The overall proportion of the sample experiencing sunburn decreased from week 4 to 24, but the effects of the interventions did not differ significantly despite the fact that the basic intervention had been differentially efficacious at 12 weeks in the prior study. Indoor tanning decreased overall from week 4 to week 52. However, a higher proportion of the basic group reported indoor tanning at 52 weeks compared with the other intervention conditions. This is a surprising finding and difficult to interpret especially because the material on indoor tanning was similar for both active interventions and was addressed briefly in the e-pamphlet intervention. In the prior study, indoor tanning did not differ by intervention group.

Goal setting and changes in attitudes addressed in the IM mediator messages were associated with improved outcomes

XSL•FO
RenderX

in the prior study. However, compared with the other intervention conditions, the enhancements that were made to the enhanced intervention including enhanced goal setting, embedded incentives, and ongoing mediator messages and video or news updates had a significantly greater impact only on overall sun protection behaviors. There were no other significant differences in effects between the basic and enhanced interventions. This may be due in part to the content of these enhancements. For example, the incentives were mostly related to sun protection (eg, discounts on sunscreen, protective clothing, hats, and sunglasses). Additionally, the incentives were discounts rather than actual money or protective items. Although participants were paid for completing surveys, we do not know whether they used the discounts to purchase sun safety items, which were relatively expensive even with the discounts. All the enhancements were intended to increase intervention engagement over time and thus intervention effects. However, maintaining engagement with digital interventions past an initial exposure is an ongoing challenge for the field. A subsequent paper from this study will explore attitudes and intervention engagement to determine their associations with outcomes. A future noninferiority trial could also assess whether the basic and enhanced interventions produce similar engagement and effects on sun protection if the updated news and video features are removed from the enhanced intervention, thus making it fully automated. Future studies may benefit from more explicit consideration and evaluation of constructs and models related to (digital) intervention engagement behaviors versus health behavior change interventions, either separately or integrated together. Another important difference between the current and prior trial is that the current trial compared 3 interventions, one of which had already been found to be efficacious, which may have resulted in a ceiling effect. The prior study included an assessment-only condition, so it may have been able to demonstrate greater effects.

The limitations of the study include the convenience sample, largely from northern US climates, and that not all participants were recruited during the spring and summer months when risk behaviors are highest. The proportion of recent indoor tanners was relatively low and may limit conclusions related to this outcome. In addition to addressing these limitations, future work should focus on interventions that are less costly to build initially, for example, by using existing social media platforms for the actual interventions. The use of pragmatic trials or more complex, yet potentially efficient, study designs such as factorials or adaptive designs might help determine which aspects of interventions are most cost-effective (eg, goal setting, mediator messages, incentives, or ongoing content updates) for which specific behaviors (eg, sunburn or indoor tanning) and populations. Although some intervention effects seemed to last the full year, more engaging and intense “boosters” may enhance the longitudinal impact. Initial and ongoing engagement and response quality remain challenges for the field.

In conclusion, the strengths of the study include the relatively large national RCT of one of the few digital sun safety interventions for the at-risk young adult population, featuring social media recruitment and longitudinal follow-ups over the course of a year. The automated interventions modified UVR exposure, sunscreen use, overall sun protection, and skin self-examination. However, the effects of the enhanced intervention differed from those of the basic intervention only on sun protection, not other outcomes.

Acknowledgments

The authors would like to thank Maame Araba Assan, MPH, and Baichen Xu, MS, for their assistance with data and manuscript preparation. This study was supported by the National Cancer Institute (grants R01CA204271 and P30CA072720), which was not involved in review or approval of the manuscript for publication.

Data Availability

The data set generated during this study is available from the corresponding author upon reasonable request.

Authors' Contributions

CJH was responsible for conceptualization, funding acquisition, and writing the original draft. YL was responsible for formal analyses and writing the original draft. OK contributed to writing by reviewing and editing. AM was responsible for project administration and writing by reviewing and editing. MR was responsible for project administration. SLM, ALY, ZN, and KG contributed to writing by reviewing and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.2).
[PDF File (Adobe PDF File), 93 KB - jmir_v26i1e55831_app1.pdf]

References

XSL•FO
RenderX

1. Watson M, Garnett E, Guy GP, Holman D. The surgeon general's call to action to prevent skin cancer. Centers for Disease Control and Prevention. Washington, DC: US Dept of Health and Human Services; 2014. URL: https://stacks.cdc.gov/view/cdc/24419 [accessed 2024-04-10]
2. Cancer facts and figures 2009. American Cancer Society. 2009. URL: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2009.html [accessed 2024-04-10]
3. Donaldson MR, Coldiron BM. No end in sight: the skin cancer epidemic continues. Semin Cutan Med Surg 2011;30(1):3-5 [FREE Full text] [doi: 10.1016/j.sder.2011.01.002] [Medline: 21540015]
4. Linos E, Swetter SM, Cockburn MG, Colditz GA, Clarke CA. Increasing burden of melanoma in the United States. J Invest Dermatol 2009;129(7):1666-1674 [FREE Full text] [doi: 10.1038/jid.2008.423] [Medline: 19131946]
5. Rogers HW, Weinstock MA, Harris AR, Hinckley MR, Feldman SR, Fleischer AB, et al. Incidence estimate of nonmelanoma skin cancer in the United States, 2006. Arch Dermatol 2010;146(3):283-287 [FREE Full text] [doi: 10.1001/archdermatol.2010.19] [Medline: 20231499]
6. Sutton A, Crew A, Wysong A. Redefinition of skin cancer as a chronic disease. JAMA Dermatol 2016;152(3):255-256 [FREE Full text] [doi: 10.1001/jamadermatol.2015.4215] [Medline: 26676205]
7. Basal and squamous cell skin cancer. American Cancer Society. 2024. URL: https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer.html [accessed 2024-04-10]
8. Baxter AJ, Hughes MC, Kvaskoff M, Siskind V, Shekar S, Aitken JF, et al. The Queensland study of Melanoma: Environmental and Genetic Associations (Q-MEGA): study design, baseline characteristics, and repeatability of phenotype and sun exposure measures. Twin Res Hum Genet 2008;11(2):183-196 [FREE Full text] [doi: 10.1375/twin.11.2.183] [Medline: 18361720]
9. Skin cancer risk factors. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/skin-cancer/risk-factors/index.html [accessed 2024-06-05]
10. Chuang TY, Brashear R. Risk factors of non-melanoma skin cancer in United States veterans patients: a pilot study and review of literature. J Eur Acad Dermatol Venereol 1999;12(2):126-132. [Medline: 10343941]
11. Goldberg MS, Doucette JT, Lim HW, Spencer J, Carucci JA, Rigel DS. Risk factors for presumptive melanoma in skin cancer screening: American Academy of Dermatology National Melanoma/Skin Cancer Screening Program experience 2001-2005. J Am Acad Dermatol 2007;57(1):60-66 [FREE Full text] [doi: 10.1016/j.jaad.2007.02.010] [Medline: 17490783]
12. Grulich AE, Bataille V, Swerdlow AJ, Newton-Bishop JA, Cuzick J, Hersey P, et al. Naevi and pigmentary characteristics as risk factors for melanoma in a high-risk population: a case-control study in New South Wales, Australia. Int J Cancer 1996;67(4):485-491 [FREE Full text] [doi: 10.1002/(SICI)1097-0215(19960807)67:4<485::AID-IJC4>3.0.CO;2-O] [Medline: 8759605]
13. Markovic SN, Erickson LA, Rao RD, Weenig RH, Pockaj BA, Bardia A, et al. Malignant melanoma in the 21st century, part 1: epidemiology, risk factors, screening, prevention, and diagnosis. Mayo Clin Proc 2007;82(3):364-380 [FREE Full text] [doi: 10.4065/82.3.364] [Medline: 17352373]
14. Oberyszyn TM. Non-melanoma skin cancer: importance of gender, immunosuppressive status and vitamin D. Cancer Lett 2008;261(2):127-136 [FREE Full text] [doi: 10.1016/j.canlet.2008.01.009] [Medline: 18267352]
15. Perkins JL, Liu Y, Mitby PA, Neglia JP, Hammond S, Stovall M, et al. Nonmelanoma skin cancer in survivors of childhood and adolescent cancer: a report from the childhood cancer survivor study. J Clin Oncol 2005;23(16):3733-3741 [FREE Full text] [doi: 10.1200/JCO.2005.06.237] [Medline: 15923570]
16. Psaty EL, Scope A, Halpern AC, Marghoob AA. Defining the patient at high risk for melanoma. Int J Dermatol 2010;49(4):362-376 [FREE Full text] [doi: 10.1111/j.1365-4632.2010.04381.x] [Medline: 20465687]
17. Qureshi AA, Zhang M, Han J. Heterogeneity in host risk factors for incident melanoma and non-melanoma skin cancer in a cohort of US women. J Epidemiol 2011;21(3):197-203 [FREE Full text] [doi: 10.2188/jea.je20100145] [Medline: 21515942]
18. Siskind V, Aitken J, Green A, Martin N. Sun exposure and interaction with family history in risk of melanoma, Queensland, Australia. Int J Cancer 2002;97(1):90-95 [FREE Full text] [doi: 10.1002/ijc.1563] [Medline: 11774248]
19. Etzkorn JR, Parikh RP, Marzban SS, Law K, Davis AH, Rawal B, et al. Identifying risk factors using a skin cancer screening program. Cancer Control 2013;20(4):248-254 [FREE Full text] [doi: 10.1177/107327481302000402] [Medline: 24077401]
20. Armstrong BK, Kricker A. How much melanoma is caused by sun exposure? Melanoma Res 1993;3(6):395-401 [FREE Full text] [doi: 10.1097/00008390-199311000-00002] [Medline: 8161879]
21. Dusza SW, Halpern AC, Satagopan JM, Oliveria SA, Weinstock MA, Scope A, et al. Prospective study of sunburn and sun behavior patterns during adolescence. Pediatrics 2012;129(2):309-317 [FREE Full text] [doi: 10.1542/peds.2011-0104] [Medline: 22271688]
22. Lazovich D, Stryker JE, Mayer JA, Hillhouse J, Dennis LK, Pichon L, et al. Measuring nonsolar tanning behavior: indoor and sunless tanning. Arch Dermatol 2008;144(2):225-230 [FREE Full text] [doi: 10.1001/archdermatol.2007.45] [Medline: 18283180]
23. Skin cancer prevention. Skin Cancer Foundation. 2023. URL: http://www.skincancer.org/prevention/sun-protection/prevention-guidelines/preventing-skin-cancer [accessed 2024-04-10]

XSL•FO
RenderX

24. Swerdlow AJ, Weinstock MA. Do tanning lamps cause melanoma? An epidemiologic assessment. J Am Acad Dermatol 1998;38(1):89-98 [FREE Full text] [doi: 10.1016/s0190-9622(98)70544-4] [Medline: 9448211]
25. Vishvakarman D, Wong JCF. Description of the use of a risk estimation model to assess the increased risk of non-melanoma skin cancer among outdoor workers in Central Queensland, Australia. Photodermatol Photoimmunol Photomed 2003;19(2):81-88 [FREE Full text] [doi: 10.1034/j.1600-0781.2003.00012.x] [Medline: 12945807]
26. Gies PH, Roy CR, Toomey S, McLennan A. Protection against solar ultraviolet radiation. Mutat Res 1998;422(1):15-22 [FREE Full text] [doi: 10.1016/s0027-5107(98)00181-x] [Medline: 9920424]
27. Moloney FJ, Collins S, Murphy GM. Sunscreens: safety, efficacy and appropriate use. Am J Clin Dermatol 2002;3(3):185-191 [FREE Full text] [doi: 10.2165/00128071-200203030-00005] [Medline: 11978139]
28. Yagerman S, Marghoob A. Melanoma patient self-detection: a review of efficacy of the skin self-examination and patient-directed educational efforts. Expert Rev Anticancer Ther 2013;13(12):1423-1431 [FREE Full text] [doi: 10.1586/14737140.2013.856272] [Medline: 24236821]
29. Melanoma. American Cancer Society. 2019. URL: https://www.cancer.org/cancer/types/cancer-in-young-adults/cancers-in-young-adults.html#:~:text=Melanoma%20is%20a%20type%20of,occur%20at%20a%20younger%20age [accessed 2024-04-10]
30. Stanton WR, Janda M, Baade PD, Anderson P. Primary prevention of skin cancer: a review of sun protection in Australia and internationally. Health Promot Int 2004;19(3):369-378 [FREE Full text] [doi: 10.1093/heapro/dah310] [Medline: 15306621]
31. MacNeal RJ, Dinulos JGH. Update on sun protection and tanning in children. Curr Opin Pediatr 2007;19(4):425-429 [FREE Full text] [doi: 10.1097/MOP.0b013e3282294936] [Medline: 17630607]
32. Lim HW, Cyr WH, DeFabo E, Robinson J, Weinstock MA, Beer JZ, et al. Scientific and regulatory issues related to indoor tanning. J Am Acad Dermatol 2004;51(5):781-784 [FREE Full text] [doi: 10.1016/j.jaad.2004.08.017] [Medline: 15523359]
33. Coups EJ, Manne SL, Heckman CJ. Multiple skin cancer risk behaviors in the U.S. population. Am J Prev Med 2008;34(2):87-93 [FREE Full text] [doi: 10.1016/j.amepre.2007.09.032] [Medline: 18201637]
34. Heckman CJ, Coups EJ, Manne SL. Prevalence and correlates of indoor tanning among US adults. J Am Acad Dermatol 2008;58(5):769-780 [FREE Full text] [doi: 10.1016/j.jaad.2008.01.020] [Medline: 18328594]
35. Freedman KS, Nelson NM, Feldman LL. Smoking initiation among young adults in the United States and Canada, 1998-2010: a systematic review. Prev Chronic Dis 2012;9:E05 [FREE Full text] [Medline: 22172172]
36. Andrews JA, Tildesley E, Hops H, Li F. The influence of peers on young adult substance use. Health Psychol 2002;21(4):349-357. [doi: 10.1037//0278-6133.21.4.349] [Medline: 12090677]
37. Galliher RV, Kerpelman JL. The intersection of identity development and peer relationship processes in adolescence and young adulthood: contributions of the special issue. J Adolesc 2012;35(6):1409-1415. [doi: 10.1016/j.adolescence.2012.09.007] [Medline: 23069622]
38. Bleakley A, Jordan AB, Strasser AA, Lazovich D, Glanz K. Testing general versus specific behavioral focus in messaging for the promotion of sun protection behaviors. Ann Behav Med 2020;54(2):108-118 [FREE Full text] [doi: 10.1093/abm/kaz032] [Medline: 31586204]
39. Darlow S, Heckman C. Results from a tailored SMS and behavior-tracking pilot study on sun-safe behaviors in young women. Health Educ Behav 2017;44(6):937-944 [FREE Full text] [doi: 10.1177/1090198117699507] [Medline: 28363246]
40. Lin JS, Eder M, Weinmann S. Behavioral counseling to prevent skin cancer: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2011;154(3):190-201 [FREE Full text] [doi: 10.7326/0003-4819-154-3-201102010-00009] [Medline: 21282699]
41. Mays D, Tercyak KP. Framing indoor tanning warning messages to reduce skin cancer risks among young women: implications for research and policy. Am J Public Health 2015;105(8):e70-e76 [FREE Full text] [doi: 10.2105/AJPH.2015.302665] [Medline: 26066932]
42. Niu Z, Jeong DC, Coups EJ, Stapleton JL. An experimental investigation of human presence and mobile technologies on college students' sun protection intentions: between-subjects study. JMIR Mhealth Uhealth 2019;7(8):e13720 [FREE Full text] [doi: 10.2196/13720] [Medline: 31452523]
43. Robinson JK, Patel S, Heo SY, Gray E, Lim J, Kwon K, et al. Real-time UV measurement with a sun protection system for warning young adults about sunburn: prospective cohort study. JMIR Mhealth Uhealth 2021;9(5):e25895 [FREE Full text] [doi: 10.2196/25895] [Medline: 33955844]
44. Saraiya M, Glanz K, Briss PA, Nichols P, White C, Das D, et al. Interventions to prevent skin cancer by reducing exposure to ultraviolet radiation: a systematic review. Am J Prev Med 2004;27(5):422-466 [FREE Full text] [doi: 10.1016/j.amepre.2004.08.009] [Medline: 15556744]
45. Williams AL, Grogan S, Clark-Carter D, Buckley E. Appearance-based interventions to reduce ultraviolet exposure and/or increase sun protection intentions and behaviours: a systematic review and meta-analyses. Br J Health Psychol 2013;18(1):182-217 [FREE Full text] [doi: 10.1111/j.2044-8287.2012.02089.x] [Medline: 22989352]
46. Share of adults in the United States who use the internet as of September 2023, by age group. Statista. 2023. URL: https://www.statista.com/statistics/266587/percentage-of-internet-users-by-age-groups-in-the-us/ [accessed 2023-04-10]

XSL•FO
RenderX

47. Mobile Fact Sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-04-10]
48. Social Media Use in 2021. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-04-10]
49. Whitaker C, Stevelink S, Fear N. The use of Facebook in recruiting participants for health research purposes: a systematic review. J Med Internet Res 2017;19(8):e290 [FREE Full text] [doi: 10.2196/jmir.7071] [Medline: 28851679]
50. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020;22(10):e22810 [FREE Full text] [doi: 10.2196/22810] [Medline: 33104015]
51. Lafferty NT, Manca A. Perspectives on social media in and as research: a synthetic review. Int Rev Psychiatry 2015;27(2):85-96 [FREE Full text] [doi: 10.3109/09540261.2015.1009419] [Medline: 25742363]
52. Topolovec-Vranic J, Natarajan K. The use of social media in recruitment for medical research studies: a scoping review. J Med Internet Res 2016;18(11):e286 [FREE Full text] [doi: 10.2196/jmir.5698] [Medline: 27821383]
53. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why are health care interventions delivered over the internet? A systematic review of the published literature. J Med Internet Res 2006;8(2):e10 [FREE Full text] [doi: 10.2196/jmir.8.2.e10] [Medline: 16867965]
54. Tate DF, Finkelstein EA, Khavjou O, Gustafson A. Cost effectiveness of internet interventions: review and recommendations. Ann Behav Med 2009;38(1):40-45 [FREE Full text] [doi: 10.1007/s12160-009-9131-6] [Medline: 19834778]
55. Niu Z, Bhurosy T, Heckman CJ. Digital interventions for promoting sun protection and skin self-examination behaviors: a systematic review. Prev Med Rep 2022;26:101709 [FREE Full text] [doi: 10.1016/j.pmedr.2022.101709] [Medline: 35529530]
56. Heckman C, Darlow S, Munshi T, Caruso C, Ritterband L, Raivitch S, et al. Development of an internet intervention to address behaviors associated with skin cancer risk among young adults. Internet Interv 2015;2(3):340-350 [FREE Full text] [doi: 10.1016/j.invent.2015.04.003] [Medline: 26640776]
57. Fishbein M. A reasoned action approach to health promotion. Med Decis Making 2008;28(6):834-844 [FREE Full text] [doi: 10.1177/0272989X08326092] [Medline: 19015289]
58. Heckman CJ, Darlow SD, Ritterband LM, Handorf EA, Manne SL. Efficacy of an intervention to alter skin cancer risk behaviors in young adults. Am J Prev Med 2016;51(1):1-11 [FREE Full text] [doi: 10.1016/j.amepre.2015.11.008] [Medline: 26810358]
59. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012;50(3):217-226 [FREE Full text] [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
60. Heckman CJ, Riley M, Khavjou O, Ohman-Strickland P, Manne SL, Yaroch AL, et al. Cost, reach, and representativeness of recruitment efforts for an online skin cancer risk reduction intervention trial for young adults. Transl Behav Med 2021;11(10):1875-1884 [FREE Full text] [doi: 10.1093/tbm/ibab047] [Medline: 34160622]
61. Be safe in the Sun. American Cancer Society. URL: https://www.cancer.org/healthy/be-safe-in-sun.html [accessed 2024-04-10]
62. Litvin EB, Abrantes AM, Brown RA. Computer and mobile technology-based interventions for substance use disorders: an organizing framework. Addict Behav 2013;38(3):1747-1756 [FREE Full text] [doi: 10.1016/j.addbeh.2012.09.003] [Medline: 23254225]
63. Morrison LG, Yardley L, Powell J, Michie S. What design features are used in effective e-health interventions? A review using techniques from critical interpretive synthesis. Telemed J E Health 2012;18(2):137-144 [FREE Full text] [doi: 10.1089/tmj.2011.0062] [Medline: 22381060]
64. Brouwer W, Oenema A, Crutzen R, de Nooijer J, de Vries NK, Brug J. An exploration of factors related to dissemination of and exposure to internet-delivered behavior change interventions aimed at adults: a delphi study approach. J Med Internet Res 2008;10(2):e10 [FREE Full text] [doi: 10.2196/jmir.956] [Medline: 18417443]
65. Crutzen R, de Nooijer J, Brouwer W, Oenema A, Brug J, de Vries NK. Internet-delivered interventions aimed at adolescents: a Delphi study on dissemination and exposure. Health Educ Res 2008;23(3):427-439 [FREE Full text] [doi: 10.1093/her/cym094] [Medline: 18209115]
66. Powell BJ, McMillen JC, Proctor EK, Carpenter CR, Griffey RT, Bunger AC, et al. A compilation of strategies for implementing clinical innovations in health and mental health. Med Care Res Rev 2012;69(2):123-157 [FREE Full text] [doi: 10.1177/1077558711430690] [Medline: 22203646]
67. Manzoni GM, Pagnini F, Corti S, Molinari E, Castelnuovo G. Internet-based behavioral interventions for obesity: an updated systematic review. Clin Pract Epidemiol Ment Health 2011;7:19-28 [FREE Full text] [doi: 10.2174/1745017901107010019] [Medline: 21552423]
68. Prestwich A, Kellar I, Parker R, MacRae S, Learmonth M, Sykes B, et al. How can self-efficacy be increased? Meta-analysis of dietary interventions. Health Psychol Rev 2014;8(3):270-285 [FREE Full text] [doi: 10.1080/17437199.2013.813729] [Medline: 25053214]

XSL•FO
RenderX

69. Crutzen R, de Nooijer J, Brouwer W, Oenema A, Brug J, de Vries NK. Strategies to facilitate exposure to internet-delivered health behavior change interventions aimed at adolescents or young adults: a systematic review. Health Educ Behav 2011;38(1):49-62 [FREE Full text] [doi: 10.1177/1090198110372878] [Medline: 21189422]
70. Giles EL, Robalino S, McColl E, Sniehotta FF, Adams J. The effectiveness of financial incentives for health behaviour change: systematic review and meta-analysis. PLoS One 2014;9(3):e90347 [FREE Full text] [doi: 10.1371/journal.pone.0090347] [Medline: 24618584]
71. Mantzari E, Vogt F, Marteau TM. Financial incentives for increasing uptake of HPV vaccinations: a randomized controlled trial. Health Psychol 2015;34(2):160-171 [FREE Full text] [doi: 10.1037/hea0000088] [Medline: 25133822]
72. Mantzari E, Vogt F, Shemilt I, Wei Y, Higgins JPT, Marteau TM. Personal financial incentives for changing habitual health-related behaviors: a systematic review and meta-analysis. Prev Med 2015;75:75-85 [FREE Full text] [doi: 10.1016/j.ypmed.2015.03.001] [Medline: 25843244]
73. Alexander GL, Divine GW, Couper MP, McClure JB, Stopponi MA, Fortman KK, et al. Effect of incentives and mailing features on online health program enrollment. Am J Prev Med 2008;34(5):382-388 [FREE Full text] [doi: 10.1016/j.amepre.2008.01.028] [Medline: 18407004]
74. Khadjesari Z, Murray E, Kalaitzaki E, White IR, McCambridge J, Thompson SG, et al. Impact and costs of incentives to reduce attrition in online trials: two randomized controlled trials. J Med Internet Res 2011;13(1):e26 [FREE Full text] [doi: 10.2196/jmir.1523] [Medline: 21371988]
75. Schierenberg A, van Amsterdam J, van den Brink W, Goudriaan AE. Efficacy of contingency management for cocaine dependence treatment: a review of the evidence. Curr Drug Abuse Rev 2012;5(4):320-331 [FREE Full text] [doi: 10.2174/1874473711205040006] [Medline: 23244344]
76. Young JM, O'Halloran A, McAulay C, Pirotta M, Forsdike K, Stacey I, et al. Unconditional and conditional incentives differentially improved general practitioners' participation in an online survey: randomized controlled trial. J Clin Epidemiol 2015;68(6):693-697 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.09.013] [Medline: 25450450]
77. Brouwer W, Kroeze W, Crutzen R, de Nooijer J, de Vries NK, Brug J, et al. Which intervention characteristics are related to more exposure to internet-delivered healthy lifestyle promotion interventions? A systematic review. J Med Internet Res 2011;13(1):e2 [FREE Full text] [doi: 10.2196/jmir.1639] [Medline: 21212045]
78. Franklin VL, Greene A, Waller A, Greene SA, Pagliari C. Patients' engagement with "Sweet Talk"—a text messaging support system for young people with diabetes. J Med Internet Res 2008;10(2):e20 [FREE Full text] [doi: 10.2196/jmir.962] [Medline: 18653444]
79. Glanz K, Schoenfeld E, Weinstock MA, Layi G, Kidd J, Shigaki DM. Development and reliability of a brief skin cancer risk assessment tool. Cancer Detect Prev 2003;27(4):311-315 [FREE Full text] [doi: 10.1016/s0361-090x(03)00094-1] [Medline: 12893080]
80. What is HINTS? National Cancer Institute. Health Information National Trends Survey. URL: http://hints.Cancer.Gov [accessed 2024-04-10]
81. Haghighat R. The development of the Brief Social Desirability Scale (BSDS). Eur J Psychol 2007;3(4):10-5964 [FREE Full text] [doi: 10.5964/ejop.v3i4.417]
82. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on amazon mechanical turk. Behav Res Methods 2014;46(4):1023-1031 [FREE Full text] [doi: 10.3758/s13428-013-0434-y] [Medline: 24356996]
83. Glanz K, Yaroch AL, Dancel M, Saraiya M, Crane LA, Buller DB, et al. Measures of sun exposure and sun protection practices for behavioral and epidemiologic research. Arch Dermatol 2008;144(2):217-222 [FREE Full text] [doi: 10.1001/archdermatol.2007.46] [Medline: 18283179]
84. Ingledew DK, Ferguson E, Markland D. Motives and sun-related behaviour. J Health Psychol 2010;15(1):8-20 [FREE Full text] [doi: 10.1177/1359105309342292] [Medline: 20064880]
85. Gardner B, Abraham C, Lally P, de Bruijn GJ. Towards parsimony in habit measurement: testing the convergent and predictive validity of an automaticity subscale of the self-report habit index. Int J Behav Nutr Phys Act 2012;9:102 [FREE Full text] [doi: 10.1186/1479-5868-9-102] [Medline: 22935297]
86. Glanz K, Mayer JA. Reducing ultraviolet radiation exposure to prevent skin cancer methodology and measurement. Am J Prev Med 2005;29(2):131-142 [FREE Full text] [doi: 10.1016/j.amepre.2005.04.007] [Medline: 16005810]
87. Glanz K, McCarty F, Nehl EJ, O'Riordan DL, Gies P, Bundy L, et al. Validity of self-reported sunscreen use by parents, children, and lifeguards. Am J Prev Med 2009;36(1):63-69 [FREE Full text] [doi: 10.1016/j.amepre.2008.09.012] [Medline: 18945582]
88. Glanz K, Gies P, O'Riordan DL, Elliott T, Nehl E, McCarty F, et al. Validity of self-reported solar UVR exposure compared with objectively measured UVR exposure. Cancer Epidemiol Biomarkers Prev 2010;19(12):3005-3012 [FREE Full text] [doi: 10.1158/1055-9965.EPI-10-0709] [Medline: 20940277]
89. O'Riordan DL, Glanz K, Gies P, Elliott T. A pilot study of the validity of self-reported ultraviolet radiation exposure and sun protection practices among lifeguards, parents and children. Photochem Photobiol 2008;84(3):774-778 [FREE Full text] [doi: 10.1111/j.1751-1097.2007.00262.x] [Medline: 18179624]

XSL•FO
RenderX

90. O'Riordan DL, Nehl E, Gies P, Bundy L, Burgess K, Davis E, et al. Validity of covering-up sun-protection habits: association of observations and self-report. J Am Acad Dermatol 2009;60(5):739-744 [FREE Full text] [doi: 10.1016/j.jaad.2008.12.015] [Medline: 19278750]
91. Bowen AM, Daniel CM, Williams ML, Baird GL. Identifying multiple submissions in internet research: preserving data integrity. AIDS Behav 2008;12(6):964-973 [FREE Full text] [doi: 10.1007/s10461-007-9352-2] [Medline: 18240015]
92. Heckman CJ, Handorf E, Auerbach MV. Prevalence and correlates of skin cancer screening among indoor tanners and nontanners. JAMA Dermatol 2018;154(5):554-560 [FREE Full text] [doi: 10.1001/jamadermatol.2018.0163] [Medline: 29617518]
93. Meade AW, Craig SB. Identifying careless responses in survey data. Psychol Methods 2012;17(3):437-455. [doi: 10.1037/a0028085] [Medline: 22506584]
94. Ge M, Durham LK, Meyer RD, Xie W, Thomas N. Covariate-adjusted difference in proportions from clinical trials using logistic regression and weighted risk differences. Drug Information J 2011;45(4):481-493 [FREE Full text] [doi: 10.1177/009286151104500409]
95. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30(4):377-399 [FREE Full text] [doi: 10.1002/sim.4067] [Medline: 21225900]
96. IVEware: Imputation and Variance Estimation Software. University of Michigan Institute for Social Research Survey Research Center. 2024. URL: https://src.isr.umich.edu/iveware/ [accessed 2024-04-10]
97. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987.
98. SAS/STAT 13.1 User's Guide the MIANALYZE Procedure. Cary, NC: SAS Instutite Inc; 2013.
99. Health and Medicine Division, Division on Earth and Life Studies, National Academies of Sciences, Engineering, and Medicine, Board on Health Sciences Policy, Board on Environmental Studies and Toxicology, Ocean Studies Board. Review of Fate, Exposure, and Effects of Sunscreens in Aquatic Environments and Implications for Sunscreen Usage and Human Health. Washington, DC: National Academies Press; 2022.

Abbreviations

BRAT: Brief Risk Assessment Tool
CONSORT: Consolidated Standards of Reporting Trials
DQS: data quality score
IM: Integrated Model of Behavioral Prediction
MI: multiple imputation
RCT: randomized controlled trial
UVR: UV radiation

Edited by S Ma; submitted 29.12.23; peer-reviewed by J Brzozowska, A King, J Makin; comments to author 14.03.24; revised version received 01.04.24; accepted 10.04.24; published 02.07.24.

Please cite as:
Heckman CJ, Mitarotondo A, Lin Y, Khavjou O, Riley M, Manne SL, Yaroch AL, Niu Z, Glanz K
Digital Interventions to Modify Skin Cancer Risk Behaviors in a National Sample of Young Adults: Randomized Controlled Trial
J Med Internet Res 2024;26:e55831
URL: https://www.jmir.org/2024/1/e55831
doi:10.2196/55831
PMID:

©Carolyn J Heckman, Anna Mitarotondo, Yong Lin, Olga Khavjou, Mary Riley, Sharon L Manne, Amy L Yaroch, Zhaomeng Niu, Karen Glanz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effects of a Complex Interactive Multimodal Intervention on Personalized Stress Management Among Health Care Workers in China: Nonrandomized Controlled Study

Wenhua Liu[1,2*], MPH; Quan Wang[2*], PhD; Danli Zheng[2*], MS; Junhua Mei[3], MD; Jiajia Lu[3], MD; Guohua Chen[3], MD, PhD; Wei Wang[2], MD, PhD; Fengfei Ding[1,2], MD

[1]Department of Pharmacology, School of Basic Medical Science, Shanghai Medical College, Fudan University, Shanghai, China

[2]Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[3]Department of Neurology, Wuhan No.1 Hospital, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Fengfei Ding, MD
Department of Pharmacology
School of Basic Medical Science, Shanghai Medical College
Fudan University
No.130 Dong An Road, Xu hui district
Shanghai, 20032
China
Phone: 86 13476255813
Email: Fengfei_ding@fudan.edu.cn

Abstract

Background: Health care workers (HCWs) frequently face multiple stressors at work, particularly those working night shifts. HCWs who have experienced distress may find it difficult to adopt stress management approaches, even if they are aware of the effects of stress and coping processes. Therefore, an individualized intervention may be required to assist distressed HCWs in bridging the “knowledge-practice” gap in stress management and effectively alleviating stress symptoms.

Objective: The main objective of this research was to compare the effects of a complex interactive multimodal intervention (CIMI) to self-guided stress management interventions on stress symptoms of distressed HCWs, as measured by physiological (heart rate variability), psychological (perceived stress, mental distress, and subjective happiness), and sleep disorder (fatigue and sleepiness) indicators.

Methods: We conducted a nonrandomized, controlled study in 2 Chinese general hospitals. The participants in this study were 245 HCWs who fulfilled at least 1 of the 3 dimensions on the Depression, Anxiety, and Stress Scale. All eligible individuals were required to complete a questionnaire and wear a 24-hour Holter device to determine the physiological signs of stress as indexed by heart rate variability at both baseline and after the intervention. The CIMI group received a 12-week online intervention with 4 components—mobile stress management instruction, a web-based WeChat social network, personalized feedback, and a nurse coach, whereas the control group simply received a self-guided intervention.

Results: After a 12-week intervention, the Perceived Stress Scale (PSS) scores reduced significantly in the CIMI group (mean difference [MD] –5.31, 95% CI –6.26 to –4.37; $P<.001$) compared to the baseline levels. The changes in PSS scores before and after the intervention exhibited a significant difference between the CIMI and control groups ($d=-0.64$; MD –4.03, 95% CI –5.91 to –2.14; $P<.001$), and the effect was medium. In terms of physiological measures, both the control group (MD –9.56, 95% CI –16.9 to –2.2; $P=.01$) and the CIMI group (MD –8.45, 95% CI –12.68 to –4.22; $P<.001$) demonstrated a significant decrease in the standard deviation of normal-to-normal intervals (SDNN) within the normal clinical range; however, there were no significant differences between the 2 groups ($d=0.03$; MD 1.11, 95% CI –7.38 to 9.59; $P=.80$).

Conclusions: The CIMI was an effective intervention for improving sleep disorders, as well as parts of the psychological stress measures in distressed HCWs. The findings provide objective evidence for developing a mobile stress management intervention that is adaptable and accessible to distressed HCWs, but its long-term effects should be investigated in future research.

Trial Registration: ClinicalTrials.gov NCT05239065; https://clinicaltrials.gov/ct2/show/NCT05239065

XSL•FO
RenderX

(J Med Internet Res 2024;26:e45422) doi:10.2196/45422

KEYWORDS

multimodal intervention; stress management; health care workers; perceived stress; autonomic nervous system; stress; management; mental health; engagement; human support; physiological stress; psychological stress; social network; mobile phone

Introduction

Overview

As a result of shift schedules, night shift health care employees face multiple workplace stressors, including insufficient rest periods, an increased risk of sleep deprivation, and mood changes [1-3]. Work-related stress has been shown to have a negative impact on employees' physical health, including a greater likelihood of developing hypertension, the advancement of cardiovascular diseases, and all-cause and coronary heart disease mortality as measured by heart rate variability (HRV) [4-7]. The link between work-related stress and psychological distress is also well established [8], and according to a recent meta-analysis, the prevalence of depression, anxiety, and stress symptoms among health care workers (HCWs) was 24.3%, 25.8%, and 45%, respectively, which was higher than in other occupational populations [9]. Furthermore, HCWs who are under stress at work are more likely to experience sleep problems such as fatigue and sleepiness [10], which can jeopardize their job performance and patient safety.

During the past decade, numerous workplace-based stress management programs have been conducted in high-income countries and achieved the desired effect in improving the mental health and stress symptoms of staff [11,12]. However, stress management interventions remained scarce in the context of low- and middle-income countries [11]. On the one hand, due to the frequent conflict between increased medical service demand and limited health care resources, HCWs in these countries experienced a higher level of work-related stress than their counterparts [13]. On the other hand, limited medical resources made it difficult for distressed HCWs to seek professional help, and the current treatment gap for mental disorders in low- and middle-income countries is estimated to be from 72% to 93%, with those from the poorest countries receiving less coverage [14]. Therefore, it is imperative to take effective measures to assist distressed HCWs in these countries in reducing their work-related stress so that they do not endure.

For the time being, several stress management techniques, such as regular physical activity, deep breathing, and mindfulness, have been recommended for occupational populations at the workplace [15-17]. Building new stress management habits, however, may be difficult for HCWs, especially those with distress symptoms. Based on the findings of a recent mixed methods study, while the majority of HCWs were aware of some of the stress effects and stress management methods, almost half of them failed to adopt stress management behaviors [18]. Individuals enduring distress symptoms such as anxiety and depression may experience loneliness, isolation, and stigma, which can lead to stress-coping failure and increased stress responses [19]. Besides, the long-term distress symptoms may diminish their self-efficacy and perceived social support and hinder them from taking advantage of stress management [20]. It appears that a "stress-distress" vicious circle exists, preventing HCWs from putting stress management knowledge into practice [21]. As a result, it is critical to assist distressed HCWs in closing the "knowing-doing" gap and benefiting from stress management.

Knowledge translation is defined as a process that occurs through social and environmental interactions and underlines that knowledge exchange must happen in an interacting social situation [22]. Among different theoretical frameworks that are applied in knowledge translation research, the most popular makes reference to the social learning theory (SLT) [23]. The SLT emphasizes 3 main elements associated with human behaviors—environmental factors, personal factors, and behavioral factors [24], and only when environmental and individual factors interact synergically toward maintainability can knowledge translate to the desired behavioral changes [25-27]. In light of the theoretical foundation outlined, creating an online intervention context may help HCWs reap the most benefit from stress management and alleviate stress symptoms [28].

The internet can be considered ideal for creating an online context in stress management intervention. First, by introducing social learning via the internet, stressed individuals may be able to change their attitudes toward the treatment, such as feeling alone, isolated, or stigmatized [29]. Second, online peer interaction can provide a supporting environment where participants may find and provide practical advice, as well as receive social connections that may be lacking in their daily lives [30]. Third, the use of a mobile health device would enable participants to self-monitor and receive feedback on their performance and tension reduction, which could be beneficial for enhancing their stress management skills [31]. Fourth, human support and guidance provided by coaching may potentially alter individuals' self-efficacy belief in stress management [32,33]. Besides, the high percentage of internet users suggests that it has potential applications in working populations—according to the most recent official report, China's internet penetration rate has reached 73%, and there are more than 1.32 billion netizens in China, 99.7% of whom access the internet via smartphones [34]. Therefore, mobile stress management intervention may be an adaptable and accessible approach for distressed HCWs.

In response to the aforementioned findings, a complex interactive multimodal intervention (CIMI) for distressed HCWs was developed. The CIMI consisted of four components: (1) mobile stress management education, (2) a web-based social network provided by a WeChat (Tencent Holdings Limited) group, (3) individualized practice feedback, and (4) personalized support and guidance provided by a nurse coach.

XSL•FO
RenderX

Objectives

Our primary objective was to figure out whether the CIMI is more effective than self-guided stress management in improving the stress symptoms of distressed HCWs, as measured by physiological (HRV indicators), psychological (perceived stress, mental distress, and subjective happiness), and sleep disorder (fatigue and sleepiness) indicators.

Methods

Study Design

The trial was a cluster-based, nonrandomized, parallel controlled study conducted in 2 hospitals in Wuhan, China, from August 11, 2021, to January 31, 2022. In order to prevent data contamination, we used a nonrandomized study design in this study. Hospital A participants received the CIMI intervention, while hospital B participants received the self-guided stress management intervention.

Recruitment and Participants

All participants were recruited between August 11, 2021, and October 31, 2021. HCWs were eligible to participate if they (1) worked the night shift in hospital A or hospital B; (2) did not plan to leave their current position within the next 6 months; (3) did not have any serious clinical diagnoses that affected stress levels or follow-up quality (ie, abnormal cardiac rhythm; heart disease including coronary artery disease, angina, and arrhythmia; cardiac pacemaker; stroke; panic attack; and cognitive impairment); and (4) had at least 1 sign of psychological distress measured by Depression, Anxiety, and Stress Scale (DASS-21; ie, the subscale of depression, anxiety, and stress was above 10 points, 7 points, and 11 points, respectively) [35]. HCWs are excluded if they (1) were unwilling to complete the 3-month follow-up or (2) were unable to do exercise because of fractures or limb abnormalities, and so forth.

Although we intended to recruit participants with a good sex and age balance, the COVID-19 pandemic caused numerous difficulties during the recruiting process. As a result, this study was unable to maintain a balanced age and sex distribution. In total, 661 participants completed the prescreening questionnaires, and 245 of them were eligible for enrollment based on the inclusion and exclusion criteria. Participants from hospital A (n=179) were assigned to the intervention group and participants from hospital B (n=66) were assigned to the control group. A total of 5 participants from the intervention group and 8 from the control group dropped out of the study during the 3-month study period due to personal reasons (Figure 1). According to Table 1, of the 245 participants, 208 (84.9%) were female, with a mean age of 33.7 (SD 0.3) years. In regard to educational background, 57 (23.3%) participants obtained graduate degrees. There were 85 (23.7%) doctors or other health care professionals and 160 (65.3%) nurses in terms of position. With respect to age, marital status, education, weekly working hours, and health status, the CIMI and control groups are comparable. In comparison to the control group, the CIMI group had a higher proportion of female participants (157/179, 87.7% vs 51/66, 77.3%; *P*=.04), those who worked for more than 10 years (98/179, 54.8% vs 23/66, 34.9%; *P*=.006), and nurses (122/179, 68.2% vs 38/66, 57.6%; *P*=.003).

Figure 1. Participant flowchart. CIMI: complex interactive multimodal intervention.

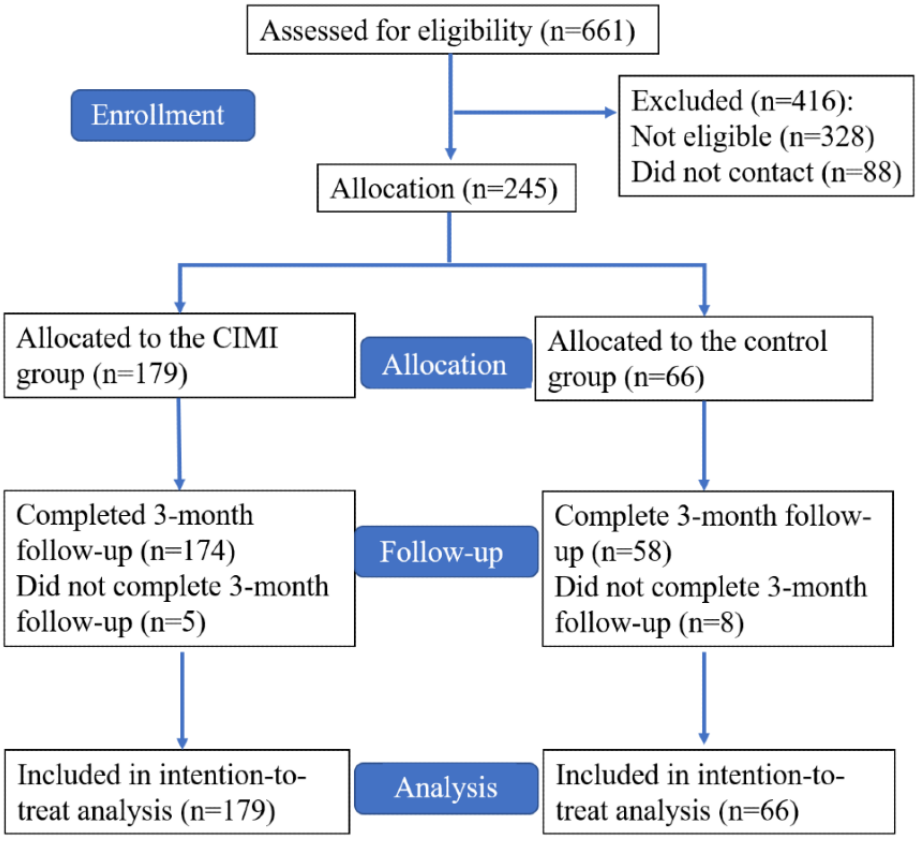

XSL•FO
RenderX

Table 1. Baseline characteristics of the participants.

Variables	Total (N=245)	Intervention (n=179)	Control (n=66)	*P* value
Age (years)[a], mean (SD)	33.7 (0.3)	33.6 (0.4)	33.9 (0.7)	.67
Sex[b], n (%)				.04
Female	208 (84.9)	157 (87.7)	51 (77.3)	
Male	37 (15.1)	22 (12.3)	15 (22.7)	
Marital status[b], n (%)				.79
Unmarried	70 (28.6)	52 (29.1)	18 (27.3)	
Married	175 (71.4)	127 (71)	48 (72.7)	
Educational level[b], n (%)				.054
≤Undergraduate	188 (76.7)	143 (79.9)	45 (68.2)	
≥Postgraduate	57 (23.3)	36 (20.1)	21 (31.8)	
Working experience (years)[b], n (%)				.006
<10	124 (50.6)	81 (45.3)	43 (65.2)	
≥10	121 (49.4)	98 (54.8)	23 (34.9)	
Role[b], n (%)				.003
Nurse	160 (65.3)	122 (68.2)	38 (57.6)	
Physician	58 (23.7)	33 (18.4)	25 (37.9)	
Other health care workers	27 (11)	24 (13.4)	3 (4.6)	
Weekly working hours[b] (hours), n (%)				.76
≤40	126 (51.4)	91 (50.8)	35 (53)	
>40	119 (48.6)	88 (49.2)	31 (47)	
Health status[b], n (%)				.47
Bad	54 (22)	38 (21.2)	16 (24.2)	
So-so	167 (68.2)	121 (67.6)	46 (69.7)	
Good	24 (9.8)	20 (11.2)	4 (6.1)	
Diagnosed disease[b], n (%)				.99
No	193 (78.8)	141 (78.8)	52 (78.8)	
Yes	52 (21.2)	38 (21.2)	14 (21.2)	

[a]Continuous variables.

[b]Categorical variables.

Power Calculation

The primary outcomes in our study were Perceived Stress Scale (PSS) scores and HRV measurements. As reported in a previous study [36], the PSS score decreased by 1.5 points after intervention in the control group. We hypothesized that the PSS score in the CIMI group would decrease by an additional 2 points compared to the control group. Therefore, at least 60 participants per group were required to detect a difference between the 2 groups with 85% power, accounting for a 20% dropout rate and a significance level of 0.025.

Similarly, based on HRV indicators, the SD of all normal-to-normal intervals (SDNN) of the HRV indicator decreased by 5 ms^2 in the control group after intervention [37]. We assumed that the CIMI group would experience a greater reduction in SDNN by 2 ms^2 compared to the control group. Therefore, a minimum of 200 participants per group were required to detect a difference between the 2 groups with 85% power, accounting for a 20% dropout rate and a significance level of 0.025. Therefore, we planned to recruit 200 participants in each group to ensure the detection of differences in both PSS and SDNN between the 2 groups.

However, the recruitment process was impeded due to the COVID-19 pressures, and we eventually enrolled 245 people who satisfied our inclusion criteria. Hence, the power calculation in this study was determined via a post hoc analysis. According to our study, PSS scores fell by 1.28 points in the control group and 5.31 points in the CIMI group between the baseline and 3

XSL•FO
RenderX

months after the intervention, respectively. Therefore, taking into account a 2-sided significance level of .05, the sample sizes of 179 in the intervention group and 66 in the control group would have more than 95% power to detect the significant difference between the 2 groups.

Procedure

First, a screening questionnaire about demographic characteristics (age, sex, marital status, educational level, working experience, role, and working hours), health status, diagnosed disease, and the DASS-21 was sent to the potential participants. After the screening, all eligible participants were required to complete a detailed baseline questionnaire and wear a 24-hour Holter device to assess the physiological indicators of stress as indexed by HRV. The day of HRV recording was selected as at least 48 hours after a night shift to ensure enough rest prior to recording. Furthermore, a sport wristband and corresponding use method were given to each participant, and participants were also asked to upload their exercise data from their wristbands to a research app on a daily basis. The research app only acted as a recording application that tracked daily information related to exercise, such as daily exercise time, consumed calories, and step counts, instead of providing any interaction between participants and researchers. In addition, a printed brochure was distributed to both the CIMI and control groups to teach them how to use 3 types of stress management techniques (ie, regular physical activity, deep breathing, and mindfulness; Multimedia Appendix 1).

During the 12-week intervention, the CIMI group was required to complete a weekly online questionnaire that contained 4 sections—daily life stressors, stress management practices, psychological symptoms, and sleep quantity and quality (Multimedia Appendix 2). The questionnaire was sent to each participant's mailbox, and the nurse coach was responsible for checking and ensuring the completion of the questionnaire, as well as reminding participants if they were unable to complete it within 3 days of receiving it. To keep contact between researchers and participants to a minimum, the control group only needed to complete the online questionnaire biweekly. All questionnaire data were collected using the REDCap (Research Electronic Data Capture; Vanderbilt University) platform.

Intervention

Participants in the CIMI group received a 12-week online intervention with the following 4 components: mobile stress management education, a web-based WeChat social network, tailored feedback, and a nurse coach.

Mobile Stress Management Education

Participants in the CIMI group were invited to join the WeChat group anonymously. A professional nurse coach delivered the evidence-based course about stress through a WeChat group every weekend for 12 weeks. In addition, 3 evidence-based stress management strategies (SMSs)—physical activity, deep breathing, and mindfulness—were introduced during every course, and to ensure participants practiced those SMSs appropriately, each participant received a package of instructional videos to guide them through the practices' techniques.

Web-Based WeChat Social Network

During each course through the WeChat group, a 30-minute online discussion session was scheduled. During the discussion, participants were invited to answer several questions about stress management and encouraged to share their photos or videos that recorded their SMS practice over the past week.

Tailored Feedback

The weekly and monthly summary reports on the stress symptoms and SMS performance were sent to the participants through email. The summary reports were developed by the research group based on the uploaded research app data from the wristbands and the weekly questionnaires completed by the participants.

Nurse Coach

The nurse coach would help participants whenever needed to (1) ensure their proper use of portable devices and continuous participation in the program, (2) evaluate their practice performance and identify possible barriers to practicing SMSs and further give specific instructions, and (3) support HCWs and encourage them to practice SMSs continuously during the intervention.

Control Group

The control group received the same wristband and research app, as well as instructions to ensure that they used the wristbands correctly. Participants in the control group were asked to fill out a brief online questionnaire every 2 weeks and upload their exercise data from the wristbands daily, but they would not receive the 4 components of intervention mentioned.

Outcomes

Primary Outcomes

Perceived Stress

The PSS-14 is a globally used 14-item scale that has been translated and validated in the Chinese population to measure the level of perceived stress [38]. The PSS-14 used a 5-point Likert scale, and each item was scored from 0 to 4. The total score ranges from 0 to 56, with higher total scores indicating a higher level of stress [39]. The Cronbach α was 0.835, demonstrating that the PSS-14 was a reliable instrument to measure the participants' perceived stress in this study.

HRV Metric

HRV is a common tool for assessing autonomic nervous system function, and it reflects the balance of the cardiovascular system controlled by the sympathetic and parasympathetic nerves of the autonomic nervous system [40]. The research and clinical device (DMS300-4, Diagnostic Monitoring Software) had an internal sampling rate of 4096 Hz, and the R-peaks were automatically detected with a precision of <1 ms. In this study, the SDNN, which is the most frequently used time domain measure of HRV, was used to assess physiological measures of stress [41,42]. The power spectrum of HRV was calculated by fast Fourier transform and expressed by ms^2 power spectrum density [41]. To avoid signal quality issues from biological (ectopic beats, arrhythmia, etc) or technological (movement,

XSL•FO
RenderX

signal interference, etc) sources, all electrocardiograph files were manually edited for detection by trained biomedical analysts. Values with differences greater than 20% were considered artifacts and the interpolation of degree zero was used for the artifact.

Secondary Outcomes

Mental Distress

The DASS-21 is a well-developed instrument used in the Chinese population for assessing the negative emotions of individuals in the past week [43]. The 21-item scale can be clustered into 3 subscales, and each subscale is rated on a 4-point Likert scale. The sum score of each subscale is multiplied by 2 and ranges from 0 to 42, with a higher score representing greater emotional health. The Cronbach α was 0.877, 0.872, and 0.805 for the depression subscale, anxiety subscale, and stress subscale, respectively, indicating excellent reliability for this scale in measuring emotional health in this study.

Subjective Happiness

The Subjective Happiness Scale (SHS) was used to assess the global subjective happiness of participants. The SHS consists of 4 items, and each item is rated on a 7-point Likert scale ranging from 1 to 7. Responses on this scale were summed and divided by 4, with higher scores indicating more happiness. The SHS has been used and validated in Chinese adults [44]. The Cronbach α was 0.683 for the SHS, demonstrating the reliability of this scale is acceptable in this study.

Fatigue Symptoms

The Fatigue Assessment Scale (FAS) is a widely used fatigue measure that includes 10 items rated on a 5-point Likert scale. This scale has been tested extensively on HCWs and demonstrated to have reliability and validity for measuring fatigue and to be free of gender bias [45]. The Cronbach α for the FAS was 0.883, indicating that the internal reliability of this scale was quite satisfactory in this study.

Sleepiness

The Epworth Sleepiness Scale (ESS) is a useful tool for evaluating daytime sleepiness in both the general population and shift workers [46]. This scale consists of 8 items, and each item is rated on a 4-point Likert scale. The Cronbach α for the ESS was 0.768, confirming that this scale is reliable for measure sleepiness in this study.

Statistical Analysis

The analysis was conducted according to the intention-to-treat principle. Continuous variables were described by mean (SD) or median (IQR), and categorical variables were represented by frequency (percentage). Baseline characteristics were compared between groups using the 2-sample *t* test for normally distributed continuous variables, nonparametric Wilcoxon rank sum tests for nonnormally distributed variables, and the chi-square test for categorical variables. A linear mixed-effects regression model with random intercepts was fitted for each primary and secondary outcome. A 2-way interaction between the groups (control vs intervention) and time (baseline vs postintervention) was modeled using restricted maximum likelihood with the Kenward-Roger denominator degrees of freedom. As proposed by the Journal of Biomedical Informatics (JBI) checklist for quasi-experimental studies (nonrandomized experimental studies), the disparities between participants in compared groups pose a risk to the internal validity of a study investigating causal linkages. Furthermore, as indicated by 2 publications published in *the Journal of Medical Internet Research*, the authors also emphasized the significance of managing confounding variables to guarantee the comparability of the intervention and control groups. Therefore, in our statistical analysis approach, we included the age, sex, marital status, education level, working experience, occupation, weekly working hours, health status, and diagnosed disease as covariates in each model, to guarantee that the impact size can be fully attributed to our intervention and not influenced by other factors.

A post hoc exploratory subgroup analysis was performed to detect the influence of age, sex, marital status, education level, working experience, occupation, and weekly working hours by adding the 3-way interaction including group, time point, and subgroup variable into the model, respectively. The adjusted mean difference (MD) with 95% CI from baseline to after the intervention for the CIMI group and the control group separately, and the between-group adjusted MD with 95% CI from baseline to after the intervention were obtained from the linear mixed-effects regression. We used Cohen *d* as the standardized effect size, which is considered small (0.2), medium (0.5), or large (0.8), and the Cohen *d* was calculated with the between-group adjusted MD divided by the SDs at baseline. The missing values of the outcomes were handled in the mixed-effects models. All analyses were performed by the Stata 15.1 software (StataCorp LLC) at a 2-sided α level of .05.

Ethical Considerations

The study was approved by the institutional review board of Tongji Medical College, Huazhong University of Science and Technology (IRB#2021S141). Before participating in the study, all participants provided informed consent in writing. All study data were deidentified to ensure participants' privacy and confidentiality. A complementary cardiovascular assessment for a close relative was given to research participants who completed the program, while no monetary rewards were offered.

Results

Primary Outcome

As shown in Table 2, the PSS scores reduced significantly in the CIMI group (MD –5.31, 95% CI –6.26 to –4.37; *P*<.001) after the intervention as compared to the baseline levels, while the PSS scores of the control group remained largely unaltered (MD –1.28, 95% CI –2.91 to 0.34; *P*=.12). The changes in PSS scores before and after the intervention exhibited a significant difference between the CIMI and control groups (d=–0.64; MD –4.03, 95% CI –5.91 to –2.14; *P*<.001), and the effect was medium.

In terms of physiological measures, both the control group (MD –9.56, 95% CI –16.9 to –2.2; *P*=.01) and the CIMI group (MD –8.45, 95% CI –12.68 to –4.22; *P*<.001) demonstrated a significant decrease in SDNN within the normal clinical range

XSL•FO
RenderX

(102-180 ms). In terms of the changes, there were no significant differences between the 2 groups (d=0.03; MD 1.11, 95% CI –7.38 to 9.59; *P*=.80).

Table 2. Estimations of adjusted mean differences and effect sizes for stress indicators (linear mixed model).

	Control group				Intervention group				Between group		
	Baseline, mean (SD)	Post intervention, mean (SD)	Change from baseline, adjusted MD[a] (95% CI)[b]	*P* value	Baseline, mean (SD)	Post intervention, mean (SD)	Change from baseline, adjusted MD (95% CI)[b]	*P* value	Change from baseline, adjusted MD (95% CI)[c]	*P* value	Cohen *d* (95% CI)
Physiological indicators											
SDNN[d,e]	156.7 (35.2)	145.7 (34.6)	–9.56 (–16.9 to –2.22)	.01	156.8 (32.5)	147 (32.6)	–8.45 (–12.68 to –4.22)	<.001	1.11 (–7.38 to 9.59)	.80	0.03 (–0.25 to 0.32)
Psychological indicators											
PSS[f]	30.7 (7.0)	29.2 (5.5)	–1.28 (–2.91 to 0.34)	.12	31.6 (6.3)	26.4 (6.3)	–5.31 (–6.26 to –4.37)	<.001	–4.03 (–5.91 to –2.14)	<.001	–0.64 (–0.93 to –0.36)
DASS-D[g]	11.4 (5.7)	10.3 (6.2)	–0.97 (–2.74 to 0.79)	.28	13.3 (8.0)	8.8 (6.8)	–4.5 (–5.52 to –3.47)	<.001	–3.53 (–5.57 to –1.48)	.001	–0.48 (–0.77 to –0.19)
DASS-A[h]	10.9 (5.6)	14.6 (6.6)	3.78 (1.97 to 5.59)	<.001	13.2 (6.3)	13.4 (6.8)	0.21 (–0.85 to 1.26)	.70	–3.57 (–5.67 to –1.47)	.001	–0.60 (–0.88 to –0.31)
DASS-S[i]	16.3 (6.4)	10.2 (6.3)	–6.09 (–8.0 to –4.17)	<.001	19.6 (6.8)	9.3 (5.9)	–10.33 (–11.45 to –9.22)	<.001	–4.25 (–6.46 to –2.03)	<.001	–0.64 (–0.93 to –0.36)
SHS[j]	3.7 (0.7)	4 (0.8)	0.26 (0.05 to 0.47)	.02	3.7 (0.9)	4.3 (1)	0.64 (0.52 to 0.76)	<.001	0.38 (0.14 to 0.63)	.002	0.47 (0.18 to 0.75)
Sleep indicators											
ESS[k]	7.6 (3.2)	7.9 (4.3)	0.5 (–0.43 to 1.42)	.29	8.7 (3.9)	7.8 (4)	–0.81(–1.35 to –0.28)	.003	–1.31 (–2.39 to –0.24)	.02	–0.35 (–0.64 to –0.07)
FAS[l]	29.3 (4.8)	27 (4.8)	–2.04 (–3.42 to –0.67)	.004	30.8 (5.6)	26.3 (6.3)	–4.52(–5.32 to –3.72)	<.001	–2.48(–4.07 to –0.89)	.002	–0.47 (–0.75 to –0.18)

[a]MD: mean difference.

[b]Change from baseline to after the intervention for intervention group and control group, respectively.

[c]Mean difference of change from baseline to after the intervention between intervention group and control group.

[d]SDNN: SD of normal-to-normal interval.

[e]The normal clinical range of SDNN was from 102 to 180 ms.

[f]PSS: Perceived Stress Scale.

[g]DASS-D: The Depression, Anxiety, and Stress-21 Scale's Depression subscale.

[h]DASS-A: The Depression, Anxiety, and Stress-21 Scale's Anxiety subscale.

[i]DASS-S: The Depression, Anxiety, and Stress-21 Scale's Stress subscale.

[j]SHS: Subjective Happiness Scale.

[k]ESS: Epworth Sleepiness Scale.

[l]FAS: Fatigue Assessment Scale.

Secondary Outcome

From baseline to after the intervention, the depression score of the CIMI group decreased by 3.53 points more than that of the control group (95% CI –5.57 to –1.48; *P*=.001; d=–0.48), and the stress score decreased by 4.25 points more than that of the control group (95% CI –6.46 to –2.03; *P*<.001; d=–0.64). Nevertheless, the anxiety score increased significantly in the control group (95% CI 1.97-5.59; *P*<.001), whereas that did not change significantly in the CIMI group (95% CI –0.85 to 1.26; *P*=.70). As compared with the control group, the CIMI group had statistically significant decreases in fatigue symptoms (MD –2.48, 95% CI –4.07 to –0.89; *P*=.002; d=–0.47) and daytime sleepiness (MD –1.31, 95% CI –2.39 to –0.24; *P*=.02; d=–0.35), in addition to a significant increase in subjective happiness (MD 0.38, 95% CI 0.14-0.63; *P*=.002; *d*=0.47), before and after the intervention (Table 2). Overall, the CIMI group had a medium effect on improving depression, stress, fatigue, and promoting subjective happiness, a small effect on improving sleepiness.

Subgroup Analysis

Subgroup analyses of the PSS score were conducted based on the age, sex, marital status, education levels, working experience, occupation, and weekly working hours of the

XSL•FO
RenderX

participants. As shown in Figure 2, the effect sizes between the CIMI group and the control group were statistically better in male than female individuals, better in doctors than nurses, and higher in participants with postgraduate degrees than those with bachelor's degrees (interaction *P* values for sex=.004, *P* values for occupation=.005, and *P* values for education=.005). There was no statistically significant difference in the effect sizes between the CIMI group and the control group by age, marital status, working experience, or weekly working hours subgroups (interaction *P* values were .92, .71, .22, and .09, respectively).

Figure 2. Subgroup analysis to detect the influence of demographic factors on treatment effects. CIMI: Complex Interactive Multimodal Intervention; MD: mean difference.

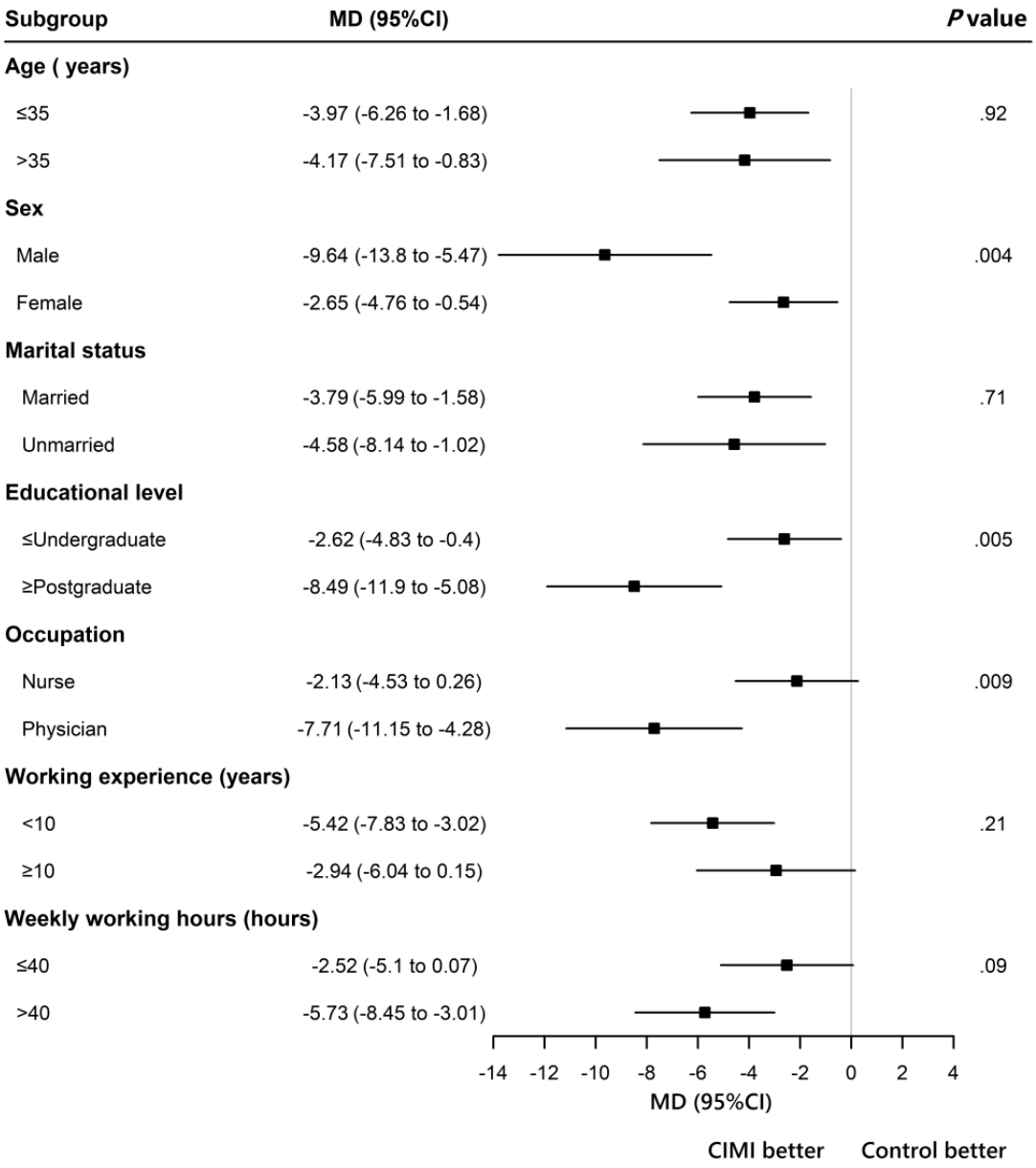

Discussion

Principal Findings

This study found that, with the exception of anxiety symptoms, most psychological markers significantly improved more in the CIMI group following a 12-week intervention than in the self-guided group. When it came to alleviating sleep disorders, the CIMI was more effective in reducing fatigue and sleepiness. Regarding physiological indications, HRV decreased in both groups but stayed within the range that is considered clinically normal. Overall, the CIMI was found to be effective for improving psychological indicators and sleep symptoms in HCWs, but not in physiological stress indicators.

To the best of our knowledge, despite the fact that numerous stress management interventions have been implemented in workplaces, none of them have been specifically designed for distressed HCWs. As previously stated, developing stress management behaviors would be difficult for HCWs with distress symptoms even if they were aware of the effects of stress and coping strategies. Hence, we developed a CIMI based

XSL•FO
RenderX

on knowledge translation theory [22] and SLT [24] to bridge the "knowing-doing" gap in stress management among distressed HCWs. The CIMI specifically created an online context with the goal of changing the HCWs' negative attitudes and skills through mobile stress management education, a web-based WeChat social network, tailored feedback, and a nurse coach. This study compared the effects of the CIMI and self-guided groups on stress symptoms using physiological, psychological, and sleep disorder indicators.

In terms of physiological indicators of stress, SDNN is a common HRV metric that represents the normal-to-normal SD. SDNN may drop significantly in disorders such as myocardial infarction, indicating collisions with HRV. With a normal SDNN range of 102 to 180 ms, SDNN swings within a specific range under physiological conditions [37,44,45]. Both the intervention and control groups showed isotropic changes in SDNN before and after intervention. There was no statistical difference between the 2 groups. This suggests that the CIMI is primarily concerned with improving subjective markers.

In aspects of psychological stress measures, the CIMI was significantly more effective in improving depression and stress symptoms and elevating the subjective happiness of HCWs. However, we found that the anxiety symptoms in the control group exaggerated significantly from baseline to after the intervention, while there were no significant changes in anxiety symptoms in the intervention group. That means the effect of the CIMI on alleviating anxiety was still unclear. Furthermore, we found that the CIMI helped improve daytime sleepiness and fatigue symptoms significantly. In a word, the CIMI was found to be significantly more effective than the control group in improving both psychological and sleep indicators of stress, as well as parts of psychological stress measures. Although previous research found that human support was more effective than self-guided mobile interventions, the majority of studies did not directly compare self-guided and human support e-therapy [47]. Therefore, this study provided objective evidence of the effect of the CIMI, a novel human support intervention, on alleviating stress symptoms among distressed HCWs.

To explore if our intervention effects varied between participants with different demographic characteristics, we further performed a subgroup analysis. We found that educational level, sex, and occupation may all influence the treatment effects of the CIMI. Participants with postgraduate degrees benefited from the intervention more than those with graduate degrees. This finding was consistent with previous research, which found that highly educated participants were more likely to engage in health-promoting behaviors and use interventions more effectively [48,49]. This could be explained by the fact that those with higher educational levels also tend to have higher socioeconomic status, allowing them to better protect themselves from the risk of poor mental health [50]. Thus, future research should consider how to assist participants with lower educational levels in making better use of interventions. Besides, we found that men benefited more from the interventions than women, which was consistent with prior studies. This may be explained by the fact that women are more likely to assume more family duties, thus they are less likely to have spare time to practice the recommended stress management techniques [51]. Hence, future research should consider how to help female HCWs get support from families so that they can spend enough time on stress management. Interestingly, we discovered that occupation influenced treatment effects, with physicians benefiting from interventions more than nurses. The findings were similar to those of Yildirim et al [52], who identified that physicians exercised more than nurses and had a higher quality of life. Given that nurses face more barriers to adopting health-promoting behaviors and have higher levels of work stress [53], future interventions may need to target this population in order to successfully manage their stress.

Overall, this study found that the CIMI was more effective than self-guided stress management in distressed HCWs. The design's uniqueness stems from the combination of multiple interventions and target populations. The integration of all elements into a computerized program allowing online crosstalk with participants could be a future CIMI modification [54,55]. It would save labor while also increasing the acceptance of interventions among distressed occupational populations.

Limitations

This study has several limitations that should be mentioned. First, the controlled, nonrandomized design may bias estimates of treatment effects and cause uncertainty in our findings. As a result, future studies can consider using randomized controlled trials to investigate our findings. Second, because of the COVID-19 pandemic, we failed to balance sex and age during our recruitment process. To avoid any bias caused by demographic characteristics, the Cohen *d* effect was calculated by dividing the between-group adjusted MD by the SDs at baseline in this study. Third, since our study only tested the effectiveness of the CIMI over a 3-month period, it was unclear how long its effects would last. As a result, more research into the long-term effects of human assistance intervention is needed. Last but not least, we recognized that we might not have been able to detect the significant effect of the CIMI on improving HRV measures since we had not been sure that our sample size complied with the requirements for HRV indicators. Therefore, we recommended that future studies examine the influence of CIMI on HRV markers using a larger sample size.

Conclusions

In this study, we discovered that the CIMI was effective at alleviating stress symptoms in distressed HCWs. This study's findings may provide objective evidence for the development of an effective mobile stress management intervention for distressed HCWs. Given that this study was a nonrandomized controlled trial that lasted 3 months among distressed HCWs, future research may consider investigating the long-term effects of the CIMI using a randomized controlled trial design.

XSL•FO
RenderX

Acknowledgments

This work was supported by the National Key R&D Program of China (2021YFC2502200). We are grateful to all health care workers who participated in this study. We sincerely acknowledge Jean-Paul Collect, who helped design the intervention. We appreciate the smart devices and compatible research application provided by Huawei Research Team. We also thank the important members of the Huawei Research Team, Runsen Wang, Yujia Han, and Rong Sheng, for their contributions to the processing of the smart device data.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

WW, FD, WL, QW, and DZ oversaw the study's design and paper preparation. WW, FD, JM, JL, and GC helped develop the study protocol and managed the study. FD, WL, QW, and DZ oversaw data science integration. All authors read and approved the final paper. Authors FD (Fengfei_ding@fudan.edu.cn) and WW (Wwang@vip.126.com) are co-corresponding authors for this article

Conflicts of Interest

None declared.

Multimedia Appendix 1
Printed booklet.
[PDF File (Adobe PDF File), 509 KB - jmir_v26i1e45422_app1.pdf]

Multimedia Appendix 2
Weekly questionnaire.
[DOCX File , 144 KB - jmir_v26i1e45422_app2.docx]

References

1. Li Y, Lv X, Li R, Wang Y, Guan X, Li L, et al. Predictors of shift work sleep disorder among nurses during the COVID-19 pandemic: a multicenter cross-sectional study. Front Public Health 2021;9:785518. [doi: 10.3389/fpubh.2021.785518]
2. Brown JP, Martin D, Nagaria Z, Verceles AC, Jobe SL, Wickwire EM. Mental health consequences of shift work: an updated review. Curr Psychiatry Rep 2020;22(2):7. [doi: 10.1007/s11920-020-1131-z] [Medline: 31955278]
3. Min A, Min H, Hong HC. Work schedule characteristics and fatigue among rotating shift nurses in hospital setting: an integrative review. J Nurs Manag 2019;27(5):884-895. [doi: 10.1111/jonm.12756] [Medline: 30737987]
4. Sara JD, Prasad M, Eleid MF, Zhang M, Widmer RJ, Lerman A. Association between work-related stress and coronary heart disease: a review of prospective studies through the job strain, effort-reward balance, and organizational justice models. J Am Heart Assoc 2018;7(9):e008073 [FREE Full text] [doi: 10.1161/JAHA.117.008073] [Medline: 29703810]
5. Kang M. Occupational risk factors for hypertension. J Hypertens 2022;40(11):2102-2110. [doi: 10.1097/HJH.0000000000003238] [Medline: 35950984]
6. Moretti Anfossi C, Ahumada Muñoz M, Tobar Fredes C, Pérez Rojas F, Ross J, Head J, et al. Work exposures and development of cardiovascular diseases: a systematic review. Ann Work Expo Health 2022;66(6):698-713 [FREE Full text] [doi: 10.1093/annweh/wxac004] [Medline: 35237787]
7. Jarczok MN, Koenig J, Wittling A, Fischer JE, Thayer JF. First evaluation of an index of low vagally-mediated heart rate variability as a marker of health risks in human adults: proof of concept. J Clin Med 2019;8(11):1940 [FREE Full text] [doi: 10.3390/jcm8111940] [Medline: 31717972]
8. van der Molen HF, Nieuwenhuijsen K, Frings-Dresen MHW, de Groene G. Work-related psychosocial risk factors for stress-related mental disorders: an updated systematic review and meta-analysis. BMJ Open 2020;10(7):e034849 [FREE Full text] [doi: 10.1136/bmjopen-2019-034849] [Medline: 32624469]
9. Salari N, Khazaie H, Hosseinian-Far A, Khaledi-Paveh B, Kazeminia M, Mohammadi M, et al. The prevalence of stress, anxiety and depression within front-line healthcare workers caring for COVID-19 patients: a systematic review and meta-regression. Hum Resour Health 2020;18(1):100 [FREE Full text] [doi: 10.1186/s12960-020-00544-1] [Medline: 33334335]
10. Lang X, Wang Q, Huang S, Feng D, Ding F, Wang W. Relations among perceived stress, fatigue, and sleepiness, and their effects on the ambulatory arterial stiffness index in medical staff: a cross-sectional study. Front Psychol 2022;13:1010647 [FREE Full text] [doi: 10.3389/fpsyg.2022.1010647] [Medline: 36389608]

XSL•FO
RenderX

11. Gray P, Senabe S, Naicker N, Kgalamono S, Yassi A, Spiegel JM. Workplace-based organizational interventions promoting mental health and happiness among healthcare workers: a realist review. Int J Environ Res Public Health 2019;16(22):4396 [FREE Full text] [doi: 10.3390/ijerph16224396] [Medline: 31717906]
12. Velana M, Rinkenauer G. Individual-level interventions for decreasing job-related stress and enhancing coping strategies among nurses: a systematic review. Front Psychol 2021;12:708696 [FREE Full text] [doi: 10.3389/fpsyg.2021.708696] [Medline: 34349711]
13. Fang P, Luo Z, Fang Z. What is the job satisfaction and active participation of medical staff in public hospital reform: a study in Hubei province of China. Hum Resour Health 2015;13:34 [FREE Full text] [doi: 10.1186/s12960-015-0026-2] [Medline: 25975721]
14. Martínez P, Rojas G, Martínez V, Lara MA, Pérez JC. Internet-based interventions for the prevention and treatment of depression in people living in developing countries: a systematic review. J Affect Disord 2018;234:193-200. [doi: 10.1016/j.jad.2018.02.079] [Medline: 29529553]
15. Lomas T, Medina JC, Ivtzan I, Rupprecht S, Eiroa-Orosa FJ. Mindfulness-based interventions in the workplace: an inclusive systematic review and meta-analysis of their impact upon wellbeing. J Posit Psychol 2018;14(5):625-640. [doi: 10.1080/17439760.2018.1519588]
16. Bischoff LL, Otto A, Hold C, Wollesen B. The effect of physical activity interventions on occupational stress for health personnel: a systematic review. Int J Nurs Stud 2019;97:94-104. [doi: 10.1016/j.ijnurstu.2019.06.002] [Medline: 31234106]
17. Tavoian D, Craighead DH. Deep breathing exercise at work: potential applications and impact. Front Physiol 2023;14:1040091 [FREE Full text] [doi: 10.3389/fphys.2023.1040091] [Medline: 36711016]
18. Wang Q, Collet J, Liu W, Lang X, Feng D, Ding F, et al. Work-related stress and stress management practice among Chinese hospital staff working night shifts: a mixed-methods study. Int J Stress Manag 2023;30(3):274-286. [doi: 10.1037/str0000281]
19. Yokota R, Okuhara T, Okada H, Goto E, Sakakibara K, Kiuchi T. Associations between stigma, cognitive appraisals, coping strategies and stress responses among Japanese women undergoing infertility treatment. Healthcare (Basel) 2022;10(10):1907 [FREE Full text] [doi: 10.3390/healthcare10101907] [Medline: 36292354]
20. Liao C, Guo L, Zhang C, Zhang M, Jiang W, Zhong Y, et al. Emergency stress management among nurses: a lesson from the COVID‐19 outbreak in China–a cross‐sectional study. J Clin Nurs 2020;30(3-4):433-442. [doi: 10.1111/jocn.15553] [Medline: 33141483]
21. Chaudhry BM, Islam A, Matthieu M. Toward designs of workplace stress management mobile apps for frontline health workers during the COVID-19 pandemic and beyond: mixed methods qualitative study. JMIR Form Res 2022;6(1):e30640 [FREE Full text] [doi: 10.2196/30640] [Medline: 34806985]
22. Esmail R, Hanson HM, Holroyd-Leduc J, Brown S, Strifler L, Straus SE, et al. A scoping review of full-spectrum knowledge translation theories, models, and frameworks. Implement Sci 2020;15(1):11 [FREE Full text] [doi: 10.1186/s13012-020-0964-5] [Medline: 32059738]
23. Thomas A, Menon A, Boruff J, Rodriguez AM, Ahmed S. Applications of social constructivist learning theories in knowledge translation for healthcare professionals: a scoping review. Implement Sci 2014;9:54 [FREE Full text] [doi: 10.1186/1748-5908-9-54] [Medline: 24885925]
24. Bandura A. In: Ross P, editor. Social Learning Theory. Hoboken, New Jersey: Prentice-Hall; 1976.
25. Beauchamp MR, Crawford KL, Jackson B. Social cognitive theory and physical activity: mechanisms of behavior change, critique, and legacy. Psychol Sport Exerc 2019;42:110-117. [doi: 10.1016/j.psychsport.2018.11.009] [Medline: 30872122]
26. Melnyk BM, Kelly SA, Stephens J, Dhakal K, McGovern C, Tucker S, et al. Interventions to improve mental health, well-being, physical health, and lifestyle behaviors in physicians and nurses: a systematic review. Am J Health Promot 2020;34(8):929-941 [FREE Full text] [doi: 10.1177/0890117120920451] [Medline: 32338522]
27. Salter KL, Kothari A. Knowledge 'Translation' as social learning: negotiating the uptake of research-based knowledge in practice. BMC Med Educ 2016;16:76 [FREE Full text] [doi: 10.1186/s12909-016-0585-5] [Medline: 26925578]
28. Cook NF, McAloon T, O'Neill P, Beggs R. Impact of a web based interactive simulation game (PULSE) on nursing students' experience and performance in life support training--a pilot study. Nurse Educ Today 2012;32(6):714-720. [doi: 10.1016/j.nedt.2011.09.013] [Medline: 22082881]
29. Alhasani M, Mulchandani D, Oyebode O, Baghaei N, Orji R. A systematic and comparative review of behavior change strategies in stress management apps: opportunities for improvement. Front Public Health 2022;10:777567 [FREE Full text] [doi: 10.3389/fpubh.2022.777567] [Medline: 35284368]
30. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
31. Heber E, Ebert DD, Lehr D, Cuijpers P, Berking M, Nobis S, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017;19(2):e32 [FREE Full text] [doi: 10.2196/jmir.5774] [Medline: 28213341]
32. Ebner K, Schulte E, Soucek R, Kauffeld S. Coaching as stress-management intervention: the mediating role of self-efficacy in a framework of self-management and coping. Int J Stress Manag 2018;25(3):209-233. [doi: 10.1037/str0000058]

XSL•FO
RenderX

33. Chaudhry B, Islam A. Design validation of a workplace stress management mobile app for healthcare workers during Covid-19 and beyond. United States: Springer International Publishing; 2022 Presented at: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; 2021; Yamaguchi URL: https://link.springer.com/chapter/10.1007/978-3-030-94822-1_17
34. The 40 internet network development statistic report of China. CNNIC. 2022. URL: http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/202202/t20220225_71727.htm [accessed 2024-06-06]
35. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol 2005;44(Pt 2):227-239. [doi: 10.1348/014466505X29657] [Medline: 16004657]
36. Balint EM, Angerer P, Guendel H, Marten-Mittag B, Jarczok MN. Stress management intervention for leaders increases nighttime SDANN: results from a randomized controlled trial. Int J Environ Res Public Health 2022;19(7):3841 [FREE Full text] [doi: 10.3390/ijerph19073841] [Medline: 35409525]
37. Hwang WJ, Jo HH. Evaluation of the effectiveness of mobile app-based stress-management program: a randomized controlled trial. Int J Environ Res Public Health 2019;16(21):4270 [FREE Full text] [doi: 10.3390/ijerph16214270] [Medline: 31684169]
38. Huang F, Wang H, Wang Z, Zhang J, Du W, Su C, et al. Psychometric properties of the perceived stress scale in a community sample of Chinese. BMC Psychiatry 2020;20(1):130 [FREE Full text] [doi: 10.1186/s12888-020-02520-4] [Medline: 32197589]
39. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. JHSB 1983;24(4):385. [doi: 10.2307/2136404]
40. Järvelin-Pasanen S, Sinikallio S, Tarvainen MP. Heart rate variability and occupational stress-systematic review. Ind Health 2018;56(6):500-511 [FREE Full text] [doi: 10.2486/indhealth.2017-0190] [Medline: 29910218]
41. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258 [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
42. Peabody J, Ryznar R, Ziesmann M, Gillman L. A systematic review of heart rate variability as a measure of stress in medical professionals. Cureus 2023;15(1):e34345 [FREE Full text] [doi: 10.7759/cureus.34345] [Medline: 36865953]
43. Jiang L, Yan Y, Jin Z, Hu M, Wang L, Song Y, et al. The depression Anxiety Stress Scale-21 in Chinese hospital workers: reliability, latent structure, and measurement invariance across genders. Front Psychol 2020;11:247 [FREE Full text] [doi: 10.3389/fpsyg.2020.00247] [Medline: 32210869]
44. Guo N, Luk TT, Ho SY, Lee JJ, Shen C, Oliffe J, et al. Problematic smartphone use and mental health in Chinese adults: a population-based study. Int J Environ Res Public Health 2020;17(3):844 [FREE Full text] [doi: 10.3390/ijerph17030844] [Medline: 32013111]
45. Barker LM, Nussbaum MA. Fatigue, performance and the work environment: a survey of registered nurses. J Adv Nurs 2011;67(6):1370-1382. [doi: 10.1111/j.1365-2648.2010.05597.x] [Medline: 21352271]
46. Kendzerska TB, Smith PM, Brignardello-Petersen R, Leung RS, Tomlinson GA. Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review. Sleep Med Rev 2014;18(4):321-331. [doi: 10.1016/j.smrv.2013.08.002] [Medline: 24135493]
47. Koelen J, Vonk A, Klein A, de Koning L, Vonk P, de Vet S, et al. Man vs. machine: a meta-analysis on the added value of human support in text-based internet treatments ("e-therapy") for mental disorders. Clin Psychol Rev 2022;96:102179 [FREE Full text] [doi: 10.1016/j.cpr.2022.102179] [Medline: 35763975]
48. Yu Y, Wang X, Wu Y, Weng W, Zhang M, Li J, et al. The benefits of psychosocial interventions for mental health in men who have sex with men living with HIV: a systematic review and meta-analysis. BMC Psychiatry 2022;22(1):440 [FREE Full text] [doi: 10.1186/s12888-022-04072-1] [Medline: 35768860]
49. Mudd AL, van Lenthe FJ, Verra SE, Bal M, Kamphuis CBM. Socioeconomic inequalities in health behaviors: exploring mediation pathways through material conditions and time orientation. Int J Equity Health 2021;20(1):184 [FREE Full text] [doi: 10.1186/s12939-021-01522-2] [Medline: 34391423]
50. Kurspahić Mujčić A, Mujčić A. The relationship between education and self-reported mental and physical health. Med Glas (Zenica) 2019;16(1):102-107 [FREE Full text] [doi: 10.17392/969-19] [Medline: 30589242]
51. Núñez-Rocha GM, López-Botello CK, Salinas-Martínez AM, Arroyo-Acevedo HV, Martínez-Villarreal RT, Ávila-Ortiz MN. Lifestyle, quality of life, and health promotion needs in Mexican University students: important differences by sex and academic discipline. Int J Environ Res Public Health 2020;17(21):8024 [FREE Full text] [doi: 10.3390/ijerph17218024] [Medline: 33142666]
52. Yildirim D, Yildirim A, Eryilmaz M. Relationship between physical activity and quality of life in health workers. Cukurova Med J 2019 Feb 21;44(2):325-333. [doi: 10.17826/cumj.451087]
53. Ross A, Bevans M, Brooks AT, Gibbons S, Wallen GR. Nurses and health-promoting behaviors: knowledge may not translate into self-care. AORN J 2017;105(3):267-275 [FREE Full text] [doi: 10.1016/j.aorn.2016.12.018] [Medline: 28241948]
54. Chaudhry BM. Sleeping with an Android. Mhealth 2017;3:7 [FREE Full text] [doi: 10.21037/mhealth.2017.02.04] [Medline: 28293622]
55. Chaudhry BM. Heart rate monitoring mobile applications. Mhealth 2016;2:17 [FREE Full text] [doi: 10.21037/mhealth.2016.04.01] [Medline: 28293594]

XSL•FO
RenderX

Abbreviations

CIMI: complex interactive multimodal intervention
DASS-21: Depression, Anxiety, and Stress Scale
ESS: Epworth Sleepiness Scale
FAS: Fatigue Assessment Scale
HCW: health care worker
HRV: heart rate variability
JBI: Journal of Biomedical Informatics
MD: mean difference
PSS: Perceived Stress Scale
REDCap: Research Electronic Data Capture
SDNN: SD of all normal-to-normal intervals
SHS: Subjective Happiness Scale
SLT: social learning theory
SMS: stress management strategy

Edited by T de Azevedo Cardoso; submitted 30.12.22; peer-reviewed by E Smoktunowicz, B Chaudhry; comments to author 23.02.23; revised version received 19.04.23; accepted 03.05.24; published 12.07.24.

Please cite as:
Liu W, Wang Q, Zheng D, Mei J, Lu J, Chen G, Wang W, Ding F
The Effects of a Complex Interactive Multimodal Intervention on Personalized Stress Management Among Health Care Workers in China: Nonrandomized Controlled Study
J Med Internet Res 2024;26:e45422
URL: https://www.jmir.org/2024/1/e45422
doi:10.2196/45422
PMID:

©Wenhua Liu, Quan Wang, Danli Zheng, Junhua Mei, Jiajia Lu, Guohua Chen, Wei Wang, Fengfei Ding. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Acceptability of the Drink Less App and the National Health Service Alcohol Advice Web Page: Qualitative Interview Process Evaluation

Melissa Oldham[1], PhD; Larisa-Maria Dina[1], MSc; Gemma Loebenberg[1], MSc; Olga Perski[2], PhD; Jamie Brown[1], PhD; Colin Angus[3], MSc; Emma Beard[1], PhD; Robyn Burton[4,5], PhD; Matt Field[6], PhD; Felix Greaves[7,8], PhD; Matthew Hickman[9], PhD; Eileen Kaner[10], PhD; Susan Michie[11], PhD; Marcus R Munafò[12,13], PhD; Elena Pizzo[14], PhD; Claire Garnett[1,12], PhD

[1]Behavioural Science and Health, University College London, London, United Kingdom
[2]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, San Diego, CA, United States
[3]School of Health and Related Research, University of Sheffield, Sheffield, United Kingdom
[4]Addictions Directorate, Office for Health Improvement and Disparities, London, United Kingdom
[5]Institute of Psychiatry, Psychology and Neuroscience, Kings College, London, United Kingdom
[6]Department of Psychology, University of Sheffield, Sheffield, United Kingdom
[7]Department of Primary Care and Public Health, Imperial College London, London, United Kingdom
[8]National Institute for Health and Care Excellence, London, United Kingdom
[9]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[10]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom
[11]Centre for Behaviour Change, University College London, London, United Kingdom
[12]School of Psychological Science, University of Bristol, Bristol, United Kingdom
[13]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[14]Department of Applied Health Research, University College London, London, United Kingdom

Corresponding Author:
Melissa Oldham, PhD
Behavioural Science and Health
University College London
1-19 Torrington Place
London, WC1E 7HB
United Kingdom
Phone: 44 20 7679 883
Email: m.oldham@ucl.ac.uk

Abstract

Background: The extent to which interventions are perceived as acceptable to users impacts engagement and efficacy.

Objective: In this study, we evaluated the acceptability of (1) the smartphone app Drink Less (intervention) and (2) the National Health Service (NHS) alcohol advice web page (usual digital care and comparator) among adult drinkers in the United Kingdom participating in a randomized controlled trial evaluating the effectiveness of the Drink Less app.

Methods: A subsample of 26 increasing- and higher-risk drinkers (Alcohol Use Disorders Identification Test score≥8) assigned to the intervention group (Drink Less; n=14, 54%; female: n=10, 71%; age: 22-72 years; White: n=9, 64%) or usual digital care group (NHS alcohol advice web page; n=12, 46%; female: n=5, 42%; age: 23-68 years: White: n=9, 75%) took part in semistructured interviews. The interview questions were mapped on to the 7 facets of acceptability according to the Theoretical Framework of Acceptability: *affective attitude*, *burden*, *perceived effectiveness*, *ethicality*, *intervention coherence*, *opportunity costs*, and *self-efficacy*. Alongside these constructs, we also included a question on *perceived personal relevance*, which previous research has linked to acceptability and engagement. Framework and thematic analysis of data was undertaken.

Results: The Drink Less app was perceived as being *ethical*, *easy*, *user-friendly*, and *effective* for the period the app was used. Participants reported particularly liking the tracking and feedback sections of the app, which they reported increased *personal*

XSL•FO
RenderX

relevance and which resulted in positive *affect* when achieving their goals. They reported no *opportunity cost*. Factors such as negative *affect* when not meeting goals and boredom led to disengagement in the longer term for some participants. The NHS alcohol advice web page was rated as being *easy* and *user-friendly* with no *opportunity costs*. However, the information presented was not perceived as being *personally relevant* or *effective* in changing drinking behavior. Most participants reported neutral or negative *affect*, most participants thought the alcohol advice web page was accessible, and some participants reported *ethical concerns* around the availability of suggested resources. Some participants reported that it had acted as a starting point or a signpost to other resources. Participants in both groups discussed motivation to change and contextual factors such as COVID-19 lockdowns, which influenced their perceived *self-efficacy* regardless of their assigned intervention.

Conclusions: Drink Less appears to be an acceptable digital intervention among the recruited sample. The NHS alcohol advice web page was generally considered unacceptable as a stand-alone intervention among the recruited sample, although it may signpost and help people access other resources and interventions.

(J Med Internet Res 2024;26:e42319) doi:10.2196/42319

KEYWORDS

alcohol reduction; digital intervention; acceptability; mobile health; mHealth; mobile phone

Introduction

Background

Drinking alcohol at increasing- and higher-risk levels is a major public health concern and contributes to health inequalities, with the most deprived groups experiencing the most harm from alcohol [1]. Fewer than 7% of increasing- and higher-risk drinkers who visited their general practitioner in the last year received face-to-face interventions in primary care to support alcohol reduction [2]. Key barriers to the delivery of these interventions by practitioners are lack of time, low confidence about discussing alcohol with patients, and lack of training [3-5]. Digital interventions, such as websites, are effective for reducing alcohol consumption compared with no intervention or minimal input controls [6]. They may overcome delivery barriers, as they potentially have a broad reach and relatively low implementation costs (once developed), so they can be delivered at scale [7]. Smartphone apps are a promising type of digital intervention, as smartphones have become increasingly affordable to end users and prevalent among the UK population [8]. However, despite the availability of hundreds of alcohol-related apps on commercial app stores, the majority have been developed without reference to scientific evidence or theory [9]. Furthermore, few have undergone evaluation in terms of their acceptability, engagement, or effectiveness [6]. It is critical to establish whether digital interventions are acceptable to end users, as acceptability impacts engagement and effectiveness [10]. Furthermore, demonstrating the acceptability of interventions can encourage public health practitioners and policy makers to promote interventions to those who would benefit from using them.

The extent to which interventions are perceived as acceptable to users and to other stakeholders such as family members, health care professionals, and policy makers affects engagement and effectiveness [10]. Acceptability sits at the core of the Technology Acceptance Model [11], which states that the perceived ease of use and perceived usefulness of a given technology positively influence use intentions. Most definitions of acceptability in digital health research primarily capture how people think and feel about a given technology [12,13], an example being “an emergent property, or a ‘gut feeling’, arising from a dynamic, complex system of emotional and cognitive components” [14]. The Theoretical Framework of Acceptability (TFA) defines acceptability as a multifaceted construct reflecting the extent to which a health care intervention is considered appropriate based on anticipated or experienced emotional and cognitive responses to the intervention [12]. Acceptability, according to the TFA, consists of 7 facets: affective attitude, burden, perceived effectiveness, ethicality, intervention coherence, opportunity costs, and self-efficacy. In addition to the facets outlined by the TFA, the extent to which an intervention is perceived as being personally relevant and tailored to the individual could be of importance when thinking about acceptability [15] and has shown to be linked to engagement [16].

This Study

Drink Less is a theory- and evidence-informed app-based intervention designed by researchers [17] to help people reduce their alcohol consumption. This study examined the acceptability of Drink Less and the National Health Service (NHS) alcohol advice web page [18] to adults drinking at increasing- and higher-risk levels in the United Kingdom following their participation in a randomized controlled trial evaluating the effectiveness of the Drink Less app [19] compared with usual digital care (eg, NHS alcohol advice web page [18]). It aimed to assess participants’ views on the acceptability of the Drink Less smartphone app and of the NHS alcohol advice web page.

Methods

Ethical Considerations

Ethics approval was obtained from University College London (UCL) Research Ethics Committee (16799/001). All participants provided informed consent before participating in the study. Data were anonymized and securely stored. All study participants were renumerated with a GBP £20 (US $25.46) Amazon (Amazon Inc) voucher to thank them for their time.

The Drink Less App

The development of Drink Less was informed by research findings and behavioral theories such as the Capability Opportunity Motivation–Behavior model of behavior change

XSL•FO
RenderX

[17]. Drink Less consists of evidence-based modules to help users change their drinking behavior: Goal Setting (setting weekly drinking reduction goals), Self-Monitoring and Feedback (monitoring alcohol consumption and seeing progress on goals), Action Planning (creating plans for dealing with difficult drinking situations), Normative Feedback (providing personalized feedback on how an individual's drinking behavior compares with the norm), Cognitive Bias Re-Training (a game for retraining users' automatic biases for alcoholic drinks), Insights (providing users with weekly and monthly feedback on alcohol consumed), Behavioural Substitution (planning to substitute drinking with a neutral behavior), and Information about Antecedents (providing users with information about situations, events, emotions, and cognitions that predict their drinking [20]).

NHS Alcohol Advice Web Page

The web page is freely accessible and appears in the top Google searches for "alcohol reduction advice" and "how to drink less alcohol" in the United Kingdom. The web page contains tips for cutting down on alcohol consumption, such as planning, setting a budget, and switching to smaller or weaker strength drinks. This is presented alongside benefits of cutting down for physical and mental health, including weight loss and improvements in mood and sleep [18]. The web page also has links to other web pages, including "alcohol support" and "the risks of drinking too much."

Study Data

This study analyzed data collected within the iDEAS (iOS Drink Less, evaluating the Effectiveness of an Alcohol Smartphone app) trial, a large-scale randomized controlled trial [19] evaluating the effectiveness and cost-effectiveness of recommending the Drink Less app compared with usual digital care (the NHS alcohol advice web page) in the United Kingdom using an embedded mixed methods process evaluation. This paper reports the analysis of the qualitative interviews assessing the acceptability of the interventions.

Participants were eligible for the iDEAS trial if they were aged ≥18 years, lived in the United Kingdom, were increasing- and higher-risk drinkers (Alcohol Use Disorders Identification Test score≥8), had access to an iOS (Apple Inc) device (ie, iPhone, iPod touch, or iPad), and wanted to drink less alcohol. Recruitment ran from July 2020 to March 2022, with the final follow-up conducted in October 2022. Recruitment occurred via a multipronged strategy, including an advertisement on the NHS website, targeted and untargeted social media advertising, radio advertising, a mail-out to a database of UK-based users of the Smoke Free app, and local advertising through health care providers.

When participants signed up for the iDEAS trial, they provided informed consent to participate in 3 web-based follow-up surveys after 1, 3, and 6 months. They were given the option to also consent to be contacted for a follow-up interview. Participants were then asked again at their 6-month follow-up whether they would be happy to be contacted for a follow-up interview about their experience of using the intervention.

Participants

Participants were selected from the group who consented at either baseline or 6-month follow-up to an interview. We identified participants purposively to interview roughly equal groups of men and women, those on low and high incomes, and from a range of ages and ethnic backgrounds. This was to ensure that the views of a diverse group of participants were represented. Researchers also purposively sampled to include people with a range of app engagement levels to avoid recruiting only highly engaged participants who may have felt more positively about the intervention. For those in the Drink Less group, engagement data were used to determine whether they had low (defined as 1-2 recorded sessions), medium (3-27 sessions), or high (≥28 sessions) engagement with the app. Participants were asked whether they used the intervention in their 1- and 6-month follow-up surveys, and these data were used to ensure that different levels of engagement within the comparator group (eg, never used vs used) were captured. When sampling, researchers filtered the data based on demographic and engagement criteria, and the first person meeting the required criteria was invited. To include the views of those who may have been less engaged in the study, those who provided consent to participate in the interview at baseline but did not complete the 6-month follow-up were invited to participate.

The final sample was 26 participants (Drink Less group: n=14, 54%; NHS alcohol advice web page [comparator] group: n=12, 46%). Recruitment ceased when meaning saturation was reached, and no further nuances or insights were found [21]. No new codes were identified in later interviews that differed from those identified in the earlier interviews. Nor did the later interviews change the meaning of any codes or themes.

Epistemological Position

We adopted a realist epistemology, assuming that our interpretation of each participant's reality is shaped through their perception of that reality. The realist epistemology assumes that meaning and experience are reflected in language [22]. This epistemology fitted our research aim of exploring acceptability judgments for the 2 interventions.

Procedure

In line with previous conceptualizations of acceptability [14], the interview topic guide (Multimedia Appendix 1) was designed to first measure an individual's gut feeling about the intervention they received (using a 5-star rating system) before exploring the 7 component facets according to the TFA [12], namely affective attitude, burden, perceived effectiveness, ethicality, intervention coherence, opportunity costs, and self-efficacy and perceived personal relevance [16]. Given that trial architecture (eg, randomization and follow-up) can be conflated with perceptions of the acceptability of the studied intervention [23], we asked participants to reflect on each separately.

We consulted a public and patient involvement group to ensure that the questions were clear for participants to answer. Public and patient involvement feedback was that the ethicality question "How fair did you think the intervention was to all possible users?" would be difficult to answer, and the suggested alternative wording was "Do you think anyone could use this

XSL•FO
RenderX

intervention?" Participants also suggested that the burden question that originally read, "How difficult did you find it to use the intervention?" be split into 2 questions focusing on how time consuming the app was and whether there were any other difficulties. Both suggested changes were implemented.

Once selected, participants were emailed and invited to attend a one-to-one semistructured interview. Interviews were conducted within 2 months of participants finishing their final follow-up survey. A total of 24 interviews took place using videoconferencing software, and 2 interviews took place on the phone (participants' preference). At the start of the interview, the researcher confirmed consent. The interview focused on perceptions of the acceptability of the interventions. The interview was led by the topic guide, exploring each facet of acceptability in turn. After each question, the interviewer prompted participants to expand on their answers or asked relevant follow-up questions to ensure that participants fully expressed their views. All interviewers conducted at least 1 pilot interview. Interviews were recorded using a Dictaphone. They took between 12 and 34 minutes to complete. Multimedia Appendix 2 contains the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [24].

Research Team and Reflexivity

Personal Characteristics of Interviewers

There were 3 interviewers. MO is a female senior research fellow at UCL and has a PhD in health psychology. She is predominantly a quantitative researcher but undertook qualitative interviewing training in advance of the interviews. LMD is a female research assistant at UCL and has an MSc in psychiatric research. She has experience in conducting interviews remotely and has undertaken training sessions in qualitative research. GL is a female research fellow (trial manager) at UCL and has an MSc in developmental and educational psychology. She has experience undertaking qualitative interviews and conducting interviews remotely by phone.

Relationship With Participants and to the Topic

All interviewers had little contact with participants before the interview, save for reminder emails sent at 1, 3, and 6 months. Interviewers may potentially also have made up to 2 reminder phone calls and sent a postcard and a letter at the 6-month follow-up. Participants may have had some knowledge of the interviewers and their roles within UCL and the trial team through study documentation. The goal of the research, to understand the acceptability of the digital intervention used, was explained to participants at the start of the interview. All 3 interviewers were blind to the outcome data at the time of the interviews.

Reflections on the Interview Process

The 3 interviewers thought that their demographic characteristics did not seem to impact the interview process. Participants seemed able to speak openly about their positive and negative experiences of using the app and web page. However, there are 2 points of interest. First, when participants reported negative points about the NHS alcohol advice web page, they often prefaced these comments by talking very positively about the NHS more generally. Second, in both groups, some participants apologized to the interviewer before reporting a negative experience or saying something they did not like about the web page or app. These points could suggest that participants experienced some social desirability bias and thought that the interviewers wanted them to report positive experiences.

Analysis

Interviews were transcribed verbatim, anonymized, and then uploaded into NVivo (version 12; QSR International) for coding and analysis. We followed a combined inductive and deductive approach. An initial coding framework was developed using a priori themes (eg, TFA facets and perceived personal relevance). Two researchers (MO and LMD) coded the first 5 interviews separately, and then an iterative process of cross-checking coding strategies and data interpretation was carried out to establish a consensus and develop a revised coding frame. Coding was further refined using an ongoing comparative method, whereby each interpretation and finding was compared with existing findings, as more transcripts were analyzed. Following initial coding, similar responses within each construct were inductively analyzed to generate content themes [22] representing how that construct contributed to the reported acceptability. Participant quotes are presented alongside the age and sex of the participants. Where duplicated, a letter has been added to the age.

Following analysis, a draft of the *Results* section of this report was shared with a subsample of 8 randomly selected participants (4 from each group). Participants were asked to comment on how well the relevant *Results* section summarized their views on the acceptability of the digital tool and whether there were key points that they thought were missing from this summary. Participants largely thought that the *Results* accurately represented their experiences, and no changes were made. Feedback is presented in Multimedia Appendix 3.

Results

Participant Characteristics

Of the 5602 trial participants, 4443 (79.31%) consented to the interview at baseline or 6-month follow-up. A total of 55 people were invited to the interview, of whom 53% (n=29) declined or did not respond, resulting in a final sample of 26 participants. Table 1 provides an overview of participant characteristics overall and by group.

A definition and an overview of themes are presented in Table 2 for both groups.

XSL•FO
RenderX

Table 1. Participant characteristics overall and by group.

	Drink Less (n=14)	NHS[a] alcohol advice web page (n=12)	All (n=26)
Sex (female), n (%)	10 (71)	5 (42)	15 (58)
Age (y), mean (SD; range)	41.64 (16.19; 22-72)	43.50 (14.08; 23-68)	42.50 (14.98; 22-72)
Ethnicity, n (%)			
Asian	1 (7)	0 (0)	1 (4)
Black	1 (7)	1 (8)	2 (8)
White	9 (64)	9 (75)	18 (69)
Mixed race	2 (14)	1 (8)	3 (12)
Other	1 (7)	1 (8)	2 (8)
Higher income[b], n (%)	10 (71)	8 (67)	18 (69)
Engagement, n (%)			
Low	2 (14)[c]	—[d]	—
Moderate	5 (36)	—	—
High	7 (50)	—	—
Ever used	—	10 (83)	—
Not used[e]	—	2 (17)[e]	—
Global acceptability, mean (SD; range[f])	3.86 (0.74; 2-5)	2.72 (1.20; 1-5)	—

[a]NHS: National Health Service.

[b]Earning more than £26,000 (US $33,093.58).

[c]These participants were originally categorized as nonusers of the app based on engagement data but self-reported briefly using the app in interviews; they may not have provided accurate linking data between the app and trial so they are categorized as having low engagement here.

[d]Not applicable.

[e]Reported never using the National Health Service (NHS) alcohol advice web page at 6 months but reported looking at it briefly at the start of the trial in the interviews.

[f]Participants were asked to judge the global acceptability of the intervention on a scale of 1 to 5 stars.

XSL•FO
RenderX

Table 2. Definition and overview of themes.

Theme	Definition of facet	Drink Less app	NHS[a] alcohol advice web page
Affective attitude	How an individual feels about the intervention	• Generally liked the app, particularly the drinking diary • Felt proud and happy when meeting goals • Some negative affect when failing to meet goals	• Most reported neutral or negative affect • Some felt that the web page was patronizing
Burden	The perceived amount of effort necessary to use intervention	• Quick and easy to use • Could tailor time spent based on time available or support required • Repetitive in the longer term	• Quick and easy to use • Some framed this as a negative aspect and thought the web page was too basic
Ethicality	The extent to which the intervention has a good fit with an individual's value system	• Generally accessible • Mixed views on whether the app would work for those who had less experience of digital tools	• Generally accessible • Concerns around the availability of treatment through the NHS as recommended on the web page and confidentiality of revealing drinking to health care professionals
Intervention coherence	The extent to which the participant understands the intervention and how it works	• The app was generally considered intuitive • Some reported a learning curve or difficulties with features • Participants reported understanding many of the mechanisms of change (eg, tracking and goal setting)	• Easy to use and navigate
Opportunity costs	The extent to which benefits, profits, or values must be given up to engage with the intervention	• Most reported no opportunity costs • Some thought that reducing drinking impacted social life • Others reported reducing lone drinking rather than social drinking, which they felt would have been more of an opportunity cost	• No opportunity costs
Perceived effectiveness	The extent to which the intervention is perceived as likely to achieve its aim	• Most thought the app helped them reduce their drinking for the time that they used it • Some contextual factors were a barrier to change (eg, the COVID-19 pandemic) • Disengagement from the app in the longer term	• Most thought that the web page did not directly help them reduce their drinking • Some found it helpful as a signposting tool to other resources
Perceived personal relevance[b]	The extent to which the intervention is suited to the participant's individual needs	• Highlighted the tailored toolbox nature of the tool—could pick the features they liked • Remote nature did not suit some, but some liked the anonymity	• Many thought they already knew the information, and it was too generic • Participants reported that the web page might be helpful for other people but was not personally suited to them
Self-efficacy	The participant's confidence that they can perform the behaviors required to participate in the intervention	• Mixed confidence in whether the app would work for them • Trust and confidence associated with UCL[c] branding • Importance of personal motivation	• Mixed confidence in whether the web page would work for them • Importance of personal motivation

[a]NHS: National Health Service.

[b]Not a facet of the Theoretical Framework of Acceptability.

[c]UCL: University College London.

XSL•FO
RenderX

The Drink Less App

Affective Attitude

Participants generally liked the Drink Less app. They reported particularly liking the tracking component (Self-Monitoring and Feedback) and the traffic light color–coded feedback on the calendar for alcohol-free, light, and heavy drinking days:

> *I quite liked the little graph that shows whether you're on track.* [Female, 60 years]
>
> *I liked the fact that it was nice and visual so the calendar, where it came up with red, green or amber I found that quite useful for me.* [Female, 38a years]

Participants reported feeling positive and proud when they were meeting their goals to reduce their consumption:

> *I was encouraged by the calendar where you have the days when you don't drink, and having it consistently green week after week, it made me feel nice and I wanted to continue doing that.* [Male, 41 years]
>
> *I did like the fact that if I logged on that I hadn't drank it sort of praised me.* [Female, 55 years]

A few participants reported negative affect when recording heavy drinking, relapsing, or failing to meet their goals. They reported not liking having a visual account of their failure:

> *I didn't want to visually see my mistakes on my phone and yeah I think judgment on myself as well.* [Female 26A years]
>
> *When you do start going backwards and drinking more and more than you know you're all you're doing is sort of putting effectively negative data into the app. And you just feel like you've let yourself down, you've let the app down and you've let your progress down.* [Male, 22 years]

Burden

Participants generally reported that the app was not time consuming to use and that they found it quite easy and user-friendly:

> *It is easy to use, it's quick it's not onerous. (Female, 47 years)*

Some participants talked about spending more or less time on the app based on the level of support they felt they needed each day or the amount of time they had available:

> *I could spend over an hour going through it and making a plan and you know for that if I thought I needed the extra support and other days it just took five minutes, maybe 10 to just log in you know what was happening.* [Female, 67 years]
>
> *Other times I'm kind of free, so I don't see maybe I'm not doing anything I'm just bored or something like that and I start using the app...so sometimes time consuming sometimes not.* [Male, 30B years]

In the longer term, using the app was described by some as repetitive. Over time, this led to negative emotion or forgetting and eventually disengagement:

> *As it went on I found it a bit of a chore and kind of forgot about it.* [Female, 26A years]
>
> *I was using it a lot at first. But then I sort of kind of lost interest in the app.* [Male, 22 years]

Ethicality

This theme is focused on fairness and accessibility. Participants reported thinking that the app was accessible for most people who had access to a smartphone:

> *Anyone who's got a smartphone and uses Apps can use it, but that isn't you know, obviously that isn't everyone.* [Female, 47 years]

Participants felt that to be fair to all users, the app should be accessible to everybody. Some participants highlighted that those who are older or less technically able might struggle to use Drink Less. Others thought it was already broadly accessible:

> *I'm sure anybody can use it, but it needs to be a bit simplified if you want people who are less tech-savvy to use it.* [Male, 41 years]
>
> *I'm sort of thinking about my mom who isn't very tech minded I think if the app was downloaded for her and she had a tablet... she'd still be able to use it as well, so yeah it's easy to use for anybody I would think.* [Female, 55 years]

Intervention Coherence

Most said that the app was intuitive:

> *I didn't have any problem with it, was very usable all the labels and options, they were all self-explanatory.* [Male, 30A years]
>
> *It was easy to download and easy to just get up, set up and start using.* [Female, 55 years]

Some participants reported a learning curve when they first started using the app, others reported specific issues with using different features of the app, such as logging cocktails for which there was no default option, entering the cost of drinks, and customizing goals:

> *In the beginning...I was a bit overwhelmed, what am I supposed to do now? Am I supposed to do this game or do that. But over time I realised that, first of all, you don't have to do any of them, it will still work.* [Male, 41 years]
>
> *The one thing that I just found a bit confusing was like, if I had a drink and I didn't know how to record it because yeah there wasn't an option for some things.* [Female, 25 years]

Participants seemed to be aware of some of the underlying behavior change components through which the app worked. Participants commented on the importance of tracking in highlighting how much they drank and reported that the praise and the traffic light function in the calendar encouraged them to have more alcohol-free days. Others commented on reflecting on the effect that alcohol had on their mood and sleep and said the app dispelled myths around alcohol helping them sleep or relax:

XSL•FO
RenderX

> *I was encouraged by the calendar where you have the days when you don't drink, and having it consistently green week after week.* [Male, 41 years]

> *Even if I wasn't putting it into the app I was aware or reminded of the numbers of units per week that are heal- that's healthy. And having you know, a glass just a glass of wine on a random night adds to that total and it's unnecessary.* [Female, 47 years]

> *People always assume, or I'd always assumed, you have a drink and it helps you sleep but actually it really doesn't so I've been able to just...think about that a little bit more.* [Female, 38 years]

Opportunity Costs

Most participants said the Drink Less app fit into their lifestyle and did not interfere with other obligations:

> *You know if I couldn't do it that day, I could always try and do it the next day or anything like that or just not do it at all.* [Female, 25 years]

> *I'm really busy. I've got young children and a full time job and all the rest of it, and it was it was something that I have no problem incorporating into my routine.* [Female, 38A years]

Some participants suggested that the using the app to cut down on their drinking had indirect opportunity costs by interfering with their social life. Participants reported not wanting to drink more than their goal amount, and avoiding some social occasions with friends where they felt they might be tempted or pressured to drink:

> *I was like I only have four units left and like weekend is when I usually drink or like meet friends and things like that so yeah it was a it was a little bit like, I was kind of losing every week in a way.* [Male, 30A years]

Other people talked about how their use of the app did not impact their willlingness to socialize in drinking settings and how they used it more to cut out lone drinking or "mindless drinking":

> *I think it made me more conscious of erm having the odd drink here and there...I can't say it helped me with my drinking if I was out socializing in the pub.* [Female, 55 years]

> *If I'm going out with or with friends I will still drink, what it knocked on the head is the having a glass of wine while I'm cooking dinner for no particular reason drinking.* [Female, 47 years]

Perceived Effectiveness

Most participants thought the app helped them drink less for the time that they used it and described different strategies of using the app to successfully cut down their consumption. These included accurately monitoring units consumed with reference to the drinking guidelines, downsizing, cutting out habitual drinking, and aiming for more alcohol-free days:

> *Overall, I do think it helped me for that time, I think the other helpful thing was also just to like see in my life like what 14 units really looks like and how small It is.* [Male, 30A years]

> *I've not cut the days I drink back, but I instead of buying a whole bottle of wine, I buy a small bottle of wine.* [Female, 72 years]

> *When you see how many alcohol units you're using per week, or you're drinking per week it yeah it just kind of reset my drinking and it has knocked on the head the mindless drinking.* [Female, 47 years]

Not all participants thought the app helped them drink less. Some described contextual barriers to them drinking less, such as the COVID-19 pandemic and Christmas. Other participants described disengaging from the app and returning to heavier drinking in the longer term:

> *Perhaps, if I hadn't started at Christmas maybe or if the situation would have been different, then maybe I would have taken more notice of it.* [Female, 60 years]

> *I used it religiously for a while and then and then after that it is hard once you drop off again like I said when you go backwards it's hard to then go back on it and start again you just can't be bothered like and frankly it's like it's really hard to find the energy to like start afresh.* [Male, 22 years]

Participants had some suggestions on how to make the app more effective for them in the longer term, which included changing the features of the app more frequently and using the app in group or health care settings:

> *Perhaps a weekly zoom meeting with 20 people that are using the app to see how people are getting on there's no shame and you can sort of just say okay well okay you've you've taken a step back, but it's fine input your data, the next day, and the next day, and this is how to actually come back.* [Male, 22 years]

Perceived Personal Relevance

Most participants reported that the app was a good fit for them. The app was commonly perceived like a toolbox with different intervention components available to select. People reported using the app in different ways and finding the components of the app that worked best for them. Some participants said that the remote nature of the app and the anonymity it offered were a good fit for them:

> *The good thing about it is that it has various tools and games and I am sure that not every one of these functions will appeal to every single user, so from my perspective, having a supermarket function was very useful.* [Male, 41 years]

> *I wasn't in a position to go to the doctor and didn't feel like seeing someone in person, so the app was a personal way of getting support anonymously.* [Male, 41 years]

Not everyone felt that the app was a good fit for them; some participants reported needing more support or not liking the digital and remote nature:

XSL•FO
RenderX

I don't think that the just reading things on the screen was something that appealed to me, personally, but may appeal to others. [Female, 60 years]

Perceived Self-Efficacy

Participants reported mixed levels of confidence in whether the app would work for them:

I was very confident it would work out for me. [Male, 30B years]

I wasn't particularly confident because I didn't know what it was and how it would you know run run out and how I could use it and what other things were in it. [Female, 67 years]

Participants reported having confidence and trust in the app based on its association with UCL:

I felt confident, I thought you know it's by UCL, I am sure its trustworthy so yeah, I was confident. [Female, 25 years]

Confident another way, I suppose I trusted it. I believe what it was telling me...I suppose I assumed by that that there'd been accurate research by the people that had developed the app. [Female, 55 years]

Participants reflected on the importance of their own motivation to change their behavior and how this determined whether they were confident in the Drink Less app:

Now, can it help me now? I don't think so because I don't have the same motivation as in February when I started. [Male, 41 years]

I think say someone I knew had a problem, and they were ready to really address it and sort it out I'd recommend it to them. Just to give it a go but. I just don't think maybe I'm not there yet ready I don't know, maybe in the future it be helpful. [Female, 26A years]

NHS Alcohol Advice Web Page

Affective Attitude

Participants generally reported neutral or negative affect when discussing the NHS alcohol advice web page. Some reported finding it basic and patronizing:

The website itself was fine it was okay I liked it no kind of bad feelings about the actual website or kind of irritations or anything. [Female, 40 years]

More like er a nagging er nagging grandmother. [Male, 68 years]

I found it vaguely informative, but a lot of it was just common sense, so in a way, in a way, some of it felt almost like a little bit patronising. [Male, 30 years]

Burden

The NHS alcohol advice web page was generally considered very easy and quick to use:

Very easy to read and clear and concise. [Male, 30 years]

However, this was also framed as a negative aspect of the web page. Some said they used the web page only briefly, as the information contained was basic. Others suggested that the lack of change or interactive features meant that they did not return to the web page:

Very, very sparse. [Male, 42 years]

I mean you know it was only a few visits to it so um you know after that you you're not going to gain any more from it. [Male, 68 years]

Ethicality

Participants reported thinking that the NHS alcohol advice web page was accessible for most people who had access to the internet:

Anyone who can use internet could use the website. [Male, 42 years]

I feel like it's very accessible to all, especially in this day and age. [Male, 23 years]

There was a range of concerns about the treatments offered by the NHS and around negative consequences if they honestly reported their alcohol consumption to a health care professional:

The help that they suggest is available isn't always available easily. [Female, 55A years]

I don't think there's enough done to help people with it [alcohol dependence] in all honesty I think it's just a case of people get stuck in some sort of rehab which is generally lumped into some sort of mental welfare ward. [Male, 36 years]

I was worried about driving lessons being taken away.. about social services, because I had a child things like that...if there was in the early days, some kind of reassurance that you could you know get a certain amount of help, without any repercussions then people would actually take part in the options that are available and suggested. [Female, 40 years]

Intervention Coherence

Participants reported confidence in navigating the NHS alcohol advice web page; the titles were clear, and it was easy to use:

It was very clear and everything's labelled nicely and kind of navigating the website is very, very simple. [Male, 23 years]

I felt I could use it pretty instinctively. [Female, 26 years]

Opportunity Costs

No participants reported that the NHS alcohol advice web page had interfered with anything else important in their life:

It wouldn't be something that I'd look at instead of day to day life. [Male, 23 years]

Perceived Effectiveness

Participants reported that while the NHS alcohol advice web page did not directly affect their drinking, they thought they would have benefited from more of a call to action:

XSL•FO
RenderX

> *It's all good and well telling me all these facts and giving me information about how to deal with drinking and what it can do to you but really it's not enough.* [Male, 23 years]

> *I sometimes wonder if they should be a bit harsher but you know like this is the consequences or something you know.* [Female, 55A years]

> *There was nothing in the NHS incentivised you to say oh right okay, but basically it was just basically textbook stuff on the internet.* [Male, 68 years]

For some participants, it served as a first step signposting to other resources or motivating them to independently search for other resources:

> *It started me on the journey and it pointed me in the right direction.* [Female, 44 years]

> *At least it made me look at other things, and eventually identify, something that was really helpful.* [Female, 61 years]

Perceived Personal Relevance

Many participants reported feeling like the NHS alcohol advice web page was not relevant to them. They thought that the information presented was too generic and would benefit from more tailored components:

> *For my my individual needs for what it was, I didn't think it was that well suited to compared to like I said if there'd been something a little bit more tailored.* [Male, 30 years]

> *A kind of screening tool at that point might have been helpful to then identify people with moderate alcohol problems to severe alcohol problems. And then you could then guide them or direct them to a more appropriate channel rather than treat everybody the same whether they drink five bottles of spirits a day or two glasses of wine.* [Female, 61 years]

Many participants talked about the NHS alcohol advice web page perhaps being a good fit for other people and acknowledged the need for the web page to be quite generic given that its purpose was to serve a varied group of people:

> *It might have served other types of persons needs but not mine. Perhaps I wanted too much from it, perhaps I wanted something too individual.* [Female, 61 years]

> *It was good and informative but quite generic. You know, which is suppose it has to be if a lot of people are using it.* [Female, 26 years]

> *But because of what it is designed to be for the NHS I suppose it needs to be that way because it'll have people who sort of drinking a bottle of vodka a day to someone who they thought I might give it a Google because, last night I hit it a little bit too hard.* [Male, 30 years]

Perceived Self-Efficacy

Participants reported mixed levels of confidence in using the NHS alcohol advice web page to reduce their drinking:

> *I don't have a lot of faith in the NHS, as much as I love the NHS, I don't have a lot of faith in the help that's available for this type of thing.* [Female, 55a years]

> *It's NHS so it's recognisable and you feel comfortable using it, because it's trusted source.* [Male, 23 years]

As in the Drink Less group, participants highlighted the importance of their own motivation to change:

> *I think that the study came at a really good time for me, that time that I was ready to face up to the fact that I needed help, and I wanted to be accountable as well.* [Female, 44 years]

Discussion

Principal Findings

Participants reported liking the Drink Less app; they particularly liked the tracking and feedback components, which they considered made the app more personally relevant to them, and felt proud and positive when they were meeting their goals. They discussed that the app functioned as a "supermarket" or toolbox, whereby they could choose and use the components of the app that worked best for them. Participants reported different strategies and goals, and most thought that the app was effective in reducing their alcohol consumption, particularly in the shorter term. This "supermarket" function extended to the depth of use, and participants also reported that use was influenced by the level of support they felt they needed that day—this could include spending more time on the app making plans and behavioral substitutions when their cravings were higher. Participants reported being confident in using the app and that it was intuitive and accessible. However, some thought that there was a learning curve at the start of using the app, and some participants reported specific difficulties in using the app, such as logging cocktails (which are not included as default options in the drinking diary) or customizing goals. Some participants reported negative affect when logging heavier drinking days or failing to achieve their goals, which led to disengagement in the longer term. Another factor reported as leading to disengagement was boredom. Some participants reported that although the app was not burdensome to use, it became something of a chore.

Participants reported that the NHS alcohol advice web page was very quick, easy, and intuitive to use and accessible to anyone with internet access. Participants reported that the web page could be a useful tool for other people, but they judged that the information contained was less personally relevant to them, and the web page was perceived as basic and generic. Some found that the web page had provided the starting point for them in reducing their alcohol consumption by signposting them to other tools or resources, whereas others thought it had not had an impact on their alcohol consumption. There were concerns raised by participants about contacting a health care professional, as some thought this would be ineffective and were not confident that the services would be available and one parent feared that social services would be contacted. Participants suggested that the web page could benefit from having some more personalized features, such as signposting for different levels of consumption.

XSL•FO
RenderX

We have avoided drawing comparisons between the 2 tools throughout this analysis because the nature of the 2 tools is different. The trial may have set up the evaluation of the acceptability of the tools to encourage participants to think of the NHS alcohol advice web page as a stand-alone or prolonged intervention. However, unlike the Drink Less app, the alcohol advice web page focuses on information provision and does not offer opportunities for personalization or engagement.

These data are consistent with work highlighting the importance of perceived personal relevance [15,16] and a recent evaluation of an app designed to help ex-service personnel reduce their drinking that highlighted credibility, ease of use, and personalization as important functions of the app [25].

Although participants perceived the app as potentially effective in helping them reduce their alcohol consumption in the shorter term, they felt that some factors, such as repetition and negative affect, led to their disengagement from the app in the longer term. Achieving prolonged engagement with digital interventions [26] and apps more widely [27] is a known challenge. Participants had some ideas, such as new features or group use, that could have boosted their engagement in the longer term. These findings suggest that the Drink Less app should be viewed as a dynamic intervention that new evidence-informed components and features could be added to. This does raise practical issues, as the app is currently managed by an academic research team with limited resources available for ongoing management and development costs. One implication of this research is, therefore, to consider how academics and third-sector partners can work together to continue to develop and maintain digital interventions with evidence of effectiveness and acceptability to end users.

Among this sample, Drink Less seems to be an acceptable intervention. However, the generalizability of these findings should be considered. Digital exclusion [28] is an important factor to consider when promoting digital interventions. Some people may be unable to access devices or data or be unable to make the most of them due to a lack of knowledge or resources or are less likely to engage with them [28]. Digital exclusion is more likely in vulnerable populations, including older people, those out of work, the most financially vulnerable people, and those who live with a condition that limits or impairs their use of communication services. Therefore, researchers and policy makers should be aware of how the introduction of digital interventions could impact widening inequalities, as digital interventions such as smartphone apps are not going to be accessible and acceptable for everyone. Metrics of variation across different protected characteristics should be measured in the roll out of all digital public health tools and explored in process work.

The topic guide was developed in collaboration with experts by experience to ensure the questions were understandable and addressed the underlying concepts. We asked participants to provide feedback on the *Results* sections (Multimedia Appendix 3), and participant feedback was positive. We did our best to avoid interviewing only participants who were more engaged with the interventions and, therefore, more likely to have favorable views, by purposively sampling participants based on a range of engagement levels. However, a limitation of this study is that we were not able to interview any participants who did not respond to 6-month follow-up in the wider iDEAS trial. Those responding at 6 months may have been more engaged with the app and the trial than those who did not respond at 6-month follow-up and, therefore, may have had more positive views of the digital interventions than those that were less engaged. In a similar vein, is possible that participants who were successful in their alcohol reduction goals would have held more positive views on the intervention. We did not measure whether participants felt they had achieved their alcohol reduction goals, and as such, this is something we cannot examine.

Conclusions

The Drink Less app appears to be an acceptable intervention for increasing- and higher-risk drinkers captured in this sample. To further this work, future research could examine the relationship between acceptability and engagement and consider how engagement with the app can be increased in the longer term. The NHS alcohol advice web page was not considered acceptable as a stand-alone intervention but may act as a positive signpost for some users.

Acknowledgments

This study is funded by the National Institute for Health and Care Research (NIHR; Public Health Research Programme, grant 127651). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

MO's salary is funded by NIHR (NIHR127651) and Medical Research Council (MR/W026430/1). CG's salary is funded by NIHR (NIHR127651) and Cancer Research UK (CRUK: C1417/A22962). EP is supported by the NIHR Applied Research Collaboration North Thames. FG is supported by the NIHR Applied Research Collaborative for Northwest London.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

MO led the design, data collection and analysis, and drafted the initial manuscript. LMD contributed to data collection and analysis. GL contributed to data collection. All authors reviewed protocols, contributed to the conceptualization and methodology, and reviewed the final manuscript.

Conflicts of Interest

JB received unrestricted funding related to smoking cessation research and sits on the scientific advisory board for the SmokeFree app. CG and MO have done paid consultancy work for the behavior change and lifestyle organization One Year No Beer, providing fact checking for blog posts. FG is employed by the Department of Health and Social Care and was previously employed by the National Institute of Health and Care Excellence. Outside the submitted work, EK has previously coauthored papers that analyzed raw market research consumer-based data provided to Newcastle University under a direct contract with Kantar Worldpanel at no cost to Newcastle University. Kantar Worldpanel received reimbursement from AB InBev to cover the costs of the data, with Kantar Wordpanel having similar commercial relationships with other customers who pay to have data collected on food and nonfood items available for sale in supermarkets and other retail outlets covered by the Worldpanel. MH is a trustee of the Society of Study of Addiction. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Interview schedule.
[DOCX File , 15 KB - jmir_v26i1e42319_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 418 KB - jmir_v26i1e42319_app2.pdf]

Multimedia Appendix 3
Public and patient involvement feedback on results.
[DOCX File , 22 KB - jmir_v26i1e42319_app3.docx]

References

1. Global status report on alcohol and health. World Health Organization. 2018. URL: https://www.who.int/publications-detail-redirect/9789241565639 [accessed 2024-04-29]
2. Brown J, West R, Angus C, Beard E, Brennan A, Drummond C, et al. Comparison of brief interventions in primary care on smoking and excessive alcohol consumption: a population survey in England. Br J Gen Pract 2016 Jan;66(642):e1-e9 [FREE Full text] [doi: 10.3399/bjgp16X683149] [Medline: 26719481]
3. Nilsen P, Kaner E, Babor TF. Brief intervention, three decades on: an overview of research findings and strategies for more widespread implementation. Nordisk Stud Alkohol Drug 2017 Apr 12;25(6):453-467. [doi: 10.1177/145507250802500608]
4. Wilson GB, Lock CA, Heather N, Cassidy P, Christie MM, Kaner EF. Intervention against excessive alcohol consumption in primary health care: a survey of GPs' attitudes and practices in England 10 years on. Alcohol Alcohol 2011;46(5):570-577 [FREE Full text] [doi: 10.1093/alcalc/agr067] [Medline: 21690169]
5. Rosário F, Santos MI, Angus K, Pas L, Ribeiro C, Fitzgerald N. Factors influencing the implementation of screening and brief interventions for alcohol use in primary care practices: a systematic review using the COM-B system and Theoretical Domains Framework. Implement Sci 2021 Jan 07;16(1):6-7 [FREE Full text] [doi: 10.1186/s13012-020-01073-0] [Medline: 33413487]
6. Kaner EF, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
7. West R, Michie S. A Guide to Development and Evaluation of Digital Behaviour Change Interventions in Healthcare. Oxford, United Kingdom: Silverback Publishing; 2018.
8. Ofcom communications market report. Office of Communications, an independent statutory corporation in the United Kingdom (Ofcom). 2018. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf [accessed 2022-03-21]
9. Crane D, Garnett C, Brown J, West R, Michie S. Behavior change techniques in popular alcohol reduction apps: content analysis. J Med Internet Res 2015 May 14;17(5):e118 [FREE Full text] [doi: 10.2196/jmir.4060] [Medline: 25977135]
10. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]

XSL•FO
RenderX

11. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
12. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
13. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 06;22(7):e17256 [FREE Full text] [doi: 10.2196/17256] [Medline: 32628122]
14. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
15. Keyworth C, Quinlivan L, Leather JZ, Armitage CJ. Exploring the acceptability of a brief online theory-based intervention to prevent and reduce self-harm: a theoretically framed qualitative study. BJPsych Open 2022 Oct 12;8(6):e184 [FREE Full text] [doi: 10.1192/bjo.2022.568] [Medline: 36221254]
16. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
17. Garnett C, Crane D, West R, Brown J, Michie S. The development of drink less: an alcohol reduction smartphone app for excessive drinkers. Transl Behav Med 2019 Mar 01;9(2):296-307 [FREE Full text] [doi: 10.1093/tbm/iby043] [Medline: 29733406]
18. Tips on cutting down. National Health Service. 2022. URL: https://web.archive.org/web/20230324141825/https://www.nhs.uk/live-well/alcohol-advice/tips-on-cutting-down-alcohol/ [accessed 2023-03-24]
19. Garnett C, Oldham M, Angus C, Beard E, Burton R, Field M, et al. Evaluating the effectiveness of the smartphone app, Drink Less, compared with the NHS alcohol advice webpage, for the reduction of alcohol consumption among hazardous and harmful adult drinkers in the UK at 6-month follow-up: protocol for a randomised controlled trial. Addiction 2021 Feb 26;116(2):412-425 [FREE Full text] [doi: 10.1111/add.15287] [Medline: 33067856]
20. Garnett C, Perski O, Michie S, West R, Field M, Kaner E, et al. Refining the content and design of an alcohol reduction app, Drink Less, to improve its usability and effectiveness: a mixed methods approach. F1000Res 2021;10:511 [FREE Full text] [doi: 10.12688/f1000research.51416.2] [Medline: 34646502]
21. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
22. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
23. Mills N, Donovan JL, Smith M, Jacoby A, Neal DE, Hamdy FC. Perceptions of equipoise are crucial to trial participation: a qualitative study of men in the ProtecT study. Control Clin Trials 2003 Jun;24(3):272-282. [doi: 10.1016/s0197-2456(03)00020-5] [Medline: 12757993]
24. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. Puddephatt JA, Leightley D, Palmer L, Jones N, Mahmoodi T, Drummond C, et al. A qualitative evaluation of the acceptability of a tailored smartphone alcohol intervention for a military population: information about drinking for ex-serving personnel (InDEx) app. JMIR Mhealth Uhealth 2019 May 24;7(5):e12267 [FREE Full text] [doi: 10.2196/12267] [Medline: 31127726]
26. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
27. Mobile app user retention rate by category. Statista. 2020. URL: https://www.statista.com/statistics/259329/ios-and-android-app-user-retention-rate/ [accessed 2024-04-29]
28. Digital exclusion: a review of Ofcom's research on digital exclusion among adults in the UK. The Office of Communications. 2022. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/234364/digital-exclusion-review-2022.pdf [accessed 2024-04-29]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
NHS: National Health Service
TFA: Theoretical Framework of Acceptability
UCL: University College London

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, S Ma; submitted 18.05.23; peer-reviewed by S Toukhy, JA Puddephatt, L Fenton; comments to author 24.02.24; revised version received 08.03.24; accepted 14.05.24; published 18.07.24.

Please cite as:
Oldham M, Dina LM, Loebenberg G, Perski O, Brown J, Angus C, Beard E, Burton R, Field M, Greaves F, Hickman M, Kaner E, Michie S, Munafò MR, Pizzo E, Garnett C
Evaluating the Acceptability of the Drink Less App and the National Health Service Alcohol Advice Web Page: Qualitative Interview Process Evaluation
J Med Internet Res 2024;26:e42319
URL: https://www.jmir.org/2024/1/e42319
doi:10.2196/42319
PMID:

©Melissa Oldham, Larisa-Maria Dina, Gemma Loebenberg, Olga Perski, Jamie Brown, Colin Angus, Emma Beard, Robyn Burton, Matt Field, Felix Greaves, Matthew Hickman, Eileen Kaner, Susan Michie, Marcus R Munafò, Elena Pizzo, Claire Garnett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Web-Based Resilience-Enhancing Program to Improve Resilience, Physical Activity, and Well-being in Geriatric Population: Randomized Controlled Trial

Yi-Chen Wu[1], PhD; Shu-Fen Shen[2], PhD; Liang-Kung Chen[3,4,5], MD, Prof Dr; Heng-Hsin Tung[4,6], Prof Dr

[1]College of Nursing and Health Sciences, Dayeh University, Changhua, Taiwan
[2]Department of Nursing, Mackay Medical College, Taipei, Taiwan
[3]Center for Geriatrics and Gerontology, Taipei Veterans General Hospital, Taipei, Taiwan
[4]Center for Healthy Longevity and Aging Sciences, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Taipei Municipal Gan-Dau Hospital, Taipei, Taiwan
[6]Department of Nursing,National Yang Ming Chiao Tung University, Taipei, Taiwan

Corresponding Author:
Heng-Hsin Tung, Prof Dr
Department of Nursing,National Yang Ming Chiao Tung University
No.155, Sec.2, Li-Nong Street
Pei-Tao district 112
Taipei, 11221
Taiwan
Phone: 886 0953005831
Email: shannontung719@gmail.com

Abstract

Background: Resilience is a protective factor in healthy aging, helping to maintain and recover physical and mental functions. The Resilience in Illness Model has proven effective in fostering resilience and well-being. Physical activity is crucial for older adults' independence and well-being, even as aging causes a progressive decline. Additionally, older adults face challenges such as spousal loss and physical disability, making preventive intervention strategies necessary.

Objective: This study aims to develop and evaluate a web-based program to enhance resilience, physical activity, and well-being among community-dwelling older adults. Additionally, we aim to gather feedback on the program's strengths and limitations.

Methods: A 4-week resilience-enhancing program was created, incorporating role-play and talk-in-interaction and focusing on 3 key skills: coping, control belief, and manageability. The program included scenarios such as becoming widowed and suffering a stroke, designed to engage older adults. A pilot test preceded the intervention. As a result of the COVID-19 pandemic, the program shifted from in-person to web-based sessions. A single-blind, parallel-group, randomized controlled trial was conducted. Participants aged over 65 years were recruited offline and randomly assigned to either an intervention or control group. A certified resilience practitioner delivered the program. Outcomes in resilience, physical activity, and well-being were self-assessed at baseline (T0), 4 weeks (T1), and 12 weeks (T2) after the program. A mixed methods approach was used to evaluate feedback.

Results: A web-based participatory program enhancing 3 skills—coping, control belief, and manageability for resilience—was well developed. Among 96 participants, 63 were randomized into the intervention group (n=31) and the control group (n=32). The mean age in the intervention group was 69.27 (SD 3.08) years and 74.84 (SD 6.23) years in the control group. Significant between-group differences at baseline were found in age (t45.6=–4.53, P<.001) and physical activity at baseline (t61=2.92, P=.005). No statistically significant between-group differences over time were observed in resilience (SE 7.49, 95% CI –10.74 to 18.61, P=.60), physical activity (SE 15.18, 95% CI –24.74 to 34.74, P=.74), and well-being (SE 3.74, 95% CI –2.68 to 11.98, P=.21) after controlling for baseline differences. The dropout rate was lower in the intervention group (2/31, 6%) compared with the control group (5/32, 16%). Moreover, 77% (24/31) of participants in the intervention group completed the entire program. Program feedback from the participants indicated high satisfaction with the web-based format and mentorship support.

Conclusions: This study demonstrated that a web-based resilience-enhancing program is appropriate, acceptable, feasible, and engaging for community-dwelling older adults. The program garnered enthusiasm for its potential to optimize resilience, physical activity, and well-being, with mentorship playing a crucial role in its success. Future studies should aim to refine program content, engagement, and delivery methods to effectively promote healthy aging in this population.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT05808491; https://clinicaltrials.gov/ct2/show/NCT05808491

(J Med Internet Res 2024;26:e53450) doi:10.2196/53450

KEYWORDS

geriatric population; community-dwelling older adult; web-based resilience-enhancing program; resilience; physical activity; well-being; pandemic.

Introduction

Population aging is a pressing global issue today, and the World Health Organization (WHO) advocates preserving intrinsic capacity and functional ability to promote healthy aging [1]. Resilience has been defined by the National Institutes of Health as a positive health outcome [2], and it remains a significant research focus aimed at facilitating healthy aging for both individuals and society at large [3,4]. Resilience is individually described as a protective factor—the capacity to overcome adversity and maintain or regain well-being [5,6]. It enables individuals to bounce back from setbacks by adapting quickly to unfavorable situations, which typically fluctuate across one's lifespan. Moreover, high resilience later in life is associated with optimal health and minimal disability [7,8], as noted even during the COVID-19 pandemic [9]. Importantly, accumulating evidence has shown that resilience interventions have produced positive effects and enhanced well-being in various populations, including the geriatric population [10-14]. However, intervention programs to promote well-being among Taiwanese community-dwelling older adults have yielded inconclusive results. For instance, a reexperiencing gratitude intervention improved well-being through the mediated effect of grateful emotions [15], whereas mental health promotion programs and reminiscence group activities enhanced well-being among institutionalized older adults [16,17].

The Resilience in Illness Model (RIM) has been shown to cultivate resilience and well-being [18,19]. Developed by Haase et al [20], this model adopts a positive health perspective to evaluate the combined contributions of biological, behavioral, and psychosocial factors in adolescents and young adults with chronic illnesses. In the RIM, there are 2 risk factors, namely, illness-related risk and defensive coping, along with 5 protective factors, including spiritual perspective, social integration, hope-derived meaning, family environment, and courageous coping. The model identifies 3 outcomes: resilience resolution, self-transcendence, and a sense of well-being [21]. According to the RIM, stressors, such as life events or disease-related health conditions, positively affect defensive coping and negatively impact courageous coping. Protective factors such as courageous coping (which includes confrontive, optimistic, and supportive strategies) and hope-derived meaning are associated with enhanced resilience [20]. Furthermore, coping, defined as behaviors that protect oneself by avoiding psychological harm from negative experiences, along with control belief, the ability to manage unexpected situations and facilitate cognitive strategies to lessen the negative consequences of adversity, and manageability, the ability to access sufficient resources to deal with adversity, are 3 key components in enhancing resilience [22-25].

Advanced age brings progressive structural and functional deterioration of physiological systems, negatively impacting an individual's daily life abilities. Consequently, older adults gradually become more dependent on support and assistance from others [26]. Physical functionality in older adults is determined by their ability to perform daily activities, and adequate physical activity is essential for maintaining or enhancing this functionality [27]. A high level of physical activity has been well-documented in numerous studies to provide significant and beneficial effects on older adults' independence, quality of life, and well-being [28-32]. In the context of healthy aging, physical function is positively associated with resilience. Additionally, physical activity influences the relationship between resilience and mental health [32,33].

In late adulthood, spousal loss is considered a significant stressor, often the most overwhelming transition, profoundly impacting both mental and physical well-being [34]. Spousal bereavement compromises older adults' well-being, often accompanied by other strains and losses such as compromised physical mobility, health, and sensory functions (eg, vision and hearing). Psychologically, it disrupts the meaning and order of daily life. The psychological adjustment in older bereaved spouses varies widely, influenced by protective factors such as resilience and personal and social resources, including social support, personality, and prior mental health [35]. Another adverse health outcome related to aging is the increased risk of chronic diseases, which contributes to higher disability incidence and functional decline among older adults. These outcomes result in illness, disability, death, and increased health care costs for individuals and nations [36]. The development of symptomatology and progression in chronic diseases are influenced by psychological processes such as stress and resilience [37]. Additionally, the prevalence of chronic illness among Taiwanese older adults aged over 65 years has reached nearly 90% [38].

COVID-19 rapidly spread worldwide from December 2019 and caused a traumatic impact on the global economy, education, hospitality, sports, leisure, and especially individual daily life [39]. Older adults were considered the most vulnerable group during the pandemic, experiencing unintended physical, mental, emotional, social, and financial consequences due to COVID-19 control strategies [40]. Moreover, studies during the pandemic showed a significant decrease in physical activity among older adults, while those who regularly engaged in moderate or vigorous physical activity demonstrated higher resilience [30,40-42]. These studies also highlighted that resilience plays a protective and buffering role in older adults' psychological, physical, social, and economic health dimensions during the recovery process from the COVID-19 pandemic [43].

XSL•FO
RenderX

Healthy aging, which promotes well-being in older age by enhancing resilience, is the process of developing and maintaining independence, purpose, vitality, and quality of life in the geriatric population despite facing unexpected adversity [29,44]. This longevity should ideally correspond to an increase in healthy, disability-free years of life [45,46]. Population aging increases the burden on health care systems globally [45]. Health services play a pivotal role in healthy aging [46], using well-designed strategies to address the specific needs of older adults. It is crucial to develop preventive intervention strategies for the aging population to mitigate adverse aging-related health outcomes. The purpose of this study was to develop a web-based resilience-enhancing program for the geriatric population, evaluate its efficacy, and gather feedback on its strengths and limitations. The research hypothesis was that the web-based resilience-enhancing program improved resilience, physical activity, and well-being in community-dwelling older adults.

Methods

Study Design and Setting

The study was a 1:1 parallel-group randomized controlled trial (RCT) using purposive sampling to recruit community-dwelling older adults. Participant selection was fully automated and randomized equally to either the intervention or control group. As a result of social distancing and quarantine during the COVID-19 pandemic, we substituted the in-person 4-week resilience-enhancing program with a web-based approach. Outcomes, including resilience, physical activity, and well-being, were measured at baseline (T0), 4 weeks (T1), and 12 weeks (T2) after completing the resilience program or after allocation.

Recruitment

We invited community-dwelling older adults aged 65 years or older, willing to participate, and able to communicate in Mandarin or another Chinese dialect in Northern Taiwan. Participants with clinically significant severe cognitive dysfunction or psychotic disorders, or those currently undergoing treatment for progressive malignant neoplasms as self-reported, were excluded. We conducted a simple recruitment briefing at a community center, where researchers used posters to promote the study. Community-dwelling older adults interested in participating received study information and a consent form individually. Upon providing consent, participants completed demographic information and a baseline survey either independently or with assistance from researchers.

Randomization and Blinding

A single-blind RCT was conducted in 2 communities, comparing a web-based resilience-enhancing program with standard clinical practice. Participants were not blinded to allocation due to the nature of the training. Eligible participants were randomized in a 1:1 ratio to either the intervention or control group using a computer-generated random assignment scheme. Block randomization with block sizes of 4 was used to maintain balance throughout the trial. An independent third party managed blinding of the characteristic study participants. The researchers involved in data analysis were blinded to the group allocation.

Intervention

The web-based resilience-enhancing program was designed with consideration for the challenges faced by older adults in later life. It consists of 3 phases that include designing and testing role-playing scenarios. The first author (YCW) completed a training program and obtained certification as a resilience practitioner before designing the study.

The in-depth training program for resilience practitioners was provided by Dr Chris Johnstone, a leading resilience trainer in the United Kingdom and author of “Seven Ways to Build Resilience: Strengthening Your Ability to Deal with Difficult Times.” The program included evidence-based resilience tools, strategies, strengths, resources, and insights tailored for health care providers in roles supporting resilience among others dealing with stress and adversity. The toolkits delivered in the resilience practitioner training program included storyboarding, emotional first aid, flexible thinking, overload management, problem-solving strategies, support strengthening, and stickability. This training course is accredited by the Association for Coaching for continuing professional development (see Multimedia Appendix 1).

Coping, control belief, and manageability were the 3 key skills in the resilience-enhancing program we developed. Coping skills were demonstrated through emotional first aid in the practitioner training program, utilizing self-compassion, mindfulness, and acceptance and commitment therapy. Control belief skills were demonstrated through flexible thinking in the practitioner training program, incorporating cognitive therapy and thinking skills taught in the Penn Resilience Program. Manageability skills were demonstrated through overload management and problem-solving strategies in the practitioner training program. All therapies and strategies used in our resilience-enhancing program have been shown to bolster resilience in the general population [6,47-49].

Design Phase

The resilience-enhancing program in this study was designed based on Haase et al’s [14,18-20] RIM and related research. We identified stressors in older adulthood as including life events (eg, spousal loss) and disease-related health conditions (eg, suffering from paralysis after a stroke). Additionally, 3 main resilience skills—coping, control belief, and manageability—were incorporated into the modified RIM (Figure 1).

XSL•FO
RenderX

Figure 1. Modified RIM showed a stressor affected defensive coping positively and affected three protective factors as courageous coping, family function and social integration negatively, whereas resilience ability, as an outcome, was enhanced by the three skills directly and other protective factors, such as courageous coping, family function, social integration, meaningfulness indirectly. Blue circles indicate risk factors, whereas red circles indicate the three skills were practiced in the program. Black circles indicate protective factors and outcome. The arrows represent in direction of interaction between two circles. The plus sign + indicates positive effects, while the minus sign—indicate negative effects.

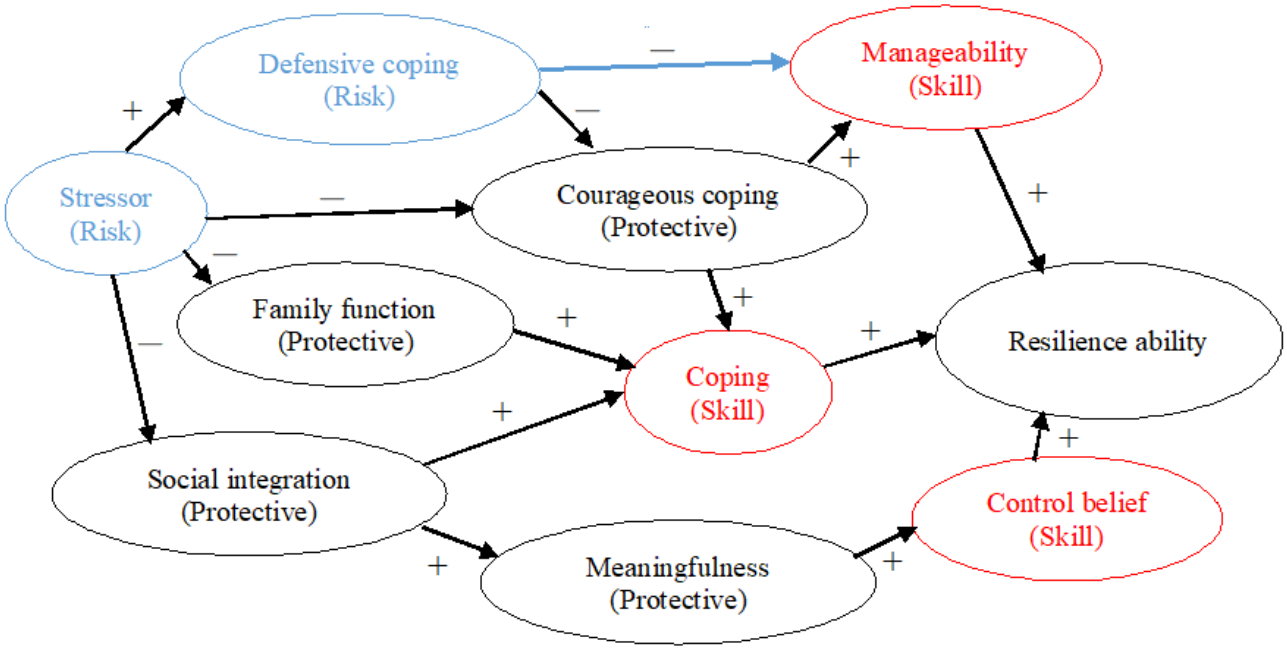

Furthermore, the learning approach we applied was situated teaching, which involves learning by doing and applying knowledge to real-life situations that older adults can relate to. Situated teaching is a social process that encourages the coconstruction and internalization of knowledge [50]. Methods used to apply situated learning included group activities, role-plays, scenario-based learning, and the use of technology, providing individuals with opportunities to practice new knowledge and skills in semirealistic contexts rather than simply reading a manual. This approach situates learning within specific social and physical environments [50,51].

In the design phase, we developed a participatory program to enhance 3 skills—coping, control belief, and manageability—for community-dwelling older adults based on a modified version of the RIM. Two scenarios were created based on the needs of community-dwelling older adults: a stressful life event of becoming widowed and a disease-related event of suffering from stroke, designed to engage participants in learning. The participatory program included 2 instructional scenarios, missions, developed scripts, and feedback for participants in 2 events, followed by resilient factor assessments to identify individual protective and risk factors (see Multimedia Appendix 2). Additionally, independent preparatory learning included 2 short films introducing resilience and demonstrating resilience skills, which were produced and provided via YouTube (Alphabet Inc./Google LLC).

Intervention Phase

The resilience-enhancing program involved weekly 60-minute role-playing and interactive talk sessions over 4 weeks. It comprised 4 sections: understanding resilience through e-learning, practicing resilience skills with a stressful life event and a disease-related event, and evaluation. We implemented a blended synchronous and asynchronous learning approach, which included independent preparatory learning with 2 microfilms (see Multimedia Appendix 3) on YouTube before 3 small-group facilitated sessions on weekly modules led by a certified resilience practitioner. Following a pilot test in January 2021, we transitioned from an in-person to a web-based approach due to COVID-19 pandemic control strategies.

After obtaining consent, we invited each eligible participant to LINE (LY Corporation) groups and divided them into small groups of 4-5 participants for convenience. The facilitated sessions were conducted in LINE group meetings with 4-5 participants each, led and guided by the certified resilience practitioner, who was also the first author (YCW). A third-party observer and assistant were present in each session. All participants received a weekly LINE message reminder to attend the sessions, regardless of their progress through the program. The purpose of these sessions was to debrief and help participants learn and demonstrate resilience skills, applying them to personal adversities in their daily lives. Additional details about the 4 sections are as follows:

- In the first section, we individually assessed 5 resilience protective factors and 2 resilience risk factors in the intervention group to understand the strengths and weaknesses of the participants. Asynchronous e-learning involved watching 2 resilience microfilms and completing a quiz with 5 questions related to the microfilms' content using registered Google Forms (Alphabet Inc./Google LLC; see Multimedia Appendix 4) before the second section.
- In the second section, the situated practice was employed using a scenario involving an older adult who had just lost a significant other. Participants mutually role-played 2 roles: a low-resilience person struggling to cope with the loss,

XSL•FO
RenderX

and a friend helping this person to clarify their needs and learn how to cope with a stressful life event. This included assessing available resources, locating support (family, social, and financial), and managing daily life. The tasks in this section included expressing negative feelings (the skill of coping), regaining control of one's life (the skill of control belief), and participating in social activities to increase positive feelings in the future (the skill of manageability). During the session, participants were invited to share their relevant experiences, and the practitioner provided feedback and guidance on enhancing resilience skills based on each individual's personal experiences.

- The third section, also utilizing situated practice, involved a scenario concerning an older adult who had just suffered paralysis after a stroke. Participants mutually role-played 2 roles: a low-resilience person unwilling to rehabilitate due to declining physical functioning and a friend providing help. The tasks in this section included accepting physical mobility impairments due to illness (the skill of coping), believing in oneself (the skill of control belief), and actively performing rehabilitation exercises (the skill of manageability). Similar to the second section, participants shared their personal experiences, and the practitioner provided individual feedback and guidance.
- In the final section, there was a review and practice of the 3 skills introduced in the 2 instructional scenarios, along with an evaluation of participants' performance. Further recommendations were provided for participants who did not perform well. The practitioner used techniques such as reframing passive to active thoughts and guided participants to recognize their capabilities in adverse situations. Emphasis was placed on the importance and benefits of becoming a resilient older adult.

In the second and third sections, each participant alternately played the role of both the friend giving coping strategy advice and the low-resilience person, as assigned by the resilience practitioner. For example, a participant who played the friend giving advice in the scenario of spousal loss would then play the role of the low-resilience person in the disease-related event scenario, and vice versa.

Evaluation Phase

To assess the efficacy of the intervention, outcome measures of participant engagement and feedback, including barriers to completion, were gathered through interviews and self-report feedback-marking sheets after the program. This focused specifically on evaluating the 3 resilience skills related to each scenario (see Multimedia Appendix 5).

Control

To serve as a comparison to the web-based resilience-enhancing program, participants in the control group received usual care, which included patient education sheets compliant with National Health Insurance regulations (see Multimedia Appendix 6). They were instructed not to attend any resilience classes or programs.

Outcome Measure

The key outcomes—resilience, physical activity, and well-being—were assessed at baseline (T0), 4 weeks (T1), and 12 weeks (T2) after completing the resilience program or after allocation. To facilitate participant convenience, telephone calls were made or stamped, self-addressed envelopes were provided for completing the self-assessed follow-up questionnaires. Additional telephone calls were made to participants who were unresponsive. Data were collected by a third party who was blinded to allocation. Demographic questionnaires, including information on age, gender, marital status, educational status, income, and other personal data, as well as details related to exercise habits such as type, frequency, and duration, were obtained for each participant.

Considering spousal loss as a stressful life event in our scenarios, it might be related to marital status among the participants. Therefore, we categorized marital status into 3 groups in our analysis: single and divorced, married, and widowed. This categorization accounts for potential differences in how individuals perceive and relate to the concept of spousal loss based on their marital status. For instance, individuals who are single or divorced may feel less connected to the concept compared with those who are married, where the scenario could be more stressful or tangible in practice. In addition, the impact of COVID-19 on monthly trends in older adults' physical activity [39-42] was investigated and categorized into 3 groups: increase, relatively stable, and decrease (see Multimedia Appendix 7).

Resilience

The Chinese version of the Resilience Scale (CRS) was chosen as the trial's key outcome measure, considering Wagnild and Young's [52] description of resilience as an internal resource akin to a protective factor. The Resilience Scale was originally developed to assess an individual's abilities that enable them to cope and face adversity successfully [52], and it was translated into Chinese by Li [53]. The validated CRS consists of 25 items assessed on a 7-point Likert scale, serving as a self-report measure for individual resilience across 5 domains: self-reliance, perseverance, equanimity, meaningfulness, and existential aloneness (score range of 25-175). A higher score indicates better resilience. The CRS has been utilized in numerous studies involving the geriatric population [54], given that older adults commonly encounter challenging adversities.

Physical Activity

Physical activity was assessed using the validated Chinese version of the Physical Activity Scale for the Elderly (PASE) [55]. Originally designed for large epidemiologic studies, the PASE measures the quantity and quality of physical activity performed by community-dwelling older adults in the previous week [56]. This assessment scale includes 12 activities that inquire about intensity, frequency, duration, and types of physical activity in recreational, household, and work-related settings, based on self-perceived performance. Each question's score was assessed using a dichotomy and/or a 4-point Likert scale, and calculated by multiplying the time spent in recreational and work-related activities or participation in

XSL•FO
RenderX

household activities by an empirically derived question weight (score range of 0-793). Higher scores indicate greater physical activity [57,58].

Well-Being

Well-being was assessed using the validated Chinese version of the Well-Being Scale (WBS) [59], adapted from the Oxford Happiness Inventory developed by Argyle [60]. It comprises 24 questions rated on a 5-point Likert scale, evaluating 4 domains: life satisfaction, interpersonal relationships, self-assurance, and physical and moral integrity. The total score ranges from 24 to 120, with higher scores indicating greater well-being. The WBS has been widely utilized in geriatric research in Taiwan [61-63].

Program Engagement and Feedback

Attendance of at least three sections was required for program completion. If a participant missed a session, they were invited to join another LINE group meeting for completion. Feedback data included interviews and self-report feedback-marking sheets. The interviews focused on participants' experiences in completing the program, their level of engagement, preferred program elements, aspects they liked and disliked, and suggestions for program improvement. After program completion, interviews were conducted with all participants in the intervention group to analyze varying levels of program completion. Self-report feedback-marking sheets highlighted 3 key issues: participants' confidence in coping (eg, facing and identifying problems, accepting current conditions), belief in themselves (eg, improving, receiving help, maintaining optimism), and ability to manage (eg, actively participating in physical and social activities, seeking support, knowing how to proceed) in adverse situations. A 4-point Likert scale was used, as indicated by Kusmaryono et al [64], to directly assess respondents on self-report feedback-marking sheets.

Sample Size Calculation

G*Power version 3.1.9.2 (Department of Criminology, University of Melbourne) was used to calculate the sample size for an *F* test of analysis of variance with repeated measures, between factors. The required effect size was determined based on the recommendations of Bartholomaeus et al [65] and calculated using the Cohen equation [66], resulting in an effect size of 0.3. Assuming a power of 80%, an α of .05, and an effect size of 0.3, the initial estimated sample size required would be 52 participants. To account for anticipated loss to follow-up in the geriatric population, an additional 20% (10/52) was recruited, bringing the total sample size to a minimum of 62 participants.

Statistical Analysis

We conducted intention-to-treat analyses on all participants (N=63), and missing data for all outcome measures were addressed through multiple imputations at the item level to minimize potential bias, following instructions reported previously [67,68]. IBM SPSS (version 24.0; IBM Corp.) was used to calculate demographics and outcome measures using means, SD, and frequencies. The significance of differences between groups at baseline was tested using independent samples *t*-tests and Pearson chi-square tests for ordinal and nominal variables, respectively. Group comparisons across the 3 time points were assessed using a 2-way repeated measures analysis of covariance. Additionally, the intraclass correlation coefficient was calculated to assess variance among time points, groups, and individuals. A linear mixed model was used to analyze the group-by-time interaction, with time points (T0, T1, and T2) and intervention (resilience vs control) as fixed effects, and individuals as a random effect, given that individual variability had the highest intraclass correlation coefficient ratio. Moreover, previous studies have demonstrated a significant impact of the COVID-19 pandemic on physical activity [31,40,42]. We accounted for the monthly trend of COVID-19 in Taiwan as a confounding factor in our analysis, categorizing it into 3 phases: escalation, stability, and decline of the epidemic. All analyses were adjusted for baseline values of the outcome measures, with statistical significance set at $P \leq .05$.

Descriptive statistics were used to summarize the program engagement data. Qualitative thematic analysis of the interview data was conducted independently by 2 authors (YCW and SFS) with over 5 years of experience in gerontology. The analysis involved a systematic process: familiarization with the transcripts, in vivo coding, categorization, and development of overarching themes to closely examine and interpret the participants' experiences with the resilience-enhancing program [69]. Three of the authors (YCW, SFS, and HHT) independently reviewed all final themes to ensure no significant patterns were overlooked.

Ethical Considerations

The trial protocol was reviewed and approved by the Institutional Review Board of National Yang-Ming University (approval number YM108150E; February 4, 2020). The protocol was registered in the ClinicalTrials.gov database with the identifier NCT05808491. Informed consent was obtained individually offline before study participation, adhering to the principles of the Declaration of Helsinki. All data were deidentified to protect the privacy and confidentiality of participants. Participants received an NT $100 (US $3.08) coupon as compensation each time they completed the survey.

Results

Completion of Enrollment Procedures

A total of 96 community-dwelling older adults underwent baseline assessment to determine eligibility. Of these, 63 eligible participants were randomly assigned to either the intervention group (n=31, 49%) or the control group (n=32, 51%) after completing the baseline assessment. All participants were contacted to complete follow-up measures at 4 and 12 weeks after the completion of the 4-week web-based resilience-enhancing program, regardless of their completion of previous measures. Figure 2 (also see Multimedia Appendix 8 [70]) presents the CONSORT (Consolidated Standards of Reporting Trials) diagram outlining the flow of participants throughout the study from February 2020 to December 2022.

XSL•FO
RenderX

Figure 2. Consolidated Standards of Reporting Trial (CONSORT) diagram.

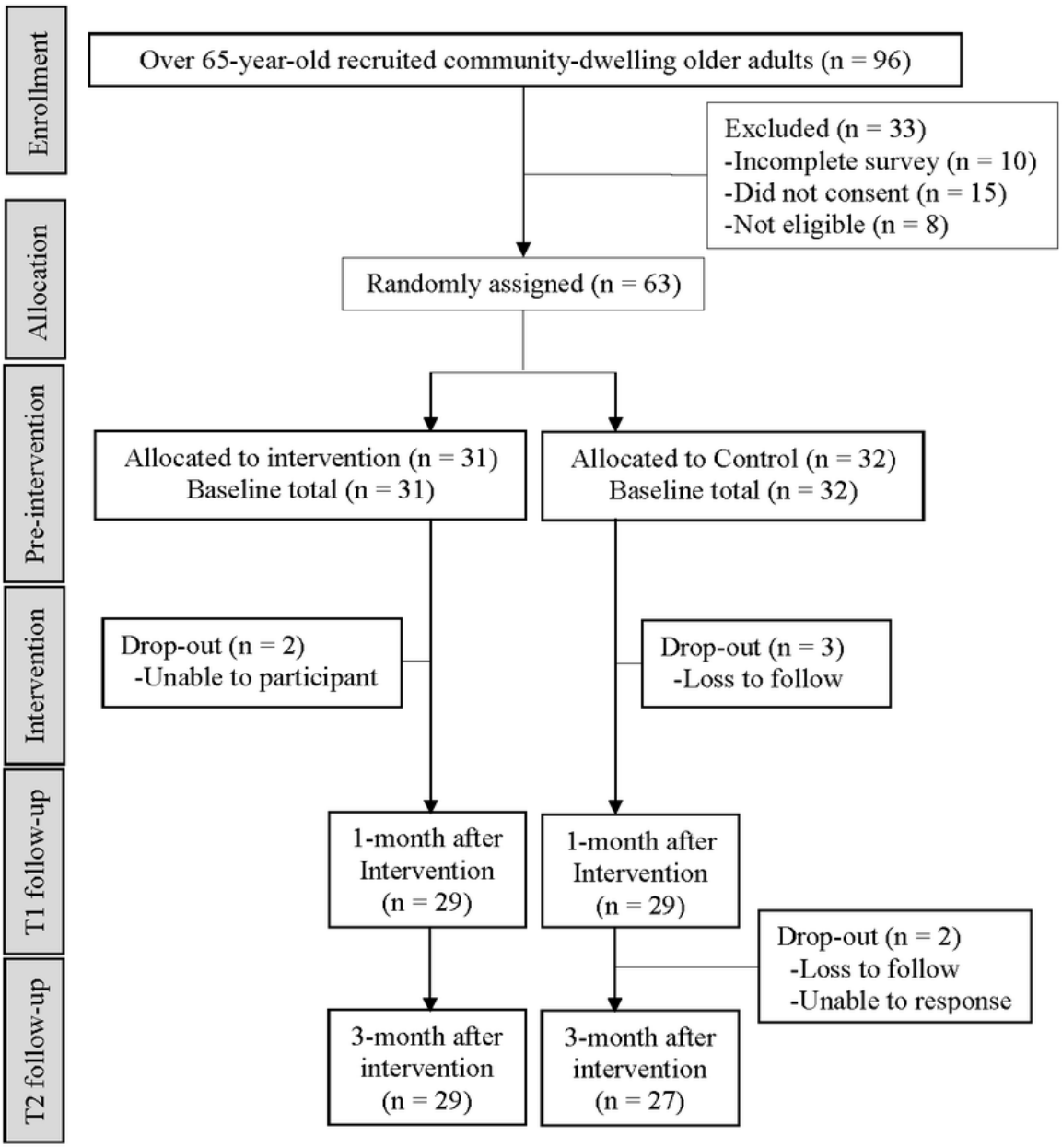

Missing Data

Of the 63 eligible participants included in the final analysis, 31 (49%) were in the intervention group and 32 (51%) were in the control group. In the intervention group, 94% (n=29) completed both the 4- and 12-week follow-up measures. In the control group, 84% (n=27) completed both follow-up measures, while 91% (n=29) completed only the 4-week follow-up measure. Three participants (9%) in the control group did not complete either the 4- or 12-week follow-up measures. The dropout rate was 6% (2/31) in the intervention group and 16% (5/32) in the control group.

Baseline Characteristics

The baseline characteristics of the participants are presented in Table 1 for each group and for the entire sample. The mean age of participants was 72.11 (SD 5.65; range 65-92) years. Most participants were female (50/63, 79%), married (43/63, 68%), and had a habit of exercise (54/63, 86%). The majority engaged in walking (exercise), with a frequency of 3-5 times per week, and sessions lasting 30-60 minutes each. Characteristics were similar between groups; however, the mean age differed significantly between the intervention (69.27 years) and control (74.86 years) groups ($t_{45.6}$=–4.53, P<.001). Table 2 presents bivariate correlations for all outcome variables at baseline for each group and the entire sample, indicating a positive association between resilience, physical activity, and well-being. Surprisingly, no relationship was found between resilience and physical activity within the control group.

XSL•FO
RenderX

Table 1. The demographic characteristics of participants.

Characteristics	Total (n=63)	Intervention (n=31)	Control (n=32)	P^a value
Number of participants randomized, n				
Age (years), mean (SD)	72.11 (5.65)	69.27 (3.08)	74.86 (6.23)	<.001
Gender, n (%)				.38
Male	13 (21)	5 (16)	8 (25)	
Female	50 (79)	26 (84)	24 (75)	
Marriage status, n (%)				.18
Single and divorced	5 (8)	4 (13)	1 (3)	
Married	43 (68)	21 (68)	22 (69)	
Widowed	15 (24)	6 (19)	9 (28)	
Educational status, n (%)				.45
Junior high school or below	22 (35)	9 (29)	13 (41)	
High school	19 (30)	12 (39)	7 (22)	
Associate's degree and above	22 (35)	10 (32)	12 (38)	
Family income per month, n (%)[b]				.94
<NTD 30,000	23 (37)	11 (35)	12 (38)	
NTD 30,000-50,000	17 (27)	9 (29)	8 (25)	
>NTD 50,000	23 (37)	11 (35)	12 (38)	
Exercise habit, n (%)				.18
No	6 (10)	2 (6)	4 (13)	
Yes				
Walking	38 (60)	20 (65)	18 (56)	
Gymnastics	7 (11)	1 (3)	6 (19)	
Swimming	4 (6)	2 (6)	2 (6)	
Other	8 (13)	6 (19)	2 (6)	
Frequency[c]				.74
0-2 times/week	9 (14)	4 (13)	5 (16)	
3-5 times/week	29 (46)	14 (45)	15 (47)	
5-7 times/week	19 (30)	11 (35)	8 (25)	
Duration[c]				.23
0-30 minutes/session	13 (21)	4 (13)	9 (28)	
30-60 minutes/session	26 (41)	14 (45)	12 (38)	
>60 minutes/session	18 (29)	11 (35)	7 (22)	

[a]*P* value between groups.

[b]NTD 1=US $0.031.

[c]A total of 57 participants had an exercise habit and were surveyed regarding the duration and frequency of their exercise habit.

XSL•FO
RenderX

Table 2. Correlation analysis of all outcome variables.

Outcomes	Total			Intervention			Control		
	Resilience	Physical activity	Well-being	Resilience	Physical activity	Well-being	Resilience	Physical activity	Well-being
Total									
Resilience									
r	1	0.21[a]	–0.81[a]	—	—	—	—	—	—
P value	__[b]	<.001	<.001	—	—	—	—	—	—
Physical activity									
r	–0.21	1	–0.25[a]	—	—	—	—	—	—
P value	<.001	—	<.001	—	—	—	—	—	—
Well-being									
r	0.81	0.25	1	—	—	—	—	—	—
P value	<.001	<.001	—	—	—	—	—	—	—
Intervention									
Resilience									
r	—	—	—	1	–0.34[a]	–0.64[a]	—	—	—
P value	—	—	—	—	<.001	<.001	—	—	—
Physical activity									
r	—	—	—	0.34	1	–0.42[a]	—	—	—
P value	—	—	—	<.001	—	<.001	—	—	—
Well-being									
r	—	—	—	0.64	0.42	1	—	—	—
P value	—	—	—	<.001	<.001	—	—	—	—
Control									
Resilience									
r	—	—	—	—	—	—	1	–0.08	–0.90[a]
P value	—	—	—	—	—	—	—	.25	<.001
Physical activity									
r	—	—	—	—	—	—	0.08	1	—
P value	—	—	—	—	—	—	.25	—	—
Well-being									
r	—	—	—	—	—	—	0.90	0.14	1
P value	—	—	—	—	—	—	<.001	.05	—

[a]The correlation is significant at a level of .05 (2-tailed).

[b]Not applicable.

In the intervention group, a majority reported positive outcomes in resilience factors: 74% (23/31) for family environment, 77% (24/31) for social integration, 81% (25/31) for positive and courageous coping strategies, 71% (22/31) for spiritual perspective, and 58% (18/31) for hope-derived meaning. Regarding risk factors, 55% (17/31) reported challenges with illness-related distress or defensive coping. Notably, more than one-third of participants reported difficulties in both illness-related distress and defensive coping, indicating potential negative impacts from chronic illness and defensive coping strategies.

Outcomes and Estimation

Resilience

These results reflect the intention-to-treat analysis, which included all participants regardless of missing data or intervention completion. The estimated marginal mean resilience scores over time are depicted in Figure 3 and Multimedia

XSL•FO
RenderX

Appendix 9. At baseline, the resilience scores were 138.84 (SD 22.51) in the intervention group and 135.78 (SD 30.09) in the control group, ranging between 84 and 174 and 53 and 174, respectively. There was no significant between-group difference observed (t_{61}=0.46, *P*=.65).

Figure 3. Estimated marginal mean score with error bars of resilience and its 5 domains over time. The red line represents the intervention group, while the blue line represents the control group. The mean scores of resilience in the intervention group were 138.84 (SD 22.51), 144.40 (SD 18.80), and 146.28 (SD 18.60), while in the control group they were 135.78 (SD 30.09), 139.29 (SD 27.61), and 140.62 (SD 29.05) at T0, T1, and T2, respectively. The mean scores of self-reliance in the intervention group were 33.29 (SD 6.28), 34.39 (SD 5.42), and 35.39 (SD 5.20), while in the control group they were 32.59 (SD 7.55), 33.69 (SD 6.80), and 34.70 (SD 6.98) at T0, T1, and T2, respectively. The mean scores of perseverance in the intervention group were 43.00 (SD 7.81), 46.01 (SD 6.72), and 46.00 (SD 7.01), while in the control group they were 42.91 (SD 9.29), 43.29 (SD 9.65), and 43.30 (SD 9.65) at T0, T1, and T2, respectively. The mean scores of equanimity in the intervention group were 33.19 (SD 5.80), 34.41 (SD 4.98), and 34.66 (SD 5.29), while in the control group they were 32.78 (SD 8.08), 33.36 (SD 7.81), and 33.42 (SD 7.92) at T0, T1, and T2, respectively. The mean scores of meaningful in the intervention group were 17.52 (SD 2.93), 18.00 (SD 2.32), and 18.01 (SD 2.63), while in the control group they were 16.22 (SD 4.48), 17.10 (SD 3.50), and 17.03 (SD 3.66) at T0, T1, and T2, respectively. The mean scores of existential aloneness in the intervention group were 11.84 (SD 2.40), 11.82 (SD 1.70), and 12.35 (SD 1.52) while in the control group they were 11.28 (SD 3.10), 11.94 (SD 2.27), and 11.85 (SD 2.53) at T0, T1, and T2, respectively.

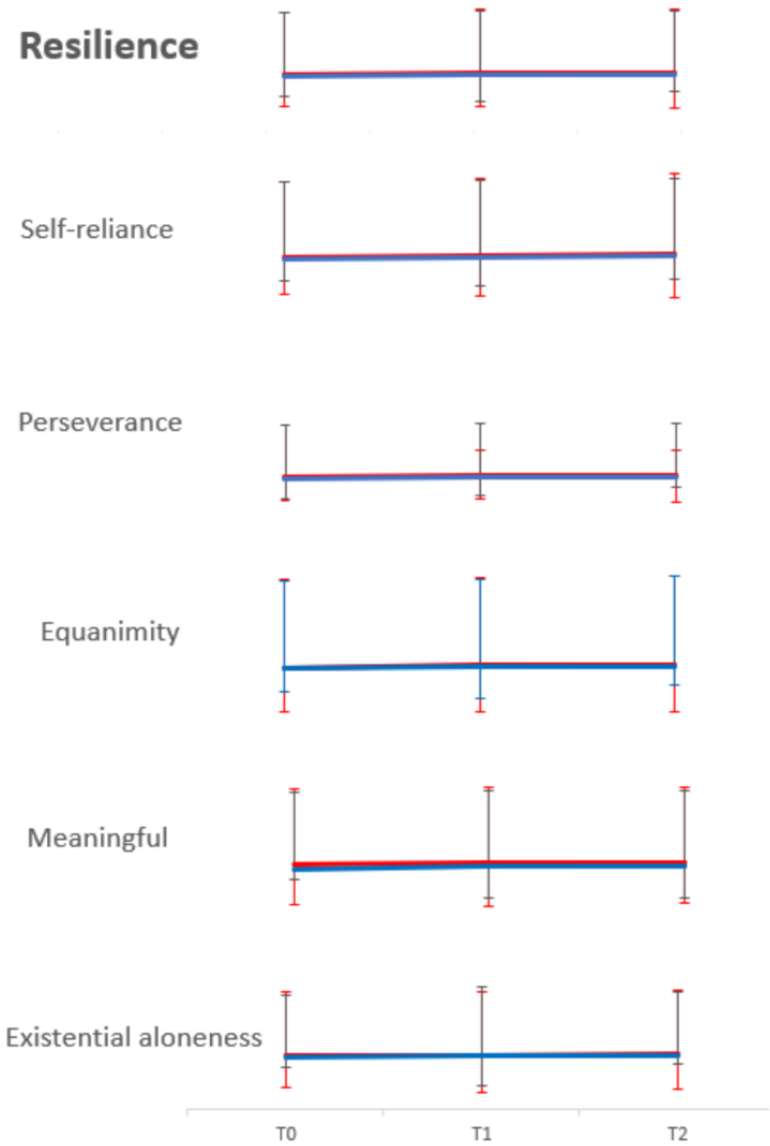

In the group-by-time interaction, there was no significant between-group difference observed over time in resilience (SE 7.49, 95% CI –10.74 to 18.61, *P*=.60), including across the 5 domains of self-reliance (SE 1.92, 95% CI –2.15 to 5.38, *P*=.40), perseverance (SE 2.62, 95% CI –4.86 to 5.40, *P*=.92), equanimity (SE 2.08, 95% CI –3.30 to 4.86, *P*=.71), meaningfulness (SE 0.92, 95% CI –1.53 to 2.08, *P*=.76), and existential aloneness (SE 0.63, 95% CI –1.22 to 1.24, *P*=.99), after controlling for age and marital status differences (Table 3).

XSL•FO
RenderX

Table 3. Mixed effects of the multilevel hierarchical model on resilience.

Outcome	Estimated marginal means	SE	95% CI	*P* value
Resilience				
Intervention	3.94	7.49	–10.74 to 18.61	.60
Group × T1[a]	0.91	4.78	–8.46 to 10.28	.85
Group × T2[b]	2.77	6.04	–9.06 to 14.60	.65
Self-reliance				
Intervention	1.61	1.92	–2.15 to 5.38	.40
Group × T1[a]	0.07	1.23	–2.34 to 2.48	.96
Group × T2[b]	–0.18	1.49	–3.11 to 2.74	.90
Perseverance				
Intervention	0.27	2.62	–4.86 to 5.40	.92
Group × T1[a]	–0.07	1.86	–3.72 to 3.58	.97
Group × T2[b]	2.60	2.09	–1.50 to 6.70	.21
Equanimity				
Intervention	0.78	2.08	–3.30 to 4.86	.71
Group × T1[a]	0.14	1.68	–3.17 to 3.45	.93
Group × T2[b]	0.71	1.82	–2.85 to 4.28	.70
Meaningful				
Intervention	0.28	0.92	–1.53 to 2.08	.76
Group × T1[a]	0.15	0.62	–1.06 to 1.36	.80
Group × T2[b]	–0.29	0.87	–1.99 to 1.42	.74
Existential aloneness				
Intervention	0.01	0.63	–1.22 to 1.24	.99
Group × T1[a]	0.62	0.54	–0.44 to 1.68	.25
Group × T2[b]	–0.08	0.70	–1.45 to 1.30	.91

[a]4 weeks after completing the resilience program

[b]12 weeks after completing the resilience program.

For group comparisons across the 3 time points, the results indicated no statistical significance in resilience ($F_{1,61}$=0.49, *P*=.49), including across the 5 domains of self-reliance ($F_{1,61}$=0.94, *P*=.34), perseverance ($F_{1,61}$=0.29, *P*=.59), equanimity ($F_{1,61}$=0.71, *P*=.40), meaningfulness ($F_{1,61}$=0.03, *P*=.87), and existential aloneness ($F_{1,61}$=0.07, *P*=.80).

Physical Activity

The estimated marginal mean score of physical activity over time is shown in Figure 4 and Multimedia Appendix 9. At baseline, the scores in the intervention and control groups were 162.16 (SD 76.25) and 111.75 (SD 60.08), respectively, ranging between 46 and 369 and 0 and 278, respectively. There was a significant between-group difference (t_{61}=2.92, *P*=.005).

XSL•FO
RenderX

Figure 4. Estimated marginal mean score with error bars of physical activity and its 3 domains with weight occupational activity over time. The red line represents the intervention group, while the blue line represents the control group. The mean scores of physical activity in the intervention group were 162.16 (SD 76.25), 166.16 (SD 70.66), and 150.01 (SD 62.53), while in the control group they were 111.75 (SD 60.08), 127.68 (SD 88.59), and 122.75 (SD 68.85) at T0, T1, and T2, respectively. The mean scores of recreational activity in the intervention group were 50.06 (SD 44.10), 58.51 (SD 46.39), and 49.46 (SD 41.25), while in the control group they were 31.22 (SD 32.43), 46.63 (SD 50.85), and 41.99 (SD 41.28) at T0, T1, and T2, respectively. The mean scores of household activity in the intervention group were 86.25 (SD 35.63), 85.14 (SD 24.17), and 80.66 (SD 33.09), while in the control group they were 66.00 (SD 38.73), 66.98 (SD 36.87), and 74.19 (SD 37.31) at T0, T1, and T2, respectively. The mean scores of weighted occupational activity in the intervention group were 25.84 (SD 30.76), 22.52 (SD 24.17), and 19.89 (SD 24.61), while in the control group they were 14.55 (SD 29.66), 14.07 (SD 24.71), and 6.57 (SD 10.71) at T0, T1, and T2, respectively.

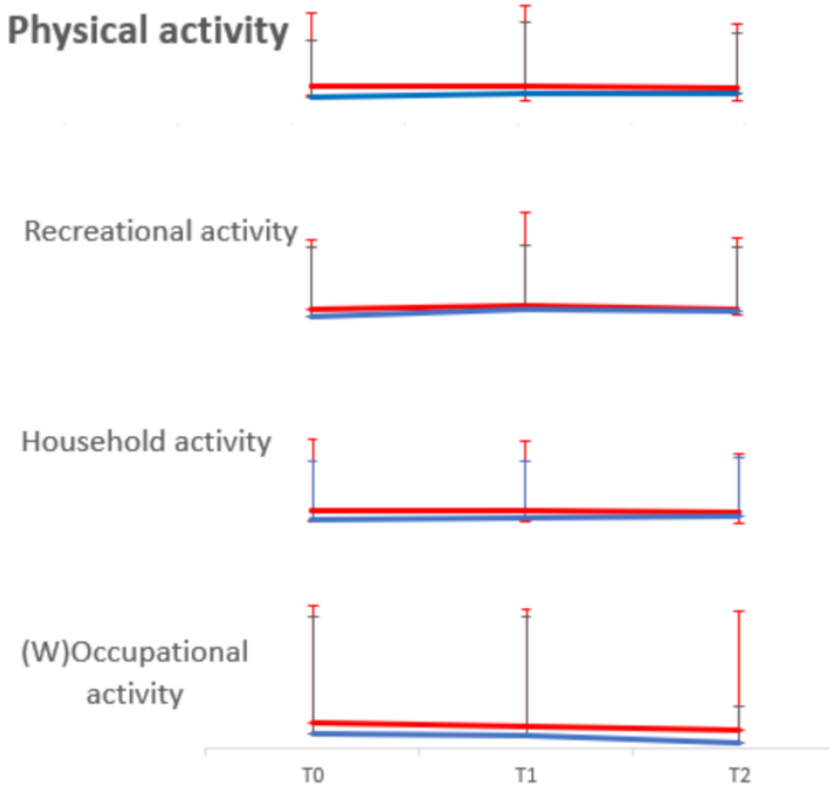

In the group × time interaction, there was no significant between-group difference over time in the change of physical activity (SE 15.18, 95% CI –24.74 to 34.74, *P*=.74), including 3 domains: recreational activity (SE 8.58, 95% CI –15.77 to 17.87, *P*=.90), household activity (SE 8.23, 95% CI –4.92 to 27.35, *P*=.17), and occupational activity (SE 5.56, 95% CI –18.13 to 3.66, *P*=.19). This analysis controlled for age difference, marital status, exercise habits, and COVID-19 monthly trends (Table 4 and Multimedia Appendix 7).

XSL•FO
RenderX

Table 4. Mixed effects of multilevel hierarchical model on the change of physical activity.

Outcome	Estimated marginal means	SE	95% CI	*P* value
Physical activity				
Intervention	5.01	15.18	–24.74 to 34.76	.74
Group × T0[a] to T1[b] duration	8.73	20.06	–30.59 to 40.05	.66
Recreational activity				
Intervention	1.05	8.58	–15.77 to 17.87	.90
Group × T0[a] to T1[b] duration	7.81	13.84	–19.31 to 34.93	.57
Household activity				
Intervention	11.21	8.23	–4.92 to 27.35	.17
Group × T0[a] to T1[b] duration	–9.93	10.91	–31.33 to 11.46	.36
Occupational activity				
Intervention	–7.23	5.56	–18.13 to 3.66	.19
Group × T0[a] to T1[b] duration	10.01	7.23	–4.17 to 24.18	.17

[a]Baseline.

[b]4 weeks after completing the resilience program.

For group comparisons at 3 time points, the results showed no statistical significance in physical activity ($F_{1,61}$=1.71, *P*=.20), including 3 domains: recreational activity ($F_{1,61}$=0.40, *P*=.53), household activity ($F_{1,61}$=2.14, *P*=.15), and occupational activity ($F_{1,61}$=0.30, *P*=.59), after controlling for baseline differences.

Well-Being

The estimated marginal mean score of well-being over time is shown in Figure 5 and Multimedia Appendix 9. At baseline, the scores in the intervention and control groups were 92.29 (SD 12.69) and 92.28 (SD 17.15), respectively, ranging between 71 and 119 and 46 and 120, respectively. There was no statistically significant difference between the groups (t_{61}=0.002, *P*=.99).

XSL•FO
RenderX

Figure 5. Estimated marginal mean score with error bars of well-being and its 4 domains over time. The red line represents the intervention group, while the blue line represents the control group. The mean scores of well-being in the intervention group were 92.29 (SD 12.69), 95.65 (SD 11.09), and 94.23 (SD 10.73), while in the control group they were 92.28 (SD 17.15), 96.29 (SD 17.38), and 95.22 (SD 15.63) at T0, T1, and T2, respectively. The mean scores of life satisfaction in the intervention group were 27.13 (SD 3.90), 27.84 (SD 3.61), and 28.06 (SD 3.82), while in the control group they were 27.06 (SD 5.54), 28.54 (SD 5.32), and 28.36 (SD 4.51) at T0, T1, and T2, respectively. The mean scores of interpersonal relationship in the intervention group were 22.74 (SD 3.54), 23.83 (SD 3.19), and 23.88 (SD 3.35), while in the control group they were 23.38 (SD 5.13), 24.88 (SD 4.81), and 24.22 (SD 4.16) at T0, T1, and T2, respectively. The mean scores of self-assurance in the intervention group were 23.94 (SD 3.23), 23.62 (SD 2.55), and 24.14 (SD 2.87), while in the control group they were 22.88 (SD 4.11), 23.38 (SD 4.15), and 23.53 (SD 4.57) at T0, T1, and T2, respectively. The mean scores of physical and moral integrity in the intervention group were 18.48 (SD 3.19), 19.82 (SD 2.78), and 18.81 (SD 2.87), while in the control group they were 18.97 (SD 4.48), 19.60 (SD 4.69), and 19.07 (SD 4.27) at T0, T1, and T2, respectively.

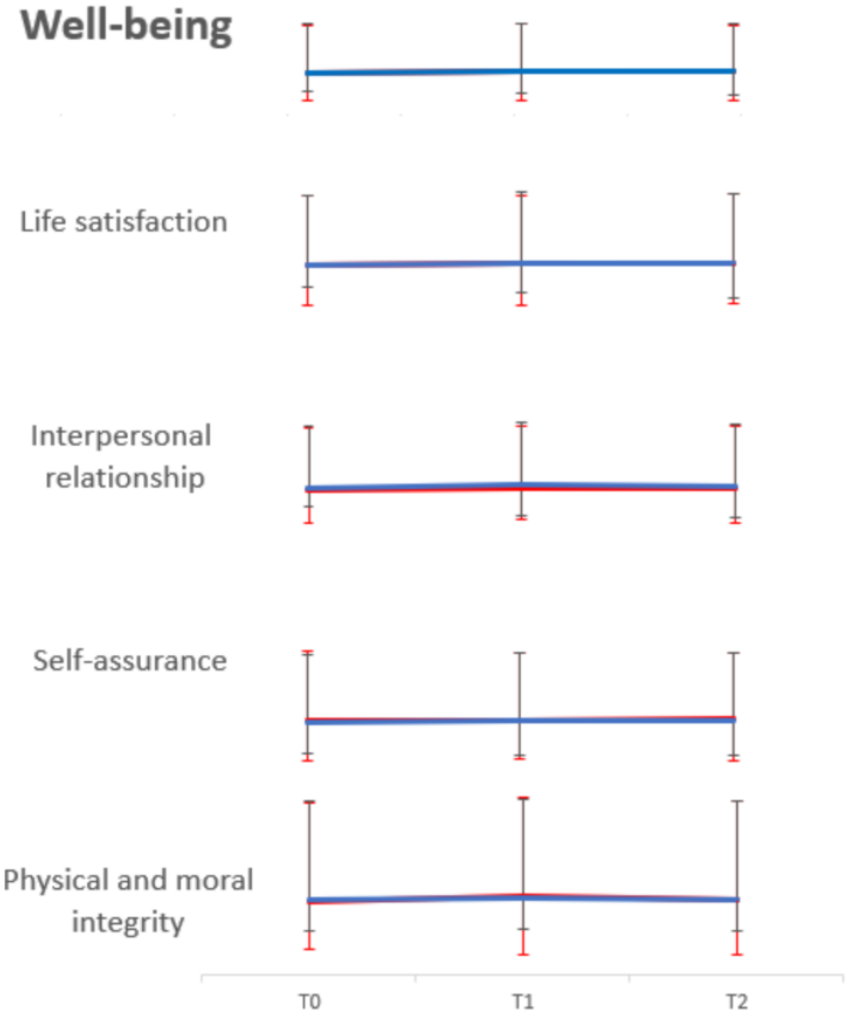

In the group × time interaction, there was no significant between-group difference over time in well-being (SE 3.74, 95% CI –2.68 to 11.98, *P*=.21), including the 4 domains of life satisfaction (SE 1.23, 95% CI –1.11 to 3.70, *P*=.29), interpersonal relationships (SE 1.06, 95% CI –0.84 to 3.30, *P*=.24), self-assurance (SE 0.97, 95% CI –0.85 to 2.94, *P*=.28), and physical and moral integrity (SE 1.07, 95% CI –0.95 to 3.25, *P*=.28), after controlling for age difference and marital status (Table 5).

XSL•FO
RenderX

Table 5. Mixed effects of multilevel hierarchical model on well-being.

Outcome	Estimated marginal means	SE	95% CI	*P* value
Well-being				
Intervention	4.65	3.74	–2.68 to 11.98	.21
Group × T1[a]	–0.30	2.07	–4.36 to 3.77	.89
Group × T2[b]	–0.81	2.89	–6.49 to 4.87	.78
Life satisfaction				
Intervention	1.30	1.23	–1.11 to 3.70	.29
Group × T1[a]	0.20	0.85	–1.48 to 1.87	.82
Group × T2[b]	–0.33	1.06	–2.41 to 1.74	.75
Interpersonal relationship				
Intervention	1.23	1.06	–0.84 to 3.30	.24
Group × T1[a]	0.78	0.64	–0.48 to 2.05	.22
Group × T2[b]	0.34	0.88	–1.38 to 2.01	.70
Self-assurance				
Intervention	1.05	0.97	–0.85 to 2.94	.28
Group × T1[a]	–0.57	0.68	–1.92 to 0.77	.40
Group × T2[b]	–1.03	0.83	–2.65 to 0.60	.22
Physical and moral integrity				
Intervention	1.14	1.07	–0.95 to 3.25	.28
Group × T1[a]	–0.70	0.75	–2.19 to 0.78	.35
Group × T2[b]	0.21	0.77	–1.30 to 1.71	.79

[a]4 weeks after completing the resilience program

[b]12 weeks after completing the resilience program.

For group comparisons at 3 time points, the results showed no statistical significance in well-being ($F_{1,61}$=1.84, P=.18), including the 4 domains of life satisfaction ($F_{1,61}$=1.79, P=.19), interpersonal relationships ($F_{1,61}$=2.88, P=.10), self-assurance (T0: $F_{1,61}$=0.24, P=.63; T1: $F_{1,61}$=0.28, P=.60; T2: $F_{1,61}$=3.43, P=.07), and physical and moral integrity ($F_{1,61}$=1.28, P=.26).

Program Feedback

Program feedback is summarized in Table 6. Out of the 31 participants randomized to the intervention group, 29 (94%) provided feedback on the program. Among them, 24 (83%) completed the entire program. For those who did not complete it, 2 of 29 (7%) cited illness or impaired function as the main barrier. Encouragingly, all participants who provided feedback reported having the most confidence to use the skills in control belief outside of the program. Additionally, 26 out of 29 (90%) expressed confidence in using coping skills, while 24 out of 29 (83%) felt confident in applying manageability skills.

XSL•FO
RenderX

Table 6. Program completion, barriers to engagement, and feedback.

Program feedback		Participants
Percentage of program completed, n (%)		
100%		24 (83)
75%		5 (17)
Reasons for noncompletion, n (%)		
Lack of time to complete the program		1 (3)
Illness impaired function to complete the program		2 (7)
Competing demands, such as work		1 (3)
Forgot to complete the program		1 (3)
Have confidence in using program strategies outside the program, n (%)		
In coping		
	Most confidence and above	26 (90)
	Some confidence and below	3 (10)
In control belief		
	Most confidence and above	29 (100)
	Some confidence and below	0 (0)
In manageability		
	Most confidence and above	24 (83)
	Some confidence and below	5 (17)

During the interviews, older participants expressed appreciation for the opportunity to participate in this facilitated program. They found the mentorships invaluable for practicing resilience skills in the face of adversity, particularly in overcoming challenges posed by the COVID-19 pandemic. Participants highlighted their learning from personal and others' experiences. The program implementation was noted as the most significant benefit for participants.

> *This course was very good and I learnt a lot from either role-play or experience shared by others. Now I learnt the ways from others' experience and the practitioner taught to help one of my relative get out of the lost after her husband died. She committed suicide after lost her husband. I didn't know how to help her at that time, but, I do now.* [Patient #9, female]

Moreover, the flexibility of the web-based digital intervention format allowed participants to attend sessions without direct contact and interact online. YouTube and LINE emerged as the most popular and familiar social media platforms. Older participants found it relatively easy to watch microfilms and participate in group meetings for program completion, even amid quarantine or symptoms of chronic disease.

Furthermore, several participants referenced prior experiences of feeling powerless when trying to support a grieving friend who had lost a spouse. They expressed frustration when their efforts to provide companionship were met with negative emotions directed inward by their friend. Additionally, those with depressive tendencies recounted instances of using defensive coping mechanisms that led them to withdraw from family and friends. One older participant shared the sentiment that "time heals all wounds," while another mentioned considering becoming a monk for spiritual support after their spouse passed away.

> *When my husband who has been in a vegetative state for many years passed away, I lost my focus in my life but felt relieved at the same time. It took me a year and half to overcome, and I had been thought about to becoming a monk. My children and friends supported me a lot during that period. I truly didn't want to bother them too much at that moment, but I appreciated their company to heal me. Now I know how to help myself recover sooner, and know the ways to help others based on my own experience and the skills learning from this program.* [Patient #2, female]

Another benefit of this facilitated program related to disease-related events is practicing mindset transformation. To assist individuals experiencing depressive states, we practiced shifting focus from physical aspects to resetting goals aimed at quickly regaining health by identifying what is most important:

> *Think about your family and children, which is the most important things for you, work hard to restore your health, change your mind toward a positive attitude, and find specific solutions to regain health. Most people discouraged in rehab period. You can start with a goal you want to achieve in the next weeks to build your confidence.* [Patient #5, male]

In addition, the program raised awareness among participants about the importance of vigilant self-care.

XSL•FO
RenderX

This program was considered ideal as it provided community-dwelling older adults with opportunities to practice and develop resilience skills, preparing them to face adversity effectively. Enhancing resilience is crucial for preventing and mitigating the severity of mental and physical health issues, especially among the geriatric population:

> *This course reminded me the ways to build resilience and practiced the skills constantly. It now was in my mind. When encountering similar problems in the future, I can use them without hesitation.* [Patient #26, female]

Several older participants suggested improvements such as delivering the program on devices with larger screens that are familiar to older adults, such as tablets or computers. They also recommended providing clear instructions on how to join the correct group and participate in sessions with shared screens, individually guiding them step by step before each session starts. Older participants who initially struggled with operating the device were assessed by the practitioner, observer, and their family. One participant who suggested computer-based delivery felt it would improve accessibility to the program content, especially if delivered on a mobile device:

> *My eyesight isn't good. It hard for me to read the sentences in the scenario via mobile device, so if I was able to assess it through a big screen that would be easier for me.* [Patient #14, female]

Discussion

Principal Findings

This study investigated the efficacy of a web-based resilience-enhancing program aimed at improving resilience, physical activity, and well-being in community-dwelling older adults. Participants in the resilience-enhancing program were compared with a control group, with postintervention assessments conducted at 4 and 12 weeks after program completion or allocation. No statistically significant between-group differences were observed over time in resilience, physical activity, and well-being after controlling for baseline differences. Interestingly, despite a small number of participants withdrawing from the study initially, all intervention group participants completed the program and both follow-up assessments, indicating high study participation and adherence rates. Participant feedback underscored the significance of enhancing resilience to cope with challenges in later life. All intervention group participants valued the resilience-enhancing program, noting its benefits in improving resilience and well-being. The findings also highlighted both the opportunities and challenges of internet-based programs for the geriatric population. Participant feedback emphasized the necessity for targeted interventions tailored to promote healthy aging.

Comparison With Prior Work

Our findings underscore the acceptability and feasibility of implementing a web-based resilience-enhancing program for community-dwelling older adults, a need identified in previous literature [7-13]. This trial was initially conducted in North Taiwan at the onset of the COVID-19 pandemic, necessitating a shift from in-person to web-based strategies for COVID-19 control. Subsequently, we conducted the study during the peak of the pandemic, when substantial damage to the physical, mental, emotional, and social dimensions of older adults was evident [31,39-42]. The geriatric population showed significant demands for recovering their health afterward [8], highlighting the potential need and responsibility for public service interventions. Despite these challenges, we successfully recruited participants, implemented the web-based resilience-enhancing program as planned, and collected data at the designated time points. Therefore, our study has provided initial evidence that implementing web-based resilience-enhancing programs with similar design features for the geriatric population is feasible. These findings pave the way for further research to design appropriate web-based resilience programs aimed at improving health and well-being in older adults.

The web-based resilience-enhancing program has demonstrated acceptability among the geriatric population in community settings, as evidenced by the low dropout rate among participants. In our study, the dropout rate in the intervention group was 6% (2/31), consistent with rates reported in prior studies utilizing virtual interventions [71], underscoring its high level of acceptance [72]. This was further confirmed by program feedback from participants in the intervention group, demonstrating high engagement and satisfaction among community-dwelling older adults. These findings underscore the genuine demand for thoughtfully evaluated web-based resilience-enhancing programs for implementation across community settings. Additionally, the dropout rate in the control group was 16% (5/32), which was more than double that of the intervention group. However, the dropout rate in our study was relatively low compared with other RCTs [73]. The disparity in dropout rates between the intervention and control groups may be attributed to the older age and poorer health status of participants in the control group [73]. Reasons for dropout included "loss to follow-up" and "inability to respond," as indicated in the CONSORT diagram, during the COVID-19 pandemic.

Encouragingly, feedback from participants in the intervention group indicated satisfaction with the 2 distinct components of the resilience-enhancing program. The web-based facilitated sessions and mentorship were particularly well received. For example, practitioners offered individual guidance on coping strategies, belief control, and managing personal bereavement. This was especially meaningful for widowed participants who shared their experiences of spousal loss, while married participants were invited to discuss related concerns during the sessions. The effectiveness of mentorship in enhancing resilience has been supported by prior research focused on resilience interventions for older adults [74,75]. These findings underscore mentorship support as a pivotal element in bolstering resilience among the geriatric population. Future studies could explore the specific impacts of individual versus group mentorship sessions, considering their potential flexibility and adaptation to the diverse physical conditions of older adults. Interviews conducted with community-dwelling older adults in the

XSL•FO
RenderX

participating trust highlighted a demand for resilience training tailored to this population. Participants identified the necessity for supportive measures to mitigate various stressors and promote healthy aging [39]. These insights suggest that the design features, content, and format of the resilience-enhancing program are suitable and well-received within this demographic, indicating potential for replication and adaptation in other populations. Furthermore, the findings of this study regarding design, engagement, and acceptability represent a significant extension of prior work. These results are promising, especially for low-cost programs that are relatively accessible to this target group.

There is limited research examining the impact of similar interventions, making direct comparisons of long-term effects in the geriatric population challenging. Despite hypothesizing that resilience, physical activity, and well-being would differ between groups after an intervention, our study did not find evidence to verify these mechanisms through the RIM. The program's ineffectiveness in increasing resilience could potentially be attributed to the necessity for participants to encounter triggering events, such as new episodes of functional deficits, the death of a spouse, retirement, or financial exploitation, to demonstrate resilience despite acquiring coping strategies during the intervention. However, we did not collect data on such triggering events. Another potential reason could be related to certain risk factors affecting older adults' learning, such as insufficient time to practice learned material, cognitive declines, and limitations in information-processing capabilities [76]. For future research aiming to quantify confidence and satisfaction after program completion in the intervention group, it is suggested to implement strategies such as pre- and postsurveys to assess the application of learned strategies.

In addition, previous research has shown that the COVID-19 pandemic led to a general decrease in physical activity during lockdown [37]. Our study was conducted during the height of the pandemic, and it revealed varying levels of physical activity among participants, particularly in the control group. This underscores the need to consider the potential impact of the pandemic as a contributing factor to the observed lack of intervention effect on physical activity. Furthermore, we observed a lack of relationship between resilience and physical activity within the control group, which contrasts with findings from prior studies [32,33]. This discrepancy may be attributed to the diverse physical activity levels among our control group participants, who engaged in both active and sedentary lifestyles and utilized different aspects of resilience, as noted by Wermelinger et al [33]. Moreover, the substitution of the in-person approach with a web-based format may have affected the intervention's effectiveness on physical activity outcomes. Future research could benefit from closely tracking changes in physical activity in response to web-based resilience-enhancing practices.

Strengths and Limitations of the Study

To our knowledge, this study represents the first implementation of a web-based, role-play, and talk-in-interaction resilience-enhancing program for community-dwelling older adults during the COVID-19 pandemic, underscoring the benefits of such interventions for promoting healthy aging [10-12]. Furthermore, the high participation and adherence rates, along with positive qualitative feedback received, highlight the successful engagement of participants and provide valuable insights into essential elements for designing and implementing digital programs tailored to community-dwelling older adults. This study lays an important foundation and offers clear guidance for optimizing web-based programs and research designs in the future, particularly in terms of ensuring the suitability of equipment for the geriatric population.

One limitation of our study was the control group design, which received standard care without social components such as virtual meetings or educational sessions with the same frequency as the intervention group. This discrepancy may have contributed to the double dropout rate observed in the control group compared with the intervention group. In addition, the study was conducted only in North Taiwan, where participants who were willing to join may have had higher social involvement [75]. The study also had a relatively small sample size and an uneven distribution of age and physical activity at baseline between groups, potentially impacting the generalizability of the findings. Attrition bias was introduced by a higher attrition rate than previously reported for older adults [12]. Future studies should aim to include participants from a broader geographical range and diverse backgrounds to enhance the generalizability of the findings.

Moreover, the outcome measures in our study were self-assessed, which could be subject to participant bias. We did not assess objective measures of health status, such as body mass index, fasting lipid profile, blood pressure, heart rate, serum creatinine, presence of other comorbid conditions, number of prescribed medications, and surgery history. Therefore, we were unable to comment on whether these variables may have influenced the intervention outcomes. Finally, conducting a study during the COVID-19 pandemic may have influenced the processes and findings of this study. Initial recruitment progress was not as anticipated, and vulnerable older adults were disproportionately affected in health, social, and economic dimensions during this global public health crisis [8,30]. Moreover, the intervention was shifted to a web-based method as part of COVID-19 control strategies, which may have influenced participant responses. These factors should be considered when implementing similar resources across health and social care settings in the postpandemic era.

Conclusions

This RCT provides compelling evidence regarding the study design, engagement, and acceptability of a web-based resilience-enhancing program for community-dwelling older adults. Furthermore, the study underscores the significance and necessity of tailored resilience-enhancing programs for older adults living independently, who encounter challenges in late adulthood and would benefit from accessible forms of structured support for healthy aging. Additionally, participants highlighted the importance of mentorship components in the web-based resilience-enhancing program, which played a pivotal role in its successful delivery. This suggests that mentorship could be an area of future research focus and practical application. Further

XSL•FO
RenderX

studies should build on these findings by optimizing program designs to ascertain whether such interventions can lead to meaningful improvements in pursuing healthy aging among this target population.

Acknowledgments

This research was funded by the Ministry of Science and Technology Republic of China (grant number MOST 109-2314-B-010-059) and the National Science and Technology Council Republic of China (grant number NSTC 110-2314-B-A49A-538). The funders had no further interest or involvement in the design, methods, subject recruitment, data collection, or preparation of the manuscript. The authors thank the technology support of microfilm production by Chen SH, Senior Vice President of National Yang Ming Chiao Tung University, and his team. Generative artificial intelligence was not used in any portion of the manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YCW was responsible for study design, execution, all statistical analyses, manuscript drafting, and critical discussions. SFS conducted participant interviews and provided data for the “Program Feedback” section. HHT contributed to study design, manuscript drafting, and critical discussions. LKC also participated in study design and critical discussions. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Resilience practitioner training.
[DOCX File , 12 KB - jmir_v26i1e53450_app1.docx]

Multimedia Appendix 2
Preintervention survey—resilience factors.
[PDF File (Adobe PDF File), 85 KB - jmir_v26i1e53450_app2.pdf]

Multimedia Appendix 3
Links to the 2 microfilms on YouTube.
[DOCX File , 13 KB - jmir_v26i1e53450_app3.docx]

Multimedia Appendix 4
Quiz on the 2 microfilms.
[PDF File (Adobe PDF File), 58 KB - jmir_v26i1e53450_app4.pdf]

Multimedia Appendix 5
Postintervention survey.
[PDF File (Adobe PDF File), 45 KB - jmir_v26i1e53450_app5.pdf]

Multimedia Appendix 6
National Health Insurance regulation documents of education sheets.
[PDF File (Adobe PDF File), 10437 KB - jmir_v26i1e53450_app6.pdf]

Multimedia Appendix 7
2020 Taiwan COVID-19 monthly trend.
[PNG File , 173 KB - jmir_v26i1e53450_app7.png]

Multimedia Appendix 8
CONSORT-eHEALTH (V 1.6.1) checklist.
[PDF File (Adobe PDF File), 2700 KB - jmir_v26i1e53450_app8.pdf]

XSL•FO
RenderX

Multimedia Appendix 9
Results of the estimated marginal mean score of resilience, physical activity, and well-being over time.
[PDF File (Adobe PDF File), 72 KB - jmir_v26i1e53450_app9.pdf]

References

1. World Health Organization (WHO). Healthy aging and functional ability. WHO. 2020. URL: https://www.who.int/news-room/questions-and-answers/item/healthy-ageing-and-functional-ability [accessed 2023-11-23]
2. Singer B, Ryff C. New Horizons in Health: An Integrative Approach. Washington (DC): National Academies Press (US); Feb 07, 2001.
3. Pruchno R, Carr D. Successful aging 2.0: resilience and beyond. J Gerontol B Psychol Sci Soc Sci 2017 Mar 01;72(2):201-203. [doi: 10.1093/geronb/gbw214] [Medline: 28364520]
4. Merchant RA, Aprahamian I, Woo J, Vellas B, Morley JE. Editorial: Resilience and successful aging. J Nutr Health Aging 2022 Jun 21;26(7):652-656 [FREE Full text] [doi: 10.1007/s12603-022-1818-4] [Medline: 35842754]
5. MacLeod S, Musich S, Hawkins K, Alsgaard K, Wicker ER. The impact of resilience among older adults. Geriatr Nurs 2016 Jul;37(4):266-272 [FREE Full text] [doi: 10.1016/j.gerinurse.2016.02.014] [Medline: 27055911]
6. Harms P, Brady L, Wood D, Silard A. Resilience and well-being. In: Diener E, Oishi S, Tay L, editors. Handbook of Well-Being. Salt Lake City, UT: DEF Publishers; 2018:604-615.
7. Wermelinger Ávila MP, Lucchetti ALG, Lucchetti G. Association between depression and resilience in older adults: a systematic review and meta-analysis. Int J Geriatr Psychiatry 2017 Mar 02;32(3):237-246. [doi: 10.1002/gps.4619] [Medline: 27805730]
8. Chen L. Older adults and COVID-19 pandemic: Resilience matters. Archives of Gerontology and Geriatrics 2020 Jul;89:104124. [doi: 10.1016/j.archger.2020.104124]
9. Tecson KM, Wilkinson LR, Smith B, Ko JM. Association between psychological resilience and subjective well-being in older adults living with chronic illness. Proc (Bayl Univ Med Cent) 2019 Oct;32(4):520-524 [FREE Full text] [doi: 10.1080/08998280.2019.1625660] [Medline: 31656409]
10. Joyce S, Shand F, Tighe J, Laurent SJ, Bryant RA, Harvey SB. Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open 2018 Jun 14;8(6):e017858 [FREE Full text] [doi: 10.1136/bmjopen-2017-017858] [Medline: 29903782]
11. Traeger L, Styklunas G, Park E, Lee M, Fricchione G, Park E. Promoting resilience and flourishing among older adult residents in community living: a feasibility study. Gerontologist 2022 Nov 30;62(10):1507-1518. [doi: 10.1093/geront/gnac031] [Medline: 35235940]
12. Wu Y, Shen S, Lee S, Chen L, Tung H. The effectiveness of mind-body approaches for enhancing resilience in older adults: a systematic review and network meta-analysis. Arch Gerontol Geriatr 2023 Jun;109:104949. [doi: 10.1016/j.archger.2023.104949] [Medline: 36796181]
13. van Agteren J, Iasiello M, Lo L. Improving the wellbeing and resilience of health services staff via psychological skills training. BMC Res Notes 2018 Dec 22;11(1):924-924 [FREE Full text] [doi: 10.1186/s13104-018-4034-x] [Medline: 30577884]
14. Henshall C, Davey Z, Srikesavan C, Hart L, Butcher D, Cipriani A. Implementation of a web-based resilience enhancement training for nurses: pilot randomized controlled trial. J Med Internet Res 2023 Feb 14;25:e43771 [FREE Full text] [doi: 10.2196/43771] [Medline: 36787181]
15. Chen Y. Enhancing the elderly's well-being: the effect of re-experiencing gratitude intervention. J Educ Psychol 2009;40(3):69-103 [FREE Full text]
16. Hsu CW. Effectiveness of the Mental Health Promotion Activities Program on Happiness and Depressive Mood for Institutionalized Elderly People in Taiwan. Thesis. Chung Shan Medical University. 2015. URL: https://www.airitilibrary.com/Article/Detail/10283684-201909-201910160012-201910160012-142-147 [accessed 2024-07-23]
17. Chen JP. The Effect of Reminiscence Group Activity Intervention on Happiness Among Long-Term Care Elderly. Thesis. University of Taipei. 2019. URL: https://hdl.handle.net/11296/j2jhw2 [accessed 2024-07-23]
18. Kelly KP, Hooke MC, Ruccione K, Landier W, Haase J. Children's Oncology Group nursing research framework. Semin Oncol Nurs 2014 Feb;30(1):17-25. [doi: 10.1016/j.soncn.2013.12.004] [Medline: 24559776]
19. Wu L, Chin C, Chen C, Lai F, Tseng Y. Development and validation of the paediatric cancer coping scale. J Adv Nurs 2011 May 24;67(5):1142-1151. [doi: 10.1111/j.1365-2648.2010.05567.x] [Medline: 21261693]
20. Haase JE, Kintner EK, Monahan PO, Robb SL. The resilience in illness model, part 1: exploratory evaluation in adolescents and young adults with cancer. Cancer Nurs 2014;37(3):E1-12 [FREE Full text] [doi: 10.1097/NCC.0b013e31828941bb] [Medline: 23519038]
21. Haase JE, Kintner EK, Robb SL, Stump TE, Monahan PO, Phillips C, et al. The Resilience in Illness Model part 2: confirmatory evaluation in adolescents and young adults with cancer. Cancer Nurs 2017;40(6):454-463 [FREE Full text] [doi: 10.1097/NCC.0000000000000450] [Medline: 27984241]
22. Pearlin LI, Schooler C. The structure of coping. J Health Soc Behav 1978 Mar;19(1):2-21. [doi: 10.2307/2136319] [Medline: 649936]

XSL•FO
RenderX

23. Skinner EA. A guide to constructs of control. Journal of Personality and Social Psychology 1996;71(3):549-570. [doi: 10.1037/0022-3514.71.3.549] [Medline: 8831161]
24. Antonovsky A. The salutogenic model as a theory to guide health promotion. Health Promot Int 1996;11(1):11-18. [doi: 10.1093/heapro/11.1.11]
25. Windle G. What is resilience? A review and concept analysis. Rev Clin Gerontol 2010 Dec 21;21(2):152-169. [doi: 10.1017/s0959259810000420]
26. American College of Sports Medicine, Chodzko-Zajko WJ, Proctor DN, Fiatarone Singh MA, Minson CT, Nigg CR, et al. American College of Sports Medicine position stand. Exercise and physical activity for older adults. Med Sci Sports Exerc 2009 Jul;41(7):1510-1530. [doi: 10.1249/MSS.0b013e3181a0c95c] [Medline: 19516148]
27. Nascimento MDM, Gouveia ?, Marques A, Gouveia BR, Marconcin P, França C, et al. The Role of Physical Function in the Association between Physical Activity and Gait Speed in Older Adults: A Mediation Analysis. IJERPH 2022 Oct 01;19(19):12581. [doi: 10.3390/ijerph191912581]
28. Daskalopoulou C, Stubbs B, Kralj C, Koukounari A, Prince M, Prina A. Physical activity and healthy ageing: a systematic review and meta-analysis of longitudinal cohort studies. Ageing Res Rev 2017 Sep;38:6-17 [FREE Full text] [doi: 10.1016/j.arr.2017.06.003] [Medline: 28648951]
29. Roberts C, Phillips L, Cooper C, Gray S, Allan J. Effect of different types of physical activity on activities of daily living in older adults: systematic review and meta-analysis. J Aging Phys Act 2017 Oct 01;25(4):653-670 [FREE Full text] [doi: 10.1123/japa.2016-0201] [Medline: 28181837]
30. Eckstrom E, Neukam S, Kalin L, Wright J. Physical activity and healthy aging. Clin Geriatr Med 2020 Nov;36(4):671-683. [doi: 10.1016/j.cger.2020.06.009] [Medline: 33010902]
31. Carriedo A, Cecchini JA, Fernandez-Rio J, Méndez-Giménez A. COVID-19, psychological well-being and physical activity levels in older adults during the nationwide lockdown in Spain. The American Journal of Geriatric Psychiatry 2020 Nov;28(11):1146-1155 [FREE Full text] [doi: 10.1016/j.jagp.2020.08.007] [Medline: 32919872]
32. Jyväkorpi SK, Urtamo A, Pitkälä KH, Strandberg TE. Nutrition, daily walking and resilience are associated with physical function in the oldest old men. J Nutr Health Aging 2018 Nov 19;22(10):1176-1182 [FREE Full text] [doi: 10.1007/s12603-018-1136-z] [Medline: 30498823]
33. Wermelinger Ávila MP, Corrêa JC, Lucchetti ALG, Lucchetti G. The role of physical activity in the association between resilience and mental health in older adults. J Aging Phys Act 2018 Apr 01;26(2):248-253. [doi: 10.1123/japa.2016-0332] [Medline: 28714789]
34. Wu-Chung EL, Leal SL, Denny BT, Cheng SL, Fagundes CP. Spousal caregiving, widowhood, and cognition: a systematic review and a biopsychosocial framework for understanding the relationship between interpersonal losses and dementia risk in older adulthood. Neurosci Biobehav Rev 2022 Mar;134:104487-104487 [FREE Full text] [doi: 10.1016/j.neubiorev.2021.12.010] [Medline: 34971701]
35. Carr D, Jeffreys JS. Spousal bereavement in later life. In: Neimeyer RA, Harris DL, Winokuer HR, Thornyon G, editors. Grief and Bereavement in Contemporary Society Bridging Research and Practice (1st Edition). London, UK: Routledge; May 18, 2011:81-92.
36. Centers for Disease Control and Prevention (CDC). Promoting health for older adults. CDC. Atlanta, GA: CDC; 2022. URL: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/promoting-health-for-older-adults.htm#print [accessed 2023-09-07]
37. Cal SF, Sá LRD, Glustak ME, Santiago MB. Resilience in chronic diseases: a systematic review. Cogent Psychology 2015 Apr 14;2(1):1024928. [doi: 10.1080/23311908.2015.1024928]
38. Ministry of Health and Welfare. Taiwan longitudinal study on aging survey report. Ministry of Health and Welfare. 2019. URL: https://www.hpa.gov.tw/Pages/Detail.aspx?nodei [accessed 2023-11-23]
39. Onyeaka H, Anumudu CK, Al-Sharify ZT, Egele-Godswill E, Mbaegbu P. COVID-19 pandemic: a review of the global lockdown and its far-reaching effects. Sci Prog 2021 Jun 01;104(2):368504211019854 [FREE Full text] [doi: 10.1177/00368504211019854] [Medline: 34061685]
40. Cocuzzo B, Wrench A, O'Malley C. Effects of COVID-19 on older adults: physical, mental, emotional, social, and financial problems seen and unseen. Cureus 2022 Sep 23;14(9):e29493 [FREE Full text] [doi: 10.7759/cureus.29493] [Medline: 36299954]
41. Elliott J, Munford L, Ahmed S, Littlewood A, Todd C. The impact of COVID-19 lockdowns on physical activity amongst older adults: evidence from longitudinal data in the UK. BMC Public Health 2022 Sep 22;22(1):1802-1802 [FREE Full text] [doi: 10.1186/s12889-022-14156-y] [Medline: 36138374]
42. Tison GH, Barrios J, Avram R, Kuhar P, Bostjancic B, Marcus GM, et al. Worldwide physical activity trends since COVID-19 onset. The Lancet Global Health 2022 Oct;10(10):e1381-e1382. [doi: 10.1016/s2214-109x(22)00361-8]
43. Beard JR, Officer A, de Carvalho IA, Sadana R, Pot AM, Michel J, et al. The world report on ageing and health: a policy framework for healthy ageing. Lancet 2016 May 21;387(10033):2145-2154 [FREE Full text] [doi: 10.1016/S0140-6736(15)00516-4] [Medline: 26520231]
44. Plácido AI, Herdeiro MT, Roque F. Health and wellbeing in aging. Int J Environ Res Public Health 2022 Jul 20;19(14):8835 [FREE Full text] [doi: 10.3390/ijerph19148835] [Medline: 35886683]

XSL•FO
RenderX

45. GBD 2019 Demographics Collaborators. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950-2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1160-1203 [FREE Full text] [doi: 10.1016/S0140-6736(20)30977-6] [Medline: 33069325]
46. World Health Organization (WHO). Integrated care for older people: guidelines on community-level interventions to manage declines in intrinsic capacity. WHO. Geneva, Switzerland: WHO; 2017. URL: https://apps.who.int/iris/ [accessed 2023-06-22]
47. Cejudo J, García-Castillo FJ, Luna P, Rodrigo-Ruiz D, Feltrero R, Moreno-Gómez A. Using a mindfulness-based intervention to promote subjective well-being, trait emotional intelligence, mental health, and resilience in women with fibromyalgia. Front Psychol 2019 Nov 14;10:2541-2541 [FREE Full text] [doi: 10.3389/fpsyg.2019.02541] [Medline: 31798502]
48. Chmitorz A, Kunzler A, Helmreich I, Tüscher O, Kalisch R, Kubiak T, et al. Intervention studies to foster resilience - a systematic review and proposal for a resilience framework in future intervention studies. Clin Psychol Rev 2018 Feb;59:78-100 [FREE Full text] [doi: 10.1016/j.cpr.2017.11.002] [Medline: 29167029]
49. Nezu AM, Nezu CM, D'Zurilla TJ. Problem-Solving Therapy: A Treatment Manual. New York, NY: Springer Publishing Company; 2013.
50. Vukovic M, Ochoa-Daderska R, Nur AY, Gródek-Szostak Z, Niemczyk A, Akilli A. Situated learning methodology: elders learning English for Europe. ResearchGate. 2022 Jul 01. URL: https://www.researchgate.net/publication/362213717_SITUATED_LEARNING_METHODOLOGY_Elders_learning_English_for_Europe [accessed 2024-07-24]
51. Zhang F, Kaufman D, Schell R, Salgado G, Seah ETW, Jeremic J. Situated learning through intergenerational play between older adults and undergraduates. Int J Educ Technol High Educ 2017 Jul 3;14(1):1-16. [doi: 10.1186/s41239-017-0055-0]
52. Wagnild G, Young H. Development and psychometric evaluation of the Resilience Scale. J Nurs Meas 1993;1(2):165-178. [Medline: 7850498]
53. Li MH. Relationships among stress coping, secure attachment, and the trait of resilience among Taiwanese college students. College Student Journal 2008;42(2, Pt A):312-325 [FREE Full text]
54. Lee S, Tung H, Peng L, Chen L, Hsu C, Huang Y. Resilience among older cardiovascular disease patients with probable sarcopenia. Arch Gerontol Geriatr 2020 Jan;86:103939. [doi: 10.1016/j.archger.2019.103939] [Medline: 31525556]
55. Wu C, Su T, Fang C, Yeh Chang M. Sleep quality among community-dwelling elderly people and its demographic, mental, and physical correlates. J Chin Med Assoc 2012 Feb;75(2):75-80 [FREE Full text] [doi: 10.1016/j.jcma.2011.12.011] [Medline: 22340741]
56. Washburn RA, Smith KW, Jette AM, Janney CA. The Physical Activity Scale for the Elderly (PASE): development and evaluation. J Clin Epidemiol 1993 Feb;46(2):153-162. [doi: 10.1016/0895-4356(93)90053-4] [Medline: 8437031]
57. Casartelli N, Bolszak S, Impellizzeri F, Maffiuletti N. Reproducibility and validity of the Physical Activity Scale for the Elderly (PASE) questionnaire in patients after total hip arthroplasty. Phys Ther 2015 Jan 01;95(1):86-94. [doi: 10.2522/ptj.20130557] [Medline: 25147185]
58. Logan S, Gottlieb B, Maitland S, Meegan D, Spriet L. The Physical Activity Scale for the Elderly (PASE) questionnaire; does it predict physical health? Int J Environ Res Public Health 2013 Aug 30;10(9):3967-3986 [FREE Full text] [doi: 10.3390/ijerph10093967] [Medline: 23999546]
59. Lin Z. The relationship among role playing, social support and well-being for adult students. Journal of Cheng Shiu University 2000;13:269-290 [FREE Full text]
60. Argyle M. The Psychology of Happiness. New York, NY: Routedge; 1987.
61. Haung KC. Research of relationship among well-being, attitudes of death and meaning in life of hospital volunteers in Kaohsiung City. Thesis. National Kaohsiung Normal University. 2003. URL: https://ndltd.ncl.edu.tw/ [accessed 2023-06-26]
62. Hsieh W, Cheng C. The effects of gender differences on the relationship of social capital and well-being: the study of medical volunteers in Hualien Tzu Chi Medical Center. J Educ Health Promot 2016 Jun(45):31-65. [doi: 10.3966/207010632016060045002]
63. Peng SL. A study of the relationship between well-being and health-promoting lifestyle in hospital elderly volunteers. Thesis. National Taipei University of Nursing and Health Sciences. 2023. URL: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=69_zIw/record?r1=1&h1=7#XXX [accessed 2023-06-26]
64. Kusmaryono I, Wijayanti D, Risqi H. Number of response options, reliability, validity, and potential bias in the use of the Likert scale education and social science research: a literature review. Int J Educ Methodol 2022 Nov 15;8(4):625-637. [doi: 10.12973/ijem.8.4.625]
65. Bartholomaeus JD, Van Agteren JEM, Iasiello MP, Jarden A, Kelly D. Positive aging: the impact of a community wellbeing and resilience program. Clinical Gerontologist 2019 Jan 18;42(4):377-386. [doi: 10.1080/07317115.2018.1561582] [Medline: 30654716]
66. Cohen J. American Psychologist 1995 Dec;50(12):1103-1103. [doi: 10.1037/0003-066x.50.12.1103]
67. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 2017 Dec 6;17(1):162-162. [doi: 10.1186/s12874-017-0442-1]

XSL•FO
RenderX

68. Twisk JW, Rijnhart JJ, Hoekstra T, Schuster NA, Ter Wee MM, Heymans MW. Intention-to-treat analysis when only a baseline value is available. Contemp Clin Trials Commun 2020 Dec;20:100684 [FREE Full text] [doi: 10.1016/j.conctc.2020.100684] [Medline: 33319119]
69. Braun V, Clarke V. Thematic Analysis: A Practical Guide. London, UK: Sage Publications; 2021.
70. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
71. Casuso-Holgado MJ, García-Muñoz C, Martín-Valero R, Lucena-Anton D, Moral-Munoz JA, Cortés-Vega MD. Dropout rate in randomised controlled trials of balance and gait rehabilitation in multiple sclerosis: is it expected to be different for virtual reality-based interventions? A systematic review with meta-analysis and meta-regression. Virtual Real 2022 Dec 07:1-17 [FREE Full text] [doi: 10.1007/s10055-022-00733-4] [Medline: 36533191]
72. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022 Apr 26;24(4):e35120 [FREE Full text] [doi: 10.2196/35120] [Medline: 35471414]
73. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
74. Loprinzi CE, Prasad K, Schroeder DR, Sood A. Stress Management and Resilience Training (SMART) program to decrease stress and enhance resilience among breast cancer survivors: a pilot randomized clinical trial. Clin Breast Cancer 2011 Dec;11(6):364-368. [doi: 10.1016/j.clbc.2011.06.008] [Medline: 21831722]
75. Wahbeh H, Nelson M. iRest meditation for older adults with depression symptoms: a pilot study. Int J Yoga Therap 2019 Nov;29(1):9-17. [doi: 10.17761/2019-00036] [Medline: 30354905]
76. Sharit J, Czaja SJ. Overcoming older adult barriers to learning through an understanding of perspectives on human information processing. J Appl Gerontol 2020 Mar 30;39(3):233-241 [FREE Full text] [doi: 10.1177/0733464818794574] [Medline: 30160573]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CRS: Chinese version of the Resilience Scale
PASE: Physical Activity Scale for the Elderly
RCT: randomized controlled trial
RIM: Resilience in Illness Model
WBS: Well-Being Scale
WHO: World Health Organization

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 07.10.23; peer-reviewed by E Vashishtha, B Greeley; comments to author 02.11.23; revised version received 27.11.23; accepted 09.07.24; published 25.07.24.

Please cite as:
Wu YC, Shen SF, Chen LK, Tung HH
A Web-Based Resilience-Enhancing Program to Improve Resilience, Physical Activity, and Well-being in Geriatric Population: Randomized Controlled Trial
J Med Internet Res 2024;26:e53450
URL: https://www.jmir.org/2024/1/e53450
doi:10.2196/53450
PMID:

©Yi-Chen Wu, Shu-Fen Shen, Liang-Kung Chen, Heng-Hsin Tung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of Mobile App–Based Cognitive Behavioral Therapy for Insomnia: Multicenter, Single-Blind Randomized Clinical Trial

Jiyoon Shin[1,2], MD, PhD; Sujin Kim[3], MD; Jooyoung Lee[4], MD; Hyerin Gu[4], MA; Jihye Ahn[4], MA; Chowon Park[5], MA; Mincheol Seo[6], MD; Jeong Eun Jeon[1], MA; Ha Young Lee[1], MA; Ji Won Yeom[7,8], MD; Sojeong Kim[9], MA; Yeaseul Yoon[7], MS; Heon-Jeong Lee[7,8*], MD, PhD; Seog Ju Kim[4*], MD, PhD; Yu Jin Lee[1*], MD, PhD

[1]Department of Psychiatry and Center for Sleep and Chronobiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Department of Psychiatry, Uijeongbu Eulji Medical Center, Eulji University School of Medicine, Gyeonggi-do, Republic of Korea

[3]Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea

[4]Department of Psychiatry, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[5]Department of Psychiatry, Severance Hospital, Seoul, Republic of Korea

[6]Department of Psychiatry, Veteran Health Service Medical Center, Seoul, Republic of Korea

[7]Chronobiology Institute, Korea University, Seoul, Republic of Korea

[8]Department of Psychiatry, Korea University College of Medicine, Seoul, Republic of Korea

[9]Department of Psychology, University of Oregon, Eugene, OR, United States

[*]these authors contributed equally

Corresponding Author:
Yu Jin Lee, MD, PhD
Department of Psychiatry and Center for Sleep and Chronobiology, Seoul National University Hospital, Seoul National University College of Medicine
103 Daehak-ro, Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 2456
Email: ewpsyche@snu.ac.kr

Abstract

Background: Cognitive behavioral therapy for insomnia (CBTi) is the first-line therapy for chronic insomnia. Mobile app–based CBTi (MCBTi) can enhance the accessibility of CBTi treatment; however, few studies have evaluated the effectiveness of MCBTi using a multicenter, randomized controlled trial design.

Objective: We aimed to assess the efficacy of Somzz, an MCBTi that provides real-time and tailored feedback to users, through comparison with an active comparator app.

Methods: In our multicenter, single-blind randomized controlled trial study, participants were recruited from 3 university hospitals and randomized into a Somzz group and a sleep hygiene education (SHE) group at a 1:1 ratio. The intervention included 6 sessions for 6 weeks, with follow-up visits over a 4-month period. The Somzz group received audiovisual sleep education, guidance on relaxation therapy, and real-time feedback on sleep behavior. The primary outcome was the Insomnia Severity Index score, and secondary outcomes included sleep diary measures and mental health self-reports. We analyzed the outcomes based on the intention-to-treat principle.

Results: A total of 98 participants were randomized into the Somzz (n=49, 50%) and SHE (n=49, 50%) groups. Insomnia Severity Index scores for the Somzz group were significantly lower at the postintervention time point (9.0 vs 12.8; t_{95}=3.85; $F_{2,95}$=22.76; η_p^2=0.13; P<.001) and at the 3-month follow-up visit (11.3 vs 14.7; t_{68}=2.61; $F_{2,68}$=5.85; η_p^2=0.03; P=.01) compared to those of the SHE group. The Somzz group maintained their treatment effect at the postintervention time point and follow-ups, with a moderate to large effect size (Cohen d=–0.62 to –1.35; P<.01 in all cases). Furthermore, the Somzz group showed better sleep efficiency (t_{95}=–3.32; $F_{2,91}$=69.87; η_p^2=0.41; P=.001), wake after sleep onset (t_{95}=2.55; $F_{2,91}$=51.81; η_p^2=0.36; P=.01), satisfaction (t_{95}=–2.05; $F_{2,91}$=26.63; η_p^2=0.20; P=.04) related to sleep, and mental health outcomes, including depression (t_{95}=2.11;

XSL•FO
RenderX

$F_{2,94}$=29.64; η_p^2=0.21; P=.04) and quality of life (t_{95}=–3.13; $F_{2,94}$=54.20; η_p^2=0.33; P=.002), compared to the SHE group after the intervention. The attrition rate in the Somzz group was 12% (6/49).

Conclusions: Somzz outperformed SHE in improving insomnia, mental health, and quality of life. The MCBTi can be a highly accessible, time-efficient, and effective treatment option for chronic insomnia, with high compliance.

Trial Registration: Clinical Research Information Service (CRiS) KCT0007292; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=22214&search_page=L

(J Med Internet Res 2024;26:e50555) doi:10.2196/50555

KEYWORDS

digital therapeutics; mobile app–based cognitive behavioral therapy for insomnia; cognitive behavioral therapy; insomnia; mental health; mobile phone

Introduction

Background

Insomnia is a highly prevalent (global prevalence of 2.3%-25.5%) sleep disorder that substantially impacts population health worldwide [1]. Individuals with insomnia often experience decreased quality of life [2] and have a higher risk of developing psychiatric illnesses [3] as well as medical conditions, including cardiometabolic and neurocognitive disorders [4].

Although cognitive behavioral therapy for insomnia (CBTi) is the first-line treatment for insomnia [5-7], its accessibility is limited. These limitations arise due to the time-consuming nature of this treatment and the shortage of trained physicians worldwide [8-10]. Therefore, many individuals face difficulties in obtaining face-to-face CBTi treatment. However, considering the busy working hours, coupled with the high penetration of smartphones in modern society, mobile app–based digital CBTi (MCBTi) may offer easy accessibility to a larger population with minimal space constraints. Digital CBTi is a type of digital therapeutics that uses qualified software programs to deliver evidence-based therapeutic interventions to patients [11] to prevent, manage, or treat medical conditions through effective therapeutic techniques.

Among the various methods of delivering digital CBTi, web-based CBTi, in which a real or web-based therapist provides feedback and supervision, is considered optimal for improving sleep efficiency, sleep onset latency (SOL), and wake after sleep onset (WASO) [12]. Although there have been numerous studies regarding the effectiveness of web-based CBTi [13], the efficacy of MCBTi for improving insomnia and mental health remains unknown [12]. Only a few studies that examined the MCBTi were designed as randomized controlled trials (RCTs) [14] and the control group did not use the same modality (eg, non–app-using control or waitlist control) [15,16]. One MCBTi study with an RCT design used the same modality of control, an active comparator app that provided feedback in chatbot form [10]; the study participants had an Insomnia Severity Index (ISI) score of ≥15, indicating a moderate degree of insomnia. Another multicenter RCT study demonstrated the effectiveness of MCBTi [17], suggesting the inclusion of mental health outcomes in future studies.

In this study, we considered a CBTi mobile app, Somzz, that can be applied to a wider range of insomniacs, ranging from those who experience mild insomnia (ISI score of ≥8) to those with severe insomnia. With this app, CBTi treatment is provided through audiovisual sleep education, guidance on relaxation therapy, and real-time customized feedback on participants' sleep behavior by analyzing the data entered into the app. Cognitive therapy through a chatbot format is used to correct dysfunctional beliefs related to sleep. In addition, because the effectiveness of sleep restriction has consistently been reported [18], daily *push-alarm* reminders were used to ensure adherence to the prescribed time in bed (TIB).

Objectives

We aimed to compare the treatment effect of Somzz with an active comparator app providing sleep hygiene education (SHE). We hypothesized that participants using Somzz would demonstrate better outcomes in measures such as the ISI, Epworth Sleepiness Scale (ESS), and sleep diary measures (including sleep efficiency, SOL, WASO, total sleep time, refreshment after sleep, and sleep satisfaction). In addition, we hypothesized that mental health–related issues, including depression, anxiety, and quality of life, would show greater improvement in the Somzz group.

Methods

Participants

This multicenter study recruited participants from Seoul National University Hospital, Korea University Anam Hospital, and Samsung Medical Center through advertisements placed in the hospitals and local community (trial registration: KCT0007292). Psychiatrists (H-JL, SJK, and YJL) used the International Classification of Sleep Disorders, Third Edition, to confirm chronic insomnia diagnoses. The participants were recruited between January 2022 and May 2022.

Participants had to meet all of the following inclusion criteria: (1) age of ≥19 years; (2) meeting the diagnostic criteria for chronic insomnia disorder according to the International Classification of Sleep Disorders, Third Edition; (3) ISI score of ≥8; (4) use of Android phones with operating system version 7.0 or above; (5) voluntary decision to participate in the clinical trial and provide written consent on the informed consent form; and (6) willingness to comply with the clinical trial protocol.

XSL•FO
RenderX

Participants were excluded from the study if any of the following exclusion criteria were met: (1) presence of sleep disorders other than chronic insomnia disorder (eg, hypersomnia, narcolepsy, or sleep-disordered breathing); (2) presence of serious medical illnesses; (3) underlying disorders that could be exacerbated by sleep restriction (eg, bipolar disorder and psychosis); (4) severe depression, defined by a Hamilton Depression Rating Scale score of ≥24; (5) presence of medications (eg, chronic use of stimulant medication) or lifestyle factors that provoke insomnia (eg, caffeine, alcohol, or tobacco addiction); (6) challenging schedules that make it difficult to receive sleep interventions (eg, night shift workers or rotating shift workers); (7) presence of auditory or cognitive impairments that make it difficult to undergo treatment using the investigational medical device for the clinical trial; (8) difficulty using universal devices such as smartphones; (9) lack of completion of electronic sleep diaries on at least 7 occasions within 10 days from the first day of electronic sleep diary entry; (10) being pregnant or lactating; (11) current participation in another clinical trial or participation in another clinical trial within 90 days from the screening date; and (12) in the judgment of the investigator, being deemed inappropriate for participation in the clinical trial due to the ethical or potential impact on the clinical trial results.

Individuals who met the study inclusion criteria were randomly assigned to either the Somzz or SHE group at a 1:1 ratio. In this single-blind RCT, the participants were blinded to group allocation. Random allocation was performed using an interactive web response system and a predetermined randomization list. Stratified block randomization was implemented by statisticians using the SAS PROC PLAN procedure; the block size was predetermined for each clinical trial site. Stratification was conducted based on the recruiting institution. The block size and seed number used were selected randomly by the responsible statistician for randomization.

The intervention and control groups had the same visit schedule, and visits were supervised by clinicians or trained psychologists. Written consent was obtained before clinical assessments (eg, vital signs, physical examination, assessment of combination therapy, completion of self-reported scales, and System Usability Scale assessment). However, all therapeutic interventions for insomnia were exclusively administered through the Somzz app. There were 6 visits in total: visit 1 involved screening for study eligibility; visit 2 (week 0) encompassed the collection of baseline demographic and ISI data; visits 4 (week 2) and 6 (posttreatment time point) involved assessment of treatment adherence and completion of sleep-related, mental health, and quality of life scales; and visits 3 (week 1) and 5 (week 4)—both telephone visits—consisted of ISI and ESS completion.

Ethical Considerations

The study protocol was approved by the Institutional Review Board for Human Subjects of Seoul National University Hospital (2109-145-1258). Detailed information about the study was provided to the participants, and written informed consent was obtained before enrollment. A transportation subsidy of 50,000 KRW (US $42) was provided for each of the four visits to the institution. The collected data were anonymized.

Sample Size Calculation

Sample size was calculated based on group differences and SDs reported in previous papers regarding digital CBTi (the Food and Drug Administration–approved "SHUT-I" program) [19,20]. Assuming an intergroup difference of –4.6 (SD 6.2), a significance level of 5%, and a power of 90%, the sample size calculation yielded 39 participants per group. Assuming a dropout rate of 20%, we intended to recruit 49 participants per group—98 participants in total.

Intervention

Somzz is a mobile app software medical device designed for the treatment of chronic insomnia. It implements the CBTi protocol, incorporating stimulus control; sleep restriction; SHE; and cognitive therapy, the first-line treatment for chronic insomnia. The app consists of six steps conveyed for 6 weeks: (1) general sleep education, (2) stimulus control and sleep restriction, (3) SHE, (4) relaxation techniques, (5) cognitive therapy, and (6) relapse prevention and termination (Table 1). The Somzz group participants were provided with real-time data-based customized feedback, daily assignments, behavioral interventions (eg, recommendations for TIB), and push notification messages (eg, when to write daily in the sleep diary and when to go to bed considering past TIB) during the 6-week intervention. The intervention was conducted in a fully automated manner. Screenshots of Somzz are presented in Figure 1.

Real-time TIB recommendations constitute a primary feature of Somzz. On the basis of each participant's sleep schedule as recorded in their sleep diary, the app calculates their sleep efficiency and prompts them to adjust their TIB as necessary. If a participant stays in bed longer than the recommended time, real-time feedback is provided in the form of alarms. In addition, as part of a daily *to-do* list, push notifications encourage participants to complete >6000 steps during the day and ensure sufficient exposure to sunlight. Somzz also incorporates a chatbot module, allowing users to change any dysfunctional beliefs related to insomnia by answering the chatbot therapist's questions. The same user interface was used for SHE, although the sleep diary module was only 6 weeks in duration and no TIB recommendations were provided. These differences in features may have led to differences in long-term treatment outcomes between the Somzz and SHE groups, as revealed by the linear mixed model analysis of ISI scores, ESS scores, and sleep efficiency.

The SHE group participants also installed the active comparator app on their phones, which comprised SHE and a daily sleep diary. However, no push notification messages or TIB prescriptions were sent to the participants. Sleep education in both apps were provided through audiovisual materials (eg, movie clips including narrations and text).

Participants were followed for 6 weeks during the intervention, and those who consented to extended follow-ups were evaluated for an additional 4 months after the intervention (with 4 additional visits per month) to assess the maintenance of treatment effects.

XSL•FO
RenderX

Table 1. Contents of the Somzz sessions.

Session	Contents	Additions
Session 1: basic education	When a patient first signs up on the app, they undergo a sleep assessment to evaluate their basic information, the severity of insomnia symptoms, and the presence of comorbid conditions. In addition, they receive basic education and set goals.	Homework
Session 2: stimulus control and sleep restriction	Somzz provides education on stimulus control, behavioral interventions, and feedback. The goal is to use the bed or bedroom solely for sleeping purposes and to engage in sleep-disruptive activities outside the bedroom. This aims to regulate sleep by adjusting the sleep environment, bedtime, and bedroom conditions in relation to sleep. It is a behavior therapy technique that associates healthy habits conducive to falling asleep with modifying maladaptive conditioning related to sleep. Avoiding activities such as watching television, work-related behaviors, and using the internet or mobile phones in bed is encouraged, and daily compliance is recorded and reviewed during CBTi[a] sessions. Somzz also provides education on sleep restriction, behavioral interventions, and feedback. In other words, we assess the patient's sleep duration through a sleep diary to implement sleep restriction. Considering the patient's sleep efficiency, we determine the target sleep duration and wake-up time. We gradually adjust the total TIB[b] by 15 minutes per week until reaching the target sleep efficiency of 85%.	Homework, feedback, and TIB recommendations
Session 3: sleep hygiene education	Somzz provides education on sleep hygiene, behavioral interventions, and feedback. This involves examining factors that can impact sleep, such as excessive consumption of coffee or alcohol, and providing education to establish healthy habits for sleep. Subsequently, we encourage daily recording of sleep habits and provide feedback on them.	Homework, feedback, and TIB recommendations
Session 4: relaxation therapy	Somzz provides education on relaxation therapy, behavioral interventions, and feedback. This includes training on relaxation techniques such as diaphragmatic breathing and progressive muscle relaxation, providing a structured self-training schedule, and offering feedback on performance.	Homework, feedback, and TIB recommendations
Session 5: cognitive therapy	Somzz provides cognitive therapy counseling and feedback. This therapy aims to address and correct automatic thoughts related to insomnia that contribute to worsening sleep. It focuses on examining and correcting dysfunctional thoughts associated with excessive worry about insomnia, anxiety arising from perceived health and psychological effects of insomnia, and thoughts characterized by excessive preoccupation with sleep duration. The therapy also targets catastrophic thinking regarding the consequences of poor sleep and generalizations attributing discomforting symptoms solely to sleep-related causes.	Homework, feedback, and TIB recommendations
Session 6: relapse prevention and termination	Somzz conducts a reassessment of symptoms before program termination and provides feedback on progress compared to the initial evaluation. We also provide feedback and conduct a relapse prevention education session followed by a final evaluation.	Homework and feedback

[a]CBTi: cognitive behavioral therapy for insomnia.

[b]TIB: time in bed.

Figure 1. Example screenshots of the Somzz app.

XSL•FO
RenderX

Outcomes and Measures

The primary outcome was the comparison of total ISI score between the Somzz and SHE groups at each clinical and postintervention follow-up visit [21]. Secondary outcomes included sleep diary and psychiatric measures. Sleep diary measures recorded via the Somzz app included SOL, sleep efficiency, total sleep time, number of awakenings during sleep, WASO, refreshment after waking up, and satisfaction with sleep. Psychiatric measures included daytime sleepiness, as indicated using the ESS [22]; dysfunctional beliefs, measured using the Dysfunctional Beliefs and Attitudes About Sleep Scale (DBAS) [23]; depressive symptom severity, measured using the Patient Health Questionnaire–9 (PHQ-9) [24]; anxiety symptom severity, measured using the Generalized Anxiety Disorder Scale–7 [25]; fatigue, measured using the Fatigue Severity Scale (FSS) [26]; and health-related quality of life, measured using the 36-item Short Form Health Survey (SF-36) [27] and the EQ-5D-5L [28].

ISI and ESS scores were obtained at each visit, and sleep diary measures were evaluated daily. Other self-report scales, including psychiatric, fatigue, and quality of life measures, were completed at baseline, the postintervention time point, and during the 4-month follow-up period. We defined remission as an ISI score of <8 after treatment; thus, a decrease in the ISI score of >7 points from baseline was defined as a treatment response.

Statistical Analysis

ISI scores, sleep diary, and psychiatric measures were compared between the 2 groups using an analysis of covariance adjusting for baseline severity and within groups using a paired 2-tailed *t* test. Regarding the calculation of effect sizes in the paired *t* test, in the following equation, s_p is the pooled s_d, n is the number of paired observations, X_i represents the difference in each paired observation, and is the mean of these differences:

(1)

Linear mixed model analysis was conducted to examine the group-time interaction regarding ISI and ESS scores and sleep diary measures. Factors associated with remission (ISI score of <8) and response (ISI score decrease from baseline of >7) were evaluated using binary logistic regression. The primary and secondary outcomes were analyzed using the intention-to-treat principle; the analysis included all participants randomly allocated to either the Somzz or SHE group. In addition, the last observation carried forward method was used to handle missing data. All analyses were conducted using R (version 4.2.0; The R Foundation). We used the R package *lme4* [29] to fit the linear mixed model. Prespecified analysis plan can be found in Multimedia Appendix 1 [30, 31].

Results

Overview

A total of 98 participants were randomly assigned to the Somzz and SHE groups, with each group consisting of 49 (50%) individuals. Of these 98 participants, 5 (5%) individuals from the Somzz group and 4 (4%) from the SHE group withdrew their consent, and 1 (1%) participant in each group did not complete the study (Figure 2). The baseline demographic data are presented in Table 2. In total, 71 participants (34/49, 69% in the Somzz group and 37/49, 76% in the SHE group) consented to participate in the extended (4-month) follow-up study.

Figure 2. Flowchart of participant enrollment. SHE: sleep hygiene education.

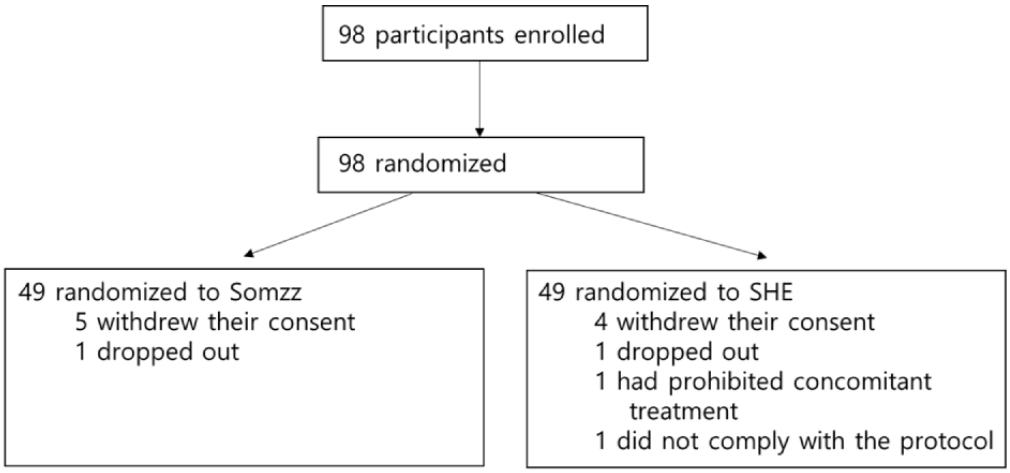

XSL•FO
RenderX

Table 2. Baseline demographic data for the Somzz and sleep hygiene education (SHE) groups.

	Somzz (n=49)	SHE (n=49)
Sex, n (%)		
Male	20 (41)	18 (37)
Female	29 (59)	31 (63)
Age (years), mean (SD)	44.1 (13.1)	40.5 (12.8)
Insomnia duration, n (%)		
3-6 months	2 (4)	4 (8)
6-12 months	4 (8)	5 (10)
1-3 years	17 (35)	12 (24)
3-5 years	7 (14)	12 (24)
Approximately 5 years	19 (39)	16 (33)
Occupation, n (%)		
Employed full-time	13 (27)	20 (41)
Employed part-time	9 (18)	8 (16)
Retired	1 (2)	2 (4)
Student	4 (8)	6 (12)
Unemployed	14 (29)	8 (16)
Other	8 (16)	5 (10)
Current medical illness, n (%)	17 (35)	17 (35)
Current use of hypnotics, n (%)	9 (18)	9 (18)
Days of use per week, mean (SD)	3.9 (2.5)	4.2 (2.6)
Caffeine use		
Current user, n (%)	35 (71)	36 (73)
Caffeine units, mean (SD)	1.7 (1.1)	1.5 (1.0)
Ex-user, n (%)	2 (4)	1 (2)
Never, n (%)	12 (24)	12 (24)
Smoking		
Current smoker, n (%)	9 (18)	4 (8)
Smoking units, mean (SD)	4.8 (4.8)	3.5 (2.4)
Duration of use (years), mean (SD)	9.9 (10.0)	11.8 (7.7)
Ex-smoker, n (%)	4 (8)	8 (16)
Never smoker, n (%)	36 (73)	37 (76)
Drinking		
Current drinker, n (%)	26 (53)	28 (57)
Drinking units, mean (SD)	2.5 (3.0)	3.5 (4.5)
Duration of use (years), mean (SD)	19.8 (12.5)	14.9 (11.4)
Ex-drinker, n (%)	1 (2)	4 (8)
Never drinker, n (%)	22 (45)	17 (35)
Psychiatric diagnosis, n (%)	5 (10)[a]	6 (12)
Depressive disorder	4 (8)	4 (8)
Anxiety disorder	1 (2)	__[b]
Panic disorder	1 (2)	1 (2)

XSL•FO
RenderX

	Somzz (n=49)	SHE (n=49)
Bipolar II disorder	1 (2)	1 (2)
Alcohol use disorder	1 (2)	—

[a]One participant had depression, panic disorder, and an alcohol use disorder, and another one was diagnosed with depression and anxiety.
[b]Not applicable.

Primary Outcomes

Comparison of the ISI scores between the Somzz and SHE groups showed that the scores of the Somzz group were lower at the postintervention time point (9.0 vs 12.8; t_{95}=3.85; $F_{2,95}$=22.76; η_p^2=0.13; *P*<.001) and at the 1-month (10.3 vs 14.2; t_{68}=3.24; $F_{2,68}$=18.63; η_p^2=0.14; *P*=.002), 2-month (10.9 vs 14.7; t_{68}=2.51; $F_{2,68}$=10.94; η_p^2=0.09; *P*=.02), and 3-month (11.3 vs 14.7; t_{68}=2.61; $F_{2,68}$=5.85; η_p^2=0.03; *P*=.01) follow-ups (Figure 3A). At the 4-month follow-up, no difference in ISI scores was found between the 2 groups (12.1 vs 13.7; t_{68}=0.95; $F_{2,68}$=6.19; η_p^2=0.07; *P*=.35). The linear mixed model analysis with a random slope revealed a significant group-time interaction effect from baseline to the postintervention time point (estimate=0.76; t_{96}=3.02; marginal R^2=0.55; *P*=.003) and from the postintervention time point to the 4-month follow-up (estimated=–0.72; $t_{70.25}$=–2.64; marginal R^2=0.47; *P*=.01). However, when considering baseline to the 4-month follow-up, the model was not significant (estimate=0.32; $t_{79.61}$=1.94; marginal R^2=0.45; *P*=.06).

The Somzz group showed treatment effects on the ISI score from intervention week 2 (Cohen *d*=–0.59; *P*=.02); the effect was sustained up to the 4-month follow-up (Cohen *d*=–0.62; *P*=.003) compared to the baseline ISI score (Figure 3B). The maximum treatment effect was observed at the postintervention time point (Cohen *d*=–1.35; *P*<.001). The SHE group also showed improvement at intervention week 2 (Cohen *d*=–0.55; *P*.004), week 4 (Cohen *d*=–0.71; *P*<.001), the postintervention time point (Cohen *d*=–0.61; *P*=.002), and the 4-month follow-up (Cohen *d*=–0.35; *P*=.04; Figure 3B).

Both the Somzz and SHE groups showed improvements in ISI scores compared to baseline. In the Somzz group, the maximum improvement in ISI score was observed at the postintervention time point (Cohen *d*=–1.35), and the intervention effect decreased gradually during the follow-up period (Figure 3B). The treatment effect was maintained at the 4-month follow-up in the Somzz group (Cohen *d*=–0.62; Figure 3B). However, there was no difference between the groups at that point (estimate=1.16; t_{68}=0.95; $F_{2,68}$=6.19, η_p^2=0.07; *P*=.35; Figure 3A).

The proportion of individuals in remission at the postintervention time point was 45% (22/49) in the Somzz group, whereas in the SHE group, the proportion was 12% (6/49; risk ratio=3.67, 95% CI 1.32-10.20; *P*.01), and the risk difference was 0.33 (95% CI 0.11-0.54; *P*.002). Regarding the intervention response, there were 57% (28/49) of responders in the Somzz group and 22% (11/49) of responders in the SHE group (risk ratio=2.55, 95% CI 1.06-6.12; *P*.04), and the risk difference was 0.35 (95% CI 0.10-0.60; *P*.006).

XSL•FO
RenderX

Figure 3. Insomnia Severity Index (ISI) scores at each time point. (A) Between-group comparison of ISI scores at each time point (analysis of covariance). (B) Within-group analysis of ISI scores at each clinical visit compared with baseline scores (2-tailed paired t test). The numbers on the bar graphs are Cohen d values. Error bars are also presented. In the Somzz and sleep hygiene education (SHE) groups, 69% (34/49) and 76% (37/49) of the participants, respectively, consented to take part in the 4-month follow-up. The postintervention time point was 6 to 9 weeks after the baseline. *P<.05; **P<.01; ***P<.001.

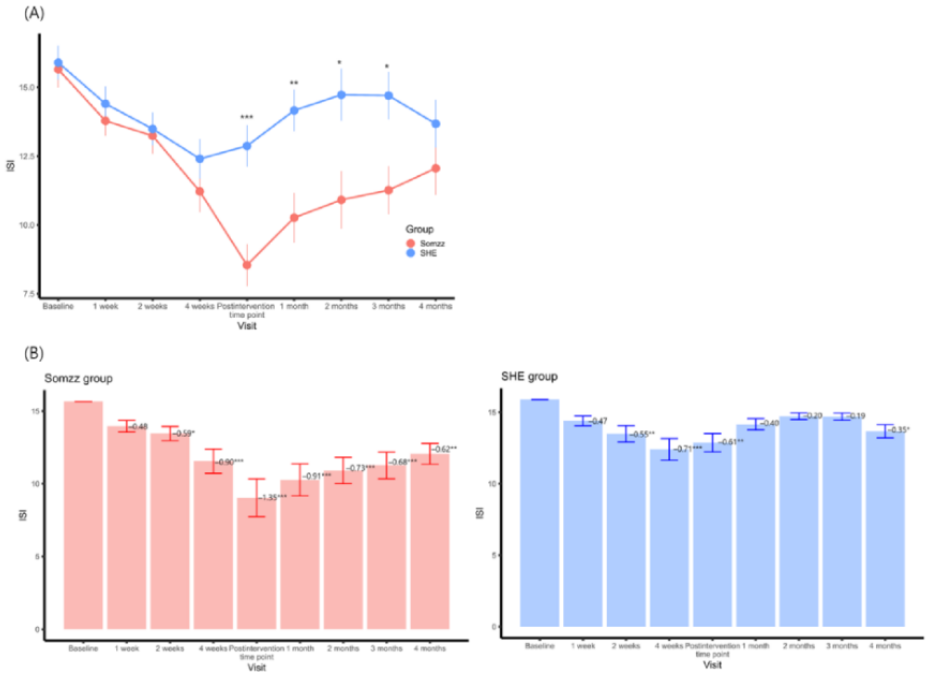

Sleep Diary Measures

The sleep diary measures were compared between the 2 groups with adjustment for baseline measures (Multimedia Appendix 2). Significant group differences were observed in sleep efficiency during treatment sessions 2 to 6 (Multimedia Appendix 2 and Figure 4). The Somzz group also had a shorter WASO and higher satisfaction levels after sleep compared to the SHE group after the intervention.

The linear mixed model analysis with a random slope revealed significant group-time interaction effects on sleep efficiency in the treatment sessions (estimate=–1.00; t_{92}=–2.70; adjusted R^2=0.71; P=.008).

Figure 4. Comparison of sleep efficiency via sleep diary measures between the Somzz and sleep hygiene education (SHE) groups. Group comparison of Insomnia Severity Index scores at each time point using analysis of covariance adjusting for the baseline score. *P<.05; **P<.01; ***P<.001.

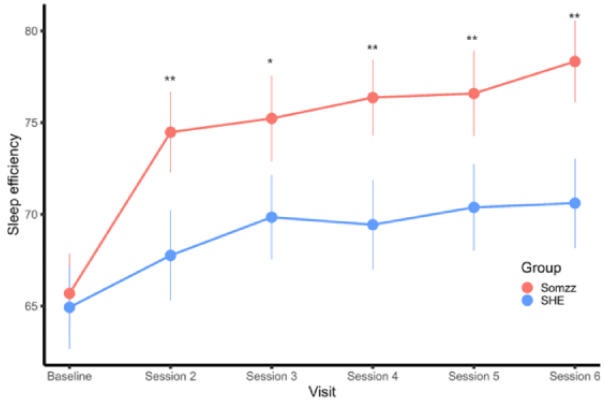

Self-Report Questionnaires

With adjustment for baseline severity, scores for self-report measures including the ESS, DBAS, PHQ-9, FSS, and SF-36 significantly differed between the Somzz and SHE groups after the intervention (Table 3). In particular, the differences in DBAS and FSS scores persisted at the 4-month follow-up (Table 3).

The linear mixed model analysis with a random slope revealed a significant group-time interaction effect on ESS from baseline to the postintervention time point (estimate=0.59; $t_{90.90}$=3.48; marginal R^2=0.51; P<.001), from the postintervention time point to the 4-month follow-up (estimate=–0.49; $t_{300.56}$=–2.58; marginal R^2=0.43; P=.01), and from baseline to the 4-month follow-up (estimate=0.28; $t_{79.11}$=3.35; marginal R^2=0.49; P=.001).

XSL•FO
RenderX

Table 3. Comparisons of self-report questionnaires between the Somzz and sleep hygiene education (SHE) groups[a].

	Somzz (n=49)		SHE (n=49)		ANCOVA[b] statistics			
	Values, mean (SD)	Cohen *d*	Values, mean (SD)	Cohen *d*	*t* test (*df*)	*F* test (*df*)	η_p^2	*P* value
ESS[c]								
Baseline	6.7 (3.9)	__[d]	6.4 (3.4)	—	—	—	—	—
Postintervention time point	4.3 (3.2)	–0.72[e]	6.6 (4.3)	0.01	3.82 (95)	32.38 (2, 94)	0.22	<.001
4-month follow-up	6.1 (3.7)	–0.51	5.8 (3.6)	–0.17	0.84 (68)	26.10 (2, 67)	0.28	.41
DBAS[f]								
Baseline	107.3 (21.2)	—	99.8 (25.8)	—	—	—	—	—
Postintervention time point	74.0 (33.7)	–1.01[g]	98.4 (20.3)	–0.10	5.32 (95)	21.55 (2, 94)	0.10	<.001
4-month follow-up	80.2 (28.5)	–1.06[g]	94.9 (22.8)	–0.17	3.35 (68)	12.97 (2, 67)	0.13	.001
PHQ-9[h]								
Baseline	9.4 (4.9)	—	10.4 (5.0)	—	—	—	—	—
Postintervention time point	6.6 (5.2)	–0.67[e]	8.7 (3.9)	–0.36	2.11 (95)	29.64 (2, 94)	0.21	.04
4-month follow-up	6.8 (4.6)	–0.62[i]	8.6 (4.8)	–0.34	1.21 (68)	9.90 (2, 67)	0.11	.23
GAD-7[j]								
Baseline	6.6 (4.5)	—	7.6 (4.6)	—	—	—	—	—
Postintervention time point	4.9 (4.5)	–0.44	6.7 (3.7)	–0.27	1.69 (95)	35.36 (2, 94)	0.25	.10
4-month follow-up	5.4 (4.6)	–0.28	6.3 (4.5)	–0.38	0.35 (68)	6.96 (2, 67)	0.09	.73
FSS[k]								
Baseline	39.8 (12.8)	—	44.8 (8.9)	—	—	—	—	—
Postintervention time point	33.6 (13.8)	–0.71[i]	44.5 (9.0)	–0.13	3.88 (95)	87.91 (2, 94)	0.41	<.001
4-month follow-up	34.5 (12.6)	–0.60	42.1 (10.0)	–0.39	2.04 (68)	41.62 (2, 67)	0.33	.046
SF-36[l]								
Baseline	65.4 (16.9)	—	62.5 (15.2)	—	—	—	—	—
Postintervention time point	72.4 (16.8)	0.58[i]	63.5 (14.1)	0.08	–3.13 (95)	54.20 (2, 94)	0.33	.002
4-month follow-up	68.9 (20.8)	0.21	66.0 (14.3)	0.19	–0.59 (68)	12.07 (2, 67)	0.15	.56
EQTOTAL[m]								
Baseline	7.6 (2.0)	—	8.2 (2.0)	—	—	—	—	—
Postintervention time point	7.3 (2.3)	–0.19	7.8 (1.9)	–0.21	0.40 (95)	44.30 (2, 94)	0.31	.69
4-month follow-up	7.2 (1.9)	–0.19	7.4 (2.6)	–0.22	0.02 (68)	4.75 (2, 67)	0.06	.98

[a]Cohen *d* represents the effect sizes of within-group comparisons (with baseline score as the reference) performed using paired *t* tests (2-tailed). *t* test and *P* values are for group comparisons performed using analysis of covariance. In the Somzz and SHE groups, 69% (34/49) and 76% (37/49) of the participants, respectively, consented to take part in the 4-month follow-up.

XSL•FO
RenderX

[b]ANCOVA: analysis of covariance.
[c]ESS: Epworth Sleepiness Scale.
[d]Not applicable.
[e]$P<.01$.
[f]DBAS: Dysfunctional Beliefs and Attitudes About Sleep Scale.
[g]$P<.001$.
[h]PHQ-9: Patient Health Questionnaire–9.
[i]$P<.05$.
[j]GAD-7: Generalized Anxiety Disorder Scale–7.
[k]FSS: Fatigue Severity Scale.
[l]SF-36: 36-item Short Form Health Survey.
[m]EQTOTAL: EQ-5D-5L.

Concurrent Medications

Medications related to sleep were used by 18% (9/49) of participants in the Somzz group and 18% (9/49) of participants in the SHE group (χ^2_1=0.0; P>.99). Specifically, mirtazapine was used by 2% (1/49) of participants in the Somzz group and none in the SHE group, and benzodiazepines were used by 8% (4/49) of participants in the Somzz group and 14% (7/49) of participants in the SHE group. Z-drugs were used by 10% (5/49) of participants in each group. Antihistamines were used by 2% (1/49) of participants in the Somzz group and 12% (6/49) of participants in the SHE group.

Factors Associated With Remission and Response in the Somzz Group

Binary regression analysis of remission in the Somzz group revealed that higher DBAS and nondrinking were associated with nonremission (DBAS: estimate=0.11, z=2.05, and P=.04; drinking: estimate=4.39, z=2.25, and P=.03). In addition, the binary regression model of responders indicated that a younger age (estimate=0.28; z=2.07; P=.04), absence of medical history (estimate=–8.01; z=–2.01; P=.045), and shorter TIB (estimate=0.03; z=2.11; P=.04) at baseline were significantly associated with the treatment response in the Somzz group.

Discussion

Principal Findings

In this study, we found that the Somzz group showed significantly greater improvements in ISI scores, ESS scores, sleep efficiency, WASO, and satisfaction after sleep at the postintervention time point compared to the SHE group. The effect on ISI scores persisted at the 4-month follow-up in the Somzz group. In addition, the Somzz group exhibited lower scores in the ESS, DBAS, PHQ-9, FSS, and SF-36 at the postintervention time point than the SHE group. Among the various methods for delivering digital or internet-based CBTi, to our knowledge, this study is the first to investigate effect of an MCBTi on sleep as well as mental health with a multicenter RCT design using an active control. The Somzz group achieved a favorable attrition rate of 12% (6/49), suggesting that Somzz is an efficacious insomnia treatment with high sustainability of use.

The Somzz group demonstrated a decrease in insomnia severity, as measured using the ISI, starting from 2 weeks after the intervention with a gradual decrease until the postintervention visit. The treatment effect was sustained until the 4-month follow-up, with a moderate to large effect size observed from the second week of the intervention to the 4-month follow-up compared to the baseline ISI score. In the Somzz group, basic education was implemented in the first week followed by stimulus control and sleep restriction in the second week. It appears that stimulus control and sleep restriction play more critical roles in improving insomnia, with a greater decrease in ISI score at week 2. Furthermore, CBTi can initially reduce sleep time, resulting in a possible aggravation of insomnia symptoms, which improve with time [32]. The active comparator app of the SHE group, on the other hand, only includes SHE and a sleep diary module. The SHE group also presented with reduced insomnia severity compared to the baseline from intervention week 3 to the postintervention time point, with a smaller effect size than that of the Somzz group. The findings indicate that SHE has some effect but not sufficient to invoke insomnia improvement [10,33,34], which can be reinforced by other components of CBTi, as indicated by a larger effect size in the Somzz group. Moreover, the impact of SHE might have been influenced by the Hawthorne effect [35] and the natural progression of insomnia [36]. Furthermore, participants with mild insomnia (ISI score of ≥8 and <15) may have partially benefited from SHE. The larger effect size of the ISI score decrease observed in the Somzz group indicates a superior treatment effect on insomnia severity compared to that of the SHE group.

Group differences in ISI scores were noted at the postintervention time point, whereas no significant differences in ISI scores were observed between the Somzz and SHE groups at weeks 1, 2, and 4. Cognitive therapy, TIB restriction, and relapse prevention were the components of the Somzz from week 4 to the postintervention time point. Assuming the “additive model” posited by a component network meta-analysis [18] where the combination therapy effect is viewed as a simple sum of effects without interaction, we presume that cognitive therapy and TIB restriction may have been important factors in improving subjective sleep quality in this study, which aligns with the results of the meta-analysis. Given the limitations of cognitive therapy provided in MCBTi compared to face-to-face CBTi, we believed that the significant effect may be due to the consistent TIB restriction over the 6 weeks in this study. In addition, significant differences in sleep efficiency and SOL between the 2 groups were observed at session 2. This suggests

XSL•FO
RenderX

that combination of TIB restriction and stimulus control may have significant effects on sleep efficiency and SOL, which is also consistent with the results of the meta-analysis [18]. However, the conclusions should be interpreted meticulously due to the possible complex interactions among the components of CBTi. Further component network meta-analyses encompassing more MCBTi studies would be helpful for more accurate conclusions.

Although insomnia severity remained significantly reduced compared to baseline at the 4-month follow-up, it showed an increasing trend after the intervention in both the Somzz and SHE groups. Group differences in ISI scores were observed at the postintervention time point and at the 1-, 2-, and 3-month follow-ups, but these differences disappeared at the 4-month follow-up. The lack of difference between the 2 groups was also observed in an earlier study in which group differences in ISI scores were observed at the 3-month follow-up but not at the 6-month follow-up [10]. These results suggest that additional booster sessions approximately 3 months after the intervention may be beneficial in sustaining the better treatment effect of MCBTi. However, the ISI score decrease was relatively maintained in the CBTi group in the previous study [10]. An internet-based CBTi program also showed sustained treatment effects on ISI scores, with a notable effect of continuously decreasing ISI scores observed at the 1-year follow-up [33]. That study did not include a relaxation session; instead, the focus was on other components of CBTi, such as sleep restriction, stimulus control, cognitive restructuring, and sleep hygiene in 6 sessions [33]. This suggests that these components, aside from relaxation therapy, play a critical role in maintaining long-term treatment effects. If we had added a session to reinforce the cognitive component [18], thus providing 7 sessions similar to a previous study on CBTi [10], the effects of CBTi may have been better sustained. In addition, our active comparator app included SHE, providing the same content as Somzz. The education had some treatment effect on insomnia severity in the SHE group, which may have obscured the differences in ISI scores between the 2 groups at the 4-month follow-up. In addition, the waxing or waning nature of insomnia disorder may have contributed to a slight insomnia improvement at the 4-month follow-up in the SHE group.

Furthermore, the effectiveness of the Somzz app could have been confounded by the natural progression of insomnia and the Hawthorne effect. A population-based study reported cumulative remission rates for insomnia disorder among good sleepers of 13.4% after 1 month and 14.2% after 3 months [36] compared with our rate of 45% (22/49) at 6 weeks after the intervention. However, because the group classification criteria differed between our study and that previous study, the results should be compared with caution. Another study investigating the natural history of insomnia over 3 years reported that 74% and 46% of the participants had persistent insomnia for 1 and 3 years, respectively [37]. Although the follow-up period in our study was shorter, and considering that insomnia symptoms may wax and wane, it can be inferred that insomnia disorder can persist without treatment. Parallel RCTs including a third group with no intervention would be optimal for excluding any possible impact of the natural progression of insomnia or the Hawthorne effect on the outcomes of Somzz.

Our study found that the Somzz group had better outcomes in sleep efficiency, WASO, and satisfaction related to sleep based on sleep diary measures compared to the SHE group, which aligns with the results of previous studies [10,33]. Daytime sleepiness measured using the ESS was also improved in the Somzz group, with a significant group difference compared to the SHE group at the postintervention time point. The Somzz group also showed treatment effects on depression and quality of life at the postintervention time point, again consistent with the results of previous studies [38-40]. Although the scores on the PHQ-9 and SF-36 were lower than those at the baseline at the 4-month follow-up, the scores showed a worsening trend in the Somzz group from the postintervention time point to the 4-month follow-up. This trend may be correlated with increased insomnia severity after the intervention, and the correlation can be attributed to the well-established bidirectional interaction of insomnia with depression and quality of life [40]. Similar results were noted in a web-based CBTi study, which showed a treatment effect on depression and suicidal ideation at week 6 but not at the 6-month follow-up when measured using the psychiatric symptom frequency scale [38].

In addition, the Somzz group had significantly lower scores on the DBAS than the SHE group, and this improvement was maintained at the 4-month follow-up. Moreover, higher DBAS scores at baseline were a significant predictor of nonremission. Dysfunctional beliefs about sleep are a crucial factor in perpetuating insomnia as they can provoke a fear of losing control over sleep, worry about the negative consequences of insomnia, and feelings of hopelessness and anxiety [23]. This, in turn, could potentially prolong insomnia symptoms. The results suggest that targeting dysfunctional beliefs may be crucial in effectively treating insomnia using CBTi. Furthermore, a shorter total TIB at baseline was a significant predictor of remission, highlighting the importance of sleep restriction through TIB prescription as a crucial intervention for treating insomnia.

Our study had an attrition rate of 12% (6/49) at the postintervention time point among participants randomized to the Somzz group. This shows relatively good compliance compared to other delivery forms of digital CBTi, with an attrition rate of 21.6% (SD 16.9%) [13], and face-to-face CBTi, with a dropout rate of 14% to 40% [41]. This suggests that the use of MCBTi has the potential to decrease the prevalence of insomnia through high compliance with treatment.

Limitations

There are several limitations to our study. First, we conducted follow-up for 4 months (18 weeks), which is a relatively short time to infer the long-term effects of Somzz. Second, the impact of Somzz might have been modulated by the Hawthorne effect or the natural progression of insomnia. However, we compared Somzz with SHE in an effort to minimize any influence of such effects. A parallel RCT including a third group with no intervention and with a longer follow-up period and larger sample size is needed to validate our results. Finally, we only used subjective measures of sleep in our study. However,

XSL•FO
RenderX

subjective sleep measures including sleep diaries are known to be reliable and valid measures of sleep and correlate with objective measures of sleep [42].

Conclusions

Our study provides strong evidence that MCBTi can effectively treat insomnia by improving sleep-related variables, depression, and social functioning. MCBTi provides a highly accessible, time-efficient treatment option with high compliance that is suitable for industrious society.

Acknowledgments

This study was supported by the Technology Innovation Infrastructure program through the Korea Institute for Advancement of Technology funded by the Ministry of Trade, Industry, and Energy (grant 20009210 [awarded to YJL]).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HJL, SJK, and YJL conceived the idea, designed and supervised the study, and critically revised the manuscript. HJL, SJK, and YJL contributed equally as corresponding authors. YJL acquired funding. JS handled data analysis, interpreted the results, and drafted and revised the manuscript. SK, JL, HG, JA, CP, MS, JEJ, HYL, JWY, SK, and YY collected and organized patient data, interpreted the results, and critically revised the manuscript. All authors have read and approved the final manuscript for submission. For correspondence, please contact HJL (leehjeong@korea.ac.kr), SJK (psychical@hanmail.net), or YJL (ewpsyche@snu.ac.kr).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prespecified analysis plan.
[DOCX File , 14 KB - jmir_v26i1e50555_app1.docx]

Multimedia Appendix 2
Sleep diary outcomes compared between the Somzz and sleep hygiene education groups.
[DOCX File , 16 KB - jmir_v26i1e50555_app2.docx]

Multimedia Appendix 3
CONSORT-EHEALTH checklist (v1.6.1).
[PDF File (Adobe PDF File), 1161 KB - jmir_v26i1e50555_app3.pdf]

References

1. Aernout E, Benradia I, Hazo JB, Sy A, Askevis-Leherpeux F, Sebbane D, et al. International study of the prevalence and factors associated with insomnia in the general population. Sleep Med 2021 Jun;82:186-192. [doi: 10.1016/j.sleep.2021.03.028] [Medline: 33957414]
2. Kyle SD, Morgan K, Espie CA. Insomnia and health-related quality of life. Sleep Med Rev 2010 Feb;14(1):69-82. [doi: 10.1016/j.smrv.2009.07.004] [Medline: 19962922]
3. Ford DE, Kamerow DB. Epidemiologic study of sleep disturbances and psychiatric disorders. An opportunity for prevention? JAMA 1989 Sep 15;262(11):1479-1484. [doi: 10.1001/jama.262.11.1479] [Medline: 2769898]
4. Taylor DJ, Lichstein KL, Durrence HH. Insomnia as a health risk factor. Behav Sleep Med 2003;1(4):227-247. [doi: 10.1207/S15402010BSM0104_5] [Medline: 15600216]
5. Edinger JD, Arnedt JT, Bertisch SM, Carney CE, Harrington JJ, Lichstein KL, et al. Behavioral and psychological treatments for chronic insomnia disorder in adults: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med 2021 Feb 01;17(2):255-262 [FREE Full text] [doi: 10.5664/jcsm.8986] [Medline: 33164742]
6. Choi H, Youn S, Um YH, Kim TW, Ju G, Lee HJ, et al. Korean clinical practice guideline for the diagnosis and treatment of insomnia in adults. Psychiatry Investig 2020 Nov;17(11):1048-1059 [FREE Full text] [doi: 10.30773/pi.2020.0146] [Medline: 33198436]
7. Riemann D, Espie CA, Altena E, Arnardottir ES, Baglioni C, Bassetti CL, et al. The European Insomnia Guideline: an update on the diagnosis and treatment of insomnia 2023. J Sleep Res 2023 Dec 28;32(6):e14035 [FREE Full text] [doi: 10.1111/jsr.14035] [Medline: 38016484]

XSL•FO
RenderX

8. Baglioni C, Altena E, Bjorvatn B, Blom K, Bothelius K, Devoto A, et al. The European Academy for Cognitive Behavioural Therapy for Insomnia: an initiative of the European Insomnia Network to promote implementation and dissemination of treatment. J Sleep Res 2020 Apr 19;29(2):e12967. [doi: 10.1111/jsr.12967] [Medline: 31856367]
9. Thomas A, Grandner M, Nowakowski S, Nesom G, Corbitt C, Perlis ML. Where are the behavioral sleep medicine providers and where are they needed? A geographic assessment. Behav Sleep Med 2016 May 09;14(6):687-698 [FREE Full text] [doi: 10.1080/15402002.2016.1173551] [Medline: 27159249]
10. Zhang C, Liu Y, Guo X, Liu Y, Shen Y, Ma J. Digital cognitive behavioral therapy for insomnia using a smartphone application in China: a pilot randomized clinical trial. JAMA Netw Open 2023 Mar 01;6(3):e234866 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.4866] [Medline: 36972049]
11. Hong JS, Wasden C, Han DH. Introduction of digital therapeutics. Comput Methods Programs Biomed 2021 Sep;209:106319. [doi: 10.1016/j.cmpb.2021.106319] [Medline: 34364181]
12. Hasan F, Tu YK, Yang CM, James Gordon C, Wu D, Lee HC, et al. Comparative efficacy of digital cognitive behavioral therapy for insomnia: a systematic review and network meta-analysis. Sleep Med Rev 2022 Feb;61:101567. [doi: 10.1016/j.smrv.2021.101567] [Medline: 34902820]
13. Soh HL, Ho RC, Ho CS, Tam WW. Efficacy of digital cognitive behavioural therapy for insomnia: a meta-analysis of randomised controlled trials. Sleep Med 2020 Nov;75:315-325. [doi: 10.1016/j.sleep.2020.08.020] [Medline: 32950013]
14. Chen YX, Hung YP, Chen HC. Mobile application-assisted cognitive behavioral therapy for insomnia in an older adult. Telemed J E Health 2016 Apr;22(4):332-334. [doi: 10.1089/tmj.2015.0064] [Medline: 26393902]
15. Koffel E, Kuhn E, Petsoulis N, Erbes CR, Anders S, Hoffman JE, et al. A randomized controlled pilot study of CBT-I Coach: feasibility, acceptability, and potential impact of a mobile phone application for patients in cognitive behavioral therapy for insomnia. Health Informatics J 2018 Mar;24(1):3-13 [FREE Full text] [doi: 10.1177/1460458216656472] [Medline: 27354394]
16. Horsch CH, Lancee J, Griffioen-Both F, Spruit S, Fitrianie S, Neerincx MA, et al. Mobile phone-delivered cognitive behavioral therapy for insomnia: a randomized waitlist controlled trial. J Med Internet Res 2017 Apr 11;19(4):e70 [FREE Full text] [doi: 10.2196/jmir.6524] [Medline: 28400355]
17. Watanabe Y, Kuroki T, Ichikawa D, Ozone M, Uchimura N, Ueno T. Effect of smartphone-based cognitive behavioral therapy app on insomnia: a randomized, double-blind study. Sleep 2023 Mar 09;46(3):zsac270. [doi: 10.1093/sleep/zsac270] [Medline: 36355920]
18. Furukawa Y, Sakata M, Yamamoto R, Nakajima S, Kikuchi S, Inoue M, et al. Components and delivery formats of cognitive behavioral therapy for chronic insomnia in adults: a systematic review and component network meta-analysis. JAMA Psychiatry 2024 Apr 01;81(4):357-365 [FREE Full text] [doi: 10.1001/jamapsychiatry.2023.5060] [Medline: 38231522]
19. 510(k) Premarket notification. U.S. Food & Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K191716 [accessed 2024-06-10]
20. Vedaa Ø, Kallestad H, Scott J, Smith OR, Pallesen S, Morken G, et al. Effects of digital cognitive behavioural therapy for insomnia on insomnia severity: a large-scale randomised controlled trial. Lancet Digit Health 2020 Aug;2(8):e397-e406 [FREE Full text] [doi: 10.1016/S2589-7500(20)30135-7] [Medline: 33328044]
21. Morin CM, Belleville G, Bélanger L, Ivers H. The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. Sleep 2011 May 01;34(5):601-608 [FREE Full text] [doi: 10.1093/sleep/34.5.601] [Medline: 21532953]
22. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991 Dec;14(6):540-545. [doi: 10.1093/sleep/14.6.540] [Medline: 1798888]
23. Morin CM, Stone J, Trinkle D, Mercer J, Remsberg S. Dysfunctional beliefs and attitudes about sleep among older adults with and without insomnia complaints. Psychol Aging 1993 Sep;8(3):463-467. [doi: 10.1037//0882-7974.8.3.463] [Medline: 8216967]
24. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
25. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
26. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989 Oct;46(10):1121-1123. [doi: 10.1001/archneur.1989.00520460115022] [Medline: 2803071]
27. Ware JEJ. SF-36 health survey update. Spine (Phila Pa 1976) 2000 Dec 15;25(24):3130-3139. [doi: 10.1097/00007632-200012150-00008] [Medline: 11124729]
28. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736 [FREE Full text] [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
29. Bates D, Maechler M, Bolker B, Walker S, Christensen RH, Singmann H, et al. lme4: linear mixed-effects models using 'Eigen' and S4. The Comprehensive R Archive Network. 2024 Jun 19. URL: https://cran.r-project.org/web/packages/lme4/index.html [accessed 2024-06-20]

XSL•FO
RenderX

30. Prescription Digital Therapeutic for the Treatment of Insomnia (SLEEP-I). National Library of Medicine. 2024 Mar 29. URL: https://classic.clinicaltrials.gov/ct2/show/results/NCT04909229 [WebCite Cache ID https://classic.clinicaltrials.gov/ct2/show/results/NCT04909229]
31. Herber R, Graehlert X, Raiskup F, Veselá M, Pillunat LE, Spoerl E. Statistical Evaluation of Correlated Measurement Data in Longitudinal Setting Based on Bilateral Corneal Cross-Linking. In: Curr Eye Res Jul 2022:995-1002.
32. Morin CM, Benca R. Chronic insomnia. Lancet 2012 Mar 24;379(9821):1129-1141. [doi: 10.1016/S0140-6736(11)60750-2] [Medline: 22265700]
33. Ritterband LM, Thorndike FP, Ingersoll KS, Lord HR, Gonder-Frederick L, Frederick C, et al. Effect of a web-based cognitive behavior therapy for insomnia intervention with 1-year follow-up: a randomized clinical trial. JAMA Psychiatry 2017 Jan 01;74(1):68-75. [doi: 10.1001/jamapsychiatry.2016.3249] [Medline: 27902836]
34. Morin CM, Culbert JP, Schwartz SM. Nonpharmacological interventions for insomnia: a meta-analysis of treatment efficacy. Am J Psychiatry 1994 Aug;151(8):1172-1180. [doi: 10.1176/ajp.151.8.1172] [Medline: 8037252]
35. Adair JG. The Hawthorne effect: a reconsideration of the methodological artifact. J Appl Psychol 1984 May;69(2):334-345. [doi: 10.1037/0021-9010.69.2.334]
36. Morin CM, Leblanc M, Ivers H, Bélanger L, Mérette C, Savard J, et al. Monthly fluctuations of insomnia symptoms in a population-based sample. Sleep 2014 Feb 01;37(2):319-326 [FREE Full text] [doi: 10.5665/sleep.3406] [Medline: 24497660]
37. Morin CM, Bélanger L, LeBlanc M, Ivers H, Savard J, Espie CA, et al. The natural history of insomnia: a population-based 3-year longitudinal study. Arch Intern Med 2009 Mar 09;169(5):447-453. [doi: 10.1001/archinternmed.2008.610] [Medline: 19273774]
38. Christensen H, Batterham PJ, Gosling JA, Ritterband LM, Griffiths KM, Thorndike FP, et al. Effectiveness of an online insomnia program (SHUTi) for prevention of depressive episodes (the GoodNight Study): a randomised controlled trial. Lancet Psychiatry 2016 Apr;3(4):333-341. [doi: 10.1016/S2215-0366(15)00536-2] [Medline: 26827250]
39. Morin C, Thorndike FP, Ojile JM, Gerwien R, Wendorf A, Maricich YA. Digital CBT-I treatment improves sleep and reduces anxiety and depression symptoms in adults with chronic insomnia: interim analysis of DREAM study. CNS Spectr 2023 Apr 14;28(2):227-228. [doi: 10.1017/s1092852923001505]
40. Bai W, Zhao YJ, Cai H, Sha S, Zhang Q, Lei SM, et al. Network analysis of depression, anxiety, insomnia and quality of life among Macau residents during the COVID-19 pandemic. J Affect Disord 2022 Aug 15;311:181-188 [FREE Full text] [doi: 10.1016/j.jad.2022.05.061] [Medline: 35594975]
41. Matthews EE, Arnedt JT, McCarthy MS, Cuddihy LJ, Aloia MS. Adherence to cognitive behavioral therapy for insomnia: a systematic review. Sleep Med Rev 2013 Dec;17(6):453-464 [FREE Full text] [doi: 10.1016/j.smrv.2013.01.001] [Medline: 23602124]
42. Morin CM, Colecchi C, Stone J, Sood R, Brink D. Behavioral and pharmacological therapies for late-life insomnia: a randomized controlled trial. JAMA 1999 Mar 17;281(11):991-999. [doi: 10.1001/jama.281.11.991] [Medline: 10086433]

Abbreviations

CBTi: cognitive behavioral therapy for insomnia
DBAS: Dysfunctional Beliefs and Attitudes About Sleep Scale
ESS: Epworth Sleepiness Scale
FSS: Fatigue Severity Scale
ISI: Insomnia Severity Index
MCBTi: mobile app–based digital cognitive behavioral therapy for insomnia
PHQ-9: Patient Health Questionnaire–9
RCT: randomized controlled trial
SF-36: 36-item Short Form Health Survey
SHE: sleep hygiene education
SOL: sleep onset latency
TIB: time in bed
WASO: wake after sleep onset

XSL•FO
RenderX

Edited by YH Lin; submitted 06.07.23; peer-reviewed by K Coteur, M Sakata; comments to author 29.02.24; revised version received 25.04.24; accepted 22.05.24; published 26.07.24.

Please cite as:
Shin J, Kim S, Lee J, Gu H, Ahn J, Park C, Seo M, Jeon JE, Lee HY, Yeom JW, Kim S, Yoon Y, Lee HJ, Kim SJ, Lee YJ
Efficacy of Mobile App–Based Cognitive Behavioral Therapy for Insomnia: Multicenter, Single-Blind Randomized Clinical Trial
J Med Internet Res 2024;26:e50555
URL: https://www.jmir.org/2024/1/e50555
doi:10.2196/50555
PMID:

©Jiyoon Shin, Sujin Kim, Jooyoung Lee, Hyerin Gu, Jihye Ahn, Chowon Park, Mincheol Seo, Jeong Eun Jeon, Ha Young Lee, Ji Won Yeom, Sojeong Kim, Yeaseul Yoon, Heon-Jeong Lee, Seog Ju Kim, Yu Jin Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Remote Exercise Training Intervention During the COVID-19 Pandemic: Randomized Controlled Trial

Keito F A Philippi[1*], MSc; Peter Zeier[2*], PhD; Alexandra Brahmer[1], Dr rer nat; Elmo W I Neuberger[1], PhD; Magdalena Sandner[2], PhD; Matthias Hagenah[1]; Thilo Porten[1]; Regina Lenz[1], BSc; David T Ochmann[1], MSc; Florian Wedekink[3], MSc; Jörg Wischhusen[3], Prof Dr; Beat Lutz[4,5], Prof Dr; Klaus Lieb[5,6], Prof Dr; Michèle Wessa[2,5*], Prof Dr; Perikles Simon[1*], Prof Dr, Dr med

[1]Department of Sports Medicine, Disease Prevention and Rehabilitation, Institute of Sports Science, Johannes Gutenberg-University Mainz, Mainz, Germany

[2]Department of Clinical Psychology and Neuropsychology, Institute for Psychology, Johannes Gutenberg-University Mainz, Mainz, Germany

[3]Department of Obstetrics and Gynecology—Experimental Tumor Immunology, University Clinics Würzburg, Würzburg, Germany

[4]Institute of Physiological Chemistry, University Medical Center, Johannes Gutenberg-University Mainz, Mainz, Germany

[5]Leibniz Institute for Resilience Research, Mainz, Germany

[6]Department of Psychiatry and Psychotherapy, University Medical Center, Johannes Gutenberg-University Mainz, Mainz, Germany

[*]these authors contributed equally

Corresponding Author:
Perikles Simon, Prof Dr, Dr med
Department of Sports Medicine, Disease Prevention and Rehabilitation
Institute of Sports Science
Johannes Gutenberg-University Mainz
Albert-Schweitzer-Straße 22
Mainz, 55128
Germany
Phone: 49 6131 3923531
Email: simonpe@uni-mainz.de

Abstract

Background: Societal measures to contain the spread of COVID-19 (eg, lockdown and contact restrictions) have been associated with decreased health and well-being. A multitude of prepandemic studies identified the beneficial effects of physical exercise on both physical and mental health.

Objective: We report on the feasibility of a remote physical exercise intervention and its stress-buffering potential in 2 untrained cohorts: a pre–COVID-19 cohort that completed the intervention in 2019 and a lockdown cohort that started the intervention shortly before pandemic-related restrictions were implemented.

Methods: In a randomized controlled trial, participants were assigned to either an intervention group (IG; pre–COVID-19 cohort: n=7 and lockdown cohort: n=9) or a control group (CG; pre–COVID-19 cohort: n=6 and lockdown cohort: n=6). IG participants received weekly individualized training recommendations delivered via web-based support. The intervention period was initially planned for 8 weeks, which was adhered to in the pre–COVID-19 cohort (mean 8.3, SD 0.5 weeks) but was extended to an average of 17.7 (SD 2.0) weeks in the lockdown cohort. Participants' health parameters were assessed before and after the intervention: aerobic capacity was measured as peak oxygen uptake (VO_{2peak}) via cardiopulmonary exercise testing. Depressive symptoms were scored via the depression subscale of the Brief Symptom Inventory-18.

Results: Dropout rates were low in both cohorts in the IG (pre–COVID-19 cohort: n=0, 0% and lockdown cohort: n=2, 16.7%) and the CG (pre–COVID-19 cohort: n=0, 0% and lockdown cohort: n=2, 20%). The mean adherence to the training sessions of the IG for both cohorts was 84% (pre–COVID-19 cohort: SD 5.5% and lockdown cohort: SD 11.6%). Aligned rank transform ANOVAs in the lockdown cohort indicated deterioration of VO_{2peak} and depressive symptoms from before to after the intervention in the CG but no longitudinal changes in the IG. Analyses in the pre–COVID-19 cohort revealed significant increases in VO_{2peak} for the IG compared to the CG (P=.04) but no intervention effects on depressive symptoms.

XSL•FO
RenderX

Conclusions: With low dropout rates and high adherence, the remote intervention was feasible for healthy adults under regular conditions and in the face of pandemic-related stressors. Moreover, our results hint at a stress-buffering effect as well as a buffering of a lockdown-induced deconditioning of remote physical exercise interventions in the pandemic scenario, which can be used in future studies to overcome equally stressful periods of life. However, due to limited statistical power, these findings should be replicated in similar scenarios.

Trial Registration: German Clinical Trials Register DRKS00018078; https://drks.de/search/en/trial/DRKS00018078

(J Med Internet Res 2024;26:e53145) doi:10.2196/53145

KEYWORDS

physical exercise; web-based; COVID-19; lockdown; aerobic capacity; depression; feasibility; aerobic; anxiety; randomized controlled trial; prepandemic cohort; mental health; pandemic; remote exercise; intervention; pre-COVID-19; lockdown cohort; pandemic-related; adult; adults

Introduction

In March 2020, infections with SARS-CoV-2 and resulting cases of COVID-19 increased sharply in Germany and elsewhere. While the World Health Organization declared the COVID-19 outbreak as a pandemic on March 11, a first peak was reached in Germany only 3 weeks later, with about 36,000 reported cases in calendar week 14, 2020. The incidence leveled off again the following summer [1,2]. One day before the World Health Organization declared the pandemic situation, all events with over 1000 attendants were canceled in Germany. Other measures such as closing daycare centers and schools, a countrywide lockdown with restaurant closures, and extensive contact restrictions followed shortly thereafter [1].

These measures affected many people in a wide variety of situations. Even though there are findings of an increase in both physical activity and sedentary behavior in a group of university students [3], a systematic review of 64 studies with over 86,000 participants [4], however, indicates that the majority of studies [5,6] report a decrease in physical activity and an increase in sedentary behavior during lockdown periods. When addressing changes in aerobic capacity potentially induced by lockdown measures, the findings are similar. Previous studies show a decrease in aerobic capacity during the COVID-19 pandemic in different subgroups of both soccer and handball players as well as high school adolescents [7-12]. To our knowledge, there are no data regarding changes in aerobic capacity in healthy untrained adults. Aerobic capacity has been shown to be a good predictor of all-cause mortality [13-16]. Consequently, a decrease in aerobic capacity may be related to an increase in mortality rates. Improving aerobic capacity via endurance training is a possible way to counteract this deterioration [17].

Multiple studies indicated the detrimental effects of the COVID-19 pandemic on both mental and physical health. A systematic review [18] reported COVID-19–related increases in the global prevalence of major depressive disorders and anxiety disorders of 27.6% and 25.6%, respectively. Interestingly, these increases were most salient in locations with decreased human mobility. In accordance with prepandemic research [19], decreased physical activity during COVID-19 lockdown periods was significantly correlated with impaired mental health [20,21]. Conversely, physical exercise was proposed to counteract the COVID-19–related detrimental effects on mental health, as described in a systematic overview of the literature [22]. This recommendation is largely based on prepandemic research linking physical exercise to improved psychological well-being, for example, reduced levels of depression [23,24], anxiety [25], and stress reactivity [26]. Additionally, research on human resilience, that is, an individual's ability to maintain or regain health in the face of adversity [27], revealed the stress-buffering potential of physical fitness via biological pathways, for example, optimizing hormonal stress response systems and minimizing excessive inflammation [28], and by psychological effects, for example, increasing self-efficacy [29].

Web-based interventions have been shown to be successful in increasing physical activity levels in adults when compared with waiting list control groups (CGs) [30-32]. Various studies with the same general intervention design as this study have proven to be both applicable and effective in increasing aerobic capacity in different patient groups [33-38]. Approaches that address the combination of web-based patient empowerment, as the ability of patients to positively influence their health and health behavior, as well as physical activity, show beneficial effects on patients with various chronic diseases [39].

In 2019, we started a randomized controlled trial that addressed the association of innate and acquired aerobic capacity with resilience in physically inactive healthy adults implementing remote exercise training for 8 weeks [40]. A first set of participants (pre–COVID-19 cohort; Figure 1) successfully completed the study protocol in 2019. We conducted preintervention assessments of a second set of participants (lockdown cohort) before any COVID-19–related restrictions were implemented. However, due to local restrictions, the original protocol could not have been pursued as planned in the lockdown cohort. We thus adapted the exercise intervention to investigate the stress-buffering potential of physical exercise when facing a naturalistic stressor, that is, pandemic-related restrictions. Here, we (1) report on the feasibility of a remote exercise intervention during the pandemic situation and (2) describe the impact of physical training on physical and mental health for both the pre–COVID-19 and lockdown cohorts. In accordance with previous research on physical exercise interventions, we expect a beneficial impact on both health parameters. More specifically, we hypothesize an increase in aerobic capacity (peak oxygen uptake [VO_{2peak}]) and a decrease in depressive symptoms in individuals receiving remote exercise training compared to a passive CG for the pre–COVID-19

XSL•FO
RenderX

cohort. For the lockdown cohort, we hypothesize a smaller decrease in VO_{2peak} and a smaller increase in depressive symptoms in physically active individuals compared to passive controls, as the exercise training would likely counteract the detrimental effects of COVID-19–related measures, which may be caused by a more sedentary daily routine.

Figure 1. (A) Original plan for IG and CG from T0 to T3. (B) Timeline of the individual participants from T0 to T3. CG: control group; IG: intervention group.

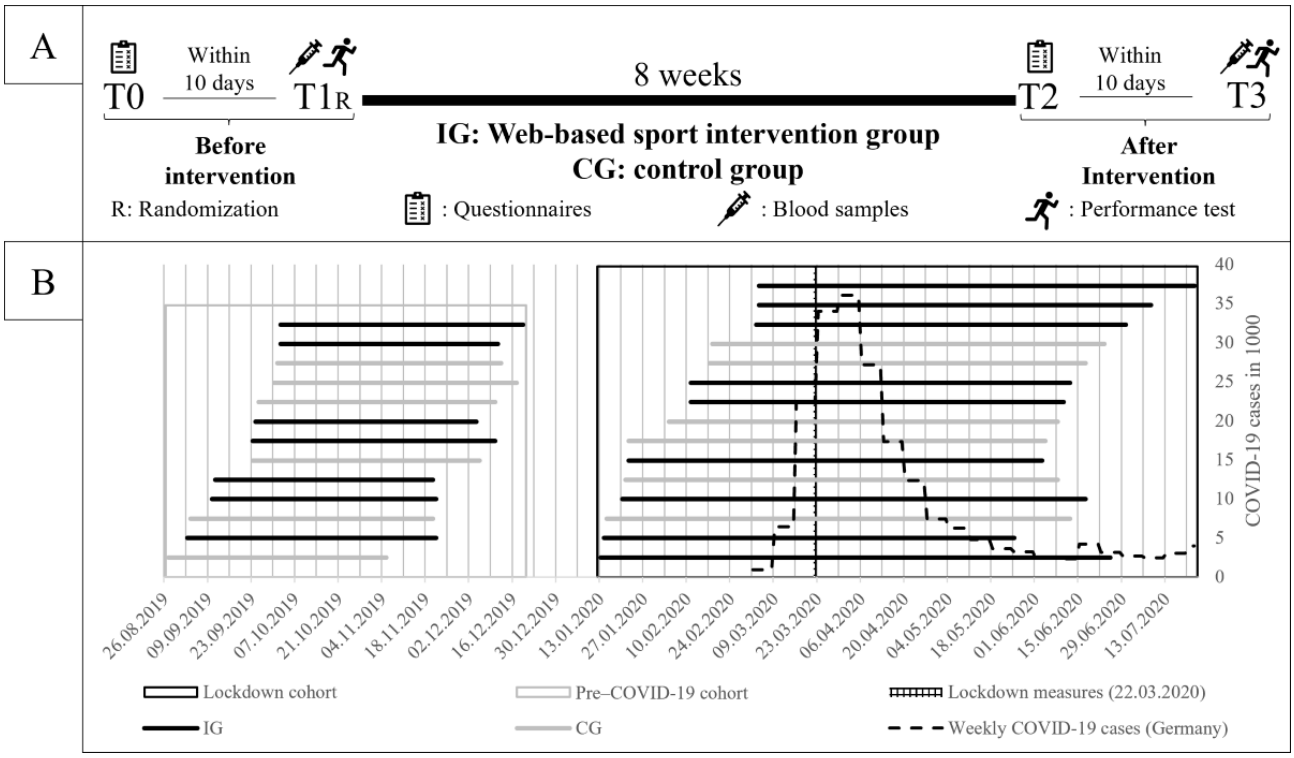

Methods

Overview

The methodology of the study was previously described in the frame of the prospective randomized controlled trial, published in a study protocol by Ochmann et al [40]. The protocol adheres to the guidelines from the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) and the CONSORT (Consolidated Standards of Reporting Trials)–eHEALTH checklists (Multimedia Appendix 1). With the outbreak of the COVID-19 pandemic, the lockdown measures introduced in Germany in March 2020 prevented the continuation of the regular study procedure as well as further enrollment of participants. To still be able to provide sports therapy for the participants included at the beginning of 2020, the initial study protocol had to be adapted, and in the following sections, mainly these amendments are described.

Participants and Study Design

The study design, illustrated in Figure 1A, included examinations before the intervention (T0: psychological assessments and T1: physiological assessments), an intervention period, and final examinations (T2: psychological assessments and T3: physiological assessments). Healthy untrained (less than 120 minutes per week being physically active) participants aged between 18 and 45 years were recruited via flyers distributed at the Johannes Gutenberg–University of Mainz and the University Medical Center campuses (for a detailed overview of inclusion and exclusion criteria, see Ochmann et al [40]). A total of 35 participants who provided written informed consent were randomly assigned to an intervention group (IG) or CG and completed the assessments before the intervention. This study includes data from a pre–COVID-19 cohort (2019) and a lockdown cohort (2020). Their difference in methodology relates mainly to the longer intervention period for the lockdown cohort, as shown in Figure 1B. A total of 7 participants had to be excluded from all analyses: 3 of them were excluded due to incomplete data sets and 4 participants dropped out of the study (for more details, see the *Results* section). This leads to a final sample of 13 participants in the pre–COVID-19 cohort (IG: n=7 and CG: n=6) and 15 participants in the lockdown cohort (IG: n=9 and CG: n=6). Further descriptions of the final study population can be found in the *Results* section.

Training Intervention

Weekly training in the IG consisted of progressive continuous and interval runs with the goal of gradually increasing aerobic capacity (for details, see Ochmann et al [40]). The trainer individually adapted the training to the weekly feedback provided by the participant via web-based support. As we did not receive any reports regarding major technical problems from the study participants, we assume that the technical use of web-based support was not a challenge from the perspective of the test participants. However, if more intensive clarification was required, a telephone consultation was arranged. After verifying that the participant complied with the training recommendations, the 2 factors of perceived ailment and load mainly determined the adjustments. Since the original training

XSL•FO
RenderX

progression had been designed for an 8-week study period, this design could not cover an indefinite period. Therefore, the trainer mainly focused on the continuation of the participants' sporting activities. As originally planned, the principle of training adjustment was also applied to the lockdown cohort. Thus, depending on the perceived ailment and load, the training was reduced, maintained, or intensified.

Measures

The German version of the Brief Symptom Inventory-18 (BSI-18) [41,42] was applied to measure psychological distress. The BSI-18 contains 3 subscales, that is, depression, anxiety, and somatization, which are assessed as sum scores of 6 items each, as well as a Global Severity Index calculated as a sum across all 18 items. All participants worked twice on the BSI-18, once before the intervention period (T0) and once after the period (T2). In accordance with our hypotheses, we limited our main analyses to the depression subscale of the BSI-18.

VO_{2peak} was determined as a main indicator of aerobic capacity. It was chosen over VO_{2max} as the main parameter since we did not expect the occurrence of a plateau in testing untrained participants [43]. VO_{2peak} was measured during a stepwise incremental running test starting at 4 km/h and increasing velocity every 3 minutes by 1.5 km/h (for further details, see the initial study protocol by Ochmann et al [40]). The determination took place once before the intervention period (T1) and once after the period (T3). Training adherence was calculated by dividing documented training sessions by recommended training sessions.

Test for SARS-CoV-2 Seroconversion

Participants in the lockdown cohort were tested for SARS-CoV-2 seroconversion to exclude an undetected infection via the enzyme-linked immunosorbent assay described by Notz et al [44]. Venous blood was drawn in tripotassium-ethylenediaminetetraacetic acid–covered Monovettes (Sarstedt) and was processed within 5 minutes after blood collection. Platelet-free plasma was prepared by 2 rounds of centrifugation for 15 minutes at 2500×g and room temperature. Plasma was aliquoted and stored at –80 °C until enzyme-linked immunosorbent assay measurement.

At T1, all participants showed no sign of seroconversion, as expected prior to the pandemic situation. At T3, seroconversion of 2 participants, who had already stated a SARS-CoV-2 infection during the intervention phase, was confirmed. Medical anamnesis as well as pulmonary function and spiroergometric tests ensured that these participants did not experience any aftermaths affecting their ability to participate in the study. Besides this, no seroconversions or stated SARS-CoV-2 infections were reported. Therefore, no participants were excluded due to the SARS-CoV-2 infection.

Statistical Analyses

SPSS (version 23; IBM Corp) was used for data analysis. Aligned rank transform (ART) ANOVAs [45] were conducted for each cohort separately to test our hypotheses, with time (before vs after the intervention) as a within-subject factor, group (IG vs CG) as a between-subject factor, and VO_{2peak} and depression scores as dependent variables. Follow-up Wilcoxon tests were performed to contrast specific time- and group-related effects.

Ethical Considerations

The study protocol follows the standards of the Declaration of Helsinki of the World Medical Association, and ethics approval was obtained by the medical association Rhineland-Palatinate (July 29, 2019; IRB approval 2019-14305). The original study was registered with the German Clinical Trials Register (DRKS00018078; October 2, 2019). All participants provided informed consent before participation and were informed about the ability to opt out of the study at any time. Furthermore, all data were pseudonymized. As compensation for complete participation, 15 smartwatches (M430, Polar Electro Oy) were raffled among all participants.

Results

Study Population and Feasibility of the Randomized Controlled Exercise Intervention

Due to the unexpected occurrence of the COVID-19 pandemic and its related restrictions, we have the unique opportunity to report on the feasibility of the conduction of a randomized controlled remote exercise intervention under these circumstances. Whereas a pre–COVID-19 cohort was able to proceed through the study protocol as planned, the intervention period and the final examinations of the lockdown cohort could not be conducted as planned. On March 16, 2020, all participants were notified about an indefinite extension of the intervention period. The local restrictions, in the form of the first lockdown, in Germany started on March 27, 2020. Figure 1B shows all participant timelines combined with a visualization of the weekly COVID-19 cases. Table 1 provides additional measurements describing the study population. Participants in the IG and CG did not show significant differences regarding VO_{2peak} (pre–COVID-19 cohort: P=.22 and lockdown cohort: P=.52) or depression scores (pre–COVID-19 cohort: P=.43 and lockdown cohort: P=.70) before the intervention.

XSL•FO
RenderX

Table 1. Description of the study population (N=28).

Characteristics	Pre–COVID-19 cohort			Lockdown cohort			Total
	CG[a]	IG[b]	Subtotal	CG	IG	Subtotal	
Sex, n (%)							
Female	3 (11)	5 (18)	8 (29)	4 (14)	5 (18)	9 (32)	17 (61)
Male	3 (11)	2 (7)	5 (18)	2 (7)	4 (14)	6 (21)	11 (39)
Age (years), mean (SD)	26.5 (9.5)	25.6 (3.9)	26.0 (6.7)	26.7 (4.6)	25.4 (5.9)	25.9 (5.3)	26.0 (5.9)
Weight (kg), mean (SD)	68.4 (13.2)	75.7 (11.1)	72.4 (12.2)	68.6 (8.4)	69.3 (9.5)	69.0 (8.8)	70.6 (10.4)
Height (cm), mean (SD)	172.2 (9.9)	173.9 (12.4)	173.1 (10.9)	170.0 (7.6)	175.4 (7.1)	173.3 (7.6)	173.2 (9.1)
BMI (kg/m^2), mean (SD)	23.2 (4.2)	25.0 (2.8)	24.2 (3.5)	23.9 (3.9)	22.5 (2.3)	23.0 (3.0)	23.6 (3.2)
Body fat (%), mean (SD)	21.5 (9.2)	26.7 (9.9)	24.3 (9.5)	22.0 (11.1)	18.0 (7.1)	19.6 (9.0)	21.8 (9.4)
Muscle mass (kg), mean (SD)	49.7 (7.9)	51.8 (9.1)	50.8 (8.3)	49.5 (3.7)	53.3 (8.9)	51.8 (7.3)	51.3 (7.7)
VO_{2peak}[c] (mL/min/kg), mean (SD)	40.4 (6.2)	35.0 (8.4)	37.5 (7.7)	37.6 (7.8)	39.9 (6.1)	39.0 (6.7)	38.3 (7.1)
Speed at IAT[d] (km/h), mean (SD)	7.8 (1.5)	7.7 (2.3)	7.7 (1.9)	6.7 (1.2)	8.5 (2.1)	7.8 (2.0)	7.8 (1.9)
Maximum heart rate (per minute), mean (SD)	192.8 (8.0)	196.0 (12.4)	194.5 (10.3)	193.3 (5.4)	201.0 (7.6)	197.9 (7.6)	196.4 (9.0)
BSI[e]—depression, mean (SD)	2.5 (1.2)	3.7 (3.4)	3.2 (2.6)	1.8 (1.7)	1.6 (1.0)	1.7 (1.3)	2.4 (2.1)
Total intervention weeks, mean (SD)	8.3 (0.5)	8.3 (0.5)	8.3 (0.5)	17.3 (1.0)	17.7 (2.0)	17.5 (1.6)	13.3 (4.8)
Adherence (%), mean (SD)	N/A[f]	84.0 (5.5)	N/A	N/A	84.0 (11.6)	N/A	84.0 (9.2)
Weekly training sessions, mean (SD)	N/A	2.3 (0.4)	N/A	N/A	2.3 (0.6)	N/A	2.3 (0.5)
Total training sessions, mean (SD)	N/A	18.6 (2.9)	N/A	N/A	41.3 (12.9)	N/A	31.4 (15.1)

[a]CG: control group.

[b]IG: intervention group.

[c]VO_{2peak}: peak oxygen uptake.

[d]IAT: individual anaerobic threshold.

[e]BSI: Brief Symptom Inventory.

[f]N/A: not applicable.

The duration of the intervention for the lockdown cohort between T1 and T2 increased on average to 17.3 (SD 1.0) weeks for the CG and 17.7 (SD 2.0) weeks for the IG. The intervention times for the pre–COVID-19 cohort in comparison lasted a mean of 8.3 (SD 0.5) weeks for the CG and 8.3 (SD 0.5) weeks for the IG. The mean number of training sessions per week, on the other hand, showed no difference (pre–COVID-19 cohort: mean 2.3, SD 0.4 and lockdown cohort: mean 2.3, SD 0.6). However, in the lockdown cohort, the training became more individualized the longer the training intervention lasted. While the framework of the original training progression was retained, the structure of the training sessions was adjusted to the individual demands. This led to a slightly higher effort for the trainer without increasing communication with the participants. For example, there were individuals who preferred to maintain a frequency of 3 units per week during the lockdown, while others preferred 5 units per week. Furthermore, in addition to endurance runs and interval units, combinations of moderate and vigorous intensities as well as Fartlek runs (individual pace and intensity variations during unstructured interval training) were included. Depending on how the participants accepted these training methods, they were recommended more or less frequently. However, the overall weekly volume was always maintained. Finally, the mean adherence, that is, the quotient of training sessions performed and recommended, did not differ between the cohorts (84% for both cohorts) but differed in the SD (pre–COVID-19 cohort: 5.5% and lockdown cohort: 11.6%; Table 1).

There were low dropout rates in both cohorts in the IG (pre–COVID-19 cohort: n=0, 0% and lockdown cohort: n=2, 16.7%) and the CG (pre–COVID-19 cohort: n=0, 0% and lockdown cohort: n=2, 20%). No participants in the pre–COVID-19 cohort and 4 participants in the lockdown cohort dropped out from the study. However, there was no study dropout specifically induced by the prolongation of the intervention period. In total, 2 participants dropped out of the CG due to a change of personal situation, and 2 participants dropped out of the IG due to medical reasons not related to the study or COVID-19. One of the latter participants dropped out 16 weeks after T1, and the other dropped out 4 weeks after T1 and before the lockdown started.

Taken together, the adaptation of the training intervention to the pandemic situation was well feasible, and the training

XSL•FO
RenderX

adherence as well as the dropout rate were adequate. Despite the small sample sizes induced by introducing a pre–COVID-19 cohort and a lockdown cohort, the study populations in the IG and the CG for both cohorts were highly similar.

Effects of Exercise Intervention on VO2peak and Depression

Since the conduction of our remote physical exercise intervention turned out feasible and the study population was highly comparable at T1, specifically regarding VO_{2peak} and depression scores, we investigated a potentially beneficial effect of the physical exercise intervention on both health variables in the next step. Since the interventions in the pre–COVID-19 and lockdown cohorts differed especially regarding training durations, we analyzed both cohorts separately. We performed nonparametric analyses to account for the small sample sizes in both cohorts. More specifically, ART [45] was conducted to investigate the main and interaction effects of group (intervention vs control) and time (before vs after the intervention) on depression and VO_{2peak}. Follow-up 1-sample Wilcoxon signed rank tests were performed to investigate changes in both health parameters within groups.

Regarding the pre–COVID-19 cohort, the ART ANOVA with VO_{2peak} as a dependent variable revealed nonsignificant main effects of time ($F_{1,11}$=0.98; P=.34; η^2=0.08) and group ($F_{1,11}$=0.01; P=.92; η^2<0.01) but a significant time×group interaction ($F_{1,11}$=5.27; P=.04; η^2=0.32). However, a second ART ANOVA on depression did not indicate significant main effects of time ($F_{1,11}$=0.45; P=.52; η^2=0.04) and group ($F_{1,11}$=0.18; P=.68; η^2=0.02) nor a significant time×group interaction effect ($F_{1,11}$=0.82; P=.39; η^2=0.07). One-sample Wilcoxon tests for each group separately indicated a significant increase in VO_{2peak} (before the intervention: mean 34.99, SD 8.43 and after the intervention: mean 37.94, SD 7.48; P=.04) in the IG (Figure 2).

Regarding the lockdown sample, the ART ANOVAs revealed significant time×group interaction effects on VO_{2peak} ($F_{1,13}$=11.44; P=.01; η^2=0.47) and a marginally significant interaction effect on depression ($F_{1,13}$=3.90; P=.07; η^2=0.23). The ART ANOVAs revealed no significant main effects of time or group on VO_{2peak} (time: $F_{1,13}$=0.04; P=.85; η^2<0.01 and group: $F_{1,13}$=0.09; P=.77; η^2=0.01) or depression scores (time: $F_{1,13}$<0.01; P=.98; η^2<0.01 and group: $F_{1,13}$=0.23; P=.64; η^2=0.02). As indicated in Figure 3, follow-up Wilcoxon tests indicated a significant decrease of VO_{2peak} (before the intervention: mean 37.58, SD 7.84 and after the intervention: mean 35.38, SD 8.16; P=.03) and a marginally significant increase in depression (before the intervention: mean 1.83, SD 1.72 and after the intervention: mean 3.67, SD 2.42; P=.08) in the CG.

Taken together, our findings indicate that while participants in the CG experienced a decrease in aerobic capacity and an increase in depression levels during the first COVID-19–induced lockdown, participants in the IG remained stable on both parameters. On the other hand, under normal conditions (pre–COVID-19 cohort), only the aerobic capacity increased significantly in the IG.

XSL•FO
RenderX

Figure 2. Results of Wilcoxon tests before and after the intervention for depression and VO_{2peak} in the pre–COVID-19 cohort. BSI: Brief Symptom Inventory; VO_{2peak}: peak oxygen uptake.

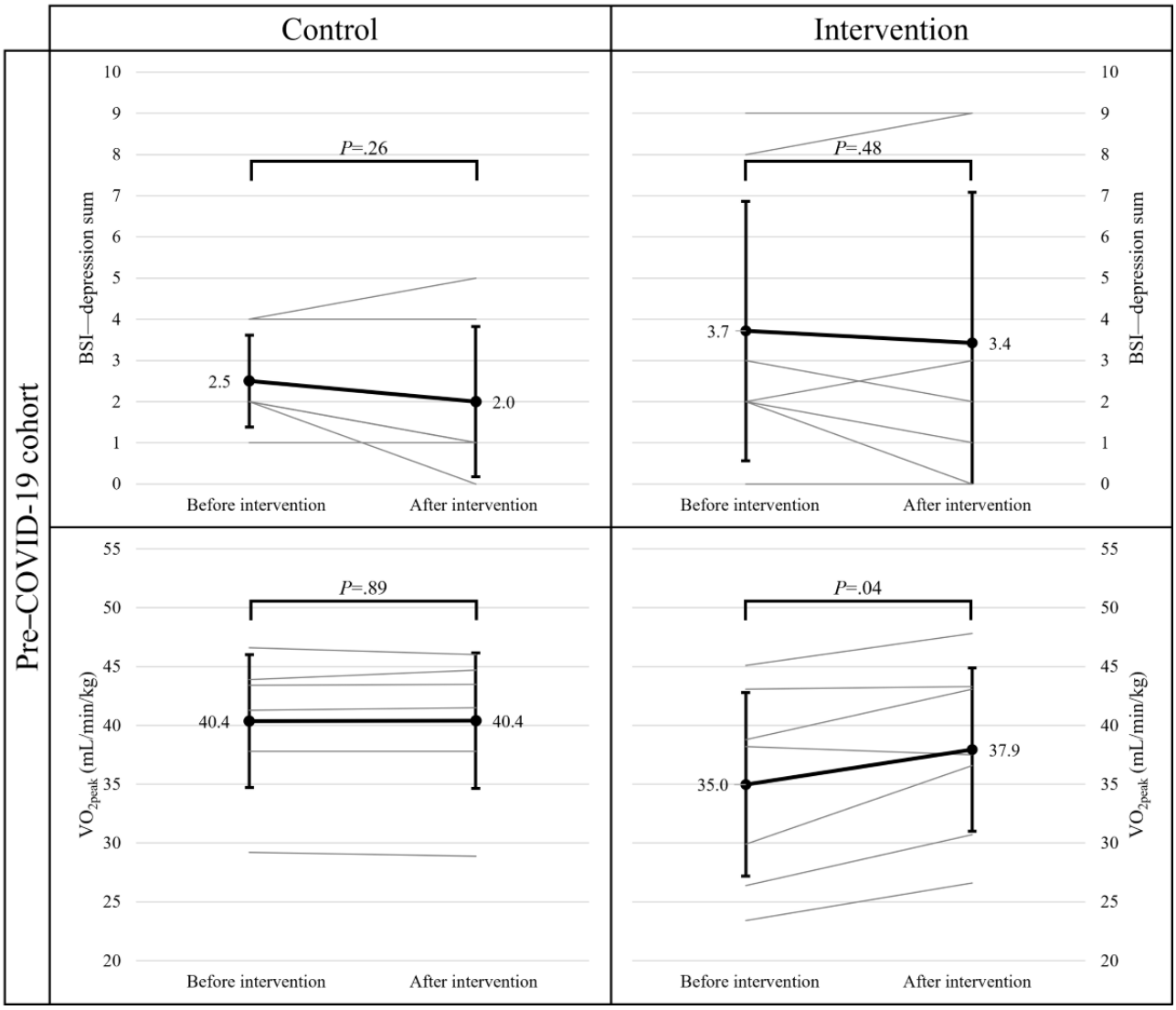

XSL•FO
RenderX

Figure 3. Results of Wilcoxon tests before and after the intervention for depression and VO_{2peak} in the lockdown cohort. BSI: Brief Symptom Inventory; VO_{2peak}: peak oxygen uptake.

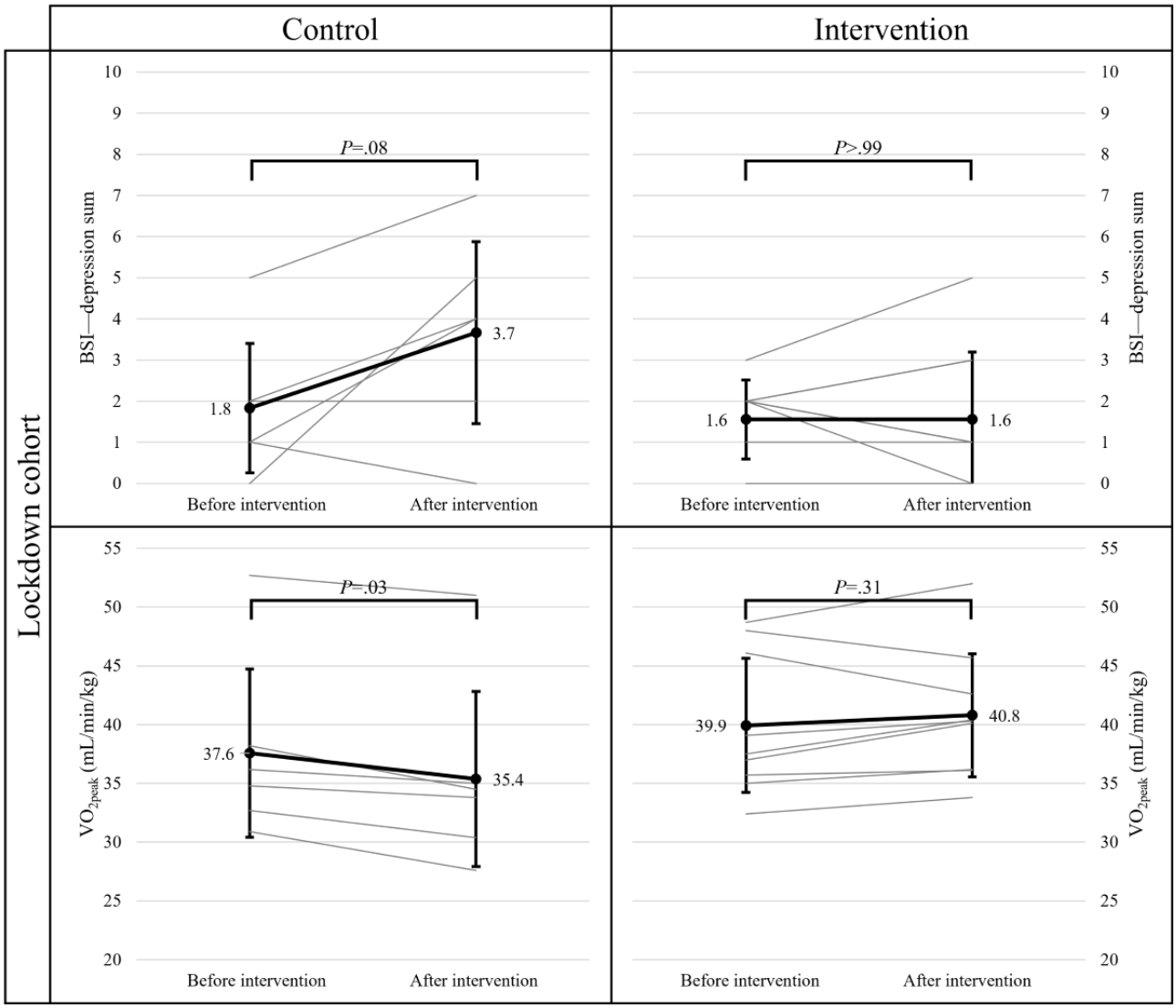

Discussion

Principal Findings

In this study, we had the unique chance to investigate the feasibility of a remote physical exercise intervention and, furthermore, its potentially beneficial effect on physical and mental health during a stress-inducing COVID-19 lockdown situation. With a low dropout rate and high adherence to the training sessions, the intervention was well feasible. Moreover, our results hint at the stress-buffering potential of the exercise intervention, although the statistical power was limited due to the small sample size.

The continuation of our study was well feasible in the prevailing pandemic situation mainly due to the use of a web-based training intervention. All training support structures implemented in this study procedure could be maintained as planned, and high adherence values and low dropout rates in the lockdown cohort were achieved. Most importantly, there was no need for physical contact, while the continuation of a program including such an approach would not have been possible. The training could simply continue as before since the trainer and participant were in contact exclusively via digital means. This remote contact allowed for fast support in case of problems, as there was no need to wait for the next training session. Changes that were necessary to keep the intervention running included the extension of the training plans from originally 8 weeks to a longer, undefined period, stronger exchange between participant and trainer, and more individualized training plans. Heart rate and other parameters were tracked continuously, and the trainer had a very detailed insight into the physiological variables during the workout, which facilitated personalization processes. Moreover, the individuals' training preferences and personal motivation could be easily considered through frequent feedback via chat messages and training protocols completed by the participants. However, this increased the trainer's workload to a certain level; but in interventions that keep contact with the participants exclusively via remote, it is important for the trainer to follow a certain regularity, provide frequent feedback, and engage with the participant on a personal level with genuine interest anyway [46]. Finally, the adaptations were feasible, and we had the opportunity to further evaluate the effect of our physical exercise intervention on physical and mental health.

XSL•FO
RenderX

Next, we investigated the stress-buffering potential of our remotely supervised physical exercise intervention facing a severely restrictive life situation in terms of daily activity and social interactions. Specifically, we were interested in the potentially beneficial effects of regular physical exercise on both VO_{2peak} and depressive symptoms in the IGs. The results of the pre–COVID-19 cohort showed the expected significant increase in aerobic capacity for the IG but no changes in the CG. The increase in VO_{2peak} in the healthy pre–COVID-19 cohort is similar to changes in patient cohorts with different diseases in comparable intervention studies [35,37]. However, the exercise intervention did not significantly decrease depression scores, most likely due to only minimal depressive symptoms before intervention. In the lockdown cohort, time×group interactions revealed a significant deterioration of VO_{2peak} and a marginally significant decrease of depression in the CG but not in the IG. While the results in the pre–COVID-19 cohort indicate the overall effectiveness of the remote training intervention, the IG of the lockdown cohort maintained their physical and mental health in the face of a naturalistic stressor, that is, the pandemic-related lockdown. The latter finding is in line with previous studies on the stress-buffering effects of physical exercise [28,29]. Accordingly, the training intervention has likely compensated for the detrimental effects of a sedentary routine due to COVID-19–related restrictions.

Comparison to Prior Work

Our findings corroborate the stress-buffering potential of physical exercise previously indicated by multiple cross-sectional studies [47]. However, only a limited number of randomized controlled trials, especially in real-life stressor settings, have been conducted so far. For example, 20 weeks of aerobic exercise training were shown to have beneficial effects on the emotional stress reactivity of healthy students [48] for the prevention of stress-related mental health impairments. In line with the current findings, future studies should focus on real-life stressors when investigating the stress-buffering potential of physical exercise interventions.

The results hint at a stress-buffering effect of physical exercise after ~18 weeks. It remains unclear whether fewer training sessions, for example, 8 weeks, as originally planned, would yield similar stress-buffering effects. Most studies on physical exercise use intervention periods of 8-12 weeks, whereas a meta-analysis by Bruner et al [48] shows stronger effects for different factors of positive youth development in studies exceeding 10-week intervention periods. Interventional studies with overfat individuals show the most beneficial effects on cardiorespiratory characteristics in interventions that last at least 12 weeks [49]. The longer intervention period may thus also have enhanced effects on aerobic capacity. Further randomized controlled trials should define the modes, volumes, and lengths of physical exercise that elicit the stress-buffering effect of routine physical exercise.

Strengths and Limitations

To our knowledge, this is the first study to investigate the benefits of a training intervention at the onset of the COVID-19 pandemic in healthy untrained participants, which measures aerobic capacity by VO_{2peak}. Specifically, participants started the intervention shortly before pandemic-related restrictions were implemented. The significant improvements in VO_{2peak} observed in the pre–COVID-19 cohort were not achieved under lockdown conditions. Instead, aerobic capacity in the IG of the lockdown cohort did not change significantly, while it decreased significantly in the CG. In line with findings from similar studies, participants in the CG of the lockdown cohort may have experienced detraining or deconditioning. Ammar et al [50] showed a detraining effect after an initial 8-week fitness-dance training in older adults with mild cognitive impairment. While detraining in general is defined as the partial or complete loss of training-induced adaptations [51], our observations show a similar effect even without initial training in untrained participants. This deconditioning effect is most likely induced by the overall lowered physical activity [52-54], which different studies show during the COVID-19 lockdowns [4,6]. To conclude, training interventions may stop or at least buffer lockdown-induced deconditioning even in untrained healthy participants.

However, as external circumstances introduced unforeseen differences between the first and second cohorts, small sample sizes limited the robustness of our findings, which should consequently be interpreted with caution. While the results indicate large effect sizes regarding the intervention on aerobic capacity in the pre–COVID-19 cohort (η^2=0.32) and on both health parameters in the lockdown cohort (VO_{2peak}: η^2=0.47 and depression: η^2=0.23), a follow-up study in a more representative sample would enable a more conclusive interpretation of these findings. Along these lines, the limited statistical power did not permit analyses of potential mediator variables. Future studies in larger samples should analyze resilience mechanisms affected by the training intervention, for example, improved hypothalamic-pituitary-adrenal axis activity or increased self-efficacy. Still, the coincidental (or serendipitous) nature of our findings renders an equivalent replication study rather impossible. Nevertheless, a follow-up study may apply a somewhat similar design by implementing the training intervention in a population that experiences a more predictable real-life stressor with similar characteristics, for example, increased sedentary behavior and less social interactions. A population of interest may, for example, consist of students facing an examination period.

Conclusions

The outbreak of the COVID-19 pandemic and the implemented lockdown periods led to an increase in sedentary behavior associated with decreased health and well-being. We had the unique opportunity to investigate the feasibility and the stress-buffering potential of a remote physical exercise intervention with healthy untrained participants under these circumstances. Our results indicate that a lockdown cohort, which started a remotely supervised physical exercise intervention shortly before pandemic-related restrictions, significantly benefited from regular physical exercise in terms of aerobic capacity and depressive symptoms. Low dropout rates and high adherence to the training, despite the prevailing lockdown-related restrictions, underscore the feasibility and

XSL•FO
RenderX

importance of remote physical exercise training approaches in such scenarios. Furthermore, our findings highlight the importance of physical exercise to cope with stressful life situations and its integration into daily life, for example, on a personal level or via occupational health management. Future randomized controlled trials regarding further relevant real-life stressors may reveal the broad stress-buffering potential of physical exercise.

Acknowledgments

The authors thank all participants for their contributions to this study. This research was funded by the Boehringer Ingelheim Foundation in support of the former Deutsche Resilienz Zentrum (now: Leibniz Institute for Resilience Research), Mainz (subproject BIS 05). The Boehringer Ingelheim Foundation was not involved in the design, management, analysis, or reporting of the study. AB was additionally funded by the Mainz Research Center for Mental Health, University Clinics Johannes Gutenberg-University, Mainz.

Authors' Contributions

PS, MW, BL, and KL conceived the original idea. AB, KFAP, DTO, EWIN, and PS conceived and planned physical exercise testing and remote training supervision. PZ, MS, and MW conceived and planned psychological testing. KFAP, AB, MH, and TP performed exercise testing and remote training supervision. PZ and MS performed psychological testing. KFAP, PZ, AB, and RL analyzed the data. FW and JW planned and performed SARS-CoV-2 seroconversion tests. KFAP, PZ, and AB wrote the manuscript in consultation with EWIN, MS, MH, TP, RL, DTO, FW, JW, KL, BL, MW, and PS. KFAP and PZ contributed equally as first authors. PS (simonpe@uni-mainz.de) and MW (wessa@uni-mainz.de) contributed equally as corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH V1.6 checklist.
[PDF File (Adobe PDF File), 1443 KB - jmir_v26i1e53145_app1.pdf]

References

1. Schilling J, Buda S, Fischer M, Goerlitz L, Grote U, Haas W, et al. Retrospektive Phaseneinteilung der COVID-19-Pandemie in Deutschland bis Februar 2021 [Article in German]. Epidemiologisches Bull 2021(15):3-12 [FREE Full text]
2. COVID-19-Fälle nach Meldewoche und Geschlecht sowie Anteile mit für COVID-19 relevanten Symptomen, Anteile Hospitalisierter/Verstorbener und Altersmittelwert/-median [Article in German]. Robert Koch Institut. 2022. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/Klinische_Aspekte.html [accessed 2022-06-29]
3. Romero-Blanco C, Rodríguez-Almagro J, Onieva-Zafra MD, Parra-Fernández ML, Del Carmen Prado-Laguna M, Hernández-Martínez A. Physical activity and sedentary lifestyle in university students: changes during confinement due to the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(18):6567 [FREE Full text] [doi: 10.3390/ijerph17186567] [Medline: 32916972]
4. Stockwell S, Trott M, Tully M, Shin J, Barnett Y, Butler L, et al. Changes in physical activity and sedentary behaviours from before to during the COVID-19 pandemic lockdown: a systematic review. BMJ Open Sport Exerc Med 2021;7(1):e000960 [FREE Full text] [doi: 10.1136/bmjsem-2020-000960] [Medline: 34192010]
5. Gallè F, Sabella EA, Ferracuti S, de Giglio O, Caggiano G, Protano C, et al. Sedentary behaviors and physical activity of Italian undergraduate students during lockdown at the time of COVID-19 pandemic. Int J Environ Res Public Health 2020;17(17):6171 [FREE Full text] [doi: 10.3390/ijerph17176171] [Medline: 32854414]
6. Luciano F, Cenacchi V, Vegro V, Pavei G. COVID-19 lockdown: physical activity, sedentary behaviour and sleep in Italian medicine students. Eur J Sport Sci 2021;21(10):1459-1468 [FREE Full text] [doi: 10.1080/17461391.2020.1842910] [Medline: 33108970]
7. Ahsan M, Javed M. Effect on aerobic capacity, body mass, and body mass index in university soccer players during the COVID-19 pandemic. Saudi J Sports Med 2021;21(3):102. [doi: 10.4103/sjsm.sjsm_28_21]
8. Alvurdu S, Baykal C, Akyildiz Z, Şenel Ö, Silva AF, Conte D, et al. Impact of prolonged absence of organized training on body composition, neuromuscular performance, and aerobic capacity: a study in youth male soccer players exposed to COVID-19 lockdown. Int J Environ Res Public Health 2022;19(3):1148 [FREE Full text] [doi: 10.3390/ijerph19031148] [Medline: 35162174]
9. Kalinowski P, Myszkowski J, Marynowicz J. Effect of online training during the COVID-19 quarantine on the aerobic capacity of youth soccer players. Int J Environ Res Public Health 2021;18(12):6195 [FREE Full text] [doi: 10.3390/ijerph18126195] [Medline: 34201095]

XSL•FO
RenderX

10. Berkowicz K, Obmiński Z, Michnik K, Czaja R. The impact of COVID-19 lockdown on aerobic capacity in female soccer players. J Kinesiol Exerc Sci 2021;31(96):49-53 [FREE Full text] [doi: 10.5604/01.3001.0015.7426]
11. Font R, Irurtia A, Gutierrez JA, Salas S, Vila E, Carmona G. The effects of COVID-19 lockdown on jumping performance and aerobic capacity in elite handball players. Biol Sport 2021;38(4):753-759 [FREE Full text] [doi: 10.5114/biolsport.2021.109952] [Medline: 34937987]
12. Sunda M, Gilic B, Peric I, Savicevic AJ, Sekulic D. Evidencing the influence of the COVID-19 pandemic and imposed lockdown measures on fitness status in adolescents: a preliminary report. Healthcare (Basel) 2021;9(6):681 [FREE Full text] [doi: 10.3390/healthcare9060681] [Medline: 34198858]
13. Park MS, Chung SY, Chang Y, Kim K. Physical activity and physical fitness as predictors of all-cause mortality in Korean men. J Korean Med Sci 2009;24(1):13-19 [FREE Full text] [doi: 10.3346/jkms.2009.24.1.13] [Medline: 19270807]
14. Strasser B, Burtscher M. Survival of the fittest: VO2max, a key predictor of longevity? Front Biosci (Landmark Ed) 2018;23(8):1505-1516 [FREE Full text] [doi: 10.2741/4657] [Medline: 29293447]
15. Sui X, LaMonte MJ, Blair SN. Cardiorespiratory fitness as a predictor of nonfatal cardiovascular events in asymptomatic women and men. Am J Epidemiol 2007;165(12):1413-1423 [FREE Full text] [doi: 10.1093/aje/kwm031] [Medline: 17406007]
16. Kodama S, Saito K, Tanaka S, Maki M, Yachi Y, Asumi M, et al. Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. J Am Med Assoc 2009;301(19):2024-2035 [FREE Full text] [doi: 10.1001/jama.2009.681] [Medline: 19454641]
17. Winett RA, Carpinelli RN. Examining the validity of exercise guidelines for the prevention of morbidity and all-cause mortality. Ann Behav Med 2000;22(3):237-245 [FREE Full text] [doi: 10.1007/BF02895119] [Medline: 11126469]
18. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
19. Teychenne M, Ball K, Salmon J. Sedentary behavior and depression among adults: a review. Int J Behav Med 2010;17(4):246-254. [doi: 10.1007/s12529-010-9075-z] [Medline: 20174982]
20. Meyer J, McDowell C, Lansing J, Brower C, Smith L, Tully M, et al. Changes in physical activity and sedentary behavior in response to COVID-19 and their associations with mental health in 3052 US adults. Int J Environ Res Public Health 2020;17(18):6469 [FREE Full text] [doi: 10.3390/ijerph17186469] [Medline: 32899495]
21. Maugeri G, Castrogiovanni P, Battaglia G, Pippi R, D'Agata V, Palma A, et al. The impact of physical activity on psychological health during COVID-19 pandemic in Italy. Heliyon 2020;6(6):e04315 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e04315] [Medline: 32613133]
22. Chtourou H, Trabelsi K, H'mida C, Boukhris O, Glenn JM, Brach M, et al. Staying physically active during the quarantine and self-isolation period for controlling and mitigating the COVID-19 pandemic: a systematic overview of the literature. Front Psychol 2020;11:1708 [FREE Full text] [doi: 10.3389/fpsyg.2020.01708] [Medline: 33013497]
23. Mead GE, Morley W, Campbell P, Greig CA, McMurdo M, Lawlor DA. Exercise for depression. Cochrane Database Syst Rev 2008(4):CD004366. [doi: 10.1002/14651858.CD004366.pub3] [Medline: 18843656]
24. Schuch FB, Vancampfort D, Firth J, Rosenbaum S, Ward PB, Silva ES, et al. Physical activity and incident depression: a meta-analysis of prospective cohort studies. Am J Psychiatry 2018;175(7):631-648 [FREE Full text] [doi: 10.1176/appi.ajp.2018.17111194] [Medline: 29690792]
25. Carek PJ, Laibstain SE, Carek SM. Exercise for the treatment of depression and anxiety. Int J Psychiatry Med 2011;41(1):15-28. [doi: 10.2190/PM.41.1.c] [Medline: 21495519]
26. Rimmele U, Zellweger BC, Marti B, Seiler R, Mohiyeddini C, Ehlert U, et al. Trained men show lower cortisol, heart rate and psychological responses to psychosocial stress compared with untrained men. Psychoneuroendocrinology 2007;32(6):627-635. [doi: 10.1016/j.psyneuen.2007.04.005] [Medline: 17560731]
27. Kalisch R, Müller MB, Tüscher O. A conceptual framework for the neurobiological study of resilience. Behav Brain Sci 2015;38:e92. [doi: 10.1017/S0140525X1400082X] [Medline: 25158686]
28. Silverman MN, Deuster PA. Biological mechanisms underlying the role of physical fitness in health and resilience. Interface Focus 2014;4(5):20140040 [FREE Full text] [doi: 10.1098/rsfs.2014.0040] [Medline: 25285199]
29. Neumann RJ, Ahrens KF, Kollmann B, Goldbach N, Chmitorz A, Weichert D, et al. The impact of physical fitness on resilience to modern life stress and the mediating role of general self-efficacy. Eur Arch Psychiatry Clin Neurosci 2022;272(4):679-692 [FREE Full text] [doi: 10.1007/s00406-021-01338-9] [Medline: 34622343]
30. Davies CA, Spence JC, Vandelanotte C, Caperchione CM, Mummery WK. Meta-analysis of internet-delivered interventions to increase physical activity levels. Int J Behav Nutr Phys Act 2012;9:52 [FREE Full text] [doi: 10.1186/1479-5868-9-52] [Medline: 22546283]
31. van den Berg MH, Schoones JW, Vlieland TPMV. Internet-based physical activity interventions: a systematic review of the literature. J Med Internet Res 2007;9(3):e26 [FREE Full text] [doi: 10.2196/jmir.9.3.e26] [Medline: 17942388]
32. Jahangiry L, Farhangi MA, Shab-Bidar S, Rezaei F, Pashaei T. Web-based physical activity interventions: a systematic review and meta-analysis of randomized controlled trials. Public Health 2017;152:36-46. [doi: 10.1016/j.puhe.2017.06.005] [Medline: 28734170]

XSL•FO
RenderX

33. Pfirrmann D, Haller N, Huber Y, Jung P, Lieb K, Gockel I, et al. Applicability of a web-based, individualized exercise intervention in patients with liver disease, cystic fibrosis, esophageal cancer, and psychiatric disorders: process evaluation of 4 ongoing clinical trials. JMIR Res Protoc 2018;7(5):e106 [FREE Full text] [doi: 10.2196/resprot.8607] [Medline: 29789277]
34. Pfirrmann D, Tug S, Brosteanu O, Mehdorn M, Busse M, Grimminger PP, et al. Internet-based perioperative exercise program in patients with Barrett's carcinoma scheduled for esophagectomy [iPEP—study] a prospective randomized-controlled trial. BMC Cancer 2017;17(1):413 [FREE Full text] [doi: 10.1186/s12885-017-3400-8] [Medline: 28615010]
35. Haller N, Lorenz S, Pfirrmann D, Koch C, Lieb K, Dettweiler U, et al. Individualized web-based exercise for the treatment of depression: randomized controlled trial. JMIR Ment Health 2018;5(4):e10698 [FREE Full text] [doi: 10.2196/10698] [Medline: 30314962]
36. Hillen B, Simon P, Schlotter S, Nitsche O, Bähner V, Poplawska K, et al. Feasibility and implementation of a personalized, web-based exercise intervention for people with cystic fibrosis for 1 year. BMC Sports Sci Med Rehabil 2021;13(1):95 [FREE Full text] [doi: 10.1186/s13102-021-00323-y] [Medline: 34412703]
37. Pfirrmann D, Huber Y, Schattenberg JM, Simon P. Web-based exercise as an effective complementary treatment for patients with nonalcoholic fatty liver disease: intervention study. J Med Internet Res 2019;21(1):e11250 [FREE Full text] [doi: 10.2196/11250] [Medline: 30602434]
38. Bossen D, Veenhof C, Dekker J, de Bakker D. The effectiveness of self-guided web-based physical activity interventions among patients with a chronic disease: a systematic review. J Phys Act Health 2014;11(3):665-677. [doi: 10.1123/jpah.2012-0152] [Medline: 23493018]
39. Kuijpers W, Groen WG, Aaronson NK, van Harten WH. A systematic review of web-based interventions for patient empowerment and physical activity in chronic diseases: relevance for cancer survivors. J Med Internet Res 2013;15(2):e37 [FREE Full text] [doi: 10.2196/jmir.2281] [Medline: 23425685]
40. Ochmann DT, Philippi KFA, Zeier P, Sandner M, Hillen B, Neuberger EWI, et al. Association of innate and acquired aerobic capacity with resilience in healthy adults: protocol for a randomized controlled trial of an 8-week web-based physical exercise intervention. JMIR Res Protoc 2021;10(11):e29712 [FREE Full text] [doi: 10.2196/29712] [Medline: 34847062]
41. Derogatis LR. The Brief Symptom Inventory-18 (BSI-18): Administration, Scoring, and Procedures Manual. Upper Saddle River, NJ: Pearson; 2000.
42. Spitzer C, Hammer S, Löwe B, Grabe HJ, Barnow S, Rose M, et al. Die Kurzform des Brief Symptom Inventory (BSI-18): erste Befunde zu den psychometrischen Kennwerten der deutschen Version. The short version of the Brief Symptom Inventory (BSI-18): preliminary psychometric properties of the German translation [Article in German]. Fortschr Neurol Psychiatr 2011;79(9):517-523. [doi: 10.1055/s-0031-1281602] [Medline: 21870312]
43. Day JR, Rossiter HB, Coats EM, Skasick A, Whipp BJ. The maximally attainable VO2 during exercise in humans: the peak vs. maximum issue. J Appl Physiol (1985) 2003;95(5):1901-1907 [FREE Full text] [doi: 10.1152/japplphysiol.00024.2003] [Medline: 12857763]
44. Notz Q, Schmalzing M, Wedekink F, Schlesinger T, Gernert M, Herrmann J, et al. Pro- and anti-inflammatory responses in severe COVID-19-induced acute respiratory distress syndrome—an observational pilot study. Front Immunol 2020;11:581338 [FREE Full text] [doi: 10.3389/fimmu.2020.581338] [Medline: 33123167]
45. Leys C, Schumann S. A nonparametric method to analyze interactions: the adjusted rank transform test. J Exp Soc Psychol 2010;46(4):684-688 [FREE Full text] [doi: 10.1016/j.jesp.2010.02.007]
46. Hinman RS, Delany CM, Campbell PK, Gale J, Bennell KL. Physical therapists, telephone coaches, and patients with knee osteoarthritis: qualitative study about working together to promote exercise adherence. Phys Ther 2016 Apr;96(4):479-493. [doi: 10.2522/ptj.20150260] [Medline: 26316529]
47. von Haaren B, Haertel S, Stumpp J, Hey S, Ebner-Priemer U. Reduced emotional stress reactivity to a real-life academic examination stressor in students participating in a 20-week aerobic exercise training: a randomised controlled trial using ambulatory assessment. Psychol Sport Exerc 2015 Sep;20:67-75. [doi: 10.1016/j.psychsport.2015.04.004]
48. Bruner MW, McLaren CD, Sutcliffe JT, Gardner LA, Lubans DR, Smith JJ, et al. The effect of sport-based interventions on positive youth development: a systematic review and meta-analysis. Int Rev Sport Exerc 2021;16(1):368-395. [doi: 10.1080/1750984x.2021.1875496]
49. Clark JE. The impact of duration on effectiveness of exercise, the implication for periodization of training and goal setting for individuals who are overfat, a meta-analysis. Biol Sport 2016;33(4):309-333 [FREE Full text] [doi: 10.5604/20831862.1212974] [Medline: 28090136]
50. Ammar A, Boukhris O, Halfpaap N, Labott BK, Langhans C, Herold F, et al. Four weeks of detraining induced by COVID-19 reverse cardiac improvements from eight weeks of fitness-dance training in older adults with mild cognitive impairment. Int J Environ Res Public Health 2021 May 31;18(11):5930 [FREE Full text] [doi: 10.3390/ijerph18115930] [Medline: 34073051]
51. Mujika I, Padilla S. Detraining: loss of training-induced physiological and performance adaptations. Part I: short term insufficient training stimulus. Sports Med 2000 Aug;30(2):79-87. [doi: 10.2165/00007256-200030020-00002] [Medline: 10966148]

XSL•FO
RenderX

52. Silva FM, Duarte-Mendes P, Rusenhack MC, Furmann M, Nobre PR, Fachada M, et al. Objectively measured sedentary behavior and physical fitness in adults: a systematic review and meta-analysis. Int J Environ Res Public Health 2020 Nov 21;17(22):8660 [FREE Full text] [doi: 10.3390/ijerph17228660] [Medline: 33233451]
53. Santos R, Mota J, Okely AD, Pratt M, Moreira C, Coelho-e-Silva MJ, et al. The independent associations of sedentary behaviour and physical activity on cardiorespiratory fitness. Br J Sports Med 2014 Oct;48(20):1508-1512. [doi: 10.1136/bjsports-2012-091610] [Medline: 23410883]
54. Lavie CJ, Ozemek C, Carbone S, Katzmarzyk PT, Blair SN. Sedentary behavior, exercise, and cardiovascular health. Circ Res 2019 Mar;124(5):799-815. [doi: 10.1161/CIRCRESAHA.118.312669] [Medline: 30817262]

Abbreviations

ART: aligned rank transform
BSI-18: Brief Symptom Inventory-18
CG: control group
CONSORT: Consolidated Standards of Reporting Trials
IG: intervention group
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
VO_{2peak}: peak oxygen uptake

Edited by T de Azevedo Cardoso, G Greco; submitted 27.09.23; peer-reviewed by Y Jang, M Pritchard; comments to author 08.12.23; revised version received 17.01.24; accepted 08.04.24; published 08.08.24.

Please cite as:
Philippi KFA, Zeier P, Brahmer A, Neuberger EWI, Sandner M, Hagenah M, Porten T, Lenz R, Ochmann DT, Wedekink F, Wischhusen J, Lutz B, Lieb K, Wessa M, Simon P
Remote Exercise Training Intervention During the COVID-19 Pandemic: Randomized Controlled Trial
J Med Internet Res 2024;26:e53145
URL: https://www.jmir.org/2024/1/e53145
doi:10.2196/53145
PMID:

©Keito F A Philippi, Peter Zeier, Alexandra Brahmer, Elmo W I Neuberger, Magdalena Sandner, Matthias Hagenah, Thilo Porten, Regina Lenz, David T Ochmann, Florian Wedekink, Jörg Wischhusen, Beat Lutz, Klaus Lieb, Michèle Wessa, Perikles Simon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sex Differences in Pain Sensitivity in a Dutch Cohort: Cross-Sectional and Web-Based Multidimensional Study

Rachel J H Smits[1], MD; Selina E I van der Wal[1], MD, PhD; Regina L M van Boekel[1], PhD; Hans Timmerman[2], PhD; Ewald M Bronkhorst[3], PhD; Diana Abrar[1], MD; Kris C P Vissers[1], MD, PhD; Esmeralda N Blaney Davidson[4*], PhD; Monique A H Steegers[5*], MD, PhD

[1]Department of Anesthesiology, Pain and Palliative Medicine, Radboud university medical center, Nijmegen, Netherlands
[2]Department of Anesthesiology, Pain Center, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[3]Department for Health Evidence, Radboud university medical center, Nijmegen, Netherlands
[4]Department of Rheumatic Diseases, Radboud university medical center, Nijmegen, Netherlands
[5]Department of Anesthesiology, Amsterdam University Medical Center, location VU Medical Center, Amsterdam, Netherlands
[*]these authors contributed equally

Corresponding Author:
Rachel J H Smits, MD
Department of Anesthesiology, Pain and Palliative Medicine
Radboud university medical center
Postbus 9101
Nijmegen, 6500 HB
Netherlands
Phone: 31 6 29698822
Email: rachel.smits@radboudumc.nl

Abstract

Background: Sex is an important factor influencing the development and treatment of chronic pain, but the extent of its influence is still unclear. Other demographic factors as well as nonpharmacological interventions might influence pain sensitivity differently in men and women.

Objective: In this study, we aimed to investigate the influence of sex and other demographic, lifestyle, behavioral, clinical, and environmental factors on pain sensitivity in the Dutch population. Different films were used to investigate how they would impact pain sensitivity and what influence sex and other variables have on the effect of this simple intervention.

Methods: We performed a study consisting of 2 parts: (1) a cross-sectional research to investigate pain sensitivity differences between men and women and the influence of other demographic variables on the pain sensitivity in a Dutch cohort and (2) an internet intervention study to determine whether a short film could skew pain sensitivity.

Results: All respondents filled in a web-based demographic questionnaire and were randomized into 4 groups. The control group filled in the Pain Sensitivity Questionnaire without watching a preliminary film. A cross-sectional analysis was performed in the control group (n=1746). The other 3 groups watched short films: one group watched a film with scenes of nature (n=2650), another group watched a film on laughing people (n=2735), and the last group watched a film on physically painful events (n=2708). Immediately after the film viewing, participants were directed to the Pain Sensitivity Questionnaire to measure their pain sensitivity. The Pain Sensitivity Questionnaire score was stated as a mean per question on the numeric rating scale from 0-1. The cross-sectional study revealed no significant differences between men and women but showed male-female differences in the Pain Sensitivity Questionnaire when specific background factors were present. Watching a short film had a positive impact on the pain sensitivity of the respondents who had chronic pain, with a higher effect observed in female respondents.

Conclusions: Scientists performing pain research need to account for factors that can influence the outcome of their study and be aware that these factors can be sex-dependent, and pain sensitivity should be analyzed accordingly. Even relatively small interventions such as watching a film can impact pain sensitivity, especially in respondents with current chronic pain. This effect can vary as well when different background factors are present. Our findings warrant further explorations of the possibilities that simple interventions bring for patients in personalized medicine.

Trial Registration: Landelijk Trial Register NTR-new NL8182; https://onderzoekmetmensen.nl/en/trial/29537

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53926) doi:10.2196/53926

KEYWORDS

pain sensitivity; pain sensitivity questionnaire; chronic pain; digital interventions; mobile health; film; sex; Dutch population; personalized medicine; individualized care

Introduction

According to the International Association for the Study of Pain, pain is a personal experience that is influenced in varying degrees by biological, psychological, and social factors [1]. Individual pain sensitivity is an important factor in the experience and treatment of pain. A higher pain sensitivity is associated with the development of chronic pain [2]. Numerous demographic variables such as age, obesity, and smoking have been associated with the development of chronic pain [3-5]. Analgesic drug treatment of chronic pain is often difficult, insufficient, and has substantial side effects, thereby emphasizing the need for more nonpharmacological preventive and treatment options [6,7].

Sex is an important factor influencing chronic pain as well [8,9]. Women report a higher prevalence of chronic pain and experience more severe and longer lasting pain compared with men in comparable situations [10,11]. Important variability has been observed in the pharmacokinetics and pharmacodynamics between sexes, with a higher analgesic effect and more side effects of opioids and other pharmacological agents in women [12,13]. Men and women have also different responses to treatment in multimodal pain management and cognitive behavioral therapy [14,15]. However, the effects of sex and other background factors on pain sensitivity and whether women or men are more sensitive to painful stimuli are unclear [16]. Moreover, demographic, lifestyle, behavioral, clinical, and environmental factors might influence pain sensitivity differently in men and women [9]. These sex differences can be linked to specific physiological mechanisms such as cerebral responses to visual, auditory, emotional, and pain processing stimuli [17]. This includes the observation of the strong influence of empathy in experimental settings on pain and higher thresholds for painful stimuli after social laughter among women [18,19]. These factors raise the question whether nonpharmacological interventions such as watching a short film might work differently between men and women.

With this study, we wanted to raise awareness in the Dutch population regarding chronic pain and the sex differences in chronic pain. With our data, we aimed to investigate the influence of sex and other demographic, lifestyle, behavioral, clinical, and environmental factors on pain sensitivity. Short films with different contents were implemented to make the study more appealing for the public and were used to investigate how films would impact pain sensitivity and what influence sex and other variables have on the effect of this simple intervention.

Methods

Recruitment

In 2017 and 2018, a national campaign was launched in the Netherlands in cooperation with the Dutch Public Broadcast Corporation (Nederlandse Publieke Omroep) to raise awareness of chronic pain and sex differences in chronic pain. This study was part of this campaign and enabled people to actively participate and gain knowledge about pain and their own pain sensitivity. This awareness creation included a broadcast on national television, several radio performances, and promotion at festivals. The website was presented through the national campaign. Volunteers were enrolled between May 4, 2017, and October 8, 2018.

Ethics Approval

This study adheres to the applicable CONSORT (Consolidated Standards of Reporting Trials) guidelines. The trial protocol was approved by the Medical Ethical Committee on Research Involving Human Subjects (Arnhem-Nijmegen NL2017-3314). The study was registered in the Dutch national trial register on November 21, 2019 (NTR-new NL8182).

Trial Oversight

This first Dutch national web-based pain research trial in the general population was a large national mixed methods study with a cross-sectional questionnaire in the Dutch population and a concomitant feasibility study with a web-based intervention. The trial was designed by a committee formed at the Radboud University Medical Center. It was supported and partly funded by the Dutch Society for Scientific Research and an independent Dutch public service broadcaster specializing in information, education, and culture. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

This study was performed in collaboration with Dutch television. This enabled us to gather data on a very large population but was associated with a tight deadline in which we had to move very quickly to get the study up and running. Therefore, our entire team was focused on ensuring that the setup of our experiment was sound, but as a result, the registration was delayed. The protocol was not altered after the gained approval by the Medical Ethical Committee, and the website was launched on May 4, 2017. No adjustments were made in the website until the last inclusion.

Population

Adults with proper understanding of the Dutch language could voluntarily participate in this study. The informed consent procedure was performed digitally. A total of 10,876 consenting volunteers participated. Filling in the demographic questionnaire was optional. After excluding the participants who did not complete the demographic questionnaire and those who gave impossible answers, 9839 respondent data were analyzed.

Trial Procedure

After registration, participants were requested to fill out a general baseline questionnaire containing demographic factors

XSL•FO
RenderX

and variables that might influence pain sensitivity. These included sex, age, weight, height, highest level of education, smoking, hours of physical activity per week, current pain level on a numeric rating scale (NRS 0: no pain to 10: worst pain imaginable), and the duration of the current pain. They were not obligated to fill in information and were allowed to skip questions; only participants who completed the questionnaires about demographics and pain sensitivity were included in the final database for the cross-sectional analysis.

After completion of the baseline questionnaire, a built-in automated complete blind randomization program allocated the respondents into 1 of the 4 groups (Figure 1). The control group filled in the Pain Sensitivity Questionnaire (PSQ) without watching a preliminary film. These data were used for the cross-sectional analysis. The other groups were shown one of the 3 films with a duration of 45 seconds. The clips in the short films were chosen via a discussion-based focus group. The first film was considered as a neutral film, displaying calm scenes of nature like gently flowing water; the second film showed contagious laughter of people of all ages; and the third film showed people of all ages enduring clearly physically painful situations. Immediately after the viewing, participants were directed to the PSQ to measure their pain sensitivity. The feasibility study was designed to assess whether a simple intervention such as watching a film could influence pain sensitivity, as measured using PSQ.

Figure 1. Enrollment and randomization of respondents. PSQ: Pain Sensitivity Questionnaire.

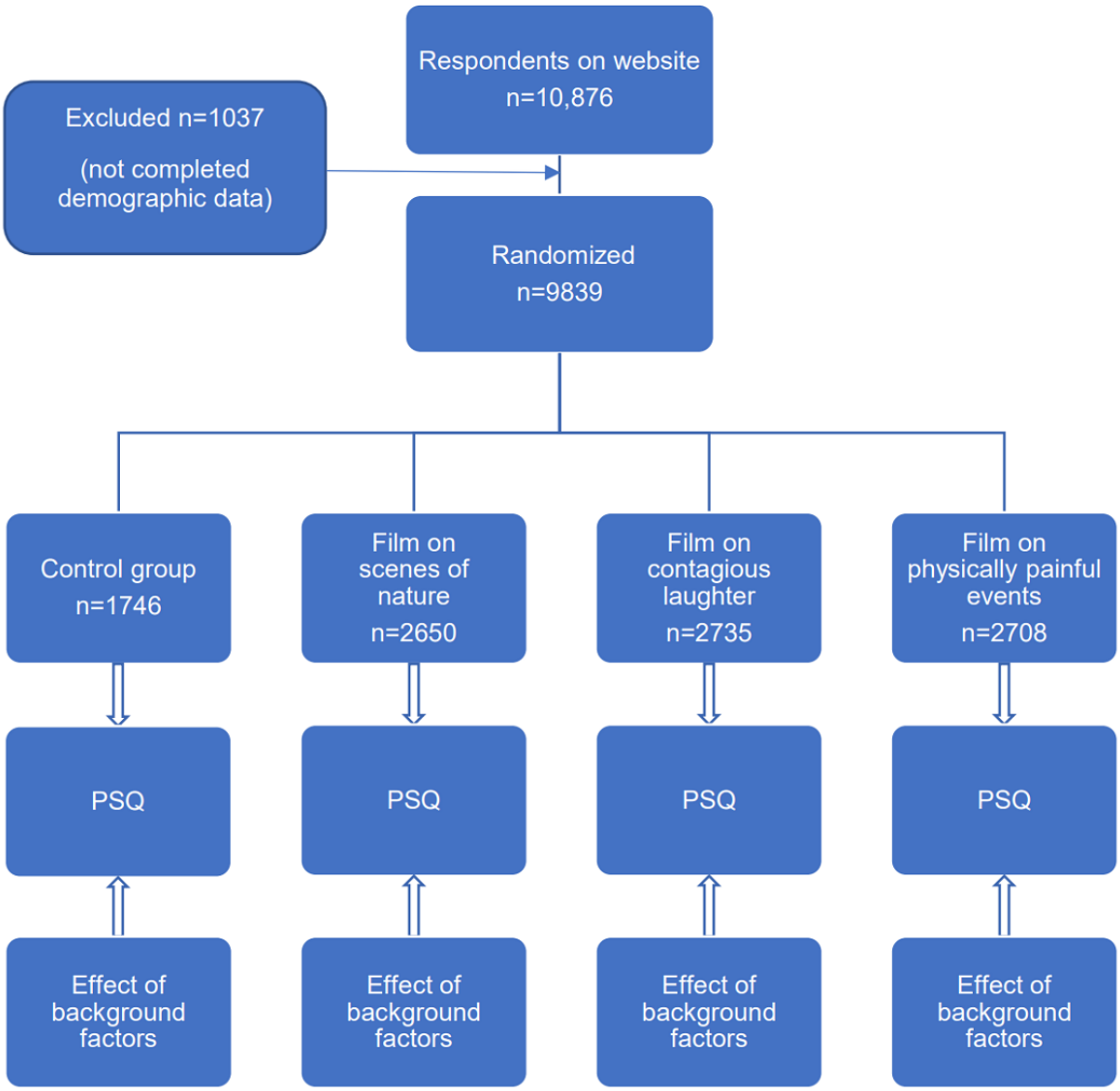

Study Outcome

The primary outcome of this study was the score on PSQ, reflecting the participant's pain sensitivity and the effect of background factors on their pain sensitivity. The PSQ was developed by Ruscheweyh and coworkers [20] to measure pain sensitivity in daily clinical practice without experimental determination of one's pain. The validity to measure perceived

XSL•FO
RenderX

pain sensitivity has been shown by significant correlations with experimental pain intensity in healthy individuals as well as in patients with chronic pain [20-22]. It has been used, among other instruments, to identify pain sensitivity in epidemiological health research and in several studies regarding postoperative and chronic pain before and after multiple surgeries [20-23].

The PSQ has been translated and validated in the healthy Dutch population [24]. It is a self-rating pain measurement tool consisting of 17 questions, which entail imagining potentially painful everyday life situations, and takes 5-10 minutes to complete. Respondents score their pain intensity of the described situation ranging from 0 (not painful) to 10 (strongest pain imaginable) on the NRS. Fourteen questions represent life situations that a majority of healthy individuals consider painful. Three questions represent situations that are generally perceived as nonpainful and are used as a nonpainful sensory reference not included in the final scoring. A higher score on the PSQ (range 0-140) corresponds to a higher sensitivity to pain, meaning the respondent rates a higher pain score for the same situation compared to a respondent with a lower score. To facilitate the interpretation of PSQ, the overall score is transformed to a score between 0 and 10; therefore, its score has the same range as the individual questions.

Statistical Analysis

In the descriptive analyses, all categorical variables were presented with counts and percentages, while the continuous variables were presented with mean and range. The effect, stated as β, was presented with a 95% CI. To analyze the relation between sex and pain sensitivity, adjusted for other background factors, a multiple linear regression model was used. In this model, full interaction between sex and all other demographic variables was added, without demand for statistical significance. Because of the large sample size, power analysis was not performed. For the same reason, we used the model as specified, without checking for multicollinearity, relying on the large sample size to compensate for potential loss of precision due to multicollinearity. The model was fully created with an “enter philosophy” in mind; all variables were entered at once.

For the feasibility study concerning the effect of exposure to various short films on the PSQ score, linear regression was applied again. Due to the smaller sample size in the control group compared to the other group sizes, the film on the calm scenes of nature was used as the control group (as a neutral audiovisual stimulus) to compare the effects of the film on contagious laughter and the film on painful events to limit the chance of subtle selection effects. Film was entered in the model as a categorical variable, with nature film as a reference. In addition to film, sex was added to the model as well, supplemented with an interaction term for sex and film.

Based on the relations found in the demographic analysis, a more comprehensive analysis of the effect of the film was undertaken. In this analysis, the base regression model was a 3-way interaction between film, sex, and NRS score, supplemented with all other demographic variables. Interactions with this last group of variables were only added if it improved the model fit significantly. Analysis was performed using SPSS Statistics for Windows (version 25.0; IBM Corp; released 2013) and R (version 3.6.2). In all analyses a P value <.05 was considered statistically significant.

Results

Cross-Sectional Analysis Results

The common variables gathered in demographic research included age, sex, BMI, smoking, exercise in hours per week, and education (Table 1). Of the 7375 female respondents, 25.9% (1908/7375) reported no pain, 27.3% (2016/7375) reported mild pain, 35.2% (2599/7375) reported moderate pain, and 10.9% (810/7375) reported severe pain. In women experiencing pain, 81.2% (4403/5425) had pain lasting more than 3 months (eg, chronic pain). In the male cohort, 39.8% (979/2464) did not report pain, 31.5% (775/2464) reported mild pain, 23.2% (571/2464) reported moderate pain, and 5.5% (134/2464) reported severe pain. The pain was present for more than 3 months in 74.8% (1111/1485) of the men reporting pain.

XSL•FO
RenderX

Table 1. Complete demographic data of all the respondents (N= 9839).

	No intervention		Nature film		Contagious laughter film		Painful event film	
	Males (n=411)	Females (n=1335)	Males (n=625)	Females (n=2025)	Males (n=700)	Females (n=2035)	Males (n=728)	Females (n=1980)
Age (years), mean (range)	49 (18-89)	44 (18-93)	49 (18-91)	44 (18-86)	49 (18-90)	44 (18-88)	50 (18-90)	43 (18-85)
BMI (kg/m²), mean (range)	25.0 (16.5-46.8)	25.9 (14.3-52.4)	25.2 (17.1-43.3)	25.5 (14.2-62.4)	25.5 (15.6-56.8)	25.6 (13.9-61.3)	25.5 (15.7-56.4)	25.8 (14.0-61.0)
Smoking, n (%)								
Yes	60 (14.6)	226 (16.9)	95 (15.2)	336 (16.6)	116 (16.6)	313 (15.4)	131 (18)	312 (15.8)
No	350 (85.2)	1104 (82.7)	529 (84.6)	1683 (83.1)	583 (83.3)	1716 (84.3)	592 (81.3)	1658 (83.7)
Missing	1 (0.2)	5 (0.4)	1 (0.2)	6 (0.3)	1 (0.1)	6 (0.3)	5 (0.7)	10 (0.5)
Practice sports, n (%)								
Yes	224 (54.5)	721 (54)	353 (56.5)	1098 (54.2)	397 (56.7)	1156 (56.8)	399 (54.8)	1085 (54.8)
No	185 (45)	609 (45.6)	270 (43.2)	920 (45.4)	303 (43.3)	873 (42.9)	329 (45.2)	892 (45)
Missing	2 (0.5)	5 (0.4)	2 (0.3)	7 (0.4)	0 (0)	6 (0.3)	0 (0)	3 (0.2)
Exercise (h/wk), n (%)								
No sport	185 (45)	609 (45.6)	270 (43.2)	920 (45.4)	303 (43.3)	873 (42.9)	329 (45.2)	892 (45)
0-1	29 (7.1)	128 (9.6)	48 (7.7)	230 (11.3)	47 (6.7)	220 (10.8)	45 (6.2)	212 (10.7)
2-3	88 (21.4)	386 (29)	149 (23.8)	541 (26.7)	172 (24.6)	593 (29.1)	181 (24.9)	582 (29.4)
3-6	77 (18.7)	157 (11.7)	105 (16.8)	259 (12.8)	112 (16)	261 (12.8)	111 (15.2)	211 (10.7)
7-10	20 (4.9)	37 (2.8)	41 (6.6)	50 (2.5)	49 (7)	64 (3.1)	48 (6.6)	64 (3.2)
More than 10	10 (2.4)	12 (0.9)	11 (1.7)	16 (0.8)	17 (2.4)	17 (0.8)	14 (1.9)	14 (0.7)
Missing	2 (0.5)	6 (0.4)	1 (0.2)	9 (0.5)	0 (0)	7 (0.3)	0 (0)	5 (0.3)
Education, n (%)								
Primary school	7 (1.7)	26 (1.9)	13 (2.1)	38 (1.9)	14 (2)	46 (2.3)	16 (2.2)	44 (2.2)
Secondary school	63 (15.3)	244 (18.3)	97 (15.5)	374 (18.5)	123 (17.6)	356 (17.5)	125 (17.2)	370 (18.7)
Secondary vocational school	84 (20.4)	381 (28.5)	134 (21.4)	556 (27.5)	158 (22.6)	551 (27.1)	161 (22.1)	540 (27.3)
Higher professional education or up	249 (60.6)	649 (48.6)	364 (58.2)	1015 (50.1)	395 (56.4)	1048 (51.5)	415 (57)	977 (49.3)
No education	3 (0.7)	19 (1.4)	11 (1.8)	25 (1.2)	4 (0.6)	24 (1.2)	11 (1.5)	27 (1.4)
Missing	5 (1.2)	16 (1.2)	6 (1)	17 (0.8)	6 (0.8)	10 (0.5)	0 (0)	22 (1.1)
Current pain score (numeric rating scale score), n (%)								
No pain	183 (44.5)	314 (23.5)	240 (38.4)	538 (26.6)	293 (41.9)	533 (26.2)	263 (36.1)	523 (26.4)
Mild (1-3)	120 (29.2)	337 (25.2)	211 (33.8)	536 (26.5)	198 (28.3)	583 (28.6)	246 (33.8)	560 (28.3)
Moderate (4-7)	84 (20.5)	500 (37.5)	138 (22.1)	723 (35.7)	175 (25)	704 (34.6)	174 (23.9)	672 (33.9)
Severe (8-10)	24 (5.8)	177 (13.3)	34 (5.4)	215 (10.6)	33 (4.7)	204 (10)	43 (5.9)	214 (10.8)
Missing	0 (0)	7 (0.5)	2 (0.3)	13 (0.6)	1 (0.1)	11 (0.5)	2 (0.3)	11 (0.6)
Duration of pain (months), n (%)								
No pain	183 (44.5)	314 (23.5)	240 (38.4)	538 (26.6)	293 (41.9)	533 (26.2)	263 (36.1)	523 (26.4)
Shorter than 3 months	55 (13.4)	155 (11.6)	95 (15.2)	278 (13.7)	91 (13)	297 (14.6)	113 (15.5)	258 (13)
3-6 months	17 (4.2)	62 (4.6)	25 (4)	71 (3.5)	26 (3.7)	84 (4.1)	36 (4.9)	65 (3.3)
Longer than 6 months	151 (36.7)	788 (59)	257 (41.1)	1120 (55.3)	286 (4.8)	1098 (53.9)	313 (43)	1115 (56.3)
Missing	5 (1.2)	16 (1.2)	8 (1.3)	18 (0.9)	4 (0.6)	23 (1.1)	3 (0.4)	19 (1)

XSL•FO
RenderX

Pain Sensitivity Is Influenced by Background Factors

A total of 1746 respondents filled in the PSQ in the control group. This group was also used for cross-sectional analysis. The control group consisted of 411 men and 1335 women. Initially, we intended to analyze men versus women in a relatively simple manner and hypothesized their PSQ scores would be different. Overall, it was not possible to state whether the pain sensitivity is different in men and women without taking the background factors into account. In doing so, we found that factors that affect the PSQ scores positively were different between men and women. The PSQ scores of women who smoked were lower than those who did not smoke (–.28). Men with a higher education had a lower PSQ score, while for women, no difference was seen in the PSQ score between different levels of education. Experiencing moderate and severe pain in women affected the PSQ scores negatively (.96 and 1.65, respectively), meaning a higher score on the PSQ. Men with severe pain also had a higher score on the PSQ, but this result was obtained in a very small group (n=24). Table 2 and Figure 2 show the effect of different background factors on the PSQ scores in men and women.

Figure 2. Effects of different background factors on the Pain Sensitivity Questionnaire scores of men and women. NRS: numeric rating scale.

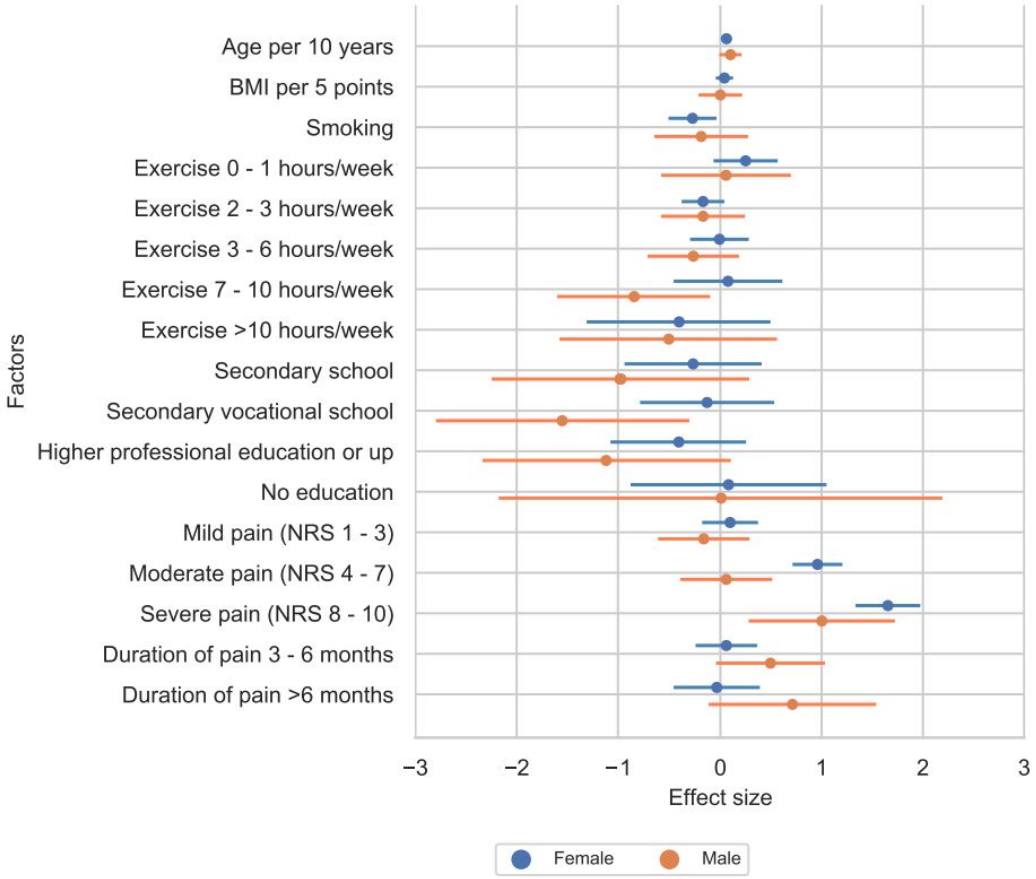

XSL•FO
RenderX

Table 2. Effect of various background factors on the Pain Sensitivity Questionnaire scores in men and women in the control group (N=1746).

Background factors per category	Males			Females		
	Value	β (95% CI)	*P* value	Value	β (95% CI)	*P* value
Total (intercept), n	411	4.38 (2.76 to 6.00)	<.001	1335	3.48 (2.64 to 4.33)	<.001
Sex, n	411	–.90 (–2.72 to .93)	.34	1335	.90 (–.93 to 2.72)	.34
Age (years), mean (range)	49 (18 to 89)	.01 (.00 to .02)	.06	44 (18 to 93)	.01 (.00 to .01)	.07
BMI (kg/m²), mean (range)	25.0 (16.5 to 46.8)	0 (–.04 to .04)	.99	25.9 (14.3 to 52.4)	.01 (–.01 to .03)	.32
Smoking, n (%)						
No	350 (85.2)	Ref[a]	Ref	1104 (82.7)	Ref	Ref
Yes	60 (14.6)	–.19 (–.65 to .27)	.42	226 (16.9)	–.28 (–.51 to –.04)	.02
Missing	1 (0.2)	N/A[b]	N/A	5 (0.4)	N/A	N/A
Exercise (h/wk), n (%)						
No sport	185 (45)	Ref	Ref	609 (45.6)	Ref	Ref
0-1	29 (7.1)	.06 (–.58 to .70)	.86	128 (9.6)	.25 (–.07 to .57)	.13
2-3	88 (21.4)	–.17 (–.59 to .24)	.42	386 (29)	–.17 (–.38 to .04)	.11
3-6	77 (18.7)	–.27 (–.72 to .18)	.25	157 (11.7)	–.01 (–.30 to .28)	.95
7-10	20 (4.9)	–.85 (–1.60 to –.10)	.03	37 (2.8)	.08 (–.46 to .61)	.78
>10	10 (2.4)	–.51 (–1.58 to .56)	.35	12 (0.9)	–.41 (–1.31 to .50)	.38
Missing	2 (0.5)	N/A	N/A	6 (0.4)	N/A	N/A
Education, n (%)						
Primary school	7 (1.7)	Ref	Ref	26 (1.9)	Ref	Ref
Secondary school	63 (15.3)	–.98 (–2.25 to .28)	.13	244 (18.3)	–.27 (–.95 to .41)	.44
Secondary vocational school	84 (20.4)	–1.55 (–2.80 to –.31)	.02	381 (28.5)	–.13 (–.79 to .53)	.70
Higher professional education or up	249 (60.6)	–1.12 (–2.34 to .10)	.07	649 (48.6)	–.41 (–1.08 to .25)	.23
No education	3 (0.7)	.01 (–2.18 to 2.19)	.99	19 (1.4)	.08 (–.89 to 1.05)	.87
Missing	5 (1.2)	N/A	N/A	16 (1.2)	N/A	N/A
Current pain score (numeric rating scale score), n (%)						
No pain	183 (44.5)	Ref	Ref	314 (23.6)	Ref	Ref
Mild (1-3)	120 (29.2)	–.16 (–.62 to .29)	.48	337 (25.4)	.10 (–.18 to .37)	.50
Moderate (4-7)	84 (20.5)	.06 (–.40 to .51)	.80	500 (37.7)	.96 (.71 to 1.20)	<.001
Severe (8-10)	24 (5.8)	1.00 (.28 to 1.72)	.01	177 (13.3)	1.65 (1.34 to 1.97)	<.001
Missing	0 (0)	N/A	N/A	7 (0.1)	N/A	N/A
Duration of pain (months), n (%)						
No pain	183 (44.5)	Ref	Ref	314 (23.5)	Ref	Ref
Shorter than 3 months	55 (13.4)	.50 (–.04 to 1.03)	.07	155 (11.6)	.06 (–.24 to .36)	.70
Longer than 3 months	168 (40.9)	.71 (–.12 to 1.54)	.09	850 (63.6)	–.04 (–.46 to .39)	.87
Missing	5 (1.2)	N/A	N/A	16 (1.2)	N/A	N/A

[a]Ref: reference.

[b]N/A: not applicable.

To clarify the results, we used a model in the form of a hypothetical 40-year-old nonsmoking respondent with a BMI of 25, who had no pain, exercises 2-3 h/wk, and had completed higher education. If this respondent was a female, she had an estimated PSQ score of 3.34. If this respondent was a male, his estimated PSQ score was 3.51 in this model. However, if this respondent had the same demographic characteristics but was experiencing moderate pain at the moment, the PSQ score for

XSL•FO
RenderX

a female respondent would be 4.29, while that for the male respondent would be 3.57. Further, if this respondent was experiencing severe pain, the PSQ score was higher in both sexes compared to the respondents without pain or with moderate pain (female 4.99, male 4.52). This exemplifies how various factors can influence pain sensitivity differentially in men and women.

Film and Pain Sensitivity

The distribution of demographics in the control and nature film groups was similar due to randomization, and there was no difference in the PSQ scores between the group that saw a film with nature and those who did not see a film before filling in the PSQ, which makes it possible to use the nature film group as a neutral film and reference, thereby overcoming the limitation of the smaller sample size of the control group. When considering the complete group, the film with contagious laughter had no significant effect on the PSQ scores in men or women. In the group of respondents who watched the film in which people endured painful situations, both men and women documented a lower score on the PSQ (–.20 for men and –.37 for women) (Figure 3). Even though the difference between men and women was not statistically significant (*P*=.07), the outcome of the differential impact that background variables have on men and women made us decide to investigate whether this would translate into a different effect of the films as well. This revealed that even the effect of a short film on pain sensitivity is dependent on some background factors and that this differs between sexes.

Figure 3. Effects of the types of films on the Pain Sensitivity Questionnaire scores of men and women.

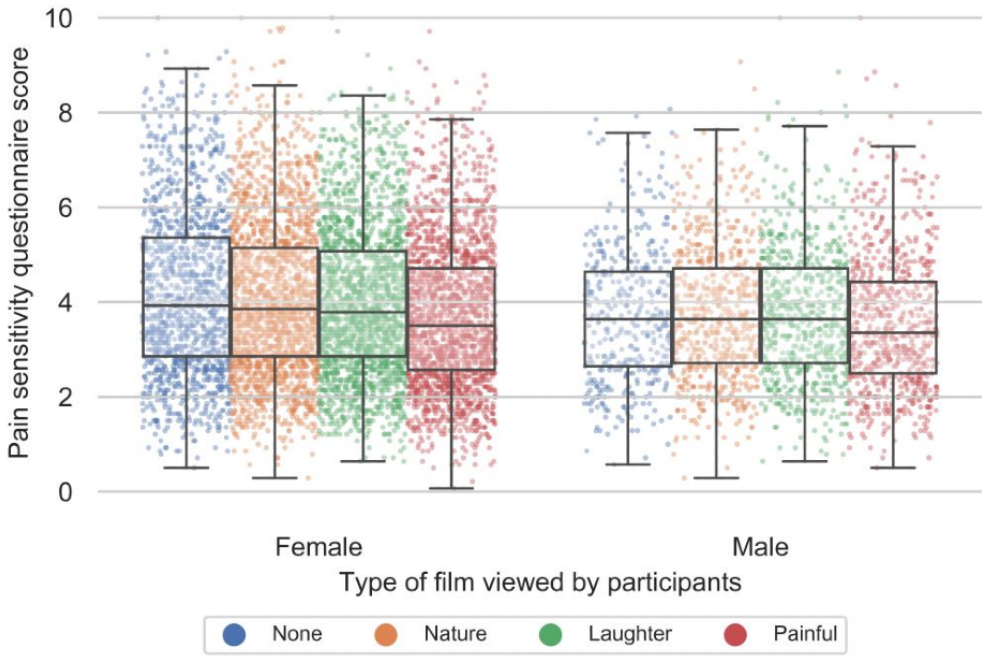

Sex, BMI, and (the degree of) current pain increased the effect of contagious laughter and painful event films. Sex and current pain were both influential factors. The combination of being female and experiencing pain had an additional lowering effect on the PSQ score. The other background factors that showed effect on the pain sensitivity did not influence the effect of the films on pain sensitivity. When circling back to the prior used hypothetical 40-year-old respondent (nonsmoking, with a BMI of 25, no current pain, exercise of 2-3 h/wk, and a higher education), we only see a minimal difference in the PSQ score between both female (3.31) and male (3.25) respondents after watching a painful event film (Figure 4, respondent 1). However, the film with painful situations had a larger positive effect in the model when the respondent was experiencing moderate pain at the moment (female 3.84, male 3.80, Figure 4, respondent 2). In the respondent who experienced severe pain, both sexes had a lower score on the PSQ after watching the contagious laughter film (female 4.68, male 4.02) and after watching the painful event film (female 4.21, male –4.20, Figure 4, respondent 3). These estimated PSQ scores and effects differ when other influencing background variables are present (Table 3, Figure 4).

XSL•FO
RenderX

Figure 4. Scores of hypothetical respondents on the Pain Sensitivity Questionnaire. L: film on laughter; P: film on painful events; PSQ: Pain Sensitivity Questionnaire.

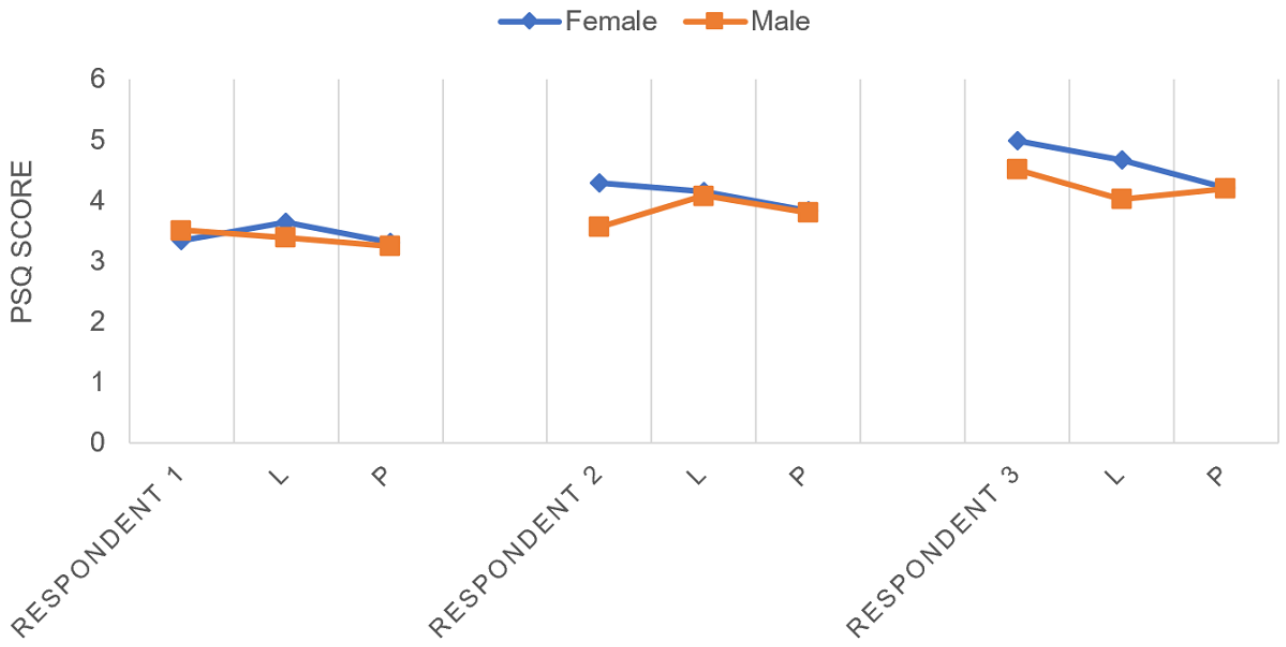

XSL•FO
RenderX

Table 3. Effects of all the factors and the effects of a combination of factors with the intervention in men and women on the Pain Sensitivity Questionnaire scores.

Factor	Effect in women, β (95% CI)	*P* value	Effect in men, β (95% CI)	*P* value
(Intercept)	3.37 (2.98 to 3.76)	<.001	3.17 (2.75 to 3.59)	<.001
Film with contagious laughter vs neutral film	–.14 (–.56 to .28)	.52	–.19 (–.65 to .28)	.43
Film with painful scenes vs neutral film	.12 (–.29 to .053)	.57	.26 (–.20 to .72)	.27
Age (years)	.01 (0 to .01)	<.001	.01 (0 to .01)	<.001
BMI (kg/m^2)	.01 (0 to .02)	.048	.01 (0 to .02)	.048
Smoking vs nonsmoking	–.23 (–.32 to –.13)	<.001	–.23 (–.32 to –.13)	<.001
Exercise 0-1 h/wk vs none	.13 (.02 to .25)	.02	.13 (.02 to .25)	.02
Exercise 2-3 h/wk vs none	0 (–.08 to .08)	.97	0 (–.08 to .08)	.97
Exercise 3-6 h/wk vs none	–.06 (–.16 to .05)	.30	–.06 (–.16 to .05)	.30
Exercise 7-10 h/wk vs none	–.27 (–.44 to –.09)	.003	–.27 (–.44 to –.09)	.003
Exercise >10 h/wk vs none	–.31 (–.63 to .01)	.06	–.31 (–.63 to .01)	.06
Secondary school vs primary school	–.10 (–.34 to .14)	.42	–.10 (–.34 to .14)	.42
Secondary vocational school vs primary school	–.14 (–.37 to .1)	.26	–.14 (–.37 to .1)	.26
Higher professional education or up vs primary school	–.36 (–.60 to –.13)	.002	–.36 (–.60 to –.13)	.002
No education vs primary school	.20 (–.18 to .57)	.29	.20 (–.18 to .57)	.29
NRS[a] mild pain score 1-3 vs no pain	.07 (–.12 to .25)	.49	.26 (–.02 to .54)	.07
NRS moderate pain score 4-7 vs no pain	.72 (.55 to .89)	<.001	.65 (.33 to .96)	<.001
NRS severe pain score 8-10 vs no pain	1.32 (1.08 to 1.56)	<.001	1.01 (.46 to 1.55)	<.001
Duration of pain (shorter than 3 months vs no pain)	.01 (–.10 to .11)	.91	.01 (–.10 to .11)	.91
Duration of pain (longer than 3 months vs no pain)	–.05 (–.22 to .13)	.61	–.05 (–.22 to .13)	.61
Film with contagious laughter and NRS mild pain score 1-3	–.16 (–.41 to .09)	.22	–.05 (–.43 to .34)	.80
Film with painful scenes and NRS mild pain score 1-3	–.03 (–.28 to .23)	.82	–.23 (–.61 to .15)	.23
Film with contagious laughter and NRS moderate pain score 4-7	–.21 (–.45 to .02)	.08	.04 (–.38 to .45)	.87
Film with painful scenes and NRS moderate pain score 4-7	–.19 (–.44 to .05)	.12	–.10 (–.52 to .32)	.64
Film with contagious laughter and NRS severe pain score 8-10	–.28 (–.62 to .07)	.12	–.37 (–1.14 to .39)	.34
Film with painful scenes and NRS severe pain score 8-10	–.42 (–.76 to –.07)	.02	–.06 (–.80 to .67)	.86
Film with contagious laughter and BMI	.01 (–.01 to .02)	.26	.01 (–.01 to .02)	.26
Film with painful scenes and BMI	–.01 (–.03 to 0)	.07	–.01 (–.03 to 0)	.07

[a]NRS: numeric rating scale.

Discussion

Principal Results

Our study results indicate that determining whether men or women have a higher sensitivity to pain requires considering the influence of other variables. These factors have differential effects on the pain sensitivity in men and women. Differences in effects are even present in a simple, although not validated, intervention, showing the importance of always considering individual factors in treatment options and future studies. Only investigating the effect of an intervention and the influence of sex provides a limited perspective on the complex multidimensional problem of pain between sexes.

Limitations

The cross-sectional analysis of demographics encompassed the entire study population, while randomization occurred before PSQ completion, resulting in a smaller group available for analyzing sex differences and the influence of background variables on the PSQ scores. Additionally, a lower number of respondents completed the questionnaire in this group compared to those who watched a short film first. In retrospect, the PSQ could have been filled in at baseline at the start of the web-based

XSL•FO
RenderX

study and after viewing the film. Although we only included complete questionnaires, we lacked control over the interpretation of the respondents and completion of the questionnaires. Furthermore, an uneven distribution of sex, with more women participating, and a high percentage of respondents experiencing chronic pain may introduce bias. People with chronic pain are more likely to be interested and to participate in this project. We addressed this concern by conducting multivariable regression analyses to correct for potential confounders.

Within-sex differences in pain sensitivity linked to gender expression have been reported previously [25]. For example, men who consider themselves as more masculine show higher thresholds for pain [26]. It is therefore very plausible that gender expression has an effect on pain perception and treatment as well, but our data were not sufficient to support or reject this hypothesis. In the previous baseline questionnaires [25,26], respondents had the option to fill in their gender if their gender identity did not correspond to their biological sex. The number of respondents who identified with this option was low and the demographic data were inconclusive. Psychological and other socioeconomic factors were not considered in our study as well. It is plausible that these factors influence pain and interventions on pain, and we would recommend taking these factors into account in future research.

We used the nature film as a neutral (control) group because there were fewer respondents in the control group. The short films were composed via a discussion-based focus group through evoked emotions, but these films were not validated. At the time, we did not have access to validated sets of films. We wanted to make our study more appealing for respondents. When looking at the smaller number of completed questionnaires in respondents who had not seen a film, the supposed appeal of the film was correct. For further research, the group could be asked about the evoked emotions in our used films for validation.

Statistical Analysis

In the case of our study, we had a set of independent variables that were potentially related with pain sensitivity. In case of redundant variables such as year of birth and age, full inclusion would give multicollinearity that can be eliminated without any downside. Either year of birth or age could be used. However, such redundancies are not present in the data. Overall, little multicollinearity is seen, with one exception. Pain duration and NRS score are correlated and show high variance inflation factors (above 10). This can be eliminated by removing either one of them from the model. However, in that case, no insights on the relation of the removed variable and pain sensitivity would be gained. Insights on the remaining variables would be confounded by the effects of the variable removed. Therefore, we decided to pay the price for allowing the multicollinearity.

Comparison With Prior Work

Women have been regarded as the more pain-sensitive sex, with previous studies reporting heightened pain responses in females. Although the greater pain sensitivity in women is evident in specific experimental settings [27,28], our large cohort study does not report a higher pain sensitivity in women with use of the PSQ. Different background factors (BMI, highest level of education, smoking level, and current pain level) influence pain sensitivity differently in women compared with that in men, as investigated in our control group (n=1746). Experiencing pain is one of the most important risk factors for developing chronic pain in another site of the body, suggesting that the presence of pain influences pain sensitivity [8]. This is supported by our data. In women who experience moderate to severe pain, the PSQ scores were notably higher. Similarly, severe pain exerted a negative influence on pain sensitivity in men, although the limited sample size of this subgroup prevented definitive conclusions. These findings align with previous research, highlighting sensitization in individuals with chronic pain and elevated PSQ scores in such patients [21,29].

Smoking emerged as another influential factor, displaying distinct effects in women. This contradicts previous results, which show that smokers experience a higher pain intensity [30]. However, the overall impact of smoking on chronic pain states remains inconclusive, with modest evidence suggesting an association [31,32]. Although people with lower education levels often experience chronic pain more frequently and with greater intensity, our data did not show an effect of education level on pain sensitivity in women [8]. In men, higher education influences the PSQ score positively, leading to a lower score on the PSQ. This may support the notion that sociocultural characteristics affect men more than women [8,33]. Contrary to previous findings, our study did not observe a positive effect of physical activity [34].

In addition to investigating background factors, we examined the impact of watching a film on pain sensitivity. We found that respondents experiencing pain had a lower score on the PSQ after watching a film with painful situations (Table 2). An explanation could be that context can relieve pain and that the situations stated in the PSQ are considered more relative for the respondents to what is shown on the film [35]. This result is in contrast to the empathy for pain theory, in which respondents experience pain themselves when watching others endure a painful situation [36,37]. There was no observed effect of the other films, contrary to previously performed research in smaller numbers on the effect of social laughter or funny films on pain [18,38]. This may suggest that context plays a crucial role in pain relief. Participants were not in a social setting with no researcher present when viewing the film—a setting that is known to affect pain perception [39,40].

The effects of audiovisual stimuli were different between sexes when considering the effects of other background factors. Sex, BMI, and (the level of) current pain influence the effects of both the contagious laughter film and the painful event film. Sex, BMI, and current pain are both independent influential factors, but being a female with current pain had an extra lowering effect on the PSQ. A systematic review on obesity and pain sensitivity suggested a correlation between higher BMI and increased pain sensitivity, although the impact of an intervention and the role of BMI remain understudied [41].

XSL•FO
RenderX

Conclusions

In our full cohort, the scores on the PSQ were not different between men and women. When exploring other factors, we do see a difference of their effects on the PSQ scores in men and women. These results underscore the importance of considering sex and sex-dependent background variables to obtain a comprehensive understanding of pain sensitivity. Hypotheses regarding pain sensitivity cannot be definitively accepted or rejected without accounting for these influential factors. Studying the effects of background factors in different treatment strategies could lead to further development of better individualized multimodal treatment for patients with pain, as has been proposed by other researchers [33,42-44].

Acknowledgments

We thank the Dutch Society for Scientific Research (NWO) and the independent Dutch public service broadcaster specializing in information, education, and culture (NTR) for partly funding and broadcasting for exposure. We thank Amon Heijne, as occasional graphic designer, for depicting the visualization of the results in figures. We thank Dominique (Kiki) Stolzenbach, as a member of the research group, and Marieke van Nieuwland for promoting and social media exposure. We also thank Fabiënne Balvers for editing the films, Basten Stokhuyzen as the web designer and developer of the website, and Diana v d Berg-van Daalen Meijer for administrative support during the research development.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Raja SN, Carr DB, Cohen M, Finnerup NB, Flor H, Gibson S, et al. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. Pain 2020 Sep 01;161(9):1976-1982 [FREE Full text] [doi: 10.1097/j.pain.0000000000001939] [Medline: 32694387]
2. Edwards RR. Individual differences in endogenous pain modulation as a risk factor for chronic pain. Neurology 2005 Aug 09;65(3):437-443. [doi: 10.1212/01.wnl.0000171862.17301.84] [Medline: 16087910]
3. Picavet HSJ, Monique Verschuren WM, Groot L, Schaap L, van Oostrom SH. Pain over the adult life course: 15-year pain trajectories-The Doetinchem Cohort Study. Eur J Pain 2019 Oct;23(9):1723-1732 [FREE Full text] [doi: 10.1002/ejp.1450] [Medline: 31257661]
4. Mundal I, Gråwe RW, Bjørngaard JH, Linaker OM, Fors EA. Prevalence and long-term predictors of persistent chronic widespread pain in the general population in an 11-year prospective study: the HUNT study. BMC Musculoskelet Disord 2014 Jun 20;15:213 [FREE Full text] [doi: 10.1186/1471-2474-15-213] [Medline: 24951013]
5. Glare P, Aubrey KR, Myles PS. Transition from acute to chronic pain after surgery. The Lancet 2019 Apr;393(10180):1537-1546. [doi: 10.1016/s0140-6736(19)30352-6]
6. Kalkman GA, Kramers C, van Dongen RT, van den Brink W, Schellekens A. Trends in use and misuse of opioids in the Netherlands: a retrospective, multi-source database study. Lancet Public Health 2019 Oct;4(10):e498-e505 [FREE Full text] [doi: 10.1016/S2468-2667(19)30128-8] [Medline: 31444001]
7. Ballantyne JC. Opioids for the treatment of chronic pain: mistakes made, lessons learned, and future directions. Anesth Analg 2017 Nov;125(5):1769-1778. [doi: 10.1213/ANE.0000000000002500] [Medline: 29049121]
8. Mills SEE, Nicolson KP, Smith BH. Chronic pain: a review of its epidemiology and associated factors in population-based studies. Br J Anaesth 2019 Aug;123(2):e273-e283 [FREE Full text] [doi: 10.1016/j.bja.2019.03.023] [Medline: 31079836]
9. Mogil JS. Qualitative sex differences in pain processing: emerging evidence of a biased literature. Nat Rev Neurosci 2020 Jul;21(7):353-365. [doi: 10.1038/s41583-020-0310-6] [Medline: 32440016]
10. Pieretti S, Di Giannuario A, Di Giovannandrea R, Marzoli F, Piccaro G, Minosi P, et al. Gender differences in pain and its relief. Ann Ist Super Sanita 2016;52(2):184-189 [FREE Full text] [doi: 10.4415/ANN_16_02_09] [Medline: 27364392]
11. Ruau D, Liu LY, Clark JD, Angst MS, Butte AJ. Sex differences in reported pain across 11,000 patients captured in electronic medical records. J Pain 2012 Mar;13(3):228-234 [FREE Full text] [doi: 10.1016/j.jpain.2011.11.002] [Medline: 22245360]
12. Niesters M, Dahan A, Kest B, Zacny J, Stijnen T, Aarts L, et al. Do sex differences exist in opioid analgesia? A systematic review and meta-analysis of human experimental and clinical studies. Pain 2010 Oct;151(1):61-68. [doi: 10.1016/j.pain.2010.06.012] [Medline: 20692097]
13. Soldin OP, Mattison DR. Sex differences in pharmacokinetics and pharmacodynamics. Clin Pharmacokinet 2009;48(3):143-157 [FREE Full text] [doi: 10.2165/00003088-200948030-00001] [Medline: 19385708]
14. Pieh C, Altmeppen J, Neumeier S, Loew T, Angerer M, Lahmann C. Gender differences in outcomes of a multimodal pain management program. Pain 2012 Jan;153(1):197-202. [doi: 10.1016/j.pain.2011.10.016] [Medline: 22100358]

XSL•FO
RenderX

15. Jensen IB, Bergström G, Ljungquist T, Bodin L, Nygren AL. A randomized controlled component analysis of a behavioral medicine rehabilitation program for chronic spinal pain: are the effects dependent on gender? Pain 2001 Mar;91(1-2):65-78. [doi: 10.1016/s0304-3959(00)00420-6] [Medline: 11240079]
16. Bartley EJ, Fillingim RB. Sex differences in pain: a brief review of clinical and experimental findings. Br J Anaesth 2013 Jul;111(1):52-58 [FREE Full text] [doi: 10.1093/bja/aet127] [Medline: 23794645]
17. Sibille KT, Kindler LL, Glover TL, Gonzalez RD, Staud R, Riley JL, et al. Individual differences in morphine and butorphanol analgesia: a laboratory pain study. Pain Med 2011 Jul;12(7):1076-1085 [FREE Full text] [doi: 10.1111/j.1526-4637.2011.01157.x] [Medline: 21668741]
18. Dunbar RIM, Baron R, Frangou A, Pearce E, van Leeuwen EJC, Stow J, et al. Social laughter is correlated with an elevated pain threshold. Proc Biol Sci 2012 Mar 22;279(1731):1161-1167 [FREE Full text] [doi: 10.1098/rspb.2011.1373] [Medline: 21920973]
19. Yang C, Decety J, Lee S, Chen C, Cheng Y. Gender differences in the mu rhythm during empathy for pain: an electroencephalographic study. Brain Res 2009 Jan 28;1251:176-184. [doi: 10.1016/j.brainres.2008.11.062] [Medline: 19083993]
20. Ruscheweyh R, Marziniak M, Stumpenhorst F, Reinholz J, Knecht S. Pain sensitivity can be assessed by self-rating: Development and validation of the Pain Sensitivity Questionnaire. Pain 2009 Nov;146(1-2):65-74. [doi: 10.1016/j.pain.2009.06.020] [Medline: 19665301]
21. Ruscheweyh R, Verneuer B, Dany K, Marziniak M, Wolowski A, Çolak-Ekici R, et al. Validation of the pain sensitivity questionnaire in chronic pain patients. Pain 2012 Jun;153(6):1210-1218. [doi: 10.1016/j.pain.2012.02.025] [Medline: 22541722]
22. Grundström H, Larsson B, Arendt-Nielsen L, Gerdle B, Kjølhede P. Associations between pain thresholds for heat, cold and pressure, and Pain Sensitivity Questionnaire scores in healthy women and in women with persistent pelvic pain. Eur J Pain 2019 Oct;23(9):1631-1639 [FREE Full text] [doi: 10.1002/ejp.1439] [Medline: 31192501]
23. Johnson AM, Smith SM. A review of general pain measurement tools and instruments for consideration of use in COPD clinical practice. Int J Chron Obstruct Pulmon Dis 2017;12:923-929 [FREE Full text] [doi: 10.2147/COPD.S119889] [Medline: 28360515]
24. Van Boekel RLM, Timmerman H, Bronkhorst EM, Ruscheweyh R, Vissers KCP, Steegers MAH. Translation, cross-cultural adaptation, and validation of the pain sensitivity questionnaire in Dutch healthy volunteers. Pain Res Manag 2020;2020:1050935 [FREE Full text] [doi: 10.1155/2020/1050935] [Medline: 32774564]
25. Vigil JM, Rowell LN, Lutz C. Gender expression, sexual orientation and pain sensitivity in women. Pain Res Manag 2014;19(2):87-92 [FREE Full text] [doi: 10.1155/2014/297060] [Medline: 24575419]
26. Alabas OA, Tashani OA, Tabasam G, Johnson MI. Gender role affects experimental pain responses: a systematic review with meta-analysis. Eur J Pain 2012 Oct;16(9):1211-1223. [doi: 10.1002/j.1532-2149.2012.00121.x] [Medline: 22434689]
27. Hashmi JA, Davis KD. Deconstructing sex differences in pain sensitivity. Pain 2014 Jan;155(1):10-13. [doi: 10.1016/j.pain.2013.07.039] [Medline: 23891901]
28. Mogil JS. Sex differences in pain and pain inhibition: multiple explanations of a controversial phenomenon. Nat Rev Neurosci 2012 Dec;13(12):859-866. [doi: 10.1038/nrn3360] [Medline: 23165262]
29. Stabell N, Stubhaug A, Flægstad T, Nielsen CS. Increased pain sensitivity among adults reporting irritable bowel syndrome symptoms in a large population-based study. Pain 2013 Mar;154(3):385-392. [doi: 10.1016/j.pain.2012.11.012] [Medline: 23320954]
30. Jakobsson U. Tobacco use in relation to chronic pain: results from a Swedish population survey. Pain Med 2008 Nov;9(8):1091-1097. [doi: 10.1111/j.1526-4637.2008.00473.x] [Medline: 18565000]
31. Ditre JW, Brandon TH, Zale EL, Meagher MM. Pain, nicotine, and smoking: research findings and mechanistic considerations. Psychol Bull 2011 Nov;137(6):1065-1093 [FREE Full text] [doi: 10.1037/a0025544] [Medline: 21967450]
32. Jamner LD, Girdler SS, Shapiro D, Jarvik ME. Pain inhibition, nicotine, and gender. Exp Clin Psychopharmacol 1998 Feb;6(1):96-106. [doi: 10.1037//1064-1297.6.1.96] [Medline: 9526150]
33. Waardenburg S, de Meij N, Brouwer B, van Zundert J, van Kuijk SM. Cross-sectional study on sex differences in chronic pain patients using the DATAPAIN cohort. Reg Anesth Pain Med 2021 Nov;46(11):997-999. [doi: 10.1136/rapm-2020-102385] [Medline: 33674427]
34. Årnes AP, Nielsen CS, Stubhaug A, Fjeld MK, Hopstock LA, Horsch A, et al. Physical activity and cold pain tolerance in the general population. Eur J Pain 2021 Mar;25(3):637-650. [doi: 10.1002/ejp.1699] [Medline: 33165994]
35. Leknes S, Berna C, Lee MC, Snyder GD, Biele G, Tracey I. The importance of context: when relative relief renders pain pleasant. Pain 2013 Mar;154(3):402-410 [FREE Full text] [doi: 10.1016/j.pain.2012.11.018] [Medline: 23352758]
36. Bruneau E, Dufour N, Saxe R. How we know it hurts: item analysis of written narratives reveals distinct neural responses to others' physical pain and emotional suffering. PLoS One 2013;8(4):e63085 [FREE Full text] [doi: 10.1371/journal.pone.0063085] [Medline: 23638181]
37. Fitzgibbon BM, Giummarra MJ, Georgiou-Karistianis N, Enticott PG, Bradshaw JL. Shared pain: from empathy to synaesthesia. Neurosci Biobehav Rev 2010 Mar;34(4):500-512. [doi: 10.1016/j.neubiorev.2009.10.007] [Medline: 19857517]

XSL•FO
RenderX

38. Zweyer K, Velker B, Ruch W. Do cheerfulness, exhilaration, and humor production moderate pain tolerance? A FACS study. Humor - International Journal of Humor Research. (1-2) 2004;17:85-111. [doi: 10.1515/humr.2004.009]
39. Kállai I, Barke A, Voss U. The effects of experimenter characteristics on pain reports in women and men. Pain 2004 Nov;112(1-2):142-147. [doi: 10.1016/j.pain.2004.08.008] [Medline: 15494194]
40. Vo L, Drummond PD. "Big girls don't cry": the effect of the experimenter's sex and pain catastrophising on pain. Scand J Pain 2021 Jul 27;21(3):617-627 [FREE Full text] [doi: 10.1515/sjpain-2020-0157] [Medline: 33565286]
41. Torensma B, Oudejans L, van Velzen M, Swank D, Niesters M, Dahan A. Pain sensitivity and pain scoring in patients with morbid obesity. Surg Obes Relat Dis 2017 May;13(5):788-795. [doi: 10.1016/j.soard.2017.01.015] [Medline: 28216116]
42. Bauer GR. Incorporating intersectionality theory into population health research methodology: challenges and the potential to advance health equity. Soc Sci Med 2014 Jun;110:10-17 [FREE Full text] [doi: 10.1016/j.socscimed.2014.03.022] [Medline: 24704889]
43. Nijs J, Leysen L, Vanlauwe J, Logghe T, Ickmans K, Polli A, et al. Treatment of central sensitization in patients with chronic pain: time for change? Expert Opin Pharmacother 2019 Nov;20(16):1961-1970. [doi: 10.1080/14656566.2019.1647166] [Medline: 31355689]
44. Zimmer Z, Fraser K, Grol-Prokopczyk H, Zajacova A. A global study of pain prevalence across 52 countries: examining the role of country-level contextual factors. Pain 2022 Sep 01;163(9):1740-1750 [FREE Full text] [doi: 10.1097/j.pain.0000000000002557] [Medline: 35027516]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
NRS: numeric rating scale
PSQ: Pain Sensitivity Questionnaire

Edited by S He; submitted 02.11.23; peer-reviewed by J Mogil, B Korwisi; comments to author 25.02.24; revised version received 19.04.24; accepted 12.06.24; published 09.08.24.

Please cite as:
Smits RJH, van der Wal SEI, van Boekel RLM, Timmerman H, Bronkhorst EM, Abrar D, Vissers KCP, Blaney Davidson EN, Steegers MAH
Sex Differences in Pain Sensitivity in a Dutch Cohort: Cross-Sectional and Web-Based Multidimensional Study
J Med Internet Res 2024;26:e53926
URL: https://www.jmir.org/2024/1/e53926
doi:10.2196/53926
PMID:

©Rachel J H Smits, Selina E I van der Wal, Regina L M van Boekel, Hans Timmerman, Ewald M Bronkhorst, Diana Abrar, Kris C P Vissers, Esmeralda N Blaney Davidson, Monique A H Steegers. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Mobile Health–Based Remote Interaction Management Intervention on the Quality of Life and Self-Management Behavior of Patients With Low Anterior Resection Syndrome: Randomized Controlled Trial

Peng Zhou[1,2*], MNS; Hui Li[3*], MSN; Xueying Pang[4*], MSN; Ting Wang[4], MSN; Yan Wang[2], MSN; Hongye He[4], MSN; Dongmei Zhuang[2], MSN; Furong Zhu[2], MNS; Rui Zhu[1], MSN; Shaohua Hu[1], PhD

[1]Department of Nursing, the First Affiliated Hospital of Anhui Medical University, Hefei, China

[2]School of Nursing, Anhui Medical University, Hefei, China

[3]College of Traditional Chinese Medicine, Bozhou University, Bozhou, China

[4]Department of Gastrointestinal Surgery, the First Affiliated Hospital of Anhui Medical University, Hefei, China

[*]these authors contributed equally

Corresponding Author:
Shaohua Hu, PhD
Department of Nursing
the First Affiliated Hospital of Anhui Medical University
218 Jixi Road
Hefei, 230009
China
Phone: 86 62922005
Email: hushaohua1003@126.com

Abstract

Background: People who undergo sphincter-preserving surgery have high rates of anorectal functional disturbances, known as low anterior resection syndrome (LARS). LARS negatively affects patients' quality of life (QoL) and increases their need for self-management behaviors. Therefore, approaches to enhance self-management behavior and QoL are vital.

Objective: This study aims to assess the effectiveness of a remote digital management intervention designed to enhance the QoL and self-management behavior of patients with LARS.

Methods: From July 2022 to May 2023, we conducted a single-blinded randomized controlled trial and recruited 120 patients with LARS in a tertiary hospital in Hefei, China. All patients were randomly assigned to the intervention group (using the "e-bowel safety" applet and monthly motivational interviewing) or the control group (usual care and an information booklet). Our team provided a 3-month intervention and followed up with all patients for an additional 3 months. The primary outcome was patient QoL measured using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30. The secondary outcomes were evaluated using the Bowel Symptoms Self-Management Behaviors Questionnaire, LARS score, and Perceived Social Support Scale. Data collection occurred at study enrollment, the end of the 3-month intervention, and the 3-month follow-up. Generalized estimating equations were used to analyze changes in all outcome variables.

Results: In the end, 111 patients completed the study. In the intervention group, 5 patients withdrew; 4 patients withdrew in the control group. Patients in the intervention group had significantly larger improvements in the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 total score (mean difference 11.51; 95% CI 10.68-12.35; Cohen d=1.73) and Bowel Symptoms Self-Management Behaviors Questionnaire total score (mean difference 8.80; 95% CI 8.28-9.32; Cohen d=1.94) than those in the control group. This improvement effect remained stable at 3-month follow-up (mean difference 14.47; 95% CI 13.65-15.30; Cohen d=1.58 and mean difference 8.85; 95% CI 8.25-9.42; Cohen d=2.23). The LARS score total score had significantly larger decreases after intervention (mean difference –3.28; 95% CI –4.03 to –2.54; Cohen d=–0.39) and at 3-month follow-up (mean difference –6.69; 95% CI –7.45 to –5.93; Cohen d=–0.69). The Perceived Social Support Scale total score had significantly larger improvements after intervention (mean difference 0.47; 95% CI 0.22-0.71; Cohen d=1.81).

XSL•FO
RenderX

Conclusions: Our preliminary findings suggest that the mobile health–based remote interaction management intervention significantly enhanced the self-management behaviors and QoL of patients with LARS, and the effect was sustained. Mobile health–based remote interventions become an effective method to improve health outcomes for many patients with LARS.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2200061317; https://tinyurl.com/tmmvpq3

(J Med Internet Res 2024;26:e53909) doi:10.2196/53909

KEYWORDS

mHealth; low anterior resection syndrome; quality of life; sphincter-preserving surgery; randomized controlled trial; mobile health; mobile phone

Introduction

Background

The Global Cancer Statistics 2020 showed that colorectal cancer ranks third in incidence of malignant tumors and second in cause of death worldwide [1]. Colorectal cancer incidence is also on the rise in China, with rectal cancer accounting for 60% of cases and middle and lower rectal cancers being the most common [2]. With the advancement of medical technology, optimal management of middle and lower rectal cancers increasingly favors sphincter-preserving surgery (SPS) [3]. This operation preserves anal function and avoids the inconvenience and pressure caused by permanent colostomy [4]. However, 70%-90% of patients after SPS struggle with long-term anorectal functional disturbances called low anterior resection syndrome (LARS) [5,6].

The presence of LARS has a severe adverse effect on the quality of life (QoL) of patients [7]. Postoperative LARS induces a spectrum of adverse physical and psychological effects in patients; for example, up to 50% of patients with LARS report toilet dependence during rehabilitation [8,9], 36% of patients experience pain, and approximately 13% of patients report high psychological distress [10,11]. Furthermore, LARS can restrict a patient's social life, leading to further impact on their QoL [12]. Recently, longitudinal studies have found that patients' QoL is still affected by LARS even 15 years after surgery [13]. Research has shown that patients can improve their QoL through methods, such as pelvic floor muscle exercises and dietary adjustments during home care; however, the effectiveness of these methods is limited by patients' lack of knowledge of LARS and rehabilitation guidance [14,15].

Owing to the frequent occurrence of LARS in patients post discharge, patients must have a high level of self-management behavior [16]. However, in China, the majority of patients have a passive response to LARS, and their self-management behavior is at a low level [17]. Enhancing self-management awareness and providing information on supportive care can improve the self-management behavior of patients with LARS [18]. Research has demonstrated that motivational interviewing (MI) enhances self-management awareness and supports behavioral change [19].

Therefore, to improve patients' QoL and self-management behaviors, providing supportive care information to patients is crucial. A qualitative exploration of patients with LARS's perspectives on information needs revealed that timely symptom management measures are critical during home-based rehabilitation [20]. However, it is difficult to maintain continuity and instantaneity with existing management measures [21,22]. Owing to current advances in mobile technology, mobile health (mHealth) has been widely considered a means of patient health management, which can improve the effects of symptoms and assist patients in timely access to the required information [23,24].

To date, remote follow-up tools for patients with LARS have yielded promising results [25]. For patients with LARS, mHealth-based remote interventions may become an effective method to assist them in improving symptoms. However, mHealth intervention measures constructed for patients with LARS are rare. Most studies have only completed the development and pilot research of remote intervention programs, leading to insufficient data on the effectiveness of remote interventions in improving patient health outcomes [26,27]. WeChat (Tencent Corp) is China's most frequently used instant messaging and social media application [28]. Evidence suggests that WeChat-based mHealth interventions effectively improve health outcomes in various health conditions [29,30].

Objective

This study aimed to assess the effectiveness of a remote digital management intervention designed for patients with LARS. The effectiveness of the intervention measure is determined by improvement in QoL, self-management behaviors, gastrointestinal symptoms, and social support. We hypothesized that the remote digital management intervention can effectively improve the health outcomes of patients with LARS.

Methods

Study Design

This study was conducted from July 15, 2022, to March 15, 2023, in Hefei, China. Our team provided a 3-month intervention and followed up with all patients for an additional 3 months. The intervention group used the "e-bowel safety" applet and received monthly MI. The control group received the usual care and was provided with a handbook containing information related to LARS. The CONSORT (Consolidated Standards of Reporting Trials) checklist is in Multimedia Appendix 1.

Ethical Considerations

This randomized controlled trial (RCT) was approved by the ethics committee of the First Affiliated Hospital of Anhui Medical University (PJ2022-07-53) and registered on the Chinese Clinical Trial Registry (ChiCTR2200061317). All data were identified with a code number to ensure the confidentiality

XSL•FO
RenderX

of the subjects' data. No compensation was provided to participants.

Participants

The patients were recruited from a tertiary hospital in Hefei, Anhui Province, China. Patients were eligible to participate in our study if they met the following criteria: age older than 18 years, a diagnosis of rectal cancer, underwent SPS, LARS scores ≥21, ostomy closure surgery performed at least 3 months prior, the ability to read and write text, and proficiency in using WeChat. Patients with chronic gastrointestinal conditions, prior or current mental health disorders, cognitive impairments, communication disorders, or those who have participated in other clinical studies are ineligible for participation in this research. When patients meeting the recruitment criteria appeared in the hospital database, the system sent recruitment information to these patients with the approval of doctors not directly involved in the research design.

In this study, the sample size was determined based on the QoL. Previous research has shown that the QoL for patients with rectal cancer is 77±19 [31]. In an RCT using the EORTC QLQ-C30, a difference of 10 points is considered clinically significant [32]. With a two-sided test level of 0.05 and 80% test efficacy, each group requires a sample size of 45. Accounting for a 20% dropout rate, 112 patients are needed.

Intervention

Our previous study provided a comprehensive description of the intervention protocol [33]. The patients in the intervention group used the "e-bowel safety" applet for 3 months. They were required to check in on the applet daily and record their daily gastrointestinal symptoms. Our "e-bowel safety" applet comprises 4 main sections: a rehabilitation plan, LARS knowledge, web-based consultation, and patient stories. The rehabilitation plan module involves the collaborative development of home dietary and exercise plans by patients and researchers. The applet features intelligent reminders to monitor daily plan completion and provide prompts. After completing the rehabilitation plan, patients must fill out a daily health diary, and researchers dynamically adjust the rehabilitation plan based on patients' feedback and physical condition. The LARS knowledge module offers evidence-based information on LARS and symptom management strategies. The web-based consultation module provides patients with an opportunity to interact with health care professionals, offering personalized guidance and feedback. The patient stories module allows patients to share symptom management experiences or engage with other patients, with all published content subject to researcher approval. Additionally, an incentive system has been designed to encourage participation. For instance, patients earn points by sharing personal stories or comments, which can later be exchanged for rewards after accumulating a certain number of points.

Moreover, our team members conducted monthly MIs with patients. MIs were led by 4 researchers with expertise in health coaching and disease management, including 1 clinical psychologist (Shangxin Zhang) and 3 registered nurses (TW, HH, and Ling Fang). The researchers engage with patients via WeChat for 30-60 minutes per call. The aim of MIs is to assist patients in setting rehabilitation goals, reinforcing self-management awareness, and promoting health behavior changes. The content of MIs is based on the interview guide determined by the research team, which guides the conversation from the initial session to explore the participant's motivation to identify the facilitating factors and barriers to achieving their health goals. The interview guide is outlined in Multimedia Appendix 2.

Patients in the control group received the usual care and were provided with a handbook containing information related to LARS. At the same time, our team members followed up with patients, using the same timing and frequency as the MI intervention group.

Randomization and Masking

This study was a single-blind, two-arm RCT. After obtaining consent from eligible patients, assistants who were not involved in the study randomly assigned them to the intervention and control groups at a 1:1 ratio. The randomization process was performed by the assistants and anonymized envelopes were used with block randomization, including block sizes randomly varying between 4 (2:2) and 6 (3:3). The research assistants (Ping Ni and Ai Wang) who collected the data were unaware of the patient assignments throughout the study. Patients used the QR codes provided by the research team to access the "e-bowel safety" applet, effectively reducing contamination between the 2 groups. Patients were blinded to their group assignments throughout the entire research process.

Quality Control and Participant Retention

Several strategies were used to ensure quality control and participant retention. Our "e-bowel safety" applet can monitor patients' plan execution and provide reminders, which ensures the daily plans are followed strictly by patients. Before the formal intervention, we conducted a pilot experiment and gathered participant feedback to enhance our plan. The specific results are included in Multimedia Appendix 3. Furthermore, patients received consistent guidance from our research assistants (Ping Ni and Ai Wang) when they had questions about the questionnaire content. Before the start of the study, all research assistants must undergo training and assessment on the use of all questionnaires by research team members. Only research assistants who pass the assessment can participate in data collection. Additionally, team members regularly check the progress of research assistants' work to ensure that they are following the questionnaire collection process, identifying issues promptly, and making corrections.

Outcome Measures

The patients' demographic and clinical information were obtained from the hospital database. Data were collected from patients using scales for their QoL, social support, self-management behaviors, and LARS scores at different time periods (0, 3, and 6 months). The research assistants (Ping Ni and Ai Wang) who collected the data assisted patients in completing questionnaires over the phone or through direct personal interaction.

XSL•FO
RenderX

Primary Outcome: QoL

The EORTC-QLQ-C30 (European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30) was used to measure QoL. This questionnaire comprises 30 items divided into 15 dimensions, including 1 dimension for QoL, 5 dimensions for functionality, 3 dimensions for symptoms, and 6 dimensions for additional symptoms. All dimension scores were linearly transformed to a scale of 0-100 points. Elevated scores on the 5 functionality dimensions and the QoL dimension were linked to improved functional status, whereas the reverse pattern was observed for the symptom dimensions and additional symptom dimensions. The Cronbach α coefficient ranged from 0.764 to 0.809 [34].

Secondary Outcome

Self-Management

The self-management behavior of patients was assessed by the Bowel Symptoms Self-Management Behaviors Questionnaire (BSSBQ). This questionnaire comprises 24 items divided into 5 functional scales, with each item scored on a scale of 0 (never) to 7 (always). Higher scores indicate better bowel symptom self-management behavior. The Cronbach α coefficient was 0.81 [17].

Bowel Function

The LARS score consists of 5 items, with a total score ranging from 0 to 42. Patients' gastrointestinal symptoms are classified into no LARS, minor LARS, and major LARS based on the total score. The LARS score is a validated instrument for assessing bowel symptoms. The Cronbach α coefficient was 0.767 [35].

Social Support

The Perceived Social Support Scale (PSSS) consists of 12 items, with each item scored on a scale of 1 (extreme disagreement) to 7 (strong consent). The total scores ranged from 12 to 84. The higher the score, the stronger the perceived social support by the patient. This scale is widely used to assess the level of social support among patients in China. The Cronbach α coefficient of this scale was 0.899 [36].

Feasibility

The feasibility of intervention was assessed through the completion status of MI sessions and the adherence to health diary entries. The 3-month intervention corresponds to 3 MI sessions and 84 days of health diary entries.

Statistical Methods

All data were analyzed using SPSS Statistics (version 23.0; IBM Corp). An intention-to-treat analysis was performed in this study. We used the last observed values of the patients to replace missing data. Chi-square analysis was used to analyze the remaining demographic characteristics, and a 2-tailed independent sample *t* test was used to analyze the age and tumor height. Descriptive data were computed, including means with SD, medians with ranges, and frequencies with proportions where appropriate. The statistical significance was established at $P<.05$ (2-tailed test). Generalized estimating equations were used to analyze changes in QoL, self-management behaviors, LARS, and social support scores at different time points. The calculation of effect sizes was performed using Cohen *d* for the mean differences at various time periods.

Results

Participant Characteristics

Initially, 60 patients were recruited in the control and intervention groups. During the study, 9 patients dropped out (dropout rate 7.5%). In the intervention group, 5 patients withdrew from the study, including 2 patients who received a reostomy because of an anastomotic fistula and 3 patients whose condition worsened. In the control group, 4 patients dropped out, including 2 patients whose condition worsened and 2 patients who refused to continue the intervention because of the side effects of chemotherapy. No statistically significant differences were observed between the patients who dropped out and those who completed all evaluations ($P=.17$). Figure 1 shows the CONSORT flowchart of this study. Table 1 demonstrates no statistically significant differences in the demographics and clinical information between the control and intervention groups at baseline.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart.

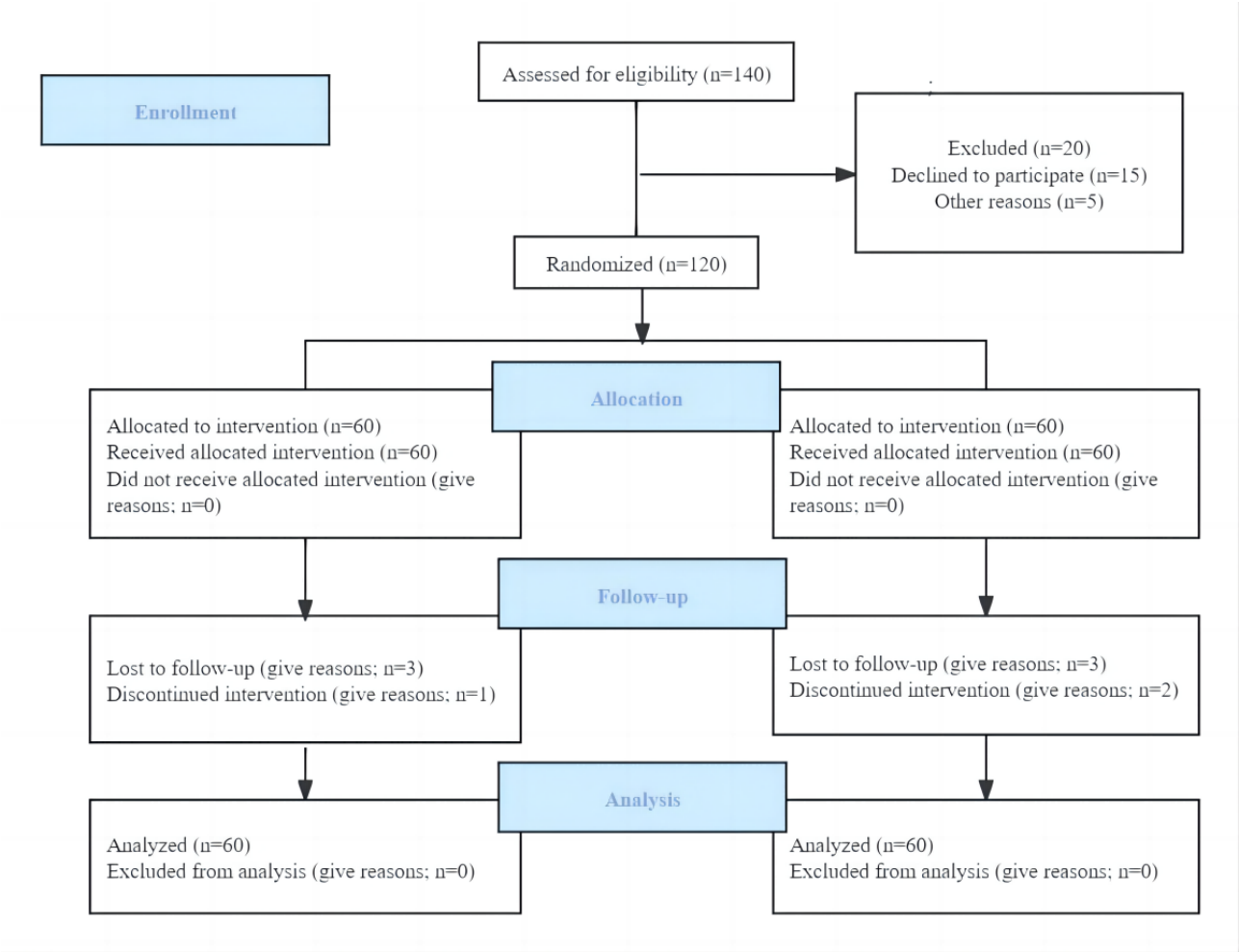

XSL•FO
RenderX

Table 1. Basic participant characteristics of the 2 groups.

Characteristics	Intervention group (n=60)	Control group (n=60)	*t* test (*df*) or chi-square value (*df*)	*P* value
Sex, n (%)			0.93 (1)	.34
Male	42 (70)	37 (62)		
Female	18 (30)	23 (38)		
Age (years), mean (SD)	62.72 (7.91)	61.78 (11.80)	0.51 (118)	.61
Education, n (%)			0.07 (2)	.96
Junior high school or lower	33 (55)	32 (53)		
High school	19 (32)	19 (32)		
College or higher	8 (13)	9 (15)		
Marital status, n (%)			0.21 (1)	.65
Married	58 (97)	57 (95)		
Single	2 (3)	3 (5)		
Tumor stage, n (%)			1.42 (3)	.70
I	14 (23)	13 (22)		
II	24 (40)	30 (50)		
III	20 (33)	15 (25)		
IV	2 (4)	2 (3)		
Tumor height, mean (SD)	7.62 (1.708)	7.80 (1.811)	–0.57 (118)	.57
Postoperative time (months), n (%)			0.378 (2)	.83
<6	18 (30)	17 (28)		
6-12	27 (45)	25 (42)		
>12	15 (25)	18 (30)		
Surgical procedures, n (%)			0.24 (1)	.62
Laparoscopy	51 (85)	49 (82)		
Laparotomy	9 (15)	11 (18)		
Surgical approach, n (%)			0.34 (1)	.56
LAR[a]	58 (9)	59 (98)		
TaTME[b]	2 (3)	1 (2)		
Temporary stoma, n (%)			0.53 (1)	.47
Yes	29 (48)	33 (55)		
No	31 (52)	27 (45)		
Chemotherapy, n (%)			0.88 (2)	.65
Preoperative	8 (13)	5 (8)		
Postoperative	49 (82)	51 (85)		
No	3 (5)	4 (7)		
Residence, n (%)			1.20 (1)	.27
Countryside	28 (47)	34 (57)		
City	32 (53)	26 (43)		
Measurements, mean (SD)				
EORTC-QLQ-C30[c]	69.67 (4.26)	69.42 (3.66)	0.35 (118)	.72
BSSBQ[d]	30.33 (1.90)	30.58 (2.01)	–0.70 (118)	.49
LARS[e] score	31.07 (3.88)	31.32 (4.73)	–0.32 (118)	.75

XSL•FO
RenderX

Characteristics	Intervention group (n=60)	Control group (n=60)	*t* test (*df*) or chi-square value (*df*)	*P* value
PSSS[f]	34.42 (1.62)	34.3 (1.48)	0.29 (118)	.77

[a]LAR: low anterior resection.

[b]TaTME: transanal total mesorectal excision.

[c]EORTC-QLQ-C30: European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30.

[d]BSSBQ: Bowel Symptoms Self-Management Behaviors Questionnaire.

[e]LARS: Low anterior resection syndrome.

[f]PSSS: Perceived Social Support Scale.

Main Evaluation Indexes

Table 2 shows that the patients' QoL improved for both groups. Patients in the intervention group demonstrated greater improvements in the EORTC-QLQ-C30 total score than those in the control group after intervention (mean difference 11.51; 95% CI 10.68-12.35; Cohen *d*=1.73). Furthermore, this improvement effect remained stable at 3-month follow-up (mean difference 14.47; 95% CI 13.65-15.30; Cohen *d*=1.58). Table 3 shows that the EORTC-QLQ-C30 total score in both groups exhibited a trend of change over the 6-month period (*P*<.001). Differences were observed between the 2 groups and the interaction between group and time. A subgroup analysis was conducted on patients receiving preoperative chemotherapy versus postoperative chemotherapy. Among the 49 patients in the intervention group and 51 in the control group undergoing postoperative chemotherapy, a nominally significant improvement in the change from baseline in the EORTC-QLQ-C30 total score at 3 months was observed compared to the control group (difference of 4.42; *P*<.001). However, this effect was not seen in patients receiving preoperative chemotherapy. The specific results are included in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Comparison of the outcomes between the 2 groups after the intervention and at 3-month follow-up.

Outcomes	Intervention group, mean (SD)	Control group, mean (SD)	Cohen *d*	GEE[a] statistical tests	
				Score, (95% CI)[b]	*P* value
EORTC-QLQ-C30[c]					
T0[d]	69.67 (4.26)	69.42 (3.66)	N/A[e]	N/A	N/A
TI[f]	83.41 (2.46)	78.71 (2.72)	1.73	11.51 (10.68 to 12.35)	<.001
T2[g]	86.22 (2.49)	81.82 (2.79)	1.58	14.47 (13.65 to 15.30)	<.001
BSSBQ[h]					
T0	30.33 (1.90)	30.58 (2.01)	N/A	N/A	N/A
TI	41.23 (2.26)	37.28 (2.04)	1.94	8.80 (8.28 to 9.32)	<.001
T2	42.25 (2.58)	36.37 (2.63)	2.23	8.85 (8.25 to 9.42)	<.001
LARS[i] score					
T0	31.07 (3.88)	31.32 (4.73)	N/A	N/A	N/A
TI	26.95 (3.51)	28.87 (4.83)	–0.39	–3.28 (–4.03 to 2.54)	<.001
T2	22.87 (3.09)	26.13 (4.67)	–0.69	–6.69 (–7.45 to 5.93)	<.001
PSSS[j]					
T0	34.42 (1.62)	34.3 (1.48)	N/A	N/A	N/A
TI	36.63 (1.44)	33.05 (1.98)	1.81	0.47 (0.22 to 0.71)	<.001
T2	34.80 (1.19)	34.40 (1.55)	0.25	0.23 (–0.20 to 0.45)	.07

[a]GEE: Generalized estimating equations.
[b]Difference in mean change from baseline to endpoint between the groups.
[c]EORTC-QLQ-C30: European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30.
[d]Baseline.
[e]N/A: Not applicable.
[f]After the intervention.
[g]3-month follow-up.
[h]BSSBQ: Bowel Symptoms Self-Management Behavior Questionnaire.
[i]LARS: Low anterior resection syndrome score.
[j]PSSS: Perceived Social Support Scale.

Table 3. Variation tendency of the EORTC-QLQ-C30[a], BSSBQ[b], LARS[c] score, and PSSS[d] in the 2 groups.

Outcomes	Group effect		Time effect		Group×time	
	F test (*df*)	*P* value	*F* test (*df*)	*P* value	*F* test (*df*)	*P* value
EORTC-QLQ-C30	68.50 (1)	<.001	53.81 (2)	<.001	27.79 (2)	<.001
BSSBQ	48.15 (1)	<.001	74.31 (2)	<.001	3.24 (2)	.03
LARS Score	7.78 (1)	.05	74.94 (2)	<.001	21.34 (2)	<.001
PSSS	29.97 (1)	<.001	14.47 (2)	.001	71.71 (2)	<.001

[a]EORTC-QLQ-C30: European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30.
[b]BSSBQ: Bowel Symptoms Self-Management Behaviors Questionnaire.
[c]LARS: Low anterior resection syndrome.
[d]PSSS: Perceived Social Support Scale.

Secondary Evaluation Indexes

Table 2 shows that the patients' self-management behavior was enhanced for both groups. The BSSBQ total score had significantly larger improvements after intervention (mean difference 8.80; 95% CI 8.28-9.32; Cohen *d*=1.94) and at 3-month follow-up (mean difference 8.85; 95% CI 8.25-9.42; Cohen *d*=2.23) between groups. The BSSBQ total score showed statistically significant time effects (*P*<.001; Table 3).

XSL•FO
RenderX

The LARS score total score had significantly larger decreases after intervention (mean difference –3.28; 95% CI –4.03 to –2.54; Cohen *d*=–0.39) and at 3-month follow-up (mean difference –6.69; 95% CI –7.45 to –5.93; Cohen *d*=–0.69). Table 3 shows that the LARS score total score in both groups exhibited a trend of change over the 6-month period. The intergroup effect exhibits homogeneity (*P*=.05).

The PSSS total score had significantly larger improvements after intervention (mean difference 0.47; 95% CI 0.22-0.71; Cohen *d*=1.81); however, the improvement in this effect did not persist at 3-month follow-up (mean difference 0.23; 95% CI –0.20 to 0.45; *P*=.07; Table 2). Table 3 shows that the PSSS total score in both groups exhibited a trend of change over the 6-month period.

Feasibility

Among the 55 patients who completed the intervention, 45 patients completed 3 MI sessions on time, 7 patients postponed 1 MI session because of scheduling conflicts, and 3 patients only completed 2 MI sessions. The mean number of attended MI sessions was 2.95 (SD 0.23). Additionally, 40 patients completed 84 health diary entries, while the remaining 11 patients did not submit completed entries or fulfill the required entries. The mean number of days of health diary entries was 82.87 (SD 3.15). We invited patients from the intervention group to complete a survey to evaluate their perceptions of the intervention's usability. In the end, 49 people completed the survey. The specific results are included in Appendix 5.

Discussion

Principal Findings

To the best of our knowledge, the “e-bowel safety” applet is the first mobile app developed for patients with LARS in China. This study offers a valuable reference point for future initiatives in mHealth interventions for patients with LARS. A mHealth-based intervention was found to be feasible and effective in helping patients with LARS relieve bowel dysfunction, improve their self-management behavior, and improve their QoL compared to usual care.

This study found that the EORTC-QLQ-C30 total score of the intervention group increased significantly more than that of the control group after the intervention, indicating that the mHealth-based remote interaction could improve the QoL of patients with LARS. These results can be attributed to multiple factors. First, uncontrollable changes in intestinal function, concerns about prognosis, and fear of the future make patients with LARS feel uncertain [37]. A sense of uncertainty influences a patient’s QoL [38]. Patients using the “e-bowel safety” applet can provide timely feedback on their problems to the medical staff and obtain solutions, which can effectively reduce the uncertainty of patients during home rehabilitation. Second, decreased bowel dysfunction severity positively affected the QoL [39]. Third, peer support reportedly enhances cancer adaptation and QoL [40]. The patients’ stories module offers a channel for communication and emotional support among patients with LARS. In this section, patients can share their experiences related to disease management or self-management and receive responses from their peers through comments.

As expected, the BSSBQ total score in the intervention group after the intervention was significantly higher than that in the control group. The findings supported our hypothesis that health-based remote interaction can enhance the self-management behavior of patients with LARS. After the intervention, the results of enhanced self-management behavior were consistent with a previous face-to-face 6-month self-management program study for LARS, which may indicate that mHealth-based remote interaction may yield intervention effects on self-management behavior similar to those observed in face-to-face interventions [41]. However, a more significant effect was observed at 3-month follow-up. This may be because monthly motivational interviews help patients adopt positive health behaviors and improve their self-management awareness [42]. Moreover, current web-based self-management information on LARS is overly intricate for patients, and the information fails to meet the patient’s needs [43]. The strength of our “e-bowel safety” applet is the credibility of the information provided and medical consultation from experts, which can meet the information needs of patients. Finally, our team members created an individualized self-management plan for each participant in the intervention group and reminded them to follow the plans on the applet, which ensured that the patients developed good habits.

Consistent with previous studies [41], this study found that the intervention group demonstrated a more significant decline in the LARS score than the control group. The LARS score also showed significant time effects, indicating that the patient’s bowel dysfunction changed significantly during the 6-month period. This may be because our team members guided patients in rehabilitation exercises and diet adjustments, which have been proven effective in improving bowel dysfunction [44-46]. Meanwhile, the severity of bowel dysfunction decreased over time [13].

Unlike those of previous studies, our findings indicated that mHealth-based remote interaction management intervention could improve the social support levels in the short term; however, sustaining a stable long-term effect on social support was not realized [47]. The patients in the study might have used the “e-bowel safety” applet only for 3 months, and the impact of the intervention on social support may not yield a residual advantage at 3-month follow-up. Furthermore, most patients’ physical and social functions gradually stabilized at 6 months. Our “e-bowel safety” applet focuses on intensive support for symptom management and lacks support knowledge for patients when symptoms plateau, which should be refined in future studies to achieve long-term effects.

In this study, MI was used to stimulate behavioral change and maintenance. The dual intervention of mHealth and MI promotes effective engagement and motivation for health behavior changes. Nearly all the patients (55/60) successfully completed the 3-month intervention and the follow-up during the intervention process, signifying that the mHealth-based remote interaction management intervention is feasible and acceptable. In addition, none of the patients in the intervention group

XSL•FO
RenderX

experienced adverse consequences caused by the intervention, indicating that the intervention was safe.

Limitations

This study has some limitations. First, this study enrolled patients from a tertiary hospital in China, which restricts the generalizability of our results. In the future, we will recruit patients from more hospitals to confirm our research findings. Second, patients were subjected to a limited 3-month follow-up period, thereby restricting our assessment of the enduring effects of the mHealth-based remote interaction management intervention on self-management behavior and QoL. Finally, patients were required to use WeChat and smartphones, which presents the potential for selection bias.

Conclusions

The mHealth-based remote interaction management intervention effectively enhanced the self-management behavior and QoL of patients with LARS, and the impact remained consistent during the 3-month follow-up. Bowel dysfunction also significantly improved throughout the entire research process. This study suggests that mHealth intervention could provide an effective and new option for many patients with LARS. Multicenter studies are necessary to establish the generalizability and effectiveness of these interventions.

Acknowledgments

This work was supported by the 2021 Anhui Higher Education Institutions Provincial Quality Engineering Project (grant 2021jyxm0718) and the Scientific Research and Cultivation project of the School of Nursing, Anhui Medical University (grant hlqm12023055).

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and onLine TeleHealth) checklist (version 1.6.1).
[PDF File (Adobe PDF File), 1193 KB - jmir_v26i1e53909_app1.pdf]

Multimedia Appendix 2
The interview guide of motivational interviewing.
[DOCX File , 13 KB - jmir_v26i1e53909_app2.docx]

Multimedia Appendix 3
Results of pilot experiment.
[DOCX File , 15 KB - jmir_v26i1e53909_app3.docx]

Multimedia Appendix 4
The results of subgroup analysis.
[DOCX File , 12 KB - jmir_v26i1e53909_app4.docx]

Multimedia Appendix 5
Comments and attitudes towards intervention of intervention group.
[DOCX File , 15 KB - jmir_v26i1e53909_app5.docx]

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Cao W, Chen H, Yu Y, Li N, Chen W. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl) 2021;134(7):783-791 [FREE Full text] [doi: 10.1097/CM9.0000000000001474] [Medline: 33734139]
3. Zheng K, Hu Q, Yu G, Zhou L, Yao Y, Zhou Y, et al. Trends of sphincter-preserving surgeries for low lying rectal cancer: a 20-year experience in China. Front Oncol 2022;12:996866 [FREE Full text] [doi: 10.3389/fonc.2022.996866] [Medline: 36568186]
4. Langenfeld SJ, Davis BR, Vogel JD, Davids JS, Temple LKF, Cologne KG, et al. The American society of colon and rectal surgeons clinical practice guidelines for the management of rectal cancer 2023 supplement. Dis Colon Rectum 2024;67(1):18-31. [doi: 10.1097/DCR.0000000000003057] [Medline: 37647138]

XSL•FO
RenderX

5. Pappou EP, Temple LK, Patil S, Smith JJ, Wei IH, Nash GM, et al. Quality of life and function after rectal cancer surgery with and without sphincter preservation. Front Oncol 2022;12:944843. [doi: 10.3389/fonc.2022.944843] [Medline: 36353560]
6. Keane C, Fearnhead NS, Bordeianou LG, Christensen P, Basany EE, Laurberg S, et al. International consensus definition of low anterior resection syndrome. Dis Colon Rectum 2020;63(3):274-284 [FREE Full text] [doi: 10.1097/DCR.0000000000001583] [Medline: 32032141]
7. Pieniowski EHA, Nordenvall C, Palmer G, Johar A, Tumlin Ekelund S, Lagergren P, et al. Prevalence of low anterior resection syndrome and impact on quality of life after rectal cancer surgery: population-based study. BJS Open 2020;4(5):935-942 [FREE Full text] [doi: 10.1002/bjs5.50312] [Medline: 32530135]
8. Sturiale A, Martellucci J, Zurli L, Vaccaro C, Brusciano L, Limongelli P, et al. Long-term functional follow-up after anterior rectal resection for cancer. Int J Colorectal Dis 2017;32(1):83-88. [doi: 10.1007/s00384-016-2659-6] [Medline: 27695976]
9. van der Heijden JAG, Thomas G, Caers F, van Dijk WA, Slooter GD, Maaskant-Braat AJG. What you should know about the low anterior resection syndrome—clinical recommendations from a patient perspective. Eur J Surg Oncol 2018;44(9):1331-1337. [doi: 10.1016/j.ejso.2018.05.010] [Medline: 29807727]
10. Qaderi SM, van der Heijden JAG, Verhoeven RHA, de Wilt JHW, Custers JAE, PLCRC study group. Trajectories of health-related quality of life and psychological distress in patients with colorectal cancer: a population-based study. Eur J Cancer 2021;158:144-155 [FREE Full text] [doi: 10.1016/j.ejca.2021.08.050] [Medline: 34666216]
11. Moon J, Garfinkle R, Zelkowitz P, Dell'Aniello S, Vasilevsky C, Brassard P, et al. Incidence and factors associated with mental health disorders in patients with rectal cancer post-restorative proctectomy. Dis Colon Rectum 2023;66(9):1203-1211. [doi: 10.1097/DCR.0000000000002744] [Medline: 37399122]
12. Bohlok A, Mercier C, Bouazza F, Galdon MG, Moretti L, Donckier V, et al. The burden of low anterior resection syndrome on quality of life in patients with mid or low rectal cancer. Support Care Cancer 2020;28(3):1199-1206. [doi: 10.1007/s00520-019-04901-2] [Medline: 31218414]
13. Pieniowski EHA, Palmer GJ, Juul T, Lagergren P, Johar A, Emmertsen KJ, et al. Low anterior resection syndrome and quality of life after sphincter-sparing rectal cancer surgery: a long-term longitudinal follow-up. Dis Colon Rectum 2019;62(1):14-20. [doi: 10.1097/DCR.0000000000001228] [Medline: 30394987]
14. Eyl RE, Xie K, Koch-Gallenkamp L, Brenner H, Arndt V. Quality of life and physical activity in long-term (≥5 years post-diagnosis) colorectal cancer survivors—systematic review. Health Qual Life Outcomes 2018;16(1):112 [FREE Full text] [doi: 10.1186/s12955-018-0934-7] [Medline: 29859108]
15. Sun V, Wendel CS, Demark-Wahnefried W, Grant M, McMullen CK, Bulkley JE, et al. Diet and behavior modifications by long-term rectal cancer survivors to manage bowel dysfunction-associated symptoms. Nutr Cancer 2019;71(1):89-99 [FREE Full text] [doi: 10.1080/01635581.2018.1524017] [Medline: 30572723]
16. Lai X, Wong FKY, Ching SSY. Review of bowel dysfunction of rectal cancer patients during the first five years after sphincter-preserving surgery: a population in need of nursing attention. Eur J Oncol Nurs 2013;17(5):681-692. [doi: 10.1016/j.ejon.2013.06.001] [Medline: 23871359]
17. Hou X, Pang D, Lu Q, Yang P, Jin S. Bowel dysfunction and self-management for bowel symptoms after sphincter-preserving surgery: a cross-sectional survey of Chinese rectal cancer patients. Cancer Nurs 2017;40(4):E9-E16. [doi: 10.1097/NCC.0000000000000393] [Medline: 27244667]
18. Pape E, Pattyn P, Van Hecke A, Somers N, Van de Putte D, Ceelen W, et al. Impact of low anterior resection syndrome (LARS) on the quality of life and treatment options of LARS—a cross sectional study. Eur J Oncol Nurs 2021;50:101878. [doi: 10.1016/j.ejon.2020.101878] [Medline: 33246248]
19. Frost H, Campbell P, Maxwell M, O'Carroll RE, Dombrowski SU, Williams B, et al. Effectiveness of motivational interviewing on adult behaviour change in health and social care settings: a systematic review of reviews. PLoS One 2018;13(10):e0204890 [FREE Full text] [doi: 10.1371/journal.pone.0204890] [Medline: 30335780]
20. Pape E, Decoene E, Debrauwere M, Van Nieuwenhove Y, Pattyn P, Feryn T, et al. Information and counselling needs of patients with major low anterior resection syndrome: a qualitative study. J Clin Nurs 2023;32(7-8):1240-1250. [doi: 10.1111/jocn.16277] [Medline: 35253296]
21. Bazzell A, Madsen LT, Dains J. Clinical management of bowel dysfunction after low anterior resection for rectal cancer. J Adv Pract Oncol 2016;7(6):618-629 [FREE Full text] [Medline: 29588867]
22. Sakr A, Sauri F, Alessa M, Zakarnah E, Alawfi H, Torky R, et al. Assessment and management of low anterior resection syndrome after sphincter preserving surgery for rectal cancer. Chin Med J (Engl) 2020;133(15):1824-1833 [FREE Full text] [doi: 10.1097/CM9.0000000000000852] [Medline: 32604174]
23. Keutzer L, Wicha SG, Simonsson US. Mobile health apps for improvement of tuberculosis treatment: descriptive review. JMIR Mhealth Uhealth 2020;8(4):e17246 [FREE Full text] [doi: 10.2196/17246] [Medline: 32314977]
24. Saeidzadeh S, Kamalumpundi V, Chi N, Nair R, Gilbertson-White S. Web and mobile-based symptom management interventions for physical symptoms of people with advanced cancer: a systematic review and meta-analysis. Palliat Med 2021;35(6):1020-1038. [doi: 10.1177/02692163211006317] [Medline: 33840271]
25. Liu F, Guo P, Su X, Cui M, Jiang J, Wang S, et al. A novel remote follow-up tool based on an instant messaging/social media app for the management of patients with low anterior resection syndrome: pilot prospective self-control study. JMIR Mhealth Uhealth 2021;9(3):e22647 [FREE Full text] [doi: 10.2196/22647] [Medline: 33739295]

XSL•FO
RenderX

26. Moon J, Monton O, Smith A, Garfinkle R, Zhao K, Zelkowitz P, et al. Interactive online informational and peer support application for patients with low anterior resection syndrome: patient survey and protocol for a multicentre randomized controlled trial. Colorectal Dis 2021;23(5):1248-1257. [doi: 10.1111/codi.15602] [Medline: 33638278]
27. Olivia M, Allister S, Jeongyoon M, Marie D, Richard G, Carol-Ann V, et al. An online educational and supportive care application for rectal cancer survivors with low anterior resection syndrome: a mixed methods pilot study. Colorectal Dis 2023;25(9):1812-1820. [doi: 10.1111/codi.16665] [Medline: 37501348]
28. Montag C, Becker B, Gan C. The multipurpose application WeChat: a review on recent research. Front Psychol 2018;9:2247 [FREE Full text] [doi: 10.3389/fpsyg.2018.02247] [Medline: 30618894]
29. Bao Y, Wang C, Xu H, Lai Y, Yan Y, Ma Y, et al. Effects of an mHealth intervention for pulmonary tuberculosis self-management based on the integrated theory of health behavior change: randomized controlled trial. JMIR Public Health Surveill 2022;8(7):e34277 [FREE Full text] [doi: 10.2196/34277] [Medline: 35834302]
30. Guo Y, Hong YA, Cai W, Li L, Hao Y, Qiao J, et al. Effect of a WeChat-based intervention (Run4Love) on depressive symptoms among people living with HIV in China: a randomized controlled trial. J Med Internet Res 2020;22(2):e16715 [FREE Full text] [doi: 10.2196/16715] [Medline: 32044751]
31. Battersby NJ, Juul T, Christensen P, Janjua AZ, Branagan G, Emmertsen KJ, et al. Predicting the risk of bowel-related quality-of-life impairment after restorative resection for rectal cancer: a multicenter cross-sectional study. Dis Colon Rectum 2016;59(4):270-280. [doi: 10.1097/DCR.0000000000000552] [Medline: 26953985]
32. Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. Evidence-based guidelines for determination of sample size and interpretation of the European organisation for the research and treatment of cancer quality of life questionnaire core 30. J Clin Oncol 2011;29(1):89-96. [doi: 10.1200/JCO.2010.28.0107] [Medline: 21098316]
33. Li H, Zhou P, Pang X, Wang T, Yin D, Fu M, et al. Mobile health-based remote interaction management intervention for patients with low anterior resection syndrome: study protocol for a randomised controlled trial. BMJ Open 2022;12(12):e066046 [FREE Full text] [doi: 10.1136/bmjopen-2022-066046] [Medline: 36564113]
34. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European organization for research and treatment of cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85(5):365-376. [doi: 10.1093/jnci/85.5.365] [Medline: 8433390]
35. Emmertsen KJ, Laurberg S. Low anterior resection syndrome score: development and validation of a symptom-based scoring system for bowel dysfunction after low anterior resection for rectal cancer. Ann Surg 2012;255(5):922-928. [doi: 10.1097/SLA.0b013e31824f1c21] [Medline: 22504191]
36. Zimet GD, Powell SS, Farley GK, Werkman S, Berkoff KA. Psychometric characteristics of the multidimensional scale of perceived social support. J Pers Assess 1990;55(3-4):610-617. [doi: 10.1080/00223891.1990.9674095] [Medline: 2280326]
37. Yin L, Fan L, Tan R, Yang G, Jiang F, Zhang C, et al. Bowel symptoms and self-care strategies of survivors in the process of restoration after low anterior resection of rectal cancer. BMC Surg 2018;18(1):35 [FREE Full text] [doi: 10.1186/s12893-018-0368-5] [Medline: 29866087]
38. Giammanco MD, Gitto L. Coping, uncertainty and health-related quality of life as determinants of anxiety and depression on a sample of hospitalized cardiac patients in Southern Italy. Qual Life Res 2016;25(11):2941-2956. [doi: 10.1007/s11136-016-1323-5] [Medline: 27245776]
39. Emmertsen K, Laurberg S, Rectal Cancer Function Study Group. Impact of bowel dysfunction on quality of life after sphincter-preserving resection for rectal cancer. Br J Surg 2013;100(10):1377-1387. [doi: 10.1002/bjs.9223] [Medline: 23939851]
40. Brodar KE, Carlisle V, Tang PY, Fisher EB. Identification and characterization of peer support for cancer prevention and care: a practice review. J Cancer Educ 2022;37(3):645-654 [FREE Full text] [doi: 10.1007/s13187-020-01861-8] [Medline: 32892278]
41. Li C, Li Z, Lu Q, Zhou Y, Qin X, Wu A, et al. The effectiveness of a self-management program of bowel dysfunction in patients with mid and low rectal cancer after sphincter-preserving surgery: a pilot randomized controlled trial. Cancer Nurs 2023;46(1):67-76. [doi: 10.1097/NCC.0000000000001065] [Medline: 35089874]
42. Seven M, Reid A, Abban S, Madziar C, Faro JM. Motivational interviewing interventions aiming to improve health behaviors among cancer survivors: a systematic scoping review. J Cancer Surviv 2023;17(3):795-804 [FREE Full text] [doi: 10.1007/s11764-022-01253-5] [Medline: 36100801]
43. Lightner A. Assessing the readability, quality and accuracy of online health information for patients with low anterior resection syndrome following surgery for rectal cancer, by Garfinkle et al. Colorectal Dis 2019;21(5):500. [doi: 10.1111/codi.14594]
44. Yuanyuan W, Shiyin H, Lei H, Ding D. Pelvic floor muscle exercises alleviate symptoms and improve mental health and rectal function in patients with low anterior resection syndrome. Front Oncol 2023;13:1168807 [FREE Full text] [doi: 10.3389/fonc.2023.1168807] [Medline: 37152027]
45. Sun V, Crane TE, Slack SD, Yung A, Wright S, Sentovich S, et al. Rationale, development, and design of the altering intake, managing symptoms (AIMS) dietary intervention for bowel dysfunction in rectal cancer survivors. Contemp Clin Trials 2018;68:61-66 [FREE Full text] [doi: 10.1016/j.cct.2018.03.010] [Medline: 29567283]

XSL•FO
RenderX

46. Sun V, Crane TE, Freylersythe S, Slack SD, Yung A, Krouse RS, et al. Altering intake and managing symptoms: feasibility of a diet modification intervention for post-treatment bowel dysfunction in rectal cancer. Clin J Oncol Nurs 2022;26(3):283-292. [doi: 10.1188/22.CJON.283-292] [Medline: 35604741]
47. Zhu J, Ebert L, Liu X, Wei D, Chan SW. Mobile breast cancer e-Support program for Chinese women with breast cancer undergoing chemotherapy (part 2): multicenter randomized controlled trial. JMIR Mhealth Uhealth 2018;6(4):e104 [FREE Full text] [doi: 10.2196/mhealth.9438] [Medline: 29712622]

Abbreviations

BSSBQ: Bowel Symptoms Self-Management Behaviors Questionnaire
CONSORT: Consolidated Standards of Reporting Trials
EORTC-QLQ-C30: European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30
LARS: low anterior resection syndrome
mHealth: mobile health
MI: motivational interviewing
PSSS: Perceived Social Support Scale
QoL: quality of life
RCT: randomized controlled trial
SPS: sphincter-preserving surgery

Edited by A Mavragani; submitted 24.10.23; peer-reviewed by V Sun, C Thomson; comments to author 13.03.24; revised version received 07.05.24; accepted 03.06.24; published 13.08.24.

Please cite as:
Zhou P, Li H, Pang X, Wang T, Wang Y, He H, Zhuang D, Zhu F, Zhu R, Hu S
Effect of a Mobile Health–Based Remote Interaction Management Intervention on the Quality of Life and Self-Management Behavior of Patients With Low Anterior Resection Syndrome: Randomized Controlled Trial
J Med Internet Res 2024;26:e53909
URL: https://www.jmir.org/2024/1/e53909
doi:10.2196/53909
PMID:

©Peng Zhou, Hui Li, Xueying Pang, Ting Wang, Yan Wang, Hongye He, Dongmei Zhuang, Furong Zhu, Rui Zhu, Shaohua Hu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association of Wearable Device–Measured Step Volume and Variability With Blood Pressure in Older Chinese Adults: Mobile-Based Longitudinal Observational Study

Han Xiao[1,2], MD; Zechen Zhou[1,2], PhD; Yujia Ma[1,2], PhD; Xiaoyi Li[1,2], PhD; Kexin Ding[1,2], PhD; Xiaotong Dai[3,4], Prof Dr; Dafang Chen[1,2], MD, PhD

[1]Department of Epidemiology and Biostatistics, Peking University, Beijing, China

[2]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China

[3]School of Sport Science, Beijing Sport University, Beijing, China

[4]Key Laboratory of Ministry of Education for Sports and Physical Health, Beijing Sport University, Beijng, China

Corresponding Author:
Dafang Chen, MD, PhD
Department of Epidemiology and Biostatistics
Peking University
38 Xueyuan Road
Haidian
Beijing, 100191
China
Phone: 86 13701248872
Email: hx624@ic.ac.uk

Abstract

Background: The paucity of evidence on longitudinal and consecutive recordings of physical activity (PA) and blood pressure (BP) under real-life conditions and their relationships is a vital research gap that needs to be addressed.

Objective: This study aims to (1) investigate the short-term relationship between device-measured step volume and BP; (2) explore the joint effects of step volume and variability on BP; and (3) examine whether the association patterns between PA and BP varied across sex, hypertension status, and chronic condition status.

Methods: This study used PA data of a prospective cohort of 3070 community-dwelling older adults derived from a mobile health app. Daily step counts, as a proxy of step volume, were derived from wearable devices between 2018 and 2022 and categorized into tertiles (low, medium, and high). Step variability was assessed using the SD of daily step counts. Consecutive daily step count recordings within 0 to 6 days preceding each BP measurement were analyzed. Generalized estimation equation models were used to estimate the individual and joint associations of daily step volume and variability with BP. Stratified analyses by sex, the presence of hypertension, and the number of morbidities were further conducted.

Results: A total of 3070 participants, with a median age of 72 (IQR 67-77) years and 71.37% (2191/3070) women, were included. Participants walked a median of 7580 (IQR 4972-10,653) steps and 5523 (IQR 3590-7820) meters per day for a total of 592,597 person-days of PA monitoring. Our results showed that higher levels of daily step volume were associated with lower BP (systolic BP, diastolic BP, mean arterial pressure, and pulse pressure). Compared with participants with low step volume (daily step counts <6000/d) and irregular steps, participants with high step volume (≥9500/d) and regular steps showed the strongest decrease in systolic BP (–1.69 mm Hg, 95% CI –2.2 to –1.18), while participants with medium step volume (6000/d to <9500/d) and regular steps were associated with the lowest diastolic BP (–1.067 mm Hg, 95% CI –1.379 to –0.755). Subgroup analyses indicated generally greater effects on women, individuals with normal BP, and those with only 1 chronic disease, but the effect pattern was varied and heterogeneous between participants with different characteristics.

Conclusions: Increased step volume demonstrated a substantial protective effect on BP among older adults with chronic conditions. Furthermore, the beneficial association between step volume and BP was enhanced by regular steps, suggesting potential synergistic protective effects of both increased step volume and step regularity. Targeting both step volume and variability through PA interventions may yield greater benefits in BP control, particularly among participants with hypertension and a higher chronic disease burden.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50075) doi:10.2196/50075

KEYWORDS

older adults; physical activity; step volume; step variability; blood pressure; wearable devices; mHealth apps; mobile health apps; mobile phone

Introduction

Background

Elevated blood pressure (BP) is a significant risk factor for cardiovascular disease and mortality and is highly prevalent in older adults [1,2]. The continuously increasing number of older individuals further leads to a growing population with high BP, which poses a significant need for both home and institutionalized BP management [1].

Evidence to date has recognized physical activity (PA) as a primary strategy for controlling BP [3,4]. For older adults, walking is the most common and feasible way of PA, which requires only a minimal level of fitness and can be done almost anywhere [5,6]. Previous epidemiological evidence has consistently shown that taking higher steps per day is associated with BP reduction and lower cardiovascular disease risk [2,3,7], implicating the potential of daily steps as a target for personalized exercise prescriptions in older adults. In accordance, the 2018 Physical Activity Guidelines Advisory Committee emphasized the importance of daily walking for its ability to improve the translation of research findings into public health recommendations [8].

However, previous studies highlighting the beneficial effects of walking are limited by their single assessments over short durations at baseline (eg, 7 days), which fail to capture consecutive daily step counts over time and may not be representative of habitual walking behavior in daily life. For example, most previous cohort studies, such as Toledo Study for Healthy Aging [9] and UK Biobank [10], measured participants' daily step levels only within 1 week at baseline and then examined their associations with outcomes during long-term follow-up periods. Recognizing this limitation, a few recent observational studies have been conducted to monitor step counts over extended periods, typically spanning 6 months to 2 years, thereby accounting for longitudinal changes in step counts over time [7,11,12]. While these studies have yielded novel insights into the associations between step counts and cardiovascular risks, little is known about the temporal relationships between daily steps and subsequent BP. This gap in knowledge is partly attributable to the absence of consecutive daily assessments of cardiovascular outcomes in current study designs. Furthermore, despite regular PA being commonly recommended by guidelines [13], previous studies have yet to quantitatively assess the variability of PA or the joint effects of PA volume and variability on BP. Indeed, structured interventions implemented by randomized controlled trials often result in homogeneity in the regularity of PA, making it challenging to detect the effects of activity variability.

Monitoring individual PA and BP consecutively in real-world scenarios and over a long period is currently a knowledge gap. Digital health technologies, including consumer-based wearable devices and mobile health (mHealth) apps, present a valuable opportunity to bridge this gap through remote data recording, uploading, and management [14]. mHealth apps also have the potential to provide tailored health information, multidimensional health data aggregation, self-monitoring, and feedback service to the users, all of which would enhance participant adherence to longitudinal studies.

Objectives

In this study with a focus on older adults, we analyzed a total of 592,597 person-days of PA records and 188,671 BP measurements collected from an mHealth app in China across a 5-year period. We aimed to (1) investigate the short-term relationship between device-measured step volume and BP; (2) explore the joint effects of step volume and variability on BP; and (3) examine whether the association patterns between PA and BP varied across sex, hypertension status, and chronic condition status. We hypothesized that older adults with different characteristics would exhibit varying responses to daily steps, and our findings have the potential to inform personalized remote health management strategies for aging populations.

Methods

Study Population

This analysis is based on a population-based cohort of community-dwelling older adults in China. Participants were recruited and managed via an mHealth app (YIDO Health app, YIDO Artificial Intelligence Technology Co, Ltd) in combination with consumer wearable devices. The YIDO mHealth app was commercially designed to provide personalized health information monitoring, health assessment, and health guidance, especially for older adults [15]. Upon system log in, users could connect wearable activity trackers to the smartphone and securely upload monitoring data to the web-based app. The analytic data were collected in an anonymous and aggregated data set without personal identifiers such as names, email addresses, or cell phone numbers. Random strings were used to identify the PA records for each day.

Retrospective analyses were conducted using 592,597 days of PA data and 188,671 BP recordings of 3070 participants from the mHealth app between 2018 and 2022. Eligible participants for the study were community-dwelling older adults who were users of the YIDO remote management system. Inclusion criteria included age ≥60 years, fluency in Chinese, ownership of a smartphone with a supported iOS or Android version, possession of disease history information, regular access to the app for uploading daily activity data throughout the study period, and informed consent before the study. Participants were excluded due to the following reasons: (1) their wrist or bicep circumference precluded comfortable use of the watch or BP cuff; (2) they had a clinical contraindication to noninvasive BP measurements; (3) incomplete step count or BP data; (4) step

XSL•FO
RenderX

count or BP outliers; and (5) not having at least 7 consecutive days of step count recordings followed by at least 1 valid measurement of BP.

Measurement

Step Volume and Variability Assessment

The participants were provided with a Lexin Mio smart bracelet (Mambo, Lexin). A Lexin Mio smart bracelet is a 3-axis acceleration sensor-based wearable device that tracks and monitors the user's steps, walking distance, heart rate, and other biometric data. Its validity has been reported previously [16]. All participants were given specific instructions to wear their watches during the whole day (24 hours) and for as many days as possible. Evidence suggests that 10 hours of wear time is sufficient to estimate daily PA during waking time [17]. To ensure the completeness of tracking, we, therefore, defined a valid day as one when the device is worn for at least 16 hours and at least 250 steps are reported. The watch wear time accounted only for the minutes reporting a valid heart rate measurement.

The most frequently used and relevant metric to reflect one's physical behaviors was the number of steps taken per day, hereon referred to as daily step count [18]. In this analysis, we computed (1) daily step counts and (2) SD of daily steps to assess PA characteristics. Daily step counts are considered as a reliable proxy of the overall volume of PA performed by a participant [14]. The SD of daily steps is considered here as an indicator of the variability of PA and as a proxy of regularity [19]. A low SD value indicates low variability and, therefore, a high regularity of the PA level. Daily step volume were categorized as low, medium, and high based on the tertile distributions of both the whole sample or specific subpopulations in terms of sex, hypertension status, or chronic morbidities status (Multimedia Appendix 1). The variability of daily steps was categorized based on the median value of the SD of daily steps (2000 steps/d): regular (SD of daily step counts<2000/d) and irregular (SD of daily step counts≥2000/d).

BP Monitoring

BP monitors (C02L, Heal Force) were provided to monitor BP during the study. To quantify BP, the user was instructed to (1) rest for at least 5 minutes before taking a measurement; (2) be seated in a comfortable position with legs uncrossed, feet flat on the floor, and back or arms supported; (3) not speak during the measurements; and (4) perform the measurement on their left arm. The resulting BP is uploaded to the user's app account. Participants could record BP at any time of day convenient to them. A BP measurement was excluded if the systolic BP (SBP) was <50 mm Hg or >300 mm Hg or if the diastolic BP (DBP) was <30 mm Hg or >140 mm Hg. Then, mean arterial pressure (MAP) was calculated using the formula MAP = (SBP + [2 * DBP])/3. Pulse pressure (PP) was calculated as the difference between SBP and DBP.

Covariates

Covariates were selected based on the literature and the biological plausibility for confounding the main associations of interest. Covariates obtained at baseline included sex, age, smoking status (never, ever, or current), and drinking status (never, ever, or current). BMI was calculated as weight (kg) divided by height squared (m^2). The presence of hypertension and type 2 diabetes was ascertained by the self-report of physician diagnosis and the use of antihypertensive and antidiabetic medications. The number of morbidities, including asthma, arthritis, cancer, chronic obstructive pulmonary disease, chronic kidney disease, coronary heart disease, diabetes, hyperlipidemia, hypertension, osteoporosis, peptic ulcer, and stroke, was used to account for the comorbidity status of participants in the study [20]. The selection of conditions was based on the Centers for Disease Control and Prevention recommendations for defining and measuring multiple chronic conditions [21]. Tracking seasons were divided into quarters as spring (March to May), summer (June to August), autumn (September to November), and winter (December to February) in this study because of the large latitude span in China. In addition, the date of recording was used to define weekdays and weekends and to adjust for the long-term time trend.

Statistical Analysis

Descriptive Analysis

Baseline characteristics of the study population were reported as median (IQR) for continuous variables and as numbers (percentages) for categorical variables.

Main Analyses on the Association of Step Volume and Variability With BP

Personal PA recordings were linked to the participants' BP measurements according to the monitoring date. To fully explore the short-term lag structure, step count recordings on the same day as BP measurement were designated as lag 0, while those from the previous day were denoted as lag 1. This approach enabled the examination of the effects of walking steps with varying lag days from lag 0 to lag 6. Cumulative PA effects were further estimated using the 7-day moving average (lag 0 to lag 6) of step counts. Therefore, only participants with at least 7 consecutive valid daily walking records preceding each BP measurement were included in the analysis (Figure 1A).

Generalized estimating equation models were applied using daily step counts categorized into tertiles (low, medium, and high). The advantages of generalized estimating equation models are that they consider the within-individual correlation in exposure and outcome due to repeated measures. They can also better handle unbalanced data, as the number of repeated measures per participant can be different from one participant to another [18]. For all models, we adjusted for the following covariates: age, sex, BMI, smoking status (never, ever, or current), drinking status (never, ever, or current), history of hypertension, history of type 2 diabetes, season, and type of record day (weekday or weekend). We also adjusted the time trend by including a natural spline of calendar dates with 5 df. This model choice was governed by the quasi-likelihood under the independence model criterion values of models.

Participants were further categorized into 6 mutually exclusive groups to investigate the joint associations of weekly average step volume (low, medium, or high) and step variability (regular

XSL•FO
RenderX

or irregular) with BP, with the group with low step volume and irregular steps as a reference. Proportions for each combined group are presented in Multimedia Appendix 2. The same models described earlier were used for data analysis.

We further conducted a series of stratified analyses by sex, the presence of hypertension, and the number of morbidities (with 1 morbidity, with 2 morbidities, and with at least 3 morbidities). When performing the stratified analyses, daily step volume and walking distance were categorized in tertiles based on the cut-offs of specific subpopulations (Multimedia Appendix 1). The joint associations of step volume and step variability across various subgroups were also examined using stratified analyses.

Figure 1. Study design for investigating the association of step volume and variability with blood pressure (BP). (A) Study design for main analyses. Physical activity (PA) recordings on the same day of BP measurement were considered as 0 days preceding the BP measurement and denoted as lag 0. In this manner, consecutive recordings of 0 to 6 days before every BP measurement were retrieved. The cumulative effects of PA were examined using the moving average value of step counts in lag 0 to 6 days. Step variability was computed as the SD of step counts in lag 0 to 6 days. (B) Study design for sensitivity analyses. Week-to-week associations were examined to reduce potential bias caused by large amounts of repeated measures and to support the credibility of our results. We focused on the weekly average PA 1 week before or on the same week of BP measurements. 1 valid week was defined as consisting of at least 3 days of BP measurement and at least 5 days of step count recordings.

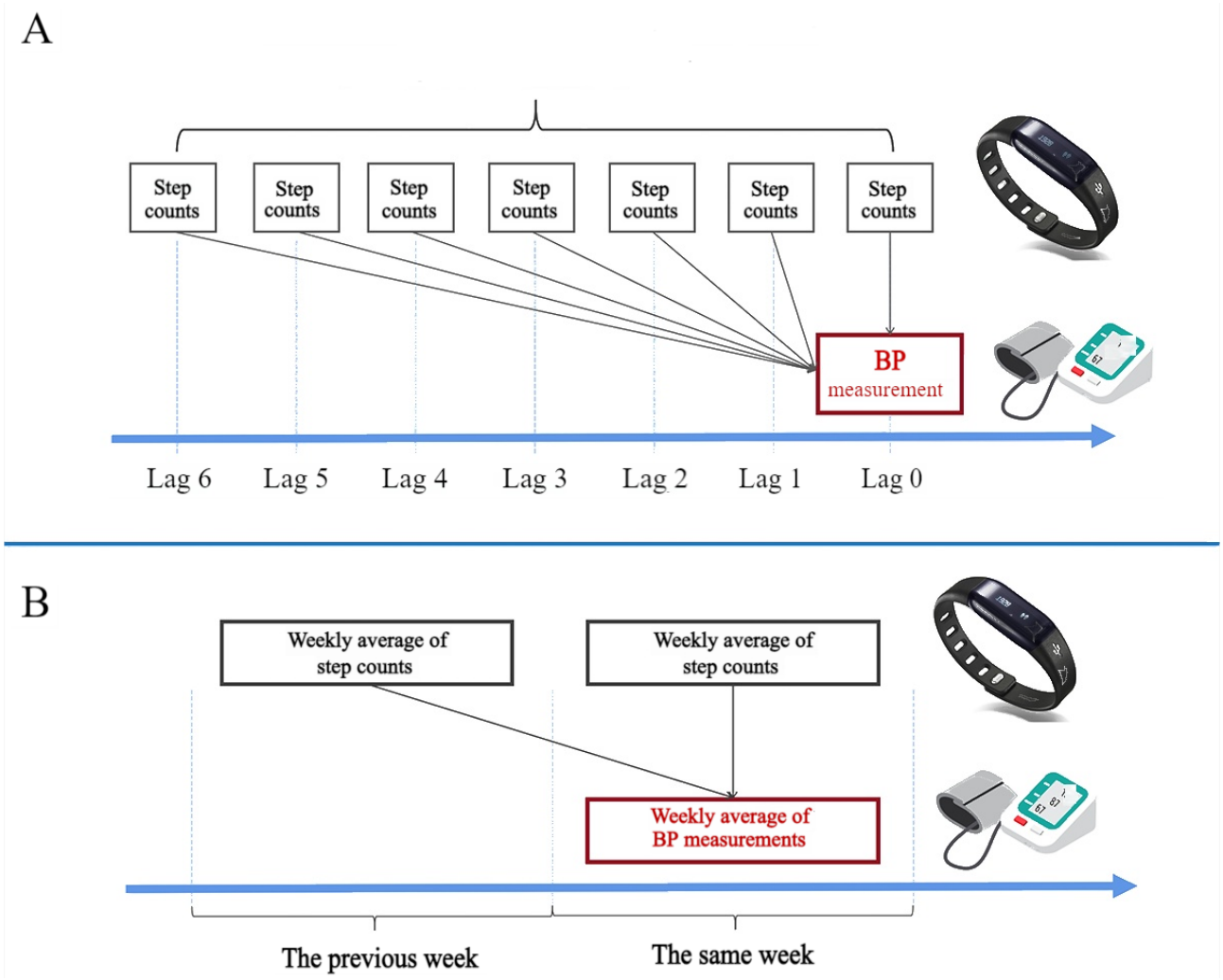

Sensitivity Analysis

A series of sensitivity analyses were performed. First, to prove the generalizability of our results, stratified analyses using categorized daily PA determined by tertiles of the whole population rather than based on subgroups were conducted. Second, the literature on using pedometers to measure PA suggests interpreting raw step counts rather than converting them to distance, although the basis for this recommendation is unclear [6]. As daily walking distance was also recorded by the wearable devices, we repeated the main analyses using walking distance as an indicator of daily PA. Third, given the large amount of repeated BP measures on the time scale of a day, we further designed week-to-week association analyses to reduce potential bias caused by repeated measures and support the credibility of our results. In this analysis, a valid week was defined as consisting of at least 3 days of BP measurements and at least 5 days of step counts and walking distance records. In addition, we considered the weekly average PA in the same week and in the previous week of BP measurement as exposure variables. Therefore, participants with 1 valid week of BP records within 2 consecutive valid weeks of PA records were considered for the analysis (Figure 1B). The analytic sample for this step was derived from the sample for the main analyses (Multimedia Appendix 3).

XSL•FO
RenderX

Statistical analyses were conducted using R (version 4.2.0; R Foundation for Statistical Computing). Significance levels were set at $P<.05$.

Ethical Considerations

Participants provided electronic consent for participation upon creating their accounts and accepting the anonymous treatment of personal data by YIDO for research purposes. The extracted data were anonymized and managed according to privacy and data protection regulations. This study was conducted in compliance with the Declaration of Helsinki principles and approved by the Sports Science Experiment Ethics Committee of Beijing Sport University (number 2024012H). All patients gave written informed consent to participate in this study. They did not receive any financial or other compensation.

Results

Characteristics of the Study Population

The participants (N=3070) had a median age of 72 (IQR 67-77) years and a median BMI of 23.9 (IQR 22.2-25.8) kg/m^2 (Table 1). Overall, 71.37% (2191/3070), 45.28% (1390/3070), and 57.26% (1758/3070) of the participants were female, had hypertension, and had at least 2 morbidities, respectively. This study comprised 188,671 accumulated person-days of BP measurements and 592,597 person-days of PA records over 1351 days of tracking. Participants monitored BP for a median of 34 (IQR 10-75) days. From the overall days studied, the median values of SBP, DBP, MAP, and PP were 125 (IQR 117-134) mm Hg, 71 (IQR 65-78) mm Hg, 89.3 (IQR 83-95) mm Hg, and 54 (IQR 46-62) mm Hg, respectively. Wearable devices for detecting PA were worn for a median of 161 (IQR 55-300) days, with participants taking a median of 7580 (IQR 4972-10,653) steps per day.

XSL•FO
RenderX

Table 1. Characteristics of the study population.

Characteristics	Overall (N=3070)	Female participants (n=2191)	Male participants (n=879)
Age (y), median (IQR)	72 (67-77)	71 (67-76)	74 (69-80)
BMI (kg/m^2), median (IQR)	23.90 (22.23-25.78)	23.83 (22.19-25.78)	24.12 (22.31-25.76)
Smoking status, n (%)			
Never	2622 (85.41)	2112 (96.39)	510 (58.02)
Ever	303 (9.87)	66 (3.01)	237 (26.96)
Current	145 (4.72)	13 (0.59)	132 (15.02)
Drinking status, n (%)			
Never	2425 (79)	1995 (91.05)	430 (48.92)
Ever	349 (11.37)	129 (5.89)	220 (25.03)
Current	296 (9.64)	67 (3.06)	229 (26.05)
Hypertension, n (%)	1390 (45.28)	965 (44.04)	425 (48.35)
Number of morbidities, n (%)			
1	1307 (42.57)	904 (41.26)	403 (45.85)
2	803 (26.16)	567 (25.88)	241 (27.42)
≥3	955 (31.11)	720 (32.86)	235 (26.73)
BP[a] measurements			
Systolic BP (mm Hg), median (IQR)	125 (117-134)	125.00 (116.50-133.00)	126.00 (118.00-135.00)
Diastolic BP (mm Hg), median (IQR)	71 (65-78)	70.50 (64.00-77.00)	72.00 (66.00-79.00)
Mean arterial pressure (mm Hg), median (IQR)	89.33 (83-95)	89.00 (82.67-94.67)	90.33 (84.33-96.00)
Pulse pressure (mm Hg), median (IQR)	54 (46-62)	54.00 (46.00-62.00)	53.00 (45.00-61.00)
Accumulative BP measurements (days), n	188,671	132,464	56,207
Measurements per person (days), median (IQR)	34 (10-75)	33 (10-74)	35 (12-76)
Physical activity records			
Daily step counts (steps), median (IQR)	7580 (4972-10,653)	7545 (4975-10,546)	7672 (4964-10,923)
Daily step volume groups, n (%)			
Low	956 (31.14)	660 (30.12)	296 (33.67)
Medium	1297 (42.25)	952 (43.45)	345 (39.25)
High	817 (26.61)	579 (26.43)	238 (27.08)
SD of daily step counts (steps), median (IQR)	2095 (1475-2854)	2109 (1499-2858)	2061 (1418-2843)
Daily walking distance (meters), median (IQR)	5523.76 (3590-7820)	5480 (3573.85-7714.56)	5640 (3630-8100)
Daily walking distance groups, n (%)			
Low	1002 (32.64)	703 (32.09)	299 (34.02)
Medium	1278 (41.63)	936 (42.72)	342 (38.91)
High	790 (25.73)	552 (25.19)	238 (27.08)
SD of daily walking distance (meters), median (IQR)	1564 (1093-2148)	1571 (1110-2149)	1545 (1058-2147)
Total days of tracking (days), n	1351	1279	1301
Accumulative activity records (days), n	592,597	417,886	174,711
Records per person (days), median (IQR)	161 (55-300)	156 (51-293)	167 (63-309)
Type of the daily records tracked in, n (%)			
Weekdays	423,119 (71.4)	298,280 (71.4)	124,839 (71.5)
Weekends	169,478 (28.6)	119,606 (28.6)	49,872 (28.5)

XSL•FO
RenderX

Characteristics	Overall (N=3070)	Female participants (n=2191)	Male participants (n=879)
Seasons of records tracked in, n (%)			
Spring (March to May)	174,169 (29.4)	122,208 (29.2)	51,961 (29.7)
Summer (June to August)	196,272 (33.1)	139,230 (33.3)	57,042 (32.6)
Autumn (September to November)	98,215 (16.6)	68,834 (16.5)	29,381 (16.8)
Winter (December to February)	123,941 (20.9)	87,614 (21.0)	36,327 (20.8)

[a]BP: blood pressure.

Association of Step Volume With BP

Compared with participants with low daily step volume (<6000 steps/d), those with higher daily step volume exhibited lower BP levels. Notably, distinctive association patterns emerged across various BP metrics. For SBP, higher daily step volume were associated with significantly lower BP values. Conversely, for DBP, the most pronounced association with step volume was observed within the medium tier of step volume, while the association attenuated for the highest step volume level (Figure 2; Multimedia Appendix 4).

For the lag structure, the step volume on the previous day (lag 1) of BP measurement showed the greatest association with BP. This association substantially attenuated at lag s of 2 to 5 days before modestly increasing again for PA effects a week prior (lag 6). We further examined the cumulative effects of PA on BP by using the moving average step volume of lag 0 to 6 days, and we found that the cumulative effects were stronger than the effect produced by a single-day exposure. Compared with a low weekly average step volume, the medium and high levels of weekly average step volume had 0.524 (95% CI –0.822 to –0.226) mm Hg and 0.913 (95% CI –1.27 to –0.556) mm Hg decreases in SBP and 0.5 (95% CI –0.721 to –0.279) mm Hg and 0.182 (95% CI –0.452 to –0.088) mm Hg decreases in DBP, respectively (Figure 2; Multimedia Appendix 4).

Figure 2. Estimated changes (95% CIs) in blood pressure (BP; systolic BP [SBP], diastolic BP [DBP], mean arterial pressure [MAP], and pulse pressure [PP]) related to daily step volume in different lag days and week averages. Step volume levels were categorized based on the tertiles of the whole sample: low (daily step counts <6000/d), medium (6000/day to <9500/d), and high (≥9500/d). Models are adjusted for age, sex, BMI, smoking status, drinking status, the presence of hypertension, the presence of type 2 diabetes, season, the type of day (weekday or weekend), and long-term trend.

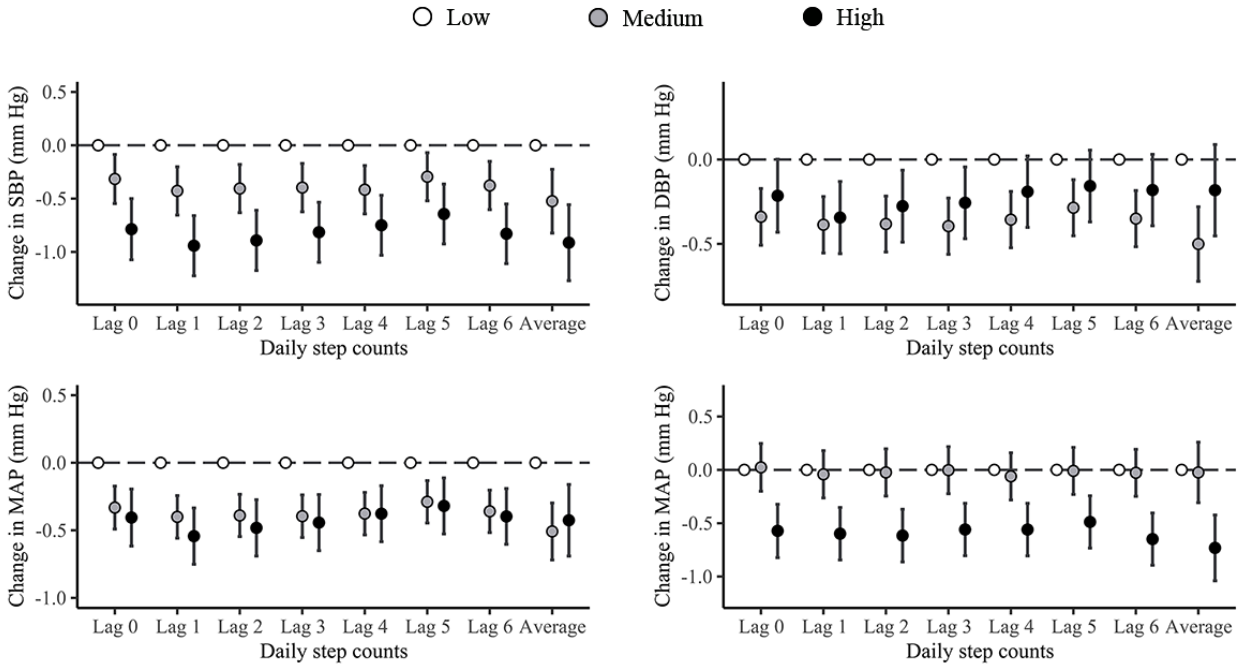

Joint Association of Step Volume and Variability With BP

Overview

The joint association estimates for each condition of step volume and variability combinations compared with the referent group with low step volume and irregular steps are illustrated in Figure 3 and Multimedia Appendix 5. Generally, participants with a more regular step pattern had a stronger effect on BP than participants with an irregular step pattern, irrespective of their step volume level. Furthermore, among participants with regular steps in a week, those who had this in combination with a high level of step volume showed the strongest association with SBP, MAP, and PP, whereas the effects on DBP were the most pronounced for medium level of daily step volume. Specifically, participants with high step volume and regular steps showed the strongest decrease in SBP (–1.69 mm Hg, 95% CI –2.2 to –1.18), MAP (–1.019 mm Hg, 95% CI –1.407 to –0.631), and

XSL•FO
RenderX

PP (–1.006 mm Hg, 95% CI –1.455 to –0.557), while participants with medium step volume and regular steps were associated with the lowest DBP (–1.067 mm Hg, 95% CI –1.379 to –0.755).

Figure 3. Estimated changes (95% CIs) in blood pressure (BP; systolic BP [SBP], diastolic BP [DBP], mean arterial pressure [MAP], and pulse pressure [PP]) related to joint effects of daily step volume and variability. Step volume levels were categorized based on the tertiles of the whole sample: low (daily step counts <6000/d), medium (6000/d to <9500/d), and high (≥9500/d). Step variability was categorized into regular (SD of daily step counts<2000/d) and irregular (≥2000/d). Models were adjusted for age, sex, BMI, smoking status, drinking status, the presence of hypertension, the presence of type 2 diabetes, season, the type of day (weekday or weekend), and long-term trend.

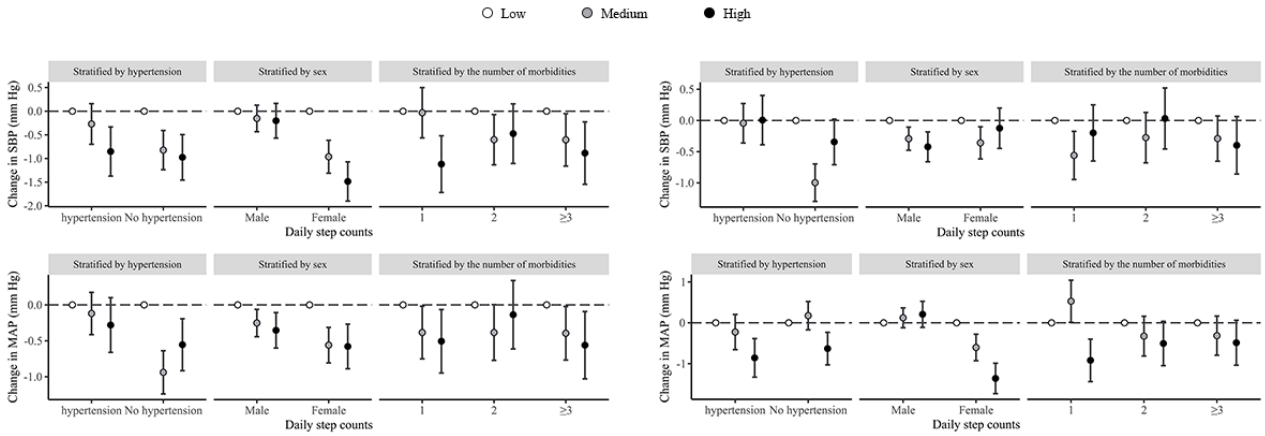

Subgroup Analyses of the Association of Step Volume and Variability With BP

In stratified analyses, step volume and variability still had substantial associations with BP in different subgroups, which were generally consistent with the results among the overall population. Despite the general beneficial relationship between PA and BP, modification effects between subgroups were also observed (Figures 4 and 5; Multimedia Appendices 6 and 7). When classified by sex, the associations of step volume and variability with BP were more apparent in female participants. Higher step volume was associated with lower DBP and MAP for both sexes, whereas it was linked to lower SBP and PP for only women.

In terms of the presence of hypertension, the effects of step volume were more pronounced in participants without hypertension. Among patients with hypertension, only those in the high tertile of step volume demonstrated significantly lower SBP and PP (for SBP: –0.852 mm Hg, 95% CI –1.371 to –0.333; for PP: –0.871 mm Hg, 95% CI –1.345 to –0.397).

Considering the number of chronic morbidities, we observed that the effects of PA on BP were more remarkable in participants with only 1 chronic disease. For example, a high step volume decreased SBP by 1.1 mm Hg (95% CI –1.696 to –0.504) in participants with 1 morbidity, while the decrease was by –0.885 mm Hg (95% CI –1.546 to –0.224) for participants with at least 3 morbidities.

Finally, joint effects of the combination of step volume and variability were further investigated among subgroups (Figure 5). Similar to the overall population, we found that the beneficial associations between step volume and BP were exacerbated by regular steps, suggesting likely synergistic protective effects of step volume and variability. Interestingly, although step volume solely seemed not to be associated with DBP and MAP among patients with hypertension, participants combination of regular steps and high step volume had significantly lower DBP (–0.789 mm Hg, 95% CI –1.417 to –0.179) and MAP (–1.190 mm Hg, 95% CI –1.782 to –0.598) compared to participants with irregular steps and low step volume. These results indicated that step volume and variability may jointly affect BP. In addition, the associations between steps and BP would vary among older adults with different demographic characteristics or disease statuses.

XSL•FO
RenderX

Figure 4. Estimated changes (95% CIs) in blood pressure (BP; systolic BP [SBP], diastolic BP [DBP], mean arterial pressure [MAP], and pulse pressure [PP]) related to the weekly average daily step volume level across sexes, the presence of hypertension, and chronic disease status. Daily step volume were categorized as low, medium, and high based on the tertile distributions of specific subpopulations. Models were adjusted for age, sex, BMI, smoking status, drinking status, the presence of hypertension, the presence of type 2 diabetes, season, the type of day (weekday or weekend), and long-term trend.

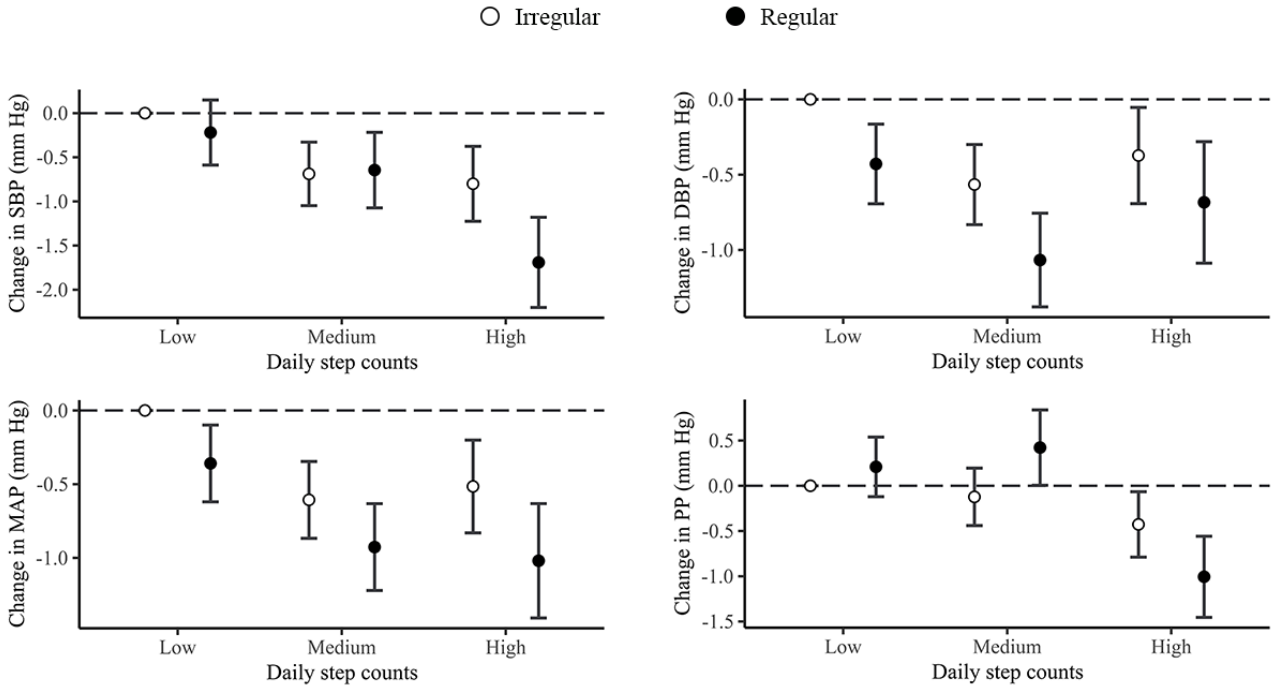

Figure 5. Joint associational patterns for step volume and variability with blood pressure (BP; systolic BP [SBP], diastolic BP [DBP], mean arterial pressure [MAP], and pulse pressure [PP]) across sexes, the presence of hypertension, and chronic disease status. Heatmap illustrated magnitudes in distinct patterns for the effects of physical activity (volume and variability) on BP across different subgroups. Daily step volume were categorized as low, medium, and high based on the tertiles of specific subpopulations. The SD of step volume over a week was categorized into regular (SD of daily step counts<2000/d) and irregular (≥2000/d). Models were adjusted for age, sex, BMI, smoking status, drinking status, the presence of hypertension, the presence of type 2 diabetes, season, the type of day (weekday or weekend), and long-term trend. **P*<.05, ***P*<.01, ****P*<001.

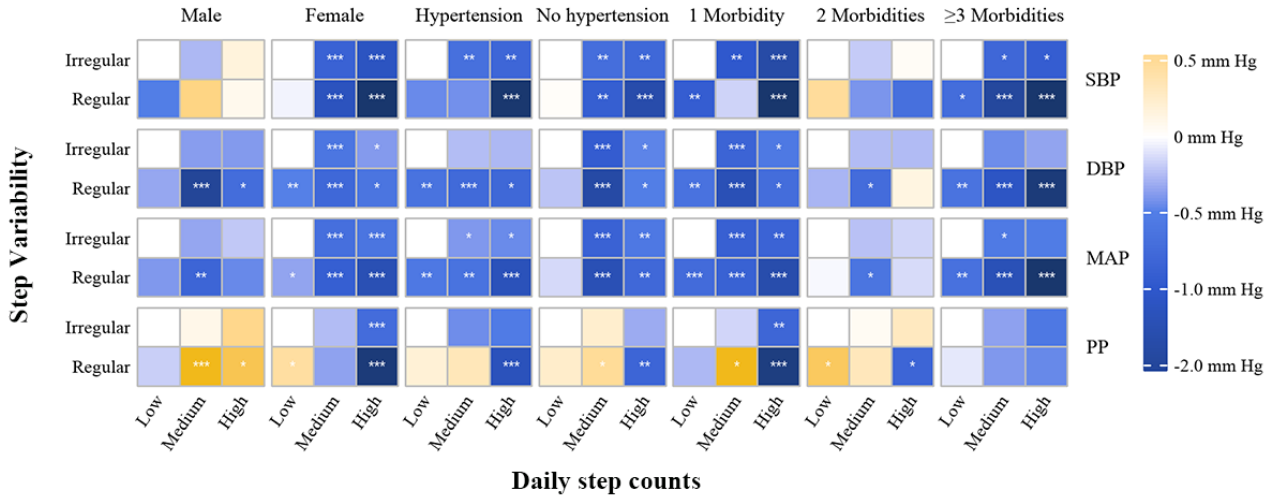

Sensitivity Analyses

We conducted several sensitivity analyses, as described earlier, to ensure the credibility of our results. First, we used daily walking distance as another indicator of PA and repeated the main analyses (Figures S1 and S2 in Multimedia Appendix 8; Multimedia Appendices 4-7). The results were generally consistent with the results using daily step volume. It is worth noting that, for some subpopulations, the significance increased when using walking distance as an independent variable (Figure S3 in Multimedia Appendix 8). For instance, for participants with at least 3 morbidities, higher daily walking distance was significantly associated with lower DBP, which was not pronounced in terms of daily step volume. Second, to prove the generalizability of our results, we repeated the stratified analyses by categorizing step volume according to the tertiles of the whole population rather than specific subpopulations. The results were generally similar to the overall results (Multimedia Appendices 9 and 10). Finally, we conducted a week-to-week association analysis to reduce repeated measures of outcomes and exposures by calculating weekly averages of BP and step volume (Multimedia Appendices 11 and 12). The week-to-week

XSL•FO
RenderX

association results were still consistent with the week-to-day association results, even though the association magnitudes were relatively lower. In addition, we found that the PA of the previous week showed greater effects on BP, compared with the PA of the same week.

Discussion

Principal Findings

To the best of our knowledge, this is one of the few studies to investigate the short-term benefits of PA. Using wearable device–measured daily step data from the YIDO mHealth app across 5 years and over 0.5 million person-days, the principal findings of our study are 3-fold. First, among community-dwelling older adults, we observed consistent negative associations between daily step volume and BP (SBP, DBP, MAP, and PP). Second, we noted a potential synergistic effect of step volume and variability on BP. Specifically, compared with irregular steps, regular steps appeared to improve the beneficial relationships between step volume and BP. Third, stratified analyses suggested that the effects of step volume and variability on BP varied across different subpopulations. When considering both step volume and variability jointly, we observed diverse associational patterns between PA and BP among older adults. Overall, our findings provided a comprehensive evaluation of the relationships of step volume and variability with BP among older adults. We underscored that a “one-size-fits-all” recommendation for step volume may not be suitable for all older adults.

Independent Association Between Step Volume and BP

Previous studies have reported inconsistent evidence on the association between step volume and BP levels [4,22-25]. In a cross-sectional sample from the Electronic Framingham Heart Study, reductions in SBP of 0.49 mm Hg and DBP of 0.36 mm Hg were found per 1000 steps of daily step volume [25]. In comparison, a study of community-dwelling older participants did not observe a significant association between step counts and either SBP or DBP [4]. We postulated that prior studies with controversial results are limited by the short and cross-sectional monitoring periods of PA, which are prone to an observer effect and may not accurately reflect long-term PA behavior in real-world scenarios [26]. The development in digital health technologies made it feasible to remotely detect PA and BP continuously and longitudinally in a free-living environment. By using mHealth technologies, this is the first study to track the PA status of older adults in their daily lives for 5 years. Our findings contributed to the existing evidence that higher step volume is associated with lower BP (SBP, DBP, MAP, and PP).

A unique aspect of our findings was that the associational patterns varied between BP metrics. Specifically, our findings suggested that a high level of step volume showed the strongest beneficial effects on SBP and PP (≥9500 steps/d). In comparison, for DBP and MAP, the most significant association appeared in participants with a medium level of step volume (6000-9500 steps/d). Similarly, previous studies have reported that the relationship between step counts and adverse events may reverse as the step counts increase [5,10]. For example, the UK Biobank study, which quantified steps over a 7-day monitoring period and assessed mortality over a median of 7.0 years, found that a higher daily step count was associated with a lower mortality risk, whereas the risk increased after taking >10,000 steps per day [10]. We hypothesized that the divergent associational patterns between various BP metrics might be attributable to their physiological disparities: higher SBP and PP levels primarily reflect arterial stiffness in older adults, whereas DBP reflects coronary flow reserve [27]. Considering the varying patterns between BP metrics, we suggest the individualization of step targets depending on which BP metrics older adults were most burdened with.

The lag structure for the short-term effects was also investigated in this study. We observed that the step volume on the previous day of BP measurement was the most significantly associated with BP levels. In addition, it showed that cumulative effects, represented by the weekly average of daily step volume, had greater effects on BP than single-day exposure. A recent study showed that increasing PA levels were linked to blunted glycemic excursions during the next day, which was similar to our results [18]. A clinical trial study with 21 adults with prehypertension showed that SBP and DBP were significantly reduced after PA for 11 hours and 10 hours, respectively [28]. Another clinical trial observed that PA has antihypertensive effects within approximately 45 minutes, and these effects increase with the intensity of the exercise [29]. Possible mechanisms underlying the BP-lowering short-term effects of PA involve a reduction in sympathetic activity, lower insulin resistance due to the release of vasodilatory substances, and a decrease in plasma renin levels [30]. From a clinical perspective, short-term (24-hour) elevations in ambulatory BP have been identified as increasing the acute risk of cardiovascular disease events, including stroke and myocardial infarction [31]. Therefore, it is plausible that interventions aimed at improving daily step counts could have meaningful clinical implications in promoting the overall cardiovascular health, particularly among older adults.

Joint Association of Step Volume and Variability With BP Level

Our study also provides novel estimations of the joint associations of step volume and variability with BP. In our study, participants taking more and regular steps per day had lower levels of BP, compared to those taking fewer and irregular steps. Although current PA guidelines generally recommend regular PA for improving fitness [8,13], few studies have investigated the effects of PA variability quantitatively or the synergistic effects of PA volume and variability quantitatively. There is only 1 recent study that reported that the greater regularity of the PA level may play an important role in short-term weight loss [14]. We believe the limited evidence on PA variability is because most previous randomized controlled trials have been based on structured interventions, where regularity is naturally part of the intervention. It is also difficult for observational studies to track the regularity of PA in daily lives based on a short period of PA monitoring.

XSL•FO
RenderX

In this study, we found that individuals with high step volume and regular steps experienced a notable decrease in SBP and DBP by 1.69 mm Hg and 0.69 mm Hg, respectively, compared to those with low step volume and irregular steps. These findings hold clinical significance, as they align with the results of a comprehensive meta-analysis evaluating the efficacy of antihypertensive agents in cardiovascular disease prevention. The meta-analysis demonstrated that reductions in SBP by 1 mm Hg and DBP by 0.5 mm Hg were associated with a 4% decrease in the incidence of stroke and a 2% decrease in coronary heart disease events [32]. Additional meta-analyses also showed that, at the population level, even an SBP decrease of 1 mm Hg could lower the risk of stroke by 5% [33]. Given the comparable magnitude between our observed BP reductions and the beneficial effects demonstrated in the previous meta-analyses, although the short-term association coefficients in our study may appear modest, they hold relevance from a public health perspective.

The concept of variability in other health-related indicators or behaviors has been extensively studied. For example, studies have investigated variability in glycemia [34]; BP [35]; and, more recently, sleep patterns [36,37]. Such research reveals that variability in health-related domains can have detrimental effects on health outcomes, where biological rhythm disruptions are often implicated in explaining these associations [36]. Accordingly, we supposed that the observed associations of step variability with BP may be explained by a similar mechanism: irregular step patterns serve as a marker of irregular health behaviors in general, including irregular timing of meals or exercise, which may contribute to circadian disruption [38,39]. Therefore, it is plausible that regular step patterns could contribute to the maintenance of balanced rest-activity rhythms and homeostatic physiological systems, thereby promoting the beneficial effects of step volume. Experimental studies are warranted to examine the potential causal mechanisms underpinning the detected relationships of PA volume and variability with BP.

Stratified Analyses of the Association of Step Volume and Variability With BP

When classified by sex, the association of step volume and variability with BP seemed to be more apparent among female participants. For male participants, higher step volume was associated with lower DBP and MAP, although this association was not as pronounced for SBP and PP. Our findings are consistent with previous literature investigating the sex-related differences in the association of PA with BP. Previous epidemiological studies reported greater PA-associated reductions in BP among women than among men [40,41]. A systematic review provided further confirmation that compared to male individuals, female individuals experience a slightly greater reduction in both SBP and DBP due to the effects of walking [42]. The sex differences in the cardiovascular system persisted throughout life. Given that the sympathetic nervous system is more active in postmenopausal women [43,44], a possible explanation is that women may be more susceptible to the effects of PA on inducing neural network plasticity, which reduces sympathetic excitation [45].

Hypertensive effects were also investigated in our study. More significant associations between step volume and BP were observed in the population without hypertension. This observation is consistent with previous research indicating stronger effects of PA on BP in individuals with normal BP [41]. One possible explanation for this discrepancy may involve vessel wall thickening and remodeling in individuals with established hypertension, as arterial stiffness is often accompanied by a reduced sensitivity to vasodilatory effects. Therefore, the vasodilatory efficacy of PA may be greater in individuals with normal BP with elastic arteries than in individuals with hypertension with stiff arteries and advanced atherosclerotic wall changes [46]. Notably, although significant associations between step volume alone and DBP or MAP were not observed among patients with hypertension, pronounced effects were evident when considering the combination of step volume and variability. This suggests that individuals with hypertension may be particularly sensitive to the regularity of PA, emphasizing the importance of considering both the volume and regularity of exercise when establishing step targets for older individuals with hypertension.

Another innovation of our study lies in stratification according to the number of morbidities. Older adults commonly experience multiple chronic conditions that cause great burdens and costs [47]. Consistent with this, our study found that all participants had at least 1 chronic disease. However, present guidelines on PA offer uniform recommendations for all adults aged >65 years, irrespective of their chronic conditions [8,13]. Our findings suggested that step volume was inversely associated with BP across all subgroups with varying numbers of morbidities, with greater impacts in participants with only 1 morbidity. Participants with a great chronic disease burden may experience other pathogenetic pathways that more significantly affect the fluctuation of BP, such as oxidative stress and mitochondrial dysfunction [48], thus obscuring the impact of steps on BP. When evaluating the joint effects of volume and variability, the associations became more apparent in patients with >2 morbidities, possibly due to their increased sensitivity to PA regularity, as discussed earlier. Overall, our results underscore heterogeneous associational patterns among older adults with diverse chronic conditions, warranting personalized activity prescriptions for the older population.

Strengths and Limitations

The main strengths of our study include a relatively large sample of community-dwelling older adults and longitudinal and continuous recordings of their PA in free-living conditions over 5 years. Wearable devices were used to measure daily steps and walking distance, which helped avoid participant errors in self-reported PA and accounted for the steps involved in any type of ambulatory activity [9]. To our knowledge, this is the first study estimating the association between step volume and BP over the entirety of an individual's monitoring period rather than a brief snapshot. We also innovatively highlighted the joint effects of step volume and step variability on BP. Moreover, multifaceted subgroup analyses were conducted to compare discrepancies in the effects of PA on BP according to different population characteristics.

XSL•FO
RenderX

Some limitations to our study must be considered in interpreting the findings. First, the observational nature of our data precludes a definitive causal interpretation of our findings, although we took several measures to minimize risks for reverse causalities and confounding, such as including a wide range of covariates and conducting sensitivity analyses. The precise mechanisms explaining the short-term relationships between PA and BP should be assessed using different study designs. Second, there were significant differences in some characteristics between retained individuals and those excluded because of missing data. In addition, the study sample comprised only individuals who uploaded their PA and BP data as users of the YIDO app, for which further selection biases may exist. Thus, the generalizability of our results to nonretained individuals may be limited, although the average step counts of our participants were similar to those reported in prior studies involving Chinese older adults [2,49]. Third, there is a potential for unmeasured confounding in our analyses. For example, we were not able to track the BMI status of participants over time. Furthermore, information on individual dietary (eg, alcohol consumption) was not available, so we could neither control for them in our analyses nor assess their potential relative contribution to PA and BP change. Further investigations with complementary data on individual diets are warranted to address these gaps. Fourth, the 3-axis acceleration used by the commercial wearable device to measure step count has been reported to have high degrees of accuracy for step counts [50]. However, evidence demonstrating the validity and reliability of step counts measured using the Lexin Mio smart bracelet is limited, so additional studies are needed to investigate possible measurement errors. Step counts can vary widely by device brand and algorithms or analytical decisions [2]. Finally, PA intensity or cadence (steps/min) was not accessible in this study, as we had access only to daily step data. Nonetheless, daily step count is an increasingly widespread and intuitive metric that can be easily used as a target for health benefits.

Conclusions

This study extends the contemporary evidence confirming that a greater number of daily steps is significantly associated with a lower level of BP (SBP, DBP, MAP, and PP). In addition, we observed that step volume and variability are jointly associated with BP. Participants with high step volume and regular steps had lower BP compared with participants with low step volume and irregular steps. Subgroup analyses further indicated heterogeneous associational patterns among older adults with various chronic conditions. For participants with hypertension and with more chronic disease burden, PA interventions combining both step volume and variability can result in more benefits in controlling BP. Collectively, these results provide an evidence base for refining activity recommendations based on an individual's characteristics. This study also provides an example of the potential clinical value of the integration of consumer wearable devices and mHealth apps.

Acknowledgments

This research was funded by the Yingdong Intelligent Technology Co, Ltd. Yingdong Intelligent Technology Co, Ltd, was not involved in the design of the study; data collection, analysis, or interpretation; writing or editing of the manuscript; or decision to submit the manuscript for publication. The results of this study are presented clearly; honestly; and without fabrication, falsification, or inappropriate data manipulation. The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sector.

Data Availability

The data sets used or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

DC and XD contributed to the design of the study, generated hypotheses, and critically reviewed the manuscript. HX contributed to the development of the search strategy, extracted and analyzed the data, and wrote and critically reviewed the manuscript. ZZ, YM, XL, and KD extracted and interpreted the data. Xiaotong Dai (daixiaotong@bsu.edu.cn) and Dafang Chen (dafangchen@bjmu.edu.cn) are co-corresponding authors of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cut-off values of daily step volume across different groups.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v26i1e50075_app1.xlsx]

Multimedia Appendix 2
Numbers and proportions for the combined groups.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v26i1e50075_app2.xlsx]

XSL•FO
RenderX

Multimedia Appendix 3
Characteristics of the study population for week-to-week associations.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v26i1e50075_app3.xlsx]

Multimedia Appendix 4
The independent associations of daily step volume and daily walking distance with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure).
[XLSX File (Microsoft Excel File), 13 KB - jmir_v26i1e50075_app4.xlsx]

Multimedia Appendix 5
The joint associations of physical activity (daily steps and walking distance) volume and variability with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure).
[XLSX File (Microsoft Excel File), 11 KB - jmir_v26i1e50075_app5.xlsx]

Multimedia Appendix 6
Stratified analyses for the independent associations of daily step volume and daily walking distance with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure).
[XLSX File (Microsoft Excel File), 14 KB - jmir_v26i1e50075_app6.xlsx]

Multimedia Appendix 7
Stratified analyses for the joint associations of physical activity (daily steps and walking distance) volume and variability with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure).
[XLSX File (Microsoft Excel File), 18 KB - jmir_v26i1e50075_app7.xlsx]

Multimedia Appendix 8
Estimated changes (95% CIs) in blood pressure (BP; systolic BP, diastolic BP, mean arterial pressure, and pulse pressure) related to daily walking distance in different lag days and week averages, joint effects of daily walking distance and variability, and weekly average daily step volume and walking distance across gender, the presence of hypertension, and chronic disease status.
[DOCX File , 700 KB - jmir_v26i1e50075_app8.docx]

Multimedia Appendix 9
Stratified analyses for the independent associations of daily step volume and daily walking distance with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure), categorized according to the tertiles of the whole population.
[XLSX File (Microsoft Excel File), 14 KB - jmir_v26i1e50075_app9.xlsx]

Multimedia Appendix 10
Stratified analyses for the joint associations of physical activity (daily steps and walking distance) volume and variability with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure), categorized according to the tertiles of the whole population.
[XLSX File (Microsoft Excel File), 17 KB - jmir_v26i1e50075_app10.xlsx]

Multimedia Appendix 11
Independent associations of daily step volume and daily walking distance with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure) in week-to-week association patterns.
[XLSX File (Microsoft Excel File), 19 KB - jmir_v26i1e50075_app11.xlsx]

Multimedia Appendix 12
Joint associations of physical activity (daily steps and walking distance) volume and variability with blood pressure (systolic blood pressure, diastolic blood pressure, mean arterial pressure, and pulse pressure) in week-to-week association patterns.
[XLSX File (Microsoft Excel File), 28 KB - jmir_v26i1e50075_app12.xlsx]

References

1. Benetos A, Petrovic M, Strandberg T. Hypertension management in older and frail older patients. Circ Res 2019 Mar 29;124(7):1045-1060 [FREE Full text] [doi: 10.1161/CIRCRESAHA.118.313236] [Medline: 30920928]

XSL•FO
RenderX

2. Lin H, Sardana M, Zhang Y, Liu C, Trinquart L, Benjamin EJ, et al. Association of habitual physical activity with cardiovascular disease risk. Circ Res 2020 Oct 23;127(10):1253-1260 [FREE Full text] [doi: 10.1161/CIRCRESAHA.120.317578] [Medline: 32842915]
3. Börjesson M, Onerup A, Lundqvist S, Dahlöf B. Physical activity and exercise lower blood pressure in individuals with hypertension: narrative review of 27 RCTs. Br J Sports Med 2016 Mar 19;50(6):356-361. [doi: 10.1136/bjsports-2015-095786] [Medline: 26787705]
4. Bravata DM, Smith-Spangler C, Sundaram V, Gienger AL, Lin N, Lewis R, et al. Using pedometers to increase physical activity and improve health: a systematic review. JAMA 2007 Nov 21;298(19):2296-2304. [doi: 10.1001/jama.298.19.2296] [Medline: 18029834]
5. Dwyer T, Ponsonby AL, Ukoumunne OC, Pezic A, Venn A, Dunstan D, et al. Association of change in daily step count over five years with insulin sensitivity and adiposity: population based cohort study. BMJ 2011 Jan 13;342(Jan 13 1):c7249. [doi: 10.1136/bmj.c7249] [Medline: 21233153]
6. Ewald B, McEvoy M, Attia J. Pedometer counts superior to physical activity scale for identifying health markers in older adults. Br J Sports Med 2010 Aug 14;44(10):756-761. [doi: 10.1136/bjsm.2008.048827] [Medline: 18625631]
7. Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, et al. Association of step counts over time with the risk of chronic disease in the All of Us Research Program. Nat Med 2022 Nov;28(11):2301-2308 [FREE Full text] [doi: 10.1038/s41591-022-02012-w] [Medline: 36216933]
8. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA 2018 Nov 20;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
9. Mañas A, Del Pozo Cruz B, Ekelund U, Losa Reyna J, Rodríguez Gómez I, Carnicero Carreño JA, et al. Association of accelerometer-derived step volume and intensity with hospitalizations and mortality in older adults: a prospective cohort study. J Sport Health Sci 2022 Sep;11(5):578-585 [FREE Full text] [doi: 10.1016/j.jshs.2021.05.004] [Medline: 34029758]
10. Del Pozo Cruz B, Ahmadi MN, Lee IM, Stamatakis E. Prospective associations of daily step counts and intensity with cancer and cardiovascular disease incidence and mortality and all-cause mortality. JAMA Intern Med 2022 Nov 01;182(11):1139-1148 [FREE Full text] [doi: 10.1001/jamainternmed.2022.4000] [Medline: 36094529]
11. Hamaya R, Fukuda H, Takebayashi M, Mori M, Matsushima R, Nakano K, et al. Effects of an mHealth app (Kencom) with integrated functions for healthy lifestyles on physical activity levels and cardiovascular risk biomarkers: observational study of 12,602 users. J Med Internet Res 2021 Apr 26;23(4):e21622 [FREE Full text] [doi: 10.2196/21622] [Medline: 33900203]
12. Hamaya R, Mori M, Miyake K, Lee IM. Association of smartphone-recorded steps over years and change in cardiovascular risk factors among working-age adults. J Am Heart Assoc 2022 Jul 19;11(14):e025689 [FREE Full text] [doi: 10.1161/JAHA.121.025689] [Medline: 35861838]
13. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
14. El Fatouhi D, Delrieu L, Goetzinger C, Malisoux L, Affret A, Campo D, et al. Associations of physical activity level and variability with 6-month weight change among 26,935 users of connected devices: observational real-life study. JMIR Mhealth Uhealth 2021 Apr 15;9(4):e25385 [FREE Full text] [doi: 10.2196/25385] [Medline: 33856352]
15. Zhang N, Zhou M, Li M, Ma G. Effects of smartphone-based remote interventions on dietary intake, physical activity, weight control, and related health benefits among the older population with overweight and obesity in China: randomized controlled trial. J Med Internet Res 2023 Apr 28;25:e41926 [FREE Full text] [doi: 10.2196/41926] [Medline: 37115608]
16. Cai Y, Wang Z, Zhang W, Kong W, Jiang J, Zhao R, et al. Estimation of heart rate and energy expenditure using a smart bracelet during different exercise intensities: a reliability and validity study. Sensors (Basel) 2022 Jun 21;22(13):4661 [FREE Full text] [doi: 10.3390/s22134661] [Medline: 35808157]
17. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008 Jan;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
18. El Fatouhi D, Héritier H, Allémann C, Malisoux L, Laouali N, Riveline JP, et al. Associations between device-measured physical activity and glycemic control and variability indices under free-living conditions. Diabetes Technol Ther 2022 Mar 01;24(3):167-177 [FREE Full text] [doi: 10.1089/dia.2021.0294] [Medline: 34648353]
19. Matura LA, Shou H, Fritz JS, Smith KA, Vaidya A, Pinder D, et al. Physical activity and symptoms in pulmonary arterial hypertension. Chest 2016 Jul;150(1):46-56 [FREE Full text] [doi: 10.1016/j.chest.2016.02.633] [Medline: 26892602]
20. Boehme MW, Buechele G, Frankenhauser-Mannuss J, Mueller J, Lump D, Boehm BO, et al. Prevalence, incidence and concomitant co-morbidities of type 2 diabetes mellitus in South Western Germany--a retrospective cohort and case control study in claims data of a large statutory health insurance. BMC Public Health 2015 Sep 03;15(1):855 [FREE Full text] [doi: 10.1186/s12889-015-2188-1] [Medline: 26334523]
21. Bowling CB, Sloane R, Pieper C, Luciano A, Davis BR, Simpson LM, et al. Association of sustained blood pressure control with multimorbidity progression among older adults. J Am Geriatr Soc 2020 Sep 05;68(9):2059-2066 [FREE Full text] [doi: 10.1111/jgs.16558] [Medline: 32501546]

XSL•FO
RenderX

22. Emberson MA, Lalande A, Wang D, McDonough DJ, Liu W, Gao Z. Effectiveness of smartphone-based physical activity interventions on individuals' health outcomes: a systematic review. Biomed Res Int 2021;2021:6296896 [FREE Full text] [doi: 10.1155/2021/6296896] [Medline: 34409104]
23. Hajna S, Ross NA, Dasgupta K. Steps, moderate-to-vigorous physical activity, and cardiometabolic profiles. Prev Med 2018 Feb;107:69-74 [FREE Full text] [doi: 10.1016/j.ypmed.2017.11.007] [Medline: 29126915]
24. Sagawa N, Rockette-Wagner B, Azuma K, Ueshima H, Hisamatsu T, Takamiya T, et al. Physical activity levels in American and Japanese men from the ERA-JUMP Study and associations with metabolic syndrome. J Sport Health Sci 2020 Mar;9(2):170-178 [FREE Full text] [doi: 10.1016/j.jshs.2019.09.007] [Medline: 32099725]
25. Sardana M, Lin H, Zhang Y, Liu C, Trinquart L, Benjamin EJ, et al. Association of habitual physical activity with home blood pressure in the electronic Framingham Heart Study (eFHS): cross-sectional study. J Med Internet Res 2021 Jun 24;23(6):e25591 [FREE Full text] [doi: 10.2196/25591] [Medline: 34185019]
26. Clemes SA, Matchett N, Wane SL. Reactivity: an issue for short-term pedometer studies? Br J Sports Med 2008 Jan;42(1):68-70. [doi: 10.1136/bjsm.2007.038521] [Medline: 18178685]
27. Sesso HD, Stampfer MJ, Rosner B, Hennekens CH, Gaziano JM, Manson JE, et al. Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. Hypertension 2000 Nov;36(5):801-807. [doi: 10.1161/01.hyp.36.5.801] [Medline: 11082146]
28. Park S, Rink LD, Wallace JP. Accumulation of physical activity leads to a greater blood pressure reduction than a single continuous session, in prehypertension. J Hypertens 2006 Sep;24(9):1761-1770. [doi: 10.1097/01.hjh.0000242400.37967.54] [Medline: 16915025]
29. Eicher JD, Maresh CM, Tsongalis GJ, Thompson PD, Pescatello LS. The additive blood pressure lowering effects of exercise intensity on post-exercise hypotension. Am Heart J 2010 Sep;160(3):513-520. [doi: 10.1016/j.ahj.2010.06.005] [Medline: 20826261]
30. Saint-Maurice PF, Troiano RP, Bassett DRJ, Graubard BI, Carlson SA, Shiroma EJ, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160 [FREE Full text] [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
31. Conen D, Bamberg F. Noninvasive 24-h ambulatory blood pressure and cardiovascular disease: a systematic review and meta-analysis. J Hypertens 2008 Jul;26(7):1290-1299. [doi: 10.1097/HJH.0b013e3282f97854] [Medline: 18550999]
32. Law MR, Morris JK, Wald NJ. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. BMJ 2009 May 19;338:b1665 [FREE Full text] [doi: 10.1136/bmj.b1665] [Medline: 19454737]
33. Grossman E. Blood pressure: the lower, the better: the con side. Diabetes Care 2011 May;34 Suppl 2(Suppl 2):S308-S312 [FREE Full text] [doi: 10.2337/dc11-s245] [Medline: 21525474]
34. Wang A, Liu X, Xu J, Han X, Su Z, Chen S, et al. Visit-to-visit variability of fasting plasma glucose and the risk of cardiovascular disease and all-cause mortality in the general population. J Am Heart Assoc 2017 Nov 29;6(12):e006757 [FREE Full text] [doi: 10.1161/JAHA.117.006757] [Medline: 29187392]
35. Nwabuo CC, Yano Y, Moreira HT, Appiah D, Vasconcellos HD, Aghaji QN, et al. Association between visit-to-visit blood pressure variability in early adulthood and myocardial structure and function in later life. JAMA Cardiol 2020 Jul 01;5(7):795-801 [FREE Full text] [doi: 10.1001/jamacardio.2020.0799] [Medline: 32293640]
36. Morales-Ghinaglia N, Larsen M, He F, Calhoun SL, Vgontzas AN, Liao J, et al. Circadian misalignment impacts the association of visceral adiposity with elevated blood pressure in adolescents. Hypertension 2023 Apr;80(4):861-871 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.122.20398] [Medline: 36876501]
37. Scott H, Lechat B, Guyett A, Reynolds AC, Lovato N, Naik G, et al. Sleep irregularity is associated with hypertension: findings from over 2 million nights with a large global population sample. Hypertension 2023 May;80(5):1117-1126. [doi: 10.1161/HYPERTENSIONAHA.122.20513] [Medline: 36974682]
38. Gabriel BM, Zierath JR. Circadian rhythms and exercise - re-setting the clock in metabolic disease. Nat Rev Endocrinol 2019 Apr;15(4):197-206. [doi: 10.1038/s41574-018-0150-x] [Medline: 30655625]
39. Smolensky MH, Hermida RC, Portaluppi F. Circadian mechanisms of 24-hour blood pressure regulation and patterning. Sleep Med Rev 2017 Jun;33:4-16. [doi: 10.1016/j.smrv.2016.02.003] [Medline: 27076261]
40. Manjoo P, Joseph L, Pilote L, Dasgupta K. Sex differences in step count-blood pressure association: a preliminary study in type 2 diabetes. PLoS One 2010 Nov 22;5(11):e14086 [FREE Full text] [doi: 10.1371/journal.pone.0014086] [Medline: 21124929]
41. Zou Q, Wang H, Su C, Du W, Ouyang Y, Jia X, et al. Longitudinal association between physical activity and blood pressure, risk of hypertension among Chinese adults: China Health and Nutrition Survey 1991-2015. Eur J Clin Nutr 2021 Feb 13;75(2):274-282. [doi: 10.1038/s41430-020-0653-0] [Medline: 32404900]
42. Lee LL, Mulvaney CA, Wong YK, Chan ES, Watson MC, Lin HH. Walking for hypertension. Cochrane Database Syst Rev 2021 Feb 24;2(2):CD008823 [FREE Full text] [doi: 10.1002/14651858.CD008823.pub2] [Medline: 33630309]
43. Joyner MJ, Charkoudian N, Wallin BG. Sympathetic nervous system and blood pressure in humans: individualized patterns of regulation and their implications. Hypertension 2010 Jul;56(1):10-16 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.109.140186] [Medline: 20497993]

XSL•FO
RenderX

44. Joyner MJ, Wallin BG, Charkoudian N. Sex differences and blood pressure regulation in humans. Exp Physiol 2016 Mar;101(3):349-355 [FREE Full text] [doi: 10.1113/EP085146] [Medline: 26152788]
45. Mueller PJ. Exercise training and sympathetic nervous system activity: evidence for physical activity dependent neural plasticity. Clin Exp Pharmacol Physiol 2007 Apr;34(4):377-384. [doi: 10.1111/j.1440-1681.2007.04590.x] [Medline: 17324153]
46. Dimeo F, Pagonas N, Seibert F, Arndt R, Zidek W, Westhoff TH. Aerobic exercise reduces blood pressure in resistant hypertension. Hypertension 2012 Sep;60(3):653-658. [doi: 10.1161/HYPERTENSIONAHA.112.197780] [Medline: 22802220]
47. Karlamangla A, Tinetti M, Guralnik J, Studenski S, Wetle T, Reuben D. Comorbidity in older adults: nosology of impairment, diseases, and conditions. J Gerontol A Biol Sci Med Sci 2007 Mar 01;62(3):296-300. [doi: 10.1093/gerona/62.3.296] [Medline: 17389727]
48. Barnes PJ. Mechanisms of development of multimorbidity in the elderly. Eur Respir J 2015 Mar;45(3):790-806 [FREE Full text] [doi: 10.1183/09031936.00229714] [Medline: 25614163]
49. Ho JY, Goggins WB, Mo PK, Chan EY. The effect of temperature on physical activity: an aggregated timeseries analysis of smartphone users in five major Chinese cities. Int J Behav Nutr Phys Act 2022 Jun 14;19(1):68 [FREE Full text] [doi: 10.1186/s12966-022-01285-1] [Medline: 35701809]
50. Park J, Ishikawa-Takata K, Tanaka S, Bessyo K, Tanaka S, Kimura T. Accuracy of estimating step counts and intensity using accelerometers in older people with or without assistive devices. J Aging Phys Act 2017 Jan;25(1):41-50. [doi: 10.1123/japa.2015-0201] [Medline: 27180730]

Abbreviations

BP: blood pressure
DBP: diastolic blood pressure
MAP: mean arterial pressure
mHealth: mobile health
PA: physical activity
PP: pulse pressure
SBP: systolic blood pressure

Edited by S Ma; submitted 19.06.23; peer-reviewed by K Madden, J Claggett, A Angelucci; comments to author 10.03.24; revised version received 08.04.24; accepted 07.06.24; published 14.08.24.

Please cite as:
Xiao H, Zhou Z, Ma Y, Li X, Ding K, Dai X, Chen D
Association of Wearable Device–Measured Step Volume and Variability With Blood Pressure in Older Chinese Adults: Mobile-Based Longitudinal Observational Study
J Med Internet Res 2024;26:e50075
URL: https://www.jmir.org/2024/1/e50075
doi:10.2196/50075
PMID:

©Han Xiao, Zechen Zhou, Yujia Ma, Xiaoyi Li, Kexin Ding, Xiaotong Dai, Dafang Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Expectations and Preferences for Digital Cessation Treatment: Multimethods Study Among Older Adults Who Smoke Cigarettes

Margaret C Fahey[1], PhD; Mathew J Carpenter[1,2], PhD; Riley O'Neal[3], BA; Kinsey Pebley[2,4], PhD; Melissa R Schick[5], PhD; Emily Ware[2,6], PharmD; Benjamin A Toll[1,2,4], PhD; Jennifer Dahne[1,2], PhD

[1]Department of Psychiatry & Behavioral Sciences, Medical University of South Carolina, Charleston, SC, United States
[2]Hollings Cancer Center, Charleston, SC, United States
[3]School of Arts & Sciences, University of South Carolina, Columbia, SC, United States
[4]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States
[5]School of Medicine, Yale University, New Haven, CT, United States
[6]Department of Pharmacy Services, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Margaret C Fahey, PhD
Department of Psychiatry & Behavioral Sciences
Medical University of South Carolina
86 Jonathan Lucas Street
Charleston, SC, 29425
United States
Phone: 1 6158389307
Email: mfahey@mtsu.edu

Related Articles:

Comment in: http://jmir.org/2024/1/e65929/

Comment in: http://jmir.org/2024/1/e67749/

Abstract

Background: To address enduring age-related tobacco disparities, it is critical to promote cessation treatment among older adults (aged 65+ years). Digital health platforms offer opportunities for wide dissemination of evidence-based behavioral cessation support. However, existing digital cessation treatments are not tailored to unique aging-related needs and preferences, resulting in low uptake. Detailed information is needed about how to best adapt these treatments for this age group.

Objective: We aimed to collect detailed, hypothesis-generating information about expectations and preferences for cessation digital treatment among older adults who smoke cigarettes.

Methods: Semistructured interviews were conducted with adults aged 65+ years currently smoking or who had quit within the past month. Interviews included open-ended questions regarding prior experiences with digital health platforms and expectations and preferences for cessation treatment via various modalities (app-delivered, texting-based, or videoconferencing counseling). Interviews also elicited questions regarding digital modalities that integrated social components (app-delivered social forums and group videoconferencing counseling). Using an iterative, team-based approach, the thematic analysis identified meaningful themes. Interviews were supplemented with quantitative measures assessing sociodemographics, digital literacy, and physical health symptoms.

Results: Participants (12/20, 60% men; 15/20, 75% White; 4/20, 20% Black or African American; 1/20, 5% Asian) were currently smoking (17/20, 85%) or had recently quit (3/20, 15%). Thematic analysis identified 3 meaningful themes across all digital modalities: convenience, accessibility, and personalization. Expected benefits of digital platforms included convenient treatment access, without reliance on transportation. Participants preferred treatments to be personalized and deliver content or strategies beyond standard education. Most (17/20, 85%) were unfamiliar with cessation apps but found them appealing given the potential for offering a novel quitting strategy. App ease of use (eg, easy navigation) was preferred. Half (10/20, 50%) would try a texting-based intervention, with many preferring texting with a counselor rather than automated messaging. Most (17/20,

XSL•FO
RenderX

85%) would use videoconferencing and expected this modality to deliver better quality counseling than via telephone. Expected videoconferencing challenges included looking presentable onscreen, technological difficulties, and privacy or security. Videoconferencing was regarded as the most personalized digital treatment, yet benefits unique to app-delivered and texting-based treatments included anonymity and access to treatment 24/7. Participants expected integrating social components into digital treatment to be useful for quit success and social connection, yet were concerned about possible interpersonal challenges.

Conclusions: Because a long history of quit attempts and familiarity with standard quitting advice is common among older adults who smoke cigarettes, digital platforms might offer appealing and novel strategies for cessation that are accessible and convenient. Overall, this population was open to trying digital cessation treatments and would prefer that these platforms prioritize ease of use and personalized content. These findings challenge the bias that older adults are uninterested or unwilling to engage with digital treatments for behavioral health.

(J Med Internet Res 2024;26:e52919) doi:10.2196/52919

KEYWORDS

older adults; digital health; digital cessation treatment; smoking cessation; digital cessation; treatment; smoke; cigarettes; cigarette; tobacco; adults; elderly; older person; older people; aging; smoking; smoke; quit; quitting; questionnaire; telehealth; treatment; treatments; behavioral health; public health; mobile phone

Introduction

In total, 1 in 7 Americans is aged 65 years or older, an age group that will comprise 22% of the US population by 2040 [1]. Cigarette smoking is a leading cause of cancer, preventable disease, and death in the United States, with tobacco morbidity and mortality disproportionately impacting older adults [2]. Cigarette smoking exacerbates numerous chronic health conditions that are more common in older age (eg, diabetes or chronic pain) [2-5]. Annual smoking-related mortality is almost twice as high for older (392,000 deaths) compared to younger (163,000 deaths) adults [2]. Although younger age groups have seen significant declines, smoking prevalence has stagnated among older adults (9%) for the past 15 years [6-9]. Despite lower smoking prevalence in older compared to younger adults, this difference is attributed to older adults being much more likely to die from smoking [8,10]. Fortunately, cessation offers significant health advantages in older age, including improved physical health and increased lifespan [2-4,11,12]. To address these age-related tobacco disparities, there is a critical need to promote the use of evidence-based cessation treatments in this age group.

More than half of older adults who smoke cigarettes want to quit [13], but only about a third (37%) use an evidence-based treatment when making a quit attempt [14]. However, when older adults do engage in evidence-based treatment, they experience comparable (and sometimes higher) quit rates compared to their younger counterparts [15]. Digital health treatment modalities offer opportunities for the wide dissemination of evidence-based behavioral cessation support (eg, mobile apps, texting programs, or counseling via videoconferencing) [16]. Digital platforms can overcome in-person barriers (eg, physical limitations or lack of transportation) common in older age, increase the availability of treatment content (eg, at times of cravings), and supplement limited provider time in medical settings [16-18]. Further, ownership of consumer technologies is increasing in this age group, as 88% of US adults aged 56-70 years and 72% aged 70+ years owned smartphones in 2022 compared to 81% and 62%, respectively, in 2021 [19]. Digital treatments effectively promote a variety of other health behaviors (eg, physical activity) in later life [20,21] and could be a means to widely disseminate cessation treatment to this high-priority population.

Despite the promise of digital health for this age group, to our knowledge, no digital cessation treatment is tailored specifically to unique aging-related needs and preferences, despite many focused on adolescent or young adult populations [16,22-24]. The lack of tailored support for this age group might contribute to low uptake. For example, the National Cancer Institute Smokefree.gov app-delivered cessation programs are free and produce quit rates comparable to other behavioral interventions, but only 2% to 3% of its users are aged 65+ years [22,25]. Further, the Washington Department of Health provides a freely available app-delivered program, yet older adults comprise <3% of those using this program (despite comprising approximately 13% of tobacco users in the state) [23]. The bias that older adults are unwilling or unable to engage in digital health platforms has contributed to their widespread exclusion in the development of digital treatments, including for cessation [26]. Consideration of older adults' needs in the design and development of digital health interventions (eg, accommodating sensory and dexterity impairments) increases their treatment engagement [27]. Thus, to address this divide in the growing field of digital cessation treatment [16,22,23], detailed information is needed about how to best adapt these treatments for older adults.

This multimethods study aimed to collect detailed, hypothesis-generating information via semistructured qualitative interviews about the expectations and preferences for cessation digital treatment among older adults (aged 65+ years) who smoke cigarettes within an academic medical center. Interviews were supplemented with quantitative measures assessing sociodemographic characteristics, digital literacy, and physical health symptoms. Outcomes were intended to inform (1) the expected benefits and challenges of using digital cessation treatments across several modalities (ie, app-delivered, texting-based, or videoconferencing counseling) and (2) preferences and suggestions for digital treatments.

XSL•FO
RenderX

Methods

Study Sample and Recruitment

Participants were recruited by 1 of 2 methods. First, leveraging Medical University of South Carolina (MUSC) electronic health record (EHR) data, patients aged 65+ years currently smoking cigarettes were identified via a study recruitment report. Consistent with institutional review board (IRB) procedures, patients were contacted by phone and email. Those interested were asked to complete a phone-based screening questionnaire to determine eligibility. Second, patients identified as aged 65+ years were referred by providers from the MUSC Tobacco Treatment Program (ie, pharmacotherapy or behavioral counseling).

Eligibility criteria included being aged >65 years and having smoked 5+ cigarettes per day on more days than not in the past month. Individuals who recently quit (eg, within the past week) but had still smoked 5+ cigarettes per day on more days than not in the past month were included. Exclusion criteria included (1) non-English speaking, (2) no access to a telephone, (3) any self-reported cognitive impairment, and (4) having another household member enrolled in this study.

Following phone eligibility screening, participants completed electronic consent (eConsent) via a Health Insurance Portability and Accountability Act–compliant REDCap (Research Electronic Data Capture; Vanderbilt University) database [28]. Participants needed one-time access to the internet to read and sign the consent form electronically. Owning a personal device with internet access (ie, smartphone) was not required for participation. Study staff spoke with participants via telephone throughout the eConsent process, offering opportunities to ask questions about the consent form and instructions on how to access the form and sign electronically via a study link (if necessary). The remainder of the study procedures (ie, questionnaires and semistructured interviews) were administered via telephone.

Procedures

Following informed consent, participants completed questionnaires and a semistructured interview (Multimedia Appendix 1).

Semistructured Interviews

Following questionnaires, participants completed semistructured interviews with the lead researcher (MCF). These telephone interviews were approximately 45 minutes in length and audio recorded. Interviews consisted of a series of open-ended questions regarding cigarette smoking and quitting history, prior experiences with digital health treatment (eg, medical appointments or cessation treatment), and expectations and preferences for cessation treatment via different digital modalities (ie, app-delivered, texting-based, or videoconferencing counseling). Interviewees were also asked about expectations and preferences for digital treatments integrating social components (ie, social forums within an app-delivered treatment or group-based videoconferencing counseling).

Measures

Demographics

Participants reported their age in years, gender, race, ethnicity, household income, education, marital status, and health insurance status.

Cigarette Use and Quitting History

Participants reported (yes or no) if they had recently quit cigarettes (defined as smoking 0 cigarettes, not even a puff, in the past 7 days), age of first cigarette, and use (yes or no) of other nicotine- and tobacco-containing products (ie, electronic cigarettes, smokeless tobacco, pipe, hookah, or cigars). Those who were currently smoking cigarettes reported the number of cigarettes they smoked per day and how soon after waking they had their first cigarette (within 5 minutes, 6-30 minutes, 30-60 minutes, and after 60 minutes). Participants also reported their motivation and confidence in quitting or remaining quit from cigarettes on a scale of 0-10, with 0 being the lowest motivation or confidence and 10 being the highest.

Physical Health

Participants reported whether any impairment or health problem limited any of their activities in any way (yes or no) using the Health-Related Quality of Life Activity Limitations Module [29]. If yes, participants reported the type of major limitation or health problem (arthritis or rheumatism, back or neck problems, fractures, or bone or joint injury, walking problem, lung or breathing problem, hearing problem, eye or vision problem, heart problem, stroke problem, hypertension or high blood pressure, diabetes, cancer, depression, anxiety, or emotional problem, and other impairment or problem). Finally, participants reported the presence of any sensory impairments (ie, visual or hearing; yes or no).

Digital Literacy

The Mobile Device Proficiency Questionnaire, a 16-item measure, assessed mobile device proficiency [30]. Scores range from 16 to 80, with higher scores indicating higher digital literacy.

Data Analysis

Recruitment was discontinued when data saturation was reached (no new themes were evident in 2 consecutive interviews) [31]. Audio-recorded interviews were deidentified by IRB-approved study staff and transcribed verbatim. An iterative approach to thematic analysis was used to code the qualitative interviews using NVivo software [32,33]. Further, 2 coders (MCF and RO) developed a codebook organized by interview prompts and the type of digital modality discussed. The same 2 coders then reviewed and coded a subsample of deidentified interviews (n=2) independently using the codebook. Using an iterative team-based approach, discrepancies were discussed, and the codebook was further refined. The codebook was updated (as needed) with each interview based on emerging themes. All deidentified interviews were independently double-coded (MCF, RO, and KP). Coders were reliable, with an agreement of 90%. An independent coder (MCF, RO, KP, or MRS) was chosen to resolve any coding discrepancies. Using NVivo software, codes facilitated the team's (MCF, RO, KP, and MRS) identification,

XSL•FO
RenderX

defining, and naming of themes meaningful to the expected benefits and challenges of digital cessation treatments, as well as preferences and considerations regarding digital design.

Ethical Considerations

All study procedures were approved by the MUSC IRB (PRO00116590). Participants provided informed consent and were compensated US $50 via Amazon gift cards for participation. Interviews were audio recorded and transcribed. All data in this paper are deidentified.

Results

Sample Overview

In total, 20 participants completed semistructured interviews and Table 1 displays sociodemographic information. The majority (n=17, 85%) were currently smoking cigarettes and 3 (15%) participants had quit within the past month. Motivation to quit or remain quit was high (mean 7.9, SD 3.1), and half (n=10, 50%) reported having a physical limitation or impairment that limited their daily activities. Most identified as White (n=15, 75%), 20% (n=4) identified as Black or African American, and 5% (n=1) identified as Asian. All were non-Hispanic, and most were men (n=12, 60%) and not concurrently using other tobacco products (n=18, 90%). Household income was varied.

XSL•FO
RenderX

Table 1. Sample characteristics (N=20).

Characteristics	Value
Age (years)	
Mean (SD)	70.7 (3)
Range	66-76
Gender, n (%)	
Men	12 (60)
Women	8 (40)
Race, n (%)	
White	15 (75)
Black or African American	4 (20)
Asian	1 (5)
Ethnicity, n (%)	
Non-Hispanic	20 (100)
Marital status, n (%)	
Married, partnered, or living as married	10 (50)
Not married	10 (50)
Household income (US $), n (%)	
<25,000	3 (15)
25,000-50,000	5 (25)
50,000-100,000	9 (45)
100,000-200,000	3 (15)
Smoking characteristics	
Recently quit, n (%)	3 (15)
Currently smoking, n (%)	17 (85)
Cigarettes per day, mean (SD)	10 (7.3)
Motivation to quit or remain quit, mean (SD)	7.9 (3.1)
Confidence to quit or remain quit, mean (SD)	6.2 (3.2)
Time to first cigarette (min), n (%)	
Within 5	1 (5)
6 to 30	7 (35)
31 to 60	5 (25)
After 60	4 (20)
Other tobacco co-use, n (%)	
Cigar	1 (5)
Electronic cigarette	1 (5)
Digital literacy, mean (SD)	62.2 (20.7)
Median	70.5
Range	16-80
Sensory impairments, n (%)	
Visual or hearing impairment	7 (35)
Physical limitation, n (%)	
Limited in everyday activities	10 (50)
Type of impairment or health problem, n (%)	

XSL•FO
RenderX

Characteristics	Value
Back or neck problem	4 (20)
Lung or breathing problem	2 (10)
Cancer	2 (10)
Fractures, or bone or joint problem	1 (5)
Walking problem	1 (5)

Digital Cessation Treatments (All Modalities)

Overview

Thematic analysis identified 3 meaningful themes found across all digital modalities (app-delivered, texting-based, or videoconferencing counseling): convenience, accessibility, and personalization (Figure 1).

Figure 1. Qualitative themes regarding expectations and preferences for digital health modalities.

Accessibility

An expected benefit of digital platforms was easy access to treatment. Many participants did not have a car or were not able to drive independently and felt burdensome to friends and family who regularly provided transportation. Others found it inconvenient to arrange transportation (eg, public transit or car service) for themselves.

> *I don't think I would come in person. That's a lot of work if you want to see me once a week. It's not like I can just get in the car and drive. I'd have to call transportation, call three days ahead, you know. Stuff like that. So, it's a lot of work.* [Aged 70 years, Black, woman, 3 cigarettes per day (reduced cigarettes per day in the past month), digital literacy 73/80]

Because of these barriers, many discussed prioritizing in-person appointments only when necessary for medical procedures (eg, bloodwork). Thus, a benefit of behavioral treatment is that it does not require in-person services for physical procedures (such as bloodwork or urinalysis). In fact, some discussed interest in behavioral cessation support only if offered through digital platforms.

> *I prefer not to try to travel in areas that I don't know. So, any place I go my husband basically drives me. I would not have done in-person only because that just puts more drain on my husband.* [Aged 67 years, White, woman, 4 cigarettes per day, digital literacy 18/80]

The physical act of traveling and sitting in waiting room chairs was painful for some individuals with physical impairments or chronic pain. Physical limitations (eg, breathing challenges or mobility impairments) were discussed as making in-person treatment uncomfortable and less preferred compared to digital health.

> *There's the wait...the lost time [of in-person treatment]. And the chance of germs, not that I'm a germaphobe...and then the chairs are not really comfortable for long-term waiting...Sometimes I tell the receptionist I'm going out to sit in the car because my car seat is more comfortable than the chairs* [Aged 70 years, White, woman, 28 cigarettes per day, digital literacy 32/80]

Convenience

Another expected benefit of digital treatment was its convenience. Interviewees felt positive about the ability to access treatment when it was best for their schedule and avoiding in-person inconveniences (eg, parking). Participants appreciated the opportunity to wear comfortable clothes or be in the comfort of their homes when receiving cessation treatment. App-delivered and texting-based modalities were considered the most convenient (compared to videoconferencing) because they did not require scheduling an appointment with a provider. Participants liked that these modalities offered access to treatment 24/7.

> *I think an app would be more convenient than anything else. And you can use it when you needed.* [Aged 73 years, White, man, 20 cigarettes per day, digital literacy 78/80]

XSL•FO
RenderX

Personalization

Across all modalities, participants discussed a preference for treatment that was personalized to their individual experience with quitting. Concerns for app-delivered and text-based treatments were that these interventions might feel automated, impersonal, and not tailored to their unique quitting needs.

> *Automated is...it makes you feel, or you don't feel like there's a personal touch there.* [Aged 70 years, White, woman, 7 cigarettes per day, digital literacy 79/80]

Interviewees preferred individualized quit plans that could address their own craving triggers. Specifically for app-delivered interventions, suggestions included tailored educational materials for their age group and personalization features (eg, notifications during self-reported craving times or customized educational materials for their reasons for continued smoking).

> *If I get anxious, or sad over something, I'll get a cigarette for my head to relax. So, I would want an app that would relate to those reasons.* [Aged 70, White, woman, 7 cigarettes per day, digital literacy 79/80]

For texting-based interventions, this sample preferred texting with a live counselor rather than receiving automated quitting tips. Participants believed automated messaging would likely be unhelpful given their long-term familiarity with standard advice to quit. Further, 1 participant who engaged with a texting-based cessation service discussed disliking the automated text messages.

> *The problem I'm having now is that I have a texting thing with [program name redacted] but it's all automated. You know they'll ask you a question and it's just uh...a question that you should think of and if you try to respond to it, it just bounces back you know...it's an automated system. Not personalized.* [Aged 68 years, White, man, recently quit, digital literacy 59/80]

Videoconference counseling was considered the most personalized form of treatment. Compared to telephone counseling, most preferred videoconferencing. Since many had had videoconferencing with their medical providers, they felt positive about seeing a provider's facial expressions during the conversation. They believed videoconferencing facilitated a better relationship with their provider and thus would provide higher quality treatment (compared to telephone counseling).

> *I would rather video...where I can see someone on the computer on video. I just, it gives you, I guess you get more from the visual aspect of a person ...in my work, before I retired, I was [delivering] veterans counseling...so I depend upon visual cues as much as I do by voice.* [Aged 69 years, Black, man, 5 cigarettes per day, digital literacy 79/80]

App-Delivered Interventions

Overview

Notably, 2 themes unique to app-delivered interventions were identified: novelty and ease of use (Figure 1). Only 1 participant had ever used an app-delivered cessation intervention; however, the majority (n=17, 85%) reported interest in using this type of treatment. Most (n=18, 90%) had used non–cessation-focused apps for a wide variety of reasons including social or communication, navigation, organization or calendar, games and entertainment, finances, shopping, and news. Only 2 interviewees were not currently using apps, both of whom had digital literacy scores <20 (out of 80) and did not own smartphones. However, both reported interest in this type of treatment.

Novelty

Many were unfamiliar with and curious about how apps could facilitate smoking cessation. Although this population used a variety of apps for other reasons (eg, entertainment, news, or navigation), fewer had used apps specifically for health reasons (eg, tracking steps). Participants were more familiar with videoconferencing and text messaging for cessation, and thus app-delivered treatments were the most unknown cessation modality. Aside from 1 participant who had used a smoking cessation app, no one had used an app-delivered treatment for any physical or mental health problem. Yet, interviewees were open to trying this novel treatment and evaluating its usefulness.

> *I think that's interesting...um I mean I'd be curious about it. That doesn't mean I'd be ongoing using an app...but I'd be curious if that makes sense. If I find it to be helpful, I'd find that very interesting for me to figure out.* [Aged 76 years, White, woman, 10 cigarettes per day, digital literacy 71/80]

Participants commonly described frustration with their inability to quit cigarettes, and an eagerness to try any means for quitting. An expected benefit of app-delivered treatments was that they offered a novel strategy for treatment.

An app was most appealing if offering novel resources beyond standard advice to quit. Most participants described long-term familiarity with standard education about the harms of smoking. Individuals were more open to trying strategies or educational materials beyond this standard content.

> *If it's going to be the same tips that you get "uh smoking is bad for your health, smoking causes cancer", et cetera, et cetera. I would want it to be something other than the routine that I've seen for the last 40 years* [Aged 70 years, White, woman, 7 cigarettes per day, digital literacy 79/80]

Ease of Use

Participants were concerned that an app-delivered treatment might be overly complicated and challenging to navigate. Interviewees referenced frustrating experiences with apps that contained too much content, did not function properly, or were difficult to understand. Some discussed enlarging font size on apps that they used for noncessation reasons. Many said they would not have the patience to engage with an app that had an overwhelming amount of content, tabs, or buttons to click.

> *Well, if it's a complicated app, most of us that are 75 years old are going to have difficulty maneuvering through it. But if it's something real simple, like going*

XSL•FO
RenderX

> *[to] the next page, next page, or click on this to learn this, you know...uh depends on how you maneuver through the app.* [Aged 75 years, White, woman, 18 cigarettes per day, digital literacy 47/80]

Texting-Based Interventions

Overview

Notably, 1 unique theme emerged for text-based interventions: inefficiency (Figure 1). All participants had at least some experience with texting before this study; however, comfort with and interest in texting ranged widely. Only 1 participant had used a texting-based cessation intervention before this study (a different participant than the one who used an app-delivered treatment). Half (n=10, 50%) said they would be interested in this type of treatment.

Inefficiency

A negative expectation of texting-based interventions was that the action of sending text messages would be inefficient for communication and treatment engagement. Many believed that they texted more slowly and with more difficulty than the average person, thus making texting frustrating. Texting was discussed as necessary for brief communication (eg, scheduling); however, sending text messages was not an enjoyable activity. Texting was considered burdensome and an unappealing option for treatment engagement.

> *I try to stay as far away from texting as I can. I'm a hunt and peck typer...I don't text a lot...I guess I can do it when I have to. But I'm not that big for it. I try not to text. I'll call you first.* [Aged 69 years, Black, man, 5 cigarettes per day, digital literacy 79/80]

Videoconferencing With Cessation Counselor

Overview

Notably, 3 unique themes for videoconferencing counseling were identified: presentability, security and privacy, and technological difficulties (Figure 1). Compared to other digital modalities, participants were most familiar with videoconferencing. More than half (n=12, 60%) had a history of videoconferencing with their medical providers. Many discussed using videoconferencing platforms for work-related or social activities. Only 1 participant had used videoconferencing for cessation counseling (a different individual from the 2 participants who had used an app-delivered and texting-based cessation treatment). Most (n=17, 85%) said they would be interested in this type of treatment.

Presentability

When videoconferencing (rather than telephone counseling or another digital modality), a negative expectation was the time and effort for on-screen presentability (eg, getting dressed or applying makeup). Although participants felt that visibly seeing their counselor would improve the quality of their cessation care, presentability was an expected barrier (depending on their time and level of motivation for the day). Interviewees also discussed feeling self-conscious when showing their face on videoconferencing screens. For app-delivered or texting-based treatments, interviewees appreciated not needing to shower or change clothes to engage with content.

> *As far as the Skype-in' and ...I don't like to do it because I'm very self-conscious of all my wrinkles. Um, and I look horrible on those screens.* [Aged 75 years, White, woman, 18 cigarettes per day, digital literacy 47/80]

Privacy and Security

A concern was that personal information shared in videoconferencing sessions could be overheard (in either the patient or provider's environment) and that counselor notes might be shared. Participants said they wanted to know where counseling notes were stored and if they would be shared with other providers. Interviewees appreciated that app-delivered or texting-based treatments might not require them to disclose personal information (eg, last name or address) and were more anonymous than videoconferencing.

> *Security. You just hear about it all the time...using technology I'm just kind of leery of it...personal information getting out. Well, they ask about your personal information, social security questions etcetera, etcetera, I would worry about this immensely.* [Aged 73 years, White, man, 20 cigarettes per day, digital literacy 78/80]

Technological Difficulties

An expected challenge was navigating technology in the setup or duration of a videoconferencing appointment. Interviewees believed technology-based difficulties would be less common with app-delivered or text-based modalities. Many had prior challenges setting up videoconferencing sessions with their medical providers or had internet connection issues during the visit. Technology-based issues were frustrating and negatively impacted the quality of care and overall experience.

> *I was trying to log into the network, and the doctor's office, they had um a link and I kept trying to get it. But evidently either I was doing something wrong, or whatever it just never worked so I kind of gave up on that...you know, maybe it was just new to me, or new to them. It just didn't work.* [Aged 69 years, Black, man, 5 cigarettes per day, digital literacy 79/80]

Digital Health Treatments With Social Components

Overview

Notably, 3 themes emerged regarding digital cessation treatments that integrated social components: helpful for quitting, connection, and interpersonal challenges (Figure 1). No participant reported prior experience with digital cessation treatments that integrated social components.

Helpful for Quitting

Some interviewees had personal experiences with in-person group counseling for other mental or physical health problems. Others had friends or family who had benefited from group counseling. Referencing these experiences, interviewees believed that listening to others quitting might be helpful for their own quit success. For app-delivered treatments, participants

XSL•FO
RenderX

commonly discussed that the main benefit of a social forum would be to learn new strategies and tips for trying to quit.

> *Never thought about that, but that might be helpful...to hear other people...the ones that are dealing with the problem...and their tried solutions...and failed solutions and why they think it worked and why they didn't think it worked and how that'd compare to my life you know* [Aged 70 years, White, woman, 28 cigarettes per day, digital literacy 32/80]

Connection

An expected benefit was the ability to build connections with others in cessation treatment. For group videoconferencing (rather than app-delivered treatments with social forums), interviewees appreciated the idea of meeting others with the same goals. Participants mentioned that others might provide accountability and encouragement during their quit attempts.

> *I know it's essential...to stop smoking its essential to network...okay and I know that because of AA and NA and the way all those 12-step programs work...there's [only] one part of it [that] is the knowledge, if you will, and then it's the fellowship, the interacting with other people trying to quit...it has an equally if not more value in um...of me facing myself...let me see if I can explain it...that I call someone before I smoke that cigarette that holds me accountable and if I reach that level of accountability with cigarettes then I might have a chance of putting em' down and I don't think there's any replacement for that personal interaction between two people trying to accomplish the same thing* [Aged 69 years, White, man, recently reducing cigarette use (down to 1 cigarette per day), digital literacy 69/80]

Interpersonal Challenges

Interviewees were concerned about potential interpersonal challenges. They felt negatively about not having a chance to speak (in group videoconferencing) if others were too talkative. Regarding social forums in apps, participants did not want their personal information shared in the case that another patient might contact them outside of the platform. Interviewees were also concerned that the anonymity of app-delivered social forums (vs group videoconferencing) would allow for negative comments and arguments.

> *I would not participate. And I mean...things unfortunately on social media turn ugly so quick it's unbelievable. Now, I do not post anything or...well rarely do I post anything on my [social media] page. I do have a business [social media] page that I post on and I run social media advertising off of. But I mean, the ugliness of social media just blows my mind...so that's just a real turn off for me...Because they're not sitting there next to the person. It's easy to sit behind a keyboard and blast somebody than to be sitting next to them* [Aged 68 years, White, man, recently quit, digital literacy 80/80]

Discussion

Principal Findings

This multimethods study explored the expectations and preferences for digital tobacco treatment among a population of older adults currently smoking or who had recently quit cigarettes. Only 3 participants had engaged with a digital tobacco cessation treatment before this study, with most unaware of these types of treatments. Technology use for non–cessation-related reasons was common (ie, texting, apps, or videoconferencing with medical providers). Digital literacy varied widely, yet the majority had relatively high scores (median 70.5, range 16-80 out of 80). Most (n=17, 85%) were interested in either an app-delivered treatment or videoconferencing counseling, and half (n=10, 50%) would try a texting-based treatment. Even individuals with lower digital literacy and who did not own smartphones were interested in digital cessation treatments. These results challenge the bias that older adults are uninterested in digital health treatments [26]. Thematic analysis identified 3 meaningful themes across all digital modalities (app-delivered, text-based, or videoconferencing counseling): convenience, accessibility, and personalization. Compared to in-person treatment, digital treatment was regarded as more accessible to those with physical impairments and without transportation. This sample felt positively about accessing treatment from their homes and avoiding in-person inconveniences (eg, parking). Yet, digital health treatments that relied on automated messaging, used a "one size fits all" approach to treatment, or delivered only standard advice to quit were unappealing. Personalized content was consistently preferred for digital treatment. This preference is consistent with other age groups, in which tailored content (eg, scheduling personal quit date) in app-delivered cessation treatment is related to app popularity and user-rated quality [34].

App-delivered treatments for cessation were appealing by offering a novel and unfamiliar behavioral strategy for quitting. This sample commonly had a long history of quit attempts and was overly familiar with the harms of smoking. Perhaps desirable for this age group, apps can provide novel and personalized strategies (eg, triggers or reasons for quitting) and content (eg, quitting in older age) beyond standard education [16]. Other ways to personalize content to experiences more common among this age group might include acknowledgment of long histories of quit attempts, low self-efficacy, negative experiences with cessation pharmacotherapy, and fatigue with standard advice. Consistent with the literature detailing aging-related considerations for digital health [35], app design suggestions included simplicity, easy navigation, and enlarged font size. This sample preferred less texting-based engagement within apps, which might be indicative of dexterity barriers for digital health treatment among this age group (eg, aging-related motor decline or stiffer joints) [35]. Thus, previous challenges with using apps were memorable and informed their preferences for future treatment. Despite usability concerns, participants were open and curious about using this modality for cessation. These findings are consistent with a pilot trial, in which older adults found a standard (not tailored for the unique needs of

XSL•FO
RenderX

older adults) cessation app to be moderately acceptable (median 3 out of 5 stars) [36]. Potential ways to further increase the acceptability of app-delivered treatments include prioritizing personalized content and design ease of use.

Texting-based interventions were the least preferred digital modality, perhaps given disinterest in the action of sending text messages. All participants had texted before this study but many found it to be inefficient for communication and unenjoyable. Participants were also uninterested in interventions that relied on automated messaging and content. These findings are inconsistent with a prior feasibility trial, in which 57% of older adults (aged 60+ years) found a text-based cessation intervention useful [37]. Notably, this text-based intervention was tailored to the individual's harm reduction progress and did not heavily rely on participant texting for engagement [37]. However, these results suggest that older adults might initially prefer other digital modalities for cessation.

Expectations for videoconferencing counseling were largely informed by past experiences videoconferencing with medical providers. Beyond the general appeal of digital health platforms, participants believed this treatment modality would provide the most personalized content. Expected challenges included navigating technological difficulties, lack of privacy or security, and appearing presentable onscreen. Despite these concerns, videoconferencing was preferred over telephone counseling given its potential to provide better quality of care. Although the literature is limited on this topic [38], recent evidence suggests that cessation videoconferencing might be more effective than telephone counseling for increasing short-term quit success [39]. Cessation counselors should consider offering videoconferencing for older patients, which might be a more appealing form of treatment for this age group.

No participant had experience with digital cessation treatments incorporating social components (group-based counseling or social forums). Expected benefits included usefulness for quitting (eg, encouragement and accountability or new tips for quitting) and social connection. Yet, participants were concerned about possible interpersonal challenges (eg, negativity in social forums). Digital social forums might benefit from moderated or asynchronous messaging for this age group. Many had positive experiences with addiction treatment (eg, Alcoholics Anonymous) and were thus open to group-based videoconferencing for cessation. Overall, because this population discussed disinterest with standard advice to quit, cessation treatment integrating social components might be novel and appealing to this age group.

Beyond disinterest in standard education and nontailored automated content, participants did not discuss specific cessation treatment components or topics that they would prefer within digital interventions. Participants might have been unable to identify preferences for more detailed content for a few reasons. First, the purpose of this study was not to elicit feedback on any existing cessation program, but rather to gain hypothesis-generating information about expectations and preferences for digital cessation interventions broadly. Thus, participants were not given any descriptions of existing treatments for which to provide feedback. Further, most of this sample had limited prior experiences with digital cessation programs and might have had difficulty envisioning these types of treatments. When developing digital cessation treatments, it will be important for future researchers to elicit more detailed and specific feedback from this population to guide the refinement process.

Results should be interpreted with study limitations in mind. This sample was recruited from an academic medical center given that older adults have much higher rates of health care use than their younger counterparts [40] and thus medical settings can play an important role in reaching older adults who smoke cigarettes. However, findings might not be generalizable to other geographic locations or populations without access to medical services (n=20, 100% of this population had health insurance). Although digital literacy varied, the majority had relatively high scores. Further, this study required 1-time access to the internet for eConsent. Although 2 participants borrowed mobile devices from family members for eConsent (as owning a smartphone or regular access to the internet was not required for participation), results might be most representative of older adults with access to technology and greater digital literacy. Even though men have higher smoking rates than women [7], our sample included 60% (n=12) men and thus might be less generalizable to women. Additionally, this study was not inclusive of all types of digital modalities. Future studies should elicit feedback regarding other digital treatments (eg, remote carbon monoxide sensors), particularly as this field continues to grow and adapt to new technologies. Finally, it will be important to ascertain the best means by which to refer older adults to digital cessation programs, as well as the preferred devices (eg, tablets as compared to desktop computers). For example, older adults might be even more likely to engage in digital programs if their medical providers refer them to these treatments. Future studies should qualitatively and systematically evaluate preferences for digital health treatments among this heterogeneous age group within large diverse samples (by gender, race, ethnicity, or socioeconomic status) and might consider eliciting feedback on existing freely available digital cessation programs.

Conclusions

Older adults are the most impacted by tobacco morbidity and mortality and, therefore, should be prioritized in the growing field of digital cessation treatment. Given a long history of quit attempts and familiarity with standard advice to quit, digital health treatments might offer appealing new behavioral approaches to quitting for this age group. This sample was interested and willing to engage with digital platforms, expecting them to be more accessible and convenient than in-person treatment. Preferences included simple and easy-to-navigate digital designs and personalized rather than automated content. Findings challenge the bias that older adults are uninterested or unwilling to engage with digital treatments for behavioral health. Clinicians and researchers should prioritize the inclusivity of older adults in the development and dissemination of digital cessation treatments.

XSL•FO
RenderX

Acknowledgments

This research is funded by the National Institute on Drug Abuse (T32DA00788, T32DA019426, and K23DA045766) and discretionary funds (principal investigator MJC). The REDCap (Research Electronic Data Capture) hosted by the South Carolina Clinical and Translational Science Institute is supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR001450).

Authors' Contributions

MCF conducted paper conceptualization and literature review, facilitated semistructured interviews and data analysis, and developed the initial draft of this paper. MJC, JD, EW, and BAT contributed to the paper's conceptualization and the study's recruitment procedures. RO, KP, and MRS helped facilitate study recruitment procedures and data analysis. All authors were involved in writing and editing the final paper.

Conflicts of Interest

BAT testifies on behalf of plaintiffs who have filed litigation against the tobacco industry. MJC serves as a paid expert consultant in litigation involving trade disputes involving e-cigarette manufacturers.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File , 15 KB - jmir_v26i1e52919_app1.docx]

References

1. Department of Health and Human Services. 2020 profile of older Americans. The Administration for Community Living. 2021. URL: https://acl.gov/sites/default/files/Profile%20of%20OA/2020ProfileOlderAmericans_RevisedFinal.pdf [accessed 2024-07-11]
2. The health consequences of smoking-50 years of progress: a report of the surgeon general. U.S. Department of Health and Human Services. 2014. URL: https://www.hhs.gov/sites/default/files/consequences-smoking-exec-summary.pdf [accessed 2024-07-11]
3. Andrews JO, Heath J, Graham-Garcia J. Management of tobacco dependence in older adults: using evidence-based strategies. J Gerontol Nurs 2004 Dec;30(12):13-24. [doi: 10.3928/0098-9134-20041201-06] [Medline: 15624692]
4. Rimer BK, Orleans CT, Keintz MK, Cristinzio S, Fleisher L. The older smoker. status, challenges and opportunities for intervention. Chest 1990 Mar;97(3):547-553. [doi: 10.1378/chest.97.3.547] [Medline: 2306958]
5. Johnson AL, Nystrom NC, Piper ME, Cook J, Norton DL, Zuelsdorff M, et al. Cigarette smoking status, cigarette exposure, and duration of abstinence predicting incident dementia and death: a multistate model approach. J Alzheimers Dis 2021;80(3):1013-1023 [FREE Full text] [doi: 10.3233/JAD-201332] [Medline: 33646160]
6. Isenberg JY, Quiñones AR, Slatore CG, Bryson WC, Thielke SM. Trends in cigarette smoking and cessation among Medicare managed care recipients, 2005-2012. Addict Behav 2016 Jul;58:155-160. [doi: 10.1016/j.addbeh.2016.02.037] [Medline: 26946446]
7. Cornelius ME, Loretan CG, Jamal A, Davis Lynn BC, Mayer M, Alcantara IC, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 5;72(18):475-483 [FREE Full text] [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
8. McAfee T, Malone RE, Cataldo J. Ignoring our elders: tobacco control's forgotten health equity issue. Tob Control 2021 Sep;30(5):479-480. [doi: 10.1136/tobaccocontrol-2021-056945] [Medline: 34389689]
9. Kleykamp BA, Kulak JA. Cigarette use among older adults: a forgotten population. Am J Public Health 2023 Jan;113(1):27-29. [doi: 10.2105/AJPH.2022.307151] [Medline: 36413703]
10. Maciosek MV, Xu X, Butani AL, Pechacek TF. Smoking-attributable medical expenditures by age, sex, and smoking status estimated using a relative risk approach. Prev Med 2015 Aug;77:162-167 [FREE Full text] [doi: 10.1016/j.ypmed.2015.05.019] [Medline: 26051203]
11. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. Smoking cessation. a report of the surgeon general. U.S. Department of Health and Human Services. 2020. URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf [accessed 2024-07-11]
12. Taylor DH, Hasselblad V, Henley SJ, Thun MJ, Sloan FA. Benefits of smoking cessation for longevity. Am J Public Health 2002 Jun;92(6):990-996. [doi: 10.2105/ajph.92.6.990] [Medline: 12036794]
13. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults—United States, 2000-2015. MMWR Morb Mortal Wkly Rep 2017 Jan 6;65(52):1457-1464 [FREE Full text] [doi: 10.15585/mmwr.mm6552a1] [Medline: 28056007]

XSL•FO
RenderX

14. Henley SJ, Asman K, Momin B, Gallaway MS, Culp MBB, Ragan KR, et al. Smoking cessation behaviors among older U.S. adults. Prev Med Rep 2019 Aug 17;16:100978 [FREE Full text] [doi: 10.1016/j.pmedr.2019.100978] [Medline: 31660285]
15. Chen D, Wu LT. Smoking cessation interventions for adults aged 50 or older: a systematic review and meta-analysis. Drug Alcohol Depend 2015 Sep 1;154:14-24 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.06.004] [Medline: 26094185]
16. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. Cochrane Database Syst Rev 2019 Oct 22;10(10):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub5] [Medline: 31638271]
17. Holmes J, Powell-Griner E, Lethbridge-Cejku M, Heyman K. Aging differently: physical limitations among adults aged 50 years and over: United States, 2001-2007. NCHS Data Brief 2009 Jul(20):1-8 [FREE Full text] [Medline: 19627658]
18. Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. J Community Health 2013 Oct;38(5):976-993 [FREE Full text] [doi: 10.1007/s10900-013-9681-1] [Medline: 23543372]
19. Kakulla B. Older adults embrace tech for entertainment and day-to-day living. American Association of Retired Persons. 2021 Dec. URL: https://doi.org/10.26419/res.00493.001 [accessed 2021-12-10]
20. Kwan RYC, Salihu D, Lee PH, Tse M, Cheung DSK, Roopsawang I, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020 Apr 21;17:7 [FREE Full text] [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
21. Buyl R, Beogo I, Fobelets M, Deletroz C, van Landuyt P, Dequanter S, et al. e-Health interventions for healthy aging: a systematic review. Syst Rev 2020 Jun 3;9(1):128 [FREE Full text] [doi: 10.1186/s13643-020-01385-8] [Medline: 32493515]
22. Prutzman YM, Wiseman KP, Grady MA, Budenz A, Grenen EG, Vercammen LK, et al. Using digital technologies to reach tobacco users who want to quit: evidence from the National Cancer Institute's Smokefree.gov initiative. Am J Prev Med 2021 Mar;60(3 Suppl 2):S172-S184 [FREE Full text] [doi: 10.1016/j.amepre.2020.08.008] [Medline: 33663705]
23. Fradkin N, Zbikowski SM, Christensen T. Analysis of demographic characteristics of users of a free tobacco cessation smartphone app: observational study. JMIR Public Health Surveill 2022 Mar 9;8(3):e32499 [FREE Full text] [doi: 10.2196/32499] [Medline: 35262491]
24. Chu KH, Matheny SJ, Escobar-Viera CG, Wessel C, Notier AE, Davis EM. Smartphone health apps for tobacco cessation: a systematic review. Addict Behav 2021 Jan;112:106616 [FREE Full text] [doi: 10.1016/j.addbeh.2020.106616] [Medline: 32932102]
25. Bricker JB, Mull KE, McClure JB, Watson NL, Heffner JL. Improving quit rates of web-delivered interventions for smoking cessation: full-scale randomized trial of WebQuit.org versus Smokefree.gov. Addiction 2018 May;113(5):914-923 [FREE Full text] [doi: 10.1111/add.14127] [Medline: 29235186]
26. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022 Dec 30;12(12):1102-1105 [FREE Full text] [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
27. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
28. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
29. Health-related quality of life (HRQOL). Centers for Disease Control and Prevention. 2018. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/hrqol/index.htm [accessed 2023-08-01]
30. Roque NA, Boot WR. A new tool for assessing mobile device proficiency in older adults: the Mobile Device Proficiency Questionnaire. J Appl Gerontol 2018 Feb;37(2):131-156 [FREE Full text] [doi: 10.1177/0733464816642582] [Medline: 27255686]
31. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Braun VC. Volume 2: research designs: quantitative, qualitative, neuropsychological, and biological. In: APA Handbook of Research Methods in Psychology. Washington, DC: America Psychological Association; 2012.
33. Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health Serv Res 2007 Aug;42(4):1758-1772 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00684.x] [Medline: 17286625]
34. Hoeppner BB, Hoeppner SS, Seaboyer L, Schick MR, Wu GWY, Bergman BG, et al. How smart are smartphone apps for smoking cessation? a content analysis. Nicotine Tob Res 2016 May;18(5):1025-1031 [FREE Full text] [doi: 10.1093/ntr/ntv117] [Medline: 26045249]
35. Boot W, Charness N, Czaja SJ, Rogers WA. Assessing needs with older adults. In: Designing for Older Adults: Case Studies, Methods, and Tools. Boca Raton, FL: CRC Press; 2020.

XSL•FO
RenderX

36. Peek J, Hay K, Hughes P, Kostellar A, Kumar S, Bhikoo Z, et al. Feasibility and acceptability of a smoking cessation smartphone app (My QuitBuddy) in older persons: pilot randomized controlled trial. JMIR Form Res 2021 Apr 14;5(4):e24976 [FREE Full text] [doi: 10.2196/24976] [Medline: 33851923]
37. Noonan D, Silva S, Njuru J, Bishop T, Fish LJ, Simmons LA, et al. Feasibility of a text-based smoking cessation intervention in rural older adults. Health Educ Res 2018 Feb 1;33(1):81-88 [FREE Full text] [doi: 10.1093/her/cyx080] [Medline: 29309599]
38. Tzelepis F, Paul CL, Williams CM, Gilligan C, Regan T, Daly J, et al. Real-time video counselling for smoking cessation. Cochrane Database Syst Rev 2019 Oct 29;2019(10):CD012659 [FREE Full text] [doi: 10.1002/14651858.CD012659.pub2] [Medline: 31684699]
39. Byaruhanga J, Paul CL, Wiggers J, Byrnes E, Mitchell A, Lecathelinais C, et al. The short-term effectiveness of real-time video counselling on smoking cessation among residents in rural and remote areas: an interim analysis of a randomised trial. J Subst Abuse Treat 2021 Dec;131:108448. [doi: 10.1016/j.jsat.2021.108448] [Medline: 34098302]
40. Institute of Medicine (US) Committee on the Future Health Care Workforce for Older Americans. Retooling for an Aging America: Building the Health Care Workforce. Washington, DC: The National Academies Press; 2008.

Abbreviations

eConsent: electronic consent
IRB: Institutional Review Board
MUSC: Medical University of South Carolina
NA: Narcotics Anonymous
REDCap: Research Electronic Data Capture

Edited by T de Azevedo Cardoso; submitted 19.09.23; peer-reviewed by S Zbikowski, S Lee; comments to author 27.03.24; revised version received 05.04.24; accepted 25.06.24; published 28.08.24.

Please cite as:
Fahey MC, Carpenter MJ, O'Neal R, Pebley K, Schick MR, Ware E, Toll BA, Dahne J
Expectations and Preferences for Digital Cessation Treatment: Multimethods Study Among Older Adults Who Smoke Cigarettes
J Med Internet Res 2024;26:e52919
URL: https://www.jmir.org/2024/1/e52919
doi:10.2196/52919
PMID:39196628

©Margaret C Fahey, Mathew J Carpenter, Riley O'Neal, Kinsey Pebley, Melissa R Schick, Emily Ware, Benjamin A Toll, Jennifer Dahne. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sociodemographic, Health, and Lifestyle-Related Characteristics Associated With the Commencement and Completion of a Web-Based Lifestyle Educational Program for People With Multiple Sclerosis: Randomized Controlled Trial

Jeanette Reece[1], MPH, PhD; Maggie Yu[1], MPSYD, MPH; William Bevens[2], PhD; Steve Simpson-Yap[1,3,4], MPH, PhD; Rebekah Davenport[5], MCP; George Jelinek[1], MD; Sandra Neate[1], MBBS

[1]Neuroepidemiology Unit, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia
[2]Institute for Clinical and Translational Science, University of California, Irvine, CA, United States
[3]The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Melbourne, Australia
[4]Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia
[5]Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia

Corresponding Author:
William Bevens, PhD
Institute for Clinical and Translational Science
University of California
843 Health Science Rd
1301 Hewitt Hall
Irvine, CA, 9269
United States
Phone: 1 5103382170
Email: wbevens@health.ucsd.edu

Abstract

Background: Digital health interventions increase access to multiple sclerosis (MS)–related knowledge for people living with MS; however, our understanding of factors associated with engagement in web-based learning is limited.

Objective: This study aims to examine associations between participant sociodemographic, health, and lifestyle-related characteristics and the commencement and completion of the Multiple Sclerosis Online Course (MSOC) in a randomized controlled trial (RCT).

Methods: An intervention course was developed based on the Overcoming MS Program—an evidence-based lifestyle modification program for MS, and a standard care course was developed based on international MS website information. An RCT was conducted to compare the effectiveness of the intervention course versus the standard care course in improving health outcomes in people living with MS. Participant data were collected from a baseline survey. Associations between baseline participant characteristics and MSOC commencement and completion, respectively, were assessed using multivariate log-binomial regression.

Results: Overall, 1893 participants enrolled in the RCT, and 45.27% (n=857) completed the baseline survey: 23.5% (n=444) in the intervention course and 21.8% (n=413) in the standard care course. Of these 857 participants, 631 (73.6%) commenced the standard care course or intervention course, and 49.1% (218/444) and 54.2% (224/413) completed the intervention course and standard care course, respectively. University education, partnered relationship status, and higher mental and physical quality of life were associated with 19%, 12%, 20%, and 22% higher rates of course commencement, respectively. Clinically significant fatigue was associated with a 10% reduction in the likelihood of commencement. Strongest associations with intervention course completion included middle and older adulthood, male sex, fatigue, and preexisting adherence to a diet program, with 96%, 27%, 24%, and 19% higher rates of completion observed, respectively, whereas higher self-efficacy was associated with up to 35% lower intervention course completion. Associations with standard care course completion included practicing meditation (20% higher completion), whereas employment was associated with 22% lower completion.

Conclusions: Sociodemographic and clinical factors, as well as lifestyle-related factors, were important factors in MSOC commencement and completion. These data may help guide the design and enhancement of digital health interventions tailored for people living with MS.

XSL•FO
RenderX

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12621001605886; https://tinyurl.com/2vyve9p9
International Registered Report Identifier (IRRID): RR2-10.1186/s12883-023-03298-0

(J Med Internet Res 2024;26:e58253) doi:10.2196/58253

KEYWORDS
multiple sclerosis; web-based education; lifestyle; randomized controlled trial; engagement; completion

Introduction

Background

There is a growing body of evidence supporting the role of modification of lifestyle-related risk factors on enhancing health outcomes in people living with multiple sclerosis (MS) [1-3]. Subsequently, the management of MS has evolved over recent years, with people living with MS placing a greater emphasis on self-management of MS and seeking lifestyle-related information as an integral part of maintaining and improving their health [4].

MS is a progressive autoimmune condition manifesting in sensory, motor, and cognitive dysfunction, the cardinal clinical elements of which are disability progression and relapse (as defined by a clinical exacerbation of symptoms involving the development of 1 or more new MS symptoms or worsening of existing symptoms lasting >48 hours, with changes in symptoms not due to extraneous conditions such as heat or illness, eg, respiratory or urinary tract infections) [5]. Modification of lifestyle-related risk factors in MS has been associated with a lower rate of relapse, reduced fatigue, disability progression, and depressive symptoms and higher quality of life (QoL) [6-12]. A program for the modification of lifestyle-related risk factors in MS, the Overcoming Multiple Sclerosis (OMS) Program [13], provides recommendations for a plant-based whole food plus seafood diet low in saturated fat, regular physical activity, vitamin D and omega-3 fatty acid supplementation, smoking cessation, and stress reduction. These recommendations have been delivered as a face-to-face educational intervention with demonstrated effectiveness, including adherence to lifestyle recommendations (a healthy diet, meditation, and vitamin D and omega-3 supplementation) 3 years after attending the retreat and associations with increased QoL in both the short (1 year) and medium (5 years) term [14-16]. However, the reliance on face-to-face delivery hinders the scalability and accessibility of the educational intervention for all people living with MS. Web-based lifestyle modification programs offer a scalable solution, overcoming commonly documented barriers such as mobility to travel outside of the home and additional financial costs [17].

Several web-based lifestyle modification interventions have been developed for people living with MS to increase MS-related knowledge [18], build resilience [19], and facilitate lifestyle changes [20]. Many have demonstrable effectiveness, such as reduced depression and fatigue [21,22]. Furthermore, interventions have been shown to improve walking ability; reduce overall neurological disability [20]; and improve depression, anxiety, and sleep [23]. Recently, the *UnderstandingMS* massive open online course (MOOC) led to increases in MS-related knowledge and health literacy [18] and lifestyle changes (diet, physical activity, and vitamin D supplementation) in people living with MS [24]. More recent studies have examined the usability of web-based education interventions, with an intervention aimed specifically at facilitating multimodal behavior change in people living with MS with moderate to severe disability found to be both practical and acceptable [25]. Similarly, a web-based nutrition education program demonstrated good acceptability using a co-design development model to incorporate the needs of people living with MS themselves [26]. As such, this nascent field of digital health for MS-related lifestyle modification has emerged as an important potential tool to support people living with MS.

Despite the potential value of digital interventions, engagement with web-based interventions compared to face-to-face interventions is generally low [27]. Course completion rates vary, with one study demonstrating that as few as 15% to 19% of people living with MS enrolled in digital health programs completed the course [28], whereas other studies have described completion rates that differ from these estimates [29]. While there has been broad implementation and acceptance in the general population, there has been limited exploration of how people living with MS use digital health technologies to support their health and well-being. There is previous evidence suggesting modest variability in the completion of digital health interventions by people living with MS [28,30]; however, factors that affect completion remain unclear. There is other evidence suggesting that the use of digital technologies varies significantly by sociodemographic factors and depending on what the technology is used for [31]. We also previously reported that sourcing lifestyle information on the web is precarious for people living with MS because critically appraising information can be difficult without professional assistance from trustworthy sources [32]. In this context, programmatic and structured web-based educational interventions developed by medical professionals and researchers may be an effective way to engage people living with MS with lifestyle modification information. An important next step is to explore factors affecting the completion of digital health interventions for people living with MS to ensure that future developments can meet their needs and deliver education effectively.

Objectives

This study aimed to identify the characteristics of people living with MS associated with the commencement and completion of the Multiple Sclerosis Online Course (MSOC) by examining sociodemographic, health, and lifestyle-related characteristics of participants enrolled in a randomized controlled trial (RCT) [33]. Study findings have the potential to increase our understanding of the barriers and enablers of the completion of

XSL•FO
RenderX

web-based, lifestyle-related interventions by people living with MS, providing a basis for the development of future digital interventions.

Methods

Ethical Considerations

This ancillary RCT is a CONSORT-R (Consolidated Standards of Reporting Trials–Routine)–compliant RCT (Multimedia Appendix 1). The protocol for the RCT was reviewed by the Australian New Zealand Clinical Trials Registry and approved on November 25, 2021 (ACTRN12621001605886). The study was reviewed by the University of Melbourne Human Research Ethics Committee and approved on November 2, 2021 (22140). Participants were invited via web-based platforms to participate in the RCT and did not receive any monetary compensation for participating in the RCT. Participants were provided with a participant information statement, and written informed consent was obtained from all participants for their data to be used for research purposes before inclusion in the RCT. The signed consent form outlined that the confidentiality of their data would be ensured as per safeguard legal requirements. For analyses and reporting, all participant data were stored in a reidentifiable format to ensure participants' privacy and confidentiality at the University of Melbourne in the form of password-protected computer databases, and only the listed investigators had access to the data.

MSOC RCT Study Design

The primary aims of the ancillary RCT (the MSOC effectiveness RCT) included examining (1) changes in the health-related QoL of people living with MS from baseline to the 6-, 12-, and 30-month follow-ups (primary outcome); and (2) changes in other health outcomes (depression, anxiety, fatigue, and disability) between baseline and the 6-, 12-, and 30-month follow-ups (secondary outcome) [33]. The secondary aims included examining changes in lifestyle from baseline to the 6-, 12-, and 30-month follow-ups. This study represents an additional analysis of baseline data collected as part of the flagship RCT and aimed to examine factors associated with commencement and completion of the MSOC. All baseline data were collected from June 2022 to July 2023. Data analysis was performed at the completion of all baseline data collection from November 2023 to December 2023.

People living with MS were invited to participate in the RCT and complete the MSOC. Data from the following participants were excluded from all analyses: (1) participants experiencing any serious comorbid chronic illness or neurological illness or injury other than MS that would threaten regular participation or significantly affect the outcome measures in its own right, such as motor neuron disease or stroke, as determined by the study investigators; and (2) participants currently taking part in another RCT.

While we did not notify participants that their data would be used specifically for this study, participants were aware that we were examining factors associated with commencement and completion of the web-based intervention outlined in the postcourse evaluation questionnaire sent to all participants upon completion of the course as per the study protocol [33].

MSOC Effectiveness RCT Study Design

The MSOC effectiveness RCT has previously been described in detail [33]. In brief, the RCT aimed to assess the effectiveness of a 6-week intervention course in improving QoL and health outcomes in people living with MS compared with a 6-week standard care course. The intervention course modules provided content adapted from the OMS evidence-based lifestyle modification program (Multimedia Appendix 2) [13]. The standard care course contained standard health recommendations sourced from international public MS society websites that aimed to reflect standard information provided by health care practitioners and MS societies.

Both courses comprised 7 modules, commencing with a *Welcome to the MSOC Study* module containing a plain-language statement and baseline survey followed by five educational modules: (1) *Introduction*, (2) *Eat well*, (3) *Sunlight and vitamin D*, (4) *Exercise*, (5) *Meditation and the mind-body connection*, (6) *Medication and family prevention*, and (7) *Conclusion*. In total, 2 modules were released each week over a 4-week period, and a further 2 weeks were provided for course completion. The feasibility of the intervention course and standard care course at delivering educational content to people living with MS has been previously demonstrated [34].

The modules of the intervention course and standard care course mirrored one another in format and style of delivery of content in terms of a combination of videos, animations, visuals, and discussions from presenters. Key differences included the intervention course's focus on specific lifestyle recommendations supplemented by video discussions, illustrations, and web resources for in-depth exploration. For instance, the intervention course advised a specific plant-based whole food diet plus seafood with very low saturated fat (<20 g/d) excluding dairy, meat, and palm and coconut oil, complemented with recipes and video discussions on selecting ingredients and adherence tips. Furthermore, for daily meditation, the intervention course offered practical video guidelines for practices of ≥30 minutes per day. In contrast, the standard care course provided general advice without detailed guidance; that is, the standard care course offered broader advice, recommending a balanced diet based on national guidelines and mentioning meditation without providing solid evidence of its efficacy in MS management.

Participants

Participants were recruited on the web via peer support Facebook groups worldwide and MS societies in Australia, Canada, Ireland, New Zealand, and the United Kingdom. Recruitment flyers were posted on Facebook, Twitter, and Instagram. Interested participants completed 2 eligibility questions at the study website confirming that they were aged ≥18 years and had received a physician-confirmed diagnosis of MS. Participants were required to speak English to be able to understand the course content.

Eligible persons were sent a link to set up an account and log in to the course platform. Participants were then allocated to

XSL•FO
RenderX

the intervention course or standard care course at a 1:1 ratio using simple randomization. Participants were also requested to complete a 166-question baseline survey on sociodemographic, health, and lifestyle-related factors. If participants did not complete the baseline survey, they were sent 2 email reminders to complete it, but this did not prevent access to the web-based course, which was provided to all enrolled participants regardless of whether they completed the baseline survey. However, only data from participants who completed the baseline survey were included in this study to address study objectives. This comprised data from 857 participant baseline surveys collected during the 5 rounds of RCT recruitment that ran between June 23, 2022, and September 4, 2023. Each of the 5 recruitment rounds used the same strategy to recruit participants; that is, recruitment for the RCT involved advertising on international MS websites and MS-related social media sites such as Facebook and Instagram.

Sociodemographic Characteristics

Data on age, sex, employment, level of education, and marital status were collected. Specifically, we queried the highest level of education (no formal schooling, primary school, secondary school, vocational training, bachelor's degree, or postgraduate degree), marital status (married; cohabitating or partnered; separated, widowed, or divorced; or single), and current work status (10 categories ranging from working full time to retired due to medical reasons or disability and work status not clearly defined). Perceived social support was measured using the 12-item Multidimensional Scale of Perceived Social Support survey [35]. A summary Multidimensional Scale of Perceived Social Support score was calculated, with higher scores indicating higher perceived support.

Health Characteristics

Height (centimeters or inches) and weight (kilograms or pounds) were used to calculate BMI, categorized as underweight (<18.5 kg/m^2), normal (18.5-24.9 kg/m^2), overweight (25.0-29.9 kg/m^2), and obese (≥30.0 kg/m^2) as per World Health Organization guidelines [36]. MS type was categorized into nonprogressive (benign or relapsing-remitting MS) and progressive (primary progressive, secondary progressive, or progressive-relapsing MS). MS duration was calculated using the year of diagnosis and baseline survey date. The number of treated comorbidities was queried using the Self-Administered Comorbidity Questionnaire [37] and categorized as 0 and ≥1. Ongoing symptoms from relapse within ≤30 days were queried (yes or no). The use of disease-modifying therapies (DMTs) was queried (yes or no), and if yes, the type of DMT was queried.

Disability was assessed using Patient-Determined Disease Steps scale scored ordinally from 0 (normal) to 8 (bed bound) [38,39] and categorized into none or mild (0-2), moderate (3-5), and severe (6-8) disability. Fatigue was measured using the 9-item Fatigue Severity Scale, with a mean score of >5 indicating clinically significant fatigue [39,40]. QoL was measured using the Multiple Sclerosis Quality of Life-54, and 2 composite scores for mental and physical QoL and 12 subdomains scored from 0 (low) to 100 (high) were calculated [41].

Self-efficacy was measured using the 6-item University of Washington Self-Efficacy Scale (UWSES) [42]. The UWSES is an item response theory–based tool designed to measure disability management self-efficacy that was originally developed for people living with MS but has also been validated for adults living with other chronic health conditions. In particular, the UWSES queries people living with MS to assess whether they believe they can manage their health condition or disability, for instance, whether they are able to keep their health condition or disability from being the center of their life or interfering with how they deal with unexpected events or social interactions. As no clinically significant cutoff for self-efficacy is reported, a summary score was calculated, with higher UWSES scores indicating greater self-efficacy.

Anxiety and depressive symptoms were measured using the 14-item self-report Hospital Anxiety and Depression Scale [43]. The Hospital Anxiety and Depression Scale includes a 2-factor structure—depression and anxiety—each measured using 7 items. Scores range from 0 (no symptoms) to 21 (most severe symptoms), with scores of 0 to 7 considered "normal," scores of 8 to 10 considered "borderline anxiety or depression," and scores of 11 to 21 considered moderate to severe anxiety or depression.

Lifestyle-Related Characteristics

Previous participation in another lifestyle course or intervention (no or yes) and undertaking a particular diet program for MS (no or yes) and, if yes, the type of lifestyle course or intervention or diet were queried. Diet quality was measured using the modified Diet Habits Questionnaire (DHQ), previously validated in people living with MS over a 24-hour recall period [44], but excluding questions on salt and alcohol [45,46]. Responses were scored and summated. Total DHQ scores were estimated as a score out of 100. Total DHQ scores range between 20 and 100, with higher DHQ scores indicating a better quality of diet. Total scores were categorized into quartiles as in the MS-related study by Kirkland et al [47], with scores of >80 indicative of a "healthier" overall reported dietary intake. Physical activity was measured using the International Physical Activity Questionnaire (IPAQ) and categorized as inactive, minimally active, and active as per IPAQ guidelines [48]. The IPAQ has been used to evaluate physical activity in people living with MS in previous studies [49], including digital interventions aimed at increasing physical activity [50]. Frequency and quantity of alcohol consumption were queried and categorized into tertiles (none [no alcohol intake], limited [≤1 standard drink per day for female individuals and ≤2 standard drinks per day for male individuals], and heavy drinking [>1 standard drink per day for female individuals and >2 standard drinks per day for male individuals]) as in previous MS-related studies [3,10]. Participants were queried as to whether they consumed meat, dairy, and vitamin D and omega-3 supplements (no or yes). Current smoking status was dichotomized as current versus never or ex-smoker. Formal meditation practice (eg, sitting meditation) in the past week was queried (no or yes).

Data Collection

This study is an additional analysis of quantitative data collected at baseline before course commencement. While not part of this

XSL•FO
RenderX

study, the ancillary RCT will collect quantitative exposure and outcome data at 6, 12, and 30 months following course completion to examine the primary and secondary aims as per the study protocol [33]. Qualitative interviewing of approximately 40 course completers across both study arms will also be performed at 1 and 12 months after course completion to develop a deeper understanding of participants' experiences of the course and any impacts of the course.

Outcome Variables

In total, 2 measures of course engagement were examined: course commencement and course completion. We examined course commencement as a proxy for measuring course engagement by evaluating the relationship between participants being interested in the course (ie, signing up to undertake the course) and following through to commence the course. Course completion is also a commonly used metric to measure acceptability of web-based interventions [28].

Course commencement was defined as completion of at least the introductory module (module ≥1). For commencement analyses, course commencers (completers of module ≥1) were compared with course noncommencers (noncompleters of module 1). Intervention course and standard care course commencement was examined collectively due to the similarity of module 1 in both courses.

Course completion was defined as the completion of modules 1 to 6 as module 7 comprised a closing session and did not provide lifestyle-related information. For completion analyses, among course commencers (completers of module ≥1), course completers (completers of modules 1-6) were compared with course noncompleters (completers of module 1 but not module 6).

Data Analysis

Associations between participants' sociodemographic, health, and lifestyle-related characteristics and course commencement were examined. In addition, among course commencers, associations between participant characteristics and course completion stratified by the standard care course and intervention course study arms were examined.

Multivariate log-binomial regression was used to examine associations, with results presented as prevalence ratios and 95% CIs. Models were adjusted for age, sex, educational level, MS type, experience of ongoing symptoms from recent relapse, physical comorbidities, disability, use of DMTs, and participation in another lifestyle course or diet program at baseline. Model covariates were selected based on significant independent associations with exposure and outcome terms in the current data set in univariate analyses, as well as from review of relevant studies in the literature [3,6]. To ensure the validity of the regression models, we conducted collinearity diagnostics, which revealed no evidence of multicollinearity among the examined variables. This was confirmed through variance inflation factor and tolerance statistics. All statistical analyses were performed using Stata (version 16.0; StataCorp). Statistical significance was defined as $P<.05$.

Results

Baseline Participant Characteristics

Participants were of a mean age of 47.0 (SD 11.7) years; 87.7% (752/857) were female, 49.5% (424/857) had underweight or normal BMI, 65.6% (562/857) were university educated, 55% (471/857) were employed, and 69.8% (598/857) were married or partnered (Table 1).

Participants resided in 53 countries, with most participants residing in North America (286/857, 33.4%), Australia or New Zealand (236/857, 27.5%), and the United Kingdom (113/857, 13.2%) and 25.9% (222/857) residing in 48 other countries (Multimedia Appendix 3).

Most participants (614/857, 71.6%) reported having nonprogressive MS; 55.4% (475/857) reported clinically significant fatigue; and 17.2% (147/857) and 31.6% (271/857) reported symptoms of severe depression and anxiety, respectively (Table 1). The mean duration since MS diagnosis was 9.4 (SD 9.1) years. In total, 52.4% (449/857) of participants had none or mild disability, and 56.7% (486/857) reported ³1 comorbidity. A total of 67.8% (581/857) were taking DMTs, with ocrelizumab (164/857, 19.1%), natalizumab (56/857, 6.5%), and ofatumumab (53/857, 6.2%) being the most commonly reported (Multimedia Appendix 4).

Regarding lifestyle-related characteristics, 26.5% (227/857) of participants were following a lifestyle program or undertaking some lifestyle modifications (OMS: 68/857, 7.9%; exercise: 43/857, 5%; stress reduction: 9/857, 1%; other health-related modifications: 9/857, 1%), and 28.1% (241/857) followed a specific diet for MS (OMS: 17/241, 7.1%; gluten free: 12/241, 5%; vegetarian/vegan: 7/241, 2.9%; Mediterranean: 2/241, 0.8%; other MS-related diets [eg, Wahls]: 2/241, 0.8%; other healthy diets [eg, low sugar]: 14/241, 5.8%). A large proportion of participants were taking vitamin D (721/857, 84.1%) and omega-3 (296/857, 34.5%) supplements, 9.3% (80/857) were current smokers, 4.6% (36/857) were heavy drinkers, 28.2% (242/857) practiced meditation, and 19.1% (164/857) engaged in "active" levels of physical activity.

XSL•FO
RenderX

Table 1. Descriptive statistics of the study sample (n=857)[a].

Characteristics	Values
Age (y), mean (SD)	47.0 (11.7)
Sex (female), n (%)	752 (87.7)
BMI (kg/m^2), n (%)	
Underweight or normal	424 (49.5)
Overweight	212 (24.8)
Obese	219 (25.6)
Country of residence, n (%)	
Australia or New Zealand	236 (27.5)
United States or Canada	286 (33.4)
United Kingdom	113 (13.2)
Other	222 (25.9)
Educational level (university), n (%)	562 (65.6)
Employment status (working), n (%)	471 (55.0)
Marital status (partnered), n (%)	598 (69.8)
MS[b] type (nonprogressive), n (%)	614 (71.6)
MS duration (y), n (%)	
≤2	245 (28.6)
3-6	203 (23.7)
7-15	215 (25.1)
>15	194 (22.6)
Disability (PDDS[c]), n (%)	
None or mild	449 (52.4)
Moderate	323 (37.7)
Severe	85 (9.9)
Depression, n (%)	
Normal (HADS–Depression[d] 0-7)	515 (60.2)
Borderline (HADS–Depression >7-10)	193 (22.6)
Severe (HADS–Depression >10-14)	147 (17.2)
Anxiety, n (%)	
Normal (HADS–Anxiety[e] 0-7)	382 (44.6)
Borderline (HADS–Anxiety >7-10)	203 (23.7)
Severe (HADS–Anxiety >10-14)	271 (31.6)
Clinically significant fatigue (mean FSS[f] >5), n (%)	475 (55.4)
Comorbidities (≥1), n (%)	486 (56.7)
Taking DMTs[g], n (%)	581 (67.8)
Following a diet program for MS, n (%)	241 (28.1)
Participating in a lifestyle program or intervention, n (%)	227 (26.5)
Vitamin D supplements (any), n (%)	721 (84.1)
Omega-3 supplements (any), n (%)	296 (34.5)
Smoking, n (%)	

XSL•FO
RenderX

Characteristics	Values
Never smoker	482 (56.2)
Ex-smoker	295 (34.4)
Current smoker	80 (9.3)
Alcohol consumption, n (%)	
None	195 (24.7)
Limited	527 (66.6)
Heavy	36 (4.6)
Meditation, n (%)	242 (28.2)
Exercise (IPAQ[h]), n (%)	
Inactive	319 (37.2)
Minimally active	374 (43.6)
Active	164 (19.1)

[a]444 recruited participants randomized to the intervention course study arm completed the baseline survey, and 413 participants in the standard care course study arm completed the baseline survey.

[b]MS: multiple sclerosis.

[c]PDDS: Patient-Determined Disease Steps.

[d]HADS–Depression: Hospital Anxiety and Depression Scale for symptoms of depression.

[e]HADS–Anxiety: Hospital Anxiety and Depression Scale for symptoms of anxiety.

[f]FSS: Fatigue Severity Scale.

[g]DMT: disease-modifying therapy.

[h]IPAQ: International Physical Activity Questionnaire.

Course Commencement

Of the 1893 participants enrolled in the RCT, 857 (45.27%) completed the baseline survey—444 (23.45%) in the intervention course and 413 (21.82%) in the standard care course. A total of 33.33% (631/1893) of the total number of enrolled participants and 73.6% (631/857) of the participants who completed the baseline survey commenced the MSOC (Figure 1). The proportions of participants who completed the baseline survey and the MSOC were similar across study arms; completion rates in the intervention course were 49.1% (218/444) versus 54.2% (224/413) in the standard care course.

Following multivariate analysis, educational level, marital status, country of residence, mental and physical QoL, and fatigue were associated with MSOC commencement (intervention course and standard care course combined; Table 2).

A university degree or being in a relationship was associated with a 19% (95% CI 7%-32%) and 12% (95% CI 1%-25%) higher likelihood of MSOC commencement, respectively. Participants residing outside Australia or New Zealand, the United States, Canada, or the United Kingdom were 16% (95% CI 4%-26%) less likely to commence the MSOC.

People living with MS with fatigue were 10% (95% CI 1%-18%) less likely to commence the course, whereas people living with MS with mental and physical QoL scores in the top quartiles were 20% (95% CI 4%-38%) and 22% (95% CI 4%-43%) more likely to commence the MSOC, respectively. On the basis of our previous work [10], higher QoL is associated with less clinical severity, better mood, and generally better well-being. Specifically, being in the higher quartiles of categorized QoL indicates that participants with greater health and well-being were approximately 20% more likely to commence the MSOC.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart of the study cohort. IC: intervention course; MS: multiple sclerosis; SCC: standard care course.

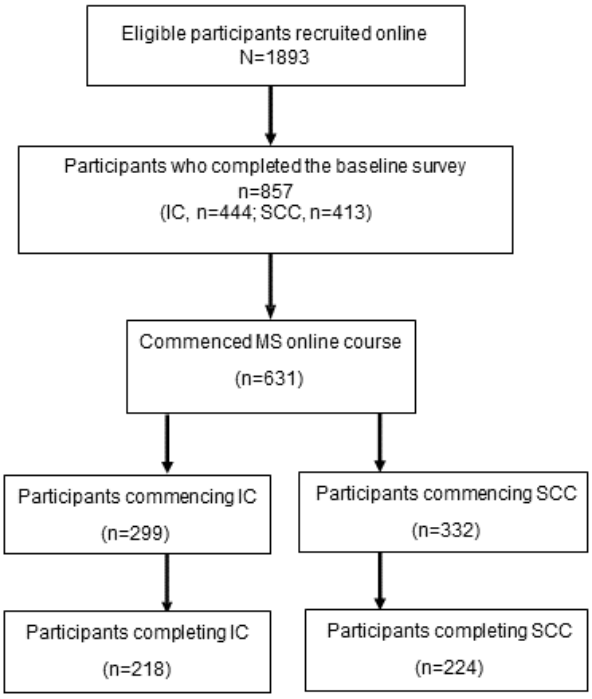

XSL•FO
RenderX

Table 2. Associations between sociodemographic, health, and lifestyle-related characteristics and course commencement[a].

Characteristics	Participants (n=613)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Sociodemographics[e]			
Age (y)			
≤35	108 (17.1)	1.00 (reference)	1.00 (reference)
36-44	172 (27.3)	0.99 (0.87-1.12)	0.99 (0.93-1.16)
45-54	173 (27.4)	1.01 (0.89-1.14)	1.08 (0.89-1.14)
≥55	178 (28.2)	1.01 (0.89-1.14)	1.11 (0.95-1.29)
Sex			
Female	558 (88.9)	1.00 (reference)	1.00 (reference)
Male	70 (11.1)	0.94 (0.82-1.08)	0.95 (0.82-1.11)
Educational level			
Below university	205 (32.5)	1.00 (reference)	1.00 (reference)
University	426 (67.5)	*1.09 (1.01-1.19)*[f]	*1.19 (1.07-1.32)*[g]
Employment status			
Not working	237 (40.2)	1.00 (reference)	1.00 (reference)
Working	352 (59.8)	1.03 (0.95-1.13)	1.02 (0.92-1.13)
In a relationship or partnered			
No	168 (26.9)	1.00 (reference)	1.00 (reference)
Yes	456 (73.1)	*1.14 (1.03-1.26)*[f]	*1.12 (1.01-1.25)*[f]
Country of residence			
Australia or New Zealand	187 (29.6)	1.00 (reference)	1.00 (reference)
United States or Canada	203 (32.2)	*0.90 (0.81-0.99)*[f]	0.93 (0.83-1.04)
United Kingdom	92 (14.6)	1.03 (0.92-1.15)	1.03 (0.92-1.15)
Other	149 (23.6)	*0.85 (0.76-0.95)*[g]	*0.84 (0.74-0.96)*[g]
BMI			
Underweight or normal	325 (51.6)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Overweight	149 (23.7)	0.92 (0.83-1.02)	0.91 (0.81-1.03)
Obese	156 (24.8)	0.93 (0.84-1.03)	0.93 (0.82-1.05)
Perceived support			
Lowest	155 (24.6)	1.00 (reference)	1.00 (reference)
Second quartile	165 (26.1)	1.06 (0.94-1.19)	0.99 (0.87-1.14)
Third quartile	147 (23.3)	1.06 (0.93-1.19)	1.03 (0.90-1.18)
Fourth quartile	164 (26)	1.10 (0.98-1.23)	1.11 (0.98-1.25)
Health[h]			
MS[i] type			
Nonprogressive	467 (78.9)	1.00 (reference)	1.00 (reference)
Progressive	125 (21.1)	0.94 (0.84-1.04)	0.89 (0.78-1.02)
MS duration since diagnosis (years)			
≤2	180 (28.5)	1.00 (reference)	1.00 (reference)
3-6	158 (25)	1.06 (0.95-1.18)	1.03 (0.91-1.17)
7-15	157 (24.9)	0.99 (0.89-1.11)	0.97 (0.85-1.10)

XSL•FO
RenderX

Characteristics	Participants (n=613)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
>15	136 (21.6)	0.95 (0.85-1.07)	0.92 (0.78-1.07)
Disability (PDDS[j])			
None or mild	338 (53.6)	1.00 (reference)	1.00 (reference)
Moderate	231 (36.6)	0.95 (0.87-1.04)	0.98 (0.87-1.05)
Severe	62 (9.8)	0.97 (0.84-1.11)	1.00 (0.84-1.18)
Mental quality of life			
Lowest	142 (22.5)	1.00 (reference)	1.00 (reference)
Second quartile	154 (24.4)	1.08 (0.96-1.23)	1.08 (0.96-1.23)
Third quartile	159 (25.2)	1.09 (0.96-1.23)	1.09 (0.96-1.23)
Fourth quartile	176 (27.9)	*1.25 (1.11-1.39)*[g]	*1.20 (1.04-1.38)*[f]
Physical quality of life			
Lowest	150 (24.3)	1.00 (reference)	1.00 (reference)
Second quartile	145 (23.5)	1.08 (0.95-1.23)	1.13 (0.98-1.32)
Third quartile	155 (25.1)	*1.13 (1.01-1.28)*[f]	1.14 (0.98-1.33)
Fourth quartile	167 (27.1)	*1.21 (1.08-1.36)*[g]	*1.22 (1.04-1.43)*[f]
Self-efficacy			
Lowest	170 (27)	1.00 (reference)	1.00 (reference)
Second quartile	167 (26.6)	1.10 (0.99-1.23)	1.10 (0.99-1.23)
Third quartile	138 (21.9)	1.07 (0.95-1.20)	1.07 (0.95-1.20)
Fourth quartile	154 (24.5)	1.05 (0.94-1.18)	1.05 (0.94-1.18)
Fatigue (FSS[k] >5)			
No	297 (47.1)	1.00 (reference)	1.00 (reference)
Yes	334 (52.9)	*0.90 (0.84-0.98)*[f]	*0.90 (0.82-0.99)*[f]
Depression			
Normal (HADS-D[l] 0-7)	396 (63)	1.00 (reference)	1.00 (reference)
Borderline (HADS-D >7-10)	137 (21.8)	0.92 (0.83-1.02)	0.93 (0.82-1.05)
Severe (HADS-D >10-14)	96 (15.3)	*0.85 (0.75-0.96)*[f]	0.86 (0.73-1.00)
Anxiety			
Normal (HADS-A[m] 0-7)	292 (46.3)	1.00 (reference)	1.00 (reference)
Borderline (HADS-A >7-10)	143 (22.7)	0.92 (0.83-1.02)	0.90 (0.79-1.02)
Severe (HADS-A >10-14)	195 (31)	0.94 (0.86-1.03)	0.97 (0.86-1.08)
Comorbidities (≥1)			
No	255 (41.7)	1.00 (reference)	1.00 (reference)
Yes	356 (58.3)	0.99 (0.91-1.07)	0.96 (0.87-1.05)
Taking DMTs[n]			
No	195 (30.9)	1.00 (reference)	1.00 (reference)
Yes	436 (69.1)	1.06 (0.97-1.16)	1.08 (0.97-1.21)
Lifestyle related[o]			
Participating in a lifestyle program			
No	465 (73.7)	1.00 (reference)	1.00 (reference)

XSL•FO
RenderX

Characteristics	Participants (n=613)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Yes	166 (26.3)	0.99 (0.90-1.09)	0.97 (0.88-1.08)
Following a diet program			
No	447 (70.8)	1.00 (reference)	1.00 (reference)
Yes	184 (29.2)	1.05 (0.97-1.15)	1.06 (0.97-1.17)
Diet quality			
Lowest	140 (22.2)	1.00 (reference)	1.00 (reference)
Second quartile	160 (25.4)	1.10 (0.97-1.25)	1.08 (0.94-1.25)
Third quartile	173 (27.4)	1.11 (0.99-1.26)	1.09 (0.95-1.26)
Fourth quartile	158 (25)	*1.21 (1.07-1.36)*[f]	1.14 (0.99-1.31)
Vitamin D supplements (any)			
No	90 (14.3)	1.00 (reference)	1.00 (reference)
Yes	541 (85.7)	1.13 (1.00-1.29)	1.17 (1.00-1.38)
Omega-3 supplements (any)			
No	406 (64.3)	1.00 (reference)	1.00 (reference)
Yes	225 (35.7)	1.05 (0.97-1.14)	1.09 (0.99-1.20)
Smoking			
Never smoker	359 (56.9)	1.00 (reference)	1.00 (reference)
Ex-smoker	223 (35.3)	1.01 (0.93-1.10)	1.01 (0.91-1.10)
Current smoker	49 (7.8)	*0.82 (0.69-0.99)*[f]	0.88 (0.72-1.09)
Alcohol consumption			
None	134 (23.5)	1.00 (reference)	1.00 (reference)
Limited	409 (71.6)	*1.13 (1.02-1.25)*[f]	1.11 (0.99-1.24)
Heavy	28 (4.9)	1.13 (0.93-1.38)	1.11 (0.88-1.39)
Meditation			
No	450 (71.3)	1.00 (reference)	1.00 (reference)
Yes	181 (28.7)	1.02 (0.94-1.12)	1.01 (0.91-1.11)
Exercise (IPAQ[p])			
Inactive	228 (36.1)	1.00 (reference)	1.00 (reference)
Minimally active	282 (44.7)	1.05 (0.96-1.15)	1.04 (0.93-1.15)

XSL•FO
RenderX

Characteristics	Participants (n=613)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Active	121 (19.2)	1.03 (0.92-1.16)	0.95 (0.82-1.11)

[a]Adjusted log-binomial regression models for age, sex, educational level, multiple sclerosis type, ongoing symptoms due to recent relapse, number of comorbidities, disability, use of DMTs, participation in another lifestyle intervention, and adherence to a specific diet. Italicized values denote significant association between characteristics and course commencement.

[b]631 participants commenced the MSOC (299 commenced the intervention course and 332 commenced the standard care course). For some variables, due to data unavailability the number of participants within this subgroup does not add up to 631, however the percentages add up to 100%.

[c]PR: prevalence ratio.

[d]aPR: adjusted prevalence ratio.

[e]Sociodemographic-related characteristics.

[f]$P<.05$.

[g]$P<.01$.

[h]Characteristics related to participants' health.

[i]MS: multiple sclerosis.

[j]PDDS: Patient-Determined Disease Steps.

[k]FSS: Fatigue Severity Scale.

[l]HADS-D: Hospital Anxiety and Depression Scale for symptoms of depression.

[m]HADS-A: Hospital Anxiety and Depression Scale for symptoms of anxiety.

[n]DMT: disease-modifying therapy.

[o]Lifestyle-related characteristics.

[p]IPAQ: International Physical Activity Questionnaire.

Course Completion

Completion rates were similar across study arms (intervention course: 218/299, 72.9%; standard care course: 224/332, 67.5%; Figure 1). However, factors associated with course completion varied between the intervention course (Table 3) and standard care course (Table 4).

Participants aged 45 to 54 years and ≥55 years were 54% (95% CI 6%-123%) and 96% (95% CI 34%-187%) more likely, respectively, to complete the intervention course than participants aged ≤35 years (Table 3). Being male and following a diet program were associated with 27% (95% CI 7%-51%) and 19% (95% CI 1%-40%) higher intervention course completion, respectively. Clinically significant fatigue was associated with 24% (95% CI 3%-48%) higher intervention course completion. Conversely, people living with MS in the top 3 quartiles of self-efficacy had 35% (95% CI 18%-49%), 26% (95% CI 9%-42%), and 27% (95% CI 7%-43%) lower intervention course completion.

Participants in the standard care course study arm who practiced meditation were 20% (95% CI 2%-41%) more likely to complete the standard care course, whereas those who were employed had 22% (95% CI 8%-30%) lower completion (Table 4).

XSL•FO
RenderX

Table 3. Associations between sociodemographic, health, and lifestyle-related characteristics and course completion in the intervention course (n=218)[a].

Characteristics	Participants (n=218)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Sociodemographics[e]			
Age (y)			
≤35	29 (13.3)	1.00 (reference)	1.00 (reference)
36-44	54 (24.8)	1.11 (0.82-1.51)	1.38 (0.94-2.02)
45-54	67 (30.7)	*1.39 (1.05-1.85)*[f]	*1.54 (1.06-2.23)*[f]
≥55	68 (31.2)	*1.48 (1.12-1.95)*[g]	*1.96 (1.34-2.87)*[h]
Sex			
Female	182 (83.5)	1.00 (reference)	1.00 (reference)
Male	36 (16.5)	*1.28 (1.09-1.51)*[g]	*1.27 (1.07-1.51)*[g]
Educational level			
Below university	62 (28.4)	1.00 (reference)	1.00 (reference)
University	156 (71.6)	1.08 (0.91-1.30)	1.14 (0.93-1.40)
Employment status			
Not working	83 (41.5)	1.00 (reference)	1.00 (reference)
Working	117 (58.5)	0.99 (0.84-1.17)	1.02 (0.84-1.25)
In a relationship or partnered			
No	61 (28.2)	1.00 (reference)	1.00 (reference)
Yes	155 (71.8)	0.90 (0.76-1.06)	1.05 (0.85-1.30)
Country of residence			
Australia or New Zealand	64 (29.4)	1.00 (reference)	1.00 (reference)
United States or Canada	66 (30.3)	0.88 (0.73-1.08)	0.85 (0.68-1.08)
United Kingdom	33 (15.1)	0.89 (0.70-1.14)	0.98 (0.77-1.25)
Other	55 (25.2)	0.91 (0.74-1.12)	0.97 (0.76-1.23)
BMI (kg/m^2)			
Underweight or normal	127 (58.5)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Overweight	48 (22.1)	0.89 (0.73-1.09)	0.86 (0.68-1.10)
Obese	42 (19.4)	0.84 (0.67-1.04)	0.95 (0.74-1.22)
Perceived support			
Lowest	52 (23.9)	1.00 (reference)	1.00 (reference)
Second quartile	61 (28)	1.09 (0.89-1.34)	1.17 (0.91-1.50)
Third quartile	54 (24.8)	1.04 (0.83-1.29)	1.21 (0.95-1.56)
Fourth quartile	51 (23.4)	0.87 (0.68-1.11)	0.97 (0.74-1.26)
Health related[i]			
MS[j] type			
Nonprogressive	159 (77.6)	1.00 (reference)	1.00 (reference)
Progressive	46 (22.4)	1.01 (0.84-1.22)	0.94 (0.76-1.18)
MS duration since diagnosis (years)			
≤2	64 (29.4)	1.00 (reference)	1.00 (reference)
3-6	48 (22)	0.88 (0.69-1.11)	0.88 (0.69-1.11)
7-15	53 (24.3)	1.04 (0.84-1.28)	1.04 (0.84-1.28)

XSL•FO
RenderX

Characteristics	Participants (n=218)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
>15	53 (24.3)	1.13 (0.92-1.38)	1.13 (0.92-1.38)
Disability (PDDS[k])			
None or mild	112 (51.4)	1.00 (reference)	1.00 (reference)
Moderate	84 (38.5)	1.11 (0.94-1.31)	1.01 (0.84-1.22)
Severe	22 (10.1)	1.23 (0.97-1.55)	0.86 (0.60-1.23)
Mental quality of life			
Lowest	44 (20.2)	1.00 (reference)	1.00 (reference)
Second quartile	52 (23.9)	1.12 (0.88-1.43)	1.02 (0.78-1.32)
Third quartile	54 (24.8)	1.12 (0.88-1.43)	0.95 (0.73-1.24)
Fourth quartile	68 (31.2)	1.16 (0.92-1.46)	0.97 (0.75-1.26)
Physical quality of life			
Lowest	50 (23.4)	1.00 (reference)	1.00 (reference)
Second quartile	52 (24.3)	1.01 (0.81-1.27)	0.93 (0.81-1.27)
Third quartile	63 (29.4)	1.11 (0.91-1.37)	1.11 (0.91-1.31)
Fourth quartile	49 (22.9)	0.90 (0.70-1.14)	0.77 (0.57-1.04)
Self-efficacy			
Lowest	72 (33.2)	1.00 (reference)	1.00 (reference)
Second quartile	51 (23.5)	*0.76 (0.62-0.95)*[f]	*0.65 (0.51-0.82)*[h]
Third quartile	47 (21.7)	0.86 (0.70-1.06)	*0.74 (0.58-0.91)*[g]
Fourth quartile	47 (21.7)	0.86 (0.70-1.06)	*0.73 (0.57-0.93)*[f]
Fatigue (FSS[l] >5)			
No	96 (44.0)	1.00 (reference)	1.00 (reference)
Yes	122 (56.0)	1.14 (0.97-1.34)	*1.24 (1.03-1.48)*[f]
Depression			
Normal (HADS-D[m] 0-7)	142 (65.1)	1.00 (reference)	1.00 (reference)
Borderline (HADS-D >7-10)	45 (20.6)	0.92 (0.75-1.13)	0.95 (0.75-1.20)
Severe (HADS-D >10-14)	31 (14.2)	1.01 (0.81-1.26)	1.02 (0.79-1.33)
Anxiety			
Normal (HADS-A[n]0-7)	104 (47.7)	1.00 (reference)	1.00 (reference)
Borderline (HADS-A >7-10)	60 (27.5)	0.98 (0.82-1.17)	1.02 (0.81-1.28)
Severe (HADS-A >10-14)	54 (24.8)	0.85 (0.70-1.04)	1.01 (0.80-1.29)
Comorbidities (≥1)			
No	88 (41.5)	1.00 (reference)	1.00 (reference)
Yes	124 (58.5)	0.97 (0.83-1.14)	0.97 (0.83-1.14)
Taking DMTs[o]			
No	71 (32.6)	1.00 (reference)	1.00 (reference)
Yes	147 (67.4)	1.10 (0.92-1.30)	1.22 (0.99-1.51)
Lifestyle related[p]			
Participating in a lifestyle program			
No	160 (73.4)	1.00 (reference)	1.00 (reference)
Yes	58 (26.6)	1.05 (0.89-1.25)	1.05 (0.89-1.25)

XSL•FO
RenderX

Characteristics	Participants (n=218)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Following a diet program			
No	146 (67.0)	1.00 (reference)	1.00 (reference)
Yes	72 (33)	*1.18 (1.01-1.38)*[f]	*1.19 (1.01-1.40)*[f]
Diet quality			
Lowest	42 (19.3)	1.00 (reference)	1.00 (reference)
Second quartile	51 (23.4)	0.98 (0.75-1.28)	1.06 (0.75-1.28)
Third quartile	58 (26.6)	1.16 (0.91-1.48)	1.03 (0.91-1.48)
Fourth quartile	67 (30.7)	*1.35 (1.08-1.69)*[g]	1.23 (0.94-1.60)
Vitamin D supplements (any)			
No	28 (12.8)	1.00 (reference)	1.00 (reference)
Yes	190 (87.2)	1.06 (0.83-1.36)	1.05 (0.78-1.39)
Omega-3 supplements (any)			
No	126 (57.8)	1.00 (reference)	1.00 (reference)
Yes	92 (42.2)	1.03 (0.88-1.20)	1.03 (0.88-1.20)
Smoking			
Never smoker	135 (61.9)	1.00 (reference)	1.00 (reference)
Ex-smoker	70 (32.1)	0.94 (0.79-1.12)	0.88 (0.73-1.07)
Current smoker	13 (6.0)	0.76 (0.52-1.13)	0.83 (0.54-1.28)
Alcohol consumption			
None	42 (21.3)	1.00 (reference)	1.00 (reference)
Limited	144 (73.1)	0.99 (0.80-1.22)	1.17 (0.92-1.47)
Heavy	11 (5.6)	1.04 (0.76-1.43)	1.15 (0.74-1.79)
Meditation			
No	146 (67)	1.00 (reference)	1.00 (reference)
Yes	72 (33)	1.16 (0.99-1.36)	1.15 (0.96-1.39)
Exercise (IPAQ[q])			
Inactive	76 (34.9)	1.00 (reference)	1.00 (reference)
Minimally active	91 (41.7)	1.02 (0.85-1.22)	0.91 (0.74-1.18)

XSL•FO
RenderX

Characteristics	Participants (n=218)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Active	51 (23.4)	1.09 (0.89-1.34)	0.94 (0.73-1.22)

[a]Adjusted log-binomial regression models for age, sex, educational level, multiple sclerosis type, ongoing symptoms due to recent relapse, number of comorbidities, disability, use of DMTs, participation in another lifestyle intervention, and adherence to a specific diet. Italicized values denote significant association between characteristics and course commencement or completion.

[b]218 participants completed the intervention course. Due to data unavailability the number of participants for certain variables (employment status, BMI, relationship status, MS type, comorbidities, physical quality of life, self-efficacy, and alcohol consumption) is <218, however the percentages for these variables add up to 100%.

[c]PR: prevalence ratio.

[d]aPR: adjusted prevalence ratio.

[e]Sociodemographic-related characteristics.

[f]$P<.05$.

[g]$P<.01$.

[h]$P<.001$.

[i]Characteristics related to participants' health.

[j]MS: multiple sclerosis.

[k]PDDS: Patient-Determined Disease Steps.

[l]FSS: Fatigue Severity Scale.

[m]HADS-D: Hospital Anxiety and Depression Scale for symptoms of depression.

[n]HADS-A: Hospital Anxiety and Depression Scale for symptoms of anxiety.

[o]DMT: disease-modifying therapy.

[p]Lifestyle-related characteristics.

[q]IPAQ: International Physical Activity Questionnaire.

XSL•FO
RenderX

Table 4. Associations between sociodemographic, health, and lifestyle-related characteristics and course completion in the standard care course (n=224)[a].

Characteristics	Participants, (n=224)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Sociodemographics[e]			
Age (y)			
≤35	35 (15.6)	1.00 (reference)	1.00 (reference)
36-44	55 (24.6)	1.05 (0.83-1.33)	1.04 (0.78-1.39)
45-54	67 (29.9)	*1.24 (1.01-1.54)*[f]	1.18 (0.90-1.54)
≥55	67 (29.9)	1.12 (0.89-1.40)	1.14 (0.86-1.50)
Sex			
Female	203 (90.6)	1.00 (reference)	1.00 (reference)
Male	21 (9.4)	1.09 (0.89-1.33)	1.09 (0.85-1.40)
Educational level			
Below university	77 (34.4)	1.00 (reference)	1.00 (reference)
University	147 (65.6)	1.03 (0.90-1.19)	1.10 (0.92-1.31)
Employment status			
Not working	94 (44.8)	1.00 (reference)	1.00 (reference)
Working	116 (55.2)	*0.79 (0.70-0.90)*[g]	*0.78 (0.70-0.92)*[h]
In a relationship or partnered			
No	61 (27.6)	1.00 (reference)	1.00 (reference)
Yes	160 (72.4)	1.01 (0.87-1.17)	0.97 (0.82-1.14)
Country of residence			
Australia or New Zealand	71 (31.7)	1.00 (reference)	1.00 (reference)
United States or Canada	73 (32.6)	1.02 (0.86-1.20)	1.01 (0.84-1.23)
United Kingdom	34 (15.2)	1.16 (0.97-1.39)	1.11 (0.89-1.39)
Other	46 (20.5)	0.98 (0.81-1.19)	1.02 (0.79-1.32)
BMI (kg/m^2)			
Underweight or normal	106 (47.3)	1.00 (reference)	1.00 (reference)
Overweight	55 (24.6)	1.03 (0.88-1.21)	1.04 (0.85-1.28)
Obese	63 (28.1)	1.01 (0.86-1.18)	0.98 (0.80-1.21)
Perceived support			
Lowest	59 (26.3)	1.00 (reference)	1.00 (reference)
Second quartile	61 (27.2)	0.98 (0.83-1.17)	1.02 (0.83-1.26)
Third quartile	51 (22.8)	0.97 (0.80-1.16)	0.99 (0.80-1.24)
Fourth quartile	53 (23.7)	0.91 (0.75-1.10)	0.98 (0.78-1.23)
Health related[i]			
MS[j] type			
Nonprogressive	166 (78.3)	1.00 (reference)	1.00 (reference)
Progressive	46 (21.7)	1.12 (0.97-1.29)	1.07 (0.88-1.31)
MS duration since diagnosis (years)			
≤2	61 (27.2)	1.00 (reference)	1.00 (reference)
3-6	59 (26.3)	1.07 (0.90-1.27)	1.07 (0.90-1.27)
7-15	54 (24.1)	0.90 (0.75-1.12)	0.90 (0.75-1.12)

XSL•FO
RenderX

Characteristics	Participants, (n=224)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
>15	50 (22.3)	1.05 (0.88-1.26)	1.05 (0.88-1.26)
Disability (PDDS[k])			
None or mild	116 (51.8)	1.00 (reference)	1.00 (reference)
Moderate	87 (38.8)	1.08 (0.95-1.23)	1.08 (0.90-1.30)
Severe	21 (9.4)	0.86 (0.65-1.13)	0.83 (0.59-1.16)
Mental quality of life			
Lowest	52 (23.2)	1.00 (reference)	1.00 (reference)
Second quartile	54 (24.1)	0.93 (0.76-1.13)	0.96 (0.78-1.18)
Third quartile	57 (25.4)	0.96 (0.79-1.15)	0.78 (0.61-1.01)
Fourth quartile	61 (27.2)	1.04 (0.87-1.23)	1.02 (0.83-1.25)
Physical quality of life			
Lowest	55 (25.1)	1.00 (reference)	1.00 (reference)
Second quartile	47 (21.5)	0.94 (0.77-1.16)	0.92 (0.71-1.19)
Third quartile	52 (23.7)	1.01 (0.84-1.23)	0.98 (0.77-1.24)
Fourth quartile	65 (29.7)	1.04 (0.87-1.24)	1.01 (0.77-1.35)
Self-efficacy			
Lowest	51 (22.8)	1.00 (reference)	1.00 (reference)
Second quartile	59 (26.3)	1.10 (0.90-1.34)	1.07 (0.83-1.38)
Third quartile	52 (23.2)	1.16 (0.95-1.41)	1.12 (0.88-1.44)
Fourth quartile	62 (27.7)	1.11 (0.91-1.35)	1.09 (0.84-1.41)
Fatigue (FSS[l] >5)			
No	106 (47.3)	1.00 (reference)	1.00 (reference)
Yes	118 (52.7)	0.98 (0.86-1.12)	1.08 (0.91-1.27)
Depression			
Normal (HADS-D[m] 0-7)	138 (61.9)	1.00 (reference)	1.00 (reference)
Borderline (HADS-D >7-10)	47 (21.1)	0.97 (0.82-1.15)	1.01 (0.83-1.22)
Severe (HADS-D >10-14)	38 (17.0)	1.01 (0.84-1.20)	0.96 (0.76-1.22)
Anxiety			
Normal (HADS-A[n] 0-7)	105 (46.9)	1.00 (reference)	1.00 (reference)
Borderline (HADS-A >7-10)	43 (19.2)	1.07 (0.91-1.26)	1.05 (0.85-1.30)
Severe (HADS-A >10-14)	76 (33.9)	0.99 (0.85-1.15)	0.95 (0.78-1.15)
Comorbidities (≥1)			
No	90 (42.1)	1.00 (reference)	1.00 (reference)
Yes	124 (57.9)	1.03 (0.90-1.19)	1.03 (0.88-1.21)
Taking DMTs[o]			
No	65 (29.0)	1.00 (reference)	1.00 (reference)
Yes	159 (71.0)	0.89 (0.78-1.02)	0.84 (0.70-1.00)
Lifestyle related[p]			
Participating in a lifestyle program			
No	164 (73.2)	1.00 (reference)	1.00 (reference)
Yes	60 (26.8)	0.98 (0.85-1.14)	0.98 (0.85-1.14)

XSL•FO
RenderX

Characteristics	Participants, (n=224)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Following a diet program			
No	161 (71.9)	1.00 (reference)	1.00 (reference)
Yes	63 (28.1)	0.97 (0.83-1.12)	0.93 (0.78-1.12)
Diet quality			
Lowest	45 (20.1)	1.00 (reference)	1.00 (reference)
Second quartile	57 (25.4)	1.21 (0.99-1.49)	1.24 (0.95-1.63)
Third quartile	64 (28.6)	1.11 (0.90-1.37)	1.15 (0.87-1.52)
Fourth quartile	58 (25.9)	1.20 (0.98-1.48)	1.19 (0.90-1.58)
Vitamin D supplements (any)			
No	31 (13.8)	1.00 (reference)	1.00 (reference)
Yes	193 (86.2)	1.10 (0.90-1.36)	1.04 (0.82-1.33)
Omega-3 supplements (any)			
No	161 (71.9)	1.00 (reference)	1.00 (reference)
Yes	63 (28.1)	0.95 (0.82-1.11)	0.95 (0.82-1.11)
Smoking			
Never smoker	125 (55.8)	1.00 (reference)	1.00 (reference)
Ex-smoker	80 (35.7)	0.90 (0.78-1.05)	0.91 (0.77-1.07)
Current smoker	19 (8.5)	1.02 (0.82-1.27)	1.07 (0.82-1.38)
Alcohol consumption			
None	45 (22.1)	1.00 (reference)	1.00 (reference)
Limited	148 (72.6)	1.01 (0.85-1.20)	1.19 (0.95-1.47)
Heavy	11 (5.4)	1.15 (0.87-1.51)	1.27 (0.89-1.83)
Meditation			
No	156 (69.6)	1.00 (reference)	1.00 (reference)
Yes	68 (30.4)	*1.15 (1.01-1.31)*[f]	*1.20 (1.02-1.41)*[f]
Exercise (IPAQ[q])			
Inactive	77 (34.4)	1.00 (reference)	1.00 (reference)
Minimally active	109 (48.7)	1.09 (0.93-1.26)	1.15 (0.95-1.40)

XSL•FO
RenderX

Characteristics	Participants, (n=224)[b], n (%)	Univariate analysis, PR[c] (95% CI)	Multivariate analysis, aPR[d] (95% CI)
Active	38 (17.0)	1.12 (0.93-1.35)	1.25 (0.97-1.62)

[a]Adjusted log-binomial regression models for age, sex, educational level, multiple sclerosis type, ongoing symptoms due to recent relapse, number of comorbidities, disability, use of DMTs, participation in another lifestyle intervention, and adherence to a specific diet. Italicized values denote significant association between characteristics and course commencement or completion.

[b]224 participants completed the standard care course. Due to data inavailability the number of participants for certain variables (employment status, relationship status, MS type, depression, comorbidities, physical quality of life, self-efficacy, and alcohol consumption) is <224, however the percentages for these variables add up to 100%.

[c]PR: prevalence ratio.

[d]aPR: adjusted prevalence ratio.

[e]Sociodemographic-related characteristics.

[f]$P<.05$.

[g]$P<.01$.

[h]$P<.001$.

[i]Characteristics related to participants' health.

[j]MS: multiple sclerosis.

[k]PDDS: Patient-Determined Disease Steps.

[l]FSS: Fatigue Severity Scale.

[m]HADS-D: Hospital Anxiety and Depression Scale for symptoms of depression.

[n]HADS-A: Hospital Anxiety and Depression Scale for symptoms of anxiety.

[o]DMT: disease-modifying therapy.

[p]Lifestyle-related characteristics.

[q]IPAQ: International Physical Activity Questionnaire.

Discussion

Principal Findings

Our study examined the rates of completion and commencement of a web-based course on modification of lifestyle-related risk factors for people living with MS. We found that conversion rates from initial enrollment to course commencement were relatively low (631/1893, 33.33%). However, commencement rates in those who had completed the baseline survey after enrollment were relatively high (631/857, 73.6%), as were completion rates for those who had commenced the course (218/444, 49.1% for the intervention course and 224/413, 54.2% for the standard care course). Our study also sheds light on the factors associated with course commencement and completion. Those factors related to course commencement of potentially practical interest included educational level, being in a relationship, and clinical factors such as QoL and fatigue. Factors of potential interest associated with completion included age of >45 years, male sex, being employed, self-efficacy, and already having modified some risk factors such as diet and undertaking stress-reducing activities before course commencement. We also acknowledge that the length of the baseline survey (166 questions) may have influenced commencement and completion rates, with more motivated participants and possibly more able people likely to have both commenced and completed the MSOC.

The 166-question baseline survey was estimated to take 45 to 60 minutes, and completion was requested (but not mandated) before commencing the course. This may have posed a barrier for some individuals as only 45.27% (857/1893) of participants enrolled in the MSOC effectiveness RCT completed the survey. Data are not available as to how many participants started but did not complete the survey (ie, those that attempted the survey but "gave up" before commencing the course). In hindsight, a qualitative analysis of people who did not complete the baseline survey would have been beneficial to determine the influence that the baseline survey had on MSOC commencement and completion rates and is a strong recommendation for future studies.

Notably, low rates of baseline survey completion are consistent with our previous MSOC feasibility RCT (42%) [34]. This may be attributed to time constraints or disease-related limitations, including visual impairment or cognitive fatigue. The baseline survey was adapted from the survey used in the Health Outcomes and Lifestyle in a Sample of People With Multiple Sclerosis study conducted by this research group [51]. In this study, the extensive nature of the survey enabled prospective analyses of associations between lifestyle-related risk factors and health outcomes. Despite substantial participant attrition over 7.5 years of follow-up, significant associations were found between lifestyle modification and improved health outcomes [3,10]. Hence, the research team, while understanding that the survey may be a barrier to completion, elected to use the baseline survey to obtain sufficient data to reach robust conclusions regarding the effect of the MSOC on QoL and health outcomes into the future. Future researchers need to make informed choices regarding the length of data collection surveys and their impact on course commencement and completion and weigh the potential benefits of increased data against the attrition and potential bias related to noncommencement. Future studies could consider the use of a shorter adaptive questionnaire to possibly enhance participation in the context of an RCT.

We also found positive associations between course commencement and a university degree, consistent with reported

XSL•FO
RenderX

findings of associations between higher education and positive attitudes toward health research and a willingness to participate [52,53]. People with higher educational levels may have had more exposure to, experience of, and capabilities with using digital technology. Those with higher educational levels may possibly have found the baseline survey less onerous due to enhanced digital skills and, therefore, less of a barrier to commencement. However, higher education was not associated with course completion.

Participants in a relationship or with higher perceived support were more likely to commence the MSOC, consistent with reported associations between supportive relationships and digital intervention engagement [54]. Furthermore, support from family, teachers, and friends that students receive, as well as organizational support, is related to increased adherence to web-based learning [55,56]. MSOC participants received organizational support provided by a full-time course coordinator whom they were able to directly email for advice and support with technical issues. They were also provided with support from peers and facilitators in the forum who would answer direct questions regarding course content and facilitate discussions. These factors may explain the association between perceived support and MSOC commencement, as well as the high rates of intervention course and standard care course completion among course commencers.

Consistent with studies highlighting the importance of health and well-being in digital health engagement [57,58], participants with higher QoL (ie, greater health and well-being) were more likely to commence the MSOC, whereas higher fatigue was associated with decreased MSOC commencement. Similarly, a recent systematic review of digital health intervention engagement found that lower health-related QoL and higher depression were significant barriers to engagement [59]. Given that improving QoL is a central aim of both this course and similar lifestyle modification programs, it is imperative for future research to identify and address the unmet needs of people living with MS who report lower QoL as they are likely to derive the greatest benefit from web-based interventions.

Intervention course and standard care course completion rates were 49.1% (218/444) and 54.2% (224/413), respectively, comparable to 60% (9/15) and 50% (8/16) in the MSOC feasibility RCT [34]. Similarly, Claflin et al [29] reported *Understanding MS* MOOC completion rates of 42%. In contrast, a recent meta-analysis of completion rates by people living with MS across 32 RCTs reported considerably lower completion rates (15%-17%) [28]. It is possible that the duration of the digital health program may affect completion rates as 28 of the 32 web-based programs in the meta-analysis spanned 8 to 52 weeks, whereas both the MSOC and the *Understanding MS* MOOC were 6-week programs.

Participants aged >45 years had higher intervention course completion. This was similar to other studies that found that middle-aged people were most likely to complete web-based lifestyle-related health interventions [54,60]. A meta-analysis of 10 RCTs examining engagement in web-based interventions for depression identified that younger age is associated with higher dropout rates and rates of completion also seem to decrease among individuals aged >65 years [61]. This phenomenon has been ascribed to various factors, including digital proficiency, physical limitations, and lack of interest [62].

Interestingly, being male was associated with higher intervention course completion, which is different from the meta-analysis findings, where the greatest dropout from web-based interventions was found among male individuals [61]. Despite only 9.4% (21/224) and 16.5% (36/218) of participants in the standard care course and intervention course being male, respectively, these proportions reflect the rates of MS diagnoses, as considerably more female than male individuals are diagnosed with MS, and are similar to other well-studied MS cohorts [63]. Moreover, it is important to highlight the small number of male individuals in our cohort, so caution needs to be taken when interpreting these study findings.

Employed participants had 22% lower standard care course completion, which was similar to a prospective pilot study that identified links between work-life balance issues and lower completion of a 15-day web-based well-being intervention [64]. Similarly, time constraints from other life responsibilities were a major reason for noncompletion of the *Understanding MS* MOOC by people living with MS [65]. However, it is not clear why employment was associated with lower standard care course completion but not intervention course completion. Perhaps participants were more engaged with the novel content of the intervention course compared with that of the standard care course, or other factors such as the greater engagement with the community forum observed for intervention course versus standard care course participants may have played a role in greater intervention course completion among employed participants. Nevertheless, as time limitations are a commonly reported reason for noncompletion of web-based interventions [64,65], appropriate measures, such as increased time to complete web-based courses, could be implemented to increase engagement by people living with MS, especially those who are employed. Furthermore, many participants who were interviewed a month after course completion mentioned that they found it difficult to find time to complete the course, with some commenting that more user-friendly platforms such as mobile devices rather than sitting at a computer would be helpful (data not shown).

Greater self-efficacy was found to be associated with lower intervention course completion. This was unexpected considering that self-efficacy is associated with commencement and maintenance of healthy behaviors, including healthy eating, physical activity, and smoking cessation [66], and increased engagement with web-based lifestyle interventions [67]. The quantitative impact of completing the intervention course on self-efficacy levels remains to be assessed. However, qualitative insights from intervention course participants who completed it suggested that participants sought information for the purposes of enhancing self-management [57], resulting in perceived improvements in self-efficacy [58]. One possible explanation is that participants with higher baseline self-efficacy levels felt necessarily equipped to independently initiate and sustain lifestyle change, whereas others persisted to improve

XSL•FO
RenderX

self-efficacy and were reinforced by noticeable improvements in self-efficacy.

Participants adhering to an MS-specific diet program had a 19% higher intervention course completion, and those practicing meditation had 20% higher standard care course completion. While not consistent across study arms, these results may suggest that people living with MS adopting a healthy lifestyle may have a greater interest or desire to learn lifestyle-related knowledge and, therefore, were more likely to complete the web-based course. Conversely, participants with a less healthy lifestyle or those lacking motivation have been shown to lose interest in multimodal lifestyle interventions [68].

Suggestions for Future Research

Our findings underscore the importance of considering individual differences in end-user characteristics when developing and delivering web-based interventions, which may be relevant in the development of interventions for other chronic illnesses. In particular, we found that people aged >45 years were more likely to complete the web-based intervention, emphasizing the importance of developing programs that are appropriate in format (eg, videos and text) and content across different ages. As we found that higher social support was associated with increased commencement, consistent with other studies [54], providing greater support for end users may increase commencement and adherence to web-based interventions, for instance, including a facilitator-run community forum component nested within the intervention, as with the MSOC. Similarly, our findings (associations between reduced QoL and reduced commencement), in line with a recent systematic review [59], emphasize the need to consider the health and well-being of the end-user population to increase engagement. While this may be difficult to address, including trained support staff such as mental health professionals or clinicians may help participants seek help or resources, which in turn may have the cascade effect of both improving well-being and increasing engagement. This is especially relevant for people living with MS and other chronic conditions such as cardiovascular disease and cancer due to the high prevalence of depression and anxiety among these populations [69-71]. Importantly, our findings highlight the need for future studies to examine strategies to overcome factors influencing low commencement and adherence.

Strengths and Limitations

Technological elements (eg, interactivity and multimedia components) of digital engagement by people living with MS have previously been examined through systematic reviews and meta-analyses [28,30], and barriers to MOOC participation were recently evaluated in a mixed methods study [65]. However, this study provides an extensive evaluation of the relationship between participant characteristics and the different stages of MSOC engagement and of the factors influencing the commencement and completion of a web-based lifestyle modification course across both study arms, which has not been previously reported.

However, there are study limitations that need to be outlined. First, as the results are based on self-reported baseline data, there is the potential for recall bias. Second, the possibility of selection bias exists as participants were possibly more motivated or interested in learning about lifestyle-related risk factors than the general MS population. Third, the generalizability of the study findings may be further limited as the study cohort specifically comprised individuals who had completed the extensive 166-question baseline survey, indicating that they were likely to be even more highly motivated than the general MS population. Fourth, the cohort were necessarily English speaking and predominantly resided in Westernized countries, so cultural differences could not be explored due to the limited sample sizes from other countries. Fifth, some of the baseline characteristics, such as meditation practice, were queried in a dichotomous manner, which may reduce the usefulness of the study findings and their implications. However, certain questions had to be restricted to avoid questionnaire burden and potentially increase study dropout. Sixth, the adjusted analysis may not control for all potential confounders, and the dichotomous nature of some of the variables could mean that adjustment was incomplete. Consequently, it is important to note that confounding may still be present. As a follow-up to this, as our results represent an exploratory analysis complementary to the primary and secondary aims of the ancillary RCT, we advise caution when interpreting the study findings given the potential for incomplete adjustment. This is especially relevant for significant variables for which the lower boundary of the 95% CI of the adjusted prevalence ratio was close to 1.

Conclusions

This study identified specific participant characteristics associated with different stages of MSOC engagement. Factors associated with course commencement included a university education and having greater perceived support and greater mental and physical QoL. Factors associated with course completion included older age, being male, and adherence to a diet program. Improved commencement rates may be obtained with shorter initial data collection surveys to decrease potential barriers to commencement depending on the aims of the study. Other potential interventions to enhance completion include ensuring adequate time for completion to address fatigue and health-related issues and the provision of technical support to participants throughout the RCT. Involving other people such as family members to complete web-based learning programs alongside people living with MS could also provide further support and facilitate course completion. Collectively, the study findings provide practical considerations for the future design, development, and implementation of digital lifestyle interventions for people living with MS. The findings also highlight the need for further quantitative and qualitative studies to provide greater depth of understanding into digital health engagement by people living with MS.

XSL•FO
RenderX

Acknowledgments

The authors gratefully acknowledge the generous contribution of the individuals currently participating in the Multiple Sclerosis Online Course randomized controlled trial. They thank their industry partners, JMAcreative, who created the web platform and developed content, and the Overcoming Multiple Sclerosis charity (United Kingdom), who financially supported the technical development of the Multiple Sclerosis Online Course via JMAcreative. This work was supported by the Neuroepidemiology Unit at the University of Melbourne, which is funded by philanthropic donors. The Neuroepidemiology Unit also receives funding from anonymous philanthropic donors. The funders had no role in the study design, analysis, interpretation, writing of the manuscript, or in the decision to publish the results.

Data Availability

The data set generated and analyzed during this study are in a reidentifiable format. This data set is not publicly due to the study currently being in progress but the data set will be available from the corresponding author on reasonable request at the completion of the study.

Authors' Contributions

JR contributed to study design, investigation, and project administration; JR and MY conceptualized the study and contributed to methodology, writing—original draft, and writing—review and editing. MY performed data curation and formal analysis. SSY and SN contributed to methodology, formal analysis, and writing—review and editing. WB and RD were responsible for investigation and writing—review and editing. GJ acquired the funding, and SN contributed to study design. GJ and SN performed the investigation and contributed to writing—review and editing. All authors contributed to drafting and revising of the manuscript and agreed to the published version of the manuscript.

Conflicts of Interest

GJ receives royalties for his books, *Overcoming Multiple Sclerosis and Recovering from Multiple Sclerosis*. GJ and SN receive royalties for their book, *The Overcoming MS Handbook*, and previously received remuneration for conducting educational workshops for people with multiple sclerosis.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1237 KB - jmir_v26i1e58253_app1.pdf]

Multimedia Appendix 2
Outline of the intervention and standard care course format.
[DOCX File , 15 KB - jmir_v26i1e58253_app2.docx]

Multimedia Appendix 3
Participants' country of residence.
[DOCX File , 16 KB - jmir_v26i1e58253_app3.docx]

Multimedia Appendix 4
Disease-modifying and prescription medications taken by Multiple Sclerosis Online Course participants.
[DOCX File , 20 KB - jmir_v26i1e58253_app4.docx]

References

1. Fitzgerald KC, Tyry T, Salter A, Cofield SS, Cutter G, Fox R, et al. Diet quality is associated with disability and symptom severity in multiple sclerosis. Neurology 2018 Jan 02;90(1):e1-11. [doi: 10.1212/WNL.0000000000004768] [Medline: 29212827]
2. Katz Sand I, Levy S, Fitzgerald K, Sorets T, Sumowski JF. Mediterranean diet is linked to less objective disability in multiple sclerosis. Mult Scler 2023 Feb 13;29(2):248-260 [FREE Full text] [doi: 10.1177/13524585221127414] [Medline: 36226971]
3. Simpson-Yap S, Neate SL, Nag N, Probst YC, Yu M, Jelinek GA, et al. Longitudinal associations between quality of diet and disability over 7.5 years in an international sample of people with multiple sclerosis. Eur J Neurol 2023 Oct 20;30(10):3200-3211. [doi: 10.1111/ene.15980] [Medline: 37433564]
4. Motl RW, Mowry EM, Ehde DM, LaRocca NG, Smith KE, Costello K, et al. Wellness and multiple sclerosis: the national MS society establishes a wellness research working group and research priorities. Mult Scler 2018 Mar 12;24(3):262-267 [FREE Full text] [doi: 10.1177/1352458516687404] [Medline: 28080254]

XSL•FO
RenderX

5. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. Ann Neurol 2001 Jul;50(1):121-127. [doi: 10.1002/ana.1032] [Medline: 11456302]
6. Hadgkiss EJ, Jelinek GA, Weiland TJ, Pereira NG, Marck CH, van der Meer DM. The association of diet with quality of life, disability, and relapse rate in an international sample of people with multiple sclerosis. Nutr Neurosci 2015 Apr 17;18(3):125-136 [FREE Full text] [doi: 10.1179/1476830514Y.0000000117] [Medline: 24628020]
7. Jelinek GA, Hadgkiss EJ, Weiland TJ, Pereira NG, Marck CH, van der Meer DM. Association of fish consumption and Ω 3 supplementation with quality of life, disability and disease activity in an international cohort of people with multiple sclerosis. Int J Neurosci 2013 Nov 03;123(11):792-800 [FREE Full text] [doi: 10.3109/00207454.2013.803104] [Medline: 23713615]
8. Marck CH, Hadgkiss EJ, Weiland TJ, van der Meer DM, Pereira NG, Jelinek GA. Physical activity and associated levels of disability and quality of life in people with multiple sclerosis: a large international survey. BMC Neurol 2014 Jul 12;14(1):143 [FREE Full text] [doi: 10.1186/1471-2377-14-143] [Medline: 25016312]
9. Simpson-Yap S, Nag N, Probst Y, Jelinek G, Neate S. Higher-quality diet and non-consumption of meat are associated with less self-determined disability progression in people with multiple sclerosis: a longitudinal cohort study. Eur J Neurol 2022 Jan 26;29(1):225-236 [FREE Full text] [doi: 10.1111/ene.15066] [Medline: 34390078]
10. Simpson-Yap S, Nag N, Probst Y, Reece JC, Jelinek GA, Neate S. Prospective associations of better quality of the diet with improved quality of life over 7.5 years in people with multiple sclerosis. Mult Scler Relat Disord 2022 Apr;60:103710 [FREE Full text] [doi: 10.1016/j.msard.2022.103710] [Medline: 35219239]
11. Simpson-Yap S, Oddy WH, Taylor B, Lucas RM, Black LJ, Ponsonby A, et al. High Prudent diet factor score predicts lower relapse hazard in early multiple sclerosis. Mult Scler 2021 Jun 23;27(7):1112-1124. [doi: 10.1177/1352458520943087] [Medline: 32701031]
12. Weiland TJ, Nag N, de Livera A, Jelinek GA, Neate SL, Bevens W, et al. Two healthy lifestyle scores are associated with lower subsequent fatigue risk using inverse probability weighting in an international longitudinal cohort of people with multiple sclerosis. Eur J Neurol 2021 Sep 13;28(9):2952-2964. [doi: 10.1111/ene.14956] [Medline: 34081818]
13. Jelinek GA. Overcoming Multiple Sclerosis: The Evidence-based 7 Step Recovery Program. 2nd edition. Crows Nest, Australia: Allen & Unwin; 2065.
14. Hadgkiss EJ, Jelinek GA, Weiland TJ, Rumbold G, Mackinlay CA, Gutbrod S, et al. Health-related quality of life outcomes at 1 and 5 years after a residential retreat promoting lifestyle modification for people with multiple sclerosis. Neurol Sci 2013 Feb 25;34(2):187-195 [FREE Full text] [doi: 10.1007/s10072-012-0982-4] [Medline: 22367222]
15. Li MP, Jelinek GA, Weiland TJ, Mackinlay CA, Dye S, Gawler I. Effect of a residential retreat promoting lifestyle modifications on health-related quality of life in people with multiple sclerosis. Qual Prim Care 2010;18(6):379-389. [Medline: 21294980]
16. Marck CH, de Livera AM, Brown CR, Neate SL, Taylor KL, Weiland TJ, et al. Health outcomes and adherence to a healthy lifestyle after a multimodal intervention in people with multiple sclerosis: three year follow-up. PLoS One 2018 May 23;13(5):e0197759 [FREE Full text] [doi: 10.1371/journal.pone.0197759] [Medline: 29791509]
17. Mayo CD, Farzam-Kia N, Ghahari S. Identifying barriers to and facilitators of health service access encountered by individuals with multiple sclerosis. Int J MS Care 2021;23(1):37-44 [FREE Full text] [doi: 10.7224/1537-2073.2020-026] [Medline: 33658905]
18. Bessing B, van der Mei I, Taylor BV, Honan CA, Blizzard L, Claflin SB. Evaluating the impact of the understanding multiple sclerosis online course on participant MS knowledge, health literacy, resilience, self-efficacy, quality of life, and MS symptom severity. Mult Scler Relat Disord 2022 Apr;60:103717. [doi: 10.1016/j.msard.2022.103717] [Medline: 35259682]
19. Bock L, Rana M, Westemeyer L, Rana M. Evaluation of a four-week online resilience training program for multiple sclerosis patients. Healthcare (Basel) 2024 Mar 09;12(6):620 [FREE Full text] [doi: 10.3390/healthcare12060620] [Medline: 38540584]
20. Motl RW, Hubbard EA, Bollaert RE, Adamson BC, Kinnett-Hopkins D, Balto JM, et al. Randomized controlled trial of an e-learning designed behavioral intervention for increasing physical activity behavior in multiple sclerosis. Mult Scler J Exp Transl Clin 2017 Oct 09;3(4):2055217317734886 [FREE Full text] [doi: 10.1177/2055217317734886] [Medline: 29051831]
21. Motl RW, Sandroff BM. Randomized controlled trial of physical activity intervention effects on fatigue and depression in multiple sclerosis: secondary analysis of data from persons with elevated symptom status. Contemp Clin Trials Commun 2020 Mar;17:100521 [FREE Full text] [doi: 10.1016/j.conctc.2020.100521] [Medline: 32211556]
22. van Kessel K, Wouldes T, Moss-Morris R. A New Zealand pilot randomized controlled trial of a web-based interactive self-management programme (MSInvigor8) with and without email support for the treatment of multiple sclerosis fatigue. Clin Rehabil 2016 May 07;30(5):454-462. [doi: 10.1177/0269215515584800] [Medline: 25952587]
23. Cavalera C, Rovaris M, Mendozzi L, Pugnetti L, Garegnani M, Castelnuovo G, et al. Online meditation training for people with multiple sclerosis: a randomized controlled trial. Mult Scler 2019 Apr 27;25(4):610-617. [doi: 10.1177/1352458518761187] [Medline: 29485319]

XSL•FO
RenderX

24. Claflin SB, Mainsbridge C, Campbell J, Klekociuk S, Taylor BV. Self-reported behaviour change among multiple sclerosis community members and interested laypeople following participation in a free online course about multiple sclerosis. Health Promot J Austr 2022 Jul 13;33(3):768-778. [doi: 10.1002/hpja.559] [Medline: 34807490]
25. Kutzinski M, Krause N, Riemann-Lorenz K, Meyer B, Heesen C. Acceptability of a digital health application to empower persons with multiple sclerosis with moderate to severe disability: single-arm prospective pilot study. BMC Neurol 2023 Oct 23;23(1):382 [FREE Full text] [doi: 10.1186/s12883-023-03434-w] [Medline: 37872471]
26. Russell RD, Black LJ, Purdue J, Daly A, Begley A. A collaborative approach to designing an online nutrition education program for people with multiple sclerosis. Disabil Rehabil 2024 Mar 12;46(5):947-956. [doi: 10.1080/09638288.2023.2186499] [Medline: 36908015]
27. Waller R, Gilbody S. Barriers to the uptake of computerized cognitive behavioural therapy: a systematic review of the quantitative and qualitative evidence. Psychol Med 2009 May 24;39(5):705-712. [doi: 10.1017/S0033291708004224] [Medline: 18812006]
28. Bevens W, Weiland T, Gray K, Jelinek G, Neate S, Simpson-Yap S. Attrition within digital health interventions for people with multiple sclerosis: systematic review and meta-analysis. J Med Internet Res 2022 Feb 09;24(2):e27735 [FREE Full text] [doi: 10.2196/27735] [Medline: 35138262]
29. Claflin SB, Campbell JA, Doherty K, Farrow M, Bessing B, Taylor BV. Evaluating course completion, appropriateness, and burden in the understanding multiple sclerosis massive open online course: cohort study. J Med Internet Res 2021 Dec 07;23(12):e21681 [FREE Full text] [doi: 10.2196/21681] [Medline: 34878985]
30. Arafah AM, Bouchard V, Mayo NE. Enrolling and keeping participants in multiple sclerosis self-management interventions: a systematic review and meta-analysis. Clin Rehabil 2017 Jun 08;31(6):809-823. [doi: 10.1177/0269215516658338] [Medline: 27401492]
31. Marrie RA, Leung S, Tyry T, Cutter GR, Fox R, Salter A. Use of eHealth and mHealth technology by persons with multiple sclerosis. Mult Scler Relat Disord 2019 Jan;27:13-19. [doi: 10.1016/j.msard.2018.09.036] [Medline: 30296732]
32. Bevens W, Davenport R, Neate S, Yu M, Jelinek P, Jelinek GA, et al. Web-based health information seeking by people living with multiple sclerosis: qualitative investigation of the multiple sclerosis online course. J Med Internet Res 2024 Feb 09;26:e53372 [FREE Full text] [doi: 10.2196/53372] [Medline: 38335016]
33. Reece JC, Yu M, Bevens W, Simpson-Yap S, Jelinek G, Jelinek P, et al. Study protocol for an online lifestyle modification education course for people living with multiple sclerosis: the multiple sclerosis online course (MSOC). BMC Neurol 2023 Jun 29;23(1):249 [FREE Full text] [doi: 10.1186/s12883-023-03298-0] [Medline: 37386385]
34. Bevens W, Weiland TJ, Gray K, Neate SL, Nag N, Simpson-Yap S, et al. The feasibility of a web-based educational lifestyle program for people with multiple sclerosis: a randomized controlled trial. Front Public Health 2022 Apr 27;10:852214 [FREE Full text] [doi: 10.3389/fpubh.2022.852214] [Medline: 35570898]
35. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The multidimensional scale of perceived social support. J Pers Assess 1988 Mar 08;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
36. Body mass index - BMI. World Health Organization. 2021. URL: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi [accessed 2024-04-29]
37. Sangha O, Stucki G, Liang MH, Fossel AH, Katz JN. The self-administered comorbidity questionnaire: a new method to assess comorbidity for clinical and health services research. Arthritis Rheum 2003 Apr 15;49(2):156-163 [FREE Full text] [doi: 10.1002/art.10993] [Medline: 12687505]
38. Hohol MJ, Orav EJ, Weiner HL. Disease steps in multiple sclerosis: a simple approach to evaluate disease progression. Neurology 1995 Feb;45(2):251-255. [doi: 10.1212/wnl.45.2.251] [Medline: 7854521]
39. Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D. Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. BMC Neurol 2013 Apr 25;13(1):37 [FREE Full text] [doi: 10.1186/1471-2377-13-37] [Medline: 23617555]
40. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989 Oct 01;46(10):1121-1123. [doi: 10.1001/archneur.1989.00520460115022] [Medline: 2803071]
41. Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. Qual Life Res 1995 Jun;4(3):187-206. [doi: 10.1007/BF02260859] [Medline: 7613530]
42. Amtmann D, Bamer AM, Cook KF, Askew RL, Noonan VK, Brockway JA. University of Washington self-efficacy scale: a new self-efficacy scale for people with disabilities. Arch Phys Med Rehabil 2012 Oct;93(10):1757-1765. [doi: 10.1016/j.apmr.2012.05.001] [Medline: 22575393]
43. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
44. Guan V, Simpson-Yap S, Nag N, Jelinek G, Neate S, Probst Y. Using online 24-h dietary methodology to validate the psychometric properties of a dietary scoring tool with an international sample of adults living with multiple sclerosis. Nutrients 2022 Oct 31;14(21):4568 [FREE Full text] [doi: 10.3390/nu14214568] [Medline: 36364830]
45. McKellar S, Horsley P, Chambers R, Bauer J, Vendersee P, Clarke C, et al. Development of the diet habits questionnaire for use in cardiac rehabilitation. Austr J Prim Health 2008;14(3):43-47. [doi: 10.1071/py08035]

XSL•FO
RenderX

46. Kasser SL, Kosma MJ. Health beliefs and physical activity behavior in adults with multiple sclerosis. Disabil Health J 2012 Oct;5(4):261-268. [doi: 10.1016/j.dhjo.2012.07.001] [Medline: 23021737]
47. Kirkland H, Campbell J, Reece J, Nag N, Probst Y, Neate S, et al. Higher diet quality is associated with short and long-term benefits on SF-6D health state utilities: a 5-year cohort study in an international sample of people with multiple sclerosis. Qual Life Res 2023 Jul 23;32(7):1883-1896 [FREE Full text] [doi: 10.1007/s11136-023-03361-w] [Medline: 36821020]
48. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
49. Silveira SL, Motl RW, Marquez DX, Lancia S, Salter A. Physical activity as a correlate of symptoms, quality of life, comorbidity, and disability status in Hispanics with multiple sclerosis. Disabil Health J 2023 Apr;16(2):101398. [doi: 10.1016/j.dhjo.2022.101398] [Medline: 36402726]
50. Dlugonski D, Motl RW, McAuley E. Increasing physical activity in multiple sclerosis: replicating internet intervention effects using objective and self-report outcomes. J Rehabil Res Dev 2011;48(9):1129-1136 [FREE Full text] [doi: 10.1682/jrrd.2010.09.0192] [Medline: 22234717]
51. Hadgkiss EJ, Jelinek GA, Weiland TJ, Pereira NG, Marck CH, van der Meer DM. Methodology of an international study of people with multiple sclerosis recruited through web 2.0 platforms: demographics, lifestyle, and disease characteristics. Neurol Res Int 2013;2013:580596-580512 [FREE Full text] [doi: 10.1155/2013/580596] [Medline: 23691313]
52. Brall C, Berlin C, Zwahlen M, Ormond KE, Egger M, Vayena E. Public willingness to participate in personalized health research and biobanking: a large-scale Swiss survey. PLoS One 2021 Apr 1;16(4):e0249141 [FREE Full text] [doi: 10.1371/journal.pone.0249141] [Medline: 33793624]
53. Altaf A, Bokhari R, Enani G, Judeeba S, Hemdi A, Maghrabi A, et al. Patients' attitudes and knowledge toward clinical trial participation. Saudi Surg J 2019;7(2):69. [doi: 10.4103/ssj.ssj_23_19]
54. Porter KJ, Thomson JL, Zoellner JM. Predictors of engagement and outcome achievement in a behavioural intervention targeting sugar-sweetened beverage intake among rural adults. Public Health Nutr 2020 Feb 04;23(3):554-563 [FREE Full text] [doi: 10.1017/S1368980019003392] [Medline: 31796129]
55. Barrot JS, Llenares II, Del Rosario LS. Students' online learning challenges during the pandemic and how they cope with them: the case of the Philippines. Educ Inf Technol (Dordr) 2021 May 28;26(6):7321-7338 [FREE Full text] [doi: 10.1007/s10639-021-10589-x] [Medline: 34075300]
56. Park JH, Choi HJ. Factors influencing adult learners' decision to drop out or persist in online learning. Educ Technol Soc 2009;12(4):207-217 [FREE Full text]
57. Bentley CL, Powell L, Potter S, Parker J, Mountain GA, Bartlett YK, et al. The use of a smartphone app and an activity tracker to promote physical activity in the management of chronic obstructive pulmonary disease: randomized controlled feasibility study. JMIR Mhealth Uhealth 2020 Jun 03;8(6):e16203 [FREE Full text] [doi: 10.2196/16203] [Medline: 32490838]
58. Burgess E, Hassmén P, Pumpa KL. Determinants of adherence to lifestyle intervention in adults with obesity: a systematic review. Clin Obes 2017 Jun 15;7(3):123-135. [doi: 10.1111/cob.12183] [Medline: 28296261]
59. Borghouts J, Eikey E, Mark G, de Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
60. Brouwer W, Oenema A, Raat H, Crutzen R, de Nooijer J, de Vries NK, et al. Characteristics of visitors and revisitors to an internet-delivered computer-tailored lifestyle intervention implemented for use by the general public. Health Educ Res 2010 Aug 06;25(4):585-595. [doi: 10.1093/her/cyp063] [Medline: 19897515]
61. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015 Oct 17;45(13):2717-2726 [FREE Full text] [doi: 10.1017/S0033291715000665] [Medline: 25881626]
62. Haase KR, Cosco T, Kervin L, Riadi I, O'Connell ME. Older adults' experiences with using technology for socialization during the COVID-19 pandemic: cross-sectional survey study. JMIR Aging 2021 Apr 23;4(2):e28010 [FREE Full text] [doi: 10.2196/28010] [Medline: 33739929]
63. Reece JC, Neate SL, Davenport RA, Milanzi E, Nag N, Bevens W, et al. Stressful life events and depression and fatigue in people with multiple sclerosis: a cross-sectional analysis of an international cohort. Acta Neurol Belg 2024 Apr 28;124(2):457-466 [FREE Full text] [doi: 10.1007/s13760-023-02390-z] [Medline: 37770791]
64. Sexton JB, Adair KC. Forty-five good things: a prospective pilot study of the Three Good Things well-being intervention in the USA for healthcare worker emotional exhaustion, depression, work-life balance and happiness. BMJ Open 2019 Mar 20;9(3):e022695 [FREE Full text] [doi: 10.1136/bmjopen-2018-022695] [Medline: 30898795]
65. Claflin S, Klekociuk S, Campbell J, Taylor B. Reasons for non-completion of a massive open online course about multiple sclerosis: a mixed methods study. Mult Scler Relat Disord 2022 Nov;67:104092. [doi: 10.1016/j.msard.2022.104092] [Medline: 35977440]

XSL•FO
RenderX

66. Newby K, Teah G, Cooke R, Li X, Brown K, Salisbury-Finch B, et al. Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychol Rev 2021 Mar 20;15(1):140-158. [doi: 10.1080/17437199.2019.1705873] [Medline: 31847702]
67. Miyamoto S, Henderson S, Young H, Ward D, Santillan V. Recruiting rural participants for a telehealth intervention on diabetes self-management. J Rural Health 2013 Oct 25;29(1):69-77 [FREE Full text] [doi: 10.1111/j.1748-0361.2012.00443.x] [Medline: 23289657]
68. Wilson K, Senay I, Durantini M, Sánchez F, Hennessy M, Spring B, et al. When it comes to lifestyle recommendations, more is sometimes less: a meta-analysis of theoretical assumptions underlying the effectiveness of interventions promoting multiple behavior domain change. Psychol Bull 2015 Mar;141(2):474-509 [FREE Full text] [doi: 10.1037/a0038295] [Medline: 25528345]
69. Beatty L, Kemp E, Binnion C, Turner J, Milne D, Butow P, et al. Uptake and adherence to an online intervention for cancer-related distress: older age is not a barrier to adherence but may be a barrier to uptake. Support Care Cancer 2017 Jun 2;25(6):1905-1914. [doi: 10.1007/s00520-017-3591-1] [Medline: 28155018]
70. Wallin E, Norlund F, Olsson EM, Burell G, Held C, Carlsson T. Treatment activity, user satisfaction, and experienced usability of internet-based cognitive behavioral therapy for adults with depression and anxiety after a myocardial infarction: mixed-methods study. J Med Internet Res 2018 Mar 16;20(3):e87 [FREE Full text] [doi: 10.2196/jmir.9690] [Medline: 29549067]
71. Peres DS, Rodrigues P, Viero FT, Frare JM, Kudsi SQ, Meira GM, et al. Prevalence of depression and anxiety in the different clinical forms of multiple sclerosis and associations with disability: a systematic review and meta-analysis. Brain Behav Immun Health 2022 Oct;24:100484 [FREE Full text] [doi: 10.1016/j.bbih.2022.100484] [Medline: 35856061]

Abbreviations

CONSORT-R: Consolidated Standards of Reporting Trials–Routine
DHQ: Diet Habits Questionnaire
DMT: disease-modifying therapy
IPAQ: International Physical Activity Questionnaire
MOOC: massive open online course
MS: multiple sclerosis
MSOC: Multiple Sclerosis Online Course
OMS: Overcoming Multiple Sclerosis
QoL: quality of life
RCT: randomized controlled trial
UWSES: University of Washington Self-Efficacy Scale

Edited by N Cahill; submitted 11.03.24; peer-reviewed by R Haase, D Ruijter; comments to author 27.03.24; revised version received 10.04.24; accepted 02.07.24; published 28.08.24.

Please cite as:
Reece J, Yu M, Bevens W, Simpson-Yap S, Davenport R, Jelinek G, Neate S
Sociodemographic, Health, and Lifestyle-Related Characteristics Associated With the Commencement and Completion of a Web-Based Lifestyle Educational Program for People With Multiple Sclerosis: Randomized Controlled Trial
J Med Internet Res 2024;26:e58253
URL: https://www.jmir.org/2024/1/e58253
doi:10.2196/58253
PMID:

©Jeanette Reece, Maggie Yu, William Bevens, Steve Simpson-Yap, Rebekah Davenport, George Jelinek, Sandra Neate. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Web-Based Coping Skills Training and Coach Support for Women Living With a Partner With an Alcohol Use Disorder: Randomized Controlled Trial

Robert G Rychtarik[1], PhD; Brian G Danaher[2], PhD; Neil B McGillicuddy[3], PhD; Milagra S Tyler[2], MA; Christopher Barrick[4], PhD; Florence Leong[3], PhD; Derek B Kosty[2], PhD

[1]Department of Psychiatry and Clinical and Research Institute on Addictions, University at Buffalo, The State University of New York, Buffalo, NY, United States

[2]Prevention Science Institute, University of Oregon, Eugene, OR, United States

[3]Clinical and Research Institute on Addictions, University at Buffalo, The State University of New York, Buffalo, NY, United States

[4]School of Nursing, University at Buffalo, The State University of New York, Buffalo, NY, United States

Corresponding Author:
Robert G Rychtarik, PhD
Department of Psychiatry and Clinical and Research Institute on Addictions
University at Buffalo, The State University of New York
1021 Main Street
Buffalo, NY, 14203
United States
Phone: 1 7168872591
Email: rgr@buffalo.edu

Abstract

Background: Individuals living with a partner with an alcohol use disorder (AUD) can experience significant psychological distress and use health care more than those without a partner with an AUD. However, the prevailing treatment system's focus on the partner and personal barriers limit these individuals from getting help for themselves. Preliminary work on a self-directed, web-based coping skills training program, *Stop Spinning My Wheels* (SSMW), shows promise in broadening available treatments for this population. In this study, we conducted a robust evaluation of SSMW primary outcomes.

Objective: The study aims to test whether women with a partner with an AUD assigned to SSMW experienced a greater reduction in negative affect (depression and anger) (1) than a usual web care (UWC) control and (2) with brief phone coach support (SSMW+coach) rather than without (SSMW only) and (3) whether baseline negative affect moderated treatment effects.

Methods: Women (mean age 45.7, SD 10.8 years; Black: 17/456, 3.7%; White: 408/456, 89.5%) were randomized to SSMW only, SSMW+coach, or UWC. Depression (Beck Depression Inventory–II) and anger (State-Trait Anger Expression Inventory 2–State Anger) were assessed at baseline, 12-week posttest, and 6- and 12-month follow-ups.

Results: Participants in all conditions decreased in depression from baseline to posttest and from baseline to follow-up; SSMW-only and SSMW+coach participants decreased in anger, but UWC participants did not. Compared to UWC participants, SSMW-only participants experienced greater anger reduction (P=.03), and SSMW+coach participants experienced a greater reduction in depression (P<.001) from baseline to posttest. However, from baseline to follow-up, only a greater, but not statistically significant (P=.052), reduction in anger occurred in SSMW+coach compared to UWC. Although the SSMW conditions did not differ from each other in negative affect outcomes (P=.06-.57), SSMW+coach had higher program engagement and satisfaction (all P<.004). Baseline negative affect did not moderate effects, although remission from baseline clinically relevant depressive symptoms (Beck Depression Inventory≥14) was higher in SSMW only (33/67, 49%; odds ratio 2.13, 95% CI 1.05-4.30; P=.03) and SSMW+coach (46/74, 62%; odds ratio 3.60, 95% CI 1.79-7.23; P<.001) than in UWC (21/67, 31%); remission rates did not differ between the SSMW conditions (P=.12).

Conclusions: The results partially supported the hypotheses. The SSMW conditions had earlier effects than UWC, but positive change in UWC mitigated the hypothesized long-term SSMW-UWC differences. The results highlight the importance of incorporating active controls in web-based clinical trials. Although SSMW+coach showed benefits over SSMW only on engagement and satisfaction measures and in the number needed to treat (5.6 for SSMW only; 3.2 for SSMW+coach), the SSMW conditions

XSL•FO
RenderX

were comparable and superior to UWC on depressive symptom remission levels. Overall, SSMW with or without a coach can reduce clinically meaningful distress and add to available treatment options for this large, underserved group.

Trial Registration: ClinicalTrials.gov NCT02984241; https://www.clinicaltrials.gov/study/NCT02984241

(J Med Internet Res 2024;26:e56119) doi:10.2196/56119

KEYWORDS

alcohol use disorder; AUD; spouses; coping skills training; eHealth; coaching; telehealth; web-based; women; randomized controlled trial; RCT; mobile phone

Introduction

Background

Individuals living with a partner with an alcohol use disorder (AUD) are at significantly greater risk of experiencing physical or emotional harm from the partner's drinking. Women are more likely than men to report such harm [1], and estimates suggest that 5% of adult women in the United States are married to or living with a partner with an AUD [2]. Spouses with (vs without) a partner with a substance use disorder (SUD) are at least 2 times as likely to experience depression, anxiety, and stress-related disorders [3] and have more social adjustment problems [4], subclinical disorders [5], and health care use [3,6-11] and higher health care costs [8]. While early views interpreted the psychological problems experienced by this population as indicative of personality and characterological issues, contemporary family stress models view it as indicative of normal individuals struggling to cope with the disorder in their partner or family [12,13]. Consistent with this latter view, much of the distress experienced particularly among those without their own SUD appears directly related to stress, burden, or the extent of problems brought on by the partner's SUD [12,14,15].

In this research report, we refer to *partner* as the person with the AUD; *spouse* refers to the individual married to or living with the person with the AUD. Daley et al [16] note that the "major health, social, and safety problem [of SUD] cannot be effectively addressed without considering the impact of SUDs on families and members, including children, and including them in treatment and recovery. While addiction is promoted as a 'family disease,' in reality many family members are not offered the opportunity to engage in treatment for their own health." A small but growing body of research suggests that spouses of partners with AUD can benefit from professional, therapist-delivered treatments designed to relieve their own personal distress [17,18]. However, institutional barriers (eg, systemic emphasis on the partner's treatment) and personal barriers (eg, spouse preoccupation with only the partner's problem, fears of partner retribution, stigmatization, and costs of clinical care) limit the overall accessibility of these face-to-face treatments. Self-help groups, mainly Al-Anon, can be helpful for this group, but Al-Anon may not be readily available, and dropout can be high [19]. Merkouris et al [17] recommended developing and evaluating a broader range of treatment modalities and intensities for spouses and caregivers, particularly self-directed, web-based treatment formats.

Only 2 studies, both pilots, of web-based treatments for spouses have been reported in the literature. Osilla et al [20] piloted a web adaptation of the Community Reinforcement and Family Training program that targeted spouses whose partners were military service members and veterans. Compared to a waitlist control condition, program participants reported significantly reduced anxiety and greater social support but no significant reduction in depressive symptoms or anger. In a feasibility pilot study, Rychtarik et al [21] conducted a randomized controlled trial (RCT), in which 89 women with a partner with an AUD were assigned to either (1) an early version of the web-based *Stop Spinning My Wheels* (SSMW) program, a coach-assisted eHealth adaptation of an empirically tested face-to-face coping skills training program focusing on spouse functioning [5]; or (2) a waitlist control. The results of an 8-week assessment of SSMW showed that women participated in their web sessions and were very satisfied with the overall program, although relatively few opted to use the optional coach support available to them. Compared to controls, SSMW participants exhibited significantly improved coping skills (effect size=1.02), fewer depressive symptoms (effect size=–0.65), and reduced situational anger (effect size=–0.70). The conditions did not differ on other secondary outcomes (ie, anxiety, anger expression, and general stress). While results are promising, we do not know whether SSMW can achieve similar results when evaluated against an active control condition. Similarly, we do not know whether coach support can improve outcomes when compared to SSMW alone.

Objectives

In this report, we build on and address the limitations of our earlier SSMW pilot with an updated SSMW version, a credible active web control, and randomization to phone coach support. This report is limited to the study's primary treatment outcomes (ie, depression and state anger) and to testing the study's 4 primary a priori hypotheses. First, we hypothesized that the fully self-directed SSMW would result in immediate and sustained reductions in negative affect relative to an active web control. Second, we hypothesized that SSMW plus low-intensity phone coaching would result in immediate and sustained negative affect reductions relative to the active control. Third, we hypothesized that SSMW with low-intensity phone coaching would promote a greater negative affect reduction than SSMW alone. Fourth, we hypothesized that baseline negative affect would moderate SSMW outcomes such that those higher in negative affect would show a greater reduction with a coach than without. Those lower in negative affect were hypothesized to benefit equally well regardless of coach support. Similarly, we explored whether baseline negative affect moderated the

XSL•FO
RenderX

effects of SSMW overall compared to the usual web care (UWC) control condition. Hypotheses 3 and 4 are informed by the supportive accountability model [22], which posits that low-intensity human coach support, goal setting, and performance monitoring can encourage web-based program follow-through. Thus, we hypothesized that the greater SSMW program engagement and use of program recommendations with a coach would result in a greater negative affect reduction than SSMW without a coach. Similarly, we hypothesized that the greater program engagement with a coach would promote greater benefit for those high in negative affect; those with lower negative affect would benefit from either SSMW condition.

Methods

Trial Design

This study was a parallel 3-group, intention-to-treat RCT that followed the CONSORT (Consolidated Standards of Reporting Trials) guidelines [23] (Multimedia Appendix 1). The trial was registered at ClinicalTrials.gov (NCT02984241) [24].

Ethical Considerations

All study procedures were approved by the institutional review board of the University at Buffalo (STUDY00000057). Electronic and verbal consent were used in the consenting process. Participants could opt out of taking part at any time. Data presented in this paper are deidentified; only grouped summary statistics are reported. Participants were paid US $120 at week 12 for completing the interim and 12-week assessments and US $50 for each 6- and 12-month assessment.

Participants

Overview

Study participants were 456 female spouses recruited from across New York state from October 13, 2019, to February 26, 2021. Radio, internet, and other media advertisements for the SSMW program directed women experiencing stress from their partner's drinking to a research project web page that contained a study description and a link to a brief web self-screening consent form and questionnaire. Spouses who completed the self-screening assessment, were found to be preliminarily eligible, and consented to proceed to the next step in screening were able to access a secure web portal to schedule a call with study staff to complete their final screening (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram. SSMW: Stop Spinning My Wheels.

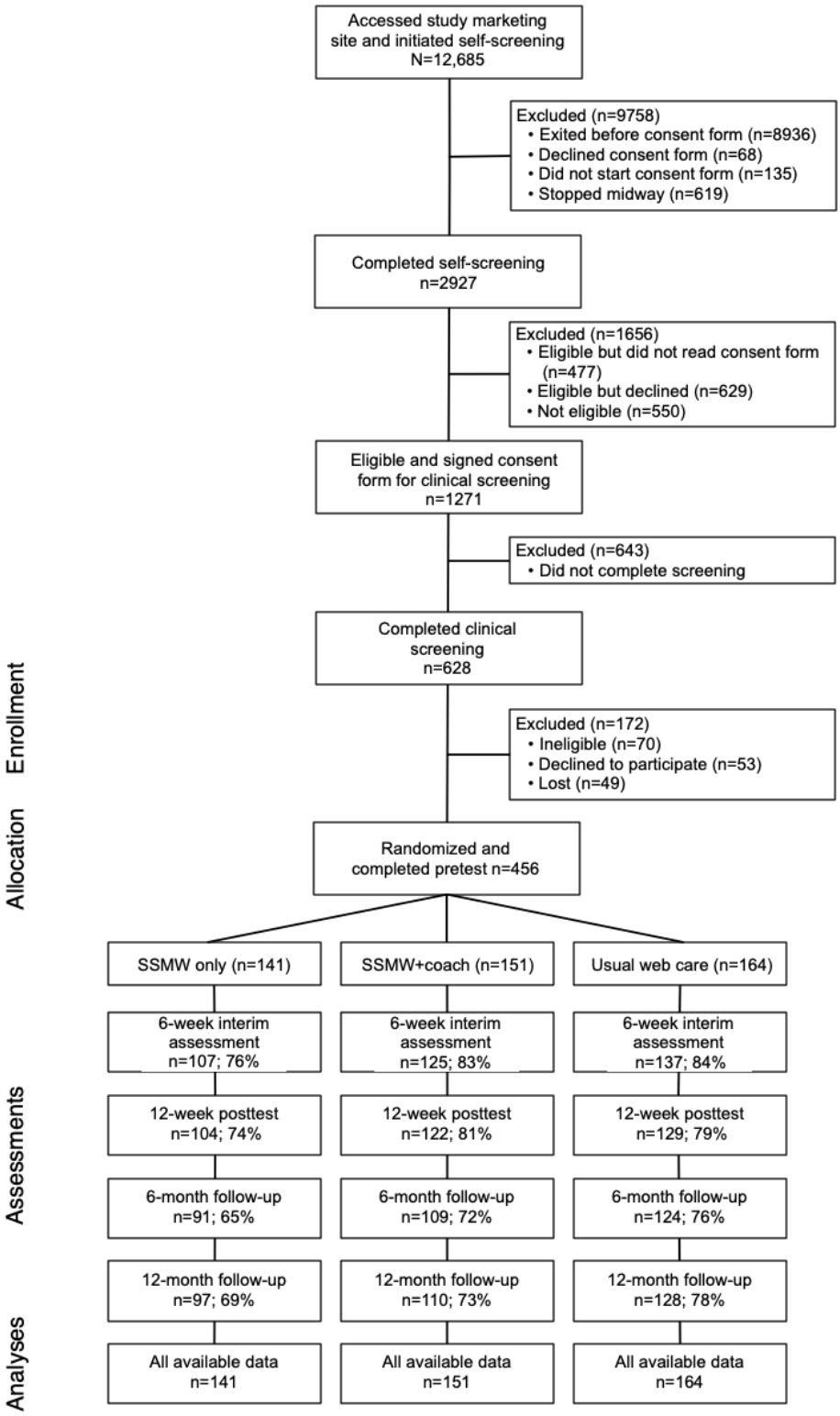

Upon completion of the final screening interview, eligible and interested individuals received log-in credentials to the study's secure main web consent form, which, when endorsed, enabled them to complete a web-administered baseline assessment and be randomized to a condition. Once randomized, participants accessed their assigned condition without charge. Throughout the screening process, individuals found ineligible or who declined to participate were provided with information on alternative treatment resources.

Recruitment Deviation

At 5 months into participant recruitment, at the onset of the COVID-19 pandemic, we paused prospective participant screenings for approximately 3 months to prepare for and accommodate study personnel's work from remote locations (eg, acquiring equipment and establishing secure remote procedures). During this time, the marketing website informed visitors of the pause in recruitment and invited them to take the web-based self-screening and, if eligible, place themselves on

XSL•FO
RenderX

a waiting list for later consideration. During this time, previously randomly assigned participants maintained access to their assigned study condition and, if assigned, coach phone support.

Eligibility

Eligibility criteria for spouses included the following: (1) age of ≥18 years; (2) residence in New York state; (3) fluency in English; (4) a woman married to or living with a male partner and, based on her report, the partner meeting criteria for past-year AUD on the Family History Assessment Module [25] and past 3-month problematic alcohol consumption, determined by a score of ≥5 on the Alcohol Use Disorders Identification Test–Consumption items [26]; (5) negative screening on the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [27] for any unremitted SUD or serious mental disorder (ie, schizophrenia or bipolar disorder) or other conditions that would interfere with participation (eg, high suicide risk); (6) no participation in previous SSMW studies; (7) negative report of having immediate fear for her life due to intimate partner violence and no experienced injuries from partner violence requiring medical attention over the previous year; and (8) internet access and a computer or tablet and smartphone.

Randomization

On completion of the baseline assessments, an urn randomization procedure [28] was used to assign each participant to 1 of 3 conditions: SSMW only, SSMW+coach, or a UWC control condition. Urn randomization was used to balance condition assignment on putative prognostic baseline variables: (1) spouse education (<12 years vs ≥12 years); (2) partner SUD treatment history (last 3 months: yes vs no); (3) spouse recent Al-Anon experience (last 3 months: yes vs no); (4) spouse psychological or mental health counseling (last 3 months: yes vs no); (5) spouse depression status (<14 vs ≥14 baseline score on the Beck Depression Inventory–II [BDI-II] [29]); (6) spouse anger status (<18 vs ≥18 baseline score on the State Anger subscale of the State-Trait Anger Expression Inventory 2 [STAXI-2] [30]); and (7) age of youngest biological, adopted, or stepchild at home (none vs 1.5-5 years vs 6-18 years). BDI-II and STAXI-2 subscale cut points correspond to mild depression and clinically significant anger, respectively [29,30]. A separate urn procedure was used to randomize participants in the SSMW+coach condition to 1 of 2 available coaches.

Procedure

Overview

Once randomized to a condition, participants were sent an automated welcoming email with a URL to create a password to access their assigned condition home page. All conditions shared a standard SSMW logo, name, and style (font and color palette) and were accessible via desktop, laptop, or tablet computer devices. Participants were encouraged to complete their assigned program during the *active treatment period*, defined as from randomization to 12 weeks. Study participants could subsequently access their web content during the follow-up period to more accurately reflect the design of many real-world internet treatments.

SSMW-Only Condition

Both SSMW conditions shared the overarching road trip theme that encouraged participants to (1) view living with a partner with an AUD as a particular challenge in one's *road trip of life*; (2) recognize that it is not possible to be completely prepared to deal with this challenge; and (3) understand that usual ways of coping with a partner's drinking problem often do not help and the failure of repeatedly trying to change a partner's behavior can cause the spouse to feel exhausted and stressed out, such as *spinning their wheels*. Similarly, the stated overall goal of both SSMW conditions was to help the spouse reduce (better manage) her stress by providing more effective skills to cope with a partner who drinks too much.

The SSMW program was an enhanced 20-session version of the self-directed SSMW program in the study by Rychtarik et al [21]. The information architecture [31,32] of SSMW defined an overall step-by-step (session-by-session) path used by study participants as they interacted with the program. Participants initially followed this tunneled approach that involved completing one session at a time in order. Participants were encouraged to interact with all content on each session web page (eg, finish each video or choose the best response from several possible video interactions) before moving to a subsequent web page. Specifically, SMS text message reminders to complete activities were used throughout, and web page *next* buttons were dormant the first 2 times a user attempted to move to the next web page before she completed all activities. After 2 tries, the *next* button was enabled, and the user was able to move forward. Finally, SSMW participants could freely access the content of any sessions they had previously completed. As a result, we describe the SSMW program as having a hybrid information architecture.

Program goals centered on training participants to use strategies and *Rules of the Road* for coping more effectively with stressful problems arising from their partner's drinking. Informed by the family stress and family interactional model [12,14,33] and using web-delivered adaptations drawn from cognitive behavioral therapy [34,35], these strategies included shifting one's focus to their own needs, managing negative thinking, adjusting expectations for change, using functional analysis to understand self- and partner behavior, and developing more effective problem-solving and communication skills.

Each session included a video host-delivered overview complemented by narrated animations, instructional text, and multiple highly realistic video portrayals of 6 women and their partners interacting in multiple real-world scenarios of stressful problems arising from the partner's drinking. Extracted video screenshots with their related text instructions can be found in Multimedia Appendix 2. Structured, interactive exercises and end-of-session journaling encouraged participants to review their personal problems with their partner's drinking and apply new skills to better manage these problems. Subsequent sessions guided the participant to develop long-term personal goals and, if interested, seek additional help through professional resources and self-help groups. SSMW participants also had 24/7 secure access to an app on their smartphones [31] that provided a repository of tailored SSMW program videos, library articles,

XSL•FO
RenderX

and personalized journal content synchronized to the participant's place in the program. The SSMW program protected user privacy in part by timing out (logging the user out of the program and displaying its log-in screen) after 60 minutes of user inactivity, defined as no clicking or typing on the web page. SSMW SMS text messages were automatically sent to participants throughout the 12 weeks. Some messages were prompted by periods of participant inactivity (eg, "Haven't seen you lately. Spun out? Just get back on the road again. Lots more useful information awaits you."), by relevance to specific SSMW sessions (eg, "Dial down strong negative feelings by managing your thoughts—watch out for Awful, All or nothing, or Poor-me thoughts."), and when progress milestones were achieved (eg, "Stop the Spinning! You've got a good start!").

SSMW+Coach Support

Content of the SSMW+coach condition was essentially the same as that described for the SSMW-only condition but with the addition of 6 scheduled support calls with an assigned SSMW coach during the initial 12-week active treatment period. Coach reminder prompts were also included in the SSMW program content, and selected SMS text messages encouraged participants to reach out to their phone coach for more help and additional phone sessions if needed. Automated emails with coach call appointment confirmations also assisted in follow-through.

While the SSMW web program provided the primary coping skills training, SSMW's phone coaching was informed by the supportive accountability model [22] and was designed to (1) support and promote spouse motivation to adhere to the program and avoid dropout and (2) offer adjunctive training as needed. Coach calls addressed motivation and dropout by dealing with practical log-in, navigation, or other problems that might discourage engagement; praising engagement; and bolstering the support and connection associated with working along with the SSMW program. Adjunctive skill training assessed the participant's understanding of program content and its applicability to the participant's life. As needed or requested by the participant, the coach could provide additional direct instruction, modeling, and role-play to help bolster and generalize the skills learned. Coach calls were designed to be no longer than 20 minutes. However, calls could last longer depending on each participant's engagement level and understanding of program content and whether problems or questions were elicited during the call. To prepare for a call, the coach used a separate coach web portal to review participant progress in completing program sessions and journal entry content.

Coaches in the SSMW+coach condition were 2 master's-level substance abuse or mental health counselors. As part of the coach fidelity and credibility protocol, coaches received intensive training in the SSMW content and phone coaching before the start of recruitment. A written phone coach manual described a framework for coaching, critical coaching skills within a supportive accountability model [22] (eg, effective use of empathic and reflective listening, eliciting commitment, use of praise, and engaging the participant collaboratively), and troubleshooting coaching issues (eg, dealing with low program adherence). In addition, to maintain consistency and minimize drift, coaches used a phone coach call checklist, digitally recorded coaching calls, and attended weekly or biweekly joint review or booster training sessions with a project supervisor to review select call recordings and discuss the status of cases. We reviewed randomly selected coach call progress notes on study completion to assess coach compliance and call recording lengths. One call was randomly chosen from each individual in a 30.1% (40/133) random sample of SSMW+coach participants who had at least one coach call; due to technical recording errors, 2 participants' call length data were unavailable. In total, 2 raters independently scored each progress note narrative on the presence (1=yes; 0=no) of the following call compliance indicators: (1) assessment of program progress, (2) engagement or encouragement, (3) help with the application of SSMW content, and (4) a cognitive behavioral versus other counseling model focus. Interrater reliability (intraclass correlation coefficient; 2-way random model for rater consistency and mean rater measurement) was 0.77. Overall compliance with the coaching protocol appeared high. Across the 2 raters, the sampled notes averaged 3.6 (SD 0.5) out of 4 compliance indicators; the median was 4 (IQR 3.5-4.0). The mean length of calls was 20.0 (SD 9.2) minutes; the median was 19.9 (IQR 11.9-25.3) minutes.

UWC Condition

Women assigned to the UWC condition received access to a text-based, menu-driven site with limited still graphics. Unlike the SSMW conditions, the program presented verbatim or paraphrased content from web searches of freely accessible information on how to deal with a partner's drinking problem. However, unlike typical search result listings, the UWC site repackaged and organized search results into appealing, meaningful subject areas and adhered to a nonlinear site architecture (ie, open to exploration). The main clickable menu topic areas included information about alcohol, AUDs, the effects of the partner's drinking on the family (her and the children, if applicable), ways to cope, and information about treatment for him and on seeking more help for herself. In addition, the site provided URL links to obtain more information from external websites (eg, Al-Anon and the National Institute on Alcohol Abuse and Alcoholism). UWC participants were able to access their program content in a 14–web page freestanding website organized into 4 major sections (*Welcome* and *About the program*, *About alcohol problems*, *His drinking and you*, and *What can you do?*). Introductory UWC material informed the participant that, while the web contains a broad range of helpful information, it is not readily available in one spot and requires a lot of work to identify. So, to make the information more accessible, the site did the web search work for her, compiling the latest content from the internet and organizing it for her use. A screenshot of the UWC introductory page can be found in Multimedia Appendix 3.

In the first 2 weeks of enrollment, program use and encouragement SMS text message prompts in the UWC condition coincided in number and type with those in the SSMW conditions. Additional encouragement SMS text messages at 6 and 12 weeks were yoked to the estimated schedule of similar messages sent to participants in the SSMW conditions. Inactivity

XSL•FO
RenderX

SMS text messages were sent on the same schedule to both SSMW conditions.

We chose this UWC condition to address several challenges of incorporating credible, active control conditions in web-based intervention research [36]. Specifically, (1) the UWC closely approximated *usual care* on the web (ie, access to free helpful content that individuals could typically find themselves using web search engines); (2) unlike a referral to available self-help groups such as Al-Anon (face-to-face or web-based), it shared the framework of web interventions in having easy access, high reach potential, and promise for high public health impact; (3) it controlled for demand characteristics and participant expectancies better than a waitlist control condition while providing a control condition available throughout the follow-up period; and (4) by providing the UWC within a self-contained website, it was possible to track participant engagement in an automated, unobtrusive manner. The UWC content was credible and useful but did not delineate step-by-step behavioral strategies.

Follow-Up and Measures

Overview

Participant demographics were measured at baseline. Participant credibility and expectancy were assessed at baseline and the 6-week interim assessment; satisfaction was assessed at the 6-week interim and the 12-week posttest. Negative affect was measured at baseline, the 12-week posttest, and the 6- and 12-month follow-ups.

Negative Affect Measures

Guided by the key findings of the SSMW pilot work, the 21-item BDI-II [29] was the primary measure of depressive symptom severity, and the State Anger subscale of the STAXI-2 [30] was the primary measure of anger severity. The latter scale measures anger intensity in the moment or current situation. It contrasts with the STAXI-2 Trait Anger subscale, which measures one's overall tendency to have an angry temperament. The State Anger subscale was chosen as the most appropriate of the 2 scales in this study. Specifically, the stress and coping model of spouse functioning, on which the SSMW program is built, views the anger observed in this population as a reaction to stress and problems brought on by the partner's drinking (ie, state anger) and not the sign of an inherently angry person (trait anger). As in the pilot study, we applied a negative inverse transformation of STAXI-2 scores to reduce skew and accommodate outliers in our statistical models. Exploratory analyses used BDI-II (score of ≥14) and STAXI-2 State Anger subscale (score of ≥18) cut points to define clinically relevant levels of depression and state anger. These points correspond to mild depression and clinically significant anger cut points as defined by Beck et al [29] and Spielberger [30], respectively.

Participant Engagement and Coach Calls

For all conditions over the entire duration of the study, the database system used by program websites collected a set of automated primary program engagement metrics continuously and unobtrusively, including the number of visits to the website, the duration of visits made to program sessions, and the date in which sessions were completed. The coach recorded the number of sessions and content covered in scheduled coach calls. The posttest mean of the 12-item Working Alliance Inventory–Short Form (WAI-SF) [37] completed by SSMW+coach participants measured the quality and strength of collaborative engagement with the coach. WAI-SF items are rated on a 7-point scale ranging from 1 (*never*) to 7 (*always*) regarding participant and coach agreement on goals (Goals subscale) and tasks (Tasks subscale) and the perceived participant-coach bond (Bond subscale). As all subscales were highly correlated with the total mean item score (0.91-0.96), we used the latter to measure participant-coach engagement.

Credibility, Expectancy, and Satisfaction

We used the Credibility/Expectancy Questionnaire (CEQ) [38] at baseline before randomization and at the 6-week interim assessment to measure participant self-reported ratings of the extent to which they found their assigned program's content credible and consistent with their preprogram expectations. Credibility scores were calculated as the mean of the first 3 items of the CEQ. Expectancy was based on the item from the CEQ that asked participants to rate the following: *By the end of the program, how much improvement in your stress do you think will occur*? Overall satisfaction with the program was measured using the Client Satisfaction Questionnaire–8 [39] at the interim assessment and the 12-week posttest.

COVID-19 Impact

We developed and administered a COVID-19 impact assessment at baseline using 30 items adapted from Behar-Zusman et al [40]. Items focused on the pandemic's effect on the household (eg, *We self-quarantined due to travel or possible exposure*) and on family relationships (eg, *You spent a lot more time taking care of or trying to keep other family members occupied*). The overall score was the sum of the items, each scored as 0=no and 1=yes. The internal consistency of this measure was 0.72. This scale was used in unplanned, exploratory analyses to assess for baseline differences between conditions on pandemic impact.

Statistical Analysis

Overview

We assessed condition differences in outcome gains from pretest to posttest using a mixed-model (multilevel) time × condition analysis [41] represented by the following equation:

$$Y_{tj} = (\gamma_{00} + \gamma_{01}C_j + \gamma_{10}T_{tj} + \gamma_{11}T_{tj}C_j) + (r_{0j} + r_{1j}T_{tj} + e_{tj})$$

Y_{tj} represents the outcome for assessment occasion t on participant j. The model included 3 predictors: time, denoted by T_{tj} (coded 0 at baseline and 1 at posttest); condition, denoted by C_j (eg, coded 0 for control and 1 for treatment); and their interaction. Our planned tests of condition differences included comparisons between each SSMW condition (SSMW+coach and SSMW only) and the UWC condition as well as comparisons between the SSMW+coach and the SSMW-only conditions. The model produces estimates of the baseline intercept for the comparison condition, γ_{00}; the difference between conditions at baseline, γ_{01}; pretest-posttest change for the comparison condition, γ_{10}; and the difference in the change

XSL•FO
RenderX

in outcomes between conditions, γ_{11} (treatment efficacy). The random effects account for participant-level variability in the intercept, r_{0j}; improvements in outcome, $r_{1j}T_{tj}$; and the residual, e_{tj}, in a time × condition model with 2 time points (e_{tj}=0) due to constraints imposed by only 2 assessments.

We extended the model to include >2 time points to test for condition effects through the 12-month follow-up assessment. This extended model assumed an unstructured covariance matrix and included outcome data at baseline (time coded as 0), 6-month follow-up (time coded as 1), and 12-month follow-up (time coded as 2). We subsequently graphed descriptive statistics for outcomes by condition and assessment time to facilitate interpretation of results.

We explored the possibility that negative affect moderated treatment effects by extending the model to include baseline scores of the BDI-II and negative inverse transformations of raw baseline scores on the STAXI-2 as well as their interactions with time, condition, and the time × condition term. Pearson *r* correlation coefficients were calculated to describe the associations between the SSMW engagement metrics and outcome improvements from baseline to posttest.

Model Estimation

Time × condition models were estimated using the SAS (version 9.4; SAS Institute) [42] PROC MIXED procedure using full-information maximum likelihood (ML) methods. ML estimation uses all available data, reducing potential bias—even in the face of substantial attrition—provided data are missing at random [43]. Compared to complete-case analyses, ML relies on relatively benign assumptions and does not introduce bias [44,45].

Interpretation of Results

To supplement *P* values in our interpretation of impact results, we reported Hedges *g* effect sizes, their 95% CIs, and model probabilities based on the Akaike information criterion [46] as recommended by the American Statistical Association [47]. Model probabilities, *w*, indicate the strength of evidence for one model when compared with others given the data at hand. Burnham et al [46] described *w* as the probability of selecting the same model with a "replicate data set from the same system" and "allow statements such as the probability of [H_A] is 0.78." Model probabilities better characterize the chance of a replicated result than *P* values. In this study, we compared models for 2 hypotheses: a model with the effect of study condition (H_A: alternate hypothesis) and one without (H_0: null hypothesis). We reported the model probability for the model with the condition effect (H_A), and with only 2 models, the model probability for H_0 is $1 - w$.

Sample Size and Power

This study was designed to detect small to moderate differences (minimum detectable effect size of 0.38 SDs) between study conditions. Power analyses assumed 450 participants (exceeded by the actual sample of 456) randomly assigned to 1 of the 3 conditions, a type-II error rate of 20% (power=0.80), a type-I error rate of 0.05 for 2-tailed tests of condition, a moderate relationship between baseline and posttest outcome measures (*r*=0.50), and 5% attrition (loss to follow-up) at posttest and 20% attrition at the 12-month follow-up.

Results

Overview

Table 1 depicts participant demographics, partner AUD characteristics, and COVID-19 impact scale scores by study condition. We found no statistically significant condition differences in baseline participant characteristics (*P*≥.11 in all cases) or outcome measures (*P*≥.08 in all cases). Figures 2 and 3 illustrate observed depressive symptom and anger outcomes, respectively, throughout the study period. Tabled descriptive statistics for primary negative affect outcomes by assessment time and study condition as depicted in Figures 2 and 3 can be found in Multimedia Appendix 4.

XSL•FO
RenderX

Table 1. Baseline participant demographic and partner alcohol use disorder characteristics and COVID-19 impact indexes by study condition[a].

	SSMW[b] only (n=141)	SSMW+coach (n=151)	UWC[c] (n=164)
Participant characteristics			
Age (y), mean (SD)	45.5 (10.2)	45.9 (10.8)	45.8 (11.4)
Years of school completed, mean (SD)	16.3 (2.3)	16.0 (2.3)	16.0 (2.5)
Race, n (%)			
American Indian	0 (0)	1 (0.7)	1 (0.6)
Asian	2 (1.4)	5 (3.3)	2 (1.2)
Black or African American	3 (2.1)	7 (4.6)	7 (4.3)
Hispanic or Latino	5 (3.5)	14 (9.3)	14 (8.5)
White	130 (92.2)	128 (84.8)	150 (91.5)
Mixed or other	6 (4.3)	10 (6.6)	4 (2.4)
Currently employed, n (%)	123 (87.2)	124 (82.1)	137 (83.5)
Primary source of income, n (%)			
Own wages	39 (27.7)	43 (28.5)	39 (23.8)
Partner wages	12 (8.5)	25 (16.6)	25 (15.2)
Own and partner wages	87 (61.7)	79 (52.3)	98 (59.8)
Other (eg, unemployment insurance)	3 (2.1)	4 (2.6)	2 (1.2)
Relationship status, n (%)			
Married	110 (78)	113 (74.8)	125 (76.2)
Cohabiting but not married	31 (22)	38 (25.2)	39 (23.8)
Child age, n (%)			
No children	77 (54.6)	95 (62.9)	99 (60.4)
<18 months	20 (14.2)	19 (12.6)	18 (11)
1.5-5 years	38 (27)	33 (21.9)	35 (21.3)
6-18 years	6 (4.3)	4 (2.6)	12 (7.3)
Al-Anon attendance (last 3 months), n (%)	8 (5.7)	15 (9.9)	14 (8.5)
Mental health counseling (last 3 months), n (%)	29 (20.6)	35 (23.2)	37 (22.6)
COVID-19 impact score, mean (SD)	7.7 (4.0)	7.8 (3.6)	7.7 (3.5)
Participant-reported partner drinking characteristics			
AUDIT-C[d] score at baseline, mean (SD)	9.7 (1.9)	9.3 (1.8)	9.7 (1.8)
FHAM[e] score at baseline, mean (SD)	9.3 (2.7)	8.9 (2.8)	9.0 (2.8)
Length of drinking problem (months), mean (SD)	210.3 (136.2)	212.7 (136.2)	209.9 (144.0)
SUD[f] treatment involvement (last 3 months), n (%)	9 (6.4)	8 (5.3)	10 (6.1)
No current legal problems, n (%)	133 (94.3)	144 (95.4)	156 (95.1)

[a]The sample sizes for each variable may not add up to the total sample size.
[b]SSMW: Stop Spinning My Wheels.
[c]UWC: usual web care.
[d]AUDIT-C: Alcohol Use Disorders Identification Test–Consumption items.
[e]FHAM: Family History Assessment Module.
[f]SUD: substance use disorder.

XSL•FO
RenderX

Figure 2. Observed mean depression scores on the Beck Depression Inventory–II (BDI-II) throughout the assessment period. SSMW: Stop Spinning My Wheels; UWC: usual web care.

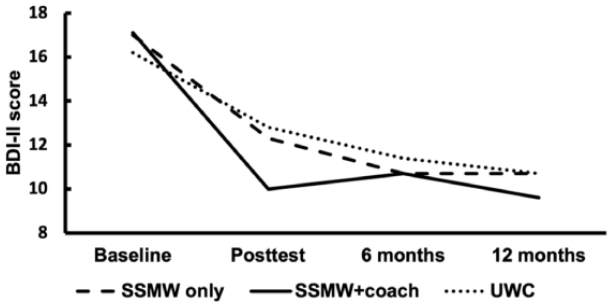

Figure 3. Observed mean, untransformed state anger scores on the State-Trait Anger Expression Inventory 2 (STAXI-2) throughout the assessment period. SSMW: Stop Spinning My Wheels; UWC: usual web care.

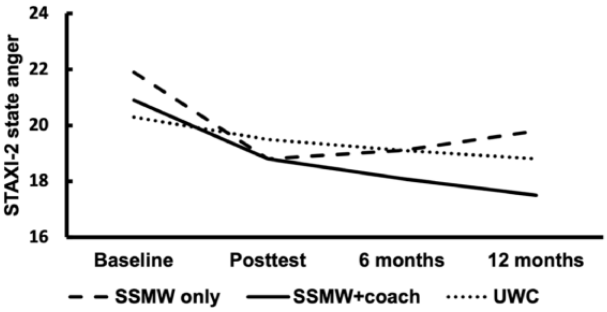

Credibility and Expectancy

Program credibility used the CEQ scale rated from 1=*Not at all useful* to 9=*Very useful*, and expectations for improvement in outcomes used a CEQ item rated on a scale from 0% to 100%. Mean credibility and expectancy ratings at baseline were similar for SSMW only (mean 6.3, SD 1.6 and mean 52.6, SD 19.9, respectively), SSMW+coach (mean 6.2, SD 1.6 and mean 52.0, SD 19.6, respectively), and UWC (mean 6.4, SD 1.6 and mean 55.5, SD 22.1, respectively). Credibility and expectancy ratings at the 6-week interim assessment stayed stable or slightly increased for the SSMW-only (mean 6.6, SD 1.9 and mean 51.4, SD 20.1, respectively) and SSMW+coach (mean 7.3, SD 1.5 and mean 61.4, SD 21.9, respectively) conditions but were lower for the UWC condition (mean 4.7, SD 2.1 and mean 38.0, SD 26.9, respectively).

Loss to Follow-Up

Rates of failure to complete scheduled study assessments (loss to follow-up) for the SSMW-only, SSMW+coach, and UWC conditions were 26.2% (37/141), 19.2% (29/151), and 21.3% (35/164), respectively, at posttest; 35.5% (50/141), 27.8% (42/151), and 24.4% (40/164), respectively, at the 6-month follow-up; and 31.2% (44/141), 27.2% (41/151), and 22% (36/164), respectively, at the 12-month follow-up (see Figure 1 for participant flow and follow-up rates). Study participants with children (vs no children) were less likely to complete their assessments at the posttest (129/185, 69.7% vs 226/271, 83.4% completion; N=456, χ^2_1=12.8, *P*<.001), 6-month follow-up (118/185, 63.8% vs 206/271, 76%; N=456, χ^2_1=8.0, *P*=.005), and 12-month follow-up (120/185, 64.9% vs 215/271, 79.3%; N=456, χ^2_1=11.8, *P*=.001). Attrition rates did not significantly differ as a function of condition (*P*=.10-.33 in all cases), other demographic characteristics (*P*=.07-.98 in all cases), or baseline outcome measures (*P*=.16-.38 in all cases). Interactions between attrition and study condition did not predict baseline outcome scores (*P*=.37-.82 in all cases). These results provide no evidence of bias due to attrition in our tests of condition effects.

Results of Hypothesis-Driven Research Questions

Research Question 1: What Was the Impact of SSMW Only Versus UWC?

The first column of Table 2 summarizes time × condition model results comparing the SSMW-only and UWC conditions on change in negative affect outcomes from baseline to posttest. SSMW-only participants achieved greater decreases than UWC participants on the STAXI-2 (*g*=–0.34, 95% CI –0.64 to –0.03; t_{303}=–2.19; *P*=.03; *w*=0.79). These results suggested that the hypothesis of a difference between conditions measured using the time × condition effect fit the data. That is, the model for STAXI-2 scores that included the time × condition interaction had a considerably higher probability (*w*=0.79) than the model without the condition difference (*w*=0.21). To illustrate the fixed effects, the model estimated a baseline intercept for the UWC condition of –0.054, a baseline-to-posttest change for the UWC condition of –0.001, a difference between conditions at baseline of 0.003, and a difference between conditions in terms of the change from baseline to posttest assessment of –0.004. We found no evidence of a condition difference between SSMW only and UWC in change in BDI-II scores from baseline to posttest (*g*=–0.14, 95% CI –0.35 to 0.06; t_{303}=–1.41; *P*=.16; *w*=0.49). From baseline to the 12-month follow-up, we found no evidence of differences between the SSMW-only and UWC conditions on change in BDI-II (*g*=–0.09, 95% CI –0.27 to 0.09;

XSL•FO
RenderX

t_{303}=–0.95; *P*=.34; *w*=0.36) or STAXI-2 (*g*=–0.06, 95% CI –0.30 to 0.15; t_{303}=–0.62; *P*=.54; *w*=0.30) scores.

Table 2. Time × condition model results comparing conditions on negative affect improvements from baseline to posttest.

Effect or statistic	Condition contrast and negative affect outcome					
	SSMW[a] only vs usual web care		SSMW+coach vs usual web care		SSMW+coach vs SSMW only	
	Depression (BDI-II[b])	Anger (STAXI-2[c])	Depression (BDI-II)	Anger (STAXI-2)	Depression (BDI-II)	Anger (STAXI-2)
Model probability (*w*)[d]	0.49	0.79	0.99	0.52	0.67	0.33
Fixed effects, parameter estimate (SE)						
Intercept	16.2[e] (0.8)	–0.054[e] (0.001)	16.2[e] (0.8)	–0.054[e] (0.001)	17.0[e] (0.8)	–0.051[e] (0.001)
Time	–3.4[e] (0.7)	–0.001 (0.001)	–3.4[e] (0.7)	–0.001 (0.001)	–4.9[e] (0.8)	–0.006[e] (0.001)
Condition	0.8 (1.2)	0.003[f] (0.002)	0.9 (1.1)	0.002 (0.001)	0.1 (1.1)	–0.001 (0.002)
Time × condition	–1.5 (1.0)	–0.004[f] (0.002)	–3.4[e] (1.0)	–0.003 (0.002)	–1.9 (1.0)	0.001 (0.002)
Variances, parameter estimate (SE)						
Participant-level intercept	69.3[e] (7.3)	0.000[e] (0.000)	63.4[e] (6.7)	0.000[e] (0.000)	55.3[e] (6.4)	0.000[e] (0.000)
Participant-level gain	32.2[e] (3.0)	0.000[e] (0.000)	31.9[e] (2.8)	0.000[e] (0.000)	31.7[e] (3.0)	0.000[e] (0.000)
Time × condition effect size and statistical significance						
Hedges *g* effect size (95% CI)	–0.14 (–0.35 to 0.06)	–0.34 (–0.64 to 0.03)	–0.36 (–0.57 to 0.15)	–0.22 (–0.52 to 0.07)	–0.22 (–0.45 to 0.01)	0.13 (–0.19 to 0.44)
P value	.16	.03	.001	.14	.06	.44

[a]SSMW: Stop Spinning My Wheels.

[b]BDI-II: Beck Depression Inventory–II.

[c]STAXI-2: State-Trait Anger Expression Inventory 2. The negative inverses of raw STAXI-2 scores were analyzed.

[d]The model probabilities (*w*) compares the strength of evidence for a model with the time × condition effect to one without. Condition was coded such that the second group listed for each contrast served as the reference category. Time was coded as 0 for baseline and 1 for posttest. Tests of fixed effects included 303, 313, and 289 df for each contrast.

[e]*P*<.001.

[f]*P*<.05.

Research Question 2: What Was the Impact of the SSMW+Coach Condition Versus UWC?

The second column of Table 2 summarizes time × condition model results comparing the SSMW+coach and UWC conditions on change in negative affect outcomes from baseline to posttest. SSMW+coach participants achieved greater decreases than UWC participants in BDI-II scores (*g*=–0.36, 95% CI –0.57 to –0.15; t_{313}=–3.40; *P*=.001; *w*=0.99). We found no difference between the SSMW+coach and UWC conditions in change in STAXI-2 scores from baseline to posttest (*g*=–0.22, 95% CI –0.52 to 0.07; t_{313}=–1.49; *P*=.14; *w*=0.52).

From baseline to the 12-month follow-up, SSMW+coach participants achieved greater decreases than UWC participants on STAXI-2 scores (*g*=–0.24, 95% CI –0.45 to 0.00; t_{313}=–1.95; *P*=.052; *w*=0.70). Though not statistically significant, the model probability with the time × condition effect (*w*=0.70) was considerably greater than the model probability without the time × condition effect (*w*=0.30), hence providing support for the hypothesized effect, albeit a small one. We found no difference between the SSMW+coach and UWC conditions on change in BDI-II scores from baseline to the 12-month follow-up (*g*=–0.15, 95% CI –0.33 to 0.03; t_{313}=–1.58; *P*=.11; *w*=0.56).

Research Question 3: Did Adding a Coach to SSMW Reduce Negative Affect More Than SSMW Alone?

The third column of Table 2 summarizes time × condition model results comparing the SSMW+coach and SSMW-only conditions on outcome improvements from baseline to posttest. The effect of SSMW+coach versus SSMW only on BDI-II scores during this period was small and was not statistically significant at conventional levels (*g*=–0.22, 95% CI –0.45 to 0.01; t_{289}=–1.86; *P*=.063; *w*=0.67). However, the model probability with the time × condition effect (*w*=0.67) was considerably greater than the model probability without the time × condition effect (*w*=0.33), hence providing support for the hypothesized effect, albeit a small one. We found no difference between the SSMW+coach and SSMW-only conditions on baseline to posttest STAXI-2 score changes (*g*=0.13, 95% CI –0.19 to 0.44; t_{289}=0.76; *P*=.45;

XSL•FO
RenderX

w=0.33). When we examined changes from baseline to the 12-month follow-up, we also found no differences between the SSMW+coach and SSMW-only conditions for BDI-II (*g*=–0.06, 95% CI –0.27 to 0.15; t_{290}=–0.57; *P*=.57; *w*=0.30) or STAXI-2 (*g*=–0.15, 95% CI –0.39 to 0.09; t_{290}=–1.22; *P*=.23; *w*=0.43) scores.

Research Question 4: Did Baseline Negative Affect Moderate Effects of Condition?

To examine whether baseline negative affect scores moderated condition effects on negative affect outcomes from baseline to posttest, the baseline score and its interaction with time, condition, and the time × condition term was added to the models used to address the first 3 research questions. In the comparison between SSMW only and UWC, we found no moderation effect for depression severity (BDI-II scores; 3-way interaction estimate=–0.05, 95% CI –0.25 to 0.16; t_{301}=–0.44; *P*=.66; *w*=0.28) or anger (STAXI-2 scores; estimate=0.03, 95% CI –0.16 to 0.21; t_{301}=0.27; *P*=.78; *w*=0.27). We also found no moderation effect in the comparison between SSMW+coach and UWC on depression severity (BDI-II scores; estimate=–0.13, 95% CI –0.33 to 0.08; t_{311}=–1.23; *P*=.22; *w*=0.43) or anger (STAXI-2 scores; estimate=–0.04, 95% CI –0.23 to 0.15; t_{311}=–0.43; *P*=.67; *w*=0.28). Finally, we found no evidence of moderation of the SSMW+coach effects relative to SSMW only on baseline depression (BDI-II scores; estimate=–0.08, 95% CI –0.30 to 0.13; t_{288}=–0.76; *P*=.45; *w*=0.32) or anger (STAXI-2 scores; estimate=–0.10, 95% CI –0.28 to 0.07; t_{288}=–1.16; *P*=.25; *w*=0.41).

Unplanned Exploratory Analyses of Clinically Relevant Change

Did Remission Rates Differ by Study Condition Among Women With Clinically Relevant Negative Affect at Baseline?

Because condition differences were most robust at the posttest, we explored condition differences in rates of remission at posttest from clinically relevant levels of depressive symptoms based on a BDI-II cutoff of 14. The analysis included 208 participants with a baseline BDI-II score of ≥14 and complete posttest data (n=67, 32.2% for SSMW only; n=74, 35.6% for SSMW+coach; and n=67, 32.2% for UWC). We examined condition differences in rates of remission using contingency table analyses with chi-square test statistics and reported odds ratios (ORs) as a measure of effect size. Rates of remission based on posttest BDI-II scores were 49% (33/67), 62% (46/74), and 31% (21/67) for the SSMW-only, SSMW+coach, and UWC conditions, respectively. Participants in the SSMW-only condition were more likely than UWC participants to have remitted (33/67, 49% vs 21/67, 31%; OR 2.13, 95% CI 1.05-4.30; *P*=.03). Similarly, participants in the SSMW+coach condition were more likely than UWC participants to have remitted (46/74, 62% vs 21/67, 31%; OR 3.60, 95% CI 1.79-7.23; *P*<.001). SSMW-only and SSMW+coach remission rates did not differ (*P*=.12). From a *number needed to treat* perspective [48], the SSMW-only condition required 5.6 participants to achieve 1 remission, whereas SSMW+coach needed only 3.2 participants.

We similarly explored condition differences in rates of remission among 175 participants with elevated anger at baseline (STAXI-2 score of ≥18) and complete posttest data (n=53, 30.3% for SSMW only; n=62, 35.4% for SSMW+coach; and n=60, 34.3% for UWC). Rates of remission based on posttest STAXI-2 scores were 51% (27/53), 55% (34/62), and 40% (24/60) for the SSMW-only, SSMW+coach, and UWC conditions, respectively. No group contrasts were statistically significant (*P*≥.10 in all cases).

Was There Differential Onset of Clinically Relevant Negative Affect Levels Among Those Without Clinically Relevant Negative Affect at Baseline?

This analysis for depression included 147 participants with a baseline BDI-II score of <14 and complete posttest data (n=37, 25.2% for SSMW only; n=48, 32.7% for SSMW+coach; and n=62, 42.2% for UWC). Rates of onset based on posttest BDI-II scores were 14% (5/37), 6% (3/48), and 10% (6/62) for the SSMW-only, SSMW+coach, and UWC conditions, respectively. No group contrasts were statistically significant (*P*≥.26 in all cases).

We also explored condition differences in rates of onset of elevated anger. The analysis included 180 participants with a baseline STAXI-2 score of <18 and complete posttest data (n=51, 28.3% for SSMW only; n=60, 33.3% for SSMW+coach; and n=69, 38.3% for UWC). Rates of onset using posttest STAXI-2 scores were 24% (12/51), 28% (17/60), and 28% (19/69) for the SSMW-only, SSMW+coach, and UWC conditions, respectively. No contrasts were statistically significant (*P*≥.57 in all cases).

Ancillary Engagement and Satisfaction Analyses

SSMW Engagement

Table 3 summarizes program engagement metrics by study condition. We examined condition differences using independent-sample, 2-tailed *t* tests and associations between engagement and outcomes using Pearson correlation coefficients. Participants in the SSMW+coach condition, compared to the SSMW-only condition, visited their program more often (t_{290}=4.57; *P*<.001) and for a longer duration (t_{290}=5.49; *P*<.001) and completed more program sessions (t_{290}=5.79; *P*<.001).

XSL•FO
RenderX

Table 3. Program engagement and satisfaction by study condition.

Program engagement metric	SSMW[a] only, mean (SD; median)	SSMW+coach, mean (SD; median)	UWC[b], mean (SD; median)
Number of visits	8.8 (7.7; 7.0)	13.4 (9.1; 12.0)	3.6 (2.6; 3.0)
Hours on the program	5.3 (5.5; 4.3)	9.7 (7.7; 8.3)	1.2 (1.8; 0.7)
Sessions completed out of 20	9.3 (7.6; 9.0)	14.1 (6.5; 16.0)	__[c]
Number of coach calls	—	3.8 (2.1; 4.0)	—
Working Alliance Inventory total score[d]	—	5.4 (1.4; 5.9)	—

[a]SSMW: Stop Spinning My Wheels.

[b]UWC: usual web care.

[c]Not applicable.

[d]Working Alliance Inventory items were rated on a 7-point scale ranging from 1 (*never*) to 7 (*always*).

The WAI-SF score in the SSMW+coach condition was positively correlated with the total number of visits to the SSMW program (r=0.31; P=.001), hours on the program across visits (r=0.27; P=.003), number of sessions completed (r=0.41; P<.001), and number of coach calls (r=0.55; P<.001). Higher WAI-SF scores also were associated with greater decreases in BDI-II scores from baseline to posttest (r=–0.18; P=.045). Notably, across both SSMW conditions, the higher number of SSMW sessions completed was associated with greater decreases in BDI-II scores from baseline to posttest (r=–0.20; P=.002), whereas other SSMW engagement metrics were not correlated with outcome gains (P≥.13 in all cases).

UWC Engagement

Participants in the UWC condition averaged 3.6 (SD 2.6) visits to their program and 1.2 (SD 1.8) hours on the program across visits. These engagement metrics were not correlated with outcome gains from baseline to posttest (P≥.67 in all cases).

Program Satisfaction

The Client Satisfaction Questionnaire–8 used a 4-point rating scale coded as 1=*Not at all satisfied* to 4=*Very satisfied*. SSMW-only and SSMW+coach participants reported being satisfied with program features at the 6-week interim assessment (mean 3.1, SD 0.6 vs mean 3.3, SD 0.5) and the 12-week posttest (mean 3.0, SD 0.8 vs mean 3.3, SD 0.6). SSMW+coach participants, compared to SSMW-only participants, reported greater satisfaction at the 6-week interim assessment (t_{229}=3.99; P<.001) and the 12-week posttest (t_{237}=2.93; P=.004). UWC participants, compared to SSMW participants, were less satisfied at the 6-week interim assessment (mean 2.5, SD 0.7 vs mean 3.2, SD 0.6; t_{366}=11.58; P<.001) and the 12-week posttest (mean 2.4, SD 0.8 vs mean 3.2, SD 0.7; t_{375}=9.83; P<.001).

Discussion

Support for Original Hypotheses

Primary outcome results partially supported the hypotheses regarding the presumed benefit of the SSMW-only condition over the UWC active control. Contrary to our hypotheses, the SSMW-only and UWC conditions showed comparable reductions from baseline depressive symptoms at posttest and follow-up (Figure 2) and did not differ. However, participants with a clinically relevant baseline level of depressive symptoms were significantly more likely than UWC participants to have remitted to a nonclinical level by the posttest. As predicted, SSMW-only participants experienced a greater baseline-to-posttest decrease in anger compared to UWC participants, whose state anger decreased little from the baseline level at either the posttest or follow-up. However, a rebound in SSMW-only anger appeared to erase this advantage when outcome trajectories were compared up to the 12-month follow-up (Figure 3). Remission rates from a clinically relevant level of anger also did not differ between conditions.

The results were more consistent with hypothesized differences between SSMW+coach and UWC. As predicted, at the posttest, SSMW+coach participants displayed a larger decrease in depressive symptoms and a higher remission rate than UWC participants (Figure 2). In addition, depressive symptom reductions at posttest in SSMW+coach participants appeared to be sustained during follow-up. However, concurrent UWC changes in depressive symptoms appeared to erase differences from the changes in the SSMW+coach condition when outcome trajectories were compared up to the 12-month follow-up (Figure 2). Contrary to predictions, SSMW+coach anger reductions did not differ significantly from those in the UWC condition at posttest, or posttest anger remission. However, SSMW+coach anger decreased further up to the 12-month follow-up, and their overall reduction in anger from baseline was greater to a small degree than that of the UWC condition (Figure 3). Overall, to varying degrees, the SSMW conditions appeared to facilitate earlier improvement in negative affect outcomes than the UWC, and those initial improvements, particularly for depressive symptoms, endured until follow-up. However, the reduction in UWC depression was particularly notable and may have dampened the emergence of predicted differences from the SSMW-only condition at posttest and from the SSMW+coach condition at follow-up. Unfortunately, this study was not designed to measure whether the change in UWC resulted from regression to the mean, UWC content, participant uptake of other treatment programs, or other factors.

Contrary to our hypothesis, adding a phone coach to SSMW generated only a small additional benefit on posttest depression, but this difference did not reach conventional levels of statistical significance. The video-rich, engaging design of the SSMW program alone may have mitigated the benefits that coach

XSL•FO
RenderX

involvement might have provided. Nevertheless, ancillary engagement and satisfaction analyses and exploratory comparisons of the individual SSMW conditions relative to UWC provided additional support for the benefits of a coach. Specifically, compared to SSMW only, SSMW+coach participants displayed higher levels of engagement and reported greater program satisfaction. Furthermore, when compared with UWC, SSMW+coach participants achieved a greater decrease in depressive symptoms at posttest, whereas those in SSMW only did not. Moreover, while SSMW-only participants reported a greater reduction than UWC participants in anger at posttest, only SSMW+coach participants reported greater anger reduction at follow-up. More importantly, SSMW+coach achieved the highest rate of remission from clinically relevant depressive symptoms (46/74, 62%) compared to either SSMW only (33/67, 49%) or UWC (21/67, 31%). Notably, although SSMW-only participants were >2 times as likely as UWC participants to remit to below depression threshold levels, SSMW+coach participants were >3.5 times as likely. Finally, the SSMW-only condition required 5.6 participants to achieve 1 remission, whereas SSMW+coach needed only 3.2 participants. These findings suggest that adding a coach can (1) significantly improve SSMW engagement and satisfaction, (2) result in clinically meaningful change compared to usual web use, and (3) possibly confer a small advantage over SSMW alone. The quality and strength of the participant-coach relationship may play an important role in these findings, as higher WAI-SF scores were associated with higher program engagement and a greater reduction in depressive symptoms.

We did not find support for the coach-driven hypothesis that baseline negative affect would moderate change in negative affect between the 2 SSMW conditions (ie, women with higher negative affect would benefit more from having a coach than from not having one). Although results suggest that individuals experiencing a wide range of symptom levels may benefit from SSMW conditions, we caution that our sample averaged in the mild range of depressive symptoms, and their outcomes might not generalize to individuals with severe depression.

Comparison to Prior Work

The results support and build on our web-based SSMW pilot study and our prior work on face-to-face coping skill training for spouses that used waitlist control conditions [5,21]. The active UWC condition in this study is a significant methodological advancement over our prior work that may present a more stringent comparison based on which to gauge SSMW effects. In this respect, it is noteworthy that the size of effects in this study was approximately half that of those in the waitlist-control pilot. These findings and observed improvement in UWC during follow-up highlight the importance of active controls and longitudinal follow-ups in web-based RCTs. This study also extends our previous work by examining SSMW effects with and without brief coach support. Finally, the large sample size relative to earlier work on coping skill training with this population allowed for more powerful tests of SSMW, coaching effects, moderating effects, and exploratory analyses of clinically relevant outcomes.

Limitations

Several study limitations may have influenced internal and external validity. First, while coaches complied with the coaching protocol and collaborative engagement was relatively high, we did not gauge overall coach skillfulness. Second, SSMW+coach participants were scheduled for 6 brief coach calls, but the optimum level of coaching required and the participants who benefit most from it are unknown. Third, although our pilot found support for anger as a primary outcome, anger measurement typically has not been included in studies of this population, and anger findings were the weakest in this study. However, feelings of anger are a common experience in this population. Future research is needed to better understand and assess anger in this group and evaluate measures of angry thoughts and feelings that are more closely associated with the stressors experienced. Fourth, our study participants were predominantly White and highly educated. More diverse samples are needed. Fifth, most of the recruitment, treatment, and follow-up occurred amid the COVID-19 pandemic. Although we found no baseline condition differences in COVID-19 impact, some participants who might otherwise not have taken part did so because they had the time to do it. Others began reevaluating their life and sought help. However, as businesses and schools reopened, follow-ups became more challenging as participants were again pressed for time. The greater attrition among women with children is also noteworthy. We speculate that these women were more challenging to follow because they had to manage more disruption in their lives and that of their children as the pandemic and its consequences waxed and waned. In addition, other pandemic-related factors not assessed may have influenced the results. Sixth, because participants in this study consented to be assigned to any of our 3 possible study conditions, our results might not generalize to women interested only in a fully self-directed program or only in a program involving a coach [49,50]. Finally, exploratory findings need to be interpreted cautiously. For these analyses, samples were smaller; the results were less consistent; and, in certain instances, the effects were smaller with wider 95% CIs than in primary analyses.

Implications

The results of this study provide empirical support for adding the self-directed web-based SSMW treatment to the limited options available for women experiencing distress from a partner's drinking, providing greater access to care for this large, underserved population. Although our findings suggest minimal phone coaching may facilitate engagement and incrementally improve outcomes, coach-facilitated web-based programs have additional requirements (eg, coach availability, scheduling, and training) and costs that could discourage broad implementation and increase costs to participants. As noted by Matthay et al [51] and embodied in the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) model [52], broadly implemented treatment programs with a smaller effect size can nonetheless have a large population health impact.

Implications for future research are informed by the National Institutes of Health Stage Model [36,53]. First, future SSMW efficacy research could focus on replicating our current findings

XSL•FO
RenderX

and possibly exploring features of adjunctive coaching (eg, who benefits most and optimal coaching levels). Second, while our hybrid design incorporated implementation elements, future research could focus on "...generalizability, implementation, cost-effectiveness, and social validity (acceptability to end users, program adopters, health care providers, policy makers)" [36]. Future analyses should also examine mediators of the SSMW treatment effect and secondary outcomes (eg, partner drinking, violence, and relationship stability).

Finally, the SSMW program's cognitive behavioral approach and engaging video-rich strategies provide a framework for adapting and delivering web care for other often hidden and underserved spouse and caregiver populations. For example, eHealth adaptations of face-to-face coping skill training programs for parents of an adolescent with an SUD [54], spouses of those with a gambling disorder [55], and caregivers of a family member with an SUD [56] may accelerate research on and rapid implementation of accessible treatments for these groups.

Acknowledgments

The research was supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award R01AA024118 to RGR and BGD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge John R Seeley, Jean Finn, Jennifer Robinson, Lara Balawejder, Anna Crouss, Steven Sanders, Diane Augustino, Jaclyn Colangelo, Paulette Giarratano, Tim Woolley (IEQ Technologies), Mike Gluck (Gluckworks), and the University at Buffalo Research Information Systems team for their invaluable work on this project. The authors also express appreciation to Kenneth Leonard for his help in facilitating the completion of this study.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RGR and BGD contributed to conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, and writing—original draft and review and editing. NBM contributed to conceptualization, data curation, investigation, methodology, project administration, resources, supervision, validation, and writing—review and editing. MST contributed to funding acquisition, investigation, methodology, project administration, resources, software, supervision, and writing—review and editing. CB contributed to conceptualization, methodology, project administration, resources, and software. FL provided research project administration. DBK contributed to data curation, formal analysis, methodology, validation, visualization, and writing—original draft and review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist.
[PDF File (Adobe PDF File), 359 KB - jmir_v26i1e56119_app1.pdf]

Multimedia Appendix 2
Extracted video screenshots with their related text instructions (excerpted from Stop Spinning My Wheels and formatted for this paper; portrayals are by paid actors).
[PDF File (Adobe PDF File), 2161 KB - jmir_v26i1e56119_app2.pdf]

Multimedia Appendix 3
Screenshot of usual web care introductory page.
[PDF File (Adobe PDF File), 4114 KB - jmir_v26i1e56119_app3.pdf]

Multimedia Appendix 4
Tabled descriptive statistics for primary negative affect outcomes by assessment time and study condition.
[PDF File (Adobe PDF File), 88 KB - jmir_v26i1e56119_app4.pdf]

References

XSL•FO
RenderX

1. Nayak MB, Patterson D, Wilsnack SC, Karriker-Jaffe KJ, Greenfield TK. Alcohol's secondhand harms in the United States: new data on prevalence and risk factors. J Stud Alcohol Drugs 2019 May;80(3):273-281 [FREE Full text] [doi: 10.15288/jsad.2019.80.273] [Medline: 31250790]
2. Alcohol use and alcohol use disorders in the United States: main findings from the 2001–2002 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). National Institutes of Health National Institute on Alcohol Abuse and Alcoholism. 2006 Jan. URL: https://www.niaaa.nih.gov/sites/default/files/NESARCDRM.pdf [accessed 2024-08-06]
3. Dawson DA, Grant BF, Chou SP, Stinson FS. The impact of partner alcohol problems on women's physical and mental health. J Stud Alcohol Drugs 2007 Jan;68(1):66-75. [doi: 10.15288/jsad.2007.68.66] [Medline: 17149519]
4. Hudson CR, Kirby KC, Clements NT, Benishek LA, Nick CE. Social adjustment of women with and without a substance-abusing partner. J Psychoactive Drugs 2014 May 06;46(2):106-113 [FREE Full text] [doi: 10.1080/02791072.2014.880536] [Medline: 25052786]
5. Rychtarik RG, McGillicuddy NB. Coping skills training and 12-step facilitation for women whose partner has alcoholism: effects on depression, the partner's drinking, and partner physical violence. J Consult Clin Psychol 2005 Apr;73(2):249-261 [FREE Full text] [doi: 10.1037/0022-006X.73.2.249] [Medline: 15796632]
6. Copello A, Templeton L, Powell J. The impact of addiction on the family: estimates of prevalence and costs. Drugs Educ Prevent Policy 2010 Nov 16;17(sup1):63-74. [doi: 10.3109/09687637.2010.514798]
7. Lipscomb HJ, Dement JM, Li L. Health care utilization of families of carpenters with alcohol or substance abuse-related diagnoses. Am J Ind Med 2003 Apr 13;43(4):361-368. [doi: 10.1002/ajim.10214] [Medline: 12645093]
8. Ray G, Mertens JR, Weisner C. The excess medical cost and health problems of family members of persons diagnosed with alcohol or drug problems. Med Care 2007 Feb;45(2):116-122. [doi: 10.1097/01.mlr.0000241109.55054.04] [Medline: 17224773]
9. Roberts KS, Brent EE. Physician utilization and illness patterns in families of alcoholics. J Stud Alcohol 1982 Jan;43(1):119-128. [doi: 10.15288/jsa.1982.43.119] [Medline: 7109611]
10. Svenson LW, Forster DI, Woodhead SE, Platt GH. Individuals with a chemical-dependent family member. Does their health care use increase? Can Fam Physician 1995 Sep;41:1488-1493 [FREE Full text] [Medline: 8520237]
11. Weisner C, Parthasarathy S, Moore C, Mertens JR. Individuals receiving addiction treatment: are medical costs of their family members reduced? Addiction 2010 Jul 08;105(7):1226-1234 [FREE Full text] [doi: 10.1111/j.1360-0443.2010.02947.x] [Medline: 20491730]
12. Moos RH, Finney JW, Cronkite RC. Alcoholism Treatment: Context, Process, and Outcome. New York, NY: Oxford University Press; 1990.
13. Orford J, Velleman R, Natera G, Templeton L, Copello A. Addiction in the family is a major but neglected contributor to the global burden of adult ill-health. Soc Sci Med 2013 Feb;78:70-77 [FREE Full text] [doi: 10.1016/j.socscimed.2012.11.036] [Medline: 23268776]
14. Rychtarik RG, McGillicuddy NB. The spouse situation inventory: a role-play measure of coping skills in women with alcoholic partners. J Fam Psychol 1997 Sep;11(3):289-300 [FREE Full text] [doi: 10.1037/0893-3200.11.3.289] [Medline: 16981014]
15. Zweben A. Problem drinking and marital adjustment. J Stud Alcohol 1986 Mar;47(2):167-172. [doi: 10.15288/jsa.1986.47.167] [Medline: 3713180]
16. Daley DC, Smith E, Balogh D, Toscaloni J. Forgotten but not gone: the impact of the opioid epidemic and other substance use disorders on families and children. Commonwealth 2018;20:2-3 [FREE Full text] [doi: 10.15367/com.v20i2-3.189]
17. Merkouris SS, Rodda SN, Dowling NA. Affected other interventions: a systematic review and meta-analysis across addictions. Addiction 2022 Sep 21;117(9):2393-2414 [FREE Full text] [doi: 10.1111/add.15825] [Medline: 35129234]
18. McCrady BS, Flanagan JC. The role of the family in alcohol use disorder recovery for adults. Alcohol Res 2021 May 6;41(1):06 [FREE Full text] [doi: 10.35946/arcr.v41.1.06] [Medline: 33981521]
19. Timko C, Laudet A, Moos RH. Al-Anon newcomers: benefits of continuing attendance for six months. Am J Drug Alcohol Abuse 2016 Jul;42(4):441-449 [FREE Full text] [doi: 10.3109/00952990.2016.1148702] [Medline: 27120262]
20. Osilla KC, Trail TE, Pedersen ER, Gore KL, Tolpadi A, Rodriguez LM. Efficacy of a web-based intervention for concerned spouses of service members and veterans with alcohol misuse. J Marital Fam Ther 2018 Apr 03;44(2):292-306 [FREE Full text] [doi: 10.1111/jmft.12279] [Medline: 28972265]
21. Rychtarik RG, McGillicuddy NB, Barrick C. Web-based coping skills training for women whose partner has a drinking problem. Psychol Addict Behav 2015 Mar;29(1):26-33 [FREE Full text] [doi: 10.1037/adb0000032] [Medline: 25347016]
22. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
23. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340(mar23 1):c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
24. Rychtarik R. e-Health coping skills training for women whose partner has a drinking problem. ClinicalTrials.gov. URL: https://www.clinicaltrials.gov/study/NCT02984241 [accessed 2023-08-31]

XSL•FO
RenderX

25. Janca A, Bucholz KK, Janca MA, Jabos-Laster L. Family history assessment module. Washington University School of Medicine. 1991. URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/document.cgi?study_id=phs000092.v1.p1&phd=1039 [accessed 2024-08-06]
26. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Alcohol Use Disorders Identification Test. Arch Intern Med 1998 Sep 14;158(16):1789-1795. [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
27. First MB, Williams JB, Karg SK, Spitzer RL. Structured Clinical Interview for DSM-5 Disorders--Clinical Version (SCID-5-CV). Arlington, VA: American Psychiatric Association; 2016.
28. Wei LJ, Lachin JM. Properties of the urn randomization in clinical trials. Control Clin Trials 1988 Dec;9(4):345-364 [FREE Full text] [doi: 10.1016/0197-2456(88)90048-7] [Medline: 3203525]
29. Beck AT, Steer RA, Brown GK. BDI-II, Beck Depression Inventory Manual. San Antonio, TX: Psychological Corporation; 1996.
30. Spielberger CD. STAXI-2: State-Trait Anger Expression Inventory–2. Odessa, FL: Psychological Assessment Resources; 1999.
31. Danaher BG, Brendryen H, Seeley JR, Tyler MS, Woolley T. From black box to toolbox: outlining device functionality, engagement activities, and the pervasive information architecture of mHealth interventions. Internet Interv 2015 Mar 01;2(1):91-101 [FREE Full text] [doi: 10.1016/j.invent.2015.01.002] [Medline: 25750862]
32. Danaher BG, McKay HG, Seeley JR. The information architecture of behavior change websites. J Med Internet Res 2005 May 18;7(2):e12 [FREE Full text] [doi: 10.2196/jmir.7.2.e12] [Medline: 15914459]
33. Hobfoll SE, Spielberger CD. Family stress: integrating theory and measurement. J Fam Psychol 1992 Dec;6(2):99-112. [doi: 10.1037/0893-3200.6.2.99]
34. Wilhelm S, Weingarden H, Ladis I, Braddick V, Shin J, Jacobson NC. Cognitive-behavioral therapy in the digital age: presidential address. Behav Ther 2020 Jan;51(1):1-14 [FREE Full text] [doi: 10.1016/j.beth.2019.08.001] [Medline: 32005328]
35. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
36. Danaher BG, Seeley JR. Methodological issues in research on web-based behavioral interventions. Ann Behav Med 2009 Aug;38(1):28-39 [FREE Full text] [doi: 10.1007/s12160-009-9129-0] [Medline: 19806416]
37. Tracey TJ, Kokotovic AM. Factor structure of the Working Alliance Inventory. Psychol Assess 1989 Sep;1(3):207-210. [doi: 10.1037/1040-3590.1.3.207]
38. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
39. Attkisson CC, Greenfield TK. The UCSF client satisfaction scales: I. The Client Satisfaction Questionnaire-8. In: Maruish ME, editor. The Use of Psychological Testing for Treatment Planning and Outcomes Assessment: Instruments for Adults. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2004:799-811.
40. Behar-Zusman V, Chavez JV, Gattamorta K. Developing a measure of the impact of COVID-19 social distancing on household conflict and cohesion. Fam Process 2020 Sep 20;59(3):1045-1059 [FREE Full text] [doi: 10.1111/famp.12579] [Medline: 32621755]
41. Murray DM. Design and Analysis of Group-randomized Trials. New York, NY: Oxford University Press; 1998.
42. Base SAS® 9.4 procedures guide, seventh edition. SAS Institute, Inc. URL: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/proc/titlepage.htm [accessed 2024-08-06]
43. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol 2009 Jan;60(1):549-576. [doi: 10.1146/annurev.psych.58.110405.085530] [Medline: 18652544]
44. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6(4):330-351. [doi: 10.1037/1082-989x.6.4.330]
45. Allison PD. Missing data. In: Millsap RE, Maydeu-Olivares A, editors. The Sage Handbook of Quantitative Methods in Psychology. Thousand Oaks, CA: SAGE Publications; 2009:72-89.
46. Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol 2010 Aug 18;65(1):23-35. [doi: 10.1007/s00265-010-1029-6]
47. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Statistician 2016 Jun 09;70(2):129-133. [doi: 10.1080/00031305.2016.1154108]
48. Furukawa TA. From effect size into number needed to treat. The Lancet 1999 May 15;353(9165):1680. [doi: 10.1016/s0140-6736(99)01163-0]
49. Milgrom J, Danaher BG, Seeley JR, Holt CJ, Holt C, Ericksen J, et al. Internet and face-to-face cognitive behavioral therapy for postnatal depression compared with treatment as usual: randomized controlled trial of MumMoodBooster. J Med Internet Res 2021 Dec 08;23(12):e17185 [FREE Full text] [doi: 10.2196/17185] [Medline: 34889742]

XSL•FO
RenderX

50. Rychtarik RG, McGillicuddy NB, Barrick C. Reaching women under stress from a partner's drinking problem: assessing interest in online help. J Technol Hum Serv 2013 Jul;31(3):185-196. [doi: 10.1080/15228835.2013.803214]
51. Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler N, et al. Powering population health research: considerations for plausible and actionable effect sizes. SSM Popul Health 2021 Jun;14:100789 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100789] [Medline: 33898730]
52. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar 29;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
53. Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science: unifying the discipline to improve the public health. Clin Psychol Sci 2014 Jan 01;2(1):22-34 [FREE Full text] [doi: 10.1177/2167702613497932] [Medline: 25821658]
54. McGillicuddy NB, Rychtarik RG, Papandonatos GD. Skill training versus 12-step facilitation for parents of substance-abusing teens. J Subst Abuse Treat 2015 Mar;50:11-17 [FREE Full text] [doi: 10.1016/j.jsat.2014.09.006] [Medline: 25306932]
55. Rychtarik RG, McGillicuddy NB. Preliminary evaluation of a coping skills training program for those with a pathological-gambling partner. J Gambl Stud 2006 Jun 14;22(2):165-178. [doi: 10.1007/s10899-006-9008-6] [Medline: 16841184]
56. Renati R, Bonfiglio NS, Rollo D. Dealing with loved ones' addiction: development of an app to cope with caregivers' stress. Int J Environ Res Public Health 2022 Nov 30;19(23):15950 [FREE Full text] [doi: 10.3390/ijerph192315950] [Medline: 36498025]

Abbreviations

AUD: alcohol use disorder
BDI-II: Beck Depression Inventory–II
CEQ: Credibility/Expectancy Questionnaire
CONSORT: Consolidated Standards of Reporting Trials
ML: maximum likelihood
OR: odds ratio
RCT: randomized controlled trial
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
SSMW: Stop Spinning My Wheels
STAXI-2: State-Trait Anger Expression Inventory 2
SUD: substance use disorder
UWC: usual web care
WAI-SF: Working Alliance Inventory–Short Form

Edited by N Cahill; submitted 09.01.24; peer-reviewed by C Roos; comments to author 23.05.24; revised version received 28.06.24; accepted 11.07.24; published 29.08.24.

Please cite as:
Rychtarik RG, Danaher BG, McGillicuddy NB, Tyler MS, Barrick C, Leong F, Kosty DB
Web-Based Coping Skills Training and Coach Support for Women Living With a Partner With an Alcohol Use Disorder: Randomized Controlled Trial
J Med Internet Res 2024;26:e56119
URL: https://www.jmir.org/2024/1/e56119
doi:10.2196/56119
PMID:

©Robert G Rychtarik, Brian G Danaher, Neil B McGillicuddy, Milagra S Tyler, Christopher Barrick, Florence Leong, Derek B Kosty. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating Patient Perspectives on Using eHealth Technologies for the Self-Management of Inflammatory Bowel Disease: Mixed Methods Study

Sander Hermsen[1,2], PhD; Danielle Tump[1], MSc; Eva Wentink[1], PhD; Marjolijn Duijvestein[3], MD, PhD

[1]OnePlanet Research Centre, Precision Health and Nutrition Group, Wageningen, Netherlands

[2]Radboud University Medical Centre, Prevention Hub, Nijmegen, Netherlands

[3]Radboud University Medical Centre, Department of Gastroenterology, Nijmegen, Netherlands

Corresponding Author:
Sander Hermsen, PhD
OnePlanet Research Centre
Precision Health and Nutrition Group
Bronland 10
Wageningen, 6708 WH
Netherlands
Phone: 31 317 791 009
Email: sander.hermsen@imec.nl

Abstract

Background: Inflammatory bowel disease (IBD) poses significant challenges for patients, requiring continuous monitoring and self-management to improve quality of life.

Objective: This study aims to investigate the viewpoints of individuals living with IBD on the use of information and communication technology (ICT) for the self-management of their condition, with a particular focus on the concept of a “smart” toilet seat as an example of ICT for IBD self-management.

Methods: We conducted an analysis of questionnaire responses obtained from 724 participants. They were encouraged to share their use cases and identify any perceived barriers associated with ICT adoption for managing their condition. To assess their responses, we used descriptive quantitative analysis, summative content analysis, and thematic qualitative analysis. We combined these results in an epistemic network analysis to look for meaningful patterns in the responses.

Results: Of the 724 participants, more than half (n=405, 55.9%) were already using various forms of ICT for IBD self-management. The primary factor influencing their use of ICT was their affinity for interacting with technology. Distinct differences emerged between individuals who were using ICT and those who were not, particularly regarding their perceived use cases and concerns.

Conclusions: This study provides valuable insights into the perspectives of individuals with IBD on the use of ICT for self-management. To facilitate wider adoption, addressing privacy concerns, ensuring data security, and establishing reliable ICT integration will be critical.

(J Med Internet Res 2024;26:e53512) doi:10.2196/53512

KEYWORDS

information and communication technology; ICT; self-management; inflammatory bowel disease; IBD; smart toilet seat; mobile phone

Introduction

Background

Inflammatory bowel disease (IBD), an umbrella term for chronic gastrointestinal diseases, including Crohn disease (CD) and ulcerative colitis (UC), affects approximately 0.3% of the population in North America and Oceania, 0.2% of the population in Europe, and rapidly rising proportions of the population in many other countries around the world [1,2]. The rising incidence has resulted in increased health care use and longer outpatient waiting lists [3]. One way to alleviate pressure on the health care system is to reduce waiting lists and health care use through patient self-management, that is, a patient’s

XSL•FO
RenderX

"ability to manage the symptoms, treatment, physical and psychosocial consequences, and lifestyle changes inherent in living with a chronic condition" [4]. Being a chronic condition, IBD entails lifelong self-management; even in remission, IBD is known to have a severe impact on mental health, with recent evidence suggesting that approximately 30% of people with IBD report symptoms of anxiety or depression [5-7]. Self-management, including the patient's ability to monitor their condition, is not only effective in relieving health care but is also well known to effectively maintain or even increase quality of life [4,8]. Self-management strategies are usually related to diet and the avoidance of perceived trigger foods [9-11], seeking support, stress reduction, experimentation, and lifestyle adaption [11,12].

In recent decades, information and communication technology (ICT) has played an ever-increasing role in the self-management of chronic conditions [13,14]. People living with IBD have been found to be generally willing to use ICT solutions for the self-management of their condition [15]. ICT solutions used for IBD self-management include websites [16]; smartphone apps [17,18]; wearable sensors [15,19]; and, lately, domotics such as "smart" toilet seats [20,21], which may soon arrive in the consumer market. Compared with standard outpatient-led care, ICT solutions (also known as eHealth technologies) have been shown to lead to improvements in relapse duration, disease activity, short-term medication adherence, quality of life and stress reduction, IBD knowledge, health care costs, and clinic visits [22].

Unfortunately, the efficacy of a majority of current ICT solutions for IBD self-management is marred by a lack of professional medical involvement and the limited use of evidence and international consensus guidelines [23]. Another important detrimental factor is a lack of patient input to inform the development of ICT solutions. Patient input is important because research shows that patients with IBD desire more information and actionable strategies with regard to managing their disease, shared decision-making, and reducing the burden of treatment and symptom control [24,25]. Furthermore, although using ICT for IBD self-management improves participants' feelings of empowerment [25], the views, needs, and desires of users, as well as the barriers they face, have as yet largely been disregarded [15,21,26]. One recent study sheds light on patient self-management strategies and needs for ICT intervention content [27], but much remains unclear about prerequisites, desires, and barriers to using ICT for the self-management of IBD, such as individual differences in, and determinants for, the use of ICT; the impact of self-management strategies and needs on the use of ICT; current ICT use; and perceived use cases.

Objectives

This study addresses the lack of patient perspectives on using ICT for the self-management of IBD in the literature by researching self-management strategies and needs, current ICT use, perceived ICT use cases, and the perceived efficacy of patients with IBD, and individual differences in ICT use in self-management. The study further elucidates perceived use cases, concerns, and barriers for a novel method to support IBD self-management by presenting participants with a description of a smart toilet seat currently under development at OnePlanet Research Center in Wageningen, Netherlands. Using low-fidelity prototypes such as descriptions or paper examples enables people to reflect on their potential use cases and barriers for ICT for IBD self-management [28,29]. The literature [21,30] shows that people differ in their use cases for smart technology, ranging from documenting aspects of their condition and using technology as a tool for diagnosis to signaling exacerbations and prescribing a course of action. Furthermore, people differ in terms of concerns and barriers, such as the transparency and visibility of technology use [31], privacy, and data agency (ie, who controls the dissemination of the data) [21]. What determines these differences, how they are interrelated, and how they relate to self-management strategies as yet remain unknown. This study contributes to addressing this gap in our knowledge.

Methods

Overview

To investigate the perspectives of people with IBD on using technology for the self-management of their condition and to identify the use cases and barriers for people with IBD regarding a smart toilet seat for IBD self-management, we performed a mixed methods study. We used a 3-part web-based questionnaire with both Likert-scale questions and questions with open-text fields. The first part of the questionnaire contained questions on demographics, health status, and disease self-management; the second part covered the use of ICT for disease self-management; and the third part focused on potential use cases and barriers for a specific ICT example, that is, a smart toilet seat. Responses to the Likert-scale questions were analyzed using descriptive statistics, correlations, and binomial general linear modeling, and these served as criteria for stratification in the qualitative analyses of the open-text questions. The answers to the open-text questions were analyzed and coded using summative content analysis and thematic qualitative analysis [32,33]. We then combined the quantitative and qualitative analyses through epistemic network analysis (ENA) [34-37], which enabled us to evaluate differences in the responses as well as their patterns across the stratification groups.

Participants

The study participants were members of the patients' association Crohn & Colitis NL (CCNL [38]). CCNL has >10,000 members and sends out a monthly email newsletter. In this newsletter, we published a brief invitation to participate in the questionnaire, which was sent out to 2274 members, of whom 1633 (71.81%) identified as female, 634 (27.88%) as male, and 6 (0.26%) as other or unknown. The average age of the participants was 48.77 (SD 15.72) years. Of the 2273 members, 568 (24.98%) had UC, 707 (31.09%) had CD, and 999 (43.93%) had not disclosed their condition when subscribing to the CCNL newsletter. The inclusion criteria for participation were as follows: having IBD (CD, UC, or as-yet-undiagnosed IBD), being aged ≥18 years, and having access to the internet and email. There were no further inclusion or exclusion criteria.

XSL•FO
RenderX

Sample size considerations were driven by the qualitative analyses. A priori sample size calculations in qualitative research are subject to conceptual debate and practical uncertainty [39]. Saturation (ie, the moment when adding more data does not lead to new insights) is often seen as a criterion for whether to include more participants once analysis has started. Although saturation as a concept is contested [39], it is still the most practical measure of information maturity. As a rule of thumb, 20 to 40 participants are usually considered sufficient to achieve saturation [40]. Another measure is information power [41]: the density, depth, and specificity of the information required affects the required sample size. In this study, the aim was relatively broad, the combination of participants was less specific for the research question, there was little theory to inform the analysis, the questionnaire methodology allowed for little dialogue, and we planned to perform cross-case analyses. Therefore, we could expect the need to include more participants for a mature data set than rule-of-thumb approaches would suggest. For the ENA approach, no sample size considerations exist as yet, but a rule-of-thumb consideration is that there should be more participants than codes in each network [33-37], and it is advisable to achieve thematic saturation within each network. In this study, the maximum number of potential networks to compare would be when comparing age groups (5 groups, resulting in 5 networks). Therefore, we aimed to include at least 200 participants—5 times as many as rules of thumb would advise—to make ENA comparisons between groups possible.

Procedure

After clicking the link in the newsletter, participants were taken to the questionnaire platform. Before entering the questionnaire, participants were presented with a consent form and asked to provide informed consent. They could exit the questionnaire at any time for any reason without any consequences. Participant data were only stored when they clicked the Send Results button at the end of the questionnaire. However, participants had the option to skip questions; any unanswered questions were recorded as missing data. Filling out the questionnaire took approximately 20 minutes. The questionnaire remained accessible for 3 weeks after participants received the invitation.

Ethical Considerations

This study was deemed exempt from ethics approval under Dutch law by the ethics review board of the Máxima Medical Centre, Veldhoven (N23.030). Participants provided informed consent before starting the questionnaire by reading and agreeing to a statement outlining the goal and contents of the questionnaire. They were informed that all data would be stored anonymously, that answers would not be analyzed on an individual level, and that they had the right to skip any question. The statement also explained that filling out the questionnaire meant providing consent for the use of their questionnaire data for analysis and in scientific and professional publications. Finally, the statement advised them that they could quit participating at any time by closing the questionnaire and refraining from submitting their responses; no data were stored before the questionnaire was completed. Neither CCNL nor the researchers had access to any names or physical or email addresses of participants. After questionnaire completion, data that could potentially pose a privacy risk, such as age and occupation details, were pseudonymized by converting quantitative data into categorical data. No compensation was offered to participants for taking part in the study.

Reflexivity

In research aimed at understanding participants' experiences, there is a potential risk of researcher bias [42]. To enhance the integrity and credibility of qualitative research, researchers must assess how intersubjective components affect data collection and analysis. An instrument for this examination is reflexivity, which refers to researchers' explicit, self-aware appraisal of their own roles [42,43].

The host institute of this study, OnePlanet Research Center, is a relatively new (it was founded 5 years ago) research institute focused on potential innovations in health and sustainability using sensor technology and artificial intelligence. One of its research programs is dedicated to gut health, with the smart toilet seat described in this study being an important part. The end goal of the program is the development of an integrated suite of sensors that informs a personal digital twin model designed to signal, measure, and prevent a range of health conditions.

SH is the principal behavioral scientist at OnePlanet Research Center, where he leads human factors research. His work focuses on the acceptability, usability, and efficacy of technological innovations designed to support healthy living. DT is a senior biomedical researcher at OnePlanet Research Center and is responsible for research projects within the smart toilet seat project. EW is the principal investigator of the smart bathroom program at OnePlanet Research Center; she leads all scientific and developmental activities for the smart toilet seat and other innovations. MD is the head of clinical research at the department of gastroenterology, Radboud University Medical Centre, Nijmegen. She specializes in treating IBD.

Questionnaire

Participants filled out a 3-part questionnaire containing 31 questions on demographics; health status and disease self-management; the use of ICT for disease self-management; and potential use cases and barriers for a specific ICT example, that is, a smart toilet seat (refer to Textbox 1). The full questionnaire, translated into English, is available in Multimedia Appendix 1.

XSL•FO
RenderX

Textbox 1. Brief details of the questionnaire. Open-text questions for qualitative analysis are displayed in italics.

Questionnaire part 1 (questions 1-17)

- Demographic data: age, sex, education level, and occupation category
- Health status: condition type (Crohn disease, ulcerative colitis, or as-yet-undiagnosed inflammatory bowel disease), years since diagnosis, the effect of the condition on everyday life, *perceived symptoms*, *comorbidities*, and recent exacerbations (the questions on symptoms and comorbidities were presented as long lists, where participants could mark checkboxes for their symptoms or comorbidities and then supplement the list with their own experiences in an open-text field)
- Disease self-management: ability to predict an exacerbation (flare), *predictors of exacerbations*, knowing the potential cause of the exacerbation, *perceived causes of exacerbations*, perceived influence on exacerbations, *self-management strategies*, and *self-management needs* (these questions were asked only if the participant indicated having had an exacerbation in the past year)

Questionnaire part 2 (questions 18-23)

- The use of information and communication technology (ICT) for condition self-management: short version of the Affinity for Technology Interaction questionnaire [44], the use of ICT for self-management, *the type of ICT used*, *perceived support for self-management through ICT*, *reasons not to use ICT*, *and perceived uses of ICT for self-management* (these questions were asked only if the participant indicated having had an exacerbation in the past year)
- A short description of a smart toilet seat, currently in development at OnePlanet Research Center; the description, which contained images of a toilet seat equipped with sensors, read (in Dutch): "Currently, OnePlanet Research Center is developing a special toilet with which you can, in the future, measure and observe all kinds of processes. The toilet performs its measurements automatically, all you need to do is sit down as on a regular toilet. In the toilet seat and elsewhere are small instruments that can measure, for instance, your heartbeat, breath, body temperature or urine or stool composition. By simply using the toilet, you can find out many things about your health. You can use an app or your smartphone, or the toilet can make a direct connection with your doctor"

Questionnaire part 3 (questions 24-31)

- Questions about the smart toilet seat: desire to use, perceived efficacy, *perceived use cases for the smart toilet seat*, *perceived barriers*, *data agency*, *who can receive data*, and *concerns*; finally, participants had the opportunity to leave a comment about the questionnaire

Analysis

Data analysis consisted of 3 steps: quantitative analysis of the results of the questions used for stratification, summative content analysis and thematic qualitative analysis of the open-text question results, and ENA of the qualitative data using the quantitative results as stratification factors.

Quantitative Analyses

Quantitative analyses were performed using RStudio for Mac (version 2023.03.0+386; Posit Software, PBC) [45]. The questionnaire data were first imported into RStudio. Where necessary, data were converted into factorial format and labeled. For numerical data, we calculated means, SDs, minimum and maximum values, and medians. For all data, we printed bar plot histograms for visual information. We then calculated the heterogenous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables. Next, we qualified the correlations by testing binomial general linear models for the 2 main research questions: what predicts the use of ICT for self-management? What predicts willingness to use the smart toilet seat innovation? Finally, all quantitative data were gathered in a data frame for use as stratification factors in the ENA step, using the *source_from_df* function from the *rock* package for R [46].

Qualitative Analyses

All source files were cleaned using the *clean_sources* function from the *rock* package. For each participant, answers to open-text questions were then grouped together per questionnaire part: (1) condition characteristics (ie, symptoms and comorbidities) and self-management (ie, predictors of exacerbations, perceived causes, self-management strategies, and self-management needs), (2) ICT use for IBD self-management (ie, which hardware and software participants currently use, perceived support for self-management through ICT, reasons not to use ICT, and desired uses of ICT), and (3) the smart toilet seat (ie, perceived use cases of the toilet, perceived barriers, data agency, who can receive data, and concerns).

All data were then manually segmented into utterances, and given a unique utterance identifier through the *prepend_ids_to_sources* function from the *rock* package. One researcher performed a summative content analysis [47] on the questions on symptoms, comorbidities, predictors of exacerbations, perceived causes, self-management strategies, self-management needs, and current technology use. Summative content analysis is well suited for counting and comparing keywords within content [47]. The researcher then performed a thematic analysis [32,33] on the results from the remaining open-text questions—*perceived support for self-management through ICT*, *reasons not to use ICT*, and *desired uses of ICT for self-management* from questionnaire part 2; *perceived use cases*, *perceived barriers*, *data agency*, *who can receive data*, and *concerns* from questionnaire part 3; and the open-text comment section—by coding and recoding all utterances using the i.rock web-based coding tool, a tool for creating human- and machine-readable coded qualitative data and then identifying themes within the codes. Thematic analysis is well suited to identify and interpret patterns of meaning within open qualitative data [33]. All coded utterances were then gathered in a data

XSL•FO
RenderX

frame to use as analysis units in the ENA step, using the *parse_sources* function from the *rock* package.

Combined Analysis: ENA

To look for meaningful patterns and differentiations between subgroups in the qualitative data, we performed an ENA using the ENA web tool (version 1.7.0) [48,49]. ENA is a technique for modeling the structure of connections in qualitative data. It assumes that (1) it is possible to systematically identify a set of meaningful features in the data (*codes*), (2) the data have local structure (*conversations*), and (3) an important feature of the data is the way that codes are *connected* to one another within conversations [34-37]. ENA models the connections among codes by quantifying the co-occurrence of codes within a given moving stanza window in the conversations (in our analysis, this moving stanza window was set to 6 utterances), producing a weighted network graph in which nodes correspond to the codes, and edges reflect the relative frequency of co-occurrence (or connection) between 2 codes. The positions of the network graph nodes are fixed, and those positions are determined by an optimization routine that minimizes the difference between the plotted points and their corresponding network centroids. Because of this coregistration of network graphs and projected space, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the projected space and explain the positions of plotted points in the space. Furthermore, the size of the nodes indicates the importance of the code for the network. We created graphs by selecting all utterances from each unique participant as *conversations*; the codes associated with one of the themes derived from the thematic qualitative analysis as *codes*; and the different values of the stratification factors derived from the quantitative analysis, plus the unique participants, as *units*.

Results

Demographics, Health Condition, and Self-Management

Of the 2274 potential respondents, 724 (31.84%) participants with IBD filled out the questionnaire (ie, approximately 0.8%-1% of all people with IBD in the Netherlands) [50]. Participant demographic information (sex, age, and education; questions 1-3) is presented in Table 1. The results of the questionnaire are discussed in the subsections that follow; the results of part 1 (condition characteristics and self-management) are available in Table 1 (participant characteristics; questions 1-3), Table 2 (condition characteristics; questions 5, 6, and 10), Table 3 (differences in symptoms between participants with CD and those with UC; questions 8 and 9), and Table 4 (self-management skills, needs, and strategies; questions 11-17). The results of part 2 (ICT use for IBD self-management) are available in Table 5 (hardware and software use; questions 18-20). The results from the thematic analyses and ENAs on perceived use cases for ICT (questions 21-23) and the smart toilet seat (questions 24-31) are presented in detail in the following subsections. As the ENA may be unfamiliar to many readers, and it results in a large number of outcomes that may be difficult to interpret, we have added bullet points summarizing the main findings in each ENA subsection. An overview of all main themes, subthemes, codes, and code frequencies is available in Multimedia Appendix 2.

Table 1. Participant characteristics (questions 1-3; n=724).

Characteristics	Participants, n (%)
Sex	
Female	489 (67.5)
Male	233 (32.2)
Other	0 (0)
Not answered	2 (0.3)
Age group (y)	
<31	76 (10.5)
31-40	109 (15.1)
41-50	117 (16.2)
51-65	261 (36)
>65	161 (22.2)
Highest education achieved	
Primary school	5 (0.7)
Primary vocational education	219 (30.2)
Secondary vocational education	72 (9.9)
Higher professional or university education	423 (58.5)
Not answered	5 (0.7)

XSL•FO
RenderX

Table 2. Condition characteristics (questions 5, 6, and 10; n=724).

Characteristics	Participants, n (%)
Condition type	
Ulcerative colitis	304 (42)
Female	182 (59.9)
Male	122 (40.1)
Crohn disease	394 (54.6)
Female	292 (74.1)
Male	102 (25.9)
As-yet-undiagnosed inflammatory bowel disease	24 (3.3)
Female	15 (62.5)
Male	9 (37.5)
Not answered	2 (0.03)
Years since diagnosis	
<1	36 (5)
1-5	128 (17.7)
6-10	134 (18.5)
10-25	237 (32.7)
>25	185 (25.6)
Not answered	4 (0.6)
Exacerbations in past year	
I have had an exacerbation	389 (53.7)
I have not had an exacerbation	283 (39.1)
I do not know	52 (9.9)

Table 3. Symptoms for which participants with ulcerative colitis (UC) and those with Crohn disease (CD) mentioned different occurrences (questions 8 and 9; n=724).

Participants' condition	Blood and slime in stool, n (%)	Tenesmus, n (%)	Frequent toilet visits, n (%)	Pain in joints, n (%)	Stomachache, n (%)
UC (n=304)	220 (72.4)	116 (38.4)	228 (75)	94 (30.9)	186 (61.2)
CD (n=395)	170 (43)	79 (20)	238 (60.2)	156 (39.5)	280 (70.9)

XSL•FO
RenderX

Table 4. Self-management skills (questions 11-17; n=283, ie, all participants who indicated having had an exacerbation in the past year).

Self-management skills	Participants, n (%)
Perceived ability to detect exacerbations	
Never	86 (30.6)
Sometimes	135 (48)
Often	47 (16.7)
Always	13 (0.1)
Not answered	2 (0.1)
Knowing what causes exacerbations	
No	79 (27.9)
Yes	203 (71.7)
Can influence an exacerbation	
No	184 (65)
Yes	99 (35)
Exacerbation predictors	
Stomachache	71 (25.1)
Fatigue	70 (24.7)
Change in stool or diarrhea	41 (14.4)
Increased toilet visits	37 (13.1)
Stress	32 (11.3)
Perceived exacerbation causes	
Stress	158 (55.8)
Diet and trigger foods	105 (37.1)
Work and life pressures	45 (15.9)
Self-management needs	
Social support	105 (37.1)
Informational support	81 (28.6)
Resilience and optimism	64 (22.6)
Action planning or practical	35 (12)
No need for support	35 (12)
Self-management strategies	
Sleep and relaxation	65 (22.9)
Healthy diet, avoid triggers	51 (18)
Stress avoidance	36 (12.7)
Changing medication	28 (9.9)

XSL•FO
RenderX

Table 5. Using information and communication technology (ICT) for inflammatory bowel disease (IBD) self-management (questions 18-20; n=724).

	Participants, n (%)	ATI[a] score						
		1	2	3	4	5	6	7
Uses ICT for IBD self-management								
Total	724 (100)	33	157	46	190	109	85	102
Yes	405 (55.9)	20	96	19	83	35	17	27
No	300 (41.4)	12	56	26	100	70	66	74
I don't know	18 (2.5)	1	3	0	7	4	2	1
Not answered	1 (0.1)	0	2	1	0	0	0	0
ICT hardware use								
Smartphone	116[b] (16)	—[c]	—	—	—	—	—	—
Wearable tracker	108 (14.9)	—	—	—	—	—	—	—
Smartwatch	99 (13.6)	—	—	—	—	—	—	—
Smart domotics	11 (1.5)	—	—	—	—	—	—	—
No ICT fits me	119 (16.4)	—	—	—	—	—	—	—
ICT software use								
MyIBDcoach	99 (13.6)	—	—	—	—	—	—	—
IBD information sites	47 (6.5)	—	—	—	—	—	—	—
Hospital-provided apps	46 (6.4)	—	—	—	—	—	—	—
Generic health apps	32 (4.4)	—	—	—	—	—	—	—

[a]ATI: Affinity for Technology Interaction.

[b]It is important to bear in mind that this does not mean that only 116 participants use a smartphone; participants with low ATI scores are more likely to talk about their ICT use for IBD self-management in more abstract terms, mentioning "smartphone" or "apps and websites," whereas participants with high ATI scores are more likely to not mention the smartphone but only mention the specific apps they are using for the self-management of their condition (eg, Strava, an activity tracking app, or DEARhealth, an IBD self-management app).

[c]Not applicable.

Saturation

The coding process took place in batches of results from 25 participants and resulted in 165 codes. In the first batch, we used 137 unique codes; the second batch yielded another 10 unique codes, the third batch another 14 codes, the fourth batch another 7 codes, and the fifth batch another 9 codes. After these 5 batches, new codes never exceeded a threshold value of 2.5%, indicating full code saturation after 125 participants. Furthermore, all cells in the ENA analyses exceeded the rule-of-thumb threshold of 40 participants.

Using ICT for IBD Self-Management

Quantitative Results

A small majority of the participants (405/724, 55.9%) reported currently using ICT for the self-management of their condition. This use had no significant correlations with any other determinant (eg, age, sex, education, and condition type), except with the score on the self-reported Affinity for Technology Interaction (ATI) questions. Participants who reported using ICT for self-management of IBD had a higher average ATI score (mean 4.61, SD 1.8031) than those who reported using no ICT (μ=3.57, r_{722}=0.2937, SE 0.03979; P<.001). To assess which factors best predicted the use of ICT for the self-management of IBD, we performed a binomial general linear model analysis. This analysis revealed a significant contribution of the ATI score to ICT use for self-management; again, participants with higher ATI scores had a greater chance of also using ICT (estimate=0.50323, SE 0.08592, z=5.857; P<.001); furthermore, there were weakly significant (P<.05) contributions of years since diagnosis, with participants diagnosed for >25 years having a slightly smaller chance of using ICT for self-management (estimate=–1.49372, SE 0.69406, z=–2.152; P=.03); and perceived influence on exacerbations, with perceived influence having a small positive effect on the chance of using ICT (estimate=0.84647, SE 0.32910, z=2.572; P=.01). Table 5 presents an overview of the results from the questions related to using technology for the self-management of IBD.

Qualitative Results: Using ICT for Self-Management (Thematic Analysis)

Thematic analysis made it possible to construct 3 main themes in perceived use cases for ICT in IBD self-management: *learning about your condition*, *living with your condition*, and *emotional and peer support*. First, participants stated that ICT could be useful in learning about their condition. Many participants mentioned using ICT to document various aspects of their condition, such as symptom occurrence and severity, as well as documenting the occurrence and impact of the

XSL•FO
RenderX

perceived determinants of their condition, such as nutrition, stress, and sleep. Some participants used this information for self-diagnosis or to prognosticate about their condition. Participants also mentioned using ICT to find information about their condition, and some participants expressed a wish that the medical profession should make greater use of ICT:

> *[I use the] digital patient contact system to keep sight on appointments, to communicate with my care team, and for insight in [blood] test results, so I know quickly if everything is under control and I can remain in the driver's seat.* [Participant 76, male, aged 35 years, CD]
>
> *Mapping your own values, so you know if anything deviates. The standards in medical measurement values do not apply to everyone. What would for most people be a regular blood pressure or body temperature would for me be a cause for alarm, so I need to know my own values. If I get a flare-up, it is practical to be able to look back at my values and identify the onset. I notice, for instance in hospital admissions, that they rely on numbers from theory too much. They should listen to the patient more.* [Participant 113, female, aged 48 years, CD]
>
> *The stool app provides a nice clear overview. The hospital app is supposed to make communication with the hospital easier and more accessible, but in my experience this does not work. I still need to call them and wait for them to call back. IBDream does not give me added value, but the hospital prefers it.* [Participant 266, female, aged 35 years, CD]

A second theme concerned the practical aspects of living with the condition. Participants mentioned using activity tracking and other forms of monitoring to ensure that they maintained a healthy lifestyle to better enable them to deal with their condition and change their behavior (*directive/behavior change*). Participants also mentioned using ICT to connect with their care professionals, which reduced the burden their medical care entails, for instance, by opting for online appointments instead of making in-person visits and asking questions through apps and emails. Finally, participants mentioned using apps to find nearby toilets:

> *Because [the app] makes it more transparent how active I am and how much progress I make, I regain confidence in my body.* [Participant 27, male, aged 36 years, UC]
>
> *An immediate overview of everything I need for a healthy life. For instance, Fitbit's recovery score. It measures sleep and activity and calculates a daily recovery score. If I have a very bad day, the app provides me with the confirmation I need to take a step back.* [Participant 373, female, aged 44 years, UC]
>
> *When I was ill, I could make connections between flare-ups, heart beats, and a deteriorating cardiovascular condition. My activity went down to 0 immediately. I could not exert myself without immediately having to run for the toilet. Since I have been capable to do sport again, I like to keep track of my progress. At work, I take between ten thousand and eighteen thousand steps per day.* [Participant 218, male, aged 30 years, UC]
>
> *I like to know if I have been active enough each day, that's why I use the activity tracker—I think being active helps me. Also, I feel exhausted quickly because of my condition, so I like to be able to check how my sleep was.* [Participant 527, female, aged 20 years, CD]

A third theme concerned using ICT for emotional and peer support. Participants mentioned using ICT to connect with, and find out about, others with their condition, as well as to stay calm and detect stress:

> *Social media: by following others with IBD [or other autoimmune conditions] I feel less alone in my process. I have dear friends, but they will never completely understand what it means to have a chronic condition.* [Participant 249, female, aged 41 years, UC]
>
> *I like to read about situations that I recognize, so I know I am not the only one. And sometimes tell my story, get support, advice, tips.* [Participant 526, male, aged 66 years, CD]
>
> *It helps me find rest, so I know everything is taken care of and I don't get stressed, for example, about not getting my medicine in time.* [Participant 88, male, aged 57 years, UC]

ENA Results: Using ICT for Self-Management

The main findings in this analysis were as follows:

- All participants saw ICT as a method to connect to health care professionals and find information.
- Participants who were already using ICT also did so to document their condition and monitor physical activity.
- Participants who were not using ICT saw ICT as a tool for diagnosis, behavior change support, and finding toilets.

To find out whether participants who were using ICT for IBD self-management differed from those who were not using ICT regarding their perceived use cases for ICT, we performed an ENA using all codes from the *perceived use cases* subgroup as codes, with scores on the binomial variable *using technology for self-management* plus participant numbers as units, and all utterances from the questionnaires as conversations, with a moving stanza window of 6 utterances. Figure 1A shows the network for the response *no* (ie, participants who reported not using ICT for self-management, n=300), and Figure 1B shows the network for the response *yes* (ie, participants who reported using ICT for self-management, n=405).

XSL•FO
RenderX

Figure 1. Epistemic network analysis results for all codes from the "perceived use cases for information and communication technology" (ICT) subgroup. (A) The network for the response No (ie, participants who reported not using ICT for inflammatory bowel disease [IBD] self-management; n=300). (B) The network for the response Yes (ie, participants who reported using ICT for IBD self-management; n=405). The abbreviations DIAG (diagnosing condition), DOC (documenting condition), LE (learning and finding information) and MPS (medical profession should do more) apply to the theme "learning about condition." The abbreviations AT (activity tracking or keeping active), CCP (connecting with care professionals), DIR (directive tracking or behavior change), FT (finding toilets), RB (reducing the burden of medical care), and SHD (supporting a healthy diet) apply to the theme "living with condition." The abbreviations CP (connecting with peers), DS (detecting stress), and SC (staying calm) apply to the theme "emotional and social support.".

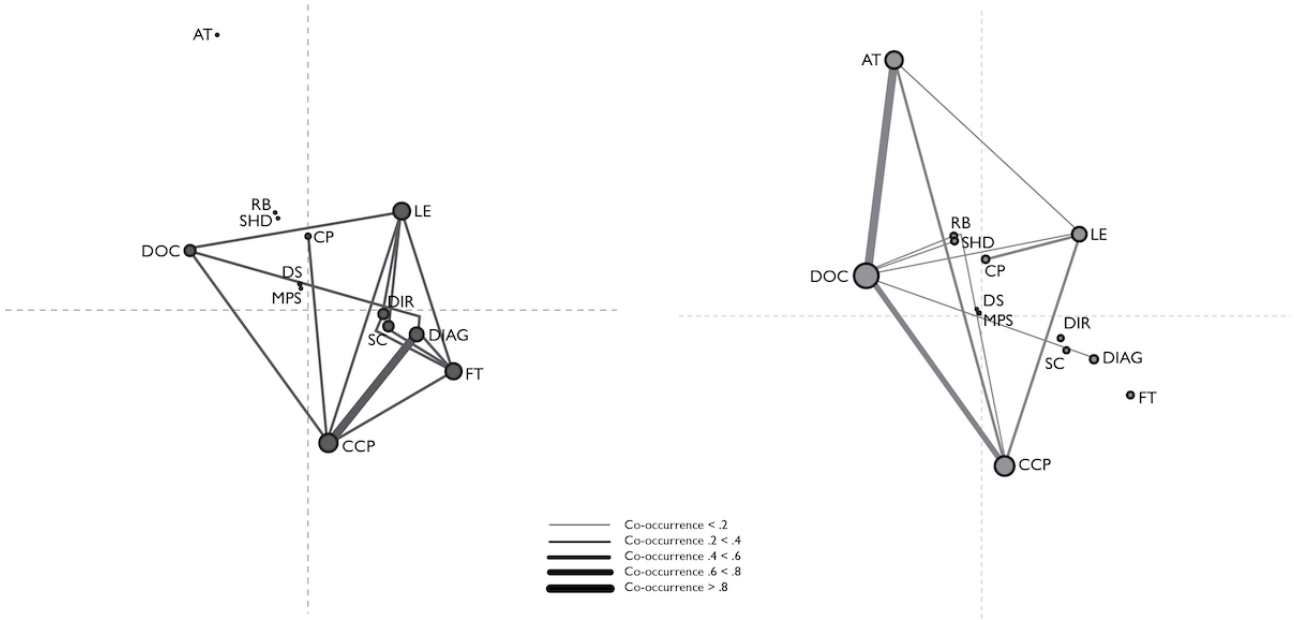

There are 3 sources of information in each graph: node size, node position, and node connections. The node size refers to how often the depicted code co-occurs with any other code, which can be seen as a measure of relative importance of the code. In Figure 1, *documenting condition*, *connecting with care professionals*, and *activity tracking/keeping active* are the codes with the most connections overall for participants who were using ICT for self-management (Figure 1B), whereas *connection with care professionals*, *learning/finding information*, *diagnosing condition*, and *finding toilets* are the codes with the most connections for participants who were not using ICT for self-management (Figure 1A). When it comes to node position, nodes at a central position in the graph have a smaller contribution to the explained variance in the model than nodes in more peripheral positions. Again, the most important contributors are *documenting condition* and *activity tracking/keeping active* for people who were using ICT; and *connecting with care professionals*, *diagnosing condition*, *learning/finding information*, and *finding toilets* for people who were not using ICT. Finally, connection strength, depicted as the thickness of the lines between the nodes, shows the relative frequency of the co-occurrence of codes. In participants who were using ICT for self-management, the strongest connections are between *documentingcondition* and *connecting with care professionals* and between *documentingcondition* and *activity tracking/keeping active*. *Finding toilets*, *staying calm*, *detecting stress*, and *directive tracking/behavior change* do not play a role in the network.

For participants who were not using ICT, the strongest connection is between *documentingcondition* and *diagnosing condition*. *Activity tracking/keeping active*, *reducing the burden of medical care*, *supporting a healthy diet*, and *medical professionals should make greater use of ICT* (eg, participants expressing an opinion that their medical professional should put in more effort) do not play a role.

Using the Smart Toilet Seat

Quantitative Results

A majority of the participants stated that they would like to try out the smart toilet seat described in the questionnaire, with 59.7% (432/724) answering *yes*, 15.9% (115/724) answering *no*, and 23.8% (172/724) answering *I don't know* to the question whether they would want to use it. When asked whether they thought the smart toilet seat would help them manage their condition, of the 724 participants, 349 (48.2%) answered *yes*, 125 (17.3%) answered *no*, and 244 (33.7%) answered *I don't know*. Wanting to use the smart toilet seat correlated significantly with age group membership, with the older age groups more reluctant to use the smart toilet seat than the younger age groups (r_{718}=–0.05512, SE 0.04477; P<.001) and those with a lower ATI score more reluctant to use the smart toilet seat (r_{718}=–0.09369, SE 0.04179; P<.001). We performed a binomial general linear model analysis to assess what determinants contribute to predicting whether participants would want to use the smart toilet seat. Responding *sometimes* on *perceived ability to predict exacerbations* was the only significant contributor to *expressed desire for smart toilet seat use* (estimate=1.54390, SE 0.50561, z=3.054; P=.002).

Qualitative Results: Smart Toilet Seat Use Cases (Thematic Analysis)

Participants mentioned a diverse range of use cases they envision for the smart toilet seat. Again, we could construct 3 themes:

XSL•FO
RenderX

learning about the condition, *living with the condition*, and *negative use cases*. In the first theme, *learning about the condition*, documenting symptoms, perceived determinants, and the effects of the condition again play a large role. Similarly, participants saw the smart toilet seat as an instrument to support diagnosis, with many participants mentioning specific diagnostic use cases such as measuring stool composition and urine composition as well as inflammatory markers such as fecal calprotectin. However, when thinking of use cases for the smart toilet seat, participants also mentioned use cases that did not come up previously when discussing the use cases of ICT in general, such as automatically and unobtrusively signaling significant changes in their condition (eg, the onset of an exacerbation):

> *It would be nice if inflammation markers in stool is available sooner, so I can take action quicker without being confined to my house immediately. Especially the inflammation markers in stool. This also makes it easier if certain nutrients affect this, and sound the alarm if inflammation markers go up without any direct increase in symptoms.* [Participant 59, female, aged 44 years, UC]

> *If it were possible to figure out where exhaustion comes from and maybe tackle the causes, or determine an energy level so I can anticipate that, I would be very interested. That would encourage me to act when I see inflammations, let me find out about bad or one-sided nutrition or moisture balance, a lack of certain bacteria, or stool composition.* [Participant 103, female, aged 40 years, CD]

> *I think it's mostly about taking away insecurities about my condition. Are my values normal or not? Do I need to contact my specialist or GP [general practitioner] yet? I would like to reduce that insecurity.* [Participant 322, female, aged 63 years, CD]

A second theme is *living with the condition*, which concerns use cases aimed at reducing the burden of medical care and informing behavioral change:

> *Predicting flare-ups and providing tips and instructions that fit my current situation [activity and body monitoring] and personal healthy daily schedule accordingly. Maybe even connect to my diary or advise me to cancel a meeting because of deteriorations. I think that early signaling and implementation of the right protective measures, a flare-up can be reduced, delayed, or even prevented. But it is hard to recognize the right signals and give priority to the right actions within the mess of everyday life. A machine or app can measure much more effectively and provide programmed actions based on measured values. That relieves me as human from the daily hassle of dealing with my condition, which would take away a lot of stress. It might even, in the long run, reduce fatigue and give more energy.* [Participant 512, male, aged 37 years, UC]

Finally, participants also mentioned a range of *negative use cases*. Of the 724 participants, 112 (15.5%) stated that they did not feel any need for the smart toilet seat, would rather not use it, or did not know whether the toilet would have any benefits for them. Some of the participants said that they could not use the toilet because of a stoma or other physical condition:

> *I think, right now, for me it would be "overkill" to use this, because I am doing fine on my current medication, for now. I would rather do my own tests, when I feel the need.* [Participant 305, female, aged 57 years, UC]

> *That too much relies on the smart measurements. If you know your own body well, you will feel that things are off, and you lose that knowledge when you solely rely on measurements.* [Participant 165, female, aged 62 years, UC]

> *I would rather live without technology. What will be, will be.* [Participant 332, male, aged 67 years, CD]

ENA Results: Smart Toilet Seat Use Cases

The main findings in this analysis were as follows:

- We compared codes occurring in 4 groups of participants: those who were not using ICT for self-management and would not like to try the smart toilet seat (N-N), those who were using ICT for self-management and would not like to try the smart toilet seat (Y-N), those who were not using ICT for self-management and would like to try the smart toilet seat (N-Y), and those who were using ICT for self-management and would like to try the smart toilet seat (Y-Y).
- Documenting their condition was seen as an important use case for the smart toilet seat by all groups, apart from group N-N.
- Group N-N often mentioned *no need for the smart toilet seat*, *would rather not use*, and *cannot use because of stoma* (as a reason).
- Group N-Y often mentioned a combination of *signaling changes and directive tracking* (ie, receiving an alarm when something is amiss and being given advice on what to do). *Measuring inflammation markers*, *using for diagnosis* (in combination with *signaling changes*), and *measuring stool composition* also occurred often.
- Group Y-N members often mentioned that they did not know what to use the toilet for and that current advice sufficed; when discussing use cases, they most often mentioned documenting their condition, combined with reducing the burden of medical care and measuring calprotectin.
- Group Y-Y most often mentioned *documenting for (self-)diagnosis*, with a strong connection with *signaling changes*. Furthermore, the group members often mentioned *measuring calprotectin*, *stool composition*, *and inflammation markers*, as well as *reducing the burden of medical care*.

To find out whether participants who were using ICT for IBD self-management differed from those who were not using ICT regarding their perceived use cases for the smart toilet seat, and

XSL•FO
RenderX

whether, within these 2 groups, participants who said they would like to use the smart toilet seat differed from those who said they did not want to use the toilet, we performed an ENA using all codes from the *smart toilet seat perceived use cases* subgroup as codes and all utterances from the questionnaires as conversations, with a moving stanza window of 6 utterances. In this ENA, we compared scores on 2 variables (*using ICT for IBD self-management* and *desire to use the smart toilet seat*) to compose four units: (1) those who were not using ICT for IBD self-management and did not want to use the smart toilet seat (N-N, n=70), (2) those who were not using ICT for IBD self-management but were interested in using the toilet (N-Y, n=141), (3) those who were using ICT for IBD self-management but did not want to use the smart toilet seat (Y-N, n=42), and (4) those who were using ICT for IBD self-management and wanted to use the smart toilet seat (Y-Y, n=282). In this analysis, participants who answered *I don't know whether I want to use the smart toilet seat* (those who were not using ICT for IBD self-management: n=86, those who were using ICT for IBD self-management: n=79) were left out. Figures 2A-2D show the networks for the 4 groups (N-N, N-Y, Y-N, and Y-Y*)*.

XSL•FO
RenderX

Figure 2. Epistemic network analysis results for all codes from the "perceived use cases for the smart toilet seat subgroup." (A) Code network for participants who were not using information and communication technology (ICT) for self-management and would not want to use the smart toilet seat (N-N). (B) Code network for participants who were not using ICT for self-management and would want to use the smart toilet seat (N-Y). (C) Code network for participants who were using ICT for self-management and would not want to use the smart toilet seat (Y-N). (D) Code network for participants who were using ICT for self-management and would want to use the smart toilet seat (Y-Y). The abbreviations DIAG (diagnosing condition), DOC (documenting condition), LE (learning or finding information), MC (measuring calprotectin), MIM (measuring inflammation markers), MSC (measuring stool composition), SIG (signaling changes), and TAI (tailoring treatment to my situation) apply to the theme "learning about condition." The abbreviations DIR (directive tracking or behavior change), LT (love of technology), and RB (reducing the burden of medical care) apply to the theme "living with condition." The abbreviations CAS (current advice suffices), CNS (cannot because of stoma), DK (don't know), NN (no need), and RN (rather not) apply to the theme "negative use cases.".

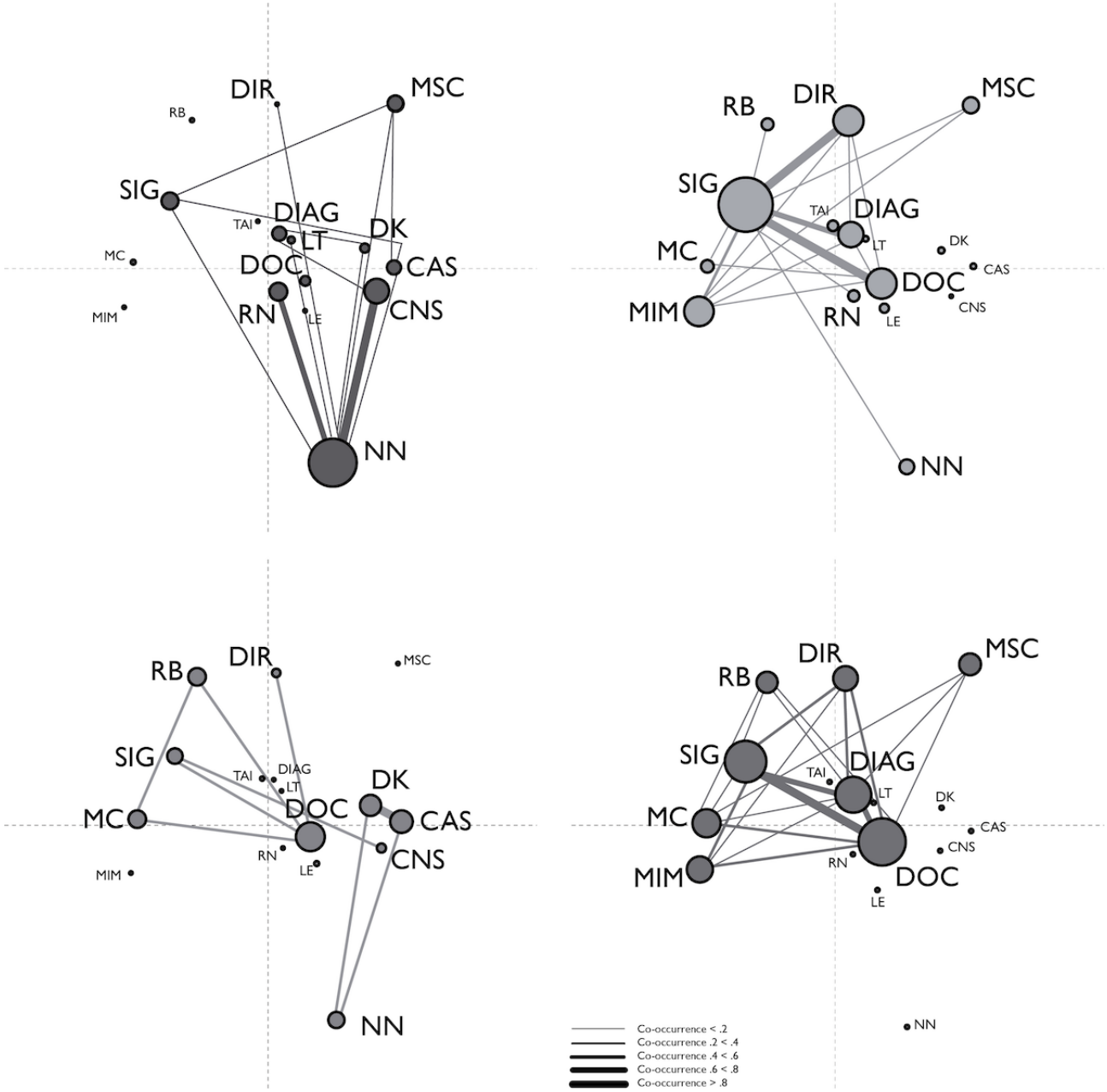

Again, information can be derived from node sizes, node positions, and connections between nodes. The node size refers to how often the depicted code co-occurs with any other code, which can be seen as a measure of the relative importance of the code; the position of the node shows the code's relative contribution to the differences between units, and the connection strength between codes shows the relative frequency of co-occurrence of codes. In the N-N group, the codes *no need*, *rather not*, and *cannot use because of stoma* have the most overall connections, with the code *no need* also making the largest contribution to differences with the other groups. When it comes to connections, the combination of *no need* with *rather not* and with *cannot use because of stoma* are the most prominent. Some nodes remain unconnected and play no role for this group: *reducing the burden of medical care*, *tailoring treatment to my situation*, *measuring calprotectin*, *measuring inflammation markers*, and *learning/finding information*. In the N-Y group, the code *signaling changes* has the most connections overall, with further well-connected nodes being *directive tracking/behavior change*, *measuring inflammation markers*,

XSL•FO
RenderX

documenting condition, and *diagnosing condition*. Nodes with the most contribution to differences with other groups are *signaling changes* and *directive tracking/behavior change*. The nodes with the strongest connections are the combinations *signaling changes+directive tracking/behavior change* and *signaling changes+documenting condition*. Nodes with no connections are *current advice suffices*, *don't know*, and *cannot because of stoma*. In the Y-N group, there are no nodes that have a distinctively larger size than others, with *documenting condition* being the largest node. Similarly, connections between nodes for this group are relatively weak, with the exception of the connection between the nodes *don't know* and *current advice suffices*. Finally, in the Y-Y group, the nodes corresponding with the codes *documenting condition*, *signaling changes*, and *diagnosingcondition* are the largest, with *signaling changes* and specific cases of diagnostic tracking (*measuring stool composition*, *measuring calprotectin*, and *measuring inflammation markers*) contributing the most to the differences with the other groups. When it comes to inflammation markers, this group mentioned the specific marker (calprotectin) more often than the general term (inflammation markers), whereas the N-Y group mentioned the general term more than the specific marker. When it comes to connection strength, the combinations *documenting condition+signaling changes* and *signaling changes+diagnosing condition* are the strongest.

Qualitative Results

Concerns and Data Agency: Summative Content Analysis

Regarding data agency, a majority of the participants (418/724, 57.7%) stated that they would like their data to be always available to themselves (eg, through a smartphone app or dashboard); 17% (123/724) wanted to receive reports only when something happened that necessitated action; 8% (58/724) would not want constant access to data, but they would want regular summaries of key information; and 3.2% (23/724) wanted reports to go directly to their health care provider only.

Concerns and Data Agency: Thematic Analysis

Besides perceived use cases, participants also mentioned perceived concerns about using the smart toilet seat. For these concerns, 4 major themes could be constructed: *privacy and data agency*, *practical concerns*, *concerns about dealing with the condition*, and *validity*. First, many participants cited *privacy and data agency* (ie, who has control over accessibility to the data) as a major concern:

> *Privacy issues; my data can end up with the wrong persons or companies, or worse: get connected—not anonymously—to my name and other personal data.* [Participant 13, female, aged 53 years, CD]
>
> *Evidence of my disease in facts. That is a relief, because you do not have to constantly think about how you are doing. The question that remains, though, is who else can see these data. You don't want to share these data with just anybody, maybe even with nobody at all.* [Participant 242, male, aged 62 years, CD]

A second theme involves practical concerns. The cost of purchasing and having a smart toilet seat, how the toilet can distinguish between different users, technological issues, breakdowns, hassle, and hygienic concerns all played a role for many participants:

> *It will probably be costly, that makes sense, but it is a barrier.* [Participant 178, female, aged 40 years, CD]
>
> *How can the smart toilet seat figure out who is on the toilet? It is important that different people can use the toilet without mixing up data.* [Participant 260, female, aged 60 years, UC]
>
> *The whole digital business, it hardly ever works for me, or only partially.* [Participant 294, male, aged 78 years, CD]

A third theme involves participants' concerns about how a smart toilet seat will affect living and dealing with their condition. Some participants mentioned a fear of unnecessary worries; they do not want a constant reminder of their illness and want to avoid overmedicalization of their condition or the replacement of real-life contact with health care professionals by machines:

> *It only leads to more unrest, I think. I don't want to be thinking about my disease all day, I want to think about being healthy.* [Participant 16, female, aged 52 years, CD]
>
> *I think that every time an abnormal measurement comes up, I would get very stressed, and that is a problem if your gut responds to stress strongly.* [Participant 320, female, aged 63 years, CD]
>
> *That in the end, a doctor's visit is deemed superfluous by health insurance. A patient is more than their excrement.* [Participant 289, female, aged 72 years, CD]

A final, fourth theme involves concerns about the *validity* of the smart toilet seat measurements. This theme encompasses concerns about the validity and reliability of measurements, especially for participants' specific situation, and unnecessary data generation:

> *I think this will not help me, because blood and excrement tests have never shown an inflammation with me.* [Participant 337, female, aged 59 years, CD]
>
> *I don't see how the information gathered here could help me manage my condition.* [Participant 538, female, aged 42 years, CD]

ENA Results: Smart Toilet Seat Use Concerns

The main findings in this analysis were as follows:

- A shared concern for all groups is privacy.
- People who were not using ICT for self-management (N-N and N-Y groups) often mentioned fear that the toilet would not work, unnecessary worries, hassle, and technical issues.
- People who would not like to try the smart toilet seat (N-N and Y-N groups) mentioned constant reminders of their illness and overmedicalization of their condition as concerns, as well as the reliability and validity of the smart toilet seat's measurements.

XSL•FO
RenderX

- The N-N group also often mentioned unnecessary data generation and the replacement of health care professionals by technology as potential downsides.
- The Y-N group also often mentioned cost, data agency, and the replacement of health care professionals by technology as potential downsides.
- The N-Y group also mentioned concerns about whether others could use the toilet without messing up the measurements and data agency as potential downsides.
- The Y-Y group had mostly practical concerns: potential issues with the technology, others using the smart toilet seat, cost, and data agency.

To find out whether participants who were using ICT for IBD self-management differed from those who were not using ICT regarding their concerns about using the smart toilet seat, and whether, within these 2 groups, participants who said that they would like to use the smart toilet seat differed from those who said they did not want to use the toilet, we performed an ENA using all codes from the *smart toilet seat concerns* subgroup as codes and all utterances from the questionnaires as conversations, with a moving stanza window of 6 utterances. In this ENA, as before, we compared scores on *using ICT for IBD self-management* and *desire to use the smart toilet seat* to compose four units: (1) N-N, n=70; (2) N-Y, n=141; (3) Y-N, n=42; and (4) Y-Y, n=82 (Figures 3A-3D, respectively).

XSL•FO
RenderX

Figure 3. Epistemic network analysis results for all codes from the "concerns about the smart toilet seat" subgroup. (A) Code network for participants who were not using information and communication technology (ICT) for self-management and would not want to use the smart toilet seat (N-N). (B) Code network for participants who were not using ICT for self-management and would want to use the smart toilet seat (N-Y). (C) Code network for participants who were using ICT for self-management and would not want to use the smart toilet seat (Y-N). (D) Code network for participants who were using ICT for self-management and would want to use the smart toilet seat (Y-Y). The abbreviations AC (agency control) and PR (privacy) apply to the theme privacy and data agency. The abbreviations AE (aesthetics and design), CO (cost), HY (hygiene), ISS (technical issues), OTH (others using the toilet), TMH (too much hassle), and UNP (unpleasant) apply to the theme "practical concerns." The abbreviations CRI (constant reminder of illness), MED (overmedicalization of condition), RHP (replacement of health care professionals by technology), and UW (unnecessary worries) apply to the theme "concerns about dealing with the condition." The abbreviations REV (reliability and validity), DNW (won't work [for me]), and UD (unnecessary data generation) apply to the theme "validity.".

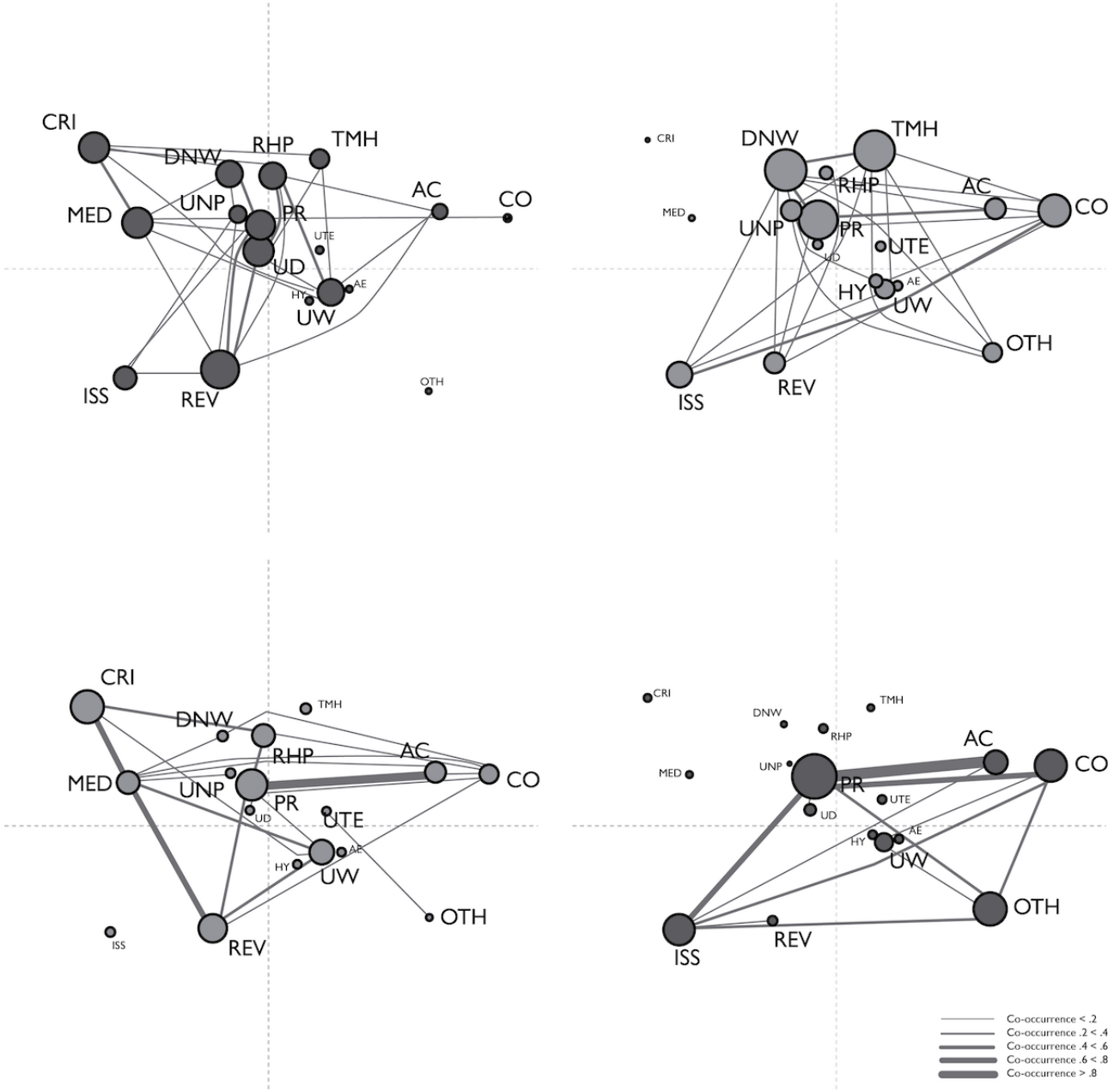

As before, information can be derived from node sizes, node positions, and connections between nodes. The ENA shows that *privacy* was a concern for every participant, and therefore the position of the node shows that it does not contribute to distinguishing between the groups. However, the connections between *privacy* and other nodes differ for the 4 participant groups. Apart from *privacy*, when it comes to node size, all codes from the theme *concerns about dealing with the condition* play an important role in the N-N group: *constant reminder of illness*, *overmedicalization of condition*, *unnecessary worries*, and *replacement of health care professionals by technology*. Furthermore, codes from the theme *validity*, such as *won't work (for me)*, *reliability and validity*, and *unnecessary data generation*, are most notable. No edges between nodes stand out particularly, although the combination of *constant reminder of illness* with *overmedicalization of condition* occurs more often than others. For the position of nodes, *constant reminder of illness*, *overmedicalization of condition*, and *reliability and validity* contribute the most to distinguishing this group from others. Not connected are nodes from the *practical concerns*

XSL•FO
RenderX

theme, such as *others using the toilet*, *hygiene*, *aesthetics and design*, and *using toilets elsewhere*. In the N-Y group, besides *privacy*, the most notable nodes are *cost*, *issues with technology*, and *too much hassle* from the *practical concerns* theme; and *won't work (for me)* from the *validity* theme. Nodes without connections are *constant reminder of illness*, *overmedicalization of condition*, *replacement of health care professionals by technology*, *aesthetics and design*, and *unnecessary datageneration*. No connections between nodes stand out in particular, although the combination of *privacy* with *agency control* and that of *cost* with *issues with technology* have stronger connections than others. Node position shows that *issues with technology*, *cost*, *won't work (for me)* and *too much hassle* contribute the most to distinguishing this group from the others. In the group Y-N, *constant reminder of illness*, *overmedicalization of condition*, *unnecessary worries*, and *replacement of health care professionalsby technology* from the *dealing with condition* theme; *reliability and validity* from the *validity* theme; *privacy* and *agency control* from the *privacy* theme; and *cost* from the *practical concerns* theme are the most prominent. Connections between *constant reminder of illness* and *overmedicalization of condition*, *overmedicalization of condition* and *reliability and validity*, and *privacy* and *agency control* stand out. As for node position, *constant reminder of illness*, *overmedicalization of condition*, *reliability and validity*, *agency control*, and *cost* contribute the most to distinguishing this group from other groups. *Issues with technology*, *too much hassle*, *unnecessary datageneration*, *hygiene*, and *aesthetics and design* have no connections with other nodes. Finally, in the Y-Y group, the node sizes of *privacy* and *agency control* from the *privacy* theme as well as *cost*, *others using the toilet*, and *issues with technology* from the *practical concerns* theme stand out. There are strong connections between *privacy* and *agency control*, *privacy* and *issues with technology*, and *privacy* and *cost*. Node position shows that *issues with technology*, *cost*, *others using the toilet*, and *agency control* contribute the most to distinguishing this group from others. Nodes from the *concerns aboutdealing with the condition* theme (eg, *constant reminder of illness*, *overmedicalization of condition*, and *replacement of health care professionals by technology*) and those from the *validity* theme (eg, *won't work* [*for me*]), have no connections.

Discussion

Principal Findings

In this study, we aimed to investigate the perspectives of people with IBD regarding the use of ICT for self-management of their condition. To do so, we constructed a questionnaire and analyzed the responses provided by 724 participants. Through this analysis, we sought to identify the use cases and barriers perceived by individuals with IBD in adopting ICT-based solutions for managing their health condition. Among the participants in the study, more than half (405/724, 55.9%) were already using various forms of ICT for IBD self-management. The primary factor influencing their current use of ICT was their affinity for interacting with technology. The questionnaire also shed light on self-management needs and skills among persons with IBD. A large majority (never: 86/283, 30.6%; sometimes: 135/283, 48%) could often not predict exacerbations, and although many participants (203/283, 71.7%) thought they knew what caused exacerbations, most (203/283, 71.7%) did not know how to influence them. Participants indicated needs for social support (105/283, 37.1%), information support (81/283, 28.6%), and resilience and optimism (64/283, 22.6%); their current strategies were sleep and relaxation (65/283, 22.9%), avoiding trigger foods and maintaining a healthy diet (51/283, 18%), stress avoidance (36/283, 12.7%), and changes in medication (28/283, 9.9%).

Through ENA, distinct differences emerged between individuals who were using ICT and those who were not using ICT, particularly regarding their perceived use cases. Participants who were using ICT viewed it as a means to document their condition, monitor physical activity, and establish connections with health care professionals to share their insights. By contrast, the individuals not using ICT regarded it as a tool to document their condition, aiming to support professional diagnosis.

Upon considering the perceived cases and concerns related to using a smart toilet seat as an example of ICT for IBD self-management, the sample population can be categorized into 4 distinct groups. First, those who were not using ICT for the self-management of their condition and had no interest in using the smart toilet seat (N-N), accounted for 13.1% (70/535) of the respondents who provided a response to the question regarding their interest in the technology. Second, those who were not using ICT but expressed a desire to use the smart toilet seat (N-Y) constituted 26.4% (141/535) of the sample. Third, individuals who were using ICT for IBD self-management but were not interested in using the smart toilet seat (Y-N) represented 7.9% (42/535) of the participants. Finally, the majority of the respondents (282/535, 52.7%) were individuals who were using ICT for IBD self-management and expressed an interest in using the smart toilet seat (Y-Y).

These groups exhibited distinct variations in perceived use cases and concerns associated with the adoption of a smart toilet seat for IBD self-management. The N-N and Y-N groups lacked clear use cases for the smart toilet seat. The N-Y group members perceived the smart toilet seat as a means to identify detrimental changes in their condition, offering directions for action, and documenting information for health care professionals' diagnoses. By contrast, although the Y-Y group members also viewed the smart toilet seat as a means to signal negative changes in their condition, they stood apart from the N-Y group in that they stated their intention to document these changes for their own self-management and decision-making.

Similarly, these groups expressed different concerns regarding the use of the smart toilet seat. The N-N group members expressed worries about the potential impact on daily life with their condition, including constant reminders of their health status. In addition, practical concerns, such as privacy and the accuracy of the measurements, were also prominent for this group. The Y-N group mainly raised concerns related to data agency and privacy issues. The N-Y group emphasized privacy concerns and practical aspects such as the potential hassle and associated costs. Finally, the Y-Y group members, again, shared concerns about privacy and data agency, along with the

XSL•FO
RenderX

apprehension of others using the smart toilet seat, potentially compromising their personal information and health data.

This research provides new insights about patient perspectives on using ICT for IBD self-management, which are currently not available in the literature. Some of the results confirm earlier work with more general participant groups [21], in which signaling and documenting also appear as important use cases, with privacy and data agency emerging as major concerns. However, this study also offers valuable insights into effectively addressing the needs of individuals who exhibit an inclination toward technology interaction but remain hesitant to embrace proposed interventions. Providing assurance regarding privacy and data agency becomes essential in catering to this particular group.

The research shows that there is also a group of people who have low affinity for technology and would therefore not want to use an ICT-based intervention. They believed that such an intervention would be too much of a hassle and that it was not for them. These opinions need to be respected. However, a substantial number of participants in this group (N-N) talked about wanting to avoid constant reminders of their condition and overmedicalization of their condition. This may be a sign of detrimental coping by hiding or denying their condition [42,51], which may impact IBD coping and self-management negatively [42,43]. It is imperative to conduct further research in this area to gain a deeper understanding of the implications of these coping strategies on IBD coping and self-management. By exploring these aspects more comprehensively, we can develop tailored approaches and interventions that better align with the needs and preferences of individuals with low technology affinity while promoting effective disease management and self-care.

When it comes to other results from this study, we see that there is a participant bias in that more female individuals than male individuals participated. Although the participant rate for female and male individuals (60% and 40%, respectively) reflects sex differences in the occurrence of IBD in the Dutch population [52,53], the occurrence of CD and UC in the sample in this paper (approximately 55% and 45%, respectively) differs from population data (approximately 40% and 60%, respectively) [52,53]. This may be caused by the aforementioned participation bias, a recurring pattern in questionnaire research [54,55]. Similarly, the literature [54,55] on questionnaire research indicates that it is relatively common for older participants to be overrepresented compared to population numbers. However, in this study, we think this is not problematic because there were sufficient male participants (233/724, 32.2%) to achieve saturation and stratification, as well as adequate representation of people aged <31 years (76) and those aged between 31 and 40 years (109). Furthermore, highly educated participants were overrepresented in the sample. Notably, those with only primary school education, who can be expected to have limited health literacy, were underrepresented. Given the fact that this group often has limited language skills [52] and is therefore less likely to participate in web-based questionnaires, further research is necessary to learn about their needs and barriers.

Finally, when it comes to use cases for ICT for IBD in general, 3 major themes were constructed: *learning about the condition*, *dealing with the condition*, and *connecting with others*. The smart toilet seat only covers the first 2 themes; the third theme is difficult to integrate in this solution because toilet-related topics are often awkward to talk about or even taboo [21]. To fulfil people's needs around connecting with others, a stand-alone solution would be preferable, especially considering the concerns that people have about privacy and data agency.

Conclusions and Recommendations

In conclusion, this study provides valuable insights into the perspectives of individuals with IBD on using ICT for self-management. The positive reception of the participants to the smart toilet seat concept suggests the potential for innovative ICT solutions in empowering patients to actively participate in their care. To facilitate wider adoption, it will be critical to address privacy concerns, ensure data security, and establish reliable ICT integration. Furthermore, the results of our research show that one-size-fits-all solutions will likely fail to reach many potential users of technological innovations for disease self-management because they do not connect to their perceived use cases or do not address their concerns. Use cases differ from simply signaling when something is amiss and providing an action perspective to informing personal scientific experiments. Patients' concerns differ from fear of overmedicalization of their condition and constant reminders of the illness to concerns about cost, privacy, and data agency. This paper not only shows that people differ but also provides evidence that there are possibilities for profiling large groups of potential users with similar needs and barriers. The results of this study and methodologically similar future studies can inform implementation and dissemination strategies for technological innovations. Future research and development should not only focus on refining and validating the technologies to enhance their effectiveness and user acceptance, with extra focus on patients with limited health literacy or low affinity with technology interaction, but also strive to shed greater light on who the user is, what their needs and requirements are, and how their lived experience of their condition informs their use of technology for self-management.

Acknowledgments

This work was financially supported by the Dutch Digestive Foundation (in Dutch: Maag Lever Darm Stichting; WOO 23-06). OnePlanet Research Center acknowledges the province of Gelderland for their funding. Special thanks go out to Menne Scherpenzeel and Daniëlle van der Horst of the patient association Crohn & Colitis NL for their insights and help setting up and disseminating the inflammatory bowel disease questionnaire, to Gjalt-Jorn Peters of Open University for their help with using the *rock* package for R, and especially to Szilvia Zörgő of Maastricht University for their methodological advice.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. SH, DT, and EW set up the questionnaire. DT supervised the data collection. SH performed the main analysis, with contributions from all authors. SH prepared the draft manuscript, and all authors contributed to the final manuscript. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full questionnaire, translated into English.
[DOCX File , 136 KB - jmir_v26i1e53512_app1.docx]

Multimedia Appendix 2
Overview of themes, subthemes, codes, and code frequencies for each part of the questionnaire. Part 1: disease characteristics and self-management; part 2: information and communication technology use in self-management; and part 3: the smart toilet seat. Codes from disease characteristics, predictors, perceived causes, self-management strategies, self-management needs, and hardware and software use are derived from the summative content analysis. All other codes are derived from the thematic analysis.
[PDF File (Adobe PDF File), 280 KB - jmir_v26i1e53512_app2.pdf]

References

1. Zhao M, Gönczi L, Lakatos PL, Burisch J. The burden of inflammatory bowel disease in Europe in 2020. J Crohns Colitis 2021 Sep 25;15(9):1573-1587 [FREE Full text] [doi: 10.1093/ecco-jcc/jjab029] [Medline: 33582812]
2. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet 2017 Dec 23;390(10114):2769-2778. [doi: 10.1016/S0140-6736(17)32448-0] [Medline: 29050646]
3. Kaplan GG. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol 2015 Dec;12(12):720-727. [doi: 10.1038/nrgastro.2015.150] [Medline: 26323879]
4. Barlow J, Wright C, Sheasby J, Turner A, Hainsworth J. Self-management approaches for people with chronic conditions: a review. Patient Educ Couns 2002;48(2):177-187. [doi: 10.1016/s0738-3991(02)00032-0] [Medline: 12401421]
5. Neuendorf R, Harding A, Stello N, Hanes D, Wahbeh H. Depression and anxiety in patients with inflammatory bowel disease: a systematic review. J Psychosom Res 2016 Aug;87:70-80. [doi: 10.1016/j.jpsychores.2016.06.001] [Medline: 27411754]
6. Byrne G, Rosenfeld G, Leung Y, Qian H, Raudzus J, Nunez C, et al. Prevalence of anxiety and depression in patients with inflammatory bowel disease. Can J Gastroenterol Hepatol 2017;2017:6496727 [FREE Full text] [doi: 10.1155/2017/6496727] [Medline: 29181373]
7. Mikocka-Walus A, Massuger W, Knowles SR, Moore GT, Buckton S, Connell W, et al. Psychological distress is highly prevalent in inflammatory bowel disease: a survey of psychological needs and attitudes. JGH Open 2020 Apr;4(2):166-171 [FREE Full text] [doi: 10.1002/jgh3.12236] [Medline: 32280760]
8. Kamp KJ, Luo Z, Holmstrom A, Given B, Wyatt G. Self-management through social support among emerging adults with inflammatory bowel disease. Nurs Res 2019;68(4):285-295 [FREE Full text] [doi: 10.1097/NNR.0000000000000354] [Medline: 31265438]
9. Fletcher PC, Schneider MA. Is there any food I can eat? Living with inflammatory bowel disease and/or irritable bowel syndrome. Clin Nurse Spec 2006;20(5):241-247. [doi: 10.1097/00002800-200609000-00011] [Medline: 16980794]
10. Jamieson AE, Fletcher PC, Schneider MA. Seeking control through the determination of diet: a qualitative investigation of women with irritable bowel syndrome and inflammatory bowel disease. Clin Nurse Spec 2007;21(3):152-160. [doi: 10.1097/01.NUR.0000270015.97457.9c] [Medline: 17495550]
11. DuBois K, Blake C, Rudisill C, Harrison S, Hébert J. P031 Perceptions among patients with ulcerative colitis: treatment and self management methods. Gastroenterology 2020 Feb;158(3):S106 [FREE Full text] [doi: 10.1053/j.gastro.2019.11.249]
12. Skrastins O, Fletcher PC. "One flare at a time": adaptive and maladaptive behaviors of women coping with inflammatory bowel disease and irritable bowel syndrome. Clin Nurse Spec 2016;30(5):E1-11. [doi: 10.1097/NUR.0000000000000229] [Medline: 27509570]
13. Lawn S, Schoo A. Supporting self-management of chronic health conditions: common approaches. Patient Educ Couns 2010 Aug;80(2):205-211. [doi: 10.1016/j.pec.2009.10.006] [Medline: 19931372]

XSL•FO
RenderX

14. Celler BG, Lovell NH, Basilakis J. Using information technology to improve the management of chronic disease. Med J Aust 2003 Sep 01;179(5):242-246. [doi: 10.5694/j.1326-5377.2003.tb05529.x] [Medline: 12924970]
15. Hirten RP, Stanley S, Danieletto M, Borman Z, Grinspan A, Rao P, et al. Wearable devices are well accepted by patients in the study and management of inflammatory bowel disease: a survey study. Dig Dis Sci 2021 Jun 23;66(6):1836-1844. [doi: 10.1007/s10620-020-06493-y] [Medline: 32705439]
16. Langille M, Bernard A, Rodgers C, Hughes S, Leddin D, van Zanten SV. Systematic review of the quality of patient information on the internet regarding inflammatory bowel disease treatments. Clin Gastroenterol Hepatol 2010 Apr;8(4):322-328. [doi: 10.1016/j.cgh.2009.12.024] [Medline: 20060070]
17. Yin AL, Hachuel D, Pollak JP, Scherl EJ, Estrin D. Digital health apps in the clinical care of inflammatory bowel disease: scoping review. J Med Internet Res 2019 Aug 19;21(8):e14630 [FREE Full text] [doi: 10.2196/14630] [Medline: 31429410]
18. Walsh A, Travis S. What's app? Electronic health technology in inflammatory bowel disease. Intest Res 2018 Jul;16(3):366-373 [FREE Full text] [doi: 10.5217/ir.2018.16.3.366] [Medline: 30090035]
19. Chong KP, Woo BK. Emerging wearable technology applications in gastroenterology: a review of the literature. World J Gastroenterol 2021 Mar 28;27(12):1149-1160 [FREE Full text] [doi: 10.3748/wjg.v27.i12.1149] [Medline: 33828391]
20. Wang XJ, Camilleri M. A smart toilet for personalized health monitoring. Nat Rev Gastroenterol Hepatol 2020 Aug;17(8):453-454. [doi: 10.1038/s41575-020-0320-x] [Medline: 32483351]
21. Hermsen S, Verbiest V, Buijs M, Wentink E. Perceived use cases, barriers, and requirements for a smart health-tracking toilet seat: qualitative focus group study. JMIR Hum Factors 2023 Aug 11;10:e44850 [FREE Full text] [doi: 10.2196/44850] [Medline: 37566450]
22. Jackson BD, Gray K, Knowles SR, De Cruz P. eHealth technologies in inflammatory bowel disease: a systematic review. J Crohns Colitis 2016 Sep;10(9):1103-1121. [doi: 10.1093/ecco-jcc/jjw059] [Medline: 26928960]
23. Con D, De Cruz P. Mobile phone apps for inflammatory bowel disease self-management: a systematic assessment of content and tools. JMIR Mhealth Uhealth 2016 Feb 01;4(1):e13 [FREE Full text] [doi: 10.2196/mhealth.4874] [Medline: 26831935]
24. Al Khoury A, Balram B, Bessissow T, Afif W, Gonczi L, Abreu M, et al. Patient perspectives and expectations in inflammatory bowel disease: a systematic review. Dig Dis Sci 2022 Jun;67(6):1956-1974 [FREE Full text] [doi: 10.1007/s10620-021-07025-y] [Medline: 34021425]
25. Walsh A, Matini L, Hinds C, Sexton V, Brain O, Keshav S, et al. Real-time data monitoring for ulcerative colitis: patient perception and qualitative analysis. Intest Res 2019 Jul;17(3):365-374 [FREE Full text] [doi: 10.5217/ir.2018.00173] [Medline: 31146510]
26. Byron C, Savage E, Lehane E, Burton A, Cornally N. Self-management interventions to support help-seeking behaviours in inflammatory bowel disease: a scoping review. Gastrointest Nurs 2022 Jul 02;20(6):30-41. [doi: 10.12968/gasn.2022.20.6.30]
27. Fawson S, Dibley L, Smith K, Batista J, Artom M, Windgassen S, et al. Developing an online program for self-management of fatigue, pain, and urgency in inflammatory bowel disease: patients' needs and wants. Dig Dis Sci 2022 Jul;67(7):2813-2826 [FREE Full text] [doi: 10.1007/s10620-021-07109-9] [Medline: 34146202]
28. Buchenau M, Suri JF. Experience prototyping. In: Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques. 2000 Presented at: DIS '00; August 17-19, 2000; New York, NY. [doi: 10.1145/347642.347802]
29. Yasar AU. Enhancing experience prototyping by the help of mixed-fidelity prototypes. In: Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology. 2007 Presented at: Mobility '07; September 10-12, 2007; Singapore, Singapore. [doi: 10.1145/1378063.1378137]
30. Rooksby J, Rost M, Morrison A, Chalmers M. Personal tracking as lived informatics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON. [doi: 10.1145/2556288.2557039]
31. O'Kane AA, Rogers Y, Blandford AE. Concealing or revealing mobile medical devices?: designing for onstage and offstage presentation. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15; April 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702453]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
33. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications; 2021.
34. Bowman D, Swiecki Z, Cai Z, Wang Y, Eagan B, Linderoth J, et al. The mathematical foundations of epistemic network analysis. In: Proceedings of the Second International Conference on Advances in Quantitative Ethnography. 2021 Presented at: ICQE 2020; February 1-3, 2021; Malibu, CA. [doi: 10.1007/978-3-030-67788-6_7]
35. Shaffer DW, Collier W, Ruis AR. A tutorial on epistemic network analysis: analyzing the structure of connections in cognitive, social, and interaction data. J Learn Anal 2016;3(3):9-45. [doi: 10.18608/jla.2016.33.3]
36. Shaffer DW, Hatfield D, Svarovsky GN, Nash P, Nulty A, Bagley E, et al. Epistemic network analysis: a prototype for 21st-century assessment of learning. Int J Learn Media 2009 May;1(2):33-53. [doi: 10.1162/ijlm.2009.0013]

XSL•FO
RenderX

37. Shaffer D, Ruis A. Epistemic network analysis: a worked example of theory-based learning analytics. In: Handbook of Learning Analytics. Edmonton, AB: Society for Learning Analytics and Research; 2017:175-187.
38. Crohn & Colitis NL. URL: https://www.crohn-colitis.nl/ [accessed 2024-08-29]
39. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qual Res Sport Exerc Health 2019 Dec 26;13(2):201-216. [doi: 10.1080/2159676x.2019.1704846]
40. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148 [FREE Full text] [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]
41. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
42. Sirois FM, Molnar DS, Hirsch JK. Self-compassion, stress, and coping in the context of chronic illness. Self Identity 2015 Jan 09;14(3):334-347. [doi: 10.1080/15298868.2014.996249]
43. Zhang M, Hong L, Zhang T, Lin Y, Zheng S, Zhou X, et al. Illness perceptions and stress: mediators between disease severity and psychological well-being and quality of life among patients with Crohn's disease. Patient Prefer Adherence 2016 Nov;Volume 10:2387-2396. [doi: 10.2147/ppa.s118413]
44. Wessel D, Attig C, Franke T. ATI-S - an ultra-short scale for assessing affinity for technology interaction in user studies. In: Proceedings of Mensch und Computer 2019. 2019 Presented at: MuC '19; September 8-11, 2019; Hamburg, Germany. [doi: 10.1145/3340764.3340766]
45. RStudio: integrated development environment for R. Posit Software. URL: https://posit.co/ [accessed 2024-08-29]
46. Peters GJ, Zorgo S. rock: reproducible open coding kit. The Comprehensive R Archive Network. 2024 Jan 21. URL: https://cran.r-project.org/web/packages/rock/index.html [accessed 2024-08-23]
47. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov 01;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
48. Marquart CL, Hinojosa C, Swiecki Z, Eagan B, Shaffer DW. Epistemic network analysis (version 1.7. 0)[software](2018). Epistemic Network Analysis. 2018. URL: http://app.epistemicnetwork.org [accessed 2024-08-23]
49. Marquart CL, Swiecki Z, Collier W, Eagan B, Woodward R, Shaffer DW. rENA: epistemic network analysis. The Comprehensive R Archive Network. 2024 Feb 23. URL: https://cran.rstudio.com/web/packages/rENA/ [accessed 2024-08-23]
50. de Groof EJ, Rossen NG, van Rhijn BD, Karregat EP, Boonstra K, Hageman I, et al. Burden of disease and increasing prevalence of inflammatory bowel disease in a population-based cohort in the Netherlands. Eur J Gastroenterol Hepatol 2016 Sep;28(9):1065-1072. [doi: 10.1097/MEG.0000000000000660] [Medline: 27281507]
51. Defenbaugh NL. Revealing and concealing ill identity: a performance narrative of IBD disclosure. Health Commun 2013;28(2):159-169. [doi: 10.1080/10410236.2012.666712] [Medline: 22545720]
52. Göritz AS, Stieger S. The impact of the field time on response, retention, and response completeness in list-based web surveys. Int J Hum Comput Stud 2009 Apr;67(4):342-348. [doi: 10.1016/j.ijhcs.2008.10.002]
53. van den Heuvel TR, Jonkers DM, Jeuring SF, Romberg-Camps MJ, Oostenbrug LE, Zeegers MP, et al. Cohort profile: the Inflammatory Bowel Disease South Limburg cohort (IBDSL). Int J Epidemiol 2017 Apr 01;46(2):e7. [doi: 10.1093/ije/dyv088] [Medline: 26045509]
54. Porter SR, Whitcomb ME. Non-response in student surveys: the role of demographics, engagement and personality. Res High Educ 2005 Mar;46(2):127-152. [doi: 10.1007/s11162-004-1597-2]
55. Slauson‐Blevins K, Johnson KM. Doing gender, doing surveys? Women's gatekeeping and men's non‐participation in multi‐actor reproductive surveys. Sociol Inq 2016 Jun 17;86(3):427-449. [doi: 10.1111/soin.12122]

Abbreviations

ATI: Affinity for Technology Interaction
CCNL: Crohn & Colitis NL
CD: Crohn disease
ENA: epistemic network analysis
IBD: inflammatory bowel disease
ICT: information and communication technology
UC: ulcerative colitis

XSL•FO
RenderX

Edited by K Williams; submitted 09.10.23; peer-reviewed by K Kamp, A Yin; comments to author 16.12.23; revised version received 29.03.24; accepted 16.04.24; published 06.09.24.

Please cite as:
Hermsen S, Tump D, Wentink E, Duijvestein M
Investigating Patient Perspectives on Using eHealth Technologies for the Self-Management of Inflammatory Bowel Disease: Mixed Methods Study
J Med Internet Res 2024;26:e53512
URL: https://www.jmir.org/2024/1/e53512
doi:10.2196/53512
PMID:39240663

©Sander Hermsen, Danielle Tump, Eva Wentink, Marjolijn Duijvestein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing the Efficacy of 2 WeChat Mini Programs in Reducing Nonmarital Heterosexual Contact by Male Factory Workers: Randomized Controlled Trial

Kechun Zhang[1*], MSc; Bolin Cao[2*], PhD; Yuan Fang[3*], PhD; Xue Liang[4], MPhil; Danhua Ye[4], MPH; Ya Qi Chen[1], MPH; Ruilan Zhong[1], BSc; He Cao[1], MSc; Tian Hu[1], BSc; Ting Li[2], BSc; Yong Cai[5], PhD; Huachun Zou[6], PhD; Zixin Wang[4], PhD

[1]Longhua District Center for Disease Control and Prevention, Shenzhen, China

[2]School of Media and Communication, Shenzhen University, Shenzhen, China

[3]Department of Health and Physical Education, The Education University of Hong Kong, Hong Kong, China (Hong Kong)

[4]Centre for Health Behaviours Research, The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[5]Public Health Department, Hongqiao International Institute of Medicine, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[6]School of Public Health, Fudan University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Zixin Wang, PhD
Centre for Health Behaviours Research, The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong
Room 508, School of Public Health, Prince of Wales Hospital
30-32 Ngan Shing Street
Hong Kong
China (Hong Kong)
Phone: 852 22528740
Fax: 852 26453098
Email: wangzx@cuhk.edu.hk

Abstract

Background: Male factory workers in China are vulnerable to HIV transmission. Commercial and nonmarital noncommercial contacts are the driving forces of heterosexual HIV transmission among male factory workers in China. There is a lack of effective HIV interventions for male factory workers in China.

Objective: The primary objective of this randomized controlled trial was to compare the efficacy of an enhanced versus the standard version of a WeChat mini program in reducing sexual intercourse with nonregular female sex partners and female sex workers among male factory workers in Shenzhen, China.

Methods: A nonblinded 2-arm parallel randomized controlled trial was conducted between December 2021 and April 2023. Participants were adult male factory workers in Shenzhen who had access to a smartphone and WeChat. Those who had oral or anal sex with a man or self-reported as HIV positive were excluded. A total of 247 participants were randomly assigned to the intervention group (n=125, 50.6%) or the control group (n=122, 49.4%); 221 (89.5%) and 220 (89.1%) completed follow-up surveys at T1 (6 months after completion of the interventions) and T2 (6 months after T1). Participants in the control group had access to the standard WeChat mini program that provided basic HIV-related knowledge and information about local free HIV testing services. Participants in the intervention group had access to the enhanced WeChat mini program. The enhanced mini program covered all the information in the standard mini program. In addition, the enhanced mini program assessed users' behaviors and invited users to watch different web-based videos on reducing nonmarital sexual contacts and promoting HIV testing based on their behavioral characteristics at months 0 and 1. The videos were developed based on in-depth interviews with male factory workers. Intention-to-treat analysis was used for outcome analyses. Multiple imputation was used to replace missing outcome values at T1 and T2.

XSL•FO
RenderX

Results: At T1, fewer participants in the intervention group reported sexual intercourse with a nonregular female sex partner in the past 6 months compared with the control group (1/125, 0.8% vs 8/122, 6.6%; relative risk=0.12, 95% CI 0.02-0.96; *P*=.02). However, there were no between-group differences in sexual intercourse with a nonregular female sex partner at T2 (10/125, 8% vs 14/122, 11.5%; *P*=.36) or sexual intercourse with a female sex worker at T1 (2/125, 1.6% vs 2/122, 1.6%; *P*=.98) or T2 (8/125, 6.4% vs 8/122, 6.6%; *P*=.96).

Conclusions: The enhanced WeChat mini program was more effective than the standard WeChat mini program in reducing sexual intercourse with nonregular female sex partners among male factory workers in the short term but not in the longer term. Improvements should be made to the WeChat mini program before implementation.

Trial Registration: ClinicalTrials.gov NCT05811611; https://clinicaltrials.gov/study/NCT05811611

(J Med Internet Res 2024;26:e49362) doi:10.2196/49362

KEYWORDS

nonmarital heterosexual contacts; male factory workers; WeChat mini program; randomized controlled trial; China; mobile phone

Introduction

Overall Situation of HIV Transmission in China

Worldwide, there were 1.5 million new HIV infections and 7.1 to 82 million other new sexually transmitted infection (STI) cases in 2020 [1,2]. In China, the HIV and STI burdens have been increasing over the last few decades [3]. As of 2020, more than 1 million people are living with HIV, and heterosexual contacts continue to be the main transmission route of HIV and other STIs in China; 69.9% and 30.6% to 70% of new HIV and other STI cases, respectively, have been attributed to heterosexual contacts [4]. A sex-specific analysis indicated that the prevalence and mortality rate of HIV in the male population were 2.9 and 3.5 times higher than those in the female population in China, respectively [5]. Factory workers in China, especially male factory workers, are vulnerable to HIV transmission [6]. In recent years, commercial and nonmarital noncommercial contacts, the driving forces of heterosexual HIV transmission in China, have been prevalent among male factory workers [7]. A recent survey of 1361 male factory workers in Shenzhen showed that 24.5% and 21.2% of them had had sexual intercourse with nonregular female sex partners and female sex workers, respectively, in the previous 6 months [7]. Approximately 80% of new HIV cases reported between 2016 and 2018 were among male migrant workers [8].

HIV-Related Interventions Implemented in China

Sustained efforts have been made by the Chinese government to enhance HIV treatment and care. Recently, the treatment coverage exceeded 90% for people living with HIV in China, along with a treatment success rate of >90% [5]. Regarding HIV risk reduction interventions, most efforts were focused on men who have sex with men and female sex workers. A large number of risk reduction interventions aimed to improve HIV-related knowledge, promote consistent condom use during anal intercourse, increase HIV testing coverage, and implement pre-exposure prophylaxis among men who have sex with men [9-13]. Systematic reviews and meta-analyses have found positive impacts of these interventions in reducing sexual risk behaviors and improving HIV testing uptake [9]. HIV risk reduction interventions targeting female sex workers mainly focus on improving consistent condom use with male clients and HIV testing uptake. A systematic review and meta-analysis of >120 studies in China found that these interventions could effectively improve consistent condom use and HIV testing uptake among female sex workers [14]. However, relatively few HIV risk reduction interventions targeted factory workers [6]. There were only 2 HIV interventions targeting female workers instead of male factory workers, who are at higher risk of HIV in China [15,16]. Although studies have shown that eHealth HIV interventions could reach more participants with low cost and can significantly reduce sexual risk behaviors and increase HIV testing among high-risk groups [17,18], no studies have applied eHealth HIV interventions for factory workers [6]. Therefore, it is necessary to examine the efficacy of eHealth interventions in reducing sexual risk behaviors among factory workers in China using a more rigorous evaluation design.

WeChat is a popular social media site in China with 800 million active users worldwide. WeChat mini programs are “mini-applications” built within the WeChat platform. WeChat allows third parties to develop such mini programs to provide advanced features to users. Users do not need to download or install these mini programs on their smartphones, they only need to scan a QR code or search to open and run these mini programs within WeChat. These features of WeChat mini programs improve their accessibility and convenience. WeChat mini programs have been increasingly used in health promotion and disease management, such as promoting physical activity [19], breastfeeding [20], preventing falls [21], and managing cancer symptoms [22]. To our knowledge, no studies have evaluated the effectiveness of WeChat mini programs in changing HIV-related behaviors. In 2021, the Longhua District Center for Disease Control and Prevention developed a WeChat mini program for factory workers, which provided HIV-related knowledge and information of HIV testing services. However, the WeChat mini program did not have any components targeting nonmarital heterosexual contacts, which are a key risk factor of HIV acquisition among factory workers. In this study, the research team enhanced the standard WeChat mini program by adding interventions that aimed to reduce nonmarital heterosexual contacts among male factory workers.

Objectives and Hypothesis

The primary objective of this randomized controlled trial (RCT) was to compare the efficacy of an enhanced version versus the standard version of the WeChat mini program in reducing sexual

XSL•FO
RenderX

intercourse with nonregular female sex partners and female sex workers in the previous 6 months at T1 (6 months after completing the interventions) and T2 (6 months after T1) among male factory workers in Shenzhen, China. The secondary objectives were to compare the between-group difference in condomless sex with nonregular female sex partners and female sex workers and HIV testing uptake at T1 and T2. We hypothesized that the enhanced version of the WeChat mini program would be more effective in changing the study outcomes than the standard version.

Methods

Study Design

The COVID-19 situation and its control measures in Shenzhen during the study period are presented in Figure 1. We reported this trial in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [23].

Figure 1. The COVID-19 situation and its control measures in Shenzhen during the study period. There were no official new case data after December 17, 2022. PCR: polymerase chain reaction.

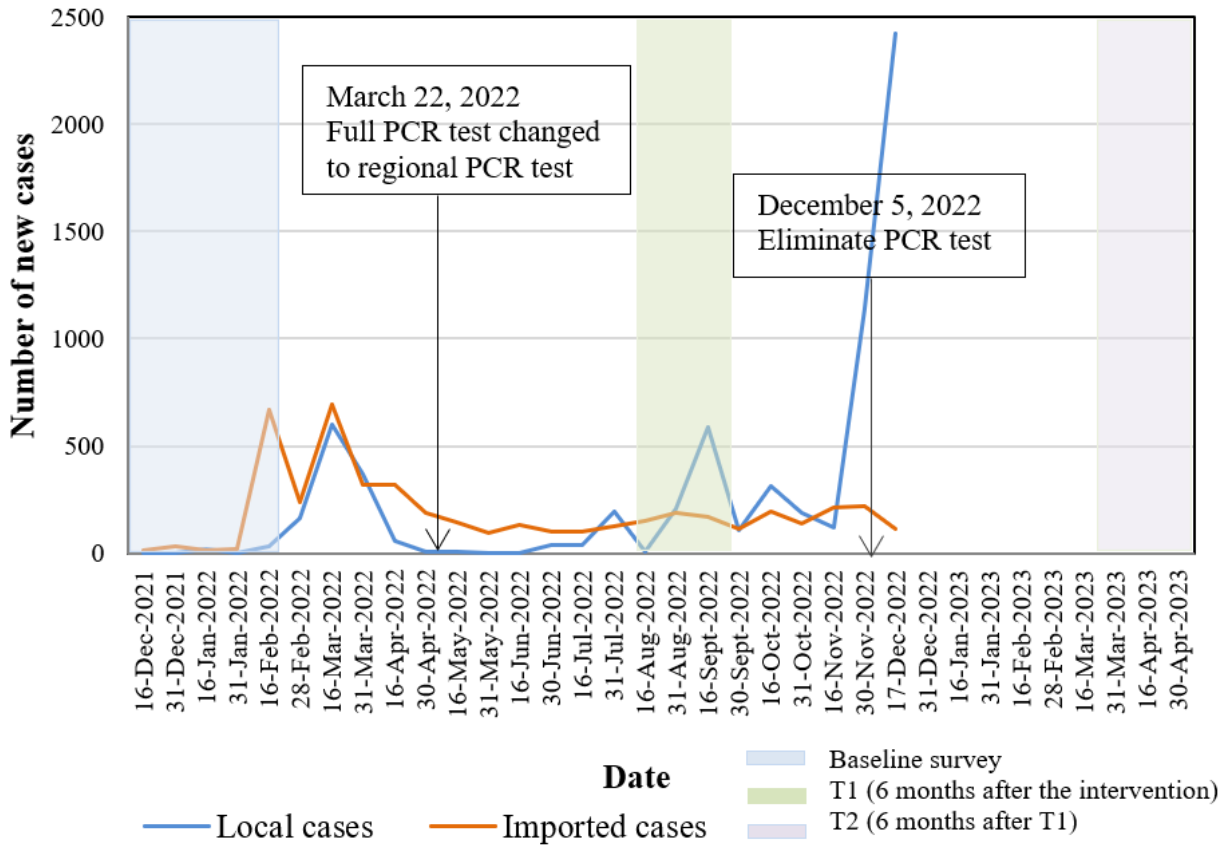

Ethical Considerations

Written informed consent was obtained from all participants. Upon completion of each survey, an e-coupon for ¥50 (US $6.50) was given to the participants as a token of appreciation. All data were stored in the Questionnaire Star (Changsha Ranxing Information Technology Company) server and protected by a password. Only the corresponding author had access to the database. Ethics approval was obtained from the ethics committee of the Longhua District Center for Disease Control and Prevention (reference 2021015).

Participants and Sample Size Planning

The inclusion criteria were (1) male sex at birth; (2) age of ≥18 years; (3) full-time employee of a factory in Shenzhen, China; (4) a smartphone with internet access; (5) having installed or willingness to install WeChat on the smartphone; and (6) willingness to complete follow-up surveys at T1 and T2. The exclusion criteria were (1) having had oral or anal sex with a man, (2) blindness or deafness, (3) inability to communicate effectively with project staff, and (4) self-report of being HIV positive. Previous studies have shown that approximately 25% of male factory workers in Shenzhen had sexual intercourse with nonregular female sex partners in the previous 6 months [7]. For planning purposes, we assumed that 25% of those in the control group would have had sexual intercourse with nonregular female sex partners or female sex workers in the previous 6 months at T1 or T2. We used a smallest detectable difference of 15% between the intervention and control groups (ie, 10% in the intervention group). Assuming the dropout rate to be 20% at T2, a total of 122 participants per group was required. The total sample size was 244 (PASS; version 11.0; NCSS LLC).

XSL•FO
RenderX

Recruitment Procedures

In a cross-sectional survey conducted between October 2019 and December 2019, the research team randomly selected 16 factories in Shenzhen from the most up-to-date registry kept by the Longhua District Center for Disease Control and Prevention (approximately 2000 factories in total) to investigate HIV-related behaviors among factory workers. These factories included 4 machinery processing plants, 3 electronic equipment manufacturers, 3 printing and dyeing plants, 2 chemical raw material factories, 1 smelter, 1 clothing factory, 1 food and beverage factory, and 1 other factory. Details of the sampling methods have been reported in published studies [7,24]. The research team obtained the consent of the owners of these 16 factories to recruit participants for this RCT. Since the COVID-19 outbreak, the Shenzhen government requested that each factory establish WeChat groups including all employees to facilitate COVID-19 control [25-27]. During the recruitment period, a coordinator posted the study information in the WeChat groups involving all employers of a participating factory. Interested workers could approach a recruitment booth set by field-workers in the factory. On-site, the field-workers briefed prospective participants about the study procedures and confirmed their eligibility. The field-workers also guaranteed anonymity and participants' right to withdraw from the study at any time, and refusals would have no consequences. Written informed consent was obtained. We developed web-based baseline and follow-up questionnaires using Questionnaire Star, a commonly used web-based survey platform in China. A link to access a web-based questionnaire was sent to participants' WeChat at baseline, T1, and T2. Each individual WeChat account was allowed to access the questionnaire once to avoid duplicate responses. The baseline survey contained 60 items (approximately 15 items per page for 4 pages) and required approximately 15 minutes to complete. The follow-up surveys at T1 and T2 had 45 items (approximately 15 items per page for 3 pages) and took approximately 10 minutes to complete. The Questionnaire Star tool performed completeness checks before the questionnaires were submitted. Participants could review and change their responses using the Back button.

Of the 458 participants who visited the recruitment booth during the recruitment period, 398 (86.9%) were eligible for participation after screening, 151 (33%) refused to participate in the study for time or other logistic reasons, and 247 (53.9%) completed the baseline survey. The response rate was 62.1% (247/398). The proportion of participants who completed the follow-up surveys was 89.5% (221/247) at T1 and 89.1% (220/247) at T2.

The Random Allocation Process

On-site at the recruitment booth, participants were randomized evenly to either the intervention group or the control group. Computer-generated random allocation codes were produced and sealed in opaque envelopes by a research staff member with no involvement in recruitment or the baseline survey. One envelope was drawn and opened in front of the participants by the field-workers. Participants were then informed by the field-workers of which group they were assigned to. Block randomization with block size of 8 was used.

Development of the Intervention Materials

Video Promoting Refusal to Engage in Sexual Behaviors With Nonregular Female Sex Partners and Female Sex Workers

The intervention mainly used an audiovisual approach, which has been used effectively in many HIV-related health promotion programs [12,13,28]. Previous studies have suggested that male factory workers might engage in hazardous sexual behaviors or seek commercial sex under peer pressure [29]. Such findings were confirmed by our previous study among male factory workers in Shenzhen [7]. Hard-to-refuse personal invitations from peers to have sex with nonregular female sex partners or female sex workers was a risk factor of such sexual risk behaviors among male factory workers [7]. Therefore, we developed a web-based video that aimed to increase male factory workers' capacity to refuse peers' invitations to have sex with nonregular female sex partners and female sex workers. To inform the contents of the video, in-depth interviews were conducted to understand male factory workers' perspective of accepting or refusing peers' invitations to have sex with nonregular female sex partners or female sex workers. In total, 5 male factory workers were recruited through purposive sampling for the in-depth interviews. Before the interviews, field-workers explained the purpose and nature of the interviews. With written informed consent, face-to-face interviews were conducted in private places in the Longhua District Center for Disease Control and Prevention and audio recorded. The interviews lasted approximately 1 hour and were transcribed. A codebook was kept to record special data and transform the data into categories to identify main themes. All informants had been invited by peers to have sex with nonregular female sex partners or female sex workers in the past, and 60% (3/5) of them had had sex with such partners. Four themes were identified as reasons to refuse such invitations: (1) perceiving a high risk of contracting HIV or other STIs through sexual behavior with nonregular female sex partners or female sex workers, (2) female individuals with HIV or STIs might not have any symptoms (look "clean"), (3) HIV or STI infection would have severe consequences on family relationships (their wives or girlfriends would break up with them, and their children or parents would be disappointed regarding their behaviors), and (4) concerns about transmitting HIV or STIs to their wives or stable girlfriends. On the basis of the qualitative findings, a panel consisting of 2 factory worker volunteers, 1 staff member of the Longhua District Center for Disease Control and Prevention, 1 behavioral health expert, and 2 health communication experts was formed to design the health communication messages in the video.

Video Promoting HIV Testing Uptake

Another video was developed to promote free HIV testing services provided by the Longhua District Center for Disease Control and Prevention. The research team conducted a survey investigating facilitators and barriers to uptake of HIV testing among male factory workers who had sexual intercourse with nonregular female sex partners or female sex workers [24]. Belief that taking up HIV testing could detect HIV infection earlier to enable better treatment outcomes was a facilitator for

XSL•FO
RenderX

these factory workers to undergo HIV testing. Concerns about being stigmatized and inconvenience were found to be barriers [24]. The aforementioned in-depth interviews of male factory workers confirmed these findings. The contents of the video addressed these facilitators and barriers. Both videos were produced by a professional team, reviewed by 5 other male factory workers, and finalized by the panel.

The Control Group

Participants scanned a QR code to run the standard version of the WeChat mini program developed by the Longhua District Center for Disease Control and Prevention. The mini program provided basic HIV-related knowledge, including (1) what are HIV and AIDS, (2) how is HIV spread, (3) how can a person reduce the risk of contracting HIV, (4) what is the treatment for HIV, and (5) what are the symptoms of HIV and AIDS. Users could also access information about free HIV testing and counseling services provided by the Longhua District Center for Disease Control and Prevention and make an appointment for HIV testing. In addition, they could contact the staff of Longhua District Center for Disease Control and Prevention for inquiry through private messaging. Figure 2 shows the standard version of the WeChat mini program in the control group.

Figure 2. Standard version of the WeChat mini program developed by the Longhua District Center for Disease Control and Prevention (control group).

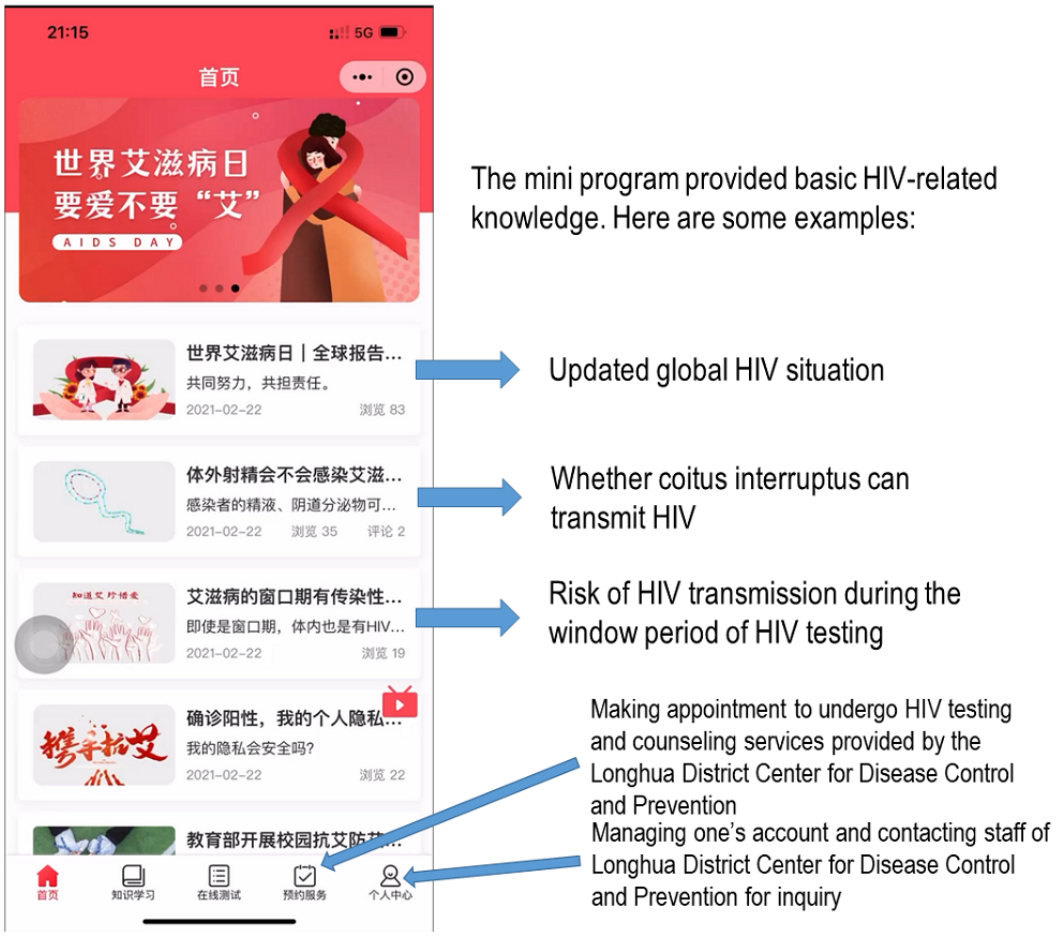

The Intervention Group

Overview

Participants scanned another QR code to run the enhanced WeChat mini program. The enhanced mini program covered all the information in the standard version of the WeChat mini program developed by the Longhua District Center for Disease Control and Prevention. In addition, the enhanced mini program assessed users' behaviors and invited users to watch different videos based on their behavioral characteristics at months 0 and 1. To avoid participants watching the same video twice, 2 slightly different versions of each video were prepared. In each session, the mini program randomly selected 1 version for the participants. Figure 3 shows the demonstration of the enhanced version of the WeChat mini program in the intervention group.

XSL•FO
RenderX

Figure 3. Enhanced version of the WeChat mini program developed by the Longhua District Center for Disease Control and Prevention (intervention group).

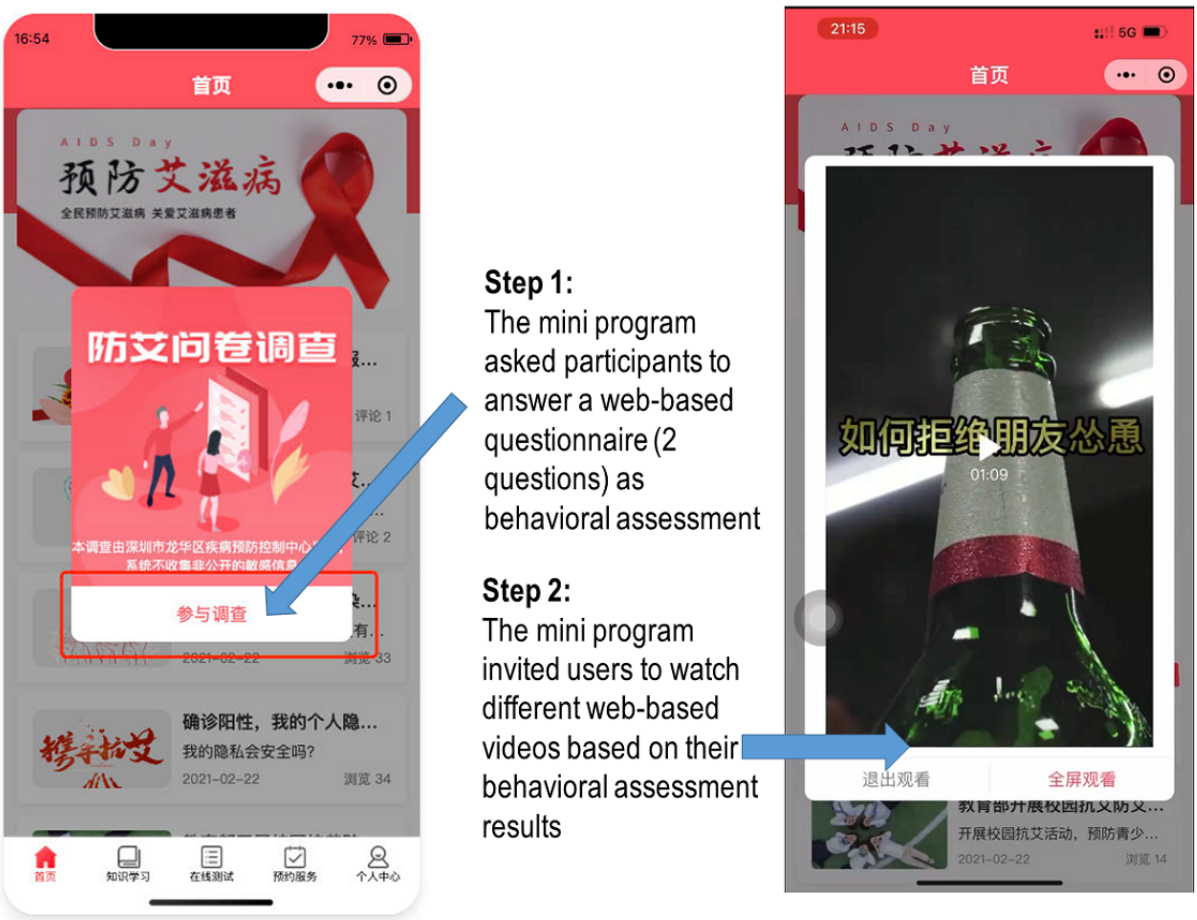

Behavioral Assessment

The mini program asked participants to answer two questions: (1) whether they had had sex with nonregular female sex partners or female sex workers in the previous month and (2) whether they intended to have sex with nonregular female sex partners or female sex workers in the following month. Participants chose their response on the screen (yes or no). If the participant answered "yes" to at least one question, he was classified as "higher risk" by the mini program. If the participant answered "no" to both questions, he was considered as "lower risk." Answering these questions was mandatory.

Video for People at Lower Risk

A video popped up on the screen to promote refusal to peers' invitations to engage in sexual behaviors with nonregular female sex partners and female sex workers. In the video, a male factory worker was trying to persuade his friend not to have sex with nonregular female sex partners and female sex workers. He emphasized that (1) female individuals with HIV might not have any symptoms, (2) there is a high risk of HIV acquisition through sexual behavior with nonregular female sex partners or female sex workers, (3) there are severe consequences of HIV infection for family relationships (breaking up with his wife, and his children or parents being disappointed regarding his behaviors), and (4) there was a high risk of transmitting HIV to his wife.

Video for People at Higher Risk

In addition to the video watched by people at lower risk, another video popped up on the screen to promote HIV testing. The contents included (1) a male factory worker introducing the benefits of taking up HIV testing, such as that it would help detect HIV infection earlier and reduce psychological burden; and (2) the same factory worker demonstrating the procedures of free HIV testing and counseling at the Longhua District Center for Disease Control and Prevention, which portrayed caring, supportive, and nonjudgmental administrators regardless of users' test results.

Fidelity Assessments

To ensure a complete exposure, the videos were formatted in such a way that the participants could not fast-forward or skip any part of them. The mini program automatically documented the number of sessions completed by participants in the intervention group.

Measures to Avoid Contamination

The videos in the enhanced version of the WeChat mini program could not be downloaded or forwarded to other people. Participants in the intervention group were asked not to

XSL•FO
RenderX

disseminate the QR code to access the enhanced mini program to other people. The research team also cross-checked the user IDs of the standard and the enhanced mini programs during the project period.

Outcome Measurements

Sexual intercourse with nonregular female sex partners and female sex workers in the previous 6 months were the primary outcomes. In total, 2 questions developed by published studies measured these behaviors at baseline, T1, and T2 (ie, "Did you have sexual intercourse with a nonregular female sex partner in the past 6 months?" and "Did you have sexual intercourse with a female sex worker in the past 6 months?"). Similarly to the published studies, a female sex worker was defined as a woman who explicitly exchanged sex for money or gifts, and a nonregular female sex partner was defined as a woman who was neither a female sex worker nor their regular female sex partner (stable girlfriend or wife) [7,24].

Secondary outcomes measured at baseline, T1, and T2 included (1) condomless sex with a nonregular female sex partner among participants who had sexual intercourse with a nonregular female sex partner in the previous 6 months, (2) condomless sex with a female sex worker among participants who had sexual intercourse with a female sex worker in the previous 6 months, and (3) uptake of any type of HIV testing among all participants in the previous 6 months. These secondary outcomes were also measured using validated questions [7,24].

Background characteristics were collected through a baseline survey, including sociodemographics and use of HIV or STI prevention services other than HIV testing.

Process Evaluation

At T1, participants in the intervention group were asked whether the contents of the videos were interesting, relevant to them, informative, helpful, valuable, persuasive, and credible. They were also asked whether they were willing to share these videos with their peers. The response categories of the process evaluation questions were 1=*strongly disagree*, 2=*disagree*, 3=*neutral*, 4=*agree*, and 5=*strongly agree*. In addition, COVID-19 vaccination status and history of confirmed COVID-19 infection were measured at both T1 and T2.

Statistical Analysis

Between-group differences in baseline characteristics were compared using chi-square tests. The baseline characteristics of participants who completed the survey at T1 or T2 and those who were lost to follow-up were also compared using chi-square tests. Intention-to-treat analysis was used for the outcome analyses. Missing data were all binary variables. As there were no missing data among participants who completed the T1 and T2 surveys, the missing rate was equal to the dropout rate. We filled the missing outcome values at T1 or T2 using multiple imputations. The Markov chain Monte Carlo method was used to impute data with an arbitrary pattern of missing values, whereas monotone methods were used to impute data with a monotone pattern of missing values. Predictors included baseline background characteristics and baseline values of these outcomes. The relative risk, absolute risk reduction, and number needed to treat and their 95% CIs were calculated using Microsoft Excel (Microsoft Corp). Logistic regression models were used to test the between-group difference in the imputed primary and secondary outcomes after controlling for baseline background characteristics, with $P<.05$ for between-group comparisons. Crude odds ratios, adjusted odds ratios, and their 95% CIs were obtained. As relationship status might affect the onset of sexual risk behaviors, subgroup analyses of the primary and secondary outcomes were conducted among participants with and without a stable girlfriend or spouse at baseline. Within-group changes in the imputed primary and secondary outcomes (month 6 vs baseline) were investigated using McNemar tests. SPSS (version 26.0; IBM Corp) was used for data analysis, and $P<.05$ was considered statistically significant.

Results

Descriptive Statistics

A nonblinded 2-arm parallel RCT was conducted between December 2021 and April 2023. The trial was registered at ClinicalTrials.gov (NCT05811611). The CONSORT (Consolidated Standards of Reporting Trials) flowchart is shown in Figure 4. At baseline, over half of the participants were aged ≤40 years (204/247, 82.6%), were internal migrants (226/247, 91.5%), were married to a woman or had a stable girlfriend (173/247, 70%), had attained tertiary education (135/247, 54.7%), had a monthly income of ¥5000-9999 (US $782.9-$1564.8; 151/247, 61.1%), lived with their partner or children in Shenzhen (139/247, 56.3%), and were frontline workers (167/247, 67.6%). In the previous 6 months, 36.4% (90/247) had had sexual intercourse with a nonregular female sex partner, and 62% (56/90) of those with a nonregular female sex partner reported condomless sex with such a partner. Of 86 participants (86/247, 34.8% of all participants) who had sexual intercourse with a female sex worker, 64 (74%) reported condomless sex with such a partner. Among the participants, 4.9% (12/247) had used HIV testing, and 19.8% (49/247) had used other HIV or STI prevention services in the previous 6 months. No significant between-group differences in these baseline characteristics were found (*P* values between .07 and .97; Table 1).

The dropout rate in the intervention group and the control group was 6.4% (8/125) and 14.8% (18/122) at T1 and 7.2% (9/125) and 14.8% (18/122) at T2, respectively. In the intervention group, participants with a lower educational level and those who were internal migrants in the control group were more likely to drop out at both T1 and T2 (Multimedia Appendices 1 and 2).

XSL•FO
RenderX

Figure 4. The CONSORT (Consolidated Standards of Reporting Trials) flowchart of the randomized controlled trial.

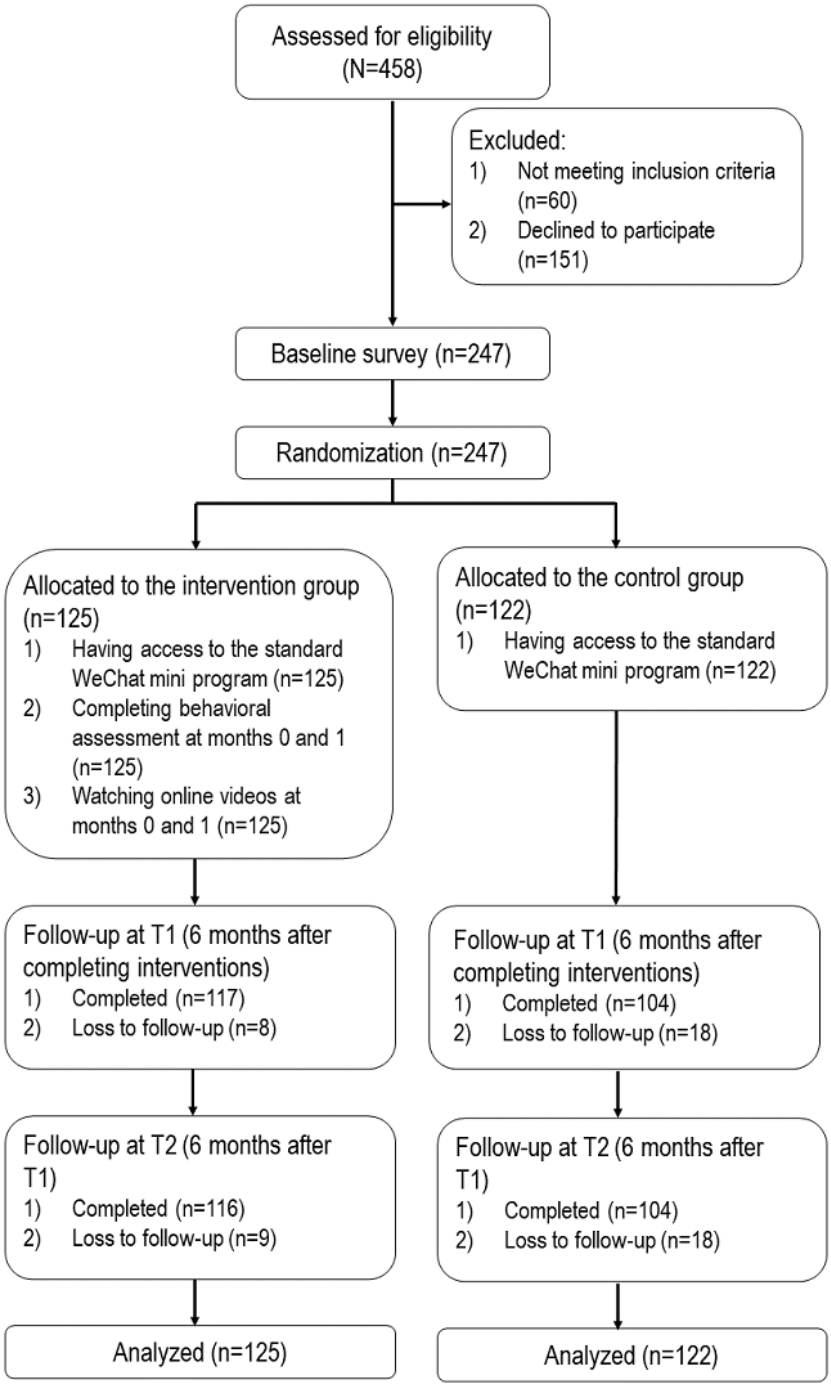

XSL•FO
RenderX

Table 1. Characteristics of the participants at baseline (N=247).

Characteristic	All, n (%)	Intervention group (n=125), n (%)	Control group (n=122), n (%)	*P* value
Sociodemographic characteristics				
Age group (y)				.22
18-30	92 (37.2)	47 (37.6)	45 (36.9)	
31-40	112 (45.3)	62 (49.6)	50 (41)	
41-50	33 (13.4)	13 (10.4)	20 (16.4)	
>50	10 (4)	3 (2.4)	7 (5.7)	
Ethnicity				.57
Han	231 (93.5)	118 (94.4)	113 (92.6)	
Minority	16 (6.5)	7 (5.6)	9 (7.4)	
Registered permanent resident of Shenzhen				.12
No	226 (91.5)	111 (88.8)	115 (94.3)	
Yes	21 (8.5)	14 (11.2)	7 (5.7)	
Relationship status				.32
Currently single	74 (30)	41 (32.8)	33 (27)	
Stable girlfriend or married to a woman	173 (70)	84 (67.2)	89 (73)	
Educational level				.72
Junior high or lower	50 (20.2)	23 (18.4)	27 (22.1)	
Senior high or equivalent	62 (25.1)	31 (24.8)	31 (25.4)	
College and higher	135 (54.7)	71 (56.8)	64 (52.5)	
Living with partner, spouse, or children in Shenzhen				.73
No	108 (43.7)	56 (44.8)	52 (42.6)	
Yes	139 (56.3)	69 (55.2)	70 (57.4)	
Monthly personal income				.07
<¥3000 (US $469.5)	12 (4.9)	9 (7.2)	3 (2.5)	
¥3000-¥4999 (US $469.5-$782.3)	69 (27.9)	31 (24.8)	38 (31.1)	
¥5000-¥9999 (US $782.9-$1564.8)	151 (61.1)	74 (59.2)	77 (63.1)	
≥¥10,000 (US $1564.9)	15 (6.1)	11 (8.8)	4 (3.3)	
Status as frontline worker or management staff				.49
Frontline worker	167 (67.6)	82 (65.6)	85 (69.7)	
Management staff	80 (32.4)	43 (34.4)	37 (30.3)	
HIV or STI[a] prevention service use in the previous 6 months				
Use of HIV testing				.97
No	235 (95.1)	119 (95.2)	116 (95.1)	
Yes	12 (4.9)	6 (4.8)	6 (4.9)	
Use of other HIV or STI prevention services (receiving free condoms or pamphlets or attending workshops or seminars)				.10
No	198 (80.2)	95 (76)	103 (84.4)	
Yes	49 (19.8)	30 (24)	19 (15.6)	
Sexual behaviors in the previous 6 months				
Sexual intercourse with nonregular female sex partners				.70
No	157 (63.6)	78 (62.4)	79 (64.8)	
Yes	90 (36.4)	47 (37.6)	43 (35.2)	

XSL•FO
RenderX

Characteristic	All, n (%)	Intervention group (n=125), n (%)	Control group (n=122), n (%)	*P* value
Sexual intercourse with female sex workers				.23
No	161 (65.2)	77 (61.6)	84 (68.9)	
Yes	86 (34.8)	48 (38.4)	38 (31.1)	
Condomless sex with nonregular female sex partners (among participants who had sexual intercourse with nonregular female sex partners in the previous 6 months)				.74
No	34 (37.8)[b]	17 (36.2)[c]	17 (39.5)[d]	
Yes	56 (62.2)[b]	30 (63.8)[c]	26 (60.5)[d]	
Condomless sex with female sex workers (among participants who had sexual intercourse with female sex workers in the previous 6 months)				.39
No	22 (25.6)[e]	14 (29.2)[f]	8 (21.1)[g]	
Yes	64 (74.4)[e]	34 (70.8)[f]	30 (78.9)[g]	

[a]STI: sexually transmitted infection.
[b]n=90.
[c]n=47.
[d]n=43.
[e]n=86.
[f]n=48.
[g]n=38.

Primary Outcomes

At T1, fewer participants in the intervention group reported sexual intercourse with a nonregular female sex partner in the previous 6 months as compared with those in the control group (1/125, 0.8% vs 8/122, 6.6%; relative risk=0.12, 95% CI 0.02-0.96; absolute risk reduction=–0.06, 95% CI –0.10 to –0.01; number needed to treat=–17.4, 95% CI –91.3 to –9.6; *P*=.02). However, there were no between-group differences in sexual intercourse with a nonregular female sex partner at T2 (10/125, 8% vs 14/122, 11.5%; *P*=.36) or sexual intercourse with a female sex worker at T1 (2/125, 1.6% vs 2/122, 1.6%; *P*=.98) or T2 (8/125, 6.4% vs 8/122, 6.6%; *P*=.96). There was no between-group difference in these primary outcomes among the subgroups of participants with and without a stable girlfriend or spouse at baseline (Table 2).

When comparing data measured at T1 or T2 with the baseline data, significant decreases in sexual intercourse with a nonregular female sex partner and with a female sex worker were observed in both the intervention group and the control group (*P*<.001 in all comparisons; Table 3). As compared with T1, significant increases in sexual intercourse with a nonregular female sex partner (*P*=.01) and with a female sex worker (*P*=.03) were observed in the intervention group at T2. A significant increase in sexual intercourse with a female sex worker was observed when comparing T2 versus T1 among participants in the control group (*P*=.03; Table 3).

XSL•FO
RenderX

Table 2. Between-group differences in the primary and secondary outcomes.

	Intervention group, n (%)	Control group, n (%)	RR[a] (95% CI)	ARR[b] (95% CI)	NNT[c] (95% CI)	*P* value
Primary outcomes						
Sexual intercourse with nonregular female sex partners						
All participants						
T1[d]	1 (0.8)[e]	8 (6.6)[f]	0.12 (0.02 to 0.96)	–0.06 (–0.10 to –0.01)	–17.4 (–91.3 to –9.6)	.02
T2[g]	10 (8)[e]	14 (11.5)[f]	0.70 (0.32 to 1.51)	–0.03 (–0.11 to 0.04)	–28.8 (–9.2 to 25.5)	.36
Subgroup of participants without a stable girlfriend or spouse at baseline						
T1	0 (0)[h]	2 (6.1)[i]	—[j]	–0.06 (–0.14 to 0.02)	–16.5 (–7.0 to 48.1)	.20
T2	5 (12.2)[h]	6 (18.2)[i]	0.68 (0.22 to 2.00)	–0.06 (–0.22 to 0.10)	–16.7 (–4.43 to 9.48)	.53
Subgroup of participants with a stable girlfriend or spouse at baseline						
T1	1 (1.2)[k]	6 (6.7)[l]	0.18 (0.02 to 1.44)	–0.06 (–0.11 to 0.001)	–18.1 (–8.9 to 661.3)	.07
T2	5 (6)[k]	8 (9)[l]	0.59 (0.21 to 1.69)	–0.03 (–0.11 to 0.05)	–32.9 (–9.22 to 20.97)	.54
Sexual intercourse with female sex workers						
All participants						
T1	2 (1.6)[e]	2 (1.6)[f]	0.98 (0.14 to 6.82)	0.0 (–0.03 to 0.03)	2541.7 (–31.4 to 32.2)	.98
T2	8 (6.4)[e]	8 (6.6)[f]	0.98 (0.38 to 2.51)	–0.02 (–0.06 to 0.06)	635.4 (–15.9 to 16.7)	.96
Subgroup of participants without a stable girlfriend or spouse at baseline						
T1	0 (0)[h]	1 (3)[i]	—	–0.03 (–0.09 to 0.03)	–33.0 (–11.26 to 35.48)	.45
T2	4 (9.8)[h]	4 (12.1)[i]	0.80 (0.22 to 2.98)	–0.02 (–0.17 to 0.12)	–42.28 (–5.98 to 8.33)	>.99
Subgroup of participants with a stable girlfriend or spouse at baseline						
T1	2 (2.4)[k]	1 (1.1)[l]	2.11 (0.20 to 22.93)	0.01 (–0.03 to 0.05)	79.53 (–37.45 to 19.28)	.48
T2	4 (4.8)[k]	4 (4.5)[l]	1.06 (0.27 to 4.10)	0.002 (–0.06 to 0.06)	373.8 (–16.67 to 15.30)	>.99
Secondary outcomes						
Condomless sex with nonregular female sex partners (among participants who had sexual intercourse with a nonregular female sex partner at each time point)						
All participants						
T1	0 (0)[m]	5 (62.5)[n]	—	0.63 (–0.96 to –0.29)	–1.6 (–3.5 to –1.0)	.24
T2	6 (60)[o]	3 (21.4)[p]	2.80 (0.91 to 8.61)	0.39 (0.01 to 0.76)	2.59 (1.32 to 72.99)	.05
Subgroup of participants without a stable girlfriend or spouse at baseline						
T1	0 (0)[q]	1 (50)[r]	—	—	—	—
T2	2 (40)[s]	1 (16.7)[t]	2.40 (0.30 to 19.33)	0.23 (–0.29 to 0.76)	4.29 (–3.55 to 1.32)	.39
Subgroup of participants with a stable girlfriend or spouse at baseline						
T1	0 (0)[m]	4 (66.7)[t]	—	–0.67 (–1.04 to –0.29)	–1.5 (–0.35 to –0.96)	.21
T2	4 (80)[s]	2 (25)[n]	3.20 (0.89 to 11.48)	0.55 (0.09 to 1.01)	1.82 (0.99 to 11.30)	.05
Condomless sex with female sex workers (among participants who had sexual intercourse with a female sex worker at each time point)						
All participants						

XSL•FO
RenderX

	Intervention group, n (%)	Control group, n (%)	RR[a] (95% CI)	ARR[b] (95% CI)	NNT[c] (95% CI)	*P* value
T1	0 (0)[r]	0 (0)[r]	—	—	—	—
T2	0 (0)[n]	0 (0)[n]	—	—	—	—
Subgroup of participants without a stable girlfriend or spouse at baseline						
T1	0 (0)[q]	0 (0)[m]	—	—	—	—
T2	0 (0)[u]	0 (0)[u]	—	—	—	—
Subgroup of participants with a stable girlfriend or spouse at baseline						
T1	0 (0)[r]	0 (0)[m]	—	—	—	—
T2	0 (0)[u]	0 (0)[u]	—	—	—	—
Uptake of any type of HIV testing						
All participants						
T1	10 (8)[e]	10 (8.2)[f]	0.98 (0.42 to 2.26)	–0.002 (–0.07 to 0.07)	–508.3 (–14.3 to 15.1)	.57
T2	10 (8)[e]	11 (9)[f]	0.89 (0.39 to 2.01)	–0.01 (–0.08 to 0.05)	–98.4 (–12.5 to 16.8)	.47
Subgroup of participants without a stable girlfriend or spouse at baseline						
T1	3 (7.3)[h]	2 (6.1)[i]	1.21 (0.21 to 6.81)	0.01 (–0.10 to 0.13)	79.59 (–9.86 to 7.91)	.83
T2	2 (4.9)[h]	2 (6.1)[i]	0.80 (0.12 to 5.41)	–0.01 (–0.12 to 0.09)	–84.56 (–8.58 to 10.76)	.82
Subgroup of participants with a stable girlfriend or spouse at baseline						
T1	7 (8.3)[k]	8 (9)[l]	0.82 (0.32 to 2.11)	–0.01 (–0.09 to 0.08)	–152.57 (–11.07 to 12.94)	.88
T2	8 (9.5)[k]	9 (10.1)[l]	0.94 (0.38 to 2.33)	–0.01 (–0.09 to 0.08)	–169.91 (–10.57 to 12.08)	.69

[a]RR: relative risk.
[b]ARR: absolute risk reduction.
[c]NNT: number needed to treat.
[d]T1: 6 months after completion of the interventions.
[e]n=125.
[f]n=122.
[g]T2: 6 months after T1.
[h]n=41.
[i]n=33.
[j]Not applicable.
[k]n=84.
[l]n=89.
[m]n=1.
[n]n=8.
[o]n=10.
[p]n=14.
[q]n=0.
[r]n=2.
[s]n=5.
[t]n=6.
[u]n=4.

XSL•FO
RenderX

Table 3. Within-group changes in primary and secondary outcomes.

	Intervention group (n=125)	Control group (n=122)
Primary outcomes		
Sexual intercourse with nonregular female sex partners		
Baseline, n (%)	47 (37.6)	43 (35.2)
T1[a], n (%)	1 (0.8)	8 (6.6)
T2[b], n (%)	10 (8)	14 (11.5)
T1 vs baseline, *P* value	<.001	<.001
T2 vs baseline, *P* value	<.001	<.001
T2 vs T1, *P* value	.01	.42
Sexual intercourse with female sex workers		
Baseline, n (%)	48 (38.4)	38 (31.1)
T1, n (%)	2 (1.6)	2 (1.6)
T2, n (%)	8 (6.4)	8 (6.6)
T1 vs baseline, *P* value	<.001	<.001
T2 vs baseline, *P* value	<.001	<.001
T2 vs T1, *P* value	.03	.03
Secondary outcome		
Uptake of any type of HIV testing		
Baseline, n (%)	6 (4.8)	6 (4.9)
T1, n (%)	10 (8)	10 (8.2)
T2, n (%)	10 (8)	11 (9)
T1 vs baseline, *P* value	.33	.51
T2 vs baseline, *P* value	.33	.34
T2 vs T1, *P* value	>.99	.73

[a]T1: 6 months after completion of the interventions.

[b]T2: 6 months after T1.

Secondary Outcomes

There was no significant between-group difference in condomless sex with a nonregular female sex partner at T1 or T2. No participants in the intervention group or the control group reported condomless sex with a female sex worker at T1 or T2. Participants in the intervention group and the control group reported similar HIV testing uptake in the previous 6 months at T1 (10/125, 8% vs 10/122, 8.2%; *P*=.57) and T2 (10/125, 8% vs 11/122, 9%; *P*=.47). There was no between-group difference in these secondary outcomes between the subgroups of participants with and without a stable girlfriend or spouse at baseline (Table 2). There were no within-group changes in HIV testing uptake in the intervention group or the control group (Table 3).

Process Evaluation

As documented by the WeChat mini program, all participants in the intervention group (125/125, 100%) completed a behavioral assessment and watched the videos at months 0 and 1. No participants in the control group accessed the enhanced mini program during the project period. Among the 117 participants in the intervention group who completed the process evaluation at T1, approximately half of them strongly agreed or agreed that contents of the videos were interesting (n=54, 46.2%), relevant to them (n=53, 45.3%), informative (n=53, 45.3%), helpful (n=57, 48.7%), valuable (n=57, 48.7%), persuasive (n=56, 47.9%), and credible (n=56, 47.9%). Moreover, 53% (62/117) of them were willing to share these videos with their peers in the future. At T1, all participants in the intervention group (117/117, 100%) and the control group (104/104, 100%) who completed the follow-up survey received 2 doses of the COVID-19 vaccination. There was no significant between-group difference in the uptake rate of the third dose of the COVID-19 vaccination (110/117, 94% in the intervention group vs 95/104, 91.3% in the control group; *P*=.44). At T2, all participants who completed the follow-up survey received 3 doses of the COVID-19 vaccination (116/116, 100% in the intervention group vs 104/104, 100% in the control group; *P*>.99). No participant reported a history of COVID-19 infection at T1. At T2, a total of 65.5% (76/116) of the participants in the intervention group and 71.2% (74/104) of the participants in the control group reported a history of COVID-19 infection (*P*=.37 for between-group comparison).

XSL•FO
RenderX

Discussion

Value and Strengths of This Study

Our study contributed to the literature by comparing the efficacy of an enhanced version versus the standard version of a WeChat mini program using an RCT design and developing a new method to reduce nonmarital heterosexual contacts among male factory workers. Our study also had the strengths of intervention materials developed based on in-depth interviews, a long follow-up duration, and a low dropout rate and was based on the findings of formative studies [7,24].

Findings of and Elaborations on the Primary Outcomes

As compared with the baseline data, significant decreases in sexual intercourse with nonregular female sex partners and with female sex workers were observed in both the intervention group and the control group at T1 and T2. Our results showed that adding health promotion videos to the standard WeChat mini program increased its efficacy in reducing sexual intercourse with nonregular female sex partners at T1 (1/125, 0.8% vs 8/122, 6.6%). However, there is a lack of long-term efficacy as no between-group differences in sexual intercourse with nonregular female sex partners or female sex workers were observed at T2. The lack of additional interventions (eg, videos, reminders, and notifications) might contribute to the absence of long-term effects on sexual risk behaviors. There are some potential strategies to improve our interventions. First, it is beneficial to have more intervention sessions and a longer intervention period. Our mini program only disseminated health promotion twice 1 month apart. A systematic review and meta-analysis showed that multiple-session interventions were more effective than single-session interventions in improving consistent condom use [14]. An intervention duration of 3 to 6 months is commonly used in successful HIV risk reduction interventions and WeChat mini programs for other health-related behaviors [30-32]. Therefore, future studies may evaluate the effectiveness of larger doses of our interventions (eg, 1 session per month for 3-6 months) in reducing sexual risk behaviors in the longer term. Second, some strategies used by successful WeChat mini programs for other types of health promotion have shed some light regarding improving our interventions. These programs made use of behavior change theories (eg, the Behavior Change Wheel or the Health Action Process Approach theory) and were equipped with real-time question-and-answer functions [31,32]. Interventions based on behavior change theories are more likely to be effective than non–theory-based ones [33]. Providing real-time answers to participants' individual questions could improve engagement and, hence, improve the effectiveness of digital health interventions [34]. In addition, the process evaluation showed that only half of the participants were satisfied with the additional health promotion videos in the enhanced WeChat mini program. Therefore, it is necessary to improve the contents of these health promotion materials. Similarly to most HIV interventions for factory workers [6], our interventions were developed using a top-down approach in which end users' involvement was limited. Making use of cocreation, which refers to collaborative public health intervention development by academics together with end users [35-37], may be useful in improving the health promotion materials. Such an approach is considered to be promising and efficient for addressing complex health issues and facilitating behavior change [35-37].

Findings of and Elaborations on the Secondary Outcome

Regarding the secondary outcomes, no between-group difference in HIV testing was observed at T1 or T2. The HIV testing rate in both groups at T1 and T2 was similar to that in the baseline data. As the prevalence of sexual intercourse with nonregular female sex partners or female sex workers at T1 and T2 was much lower than that in the baseline data, participants might perceive that they were at low risk of HIV infection and, hence, had no need to undergo HIV testing.

Impact of the COVID-19 Pandemic

The changes in the COVID-19 situation and its control measures had impacts on our study outcomes. At baseline (December 2021 to February 2022), the COVID-19 situation was stable and under control in Shenzhen. Starting from April 2022, the city was continuously hit by a COVID-19 outbreak caused by Omicron variants BA2 or BA5. In response to the outbreak, the city implemented strict control measures, including regular universal COVID-19 testing, territory lockdown, and closure of entertainment premises [38]. The implementation of these control measures increased difficulties for male factory workers to seek nonregular female sex partners or female sex workers. When China started to relieve its COVID-19 control measures in December 2022, the number of new cases increased dramatically. It was estimated that >70% of all people in China were infected with SARS-CoV-2 during the huge wave between December 2022 and February 2023 [39]. It was likely that most of our participants were infected with SARS-CoV-2 during this period. The discomfort and symptoms of COVID-19 might reduce their motivation to have sex with nonregular female sex partners or female sex workers. The aforementioned situation could partially explain why the magnitude of decreases in sexual risk behaviors in our study (T1 vs baseline: 28.6%-36.8%; T2 vs baseline: 24.5%-36.8%) was much larger than those of previous interventions (posttest vs pretest: 3.1%-14%) targeting factory workers [40-44]. Moreover, facility-based HIV testing was seriously affected by the COVID-19 outbreak in China [45-48]. Some governmental HIV testing facilities were suspended during the pandemic as the government reallocated resources to focus on COVID-19 control. The decreased access to facility-based HIV testing might also contribute to the low HIV testing uptake observed in our study.

Limitations of the Study

Our study also had other limitations. First, although we measured COVID-19 vaccination status and history of confirmed COVID-19 infection at both T1 and T2, we did not ask about the direct impact of the COVID-19 situation on participants' sexual behaviors. The strict COVID-19 measures implemented in the study site applied to everyone. As this study was an RCT, the impact of the COVID-19 situation on the study outcomes was expected to be similar between the intervention and control groups. Second, we did not have a control group without access to the WeChat mini program. Therefore, it was unclear whether

XSL•FO
RenderX

the standard or enhanced WeChat mini program would be more effective than no intervention in reducing sexual risk behaviors among factory workers. Third, similarly to most interventional studies targeting factory workers, our participants were recruited through convenience sampling. Information from factory workers who refused to join this study was unavailable. As compared with a larger and more representative sample of male factory workers in the same city [7], our participants were older (155/247, 62.9% vs 49.1% aged >30 years) and more likely to have a stable girlfriend or be married to a woman (173/247, 70% vs 40.9%), have sexual intercourse with a nonregular female sex partner (90/247, 36.4% vs 24.5%) and with a female sex worker (86/247, 34.8% vs 21.3%), and report condomless sex with a nonregular female sex partner (56/90, 62% vs 21.3%) or with a female sex worker (64/86, 74% vs 13.5%). Therefore, selection bias existed. Our results were more applicable to male factory workers who were older, had a stable partner, and engaged in sexual risk behaviors. Future studies should explore the effectiveness of our interventions in a more representative sample of male factory workers. Fourth, only male factory workers in 1 Chinese city were included in this study. Generalization should be made cautiously to other Chinese cities. Fifth, self-reported responses might cause report bias despite the anonymous nature of the study. Participants might underreport sexual risk behaviors and overreport HIV testing uptake due to social desirability. Verification of these self-reported responses was impossible. Moreover, a long-distance relationship (eg, a stable partner who lives in another city) may have a different impact on the study outcomes as compared with a short-distance relationship (eg, living with a stable partner in the same city). In the baseline survey, we did not measure whether participants who had a stable girlfriend or were married to a woman were living with either their girlfriend or their spouse in Shenzhen. It is possible that some of them were living with their children but not with their girlfriend or spouse. Furthermore, attrition bias might exist. Participants with a lower education level in the intervention group and those who were internal migrants in the control group were more likely to drop out. As educational level and being an internal migrant were not significant factors of sexual intercourse with nonregular female sex partners or female sex workers [7], the impact of attrition on the study outcomes might be limited.

Conclusions

In summary, the RCT findings showed that the enhanced version of the WeChat mini program with web-based videos addressing risk factors of sexual intercourse with nonregular female sex partners and female sex workers and promoting HIV testing was more effective than the standard version in reducing sexual intercourse with nonregular female sex partners in the short term but not in the longer term. Improvements should be made to the WeChat mini program before implementation.

Acknowledgments

This study was funded by the High Level Project of Medicine in Longhua, Shenzhen (HLPM2019090201015), and the Key Discipline of Infectious Diseases Control and Prevention of Long Hua (grant 2020-2014). The authors did not use generative artificial intelligence in any portion of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of baseline characteristics between participants who completed the follow-up evaluation at T1 (6 months after completion of the intervention) and those who were lost to follow-up.
[DOCX File , 26 KB - jmir_v26i1e49362_app1.docx]

Multimedia Appendix 2
Comparison of baseline characteristics between participants who completed the follow-up evaluation at T2 (6 months after T1; 6 months after completion of the intervention) and those who were lost to follow-up.
[DOCX File , 26 KB - jmir_v26i1e49362_app2.docx]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1182 KB - jmir_v26i1e49362_app3.pdf]

References

1. HIV and AIDS. World Health Organization. 2024 Jul 22. URL: https://www.who.int/news-room/fact-sheets/detail/hiv-aids [accessed 2024-08-27]
2. Sexually transmitted infections (STIs). World Health Organization. 2024 May 21. URL: https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis) [accessed 2024-08-27]
3. China statistical yearbook. National Bureau of Statistics of China. 2020. URL: https://www.stats.gov.cn/sj/ndsj/2020/indexeh.htm [accessed 2022-07-20]

XSL•FO
RenderX

4. Dong Z, Pan X, Cai C, Qin Q, Gao GF, Lyu F. Spatial analysis of people living with HIV/AIDS transmitted through commercial heterosexual contact or non-marital non-commercial heterosexual contact - China, 2018. China CDC Wkly 2021 Apr 09;3(15):316-319 [FREE Full text] [doi: 10.46234/ccdcw2021.086] [Medline: 34594874]
5. Dou Z, Luo Y, Zhao Y, Zheng X, Han M. Trends in mortality and prevalence of reported HIV/AIDS cases - China, 2002-2021. China CDC Wkly 2023 Oct 20;5(42):943-947 [FREE Full text] [doi: 10.46234/ccdcw2023.177] [Medline: 38026097]
6. Chen D, Luo G, Meng X, Wang Z, Cao B, Yuan T, et al. Efficacy of HIV interventions among factory workers in low- and middle-income countries: a systematic review. BMC Public Health 2020 Aug 28;20(1):1310 [FREE Full text] [doi: 10.1186/s12889-020-09333-w] [Medline: 32859178]
7. Zhang K, Chen S, Zhu S, Fang Y, Zou H, Cai Y, et al. Multifaceted determinants of sexual intercourse with non-regular female sex partners and female sex workers among male factory workers in China-a cross-sectional survey. Int J Environ Res Public Health 2022 Nov 30;19(23):16008 [FREE Full text] [doi: 10.3390/ijerph192316008] [Medline: 36498089]
8. Tan W, Liu S, Zhao J. Modes of HIV transmission through heterosexual contact among HIV/AIDS cases in Shenzhen city. Practical Preventive Medicine 2019;26:1411 [FREE Full text]
9. Lu H, Liu Y, Dahiya K, Qian HZ, Fan W, Zhang L, et al. Effectiveness of HIV risk reduction interventions among men who have sex with men in China: a systematic review and meta-analysis. PLoS One 2013 Aug 30;8(8):e72747 [FREE Full text] [doi: 10.1371/journal.pone.0072747] [Medline: 24137497]
10. Wang H, Wang Z, Huang X, Chen Y, Wang H, Cui S, et al. Association of HIV preexposure prophylaxis use with HIV incidence among men who have sex with men in China: a nonrandomized controlled trial. JAMA Netw Open 2022 Feb 01;5(2):e2148782 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.48782] [Medline: 35171258]
11. Liu Z, Bao R, Zhang X, Wang H, Wang Z, Zhang J, et al. A brief smartphone-based intervention significantly improved pre-exposure prophylaxis adherence among Chinese men sex with men: findings of a randomized controlled trial. J Infect 2022 Feb;84(2):248-288. [doi: 10.1016/j.jinf.2021.12.045] [Medline: 34990708]
12. Wang Z, Lau JT, Ip TK, Yu Y, Fong F, Fang Y, et al. Two web-based and theory-based interventions with and without brief motivational interviewing in the promotion of human papillomavirus vaccination among Chinese men who have sex with men: randomized controlled trial. J Med Internet Res 2021 Feb 02;23(2):e21465 [FREE Full text] [doi: 10.2196/21465] [Medline: 33528372]
13. Wang Z, Lau JT, Ip M, Ho SP, Mo PK, Latkin C, et al. A randomized controlled trial evaluating efficacy of promoting a home-based HIV self-testing with online counseling on increasing HIV testing among men who have sex with men. AIDS Behav 2018 Jan 22;22(1):190-201. [doi: 10.1007/s10461-017-1887-2] [Medline: 28831616]
14. Chow EP, Tung K, Tucker JD, Muessig KE, Su S, Zhang X, et al. Behavioral interventions improve condom use and HIV testing uptake among female sex workers in China: a systematic review and meta-analysis. AIDS Patient Care STDS 2015 Aug;29(8):454-460. [doi: 10.1089/apc.2015.0043] [Medline: 26217931]
15. Qian X, Smith H, Huang W, Zhang J, Huang Y, Garner P. Promoting contraceptive use among unmarried female migrants in one factory in Shanghai: a pilot workplace intervention. BMC Health Serv Res 2007 May 31;7(1):77 [FREE Full text] [doi: 10.1186/1472-6963-7-77] [Medline: 17540025]
16. Zhu C, Geng Q, Chen L, Yang H, Jiang W. Impact of an educational programme on reproductive health among young migrant female workers in Shenzhen, China: an intervention study. Int J Behav Med 2014 Aug 19;21(4):710-718. [doi: 10.1007/s12529-014-9401-y] [Medline: 24643443]
17. Govender K, Beckett S, Masebo W, Braga C, Zambezi P, Manhique M, et al. Effects of a short message service (SMS) intervention on reduction of HIV risk behaviours and improving HIV testing rates among populations located near roadside wellness clinics: a cluster randomised controlled trial in South Africa, Zimbabwe and Mozambique. AIDS Behav 2019 Nov 15;23(11):3119-3128. [doi: 10.1007/s10461-019-02427-6] [Medline: 30771133]
18. El-Bassel N, Gilbert L, Goddard-Eckrich D, Chang M, Wu E, Hunt T, et al. Efficacy of a group-based multimedia HIV prevention intervention for drug-involved women under community supervision: project WORTH. PLoS One 2014 Nov 5;9(11):e111528 [FREE Full text] [doi: 10.1371/journal.pone.0111528] [Medline: 25372149]
19. Duan Y, Li X, Guo L, Liang W, Shang B, Lippke S. A WeChat mini program-based intervention for physical activity, fruit and vegetable consumption among Chinese cardiovascular patients in home-based rehabilitation: a study protocol. Front Public Health 2022 Mar 22;10:739100 [FREE Full text] [doi: 10.3389/fpubh.2022.739100] [Medline: 35392478]
20. Huo J, Wu X, Gu C, Yu Z, Zhang J, Chen X, et al. Using a WeChat mini-program-based lactation consultant intervention to increase the consumption of mother's own milk by preterm infants in the neonatal intensive care unit: a study protocol for a cluster randomized controlled trial. Trials 2021 Nov 24;22(1):834 [FREE Full text] [doi: 10.1186/s13063-021-05731-6] [Medline: 34819117]
21. Ye Q, Yang Y, Yao M, Yang Y, Lin T. Effects of teach-back health education (TBHE) based on WeChat mini-programs in preventing falls at home for urban older adults in China: a randomized controlled trial. BMC Geriatr 2022 Jul 23;22(1):611 [FREE Full text] [doi: 10.1186/s12877-022-03297-9] [Medline: 35870888]
22. Liu Y, Li D, Ruan H, Hu Y, Shen N. Development and usability test of a symptom management WeChat Mini Program for parents of children with cancer. Asia Pac J Oncol Nurs 2022 Dec;9(12):100166 [FREE Full text] [doi: 10.1016/j.apjon.2022.100166] [Medline: 36479577]

XSL•FO
RenderX

23. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
24. Zhang K, Chan PS, Li X, Fang Y, Cai Y, Zou H, et al. Low behavioral intention to use any type of HIV testing and HIV self-testing among migrant male factory workers who are at high risk of HIV infection in China: a secondary data analysis. Int J Environ Res Public Health 2023 Mar 13;20(6):5029 [FREE Full text] [doi: 10.3390/ijerph20065029] [Medline: 36981938]
25. Pan Y, Fang Y, Xin M, Dong W, Zhou L, Hou Q, et al. Self-reported compliance with personal preventive measures among Chinese factory workers at the beginning of work resumption following the COVID-19 outbreak: cross-sectional survey study. J Med Internet Res 2020 Sep 29;22(9):e22457 [FREE Full text] [doi: 10.2196/22457] [Medline: 32924947]
26. Yuan J, Cao B, Zhang C, Chan PS, Xin M, Fang Y, et al. Changes in compliance with personal preventive measures and mental health status among Chinese factory workers during the COVID-19 pandemic: an observational prospective cohort study. Front Public Health 2022 Mar 10;10:831456 [FREE Full text] [doi: 10.3389/fpubh.2022.831456] [Medline: 35359764]
27. Pan Y, Xin M, Zhang C, Dong W, Fang Y, Wu W, et al. Associations of mental health and personal preventive measure compliance with exposure to COVID-19 information during work resumption following the COVID-19 outbreak in China: cross-sectional survey study. J Med Internet Res 2020 Oct 08;22(10):e22596 [FREE Full text] [doi: 10.2196/22596] [Medline: 32936776]
28. Wang Z, Yang L, Hao C, Jiang H, Zhu J, Luo Z, et al. A randomized controlled trial evaluating efficacy of a brief setting-based and theory-based intervention promoting voluntary medical male circumcision among heterosexual male sexually transmitted disease patients in China. AIDS Behav 2019 Sep 18;23(9):2453-2466. [doi: 10.1007/s10461-019-02610-9] [Medline: 31321636]
29. Nguyen VH, Dunne MP, Debattista J, Nguyen TH, Dao TM. Social contexts of risk behaviors for HIV among male, unskilled, unregistered laborers in urban Vietnam. Qual Health Res 2012 Jul 07;22(7):871-879. [doi: 10.1177/1049732311424510] [Medline: 22068045]
30. de Vries A, den Daas C, Willemstein IJ, de Wit JB, Heijne JC. Interventions promoting condom use among youth: a systematic review. J Adolesc Health 2024 Apr;74(4):644-656 [FREE Full text] [doi: 10.1016/j.jadohealth.2023.11.014] [Medline: 38085203]
31. Chen D, Zhang H, Wu J, Xue E, Guo P, Tang L, et al. Effects of an individualized mHealth-based intervention on health behavior change and cardiovascular risk among people with metabolic syndrome based on the behavior change wheel: quasi-experimental study. J Med Internet Res 2023 Nov 29;25:e49257 [FREE Full text] [doi: 10.2196/49257] [Medline: 38019579]
32. Wu W, Hu L, Chen Y, Cao F, Ding S, Wu T, et al. Effectiveness of an online application of the health action process approach (HAPA) theory on oral hygiene intervention in young adults with fixed orthodontic appliances: a randomized controlled trial. BMC Oral Health 2022 May 19;22(1):192 [FREE Full text] [doi: 10.1186/s12903-022-02219-w] [Medline: 35590291]
33. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl Psychol 2008 Jul 08;57(4):660-680. [doi: 10.1111/j.1464-0597.2008.00341.x]
34. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Feb 19;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
35. Leask CF, Sandlund M, Skelton DA, Altenburg TM, Cardon G, Chinapaw MJ, et al. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019 Jan 9;5(1):2 [FREE Full text] [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]
36. Galvagno M, Dalli D. Theory of value co-creation: a systematic literature review. Manag Serv Qual Int J 2014;24(6):643-683. [doi: 10.1108/MSQ-09-2013-0187]
37. Sanders EB, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
38. Information disclosure. Public Hygiene and Health Commission of Shenzhen Municipality. URL: http://wjw.sz.gov.cn/yqxx/index_42.html [accessed 2023-05-24]
39. WHO COVID-19 dashboard. World Health Organization. URL: https://covid19.who.int/region/wpro/country/cn [accessed 2023-05-24]
40. Higginbotham N. An education intervention to increase the use of condoms for the prevention of HIV infection among male factory workers in Khon Kaen. J Clin Epidemiol 1996;49:13S. [doi: 10.1016/0895-4356(96)89201-8]
41. Sakondhavat C, Sittitrai W, Soontharapa S, Werawatanakul Y, Pinitsoontorn P. AIDS education and intervention trials among youths in factories: a pilot project. J Med Assoc Thai 1998 Nov;81(11):872-878. [Medline: 9803088]
42. Ng'weshemi JZ, Boerma JT, Pool R, Barongo L, Senkoro K, Maswe M, et al. Changes in male sexual behaviour in response to the AIDS epidemic: evidence from a cohort study in urban Tanzania. AIDS 1996 Oct;10(12):1415-1420. [doi: 10.1097/00002030-199610000-00015] [Medline: 8902072]

XSL•FO
RenderX

43. Sahlu T, de Wit TR, Tsegaye A, Mekonnen Y, Beyene A, Hailu B, et al. Low incidence of syphilis among factory workers in Ethiopia: effect of an intervention based on education and counselling. Sex Transm Infect 2002 Apr;78(2):123-126 [FREE Full text] [doi: 10.1136/sti.78.2.123] [Medline: 12081173]
44. Mekonnen Y, Sanders E, Aklilu M, Tsegaye A, Rinke de Wit TF, Schaap A, et al. Evidence of changes in sexual behaviours among male factory workers in Ethiopia. AIDS 2003 Jan 24;17(2):223-231. [doi: 10.1097/00002030-200301240-00013] [Medline: 12545083]
45. Booton RD, Fu G, MacGregor L, Li J, Ong JJ, Tucker JD, et al. The impact of disruptions due to COVID-19 on HIV transmission and control among men who have sex with men in China. J Int AIDS Soc 2021 Apr;24(4):e25697 [FREE Full text] [doi: 10.1002/jia2.25697] [Medline: 33821553]
46. Zhang K, Chen S, Chan PS, Fang Y, Cao H, Chen H, et al. Changes in HIV testing utilization among Chinese men who have sex with men during the COVID-19 Pandemic in Shenzhen, China: an observational prospective cohort study. Front Med (Lausanne) 2022;9:842121 [FREE Full text] [doi: 10.3389/fmed.2022.842121] [Medline: 35755039]
47. Suen YT, Chan RC, Wong EM. An exploratory study of factors associated with difficulties in accessing HIV services during the COVID-19 pandemic among Chinese gay and bisexual men in Hong Kong. Int J Infect Dis 2021 May;106:358-362 [FREE Full text] [doi: 10.1016/j.ijid.2021.04.005] [Medline: 33845197]
48. Zhang KC, Fang Y, Cao H, Chen H, Hu T, Chen YQ, et al. The impacts of the COVID-19 pandemic on HIV testing utilization among men who have sex with men in China: cross-sectional online survey. JMIR Public Health Surveill 2022 May 25;8(5):e30070 [FREE Full text] [doi: 10.2196/30070] [Medline: 35486811]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
RCT: randomized controlled trial
STI: sexually transmitted infection

Edited by S Ma; submitted 25.05.23; peer-reviewed by Y Zhou, J Xu; comments to author 05.04.24; revised version received 25.04.24; accepted 25.06.24; published 09.09.24.

Please cite as:
Zhang K, Cao B, Fang Y, Liang X, Ye D, Chen YQ, Zhong R, Cao H, Hu T, Li T, Cai Y, Zou H, Wang Z
Comparing the Efficacy of 2 WeChat Mini Programs in Reducing Nonmarital Heterosexual Contact by Male Factory Workers: Randomized Controlled Trial
J Med Internet Res 2024;26:e49362
URL: https://www.jmir.org/2024/1/e49362
doi:10.2196/49362
PMID:39250213

©Kechun Zhang, Bolin Cao, Yuan Fang, Xue Liang, Danhua Ye, Ya Qi Chen, Ruilan Zhong, He Cao, Tian Hu, Ting Li, Yong Cai, Huachun Zou, Zixin Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Web-Based Antiretroviral Therapy Adherence Intervention (Thrive With Me) in a Community-Recruited Sample of Sexual Minority Men Living With HIV: Results of a Randomized Controlled Study

Keith J Horvath[1], PhD; Sara Lammert[2], PhD; Darin Erickson[2], PhD; K Rivet Amico[3], PhD; Ali J Talan[4], DrPH; Ore Shalhav[5], MPH; Christina J Sun[6], PhD; H Jonathon Rendina[7], PhD

[1]Department of Psychology, San Diego State University, San Deigo, CA, United States
[2]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, United States
[3]Health Behavior and Health Education, School of Public Health, University of Michigan, Ann Arbor, MI, United States
[4]Whitman-Walker Institute, Washington, DC, MD, United States
[5]City University of New York - Hunter College, New York City, NY, United States
[6]College of Nursing, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[7]Department of Epidemiology, Milken Institute School of Public Health, George Washington University, Washington DC, MD, United States

Corresponding Author:
Keith J Horvath, PhD
Department of Psychology
San Diego State University
6363 Alvarado Court
Suite 250
San Deigo, CA, 92120
United States
Phone: 1 6196943346
Email: khorvath@sdsu.edu

Abstract

Background: Most new HIV infections are attributed to male-to-male sexual contact in the United States. However, only two-thirds of sexual minority men living with HIV achieve an undetectable viral load (UVL). We tested a web-based antiretroviral therapy adherence intervention called Thrive with Me (TWM) with core features that included medication self-monitoring and feedback, HIV and antiretroviral therapy information, and a peer-to-peer exchange.

Objective: We assessed the efficacy of TWM on HIV UVL among adult (aged ≥18 years) sexual minority men. Moreover, we assessed the impact of overall engagement and engagement with specific intervention features on HIV UVL.

Methods: In total, 401 sexual minority men (mean age 39.1, SD 10.8 y; 230/384, 59.9% African American) in New York City were recruited between October 2016 and December 2019 and randomized to receive TWM (intervention) or a weekly email newsletter (control) for 5 months. Computerized assessments occurred at baseline and months 5, 11, and 17. The primary outcome was a dichotomous measure of HIV UVL (≤20 copies/µL). Generalized estimating equations with robust SEs were used to assess the effect of the TWM intervention on HIV UVL over the follow-up period in an unadjusted model and a model adjusted for baseline differences and then stratified by baseline recent drug use urinalysis. In secondary analyses, generalized linear models were used to estimate risk differences in the association of overall engagement with TWM (the sum of the number of days participants accessed ≥1 screen of the TWM intervention out of a possible 150 days) and engagement with specific TWM components on HIV UVL throughout the 17-month intervention period.

Results: Participant retention was 88.5% (355/401; month 5), 81.8% (328/401; month 11), and 80.3% (322/401; month 17). No consistent differences in HIV UVL were found between those randomized to receive TWM or the control at the 5- (difference-in-differences [DD]=–7.8, 95% CI –21.1 to 5.5), 11- (DD=–13.9, 95% CI –27.7 to 0.04), or 17-month (DD=–8.2, 95% CI –22.0 to 5.7) time points, or when stratified by baseline recent drug use. However, those TWM-assigned participants with high overall levels of engagement (in the upper 25th percentile) were more likely to have an HIV UVL at the end of the 5-month active intervention period compared to those with low engagement (below the 75th percentile; risk difference=17.8,

XSL•FO
RenderX

95% CI 2.5-33.0) or no engagement (risk difference=19.4, 95% CI 3.3-35.5) in the intervention. Moreover, high engagement with the peer-to-peer exchange was associated with HIV UVL over time in unadjusted models.

Conclusions: TWM did not have overall impacts on HIV UVL; however, it may assist some sexual minority men who are highly engaged with this web-based intervention in achieving HIV viral suppression.

Trial Registration: ClinicalTrials.gov NCT02704208; https://clinicaltrials.gov/study/NCT02704208

(J Med Internet Res 2024;26:e53819) doi:10.2196/53819

KEYWORDS

HIV; antiretroviral therapy; ART; mobile health; mHealth; intervention; men who have sex with men

Introduction

Background

Sexual minority men, especially Black and Latinx men, are at highest risk of acquiring HIV in the United States [1]. It is estimated that 71% of all cases of HIV in the United States (and 89% among male individuals) in 2020 were attributed to male-to-male sexual contact, most of which occurred among Black (39%) and Hispanic or Latino (31%) sexual minority men [1]. Disparities persist in HIV treatment outcomes. Among people living with HIV, it is clear that greater efforts are needed to reach the Ending the HIV Epidemic initiative goal of 90% of people living with HIV being virally suppressed by 2030 [2]. High proportions of people living with HIV achieving an undetectable viral load (UVL) is important to both optimize individual health and substantially reduce onward transmission [3-5]. However, among male individuals living with HIV attributed to male-to-male sexual contact in 2020, a total of 82% were linked to care within 1 month of their diagnosis, 76% received any HIV medical care, and 67% achieved viral suppression in 2020 [6]. For these reasons, there is a continuing need for innovative antiretroviral therapy (ART) adherence interventions for sexual minority men living with HIV.

Sexual minority men living with HIV must navigate unique individual, social, and structural challenges—including intersectional stigma, racism, and economic and legal hardships—throughout the HIV continuum of care [7]. Drug use, particularly stimulant use [8], has been consistently associated with HIV seroconversion [9,10], as well as poorer ART adherence that leads to suboptimal rates of viral suppression [11-14]. Stimulant-using sexual minority men have greater difficulties achieving and sustaining viral suppression, which could unnecessarily quicken clinical HIV progression [15-17] and increase risk of onward HIV transmission [18-20].

Efforts to develop and test eHealth and, more narrowly, mobile health (mHealth) interventions to address the HIV prevention and care continuum have expanded in recent years [21-23]. Reviews of eHealth- and mHealth-based ART adherence interventions have found that most have pilot or quasi-experimental designs, and the results of these studies suggest a high potential of eHealth and mHealth approaches to improve ART adherence [21,22]. However, the results of larger randomized controlled trials of technology-based ART adherence interventions are mixed. One study showed significant reductions in plasma HIV viral load (VL) in the treatment arm compared to the control arm [24]. Other studies have shown impacts on electronic dose monitoring or self-reported adherence outcomes but not on plasma HIV VL [25-27] or only on plasma HIV VL for a subset of participants [28] or no impacts on plasma HIV VL or adherence [29]. For these reasons, further efforts are needed to understand how eHealth or mHealth interventions can be optimized to have impacts on clinically meaningful outcomes.

In addition to assessing overall intervention outcomes, it is critical that we have a better understanding of the effects of engagement with digital interventions on treatment outcomes [30]. Intervention engagement metrics (eg, the amount, frequency, duration, and depth of engagement with the entire or parts of a digital intervention) through the collection of paradata (ie, data collected on users' actual activity in a digital intervention) may provide a more nuanced understanding of what features work for whom to advance digital intervention science [30-32]. A recent review of the use and findings of digital HIV prevention and treatment interventions for sexual and gender minority persons from 2017 to 2023 showed that, of the 33 digital HIV interventions published in those years, 19 reported paradata engagement metrics. However, only 6 studies reported the associations between intervention engagement and intermediate or primary study outcomes [30], of which 4 showed a positive association between intervention engagement (defined in a variety of ways) and health promotion behaviors [30]. Although these findings are encouraging, more information across a broad spectrum of digital HIV prevention and treatment interventions is needed to guide further investments into these approaches.

Objectives

In this paper, we report the results of the Thrive with Me (TWM) web-based intervention [33] in a large community sample of sexual minority men living with HIV. We hypothesized that a higher proportion of participants in the TWM intervention than control participants would have HIV UVL after the intervention and that participants in the TWM intervention who had recently used illicit drugs would demonstrate the greatest improvements in HIV UVL after the intervention compared to participants who had not recently used illicit drugs. As a secondary aim, we examined the effect of level of engagement with the TWM intervention and its unique components on HIV UVL among those assigned to the TWM study arm. Such an analysis, although infrequently done [30,31], is critical to increase understanding of the degree to which exposure to eHealth or mHealth interventions and their individual components may be associated with HIV outcomes.

XSL•FO
RenderX

Methods

Participants and Procedures

Sexual minority men residing in the New York City metropolitan area were recruited between October 2016 and April 2018 by staff at a lesbian, gay, bisexual, transgender, and queer health research center at Hunter College, City University of New York. Interested persons completed a web-based screening survey that assessed eligibility across multiple studies; next, persons who met the basic criteria (eg, living with HIV) for TWM were contacted to complete a second screener specific to the TWM study to determine eligibility. Eligibility criteria were (1) identified current gender as male, (2) report of having had sex with a man in the previous year, (3) HIV-positive serostatus, (4) self-report of a detectable VL in the previous year or ART adherence of <90% in the previous 30 days, (5) English-language proficiency, (6) the ability to send or receive SMS text messages, and (7) internet access over the course of the active intervention (5 months). Recruitment included a 50% target of self-reported street drug use (ie, powder cocaine, crack cocaine, painkillers, methamphetamine, heroin, hallucinogens, recreationally used prescription drugs, ketamine, methylenedioxymethamphetamine, and poppers). Participants meeting the inclusion criteria were invited to attend an in-person enrollment visit that included completing a consent process with web-based documentation of consent, a computerized survey, urine-based drug use test, and a blood draw to assess HIV VL at the research offices at baseline and months 5 (immediately after the intervention), 11, and 17.

Randomization and Blinding

Following consent and completion of the enrollment activities described previously, research staff used a computer algorithm to randomize participants to either the TWM arm or control arm (1:1) using blocks of 20 (10 for the intervention and 10 for the control group). The same computer algorithm stratified enrollment by recent drug use (ie, reporting use of powder cocaine, crack cocaine, painkillers, methamphetamine, heroin, hallucinogens, prescription drugs used recreationally, ketamine, methylenedioxymethamphetamine, or poppers in the previous 30 days) and nonrecent drug use based on their self-reported substance use in the baseline survey. Study arms were described to participants as “Group 1” and “Group 2”; however, study staff were not blinded as to which arm the participants were assigned to.

The TWM and Control Groups

Participants were onboarded to either the TWM or control condition as the final step during the enrollment visit. Participants randomized to the TWM study arm created a username and password that would give them free access to the intervention, and study staff provided information about each part of the intervention, described in the following paragraph. Participants randomized to the control condition were told that they should receive the first email with the newsletter that week.

TWM is a 5-month (150 days) mobile-optimized web-based intervention to improve ART adherence among adult (aged ≥18 years) HIV-positive sexual minority men residing in New York City with detectable VL or problematic ART adherence at baseline. An earlier pilot trial of TWM demonstrated feasibility, acceptability, and preliminary impact [34]. TWM is grounded in the Information-Motivation-Behavioral Skills (IMB) model [35,36] and has the following core components: (1) a peer exchange (similar to Facebook) that allows participants to post and comment on text, pictures, and videos in an unstructured format; (2) brief, tailored HIV and ART adherence informational articles (called Thrive Tips; 300 were created in total) that were delivered daily, with the ability to search past content; (3) daily SMS text message–based ART reminders that also provide an option for participants to report (through texting back) their ART adherence and overall mood for that day; and (4) weekly reflection on adherence, mood, and substance use. In addition, the TWM site has a profile page to update settings and viewpoints and badges earned, an “about us” page, and pages to explain how to use the site. Participants randomized to the TWM intervention were encouraged during the enrollment visit to use the intervention regularly; however, they were not required to meet a minimum engagement standard. A weekly draw for a US $25 gift card for those who had used the site ≥5 times in the previous 10 days was held during the active study period.

Participants randomized to the control condition received a weekly email for 21 consecutive weeks that contained a link that, when clicked on, opened a web page with information on a specific topic related to living with HIV and improving general well-being (eg, HIV disclosure and managing stress) but not specifically about ways to improve ART adherence. An example weekly newsletter on the topic of managing stress is shown in Multimedia Appendix 1. More information about the TWM intervention and control arms is available elsewhere [33].

Measures

Laboratory Measures

Plasma HIV-1 VL

Blood draws were collected at baseline and the 5-, 11-, and 17-month follow-ups at the Hunter research offices by a trained phlebotomist and assayed by Quest Diagnostics for the presence of plasma HIV-1 RNA. To assess the primary outcome, we used anything below the lower limit of detection of 20 copies per microliter to define HIV UVL.

Drug Use

Participants completed urine screens at each assessment time point using the Integrated E-Z Split Key Cup II-5 panel (Innovacon Laboratories) to detect the following substances: tetrahydrocannabinol (THC; marijuana), methamphetamine, amphetamine, cocaine, and opioids. This test can detect methamphetamine, amphetamine, cocaine, and opioids from 1 to 4 days after use and THC (marijuana) for up to 30 days after use. For this analysis, a positive baseline urinalysis was defined as any detectable measure of methamphetamine, amphetamine, cocaine, or opioids.

XSL•FO
RenderX

Survey Measures

Sociodemographic Characteristics

At baseline, participants were asked common sociodemographic factors, including age (in years), race and ethnicity, highest level of education, and current employment status. Additional HIV history data were collected, including 30-day ART adherence.

Psychosocial Characteristics

The following psychosocial characteristics were assessed using a computerized survey at each assessment point, of which baseline measures were used for the purpose of characterizing participants in this study. Depressive symptoms were assessed using the 10-item Center for Epidemiologic Studies Depression Scale [37] and dichotomized into depressive symptoms (score of ≥10) and no depressive symptoms (score of <10). ART-related information (9 items), personal and social motivation (10 items), and behavioral skills (14 items) were assessed using the IMB ART Adherence Questionnaire [38]. HIV stigma, including internalized (6 items), anticipated (9 items), and enacted (9 items) stigma was assessed using the HIV Stigma Scale [39]. Social support, including emotional or information support (8 items), tangible support (4 items), affectionate support (3 items), positive social interaction (3 items), and an overall social support score, was assessed using the Medical Outcome Study Social Support Survey [40]. Life chaos and perceived stress were assessed using the 6-item Life Chaos Scale [41] and the 14-item Perceived Stress Scale [42], respectively. Finally, alcohol use was assessed using the Alcohol Use Disorders Identification Test [43].

TWM Intervention Engagement Measures

Overview

TWM engagement data were only collected from individuals randomized to receive the TWM intervention (202/401, 50.4%) during the 5-month (150 days) active intervention period. Engagement data were collected via a customized back-end website accessible only to study staff and investigators that provided these data in CSV files that were updated as participants engaged with the intervention. These CSV files were downloaded after study participants were no longer active in the TWM intervention arm and cleaned for analyses.

Overall TWM Engagement

To quantify overall engagement with the TWM intervention, we summed the number of days on which a participant accessed at least one screen of the TWM intervention out of a possible 150 days. Those who did not log onto the TWM intervention during the intervention period were categorized as nonengagers, whereas those who accessed at least one screen of the TWM intervention were considered engagers. Next, those who engaged with TWM were further categorized as having high or low engagement based on dichotomization at the 75th percentile. While there are no established thresholds for definitions of engagement with digital interventions, we chose this level of engagement (as opposed to other possible options, such as the 50th percentile) to represent participants who were more obviously active in the TWM intervention than, for example, a participant who was above the 50th percentile but not by much (eg, the 55th percentile). Using this cutoff, engagement with TWM was categorized as follows: nonengagers (0 days active), low engagement with TWM (1-33 days active), and high engagement with TWM (≥34 active days).

Engagement With Individual TWM Components

We calculated engagement with three individual TWM components: (1) peer exchanges, (2) Thrive Tips (ie, brief informational pieces of content), (3) and ART adherence self-monitoring. Engagement with the asynchronous peer exchanges was measured using the sum of the number of unique wall posts and comments made by a user during the TWM intervention period (ie, up to 150 days). Engagement with the Thrive Tips was calculated as the number of unique Thrive Tips that participants viewed (possible range 0-300). Engagement with the ART adherence self-monitoring feature was the number of days on which participants reported their ART adherence (possible range 0-150). Participants were categorized as having high or low engagement with the individual TWM components based on dichotomization of each component at the 75th percentile.

Statistical Analyses

Descriptive statistics (means and SDs and frequencies and percentages) were generated to describe the participants in the TWM study. Appropriate statistical tests were used to compare numeric (2-tailed *t* test) and categorical (chi-square test) variables between the TWM intervention and control arms at baseline. All analyses were conducted using Stata (version 14; StataCorp), and significance was established at $P<.05$.

To account for missing data during follow-up, we used multiple imputation. Demographic, psychosocial, and substance use factors that were associated with having an HIV UVL (<20 copies/μL) at baseline were used to impute missing variables at subsequent time points and HIV UVLs at the 5-, 11-, and 17-month follow-ups.

We used generalized estimating equations with robust SEs to assess the effect of the TWM intervention on HIV UVL over the follow-up period in an unadjusted model and a model adjusted for baseline differences and then stratified by baseline urinalysis (ie, positive vs negative drug use). The main effects of interest were the condition-by-time interaction terms at months 5, 11, and 17. The Stata *xtgee* command was used to calculate the difference-in-differences (DD) estimate as the proportion of those with an HIV UVL between participants in the TWM intervention arm and those in the control arm. Described in more detail elsewhere [44], a DD estimation first calculates the difference between each follow-up time point and baseline estimates within each study arm; using those within-arm difference estimates, the DD estimate (and the associated CI) is arrived at by finding the difference between study arms for each assessment time point.

To assess the impact of engagement with the TWM intervention on viral suppression, we modeled overall engagement with the TWM intervention as well as engagement with the 3 TWM components described previously on HIV UVL (<20 copies/μL) throughout the 17 months for participants assigned to the TWM intervention (202/401, 50.4%). Generalized linear models with

XSL•FO
RenderX

robust SEs were used to estimate risk differences (RDs) and 95% CIs of the association of overall engagement with TWM and engagement with specific TWM components on viral suppression throughout the 17-month intervention period, including unadjusted models and models adjusted for differences between TWM users and nonusers.

Ethical Considerations

All study procedures were approved by the University of Minnesota Institutional Review Board and Hunter College, City University of New York, Institutional Review Board (IRB Study 1120). In addition, a Data Safety and Monitoring Board was established to provide regular oversight of research practices and activities to protect human participants and the integrity of the data in the study. Written informed consent was provided by all study participants, and data were deidentified before analyses for this study. Participants were paid US $50 in cash at the baseline and 5-, 11-, and 17-month assessments. Participants in the intervention group were also eligible to win a weekly prize draw for a US $25 electronic gift card if they were active on the TWM website for 5 out of the 10 preceding days. This study was registered on the national registry of clinical trials at ClinicalTrials.gov (NCT02704208).

Results

A total of 401 men who have sex with men living with HIV were recruited to participate in the TWM study. Following a 1:1 randomization, 50.4% (202/401) of the participants were randomized to receive the TWM intervention, whereas 49.6% (199/401) were randomized to receive the control condition (Figure 1).

Figure 1. Thrive with Me (TWM) intervention trial CONSORT (Consolidated Standards of Reporting Trials) diagram. ART: antiretroviral therapy; mHealth: mobile health; VL: viral load.

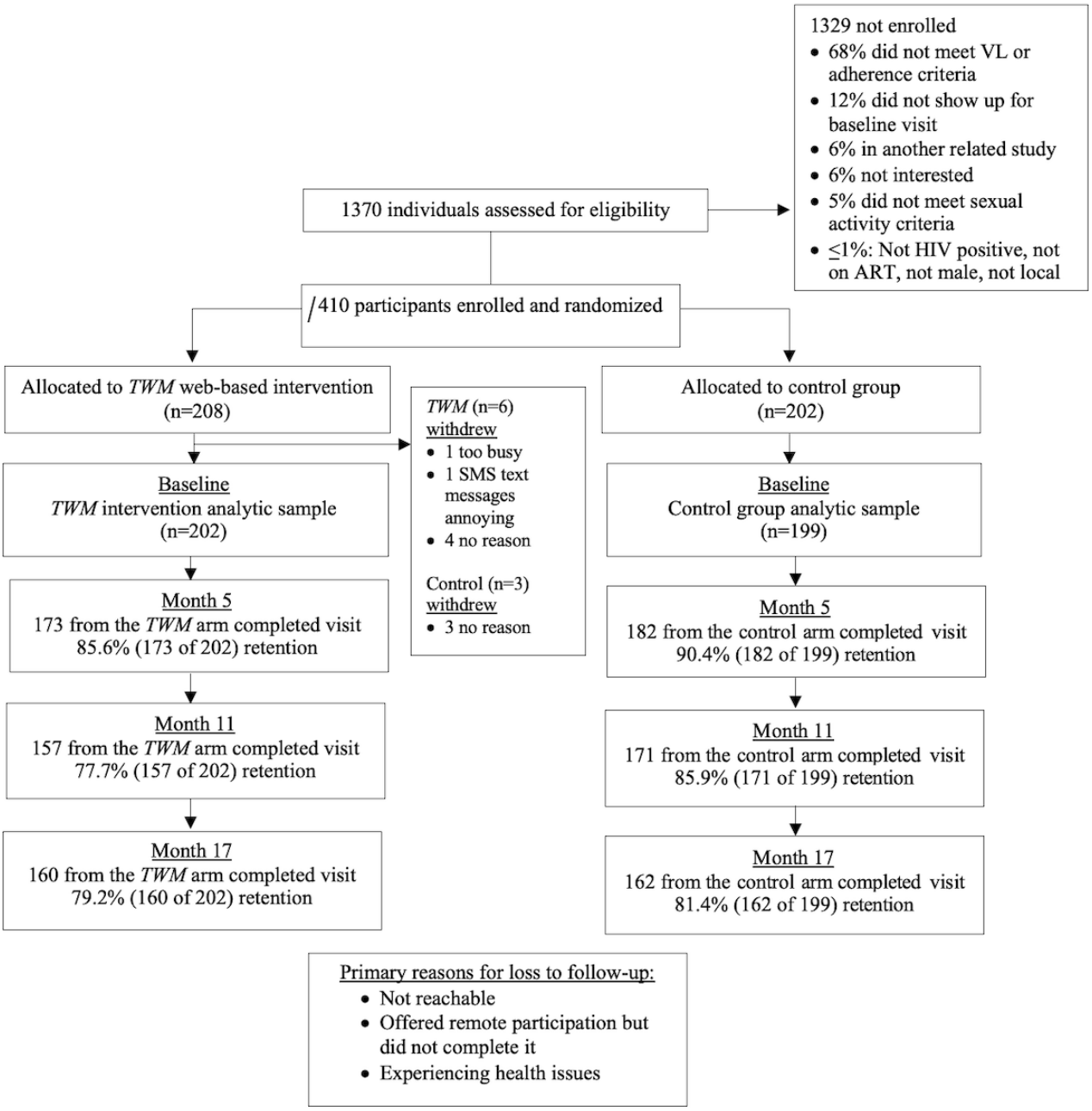

XSL•FO
RenderX

Baseline Characteristics

Table 1 shows participant characteristics for the full sample and by study arm. The average age of the participants was 39.1 (SD 10.8) years. More than half (230/384, 59.9%) of the participants identified as African American or Black individuals, 29.4% (113/384) identified as White individuals, and almost 8% (31/384, 8.1%) identified as more than one race. In addition, more than a quarter (108/401, 26.9%) identified as Hispanic or Latino. More than 40% (175/397, 44.1%) were employed full time or part time, and an additional 39.3% (156/397) were unemployed. Overall, approximately three-quarters of the participants (303/400, 75.8%) had some college education or a college degree, with a higher proportion of participants in the control than the TWM arm having a college degree (84/199, 42.2% vs 52/201, 25.9%, respectively). Nearly half (194/401, 48.4%) of the participants self-reported depressive symptoms at baseline. No group differences were found for perceived stress, HIV stigma, social support, or life chaos.

Excluding marijuana, approximately one-quarter of the participants (113/401, 28.2%) at baseline had detectable levels of substance use from urinalysis. Over 10% of the sample exhibited baseline levels of methamphetamine (58/401, 14.5%), cocaine (52/401, 13%), and amphetamine (45/401, 11.2%); however, only 1.2% (5/401) of the participants tested positive for opioids. A total of 41.1% (165/401) of the participants tested positive for THC at baseline. More than one-quarter of the participants (116/401, 28.9%) self-reported hazardous or harmful alcohol use or alcohol dependence.

At baseline, participants reported an average percentage of ART adherence in the previous 30 days of 87.6% (SD 17.6%), with no group differences found in this or the IMB ART Adherence Questionnaire ART-related information, motivation, or behavioral skills variables.

XSL•FO
RenderX

Table 1. Baseline characteristics of the participants in the Thrive with Me (TWM) intervention and control conditions.

	Total (N=401)	TWM study arm (n=202)	Control study arm (n=199)
Demographics			
Age (y), mean (SD)	39.1 (10.8)	40.1 (10.8)	38.1 (10.6)
Race, n (%)[a]			
African American or Black	230 (59.9)	123 (63.7)	107 (56)
American Indian or Alaska Native	5 (1.3)	1 (0.5)	4 (2.1)
Asian	3 (0.8)	1 (0.5)	2 (1)
Native Hawaiian or Pacific Islander	2 (0.5)	2 (1)	0 (0)
White	113 (29.4)	54 (28)	59 (30.9)
>1 Race	31 (8.1)	12 (6.2)	19 (9.9)
Hispanic, n (%)	108 (26.9)	62 (30.7)	46 (23.1)
Educational level, n (%)[b]			
High school or lower	97 (24.3)	59 (29.4)	38 (19.1)
Some college education, associate degree, and technical degree	167 (41.8)	90 (44.8)	77 (38.7)
College, postgraduate, and professional degree	136 (34)	52 (25.9)	84 (42.2)
Employment status, n (%)[c]			
Full time	87 (21.9)	41 (20.5)	46 (23.4)
Part time	88 (22.2)	45 (22.5)	43 (21.8)
Unemployed	156 (39.3)	77 (38.5)	79 (40.1)
Disabled	63 (15.9)	35 (17.5)	28 (14.2)
Retired	3 (0.8)	2 (0.1)	1 (0.5)
VL[d] and ART[e] adherence measures			
VL (≤20 copies/μL), n (%)[f]			
Detectable VL	154 (38.5)	74 (36.8)	80 (40.2)
Undetectable VL	246 (61.5)	127 (63.2)	119 (59.8)
30-day adherence, mean (SD)	87.6 (17.6)	87.2 (18.6)	88.0 (16.6)
IMB[g] scale for ART adherence, mean (SD)			
Information	36.7 (5.8)	37.0 (5.8)	36.4 (5.9)
Motivation	34.0 (8.4)	34.8 (8.1)	33.2 (8.5)
Behavior	48.2 (8.5)	48.8 (8.3)	47.6 (8.8)
Drug and alcohol use measures			
Positive urinalysis[h], n (%)	113 (28.2)	65 (32.2)	48 (24.1)
Marijuana use	165 (41.1)	88 (43.6)	77 (38.7)
Methamphetamine use	58 (14.5)	31 (15.3)	27 (13.6)
Amphetamine use	45 (11.2)	25 (12.4)	20 (10.1)
Cocaine use	52 (13)	33 (16.3)	19 (9.5)
Opioid use	5 (1.2)	2 (1)	3 (1.5)
Hazardous or harmful alcohol use, n (%)	116 (28.9)	61 (30.2)	55 (27.6)
Additional sociodemographics			
Depressive symptoms, n (%)	194 (48.4)	100 (49.5)	94 (47.2)
Perceived Stress Scale score, mean (SD)	16.8 (7.3)	17.0 (7.3)	16.6 (7.4)

XSL•FO
RenderX

	Total (N=401)	TWM study arm (n=202)	Control study arm (n=199)
HIV Stigma Scale score, mean (SD)			
Internalized stigma	2.2 (1.1)	2.2 (1.1)	2.2 (1.1)
Anticipated stigma	2.0 (0.9)	2.0 (0.9)	2.0 (0.8)
Enacted stigma	1.5 (0.7)	1.5 (0.7)	1.6 (0.7)
Medical Outcome Study Social Support Survey score, mean (SD)			
Emotional support	3.5 (1.2)	3.6 (1.2)	3.5 (1.2)
Affectionate support	3.5 (1.3)	3.5 (1.3)	3.4 (1.4)
Tangible support	3.1 (1.4)	3.1 (1.4)	3.1 (1.4)
Social interaction support	3.6 (1.3)	3.6 (1.2)	3.5 (1.3)
Overall social support	3.4 (1.1)	3.5 (1.1)	3.4 (1.2)
HIV life chaos (Life Chaos Scale), mean (SD)	15.9 (4.8)	15.8 (4.7)	15.9 (4.8)

[a]17 participants did not report race.

[b]1 participant did not report educational level.

[c]4 participants did not report employment status.

[d]VL: viral load.

[e]ART: antiretroviral therapy.

[f]1 participant did not undergo VL testing.

[g]IMB: Information-Motivation–Behavioral Skills model.

[h]A positive urinalysis was defined as any detectable measure of methamphetamine, amphetamine, cocaine, or opioids.

Study Retention

Overall, nearly 90% of the participants were retained at month 5 (355/401, 88.5%), whereas 81.8% (328/401) and 80.3% (322/401) were retained at months 11 and 17, respectively. Differences in retention were found at month 11 (intervention: 157/202, 77.7% vs control: 171/199, 85.9%). On average, those who were retained across both study arms at month 11 had lower affectionate support (mean: 3.4; SD: 1.4) compared to those who were not retained (mean: 3.8; SD: 1.3). No other sociodemographic differences by study arm were found. Retention by treatment and control group can be found in the CONSORT (Consolidated Standards of Reporting Trials) [45] flow diagram in Figure 1.

The Effect of TWM on HIV VL for the Overall Sample and Stratified by Baseline Drug Use

Approximately two-thirds of the participants had an HIV UVL at baseline (266/401, 66.3% in the adjusted model), as shown in Table 2. The proportion of all participants with an HIV UVL remained relatively consistent over the follow-up period (month 5: 239/401, 59.6%; month 11: 224/401, 55.9%; month 17: 219/401, 54.5%).

Overall, the proportion of participants with HIV UVL decreased from baseline to month 17 across both study arms. In unadjusted models, there was no difference in HIV UVL between those randomized to receive the TWM intervention and those randomized to the control group at any of the follow-up time points, as shown in Table 3. Similarly, adjusting for baseline educational level, there was no difference in HIV UVL between those randomized to receive the TWM intervention and the control arm at the 5- and 17-month assessments. However, at the 11-month assessment, participants randomized to the TWM intervention were less likely to have an HIV UVL compared to those randomized to the control group in the adjusted model (DD=–13.9 95% CI –27.7 to –0.04), as shown in Table 3.

When stratified by baseline drug use, the proportion of participants with HIV UVL was higher among those with a negative urinalysis compared to those with a positive urinalysis at baseline. Among those with a positive urinalysis at baseline, individuals randomized to the TWM intervention were less likely only at the 11-month assessment to have an HIV UVL compared to those randomized to the control group in both unadjusted (DD=–26.8, 95% CI –49.9 to –3.6) and adjusted (DD=–27.2, 95% CI –52.5 to –1.9) models, as shown in Table 3. However, this difference was not found at any other time point for those with a positive urinalysis or at any time point for participants with a negative baseline urinalysis.

XSL•FO
RenderX

Table 2. The effect of Thrive with Me (TWM) on HIV-1 undetectable viral load (UVL) for the full sample and by study arm.

	Baseline UVL, estimate (95% CI)	Month 5 UVL, estimate (95% CI)	Month 11 UVL, estimate (95% CI)	Month 17 UVL, estimate (95% CI)
TWM intervention (n=202)				
Full sample				
Unadjusted	63.1 (56.4-59.7)	57.9 (50.7-65.0)	54.2 (46.9-61.6)	52.9 (45.2-60.6)
Adjusted	66.3 (59.8-72.7)	59.6 (52.4-66.8)	55.9 (48.5-63.3)	54.5 (46.8-62.2)
Positive drug urinalysis				
Unadjusted	48.5 (36.2-60.7)	42.1 (29.3-54.8)	30.5 (17.5-43.5)	34.2 (21.0-47.3)
Adjusted	51.1 (39.0-63.2)	44.8 (31.6-58.1)	31.0 (17.5-44.6)	35.4 (21.6-49.2)
Negative drug urinalysis				
Unadjusted	70.1 (62.4-77.7)	66.4 (58.2-74.6)	65.0 (56.2-73.8)	61.6 (52.8-70.5)
Adjusted	73.1 (65.7-80.4)	67.4 (59.4-75.5)	66.8 (58.2-75.3)	62.7 (54.0-71.5)
Control group (n=199)				
Full sample				
Unadjusted	59.8 (53.0-66.6)	60.9 (53.9-67.9)	63.9 (56.5-71.2)	56.8 (49.2-64.5)
Adjusted	58.1 (51.6-64.7)	59.3 (52.6-65.9)	61.6 (54.6-68.7)	54.5 (47.1-62.0)
Positive drug urinalysis				
Unadjusted	47.9 (33.8-62.0)	54.1 (38.9-69.3)	56.7 (42.2-71.1)	46.7 (30.5-62.8)
Adjusted	46.0 (32.9-59.1)	51.5 (37.8-65.2)	53.1 (39.8-66.4)	42.2 (27.4-57.0)
Negative drug urinalysis				
Unadjusted	63.6 (55.9-71.3)	62.6 (54.7-70.6)	66.6 (58.3-74.8)	60.4 (51.6-69.2)
Adjusted	62.1 (54.6-69.6)	61.6 (53.8-69.3)	64.9 (56.9-72.9)	58.9 (50.3-67.6)

Table 3. Difference-in-differences estimates of the effect of Thrive with Me (TWM) on HIV-1 undetectable viral load (UVL) for the full sample stratified by positive and negative drug urinalysis.

Difference-in-differences estimates[a]	Baseline to month 5, estimate (95% CI)	Baseline to month 11, estimate (95% CI)	Baseline to month 17, estimate (95% CI)
Full sample (N=401)			
Unadjusted	–6.4 (–20.1 to 7.4)	–12.9 (–27.1 to 1.4)	–7.2 (–21.5 to 7.0)
Adjusted	–7.8 (–21.1 to 5.5)	–13.9 (–27.7 to –0.04)	–8.2 (–22.0 to 5.7)
Positive drug urinalysis			
Unadjusted	–12.5 (–34.7 to 9.6)	–26.8 (–49.9 to –3.6)	–13.1 (–38.5 to 12.4)
Adjusted	–11.7 (–37.8 to 14.3)	–27.2 (–52.5 to –1.9)	–11.9 (–39.0 to 15.2)
Negative drug urinalysis			
Unadjusted	–2.8 (–18.5 to 13.0)	–8.1 (24.4 to 8.3)	–5.3 (–21.6 to 11.1)
Adjusted	–5.1 (–20.4 to 10.2)	–9.1 (–24.9 to 6.8)	–7.1 (–23.1 to 8.8)

[a]The between–study arm difference calculated from the difference between each follow-up assessment and baseline within study arms.

The Effect of Engagement With TWM on HIV VL for Participants Randomized to the TWM Intervention

Of the 202 participants who were randomized to receive the TWM intervention, 110 (54.5%) accessed the TWM intervention at least once during the 5-month intervention (ie, engagers), whereas 92 (45.5%) did not use any of the TWM intervention components (ie, nonengagers). There were no demographic differences between TWM engagers and nonengagers. However, participants who used the TWM intervention were less likely to have a detectable VL (32/110, 29.1%) at baseline compared to those who did not use the intervention (42/92, 46%). TWM engagers were also less likely to have a positive urinalysis for methamphetamine (10/110, 9.1% vs 21/92, 23%) and amphetamines (6/110, 5.5% vs 19/92, 21%) compared to nonengagers.

XSL•FO
RenderX

Table 4 shows the median frequency and IQR for overall engagement and engagement with each of the 3 individual components (frequency of asynchronous peer exchanges, number of unique Thrive Tips viewed, and daily ART self-monitoring) by whether participants were considered low engagers (75th percentile) or high engagers (25th percentile) in each of the variables. With the exception of daily ART self-monitoring, substantial differences in engagement were found between low and high engagers. For example, low engagers had an average of only 7 active days in the intervention, whereas high engagers had an average of 78 active days. While differences in daily ART self-monitoring were evident among low and high engagers (100 days for low engagers vs 144 days for high engagers), they were not as stark as with other features.

Table 4. Frequency of Thrive with Me (TWM) user engagement among participants randomized to TWM and who accessed it at least once.

TWM user engagement	TWM-engaged participants (n=110), median (IQR)	Low engagers[a], median (IQR)	High engagers[a], median (IQR)
Overall engagement—number of active days	16 (5-33)	9 (4-20.5)[b]	77.5 (40-115)[c]
Number of peer exchanges[d]	7 (2-34)	4 (2-9)[e]	135 (58-195)[f]
Wall posts	3 (1-13)	2 (1-5)[e]	34.5 (20-68)[f]
Comments	3 (0-12)	1 (0-5)[e]	91 (29-142)[f]
Number of unique Thrive Tips accessed	6 (2-32)	4 (2-10)[g]	121 (78-168)[h]
Number of ART[i] self-monitoring days	117 (51-137)	100 (41-124)[j]	144.5 (142-147)[c]

[a]Engagement dichotomized at the 75th percentile for each user engagement variable.
[b]n=84.
[c]n=26.
[d]Excluding 1 individual.
[e]n=83.
[f]n=27.
[g]n=75.
[h]n=25.
[i]ART: antiretroviral therapy.
[j]n=79.

Multimedia Appendix 2 shows the effect of overall engagement with TWM on HIV UVL, as well as the effect of engagement with the individual components of TWM on HIV UVL, at each assessment time point. At month 5 (the end of the intervention period), 73% (19/26) of TWM users with overall high engagement (ie, were in the 25th percentile of active intervention use days) were virally suppressed compared to 54% (45/84) of those categorized as having low engagement and 40% (37/92) of nonengagers. At month 5, high engagers were more likely to be virally suppressed compared to nonengagers in both the unadjusted (RD=31.9, 95% CI 12.6-51.3) and adjusted (RD=19.4, 95% CI 3.3-35.5) models. Similarly, high engagers were more likely to be virally suppressed compared to low engagers in the unadjusted (RD=23.4, 95% CI 4.3-42.4) and adjusted (RD=17.8, 95% CI 2.5-33.0) models (not shown in the table). However, no relationship was found between engagement categories and viral suppression at months 11 or 17.

Notably, 70% (19/27) of those categorized as having high engagement with peer exchanges (ie, were in the 25th percentile) were virally suppressed at month 5 compared to 54% (45/83) of those with low engagement and 40% (37/92) of nonengagers, as shown in Multimedia Appendix 2. High engagers with the peer exchange component were more likely to be virally suppressed than participants who did not engage with the peer exchanges at all at months 5 (RD=25.3, 95% CI 4.9-45.7), 11 (RD=22.5, 95% CI 1.1-44.0), and 17 (RD=23.4, 95% CI 2.2-44.7) in unadjusted models, although these effects were not found in adjusted models. We found no differences in viral suppression between engagement categories when examining the Thrive Tips or ART self-monitoring components at any assessment time point.

Discussion

Principal Findings

We did not find that the 5-month TWM web-based intervention delivered to a community-recruited sample of sexual minority men living with HIV overall improved their VL outcomes over 17 months, nor did we find that the intervention showed greater impact for participants with recent drug use at study entry. However, overall high engagement with the intervention compared to no engagement or low engagement among participants assigned to the TWM intervention arm was associated with having an HIV UVL at month 5, which was the active intervention use period (after which access to the intervention ended).

Behavioral ART adherence interventions have had mixed success in modifying adherence behaviors and reducing VL. In a meta-analysis of interventions cotargeting one or more syndemic factors (eg, mental health and substance use) and HIV

XSL•FO
RenderX

prevention or treatment outcomes, significant effects were found for sexual risk reduction interventions but not for ART adherence interventions [46]. SMS text messaging and voice (eg, telephone-based) interventions have shown some success for improving adherence or viral suppression in lower-income countries [47]. However, mobile apps and computerized interventions most commonly used in higher-income countries, with the exception of a computerized intervention by Kurth et al [24], either lack sufficient power to be conclusive [47] or have not yet demonstrated effects on plasma HIV VL [25-29].

An important factor that may drive the degree to which digital ART adherence interventions are impactful is whether they are embedded in other aspects of HIV clinical care. TWM was designed and tested as a stand-alone ART adherence intervention that could supplement the HIV clinical care of sexual minority men living with HIV. In contrast, the computerized ART adherence intervention by Kurth et al [24] was conducted in 4 sessions during regular HIV clinic appointments and demonstrated positive impacts on HIV VL [24]. This arrangement may have provided a more successful context to engage participants with the intervention than mobile apps or computerized interventions that rely on participants engaging with them outside the clinic context. However, providing technology-based interventions in the clinic may diminish the potential benefits of these types of interventions, including lower implementation costs (ie, requiring less staff time), circumventing clinic-based stigma, and direct-to-consumer availability [21,48,49]. A better understanding of the trade-offs of the venues where technology-based interventions are deployed and the accompanying staffing requirements is needed.

While results did not support that the overall sample benefited from the intervention, those TWM-assigned participants with high levels of engagement (in the 25th percentile) were more likely to have an HIV UVL at the end of the 5-month active intervention period compared to those with low engagement (below the 75th percentile) or no engagement at all with the intervention. This finding is consistent with those of an ART adherence gaming intervention to improve rates of viral suppression among youth aged 16 to 24 years living with HIV called Epic Allies [29]. In that study, participants were enrolled primarily in HIV clinics that served a large number of youths, but engagement with the app occurred outside the clinic over 26 weeks. While no intervention effects on HIV VL were shown in an intention-to-treat analysis, youths who used the mobile app ≥4 days a week were 56% more likely to achieve viral suppression compared to youths who used the app less often at the 3-month assessment; however, those effects dissipated in later follow-up assessments. This and other studies showing that the impacts of mHealth interventions dissipate over time [50] suggest that optimizing technology-based ART adherence interventions will require robust management of participants to get and keep them engaged with the intervention.

Little is known about the impact of individual components of eHealth and mHealth interventions on HIV VL even though having a better understanding of which components may be the most impactful is needed [31]. A previous study by Bonett et al [51] found that engagement (measured as total time spent) with different features of the mHealth intervention among HIV-negative youth was associated with sociodemographic and behavioral factors (eg, participants who spent more time using the HIV and sexually transmitted infection testing site locator were more likely to be single). In the TWM arm of our study, we found high proportions of HIV UVL among users with high engagement with the peer exchange feature (both posting and commenting) in the unadjusted models at all follow-up time points, although not when adjusted for baseline viral suppression and positive drug urinalysis. Among high-engagement participants, all of them posted at least one wall comment, and 75% (20/27) of them responded to another participant's post. The content of peer exchanges on social platforms focused on health promotion among people living with HIV may play a critical role. A previous qualitative analysis of the peer exchanges in TWM showed that social support was the most common theme from the analysis of peer exchanges, with half of social support exchanges coded as seeking social support and half as providing social support [52]. Peer exchanges also included issues regarding HIV treatment and care, such as challenges taking ART, adherence strategies, discussing HIV treatment including laboratory tests or side effects, and participants' relationships with health care providers [52]. A recent study by Bauermeister et al [53] found that young (aged 18-30 years) Black sexual minority men who discussed experiences of stigma in a peer exchange forum during an mHealth intervention trial showed decreased anticipated stigma over time, whereas those whose discussions focused on sexuality-related stigma reported increased levels of internalized homophobia and sexual prejudice. Together, the results of the latter study and our TWM study suggest that a greater understanding of how peer exchanges may impact HIV care outcomes and other dimensions of well-being is warranted.

Study Limitations

This study has important limitations that contextualize the findings. First, approximately 39% of the participants (154/401, 38.5%) entered the study with detectable HIV VL, which may have hindered our ability to demonstrate effects on VL as many participants started the study virally suppressed. As with many ART adherence studies, we balanced our ability to recruit our sample, who were required to self-report adherence problems, with more rigorous study inclusion (eg, detectable VL). Nonetheless, our findings suggest potential positive impact among those using the TWM intervention. Second, participants were recruited in the New York City area, which may have better technology infrastructure, higher technology adoption rates, and other unique considerations that differ from those in other regions of the United States. Therefore, the results may not generalize to other areas of the United States. Third, while our findings suggest a possible benefit of TWM for high-engaging participants, only approximately half (110/202, 54.5%) of those randomized to the TWM intervention actually engaged with it. We only incentivized intervention use through gift card draws, and therefore, this approach may mimic "real-world" use of this type of computerized intervention in the context of deploying minimal staffing to engage participants. Finally, improvement in VL may be due to an exogenous factor unrelated to study participation for men who chose to access

XSL•FO
RenderX

TWM. For this reason, caution should be used when interpreting study findings for this group of men.

Conclusions

Keeping limitations in mind, we believe that the results fill a critical gap in eHealth and mHealth ART adherence intervention science by adding to evidence that these types of interventions may need to be developed alongside in-person or internet-based support to have the greatest impacts on HIV VL. Using stepped-care models [54] to identify which participants need more intensive support holds promise in this area. Aside from using different intervention designs to meet the unique needs of users, processes for engaging users early and sustaining that engagement over time continue to be a high priority to fully realize the potential of technology-based ART adherence interventions. Finally, efforts should be made to expand our understanding of how peer exchange forum use may have positive impacts across a range of clinical and mental health domains.

Acknowledgments

The authors would like to thank the participants for the time they offered during the conduct of this study. This work was supported by a grant from the National Institute on Drug Abuse (R01DA039950). The content is solely the responsibility of the authors and does not represent the official views of the funding agencies.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

KJH contributed to the conceptualization, funding acquisition, investigation, methodology, project administration, supervision, and writing the original draft. SL contributed to the data curation, formal analysis, and writing the original draft. DE contributed to the conceptualization, investigation, methodology, formal analysis, and review and editing of the manuscript. KRA contributed to the conceptualization, investigation, methodology, and review and editing of the manuscript. AJT contributed to the project administration, resources, supervision, and review and editing of the manuscript. OS contributed to the project administration, resources, supervision, and review and editing of the manuscript. CJS contributed to the review and editing of the manuscript. HJR contributed to the project administration, resources, supervision, and review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example newsletter received by participants randomized to the control group.
[PDF File (Adobe PDF File), 3465 KB - jmir_v26i1e53819_app1.pdf]

Multimedia Appendix 2
The impact user engagement in the Thrive with Me intervention on HIV undetectable viral load.
[PDF File (Adobe PDF File), 167 KB - jmir_v26i1e53819_app2.pdf]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 266 KB - jmir_v26i1e53819_app3.pdf]

References

1. HIV surveillance report, 2020. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html [accessed 2024-04-29]
2. Key strategies in the plan. HIV.gov. URL: https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/key-strategies [accessed 2024-04-29]
3. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Antiretroviral therapy for the prevention of HIV-1 transmission. N Engl J Med 2016 Sep;375(9):830-839. [doi: 10.1056/nejmoa1600693]
4. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, HPTN 052 Study Team. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 2011 Aug 11;365(6):493-505 [FREE Full text] [doi: 10.1056/NEJMoa1105243] [Medline: 21767103]
5. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, van Lunzen J, PARTNER Study Group. Sexual activity without condoms and risk of HIV transmission in serodifferent couples when the HIV-positive partner is using suppressive

XSL•FO
RenderX

antiretroviral therapy. JAMA 2016 Jul 12;316(2):171-181 [FREE Full text] [doi: 10.1001/jama.2016.5148] [Medline: 27404185]
6. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data: United States and 6 dependent areas, 2020. Centers for Disease Control and Prevention. URL: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html [accessed 2024-04-29]
7. Mayer KH, Nelson L, Hightow-Weidman L, Mimiaga MJ, Mena L, Reisner S, et al. The persistent and evolving HIV epidemic in American men who have sex with men. Lancet 2021 Mar 20;397(10279):1116-1126 [FREE Full text] [doi: 10.1016/S0140-6736(21)00321-4] [Medline: 33617771]
8. Colfax G, Shoptaw S, Colfax G, Shoptaw S. The methamphetamine epidemic: implications for HIV prevention and treatment. Curr HIV/AIDS Rep 2005 Dec;2(4):194-199. [doi: 10.1007/S11904-005-0016-4]
9. Plankey MW, Ostrow DG, Stall R, Cox C, Li X, Peck JA, et al. The relationship between methamphetamine and popper use and risk of HIV seroconversion in the multicenter AIDS cohort study. J Acquir Immune Defic Syndr 2007 May 01;45(1):85-92 [FREE Full text] [doi: 10.1097/QAI.0b013e3180417c99] [Medline: 17325605]
10. Carey JW, Mejia R, Bingham T, Ciesielski C, Gelaude D, Herbst JH, et al. Drug use, high-risk sex behaviors, and increased risk for recent HIV infection among men who have sex with men in Chicago and Los Angeles. AIDS Behav 2009 Dec 23;13(6):1084-1096. [doi: 10.1007/s10461-008-9403-3] [Medline: 18498049]
11. Carrico AW, Hunt PW, Neilands TB, Dilworth SE, Martin JN, Deeks SG, et al. Stimulant use and viral suppression in the era of universal antiretroviral therapy. J Acquir Immune Defic Syndr 2019 Jan 01;80(1):89-93 [FREE Full text] [doi: 10.1097/QAI.0000000000001867] [Medline: 30272634]
12. Carrico AW, Riley ED, Johnson MO, Charlebois ED, Neilands TB, Remien RH, et al. Psychiatric risk factors for HIV disease progression: the role of inconsistent patterns of antiretroviral therapy utilization. J Acquir Immune Defic Syndr 2011 Feb 01;56(2):146-150 [FREE Full text] [doi: 10.1097/QAI.0b013e318201df63] [Medline: 21116186]
13. Ellis RJ, Childers ME, Cherner M, Lazzaretto D, Letendre S, Grant I, HIV Neurobehavioral Research Center Group. Increased human immunodeficiency virus loads in active methamphetamine users are explained by reduced effectiveness of antiretroviral therapy. J Infect Dis 2003 Dec 15;188(12):1820-1826. [doi: 10.1086/379894] [Medline: 14673760]
14. Horvath KJ, Carrico AW, Simoni J, Boyer EW, Amico KR, Petroll AE. Engagement in HIV medical care and technology use among stimulant-using and nonstimulant-using men who have sex with men. AIDS Res Treat 2013;2013:121352-121311 [FREE Full text] [doi: 10.1155/2013/121352] [Medline: 23864944]
15. Carrico AW. Substance use and HIV disease progression in the HAART era: implications for the primary prevention of HIV. Life Sci 2011 May 23;88(21-22):940-947. [doi: 10.1016/j.lfs.2010.10.002] [Medline: 20934437]
16. Carrico AW, Shoptaw S, Cox C, Stall R, Li X, Ostrow DG, et al. Stimulant use and progression to AIDS or mortality after the initiation of highly active antiretroviral therapy. J Acquir Immune Defic Syndr 2014 Dec 15;67(5):508-513 [FREE Full text] [doi: 10.1097/QAI.0000000000000364] [Medline: 25271387]
17. Adams JW, Bryant KJ, Edelman JE, Fiellin DA, Gaither JR, Gordon AJ, et al. Association of cannabis, stimulant, and alcohol use with mortality prognosis among HIV-infected men. AIDS Behav 2018 Apr 8;22(4):1341-1351 [FREE Full text] [doi: 10.1007/s10461-017-1905-4] [Medline: 28887669]
18. Mayer KH, Skeer MR, O'Cleirigh C, Goshe BM, Safren SA. Factors associated with amplified HIV transmission behavior among American men who have sex with men engaged in care: implications for clinical providers. Ann Behav Med 2014 Apr 20;47(2):165-171 [FREE Full text] [doi: 10.1007/s12160-013-9527-1] [Medline: 23873338]
19. Johnson MO, Carrico AW, Chesney MA, Morin SF. Internalized heterosexism among HIV-positive, gay-identified men: implications for HIV prevention and care. J Consult Clin Psychol 2008 Oct;76(5):829-839 [FREE Full text] [doi: 10.1037/0022-006X.76.5.829] [Medline: 18837600]
20. Morin SF, Myers JJ, Shade SB, Koester K, Maiorana A, Rose CD. Predicting HIV transmission risk among HIV-infected patients seen in clinical settings. AIDS Behav 2007 Sep 19;11(5 Suppl):S6-16 [FREE Full text] [doi: 10.1007/s10461-007-9253-4] [Medline: 17577655]
21. Nelson KM, Perry NS, Horvath KJ, Smith LR. A systematic review of mHealth interventions for HIV prevention and treatment among gay, bisexual, and other men who have sex with men. Transl Behav Med 2020 Oct 12;10(5):1211-1220 [FREE Full text] [doi: 10.1093/tbm/ibaa007] [Medline: 33044531]
22. Muessig KE, LeGrand S, Horvath KJ, Bauermeister JA, Hightow-Weidman LB. Recent mobile health interventions to support medication adherence among HIV-positive MSM. Curr Opin HIV AIDS 2017 Sep;12(5):432-441 [FREE Full text] [doi: 10.1097/COH.0000000000000401] [Medline: 28639990]
23. Horvath KJ, Walker T, Mireles L, Bauermeister JA, Hightow-Weidman L, Stephenson R. A systematic review of technology-assisted HIV testing interventions. Curr HIV/AIDS Rep 2020 Aug 08;17(4):269-280 [FREE Full text] [doi: 10.1007/s11904-020-00506-1] [Medline: 32507984]
24. Kurth AE, Spielberg F, Cleland CM, Lambdin B, Bangsberg DR, Frick PA, et al. Computerized counseling reduces HIV-1 viral load and sexual transmission risk: findings from a randomized controlled trial. J Acquir Immune Defic Syndr 2014 Apr 15;65(5):611-620 [FREE Full text] [doi: 10.1097/QAI.0000000000000100] [Medline: 24384803]

XSL•FO
RenderX

25. Schnall R, Sanabria G, Jia H, Cho H, Bushover B, Reynolds N, et al. Efficacy of an mHealth self-management intervention for persons living with HIV: the WiseApp randomized clinical trial. J Am Med Inform Assoc 2023 Feb 16;30(3):418-426 [FREE Full text] [doi: 10.1093/jamia/ocac233] [Medline: 36469808]
26. Amico KR, Lindsey JC, Hudgens M, Dallas R, Horvath KJ, Dunlap A, TERA Study Team. Randomized controlled trial of a remote coaching mHealth adherence intervention in youth living with HIV. AIDS Behav 2022 Dec 07;26(12):3897-3913 [FREE Full text] [doi: 10.1007/s10461-022-03717-2] [Medline: 35670987]
27. Garofalo R, Kuhns LM, Hotton A, Johnson A, Muldoon A, Rice D. A randomized controlled trial of personalized text message reminders to promote medication adherence among HIV-positive adolescents and young adults. AIDS Behav 2016 May 11;20(5):1049-1059 [FREE Full text] [doi: 10.1007/s10461-015-1192-x] [Medline: 26362167]
28. Lewis MA, Harshbarger C, Bann C, Marconi VC, Somboonwit C, Piazza MD, Positive Health Check Study Team. Effectiveness of an interactive, highly tailored "video doctor" intervention to suppress viral load and retain patients with HIV in clinical care: a randomized clinical trial. J Acquir Immune Defic Syndr 2022 Sep 01;91(1):58-67 [FREE Full text] [doi: 10.1097/QAI.0000000000003045] [Medline: 35972854]
29. Hightow-Weidman L, Muessig KE, Egger JR, Vecchio A, Platt A. Epic allies: a gamified mobile app to improve engagement in HIV care and antiretroviral adherence among young men who have sex with men. AIDS Behav 2021 Aug 19;25(8):2599-2617 [FREE Full text] [doi: 10.1007/s10461-021-03222-y] [Medline: 33740213]
30. Choi SK, Muessig KE, Hightow-Weidman LB, Bauermeister JA. Paradata: measuring engagement in digital HIV interventions for sexual and gender minorities. Curr HIV/AIDS Rep 2023 Dec;20(6):487-501. [doi: 10.1007/s11904-023-00679-5] [Medline: 37930613]
31. Bauermeister JA, Golinkoff JM, Muessig KE, Horvath KJ, Hightow-Weidman LB. Addressing engagement in technology-based behavioural HIV interventions through paradata metrics. Curr Opin HIV AIDS 2017 Sep;12(5):442-446 [FREE Full text] [doi: 10.1097/COH.0000000000000396] [Medline: 28617711]
32. Hightow-Weidman LB, Bauermeister JA. Engagement in mHealth behavioral interventions for HIV prevention and care: making sense of the metrics. Mhealth 2020;6:7 [FREE Full text] [doi: 10.21037/mhealth.2019.10.01] [Medline: 32190618]
33. Horvath KJ, Amico KR, Erickson D, Ecklund AM, Martinka A, DeWitt J, et al. Thrive with me: protocol for a randomized controlled trial to test a peer support intervention to improve antiretroviral therapy adherence among men who have sex with men. JMIR Res Protoc 2018 May 31;7(5):e10182 [FREE Full text] [doi: 10.2196/10182] [Medline: 29853437]
34. Horvath KJ, Oakes JM, Rosser BR, Danilenko G, Vezina H, Amico KR, et al. Feasibility, acceptability and preliminary efficacy of an online peer-to-peer social support ART adherence intervention. AIDS Behav 2013 Jul 4;17(6):2031-2044 [FREE Full text] [doi: 10.1007/s10461-013-0469-1] [Medline: 23553347]
35. Fisher JD, Fisher WA, Bryan AD, Misovich SJ. Information-motivation-behavioral skills model-based HIV risk behavior change intervention for inner-city high school youth. Health Psychol 2002;21(2):177-186. [doi: 10.1037//0278-6133.21.2.177]
36. Fisher WA, Fisher JD, Harman JJ. The information-motivation-behavioral skills model: a general social psychological approach to understanding and promoting health behavior. In: Suls J, Wallston KA, editors. Social Psychological Foundations of Health and Illness. New York, NY: Wiley-Blackwell; 2003:82-106.
37. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 2016 Jul 26;1(3):385-401. [doi: 10.1177/014662167700100306]
38. Amico KR, Barta W, Konkle-Parker DJ, Fisher JD, Cornman DH, Shuper PA, et al. The information-motivation-behavioral skills model of ART adherence in a Deep South HIV+ clinic sample. AIDS Behav 2009 Feb;13(1):66-75 [FREE Full text] [doi: 10.1007/s10461-007-9311-y] [Medline: 17876697]
39. Earnshaw VA, Smith LR, Chaudoir SR, Amico KR, Copenhaver MM. HIV stigma mechanisms and well-being among PLWH: a test of the HIV stigma framework. AIDS Behav 2013 Jun;17(5):1785-1795 [FREE Full text] [doi: 10.1007/s10461-013-0437-9] [Medline: 23456594]
40. Sherbourne CD, Stewart AL. The MOS social support survey. Soc Sci Med 1991 Jan;32(6):705-714. [doi: 10.1016/0277-9536(91)90150-b] [Medline: 2035047]
41. Wong MD, Sarkisian CA, Davis C, Kinsler J, Cunningham WE. The association between life chaos, health care use, and health status among HIV-infected persons. J Gen Intern Med 2007 Sep 28;22(9):1286-1291 [FREE Full text] [doi: 10.1007/s11606-007-0265-6] [Medline: 17597350]
42. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
43. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction 1993 Jun 24;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
44. Lechner M. The estimation of causal effects by difference-in-difference methods. Found Trends Econom 2011;4(3):165-224 [FREE Full text]
45. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
46. Pantalone DW, Nelson KM, Batchelder AW, Chiu C, Gunn HA, Horvath KJ. A systematic review and meta-analysis of combination behavioral interventions co-targeting psychosocial syndemics and HIV-related health behaviors for sexual

XSL•FO
RenderX

minority men. J Sex Res 2020 Jul 20;57(6):681-708 [FREE Full text] [doi: 10.1080/00224499.2020.1728514] [Medline: 32077326]
47. Lee SB, Valerius J. mHealth interventions to promote anti-retroviral adherence in HIV: narrative review. JMIR Mhealth Uhealth 2020 Aug 28;8(8):e14739 [FREE Full text] [doi: 10.2196/14739] [Medline: 32568720]
48. Muessig KE, Nekkanti M, Bauermeister J, Bull S, Hightow-Weidman LB. A systematic review of recent smartphone, internet and web 2.0 interventions to address the HIV continuum of care. Curr HIV/AIDS Rep 2015 Mar 28;12(1):173-190 [FREE Full text] [doi: 10.1007/s11904-014-0239-3] [Medline: 25626718]
49. Sullivan PS, Hightow-Weidman L. Mobile apps for HIV prevention: how do they contribute to our epidemic response for adolescents and young adults? Mhealth 2021 Apr;7:36 [FREE Full text] [doi: 10.21037/mhealth-20-71] [Medline: 33898605]
50. Horvath KJ, Lammert S, MacLehose RF, Danh T, Baker JV, Carrico AW. A pilot study of a mobile app to support HIV antiretroviral therapy adherence among men who have sex with men who use stimulants. AIDS Behav 2019 Nov 15;23(11):3184-3198 [FREE Full text] [doi: 10.1007/s10461-019-02597-3] [Medline: 31309348]
51. Bonett S, Connochie D, Golinkoff JM, Horvath KJ, Bauermeister JA. Paradata analysis of an eHealth HIV testing intervention for young men who have sex with men. AIDS Educ Prev 2018 Oct;30(5):434-447 [FREE Full text] [doi: 10.1521/aeap.2018.30.5.434] [Medline: 30332307]
52. Sun CJ, Shato T, Steinbaugh A, Pradeep S, Rivet Amico K, Horvath K. Virtual voices: examining social support exchanged through participant-generated and unmoderated content in a mobile intervention to improve HIV antiretroviral therapy adherence among GBMSM. AIDS Care 2023 Jan 13;35(1):7-15 [FREE Full text] [doi: 10.1080/09540121.2022.2038364] [Medline: 35156485]
53. Bauermeister JA, Muessig KE, LeGrand S, Flores DD, Choi SK, Dong W, et al. HIV and sexuality stigma reduction through engagement in online forums: results from the HealthMPowerment intervention. AIDS Behav 2019 Mar 18;23(3):742-752 [FREE Full text] [doi: 10.1007/s10461-018-2256-5] [Medline: 30121727]
54. Reback CJ, Rusow JA, Cain D, Benkeser D, Arayasirikul S, Hightow-Weidman L, et al. Technology-based stepped care to stem transgender adolescent risk transmission: protocol for a randomized controlled trial (TechStep). JMIR Res Protoc 2020 Aug 13;9(8):e18326 [FREE Full text] [doi: 10.2196/18326] [Medline: 32788149]

Abbreviations

ART: antiretroviral therapy
CONSORT: Consolidated Standards of Reporting Trials
DD: difference-in-differences
IMB: Information-Motivation-Behavioral Skills
mHealth: mobile health
RD: risk difference
THC: tetrahydrocannabinol
TWM: Thrive with Me
UVL: undetectable viral load
VL: viral load

Edited by A Mavragani; submitted 25.10.23; peer-reviewed by K Ingersoll, SK Choi; comments to author 20.03.24; revised version received 27.06.24; accepted 24.07.24; published 30.09.24.

Please cite as:
Horvath KJ, Lammert S, Erickson D, Amico KR, Talan AJ, Shalhav O, Sun CJ, Rendina HJ
A Web-Based Antiretroviral Therapy Adherence Intervention (Thrive With Me) in a Community-Recruited Sample of Sexual Minority Men Living With HIV: Results of a Randomized Controlled Study
J Med Internet Res 2024;26:e53819
URL: https://www.jmir.org/2024/1/e53819
doi:10.2196/53819
PMID:

©Keith J Horvath, Sara Lammert, Darin Erickson, K Rivet Amico, Ali J Talan, Ore Shalhav, Christina J Sun, H Jonathon Rendina. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of WeChat-Based Digital Intervention Versus Metformin in Women With Polycystic Ovary Syndrome: Randomized Controlled Trial

Diliqingna Dilimulati[1*], PhD; Xiaowen Shao[2*], PhD; Lihua Wang[3*], PhD; Meili Cai[1], PhD; Yuqin Zhang[1], PhD; Jiayi Lu[4], MD; Yao Wang[5], PhD; Hongying Liu[6], PhD; Ming Kuang[6], MSc; Haibing Chen[1], PhD; Manna Zhang[1*], PhD; Shen Qu[1], PhD

[1]Department of Endocrinology and Metabolism, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China
[2]Department of Obstetrics and Gynecology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China
[3]Department of Endocrinology and Metabolism, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China
[4]Department of Clinical Laboratory, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China
[5]Department of Dermatology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China
[6]Hangzhou Kang Ming Information Technology Co., Ltd, Hangzhou, China
[*]these authors contributed equally

Corresponding Author:
Manna Zhang, PhD
Department of Endocrinology and Metabolism
Shanghai Tenth People's Hospital
Tongji University School of Medicine
No. 301 Middle Yanchang Road
Shanghai, 200072
China
Phone: 86 13774448495
Fax: 86 021 66301004
Email: mannazhang@126.com

Abstract

Background: The first-line treatment for polycystic ovary syndrome (PCOS) is lifestyle modification. However, it is currently unknown whether digital medicine can assist patients with PCOS in maintaining a healthy lifestyle while alleviating PCOS symptoms.

Objective: This study aims to evaluate the efficacy of WeChat-based digital intervention versus metformin treatment in women with PCOS and insulin resistance.

Methods: A total of 80 women with PCOS and insulin resistance were recruited from an endocrinology clinic and randomly assigned to receive either a WeChat-based digital intervention (n=40, 50%) or metformin (n=40, 50%) for 12 weeks. The WeChat-based digital intervention consisted of 3 modules; a coach assisted the patients in using the intervention. The primary outcome was the change in a homeostatic model assessment for insulin resistance. At baseline and after the 12-week intervention, anthropometric parameters, menstruation frequency, sex hormone levels, metabolic factors, and body fat distribution were measured in the clinic. Furthermore, self-assessed web-based questionnaires on diet, exercise, sleep, anxiety, and depression were obtained.

Results: A total of 72 participants completed the follow-up (for a 90% follow-up rate), including 35 of 40 (88%) participants from the digital intervention group and 37 of 40 (93%) participants from the metformin group. The homeostatic model assessment for insulin resistance in the digital intervention group was significantly improved after 12 weeks of treatment with a mean change of –0.93 (95% CI –1.64 to –0.23), but no statistical difference was observed between the groups (least squares mean difference –0.20; 95% CI –0.98 to 0.58; P=.62). Both digital intervention and metformin treatment significantly improved menstruation frequency (digital intervention: P<.001; metformin: P<.001) and reduced body weight (digital intervention: P<.001; metformin: P<.001) and total fat mass (digital intervention: P<.001; metformin: P<.001). Furthermore, the digital intervention had a significant advantage over metformin in improving waist circumference (least squares mean difference –1.84; 95% CI –3.44 to –0.24; P=.03),

XSL•FO
RenderX

waist-to-hip ratio (least squares mean difference –0.02; 95% CI –0.03 to 0.00; *P*=.03), total fat mass (least squares mean difference –1.59; 95% CI –2.88 to –0.30; *P*=.02), and dehydroepiandrosterone sulfate (least squares mean difference –69.73; 95% CI –129.70 to –9.75; *P*=.02). In terms of safety, the main adverse events were sensations of hunger in the digital intervention group (2/40, 5%) and gastrointestinal adverse events in the metformin group (12/40, 30%).

Conclusions: Our data suggest that digital intervention is an effective treatment option for patients with PCOS, with an efficacy comparable to that of metformin, and that it can also alleviate the negative effects of medications and make it easier and more efficient to adhere to lifestyle treatments. WeChat-based digital interventions have the potential to provide a new path for the improvement and health of women with PCOS in China.

Trial Registration: ClinicalTrials.gov NCT05386706; https://clinicaltrials.gov/study/NCT05386706

(J Med Internet Res 2024;26:e55883) doi:10.2196/55883

KEYWORDS

polycystic ovary syndrome; insulin resistance; digital intervention; metformin; women's health

Introduction

Background

Approximately 10% to 13% of women of reproductive age worldwide have polycystic ovary syndrome (PCOS) [1]. The prevalence of PCOS is increasing globally because of changes in environmental variables and lifestyle, as well as increased disease awareness and detection rates. PCOS is characterized by a range of heterogeneous symptoms with varying severities, including abnormal menstruation, hyperandrogenism (hirsutism or hyperandrogenemia), and ovarian dysfunction (oligo-ovulation or polycystic ovary morphology) [2]. One of the most prevalent endocrine disorders among patients with PCOS is insulin resistance, with variable degrees of insulin resistance present in approximately 75% of patients with PCOS [3]. For such patients, insulin resistance reduction is essential in the management of various endocrine and metabolic disorders [4].

The first-line treatment for PCOS is lifestyle intervention, and studies have demonstrated that positive lifestyle changes can improve hyperandrogenemia, menstrual cycle, hirsutism, and insulin resistance in the affected patients [5]. However, lifestyle intervention is associated with poor compliance and low sustainability. Nevertheless, there are no indicators to determine whether patients are making enough progress in improving their lifestyle. Studies have demonstrated that digital therapy accessed on mobile phones, which are convenient and accessible, can effectively improve behavioral changes in diet, exercise, and medication adherence while predicting disease progression, reducing the frequency of disease-related symptoms, and promoting effective disease management [6,7]. Regarding PCOS, which is a chronic disease, there are limited digital treatment methods. A previous 3-component lifestyle intervention consisting of diet, exercise, and cognitive behavioral therapy, facilitated using texting to track patients' food intake, physical activity, and mood, can effectively improve eating disorder and mood in patients with PCOS and contribute to weight loss [8,9]. This study focused on the effect of digital intervention compared with metformin in improving homeostatic model assessment for insulin resistance (HOMA-IR) levels and other metabolic and reproductive indicators.

Objectives

This was a single-center, prospective, randomized controlled clinical trial designed to determine whether digital intervention was effective in patients with PCOS after 12 weeks of treatment. Evaluations were conducted on improvements in PCOS-related clinical parameters and side effects for both groups. Simultaneously, a patient satisfaction survey was conducted for the digital intervention. This study aimed to confirm the efficacy and safety of digital intervention in the treatment of PCOS.

Methods

Study Design and Participants

A total of 80 women with PCOS (including 73 newly diagnosed and 7 previously diagnosed women with PCOS) from the endocrinology clinic of a tertiary affiliated hospital of Tongji University were enrolled between June 21, 2022, and August 12, 2023, in this single-center, unblinded, 2-arm parallel study with 1:1 allocation. PCOS was diagnosed in patients attending PCOS outpatient clinics, and eligible patients were invited to participate in this study. The CONSORT eHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist is presented in Multimedia Appendix 1.

The inclusion criteria were being aged 18 to 45 years, having a clinical diagnosis of PCOS according to the 2003 Rotterdam diagnostic criteria [10], and having a HOMA-IR score ≥1.8 (in Asian countries, a HOMA-IR score ≥1.8 can diagnose insulin resistance and predict the occurrence of metabolic syndrome [11]). The exclusion criteria included the inability to use WeChat; hyperthyroidism or hypothyroidism; severe abnormal liver function (liver enzyme levels >3 times the standard limit); severe abnormal renal function (serum creatinine level >123.8 μmol/L or estimated glomerular filtration rate <45 mL/min/1.73 m^2); congenital adrenal hyperplasia, hyperprolactinemia, or adrenal tumor; severe infection, severe anemia, neutropenia, and other blood system diseases; type 1 diabetes, monogenic mutation diabetes, or diabetes due to pancreatic damage or other secondary diabetes; existence of mental illness or dementia; history of using contraceptives, metformin, glucagon-like peptide-1 analogs, pioglitazone, all classes of antidepressant medications, and other drugs in the past 3 months; pregnancy

XSL•FO
RenderX

or intention to become pregnant within 6 months; participation in another clinical trial within 3 months; and declining to participate. Moreover, perimenopausal or menopausal women were excluded from this study. As this last exclusion criterion was not reported in the ClinicalTrials.gov registration (NCT05386706), this exclusion was a protocol deviation.

Ethical Considerations

The study protocol was approved by the ethics committee of the Shanghai Tenth People's Hospital (SHSY-IEC-5.0/21K191/P02); the study was registered in ClinicalTrials.gov (NCT05386706); and the name of the trial registry is Digital Intervention PCOS. No changes were made to this clinical trial after registration. All participants provided a signed written informed consent. Every participant was informed that their participation in the study was completely voluntary, that the information gathered for the study would be kept private, and that they could revoke their agreement at any moment without it having an adverse effect on their care. Each participant was given a unique study ID for data entry, maintenance, and analysis, and their personal information was anonymized.

Randomization and Procedure

Following baseline evaluation, based on a predefined, computer-generated number in a 1:1 assignment, the participants were randomly assigned to receive either digital intervention or metformin. Allocation concealment was achieved by using opaque, sealed envelopes with sequential numbers. One researcher created the random number sequence for this investigation, and another researcher enrolled and assigned participants. During intervention, both groups were aware of their allocation. After intervention, the percentage of adherence to the intervention was calculated as adherence days divided by intervention days.

Digital Intervention Group

The PCOS guidelines recommend lifestyle interventions as the first-line treatment for PCOS (diet, exercise, sleep, and mood) [12,13]. Cognitive behavioral therapy is a type of treatment that combines behavioral and cognitive methods [14]. Prior studies evaluated the effectiveness of 20 cognitive behavioral therapy lifestyle courses with physical therapy, healthy eating, and 9 months of extra feedback via a mobile SMS text messaging service to conventional treatment [8]. In this regard, we designed a digital intervention method based on the WeChat mini program. The mini program was pretested with 10 users for usability and face validity before being used in the intervention. The WeChat mini program was designed by the Endocrinology and Metabolism Department of Shanghai Tenth People's Hospital and developed by Shanghai Zoey Information Technology Co, Ltd. Patients with PCOS assigned to the digital intervention group can register a new account for free through WeChat, and they need to use the WeChat mini program for 12 weeks. As presented in Figure 1, the program consists of 3 modules. In module 1, patients can watch 2 videos per week, with each video being about 5 minutes long; a total of 24 videos can be watched over a 12-week period. These videos describe the 4 elements of a healthy lifestyle that should be prioritized by patients with PCOS: healthy diet, regular exercise, good sleep, and stress reduction. The video specifically explains how the 4 elements affect the occurrence and development of PCOS, what can be effectively done to achieve a good lifestyle, and how to adhere to these lifestyles. Multimedia Appendix 2 presents the title and main content of each video. In module 2, patients need to record their menstruation during their menstrual period (duration of menstruation, amount of bleeding, body temperature, and mood) and record their daily weight (weight change value, BMI, and weight status appearing after weight submission); daily exercise time (according to different exercise intensities, low-, moderate-, and high-intensity exercises were recorded, and the type of exercise needed to be selected); daily sleep time (bedtime and wake-up time); and daily meditation time (also provides 24 pieces of meditation music). Simultaneously, the list of recommended food and food to avoid is provided in the food recommendation part. After recording, the patients can see a statistical chart of their weight and exercise in module 2. To allow patients to sufficiently understand the important role of lifestyle improvement in PCOS treatment, we have set up module 3, which they can selectively read by themselves, including an overall introduction to PCOS; healthy lifestyles suitable for PCOS; different dietary patterns (including low-carbohydrate, high-protein, and restricted energy-balanced diets); choices of different exercises (the intensity of each exercise and the amount of energy that can be consumed); and importance of sleep hygiene (the dangers of circadian disruption). Meanwhile, contact between the patient and doctor through WeChat's electronic message was synchronized, and communication was scheduled every Monday. The doctor acts as the patient's coach, downloading the patient's record data every week; analyzing the patient's performance that week; summarizing the weekly exercise time, weight, sleep, and menstrual period and sending it to the patient; and suggesting possible improvements. In addition, the doctor asks the patients if they encountered problems that cannot be solved and provides support and encouragement. While using the digital intervention, 5 patients experienced temporary system problems (4 patients had video playback issues and 1 patient had data entry issues). After communicating with the programmer, we optimized and improved the system in time and solved the problems encountered by the patients, but the content has not changed.

XSL•FO
RenderX

Figure 1. Overview of the components of the digital intervention.

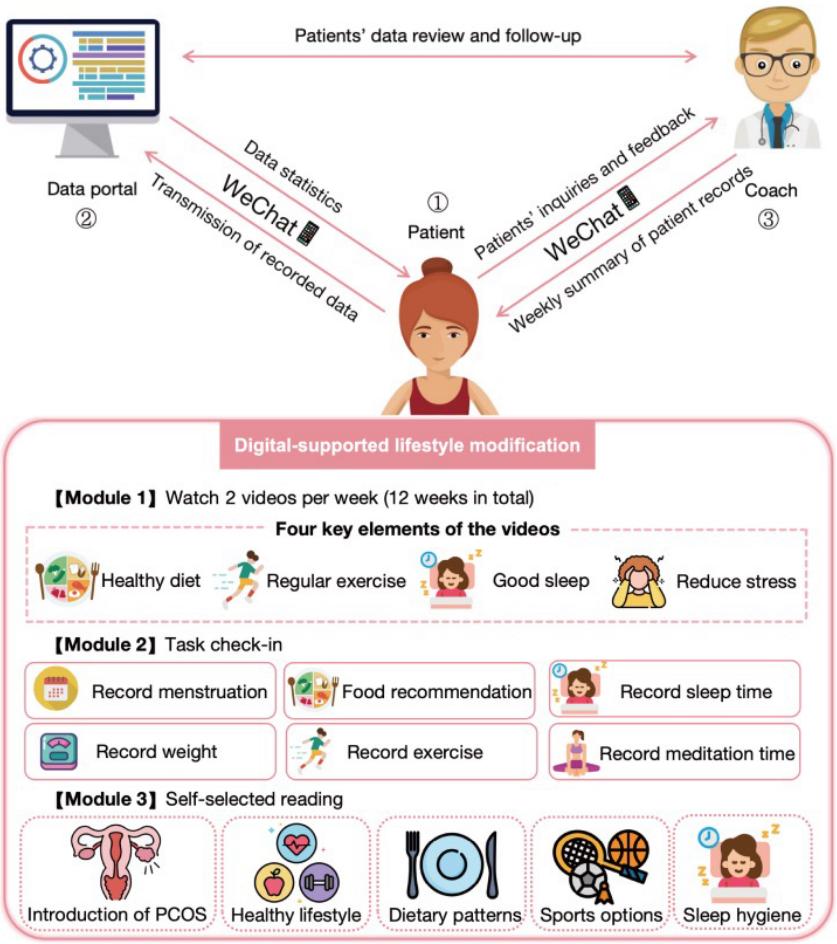

Metformin Group

Metformin, which improves ovulation and enhances insulin sensitivity, is the first-line insulin-sensitizing medication used to treat PCOS; it was used as a treatment drug in the control group in this study [15]. Metformin should be used for at least 3 months after completion depending on the patient's ovulation recovery, androgen status, and metabolic clinical indicators [16,17]. During the 12-week intervention, the metformin group was administered 1000 mg of metformin (Glucophage, Merck KGaA) per day. In previous clinical trials, the dose of metformin ranged from 850 mg to 2000 mg per day [18]. To reduce the side effects of drugs, the dose of 1000 mg per day was selected for this study, and a doctor educated the patients regarding healthy lifestyles during the first visit.

Measurements

Clinical and Laboratory Measurements

The primary outcome was HOMA-IR alterations, whereas the secondary outcomes were changes in menstrual cycles, metabolic indicators, sex hormones, body composition, and degree of fibrosis and steatosis in the liver. Furthermore, body weight, waist circumference, and hip circumference were collected. Body weight (kg)/height (m^2) was used to determine the BMI, and waist circumference (cm)/hip circumference (cm) was used to determine the waist-to-hip ratio. Menstrual cycle is defined as the total number of menstruations in the past 12 months. The following metabolic variables were assessed: fasting blood glucose, fasting insulin, total cholesterol, triglycerides, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, alanine aminotransferase (ALT), aspartate aminotransferase (AST), creatinine, uric acid, sex hormone–binding globulin (SHBG), and glycated hemoglobin A1c (HbA_{1c}). A 75-g oral glucose tolerance test was used to evaluate the postprandial blood glucose and postprandial insulin levels at 120 minutes. Sex hormones included the luteinizing hormone (LH), follicle-stimulating hormone, total testosterone, free testosterone, androstenedione, anti-Müllerian hormone, and dehydroepiandrosterone sulfate (DHEAS). (Fasting blood glucose [mmol/L]×fasting insulin [mU/L])/22.5 was used to calculate the HOMA-IR levels [19]. Transient elastography (Echosens) was used to measure the controlled attenuation parameter and determine the liver stiffness measurement, which represent the extent of liver steatosis and fibrosis, respectively. Dual-energy x-ray absorptiometry (Hologic) was used to measure body composition. At baseline and after the 12-week intervention, all clinical and laboratory tests were conducted at the clinic.

Questionnaires

At baseline and after the 12-week intervention, all participants were required to complete the self-assessed web-based questionnaires, including the 21-item Three-Factor Eating Questionnaire (TFEQ-R21), International Physical Activity Questionnaire (IPAQ), Pittsburgh Sleep Quality Index (PSQI),

XSL•FO
RenderX

and Hospital Anxiety and Depression Scale (HADS). Furthermore, participants in the digital intervention group filled out a self-assessed web-based satisfaction survey 12 weeks after the intervention, and the 5 questions in the questionnaire are shown in Multimedia Appendix 3.

TFEQ-R21 is a questionnaire used to examine eating behavior [20]. The 3 scales comprising the TFEQ-R21 are the cognitive restraint scale (6 items), uncontrolled eating scale (9 items), and emotional eating scale (6 items). These scales cover different aspects of eating behavior. The higher the score, the more severe the behavior of cognitive restraint, uncontrolled eating, or emotional eating.

The translated short form of the IPAQ was used to self-evaluate physical activity, and there was a 7-day recall period. The usefulness of the IPAQ has been validated. It provides data on the amount of time spent each week on several physical activity domains, including low, moderate, and vigorous intensity [21,22]. Metabolic equivalent (MET) is a measure of energy expenditure expressed as a multiple of resting energy costs and can be used to quantify the intensity of different activities. By calculating MET×days×daily time, standardized techniques were used to convert time, days per activity, and intensity to MET minute per week scores. One minute of walking, 1 minute of general moderate-intensity activity, and 1 minute of vigorous exercise were equivalent to 3.3, 4.0, and 8.0 METs, respectively.

The PSQI is a self-report assessment tool that measures sleep quality over a 30-day period [23]. The scale yields a global score and 7 component scores, namely, subjective sleep quality, sleep latency, duration, efficiency, sleep disruptions, use of sleeping pills, and dysfunction throughout the day. Every element receives a score between 0 and 3, and the overall result can range from 0 to 21. A higher number indicates lower sleep quality.

A 14-item test called the HADS is used to assess an individual's anxiety and depression levels [24]. Of 14 items, 7 are associated with depression, and the remaining items are associated with anxiety. The 4 response possibilities on each item range from 0 to 3, enabling the subscales of anxiety and depression (HADS-A and HADS-D) to have independent total scores, each ranging from 0 to 21. While the scores of 0 to 7 are considered as "normal" and 8 to 10 are considered as "borderline," scores of 11 or more on each subscale are considered to indicate serious instances of psychological morbidity.

Sample Size Estimation

On the basis of previous research [17,25,26], we calculated the mean difference in the HOMA-IR levels before and after the metformin intervention as 1.0 and the SD of the outcome variable as 1.0. Alternatively, we predicted the mean difference in the HOMA-IR levels before and after the digital intervention to be 2.0 and the SD of the outcome variable to be 1.0. Our power analysis revealed that according to a web-based calculator (MedSci Sample Size tools), with a 30% sample loss rate, a minimum of 30 participants were required for each group to achieve a significant outcome (α=.05, power=0.90); the total sample size of the study was 60.

Statistical Analysis

The intention-to-treat principle was used for all data analyses, and SPSS (version 25; IBM Corp) was used for all statistical analyses. Continuous variables were expressed as means with 95% CIs. Mean differences between the two treatment groups are displayed as least squares mean changes and 95% CIs. To evaluate the group differences in the primary and continuous secondary end points, we used an analysis of covariance model in which the treatment was a fixed effect and the corresponding baseline value was a covariate. Fisher exact test or chi-square test was used where applicable to analyze categorical variables. A P value <.05 was considered to be statistically significant.

Results

Participant Characteristics

The process of recruiting ran from June 2022 to May 2023, and the follow-up assessments were completed in August 2023. Of the 137 outpatients who had PCOS from June 2022 to August 2023, a total of 80 met the admission criteria and volunteered to participate in the study (Figure 2). They were randomly divided into the digital intervention group (40/80, 50%) and metformin group (40/80, 50%). The dropout rates were 13% (5/40) and 8% (3/40) for the digital intervention and metformin groups, respectively. The digital intervention group used a WeChat mini program for 12 weeks, whereas the metformin group received daily doses of 1000 mg of metformin also for 12 weeks. As shown in Multimedia Appendix 4, the adherence rate of the digital intervention group was 89% (95% CI 83%-95%), which was calculated by the participants' recorded data on the WeChat mini program. On the basis of the number of days participants took metformin within 12 weeks, the adherence rate of the metformin group was 98% (95% CI 96%-99%). The adherence rate was higher in the metformin group than in the digital intervention group (P=.003).

Table 1 presents the clinical characteristics measured at baseline. No differences were observed between the two groups in terms of age, body measurements, menstrual cycles, sex hormones, glucose and lipid metabolism, uric acid, body fat distribution, hepatic steatosis, and hepatic fibrosis levels. However, the digital intervention group had a higher mean AST of 25.66 U/L (95% CI 20.78-30.53) compared with the metformin group, which had a mean AST of 19.23 U/L (95% CI 17.00-21.45). Furthermore, there were no appreciable variations in terms of sleep quality, anxiety, depression, eating habits, weekly physical activity, and daily sitting time between the two groups.

XSL•FO
RenderX

Figure 2. Flow diagram.

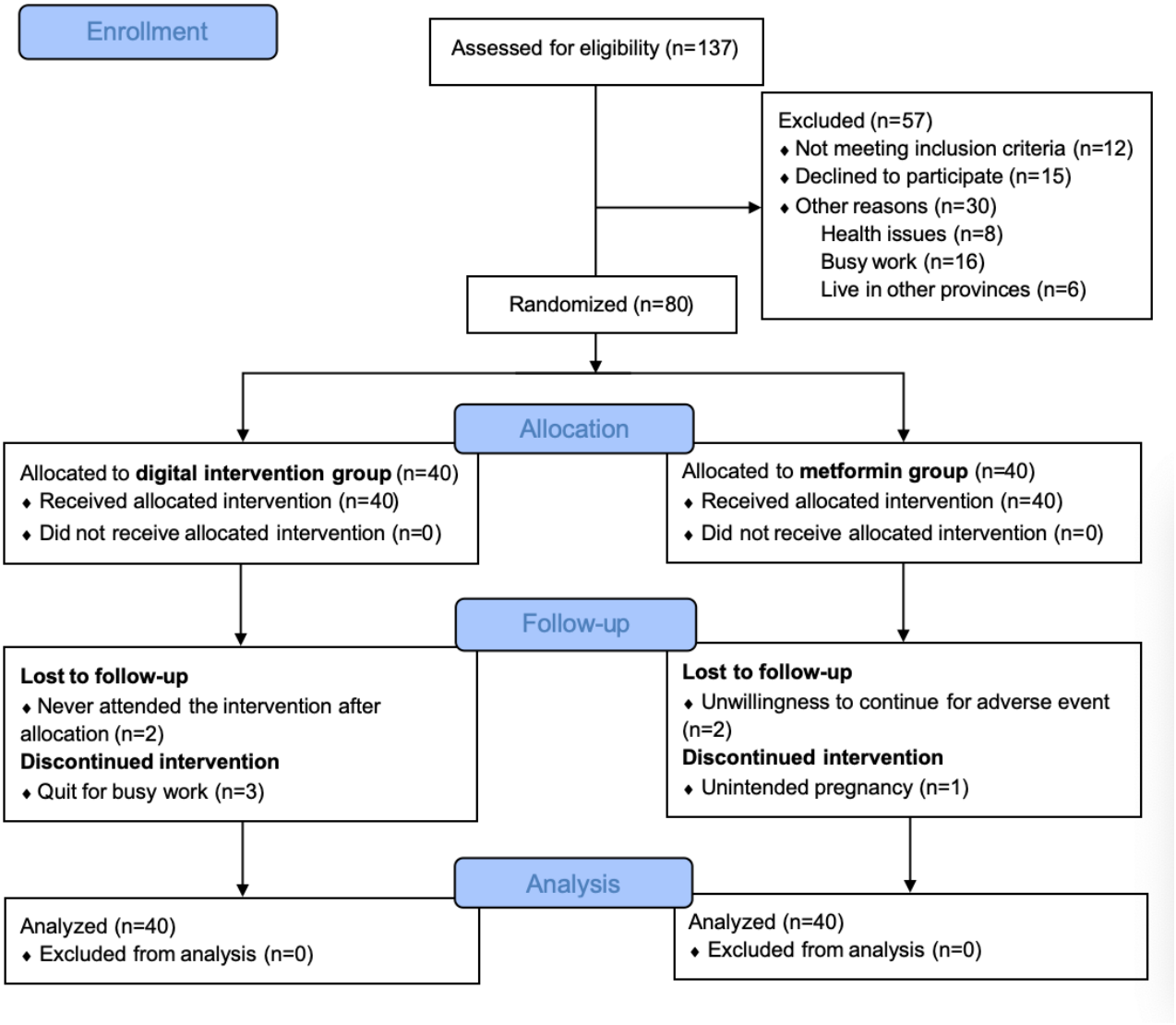

XSL•FO
RenderX

Table 1. Clinical characteristics of participants measured in the baseline.

Variables	Digital intervention (n=40), mean (95% CI)	Metformin (n=40), mean (95% CI)	*P* value
Age (y)	27.60 (26.22-28.98)	27.95 (26.26-29.64)	.75
Weight (kg)	67.89 (64.00-71.78)	69.23 (65.78-72.68)	.61
BMI (kg/m^2)	26.22 (24.81-27.63)	26.30 (25.12-27.48)	.93
Waist circumference (cm)	85.63 (82.22-89.03)	84.85 (81.74-87.97)	.74
Hip circumference (cm)	96.57 (94.30-98.84)	98.53 (96.13-100.93)	.23
Waist-to-hip ratio	0.89 (0.86-0.91)	0.86 (0.84-0.88)	.12
Menstrual cycles (number/y)	8.20 (7.20-9.20)	7.82 (6.86-8.78)	.58
Glycated hemoglobin A1c (%)	5.55 (5.43-5.66)	5.54 (5.42-5.66)	.93
Homeostatic model assessment of insulin resistance	4.01 (3.18-4.84)	4.47 (3.54-5.40)	.45
Fasting blood glucose (mmol/L)	4.95 (4.82-5.08)	4.90 (4.70-5.11)	.71
Postprandial blood glucose (mmol/L)	7.60 (6.92-8.27)	7.30 (6.67-7.93)	.52
Fasting insulin (mU/L)	17.82 (14.45-21.20)	19.81 (16.72-22.90)	.38
Postprandial insulin (mU/L)	157.23 (123.22-191.23)	134.35 (103.68-165.01)	.31
Alanine aminotransferase (U/L)	35.63 (25.84-45.43)	26.12 (20.96-31.28)	.09
Aspartate aminotransferase (U/L)	25.66 (20.78-30.53)	19.23 (17.00-21.45)	.02[a]
Creatinine (μmol/L)	58.02 (54.67-61.37)	57.96 (55.53-60.39)	.98
Uric acid (μmol/L)	372.63 (343.10-402.15)	363.88 (337.28-390.47)	.66
Total cholesterol (mmol/L)	4.98 (4.65-5.31)	4.82 (4.57-5.07)	.45
Triglycerides (mmol/L)	1.45 (1.22-1.68)	1.62 (1.29-1.94)	.39
Low-density lipoprotein cholesterol (mmol/L)	3.22 (2.89-3.54)	2.99 (2.79-3.19)	.24
High-density lipoprotein cholesterol (mmol/L)	1.24 (1.03-1.45)	1.15 (1.06-1.24)	.44
Luteinizing hormone (IU/L)	9.67 (7.61-11.72)	10.33 (7.02-13.64)	.73
Follicle-stimulating hormone (IU/L)	5.43 (4.92-5.94)	5.01 (4.38-5.63)	.30
Total testosterone (ng/mL)	1.83 (1.64-2.02)	1.74 (1.51-1.97)	.54
Free testosterone (pg/mL)	4.42 (3.71-5.13)	4.74 (3.76-5.72)	.59
Androstenedione (ng/mL)	5.68 (4.99-6.37)	5.64 (4.71-6.57)	.95
Dehydroepiandrosterone sulfate (μg/dL)	313.01 (263.32-362.70)	289.74 (245.61-333.87)	.49
Sex hormone–binding globulin (nmol/L)	31.58 (20.92-42.24)	27.58 (22.63-32.53)	.53
Anti-Müllerian hormone (ng/mL)	7.05 (5.59-8.50)	6.14 (5.29-6.99)	.28
Total body fat (%)	43.93 (42.24-45.63)	43.21 (41.73-44.68)	.51
Total body lean mass (%)	52.39 (50.69-54.09)	53.16 (51.67-54.65)	.49
Total fat mass (kg)	29.15 (26.80-31.50)	29.17 (27.23-31.10)	.99
Total lean mass (kg)	34.28 (32.49-36.08)	35.78 (33.81-37.76)	.26
Subcutaneous adipose tissue mass (kg)	1.59 (1.43-1.75)	1.59 (1.45-1.73)	.95
Visceral adipose tissue mass (kg)	0.63 (0.55-0.72)	0.62 (0.56-0.69)	.82
Controlled attenuation parameter (dB/m)	308.68 (291.03-326.32)	316.9 (298.59-335.21)	.52
Liver stiffness measurement (kPa)	5.42 (4.89-5.95)	4.97 (4.62-5.31)	.15
Pittsburgh Sleep Quality Index score	6.35 (5.08-7.02)	6.05 (5.46-7.24)	.30
Hospital Anxiety and Depression–Anxiety Subscale	6.18 (4.82-7.53)	5.13 (3.78-6.47)	.27

XSL•FO
RenderX

Variables	Digital intervention (n=40), mean (95% CI)	Metformin (n=40), mean (95% CI)	*P* value
Hospital Anxiety and Depression–Depression Subscale	5.60 (4.37-6.83)	5.43 (4.16-6.69)	.18
Uncontrolled eating	20.50 (18.62-22.38)	20.50 (18.70-22.30)	>.99
Cognitive restraint	15.80 (14.72-16.88)	16.13 (14.98-17.28)	.67
Emotional eating	14.38 (12.53-16.22)	14.29 (12.67-15.90)	.94
Total physical activity (metabolic equivalent minutes/wk)	1061.20 (696.61-1425.79)	1538.80 (1003.10-2074.51)	.13
Daily sitting time (min)	465.79 (423.22-508.36)	418.48 (345.91-491.06)	.24

[a]Italicized values indicate *P*<.05.

Overall Intervention Effect Between Groups

Primary Outcome: HOMA-IR

After 12 weeks of treatment, both digital intervention and metformin reduced the HOMA-IR levels (Table 2). The mean change in HOMA-IR from baseline to week 12 was –0.93 (95% CI –1.64 to –0.23) for the digital intervention group and –1.07 (95% CI –2.04 to –.09) for the metformin group. There was no significant difference in HOMA-IR levels between the groups (least squares mean difference –0.20; 95% CI –0.98 to 0.58; *P*=.62), indicating that the improvements in both groups in insulin resistance were comparable.

XSL•FO
RenderX

Table 2. Comparison of clinical characteristics between the groups before and after the 12-week study period.

Variables	Digital intervention (n=40), LS[a] mean (95% CI)	Metformin (n=40), LS mean (95% CI)	Treatment difference (digital intervention–metformin), LS mean (95% CI)	*P* value
Homeostatic model assessment of insulin resistance	*–0.93 (–1.64 to –0.23)*	*–1.07 (–2.04 to –0.09)*	*–0.20 (–0.98 to 0.58)*	*.62* [b]
Weight (kg)	–2.47 (–3.48 to –1.46)	–2.21 (–3.14 to –1.28)	–0.33 (–1.70 to 1.03)	.63
BMI (kg/m^2)	–0.95 (–1.34 to –0.56)	–0.85 (–1.21 to –0.50)	–0.10 (–0.63 to 0.42)	.69
Waist circumference (cm)	–4.52 (–5.89 to –3.14）	–2.66 (–3.65 to –1.66）	–1.84 (–3.44 to –0.24）	.03
Hip circumference (cm)	–1.98 (–2.87 to –1.09)	–1.80 (–2.45 to –1.15)	–0.29 (–1.37 to 0.79)	.60
Waist-to-hip ratio	–0.03 (–0.04 to –0.02)	–0.01 (–0.02 to 0.00)	–0.02 (–0.03 to 0.00)	.03
Menstrual cycles (no/yr)	1.43 (1.10 to 1.75)	1.51 (1.23 to 1.79)	–0.02 (–0.42 to 0.37)	.91
Glycated hemoglobin A1c (%)	–0.21 (–0.31 to –0.11)	–0.19 (–0.28 to –0.10)	0.00 (–0.11 to 0.11)	.97
Fasting blood glucose (mmol/L)	–0.18 (–0.31 to –0.05)	–0.28 (–0.49 to –0.08)	0.11 (–0.07 to 0.29)	.23
Postprandial blood glucose (mmol/L)	–0.40 (–1.30 to 0.50)	0.02 (–0.68 to 0.71)	–0.03 (–0.90 to 0.85)	.95
Fasting insulin (mU/L)	–3.47 (–6.31 to –0.64)	–3.37 (–6.70 to –0.04)	–1.21 (–4.64 to 2.22)	.48
Postprandial insulin (mU/L)	–59.87 (–106.61 to –13.14)	–7.86 (–37.42 to 21.70)	–33.95 (–76.23 to 8.32)	.11
Alanine aminotransferase (U/L)	–11.36 (–18.89 to –3.83)	–2.81 (–6.37 to 0.76)	–4.34 (–8.75 to 0.08)	.05
Aspartate aminotransferase (U/L)	–5.24 (–10.09 to –0.39)	–1.25 (–3.23 to 0.73)	0.49 (–1.97 to 2.95)	.69
Creatinine (μmol/L)	–1.34 (–4.31 to 1.64)	–2.23 (–4.19 to –0.27)	0.60 (–2.50 to 3.69)	.70
Uric acid (μmol/L)	–9.88 (–34.29 to 14.52)	–3.34 (–21.93 to 15.25)	–3.48 (–29.83 to 22.88)	.79
Total cholesterol (mmol/L)	–0.16 (–0.33 to 0.01)	–0.07 (–0.27 to 0.13)	–0.04 (–0.29 to 0.21)	.74
Triglycerides (mmol/L)	–0.20 (–0.41 to 0.01)	–0.15 (–0.41 to 0.11)	–0.13 (–0.41 to 0.16)	.38
Low-density lipoprotein cholesterol (mmol/L)	–0.13 (–0.29 to 0.04)	0.01 (–0.15 to 0.16)	–0.08 (–0.30 to 0.13)	.44
High-density lipoprotein cholesterol (mmol/L)	–0.14 (–0.38 to 0.09)	–0.01 (–0.07 to 0.05)	–0.00 (–0.10 to 0.09)	.95
Luteinizing hormone (IU/L)	1.72 (–1.49 to 4.94)	–1.65 (–5.17 to 1.86)	3.65 (1.13 to 6.18)	.01
Follicle-stimulating hormone (IU/L)	0.08 (–0.61 to 0.76)	–0.15 (–0.94 to 0.63)	0.43 (–0.42 to 1.28)	.32
Total testosterone (ng/mL)	–0.18 (–0.37 to 0.02)	–0.15 (–0.34 to 0.03)	0.02 (–0.22 to 0.25)	.87
Free testosterone (pg/mL)	s0.24 (–0.99 to 0.51)	–0.77 (–1.44 to –0.10)	0.33 (–0.42 to 1.07)	.38
Androstenedione (ng/mL)	0.11 (–0.89 to 1.12)	0.27 (–0.66 to 1.21)	–0.11 (–1.39 to 1.18)	.87
Dehydroepiandrosterone sulfate (μg/dL)	–17.41 (–55.20 to 20.38)	59.13 (4.34 to 113.92)	–69.73 (–129.70 to –9.75)	.02
Sex hormone–binding globulin (nmol/L)	–1.24 (–11.11 to 8.63)	6.52 (2.44 to 10.60)	–4.61 (–13.81 to 4.59)	.32
Anti-Müllerian hormone (ng/mL)	–0.31 (–1.25 to 0.62)	–0.21 (–0.78 to 0.35)	0.06 (–0.99 to 1.11)	.91
Total body fat (%)	–4.53 (–5.76 to –3.30)	–3.31 (–4.48 to –2.13)	–0.96 (–2.43 to 0.51)	.20
Total body lean mass (%)	4.21 (3.01 to 5.40)	3.22 (2.06 to 4.39)	0.65 (–0.77 to 2.07)	.36
Total fat mass (kg)	–4.44 (–5.60 to –3.29)	–2.86 (–3.66 to –2.05)	–1.59 (–2.88 to –0.30)	.02
Total lean mass (kg)	0.86 (–0.24 to 1.96)	1.02 (0.15 to 1.89)	–0.49 (–1.77 to 0.78)	.44
Subcutaneous adipose tissue mass (kg)	–0.27 (–0.39 to –0.15)	–0.15 (–0.26 to –0.05)	–0.11 (–0.26 to 0.03)	.13
Visceral adipose tissue mass (kg)	–0.12 (–0.18 to –0.06)	–0.09 (–0.13 to –0.05)	–0.02 (–0.09 to 0.04)	.47
Controlled attenuation parameter (dB/m)	–27.00 (–46.02 to –7.98)	–16.75 (–35.27 to 1.77)	–11.34 (–32.18 to 9.51)	.28

XSL•FO
RenderX

Variables	Digital intervention (n=40), LS[a] mean (95% CI)	Metformin (n=40), LS mean (95% CI)	Treatment difference (digital intervention–metformin), LS mean (95% CI)	*P* value
Liver stiffness measurement (kPa)	–0.76 (–1.44 to –0.09)	–0.15 (–0.70 to 0.40)	–0.37 (–1.17 to 0.42)	.35
Pittsburgh Sleep Quality Index score	–0.70 (–1.66 to 0.27)	–1.12 (–2.17 to –0.08)	0.40 (–0.93 to 1.72)	.55
Hospital Anxiety and Depression–Anxiety Subscale	–1.33 (–2.24 to –0.43)	–0.55 (–1.52 to 0.43)	–0.53 (–1.75 to 0.68)	.38
Hospital Anxiety and Depression–Depression Subscale	–1.55 (–2.48 to –0.61)	–1.45 (–2.47 to –0.44)	–0.17 (–1.37 to 1.03)	.78
Uncontrolled eating	–2.18 (–3.73 to –0.63)	–2.30 (–3.97 to –0.63)	0.11 (–1.78 to 1.99)	.91
Cognitive restraint	1.88 (0.63 to 3.13)	1.27 (0.08 to 2.46)	0.33 (–1.18 to 1.85)	.66
Emotional eating	–0.91 (–1.80 to –0.02)	–1.42 (–2.26 to –0.59)	0.55 (–0.60 to 1.69)	.35
Total physical activity (metabolic equivalent minutes/wk)	642.69 (176.84 to 1108.54)	–229.23 (–880.15 to 421.69)	373.93 (–218.34 to 966.20)	.21
Daily sitting time (minutes)	–32.81 (–94.76 to 29.13)	17.14 (–37.16 to 71.45)	–26.75 (–94.55 to 41.06)	.43

[a]LS: least squares.

[b]The italicized values are the *P* value of the primary outcome.

Secondary Outcomes

After the 12-week intervention, both groups exhibited substantial improvements in body measurements, including body weight, BMI, waist circumference, hip circumference, and waist-to-hip ratio. The digital intervention group exhibited higher improvements in waist circumference (least squares mean difference –1.84 cm; 95% CI –3.44 to –0.24; *P*=.03) and waist-to-hip ratio (least squares mean difference –0.02; 95% CI –0.03 to 0.00; *P*=.03) than the metformin group. In addition, the number of annual menstrual cycles increased in both groups (digital intervention: 1.43, 95% CI 1.10-1.75; metformin: 1.51, 95% CI 1.23-1.79), but the difference was not significant (least squares mean difference –0.02; 95% CI –0.42 to 0.37; *P*=.91).

Considering glucose metabolism, the HbA_{1c}, fasting blood glucose, and fasting insulin of both groups significantly improved, and there was no difference in these indices between the two groups. After the 12-week intervention, the lipid metabolism and uric acid levels did not change in either group. After the digital intervention, the ALT, AST, controlled attenuation parameter, and liver stiffness measurement significantly improved; however, no significant difference was observed between the groups. In the metformin group, the creatinine levels decreased, and there was no difference between the groups. In terms of sex hormone–related indicators, the metformin group exhibited a significant decrease in free testosterone levels (–0.77 pg/mL; 95% CI –1.44 to –0.10) but an increase in DHEAS (59.13 µg/dL; 95% CI 4.34-113.92) and SHBG (6.52 nmol/L; 95% CI 2.44-10.60) levels. Although there was no significant change in DHEAS levels in the digital intervention group, there was a decrease compared with the metformin group (least squares mean difference: –69.73 µg/dL, 95% CI –129.70 to –9.75; *P*=.02). In addition, the LH level did not significantly change in either group, but it increased in the digital intervention group compared with the metformin group (least squares mean difference 3.65 IU/L; 95% CI 1.13-6.18; *P*=.005).

After the 12-week intervention, both groups exhibited a notable improvement in body fat distribution. The total body fat, total fat mass, subcutaneous adipose tissue mass, and visceral adipose tissue mass all decreased, whereas the total body lean mass increased in both groups. The digital intervention group exhibited a more significant decrease in total fat mass than the metformin group (least squares mean difference –1.59 kg; 95% CI –2.88 to –0.30; *P*=.02). Furthermore, the metformin group exhibited an increase in total body lean mass, with no difference between the groups.

Digital intervention improved both the HADS-A and HADS-D scores, whereas metformin only improved the HADS-D scores, with no differences between the groups. Both groups exhibited improved eating behaviors, including a decrease in uncontrolled eating, an increase in cognitively restrictive eating, and a decrease in emotional eating, with no differences between the groups. In terms of sleep quality, the mean change in PSQI from baseline to week 12 was –0.70 (–1.66 to 0.27) for the digital intervention group and –1.12 (–2.17 to –0.08) for the metformin group, but there was no difference between the groups (least squares mean difference 0.40; 95% CI –0.93 to 1.72; *P*=.55). The digital intervention effectively increased the weekly total physical activity but did not significantly differ from metformin (least squares mean difference 373.93 MET minutes per week; 95% CI –218.34 to 966.20; *P*=.22).

Adverse Events

During the study period, no fatalities or major adverse events were reported (Table 3). In the metformin group, the discontinuation rate due to adverse events was 5% (2/40). The rates of adverse events during the treatment course were 8% (3/40) and 35% (14/40) in the digital intervention and metformin groups, respectively. A minimum of 1 gastrointestinal

XSL•FO
RenderX

treatment-emergent adverse event was reported by approximately 30% of the patients in the metformin group. The most frequent gastrointestinal treatment-emergent adverse events were diarrhea (20%), nausea (15%), decreased appetite (8%), and abdominal distension (5%). Most gastrointestinal treatment–emergent adverse events were mild to moderate in severity, with symptoms being noticeable in the first week and then gradually improving over the first month. In the metformin group, severe hypersensitivity reaction and dizziness were reported in one case each. Furthermore, 5% of the participants who received the digital intervention reported feeling hungry and 3% reported feeling dizzy.

Table 3. Adverse events and safety data during the treatment in 2 groups.

Event	Digital intervention (n=40), n (%)	Metformin (n=40), n (%)
Gastrointestinal adverse events	0 (0)	12 (30)
Diarrhea	0 (0)	8 (20)
Nausea	0 (0)	6 (15)
Decreased appetite	0 (0)	3 (8)
Abdominal distension	0 (0)	2 (5)
Vomiting	0 (0)	1 (3)
Upper abdominal pain	0 (0)	1 (3)
Constipation	0 (0)	1 (3)
Sensations of hunger	2 (5)	0 (0)
Headache	0 (0)	0 (0)
Hypoglycemia	0 (0)	0 (0)
Severe hypersensitivity reaction	0 (0)	1 (3)
Dizziness	1 (3)	1 (3)

Satisfaction With the Program

We also conducted a satisfaction survey for patients with PCOS who completed the 12-week digital intervention (n=35). As presented in Multimedia Appendix 3, in total, 80% of the users reported feeling satisfied overall, 89% of the users thought it saves time, 74% of users thought their PCOS-related issues were resolved, 86% of the users believed that they had a more comprehensive understanding of PCOS, and 80% of the users would recommend the program to others. Of the 35 women who completed the follow-up, 31 responded, with 25 (81%) saying the coaching was most effective, 4 (13%) saying knowledge acquisition was most effective, and 2 (6%) saying tracking was most effective.

Discussion

Principal Findings

Our results provide evidence that WeChat-based digital therapies significantly improve self-management and the metabolic and reproductive aspects of the disease among patients with PCOS. While the WeChat-based digital intervention was not statistically different from metformin treatment in terms of lowering HOMA-IR levels, it was more effective in reducing waist circumference, waist-to-hip ratio, total fat mass, and DHEAS. In terms of safety, the main adverse events in the digital intervention group were not as high as those in the metformin group. With a randomized controlled trial design and a 12-week follow-up, the study's findings offer the first convincing proof of the efficacy and safety of self-supervised, WeChat-based digital intervention to improve PCOS.

Using gold standard procedures, it was estimated that approximately 75% of patients with PCOS had insulin resistance and that the endocrine and reproductive characteristics of PCOS can be brought on by insulin resistance [27]. In clinical studies, HOMA-IR is frequently used as a surrogate marker to detect insulin resistance. Multiple studies have shown that alleviating insulin resistance is closely associated with improving menstrual cycles, ovulation, and in vitro fertilization results in PCOS [28-30]. Consequently, HOMA-IR was selected in this study as the major end point as it plays a significant role in the onset and progression of PCOS. This study demonstrated that, in patients with PCOS with insulin resistance, HOMA-IR can be significantly decreased by digital intervention. Prior research has shown that when several therapy methods (such as metformin, rosiglitazone, and acupuncture) were given for approximately 12 weeks, patients with PCOS experienced improvements in their hyperandrogenemia, weight loss, and metabolic and biochemical parameters [17,25,26].

Furthermore, the study showed that, in the same population, both digital intervention and metformin therapy reduced body weight, BMI, waist circumference, hip circumference, and waist-to-hip ratio; however, the former was associated with a greater reduction in waist circumference and waist-to-hip ratio levels than the latter. Similarly, previous research showed that metformin did not significantly reduce waist circumference as much as lifestyle change did [31,32]. In a recent study, anthropometric indicators, including body weight, waist circumference, and fat mass, significantly improved in the web-based lifestyle intervention group compared with the control group [33]. These results may be attributed to an increase in patients' physical activities, as only the digital intervention

XSL•FO
RenderX

group experienced an increase in weekly total physical activity after the intervention in this study.

Furthermore, the menstrual frequency, HbA_{1c}, fasting blood glucose, and fasting insulin of both groups significantly improved; however, there were no significant differences between the groups. Consistent with our results, no significant differences were observed in menstrual patterns between the groups in a previous randomized controlled study of metformin versus lifestyle changes [32]. Another study demonstrated that metformin therapy improved anthropometric parameters, menstrual cycles, and glucose metabolism [34]. Alternatively, the digital intervention significantly reduced the HbA_{1c} levels in patients with prediabetes and diabetes [35,36]. In a study of digital intervention in patients with obesity, the intervention group exhibited significant improvements in HbA_{1c}, fasting blood glucose, and HOMA-IR levels while retaining muscle mass and reducing body fat and obtained equivalent outcomes to an intensive obesity management program [37].

The digital intervention group exhibited improved liver function (ALT and AST), liver steatosis, and liver fibrosis but did not show a statistically significant difference from the metformin group. Previous studies have shown that nonalcoholic fatty liver disease and liver function in patients with PCOS could be improved by diet control and exercise [38], and patients in the digital intervention group probably changed their diet and exercise routines because of these benefits in our study. In this study, the creatinine levels also decreased in the metformin group but not in the digital intervention group. Consistently, metformin decreased the serum creatinine levels and renal oxidative damage and fibrosis in a study on diabetic mice [39].

In terms of sex hormones, LH was elevated in the digital intervention group compared with the metformin group, but there was no difference in the LH levels before and after treatment in either group. This may be because not all reproductive hormones in this study were measured during menstruation. As the DHEAS levels increased after metformin treatment, the DHEAS levels decreased in the comparison of differences between the groups (digital intervention–metformin). Metformin therapy has been previously reported to increase DHEAS levels in women with infertility due to PCOS [40]. As metformin can improve insulin sensitivity and adrenal cells have insulin receptors [41], insulin may have an impact on peripheral or adrenal tissue DHEAS metabolism. Meanwhile, the free testosterone levels decreased and the SHBG levels increased in the metformin group. In previous studies, both insulin resistance status improvement and circulating insulin reduction have been shown to increase serum SHBG levels, thereby reducing the blood level of free testosterone in patients with PCOS [34,42].

Because central obesity is frequently associated with patients with PCOS [43], we examined the distribution of body fat in these patients. The percentage of total body fat, percentage of total body lean mass, total fat mass, subcutaneous adipose tissue mass, and visceral adipose tissue mass were improved significantly in both groups, but the decline of the total fat mass was more pronounced in the digital intervention group. This may be explained by the metformin group receiving oral lifestyle guidance. Similar to our results, a meta-analysis suggests that lifestyle interventions are the best way to improve body composition and cardiorespiratory health in women with PCOS [44].

Using the questionnaire, we determined changes in the patients' sleep quality, mood (anxiety and depression), eating behaviors, and exercise condition. Both groups exhibited significantly improved depression and eating behaviors. The digital intervention improved patients' anxiety and level of physical activity. In a recent study of digital intervention for other diseases, it effectively improved physical activity, depression, anxiety, and eating behavior [45-47]. Metformin, alternatively, significantly improved sleep quality in our study. Studies have demonstrated that in people with type 2 diabetes, metformin therapy improved sleep quality [48]. In addition, it reduced social stress–related anxiety-like behaviors [49].

In terms of safety, the main adverse events in the digital intervention groups were sensations of hunger (5%) and dizziness (3%). The main adverse events in the metformin groups were gastrointestinal treatment–emergent adverse events (30%). Overall, the digital intervention was found to have fewer side effects, which is advantageous. In the satisfaction questionnaire, most of the patients who completed the follow-up of the digital intervention were satisfied with this treatment and would recommend it to other patients with PCOS. On the basis of the questionnaire's results, the patients believed that the digital intervention could provide a more thorough understanding of PCOS, solve PCOS-related issues, and save more time. This suggests that the digital intervention is feasible and beneficial while saving doctors' time and being beneficial for more patients with PCOS. Regarding adherence, while the digital intervention helped patients follow through on lifestyle modifications more successfully, metformin adherence outperformed the digital intervention.

In the digital intervention group, 81% of patients felt that the coaching was most effective. In previous studies, for women with overweight or obesity and PCOS, lifestyle therapies with added SMS text messaging assistance appear to aid in enhancing physical fitness, activity, and aerobic capacity as well as minimize sedentary behavior. It shows that supervision and communication play an important role in maintaining a healthy lifestyle.

Limitations

However, this study has several limitations. First, the selected population had insulin-resistant PCOS and thus does not represent the entire population. Second, sex hormone indicators were not monitored during menstruation, which affected the judgment of sex hormone results. Third, the participants in this study were not blind, and multiple evaluations of the results increased the chances of type 1 errors. Fourth, no formal evaluation of the cost of the intervention was conducted. Finally, there is a lack of long-term follow-up data to assess whether these changes will persist over time. Thus, multicenter, large-sample studies are warranted to further elucidate the efficacy of digital intervention.

XSL•FO
RenderX

Conclusions

This study summarizes and confirms the feasibility and application of digital intervention for women with PCOS. The results of the study indicated that the intervention is promising in improving self-management behavior and the metabolic and reproductive features of PCOS. Furthermore, in terms of reducing waist circumference, waist-to-hip ratio, total fat mass, and DHEAS, the digital intervention was more beneficial than metformin. Regarding safety, there were fewer adverse events in the digital intervention group than the metformin group.

Acknowledgments

This research has been supported by funding from the Shanghai Tenth People's Hospital Funding for clinical research projects YNCR2C013; the National Nature Science Foundation of China (grants 81601269, 82170861, 82103713, and 81970677); and the SHDC Clinical Research Plan (grant SHDC2020CR1017B).

The authors acknowledge that this study was partly financed by Hangzhou Kang Ming Information Technology Co, Ltd, and the WeChat mini program used in this study was provided by Shanghai Zoey Information Technology Co, Ltd.

Authors' Contributions

MZ, SQ, and HC designed the study. DD, MC, XS, LW, and YW collected the data. DD, YZ, HL, MK, and JL analyzed the clinical data. DD and MZ drafted the manuscript. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.
[PDF File (Adobe PDF File), 2176 KB - jmir_v26i1e55883_app1.pdf]

Multimedia Appendix 2
The content of 24 videos.
[DOCX File , 17 KB - jmir_v26i1e55883_app2.docx]

Multimedia Appendix 3
Satisfaction survey on digital intervention (n=35).
[DOCX File , 13 KB - jmir_v26i1e55883_app3.docx]

Multimedia Appendix 4
The adherence rate of the digital intervention group and metformin group.
[DOCX File , 272 KB - jmir_v26i1e55883_app4.docx]

References

1. Teede HJ, Tay CT, Laven J, Dokras A, Moran LJ, Piltonen TT, International PCOS Network. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Fertil Steril 2023 Oct;120(4):767-793. [doi: 10.1016/j.fertnstert.2023.07.025] [Medline: 37589624]
2. Lizneva D, Suturina L, Walker W, Brakta S, Gavrilova-Jordan L, Azziz R. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertil Steril 2016 Jul;106(1):6-15 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.05.003] [Medline: 27233760]
3. Tong C, Wu Y, Zhang L, Yu Y. Insulin resistance, autophagy and apoptosis in patients with polycystic ovary syndrome: association with PI3K signaling pathway. Front Endocrinol (Lausanne) 2022;13:1091147 [FREE Full text] [doi: 10.3389/fendo.2022.1091147] [Medline: 36589825]
4. Zhao H, Zhang J, Cheng X, Nie X, He B. Insulin resistance in polycystic ovary syndrome across various tissues: an updated review of pathogenesis, evaluation, and treatment. J Ovarian Res 2023 Jan 11;16(1):9 [FREE Full text] [doi: 10.1186/s13048-022-01091-0] [Medline: 36631836]
5. Moran LJ, Tassone EC, Boyle J, Brennan L, Harrison CL, Hirschberg AL, et al. Evidence summaries and recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome: lifestyle management. Obes Rev 2020 Oct;21(10):e13046. [doi: 10.1111/obr.13046] [Medline: 32452622]
6. Lugones-Sanchez C, Recio-Rodriguez JI, Agudo-Conde C, Repiso-Gento I, G Adalia E, Ramirez-Manent JI, EVIDENT 3 Investigators. Long-term effectiveness of a smartphone app combined with a smart band on weight loss, physical activity,

XSL•FO
RenderX

and caloric intake in a population with overweight and obesity (evident 3 study): randomized controlled trial. J Med Internet Res 2022 Feb 01;24(2):e30416 [FREE Full text] [doi: 10.2196/30416] [Medline: 35103609]
7. Liu F, Song T, Yu P, Deng N, Guan Y, Yang Y, et al. Efficacy of an mhealth app to support patients' self-management of hypertension: randomized controlled trial. J Med Internet Res 2023 Dec 19;25:e43809 [FREE Full text] [doi: 10.2196/43809] [Medline: 38113071]
8. Jiskoot G, Dietz de Loos A, Beerthuizen A, Timman R, Busschbach J, Laven J. Long-term effects of a three-component lifestyle intervention on emotional well-being in women with Polycystic Ovary Syndrome (PCOS): a secondary analysis of a randomized controlled trial. PLoS One 2020;15(6):e0233876 [FREE Full text] [doi: 10.1371/journal.pone.0233876] [Medline: 32479544]
9. Jiskoot G, de Loos AD, Timman R, Beerthuizen A, Laven J, Busschbach J. Changes in eating behavior through lifestyle treatment in women with polycystic ovary syndrome (PCOS): a randomized controlled trial. J Eat Disord 2022 May 17;10(1):69 [FREE Full text] [doi: 10.1186/s40337-022-00593-y] [Medline: 35581616]
10. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil Steril 2004 Jan;81(1):19-25. [doi: 10.1016/j.fertnstert.2003.10.004] [Medline: 14711538]
11. Esteghamati A, Ashraf H, Khalilzadeh O, Zandieh A, Nakhjavani M, Rashidi A, et al. Optimal cut-off of homeostasis model assessment of insulin resistance (HOMA-IR) for the diagnosis of metabolic syndrome: third national surveillance of risk factors of non-communicable diseases in Iran (SuRFNCD-2007). Nutr Metab (Lond) 2010 Apr 07;7:26 [FREE Full text] [doi: 10.1186/1743-7075-7-26] [Medline: 20374655]
12. Al Wattar BH, Fisher M, Bevington L, Talaulikar V, Davies M, Conway G, et al. Clinical practice guidelines on the diagnosis and management of polycystic ovary syndrome: a systematic review and quality assessment study. J Clin Endocrinol Metab 2021 Jul 13;106(8):2436-2446 [FREE Full text] [doi: 10.1210/clinem/dgab232] [Medline: 33839790]
13. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol 2018 May;14(5):270-284. [doi: 10.1038/nrendo.2018.24] [Medline: 29569621]
14. Stallard P. Evidence-based practice in cognitive-behavioural therapy. Arch Dis Child 2022 Feb;107(2):109-113. [doi: 10.1136/archdischild-2020-321249] [Medline: 34266878]
15. Melin JM, Forslund M, Alesi SJ, Piltonen T, Romualdi D, Spritzer PM, et al. Effects of different insulin sensitisers in the management of polycystic ovary syndrome: a systematic review and meta-analysis. Clin Endocrinol (Oxf) 2023 Nov 07;100(2):149-163. [doi: 10.1111/cen.14983] [Medline: 37933831]
16. Morley LC, Tang T, Yasmin E, Norman RJ, Balen AH. Insulin-sensitising drugs (metformin, rosiglitazone, pioglitazone, D-chiro-inositol) for women with polycystic ovary syndrome, oligo amenorrhoea and subfertility. Cochrane Database Syst Rev 2017 Nov 29;11(11):CD003053 [FREE Full text] [doi: 10.1002/14651858.CD003053.pub6] [Medline: 29183107]
17. Sathyapalan T, Cho LW, Kilpatrick ES, Coady A, Atkin SL. Metformin maintains the weight loss and metabolic benefits following rimonabant treatment in obese women with polycystic ovary syndrome (PCOS). Clin Endocrinol (Oxf) 2009 Jan;70(1):124-128. [doi: 10.1111/j.1365-2265.2008.03345.x] [Medline: 19128368]
18. Løvvik TS, Carlsen SM, Salvesen Ø, Steffensen B, Bixo M, Gómez-Real F, et al. Use of metformin to treat pregnant women with polycystic ovary syndrome (PregMet2): a randomised, double-blind, placebo-controlled trial. Lancet Diabetes Endocrinol 2019 Apr;7(4):256-266. [doi: 10.1016/S2213-8587(19)30002-6] [Medline: 30792154]
19. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985 Jul;28(7):412-419. [Medline: 3899825]
20. Karlsson J, Persson LO, Sjöström L, Sullivan M. Psychometric properties and factor structure of the Three-Factor Eating Questionnaire (TFEQ) in obese men and women. Results from the Swedish Obese Subjects (SOS) study. Int J Obes Relat Metab Disord 2000 Dec;24(12):1715-1725. [Medline: 11126230]
21. Papathanasiou G, Georgoudis G, Georgakopoulos D, Katsouras C, Kalfakakou V, Evangelou A. Criterion-related validity of the short international physical activity questionnaire against exercise capacity in young adults. Eur J Cardiovasc Prev Rehabil 2010 Aug;17(4):380-386. [doi: 10.1097/HJR.0b013e328333ede6] [Medline: 19940775]
22. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
23. Buysse DJ, Reynolds 3rd CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [Medline: 2748771]
24. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [Medline: 6880820]
25. Yilmaz M, Bukan N, Ayvaz G, Karakoç A, Törüner F, Cakir N, et al. The effects of rosiglitazone and metformin on oxidative stress and homocysteine levels in lean patients with polycystic ovary syndrome. Hum Reprod 2005 Dec;20(12):3333-3340. [doi: 10.1093/humrep/dei258] [Medline: 16123091]

XSL•FO
RenderX

26. Wen Q, Hu M, Lai M, Li J, Hu Z, Quan K, et al. Effect of acupuncture and metformin on insulin sensitivity in women with polycystic ovary syndrome and insulin resistance: a three-armed randomized controlled trial. Hum Reprod 2022 Mar 01;37(3):542-552 [FREE Full text] [doi: 10.1093/humrep/deab272] [Medline: 34907435]
27. Moghetti P, Tosi F. Insulin resistance and PCOS: chicken or egg? J Endocrinol Invest 2021 Feb;44(2):233-244. [doi: 10.1007/s40618-020-01351-0] [Medline: 32648001]
28. Yilmaz M, Biri A, Karakoç A, Törüner F, Bingöl B, Cakir N, et al. The effects of rosiglitazone and metformin on insulin resistance and serum androgen levels in obese and lean patients with polycystic ovary syndrome. J Endocrinol Invest 2005 Dec;28(11):1003-1008. [doi: 10.1007/BF03345339] [Medline: 16483179]
29. Zhang D, Yang X, Li J, Yu J, Wu X. Effect of hyperinsulinaemia and insulin resistance on endocrine, metabolic and fertility outcomes in women with polycystic ovary syndrome undergoing ovulation induction. Clin Endocrinol (Oxf) 2019 Sep;91(3):440-448. [doi: 10.1111/cen.14050] [Medline: 31222771]
30. An Y, Sun Z, Zhang Y, Liu B, Guan Y, Lu M. The use of berberine for women with polycystic ovary syndrome undergoing IVF treatment. Clin Endocrinol (Oxf) 2014 Mar;80(3):425-431. [doi: 10.1111/cen.12294] [Medline: 23869585]
31. Karimzadeh MA, Javedani M. An assessment of lifestyle modification versus medical treatment with clomiphene citrate, metformin, and clomiphene citrate-metformin in patients with polycystic ovary syndrome. Fertil Steril 2010 Jun;94(1):216-220 [FREE Full text] [doi: 10.1016/j.fertnstert.2009.02.078] [Medline: 19463994]
32. Curi DD, Fonseca AM, Marcondes JA, Almeida JA, Bagnoli VR, Soares JM, et al. Metformin versus lifestyle changes in treating women with polycystic ovary syndrome. Gynecol Endocrinol 2012 Mar;28(3):182-185. [doi: 10.3109/09513590.2011.583957] [Medline: 22309675]
33. Kohl J, Brame J, Centner C, Wurst R, Fuchs R, Sehlbrede M, et al. Effects of a web-based lifestyle intervention on weight loss and cardiometabolic risk factors in adults with overweight and obesity: randomized controlled clinical trial. J Med Internet Res 2023 Jun 27;25:e43426 [FREE Full text] [doi: 10.2196/43426] [Medline: 37368484]
34. Xing C, Zhao H, Zhang J, He B. Effect of metformin versus metformin plus liraglutide on gonadal and metabolic profiles in overweight patients with polycystic ovary syndrome. Front Endocrinol (Lausanne) 2022;13:945609 [FREE Full text] [doi: 10.3389/fendo.2022.945609] [Medline: 36060969]
35. Katula JA, Dressler EV, Kittel CA, Harvin LN, Almeida FA, Wilson KE, et al. Effects of a digital diabetes prevention program: an RCT. Am J Prev Med 2022 Apr;62(4):567-577 [FREE Full text] [doi: 10.1016/j.amepre.2021.10.023] [Medline: 35151522]
36. Hsia J, Guthrie NL, Lupinacci P, Gubbi A, Denham D, Berman MA, et al. Randomized, controlled trial of a digital behavioral therapeutic application to improve glycemic control in adults with type 2 diabetes. Diabetes Care 2022 Dec 01;45(12):2976-2981 [FREE Full text] [doi: 10.2337/dc22-1099] [Medline: 36181554]
37. Moravcová K, Karbanová M, Bretschneider MP, Sovová M, Ožana J, Sovová E. Comparing digital therapeutic intervention with an intensive obesity management program: randomized controlled trial. Nutrients 2022 May 10;14(10):2005 [FREE Full text] [doi: 10.3390/nu14102005] [Medline: 35631145]
38. Wang D, He B. Current perspectives on nonalcoholic fatty liver disease in women with polycystic ovary syndrome. Diabetes Metab Syndr Obes 2022;15:1281-1291 [FREE Full text] [doi: 10.2147/DMSO.S362424] [Medline: 35494531]
39. Han YC, Tang SQ, Liu YT, Li AM, Zhan M, Yang M, et al. AMPK agonist alleviate renal tubulointerstitial fibrosis via activating mitophagy in high fat and streptozotocin induced diabetic mice. Cell Death Dis 2021 Oct 09;12(10):925 [FREE Full text] [doi: 10.1038/s41419-021-04184-8] [Medline: 34628484]
40. Kjøtrød SB, Sunde A, von Düring V, Carlsen SM. Possible metformin effect on adrenal androgens during pretreatment and IVF cycle in women with polycystic ovary syndrome. Fertil Steril 2009 Feb;91(2):500-508 [FREE Full text] [doi: 10.1016/j.fertnstert.2007.11.069] [Medline: 18304542]
41. Schriock ED, Buffington CK, Hubert GD, Kurtz BR, Kitabchi AE, Buster JE, et al. Divergent correlations of circulating dehydroepiandrosterone sulfate and testosterone with insulin levels and insulin receptor binding. J Clin Endocrinol Metab 1988 Jun;66(6):1329-1331. [doi: 10.1210/jcem-66-6-1329] [Medline: 2967305]
42. Samadi Z, Bambaeichi E, Valiani M, Shahshahan Z. Evaluation of changes in levels of hyperandrogenism, hirsutism and menstrual regulation after a period of aquatic high intensity interval training in women with polycystic ovary syndrome. Int J Prev Med 2019;10:187 [FREE Full text] [doi: 10.4103/ijpvm.IJPVM_360_18] [Medline: 31807257]
43. Zhu S, Li Z, Hu C, Sun F, Wang C, Yuan H, et al. Imaging-based body fat distribution in polycystic ovary syndrome: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 2021;12:697223 [FREE Full text] [doi: 10.3389/fendo.2021.697223] [Medline: 34566888]
44. Haqq L, McFarlane J, Dieberg G, Smart N. The effect of lifestyle intervention on body composition, glycemic control, and cardiorespiratory fitness in polycystic ovarian syndrome: a systematic review and meta-analysis. Int J Sport Nutr Exerc Metab 2015 Dec;25(6):533-540. [doi: 10.1123/ijsnem.2013-0232] [Medline: 24668721]
45. Carl JR, Miller CB, Henry AL, Davis ML, Stott R, Smits JA, et al. Efficacy of digital cognitive behavioral therapy for moderate-to-severe symptoms of generalized anxiety disorder: a randomized controlled trial. Depress Anxiety 2020 Dec 29;37(12):1168-1178. [doi: 10.1002/da.23079] [Medline: 32725848]
46. Fitzsimmons-Craft EE, Taylor CB, Graham AK, Sadeh-Sharvit S, Balantekin KN, Eichen DM, et al. Effectiveness of a digital cognitive behavior therapy-guided self-help intervention for eating disorders in college women: a cluster randomized

XSL•FO
RenderX

clinical trial. JAMA Netw Open 2020 Aug 03;3(8):e2015633 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15633] [Medline: 32865576]
47. Salunkhe VA, Sinha N, Ahlqvist E, Prasad RB, Johansson S, Abrahamsson B, et al. Digital lifestyle treatment improves long-term metabolic control in type 2 diabetes with different effects in pathophysiological and genetic subgroups. NPJ Digit Med 2023 Oct 26;6(1):199 [FREE Full text] [doi: 10.1038/s41746-023-00946-0] [Medline: 37884680]
48. Kajbaf F, Fendri S, Basille-Fantinato A, Diouf M, Rose D, Jounieaux V, et al. The relationship between metformin therapy and sleep quantity and quality in patients with Type 2 diabetes referred for potential sleep disorders. Diabet Med 2014 May;31(5):577-580. [doi: 10.1111/dme.12362] [Medline: 24236897]
49. Zhang YM, Zong HC, Qi YB, Chang LL, Gao YN, Zhou T, et al. Anxiolytic effect of antidiabetic metformin is mediated by AMPK activation in mPFC inhibitory neurons. Mol Psychiatry 2023 Sep;28(9):3955-3965 [FREE Full text] [doi: 10.1038/s41380-023-02283-w] [Medline: 37798418]

Abbreviations

ALT: alanine aminotransferase
AST: aspartate aminotransferase
CONSORT eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DHEAS: dehydroepiandrosterone sulfate
HADS: Hospital Anxiety and Depression Scale
HbA1c: glycated hemoglobin A1c
HOMA-IR: homeostatic model assessment for insulin resistance
IPAQ: International Physical Activity Questionnaire
LH: luteinizing hormone
MET: metabolic equivalent
PCOS: polycystic ovary syndrome
PSQI: Pittsburgh Sleep Quality Index
SHBG: sex hormone–binding globulin
TFEQ-R21: 21-item Three-Factor Eating Questionnaire

Edited by N Cahill; submitted 30.12.23; peer-reviewed by E Dressler, X Cheng; comments to author 28.02.24; revised version received 14.04.24; accepted 15.08.24; published 02.10.24.

Please cite as:
Dilimulati D, Shao X, Wang L, Cai M, Zhang Y, Lu J, Wang Y, Liu H, Kuang M, Chen H, Zhang M, Qu S
Efficacy of WeChat-Based Digital Intervention Versus Metformin in Women With Polycystic Ovary Syndrome: Randomized Controlled Trial
J Med Internet Res 2024;26:e55883
URL: https://www.jmir.org/2024/1/e55883
doi:10.2196/55883
PMID:

©Diliqingna Dilimulati, Xiaowen Shao, Lihua Wang, Meili Cai, Yuqin Zhang, Jiayi Lu, Yao Wang, Hongying Liu, Ming Kuang, Haibing Chen, Manna Zhang, Shen Qu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancing Patient Safety Through an Integrated Internet of Things Patient Care System: Large Quasi-Experimental Study on Fall Prevention

Ming-Huan Wen[1,2], MSN, RN; Po-Yin Chen[3], PhD; Shirling Lin[1], MN, RN; Ching-Wen Lien[1], MSN, RN; Sheng-Hsiang Tu[1], MSN, RN; Ching-Yi Chueh[1], BSN, RN; Ying-Fang Wu[1], BSN, RN; Kelvin Tan Cheng Kian[4], PhD; Yeh-Liang Hsu[5], PhD; Dorothy Bai[6], PhD

[1]Department of Nursing, Taipei Veterans General Hospital, Taipei, Taiwan
[2]School of Nursing, National Yang Ming Chiao Tung University, Taipei, Taiwan
[3]Department of Physical Therapy and Assistive Technology, National Yang Ming Chiao Tung University, Taipei, Taiwan
[4]S R Nathan School of Human Development, Singapore University of Social Sciences, Singapore, Singapore
[5]Gerontechnology Research Center, Yuan Ze University, Taoyuan, Taiwan
[6]School of Gerontology and Long-Term Care, College of Nursing, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Dorothy Bai, PhD
School of Gerontology and Long-Term Care
College of Nursing
Taipei Medical University
No.250, Wuxing Street, Xinyi District
Taipei, 110
Taiwan
Phone: 886 227361661 ext 6332
Fax: 886 223772842
Email: dbai@tmu.edu.tw

Abstract

Background: The challenge of preventing in-patient falls remains one of the most critical concerns in health care.

Objective: This study aims to investigate the effect of an integrated Internet of Things (IoT) smart patient care system on fall prevention.

Methods: A quasi-experimental study design is used. The smart patient care system is an integrated IoT system combining a motion-sensing mattress for bed-exit detection, specifying different types of patient calls, integrating a health care staff scheduling system, and allowing health care staff to receive and respond to alarms via mobile devices. Unadjusted and adjusted logistic regression models were used to investigate the relationship between the use of the IoT system and bedside falls compared with a traditional patient care system.

Results: In total, 1300 patients were recruited from a medical center in Taiwan. The IoT patient care system detected an average of 13.5 potential falls per day without any false alarms, whereas the traditional system issued about 11 bed-exit alarms daily, with approximately 4 being false, effectively identifying 7 potential falls. The bedside fall incidence during hospitalization was 1.2% (n=8) in the traditional patient care system ward and 0.1% (n=1) in the smart ward. We found that the likelihood of bedside falls in wards with the IoT system was reduced by 88% (odds ratio 0.12, 95% CI 0.01-0.97; *P*=.047).

Conclusions: The integrated IoT smart patient care system might prevent falls by assisting health care staff with efficient and resilient responses to bed-exit detection. Future product development and research are recommended to introduce IoT into patient care systems combining bed-exit alerts to prevent inpatient falls and address challenges in patient safety.

(J Med Internet Res 2024;26:e58380) doi:10.2196/58380

KEYWORDS

patient safety; falls; fall prevention; fall risk; sensors; Internet of Things; bed-exit alert; motion-sensing mattress system; care quality; quality improvement; ubiquitous health; mHealth

XSL•FO
RenderX

Introduction

The challenge of preventing in-patient falls remains one of the most critical concerns in health care. It was reported that the fall rate ranged from approximately 2 to 14 per 1000 patient days depending on the hospital setting and patient characteristics [1-3]. Approximately 15%-50% of patients who fall sustain fall-related injuries, including significant injuries (eg, fractures and lacerations), with approximately 4%-6% leading to severe injuries, which can result in activity restrictions, comorbidities, and even death [2]. In addition, falling is often followed by a prolonged length of stay, which may lead to extra costs and a higher likelihood of patients' discharge to long-term institutions [4,5]. An additional 11.5 days of length of stay was estimated to occur after a fall [4], with an average cost increase of 61% [6]. It was estimated that medical costs related to both fatal and nonfatal falls in the United States were approximately US $50.0 billion [7]. The US Centers for Medicare and Medicaid Services advised that falls during hospitalization that can reasonably be prevented using evidence-based care should never occur and stopped paying for such preventable conditions since 2008 [3,8].

Various factors are linked to falls, encompassing intrinsic factors such as a history of falls, deficits in gait or balance or vision, chronic diseases, and medication, and extrinsic factors such as inadequate grab bars in bathrooms or toilets, insufficient lighting, bed height, poorly maintained floor surfaces, and incorrect use of assistive devices [9]. Numerous programs to prevent falls have been implemented and studied, encompassing fall risk assessments, patient and family education, technical interventions, and evaluations following a fall [10,11]. The use of technological interventions, such as call buttons, bedside rails, bed-exit detection systems, and environment redesign, among high-risk groups for falling is becoming increasingly common [12,13]. The effectiveness of these technical intervention programs in individual studies varies, and a pooled estimate of multiple studies concluded that there was no statistically significant effect [13-16].

Bed-exit detection systems have become increasingly used technological solutions for inpatient fall prevention [12,15]. Bed-related activities of inpatients were considered and reported to be an essential factor in the clinic, with growing concerns about issues such as a lack of human resources, physical restraints, and pressure injury reduction [17,18]. Bed-exit detection systems include various forms, such as motion-sensing mattresses or pad systems, infrared systems, cameras, and wearable devices [12,14,16,19]. Despite the fact that these solutions are commonly used in clinical practice to prevent falls, the evidence for their effectiveness is insufficient and inconclusive [20-24]. Two studies found that results were positive [20,22], other studies found no significant result [21,24], and 1 study found that there was no effect after controlling for confounders [23]. In addition to the inconclusive findings of the effect of alerting technologies on fall prevention, false alarms remain a barrier to their application in clinical practice [25,26].

This study aims to examine the effect of an innovative smart patient care system (SPCS) on inpatient fall prevention. The specific objective was to compare the inpatient fall incidences between patients who were admitted to wards that used a traditional patient care system (TPCS) and those admitted to wards that used an SPCS.

Methods

Recruitment

We used a quasi-experimental study design to investigate relationships between an Internet of Things (IoT) patient call system and inpatient fall occurrence. Participants were recruited from 4 wards of the Gastroenterology and Hepatology Department of a medical center in Taiwan over a 12-month period. The department's acquisition of the smart system, prompted by rising fall rates and a high incidence of false alarms with the traditional system, provided a timely opportunity for a convenient sample. This enabled us to investigate the impact of an IoT patient care system on patient safety. With the institutional review board (IRB) waiving the requirement for informed consent, we recruited all patients admitted to this department during the study period, resulting in a total of 1300 participants. The sample size was determined prior to recruitment, with a significance level set at 5% and a desired power level of 80%. Recruitment concluded upon reaching this predetermined sample size. The inclusion criteria were patients administrated to 4 wards of the Gastroenterology and Hepatology Department who were aged over 20 years and who spoke Chinese or Taiwanese. Exclusion criteria were patients who were bedridden during hospitalization, and those with medical conditions that would preclude the use of the SPCS, for example, patients using an air cushion bed.

Fall Risk and Injury Assessments

Apart from the different patient care systems, participants in all 4 wards received standard care following departmental health care guidelines. To compare the characteristics of participants from the TPCS and SPCS wards, we measured their age, sex, and length of stay in the ward together with a fall risk assessment. In the Gastroenterology and Hepatology Department, health care staff were required to complete the fall risk assessment test for all patients using comprehensive fall risk scores within 24 hours after admission. The comprehensive fall risk scores measured were based on scores of the St Thomas's Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) plus the polypharmacy situation. The STRATIFY instrument is a simple fall risk-assessment tool consisting of 5 items: history of falling, patient agitation, visual impairment affecting everyday function, need for frequent toileting, and transfer ability and mobility [27]. Although STRATIFY has been widely used to assess fall risks since its development, its ability to predict falls among older adults is unsatisfactory. A multicenter study that recruited 2568 patients showed that although the STRATIFY showed satisfactory sensitivity (≥84%) and a high negative predictive value (≥99%) for the total sample, the sensitivity (approximately 52%-69%) and specificity (approximately 55%-71%) for patients admitted to geriatric wards and for patients aged 75 years and older were moderate to low [28]. For the comprehensive fall risk scores of the site of this study, the polypharmacy status (yes or no, with "yes" defined as using 4 or more different medicines when recruited)

XSL•FO
RenderX

was added, which was shown to be a significant predictor in previous studies [29]. At the study site, a patient was considered to have polypharmacy if he or she used 4 or more different medicines at the same time. The total comprehensive fall risk score corresponded to the sum of the STRATIFY score and the polypharmacy status (yes=1 and no=0) and ranged from 0 to 6. A score of 3 or greater was accepted as a high risk of falls.

According to the Taiwan Quality Indicator Project followed by most hospitals in Taiwan, the severity classification of fall injuries includes 3 grades: grade 1 (minor injuries such as scratches, abrasions, small skin tears, or cuts that only require observation or minimal treatment); grade 2 (injuries such as sprains, large or deep cuts, or small abrasions that require medical or nursing intervention such as ice, bandaging, stitching, or splinting); and grade 3 (injuries such as fractures, loss of consciousness, or changes in body posture that require medical intervention and consultation, and may significantly affect the patient's condition or lead to an extended hospital stay).

Instruments

The TPCS is the current patient care system used in most wards at this medical center and has been used for over 30 years. The SPCS was set up in 1 ward of the Gastroenterology and Hepatology Department, and all relevant health care staff in that ward were subsequently trained to use this system. Both the TPCS and SPCS were used as systems for receiving and managing patient calls or bed-exit alerts. Details of the differences between the TPCS and SPCS were described in a previously published article [30]. The SPCS, which is an integrated IoT patient care system, comprises a motion-sensing mattress for bed-exit monitoring, a patient call button, the health care staff's mobile devices, and the health care staff scheduling system. Integrated patient care provided by the SPCS aims to minimize hospital-acquired falls and reduce the care burden on health care staff.

The signal transduction pathway of the SPCS is shown in Figure 1. In this pathway, there are several advantages of the SPCS compared to the TPCS. First, in the SPCS, the type of patient calls such as normal, emergency, and bed-exit, can be recognized, and a signal is sent to health care staff's mobile devices to help them receive the alert without the limitation of location. Second, the information exchange system in SPCS is connected to the scheduling system in the hospital so that only the health care staff on duty and responsible for a certain patient will receive such alarms. Third, this system also allows immediate communication between health care staff and patients. When the health care staff receive an alarm, they can directly speak to the patient, when necessary, by phone, and the patient can hear them by the speaker near the call button or bed. For example, if a health care staff member receives a bed-exit alert but cannot immediately provide assistance, he or she can answer the alarm by phone and ask the patient not to leave the bed until he or she comes to assist.

Figure 1. The signal transduction pathway of the smart patient care system.

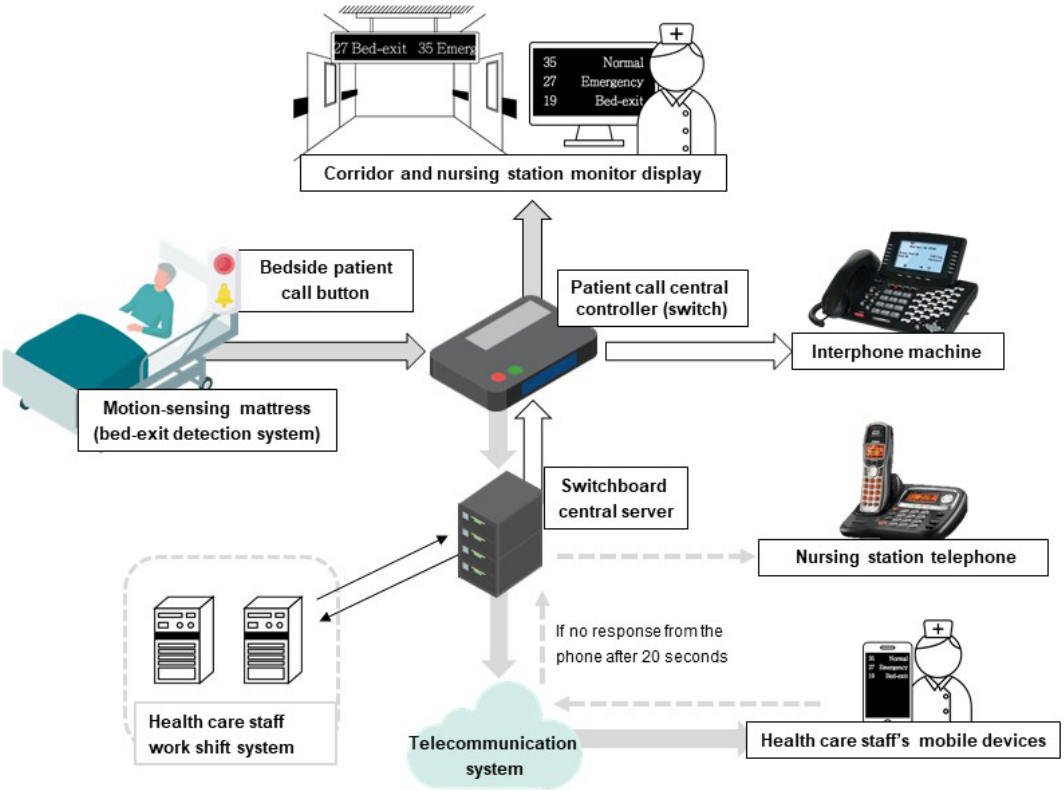

The bed-exit alert is used for patients at high risk of falling. All call alarms can be canceled at a distance by monitors or phones, except for the bed-exit alert which needs to be canceled by pressing a bedside button to ensure that actual assistance has arrived. In wards with TPCSs, a movable sensing mat was used for the bed-exit alert which was activated in the same way as patient calls. Again, only the bed number is shown, and health care staff are unable to tell whether it is a bed alarm or a patient

XSL•FO
RenderX

call. In wards with SPCSs, a motion-sensing mattress plays an important role in notifying health care staff of a patient's real-time in-bed positions. The IT structure of the fall-preventing motion-sensing mattress system in SPCS consists of 30 sensing areas, a sensor board, a wireless router, and a server (Figure 2). Sensing data are collected when users lie or sit in bed via the 30 sensing areas of membrane switches and then sent to the sensor board which contains a multiprocessor communications unit and a Bluetooth Low Energy module.

Figure 2. The information technology structure of the bed-exit detection system in the smart patient care system. API: application programming interface; APNS: Apple Push Notification service; AWS: Amazon Web Services; BLE: Bluetooth Low Energy; FCM: Firebase Cloud Messaging; MCU: multiprocessor communications unit; MQTT: message queuing telemetry transport.

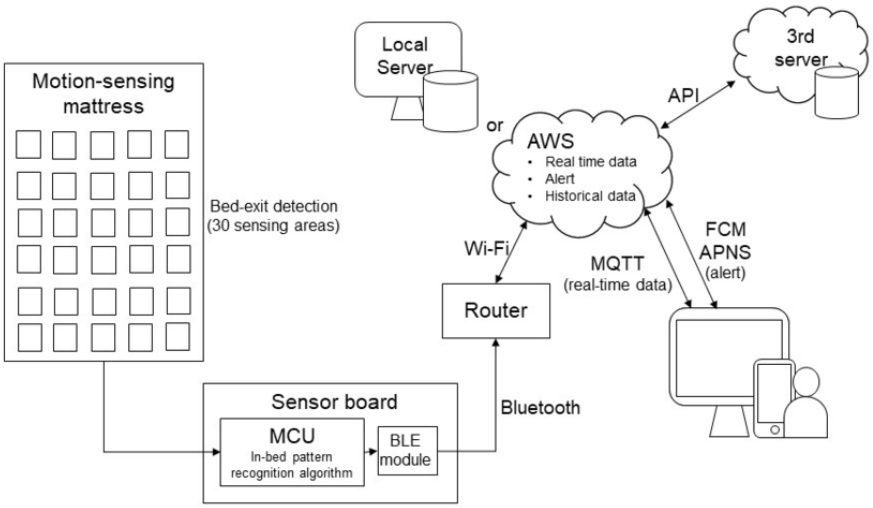

The motion-sensing mattress uses a multilayer perceptron neural network, a machine learning method, to generate the algorithm to identify a user's position, including on-bed, bed-edge, and off-bed. The final algorithm is loaded onto the sensor board, and position pattern results are sent to the wireless router by the Bluetooth Low Energy module and then sent to the database by Wi-Fi. In this system, Amazon Web Services or sometimes a local server is normally used as the server, and a third server can also access the data via an application programming interface with permission. On one hand, the server sends position pattern results to health care staff's mobile devices or computers for real-time event display by message queuing telemetry transport and sends the real-time position alert by Firebase Cloud Messaging or Apple Push Notification service. On the other hand, the server is used as the database for further analysis of historical data or user or device information input and settings.

Statistical Analysis

Descriptive analyses were used to describe the characteristics of participants using the mean, frequency, and percentage depending on the variables. Chi-square tests, Fisher exact tests, and 2-tailed *t* tests were also conducted to compare participant characteristics in the TPCS and SPCS groups. Fisher exact tests were used to compare whether there was a significant difference in the occurrence of falling between the 2 groups. In addition, unadjusted and adjusted logistic regression models were used to investigate the effect of the SPCS on fall prevention during hospitalization. The confounding variables included in the regression model were age, sex, length of hospital stay, and fall risk. The Hosmer-Lemeshow goodness-of-fit test was used to assess the adequacy of the logistic models, and the variance inflation factor was used to assess multicollinearity [31,32]. All data analyses were performed using the statistical software package Stata statistical software (version 16; StataCorp). A nominal significance level of .05 and a power of 80% were used throughout the analysis.

Ethical Considerations

The study procedures were reviewed and approved by the IRB of Taipei Veterans General Hospital, Taipei, Taiwan (2017-07-017B). The IRB waived the requirement for informed consent. Additionally, all data collected for this study were anonymized to ensure the privacy and confidentiality of participants. No compensation was provided to participants for their involvement in the study.

Results

In total, 1300 patients (from 4 wards of the Gastroenterology and Hepatology Department) were recruited, followed up, and analyzed in this study (Figure 3). Participants' demographic characteristics and the incidence of bedside falls were compared between the TPCS and SPCS wards (Table 1). The average age was 67.02 years in the TPCS ward and 67.81 years in the SPCS ward. There was no significant difference in the age distribution between the TPCS and SPCS wards but with some difference in sex. In the TPCS ward, approximately 466 (71.7%) participants were male, compared to about 305 (46.9%) participants in the SPCS ward. Among all participants, 733 (56.4%) had a stay of less than 1 week, while 246 (18.7%) stayed for more than 2 weeks, with no significant differences between the 2 groups. Fall risks of the 2 groups were found to significantly differ, with 320 participants (49.2%) recognized as being at some level of risk in the TPCS ward, whereas in the SPCS ward, 365 participants (56.2%) were considered at risk. The total bedside fall incidence during hospitalization was 0.7% (n=9), with 8 (1.2%) in the TPCS ward and 1 (0.1%) in the

XSL•FO
RenderX

SPCS ward. The participants in the SPCS group who fell had no clinical injury. Among the 8 participants who fell in the TPCS ward, 3 had no clinical injury, 3 had a grade 1 injury, 1 had a grade 2 injury, and 1 had a grade 3 injury. Results from Fisher exact test suggested a significant difference in the fall incidences between the TPCS and SPCS wards.

Figure 3. Flow diagram of participants. SPCS: smart patient care system; TPCS: traditional patient care system.

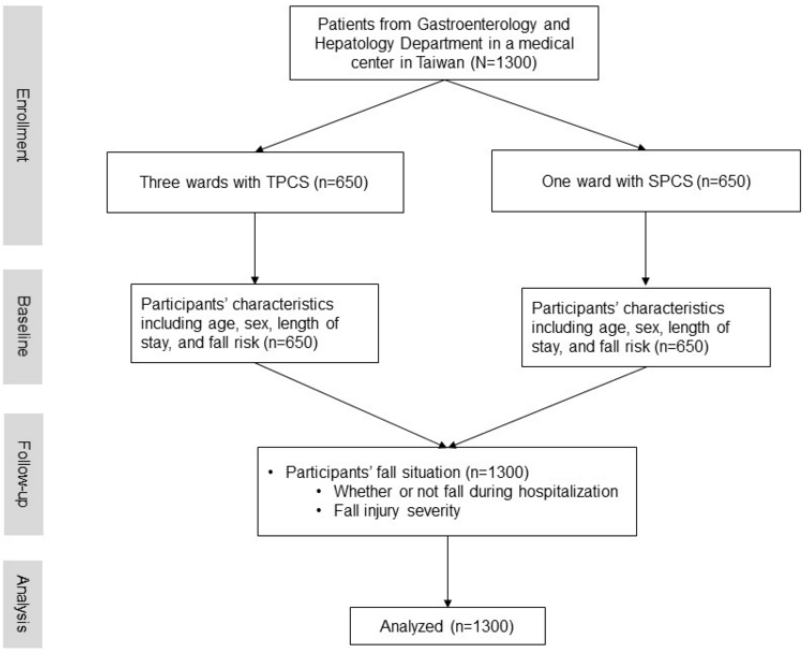

Table 1. Characteristics of participants (N=1300).

	Total	TPCS[a] (n=650)	SPCS[b] (n=650)	*P* value
Age (years), mean (SD)	66.93 (14.97)	67.02 (14.92)	67.81 (15.01)	.46
Sex, n (%)				<.001
Male	771 (59.3)	466 (71.7)	305 (46.9)	
Female	529 (40.7)	184 (28.3)	345 (53.1)	
Length of stay (days), n (%)				.06
<7	733 (56.4)	348 (53.6)	385 (59.2)	
8-14	324 (24.9)	166 (25.5)	158 (24.3)	
>15	243 (18.7)	136 (20.9)	107 (16.5)	
Fall risk scores, n (%)				.02
0	615 (47.3)	330 (50.8)	285 (43.8)	
1	150 (11.6)	79 (12.1)	71 (10.9)	
2	354 (27.2)	165 (25.4)	189 (29.1)	
>3	181 (13.9)	76 (11.7)	105 (16.2)	
Fall, n (%)				.04
No	1291 (99.3)	642 (98.8)	649 (99.9)	
Yes	9 (0.7)	8 (1.2)	1 (0.1)	

[a]TPCS: traditional patient care system.

[b]SPCS: smart patient care system.

In the SPCS ward, the IoT patient care system equipped with a 3-stage bed-exit alarm system detected an average of 13.5 potential falls per day, with no false alarms recorded throughout the study period. In contrast, the TPCS wards issued approximately 11 alarms daily, of which roughly 4 were false, effectively identifying 7 potential falls using the traditional system. The IoT system enhanced operational efficiency by transmitting alarm types directly to health care staff's mobile devices and monitors located in the corridors. This capability reduced the proportion of alarms addressed at the nursing station

XSL•FO
RenderX

to 10.9% in the SPCS wards, with the majority being managed proximally to the health care staff's locations at the time of alert. This system design significantly reduced workflow interruptions compared to the TPCS wards, where approximately 28.2% of the alarms required health care staff to communicate with patients via the nursing station telephone or return to verify alarm types, thus disrupting other duties.

We used unadjusted and adjusted logistic regression models to investigate the relationship of the use of the SPCS with falling (Table 2). We found that the likelihood of falling in the SPCS ward was reduced by 88% (odds ratio [OR] 0.12, 95% CI 0.02-0.99; *P*=.049) compared to the TPCS ward. After adjusting for confounding variables, this result remained significant and showed that the SPCS was associated with an 88% reduced risk of falling (OR 0.12, 95% CI 0.01-0.97; *P*=.047). In addition, with one more score of the comprehensive fall risk scores, the likelihood of actual falling was significantly increased 2.88 times (OR 2.88, 95% CI 1.28-6.48; *P*=.01) even after adjusting for confounding variables. Results of the Hosmer-Lemeshow goodness-of-fit tests for the adjusted logistic model indicated that it was a good fit for the data. Variance inflation factor values also indicated a low degree of multicollinearity.

Table 2. Unadjusted and adjusted OR[a] of fall according to the wards and participants' characteristics (N=1300).

	Unadjusted		Adjusted	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Group				
TPCS[b]	1.00 (—[c])	—	1.00 (—)	—
SPCS[d]	0.12 (0.02-0.99)	.049	0.12 (0.01-0.97)	.047
Fall risk scores[e]	2.96 (1.39-6.33)	.005	2.88 (1.28-6.48)	.01
Age	1.05 (1.00-1.11)	.048	1.03 (0.97-1.08)	.33
Sex				
Male	1.00 (—)	—	1.00 (—)	—
Female	0.41 (0.09-2.00)	.27	0.55 (0.11-2.77)	.47
Length of stay (days)				
<7	1.00 (—)	—	1.00 (—)	—
8-14	2.27 (0.46-11.33)	.32	1.35 (0.26-7.03)	.72
>15	3.04 (0.61-15.17)	.18	1.37 (0.26-7.18)	.71

[a]OR: odds ratio.

[b]TPCS: traditional patient care system.

[c]—: not applicable/not available.

[d]SPCS: smart patient care system.

[e]The total comprehensive fall risk score corresponds to the sum of St Thomas's Risk Assessment Tool in Falling Elderly Inpatients score and the polypharmacy status (yes=1, no=0) and ranges between 0 and 6. A score of 3 points or higher predicts a higher likelihood of falling.

Discussion

Principal Findings

This study investigated an integrated IoT patient care system's effect on inpatient fall prevention with a large prospective cohort. The SPCS is an IoT patient care system that consists of bed-exit monitoring by a motion-sensing mattress, patient call button, health care staff's mobile devices, and health care staff scheduling system for integrated patient care aimed at reducing hospital-acquired falls and health care staff's care burdens. The advantages of the SPCS were that it specifies the source of a signal as general, emergency, and bed-exit alerts; the health care staff can immediately receive an alarm and communicate with the patient via a mobile device. The 3-stage bed-exit alerts of the motion-sensing mattress help the health care staff determine a patient's intention of leaving the bed and offering help if necessary. We found that the incidence of hospital-acquired falls was significantly reduced by 88% in the ward using the SPCS compared to that using the TPCS.

It is common for falls to occur around beds in both hospital and household settings [1-3,33,34]. Falls also frequently occur between 5 PM and 7 AM in rooms with lower staffing levels [19]. Similarly, a recent report from the Department of Health in Taiwan showed that of all inpatient falls, approximately 78.9% occurred between 6 PM and 6 AM, and around one-quarter of falls were related to getting in or out of bed [35]. Increasing the human resources or workload of the health care staff is less likely to prevent bed-related falls given the consideration that human resources are most likely to be short in most conditions [36]. An alternative and rapidly emerging strategy for fall prevention is bed-exit detection systems that provide a possible solution by detecting and alerting patients and health care staff by providing alarms when a patient is at high risk of falling when trying to leave their bed unassisted [14,37,38].

XSL•FO
RenderX

The bed-exit detection system is commonly used in clinical practice, and studies even observed a significant increase in their use after the US Centers for Medicare and Medicaid Services stopped paying for preventable hospital-acquired falls and shifted the burden to hospitals [39]. Bed-exit detection systems can be categorized as ambient, wearable types and various forms of mattress or pad systems, wearable devices, and video systems using infrared sensors or cameras. Although wearable devices are usually portable and easy to use, they are usually designed to detect a fall event instead of predicting it and sending an alert before the fall happens, which limits their application in fall prevention and makes them less effective in preventing falls [40-42]. Although some researchers have been trying to design body-worn devices that detect the rising attempts, patients' adherence to wearing such devices remains a barrier [22]. Video monitoring is another kind of bed-exit detection system to predict the occurrence of falls, the acceptance of which by the public, however, is generally lower than 50% in various settings, let alone considering potential legal issues of such a system [43].

With less adherence and acceptance problems, mattress or pad systems provide a proper solution to reduce falls by sending the health care staff alerts when at-risk patients attempt to leave their beds without assistance (nonrestrictive). Bed-exit alerts usually use pressure-sensitive sensors to collect pressure signals of a patient's weight, and alarms are triggered when a patient tries to get up from the bed and the pressure on the sensor is relieved [37,44]. Despite their widespread use, the evidence for the effectiveness of bed-exit alerts in preventing falls is insufficient and inconsistent [16,20,45]. Two before-and-after studies showed that bed-exit alerts were associated with reduced fall events by 18% and 54% [23]. One cluster randomized trial also showed their significant effectiveness [46]. In contrast, many other studies did not find a significant reduction in fall rates when using bed-exit alerts [20,21,24,42,46]. A clinical trial study that recruited 70 participants from a geriatric ward showed that a bed-exit alert using pressure-sensitive sensors had no significant effect on fall prevention [24]. A more recent clustered randomized trial that recruited 27,672 inpatients from 16 nursing units in an urban community hospital in the United States suggested no significant effect on hospital-acquired falls or the use of physical restraints [21]. Another clustered randomized study showed no effect [42].

In this study, we observed a significant association between the SPCS and a reduction in hospital-acquired falls. Apart from the fall rate being low and it being hard to detect a significant effect [21,24,44], there are several plausible explanations for why the SPCS was significantly effective in fall prevention. First, it is common for pressure-sensitive sensors to send alarm signals after a patient has already exited the bed, which leads to limited time for health care staff to come and assist so that a fall is more likely to occur [43]. Moreover, some sensors are designed to delay the alert for about 2 seconds after detecting the pressure change to reduce false alarms, which, however, could also be a barrier to providing timely assistance by health care staff [44]. The system in this study adopted an algorithm that allows the mattress to send signals in 3 stages: a patient sits on the bed, sits on the edge of the bed, and exits the bed. The health care staff can choose whether to receive the early-stage notifications according to a patient's comprehensive situation of fall risks and family assistance. This graduated sequence of bed-exit notifications provides health care staff additional time to respond to the alerts. A pilot study that investigated the effect of a medical IoT system for fall prevention using similar technology to provide early notifications when a patient attempted to leave the bed showed no bed-related falls during 234 patient days [47].

One of the principal challenges encountered in implementing this study was designing the IoT system that could be integrated into health care staff's daily routines. The goal was to enhance patient safety through the use of this system, rather than imposing additional burdens that might lead to its eventual abandonment despite its innovative potential. In addition to sending alerts to a dashboard at the nursing station or a monitor in the corridor, the SPCS can also notify health care staff via an app on their mobile devices so that they can receive the alerts wherever and whenever it is necessary. The system is connected to the internet at the hospital as well as the health care staff scheduling system so that the health care staff will only receive an alert when they are on duty. Similarly, a study in an acute-care setting showed that the use of an alarm system with a portable beeper carried by nursing staff was associated with a reduction in fall incidence [23]. Although there is a clinical strategy to move patients at high risk of falling to units near the nursing station for better monitoring, on one hand, this could increase the workload of administration arrangements, and on the other hand, there is a lack of evidence of its effectiveness. With a portable device that can receive alerts, patients' locations would not influence the nursing staff's response times [47]. With the IoT system, the portable and immediate reception of an alert would help health care staff decrease the response time and provide help in a timely manner whenever necessary without the patient being placed near the nursing station.

Furthermore, it was shown that health care staff's experiences with bed-exit alarm systems influenced their effectiveness. The SPCS has an immediate communication function through which health care staff can talk to patients on a mobile device once they receive an alert. Health care staff have multiple responsibilities and need to decide the priority of a new task when it comes in. Without enough information on the source and emergency level of a patient call, health care staff are more likely to assume it is nonurgent and choose to handle it after they finish what they are doing, which may lead to a fall when immediate assistance is truly needed [48]. For example, when health care staff receive a bed-exit alert, but they cannot drop what they are doing, it will help if they can talk to the patient at high risk of falling to instruct them to wait until they can come to assist or maybe ask other colleagues to help handle it. Otherwise, a patient call signal can more easily be ignored especially when the purpose of the signal is not specific [49]. With the immediate communication function of the SPCS, bed exits are no longer an extra burden but are well integrated into the health care process.

Last but not least, false alarms are commonly reported with existing bed-exit alert systems. It was shown that false alarms accounted for greater than 80% of all bed-exit alerts, leading to

XSL•FO
RenderX

alarm fatigue, which can increase the health care staff's work burden, lower their willingness to use the system, and ultimately result in patients falling [25,50,51]. Excessive false alarms may lead to alarm fatigue in which health care staff will occasionally ignore an alarm. A multicenter study showed that over 70% of monitor alarms were false alarms, and only 5.9% of all alarms received a response from health care staff [52]. In addition, false alarms may frequently disrupt patients' sleep and negatively influence their recovery [53]. Our previous results showed that it took around 2.5 times longer for health care staff to respond to bed-exit alerts in TPCS wards than that in SPCS wards [30]. The high predictive positive value of the SPCS significantly improved health care staff's willingness to use the system, which may be another reason that the SPCS had a positive effect on fall prevention [54].

Strengths and Limitations

This was a large prospective cohort study that examined the association between an IoT patient care system and inpatient falls. Despite this, it has some limitations. First, the study was not population-based and used a convenience sample, which may introduce selection bias and limit the generalizability of the findings to a broader population. Additionally, the sample was drawn from a single site, further restricting the applicability of the results to other settings. Generalization of the findings to the entire inpatient population would be limited. The proportion of female participants was higher in the SPCS group than in the TPCS group. The existing evidence shows that sex is not statistically associated with inpatient falls [55,56]. In addition, health care staff were aware that the effect of SPCS on falls was being studied, and it is possible that they were more vigilant when the bed alarms occurred according to the Hawthorne effect [57]. However, this study lasted for around 1 year, and it is likely that the active responses to patient calls including bed-exit alerts were integrated into their daily practice, which was one of the purposes of the system. Another potential limitation of this study was the use of logistic regression to analyze fall events, which are relatively rare in our dataset. Logistic regression models can sometimes underestimate the likelihood of rare events occurring. This underestimation may affect the accuracy of the predicted probabilities; therefore, our results should be interpreted with caution.

Conclusions

Bed-exit detection systems have been increasingly used around the world for inpatient fall prevention. They are also commonly combined with patient care systems to help notify health care staff of bed-exit alerts. Unfortunately, many of the current bed-exit detection and patient care systems are connected in a traditional signal transduction way instead of a more efficient method using IoT techniques. The SPCS introduced in this study integrated a motion-sensing mattress with bed-exit alerts as well as a patient call button, health care staff's mobile devices, and a health care staff scheduling system. The design of this IoT patient care system allows health care staff to immediately receive necessary bed-exits alerts by mobile devices without location limitations, which will help them provide timely assistance for at-risk patients or easily cope with an alarm while they are in the middle of other duties. As a result, the efficient and resilient responses to bed-exit detection are expected to help with fall prevention in hospitals. Future research should focus on refining IoT integration techniques within health care systems, enhancing predictive analytics for early detection, expanding interoperability standards, and assessing the longitudinal impact of IoT-enabled systems on patient safety and care efficiency, particularly in reducing hospital-acquired falls.

Acknowledgments

This work was supported by Taipei Medical University (grant TMU111-AE1-B20), Taipei Veterans General Hospital (grant V107A-029), and the National Science and Technology Council of Taiwan (grant NSTC112-2622-E-038-007; MOST-110-2923-E-155-002-MY3).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MHW developed the concepts and methodologies. SL, CWL, SHT, CYC, and YFW participated in the study design and data collection. DB supervised the project, advised on its concepts and methodologies, and conducted the statistical analysis with PYC and YLH. MHW, PYC, YLH, and DB drafted the manuscript. All authors revised the manuscript and approved the submitted version.

Conflicts of Interest

None declared.

References

1. Qian XX, Chen Z, Fong DYT, Ho M, Chau PH. Post-hospital falls incidence and risk factors among older adults: a systematic review and meta-analysis. Age Ageing 2022;51(1):afab209. [doi: 10.1093/ageing/afab209] [Medline: 34718373]

XSL•FO
RenderX

2. Hoffman G, Franco N, Perloff J, Lynn J, Okoye S, Min L. Incidence of and county variation in fall injuries in US residents aged 65 years or older, 2016-2019. JAMA Netw Open 2022;5(2):e2148007 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.48007] [Medline: 35147689]
3. Bouldin ELD, Andresen EM, Dunton NE, Simon M, Waters TM, Liu M, et al. Falls among adult patients hospitalized in the United States: prevalence and trends. J Patient Saf 2013;9(1):13-17 [FREE Full text] [doi: 10.1097/PTS.0b013e3182699b64] [Medline: 23143749]
4. Dunne TJ, Gaboury I, Ashe MC. Falls in hospital increase length of stay regardless of degree of harm. J Eval Clin Pract 2014;20(4):396-400. [doi: 10.1111/jep.12144] [Medline: 24814338]
5. Fields J, Alturkistani T, Kumar N, Kanuri A, Salem DN, Munn S, et al. Prevalence and cost of imaging in inpatient falls: the rising cost of falling. Clinicoecon Outcomes Res 2015;7:281-286 [FREE Full text] [doi: 10.2147/CEOR.S80104] [Medline: 26082653]
6. Spetz J, Jacobs J, Hatler C. Cost effectiveness of a medical vigilance system to reduce patient falls. Nurs Econ 2007;25(6):333-338, 352. [Medline: 18240834]
7. Florence CS, Bergen G, Atherly A, Burns E, Stevens J, Drake C. Medical costs of fatal and nonfatal falls in older adults. J Am Geriatr Soc 2018;66(4):693-698 [FREE Full text] [doi: 10.1111/jgs.15304] [Medline: 29512120]
8. Rosenthal MB. Nonpayment for performance? Medicare's new reimbursement rule. N Engl J Med 2007;357(16):1573-1575. [doi: 10.1056/NEJMp078184] [Medline: 17942869]
9. Morris ME, Webster K, Jones C, Hill AM, Haines T, McPhail S, et al. Interventions to reduce falls in hospitals: a systematic review and meta-analysis. Age Ageing 2022;51(5):afac077 [FREE Full text] [doi: 10.1093/ageing/afac077] [Medline: 35524748]
10. Thomas E, Battaglia G, Patti A, Brusa J, Leonardi V, Palma A, et al. Physical activity programs for balance and fall prevention in elderly: a systematic review. Medicine 2019;98(27):e16218 [FREE Full text] [doi: 10.1097/MD.0000000000016218] [Medline: 31277132]
11. Cameron ID, Dyer SM, Panagoda CE, Murray GR, Hill KD, Cumming RG, et al. Interventions for preventing falls in older people in care facilities and hospitals. Cochrane Database Syst Rev 2018;9(9):CD005465 [FREE Full text] [doi: 10.1002/14651858.CD005465.pub4] [Medline: 30191554]
12. Tanwar R, Nandal N, Zamani M, Manaf AA. Pathway of trends and technologies in fall detection: a systematic review. Healthcare 2022;10(1):172 [FREE Full text] [doi: 10.3390/healthcare10010172] [Medline: 35052335]
13. Ren L, Peng Y. Research of fall detection and fall prevention technologies: a systematic review. IEEE Access 2019;7:77702-77722. [doi: 10.1109/access.2019.2922708]
14. Jähne-Raden N, Kulau U, Marschollek M, Wolf K. INBED: a highly specialized system for bed-exit-detection and fall prevention on a geriatric ward. Sensors 2019;19(5):1017 [FREE Full text] [doi: 10.3390/s19051017] [Medline: 30818871]
15. Gottlieb S, Makower J. A role for entrepreneurs: an observation on lowering healthcare costs via technology innovation. Am J Prev Med 2013;44:S43-S47 [FREE Full text] [doi: 10.1016/j.amepre.2012.09.006] [Medline: 23195166]
16. Bruyneel M, Libert W, Ninane V. Detection of bed-exit events using a new wireless bed monitoring assistance. Int J Med Inform 2011;80(2):127-132. [doi: 10.1016/j.ijmedinf.2010.10.007] [Medline: 21051278]
17. Tzeng HM, Yin CY. The extrinsic risk factors for inpatient falls in hospital patient rooms. J Nurs Care Qual 2008;23(3):233-241. [doi: 10.1097/01.NCQ.0000324588.91837.84] [Medline: 18562866]
18. Hitcho EB, Krauss MJ, Birge S, Claiborne Dunagan W, Fischer I, Johnson S, et al. Characteristics and circumstances of falls in a hospital setting: a prospective analysis. J Gen Intern Med 2004;19(7):732-739 [FREE Full text] [doi: 10.1111/j.1525-1497.2004.30387.x] [Medline: 15209586]
19. Ranasinghe DC, Shinmoto Torres RL, Hill K, Visvanathan R. Low cost and batteryless sensor-enabled radio frequency identification tag based approaches to identify patient bed entry and exit posture transitions. Gait Posture 2014;39(1):118-123. [doi: 10.1016/j.gaitpost.2013.06.009] [Medline: 23850327]
20. Capezuti E, Brush BL, Lane S, Rabinowitz HU, Secic M. Bed-exit alarm effectiveness. Arch Gerontol Geriatr 2009;49(1):27-31 [FREE Full text] [doi: 10.1016/j.archger.2008.04.007] [Medline: 18508138]
21. Shorr RI, Chandler AM, Mion LC, Waters TM, Liu M, Daniels MJ, et al. Effects of an intervention to increase bed alarm use to prevent falls in hospitalized patients: a cluster randomized trial. Ann Intern Med 2012;157(10):692-699 [FREE Full text] [doi: 10.7326/0003-4819-157-10-201211200-00005] [Medline: 23165660]
22. Wolf KH, Hetzer K, zu Schwabedissen HM, Wiese B, Marschollek M. Development and pilot study of a bed-exit alarm based on a body-worn accelerometer. Z Gerontol Geriatr 2013;46(8):727-733. [doi: 10.1007/s00391-013-0560-2] [Medline: 24271253]
23. Diduszyn J, Hofmann MT, Naglak M, Smith DG. Use of a wireless nurse alert fall monitor to prevent inpatient falls. J Sci Commun 2008;15(6):293.
24. Tideiksaar R, Feiner CF, Maby J. Falls prevention: the efficacy of a bed alarm system in an acute-care setting. Mt Sinai J Med 1993;60(6):522-527. [Medline: 8121429]
25. Siebig S, Kuhls S, Imhoff M, Gather U, Schölmerich J, Wrede CE. Intensive care unit alarms—how many do we need? Crit Care Med 2010;38(2):451-456. [doi: 10.1097/CCM.0b013e3181cb0888] [Medline: 20016379]

XSL•FO
RenderX

26. Cvach M. Monitor alarm fatigue: an integrative review. Biomed Instrum Technol 2012;46(4):268-277. [doi: 10.2345/0899-8205-46.4.268] [Medline: 22839984]
27. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 1997;315(7115):1049-1053 [FREE Full text] [doi: 10.1136/bmj.315.7115.1049] [Medline: 9366729]
28. Milisen K, Staelens N, Schwendimann R, De Paepe L, Verhaeghe J, Braes T, et al. Fall prediction in inpatients by bedside nurses using the St. Thomas's Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) instrument: a multicenter study. J Am Geriatr Soc 2007;55(5):725-733. [doi: 10.1111/j.1532-5415.2007.01151.x] [Medline: 17493192]
29. Dhalwani NN, Fahami R, Sathanapally H, Seidu S, Davies MJ, Khunti K. Association between polypharmacy and falls in older adults: a longitudinal study from England. BMJ Open 2017;7(10):e016358 [FREE Full text] [doi: 10.1136/bmjopen-2017-016358] [Medline: 29042378]
30. Wen MH, Bai D, Lin S, Chu CJ, Hsu YL. Implementation and experience of an innovative smart patient care system: a cross-sectional study. BMC Health Serv Res 2022;22(1):126 [FREE Full text] [doi: 10.1186/s12913-022-07511-7] [Medline: 35093036]
31. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16(9):965-980. [doi: 10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.0.co;2-o] [Medline: 9160492]
32. Hair JF, Ringle CM, Sarstedt M. PLS-SEM: indeed a silver bullet. J Mark Theory Pract 2011;19(2):139-151. [doi: 10.2753/MTP1069-6679190202]
33. Rapp K, Becker C, Cameron ID, König HH, Büchele G. Epidemiology of falls in residential aged care: analysis of more than 70,000 falls from residents of bavarian nursing homes. J Am Med Dir Assoc 2012;13(2):187.e1-187.e6. [doi: 10.1016/j.jamda.2011.06.011] [Medline: 21816682]
34. Hanger HC, Ball MC, Wood LA. An analysis of falls in the hospital: Can we do without bedrails? J Am Geriatr Soc 1999;47(5):529-531. [doi: 10.1111/j.1532-5415.1999.tb02565.x] [Medline: 10323644]
35. Department of Health of Taiwan. Taiwan patient safety reporting system annual report 2021. URL: https://www.patientsafety.mohw.gov.tw/xmfile?xsmsid=0M098412292297480494 [accessed 2023-10-16]
36. Chiang HY, Hsiao YC, Lee HF. Predictors of hospital nurses' safety practices: work environment, workload, job satisfaction, and error reporting. J Nurs Care Qual 2017;32(4):359-368. [doi: 10.1097/NCQ.0000000000000240] [Medline: 27875383]
37. Kosse NM, Brands K, Bauer JM, Hortobagyi T, Lamoth CJC. Sensor technologies aiming at fall prevention in institutionalized old adults: a synthesis of current knowledge. Int J Med Inform 2013;82(9):743-752. [doi: 10.1016/j.ijmedinf.2013.06.001] [Medline: 23845790]
38. Ward-Smith P, Barrett L, Rayson K, Govro K. Effectiveness of a bed alarm system to predict falls in an acute care setting. Clin Nurs Stud 2014;3(1):1-4. [doi: 10.5430/cns.v3n1p1]
39. Fehlberg EA, Lucero RJ, Weaver MT, McDaniel AM, Chandler AM, Richey PA, et al. Impact of the CMS no-pay policy on hospital-acquired fall prevention related practice patterns. Innov Aging 2017;1(3):igx036 [FREE Full text] [doi: 10.1093/geroni/igx036] [Medline: 29911187]
40. Nyan MN, Tay FEH, Murugasu E. A wearable system for pre-impact fall detection. J Biomech 2008;41(16):3475-3481. [doi: 10.1016/j.jbiomech.2008.08.009] [Medline: 18996529]
41. Jung S, Hong S, Kim J, Lee S, Hyeon T, Lee M, et al. Wearable fall detector using integrated sensors and energy devices. Sci Rep 2015;5:17081 [FREE Full text] [doi: 10.1038/srep17081] [Medline: 26597423]
42. Cumming RG, Sherrington C, Lord SR, Simpson JM, Vogler C, Cameron ID, et al. Cluster randomised trial of a targeted multifactorial intervention to prevent falls among older people in hospital. BMJ 2008;336(7647):758-760 [FREE Full text] [doi: 10.1136/bmj.39499.546030.BE] [Medline: 18332052]
43. Ziefle M, Himmel S, Wilkowska W. When your living space knows what you do: acceptance of medical home monitoring by different technologies. 2011 Presented at: Proceedings of the 7th Conference on Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society: Information Quality in e-Health; November 25-26, 2011; Berlin, Germany p. 25-26.
44. Kwok T, Mok F, Chien WT, Tam E. Does access to bed-chair pressure sensors reduce physical restraint use in the rehabilitative care setting? J Clin Nurs 2006;15(5):581-587. [doi: 10.1111/j.1365-2702.2006.01354.x] [Medline: 16629967]
45. Hilbe J, Schulc E, Linder B, Them C. Development and alarm threshold evaluation of a side rail integrated sensor technology for the prevention of falls. Int J Med Inform 2010;79(3):173-180. [doi: 10.1016/j.ijmedinf.2009.12.004] [Medline: 20083427]
46. Dykes PC, Carroll DL, Hurley A, Lipsitz S, Benoit A, Chang F, et al. Fall prevention in acute care hospitals: a randomized trial. JAMA 2010;304(17):1912-1918 [FREE Full text] [doi: 10.1001/jama.2010.1567] [Medline: 21045097]
47. Balaguera HU, Wise D, Ng CY, Tso H, Chiang W, Hutchinson AM, et al. Using a medical Intranet of Things system to prevent bed falls in an acute care hospital: a pilot study. J Med Internet Res 2017;19(5):e150 [FREE Full text] [doi: 10.2196/jmir.7131] [Medline: 28473306]
48. Graham KC, Cvach M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. Am J Crit Care 2010;19(1):28-34; quiz 35. [doi: 10.4037/ajcc2010651] [Medline: 20045845]

XSL•FO
RenderX

49. Ruppel H, De Vaux L, Cooper D, Kunz S, Duller B, Funk M. Testing physiologic monitor alarm customization software to reduce alarm rates and improve nurses' experience of alarms in a medical intensive care unit. PLoS One 2018;13(10):e0205901 [FREE Full text] [doi: 10.1371/journal.pone.0205901] [Medline: 30335824]
50. Atzema C, Schull MJ, Borgundvaag B, Slaughter GR, Lee CK. ALARMED: adverse events in low-risk patients with chest pain receiving continuous electrocardiographic monitoring in the emergency department. A pilot study. Am J Emerg Med 2006;24(1):62-67. [doi: 10.1016/j.ajem.2005.05.015] [Medline: 16338512]
51. Tsien CL, Fackler JC. Poor prognosis for existing monitors in the intensive care unit. Crit Care Med 1997;25(4):614-619. [doi: 10.1097/00003246-199704000-00010] [Medline: 9142025]
52. Chambrin MC, Ravaux P, Calvelo-Aros D, Jaborska A, Chopin C, Boniface B. Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis. Intensive Care Med 1999;25(12):1360-1366. [doi: 10.1007/s001340051082] [Medline: 10660842]
53. Matthews EE. Sleep disturbances and fatigue in critically ill patients. AACN Adv Crit Care 2011;22(3):204-224 [FREE Full text] [doi: 10.1097/NCI.0b013e31822052cb] [Medline: 21808157]
54. Murray GR, Cameron ID, Cumming RG. The consequences of falls in acute and subacute hospitals in Australia that cause proximal femoral fractures. J Am Geriatr Soc 2007;55(4):577-582. [doi: 10.1111/j.1532-5415.2007.01102.x] [Medline: 17397437]
55. Oliver D, Healey F, Haines TP. Preventing falls and fall-related injuries in hospitals. Clin Geriatr Med 2010;26(4):645-692. [doi: 10.1016/j.cger.2010.06.005] [Medline: 20934615]
56. Ganz DA, Bao Y, Shekelle PG, Rubenstein LZ. Will my patient fall? JAMA 2007;297(1):77-86. [doi: 10.1001/jama.297.1.77] [Medline: 17200478]
57. McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P. The hawthorne effect: a randomised, controlled trial. BMC Med Res Methodol 2007;7:30 [FREE Full text] [doi: 10.1186/1471-2288-7-30] [Medline: 17608932]

Abbreviations

IoT: Internet of Things
IRB: institutional review board
OR: odds ratio
SPCS: smart patient care system
STRATIFY: St Thomas's Risk Assessment Tool in Falling Elderly Inpatients
TPCS: traditional patient care system

Edited by A Mavragani; submitted 14.03.24; peer-reviewed by J Walsh, MII Alghamdi; comments to author 13.04.24; revised version received 17.04.24; accepted 23.08.24; published 03.10.24.

Please cite as:
Wen MH, Chen PY, Lin S, Lien CW, Tu SH, Chueh CY, Wu YF, Tan Cheng Kian K, Hsu YL, Bai D
Enhancing Patient Safety Through an Integrated Internet of Things Patient Care System: Large Quasi-Experimental Study on Fall Prevention
J Med Internet Res 2024;26:e58380
URL: https://www.jmir.org/2024/1/e58380
doi:10.2196/58380
PMID:39361417

©Ming-Huan Wen, Po-Yin Chen, Shirling Lin, Ching-Wen Lien, Sheng-Hsiang Tu, Ching-Yi Chueh, Ying-Fang Wu, Kelvin Tan Cheng Kian, Yeh-Liang Hsu, Dorothy Bai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effectiveness of a Digital Mental Fitness Program (Positive Intelligence) on Perceived Stress, Self-Compassion, and Ruminative Thinking of Occupational Therapy Undergraduate Students: Longitudinal Study

Shermain Puah[1], BSc, MSc; Ching Yee Pua[1], BSc; Jing Shi[2], BHSc, MSc(OT), PhD; Sok Mui Lim[1,2], BSc, GCHEd, PhD

[1]SIT Teaching and Learning Academy (STLA), Singapore Institute of Technology, Singapore, Singapore
[2]Health and Social Sciences, Singapore Institute of Technology, Singapore, Singapore

Corresponding Author:
Shermain Puah, BSc, MSc
SIT Teaching and Learning Academy (STLA)
Singapore Institute of Technology
1 Punggol Coast Road
Singapore, 828608
Singapore
Phone: 65 6592 7637
Email: shermain.puah@singaporetech.edu.sg

Abstract

Background: Health care students often endure numerous stressors throughout their undergraduate education that can have lasting negative effects on their mental well-being. Positive Intelligence (PQ) is a digital mental fitness program designed to enhance self-mastery and help individuals reach their potential by strengthening various “mental muscles.”

Objective: This study aims to evaluate the effectiveness of a 6-week app-delivered PQ program in reducing perceived stress, increasing self-compassion, and decreasing rumination tendencies among health care undergraduates. We hypothesized that students would show reductions in perceived stress, increases in self-compassion, and decreases in rumination tendencies by the end of the PQ program, compared with their preprogram scores. We adopted an exploratory approach for the 5-month follow-up due to the limited research consensus on the sustained effects of app-based programs over varying periods.

Methods: The PQ program includes weekly hour-long videos, weekly group meetings, and daily 15-minute app-guided practices. Participants were first-year students from the occupational therapy program at a university in Singapore. Participants completed surveys measuring self-compassion, perceived stress, and rumination levels before and after the PQ program, and again at a 5-month follow-up. Data were analyzed using repeated measures ANOVA to assess differences across the pretest, immediate posttest, and follow-up posttest.

Results: Out of 87 students enrolled in the study, the final sample consisted of 64 students (n=47, 73%, female; mean age 23 years, SD 5.06 years) with complete data. At the end of the 6 weeks, students exhibited significant increases in self-compassion (before the intervention: mean 3.07, SD 0.35; after the intervention: mean 3.34, SD 0.35; $P<.001$) and reductions in rumination tendencies (before the intervention: mean 3.57, SD 0.40; after the intervention: mean 3.27, SD 0.34; $P<.001$). However, no significant change in perceived stress levels was observed (before the intervention: 2.99, SD 0.14; after the intervention: mean 2.97, SD 0.16; $P=.50$). These effects were not influenced by the daily app-based practice of PQ exercises, and there were no sustained effects on self-compassion (mean 3.17, SD 0.27; $P=.09$) or rumination tendencies (mean 3.42, SD 0.38; $P=.06$) at the 5-month follow-up. Additionally, there was a significant increase in perceived stress at follow-up (mean 3.17, SD 0.21; $P<.001$) compared with pre- and postintervention levels.

Conclusions: The PQ program did not directly alter stress perceptions but may have reframed students’ automatic negative thought processes, increased their awareness of self-sabotaging behaviors, and enhanced their self-compassion while reducing ruminative thinking. These findings highlight the importance of self-awareness for students’ well-being. Students can benefit from practices such as mindfulness and peer discussions to enhance self-compassion and reduce rumination. Educators trained

XSL•FO
RenderX

in the PQ program can foster a supportive environment that encourages self-compassion, challenges negative self-talk, and helps students manage stress.

(J Med Internet Res 2024;26:e49505) doi:10.2196/49505

KEYWORDS

mental health; students; digital wellness; mobile health (mHealth); perceived stress; self-compassion; rumination

Introduction

Background

A growing body of literature highlights the increasing levels of stress, anxiety, and psychological distress experienced by undergraduate students in health care disciplines [1]. Common stressors during their education include clinical workload, expectations during clinical placements, academic program demands, insufficient support to manage both clinical and academic responsibilities, and the emotional toll of working with patients suffering from severe conditions [2-4]. For some students, the transition to university itself can be overwhelming [5].

The psychological and emotional distress resulting from these challenges and demands can significantly impact the mental well-being of health care students, not only during their education but also potentially extending into their transition into the workforce as health care professionals [6,7]. Given the potential impact on work-relevant functioning, helping students develop coping strategies to mitigate these effects throughout their professional lives is an important area of investigation [7,8]. Therefore, efforts should focus on teaching self-care and promoting well-being at the undergraduate level, supporting students in building habits that will help them manage stress and challenges effectively.

Positive Intelligence

Positive Intelligence (PQ), a method developed by Chamine [9] at Stanford University as a personal development tool, aims to build mastery over one's mind to fulfill one's potential (ie, mental fitness). The PQ program is designed to fit into the busy schedules of modern individuals by offering bite-sized practice exercises and videos over a short period (6 weeks). Although it has been applied to MBA (master of business administration) students, athletes, executive teams, and chief executive officers [10], its use in the fast-paced and demanding health care setting remains largely unexplored.

According to Chamine's findings [9], throughout life, every experience we have helps to develop and solidify certain automatic mental attitudes and habits. These mental habits can either work in our favor or against us, placing us in a state of "thriving" or "survival." Two opposing states of mind, namely, the "sage" and the "saboteurs," are in a constant tug of war, as they are wired for very different primary purposes. The "sage" is wired for creativity, empathy, expanding perspectives, and fostering positive emotions. By contrast, the "saboteurs" trigger survival instincts, narrow perspectives, and generate negative emotions. The risk of remaining under the influence of one's "saboteurs" is the missed opportunities that arise in challenging situations due to an inability to see beyond the negativity. The goal of PQ is to build mental fitness by strengthening "mental muscles"—choosing "sage" responses and intercepting the tendencies of one's "saboteurs."

The PQ program is unique in that it goes beyond training mindfulness. It incorporates additional components such as increasing awareness of one's inner critics (or "saboteurs"), learning to adopt a perspective that sees opportunities and gifts in challenges (the "sage" perspective), and participating in group support through "Pod" meetings. These elements work together to create a comprehensive and holistic experience. The PQ program spans 6 weeks and is primarily delivered through a mobile app, though some components may also be accessed via a web browser. A "muscle" score is cumulatively logged in the app whenever an individual practices the "self-command muscle" (ie, the ability to self-regulate thoughts and feelings), the "saboteur interceptor muscle" (ie, the ability to intercept and disarm the false messages of "saboteurs"), or the "sage muscle" (ie, the ability to use wisdom and insight to handle challenges) over the 6 weeks [11]. The implementation of the PQ program in this study is detailed in the "Methods" section.

According to Positive Intelligence [11], the PQ program has been utilized by over 500,000 individuals from diverse backgrounds and ages, across as many as 50 countries. However, there is limited scholarly research assessing the effectiveness of the PQ program in providing students with the mental fitness needed to manage stressors during their undergraduate education. We propose that the PQ program may effectively enhance health care students' self-compassion, reduce perceived stress, and decrease rumination tendencies. These variables—perceived stress, self-compassion, and rumination—were chosen for this study due to their critical roles in the psychological well-being of health care students and their importance in protecting against emotional distress [12-14]. Additionally, these variables align with the goals of the PQ program. The following section discusses the relevance of these variables in relation to the program and overall well-being.

Perceived Stress

According to Lazarus [15], perceived stress is defined as a subjective emotional experience in which an individual feels that the demands placed upon them exceed their ability to cope. An individual's perception of a stressful event is crucial, as it influences how uncontrollable and overwhelming they perceive the situation to be [16]. Among university students, elevated stress levels are more common in health care students compared with their non–health care counterparts, who are also at a higher risk of developing mental health issues [17].

Stress management is identified as a key area of impact when individuals practice PQ and learn to intercept their "saboteurs"

XSL•FO
RenderX

and activate their "sage" [11]. According to research conducted by the Positive Intelligence company [11], stress is viewed as a negative emotion resulting from the habitual activation of one's "saboteurs." These "saboteurs" are responsible for generating many negative emotions. While these emotions can be useful in alerting individuals to dangers or issues when activated briefly, they impair one's ability to think clearly and make optimal decisions [11]. By practicing the act of focusing attention on a particular sensation or object (the "self-command muscle") and intercepting one's inner critics (the "saboteur interceptor muscle"), individuals gain greater control over their mind and thoughts. This process enables them to adopt a more effective perspective (the "sage perspective").

Self-Compassion

Self-compassion involves extending empathy and support to oneself during times of suffering. Neff [18,19] conceptualized self-compassion as comprising 3 core components: kindness toward oneself versus self-judgment, a sense of common humanity versus isolation, and mindfulness versus overexaggeration when dealing with painful experiences. Our inner critics often lead us to harbor negative messages about ourselves, resulting in self-evaluation that is harsh and unkind. With self-compassion, we adopt a compassionate stance toward ourselves, allowing us to view our sufferings, lives, and experiences from a broader and more neutral perspective [20]. Existing literature consistently reports that higher levels of self-compassion are strongly associated with a lower risk of mental health issues and greater psychological flourishing [20].

The ability to intercept one's inner critics is essential for achieving greater self-acceptance and self-compassion. According to PQ, one prominent "saboteur" is the "judge." The "judge" manifests unconsciously when we evaluate ourselves, others, and the situations we encounter [9]. Although the primary role of the "judge" is to discern potentially harmful scenarios, habitual judgment can lead us to judge everyone and everything unconsciously, including ourselves, in an attempt to gain a sense of security or control [21]. This can eventually lead to various issues, such as negative impacts on working relationships and increased isolation. Therefore, becoming aware of the key inner critic, the "judge," and practicing intercepting its negative evaluations of oneself, others, and external events are believed to enhance self-compassion and empathy [22].

Rumination

Rumination is defined as the process of being in a negative affective state and repeatedly thinking about the causes, signs, meanings, and consequences of that state [23]. This process is typically self-referential and past-oriented, with individuals continually reflecting on their past to understand the origins and causes of their current situation and relating it to their present self [24,25]. Research shows that rumination is strongly correlated with negative mood, as well as anxiety and depressive symptoms [26-28].

Rumination exemplifies how our mind can engage in a nondeliberate thinking process. By contrast, PQ emphasizes practicing mindfulness and the "self-command muscle" [11]. This approach involves a deliberate process where attention is focused on being present, rather than allowing the mind to wander during rumination. Indeed, mindfulness is believed to disrupt the fixation on negative thoughts, which in turn alleviates rumination [29]. Additionally, becoming aware of our unconscious judgments of ourselves, others, and events, and practicing intercepting these negative evaluations by the "judge saboteur," may lead to a reduction in self-referential rumination [22].

Existing Research

Given the current lack of scholarly literature examining the effects of the PQ intervention on psychological and emotional factors, this study also draws on existing research investigating similar interventions. Specifically, it references studies on mindfulness-based mobile apps and their effects on perceived stress, self-compassion, and rumination [13,14,30-32]. The selection of perceived stress, self-compassion, and rumination as variables in this study is justified by their significant impact on the mental well-being of health care students and their alignment with the objectives of the PQ program. By focusing on these variables, the study aims to provide comprehensive insights into the program's potential to enhance mental fitness and well-being, thereby equipping students with the tools needed to thrive in their demanding educational and professional environments.

Recent literature has recognized the inverse relationship between mindfulness and rumination. Goldberg et al [33] evaluated the effects of mindfulness training, combined with self-compassion practice through introspection, using an app-delivered mindfulness meditation program. The study found that students' self-reported levels of empathy, self-compassion, self-reflection, and rumination improved following the intervention. In a waitlist control experimental study, Huberty et al [34] assessed the impact of an 8-week mindfulness meditation mobile app on university students' perceived stress, mindfulness, self-compassion, and other health behaviors. The authors reported a significant reduction in stress and an increase in self-compassion immediately following the intervention, with these effects sustained at a 1-month follow-up. However, the short follow-up period limits the ability to draw long-term conclusions.

Similar to the PQ mobile app, the study by Huberty et al [34] used the Calm app (Calm.com, Inc.) mobile app, which provided users with daily guided (or unguided) mindfulness meditation exercises. Therefore, we can similarly expect that participants in the PQ program might experience a reduction in stress and increases in self-compassion from the 15 minutes of daily mindfulness practice. The Calm mobile app primarily aims to provide mindfulness meditation to users. Unlike the fixed 6-week program structure of PQ, the aforesaid app allows users to choose from various topics or weekly series targeting different goals (eg, managing stress or increasing self-esteem). Additionally, the exercises vary in length, instruction, and mindfulness principles.

Previous studies on web-based or app-based mindfulness interventions have also reported similar reductions in stress and anxiety after several weeks [13,30-32]. For example, Orosa-Duarte and colleagues [14] compared the effects of a

XSL•FO
RenderX

mindfulness-based mobile app with those of an in-person mindfulness training program on anxiety, empathy, self-compassion, and mindfulness among a group of health care students. The authors found that delivering mindfulness training through a mobile app can be as effective as in-person training in reducing anxiety and increasing self-compassion among health care students. These results align with recent systematic reviews and meta-analyses, which have shown that online mindfulness interventions are beneficial for enhancing self-compassion [35] and reducing stress, depression, and anxiety-related symptoms after intervention [36].

Using a chatbot developed based on the self-help approach by Chamine [9], Kroon [37] used this intervention to help health care workers identify and become aware of their inner critics (saboteurs) and address these critics to enhance self-compassion and reduce compassion fatigue. The author reported that a common theme in interviews with participants was an increased awareness of their inner critics and the negative impact these critics had on their self-compassion and self-kindness. Similarly, a previous study investigating the effects of self-compassion meditation in caregivers found that increasing awareness of one's inner judgmental critic led to a greater recognition of the need to show compassion toward both oneself and others [38].

The PQ program addresses a common barrier to engagement in mobile health (mHealth) interventions—lack of social connectedness—through its "Pod" group meetings. A review by Borghouts et al [39] highlighted numerous studies showing that participants' sense of social connectedness (eg, regular interactions with peers or a personal therapist) enhances user engagement. When participants could emotionally connect with peers by openly sharing problems or relating to others' experiences, their motivation to continue using mental health apps increased [40,41].

Study Objectives

This study aims to evaluate the efficacy of the 6-week PQ program, which includes increasing awareness of inner critics (saboteurs) in one's personal life, training intentional and deliberate control of one's mind and thoughts (self-command muscle), intercepting inner critics (saboteur interceptor muscle), and adopting a different perspective by transforming difficult situations or challenges into specific gifts or opportunities (sage muscle). We will measure health care students' self-compassion, perceived stress, and rumination tendencies before and after the PQ program, and again at a follow-up a few months later, to assess the program's efficacy.

This study addresses the following research question (RQ): "What is the effect of undergoing the 6-week PQ program on health care students' perceived stress, self-compassion, and rumination tendencies, measured before and after the program?" (RQ1). Based on existing findings, we expect that students will show decreases in perceived stress, increases in self-compassion, and decreases in rumination tendencies immediately after completing the PQ program, compared with their scores before the program. However, there has been less consensus on the sustained effects of app-based programs at follow-up after varying periods [26]. Therefore, we pose the following RQ: "To what extent does undergoing the 6-week PQ program benefit health care students' perceived stress, self-compassion, and rumination tendencies at a 5-month follow-up compared with baseline measures before the program?" (RQ2).

A subquestion is as follows: "To what extent do changes in self-compassion, perceived stress, and rumination tendencies depend on whether students have consistently engaged in daily practice throughout the PQ program, as indicated by their cumulative "PQ muscle" count at the end of the 6 weeks?" (RQ3). Addressing this RQ will clarify whether consistent app-based practice contributes to changes in the outcome variables. In line with previous studies on web-based or app-based mindfulness interventions [13,30-32], we recognize that the effectiveness of the PQ program for perceived stress, self-compassion, and rumination tendencies requires a conscious effort from students. Therefore, we hypothesize that students with a higher cumulative average "PQ muscle" count after the program will exhibit a greater decrease in perceived stress and rumination tendencies and a greater increase in self-compassion.

Methods

Sample

This study used a convenience sampling method, with participants drawn from the freshmen cohort of the occupational therapy degree program at a Singapore-based university. All 98 students were beginning their first trimester of occupational therapy studies. No exclusion criteria were applied, so all students in the cohort were eligible as potential participants. As the 6-week PQ program was integrated into a module, no additional incentives were provided for completing the program or participating in the research.

Intervention: PQ Program

The PQ program consists of 3 core modules spread over 6 weeks and is completed individually on the PQ mobile app at the participant's convenience, with the exception of scheduled small group (Pod) meetings. The program includes 15 minutes of daily practice (eg, "Focus of the Day" or "Coach Challenges"); weekly hour-long videos, which are broken down into short segments, with each segment lasting between 3 and 15 minutes; and weekly 45-minute small group meetings (Figure 1). The 15 minutes of daily practice can be divided into several practice blocks (PQ reps) lasting from 10 seconds to a few minutes [42]. These blocks involve brief mindfulness practices [43], which may include focusing attention on a specific sensation or object, such as breathing, tactile sensations, or sounds, or bringing into awareness what one is experiencing as they arise in the present moment.

XSL•FO
RenderX

Figure 1. Details of the PQ program. Adapted from Positive Intelligence, LLC (personal communication, June 14, 2022). PQ: Positive Intelligence.

	Mon	Tues	Wed	Thu	Fri	Sat	Sun
Prep Week	**Introduction to program**	**Download App / Take Saboteur Assessment**				**Watch Week 1 Video** (60 minutes)	
Week 1: Boost Self-Command	**Pod Meeting** (50-60 minutes)					**Watch Week 2 Video** (60 minutes)	
	Focus of the Day, Coach Challenges, Reflection 2 minutes every 3 hours						
Week 2: Intercept the Judge	**Pod Meeting** (50-60 minutes)					**Watch Week 3 Video** (60 minutes)	
	Focus of the Day, Coach Challenges, Reflection 2 minutes every 3 hours						
Week 3: Accomplice Saboteurs	**Pod Meeting** (50-60 minutes)					**Watch Week 4 Video** (60 minutes)	
	Focus of the Day, Coach Challenges, Reflection 2 minutes every 3 hours						
Week 4: Shift to Sage	**Pod Meeting** (50-60 minutes)					**Watch Week 5 Video** (60 minutes)	
	Focus of the Day, Coach Challenges, Reflection 2 minutes every 3 hours						
Week 5: Boost Sage Powers	**Pod Meeting** (50-60 minutes)					**Watch Week 6 Video** (60 minutes)	
	Focus of the Day, Coach Challenges, Reflection 2 minutes every 3 hours						
Week 6: Taking Actions	**Pod Meeting** (50-60 minutes)					**Watch Week 7 Video** (60 minutes) *This video does not cover a new module, but provides a plan and suggestions for continuing practice of the PQ® concepts.	
	Focus of the Day, Coach Challenges, Reflection 2 minutes every 3 hours						
Week 7: Applying PQ	**Pod Meeting** (50-60 minutes)						

In the first week, participants focus on the concept of the "self-command muscle," practicing the development of intentional and deliberate control over their minds and thoughts through "PQ reps." As participants strengthen their "self-command muscle," they can then integrate "PQ reps" with cognitive reframing techniques to effectively intercept their "saboteurs" and activate "sage" responses [42].

In the second and third weeks, participants focus on the "saboteur interceptor muscle," practicing how to intercept and shut down their inner critics (saboteurs). During the third week, the app customizes content based on each participant's top-ranking "saboteur" from the Saboteur Assessment, providing targeted exercises to further develop their "saboteur interceptor muscle." Inner critics (or saboteurs) are defined as internal voices that fuel negative messages and evaluations of oneself, typically developed unconsciously over time [9]. These inner critics are shaped and reinforced by interactions with close others or people in one's environment; experiences of criticism, neglect, or inconsistency in these interactions can lead individuals to form negative self-evaluations (eg, feeling inadequate, unworthy, or useless) [9,44].

Finally, in the fourth to sixth weeks, participants focus on developing the "sage muscle," practicing how to adopt the "sage" perspective by converting difficult situations or challenges into specific gifts or opportunities. For example, the "Three Gifts" technique helps individuals transform challenges or disappointments into valuable opportunities by leveraging them to cultivate personal strengths, gain new insights, or initiate inspired actions as a direct outcome of adversity [42].

Procedure

All students were instructed to download the PQ app from Google Play (Alphabet Inc.) or Apple App Store (Apple Inc.). The rationale and aims of the PQ program were explained to the students through onboarding materials, which included an overview of the program, an introduction to the app's functions, and details on how the "Pod" groups should operate. Students were assigned to small groups of 5 or 6, known as "Pods." They were encouraged to watch the weekly videos by Monday each week in preparation for their virtual "Pod" discussions on Tuesday, which were scheduled for an hour after the module lecture. Students were also encouraged to practice for at least 15 minutes a day using the "Focus of the Day" and "Coach Challenges" features and to log their practice time on the app. Before officially starting the program, students were required to complete the Saboteur Assessment online [45,46], which ranks the dominance of each of the 9 saboteurs [21] in their personal lives.

In addition to the Saboteur Assessment, students completed an online survey questionnaire measuring perceived stress, self-compassion, and rumination as a baseline (pretest). Before taking the survey, students were first asked to consent or decline participation in the study. They were presented with an information sheet detailing the study's purpose, process, and the differences between research participants and nonparticipants, followed by an informed consent form.

Participants completed the same survey questionnaire immediately after the intervention (immediate posttest). To assess the sustained effects of the program on perceived stress, self-compassion, and rumination levels, participants took the questionnaire again 5 months after completing the program (follow-up posttest). During this follow-up, participants were also asked how frequently they continued to use the PQ mobile app since the completion of the 6-week PQ course.

XSL•FO
RenderX

Measures

Measurement of Stress, Self-Compassion, and App Usage Over 6 Weeks

A survey questionnaire measuring perceived stress, self-compassion, and rumination was administered at 3 time points. App usage was measured objectively using the PQ app, which logs a cumulative "muscle" score whenever an individual practices the "self-command muscle," "saboteur interceptor muscle," or "sage muscle" over the 6 weeks. PQ reps are automatically recorded on the app as an indicator of use. Users also have the option to manually enter their mindfulness practice if they do not use the standard app exercises.

Perceived Stress

The 14-item Perceived Stress Scale (PSS-14; [47]) assesses the frequency with which individuals have experienced perceived stress over the past 2 months (eg, "How often have you felt that difficulties were piling up so high that you could not overcome them?"). All items are rated on a 5-point Likert scale (1=never to 5=always), with higher scores indicating greater perceived stress experienced over the past 2 months. Cronbach α for this sample was >0.84 at all measurement time points, indicating the high reliability of the measure. This is consistent with existing evidence demonstrating that the PSS-14 is a reliable and valid measure of perceived stress across various samples of university students [48-50].

Self-Compassion

The 12-item Self-Compassion Scale–Short Form (SCS-SF; [51]) assesses how respondents view their actions toward themselves during difficult times (eg, "When times are really difficult, I tend to be tough on myself"). Items are rated on a 5-point scale (1=strongly disagree to 5=strongly agree), with higher scores indicating greater self-compassion. Cronbach α for this sample ranged from 0.78 to 0.83 at all measurement time points, indicating the high reliability for this measure in our study. This aligns with previous findings that demonstrate the 12-item SCS-SF's robust psychometric properties, including high internal reliability [52,53] and strong face, content, convergent, and divergent validity [51,54].

Rumination

The 12-item Rumination subscale of the Reflection Rumination Questionnaire (RRQ-Rum; [55]) was used to assess rumination (eg, "Long after an argument or disagreement is over, my thoughts keep going back to what happened"). Items are rated on a 5-point scale (1=strongly disagree to 5=strongly agree), with higher scores indicating greater rumination tendencies. Cronbach α for this sample was >0.88 at all measurement time points, demonstrating the high reliability of the measure. This is consistent with previous studies that report high internal reliability [56,57] and robust convergent and divergent validity [58,59].

Data Analysis

This study aims to evaluate the efficacy of the 6-week PQ program on health care students' perceived stress, self-compassion, and rumination tendencies. To assess differences in perceived stress, self-compassion, and rumination scores across the pretest, immediate posttest, and follow-up posttest, a repeated measures ANOVA was conducted. The data set was cleaned and prepared for analysis, including only data from participants who completed measurements at all 3 time points. Before conducting the analyses, Shapiro-Wilk tests were performed to assess the assumption of normality, and Mauchly tests were used to check for violations of the assumption of sphericity. This analysis will help determine whether there are statistically significant differences in students' perceived stress, self-compassion, and rumination tendencies from pre-intervention to postintervention and at follow-up 5 months later. For each outcome variable, pairwise comparisons using the Tukey honestly significant difference (HSD) test were conducted when significant differences were observed.

To address RQ3, which asks, "To what extent does the change in self-compassion, perceived stress, and rumination tendencies scores depend on whether students have consistently engaged in daily practice throughout the PQ program, as indicated by their cumulative 'PQ muscle' count at the end of the six weeks?" A linear regression analysis with an interaction term was conducted to test for moderation effects. Posttest scores for each of the outcome variables—perceived stress, self-compassion, and rumination—were regressed on their respective pretest scores. In each linear regression model, the cumulative PQ muscle count was included as a moderator.

Ethical Considerations

Ethical approval for data collection was obtained from the Institutional Review Board at the Singapore Institute of Technology (approval number 2022157). All students in the cohort were enrolled in the PQ program but had the option to decline consent for their data to be used for research purposes. The first author (SP), a nonteaching staff member, informed students of the opportunity to participate in the research study both during lecture time and via email. Informed consent was obtained from participants before their involvement in the study and students were assured that their data would be used solely for research purposes. Research data were pseudonymized to protect participant confidentiality.

Results

Overview

Ultimately, 87 out of 98 students from the occupational therapy freshmen cohort signed the informed consent form. Of these, 76 students completed the pretest, which measured self-compassion, perceived stress, and rumination tendencies. Following the 6-week PQ program, 78 students completed the same questionnaire for the immediate posttest. At follow-up, 87 students completed the questionnaire 5 months after the completion of the 6-week PQ program (follow-up posttest).

Of these students, 64 could be matched across time, having completed the pretest, immediate posttest, and the follow-up posttest. Of the 64 students analyzed, 47 (73%) were female (mean age 23 years, SD 5.06 years). The distribution of the cohort also had a female majority, with 76 out of 98 students (78%) being female.

XSL•FO
RenderX

Primary Outcomes

Perceived Stress

The assumptions of normality (Shapiro-Wilk tests: pretest, *P*=.37; posttest, *P*=.55; and after 5 months, *P*=.21) and sphericity (Mauchly test: W=0.96, *P*=.24) were met. A repeated measures ANOVA revealed a statistically significant difference in perceived stress scores across the 3 time points ($F_{2,126}$=10.55, *P*<.001, η^2_g=0.05). Contrary to our hypothesis, pairwise comparisons using the Tukey HSD test showed no significant change in students' perceived stress ratings after the 6-week PQ program (mean 2.97, SD 0.16; *P*=.50) compared with their baseline scores measured before the start of the program (mean 2.99, SD 0.14). Instead, we found a significant increase in students' ratings of perceived stress at follow-up (mean 3.17, SD 0.21; *P*<.001) from perceived stress scores measured before and after the intervention (Figure 2).

Figure 2. Changes in students' stress, self-compassion, and rumination from pretest to posttest to 5-month follow-up.

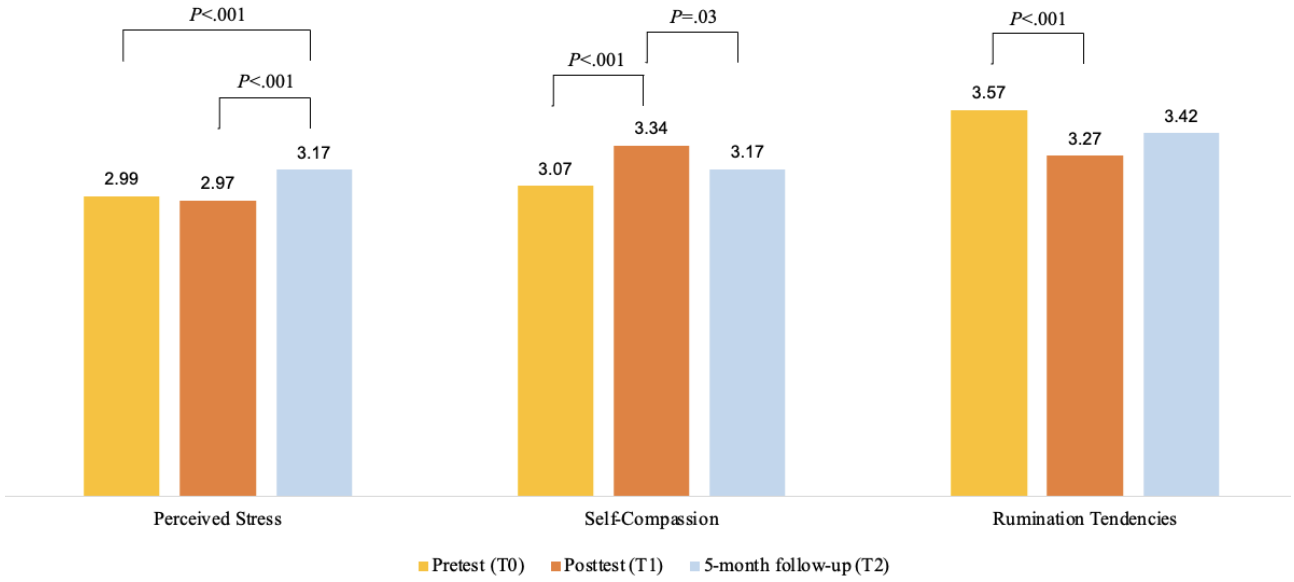

Self-Compassion

The Shapiro-Wilk test confirmed that the assumption of normality was met for pretest scores (*P*=.73) but a slight departure from normality for the other time points (posttest, *P*=.02; after 5 months, *P*=.01); however, visual inspection of the Q-Q plots suggested that the deviations were minor. The assumption of sphericity (Mauchly test: W=.97, *P*=.45) was met. A repeated measures ANOVA found a statistically significant difference in self-compassion ratings across the 3 time points ($F_{2,126}$=9.83, *P*<.001, η^2_g=0.05). As hypothesized, pairwise comparisons with the Tukey HSD test showed a significant increase in students' self-compassion ratings after the 6-week PQ program (mean 3.34, SD 0.35; *P*<.001) compared with self-compassion scores before the program (mean 3.07, SD 0.35). However, at follow-up (mean 3.17, SD 0.27), self-compassion ratings were no different from that of scores before the program (*P*=.09) but were significantly less than scores immediately after the program (*P*=.03).

Rumination

We confirmed that the assumptions of normality (Shapiro-Wilk tests: pretest, *P*=.23; posttest, *P*=.14; and after 5 months, *P*=.57) and sphericity (Mauchly test: W=.97, *P*=.44) were met. A repeated measures ANOVA found a statistically significant difference in rumination scores across the 3 time points ($F_{2,126}$=8.48, *P*<.001, η^2_g=0.04). In line with our hypothesis, pairwise comparisons with the Tukey HSD test showed a significant decrease in students' rumination tendencies after the 6-week PQ program (mean 3.27, SD 0.34; *P*<.001) compared with rumination tendencies before the program (mean 3.57, SD 0.40). However, at follow-up (mean 3.42, SD 0.38; *P*=.06), rumination scores were no different from that of scores before the program or immediately after the program.

App Usage

Results of the moderated regression analysis revealed no significant interaction effect between pretest scores of perceived stress, self-compassion, or rumination and PQ muscle counts on immediate posttest scores for these variables (*P*=.62 for perceived stress, *P*=.65 for self-compassion, and *P*=.42 for rumination; Table 1). This means that contrary to our hypothesis, changes in perceived stress, self-compassion, and rumination scores from pretest to posttest did not depend on whether students engaged consistently in daily app-based practices throughout the PQ program, as indicated by their cumulative "PQ muscles" logged in the app.

At the 5-month follow-up, participants were asked about their continued use of the PQ mobile app since the completion of the 6-week program. Of the 64 students who completed the follow-up survey, only 1 (2%) participant reported using the app "sometimes." By contrast, 46 (72%) participants indicated they had "never" used the app and 17 (27%) participants indicated they had "seldom" used it.

XSL•FO
RenderX

Table 1. Results of the moderated regression analyses.

Regression analysis	β (SE)	*P* value
Perceived stress		
Intercept	3.00 (0.04)	<.001
Pretest perceived stress	0.67 (0.11)	<.001
PQ[a] muscle	0.00 (0.00)	.19
Pretest perceived stress × PQ muscle	0.00 (0.00)	.62
Self-compassion		
Intercept	3.30 (0.05)	<.001
Pretest self-compassion	0.41 (0.08)	<.001
PQ muscle	0.00 (0.00)	.07
Pretest self-compassion × PQ muscle	0.00 (0.00)	.65
Rumination		
Intercept	3.30 (0.06)	<.001
Pretest rumination	0.57 (0.09)	<.001
PQ muscle	0.00 (0.00)	.89
Pretest rumination × PQ muscle	0.00 (0.00)	.42

[a]PQ: Positive Intelligence.

Discussion

Principal Findings

This study investigated whether a 6-week app-delivered mindfulness-based mental fitness program, PQ, could decrease perceived stress levels, increase self-compassion, and reduce rumination tendencies among health care undergraduates. Our findings partially support our hypotheses: by the end of the 6 weeks, students exhibited significant increases in self-compassion and reductions in rumination tendencies. However, there was no significant effect on perceived stress levels. At the 5-month follow-up after completing the PQ program, we observed that students' self-compassion and rumination scores were no different from their baseline measures (pretest). Additionally, perceptions of stress had increased at follow-up. Furthermore, there was no evidence that more consistent daily PQ practice, as tracked through the app, resulted in greater changes in self-compassion, perceived stress, or rumination tendencies.

Changes to Perceived Stress, Self-Compassion, and Rumination

A particularly interesting and unexpected finding was the significant benefit of the PQ program on self-compassion and rumination levels, though not on stress. The lack of an effect on stress might be attributed to the timing of the study; our sample consisted of students in their first year and first trimester of undergraduate studies. As the participants began the PQ program and completed the pretest surveys, they were also just starting their first weeks of freshman year. Some students may have taken a long break from studying before starting their degree program, so our baseline measure might not have been the most representative of their perceived stress levels. At the same time, immediate posttest surveys were distributed during a period when several assignments were due. Consequently, although our findings did not support our hypothesis, we are encouraged by the fact that students did not report a significant increase in stress immediately after completing the program, despite the rising academic workload. We recommend that future studies include a control group to better investigate the true effect of the PQ program on perceived stress, as this would provide a clearer understanding of how the PQ program functions [60].

As perceptions of stress depend on life and work demands, it is likely that the PQ program did not directly alter students' perceptions of stress. Instead, it may have reframed students' automatic negative thought processes, increased their awareness of potential self-sabotage, and helped them develop greater self-compassion and reduced ruminative thinking. Rumination tendencies significantly decreased from the pretest to the posttest as hypothesized, consistent with previous research indicating that app-delivered mindfulness-based practices can reduce rumination [27,29]. Additionally, we observed a significant increase in self-compassion ratings after the 6-week program. This aligns with previous studies, such as Frewen et al [61], which found that mindfulness practice enhances an individual's ability to let go of negative automatic thoughts. As humans, it is natural to fixate on negative thoughts and struggle to break free from the cycle of rumination. Practicing the ability to "be present" in the here and now, without getting distracted or lost in repetitive thoughts about worries or situations, enhances psychological and emotional regulation and reduces automatic negative thinking [61,62].

However, if mindfulness practice alone were responsible for the intervention effects on self-compassion and ruminative thinking, we would have expected to see a notable difference

XSL•FO
RenderX

in outcomes between students who logged more daily PQ exercises and those who logged fewer. Instead, the study results showed no evidence that more consistent daily PQ practice, as tracked through the app, led to greater changes in self-compassion, perceived stress, or rumination tendencies. In other words, app usage did not moderate the changes in self-compassion, perceived stress, or rumination scores between the pretest and posttest. Therefore, our findings cannot be solely attributed to mindfulness practice alone. Other aspects of the PQ program—such as the awareness gained from the Saboteur Assessment, the weekly videos, and the weekly "Pod" meetings—may have also contributed to the observed intervention effects. Moreover, it is possible that students engaged in PQ or similar mindfulness practices outside of the app, which means that the tracked app usage might not accurately reflect the total amount of practice. However, without an active control group, our study is limited in the comparisons and conclusions we can draw.

Previous research has shown that the PQ program's focus on awareness of one's inner critics leads to greater self-acceptance and encourages withholding judgment toward oneself and others [37,38]. This awareness fosters empathy and compassion, helping individuals recognize that self-sabotaging patterns and tendencies are often driven by underlying struggles, such as unmet needs, vulnerabilities, or past trauma [63]. Indeed, Stjernswärd and Hansson [38] found that increasing awareness of participants' self-sabotaging inner critics led to greater compassion and acceptance toward themselves, others, and external situations. Furthermore, gaining awareness of these inner critics is essential for effectively interrupting negative thinking patterns. This approach helps reduce self-referential rumination tendencies by enabling individuals to recognize sabotaging thoughts or behaviors and consciously redirect their attention to more constructive thinking [64]. Our findings suggest that the novel metacognitive awareness of one's inner critics (saboteurs) may have broadened students' perceptions of how they were self-sabotaging their relationships with others and themselves, thereby contributing to the observed intervention effects.

Based on existing literature, we also speculate that social connections formed during the "Pod" meetings of the PQ program facilitated greater reflection and introspection. This social interaction likely enhanced participants' awareness of their inner critics and how these critics negatively impacted them. In a study examining the usefulness of mHealth apps from the perspective of health care workers, Yoon et al [65] found that these workers frequently reported a need for ongoing social support and desired access to an in-app peer support community. According to nudge theory, social influence acts as a significant push factor (nudge) by fostering a sense of belonging to a community with shared goals, thereby enhancing the effectiveness of mHealth apps [65,66]. Therefore, being part of a "Pod" social group likely contributed to sustained engagement in the program, as well as reductions in rumination tendencies and increases in self-compassion. However, because "Pod" group meetings were mandatory and scheduled weekly in this study, future research should explore whether these results are generalizable to settings where "Pod" meetings are nonmandatory and rely on voluntary group participation.

Five months after completing the PQ program, when the "Pod" groups were no longer scheduled, rumination tendencies and self-compassion scores were no longer significantly different from pretest levels. In other words, the PQ program's effects on rumination tendencies and self-compassion did not persist after 5 months. Given the mechanisms cited for the changes in self-compassion and rumination tendencies immediately after the PQ program, it is not surprising that these effects were not sustained. As the initial insights from becoming aware of one's self-sabotaging inner critics faded and the "Pod" meetings ceased, the sustained impact on self-compassion and ruminative thinking diminished. However, a more likely explanation for the lack of sustained effects from the PQ program is that a significant percentage of students (46/64, 72%) did not continue using the PQ mobile app regularly after the completion of the 6-week program. Research indicates that it is common for users to engage with mHealth apps only for a short period and that sustained use typically requires clear, ongoing benefits. Approximately 74% of mHealth app users discontinue use by the tenth session [67]. Future studies should explore strategies to motivate health care students to continue benefiting from the PQ app beyond the regular program.

Practical Implications

Our findings underscore the importance of fostering self-awareness of inner critics to enhance students' well-being. This core component of the PQ program may have universal applicability across diverse student populations, not just health care students, though further research is needed to confirm this. Encouraging students to engage in practices such as mindfulness and journaling can help them identify and understand their inner critics. Educators can play a crucial role in promoting self-compassion and well-being among students. By being trained in the PQ program, educators can model and cultivate self-awareness in their classrooms. Creating a supportive and nonjudgmental environment allows educators to teach students how to develop self-compassion, challenge negative self-talk, and manage stress effectively. Additionally, integrating principles of self-compassion and well-being into the curriculum can promote lasting awareness and mindfulness, rather than relying solely on a structured program.

Strengths

This study has several notable strengths. First, it utilized a mindfulness-based PQ training program delivered through a smartphone app, providing a highly scalable, accessible, engaging, and convenient way for participants to regularly access the program, compared with traditional in-person mindfulness classes. Second, the study focused on a specific group of participants: occupational therapy undergraduates. This is a unique population that experiences high levels of stress and could benefit significantly from mindfulness-based interventions. Third, the study used a longitudinal design, enabling us to assess the effects of the intervention over an extended period, up to 5 months after the program concluded. Lastly, the intervention in this study was a 6-week app-delivered PQ program, which has been used with MBA students, athletes,

XSL•FO
RenderX

executive teams, and chief executive officers [10]. This novel intervention has not been extensively studied, particularly within the health care student population, making our findings especially valuable.

Limitation

One limitation of this study is the absence of a comparable intervention group to control for potential confounding factors. Future research should include an active control condition to isolate the specific effects of the PQ program and determine which elements are most effective in enhancing self-compassion and reducing stress and rumination among health care students.

Conclusions

This study examined the effects of the 6-week PQ program on perceived stress, self-compassion, and rumination tendencies in health care undergraduates. While the findings did not support the hypothesis that the PQ program would reduce perceived stress levels, there were significant increases in self-compassion and reductions in rumination tendencies immediately following the intervention. However, there was no significant effect on self-compassion, rumination tendencies, or perceived stress at the 5-month follow-up, and no evidence that increased daily PQ practice led to greater changes. Our findings suggest that the PQ program helped reframe students' automatic negative thought processes through mindfulness practice. Additionally, the novel awareness of saboteurs and the weekly "Pod" group meetings may have contributed to the observed intervention effects. Our findings contribute to existing evidence that increased self-acceptance and the practice of withholding judgment toward oneself and others are essential for enhancing self-compassion and reducing rumination. We recommend further research with an active control group to better understand the actual impact of the PQ program and to identify which specific components of the program produce the most significant effects.

Acknowledgments

The authors thank all students who completed the survey and for their willingness to participate in this study.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SP and CYP carried out data collection. SP analyzed the data and wrote the manuscript with feedback from CYP, SML, and JS. SML supervised the project. All authors contributed substantially to the design and conceptualization of the study, reviewed the paper, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. McConville J, McAleer R, Hahne A. Mindfulness training for health profession students-the effect of mindfulness training on psychological well-being, learning and clinical performance of health professional students: a systematic review of randomized and non-randomized controlled trials. Explore (NY) 2017 Jan;13(1):26-45. [doi: 10.1016/j.explore.2016.10.002] [Medline: 27889445]
2. de Witt PA, Monareng L, Abraham AA, Koor S, Saber R. Resilience in occupational therapy students. S Afr J Occup Ther 2019;49(2):33-41. [doi: 10.17159/23103833/2019/vol49n2a6]
3. Delany C, Miller KJ, El-Ansary D, Remedios L, Hosseini A, McLeod S. Replacing stressful challenges with positive coping strategies: a resilience program for clinical placement learning. Adv Health Sci Educ Theory Pract 2015 Dec 3;20(5):1303-1324. [doi: 10.1007/s10459-015-9603-3] [Medline: 25835325]
4. Nerdrum P, Rust en T, Helge R nnestad M. Psychological distress among nursing, physiotherapy and occupational therapy students: a longitudinal and predictive study. Scandinavian Journal of Educational Research 2009 Aug;53(4):363-378. [doi: 10.1080/00313830903043133]
5. Morrissette P, Doty-Sweetnam K. Safeguarding student well-being: establishing a respectful learning environment in undergraduate psychiatric/mental health education. J Psychiatr Ment Health Nurs 2010 Aug;17(6):519-527. [doi: 10.1111/j.1365-2850.2010.01551.x] [Medline: 20633079]
6. Dunn LB, Iglewicz A, Moutier C. A conceptual model of medical student well-being: promoting resilience and preventing burnout. Acad Psychiatry 2008 Feb 01;32(1):44-53. [doi: 10.1176/appi.ap.32.1.44] [Medline: 18270280]
7. Robins TG, Roberts RM, Sarris A. The role of student burnout in predicting future burnout: exploring the transition from university to the workplace. Higher Education Research & Development 2017 Jun 29;37(1):115-130. [doi: 10.1080/07294360.2017.1344827]

XSL•FO
RenderX

8. O'Driscoll M, Sahm LJ, Byrne H, Lambert S, Byrne S. Impact of a mindfulness-based intervention on undergraduate pharmacy students' stress and distress: quantitative results of a mixed-methods study. Curr Pharm Teach Learn 2019 Sep;11(9):876-887. [doi: 10.1016/j.cptl.2019.05.014] [Medline: 31570124]
9. Chamine S. Positive Intelligence: Why Only 20% of Teams and Individuals Achieve Their True Potential and How You Can Achieve Yours. Austin, TX: Greenleaf Book Group Press; 2012.
10. About | Positive Intelligence. Positive Intelligence. 2021. URL: https://www.positiveintelligence.com/about/ [accessed 2024-07-29]
11. The science of positive intelligence. Positive Intelligence. 2021. URL: https://www.positiveintelligence.com/science/ [accessed 2024-07-29]
12. Braun SE, Dow A, Loughan A, Mladen S, Crawford M, Rybarczyk B, et al. Mindfulness training for healthcare professional students: a waitlist controlled pilot study on psychological and work-relevant outcomes. Complement Ther Med 2020 Jun;51:102405-102407. [doi: 10.1016/j.ctim.2020.102405] [Medline: 32507424]
13. Krusche A, Cyhlarova E, Williams JMG. Mindfulness online: an evaluation of the feasibility of a web-based mindfulness course for stress, anxiety and depression. BMJ Open 2013 Nov 29;3(11):e003498 [FREE Full text] [doi: 10.1136/bmjopen-2013-003498] [Medline: 24293203]
14. Orosa-Duarte ?, Mediavilla R, Muñoz-Sanjose A, Palao ?, Garde J, López-Herrero V, et al. Mindfulness-based mobile app reduces anxiety and increases self-compassion in healthcare students: a randomised controlled trial. Medical Teacher 2021 Feb 28;43(6):686-693. [doi: 10.1080/0142159x.2021.1887835]
15. Lazarus RS. Theory-based stress measurement. Psychological Inquiry 1990 Jan;1(1):3-13. [doi: 10.1207/s15327965pli0101_1]
16. Epel ES, Crosswell AD, Mayer SE, Prather AA, Slavich GM, Puterman E, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169 [FREE Full text] [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
17. Koura M, Al-Dabal B, Rasheed P, Al-Sowielem L, Makki S. Prehypertension among young adult females in Dammam, Saudi Arabia. East Mediterr Health J 2012 Jul 01;18(7):728-734 [FREE Full text] [doi: 10.26719/2012.18.728] [Medline: 22891521]
18. Neff K. Self-compassion: an alternative conceptualization of a healthy attitude toward oneself. Self and Identity 2003 Apr;2(2):85-101. [doi: 10.1080/15298860309032]
19. NEFF KD. The development and validation of a scale to measure self-compassion. Self and Identity 2003 Jul;2(3):223-250. [doi: 10.1080/15298860309027]
20. Germer CK, Neff KD. Self-compassion in clinical practice. J Clin Psychol 2013 Aug 17;69(8):856-867. [doi: 10.1002/jclp.22021] [Medline: 23775511]
21. Positive I. Take the Saboteur Assessment. The Accomplice Saboteurs. 2021. URL: https://www.positiveintelligence.com/saboteurs/ [accessed 2024-07-29]
22. The neuroscience of positive intelligence. Positive Intelligence. 2016. URL: https://www.positiveintelligence.com/wp-content/uploads/2022/05/neuroscience-of-positive-intelligence-2022-v1-0.pdf [accessed 2024-07-29]
23. Ramel1 W, Goldin PR, Carmona PE, McQuaid JR. The effects of mindfulness meditation on cognitive processes and affect in patients with past depression. Cognitive Therapy and Research 2004 Aug;28(4):433-455. [doi: 10.1023/b:cotr.0000045557.15923.96]
24. Didriksen L. A contemplative perspective on mind wandering. In: Dario N, Tateo L, editors. New Perspectives on Mind-Wandering. Cham, Switzerland: Springer International Publishing; 2022:175-199.
25. Nejad AB, Fossati P, Lemogne C. Self-referential processing, rumination, and cortical midline structures in major depression. Front Hum Neurosci 2013 Oct 10;7:666-669 [FREE Full text] [doi: 10.3389/fnhum.2013.00666] [Medline: 24124416]
26. Aldao A, Nolen-Hoeksema S, Schweizer S. Emotion-regulation strategies across psychopathology: a meta-analytic review. Clin Psychol Rev 2010 Mar;30(2):217-237. [doi: 10.1016/j.cpr.2009.11.004] [Medline: 20015584]
27. Kiran I, Ahsan S, Zaheer S, Naz A. Impact of self-generated thoughts on anxiety symptoms among university students: mediating role of rumination. Pakistan Journal of Psychological Research 2020 May 20;35(1):37-51. [doi: 10.33824/pjpr.2020.35.1.3]
28. Martin LL, Tesser A. Some ruminative thoughts. In: Wyer, Jr RS, editor. Ruminative Thoughts. New York, NY: Psychology Press; 1996:1-47.
29. Hawley LL, Schwartz D, Bieling PJ, Irving J, Corcoran K, Farb NAS, et al. Mindfulness practice, rumination and clinical outcome in mindfulness-based treatment. Cogn Ther Res 2013 Oct 8;38(1):1-9. [doi: 10.1007/s10608-013-9586-4]
30. El Morr C, Ritvo P, Ahmad F, Moineddin R, Team M. Effectiveness of an 8-week web-based mindfulness virtual community intervention for university students on symptoms of stress, anxiety, and depression: randomized controlled trial. JMIR Ment Health 2020 Jul 17;7(7):e18595 [FREE Full text] [doi: 10.2196/18595] [Medline: 32554380]
31. Kranenburg LW, Gillis J, Mayer B, Hoogendijk WJG. The effectiveness of a nonguided mindfulness app on perceived stress in a nonclinical Dutch population: randomized controlled trial. JMIR Ment Health 2022 Mar 18;9(3):e32123 [FREE Full text] [doi: 10.2196/32123] [Medline: 35302504]

XSL•FO
RenderX

32. Ritvo P, Ahmad F, El Morr C, Pirbaglou M, Moineddin R, Team MVC. A mindfulness-based intervention for student depression, anxiety, and stress: randomized controlled trial. JMIR Ment Health 2021 Jan 11;8(1):e23491 [FREE Full text] [doi: 10.2196/23491] [Medline: 33264098]
33. Goldberg SB, Imhoff-Smith T, Bolt DM, Wilson-Mendenhall CD, Dahl CJ, Davidson RJ, et al. Testing the efficacy of a multicomponent, self-guided, smartphone-based meditation app: three-armed randomized controlled trial. JMIR Ment Health 2020 Nov 27;7(11):e23825 [FREE Full text] [doi: 10.2196/23825] [Medline: 33245288]
34. Huberty J, Green J, Glissmann C, Larkey L, Puzia M, Lee C. Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: randomized controlled trial. JMIR Mhealth Uhealth 2019 Jun 25;7(6):e14273 [FREE Full text] [doi: 10.2196/14273] [Medline: 31237569]
35. Linardon J. Can acceptance, mindfulness, and self-compassion be learned by smartphone apps? A systematic and meta-analytic review of randomized controlled trials. Behav Ther 2020 Jul;51(4):646-658. [doi: 10.1016/j.beth.2019.10.002] [Medline: 32586436]
36. Ahmad F, Wang J, Morr CE. Online mindfulness interventions: a systematic review. In: Morr CE, editor. Novel Applications of Virtual Communities in Healthcare Settings. Hershey, PA: IGI Global; 2017:1-27.
37. Kroon Y. The appreciation of the iThrive chatbot and its impact on self-compassion and compassion fatigue in healthcare workers: a mixed method feasibility study. Master Thesis. University of Twente. 2018. URL: http://essay.utwente.nl/75151/ [accessed 2024-09-11]
38. Stjernswärd S, Hansson L. User value and usability of a web-based mindfulness intervention for families living with mental health problems. Health Soc Care Community 2017 Mar 17;25(2):700-709. [doi: 10.1111/hsc.12360] [Medline: 27189237]
39. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
40. Berry N, Lobban F, Bucci S. A qualitative exploration of service user views about using digital health interventions for self-management in severe mental health problems. BMC Psychiatry 2019 Jan 21;19(1):35-13 [FREE Full text] [doi: 10.1186/s12888-018-1979-1] [Medline: 30665384]
41. Kenny R, Dooley B, Fitzgerald A. Developing mental health mobile apps: exploring adolescents' perspectives. Health Informatics J 2016 Jun;22(2):265-275 [FREE Full text] [doi: 10.1177/1460458214555041] [Medline: 25385165]
42. Taylor S. How to increase your mental fitness in 15 minutes a day. Medium. 2023. URL: https://medium.com/@thrivewise/you-take-10-000-steps-a-day-for-physical-fitness-now-try-100-pq-reps-a-day-for-mental-fitness-31dc1e2f2d19 [accessed 2024-07-29]
43. Lutz A, Slagter HA, Dunne JD, Davidson RJ. Attention regulation and monitoring in meditation. Trends Cogn Sci 2008 Apr;12(4):163-169 [FREE Full text] [doi: 10.1016/j.tics.2008.01.005] [Medline: 18329323]
44. Green EJ, Elliott K, Myrick AC. Deepening couples counseling through inner critic dialogue. The Family Journal 2013 Dec 09;22(2):156-161. [doi: 10.1177/1066480713513549]
45. Chamine S, Barnhart M. Research foundation for positive intelligence assessments. Positive Intelligence. 2022. URL: https://www.positiveintelligence.com/wp-content/uploads/2022/07/Whitepaper-Research-Foundation-for-Positive-Intelligence-Assessments.pdf [accessed 2024-07-29]
46. Positive I. Assessments. Positive Intelligence. URL: https://assessment.positiveintelligence.com/saboteur/overview [accessed 2024-07-29]
47. Cohen S, Kamarck T, Mermelstein R. Perceived stress scale. In: Cohen S, Kessler RC, Gordon LU, editors. Measuring Stress: A Guide for Health and Social Scientists. New York, NY: Oxford University Press; 1994:138-139.
48. Andreou E, Alexopoulos EC, Lionis C, Varvogli L, Gnardellis C, Chrousos GP, et al. Perceived Stress Scale: reliability and validity study in Greece. Int J Environ Res Public Health 2011 Aug 11;8(8):3287-3298 [FREE Full text] [doi: 10.3390/ijerph8083287] [Medline: 21909307]
49. Huang F, Wang H, Wang Z, Zhang J, Du W, Su C, et al. Psychometric properties of the perceived stress scale in a community sample of Chinese. BMC Psychiatry 2020 Mar 20;20(1):130-137 [FREE Full text] [doi: 10.1186/s12888-020-02520-4] [Medline: 32197589]
50. Tambol Z, Abu Bakar AY, Mahmud MI. Validity and reliability of the Malaysian Perceived Stress Scale (PSS) using Rasch Measurement Model. IJARBSS 2021 Aug 12;11(8):667-677. [doi: 10.6007/ijarbss/v11-i8/10770]
51. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011 Jun 08;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
52. Dev V, Fernando AT, Consedine NS. Self-compassion as a stress moderator: a cross-sectional study of 1700 doctors, nurses, and medical students. Mindfulness (N Y) 2020 Feb 11;11(5):1170-1181 [FREE Full text] [doi: 10.1007/s12671-020-01325-6] [Medline: 32435318]
53. Póka T, Barta A, Mérő L. Self-compassion mediates the relationships between university students' mindfulness, dysfunctional attitudes, and various distress and well-being indicators. Journal of Research in Higher Education 2023 Dec 21;7(2):35-66. [doi: 10.24193/jrhe.2023.2.2]
54. Neff KD. The Self-Compassion Scale is a valid and theoretically coherent measure of self-compassion. Mindfulness 2015 Dec 18;7(1):264-274. [doi: 10.1007/s12671-015-0479-3]

XSL•FO
RenderX

55. Trapnell PD, Campbell JD. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. J Pers Soc Psychol 1999 Feb;76(2):284-304. [doi: 10.1037//0022-3514.76.2.284] [Medline: 10074710]
56. DaSilveira A, DeSouza ML, Gomes WB. Self-consciousness concept and assessment in self-report measures. Front Psychol 2015 Jul 03;6:930 [FREE Full text] [doi: 10.3389/fpsyg.2015.00930] [Medline: 26191030]
57. Svendsen JL, Osnes B, Binder P, Dundas I, Visted E, Nordby H, et al. Trait self-compassion reflects emotional flexibility through an association with high vagally mediated heart rate variability. Mindfulness (N Y) 2016 Jun 2;7(5):1103-1113 [FREE Full text] [doi: 10.1007/s12671-016-0549-1] [Medline: 27642372]
58. Harrington R, Loffredo DA. Insight, rumination, and self-reflection as predictors of well-being. J Psychol 2011 Nov 30;145(1):39-57. [doi: 10.1080/00223980.2010.528072] [Medline: 21290929]
59. Siegle GJ, Moore PM, Thase ME. Rumination: one construct, many features in healthy individuals, depressed individuals, and individuals with lupus. Cognitive Therapy and Research 2004 Oct;28(5):645-668. [doi: 10.1023/b:cotr.0000045570.62733.9f]
60. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled Trials. JMIR Ment Health 2021 Jul 19;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
61. Frewen PA, Evans EM, Maraj N, Dozois DJA, Partridge K. Letting go: mindfulness and negative automatic thinking. Cogn Ther Res 2007 May 9;32(6):758-774. [doi: 10.1007/s10608-007-9142-1]
62. Eckland NS, Berenbaum H. Emotional awareness in daily life: exploring its potential role in repetitive thinking and healthy coping. Behav Ther 2021 Mar;52(2):338-349. [doi: 10.1016/j.beth.2020.04.010] [Medline: 33622504]
63. Ke T, Barlas J. Thinking about feeling: using trait emotional intelligence in understanding the associations between early maladaptive schemas and coping styles. Psychol Psychother 2020 Mar 28;93(1):1-20 [FREE Full text] [doi: 10.1111/papt.12202] [Medline: 30369013]
64. Kang Y, Gruber J, Gray JR. Mindfulness and de-automatization. Emotion Review 2012 Nov 07;5(2):192-201. [doi: 10.1177/1754073912451629]
65. Yoon S, Goh H, Nadarajan GD, Sung S, Teo I, Lee J, et al. Perceptions of mobile health apps and features to support psychosocial well-being among frontline health care workers involved in the COVID-19 pandemic response: qualitative study. J Med Internet Res 2021 May 31;23(5):e26282 [FREE Full text] [doi: 10.2196/26282] [Medline: 33979296]
66. Wiecek E, Torres-Robles A, Cutler RL, Benrimoj SI, Garcia-Cardenas V. Impact of a multicomponent digital therapeutic mobile app on medication adherence in patients with chronic conditions: retrospective analysis. J Med Internet Res 2020 Aug 12;22(8):e17834 [FREE Full text] [doi: 10.2196/17834] [Medline: 32784183]
67. Motivating patients to use smartphone health apps. Consumer Health Information Corporation. 2011. URL: http://www.consumer-health.com/press/2008/NewsReleaseSmartPhoneApps.php [accessed 2024-07-29]

Abbreviations

HSD: honestly significant difference
mHealth: mobile health
PQ: Positive Intelligence
PSS-14: 14-item Perceived Stress Scale
RQ: research question
RRQ-Rum: Reflection Rumination Questionnaire
SCS-SF: Self-Compassion Scale–Short Form

Edited by T de Azevedo Cardoso; submitted 28.06.23; peer-reviewed by A Slezackova, R Nooripour; comments to author 09.05.24; revised version received 25.06.24; accepted 08.08.24; published 07.10.24.

Please cite as:
Puah S, Pua CY, Shi J, Lim SM
The Effectiveness of a Digital Mental Fitness Program (Positive Intelligence) on Perceived Stress, Self-Compassion, and Ruminative Thinking of Occupational Therapy Undergraduate Students: Longitudinal Study
J Med Internet Res 2024;26:e49505
URL: https://www.jmir.org/2024/1/e49505
doi:10.2196/49505
PMID:39374067

©Shermain Puah, Ching Yee Pua, Jing Shi, Sok Mui Lim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of a Digital Health Intervention Leveraging Reinforcement Learning: Results From the Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) Randomized Clinical Trial

Adrian Aguilera[1,2], PhD; Marvyn Arévalo Avalos[1], PhD; Jing Xu[3], PhD; Bibhas Chakraborty[3], PhD; Caroline Figueroa[4], PhD; Faviola Garcia[5], BA; Karina Rosales[1], BA; Rosa Hernandez-Ramos[6], BA; Chris Karr[7], BA; Joseph Williams[8], PhD; Lisa Ochoa-Frongia[9], MD; Urmimala Sarkar[9], MD; Elad Yom-Tov[10], PhD; Courtney Lyles[11,12], PhD

[1]School of Social Welfare, University of California Berkeley, Berkeley, CA, United States
[2]Department of Psychiatry and Behavioral Sciences, University of California-San Francisco, San Francisco, CA, United States
[3]Centre for Quantitative Medicine, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore
[4]Faculty of Technology, Policy, and Management, Delft Technical University, Delft, Netherlands
[5]Action Research Center (ARC) for Health Equity, Department of Medicine, University of California-San Francisco, San Francisco, CA, United States
[6]Department of Psychological Sciences, University of California Irvine, Irvine, CA, United States
[7]Audacious Software, Chicago, IL, United States
[8]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[9]Department of Medicine, University of California–San Francisco, San Francisco, CA, United States
[10]Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel
[11]Center for Healthcare Policy and Research, UC Davis Health, Sacramento, CA, United States
[12]Department of Public Health Sciences, UC Davis School of Medicine, Sacramento, CA, United States

Corresponding Author:
Adrian Aguilera, PhD
School of Social Welfare
University of California Berkeley
205 Haviland Hall
Berkeley, CA, 94709
United States
Phone: 1 510 642 8564
Email: aguila@berkeley.edu

Abstract

Background: Digital and mobile health interventions using personalization via reinforcement learning algorithms have the potential to reach large number of people to support physical activity and help manage diabetes and depression in daily life.

Objective: The Diabetes and Mental Health Adaptive Notification and Tracking Evaluation (DIAMANTE) study tested whether a digital physical activity intervention using personalized text messaging via reinforcement learning algorithms could increase step counts in a diverse, multilingual sample of people with diabetes and depression symptoms.

Methods: From January 2020 to June 2022, participants were recruited from 4 San Francisco, California–based public primary care clinics and through web-based platforms to participate in the 24-week randomized controlled trial. Eligibility criteria included English or Spanish language preference and a documented diagnosis of diabetes and elevated depression symptoms. The trial had 3 arms: a Control group receiving a weekly mood monitoring message, a Random messaging group receiving randomly selected feedback and motivational text messages daily, and an Adaptive messaging group receiving text messages selected by a reinforcement learning algorithm daily. Randomization was performed with a 1:1:1 allocation. The primary outcome, changes in daily step counts, was passively collected via a mobile app. The primary analysis assessed changes in daily step count using a linear mixed-effects model. An a priori subanalysis compared the primary step count outcome within recruitment samples.

Results: In total, 168 participants were analyzed, including those with 24% (40/168) Spanish language preference and 37.5% (63/168) from clinic-based recruitment. The results of the linear mixed-effects model indicated that participants in the Adaptive

XSL•FO
RenderX

arm cumulatively gained an average of 3.6 steps each day (95% CI 2.45-4.78; *P*<.001) over the 24-week intervention (average of 608 total steps), whereas both the Control and Random arm participants had significantly decreased rates of change. Postintervention estimates suggest that participants in the Adaptive messaging arm showed a significant step count increase of 19% (606/3197; *P*<.001), in contrast to 1.6% (59/3698) and 3.9% (136/3480) step count increase in the Random and Control arms, respectively. Intervention effectiveness differences were observed between participants recruited from the San Francisco clinics and those recruited via web-based platforms, with the significant step count trend persisting across both samples for participants in the Adaptive group.

Conclusions: Our study supports the use of reinforcement learning algorithms for personalizing text messaging interventions to increase physical activity in a diverse sample of people with diabetes and depression. It is the first to test this approach in a large, diverse, and multilingual sample.

Trial Registration: ClinicalTrials.gov NCT03490253; https://clinicaltrials.gov/study/NCT03490253

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2019-034723

(J Med Internet Res 2024;26:e60834) doi:10.2196/60834

KEYWORDS

digital health; physical activity; mobile phone; text messages; SMS; steps; walking; diabetes; depression; reinforcement learning; exercise; machine learning

Introduction

Noncommunicable diseases such as type 2 diabetes and depression are significant public health problems individually and interact to worsen outcomes for each other [1]. Depression is associated with unhealthy behaviors such as physical inactivity and poor dietary habits, which are also risk factors for the development of type 2 diabetes [2]. The psychosocial burden of living with a chronic illness such as diabetes can contribute to the development and worsening of depression—leading to a cycle of multiplicative negative health effects. Diabetes and depression exhibit racial and socioeconomic disparities stemming from known barriers from the individual to the structural levels, such as the stressors due to discrimination and poverty, difficulty in accessing care related to language barriers, competing health demands, and beyond [3,4].

Physical activity is a core intervention for both type 2 diabetes and depression, as both are associated with low physical activity levels [5,6]. Increased physical activity has positive effects on both conditions, including improved glycemic control in diabetes and reduced symptoms of depression [7]. Behavioral activation interventions that incorporate physical activity as a core element are among the most effective for treating depression, relative to other behavioral and psychological approaches [8,9]. Targeting physical activity is thus an efficient method of improving outcomes in both diabetes and depression.

Digital and mobile health interventions have the potential to reach a large number of people to support physical activity and help manage diabetes and depression in daily life, with a large research base using digital technology to improve and personalize health behavior programs [10,11]. Machine learning algorithms can help personalize and optimize interventions by tailoring them to the individual's specific needs, preferences, and capabilities. Reinforcement learning algorithms, in particular, have the potential to create adaptive and dynamic interventions focused on a goal or reward that can adjust based on an individual's responses and behaviors [12]. In previous work these algorithms increased physical activity among patients with diabetes [13]. However, novel data science methods such as personalization via reinforcement learning algorithms typically do not include data from marginalized populations (ie, ethnic and racialized minoritized populations and those from low-income backgrounds) [14]. It is crucial that digital platforms and newer algorithms are built with diverse populations to increase their generalizability and effectiveness, particularly to make an impact on entrenched health inequities.

The Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) study aimed to address these gaps by developing and evaluating a reinforcement learning-based digital health app co-designed with English- and Spanish-speaking adults with comorbid diabetes and depression symptoms. The app used reinforcement learning algorithms to tailor daily text messages based on feedback type, motivation type, and message timing. The algorithm training data included demographic (eg, age), clinical data (eg, baseline depression scores), and contextual data (eg, timing and category of messages previously sent) [15]. Here, we present the efficacy of the DIAMANTE intervention, in a 3-arm randomized trial, on daily step count outcomes.

Methods

Study Design

We conducted a randomized controlled trial (RCT) with 3 arms from January 2020 to June 2022. We used the Standard Protocol Items: Recommendations for Interventional Trials checklist for reporting our findings [16]. A study protocol and a correction to the protocol outlining changes in recruitment approach were published in 2020 and 2023, respectively [15,17].

Participants

RCT recruitment occurred on 2 tracks. Clinic-based patient recruitment at the San Francisco Health Network (SFHN, the public health care delivery system for the city and county of San Francisco, California) was our first method of enrolling study participants. These clinic-based strategies were

XSL•FO
RenderX

significantly impacted by the COVID-19 pandemic. Prior to March 2020 and after California's reopening in June 2021, patients were recruited by direct provider referrals, with in-person clinic recruitment timed to eligible patient visits. To identify potentially in-person eligible patients, we asked for permission from known providers to access patient lists and then reviewed and identified patients with documented electronic health records diagnosis of diabetes and elevated depression symptoms (depression diagnosis or Patient Health Questionnaire (PHQ-8) score of 5 or higher in the past 5-year period) and whether they were adults aged 18 years or older. Eligible patients were required to use text messaging and own a smartphone in order to download the pedometer app into their phone. Patients were excluded if they were unable to walk or were pregnant at the time of the study. Researchers then contacted patients directly (via in-person visits and via phone calls) to determine interest in joining the DIAMANTE study and review study eligibility criteria. Interested individuals were invited for a study visit that included obtaining informed consent and completing a baseline survey. All patients were offered assistance in downloading the DIAMANTE app onto their phone and sent test text messages back to our system. The researcher established a baseline plan for physical activity goals with the patient (ie, average 4000 steps daily) and instructed patients to have the app open at all times. To replicate real-world conditions and increase generalizability, participants were not provided with any instructions about how much, when, or where in their body to carry their phones. All participants were invited back at 24-weeks to complete a follow-up survey when the active intervention period ended.

We added a second web-based recruitment strategy during the COVID-19 pandemic, given that in-clinic recruitment was not possible for many months during the active enrollment period for our RCT. Specifically, between March 2020 and October 2021, we recruited through social media advertisements via advertisements posted on Facebook, Google Ads, and Craigslist. Potential participants completed a screening survey to assess eligibility (ie, self-reported diabetes diagnosis and PHQ-8). In addition to the exclusion criteria, if participants were unable to walk or currently pregnant, we also excluded those who did not have an eligible phone number, were outside of the United States, or those who failed the CAPTCHA requirements to verify their identity on the web. Web-based participants were enrolled remotely, offering Zoom (Zoom Video Communications, Inc) meeting and phone assistance to complete all enrollment procedures (eg, onboarding, downloading the app, and setting a step count goal) as needed. All informed consent and survey data collection matched our in-clinic recruitment procedures.

Ethical Considerations

The study was approved by the University of California San Francisco (UCSF) institutional review board (17-22608). The project coordinator and research assistants were responsible for managing patient data collection. Self-reported participant data were stored on REDCap (Research Electronic Data Capture; Vanderbilt University), surveys were stored on UCSF Qualtrics, and daily step count and text messaging data were stored on the HealthySMS platform [18]. To maintain patient privacy and confidentiality only institutional review board–approved research staff had access to the data collection platforms. Data were downloaded from their respective servers and stored in a secure UCSF Box Folder. All participants gave written or electronic informed consent in the language of their choice, English or Spanish, prior to participating in the study. Patients received a compensation of US $40 in cash for participating in the baseline questionnaire and an additional US $70 in cash for the 24-week follow-up questionnaire.

Randomization

Randomization was performed with block randomization (block size 3) and a 1:1:1 allocation into arms (adaptive messaging, random messaging, and app only), stratified by patient language preference (English vs Spanish). Upon sign-up, participants were automatically randomized via a randomization table in the HealthySMS system (developed by author CK) into one of the study conditions after they were onboarded into the study, thereby ensuring allocation concealment. Patients were informed of the nature of the app and frequency of the messages they would be receiving. The necessity of these steps made it unfeasible to fully blind participants or research staff; however, the final data analyses were completely blinded.

Procedures

In brief, the DIAMANTE study compared the effectiveness of different text messaging strategies on physical activity (measured by step counts on participants' smartphones) over an active 24-weeks intervention period. We created a custom mobile phone app titled "DIAMANTE" developed by Audacious Software for this study. This application passively tracks step counts by pooling from Google Fit, Apple HealthKit, or the built-in pedometer on patients' phones. We used a text messaging platform HealthySMS, previously developed by Dr Aguilera and Audacious Software, to integrate with the DIAMANTE app and send text messages by intervention arm. The DIAMANTE app only needed to be installed once and then remain open consistently. The app was designed in English and Spanish versions and was freely available as a download from the Apple App Store and Android Google Play App.

Figure 1 shows the different intervention groups during the trial period. The three comparator arms were (1) Control arm, who received only 1 weekly message inquiring about their mood over the past week, (2) Random messaging arm, who received daily randomly selected messages (1 feedback message on their step count [eg, "You walked 4000 steps yesterday"] and 1 motivational message about their health [eg, "Going for a walk can improve your mood and clear your mind]) from a preexisting bank of messages, and (3) Adaptive messaging arm, who received daily messages from the same feedback and motivational messaging banks, with message categories and times selected by a reinforcement learning algorithm. The Random and Adaptive arms also received the weekly mood message. In summary, the feedback messages were selected from 5 message types (step number, step number plus encouragement or reinforcement, relative [walked more or less relative to goal], relative [achieved or did not achieve goal], or no message), and motivation messages could be from 4 categories of messages based on the COM-B (capability,

XSL•FO
RenderX

opportunity, motivation, and behavior) model (each with 18 message options in these categories, or a "no message" option). Messages were generated from behavioral science theoretical domains and final messages were selected based on user-centered design methods [19]; examples are shown in Table 1. Moreover, patients and end users provided feedback on the text messages, mobile app interface, and study protocols during user-centered design phases of the study [20].

Figure 1. Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) CONSORT (Consolidated Standards of Reporting Trials) diagram.

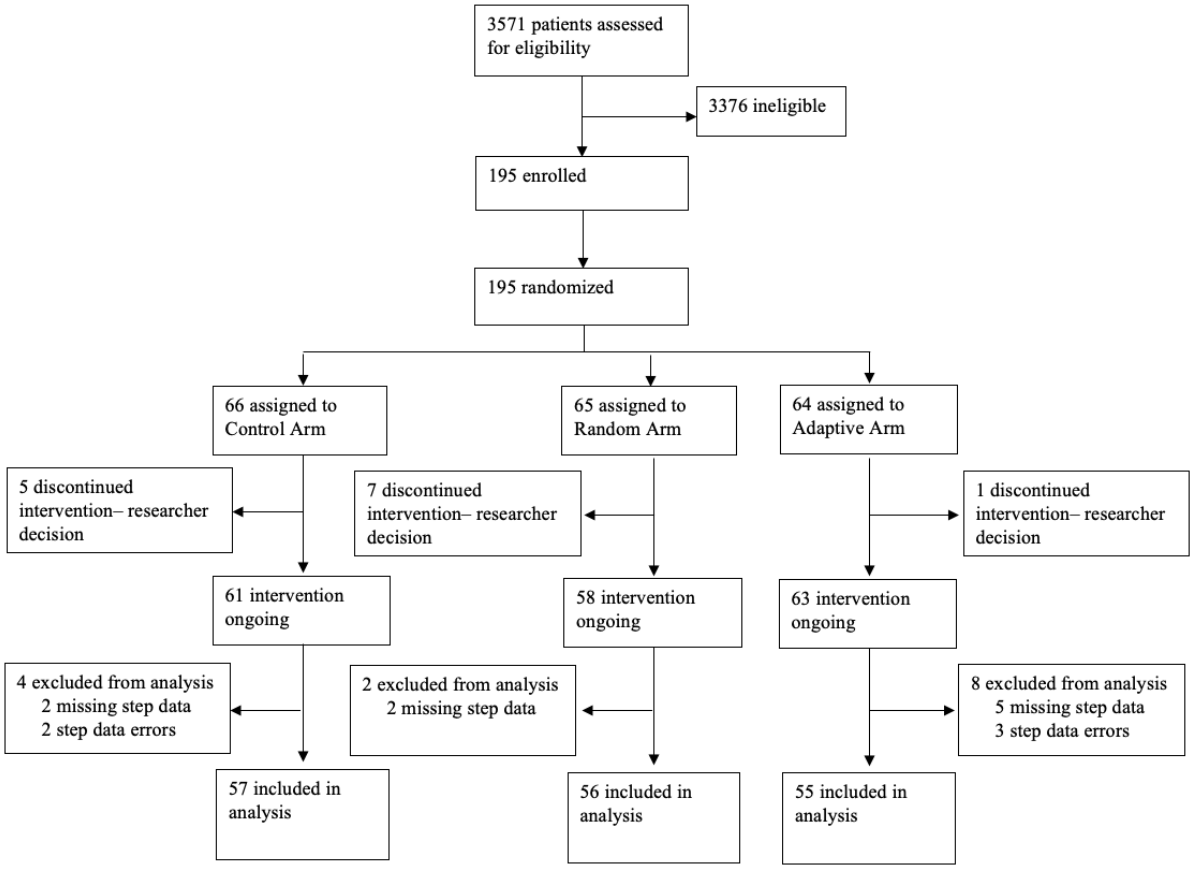

XSL•FO
RenderX

Table 1. Feedback and motivational messages and algorithm message choices.

	Examples	Random (n=51), %	Adaptive (n=53), %	*P* value
A: Feedback messages				
0. No feedback message		19.5	6.0	<.001
1. Reaching goal	"Yesterday, you did not reach your goal"	19.9	20.8	.35
2. Steps walked yesterday	"Yesterday, you walked 3824 steps"	20.5	28.1	<.001
3. Walked more/less today than yesterday	"Yesterday, you walked more than your goal"	19.7	25.6	<.001
4. Steps walked yesterday, plus a positive/negative motivational message	"You walked 8000 steps yesterday. Great job!"	20.4	19.5	.54
B: Motivational messages				
0. No message	__[a]	24.3	9.3	<.001
1. Capability, describes the physical and psychological benefits of walking and exercise	"Doing more physical activity can help reduce feelings of fatigue"	24.8	37.2	<.001
2. Motivation, meant to increase self-confidence and the belief that one is capable of walking even in the face of challenges	"You have made changes to improve your health before, you can do it again"	25.6	19.3	<.001
3. Opportunity, physical and social environment cues that make it more likely to engage in walking	"It there a local park you have been waiting to visit? Use it as an opportunity to get out of the house and do more steps!"	25.2	34.2	<.001

[a]Not applicable.

We developed the reinforcement learning algorithm (a type of machine learning) based on previous work [15]. The reinforcement learning algorithm had a "cold-start" beginning with a randomly selected message for each patient and used ongoing participant data for personalizing (1) the feedback message category selection, (2) the motivational message category selection, and (3) the timing of the message delivery daily (into 4 time periods from 8 AM to 8 PM)—upweighting messages and times that had a higher probability of increasing steps. As described previously, we used reinforcement learning algorithms for contextual multiarmed bandit problems [21], as these can maximize cumulative rewards in sequential decision tasks (as here, which sequences of messages optimally promote the highest step count in the upcoming day). We also used Thompson Sampling, a Bayesian method that can handle small amounts of data [22], which allowed the algorithm to continuously learn which feedback and motivational messages were effective for a user, based on contextual features like their previous step counts and which messages were sent previously, as well as participant data such as demographic and clinical characteristics (such as age, gender, and depressive symptoms scores collected at baseline). In summary, each morning within the adaptive intervention arm, the algorithms evaluated which categories of feedback and motivational messages would likely increase the step count for each participant in the upcoming day and at which time period the messages should be delivered to increase step count. The rates of algorithm message choices are shown in Table 1.

Researchers monitored the incoming step data to ensure that the system was collecting data to inform the algorithm and adaptive messages. Patients in all groups received reminders to open the app if no data were being transmitted from the DIAMANTE app to the HealthySMS system. An automated algorithm within HealthySMS triggered daily text messages to participants if the system did not receive step count data from the DIAMANTE app. The research team contacted the patients by phone for troubleshooting and to remind them to keep the app open in the background if patients' phones were not transmitting data for more than 3 days (after the automated SMS reminders). In addition, patients in all groups could reply "STOP" or "PARAR" if they wished to stop receiving messages.

Outcomes

Our primary outcome for the RCT, total change in daily step counts, was passively collected by a mobile phone application during the time that patients remain in the intervention (using Apple Healthkit, Google Fit, and pedometer data from participants' phones). We anticipated some variability between methods of data collection but considered this to be a trade-off for greater generalizability and broader implementation. As specified in our Protocol paper [15], the a priori sample size was 276 patients to account for the 3 intervention arms and with 80% power and up to 15% (41/276) participant dropout. Similarly, prior to analysis, participants were required to have at least 28 days of step count data during the intervention in order to be included in the primary analysis, given that the algorithm needed sufficient step count data to be able to personalize content for each participant. Data quality was evaluated for all participants upon study exit (blinded by study arm), with removal of 3 participants with less than 28 days of step count data, 6 participants with no step count data, and 5 participants whose data signaled a major discrepancy (>25% of available daily step counts greater than 20,000 steps), as this may have indicated a flaw in the DIAMANTE app step count data extraction.

XSL•FO
RenderX

Statistical Analysis

First, we reported on the overall demographics of the full sample and by intervention arm and overall step counts within the sample. Our per protocol primary analysis included all participants who had at least 28 days of steep count data and used longitudinal regression to determine the impact of the treatment arms on the change in step count for each participant. We analyzed changes in daily step count using a linear mixed-effects model (LMM) with an intervention variable (Adaptive arm as reference group, Control arm, and Random arm) and a time variable (day, Control × day, Random × day) as fixed effects and a random intercept for each participant. We plotted the raw step count data and the residuals to examine their distributions via histograms. As specified in our protocol, we also completed 1 a priori subanalysis: the impact of the intervention by recruitment sample (SFHN clinic-based recruitment vs web-based recruitment). We analyzed descriptive data in SPSS (version 28; IBM Corp) and the LMM analysis using the "lme4" package in R.

Results

The research team recruited a total of 195 participants into the RCT between January 2020 and June 2022. The CONSORT (Consolidated Standards of Reporting Trials) diagram is shown in Figure 1. A total of 13 participants were dropped by the researchers due not installing or immediately deleting the DIAMANTE app, and another 14 were excluded because of insufficient (n=9) or incorrect (n=5) step count data. The final analytical sample included 168 participants, each of them is observed for 168 days. More than one-third of participants were recruited from the SFHN (63/168, 37.5%) and the remaining were recruited via the web (105/168, 62.5%). Baseline demographic characteristics stratified by the intervention arm are reported in Table 2. Overall, the sample participants were middle-aged (mean age 49, SD 12.1 years), predominantly female (104/168, 62%), and majority were English speakers (128/168, 76%). The sample was diverse in terms of race or ethnicity, education, and employment status. Finally, due to data transmission errors (eg, not having the app open) there was an average of 20 missing days of step data (SD 32.72) for study participants, and the average step count within the sample was 3221 steps at baseline and 3783 at follow-up. By arm, participants in the Control arm had an average of 3701 steps (SD 3054.53) on day 1 of the intervention and an average of 26 days of missing data (SD 36.91); participants in the Random arm had an average of 3471 steps on day 1 (SD 3274.78) and an average of 20 days of missing data (SD 32.54). Finally, Adaptive arm participants had an average of 2277 steps on day 1 (SD 2250.87) and an average of 14 days of missing data (SD 27.2).

XSL•FO
RenderX

Table 2. Baseline demographics.

	Control (n=57)	Random (n=56)	Adaptive (n=55)	Total (N=168)
Age (years), mean (SD)	51 (13.2)	49 (11.0)	48 (12.1)	49 (12.1)
Sex, n (%)				
Female	35 (61)	38 (68)	31 (56)	104 (62)
Male	20 (35)	18 (32)	23 (42)	61 (36)
Language, n (%)				
English	42 (74)	44 (79)	42 (76)	128 (76)
Spanish	15 (26)	12 (21)	13 (24)	40 (24)
Race/ethnicity, n (%)				
Asian or Pacific Islander	4 (7)	4 (7)	4 (7)	12 (7)
Black or African American	11 (19)	9 (16)	7 (13)	27 (16)
White or Caucasian	14 (25)	17 (30)	19 (34)	50 (30)
Latinx or Hispanic	22 (39)	20 (36)	22 (40)	64 (38)
Multiracial/Ethnic	6 (11)	6 (11)	3 (6)	15 (9)
Education, n (%)				
Some high school or less	11 (19)	10 (18)	8 (14)	29 (17)
High school graduate	7 (12)	8 (14)	9 (16)	24 (14)
Some college or technical school	9 (16)	17 (30)	18 (33)	44 (26)
College graduate	24 (42)	16 (29)	15 (27)	55 (33)
Graduate degree	6 (11)	5 (9)	5 (9)	16 (10)
Employment, n (%)				
Full time (>35 hours/week)	19 (33)	19 (34)	22 (40)	60 (36)
Part-time (<35 hours/week)	9 (16)	15 (27)	5 (9)	29 (17)
Homemaker	2 (4)	3 (5)	6 (11)	11 (6)
Unemployed	8 (14)	9 (16)	10 (18)	27 (16)
Disabled/on disability	11 (19)	6 (11)	10 (18)	27 (16)
Retired	7 (12)	4 (7)	1 (2)	12 (7)
Depression scores				
PHQ-8[a], mean (SD)	10.23 (6.54)	12.0 (6.67)	10.42 (6.0)	10.88 (6.42)

[a]PHQ-8: Patient Health Questionnaire-8.

The results of the primary LMM are shown in Table 3. The results indicate that on average, at the start of the intervention, participants in the Adaptive group had 3197 steps, the Control group had 3480 steps, and Random group had 3698 steps. The random intercept effects were estimated at 5,419,080.25 (SD 2237.39), indicating a large amount of variability in step count between participants. There were no statistically significant differences in baseline step count between intervention groups. For the Adaptive group there was a significant increase of 3.6 steps (P<.001) each day of the 24-week intervention; cumulatively and estimating for the 168-day time period, participants in the Adaptive group would have gained an average of 606 steps indicating a 19% increase in daily step count. The Control (P=.001) and Random (P<.001) groups had statistically significantly lower rates of change (albeit positive) in daily step count relative to Adaptive group. The Control group increased linearly by 0.81 steps per day for a total of 136 steps or an increase of 3.9% daily steps from day 1 to day 168 of the intervention. Finally, the Random group increased linearly by 0.35 steps per day for a total of 59 steps per day indicating an increase of 1.6% daily steps from day 1 to day 168 of the intervention (Figure 2).

XSL•FO
RenderX

Table 3. Linear mixed-effects model daily step count.

Parameter	Estimate	95% CI lower limit	95% CI upper limit	*P* value
Intercept (Adaptive)	3196.79	2568.03	3825.54	<.001
Control	283.32	–598.67	1165.32	.53
Random	500.92	–384.54	1386.36	.27
Day 1	3.61	2.45	4.78	<.001
Control (day 1)	–2.80	–4.47	–1.13	.001
Random (day 1)	–3.26	–4.92	–1.60	<.001

Figure 2. Linear trajectory of step count over 168 days of study.

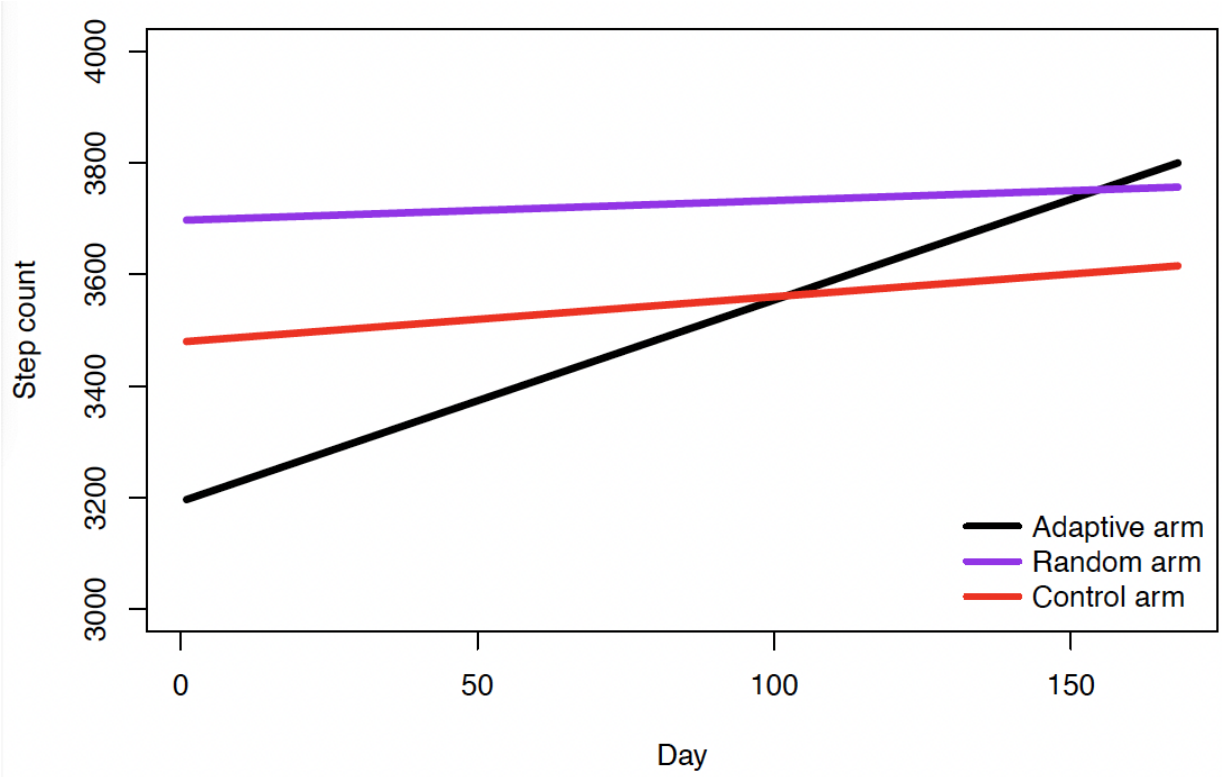

Based on our a priori subanalysis, the LMM results indicated statistically significant positive rates of change in daily step count for Adaptive group participants in both the SFHN subsample (average baseline steps 3278, daily rate of change 6.0 steps [*P*<.001], 30.6% increase in daily step count post intervention) and the web-based subsample (average baseline steps 3156, daily rate of change 2.3 steps (*P*=.002), 12.2% increase in daily step count post intervention). These results were in line with the overall analysis. However, in the SFHN sample the daily steps rate of change for the Random and Control groups was statistically significantly different and negative relative to the Adaptive group. Whereas for the web-based sample the daily rate of change for both the Random and Control groups was positive but not statistically significantly different than that of the Adaptive group.

Discussion

Our study found that applying a reinforcement learning algorithm for personalizing text messaging interventions is an effective approach for increasing physical activity in a diverse sample of people with diabetes and depression. We found that significant step count increases over 6 months (168 days) in the adaptive, reinforcement learning intervention arm, as compared with participants who received randomly selected text message content (Random messaging arm) or only a weekly mood self-monitoring message (Control arm). We found that reinforcement learning algorithms can increase the effectiveness of a physical activity–based digital health intervention in a diverse, real-world sample with comorbid diabetes and depression symptoms. Our findings are particularly significant because most digital health studies in vulnerable populations to date have been pilot studies, and the use of machine learning methods for personalization is rarely applied to low-income and Spanish-speaking populations. These findings also support previous studies using a similar reinforcement learning algorithms [12].

Personalization (particularly those based on content that might be more meaningful for participants) has been one of the core challenges for the digital health field. Until now, many digital health studies show high dropout and low engagement of digital

XSL•FO
RenderX

health interventions, also linked to a lack of personalization [23]. Here, we show that reinforcement learning holds promise for the future of digital health, especially given the rapid recent advancements in artificial intelligence. Our study importantly personalized the content using existing health behavior theory (COM-B), such as motivational content as opportunity cues for behavior change, which are likely important components for future machine learning interventions [24]. In this study, it appears that participants in the Adaptive messaging arm received messages that were more relevant in content as evidenced by the varying rates of message selection in the Adaptive arm versus the Random arm of the intervention. Furthermore, the real-world implementation of this study demonstrates the importance of establishing both efficacy and effectiveness of machine learning interventions [25].

We did find some differences in our primary step count outcome by recruitment source, with the clinic-based sample receiving more benefit from the intervention than the web-based recruitment sample (30% overall step count gain vs 12% overall step count gain among Adaptive group participants). The significant clinic-based sample findings are particularly of note, given that this population is primarily insured via Medicaid and receiving primary care within a public health care delivery system, which is reflected in the much higher proportion of Spanish speakers and individuals with lower educational attainment who typically do not have access to these types of interventions. There are various possible explanations for these differences that should be studied in future work including reaching a population that has less access to these types of interventions, better integration into care, or benefiting from co-design with similar populations.

A key strength of this study is the inclusion of a diverse sample in terms of race and ethnicity and language, education, and employment status. As machine learning algorithms continue to be developed and used in health interventions, it is crucial that the basis for these algorithmic approaches includes diverse and real-world samples so that algorithmic biases are minimized and their findings and potential benefits can be broadly shared. The results of this RCT provide valuable insights into the potential of adaptive messaging interventions to increase physical activity, especially in groups that are typically excluded from digital health research. If this intervention is widely disseminated and implemented, it can decrease health inequities by providing a type of personalized care for communities historically and presently marginalized.

Our study also highlights the need for further research to understand how digital health interventions can be tailored to different populations to maximize their effectiveness. Our future research will further investigate the parameters that influenced algorithmic decision-making, regarding the patterns of sending feedback messages, motivational messages, and timing of message delivery. It is also important to understand whether overall categories of messages were more impactful on behavior change and to understand how the algorithm can be refined to be more effective in achieving the goal of increasing physical activity in diverse samples.

Limitations of this study include challenges with data accuracy from phones. We chose to emphasize replicability in the real word and thus did not provide users with wearable devices or standardize the mobile phones that users had to have. This resulted in variability in data collection and some data collection errors. Furthermore, in order for these types of interventions to scale, we must maximize adoption and relevance (eg, use of devices that users already own) alongside precision (eg, “better” measurement using new devices).

Another limitation is that we did not specifically assess engagement with the messages. Similar to other text message interventions, we do not know whether or when people actually read messages; we only know when messages were sent. Finally, due to the unforeseen circumstances from the COVID-19 pandemic, recruitment number for our project was also less than our goal stated in the RCT protocol; however, our effect size was still large enough to be detected in the primary analysis.

The DIAMANTE intervention has shown that reinforcement learning algorithms can be used to improve the personalization of physical activity interventions in a diverse, multilingual sample with diabetes and depression symptoms. Moving forward, artificial intelligence and machine learning interventions will rapidly expand, and real-world studies targeting diverse end users are critical, especially when built upon co-design with end users.

Acknowledgments

This study received funding from the Agency for Healthcare Research and Quality (AHRQ, R01HS025429).

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AA and CL were responsible for conceptualization methodology; project administration; writing—original draft, review, and editing; and funding acquisition. MRAA conducted analysis; data curation; and writing—original draft, review, and editing. XJ conducted analysis. BC and CF were responsible for design and analysis. FG and KR contributed to writing—original draft and project administration. RH-R performed writing—review and editing and project administration. CK contributed software. JW

XSL•FO
RenderX

contributed to methodology and resources. LO-F provided resources and consultation. US was responsible for writing—review and editing, and consultation. EY-T assisted with writing—review and editing, and methodology.

Conflicts of Interest

AA is the owner of HealthySMS and licenses the software to other researchers. CL has conducted separate research projects with digital health companies and was a visiting researcher at Google from 2022 to 2023 on the health equity team. The rest of the coauthors do not have any conflicts of interest to disclose.

Multimedia Appendix 1
CONSORT eHealth Checklist (v1.6.1).
[PDF File (Adobe PDF File), 1190 KB - jmir_v26i1e60834_app1.pdf]

References

1. Bădescu SV, Tătaru C, Kobylinska L, Georgescu EL, Zahiu DM, Zăgrean AM, et al. The association between diabetes mellitus and depression. J Med Life 2016;9(2):120-125 [FREE Full text] [Medline: 27453739]
2. Cho YG. The interrelationship between diabetes and depression. Korean J Fam Med 2014;35(3):109-110 [FREE Full text] [doi: 10.4082/kjfm.2014.35.3.109] [Medline: 24921028]
3. Everson SA, Maty SC, Lynch JW, Kaplan GA. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. J Psychosom Res 2002;53(4):891-895. [doi: 10.1016/s0022-3999(02)00303-3] [Medline: 12377299]
4. Hansen MC, Cabassa LJ. Pathways to depression care: help-seeking experiences of low-income Latinos with diabetes and depression. J Immigr Minor Health 2012;14(6):1097-1106 [FREE Full text] [doi: 10.1007/s10903-012-9590-x] [Medline: 22367667]
5. Dubash S. The interplay of depression symptoms and physical activity: bidirectional insights from 25-years of the Americans' changing lives panel. Ment Health Phys Act 2024;26:100599. [doi: 10.1016/j.mhpa.2024.100599]
6. Morrato EH, Hill JO, Wyatt HR, Ghushchyan V, Sullivan P. Physical activity in U.S. adults with diabetes and at risk for developing diabetes, 2003. Diabetes Care 2007;30(2):203-209. [doi: 10.2337/dc06-1128] [Medline: 17259482]
7. Narita Z, Inagawa T, Stickley A, Sugawara N. Physical activity for diabetes-related depression: a systematic review and meta-analysis. J Psychiatr Res 2019;113:100-107. [doi: 10.1016/j.jpsychires.2019.03.014] [Medline: 30928617]
8. Dang K, Ritvo P, Katz J, Gratzer D, Knyahnytska Y, Ortiz A, et al. The role of daily steps in the treatment of major depressive disorder: secondary analysis of a randomized controlled trial of a 6-month internet-based, mindfulness-based cognitive behavioral therapy intervention for youth. Interact J Med Res 2023;12:e46419 [FREE Full text] [doi: 10.2196/46419] [Medline: 38064262]
9. Lambert JD, Greaves CJ, Farrand P, Price L, Haase AM, Taylor AH. Web-based intervention using behavioral activation and physical activity for adults with depression (The eMotion Study): pilot randomized controlled trial. J Med Internet Res 2018;20(7):e10112 [FREE Full text] [doi: 10.2196/10112] [Medline: 30012547]
10. Lee JA, Choi M, Lee SA, Jiang N. Effective behavioral intervention strategies using mobile health applications for chronic disease management: a systematic review. BMC Med Inform Decis Mak 2018;18(1):12 [FREE Full text] [doi: 10.1186/s12911-018-0591-0] [Medline: 29458358]
11. De Santis KK, Jahnel T, Matthias K, Mergenthal L, Al Khayyal H, Zeeb H. Evaluation of digital interventions for physical activity promotion: scoping review. JMIR Public Health Surveill 2022;8(5):e37820 [FREE Full text] [doi: 10.2196/37820] [Medline: 35604757]
12. Yom-Tov E, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Hochberg I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. J Med Internet Res 2017;19(10):e338 [FREE Full text] [doi: 10.2196/jmir.7994] [Medline: 29017988]
13. Hochberg I, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Yom-Tov E. Encouraging physical activity in patients with diabetes through automatic personalized feedback via reinforcement learning improves glycemic control. Diabetes Care 2016;39(4):e59-e60. [doi: 10.2337/dc15-2340] [Medline: 26822328]
14. Lyles CR, Nguyen OK, Khoong EC, Aguilera A, Sarkar U. Multilevel determinants of digital health equity: a literature synthesis to advance the field. Annu Rev Public Health 2023;44:383-405 [FREE Full text] [doi: 10.1146/annurev-publhealth-071521-023913] [Medline: 36525960]
15. Aguilera A, Figueroa CA, Hernandez-Ramos R, Sarkar U, Cemballi A, Gomez-Pathak L, et al. mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE study. BMJ Open 2020;10(8):e034723 [FREE Full text] [doi: 10.1136/bmjopen-2019-034723] [Medline: 32819981]
16. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013;158(3):200-207 [FREE Full text] [doi: 10.7326/0003-4819-158-3-201302050-00583] [Medline: 23295957]

XSL•FO
RenderX

17. Aguilera A, Figueroa CA, Hernandez-Ramos R, Sarkar U, Cemballi A, Gomez-Pathak L, et al. mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE study. BMJ Open 2020;10(8):e034723 [FREE Full text] [doi: 10.1136/bmjopen-2019-034723] [Medline: 32819981]
18. Aguilera A, Bruehlman-Senecal E, Demasi O, Avila P. Automated text messaging as an adjunct to cognitive behavioral therapy for depression: a clinical trial. J Med Internet Res 2017;19(5):e148 [FREE Full text] [doi: 10.2196/jmir.6914] [Medline: 28483742]
19. Pathak LE, Aguilera A, Williams JJ, Lyles CR, Hernandez-Ramos R, Miramontes J, et al. Developing messaging content for a physical activity smartphone app tailored to low-income patients: user-centered design and crowdsourcing approach. JMIR Mhealth Uhealth 2021;9(5):e21177 [FREE Full text] [doi: 10.2196/21177] [Medline: 34009130]
20. Avila-Garcia P, Hernandez-Ramos R, Nouri SS, Cemballi A, Sarkar U, Lyles CR, et al. Engaging users in the design of an mHealth, text message-based intervention to increase physical activity at a safety-net health care system. JAMIA Open 2019;2(4):489-497 [FREE Full text] [doi: 10.1093/jamiaopen/ooz049] [Medline: 32025646]
21. Tewari A, Murphy S. In: Rehg JM, Murphy SA, Kumar S, editors. From Ads to Interventions: Contextual Bandits in Mobile Health. Switzerland: Springer International Publishing AG; 2017:495-517.
22. Agrawal S, Goyal N. Analysis of thompson sampling for the multi-armed bandit problem. arXiv. Preprint posted online November 8, 2011:1111.1797 2011:1-21. [doi: 10.48550/arXiv.1111.1797]
23. Gan DZQ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022;8:20552076221098268 [FREE Full text] [doi: 10.1177/20552076221098268] [Medline: 35677785]
24. McEwan D, Beauchamp MR, Kouvousis C, Ray CM, Wyrough A, Rhodes RE. Examining the active ingredients of physical activity interventions underpinned by theory versus no stated theory: a meta-analysis. Health Psychol Rev 2019;13(1):1-17. [doi: 10.1080/17437199.2018.1547120] [Medline: 30412685]
25. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. J Med Internet Res 2019;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]

Abbreviations

COM-B: capability, opportunity, motivation, and behavior
CONSORT: Consolidated Standards of Reporting Trials
DIAMANTE: Diabetes and Mental Health Adaptive Notification Tracking and Evaluation
LMM: linear mixed-effects model
PHQ-8: Patient Health Questionnaire-8
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
SFHN: San Francisco Health Network
UCSF: University of California San Francisco

Edited by G Eysenbach, N Cahill; submitted 23.05.24; peer-reviewed by L Bell, T Qian, S Ren, K Watanabe; comments to author 14.06.24; revised version received 05.07.24; accepted 23.08.24; published 08.10.24.

Please cite as:
Aguilera A, Arévalo Avalos M, Xu J, Chakraborty B, Figueroa C, Garcia F, Rosales K, Hernandez-Ramos R, Karr C, Williams J, Ochoa-Frongia L, Sarkar U, Yom-Tov E, Lyles C
Effectiveness of a Digital Health Intervention Leveraging Reinforcement Learning: Results From the Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) Randomized Clinical Trial
J Med Internet Res 2024;26:e60834
URL: https://www.jmir.org/2024/1/e60834
doi:10.2196/60834
PMID:39378080

©Adrian Aguilera, Marvyn Arévalo Avalos, Jing Xu, Bibhas Chakraborty, Caroline Figueroa, Faviola Garcia, Karina Rosales, Rosa Hernandez-Ramos, Chris Karr, Joseph Williams, Lisa Ochoa-Frongia, Urmimala Sarkar, Elad Yom-Tov, Courtney Lyles. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Influence of Physical Activity and Diet Mobile Apps on Cardiovascular Disease Risk Factors: Meta-Review

Erica Bushey[1], MS; Yin Wu[1,2], PhD; Alexander Wright[1], MS; Linda Pescatello[1], PhD

[1]University of Connecticut, Storrs, CT, United States

[2]Hartford Hospital, Hartford, CT, United States

Corresponding Author:
Erica Bushey, MS
University of Connecticut
2098 Hillside Rd
Storrs, CT
United States
Phone: 1 860 486 0008
Email: erica.bushey@uconn.edu

Abstract

Background: The literature on whether physical activity (PA) and PA and diet (PA+Diet) mobile apps improve cardiovascular disease (CVD) risk factors is promising.

Objective: The aim of this meta-review is to provide an evidence synthesis of systematic reviews and meta-analyses examining the influence of PA and PA+Diet apps on the major CVD risk factors.

Methods: We systematically searched 5 databases until January 12, 2022. Included systematic reviews and meta-analyses (1) reported the CVD risk factor outcomes of BMI, waist circumference, body weight, blood pressure (BP), hemoglobin A_{1c} (HbA_{1c}), fasting blood glucose, blood lipids, or PA; (2) enrolled healthy participants ≥18 years who may or may not have the metabolic syndrome, diabetes mellitus, or preexisting CVD risk factors; (3) reviewed PA or PA+Diet app interventions integrating behavioral change techniques (BCT) to deliver their information; and (4) had a nonapp control.

Results: In total, 17 reviews (9 systematic reviews and 8 meta-analyses) published between 2012 and 2021 qualified. Participants were middle-aged, mostly women ranging in number from 10 to 62,219. Interventions lasted from 1 to 24 months, with the most common behavioral strategies being personalized feedback (n=8), self-monitoring (n=7), and goal setting (n=5). Of the PA app systematic reviews (N=4), the following CVD risk factors improved: body weight and BMI (n=2, 50%), BP (n=1, 25%), HbA_{1c} (n=1, 25%), and blood lipids (n=1, 25%) decreased, while PA (n=4, 100%) increased. Of the PA+Diet app systematic reviews (N=5), the following CVD risk factors improved: body weight and BMI (n=3, 60%), BP (n=1, 20%), and HbA_{1c} (n=3, 60%) decreased, while PA (n=3, 60%) increased. Of the PA app meta-analyses (N=1), the following CVD risk factors improved: body weight decreased (–0.73 kg, 95% CI –1.45 to –0.01; *P*=.05) and PA increased by 25 minutes/week (95% CI 0.58-1.68; *P*<.001), while BMI (–0.09 kg/m^2, 95% CI –0.29 to 0.10; *P*=.35) and waist circumference (–1.92 cm, 95% CI –3.94 to 0.09; *P*=.06) tended to decrease. Of the PA+Diet app meta-analyses (n=4), the following CVD risk factors improved: body weight (n=4, 100%; from –1.79 kg 95% CI –3.17 to –0.41; *P*=.01 to –2.80 kg 95% CI –4.54 to –1.06, *P*=.002), BMI (n=1, 25%; –0.64 kg/m^2, 95% CI –1.09 to –0.18; *P*=.01), waist circumference (n=1, 25%; –2.46 cm, 95% CI –4.56 to –0.36; *P*=.02), systolic/diastolic BP (n=1, 25%; –4.22/–2.87 mm Hg, 95% CI –6.54 to –1.91/ –4.44 to –1.29; *P*<.01), and HbA_{1c} (n=1, 25%; –0.43%, 95% CI –0.68 to –0.19; *P*<.001) decreased.

Conclusions: PA and PA+Diet apps appear to be most consistent in improving PA and anthropometric measures with favorable but less consistent effects on other CVD risk factors. Future studies are needed that directly compare and better quantify the effects of PA and PA+Diet apps on CVD risk factors.

Trial Registration: PROSPERO CRD42023392359; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=392359

(J Med Internet Res 2024;26:e51321) doi:10.2196/51321

KEYWORDS

physical activity; diet; mobile applications; obesity; hypertension; dyslipidemia; diabetes; mobile phone

XSL•FO
RenderX

Introduction

Background

Cardiovascular Disease (CVD) is the leading cause of death in the United States and the world [1], with an estimated 1 in 3 adults dying from CVD [2]. Between 2018 and 2019, direct and indirect costs of CVD were US $407.3 billion, making it the costliest disease in the United States [3]. Identifying those at risk for CVD is important due to the significant socioeconomic burden CVD imposes [1]. The major CVD risk factors are physical inactivity, obesity, diabetes mellitus (DM), dyslipidemia, and hypertension, with a prevalence ranging from 11% for DM to 75% for physical inactivity among US adults [1,2,4-8]. Half of US adults have 1 or more CVD risk factors [4-8]. Professional guidelines recommend lifestyle modifications, notably physical activity (PA) and diet, as critical first steps to prevent and treat CVD and its risk factors [2,9-11].

Mobile health apps have proliferated because of their accessibility, affordability, increased smartphone usage, and improved technology [12]. The global health care mobile app market size was valued at US $17.92 billion in 2019 and is expected to grow at a mean annual growth rate of 45% from 2020 to 2027 [12]. PA and Diet (PA+Diet) apps dominate the health App market with a share of 54.8% in 2021 [13]. Indeed, the market size of fitness and health apps is expected to increase globally by 46% from US $17.92 billion between 2020 and 2027 [13]. Evidence is growing that PA and PA+Diet apps have key functions that favorably impact CVD risk factors such as PA and hypertension [9-11,14-19]. These features include enabling patient education and selfcare through the use of self-monitoring and self-management, remote monitoring of patients with clinician online care encounters through digital platforms, and research to determine intervention effectiveness [14].

Due to the popularity of PA and PA+Diet apps, the number of systematic reviews, meta-analyses, and randomized controlled trials (RCT) investigating the influence of these apps on CVD risk factors has increased substantially [20-26]. Indeed, we conducted searches in PubMed with terms related to the various types of reviews on January 11, 2023, and found 1011 “systematic reviews, meta-analyses, and RCTs involving PA apps” from 2010 to 2023; and 268 “systematic reviews, meta-analyses, and RCTs involving PA+Diet apps” from 2009 to 2023 [20]. A meta-review is a growing type of evidence synthesis that gathers systematic reviews and meta-analyses and evaluates and summarizes the research questions being examined [27]. Surprisingly, to the best of our knowledge, there are no meta-reviews evaluating the impact of PA and PA+Diet apps on CVD risk factors.

Purpose

We performed this meta-review to provide the first evidence synthesis of systematic reviews and meta-analyses examining the influence of PA and PA+Diet apps on the major CVD risk factors of physical inactivity, obesity, DM, dyslipidemia, and hypertension.

Methods

Search Strategy

This meta-review is reported consistent with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [28,29]. The PRISMA diagram of the search strategy is found in Figure 1. With the assistance of a medical librarian, we systematically searched 5 databases (PubMed, Scopus, SportDiscus, Cumulative Index to Nursing and Health Literature, and Cochrane Library) from inception to January 12, 2022 (see Multimedia Appendix 1 for the detailed list of search terms). The protocol is registered in PROSPERO (CRD42023392359).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) search strategy.

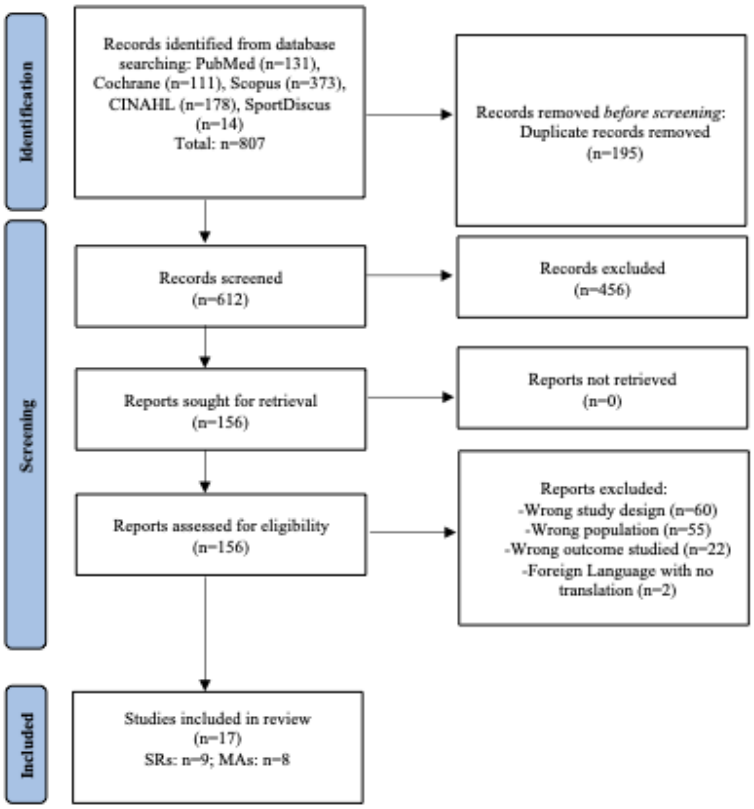

To be classified as an app, the app integrated a software program designed to run on a mobile device such as a phone or tablet that reported CVD risk factor outcomes [13]. A PA intervention within a PA app included strategies to increase PA that ranged from prescribed exercise programs, reminders to exercise, or recommendations to meet the PA guidelines of 150 to 300 minutes/week of moderate-intensity or 75 to 150 minutes/week of vigorous-intensity aerobic PA or some combination of both [6,30]. A diet intervention within a PA+Diet app included strategies to improve eating behaviors that range from keeping daily food logs, nutritional counseling, and meeting the dietary guidelines of a healthy dietary pattern to meet nutrient needs, help achieve a healthy body weight, and reduce the risk of chronic disease [5,31]. PA and PA+Diet app interventions also included behavioral change techniques (BCT) (ie, goal setting and social support) to deliver their intervention [12].

Included systematic reviews and meta-analysis of PA and PA+Diet apps: (1) reported the CVD risk factor outcomes of anthropometrics (ie, BMI, waist circumference, and body weight); blood pressure (BP; ie, systolic blood pressure [SBP] and diastolic blood pressure [DBP]); glycemic biomarkers (ie, fasting blood glucose, hemoglobin A_{1c} [HbA_{1c}]), blood lipid-lipoproteins (ie, total cholesterol, triglycerides, high-density lipoprotein [HDL], and low-density lipoprotein [LDL]), or PA volume (ie, minutes/week, steps/day); (2) enrolled healthy individuals ≥18 years who may or may not have metabolic syndrome, diabetes mellitus, or preexisting CVD risk factors; (3) reviewed PA or PA+Diet app interventions integrating BCTs to deliver their information; and (4) had a non-app control that consisted of usual care or a waitlist.

Data Extraction and Coding

A systematic review screening tool, Rayyan, was used to triage the 807 potentially qualifying studies [32]. After duplicates were removed, the 612 potentially relevant remaining reports were screened first by title, abstract, and full text (more details in Figure 1). For the 17 qualifying reports (n=9 systematic reviews and n=8 meta-analyses), data were extracted related to the study characteristics (eg, publication year), sample characteristics (eg, gender, age), intervention characteristics (eg, app features), and outcomes (eg, BMI, waist circumference, BP). Reference lists of all included studies were manually searched and cross-referenced for additional reports. Data were extracted from the qualifying studies using a standardized coding form and coding manual previously developed and used by members of our research team [33]. A total of 2 trained coders (EB, AW) independently extracted and coded study information. All disagreements were resolved through discussion (with a third investigator if necessary). The outcomes of the PA and PA+Diet app systematic reviews (n=9) and meta-analyses (n=8) are presented in Table 1. For the systematic reviews, CVD risk factor outcomes were coded as either the CVD risk factor outcomes favored PA or PA+Diet apps over control, or there was no difference in CVD risk factor outcomes between PA or PA+Diet apps over control. For the MAs, CVD risk factor outcomes were coded by the standardized and unstandardized effect sizes. We categorized the effect size as (1) Favors PA or PA+Diet apps when the effect size was negative (or positive for HDL and PA) and statistically significant, (2) Favors Control when the effect size was positive (or negative for HDL and PA) and statistically significant, and (3) No Difference between PA

XSL•FO
RenderX

or PA+Diet apps vs control when the effect size was not statistically significant.

Table 1. Physical activity, and physical activity and diet app systematic reviews (n=9) and meta-analyses (n=8) cardiovascular disease risk factor outcomes.

Outcome	PA[a] apps		PA and diet apps	
	Systematic reviews (n=4)	Meta-analyses (n=1)	Systematic reviews (n=5)	Meta-analyses (n=7)
BMI	—[b]	X[c]	X	X
Body weight	X	X	X	X
Waist circumference	—	—	—	X
Blood pressure	X	—	X	X
Hemoglobin A_{1c}	X	—	X	X
Fasting blood glucose	—	—	—	X
Total cholesterol	X	—	—	—
Triglycerides	—	—	—	—
High-density lipoprotein	—	—	—	X
Low-density lipoprotein	X	—	—	—
Physical activity	X	X	X	X

[a]PA: physical activity.

[b]Not available.

[c]X indicates which outcomes were assessed.

Study Methodological Quality Assessment

Study methodological quality was assessed by 2 trained coders (EB and AW) independently using A Measurement Tool for Assessment of Multiple Systematic Reviews (AMSTAR) checklist containing 18 items [34]. Systematic reviews meeting <9 (50%) of the items were rated as low, those meeting between 9 and <13 (50%-69%) of the items were rated as moderate, and those meeting >13 (>70%) were rated as high study methodological quality. All disagreements were resolved through discussion (with a third investigator if necessary). See Multimedia Appendix 2 for a summary of the methodological quality assessment from the AMSTAR Checklist.

Results

The PRISMA diagram of the search strategy is found in Figure 1 (checklist in Multimedia Appendix 3). We identified 807 potentially qualifying systematic reviews and meta-analyses. Only 1 primary-level study was included in 4 of the identified reviews, so the overlap in primary studies across the included reviews was minimal [35]. Out of these, we excluded 195 duplicates, 456 articles based on title and abstract, and 156 articles on full-text review. We identified 17 reports (9 systematic reviews and 8 meta-analyses) involving 269 studies satisfying the inclusion criteria. Of the 9 systematic reviews, 4 evaluated PA apps [36-39], and 5 evaluated PA+Diet apps [40-44]. Out of the 8 meta-analyses, 1 evaluated PA apps [45], and 7 evaluated PA+Diet apps [46-52].

Study Characteristics

Multimedia Appendix 4 contains the descriptions of the sample, study, and intervention characteristics of the qualifying systematic reviews or meta-analyses. The systematic reviews (n=9) and meta-analyses (n=8) involving PA and PA+Diet apps were published between 2012 and 2021 with most research groups located in the United States (5/17, 29%) and United Kingdom (5/17, 29%) followed by Spain (3/17, 18%), Taiwan (1/17, 6%), Canada (1/17, 6%), Korea (1/17, 6%), and Australia (1/17, 6%). The systematic reviews (n=4) and meta-analysis (n=1) involving PA apps only were published between 2018 and 2021, with most research groups located in the United States (2/5, 40%) followed by Canada (1/5, 20%), the United Kingdom (1/5, 20%), and Australia (1/5, 20%). The systematic reviews (n=5) and meta-analysis (n=7) involving PA+Diet apps were published between 2012 and 2021, with most research groups located in the United Kingdom (4/12, 33%), Spain (3/12, 25%), and the United States (3/12, 25%) followed by Taiwan (1/12, 8%), and Korea (1/12, 8%).

Sample Characteristics

The systematic reviews (n=9) and meta-analyses (n=8) involving PA and PA+Diet apps had participants ranging in number from 10 to 62,219 [37,42,45-47,50] who were mostly middle-aged women ranging in age from 18 to 67 years [37,42,45-47,50]. No qualifying systematic reviews and meta-analyses involving PA and PA+Diet apps disclosed the ethnicity or race or the baseline CVD risk factors values of the study participants. The systematic reviews (n=4) and meta-analysis (n=1) involving PA apps only (n=5) had participants ranging in age from 10 to 69, 219 who were mostly middle-aged women ranging in age from 18 to 65 years. The systematic reviews (n=5) and meta-analysis (n=7) involving PA+Diet apps (n=12) had participants ranging in number from 10 to 1,386, who were mostly middle-aged women ranging in age from 18 to 67 years. The systematic reviews (n=2) and meta-analysis (n=8) involving

XSL•FO
RenderX

PA and PA+Diet apps conducted risk of bias assessments, 8 of which used the Cochrane Collaboration tool for assessing the risk of bias in randomized trials. Overall, the risk of bias between the 10 reviews was low to moderate.

Intervention Characteristics

The systematic reviews (n=9) and meta-analyses (n=8) involving PA and PA+Diet apps had interventions that lasted between 1 and 24 months, with the most common durations being 6 weeks (3/17, 18%) and 8 weeks (3/17, 18%). The systematic reviews (n=4) and meta-analyses (n=1) involving PA apps only lasted between 6 weeks to 12 months, with the most common duration being 6 weeks (2/5, 40%). The systematic reviews (n=5) and meta-analyses (n=7) involving PA+Diet apps lasted between 4 weeks to 24 months, with the most common duration being 8 weeks (3/12, 25%). The systematic reviews (n=9) and meta-analyses (n=8) involving PA and PA+Diet apps used personalized feedback (8/17, 47%), self-monitoring (7/17, 41%), and goal setting (5/17, 29%) as the most common behavioral strategies. Other behavioral strategies included education (2/17, 12%), PA recommendations (1/17, 6%), reminders (1/17, 6%), and social support (1/17, 6%). The systematic reviews (n=4) and meta-analyses (n=1) involving PA apps only used personalized feedback (1/5, 20%), goal setting (1/5, 20%), and PA reminders (1/5, 20%) as the most common behavioral strategies. The systematic reviews (n=5) and meta-analyses (n=7) involving PA+Diet apps used personalized feedback (7/12, 58%), self-monitoring (7/12, 58%), and goal setting (4/12, 33%) as the most common behavioral strategies. The systematic reviews (n=9) and meta-analyses (n=8) involving PA and PA+Diet apps used either web-based applications (12/17, 71%) or mobile phone apps (11/17, 65%) as their platform. The systematic reviews (n=4) and meta-analyses (n=1) involving PA apps only (n=5) used either mobile phone apps (3/5, 60%) or web-based applications (2/5, 40%) as their platform. The systematic reviews (n=5) and meta-analysis (n=7) involving PA+Diet apps used either mobile phone apps (8/12, 67%) or web-based applications (10/12, 83%) as their platform. Of note, the systematic reviews (n=9) and meta-analyses (n=8) involving PA and PA+Diet apps did not disclose the adherence rate of how frequently the participants used the interventions.

CVD Risk Factors

The systematic reviews (n=9) involving PA and PA+Diet apps for the CVD risk factors outcomes are presented in Table 2. Of the systematic reviews (n=4) involving PA apps, the following CVD risk factors improved: body weight and BMI (2/4, 50%), BP (1/4, 25%), HbA_{1c} (1/4, 25%), and blood lipids (1/4, 25%) decreased, while PA (4/4, 100%) increased. Of the systematic reviews (n=5) involving PA+Diet apps, the following CVD risk factors improved: body weight and BMI (3/5, 60%), BP (1/5, 20%), and HbA_{1c} (3/5, 60%) decreased, while PA (3/5, 60%) increased.

Table 2. The influence of physical activity, and physical activity and diet apps on cardiovascular disease risk factors outcomes among the qualifying systematic reviews (n=9).

Outcome	PA[a] apps		PA + Diet apps	
	Favors intervention, n	No difference intervention vs Control, n	Favors intervention, n	No difference in the intervention vs Control, n
BMI	1 [38]	1	1 [41]	2 [42,44]
Body weight	2 [36,38]	__[b]	2 [43,44]	—
Waist circumference	—	—	—	—
Systolic blood pressure	1 [38]	—	1 [44]	2 [41,42]
Diastolic blood pressure	—	—	1 [44]	2 [41,42]
Hemoglobin A_{1c}	—	—	3 [40-42]	—
Fasting blood glucose	—	—	—	—
Total cholesterol	—	—	—	—
Triglycerides	—	—	—	1 [41]
High-density lipoprotein	1 [38]	—	—	—
Low-density lipoprotein	1 [38]	—	—	—
Physical activity measures	4 [36-39]	—	3 [40,41,44]	1 [42]

[a]PA: physical activity.

[b]Not available.

We categorized the effect sizes as (1) Favors PA or PA+Diet apps when the effect size was negative (or positive for HDL and PA) and statistically significant; (2) Favors Control when the effect size was positive (or negative for HDL and PA) and statistically significant; and (3) No difference between PA or PA+Diet apps vs Control when the effect size was not statistically significant. No systematic review was reported to favor control.

The meta-analyses (n=8) involving PA and PA+Diet apps unstandardized and standardized effect sizes for the CVD risk factors outcomes are presented in Table 3 and Figure 2. Of the meta-analyses reporting unstandardized effect sizes (n=1)

XSL•FO
RenderX

involving PA apps, these CVD risk factors improved: body weight decreased (–0.73 kg, 95% CI –1.45 to –0.01; *P*=.05) and PA increased by 25 minutes/week (95% CI 0.58-1.68; *P*<.001) while BMI (–0.09 kg/m^2, 95% CI –0.29 to 0.10; *P*=.35) and waist circumference (–1.92 cm, 95% CI –3.94 to 0.09; *P*=.06) tended to decrease. Of the meta-analyses reporting unstandardized effect sizes (n=4) involving PA+Diet apps, these CVD risk factors improved: body weight (4/4, 100%; effect sizes ranged from –1.79 kg (95% CI –3.17 to –0.41; *P*=.01) to –2.27 kg (95% CI –3.64 to –0.90; *P*<.01), BMI (1/4, 25%; effect size: –0.64 kg/m^2, 95% CI –1.09 to –0.18, *P*=.01), waist circumference (1/4, 25%; effect size: –2.46 cm, 95% CI –4.56 to –0.36, *P*=.02), systolic/diastolic BP (1/4, 25%; effect size: –4.22/–2.87 mm Hg, 95% CI –6.54 to –1.91/ –4.44 to –1.29; *P*<.01), and HbA_{1c} (1/4, 25%; effect size: –0.43%, 95% CI –0.68 to –0.19, *P*<.001) decreased.

XSL•FO
RenderX

Table 3. The influence of physical activity and physical activity and diet apps on the cardiovascular disease risk factors among the qualifying meta-analyses (n=8).

Outcome	PA[a] apps effect sizes		PA + Diet apps effect sizes	
	Favors intervention	No difference intervention vs Control[b]	Favors intervention	No difference intervention vs Control[b]
BMI	__[c]	n=1 [45]; unstandardized effect size: –0.09 kg/m^2, 95% CI –0.29 to 0.10; *P*=.35	n=2 [46,48]; standardized effect size: –0.46, 95% CI –0.68 to –0.23; *P*=.04; unstandardized effect size: –0.64 kg/m^2, 95% CI –1.09 to –0.18; *P*=.01	n=1 [52]; standardized effect size: –0.14 95% CI –0.51 to 0.23; *P*=.45
Body weight	n=1 [45]; unstandardized effect size: –0.73 kg, 95% CI –1.45-0.01; *P*=.05	—	n=6 [47-52]; standardized effect size n=2: –0.37, 95% CI –0.54 to –0.19; *P*<.001 to –0.43, 95% CI –0.252 to –0.609; *P*<.01; unstandardized effect size n=4: –1.79 kg 95% CI –3.17 to 0.41; *P*=.01 to –2.27 kg 95% CI –3.64 to –0.90; *P*<.01	—
Waist circumference	—	n=1 [45]; unstandardized effect size: –1.92 cm, 95% CI –3.94 to 0.09; *P*=.06	n=2 [46,48]; standardized effect size: –0.54, 95% CI –0.85 to –0.23; *P*=.001; unstandardized effect size: –2.46 cm 95% CI –4.56 to –0.36; *P*=.02	—
Systolic blood pressure	—	—	n=2 [46,48]; standardized effect size: –0.43, 95% CI –0.77 to –0.10; *P*<.001; unstandardized effect size: –4.22 mm Hg 95% CI –6.54 to –1.91; *P*<.01	—
Diastolic blood pressure	—	—	n=2 [46,48]; standardized effect size: –0.44, 95% CI –0.74 to –0.14; *P*=.002; unstandardized effect size: –2.87 mm Hg 95% CI –4.44 to –1.29; *P*<.01	—
Hemoglobin A_{1c}	—	—	n=1 [49]; unstandardized effect size: –0.43% 95% CI –0.6 to –0.19; *P*<.001	n=1 [46]; standardized effect size: –0.35, 95% CI –0.82 to 0.12, *P*<.001
Fasting blood glucose	—	—	n=1 [46]; standardized effect size: –0.39, 95% CI –0.73 to –0.06; *P*=.002	—
Total cholesterol	—	—	—	n=1 [46]; standardized effect size: –0.06, 95% CI –0.23 to 0.11; *P*=.47
Triglycerides	—	—	—	n=1 [46]; standardized effect size: –0.20, 95% CI –0.35 to –0.04; *P*=.80
High-density lipoprotein	—	—	n=1 [46]; standardized effect size: 0.23, 95% CI 0.07-0.38; *P*=.50	—
Low-density lipoprotein	—	—	—	n=1 [46]; standardized effect size: –0.04, 95% CI –0.21 to 0.13; *P*=.96
Physical activity measures	n=1 [45]; standardized effect size: 1.13 (~25 minutes/week moderate to vigorous PA), 95% CI 0.58-1.68; *P*<.001	—	n=2 [49,52]; standardized effect size: 0.21, 95% CI 0.05-0.37; *P*=.01 to 2.59, 95% CI 1.00-4.81; *P*=.001	—

[a]PA: physical activity.

[b]No effect size was reported to favor control.

[c]Not available.

XSL•FO
RenderX

Figure 2. The number of statistically significant unstandardized and standardized effect sizes reported in the meta-analysis (n=8) for the improvement in cardiovascular disease risk factors by physical activity (PA) and physical activity and diet (PA+Diet) apps.

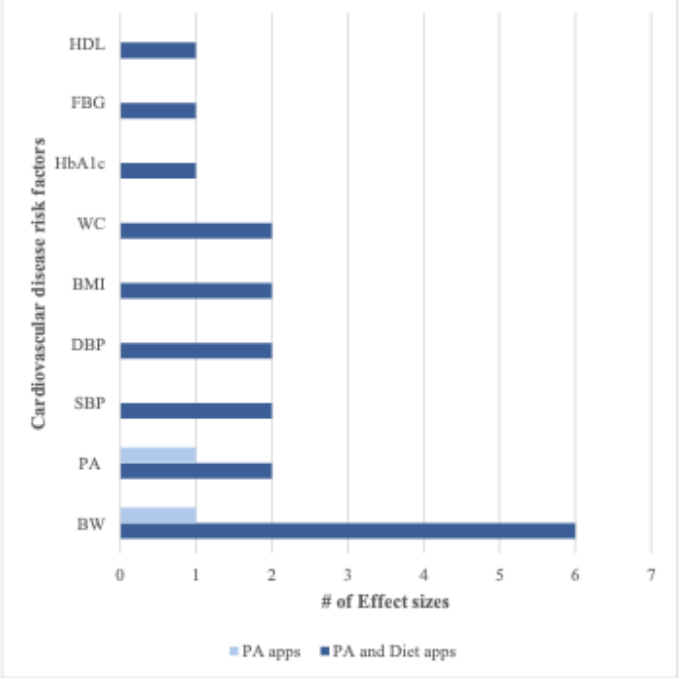

*%= # of statistically significant effect sizes for the cardiovascular disease risk factor/ # of effect sizes reported for the cardiovascular disease risk factor x100.

Of the meta-analyses reporting standardized effect sizes (n=3) involving PA+Diet apps, these CVD risk factors improved: body weight (2/3, 66%; effect sizes ranged from –0.37 (95% CI –0.54 to –0.19; *P*<.001) to –0.43 (95% CI –0.252 to –0.609; *P*<.01), BMI (1/3, 33%; effect size: –0.46, 95% CI –0.68 to –0.23; *P*=.04), waist circumference (1/3, 33%; effect size: –0.54, 95% CI –0.85 to –0.23; *P*=.001), systolic/diastolic BP (1/3, 33%; effect size: –0.43/–0.44, 95% CI –0.77 to –0.10/ –0.74 to –0.14; *P*<.001; *P*=.002), fasting blood glucose (1/3, 33%; effect size: –0.39, 95% CI –0.73 to –0.06; *P*=.002), HDL (1/3, 33%; effect size: 0.23, 95% CI 0.07-0.38, *P*=.50), and PA (2/3, 66%; effect sizes ranged from 0.21 (95% CI 0.05-0.37, *P*=.01) to 2.59 (95% CI 1.00-4.81, *P*=.001); while HbA_{1c} (1/3, 33%; effect size: –0.35, 95% CI –0.82 to 0.12; *P*<.001), total cholesterol (1/3, 33%; effect size: –0.06, 95% CI –0.23 to 0.11; *P*=.47), triglycerides (1/3, 33%; effect size: –0.20, 95% CI –0.35 to –0.04; *P*=.80), and LDL (1/3, 33%; effect size: –0.04, 95% CI –0.21 to 0.13; *P*=.96) tended to decrease.

Study Methodological Quality Assessment

The overall methodological study quality of the literature was moderate to high, with 35% scoring low, 29% moderate, and 35% high on the AMSTAR Checklist. Given the overall study, the methodological study quality was moderate to high; we conclude there is moderate certainty of evidence in our findings. See Multimedia Appendix 2 for a summary of the methodological quality assessment from the AMSTAR Checklist.

Discussion

Principal Findings

Our meta-review provides the first evidence synthesis of systematic reviews and meta-analyses examining the influence of PA and PA+Diet apps on the major CVD risk factors of physical inactivity, obesity, DM, dyslipidemia, and hypertension. For the systematic reviews involving PA apps, body weight and BMI improved compared to control. For the meta-analyses involving PA apps, body weight decreased by 0.73 kg compared with control. For the systematic reviews involving PA apps, HbA_{1c} improved compared with control. For the systematic reviews involving PA apps, blood lipids improved compared to control. For the systematic reviews involving PA apps, PA improved compared to control. For the meta-analyses involving PA apps, PA increased by 25 minutes/week compared to control. For the systematic reviews involving PA+Diet apps, body weight and BMI improved compared to control. For the meta-analyses involving PA+Diet apps, BMI decreased by 0.64 kg/m^2, and waist circumference decreased by 2.46 cm compared to control. For the systematic reviews involving PA+Diet apps, BP improved compared to control. For the meta-analyses involving PA+Diet apps, BP decreased by 3-4 mm Hg compared to control. For the systematic reviews involving PA+Diet apps, HbA_{1c} improved compared to control. For the meta-analyses involving PA+Diet Apps, HbA_{1c} decreased by 0.40 units compared to control. For the systematic reviews involving PA+Diet apps, PA improved compared to control. We found that 70% of the included systematic reviews and meta-analyses had moderate to high study methodological quality ratings on

XSL•FO
RenderX

the AMSTAR Checklist [34]. Our meta-review suggests that both PA and PA+Diet apps resulted in clinically meaningful improvements in major CVD risk factor outcomes. However, the size of the literature involving systematic reviews and meta-analyses of PA+Diet apps (n=12) is larger than PA apps (n=5), and the findings were more robust for anthropometric and PA outcomes than the other major CVD risk factor outcomes of BP, fasting blood glucose, HbA_{1c}, and HDL.

Comparison With Previous Works

Consistent with our findings, other studies have shown that even modest reductions in anthropometric CVD risk factor outcomes (ie, body weight, BMI, and waist circumference) can lead to clinically important improvements in other CVD risk factor outcomes [17,53-55]. Aucott et al [56] found a 4.5-kg decrease in body weight was associated with improvements in multiple CVD risk factors, including BP, blood lipids, and fasting blood glucose levels. Aune et al [15] and Zhou [57] found a 1-kg/m^2 decrease in BMI and a 1-cm decrease in waist circumference [57] associated with significant improvements in BP, total cholesterol, fasting blood glucose, and triglyceride levels [15] and a 2% to 4% decrease in the risk of developing CVD [57]. The decrease in body weight of 2.27 kg, BMI of 0.64 kg/m^2, and waist circumference of 2.46 cm resulting from PA+Diet apps reported by McMahon et al [48] in our meta-review suggest the improvements in measures of body composition they observed have the potential to positively impact other CVD risk factors. Lear et al [54] found that participating in 30 minutes of daily PA was associated with a 27% lower risk of CVD incidence and a 30% lower risk of CVD mortality than being physically inactive. The finding of increases in PA of 25 minutes/week resulting from PA apps reported by Cotie et al [45] in our meta-review is not as large as Lear et al [54] but this amount of weekly PA would have positive effects on the risk of CVD incidence and mortality as well [45]. Lewington et al [58] found a 3-mm Hg reduction in SBP and a 2-mm Hg reduction in DBP led to a 5%-6% reduction in the risk of CVD events. The finding of decreases in BP of 3-4 mm Hg resulting from PA+Diet apps reported by McMahon et al. [48] in our meta-review suggests they, too, may reduce the risk of CVD events. Improving one or multiple CVD risk factors can significantly reduce the risk of developing CVD [9-11]. A meta-analysis performed on the influence of body weight and BP found that body weight reductions of 5.1 kg through diet, PA, or both reduced SBP by –4.44 mm Hg and DBP by –3.57 mm Hg [10]. In addition, increases in PA tend to increase HDL by 1-2 mg/dl and decrease triglycerides by –4 to 12 mg/dl [11]. Collectively, our meta-review findings provide promising evidence that PA and PA+Diet apps that include behavioral change strategy features may lead to clinically important improvements in major CVD risk factor outcomes and overall cardiovascular health.

The PA and PA+Diet apps systematic reviews (n=9) and meta-analyses (n=8) in our meta-review integrated a variety of behavioral change strategies, the most common of which were personalized feedback (8/17, 47%), self-monitoring (7/17, 41%), and goal setting (5/17, 29%). The systematic reviews (n=4) and meta-analyses (n=1) involving PA apps primarily used personalized feedback (1/4, 20%), goal setting (1/4, 20%), and PA reminders (1/4, 20%), while the systematic reviews (n=5) and meta-analyses (n=7) involving PA+Diet apps primarily used personalized feedback (7/12, 58%), self-monitoring (7/12, 58%), and goal setting (4/12, 33%). All major professional organizations recommend the integration of behavioral change strategies in interventions to promote healthy lifestyles [59]. Michie et al [60] developed the Behavior v1 Change Technique Taxonomy, a classification system based upon an expert consensus of the reliable and transparent behavior change strategies that can be used by practitioners with confidence. Out of the 93 distinct behavioral change strategies included in their classification system, those with some of the highest levels of interrater agreement were self-monitoring, goal setting, and feedback. The systematic reviews and meta-analyses in our meta-review, consistent with their overall moderate to high study methodological quality, integrated some of the most highly regarded behavioral change strategies from this classification scheme within their PA and PA+Diet app interventions. However, the most common intervention duration was relatively short (6 and 8 weeks). Given the shorter duration, it remains unclear to what extent the behavioral change strategies were an “active ingredient” in the CVD risk factor improvements that resulted from the PA and PA+Diet app interventions [60]. It is also unclear from the systematic reviews and meta-analyses in out meta-review if there are long-term effects that are sustained following these PA and PA+Diet app interventions [60].

CVD remains the leading cause of death in the United States, accounting for 928,741 deaths in the year 2020 [1]. The PA and PA+Diet app market to prevent and treat CVD and its risk factors has escalated because they are cost-effective, easy to use, accessible, and time-efficient [12-14]. The mobile health app market was valued at US $25.92 billion in 2020 and is expected to grow to US $348.98 billion by 2027 [12]. PA+Diet apps have dominated the mobile health app market with a share of 54.6% in 2022 [13]. The favorable findings from this meta-review on the use of PA and PA+Diet apps to improve major CVD risk factors combined with the growth of the mobile health app market indicates the future potential of this industry to favorably impact CVD and its risk factors is untapped [12-14].

Strengths and Limitations

There are limitations to this meta-review. First, baseline values of the CVD risk factor outcomes were not disclosed. It is well established that the response of a CVD risk factor is a direct function of baseline values [61], except perhaps HDL [62]. Baseline values are necessary for the interpretation of the response to health outcomes as the response of a health outcome is a direct function of the baseline value, which is also needed to determine the relative percentage of improvement over the course of the app interventions. Since baseline values were not provided, it is difficult to fully appreciate the clinical implications of the resultant CVD risk factor changes. Second, the systematic reviews and meta-analyses did not disclose ethnicity, nor did they report the frequency of app use within the intervention period. Third, non-English studies were omitted from this review. Despite these limitations, our meta-review has several strengths. The systematic search methods adhered to the contemporary PRISMA standards [28,29] and included

XSL•FO
RenderX

the assessment of the study's methodological quality [34]. Finally, to the best of our knowledge, this meta-review provides the first evidence synthesis of systematic reviews and meta-analyses examining the influence of PA and PA+Diet apps on the major CVD risk factors of physical inactivity, obesity, DM, dyslipidemia, and hypertension.

Future Directions

In conclusion, we found PA and PA+Diet apps lead to improvements in anthropometric and PA outcomes with favorable but less consistent effects on other CVD risk factors outcomes of BP, fasting blood glucose, HbA_{1c}, and HDL. These findings hold promise because the prevalence of CVD and its risk factors are so high [1,2,4-8], and the health apps market is proliferating [13]. Despite the positive findings of our meta-review, there is a need for future RCTs that directly compare and quantify the effects of PA and PA+Diet apps on CVD risk factor outcomes and evaluate which behavioral strategies within them are most effective. In this way, the prevention and treatment of CVD can capitalize on the momentum of the rapidly growing health app market.

Acknowledgments

We would like to express gratitude to Hilary Kraus, MSI, Research Services Librarian at the University of Connecticut, for her assistance in the development of our search strategy. We would also like to thank the University of Connecticut Center for Excellence in Teaching and Learning (CETL) for their support.

Disclaimer

Generative artificial intelligence was not used in any portion of the manuscript writing.

Authors' Contributions

EB performed conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, visualization, writing (original draft), and writing (review and editing). YW contributed to conceptualization, formal analysis, methodology, project administration, resources, supervision, validation, and writing (original draft). AW performed investigation and writing (review and editing). LP contributed to conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing (original draft), and writing (review and editing).

Conflicts of Interest

LSP is the founder and sole proprietor of P3-EX LLC, which could potentially benefit from this research.

Multimedia Appendix 1
A Summary of the Secondary Literature Systematic Search Strategy.
[DOCX File , 19 KB - jmir_v26i1e51321_app1.docx]

Multimedia Appendix 2
A Summary of the Assessment of the Study Methodological Quality of the Included Meta-Analyses (N=17).
[DOCX File , 31 KB - jmir_v26i1e51321_app2.docx]

Multimedia Appendix 3
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 197 KB - jmir_v26i1e51321_app3.pdf]

Multimedia Appendix 4
A Summary of the Sample, Study, and Intervention Characteristics of the included RCTs within the Qualifying Systematic Reviews and Meta-Analyses (N=17).
[DOCX File , 26 KB - jmir_v26i1e51321_app4.docx]

References

1. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2023 update: a report from the American heart association. Circulation 2023;147(8):e93-e621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
2. NA NA. National center for health statistics. About multiple cause of death, 1999-2020. CDC 2022.

XSL•FO
RenderX

3. Cardiovascular Disease: A Costly Burden for American- Projections Through 2035. 2017. URL: https://www.heart.org/-/media/Files/About-Us/Policy-Research/Fact-Sheets/Public-Health-Advocacy-and-Research/CVD-A-Costly-Burden-for-America-Projections-Through-2035.pdf [accessed 2024-09-13]
4. National diabetes statistics report website. CDC. 2022. URL: https://www.cdc.gov/diabetes/php/data-research/index.html [accessed 2024-09-13]
5. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity and severe obesity among adults: United States, 2017-2018. NCHS Data Brief 2020(360):1-8 [FREE Full text] [Medline: 32487284]
6. Adult physical inactivity prevalence maps by race/ethnicity. CDC. 2022. URL: https://archive.cdc.gov/www_cdc_gov/physicalactivity/data/inactivity-prevalence-maps/2015-2018.html#:~:text=Hispanic%20adults%20(31.7%25)%20had,Hispanic%20White%20adults%20(23.4%25) [accessed 2024-09-13]
7. Goff DC, Bertoni AG, Kramer H, Bonds D, Blumenthal RS, Tsai MY, et al. Dyslipidemia prevalence, treatment, and control in the Multi-Ethnic study of atherosclerosis (MESA): gender, ethnicity, and coronary artery calcium. Circulation 2006;113(5):647-656. [doi: 10.1161/CIRCULATIONAHA.105.552737] [Medline: 16461837]
8. Hypertension cascade: hypertension prevalence, treatment and control estimates among U.S. adults aged 18 years and older applying the criteria from the American college of cardiology and American heart association's 2017 hypertension guideline. CDC. URL: https://millionhearts.hhs.gov/data-reports/hypertension-prevalence_2018.html [accessed 2024-09-13]
9. Lloyd-Jones DM, Allen NB, Anderson CA, Black T, Brewer LC, Foraker RE, American Heart Association. Life's essential 8: updating and enhancing the American heart association's construct of cardiovascular health: a presidential advisory from the American heart association. Circulation 2022;146(5):e18-e43 [FREE Full text] [doi: 10.1161/CIR.0000000000001078] [Medline: 35766027]
10. Neter JE, Stam BE, Kok FJ, Grobbee DE, Geleijnse JM. Influence of weight reduction on blood pressure: a meta-analysis of randomized controlled trials. Hypertension 2003;42(5):878-884. [doi: 10.1161/01.HYP.0000094221.86888.AE] [Medline: 12975389]
11. Barone Gibbs B, Hivert M, Jerome GJ, Kraus WE, Rosenkranz SK, Schorr EN, American Heart Association Council on Lifestyle and Cardiometabolic Health, Council on Cardiovascular and Stroke Nursing, Council on Clinical Cardiology. Physical activity as a critical component of first-line treatment for elevated blood pressure or cholesterol: who, what, and how?: a scientific statement from the American heart association. Hypertension 2021;78(2):e26-e37 [FREE Full text] [doi: 10.1161/HYP.0000000000000196] [Medline: 34074137]
12. Healthcare Mobile Applications Market Size Report, 2020-2027. 2020. URL: https://www.grandviewresearch.com/industry-analysis/healthcare-mobile-applications-market [accessed 2024-09-13]
13. Fitness App Market Size & Share Report, 2022-2030. URL: https://www.linkedin.com/pulse/fitness-app-market-size-research-report-2023-2030-jynof [accessed 2024-09-13]
14. Whitelaw S, Pellegrini DM, Mamas MA, Cowie M, Van Spall HGC. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
15. Aune D, Schlesinger S, Norat T, Riboli E. Body mass index, abdominal fatness, and the risk of sudden cardiac death: a systematic review and dose-response meta-analysis of prospective studies. Eur J Epidemiol 2018;33(8):711-722 [FREE Full text] [doi: 10.1007/s10654-017-0353-9] [Medline: 29417316]
16. Alonso-Domínguez R, Patino-Alonso MC, Sánchez-Aguadero N, García-Ortiz L, Recio-Rodríguez JI, Gómez-Marcos MA. Effect of a multifactorial intervention on the increase in physical activity in subjects with type 2 diabetes mellitus: a randomized clinical trial (EMID Study). Eur J Cardiovasc Nurs 2019;18(5):399-409. [doi: 10.1177/1474515119835048] [Medline: 30808196]
17. Carr LJ, Bartee RT, Dorozynski C, Broomfield JF, Smith ML, Smith DT. Internet-delivered behavior change program increases physical activity and improves cardiometabolic disease risk factors in sedentary adults: results of a randomized controlled trial. Prev Med 2008;46(5):431-438. [doi: 10.1016/j.ypmed.2007.12.005] [Medline: 18207228]
18. Allen JK, Stephens J, Dennison Himmelfarb CR, Stewart KJ, Hauck S. Randomized controlled pilot study testing use of smartphone technology for obesity treatment. J Obes 2013;2013:151597 [FREE Full text] [doi: 10.1155/2013/151597] [Medline: 24392223]
19. Bennett GG, Herring SJ, Puleo E, Stein EK, Emmons KM, Gillman MW. Web-based weight loss in primary care: a randomized controlled trial. Obesity (Silver Spring) 2010;18(2):308-313 [FREE Full text] [doi: 10.1038/oby.2009.242] [Medline: 19696764]
20. National library of medicine (US). NCBI. 1988. URL: https://www.ncbi.nlm.nih.gov/ [accessed 2024-09-13]
21. Scholten RJPM, Clarke M, Hetherington J. The Cochrane Collaboration. Eur J Clin Nutr 2005;59 Suppl 1:S147-S149; discussion S195. [doi: 10.1038/sj.ejcn.1602188] [Medline: 16052183]
22. CINAHL plus with full text. EBSCO Information Services, Inc. 1961. URL: https://www.ebsco.com/products/research-databases/cinahl-plus-full-text [accessed 2024-09-13]
23. Scopus. Elsevier. 2004. URL: https://en.wikipedia.org/wiki/Elsevier [accessed 2024-09-13]
24. SPORTDiscus. EBSCO Information Services, Inc. 1970. URL: https://econtents.bc.unicamp.br/pas/index.php/jad/article/view/302 [accessed 2024-09-13]

XSL•FO
RenderX

25. Spaulding EM, Marvel FA, Piasecki RJ, Martin SS, Allen JK. User engagement with smartphone apps and cardiovascular disease risk factor outcomes: systematic review. JMIR Cardio 2021;5(1):e18834 [FREE Full text] [doi: 10.2196/18834] [Medline: 33533730]
26. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
27. Johnson BT. Toward a more transparent, rigorous, and generative psychology. Psychol Bull 2021;147(1):1-15. [doi: 10.1037/bul0000317] [Medline: 33464105]
28. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
29. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
30. 2008 Physical activity guidelines for Americans. U.S. Department of Health and Human Services. 2009. URL: https://health.gov/sites/default/files/2019-09/paguide.pdf [accessed 2024-09-13]
31. Dietary guidelines for Americans. U.S. Department of Health and Human Services. 2016. URL: https://health.gov/our-work/nutrition-physical-activity/dietary-guidelines [accessed 2024-09-13]
32. Rayyan Intelligent Systematic Review. 2022. URL: https://www.rayyan.ai/ [accessed 2024-09-13]
33. Wu Y, Johnson BT, Acabchuk RL, Chen S, Lewis HK, Livingston J, et al. Yoga as antihypertensive lifestyle therapy: a systematic review and meta-analysis. Mayo Clin Proc 2019;94(3):432-446. [doi: 10.1016/j.mayocp.2018.09.023] [Medline: 30792067]
34. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
35. Glasgow RE, Boles SM, McKay H, Feil EG, Barrera M. The D-Net diabetes self-management program: long-term implementation, outcomes, and generalization results. Prev Med 2003;36(4):410-419. [doi: 10.1016/s0091-7435(02)00056-7] [Medline: 12649049]
36. Shannon HH, Joseph R, Puro N, Darrell E. Use of technology in the management of obesity: a literature review. Perspect Health Inf Manag 2019;16(Fall):1c [FREE Full text] [Medline: 31908626]
37. Howland C, Wakefield B. Assessing telehealth interventions for physical activity and sedentary behavior self-management in adults with type 2 diabetes mellitus: an integrative review. Res Nurs Health 2021;44(1):92-110 [FREE Full text] [doi: 10.1002/nur.22077] [Medline: 33091168]
38. Buckingham SA, Williams AJ, Morrissey K, Price L, Harrison J. Mobile health interventions to promote physical activity and reduce sedentary behaviour in the workplace: a systematic review. Digit Health 2019;5:2055207619839883 [FREE Full text] [doi: 10.1177/2055207619839883] [Medline: 30944728]
39. Daryabeygi-Khotbehsara R, Shariful Islam SM, Dunstan D, McVicar J, Abdelrazek M, Maddison R. Smartphone-based interventions to reduce sedentary behavior and promote physical activity using integrated dynamic models: systematic review. J Med Internet Res 2021;23(9):e26315 [FREE Full text] [doi: 10.2196/26315] [Medline: 34515637]
40. Connelly J, Kirk A, Masthoff J, MacRury S. The use of technology to promote physical activity in type 2 diabetes management: a systematic review. Diabet Med 2013;30(12):1420-1432. [doi: 10.1111/dme.12289] [Medline: 23870009]
41. Cotter AP, Durant N, Agne AA, Cherrington AL. Internet interventions to support lifestyle modification for diabetes management: a systematic review of the evidence. J Diabetes Complications 2014;28(2):243-251 [FREE Full text] [doi: 10.1016/j.jdiacomp.2013.07.003] [Medline: 24332469]
42. Veazie S, Winchell K, Gilbert J, Paynter R, Ivlev I, Eden KB, et al. Rapid evidence review of mobile applications for self-management of diabetes. J Gen Intern Med 2018;33(7):1167-1176 [FREE Full text] [doi: 10.1007/s11606-018-4410-1] [Medline: 29740786]
43. Coons MJ, Demott A, Buscemi J, Duncan JM, Pellegrini CA, Steglitz J, et al. Technology interventions to curb obesity: a systematic review of the current literature. Curr Cardiovasc Risk Rep 2012;6(2):120-134 [FREE Full text] [doi: 10.1007/s12170-012-0222-8] [Medline: 23082235]
44. Puigdomenech Puig E, Robles N, Saigí-Rubió F, Zamora A, Moharra M, Paluzie G, et al. Assessment of the efficacy, safety, and effectiveness of weight control and obesity management mobile health interventions: systematic review. JMIR Mhealth Uhealth 2019;7(10):e12612 [FREE Full text] [doi: 10.2196/12612] [Medline: 31654566]
45. Cotie LM, Prince SA, Elliott CG, Ziss MC, McDonnell LA, Mullen KA, et al. The effectiveness of eHealth interventions on physical activity and measures of obesity among working-age women: a systematic review and meta-analysis. Obes Rev 2018;19(10):1340-1358. [doi: 10.1111/obr.12700] [Medline: 30156044]
46. Sequi-Dominguez I, Alvarez-Bueno C, Martinez-Vizcaino V, Fernandez-Rodriguez R, Del Saz Lara A, Cavero-Redondo I. Effectiveness of mobile health interventions promoting physical activity and lifestyle interventions to reduce cardiovascular

XSL•FO
RenderX

risk among individuals with metabolic syndrome: systematic review and meta-analysis. J Med Internet Res 2020;22(8):e17790 [FREE Full text] [doi: 10.2196/17790] [Medline: 32865503]
47. Cavero-Redondo I, Martinez-Vizcaino V, Fernandez-Rodriguez R, Saz-Lara A, Pascual-Morena C, Álvarez-Bueno C. Effect of behavioral weight management interventions using lifestyle mhealth self-monitoring on weight loss: a systematic review and meta-analysis. Nutrients 2020;12(7):1977 [FREE Full text] [doi: 10.3390/nu12071977] [Medline: 32635174]
48. McMahon J, Thompson DR, Pascoe MC, Brazil K, Ski CF. eHealth interventions for reducing cardiovascular disease risk in men: a systematic review and meta-analysis. Prev Med 2021;145:106402. [doi: 10.1016/j.ypmed.2020.106402] [Medline: 33388336]
49. Kuo CC, Su Y, Lin C. A systematic review and meta-analysis: effectiveness of internet empowerment-based self-management interventions on adults with metabolic diseases. J Adv Nurs 2018;74(8):1787-1802. [doi: 10.1111/jan.13574] [Medline: 29574977]
50. Lyzwinski LN. A systematic review and meta-analysis of mobile devices and weight loss with an intervention content analysis. J Pers Med 2014;4(3):311-385 [FREE Full text] [doi: 10.3390/jpm4030311] [Medline: 25563356]
51. Sherrington A, Newham JJ, Bell R, Adamson A, McColl E, Araujo-Soares V. Systematic review and meta-analysis of internet-delivered interventions providing personalized feedback for weight loss in overweight and obese adults. Obes Rev 2016;17(6):541-551 [FREE Full text] [doi: 10.1111/obr.12396] [Medline: 26948257]
52. Kim HN, Seo K. Smartphone-based health program for improving physical activity and tackling obesity for young adults: a systematic review and meta-analysis. Int J Environ Res Public Health 2019;17(1):15 [FREE Full text] [doi: 10.3390/ijerph17010015] [Medline: 31861359]
53. Khunti K, Davies M, Majeed A, Thorsted BL, Wolden ML, Paul SK. Hypoglycemia and risk of cardiovascular disease and all-cause mortality in insulin-treated people with type 1 and type 2 diabetes: a cohort study. Diabetes Care 2015;38(2):316-322. [doi: 10.2337/dc14-0920] [Medline: 25492401]
54. Lear SA, Hu W, Rangarajan S, Gasevic D, Leong D, Iqbal R, et al. The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: the PURE study. Lancet 2017;390(10113):2643-2654 [FREE Full text] [doi: 10.1016/S0140-6736(17)31634-3] [Medline: 28943267]
55. Pradal-Cano L, Lozano-Ruiz C, Pereyra-Rodríguez JJ, Saigí-Rubió F, Bach-Faig A, Esquius L, et al. Using mobile applications to increase physical activity: a systematic review. Int J Environ Res Public Health 2020;17(21):8238 [FREE Full text] [doi: 10.3390/ijerph17218238] [Medline: 33171871]
56. Aucott LS, Poobalan A, Smith WCS, Avenell A, Jung R, Broom J. Effects of weight loss in overweight/obese individuals and long-term hypertension outcomes: a systematic review. Hypertension 2005;45(6):1035-1041. [doi: 10.1161/01.HYP.0000165680.59733.d4] [Medline: 15897373]
57. Zhou M. Waist circumference and risk of cardiovascular disease: a meta-analysis of prospective studies. Journal of Cardiovascular Disease Research 2019;17(4).
58. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R, Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet 2002;360(9349):1903-1913. [doi: 10.1016/s0140-6736(02)11911-8] [Medline: 12493255]
59. Suls J, Mogavero JN, Falzon L, Pescatello LS, Hennessy EA, Davidson KW. Health behaviour change in cardiovascular disease prevention and management: meta-review of behaviour change techniques to affect self-regulation. Health Psychol Rev 2020;14(1):43-65 [FREE Full text] [doi: 10.1080/17437199.2019.1691622] [Medline: 31707938]
60. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
61. Pescatello LS, Guidry MA, Blanchard BE, Kerr A, Taylor AL, Johnson AN, et al. Exercise intensity alters postexercise hypotension. J Hypertens 2004;22(10):1881-1888. [doi: 10.1097/00004872-200410000-00009] [Medline: 15361758]
62. Zmuda JM, Yurgalevitch SM, Flynn MM, Bausserman LL, Saratelli A, Spannaus-Martin DJ, et al. Exercise training has little effect on HDL levels and metabolism in men with initially low HDL cholesterol. Atherosclerosis 1998;137(1):215-221 [FREE Full text] [doi: 10.1016/s0021-9150(97)00257-8] [Medline: 9568755]

Abbreviations

AMSTAR: A Measurement Tool for Assessment of Multiple Systematic Reviews
BCT: behavioral change technique
BP: blood pressure
CVD: cardiovascular disease
DBP: diastolic blood pressure
DM: diabetes mellitus
HbA1c: hemoglobin A1c
HDL: high-density lipoprotein
LDL: low-density lipoprotein

XSL•FO
RenderX

PA: physical activity
PA+Diet: physical activity and diet
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SBP: systolic blood pressure

Edited by N Cahill; submitted 27.07.23; peer-reviewed by M Gasmi , B Bock; comments to author 27.02.24; revised version received 04.03.24; accepted 23.08.24; published 09.10.24.

Please cite as:
Bushey E, Wu Y, Wright A, Pescatello L
The Influence of Physical Activity and Diet Mobile Apps on Cardiovascular Disease Risk Factors: Meta-Review
J Med Internet Res 2024;26:e51321
URL: https://www.jmir.org/2024/1/e51321
doi:10.2196/51321
PMID:39382958

©Erica Bushey, Yin Wu, Alexander Wright, Linda Pescatello. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing a Digital Health Check With Traditional Nurse-Led Health Examinations Among Long-Term Unemployed Individuals: Comparison Study

Venla Raussi[1], BM; Sari Kujala[2], PhD; Iiris Hörhammer[2], DSc; Kaisa Savolainen[2], DA, MSc; Reija Autio[3], MSc, DSc; Tuomas Koskela[4,5], MD, PhD

[1]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[2]Department of Computer Science, Aalto University, Helsinki, Finland
[3]Faculty of Social Sciences (Health Sciences), Tampere University, Tampere, Finland
[4]Department of General Practice, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[5]The Wellbeing Services County of Pirkanmaa, Tampere, Finland

Corresponding Author:
Venla Raussi, BM
Faculty of Medicine and Health Technology
Tampere University
Arvo Ylpön katu 34
Tampere, 33520
Finland
Phone: 358 403564154
Email: venla.raussi@tuni.fi

Abstract

Background: A digital health check can be used to screen health behavior risks in the population, help health care professionals with standardized risk estimation for their patients, and motivate a patient to change unhealthy behaviors. Long-term unemployed individuals comprise a particular subgroup with an increased risk of lifestyle-related diseases.

Objective: This study aims to investigate the clinical utility of a general digital health examination, the STAR Duodecim Health Check and Coaching Program (STAR), which was developed in Finland, in the targeted screening of long-term unemployed individuals. For this purpose, we compared health challenges identified by a digital health check with those identified by a nurse during a face-to-face health check for unemployed individuals.

Methods: In this comparison study, 49 unemployed participants attending a health check were recruited from two Finnish primary health care centers. The participants used STAR and attended a nurse's health check. Data were collected by surveys with multiple-choice and open-ended questions from the participants, nurses, and a study assistant who observed the session. The nurses were asked to name the three most significant health challenges for each participant. These health challenges were categorized into health challenges corresponding to STAR and these were compared with each other. Percentages of agreement between STAR and nurses were calculated. Sensitivity and specificity, as well as Cohen κ with *P* values and CIs, were computed for agreement.

Results: STAR identified a total of 365 health challenges, an average of 7.4 (SD 2.5) health challenges per participant (n=49). The nurses named a total of 160 health challenges (n=47). In 53% (95% CI 38.1-67.9; n=25) of cases, STAR identified all categorized health challenges named by nurses. In 64% (95% CI 48.5-77.3; n=30) of cases, STAR identified at least 2/3 of the health challenges identified by nurses. Cohen κ was 0.877 ($P<.001$) for alcohol, indicating almost perfect agreement, and 0.440 ($P<.001$) for smoking and 0.457 ($P=.001$) for cholesterol, indicating moderate agreement. STAR left a total of 89 health challenges, an average of 1.8 (SD 1.1) per participant, uncategorized because STAR lacked an answer to the question or questions required for the classification of a certain health challenge. The participants did not always add information on their blood pressure (n=36, 74%), cholesterol (n=22, 45%), and waist circumference (n=15, 31%).

Conclusions: In conclusion, STAR identified most of the health challenges identified by nurses but missed some essential ones. Participants did not have information on measurements, such as blood pressure and cholesterol values, which are pivotal to STAR in assessing cardiovascular risks. Using the tool for screening or as a part of a traditional health check with necessary measurements

XSL•FO
RenderX

and dialog with health care professionals may improve the risk assessments and streamline the health checks of unemployed individuals.

International Registered Report Identifier (IRRID): RR2-10.2196/27668

(J Med Internet Res 2024;26:e49802) doi:10.2196/49802

KEYWORDS

chronic illnesses; eHealth; health care services; lifestyle; long-term unemployment; digital health check; primary prevention; risk assessment; risk factors; prevention; screening

Introduction

Background

Long-term illnesses and multimorbidity have become more common, thus reducing quality of life and increasing the demand for health care services [1,2]. Lifestyle choices have a significant impact on the expected onset of diseases, age of death, risk factors concerning long-term illnesses, and multimorbidity [1,3-5]. Preventable lifestyle-related risk factors affecting chronic morbidity and mortality have been recognized, most notably smoking, the harmful use of alcohol, physical inactivity, and an unhealthy diet [5,6].

So-called eHealth uses digital information and communication technologies for health, demonstrating the growing potential to make health services more accessible, efficient, and cost-effective [7,8]. Digital health checks, an eHealth tool aimed at assessing lifestyle-related risk factors, could improve primary prevention in health care [9-11]. A digital health check can be used to screen health behavior risks in the population, help health care professionals with standardized risk estimation for their patients, and motivate a patient to change unhealthy behaviors [9]. Web-based interventions focusing on health behavior-related risks have been reported to have an overall positive effect on the user's health, resulting in positive behavior changes [11-13]. Assessing multiple lifestyle-related risks at the same time provides an opportunity to review one's health comprehensively and target multiple health-related risk behaviors simultaneously [9]. Such interventions have been well-received by patients compared to interventions targeting only one health-related behavior [9,12-14].

The STAR Duodecim Health Check and Coaching Program (hereafter, STAR) is a general digital health examination developed by Duodecim Publishing Company Ltd and the Finnish Institute of Health and Welfare [15-17]. The abbreviation STAR comes from the Finnish word for a digital health check. STAR gives users a report that includes an evaluation of their life expectancy and an estimated risk for developing common long-term illnesses based on questions about personal characteristics, health information, lifestyle, mental well-being, and relationships. In addition, STAR provides coaching courses from which users can choose the most suitable based on STAR's recommendations. Thus, STAR provides tools for further improvement and long-term tracking of health. Users can set personal goals, participate in coaching, and follow up on the development of their health over time. STAR and its report are further described in the study protocol and Multimedia Appendix 1 [18]. Previous studies of STAR have mainly focused on creating a persuasive system design [18-21].

STAR's life expectancy evaluation and risk evaluations are based on previous Finnish studies, namely the Finriski, Autoklinikka, and Minisuomi studies [15,22-24]. The information provided by risk calculators can help health care professionals identify risk categories more accurately and improve the likelihood of prescribing medicine to high-risk patients, thereby helping with decision-making [25]. On the other hand, there is a huge variety of health risk calculators available digitally. A systematic review of digital cardiovascular disease risk calculators found wide variation in risk assessment models, risk presentation, and results [26]. This study also found the risk calculators to have overall poor actionability, and that the available risk calculators often lack clinical validity [26].

Long-term unemployed individuals comprise a particular subgroup with an increased risk of lifestyle-related diseases [27]. Long-term unemployment is linked to greater than average morbidity, earlier expected age of death, and increased risk of mortality [27,28]. Long-term unemployment is defined as having been unemployed for 12 months or more [29]. The duration of unemployment increases the burden of disease [30]. Unemployment also affects self-assessed health negatively, and the strongest association is found in people with a lower socioeconomic status, weak social networks, and health-related reasons for unemployment [31]. There is also some evidence that unemployed people use preventive services less in Finland [32]. There have been studies on digital health checks and internet-based risk assessments of subgroups, such as the employed, but there have been few studies focusing on digital health checks for unemployed individuals [33].

Although the general health checks for the unselected population may not be cost-effective in reducing illness and mortality [34], targeted screening could be useful if it leads to action (by unemployed individuals or by the service system) [32]. The lack of benefit of health checks may be due to the fact that those who would need them the most do not participate in them [34]. Furthermore, there is not enough evidence on the potential clinical utility of digital health checks in targeted health risk screening. While digital health checks performed by citizens themselves may hold potential for the low-cost screening of targeted populations, more evidence is needed regarding their ability to identify health risks in these subgroups.

The Goal of This Study

This study aims to investigate the clinical utility of a digital health check in screening targeted at long-term unemployed individuals. For this purpose, we compare the health challenges

XSL•FO
RenderX

identified by a digital health check with those identified by nurses during a face-to-face health check for unemployed individuals. We report the agreement and differences in health challenges identified by the digital health check and the nurse's check.

Methods

Recruitment

The inclusion criteria for participation in the study were unemployment for at least 12 months, age older than 18 years, and participation in a health check for long-term unemployed persons. Finland has a public health care system organized and financed by welfare counties, and every resident is entitled to receive social, health, and medical services [35,36]. According to the health care law, welfare counties are obliged to organize health checks for unemployed individuals [37]. The purpose of these health checks is to promote health and support the ability to function and work [38]. The initiative for a health check can come from the unemployed person, unemployment services, or the municipal social welfare administration.

We recruited 49 participants in total: 45 from Tampere and 4 from Espoo. The characteristics of the health check participants are described in Table 1. Three nurses from Tampere and three nurses from Espoo health centers participated in the study. Two participants were excluded from the health challenge analysis due to a lack of data from the nurses. In these cases, the participants skipped the nurse's health examination or the nurse did not fill out the professional questionnaire. As a result, data were obtained from 47 participants.

Table 1. The characteristics of the health check participants.

Characteristics	Participants (N=49)
Age (in years), mean (SD)	47.63 (10.38)
Sex, n (%)	
Male	28 (57)
Female	21 (43)
Unemployment time in months, mean (SD)[a]	43.6 (55.28)
What did you do before you became unemployed?, n (%)	
Studying	6 (12)
On sick leave	3 (6)
Working	32 (65)
Something else	8 (16)
Highest education, n (%)	
Elementary school	9 (18)
Vocational education	23 (47)
Upper secondary education	9 (18)
Bachelor's degree	5 (10)
Master's degree	3 (6)
Do you have a long-term illness diagnosed by a doctor?, n (%)[b]	
No	15 (31)
Yes	29 (59)
Don't know	6 (12)

[a]Information is missing from three (3).

[b]One (1) answered both yes and I don't know.

Procedure

The recruitment process started when the health center assistant (or in Espoo, a nurse) booked an appointment for an unemployed person to nurse's health check and told them about the possibility of participating in this study. Those who expressed their willingness to participate in the study were scheduled for a health check on the research day.

The participants attended a health check for unemployed individuals at the local health center, gave their consent on a consent form, and filled out participant questionnaire 1. Each participant filled out STAR and read its report while a study assistant observed and filled out the observer's questionnaire. After reading the report, the participant filled out participant questionnaire 2. Next, the participant was directed to the nurse's health check. After the nurse's health check, the nurse filled out the professional questionnaire 1 before reading the STAR report. In professional questionnaire 1, the nurse was asked to name the participant's three most significant health challenges. Then the nurse read the STAR report and filled out professional

XSL•FO
RenderX

questionnaire 2. A flowchart of the recruitment process and the content of the questionnaires is described in more detail in the study protocol [18].

The initial idea was that the order of STAR and the nurse's health check would be reversed after every 10 checkups [18] but implementing this proved difficult due to the nurses' schedules, hence only a few cases (n=3) were assessed in reverse order.

The recruitment took place at two Finnish public health centers in Espoo and Tampere. Espoo and Tampere are the second and third largest cities in Finland, with a population of 290,000 and 240,000, respectively [39].

After completing STAR, the participant received a report from STAR including a list of personal health challenges identified from the answers. The term "health challenge" means a medical condition, disease, habit, or lifestyle that poses the risk of disease or medical ailment. In the STAR report, health challenges are categorized into red, yellow, and green categories. STAR defines the categories as follows: red="Please check—this is essential for your health"; yellow="Please check and pay attention"; and green="Great, continue with the same pattern." The STAR classification is based on the patient's answers to the STAR questions. According to the limit values, STAR classifies a total of 17 different health challenges into these aforementioned color-coded categories. The 17 health challenges are BMI, waist circumference, exercise, diet, sleep, stress, mental resources, community action, alcohol, smoking, cholesterol, diabetes, blood pressure, oral health, ability to work, relationships, and family.

The total number of STAR health challenges per participant is not limited; STAR can categorize any number of the 17 different health challenges into red or yellow health challenges. In contrast, a nurse was asked to name only the three most significant health challenges in the questionnaire in this study.

If the user has not answered a question or questions about a specific health challenge, STAR cannot classify it into any of the previous color-coded categories. In this case, this health challenge appears as a gray category in the STAR report. STAR announces to the user that there is no classification information for the health challenge in question, and STAR cannot take a position on it in this case.

Analysis Methods

The three top health challenges named by the nurse in the health check were compared to the health challenges found in the STAR report for the same person. These three health challenges named by the nurse were first classified according to the corresponding 17 categories in the STAR report so that they could be compared. Health challenges named by the nurse that could not be categorized into the same categories as in STAR remained uncategorized. They are referred to hereafter as "uncategorized."

We calculated the cases in which STAR identified all the same health challenges as nurses did and the cases in which STAR identified at least 2/3 of the health challenges determined by the nurses. We also report the top four health challenges named by the nurse and STAR, respectively, the new health challenges STAR found, and any health challenges missed. We excluded work ability from the analyses because STAR does not classify work ability if the participant has answered that he or she is unemployed. In addition, we calculated for each health challenge the accuracy, sensitivity, specificity, and agreement with Cohen κ value with CIs and *P* value [40]. We performed a power analysis to determine the sample size necessary for detecting a moderate or higher agreement in our study. The analysis showed that 47 participants per condition would provide sufficient power (80%) to identify this agreement with a significance level of 0.05. We used R (version 4.0.1; R Core Team) [41] with the packages caret [42] and irr [43] for computing the sensitivity, specificity, accuracy, and κ values. In all statistical analyses, we considered *P* values <.05 to be statistically significant.

Ethical Considerations

This study was approved by the Ethics Committee of the Expert Responsibility Area of Tampere University Hospital in June 2020 (ETL Code R20067). All participants signed the written informed consent form to take part in the study. They were informed about the purpose of the research, the expected duration of their participation, that their participation was voluntary, and that they could discontinue it at any time without it causing them any harm. Participants' consent forms and all questionnaires collected in the study were archived. All questionnaires collected in the study were anonymized with numerical codes. Participants were not reimbursed and did not receive compensation for participating in the study.

Results

Overview

STAR predicted an average life expectancy of 78.1 (SD 6.8) years for the participants. STAR identified a total of 365 health challenges (n=49), an average of 7.4 (SD 2.5) health challenges per participant (Table 2). Nurses named a total of 160 top 3 significant health challenges (n=47; Table 2). Of these, 128 health challenges determined by the nurses were categorized into the same 17 categories as in the STAR report (Table 2 and Figure 1). The remaining 32 health challenges named by the nurses could not be categorized in the STAR categories (Table 2 and Figure 1). STAR left a total of 89 health challenges, an average of 1.8 (SD 1.1) per participant, uncategorized in the color-coded categories because STAR lacked an answer to the question or questions required for classification for a certain health challenge (Table 2).

In 53% (n=25) of cases, STAR identified all categorized health challenges named by nurses (95% CI 38.1-67.9). In 64% (n=30) of cases, STAR identified at least 2/3 of the health challenges identified by the nurse (95% CI 48.5-77.3).

Nurses most often named mental resources (n=27, 57%), blood pressure (n=14, 30 I made the requested changes.), cholesterol (n=13, 28%), and BMI (n=10, 21%) as health challenges. STAR most often named diet (n=45, 96%), community action (n=38, 81%), BMI (n=34, 72%), and mental resources (n=32, 68%) as health challenges. Community action and diet were seldomly named in the top three health challenges by the nurses. Nurses

XSL•FO
RenderX

did not name waist circumference, relationship, or family among the health challenges (Table 3).

Table 2. Characteristics of the health challenges named by STAR and the nurses.

	STAR (n=49)	Nurses (n=47)
Health challenges, number (mean, SD)	365 (7.4, 2.5)	160 (3.4, 0.9)
STAR red[a], number (mean, SD)	189 (3.9, 2.3)	N/A[b]
STAR yellow[c], number (mean, SD)	176 (3.6, 1.7)	N/A
Nurses' categorized[d], number (mean, SD)	N/A	128 (2.7, 1.1)
Nurses' uncategorized[d], number (mean, SD)	N/A	32 (0.7, 0.8)
STAR uncategorized[d], number (mean, SD)	89 (1.8, 1.1)	N/A

[a]Red: Please check—this is essential for your health.
[b]N/A: not applicable
[c]Yellow: Please check and pay attention.
[d]Able to categorize according to the corresponding 17 categories in the STAR report.

Figure 1. Flowchart of the different health challenges identified by the nurses and the digital health check (STAR) in the study.

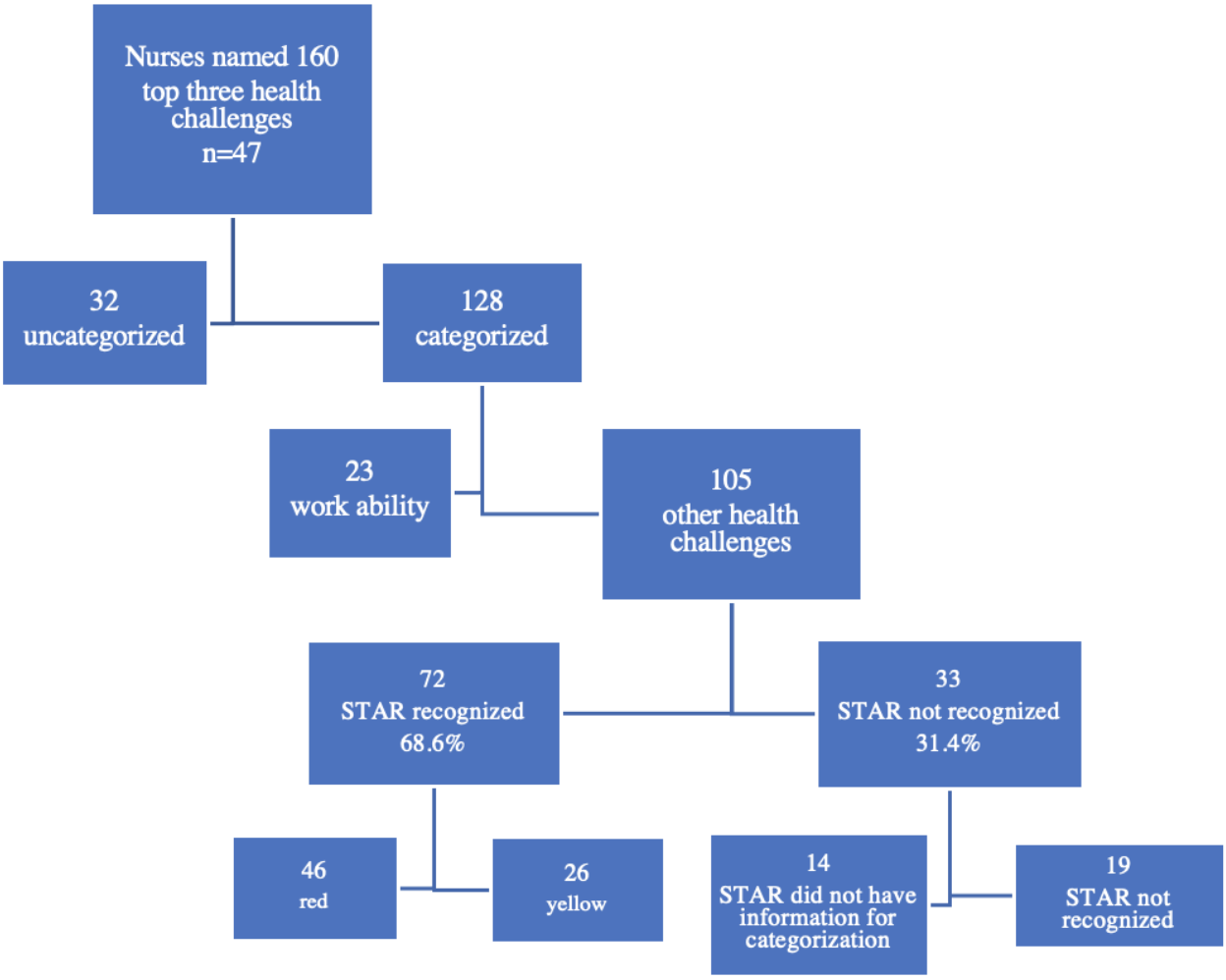

XSL•FO
RenderX

Table 3. The matches of the health challenges in the nurse's and STAR's health checks (N=47).

	Nurse named	STAR named	Accuracy[a]	Sensitivity	Specificity	Cohen κ (95% CI)	*P* value (κ)
BMI	10	34	0.404	0.8	0.297	0.052 (–0.105 to 0.208)	.54
Waist circumference	0	23	0.50	N/A[b]	N/A	N/A	N/A
Exercise	6	20	0.617	0.667	0.610	0.138 (–0.082 to 0.359)	.20
Diet	7	45	0.191	1	0.050	0.015 (–0.009 to 0.039)	.55
Sleep	5	17	0.660	0.600	0.667	0.130 (–0.106 to 0.366)	.24
Stress	3	28	0.468	1	0.432	0.088 (–0.012 to 0.189)	.14
Mental resources	27	32	0.638	0.778	0.450	0.235 (–0.041 to 0.512)	.09
Community action	2	38	0.191	0.5	0.178	–0.034 (–0.12 to 0.053)	.26
Alcohol	4	5	0.979	1	0.977	0.877 (0.641 to 1)	<.001
Smoking	7	18	0.766	1	0.725	0.440 (0.2 to 0.68)	<.001
Cholesterol	13	8	0.809	0.462	0.941	0.457 (0.166 to 0.748)	.001
Diabetes	6	17	0.681	0.667	0.683	0.196 (–0.056 to 0.448)	.09
Blood pressure	14	9	0.681	0.286	0.848	0.150 (–0.144 to 0.443)	.29
Oral health	1	17	0.617	0	0.630	–0.042 (–0.122 to 0.038)	.45
Relationship	0	14	0.702	N/A	N/A	N/A	N/A
Family	0	9	0.801	N/A	N/A	N/A	N/A
Total	105	334	0.608	0.686	0.595	0.1467 (0.09189 to 0.2016)	<.001

[a]Cases the nurse and STAR agreed (=both named or neither named a health challenge).

[b]Cannot be computed since nurses did not find any of the participants to have these conditions.

Of the 105 categorized top three health challenges named by the nurses, STAR recognized 72 (69%) challenges; 46 (64%) challenges as red and 26 (36%) challenges as yellow (Figure 1). STAR missed 33 (31%) of the 105 challenges categorized as the top three health challenges by the nurses; STAR did not recognize 19 health challenges named by the nurses, and it was not even possible for STAR to recognize 14 of the health challenges named by the nurses because STAR lacked an answer to the question or questions required for categorization (Figure 1). The 14 health challenges named by nurses that STAR could not recognize were mental resources (n=1), cholesterol (n=4), and blood pressure (n=9).

STAR and the nurses recognized alcohol (n=46, 98%), cholesterol (n=38, 81%), and smoking (n=36, 77%) as health challenges very similarly (Table 3). Participants were not able to respond to all questions due to a lack of knowledge of some measurement values, and this had an impact on STAR's ability to recognize blood pressure and cholesterol, which affected the agreement between the nurses and STAR. Cohen κ was 0.877 (*P*<.001) for alcohol indicating almost perfect agreement and 0.440 (*P*<.001) for smoking and 0.457 (*P*=.001) for cholesterol indicating moderate agreement (Table 3).

STAR named a total of 262 new health challenges (127 red and 135 yellow) that were not named by the nurses. STAR most commonly named diet (15 times) and community action (35 times) as red health challenges, while the nurses did not name them among the three most significant health challenges.

STAR left a total of 89 health challenges, an average of 1.8 (SD 1.1) per participant, unassigned according to the color-coded categories because STAR lacked an answer to the question or questions required for classification to a certain health challenge (Table 2). This means that STAR does not analyze and categorize a health challenge into any color-coded category in the STAR report if it lacks the information necessary for classification, such as information about waist circumference or cholesterol values, as mentioned above. The participants did not most often add information about their blood pressure (n=36, 74%), cholesterol (n=22, 45%), and waist circumference (n=15, 31%). Additionally, 22.4% (n=22) of the participants had not filled in information about diabetes in STAR, while diet, stress, mental resources, relationships, and family were missing each from one participant (n=1, 2%).

Health Challenges Identified by the Nurse That Could Not Be Categorized

The 32 health challenges named by the nurse that could not be classified into the 17 categories of the STAR report were not identified by STAR. Musculoskeletal diseases were the most common of them. The nurses have named musculoskeletal diseases 9 times among the health challenges, but STAR was unable to identify them. In addition, STAR does not classify work ability into 3 different color-coded categories if the participant has responded that he or she is not employed. However, the nurses mentioned work ability as a health challenge in 49% (n=23) of the participants, which caused a discrepancy (Figure 1). There were also some specific diseases

XSL•FO
RenderX

and symptoms that STAR could not recognize, for example, migraine, chronic obstructive pulmonary disease, fibromyalgia, and impaired hearing. Nurses also mentioned various psychosocial problems among the health challenges, for example, autism and fear of social situations, which STAR was unable to identify. There were also some laboratory values, problems in taking care of the participant's own health, and diseases that run in the family, which the nurses had mentioned, but STAR was unable to identify.

Discussion

Principal Results

STAR recognized the health challenges named by the nurse well if the particular health challenge was covered in the questions of STAR. In half of the cases, STAR identified all the health challenges identified by the nurses. Alcohol consumption, smoking issues, and elevated cholesterol levels in particular were recognized similarly. On the other hand, the existence of certain health challenges was often left unassessed by STAR due to a lack of input from the patients. The participants most often lacked information about their blood pressure, cholesterol, and waist circumference. Furthermore, musculoskeletal diseases were not covered in STAR questions and were not recognized as health challenges in the STAR report, but they were listed among the top three health challenges by the nurses. On the other hand, STAR sensitively named diet and community action among the health challenges.

The life expectancies of unemployed individuals in this study group (average 78.1, SD 6.8 years) did not differ substantially from the national average life expectancy at birth [44]. However, the life expectancies of the participants in the STAR reports were probably overestimated, as many participants lacked information on essential risk factors to be taken into account in estimations of their life expectancy. Above all, the majority lacked information on blood pressure. High blood pressure is known to be a significant factor in reduced life expectancy [45].

Comparison With Prior Work

Evidence on the impacts of digital health checks is limited. Especially, there is no previous research on comparisons between digital health checks and nurse-led health examinations. However, in a feasibility study in England, it was found that digital health checks for reducing alcohol intake among employees appeared feasible and acceptable, but the study included low participation rates, potentially attracting "worried well" employees rather than those at greatest health risk [46]. Our study complements the perception that hazardous alcohol use can be identified well with digital health checks and our study increases the knowledge of the uniformity of the assessments of digital health checks and nurse examinations regarding the identification of alcohol use.

Furthermore, a scoping review of the usability and utility of eHealth for physical activity counseling in primary health care centers found technical problems and the complexity of programs to be notable usability barriers to eHealth [47]. In this study, the unemployed participants' ignorance of their health-related parameters, such as blood pressure, cholesterol, and waist circumference, among other things, had a significant impact on STAR's recommendations. In an Australian study, it was found that most web-based heart age calculator users did not know their cholesterol values but knew their blood pressure values [48]. However, in this study, differently, almost half of the participants did not know their cholesterol values and 3 of 4 participants did not know their blood pressure values.

Strengths and Limitations

The strength of the study was the prospective design in a real-life setting focusing on potential high-risk people that could potentially benefit from the health check, and there is little previous research in regard to targeted screening among this particular group. In addition, as a reference for the digital health checks, we had the nurses' assessments of the top three health challenges for the same participants.

The limitation of the study was the small number of participants, which may have had an impact on the results. It was slow to gather study material due to the tight schedules in the participating health centers. There was a shortage of nurses during the COVID-19 pandemic, which slowed down the data collection. Furthermore, in many cases motivating unemployed individuals to participate in this study was difficult, possibly due to their individual and complex situations and their low interest in digital applications. In addition, many of the participants did not show up for the health examination. However, the sample size was sufficient for statistics concerning moderate or higher levels of agreement between the two measures.

The nurses were requested to name only the top three health challenges; hence an exact comparison of the health challenges between the nurses and STAR was not possible. Furthermore, deviating from the instructions, some nurses named more than three health challenges per participant but this did not have an impact on agreement analyses. In addition, the use of a digital tool as an intervention may have had an impact on the participants' thoughts about their health before the nurse's visit. Furthermore, there is a risk of selection bias because people who are comfortable with digital tools will probably more readily agree to participate compared with those who are not familiar with such technology. People who struggle with computers and technology may have refused to participate, even though they could provide important information concerning usability and the user experience. It has been reported that the users of eHealth interventions are more likely to be highly educated and have a healthier lifestyle than average, while those who could benefit the most are not using them [49,50].

The health challenges named by the nurses may have differed from those in the STAR report due to differences in terminology. In addition, the interpretation of the open-answered terms was challenging. It proved difficult to classify the health challenges named by the nurses and STAR as being exactly correspondent.

Implications

STAR recognized quite well the health challenges named by the nurses but missed some essential ones. This reflects the focus of STAR on cardiovascular risk factors, which seem to cover the health and well-being concerns of unemployed

XSL•FO
RenderX

individuals only partially. In Finland, for instance, the greatest health challenges related to the subjective inability to work are musculoskeletal and mental disorders [51]. The nurses often named musculoskeletal complaints as health challenges, but STAR was unable to identify them. In the future, STAR could be developed to identify health challenges related to musculoskeletal health as well due to their commonness and impact on health and disability [52]. Furthermore, participants did not remember or did not have information on crucial cardiovascular risk factors (eg, blood pressure, cholesterol values, and waist circumference), which had an impact on risk assessments. It is obvious that STAR would better identify health challenges if the participants had answered all the questions in STAR and known all the values needed in the answers. In this case, the report would have been also more plausible and reliable for the participant. Before filling out the STAR health check, the important values should be reminded to the user. In addition, the focus of the nurse's health check differs from the digital health check to some extent. STAR is focused on assessing cardiovascular risk and mental well-being; hence, a more comprehensive health-related assessment may have been missed in the digital health check. STAR did not identify specific diseases or symptoms, such as migraine, chronic obstructive pulmonary disease, and impaired hearing, because STAR is instead specially designed to identify risk factors associated with lifestyle, substance use, cardiovascular health, and mental well-being and the impact of various parameters (eg, blood pressure and cholesterol) on health and risk of illness.

STAR identified significantly more health challenges (an average of 7.4, SD 2.5, health challenges per participant) than the nurse was asked to name. The large number of health challenges named by STAR may be unmotivating from the participant's point of view. Although STAR prioritizes health challenges by classifying them into red and yellow, STAR could provide a more precise prioritization to enable the user to know which of the health challenges would be the most important to fix. This could also help nurses to assess which health challenge is the most significant for the patient.

According to our results, STAR may be used as a screening tool before the nurse's face-to-face health checks with unemployed individuals. This study does not give answers to the question of who needs a face-to-face health check in this group. However, it seems evident that a nurse's health check covers various aspects, especially in complex cases. Face-to-face encounters and interactions may facilitate the discussion of patient-centered health issues, which may be missed in digital screening. However, a digital health check could help to identify in a structured manner the most relevant health issues and identify the patients whose needs are the greatest. This is in line with Nordic general practice core values and principles [53]. Our results reveal that an effort should be put into providing support for self-measuring health indicators essential for digital health checks.

Conclusions

In conclusion, STAR identified most of the health challenges identified by the nurses but missed some essential ones. STAR is focused on assessing the risks of cardiovascular diseases and mental well-being, and some other aspects of health were missed. Users did not have information on crucial risk factors (eg, blood pressure and cholesterol values), which are pivotal in assessing cardiovascular risks. Using the tool for screening or as a part of a traditional health check with necessary measurements and dialog with health care professionals may improve the risk assessments and streamline health checks of unemployed individuals.

Acknowledgments

We would like to thank Antti Virta from Duodecim, who assisted with STAR and helped to determine the study setting and the study objectives. We are also grateful to Osmo Saarelma, who helped to get started with this idea, student Henna Kuhlberg, who arranged the protocol, and finally, student Onni Koskela, who helped with data collection. We also would like to thank the nurses who performed health checks at the health checkup units of unemployed individuals in Tampere and Espoo health centers. We also thank Samuli Känsälä, who helped with data collection in Espoo. This study was supported by the Strategic Research Council at the Academy of Finland (grants 352501 and 352503), and by the Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital by (projects 9AA031 and 9AC037).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to limitations of ethical approval and privacy policy, but they are available from the corresponding author upon reasonable request.

Authors' Contributions

TK, IH, and SK have designed this study. VR performed the data collection in Tampere, VR and RA performed the statistical analyses, and VR has been the major contributor to the manuscript. Samuli Känsälä and KS performed the data collection in Espoo. VR, TK, SK, IH, and KS have contributed to the interpretation of the data, and supervised and critically commented on the manuscript. All authors have read and approved the manuscript.

Conflicts of Interest

TK worked as a salaried part-time editor for Duodecim Medical publications until 2018.

XSL•FO
RenderX

Multimedia Appendix 1
The questionnaires used in the study.
[DOCX File , 16 KB - jmir_v26i1e49802_app1.docx]

References

1. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 2012;380(9836):37-43 [FREE Full text] [doi: 10.1016/S0140-6736(12)60240-2] [Medline: 22579043]
2. Glynn LG, Valderas JM, Healy P, Burke E, Newell J, Gillespie P, et al. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. Fam Pract 2011;28(5):516-523. [doi: 10.1093/fampra/cmr013] [Medline: 21436204]
3. Härkänen T, Kuulasmaa K, Sares-Jäske L, Jousilahti P, Peltonen M, Borodulin K, et al. Estimating expected life-years and risk factor associations with mortality in Finland: cohort study. BMJ Open 2020;10(3):e033741 [FREE Full text] [doi: 10.1136/bmjopen-2019-033741] [Medline: 32152164]
4. Wikström K, Lindström J, Harald K, Peltonen M, Laatikainen T. Clinical and lifestyle-related risk factors for incident multimorbidity: 10-year follow-up of Finnish population-based cohorts 1982-2012. Eur J Intern Med 2015;26(3):211-216. [doi: 10.1016/j.ejim.2015.02.012] [Medline: 25747490]
5. Global status report on noncommunicable diseases. WHO. 2010. URL: https://apps.who.int/iris/bitstream/handle/10665/44579/9789240686458_eng.pdf?sequence=1 [accessed 2021-05-22]
6. Global health risks: mortality and burden of disease attributable to selected major risks. WHO. Geneva; 2009. URL: https://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf [accessed 2021-05-20]
7. eHealth at WHO. WHO. URL: https://www.who.int/ehealth/about/en/ [accessed 2021-05-20]
8. Eland-de Kok P, van Os-Medendorp H, Vergouwe-Meijer A, Bruijnzeel-Koomen C, Ros W. A systematic review of the effects of e-health on chronically ill patients. J Clin Nurs 2011;20(21-22):2997-3010. [doi: 10.1111/j.1365-2702.2011.03743.x] [Medline: 21707807]
9. Carey M, Noble N, Mansfield E, Waller A, Henskens F, Sanson-Fisher R. The role of eHealth in optimizing preventive care in the primary care setting. J Med Internet Res 2015;17(5):e126 [FREE Full text] [doi: 10.2196/jmir.3817] [Medline: 26001983]
10. Barbabella F, Melchiorre M, Quattrini S, Papa R, Lamura G, On behalf of the ICARE4EU consortium. How can eHealth improve care for people with multimorbidity in Europe? Health Systems and Policy Analysis. URL: http://www.icare4eu.org/pdf/PB_25.pdf [accessed 2021-05-22]
11. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010;51(3-4):214-221 [FREE Full text] [doi: 10.1016/j.ypmed.2010.06.004] [Medline: 20558196]
12. Schulz DN, Kremers SPJ, Vandelanotte C, van Adrichem MJG, Schneider F, Candel MJJM, et al. Effects of a web-based tailored multiple-lifestyle intervention for adults: a two-year randomized controlled trial comparing sequential and simultaneous delivery modes. J Med Internet Res 2014;16(1):e26 [FREE Full text] [doi: 10.2196/jmir.3094] [Medline: 24472854]
13. Kypri K, McAnally HM. Randomized controlled trial of a web-based primary care intervention for multiple health risk behaviors. Prev Med 2005;41(3-4):761-766. [doi: 10.1016/j.ypmed.2005.07.010] [Medline: 16120456]
14. Vandelanotte C, Reeves MM, Brug J, De Bourdeaudhuij I. A randomized trial of sequential and simultaneous multiple behavior change interventions for physical activity and fat intake. Prev Med 2008;46(3):232-237. [doi: 10.1016/j.ypmed.2007.07.008] [Medline: 17707079]
15. Tutustu STAR-palveluun tarkemmin. Duodecim STAR. URL: https://star3.duodecim.fi/fi-FI/info/introduction [accessed 2021-05-22]
16. Duodecim STAR. URL: https://star3.duodecim.fi/fi-FI/me [accessed 2021-05-22]
17. For the general public. Duodecim Products. URL: https://www.duodecim.fi/english/products/for-everyone-else/ [accessed 2021-05-22]
18. Kuhlberg H, Kujala S, Hörhammer I, Koskela T. STAR Duodecim eHealth tool to recognize chronic disease risk factors and change unhealthy lifestyle choices among the long-term unemployed: protocol for a mixed methods validation study. JMIR Res Protoc 2021;10(6):e27668 [FREE Full text] [doi: 10.2196/27668] [Medline: 34061041]
19. Lehto T, Kukkonen H, Pätiälä T, Saarelma O. Virtual health coaching for consumers: a persuasive systems design perspective. Int J Networking Virtual Organ 2013;13(1):24-41. [doi: 10.1504/IJNVO.2013.058440]
20. Lehto T, Oinas-Kukkonen H, Pätiälä T, Saarelma O. Virtual health check and coaching: insights from the consumers and implications for persuasive design. In: Communications in Computer and Information Science. 2012 Presented at: 4th International Conference on Well-Being in the Information Society; August 22-25, 2012; Turku, Finland. [doi: 10.1007/978-3-642-32850-3_3]
21. Torniainen-Holm M, Pankakoski M, Lehto T, Saarelma O, Mustonen P, Joutsenniemi K, et al. The effectiveness of email-based exercises in promoting psychological wellbeing and healthy lifestyle: a two-year follow-up study. BMC Psychol 2016;4(1):21 [FREE Full text] [doi: 10.1186/s40359-016-0125-4] [Medline: 27184251]

XSL•FO
RenderX

22. The national FINRISK study. Finnish Institute for Health and Welfare. URL: https://thl.fi/en/research-and-development/research-and-projects/the-national-finrisk-study [accessed 2021-05-22]
23. Finnish mobile clinic health survey. Finnish Institute for Health and Welfare. URL: https://thl.fi/en/research-and-development/research-and-projects/finnish-mobile-clinic [accessed 2021-05-22]
24. Mini-Finland health survey. Finnish Institute for Health and Welfare. URL: https://thl.fi/en/research-and-development/research-and-projects/finnish-mobile-clinic/mini-finland-health-survey [accessed 2021-05-22]
25. Bonner C, Fajardo MA, Doust J, McCaffery K, Trevena L. Implementing cardiovascular disease prevention guidelines to translate evidence-based medicine and shared decision making into general practice: theory-based intervention development, qualitative piloting and quantitative feasibility. Implement Sci 2019;14(1):86 [FREE Full text] [doi: 10.1186/s13012-019-0927-x] [Medline: 31466526]
26. Bonner C, Fajardo MA, Hui S, Stubbs R, Trevena L. Clinical validity, understandability, and actionability of online cardiovascular disease risk calculators: systematic review. J Med Internet Res 2018;20(2):e29 [FREE Full text] [doi: 10.2196/jmir.8538] [Medline: 29391344]
27. Roelfs DJ, Shor E, Davidson KW, Schwartz JE. Losing life and livelihood: a systematic review and meta-analysis of unemployment and all-cause mortality. Soc Sci Med 2011;72(6):840-854 [FREE Full text] [doi: 10.1016/j.socscimed.2011.01.005] [Medline: 21330027]
28. Heponiemi T, Wahlström M, Elovainio M, Sinervo T, Anna-Mari A, Keskimäki I. Katsaus työttömyyden ja terveyden välisiin yhteyksiin (A review about the connection of unemployment and health by Ministry of Employment and the Economy of Finland, only available in Finnish). In: Tampereen yliopisto. Helsinki: Työ- ja elinkeinoministeriön julkaisuja, Työ ja yrittäjyys; 2008.
29. Long-term unemployment rate. OECD. URL: https://data.oecd.org/unemp/long-term-unemployment-rate.htm [accessed 2021-05-22]
30. Herbig B, Dragano N, Angerer P. Health in the long-term unemployed. Dtsch Arztebl Int 2013;110(23-24):413-419 [FREE Full text] [doi: 10.3238/arztebl.2013.0413] [Medline: 23837086]
31. Norström F, Virtanen P, Hammarström A, Gustafsson PE, Janlert U. How does unemployment affect self-assessed health? A systematic review focusing on subgroup effects. BMC Public Health 2014;14:1310 [FREE Full text] [doi: 10.1186/1471-2458-14-1310] [Medline: 25535401]
32. Klavus J, Forma L, Partanen J. Need and utilization of primary health care among long-term unemployed Finns. J Health Soc Sci 2018;3(3):231-242. [doi: 10.19204/2018/ndnd1]
33. Howarth A, Quesada J, Silva J, Judycki S, Mills PR. The impact of digital health interventions on health-related outcomes in the workplace: a systematic review. Digit Health 2018;4:2055207618770861 [FREE Full text] [doi: 10.1177/2055207618770861] [Medline: 29942631]
34. Krogsbøll LT, Jørgensen KJ, Larsen CG, Gøtzsche PC. General health checks in adults for reducing morbidity and mortality from disease: Cochrane systematic review and meta-analysis. BMJ 2012;345:e7191 [FREE Full text] [doi: 10.1136/bmj.e7191] [Medline: 23169868]
35. Healthcare system in Finland. EU-healthcare.fi. URL: https://www.eu-healthcare.fi/healthcare-in-finland/healthcare-system-in-finland/ [accessed 2021-05-22]
36. Finnish health care act, section 13, health counselling and health checks. Finlex. URL: https://www.finlex.fi/en/laki/kaannokset/2010/en20101326_20131293.pdf [accessed 2021-05-22]
37. Terveydenhuoltolaki, 13§. Finlex. URL: https://finlex.fi/fi/laki/ajantasa/2010/20101326 [accessed 2023-03-30]
38. Tampereen kaupunki. Terveystarkastukset. URL: https://www.tampere.fi/sosiaali-ja-terveyspalvelut/terveyspalvelut/omahoito-ja-terveysneuvonta/terveystarkastukset.html [accessed 2021-05-22]
39. Population in the largest municipalities in Finland. Statistics Finland. URL: https://www.stat.fi/tup/suoluk/suoluk_vaesto_en.html [accessed 2021-05-22]
40. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 2017;5:307 [FREE Full text] [doi: 10.3389/fpubh.2017.00307] [Medline: 29209603]
41. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.r-project.org/ [accessed 2024-09-10]
42. Kuhn M. Building predictive models in R using the caret package. J Stat Software 2008;28(5):1-26 [FREE Full text] [doi: 10.18637/jss.v028.i05]
43. Gamer M, Lemon J, Singh IFM. irr: various coefficients of interrater reliability and agreement. R package version 0.84.1. 2019. URL: https://CRAN.R-project.org/package=irr [accessed 2024-09-10]
44. Global health estimates: life expectancy and healthy life expectancy. World Health Organization. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy [accessed 2023-03-30]
45. Franco OH, Peeters A, Bonneux L, de Laet C. Blood pressure in adulthood and life expectancy with cardiovascular disease in men and women: life course analysis. Hypertension 2005;46(2):280-286. [doi: 10.1161/01.HYP.0000173433.67426.9b] [Medline: 15983235]

XSL•FO
RenderX

46. Khadjesari Z, Newbury-Birch D, Murray E, Shenker D, Marston L, Kaner E. Online health check for reducing alcohol intake among employees: a feasibility study in six workplaces across England. PLoS One 2015;10(3):e0121174 [FREE Full text] [doi: 10.1371/journal.pone.0121174] [Medline: 25798596]
47. Wattanapisit A, Tuangratananon T, Wattanapisit S. Usability and utility of eHealth for physical activity counselling in primary health care: a scoping review. BMC Fam Pract 2020;21(1):229 [FREE Full text] [doi: 10.1186/s12875-020-01304-9] [Medline: 33158430]
48. Bonner C, Raffoul N, Battaglia T, Mitchell JA, Batcup C, Stavreski B. Experiences of a national web-based heart age calculator for cardiovascular disease prevention: user characteristics, heart age results, and behavior change survey. J Med Internet Res 2020;22(8):e19028 [FREE Full text] [doi: 10.2196/19028] [Medline: 32763875]
49. Reiners F, Sturm J, Bouw LJ, Wouters EJ. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019;16(4):645 [FREE Full text] [doi: 10.3390/ijerph16040645] [Medline: 30795623]
50. Brouwer W, Oenema A, Raat H, Crutzen R, de Nooijer J, de Vries NK, et al. Characteristics of visitors and revisitors to an internet-delivered computer-tailored lifestyle intervention implemented for use by the general public. Health Educ Res 2010;25(4):585-595. [doi: 10.1093/her/cyp063] [Medline: 19897515]
51. Pihlajamäki M, Uitti J, Arola H, Korhonen M, Nummi T, Taimela S. Self-reported health problems in a health risk appraisal predict permanent work disability: a prospective cohort study of 22,023 employees from different sectors in Finland with up to 6-year follow-up. Int Arch Occup Environ Health 2020;93(4):445-456 [FREE Full text] [doi: 10.1007/s00420-019-01497-2] [Medline: 31786668]
52. Fact sheets/musculoskeletal health: life expectancy and healthy life expectancy. WHO. URL: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions [accessed 2023-12-20]
53. Nordic Federation of General Practice (NFGP). Core values and principles of nordic general practice/family medicine. Scand J Prim Health Care 2020;38(4):367-368 [FREE Full text] [doi: 10.1080/02813432.2020.1842674] [Medline: 33284030]

Abbreviations

STAR: STAR Duodecim Health Check and Coaching Program

Edited by T de Azevedo Cardoso; submitted 09.06.23; peer-reviewed by R Davey, L Guo; comments to author 02.11.23; revised version received 22.12.23; accepted 26.07.24; published 16.10.24.

Please cite as:
Raussi V, Kujala S, Hörhammer I, Savolainen K, Autio R, Koskela T
Comparing a Digital Health Check With Traditional Nurse-Led Health Examinations Among Long-Term Unemployed Individuals: Comparison Study
J Med Internet Res 2024;26:e49802
URL: https://www.jmir.org/2024/1/e49802
doi:10.2196/49802
PMID:39412874

©Venla Raussi, Sari Kujala, Iiris Hörhammer, Kaisa Savolainen, Reija Autio, Tuomas Koskela. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Outcomes of a Remotely Delivered Complementary and Integrative Health Partnered Intervention to Improve Chronic Pain and Posttraumatic Stress Disorder Symptoms: Randomized Controlled Trial

Jolie N Haun[1,2], EdS, PhD; Christopher A Fowler[1,3], PhD; Hari H Venkatachalam[1], MPH, CPH; Amy C Alman[4], MSPH, PhD; Lisa M Ballistrea[1], MSPT, DPT; Tali Schneider[1], PhD; Rachel C Benzinger[1], BA; Christine Melillo[1], PhD; Neil B Alexander[5,6], MD; S Angel Klanchar[1], RN; William A Lapcevic[1], MPH, MS; Matthew J Bair[7,8,9], MS, MD; Stephanie L Taylor[10,11,12], MPH, PhD; Jennifer L Murphy[13], PhD; Dustin D French[14,15,16], PhD

[1]Research and Development Service, James A Haley Veterans' Hospital, Tampa, FL, United States

[2]Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States

[3]Department of Psychiatry and Behavioral Neurosciences, University of South Florida, Tampa, FL, United States

[4]College of Public Health, University of South Florida, Tampa, FL, United States

[5]Geriatric Research Education and Clinical Center, VA Ann Arbor Healthcare System, Ann Arbor, MI, United States

[6]Division of Geriatric and Palliative Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States

[7]Health Services Research and Development Center for Health Information and Communication, Roudebush VA Medical Center, Department of Veterans Affairs, Indianapolis, IN, United States

[8]Department of Medicine, School of Medicine, Indiana University, Indianapolis, IN, United States

[9]Regenstrief Institute Inc, Indianapolis, IN, United States

[10]Health Services Research and Development Center for the Study of Healthcare Innovation, Implementation, and Policy, VA Greater Los Angeles Healthcare System, Department of Veterans Affairs, Los Angeles, CA, United States

[11]Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, United States

[12]Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, CA, United States

[13]National Pain Management, Opioid Safety, and Prescription Drug Monitoring Program, Specialty Care Program Office, Veterans Health Administration, Washington, DC, DC, United States

[14]Health Services Research and Development Center of Innovation for Complex Chronic Healthcare, Edward Hines Jr VA Hospital, Department of Veterans Affairs, Hines, IL, United States

[15]Center for Health Services and Outcomes Research, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[16]Departments of Ophthalmology and Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

Corresponding Author:
Rachel C Benzinger, BA
Research and Development Service
James A Haley Veterans' Hospital
8900 Grand Oak Circle
Tampa, FL, 33637
United States
Phone: 1 813 558 7622
Email: rachel.benzinger@VA.Gov

Abstract

Background: Nonpharmacological interventions for veterans are needed to help them manage chronic pain and posttraumatic stress disorder (PTSD) symptoms. Complementary and integrative health (CIH) interventions such as Mission Reconnect (MR) seek to provide veterans with the option of a partnered, self-directed intervention that teaches CIH skills remotely to support symptom management.

Objective: The purpose of this study was to describe the physical, psychological, and social outcomes of a self-directed mobile- and web-based CIH intervention for veterans with comorbid chronic pain and PTSD and their partners and qualitatively examine their MR user experience.

XSL•FO
RenderX

Methods: A sample of veteran-partner dyads (n=364) were recruited to participate in a mixed methods multisite waitlist control randomized controlled trial to measure physical, psychological, and social outcomes, with pain as the primary outcome and PTSD, depression, stress, sleep, quality of life, and relationships as secondary outcomes. Linear mixed models were constructed for primary and secondary patient-reported outcomes. The quantitative analysis was triangulated using qualitative interviews from a subsample of dyads (n=35) to examine participants' perceptions of their program experience.

Results: Dyads were randomized to 2 groups: intervention (MR; 140/364, 38.5%) and waitlist control (136/364, 37.4%). No significant change was observed in overall pain, sleep, PTSD, quality of life, relationship satisfaction, overall self-compassion, or compassion for others. A significant reduction in pain interference in mood (P=.008) and sleep (P=.008) was observed among the veteran MR group that was not observed in the waitlist control group. We also observed a positive effect of the MR intervention on a reduction in negative affect associated with pain (P=.049), but this effect did not exceed the adjusted significance threshold (P=.01). Significant improvements were also observed for partners in the affection (P=.007) and conflict (P=.001) subdomains of the consensus and satisfaction domains. In contrast to quantitative results, qualitative data indicated that intervention impacts included improved sleep and reduced pain, anxiety, and stress and, in contrast to the survey data, overall improvement in PTSD symptoms and social relationships. Participants' overall impressions of MR highlight usability and navigation, perceptions on packaging and content, and barriers to and facilitators of MR use.

Conclusions: Adjunctive CIH-based modalities can be delivered using web and mobile apps but should be developed and tailored using established best practices. MR may be beneficial for veterans with pain and PTSD and their partners. Further pragmatic trials and implementation efforts are warranted.

Trial Registration: ClinicalTrials.gov NCT03593772; https://clinicaltrials.gov/study/NCT03593772

International Registered Report Identifier (IRRID): RR2-10.2196/13666

(J Med Internet Res 2024;26:e57322) doi:10.2196/57322

KEYWORDS

posttraumatic stress disorder; PTSD; pain; veteran; attrition; complementary and integrative health; CIH; randomized controlled trial; chronic pain; remote intervention; dyad; mobile health

Introduction

Background

Chronic pain is an ongoing, debilitating morbidity that impacts an estimated 50 million Americans, roughly 20% of the US population [1]. Compared to the general population, veterans disproportionally experience chronic pain overall (29.1% vs 19.5%) [1], and their incidence of severe chronic pain is higher (9.7% vs 6.4%) [2]. In veterans, chronic pain commonly co-occurs with posttraumatic stress disorder (PTSD), which can complicate therapy options and exacerbate symptom severity [3-5]. In veteran clinical pain populations, the estimated PTSD prevalence is 11.7% [6]. Co-occurrence of chronic pain and PTSD is associated with poor outcomes, such as elevated PTSD symptoms, pain intensity, sleep issues, pain catastrophizing beliefs, and psychological health outcomes [7-9].

Driven by need for nonpharmacological interventions, evidence, and veteran demand, Subtitle C of the 2016 Comprehensive Addiction and Recovery Act mandated expansion of research, education, and delivery of complementary and integrative health (CIH) for veterans receiving pain and mental health services [10]. Following the Comprehensive Addiction and Recovery Act, Veterans Health Administration (VHA) Directive 1137 [11] supported transformation to a whole health system of care to support proactive delivery of evidence-based CIH modalities to address veterans' health and wellness needs while also creating alternatives for nonpharmacological options in a national opioid epidemic [11]. In this transformation to whole health, 8 evidence-based CIH therapies were identified and included at the Department of Veterans Affairs (VA) facilities as medical benefits—acupuncture, meditation, tai chi, yoga, massage therapy, guided imagery, biofeedback, and clinical hypnosis [12-18]. These CIH modalities have demonstrated significant impacts in military and veteran populations in managing pain intensity and physical functioning [19].

While whole health and CIH services provide an evidence-based biopsychosocial integrative approach to pain management [20], examining the use of CIH-based mobile apps and resources to complement in-person clinical care [21] is warranted. Mission Reconnect (MR) is a remotely delivered, mobile and web-based CIH-based didactic self-care intervention (ie, education, massage therapy, meditation, and positive psychology) that leverages social relationships between veterans and their partners and has demonstrated improvements in pain and PTSD measures among veterans and service members in a powered randomized controlled trial conducted in a community-based setting. The MR intervention has been further detailed in a previous publication [9].

Massage therapy, the primary partnered CIH component of MR, has been examined extensively as a nonpharmacologic approach to pain management [22-26] in diverse populations. More recently, the VA has supported efforts to examine the potential effects of massage therapy [27,28], with results indicating potential to support pain management. Massage therapy and mindfulness have shown low- to moderate-quality evidence, indicating improvement in low back pain, short term (ie, immediate effects after the treatment) [24,29-37], cancer pain [25], and surgical pain [26]. For other areas of pain (eg, shoulder and neck), the evidence is conflicting, with primarily low-quality, short-term studies [30,35,38]. Although the field

XSL•FO
RenderX

has shown interest in the use of massage therapy with other CIH modalities to manage PTSD-related symptoms, currently, research is limited [39-45]. Although a few studies have identified the potential benefit of the use of massage therapy and CIH-related modalities [39-41] to treat PTSD, others have indicated insignificant effects [42,44,45]. MR also provides "movement" exercises for veterans and their partners; this inclusion is based on the documented benefit of CIH such as yoga and tai chi for managing pain [15,46-50], including short-term and long-term impacts on pain in diverse populations [51].

Finally, MR also includes a mindfulness-based component, which has been associated with short-term effects on pain intensity and physical functioning [33,52-55]. Several reviews have indicated that mindfulness-based interventions have significant effects on neck pain [46] and low back pain specifically [48,56-61]; meta-analytic results indicate large effect sizes for improving pain intensity, unpleasantness, and interference [62,63]. In contrast, a recent review in a veteran population indicated that mindfulness-based stress reduction therapies had significant impacts on measures of PTSD, depression, general psychological symptoms, quality of life (QOL) and functioning, and mindfulness, but not on physical health and did not sustain effects at follow-up [64]. A separate meta-analysis examining meditation for improving depression symptoms in participants with chronic pain indicated small effect sizes and poor study quality [65]. There is evidence suggesting that mind-body exercises (eg, yoga, tai chi, and qigong) have significant impacts on PTSD [66], depressive, and anxiety symptoms [67-69]. In general, mindfulness-based (eg, meditation) trials have demonstrated significant improvements in PTSD symptoms [70-76].

Specific to the veteran population, as early as 2011, a VA evidence review indicated that a review of studies reported short-term effects of CIH interventions on PTSD symptoms [77]. Meditation interventions have been shown to have small– to medium–effect size improvements for alleviating military-related PTSD symptoms [78]. Reviews of the literature provide evidence of significant effects using meditation-based therapy for managing PTSD symptoms in military and veteran populations [78-80]. Although the evidence is compelling, more work demonstrating the effects of CIH for managing PTSD-related symptoms using standardized protocols and high-quality reporting are warranted.

As research is needed to examine the use of CIH modalities to manage pain and PTSD in clinical populations, examination of the use of mobile and web-based technology to deliver CIH interventions is also warranted. In a scoping review of the literature examining the use of mobile and web-based resources to deliver CIH, of the 330 articles included, only 2 (excluding previous MR publications) included relevant partnered interventions [81]. A previous feasibility study of caregiver-provided massage therapy found statistically significant improvements in veteran pain, stress, and fatigue [82]. Another pilot partnered intervention using remotely telephone-delivered symptom management found improvement in pain interference and psychological distress for patient-caregiver dyads that had greater assertive communication but, surprisingly, increased pain and fatigue interference for dyads with greater mindfulness practice [83]. We are not aware of any publications on completed remotely delivered partnered interventions since our published scoping review; a search of the literature revealed only 2 protocol papers [84,85].

Specific to MR, a previous community-based pilot study demonstrated that MR use was associated with pretest-posttest improvements in pain intensity and PTSD symptoms among a sample of postdeployment service members and a consenting partner of their choosing (97.7% spouses or life partners) [86]. A randomized controlled trial of community-based post-9/11 combat veterans and their self-selected partners (94.4% spouses or life partners) found that pain intensity and PTSD symptoms improved when MR was used alone or as an adjunct to a relationship enhancement intervention compared to a waitlist control group [87]. Of note, despite PTSD symptom scores often exceeding clinically recommended cutoffs for diagnostic screening in the observed samples, formal chronic pain and PTSD diagnoses were not used as inclusion criteria in either study. Given the demonstrated efficacy of MR in reducing pain intensity and PTSD symptoms, the purpose of this study was to examine the effects of MR among a sample of veterans diagnosed with co-occurring chronic pain and PTSD within the VA system.

Objectives

This paper presents findings from a 4-year mixed methods randomized waitlist-controlled trial aimed to test MR for chronic pain and PTSD symptom management [9]. The aims of this research were to (1) determine MR effectiveness for physical (pain and sleep), PTSD (intrusion, arousal, avoidance, and numbing), and psychological (depression, stress, and anxiety) symptoms and global health (QOL); (2) determine MR effectiveness for social (relationship satisfaction and compassion for the self or others) outcomes among veterans and their partners; and (3) describe the veteran- and partner-perceived value of MR in a subsample of dyads using qualitative interviews to explore perceived outcomes and identify recommendations for program improvement.

Methods

Ethical Considerations

Study procedures and all research activities were reviewed for human participants protection and approved (Pro00035440) by the institutional review boards from the universities of South Florida, Michigan, and Washington. All participants consented before taking part in the randomized controlled trial. To incentivize participation, veterans were compensated with US $20 for each of the 4 assessments and US $5 per week for providing MR use and pain rating data. Interview participants received an additional US $20, with a total compensation of up to US $140. The methodology, consent and recruitment process, and baseline attrition have been described in detail elsewhere [9,88]. The study's methods and results have been reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (Multimedia Appendix 1) [89]. Data were deidentified to protect participants'

XSL•FO
RenderX

identities. A plan describing data management and access was developed to outline the applicable data management and access details for the project.

Design

A randomized controlled trial with a mixed methods design was used in which veterans and their partners were randomly assigned to the intervention (MR) and waitlist control arms. Repeated measurements of patient-reported outcomes were assessed within participants at weeks 1 to 8 and, finally, 16 weeks after assignment. Qualitative interviews were conducted 2 months after intervention completion. Including qualitative data to complement the quantitative findings contributed to a deeper understanding of participant experiences with the intervention. Intention-to-treat principles guided all quantitative and qualitative analyses.

Sample

Veteran-partner dyads (n=364) were recruited from 3 urban VA medical centers located in the southeastern, midwestern, and northwestern regions of the continental United States. Partners were family members, friends, and significant others chosen by the veteran to participate in the MR program and study. Veterans were recruited and consented based on a confirmed comorbid diagnosis of PTSD and chronic pain. Participants were required to be aged ≥18 years, English speaking, and capable of performing MR intervention practices and fulfilling the technology requirements. Additional eligibility requirements were absence of reported aggression or violence within the participant dyad, no history of moderate to severe traumatic brain injury (TBI), no sensory or physical dysfunction that may interfere with study activities (ie, cognitive, hearing, and vision), no recent or current diagnosis of or inpatient treatment for psychotic or substance use disorder, and no technology barriers interfering with MR and Qualtrics (Qualtrics International Inc) [90] survey access (eg, lack of access to internet or email) [9].

The number of dyads was selected based on having adequate statistical power on the pain outcome measure to detect a small to medium effect size (Cohen d=0.38) given an anticipated 20% attrition rate. However, the observed attrition rate was roughly 33%, and the number of dyads recruited increased from the original target of 228 to a total of 364 dyads [88]. Participants in the qualitative inquiry comprised a subsample of 35 dyads that were recruited from the intervention arm of the parent MR study. Participants agreed to be interviewed during the consenting process. Purposive sampling methodology was applied to select participants who self-reported use of the MR website and completion of study activities within the highest and lowest terciles of use. As the number of individuals who did not respond to use questions increased, we added a randomization option to ensure that we had enough participants for data saturation.

Materials

Screening Measures

A 16-item structured interview questionnaire was used to assess eligibility criteria (see the previous section) and randomization stratum (site, concurrent PTSD, and chronic pain treatment) for veteran-partner dyads. All 16 interview items were binary (*yes* or *no*) [9,88]. The Ohio State University TBI Identification Method was used to assess participants' TBI history [91]. This 8-item semistructured interview assesses self-report of head or high-impact injuries to determine participants' TBI history. Follow-up probes were used on ≤3 of the most severe injuries. Participants were excluded if they self-reported losing consciousness for 30 minutes to 24 hours (moderate TBI) or ≥24 hours (severe TBI) [92].

Patient-Reported Outcome Measures

All patient-reported outcome measures were collected using Qualtrics, a secure cloud-based data collection platform [90]. The Pain Outcomes Questionnaire–For Veterans (POQ-VA) [93]; Defense and Veterans Pain Rating Scale (DVPRS) [94]; and 3 single-item scales for pain, stress, and tension scale (PST) [9] were used as primary outcomes to assess the biopsychosocial impact of chronic pain. The POQ-VA and DVPRS, established VA pain measures, were administered at baseline and at the 4-, 8-, and 16-week time points. The single-item PST measures were administered at baseline and at the week 1 to week 8 time points, and again at week 16. The single-item measures for the PST were used to assess weekly fluctuations and assess short-term trends.

Veterans' PTSD symptom severity was examined using the PTSD Checklist for the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [95]. Secondary psychological health outcomes included the Beck Depression Inventory–II (BDI-II), which was used to measure veterans' depression severity [96]. Veteran sleep disturbance was measured using the Pittsburgh Sleep Quality Index (PSQI) [97]. The Perceived Stress Scale (PSS) measured veterans' self-reported stress levels [98]. Finally, health-related QOL (HRQoL) was measured using the 12-Item Short Form Health Survey (SF-12) [99]. Descriptions, number of items, response scales, and scoring are described elsewhere [9,88]. The PTSD Checklist for the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; PSQI; PSS; SF-12; and BDI-II were administered at baseline and at the 4-, 8-, and 16-week time points (Multimedia Appendix 2).

Relationship Measures

Assessment of dyads incorporated the Revised Dyadic Adjustment Scale (RDAS), which is a measure of relationship satisfaction [100]. The Self-Compassion Scale (SCS) was used for assessing the compassion of participants during difficult times [101], and the Compassionate Love Scale (CLS) was used to assess compassion and altruistic love for close family and friends [102]. The RDAS, SCS, and CLS were administered at baseline and at the 4-, 8-, and 16-week time points.

Qualitative Measures

Qualitative semistructured interviews were conducted to elicit rich descriptions of participant experiences accessing and using MR, including their overall impressions, perceived outcomes, and recommendations for program enhancement. Given the body-mind focus of the MR program, the study applied the biopsychosocial model as a guiding framework to explore the experiences of the program participants, recognizing the

XSL•FO
RenderX

multidimensionality of health and the interrelationships among the biological, psychological, and social components that contribute to one's health and well-being [103,104].

Onboarding Procedures

Block stratified randomization (site, concurrent PTSD, and chronic pain treatment) was used to assign consented dyads to the MR or waitlist control arms. Participants were then emailed a link to the MR website to create a profile and complete onboarding procedures. Participants who stagnated in the onboarding process were contacted up to 3 times, and study staff were available to assist participant dyads to complete onboarding. Following successful completion of the MR onboarding process, participant dyads were emailed a Qualtrics link to complete baseline survey and demographic measures. Dyads that completed the baseline survey were then unblinded to their assigned study condition.

Qualitative Interview Procedure

Participant interviews were scheduled after the first 2 months after completion of the study intervention. Participant dyads completed a structured interview in person or over the phone to determine their eligibility and complete the informed consent process. Interviews were scheduled for 60 minutes, and consisted of 6 semistructured questions to allow for probing of the main topics. The main topics encompassed participants' experiences participating in the MR program and using its various tools (eg, which tools were used and how), and suggestions on improving the use of MR and making it available within the VA. Data collection for the qualitative analysis began in January 2020 and was completed in May 2022.

Data Analyses

Descriptive Analyses

Descriptive data are presented as frequencies and percentages for categorical data, means and SDs for normally distributed continuous data, or medians and IQRs for nonnormally distributed continuous data. Baseline group comparisons were performed using 2-tailed chi-square tests, independent-samples *t* tests or Mann-Whitney *U* tests, and Fisher exact tests. Descriptive analyses and group comparisons were conducted using SAS (version 9.4; SAS Institute) [105].

Missing Data Analysis

Missing item-level scale data were imputed using predictive mean matching, a method that selects closest "donor" observations based on characteristics of sex, age, race, educational attainment, relationship status, relationship duration, computer and internet use, treatment group, and survey completion time point. Missing data were not imputed if >50% of the items were missing from a scale. Missing data imputation was additionally performed for impossible observations (eg, negative time asleep) for the PSQI instrument. Predictive mean matching was performed using the *MICE* package in R (version 4.1.2; R Foundation for Statistical Computing) [105-107].

Mixed Models for Primary and Secondary Outcomes

Linear mixed models were constructed for primary (pain) and secondary outcomes with fixed-effects terms for treatment group (MR vs waitlist control) and time (as days since activation in the protocol) and a treatment group × time interaction term to test whether the rate of change in the instrument score differed by treatment group. Random effects for the intercept and time were also included in all models. Mixed model analyses were conducted using SAS (version 9.4) [105]. The α *P* value for statistical significance was set at .05 for the primary outcome (pain) and its subscales and .01 for the secondary outcomes. Borderline statistical significance was considered to be <.10. Standardized mean difference effect size recommendations were adopted from a preregistered systematic review and meta-analysis [108] of pain management interventions to improve affect that align with MR and mindfulness principles (≤0.32=small; 0.33-0.55=moderate; ≥0.56=large).

Qualitative Analysis

Data content analysis followed inductive and deductive coding methods. A qualitative codebook consisting of known constructs from the literature and those that emerged inductively from the data was created. Emerging codes were added until no new codes emerged from the data. In total, 2 trained qualitative researchers coded the data using the qualitative data management software ATLAS.ti (version 22; ATLAS.ti Scientific Software Development GmbH) [109]. To establish intercoder reliability at 80% [110,111], 2 qualitative researchers coded 20% of the interviews (initially every fifth transcript and then randomly to ensure coding consistency) separately and then compared codes to determine percentage of agreement [112]. Coded text in ATLAS.ti (version 22) was exported and further analyzed in Microsoft Excel 365 spreadsheets (Microsoft Corp) to develop themes. To enhance credibility, interim findings were presented routinely during quarterly stakeholder meetings to garner stakeholder input and facilitate refinement of findings.

Results

Participant Characteristics

A flow diagram of the recruitment process leading to baseline survey completion has been published elsewhere (and can be reproduced) under the terms of Creative Commons Attribution 4.0 license [9,88] (Figure 1). Demographic information was examined for veterans in the MR (n=140) and waitlist control (n=136) conditions who completed baseline surveys. Veterans (n=276) had an average age of 56.54 (SD 13.74) years; were primarily male (201/276, 72.8%), White (198/276, 71.7%), non-Hispanic (236/276, 85.5%), and married or partnered (207/276, 75%); had attended or completed college or vocational school (256/276, 92.8%); and reported daily computer (167/276, 60.5%) and internet (222/276, 80.4%) use. No veteran differences were observed between the MR and waitlist control conditions across demographic characteristics (*P*≥.38 in all cases).

Partners in the MR (n=138) and waitlist control (n=134) conditions were also examined. Partners (n=272) had an average age of 52.56 (SD 14.12) years; were primarily female (222/272, 81.6%), White (206/272, 75.7%), non-Hispanic (244/272, 89.7%), and married or partnered (212/272, 77.9%); had attended or completed college or vocational school (226/272,

XSL•FO
RenderX

83.1%); and reported daily computer (177/272, 65.1%) and internet (239/272, 87.9%) use. No partner demographic differences were observed between the MR and waitlist control conditions ($P \geq .19$ in all cases). Demographic characteristics by study condition are presented in Table 1 [88]. Veteran-selected partners in the sample represented four basic categories: (1) partner or spouse, (2) other family member (child or sibling), (3) friend or caregiver, or (4) not identified. A plurality of the partners were partners or spouses in the treatment (96/196, 49%) and waitlist control (76/167, 45.5%) groups. Statistics for partner categories are illustrated in Table 2. Descriptive statistics for patient-reported outcome measures by condition are reported in Table 3.

Figure 1. Mission Reconnect recruitment flow diagram. ICD: International Statistical Classification of Diseases and Related Health Problems; MR: Mission Reconnect; NOPP: notice of privacy practices; PTSD: posttraumatic stress disorder; TBI: traumatic brain injury; VA: Department of Veterans Affairs.

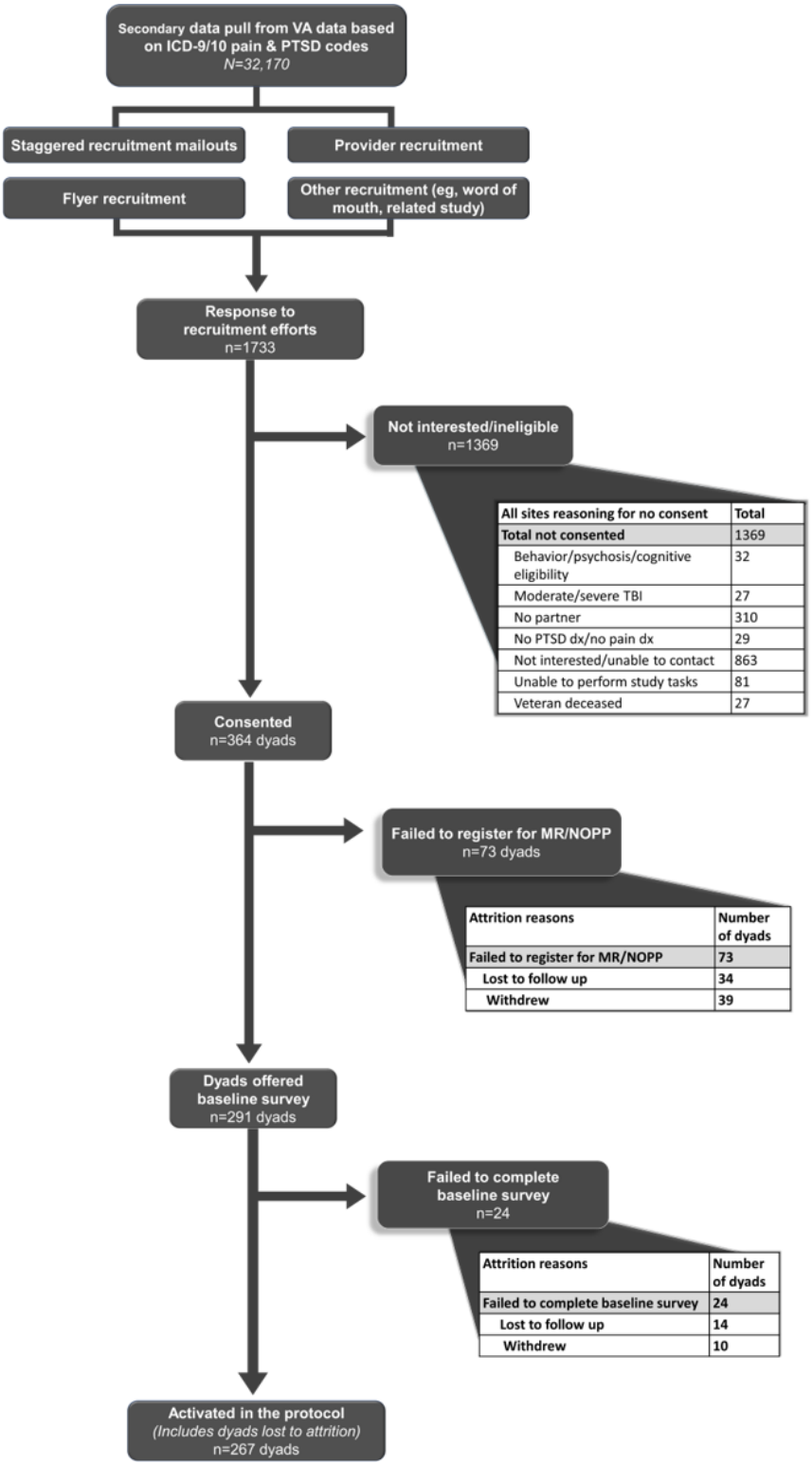

XSL•FO
RenderX

Table 1. Baseline demographic characteristics of participant dyads by study arm[a].

Characteristic	Veterans			Partners		
	Mission Reconnect (n=140)	Waitlist control (n=136)	*P* value[b]	Mission Reconnect (n=138)	Waitlist control (n=134)	*P* value
Age (y), mean (SD)	55.80 (14.13)	57.28 (13.35)	.38	52.20 (13.72)	52.92 (14.57)	.68
Sex, n (%)			.48			.19
Female	39 (27.9)	33 (24.3)		107 (77.5)	115 (85.8)	
Male	99 (70.7)	102 (75)		29 (21)	17 (12.7)	
Intersex	1 (0.7)	0 (0)		1 (0.7)	1 (0.7)	
Missing or declined to respond	1 (0.7)	1 (0.7)		1 (0.7)	1 (0.7)	
Race, n (%)			.94			.97
African American or Black	19 (13.6)	22 (16.2)		20 (14.5)	21 (15.7)	
American Indian or Alaska native	3 (2.1)	1 (0.7)		1 (0.7)	0 (0)	
Asian	1 (0.7)	1 (0.7)		5 (3.6)	4 (3)	
Multiracial	9 (6.4)	8 (5.9)		3 (2.2)	2 (1.5)	
Native Hawaiian or Pacific Islander	0 (0)	0 (0)		1 (0.7)	1 (0.7)	
White	102 (72.9)	96 (70.6)		104 (75.4)	102 (76.1)	
Other	4 (2.9)	6 (4.4)		2 (1.4)	1 (0.7)	
Missing or declined to respond	2 (1.4)	2 (1.5)		2 (1.4)	3 (2.2)	
Ethnicity, n (%)			.43			.94
Hispanic	13 (9.3)	10 (7.4)		11 (8)	9 (6.7)	
Non-Hispanic	121 (86.4)	115 (84.6)		123 (89.1)	121 (90.3)	
Missing or declined to respond	6 (4.3)	11 (8.1)		4 (2.9)	4 (3)	
Marital status, n (%)			.48			.65
Married or partnered	109 (77.9)	98 (72.1)		111 (80.4)	101 (75.4)	
Divorced, separated, or widowed	26 (18.6)	33 (24.3)		13 (9.4)	15 (11.2)	
Single or never married	4 (2.9)	3 (2.2)		12 (8.7)	15 (11.2)	
Missing or declined to respond	1 (0.7)	2 (1.5)		2 (1.4)	3 (2.2)	
Educational level, n (%)			.81			.30
Lower than high school	0 (0)	0 (0)		0 (0)	1 (0.7)	
High school	11 (7.9)	6 (4.4)		18 (13)	24 (17.9)	
Some college or vocational school	36 (25.7)	39 (28.7)		44 (31.9)	30 (22.4)	
Associate's degree	35 (25)	29 (21.3)		20 (14.5)	27 (20.1)	
Bachelor's degree	32 (22.9)	35 (25.7)		33 (23.9)	27 (20.1)	
Graduate degree	26 (18.6)	25 (18.4)		22 (15.9)	23 (17.2)	
Missing or declined to respond	1 (0.7)	2 (1.5)		1 (0.7)	2 (1.5)	
Daily computer use, n (%)	87 (62.1)	80 (58.8)	.74	94 (68.1)	83 (61.9)	.51
Daily internet use, n (%)	113 (80.7)	109 (80.1)	.83	120 (87)	119 (88.8)	.89

XSL•FO
RenderX

Characteristic	Veterans			Partners		
	Mission Reconnect (n=140)	Waitlist control (n=136)	*P* value[b]	Mission Reconnect (n=138)	Waitlist control (n=134)	*P* value
Years in relationship with partner, n (%)			.64			.84
<10	50 (35.7)	42 (30.9)		44 (31.9)	39 (29.1)	
10-29	48 (34.3)	52 (38.2)		52 (37.7)	53 (39.6)	
≥30	39 (27.9)	41 (30.1)		39 (28.3)	41 (30.6)	
Missing or declined to respond	3 (2.1)	1 (0.7)		3 (2.2)	1 (0.7)	

[a]Percentages may not add up to 100 exactly due to rounding error.

[b]*P* values obtained from 2-tailed *t* test, chi-square test, or Fisher exact test.

Table 2. Partner participant relationship to the veteran (n=363).

Relationship to the veteran	Mission Reconnect, n (%)	Waitlist control, n (%)	Total, n (%)
Partner, spouse, or significant other	96 (26.4)	76 (20.9)	172 (47.4)
Other family member (eg, child or sibling)	12 (3.3)	13 (3.6)	25 (6.9)
Friend or caregiver	6 (1.7)	6 (1.7)	12 (3.3)
Unknown	82 (22.6)	72 (19.8)	154 (42.4)
Total	196 (54)	167 (46)	363 (100)[a]

[a]One dyad was not assigned to a study arm before they withdrew.

XSL•FO
RenderX

Table 3. Baseline veteran- and partner-reported outcome mean scores by study arm.

Scores	Cronbach α	Veterans[a]			Partners		
		Mission Reconnect	Waitlist control	*P* value[b]	Mission Reconnect	Waitlist control	*P* value
Pain							
POQ-VA[c]							
Total, mean (SD)	0.87	97.25 (35.64)	87.57 (31.88)	.02[d]	N/A[e]	N/A	N/A
Pain intensity, mean (SD)	—[f]	6.25 (1.91)	6.07 (1.92)	.45	N/A	N/A	N/A
ADL[g] interference, median (IQR)	0.93	8.00 (1.50-20.00)	5.50 (2.00-19.00)	.35	N/A	N/A	N/A
Kinesiophobia, mean (SD)	0.55	11.38 (4.95)	10.88 (4.88)	.41	N/A	N/A	N/A
Mobility interference, mean (SD)	0.89	21.63 (11.54)	19.78 (10.78)	.17	N/A	N/A	N/A
Negative affect, mean (SD)	0.86	31.93 (11.59)	28.61 (11.50)	.02[e]	N/A	N/A	N/A
Vitality, mean (SD)	0.76	20.60 (5.74)	18.21 (5.71)	<.001[e]	N/A	N/A	N/A
DVPRS[h], mean (SD)							
Pain intensity	—	5.91 (1.69)	5.81 (1.80)	.64	N/A	N/A	N/A
Activity interference	—	6.15 (2.29)	6.07 (2.36)	.77	N/A	N/A	N/A
Sleep interference	—	6.25 (2.73)	5.65 (2.81)	.07	N/A	N/A	N/A
Mood	—	6.36 (2.58)	5.91 (2.54)	.14	N/A	N/A	N/A
Stress	—	6.29 (2.71)	5.96 (2.54)	.29	N/A	N/A	N/A
PST[i], median (IQR)							
Pain intensity	—	3.00 (1.00)	3.00 (1.00)	.19	N/A	N/A	N/A
Muscle tension	—	4.00 (1.00)	4.00 (2.00)	.50	N/A	N/A	N/A
Stress	—	3.00 (1.00)	3.00 (2.00)	.06	N/A	N/A	N/A
Psychological, mean (SD)							
PTSD[j] (PCL-5[k])	0.94	45.06 (15.92)	42.25 (16.57)	.15	N/A	N/A	N/A
Depression (BDI-II[l])	0.93	27.05 (13.10)	25.13 (12.58)	.22	N/A	N/A	N/A
Stress (PSS[m])	0.88	23.67 (6.77)	22.49 (6.65)	.14	N/A	N/A	N/A
Sleep (PSQI[n])	0.80	12.84 (2.92)	12.79 (3.12)	.90	N/A	N/A	N/A
Quality of life, mean (SD)							
Physical health (SF-12[o])	0.84	31.30 (8.11)	31.55 (7.38)	.63	N/A	N/A	N/A
Mental health (SF-12)	0.82	34.44 (8.66)	36.05 (8.74)	.12	N/A	N/A	N/A
Relationship							
Satisfaction (RDAS[p]): total, mean (SD)	0.86	3.35 (0.72)	3.44 (0.66)	.29	3.42 (0.66)	3.52 (0.67)	.22
Cohesion, mean (SD)	0.77	2.73 (0.96)	2.94 (1.05)	.10	2.86 (0.87)	2.86 (0.86)	.08
Consensus, median (IQR)	0.80	3.67 (3.0-4.17)	3.83 (3.33-4.17)	.47	3.83 (1.00)	4.00 (0.73)	.01[a]
Satisfaction, median (IQR)	0.78	3.75 (3.25-4.25)	3.75 (3.25-4.25)	.61	3.75 (1.00)	3.75 (1.00)	.71
Self-compassion (SCS[q]): total, mean (SD)	0.94	2.60 (0.77)	2.71 (0.80)	.22	3.21 (0.78)	3.24 (0.76)	.71

XSL•FO
RenderX

Scores	Cronbach α	Veterans[a]			Partners		
		Mission Reconnect	Waitlist control	*P* value[b]	Mission Reconnect	Waitlist control	*P* value
Common humanity, mean (SD)	0.79	2.71 (0.93)	2.83 (0.95)	.31	3.29 (1.01)	3.39 (1.00)	.39
Isolation, mean (SD)	0.82	2.56 (1.05)	2.61 (1.11)	.70	3.24 (1.08)	3.16 (0.97)	.55
Mindfulness, mean (SD)	0.80	2.88 (0.88)	3.09 (0.95)	.05	3.46 (0.95)	3.52 (0.88)	.65
Overidentification, mean (SD)	0.79	2.65 (1.04)	2.75 (0.95)	.44	3.19 (1.00)	3.20 (0.94)	.92
Self-kindness, mean (SD)	0.84	2.44 (0.91)	2.63 (1.00)	.09	3.11 (0.93)	3.18 (0.97)	.55
Self-judgment, mean (SD)	0.84	2.43 (0.97)	2.44 (0.96)	.83	3.02 (0.99)	3.06 (0.93)	.93
Compassion for others (CLS[r]), median (IQR)	0.95	5.81 (4.90-6.55)	5.95 (4.77-6.64)	.22	6.29 (0.97)	6.19 (1.09)	.16

[a]3 veterans were excluded for incomplete data.

[b]*P* values were obtained from independent-sample 2-tailed *t* tests or Mann-Whitney *U* tests.

[c]POQ-VA: Pain Outcomes Questionnaire–For Veterans.

[d]*P*<.05.

[e]N/A: not applicable.

[f]Cronbach α cannot be calculated for single-item measures.

[g]ADL: activity of daily living.

[h]DVPRS: Defense and Veterans Pain Rating Scale.

[i]PST: pain, stress, and tension scale.

[j]PTSD: posttraumatic stress disorder.

[k]PCL-5: PTSD Checklist for the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

[l]BDI-II: Beck Depression Inventory–II.

[m]PSS: Perceived Stress Scale.

[n]PSQI: Pittsburgh Sleep Quality Index.

[o]SF-12: 12-Item Short Form Health Survey.

[p]RDAS: Revised Dyadic Adjustment Scale.

[q]SCS: Self-Compassion Scale.

[r]CLS: Compassionate Love Scale.

Primary Outcomes

Improvements in total pain were not observed as measured using the POQ-VA (*P*=.38), DVPRS (*P*=.17), or PST (*P*=.19). In analyses of the subdomains of the POQ-VA (pain rating, impairment of mobility, *P*=.53; impairment of activities of daily living, *P*=.69; impairment of vitality, *P*=.17; and fear, *P*=.13), no significant improvements were detected over time, although a significant decline in the MR group relative to the waitlist control group was observed for negative affect (β=–.020; SE 0.010; *P*=.049; Table 4 and Figure 2). Similarly, while there was no change in pain interference with activity or stress for either group as measured using the DVPRS, there were significant reductions in pain interference in sleep (β=–.006; SE 0.002; *P*=.008) and mood (β=–.006; SE 0.002; *P*=.008) among the MR group compared to the waitlist control group (Figures 3A and 3B, respectively). For PTSD, no treatment effect was observed for the MR group compared to the waitlist control group over the follow-up period (*P*=.48).

XSL•FO
RenderX

Table 4. Fixed-effects estimates for select patient-reported Mission Reconnect (MR) outcomes.

Patient-reported outcome measure	Fixed effects β (SE; 95% CI[b])	*P* value	SMD[a]
Veterans			
POQ[c] (negative affect): time × treatment group (MR vs WC[d])	–0.020 (0.010; –0.039 to 0.000)	.049	–0.13
DVPRS[e] (sleep): time × treatment group (MR vs WC)	–0.006 (0.002; –0.011 to –0.002)	.008	–0.16
DVPRS (mood): time × treatment group (MR vs WC)	–0.006 (0.002; –0.011 to –0.002)	.008	–0.16
PST[f] (stress): time × treatment group (MR vs WC)	–0.002 (0.001; –0.004 to 0.000)	.10	–0.10
SF-12[g] (mental health): time × treatment group (MR vs WC)	0.017 (0.008; 0.002 to 0.033)	.03	0.13
SCS[h] (overidentification): time × treatment group (MR vs WC)	0.002 (0.001; 0.000 to 0.004)	.04	0.14
Partners			
RDAS[i] (total score): time × treatment group (MR vs WC)	0.021 (0.008; 0.005 to 0.038)	.01	0.19
RDAS (total consensus): time × treatment group (MR vs WC)	0.010 (0.004; 0.002 to 0.019)	.02	0.16
RDAS (consensus [affection subdomain]): time × treatment group (MR vs WC)	0.006 (0.002; 0.002 to 0.010)	.007	0.20
RDAS (total satisfaction): time × treatment group (MR vs WC)	0.004 (0.002; 0.000 to 0.009)	.07	0.07
RDAS (satisfaction [conflict subdomain]): time × treatment group (MR vs WC)	0.005 (0.001; 0.002 to 0.007)	.001	0.22

[a]SMD: standardized mean difference.

[b]Fixed-effects estimates for interaction term of time (per day) × treatment group (MR vs waitlist control) indicating that the rate of change over time differed between the MR group and the waitlist control group. Only outcomes that were found to be statistically or borderline significant ($P \leq .10$) are included in the table.

[c]POQ: Pain Outcomes Questionnaire–For Veterans.

[d]WC: waitlist control.

[e]DVPRS: Defense and Veterans Pain Rating Scale.

[f]PST: pain, stress, and tension scale.

[g]SF-12: 12-Item Short Form Health Survey.

[h]SCS: Self-Compassion Scale.

[i]RDAS: Revised Dyadic Adjustment Scale.

Figure 2. Predicted values for Pain Outcomes Questionnaire–For Veterans (POQ-VA) negative affect subdomain scores over time by treatment group. MR: Mission Reconnect; WC: waitlist control.

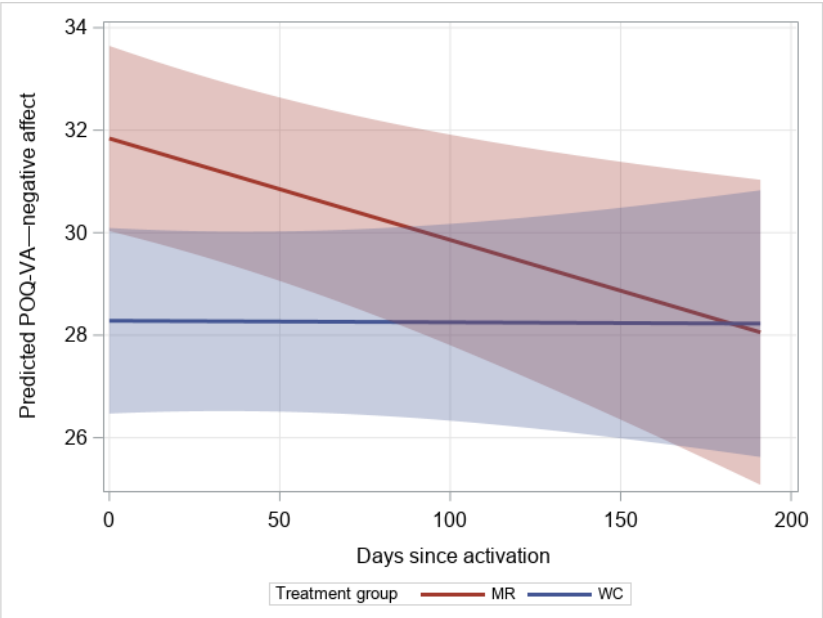

XSL•FO
RenderX

Figure 3. Predicted values for Defense and Veterans Pain Rating Scale (DVPRS) sleep (left panel) and mood (right panel) subdomain scores over time by treatment group. MR: Mission Reconnect; WC: waitlist control.

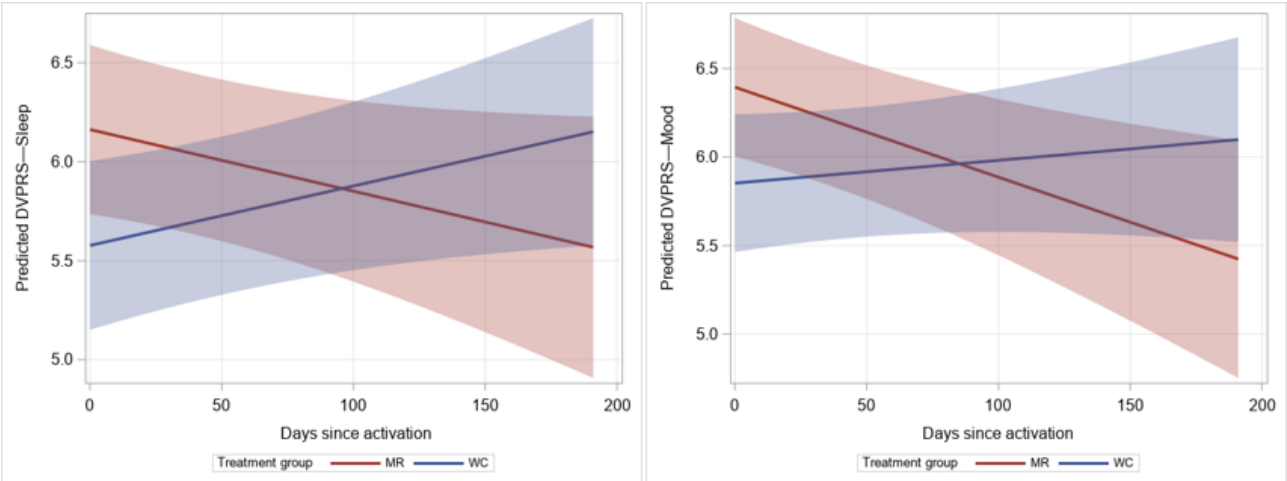

Secondary Outcomes

For total stress measured using the PST, we observed a reduction in stress in the MR group relative to the waitlist control group that was of borderline significance (β=–.002; SE 0.001; *P*>.99; Table 4); however, no effect of the MR intervention was observed for stress as measured using the PSS (*P*=.18). We also did not observe an effect of the MR intervention on other secondary outcomes, including muscle tension (PST; *P*=.73), depression (BDI-II; *P*=.20), sleep (PSQI; *P*=.24), and physical health (SF-12; *P*=.84). Mental health, as assessed using the SF-12, improved in the MR group relative to the waitlist control group (β=.017; SE 0.008; *P*=.03; Table 4), although this was considered to be of borderline significance (with the adjusted *P* value).

Among veterans, the average total SCS score did not differ by treatment group (*P*=.13). Similar results were found for the partners, with no difference by treatment group (*P*=.70). Many of the subdomain scores [self-kindness (*P*=.37), self-judgment (*P*=.26), common humanity (*P*=.33), isolation (*P*=.17), and mindfulness (*P*=.11)] did not differ by treatment group for either veterans or partners. For veterans, overidentification with personal failures and shortcomings significantly improved in the MR group compared to the waitlist control group (β=.002; SE 0.001; *P*=.04; Table 4). No effect of the MR intervention was observed for overidentification among partners (*P*=.76). Similarly, no effect was observed for either veterans or partners (*P*=.48) on the CLS (*P*=.24) and for veterans on the RDAS (*P*=.72).

We observed significant or borderline significant findings on the RDAS (ie, relationship satisfaction) for partners. Overall relationship satisfaction increased for partners in the MR group relative to those in the waitlist control group (β=.02; SE 0.01; *P*=.01; Table 4), as well as in the domains of consensus (β=.01; SE 0.004; *P*=.02) and satisfaction (β=.004; SE 0.002; *P*=.07), although these results were of borderline significance. These findings are likely driven by significant improvements in the affection (β=.01; SE 0.002; *P*=.007) and conflict (β=.005; SE 0.001; *P*=.001) subdomains of the consensus and satisfaction domains (Figure 4). No effect of the treatment was observed on the subdomains of consensus (decision-making, *P*=.14 and values, *P*=.63) and satisfaction (stability, *P*=.72), as well as the cohesion domain and its subdomains (activities, *P*=.11 and discussion, *P*=.16).

Figure 4. Predicted values for Revised Dyadic Adjustment Scale (RDAS) affection (left panel) and conflict (right panel) subdomain scores among partners over time by treatment group. MR: Mission Reconnect; WC: waitlist control.

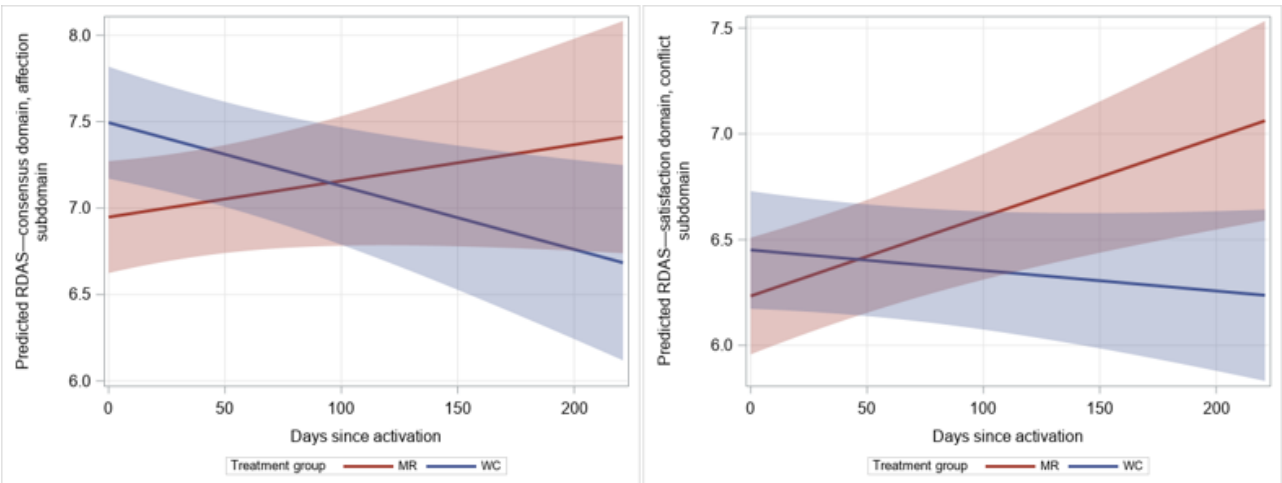

XSL•FO
RenderX

MR Satisfaction

Veteran and partner satisfaction with MR was relatively high at weeks 4, 8, and 16. For all 11 satisfaction items, median ratings ranged from 7 to 9 on a 10-point scale, with higher scores indicating greater satisfaction. Participant experiences with MR were further examined using qualitative data, as presented in the following section.

Qualitative Results

The interview subsample reflected the larger randomized controlled trial sample, representing veterans (n=35) with an average age of 53.94 (SD 12.76) years who were primarily male (28/35, 80%), non-Hispanic (29/35, 83%), and married or partnered (25/35, 71%); had attended or completed college or vocational school (32/35, 91%); and reported daily computer (25/35, 71%) and internet (29/35, 83%) use. Partner interviewees (n=35) had an average age of 49.89 (SD 13.88) years; were primarily female (29/35, 83%), non-Hispanic (31/35, 89%), and married or partnered (26/35, 74%); had attended or completed college or vocational school (33/35, 94%); and reported daily computer (21/35, 60%) and internet (29/35, 83%) use (Multimedia Appendix 3).

Overall Impressions of MR

The overall impression from participant qualitative interviews was that veterans and their partners valued the program, which corresponded with the quantitative satisfaction data. They reported that the program was inclusive, supported skill development, promoted a sense of positivity, and reinforced users' existing practices. Most participants reported scheduling MR activities and incorporating them into their daily routines. Nearly half reported using the app. Usability and navigation issues were relevant to user experience, consistent with similar apps. There were reports that indicated that the benefits such as portability and remote access from anywhere were highly valued. In addition, navigation and accessibility were relevant; however, reports were mixed. Participants requested guidance and support, which was provided as needed. An in-person component was recommended for programmatic implementation.

Barriers to using MR included personal characteristics or situations or not finding time to integrate the techniques into their lives. In addition, as with most video and audio-based interventions, length of time to complete the content was an issue for users. Other recommendations included providing additional content, such as shorter videos, content for family (including kids), additional meditations, and advanced material; accommodating different learning styles; and organizing content by symptoms and having a questionnaire to help users identify what they should prioritize. Facilitators of using MR included having a conducive environment, experiencing benefits, and the program content and structure.

Effect of MR

Veterans and their partners reported that the program was helpful and yielded positive outcomes. Veterans specifically indicated alleviation of symptoms related to pain, anxiety, and PTSD. In addition, veterans and partners reported improvement in QOL and influence on their relationship and communication. Multimedia Appendix 4 quantitatively counts veterans' and their partners' predominantly positive endorsement of outcome domains, whereas Multimedia Appendix 5 illustrates exemplary quotes by subdomain and participant type across all dimensions of the biopsychosocial model, including biological effects (ie, pain and sleep), psychological effects (ie, PTSD symptoms and anxiety), and social effects (ie, relationships with others). These domains and subdomains are illustrated in the following paragraphs.

Biological Effects: Pain and Sleep

Veterans talked about MR and their pain throughout the interviews. Some veteran participants reported that they learned activities to alleviate pain, although often the relief was temporary or inconsistent. Some participants reported trying several of the activities and obtaining skills that helped them cope with the pain. Overall, specific MR activities, such as partner massage and the activities "loosening your body" and "movement to stillness," were mentioned as a source of muscle release and relaxation. Reduced pain was linked to an overall improved attitude and QOL. For some participants who were not able to take or preferred not taking pain medication, the CIH approaches served as an effective nonpharmacological solution.

While some indicated a reduced pain intensity, activities made pain worse for other participants. For instance, the Waking up the Body exercise that incorporates rhythmical patting of meridians was too painful for some participants to incorporate into their routine. Massage of some body parts or with too hard a touch was also reported by some to make their pain worse.

The effect of MR on sleep was discussed minimally. In general, participants felt that the MR activities helped them fall asleep more easily; a few discussed "improved sleep." MR activities, such as intentional yawning to release tension (Reset and Refresh), guided meditation (Deep Relaxation), and swaying movement to align body and mind (Movement to Stillness) contributed to improved sleep.

Psychological Effects: PTSD, Anxiety, and Depression

Few participants made a direct connection between MR activities and PTSD symptoms specifically. Of those who did, participants described experiencing PTSD symptoms as being "stuck" in the past, reliving the feelings around a traumatic event. They reported a benefit to getting back in touch with and learning to live in the present. They reported being more aware and, thus, able to bring their focus to their current situation and environment. This shift in focus was reported to help limit negative PTSD symptoms. Of the participants who referenced PTSD-related nightmares, most reported using MR activities (mindfulness, quieting the mind, and massage therapy) to relieve nightmares and associated feelings.

Participants reported that MR was helpful in managing mental health symptoms, such as improving their anxiety. Veterans described using the videos and MR activities in a variety of ways (website, app, or saved to the cloud) to help manage anxiety and depression. Others reported accessing MR videos to manage PTSD symptoms during an event. Participants reported that the meditation and mindfulness activities were

XSL•FO
RenderX

also helpful and created a sense of quiet and calm, although these benefits may not necessarily last beyond the practice time. Participants reported better self-awareness that ultimately helped them understand how their PTSD affects their communication. With this self-awareness, they were able to change some of their communication and, therefore, improved their relationships and a sense of connection with others. Notably, some adverse effects were attributed to the study activities. Several veterans reported that completing the study survey items, especially those measuring PTSD symptoms, created stress and, at times, exacerbated their symptoms.

Social Effects: Relationships With Others

Participants reported that their use of MR and the CIH activities was beneficial for the relationship with their study partner. Participants overwhelmingly talked about improved communication, reporting that their communication occurred more often and was more productive. Improved communication, connection, and understanding of each other were often cited as outcomes of their study coparticipation. MR activities required scheduling time together in otherwise busy lives. This scheduled time was seen as beneficial to reconnecting with partners. Activities often required doing things outside their daily patterns, creating more interaction between dyads and promoting more touch with partners. The interaction among partners indicated an improved connection and understanding. Participants reported that MR activities brought about a better understanding of the self (eg, meditation). This enhanced awareness of their own feelings facilitated improved communication and understanding among partners.

Effects on QOL

Several participants made general statements about MR's impact on their general QOL, such as "My life is better now." Participants spoke to the general improved QOL experience through using the MR platform and CIH activities. Those who spoke of improved QOL often spoke of adopting a positive attitude, especially at the beginning of the day; reconnecting or "centering" oneself; and feeling more relaxed.

Discussion

Principal Findings

Making nonpharmacological treatments such as CIH available remotely can help reduce overreliance on pain medications and promote patient-centered care, and can increase access and use of nonpharmacological pain treatments [113]. There is evidence that CIH can be used for pain management with little evidence of negative effects and can be easily implemented remotely with impactful results [21]. The purpose of this waitlist control mixed methods randomized controlled trial was to describe the physical, psychological, and social outcomes associated with the use of a self-directed remotely delivered CIH education intervention for veterans with comorbid chronic pain and PTSD and their partners.

Although participants reported high levels of satisfaction with MR, comparisons did not show an effect on our primary outcome, pain, as measured using the total POQ-VA score, DVPRS, or PST. Our findings diverge from the literature establishing the use of CIH, including massage therapy, movement, and mindfulness, to improve pain outcomes such as pain intensity and physical functioning in veteran and nonveteran populations [12,13,15,19]. We found significant small–effect size improvements on the subdomains of negative affect (POQ-VA) and pain interference on mood and sleep (DVPRS). We also observed a small–effect size improvement for mental HRQoL (SF-12 score) consistent with a meta-analysis that found a comparable small effect size improvement for massage therapy and improved overall HRQoL (using the SF-12 and 36-Item Short Form Health Survey) [24]. However, low certainty of evidence has been found for massage therapy in improving overall HRQoL among people with chronic pain [38]. For partners, the treatment effect was nearly statistically significant for overall relationship satisfaction and the domains of consensus and satisfaction (RDAS), with small effect sizes being observed. Statistically significant small–effect size improvements in the subdomains of affection and conflict were noted.

In contrast to our results, in a review of 2 randomized trials for co-occurring chronic pain and PTSD, pooled analyses found nonsignificant changes in pain intensity and interference but significant improvement in PTSD symptoms [72]. Interestingly, the authors noted that interventions targeting co-occurring chronic pain and PTSD typically target the latter, which may explain the limited effects on pain-related outcomes. The nonsignificant improvement in pain reduction during MR was inconsistent, with specific meta-analytic evidence supporting massage therapy over passive and active interventions for chronic pain overall [24,114]. However, the broader literature on massage therapy for pain management is mixed and inconclusive. Certain meta-analyses examining specific chronic pain conditions have found moderate evidence for low back [12,13] and neck pain [32,37] and strong evidence for shoulder pain [31]. An additional meta-analysis examining massage therapy in pooled analyses with other types of CIH found support for pain reduction [17]. Conversely, a recent Cochrane review [38] and evidence maps [15,35] indicate less certainty regarding massage therapy for pain reduction. Giannitrapani et al [15] found support for the benefits of massage therapy on certain types of pain (eg, general chronic pain) but unclear evidence for others (eg, neck pain), with variability across studies, pain location, and massage type. Although evidence in the literature is compelling, our powered randomized controlled trial did not find strong support for partnered massage for pain reduction. Our findings could be resultant of not focusing on a location-specific pain condition or the partnered delivery of massage versus professional massage, as is often practiced in CIH trials.

Mindfulness-based interventions are often used with chronic pain populations given their low intensity and limited adverse outcomes [12,114]. This study did not support existing meta-analytic evidence for pain reduction among people with chronic neck [46] and low back pain, [48,56-61] or depression reduction [65] among general chronic pain despite mindfulness being an important MR component. However, the MR intervention was associated with significant small–effect size improvements in pain interference with mood and sleep.

XSL•FO
RenderX

Although consistent, previous meta-analyses have found medium–effect size improvements in pain interference for massage therapy compared to inactive controls [62,63]. Previous reviews support massage therapy and mindfulness-based stress reduction techniques as moderately effective at reducing pain and improving function over the short term [22,24-26] but may not produce similar moderate to large effects for other types of pain [22,36]. Furthermore, this evidence is largely based on immediate, short-term, and pretest-posttest intervention effects, with limited evidence from studies to support follow-up effects [24,31,32,37,61]. As mentioned previously, these distinctions may inform why we did not find a significant reduction in pain as our population did not present a homogenous pain source. There is limited published research on the effects of nonpharmacologic therapies on subdomains of pain (negative affect and pain interference on mood and sleep), so these results warrant further examination.

Significant improvements for PTSD were not observed in this trial. This is inconsistent with systematic review and meta-analytic literature that indicates the promise of CIH approaches, including mindfulness and mind-body interventions, for PTSD and mental health symptoms (eg, depression and anxiety), which are directly relevant for MR [67,70,73,74,76]. However, several of these studies suggest caution due to possible publication bias [69] as well as studies being underpowered [74,79,80] and having low methodological quality [66,68,70,73]. Furthermore, studies have also demonstrated that evidence from randomized controlled trials examining mindfulness and mind-body interventions for PTSD symptom improvement is generally positive [48,67] or mixed [66,69,80]. Perhaps part of the reason for inconsistent findings across trials that include MR is that CIH interventions may require tailoring for these unique patient populations [68].

Our findings do not support the literature suggesting that CIH, such as meditation and breathing exercises, can be an effective self-management approach to PTSD symptoms for veterans [79,80]. However, Haider et al [79] found consistent evidence in the literature suggesting improvement in PTSD symptoms for veterans. Cushing and Braun [80] found evidence for improving PTSD and mental health symptoms, including depression, as well as evidence for improved sleep quality, which was supported in this trial. It is notable that sample sizes, methodological rigor, and limited reporting of effect sizes are identified as limiting factors in efforts to compare results across trials [79,80].

It is critical to note that the findings of this trial are relatively inconsistent with those of previous MR studies. A pilot study of postdeployment National Guard service members found significant pretest-posttest improvements in pain, tension, and depression. A follow-up randomized controlled trial found significant improvement in veterans' pain, PTSD, depression, sleep, tension, and self-compassion from baseline to after MR. The findings of our study only replicated improvements in sleep and a nonsignificant trend for negative affect (eg, depression). Effect sizes were unavailable for direct comparison. Of note, the differences among these studies may be accounted for by the fact that previous MR trials used nonclinical samples, which may have less severe conditions than the co-occurring pain and PTSD criteria for our trial. Still, our study did expand the MR evidence base supporting positive relationship outcomes, including small–effect size improvements in affection and reduced conflict, as well as a borderline significant effect for relationship satisfaction.

Convergences and divergences of quantitative and qualitative data were documented in the findings. In interviews, veterans and partners indicated benefits of the program and reported ease of incorporation of activities into their daily life activities. Veterans mainly highlighted physical benefits of the program specifically related to pain and PTSD symptoms. Veterans and partners indicated positive effects of the program on mental health, social health, and QOL. In addition, they mentioned that they slept better and felt more energized and refreshed. These findings are in line with those of previous publications on the benefits of remotely delivered CIH for veterans. Consistent with the quantitative data, the survey and interview data indicated improvement in pain interference and overidentification for veteran participants, as exemplified by the following quote:

> *Yeah, I think it's helped me—you know, being able to connect with my wife. It's helped me to feel more comfortable sharing with her when I have thoughts or struggles, or I'm just kind of feeling out of it. You know, I feel like I can—because we're more connected, I can share with her.*

Interestingly, interview data indicated improvement for partners in these same domains, although the quantitative data did not indicate differences among partners between the study arms. In addition, interview data indicated meaningful improvement for overall veteran-reported pain and PTSD symptoms, which was not noted in the quantitative findings. The inconsistencies between the quantitative and qualitative findings highlight a need for future research to identify appropriate CIH measures for veteran and partner populations [115,116].

The VHA is well positioned to provide nonpharmacological treatment for pain [116,117]. Nonpharmacological treatment for pain is associated with reducing opioid use and overall improved outcomes, although the quality [14,18,118] of evidence varies. Veterans who use VHA services have a greater likelihood of receiving nonpharmacological treatment, including educational classes and mind-body therapy, for chronic pain [119]. Given limitations in staffing and potential wait times for new patient appointments, MR may improve access and provide timely, patient-centered care as a fully remote intervention.

Implementation of MR as a fully remote mobile and web-based intervention has not been attempted in previous studies. It is notable that, in the study by Kahn et al [87], there was an in-person component in one of the treatment arms. We did not replicate the in-person component because the data from the aforementioned study did not indicate a significant benefit. In addition, due to the coincidence of data collection for this study with the COVID-19 outbreak, an in-person component would not have been feasible. It is possible that the lack of an in-person component impacted intervention engagement and modality use and, thereby, intervention outcomes [120].

XSL•FO
RenderX

Interview data highlighted MR's effects on PTSD symptoms and perceived improvements in communication, relationship quality, and sense of connectedness. Participation in the program did impact pain and interference, but mainly, the program supported acquisition of coping mechanisms. With regard to effectiveness on outcomes, effects may have been related to dose response. Despite reporting an overall positive experience and satisfaction with MR, participant recommendations for improvements in usability and engagement could enhance program use and attract more sustained users.

Future Research and Implementation

After nearly a decade, the VA continues its efforts to understand the impact of the whole health and CIH modalities on the health and well-being of veterans [116,117]. More evidence is needed to determine how to optimize the integration of whole health and CIH into health care to meet veterans' needs and values, particularly of veterans with PTSD [121]. Furthermore, there is evidence suggesting that, when service members engage in CIH approaches for pain while on active duty, they have a lower risk of substance abuse, overdose, and suicide attempt later in life [118]. As such, promoting the use of whole health and CIH throughout the military and veteran health care trajectories may be critical for improving outcomes, particularly for veterans with pain and PTSD.

Adjunctive modalities such as MR can help veterans manage chronic pain and PTSD symptoms, but further research using pragmatic trials [122] and implementation efforts [123] are warranted. The findings warrant a tailored approach for vulnerable at-risk populations, such as veterans with comorbid chronic pain and PTSD. As we advance the science to determine the risks and benefits of CIH-based interventions specifically for PTSD, some features of CIH modalities, such as massage therapy and touch, may cause emotional discomfort for this vulnerable population. Future studies should aim to (1) increase sampling to allow for analysis of stratified pain or PTSD profiles; (2) explore divergent findings specifically related to pain, PTSD, and relationship outcomes; (3) reassess the fit of outcome measures to address sensitivity to change of outcomes, such as connectedness; (4) assess professional versus partnered massage with attention to those who experience PTSD symptoms; (5) conduct an intervention component and dosage evaluation; and (6) evaluate the CIH effects on opioid use [49] and misuse. Consideration of a standardized approach to measuring patient self-reported use and outcomes is also warranted to support advancing the science through the continued application of cross-study meta-analyses [115].

When planning future research with this vulnerable population, the increased risk of frustration with onboarding and data collection processes should be considered. Authors recommend simplifying onboarding processes and providing personal support throughout the study and increasing automation, reminders, and project navigators (electronic and human). With studies relevant to PTSD, revisiting trauma is a risk—as such, we recommend a rigorous screening and consent process that clearly prepares patients for the types of emotional, mental, and physical health data they will be providing to manage expectations and burden.

The findings indicate that an in-person component in tandem with the intervention is warranted to provide personalized support. Recommended strategies included personalized recommendations based on user preferences, needs, and progress; tailoring content to individual pain and PTSD profiles; and fostering a sense of community among users by integrating social features to create a supportive network. Future implementation will benefit from efforts to improve 508 compliance standards set for federal agencies to ensure technology features are accessible [124] for optimized usability and navigation.

Strengths and Limitations

Although the findings are compelling, the balance of strengths and study limitations should be considered. First, this trial was launched during the COVID-19 pandemic. Consistent with the literature [120], many adjustments were required for nonpharmacological pain management trials during the pandemic; however, the remote nature of MR minimized disruption to the trial process and outcomes. Nonetheless, we acknowledge that the pandemic could have presented confounding factors (eg, social isolation and pandemic-related competing demands) that may have influenced the study findings. Second, regarding mobile and web-based delivery, the data represented 3 distinct sites; however, participants from each site represent a larger geographic area that comprises rural and nonrural areas across most of the United States. Third, participants reported that "too many" surveys were required at each time point. At the time of this study, no single instrument existed to capture the CIH patient experience. Including a single survey for CIH patient experience [123] may have limited survey burden, thus improving the reliability of the findings. Fourth, we were unable to examine dose effect (eg, more frequent or longer use of the MR program could result in improved outcomes) due to the lack of an effective measure of frequency and duration of use and lack of consistent participant reporting. Fifth, we did not further adjust for demographic variables, including those used for the missing data imputation, in the final models due to the relatively small sample size and general lack of significant findings. However, as randomization is expected to result in balanced groups, it is generally not necessary to further adjust for baseline characteristics in randomized clinical trials, and baseline demographics did not differ between groups. We also did not consider, in the final models, concurrent treatments for pain and PTSD that may have had synergetic or deleterious effects on the program's impact on patient-reported outcomes. Sixth, regarding the qualitative results, while we achieved thematic saturation, the findings may not be generalizable to other CIH programs, and insights may have been overlooked from the veterans or partners who participated in the larger parent randomized controlled trial but did not participate in the interviews.

Conclusions

MR offers a CIH-based intervention for veterans with chronic pain and PTSD and their partners, with significant impacts on pain interference and relationship quality. Additional design and content modification can enhance the program's fit and use for this population and may generate greater uptake. Data

XSL•FO
RenderX

findings can inform MR program enhancements, protocol design for subsequent pragmatic trials in diverse populations, and implementation of MR and other nonpharmacological CIH-based remotely delivered partnered programs in clinical settings. The findings warrant future efforts to explore the role of pain interference in pain management interventions, impacts on relationship outcomes, and cost-effectiveness of the program.

Acknowledgments

The Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development and Rehabilitation Research and Development Service (822-MR-35876) supported the development of this manuscript. This study is registered on ClinicalTrials.gov as NCT03593772. ChatGPT or any other generative artificial intelligence was not used in the development of manuscript content, and the contents do not represent the views of the Department of Veterans Affairs or the US government. The corresponding author, as a US government employee, has the right to grant on behalf of all authors and does grant on behalf of all authors a nonexclusive worldwide license to the publishers and its licensees in perpetuity, in all forms, formats, and media (whether known now or created in the future), to (1) publish, reproduce, distribute, display, and store the contribution; (2) translate the contribution into other languages, create adaptations and reprints, and include the Contribution in collections, as well as create summaries, extracts, or abstracts of the contribution; (3) create any other derivative work based on the contribution; (4) exploit all subsidiary rights in the contribution; (5) include electronic links from the contribution to third-party material wherever it may be located; and (6) license any third party to do any or all of the aforementioned.

Data Availability

The deidentified anonymized datasets included in this manuscript are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6.x).
[PDF File (Adobe PDF File), 4327 KB - jmir_v26i1e57322_app1.pdf]

Multimedia Appendix 2
Patient-reported outcome measures administered.
[DOCX File , 17 KB - jmir_v26i1e57322_app2.docx]

Multimedia Appendix 3
Demographic characteristics of Mission Reconnect participants who completed qualitative interviews.
[DOCX File , 17 KB - jmir_v26i1e57322_app3.docx]

Multimedia Appendix 4
Outcome domains endorsed in interviews by unique respondent type.
[DOCX File , 14 KB - jmir_v26i1e57322_app4.docx]

Multimedia Appendix 5
Veteran and partner dyad quotes representing qualitative domains.
[DOCX File , 18 KB - jmir_v26i1e57322_app5.docx]

References

1. Dahlhamer J, Lucas J, Zelaya C, Nahin R, Mackey S, DeBar L, et al. Prevalence of chronic pain and high-impact chronic pain among adults - United States, 2016. MMWR Morb Mortal Wkly Rep 2018 Sep 14;67(36):1001-1006 [FREE Full text] [doi: 10.15585/mmwr.mm6736a2] [Medline: 30212442]
2. Nahin RL. Severe pain in veterans: the effect of age and sex, and comparisons with the general population. J Pain 2017 Mar;18(3):247-254 [FREE Full text] [doi: 10.1016/j.jpain.2016.10.021] [Medline: 27884688]
3. López-Martínez AE, Reyes-Pérez Á, Serrano-Ibáñez ER, Esteve R, Ramírez-Maestre C. Chronic pain, posttraumatic stress disorder, and opioid intake: a systematic review. World J Clin Cases 2019 Dec 26;7(24):4254-4269 [FREE Full text] [doi: 10.12998/wjcc.v7.i24.4254] [Medline: 31911906]

XSL•FO
RenderX

4. Murphy JL, Driscoll MA, Odom AS, Hadlandsmyth K. Posttraumatic stress disorder and chronic pain. PTSD Res Q 2017;33(2):1-10 [FREE Full text]
5. Benedict TM, Keenan PG, Nitz AJ, Moeller-Bertram T. Post-traumatic stress disorder symptoms contribute to worse pain and health outcomes in veterans with PTSD compared to those without: a systematic review with meta-analysis. Mil Med 2020 Sep 18;185(9-10):e1481-e1491. [doi: 10.1093/milmed/usaa052] [Medline: 32248229]
6. Siqveland J, Hussain A, Lindstrøm JC, Ruud T, Hauff E. Prevalence of posttraumatic stress disorder in persons with chronic pain: a meta-analysis. Front Psychiatry 2017 Sep 14;8:164 [FREE Full text] [doi: 10.3389/fpsyt.2017.00164] [Medline: 28959216]
7. Åkerblom S, Perrin S, Rivano Fischer M, McCracken L. The relationship between posttraumatic stress disorder and chronic pain in people seeking treatment for chronic pain: the mediating role of psychological flexibility. Clin J Pain 2018 Jun;34(6):487-496. [doi: 10.1097/AJP.0000000000000561] [Medline: 29016388]
8. Alschuler KN, Otis JD. Coping strategies and beliefs about pain in veterans with comorbid chronic pain and significant levels of posttraumatic stress disorder symptoms. Eur J Pain 2012 Feb 06;16(2):312-319. [doi: 10.1016/j.ejpain.2011.06.010] [Medline: 22323383]
9. Haun JN, Ballistrea LM, Melillo C, Standifer M, Kip K, Paykel J, et al. A mobile and web-based self-directed complementary and integrative health program for veterans and their partners (mission reconnect): protocol for a mixed-methods randomized controlled trial. JMIR Res Protoc 2019 May 13;8(5):e13666 [FREE Full text] [doi: 10.2196/13666] [Medline: 31094345]
10. S.524 - comprehensive addiction and recovery act of 2016. Government of USA. URL: https://www.congress.gov/bill/114th-congress/senate-bill/524/text [accessed 2024-04-29]
11. VA directive 1137 — provision of complementary and integrative health. US Department of Veterans Affairs. URL: https://www.va.gov/WHOLEHEALTH/professional-resources/clinician-tools/cih.asp [accessed 2024-04-29]
12. Chou R, Huffman LH, American Pain Society, American College of Physicians. Nonpharmacologic therapies for acute and chronic low back pain: a review of the evidence for an American Pain Society/American College of Physicians clinical practice guideline. Ann Intern Med 2007 Oct 02;147(7):492-504 [FREE Full text] [doi: 10.7326/0003-4819-147-7-200710020-00007] [Medline: 17909210]
13. Furlan AD, Yazdi F, Tsertsvadze A, Gross A, Van Tulder M, Santaguida L, et al. A systematic review and meta-analysis of efficacy, cost-effectiveness, and safety of selected complementary and alternative medicine for neck and low-back pain. Evid Based Complement Alternat Med 2012;2012:953139-953161 [FREE Full text] [doi: 10.1155/2012/953139] [Medline: 22203884]
14. Peterson K, Anderson J, Ferguson L, Mackey K. Evidence brief: the comparative effectiveness of selected complementary and integrative health (CIH) interventions for preventing or reducing opioid use in adults with chronic neck, low back, and large joint pain. Department of Veterans Affairs (US). 2016. URL: https://www.ncbi.nlm.nih.gov/books/NBK409241/#:~:text=Possible%20roles%20for%20CIH%20in,for%20pain%20exacerbations%20during%20stable [accessed 2024-04-29]
15. Giannitrapani KF, Holliday JR, Miake-Lye IM, Hempel S, Taylor SL. Synthesizing the strength of the evidence of complementary and integrative health therapies for pain. Pain Med 2019 Sep 01;20(9):1831-1840. [doi: 10.1093/pm/pnz068] [Medline: 31070752]
16. Report of the commission on creating options for veterans' expedited recovery. Cover Commission. 2020. URL: https://www.va.gov/COVER/docs/COVER-Commission-Final-Report-2020-01-24.pdf [accessed 2024-08-06]
17. Dyer NL, Surdam J, Dusek JA. A systematic review of practiced-based research of complementary and integrative health therapies as provided for pain management in clinical settings: recommendations for the future and a call to action. Pain Med 2022 Jan 03;23(1):189-210 [FREE Full text] [doi: 10.1093/pm/pnab151] [Medline: 34009391]
18. Soares Fonseca L, Pereira Silva J, Bastos Souza M, Gabrich Moraes Campos M, de Oliveira Mascarenhas R, de Jesus Silva H, et al. Effectiveness of pharmacological and non-pharmacological therapy on pain intensity and disability in older people with chronic nonspecific low back pain: a systematic review with meta-analysis. Eur Spine J 2023 Sep 18;32(9):3245-3271. [doi: 10.1007/s00586-023-07857-4] [Medline: 37464184]
19. Donahue ML, Dunne EM, Gathright EC, DeCosta J, Balletto BL, Jamison RN, et al. Complementary and integrative health approaches to manage chronic pain in U.S. military populations: results from a systematic review and meta-analysis, 1985-2019. Psychol Serv 2021 Aug;18(3):295-309 [FREE Full text] [doi: 10.1037/ser0000417] [Medline: 32134305]
20. VA/DoD clinical practice guideline for the use of opioids in the management of chronic pain. Department of Veterans Affairs (US). URL: https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOpioidsCPG.pdf [accessed 2024-06-11]
21. Mullur RS, Kaur Cheema SP, Alano RE, Chang LE. Tele-integrative medicine to support rehabilitative care. Phys Med Rehabil Clin N Am 2021 May;32(2):393-403. [doi: 10.1016/j.pmr.2020.12.006] [Medline: 33814064]
22. Patel KC, Gross A, Graham N, Goldsmith CH, Ezzo J, Morien A, et al. Massage for mechanical neck disorders. Cochrane Database Syst Rev 2012 Sep 12(9):CD004871. [doi: 10.1002/14651858.CD004871.pub4] [Medline: 22972078]
23. Piper S, Shearer HM, Côté P, Wong JJ, Yu H, Varatharajan S, et al. The effectiveness of soft-tissue therapy for the management of musculoskeletal disorders and injuries of the upper and lower extremities: a systematic review by the Ontario Protocol for Traffic Injury management (OPTIMa) collaboration. Man Ther 2016 Feb;21:18-34. [doi: 10.1016/j.math.2015.08.011] [Medline: 26386912]

XSL•FO
RenderX

24. Crawford C, Boyd C, Paat CF, Price A, Xenakis L, Yang E, Evidence for Massage Therapy (EMT) Working Group. The impact of massage therapy on function in pain populations-a systematic review and meta-analysis of randomized controlled trials: part I, patients experiencing pain in the general population. Pain Med 2016 Jul 01;17(7):1353-1375 [FREE Full text] [doi: 10.1093/pm/pnw099] [Medline: 27165971]
25. Boyd C, Crawford C, Paat CF, Price A, Xenakis L, Zhang W, Evidence for Massage Therapy (EMT) Working Group. The impact of massage therapy on function in pain populations-a systematic review and meta-analysis of randomized controlled trials: part II, cancer pain populations. Pain Med 2016 Aug 10;17(8):1553-1568 [FREE Full text] [doi: 10.1093/pm/pnw100] [Medline: 27165967]
26. Boyd C, Crawford C, Paat CF, Price A, Xenakis L, Zhang W, Evidence for Massage Therapy (EMT) Working Group. The impact of massage therapy on function in pain populations-a systematic review and meta-analysis of randomized controlled trials: part III, surgical pain populations. Pain Med 2016 Sep 10;17(9):1757-1772 [FREE Full text] [doi: 10.1093/pm/pnw101] [Medline: 27165970]
27. Miake-Lye I, Lee J, Lugar T, Taylor S, Shanman R, Beroes J, et al. Massage for pain: an evidence map. Department of Veterans Affairs (US). 2016. URL: https://www.ncbi.nlm.nih.gov/books/NBK424144/ [accessed 2024-06-04]
28. Mak S, Allen J, Begashaw M, Miake-Lye I, Beroes-Severin J, Lawson E, et al. Evidence map of massage therapy: update from 2018-2023. Department of Veterans Affairs (US). 2023. URL: https://www.hsrd.research.va.gov/publications/esp/evidence-massage.pdf [accessed 2024-04-29]
29. Cherkin DC, Sherman KJ, Deyo RA, Shekelle PG. A review of the evidence for the effectiveness, safety, and cost of acupuncture, massage therapy, and spinal manipulation for back pain. Ann Intern Med 2003 Jun 03;138(11):898-906. [doi: 10.7326/0003-4819-138-11-200306030-00011] [Medline: 12779300]
30. Tsao JC. Effectiveness of massage therapy for chronic, non-malignant pain: a review. Evid Based Complement Alternat Med 2007 Jun;4(2):165-179 [FREE Full text] [doi: 10.1093/ecam/nel109] [Medline: 17549233]
31. Kong LJ, Zhan HS, Cheng YW, Yuan WA, Chen B, Fang M. Massage therapy for neck and shoulder pain: a systematic review and meta-analysis. Evid Based Complement Alternat Med 2013;2013:613279 [FREE Full text] [doi: 10.1155/2013/613279] [Medline: 23533504]
32. Cheng YH, Huang GC. Efficacy of massage therapy on pain and dysfunction in patients with neck pain: a systematic review and meta-analysis. Evid Based Complement Alternat Med 2014;2014:204360 [FREE Full text] [doi: 10.1155/2014/204360] [Medline: 24695806]
33. Hilton L, Hempel S, Ewing BA, Apaydin E, Xenakis L, Newberry S, et al. Mindfulness meditation for chronic pain: systematic review and meta-analysis. Ann Behav Med 2017 Apr 22;51(2):199-213 [FREE Full text] [doi: 10.1007/s12160-016-9844-2] [Medline: 27658913]
34. Qaseem A, Wilt TJ, McLean RM, Forciea MA, Clinical Guidelines Committee of the American College of Physicians. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2017 Apr 04;166(7):514-530. [doi: 10.7326/M16-2367] [Medline: 28192789]
35. Miake-Lye IM, Mak S, Lee J, Luger T, Taylor SL, Shanman R, et al. Massage for pain: an evidence map. J Altern Complement Med 2019 May;25(5):475-502 [FREE Full text] [doi: 10.1089/acm.2018.0282] [Medline: 30892910]
36. Skelly AC, Chou R, Dettori JR, Turner JA, Friedly JL, Rundell SD, et al. Noninvasive nonpharmacological treatment for chronic pain: a systematic review update. Agency for Healthcare Research and Quality (AHRQ). URL: https://www.ncbi.nlm.nih.gov/books/NBK556229/ [accessed 2024-06-05]
37. Wang SQ, Jiang AY, Gao Q. Effect of manual soft tissue therapy on the pain in patients with chronic neck pain: a systematic review and meta-analysis. Complement Ther Clin Pract 2022 Nov;49:101619. [doi: 10.1016/j.ctcp.2022.101619] [Medline: 35988324]
38. Gross AR, Lee H, Ezzo J, Chacko N, Gelley G, Forget M, et al. Massage for neck pain. Cochrane Database Syst Rev 2012 Sep 12(9):CD004871. [doi: 10.1002/14651858.CD004871.pub4] [Medline: 22972078]
39. Libretto S, Hilton L, Gordon S, Zhang W, Wesch J. Effects of integrative PTSD treatment in a military health setting. Energy Psychol Theory Res Treat 2015 Nov 1;7(2):33-44. [doi: 10.9769/EPJ.2015.7.2.SL]
40. Rosenow M, Munk N. Massage for combat injuries in veteran with undisclosed PTSD: a retrospective case report. Int J Ther Massage Bodywork 2020 Dec 08;14(1):4-11 [FREE Full text] [doi: 10.3822/ijtmb.v14i1.555]
41. Field T, Seligman S, Scafidi F, Schanberg S. Alleviating posttraumatic stress in children following hurricane Andrew. J Appl Dev Psychol 1996 Jan;17(1):37-50. [doi: 10.1016/S0193-3973(96)90004-0]
42. Price C. Body-oriented therapy in recovery from child sexual abuse: an efficacy study. Altern Ther Health Med 2005;11(5):46-57 [FREE Full text] [Medline: 16189948]
43. Ferguson PE, Persinger D, Steele M. Resolving dilemmas through bodywork. Int J Ther Massage Bodywork 2010 Mar 17;3(1):41-47 [FREE Full text] [doi: 10.3822/ijtmb.v3i1.74] [Medline: 21589702]
44. Research: massage for female veterans with PTSD. American Massage Therapy Association. URL: https://www.amtamassage.org/publications/massage-therapy-journal/massage-for-female-veterans-with-ptsd/ [accessed 2024-04-29]
45. Sumpton B, Baskwill A. A series of case reports regarding the use of massage therapy to improve sleep quality in individuals with post-traumatic stress disorder (PTSD). Int J Ther Massage Bodywork 2019 Sep 17;12(4):3-9. [doi: 10.3822/ijtmb.v12i4.381]

XSL•FO
RenderX

46. Mueller J, Weinig J, Niederer D, Tenberg S, Mueller S. Resistance, motor control, and mindfulness-based exercises are effective for treating chronic nonspecific neck pain: a systematic review with meta-analysis and dose-response meta-regression. J Orthop Sports Phys Ther 2023 Aug;53(8):420-459. [doi: 10.2519/jospt.2023.11820] [Medline: 37339388]
47. Anheyer D, Haller H, Lauche R, Dobos G, Cramer H. Yoga for treating low back pain: a systematic review and meta-analysis. Pain 2022 Apr 01;163(4):e504-e517. [doi: 10.1097/j.pain.0000000000002416] [Medline: 34326296]
48. Zou L, Zhang Y, Yang L, Loprinzi PD, Yeung AS, Kong J, et al. Are mindful exercises safe and beneficial for treating chronic lower back pain? A systematic review and meta-analysis of randomized controlled trials. J Clin Med 2019 May 08;8(5):628 [FREE Full text] [doi: 10.3390/jcm8050628] [Medline: 31072005]
49. Lin YC, Wan L, Jamison RN. Using integrative medicine in pain management: an evaluation of current evidence. Anesth Analg 2017 Dec;125(6):2081-2093. [doi: 10.1213/ANE.0000000000002579] [Medline: 29189365]
50. Cramer H, Klose P, Brinkhaus B, Michalsen A, Dobos G. Effects of yoga on chronic neck pain: a systematic review and meta-analysis. Clin Rehabil 2017 Nov 09;31(11):1457-1465. [doi: 10.1177/0269215517698735] [Medline: 29050510]
51. Cramer H, Lauche R, Haller H, Dobos G. A systematic review and meta-analysis of yoga for low back pain. Clin J Pain 2013;29(5):450-460. [doi: 10.1097/ajp.0b013e31825e1492]
52. Bawa FL, Mercer SW, Atherton RJ, Clague F, Keen A, Scott NW, et al. Does mindfulness improve outcomes in patients with chronic pain? Systematic review and meta-analysis. Br J Gen Pract 2015 Jun;65(635):e387-e400 [FREE Full text] [doi: 10.3399/bjgp15X685297] [Medline: 26009534]
53. Veehof MM, Trompetter HR, Bohlmeijer ET, Schreurs KM. Acceptance- and mindfulness-based interventions for the treatment of chronic pain: a meta-analytic review. Cogn Behav Ther 2016 Jan 28;45(1):5-31. [doi: 10.1080/16506073.2015.1098724] [Medline: 26818413]
54. Pei JH, Ma T, Nan RL, Chen H, Zhang Y, Gou L, et al. Mindfulness-based cognitive therapy for treating chronic pain a systematic review and meta-analysis. Psychol Health Med 2021 Mar 26;26(3):333-346. [doi: 10.1080/13548506.2020.1849746] [Medline: 33241941]
55. Sharpe L, Richmond B, Menzies RE, Forrest D, Crombez G, Colagiuri B. A synthesis of meta-analyses of mindfulness-based interventions in pain. Pain 2024 Jan 01;165(1):18-28. [doi: 10.1097/j.pain.0000000000002997] [Medline: 37578535]
56. Petrucci G, Papalia GF, Russo F, Vadalà G, Piredda M, De Marinis MG, et al. Psychological approaches for the integrative care of chronic low back pain: a systematic review and metanalysis. Int J Environ Res Public Health 2021 Dec 22;19(1):60 [FREE Full text] [doi: 10.3390/ijerph19010060] [Medline: 35010319]
57. Lin TH, Tam KW, Yang YL, Liou TH, Hsu TH, Rau CL. Meditation-based therapy for chronic low back pain management: a systematic review and meta-analysis of randomized controlled trials. Pain Med 2022 Sep 30;23(10):1800-1811. [doi: 10.1093/pm/pnac037] [Medline: 35226059]
58. Soares LO, Ferreira GE, Costa LO, Nogueira LC, Meziat-Filho N, Reis FJ. Meditation for adults with non-specific low back pain: a systematic review and meta-analysis. Scand J Pain 2022 Jan 27;22(1):26-39 [FREE Full text] [doi: 10.1515/sjpain-2021-0096] [Medline: 34516731]
59. Soundararajan K, Prem V, Kishen TJ. The effectiveness of mindfulness-based stress reduction intervention on physical function in individuals with chronic low back pain: systematic review and meta-analysis of randomized controlled trials. Complement Ther Clin Pract 2022 Nov;49:101623. [doi: 10.1016/j.ctcp.2022.101623] [Medline: 35779457]
60. Schmidt H, Pilat C. Effects of meditation on pain intensity, physical function, quality of life and depression in adults with low back pain - a systematic review with meta-analysis. Complement Ther Med 2023 Mar;72:102924 [FREE Full text] [doi: 10.1016/j.ctim.2023.102924] [Medline: 36709927]
61. Paschali M, Lazaridou A, Sadora J, Papianou L, Garland EL, Zgierska AE, et al. Mindfulness-based interventions for chronic low back pain: a systematic review and meta-analysis. Clin J Pain 2024 Feb 01;40(2):105-113. [doi: 10.1097/AJP.0000000000001173] [Medline: 37942696]
62. Furlan AD, Giraldo M, Baskwill A, Irvin E, Imamura M. Massage for low-back pain. Cochrane Database Syst Rev 2015 Sep 01(9):CD001929. [doi: 10.1002/14651858.CD001929.pub3] [Medline: 26329399]
63. Garland EL, Brintz CE, Hanley AW, Roseen EJ, Atchley RM, Gaylord SA, et al. Mind-body therapies for opioid-treated pain: a systematic review and meta-analysis. JAMA Intern Med 2020 Jan 01;180(1):91-105 [FREE Full text] [doi: 10.1001/jamainternmed.2019.4917] [Medline: 31682676]
64. Goldberg SB, Riordan KM, Sun S, Kearney DJ, Simpson TL. Efficacy and acceptability of mindfulness-based interventions for military veterans: a systematic review and meta-analysis. J Psychosom Res 2020 Nov;138:110232 [FREE Full text] [doi: 10.1016/j.jpsychores.2020.110232] [Medline: 32906008]
65. Sud A, Lai KS, Cheng DK, Chung C, Pico-Espinosa OJ, Rice DB. Mind-body interventions for depressive symptoms in chronic pain: a systematic review of meta-analyses. Pain Physician 2021 Jan;24(1):61-72 [FREE Full text] [Medline: 33400428]
66. Kaplan J, Somohano VC, Zaccari B, O'Neil ME. Randomized controlled trials of mind-body interventions for posttraumatic stress disorder: a systematic review. Front Psychol 2023 Jan 24;14:1219296 [FREE Full text] [doi: 10.3389/fpsyg.2023.1219296] [Medline: 38327501]

XSL•FO
RenderX

67. Zhu L, Li L, Li X, Wang L. Mind-Body exercises for PTSD symptoms, depression, and anxiety in patients with PTSD: a systematic review and meta-analysis. Front Psychol 2021 Jan 18;12:738211 [FREE Full text] [doi: 10.3389/fpsyg.2021.738211] [Medline: 35153889]
68. Niles BL, Mori DL, Polizzi C, Pless Kaiser A, Weinstein ES, Gershkovich M, et al. A systematic review of randomized trials of mind-body interventions for PTSD. J Clin Psychol 2018 Sep 10;74(9):1485-1508 [FREE Full text] [doi: 10.1002/jclp.22634] [Medline: 29745422]
69. Tan L, Strudwick J, Deady M, Bryant R, Harvey SB. Mind-body exercise interventions for prevention of post-traumatic stress disorder in trauma-exposed populations: a systematic review and meta-analysis. BMJ Open 2023 Jul 12;13(7):e064758 [FREE Full text] [doi: 10.1136/bmjopen-2022-064758] [Medline: 37438059]
70. Hopwood TL, Schutte NS. A meta-analytic investigation of the impact of mindfulness-based interventions on post traumatic stress. Clin Psychol Rev 2017 Nov;57:12-20. [doi: 10.1016/j.cpr.2017.08.002] [Medline: 28806536]
71. Liu Q, Zhu J, Zhang W. The efficacy of mindfulness-based stress reduction intervention 3 for post-traumatic stress disorder (PTSD) symptoms in patients with PTSD: a meta-analysis of four randomized controlled trials. Stress Health 2022 Oct 09;38(4):626-636. [doi: 10.1002/smi.3138] [Medline: 35253353]
72. Goldstein E, McDonnell C, Atchley R, Dorado K, Bedford C, Brown RL, et al. The impact of psychological interventions on posttraumatic stress disorder and pain symptoms: a systematic review and meta-analysis. Clin J Pain 2019 Aug;35(8):703-712. [doi: 10.1097/AJP.0000000000000730] [Medline: 31145146]
73. Hilton L, Maher AR, Colaiaco B, Apaydin E, Sorbero ME, Booth M, et al. Meditation for posttraumatic stress: systematic review and meta-analysis. Psychol Trauma 2017 Jul;9(4):453-460. [doi: 10.1037/tra0000180] [Medline: 27537781]
74. Gallegos AM, Crean HF, Pigeon WR, Heffner KL. Meditation and yoga for posttraumatic stress disorder: a meta-analytic review of randomized controlled trials. Clin Psychol Rev 2017 Dec;58:115-124 [FREE Full text] [doi: 10.1016/j.cpr.2017.10.004] [Medline: 29100863]
75. Laplaud N, Perrochon A, Gallou-Guyot M, Moens M, Goudman L, David R, et al. Management of post-traumatic stress disorder symptoms by yoga: an overview. BMC Complement Med Ther 2023 Jul 21;23(1):258 [FREE Full text] [doi: 10.1186/s12906-023-04074-w] [Medline: 37480017]
76. Wahbeh H, Senders A, Neuendorf R, Cayton J. Complementary and alternative medicine for posttraumatic stress disorder symptoms: a systematic review. Focus (Am Psychiatr Publ) 2018 Jan;16(1):98-112 [FREE Full text] [doi: 10.1176/appi.focus.16102] [Medline: 32015705]
77. Strauss JL, Coeytaux R, McDuffie J, Nagi A, Williams JW. Efficacy of complementary and alternative medicine therapies for posttraumatic stress disorder. Department of Veterans Affairs (US). URL: https://www.ncbi.nlm.nih.gov/books/NBK82774/ [accessed 2024-04-29]
78. Sun LN, Gu JW, Huang LJ, Shang Z, Zhou Y, Wu L, et al. Military-related posttraumatic stress disorder and mindfulness meditation: a systematic review and meta-analysis. Chin J Traumatol 2021 Jul;24(4):221-230 [FREE Full text] [doi: 10.1016/j.cjtee.2021.05.003] [Medline: 34099359]
79. Haider T, Dai CL, Sharma M. Efficacy of meditation-based interventions on post-traumatic stress disorder (PTSD) among veterans: a narrative review. Adv Mind Body Med 2021;35(1):16-24. [Medline: 33513582]
80. Cushing RE, Braun KL. Mind-body therapy for military veterans with post-traumatic stress disorder: a systematic review. J Altern Complement Med 2018 Feb;24(2):106-114. [doi: 10.1089/acm.2017.0176] [Medline: 28880607]
81. Haun JN, Alman AC, Jean-Baptiste E, Melillo C, McMahon-Grenz J, Paykel JM. Delivery of complementary and integrative health using virtual health resources: a scoping review. J Integr Complement Med 2022 Nov 01;28(11):851-861 [FREE Full text] [doi: 10.1089/jicm.2021.0458] [Medline: 35819410]
82. Kozak L, Vig E, Simons C, Eugenio E, Collinge W, Chapko M. A feasibility study of caregiver-provided massage as supportive care for veterans with cancer. J Support Oncol 2013 Sep;11(3):133-143. [doi: 10.12788/j.suponc.0008] [Medline: 24400393]
83. Winger JG, Rand KL, Hanna N, Jalal SI, Einhorn LH, Birdas TJ, et al. Coping skills practice and symptom change: a secondary analysis of a pilot telephone symptom management intervention for lung cancer patients and their family caregivers. J Pain Symptom Manage 2018 May;55(5):1341-9.e4 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.01.005] [Medline: 29366911]
84. Harris M, Van Houtven C, Hastings S. Development of a home-based stress management toolkit for dementia caring dyads: protocol for a pilot intervention development and feasibility study. JMIR Res Protoc 2022 Dec 14;11(12):e43098 [FREE Full text] [doi: 10.2196/43098] [Medline: 36399169]
85. McDonnell KK, Andrews JO, Yang CH, Newsome BR, Weinkle E, Davis JE, et al. Study protocol for the Breathe Easier trial: a pilot RCT of a dyad-based, multiple-behavior intervention for improving physical and emotional health in survivors facing lung cancer. Integr Cancer Ther 2023 Nov 27;22:15347354231212876 [FREE Full text] [doi: 10.1177/15347354231212876] [Medline: 38009546]
86. Collinge W, Kahn J, Soltysik R. Promoting reintegration of national guard veterans and their partners using a self-directed program of integrative therapies: a pilot study. Mil Med 2012 Dec;177(12):1477-1485 [FREE Full text] [doi: 10.7205/milmed-d-12-00121] [Medline: 23397692]

XSL•FO
RenderX

87. Kahn JR, Collinge W, Soltysik R. Post-9/11 veterans and their partners improve mental health outcomes with a self-directed mobile and web-based wellness training program: a randomized controlled trial. J Med Internet Res 2016 Sep 27;18(9):e255 [FREE Full text] [doi: 10.2196/jmir.5800] [Medline: 27678169]
88. Haun JN, Venkatachalam HH, Fowler CA, Alman AC, Ballistrea LM, Schneider T, et al. Mobile and web-based partnered intervention to improve remote access to pain and posttraumatic stress disorder symptom management: recruitment and attrition in a randomized controlled trial. J Med Internet Res 2023 Oct 03;25:e49678 [FREE Full text] [doi: 10.2196/49678] [Medline: 37788078]
89. Eysenbach G, CONSORT- E. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
90. Strauss. Home page. Qualtrics. URL: https://www.qualtrics.com [accessed 2024-04-29]
91. Corrigan JD, Bogner J. Initial reliability and validity of the Ohio State University TBI Identification Method. J Head Trauma Rehabil 2007;22(6):318-329. [doi: 10.1097/01.HTR.0000300227.67748.77] [Medline: 18025964]
92. Bogner J, Corrigan JD. Reliability and predictive validity of the Ohio State University TBI identification method with prisoners. J Head Trauma Rehabil 2009;24(4):279-291. [doi: 10.1097/HTR.0b013e3181a66356] [Medline: 19625867]
93. Clark ME, Gironda RJ, Young RW. Development and validation of the Pain Outcomes Questionnaire-VA. J Rehabil Res Dev 2003;40(5):381-395 [FREE Full text] [doi: 10.1682/jrrd.2003.09.0381] [Medline: 15080223]
94. Buckenmaier CC, Galloway KT, Polomano RC, McDuffie M, Kwon N, Gallagher RM. Preliminary validation of the Defense and Veterans Pain Rating Scale (DVPRS) in a military population. Pain Med 2013 Jan 01;14(1):110-123. [doi: 10.1111/j.1526-4637.2012.01516.x] [Medline: 23137169]
95. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. J Trauma Stress 2015 Dec;28(6):489-498. [doi: 10.1002/jts.22059] [Medline: 26606250]
96. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation; 1996.
97. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
98. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
99. Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996 Mar;34(3):220-233. [doi: 10.1097/00005650-199603000-00003] [Medline: 8628042]
100. Busby DM, Christensen C, Crane DR, Larson JH. A revision of the dyadic adjustment scale for use with distressed and nondistressed couples: construct hierarchy and multidimensional scales. J Marital Family Therapy 2007 Jun 08;21(3):289-308. [doi: 10.1111/j.1752-0606.1995.tb00163.x]
101. Neff KD. The development and validation of a scale to measure self-compassion. Self Identity 2003 Jul;2(3):223-250. [doi: 10.1080/15298860309027]
102. Sprecher S, Fehr B. Compassionate love for close others and humanity. J Soc Pers Relat 2005 Oct 01;22(5):629-651. [doi: 10.1177/0265407505056439]
103. Engel GL. The clinical application of the biopsychosocial model. Am J Psychiatry 1980 May;137(5):535-544. [doi: 10.1176/ajp.137.5.535] [Medline: 7369396]
104. Gatchel RJ, Peng YB, Peters ML, Fuchs PN, Turk DC. The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychol Bull 2007 Jul;133(4):581-624. [doi: 10.1037/0033-2909.133.4.581] [Medline: 17592957]
105. Strauss JL. SAS® 9.4 — today and tomorrow. SAS. URL: https://support.sas.com/software/94/ [accessed 2024-04-29]
106. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Soft 2011;45(3):1-67. [doi: 10.18637/jss.v045.i03]
107. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org [accessed 2024-04-29]
108. Ong AD, Thoemmes F, Ratner K, Ghezzi-Kopel K, Reid MC. Positive affect and chronic pain: a preregistered systematic review and meta-analysis. Pain 2020 Jun 10;161(6):1140-1149 [FREE Full text] [doi: 10.1097/j.pain.0000000000001828] [Medline: 32049895]
109. Master your research projects with the power of AI. ATLAS.ti. URL: https://atlasti.com/ [accessed 2024-04-29]
110. Miles MB, Huberman AM. Qualitative Data Analysis: An Expanded Sourcebook. New York, NY: Sage Publication; 1994.
111. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922. [doi: 10.1177/1609406919899220]
112. Huberman AM, Miles MB. The Qualitative Researcher's Companion. Thousand Oaks, CA: Sage Publications; 2002.
113. Fritz JM, Davis AF, Burgess DJ, Coleman B, Cook C, Farrokhi S, et al. Pivoting to virtual delivery for managing chronic pain with nonpharmacological treatments: implications for pragmatic research. Pain 2021 Jun 01;162(6):1591-1596 [FREE Full text] [doi: 10.1097/j.pain.0000000000002139] [Medline: 33156148]

XSL•FO
RenderX

114. Tick H, Nielsen A, Pelletier KR, Bonakdar R, Simmons S, Glick R, Pain Task Force of the Academic Consortium for Integrative Medicine and Health. Evidence-based nonpharmacologic strategies for comprehensive pain care: the consortium pain task force white paper. Explore (NY) 2018;14(3):177-211 [FREE Full text] [doi: 10.1016/j.explore.2018.02.001] [Medline: 29735382]
115. Taylor SL, Elwy AR, Bokhour BG, Coggeshall SS, Cohen A, Der-Martirosian C, et al. Measuring patient-reported use and outcomes from complementary and integrative health therapies: development of the complementary and integrative health therapy patient experience survey. Glob Adv Integr Med Health 2024 Apr 06;13:27536130241241259 [FREE Full text] [doi: 10.1177/27536130241241259] [Medline: 38585239]
116. Bokhour BG, Haun JN, Hyde J, Charns M, Kligler B. Transforming the veterans affairs to a whole health system of care: time for action and research. Med Care 2020 Apr;58(4):295-300. [doi: 10.1097/MLR.0000000000001316] [Medline: 32044866]
117. Kligler B, Hyde J, Gantt C, Bokhour B. The whole health transformation at the veterans health administration: moving from "what's the matter with you?" to "what matters to you?". Med Care 2022 May 01;60(5):387-391. [doi: 10.1097/MLR.0000000000001706] [Medline: 35283434]
118. Meerwijk EL, Larson MJ, Schmidt EM, Adams RS, Bauer MR, Ritter GA, et al. Nonpharmacological treatment of army service members with chronic pain is associated with fewer adverse outcomes after transition to the veterans health administration. J Gen Intern Med 2020 Mar 28;35(3):775-783 [FREE Full text] [doi: 10.1007/s11606-019-05450-4] [Medline: 31659663]
119. Mannes ZL, Stohl M, Fink DS, Olfson M, Keyes KM, Martins SS, et al. Non-pharmacological treatment for chronic pain in US veterans treated within the veterans health administration: implications for expansion in US healthcare systems. J Gen Intern Med 2022 Nov 19;37(15):3937-3946 [FREE Full text] [doi: 10.1007/s11606-021-07370-8] [Medline: 35048300]
120. Midboe AM, Javier SJ, Salsbury SA, Katsovich L, Burgess DJ, King HA, et al. Impact of COVID-19 pandemic on nonpharmacological pain management trials in military and veteran healthcare settings: an evaluation informed by implementation science. Transl Behav Med 2023 Aug 11;13(8):601-611 [FREE Full text] [doi: 10.1093/tbm/ibad015] [Medline: 37030012]
121. Reed DE, Engel CC, DeFaccio R, Gaj L, Douglas JH, Williams RM, et al. Examining the Veterans Health Administration whole health model of care within the context of posttraumatic stress disorder. Psychol Serv 2024 May;21(2):224-234. [doi: 10.1037/ser0000822] [Medline: 38079475]
122. Kerns RD, Brandt CA. NIH-DOD-VA pain management collaboratory: pragmatic clinical trials of nonpharmacological approaches for management of pain and co-occurring conditions in veteran and military health systems: introduction. Pain Med 2020 Dec 12;21(Suppl 2):S1-S4. [doi: 10.1093/pm/pnaa358] [Medline: 33313722]
123. Becker WC, DeBar LL, Heapy AA, Higgins D, Krein SL, Lisi A, et al. A research agenda for advancing non-pharmacological management of chronic musculoskeletal pain: findings from a VHA state-of-the-art conference. J Gen Intern Med 2018 May 9;33(Suppl 1):11-15 [FREE Full text] [doi: 10.1007/s11606-018-4345-6] [Medline: 29633136]
124. Rehabilitation act. U.S. Access Board. URL: https://www.access-board.gov/about/law/ra.html [accessed 2024-04-29]

Abbreviations

BDI-II: Beck Depression Inventory–II
CIH: complementary and integrative health
CLS: Compassionate Love Scale
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DVPRS: Defense and Veterans Pain Rating Scale
HRQoL: health-related quality of life
MR: Mission Reconnect
POQ-VA: Pain Outcomes Questionnaire–For Veterans
PSQI: Pittsburgh Sleep Quality Index
PSS: Perceived Stress Scale
PST: pain, stress, and tension scale
PTSD: posttraumatic stress disorder
QOL: quality of life
RDAS: Revised Dyadic Adjustment Scale
SCS: Self-Compassion Scale
SF-12: 12-item Short Form Health Survey
TBI: traumatic brain injury
VA: Department of Veterans Affairs
VHA: Veterans Health Administration

XSL•FO
RenderX

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 13.02.24; peer-reviewed by V Ameral, R Nooripour; comments to author 29.05.24; revised version received 19.06.24; accepted 31.08.24; published 18.10.24.

Please cite as:
Haun JN, Fowler CA, Venkatachalam HH, Alman AC, Ballistrea LM, Schneider T, Benzinger RC, Melillo C, Alexander NB, Klanchar SA, Lapcevic WA, Bair MJ, Taylor SL, Murphy JL, French DD
Outcomes of a Remotely Delivered Complementary and Integrative Health Partnered Intervention to Improve Chronic Pain and Posttraumatic Stress Disorder Symptoms: Randomized Controlled Trial
J Med Internet Res 2024;26:e57322
URL: https://www.jmir.org/2024/1/e57322
doi:10.2196/57322
PMID:39422992

©Jolie N Haun, Christopher A Fowler, Hari H Venkatachalam, Amy C Alman, Lisa M Ballistrea, Tali Schneider, Rachel C Benzinger, Christine Melillo, Neil B Alexander, S Angel Klanchar, William A Lapcevic, Matthew J Bair, Stephanie L Taylor, Jennifer L Murphy, Dustin D French. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study

Angela Karellis[1], PhD; Duncan Webster[2], MD; Jean Boulanger[3], RN; Kate Harland[2], RN; Paige Feltmate[2], RN; Stefanie Materniak[2], MSc; Gabriel Daunais-Laurin[3], BA; Christine Mesa[4], MSc; Olivia Vaikla[1], BSc; John Kim[4], PhD; Nitika Pant Pai[1,5], MPH, MD, PhD

[1]Infectious Diseases and Immunity in Global Health, Research Institute of the McGill University Health Centre, Montreal, QC, Canada

[2]Centre for Research, Education & Clinical Care of At-Risk Populations, Saint John, NB, Canada

[3]REZO, Montreal, QC, Canada

[4]National Laboratory for HIV Reference Services, Public Health Agency of Canada, Winnipeg, MB, Canada

[5]Faculty of Medicine, McGill University, Montreal, QC, Canada

Corresponding Author:
Nitika Pant Pai, MPH, MD, PhD
Infectious Diseases and Immunity in Global Health
Research Institute of the McGill University Health Centre
5252 boul de Maisonneuve Ouest
Montreal, QC, H4A3S5
Canada
Phone: 1 514 934 1934 ext 44729
Email: nitika.pai@mcgill.ca

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e67994

Abstract

Background: With the need to expedite the detection of multiple sexually transmitted blood-borne infections (STBBIs), there is an increased demand for digital innovations and tests that improve the efficiency of point-of-care testing in outreach community settings. Multiplexed testing is being offered to at-risk populations by frontline health care professionals.

Objective: With this in mind, we evaluated AideSmart! (for health aides or health care workers), an integrated app and platform solution that enables multiplexed STBBI screening at the point of care, for feasibility, preference, accuracy, and impact. With AideSmart!, we provided trained health care workers with the ability to perform rapid multiplexed tests; offer STBBI pre- or posttest counseling; maintain quality assurance of testing; facilitate linkages to care; and enhance health care provider–patient communication, data documentation, and patient engagement through the multiplexed STBBI screening and linkage process. We evaluated the performance of multiplexed tests.

Methods: In a cross-sectional study conducted during the COVID-19 pandemic, we recruited at-risk populations in Canada from community-based organizations in Montreal, Quebec, and Saint John, New Brunswick, with unknown serostatus for HIV, hepatitis C virus (HCV), and syphilis. Following orientation and pretest counseling with AideSmart!, we performed multiplexed tests, simultaneously screening for HIV, HCV, and syphilis, using 2 rapid investigational multiplexed tests (Chembio for HIV and syphilis and MedMira for HIV, HCV, and syphilis) followed by confirmatory testing from local and national laboratories.

Results: Of the 401 participants, all (100%) accepted the AideSmart! multiplexed strategy: 59.4% (148/249) preferred multiplexed rapid tests over laboratory tests, and 56.6% (141/249) preferred receiving same-day test results. Rapid test results were obtained in 15 minutes (vs laboratory tests at 4-24 days). A total of 29 new infections (n=27, 93% HCV; n=1, 3% syphilis; and n=1, 3% HIV) were identified and treated within a week. Feasibility of the strategy (completion of testing and linkages to care) was at

XSL•FO
RenderX

76.1% (305/401). Health care professionals provided positive feedback and emphasized the importance of knowing one's negative or positive serostatus, especially during a pandemic. Multiplexed rapid tests' specificity (against laboratory reference standards) exceeded 98% (98.7%-100%) for all pathogens and devices. An electronic reader, used by the Chembio rapid test, enhanced sensitivity (HIV: 100%, 95% CI 79.4%-100%; syphilis: 86.8%, 95% CI 71.9%-95.6% [Chembio] vs HIV: 100%, 95% CI 78.2%-100%; HCV: 90.3%, 95% CI 80.1%-96.4%; and syphilis: 57.9%, 95% CI 40.8%-73.7% [MedMira]).

Conclusions: The AideSmart! digital multiplexed rapid screening strategy for health care workers facilitated STBBI testing for multiple STBBIs and arranged for pre- or posttest counseling and rapid linkages with high feasibility and acceptability. Electronic readers enhanced the diagnostic performance of multiplexed biomarker tests. This study generated data in support of digital multiplexed strategies in digitally enabled settings for at-risk populations nationally and worldwide.

(J Med Internet Res 2024;26:e55075) doi:10.2196/55075

KEYWORDS

sexually transmitted blood-borne infections; rapid multiplexed biomarker-based screening; digital strategy; feasibility; preference; impact

Introduction

Background

In Canada, at-risk populations are disproportionately impacted by sexually transmitted blood-borne infections (STBBIs). These populations include men who have sex with men; Indigenous, immigrant, and refugee populations; and people who inject drugs, among others [1-3]. The standard of care with regard to STBBI diagnostics entails laboratory-based testing, sometimes requiring invasive specimen collection, and providing results at a subsequent clinic visit several days later based on geographical location. Location and availability of tests impact the turnaround time to obtain test results, which affects timely counseling and initiation of treatment and care [4,5].

Challenges in STBBI Diagnostics and Care for At-Risk Populations

Following the COVID-19 pandemic and the provision of self-tests that led to an improved understanding of testing and screening services, at-risk populations desire convenient and timely access to health care services for STBBIs, including demands for rapid testing, confirmation, counselling, and initiation of treatment in 1 to 2 visits, if possible [6]. However, the current status quo, with health care worker shortages and demands on professional time, limits patient access to sought services, namely due to the lack of dedicated family physicians, translating to many missed opportunities in STBBI screening [7-9].

In addition, marginalization of at-risk underserved populations, together with stigma, discrimination, and systemic racism, impacts health care seeking and timely screening for these important STBBIs despite availability of testing, further perpetuating health care inequity [10].

In parallel, in current times, in the context of acute health care professional shortages nationwide, populations demand the use of innovative patient-centered strategies that improve efficiency and costs of service delivery [11-13]. Mobile, rapid, connected digital screening programs fill health service delivery gaps in STBBI care, including efficient service delivery, data documentation, and linkage arrangement when screened using point-of-care (POC) technology-based services [14].

Innovative Strategies for Improving STBBI Care

As mentioned in the Lancet Commission on Diagnostics, digitization can improve access to screening and linkage and retention pathways for underserved populations [15]. Together with increasing access to timely diagnostics, the commission has called for an investment in the training of health care professionals using innovative solutions. The proposed solution is an investment to fill gaps in training, data documentation, communication, and linkages using POC technologies that will improve service delivery. To reach global targets, evidence to support the use of innovative and smarter connected screening solutions across different contexts and populations is much needed, for example, a study in a high-income context [6].

Since 2009, our laboratory has been at the forefront of creating digital process innovations in this space of POC diagnostics. We have developed screening digital solutions for HIV and hepatitis C virus (HCV) self-testing, multiplexed screening for STBBIs, COVID-19 self-testing, and sexual and reproductive health services. Innovative digital strategies can not only improve rapid STBBI screening and linkage of testers but also help in training, counselling, and data management by health care workers while optimizing the screening pathway. Digital strategies also address the gaps in data and increase efficiency by streamlining workflows thus enabling task shifting, all while providing data to guide health care service delivery [16].

The AideSmart! app- and platform-based solution integrated with multiplexed POC technologies, called the AideSmart! multiplexed strategy, is designed for health aides (frontline health care workers) and is one such unique solution.

The AideSmart! strategy was first pilot-tested in India in 2018 in pregnant women in a Grand Challenges Canada–funded study. In that study, the strategy was deemed successful in detecting 239 new infections or conditions (HIV, hepatitis B or C, and syphilis) and helped in reducing vertical transmission to infants [17]. While the pilot study showed the effectiveness of the AideSmart! strategy in rural India, a middle-income country, we wished to examine its feasibility in Canada, a high-income country with vastly different at-risk subpopulations with varying prevalence of STBBIs.

XSL•FO
RenderX

In this Canadian Institutes of Health Research–funded study, we adapted the innovative solution to meet the needs of Canadian vulnerable populations (people who inject drugs; sexually transmitted disease [STD] clinic attendees; lesbian, gay, bisexual, transgender, and queer individuals; and refugee and immigrant populations). We localized the digital platform into French, tailored the content to align with Canadian guidelines, and integrated the newest versions of multiplexed assays to offer a biomarker-based strategy to populations.

With a focus on global portability and uptake and the aim of introducing scalable solutions that address data and service delivery gaps using POC technologies for STBBI screening, counselling, and linkage to care, we sought to evaluate AideSmart! within a high-income setting using multiplexed strategies that are slowly changing the landscape of delivery of STBBI diagnostics at the POC across global settings.

Methods

Study Design and Objectives

In this cross-sectional study, our primary aim was to evaluate the feasibility, preference, and acceptability of the AideSmart! strategy among at-risk populations in Canada. Our secondary aim was to estimate the accuracy of the 2 investigational rapid multiplexed devices and demonstrate their impact on detection of new infections.

Ethical Considerations

Ethics approval was obtained from the Research Institute of the McGill University Health Centre (2020-6048) and from the Horizon Health Network Research Ethics Board at the Saint John Regional Hospital (2020-2963). An informed consent form (ICF) was presented to each participant before study entry in their preferred language (English or French). The ICF described the ability to withdraw consent or opt out during the study in the event that participants no longer wished to continue the study. Participants received financial compensation during the study of CAD $25 (US $18.51) for completing visits 1 and 2; an extra CAD $10 (US $7.41) was provided if a third visit was required and completed.

Furthermore, to ensure data privacy and security, all data recorded on the AideSmart! app were encrypted (3-layer encoded) and protected on a secure Health Insurance Portability and Accountability Act (HIPAA)–compliant server. A keyed access to the server was provided to the principal investigator of the study. A study ID was created to store data. It contained a site-specific code and a randomly generated participant ID. No names were revealed to the data analysts and the staff conducting the study at any point of the study. This anonymized, deidentified data protection process guaranteed participant confidentiality and has been successfully deployed in our national and global studies. The master list that linked names with study IDs was available only to the clinicians responsible for patient care and was safeguarded for the entire duration of the study.

Study Population

Eligible recruited participants were adults with unknown serostatus for HIV, HCV, and syphilis, including but not limited to (1) people who inject drugs, (2) men who have sex with men, and (3) immigrant populations, in Quebec and New Brunswick. Individuals in treatment for any coinfection or with any urgent medical condition requiring hospitalization were excluded.

Participants were recruited from 2 community-based service delivery sites in Quebec and New Brunswick: RÉZO, Montreal, Quebec (May 2021 to February 2022), and the Centre for Research, Education, and Clinical Care of At-Risk Populations (RECAP), Saint John, New Brunswick (March 2021 to August 2022). Convenience sampling was deployed.

Recruitment was managed by the community-based organizations. Social media platforms (Facebook and Grindr) were used to spread the word about the study. Posters were placed in community clinics to increase awareness of the study, and flyers and handouts were printed that explained the purpose of the study (approved by the Research Ethics Board) and shared with clinic attendees and in local magazines. Health care staff participated in community outreach events in the area to increase word of mouth regarding the study, leveraging their existing community-based networks. Enrolled participants were offered to partake in a “refer a friend” program to invite additional persons to participate. Existing relationships with communities were thus leveraged to increase recruitment.

RÉZO is a community organization active since 1991 that offers gay and bisexual cis- and transgender men various free health and wellness promotion programs. RÉZO provides HIV or HCV and STBBI prevention and testing [18].

RECAP is a community-based harm reduction clinic dedicated to improving prevention, diagnosis, and treatment of HCV in at-risk populations, many with substance use disorders. This nonprofit organization, active since 2014, provides opioid agonist therapy, harm reduction counseling, safe supplies, and clinic- and outreach-based screening for STBBIs, among other services [19].

AideSmart! STBBI Rapid Multiplexed Screening Strategy

The user-friendly AideSmart! app’s main page directs health aides to several tabs offering training on multiplexed tests, counselling, and recommendations for all STBBIs for both health care workers and testers. The Overview tab is shown in Figure 1, which presents the overall aim of the AideSmart! app and participant eligibility for study inclusion. The app also provides education and screening and testing information related to rapid multiplexed tests and arranges for communication with patients and health care providers while arranging for rapid linkage to confirmatory testing, treatment, and clinical care. As shown in Figure 2, the AideSmart! app was designed to streamline and simplify STBBI diagnostic care, document and store digital data, and link participants to care, all while maintaining high patient engagement and communication with various stakeholders at all times. The app- or web-based strategy is open access, portable, and customizable to both technologies: multiplexed platforms and rapid biomarker-based multiplexed

XSL•FO
RenderX

assays. It is currently being adapted and tested using multiplexed molecular platforms in an India-Canada Centre for Innovative Multidisciplinary Partnerships to Accelerate Community Transformation and Sustainability–funded study in India.

The AideSmart! platform and app solution (open access, McGill University; copyright Report of Invention 16126, 2016) was integrated with 2 investigational lateral flow assays: Multiplo Rapid *Treponema pallidum* Antibody Test (MedMira Inc) and Dual Path Platform HIV-syphilis assay (Chembio Diagnostic Systems, Inc) [17]. Of note, the Chembio test for HIV or syphilis was operated using a quantitative reader that removed ambiguity associated with a visual test result interpretation, whereas the MedMira test relied on visual interpretation of test results.

In this study, these tests were used to independently evaluate accuracy and evaluate some prototypes of the rapid tests. Both rapid index tests required health care professional–collected finger stick blood–based samples. Confirmatory testing was performed using both the local laboratory and dried blood spot (DBS) tests as reference standard tests on each participant. Local laboratory testing per standardized algorithms (Textbox 1) was conducted in Quebec and New Brunswick. DBS specimens were sent for analysis at the National Laboratory for HIV Reference Services in Winnipeg, Manitoba.

The AideSmart! app-based program connected stakeholders throughout the study. The digital app-based platform detailed study procedures, training on rapid testing procedures, pretest and posttest combined counseling for STBBIs, clinical care and coordination for those who receive a positive test result, treatment, and follow-up. For counseling and treatment, Canada-specific guidelines for STBBIs were included in the app-based program to provide targeted information to participants [20].

Figure 1. Screenshot of the AideSmart! app overview page.

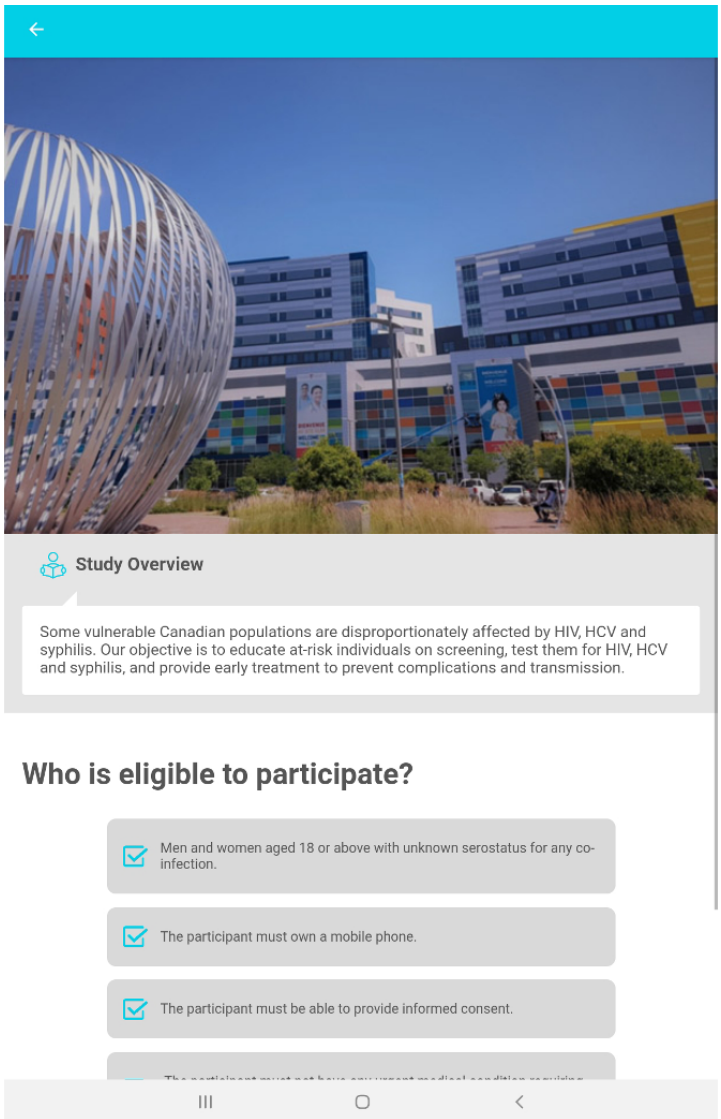

XSL•FO
RenderX

Figure 2. AideSmart! rapid multiplexed testing strategy to simultaneously screen for populations at risk of contracting sexually transmitted blood-borne infections in Canada.

Textbox 1. Local laboratory screening algorithms in Canada for HIV, syphilis, and hepatitis C virus (HCV).

HIV

- Geenius HIV 1/2 test (Quebec) or Architect HIV 1/2 Ag/Ab Combo assay (New Brunswick)
- If reactive: confirmatory p24 enzyme immunoassay (EIA)
- If reactive: specimen sent to the National Microbiology Laboratory for final analysis/confirmation

Syphilis

- Enzyme immunoassay (EIA)
- If reactive: Rapid plasma regain (RPR) test
- If needed, *Treponema pallidum* particle agglutination (TP-PA) and INNO-LIA tests

HCV

- Anti-HCV screen
- If reactive: qualitative RNA test
- If reactive, viral load measured
- Note: If either test yielded non-reactive results but the test was conducted in the window period following suspected HCV exposure, follow-up test was recommended three months later

Study Procedures and Intervention

We recruited throughout the COVID-19 pandemic, and communication was facilitated through SMS text messages and calls enabled by our AideSmart! app strategy.

At visit 1, following an explanation of the study objectives and study procedures by the health care professionals, consenting participants were enrolled in the study. Using the AideSmart! innovation, health care professionals conducted pretest counseling aiming to educate participants regarding STBBIs, rapid and laboratory-based testing methods, and posttest linkage and treatment options for each STBBI.

Participants had the opportunity to ask questions as needed. Next, rapid MedMira and Chembio tests were performed per manufacturer instructions. As both MedMira and Chembio were considered investigational devices (not approved in Canada yet), test results were declared following receipt of local laboratory test results, although turnaround time to obtain rapid test results was recorded in real time. Health care professionals documented deidentified data on participants' demographics, risk profile, and rapid test results.

XSL•FO
RenderX

Between visits 1 and 2, health care professionals and participants communicated using participants' preferred method of communication. Health care professionals sent SMS text message reminders to confirm the date and time of the follow-up appointment, also allowing for sufficient time for confirmatory test results to be expedited from local laboratories. All participants returned for a second visit, during which both positive and negative test results were declared in a posttest counseling or harm reduction session aided by the AideSmart! app. Health care professionals documented rapid and confirmatory test completion and preference data. Participants with nonreactive test results were encouraged to seek repeat testing should future exposure be expected. In case of one or more reactive result or results, linkage to treatment was arranged, either in-house or at a nearby clinic.

For follow-up and retention in visit 3, health care professionals coordinated another visit for those who received a positive test result.

Feasibility

Feasibility was computed using the metric completion rate, defined as the number of participants who successfully completed the procedure that consisted of pretest counselling and testing (visit 1) and posttest counselling and linkage to treatment initiation (visit 2; numerator) divided by the total number of participants who consented to take part at baseline (denominator).

The primary end point was used for sample size calculation. On the basis of our previous evaluations, with an assumed completion rate of 70%, a sample size of 323 was deemed sufficient with 95% level of confidence and 5% precision [21].

Acceptability was defined as the number of participants who consented to the strategy (numerator) over the number of eligible participants who were approached (denominator).

Preference

Preference was defined as the number of participants who indicated a preference for a patient-reported outcome (including preference of timing to receive test results and retest preference) divided by the number of participants who consented to take part in the study (denominator). Satisfaction of POC and conventional testing was further recorded categorically (*very satisfied*, *satisfied*, *neutral*, *dissatisfied*, and *very dissatisfied*).

Health Care Provider Perspectives

Health care professional feedback was collected to gain perspective regarding their views on multiplexed rapid testing. The following questions were asked: "What do you think about multiplex tests?" "Do you think they will have a future in Canada?" "If approved, do you think they will be useful for populations they serve?" "Is it helpful for people to know if they test negative?"

Impact

Impact was defined as the number of newly diagnosed infections (numerator) relative to the number of patients who consented and were screened (denominator).

Impact was also recorded using the metric that captures turnaround time, defined as the mean time recorded in minutes or days for test results to be obtained so as to inform the next step in the screening pathway.

Diagnostic Accuracy

Diagnostic accuracy performance metrics (sensitivity, specificity, and 95% CIs) for each pathogen were tabulated for index tests (Chembio and MedMira) separately and compared to both reference standard tests (local laboratory algorithms and DBS).

Participants' sociodemographic profiles were tabulated.

Outcome data (feasibility, acceptability, preference, and impact) were calculated in proportions with 95% CIs. All analyses were performed in Stata (version 10; StataCorp) [22] and R (version 4.1.3; R Foundation for Statistical Computing) [23].

Results

Study Population

In total, we recruited 401 participants (n=237, 59.1% from RÉZO and n=164, 40.9% from RECAP).

Overall, 71.6% (282/394) of participants were men, 21.6% (85/394) were women, and 6.9% (27/394) identified as another gender. In terms of ethnicity, 5.6% (22/394) self-reported being Indigenous individuals, 5.8% (23/394) self-reported being Asian or Mediterranean individuals, 9.6% (38/394) self-reported being Latin American or Caribbean individuals, 5.3% (21/394) self-reported being African individuals, and 73.6% (290/394) self-reported being White individuals.

Most participants (165/394, 41.6%) were young (aged 18-34 years), 59.4% (234/394) were doing part-time work or were unemployed, and 65.5% (258/394) earned <CAD $2000 (US $1480.26) per month.

Although 86% (339/394) of the participants had completed high school or advanced degrees, nearly half (172/394, 43.7%) of the participants were unemployed (Table 1).

In the previous year, with respect to past testing history, nearly half of the participants had been tested for HIV (184/392, 46.9%), HCV (185/392, 47.2%), and syphilis (172/391, 43.9%). A subset of participants (169/385, 43.9%) reported a previous STD, the most common being chlamydia (72/169, 42.6%), syphilis (48/169, 28.4%), and gonorrhea (47/169, 27.8%).

Regarding risk profile, while most (288/394, 73.1%) reported being sexually active, many (247/394, 62.7%) reported no stable partners. Approximately half (159/379, 42%) reported 2 to 5 sexual partners in the previous 6 months, and half (167/370, 45.1%) reported occasional condom use. Approximately half had never been diagnosed with a previous STD (204/385, 53%), had never injected recreational drugs (250/392, 63.8%), or were not alcohol users (206/393, 52.4%). Among people who injected drugs, most (83/139, 59.7%) had shared needles in the past.

Men presented with a riskier baseline status. Of the 60 participants with at least 6 recent sexual partners, 47 (78%) were men, whereas 13 (22%) were women or identified as

XSL•FO
RenderX

another gender. Moreover, a higher proportion of men (122/272, 44.9%) reported a previous STD compared to women and those who identified as another gender (38/112, 33.9%).

In addition, the 43.9% (169/385) of participants who had a previous STD presented with riskier baseline sexual behaviors. For instance, 20.7% (35/169), 49.1% (83/169), and 20.7% (35/169) reported always, sometimes, and never using a condom, respectively. Moreover, 73.4% (124/169) had 0 to 5 recent sexual partners, and 24.9% (42/169) had ≥6 partners. Conversely, participants with no previous recorded STDs (216/385, 56.1%) had less risky behaviors, including a higher rate of protected sex (67/216, 31%) and up to 5 recent sexual partners (191/216, 88.4%). However, participants with a previous STD seemed more aware of this riskier behavior, as indicated by higher previous testing rates (14%-19% vs 7%-8%).

XSL•FO
RenderX

Table 1. Demographics and medical history of recruited study populations to the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (N=394).

Criteria	Participants, n (%)
Age (y)	
18-24	32 (8.1)
25-34	133 (33.8)
35-44	112 (28.4)
45-54	63 (16)
≥55	54 (13.7)
Gender	
Man	282 (71.6)
Woman	85 (21.6)
Another gender	16 (4.1)
No response	11 (2.8)
Descent	
African	21 (5.3)
Asian or South Asian	10 (2.5)
European or North American	290 (73.6)
Latin American or Caribbean	38 (9.6)
Mediterranean	13 (3.3)
North American Indigenous	22 (5.6)
Educational level	
Did not complete primary schooling	55 (14)
High school to preuniversity education	163 (41.4)
Technical degree	55 (14)
Postgraduate degree	121 (30.7)
Employment status	
Employed full time	130 (33)
Employed part time	62 (15.7)
Not employed	172 (43.7)
Retired	11 (2.8)
Not willing to work	19 (4.8)
Monthly income	
<CAD $2000 (US $1480.26)	258 (65.5)
CAD $2001-$4000 (US $1481-$2960.52)	105 (26.6)
CAD $4001-$6000 (US $2961.26-$4440.78)	22 (5.6)
CAD $6001-$8000 (US $4441.52-$5921.04)	7 (1.8)
≥CAD $8001 (US $5921.78)	2 (0.5)
Currently partnered	
Yes	147 (37.3)
No	247 (62.7)
Sexual activity	
Yes	288 (73.1)
No	88 (22.3)

XSL•FO
RenderX

Criteria	Participants, n (%)
I do not wish to answer	18 (4.6)
Condom use (n=370)	
Never	98 (26.5)
Always	105 (28.4)
Sometimes	167 (45.1)
Recent sexual partners (past 6 months; n=379)	
0	66 (17.4)
1	94 (24.8)
2-5	159 (42)
6-10	38 (10)
≥11	22 (5.8)
Previous STDs[a] (n=385)	
Yes	169 (43.9)
No	204 (53)
I do not know	12 (3.1)
Type of previous STD[b] (n=169)	
HIV	10 (6)
Herpes simplex	3 (1.8)
Chlamydia	72 (43.1)
Gonorrhea	47 (28.1)
Syphilis	48 (28.4)
Other	25 (15)
Previous recreational injection drug use (n=392)	
Yes	139 (35.5)
No	250 (63.8)
I do not wish to answer	3 (0.8)
If yes to previous recreational injection drug use, were needles shared? (n=139)	
Yes	83 (59.7)
No	56 (40.3)
Alcohol use (n=393)	
No	206 (52.4)
1-2 times per week	122 (31)
3-5 times per week	65 (16.5)
Previous HIV test (n=392)	
No	54 (13.8)
Yes, <6 months ago	137 (34.9)
Yes, 6 months to 1 year ago	47 (12)
Yes, >1 year ago	154 (39.3)
Previous HCV [c] test (n=392)	
No	41 (10.5)
Yes, <6 months ago	137 (34.9)
Yes, 6 months to 1 year ago	48 (12.2)

XSL•FO
RenderX

Criteria	Participants, n (%)
Yes, more than 1 year ago	166 (42.3)
Previous syphilis test (n=391)	
Yes, <6 months ago	125 (32)
Yes, 6 months to 1 year ago	47 (12)
Yes, more than 1 year ago	219 (56)

[a]STD: sexually transmitted disease.

[b]Patients may have reported >1 previous STD.

[c]HCV: hepatitis C virus.

Feasibility

All participants approached (401/401, 100%) accepted the AideSmart! strategy. Completion of AideSmart! over 2 visits was 76.1% (305/401).

The mean turnaround time to receive multiplexed rapid test results was <15 minutes (mean 13.9, SD 2.1 min for Chembio and mean 13.5, SD 2.3 min for MedMira). In comparison, mean turnaround time for local laboratory test results ranged from 4 days to 3 weeks (HIV p24, mean 4.4, SD 3.0 days; HCV RNA, mean 9.0, SD 6.2 days; HIV RNA, mean 11.2, SD 5.1 days; *T pallidum* particle agglutination, mean 12.4, SD 6.5 days; DBS, mean 24.2, SD 5.8 days).

Preference

Nearly all participants (243/247, 98.4%) reported high or very high satisfaction with the rapid testing and counseling through the AideSmart! strategy versus 95.8% (226/236) who reported satisfaction with the conventional laboratory-based strategy. Participants reported comparable fear, pain, and discomfort between the rapid and conventional strategies (Table 2).

Preference to obtain same-day test results using rapid testing was expressed by 56.6% (141/249) of the participants. Preference for follow-up through various digital health supports varied—the most popular response was app-based secure messaging (84/250, 33.6%), followed by face-to-face visits (60/250, 24%), phone calls (39/250, 15.6%), and SMS text messaging (34/250, 13.6%); 13.2% (33/250) had no preference. As a result, unsurprisingly, 93.6% (234/250) of the participants reported that they would recommend rapid testing to friends.

XSL•FO
RenderX

Table 2. Acceptability, feasibility, and preference results of the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (2021-2022).

Criteria	Participants, n (%)
Acceptability (n=401)	
Yes	401 (100)
Completion of pretest counseling, testing, posttest counseling, and linkage to treatment (N=401)	
Yes	305 (76.1)
No	96 (23.9)
Point-of-care testing	
Fear[a] (n=249)	
No fear (1)	161 (64.7)
2	53 (21.3)
3	19 (7.6)
4	9 (3.6)
Highest fear (5)	7 (2.8)
Pain[b] (n=248)	
No pain (1)	171 (69)
2	61 (24.6)
3	15 (6)
4	1 (0.4)
Discomfort[c] (n=248)	
No discomfort (1)	185 (74.6)
2	49 (19.8)
3	11 (4.4)
4	2 (0.8)
Highest discomfort (5)	1 (0.4)
Satisfaction (n=247)	
Very satisfied	213 (86.2)
Satisfied	30 (12.1)
Neutral or not sure	4 (1.6)
Conventional testing	
Fear[a] (n=235)	
No fear (1)	152 (64.7)
2	46 (19.6)
3	22 (9.4)
4	10 (4.3)
Highest fear (5)	5 (2.1)
Pain[b] (n=236)	
No pain (1)	162 (68.6)
2	62 (26.3)
3	11 (4.7)
4	1 (0.4)
Discomfort[c] (n=236)	

XSL•FO
RenderX

Criteria	Participants, n (%)
No discomfort (1)	175 (74.2)
2	44 (18.6)
3	14 (5.9)
4	2 (0.8)
Highest discomfort (5)	1 (0.4)
Satisfaction (n=236)	
Very satisfied	192 (81.4)
Satisfied	34 (14.4)
Neutral or not sure	9 (3.8)
Dissatisfied	1 (0.4)
Preference of timing to receive test results (n=249)	
1 day	141 (56.6)
<1 week	99 (39.8)
<2 weeks	7 (2.8)
≥2 weeks	2 (0.8)
Retest preference (n=249)	
Rapid test	148 (59.4)
Phlebotomy	67 (26.9)
No preference	34 (13.7)
Rapid test recommendation to friends (n=250)	
No	7 (2.8)
Yes	234 (93.6)
Not sure	9 (3.6)
Preference for follow-up (n=250)	
Phone call	39 (15.6)
SMS text messaging	34 (13.6)
App-based secure messaging	84 (33.6)
Face-to-face	60 (24)
No preference	33 (13.2)

[a]Mean score for point-of-care: 1.6 (SD 1.0); conventional: 1.6 (SD 1.0).
[b]Mean score for point-of-care: 1.4 (SD 0.6); conventional: 1.4 (SD 0.6).
[c]Mean score for point-of-care: 1.3 (SD 0.6); conventional: 1.3 (SD 0.7).

Health Care Provider Perspectives

Health care professionals were questioned, and their perspectives were tabulated qualitatively (Textbox 2). Across all queries, favorable responses were conveyed. For instance, health care professionals deemed multiplexed tests to be user-friendly and accurate when compared to reference standard tests. The overall perspective was that they would add value in Canada if implemented in routine care in community settings, although it was noted that acquiring a proper patient medical history was important to identify new infections. The rapid multiplexed test use was considered useful whether the test result was negative or positive; a negative status provided multiple benefits as patients' anxiety would diminish and the result would serve as a baseline for subsequent testing.

XSL•FO
RenderX

Textbox 2. Health care provider perspective of multiplexed tests in the context of the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (2021-2022).

What do you think about multiplexed tests?

- "I really like the multiplex tests. They are super user friendly and I think they have immense potential for providing comprehensive screening to vulnerable populations."
- "They should always be paired with such a positive blood test."
- "I enjoy using multiplex tests, they are simple to use and provide quick results. I also have found that the results have been consistent with bloodwork results, or dried blood spots."

Do you think they will have a future in Canada?

- "I think they have huge potential in Canada. I can see them being used in the community setting to provide on the spot screening."
- "Yes, but it's important to be careful for customers with one of the diagnoses already known."
- "I believe that they will have a future in Canada, our clinic has been using point of care testing for HCV screening before the multiplex study, and I believe the multiplex tests are an improvement as we can test for two/three bloodborne pathogens as opposed to one."
- "It will take some time to get approved."

If approved, do you think they will be useful for the populations they serve?

- "In the population I work with it can be challenging to collect bloodwork, and patients are hesitant to have it done. With the multiplex test it requires minimal blood, and is noninvasive, which makes the patients more likely to participate."
- "If approved the multiplex test will be useful in our population, as it is important for patients to have results on site."

Is it helpful for people to know if they test negative?

- "This is the reason why they take the test; people want to know if they are not contagious."
- "It provides relief and ease of mind to the patient. Furthermore, it serves as a baseline for rescreening."
- "I think knowing a negative status is just as important as knowing a positive status."

Impact

With respect to new infections, 29 new infections (n=27, 93% HCV; n=1, 3% HIV; and n=1, 3% syphilis) were detected using the AideSmart! multiplexed strategy in at-risk populations. Of the 29 new infections, 8 (28%) spontaneously cleared and required no further follow-up. Among the remaining 21 infected participants, 12 (57%) received follow-up treatment and care in visit 3, a total of 2 (10%) died, 4 (19%) were incarcerated, and 3 (14%) were completely lost to follow-up. Furthermore, 94.7% (380/401) of the participants did not require a visit 3 follow-up (participants who tested negative for all 3 pathogens, 372/380, 97.9%; and those who tested positive for HCV but the virus spontaneously cleared and they required no treatment, 8/380, 2.1%).

While the study recruitment occurred during the COVID-19 pandemic, with this strategy, we were able to track 99.3% (398/401) of the participants throughout the study. The pandemic-induced restrictions made it difficult for us to keep participants in care; despite the attempts made by the staff to retain these patients, loss to follow-up was inevitable.

Of note, 38% (11/29) of participants who tested positive had had a previous STD, and 38% (11/29) reported never using condoms at baseline. Interestingly, with the exception of 14% (4/29) of the participants, who were not comfortable disclosing their past sexual history, all individuals with new infections (25/29, 86%) had 0 to 5 recent sexual partners, thereby indicating that even persons with stable or few partners are at risk of contracting a sexually transmitted infection (STI). Moreover, most participants with new infections were unemployed (22/29, 76%), identified as a man (21/29, 72%), and aged ≥35 years (20/29, 69%).

Diagnostic Accuracy

The diagnostic accuracy of both rapid tests is presented in Tables 3 and 4, compared to reference standard tests through local laboratory assays in Table 3 and DBS in Table 4.

For HIV, both MedMira and Chembio reported a high sensitivity and specificity (>98%) when compared to local laboratory tests as a reference standard. With DBS as a reference standard alone, the sensitivity of the MedMira test for HIV was 93.8% (95% CI 69.8%-99.8%).

For HCV, the MedMira test was the sole rapid test that detected HCV antibodies; it reported a high sensitivity at 91.2% (95% CI 84.3%-95.7%) and a high specificity at >99% (95% CI 97.5%-99.9%).

For syphilis, against local laboratory tests, for both rapid tests, specificity was high (>99%). However, the sensitivity of the Chembio test varied (relative to local laboratory tests: 86.8%, 95% CI 71.9%-95.6%; relative to DBS: 80%, 95% CI 64.4%-91%). MedMira's sensitivity varied as well (relative to local laboratory tests: 57.9%, 95% CI 40.8%-73.7%; relative to DBS: 55%, 95% CI 38.5%-70.7%). This statement holds true while comparing test results between the rapid tests and both reference standard tests.

XSL•FO
RenderX

Table 3. Diagnostic accuracy of the Chembio and MedMira point-of-care (POC) tests in comparison to local laboratory tests in the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (2021-2022).

POC test	Sensitivity, % (95% CI)	Specificity, % (95% CI)
HIV		
Chembio	100.00 (79.41-100.00)	98.71 (96.72-99.65)
MedMira	100.00 (78.20-100.00)	100.00 (98.82-100.00)
Syphilis		
Chembio	86.84 (71.91-95.59)	99.65 (98.06-99.99)
MedMira	57.89 (40.82-73.69)	100.00 (98.72-100.00)
Hepatitis C		
Chembio	—[a]	—
MedMira	90.32 (80.12-96.37)	99.62 (97.92-99.99)

[a]Not applicable.

Table 4. Diagnostic accuracy of the Chembio and MedMira point-of-care (POC) tests in comparison to dried blood spot tests in the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (2021-2022).

POC test	Sensitivity, % (95% CI)	Specificity, % (95% CI)
HIV		
Chembio	100.00 (76.84-100.00)	98.94 (97.30-99.71)
MedMira	93.75 (69.77-99.84)	100.00 (99.02-100.00)
Syphilis		
Chembio	80.00 (64.35-90.95)	99.03 (97.19-99.80)
MedMira	55.00 (38.49-70.74)	100.00 (98.81-100.00)
Hepatitis C		
Chembio	—[a]	—
MedMira	91.15 (84.33-95.67)	99.29 (97.46-99.91)

[a]Not applicable.

Discussion

Principal Findings

In this study, we report high acceptability (401/401, 100%), satisfaction (393/401, 98%), and feasibility (305/401, 76.1%) for the AideSmart! multiplexed strategy. Participants preferred multiplexed rapid testing (148/249, 59.4%) over conventional laboratory-based testing, primarily due to its ability to provide a test result (negative or positive) in a shorter turnaround time.

Study findings demonstrate the strategy's relevance for Canadian at-risk populations, as illustrated by the importance of obtaining a negative test result in at-risk populations and its overall impact in detecting new infections. With respect to impact, at the RECAP site, we detected most new infections (26/29, 90%), all of which were HCV, whereas at RÉZO, we detected 3 new infections (n=1, 33% HIV; n=1, 33% HCV; and n=1, 33% syphilis). Furthermore, all participants obtained laboratory-confirmed test results, were engaged throughout the process, stayed in communication with health care aides regarding their test results, and were counselled in time. Furthermore, treatment decisions were expedited in 3 visits for participants who tested positive. Most participants (377/401, 94%) did not require treatment as many infections spontaneously resolved. Although many participants who required follow-up (12/21, 57%) were linked and retained in care in visit 3, due to COVID-19 restrictions and repeat lockdowns, tracking 43% (9/21) of the participants was difficult. However, we were notified that 22% (2/9) died and 44% (4/9) were incarcerated. Despite the best efforts of our staff, 33% (3/9) left the province or were lost to follow-up. Overall, using this digital strategy, 99.3% (398/401) of our participants were tracked throughout the pandemic.

To the health care providers, knowledge of serostatus for STBBIs with rapid tests allowed for planning of screening and confirmatory testing and counselling for these participants at a time during the pandemic when services were interrupted for many STBBIs. This positive impact can be further enhanced by a rapid turnaround time of 15 minutes (once approved and implemented) for preliminary test results, as opposed to a waiting time of 4 to 24 days to obtain laboratory-based confirmatory results. The long waiting time carries the possibility of losing at-risk populations.

Health care providers underscored the importance of declaring both positive and negative test results using multiplexed rapid

XSL•FO
RenderX

test results. Indeed, as the study population comprised at-risk individuals, the knowledge of one's negative status is crucial for infection control. Although false positive results are troublesome, false negative results can lead to more devastating outcomes when used for screening, including a false sense of security and higher likelihood of engaging in risky behaviors fueling forward transmission. Health care providers expressed positive feedback regarding the multiplexed testing strategy and were convinced of their use in future screening initiatives. The digital strategy allowed for the tracking of participants' care throughout the care pathway and maximized engagement of health care workers and participants during the COVID-19 pandemic, especially when health care professionals knew and used patients' preferred method of communication.

One very encouraging finding was the consistently high specificity of the rapid tests for the pathogens, often approaching 99% to 100% with biomarker-based tests, suggesting that both tests can be used with confidence to rule out disease. This property was noted by our health care providers. As such, negative test results could be considered negative unless proven otherwise through a second confirmatory laboratory test. This saves time and energy for health care systems and allows for task shifting and for focus to shift to new infections based on individuals' history and presentation. This was relevant during the pandemic, when the possibility of bringing underserved populations to screen for infections, in the context of health care staff shortages was very low.

This study is the first in Canada to assess the diagnostic accuracy of the multiplexed Chembio Dual Path Platform assay (with an electronic reader) and of MedMira's investigational Multiplo Rapid *T pallidum*, HIV, and HCV Antibody Test in Canada and the first to demonstrate the success of a digital connected strategy for multiplexed biomarker tests for triple infections. We observed that, to increase the accuracy (sensitivity) of detection, electronic readers were extremely useful. They helped reduce the ambiguity faced in field implementation with respect to rapid test result interpretation; this has been noted in other studies or sites. With electronic readers, health care professionals could state with more confidence that observed positive results were true positives and that observed negative results were true negatives. This was observed with the syphilis and HIV Chembio test results in this study. Rapid tests, if used in combination with electronic readers, will be extremely useful for expanded access and screening initiatives in outreach and community-based settings where laboratory-based tests are not readily available in a rapid time and where the endemicity of these infections in subpopulations is high and the chance of losing a tester to follow-up is also high. An important strength of the MedMira test is the ability to accurately screen for HCV using a blood sample in a short span of time. However, the Chembio test demonstrated an improved and increased sensitivity for HIV and syphilis with the quantitative reader, increasing confidence in the test result. MedMira fell short, particularly in screening for syphilis; the Chembio test performed very well with an electronic reader. As both are lateral flow assays, a visual assessment is required to interpret findings. However, the use of an electronic reader with the Chembio test improves the health care professional's task and reduces incorrect interpretation. With many rapid assays without electronic readers, faint test lines or dots are missed, impacting sensitivity estimations.

Last year, global diagnostic organizations such as the Foundation for Innovative New Diagnostics (FIND), Geneva [24], and the World Health Organization (WHO) released their new guidelines. These guidelines have called for a greater use of electronic readers integrated with POC tests, apps, and digital supports that can enhance an accurate reading of rapid test results—an attribute that was noted as an advantage with connected rapid biomarker-based tests [25,26].

Diagnostic accuracy with biomarker-based assays varies between devices, for pathogens, against different reference standard tests, and for populations with varying prevalence of coinfections, as is the case with many rapid tests [5,27]. It is important to note additional factors that may impact the sensitivity of a multiplexed test. To corroborate this, in our previous studies with multiplexed tests without apps in India and Montreal, we observed that factors such as the immunogenicity of the patient, level of immune compromise, presence, prevalence, and endemicity of STBBIs, number of biomarkers present and type of device, and incidence of STBBIs in the population tested impacted the avidity that in turn affected the sensitivity of biomarker-based rapid tests [27]. In a few editorials, we have discussed the usefulness of these assays for subpopulations in national and global settings [5,27-29]. We found that 2-pathogen multiplexed biomarker rapid tests were easier to implement than a 3-pathogen test [30-33]. On the basis of our experience, we recommend the use of 2-pathogen rapid biomarker-based tests together with quantitative readers that help improve test interpretation capacity at the POC. As both rapid tests detected antibodies, several positive rapid test results were associated with treated previous infections.

Of note, the availability of medical history and risk profiles or electronic medical records on file and information collected using the AideSmart! app becomes especially important and is relevant as it helps make sense of a new diagnosis of infections or reinfections in outreach settings.

While no study to the best of our knowledge has presented the field-based diagnostic accuracy of a connected multiplexed strategy using biomarker tests, few HIV or syphilis rapid tests are commercially available. Similar to our results, their specificity is generally very high (>98%), but sensitivities vary [33]. For instance, the sensitivity of the bioLytical INSTI test is 56.8% for syphilis but is high (98.8%) for HIV [34]. The SD Bioline test yields an impressive sensitivity of 86.4% to 100% for syphilis and 99.1% to 100% for HIV [30-32,35]. However, the Biosynex Triplex HIV, HCV, and HBsAg test has reported an optimal (100%) diagnostic accuracy for all 3 pathogens [36]. We recommend the use of electronic readers for the tests reporting poor sensitivity to enhance their performance.

While multiplexed rapid biomarker-based tests offer a potential to screen for multiple pathogens in a short turnaround time, saving time and money to the health care system and the testers, they also catalyze the screening pathway for the populations who need multiplexed rapid biomarker-based tests through their availability and accuracy. To aid health care professionals in

XSL•FO
RenderX

selecting the most appropriate test in their setting, we have published a systematic review that documents the accuracy of all rapid multiplexed assays and platforms [33]. While rapid lateral flow assay tests were used in this study, it is worth noting that platform devices can also yield rapid same-day test results. Multiplexed platform diagnostic technologies have been shown to be highly accurate using metrics of diagnostic performance (sensitivity and specificity). Certain devices yield such high diagnostic accuracy that they can be used to confirm infection [33], obviating the need to conduct or call for further laboratory-based tests. Portable platforms have the potential to transform the testing or treatment for these infections in community-based outreach settings; however, their need for laboratory infrastructure prevents their use in field settings. That said, the test results are available in 1 to 2 visits, and their potential to bring individuals to linkage if set up judiciously remains high. Such combination strategies of connected app-based rapid tests and portable platform or biomarker tests with electronic readers, if used together, can be useful to achieve elimination targets for HIV, HCV, and syphilis. Platform technologies detect nucleic acids as opposed to antibodies, and their cartridges remain expensive, but they will be more useful in diagnosing and treating new infections. However, these larger benchtop devices are associated with limited portability and high cost, and some require infrastructure such as air conditioning, which impedes their use in high temperature–prone outreach settings. While the benefits of platform devices may outweigh their drawbacks for some centers, biomarker-based flow-through multiplexed devices are cost-effective, affordable, and easy to use and represent an important and viable diagnostic option for populations with limited access to testing, at the very least to rule out infection given their high specificity.

Study Strengths and Relevance

This study is the first to offer, evaluate, and generate evidence for the use of an integrated app-based multiplexed rapid STBBI testing strategy for at-risk Canadian populations, namely, people who inject drugs and immigrant and refugee populations. This study was conducted during and participants were recruited throughout the COVID-19 pandemic. For this study, the AideSmart! app was tailored to the Canadian context and test devices to enhance its global portability. This study represented key populations that typically need or desire rapid and regular STBBI screening. Participants were recruited from 2 sites, of which one specializes in HCV screening and care and the other provides services to immigrant and refugee populations. Our findings highlight that the AideSmart! multiplexed strategy impacted detection of new HCV, syphilis, and HIV infections with a rapid turnaround time; helped initiate counselling, linkage, and pre-exposure prophylaxis plans for those who tested both negative and positive; and helped track 99.3% (398/401) of our participants throughout the pandemic.

Study Limitations

The study findings should be interpreted cautiously considering certain design limitations. Although our study design aimed to minimize biases, convenience sampling raises the potential for possible selection bias (sampling bias). Individuals willing to undergo STBBI testing may be more likely to be infected with an STBBI as these persons are more aware of their riskier lifestyle. This assumption was corroborated through our findings—study participants with a previous STD reported higher past testing rates when compared to participants without a previous STD. However, even persons with less risky lifestyles are at risk of contracting an STBBI, which illustrates the importance of increasing awareness among various populations to increase testing and, therefore, detection and treatment. We attempted to reduce misclassification bias by declaring results only upon receipt of laboratory-confirmed test results. Detection, ascertainment, and verification biases were minimized through an independent verification of test results from 3 laboratories.

As the study was entirely conducted during the COVID-19 pandemic, recruitment and retention presented many challenges, which led to delays in study completion. These were namely due to stay-at-home orders combined with individuals' hesitancy to get tested during such uncertain times. Furthermore, while rapid test results were obtained in a few minutes, they could only be provided at visit 2 due to the investigational nature of the rapid tests and the need to wait for the test results from the laboratories, as well as the directives regarding clinic visits during the COVID-19 pandemic. The pandemic restrictions may have contributed to participants' attrition at visit 2. However, it is worth noting that, in clinical care, the use of POC antibody-based tests limits the ability to provide definitive same-day positive screening results as further confirmatory testing is required. However, if confirmatory testing is expedited, treatment initiation is possible in visit 2, as was done in our study.

Implications for Practice

Going one step further from rapid testing, the open access digital AideSmart! strategy is able to integrate multiplexed tests and platforms, thereby optimizing connected initiatives and saving health aide time in field or practice settings. Connected diagnostics and digitization represent an encouraging development during the COVID-19 pandemic and simplify notification of test results [15,37]. Connected solutions increase convenience and expedited access to screen, confirm, treat, and retain individuals in care, encouraging task shifting. This study illustrates the value and impact of an innovative, connected strategy for quality screening, patient engagement, data management, communication, linkages to care, and tracking of participants during the pandemic [38-40]. In addition to testing, the possibility to train health care professionals in each setting allows for quality control and quality assurance, which bodes well for certification and helps maintain a high quality of rapid POC tests, which is a concern in field implementation. Moreover, it allows for efficiency in data management and timely linkage to services.

The combination of accurate rapid tests and connected digital solutions fulfills several stakeholder needs. While individuals being tested generally seek highly accurate results with a quick turnaround time, the use of rapid tests together with a digital solution will significantly impact diagnostic care in several ways. Tasks are alleviated for health care professionals as the app-based program provides testing guidance, allows for result documentation, and improves stakeholder communication.

XSL•FO
RenderX

Additional studies have shown the benefits of digital health in STBBI care. A scoping review by Cao et al [16] illustrated that digital platforms can improve STI prevention, accessibility and acceptability of testing, understanding and uptake of pre-exposure prophylaxis, surveillance, and postdiagnostic care (such as linkage and retention in care) for individuals who tested positive for at least one STI. This endeavor aligns with several global initiatives, such as those by the WHO [41] and the FIND [42], to integrate the use of digital health in health care services, including diagnostic care.

In terms of the generalizability of the findings, this study builds upon a previous AideSmart! study (Grand Challenges Canada funded) in India [17], a middle-income country, in a different at-risk population: pregnant women who were tested for bacterial and viral STBBIs. With a Canadian Institutes of Health Research–funded Canadian evaluation, we can confidently state that this rapid multiplexed screening strategy could be applied to populations desiring STBBI screening in high- and middle-income settings, including people who inject drugs; pregnant women; STD clinic attendees; and immigrant, refugee, and gender diverse populations. With new funding from the National Institutes of Health, in the coming years, we will aim to expand this program for multiple domains and infections and adopt it for use among health care workers in India [43], Uganda, and Ecuador [44].

Moreover, in Canada, while there is currently only 1 approved multiplexed test (INSTI) that is commercially available, it has a sensitivity of 98.85% (95% CI 93.4%-100%) and a specificity of 100% (95% CI 98.1%-100%) for HIV but a low sensitivity of 56.8% (95% CI 44.7%-68.2%) and a specificity of 98.5% (95% CI 95.7%-99.7%) for syphilis [34]. There is an imperative need to approve other high-performing tests such as Chembio (for syphilis and HIV) and Multiplo (for HCV and HIV) to meet the demand for multiplexed tests but also to drive down the prices of rapid tests. These tests can help expand the offer to screen at-risk Canadians, who regularly miss out on STBBI screening. Canada is presently facing an outbreak of syphilis in the Canadian Prairies and numerous challenges in timely screening of HCV-affected populations. This independent evaluation is relevant for such considerations.

Rapid test approvals can aid the achievement of elimination targets for syphilis, HCV, and HIV, particularly to meet the intensified demand for rapid testing in several provinces in a post–COVID-19 pandemic era. Demand for portable molecular platform-based tests is also high in high-volume clinics and community-based settings. While POC tests are useful to provide rapid results, molecular platforms combine the benefits of POC and conventional tests with DNA and RNA detection with a high diagnostic accuracy, and results could be provided for test-and-treat strategies without the need for further confirmation, offering a multitude of savings to the health care system [45,46]. This must be considered to reduce the burden of individuals living with undiagnosed infections in Canada and the world.

Conclusions

We conclude that the AideSmart! integrated multiplexed STBBI rapid screening strategy offered by health aides was deemed acceptable, accurate, and feasible while impacting the detection of multiple STBBIs in a rapid turnaround time, all while satisfying end-user and health care provider preferences. This study demonstrated the utility of connected digital solutions and multiplexed POC testing assays with electronic readers, which can facilitate the detection of infections at the POC in challenging circumstances and among vulnerable populations. A total of 99.3% (398/401) of at-risk participants were tracked throughout the pandemic. The app-based strategy engaged participants and health care providers, demonstrating the relevance of digitally enabled diagnostic strategies in stressful pandemic times. While multiplexed assays with electronic readers provided accurate test results at the POC in a rapid turnaround time, impacting linkage to clinical decisions, the negative test results were also reassuring to health care providers involved in the care of these populations. This study is relevant not only for Canadians but also for global populations. Aligned with the WHO's One Health approach [47], this digital strategy can be used to serve the hardest-to-reach undiagnosed marginalized populations worldwide to help meet elimination targets for triple infections.

Acknowledgments

The authors thank all the study participants who took part during the COVID-19 pandemic. This study was funded by the Canadian Institutes of Health Research (grant PJT153149). The study sponsor had no role in the study design; the collection, analysis, and interpretation of the data; the writing of the report; or the decision to submit the paper for publication. NPP acknowledges the current ongoing support from the Fonds de recherche du Québec – Santé (Senior Scientist award 252845; Distinguished Research Scholar award 324154) for the duration of the study. She acknowledges the support of the India-Canada Centre for Innovative Multidisciplinary Partnerships to Accelerate Community Transformation and Sustainability (CNO 3072) and the US National Institutes of Health Point of Care Technologies for Nutrition, Infection and Cancer for Global Health grant (U54EB034654) for technology-based projects. AideSmart! is an open access app funded by Grand Challenges Canada and the Canadian Institutes of Health Research and recently, the US NIH.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

AK managed the study, contributed to study design, conducted data analysis, and prepared the manuscript. DW oversaw study completion and recruitment at the Centre for Research, Education, and Clinical Care of At-Risk Populations (RECAP) site and provided treatment to infected patients. JB conducted testing and coordinated treatment linkage at the RÉZO site. KH and PF conducted testing and coordinated treatment linkage at the RECAP site. SM oversaw project management at the RECAP site. GD-L oversaw project management at the RÉZO site. CM and JK led confirmatory testing at the National Labs. OV prepared the manuscript. NPP developed the study concept, oversaw project management, and prepared the manuscript. All authors contributed to the writing of the initial drafts or the final editing of the final drafts of the manuscript. All authors had full access to the data in the study, and the corresponding author and first author take responsibility for the data and analyses related to the project.

Conflicts of Interest

NPP is the creator of AideSmart! App, an open access app (McGill University; copyright ROI 16126; June 3rd, 2016), and shares copyright with her former trainees, Jana Daher and Sean Yaphe.

References

1. Bachmann LH, Grimley DM, Waithaka Y, Desmond R, Saag MS, Hook EW3. Sexually transmitted disease/HIV transmission risk behaviors and sexually transmitted disease prevalence among HIV-positive men receiving continuing care. Sex Transm Dis 2005 Jan;32(1):20-26. [doi: 10.1097/01.olq.0000148293.81774.e4] [Medline: 15614117]
2. Perez-Brumer AG, Konda KA, Salvatierra HJ, Segura ER, Hall ER, Montano SM, et al. Prevalence of HIV, STIs, and risk behaviors in a cross-sectional community- and clinic-based sample of men who have sex with men (MSM) in Lima, Peru. PLoS One 2013;8(4):e59072 [FREE Full text] [doi: 10.1371/journal.pone.0059072] [Medline: 23634201]
3. Syphilis among gay, bisexual, two-spirit and other men who have sex with men: a resource for population-specific prevention. Government of Canada. URL: https://www.canada.ca/en/public-health/services/infectious-diseases/sexual-health-sexually-transmitted-infections/syphilis-resource-population-specific-prevention.html [accessed 2022-12-21]
4. Mashamba-Thompson TP, Jama NA, Sartorius B, Drain PK, Thompson RM. Implementation of point-of-care diagnostics in rural primary healthcare clinics in South Africa: perspectives of key stakeholders. Diagnostics (Basel) 2017 Jan 08;7(1):3 [FREE Full text] [doi: 10.3390/diagnostics7010003] [Medline: 28075337]
5. Pant Pai N, Peeling RW, Smith BD, Dowdy D. Point-of-care tests for HIV, related coinfections, and blood-borne infections. AIDS Res Treat 2014;2014:625082 [FREE Full text] [doi: 10.1155/2014/625082] [Medline: 24895531]
6. Rogers B, Tao J, Murphy M, Chan PA. The COVID-19 pandemic and sexually transmitted infections: where do we go from here? Sex Transm Dis 2021 Jul 01;48(7):e94-e96 [FREE Full text] [doi: 10.1097/OLQ.0000000000001445] [Medline: 33872225]
7. Shukla P, Lee M, Whitman SA, Pine KH. Delay of routine health care during the COVID-19 pandemic: a theoretical model of individuals' risk assessment and decision making. Soc Sci Med 2022 Aug;307:115164 [FREE Full text] [doi: 10.1016/j.socscimed.2022.115164] [Medline: 35816834]
8. Pujolar G, Oliver-Anglès A, Vargas I, Vázquez ML. Changes in access to health services during the COVID-19 pandemic: a scoping review. Int J Environ Res Public Health 2022 Feb 03;19(3):1749 [FREE Full text] [doi: 10.3390/ijerph19031749] [Medline: 35162772]
9. Tao J, Napoleon SC, Maynard MA, Almonte A, Silva E, Toma E, et al. Impact of the COVID-19 pandemic on sexually transmitted infection clinic visits. Sex Transm Dis 2021 Jan;48(1):e5-e7 [FREE Full text] [doi: 10.1097/OLQ.0000000000001306] [Medline: 33181578]
10. Garcia PJ, Miranda AE, Gupta S, Garland SM, Escobar ME, Fortenberry JD. The role of sexually transmitted infections (STI) prevention and control programs in reducing gender, sexual and STI-related stigma. EClinicalMedicine 2021 Mar;33:100764 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100764] [Medline: 33718850]
11. Chattu VK, Lopes CA, Javed S, Yaya S. Fulfilling the promise of digital health interventions (DHI) to promote women's sexual, reproductive and mental health in the aftermath of COVID-19. Reprod Health 2021 Jun 04;18(1):112 [FREE Full text] [doi: 10.1186/s12978-021-01168-x] [Medline: 34088319]
12. Lunt A, Llewellyn C, Bayley J, Nadarzynski T. Sexual healthcare professionals' views on the rapid provision of remote services at the beginning of COVID-19 pandemic: a mixed-methods study. Int J STD AIDS 2021 Oct 09;32(12):1138-1148 [FREE Full text] [doi: 10.1177/09564624211023018] [Medline: 34106016]
13. Naeem F, Karellis A, Nair S, Routy JP, Yansouni CP, Kim J, et al. Multiplexed technologies for sexually transmitted infections: global evidence on patient-centered and clinical health outcomes. BMJ Glob Health 2021 Jul 23;6(7):e005670 [FREE Full text] [doi: 10.1136/bmjgh-2021-005670] [Medline: 34301675]
14. Habel MA, Sullivan P, Hall C, Aral S. Remote health: optimizing the delivery of sexual health care. Sexual Trans Dis 2022 Mar 21;49(11S):S1-S6. [doi: 10.1097/olq.0000000000001618]
15. Fleming KA, Horton S, Wilson ML, Atun R, DeStigter K, Flanigan J, et al. The Lancet Commission on diagnostics: transforming access to diagnostics. The Lancet 2021 Nov 27;393(10315):1997-2050.

XSL•FO
RenderX

16. Cao BB, Bao H, Oppong E, Feng S, Smith KM, Tucker JD, et al. Digital health for sexually transmitted infection and HIV services: a global scoping review. Curr Opin Infect Dis 2020 Feb;33(1):44-50 [FREE Full text] [doi: 10.1097/QCO.0000000000000619] [Medline: 31789695]
17. Pant Pai N, Daher J, Prashanth HR, Shetty A, Sahni RD, Kannangai R, et al. Will an innovative connected AideSmart! app-based multiplex, point-of-care screening strategy for HIV and related coinfections affect timely quality antenatal screening of rural Indian women? Results from a cross-sectional study in India. Sex Transm Infect 2019 Mar;95(2):133-139 [FREE Full text] [doi: 10.1136/sextrans-2017-053491] [Medline: 30322858]
18. Santé et mieux-être des hommes gais et bisexuels, cis et trans. RÉZO. URL: https://www.rezosante.org/ [accessed 2023-01-11]
19. Centre for Research, Education and Clinical Care of At-Risk Populations homepage. Centre for Research, Education & Clinical Care of At-Risk Populations. URL: http://www.recapsj.ca/ [accessed 2023-01-11]
20. Sexually transmitted and blood-borne infections: guides for health professionals. Government of Canada. URL: https://www.canada.ca/en/public-health/services/infectious-diseases/sexual-health-sexually-transmitted-infections/canadian-guidelines.html [accessed 2022-12-21]
21. Dhand NK, Khatkar MS. Sample size calculator for comparing two paired proportions. Statulator. 2014. URL: http://statulator.com/SampleSize/ss2PP.html [accessed 2022-12-19]
22. Stata statistical software: release 17. StataCorp LLC. 2021. URL: https://www.stata.com/stata-news/news36-2/ [accessed 2024-09-10]
23. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.R-project.org [accessed 2024-09-10]
24. FIND. URL: https://www.finddx.org/ [accessed 2024-03-30]
25. Target product profile for readers of rapid diagnostic tests. World Health Organization. 2023 Feb 24. URL: https://www.who.int/publications/i/item/9789240067172 [accessed 2024-09-10]
26. RDT digital companion apps. FIND. URL: https://www.finddx.org/what-we-do/cross-cutting-workstreams/digital-health/rdt-digital-companion-apps/#:~:text=The%20RDT%20reader%20app%2C%20co,now%20integrated%20into%20the%20RDToolkit [accessed 2024-09-10]
27. Pai NP, Dhurat R, Potter M, Behlim T, Landry G, Vadnais C, et al. Will a quadruple multiplexed point-of-care screening strategy for HIV-related co-infections be feasible and impact detection of new co-infections in at-risk populations? Results from cross-sectional studies. BMJ Open 2014 Dec 15;4(12):e005040 [FREE Full text] [doi: 10.1136/bmjopen-2014-005040] [Medline: 25510882]
28. Smallwood M, Pant Pai N. Improving the quality of diagnostic studies evaluating point of care tests for acute HIV infections: problems and recommendations. Diagnostics (Basel) 2017 Mar 04;7(1):13 [FREE Full text] [doi: 10.3390/diagnostics7010013] [Medline: 28273857]
29. Pant Pai N, Daher J. Multiplexed testing for HIV and related bacterial and viral co-infections at the point-of-care: quo vadis? Expert Rev Mol Diagn 2015 Apr 20;15(4):463-469. [doi: 10.1586/14737159.2015.1021335] [Medline: 25795042]
30. Omoding D, Katawera V, Siedner M, Boum Y2. Evaluation of the SD Bioline HIV/Syphilis Duo assay at a rural health center in Southwestern Uganda. BMC Res Notes 2014 Oct 22;7:746 [FREE Full text] [doi: 10.1186/1756-0500-7-746] [Medline: 25339379]
31. Shakya G, Singh DR, Ojha HC, Ojha CR, Mishra SK, Malla K, et al. Evaluation of SD Bioline HIV/syphilis duo rapid test kits in Nepal. BMC Infect Dis 2016 Aug 26;16(1):450 [FREE Full text] [doi: 10.1186/s12879-016-1694-9] [Medline: 27566067]
32. Bristow CC, Adu-Sarkodie Y, Ondondo RO, Bukusi EA, Dagnra CA, Oo KY, et al. Multisite laboratory evaluation of a dual human immunodeficiency virus (HIV)/syphilis point-of-care rapid test for simultaneous detection of HIV and syphilis infection. Open Forum Infect Dis 2014 Mar;1(1):ofu015 [FREE Full text] [doi: 10.1093/ofid/ofu015] [Medline: 25734088]
33. Karellis A, Naeem F, Nair S, Mallya SD, Routy JP, Gahagan J, et al. Multiplexed rapid technologies for sexually transmitted infections: a systematic review. Lancet Microbe 2022 Apr;3(4):e303-e315 [FREE Full text] [doi: 10.1016/S2666-5247(21)00191-9] [Medline: 35544067]
34. Stafylis C, Bristow CC, Natoli LJ, Salow KR, Davidson E, Granados Y, et al. Field evaluation of a dual rapid human immunodeficiency virus and treponemal syphilis rapid test in community-based clinics in Los Angeles and New York. Diagn Microbiol Infect Dis 2019 Apr;93(4):325-328 [FREE Full text] [doi: 10.1016/j.diagmicrobio.2018.10.002] [Medline: 30704779]
35. Lodiongo DK, K Bior B, W Dumo G, S Katoro J, Mogga JJ, Lokore ML, et al. Field evaluation of SD BIOLINE HIV/syphilis duo assay among pregnant women attending routine antenatal care in Juba, South Sudan. PLoS One 2018;13(10):e0205383 [FREE Full text] [doi: 10.1371/journal.pone.0205383] [Medline: 30304043]
36. Kalla GC, Voundi EV, Guiadem R, Iii FA, Bélec L, Mbopi-Keou FX. Mass campaigns for HIV, HBV (HBsAg) and HCV screening by multiplex rapid diagnostic test in sub-Saharan Africa using mobile units: the game changer. Int J Infect Dis 2019 Feb;79:107. [doi: 10.1016/j.ijid.2018.11.265]
37. Vasudevan S, Saha A, Tarver ME, Patel B. Digital biomarkers: convergence of digital health technologies and biomarkers. NPJ Digit Med 2022 Mar 25;5(1):36 [FREE Full text] [doi: 10.1038/s41746-022-00583-z] [Medline: 35338234]

XSL•FO
RenderX

38. Diagnostics access initiative to achieve the 90-90-90 treatment target. The Joint United Nations Programme on HIV and AIDS. URL: https://www.unaids.org/sites/default/files/media_asset/20150422_diagnostics_access_initiative.pdf [accessed 2022-12-21]
39. Trubnikov M, Yan P, Archibald C. Estimated prevalence of Hepatitis C Virus infection in Canada, 2011. Can Commun Dis Rep 2014 Dec 18;40(19):429-436 [FREE Full text] [doi: 10.14745/ccdr.v40i19a02] [Medline: 29769874]
40. Summary: estimates of HIV incidence, prevalence and proportion undiagnosed in Canada, 2014. Government of Canada. 2014. URL: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/summary-estimates-hiv-incidence-prevalence-proportion-undiagnosed-canada-2014.html [accessed 2022-12-21]
41. Global strategy on digital health 2020-2025. World Health Organization. 2021 Aug 18. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-09-10]
42. Testing for healthy and safe lives. Foundation for Innovative New Diagnostics. 2021. URL: https://www.finddx.org/wp-content/uploads/2022/12/20221219_strategy_2021_FV_EN.pdf [accessed 2024-09-23]
43. IC-IMPACTS 2021-2022 annual report. IC-IMPACTS. 2022. URL: https://ic-impacts.com/wp-content/uploads/2022/11/IC-IMPACTS-2021-2022-Annual-Report-Web.pdf [accessed 2024-09-10]
44. PORTENT: a center for point of care technologies for nutrition, infection and cancer for global health. Cornell University. URL: https://pocglobalhealth.cornell.edu/ [accessed 2024-09-10]
45. Gaydos CA, Manabe YC, Melendez JH. A narrative review of where we are with point-of-care sexually transmitted infection testing in the United States. Sex Transm Dis 2021 Aug 01;48(8S):S71-S77 [FREE Full text] [doi: 10.1097/OLQ.0000000000001457] [Medline: 34110728]
46. Dorward J, Naidoo J, Moodley P, Sookrajh Y, Samsunder N, Sayed F, et al. Diagnostic accuracy of the Rapid Xpert HIV-1 Viral Load XC, Xpert HIV-1 Viral Load, and m-PIMA HIV-1/2 Viral Load in South African Clinics. J Acquir Immune Defic Syndr 2022 Oct 01;91(2):189-196 [FREE Full text] [doi: 10.1097/QAI.0000000000003037] [Medline: 36094486]
47. One health. World Health Organization. URL: https://www.who.int/health-topics/one-health#tab=tab_1 [accessed 2024-09-10]

Abbreviations

DBS: dried blood spot
FIND: Foundation for Innovative New Diagnostics
HCV: hepatitis C virus
HIPAA: Health Insurance Portability and Accountability Act
POC: point-of-care
RECAP: Centre for Research, Education, and Clinical Care of At-Risk Populations
STBBI: sexually transmitted blood-borne infection
STD: sexually transmitted disease
STI: sexually transmitted infection
WHO: World Health Organization

Edited by A Mavragani; submitted 04.12.23; peer-reviewed by F Regalado, Q Fang; comments to author 14.03.24; revised version received 06.05.24; accepted 20.08.24; published 18.10.24.

Please cite as:
Karellis A, Webster D, Boulanger J, Harland K, Feltmate P, Materniak S, Daunais-Laurin G, Mesa C, Vaikla O, Kim J, Pant Pai N Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study
J Med Internet Res 2024;26:e55075
URL: https://www.jmir.org/2024/1/e55075
doi:10.2196/55075
PMID:

©Angela Karellis, Duncan Webster, Jean Boulanger, Kate Harland, Paige Feltmate, Stefanie Materniak, Gabriel Daunais-Laurin, Christine Mesa, Olivia Vaikla, John Kim, Nitika Pant Pai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Monitoring Frailty Through a Mobile/Web-Based Application and a Sensor Kit to Prevent Functional Decline in Frail and Prefrail Older Adults: FACET (Frailty Care and Well Function) Pilot Randomized Controlled Trial

Myriam Valdés-Aragonés[1], MD; Rodrigo Pérez-Rodríguez[2], PhD; José Antonio Carnicero[3], PhD; Pedro A Moreno-Sánchez[4], PhD; Myriam Oviedo-Briones[3], MD; Elena Villalba-Mora[5,6], PhD; Pedro Abizanda-Soler[7,8], MD, PhD; Leocadio Rodríguez-Mañas[1,8], MD, PhD

[1]Geriatrics Service, Getafe University Hospital, Getafe, Spain
[2]Intelligent Robotics Lab, Universidad Rey Juan Carlos, Fuenlabrada, Spain
[3]Biomedical Research Foundation, Getafe University Hospital, Getafe, Spain
[4]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[5]Human-Computer Interaction Laboratory, Centre for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain
[6]Center of Biomedical Research on Bioengineering, Biomaterials and Nanomedicine, Instituto de Salud Carlos III, Madrid, Spain
[7]Geriatrics Service, Albacete University Hospital, Albacete, Spain
[8]Center of Biomedical Research on Frailty and Healthy Aging, Instituto de Salud Carlos III, Madrid, Spain

Corresponding Author:
Rodrigo Pérez-Rodríguez, PhD
Intelligent Robotics Lab
Universidad Rey Juan Carlos
Camino del Molino 5
28942
Fuenlabrada
Spain
Phone: 34 914888401
Email: rodrigo.perez@urjc.es

Abstract

Background: Frailty represents a state of susceptibility to stressors and constitutes a dynamic process. Untreated, this state can progress to disability. Hence, timely detection of alterations in patients' frailty status is imperative to institute prompt clinical interventions and impede frailty progression. With this aim, the FACET (Frailty Care and Well Function) technological ecosystem was developed to provide clinically gathered data from the home to a medical team for early intervention.

Objective: The aim of this study was to assess whether the FACET technological ecosystem prevents frailty progression and improves frailty status, according to the frailty phenotype criteria and Frailty Trait Scale-5 items (FTS-5) at 3 and 6 months of follow-up.

Methods: This randomized clinical trial involved 90 older adults aged ≥70 years meeting 2 or more Fried frailty phenotype criteria, having 4 or more comorbidities, and having supervision at home. This study was conducted between August 2018 and June 2019 at the geriatrics outpatient clinics in Getafe University Hospital and Albacete University Hospital. Participants were randomized into a control group receiving standard treatment and the intervention group receiving standard treatment along with the FACET home monitoring system. The system monitored functional tests at home (gait speed, chair stand test, frailty status, and weight). Outcomes were assessed using multivariate linear regression models for continuous response and multivariate logistic models for dichotomous response. *P* values less than .05 were considered statistically significant.

Results: The mean age of the participants was 82.33 years, with 28% (25/90) being males. Participants allocated to the intervention group showed a 74% reduction in the risk of deterioration in the FTS-5 score (*P*=.04) and 92% lower likelihood of worsening by 1 point according to Fried frailty phenotype criteria compared to the control group (*P*=.02) at 6 months of follow-up. Frailty

XSL•FO
RenderX

status, when assessed through FTS-5, improved in the intervention group at 3 months (*P*=.004) and 6 months (*P*=.047), while when the frailty phenotype criteria were used, benefits were shown at 3 months of follow-up (*P*=.03) but not at 6 months.

Conclusions: The FACET technological ecosystem helps in the early identification of changes in the functional status of prefrail and frail older adults, facilitating prompt clinical interventions, thereby improving health outcomes in terms of frailty and functional status and potentially preventing disability and dependency.

Trial Registration: ClinicalTrials.gov NCT03707145; https://clinicaltrials.gov/study/NCT03707145

(J Med Internet Res 2024;26:e58312) doi:10.2196/58312

KEYWORDS

frailty; functional status; older adults; new technologies; sensor; monitoring system; information and communication technologies; mobile app; sensor kit; sensors; technological ecosystem; clinical intervention

Introduction

Functional status is one of the best indicators of health condition and predicts poor outcomes better than morbidity in older people [1]. In the pathway leading from robustness to disability, frailty is a dynamic process, and transitions that occur in the state of frailty happen bidirectionally: from better to worse states on the spectrum of frailty as well as toward disability [2]. Life expectancy has increased in the recent decades [3], leading to an increase in the older population, which, in turn, has led to an increase in the number of frail people. Therefore, prevention and treatment of frailty have generated great interest since it represents a challenge for health systems [4]. Several studies have shown that multimodal interventions, that is, those targeting multiple factors are effective both in primary (prevent or delay the debut of frailty) and secondary prevention (treating the syndrome when it has appeared with the aim of reversing it as much as possible) of frailty, particularly in older individuals with multimorbidity and high health care utilization [5,6].

The current medical practice to assess, treat, and monitor frailty is performed through periodic visits of the patient to health care settings. Between these visits, there is no information about the changing status of the patient's condition, adherence to treatment, or the response to it. In recent years, there has been a digital transformation in health care models, and these new technologies could provide aids to measure, monitor, and treat frailty in the older population to prevent health adverse events. Different studies have been performed within the framework of information and communication technologies and frailty treatment and prevention, but the results obtained are not conclusive, given that the technologies used are very different, and the results are very diverse [7,8].

With this approach in mind, we developed the FACET (Frailty Care and Well Function) technological ecosystem to improve the effectiveness of management of frail patients. A technological ecosystem refers to a complex network of interrelated technologies, services, and stakeholders that interact and depend on each other to create a unified and functional environment. This system automatically collects information with high predictive power for adverse events at home. This information comprises speed, power in the lower limbs, and involuntary weight loss. This valuable information is complemented with questionnaires to assess nutritional and other data about the functional status, which are uploaded to the platform. Data are provided to the geriatric team through the FACET technological ecosystem.

This research work aims to evaluate the impact of the FACET technological ecosystem when supporting a comprehensive geriatric intervention and follow-up. To do so, a randomized pilot study was designed to assess whether the information provided by the remote sensors helps early detection of functional changes, thereby promoting an adjusted multimodal intervention in prefrail and frail older adults compared to usual care during a 6-month period.

Methods

Trial Design

This is a multicenter, randomized, simple blind intervention study, with a duration of 11 months (5 months for recruitment and 6 for intervention). This study was conducted from August 2018 to June 2019 in the geriatric outpatient clinics of 2 Spanish hospitals (Getafe University Hospital and Albacete University Hospital). This study was registered in ClinicalTrials.gov (NCT03707145).

Participants and Randomization

Participant Recruitment

Participants were recruited in-person and by telephone from several settings, including the geriatric outpatient clinic, acute care unit (after discharge), and primary care outpatient clinics. Initially, a prescreening interview was performed in the outpatient geriatric facility to identify potentially eligible patients who met the following inclusion criteria: (1) age ≥70 years, (2) living at home, (3) having a caregiver or supervision at home, (4) Barthel index ≥90, (5) having at least 4 comorbidities, (5) prefrail older adults meeting 1 or 2 Fried frailty phenotype criteria, and (6) frail older adults meeting 3, 4, or 5 Fried frailty phenotype criteria.

The exclusion criteria were (1) inadequate home infrastructure to host the required technology, (2) inability to understand how to use the FACET system, (3) illness that impedes performing the prescribed therapy or follow-up (acute myocardial infarction in the last 3 months, unstable cardiovascular disease, terminal disease of <12 months of life expectancy, other pathologies involving clinical instability), (4) alcohol/drug abuse, (5) living with a participant in the trial, and (6) participation in other interventional clinical studies at the same time. Finally, to

XSL•FO
RenderX

determine if a participant could or could not understand and use the FACET system, the following procedure was performed:

1. Provide indications on how to react to a reminder to perform an action through a mobile device of identical conditions to the one that will be used during the experimentation.
2. Provide indications on how to fill out a test with the mobile device.
3. Generation of a reminder to fill out a test different from the previously shown.

If the potential participant reacted to the alarm and completed the test by him/herself, he/she was considered a suitable candidate. Once it was verified that the participants met the eligibility criteria, the informed consent was signed, and then they were randomized to the intervention or the control group.

Settings and Locations Where the Data Were Collected

Recruitment, assessment, follow-up, and treatment were performed in parallel in 2 institutions: Getafe University Hospital and Albacete University Hospital. Researchers in charge of performing the intervention or collecting the data along with the follow-up were different. Although information gathered by the FACET technological ecosystem was revised every week by a nonblinded geriatrician, who after reviewing it, designed and performed the different needed interventions during the development of the study in the intervention group, data about the outcomes were collected through face-to-face interviews by another geriatrician who was blinded regarding the branch where the participant had been randomly allocated every 3 months.

Outcomes

Primary Outcomes

The primary outcomes of this study were as follows: (1) to assess whether the FACET monitoring system prevents the worsening of frailty status according to the Frailty Trait Scale-5 items (FTS-5) (Table S1 in Multimedia Appendix 1) and the Fried frailty phenotype criteria (Table S2 in Multimedia Appendix 1) at 3 and 6 months and (2) to assess whether the FACET monitoring system improves frailty status according to FTS-5 and the Fried frailty phenotype criteria at 3 and 6 months. The outcomes were assessed through the following criteria: frailty worsening and improvement.

Assessment of Primary Outcomes

Frailty Worsening

The worsening in frailty status was evaluated according to 2 different criteria: (1) changes in the score in FTS-5 (worsening of 2.5 points or more) [9] and worsening by 1 criterion according to Fried frailty phenotype criteria [10] and (2) by analyzing transitions from nonfrail to frail according to FTS-5 [11] and from prefrail to frail when Fried frailty phenotype criteria were used [12]. Measurements were performed at baseline, 3 months, and 6 months of follow-up.

Frailty Improvement

Improvements in frailty were quantified according to 2 sets of criteria: (1) within FTS-5, based on an improvement of 2.5 points or more [9] and decrease by at least 1 criterion of the Fried frailty phenotype criteria [10] and (2) analyzing transitions from frail to nonfrail according to FTS-5 [11] and either from frail to prefrail or robust or from prefrail to robust according to Fried frailty phenotype criteria [12].

Secondary Outcomes

The secondary outcomes were as follows:

1. Changes in quality of life assessed using EQ-5D-5L at the initial visit, at 3 months, and after 6 months of follow-up.
2. Changes in the use of health resources assessed by open questions on number of visits to the emergency room, number of visits to the primary care physician, number of visits to nursing staff, number of visits to primary care, number of falls, and hospital admissions in the last 6 months. These data were collected at the initial visit and after 6 months of follow-up.

Hypothesis

The combination of information and communication technologies improves the monitoring of the patients at home with an impact on the frailty status.

Variables

Data were collected from participants during face-to-face visits in outpatient clinics at baseline, 3 months, and 6 months, except for the cognitive test that was assessed at baseline and at 6 months. The following data were collected: (1) demographic data, (2) comorbidities, (3) medication, (4) Barthel index [13], (5) frailty phenotype criteria [12], (6) FTS-5 [11], (7) gait speed in 6 meters [14], (8) Mini Mental Status Examination [15], (9) EQ-5D-5L [16], and (10) use of health resources in the last 6 months, that is, number of emergency visits, number of hospital admissions, number of visits to primary care physician, number of visits to nursing staff, number of visits to specialist physician, and number of falls.

Technological Ecosystem Description

Overview of the FACET Monitoring System

The FACET monitoring system is a technological ecosystem developed jointly by the Getafe University Hospital and Universidad Politécnica of Madrid that lies on the technological substrate of a Europe Union-funded (European Institute of Innovation and Technology-Health) research and innovation project: FACET. This project produced a mature and low-cost home monitoring system that aims at preventing functional decline among prefrail and frail older adults by detecting early functional changes and by generating alerts to promote early interventions to prevent potential adverse events that may lead to disability and dependency. Furthermore, the FACET technological ecosystem provides a means to interconnect the geriatric care team with older adults (see Figure 1). It is important to remark that the technology under study was fully developed following a cocreation approach [17,18], integrating knowledge and experience from all relevant actors in the process: older adults, health care professionals, informal caregivers, and technology experts.

XSL•FO
RenderX

Figure 1. The FACET (Frailty Care and Well Function) system: components and users.

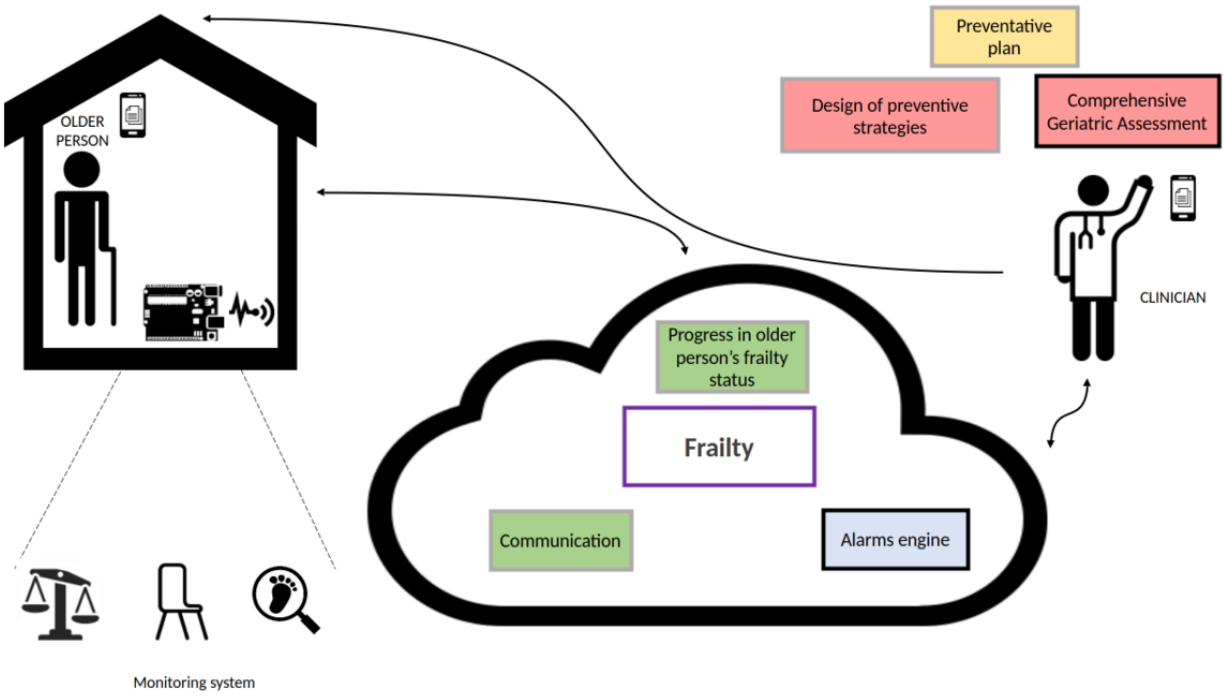

FACET Monitoring System

The FACET technology has 2 different components (Figure 1): (1) home monitoring subsystem for patients and (2) web interface for professionals.

Home Monitoring Subsystem for Patients

Mobile App for Older Adults

The mobile app (Figure 2) in the monitoring subsystem offers the following functionalities: continuous frailty follow-up using a home monitoring kit that produces information (gait speed, power in the lower limbs, involuntary weight loss) that is later processed to trigger potential deterioration alerts [19,20]; access to a customized therapeutic multicomponent intervention provided by the nonblinded geriatrician (medical treatment, VIVIFRAIL physical activity program [21], and nutritional recommendations); retrieving their own data on usage; communication with a geriatrician via asynchronous channels; notifications on pertinent alarms related to health to activate early interventions, with the goal of preventing disability; and reminders to perform the tests that must be performed at home. The participants in the intervention group accessed the mobile app through a tablet with 4G connection, provided by the study.

Figure 2. Mobile app in this study.

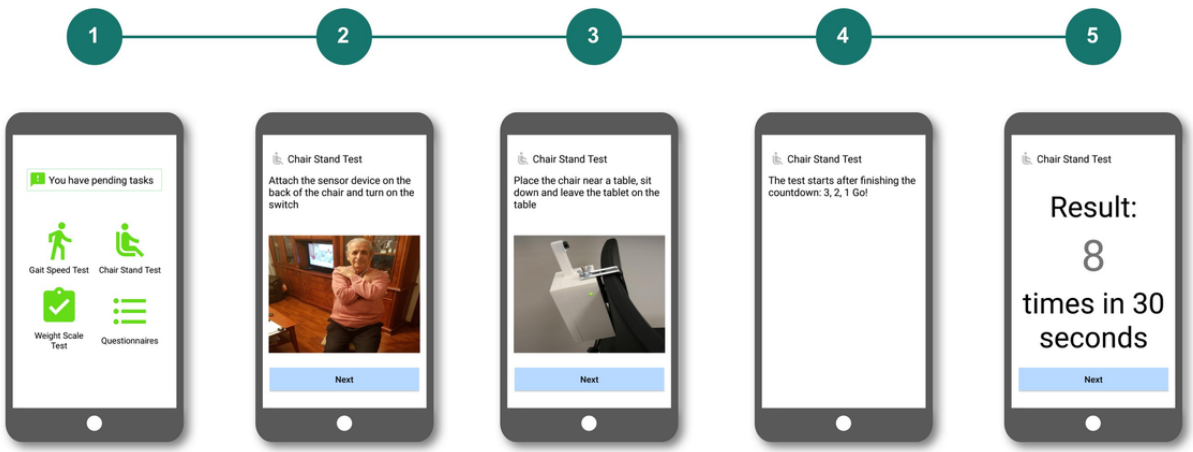

XSL•FO
RenderX

Monitoring System

The interaction with the home monitoring subsystem is handled by a mobile app that acts as a guiding element to the older person, as a data concentrator (Bluetooth connection with the monitoring kit) (Figure 1) and as data input point, not only enabling the older adults using the sensors but also completing a set of questionnaires to enrich the information handled by the clinical professionals. The home monitoring kit has been designed to measure variables with high predictive values for adverse events. This kit consists of a gait-speed sensor (Figure 3) [20], a sensor to indirectly (through the chair stand test) measure power in the lower limbs (Figure 4) [19], and a wireless commercial weight scale to measure involuntary weight loss. This information was enriched to build a short comprehensive geriatric assessment thought different questionnaires, which were tailored to language appropriate for the study's target population. The questionnaires that patients complete through the monitoring system is based on the Fried frailty phenotype criteria [12], Mini Nutritional Assessment [22], and Barthel index [23]. Information collected by this home monitoring kit is processed to trigger potential deterioration alarms.

Figure 3. Gait speed sensor.

XSL•FO
RenderX

Figure 4. Lower limbs power sensor.

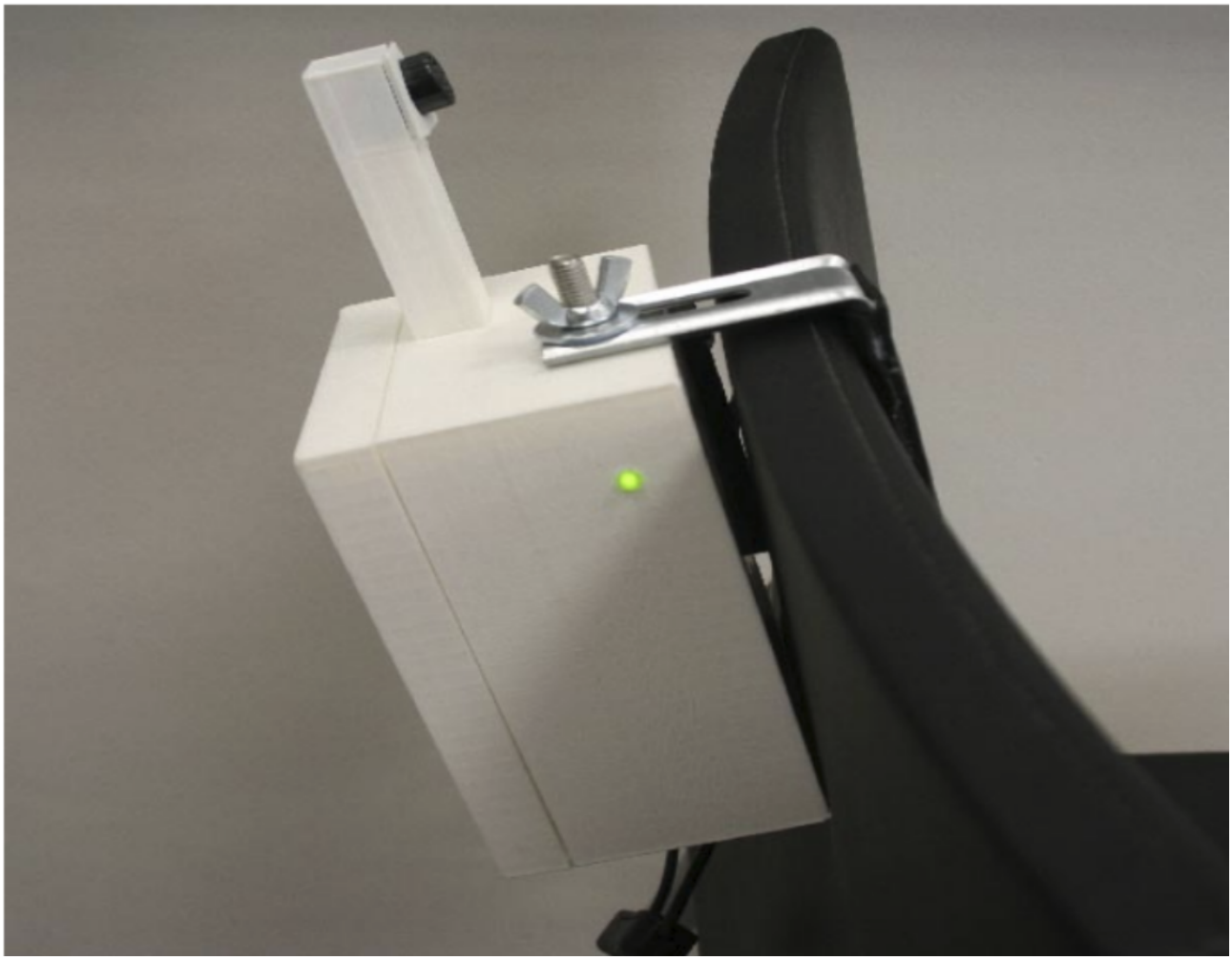

Web Interface for Professionals

The web interface provides essential infrastructure to offer functionalities to health care professionals. These functionalities include storing and accessing clinical information, asynchronous messaging between patients and the geriatrician team, displaying alarms based on patient monitoring results, tracking patient's clinical progress (questionnaires, functional test, etc), as well as prescribing and modifying treatments.

Sample Size

There are no similar studies allowing for a priori estimation of the necessary sample size. Therefore, an initial recruitment target of 90 participants was set based on the following criteria:

1. There are 2 main groups of interest: frail and prefrail individuals.
2. Based on commonly used standards, the recruitment goal is set at 20 participants per group of interest, totaling 40 participants for the control group and 40 for the intervention group.
3. Assuming a 10%-15% loss from the above estimation, the target sample size is established at 90 participants (N=90).

Randomization

For participant allocation into either the control or the intervention group, a stratified randomization by age (70-85 years, >85 years), sex (male/female), diagnosis (frail and prefrail among Fried frailty phenotype criteria), and educational level (higher education, illiterate, others) was performed to ensure that the 2 research arms were properly balanced. Participants were allocated using the MINIM tool [24] configured according to the study needs. Randomization was performed by the team members not directly involved in the development of the clinical trial. A recruitment target of 90 participants was established: 44 participants were allocated to the control group and 46 to the intervention group.

Three professional roles were involved in this study:

1. Nonblinded geriatricians to check and verify participant's eligibility based on the inclusion and exclusion criteria, explain the study details and obtain informed consent before randomization, monitor progress through the FACET system, and adjust the treatment as needed along the study.
2. Blinded geriatricians who participated in the prerandomization tasks of the study. They did not receive information about the arm to which the patient was randomly allocated. They also recorded the participant's evaluations at the baseline visit, 3 months, and 6 months. These researchers did not have access to the data from the FACET monitoring system.
3. Biomedical engineers conducted training sessions with the participants regarding the use of the FACET monitoring

XSL•FO
RenderX

system, installation of technology at home, and resolution of technical problems during the development of the study.

Interventions

After signing the informed consent, the participants were allocated into the control or the intervention group. The control group (n=44) received usual geriatric care through classical ways (comprehensive geriatric assessment; adjustment of polypharmacy; physical, cognitive, and nutritional prescription done face-to-face in classical patient visits). Participants in the intervention group (n=46) received the usual health care by a geriatric team, but they were supported by the information provided by the FACET monitoring system.

Participants were assessed to collect the variables previously described at baseline and after 3 and 6 months of follow-up. After the baseline evaluation, a treatment plan was designed with the following core components: medication adjustment, prescription of physical exercise based on the VIVIFRAIL guide [21], and dietary recommendations. In the intervention group, the participants were periodically and remotely supervised by their nonblinded geriatrician who scheduled the questionnaires and tests that the participants must perform at home to monitor the progress in their conditions. Frailty scale, chair stand test, gait speed (2.4 meters), and weight measurements were performed once weekly. Barthel index measurements and frequently asked questions were administered once every 2 weeks, and the Mini Nutritional Assessment-Short Form was administered once monthly. These sets of information were captured by the system and stored in the project-dedicated server. The system provides alerts to the clinician (nonblinded geriatrician) when pre-established changes were detected.

Nonblinded geriatricians checked the platform daily to see if any alarm had been generated. In case an impairment was detected, the geriatrician would call by phone to check the patient's health status and assess them. The nonblinded geriatrician, after this phone call, could provide an appointment to meet the patient face-to face if needed. When clinically indicated, he/she made changes to the treatment. Moreover, the participants and the nonblinded researcher could contact through a basic asynchronous communication module along the study. Those in the control group followed the usual guidelines without technological monitoring.

Technological Revisions and Updating

The technological solution was frozen during the intervention, which only fixed malfunctions without interfering with the functionalities and services provided.

Statistical Analysis

Statistical analysis was performed on an intention-to-treat basis. Summary statistics were presented as mean (SD) and n (%). Mann-Whitney and chi-square tests were performed to verify the hypothesis of the randomization scheme. To assess the main and secondary objectives, we compared both interventions (with and without technology) by using multivariate linear regression models for continuous response, multivariate logistic models for dichotomous response (improvement and worsening events), and Poisson models for count responses (ie, visits to the doctor). The analysis was adjusted for basal functional status (FTS-5) and gender. All the analyses were performed using R software (R Foundation for Statistical Computing) for Windows version 4.1.2. Statistical significance was set at $P<.05$.

Ethics Approval

This study followed the principles of the Declaration of Helsinki and was approved by the Getafe Hospital and Albacete Hospital Clinical Research ethics committees. This study was approved by the Getafe Hospital Ethics Committee (PY: 17/85). All participants provided written informed consent before screening (Multimedia Appendix 2). Participants' data were anonymized, and data correspondence was stored in a secure digital file supervised by a study engineer. There was no financial compensation.

Results

Participant Recruitment

Of the 281 individuals who were initially evaluated, 191 were excluded because they did not meet the inclusion criteria (n=133), they refused to participate in the study (n=36), and of other reasons (n=22). Recruitment ceased when 90 participants were enrolled. Randomization resulted in the allocation of 44 participants to the control group and 46 participants to the intervention group. In the third month of follow-up, 3 participants from the control group and 7 in the intervention group refused to continue in the study. In the follow-up at 6 months, 2 participants from the intervention group dropped out. Finally, from the original sample, 41 participants in the control group and 37 in the intervention group completed the full study (Figure 5).

XSL•FO
RenderX

Figure 5. CONSORT (Consolidated Standards of Reporting Trials) flowchart diagram. ITT: intention-to-treat.

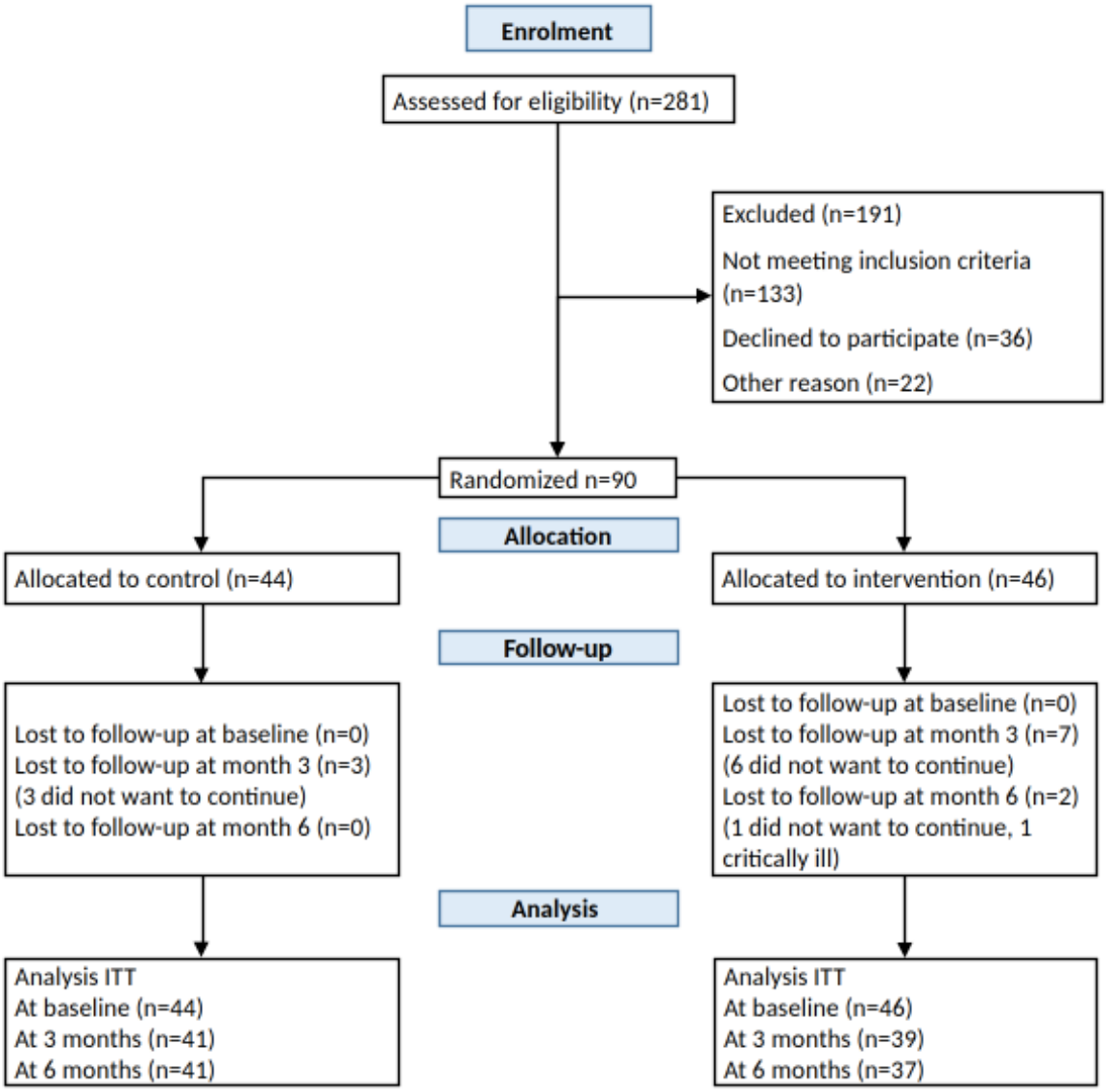

Baseline Data

The baseline characteristics of the participants are summarized in Table 1. Both groups exhibited similar baseline traits. The mean age of our study population was 82.33 (SD 5.91) years. The majority were women (65/90, 72%), had limited education (either no education or only primary education), and did not use technology daily. Participants in both groups showed independence in basic activities of daily living, with a mean Barthel index of 94.11 (SD 3.72). Participants met a mean of 2.73 (SD 0.86) criteria, and 50% (45/90) were categorized as frail and 50% (45/90) as prefrail. The mean score of FTS-5 was 22.89 (SD 6.14). The Charlson index data indicated that our patients exhibited high levels of comorbidity. Additionally, the recruited patients were characterized by polypharmacy. Data on functional status, gait speed, Short Physical Performance Battery, and Timed Up and Go indicated mild levels of physical impairment. The mean cognitive status values indicated very mild cognitive impairment. No significant differences were observed between the 2 groups, confirming the adequacy of the randomization process.

XSL•FO
RenderX

Table 1. Overall group characteristics.

	Total (N=90)	Intervention group (n=46)	Control group (n=44)	*P* value
Age (years), mean (SD)	82.33 (5.91)	82.11 (5.42)	82.56 (6.43)	.77
Males, n (%)	25 (28)	14 (30)	11 (25)	.56
Studies, n (%)				.14
No education	32 (36)	20 (43)	12 (28)	
Primary education	43 (48)	20 (43)	23 (53)	
Secondary education	11 (12)	5 (11)	6 (14)	
Tertiary education	3 (43)	1 (32)	2 (55)	
Experience with technology, n (%)				.32
Never	56 (63)	30 (65)	26 (60)	
1-2 times	5 (66)	4 (999)	1 (32)	
Occasional use	6 (77)	3 (77)	3 (77)	
Daily use	22 (25)	9 (20)	13 (30)	
Charlson index, mean (SD)	5.43 (1.83)	5.64 (1.93)	5.17 (1.69)	.31
Number of drugs, mean (SD)	9.44 (3.74)	9.91 (3.91)	8.95 (3.58)	.40
Mini Mental State Examination, mean (SD)	26.81 (3.04)	26.98 (2.68)	26.64 (3.39)	.93
Mini Nutritional Assessment, mean (SD)	24.63 (3.73)	24.49 (3.82)	24.78 (3.66)	.74
Barthel index, mean (SD)	94.11 (3.72)	94.24 (3.94)	93.98 (3.51)	.81
Short Physical Performance Battery, mean (SD)	7.21 (2.54)	7.41 (2.36)	7.0 (2.72)	.46
Gait speed 6 m, mean (SD)	0.71 (0.22)	0.72 (0.25)	0.69 (0.2)	.92
Frailty phenotype criteria				
Number of criteria, mean (SD)	2.73 (0.86)	2.74 (0.80)	2.73 (0.92)	.84
Prefrail, n (%)	45 (50)	22 (48)	23 (52)	.67
Frail, n (%)	45 (50)	24 (52)	21 (48)	.67
Frailty Trait Scale-5 items				
Mean score, mean (SD)	22.89 (6.14)	23.40 (6.56)	22.38 (5.72)	.64
Frail, n (%)	27 (30)	14 (32)	13 (29)	.82
Nonfrail, n (%)	63 (70)	32 (68)	31 (71)	.82
EQ-5D-5L, mean (SD)	56.07 (25.3)	54.15 (24.78)	58.07 (25.92)	.36

Primary Outcomes and Estimation

Frailty Worsening

Changes in FTS-5 Score and Worsening of 1 criterion in the Fried Frailty Phenotype Criteria

When we compared the intervention and control groups along the time through changes in FTS-5, we observed a 77% reduction in the risk of deterioration at the limits of statistical significance (odds ratio [OR] 0.23, 95% CI 0.05-1.09; *P*=.06) at 3 months in the intervention group that reached statistical significance at 6 months of follow-up, with a 74% reduction in the risk of deterioration (OR 0.26, 95% CI 0.07-0.98; *P*=.04) (Table 2). When analyzing data based on 1-point worsening according to the Fried frailty phenotype criteria, the results showed that the intervention group had a 92% lower likelihood of worsening compared to that in the control group at 6 months of follow-up (OR 0.08, 95% CI 0.01-0.67; *P*=.02) (Table 2). As there were no events at 3 months in any of the 2 groups, it was not possible to assess the effect.

XSL•FO
RenderX

Table 2. Frailty worsening assessed by the Frailty Trait Scale-5 items (worsening of 2.5 points) and changes in 1 point in the Fried frailty phenotype criteria.

	Odds ratio (95% CI)	*P* value
Frailty Trait Scale-5 items		
Month 0-month 3	0.23 (0.05-1.09)	.06
Month 0-month 6	0.26 (0.07-0.98)	.04
Fried Frailty Criteria		
Month 0-month 3	None[a]	—[b]
Month 0-month 6	0.08 (0.01-0.67)	.02

[a]The frailty status of none of the participants worsened (Fried criteria) at 3 months of follow-up.
[b]Not available.

Transitions From Nonfrail to Frail According to FTS-5 and From Prefrail to Frail by Fried Frailty Phenotype Criteria

We did not find significant changes in frailty transitions at either 3 months or 6 months on either of the 2 scales (see Table 3).

Table 3. Transitions from nonfrail to frail by Frailty Trait Scale-5 items and transition from prefrail to frail by Fried frailty phenotype criteria.

	Odds ratio (95% CI)	*P* value
Frailty Trait Scale-5 items		
Month 0-month 3	0.52 (0.16-1.66)	.27
Month 0-month 6	0.76 (0.24-2.42)	.64
Fried Frailty Criteria		
Month 0-month 3	None[a]	—[b]
Month 0-month 6	0.43 (0.03-5.74)	.52

[a]The frailty status of none of the participants worsened (Fried criteria) at 3 months of follow-up.
[b]Not available.

Frailty Improvement

Frailty Status

There was a higher likelihood of improvement in frailty status through FTS-5 in the intervention group compared to the control group at 3 months, with an OR of 4.16 (95% CI 1.57-11.03; *P*=.004). The benefits were retained for 6 months, with an OR of 2.63 (95% CI 1.004-6.90; *P*=.047). According to Fried frailty phenotype criteria, there was an improvement in the intervention group compared to the control group, with an OR of 4.50 (95% CI 1.17-17.38; *P*=.03) at 3 months of follow-up. The improvement was not observed at the 6-month follow-up (Table 4).

Table 4. Frailty improvement assessed by changes in 2.5 points in Frailty Trait Scale-5 items and changes in 1 criterion by Fried frailty phenotype criteria.

	Odds ratio (95% CI)	*P* value
Frailty Trait Scale-5 items		
Month 0-month 3	4.16 (1.57-11.03)	.004
Month 0-month 6	2.63 (1.004-6.90)	.047
Fried Frailty Criteria		
Month 0-month 3	4.50 (1.17-17.38)	.03
Month 0-month 6	1.37 (0.37-5.03)	.63

XSL•FO
RenderX

Transitions From Frail to Nonfrail According to FTS-5 and From Prefrail to Robust or From Frail to Prefrail or Robust by Fried Frailty Phenotype Criteria

We did not observe statistically significant results in FTS-5 at 3 or 6 months of follow-up in the transitions from frail to nonfrail according to FTS-5 and from prefrail to robust or from frail to prefrail or robust by the Fried frailty phenotype criteria. However, when we analyzed transitions by the Fried frailty phenotype criteria, the results indicated a marginal higher likelihood of improvement in frailty status occurring in the intervention group compared to that in the control group, with an OR of 3.10 (95% CI 1.01-9.54; *P*=.049). These benefits did not persist at 6 months, with an OR of 1.50 (95% CI 0.54-4.13; *P*=.44) (Table 5).

These transitions are shown in more detail in Table 6 (FTS-5) and Table 7 (Fried frailty criteria).

Table 5. Transitions from frail to nonfrail by Frailty Trait Scale-5 items and transitions from prefrail to robust or from frail to prefrail or robust by Fried frailty phenotype criteria.

	Odds ratio (95% CI)	*P* value
Frailty Trait Scale-5 items		
Month 0-month 3	1.93 (0.60-6.21)	.27
Month 0-month 6	1.32 (0.41-4-19)	.64
Fried Frailty Criteria		
Month 0-month 3	3.10 (1.01-9.54)	.049
Month 0-month 6	1.50 (0.54-4.13)	.44

Table 6. Frailty transitions, as shown by Frailty Trait Scale-5 items.

	Worsening, n (%)	No changes, n (%)	Improving, n (%)
Month 0-month 3			
Control (n=41)	8 (20)	19 (46)	14 (34)
Intervention (n=39)	2 (5)	15 (39)	22 (56)
Month 0-month 6			
Control (n=41)	9 (22)	13 (32)	19 (46)
Intervention (n=37)	3 (8)	10 (27)	24 (65)

XSL•FO
RenderX

Table 7. Frailty transitions, as shown by the Fried frailty phenotype criteria.

	Robust, n (%)	Prefrail, n (%)	Frail, n (%)
Month 0-month 3			
Robust			
Control (n=0)	0 (0)	0 (0)	0 (0)
Intervention (n=0)	0 (0)	0 (0)	0 (0)
Prefrail			
Control (n=22)	2 (9)	20 (91)	0
Intervention (n=18)	2 (11)	16 (89)	0 (0)
Frail			
Control (n=19)	0 (0)	7 (37)	12 (63)
Intervention (n=21)	1 (5)	14 (67)	6 (29)
Month 0-month 6			
Robust			
Control (n=0)	0 (0)	0 (0)	0 (0)
Intervention (n=0)	0 (0)	0 (0)	0 (0)
Prefrail			
Control (n=22)	3 (14)	17 (77)	2 (9)
Intervention (n=19)	4 (21)	14 (74)	1 (5)
Frail			
Control (n=19)	0 (0)	9 (47)	10 (53)
Intervention (n=18)	1 (6)	9 (50)	8 (44)

In this same regard, as a whole, the mean improvement in Fried's criteria was marginally higher in the intervention group at 3 months (OR 0.34, 95% CI –0.05 to 0.74; *P*=.09) and 6 months (OR 0.42, 95% CI –0.02 to 0.87; *P*=.06), while it reached statistical significance when FTS-5 was used, both at 3 months (OR 2.85, 95% CI 0.92-4.77; *P*=.005) and 6 months (OR 2.10, 95% CI 0.07-4.14; *P*=.04) (Table S3 in Multimedia Appendix 1).

Secondary Outcomes

We did not find significant changes in any of the secondary outcomes, except for falls. We did not detect changes in the visits to the emergency room, hospitalizations, visits to primary care physician and nurse, number of falls, or quality of life (Table S4 in Multimedia Appendix 1).

Discussion

Principal Findings

The main finding of our study is that the use of the FACET technological ecosystem over a 6-month period prevents impairment in older adults and improves their frailty status, irrespective of the method used to assess it, and prevents the progression of frailty according to FTS-5. Additionally, our data suggest that this benefit is observed quite early with the use of the FACET technological ecosystem, and it is maintained after 6 months of follow-up, although with a tendency to moderate its impact.

FACET per se does not provide any intervention (changes in physical exercise, nutrition, drugs, etc), which is decided by the geriatrician, but only monitors the functional variables. Therefore, intensive monitoring that allows early detection of functional deterioration by generating alerts to the geriatrician is useful for obtaining benefits in this prefrail and frail population, no matter the type of intervention provided. This finding reinforces the concept of a low functional reserve classically linked to the presence of frailty [6]—a fact that highlights the need of an intervention as early as possible to avoid further deteriorations that are quickly developed in the absence of any intervention.

Although data are consistent across frailty scales, FTS-5 shows the most reliable data. FTS-5 assesses a continuous multisystemic gradient, which makes it a more sensitive scale to changes and enables better detection of biological dysfunction, ranging from identifying the most robust to the most vulnerable individuals [11,25]. FTS-5 exhibits a heightened sensitivity in identifying even the slightest alterations in frailty status, which could justify the results obtained. FTS-5 and Fried frailty phenotype criteria show a low agreement, as recently reported in a study that evaluated the different tools that assess frailty [26], raising the possibility that they capture different dimensions of frailty [27,28]. In this regard, some researchers postulate the existence of different types of frailty [27].

XSL•FO
RenderX

Comparison With Prior Works

All previous studies [8,29,30] conducted in this field differ from ours in their inability to integrate information acquired through novel technologies with comprehensive geriatric assessments, aiming to enhance the evaluation and intervention processes for frailty. Numerous clinical studies [8,31-33] have been performed to enhance frailty diagnosis and treatment through information and communication technologies. However, results from these studies vary significantly, reflecting a wide array of devices employed for these purposes. This diversity complicates drawing definitive conclusions about the practical use and implementation of new technologies in clinical settings [8].

Most studies [33-35] assess frailty based on the frailty phenotype, while others [36] employ the frailty index. A study similar to ours was conducted in the United States by Upatising et al [29], involving 205 older adults (100 men) with an average age of 80.4 years, recruited from primary and community care. That study aimed to assess whether 12-month home monitoring could prevent frailty and mortality in patients with clinical issues. The devices utilized included a remote surveillance system, a health guide placed in the home, and other peripheral equipment connected to the health care system. Parameters such as heart rate, blood pressure, oxygen saturation, glucose level, and weight were monitored at home. Results showed that home monitoring did not reduce the functional decline, as measured by frailty and mortality rates, among older adults. Moreover, the home monitoring system itself did not induce changes in patients' functional status. Instead, the study indicated a need for re-evaluating the organizational model. A major limitation of that study was the absence of a clinical intervention, as they only detect status that did not promote an intervention, while in our study, the detection of an impairment promoted an alert motivating a contact between the clinician and the patient. This fact may at least partially explain the differences in our findings. In fact, simply monitoring should not be expected to produce any effect without intervention.

Another randomized clinical trial conducted by Geraedts et al [31] investigated whether frail older individuals using a portable physical activity sensor showed increased adherence and efficacy in a personally tailored home-based physical activity program. Feedback was provided to the participants, including videos demonstrating the exercises. That study concluded that patients using these sensors exhibited superior mobility outcomes compared to the control group.

Strengths and Limitations

This study has both strengths and limitations. One of its primary strengths lies in its internal validity, given its status as a randomized controlled trial. The control and intervention groups were meticulously balanced and comparable at the baseline, enhancing the study's credibility. Moreover, this study stands as the first of its kind to assess the effectiveness of an intervention integrating a home monitoring system with a classically provided comprehensive geriatric assessment and intervention. This assessment evaluates cognitive, affective, and physical domains; nutritional status; and social circumstances of older individuals, while also attempting to reshape the existing organizational model. The participants were older adults, which provides the study with validity when implementing the results in clinical practice. Another strength lies in the cocreation of the technology with older individuals, ensuring high adherence to the FACET technological ecosystem [18].

However, our study is not without its limitations. There were several dropouts in the intervention group, potentially attributed to the low digital literacy of the participants, who might have felt overwhelmed by the technology. Another limitation stemmed from the challenge of maintaining blinding throughout the study. Some participants revealed their research group affiliation during evaluation visits, although the researcher remained unaware of the previous assessments.

Conclusion

Our findings demonstrate that the FACET technological ecosystem effectively helps in the early identification of changes in the functional status of prefrail and frail older adults, facilitating prompt clinical interventions, thereby improving health outcomes in terms of frailty and functional status, which may prevent disability and dependency. We believe that the integration of information and communication technologies such the one used in this study is an asset in routine clinical practice. Nevertheless, further studies are warranted to advance our understanding in this area.

Acknowledgments

We would like to thank all the partners participating in the FACET (Integrated supportive services/products to promote Frailty Care and Well Function, PGA 16003) project (Abbott, Atos, Genesis Biomed, GMV Quirón Prevención, and University of Aberystwyth). This research work was supported by the FACET project, funded by the European Institute of Innovation and Technology-Health.

Data Availability

Restrictions apply to the availability of the data used in this study due to the private nature of the health information contained. These data are not publicly available but are available from the authors upon reasonable request and with permission of Getafe University Hospital.

XSL•FO
RenderX

Authors' Contributions

LR-M generated the original idea and coordinated with the team. RPR, PAM-S, and EV-M conceptualized and materialized the technological approach. MV-A, RPR, PAM-S, MO-B, and LR-M designed the trial. RPR led and supervised the correct execution of the experiment, with support from PAM-S, MV-A, MO-B, PAM-S, and EV-M. JAC performed the statistical analysis. MV-A, MO-B, and LR-M analyzed the clinical results. All authors participated in writing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data.
[DOCX File , 49 KB - jmir_v26i1e58312_app1.docx]

Multimedia Appendix 2
Informed consent form.
[DOCX File , 41 KB - jmir_v26i1e58312_app2.docx]

Multimedia Appendix 3
CONSORT eHEALTH checklist (V.1.6.2).
[PDF File (Adobe PDF File), 119 KB - jmir_v26i1e58312_app3.pdf]

References

1. Hoogendijk EO, Romero L, Sánchez-Jurado PM, Flores Ruano T, Viña J, Rodríguez-Mañas L, et al. A new functional classification based on frailty and disability stratifies the risk for mortality among older adults: The FRADEA study. J Am Med Dir Assoc 2019 Sep;20(9):1105-1110. [doi: 10.1016/j.jamda.2019.01.129] [Medline: 30853426]
2. Rodríguez-Laso Á, García-García FJ, Rodríguez-Mañas L. Transitions between frailty states and its predictors in a cohort of community-dwelling Spaniards. J Am Med Dir Assoc 2022 Mar;23(3):524.e1-524.e11. [doi: 10.1016/j.jamda.2021.07.008] [Medline: 34389334]
3. European C. Underlying assumptions and projection methodologies. The 2021 Ageing Report 2021:1-238. [doi: 10.2765/733565]
4. Gigantesco A, Giuliani M. Quality of life in mental health services with a focus on psychiatric rehabilitation practice. Ann Ist Super Sanita 2011;47(4):363-372 [FREE Full text] [doi: 10.4415/ANN]
5. Mazya AL, Garvin P, Ekdahl AW. Outpatient comprehensive geriatric assessment: effects on frailty and mortality in old people with multimorbidity and high health care utilization. Aging Clin Exp Res 2019 Apr;31(4):519-525 [FREE Full text] [doi: 10.1007/s40520-018-1004-z] [Medline: 30039453]
6. Angulo J, El Assar M, Álvarez-Bustos A, Rodríguez-Mañas L. Physical activity and exercise: Strategies to manage frailty. Redox Biol 2020 Aug;35:101513 [FREE Full text] [doi: 10.1016/j.redox.2020.101513] [Medline: 32234291]
7. Schwenk M, Howe C, Saleh A, Mohler J, Grewal G, Armstrong D, et al. Frailty and technology: a systematic review of gait analysis in those with frailty. Gerontology 2014;60(1):79-89 [FREE Full text] [doi: 10.1159/000354211] [Medline: 23949441]
8. Gallucci A, Trimarchi PD, Abbate C, Tuena C, Pedroli E, Lattanzio F, et al. ICT technologies as new promising tools for the managing of frailty: a systematic review. Aging Clin Exp Res 2021 Jun;33(6):1453-1464. [doi: 10.1007/s40520-020-01626-9] [Medline: 32705589]
9. El Assar M, Carnicero JA, Angulo J, Cámara-Hernández V, García-García FJ, Rodríguez-Mañas L. Fat mass accounts for insulin resistance impact on functional decline and mortality in nondiabetic older adults. J Am Med Dir Assoc 2024 Mar;25(3):448-453. [doi: 10.1016/j.jamda.2023.09.015] [Medline: 37898163]
10. Singer JP, Diamond JM, Gries CJ, McDonnough J, Blanc PD, Shah R, et al. Frailty phenotypes, disability, and outcomes in adult candidates for lung transplantation. Am J Respir Crit Care Med 2015 Dec 01;192(11):1325-1334. [doi: 10.1164/rccm.201506-1150oc]
11. García-García FJ, Carnicero JA, Losa-Reyna J, Alfaro-Acha A, Castillo-Gallego C, Rosado-Artalejo C, et al. Frailty Trait Scale-Short Form: A frailty instrument for clinical practice. J Am Med Dir Assoc 2020 Sep;21(9):1260-1266.e2. [doi: 10.1016/j.jamda.2019.12.008] [Medline: 32005416]
12. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci actions Search in PubMed Search in NLM Catalog Add to Search 2001 Mar;56(3):146-156 [FREE Full text] [doi: 10.1016/b978-0-323-03432-6.50096-9]
13. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL index: a reliability study. Int Disabil Stud 1988;10(2):61-63. [doi: 10.3109/09638288809164103] [Medline: 3403500]

XSL•FO
RenderX

14. Graham JE, Ostir GV, Kuo Y, Fisher SR, Ottenbacher KJ. Relationship between test methodology and mean velocity in timed walk tests: a review. Arch Phys Med Rehabil 2008 May;89(5):865-872 [FREE Full text] [doi: 10.1016/j.apmr.2007.11.029] [Medline: 18452733]
15. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
16. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990 Dec;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9] [Medline: 10109801]
17. Pérez-Rodríguez R, Villalba-Mora E, Valdés-Aragonés M, Ferre X, Moral C, Mas-Romero M, et al. Usability, user experience, and acceptance evaluation of CAPACITY: A technological ecosystem for remote follow-up of frailty. Sensors 2021 Sep 27;21(19):6458. [doi: 10.3390/s21196458]
18. Villalba-Mora E, Ferre X, Pérez-Rodríguez R, Moral C, Valdés-Aragonés M, Sánchez-Sánchez A, et al. Home monitoring system for comprehensive geriatric assessment in patient's dwelling: system design and UX evaluation. Front Digit Health 2021;3:659940 [FREE Full text] [doi: 10.3389/fdgth.2021.659940] [Medline: 34713133]
19. Cobo A, Villalba-Mora E, Pérez-Rodríguez R, Ferre X, Escalante W, Moral C, et al. Automatic and real-time computation of the 30-seconds chair-stand test without professional supervision for community-dwelling older adults. Sensors (Basel) 2020 Oct 14;20(20):5813 [FREE Full text] [doi: 10.3390/s20205813] [Medline: 33066673]
20. Ferre X, Villalba-Mora E, Caballero-Mora M, Sanchez A, Aguilera W, Garcia-Grossocordon N, et al. Gait speed measurement for elderly patients with risk of frailty. Mobile Information Systems 2017;2017:1-11. [doi: 10.1155/2017/1310345]
21. Izquierdo M, Rodriguez-Mañas L, Sinclair A. Editorial: what is new in exercise regimes for frail older people - how does the Erasmus Vivifrail project take us forward? J Nutr Health Aging 2016;20(7):736-737 [FREE Full text] [doi: 10.1007/s12603-016-0702-5] [Medline: 27499307]
22. Rubenstein LZ, Harker JO, Salva A, Guigoz Y, Vellas B. Screening for undernutrition in geriatric practice: developing the short-form mini-nutritional assessment (MNA-SF). The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 2001 Jun 01;56(6):M366-M372. [doi: 10.1093/gerona/56.6.M366]
23. Pfeffer RI, Kurosaki TT, Harrah CH, Chance JM, Filos S. Measurement of functional activities in older adults in the community. J Gerontol 1982 May;37(3):323-329. [doi: 10.1093/geronj/37.3.323] [Medline: 7069156]
24. Saghaei M. An overview of randomization and minimization programs for randomized clinical trials. J Med Signals Sens 2011 Jan;1(1):55-61 [FREE Full text] [Medline: 22606659]
25. García-García FJ, Carcaillon L, Fernandez-Tresguerres J, Alfaro A, Larrion JL, Castillo C, et al. A new operational definition of frailty: the Frailty Trait Scale. J Am Med Dir Assoc 2014 May;15(5):371.e7-371.e13. [doi: 10.1016/j.jamda.2014.01.004] [Medline: 24598478]
26. Oviedo-Briones M, Laso ?, Carnicero JA, Cesari M, Grodzicki T, Gryglewska B, et al. A comparison of frailty assessment instruments in different clinical and social care settings: The Frailtools project. J Am Med Dir Assoc 2021 Mar;22(3):607.e7-607.e12. [doi: 10.1016/j.jamda.2020.09.024] [Medline: 33162359]
27. Huang S, Tange C, Otsuka R, Nishita Y, Peng L, Hsiao F, et al. Subtypes of physical frailty and their long-term outcomes: a longitudinal cohort study. J Cachexia Sarcopenia Muscle 2020 Oct;11(5):1223-1231 [FREE Full text] [doi: 10.1002/jcsm.12577] [Medline: 32558267]
28. Aguayo GA, Donneau A, Vaillant MT, Schritz A, Franco OH, Stranges S, et al. Agreement between 35 published frailty scores in the general population. Am J Epidemiol 2017 Aug 15;186(4):420-434 [FREE Full text] [doi: 10.1093/aje/kwx061] [Medline: 28633404]
29. Upatising B, Hanson, Kim, Cha, Yih Y, Takahashi P. Effects of home telemonitoring on transitions between frailty states and death for older adults: a randomized controlled trial. IJGM 2013 Mar:145. [doi: 10.2147/ijgm.s40576]
30. Dasenbrock L, Heinks A, Schwenk M, Bauer JM. Technology-based measurements for screening, monitoring and preventing frailty. Z Gerontol Geriatr 2016 Oct;49(7):581-595. [doi: 10.1007/s00391-016-1129-7] [Medline: 27637581]
31. Geraedts HAE, Zijlstra W, Van Keeken HG, Zhang W, Stevens M. Validation and user evaluation of a sensor-based method for detecting mobility-related activities in older adults. PLoS One 2015;10(9):e0137668 [FREE Full text] [doi: 10.1371/journal.pone.0137668] [Medline: 26361009]
32. Galán-Mercant A, Cuesta-Vargas AI. Mobile Romberg test assessment (mRomberg). BMC Res Notes 2014 Sep 12;7:640 [FREE Full text] [doi: 10.1186/1756-0500-7-640] [Medline: 25217250]
33. Galán-Mercant A, Cuesta-Vargas AI. Differences in trunk accelerometry between frail and nonfrail elderly persons in sit-to-stand and stand-to-sit transitions based on a mobile inertial sensor. JMIR Mhealth Uhealth 2013 Aug 16;1(2):e21 [FREE Full text] [doi: 10.2196/mhealth.2710] [Medline: 25098977]
34. Martínez-Ramírez A, Lecumberri P, Gómez M, Rodriguez-Mañas L, García FJ, Izquierdo M. Frailty assessment based on wavelet analysis during quiet standing balance test. J Biomech 2011 Aug 11;44(12):2213-2220. [doi: 10.1016/j.jbiomech.2011.06.007] [Medline: 21719016]
35. Fontecha J, Hervás R, Bravo J, Navarro FJ. A mobile and ubiquitous approach for supporting frailty assessment in elderly people. J Med Internet Res 2013 Sep 04;15(9):e197 [FREE Full text] [doi: 10.2196/jmir.2529] [Medline: 24004497]
36. Mitnitski AB, Mogilner AJ, Rockwood K. Accumulation of deficits as a proxy measure of aging. ScientificWorldJournal 2001 Aug 08;1:323-336 [FREE Full text] [doi: 10.1100/tsw.2001.58] [Medline: 12806071]

XSL•FO
RenderX

Abbreviations

FACET: Frailty Care and Well Function
FTS-5: Frailty Trait Scale-5 items
OR: odds ratio

Edited by S Ma; submitted 12.03.24; peer-reviewed by M Malaguarnera, JL Sanchez-Sanchez; comments to author 09.05.24; revised version received 21.05.24; accepted 18.06.24; published 22.10.24.

Please cite as:
Valdés-Aragonés M, Pérez-Rodríguez R, Carnicero JA, Moreno-Sánchez PA, Oviedo-Briones M, Villalba-Mora E, Abizanda-Soler P, Rodríguez-Mañas L
Effects of Monitoring Frailty Through a Mobile/Web-Based Application and a Sensor Kit to Prevent Functional Decline in Frail and Prefrail Older Adults: FACET (Frailty Care and Well Function) Pilot Randomized Controlled Trial
J Med Internet Res 2024;26:e58312
URL: https://www.jmir.org/2024/1/e58312
doi:10.2196/58312
PMID:39436684

©Myriam Valdés-Aragonés, Rodrigo Pérez-Rodríguez, José Antonio Carnicero, Pedro A Moreno-Sánchez, Myriam Oviedo-Briones, Elena Villalba-Mora, Pedro Abizanda-Soler, Leocadio Rodríguez-Mañas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Universal Digital Stress Management Intervention for Employees: Randomized Controlled Trial with Health-Economic Evaluation

Johanna Freund[1,2], PhD; Filip Smit[3,4,5], PhD; Dirk Lehr[6], PhD; Anna-Carlotta Zarski[1,2,7], PhD; Matthias Berking[1], PhD; Heleen Riper[5,8,9], PhD; Burkhardt Funk[10], PhD; David Daniel Ebert[2], PhD; Claudia Buntrock[1,11], PhD

[1]Department of Clinical Psychology and Psychotherapy, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]School of Medicine and Health, Technical University of Munich, Munich, Germany
[3]Department of Mental Health and Prevention, Trimbos Institute, Utrecht, Netherlands
[4]Department of Epidemiology and Biostatistics, Amsterdam University Medical Centers, Amsterdam, Netherlands
[5]Department of Clinical, Neuro and Developmental Psychology, VU University, Amsterdam, Netherlands
[6]Department of Health Psychology and Applied Biological Psychology, Leuphana University of Lueneburg, Lueneburg, Germany
[7]Division of eHealth in Clinical Psychology, Department of Clinical Psychology, Philipps University of Marburg, Marburg, Germany
[8]Department of Psychiatry, VU University, Amsterdam, Netherlands
[9]Amsterdam Public Health Research Institute, VU University, Amsterdam, Netherlands
[10]Institute of Information Systems, Leuphana University of Lueneburg, Lueneburg, Germany
[11]Institute of Social Medicine and Health Systems Research, Otto von Guericke University Magdeburg, Magdeburg, Germany

Corresponding Author:
Dirk Lehr, PhD
Department of Health Psychology and Applied Biological Psychology
Leuphana University of Lueneburg
Universitätsallee 1
Lueneburg, 21335
Germany
Phone: 49 4131 677 2720
Email: lehr@leuphana.de

Abstract

Background: Stress is highly prevalent and known to be a risk factor for a wide range of physical and mental disorders. The effectiveness of digital stress management interventions has been confirmed; however, research on its economic merits is still limited.

Objective: This study aims to assess the cost-effectiveness, cost-utility, and cost-benefit of a universal digital stress management intervention for employees compared with a waitlist control condition within a time horizon of 6 months.

Methods: Recruitment was directed at the German working population. A sample of 396 employees was randomly assigned to the intervention group (n=198) or the waitlist control condition (WLC) group (n=198). The digital stress management intervention included 7 sessions plus 1 booster session, which was offered without therapeutic guidance. Health service use, patient and family expenditures, and productivity losses were self-assessed and used for costing from a societal and an employer's perspective. Costs were related to symptom-free status (PSS-10 [Perceived Stress Scale] score 2 SDs below the study population baseline mean) and quality-adjusted life years (QALYs) gained. The sampling error was handled using nonparametric bootstrapping.

Results: From a societal perspective, the digital intervention was likely to be dominant compared with WLC, with a 56% probability of being cost-effective at a willingness-to-pay (WTP) of €0 per symptom-free person gained. At the same WTP threshold, the digital intervention had a probability of 55% being cost-effective per QALY gained relative to the WLC. This probability increased to 80% at a societal WTP of €20,000 per QALY gained. Taking the employer's perspective, the digital intervention showed a probability of a positive return on investment of 78%.

Conclusions: Digital preventive stress management for employees appears to be cost-effective societally and provides a favorable return on investment for employers.

Trial Registration: German Clinical Trials Register DRKS00005699; https://drks.de/search/en/trial/DRKS00005699

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48481) doi:10.2196/48481

KEYWORDS

economic evaluation; cost-effectiveness; cost-utility; cost-benefit; return-on-investment; employees; universal prevention; internet-based; stress management

Introduction

In Europe, up to 27% of the working population suffers from stress [1]. Stress is often caused by work-related factors, including high perceived work demands, little work control, and little support from coworkers and supervisors [2]. Stress is linked to numerous diseases, including mental health problems and psychiatric diseases [3]. Besides the great burden of disease for the individual, stress is linked to formidable costs for employers as well as society as a whole. Due to stress and stress-related disorders, individuals suffer from impairment at work and lower productivity, are absent for more days from work, and use health services at a higher rate [4]. Lazarus' transactional model of stress [5] delineates 2 distinct coping strategies. On the one hand, problem-focused coping involves actively influencing a stressful situation positively by using cognitive or behavioral efforts. On the other hand, emotion-focused coping primarily serves the purpose of managing challenging emotions, such as anger, disappointment, and sadness, in response to the specific situation.

The efficacy of stress management interventions has been confirmed in numerous meta-analyses of randomized trials in the general population [6] and in occupational settings [7]. However, high levels of psychological stress among employees are omnipresent and remain largely untreated [8]. Easily accessible and highly scalable digital interventions independent of time and place represent a promising approach to lowering the threshold for use compared with face-to-face interventions [9]. A recent meta-analysis provides evidence that digital or internet-based stress management interventions (iSMIs) are effective in terms of stress reduction in adults with small-to-medium effect sizes at posttreatment (Cohen d=0.43, 95% CI 0.31-0.51) [10]. In the workplace, in particular, a universal prevention approach is especially desirable because more employees can be reached without previous screening, which might be costly [11]. Compared with selective prevention with a focus on groups of people at increased risk or indicated prevention with a focus on individuals with elevated symptoms, universal prevention aims to reach the entire population regardless of any risk status [12]. A scalable digital intervention for the universal prevention of stress in the working population has the potential for substantial reach. Furthermore, with a focus on the entire working population, individuals who do not want to disclose symptoms due to fear of stigmatization can also be reached [13]. Concerning the efficacy of a universal iSMI for employees, the findings from a pragmatic randomized controlled trial indicate significantly reduced perceived stress with medium-to-large effect sizes both at posttreatment (Cohen d=0.71, 95% CI 0.51-0.91) and at 6-month follow-up (Cohen d=0.61, 95% CI 0.41-0.81) in the iSMI group when compared with a waitlist control condition (WLC) [14].

Yet, although the effectiveness of iSMI has been demonstrated, research on its economic merits is still limited. Evidence suggests that an indicated iSMI to proactively prevent the onset of stress in employees represents good value for money from both a societal and an employer's perspective [15,16]. However, no study has yet evaluated the cost-effectiveness of a universal iSMI in the working population. This study, thus, aimed to evaluate the cost-effectiveness and cost-utility of a universal iSMI for stress in employees compared with a WLC from a societal perspective and the cost-benefit from the employer's perspective over a time horizon of 6 months. The clinical effectiveness of this iSMI has already been established [14].

Methods

Study Design

We carried out the health-economic evaluation alongside a 2-arm randomized controlled trial (RCT), comparing the effects of a self-guided iSMI for stress prevention with a WLC. Detailed information about the study design can be found elsewhere [14]. We carried out and reported the health economic evaluation in accordance with the guidelines of the International Society for Pharmacoeconomics and Outcomes Research [17] and the Consolidated Health Economic Evaluation Reporting Standards statement [18] (more details in Multimedia Appendix 1).

Recruitment

Recruitment was carried out as part of the occupational health program of a large German health insurance company (BARMER) in a way similar to the intended implementation of the intervention in routine practice in the future. The goal was to recruit individuals from the general working population, not just individuals insured by this health insurance company. This was done primarily through reports in the member magazine of the health insurance company and the insurance company's occupational health and safety management staff, who informed the human resources departments of collaborating companies about their employees' possibility to participate in the study. Interested individuals who completed a web-based screening questionnaire and met eligibility criteria were asked to fill out the informed consent form. To best reflect the routine conditions, the inclusion and exclusion criteria were reduced to a minimum. Individuals who were included (1) were 18 years and older, (2) were currently employed, and (3) had internet access and a valid email address. Exclusion criteria only included (1) a risk of suicide as indicated by a score of greater than 1 on the Beck Depression Inventory suicide item [14] or (2) any diagnosis with psychosis or dissociative symptoms (self-reported). If participants were excluded from the study, they were provided with information about alternative treatment options available in routine care.

Overall, 396 participants were included in the study, with 198 randomized to either the iSMI or WLC condition. All

XSL•FO
RenderX

participants completed the assessment at baseline, while 313 participants (79%) provided data at 6-month follow-up. Study dropout was not statistically significantly associated with any sociodemographic characteristics or initial perceived stress level. The average participant was 41.76 (SD 10.09) years old, female (302/396, 76%), highly educated (285/396, 72%), and employed full-time (296/396, 75%) with a working experience of 17.58 (SD 10.36) years. Almost half of the participants worked in a management position (169/396, 43%).

Randomization and Masking

Overview

Study participants were randomly assigned to the invention or control group in a 1:1 ratio by an independent researcher not otherwise involved in the study. Randomization took place using a computer-based random numbers table (Randlist) to ensure equal sample sizes for both conditions. Detailed information about the randomization procedure can be found elsewhere [14]. During the randomization process, the assignment was hidden from the participants as well as the researchers involved in recruitment and study administration. After randomization, participants were not blind to the study conditions due to the nature of the intervention.

Control Condition

In both study conditions, the participants had full access to treatment as usual (TAU). We did not interfere in TAU. Rather, we tried to maintain a naturalistic TAU state in order to represent routine care as much as possible.

Intervention

Participants in the intervention group (IG) received the iSMI GET.ON Stress. This iSMI entails 7 regular modules and 1 additional booster session for reviewing the most relevant content. Each module consists of psychoeducation, strategies for problem-solving, emotion regulation techniques, and plans for the future. It was recommended to complete 1-2 modules per week. The iSMI is based on the transactional model of stress [14]. The training includes interactive education, exercises, testimonials, and audio and video files. The content of the intervention is tailored to the individual needs and interests of the participants since several choices were made available through different answer options. In order to integrate the new knowledge sustainably into everyday life, homework, behavior planning, and an online diary are parts of the intervention. There was no therapeutic guidance provided. However, the participants had the opportunity to receive automatic text messages on their mobile phones. Participants could choose between light support with 1 text message every other day or intensive support with 2 or 3 text messages per day. The text messages contained very short exercises that should be carried out in everyday life to support the transfer from training to real life. More details about the intervention can be found elsewhere [14]. In this study, IG participants completed, on average, 5.23 (SD 2.74) sessions. Out of 198 participants, 94 (47%) participants finished all 7 modules, while 66 (33%) participants completed the additional booster session.

Outcome Measurements

Health-Related Outcome

The health outcome in the cost-effectiveness analysis (CEA) was symptom-free status based on the Perceived Stress Scale (PSS) and defined as a score at a 6-month follow-up of 2 SDs below the baseline mean of the study population (22.65, SD 5.63) [19].

Quality-Adjusted Life Years

Quality-adjusted life years (QALYs) were used as a health outcome in the cost-utility analysis (CUA). QALYs were based on the 35-item version of the Assessment of Quality of Life (AQoL-8D), which was assessed at baseline and a 6-month follow-up and is a reliable and validated instrument [20]. In total, 8 dimensions of health-related quality of life were covered (ie, independent living, relationships, mental health, coping, pain, senses, self-worth, and happiness) and preference-based valuations of health states (utilities) on a scale of 0 (death) to 1 (perfect health) were generated, using the time trade-off method [21]. Cumulative QALYs gains over the study's follow-up time of 6 months were estimated by calculating the area under the curve (AUC) of linearly interpolated AQoL-8D utilities between measurement points to cover the whole follow-up period.

Costs

Resource Use and Costing

We used the Trimbos and Institute for Medical Technology Assessment "Treatment Inventory of Costs in Patients with psychiatric disorders" (TiC-P) questionnaire to collect data on health care use, patient and family costs, and productivity losses [22]. The TiC-P is a retrospective questionnaire with a 3-month recall period and has been used in a similar study [15]. Costs were expressed in Euro and indexed from 2011 to 2013, the year the study was conducted, based on the German consumer price index (index factor 1.04) [23]. Costs were converted to pound sterling (£) using the purchasing power parities reported by the Organization for Economic Cooperation and Development [24]. For the reference year 2013, €1 was equated to £0.85 (A conversion rate of 1.33 in 2013 can be used for the conversion from € to US $).

Intervention Costs

At the time of conducting the study, the market price of the digital intervention provided by the GET.ON Institute, a commercial health care service provider was €99 (£84; ie, US $108.62) per participant, including costs for text messages, costs for website maintenance and hosting, technical support, and overheads.

Health Care Costs

We used 2 German guidelines for calculating health care costs [25,26]. Health care costs on a per-participant level were based on available lists of unit costs [26]. Unit costs were as follows: €20.92 (£17.78; ie, US $22.95) for a visit to the general practitioner, €46.55 (£39.57; ie, US $51.07) for a session with a psychiatrist, and €81.44 (£69.22; ie, US $89.36) with a psychotherapist, respectively. Costs per contact for allied health

XSL•FO
RenderX

services (eg, physiotherapist) were valued at €17.08 (£14.52; ie, US $18.74). Hospital stays were computed at €335.52 (£285.19; ie, US $368.13) for an in-patient day in a psychiatric hospital and €306.41 (£260.45; ie, US $336.19) for an in-patient day in a hospital for psychosomatic medicine and psychotherapy. Costs were estimated by multiplying the units of resource use with corresponding unit costs. The costs of prescribed medication were based on Lauer-Taxe [27].

Patient and Family Costs

Out-of-pocket payments were directly obtained from participants. Costs for traveling were valued at €0.30 (£0.27; ie, US $0.33) per kilometer. Productivity losses from unpaid work (eg, household chores, shopping, and child care) and informal care were valued using the proxy good method (eg, price of a close market substitute: domestic help). The average gross wage of domestic help per hour was estimated at €18.33 (£15.58; ie, US $20.11) per hour.

Productivity Costs

Costs due to absenteeism (ie, days not worked) were valued according to the human capital method [28]. Lost working days due to absenteeism were valued at the gross average income of participants per day. Lost working days due to presenteeism (ie, reduced efficiency while at work) were computed by taking into account the number of working days for which the participant reported a reduced work performance weighted by an inefficiency score for those days (Osterhaus method) [29].

Statistical Analysis

Health-Related Outcome, QALYs, and Costs

Analyses were conducted according to the intention-to-treat principle using Stata (version 16; StataCorp) [30]. Missing data were imputed using multiple imputations by chained equations (MICE) using predictive mean matching to account for the skewed distribution of cost and utility data. The imputation model was stratified by the study arm and included demographic data (eg, age, gender, marital status, and education) alongside health-related outcome variables at baseline (eg, utility values and perceived stress). The number of imputed datasets was at least equal to the percentage of incomplete cases (ie, m=30). Analyses, as described below, were performed on each dataset separately, and results were pooled using Rubin's rules.

Economic Evaluation: Societal Perspective

Disaggregated and total costs from the employer's and societal perspectives, as well as QALYs per study group, were assessed with a set of ordinary least square regression equations, the latter adjusted for baseline utility values. Group differences in symptom-free status were tested using logistic regression.

From a societal perspective, the incremental cost-effectiveness ratio (ICER) was calculated as the extra costs per additional symptom-free participant or QALY gained, respectively. We bootstrapped seemingly unrelated regression equations (SURE) models to generate 2500 simulations of incremental cost and effect pairs while allowing for correlated residuals of the cost and effect equations and adjusting for potential confounders (eg, initial utility values in the effect equation). Based on the bootstrapped SURE models, bias-corrected and accelerated 95% CIs were obtained for incremental costs and effects. Bootstrapped cost and effect pairs were plotted on cost-effectiveness planes to graphically represent the uncertainty surrounding the ICERs. Cost-effectiveness acceptability curves (CEACs) were created to depict the probability of the intervention being cost-effective compared with the control condition for varying willingness-to-pay (WTP) thresholds.

To determine subgroups in which the intervention was particularly cost-effective, the net-benefit regression framework (NBRF) was used [31]. In the NBRF, the treatment dummy, prognostically relevant baseline characteristics, and their interactions are regressed on net benefit. Net benefit (NB) is defined as NB = (E*λ)–C, where E denotes the effects per participant (eg, QALYs), C denotes the costs per participant, and λ is the willingness-to-pay for a unit of effect (ie, €20,000 [ie, US $21,944] per QALY gained). Analyses were conducted using gender, age, education, marital status, work experience, previous health training or psychotherapy, level of perceived stress, resilience, agreeableness, psychological strain, and self-regulation competencies as independent variables.

Economic Evaluation: Employer's Perspective

From the employer's perspective, cost-benefit analyses were performed using 2 metrics: (1) net benefits (NB = benefits – costs; amount of money gained after costs are taken into account) and (2) return-on-investment (ROI = [(Benefits–Costs)/Costs] × 100%; percentage of profit per Euro invested), where costs are defined as intervention costs and benefits as the difference in productivity costs between the iSMI group and the control condition. The metrics were estimated by bootstrapping linear regression models (N=2500). The probability of a positive financial return was assessed by the proportion of positive estimates (eg, NB>0, ROI>0%).

Sensitivity Analyses

To test the robustness of the base case findings, we performed 3 sensitivity analyses. First, we applied Winsorizing, where extreme values of cost outliers (eg those above the 95th percentile) were replaced by the value at the 95th percentile. Second, we varied the costs of the intervention (ie, €299; ie, US $328.06) to reflect uncertainties about the actual market price. Third, we used another instrument, the EQ-5D-3L [32], to assess health-related quality of life instead of the AQoL-8D. While the first 2 sensitivity analyses were applied to both the societal and employer's perspectives, the latter was only done from the societal perspective.

Ethical Considerations

The study was approved by the ethics committee of the University of Marburg (Germany [Aktenzeichen 2014-30k (date: 08/14/2014]) and registered in the German clinical trials register (DRKS00005699) on December 12, 2014.

Results

Health-Related Outcomes and QALYs

Participants in the iSMI condition had a statistically significantly higher probability for symptom-free status at 6-month follow-up with an odds ratio of 5.14 (95% CI 3.23-8.18) compared with

XSL•FO
RenderX

WLC based on bootstrapped data. The mean of the cumulative QALYs was higher in the iSMI condition (0.321 QALYs, 95% CI 0.32-0.33) compared with the WLC (0.306 QALYs, 95% CI 0.30-0.31). Adjusted incremental differences in QALYs between the iSMI condition and the WLC were statistically significant (Δ(e)=0.015 QALYs, 95% CI 0.01-0.02).

Costs

Baseline costs were slightly higher in the iSMI condition (€2283, 95% CI 1916-2650 [US $2504.91, 95% CI 2102.24-2907.58]) but comparable to the WLC (€1936, 95% CI 1569-2303 [US $21.24.18, 95% CI 1721.51-2526.85]). The imputed mean 6-month cumulative per-participant costs (in €) separately for various cost categories by study condition are presented in Table 1. Direct medical costs, as well as patient and family costs, were similar in both groups. In the iSMI group, costs for absenteeism were higher than costs due to presenteeism. The opposite was seen in the WLC. Employer's costs were slightly higher in the WLC compared with the iSMI (incremental difference of €76, 95% CI −667 to 515 [US $83.39, 95% CI –731.83 to 565.06]). Average total costs were comparable in both groups, with €3195 (US $3505.55; 95% CI 2661-3729 [95% CI 2919.65-4091.46]) in the iSMI condition and €3233 (US $3547.25; 95% CI 2722-3744 [95% CI 2986.58-4107.92]) in the WLC, resulting in an incremental difference of €38 (US $41.69; 95% CI −771 to 695 [95% CI –845.94 to 762.55]) in favor of the iSMI group.

Table 1. Imputed mean cumulative per-participant costs (in €) by condition over a 6-month follow-up period.

	Intervention group (n=198)		Control group (n=198)		Incremental difference	
	Mean, €[a]	95% CI	Mean, €	95% CI	Mean, €	95% CI
Direct medical costs						
Intervention costs	99	__[b]	—	—	99	—
General practitioner	52	42 to 62	50	41 to 58	2	−11 to 16
Mental health care	126	67 to 185	227	168 to 286	−101	−184 to −18
Antidepressants	4	−5 to 14	15	5 to 25	−10	−24 to 4
Allied health services[c]	41	23 to 60	30	13 to 46	11	−13 to 36
Patient and family costs						
Over-the-counter drugs	20	13 to 28	26	19 to 34	−6	−16 to 5
Out of pocket expenses[d]	106	56 to 156	104	61 to 148	2	−63 to 66
Travel	11	6 to 16	15	10 to 20	−4	−11 to 4
Unpaid work	159	90 to 229	187	122 to 252	−28	−125 to 70
Informal care	330	195 to 466	194	92 to 297	136	−35 to 307
Domestic help	168	89 to 248	133	44 to 223	35	−88 to 159
Productivity costs						
Absenteeism	1217	897 to 1538	884	584 to 1184	334	−102 to 770
Presenteeism	859	616 to 1103	1368	1128 to 1609	−509	−852 to −166
Employer's perspective						
Intervention costs + productivity costs	2176	1747 to 2605	2252	1841 to 2663	−76	−667 to 515
Societal perspective						
Total societal costs[e]	3195	2661 to 3729	3233	2722 to 3744	−38	−771 to 695

[a]€1=US $1.10.

[b]Not available.

[c]For example, massage, physiotherapist, and occupational therapist.

[d]For example, allied health services without prescription.

[e]Includes all cost categories based on a bootstrapped (n=5000) linear regression model. Columns may not add up correctly due to rounding.

XSL•FO
RenderX

Economic Evaluation

Societal Perspective

Cost-Effectiveness

Table 2 shows the incremental costs, effects, mean ICERs, and the distribution of cost and effect pairs on the cost-effectiveness plane based on 2500 bootstrap simulations. Cost-effectiveness analysis revealed that the iSMI generated more symptom-free individuals (Δ[E]=0.39; 95% CI 0.29-0.48) at lower costs (Δ[C]=–€38 [£953; ie, US $1246.51]; 95% CI €–705 to €692 [ie, 95% CI US $–773.32 to $759.06]) relative to the WLC. With regard to the cost-effectiveness plane, 56% of the bootstrapped ICERs fell in the southeast quadrant, indicating a 56% probability that the intervention dominates WLC (Figure 1). The remaining 44% of ICERs fell in the northeast quadrant, demonstrating a 44% probability that the intervention leads to greater health gains but at higher costs than WLC. At a willingness-to-pay of €1500 (US $1645.37) per additional symptom-free person, the iSMI showed a probability of 94% being cost-effective compared with the WLC (Figure 2).

Table 2. Results from the societal perspective (main and sensitivity analysis) based on 2500 bootstrap simulations.

Outcome	Incremental costs, € (95% CI)	Incremental effects, points (95% CI)	Incremental cost-effectiveness ratio, € or points (95% CI)[a]	Distribution over the cost-effectiveness plane (%)			
				North-east quadrant[b]	South-east quadrant[c]	South-west quadrant[d]	North-west quadrant[e]
Main analysis							
Symptom-free status (0/1)	−38 (−705 to 692)	0.39 (0.29 to 0.48)	Dominant	44	56	0	0
QALYs[f] (range: 0-1)	−38 (−735 to 687)	0.015 (0.01 to 0.02)	Dominant	45	55	0	0
Sensitivity analysis 1[g]							
Symptom-free status (0/1)	−110 (−652 to 477)	0.39 (0.29 to 0.48)	Dominant	34	66	0	0
QALYs[f] (range: 0-1)	−110 (−652 to 477)	0.025 (0.01 to 0.04)	Dominant	34	66	0	0
Sensitivity analysis 2[h]							
Symptom-free status (0/1)	162 (−505 to 892)	0.39 (0.29 to 0.48)	€418	89	11	0	0
QALYs[f] (range: 0-1)	162 (−505 to 892)	0.025 (0.01 to 0.04)	€6,515	89	11	0	0
Sensitivity analysis 3[i]							
QALYs[f] (range: 0-1)	−38 (−735 to 687)	0.011 (0.00 to 0.02)	Dominant	45	55	0	0

[a]In accordance with ISPOR best practice guidelines on "Model Parameter Estimation and Uncertainty," we do not report any negative incremental cost effectiveness. Ratios (ICERs) since they are meaningless. Instead, we use the term "dominant," which means that the intervention has a higher impact and comparatively lower cost with the WLC.

[b]The northeast quadrant of the CE plane, indicating that intervention is more effective and more costly.

[c]The southeast quadrant of the CE plane, indicating that intervention is more effective and less costly.

[d]The southwest quadrant of the CE plane, indicating that intervention is less effective and less costly.

[e]The northwest quadrant of the CE plane, indicating that intervention is less effective and more costly.

[f]QALYs: quality-adjusted life years.

[g]Sensitivity analysis 1 analyses for winzorizing cost outlier to 95% percentiles.

[h]Sensitivity analysis 2 analyses adding intervention costs of €299 (US $327.98; instead of €99 [US $108.59]).

[i]Sensitivity analysis 3 analyses for EuroQol for quality-adjusted life years.

XSL•FO
RenderX

Figure 1. Scatterplot of 5000 replicates of the incremental cost-effectiveness ratio (mean differences in costs and symptom-free status) on the cost-effectiveness plane from the societal perspective: iSMI versus WLC.

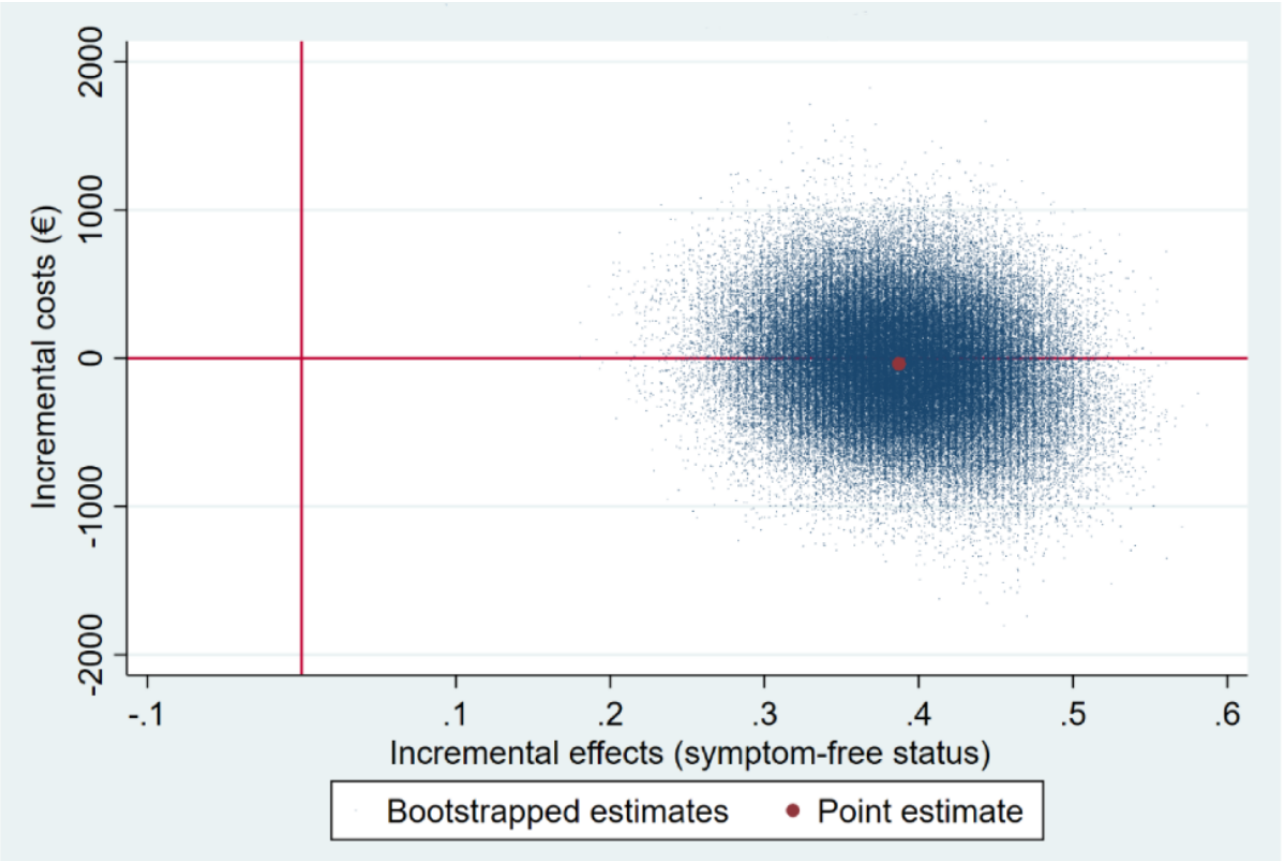

Figure 2. Cost-effectiveness acceptability curves (symptom-free status) from the societal perspective.

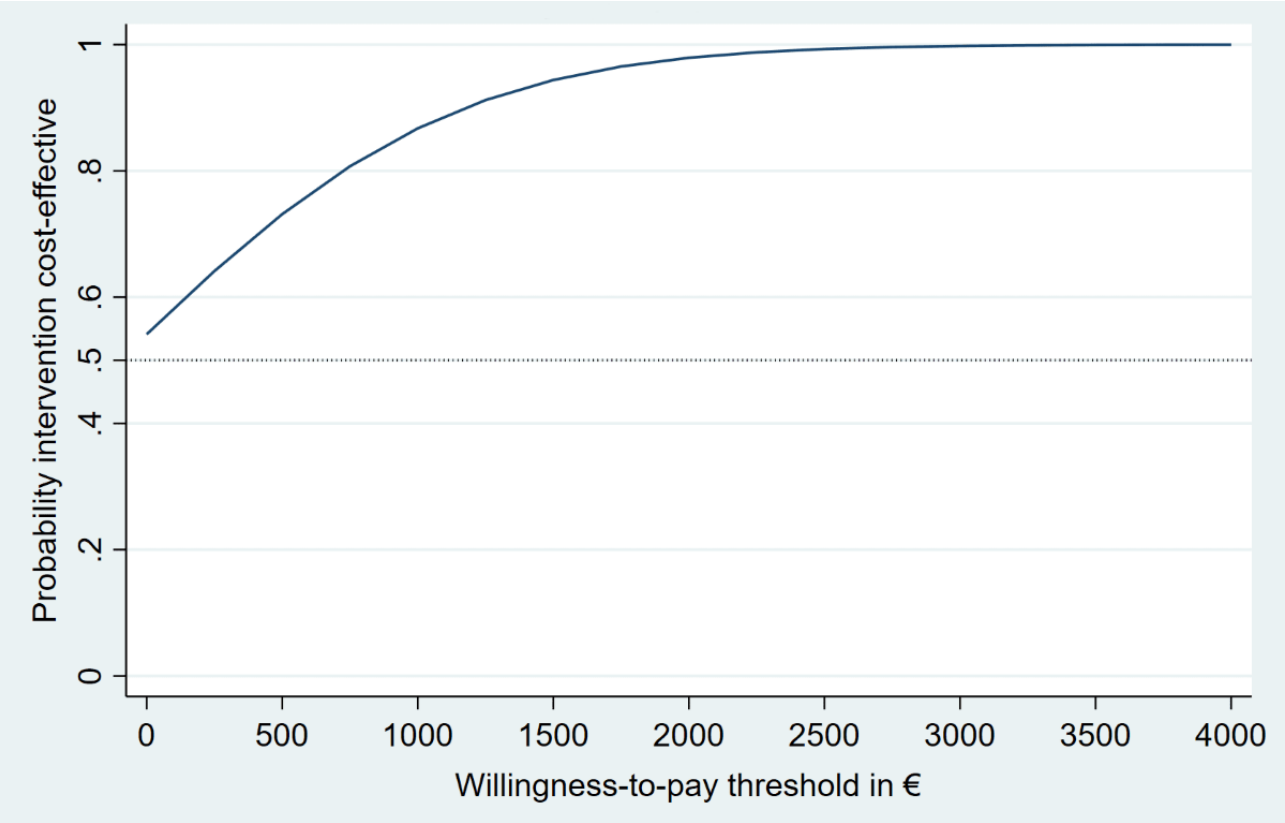

XSL•FO
RenderX

Cost-Utility

Cost-utility analysis revealed similar results (Table 2). Again, the iSMI was likely to be dominant relative to the WLC with more QALY gains at lower costs, resulting in a 55% probability that the iSMI is cost-effective at a societal willingness-to-pay threshold of €0 (Figure 3). This probability increased to 80% at a societal willingness-to-pay per QALY gained of €20,000 (US $21,938.20; Figure 4). Using NBRF, no subgroups were identified for which iSMI was significantly more (or less) cost-effective (all *P*-values ≥.05 for both WTPs of 0€ and €20,000 [US $21,938.20] per QALY gained, respectively).

Figure 3. Scatterplot of 5000 replicates of the incremental cost-effectiveness ratio (mean differences in costs and QALYs) on the cost-effectiveness plane from the societal perspective: iSMI versus WLC. QALYs: quality-adjusted life years.

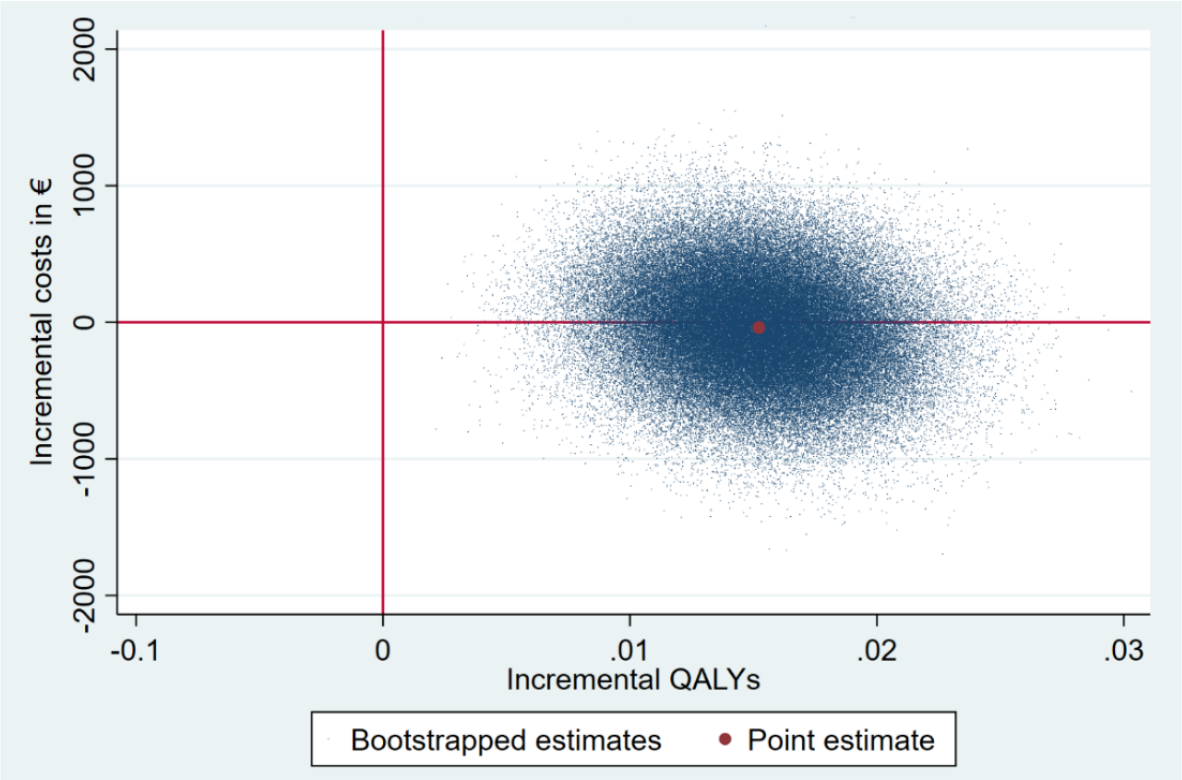

Figure 4. Cost-effectiveness acceptability curves (QALYs) from the societal perspective. QALYs: quality-adjusted life years.

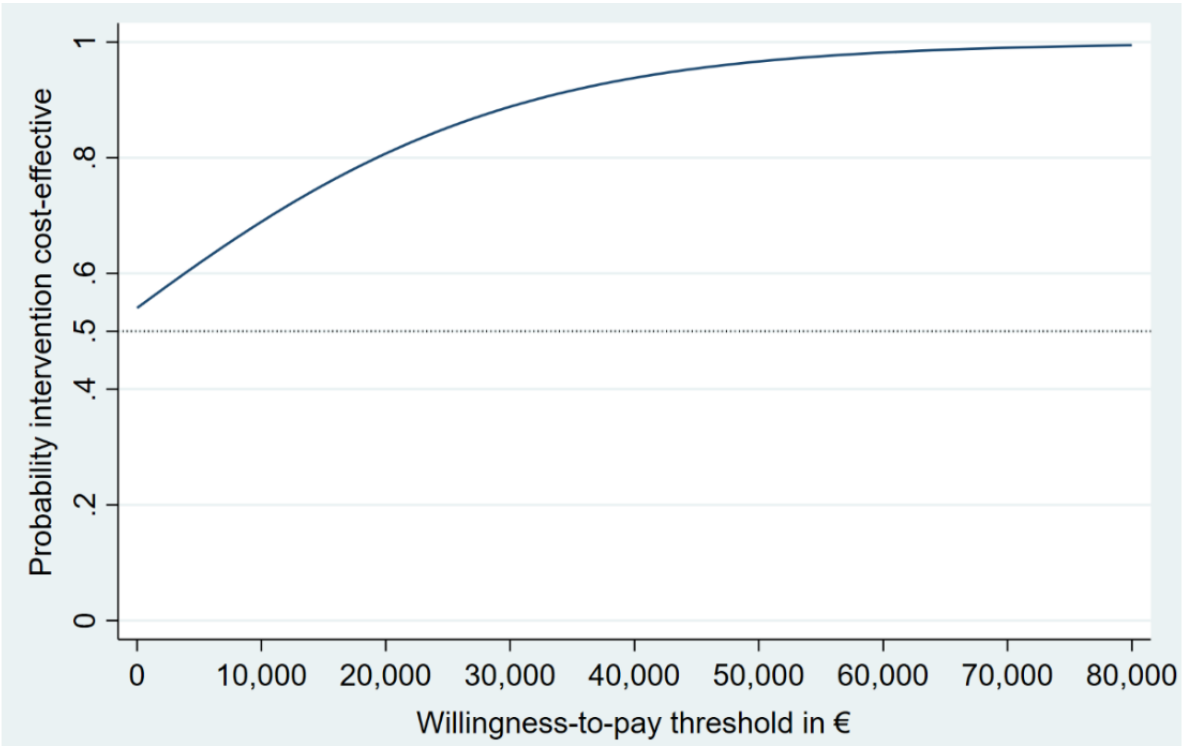

XSL•FO
RenderX

Employer's Perspective

The iSMI condition showed a net benefit per participant of €76 (£65; ie, US $83.37; 95% CI €−498 to 665 [ie, 95% CI US $–546.26 to $729.45]) and a benefit-to-cost ratio of 1.77 (95% CI €−4.03 to 7.72 [ie, 95% CI US $–4.42 to $8.47]). The ROI was 77% (95% CI −503% to 672%), respectively. The probability of a positive ROI was 78% for the iSMI condition (Table 3).

Table 3. Results from the employer's perspective (main and sensitivity analyses) of adjusted cost-benefit analyses based on 5000 bootstrapped linear regression models.

	Costs[a]		Benefits[b]		Financial returns				
	Total	95% CI	Total	95% CI	NB[c]	95% CI	ROI[d] (%)	95% CI	P[e] (%)
Main analysis									
Unguided intervention	99	__[f]	175	−399 to 764	76	−498 to 665	77	−50 to 672	78
Sensitivity analysis 1[g]									
Unguided intervention	99	—	217	−226 to 687	118	−325 to 588	120	−329 to 594	96
Sensitivity analysis 2[h]									
Unguided intervention	299	—	175	−399 to 764	−124	−698 to 465	−41	−233 to 156	14

[a]Includes intervention costs.

[b]Benefits are the difference in productivity costs between the intervention group and the control condition.

[c]NB: Net benefit linear regression models adjusted for baseline costs due to absenteeism and presenteeism.

[d]ROI: Return on investment linear regression models adjusted for baseline costs due to absenteeism and presenteeism.

[e]Probability of positive return on investment.

[f]Not available.

[g]Sensitivity analysis 1 analyses for winzorizing cost outlier to 95% percentiles.

[h]Sensitivity analysis 2 analyses adding intervention costs of €299 (ie, US $ 327.98; instead of €99 [US $108.59]).

Sensitivity Analyses

The results of the sensitivity analyses are summarized in Table 2 and Table 3. Winsorizing cost outliers led to a slightly higher probability that the intervention produces higher health gains at lower costs than WLC with regard to symptom-free status (66%) and QALYs (65%) at a societal WTP threshold of €0. The net benefit increased when cost outliers were winzorized, and the probability of a positive return on investment was 96% for the iSMI. Increasing intervention costs up to 3 times (€299 [US $327.98] instead of €99 [108.59]), this probability decreased to 33% regarding symptom-free status and QALYs gained at a WTP of €0. However, at a WTP of €20,000 [US $21,938.20] per QALY gained, the probability of being cost-effective was comparable to the main analysis (eg, 78%). Return on investment became negative when intervention costs were increased up to 3 times, and the probability of a positive financial return decreased to 14% for the iSMI. Using the EQ-5D-3L resulted in a probability of 55% regarding QALYs gained at a WTP of €0, identical to the main analysis.

Discussion

Principal Findings

This study evaluated the cost-effectiveness and cost-utility of a universal unguided digital stress management intervention for employees from a societal perspective and the cost-benefit from the employer's perspective compared with a waitlist control condition over a 6-month time horizon. From a societal perspective, the iSMI had a high probability of being cost-effective (eg, 80% at a WTP of €20,000 [US $21,938.20] per QALY gained). From an employer's perspective, the iSMI had a high probability of a positive return on investment with 78%.

Comparison With Previous Work

Evidence for economic evaluations of universal digital prevention of mental disorders is scarce. To our knowledge, this is the first economic evaluation of a universal iSMI to reduce stress in employees using a societal and an employer's perspective. The results of our study are in line with other economic evaluations of the same iSMI for stressed employees in the field of indicated prevention (PSS≥22) in which e-coaches provided personalized feedback throughout the intervention [15]. Kählke et al [15] demonstrated that more QALYs were generated for lower costs in the guided iSMI compared with a waitlist control condition, indicating a similar probability of 76% compared with our study that the intervention was cost-effective compared with WLC at a societal WTP of €20,000 [US $21,938.20] per QALY gained. A similar probability of 71% at a societal WTP of US $25,000 was also found by Lindsäter et al [33], who evaluated therapist-guided internet-based Cognitive Behavioral Therapy for stress-related disorders compared to WLC.

Results from the employer's perspective are comparable to the study by Ebert et al [16], who examined the cost-benefit of the

XSL•FO
RenderX

same iSMI as Kählke et al [15]. The analysis of Ebert et al [16] yielded a similar benefit-to-cost ratio of 1.6 (95% CI €−1.2-4.5 [US $–1.32 to $4.94]) compared with a benefit-to-cost ratio of 1.77 (95% CI €−4.0 to 7.7 [US $–4.39 to $8.45]) in this study. Our results from the employer's perspective are also in line with findings from a recent systematic review indicating that addressing mental health in employees improves both their well-being and productivity [34]. Regarding the ROI analyses, our findings compare favorably to a systematic review of face-to-face health promotion interventions at the workplace (12 RCTs) that showed, on average, a negative ROI (ROI=–0.22, 95% CI 0.27-0.16; min=−4.3, max=5) [35]. van Dongen et al [36] also showed a negative return on investment of a combined social and physical environmental intervention in office employees.

Limitations

The following limitations have to be considered. First, contrary to the pharmaco-economic guidelines recommending TAU as a comparator [37], the iSMI was compared with a wait-listed control condition, albeit with unrestricted access to usual care. Second, due to the limited time horizon of 6 months, longer-term costs and effects caused by chronic stress (eg, the onset of a new health disorder or staff turnover) could not be analyzed. Future studies should examine whether treatment effects and costs are sustained over a longer period of time or decline. Third, preference-based utility values were only evaluated at baseline and after 6 months, and thus, an immediate treatment effect was not assessed. Fourth, costs due to presenteeism were only assessed using the Osterhaus method. However, this method tends to overestimate costs because it assumes a 1:1 relationship between work hours lost and productivity losses, but that relationship might be better described as an elastic relationship, for example, 100% to 90%. Future studies should also include alternative methods (eg, Health and Labor Questionnaire method) to calculate costs due to presenteeism. Fifth, self-report questionnaires were used, which may have led to effects related to social desirability and response bias [38]. Future studies could consider claims data from health insurance companies. Sixth, the sample is characterized by a high proportion of female (302/396, 76.3%) as well as highly educated (285/396, 72%) participants, limiting the generalizability of the study findings. However, women with higher education are often the typical target group who take part in preventive internet-based interventions [39].

Clinical Implications

Evidence-based recommendations on the cost-effectiveness of interventions can help inform decision-makers from the societal and employer's perspective when choosing preventive stress interventions. The findings of this study support the hypothesis that a universal prevention approach could be a cost-effective strategy to reduce the adverse consequences of work-related stress besides its shown effectiveness [14]. However, it must be taken into account that the participants in the study already had a rather high initial stress level on average. That is, we cannot say whether the cost-effectiveness results will hold up if increased numbers of participants with low levels of stress participate in the intervention. An unguided universal iSMI might be appropriate from an economic perspective since there are no costs for screening compared with an iSMI for indicated prevention and additionally no costs for therapeutic guidance compared with guided interventions. Since health care professionals and available resources are often limited, unguided interventions have the potential to be implemented in routine occupational care on a large scale. Meta-analytic evidence shows that unguided interventions generate similar effects compared with guided interventions in individuals with low symptom severity [40]. However, the evidence on this is heterogeneous [41]. Therefore, future studies should directly compare the cost-effectiveness of guided compared with unguided iSMIs. From a health economic point of view, no moderators were revealed in NBRF analyses. However, this could also have been due to the small sample size and should be further researched in an individual participant data meta-analysis in the future.

Conclusions

A universal unguided digital stress management intervention appears to be cost-effective (ie, the health effects achieved represent good value for the invested money) as well as offering a favorable cost-to-benefit ratio (ie, the financial gains outweigh the intervention costs, so the return on investment is positive). Further research is needed to examine the long-term effects on cost-effectiveness and to compare digital stress management intervention with standard care. Unguided digital stress prevention carries the promise of being scalable, which would leverage the cost-effectiveness and positive return on investment of this type of universal stress prevention in the working population.

Acknowledgments

This publication was funded by the German Research Foundation (DFG). The funders did not have a role in the study design, data collection, analysis, and interpretation of results or the decision to publish the study results.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available but are available from the corresponding author on reasonable request by qualified scientific and medical researchers for legitimate research purposes. The data will be shared under a data-sharing agreement once the results of the health-economic evaluation are published. Requests should be sent to the last author.

XSL•FO
RenderX

Authors' Contributions

BF and MB obtained funding for the study. DDE and DL designed the randomized controlled trial. DDE, DL, HR, and MB developed the intervention content. ACZ was responsible for the recruitment of participants, data collection, and data curation, supervised by DL and DDE. CB developed the analysis plan for the health-economic evaluation, which was supported by FS. CB and JF accessed and verified the data. MB provided the analysis tool. CB conducted the analyses. CB, JF, and FS interpreted the data. JF drafted the manuscript and was supervised by CB. All authors participated in the critical review of the manuscript and approved the final manuscript.

Conflicts of Interest

BF and DDE are stakeholders of the GET.ON Institute (HelloBetter), which aims to implement scientific findings related to digital health interventions into routine care. MB is a cofounder and stakeholder of mentalis GmbH, a provider of digital mental health care products and services. DDE has served as a consultant and on the scientific advisory boards of Sanofi, Novartis, Minddistrict, Lantern, Schoen Kliniken, Ideamed, and German health insurance companies (BARMER, Techniker Krankenkasse) and a number of federal chambers for psychotherapy. ACZ receives license fees for a digital health application (DiGA) for sexual dysfunction implemented in routine care in Germany. She reports receiving fees for delivering presentations at scientific conferences and for producing expert videos. JF, FS, DL, HR, and CB report no conflicts of interest.

Multimedia Appendix 1
CHEERS (Consolidated Health Economic Evaluation Reporting Standards) checklist.
[DOCX File , 23 KB - jmir_v26i1e48481_app1.docx]

References

1. Eurofound. In: Developments in Working Life in Europe: EurWORK Annual Review 2016. Luxembourg: Publications Office of the European Union; 2017.
2. Nieuwenhuijsen K, Bruinvels D, Frings-Dresen M. Psychosocial work environment and stress-related disorders, a systematic review. Occup Med (Lond) 2010;60(4):277-286. [doi: 10.1093/occmed/kqq081] [Medline: 20511268]
3. Duric V, Clayton S, Leong ML, Yuan L. Comorbidity factors and brain mechanisms linking chronic stress and systemic illness. Neural Plast 2016;2016:5460732 [FREE Full text] [doi: 10.1155/2016/5460732] [Medline: 26977323]
4. Jensen I, Arapovic-Johansson Z, Aboagye E. The cost-effectiveness analysis of the productivity measurement and enhancement system intervention to reduce employee work-related stress and enhance work performance. Int J Environ Res Public Health 2022;19(4):2431 [FREE Full text] [doi: 10.3390/ijerph19042431] [Medline: 35206615]
5. Lazarus RS, Folkman S. Stress, Appraisal, and Coping. New York: Springer publishing company; 1984.
6. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
7. Heckenberg RA, Eddy P, Kent S, Wright BJ. Do workplace-based mindfulness meditation programs improve physiological indices of stress? A systematic review and meta-analysis. J Psychosom Res 2018;114:62-71. [doi: 10.1016/j.jpsychores.2018.09.010] [Medline: 30314581]
8. Hilton MF, Whiteford HA, Sheridan JS, Cleary CM, Chant D, Wang P, et al. The prevalence of psychological distress in employees and associated occupational risk factors. J Occup Environ Med 2008;50(7):746-757. [doi: 10.1097/JOM.0b013e31817e9171] [Medline: 18617830]
9. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
10. Heber E, Ebert DD, Lehr D, Cuijpers P, Berking M, Nobis S, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017;19(2):e32 [FREE Full text] [doi: 10.2196/jmir.5774] [Medline: 28213341]
11. McLaughlin KA. The public health impact of major depression: a call for interdisciplinary prevention efforts. Prev Sci 2011;12(4):361-371 [FREE Full text] [doi: 10.1007/s11121-011-0231-8] [Medline: 21732121]
12. Muñoz RF, Cuijpers P, Smit F, Barrera AZ, Leykin Y. Prevention of major depression. Annu Rev Clin Psychol 2010;6:181-212. [doi: 10.1146/annurev-clinpsy-033109-132040] [Medline: 20192789]
13. Couser GP. Challenges and opportunities for preventing depression in the workplace: a review of the evidence supporting workplace factors and interventions. J Occup Environ Med 2008;50(4):411-427. [doi: 10.1097/JOM.0b013e318168efe2] [Medline: 18404014]
14. Ebert DD, Franke M, Zarski AC, Berking M, Riper H, Cuijpers P, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021;23(12):e22107 [FREE Full text] [doi: 10.2196/22107] [Medline: 34941541]

XSL•FO
RenderX

15. Kählke F, Buntrock C, Smit F, Berking M, Lehr D, Heber E, et al. Economic evaluation of an Internet-based stress management intervention alongside a randomized controlled trial. JMIR Ment Health 2019;6(5):e10866 [FREE Full text] [doi: 10.2196/10866] [Medline: 31094355]
16. Ebert DD, Kählke F, Buntrock C, Berking M, Smit F, Heber E, et al. A health economic outcome evaluation of an internet-based mobile-supported stress management intervention for employees. Scand J Work Environ Health 2018;44(2):171-182 [FREE Full text] [doi: 10.5271/sjweh.3691] [Medline: 29144535]
17. Ramsey SD, Willke RJ, Glick H, Reed SD, Augustovski F, Jonsson B, et al. Cost-effectiveness analysis alongside clinical trials II-An ISPOR good research practices task force report. Value Health 2015;18(2):161-172 [FREE Full text] [doi: 10.1016/j.jval.2015.02.001] [Medline: 25773551]
18. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, ISPOR Health Economic Evaluation Publication Guidelines-CHEERS Good Reporting Practices Task Force. Consolidated health economic evaluation reporting standards (CHEERS)--explanation and elaboration: a report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. Value Health 2013;16(2):231-250 [FREE Full text] [doi: 10.1016/j.jval.2013.02.002] [Medline: 23538175]
19. Reis D, Lehr D, Heber E, Ebert DD. The German version of the perceived stress scale (PSS-10): evaluation of dimensionality, validity, and measurement invariance with exploratory and confirmatory bifactor modeling. Assessment 2019;26(7):1246-1259. [doi: 10.1177/1073191117715731] [Medline: 28627220]
20. Richardson J, Iezzi A, Khan MA, Maxwell A. Validity and reliability of the assessment of quality of life (AQoL)-8D multi-attribute utility instrument. Patient 2014;7(1):85-96 [FREE Full text] [doi: 10.1007/s40271-013-0036-x] [Medline: 24271592]
21. Richardson J, Sinha K, Iezzi A, Khan MA. Modelling utility weights for the assessment of quality of life (AQoL)-8D. Qual Life Res 2014;23(8):2395-2404. [doi: 10.1007/s11136-014-0686-8] [Medline: 24719017]
22. Bouwmans C, De Jong K, Timman R, Zijlstra-Vlasveld M, Van der Feltz-Cornelis C, Tan Swan S, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). BMC Health Serv Res 2013;13:217 [FREE Full text] [doi: 10.1186/1472-6963-13-217] [Medline: 23768141]
23. German Consumer Price Index. DESTATIS. URL: https://www.destatis.de/EN/Themes/Economy/Prices/Consumer-Price-Index/_node.html [accessed 2021-03-15]
24. Exchange rates (indicator). OECD. 2015. URL: https://tinyurl.com/76er8745 [accessed 2022-09-14]
25. Krauth C, Hessel F, Hansmeier T, Wasem J, Seitz R, Schweikert B. Empirische bewertungssätze in der gesundheitsökonomischen evaluation - ein Vorschlag der AG methoden der gesundheitsökonomischen evaluation (AG MEG). Gesundheitswesen 2005;67(10):736-746. [doi: 10.1055/s-2005-858698] [Medline: 16235143]
26. Bock JO, Brettschneider C, Seidl H, Bowles D, Holle R, Greiner W, et al. Calculation of standardised unit costs from a societal perspective for health economic evaluation. Gesundheitswesen 2015;77(1):53-61. [doi: 10.1055/s-0034-1374621] [Medline: 25025287]
27. Lauer Taxe. WEBAPO. 2009. URL: https://de.wikipedia.org/wiki/Lauer-Taxe [accessed 2014-06-24]
28. Drummond M, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press; 2015.
29. Osterhaus JT, Gutterman DL, Plachetka JR. Healthcare resource and lost labour costs of migraine headache in the US. Pharmacoeconomics 1992;2(1):67-76. [doi: 10.2165/00019053-199202010-00008] [Medline: 10146980]
30. Stata Version 16. College Station, TX: Stata Corp URL: https://www.stata.com/stata-news/news35-1/updates-to-stata16/ [accessed 2022-09-14]
31. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Econ 2002;11(5):415-430. [doi: 10.1002/hec.678] [Medline: 12112491]
32. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9] [Medline: 10109801]
33. Lindsäter E, Axelsson E, Salomonsson S, Santoft F, Ljótsson B, Åkerstedt T, et al. Cost-effectiveness of therapist-guided internet-based cognitive behavioral therapy for stress-related disorders: secondary analysis of a randomized controlled trial. J Med Internet Res 2019;21(9):e14675 [FREE Full text] [doi: 10.2196/14675] [Medline: 31586370]
34. de Oliveira C, Cho E, Kavelaars R, Jamieson M, Bao B, Rehm J. Economic analyses of mental health and substance use interventions in the workplace: a systematic literature review and narrative synthesis. Lancet Psychiatry 2020;7(10):893-910. [doi: 10.1016/S2215-0366(20)30145-0] [Medline: 32949521]
35. Baxter S, Sanderson K, Venn AJ, Blizzard CL, Palmer AJ. The relationship between return on investment and quality of study methodology in workplace health promotion programs. Am J Health Promot 2014;28(6):347-363. [doi: 10.4278/ajhp.130731-LIT-395] [Medline: 24977496]
36. van Dongen JM, Coffeng JK, van Wier MF, Boot CRL, Hendriksen IJM, van Mechelen W, et al. The cost-effectiveness and return-on-investment of a combined social and physical environmental intervention in office employees. Health Educ Res 2017;32(5):384-398. [doi: 10.1093/her/cyx055] [Medline: 28931167]

XSL•FO
RenderX

37. Sacristán JA, Abellán-Perpiñán JM, Dilla T, Soto J, Oliva J. Some reflections on the use of inappropriate comparators in CEA. Cost Eff Resour Alloc 2020;18:29 [FREE Full text] [doi: 10.1186/s12962-020-00226-8] [Medline: 32874138]
38. van DMT. Faking it: social desirability response bias in self-report research. Aust J Adv Nurs 2008;25(4):40-48.
39. Kelders SM, Bohlmeijer ET, Van Gemert-Pijnen JE. Participants, usage, and use patterns of a web-based intervention for the prevention of depression within a randomized controlled trial. J Med Internet Res 2013;15(8):e172 [FREE Full text] [doi: 10.2196/jmir.2258] [Medline: 23963284]
40. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
41. Domhardt M, Geßlein H, von Rezori RE, Baumeister H. Internet- and mobile-based interventions for anxiety disorders: a meta-analytic review of intervention components. Depress Anxiety 2019;36(3):213-224. [doi: 10.1002/da.22860] [Medline: 30450811]

Abbreviations

AQoL-8D: Assessment of Quality of Life
AUC: area under the curve
CEA: cost-effectiveness analysis
CEAC: cost-effectiveness acceptability curve
CUA: cost-utility analysis
ICER: incremental cost-effectiveness ratio
IG: intervention group
iSMI: internet-based stress management intervention
MICE: multiple imputations by chained equations
NB: net benefit
NBRF: net-benefit regression framework
PSS: Perceived Stress Scale
QALY: quality-adjusted life year
RCT: randomized controlled trial
ROI: return-on-investment
SURE: seemingly unrelated regression equations
TAU: treatment as usual
TiC-P: Treatment Inventory of Costs in Patients with Psychiatric Disorders
WLC: waitlist control condition
WTP: willingness-to-pay

Edited by S Ma; submitted 25.04.23; peer-reviewed by C de Miquel, F Lopez Segui, M Gasmi , S Kaur; comments to author 09.01.24; revised version received 29.01.24; accepted 28.06.24; published 22.10.24.

Please cite as:
Freund J, Smit F, Lehr D, Zarski AC, Berking M, Riper H, Funk B, Ebert DD, Buntrock C
A Universal Digital Stress Management Intervention for Employees: Randomized Controlled Trial with Health-Economic Evaluation
J Med Internet Res 2024;26:e48481
URL: https://www.jmir.org/2024/1/e48481
doi:10.2196/48481
PMID:

©Johanna Freund, Filip Smit, Dirk Lehr, Anna-Carlotta Zarski, Matthias Berking, Heleen Riper, Burkhardt Funk, David Daniel Ebert, Claudia Buntrock. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Behavior Change Techniques Used in Self-Management Interventions Based on mHealth Apps for Adults With Hypertension: Systematic Review and Meta-Analysis of Randomized Controlled Trials

You Zhou[1,2*], MSN; Si-Jia Li[1*], MSN; Ren-Qian Huang[3*], MM; Hao-Ming Ma[1], PhD; Ao-Qi Wang[1], MSN; Xing-Yi Tang[1], PhD; Run-Yuan Pei[1], BSN; Mei-Hua Piao[1], PhD

[1]School of Nursing, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China

[2]Department of Nursing, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, China

[3]Changhai Clinical Research Unit, Changhai Hospital, Naval Medical University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Mei-Hua Piao, PhD
School of Nursing
Chinese Academy of Medical Sciences
Peking Union Medical College
33 Badachu Road
Shijignshan District
Beijing, 100144
China
Phone: 86 13522112889
Email: piaomeihua@nursing.pumc.edu.cn

Abstract

Background: Hypertension has become an important global public health challenge. Mobile health (mHealth) intervention is a viable strategy to improve outcomes for patients with hypertension. However, evidence on the effect of mHealth app interventions on self-management in patients with hypertension is yet to be updated, and the active ingredients promoting behavior change in interventions remain unclear.

Objective: We aimed to evaluate the effect of mHealth app self-management interventions on blood pressure (BP) management and investigate the use of behavior change techniques (BCTs) in mHealth app interventions.

Methods: We conducted a literature search in 6 electronic databases from January 2009 to October 2023 for studies reporting the application of mHealth apps in self-management interventions. The Cochrane Risk of Bias (version 2) tool for randomized controlled trials was used to assess the quality of the studies. BCTs were coded according to the Taxonomy of BCTs (version 1). The extracted data were analyzed using RevMan5.4 software (Cochrane Collaboration).

Results: We reviewed 20 studies, of which 16 were included in the meta-analysis. In total, 21 different BCTs (mean 8.7, SD 3.8 BCTs) from 12 BCT categories were reported in mHealth app interventions. The most common BCTs were *self-monitoring of outcomes of behavior*, *feedback on outcomes of behavior*, *instruction on how to perform the behavior*, and *pharmacological support*. The mHealth app interventions resulted in a –5.78 mm Hg (95% CI –7.97 mm Hg to –3.59 mm Hg; $P<.001$) reduction in systolic BP and a –3.28 mm Hg (95% CI –4.39 mm Hg to –2.17 mm Hg; $P<.001$) reduction in diastolic BP. The effect of interventions on BP reduction was associated with risk factors, such as hypertension, that were addressed by the mHealth app intervention (multiple risk factors vs a single risk factor: –6.50 mm Hg, 95% CI –9.00 mm Hg to –3.99 mm Hg vs –1.54 mm Hg, 95% CI –4.15 mm Hg to 1.06 mm Hg; $P=.007$); the presence of a theoretical foundation (with vs without behavior change theory: –10.06 mm Hg, 95% CI –16.42 mm Hg to –3.70 mm Hg vs –4.13 mm Hg, 95% CI –5.50 to –2.75 mm Hg; $P=.07$); intervention duration (3 vs ≥6 months: –8.87 mm Hg, 95% CI –10.90 mm Hg to –6.83 mm Hg vs –5.76 mm Hg, 95% CI –8.74 mm Hg to –2.77 mm Hg; $P=.09$); and the number of BCTs (≥11 vs <11 BCTs: –9.68 mm Hg, 95% CI –13.49 mm Hg to –5.87 mm Hg vs –2.88 mm Hg, 95% CI –3.90 mm Hg to –1.86 mm Hg; $P<.001$).

XSL•FO
RenderX

Conclusions: The self-management interventions based on mHealth apps were effective strategies for lowering BP in patients with hypertension. The effect of interventions was influenced by factors related to the study's intervention design and BCT.

(J Med Internet Res 2024;26:e54978) doi:10.2196/54978

KEYWORDS

hypertension; mHealth; app; behavior change technique; systematic review; meta-analysis; mobile phone

Introduction

Background

Hypertension is a chronic disease characterized by persistent elevation of systemic arterial blood pressure (BP) and is considered the most important risk factor for cardiovascular diseases [1]. Epidemiological evidence suggests that approximately 31.1% of adults worldwide had hypertension in 2010 [2]. With the acceleration of population aging and the increase in lifestyle risk factors such as alcohol consumption, obesity, lack of physical activity, and unhealthy diets, the global prevalence of hypertension has shown a significant upward trend [3]. The Lancet Commission on hypertension has stated that primordial, primary, and secondary prevention of hypertension in the life course should be implemented to address the global burden of increased BP [4]. Therefore, people with risk of hypertension and patients with hypertension need to carry out BP management including lifestyle modifications addressing various lifestyle risk factors and medication management after diagnosis.

Mobile health (mHealth) refers to medical and public health practices supported by mobile devices, such as mobile phones, patient monitoring devices, and other wireless devices [5]. Considering its advantages, such as low cost and ease of use, the development of self-management solutions for patients with hypertension based on mHealth apps has become a frontier and hot spot of research. Several meta-analyses have demonstrated that mHealth interventions are effective for patients with hypertension and can improve clinical outcomes [6-11]. For example, a meta-analysis including SMS text messaging, smartphone apps, and website interventions showed that mHealth interventions, particularly smartphone apps, were associated with clinical reductions in systolic BP (SBP) [10]. Another meta-analysis showed that mHealth apps significantly improved medication adherence of patients with hypertension and increased patients' perceived confidence, treatment self-efficacy, acceptance of technology, and knowledge about health issues [11]. However, all the aforementioned studies lacked further analysis to identify the study design factors and active ingredients in the interventions or apps that promote self-management effectiveness.

Behavior change techniques (BCTs) are observable, replicable, and irreducible components of interventions designed to alter or redirect causal processes that regulate behavior [12]. By specifying BCTs in interventions, researchers can identify the active ingredients in interventions, synthesize evidence, replicate interventions, and even optimize them [13]. Since the pandemic, remote self-management interventions for patients with hypertension based on mHealth apps have exploded, but there are great differences among studies in terms of intervention design, app contents, and study findings, which raise potential concerns about the generalizability of mHealth app–based interventions for hypertension. The presence of BCTs offers the possibility of optimizing the study design of the aforementioned studies and improving the effectiveness of the interventions. Previous literature reviews have described the efficacy of BCTs in mHealth for self-management of diabetes [14]. However, to the best of our knowledge, there is limited literature exploring the active ingredients of mHealth interventions for patients with hypertension. The status and role of BCTs in mHealth app self-management interventions for patients with hypertension are unclear.

This Study

This study aimed to (1) evaluate the true effect of mHealth app self-management interventions on BP management in patients with hypertension, (2) comprehensively investigate the status of the use of BCTs in mHealth app interventions for patients with hypertension, and (3) further analyze what factors can influence the effect of interventions.

Methods

Overview

This systematic review and meta-analysis were carried out following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) [15], and the review protocol was registered in PROSPERO (CRD42023483746).

Search Strategy and Selection Criteria

With the assistance of a professional librarian, a systematic literature search was conducted in PubMed, Web of Science, Embase, American Psychological Association PsycINFO, CINAHL Plus, and Cochrane Central Register of Controlled Trials. In brief, we included randomized controlled trials (RCT) applying mHealth app self-management interventions among patients with hypertension and those published in English. Given that digital health apps started to become widely adopted in 2009, we set the search time from January 1, 2009, to October 15, 2023 [10]. We also performed a hand search of the reference lists of included studies and reviews related to the topic of this study to identify additional studies. The detailed search strategy is provided in Table S1 in Multimedia Appendix 2.

There were 5 eligibility criteria. The first was population: the study population was adults with a primary diagnosis of hypertension. We excluded studies involving minors, participants with pregnancy-related hypertension, and mixed patient populations without stratified results. The second eligibility criterion was intervention: the primary intervention was based on an mHealth app that can run on mobile or wearable

XSL•FO
RenderX

devices such as smartphones, tablets, or smartwatches, and the duration of interventions was >4 weeks or 1 month. We excluded studies involving an intervention mainly based on phone calls, text messages, and website programs, as well as apps with only notification, self-monitoring, and counseling functions (interventions based on such apps may lack sufficient complexity and comprehensiveness to accurately reflect the true effect of mHealth app interventions on BP self-management). The third eligibility criterion was control: the study consisted of at least 1 intervention group and 1 control group that incorporated usual or standard care. Studies in which the control group applied the mHealth app were excluded. The fourth eligibility criterion was outcome: the primary outcomes were SBP, diastolic BP (DBP), or both; the secondary outcomes were clinical or self-reported indicators related to hypertension self-management, including but not restricted to the proportion of BP in control and medication adherence. the fifth eligibility criterion was study design and publication type: peer-reviewed RCTs, pilot studies, and cluster RCTs. We excluded nonrandomized, noncontrolled, and observational studies; case reports; systematic reviews; meta-analyses; conference abstracts; protocols; preprints; and studies without full text. In addition, we also excluded studies in which the intervention was not well described (eg, low-quality studies with overly simplistic descriptions of intervention methods that failed to identify intervention contents) to allow for valid coding of BCTs.

Study Selection and Data Extraction

All identified studies were imported into EndNote X9 (Clarivate Analytics) software to remove duplicates and then imported into the Rayyan web platform (Rayyan Systems Inc) for eligibility review. The review process involved 2 rounds of screening: 2 reviewers (YZ and SJL) first screened the titles, abstracts, and keywords independently, and then the same 2 reviewers reviewed the full texts of the studies meeting the eligibility criteria to determine the list of included studies. Any disagreements in the review process were resolved through consultation and discussion with a senior reviewer (MHP).

Two reviewers (YZ and RQH) independently performed data extraction using a Microsoft Excel form developed following the guidelines in the Cochrane Handbook for Systematic Reviews of Interventions [16]. Any disagreements were resolved through rechecking the original research and discussion with another reviewer (MHP). The information recorded was (1) basic characteristics of publication: title, author, year, country of publication, and journal; (2) study details: study design, sample size, retention rate, intervention duration, detailed content of intervention and control group, and primary and secondary outcomes; (3) participants' characteristics: age, sex ratio, education, ethnicity, and diagnostic criteria of hypertension; (4) outcomes: duration of follow-up, critical outcomes in BP changes, including the mean, SDs, SEs, and 95% CIs in baseline and follow-up, and validated measurements used for self-reported outcomes; and (5) theoretical foundations and BCTs applied in mHealth app interventions.

Coding of BCTs

The coding of BCTs was performed according to the Taxonomy of BCTs (version 1) proposed by Michie et al [12], which summarized 93 BCTs into 16 categories and provided a standardized framework for identifying the BCTs used in behavior change interventions. In total, 2 reviewers (YZ and SJL) coded the BCTs independently after completing web-based training [17]. Evidence supporting BCT coding was from articles, supplementary materials, protocols, and secondary analysis publications. Any discrepancies in the coding process were resolved by discussion with another reviewer (MHP) until unanimity was achieved.

Risk of Bias and Grade of Evidence Assessment

The revised Cochrane Risk of Bias (version 2) tools for RCTs and cluster RCTs were used to assess the risk of bias in (1) the randomization process, (2) deviations from the intended interventions, (3) missing outcome data, (4) measurement of the outcome, and (5) selection of the reported result of the included studies [18]. The Grading of Recommendation, Assessment, Development, and Evaluation criteria were used to assess the quality of the overall evidence in (1) risk of bias, (2) inconsistency, (3) indirectness, (4) imprecision, and (5) publication bias, and the quality of evidence could be classified as high, moderate, low, or very low [19]. In total, 2 reviewers (YZ and RQH) reviewed and rated the studies independently, and any disagreements were resolved via discussion with another reviewer (MHP).

Data Synthesis and Analysis

A meta-analysis was performed to evaluate the pooled effect size of an mHealth app intervention on BP reduction. Given the potential bias caused by differences in BP levels between the intervention and control groups at baseline, we decided to include data on within-group changes in the mean and SD values of SBP and DBP in each group at follow-up in the analyses.

For studies in which the data were not available from articles, supplementary materials, and secondary analysis publications, the researcher attempted to contact the corresponding authors for necessary data. For studies that only reported SE or 95% CI or did not report the within-group changes in mean and SD values for BP, the researcher transformed the SE and 95% CI data and estimated the SD values according to the Cochrane handbook [16]. Studies in which analytical data were unavailable based on the aforementioned methods were excluded from the analysis.

Due to the considerable heterogeneity between the included studies, the random-effects model was considered appropriate to synthesize the effect sizes and SDs [20]. Heterogeneity was quantified using the Cochran Q test and Higgins I^2 statistics, and the I^2 values <25%, 25% to 75%, and >75% were considered low, medium, and high heterogeneity, respectively [21]. The bias of publication was evaluated by using the Egger test and visualization of the funnel plot.

Subgroup analyses were conducted to explore the impact of intervention and study design factors (such as the risk factors of hypertension that the mHealth app intervention addressed, the presence of a theoretical foundation, and intervention duration) and the number of BCTs on the effect size of an mHealth app intervention on BP levels and to evaluate the

XSL•FO
RenderX

possible sources of heterogeneity. Review Manager software (version 5.4.1) was used to perform meta- and subgroup analyses, and Stata software (version 15.1; StataCorp) was applied to draw the funnel plot.

Results

Literature Screening

The process of study identification and screening is outlined in Figure 1. A total of 1668 records were extracted from the initial literature search, and 5 records were identified from hand searching. After removing duplicates, 831 (49.67%) records were screened for titles and abstracts. Overall, 35 (4.2%) full-text studies were assessed for eligibility based on the inclusion and exclusion criteria. Finally, 20 (2.4%) studies were considered eligible and included in the systematic review, and 16 (1.9%) studies with available data were included in the meta-analysis.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process. RCT: randomized controlled trial.

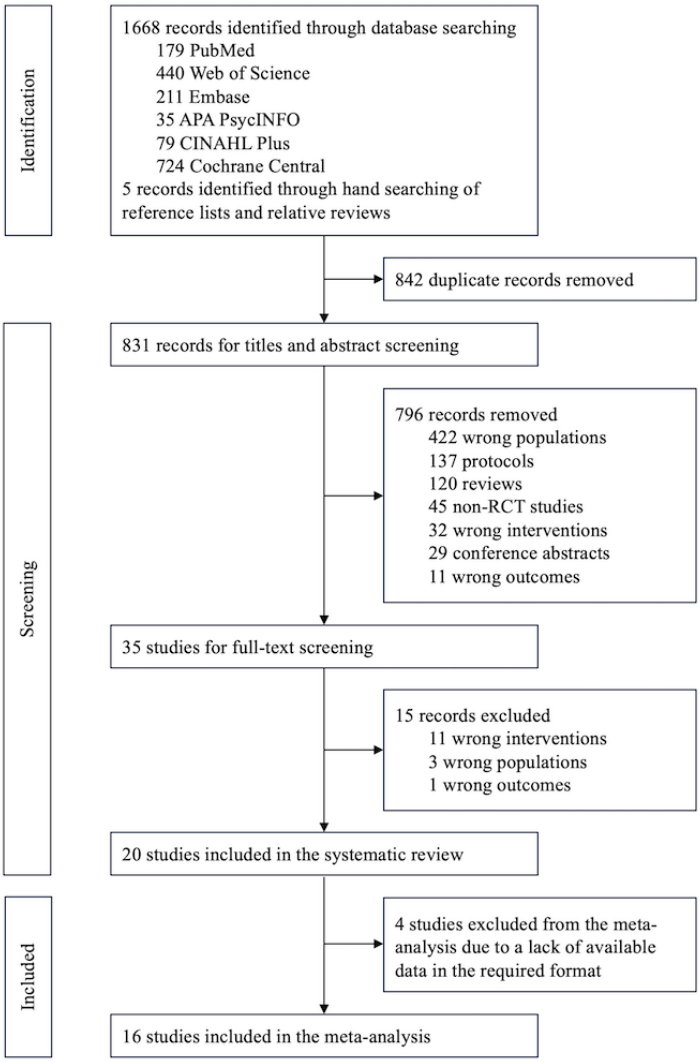

Study Characteristics

The characteristics of the included studies are presented in Table 1. A total of 20 studies were conducted in 9 countries, including China (6/20, 30%) [22-27], the United States (6/20, 30%) [28-33], Iran (2/20, 10%) [34,35], the United Kingdom (1/20, 5%) [36], Spain (1/20, 5%) [37], Germany (1/20, 5%) [38], Japan (1/20, 5%) [39], Palestine (1/20, 5%) [40], and Jordan (1/20, 5%) [41]. All studies were published after 2017, especially between 2020 and 2023. Of the 20 studies, 16 (80%) studies [22-24,26-28,30-36,39-41] used a parallel group design and 4 (20%) studies [25,29,37,38] used a cluster design. In total, 2 (10%) studies [33,36] addressed salt intake reduction and 2 (10%) studies [29,30] focused on medication management. In total, 2 (10%) studies [23,32] were conducted with ethnic minority groups and underserved and vulnerable populations. The number of participants in each study ranged from 30 to 636, and the duration of intervention ranged from 1.5 to 12 months. A total of 4168 participants were enrolled in the studies, with a retention rate of 85.2% (3551/4168).

XSL•FO
RenderX

Table 1. Characteristics of included studies (N=20).

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
Abu-El-Noor et al [40], 2021	Palestine	RCT[a]	A smartphone app that included reminders for taking medication, follow-up appointments, educational information about hypertension management, and records of BP[b] readings	—[c]	3	191/218 (87.6)	• MAd (measured by Hill-Bone CHBPTSe) improved from 15.64 to 11.73 (intervention) and 15.92 to 13.98 (control; *P*<.001)	• Hill-Bone CHBPTS improved from 30.82 to 23.40 (intervention) and 31.10 to 27.38 (control; *P*<.001) • Diet adherence improved from 10.99 to 8.36 (intervention) and 10.90 to 9.65 (control; *P*=.001) • Appointment adherence improved from 4.24 to 3.30 (intervention) and 4.28 to 3.76 (control)	Effective for adherence
Alsaqer and Bebis [41], 2022	Jordan	RCT	4 smartphone apps that encouraged self-monitoring of BP and record readings, adherence to medication, deep breathing exercises, and walking and counting steps daily; education for hypertension self-care; a public health nursing intervention that included telephone follow-up	—	3	74/80 (92)	• ΔSBPf: –14 mm Hg (intervention) and –7.75 mm Hg (control; *P*=.001) • ΔDBPg: –2.65 mm Hg (intervention) and –0.38 mm Hg (control; *P*=.14)	• SC-HIh—maintenance improved from 37.06 to 67.01 (intervention) and 33.93 to 44.52 (control; *P*=.001) • SC-HI—monitoring improved by 55.29 to 73.04 (intervention) and 52.70 to 54.59 (control; *P*=.001) • SC-HI—confidence improved from 41.79 to 82.06 (intervention) and 40.12 to 40.85 (control; *P*=.001) • Changes in all dimensions in SF-36i were significant after the intervention	Effective for SBP but not DBP and self-care

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
Bozorgi et al [34], 2021	Iran	RCT	A smartphone app that included BP records, reminders for drug consumption, visit date and BP measurement, a healthy diet and weight loss plans, disease knowledge, motivational and supportive programs for smoking cessation, critical BP alarm, customized messages about adherence to treatment, and usual treatment	—	2	118/120 (98.3)	• MA (measured by the 14-item Hill-Bone Scale) improved from 58.5 to 65.1 (intervention) and 59.1 to 59.7 (control)	• ΔMAPj: –12.6 mm Hg (intervention) and –16 mm Hg (control) • Adherence to a low-salt diet improved from 14.3 to 18.4 (intervention) and 15.8 to 17.3 (control) • Adherence to a low-fat diet improved from 16.0 to 17.86 (intervention) and 15.8 to 17.13 (control) • ΔVPAk: 32.8 min/wk (intervention) and –21.8 min/wk (control) • ΔMPAl: 91.5 min/wk (intervention) and 41.6 min/wk (control)	Effective for adherence and PA[m]
Chandler et al [28], 2019	United States	Small-scale efficacy RCT	The Smartphone Med Adherence Stops Hypertension intervention that included a smartphone app, global systems for a mobile electronic medication tray, and a Bluetooth-enabled BP device	Self-determination theory	9	54/56 (96)	• ΔSBP: –30.5 mm Hg (intervention) and –5 mm Hg (control) • ΔDBP: –12.6 mm Hg (intervention) and –5.2 mm Hg (control) • The percentage of participants with controlled SBP (<140 mm Hg) improved from 0% to 92.3% (intervention) and 0% to 27.8% (control; *P*=.001)	• MA (measured by MMASn) improved from 6.83 to 9.81 (intervention) and decreased from 6.99 to 6.84 (control; *P*<.001)	Effective for BP and MA
Dorsch et al [33], 2020	United States	Single-center, prospective pilot RCT	The LowSalt4Life mobile app that included just-in-time tailored messages that promote behavioral changes when the participant entered stores and restaurants, and easy scan and search for the foods at stores and restaurants to find options containing lower sodium content	Theory planned behavior; self-regulation theory	2	48/50 (96)	• ΔSBP: –7.5 mm Hg (intervention) and –0.7 mm Hg (control; *P*=.12)	• ΔKawasaki estimated 24-h urinary excretion of sodium –462 mg (intervention) and 381 mg (control; *P*=.03) • Δ24-h urinary excretion of sodium –637 mg (intervention) and –322 mg (control; *P*=.47)	Effective for salt intake
Frias et al [29], 2017	United States			—	3	105/109 (96.3)			Effective for BP

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
		Prospective, cluster, pilot RCT	The DMO° intervention included a smartphone app, medicines coencapsulated with an ingestible sensor, an adhesive wearable sensor patch, and a provider web portal and education and counseling from investigators				• ΔSBP (in 4 weeks): –21.8 mm Hg (intervention) and –12.7 mm Hg (control)	• ΔSBP (in 12 weeks): –20.9 mm Hg (intervention) and –15.2 mm Hg (control) • ΔDBP (in 4 weeks): –9.0 mm Hg (intervention) and –5.9 mm Hg (control) • ΔDBP (in 12 weeks): –8.6 mm Hg (intervention) and –5.8 mm Hg (control) • The percentage of participants with controlled BP (<140/90 mm Hg; in 4 weeks): 81.2% (intervention) and 33.3% (control); in 12 weeks: 80.0% (intervention) and 51.7% (control) • ΔGlycated hemoglobin A1c (in 12 weeks): –0.19 mmol/L (intervention) and +0.26 mmol/L (control) • The overall MA when using DMO was ≥80%	
Gong et al [27], 2020	China	Multicenter RCT	A smartphone app that provided reminders of drug dose and BP measurement, and scientific information and suggestions about hypertension	—	6	443/480 (92.3)	• ΔSBP: –8.99 mm Hg (intervention) and –5.92 mm Hg (control) • ΔDBP: –7.04 mm Hg (intervention) and –4.14 mm Hg (control) • The percentage of participants with controlled BP (<140/90 mm Hg) improved from 39% to 77% (intervention) and 39% to 67% (control; *P*=.01)	• MA (measured by MMAS) was 55% (low), 42% (medium), and 3% (high) in the intervention group and 68% (low), 30% (medium), and 2% (high) in the control group (*P*=.004)	Effective for BP and MA
Kario et al [39], 2021	Japan	Multicenter pilot RCT	A smartphone app that included a personalized lifestyle - modification program for lowering BP and standard lifestyle modification	—	6	140/146 (95.9)			Not effective for BP

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
							• ΔSBP (24-h ABPMp at 24 weeks): –0.47 mm Hg (intervention) and –0.042 mm Hg (control; *P*=.78) • ΔDBP (24-h ABPM at 24 weeks): –1.3 mm Hg (intervention) and –0.2 mm Hg (control; *P*=.39)	• ΔSBP (24-h ABPM at 16 weeks): 0.096 mm Hg (intervention) and –0.29 mm Hg (control; *P*=.88) • ΔSBP (home BP at 16 weeks): –4.1 mm Hg (intervention) and –0.96 mm Hg (control; *P*=.06) • ΔSBP (home BP self-monitoring at 24 weeks): –5.2 mm Hg (intervention) and –2.0 mm Hg (control; *P*=.07) • ΔDBP (nighttime ABPM at 24 weeks): –3.2 mm Hg (intervention) and –0.042 mm Hg (control; *P*=.04) • There were no significant changes in body weight, BMI, or waist circumference	
Leupold et al [38], 2023	Germany	Prospective cluster RCT	The PIA app that included transmission of BP measurements, graphic display of BP over time with an individual target range, a medication plan, ordering of prescription refills, video education, and links to BP-related information	—	12	525/636 (82.5)	• ΔSBP: –21.1 mm Hg (intervention) and –15.5 mm Hg (control; *P*<.001) • ΔDBP: –11.3 mm Hg (intervention) and –8.2 mm Hg (control; *P*=.40) • The percentage of participants with controlled BP (<140/90 mm Hg) was 62.6% (intervention) and 44.6% (control; *P*<.001)	—	Effective for BP
Li et al [25], 2019	China	Cluster RCT	Self-management intervention based on WeChat and included health education, health promotion, group chat, and BP monitoring	Self-efficacy theory	6	243/462 (52.6)			Effective for BP and self-management

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
							• ΔSBP: –5.5 mm Hg (intervention) and 1.6 mm Hg (control; *P*<.001) • ΔDBP: –1.3 mm Hg (intervention) and 2.1 mm Hg (control; *P*=.01) • The percentage of participants with controlled BP (<140/90 mm Hg) improved from 60.9% to 83.6% (intervention) and decreased from 69.2% to 63.6% (control; *P*<.001)	• ΔScore of hypertension: knowledge 2.3 (intervention) and 0.8 (control) • ΔScore of self-efficacy 0.8 (intervention) and –0.6 (control) • ΔScore of self-management: 7.3 (intervention) and –1.4 (control; *P*<.001) • ΔScore of social support: 0.4 (intervention) and 0.7 (control)	
Ma et al [26], 2022	China	RCT	A smartphone app that included health education, individual self-care planning, daily records, and an automated weekly health report and nurse-led individual education and consultation sessions	—	3	191/210 (91)	• ΔSBP: –11.63 mm Hg (intervention) and –1.01 mm Hg (control; *P*<.001) • ΔDBP: –5.53 mm Hg (intervention) and 1.69 mm Hg (control; *P*<.001) • The percentage of participants with controlled BP (<140/90 mm Hg) improved from 14.3% to 31.43% (intervention) and 7.6% to 8.57% (control; *P*=.003)	• ΔWeight: –1.16 kg (intervention) and –0.03 kg (control) • ΔBMI: –0.50 kg/m2 (intervention) and –0.09 kg/m2 (control) • ΔWCq: –3.02 cm (intervention) and 0.82 cm (control) • Self-care behavior (measured by HBP-HCPr) improved from 53.11 to 61.50 (intervention) and 54.82 to 54.94 (control); self-care motivation (measured by HBP-HCP) improved from 53.79 to 60.89 (intervention) and 55.02 to 55.27 (control); self-care–self-efficacy (measured by HBP-HCP) improved from 55.53 to 63.32 (intervention) and 55.60 to 56.13 (control)	Effective for BP and self-care
Márquez Contreras et al [37], 2018	Spain	Cluster RCT	A smartphone app that included personal data records, recommendations for target BP, physician's medication advice, reminder alarms, calendar of appointments or events, and a record of BP measurement results	—	12	148/154 (96.1)	• ΔSBP: –2.5 mm Hg (intervention) and –0.07 mm Hg (control; *P*<.001) • ΔDBP: –3.14 mm Hg (intervention) and –0.5 mm Hg (control; *P*<.001)	• Pharmacological therapeutic adherence was significantly improved after the intervention	Effective for BP and MA

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
Morawski et al [30], 2018	United States	RCT	The Medisafe mobile app that included reminder alerts of medication, adherence reports, tracks of BP, and optional peer support	—	3	411/412 (99.8)	• ΔSBP: –10.6 mm Hg (intervention) and –10.1 mm Hg (control; *P*=.78) • MA (measured by MMAS-8) improved from 6.0 to 6.3 (intervention) and maintained 5.7 (control; *P*=.001)	• The percentage of participants with controlled BP (<140/90 mm Hg) improved from 0% to 35.8% (intervention) and 0% to 37.9% (control; *P*=.69)	Effective for MA but not SBP
Najafi Ghezeljeh et al [35], 2018	Iran	RCT	A Telegram group that included self-management education and contact and communication	—	1.5	50/50 (100)	• Self-management behavior (measured by HSM-BQs) improved from 2.23 to 3.73 (intervention) and decreased from 1.84 to 1.83 (control)	—	Effective for self-management
Persell et al [31], 2020	United States	RCT	Home BP self-monitoring and hypertension personal control program that included medication reminders, hypertension education, DASH[t] diet encouragements, BP measuring reminders, coaching about PA, sleep tracks, stress management education, record reminders, and customized communication	Cognitive behavioral therapy	6	297/333 (88.2)	• ΔSBP: –8.3 mm Hg (intervention) and –6.8 mm Hg (control; *P*=.16) • ΔDBP: –4.3 mm Hg (intervention) and –3.6 mm Hg (control; *P*=.61) • The percentage of participants with controlled BP (<140/90 mm Hg) improved from 36% to 72% (intervention) and 41% to 78% (control; *P*=.66)	• Self-confidence in controlling BP significantly improved after the intervention (*P*<.001)	Not effective for BP
Payne Riches et al [36], 2021	United Kingdom	RCT	The SaltSwap mobile app that included brief advice on encouraging individuals to swap to lower-salt alternatives, buying fewer high-salt foods, and using less salt when cooking and face-to-face behavioral advice and support provided by a health care professional	Behavior change wheel	1.5	45/47 (96)	• ΔSBP: –1.0 mm Hg (intervention) and –1.1 mm Hg (control; *P*=.82) • ΔDBP: –1.0 mm Hg (intervention) and 2.3 mm Hg (control; *P*=.23)	• Salt intake: –0.2 g/d (intervention) and –1.0 g/d (control; *P*=.68); purchased salt: –0.0 g/100 g (intervention) and –0.1 g/100 g (control; *P*=.16)	Not effective for BP and salt intake
Sun et al [24], 2020	China	RCT	3 WeChat groups (according to cardiovascular risk factors) that included health education, health behavior promotion, group chats, and BP monitoring	—	3	117/120 (98)	• ΔSBP: –10.92 mm Hg (intervention) and –3.43 mm Hg (control; *P*<.001) • ΔDBP: –5.68 mm Hg (intervention) and –2.23 mm Hg (control; *P*=.07)		Effective for BP and MA

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
								• Changes in TCu and LDL-Cv were significant after the intervention • MA (measured by MMAS-8) improved from 2.28 to 3.30 (intervention) and 2.35 to 2.38 (control) • Self-management behavior (measured by HPSM-BRSw) improved from 72.27 to 74.57 (intervention) and 70.82 to 70.85 (control; *P*<.001) • ΔBMI: –0.49 kg/m2 (intervention; *P*<.001) and –0.06 kg/m2 (control)	
Zha et al [32], 2020	United States	Pilot RCT	The iHealth MyVitals mobile app included tracks and analysis of key health measurements and instant feedback, helping users self-monitor and manage BP, and standard hypertension management	—	6	25/30 (83)	• ΔSBP: –8.39 mm Hg (intervention) and –4.79 mm Hg (control; *P*=.01) • ΔDBP: –2.76 mm Hg (intervention) and –2.2 mm Hg (control)	• MA self-efficacy (measured by MASESx) improved from 64.85 to 69.17 (intervention) and decreased from 64.75 to 61.00 (control; *P*=.06)	Effective for BP and self-management
Zhang et al [22], 2022	China	RCT	A smartphone app that included reminder alerts, access to historical data, and real-time chat and wearable device that can trace steps, heart rate, BP, and sleeping hours	—	6	192/307 (62.5)	• ΔSBP: 3.2 mm Hg (intervention) and 6.4 mm Hg (control) • ΔDBP: 2.7 mm Hg (intervention) and 5.4 mm Hg (control)	• ΔBMI: 0.4 kg/m2 (intervention) and 0.6 kg/m2 (control) • ΔWeight: 1.1 kg (intervention) and 1.5 kg (control)	Effective for BP
Zhang et al [23], 2023	China	RCT	A smartphone app that included reminder alerts, adherence reports, medical instruction, and optional family support and monitoring a wearable device (model unknown)	Theory of planned behavior	3	134/148 (90.5)	• ΔSBP: –8.52 mm Hg (intervention) and –1.25 mm Hg (control; *P*=.01) • ΔDBP: –0.42 mm Hg (intervention) and –0.01 mm Hg (control)		Effective for SBP but not DBP

XSL•FO
RenderX

Study, year	Country	Study design	Intervention content	Behavior change theory	Duration (months)	Retention, n/N (%)	Primary outcomes	Secondary outcomes	Conclusion
								• ΔWC: –2.14 cm (intervention) and –0.25 cm (control) • ΔHCy: –0.30 cm (intervention) and –0.01 cm (control; *P*=.08) • Self-efficacy improved: 12.89 (intervention) and 5.43 (control) • Hypertension compliance improved: 7.35 (intervention) and 3.01 (control) • Physical health improved: 12.21 (intervention) and 1.54 (control) • Mental health improved: 13.17 (intervention) and 2.55 (control)	

[a]RCT: randomized controlled trial.
[b]BP: blood pressure.
[c]Not applicable.
[d]MA: medication adherence.
[e]CHBPTS: Compliance to High Blood Pressure Therapy Scale.
[f]SBP: systolic blood pressure.
[g]DBP: diastolic blood pressure.
[h]SC-HI: Self-Care of Hypertension Inventory.
[i]SF-36: 36-Item Short Form Survey.
[j]MAP: mean arterial pressure.
[k]VPA: vigorous physical activity.
[l]MPA: moderate physical activity.
[m]PA: physical activity.
[n]MMAS: Morisky Medication Adherence Scale.
[o]DMO: digital medicine offering.
[p]ABPM: ambulatory blood pressure monitoring.
[q]WC: waist circumference.
[r]HBP-HCP: Hypertension Self-care Profile.
[s]HSMBQ: Hypertension Self-Management Behavior Questionnaire.
[t]DASH: Dietary Approaches to Stop Hypertension.
[u]TC: total cholesterol.
[v]LDL-C: low-density lipoprotein cholesterol.
[w]HPSMBRS: Hypertension Patients Self-Management Behavior Rating Scale.
[x]MASES: Medication Adherence Self-Efficacy Scale.
[y]HC: hip circumference.

Risk of Bias and Grade of Evidence

Of the 20 included studies, 15 (75%) [22-25,27-29,32-36,39-41] were judged as high risk of bias, 1 (5%) [37] was judged as having some concerns, and 4 (20%) [26,30,31,38] were judged as low risk of bias (Figure 2 [22-41]). Overall, the risk of bias primarily existed in the randomization process and measurement of the outcome. In terms of the randomization process, 50% (10/20) of the studies [23,25,27,29,32-34,36,40,41] were rated as high risk, primarily because of not following the principle of, or not providing details about, randomization and allocation concealment. Regarding the measurement of the outcome, 40% (8/20) of the studies [24,27,28,33-36,40] were assessed as high

XSL•FO
RenderX

risk mainly due to the use of patients' self-reported subjective outcomes and the lack of disclosing whether the outcome measurers were aware of the intervention patients received. In addition, the lack of blinding among patients and between patients and investigators during the intervention process (although often difficult to achieve in mHealth interventions), high dropout rates and not conducting intention-to-treat analyses, and selective disclosure of predetermined outcome data were also important sources of risk of bias. The result of the Egger test (P=.16) and visualization of the funnel plot (Figure S1 in Multimedia Appendix 2) indicated that there was no significant risk of publication bias in the included studies.

Figure 2. Risk of bias assessment of the included studies [22-41] using the Cochrane Risk of Bias (version 2) tool. RCT: randomized controlled trial.

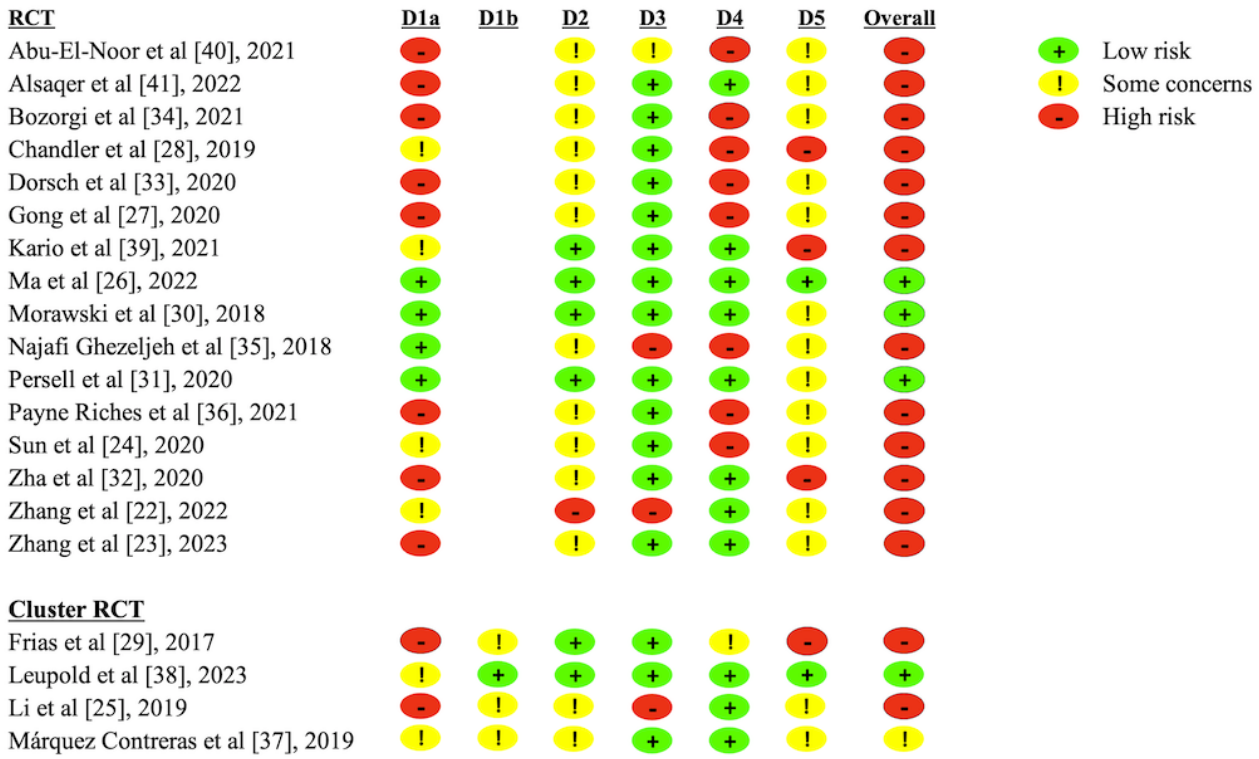

The results of the Grading of Recommendation, Assessment, Development, and Evaluation assessment are presented in Table 2. In terms of SBP, the quality of evidence was low, and for DBP, the quality of evidence was moderate.

Table 2. Grading of Recommendation, Assessment, Development, and Evaluation assessment.

Blood pressure	Studies, n	Quality assessment					Patients (E[a]/C[b])	Effect, mean difference (95% CI)	Overall certainty of evidence
		Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias			
SBP[c]	16	Serious[d]	Serious[e]	Not serious	Not serious	None	1568/1516	–5.78 (–7.97 to 3.59)	Low
DBP[f]	15	Serious[d]	Not serious	Not serious	Not serious	None	1359/1316	–3.28 (–4.39 to –2.17)	Moderate

[a]E: experimental group.

[b]C: control group.

[c]SBP: systolic blood pressure.

[d]The included studies were judged as have a high risk of bias overall. Some of the studies had evident risks in the randomization process and measurement of the outcome.

[e]Cochran Q test and Higgins I^2 suggested significant heterogeneity between studies.

[f]DBP: diastolic blood pressure.

Behavior Change Theories and Techniques

Table 1 and Table S2 in Multimedia Appendix 2 summarize the behavior change theories and BCTs used in the included studies. Of the 20 studies, 6 (30%) [23,25,28,31,33,36] reported 7 behavior change theories in their interventions and the remaining 14 (70%) did not include a theoretical foundation. Theories involved in the interventions included

XSL•FO
RenderX

self-determination theory [28], theory of planned behavior [23,33], self-regulation theory [33], self-efficacy theory [25], cognitive behavioral theory [31], and behavior change wheel [36]. Of the 6 studies, 1 (17%) used 2 theories [33] and the remaining 5 (83%) were based on a single theory.

A total of 21 different BCTs from 12 BCT categories were reported in the included 20 studies. The mean number of BCTs was 8.7 (SD 3.8), with a range of 2 to 17, accounting for 2% (2/93) to 18% (17/93) of the total 93 BCTs. The frequently used BCT clusters (used in ≥10 studies) were *goals and planning*, *feedback and monitoring*, *shaping knowledge*, *associations*, *comparison of outcomes*, and *regulation*. The most common BCTs were *self-monitoring of outcomes of behavior* (17/20, 85%), *feedback on outcomes of behavior* (15/20, 75%), *instruction on how to perform the behavior* (15/20, 75%), *pharmacological support* (15/20, 75%), *biofeedback* (14/20, 70%), *prompts/cues* (13/20, 65%), *credible source* (13/20, 65%), and *action planning* (11/20, 55%). There were 4 BCTs used only in 10% (2/20) of the studies: *problem solving*, *information about health consequences*, *behavior substitution*, and *conserving mental resources*.

Effects of an mHealth App Intervention on BP

Effect of an mHealth App Intervention on SBP

In total, 16 (80%) of the 20 studies reported the effects of an mHealth app intervention on SBP. A total of 1568 participants from the intervention group and 1516 participants from the control group were included in the meta-analysis. As presented in Figure 3, the mHealth app intervention resulted in a –5.78 mm Hg (95% CI –7.97 mm Hg to –3.59 mm Hg) reduction in SBP. The heterogeneity was significant (I^2=82%; P<.001) between the studies.

Figure 3. Forest plot of the overall effect of a mobile health app intervention on systolic blood pressure.

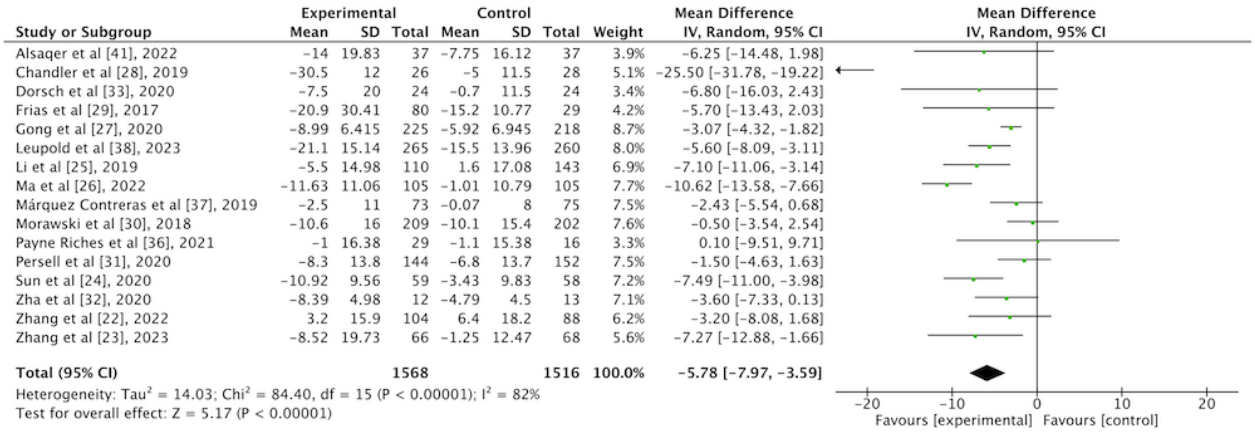

Effect of an mHealth App Intervention on DBP

As shown in Figure 4, in total, 15 (75%) of the 20 studies reporting the effect of an mHealth app intervention on DBP including 1359 participants from the intervention group and 1316 participants from the control group were included in the analysis. Overall, the mHealth app intervention resulted in a –3.28 mm Hg (95% CI –4.39 mm Hg to –2.17 mm Hg) reduction in DBP with a medium heterogeneity (I^2=58%; P=.003).

Figure 4. Forest plot of the overall effect of a mobile health app intervention on diastolic blood pressure.

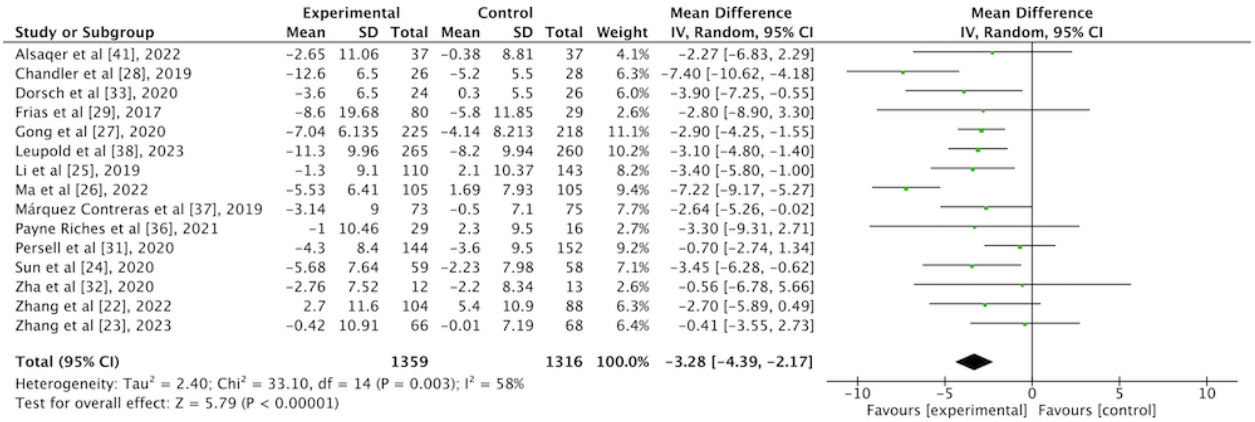

Subgroup Analyses

Subgroup analyses were subsequently conducted to explore the impact of factors, such as the risk factors of hypertension that mHealth app interventions addressed (addressing a single risk factor vs addressing multiple risk factors), the presence of a theoretical foundation (with behavior change theories vs without behavior change theories), duration of the intervention (<6 vs ≥6 months), and the number of BCTs (≥11 vs<11), on the effect size of an mHealth app intervention on SBP.

XSL•FO
RenderX

Factors Addressed in an mHealth App Intervention

Of the 16 studies that reported the effects of an mHealth app intervention on SBP, 12 (75%) studies [22-28,31,32,37,38,41] carried out interventions (eg, education of disease; BP monitoring and alerts; instructions on diet, salt intake, exercise, sleep, and stress management; and medication reminders or management) addressing multiple physiological and behavioral risk factors of hypertension, and the remaining 4 (25%) studies [29,30,33,36] carried out interventions (eg, reduction of salt intake or medication adherence management) addressing a single physiological or behavioral risk factor. As illustrated in Figure S2 in Multimedia Appendix 2, compared with the interventions addressing a single risk factor of hypertension (–1.54 mm Hg, 95% CI –4.15 mm Hg to 1.06 mm Hg), interventions addressing multiple risk factors of hypertension resulted in a better reduction in SBP (–6.50 mm Hg, 95% CI –9.00 mm Hg to –3.99 mm Hg) and the difference between subgroups was statistically significant (P=.007). However, the heterogeneity between studies addressing multiple risk factors remained significant (I^2=86%; P<.001).

Given the obvious heterogeneity in studies with interventions addressing multiple physiological and behavioral risk factors of hypertension and that hypertension is a disease associated with multiple risk factors, the effect of mHealth apps addressing a single physiological or behavioral risk factor may differ from the effect of those addressing multiple risk factors. Therefore, subsequent subgroup analyses were conducted based on the 12 studies addressing multiple physiological and behavioral risk factors of hypertension.

Use of Behavior Change Theory

Of the 12 studies, 5 (42%) studies [23,25,26,28,31] used behavior change theory and 7 (58%) studies [22,24,27,32,37,38,41] did not disclose the theoretical foundation. As presented in Figure S3 in Multimedia Appendix 2, although the difference between subgroups was not statistically significant (P=.07), interventions with behavior change theory resulted in a –10.06 mm Hg (95% CI –16.42 mm Hg to –3.70 mm Hg) reduction in SBP, which is better than without behavior change theory with a –4.13 mm Hg (95% CI –5.50 mm Hg to –2.75 mm Hg) reduction in SBP. The heterogeneity among studies with behavior change theory was considerable (I^2=92%; P<.001).

Duration of Intervention

In the 12 studies, the duration of intervention ranged from 3 to 12 months, with a mean duration of 6.25 months. Among them, 4 (33%) studies [23,24,26,41] had a duration of 3 months, and the other 8 (67%) studies [22,25,27,28,31,32,37,38] had a duration of at least 6 months. The results of the subgroup analysis showed that the effect of the 3-month intervention duration (–8.87 mm Hg, 95% CI –10.90 mm Hg to –6.83 mm Hg) on reducing SBP was superior to those with a longer duration (–5.76 mm Hg, 95% CI –8.74 mm Hg to –2.77 mm Hg; P=.09). Apparent heterogeneity persisted between studies with interventions lasting longer than 6 months (I^2=86%; P<.001; Figure S4 in Multimedia Appendix 2).

Use of BCTs

A total of 19 different BCTs were used in 12 mHealth app interventions, with the number ranging from 2 to 12 and an average of 10.58 BCTs per intervention. In total, 7 (58%) of the 12 studies [23-26,28,38,41] used at least 11 BCTs. As shown in Figure S5 in Multimedia Appendix 2, interventions using at least 11 BCTs (–9.68 mm Hg, 95% CI –13.49 mm Hg to –5.87 mm Hg) had a statistically significant effect on the reduction in SBP compared to studies using fewer BCTs (–2.88 mm Hg, 95% CI –3.90 mm Hg to –1.86 mm Hg; P<.001).

Discussion

Principal Findings

In this study, we conducted a systematic review of 20 studies about mHealth app interventions for hypertension and performed a meta-analysis of 16 (80%) of the studies. The results indicated that mHealth app interventions resulted in a significant reduction in SBP (P<.001) and DBP (P<.001) compared to usual care, and the effect size was influenced by factors of intervention design (eg, presence of a theoretical foundation, intervention duration, and number of BCTs) and app contents (eg, the risk factors of the hypertension app addressed). Our study further demonstrates the effectiveness of mHealth app interventions in hypertension self-management and, for the first time, provides an interpretation of the active ingredients in such interventions from a BCT perspective. However, significant differences in intervention designs and the number and selection of BCTs across studies also indicated that these interventions may not be generalized in different social settings, and there is currently a lack of a unified guidance framework in mHealth interventions for hypertension.

Comprehensive intervention based on antihypertensive medications combined with lifestyle modifications is considered a standard strategy for the management of hypertension [42,43]. Nonpharmacological management is a multidimensional task that includes weight loss, the Dietary Approaches to Stop Hypertension diet, sodium reduction, potassium supplementation, increased physical activity, and reduction in tobacco and alcohol consumption [44]. As we guessed, due to covering more lifestyle modifications, mHealth apps addressing multiple physiological and behavioral risk factors of hypertension and offering more functions were more effective in lowering SBP. This finding was consistent with previous studies conducted in various populations [45-48]. Notably, the interventions in studies by Chandler et al [28] and Frias et al [29] resulted in more reduced SBP in absolute values compared to other mHealth app interventions. In those studies, the researchers applied medication trays with a series of reminder signals and ingestible sensors to enhance medication management, indicating that the strategies of enhanced reminders and tracking of medication behavior might be potential ways to improve medication management. Epidemiological evidence suggests that nonadherence to antihypertensive medications is as high as 27% to 40% globally, and there is still significant room for improvement in patients' medication behavior [49]. Therefore, developing a multicomponent mHealth app for lifestyle modifications and

XSL•FO
RenderX

conducting intensive intervention for medication behavior may be a viable direction for future mHealth interventions in hypertension.

Intervention duration is a key factor in intervention design, and it was noteworthy that, in our study, 20% (4/20) of the studies with an intervention duration of 3 months were more effective in reducing SBP than studies lasting ≥6 months. This finding was consistent with a previous study, in which Ma et al [50] found that the effect of a habit formation intervention on physical activity habits was better if the intervention duration was <12 weeks. Given that the contents of mHealth app interventions primarily involve lifestyle modifications and the formation of healthy living habits, this phenomenon seems to be partly explained by habit formation. The law of automaticity change in the formation of new habits indicates that habit strength usually reaches a peak of automaticity at approximately 12 weeks and gradually weakens over the following period [50,51]. In addition, of these 4 studies, 2 (50%) [24,26] involved educational curriculum programs that progressed over time. We suggested that, in addition to habit formation, the enhancement of intrinsic motivation for behavior change by planned sessions on disease, medications, and coping may also contribute to BP management. Therefore, how to promote the development of healthy living habits and maintain the habits and motivation of behavior change over a long period of the intervention will be an issue to consider in future study designs of similar interventions.

A theory is a set of interrelated concepts, definitions, and propositions that explain or predict events or situations by specifying relations among variables [52]. The existing view is that effective interventions used to promote healthy lifestyles and reduce risky behaviors are inseparable from the evidence of theories [53]. Similar to our findings, theory-based mHealth interventions have been found more effective in other chronic diseases and behavioral change studies [14,54]. A possible explanation is that, based on theories, researchers can identify causal factors associated with behavior change and the pathways through which behavior change occurs. Furthermore, this enhances treatment fidelity, allowing for a more comprehensive design of the protocol, early detection of errors and protocol deviations, and improvement in treatment retention, to maximize the effectiveness of interventions [43,55]. In our review, 6 studies disclosed the theories used, and the theory use rate was 30% (6/20). The lack of a theoretical foundation seems to have become a common phenomenon in current behavior change interventions [56,57]. We believe that this phenomenon needs urgent attention. With the development of mobile information technology, an increasing number of interventions are being based on mobile, smart, and wearable devices and are being transferred from health care institutions to patients' homes. In the absence of face-to-face communication and strict supervision in telemedicine, holistic study designs based on theories will be particularly important to improve the effectiveness and generalization of the interventions. Future researchers can try to design intervention protocols based on multiple theories and integrate these theories to explore the best model for mHealth app interventions for hypertension.

Research evidence, including our study, has suggested that using more BCTs may be associated with better outcomes of behavior change [58,59]. However, some studies found that fewer techniques and the right combinations of techniques are more effective [60-62]. Indeed, due to the small effect of a single BCT, the fact that BCTs are often present in combinations in interventions, and the possibility of interactions between BCTs, determining which specific BCT, the combinations of BCTs, and the number of BCTs that are effective for a given behavior is a challenge [63]. Therefore, based on applying a certain number of BCTs and effective BCT combinations, customizing an intervention to the patient's behavior change needs, including contents, duration, and delivery, may be another possible method to improve the effectiveness of the interventions.

This study also had some limitations. First, we intended to evaluate the effectiveness of mHealth app interventions on several subjective and objective outcomes including BP, medication adherence, self-efficacy, etc. However, due to the limited number of included studies and significant heterogeneity of subjective outcome measures, we ultimately only reported the result of BP management, which was inconsistent with the published protocol. In addition, although we attempted to explore the source of heterogeneity through subgroup analyses, heterogeneity remained significant in some subgroups, and because the included studies were all published after 2017, we also failed to explore the changes in the effectiveness of mHealth app interventions over time. It is recommended that future studies incorporate additional studies and conduct extensive subgroup analyses to further explore the impact of mHealth app interventions on self-management–related outcomes in patients with hypertension. Second, in this study, we only included app-based mHealth interventions and excluded studies based on phone calls, text messages, and website programs, which also contained BCTs, potentially leading to bias in the assessment of effective BCTs. Third, the evidence for BCT coding relies on information about intervention contents from available articles, supplementary materials, and secondary analysis publications, which were underreported or roughly outlined. Furthermore, despite the coding process being performed by 2 researchers independently with good consistency, BCT coding is inevitably susceptible to researchers' subjective judgments. Therefore, the efficacy of BCTs for hypertension self-management still requires further validation in future studies. Finally, while our study coded BCTs and initially explored the effectiveness of the number of BCTs in SBP reduction, considering the wide variations of intervention contents and specific forms, frequencies, and intensities of BCTs across studies, we failed to further quantitatively investigate the effect of a single BCT or a combination of BCTs on BP reduction. Self-management of hypertension is a complex and multidimensional intervention, involving lifestyle modifications and medication management, and is affected by factors such as patients' knowledge, intention, self-efficacy, and environment. It is foreseeable that more BCTs will be found effective in self-management of hypertension. Future research could further explore other BCTs and combinations of BCTs to provide references for the development of relevant mHealth interventions and apps.

XSL•FO
RenderX

Conclusions

This systematic review and meta-analysis further confirmed the effectiveness of mHealth app self-management interventions for hypertension and identified the BCTs used in the interventions. Our study found that mHealth app interventions can lead to a reduction in SBP and DBP compared to usual care; factors related to the intervention and study design, such as the risk factors of hypertension the mHealth app intervention addressed, the presence of a theoretical foundation, intervention duration, and the number of BCTs, were associated with the effect sizes of BP reduction; and the most commonly used BCTs included *self-monitoring of outcomes of behavior*, *feedback on outcomes of behavior*, *instruction on how to perform the behavior*, and *pharmacological support*. On the basis of the findings of our study, future research can optimize the intervention designs and use more BCTs and BCT combinations to develop more effective mHealth apps and interventions for hypertension management.

Acknowledgments

This study was funded by the NonProfit Central Research Institute Fund of the Chinese Academy of Medical Sciences (grant #2023-RC320-01).

The authors would like to thank all authors of the included studies.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YZ and MHP contributed to conceptualization. YZ, SJL, and RQH contributed to the methodology. YZ and HMM were involved in software analyses. AQW and RYP contributed to validation. YZ and AQW contributed to the formal analysis. YZ, SJL, and RQH contributed to the investigation. YZ and XYT contributed to data curation. YZ and SJL contributed to writing the original draft. YZ and MHP contributed to reviewing and editing the manuscript. MHP contributed to supervising the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e54978_app1.docx]

Multimedia Appendix 2
Supplementary tables and figures.
[DOCX File , 2014 KB - jmir_v26i1e54978_app2.docx]

References

1. Colafella KM, Denton KM. Sex-specific differences in hypertension and associated cardiovascular disease. Nat Rev Nephrol 2018 Mar;14(3):185-201. [doi: 10.1038/nrneph.2017.189] [Medline: 29380817]
2. Mills KT, Bundy JD, Kelly TN, Reed JE, Kearney PM, Reynolds K, et al. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. Circulation 2016 Aug 09;134(6):441-450 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.018912] [Medline: 27502908]
3. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nat Rev Nephrol 2020 Apr;16(4):223-237 [FREE Full text] [doi: 10.1038/s41581-019-0244-2] [Medline: 32024986]
4. Olsen MH, Angell SY, Asma S, Boutouyrie P, Burger D, Chirinos JA, et al. A call to action and a lifecourse strategy to address the global burden of raised blood pressure on current and future generations: the Lancet Commission on hypertension. Lancet 2016 Nov 26;388(10060):2665-2712. [doi: 10.1016/S0140-6736(16)31134-5] [Medline: 27671667]
5. Aovare P, Abdulai K, Laar A, van der Linden EL, Moens N, Richard E, et al. Assessing the effectiveness of mHealth interventions for diabetes and hypertension management in Africa: systematic review and meta-analysis. JMIR Mhealth Uhealth 2023 Aug 29;11:e43742 [FREE Full text] [doi: 10.2196/43742] [Medline: 37646291]
6. Zhou L, He L, Kong Y, Lai Y, Dong J, Ma C. Effectiveness of mHealth interventions for improving hypertension control in uncontrolled hypertensive patients: a meta-analysis of randomized controlled trials. J Clin Hypertens (Greenwich) 2023 Jul;25(7):591-600 [FREE Full text] [doi: 10.1111/jch.14690] [Medline: 37409556]
7. Xu H, Long H. The effect of smartphone app-based interventions for patients with hypertension: systematic review and meta-analysis. JMIR Mhealth Uhealth 2020 Oct 19;8(10):e21759 [FREE Full text] [doi: 10.2196/21759] [Medline: 33074161]

XSL•FO
RenderX

8. Kassavou A, Wang M, Mirzaei V, Shpendi S, Hasan R. The association between smartphone app-based self-monitoring of hypertension-related behaviors and reductions in high blood pressure: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Jul 12;10(7):e34767 [FREE Full text] [doi: 10.2196/34767] [Medline: 35819830]
9. Li R, Liang N, Bu F, Hesketh T. The effectiveness of self-management of hypertension in adults using mobile health: systematic review and meta-analysis. JMIR Mhealth Uhealth 2020 Mar 27;8(3):e17776 [FREE Full text] [doi: 10.2196/17776] [Medline: 32217503]
10. Siopis G, Moschonis G, Eweka E, Jung J, Kwasnicka D, Asare BY, et al. Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with hypertension: a systematic review and meta-analysis of randomised controlled trials. Lancet Digit Health 2023 Mar;5(3):e144-e159 [FREE Full text] [doi: 10.1016/S2589-7500(23)00002-X] [Medline: 36828607]
11. Mikulski BS, Bellei EA, Biduski D, De Marchi AC. Mobile health applications and medication adherence of patients with hypertension: a systematic review and meta-analysis. Am J Prev Med 2022 Apr;62(4):626-634. [doi: 10.1016/j.amepre.2021.11.003] [Medline: 34963562]
12. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
13. Wang X, Sun J, Yin X, Zou C, Li H. Effects of behavioral change techniques on diet and physical activity in colorectal cancer patients: a systematic review and meta-analysis. Support Care Cancer 2022 Dec 14;31(1):29. [doi: 10.1007/s00520-022-07511-7] [Medline: 36515770]
14. El-Gayar O, Ofori M, Nawar N. On the efficacy of behavior change techniques in mHealth for self-management of diabetes: a meta-analysis. J Biomed Inform 2021 Jul;119:103839 [FREE Full text] [doi: 10.1016/j.jbi.2021.103839] [Medline: 34139330]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0. The Cochrane Collaboration. 2011 Mar. URL: https://handbook-5-1.cochrane.org/ [accessed 2023-01-23]
17. Online training. BCT Taxonomy V1. URL: https://www.bct-taxonomy.com/ [accessed 2024-09-30]
18. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
19. Grade homepage. Grade Working Group. URL: https://www.gradeworkinggroup.org/ [accessed 2024-09-30]
20. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
21. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
22. Zhang Y, Tao Y, Zhong Y, Thompson J, Rahmani J, Bhagavathula AS, et al. Feedback based on health advice via tracing bracelet and smartphone in the management of blood pressure among hypertensive patients: a community-based RCT trial in Chongqing, China. Medicine (Baltimore) 2022 Jul 15;101(28):e29346 [FREE Full text] [doi: 10.1097/MD.0000000000029346] [Medline: 35839004]
23. Zhang YT, Tan XD, Wang Q. Effectiveness of a mHealth intervention on hypertension control in a low-resource rural setting: a randomized clinical trial. Front Public Health 2023 Mar 1;11:1049396 [FREE Full text] [doi: 10.3389/fpubh.2023.1049396] [Medline: 36935728]
24. Sun YQ, Jia YP, Lv JY, Ma GJ. The clinical effects of a new management mode for hypertensive patients: a randomized controlled trial. Cardiovasc Diagn Ther 2020 Dec;10(6):1805-1815 [FREE Full text] [doi: 10.21037/cdt-20-589] [Medline: 33381425]
25. Li X, Li T, Chen J, Xie Y, An X, Lv Y, et al. A WeChat-based self-management intervention for community middle-aged and elderly adults with hypertension in Guangzhou, China: a cluster-randomized controlled trial. Int J Environ Res Public Health 2019 Oct 23;16(21):4058 [FREE Full text] [doi: 10.3390/ijerph16214058] [Medline: 31652688]
26. Ma Y, Cheng HY, Sit JW, Chien WT. The effects of a smartphone-enhanced nurse-facilitated self-care intervention for Chinese hypertensive patients: a randomised controlled trial. Int J Nurs Stud 2022 Oct;134:104313. [doi: 10.1016/j.ijnurstu.2022.104313] [Medline: 35802960]
27. Gong K, Yan YL, Li Y, Du J, Wang J, Han Y, et al. Mobile health applications for the management of primary hypertension: a multicenter, randomized, controlled trial. Medicine (Baltimore) 2020 Apr;99(16):e19715 [FREE Full text] [doi: 10.1097/MD.0000000000019715] [Medline: 32311957]
28. Chandler J, Sox L, Kellam K, Feder L, Nemeth L, Treiber F. Impact of a culturally tailored mHealth medication regimen self-management program upon blood pressure among hypertensive Hispanic adults. Int J Environ Res Public Health 2019 Apr 06;16(7):1226 [FREE Full text] [doi: 10.3390/ijerph16071226] [Medline: 30959858]

XSL•FO
RenderX

29. Frias J, Virdi N, Raja P, Kim Y, Savage G, Osterberg L. Effectiveness of digital medicines to improve clinical outcomes in patients with uncontrolled hypertension and type 2 diabetes: prospective, open-label, cluster-randomized pilot clinical trial. J Med Internet Res 2017 Jul 11;19(7):e246 [FREE Full text] [doi: 10.2196/jmir.7833] [Medline: 28698169]
30. Morawski K, Ghazinouri R, Krumme A, Lauffenburger JC, Lu Z, Durfee E, et al. Association of a smartphone application with medication adherence and blood pressure control: the MedISAFE-BP randomized clinical trial. JAMA Intern Med 2018 Jun 01;178(6):802-809 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0447] [Medline: 29710289]
31. Persell SD, Peprah YA, Lipiszko D, Lee JY, Li JJ, Ciolino JD, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. JAMA Netw Open 2020 Mar 02;3(3):e200255 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0255] [Medline: 32119093]
32. Zha P, Qureshi R, Porter S, Chao YY, Pacquiao D, Chase S, et al. Utilizing a mobile health intervention to manage hypertension in an underserved community. West J Nurs Res 2020 Mar;42(3):201-209. [doi: 10.1177/0193945919847937] [Medline: 31057081]
33. Dorsch MP, Cornellier ML, Poggi AD, Bilgen F, Chen P, Wu C, et al. Effects of a novel contextual just-in-time mobile app intervention (LowSalt4Life) on sodium intake in adults with hypertension: pilot randomized controlled trial. JMIR Mhealth Uhealth 2020 Aug 10;8(8):e16696 [FREE Full text] [doi: 10.2196/16696] [Medline: 32663139]
34. Bozorgi A, Hosseini H, Eftekhar H, Majdzadeh R, Yoonessi A, Ramezankhani A, et al. The effect of the mobile "blood pressure management application" on hypertension self-management enhancement: a randomized controlled trial. Trials 2021 Jun 24;22:413. [doi: 10.1186/s13063-021-05270-0]
35. Najafi Ghezeljeh T, Sharifian S, Nasr Isfahani M, Haghani H. Comparing the effects of education using telephone follow-up and smartphone-based social networking follow-up on self-management behaviors among patients with hypertension. Contemp Nurse 2018 Mar 05;54(4-5):362-373. [doi: 10.1080/10376178.2018.1441730]
36. Payne Riches S, Piernas C, Aveyard P, Sheppard JP, Rayner M, Albury C, et al. A mobile health salt reduction intervention for people with hypertension: results of a feasibility randomized controlled trial. JMIR Mhealth Uhealth 2021 Oct 21;9(10):e26233. [doi: 10.2196/26233]
37. Márquez Contreras E, Márquez Rivero S, Rodríguez García E, López-García-Ramos L, Carlos Pastoriza Vilas J, Baldonedo Suárez A, et al. Specific hypertension smartphone application to improve medication adherence in hypertension: a cluster-randomized trial. Curr Med Res Opin 2018 Dec 05;35(1):167-173. [doi: 10.1080/03007995.2018.1549026]
38. Leupold F, Karimzadeh A, Breitkreuz T, Draht F, Klidis K, Grobe T, et al. Digital redesign of hypertension management with practice and patient apps for blood pressure control (PIA study): a cluster-randomised controlled trial in general practices. eClinicalMedicine 2023 Jan;55:101712. [doi: 10.1016/j.eclinm.2022.101712]
39. Kario K, Nomura A, Kato A, Harada N, Tanigawa T, So R, et al. Digital therapeutics for essential hypertension using a smartphone application: a randomized, open‐label, multicenter pilot study. J Clin Hypertens 2021 Jan 23;23(5):923-934. [doi: 10.1111/jch.14191]
40. Abu-El-Noor N, Aljeesh Y, Bottcher B, Abu-El-Noor M. Impact of a mobile phone app on adherence to treatment regimens among hypertensive patients: a randomised clinical trial study. European Journal of Cardiovascular Nursing. Eur J Cardiovasc Nurs 2021 Jun;20(5):428-435. [doi: 10.1177/1474515120938235]
41. Alsaqer K, Bebis H. Self-care of hypertension of older adults during COVID-19 lockdown period: a randomized controlled trial. Clin Hypertens 2022 Jul 15;28:21. [doi: 10.1186/S40885-022-00204-7]
42. Burnier M. Medication adherence and persistence as the cornerstone of effective antihypertensive therapy. Am J Hypertens 2006 Nov;19(11):1190-1196. [doi: 10.1016/j.amjhyper.2006.04.006] [Medline: 17070434]
43. Bartholomew LK, Mullen PD. Five roles for using theory and evidence in the design and testing of behavior change interventions. J Public Health Dent 2011 Mar 18;71 Suppl 1(s1):S20-S33. [doi: 10.1111/j.1752-7325.2011.00223.x] [Medline: 21656946]
44. Whelton PK, Carey RM, Aronow WS, Casey DEJ, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension 2018 Jun;71(6):e13-115 [FREE Full text] [doi: 10.1161/HYP.0000000000000065] [Medline: 29133356]
45. Fu J, Liu Y, Zhang L, Zhou L, Li D, Quan H, et al. Nonpharmacologic interventions for reducing blood pressure in adults with prehypertension to established hypertension. J Am Heart Assoc 2020 Oct 20;9(19):e016804 [FREE Full text] [doi: 10.1161/JAHA.120.016804] [Medline: 32975166]
46. Cotie LM, Prince SA, Elliott CG, Ziss MC, McDonnell LA, Mullen KA, et al. The effectiveness of eHealth interventions on physical activity and measures of obesity among working-age women: a systematic review and meta-analysis. Obes Rev 2018 Oct;19(10):1340-1358. [doi: 10.1111/obr.12700] [Medline: 30156044]
47. Hui CY, Walton R, McKinstry B, Jackson T, Parker R, Pinnock H. The use of mobile applications to support self-management for people with asthma: a systematic review of controlled studies to identify features associated with clinical effectiveness and adherence. J Am Med Inform Assoc 2017 May 01;24(3):619-632 [FREE Full text] [doi: 10.1093/jamia/ocw143] [Medline: 27694279]

XSL•FO
RenderX

48. Ramallo-Fariña Y, García-Bello MA, García-Pérez L, Boronat M, Wägner AM, Rodríguez-Rodríguez L, et al. Effectiveness of internet-based multicomponent interventions for patients and health care professionals to improve clinical outcomes in type 2 diabetes evaluated through the INDICA study: multiarm cluster randomized controlled trial. JMIR Mhealth Uhealth 2020 Nov 02;8(11):e18922 [FREE Full text] [doi: 10.2196/18922] [Medline: 33136059]
49. Lee EK, Poon P, Yip BH, Bo Y, Zhu MT, Yu CP, et al. Global burden, regional differences, trends, and health consequences of medication nonadherence for hypertension during 2010 to 2020: a meta-analysis involving 27 million patients. J Am Heart Assoc 2022 Sep 06;11(17):e026582 [FREE Full text] [doi: 10.1161/JAHA.122.026582] [Medline: 36056737]
50. Ma H, Wang A, Pei R, Piao M. Effects of habit formation interventions on physical activity habit strength: meta-analysis and meta-regression. Int J Behav Nutr Phys Act 2023 Sep 12;20(1):109 [FREE Full text] [doi: 10.1186/s12966-023-01493-3] [Medline: 37700303]
51. Lally P, van Jaarsveld CH, Potts HW, Wardle J. How are habits formed: modelling habit formation in the real world. Eur J Soc Psychol 2009 Jul 16;40(6):998-1009. [doi: 10.1002/ejsp.674]
52. Glanz K, Bishop DB. The role of behavioral science theory in development and implementation of public health interventions. Annu Rev Public Health 2010;31:399-418. [doi: 10.1146/annurev.publhealth.012809.103604] [Medline: 20070207]
53. Lippke S, Ziegelmann JP. Theory - based health behavior change: developing, testing, and applying theories for evidence - based interventions. Appl Psychol 2008 Jul 08;57(4):698-716. [doi: 10.1111/j.1464-0597.2008.00339.x]
54. Peng S, Yuan F, Othman AT, Zhou X, Shen G, Liang J. The effectiveness of e-health interventions promoting physical activity and reducing sedentary behavior in college students: a systematic review and meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2022 Dec 25;20(1):318 [FREE Full text] [doi: 10.3390/ijerph20010318] [Medline: 36612643]
55. Borrelli B. The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. J Public Health Dent 2011;71 Suppl 1:S52-S63. [Medline: 21656954]
56. Al-Durra M, Torio MB, Cafazzo JA. The use of behavior change theory in internet-based asthma self-management interventions: a systematic review. J Med Internet Res 2015 Apr 02;17(4):e89 [FREE Full text] [doi: 10.2196/jmir.4110] [Medline: 25835564]
57. Salas-Groves E, Galyean S, Alcorn M, Childress A. Behavior change effectiveness using nutrition apps in people with chronic diseases: scoping review. JMIR Mhealth Uhealth 2023 Jan 13;11:e41235 [FREE Full text] [doi: 10.2196/41235] [Medline: 36637888]
58. Avery L, Flynn D, van Wersch A, Sniehotta FF, Trenell MI. Changing physical activity behavior in type 2 diabetes: a systematic review and meta-analysis of behavioral interventions. Diabetes Care 2012 Dec;35(12):2681-2689 [FREE Full text] [doi: 10.2337/dc11-2452] [Medline: 23173137]
59. Tang MY, Smith DM, Mc Sharry J, Hann M, French DP. Behavior change techniques associated with changes in postintervention and maintained changes in self-efficacy for physical activity: a systematic review with meta-analysis. Ann Behav Med 2019 Aug 16;53(9):801-815. [doi: 10.1093/abm/kay090] [Medline: 30534971]
60. Evangelidis N, Craig J, Bauman A, Manera K, Saglimbene V, Tong A. Lifestyle behaviour change for preventing the progression of chronic kidney disease: a systematic review. BMJ Open 2019 Oct 28;9(10):e031625 [FREE Full text] [doi: 10.1136/bmjopen-2019-031625] [Medline: 31662393]
61. Dombrowski SU, Sniehotta FF, Avenell A, Johnston M, MacLennan G, Araújo-Soares V. Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. Health Psychol Rev 2012 Mar;6(1):7-32. [doi: 10.1080/17437199.2010.513298]
62. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
63. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018 Mar 01;8(2):212-224 [FREE Full text] [doi: 10.1093/tbm/ibx019] [Medline: 29381786]

Abbreviations

BCT: behavior change technique
BP: blood pressure
DBP: diastolic blood pressure
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SBP: systolic blood pressure

XSL•FO
RenderX

Edited by E Sezgin; submitted 29.11.23; peer-reviewed by B Green, S Payne; comments to author 10.03.24; revised version received 29.05.24; accepted 19.08.24; published 22.10.24.

Please cite as:
Zhou Y, Li SJ, Huang RQ, Ma HM, Wang AQ, Tang XY, Pei RY, Piao MH
Behavior Change Techniques Used in Self-Management Interventions Based on mHealth Apps for Adults With Hypertension: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2024;26:e54978
URL: https://www.jmir.org/2024/1/e54978
doi:10.2196/54978
PMID:

©You Zhou, Si-Jia Li, Ren-Qian Huang, Hao-Ming Ma, Ao-Qi Wang, Xing-Yi Tang, Run-Yuan Pei, Mei-Hua Piao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of HIV Self-Testing on Testing Promotion and Risk Behavior Reduction Among Transgender Women in China: Randomized Controlled Trial

Yan-Yan Zhu[1,2,3*], BN; Ze-Hao Ye[4*], MSc; Zhen-Xing Chu[1,2,3*], PhD; Yingjie Liu[1,2,3], PhD; Jie Wei[1,2,3], MPH; Le Jia[1,2,3], MPH; Yong-Jun Jiang[1,2,3], PhD; Hong Shang[1,2,3*], PhD; Qing-Hai Hu[1,2,3*], PhD

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, NHC Key Laboratory of AIDS Prevention and Treatment, National Clinical Research Center for Laboratory Medicine, The First Hospital of China Medical University, China Medical University, Shenyang, China

[2]Key Laboratory of AIDS Immunology, Chinese Academy of Medical Sciences, Shenyang, China

[3]Key Laboratory of AIDS Immunology of Liaoning Province, Shenyang, China

[4]Ningbo Municipal Centre for Disease Control and Prevention, Ningbo, China

[*]these authors contributed equally

Corresponding Author:
Qing-Hai Hu, PhD
State Key Laboratory for Diagnosis and Treatment of Infectious Diseases
NHC Key Laboratory of AIDS Prevention and Treatment, National Clinical Research Center for Laboratory Medicine
The First Hospital of China Medical University, China Medical University
155 Nanjing North Street
Heping District
Shenyang, 110001
China
Phone: 86 024 83282192
Email: qhhu@cmu.edu.cn

Abstract

Background: To date, no randomized controlled trials have specifically addressed behavior changes after HIV self-testing (HIVST) among transgender women.

Objective: This study aims to evaluate the effects of HIVST on changes in HIV testing behavior, frequency of condomless sex, and partner numbers among transgender women in China.

Methods: Participants were recruited from 2 Chinese cities using both online and offline methods. Transgender women were randomly assigned to receive an HIVST intervention. Data from the previous 3 months were collected at baseline, 3 months, and 6 months. The primary outcome was the mean change in the number of HIV tests among transgender women during the 6-month follow-up. An intention-to-treat analysis was conducted. The statistical analysis used analysis of covariance and linear mixed-effects models.

Results: From February to June 2021, and 255 transgender women were recruited, of which only 36.5% (93/255) had a steady job, and 27.1% (69/255) earned less than US $414.9 of income per month. They were randomly assigned to the intervention (n=127) and control (n=128) groups. At 6 months, the mean number of HIV tests was 2.14 (95% CI 1.80-2.48) in the intervention group and 1.19 (95% CI 0.99-1.40) in the control group (P<.001), with increases of 0.84 (95% CI 0.54-1.14) and 0.11 (95% CI –0.19-0.41) over 6 months, respectively. The net increase was 0.73 (95% CI 0.31-1.15; P<.001), with a similar adjusted result. No significant differences in the frequency of condomless sex or partner numbers were observed between the 2 groups.

Conclusions: HIVST is an effective strategy for enhancing regular HIV testing behavior among transgender women in China. This strategy should be combined with measures to address the financial vulnerability of the transgender women community to reduce subsequent risk behaviors, including condomless sex.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000039766; https://www.chictr.org.cn/showproj.html?proj=61402

(J Med Internet Res 2024;26:e58591) doi:10.2196/58591

XSL•FO
RenderX

KEYWORDS

HIV; HIV self-testing; testing behavior; sexual behaviours; transgender women; sexual health; mobile phone

Introduction

Transgender women are individuals assigned male at birth who identify as female and have a disproportionate demographic burden of the HIV worldwide; a meta-analysis encompassing data from 15 countries yielded an estimated HIV prevalence of 19.1% among transgender women, with an OR of 48.8 for HIV infection relative to the general population [1]. Data collated between 2018 and 2022 from 9 countries in Asia and the Pacific indicated a median HIV prevalence among transgender women of 3.9%, which was 19.5 times higher than that among general adults (0.2%) [2]. Notably, a study from Shenyang, China, found an HIV prevalence of 24.7% among transgender women [3], significantly exceeding that of other key populations, such as men who have sex with men (MSM) in China [4]. Minority stress and adverse mental health conditions frequently lead transgender women to engage in condomless sex with both regular and casual partners [5-7], elevating their risk of HIV transmission. HIV testing is the gateway to HIV prevention and treatment services; however, transgender women often encounter considerable obstacles to accessing HIV testing services. In addition to financial problems [8], issues of confidentiality, stigma, and discrimination by health care providers in facility-based HIV testing settings are prevalent [9-12], thereby hindering the use of HIV testing and related health care services by transgender women and obstructing efforts to eradicate AIDS.

To achieve the UNAIDS 95-95-95 targets for HIV [2], strategies must prioritize transgender women, who exhibit higher rates of prevalence and transmission risk. HIV self-testing (HIVST) offers greater convenience and privacy than conventional facility-based testing, mitigating stigma and discrimination from health care providers [13]. Randomized controlled trials (RCTs) have confirmed the efficacy and safety of HIVST in scaling up testing services, primarily among MSM and their partners [14-16]. The World Health Organization endorsed HIVST in 2016 to expand testing services [17], which were subsequently incorporated into the *National Guidelines for Detection of HIV/AIDS* in China [18]. Despite the high acceptability of HIVST among transgender women [19,20], high-quality evidence is needed to assess its impact on HIV testing behaviors.

Furthermore, studies have reported inconsistent findings on sexual behaviors and partner numbers post HIVST among MSM and female sex workers [21-23]. A Chinese RCT noted a marked reduction in sexual partners following HIVST among MSM but no change in condomless sex [16]. Research on the post-HIVST behavioral changes among transgender women is scant, with only 1 qualitative study yielding inconclusive results regarding condom use [24]. Additional research is needed to investigate the role of HIVST in health promotion strategies.

We conducted this RCT to evaluate the effects of HIVST on regular HIV testing behaviors among transgender women in China over 6 months. Secondary objectives included assessing the effect of HIVST on the frequency of condomless sex and partner numbers.

Methods

Study Design and Participant Eligibility

An open-label RCT was conducted at 2 sites in China (Shenyang and Beijing), using both online and offline recruitment methods (the distribution of 255 transgender women is presented in Figure S1 in Multimedia Appendix 1). The first participant was enrolled in this trial on February 25, 2021, and the last on June 24, 2021. This trial was ended on January 5, 2022. Digital informed consent was obtained from all participants. The trial protocol (Multimedia Appendix 2) was approved by the ethics committee of the First Hospital of China Medical University in Shenyang and adhered to the CONSORT (Consolidated Standards of Reporting Trials) 2010 guidelines as detailed in Multimedia Appendix 3.

We recruited transgender women through community-based organizations (CBOs) at both study sites. An initial online questionnaire screened potential participants referred by CBOs, as well as 10 preconfirmed transgender women. The inclusion criteria were age between 18 and 65 years, assigned male at birth, and self-identification as female. Verification of transgender women identity included responses to specific questions and, when necessary, photographic evidence in women's attire. Eligible candidates also reported proficiency in using smartphones and engagement in sexual activity within the past 6 months. The exclusion criteria encompassed a positive HIV status before enrollment and refusal to participate in this trial. All qualifying HIV-negative transgender women were subsequently admitted as participants. Throughout this trial, stringent measures were used to preserve the confidentiality of participants' personal information. All participants retained the right to withdraw from this trial at any point without incurring any negative repercussions.

Randomization and Masking

We simply randomized participants to either the intervention or control group (1:1). Random numbers were generated using the RANDBETWEEN function in Microsoft Excel 2021 and subsequently concealed with sequentially numbered, opaque envelopes. The investigator opened these envelopes immediately before allocation. Each participant received a unique identification number based on their sequence of enrollment (1-128) and group assignment, with numbers in the intervention group and the control group commencing at "500" and "600", respectively (eg, 500007 and 600007). Due to the nature of the distinct interventions, blinding of participants was not feasible.

Procedures

Participants were instructed to add a WeChat account created specifically for this trial. Candidates were asked to complete an online baseline questionnaire distributed by study staff through the dedicated WeChat account and undergo HIVST we provided to exclude those with positive results, who were then referred for confirmatory testing and HIV care. To enhance the acceptability of the questionnaire, we invited 10 transgender

XSL•FO
RenderX

women to revise sensitive content and inappropriate expressions before this trial.

After randomization, participants assigned to the intervention group could apply for 1 to 5 complimentary quadruplex self-testing kits from Wondlfo, Guangzhou, China (details are described in Multimedia Appendix 2), both at baseline and at the 3-month follow-up. Each kit had an attached label containing the participant's unique identification number, HIVST kit number, and a QR code for anonymous result feedback. HIVST kits and instruction cards were packaged without any personal or HIV-related information. We also sent operation instructions through WeChat, including online support for any inquiries. Participants could report their HIVST results by scanning the QR code and uploading photos of the test and label. Participants in both the intervention and control groups attended follow-up visits at 3 and 6 months; during this trial, participants received regular HIV-related health information (eg, free HIV testing locations, window period for HIV seroconversion, benefits of regular HIV testing). Condoms and lubricants were offered to all participants at no cost. Confirmatory testing at clinics was arranged for individuals with positive HIVST kit results; those with confirmatory positive results were linked to HIV care.

Participants were followed up online at 3-month intervals, and online surveys were conducted to assess their HIV testing experiences and related risk behaviors during the previous 3 months. Follow-up questionnaires were disseminated through the dedicated WeChat account. During follow-up, the identity of participants was confirmed using the unique identification number assigned to each individual at the time of enrollment. Participants were asked to upload an image of their test results within 48 hours using the QR code provided for feedback. Reminder messages were sent to participants who failed to upload their results within 1 week. Participants who failed to upload their results within 2 weeks became ineligible to request additional kits. Each participant received financial compensation of US $6.9 for completing each follow-up survey, with an additional US $1.4 awarded for the successful upload of a test result photo.

Outcomes

We recommended that individuals at risk of HIV have regular HIV testing every 3 months. At the 3- and 6-month follow-ups, we collected the number of HIV tests over the past 3 months to evaluate the behaviors associated with regular HIV testing. The primary outcome was the mean changes in the number of HIV tests among transgender women during 6 months of follow-up, which included both HIVST and facility-based HIV testing. The secondary outcomes included the mean changes in the number of HIV tests during 3 months of follow-up, the mean changes in the frequency of condomless sex, and the number of partners during 6 months of follow-up. The results were self-reported by transgender women. Data were collected through online questionnaires at baseline and at the 3- and 6-month follow-ups, including sociodemographic characteristics, perception of HIVST, number of sexual partners, sexual behaviors, and frequency of HIV testing in the past 3 months. We reminded transgender women by WeChat text messages to complete follow-up questionnaires.

Statistical Analysis

We used PASS (version 15.0; NCSS, LLC) to calculate the sample size based on a repeated measures design with 2 time points. Enrollment of 103 participants per group was projected to provide 90% power to detect a mean difference of 0.35 in the number of HIV tests per 3 months [15] between groups (2-tailed α=.05). At least 115 patients per group were needed assuming an anticipated 10% dropout rate.

An intention-to-treat analysis was conducted, in which outcomes were compared between participants according to their randomization assignment rather than their actual application for the intervention. To evaluate the effect of HIVST on mean changes in the number of HIV tests among transgender women (primary outcome) and their partner number and frequency of condomless sex (secondary outcome), we compared the outcome variables between the 2 groups. We also evaluated the number of HIV tests by recoding the number of tests into 2 binary variables: yes or no.

Net differences in mean changes from baseline in outcome variables between groups were analyzed using covariance analysis, with the group as a factor and baseline value as a covariate. Adjusted net differences were adjusted for important covariables, including age, ethnicity, education, occupation, income, marital status, transient population, and baseline value (HIV test number, HIV testing history, partner number, or sexual behavior number). A step-down approach was used to find a significant result; the primary and secondary outcomes were tested sequentially to control the familywise type I error rate according to the following order: mean changes in the number of HIV tests at 6 months (primary outcome), mean changes in the number of HIV tests at 3 months, mean changes in the sexual partner number at 6 months, and mean changes in the frequency of condomless sex at 6 months. Statistical significance (2-sided $P<.05$) at each step was required to test the next hypothesis. We used a linear mixed-effects model with individuals as a random effect to evaluate the main effects of the intervention and interaction effects of time on the primary outcome.

Subgroup analyses for the primary outcome were performed based on age, education, occupation, income, marital status, and transient population. For sensitivity analysis, the results with imputation are explored. We assumed that data were missing completely at random and imputed the number of HIV tests, the number of sexual behaviors, and the number of sexual partners through predictive mean matching. Similar distributions of variable values before and after imputation are plotted in Figure S2 in Multimedia Appendix 1. All statistical analyses were performed using R software (version 4.2.2; R Foundation for Statistical Computing).

Ethical Considerations

This trial received approval from the ethics committee of the First Hospital of China Medical University in Shenyang, China (Approval number: (2020)2020-326-2). Participants were provided with digital informed consent and were given a unique identification number for confidentiality.

XSL•FO
RenderX

Results

Study Participants and Baseline Characteristics

From February 2021 to June 2021, a total of 255 participants who identified as transgender women were enrolled in this trial, with 127 assigned to the intervention group and 128 assigned to the control group (Figure 1). During the 6 months, 20 participants in the intervention group and 20 in the control group were lost to follow-up. The mean age of the participants was 31 (SD 9) years; out of the 255 participants, 140 (54.9%) participants had an education level of college or above, 199 (78%) participants were single, 93 (36.5%) participants had a steady job, and 69 (27.1%) participants earned less than US $414.9 per month. Out of the 255 participants, 180 (70.6%) had a history of HIV testing within the past 3 months, the mean number of HIV tests was 1.2 (SD 1.2), and 30 (11.8%) had a history of pre-exposure prophylaxis uptake (15 in the intervention group). The baseline characteristics of transgender women were similar between the groups (Table 1).

XSL•FO
RenderX

Figure 1. Flow chart of the randomized controlled trial. CBOs: community-based organizations; HIVST: HIV self-testing.

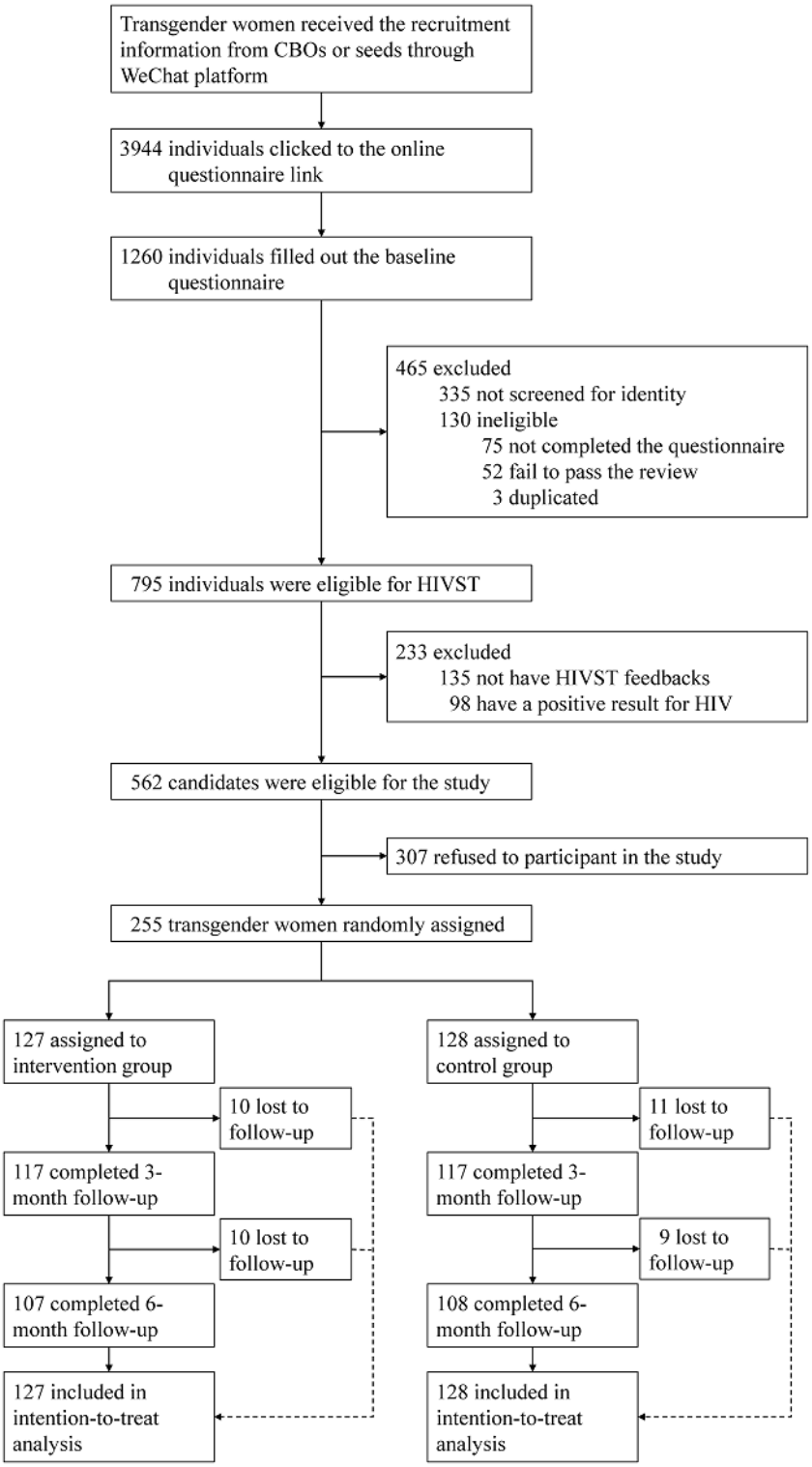

XSL•FO
RenderX

Table 1. Demographic and behavioral characteristics of transgender women at baseline.

Characteristics	Intervention (N=127)	Control (N=128)	*P* value
Recruitment method, n (%)			.74
Facility-based	94 (74)	98 (76.6)	
Web-based	33 (26)	30 (23.4)	
Age (years), n (%)			.88
18-25	35 (27.6)	37 (28.9)	
26-35	50 (39.4)	54 (42.2)	
36-45	30 (23.6)	25 (19.5)	
≥45	12 (9.4)	12 (9.4)	
Ethnicity, n (%)			.73
Han	112 (88.2)	110 (85.9)	
Others	15 (11.8)	18 (14.1)	
Education, n (%)			.34
College or above	74 (58.3)	66 (51.6)	
High school or below	53 (41.7)	62 (48.4)	
Occupation, n (%)			.13
Steady job	53 (41.7)	40 (31.2)	
Freelancer	31 (24.4)	47 (36.7)	
Student	20 (15.7)	22 (17.2)	
Others	23 (18.1)	19 (14.8)	
Monthly income (US $), n (%)			.76
<414.9	36 (28.3)	33 (25.8)	
414.9-829.6	55 (43.3)	51 (39.8)	
829.7-1244.5	20 (15.7)	26 (20.3)	
≥1244.6	16 (12.6)	18 (14.1)	
Marital status, n (%)			.45
Single	95 (74.8)	104 (81.2)	
Cohabitating	21 (16.5)	15 (11.7)	
Married	11 (8.7)	9 (7)	
Transient population, n (%)			.66
No	72 (56.7)	77 (60.2)	
Yes	55 (43.3)	51 (39.8)	
HIV testing over the past 3 months, n (%)			.82
No	36 (28.3)	39 (30.5)	
Yes	91 (71.7)	89 (69.5)	
HIV test numbers over the past 3 months, Mean (SD)	1.3 (1.3)	1.1 (1.1)	.45
Partner numbers over the past 3 months[a], Mean (SD)	6.7 (14.4)	5.9 (12.6)	.61
Condomless sex numbers over the past 3 months[a], Mean (SD)	1.8 (4.3)	1.7 (3.3)	.80

[a]Illogical values (1%) were deleted and imputed with medians.

Effects Analysis of Primary and Secondary Outcomes

Over 6 months, the number of HIV tests in the intervention group increased by an average of 0.84 tests (95% CI 0.54-1.14) and by 0.11 tests (95% CI –0.19 to 0.41) in the control group (Table 2). The net increase was 0.73 tests (95% CI 0.31-1.15; *P*<.001). After adjustment for potential covariates, this difference remained significant at 0.82 tests (95% CI 0.44-1.19;

XSL•FO
RenderX

P<.001). At 6 months, the average number of HIV tests was 2.14 (95% CI 1.80-2.48) in the intervention group and 1.19 (95% CI 0.99-1.40) in the control group (*P*<.001; Figure 2A; Table S1 in Multimedia Appendix 1). The main effect of the intervention over the 6 months compared with the control group was an increase of 0.77 tests (95% CI 0.47-1.07; *P*<.001) when the individual was included as a random effect in a linear mixed-effects model. The increase in the number of HIV tests among transgender women was significantly higher in the intervention group than in the control group at 6 months; however, the difference between groups remained stable over time (*P*>.05 for interaction between intervention and follow-up time; Figure 2A).

At 6 months, 93 (86.9%) of 127 transgender women in the intervention group and 77 (71.3%) of 128 transgender women in the control group had been tested for HIV in the previous 3 months, with a group difference of 15.6% (95% CI 4.8-26.4; *P*=.005; Table 2). The proportion of transgender women tested for HIV in the previous 3 months remained steady during the 6-month follow-up (*P*>.05 for interaction between intervention and follow-up time; Figure 2B). None of them were detected with an HIV-positive result. No significant differences were observed in changes in the average number of sexual partners or condomless sex from baseline to 6 months between the 2 groups (Table 2; Figures 2C and 2D).

Table 2. Effectiveness of HIV self-testing on outcomes among 255 transgender women.

	Proportion or mean change (95% CI)[a]		Net difference (95% CI)	*P* value	Adjusted net difference[b] (95% CI)	*P* value
	Intervention	Control				
Primary outcome						
Change in the number of HIV tests at 6 months.	0.84 (0.54 to 1.14)	0.11 (–0.19 to 0.41)	0.73 (0.31 to 1.15)	<.001	0.81 (0.44 to 1.18)	<.001
Secondary outcomes						
Change in the number of HIV tests at 3 months.	0.82 (0.57 to 1.08)	0.22 (–0.03 to 0.48)	0.60 (0.24 to 0.96)	.001	0.61 (0.28-0.93)	<.001
Proportion of HIV testing at 6 months.	86.9% (79.3 to 94.6)	71.3% (63.7 to 78.9)	15.6% (4.8 to 26.4)	.005	13.5% (3.1 to 23.9)	.01
Proportion of HIV testing at 3 months.	89.7% (82.8 to 96.7)	72.6% (65.7 to 79.6)	17.1% (7.2 to 27)	<.001	16.8% (7.3 to 26.3)	<.001
Change in partner numbers at 6 months.	–2.93 (–5 to –0.87)	–1.85 (–3.91 to 0.21)	–1.08 (–4 to 1.83)	.47	0.45 (–1.52 to 2.41)	.66
Change in the frequency of condomless sex at 6 months.	–1.11 (–1.81 to –0.42)	–0.59 (–1.28 to 0.10)	–0.52 (–1.50 to 0.46)	.30	0.05 (–0.34 to 0.44)	.80

[a]Indicators obtained at every visit through the questionnaire are situations in the previous 3 months (eg, the number of HIV tests in the previous 3 months were obtained at baseline, 3 months, and 6 months, respectively). The change of indicators in the outcomes are all based on values at baseline.

[b]Adjusted for age, ethnicity, education, occupation, income, marital, transient population, HIV testing number, HIV testing history, partner number, or sexual behavior number at baseline.

XSL•FO
RenderX

Figure 2. Behavior changes during the 6-month follow-up in the intervention and control groups.

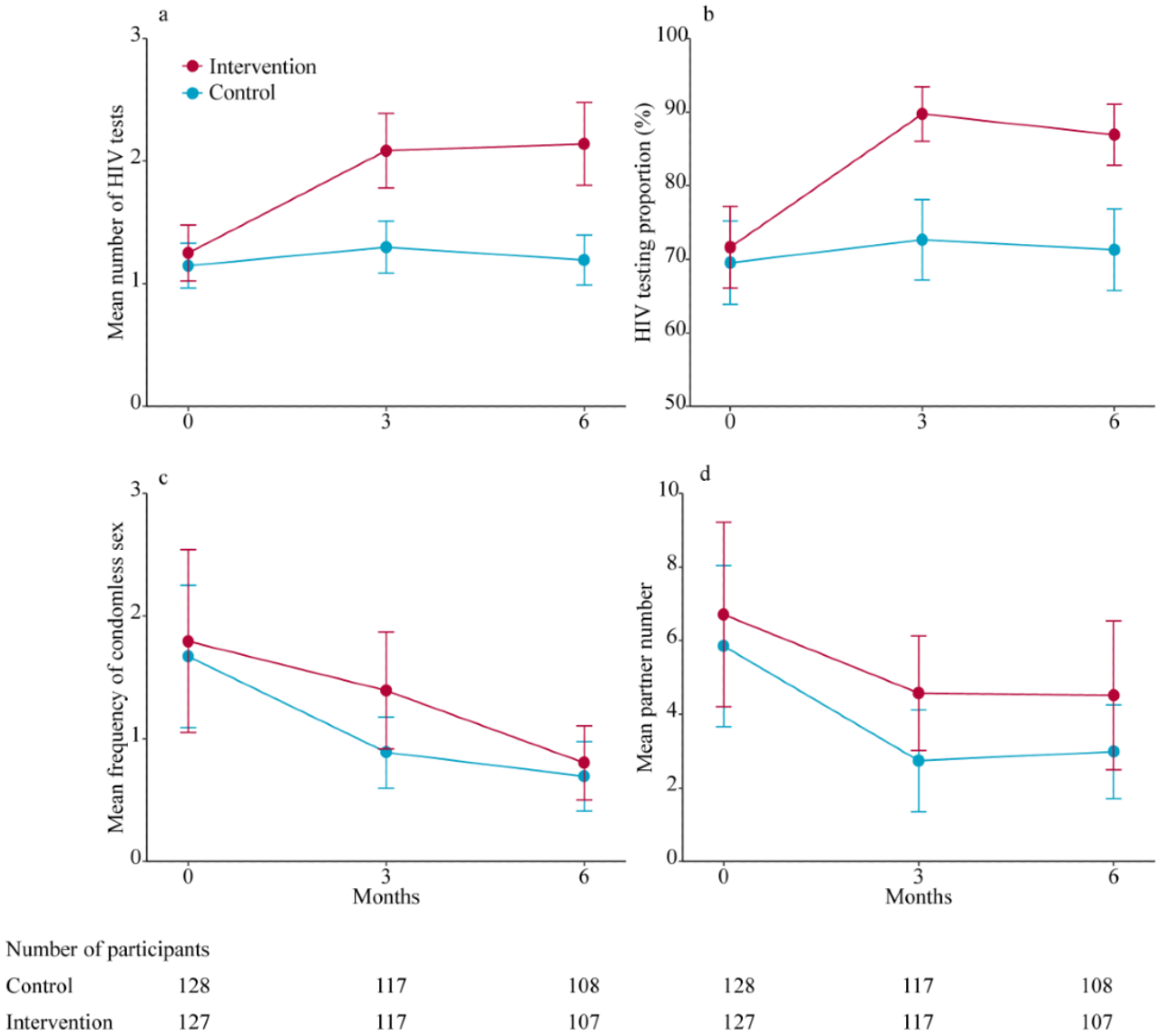

Sensitivity Analysis of the Primary Outcome

Mean differences in the number of HIV tests among transgender women between the groups at 6 months remained consistent across subgroups, including age, education, occupation, income, marital status, and transient population (Figure 3). After imputation for missing data, the results for the primary outcome remained consistent; the mean changes in the number of HIV tests among transgender women at 6 months were 0.84 (95% CI 0.56-1.11) in the intervention group and 0.10 (95% CI –0.17 to 0.38) in the control group, with a significant group difference of 0.73 (95% CI 0.34-1.12; P<.001; Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 3. Difference in the mean number of HIV tests at 6 months by subgroup. Error bars indicate 95% CIs.

	Intervention group		Control group			
	Numbers of participants	Mean change (95% CI)	Numbers of participants	Mean change (95% CI)	Adjusted group difference (95% CI)	*P* for interaction
All people	127	0.84 (0.54-1.14)	128	0.11 (−0.19-0.41)	0.82 (0.44-1.19)	
Age						
<35	85	0.75 (0.37-1.13)	91	0.12 (−0.24-0.48)	0.82 (0.34-1.30)	.91
≥35	42	1.00 (0.51-1.49)	37	0.09 (−0.44-0.62)	0.83 (0.24-1.43)	
Education						
>High school	74	1.00 (0.59-1.41)	66	−0.04 (−0.46-0.39)	1.09 (0.55-1.64)	.06
≤High school	53	0.62 (0.19-1.06)	62	0.28 (-0.13-0.68)	0.53 (0.05-1.01)	
Occupation						
Steady job	53	0.54 (0.14-0.95)	40	0.10 (−0.40-0.60)	0.48 (−0.09-1.05)	
Freelancer	31	1.19 (0.52-1.85)	47	0.34 (−0.20-0.88)	1.17 (0.45-1.89)	.45
Others	43	0.97 (0.44-1.50)	41	−0.14 (−0.65-0.38)	0.66 (−0.01-1.32)	
Income (US $)						
<829.7	91	0.72 (0.39-1.06)	84	0.05 (−0.28-0.39)	0.76 (0.34-1.17)	.52
≥829.7	36	1.13 (0.50-1.76)	44	0.24 (−0.37-0.85)	0.96 (0.15-1.78)	
Marital status						
Single	95	0.78 (0.42-1.14)	104	0.12 (−0.23-0.46)	0.74 (0.31-1.17)	.16
Non-single	32	1.04 (0.53-1.55)	24	0.10 (−0.47-0.66)	0.97 (0.28-1.66)	
Transient population						
Yes	55	0.79 (0.24-1.33)	51	0.23 (−0.34-0.79)	0.78 (0.09-1.48)	.95
No	72	0.88 (0.55-1.22)	77	0.03 (−0.29-0.35)	0.86 (0.46-1.27)	

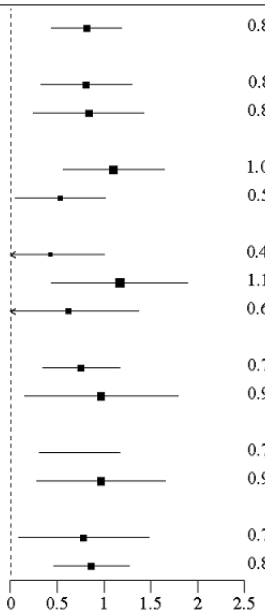

Discussion

Principal Findings

This open-label RCT demonstrated that HIVST was effective in improving HIV testing uptake among transgender women in China. There was a greater mean increase in the number of HIV tests among transgender women in the intervention group compared with those in the control group. A significant net increase in the proportion of HIV testing attributable to HIVST was observed between the 2 groups. However, its reliability as a strategy for reducing risk behaviors is questionable.

Comparison With Previous Work

In this trial, we observed an almost eightfold difference between the 2 groups in the mean change in the number of HIV tests over 6 months. The intervention led to a notable increase in HIV testing frequency over the previous 3 months (2.14 times per participant), which was greater than that in a trial among MSM in Seattle, Washington [15] and another trial among MSM in Hunan, China [14]; MSM randomized to HIVST reported an average 5.30 times of HIV testing frequency over 15 months (average 1.06 times per participant over 3 months) and an average 3.75 times of HIV testing frequency over 12 months (average 0.94 times per participant over 3 months), respectively. Transgender individuals experience additional disparities compared with their cisgender counterparts, such as increased risk behaviors, stigma, mental health issues, and negative health care experiences (potentially avoiding health care due to fear of mistreatment) [11,25]. This may account for the higher demand among transgender women for HIVST, a method that offers convenience and confidentiality. Similar to previous findings in other populations [16,26], we showed a significantly higher proportion of transgender women in the intervention group who underwent HIV testing. Considering the varied contexts and populations, these studies reported inconsistent effect sizes. Furthermore, the heightened frequency and proportion of regular HIV testing in this trial remained stable at the 3- and 6-month follow-ups, reflecting a sustained high demand for HIV testing among transgender women. Transgender individuals are often left behind in HIV-related programs in Asia and the Pacific [27]. This trial provides further evidence from China that HIVST is effective at establishing and improving regular HIV testing behavior among transgender women.

Despite the effectiveness of HIVST, the cost is one of the main concerns regarding its wider adoption by transgender women [19], particularly among transgender female sex workers [8]. In this trial, 35.7% of transgender women had a steady job and 27.1% had a low-level income. A survey in the United States reported that 19% of transgender individuals have engaged in sex work in their lifetimes and that nearly 29% live in poverty compared with 12% of the general US population [25]. Regular use of HIVST, not to mention partner screening, may be unaffordable for them [24]. Strategies to reduce or subsidize costs are crucial for future HIVST use among transgender women. Linkage to confirmatory testing and subsequent antiretroviral therapy (ART) initiation following a positive HIVST result is another critical challenge. An observational study in Bangkok revealed that nearly half of MSM and transgender women who first tested positive by HIVST failed to start ART [28]. Notably, community-based ART initiation, such as mobile- and home-based platforms, could have the potential to achieve higher ART initiation rates [29,30] and could also facilitate the provision of prevention services (eg, pre-exposure prophylaxis) following HIVST.

XSL•FO
RenderX

Previous studies reported inconsistent results on partner number or frequency of condomless sex following HIVST use [21-23]. In a qualitative study, transgender women reported varied decisions on condom use after receiving negative HIVST results from partners [24]. In this trial, we failed to find a statistically significant reduction in partner number or frequency of condomless sex among transgender women over a 6-month follow-up period. It may be unreliable to consider HIVST as a standalone strategy for promoting condom use among transgender women, as it appears to be more influenced by risk perception [31] and partner type [32].

Transgender women frequently occupy a vulnerable position within partnerships; the reasons for condomless sex include trust in their regular partners and opposition from casual and commercial partners [32]. Discrimination often leads to unemployment among transgender women worldwide, which sometimes makes them involved in high-risk sexual behaviors for income, including having multiple sex partners, condomless sex, and commercial sex [25,27]. Therefore, in addition to increasing the availability of condoms, interventions should address the financial vulnerability of the transgender women community to improve their ability to negotiate condom use within partnerships. For instance, the Avahan program has accomplished this through financial literacy training, assistance with opening savings accounts, and guidance on financial investments [33].

Failure to establish regular HIV testing behavior in key populations will easily lead to missed opportunities for ART initiation and then prolong the pandemic indefinitely, causing the affected community and society to incur substantial costs.

Strengths and Limitations

In this trial, we first showed the effects of HIVST on improving regular HIV testing behavior and subsequent changes in condomless sex and partner numbers among transgender women in China. However, this trial has several limitations. First, there was potential reporting bias due to self-defined identities among transgender women and the absence of standardized tools for assessing gender identity, as well as our partial reliance on self-reported HIVST results from participants. Second, the mixed online and offline recruitment in this trial may have introduced selection bias from diverse district sources. Furthermore, in this unblinded trial, participant recruitment through CBOs and 10 transgender women seeds may have led to contamination between the intervention and control groups, potentially reducing the difference in intervention effects. Third, most participants recruited offline by 2 CBOs in Shenyang and Beijing may have represented a subpopulation of MSM and engaged in other HIV testing programs, which may have overestimated regular HIV testing behavior among transgender women at baseline. Fourth, the reasons that candidates refused to participate in our trial included no fixed abode, which makes it hard to receive HIVST kits, unwilling to disclose their address, unwilling to participate in regular visits, which may affect their daily lives, and so on; regretfully, we did not systematically collect these causes. Furthermore, understanding why participants did not upload their results would be helpful to inform future approaches to developing research queries for this population. Regretfully, we did not collect specific reasons. Fifth, this trial was conducted during the COVID-19 pandemic. Public health restrictions in China affect the activities of most individuals, including transgender women. Therefore, sexual activity and healthcare access were, to a certain degree, limited among transgender women from both groups. However, the degree of the limitation, we think, is balanced between the 2 groups in this RCT, which will not affect the effect of the intervention we observed between the 2 groups. Finally, our sample presented a regional concentration in Shenyang and Beijing (Figure S1 in Multimedia Appendix 1), limiting the generalizability of our findings.

Conclusions

This trial demonstrated that HIVST resulted in statistically significant improvements in HIV testing uptake among transgender women in China. This effective intervention strategy should be integrated with actions addressing the financial vulnerability of the transgender women community and strategies for treatment and prevention services to improve condom use and linkages to ART or pre-exposure prophylaxis.

Acknowledgments

We thank all participants and staff associated with participating in the trial. We especially thank Yang Guang Studio in Shenyang and Tian Yuan Studio in Beijing, who were closely involved in our work. The study was supported by the National Natural Science Foundation of China (82073620) and the Liaoning Revitalization Talents Program (XLYCYSZX1904). The funders had no role in trial design, data collection, analysis, and interpretation. All authors approve the manuscript and consent to its publication.

Authors' Contributions

YYZ, ZHY, and ZXC contributed equally to this work as the first authors. HS and QHH had the idea and designed the study. ZXC, YJL, JW, and LJ recruited patients and collected the data. YYZ, ZHY, and ZXC analyzed and interpreted the data. YYZ, ZXC, and YJL had a check for the analysis. YYZ and ZHY wrote the first draft of the article. HS, YJJ, and QHH revised the manuscript. HS and QHH contributed equally as corresponding authors. All authors had unrestricted access to the data and participated in data interpretation. All the authors contributed to subsequent drafts and agreed to submit the article for publication.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 216 KB - jmir_v26i1e58591_app1.docx]

Multimedia Appendix 2
Study protocol.
[DOCX File , 998 KB - jmir_v26i1e58591_app2.docx]

Multimedia Appendix 3
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[DOC File , 91 KB - jmir_v26i1e58591_app3.doc]

References

1. Baral SD, Poteat T, Strömdahl S, Wirtz AL, Guadamuz TE, Beyrer C. Worldwide burden of HIV in transgender women: a systematic review and meta-analysis. Lancet Infect Dis 2013 Mar;13(3):214-222. [doi: 10.1016/S1473-3099(12)70315-8] [Medline: 23260128]
2. The path that ends AIDS: UNAIDS Global AIDS update 2023. UNAIDS. URL: https://tinyurl.com/ypsb393x [accessed 2023-11-14]
3. Wang H, Zeng C, Li X, Chang R, Yu X, Xu C, et al. Violence victimization and condomless anal intercourse among transgender women in China: mediating role of depression. AIDS Behav 2021 May;25(5):1597-1605. [doi: 10.1007/s10461-020-03076-w] [Medline: 33215274]
4. Global data on HIV epidemiology and response. UNAIDS. 2022. URL: https://aidsinfo.unaids.org/ [accessed 2023-11-14]
5. She R, Mo PK, Ma T, Liu Y, Lau JT. Impact of minority stress and poor mental health on sexual risk behaviors among transgender women sex workers in Shenyang, China. AIDS Behav 2021 Jun;25(6):1790-1799. [doi: 10.1007/s10461-020-03108-5] [Medline: 33438152]
6. Shan D, Yu M, Yang J, Zhuang M, Ning Z, Liu H, et al. Correlates of HIV infection among transgender women in two Chinese cities. Infect Dis Poverty 2018 Dec 01;7(1):123 [FREE Full text] [doi: 10.1186/s40249-018-0508-2] [Medline: 30509315]
7. Wang H, Chang R, Shen Q, Tsamlag L, Zhang S, Shi Y, et al. Information-Motivation-Behavioral Skills model of consistent condom use among transgender women in Shenyang, China. BMC Public Health 2020 Mar 26;20(1):394 [FREE Full text] [doi: 10.1186/s12889-020-08494-y] [Medline: 32216775]
8. Giguere R, Lopez-Rios J, Frasca T, Lentz C, Balán IC, Dolezal C, et al. Use of HIV self-testing kits to screen clients among transgender female sex workers in New York and Puerto Rico. AIDS Behav 2020 Feb;24(2):506-515 [FREE Full text] [doi: 10.1007/s10461-019-02730-2] [Medline: 31865516]
9. Grant JM, Mottet LA, Justin T, Min D, Jack H, Jody LH, et al. Injustice at every turn: a report of the national transgender discrimination survey. Washington: National Center for Transgender Equality and National Gay and Lesbian Task Force. 2011. URL: https://transequality.org/sites/default/files/docs/resources/NTDS_Report.pdf [accessed 2023-11-14]
10. Socías ME, Marshall BDL, Arístegui I, Romero M, Cahn P, Kerr T, et al. Factors associated with healthcare avoidance among transgender women in Argentina. Int J Equity Health 2014 Sep 27;13(1):81 [FREE Full text] [doi: 10.1186/s12939-014-0081-7] [Medline: 25261275]
11. Hibbert MP, Wolton A, Weeks H, Ross M, Brett CE, Porcellato LA, et al. Psychosocial and sexual factors associated with recent sexual health clinic attendance and HIV testing among trans people in the UK. BMJ Sex Reprod Health 2020 Apr;46(2):116-125 [FREE Full text] [doi: 10.1136/bmjsrh-2019-200375] [Medline: 31666303]
12. She R, Mo PKH, Cai Y, Ma T, Liu Y, Lau JTF. Mental health service utilisation among transgender women sex workers who are at risk of mental health problems in Shenyang, China: An application of minority stress theory. Health Soc Care Community 2022 Jul;30(4):e981-e993. [doi: 10.1111/hsc.13501] [Medline: 34240507]
13. Witzel TC, Wright T, McCabe L, Gabriel MM, Wolton A, Gafos M, et al. Impact and acceptability of HIV self-testing for trans men and trans women: A mixed-methods subgroup analysis of the SELPHI randomised controlled trial and process evaluation in England and Wales. EClinicalMedicine 2021 Feb;32:100700 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100700] [Medline: 33681732]
14. Zhang C, Koniak-Griffin D, Qian H, Goldsamt LA, Wang H, Brecht M, et al. Impact of providing free HIV self-testing kits on frequency of testing among men who have sex with men and their sexual partners in China: A randomized controlled trial. PLoS Med 2020 Oct;17(10):e1003365 [FREE Full text] [doi: 10.1371/journal.pmed.1003365] [Medline: 33035206]
15. Katz DA, Golden MR, Hughes JP, Farquhar C, Stekler JD. HIV self-testing increases HIV testing frequency in high-risk men who have sex with men: a randomized controlled trial. J Acquir Immune Defic Syndr 2018 Aug 15;78(5):505-512 [FREE Full text] [doi: 10.1097/QAI.0000000000001709] [Medline: 29697595]

XSL•FO
RenderX

16. Wang Z, Lau JTF, Ip M, Ho SPY, Mo PKH, Latkin C, et al. A randomized controlled trial evaluating efficacy of promoting a home-based HIV self-testing with online counseling on increasing HIV testing among men who have sex with men. AIDS Behav 2018 Jan;22(1):190-201. [doi: 10.1007/s10461-017-1887-2] [Medline: 28831616]
17. WHO Guidelines Approved by the Guidelines Review Committee. Guidelines on HIV Self-Testing and Partner Notification:Supplement to consolidated guidelines on HIV testing services. Geneva: World Health Organization; 2016 Dec. URL: https://iris.who.int/handle/10665/251655 [accessed 2023-11-15]
18. Chinese CDC. National guideline for detection of HIV/AIDS (version 2021). Chinese Journal of AIDS & STD 2021;27(11):1182-1201. [doi: 10.13419/j.cnki.aids.2021.11.02]
19. Lippman SA, Moran L, Sevelius J, Castillo LS, Ventura A, Treves-Kagan S, et al. Acceptability and feasibility of HIV self-testing among transgender women in San Francisco: a mixed methods pilot study. AIDS Behav 2016 Apr;20(4):928-938 [FREE Full text] [doi: 10.1007/s10461-015-1236-2] [Medline: 26511864]
20. Altaf A, Pasha MSK, Majeed A, Ali W, Alaama AS, Jamil MS. Acceptability and feasibility of HIV self-testing among transgender people in Larkana, Pakistan: Results from a pilot project. PLoS One 2022;17(7):e0270857 [FREE Full text] [doi: 10.1371/journal.pone.0270857] [Medline: 35802646]
21. Yan X, Su H, Zhang B, Li Y, Zhang L, Jia Z. Adherence of HIV self-testing among men who have sex with men in China: longitudinal study. J Med Internet Res 2020 Sep 17;22(9):e19627 [FREE Full text] [doi: 10.2196/19627] [Medline: 32940619]
22. Ortblad KF, Chanda MM, Mwale M, Haberer JE, McConnell M, Oldenburg CE, et al. Perceived knowledge of HIV-negative status increases condom use among female sex workers in Zambian transit towns. AIDS Patient Care STDS 2020 Apr;34(4):184-192 [FREE Full text] [doi: 10.1089/apc.2019.0266] [Medline: 32324483]
23. Tang W, Huang W, Lu H, Cao B, Wu D, Ong J, et al. What happens after HIV self-testing? Results from a longitudinal cohort of Chinese men who have sex with men. BMC Infect Dis 2019 Sep 14;19(1):807 [FREE Full text] [doi: 10.1186/s12879-019-4455-8] [Medline: 31521123]
24. Rael CT, Giguere R, Lopez-Rios J, Lentz C, Balán IC, Sheinfil A, et al. Transgender women's experiences using a home HIV-testing kit for partner-testing. AIDS Behav 2020 Sep;24(9):2732-2741 [FREE Full text] [doi: 10.1007/s10461-020-02829-x] [Medline: 32193729]
25. Friedman SR, Perlis T, Des Jarlais DC. Laws prohibiting over-the-counter syringe sales to injection drug users: relations to population density, HIV prevalence, and HIV incidence. Am J Public Health 2001 May;91(5):791-793. [doi: 10.2105/ajph.91.5.791] [Medline: 11344889]
26. Mulubwa C, Hensen B, Phiri MM, Shanaube K, Schaap AJ, Floyd S, HPTN 071 (PopART) Study Team. Community based distribution of oral HIV self-testing kits in Zambia: a cluster-randomised trial nested in four HPTN 071 (PopART) intervention communities. Lancet HIV 2019 Feb;6(2):e81-e92 [FREE Full text] [doi: 10.1016/S2352-3018(18)30258-3] [Medline: 30584047]
27. The gap report. UNAIDS. 2014. URL: https://tinyurl.com/yc7pusnf [accessed 2023-11-14]
28. Phanuphak N, Jantarapakde J, Himmad L, Sungsing T, Meksena R, Phomthong S, et al. Linkages to HIV confirmatory testing and antiretroviral therapy after online, supervised, HIV self-testing among Thai men who have sex with men and transgender women. J Int AIDS Soc 2020 Jan;23(1):e25448 [FREE Full text] [doi: 10.1002/jia2.25448] [Medline: 31960589]
29. Tun W, Apicella L, Casalini C, Bikaru D, Mbita G, Jeremiah K, et al. Community-based antiretroviral therapy (ART) delivery for female sex workers in Tanzania: 6-month ART initiation and adherence. AIDS Behav 2019 Sep;23(Suppl 2):142-152 [FREE Full text] [doi: 10.1007/s10461-019-02549-x] [Medline: 31197700]
30. MacPherson P, Lalloo DG, Webb EL, Maheswaran H, Choko AT, Makombe SD, et al. Effect of optional home initiation of HIV care following HIV self-testing on antiretroviral therapy initiation among adults in Malawi: a randomized clinical trial. JAMA 2014;312(4):372-379 [FREE Full text] [doi: 10.1001/jama.2014.6493] [Medline: 25038356]
31. De Santis JP, Hauglum SD, Deleon DA, Provencio-Vasquez E, Rodriguez AE. HIV risk perception, HIV knowledge, and sexual risk behaviors among transgender women in South Florida. Public Health Nurs 2017 May;34(3):210-218 [FREE Full text] [doi: 10.1111/phn.12309] [Medline: 27921325]
32. Moayedi-Nia S, Taheri L, Hosseini Rouzbahani N, Rasoolinejad M, Nikzad R, Eftekhar Ardebili M, et al. HIV prevalence and sexual behaviors among transgender women in Tehran, Iran. AIDS Behav 2019 Jun;23(6):1590-1593. [doi: 10.1007/s10461-018-02380-w] [Medline: 30734211]
33. Mahapatra B, Walia M, Patel SK, Battala M, Mukherjee S, Patel P, et al. Sustaining consistent condom use among female sex workers by addressing their vulnerabilities and strengthening community-led organizations in India. PLoS One 2020;15(7):e0235094 [FREE Full text] [doi: 10.1371/journal.pone.0235094] [Medline: 32609731]

Abbreviations

ART: antiretroviral therapy
CBO: community-based organization
CONSORT: Consolidated Standards of Reporting Trials
HIVST: HIV self-testing
MSM: men who have sex with men

XSL•FO
RenderX

RCT: randomized controlled trial

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by A Lale, G Downer; comments to author 29.06.24; revised version received 29.07.24; accepted 07.09.24; published 29.10.24.

Please cite as:
Zhu YY, Ye ZH, Chu ZX, Liu Y, Wei J, Jia L, Jiang YJ, Shang H, Hu QH
Effects of HIV Self-Testing on Testing Promotion and Risk Behavior Reduction Among Transgender Women in China: Randomized Controlled Trial
J Med Internet Res 2024;26:e58591
URL: https://www.jmir.org/2024/1/e58591
doi:10.2196/58591
PMID:

©Yan-Yan Zhu, Ze-Hao Ye, Zhen-Xing Chu, Yingjie Liu, Jie Wei, Le Jia, Yong-Jun Jiang, Hong Shang, Qing-Hai Hu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of King's Theory of Goal Attainment in Blood Glucose Management for Newly Diagnosed Patients With Type 2 Diabetes: Randomized Controlled Trial

Man Yan[1], MSc; Yingchun Yu[1], MSc; Shuping Li[1], BSc; Peiling Zhang[1], BSc; Jiaxiang Yu[2], PhD

[1]Changzhou Hospital of Traditional Chinese Medicine, Changzhou, China

[2]Department of Nursing and Rehabilitations, Faculty of Medicine and Health Sciences, University Putra Malaysia, Serdang, Malaysia

Corresponding Author:
Jiaxiang Yu, PhD
Department of Nursing and Rehabilitations
Faculty of Medicine and Health Sciences
University Putra Malaysia
Administrative building, 3rd Fl.
Serdang, 43400
Malaysia
Phone: 60 0178736091
Email: yuge1185@gmail.com

Abstract

Background: Diabetes poses a significant public health challenge in China and globally, with the number of patients expected to reach 592 million by 2035, notably in Asia. In China alone, an estimated 140 million individuals are living with diabetes, and a significant portion is nonadherent to medications, underscoring the urgency of effective management strategies. Recognizing the necessity of early and comprehensive management for newly diagnosed patients with type 2 diabetes, this study leverages an online teach-back method and "Internet + Nursing" platform based on King's Theory of Goal Attainment. The approach aims to enhance glycemic control and reduce fear and misconceptions about the disease, addressing both the educational and emotional needs of the patients.

Objective: The primary aim of this study was to assess the effectiveness of King's Goal Attainment Theory in the management of newly diagnosed patients with type 2 diabetes. This research sought to develop a collaborative model for blood glucose management, integrating the expertise and roles of physicians, nurses, and patients. The model is designed to enhance the synergy in health care provision, ensuring a comprehensive approach to diabetes management.

Methods: In this study conducted at Changzhou Traditional Chinese Medicine Hospital between January 2022 and February 2023, eligible patients were randomized into a control group or an online feedback group. The control group received standard care, while the online feedback group participated in a King's Theory of Goal Attainment–based online teach-back program, enhanced by "Internet + Nursing" strategies. This included an interactive platform for goal planning, video content sharing, comprehension assessment, misconception correction, and patient-driven recaps of disease information. Health monitoring was facilitated through the "Internet + Nursing" platform. The study focused on comparing changes in glucose metabolism and emotional disorder symptoms between the groups to evaluate the intervention's effectiveness.

Results: Following a 24-week intervention, we observed significant differences in key metrics between the online feedback group and the control group, each comprising 60 participants. The online feedback group demonstrated significant reductions in fasting plasma glucose, 2-hour postprandial glucose, and hemoglobin A_{1c} ($P<.05$). Additionally, there was a notable decrease in hypoglycemia-related anxiety and alexithymia within this group. Conversely, the control group maintained relatively higher values for these metrics at the same time point ($P<.05$). These findings underscore the efficacy of online feedback in managing glycemic control and reducing psychological distress associated with hypoglycemia.

Conclusions: The online teaching-back method, guided by King's Theory of Goal Attainment, effectively enhances glycemic control, reducing fasting plasma glucose, 2-hour postprandial glucose, and hemoglobin A_{1c} levels in newly diagnosed patients with type 2 diabetes. Simultaneously, it alleviates hypoglycemia-related anxiety and mitigates alexithymia. This approach merits widespread promotion and implementation in clinical settings.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400079547; https://www.chictr.org.cn/showproj.html?proj=208223

(J Med Internet Res 2024;26:e59142) doi:10.2196/59142

KEYWORDS

King's Theory of Goal Attainment; online feedback approach; newly diagnosed patients with type 2 diabetes mellitus; blood glucose control; type 2 diabetes; diabetes mellitus; blood glucose

Introduction

Diabetes has emerged as a major challenge in China's comprehensive prevention and management strategy, reflecting a global crisis with escalating prevalence worldwide [1,2]. The number of patients with diabetes is projected to rise to 592 million by 2035, with Asia, particularly China and India, at the epicenter of this pandemic. China currently leads with an estimated 140 million patients with diabetes [3,4]. Notably, a third of individuals with diabetes are nonadherent to prescribed medications. This highlights the severity and public health implications of diabetes management, now recognized as a critical global health issue.

Effective diabetes management is akin to a prolonged struggle [5,6]. There is a consensus among international scholars on the necessity of early, community-based, and multidimensional comprehensive management [7]. Particularly, patients with newly diagnosed type 2 diabetes often lack disease-specific knowledge and have misconceptions, leading to fear and inaccurate expectations about treatment outcomes.

At present, the multidisciplinary collaboration model has been widely recommended as an effective strategy for managing diabetes. Experts and scholars globally advocate for transforming diabetes care from traditional doctor-led approaches to multidisciplinary health care models. This transformation enables patients to work with professional health teams, including diabetes specialist nurses, psychological counselors, nutrition specialist nurses, and endocrinologists, to achieve optimal diabetes management through the implementation of healthy lifestyles and appropriate treatment interventions. Numerous studies have confirmed that personalized medical and nursing collaboration models can significantly improve patients' quality of life and reduce risk factors [8-10].

To address these challenges, this study introduces an online teach-back method using a bidirectional information model, implementing a dynamic, stepwise approach for subtle health education [11]. Using the "Internet + Nursing" platform, we developed an intervention based on King's Theory of Goal Attainment. This strategy aims to improve glycemic control and alleviate the fear of hypoglycemia and alexithymia in patients with newly diagnosed type 2 diabetes.

Methods

Study Population

This study involved first-onset patients with type 2 diabetes from Changzhou Traditional Chinese Medicine Hospital, recruited from January 2022 to February 2023, who met specific inclusion and exclusion criteria. Participants were randomly divided into a control group and an online feedback group using a random number table method. Inclusion criteria included (1) diagnosis of type 2 diabetes according to the 2017 Chinese Guidelines for the Prevention and Treatment of Diabetes; (2) ability to use basic mobile devices; (3) self-care capability, effective communication, and voluntary participation; and (4) access to a home blood glucose meter for self-monitoring. Exclusion criteria encompassed (1) acute diabetic complications or mental illnesses; (2) severe organ dysfunction; (3) malignant tumors; and (4) cognitive dysfunction, determined by Mini-Mental State Scale (MMSE) scores below 24 points. Additional exclusions during the study were due to personal dropout, withdrawal requests during follow-up, or unavailability for 2 consecutive follow-ups.

Ethical Considerations

This study was conducted with the ethics approval of the Ethics Committee of Changzhou Hospital of Traditional Chinese Medicine, in strict accordance with the guidelines and regulations governing research involving human participants (2022-L-L-035(s)). Informed consent was obtained from all participants involved in the study. The confidentiality of all personal information and data collected during the study was assured.

Control Group

Specialist diabetes nurses from the research team conducted comprehensive assessments of enrolled patients and established individual patient files. These files informed the development of personalized plans for diet, exercise, and glucose monitoring tailored to each patient's general health information and blood sugar levels. Additionally, patients received individualized guidance plan record manuals and educational logs.

Routine health education was provided to all participants. This includes monthly patient education classes organized by the department, personalized educational sessions, distribution of diabetes-related materials, and regular follow-up visits as part of the ongoing care.

Online Feedback Group

In addition to routine care, our hospital uses the "Internet + Nursing" platform to monitor patients' disease status using a feedback approach. Using this platform, we established an interactive station that leverages the King's standard theory for online teaching. This approach was further enhanced by problem-based learning techniques. We organize weekly group discussion activities, with the timing of these sessions coordinated and mutually agreed upon by team members. The research team was dedicated to managing and educating patients,

XSL•FO
RenderX

providing timely responses to queries, conducting real-time supervision, and facilitating reverse retellings for patients.

Management Process and Content

Informed Consent

Initially, we secured informed consent from the patient to ensure ethical standards and patient autonomy were upheld.

Comprehensive Assessment

We conducted a detailed questionnaire survey to understand the patient's diabetes-related knowledge, self-management practices, lifestyle, and concerns about hypoglycemia. The purpose of this assessment was to gather comprehensive information to tailor the educational intervention accurately. The questionnaire was designed to capture critical data, which was then used in follow-up interviews to formulate a personalized goal-reaching plan. This approach helped to reduce the occurrence of bias and ensured that the final plan was well-suited to each patient's needs.

Collaborative Goal Setting

We used a targeted questioning approach, such as the following:

> *Based on our established objectives, what areas do you need our assistance with for understanding? Which resources do you prefer for learning, e.g., videos or brochures?*

This facilitated a mutual planning process to meet the patient's goals, emphasizing cooperative efforts to achieve optimal glycemic control. The effectiveness of these tools was evaluated based on patient engagement and knowledge retention.

Educational Implementation

The educational strategy involved a structured process beginning with the distribution of video content, followed by organized review sessions, assessment of comprehension, clarification of misconceptions, and concluding with the patient's recapitulation of disease-related information. This methodical approach ensured patients were actively engaged in their learning journey, promoting better health outcomes.

Duration and Frequency

The intervention spanned 24 weeks, with weekly sessions lasting between 20 and 30 minutes each.

Content and Structure

Initial Sessions (1-3)

It focused on imparting knowledge about diabetes, facilitated by diabetes specialist nurses and nutritionists.

Middle Sessions (4-9)

It covered self-management skills, including diet, medication, self-monitoring, and safety practices.

Psychological Support (10-12)

It was conducted by a senior psychotherapist, these sessions address the patient's psychological well-being.

Doubt Clearance (13-16)

Doubt clearance had a diabetes specialist nurse engaged in question-and-answer sessions to resolve any patient queries.

Personalized Support (17-20)

Further individualized attention was given to strengthen knowledge and self-management skills.

Review and Summary (21-24)

Final sessions focus on answering outstanding questions and retrospective summarization. The intervention uses a hybrid of online one-on-one and group guidance for the initial 20 sessions. Specific strategies included the following.

Explanation

Using platforms like WeChat (Tencent Holdings Limited) for sharing video content, ensuring explanations were in layman's terms to enhance comprehension.

Assessment

The teach-back method was used by asking patients or family members to paraphrase educational content and demonstrate learned techniques, assess the patient's understanding of diabetes care–related knowledge and the accuracy of the blood sugar measurement process. For example, adopt an approachable attitude when asking the patient or family members questions like, "After attending this education session, can you tell me which fruits are safe to eat and which beverages should be avoided?" or "How and when do you usually measure your blood sugar?"

Clarification

Responsive feedback was provided, with misunderstandings addressed through further education and demonstration, ensuring clarity and accurate knowledge transfer.

Understanding

Open-ended questions facilitated the evaluation of comprehensive understanding and signaled the conclusion of the educational round when patients could independently recount the educational content.

Observation Indicators

Laboratory Biochemical Indicators

Blood samples were collected to assess the efficacy of the intervention on glucose metabolism. Specifically, 3 mL of venous blood was drawn from patients 1 day before the intervention and another 3 mL of fasting venous blood the following morning postintervention. Samples were centrifuged to separate serum, which was then stored at –20 °C until analysis. Key glucose metabolism markers analyzed included fasting blood glucose (FBG), 2-hour postprandial glucose (2hPG), and hemoglobin A_{1c} (HbA_{1c}).

Chinese Version of the Hypoglycemia Fear Survey II

The Chinese Version of the Hypoglycemia Fear Survey II (CHFS-II), adapted for Chinese populations by Mu Chun et al [12], is based on the original Hypoglycemia Fear Survey II. This version consists of the Chinese Hypoglycemia

XSL•FO
RenderX

Fear-Behavior Scale (CHFS-II-BS) and the Chinese Hypoglycemia Fear-Worry Scale (CHFS-II-WS), comprising 15 and 18 items, respectively. It assesses patients' fear and anxiety related to hypoglycemia experiences, using a 5-point Likert scale ranging from 0 ("never") to 4 ("always"). Scores span from 0 to 52, with higher scores indicating greater fear of hypoglycemia. The scale's internal consistency (Cronbach α) is 0.904, with a test-retest reliability of 0.911.

Toronto Alexithymia Scale

The Chinese adaptation of the Toronto Alexithymia Scale (TAS-20), translated by Jinyao et al [13], serves as a reliable measure for alexithymia. It features a Cronbach α of 0.83 and a test-retest reliability of 0.87. The scale, encompassing 20 items across 3 dimensions—difficulty identifying feelings, difficulty describing feelings, and externally oriented thinking—is scored on a 5-point scale with a total score range of 20 to 100. A score of 61 or above indicates the presence of alexithymia.

Quality Control Measures

Before the commencement of the project, all research personnel underwent standardized training to ensure a consistent understanding and application of the study procedures. To mitigate the risk of information distortion, standardized directives were used alongside a 3-tier data-cleaning methodology for meticulous data collection. Participant selection adhered rigorously to predefined inclusion and exclusion criteria, minimizing selection bias.

To further reduce sampling errors, participants were allocated into groups using a randomized method. Quality control of inspections was ensured through dedicated staff oversight and the regular calibration of instruments. Data integrity was reinforced by dual-entry verification, with 2 individuals independently recording and checking the data. Additionally, 15% of the data set was randomly selected for further validation to uphold data reliability.

Throughout the study, regular team meetings facilitated prompt communication among researchers, essential for the successful execution of the project. The research team consistently monitored treatment outcomes at predetermined intervals, systematically reviewed and verified data, and performed secondary analyses or exclusions as necessary to guarantee data precision.

Statistical Analysis Methodology

Data analysis was conducted using the SPSS statistical software (version 26.0; IBM Corp). Categorical data were summarized as frequencies and percentages. Continuous variables, including patients' FBG, 2-hour postprandial blood glucose, and glycosylated hemoglobin levels, were expressed as mean (SD). The independent sample *t* test was applied for the comparison of continuous variables between 2 groups, while the chi-square test was used for the analysis of categorical data. A *P* value of less than .05 was considered indicative of a statistically significant difference.

Results

Assessment Observations

Overview

A comprehensive assessment was conducted on 120 newly diagnosed patients with type 2 diabetes at the onset of the intervention, as described in the CONSORT (Consolidated Standards for Reporting Trials) diagram in Figure 1 (Multimedia Appendix 1). This assessment aimed to evaluate the participants' foundational knowledge of diabetes and their management practices. Baseline data between the 2 groups (each comprising 60 participants) showed no statistically significant difference ($P>.05$), confirming their comparability. Detailed baseline data for both groups are presented in Table 1.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards for Reporting Trials) diagram of participation flow in the study.

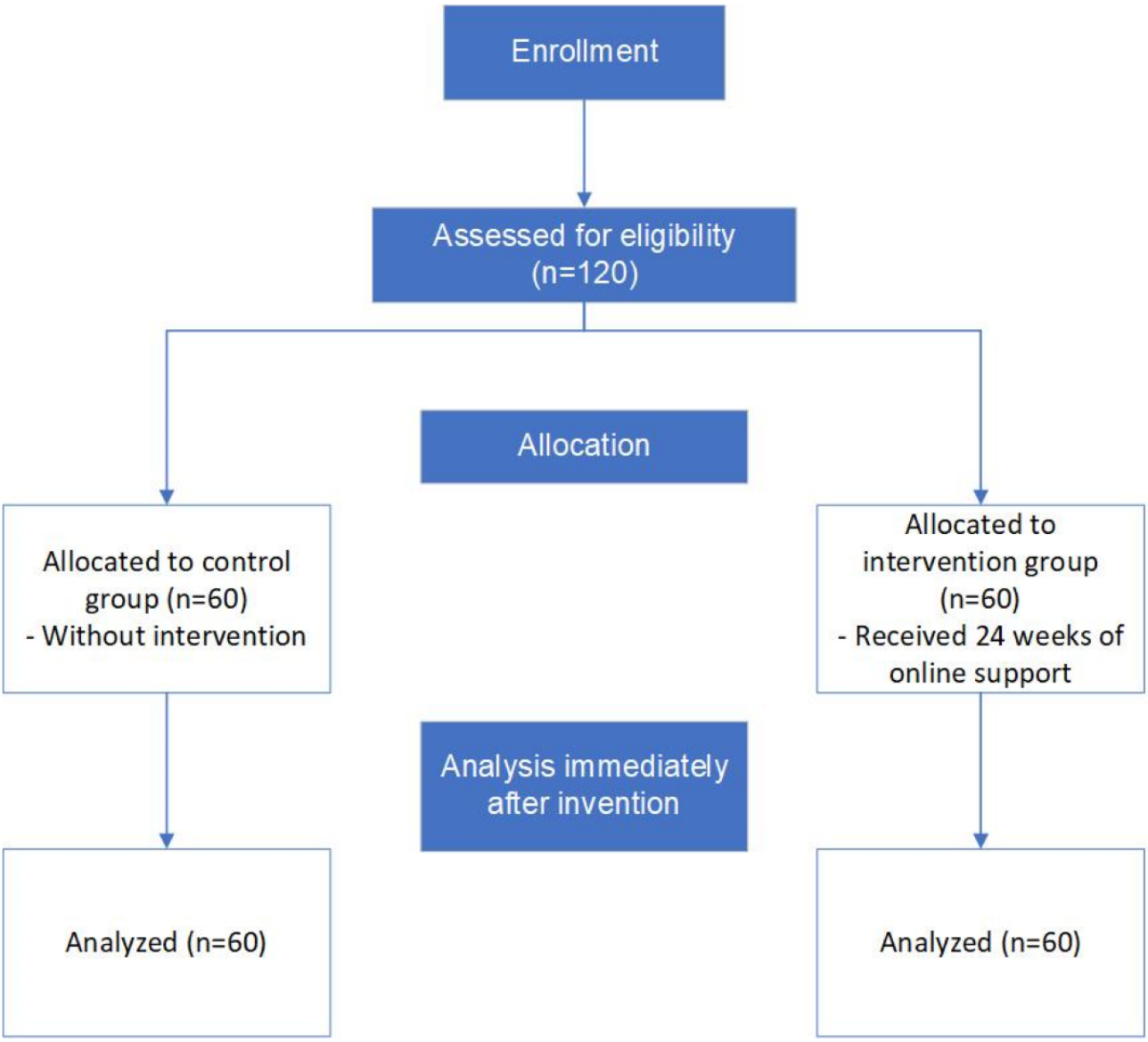

XSL•FO
RenderX

Table 1. Comparison of sociodemographic data of the 2 groups of research subjects.

Variables	Control group (n=60)	Online feedback group (n=60)	Statistical value	*P* value
Sex, n (%)			0.14[a]	>.05
Male	36 (60)	34 (57)		
Female	24 (40)	26 (43)		
Age (years), mean (SD)	45.17 (8.60)	47.88 (11.39)	–1.471 (118)[b]	.93
Education, n (%)			1.768[c]	.62
Primary school and below	6 (10)	3 (5)		
Junior high school and below	6 (10)	9 (15)		
High school or secondary school	27 (45)	25 (42)		
College and above	21 (35)	23 (38)		
Marital status, n (%)			0.789[a]	.67
Unmarried	5 (8)	7 (12)		
Married	49 (82)	45 (75)		
Divorced or widowed	6 (10)	8 (13)		
Payment method for medical expenses, n (%)			0.976[a]	.61
Municipal medical insurance	34 (57)	29 (48)		
District medical insurance	18 (30)	20 (33)		
Other	8 (13)	11 (18)		
Labor intensity, n (%)			0.311[c]	.86
Mild	20 (33)	29 (37)		
Moderate	34 (57)	20 (52)		
Severe	6 (10)	11 (12)		

[a]c^2.

[b]*t* test value.

[c]*Z* value.

Fear of Hypoglycemia

A significant proportion of participants (n=86, 72%) reported a fear of hypoglycemia. This fear often stemmed from a lack of understanding about how to manage low blood sugar levels and the potential consequences of hypoglycemia.

Misunderstandings About Dietary Requirements

Approximately 78 (65%) participants had misconceptions regarding dietary requirements. These misunderstandings included incorrect beliefs about carbohydrate restrictions, meal timing, and the role of diet in managing blood glucose levels.

Poor Adherence to Medication Regimens

About 70 (58%) participants demonstrated poor adherence to their prescribed medication regimens. This nonadherence was largely due to misunderstandings about the effects and importance of their medications, leading to inconsistent usage and suboptimal blood glucose control.

These observations were instrumental in customizing the educational content and intervention strategies for each patient. By addressing specific knowledge gaps and fears, the intervention was able to provide targeted education and support, thereby improving the overall effectiveness of diabetes management for the participants.

Among the participants, 78 (65%) opted to use videos as their primary learning tool, while 42 (35%) chose brochures. The video group demonstrated a higher level of engagement and better retention of information, as evidenced by their active participation in follow-up sessions and accurate recapitulation of disease-related information. Comparatively, the brochure group also showed improvements but at a slightly lower rate. The impact of these tools on glycemic control was significant, with the video group showing a more substantial reduction in FBG and HbA_{1c} levels compared to the brochure group ($P<.05$).

Comparative Analysis of Pre- and Postintervention Glycemic Indices

Following a 24-week intervention, both patient groups exhibited significant reductions in FBG scores ($P<.05$), with a notably greater decrease observed within the online feedback group ($P<.05$; refer to Table 2). In the online teaching group, pre- and postintervention comparisons revealed statistically significant differences in FBG, 2hPG, and HbA_{1c} scores ($P<.05$). Similar significant changes were observed in the control group for these

XSL•FO
RenderX

indices (*P*<.05). Postintervention, the glycemic indices of the online feedback group were significantly lower than those of the control group.

Table 2. Comparison of glucose metabolism index scores between the 2 groups before and after intervention.

Group	Count, n	FBG[a], mean (SD)		*t* test (*df*)	*P* value	2hPG[b], mean (SD)		*t* test (*df*)	*P* value	HbA_{1c}[c], mean (SD)		*t* test (*df*)	*P* value
		Before intervention[d]	After intervention[e]			Before intervention[f]	After intervention[g]			Before intervention[h]	After intervention[i]		
Control group	60	8.54 (2.32)	6.43 (1.45)	5.974 (59)	<.001	13.64 (4.34)	12.54 (3.35)	1.554 (59)	.12	7.94 (2.92)	7.07 (1.89)	1.934 (59)	.06
Online feedback group	60	8.39 (2.04)	5.74 (1.66)	7.804 (59)	<.001	13.31 (3.72)	10.88 (2.59)	4.153 (59)	<.001	7.89 (2.84)	6.03 (1.66)	4.38 (59)	<.001

[a]FBG: fasting blood glucose.

[b]2hPG: 2-hour postprandial glucose.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]t_{59}=0.376; *P*=.71.

[e]t_{59}=2.425; *P*=.02.

[f]t_{59}=0.447; *P*=.66.

[g]t_{59}=3.037; *P*=.003.

[h]t_{59}=0.095; *P*=.92.

[i]t_{59}=3.203; *P*=.002.

Comparative Analysis of Pre- and Postintervention Hypoglycemia Fear Scores

Following a 6-month intervention period, a notable decline in CHFS II -BS and CHFS II -WS scores was observed across both patient groups, with the reduction being significantly greater in the online feedback group (*P*<.05). Furthermore, after the intervention, scores for both CHFS II -BS and CHFS II -WS in the online feedback group were significantly lower than those recorded in the control group (*P*<.05), as detailed in Table 3.

Table 3. Comparison of hypoglycemic fear scores between the 2 groups before and after intervention.

Group	Count, n	CHFS II -BS[a], mean (SD)		*t* test (*df*)	*P* value	CHFS II -WS[b], mean (SD)		*t* test (*df*)	*P* value
		Before intervention[c]	After intervention[d]			Before intervention[e]	After intervention[f]		
Control group	60	24.53 (5.62)	20.43 (4.15)	4.546 (59)	<.001	18.23 (5.24)	16.51 (4.12)	1.999 (59)	.048
Online feedback group	60	23.29 (6.24)	17.54 (3.97)	6.022 (59)	<.001	19.34 (5.36)	12.39 (4.59)	7.629 (59)	<.001

[a]CHFS-II-BS: Chinese Hypoglycemia Fear-Behavior Scale.

[b]CHFS-II-WS: Chinese Hypoglycemia Fear-Worry Scale.

[c]t_{59}=1.144; *P*=.25.

[d]t_{59}=3.898; *P*<.001.

[e]t_{59}=–1.147; *P*=.25.

[f]t_{59}=5.174; *P*<.001.

Comparative Analysis of Pre- and Postintervention Alexithymia Levels

Before the intervention, baseline alexithymia levels were comparable between the groups. Post intervention, comparative analysis revealed a significant reduction in TAS-20 scores within both cohorts, indicative of decreased alexithymia levels. Notably, the reduction was more pronounced in the group receiving the specialized intervention (*P*<.05), compared to the control group (Table 4).

XSL•FO
RenderX

Table 4. Comparison of levels of affective disorder between the 2 groups before and after intervention.

Group	Count, n	TAS-20[a], mean (SD)		*t* test (*df*)	*P* value
		Before intervention[b]	After intervention[c]		
Control group	60	60.54 (9.72)	57.43 (7.55)	1.953 (59)	.05
Online feedback group	60	59.19 (10.04)	51.54 (6.67)	4.916 (59)	<.001

[a]TAS-20: Toronto Alexithymia Scale.

[b]t_{59}=0.748; *P*=.46.

[c]t_{59}=–4.529; *P*<.001.

Discussion

Impact of Comprehensive Assessment

The detailed baseline assessment provided a nuanced understanding of the patient's initial conditions and challenges. The high prevalence of hypoglycemia fear and misconceptions about diabetes management underscored the necessity for targeted educational interventions. By identifying these specific needs early, the intervention could be more precisely tailored to address the gaps in knowledge and self-management skills, thereby enhancing the overall effectiveness of the program.

Effectiveness of Educational Tools

The study found that videos were more effective than brochures in enhancing patient engagement and knowledge retention. Patients who used videos showed a greater reduction in glycemic indices and better comprehension of diabetes management principles. This suggests that video-based education may be a more powerful tool for patient education in diabetes management compared to traditional brochures.

Advancing Diabetes Management Through Information Technology: a Superior Alternative to Traditional Models

Diabetes management necessitates comprehensive care strategies and meticulous glycemic control. Patients increasingly seek personalized, home-based guidance from health care professionals, underscoring the pivotal role of home and community settings in rehabilitation efforts [14,15]. Our preliminary research revealed a significant demand for home-based care among patients with diabetes, paralleling the findings of Simsek et al [16] regarding chronic disease management in the older population. This "sinking" care model, designed to address both physical and psychological needs, faces challenges due to slow development, service uniformity, process inconsistencies, and personnel shortages, struggling to meet the escalating care demands [17].

Professor Sika of the Mayo School of Medicine emphasizes the current period as an optimal time to leverage "Internet+" technologies in the home health and care of older people, driven by both national policies and societal needs [18]. The era of big data heralds a paradigm shift in health management, characterized by comprehensive interventions, online consultations, risk identification, and chronic disease management, facilitated by full-cycle coverage, precision, wide reach, and rapid feedback. The integration of health monitoring systems and medical data platforms enables dynamic patient monitoring and personalized guidance, bridging the gap between health care providers and patients [19].

Community health service agencies serve as primary gateways for chronic disease management, with online platforms enhancing coordination, organization, and multidimensional monitoring. This model addresses the challenges of professionalism in remote health management and fosters resource sharing, interoperability, and effective integration of health technical personnel. For patients with newly diagnosed type 2 diabetes, digital platforms and information systems offer tailored health management guidance, marking a significant improvement in patient care.

Enhancing Glycemic Control in Newly Diagnosed Type 2 Diabetes Mellitus Through Online Feedback: an Application of King's Goal Attainment Theory

The study's findings reveal that using an online teach-back method, underpinned by King's Goal Attainment Theory, leads to significant improvements in glycemic control among patients with first-onset type 2 diabetes mellitus (T2DM). Notably, FBG, 2hPG, and HbA_{1c} levels exhibited a consistent decline across both study groups, with the online feedback group showing superior outcomes. This suggests that the online feedback approach not only enhances patients' glucose metabolism indices but also contributes to an improved quality of life and reduced risk of diabetes-related complications. Despite these positive trends, the control group did not demonstrate statistically significant changes in 2hPG and HbA_{1c} levels (P>.05), potentially attributable to the limited duration of observation and the inherently slow response of glycemic indicators.

The online feedback method, aligned with King's theory, leverages recent dietary records to offer personalized nutritional guidance, promoting healthier dietary choices, regular physical activity, and mood regulation. Such comprehensive support underscores the method's efficacy in aiding patients to meet glycemic targets. Supporting this, the study by Celik et al [20] on diabetes education among hospitalized patients with type 2 diabetes in Turkey found that repeated, multimodal diabetes education significantly improved self-care and glycemic control (P<.05), mirroring the outcomes observed in our investigation.

King's Goal Attainment Theory, central to our intervention, champions a patient-centered approach, emphasizing active patient engagement in health management through ongoing interactive communication with health care providers. This dynamic process fosters a collaborative effort toward achieving

XSL•FO
RenderX

rehabilitative goals, thereby enhancing patients' understanding of their condition and adherence to treatment regimens outside the hospital setting. Recognizing the varied needs and motivations across different stages of behavioral change, our approach advocates for tailored interventions to support the development of healthier behaviors effectively.

Given the characteristics of type 2 diabetes, including short-term hospitalizations, long-term outpatient care, and the need for effective community management, a tripartite model encompassing hospital, community, and family interventions is proposed. This model aims for comprehensive monitoring and follow-up, ensuring a cohesive support system for patients across various care settings.

Impact of Online Teach-Back Method on Psychological Control in Newly Diagnosed T2DM

Diabetic alexithymia and fear represent prevalent adverse psychological responses impacting patients with diabetes, often undermining self-management practices and overall quality of life [21]. Despite their significance, these emotional challenges are seldom recognized and addressed within clinical settings. This study uses a structured 6-step intervention mapping process—encompassing a logical model of the problem, program outcomes and objectives, design, production, execution plan, and evaluation strategy [22]. This methodology is grounded in educational and psychological frameworks, notably Bloom's Taxonomy of Educational Objectives and Bandura's Social Cognitive Theory, aiming to bolster psychological regulation among patients.

Our findings indicate that patients with T2DM exhibit moderate levels of alexithymia, aligning with the observations of Celik et al [23]. Postintervention assessments revealed significant improvements in the online feedback group, with notable reductions in the scores of the CHFS II -WS, CHFS II -BS, and TAS-20, compared to the control group ($P<.05$). These outcomes underscore the efficacy of the online teach-back approach, informed by King's Goal Attainment Theory, in enhancing coping mechanisms and life quality for patients with initial T2DM.

This study highlights the pivotal role of King's theory in addressing the dual challenges of limited medical resources and patients' diminished self-awareness, which often impede effective diabetes self-management. By leveraging this theoretical framework, our intervention promotes the dissemination of essential disease information and health education resources, fostering a patient-centered, interactive communication model. The observed significant amelioration in alexithymia and hypoglycemia fear levels substantiates the model's potential in elevating psychological well-being and control among patients with diabetes, thereby affirming the transformative impact of integrating high-quality, theory-based online feedback mechanisms in diabetes care.

Limitations

First, the constrained sample size poses challenges to the analytical precision, underscoring the necessity for larger cohorts to validate findings and examine nuanced changes in T2DM-specific health indicators. Second, there exists an imperative for future research to establish localized, comprehensive health management models that encompass a broader spectrum of chronic conditions, thereby facilitating a more holistic approach to patient care. Finally, the potential influence of extraneous social factors, including occupational status and comorbid conditions, on the study outcomes cannot be overlooked. Future investigations should aim to meticulously refine inclusion and exclusion criteria to mitigate such variables' impact.

Conclusions

The application of the online teach-back method, grounded in King's Goal Attainment Theory, emphasizes dynamic online communication. This approach transitions from leveraging the potential of intervention strategies to fostering patients' subjective initiative and internal motivation for behavioral change. Such a methodological shift holds substantial promise for enhancing health outcomes in patients with newly diagnosed T2DM.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards for Reporting Trials) checklist.
[PDF File (Adobe PDF File), 66 KB - jmir_v26i1e59142_app1.pdf]

References

1. Kim HJ, Kim K. Blood pressure target in type 2 diabetes mellitus. Diabetes Metab J 2022;46(5):667-674 [FREE Full text] [doi: 10.4093/dmj.2022.0215] [Medline: 36193727]
2. Bonora E, Cataudella S, Marchesini G, Miccoli R, Vaccaro O, Fadini GP, under the mandate of the Italian Diabetes Society. Incidence of diabetes mellitus in Italy in year 2018. A nationwide population-based study of the ARNO diabetes observatory. Nutr Metab Cardiovasc Dis 2021;31(8):2338-2344. [doi: 10.1016/j.numecd.2021.04.017] [Medline: 34074587]
3. Verma AK, Goyal Y, Bhatt D, Dev K, Alsahli MA, Rahmani AH, et al. A compendium of perspectives on diabetes: a challenge for sustainable health in the modern era. Diabetes Metab Syndr Obes 2021;14:2775-2787 [FREE Full text] [doi: 10.2147/DMSO.S304751] [Medline: 34168477]

XSL•FO
RenderX

4. Harris-Hayes M, Schootman M, Schootman JC, Hastings MK. The role of physical therapists in fighting the type 2 diabetes epidemic. J Orthop Sports Phys Ther 2020;50(1):5-16 [FREE Full text] [doi: 10.2519/jospt.2020.9154] [Medline: 31775555]
5. Heflin C, Hodges L, Ojinnaka CO, Arteaga I. Hypertension, diabetes and medication adherence among the older supplemental nutritional assistance program population. J Appl Gerontol 2022;41(3):780-787 [FREE Full text] [doi: 10.1177/07334648211022493] [Medline: 34137290]
6. Wang Z, Wu Y, Wu J, Wang M, Wang X, Wang J, et al. Trends in prevalence and incidence of type 2 diabetes among adults in Beijing, China, from 2008 to 2017. Diabet Med 2021;38(9):e14487. [doi: 10.1111/dme.14487] [Medline: 33278034]
7. Samson SL, Vellanki P, Blonde L, Christofides EA, Galindo RJ, Hirsch IB, et al. American Association of Clinical Endocrinology consensus statement: comprehensive type 2 diabetes management algorithm - 2023 update. Endocr Pract 2023;29(5):305-340. [doi: 10.1016/j.eprac.2023.02.001] [Medline: 37150579]
8. Tu Q, Lin S, Hyun K, Hafiz N, Manandi D, Koh AS, et al. The effects of multidisciplinary collaborative care on cardiovascular risk factors among patients with diabetes in primary care settings: a systematic review and meta-analysis. Prim Care Diabetes 2024;18(4):381-392 [FREE Full text] [doi: 10.1016/j.pcd.2024.05.003] [Medline: 38852029]
9. Cebrián-Cuenca AM, Moreno-Pérez O, Campuzano-Ruiz R, Soler MJ, García de Lucas MD, Orozco-Beltrán D. Multidisciplinary panel consensus for the management of patients with type 2 diabetes: a Delphi study. Arch Med Res 2024;55(1):102923. [doi: 10.1016/j.arcmed.2023.102923] [Medline: 38141271]
10. Pontoppidan JRN, Nielsen EE, Olsen MH, Skjødt MK, Christensen JO, Raymond IE, et al. A multidisciplinary, shared care clinic using personalized medicine and coordinated care in patients with concomitant type 2 diabetes and cardiovascular disease. Protocol and baseline characteristics. Prev Med Rep 2024;38:102594 [FREE Full text] [doi: 10.1016/j.pmedr.2024.102594] [Medline: 38283962]
11. Fylan B, Munro J, O'Hara JK, Khatoon B, Lawton R. Developing a research community within an online healthcare feedback platform. Health Expect 2023;26(2):705-714 [FREE Full text] [doi: 10.1111/hex.13696] [Medline: 36661042]
12. Chun M. Revision of the Chinese Version of the Fear of Hypoglycemia Questionnaire and Its Application in Patients With Type 2 Diabetes: Tianjin. China: Tianjin Medical University; 2015.
13. Jinyao Y, Shuqiao Y, Xiongzhao Z. The Chinese version of the TAS-20: reliability and validity. Chin Ment Health J 2003;17(11):763-767.
14. Faridani L, Abazari P, Heidarpour M, Melali H, Akbari M. The effect of home care on readmission and mortality rate in patients with diabetes who underwent general surgeries. J Educ Health Promot 2021;10:418 [FREE Full text] [doi: 10.4103/jehp.jehp_81_21] [Medline: 35071624]
15. Smith JM, Jarrín OF, Lin H, Tsui J, Dharamdasani T, Thomas-Hawkins C. Racial disparities in post-acute home health care referral and utilization among older adults with diabetes. Int J Environ Res Public Health 2021;18(6):3196 [FREE Full text] [doi: 10.3390/ijerph18063196] [Medline: 33808769]
16. Simsek H, Erkoyun E, Akoz A, Ergor A, Ucku R. Unmet health and social care needs and associated factors among older people aged ≥ 80 years in Izmir, Turkey. East Mediterr Health J 2021;27(8):772-781 [FREE Full text] [doi: 10.26719/emhj.21.009] [Medline: 34486713]
17. Wang Q, Liu J, Zheng Y. Evolutionary game analysis of community elderly care service regulation in the context of "Internet +". Front Public Health 2022;10:1093451 [FREE Full text] [doi: 10.3389/fpubh.2022.1093451] [Medline: 36620239]
18. Sica G, Harker-Murray P, Montori VM, Smith SA, EVIDENS Groups. Adherence of published diabetes mellitus practice guidelines to methodologic standards of guideline development. Endocrinol Metab Clin North Am 2002;31(3):819-828. [doi: 10.1016/s0889-8529(02)00016-6] [Medline: 12227133]
19. Eke R, Jannat KT, Yang XT, Parton JM. Disparity and factors associated with internet health information seeking among US adults living with diabetes mellitus: cross-sectional study. J Med Internet Res 2022;24(5):e32723 [FREE Full text] [doi: 10.2196/32723] [Medline: 35635741]
20. Celik S, Olgun N, Yilmaz FT, Anataca G, Ozsoy I, Ciftci N, et al. Assessment the effect of diabetes education on self-care behaviors and glycemic control in the Turkey nursing diabetes education evaluating project (TURNUDEP): a multi-center study. BMC Nurs 2022;21(1):215 [FREE Full text] [doi: 10.1186/s12912-022-01001-1] [Medline: 35932036]
21. Polhuis KCMM, van Bennekom E, Bot M, Nefs G, Vaandrager L, Habibovic M, et al. Flourishing mental health and lifestyle behaviours in adults with type 1 and type 2 diabetes mellitus: results from the diabetes MILES - the Netherlands study. J Psychosom Res 2022;160:110950 [FREE Full text] [doi: 10.1016/j.jpsychores.2022.110950] [Medline: 35691119]
22. Fernandez ME, Ten Hoor GA, van Lieshout S, Rodriguez SA, Beidas RS, Parcel G, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158 [FREE Full text] [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
23. Celik S, Taskin Yilmaz F, Yurtsever Celik S, Anataca G, Bulbul E. Alexithymia in diabetes patients: its relationship with perceived social support and glycaemic control. J Clin Nurs 2022;31(17-18):2612-2620. [doi: 10.1111/jocn.16088] [Medline: 34658093]

Abbreviations

2hPG: 2-hour postprandial glucose

XSL•FO
RenderX

CHFS-II: Chinese Version of the Hypoglycemia Fear Survey II
CHFS-II-BS: Chinese Hypoglycemia Fear-Behavior Scale
CHFS-II-WS: Chinese Hypoglycemia Fear-Worry Scale
CONSORT: Consolidated Standards for Reporting Trials
FBG: fasting blood glucose
HbA1c: hemoglobin A1c
MMSE: Mini-Mental State Scale
T2DM: type 2 diabetes mellitus
TAS-20: Toronto Alexithymia Scale

Edited by A Schwartz; submitted 03.04.24; peer-reviewed by W Zhao, AN Ali; comments to author 27.07.24; revised version received 29.07.24; accepted 30.08.24; published 31.10.24.

Please cite as:
Yan M, Yu Y, Li S, Zhang P, Yu J
Effectiveness of King's Theory of Goal Attainment in Blood Glucose Management for Newly Diagnosed Patients With Type 2 Diabetes: Randomized Controlled Trial
J Med Internet Res 2024;26:e59142
URL: https://www.jmir.org/2024/1/e59142
doi:10.2196/59142
PMID:

©Man Yan, Yingchun Yu, Shuping Li, Peiling Zhang, Jiaxiang Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness and Feasibility of Digital Pulmonary Rehabilitation in Patients Undergoing Lung Cancer Surgery: Systematic Review and Meta-Analysis

Taiping Lu[1,2*], BS; Ting Deng[1,2*], BS; Yangyang Long[2], BS; Jin Li[3], BS; Anmei Hu[3], BS; Yufan Hu[4], MSc; Li Ouyang[3], BS; Huiping Wang[1], PhD; Junliang Ma[3], MD; Shaolin Chen[1,2,5], MSc; Jiale Hu[6], PhD

[1]Nursing Department, Affiliated Hospital of Zunyi Medical University, Zunyi, China
[2]School of Nursing, Zunyi Medical University, Zunyi, China
[3]Thoracic Surgery Department, Affiliated Hospital of Zunyi Medical University, Zunyi, China
[4]Department of Oncology, Shenzhen People's Hospital, Shenzhen, China
[5]School of Nursing, Philippine Women's University, Manila, Philippines
[6]Department of Nurse Anesthesia, College of Health Professions, Virginia Commonwealth University, Virginia, American Samoa
[*]these authors contributed equally

Corresponding Author:
Shaolin Chen, MSc
Nursing Department
Affiliated Hospital of Zunyi Medical University
Number 149, Dalian Road, Huichuan District
Zunyi, 563000
China
Phone: 86 13762910513
Email: 30363284@qq.com

Abstract

Background: Pulmonary rehabilitation (PR) has been shown to effectively support postsurgical recovery in patients with lung cancer (LC) at various stages. While digital PR programs offer a potential solution to traditional challenges, such as time and space constraints, their efficacy and feasibility for patients undergoing LC surgery remain unclear.

Objective: This systematic review aims to assess the feasibility and effectiveness of digital PR programs for individuals undergoing LC surgery.

Methods: A systematic review was conducted, retrieving data from 6 English and 4 Chinese databases from their inception to January 1, 2024. References in related studies were also manually reviewed. The primary outcomes assessed were physical capacity, lung function, and the incidence of postoperative pulmonary complications (PPCs). The secondary outcomes were compliance, hospital stay, chest tube duration, anxiety, depression, and quality of life. Where applicable, recruitment and withdrawal rates were also evaluated. Meta-analysis and descriptive analysis were used to assess the outcomes.

Results: A total of 5 randomized controlled trials and 6 quasi-experimental studies (n=1063) were included, with 4 studies being included in the meta-analyses. Our meta-analyses revealed that digital PR reduced the decline in 6-minute walk distance (6-MWD) by an average of 15 m compared with routine PR programs from admission to discharge, demonstrating a clinically significant improvement in physical capacity (mean difference –15.00, 95% CI –25.65 to –4.34, P=.006). Additionally, digital PR was associated with a reduction (26/58, 45%) in the likelihood of PPCs (risk ratio 0.45, 95% CI 0.30-0.66, P<.001) and a reduction of 1.53 days in chest tube duration (mean difference –1.53, 95% CI –2.95 to –0.12, P=.03), without a statistically significant effect on postoperative hospital stay (mean difference –1.42, 95% CI –3.45 to 0.62, P=.17). Descriptive analyses suggested that digital PR has the potential to improve knowledge, lung function, quality of life, and self-efficacy, while reducing depression and anxiety. Notably, digital PR was found to be a safe, feasible, and acceptable supplementary intervention. Despite challenges with low recruitment, digital PR enhanced exercise compliance, increased patient satisfaction, and lowered dropout rates.

Conclusions: This systematic review is the first comprehensive analysis to suggest that digital PR is a safe, feasible, acceptable, and effective intervention for promoting recovery in patients with LC after surgery. Digital PR has the potential to be a valuable

XSL•FO
RenderX

supplement, expanding access to traditional PR programs. Future research should prioritize the development of interactive and inclusive digital solutions tailored to diverse age groups and educational backgrounds. Rigorous studies, including large-scale, high-quality randomized controlled trials with detailed protocols and robust methodologies, are needed to assess the short-, medium-, and long-term efficacy of digital PR, ensuring reproducibility in future research.

Trial Registration: PROSPERO CRD42023430271; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=430271

(J Med Internet Res 2024;26:e56795) doi:10.2196/56795

KEYWORDS

app-based; digital rehabilitation; internet-based intervention; lung cancer; perioperative pulmonary rehabilitation; systematic review; telerehabilitation

Introduction

Background

Lung cancer (LC) is the most prevalent cancer worldwide and the leading cause of cancer-related mortality [1]. The importance of perioperative pulmonary rehabilitation (PR) for patients undergoing LC surgery cannot be overstated. Evidence demonstrates its effectiveness in significantly reducing the incidence of postoperative pulmonary complications (PPCs) [2]; improving exercise capacity, pulmonary function, and overall quality of life after lung resection [3]; and enhancing self-management [4]. However, the current implementation of perioperative PR remains suboptimal [5,6]. A key challenge related to traditional PR programs is their reliance on face-to-face education. This approach is hindered by a shortage of well-trained health care professionals, limited time and resources for education, insufficient support services for patients with LC [7,8], and difficulties in providing adequate supervision [9,10]. Notably, patients in rural and remote areas have limited access to rehabilitation teams after discharge [10], highlighting the critical need for innovative approaches to enhance the effectiveness and accessibility of PR for patients undergoing LC surgery.

The rapid popularization of the internet and mobile apps has made tele–health care easily accessible, providing patients with convenient, flexible, and extended access to evidence-based interventions for self-management, surveillance, and supportive care, especially during the pandemic [6]. Although the terms “telehealth” and “digital health” share common features, telehealth often refers more narrowly to synchronous interactions between health care providers and patients, while digital health encompasses a broader range of digital technologies in health care [11]. In this review, the term “digital PR,” also referred to as “internet-based PR,” is defined as the utilization of digital technologies or devices in PR. This includes the Internet of Things, computing platforms, connectivity, software or apps, remote monitoring, wearable devices, virtual reality, augmented reality, and sensors for health care purposes, aligning with the World Health Organization’s definition [12]. Health care providers communicate with patients through various methods, including videoconferencing, video communication systems, WeChat, and Facebook, extending beyond traditional telephone and SMS text message interactions [13].

Compared with traditional PR, digital PR offers significant advantages, including secure remote storage and transmission of tailored PR educational materials, real-time monitoring of vital parameters and health care data, guidance for patient PR practices, and improved adherence to PR programs, all unhampered by temporal or geographical constraints. This approach reduces time and costs while enhancing patients’ access to PR programs, ensuring convenience, flexibility, and extended reach [14]. Researchers have explored the efficacy of various internet-based PR programs in patients with preoperative or postoperative LC [15,16]. However, digital rehabilitation also has some disadvantages, including technical issues [17], quality assurance concerns [18], and challenges related to impersonality [19]. Furthermore, there is ongoing debate regarding the effectiveness of telerehabilitation in treating other chronic diseases [6,10]. Notably, there are no published systematic reviews that focus on the feasibility and benefits of digital PR for patients with LC before or after surgery, nor are there any meta-analysis data reported. To our knowledge, no studies have been published to address this gap.

Objectives

This systematic review aims to explore the feasibility and effectiveness of digital PR. Additionally, we sought to determine whether it could serve as an effective alternative to replace or supplement traditional PR. Successful findings could provide valuable insights for future clinical research and practice, assist governments and policy makers in resource allocation, and promote the development of digital PR programs to improve the overall quality of health care services.

Methods

Design

This systematic review was registered on the PROSPERO platform (CRD42023430271), and there were no deviations from the registered protocol. The review procedures were independently conducted by 2 authors (TPL and TD), with any disagreements resolved through consensus or consultation with the third and fourth senior authors (SLC and JH).

Ethics Considerations

As our study was a systematic evaluation, no ethical review was conducted.

Eligibility Criteria

The inclusion criteria were as follows: (1) patients diagnosed with LC undergoing surgery across various phases (*population*); (2) at least one of the following digital technologies was used

XSL•FO
RenderX

in PR: Internet of Things, computing platforms, connectivity, software or apps, remote monitoring, online video dissemination, wearable devices, virtual reality, augmented reality, and sensors for health care purposes (*interventions*); (3) the control group received conventional PR through face-to-face interactions, paper documents (such as handouts and brochures), or telephone follow-up. Quasi-experimental studies without comparisons were also considered (*comparisons*); (4) primary outcomes included physical capacity measured by the 6-minute walk distance (6-MWD) test, lung function, and the incidence of PPCs [20]. Secondary outcomes included compliance, length of hospital stay, duration of chest tube indwelling, levels of anxiety or depression, and quality of life (*outcomes*); and (5) randomized controlled trials (RCTs) and quasi-experimental studies (*study designs*).

Exclusion criteria were as follows: (1) interventions in the intervention group were delivered without the internet, solely through face-to-face interactions or telephone calls (*interventions*); (2) incomplete data or unclear outcome effects (*data*); (3) duplicate or unavailable full texts (*publications*); and (4) publications in languages other than English and Chinese (languages).

Data Sources and Search Strategy

The search strategy was developed with the assistance of a library specialist. Databases searched included PubMed, Cochrane Library, Embase, Web of Science, MEDLINE, and CINAHL, along with Chinese databases such as CNKI, CBM, Wan Fang Database, and the China Science and Technology Journal Database (VIP Database), covering the period from inception to January 1, 2024. Detailed search strategies for each database and full search strings are provided in Multimedia Appendix 1. Additionally, references from the included studies and other relevant reviews were manually retrieved.

Study Selection and Data Extraction

The EndNote software (Clarivate Analytics) was used for literature management and screening. After removing duplicate articles, the titles and abstracts of the remaining articles were reviewed. Subsequently, the full texts of potentially eligible studies were carefully examined based on the eligibility criteria to determine inclusion. Data extraction followed the Template for Intervention Description and Replication (TIDieR) checklist [21] and the TIDieR-telehealth framework checklist [22]. The extraction included the following elements: author, publication year and region, study design, setting, eligibility criteria, participant characteristics, sample size, intervention details for each group, follow-up information, detailed outcome measures and time points, and outcome data.

Risk of Bias Assessment

The risk of bias in RCTs was evaluated using the revised Cochrane Risk of Bias Tool version 2.0 [23], which comprises 5 domains with responses categorized as “yes,” “probably yes,” “probably no,” “no,” or “no information.” Overall risk was classified as “low risk of bias,” “some concerns,” or “high risk of bias” [23]. For quasi-experimental studies, the Joanna Briggs Institute’s appraisal tool [24] was utilized, consisting of nine 9 questions scored as “yes” (1 point), “no” (0 points), “unclear” (0 points), or “not applicable” (0 points). Scores of less than 5 out of 9 indicated low methodological quality [24].

Statistical Analysis

All statistical analyses were conducted using Review Manager 5.4 software (RevMan, The Cochrane Collaboration). When appropriate, mean differences and 95% CIs were reported for continuous outcomes (eg, 6-MWD), while risk ratios and 95% CIs were provided for dichotomous outcomes (eg, the incidence of PPCs). A meta-analysis was conducted only when the outcome variables being assessed were the same, the timing of outcome measurements was comparable, and the interventions were consistent across studies. If clinical heterogeneity was present in any of these conditions, a descriptive analysis was performed instead of a meta-analysis. Results were visualized using forest plots, and heterogeneity was assessed using I^2 statistics. The fixed-effects model was applied for data combination when I^2 was less than 50%; otherwise, the random-effects model was used for the meta-analysis [25]. Publication bias was not assessed due to the limited number of included studies.

Results

Selection Process

A total of 2180 articles were initially retrieved from electronic databases, with an additional 26 articles found through manual searches. After a thorough screening process, 11 trials [15,16,26-34] were included, comprising 5 RCTs [27,28,30-32] and 6 quasi-experimental studies [15,16,26,29,33,34] (2 single-arm before-after studies and 1 pilot study). Four studies [26,27,30,34] underwent meta-analysis, while descriptive analyses were conducted for the remaining 7 studies. The selection process is visually presented in the flowchart (Figure 1). Our systematic review rigorously follows the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) checklist [35], with details provided in Multimedia Appendix 2 (also see [35]).

XSL•FO
RenderX

Figure 1. PRISMA 2020 Flow Diagram of Study Selection.

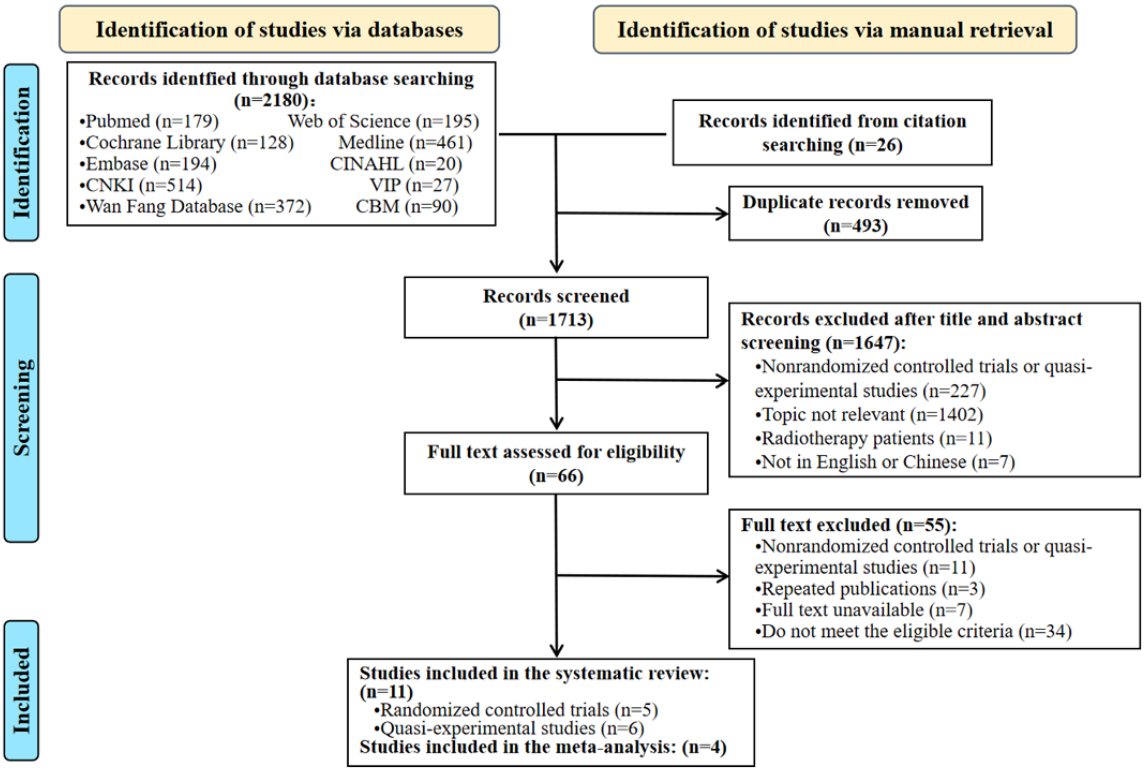

Characteristics of Included Studies

The detailed characteristics of the included studies are presented in Multimedia Appendix 3 (see also [15,16,26-34]): 11 studies, involving a total of 1063 patients with LC (control group: n=501; intervention group: n=562) and 22 family caregivers, were conducted across 4 countries: 6 in China, 2 in the United States, 2 in Korea, and 1 in Britain; 8 studies were published in English and 3 in Chinese. Of these, 10 studies focused solely on patients undergoing LC surgery, while 1 study included both patients undergoing LC surgery and their family caregivers. Of the 1085 participants, 595 (54.84%) were men and 425 (39.17%) were women. The gender was not reported by 65 participants. The average age of the participants ranged from 50.35 to 71.80 years, with sample sizes ranging from 18 to 104.

One study [16] implemented internet-based PR during the preoperative period, 7 studies [15,26-28,30,33,34] focused on the perioperative phases, and 3 studies [29,31,32] were conducted after surgery. The intervention strategies for both groups exhibited considerable heterogeneity among the studies. Seven studies [15,26,27,30,31,33,34] compared digital PR with face-to-face programs, 3 studies [27,30,33] utilized online video or multimedia education, and 2 studies [28,32] implemented PR using an app in both groups. Eight studies [15,16,26,28,29,31,32,34] focused on PR utilizing mobile apps, with 7 of these [15,16,26,28,29,32,34] also incorporating wearable devices. The apps included the WeChat app [26,31], Efil Breath app [32], Fit4Surgery app [15], Smart After-Care app [29], wrist-worn Garmin vívoactive fitness device app [16], and a wearable pedometer app [34]. Additionally, the wearable devices featured a pulse oximeter [15,29,32], muscle oxygen detector [28], heart rate monitor [16,28], digital sphygmomanometer [29], digital spirometer [29], and pedometer device [34]. Eight studies [15,26-31,33] integrated face-to-face PR with digital tools.

There was significant variation among the studies regarding the modality, intensity, duration, frequency, and supervision of digital PR programs. Intervention lengths ranged from during hospitalization to 18 months, with exercise sessions lasting from 3 minutes to 1 hour, occurring 2-5 days per week, and follow-up durations extending up to 48 months. The key intervention strategies for each included study are outlined in Table 1. Additionally, none of the included studies reported adherence to the TIDieR or TIDieR-telehealth checklist. The details of the intervention protocols for both groups and each study are comprehensively presented in Multimedia Appendix 4 (also see [15,16,26-34]), following the TIDieR guidelines.

XSL•FO
RenderX

Table 1. Key intervention strategies of the included studies.

Study	Interventions for the control group	Interventions for the intervention group	Digital types
Ji et al [32]	• Fixed pulmonary rehabilitation with an app; and aerobic exercise (walking) + resistance exercises • Providers: lung cancer specialists and nurses • Frequency: every day • Duration: No • Intensity: moderate continuous • Apps for a personalized mobile pulmonary rehabilitation platform (exercises, testing, and monitoring but no recording of breathing difficulty to adapt exercise level) • Initiate time: NAa • Length: 12 weeks • Supervision: No • Followed up: 12 weeks	• Completed pulmonary rehabilitation with the app for the first 6 weeks; used an interactive app for the remaining 6 weeks; additionally, aerobic exercise (walking) + resistance exercises • Providers: lung cancer specialists and nurses • Frequency: every day • Duration: No • Intensity: moderate continuous • Apps for a personalized mobile pulmonary rehabilitation platform (exercises, testing, monitoring, and recording the degree of breathing difficulty to adapt exercise level) • Initiate time: NA • Length: 12 weeks • Supervision: No • Followed up: 12 weeks	• A personalized mobile health–based pulmonary rehabilitation platform: app and patient monitoring website; and a personalized mobile health–based pulmonary rehabilitation platform: Efil • Breath app and patient monitoring website; and wearable pulse oximeter.
Sui et al [31]	• Usual care + simple education + rehabilitation guidance through a simple session by face-to-face education manual • Providers: physicians and nurses • Frequency: NA • Duration: NA • Intensity: light • Initiate time: after surgery • Length: 12 months • Supervision: No • Followed up: 60 months by telephone or clinic visit	• Postoperative pulmonary rehabilitation: health education (weeks 1-12, once/week); and rehabilitation exercise guidance: walking + aerobic exercise (weeks 13-52, once/week) and psychological support (12 months, once/2 weeks) • Providers: trained nurses • Frequency: every day • Duration: NA • Intensity: moderate • Initiate time: after surgery • Length: 12 months • Supervision: daily walking supervision by WeChat once a week for 12 months • Followed up: 48 months (total 60 months) by WeChat app	• WeChat app (eg, video course, nurse supervision)
Chu et al [27]	• Routine perioperative care • Providers: NA • Frequency: NA • Duration: NA • Intensity: NA • Initiate time: after surgery • Length: No • Supervision: No • Followed up: No	• Perioperative pulmonary rehabilitation: routine care + scanning QR code on a smartphone to watch breathing exercise video + interactive video call education • Providers: rehabilitation manager and respiratory specialist • Frequency: once daily from 5 to 8 days before surgery and 3 times daily within 10 days after discharge • Duration: 20 minutes each time • Intensity: light • Initiate time: 5-8 days before surgery • Length: 1 month • Supervision: No • Followed up: No	• Online video + video call
Li et al [30]	• Traditional face-to-face pulmonary rehabilitation: instructions and demonstrations of pulmonary rehabilitation techniques, including diaphragmatic breathing exercises, pursed lips breathing exercises, balloon blowing, and other breathing exercise techniques • Providers: research assistants • Frequency: NA • Duration: NA • Intensity: NA • Initiate time: at admission • Length: during hospitalization • Supervision: No • Followed up: NA	• Watched an animated pulmonary rehabilitation video downloaded on an iPad; the education content was similar to the control group; breathing exercises; and patients performed breathing exercises independently • Providers: research assistants • Frequency: twice a day • Duration: 31 minutes • Intensity: light • Initiate time: at admission • Length: during hospitalization • Supervision: bedside teach-back twice a day • Followed up: NA	• Online video

XSL•FO
RenderX

Study	Interventions for the control group	Interventions for the intervention group	Digital types
Liu and Pan [28]	• Portable wearable device with a mobile app monitoring system—preoperative: abdominal and deep breathing training + coughing + balloon blowing (every 4 hours); postoperative: isometric muscle, lower limb flexion, and extension exercise (day 1); arm raising and bed cycling (day 2); getting out of bed and shoulder and all-around exercise (day 3); and all-around exercise (day 4, 3-5 times/day, 5-8 minutes/time) and home exercise (3 times/week, 30 minutes/time) • Providers: nursing staff • Frequency: every day • Duration: 5-30 minutes/session • Intensity: light to moderate • Initiate time: before surgery • Length: perioperative period • Supervision: wearable device • Followed up: No	• Perioperative pulmonary rehabilitation same as the control group + self-confidence cultivation (direct experience + alternative experience [1 time/week, l hour/time] + verbal support + emotional counseling) • Providers: nursing staff • Frequency: every day • Duration: 5-60 minutes/session • Intensity: light to moderate • Initiate time: before surgery • Length: perioperative period • Supervision: wearable device • Followed up: No	• Wearable devices with muscle oxygen detector and a heart rate monitoring app
Sun et al [33]	• Usual care • Providers: NA • Frequency: NA • Duration: NA • Intensity: NA • Initiate time: before surgery • Length: No • Supervision: No • Followed up: telephone follow-up 2-4 weeks after discharge	• Pre- and postoperative pulmonary rehabilitation: The intervention is based on The Chronic Care Self-Management Model; traditional (information and technical skills) + self-management education; the intervention included videos, manuals, and postdischarge phone calls at home, containing different media with different learning modalities • Providers: researcher • Frequency: every day • Duration: NA • Intensity: NA • Initiate time: 3-7 days before surgery • Length: 3-7 days before surgery and day 7 after discharge • Supervision: No • Followed up: telephone follow-up 2-4 weeks after discharge	• Multimedia care model (online video)
Kadiri et al [15]	• Local chronic obstructive pulmonary disease rehabilitation classes; and strength and aerobic exercise for the upper and lower body • Providers: the medical team • Frequency: twice a week • Duration: 90 min • Intensity: light to moderate • Initiate time: before surgery • Length: 6 weeks after surgery • Supervision: No • Followed up: 6 weeks after surgery	• The “Fit4 surgery” app is based on a home rehabilitation structured exercise program (with integrated patient and clinician biofeedback), including 10 exercises for the upper and lower body as well as aerobic and strength exercises (>3 minutes) • Providers: the medical team • Frequency: NA • Duration: at least 3 minutes each time • Intensity: moderate • Initiate time: before surgery • Length: 18 months • Supervision: the pulse oximeter and cloud-based server • Followed up: 18 months	• “Fit4 surgery” app with a Bluetooth-enabled pulse oximeter and SIM card
Finley et al [16]	• NA	• Any moderately intense, aerobic physical activity; and surgeon-delivered exercise prescription + an activity tracker • Providers: surgeon and the project coordinator • Frequency: 5 days each week • Duration: 30 minutes a day • Intensity: moderate intense • Initiate time: before surgery • Length: preoperative • Supervision: Garmin vívoactive heart rate device • Followed up: 16 weeks after surgery	• Wrist-worn Garmin vívoactive heart rate monitoring device with the app

XSL•FO
RenderX

Study	Interventions for the control group	Interventions for the intervention group	Digital types
Yang et al [29]	• NA	• Smart after-care app; recording of vital signs; complete a daily subjective symptom survey; utilize video clips for pulmonary rehabilitation; and patients receive personalized diet and nutrition information • Providers: rehabilitation specialists • Frequency: every day • Duration: NA • Intensity: NA • Initiate time: after surgery • Length: 12 weeks • Supervision: self-monitoring devices including digital sphygmomanometer, finger pulse oximeter, and digital spirometer • Followed up: weekly phone calls and every 6 weeks participants returned to the clinic for follow-up for a total of 3 months	• Smart after-care app with self-monitoring devices
Qin et al [26]	• Routine pulmonary rehabilitation health education: lectures and pulmonary rehabilitation training guideline videos played on television every day; breathing and limb exercises (2-3 times/day, 10-15 minutes/time); and advice strengthening exercises (30 minutes/time) • Providers: doctors and nurses • Frequency: every day • Duration: 10-30 minutes • Intensity: moderate to intense • Initiate time: at admission • Length: during hospitalization • Supervision: No • Followed up: No	• Pulmonary rehabilitation program based on the internet: an online pulmonary rehabilitation health education course (video + documents + picture); before admission: face-to-face WeChat group education follows a daily training schedule (average 1-2 weeks before admission); during hospitalization: face-to-face education on respiratory, limbs, and resistance exercises + pulmonary rehabilitation delivered via the app; and after discharge: follow-up using the app and official account and face-to-face education • Providers: doctors, head nurses, and charge nurses • Frequency: 2-3 times a day • Duration: 10-15 minutes • Intensity: moderate • Initiate time: before admission • Length: perioperative period • Supervision: WeChat group regular reminder + check in the training schedule every day • Followed up: one-to-one telephone or home follow-up (weeks 1, 4, and 8 after discharge)	• WeChat app + official account
Chen et al [34]	• Routine nursing: close monitoring of the condition; routine anti-infection, fluid infusion, and nutritional support; medication guidance; and guidance in getting out of bed • Providers: NA • Frequency: NA • Duration: NA • Intensity: NA • Initiate time: NA • Length: NA • Supervision: No • Followed up: No	• Pulmonary rehabilitation based on a wearable device pedometer, bed limb activities + respiratory exercise function (5 minutes/time, 3-5 times/day), walking exercises (twice a day), and inspiratory muscle strength training • Providers: NA • Frequency: every day • Duration: 5-15 minutes of each exercise • Intensity: light to moderate • Initiate time: NA • Length: NA • Supervision: wearable device pedometer • Followed up: No	• Wearable device pedometer

[a]NA: not available.

Risk of Bias Assessment

All 5 RCTs were assessed to have an overall high risk of bias. Specifically, 1 study [32] was classified as high risk due to insufficient information on the random allocation process and concealment. Four studies [27,28,30,32] were deemed high risk because they lacked details on blinding and strategies to prevent contamination. Additionally, 1 study [32] showed a high risk of bias related to missing outcome data, while another study [31] exhibited a high risk of bias due to a lack of blinding in outcome measures. The risk of bias graph for each RCT is displayed in Figure 2 (also see [27,28,30-32]), and a summary of the risk of bias in RCTs is provided in Multimedia Appendix 5 (also see [15,16,26-34]).

XSL•FO
RenderX

Figure 2. Bias risk assessment of each study.

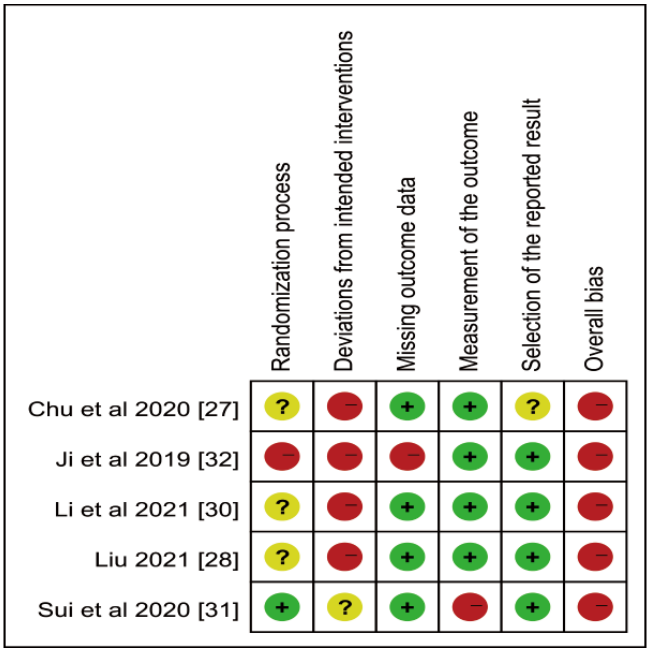

Among the 6 quasi-experimental studies, 1 [15] was assessed as having an overall high risk of bias, whereas the remaining 5 studies were at low risk. Details on the risk assessment of individual quasi-experimental studies are displayed in Multimedia Appendix 5.

Effects of the Digital PR Program

Physical Capacity

Although 6 studies [16,26,28,30,32,34] reported a 6-MWD at various time points, only 2 studies [26,30] focusing on the preoperative phase were suitable for meta-analysis. The analysis revealed that the change in 6-MWD from admission to discharge in the intervention group was statistically shorter than that in the control group (Figure 3; also see [26,30]). The mean difference was 15 m, with a 95% CI ranging from 4.34 to 25.65 m (P=.006). One study indicated no significant difference between the 2 PR apps in postsurgery patients during unclear phases [32]. Interestingly, 2 studies [26,30] demonstrated a decrease in 6-MWD at discharge compared with admission, whereas 5 studies [16,26,28,30,34] reported a significant improvement.

Figure 3. Effects of digital PR on the change in 6-WMD from admission to discharge in patients undergoing LC surgery.

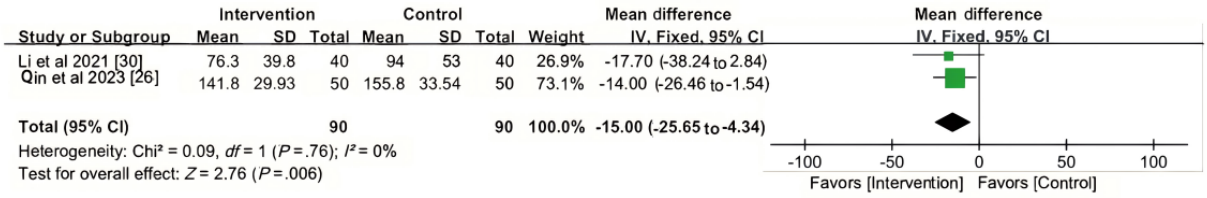

Mobile app–based PR demonstrated significant improvements in 2-MWD [29], lower limb muscle strength [29], modified Medical Research Council (mMRC) scores [32], and Chronic Obstructive Pulmonary Disease Assessment Test scores [34], although there was no improvement in upper limb muscle strength [29] after the intervention.

Lung Function

Only 2 studies [28,34] reported lung function indicators, each using diverse metrics and reassessment days that could not be combined. Therefore, a descriptive analysis was applied. The forced vital capacity (FVC), forced expiratory volume in 1 second (FEV_1), and the ratio of FEV_1 to the percent predicted for FVC (FEV_1/FVC) in the wearable devices group were statistically better than those in the routine group (P<.001) [34]. Additionally, PR that included an app and self-confidence cultivation significantly increased the peak expiratory flow rate compared with the app alone [28].

Postoperative Pulmonary Complications

Three studies [15,27,30] reported the incidence of PPCs. One study [15] recorded a PPC rate of 9.7% following the app-based intervention. A meta-analysis of 2 studies [27,30] found that patients with LC who received online video interventions had a 45% reduced risk (risk ratio 0.45, 95% CI 0.30-0.66, P<.001) of experiencing PPCs compared with those receiving traditional face-to-face interventions (Figure 4; also see [27,30]).

XSL•FO
RenderX

Figure 4. Effects of online video PR education on PPCs in patients undergoing LC surgery.

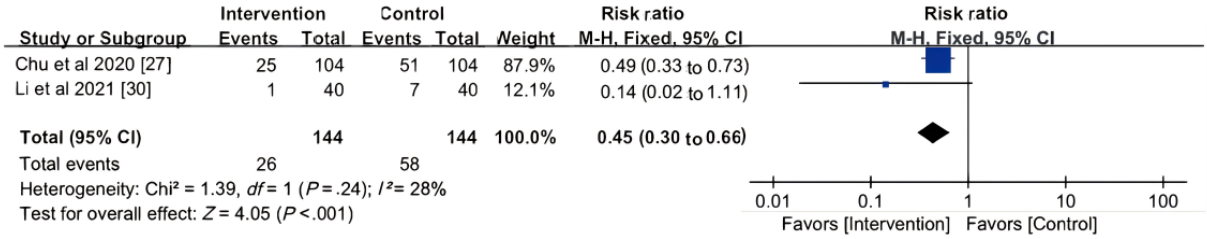

Duration of Chest Tube Placement and Postoperative Hospital Stay

Four studies [26,27,30,34] reported the duration of chest drainage tube indwelling, while 3 studies [26,27,34] documented the length of postoperative hospital stay. The meta-analysis revealed a statistically significant reduction in chest tube duration by 1.53 days (95% CI –0.12 to –2.95, *P*=.03) in the intervention group compared with the control group (Figure 5; also see [26,27,30,34]). However, no significant difference (95% CI –3.45 to 0.62, *P*=.17) was found between the groups regarding the length of postoperative hospital stay (Figure 6; also see [26,27,34]).

Figure 5. Effects of digital PR program on chest tube duration in patients undergoing LC surgery.

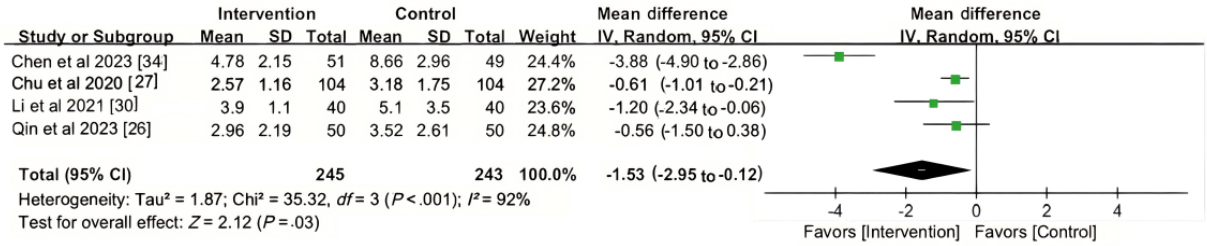

Figure 6. Effects of digital PR program on postoperative hospital stay in patients undergoing LC surgery.

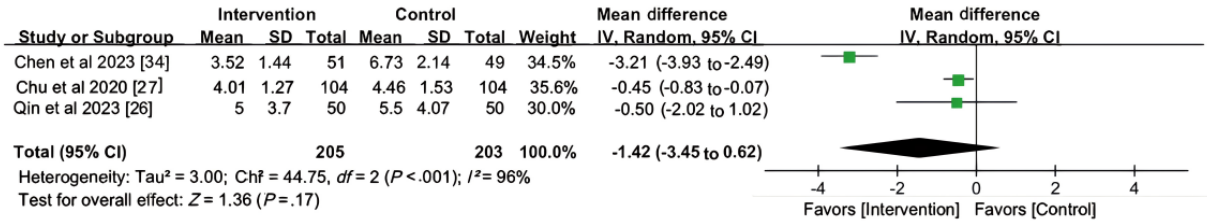

Quality of Life

Quality of life was assessed using various tools and time measurements in 6 studies [15,29,31-34]. Four studies [15,31,33,34] demonstrated that digital PR improved quality of life compared with the control group, including benefits for family caregivers [33]. Notably, app-based PR enhanced overall quality of life [32] and physical function, although it did not significantly improve symptoms [29] after a 12-week intervention. Furthermore, there was no significant difference (*P*=.99) between the 2 app groups [32].

Depression and Anxiety

Two studies [31,34] demonstrated that interventions using an app or wearable device pedometer significantly alleviated anxiety (*P*=.001) and depression (*P*=.01) compared with the control group.

Other Outcomes

Three studies [26,30,33] found that online video or app-based PR improved clients' knowledge compared with routine education. Additionally, 2 studies [28,33] showed a significant improvement in self-efficacy after the intervention (*P*=.001 and *P*=.1, respectively). Moreover, 3 studies [26,29,32] indicated that app-based PR led to higher patient satisfaction, particularly with interactive apps, compared with routine education.

Feasibility and Safety

Five studies [15,16,30,32,33] reported on the feasibility of utilizing the digital PR program. Two studies [32,33] indicated relatively low recruitment rates of 40.5% and 70%, citing difficulties in mastering the technique, as well as a lack of appropriate equipment and time. Additionally, 2 studies [16,30] reported compliance rates for the digital PR schedule at 79% and 48.3%, with higher respiratory exercise compliance observed in the animation group. Another study [33] reported high patient satisfaction along with favorable acceptability and usability ratings from both patients and family caregivers. Withdrawals were noted in all studies, with only 4 studies experiencing participant losses [15,29,31,33]. The withdrawal rate in the intervention group ranged from 10% to 32%, which was statistically lower than the control group's withdrawal rate, which varied from 20% to 79%. No adverse events were reported across any of the studies. Detailed information is presented in Multimedia Appendix 6 (see also [15,16,26-34]).

XSL•FO
RenderX

Discussion

Principal Findings

To our knowledge, this work is the first systematic review exploring digital PR for patients with LC before and after surgery. The initial meta-analyses indicated that digital PR improved the 6-MWD by an average of 15 m, with a 95% CI ranging from 4.34 to 26.65 m [26,30]. This change represents a clinically significant improvement in physical capacity for patients, meeting the threshold for minimal clinically important differences. The second meta-analysis indicated that digital PR was associated with a (26/58, 45%) lower likelihood of PPCs and suggested that digital PR reduced the risk of these complications by more than half (ie, a 55% reduction) [27,30]. The final analyses identified a statistically significant reduction in chest tube duration of 1.53 days [26,27,30,34], but found no effect on postoperative hospital stay for patients undergoing LC surgery [26,27,34].

Descriptive data analyses revealed that digital PR has the potential to enhance knowledge [26,30,33], physical capacity [16,26,28-30,32,34], lung function [28,34], quality of life [15,31,33,34], and self-efficacy [28,33], while also reducing symptoms of depression and anxiety [31,34]. Furthermore, digital PR is a safe, feasible, and acceptable supplementary intervention for patients undergoing LC surgery [15,16,30,32,33]. Despite challenges with low recruitment during the enrollment phase [15,29,31,33], digital PR has been shown to improve exercise compliance [30], enhance patient satisfaction [26,29,32], and reduce dropout rates.

Interpretation of the Findings

Our systematic review identified mobile apps as the primary intervention method, frequently combined with wearables such as pulse oximeters and heart rate monitors. However, emerging technologies such as virtual reality, augmented reality, exergame training, and intelligent robotic systems were not explored. These technologies have demonstrated potential in other rehabilitation contexts, such as improving exercise compliance for patients with chronic obstructive pulmonary disease through virtual reality–based PR [36] and reducing fall risk for older adults with exergame step training [37]. Aldebaran Robotics' Natural Another One humanoid has successfully guided older adult patients through rehabilitation exercises, evaluating their performance for near real-time processing [38]. Future research should focus on developing integrated intelligent software and devices specifically designed for PR programs tailored to patients undergoing LC surgery. These innovations could improve education, guidance, supervision, interactivity, and cost-effectiveness, thereby optimizing the rehabilitation experience. Additionally, they may have the potential to replace traditional PR guidance in the future.

Patients undergoing LC surgery frequently experience a decline in lung function, activity intolerance, and a reduction in quality of life, with recovery taking 6-12 months. Preoperative PR is crucial for recovery; however, our review found limited research focusing on this stage, revealing only a modest improvement in walking distance. Given the short time frame and the nature of home-based preoperative care, implementing effective PR can be challenging. Mobile PR presents a promising solution, warranting further investigation through RCTs.

None of the included studies reported adherence to the TIDieR or TIDieR-telehealth checklist. Furthermore, all studies lacked critical intervention details, including information on providers, frequency, duration, intensity, initiation time, length of intervention, supervision, and follow-up methods in both the intervention and control groups. Additionally, the studies assessed various outcome variables at different time points, leading to significant clinical heterogeneity. Consequently, only 4 outcomes could be synthesized in the meta-analysis, while the remaining outcomes could only be described narratively. This variability may impact the replicability and generalizability of digital PR programs.

A total of 7 perioperative studies [15,26-28,30,33,34] and 3 post-surgical studies [29,31,32] investigated digital PR for patients with LC; however, the intervention durations and follow-up periods varied significantly. Because of the diversity in outcome indicators and assessment time points, only the 6-MWD at discharge and the incidence of PPCs could be combined for analysis, and both demonstrated significant differences. The observed improvement in physical activity aligns with findings from digital PR studies for chronic obstructive pulmonary disease [39] and chronic musculoskeletal conditions [40,41]. However, other indicators exhibited high heterogeneity, which limited the ability to conduct a descriptive analysis. To strengthen the evidence, future research should evaluate the short-, medium-, and long-term effects of digital PR.

The effectiveness of digital health systems relies on their feasibility, usability, and acceptability [42]. However, most of the included studies targeted older adult participants who were already smartphone-savvy, which limited insights into program acceptability across a broader population. Only 4 studies reported on these aspects, and none addressed the experiences of providers. Despite low recruitment rates (40%-70%) [32,33], participants in digital programs demonstrated better exercise compliance [16,30] and lower dropout rates [15,29,31,33] compared with those receiving face-to-face education. However, research indicates that telemedicine interventions do not significantly enhance patients' adherence to exercise [39,43]. Low attendance in digital health programs among older adults is often attributed to unfamiliarity with technology [44], scarcity of devices, and time constraints [32,33]. However, once engaged, patients typically find it easy to access educational resources and supervision through apps, wearables, and the internet. Our research indicated that app-based PR programs were highly satisfying and user-friendly for both patients and caregivers. The combination of face-to-face and digital interventions in the 7 included studies leveraged the advantages of both approaches, thereby improving patient adherence. However, challenges persist in implementation, including feelings of being overwhelmed by devices, lack of internet access, or poor-quality video/audio [45]. Additionally, concerns about sensor accuracy [17] and perceptions of ineffectiveness [25,29] can hinder participation. Future programs should take providers' experiences into account and adapt to these obstacles

XSL•FO
RenderX

to enhance attendance and compliance with digital rehabilitation programs.

All 5 included RCTs were assessed to be at high risk of bias. Among the 6 quasi-experimental studies, 2 used single-arm designs, and 1 was also deemed to be at high risk of bias. This deficiency in high-quality studies may affect the overall level of evidence. Therefore, further well-designed RCTs with robust methodologies are essential to provide clearer insights into the effectiveness of digital PR in this context.

Strengths and Limitations

Our systematic review possesses several notable strengths. First, it represents the first comprehensive analysis that highlights the benefits of digital PR for the recovery of patients undergoing LC surgery, addressing both effectiveness and feasibility. Second, we provided a detailed description of each study based on the TIDieR-telehealth framework, underscoring the necessity for future studies to enhance replicability and generalizability. Lastly, our work incorporated evidence from both English and Chinese sources, reflecting a diverse range of cultural and social contexts.

There are several limitations to consider. First, the limited number of high-quality RCTs with relatively small sample sizes may compromise the robustness of the evidence. Second, while the primary focus was on patients undergoing LC surgery, the variability in intervention initiation, follow-up durations, and chemotherapy administration across studies may have contributed to high heterogeneity. To address heterogeneity, we adhered to the Cochrane Handbook recommendations [46] and consolidated evidence from studies with similar starting phases and interventions, ensuring the rigor and reliability of our results. For instance, we summarized data for the preoperative, perioperative, and postoperative phases, as well as different follow-up time points. Meta-analysis was conducted only when the studies involved similar patient populations, interventions, and measurement time points. In instances of significant heterogeneity, we opted for descriptive analysis. Despite these efforts, considerable heterogeneity among the studies persisted. Third, the lack of standardization in reporting interventions according to the TIDieR checklist led to significant variations in study design, methodology, and outcome measures. This heterogeneity—including differences in settings, participant characteristics, intervention providers, and digital program protocols (eg, the type of digital technology and delivery; initiation timing; exercise specifics such as location, duration, frequency, intensity, supervision, guidance, modifications; and strategies to enhance fidelity)—complicates interpretation and makes it challenging to draw robust conclusions. Finally, by limiting the search to Chinese and English literature, we may have excluded valuable research published in other languages, which could further restrict the generalizability of our findings.

Recommendations for Future Studies and Clinical Practice

First, future research should prioritize developing interactive and inclusive digital solutions that cater to diverse age groups and education levels. Second, it is essential to identify obstacles in both face-to-face and digital PR before developing any tools. Third, hybrid models that combine both methods should be explored to address challenges, enhance attendance, and ensure compliance with the PR program. Lastly, large, high-quality RCTs are needed, with a focus on clarifying blinding, allocation concealment, comparability, and specific protocol details, including duration, intensity, frequency, supervision, and adequate follow-up. Outcomes should be assessed at different time points to evaluate the short-, medium-, and long-term feasibility and effectiveness of various digital PR protocols. Key recommendations for health care practices and policies should be emphasized in future research, including the development of evidence-based digital PR guidelines to standardize the use of digital tools. Additionally, integrating digital programs into in-person rehabilitation can help address limitations of face-to-face interactions, while formulating telehealth policies will enhance patient access and options. Finally, future studies should adhere to the TIDieR [21] and TIDieR-telehealth frameworks [22], as well as relevant reporting guidelines such as CONSORT (Consolidated Standards of Reporting Trails) [47] and the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) for quasi-experimental studies [48], to ensure robust and replicable research.

Conclusions

This systematic review highlights the safety, feasibility, and efficacy of digital PR programs for patients undergoing LC surgery. Although challenges related to recruitment and attendance persist during enrollment, these programs have demonstrated benefits in enhancing physical activity, reducing the occurrence of PPCs, and shortening the duration of chest tube placement. Furthermore, digital PR programs offer potential improvements in exercise compliance, lung function, quality of life (excluding symptoms), self-efficacy, patient satisfaction, and mental health. While rigorous research is still necessary, our findings suggest that digital PR can serve as a valuable supplement to expand access to rehabilitation. These programs provide flexible, self-directed exercise options, improve continuity through extended supervision, and enhance recovery for patients undergoing LC surgery.

Acknowledgments

This study was supported by the Science and Technology Foundation of Guizhou Provincial Health Commission (grant numbers 2025GZWJKJXM1666 and gzwkj 2023-546), the Foundation of Science and Technology Department of Guizhou Province (grant numbers LC [2022] 008), and the National Undergraduate Innovative and Entrepreneurial Training Program of 2022 (grant number 202210661279).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 37 KB - jmir_v26i1e56795_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines checklist.
[DOCX File , 29 KB - jmir_v26i1e56795_app2.docx]

Multimedia Appendix 3
Basic characteristics of the included studies.
[DOCX File , 22 KB - jmir_v26i1e56795_app3.docx]

Multimedia Appendix 4
The TIDieR (Template for Intervention Description and Replication) checklist for the included studies.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v26i1e56795_app4.xlsx]

Multimedia Appendix 5
Methodological quality of each randomized controlled trial and quasi-experimental study.
[DOCX File , 87 KB - jmir_v26i1e56795_app5.docx]

Multimedia Appendix 6
Main outcome indicators and intervention effects of the included studies.
[DOCX File , 23 KB - jmir_v26i1e56795_app6.docx]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Mao X, Ni Y, Niu Y, Jiang L. The clinical value of pulmonary rehabilitation in reducing postoperative complications and mortality of lung cancer resection: a systematic review and meta-analysis. Front Surg 2021;8:685485 [FREE Full text] [doi: 10.3389/fsurg.2021.685485] [Medline: 34646857]
3. Tenconi S, Mainini C, Rapicetta C, Braglia L, Galeone C, Cavuto S, et al. Rehabilitation for lung cancer patients undergoing surgery: results of the PUREAIR randomized trial. Eur J Phys Rehabil Med 2021 Dec;57(6):1002-1011 [FREE Full text] [doi: 10.23736/S1973-9087.21.06789-7] [Medline: 34042410]
4. McCarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2015 Feb 23;2015(2):CD003793 [FREE Full text] [doi: 10.1002/14651858.CD003793.pub3] [Medline: 25705944]
5. Hui Y, Saisai L, Jingru C, Ruiyun C, Yaxin Q. A survey on the knowledge, attitude and practice of perioperative lung rehabilitation for lung cancer among nurses of Thoracic Surgery Department in 10 Class I hospitals in Henan Province. Chinese Journal of Modern Nursing 2023;29(26):3542-3548. [doi: 10.3760/cma.j.cn115682-20221203-05821]
6. Cox NS, Dal Corso S, Hansen H, McDonald CF, Hill CJ, Zanaboni P, et al. Telerehabilitation for chronic respiratory disease. Cochrane Database Syst Rev 2021 Jan 29;1(1):CD013040 [FREE Full text] [doi: 10.1002/14651858.CD013040.pub2] [Medline: 33511633]
7. Chan SC, Sekhon JK, Engkasan JP, Nathan JJ, Mirza FT, Liew SM, RESPIRE Collaborators. Barriers and challenges of implementing pulmonary rehabilitation in Malaysia: stakeholders' perspectives. J Glob Health 2021;11:02003 [FREE Full text] [doi: 10.7189/jogh.11.02003] [Medline: 34737852]
8. Gao LJ, Zhao HM. [A nationwide online questionnaire survey of the understanding and implementation of pulmonary rehabilitation at all levels of medical institutions in China in 2017]. Zhonghua Jie He He Hu Xi Za Zhi 2019 Apr 12;42(4):275-278. [doi: 10.3760/cma.j.issn.1001-0939.2019.04.006] [Medline: 30955285]
9. Lee JH, Lee HY, Jang Y, Lee JS, Oh Y, Lee S, et al. Efficacy of unsupervised home-based pulmonary rehabilitation for patients with chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis 2020;15:2297-2305 [FREE Full text] [doi: 10.2147/COPD.S268683] [Medline: 33061347]

XSL•FO
RenderX

10. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020 Jan 31;1(1):CD010255 [FREE Full text] [doi: 10.1002/14651858.CD010255.pub3] [Medline: 32002991]
11. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023 May;5(5):e316-e327 [FREE Full text] [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
12. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
13. Albahri A, Alwan JK, Taha ZK, Ismail SF, Hamid RA, Zaidan A, et al. IoT-based telemedicine for disease prevention and health promotion: state-of-the-art. Journal of Network and Computer Applications 2021 Jan 01;173:102873. [doi: 10.1016/j.jnca.2020.102873]
14. Fekete M, Fazekas-Pongor V, Balazs P, Tarantini S, Nemeth AN, Varga JT. Role of new digital technologies and telemedicine in pulmonary rehabilitation: smart devices in the treatment of chronic respiratory diseases. Wien Klin Wochenschr 2021 Nov;133(21-22):1201-1207 [FREE Full text] [doi: 10.1007/s00508-021-01930-y] [Medline: 34460006]
15. Kadiri SB, Kerr AP, Oswald NK, Budacan A, Flanagan S, Golby C, et al. Fit 4 surgery, a bespoke app with biofeedback delivers rehabilitation at home before and after elective lung resection. J Cardiothorac Surg 2019 Jul 05;14(1):132 [FREE Full text] [doi: 10.1186/s13019-019-0951-6] [Medline: 31277671]
16. Finley DJ, Stevens CJ, Emond JA, Batsis JA, Fay KA, Darabos C, et al. Potential effectiveness of a surgeon-delivered exercise prescription and an activity tracker on pre-operative exercise adherence and aerobic capacity of lung cancer patients. Surg Oncol 2021 Jun;37:101525 [FREE Full text] [doi: 10.1016/j.suronc.2021.101525] [Medline: 33813267]
17. Lee AYL, Wong AKC, Hung TTM, Yan J, Yang S. Nurse-led telehealth intervention for rehabilitation (telerehabilitation) among community-dwelling patients with chronic diseases: systematic review and meta-analysis. J Med Internet Res 2022 Nov 02;24(11):e40364 [FREE Full text] [doi: 10.2196/40364] [Medline: 36322107]
18. Barsom EZ, Feenstra TM, Bemelman WA, Bonjer JH, Schijven MP. Coping with COVID-19: scaling up virtual care to standard practice. Nat Med 2020 May;26(5):632-634. [doi: 10.1038/s41591-020-0845-0] [Medline: 32405054]
19. Ha DM, Nunnery MA, Klocko RP, Haverhals LM, Bekelman DB, New ML, et al. Lung cancer survivors' views on telerehabilitation following curative intent therapy: a formative qualitative study. BMJ Open 2023 Jun 23;13(6):e073251 [FREE Full text] [doi: 10.1136/bmjopen-2023-073251] [Medline: 37355268]
20. Fernandez FG, Falcoz PE, Kozower BD, Salati M, Wright CD, Brunelli A. The Society of Thoracic Surgeons and the European Society of Thoracic Surgeons general thoracic surgery databases: joint standardization of variable definitions and terminology. Ann Thorac Surg 2015 Jan;99(1):368-376. [doi: 10.1016/j.athoracsur.2014.05.104] [Medline: 25555970]
21. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. [Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide]. Gesundheitswesen 2016 Mar;78(3):175-188. [doi: 10.1055/s-0041-111066] [Medline: 26824401]
22. Rhon DI, Fritz JM, Kerns RD, McGeary DD, Coleman BC, Farrokhi S, et al. TIDieR-telehealth: precision in reporting of telehealth interventions used in clinical trials - unique considerations for the Template for the Intervention Description and Replication (TIDieR) checklist. BMC Med Res Methodol 2022 Jun 02;22(1):161 [FREE Full text] [doi: 10.1186/s12874-022-01640-7] [Medline: 35655144]
23. Higgins JPT, Savovi J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: Cochrane; 2019.
24. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis. Adelaide, SA, Australia: JBI; 2024.
25. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol 2011 Apr 07;11:41 [FREE Full text] [doi: 10.1186/1471-2288-11-41] [Medline: 21473747]
26. Mengxia Q, Yan Q, Rong X, Cuiping D, Chun C. The application of internet based whole health education of pulmonary rehabilitation in patients undergoing lung cancer surgery. Chinese Journal of Nursing 2023;58(15):1808-1814. [doi: 10.3761/j.issn.0254-1769.2023.15.003]
27. Yunqian C, Yijing L, Huijin G, Jing S, Kaihua L. Effect of perioperative video-mission mode breathing exercise on patients with thoracoscopic lung cancer surgery. Chinese Journal of Surgical Oncology 2020;12(03):207-210. [doi: 10.3969/j.issn.1674-4136.2020.03.006]
28. Qing L, Haifang P. Effects of remote supportive exercise combined with self-confidence development on stigma and fatigue in elderly patients undergoing radical lung cancer surgery. Chinese Journal of Gerontology 2021;41(16):3570-3573. [doi: 10.3969/j.issn. 1005-9202.2021.16.050]
29. Yang HC, Chung SH, Yoo JS, Park B, Kim MS, Lee JM. Evaluation of a smart after-care program for patients with lung cancer: a prospective, single-arm pilot study. J Chest Surg 2022 Apr 05;55(2):108-117 [FREE Full text] [doi: 10.5090/jcs.21.138] [Medline: 35135903]

XSL•FO
RenderX

30. Li J, Davies M, Ye M, Li Y, Huang L, Li L. Impact of an animation education program on promoting compliance with active respiratory rehabilitation in postsurgical lung cancer patients: a randomized clinical trial. Cancer Nurs 2021;44(2):106-115. [doi: 10.1097/NCC.0000000000000758] [Medline: 31714266]
31. Sui Y, Wang T, Wang X. The impact of WeChat app-based education and rehabilitation program on anxiety, depression, quality of life, loss of follow-up and survival in non-small cell lung cancer patients who underwent surgical resection. Eur J Oncol Nurs 2020 Apr;45:101707. [doi: 10.1016/j.ejon.2019.101707] [Medline: 32078926]
32. Ji W, Kwon H, Lee S, Kim S, Hong JS, Park YR, et al. Mobile health management platform-based pulmonary rehabilitation for patients with non-small cell lung cancer: prospective clinical trial. JMIR Mhealth Uhealth 2019 Jun 21;7(6):e12645 [FREE Full text] [doi: 10.2196/12645] [Medline: 31228180]
33. Sun V, Raz DJ, Ruel N, Chang W, Erhunmwunsee L, Reckamp K, et al. A multimedia self-management intervention to prepare cancer patients and family caregivers for lung surgery and postoperative recovery. Clin Lung Cancer 2017 May;18(3):e151-e159 [FREE Full text] [doi: 10.1016/j.cllc.2017.01.010] [Medline: 28233696]
34. Chen Z, Wang Y, Qiu P, Tan L, Hu Y. Pulmonary rehabilitation exercise based on wearable device pedometer improved lung cancer patients with impaired pulmonary function. Altern Ther Health Med 2024 Jan;30(1):78-82 [FREE Full text] [Medline: 37820671]
35. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
36. Jung T, Moorhouse N, Shi X, Amin MF. A virtual reality-supported intervention for pulmonary rehabilitation of patients with chronic obstructive pulmonary disease: mixed methods study. J Med Internet Res 2020 Jul 07;22(7):e14178 [FREE Full text] [doi: 10.2196/14178] [Medline: 32673224]
37. Sturnieks DL, Hicks C, Smith N, Ratanapongleka M, Menant J, Turner J, et al. Exergame and cognitive training for preventing falls in community-dwelling older people: a randomized controlled trial. Nat Med 2024 Jan;30(1):98-105. [doi: 10.1038/s41591-023-02739-0] [Medline: 38228913]
38. Simonov M, Delconte G. Humanoid assessing rehabilitative exercises. Methods Inf Med 2015;54(2):114-121. [doi: 10.3414/ME13-02-0054] [Medline: 24986076]
39. Koh JH, Chong LCY, Koh GCH, Tyagi S. Telemedical interventions for chronic obstructive pulmonary disease management: umbrella review. J Med Internet Res 2023 Feb 16;25:e33185 [FREE Full text] [doi: 10.2196/33185] [Medline: 36795479]
40. Zangger G, Bricca A, Liaghat B, Juhl CB, Mortensen SR, Andersen RM, et al. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023 Jul 06;25:e46439 [FREE Full text] [doi: 10.2196/46439] [Medline: 37410534]
41. Bargeri S, Castellini G, Vitale JA, Guida S, Banfi G, Gianola S, et al. Effectiveness of telemedicine for musculoskeletal disorders: umbrella review. J Med Internet Res 2024 Feb 02;26:e50090 [FREE Full text] [doi: 10.2196/50090] [Medline: 38306156]
42. Debelle H, Packer E, Beales E, Bailey HGB, Mc Ardle R, Brown P, et al. Feasibility and usability of a digital health technology system to monitor mobility and assess medication adherence in mild-to-moderate Parkinson's disease. Front Neurol 2023;14:1111260 [FREE Full text] [doi: 10.3389/fneur.2023.1111260] [Medline: 37006505]
43. Davergne T, Meidinger P, Dechartres A, Gossec L. The effectiveness of digital apps providing personalized exercise videos: systematic review with meta-analysis. J Med Internet Res 2023 Jul 13;25:e45207 [FREE Full text] [doi: 10.2196/45207] [Medline: 37440300]
44. Pang L, Liu Z, Lin S, Liu Z, Liu H, Mai Z, et al. The effects of telemedicine on the quality of life of patients with lung cancer: a systematic review and meta-analysis. Ther Adv Chronic Dis 2020;11:2040622320961597 [FREE Full text] [doi: 10.1177/2040622320961597] [Medline: 33101621]
45. Sinha Gregory N, Shukla AP, Noel JJ, Alonso LC, Moxley J, Crawford AJ, et al. The feasibility, acceptability, and usability of telehealth visits. Front Med (Lausanne) 2023;10:1198096 [FREE Full text] [doi: 10.3389/fmed.2023.1198096] [Medline: 37538312]
46. Deeks JJ, Higgins JPT, Altman DG. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: Cochrane; 2023.
47. Cuschieri S. The CONSORT statement. Saudi J Anaesth 2019 Apr;13(Suppl 1):S27-S30 [FREE Full text] [doi: 10.4103/sja.SJA_559_18] [Medline: 30930716]
48. Vlahov D. Transparent Reporting of Evaluations with Nonrandomized Designs (TREND). J Urban Health 2004 Jun;81(2):163-164 [FREE Full text] [doi: 10.1093/jurban/jth099] [Medline: 15136648]

Abbreviations

6-MWD: 6-minute walk distance
CONSORT: Consolidated Standards of Reporting Trails
FEV1: forced expiratory volume in 1 second

XSL•FO
RenderX

FVC: forced vital capacity
LC: lung cancer
mMRC: modified Medical Research Council
PPC: postoperative pulmonary complication
PR: pulmonary rehabilitation
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
RCT: randomized controlled trial
TIDieR: Template for Intervention Description and Replication
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs

Edited by N Cahill; submitted 26.01.24; peer-reviewed by S Lampridis, PC Chang, X Mao; comments to author 21.05.24; revised version received 07.06.24; accepted 25.09.24; published 11.11.24.

Please cite as:
Lu T, Deng T, Long Y, Li J, Hu A, Hu Y, Ouyang L, Wang H, Ma J, Chen S, Hu J
Effectiveness and Feasibility of Digital Pulmonary Rehabilitation in Patients Undergoing Lung Cancer Surgery: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e56795
URL: https://www.jmir.org/2024/1/e56795
doi:10.2196/56795
PMID:

©Taiping Lu, Ting Deng, Yangyang Long, Jin Li, Anmei Hu, Yufan Hu, Li Ouyang, Huiping Wang, Junliang Ma, Shaolin Chen, Jiale Hu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Offer of the Smoke Free Smartphone App Compared With No Intervention for Smoking Cessation: Pragmatic Randomized Controlled Trial

Sarah Jackson[1], PhD; Dimitra Kale[1], PhD; Emma Beard[1], PhD; Olga Perski[1], PhD; Robert West[1], PhD; Jamie Brown[1], PhD

Department of Behavioural Science and Health, University College London, London, United Kingdom

Corresponding Author:
Sarah Jackson, PhD
Department of Behavioural Science and Health
University College London
1-19 Torrington Place
London, WC1E 6BT
United Kingdom
Phone: 44 2076795634
Fax: 44 2076795634
Email: s.e.jackson@ucl.ac.uk

Abstract

Background: Digital technologies offer the potential for low-cost, scalable delivery of interventions to promote smoking cessation.

Objective: We aimed to evaluate the effectiveness of the offer of Smoke Free—an evidence-informed, widely used app—for smoking cessation versus no support.

Methods: In this 2-arm randomized controlled trial, 3143 motivated adult smokers were recruited online between August 2020 and April 2021 and randomized to receive an offer of the Smoke Free app plus follow-up (intervention arm) versus follow-up only (comparator arm). Both groups were shown a brief message at the end of the baseline questionnaire encouraging them to make a quit attempt. The primary outcome was self-reported 6-month continuous abstinence assessed 7 months after randomization. Secondary outcomes included quit attempts in the first month post randomization, 3-month continuous abstinence assessed at 4 months, and 6-month continuous abstinence at 7 months among those who made a quit attempt. The primary analysis was performed on an intention-to-treat (ITT) analysis basis. Sensitivity analyses included (1) restricting the intervention group to those who took up the offer of the app, (2) using complete cases, and (3) using multiple imputation.

Results: The effective follow-up rate for 7 months was 41.9%. The primary analysis showed no evidence of a benefit of the intervention on rates of 6-month continuous abstinence (intervention 6.8% vs comparator 7.0%; relative risk 0.97, 95% CI 0.75-1.26). Analyses of all secondary outcomes also showed no evidence of a benefit. Similar results were observed on complete cases and using multiple imputation. When the intervention group was restricted to those who took up the offer of the app (n=395, 25.3%), participants in the intervention group were 80% more likely to report 6-month continuous abstinence (12.7% vs 7.0%; relative risk 1.80, 95% CI 1.30-2.45). Equivalent subgroup analyses produced similar results on the secondary outcomes. These differences persisted after adjustment for key baseline characteristics.

Conclusions: Among motivated smokers provided with very brief advice to quit, the offer of the Smoke Free app did not have a detectable benefit for cessation compared with follow-up only. However, the app increased quit rates when smokers randomized to receive the app downloaded it.

Trial Registration: ISRCTN ISRCTN85785540; https://www.isrctn.com/ISRCTN85785540

International Registered Report Identifier (IRRID): RR2-https://onlinelibrary.wiley.com/doi/full/10.1111/add.14652

(J Med Internet Res 2024;26:e50963) doi:10.2196/50963

KEYWORDS

randomized controlled trial; smartphone app; smoking cessation; digital intervention; tobacco; mobile phone

XSL•FO
RenderX

Introduction

Tobacco remains one of the leading causes of disease and preventable death worldwide, killing more than 8 million people each year [1]. A range of evidence-based support options are available to help smokers quit [2-4], but barriers including a lack of time and (in countries where free support is unavailable) inability to pay mean that many smokers do not access this support [5-7]. Digital technologies offer the potential for low-cost, scalable delivery of interventions to promote smoking cessation. More smokers have turned to remote options for cessation support since the COVID-19 pandemic began, including smartphone apps [8], but evidence of the effectiveness of existing apps is lacking. This study aimed to assess the effectiveness of the offer of a widely used app for smoking cessation compared with no support.

Digital support for smoking cessation has the potential to contribute to meaningful reductions in smoking prevalence in countries around the world. In particular, the past decade has seen a surge in smartphone apps offering support for smokers who want to quit [9]. Combined with increasing levels of smartphone ownership (currently estimated at 3.8 billion worldwide [10]; 74% of the adult population [11]), these apps can reach large numbers of smokers. However, the potential of apps for promoting cessation is not yet being realized. A 2019 Cochrane review of mobile phone interventions for smoking cessation [12] identified just 5 studies (total N=3079) that compared a smoking cessation smartphone app with lower-intensity smoking cessation support. The pooled data provided no evidence that apps increased the likelihood of cessation (relative risk [RR] 1.00, 95% CI 0.66-1.52), but the evidence was judged to be of very low certainty limiting confidence in the effect estimate. The authors called for more large-scale randomized controlled trials (RCTs) to establish whether smartphone app interventions are effective for smoking cessation. In 2020, a large RCT [9] (n=2415) tested the efficacy of a smoking cessation app based on acceptance and commitment therapy compared with a simpler app informed by US clinical practice guidelines. Self-reported quit rates at 12 months were higher among participants randomized to use the acceptance and commitment therapy–based app (28.2% vs 21.1%, odds ratio 1.49, 95% CI 1.22-1.83). To our knowledge, no published RCTs have compared apps designed to provide ongoing support with unaided quitting.

With over 6 million downloads to date and 70,000 new users each month, Smoke Free is one of the world's most widely used smoking cessation apps. The app is evidence-informed and available for iOS and Android OS. In a large exploratory RCT (n=28,112) conducted between 2013 and 2015, the full version of the app increased 3-month self-reported continuous abstinence rates compared with a reduced version (odds ratio 1.90, 95% CI 1.53-2.37) [13]. However, this result was limited by relatively short-term outcomes and very low follow-up rates (8.5% and 6.5% in the intervention and control conditions, respectively) with no active attempts to recontact participants. Smoke Free therefore constitutes a useful test bed for assessing the effectiveness of a smartphone app for smoking cessation versus unaided quitting.

This paper describes the results of the App for Smoking Cessation Evaluation Trial [14], a 2-arm RCT designed to evaluate the effectiveness of the offer of the Smoke Free app in increasing rates of tobacco smoking cessation compared with follow-up only. We adopted a pragmatic design to provide information on the usefulness of this app in real-world settings. Given that, in an effectiveness trial of this nature, not all participants who are offered the intervention will take it up, we were therefore testing the offer rather than the actual use of the app. The primary research question was:

1. How effective is an offer to use the app plus follow-up (intervention) compared with no offer of the app and follow-up only (comparator) in promoting self-reported smoking cessation for at least 6 months, assessed 7 months after enrollment?

We also addressed several secondary research questions:

1. How effective is the intervention in promoting:
   - At least one quit attempt in the 1 month following enrollment in this study?
   - Smoking cessation for at least 3 months, assessed 4 months after enrollment?
   - Smoking cessation for at least 6 months among those who make at least one quit attempt in the 1 month following enrollment in this study, assessed 7 months after enrollment?
   - Downloading or using the Smoke Free app at least once, assessed 7 months after enrollment?
2. Do the answers to the primary research question and secondary research questions differ according to smokers' gender, level of cigarette addition, age, education, financial situation, or prior experience with a smoking cessation smartphone app?

Methods

Overview

A summary timeline of trial procedures is shown in Multimedia Appendix 1.

Design

The app for the smoking cessation evaluation trial was a 2-arm individual RCT. This study's protocol [14] and analysis plan [15] were preregistered (ISRCTN85785540). An independent Trial Steering Committee provided overall supervision of the trial.

Setting

This study was conducted online using the Qualtrics survey platform, with no restriction on location.

Participants

Inclusion criteria were current cigarette smoker, aged ≥18 years, English speaker, owns a smartphone, provided a valid email address not previously used by another participant, interested in making a quit attempt within the next month, and willing to be followed up by email and complete online questionnaires.

XSL•FO
RenderX

Eligibility was assessed via screening questions embedded at the start of the baseline study questionnaire on Qualtrics. Those who did not meet the inclusion criteria were directed to the NHS Smokefree website [16] for resources to help with quitting smoking.

Participants were not provided with any financial compensation, but we offered to donate £10 (US $12) to a cancer charity on behalf of each participant who responded to the final follow-up.

Sample Size

The intended sample size was decided a priori based on achieving 90% power to detect a RR ≥1.5 with an α of $P<.05$, 1 - tailed, and a quit rate of 6.0% in the comparator group. This led to a target sample size of 3116; 1558 per group.

We note that we amended our original power calculation as reported in the published protocol [14]. Details of these amendments and the rationale underlying them are available on the Open Science Framework [15]. The final sample size target was approved by the Trial Steering Committee on January 6, 2021, and registered on ISRCTN (International Standard Randomised Controlled Trial Number) in January 2021 after we had randomized 2798 participants and before any of the primary outcome data (6-month continuous abstinence assessed at 7 months) were collected.

Ethical Considerations

Ethical approval for the trial was obtained from the University College London Research Ethics Committee (CEHP/2020/579). All participants were provided with a summary of this study and their right to withdraw on the landing page of the baseline survey on Qualtrics. They provided informed consent by selecting "Yes: I confirm I have read the information about the study and wish to participate."

Recruitment

Recruitment took place between August 2020 and April 2021. Participants were recruited via advertisements on social media (Facebook and Twitter; Multimedia Appendix 1) and a mailing list of smokers who had previously signed up for the Smoke Free app and had agreed to be contacted. We emailed people on the mailing list who signed up to the app >6 months previously with an invitation to participate in this study. Response to these emails was low (~5/1000). Most of our participants were recruited via Facebook adverts (at an average cost of £3 [US $3.92]/participant), which linked to the Qualtrics baseline questionnaire (with embedded consenting procedure and screening questions).

Randomization

Consenting participants who met the eligibility criteria were randomized after completing the baseline questionnaire. Randomization was 1:1 at the individual level with no restriction (ie, no blocking) and was automated within Qualtrics, such that each participant was shown at random either the intervention message including the offer of the Smoke Free app or the comparator message after the final questionnaire question.

All investigators were blinded to participants' treatment allocation until all data had been collected. The data were analyzed blindly by the trial statistician (EB).

Interventions

Comparator

After consenting and completing the baseline questionnaire, participants in the comparator condition were shown a final screen with a brief message encouraging them to make a quit attempt within the next 4 weeks and reminding them of the importance of responding to follow-up requests designed to track their progress (Multimedia Appendix 1). This same message was also emailed to them immediately afterward.

Intervention: Smoke Free App

Participants in the intervention condition received the same advice as those in the comparator condition plus the offer of the full version of the Smoke Free app free of charge, encouragement to use the app, and a link to download it. This same message and information on how to access the app were also emailed to them immediately afterward.

The Smoke Free app is based on behavior change techniques that would be expected from theory [13] and evidence with face-to-face support [17,18] to aid smoking cessation. It guides smokers through the first 3 months of their quit attempt by helping them maintain their resolve by setting a clear goal, monitoring their progress toward that goal, and becoming aware of the benefits of being smoke-free achieved to date. More detail is provided in Multimedia Appendix 1.

Follow-Up Data Collection

Follow-up data was collected between September 2020 and December 2021, via online questionnaires 1, 4, and 7 months after study enrollment. Invitations to complete the 1- and 4-month surveys were sent via email (automated within Qualtrics). For the final (7-month) follow-up, to boost response rates for our primary outcome, we contacted participants up to 6 times over 2 weeks. First, they were invited via email within Qualtrics. Next, a further email invitation was sent from one of the research team's personal email addresses (to reduce the "spam" rating of the email). Then participants who provided their phone numbers were contacted via SMS text messaging (between March and August 2021) or telephone call (between August and December 2021; a change implemented to boost the response rate) and asked to respond "yes" or "no" to the key outcome assessment. Finally, up to 3 further emails were sent asking the same question as the SMS text messaging, prompting participants to a direct "yes" or "no" response via email. Participants who responded to any of the invitations or reminders were not contacted further.

Measures

Participant Baseline Characteristics

The baseline questionnaire assessed the following: email address, mobile phone number (optional), smartphone ownership, motivation to quit in the next month (Motivation To Stop Scale [19]), willingness to complete online questionnaires after 1, 4 and 7 months, age (18-34, 35-64, or

XSL•FO
RenderX

≥65 years), gender (male or female), education (any or no educational qualifications gained aged ≥16 years), financial status (live comfortably, meet needs with a little left, just meet basic expenses, or do not meet basic expenses) [20], country of residence, first language (English or other), number of cigarettes smoked per day, level of cigarette addiction (first cigarette after waking within 5 minutes, 6-30 minutes, 31-60 minutes, or >60 minutes), history of serious quit attempts (never or yes, not in the past year, or yes in the past year), and past and current use of support for smoking cessation (prescription nicotine replacement therapy or nicotine replacement therapy bought over-the-counter, varenicline, bupropion, face-to-face behavioral support, telephone support, written self-help materials, websites, or apps). In an exploratory addition to outcome assessment, given evidence that heart rate declines substantially when smokers stop [21], participants who had a heart rate monitoring device (eg, FitBit or Apple watch) were asked to report their average resting heart rate as measured by this device.

Outcomes

The primary outcome was the percentage of participants reporting not having smoked for 6 months at the 7-month follow-up. This was assessed in the online questionnaire with the question: "Have you smoked any cigarettes in the past 6 months?" with response options "none at all," "between 1 and 5," and "more than 5." In line with the Russell Standard for self-report of smoking abstinence [22], the former 2 responses were collapsed for analysis, with data coded 1 for respondents reporting smoking no more than 5 cigarettes in the past 6 months and 0 for those reporting smoking more than 5 cigarettes. Where participants did not respond to the invitations to complete the questionnaire, the question was simplified to "Have you smoked more than 5 cigarettes in the past 6 months?" and participants were asked to reply "yes" or "no" via SMS text messaging, telephone, or email. Based on the intention-to-treat (ITT) analysis principle, those who did not respond to follow-up attempts were retained in the analyses and classified as continuing smokers [22].

Secondary outcomes were the percentage of participants reporting (1) quit attempts at 1-month follow-up, defined as having made a serious quit attempt in the last 4 weeks (assessed with the question: "Have you made a serious attempt to quit smoking in the last 4 weeks? Please include any attempt that you are currently making [yes=1/no=0]," (2) smoking cessation for at least 3 months at the 4-month follow-up (assessed with the question: "Have you smoked a single puff on a cigarette in the past 3 months? [yes=0/no=1]"), (3) smoking cessation for at least 6 months at the 7-month follow-up in those who made a quit attempt (assessed with the question: "Have you smoked any cigarettes in the past 6 months? [none at all=1/between 1 and 5=1/more than 5=0]"), and (4) app use, defined as downloading or using the Smoke Free app at least once at any point during this study's period (assessed at the 7-month follow-up with the question: "In the last 7 months, have you downloaded or used the Smoke Free app (pictured) at least once? [yes=1/no=0]").

Statistical Analyses

Overview

We followed our preregistered analysis plan [14], with 2 amendments registered on the Open Science Framework before running the analysis [15] and 2 unplanned sensitivity analyses after running the analyses. Details of amendments to the analysis plan are summarized in Multimedia Appendix 2.

All variables were collected primarily online and entered automatically into a Qualtrics database. From this database, a user-specified Excel (Microsoft Corp) file was downloaded, subjected to basic processing and recoding, and integrated with responses provided by text messages. On completion, data were analyzed blind to intervention allocation using R Studio (version 4.2.1; R Foundation).

Primary Analyses

Our primary analyses used an ITT analysis approach, which treated those with missing data as smoking [22]. We used log-binomial regression to calculate the RR and 95% CI of each primary and secondary outcome in the intervention group versus the comparator group.

Moderation Analyses

For each primary and secondary end point, we ran a series of log-binomial regression models in which we added 2-way interactions between group and gender, cigarette addiction, age (<35/≥35 years to broadly distinguish between younger and older participants with the specific threshold corresponding to the age at which quitting avoids most of the excess mortality from smoking [23]), education, financial situation, and previous experience with using a smoking cessation app (based on self-reported use ever of cessation aids at baseline). The results are reported in Multimedia Appendix 3.

Sensitivity Analyses

We repeated our primary and secondary analyses (1) restricting the intervention group to participants who took up the offer of free full access to the Smoke Free app, which was self-reported or verified by matching the email address used to log in to the app to the one provided in the baseline questionnaire (with and without adjustment for key baseline characteristics); (2) restricting both groups to participants who were successfully followed up; and (3) using multiple imputation to impute missing outcomes data. In an exploratory analysis, we also repeated the analyses defining successful quits as self-reported abstinence plus a reduction in mean resting heart rate of ≥5 beats per minute, based on the lower 95% CI for the difference in resting heart rate between people smoking as usual and not smoking in a previous study [24]. Finally, we reanalyzed our primary outcome assuming different rates of abstinence in those not followed up on.

To aid interpretation of the strength of evidence for associations, we calculated Bayes factors [25] to differentiate between evidence for an effect, no effect, and data insensitivity. We used a half-normal distribution, the mode at 0 (no effect), and the SD equal to the expected effect size used in the sample size calculation (RR 1.5).

XSL•FO
RenderX

Results

Overview

A total of 3143 eligible participants were recruited, completed the baseline assessment, and were randomized to the intervention or comparator condition. Figure 1 shows the numbers allocated to each group and followed-up. Baseline sociodemographic and smoking characteristics were similar across groups (Multimedia Appendix 4), consistent with successful randomization.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart.

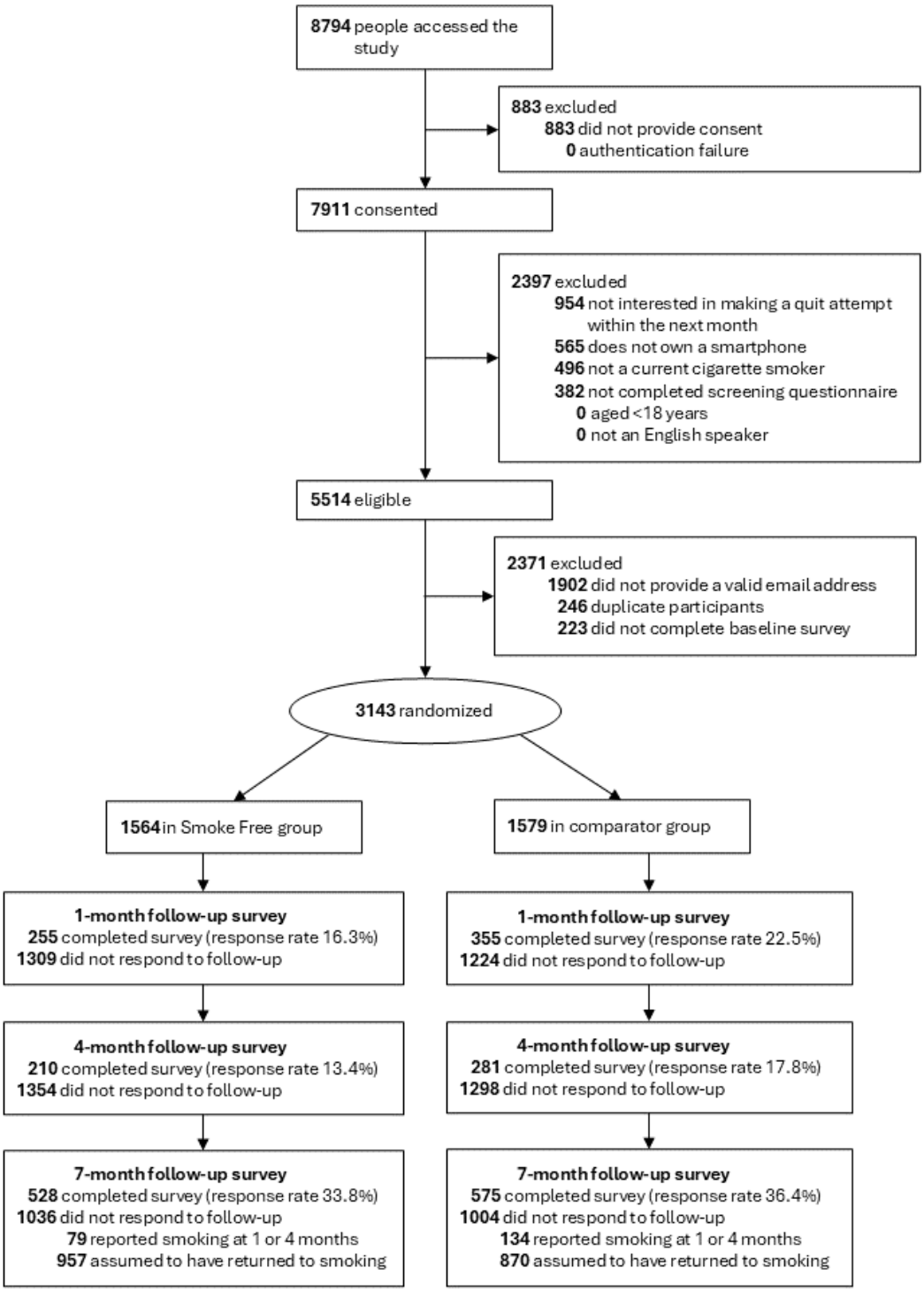

Response rates to the 1-month, 4-month, and 7-month follow-ups were 19.4%, 15.6%, and 35.1%, respectively; 16.3%, 13.4%, and 33.8% in the Smoke Free group, and 22.5%, 17.8%, and 36.4% in the comparator group (Figure 1). Of the 2040 participants who did not respond directly to the final follow-up, 213 (10.4%) individuals reported smoking in a previous follow-up assessment, which would have classified them as smokers by our continuous abstinence primary outcome,

XSL•FO
RenderX

meaning the effective follow-up rate for the primary outcome was 41.9% (2040 – 213 = 1827; 3143 – 1827 = 1316; 1316/3143); 38.8% in the Smoke Free group and 44.9% in the comparator group.

Primary Outcome

Table 1 shows results for the primary outcome of 6-month continuous abstinence, assessed at 7-month follow-up, using an ITT analysis approach. Overall, 6.9% (218/3143) of participants reported 6-month continuous abstinence. The rate of smoking cessation was similar between participants in the Smoke Free and comparator groups (6.8% vs 7.0%, respectively). The Bayes factor favored no effect. Moderation analyses showed no significant difference in treatment effect by any characteristic (Multimedia Appendix 3).

Table 1. Log-binomial regression analyses of treatment effect on primary and secondary outcomes[a].

	Comparator	Offered Smoke Free app			
	n/N (%)	n/N (%)	RR[b] (95% CI)	*P* value	BF[c,d]
Primary outcome					
6-month continuous abstinence[e]	111/1579 (7.03)	107/1564 (6.84)	0.97 (0.75 to 1.26)	.84	0.26
Secondary outcomes					
Making at least one quit attempt[f]	242/1579 (15.33)	180/1564 (11.51)	0.75 (0.63 to 0.90)	.002	0.05
3-month continuous abstinence[g]	54/1579 (3.42)	45/1564 (2.88)	0.84 (0.57 to 1.24)	.38	0.26
6-month continuous abstinence among those who tried to quit[e,h]	48/242 (19.8)	43/180 (23.9)	1.20 (0.83 to 1.73)	.32	1.01
Reported downloading or using the Smoke Free app at least once[e]	51/1579 (3.23)	85/1564 (5.43)	1.68 (1.20 to 2.38)	.003	28.6

[a]Intention-to-treat analysis.

[b]RR: relative risk (reference group: comparator).

[c]BF: Bayes factor.

[d]Bayes factors ≥3 can be interpreted as evidence for the alternative hypothesis, ≤1/3 as evidence for the null hypothesis, and Bayes factors between 1/3 and 3 suggest the data are insensitive to distinguishing the alternative hypothesis from the null [26,27].

[e]Assessed at 7-month follow-up.

[f]Assessed at 1-month follow-up.

[g]Assessed at 4-month follow-up.

[h]Tried to quit: reported making at least one quit attempt at 1-month follow-up.

Secondary Outcomes

Table 1 also summarizes results relating to the secondary outcomes. Participants in the Smoke Free group had a 25% lower risk of reporting a quit attempt compared with those in the comparator group. There was no statistically significant difference in 3-month continuous abstinence rates between groups or 6-month continuous abstinence among those who tried to quit (Bayes factors for these outcomes favored no effect). Those in the Smoke Free group were 68% more likely to report having downloaded or used the Smoke Free app at least once during this study's period than those in the comparator group, although self-reported rates were very low in both groups (5.4% vs 3.2%, respectively).

Sensitivity Analyses

Table 2 shows results for primary and secondary outcomes restricting the intervention group to those who took up the offer of the Smoke Free app. Despite just 85/1564 (5.3%) participants in the intervention group recalling, at 7-month follow-up, having downloaded or used the Smoke Free app at least once, matching email address log-ins were verified for 355/1564 (22.7%) participants in the intervention group—indicating they had (at least briefly) taken up the offer of the app after the baseline survey. This discrepancy is likely to be largely due to loss to follow-up; most of the participants either did not respond to the 7-month follow-up or only provided data on the primary outcome. Combined, 395/1564 (25.3%) participants in the intervention group either self-reported or had verified app use. Multimedia Appendix 5 compares the baseline characteristics of these intervention participants with those of the comparator group. The 6-month continuous abstinence rate among the subset of the intervention group who took up the offer of the app was 80% higher than the comparator group (12.7% vs 7.0%; Table 2). After adjustment for baseline covariates, the difference between groups attenuated to 60% but remained statistically significant ($P<.05$; Table 2). Among this subset of the intervention group, the rate of quit attempts was 40% higher than the comparator group (21.5% vs 15.3%), 3-month continuous abstinence was 85% higher (6.3% vs 3.4%), and 6-month continuous abstinence among those who made a quit attempt was 60% higher (30.3% vs 19.8%); these differences remained statistically significant after adjustment for covariates ($P<.05$; Table 2).

XSL•FO
RenderX

Table 2. Sensitivity analysis: restricting the intervention group to those who took up the offer of the app (self-reported or verified)[a].

	Comparator	Offered Smoke Free app	Unadjusted			Adjusted for baseline characteristics[b]		
	n/N (%)	n/N (%)	RR[c] (95% CI)	*P* value	BF[d,e]	RR (95% CI)	*P* value	BF[e]
Primary outcome								
6-month continuous abstinence[f]	111/1579 (7.03)	50/395 (12.7)	1.80 (1.30-2.45)	<.001	>100	1.60 (1.16-2.18)	.003	34.1
Secondary outcomes								
Making at least one quit attempt[g]	242/1579 (15.33)	85/395 (21.5)	1.40 (1.12-1.74)	.003	38.1	1.33 (1.06-1.64)	.01	5.28
3-month continuous abstinence[h]	54/1579 (3.42)	25/395 (6.3)	1.85 (1.15-2.90)	.009	14.9	1.62 (1.01-2.53)	.04	5.25
6-month continuous abstinence among those who tried to quit[f,i]	48/242 (19.8)	27/85 (32)	1.60 (1.05-2.37)	.02	7.94	1.48[j] (1.00-2.21)	.05	3.73

[a]Intention-to-treat analysis.

[b]Adjusted for age, financial status, level of addiction (time to first cigarette after waking), and current use of evidence-based support (nicotine replacement therapy on prescription, varenicline, bupropion, face-to-face support, or e-cigarettes).

[c]RR: relative risk (reference group: comparator).

[d]BF: Bayes factor.

[e]Bayes factors ≥3 can be interpreted as evidence for the alternative hypothesis, ≤1/3 as evidence for the null hypothesis, and Bayes factors between 1/3 and 3 suggest the data are insensitive to distinguish the alternative hypothesis from the null [26,27].

[f]Assessed at 7-month follow-up.

[g]Assessed at 1-month follow-up.

[h]Assessed at 4-month follow-up.

[i]Tried to quit: reported making at least one quit attempt at 1-month follow-up.

[j]There were convergence issues for this model. To ensure convergence: (1) iterations were increased to 1000 and the expectation-maximization algorithm was used rather than the default iteratively reweighted least squares algorithm [28], and (2) evidence-based support was excluded as a covariate.

Table 3 shows results for primary and secondary outcomes with analyses rerun (1) on complete cases and (2) with missing data imputed using multiple imputation. In both analyses, there was no statistically significant difference between groups on quitting outcomes (Bayes factors indicated the data were insensitive or favored no effect). Participants in the intervention group were significantly more likely to report having downloaded or used the Smoke Free app at least once.

XSL•FO
RenderX

Table 3. Sensitivity analyses: complete cases and multiple imputation.

	Comparator n/N (%)	Offered Smoke Free app n/N (%)	RR[a] (95% CI)	*P* value	BF[b,c]
Complete cases					
Primary outcome					
6-month continuous abstinence[d]	111/712 (15.6)	107/609 (17.6)	1.13 (0.88-1.44)	.33	0.75
Secondary outcomes					
Making at least one quit attempt[e]	242/354 (68.4)	180/255 (70.6)	1.03 (0.93-1.15)	.56	0.22
3-month continuous abstinence[f]	54/438 (12.3)	45/309 (14.6)	1.18 (0.81-1.70)	.38	0.91
6-month continuous abstinence among those who tried to quit[d,g]	48/225 (21.3)	43/160 (26.9)	1.26 (0.88-1.80)	.21	1.41
Reported downloading or using the Smoke Free app at least once[d]	51/212 (24.1)	85/176 (48.3)	2.01 (1.52-2.69)	<.001	>100
Multiple imputation					
Primary outcome					
6-month continuous abstinence[d]	20.01 (316/1579)	22.38 (350/1564)	1.12 (0.94-1.33)	.21	0.84
Secondary outcomes					
Making at least one quit attempt[e]	63.46 (1002/1579)	61.32 (959/1564)	0.97 (0.87-1.07)	.49	0.08
3-month continuous abstinence[f]	17.04 (269/1579)	19.12 (299/1564)	1.12 (0.91-1.39)	.29	0.72
6-month continuous abstinence among those who tried to quit[d,g]	23.58 (241/1022)	26.5 (254/959)	1.22 (0.85-1.74)	.28	1.14
Reported downloading or using the Smoke Free app at least once[d]	26.35 (416/1579)	45.27 (708/1564)	1.72 (1.39-2.13)	<.001	>100

[a]RR: relative risk (reference group: comparator).

[b]BF: Bayes factor.

[c]Bayes factors ≥3 can be interpreted as evidence for the alternative hypothesis, ≤1/3 as evidence for the null hypothesis, and Bayes factors between 1/3 and 3 suggest the data are insensitive to distinguish the alternative hypothesis from the null [26,27].

[d]Assessed at 7-month follow-up.

[e]Assessed at 1-month follow-up.

[f]Assessed at 4-month follow-up.

[g]Tried to quit: reported making at least one quit attempt at 1-month follow-up.

Just 108 participants reported their resting heart rate at baseline and 7-month follow-up. When we defined successful quits as self-reported abstinence plus a reduction in mean resting heart rate of ≥5 beats per minute, the rate of 6-month continuous abstinence did not differ significantly between intervention and comparator arms (5/1564, 0.32% vs 9/1579, 0.57%; RR 0.56, 95% CI 0.17-1.62, *P*=.30). The Bayes factor indicated the data were insensitive (Bayes factor=0.51).

Assuming different rates of abstinence among participants who were lost to follow-up had little effect on our primary outcome (Multimedia Appendix 6).

Discussion

Among motivated smokers provided with very brief advice to quit, there was no significant difference in 6-month quit rates between participants randomized to receive the offer of the Smoke Free app plus follow-up and those randomized to follow-up only. This result was observed across analyses using ITT analysis, complete cases, and multiple imputation, and there were similar results on the secondary outcomes. However, when the intervention group was restricted to those who took up the offer of the Smoke Free app a significant benefit of treatment was observed, with participants in the intervention group being 80% more likely to report abstinence than those in the comparator group on the primary outcome, with similar results on secondary outcomes. This was only partly explained by differences in baseline characteristics, with the effect remaining at 60% (a statistically significant difference) after adjustment for age, financial status, level of addiction, and current use of evidence-based cessation support.

There was no significant difference in 6-month continuous abstinence among those who tried to quit in ITT analysis, complete case, and multiply imputed analyses. These analyses

XSL•FO
RenderX

were limited by the low response to the 1-month follow-up survey (which assessed quit attempts) and Bayes factors indicated the data were insensitive. The data showed a benefit of treatment when the intervention group was restricted to those who took up the offer of the Smoke Free app, with the intervention group 60% more likely to report abstinence than those in the comparator group.

ITT analyses indicated a lower rate of quit attempts in the first 4 weeks in the intervention group compared with the comparator. However, this was not consistently observed across sensitivity analyses that used complete cases or multiple imputation, which showed no significant effect (with Bayes factors favoring no difference). It is likely the ITT analysis result was an artifact resulting from the lower response rate to the 1-month follow-up survey in the intervention versus comparator group (16% vs 23%, meaning a greater proportion of the intervention group were assumed not to have made a quit attempt) rather than any genuine difference between the groups. Indeed, when the intervention group was restricted to those who took up the offer of the Smoke Free app, a significant benefit of treatment was observed, with participants in the intervention group 40% more likely to report attempting to quit.

There was no statistically significant difference in 3-month continuous abstinence rates between groups. Analyses of this outcome were limited by the low response rate to the 4-month follow-up survey (16%). Bayes factors indicated the data were insensitive for all analyses except the ITT analysis, which favored no effect, and the analysis restricting the intervention group to those who took up the offer of the app, which showed a significant benefit of treatment (85% more likely to report 3-month continuous abstinence).

Participants in the intervention group were significantly more likely to report having downloaded or used the Smoke Free app at least once during this study's period, although uptake of the offer of the app was low across self-report (5%) and validated (23%) measures (25% overall). Prevalence of self-reported uptake was suppressed by the low response to the final follow-up survey, particularly because many responders (ie, those who responded via email or telephone) only provided data on the primary outcome. However, the validated measure of treatment uptake will have captured most of the participants who took up the offer of the app, as long as they used the same email address to sign up for this study and register for the app.

To our knowledge, this is the first study to test the effectiveness of a smoking cessation app compared with unaided quitting. It differs from other large trials of smoking cessation apps [9] not only in its comparator group (ie, follow-up only rather than active treatment) but also in the way it was advertised, making no reference to smartphone apps until participants were enrolled. It did not aim to target smokers interested in quitting with the support of an app, but rather any smoker motivated to make a quit attempt. Thus, the relatively low uptake of the offer of the app in the intervention group is not surprising. Representative observational data show low rates of adoption of digital aids for smoking cessation, with fewer than 3% of smokers who have tried to quit reporting using a digital cessation aid (app or website) [29]. Our results suggest that while not every smoker will be interested in trying them, the use of smoking cessation apps can be increased by directing smokers to this type of support (395/1564 [25.26%] of those offered registered an account with the app). Our data also suggest that the Smoke Free app boosted quit rates among smokers who used it. However, it is possible that this may have reflected participants who downloaded the app as being more serious and motivated about quitting (as indicated by acting and downloading or using the app) and that other cessation aids could have produced similar results (vs no aid). Given the wide reach of smartphone apps, it is possible that initiatives to increase smokers' awareness of smoking cessation apps could have a meaningful impact on rates of cessation at the population level even if only a minority of smokers take up the use of app-based support (even a small percentage of a very large number can be a large number). Analyses based on complete cases and multiple imputed data indicated offering the app increased the risk of 6-month continuous abstinence by 10% versus follow-up only. While this would generally be considered a small effect size for a behavioral intervention, small effects of treatments that aid smoking cessation can be clinically significant because of the very large health gains that accrue from stopping smoking [30]. Moreover, offering the app to all smokers at a population level (a low-cost, highly scalable intervention) could result in a large number of successful quits. Future trials should prioritize establishing the relative effectiveness and cost-effectiveness of apps compared with other evidence-based smoking cessation treatments.

Strengths of this study include the large sample size, the wide geographic scope, and the 6-month follow-up duration [22]. The pragmatic design offers real-world insights, focusing on the offer rather than the use of the app, as not every smoker will want to use digital support or apps. There were also several limitations. First, there was a high rate of attrition. Given the pragmatic design and light-touch intervention, we anticipated that response to follow-up attempts would be relatively low [14] and concentrated our limited resources on maximizing response to the final follow-up at 7 months. Second, our study should have been powered for a smaller effect size given the low rate of uptake in the intervention group. With just 25.26% (395/1564) of participants taking up the offer of the app and an observed effect size of RR ~1.3-1.5 among this group, future trials of this nature would need to be powered to detect smaller effects (RR ~1.1), requiring samples in the region of 64,000 participants. This is not unfeasible given the numbers involved in some digital trials [31]. Future studies could investigate barriers to app use to explore the low rate of uptake of the offer of free access to a paid app in a motivated group of smokers. Third, we did not undertake remote biochemical data collection to verify abstinence. While biochemical verification is widely considered the gold standard for evaluating cessation outcomes [22,32], the Society for Research on Nicotine and Tobacco Subcommittee on Biochemical Verification has advised that in large-scale population-based trials such as this one, where face-to-face contact is limited and data are optimally collected online, the added precision gained by biochemical verification may be offset by methodological problems in such a way that its use is not required and may not be desirable [32,33]. Given our study's large, geographically dispersed participant sample,

XSL•FO
RenderX

collecting biological samples or conducting remote observation of rapid tests would have been costly and logistically challenging. It may also have reduced the representativeness of the sample (if smokers unwilling to provide samples were ineligible to participate) or increased the rate of missing outcome data (if logistical complexity and participant burden reduced the likelihood of follow-up response or resulted in unusable samples) [34]. We explored the possibility of verifying abstinence via a reduction in self-reported resting heart rate, but only a small minority of participants provided this data and the analysis was insensitive. Collecting this data was not an important aspect of our follow-up strategy and it may be feasible in future trials with additional incentive. Finally, while we recruited an international sample, with no restrictions on location, the majority (97.26%, 3057/3143) of participants were from 5 high-income Western countries, which may limit generalizability. Further investigation is required in low- and middle-income countries, where the potential benefits of smoking cessation apps may be greater in the absence of comprehensive and affordable cessation support.

In conclusion, among motivated smokers provided with very brief advice to quit, the Smoke Free app did not have a detectable benefit for cessation compared with follow-up only. However, the app increased quit rates among smokers who were randomized to receive it and who downloaded it.

Acknowledgments

We gratefully acknowledge David Crane, PhD, creator of the Smoke Free app, for making this research possible. We also thank our Trial Steering Committee for their thoughtful guidance throughout the project. The funding was by Cancer Research UK (PRCRPG-Nov21\100002).

Data Availability

Data are available on the Open Science Framework [15].

Conflicts of Interest

EB and JB have received unrestricted research funding from Pfizer, and JB only from J&J, who manufacture smoking cessation medications. RW has undertaken research and consultancy for and receives travel funds and hospitality from manufacturers of smoking cessation medications (Pfizer, GlaxoSmithKline, and Johnson and Johnson). OP, RW, and JB are unpaid members of the scientific steering group of the Smoke Free mobile app. All authors declare no financial links with tobacco companies, e-cigarette manufacturers, or their representatives.

Multimedia Appendix 1
Additional information on methods.
[PDF File (Adobe PDF File), 261 KB - jmir_v26i1e50963_app1.pdf]

Multimedia Appendix 2
Amendments to the analysis plan.
[PDF File (Adobe PDF File), 133 KB - jmir_v26i1e50963_app2.pdf]

Multimedia Appendix 3
Results of moderation analyses.
[PDF File (Adobe PDF File), 221 KB - jmir_v26i1e50963_app3.pdf]

Multimedia Appendix 4
Baseline characteristics.
[PDF File (Adobe PDF File), 200 KB - jmir_v26i1e50963_app4.pdf]

Multimedia Appendix 5
Baseline characteristics, restricting intervention group to those who took up the offer of the app.
[PDF File (Adobe PDF File), 193 KB - jmir_v26i1e50963_app5.pdf]

Multimedia Appendix 6
Impact of assuming different rates of abstinence in participants lost to follow-up.
[PDF File (Adobe PDF File), 136 KB - jmir_v26i1e50963_app6.pdf]

Multimedia Appendix 7

XSL•FO
RenderX

CONSORT 2010 checklist. CONSORT: Consolidated Standards of Reporting Trials.
[PDF File (Adobe PDF File), 780 KB - jmir_v26i1e50963_app7.pdf]

References

1. Tobacco. WHO. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/tobacco [accessed 2018-12-02]
2. Cahill K, Stevens S, Perera R, Lancaster T. Pharmacological interventions for smoking cessation: an overview and network meta-analysis. Cochrane Database Syst Rev 2013;2013(5):CD009329 [FREE Full text] [doi: 10.1002/14651858.CD009329.pub2] [Medline: 23728690]
3. Hartmann-Boyce J, Chepkin S, Ye W, Bullen C, Lancaster T. Nicotine replacement therapy versus control for smoking cessation. Cochrane Database Syst Rev 2018;5(5):CD000146 [FREE Full text] [doi: 10.1002/14651858.CD000146.pub5] [Medline: 29852054]
4. Hartmann-Boyce J, Livingstone-Banks J, Ordóñez-Mena J, Fanshawe T, Lindson N, Freeman S. Behavioural interventions for smoking cessation: an overview and network meta-analysis. Cochrane Database Syst Rev 2021:1-77. [doi: 10.1002/14651858.cd013229]
5. van Wijk EC, Landais LL, Harting J. Understanding the multitude of barriers that prevent smokers in lower socioeconomic groups from accessing smoking cessation support: a literature review. Prev Med 2019;123:143-151. [doi: 10.1016/j.ypmed.2019.03.029] [Medline: 30902700]
6. Kale D, Gilbert H, Sutton S. An exploration of the barriers to attendance at the English stop smoking services. Addict Behav Rep 2019;9:100141 [FREE Full text] [doi: 10.1016/j.abrep.2018.10.005] [Medline: 31193736]
7. Kwah KL, Fulton EA, Brown KE. Accessing national health service stop smoking services in the UK: a COM-B analysis of barriers and facilitators perceived by smokers, ex-smokers and stop smoking advisors. Public Health 2019;171:123-130. [doi: 10.1016/j.puhe.2019.03.012] [Medline: 31128557]
8. Jackson SE, Cox S, Shahab L, Brown J. Prevalence of use and real-world effectiveness of smoking cessation aids during the COVID-19 pandemic: a representative study of smokers in England. Addiction 2022;117(9):2504-2514 [FREE Full text] [doi: 10.1111/add.15903] [Medline: 35491743]
9. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. JAMA Intern Med 2020;180(11):1472-1480 [FREE Full text] [doi: 10.1001/jamainternmed.2020.4055] [Medline: 32955554]
10. Turner A. How many smartphones are in the world?. 2021. URL: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world [accessed 2021-09-30]
11. Ang C. Visualizing the world's population by age group. Visual Capitalist. 2021. URL: https://www.visualcapitalist.com/the-worlds-population-2020-by-age/ [accessed 2021-09-30]
12. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. Cochrane Database Syst Rev 2019;10(10):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub5] [Medline: 31638271]
13. Crane D, Ubhi HK, Brown J, West R. Relative effectiveness of a full versus reduced version of the ‘Smoke Free’ mobile application for smoking cessation: a randomised controlled trial. F1000Res 2018;7:1524. [doi: 10.12688/f1000research.16148.1]
14. Jackson SE, Perski O, Crane D, Michie S, West R, Brown J. Effectiveness of an offer of the smoke free smartphone application for smoking cessation: protocol for a randomized controlled trial. Addiction 2019;114(11):2078-2086. [doi: 10.1111/add.14652] [Medline: 31083767]
15. Jackson S, Brown J, Kale D, Perski O, Beard E. ASSET trial. 2020. URL: https://osf.io/umec4/ [accessed 2023-07-18]
16. Quit smoking—better health. NHS. 2020. URL: https://www.nhs.uk/better-health/quit-smoking/ [accessed 2023-07-18]
17. West R, Evans A, Michie S. Behavior change techniques used in group-based behavioral support by the English stop-smoking services and preliminary assessment of association with short-term quit outcomes. Nicotine Tob Res 2011;13(12):1316-1320. [doi: 10.1093/ntr/ntr120] [Medline: 21742650]
18. Ubhi HK, Michie S, Kotz D, van Schayck OCP, Selladurai A, West R. Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features. Transl Behav Med 2016;6(3):410-417 [FREE Full text] [doi: 10.1007/s13142-015-0352-x] [Medline: 27528530]
19. Kotz D, Brown J, West R. Predictive validity of the Motivation To Stop Scale (MTSS): a single-item measure of motivation to stop smoking. Drug Alcohol Depend 2013;128(1-2):15-19 [FREE Full text] [doi: 10.1016/j.drugalcdep.2012.07.012] [Medline: 22943961]
20. Williams VF, Smith AA, Villanti AC, Rath JM, Hair EC, Cantrell J, et al. Validity of a subjective financial situation measure to assess socioeconomic status in US young adults. J Public Health Manag Pract 2017;23(5):487-495. [doi: 10.1097/PHH.0000000000000468] [Medline: 27798530]
21. Minami J, Ishimitsu T, Matsuoka H. Effects of smoking cessation on blood pressure and heart rate variability in habitual smokers. Hypertension 1999;33(1 Pt 2):586-590. [doi: 10.1161/01.hyp.33.1.586] [Medline: 9931170]
22. West R, Hajek P, Stead L, Stapleton J. Outcome criteria in smoking cessation trials: proposal for a common standard. Addiction 2005;100(3):299-303. [doi: 10.1111/j.1360-0443.2004.00995.x] [Medline: 15733243]

XSL•FO
RenderX

23. Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ 2004;328(7455):1519 [FREE Full text] [doi: 10.1136/bmj.38142.554479.AE] [Medline: 15213107]
24. Herbec A, Parker E, Ubhi H, Raupach T, West R. Decrease in resting heart rate measured using smartphone apps to verify abstinence from smoking: an exploratory study. Nicotine Tob Res 2020;22(8):1424-1427 [FREE Full text] [doi: 10.1093/ntr/ntaa021] [Medline: 31971595]
25. Beard E, Dienes Z, Muirhead C, West R. Using Bayes factors for testing hypotheses about intervention effectiveness in addictions research. Addiction 2016;111(12):2230-2247 [FREE Full text] [doi: 10.1111/add.13501] [Medline: 27347846]
26. Dienes Z. Using Bayes to get the most out of non-significant results. Front Psychol 2014;5:781 [FREE Full text] [doi: 10.3389/fpsyg.2014.00781] [Medline: 25120503]
27. Jeffreys H. The Theory of Probability. Oxford: Clarendon Press; 1961.
28. Schwendinger F, Grün B, Hornik K. A comparison of optimization solvers for log binomial regression including conic programming. Comput Stat 2021;36(3):1721-1754. [doi: 10.1007/s00180-021-01084-5]
29. Perski O, Jackson SE, Garnett C, West R, Brown J. Trends in and factors associated with the adoption of digital aids for smoking cessation and alcohol reduction: a population survey in England. Drug Alcohol Depend 2019;205:107653 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.107653] [Medline: 31675544]
30. West R. The clinical significance of "small" effects of smoking cessation treatments. Addiction 2007;102(4):506-509. [doi: 10.1111/j.1360-0443.2007.01750.x] [Medline: 17362283]
31. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
32. Benowitz NL, Bernert JT, Foulds J, Hecht S, Jacob P, Jarvis MJ, et al. Biochemical verification of tobacco use and abstinence: 2019 update. Nicotine Tob Res 2020;22(7):1086-1097 [FREE Full text] [doi: 10.1093/ntr/ntz132] [Medline: 31570931]
33. SRNT Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. Nicotine Tob Res 2002;4(2):149-159. [doi: 10.1080/14622200210123581] [Medline: 12028847]
34. Heffner JL, McClure JB. Commentary on Graham et al.: biochemical verification of abstinence in remotely conducted smoking cessation trials should not be a universal design requirement for rigor. Addiction 2022;117(4):1047-1048. [doi: 10.1111/add.15803]

Abbreviations

ITT: intention-to-treat analysis
ISRCTN: International Standard Randomised Controlled Trial Number
RCT: randomized controlled trial
RR: relative risk

Edited by A Schwartz, T de Azevedo Cardoso; submitted 18.07.23; peer-reviewed by S Toukhy, E Ross; comments to author 09.02.24; revised version received 26.02.24; accepted 29.08.24; published 15.11.24.

Please cite as:
Jackson S, Kale D, Beard E, Perski O, West R, Brown J
Effectiveness of the Offer of the Smoke Free Smartphone App Compared With No Intervention for Smoking Cessation: Pragmatic Randomized Controlled Trial
J Med Internet Res 2024;26:e50963
URL: https://www.jmir.org/2024/1/e50963
doi:10.2196/50963
PMID:

©Sarah Jackson, Dimitra Kale, Emma Beard, Olga Perski, Robert West, Jamie Brown. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Comment Slant and Comment Tone on Digital Health Communication Among Polarized Publics: A Web-Based Survey Experiment

Fangcao Lu[1*], PhD; Caixie Tu[2*], PhD

[1]Department of Applied Social Sciences, Faculty of Health and Social Sciences, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China (Hong Kong)

[2]School of Journalism and Communication, Shanghai University, Shanghai, China

[*]all authors contributed equally

Corresponding Author:
Caixie Tu, PhD
School of Journalism and Communication, Shanghai University
Shanghai University
149 Yanchang Road
Shanghai, 200072
China
Phone: 852 56207827
Email: cxtu@shu.edu.cn

Abstract

Background: Public attitudes toward health issues are becoming increasingly polarized, as seen in social media comments, which vary from supportive to oppositional and frequently include uncivil language. The combined effects of comment slant and comment tone on health behavior among a polarized public need further examination.

Objective: This study aims to examine how social media users' prior attitudes toward mask wearing and their exposure to a mask-wearing–promoting post, synchronized with polarized and hostile discussions, affect their compliance with mask wearing.

Methods: The study was a web-based survey experiment with participants recruited from Amazon Mechanical Turk. A total of 522 participants provided consent and completed the study. Participants were assigned to read a fictitious mask-wearing–promoting social media post with either civil anti–mask-wearing comments (130/522, 24.9%), civil pro–mask-wearing comments (129/522, 24.7%), uncivil anti–mask-wearing comments (131/522, 25.1%), or uncivil pro–mask-wearing comments (132/522, 25.3%). Following this, the participants were asked to complete self-assessed questionnaires. The PROCESS macro in SPSS (model 12; IBM Corp) was used to test the 3-way interaction effects between comment slant, comment tone, and prior attitudes on participants' presumed influence from the post and their behavioral intention to comply with mask-wearing.

Results: Anti–mask-wearing comments led social media users to presume less influence about others' acceptance of masks (B=1.49; $P<.001$; 95% CI 0.98-2.00) and resulted in decreased mask-wearing intention (B=0.07; $P=.03$; 95% CI 0.01-0.13). Comment tone with incivility also reduced compliance with mask-wearing (B=–0.44; $P=.02$; 95% CI –0.81 to –0.07). Furthermore, polarized attitudes had a direct impact (B=0.86; $P<.001$; 95% CI 0.45-1.26) and also interacted with both the slant and tone of comments, influencing mask-wearing intention.

Conclusions: Pro–mask-wearing comments enhanced presumed influence and compliance of mask-wearing, but incivility in the comments hindered this positive impact. Antimaskers showed increased compliance when they were unable to find civil support for their opinion in the social media environment. The findings suggest the need to correct and moderate uncivil language and misleading information in online comment sections while encouraging the posting of supportive and civil comments. In addition, information literacy programs are needed to prevent the public from being misled by polarized comments.

(J Med Internet Res 2024;26:e57967) doi:10.2196/57967

KEYWORDS

comments slant; incivility; social media; influence of presumed influence; health compliance; mask wearing; web survey

XSL•FO
RenderX

Introduction

Background

Escalating political polarization is increasingly reflected in public attitudes toward health issues, despite health traditionally being a nonpartisan, science-based sector [1]. This polarization is particularly noticeable in social media discussions, where health-related posts often elicit a spectrum of public responses in the comments section, ranging from support to opposition [2,3]. Alarmingly, these comments frequently include incivility [4], such as profanity, name-calling, or shouting [5]. Particularly, social media has become a crucial tool for health communication, allowing health institutions to initiate campaigns and individual users to disseminate these campaign messages [6]. The prevalence of polarization and incivility in the comments accompanying the ubiquitous health campaigns on social media necessitates an understanding of their impact on health-related compliance behavior. This understanding is crucial for guiding public health promotion and enhancing the effectiveness of digital health communication.

Previous research has indicated that individuals' exposure to opposing or uncivil comments in health promotion posts can independently reduce their compliance with the promoted health behaviors [7-9]. However, few studies have examined how the 2 attributes of comments interact and exert joint effects. Indeed, incivility might reduce the effects of comment slant as individuals may attribute low credibility to the commenters [10], thereby being less affected by them. The combined influence of comment slant and tone on health-related compliance behavior warrants an examination, as these 2 attributes of comments often occur together [11,12]. The findings contribute to a nuanced understanding of how social media users' interaction and active participation, specifically polarized and hostile online discourse, affect digital health practices.

Health campaigns typically influence compliance behavior indirectly, rather than directly. The influence of presumed influence (IPI) model provides a theoretical explanation for this indirect effect, suggesting that individuals' perceived media influence on others drives their compliance with promoted behaviors [13]. Even on social media, where cues (eg, view counts and comments) directly trigger normative influence to affect individuals [14], the IPI process remains significant for understanding their behavior [15]. Studies primarily suggest that individuals' perceptions of a message's influence on others are based on their assumptions about others' exposure to that message [16,17]. This other-consciousness perspective highlights the effects of comments as they represent others' responses to a post. Therefore, it is essential to explore how user comments with varying slants and tones can cultivate individuals' perceptions of the campaign message's influence on others and their compliance with health behaviors.

In a polarized environment, individuals often have strong prior attitudes. These polarized attitudes can also influence how they perceive the media's impact on others, an approach known as the self-centered perspective of the IPI process [15]. While studies have started to delve into the effects of the self-centered perspective on the IPI process, whether and how this perspective introduces changes in the current predominant other-consciousness perspective remains unexplored. Notably, individuals may react differently to incivility in comments, showing more tolerance for comments that align with their opinions [18,19]. It implies that an individual's polarized attitude can influence not only their perceived influence of the digital health campaign on others but also the effects of comments on their presumed influence of the campaign. By considering the separate and combined effects of the other-consciousness and self-centered perspectives, the dual nature of the IPI model offers a comprehensive understanding of the social psychological process through which people respond to mediated health communication in a polarized and hostile online environment.

This study was conducted in September 2020, during the early stage of the COVID-19 pandemic in the United States, when wearing masks to combat COVID-19 was controversial due to the polarized political environment [20]. Although numerous posts on social media advocated mask-wearing as an effective measure against the virus's spread, comments on these posts were predominantly polarized [21], and about 1 in 5 of these comments exhibited incivility [22]. Such opinion climates have been related to the public's noncompliance with COVID-19 mitigation guidelines [23], which leads to heightened virus transmission and COVID-19–related deaths. This situation provides an appropriate context to examine the impact of digital health promotion in the polarized and hostile digital space. Given that, we conducted a between-subjects experiment with a 2 (comment slant: pro–mask-wearing vs anti–mask-wearing) × 2 (comment tone: civil vs uncivil) design by manipulating comments accompanying a social media post for mask-wearing. Participants' prior attitude was included as a moderator. Given the proliferation of digital health campaigns and the increasing polarized and hostile opinion climates, public health practitioners can benefit from the findings to boost the effectiveness of digital health communication.

Presumed Influence and Health Campaigns

The IPI model comprises 3 components: presumed exposure, presumed influence, and the IPI [13]. *Presumed exposure* refers to individuals' exposure to media content serving as a foundation for their inference about others' exposure to the same content. *Presumed influence* indicates that, in turn, individuals' presumption of others' exposure triggers their presumption that the media content will influence those others. Finally, *IPI* refers to individuals' alignment of their reactions to the presumed influence on others.

To date, discussions on IPI tend to focus on how people accommodate or rectify media messages. Rectifying behavior refers to individuals taking actions to protect others from harmful media effects or to magnify desirable media effects on others [24]. Accommodation reactions are more widely studied in the context of health campaigns, where individuals adapt themselves to the social environment [24]. To assess group or social norms, people often form perceptions about media influence on others and draw conclusions based on these perceived influences [25,26]. The more individuals believe that

XSL•FO
RenderX

others adopt a particular behavior, the more likely they are to think that the behavior is normative. A desire to fit in with the group or social pressure then motivates them to adopt the same behavior [27,28]. The IPI model has been extensively tested in the context of health communication. The compliance behavior has been examined in the context of condom use, healthy diet, regular exercise, antismoking, excessive drinking, e-cigarette use, and COVID-19 pandemic protective behavior [25-28].

Other-Consciousness Perspective: Comments and Presumed Influence

Overview

Social media has served as an integral arena for organizations and individuals to share health information [6]. The commentary feature provided by social media transforms audiences from passive information receivers to active users who interact with these health messages. Comments on a public health campaign message often reflect commenters' support or opposition to the message, indicating their slants. Comments can express approval of the campaign by presenting supportive views or can be disapproving by presenting challenging views [29,30]. The slants of comments accompanying a message are likely to affect people's presumptions of the message's influence on others [7]. This effect can be explained by the exemplification theory. According to the theory, exemplars refer to the opinions or experiences of a person involved in an issue [31]. Exemplars are concrete and easy to process and remember. Thus, people tend to form judgments and beliefs about an issue based on available exemplars. Comments below a message serve as vivid exemplars of the audience's opinions on the message. When gauging the influence of a social media message on others, a person may perceive comments below the message as representations of the entire audience's reaction to the message [32,33].

Previous studies have found that the slant of comments accompanying social media health campaigns affects individuals' perceptions of the campaigns' influence on others. When individuals were exposed to supportive comments below a Facebook post promoting COVID-19 vaccination, they perceived a greater influence of the post on others' acceptance of COVID-19 vaccination than when exposed to disapproving comments about the post [7]. Similarly, when social media users encounter pro–mask-wearing comments rather than anti–mask-wearing comments below a mask-promoting post, they are likely to perceive more influence of the post on other users' acceptance of mask-wearing. The perceived influence of such a media message on others may further lead the users to comply with the behavior promoted by the message [34]. Accordingly, the following hypotheses are proposed:

- Hypothesis 1a: social media users will have weaker intentions to wear masks when exposed to anti–mask-wearing comments below a mask-promoting post than when exposed to pro–mask-wearing comments.
- Hypothesis 1b: the association between comment slants and intentions to wear masks will be mediated by social media users' perception of the influence of the mask-promoting post on others.

Although the commentary feature facilitates social media users' expressions of personal opinions, comments are often loaded with incivility. Comments are considered to contain incivility when expressed in an impolite and disrespectful tone [5]. Uncivil comments associated with a message can induce a "nasty effect," a belief that if comments below a message contain incivility, the message must be bad [35]. Readers tend to believe that the original post, juxtaposed with the uncivil comments, is biased, of low quality, uncivil, and from a noncredible source [35-38].

Research on the "nasty effect" has also extended the spillover effects of comments' incivility to audiences' perception of a media message's influence on others. Waddell and Bailey [39] found a belief in audiences' minds that "if others' comments are uncivil then they must not have been affected by the content." Uncivil comments reveal conflicts among people with different opinions on an issue, rather than their elaboration and information processing of the issue discussed in the main message. When exposed to uncivil comments rather than civil ones left on a media message, people tend to believe that others reinforce their prior views rather than reading, deliberating, and being influenced by the adjacent media message. Accordingly, social media users exposed to uncivil comments on a mask-promoting post are expected to presume that the post exerts less influence on others' acceptance of mask-wearing than when exposed to civil comments. The perception of less influence of the post on others, in turn, reduces social media users' behavioral intention to wear masks. We thus propose the following hypotheses.

- Hypothesis 2a: social media users will have weaker intentions to wear masks when exposed to uncivil comments below a mask-promoting post than when exposed to civil comments.
- Hypothesis 2b: the association between comment tone and intentions to wear masks will be mediated by social media users' perception of the influence of the mask-promoting post on others.

Self-Centered Perspective: Polarized Attitudes and Presumed Influence

Individuals' perceptions of a health campaign's influence on others may be affected by their prior attitudes toward the campaign's advocacy. This effect can be explained by the "looking-glass perception," which suggests that people's social perceptions are often self-centric, and people tend to use their own opinions to estimate those of others [40,41]. They believe that situational factors are similar between themselves and others. Therefore, they tend to amplify their prior attitudes to their perceived social consensus on related issues [42].

Previous studies have provided support for the idea that presumed influence may be self-centric. For example, the robust causal chain from self-exposure, presumed exposure, to presumed influence was found to result from the order of questions. When the order of questions (self-variable → other variable → presumed influence on others → behavior) was reversed (other variable → self-variable → presumed influence on the self → behavior), the causal chain conflicted with the IPI process [43]. The finding suggests that the self may serve as an anchor for projecting presumed influence on others.

XSL•FO
RenderX

Another study found that the more individuals relate themselves to the message and consider it real, the greater they perceive the message to elicit an influence on its audiences [15].

Extrapolating from this self-centered perspective, individuals' prior attitudes toward a health campaign's advocacy may predict their estimation of the campaign's influence on others. People tend to accept information that is consistent with their prior beliefs [44]. When individuals encounter a health message consistent with their prior attitudes, they are more willing to acknowledge that they are influenced by the message and accept its view. In contrast, individuals are more likely to reject the message when they have inconsistent attitudes toward it [45-47]. Accordingly, individuals with favorable attitudes toward mask-wearing are likely to perceive that others, like themselves, also agree with the mask-promoting message and will be influenced by it. Individuals with unfavorable attitudes toward mask-wearing are likely to believe that others, similar to themselves, reject the message and are immune to it. The perception that the mask-promoting post has affected others, in turn, influences individuals' behavioral intention to wear masks.

In addition to the partial mediating role of presumed influence, the positive association between attitudes and behavioral intentions has been sufficiently addressed [48]. Attitudes toward a health behavior can inspire individuals' intention to perform the behavior. Thus, the following 2 hypotheses are proposed:

- Hypothesis 3a: social media users will have weaker intentions to wear masks when they have unfavorable attitudes toward mask-wearing than when they have favorable attitudes.
- Hypothesis 3b: the association between prior attitudes toward mask-wearing and intentions to wear masks will be partially mediated by social media users' perception of the influence of the mask-promoting post on others.

The Interaction of Social Media Comments and Polarized Attitudes

The slant and tone of social media comments below a mask-promoting message are likely to interact and affect social media users' presumption of the message's influence and subsequently their health compliance. The content of the message provides important cues that help users form impressions of the senders. Previous studies reveal that encountering uncivil comments under a news article led to negative perceptions and less perceived credibility of the commenters [10]. While source credibility has long been recognized as a key factor in persuasiveness [49], a lack of credibility among commenters may cause uncivil comments to signal that the message has less influence on others. The reduced presumed influence, in turn, is less likely to drive behavioral change.

In other words, social media users' exposure to civil pro–mask-wearing comments on a mask-wearing post facilitates their perception that the post poses an influence on others' acceptance of mask-wearing and stimulates their compliance with mask-wearing. In contrast, exposure to uncivil pro–mask-wearing comments on a post is likely to decrease this presumed influence of the post and their intentions to comply with mask-wearing. In addition, social media users' exposure to civil anti–mask-wearing comments on a mask-wearing post can reduce their presumed influence of the post on others' acceptance of mask-wearing and their compliance with mask-wearing. Conversely, uncivil anti–mask-wearing comments can offset these negative effects to some extent by maintaining the social media users' presumed influence and behavioral intention of mask-wearing. The following 2 hypotheses are therefore proposed:

- Hypothesis 4a: comment tone will moderate the effect of comment slant on social media users' intentions to wear masks, such that the effect of comment slant on behavioral intention will be stronger when comments are expressed in a civil manner compared to in an uncivil manner.
- Hypothesis 4b: the interaction effect of comment slant and comment tone on social media users' intentions to wear masks will be mediated by their perception of the influence of the mask-promoting post on others.

It is also likely that there is an interaction between comment slant, comment tone, and prior attitudes. Individuals may have more tolerance for comments that align with their prior attitudes, and they may overlook the incivility and aggressiveness in the comments [19]. This can be explained by the Social Identity Theory [50], which posits that individuals categorize themselves and others into in groups and out groups based on shared characteristics or beliefs, and this categorization influences their attitudes and behaviors. Individuals may categorize comments into in-group comments (those that align with their prior attitudes) and out-group comments (those that contradict their prior attitudes). They are more likely to favor in-group comments and perceive them in a positive light (ie, less uncivil) than out-group comments, as these comments reinforce their social identity and validate their prior attitudes.

Experimental studies have suggested that individuals would rate a comment that supported their prior attitudes as civil, even though it contained incivility. However, they still recognized the incivility in comments that were against their prior attitudes [18]. Thus, comment tone may only function or produce a relatively greater effect on presumed influence and health-related compliance behavior when social media users' prior attitudes are inconsistent with comment slant. In contrast, when social media users' prior attitudes are consistent with comment slant, they may ignore the incivility in these comments but perceive it as civil. Therefore, the impact of comment tone on presumed influence and behavioral intention would be discounted or become nonsignificant. We propose the 2 hypotheses below:

- Hypothesis 5a: there is an interaction among comment slant, comment tone, and prior attitudes on social media users' intentions to wear masks, such that the influence of incivility will affect the influence of comments that reveal a slant inconsistent with social media user' prior attitudes on their behavioral intention to wear masks, but it will not affect the influence of comments that reveal a slant consistent with their preexisting attitudes.
- Hypothesis 5b: the impact of the interaction of comment slant, comment tone, and prior attitudes on social media users' intentions to wear masks will be mediated by their

XSL•FO
RenderX

perception of the influence of the mask-promoting post on others.

In summary, this study aims to investigate how social media users' polarized attitudes toward mask-wearing and their exposure to a mask-promoting post synchronized with user comments, independently or collectively, affect their compliance with mask-wearing.

Methods

Experimental Design

The study used a web-based between-subjects survey experiment with a 2 (comment slant: pro–mask-wearing vs anti–mask-wearing) × 2 (comment tone: civil vs uncivil) design. Participants were recruited from Amazon Mechanical Turk (MTurk), a crowdsourcing platform that allows individuals to outsource tasks, including web-based experiment participation, to registered workers [51]. There has been a long methodological discussion about the data quality obtained from MTurk. While some studies criticize the quality of data obtained from this platform, some indicate that MTurk is a feasible platform for online data collection, especially when strict criteria are applied [52,53]. Therefore, to ensure the quality of our data, we established specific criteria for participant selection (ie, the number of the participants' approved assignments was >5000, the participants' approval rating was >95%, and the participants were in the United States). In addition, we incorporated 2 attention checks (ie, select a specific word from the given options). The participation was immediately terminated when participants failed to pass attention checks.

Ethical Considerations

Participants were recruited from September 29 to October 1, 2020. The study was reviewed and approved by the Human Subjects Ethics Subcommittee of the City University of Hong Kong (2020-55359071) before data collection. In the recruitment announcement posted on MTurk, we informed participants that (1) this study examined their knowledge of and attitudes toward mask-wearing; (2) the participation was fully anonymous, and their self-reported data would be kept confidential; and (3) they could leave the study any time if they wanted. After each participant clicked to agree to a written consent form, which again highlighted these ethical considerations, they continued to participate in the survey. Informed consent was obtained from all participants.

Stimuli

A mask-promoting post was created and embedded in a fictitious health organization's Facebook page, as Facebook is widely used by health organizations to promote health initiatives. The post was created based on the guidelines about mask-wearing posted on the official website of the Centers for Disease Control and Prevention in the United States to ensure external validity [54,55]. It was developed using the standard format of fear appeal commonly used in health communication campaigns. To prevent the post from being perceived as an unintended threat to individuals' freedom, which could undermine the campaign's effectiveness, we framed it as a low-threat fear appeal by using mild and polite language to recommend mask-wearing [56]. The content and layouts of the post were kept identical across all conditions.

Prior research indicates that exposure to >4 comments does not increase the effect of comment tone [57]. Therefore, we encapsulated 4 comments below the post for each condition. Comment slant was initially created based on actual Facebook users' expressions on mask-wearing. Across the 2 conditions of comment slant, we matched 2 comments, 1 in each condition, that focused on the same aspects of mask-wearing but expressed opposite opinions and also maintained similar levels of lengths, expression style, and argument strength of the comments. We repeated this procedure for the other comments. This allowed us to generate civil pro–mask-wearing and anti–mask-wearing comments without incurring confounding factors.

Comment tone was manipulated by following the definition of incivility by Coe et al [5]. We added incivility to the previously created comments to derive uncivil pro–mask-wearing and anti–mask-wearing comments. The post and examples of comments used as stimuli are presented in Multimedia Appendix 1.

Experimental Procedure

The experiment was conducted using the web-based survey software Qualtrics. Before fielding the questionnaire, the survey, including the stimuli and measures, was proofread by 3 native speakers to ensure readability and validity. The technical functionality of the survey platforms and settings was tested by 5 student assistants. The number of items per page and the total pages of the questionnaire distributed were adjusted by Qualtrics based on the devices each participant used, thereby resulting in variations among participants.

All eligible participants could access the survey link posted on the MTurk assignment page. After providing consent for participation, participants were first asked to report their prior attitudes toward mask-wearing, social media use frequency, and mask-wearing practices. The randomizer of Qualtrics enabled us to randomly assign each participant to 1 of the 4 experimental conditions. After being exposed to the stimuli, participants were asked to indicate their responses to the variables of interest, provide demographic information, and answer manipulation check questions. Participants were allowed to review and change their answers using a "back" button at any time before submitting their responses.

The question regarding participants' prior attitudes toward masks served as a screening item. Participants were asked to rate the extent to which wearing a mask in public during the COVID-19 pandemic was favorable or unfavorable on a 7-point scale (1=very unfavorable, 4=neither unfavorable nor favorable, and 7=very favorable). Participants were categorized as antimaskers (ie, scores <4) and promaskers (ie, scores >4). As this study focused on the effects of polarized attitudes on presumed influence and compliance behavior, participants with neutral attitudes (ie, scores=4) were directed to the end of the survey.

Participants

A total of 1501 participants provided consent and started the survey, with 522 (34.78%) participants completing all the

XSL•FO
RenderX

questions and being included in the final analysis. The view rate was 84.01% (1503/1789), the participation rate was 99.53% (1496/1503), and the completion rate was 34.78% (522/1501). Upon completion of the study, each participant received a debriefing and an incentive of US $0.72. As each worker on MTurk has a unique ID, a unique visitor is defined by the unique ID, rather than cookies used. We also checked IP addresses to ensure that each participant was a unique site visitor. The survey, as an MTurk task, was displayed only once to each participant to avoid repeated registrations. The CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) form [58] and the Checklist for Reporting Results of Internet e-Surveys (CHERRIES) form [59] are presented in Multimedia Appendices 2 and 3, respectively, for further clarity.

Participants in the final sample were aged 21 to 77 (mean 41.58, SD 12.35) years. More than half of them were men (291/522, 55.7%). Of the 522 participants, 246 (47.1%) had completed college as their highest level of education, and 416 (79.7%) identified themselves as White. Most participants (317/522, 60.8%) reported that their annual family income ranged from US$20,000 to $74,999. In terms of political identification, 41.5% (217/522) of the participants identified themselves as Democrats, followed by 40.2% (210/522) as Republicans and 18.2% (95/522) as neither Republicans nor Democrats.

Measures

The measure of presumed influence was adapted from a previous study [60]. Participants were asked to indicate the extent to which they agreed that the social media post of mask promoting had made other people support mask-wearing in public during the COVID-19 pandemic, using a 7-point scale (1=strongly disagree and 7=strongly agree; mean 4.43, SD 2.00). They were also asked to evaluate whether the post had negatively or positively affected others' attitudes toward mask-wearing, using a 7-point scale (1=in a very negative manner and 7=in a very positive manner; mean 4.20, SD 1.93). These 2 items were highly correlated and were averaged to form the measure of presumed influence (r=0.77; P<.001; mean 4.32, SD 1.85).

We measured participants' behavioral intention to wear masks as compliance with health campaigns by adapting the measure used by Dillard and Shen [56]. Participants were asked to estimate the likelihood that they would wear a mask in public in the next week using a 7-point scale, ranging from 1=definitely will not to 7=definitely will (mean 5.61, SD 1.90).

Before being exposed to the experimental stimuli, participants were asked to report their social media use frequency and mask-wearing practices in the last week. Responses to the 2 questions were rated on a 5-point scale, where 1 meant never and 5 meant nearly always (social media use frequency: mean 3.79, SD 0.93 and mask-wearing practices: mean 3.29, SD 1.04).

Preliminary Statistical Analyses

For randomization checks, a series of 1-way ANOVA were conducted to test the differences in continuous variables, and several chi-square analyses were conducted to test the differences in categorical variables across conditions. In addition, participants were categorized into 2 groups (ie, antimaskers and promaskers) based on the screening question. The 1-way ANOVA and chi-square analyses were repeated to test the differences in demographic variables between antimaskers and promaskers.

Statistical Analyses for Manipulation Checks

After exposure to experimental materials, participants were asked to report whether they had read the comments below the post. Next, participants were asked to indicate the extent to which they thought the comments were favorable to the post using a 7-point scale (1=very unfavorable and 7=very favorable). We used an independent-samples 2-tailed t test to check the difference in perceived slant of comment between participants in the pro–mask-wearing comments condition and those in the anti–mask-wearing comments condition. Descriptive information (ie, mean and SD), t value, dfs, and P value were reported to illustrate the difference. Then, 1-sample t tests were conducted to indicate whether participants' perceived slant of comment significantly deviated from the midpoint of the scale (ie, 4). We reported t value, dfs, and P value to indicate the difference.

Furthermore, participants were asked to rate the degree of comment incivility using a 7-point scale (1=very uncivil and 7=very civil). We used independent-samples t tests to check the difference in perceived civility of comments between participants in the civil comments condition and those in the uncivil comments condition. Descriptive information (ie, mean and SD), t value, dfs, and P value were reported to illustrate the difference. Then, 1-sample t tests were conducted to indicate whether participants' perceived slants of comment significantly deviated from the midpoint of the scale (ie, 4). We reported t value, dfs, and P value to indicate the difference.

Statistical Analyses for Hypotheses Testing

To test the proposed hypotheses concurrently, we used the PROCESS macro (model 12). The PROCESS macro is a regression path analysis modeling tool used to conduct mediation, moderation, and conditional process analysis; it is widely applied in the fields of social, business, and health sciences [61]. Its model 12 tests moderated mediation models. In this study, behavioral intention to wear masks was included as the dependent variable. Prior attitude (0=anti–mask-wearing and 1=pro–mask-wearing) was entered as the independent variable, and comment slant (0=anti–mask-wearing and 1=pro–mask-wearing) and comment tone (0=uncivil and 1=civil) were included as moderators. Participants' demographics (ie, age, gender, education, income, race, and political identification), mask-wearing frequency, and social media use frequency were included as covariates. Missing values were replaced by mean scores. We reported the unstandardized coefficient (B), unstandardized SE, P value, and 95% CI, which indicate the effects of participants' prior attitudes, comments slant, comment tone, and presumed influence on their intention to wear masks. In addition, the effect size, SE, and 95% CI were reported to show the conditional direct and indirect effects of comment slant, comment tone, and prior attitudes on behavioral intention.

XSL•FO
RenderX

Statistical Analyses for Sensitivity Analysis

Two sensitivity analyses were conducted. First, we calculated attitude extremity by subtracting 4 from the value chosen by promaskers in the screening question and by subtracting the value chosen by antimaskers from 4 (ie, 1=low extremity, 2=medium extremity, and 3=high extremity). We controlled for this variable in sensitivity analysis 1. Second, we added the variables stepwise to the regression models—main effects first and then the interaction terms—to better demonstrate the main effects in sensitivity analysis 2.

Results

Preliminary Analyses

A CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the participants is presented in Figure 1. The demographic information across groups is presented in Multimedia Appendix 4. A series of 1-way ANOVAs indicated that there were no significant differences in participants' age (P=.77), education (P=.37), and annual income (P=.54) across conditions. Chi-square analyses also showed no significant differences in participants' gender (P=.42), race (P=.97), and political identification (P=.21) across conditions. In addition, among 522 participants, 269 (51.5%) had unfavorable attitudes toward mask-wearing (ie, antimaskers), whereas 253 (48.5%) held favorable attitudes toward mask-wearing (ie, promaskers). No significant differences in age (P=.91), gender (P=.91), and annual income (P=.51) were found between antimaskers and promaskers. However, promaskers (mean 5.62, SD 1.04) reported higher levels of education than antimaskers (mean 5.32, SD 1.19; $t_{517.16}$=3.05; P=.002). Therefore, basic demographic factors were controlled in later analysis to adjust for the differences in the sample.

Figure 1. Participation flowchart.

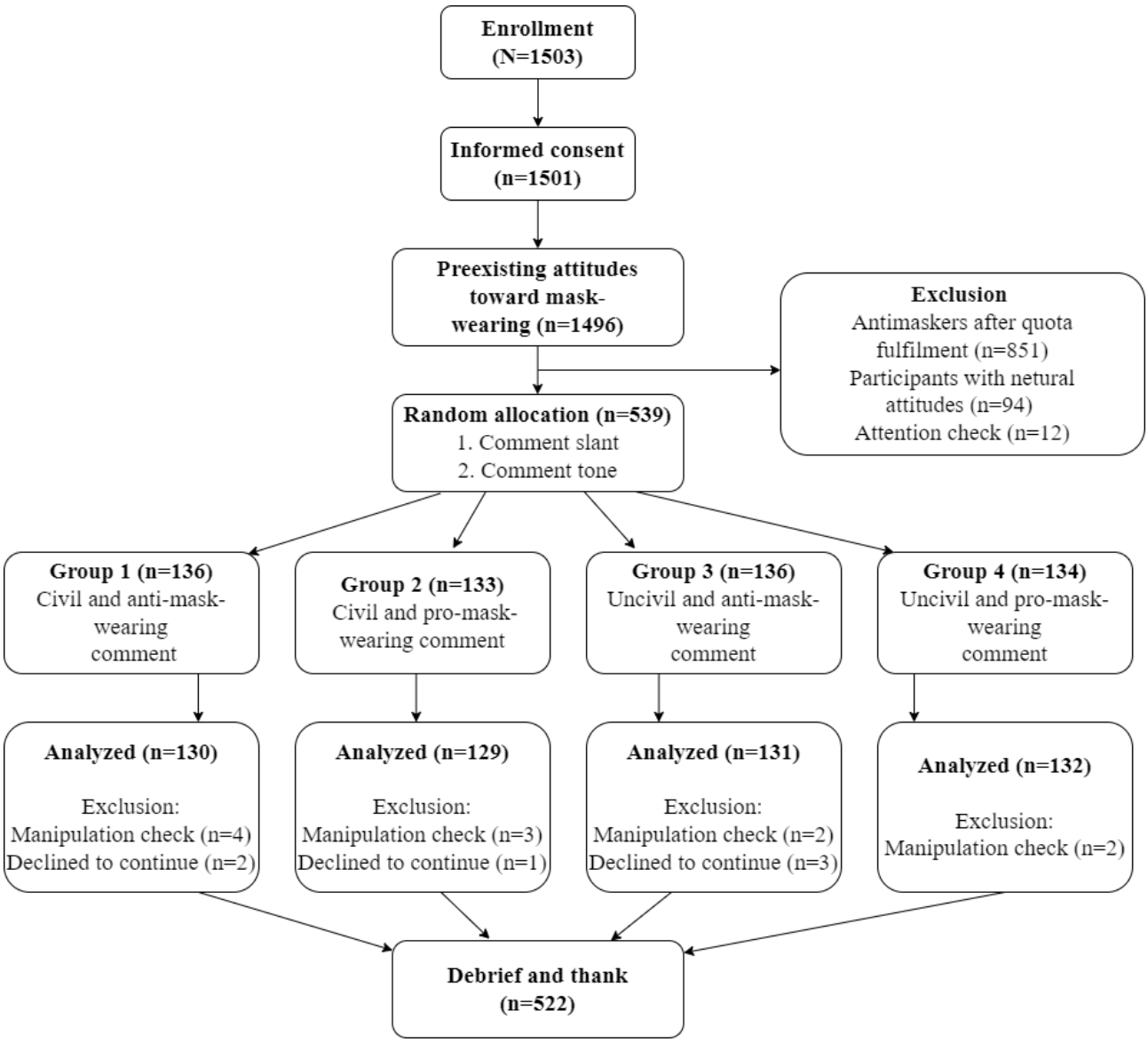

XSL•FO
RenderX

Manipulation Checks

Those who reported not reading comments were excluded (11/539, 2%). Participants in the pro–mask-wearing comments condition considered the comments to be more favorable to the post (mean 5.50, SD 1.85) than those in the anti–mask-wearing comments condition (mean 1.70, SD 1.51; $t_{499.56}$=25.75; P<.001). One-sample t tests indicated that both participants exposed to pro–mask-wearing comments (t_{260}=13.10; P<.001) or anti–mask-wearing comments (t_{260}=24.68; P<.001) perceived the comments to significantly deviate from the midpoint of the scale (ie, 4). Next, participants in the civil comments condition considered the comments more civil (mean 4.28, SD 1.88) than those in the uncivil condition (mean 1.92, SD 1.56; $t_{500.29}$=15.53; P<.001). One-sample t tests showed that both participants exposed to civil comments (t_{258}=2.38; P=.02) and uncivil comments (t_{262}=21.53; P<.001) perceived the comment tone to significantly deviate from the midpoint of 4.

Hypotheses Testing

The results of hypotheses testing for the separate and combined effects of comment slant, comment tone, and prior attitudes on presumed influence and mask-wearing intention are reported in Table 1.

Table 1. The influence of comment slant, comment tone, and prior attitudes on presumed influence and behavioral intention to wear masks[a].

	Presumed influence		Mask-wearing intention	
	B (SE; 95% CI)	*P* value	B (SE; 95% CI)	*P* value
Age	–0.00 (0.01; –0.01 to 0.01)	.81	0.01 (0.00; –0.00 to 0.01)	.19
Gender	0.03 (0.13; –0.23 to 0.30)	.81	0.16 (0.10; –0.03 to 0.35)	.09
Education	0.13 (0.06; 0.01 to 0.25)	.04	–0.09 (0.05; –0.18 to –0.01)	.04
Income	–0.11 (0.04; –0.18 to –0.04)	.003	0.02 (0.03; –0.03 to 0.08)	.40
Race	–0.09 (0.17; –0.42 to 0.24)	.59	0.20 (0.12; –0.04 to 0.44)	.10
Republican	0.15 (0.19; –0.23 to 0.53)	.43	0.02 (0.14; –0.26 to 0.29)	.91
Democrat	0.52 (0.20; 0.14 to 0.90)	.008	0.03 (0.14; –0.25 to 0.31)	.82
Mask-wearing frequency	0.22 (0.06; 0.09 to 0.35)	.001	1.08 (0.05; 0.99 to 1.71)	<.001
Social media use frequency	0.13 (0.07; –0.02 to –0.27)	.08	0.01 (0.05; –0.09 to 0.11)	.87
Prior attitude	0.19 (0.28; –0.37 to 0.74)	.51	0.86 (0.20; 0.45 to 1.26)	<.001
Comment slant	1.49 (0.26; 0.98 to 2.00)	<.001	–0.06 (0.19; –0.44 to 0.32)	.74
Comment tone	0.63 (0.26; 0.12 to 1.14)	.02	–0.44 (0.19; –0.81 to –0.07)	.02
Prior attitude × comment slant	0.27 (0.37; –0.46 to 1.00)	.47	–0.15 (0.27; –0.68 to 0.38)	.58
Prior attitude × comment tone	–0.26 (0.37; –1.00 to 0.47)	.48	0.34 (0.27; –0.19 to 0.87)	.21
Comment slant × comment tone	0.09 (0.36; –0.62 to 0.81)	.80	0.79 (0.26; 0.28 to 1.31)	.003
Prior attitude × comment slant × comment tone	0.55 (0.53; –0.48 to 1.59)	.30	–0.84 (0.38; –1.59 to –0.09)	.03
Presumed influence	__[b]	—	0.07 (0.03; 0.01 to 0.13)	.03

[a]Model summary: presumed influence, $F_{16,505}$=18.99; mask-wearing intention, $F_{17,504}$=67.01.

[b]Not applicable (at this stage, the presumed influence is the dependent variable).

The results of bootstrapping for the conditional direct and indirect effects of comment slant, comment tone, and prior attitudes on behavioral intention to wear masks are summarized in Table 2.

XSL•FO
RenderX

Table 2. The conditional direct and indirect effects of comment slant, comment tone, and prior attitudes on behavioral intention.

	Indirect effects (mediator: presumed influence), effect size (SE; 95% CI)	Direct effects, effect size (SE ; 95% CI)
Comment slant		
Uncivil × antimaskers	0.10 (0.05; 0.01 to 0.21)	–0.06 (0.19; –0.44 to 0.32)
Uncivil × promaskers	0.12 (0.06; 0.01 to 0.25)	–0.21 (0.20; –0.60 to 0.18)
Civil × antimaskers	0.11 (0.06; 0.01 to 0.23)	0.73 (0.19; 0.35 to 1.10)
Civil × promaskers	0.16 (0.08; 0.01 to 0.33)	–0.26 (0.21; –0.67 to 0.15)
Comment tone		
Anti–mask-wearing × antimaskers	0.04 (0.03; –0.00 to 0.11)	–0.44 (0.19; –0.81 to –0.07)
Anti–mask-wearing × promaskers	0.03 (0.03; –0.02 to 0.09)	–0.10 (0.20; –0.48 to 0.28)
Pro–mask-wearing × antimaskers	0.05 (0.03; 0.001 to 0.120)	0.35 (0.19; –0.02 to 0.72)
Pro–mask-wearing × promaskers	0.07 (0.04; 0.003 to 0.149)	–0.15 (0.19; –0.53 to 0.23)
Prior attitudes		
Anti–mask-wearing × uncivil	0.01 (0.03; –0.04 to 0.07)	0.86 (0.20; 0.45 to 1.26)
Anti–mask-wearing × civil	–0.01 (0.02; –0.05 to 0.04)	1.20 (0.20; 0.79 to 1.60)
Pro–mask-wearing × uncivil	0.03 (0.03; –0.01 to 0.09)	0.71 (0.20; 0.32 to 1.10)
Pro–mask-wearing × civil	0.05 (0.03; 0.00 to 0.11)	0.21 (0.20; –0.19 to 0.60)

As for hypotheses 1a and 1b, the regression results in Table 1 showed that there was no significant association between comment slant and behavioral intention (B=–0.06; *P*=.74). Hence, hypothesis 1a was not supported. Nevertheless, we found that compared with anti–mask-wearing comments, pro–mask-wearing comments were found to increase presumed influence (B=1.49; *P*<.001), and this presumed influence was positively associated with participants' behavioral intention to wear masks (B=0.07; *P*=.03). The bootstrapping results showed that the direct effect of comment slant on behavioral intention was significant only among antimaskers who read civil comments (B=0.73, SE 0.79; 95% CI 0.35-1.10). Comment slant posed an indirect influence on behavioral intention through the mediation of presumed influence, regardless of participants' prior attitudes or comment tone (Table 2). Hence, hypothesis 1b was supported by the data.

Next, for hypotheses 2a and 2b, results in Table 1 showed that there was a significant but negative association between comment tone and behavioral intention to wear masks (B=–0.44; *P*=.02). Hence, hypothesis 2a was not supported. As for hypothesis 2b, the effect of comment tone on presumed influence was positively significant (B=0.63; *P*=.02), and the association between presumed influence and intention to wear masks was also positively significant (B=0.07; *P*=.03). The bootstrapping results showed that comment tone posed a direct influence on behavioral intention to wear masks only when antimaskers encountered anti–mask-wearing comments (B=–0.44, SE 0.19; 95% CI –0.81 to –0.07). In addition, comment tone posed an influence on behavioral intention via the mediating effects of presumed influence when the comments were pro–mask-wearing, regardless of participants' prior attitudes (Table 2). Hence, hypothesis 2b was partially supported.

As for hypotheses 3a and 3b, the results in Table 1 showed that the direct effect of prior attitudes on behavioral intention to wear masks was significant (B=0.86; *P*<.001). Hence, hypothesis 3a received support. However, the effect of prior attitudes on presumed influence was not significant (B=0.19; *P*=.51). The bootstrapping results (Table 2) indicated that as long as the comments were uncivil or anti–mask-wearing, participants' prior attitudes were directly associated with their behavioral intention. Only when comments were pro–mask-wearing and civil, prior attitudes affected behavioral intention through presumed influence. Hence, we mostly could not corroborate hypothesis 3b.

Regarding hypothesis 4a, Table 1 shows that the interaction had a significant and direct effect on behavioral intention (B=0.79; *P*=.003). As shown in Figure 2, when expressed in a civil way, opposing comments (mean 5.46, SD 0.10) decreased the mask-wearing intention than supporting comments (mean 5.70, SD 0.10); while when expressed in an uncivil way, the effects of comments slant was reversed such that supporting comments (mean 5.59, SD 0.10) decreased the mask-wearing intention, compared with opposing comments (mean 5.73, SD 0.10). Hence, hypothesis 4a was supported. For hypothesis 4b, the results showed that the interaction between comment slant and comment tone did not significantly predict presumed influence (B=0.09; *P*=.80). Hence, hypothesis 4b was not supported.

XSL•FO
RenderX

Figure 2. The interaction effect of comment slant and comment tone on the behavioral intention to wear masks.

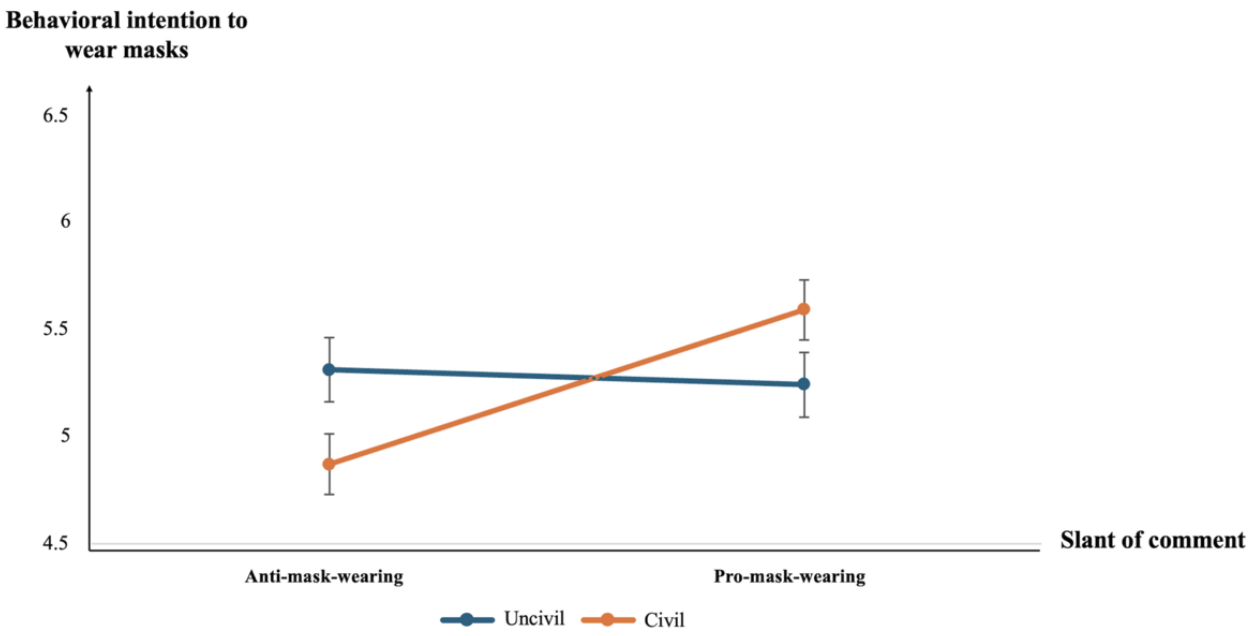

For hypotheses 5a and 5b, the results (Table 1) showed that the interaction was significant for behavioral intention (B=–0.84; *P*=.03) but not for presumed influence (B=0.55; *P*=.30). However, the effect of the interaction on behavioral intention was different from what we expected. As shown in Figure 3, for promaskers, behavioral intention to wear masks remained similar when they saw uncivil comments or civil comments, regardless of whether the comments were anti–mask-wearing (civil: mean 6.06, SD 0.15 and uncivil: mean 6.16, SD 0.14; *P*=.61) or pro–mask-wearing (civil: mean 5.80, SD 0.15 and uncivil: mean 5.95, SD 0.14; *P*=.45). In contrast, among antimaskers, their behavioral intention remained similar when they viewed uncivil pro–mask-wearing comments (mean 5.24, SD 0.13) and civil pro–mask-wearing comments (mean 5.59, SD 0.14; *P*=.06). Nevertheless, their behavioral intention was stronger when they read uncivil anti–mask-wearing comments (mean 5.31, SD 0.15) compared to civil anti–mask-wearing comments (mean 4.87, SD 0.14; *P*=.02). Hence, hypothesis 5a was partially supported, but hypothesis 5b was not supported.

Figure 3. The interaction effect of comment slant, comment tone, and prior attitudes on the behavioral intention to wear masks.

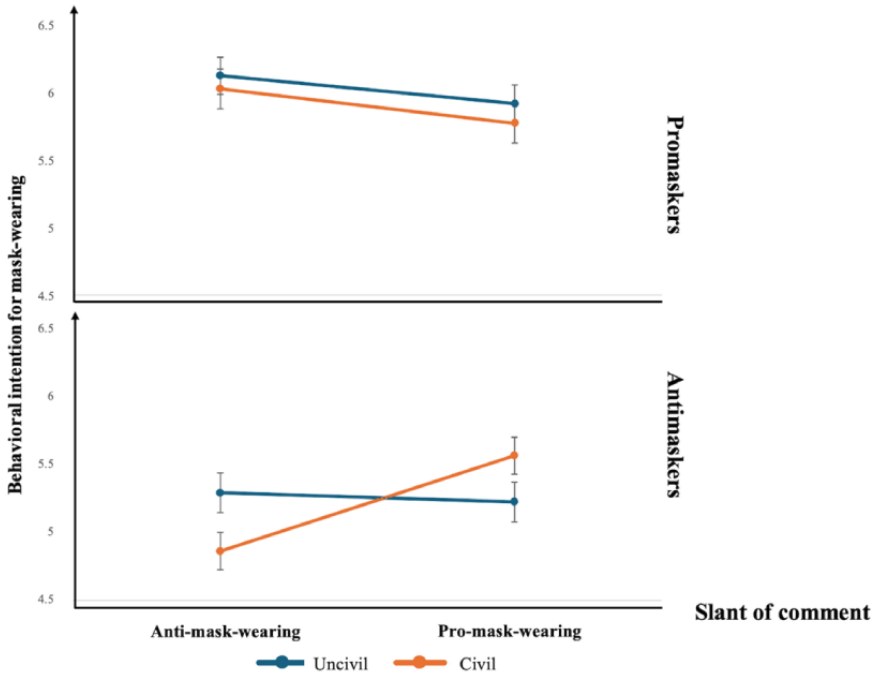

XSL•FO
RenderX

Sensitivity Analyses

Results from the sensitivity analyses, which included attitude extremity as an additional control variable and involved applying stepwise multiple linear regression (Multimedia Appendices 5-8) were consistent with the main results.

Discussion

Principal Findings

This study investigated how polarized and hostile user comments below a health campaign message on social media and social media users' polarized attitudes concurrently affected their perception of the campaign's influence on others and their compliance with the promoted health behavior. Results showed that compared with anti–mask-wearing comments, pro–mask-wearing comments enhanced presumed influence and health compliance of mask-wearing, but incivility in the comments hindered the positive impact of pro–mask-wearing comments. Antimaskers demonstrated increased compliance when they were unable to find civil support for their opinion in the social media environment. The summary of the research hypotheses and corresponding results are presented in Table 3.

Table 3. Summary of hypotheses and results.

Hypotheses	Results
Hypothesis 1a: social media users will have weaker intentions to wear masks when exposed to anti–mask-wearing comments below a mask-promoting post than when exposed to pro–mask-wearing comments.	Not supported
Hypothesis 1b: the association between comment slant and intentions to wear masks will be mediated by social media users' perception of the influence of the mask-promoting post on others.	Supported
Hypothesis 2a: social media users will have weaker intentions to wear masks when exposed to uncivil comments below a mask-promoting post than when exposed to civil comments.	Not supported
Hypothesis 2b: the association between comment tone and intentions to wear masks will be mediated by social media users' perception of the influence of the mask-promoting post on others.	Supported only when the mask-promoting post was accompanied by pro–mask-wearing comment
Hypothesis 3a: social media users will have weaker intentions to wear masks when they have unfavorable attitudes toward mask-wearing than when they have favorable attitudes.	Supported
Hypothesis 3b: the association between prior attitudes toward mask-wearing and intentions to wear masks will be mediated by social media users' perception of the influence of the mask-promoting post on others.	Supported only when the mask-promoting post was accompanied by pro–mask-wearing and civil comment
Hypothesis 4a: comment tone will moderate the effect of comment slant on social media users' intentions to wear masks, such that the effect of comment slant on behavioral intention will be stronger when comments are expressed in a civil manner compared to in an uncivil manner.	Supported
Hypothesis 4b: the interaction effect of comment slant and comment tone on social media users' intentions to wear masks will be mediated by their perception of the influence of the mask-promoting post on others.	Not supported
Hypothesis 5a: there is an interaction among comment slant, comment tone, and prior attitudes on social media users' intentions to wear masks, such that the influence of incivility will affect the influence of comments that reveal a slant inconsistent with social media user' prior attitudes on their behavioral intention to wear masks, but it will not affect the influence of comments that reveal a slant consistent with their preexisting attitudes.	Supported only among antimaskers
Hypothesis 5b: the impact of the interaction of comment slant, comment tone, and prior attitudes on social media users' intentions to wear masks will be mediated by their perception of the influence of the mask-promoting post on others.	Not supported

Results and Comparison With Prior Work

First, comment slant remained a cornerstone driving individuals' presumed influence of the mask-promoting post and their compliance with mask-wearing. Compared to pro–mask-wearing comments, anti–mask-wearing comments always reduced participants' presumed influence of mask-wearing posts, which further weakened their behavioral intention to wear masks, regardless of comment tone and their prior attitudes toward mask-wearing. These findings suggest that comments can serve as a source of misleading information. Although attitudinal consensus is inferred from the comments left by anonymous and limited others, these comments may lead social media users to develop inaccurate beliefs that the comments reflect public opinion from people in general. These beliefs may influence their health-related compliance behaviors.

In addition, incivility affected the presumed influence of a health message, but only when the comments below the message expressed supportive opinions. It is likely that pro–mask-wearing comments below the health message signal the presumed influence of the message on others' acceptance, and incivility acts as a negative cue that hinders the exemplification effect and indicates that the highly homogeneous and consistent opinion environment depicted in the comments may not be accurate. In contrast, anti–mask-wearing comments have explicitly represented others' resistance to the main health

XSL•FO
RenderX

message, and the presence of incivility only signals a similar cue of others' resistance.

These effects of comment slant and comment tone advance the other-consciousness perspective of the IPI process in the context of digital health campaigns. Individuals' perception of others' reception of a media message is influenced by affordances offered by social media. Even when information on source credibility or audience size is absent, the presumed influence of social media messages still adjusts individuals' compliance behavior accordingly. Social media users take the roles of both content producers and commenters. The opinion environment is highly prone to produce and spread misleading information due to the lack of professional gatekeepers and polarized opinion climate. The IPI model demonstrates a psychological process through which individuals' exposure to health information and relevant discussions on social media affects their compliance with promoted health behaviors. Therefore, it is crucial to consider both the direct and indirect effects of social media comments below health-persuasive messages on public health outcomes when examining the persuasiveness of digital health communication.

With concrete clues about others' reactions to health persuasion obtained from comments, individuals no longer rely solely on their prior attitudes to infer the influence of a health campaign on others. These findings somewhat challenge the self-centric perspective of IPI. This change can be explained by the evolving media landscape. Previous studies support the self-centric perspective of IPI in the context of traditional media, where traditional media audiences have limited access to others' reactions to a message and are compelled to rely on their prior attitudes for inference. In social media, users can directly see others' reactions to a message. They no longer need to fully rely on personal attitudes to infer media influence on others. Comments serve as crucial sources for them to infer the influence of social media posts on others.

Only when approving and civil comments are present, prior attitudes can affect behavioral intentions through individuals' presumed media influence on others. One possible explanation is that individuals in general are subject to negative bias, that is, they are particularly susceptible to information that contains negativity or risks. Anti–mask-wearing comments or incivility impressed and influenced participants because these comments might exaggerate the negative side of mask-wearing and demonstrate hostility among commenters. Therefore, individuals' perception of others' reactions to the main message is believed to be influenced by negative cues rather than prior attitudes. Civil pro–mask-wearing comments suggested no cues of negativity, and individuals then relied on their prior attitudes to infer the perception of the post's influence on others.

In most cases, favorable prior attitudes toward mask-wearing directly enhance individuals' behavioral intention compared to unfavorable prior attitudes. Nevertheless, the influence of prior attitudes on behavioral intentions can be altered by social media comments ensuing digital health communication. Specifically, civil pro–mask-wearing comments directly enhance antimaskers' behavioral intention to wear masks more than uncivil pro–mask-wearing comments, whereas uncivil anti–mask-wearing comments turn out to enhance antimaskers' behavioral intention to wear masks more than civil anti–mask-wearing comments. An explanation is that individuals may psychologically dissociate themselves from a group whose members belong to a relatively inferior group [62]; incivility is seen as impolite and undesirable, and individuals may avoid belonging to a group whose members are rude and uncivil. The findings indicate that individuals engage in biased information processing only when they find civil support for their prior opinions, regardless of whether the support is narrated in the main message or in the comments.

Relatedly, while we suspected that antimask attitudes differing from the post advocacy would be associated with less presumed influence, there is a possibility that opponents of mask-wearing may adhere to conspiracy theories. Such individuals might suspect that everyone around them has been brainwashed by governmental health campaigns, thereby leading to very high presumed influence. In other words, there might be a curvilinear relationship between prior attitudes and presumed influence or a linear relationship between attitude strength and presumed influence among antimaskers. Therefore, we conducted additional tests and found that these possibilities were not supported by our data. These findings suggest that in the era of new media, where user responses to health campaigns are publicly visible, judgments about the presumed influence of a post rely more on these visible examples than on personal prior attitudes.

Limitations and Future Directions

This study has several limitations that should be acknowledged. First, we edited the comments to maintain consistent argument strength across conditions, and therefore, the level of perceived authenticity in the comments may differ. Furthermore, we used default Facebook avatars in the experimental stimuli. Uncivil social media comments coupled with default avatars may be regarded as bot accounts, given the heavily politicized discussion on mask-wearing in the United States. The perception of commenters as bots may affect the presumed influence accordingly. These two aspects suggest that the perceived unrealism of the stimuli, particularly the user comments created in this study, may reduce the validity of the findings. Given that, future studies would benefit from measuring the perceived realism of comments and controlling it as a covariate in the analyses.

Second, this study focuses on the effects of user comments and prior attitudes, leaving the main effectiveness of the health campaign post unexamined. Likewise, the interaction effects between the post and its accompanying comments on polarized publics' presumed influence and behavioral intentions remain unexplored. The combined effects of comment slant and comment tone may vary depending on the post presented together with the comments. The lack of examination of the interplay between comments and the post may hinder a nuanced understanding of the combined effects of digital information. Future studies are encouraged to consider the effectiveness of a post and its interaction with comments.

Third, participants were required to read the post and accompanying comments, which may not reflect real-life

XSL•FO
RenderX

scenarios where individuals may choose whether to browse the information or not. Participants may generate bias through the procedure of providing informed consent and reading the survey questions, influencing their later answers. These factors could also affect the validity of the study findings. Future research should use experimental designs that better reflect real-world settings.

Fourth, although the IPI model has long been used in health communication research and is valuable for addressing specific questions in this study, it primarily focuses on the indirect effects of health campaigns. However, within the context of public health communication, there are various alternative theoretical explanations for the effectiveness or ineffectiveness of health campaigns. For instance, fear appeals suggest that how information is presented by the supply side of communication can influence individuals' emotional reactions and health behavior changes [63]. Psychological reactance can be another relevant concept with regard to campaign failure from the recipients' perspective. When individuals perceive health campaigns to threaten their behavioral freedom, they react in ways contrary to the campaign's intent, resulting in communication failure [56]. In other words, the findings from this study should be interpreted as 1 aspect of evaluating the effectiveness of health campaigns. To gain a comprehensive understanding of their effectiveness, these findings should be integrated with insights from other theoretical perspectives.

Conclusions and Implications

Despite these limitations, our study suggests that online health campaigns may yield desirable outcomes when civil and supportive comments are present. Moreover, social media users often engage in biased processing of health persuasion and rely heavily on their prior attitudes to guide their subsequent compliance behaviors. Unfavorable prior attitudes toward health behaviors can harm the effects of digital health communication only when individuals find civil and consistent evidence supporting their unfavorable opinions. Therefore, it is beneficial to encourage social media users to leave civil and supportive comments on digital health campaigns. In addition, misinformation and incivility in online comment sections should be moderated by relevant media platforms. Moreover, relevant information literacy programs should be delivered to the public to prevent them from being misled by biased user comments. Theoretically, this study explores the other-consciousness and self-centered perspectives of presumed influence in the context of social media health campaigns, where messages are presented together with extensive polarized and hostile user comments. People rely on online commentary and their prior attitudes to infer the presumed influence of health campaigns.

Acknowledgments

Generative AI (ie, ChatGPT-4) was used for proofreading the manuscript and improving readability, but it was not used in any part of the manuscript's generation. This study did not receive any funding.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

FL conceived the study, collected and analyzed the data, and drafted the manuscript. CT validated the study and proofread the manuscript. Both FL and CT contributed to review, revision, and editing and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the experimental stimuli.
[DOCX File , 636 KB - jmir_v26i1e57967_app1.docx]

Multimedia Appendix 2
Checklist for reporting of web-based and mobile randomized controlled trials (CONSORT-EHEALTH [Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth]).
[PDF File (Adobe PDF File), 1231 KB - jmir_v26i1e57967_app2.pdf]

Multimedia Appendix 3
Checklist for Reporting Results of Internet e-Surveys (CHERRIES).
[DOCX File , 17 KB - jmir_v26i1e57967_app3.docx]

Multimedia Appendix 4
Participants' demographics across groups.

XSL•FO
RenderX

[DOCX File , 16 KB - jmir_v26i1e57967_app4.docx]

Multimedia Appendix 5
Sensitivity analysis 1 (regression results).
[DOCX File , 15 KB - jmir_v26i1e57967_app5.docx]

Multimedia Appendix 6
Sensitivity analysis 1 (direct and indirect effects tests).
[DOCX File , 15 KB - jmir_v26i1e57967_app6.docx]

Multimedia Appendix 7
Sensitivity analysis 2 (dependent variable: presumed influence).
[DOCX File , 15 KB - jmir_v26i1e57967_app7.docx]

Multimedia Appendix 8
Sensitivity analysis 2 (dependent variable: mask-wearing intention).
[DOCX File , 16 KB - jmir_v26i1e57967_app8.docx]

References

1. Rao A, Morstatter F, Hu M, Chen E, Burghardt K, Ferrara E, et al. Political partisanship and antiscience attitudes in online discussions about COVID-19: Twitter content analysis. J Med Internet Res 2021 Jun 14;23(6):e26692 [FREE Full text] [doi: 10.2196/26692] [Medline: 34014831]
2. Oksanen A, Garcia D, Sirola A, Näsi M, Kaakinen M, Keipi T, et al. Pro-anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. J Med Internet Res 2015 Nov 12;17(11):e256 [FREE Full text] [doi: 10.2196/jmir.5007] [Medline: 26563678]
3. Lenti J, Mejova Y, Kalimeri K, Panisson A, Paolotti D, Tizzani M, et al. Global misinformation spillovers in the vaccination debate before and during the COVID-19 pandemic: multilingual Twitter study. JMIR Infodemiology 2023 May 24;3:e44714 [FREE Full text] [doi: 10.2196/44714] [Medline: 37223965]
4. Stevens H, Rasul ME, Oh YJ. Emotions and incivility in vaccine mandate discourse: natural language processing insights. JMIR Infodemiology 2022 Sep 13;2(2):e37635 [FREE Full text] [doi: 10.2196/37635] [Medline: 36188420]
5. Coe K, Kenski K, Rains SA. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. J Commun 2014 Jun 16;64(4):658-679. [doi: 10.1111/jcom.12104]
6. Kite J, Chan L, MacKay K, Corbett L, Reyes-Marcelino G, Nguyen B, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25:e46345 [FREE Full text] [doi: 10.2196/46345] [Medline: 37450325]
7. Lu F, Sun Y. COVID-19 vaccine hesitancy: the effects of combining direct and indirect online opinion cues on psychological reactance to health campaigns. Comput Human Behav 2022 Feb;127:107057 [FREE Full text] [doi: 10.1016/j.chb.2021.107057] [Medline: 34707328]
8. Lu F, Sun Y, Oktavianus J. Resistance to masks during the COVID-19 pandemic: how user comments drive psychological reactance to health campaigns. Health Commun 2023 Dec 21;38(14):3193-3206. [doi: 10.1080/10410236.2022.2141045] [Medline: 36411522]
9. Sun Y, Lu F. How misinformation and rebuttals in online comments affect people's intention to receive COVID-19 vaccines: the roles of psychological reactance and misperceptions. Journal Mass Commun Q 2023 Mar 31;100(1):145-171 [FREE Full text] [doi: 10.1177/10776990221084606] [Medline: 36814707]
10. Wang S. The influence of anonymity and incivility on perceptions of user comments on news websites. Mass Commun Soc 2020 Jul 07;23(6):912-936. [doi: 10.1080/15205436.2020.1784950]
11. Lee FL, Liang H, Tang GK. Online incivility, cyberbalkanization, and the dynamics of opinion polarization during and after a mass protest event. Int J Commun 2019;13(2019) [FREE Full text]
12. Kim Y, Kim Y. Incivility on Facebook and political polarization: the mediating role of seeking further comments and negative emotion. Comput Human Behav 2019 Oct;99:219-227. [doi: 10.1016/j.chb.2019.05.022]
13. Gunther AC, Storey JD. The influence of presumed influence. J Commun 2003 Jun 1;53(2):199-215. [doi: 10.1111/j.1460-2466.2003.tb02586.x]
14. Chung JE. Peer influence of online comments in newspapers: applying social norms and the social identification model of deindividuation effects (SIDE). Soc Sci Comput Rev 2018 Jun 06;37(4):551-567. [doi: 10.1177/0894439318779000]
15. Cho H, Shen L, Peng L. Examining and extending the influence of presumed influence hypothesis in social media. Media Psychol 2021 Feb 18;24(3):413-435 [FREE Full text] [doi: 10.1080/15213269.2020.1729812] [Medline: 34421390]
16. Gunther AC, Bolt D, Borzekowski DL, Liebhart JL, Dillard JP. Presumed influence on peer norms: how mass media indirectly affect adolescent smoking. J Commun 2006;56(1):52-68. [doi: 10.1111/j.1460-2466.2006.00002.x]

XSL•FO
RenderX

17. Yoo W, Yang J, Cho E. How social media influence college students' smoking attitudes and intentions. Comput Human Behav 2016 Nov;64:173-182 [FREE Full text] [doi: 10.1016/j.chb.2016.06.061] [Medline: 27956757]
18. Kluck JP, Krämer NC. Appraising uncivil comments in online political discussions: how do preceding incivility and senders' stance affect the processing of an uncivil comment? Commun Res 2022 Nov 13;50(4):453-479. [doi: 10.1177/00936502221113812]
19. Pascual-Ferrá P, Alperstein N, Barnett DJ, Rimal RN. Toxicity and verbal aggression on social media: polarized discourse on wearing face masks during the COVID-19 pandemic. Big Data Soc 2021 Jun 10;8(1). [doi: 10.1177/20539517211023533]
20. Yiannakoulias N, Darlington JC, Slavik CE, Benjamin G. Negative COVID-19 vaccine information on Twitter: content analysis. JMIR Infodemiology 2022 Aug 29;2(2):e38485 [FREE Full text] [doi: 10.2196/38485] [Medline: 36348980]
21. Majó-Vázquez S, Nielsen R, Verdú J, Rao N, Domenico N, Papaspiliopoulos O. Volume and patterns of toxicity in social media conversations during the COVID-19 pandemic. Reuters Institute for the Study of Journalism. 2020 Jul 09. URL: https://ora.ox.ac.uk/objects/uuid:e3862b70-2c75-47bf-a403-c7d7eaae4702 [accessed 2024-10-10]
22. Block RJ, Burnham M, Kahn K, Peng R, Seeman J, Seto C. Perceived risk, political polarization, and the willingness to follow COVID-19 mitigation guidelines. Soc Sci Med 2022 Jul;305:115091 [FREE Full text] [doi: 10.1016/j.socscimed.2022.115091] [Medline: 35690035]
23. Ho SS, Lee EW, Ng K, Leong GS, Tham TH. For fit's sake: a norms-based approach to healthy behaviors through influence of presumed media influence. Health Commun 2016 Sep 22;31(9):1072-1080. [doi: 10.1080/10410236.2015.1038772] [Medline: 26799846]
24. Chia SC, Sun Y, Lu F, Gudmundsdottir A. Doxing, regulation, and privacy protection: expanding the behavioral consequences of the third-person effect. Asian J Commun 2023 Mar 27;33(3):289-307. [doi: 10.1080/01292986.2023.2194883]
25. Hong Y, Kim S. Influence of presumed media influence for health prevention: how mass media indirectly promote health prevention behaviors through descriptive norms. Health Commun 2020 Dec;35(14):1800-1810. [doi: 10.1080/10410236.2019.1663585] [Medline: 31496288]
26. Lin CA, Xu X, Dam L. Information source dependence, presumed media influence, risk knowledge, and vaccination intention. Atl J Commun 2020 Feb 11;29(2):53-64. [doi: 10.1080/15456870.2020.1720022]
27. Tu C, Li X. Personal versus societal risk: examining social media influence on individual and collective behaviors for COVID-19 containment. Soc Media Society 2024 Jan 27;10(1). [doi: 10.1177/20563051231224402]
28. Shi R, Messaris P, Cappella JN. Effects of online comments on smokers' perception of anti-smoking public service announcements. J Comput Mediat Commun 2014 Jul 10;19(4):975-990 [FREE Full text] [doi: 10.1111/jcc4.12057] [Medline: 25561825]
29. Tu C, Li X. Social support and reputational defense strategies of Chinese social networking site users in Sun Yang's controversy. Public Relat Rev 2022 Jun;48(2):102181. [doi: 10.1016/j.pubrev.2022.102181]
30. Winter S, Brückner C, Krämer NC. They came, they liked, they commented: social influence on facebook news channels. Cyberpsychol Behav Soc Netw 2015 Aug;18(8):431-436. [doi: 10.1089/cyber.2015.0005] [Medline: 26252927]
31. Zillmann D. Exemplification effects in the promotion of safety and health. J Commun 2006;56(suppl_1):S221-S237. [doi: 10.1111/j.1460-2466.2006.00291.x]
32. Lee EJ, Jang YJ. What do others' reactions to news on internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. Commun Res 2010 Jul 16;37(6):825-846. [doi: 10.1177/0093650210376189]
33. Waddell TF, Sundar SS. Bandwagon effects in social television: how audience metrics related to size and opinion affect the enjoyment of digital media. Comput Human Behav 2020 Jun;107:106270. [doi: 10.1016/j.chb.2020.106270]
34. Chia SC, Wen N. College men's third-person perceptions about idealized body image and consequent behavior. Sex Roles 2010 Jul 24;63(7-8):542-555. [doi: 10.1007/s11199-010-9833-z]
35. Anderson AA, Yeo SK, Brossard D, Scheufele DA, Xenos MA. Toxic talk: how online incivility can undermine perceptions of media. Int J Public Opin Res 2018;30(1):156-168. [doi: 10.1093/ijpor/edw022]
36. Prochazka F, Weber P, Schweiger W. Effects of civility and reasoning in user comments on perceived journalistic quality. Journal Stud 2016 Mar 22;19(1):62-78. [doi: 10.1080/1461670x.2016.1161497]
37. Bode L, Vraga EK, Tully M. Do the right thing: tone may not affect correction of misinformation on social media. Harvard Kennedy School (HKS) Misinformation Review. 2020 Jun 11. URL: https://misinforeview.hks.harvard.edu/article/do-the-right-thing-tone-may-not-affect-correction-of-misinformation-on-social-media/ [accessed 2024-10-10]
38. Masullo GM, Lu S, Fadnis D. Does online incivility cancel out the spiral of silence? A moderated mediation model of willingness to speak out. New Media Soc 2020 Sep 10;23(11):3391-3414. [doi: 10.1177/1461444820954194]
39. Waddell TF, Bailey A. Inspired by the crowd: the effect of online comments on elevation and universal orientation. Commun Monogr 2017 Sep 13;84(4):534-550. [doi: 10.1080/03637751.2017.1369137]
40. Fields JM, Schuman H. Public beliefs about the beliefs of the public. Public Opin Q 1976;40(4):427-448. [doi: 10.1086/268330]
41. Cherniak C, Nisbett R, Ross L. Human inference: strategies and shortcomings of social judgment. Philos Rev 1983 Jul;92(3):462-465. [doi: 10.2307/2184495]
42. Marks G, Miller N. Ten years of research on the false-consensus effect: an empirical and theoretical review. Psychol Bull 1987;102(1):72-90. [doi: 10.1037//0033-2909.102.1.72]

XSL•FO
RenderX

43. Shen L, Huggins C. Testing the model of influence of presumed influence in a boundary condition: the impact of question order. Hum Commun Res 2013 Jun 18;39(4):470-491. [doi: 10.1111/hcre.12013]
44. Kunda Z. The case for motivated reasoning. Psychol Bull 1990;108(3):480-498. [doi: 10.1037//0033-2909.108.3.480]
45. Vraga EK, Kim SC, Cook J. Testing logic-based and humor-based corrections for science, health, and political misinformation on social media. J Broadcast Electron Media 2019 Sep 20;63(3):393-414. [doi: 10.1080/08838151.2019.1653102]
46. Kim H, Seo Y, Yoon HJ, Han JY, Ko Y. The effects of user comment valence of Facebook health messages on intention to receive the flu vaccine: the role of pre-existing attitude towards the flu vaccine and psychological reactance. Int J Advert 2021 Jan 04;40(7):1187-1208. [doi: 10.1080/02650487.2020.1863065]
47. Georgiou N, Delfabbro P, Balzan R. COVID-19-related conspiracy beliefs and their relationship with perceived stress and pre-existing conspiracy beliefs. Pers Individ Dif 2020 Nov 01;166:110201 [FREE Full text] [doi: 10.1016/j.paid.2020.110201] [Medline: 32565592]
48. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
49. Pornpitakpan C. The persuasiveness of source credibility: a critical review of five decades' evidence. J Appl Soc Pyschol 2006 Jul 31;34(2):243-281. [doi: 10.1111/j.1559-1816.2004.tb02547.x]
50. Hogg MA. A social identity theory of leadership. Pers Soc Psychol Rev 2001 Aug 01;5(3):184-200. [doi: 10.1207/s15327957pspr0503_1]
51. Aguinis H, Villamor I, Ramani RS. MTurk research: review and recommendations. J Manag 2020 Nov 19;47(4):823-837. [doi: 10.1177/0149206320969787]
52. Mortensen K, Hughes TL. Comparing Amazon's mechanical Turk platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med 2018 Apr 4;33(4):533-538 [FREE Full text] [doi: 10.1007/s11606-017-4246-0] [Medline: 29302882]
53. Kennedy R, Clifford S, Burleigh T, Waggoner PD, Jewell R, Winter NJ. The shape of and solutions to the MTurk quality crisis. Political Sci Res Methods 2020 Apr 24;8(4):614-629. [doi: 10.1017/psrm.2020.6]
54. Community respirators and masks. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/niosh/topics/publicppe/community-ppe.html [accessed 2024-08-29]
55. About COVID-19. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/covid/about/index.html [accessed 2024-08-28]
56. Dillard JP, Shen L. On the nature of reactance and its role in persuasive health communication. Commun Monogr 2005 Jun;72(2):144-168. [doi: 10.1080/03637750500111815]
57. Rösner L, Winter S, Krämer NC. Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. Comput Human Behav 2016 May;58:461-470. [doi: 10.1016/j.chb.2016.01.022]
58. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
59. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
60. Tsfati Y, Cohen J, Gunther AC. The influence of presumed media influence on news about science and scientists. Sci Commun 2010 Sep 14;33(2):143-166. [doi: 10.1177/1075547010380385]
61. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: Guilford Press; 2017.
62. Hornsey MJ. Social identity theory and self‐categorization theory: a historical review. Soc Personal Psychol Compass 2008 Jan 14;2(1):204-222. [doi: 10.1111/j.1751-9004.2007.00066.x]
63. Moussaoui LS, Claxton N, Desrichard O. Fear appeals to promote better health behaviors: an investigation of potential mediators. Health Psychol Behav Med 2021 Jul 06;9(1):600-618 [FREE Full text] [doi: 10.1080/21642850.2021.1947290] [Medline: 34285825]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IPI: influence of presumed influence
MTurk: Mechanical Turk

XSL•FO
RenderX

Edited by A Schwartz; submitted 01.03.24; peer-reviewed by M Marelić, D Thiele; comments to author 12.06.24; revised version received 26.06.24; accepted 12.09.24; published 15.11.24.

Please cite as:
Lu F, Tu C
The Impact of Comment Slant and Comment Tone on Digital Health Communication Among Polarized Publics: A Web-Based Survey Experiment
J Med Internet Res 2024;26:e57967
URL: https://www.jmir.org/2024/1/e57967
doi:10.2196/57967
PMID:

©Fangcao Lu, Caixie Tu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Exclusion and Cognitive Function in Elderly Populations in Developing Countries: Insights Derived From 2 Longitudinal Cohort Studies

Sainan Duan[1,2,3*], MD; Dongxu Chen[1,2*], PhD; Jinping Wang[3,4], MD; Mohammed Sharooq Paramboor[1,2,3], MD; Zhen Xia[5], MD; Wanting Xu[6], MD; Kun Han[1,2], PhD; Tao Zhu[3,4], PhD; Xiaoqin Jiang[1,2], MD

[1]Department of Anesthesiology, West China Second University Hospital, Sichuan University, Chengdu, China
[2]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, China
[3]Department of Anesthesiology, West China Hospital, Sichuan University, Chengdu, China
[4]The Research Units of West China (2018RU012)-Chinese Academy of Medical Sciences, West China Hospital, Sichuan University, Chengdu, China
[5]West China School of Medicine, Sichuan University, Chengdu, China
[6]Department of Anesthesiology, Shengli Oilfield Central Hospital, Dongying, China
[*]these authors contributed equally

Corresponding Author:
Xiaoqin Jiang, MD
Department of Anesthesiology
West China Second University Hospital
Sichuan University
No 20, Section 3
South Renmin Road
Chengdu, 610041
China
Phone: 86 18180609218
Email: 1598862657jxq@scu.edu.cn

Abstract

Background: Cognition disorders not only lead to adverse health consequences but also contribute to a range of socioeconomic challenges and diminished capacity for performing routine daily activities. In the digital era, understanding the impact of digital exclusion on cognitive function is crucial, especially in developing countries.

Objective: This study aimed to evaluate the association between digital exclusion and cognitive function among elderly populations in developing countries.

Methods: Using data from CHARLS (China Health and Retirement Longitudinal Study) from 2011 to 2020 and MHAS (Mexican Health & Aging Study) from 2012 to 2021, we defined digital exclusion as self-reported absence from the internet. Cognitive function was assessed through 5 tests: orientation, immediate verbal recall, delayed verbal recall, serial 7s, and figure recall. Cognitive function was assessed in 2 categories: worse cognition (a categorical variable that classifies cognition as either better or worse compared to the entire cohort population) and cognitive scores (a continuous variable representing raw cognitive scores across multiple follow-up waves). Logistic regression analyses and generalized estimating equation (GEE) analyses were used to examine the relationship between cognitive function and digital exclusion, adjusting for potential confounders, including demographics, lifestyle factors, history of chronic diseases, basic activities of daily living (BADL) disability, instrumental activities of daily living (IADL) disability, and basic cognitive abilities.

Results: After excluding participants with probable cognitive impairment at baseline and those who did not have a complete cognitive assessment in any given year (ie, all tests in the cognitive assessment must be completed in any follow-up wave), a total of 24,065 participants in CHARLS (n=11,505, 47.81%) and MHAS (n=12,560, 52.19%) were included. Of these, 96.78% (n=11,135) participants in CHARLS and 70.02% (n=8795) in MHAS experienced digital exclusion. Adjusted logistic regression analyses revealed that individuals with digital exclusion were more likely to exhibit worse cognitive performance in both CHARLS (odds ratio [OR] 2.04, 95% CI 1.42-2.99; $P<.001$) and MHAS (OR 1.40, 95% CI 1.26-1.55; $P<.001$). Gender and age did not significantly modify the relationship between digital exclusion and worse cognition (intervention $P>.05$). The fully adjusted mean

XSL•FO
RenderX

differences in global cognitive scores between the 2 groups were 0.98 (95% CI 0.70-1.28; *P*<.001) in CHARLS and 0.50 (95% CI 0.40-0.59; *P*<.001) in MHAS.

Conclusions: A substantial proportion of older adults, particularly in China, remain excluded from internet access. Our study examined longitudinal changes in cognitive scores and performed cross-sectional comparisons using Z-score standardization. The findings suggest that digital exclusion is linked to an increased risk of cognitive decline among older adults in developing countries. Promoting internet access may help mitigate this risk and support better cognitive health in these populations.

(J Med Internet Res 2024;26:e56636) doi:10.2196/56636

KEYWORDS

digital exclusion; cognition; cohort study; developing country; China Health and Retirement Longitudinal Study; CHARLS; ; MHAS; cognitive decline

Introduction

The worldwide phenomenon of population aging presents significant health and socioeconomic challenges. Notably, China, home to the world's largest older adult population, is projected to see over 479 million individuals aged 60 years and above by 2050 [1]. The prevalence of dementia among older adults in China is anticipated to escalate to 18 million by 2030 [2]. In Western developing countries, such as Mexico, the prevalence of dementia among people over 65 years old (7.1%-11.5%) is increasing compared to Europe and the United States [3]. These trends underscore the urgent need for effective public health strategies focusing on dementia prevention and treatment. Cognitive decline, characterized by a deterioration in memory and other cognitive abilities, manifests as a noticeable reduction in cognitive performance compared to previous levels [4], poses substantial socioeconomic challenges, and affects the ability to perform daily activities. It is further exacerbated by factors such as chronic conditions (eg, hypertension or hearing impairment) [5]. However, significant gaps remain in understanding modifiable risk factors and their potential role in the prevention or decelerating cognitive decline.

With technological advancements, the internet has become an essential digital platform for communication, information access [6], and medical consultation [7]. However, digital exclusion—defined as the lack of access to information and communications technologies, including the internet [8]—remains prevalent. In more prosperous countries with well-established market economies, internet usage rates are high [9], but in some developing countries, the proportion of internet users is significantly lower, particularly among the elderly. For example, digital exclusion among the elderly reaches 97% in China and 66% in Mexico, significantly exceeding 25%-65% typical in other nations [10]. The association between mental health and internet use has been extensively explored [11-13], although the association remains unclear. Additionally, research indicated that middle-aged and older adults who use the internet have a lower risk of developing chronic diseases compared to those who do not use the internet [14]. Digital exclusion among older adults is associated with functional dependence, as evidenced by 5 longitudinal cohort studies [10]. Previous studies have suggested that in developed countries, such as England [15], Australia [16], and the United States [17], internet access can influence cognitive function in older individuals. The potential mechanisms for explaining the association between digital exclusion and cognition could be explained as follows: the internet serves as a critical tool for the elderly population, enabling access to up-to-date health management information, medication purchases, and health devices procurement [18]. It also facilitates timely consultations with health care professionals and real-time data monitoring, which are essential for effective health management [19,20]. Additionally, vulnerable digital excluders, such as those with lower levels of social communication with colleagues or family members, are more likely to experience poor health status compared to those who are not in a vulnerable position. This observation aligns with studies highlighting that internet engagement promotes social activities that can alleviate loneliness and enhance overall health [21]. Thus, the internet has emerged as an alternative means to real-life interactions, helping reduce social isolation and enhance the quality of life for older adults [22]. However, there is limited research on the association between digital exclusion and cognitive performance among the elderly in developing countries. Liu et al [23] demonstrated a correlation between digital exclusion and cognitive impairment in a cross-sectional study of 10,325 people in China, but no causal relationships could be established. Through cross-sectional and longitudinal data, Li et al [24] revealed that internet use can help prevent the contraction of the pallidum and thus mitigate the decline of cognitive function in elderly Chinese individuals. However, this study only conducted a 1-year follow-up [24]. Jin et al [25] conducted a similar study that was still limited to China. The methodological limitations of many existing studies encompass varied definitions of cognitive decline [23-25], cross-sectional design [23,24], or geographic limitations [23-25], as well as insufficient analysis of risk factors in developing countries, thus obscuring the actual role of internet use on cognitive decline risk. Similar attempts are also needed to elucidate the role of digital exclusion in these regions.

CHARLS (China Health and Retirement Longitudinal Study) and MHAS (Mexican Health & Aging Study) cohorts, comprising large, prospective longitudinal data sets with a wealth of variables and extensive follow-up information, represented ideal study populations for this study. Both China and Mexico, as developing countries, exhibit high rates of digital exclusion, further underscoring the relevance of these cohorts for examining the impacts of this association. In addition, differences in regional, demographic, and educational characteristics between the 2 cohorts not only allow for a robust verification of study results but also enhance their

XSL•FO
RenderX

generalizability. Thus, we used comprehensive data from CHARLS and MHAS to examine the impact of digital exclusion on cognitive function among older adults in developing countries. Furthermore, we specifically incorporated the latest CHARLS Wave 5 National Follow-up Survey data of 2020, released on November 16, 2023.

Methods

Ethical Considerations

CHARLS was approved by Peking University (IRB00001052-11015), and MHAS was approved by the University of Texas Medical Branch in the United States, the Instituto Nacional de Estadística y Geografía (INEGI), and the Instituto Nacional de Salud Pública (INSP) in Mexico. Written informed consent was obtained from all participants prior to the survey. This study adhered to the ethical guidelines outlined in the Declaration of Helsinki. There was no requirement for additional ethics approval for approved data users in CHARLS and MHAS [26,27].

Data Source and Population

CHARLS, a comprehensive longitudinal survey, targets individuals aged 45 years or older and their spouses across 150 counties/districts and 450 villages/urban areas in 28 provinces of China, including assessments of the social, economic, and health conditions of community residents [26]. MHAS is a national longitudinal study focusing on adults aged 50 years or above in Mexico, aimed at prospectively evaluating the impact of disease on health, function, and mortality among Mexican adults [27]. Data for this study were sourced from the publicly accessible CHARLS and MHAS databases. This study incorporated data from 5 waves of CHARLS (2011 wave 1, 2013 wave 2, 2015 wave 3, 2018 wave 4, 2020 wave 5) and 4 waves from MHAS (2012 wave 3, 2015 wave 4, 2018 wave 5, and 2021 wave 6).

Participants with memory-related disorders, brain damage, or stroke at baseline (2011 wave 1 in CHARLS, 2012 wave 3 in MHAS) were excluded. Also excluded were those lacking digital evaluation at baseline or those without a single complete cognitive assessment in subsequent waves. Finally, the analyses included 11,505 eligible participants from CHARLS and 12,560 from MHAS.

Ascertainment of Digital Exclusion

Data regarding digital exclusion were obtained via self-completed questionnaires. In CHARLS, digital exclusion was assessed with question DA056: “Have you used the internet in the past month?” For MHAS, where individual internet usage was not directly assessed, an alternative question (J18), “In this house do you have Internet?” was used to infer digital exclusion [10]. A response of no to both questions was categorized as evidence of digital exclusion, whereas a response of yes indicated inclusion.

Assessment of Cognitive Function

Cognitive functional status was assessed using 2 methods: categorical (based on the standardized *Z*-score, categorizing cognition as either better or worse cognition compared to the entire cohort population) and continuous (cognitive raw scores in multiple follow-up waves). Cognitive scores included assessments across 5 tests: orientation, immediate verbal recall, delayed verbal recall, serial 7s, and figure recall [28,29]. Notably, the serial 7s component was unavailable in the MHAS cohort. Orientation assessment in CHARLS covered the year, month, day, day of the week, and seasonal, for a total of 5 points. In MHAS, the orientation assessment excluded the day of the week and season. Immediate and delayed verbal recall tasks involved recalling unrelated words, with 1 point given for each word recalled. The maximum score for immediate verbal recall was 10 points in CHARLS and 8 in MHAS, while delayed verbal recall had a ceiling of 10 points in CHARLS and 8 in MHAS. Figure recall involved participants replicating a previously given graph, scoring 1 point in CHARLS and up to 6 points in MHAS. The serial 7s task required counting backward from 100 by 7 five times, with each correct response earning 1 point. Overall, the global cognitive scores ranged from 0 to 31 in CHARLS and from 0 to 25 in MHAS, with higher scores indicating superior cognitive performance [30]. Detailed methodologies of cognitive function assessments in both CHARLS and MHAS are documented in Tables S1 and S2 in Multimedia Appendix 1, respectively.

In this study, worse cognition was used as a proxy for assessing cognitive function in comparison to others at the follow-up endpoint. Worse cognition was accessed in 2 steps. First, the cognitive score was obtained from all follow-up waves, and then the raw score was standardized to the baseline to obtain the *Z*-score [30]. For instance, in CHARLS, the mean (SD) of global cognitive scores at baseline was 14.91 (1.35). Next, the *Z*-score was calculated as (last complete raw cognitive score – 14.91)/1.35, with positive *Z*-values indicating better cognitive ability than the mean population and negative *Z*-values suggesting worse cognition [29]. Worse cognition was used to assess performance in comparison to other participants, and the methodology for this assessment was based on previous published studies [29,30].

Covariates

Covariates were identified through literature reviews and included demographics, lifestyle, history of chronic diseases, basic activities of daily living (BADL) disability, instrumental activities of daily living (IADL) disability, and anthropometric measures. Demographics included age, gender, marital status (single, married or cohabiting, and divorced or widowed), education level, and residence (rural or urban). Education levels were divided into lower secondary (illiteracy, primary school, and junior high schools), upper secondary and vocational training, and tertiary (bachelor’s degree or above). Lifestyle factors assessed were smoking and drinking status. A history of chronic diseases encompassed hypertension, diabetes, dyslipidemia, heart disease, respiratory illness, liver disease, kidney disease, digestive disease, arthritis, rheumatism, and cancer. BADL were assessed using the Katz scale with 6 items: continence, dressing, bathing, feeding, transferring, and going to the toilet [31]. According to Lawton and Brody [32], IADL include housekeeping, preparing hot meals, shopping, managing money, taking medications, and using the telephone. BADL disability and IADL disability were defined as having difficulties

XSL•FO
RenderX

in doing any item in BADL and IADL, respectively [33]. Anthropometric measures included the hand grip test (kg) and waist circumference (cm). The BMI was calculated as the measure of body weight in kilograms divided by the measure of body height in meters squared. Cognitive scores at baseline were considered candidate covariates in statistical analysis. Additional information about the covariates in CHARLS and MHAS is presented in Tables S3 and S4 in Multimedia Appendix 1, respectively.

Statistical Analysis

Categorical variables were described as percentages, with missing values indicated by "unknown," and continuous variables were expressed as mean (SD) or median (IQR; 25th-75th percentiles, Q1-Q3). Continuous variables were compared between groups using the Wilcoxon rank-sum test, and categorical variables were compared using the Pearson chi-squared test or the Fisher exact test. To identify covariates with high multicollinearity, a generalized variance inflation factor (GVIF) analysis was conducted [34-36], with adjusted generalized standard error inflation factor (aGSIF) values below 1.6 deemed acceptable [36]. We assessed the associations between digital exclusion and worse cognition using odds ratios (ORs), with 95% CIs derived from logistic regression models, partly (models 1-3) or fully (model 4), adjusted for age at baseline (continuous variable), gender (male or female), BMI group (≤24 kg/m^2 or >24 kg/m^2), marital status (single, married, or divorced), education level (lower secondary, upper secondary, or tertiary), residence (rural or urban), smoking status (current or never), drinking status (ever, current, or never), BADL and IADL disabilities (yes or no), global cognitive scores at baseline (continuous variable), and chronic diseases (yes or no). Furthermore, to detect the modification effect of age and sex on the studied associations, we performed subgroup analyses by age group (<50, 50-59, or >59 years) and sex (male or female).

Generalized estimating equations (GEEs) extend the generalized linear model to allow further longitudinal data analysis [37]. We appointed exchangeable structures as working correlation structures and analyzed raw cognitive scores (including the scores and total scores for each section of the cognitive tests) with GEE models [38] to evaluate the differences in the changes between digital exclusive and inclusive cognitive scores of follow-up waves [39]. In addition, the interaction of the time variable with cognitive scores was conducted to examine whether cognitive scores vary by digital exclusion [40].

R software (v4.3.0; R Core Team) was used for all statistical data analyses. The *geepack* package was used to perform GEE analysis. Two-tailed $P<.05$ was considered statistically significant.

Results

Baseline Characteristics of Participants

The final analysis included a total of 24,065 participants. In CHARLS, 3826 (15.9%) samples were excluded due to incomplete follow-up data, resulting in a final data set of 11,505 (47.81%) observations. In MHAS, 2724 (11.32%) samples were excluded due to incomplete follow-up data, leaving a final data set of 12,560 (52.19%) observations, as detailed in Tables 1 and 2 and in Figure S1 in Multimedia Appendix 1. The missing data for the covariate are referenced in Figure S2 in Multimedia Appendix 1. The mean age of participants was 57.28 years (SD 8.76) in CHARLS and 62.43 years (SD 10.01) in MHAS. A significant divergence was observed between the 2 cohorts in various measured parameters. Participants in CHARLS typically exhibited poorer health outcomes, as evidenced by higher comorbidity rates and more prevalent unhealthy lifestyles. Conversely, the MHAS cohort, characterized by an older age profile and a higher BMI, demonstrated increased challenges in BADL disability.

XSL•FO
RenderX

Table 1. Descriptive statistics in CHARLS[a] and MHAS[b].

Characteristics	CHARLS (n=11,505)	MHAS (n=12,560)
Age (years), mean (SD)	57.28 (8.76)	62.43 (10.01)
Gender, n (%)		
Male	5666 (49.25)	5170 (41.16)
Female	5839 (50.75)	7390 (58.84)
BMI (kg/m^2), mean (SD)	23.93 (10.76)	29.04 (5.26)
Marital status, n (%)		
Single	74 (0.64)	554 (4.41)
Married or cohabiting	10421 (90.58)	9110 (72.53)
Divorced or widowed	1010 (8.78)	2896 (23.06)
Education, n (%)		
Lower secondary	9811 (85.28)	9895 (78.78)
Upper secondary and vocational training	311 (2.70)	1513 (12.05)
Tertiary	1379 (11.99)	1099 (8.75)
Unknown	4 (0.03)	53 (0.42)
Region of residence, n (%)		
Rural	5374 (46.71)	—[c]
Town	4106 (35.69)	—
Unknown	2025 (17.60)	—
Smoking status, n (%)		
Current	3612 (31.40)	1552 (12.36)
Never	6944 (60.36)	8002 (63.71)
Ever	945 (8.21)	3005 (23.93)
Unknown	4 (0.03)	1 (0.01)
Drinking status, n (%)		
Current	3977 (34.57)	3060 (24.36)
Never	7526 (65.42)	9497 (75.61)
Unknown	2 (0.02)	3 (0.02)
Chronic disease, n (%)		
Hypertension	2560 (22.25)	5211 (41.49)
Diabetes	608 (5.28)	2630 (20.94)
Dyslipidemia	1061 (9.22)	—
Heart disease	1316 (11.44)	371 (2.95)
Respiratory illness	1203 (10.46)	661 (5.26)
Liver disease	456 (3.96)	—
Kidney disease	714 (6.21)	—
Digestive disease	2567 (22.31)	—
Arthritis or rheumatism	3696 (32.13)	1630 (12.98)
Cancer	101 (0.88)	229 (1.82)
BADL[d] disability, n (%)		
Yes	670 (5.82)	1700 (13.54)
No	4670 (40.59)	5222 (41.58)

XSL•FO
RenderX

Characteristics	CHARLS (n=11,505)	MHAS (n=12,560)
Unknown	6165 (53.59)	5638 (44.89)
IADL[e] disability, n (%)		
Yes	844 (7.34)	844 (6.72)
No	10655 (92.61)	10512 (83.69)
Unknown	6 (0.05)	1204 (9.59)
Hand grip test (k g)		
Mean (SD)	30.93 (10.05)	25.66 (8.65)
Median (IQR; quartiles 1-3)	30.00 (24.00-38.00)	25.00 (20.00-31.00)
Waist circumference, cm		
Mean (SD)	84.52 (12.58)	98.29 (12.46)
Median (IQR; quartiles 1-3)	85.00 (78.00-92.00)	97.70 (90.00-105.80)

[a]CHARLS: China Health and Retirement Longitudinal Study.

[b]MHAS: Mexican Health & Aging Study.

[c]Not applicable.

[d]BADL: basic activities of daily living. BADL disability was defined as having difficulties in continence, dressing, bathing, feeding, transferring, or going to the toilet.

[e]IADL: instrumental activities of daily living. IADL disability was defined as having difficulties in housekeeping, preparing hot meals, shopping, managing money, or taking medications.

Table 2. Baseline cognitive function in CHARLS[a] and MHAS[b,c].

Cognitive function	CHARLS (n=11,505)	MHAS (n=12,560)
Global cognition		
Mean (SD)	16.54 (4.53)	17.74 (3.45)
Median (IQR; quartiles 1-3)	17.00 (13.00-20.00)	18.33 (15.67-20.33)
Orientation		
Mean (SD)	4.32 (0.93)	2.55 (0.75)
Median (IQR; quartiles 1-3)	5.00 (4.00-5.00)	3.00 (2.00-3.00)
Immediate verbal recall		
Mean (SD)	4.30 (1.64)	4.88 (1.18)
Median (IQR; quartiles 1-3)	4.00 (3.00-5.00)	5.00 (4.00-5.67)
Delayed verbal recall		
Mean (SD)	3.37 (1.94)	4.60 (1.98)
Median (IQR; quartiles 1-3)	3.00 (2.00-5.00)	5.00 (4.00-5.67)
Serial 7s		
Mean (SD)	3.30 (1.82)	—[d]
Median (IQR; quartiles 1-3)	4.00 (1.00-5.00)	—
Figure recall		
Mean (SD)	0.72 (0.45)	5.59 (0.99)
Median (IQR; quartiles 1-3)	1.00 (0.00-1.00)	6.00 (6.00-6.00)

[a]CHARLS: China Health and Retirement Longitudinal Study.

[b]MHAS: Mexican Health & Aging Study.

[c]Scores are presented in different ranges due to different assessment methods. CHARLS: global cognition (0-31 points), orientation (0-5 points), immediate verbal recall (0-10 points), delayed verbal recall (0-10 points), serial 7s (0-5 points), and figure recall (0-1 points). MHAS: global cognition (0-25 points), orientation (0-3 points), immediate verbal recall (0-8 points), delayed verbal recall (0-8 points), and figure recall (0-6 points).

[d]Not applicable.

XSL•FO
RenderX

A high prevalence of digital exclusion was reported by 96.78% (n=11,135) participants in CHARLS and 70.02% (n=8795) in MHAS. The characteristics of participants, divided into digitally excluded and included groups, are summarized in Table S5 in Multimedia Appendix 1. Significant differences were observed in several demographic and health-related parameters between these groups in both cohorts. Briefly, participants experiencing digital exclusion tended to be older, had lower levels of education, and were more likely to suffer from disabilities related to daily living activities in both cohorts.

Association Between Digital Exclusion and Worse Cognition

Table 3 presents the relationship between digital exclusion and cognitive performance. When categorizing cognitive function based on overall cognitive scores (indicative of worse cognition), a higher incidence of worse cognition was observed among digitally excluded participants (CHARLS: n=5819, 52.26%, vs n=41, 11.08%; MHAS: n=5371, 61.08%, vs n=1377, 35.56%). In CHARLS, the age-adjusted, sex-adjusted, BMI group–adjusted, and baseline cognitive score–adjusted OR for worse cognition in digitally excluded participants was 3.37 (95% CI 2.39-4.89), as shown in model 1 in Table S6 in Multimedia Appendix 1). This OR reduced to 2.06 (95% CI 1.44-3.03) with further adjustments for additional confounders (models 2 and 3, Table S6 in Multimedia Appendix 1), and the fully adjusted OR (model 4, Table S6 in Multimedia Appendix 1) was 2.04 (95% CI 1.42-2.99). In MHAS, the association remained statistically significant (OR 1.40, 95% CI 1.26-1.55), with similar ORs observed across models with varying levels of adjustment (Table S6 in Multimedia Appendix 1). The elevated risk of worse cognition in individuals with prepandemic psychiatric disorders did not differ by age and sex in both CHARLS and MHAS (Figure 1). Table S7 in Multimedia Appendix 1 shows that all aGSIF values were close to 1 in CHARLS and MHAS, indicating the absence of multicollinearity among confounders in both cohorts.

Table 3. Cognition outcome in CHARLS[a] and MHAS[b] stratified by exposure.[c]

Cognition	CHARLS (n=11,505)		MHAS (n=12,560)	
	Digital exclusion (n=11,135)	Digital inclusion (n=370)	Digital exclusion (n=8794)	Digital inclusion (n=3766)
Worse cognition (measured by category), n (%)	5819 (52.26)	41 (11.08)	5371 (61.08)	1377 (36.56)
Global cognition (measured by scores); CHARLS *P*<.001, MHAS *P*<.001				
Mean (SD)	15.61 (5.88)	21.54 (3.75)	16.27 (3.94)	18.35 (3.31)
Median (IQR; quartiles 1-3)	16.00 (12.00-20.00)	22.00 (19.33-24.33)	16.67 (14.00-19.00)	18.67 (16.67-20.67)
Orientation (measured by scores); CHARLS *P*<.001, MHAS *P*<.001				
Mean (SD)	4.00 (1.14)	4.71 (0.57)	2.40 (0.88)	2.68 (0.66)
Median (IQR; quartiles 1-3)	4.00 (3.00-5.00)	5.00 (5.00-5.00)	3.00 (2.00-3.00)	3.00 (3.00-3.00)
Immediate verbal recall (measured by scores); CHARLS *P*<.001, MHAS *P*<.001				
Mean (SD)	4.12 (2.00)	5.82 (1.52)	4.54 (1.21)	5.19 (1.16)
Median (IQR; quartiles 1-3)	4.00 (3.00-5.67)	6.00 (5.00-7.00)	4.67 (3.67-5.33)	5.33 (4.33-6.00)
Delayed verbal recall (measured by scores); CHARLS *P*<.001, MHAS *P*<.001				
Mean (SD)	4.13 (2.60)	6.38 (1.83)	4.00 (1.99)	4.75 (1.84)
Median (IQR; quartiles 1-3)	4.00 (2.00-6.00)	6.00 (5.00-8.00)	4.00 (3.00-5.00)	5.00 (4.00-6.00)
Serial 7s (measured by scores); CHARLS *P*<.001				
Mean (SD)	2.87 (1.88)	3.92 (1.54)	—[d]	—
Median (IQR; quartiles 1-3)	3.00 (1.00-5.00)	5.00 (3.00-5.00)	—	—
Figure recall (measured by scores); CHARLS *P*<.001, MHAS *P*<.001				
Mean (SD)	0.49 (0.53)	0.71 (0.51)	5.33 (1.25)	5.73 (0.81)
Median (IQR; quartiles 1-3)	0.00 (0.00-1.00)	1.00 (0.00-1.00)	6.00 (5.00-6.00)	6.00 (5.00-6.00)

[a]CHARLS: China Health and Retirement Longitudinal Study.

[b]MHAS: Mexican Health & Aging Study.

[c]*P* values are derived from the Wilcoxon rank-sum test.

[d]Not applicable.

XSL•FO
RenderX

Figure 1. Subgroup analysis of digital exclusion and worse cognition by age group and sex in (A) CHARLS and (B) MHAS. Values are adjusted for age, gender, BMI group, marital status, education level, residence, smoking status, drinking status, BADL and IADL disabilities, chronic diseases, and baseline cognitive scores. BADL: basic activities of daily living; CHARLS: China Health and Retirement Longitudinal Study; IADL: instrumental activities of daily living; MHAS: Mexican Health & Aging Study; OR: odds ratio.

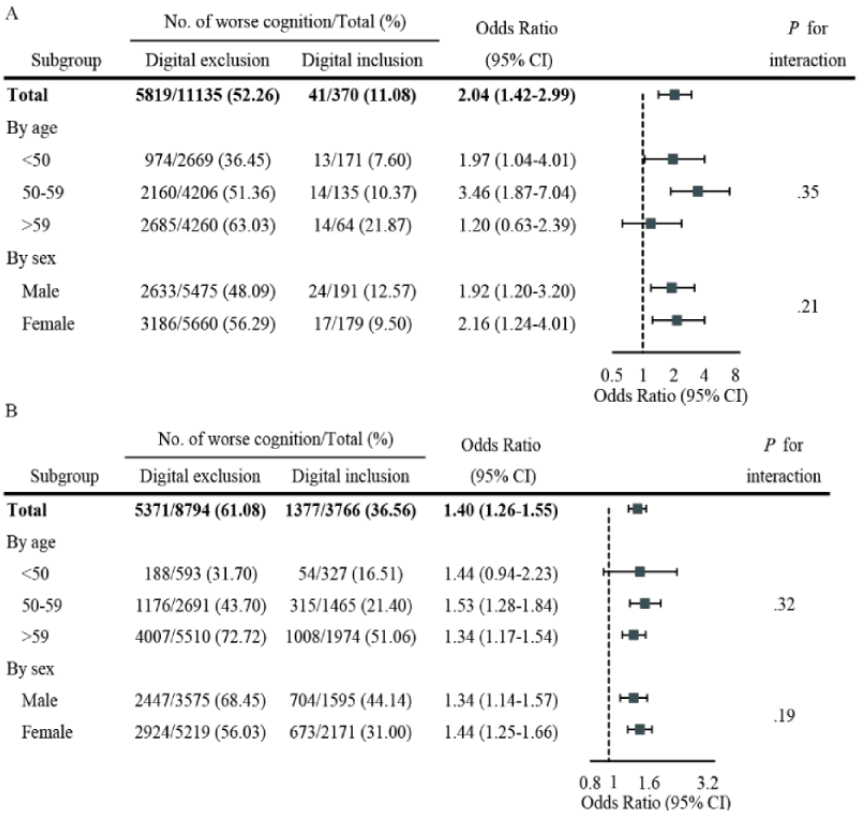

Association Between Digital Exclusion and Cognitive Scores Throughout the Follow-Up Waves

In analyzing the cognitive score performance, participants experiencing digital exclusion had lower global cognitive scores than those experiencing digital inclusion. The overall difference in global cognitive scores was 0.98 (95% CI 0.70-1.28) in CHARLS and 0.50 (95% CI 0.40-0.59) in MHAS, as detailed in Tables S8 and S9 in Multimedia Appendix 1, respectively, and Figure 2. GEE models indicated that participants in the digital exclusion group were significantly more likely to have lower cognitive scores during follow-up periods (2013-2020 in CHARLS and 2015-2021 in MHAS), albeit with marginal differences (Tables S8 and S9 in Multimedia Appendix 1, respectively). Overall group differences in individual components of the expanded composite outcome were largely consistent. These components included orientation (CHARLS: 0.13, 95% CI 0.09-0.19; MHAS: 0.10, 95% CI 0.08-0.12), immediate verbal recall (CHARLS: 0.33, 95% CI 0.19-0.48; MHAS: 0.18, 95% CI 0.15-0.22), delayed verbal recall (CHARLS: 0.47, 95% CI 0.28-0.65; MHAS: 0.21, 95% CI 0.15-0.27), serial 7s (CHARLS: 0.27, 95% CI 0.12-0.42), and figure recall (CHARLS: 0.09, 95% CI 0.07-0.12; MHAS: 0.15, 95% CI 0.12-0.17), as shown in Figure 3 and Tables S8 and S9 in Multimedia Appendix 1, respectively.

Figure 2. Average follow-up global cognitive scores in (A) CHARLS and (B) MHAS. Values are adjusted for baseline global cognitive scores, age, gender, BMI group, marital status, education level, residence, smoking status, BADL disability, IADL disability, and chronic diseases. BADL: basic activities of daily living; CHARLS: China Health and Retirement Longitudinal Study; IADL: instrumental activities of daily living; MHAS: Mexican Health & Aging Study.

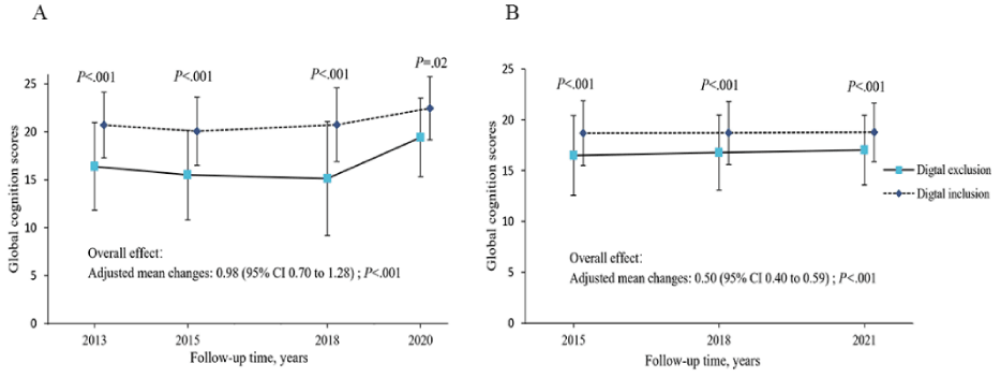

XSL•FO
RenderX

Figure 3. Average follow-up tests' cognitive scores in CHARLS and MHAS. (A) Orientation scores in CHARLS, (B) orientation scores in MHAS, (C) immediate verbal recall scores in CHARLS, (D) immediate verbal recall scores in MHAS, (E) delayed verbal recall scores in CHARLS, (F) delayed verbal recall scores in MHAS, (G) figure recall scores in CHARLS, (H) figure recall scores in MHAS, and (I) serial 7s scores in CHARLS. Values are adjusted for baseline cognitive scores, age, gender, BMI group, marital status, educational level, residence, smoking status, BADL and IADL disabilities, and chronic diseases. BADL: basic activities of daily living; CHARLS: China Health and Retirement Longitudinal Study; IADL: instrumental activities of daily living; MHAS: Mexican Health & Aging Study.

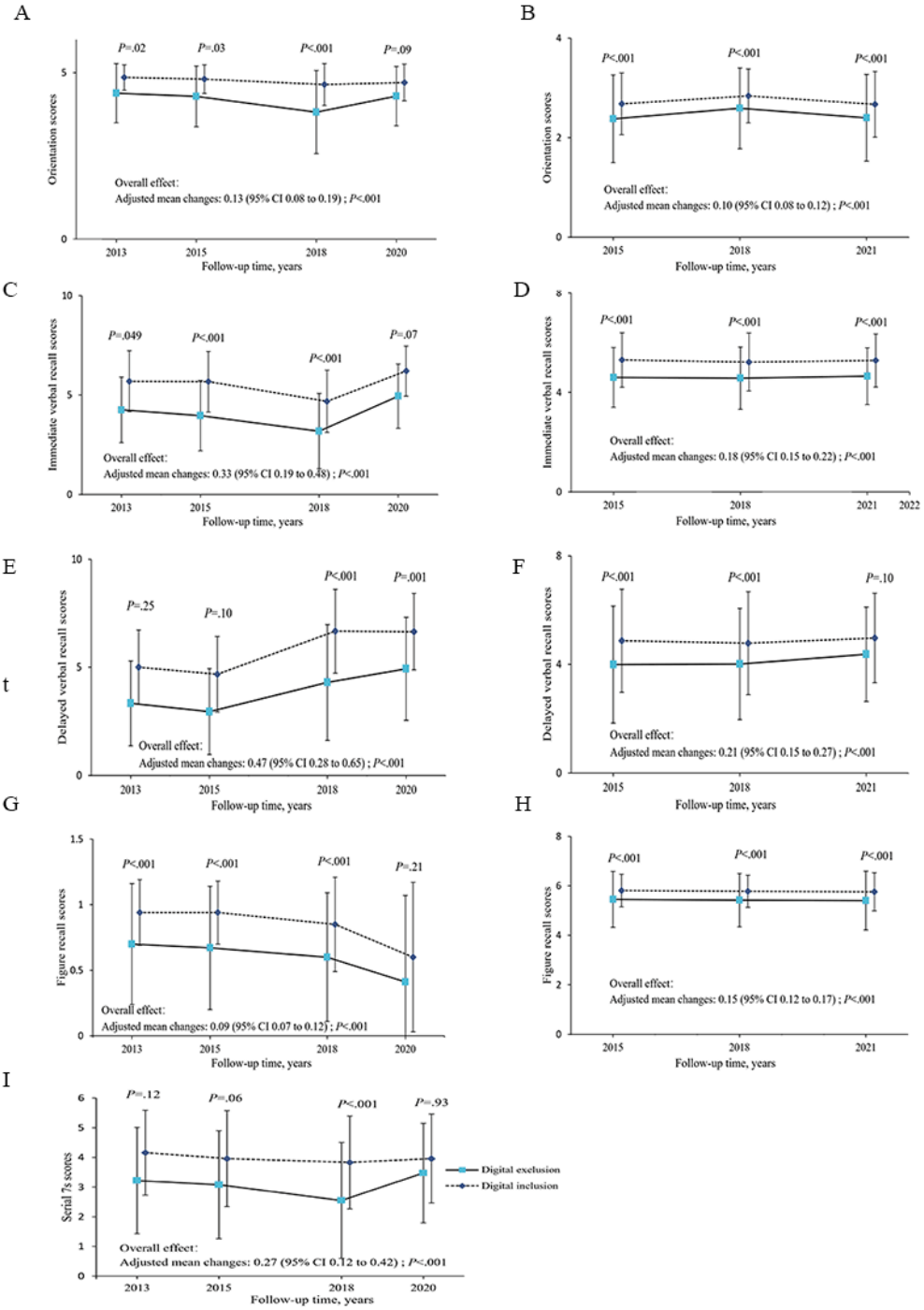

Discussion

Principal Findings

As the prevalence of cognitive dysfunction in older populations increases, with no effective medical treatment available to cure or delay its progression, addressing modifiable risk factors becomes crucial. Considering that up to 40% of dementia cases may be attributable to modifiable risk factors [41], identifying interventions to slow cognitive decline, a potential precursor to dementia, is imperative. Our study examined the longitudinal relationship between digital exclusion and cognitive performance among 11,505 middle-aged and elderly Chinese participants and 12,560 older adults in Mexico. Digital exclusion rates ranged from 96.78% in China and 70.02% in Mexico. We observed that the elderly experiencing digital exclusion were more likely to exhibit worse cognitive function compared to

XSL•FO
RenderX

others, and this association persisted even after adjusting for potential covariates and was found to be independent of gender and age.

Cognitive assessments are crucial for identifying cognitive impairment. The Mini-Mental State Exam (MMSE), which is widely used, has shown limited sensitivity in detecting mild cognitive impairment (MCI) [42]. Conversely, the Montreal Cognitive Assessment (MoCA) is a more effective tool for early MCI detection, incorporating tasks such as memory recall, delayed recall, and visuospatial and language tasks [42]. However, MoCA scores are significantly influenced by the individual's education level [43]. Our study, informed by prior research [28,29,44], evaluated cognition across multiple tests, including memory, executive functioning, visuospatial abilities, and orientation. Generally, the cognitive function of the 2 groups was evaluated based on the orientation test, the immediate/delayed verbal recall test, and the figure recall test. The serial 7s test was not included in MHAS. Due to the diverse cultural backgrounds of the 2 countries, the verbal recall test set different words and total scores in the 2 cohorts, and a similar situation was found in the figure recall test. After standardizing cognitive test scores to a range of 0-1 and distinguishing between higher and lower cognitive scores using a threshold of 0.5, we found that the score distribution of the same type of cognitive test in CHARLS and MHAS is inconsistent. (Figure S3 in Multimedia Appendix 1). Although the differing distribution of cognitive scores across the CHARLS and MHAS waves provides a unique opportunity to generalize our findings across cultures, it also highlights the challenge of maintaining uniformity in test application across diverse cohorts. For future research, standardizing cognitive assessments across varied cohorts will be crucial.

In MHAS, we observed that although the cognitive scores of the digital inclusion group were generally higher than those of the digital exclusion group (follow-up waves 4-6, 2015-2021), there appeared to be no significant difference in cognitive scores among different waves (follow-up waves 4-6, 2015-2021). Through the questionnaire study, we consider that this consistency may be attributed to the uniformity of assessment questions across different waves, which guarantees the stability of assessment results. Repeatedly conducting the same test inevitably leads to the formation of memories, although it remains uncertain whether time can erase those memories. We observed a trend of declining cognitive scores in the tests of orientation, immediate verbal recall, serial 7s, and figure recall in CHARLS (follow-up waves 2-4, 2013-2018). To capture the nuances of cognition, various word lists were used to evaluate immediate and delayed verbal recall in the follow-up waves of CHARLS. In wave 5 (2020), words more aligned with Chinese cultural contexts were introduced, such as substituting "queen" with "president." Additionally, wave 5 (2020) considered dialect tolerability, leading to higher scores in verbally oriented tasks. These methodological adjustments resulted in increased scores in orientation, word recall, and serial 7s in wave 5 compared to earlier waves. Despite these changes, significant disparities in cognitive scores persisted between digitally inclusive and exclusive groups. The use of varied assessment techniques indirectly enhanced the robustness of our findings, underscoring the impact of digital exclusion on poor cognitive performance.

Research in developing countries across Eastern and Western cultures indicates that older individuals facing digital exclusion are at a heightened risk of cognitive decline. This association is especially marked in China, where digital exclusion rates are notably higher. Analysis of CHARLS reveals a progressive widening of the cognitive score gap between digitally excluded and included individuals, particularly evident from follow-up waves 2-4 (Figure 2). In contrast, this gap remained stable across similar periods in MHAS. The observed heterogeneity between cohorts may stem from lower education levels in CHARLS compared to MHAS (85.28% vs 78.78% with lower secondary education) and a higher prevalence of comorbidities, such as arthritis or rheumatism, which may impede functional exercise and social engagement, exacerbating cognitive decline. Further investigation is warranted to explore how digital exclusion may accelerate cognitive decline among the elderly, especially across varying levels of exclusion.

Underlying Mechanisms

The mechanisms linking digital exclusion to diminished cognitive performance are not fully understood, yet several hypotheses have been proposed. One explanation involves the activation of brain regions during internet text reading, such as the left inferior frontal, temporal, posterior cingulate, parietal, and occipital regions, which are associated with language, reading, memory, and visual abilities [45]. Online cognitive function training, particularly for elderly patients at high risk of cognitive decline or dementia, has been shown to enhance cognitive scores, including memory and executive functions, more effectively than traditional face-to-face training [46]. Additionally, internet usage, which often involves searching activities, engages more neural circuitry compared to mere reading of text pages, especially in internet-savvy older adults (aged 55-76 years) [47]. Another interesting avenue of research is the study of astrocytes in mice, which showed that their calcium activity gradually increases when adapting to a new environment, thereby enhancing cognitive ability [48]. Investigating whether a similar process occurs in individuals newly introduced to the internet could provide further insights into the relationship between digital inclusion and cognitive performance. Digital exclusion intensifies the disparities in medical resource distribution, consequently hastening cognitive decline among the elderly [49]. For instance, patients with impaired vestibular function can benefit from online training programs designed for recovery [50,51]. However, elderly individuals who are unfamiliar with or uncomfortable using digital technology face significant barriers in accessing these online medical services. Compounding this, conditions such as hearing loss can further hinder the effectiveness of these digital interventions, potentially accelerating cognitive decline [52]. Moreover, limited internet access restricts opportunities for elderly individuals to make friends online, further isolating them and exacerbating cognitive deterioration [53]. However, internet users are generally more engaged in social activities, which can reduce social loneliness and enhance cognitive functions. To address the imbalanced distribution of medical resources caused by the digital divide among the elderly, the Chinese government

XSL•FO
RenderX

has implemented "integrated medical services and elderly care" (IMSEC). These include the Home IMSEC model, the Community IMSEC model, the Institutional IMSEC model, and the Internet Plus IMSEC model [54]. The Japanese government has proposed a similar policy [55].

Strengths and Limitations

Previous studies have often been limited to cross-sectional analyses focusing on a single follow-up wave [23]. In contrast, our study leveraged longitudinal cohorts to deepen understanding. In MHAS, we assessed cognitive differences between populations affected by digital exclusion and those with data inclusion across 3 follow-up waves. Meanwhile, CHARLS provided 4 waves of follow-up data. Although existing studies have established a link between digital exclusion and cognitive function, our longitudinal approach further bolsters the argument for the relationship in the elderly population. Furthermore, prior research has typically concentrated on single-country analyses [15-17,24,25]. Our study, however, enhanced the generalizability of these findings, using a cross-cultural, longitudinal design with individual-level cohorts from 2 developing countries, China and Mexico. Our study also built on the existing literature by incorporating more follow-up waves and broadening the geographic scope, thus offering a deeper understanding of digital exclusion's impact on cognitive health in older populations. Additionally, the use of GEE models, which account for correlations among multiple waves of longitudinal data, reduces the potential for misestimation. Furthermore, compared to other studies [24,25], our research not only examined the longitudinal changes in participants' cognitive scores over the years but also performed a horizontal comparison of population cognitive function using *Z*-score standardization. It not only considered the overall cognitive level of the population but also combined its own changes, which makes our research conclusions more reliable. Overall, our study expanded upon previous research findings by incorporating additional follow-up waves and broadening the geographical scope of the study areas, thereby offering a more comprehensive understanding of the impact of digital exclusion on cognitive health in older individuals.

The findings of our study were subject to several limitations. First, the assessment of digital exclusion was based on self-reported data. Despite the data collectors being well trained, there remains a potential for information bias due to variations in patients' willingness to provide accurate answers. This variability could have unknown impacts on the observed associations, potentially skewing the results. Second, longitudinal data frequently suffer from incompleteness. Despite efforts to select a representative sample, not all respondents participated in the entire follow-up waves. In CHARLS, 10.47% of the participants were observed for a period of 2 years, 18.61% for 4 years, 30.33% for 7 years, and 40.59% for 9 years. Similarly, in MHAS, 15.45% were followed for 3 years, 18.10% for 6 years, and 66.46% for 9 years. Although previous research has indicated that attrition in cohort studies involving older individuals may not necessarily imply bias, this remains a consideration [56]. Third, although the GEE provides unbiased population-averaged regression coefficients [57], this analysis did not establish a causal relationship between internet use and cognitive abilities, nor did it rule out the possibility of reverse causality, where lower cognitive function might reduce internet use among older adults. Fourth, our study only enrolled 2 cohorts from developing countries. Therefore, the generalizability of our findings to other developing nations or regions with different socioeconomic and cultural contexts needs further evaluation. Although our study adjusted for common factors known to affect cognition—including baseline cognitive level, age, gender, education, marital status, lifestyle, chronic diseases, and basic daily living abilities—there remains the possibility of residual or unmeasured confounding. Factors such as genetic predisposition, cognitive reserve, or additional lifestyle variables could still influence the results. Additionally, our study's design does not permit the establishment of a causal relationship between internet use and cognitive abilities. Internet use is also relevant in terms of social connections and a sense of belonging in society, which in turn could affect cognition. Thus, there is also the potential for reverse causality, as lower cognitive function may decrease the likelihood of internet use among older adults. To clarify these associations, further research involving the design of studies that use instrumental variable analyses may be necessary to rigorously test and potentially demonstrate a causal relationship.

Conclusion

The findings of our study suggest that internet access may be associated with a reduced risk of cognitive deterioration in the elderly, highlighting the potential benefits of tailored digital inclusion strategies to promote active aging. However, there is a need for more robust evidence to substantiate the positive effects of digital inclusion in older adults. Future research should include interventional trials expanded across a broader range of countries and regions. Such studies are essential to develop customized strategies that effectively bridge the digital divide, thereby potentially enhancing cognitive health and overall well-being in aging populations.

Acknowledgments

We would like to acknowledge the China Health and Retirement Longitudinal Study and the Mexican Health & Aging Study for providing data. The study was supported by grants from the Sichuan Science and Technology Program (2023YFS0295).

Data Availability

The China Health and Retirement Longitudinal Study and the Mexican Health & Aging Study data are available online upon request [26,27]. The source codes are available from the corresponding author.

XSL•FO
RenderX

Authors' Contributions

XJ and TZ contributed equally; they designed the study and edited and reviewed the manuscript. SD and DC carried out data extraction and statistical analysis. JW, MSP, and ZX prepared the manuscript. WX and KH searched the literary background and defined the intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 1258 KB - jmir_v26i1e56636_app1.docx]

References

1. Larson EB, Yaffe K, Langa KM. New insights into the dementia epidemic. N Engl J Med 2013 Dec 12;369(24):2275-2277. [doi: 10.1056/nejmp1311405]
2. Chan KY, Wang W, Wu JJ, Liu L, Theodoratou E, Car J, Global Health Epidemiology Reference Group (GHERG). Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990-2010: a systematic review and analysis. Lancet 2013 Jun 08;381(9882):2016-2023 [FREE Full text] [doi: 10.1016/S0140-6736(13)60221-4] [Medline: 23746902]
3. Parra MA, Baez S, Sedeño L, Gonzalez Campo C, Santamaría-García H, Aprahamian I, et al. Dementia in Latin America: paving the way toward a regional action plan. Alzheimers Dement 2021 Feb;17(2):295-313 [FREE Full text] [doi: 10.1002/alz.12202] [Medline: 33634602]
4. Jessen F, Amariglio RE, van Boxtel M, Breteler M, Ceccaldi M, Chételat G, Subjective Cognitive Decline Initiative (SCD-I) Working Group. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. Alzheimers Dement 2014 Nov;10(6):844-852 [FREE Full text] [doi: 10.1016/j.jalz.2014.01.001] [Medline: 24798886]
5. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020 Aug 08;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
6. Kozyreva A, Lewandowsky S, Hertwig R. Citizens versus the internet: confronting digital challenges with cognitive tools. Psychol Sci Public Interest 2020 Dec;21(3):103-156 [FREE Full text] [doi: 10.1177/1529100620946707] [Medline: 33325331]
7. Guo Y, Zu L, Chen D, Zhang H. Research on influencing factors of satisfaction with the use of public health internet platform: evidence from Ding Xiang Doctor (DXY) internet medical platform. Int J Environ Res Public Health 2023 Jan 27;20(3):2276 [FREE Full text] [doi: 10.3390/ijerph20032276] [Medline: 36767643]
8. Seifert A. The digital exclusion of older adults during the COVID-19 pandemic. J Gerontol Soc Work 2020;63(6-7):674-676. [doi: 10.1080/01634372.2020.1764687] [Medline: 32401181]
9. Korkmaz Ö, Erer E, Erer D. Internet access and its role on educational inequality during the COVID-19 pandemic. Telecomm Policy 2022 Jun;46(5):102353 [FREE Full text] [doi: 10.1016/j.telpol.2022.102353] [Medline: 35437338]
10. Lu X, Yao Y, Jin Y. Digital exclusion and functional dependence in older people: findings from five longitudinal cohort studies. EClinicalMedicine 2022 Dec;54:101708 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101708] [Medline: 36353265]
11. Yang Y, Zeng D, Yang F. Internet use and subjective well-being of the elderly: an analysis of the mediating effect based on social capital. Int J Environ Res Public Health 2022 Sep 24;19(19):12087 [FREE Full text] [doi: 10.3390/ijerph191912087] [Medline: 36231387]
12. Zhang H, Wang H, Yan H, Wang X. Impact of internet use on mental health among elderly individuals: a difference-in-differences study based on 2016-2018 CFPS data. Int J Environ Res Public Health 2021 Dec 23;19(1):101 [FREE Full text] [doi: 10.3390/ijerph19010101] [Medline: 35010361]
13. Yang H, Wu Y, Lin X, Xie L, Zhang S, Zhang S, et al. Internet use, life satisfaction, and subjective well-being among the elderly: evidence from 2017 China General Social Survey. Front Public Health 2021;9:677643 [FREE Full text] [doi: 10.3389/fpubh.2021.677643] [Medline: 34268289]
14. Li P, Zhang C, Gao S, Zhang Y, Liang X, Wang C, et al. Association between daily internet use and incidence of chronic diseases among older adults: prospective cohort study. J Med Internet Res 2023 Jul 17;25:e46298 [FREE Full text] [doi: 10.2196/46298] [Medline: 37459155]
15. Xavier AJ, d'Orsi E, de Oliveira CM, Orrell M, Demakakos P, Biddulph JP, et al. English longitudinal study of aging: can internet/e-mail use reduce cognitive decline? J Gerontol A Biol Sci Med Sci 2014 Sep;69(9):1117-1121 [FREE Full text] [doi: 10.1093/gerona/glu105] [Medline: 25116923]
16. Almeida OP, Yeap BB, Alfonso H, Hankey GJ, Flicker L, Norman PE. Older men who use computers have lower risk of dementia. PLoS One 2012;7(8):e44239 [FREE Full text] [doi: 10.1371/journal.pone.0044239] [Medline: 22937167]
17. Tun PA, Lachman ME. The association between computer use and cognition across adulthood: use it so you won't lose it? Psychol Aging 2010 Sep;25(3):560-568 [FREE Full text] [doi: 10.1037/a0019543] [Medline: 20677884]

XSL•FO
RenderX

18. Shen Y, Wang F, Zhang X, Zhu X, Sun Q, Fisher E, et al. Effectiveness of internet-based interventions on glycemic control in patients with type 2 diabetes: meta-analysis of randomized controlled trials. J Med Internet Res 2018 May 07;20(5):e172 [FREE Full text] [doi: 10.2196/jmir.9133] [Medline: 29735475]
19. McManus RJ, Little P, Stuart B, Morton K, Raftery J, Kelly J, HOME BP Investigators. Home and Online Management and Evaluation of Blood Pressure (HOME BP) using a digital intervention in poorly controlled hypertension: randomised controlled trial. BMJ 2021 Jan 19;372:m4858 [FREE Full text] [doi: 10.1136/bmj.m4858] [Medline: 33468518]
20. Green BB, Cook AJ, Ralston JD, Fishman PA, Catz SL, Carlson J, et al. Effectiveness of home blood pressure monitoring, web communication, and pharmacist care on hypertension control: a randomized controlled trial. JAMA 2008 Jun 25;299(24):2857-2867 [FREE Full text] [doi: 10.1001/jama.299.24.2857] [Medline: 18577730]
21. Duplaga M, Szulc K. The association of internet use with wellbeing, mental health and health behaviours of persons with disabilities. Int J Environ Res Public Health 2019 Sep 04;16(18):3252 [FREE Full text] [doi: 10.3390/ijerph16183252] [Medline: 31487949]
22. The Lancet Healthy Longevity. Tackling the digital divide. Lancet Healthy Longev 2021 Oct;2(10):e601. [doi: 10.1016/s2666-7568(21)00233-6]
23. Liu X, Wang X, Zhang H, Pei M, Li N. Relationship between digital exclusion and cognitive impairment in Chinese adults. Front Aging Neurosci 2023 Jul 3;15:1194348 [FREE Full text] [doi: 10.3389/fnagi.2023.1194348] [Medline: 37465320]
24. Li W, Yue L, Xiao S. Association between internet use, cognitive function, and globus pallidus volumes: a study among the elderly in Chinese communities. Front Public Health 2022;10:886974 [FREE Full text] [doi: 10.3389/fpubh.2022.886974] [Medline: 35646776]
25. Jin Y, Jing M, Ma X. Effects of digital device ownership on cognitive decline in a middle-aged and elderly population: longitudinal observational study. J Med Internet Res 2019 Jul 29;21(7):e14210 [FREE Full text] [doi: 10.2196/14210] [Medline: 31359864]
26. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol 2014 Feb;43(1):61-68 [FREE Full text] [doi: 10.1093/ije/dys203] [Medline: 23243115]
27. Wong R, Michaels-Obregon A, Palloni A. Cohort profile: the Mexican Health and Aging Study (MHAS). Int J Epidemiol 2017 Apr 01;46(2):e2 [FREE Full text] [doi: 10.1093/ije/dyu263] [Medline: 25626437]
28. Yao Y, Wang K, Xiang H. Association between cognitive function and ambient particulate matters in middle-aged and elderly Chinese adults: evidence from the China Health and Retirement Longitudinal Study (CHARLS). Sci Total Environ 2022 Jul 01;828:154297 [FREE Full text] [doi: 10.1016/j.scitotenv.2022.154297] [Medline: 35288137]
29. Lin L, Cao B, Chen W, Li J, Zhang Y, Guo VY. Association of adverse childhood experiences and social isolation with later-life cognitive function among adults in China. JAMA Netw Open 2022 Nov 01;5(11):e2241714 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.41714] [Medline: 36367722]
30. Levine DA, Galecki AT, Langa KM, Unverzagt FW, Kabeto MU, Giordani B, et al. Trajectory of cognitive decline after incident stroke. JAMA 2015 Jul 07;314(1):41-51 [FREE Full text] [doi: 10.1001/jama.2015.6968] [Medline: 26151265]
31. Fuentes-García A. Katz activities of daily living scale. In: Michalos AC, editor. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer Netherlands; 2014:3465-3468.
32. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist 1969;9(3):179-186. [Medline: 5349366]
33. Wu C. The mediating and moderating effects of depressive symptoms on the prospective association between cognitive function and activities of daily living disability in older adults. Arch Gerontol Geriatr 2021;96:104480 [FREE Full text] [doi: 10.1016/j.archger.2021.104480] [Medline: 34274875]
34. Madlock-Brown C, Wilkens K, Weiskopf N, Cesare N, Bhattacharyya S, Riches NO, et al. Clinical, social, and policy factors in COVID-19 cases and deaths: methodological considerations for feature selection and modeling in county-level analyses. BMC Public Health 2022 Apr 14;22(1):747 [FREE Full text] [doi: 10.1186/s12889-022-13168-y] [Medline: 35421958]
35. Fox J, Weisberg S. An R Companion to Applied Regression. Thousand Oaks, CA: SAGE Publications; 2011.
36. Nahhas RW. Introduction to Regression Methods for Public Health Using R. London, UK: Chapman & Hall; 2024.
37. Liu Y, Cao X, Gu N, Yang B, Wang J, Li C. A prospective study on the association between grip strength and cognitive function among middle-aged and elderly Chinese participants. Front Aging Neurosci 2019 Sep 10;11:250 [FREE Full text] [doi: 10.3389/fnagi.2019.00250] [Medline: 31551762]
38. Kroon FPB, Kortekaas MC, Boonen A, Böhringer S, Reijnierse M, Rosendaal FR, et al. Results of a 6-week treatment with 10 mg prednisolone in patients with hand osteoarthritis (HOPE): a double-blind, randomised, placebo-controlled trial. Lancet 2019 Nov;394(10213):1993-2001. [doi: 10.1016/s0140-6736(19)32489-4]
39. Chou C, Liaw J, Chen C, Liou Y, Wang C. Effects of a case management program for women with pregnancy-induced hypertension. J Nurs Res 2021 Aug 25;29(5):e169. [doi: 10.1097/jnr.0000000000000450] [Medline: 34432727]
40. Hua J, Qiao Y, Ke C, Shen Y. Higher visit-to-visit total cholesterol variability is associated with lower cognitive function among middle-aged and elderly Chinese men. Sci Rep 2020 Sep 23;10(1):15555 [FREE Full text] [doi: 10.1038/s41598-020-72601-7] [Medline: 32968174]

XSL•FO
RenderX

41. Hafdi M, Hoevenaar-Blom MP, Richard E. Multi-domain interventions for the prevention of dementia and cognitive decline. Cochrane Database Syst Rev 2021 Nov 08;11(11):CD013572 [FREE Full text] [doi: 10.1002/14651858.CD013572.pub2] [Medline: 34748207]
42. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
43. Adana Díaz L, Arango A, Parra C, Rodríguez-Lorenzana A, Yacelga-Ponce T. Impact of educational level on versions (basic and complete) of the Montreal Cognitive Assessment. Dement Geriatr Cogn Disord 2021 Sep 28;50(4):341-348. [doi: 10.1159/000518747] [Medline: 34808623]
44. Jessen F, Amariglio RE, Buckley RF, van der Flier WM, Han Y, Molinuevo JL, et al. The characterisation of subjective cognitive decline. Lancet Neurol 2020 Mar;19(3):271-278 [FREE Full text] [doi: 10.1016/S1474-4422(19)30368-0] [Medline: 31958406]
45. Small GW, Lee J, Kaufman A, Jalil J, Siddarth P, Gaddipati H, et al. Brain health consequences of digital technology use . Dialogues Clin Neurosci 2020 Jun;22(2):179-187 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
46. Cuevas H, Carter S. Online cognitive training: an adaptation of the memory, attention, and problem solving skills for persons with diabetes intervention. Comput Inform Nurs 2020 Aug 10;39(3):162-169. [doi: 10.1097/CIN.0000000000000663] [Medline: 32769625]
47. Small GW, Moody TD, Siddarth P, Bookheimer SY. Your brain on Google: patterns of cerebral activation during internet searching. Am J Geriatr Psychiatry 2009 Feb;17(2):116-126. [doi: 10.1097/jgp.0b013e3181953a02]
48. Doron A, Rubin A, Benmelech-Chovav A, Benaim N, Carmi T, Refaeli R, et al. Hippocampal astrocytes encode reward location. Nature 2022 Sep 31;609(7928):772-778. [doi: 10.1038/s41586-022-05146-6] [Medline: 36045289]
49. Merkel S, Hess M. The use of internet-based health and care services by elderly people in Europe and the importance of the country context: multilevel study. JMIR Aging 2020 Jun 03;3(1):e15491 [FREE Full text] [doi: 10.2196/15491] [Medline: 32490837]
50. Geraghty AWA, Kirby S, Essery R, Little P, Bronstein A, Turner D, et al. Internet-based vestibular rehabilitation for adults aged 50 years and over: a protocol for a randomised controlled trial. BMJ Open 2014 Jul 22;4(7):e005871 [FREE Full text] [doi: 10.1136/bmjopen-2014-005871] [Medline: 25052178]
51. van Vugt VA, van der Wouden JC, Bosmans JE, Smalbrugge M, van Diest W, Essery R, et al. Guided and unguided internet-based vestibular rehabilitation versus usual care for dizzy adults of 50 years and older: a protocol for a three-armed randomised trial. BMJ Open 2017 Jan 20;7(1):e015479 [FREE Full text] [doi: 10.1136/bmjopen-2016-015479] [Medline: 28110290]
52. Slade K, Plack CJ, Nuttall HE. The effects of age-related hearing loss on the brain and cognitive function. Trends Neurosci 2020 Oct;43(10):810-821 [FREE Full text] [doi: 10.1016/j.tins.2020.07.005] [Medline: 32826080]
53. Kuiper JS, Zuidersma M, Zuidema SU, Burgerhof JGM, Stolk RP, Oude Voshaar RC, et al. Social relationships and cognitive decline: a systematic review and meta-analysis of longitudinal cohort studies. Int J Epidemiol 2016 Aug 06;45(4):1169-1206. [doi: 10.1093/ije/dyw089] [Medline: 27272181]
54. Qin S, Zhou M, Cheng Y, Zhao J, Ding Y. Choice preference of middle-aged and elderly people on integrated medical services and elderly care model: a cross-sectional study. Inquiry 2024;61:469580231224345 [FREE Full text] [doi: 10.1177/00469580231224345] [Medline: 38281995]
55. Sudo K, Kobayashi J, Noda S, Fukuda Y, Takahashi K. Japan's healthcare policy for the elderly through the concepts of self-help (Ji-jo), mutual aid (Go-jo), social solidarity care (Kyo-jo), and governmental care (Ko-jo). Biosci Trends 2018 Mar 18;12(1):7-11 [FREE Full text] [doi: 10.5582/bst.2017.01271] [Medline: 29479017]
56. Lacey RJ, Jordan KP, Croft PR. Does attrition during follow-up of a population cohort study inevitably lead to biased estimates of health status? PLoS One 2013 Dec 30;8(12):e83948 [FREE Full text] [doi: 10.1371/journal.pone.0083948] [Medline: 24386313]
57. Ghisletta P, Spini D. An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. J Educ Behav Stat 2004 Dec 01;29(4):421-437. [doi: 10.3102/10769986029004421]

Abbreviations

aGSIF: adjusted generalized standard error inflation factor
BADL: basic activities of daily living
CHARLS: China Health and Retirement Longitudinal Study
GEE: generalized estimating equation
GVIF: generalized variance inflation factor
IADL: instrumental activities of daily living
IMSEC: integrated medical services and elderly care

XSL•FO
RenderX

MCI: mild cognitive impairment
MHAS: Mexican Health & Aging Study
MoCA: Montreal Cognitive Assessment
OR: odds ratio

Edited by A Coristine; submitted 24.01.24; peer-reviewed by A Azizan, P Hwang, L Cachon-Alonso; comments to author 29.05.24; revised version received 10.07.24; accepted 16.10.24; published 15.11.24.

Please cite as:
Duan S, Chen D, Wang J, Paramboor MS, Xia Z, Xu W, Han K, Zhu T, Jiang X
Digital Exclusion and Cognitive Function in Elderly Populations in Developing Countries: Insights Derived From 2 Longitudinal Cohort Studies
J Med Internet Res 2024;26:e56636
URL: https://www.jmir.org/2024/1/e56636
doi:10.2196/56636
PMID:

©Sainan Duan, Dongxu Chen, Jinping Wang, Mohammed Sharooq Paramboor, Zhen Xia, Wanting Xu, Kun Han, Tao Zhu, Xiaoqin Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Investigating the Effectiveness of Technology-Based Distal Interventions for Postpartum Depression and Anxiety: Systematic Review and Meta-Analysis

Sarah P Brocklehurst[1], BPsych (Hons), MClin Psych; Alyssa R Morse[2], BPsych (Hons), PhD; Tegan Cruwys[1], PhB (Hons), MClin Psych, PhD; Philip J Batterham[2], BPsych (Hons), MPH, PhD; Liana Leach[3], BAppPsych (Hons), PhD; Alysia M Robertson[1], BPsych (Hons); Aseel Sahib[1], BPsych (Hons); Colette T Burke[1], BPsych (Hons); Jessica Nguyen[1], BPsych (Hons), MClin Psych; Alison L Calear[2], BAppPsych (Hons), PhD

[1]School of Medicine and Psychology, The Australian National University, Canberra, Australia

[2]Centre for Mental Health Research, The Australian National University, Canberra, Australia

[3]National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia

Corresponding Author:
Alison L Calear, BAppPsych (Hons), PhD
Centre for Mental Health Research
The Australian National University
63 Eggleston Road
Acton
Canberra, 2601
Australia
Phone: 61 2 6125 8406
Fax: 61 2 6125 0733
Email: alison.calear@anu.edu.au

Abstract

Background: Postpartum anxiety and depression are common in new parents. While effective interventions exist, they are often delivered in person, which can be a barrier for some parents seeking help. One approach to overcoming these barriers is the delivery of evidence-based self-help interventions via websites, smartphone apps, and other digital media.

Objective: This study aims to evaluate the effectiveness of technology-based distal interventions in reducing or preventing symptoms of postpartum depression or anxiety in male and female birth and adoptive parents, explore the effectiveness of technology-based distal interventions in increasing social ties, and determine the level of adherence to and satisfaction with technology-based distal interventions.

Methods: A systematic review and series of meta-analyses were conducted. Three electronic bibliographic databases (PsycINFO, PubMed, and Cochrane Library) were searched for randomized controlled trials evaluating technology-based distal interventions for postpartum depression or anxiety in birth and adoptive parents. Searches were updated on August 1, 2023, before conducting the final meta-analyses. Data on trial characteristics, effectiveness, adherence, satisfaction, and quality were extracted. Screening and data extraction were conducted by 2 reviewers. Risk of bias was assessed using the Joanna Briggs Institute quality rating scale for randomized controlled trials. Studies were initially synthesized qualitatively. Where possible, studies were also quantitatively synthesized through 5 meta-analyses.

Results: Overall, 18 articles met the inclusion criteria for the systematic review, with 14 (78%) providing sufficient data for a meta-analysis. A small significant between-group effect on depression favored the intervention conditions at the postintervention (Cohen d=–0.28, 95% CI –0.41 to –0.15; P<.001) and follow-up (Cohen d=–0.27, 95% CI –0.52 to –0.02; P=.03) time points. A small significant effect on anxiety also favored the intervention conditions at the postintervention time point (Cohen d=–0.29, 95% CI –0.48 to –0.10; P=.002), with a medium effect at follow-up (Cohen d=–0.47, 95% CI –0.88 to –0.05; P=.03). The effect on social ties was not significant at the postintervention time point (Cohen d=0.04, 95% CI –0.12 to 0.21; P=.61). Effective interventions tended to be web-based cognitive behavioral therapy programs with reminders. Adherence varied considerably between studies, whereas satisfaction tended to be high for most studies.

XSL•FO
RenderX

Conclusions: Technology-based distal interventions are effective in reducing symptoms of postpartum depression and anxiety in birth mothers. Key limitations of the reviewed evidence include heterogeneity in outcome measures, studies being underpowered to detect modest effects, and the exclusion of key populations from the evidence base. More research needs to be conducted with birth fathers and adoptive parents to better ascertain the effectiveness of interventions in these populations, as well as to further assess the effect of technology-based distal interventions on social ties.

Trial Registration: PROSPERO CRD42021290525; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=290525

(J Med Internet Res 2024;26:e53236) doi:10.2196/53236

KEYWORDS

postpartum; depression; anxiety; birth; adoptive; parents; mobile phone

Introduction

Background

Postpartum depression and anxiety can be experienced by any parent, including birth or adoptive mothers and fathers, as they welcome their new child into their family [1]. Postpartum depression and anxiety are defined in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, as a major depressive disorder or generalized anxiety disorder with postpartum onset [2]. In clinical settings and research, it is well accepted that symptom onset can occur up to 12 months after birth or adoption [3,4].

Postpartum depression is quite common. For birth parents, it is estimated that approximately 10% of fathers [5] and 13% of mothers [6] experience postpartum depression. For adoptive parents, it impacts approximately 11% of fathers [7] and 8.8% of mothers [6]. Postpartum anxiety is also relatively common, impacting approximately 4.4% to 10.8% of parents [8]. There are many risk factors that contribute to parents' vulnerability to developing postpartum depression and anxiety symptoms. These include low self-esteem; low income; history of mood disorders; young age; a negative cognitive attributional style; and stressful life events, including marital strain, past miscarriage, and childhood sexual abuse [9]. New parents are also often at higher risk of social isolation due to fatigue and limited spare time [10]. This social isolation may negatively affect their mental health, exacerbating the chances of developing postpartum depression or anxiety. The development of anxiety and depression symptoms may also increase social withdrawal due to feelings of incompetence and worthlessness, creating a cycle in which poorer mental health and social isolation fuel each other [10].

Preventing and treating postpartum depression and anxiety is paramount as they can have significant short- and long-term effects on parents and children [11]. Well-evidenced short-term effects include sleep disturbance [12], poorer parent-infant attachment, and partner relationship dissatisfaction [11], whereas long-term effects can include poorer cognitive development for the infant [13], the breakdown of close relationships [11], and challenges in parental responsiveness to infant cues such as facial expressions [14].

There is a large body of research on interventions to prevent and treat postpartum depression and anxiety. This research provides support for a number of therapeutic approaches aimed at preventing and reducing symptoms, including cognitive behavioral therapy (CBT) [15] and interpersonal psychotherapy (IPT) [16]. In addition, some interventions have targeted social isolation as their mechanism of action to reduce postpartum depression and anxiety symptoms [17].

While effective interventions for postpartum depression and anxiety exist, they are often delivered in person or with the direct and immediate involvement of a therapist (eg, over videoconferencing, phone calls, or SMS text messaging). The need for client and therapist interaction to be in person or in real time creates barriers for many parents who need help. These barriers include not having the transport to get to a session, not having a babysitter or feeling uncomfortable leaving their child to go to the sessions or take a phone call, and the stigma attached to seeking help [18]. These barriers potentially reduce the number of parents receiving evidence-based treatments for their postpartum depression and anxiety [18]. In addition, professional mental health services are often overwhelmed, resulting in challenges obtaining an appointment for in-person therapy [19]. Therefore, it is important to identify and develop more cost- and resource-effective interventions to increase accessibility.

One approach to overcome these barriers is the delivery of evidence-based self-help via websites, smartphone apps, and chatbot interventions. These can be completed at a time and place most convenient for the user, reducing the need for travel or fitting appointments around feeding and sleeping schedules. Some of these interventions have been empirically tested, and there is evidence that these methods can be effective in reducing symptoms of postpartum anxiety and depression [20] and tend to be well accepted by users [21].

This Research

The aims of this systematic review and meta-analysis were to identify and assess the effectiveness of technology-based distal interventions for postpartum depression and anxiety in male and female birth and adoptive parents. We defined distal interventions as those that are delivered remotely without the direct and immediate input of a therapist or support person. Distal interventions could include podcasts, mobile apps, automated chatbots, and self-help web-based programs. Technology-based interventions were defined as programs that are accessed or downloaded through the internet. The effectiveness of each intervention to reduce symptoms of postpartum depression and anxiety, as well as the adherence to and satisfaction with the intervention, was assessed. Given the important role of social isolation in the development of postpartum depression and anxiety, this review also aimed to

XSL•FO
RenderX

explore whether technology-based distal interventions for postpartum depression and anxiety are effective in increasing social ties. Although multiple systematic reviews and meta-analyses have been conducted on technology-based interventions for postnatal anxiety and depression, they tend to only include birth mothers, and none focus on distal interventions [22-24]. Therefore, this is the first study to review the effectiveness of distal and technology-based interventions on birth and adoptive mothers and fathers.

Methods

Protocol and Registration

This review was prospectively registered with PROSPERO (CRD42021290525; December 10, 2021).

Search Strategy

A comprehensive literature search was conducted in the following electronic bibliographic databases: PsycINFO, PubMed, and Cochrane Library. The search included a combination of five key blocks of terms related to the main objective of the review: (1) anxiety or depression (eg, depress* and anxi*), (2) postnatal (eg, perinatal and parent), (3) intervention (eg, program and cognitive therapy), (4) distal and technology based (eg, remote and web*site*), and (5) trial (eg, experiment and evaluation). Search terms were generated, trialed, and revised by the research team, with additional search terms identified from relevant research literature. Terms were entered in the appropriate search fields (eg, title, abstract, keywords or text words, and subject headings) and adapted to meet the requirements of each database. Multimedia Appendix 1 provides a full search strategy example, and a complete list of search terms can be found in the preregistration.

There were no publication date restrictions. Only English-language studies were included. One reviewer ran the search in all databases. The reference lists of the included studies and previous literature reviews in this field were hand searched. An exploratory search was conducted on November 30, 2021, and the searches were updated on April 17, 2023, and again on August 1, 2023, before conducting the final meta-analyses.

Study Selection

Studies were considered for inclusion in the review and meta-analysis if (1) they were a randomized controlled trial (RCT) where an active intervention was compared to a treatment-as-usual, no intervention, waitlist, or attention control condition; (2) the evaluated intervention was primarily designed to reduce or prevent the depression or anxiety of parents of any gender, age, or nationality who were in the postpartum period (ie, first 12 months) following the birth of their child, including birth and adoptive mothers and fathers; (3) the intervention was delivered distally in the community without the direct and immediate input of a therapist or support person and was technology based (eg, self-help website, podcast, or mobile app; communication could be included if it was automated, infrequent [ie, not scheduled meetings], or optional [eg, SMS text message reminders]); (4) the intervention was self-guided but delivered therapeutic content (eg, CBT or psychoeducation); (5) the primary outcome of the study was the effect of the intervention on symptoms of anxiety or depression (collected via either validated self-report measures or clinical interview) at the postintervention and follow-up time points (if measured); and (6) the study was published in an English-language peer-reviewed journal. Noninferiority RCTs were excluded to ensure comparability of effect sizes. Gray literature, non–peer-reviewed journal articles, and book chapters were excluded, as were conference abstracts and proceedings, dissertations, editorials, viewpoints, perspectives, reviews, and commentaries.

Data Extraction

This systematic review is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Multimedia Appendix 2). The screening, risk-of-bias assessment, and data extraction were managed using the Covidence software (Veritas Health Innovation) [25].

The titles and abstracts of all studies were screened independently by the first author and a second reviewer to identify studies that potentially met the inclusion criteria. All reviewers were briefed and provided with the detailed review protocol before screening abstracts. The full texts of potentially eligible studies were retrieved and independently assessed for eligibility by the first author and a second reviewer in accordance with the review protocol. Any discrepancies between reviewers were resolved through discussion and, if necessary, consultation with a third reviewer. Studies that did not meet the inclusion criteria were excluded, and the reasons for exclusion were recorded.

Data were extracted from the included studies for assessment of study quality and evidence synthesis. Extracted information included study details (authors, year of study, country of study, research design, and type of control); participant characteristics (age, gender, sample size, and recruitment setting or method); intervention details (content, mode of delivery, number of sessions, and length of sessions); outcome measures and time points; and primary (ie, depression and anxiety symptoms) and secondary (ie, social ties) outcome data, including effect sizes, adherence and completion rates (eg, percentage of modules completed), and satisfaction scores (eg, self-report Likert scales). The secondary outcome, social ties, included measures of social support, social isolation, belonging, loneliness, social participation, social capital, and social functioning. The first author and a second reviewer extracted the data independently using a coding form, and any discrepancies identified were resolved through discussion (with a third reviewer where necessary). The pro forma coding sheet was tested for clarity before implementation, with all reviewers provided with a briefing before data extraction.

Risk-of-Bias Assessment

In total, 2 reviewers independently assessed the risk of bias of the included studies using the Joanna Briggs Institute quality rating scale for RCTs [26]. The following two items were removed: (1) “Were participants blind to treatment assignment” and (2) “Were those delivering the treatment blind to treatment assignment.” These items were not relevant for this review as

XSL•FO
RenderX

the distal nature of the interventions meant that participants were automatically unblinded to their condition.

Effect Size Calculations

Effect sizes were calculated for depression, anxiety, and social ties at the postintervention and follow-up time points (where measured and reported). The final follow-up period reported was chosen for effect size calculations. Furthermore, where multiple measures of depression and anxiety were used, the most commonly used measures in research and practice were chosen for effect size calculations (ie, Edinburgh Postnatal Depression Scale [EPDS] and Generalized Anxiety Disorder–7). Cohen *d* [27] was calculated by 3 reviewers and cross checked with the meta-analysis output. Cohen [27] reports the conventional levels of effect sizes as small (Cohen d=0.2), medium (Cohen d=0.5), and large (Cohen d=0.8).

Data Synthesis and Meta-Analysis

The included studies were initially synthesized qualitatively, with a narrative summary describing the main characteristics and results (including effect sizes). The summary focused on participant characteristics, information about the intervention, and risk of bias. On the basis of available data, studies were also quantitatively synthesized through 5 meta-analyses of intervention effects—for depression and anxiety at the postintervention and follow-up time points and for social ties only at the postintervention time point (there were insufficient studies with follow-up time points for this outcome). The meta-analyses yielded statistical summaries of the effects of the interventions at the different time points. A total of 78% (14/18) of the studies were eligible for inclusion in at least one of the meta-analyses, with varying numbers in each analysis based on whether the study measured and reported the required data. Where a journal article did not provide the required data, the authors were contacted twice over a 1-month period with a request for this information. The authors of 2 studies did not respond, and one did not provide appropriate information due to incompatible statistical methods; therefore, these studies were excluded from the meta-analyses.

RevMan (The Cochrane Collaboration) [28] was used to conduct the meta-analyses, which all used a random-effects model due to the expected high heterogeneity, allowing for the differences between the studies to be modeled [29]. In addition, the standardized mean difference (SMD) was used as multiple measurement tools were implemented to measure depression, anxiety, and social ties in the 14 studies [30]. Between-group heterogeneity was tested using the I^2 statistic in RevMan. The I^2 statistic and accompanying CIs indicate the level of difference between the studies by reporting heterogeneity as a percentage [31]. The levels were categorized as low (I^2=25%), moderate (I^2=50%), and high (I^2=75%) heterogeneity [31]. Subgroup analyses were considered given the high level of heterogeneity; however, it was deemed not appropriate due to the insufficient number of studies. High heterogeneity was expected due to the many differences between the studies, including the measures used, intervention length, onset of the intervention, follow-up timing, country of origin, intervention formats, content, theoretical framework, and the fact that 11% (2/18) of the studies included couples [32]. Funnel plots were also created in RevMan to assess for publication bias, which is evident if the funnel plot is asymmetrical [33].

Results

Study Selection

Figure 1 presents the PRISMA flow diagram showing the flow of studies through the different phases of the systematic review, including identification, screening, eligibility, and inclusion. The 3 searches (original and 2 updates) identified a total of 6487 studies, of which 2192 (33.79%) were excluded as duplicates. All the remaining papers were screened and coded by 2 reviewers. A total of 4295 articles were screened by their title and abstract, with 4161 (96.88%) being deemed irrelevant and, therefore, excluded. Following screening, 134 full-text articles were collected and reviewed, of which 116 (86.6%) were excluded, and the reasons for exclusion were noted. In total, 18 papers were deemed eligible and coded for inclusion in the systematic review. Of the 18 included studies, 14 (78%) provided sufficient data for inclusion in at least one of the meta-analyses.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. RCT: randomized controlled trial.

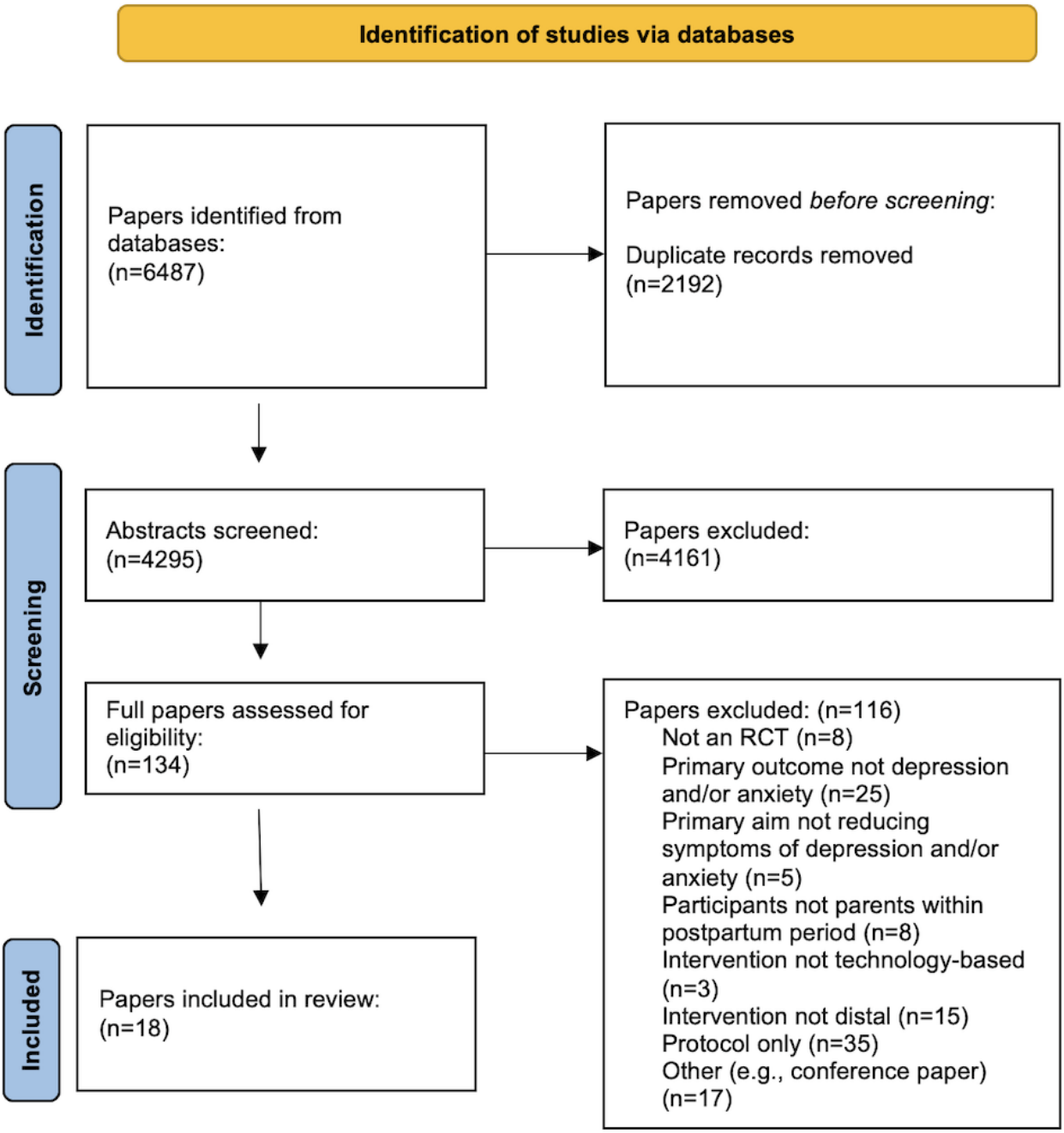

Intervention Characteristics

Overview

Table 1 shows the intervention characteristics of the included studies. Overall, there were 18 studies deemed eligible for this review. This included studies conducted in 12 different regions between 2013 and 2023, with the largest number of studies conducted in Australia (3/18, 17%) and China (3/18, 17%).

XSL•FO
RenderX

Table 1. Intervention characteristics.

Intervention	Study	Region	Content	Length	Delivery mode	Host	Automated, infrequent, or optional communication
Mothers and Babies Course (Curso Mamas y Bebes)	Barrera et al [34]	Global sample	CBT[a]	8 lessons	Web	University	No
Parents Interacting with Infants	Boyd et al [35]	United States	CBT and education	8 modules over 8 weeks	Web	Not for profit	Yes: optional forum with other participants and facilitators
Be a Mom	Carona et al [36]	Portugal	CBT	5 modules over 5 weeks	Web	University	Yes: automated reminders and phone calls assessing IT and program questions and difficulties
iParent	Chan et al [37]	Hong Kong	Education	The app could be used from the first visit to the antenatal clinic until childbirth	Mobile app	Hospital	Yes: optional, unscheduled messages from a health professional answering questions related to pregnancy, childbirth, and infant health and care
MomMoodBooster2	Danaher et al [38]	United States	CBT	6 sessions over a 12-week active treatment phase; participants could continue using it for 7 months	Web	University	Yes: 1 nonclinical call to resolve any IT difficulties and 1 call after the intervention for feedback
Be a Mom	Fonseca et al [39]	Portugal	CBT	5 modules; 1 module per week recommended but could be completed at participants' own pace	Web	University	Yes: reminders and emails for technical support
Mindful Self-Compassion on Program	Guo et al [40]	China	CBT	6-week program with 10 hours of training over 36 episodes (6 episodes per week)	Web	Hospital	No
Mamma Mia	Haga et al [41]	Norway	IPT[b] and education	44 sessions over an 11.5-month period	Web	Hospital	No
MamaKits Online	Heller et al [42]	The Netherlands	Education and problem-solving treatment	5 modules over 5 weeks	Web	University and hospital	Yes: reminders and optional, unscheduled messages or phone calls with coaches to answer questions and receive feedback for homework tasks
Baby Steps	Kavanagh et al [43]	Australia	Education	9 modules with no advised pace	Web	University	Yes: reminders from a support person
__[c]	Lennard et al [44]	Australia and New Zealand	Compassion-focused therapy and education	Unlimited access to online resources over 8 weeks	Web	University	Yes: reminders
MUMentum Postnatal	Loughnan et al [45]	Australia	CBT	3 sessions over 6 weeks	Web	University and hospital	Yes: reminders and unscheduled messages from a health professional (only if participants scored highly on distress questionnaires as a safety protocol)
Luna Baby	Nishi et al [46]	Japan	CBT	6 sessions delivered weekly	Mobile app	University and preexisting mobile app	Yes: automated reminders
NetMums	O'Mahen et al [21]	United Kingdom	CBT and behavioral activation	11 sessions over 15 weeks	Web	University	Yes: optional online chat rooms with other participants and health professionals

XSL•FO
RenderX

Intervention	Study	Region	Content	Length	Delivery mode	Host	Automated, infrequent, or optional communication
CareMom	Qin et al [47]	China	CBT	28 daily challenges completed over 4 weeks	Mobile app	Hospital	Yes: reminders
Home but not Alone	Shorey et al [48]	Singapore	Education	App available over 4 weeks	Mobile app	University and hospital	No
Spirits Healing App	Sun et al [49]	China	MBCT[d]	8 sessions over 8 weeks	Mobile app	University and hospital	Yes: control group had personalized communication and intervention group had automated reminders
—	Zhang et al [50]	China	Mindfulness	6 modules delivered weekly	Mobile app	University	Yes: standardized reminders

[a]CBT: cognitive behavioral therapy.

[b]IPT: interpersonal psychotherapy.

[c]Intervention name was not provided.

[d]MBCT: mindfulness-based cognitive therapy.

Content

Of the 18 studies identified in the review, most (11/18, 61%) assessed a CBT-based intervention. A total of 17% (3/18) of the studies combined CBT with another therapeutic approach (eg, CBT with psychoeducation) [35] or delivered specific components of CBT (eg, behavioral activation or mindfulness-based CBT) [49,51]. One study assessed a mindfulness intervention [50]. The remaining studies (6/18, 33%) were education-focused interventions, with 50% (3/6) of these studies combining psychoeducation with non-CBT therapeutic approaches, specifically IPT [41], problem-solving therapy [42], and compassion-focused therapy [44].

Format and Mode of Delivery

All the interventions (18/18, 100%) were delivered to the individual (vs being group based), and most of the studies (12/18, 67%) tested online interventions that were directly accessed or downloaded from the internet. The remaining studies (6/18, 33%) evaluated interventions that were delivered as mobile apps.

Automated, Infrequent, or Optional Communication

Most of the interventions (14/18, 78%) included optional communication with the participants. A significant proportion of this communication was reminders to complete the program (10/18, 56%). A total of 11% (2/18) of the trials [35,51] included optional online forums with other participants and health professionals or facilitators. In addition, 11% (2/18) of the studies [37,42] had messages from health professionals answering questions related to pregnancy and birth. One study (1/18, 6%) [45] allowed for communication with a health professional if the participant expressed high levels of distress in their questionnaire as a safety precaution. In total, 11% (2/18) of the studies [36,38] provided nonclinical phone calls to seek feedback on any technological issues and the participants' experience with the program. The remaining studies (4/18, 22%) provided no communication as part of the intervention or trial.

Intervention Host

Most of the interventions (13/18, 72%) were offered by a university. Of those 13 studies, 4 (31%) were cohosted by a hospital, and 1 (8%) was cohosted by a preexisting mobile app. Of the remaining 5 interventions, 4 (80%) were hosted by hospitals, and 1 (20%) was hosted by a not-for-profit child health organization.

Length of the Intervention

The length of the interventions included in this review was reported in terms of duration or number of intervention modules. Of those that reported the number of modules (15/18, 83%), 8 modules (3/18, 17%), 6 modules (3/18, 17%), and 5 modules (3/18, 17%) were the most common. The remaining 40% (6/15) of the interventions had 3 to 44 modules. Regarding the time allowed to complete the modules, 11% (2/18) of the studies did not specify a suggested time frame, 39% (7/18) of the studies recommended 1 module per week for a set duration, 11% (2/18) of the studies recommended multiple modules per week for a set duration, 22% (4/18) of the studies recommended 1 module over multiple weeks, and 17% (3/18) of the studies provided unlimited access over a set number of weeks. The overall intervention duration ranged from 4 to 11.5 weeks, with 6 weeks being the most common (4/18, 22%) followed by 8 weeks (3/18, 17%) and 5 weeks (3/18, 17%).

Trial Characteristics and Outcomes

Overview

Table 2 shows the trial characteristics and outcomes. Overall, the interventions included 11,802 participants. Of those, 5916 were part of the intervention conditions, and 5886 were part of the control conditions. The sample sizes of each study ranged from 24 to 5017, with a median sample size of 221.

XSL•FO
RenderX

Table 2. Trial characteristics and outcomes.

Study	Recruitment setting	Control group	Gender	Participant (parent) age (y) and child age (months), mean (SD)	Target population	Sample size	Outcomes	Quality rating score
Barrera et al [34]	Online advertising	Attention	100% female	Parent: 30.19 (5.57); child: not reported	Universal: birth mothers of first or subsequent children	Intervention: 57; control: 54	• Depression: small effect (Cohen d=–0.04) at the postintervention time point and medium effect size (Cohen d=–0.54) at the 6-month follow-up. No significant effect (*P*=.11). • Satisfaction: 28/57 of the participants rated the lesson material. In total, 88.9% indicated that the lessons were helpful for managing mood changes, whereas the content was rated as easy to understand (mean 4.12, SD 1.02) and highly useful (mean 4.20, SD 1.10).	8/11
Boyd et al [35]	Pediatric primary care clinic	Attention	100% female	Parent: 26.4 (1.9); child: 2.7 (0.2)	Indicated: birth mothers	Intervention: 12; control: 12	• Depression: large significant effect (Cohen d=–0.82) at the postintervention time point (*P*<.01). • Adherence: all the mothers (100%) in the social media group attended at least one session. Average attendance was 83%, and average participant commenting was 73%. • Satisfaction: the mean ratings of individual sessions ranged from 3.6 to 4.4 out of 5, demonstrating favorable scores. The mean satisfaction score for the overall intervention was 4.54 (SD 0.78; 91%).	6/11
Carona et al [36]	Online advertising	Waitlist and TAU[a]	100% female	Parent: 32.71 (4.53); child: 2.03 (0.96)	Selective: birth mothers of first or subsequent children with or without a history of mental illness	Intervention: 542; control: 511	• Depression: medium significant effect (Cohen d=–0.42) at the postintervention time point (*P*<.001). • Anxiety: medium significant effect (Cohen d=–0.41) at the postintervention time point (*P*<.001).	9/11
Chan et al [37]	Hospital	TAU	100% female	Parent: 31.25 (4.55); child: not reported	Indicated and selective: birth mothers of first children with or without a history of mental illness	Intervention: 330; control: 330	• Depression: small significant effect (Cohen d=–0.13) at the postintervention time point (*P*=.049). • Anxiety: small nonsignificant effect (Cohen d=0.05) at the postintervention time point (*P*=.94).	10/11
Danaher et al [38]	Prenatal clinic and hospital	TAU	100% female	Parent: 31.9 (5.3); child: not reported	Indicated: birth mothers of first and subsequent children with a history of mental illness	Intervention: 96; control:95	• Depression: small significant effect (Cohen d=–0.33) at the postintervention time point (*P*=.003). • Anxiety: medium nonsignificant effect (Cohen d=–0.39) at the postintervention time point (*P*=.28). • Adherence: mean 10.3 (SD 8.7) total program visits; mean 4.3 (SD 2) session visits; 49% viewed all 6 sessions • Satisfaction: 96% rated the program as somewhat to extremely easy to use; 83% rated it as somewhat to extremely helpful; 93% would recommend the program	10/11

XSL•FO
RenderX

Study	Recruitment setting	Control group	Gender	Participant (parent) age (y) and child age (months), mean (SD)	Target population	Sample size	Outcomes	Quality rating score
Fonseca et al [39]	Hospital and online advertising	Waitlist	100% female	Parent: 32.58 (4.82); child: 2 (0.89)	Universal: birth mothers of first or subsequent children with or without a history of mental illness	Intervention: 98; control: 96	• Depression: small nonsignificant effect size (Cohen d=0.01) at the postintervention time point (*P*=.61). • Anxiety: small nonsignificant effect (Cohen d=–0.08) at the postintervention time point (*P*=.90). • Social ties: small nonsignificant effect (Cohen d=–0.14) at the postintervention time point (*P*=.51). • Adherence: 41.8% of the participants completed the program. • Satisfaction: compared to noncompleters, completers were significantly more satisfied (*P*<.001), had a higher intention to use the intervention if needed (*P*<.001), and perceived the program as useful (*P*=.006). No significant difference was reported in the perceived demandingness of using the program (*P*=.62).	8/11
Guo et al [40]	Prenatal clinic	Waitlist	100% female	Parent: 30.6 (5.95); child: not reported	Indicated: birth mothers	Intervention: 157; control: 157	• Depression: significant effect at the 3-month and 1-year follow-ups (*P*<.01). • Anxiety: not reported. • Adherence: the overall attendance rate was 91.8%, with no significant difference between the groups (*P*=.56). • Satisfaction: 95% of the participants filled out the posttest survey, the results of which indicated high acceptability.	9/11
Haga et al [41]	Prenatal clinic and hospital	TAU	100% female	Parent: 31 (4.6); child: not reported	Universal: birth mothers of first or subsequent children; mothers with or without a history of mental illness	Intervention: 678; control: 664	• Depression: small effect size (Cohen d=–0.10) at the postintervention time point. The intervention group showed less severe depression symptoms than participants in the control group on all measurement occasions after baseline. This difference was statistically significant at gestational week 37 (*P*=.008) and 6 weeks post partum (*P*=.03). • Adherence: 33% of the participants in the intervention group completed all 44 sessions, 51% completed ≥36 sessions, and 6% completed no sessions.	9/11
Heller et al [42]	Community advertising, prenatal clinic, online advertising, and obstetricians and midwives	TAU	100% female	Parent: 32.08 (4.61); child: not reported	Indicated: birth mothers of first or subsequent children; mothers with or without a history of mental illness	Intervention: 79; control: 80		10/11

XSL•FO
RenderX

Study	Recruitment setting	Control group	Gender	Participant (parent) age (y) and child age (months), mean (SD)	Target population	Sample size	Outcomes	Quality rating score
							• Depression: small nonsignificant effect sizes reported at the postintervention time point and 6-week postpartum follow-up using the CES-Db (postintervention time point: Cohen d=0.09; 6-week postpartum follow-up: Cohen d=–0.27) and EPDSc (postintervention time point: Cohen d=0.11; 6-week postpartum follow-up: Cohen d=–0.12). • Anxiety: small nonsignificant effect sizes were reported at the postintervention time point (Cohen d=–0.05) and 6-week postpartum follow-up (Cohen d=–0.18). • Adherence: 47% completed all modules. • Satisfaction: 87% were satisfied with the help they received, and 74% would recommend the intervention to others.	
Kavanagh et al [43]	Community advertising, prenatal clinic, hospital, and referral	Attention	50% female and 50% male	Parent: 32.2 (4.4); child: not reported	Universal: birth mothers and fathers (coparenting couples) of first children with or without a history of mental illness	Intervention: 124; control: 124	• Depression: small nonsignificant effect sizes were reported at the postintervention time point for mothers (Cohen d=0.01) and fathers (Cohen d=0.01) and at the 6-month follow-up for mothers (Cohen d=0.14) and fathers (Cohen d=0.14). • Social ties: small nonsignificant effect sizes were reported at the postintervention time point for mothers (Cohen d=0.10) and fathers (Cohen d=0.10) and at the 6-month follow-up for mothers (Cohen d=0.14) and fathers (Cohen d=0.17). • Adherence: 37.3% of participants accessed their programs more than once, with higher rates for mothers (53.6%) than fathers (20.9%). • Satisfaction: satisfaction with the program was high among the participants who accessed it at least once (median satisfaction score 75/100; 92% satisfaction score ≥50/100), with no significant differences due to treatment or parent gender.	9/11
Lennard et al [44]	Online advertising and forums	Waitlist	100% female	Parent: 32.56 (3.96); child: 9.54 (6.59)	Universal: birth mothers of first or subsequent children with or without a history of mental illness	Intervention: 231; control: 239		7/11

XSL•FO
RenderX

Study	Recruitment setting	Control group	Gender	Participant (parent) age (y) and child age (months), mean (SD)	Target population	Sample size	Outcomes	Quality rating score
							• Depression: small nonsignificant effect size (Cohen d=–0.21) at the postintervention time point (*P*=.11). • Anxiety: small nonsignificant effect size (Cohen d=0.02) at the postintervention time point (*P*=.45). • Adherence: 64.9% of the intervention group reported watching the psychoeducational video at least once (18.7% more than once) and did the guided self-compassion exercise (13.2% more than once). • Satisfaction: overall, 69.3% reported attempting to apply the strategies in their own lives. A total of 47.3% reported feeling that they had become more self - compassionate over the study period. Most agreed that self - compassion would be helpful for new mothers coping with challenging birth (90.1%) and breastfeeding (83.5%) experiences. Similarly, 75.8% of women reported that the SMS text message reminders were useful. Finally, 80.2% would recommend this type of intervention to others.	
Loughnan et al [45]	Community advertising, prenatal clinic, online advertising, and online forum	TAU	100% female	Parent: 32.56 (4.53); child: 4.55 (3.05)	Indicated: birth mothers of first or subsequent children with or without a history of mental illness	Intervention: 69; control: 62	• Depression: large significant effect sizes using the PHQ-9d (Cohen d=–1) and EPDS (Cohen d=–0.91) at the postintervention time point (*P*<.05). Large significant effect sizes using the PHQ-9 (Cohen d=–0.86) and EPDS (Cohen d=–0.99) at the 4-week follow-up (*P*<.05). • Anxiety: large significant effect size (Cohen d=–0.78) at the postintervention time point (*P*<.05). A large significant effect size (Cohen d=–1.15) at the 4-week follow-up (*P*<.05). • Adherence: a total of 46 women completed all 3 lessons of treatment (46/61, 75% completion rate). Of those in iCBTe, 82% completed posttreatment questionnaires, and 61% completed follow-up questionnaires. Of those in TAU, 85% and 76% provided posttreatment and follow-up data, respectively. • Satisfaction: 86% judged the quality of the program as good to excellent. A total of 80% reported being mostly to very satisfied with the program	9/11
Nishi et al [46]	Message sent through preexisting mobile app	No intervention	100% female	Parent: 30.44 (4.6); child: not reported		Intervention: 2509; control: 2508		9/11

XSL•FO
RenderX

Study	Recruitment setting	Control group	Gender	Participant (parent) age (y) and child age (months), mean (SD)	Target population	Sample size	Outcomes	Quality rating score
					Universal: birth mothers of first and subsequent children with or without a history of mental illness		• Depression: 0 cases had an MDEf at baseline, and 59 cases (2.35%) in the intervention group and 73 cases (2.91%) in the control group experienced the onset of an MDE during the intervention and 3-month postpartum follow-up period. • Adherence: 37.2% completed all 6 modules. Module completion rates ranged from 55.9% to 79.5%.	
O'Mahen et al [21]	Online advertising	TAU	100% female	Parent: 32.3 (4.7); child: not reported	Indicated: birth mothers of first or subsequent children with or without a history of mental illness	Intervention: 462; control: 448	• Depression: medium significant effect (Cohen d=–0.55) at the postintervention time point (*P*<.01). • Adherence: number of sessions viewed ranged between 95 and 1310, with a decrease each session except for session 7. Engagement with the treatment chat room and online clinic was low. A total of 7% (32/462) of the women posted on the chat room. • Satisfaction: key acceptability endorsements were regarding the flexible and convenient delivery of the treatment and helping women help themselves, although the women also noted that they struggled to keep up with the program.	8/11
Qin et al [47]	Hospital	Waitlist and TAU	100% female	Parent: 31.9 (3.62); child: not reported	Healthy population: birth mothers of first and subsequent children with no history of mental illness	Intervention: 57; control:55	• Depression: medium significant effect (Cohen d=–0.55) at the postintervention time point (*P*=.04). • Anxiety: small nonsignificant effect (Cohen d=–0.29) at the postintervention time point (*P*=.19). • Adherence: 90% completed all 28 daily challenges. • Satisfaction: overall satisfaction—mean 4.58/5 (SD 0.74); recommendation to a friend—mean 4.54/5 (SD 0.80); relatedness to life—mean 4.44/4 (SD 0.62); application of content to life—mean 4.44/5 (SD 0.58)	5/11
Shorey et al [48]	Hospital	TAU	50% female and 50% male	Parent: 32.66 (5.03); child: not reported	Universal: birth mothers and fathers (coparenting couples) of first or subsequent children with or without a history of mental illness	Intervention: 126; control: 124	• Depression: no significant effect. • Social ties: the intervention group had statistically significant improvements for social support at 4 weeks post partum compared with the control group. This occurred regarding support from spouses (*P*<.001) and other sources (*P*<.001). • Satisfaction: most of the participants in the intervention group felt satisfied with the mHealthg app intervention (n=97, 77%). Most of the participants in the intervention group also stated that they benefited from the mHealth app intervention (n=94, 74.6%)	9/11

XSL•FO
RenderX

Study	Recruitment setting	Control group	Gender	Participant (parent) age (y) and child age (months), mean (SD)	Target population	Sample size	Outcomes	Quality rating score
Sun et al [49]	Hospital	Attention	100% female	Parent: 29.91 (4.02); child: not reported	Indicated: birth mothers of first or subsequent children with or without a history of mental illness	Intervention: 84; control: 84	• Depression: a medium nonsignificant effect size was reported (Cohen d=–0.48) at the postintervention time point (*P*=.25), and a small nonsignificant effect size was reported (Cohen d=0.11) at the 6-week postpartum follow-up. • Anxiety: small nonsignificant effect sizes (Cohen d=–0.27) at the postintervention time point (*P*=.75) and at the 6-week postpartum follow-up (Cohen d=–0.08). • Adherence: 8% completed the intervention.	9/11
Zhang et al [50]	Prenatal clinic	TAU	100% female	Parent: 30.29 (4.29); child: 0-6 months	Indicated: birth mothers of first and subsequent children with a history of mental illness	Intervention: 80; control: 80	• Depression: medium significant effect (Cohen d=–0.55) at the postintervention time point (*P*<.001) and final 6-month postpartum follow-up (Cohen d=–0.48; *P*<.001). Significant effects reported at follow-ups—gestational weeks 36-37 (*P*<.001), 6 weeks post partum (*P*<.001), and 3 months post partum (*P*=.001). • Anxiety: large significant effect (Cohen d=–0.85) at the postintervention time point (*P*<.001) and a medium effect (Cohen d=–0.53) at the final 6-month postpartum follow-up (*P*=.03). Significant effects reported at gestational weeks 36-37 (*P*<.001), 6 weeks post partum (*P*<.001), and 3 months post partum (*P*=.02).	10/11

[a]TAU: treatment as usual.

[b]CES-D: Center for Epidemiologic Studies Depression Scale.

[c]EPDS: Edinburgh Postnatal Depression Scale.

[d]PHQ-9: Patient Health Questionnaire–9.

[e]iCBT: internet-based cognitive behavioral therapy.

[f]MDE: major depressive episode.

[g]mHealth: mobile health.

Recruitment Setting and Clinical Interview

Most of the studies (11/18, 61%) used 1 recruitment setting, whereas the remaining studies (7/18, 39%) used a combination of recruitment settings. Across the studies, participants were recruited through hospitals (8/18, 44%), online advertising (7/18, 39%), online forums (2/18, 11%), prenatal clinics (7/18, 39%), community advertising (3/18, 17%), and messaging via an existing mobile app (1/18, 6%). One study included a clinical interview to determine eligibility in the form of a web-based and self-administered World Health Organization Composite International Diagnostic Interview [46].

Participant Age and Gender and Target Population

The target populations varied across the studies. Indicated interventions (ie, delivered to individuals with elevated symptoms) were the most common (8/18, 44%), followed by universal interventions (7/18, 39%), selective interventions (1/18, 6%), combined indicated and selective prevention interventions (1/18, 6%), and interventions targeted at a healthy population (1/18, 6%). The vast majority of studies (16/18, 89%) included mothers only, with only 11% (2/18) of the studies including coparenting mothers and fathers. All 18 studies were targeted at birth parents only, with most (14/18, 78%) including both first-time parents and those with ≥2 children. A small proportion of studies (2/18, 11%) targeted first-time parents only. A total of 33% (6/18) of the studies reported on the age

XSL•FO
RenderX

of the children at the time of the intervention, with the mean age ranging from 0 to 9.54 months. Of these 6 studies, 3 (50%) included children with a mean age of 2 months.

Regarding parental mental health history, most studies (13/18, 72%) indicated that their participants were parents both with a history of mental illness and without. A total of 11% (2/18) of the studies targeted parents with a history of mental illness, and 6% (1/18) of the studies enrolled only participants with no history of mental illness. The remaining studies (2/18, 11%) did not report on these factors. Mean parent age ranged between 26.4 and 39.9 years across the studies.

Control Group and Randomization

Of the 18 studies included, almost half (n=10, 56%) had a "treatment as usual" control group, making it the most common type. Treatment as usual tended to include information about pregnancy and childbirth alongside regular appointments with health professionals. The second most common type of control group was "attention" control groups (4/18, 22%), which involved receiving information about the postpartum period, and waitlist control groups (5/18, 28%). One study included a "no intervention" control, which differed from the waitlist and treatment-as-usual controls by providing no treatment, including after the study. A total of 89% (16/18) of the studies used individual randomization, and the remaining 11% (2/18) of the studies, which included couples, used stratified randomization. Participants completed the interventions individually, including couples, although the latter could discuss the intervention.

Measurement Time Point

The measurement time points varied between the studies. Exclusively pre- and postintervention measures were evident in 44% (8/18) of the studies. The remaining 56% (10/18) of the studies included pre- and postintervention measures in addition to measures at other time points, ranging from 4 weeks to 6 months after the intervention. The studies included preintervention measurements during pregnancy (11/18, 61%) and post partum (7/18, 39%) depending on the timing of the intervention regarding antenatal and postnatal onset.

Outcomes

Depression

Depression was measured in all the studies (18/18, 100%), with the most frequently used scale being the EPDS (14/18, 78%). A total of 29% (4/14) of these studies used additional depression measures, including the Beck Depression Inventory (1/4, 25%), Patient Health Questionnaire–9 (1/4, 25%), and Center for Epidemiologic Studies Depression Scale (2/4, 50%). Another 29% (4/14) of these studies used the Beck Depression Inventory–II (1/4, 25%); Depression, Anxiety, and Stress Scales–21 (DASS-21; 1/4, 25%); Patient Health Questionnaire–9 (1/4, 25%); and World Health Organization Composite International Diagnostic Interview (1/4, 25%).

All 18 studies reported a decrease in symptoms in the intervention group at the postintervention time point. This decrease was significant in 50% (9/18) of the studies when compared to the control groups, with their effect sizes ranging from Cohen d=–0.10 [41] to Cohen d=–0.91 [45] (median –0.55). Among these 9 studies, multiple control group types were used, including attention (n=1, 11%), treatment as usual (n=6, 67%), and a combination of waitlist and treatment as usual (n=2, 22%). A total of 50% (9/18) of the studies reported a follow-up period of between 4 weeks and 6 months, with all indicating a decrease in symptoms of depression in the intervention group. Of these 9 studies, 4 (44%) reported a significant difference compared to the control group, with effect sizes ranging from Cohen d=–0.10 [41] at the 6-month follow-up to Cohen d=–0.99 [45] at the 4-week follow-up (median –0.48).

Anxiety

Anxiety was measured in 61% (11/18) of the studies. The most commonly used anxiety scales were the Generalized Anxiety Disorder–7 (4/11, 36%), DASS-21 (3/11, 27%), and Hospital Anxiety and Depression Scale–A (3/11, 27%). One study used the State-Trait Anxiety Inventory (1/11, 9%).

All studies reported a reduction in symptoms of anxiety in the intervention groups at the postintervention time point, with 27% (3/11) of the studies [36,45,50] reporting significant effects with moderate to large effect sizes relative to the control group (Cohen d=–0.41, Cohen d=–0.78, and Cohen d=–0.85). Of these 3 studies, all (100%) used a treatment-as-usual control condition, and 1 (33%) additionally included a waitlist control condition. In total, 4 studies reported a follow-up period of between 4 weeks and 6 months, of which 2 (50%) [45,50] reported a significant effect (Cohen d=–1.15 at the 4-week follow-up; Cohen d=–0.53 at the 6-month follow-up).

Social Ties

Social ties were measured in 17% (3/18) of the studies. Each study used a different measure, which included the Portuguese version of the Revised Dyadic Adjustment Scale, short Medical Outcomes Study–Social Support Survey, and the Perceived Social Support for Parenting Scale.

Of these studies, only the one by Shorey et al [48] reported a significant increase in social support at the postintervention time point; however, it did not provide sufficient data to calculate the size of the effect. In addition, only the study by Kavanagh et al [43] assessed social ties at the 6-month follow-up, noting a nonsignificant effect, resulting in insufficient studies to complete a meta-analysis for this time point.

Adherence

Adherence was measured in 72% (13/18) of the studies using differing methods. Multiple studies (9/18, 50%) assessed adherence using data collected automatically by the intervention program, including number of modules completed, pages accessed, log-ins, time between log-ins, time spent on the app, and completion rate. A total of 17% (3/18) of the studies administered a self-report questionnaire after the intervention, which included various questions about the use of the app and how they used what they learnt in their lives. One study included a viewing session on social media and the chance to post comments and used this interaction as an adherence measure. Among the 13 studies measuring adherence, attendance and program access ranged from 7% [21] to 91.8% [40] (n=5, 38% of the studies; median 64.9%). Completion of the intervention

XSL•FO
RenderX

ranged from 8% [49] to 90% [47] of the participants (8/13, 62% of the studies; median 44.4%).

Satisfaction

Satisfaction was measured in 72% (13/18) of the studies. The most common method of measurement (6/13, 46% of the studies) was administering postintervention scales created by the researchers, including questions about how useful and understandable the program material was, how likely the participants were to use it in the future, and their satisfaction with the intervention. Individualized postintervention self-report questionnaires were also commonly used (4/13, 31%). In total, 2 validated questionnaires used by 3 studies were the Client Satisfaction Questionnaire (n=1, 33%) and Telemedicine Satisfaction Questionnaire (n=1, 33%). The remaining study used completion rate as an indicator of the acceptability of the intervention.

Among the 6 studies who measured it, participant satisfaction ranged from 77% [48] to 92% [47] (n=5, 83% of the studies; median 87%), whereas 74% [42] to 93% [38] of the participants (n=4, 67% of the studies; median 85.6%) reported that they would recommend the intervention to others. Reports of benefitting from the program and finding it helpful ranged between 74.6% and 88.9% (3/13, 23% of the studies; median 83%). Fonseca et al [39] noted that those who completed the intervention were more satisfied with the program, had a higher intention to use the program in the future if needed, and perceived the intervention as useful at a significantly higher level than those allocated to the intervention condition who did not complete the program.

Quality Rating Score

The quality rating scores are presented in Table 2 and ranged between 6 and 10 out of 11. The most frequent score was 9 (8/18, 44% of the studies; median 9), indicating generally high methodological quality. The criteria most commonly met were true randomization, outcomes measured in the same way across groups, and outcomes measured reliably (18/18, 100%), and the criteria least commonly met were follow-up complete and differences between groups at follow-up described and analyzed (11/18, 61%). Multimedia Appendix 3 [21,34-50] provides detailed quality rating scores.

Synthesis of Results

Primary Analysis

Figures 2-6 present the forest plots for the 5 meta-analyses conducted to assess the effect of the interventions on depression, anxiety, and social ties.

Figure 2. Depression meta-analysis and forest plot at the postintervention time point. IV: inverse variance. [21,35-39,41-45,47,49-50].

Study or Subgroup	Intervention Mean	Intervention SD	Intervention Total	Control Mean	Control SD	Control Total	Weight	Std. Mean Difference IV, Random, 95% CI
Boyd et al., 2019	20.2	2.2	12	23.3	4.9	8	1.6%	−0.85 [−1.79, 0.09]
Carona 2023	8.75	4.53	267	10.78	5.06	445	9.1%	−0.42 [−0.57, −0.26]
Chan et al., 2019	5.3	4.4	218	5.9	4.7	225	8.6%	−0.13 [−0.32, 0.05]
Danaher 2023	5.78	4.42	86	7.48	5.67	92	6.8%	−0.33 [−0.63, −0.04]
Fonseca et al., 2020	6.91	3.63	65	6.87	3.71	82	6.3%	0.01 [−0.31, 0.34]
Haga et al., 2019	4	4	381	4.4	4.3	466	9.3%	−0.10 [−0.23, 0.04]
Heller et al., 2020	9.5	5.6	62	8.9	5.5	69	6.0%	0.11 [−0.24, 0.45]
Kavanagh et al., 2021 – Fathers	4.28	3.23	102	4.26	3.33	108	7.2%	0.01 [−0.26, 0.28]
Kavanagh et al., 2021 – Mothers	6.28	3.28	105	6.26	3.36	110	7.2%	0.01 [−0.26, 0.27]
Lennard et al., 2021	3.17	3.32	94	3.98	4.22	154	7.4%	−0.21 [−0.46, 0.05]
Loughnan et al., 2019	8.82	4.96	50	13.34	4.96	47	5.0%	−0.90 [−1.32, −0.49]
O'Mahen et al., 2013	10.94	5.57	181	14.28	6.63	162	8.1%	−0.55 [−0.76, −0.33]
Qin 2022	2.71	2.75	52	4.55	3.87	53	5.4%	−0.54 [−0.93, −0.15]
Sun et al., 2021	6.49	4.497	63	9.09	6.241	54	5.7%	−0.48 [−0.85, −0.11]
Zhang 2023	5.21	4.46	78	7.86	5.07	79	6.4%	−0.55 [−0.87, −0.23]
Total (95% CI)			1816			2154	100.0%	−0.28 [−0.41, −0.15]

Heterogeneity: Tau² = 0.04; Chi² = 50.35, df = 14 (P < 0.00001); I² = 72%
Test for overall effect: Z = 4.25 (P < 0.0001)

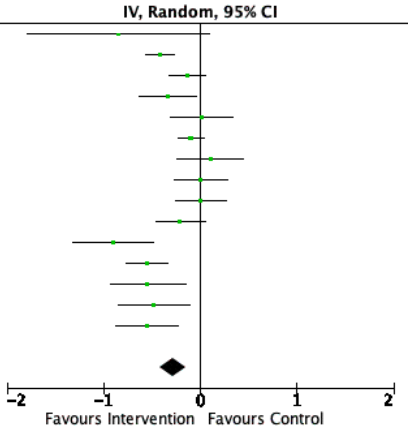

Figure 3. Depression meta-analysis and forest plot at follow-up. IV: inverse variance. [42-43,45,49-50].

Study or Subgroup	Intervention Mean	Intervention SD	Intervention Total	Control Mean	Control SD	Control Total	Weight	Std. Mean Difference IV, Random, 95% CI
Heller et al., 2020	8	5.2	54	8.7	5.9	65	16.4%	−0.12 [−0.49, 0.24]
Kavanagh et al., 2021 – Fathers	3.62	3.26	104	4.09	3.26	104	19.2%	−0.14 [−0.42, 0.13]
Kavanagh et al., 2021 – Mothers	5.07	3.33	108	5.54	3.36	110	19.5%	−0.14 [−0.41, 0.13]
Loughnan et al., 2019	8.01	4.05	37	12.13	4.22	42	13.2%	−0.99 [−1.45, −0.52]
Sun et al., 2021	6.77	4.693	52	6.25	5.098	40	14.8%	0.11 [−0.31, 0.52]
Zhang 2023	5.54	5.44	69	8.45	6.53	66	17.0%	−0.48 [−0.82, −0.14]
Total (95% CI)			424			427	100.0%	−0.27 [−0.52, −0.02]

Heterogeneity: Tau² = 0.06; Chi² = 15.73, df = 5 (P = 0.008); I² = 68%
Test for overall effect: Z = 2.15 (P = 0.03)

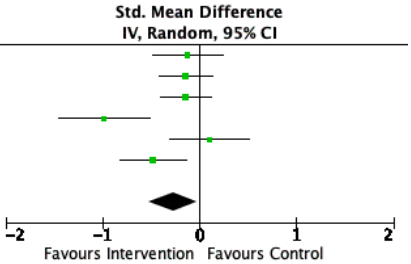

XSL•FO
RenderX

Figure 4. Anxiety meta-analysis and forest plot at the postintervention time point. IV: inverse variance. [36-39, 42, 44-45, 47, 49-50].

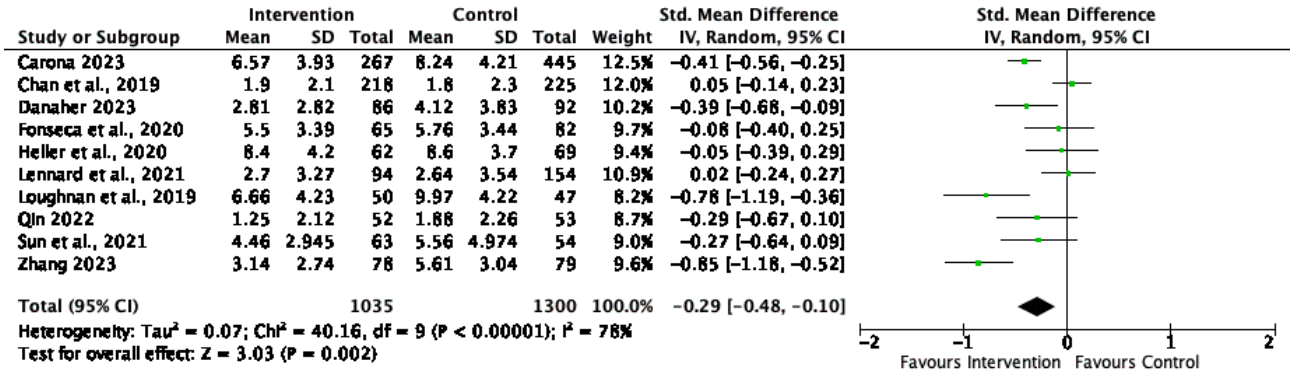

Figure 5. Anxiety meta-analysis and forest plot at follow-up. IV: inverse variance. [42, 45, 49-50].

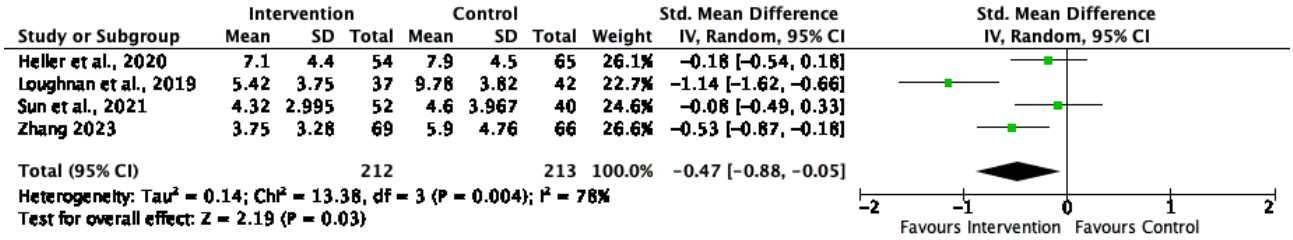

Figure 6. Social tie meta-analysis and forest plot at the postintervention time point. IV: inverse variance. [39, 43].

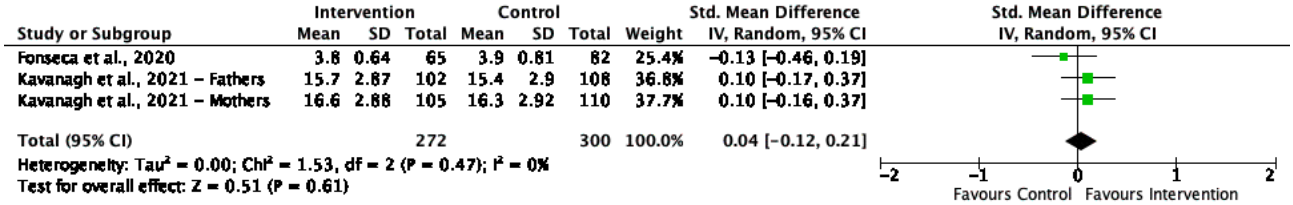

Depression

The overall effect size for depression at the postintervention time point was small and significant (SMD=–0.28, 95% CI –0.41 to –0.15; *P*<.001), with high and significant heterogeneity (I^2=72%; *P*<.001). The effect size at follow-up was small and significant (SMD=–0.27, 95% CI –0.52 to –0.02; *P*=.03), with high and significant heterogeneity (I^2=68%; *P*=.008).

Anxiety

The effect size for anxiety at the postintervention time point was small and significant (SMD=–0.29, 95% CI –0.48 to –0.10; *P*=.002), with high and significant heterogeneity (I^2=78%; *P*<.001). The effect size at follow-up was medium and significant (SMD=–0.47, 95% CI –0.88 to –0.05; *P*=.03), with high and significant heterogeneity (I^2=78%; *P*=.004).

Social Ties

The effect size for social ties at the postintervention time point was small and nonsignificant (SMD=0.04, 95% CI –0.12 to 0.21; *P*=.61), with low and nonsignificant heterogeneity (I^2=0%; *P*=.47). Due to a lack of studies, a meta-analysis for the follow-up time point could not be conducted.

Publication Bias

Publication bias was analyzed using funnel plots for each meta-analysis (Multimedia Appendices 4-8). There appeared to be symmetry in these funnel plots, indicating minimal publication bias.

Discussion

Principal Findings and Comparison to Prior Work

This review and series of meta-analyses aimed to assess the effectiveness of technology-based distal interventions in reducing or preventing symptoms of postpartum depression or anxiety and increasing social ties and to determine the level of adherence and satisfaction associated with these interventions. Overall, 18 relevant studies were identified, 14 (78%) of which were able to be included in one or more of the meta-analyses. The overall results of the meta-analyses were promising, with significant effects in favor of the intervention condition found for depression and anxiety at the postintervention and follow-up time points. These findings align with the outcomes of previous systematic reviews and meta-analyses of digital mental health interventions for perinatal depression [22-24] and anxiety in women [23,24], which also found small to medium overall effects in favor of the intervention condition. This review extends the findings of previous reviews through its assessment

XSL•FO
RenderX

of interventions for both mothers and fathers and by expanding the focus beyond CBT-based interventions.

Most of the studies that showed efficacy in this review compared the focal intervention to an attention or treatment-as-usual control condition. Overall, the findings suggest that distally delivered technology-based interventions can be effective in reducing symptoms of depression and anxiety primarily among birth mothers and provide support for their ongoing development and implementation as an alternative treatment and prevention approach to face-to-face services, which report broadly similar findings [52,53]. Given the small number of trials that used attention or active control conditions, future studies should consider using such controls to further strengthen the evidence base.

No significant effect was observed in the meta-analyses for social ties at the postintervention time point, although the findings were in the expected direction in favor of the intervention condition. Fewer studies assessed social ties in comparison to depression and anxiety, which may have contributed to the nonsignificant effects on this outcome. Previous research has shown that strong social ties tend to have a positive impact on postpartum mental health [54,55]. In addition, social isolation has been reported as a contributing factor to developing and maintaining depression and anxiety symptoms [10]. Despite this, few papers that were eligible for this review measured social ties. This resulted in not enough research to draw strong conclusions; however, it was indicated that couples participating in interventions at the same time potentially allows for increased social support [43]. This may be a beneficial area for future research to investigate to determine whether coparenting mothers and fathers' postpartum mental health could be improved by couples completing interventions at the same time and then discussing the materials. However, it may also be the case that other forms of social or group interventions are more efficient in promoting social ties among new parents [56].

Two studies, those by Loughnan et al [45] and Boyd et al [35], produced large and significant effects for reducing symptoms of depression. These studies were both conducted with birth mothers in Western countries using online CBT programs compared to treatment-as-usual or attention control conditions. These interventions also included a form of optional communication, including messages reminding participants to complete the program, messages from health professionals if participant distress was noted as high, and optional access to a forum with other participants and facilitators. In addition, they were both short-term interventions spanning between 6 and 8 weeks. The large and significant effects demonstrated by these 2 studies suggest that these particular interventions may include effective characteristics for reducing postpartum depression or anxiety that warrant further investigation. A key area may be determining whether a limited amount of optional communication, as optimized in the interventions by Loughnan et al [45] and Boyd et al [35], is more beneficial for reducing symptoms of depression than no communication. Given this finding, it is recommended that future research directly compare the benefits of communication by having 2 intervention conditions, with one receiving communication and the other not.

The study by Zhang et al [50] produced a large and significant effect size for reducing anxiety symptoms. This study was conducted in China and focused on mindfulness via a mobile app. It included 6 modules delivered weekly and provided standardized reminder messages. This large and significant effect size indicates that the features of this intervention may be effective in reducing symptoms of anxiety using distal and technology-based interventions. As with the studies by Loughnan et al [45] and Boyd et al [35], the intervention was short term, included communication via a reminder message, and was tested among birth mothers. However, there were some differences among the 3 studies, including the fact that one was a mindfulness-based mobile app in China being used for anxiety and the other 2 were online CBT interventions in Western countries being used for depression. Future research may benefit from testing the interventions' effectiveness on wider populations, including birth fathers and adoptive parents, which may assist in understanding whether these intervention characteristics are effective for populations other than birth mothers.

The overall methodological quality of the studies was high, indicating that measures were in place to avoid potential bias, including true randomization methods and measuring outcomes reliably, and that study authors adequately reported on their study methods. Some areas that could be improved include ensuring that the groups were similar at baseline and including follow-up measures in the study design.

Despite the scope of this review including interventions for postpartum depression and anxiety for birth and adoptive mothers and fathers, the literature search did not find any relevant papers for adoptive parents or fathers and found only 2 studies that assessed interventions targeted at both parents in coupled families. This finding suggests that there is a clear gap in the provision of interventions to support adoptive families and birth fathers who may be experiencing postpartum anxiety and depression. There is a need to develop and evaluate such interventions given the rates of disorder in these populations [6,7]. Excluding these populations from intervention research may increase their risk of poor mental health outcomes, including sleep disturbance, the breakdown of relationships, and challenges bonding with their children, in addition to developing postpartum depression and anxiety [11,12,14].

Limitations

The results of this review should be interpreted in light of the limitations of both this paper and the included studies. First, there was high and significant heterogeneity in the 4 depression and anxiety meta-analyses. This was expected due to the many differences between the studies, including the measures used, intervention length, onset of intervention, participants, use of prevention and treatment interventions, adherence, follow-up timing, country where the studies were conducted, intervention formats, content, and theoretical perspectives. Given this, subgroup and moderator analyses were considered; however, they were deemed inappropriate due to having an insufficient number of studies. Some heterogeneity is unavoidable due to

XSL•FO
RenderX

differences in populations and intervention characteristics. However, heterogeneity in measurement of anxiety, depression, social ties, adherence, and satisfaction may be addressed through further evaluation of measurement properties, harmonization of measures [57], and consideration of the limitations of some common measures [58].

Second, most studies had different follow-up measurement time points. Therefore, the statistics used for the follow-up meta-analyses were collected at different times after the intervention. This may result in some interventions appearing more or less effective than others based on the timing of the measures. Future research may benefit from determining the optimal time to measure follow-ups (eg, timing based on when participants gave birth or after the intervention has been administered or drawing on theory and evidence on the timing for the development of depression and anxiety during this period to inform these decisions), completing subgroup analyses based on follow-up time points for future meta-analyses, or conducting individual patient data meta-analyses that can accommodate variable timing in outcome measures.

A limitation of the studies included is that many of them were underpowered to detect modest effects. With universal prevention interventions, it is unlikely that medium or large effects will be identified as a large proportion of the sample will not have scope for a reduction in symptoms. Therefore, future research would benefit from the use of relatively large samples. Larger samples may also enable more nuanced examination of which groups of postpartum parents benefit most from internet-based interventions and when the interventions should be optimally delivered.

The EPDS was the most commonly used measure of depression in the included studies. This is unsurprising given that it has been largely used in previous research and postpartum support services. However, it has been criticized for its inclusion of ambiguous items, exclusion of particular forms of distress, challenges with scoring, poor predictive ability, and limited detection of depressive and anxious symptoms in men [58]. Future research may consider including either multiple or multidimensional measures (eg, the DASS-21) so that findings related to both anxiety and depression can be tested and triangulated.

Some relevant studies may not have been identified due to only including papers written in English and published in peer-reviewed journals. In addition, due to manually screening papers, relevant studies may have been excluded. However, double screening and coding protocols were in place to minimize the risk of this occurring. Further, incomplete data prevented the inclusion in the meta-analyses of several of the studies included in this review. This resulted in relevant papers not having their interventions assessed for effectiveness on symptom reduction. Although the corresponding authors were contacted with requests for data, they either did not respond or provided data that were incompatible with the meta-analysis due to the statistical analysis methods used.

Practical Implications

There are multiple key practical implications from this review. Primarily, the results indicate promising findings to prevent and reduce postpartum anxiety and depression using distal technology-based interventions, which can be more readily accessed and widely disseminated to those in need. The provision of such interventions could allow more parents to access support when it is most needed, with reductions in anxiety and depression leading to a range of positive effects on parents and their children [20]. These interventions can also overcome the stigma often associated with mental health problems and help seeking. As such, there is a clear need to promote these interventions within health care settings and among new parents to ensure access to and awareness of them. Future research may endeavor to compare the effect of distal technology-based interventions with in-person therapies rather than control groups to further determine the comparable effectiveness of these intervention methods.

In addition, this review did not identify any distal technology-based approaches other than online interventions or mobile apps, such as podcasts or artificial intelligence–based interventions [59]. Emerging technologies may build on the benefits of web-based and mobile app interventions by additionally providing an option for those with low literacy or potentially tailoring content to individual needs and preferences. The use of audio interventions (eg, podcasts) may also increase adherence as parents can listen to the content while completing other daily activities, such as feeding their child, driving, or exercising. This may be a beneficial direction for future research to explore to allow for additional flexibility in interventions for postpartum anxiety and depression.

Finally, this review provided insights regarding the content of distal technology-based interventions. Most of the interventions were CBT based, with multiple incorporating mindfulness, and these included the 3 most effective interventions in reducing symptoms of depression and anxiety when compared to the control groups [35,45,50]. This indicates that CBT-based and mindfulness interventions can be beneficial for parents in the postpartum period, which aligns with research on nondistal or non–technology-based interventions [60,61]. In contrast, only 1 study used IPT despite this being a well-accepted therapeutic technique for this population in nondistal or non–technology-based interventions [62,63]. Future research would benefit from further assessing IPT-based distal technology-based interventions.

Conclusions

Overall, the findings of this review and the meta-analyses suggest that distal technology-based interventions are effective in reducing symptoms of postpartum anxiety and depression when compared to a control condition for birth parents, particularly mothers. Although the overall effect sizes were small to medium, there was a consistent reduction in symptoms in the intervention groups. Short-term interventions with optional communication channels, such as reminders to complete the program, were associated with the largest effects. Further research is required to determine what factors make some interventions more effective than others and apply this to

XSL•FO
RenderX

adoptive parents, birth fathers, and couples in addition to birth mothers. Furthermore, while there is a growing body of research investigating distal interventions for postpartum depression and anxiety, there has been limited focus on social ties. This is an area for future research given the prevalence of social isolation in the perinatal period and its contribution to distress.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. ALC is supported by a National Health and Medical Research Council fellowship (1173146). TC is supported by a National Health and Medical Research Council fellowship (1173270). No generative AI was used in any portion of the manuscript writing.

Authors' Contributions

SPB contributed to conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing the original draft, reviewing and editing the manuscript, visualization, and project administration. ARM contributed to conceptualization, methodology, investigation, reviewing and editing the manuscript, visualization, supervision, and project administration. TC contributed to conceptualization, methodology, investigation, reviewing and editing the manuscript, visualization, and supervision. PJB contributed to conceptualization, methodology, investigation, reviewing and editing the manuscript, visualization, and supervision. LL contributed to conceptualization, methodology, investigation, reviewing and editing the manuscript, visualization, and supervision. AR contributed to investigation and reviewing and editing the manuscript. AS contributed to investigation and reviewing and editing the manuscript. CTB contributed to investigation and reviewing and editing the manuscript. JN contributed to investigation and reviewing and editing the manuscript. ALC contributed to conceptualization, methodology, investigation, reviewing and editing the manuscript, visualization, supervision, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search string example.
[DOCX File , 18 KB - jmir_v26i1e53236_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 2957 KB - jmir_v26i1e53236_app2.pdf]

Multimedia Appendix 3
Quality ratings.
[DOCX File , 18 KB - jmir_v26i1e53236_app3.docx]

Multimedia Appendix 4
Funnel plot of depression effect size data at the postintervention time point.
[DOCX File , 34 KB - jmir_v26i1e53236_app4.docx]

Multimedia Appendix 5
Funnel plot of depression effect size data at follow-up.
[DOCX File , 31 KB - jmir_v26i1e53236_app5.docx]

Multimedia Appendix 6
Funnel plot of anxiety effect size data at the postintervention time point.
[DOCX File , 32 KB - jmir_v26i1e53236_app6.docx]

Multimedia Appendix 7
Funnel plot of anxiety effect size data at follow-up.
[DOCX File , 31 KB - jmir_v26i1e53236_app7.docx]

Multimedia Appendix 8
Funnel plot of social tie effect size data at the postintervention time point.

XSL•FO
RenderX

[DOCX File , 30 KB - jmir_v26i1e53236_app8.docx]

References

1. Foli KJ, South SC, Lim E, Jarnecke AM. Post-adoption depression: parental classes of depressive symptoms across time. J Affect Disord 2016 Aug;200:293-302 [FREE Full text] [doi: 10.1016/j.jad.2016.01.049] [Medline: 27155072]
2. Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition. New York, NY: American Psychiatric Association; 2013.
3. Almond P. Postnatal depression: a global public health perspective. Perspect Public Health 2009 Sep 15;129(5):221-227. [doi: 10.1177/1757913909343882] [Medline: 19788165]
4. Jones BA, Griffiths KM, Christensen H, Ellwood D, Bennett K, Bennett A. Online cognitive behaviour training for the prevention of postnatal depression in at-risk mothers: a randomised controlled trial protocol. BMC Psychiatry 2013 Oct 17;13(1):265 [FREE Full text] [doi: 10.1186/1471-244X-13-265] [Medline: 24131528]
5. Paulson JF, Bazemore SD. Prenatal and postpartum depression in fathers and its association with maternal depression: a meta-analysis. JAMA 2010 May 19;303(19):1961-1969. [doi: 10.1001/jama.2010.605] [Medline: 20483973]
6. Mott SL, Schiller CE, Richards JG, O'Hara MW, Stuart S. Depression and anxiety among postpartum and adoptive mothers. Arch Womens Ment Health 2011 Aug 3;14(4):335-343 [FREE Full text] [doi: 10.1007/s00737-011-0227-1] [Medline: 21725836]
7. Foli KJ, South SC, Lim E. Rates and predictors of depression in adoptive mothers: moving toward theory. ANS Adv Nurs Sci 2012;35(1):51-63. [doi: 10.1097/ANS.0b013e318244553e] [Medline: 22293610]
8. Misri S, Abizadeh J, Sanders S, Swift E. Perinatal generalized anxiety disorder: assessment and treatment. J Womens Health (Larchmt) 2015 Sep;24(9):762-770 [FREE Full text] [doi: 10.1089/jwh.2014.5150] [Medline: 26125602]
9. Leigh B, Milgrom J. Risk factors for antenatal depression, postnatal depression and parenting stress. BMC Psychiatry 2008 Apr 16;8:24 [FREE Full text] [doi: 10.1186/1471-244X-8-24] [Medline: 18412979]
10. Letourneau NL, Dennis C, Benzies K, Duffett-Leger L, Stewart M, Tryphonopoulos PD, et al. Postpartum depression is a family affair: addressing the impact on mothers, fathers, and children. Issues Ment Health Nurs 2012 Jul 03;33(7):445-457. [doi: 10.3109/01612840.2012.673054] [Medline: 22757597]
11. Wilkinson RB, Mulcahy R. Attachment and interpersonal relationships in postnatal depression. J Reprod Infant Psychol 2010 May 20;28(3):252-265. [doi: 10.1080/02646831003587353]
12. Sreejith V, Jose G, Athulya L. Association between quality of sleep and depression in postpartum women. Int J Phys Educ Sports Health 2021;8:1-3. [doi: 10.22271/kheljournal]
13. Murray L, Arteche A, Fearon P, Halligan S, Croudace T, Cooper P. The effects of maternal postnatal depression and child sex on academic performance at age 16 years: a developmental approach. J Child Psychol Psychiatry 2010 Oct 14;51(10):1150-1159. [doi: 10.1111/j.1469-7610.2010.02259.x] [Medline: 20840504]
14. Stein A, Arteche A, Lehtonen A, Craske M, Harvey A, Counsell N, et al. Interpretation of infant facial expression in the context of maternal postnatal depression. Infant Behav Dev 2010 Jun;33(3):273-278 [FREE Full text] [doi: 10.1016/j.infbeh.2010.03.002] [Medline: 20381873]
15. Huang L, Zhao Y, Qiang C, Fan B. Is cognitive behavioral therapy a better choice for women with postnatal depression? A systematic review and meta-analysis. PLoS One 2018;13(10):e0205243. [doi: 10.1371/journal.pone.0205243] [Medline: 30321198]
16. Weissman MM, Markowitz JC, Klerman GL. Pregnancy, miscarriage, and postpartum depression. In: Weissman MM, Markowitz JC, Klerman GL, editors. Clinician's Quick Guide to Interpersonal Psychotherapy. Oxford, UK: Oxford University Press; 2007:94-97.
17. Ramsauer B, Mühlhan C, Mueller J, Schulte-Markwort M. Parenting stress and postpartum depression/anxiety in mothers with personality disorders: indications for differential intervention priorities. Eur Child Adolesc Psychiatry 2016 Jun 23;25(6):589-600. [doi: 10.1007/s00787-015-0775-4] [Medline: 26400074]
18. Stewart DE, Vigod S. Postpartum depression. N Engl J Med 2016 Dec;375(22):2177-2186. [doi: 10.1056/nejmcp1607649]
19. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry 2020 Oct;19(3):313-327 [FREE Full text] [doi: 10.1002/wps.20769] [Medline: 32931106]
20. Milgrom J, Danaher BG, Gemmill AW, Holt C, Holt CJ, Seeley JR, et al. Internet cognitive behavioral therapy for women with postnatal depression: a randomized controlled trial of MumMoodBooster. J Med Internet Res 2016 Mar 07;18(3):e54 [FREE Full text] [doi: 10.2196/jmir.4993] [Medline: 26952645]
21. O'Mahen HA, Woodford J, McGinley J, Warren FC, Richards DA, Lynch TR, et al. Internet-based behavioral activation--treatment for postnatal depression (Netmums): a randomized controlled trial. J Affect Disord 2013 Sep 25;150(3):814-822. [doi: 10.1016/j.jad.2013.03.005] [Medline: 23602514]
22. Chen C, Wang X, Xu H, Li Y. Effectiveness of digital psychological interventions in reducing perinatal depression: a systematic review of meta-analyses. Arch Womens Ment Health 2023 Aug;26(4):423-439 [FREE Full text] [doi: 10.1007/s00737-023-01327-y] [Medline: 37231116]

XSL•FO
RenderX

23. Ching H, Chua JY, Chua JS, Shorey S. The effectiveness of technology-based cognitive behavioral therapy on perinatal depression and anxiety: a systematic review and meta-analysis. Worldviews Evid Based Nurs 2023 Oct 09;20(5):451-464. [doi: 10.1111/wvn.12673] [Medline: 37559467]
24. Lau Y, Chew HS, Ang WH, Ang WW, Yeo CY, Lim GZ, et al. Effects of digital health interventions on the psychological outcomes of perinatal women: umbrella review of systematic reviews and meta-analyses. Health Psychol Rev 2024 Jun 15;18(2):229-254. [doi: 10.1080/17437199.2023.2185654] [Medline: 36919443]
25. Better systematic review management. Covidence. URL: https://www.covidence.org/ [accessed 2024-04-29]
26. Barker T, Stone JC, Sears K, Klugar M, Tufanaru C, Leonardi-Bee J, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth 2023 Mar 01;21(3):494-506. [doi: 10.11124/JBIES-22-00430] [Medline: 36727247]
27. Cohen J. The effect size index. In: Cohen J, editor. Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge; 1988:25-26.
28. RevMan. The Cochrane Collaboration. URL: https://revman.cochrane.org/ [accessed 2024-04-29]
29. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011 Feb 10;342(feb10 2):d549. [doi: 10.1136/bmj.d549] [Medline: 21310794]
30. Higgins JP, Li T, Deeks JJ. Choosing effect measures and computing estimates of effect. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. 3rd edition. New York, NY: The Cochrane Collaboration; 2022.
31. Werner-Seidler A, Spanos S, Calear AL, Perry Y, Torok M, O'Dea B, et al. School-based depression and anxiety prevention programs: an updated systematic review and meta-analysis. Clin Psychol Rev 2021 Nov;89:102079 [FREE Full text] [doi: 10.1016/j.cpr.2021.102079] [Medline: 34571372]
32. Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol 2012 Jul 30;12:111 [FREE Full text] [doi: 10.1186/1471-2288-12-111] [Medline: 22846171]
33. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
34. Barrera AZ, Wickham RE, Muñoz RF. Online prevention of postpartum depression for Spanish- and English-speaking pregnant women: a pilot randomized controlled trial. Internet Interv 2015 Sep 01;2(3):257-265 [FREE Full text] [doi: 10.1016/j.invent.2015.06.002] [Medline: 26273567]
35. Boyd RC, Price J, Mogul M, Yates T, Guevara JP. Pilot RCT of a social media parenting intervention for postpartum mothers with depression symptoms. J Reprod Infant Psychol 2019 Jul;37(3):290-301. [doi: 10.1080/02646838.2018.1556788] [Medline: 30556428]
36. Carona C, Pereira M, Araújo-Pedrosa A, Canavarro MC, Fonseca A. The efficacy of be a mom, a web-based intervention to prevent postpartum depression: examining mechanisms of change in a randomized controlled trial. JMIR Ment Health 2023 Mar 17;10:e39253 [FREE Full text] [doi: 10.2196/39253] [Medline: 36930182]
37. Chan KL, Leung WC, Tiwari A, Or KL, Ip P. Using smartphone-based psychoeducation to reduce postnatal depression among first-time mothers: randomized controlled trial. JMIR Mhealth Uhealth 2019 May 14;7(5):e12794 [FREE Full text] [doi: 10.2196/12794] [Medline: 31094354]
38. Danaher BG, Seeley JR, Silver RK, Tyler MS, Kim JJ, La Porte LM, et al. Trial of a patient-directed eHealth program to ameliorate perinatal depression: the MomMoodBooster2 practical effectiveness study. Am J Obstet Gynecol 2023 Apr;228(4):453.e1-453.10 [FREE Full text] [doi: 10.1016/j.ajog.2022.09.027] [Medline: 36174746]
39. Fonseca A, Alves S, Monteiro F, Gorayeb R, Canavarro MC. Be a mom, a web-based intervention to prevent postpartum depression: results from a pilot randomized controlled trial. Behav Ther 2020 Jul;51(4):616-633. [doi: 10.1016/j.beth.2019.09.007] [Medline: 32586434]
40. Guo L, Zhang J, Mu L, Ye Z. Preventing postpartum depression with mindful self-compassion intervention: a randomized control study. J Nerv Ment Dis 2020 Feb;208(2):101-107. [doi: 10.1097/NMD.0000000000001096] [Medline: 31868776]
41. Haga SM, Drozd F, Lisøy C, Wentzel-Larsen T, Slinning K. Mamma Mia - a randomized controlled trial of an internet-based intervention for perinatal depression. Psychol Med 2019 Aug;49(11):1850-1858. [doi: 10.1017/S0033291718002544] [Medline: 30191779]
42. Heller HM, Hoogendoorn AW, Honig A, Broekman BF, van Straten A. The effectiveness of a guided internet-based tool for the treatment of depression and anxiety in pregnancy (MamaKits online): randomized controlled trial. J Med Internet Res 2020 Mar 23;22(3):e15172 [FREE Full text] [doi: 10.2196/15172] [Medline: 32202505]
43. Kavanagh DJ, Connolly J, Fisher J, Halford WK, Hamilton K, Hides L, et al. The baby steps web program for the well-being of new parents: randomized controlled trial. J Med Internet Res 2021 Nov 26;23(11):e23659 [FREE Full text] [doi: 10.2196/23659] [Medline: 34842534]
44. Lennard GR, Mitchell AE, Whittingham K. Randomized controlled trial of a brief online self-compassion intervention for mothers of infants: effects on mental health outcomes. J Clin Psychol 2021 Mar 15;77(3):473-487. [doi: 10.1002/jclp.23068] [Medline: 33063321]

XSL•FO
RenderX

45. Loughnan SA, Butler C, Sie AA, Grierson AB, Chen AZ, Hobbs MJ, et al. A randomised controlled trial of 'MUMentum postnatal': internet-delivered cognitive behavioural therapy for anxiety and depression in postpartum women. Behav Res Ther 2019 May;116:94-103. [doi: 10.1016/j.brat.2019.03.001] [Medline: 30877878]
46. Nishi D, Imamura K, Watanabe K, Obikane E, Sasaki N, Yasuma N, et al. The preventive effect of internet-based cognitive behavioral therapy for prevention of depression during pregnancy and in the postpartum period (iPDP): a large scale randomized controlled trial. Psychiatry Clin Neurosci 2022 Nov;76(11):570-578 [FREE Full text] [doi: 10.1111/pcn.13458] [Medline: 36114656]
47. Qin X, Liu C, Zhu W, Chen Y, Wang Y. Preventing postpartum depression in the early postpartum period using an app-based cognitive behavioral therapy program: a pilot randomized controlled study. Int J Environ Res Public Health 2022 Dec 15;19(24):16824 [FREE Full text] [doi: 10.3390/ijerph192416824] [Medline: 36554704]
48. Shorey S, Lau Y, Dennis CL, Chan YS, Tam WW, Chan YH. A randomized-controlled trial to examine the effectiveness of the 'Home-but not Alone' mobile-health application educational programme on parental outcomes. J Adv Nurs 2017 Sep 07;73(9):2103-2117. [doi: 10.1111/jan.13293] [Medline: 28276086]
49. Sun Y, Li Y, Wang J, Chen Q, Bazzano AN, Cao F. Effectiveness of smartphone-based mindfulness training on maternal perinatal depression: randomized controlled trial. J Med Internet Res 2021 Jan 27;23(1):e23410 [FREE Full text] [doi: 10.2196/23410] [Medline: 33502326]
50. Zhang X, Li Y, Wang J, Mao F, Wu L, Huang Y, et al. Effectiveness of digital guided self-help mindfulness training during pregnancy on maternal psychological distress and infant neuropsychological development: randomized controlled trial. J Med Internet Res 2023 Feb 10;25:e41298. [doi: 10.2196/41298] [Medline: 36763452]
51. O'Mahen HA, Grieve H, Jones J, McGinley J, Woodford J, Wilkinson EL. Women's experiences of factors affecting treatment engagement and adherence in internet delivered behavioural activation for postnatal depression. Internet Interv 2015 Mar;2(1):84-90. [doi: 10.1016/j.invent.2014.11.003]
52. Clinkscales N, Golds L, Berlouis K, MacBeth A. The effectiveness of psychological interventions for anxiety in the perinatal period: a systematic review and meta-analysis. Psychol Psychother 2023 Jun 11;96(2):296-327. [doi: 10.1111/papt.12441] [Medline: 36504355]
53. Stephens S, Ford E, Paudyal P, Smith H. Effectiveness of psychological interventions for postnatal depression in primary care: a meta-analysis. Ann Fam Med 2016 Sep 12;14(5):463-472 [FREE Full text] [doi: 10.1370/afm.1967] [Medline: 27621164]
54. Cho H, Lee K, Choi E, Cho HN, Park B, Suh M, et al. Association between social support and postpartum depression. Sci Rep 2022 Feb 24;12(1):3128 [FREE Full text] [doi: 10.1038/s41598-022-07248-7] [Medline: 35210553]
55. Seymour-Smith M, Cruwys T, Haslam SA. More to lose? Longitudinal evidence that women whose social support declines following childbirth are at increased risk of depression. Aust N Z J Public Health 2021 Aug;45(4):338-343 [FREE Full text] [doi: 10.1111/1753-6405.13099] [Medline: 33818864]
56. Seppälä T, Riikonen R, Stevenson C, Paajanen P, Repo K, Finell E. Intragroup contact with other mothers living in the same neighborhood benefits mothers' life satisfaction: the mediating role of group identification and social support. J Community Psychol 2023 Apr;51(3):1365-1377. [doi: 10.1002/jcop.22960] [Medline: 36326076]
57. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess 2014 Jun;26(2):513-527 [FREE Full text] [doi: 10.1037/a0035768] [Medline: 24548149]
58. Matthey S, Agostini F. Using the Edinburgh postnatal depression scale for women and men-some cautionary thoughts. Arch Womens Ment Health 2017 Apr 11;20(2):345-354. [doi: 10.1007/s00737-016-0710-9] [Medline: 28078472]
59. Karing C. The efficacy of online mindfulness-based interventions in a university student sample: videoconference- or podcast-delivered intervention. Appl Psychol Health Well Being 2023 May;15(2):740-756. [doi: 10.1111/aphw.12408] [Medline: 36214182]
60. Lavender TJ, Ebert L, Jones D. An evaluation of perinatal mental health interventions: an integrative literature review. Women Birth 2016 Oct;29(5):399-406. [doi: 10.1016/j.wombi.2016.04.004] [Medline: 27118000]
61. Marrs J. Evaluation of the impact of a CBT-based group on maternal postnatal mental health difficulties. J Reprod Infant Psychol 2013 Feb;31(1):72-80. [doi: 10.1080/02646838.2012.747676]
62. Mulcahy R, Reay RE, Wilkinson RB, Owen C. A randomised control trial for the effectiveness of group interpersonal psychotherapy for postnatal depression. Arch Womens Ment Health 2010 Apr 21;13(2):125-139. [doi: 10.1007/s00737-009-0101-6] [Medline: 19697094]
63. Sockol LE. A systematic review and meta-analysis of interpersonal psychotherapy for perinatal women. J Affect Disord 2018 May;232:316-328. [doi: 10.1016/j.jad.2018.01.018] [Medline: 29501991]

Abbreviations

CBT: cognitive behavioral therapy
DASS-21: Depression, Anxiety, and Stress Scales–21
EPDS: Edinburgh Postnatal Depression Scale

XSL•FO
RenderX

IPT: interpersonal psychotherapy
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by N Cahill; submitted 29.09.23; peer-reviewed by K Wislocki, D Liu; comments to author 23.01.24; revised version received 14.03.24; accepted 25.09.24; published 19.11.24.

Please cite as:
Brocklehurst SP, Morse AR, Cruwys T, Batterham PJ, Leach L, Robertson AM, Sahib A, Burke CT, Nguyen J, Calear AL
Investigating the Effectiveness of Technology-Based Distal Interventions for Postpartum Depression and Anxiety: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e53236
URL: https://www.jmir.org/2024/1/e53236
doi:10.2196/53236
PMID:

©Sarah P Brocklehurst, Alyssa R Morse, Tegan Cruwys, Philip J Batterham, Liana Leach, Alysia M Robertson, Aseel Sahib, Colette T Burke, Jessica Nguyen, Alison L Calear. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Usability and Effectiveness of eHealth and mHealth Interventions That Support Self-Management and Health Care Transition in Adolescents and Young Adults With Chronic Disease: Systematic Review

ZhiRu Li[1], BSN, RN; FangYan Lu[1], MSN, RN; JingYun Wu[1], BSN, RN; RuiJie Bao[1], BSN, RN; YuXin Rao[1], BSN, RN; Yun Yang[1], BSN, RN; Huafen Wang[1], MSN, RN

Nursing Department, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

Corresponding Author:
Huafen Wang, MSN, RN
Nursing Department
The First Affiliated Hospital
Zhejiang University School of Medicine
79 Qingchun Road
Hangzhou, 310000
China
Phone: 86 13516805833
Fax: 86 14223203400
Email: 2185015@zju.edu.cn

Abstract

Background: With advances in medical technologies, more children with chronic diseases are now living on into adulthood. The development of proficient self-management skills is essential for adolescents and young adults to transition from pediatric to adult health care services. An innovative way to improve the current care model and foster self-management skills could be through eHealth or mHealth (mobile health) interventions, in particular, when considering the rising ownership of digital technology by adolescents and young adults.

Objective: This systematic review aimed to evaluate the features, acceptability, usability, engagement, and intervention efficacy of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease.

Methods: This review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines (registration number CRD42023378355). We systematically searched the MEDLINE complete, Embase, Cochrane Library, CINAHL complete, and ProQuest Health & Medical complete. We considered only articles published in or since 2019, as we aimed to extend the data collected by 2 previous systematic reviews.

Results: A total of 16,752 studies were screened. After removing duplicates, 14,507 studies were excluded based on the title and abstract. Ultimately, 22 studies were included. The interventions ranged from simple text messages to complex interventions involving web-based games and engagement of health care providers, which were summarized into 6 themes: medication monitoring and reminders, symptom tracking and monitoring, management goal setting, knowledge education and self-management skills training, incentives and reinforcement, and communication. Most adolescents and young adults felt the eHealth and mHealth interventions were feasible, as they were convenient, easy to use, and accessible in the context of helping manage their health. However, user engagement was variable and presented a gradual decline in youth engagement with these apps over time. Barriers that prevent user engagement are diverse, such as time-consuming video uploads, noncontinuous access to a phone, reading literacy levels, language, and false impressions. Moreover, adolescents and young adults had different preferred styles of message delivery and functions, especially the engaging elements, disease-specific information, and opportunities to communicate with peers, health care providers, and app teams.

Conclusions: There remains limited data about the effectiveness of eHealth and mHealth interventions facilitating the self-management and health care transition of adolescents and young adults with chronic diseases. Based on the available evidence, they were receptive to and interested in receiving information and managing their health using mobile apps or websites. Considering

XSL•FO
RenderX

adolescents and young adults had different preferred styles of message delivery and features, to improve user engagement and provide focused interventions, it would be better to involve them early in the design process to identify their specific needs, as well as collaborate with health care providers and app teams to obtain suggestions.

(J Med Internet Res 2024;26:e56556) doi:10.2196/56556

KEYWORDS

eHealth; mHealth; mobile health; chronic disease; healthcare transition; self-management; adolescents; young adults; chronic illness; systematic review; digital health; health education; social support; symptom tracking; monitoring; effectiveness

Introduction

Background

Chronic diseases in childhood refer to children aged 0-18 years who endure incurable diseases diagnosed by reproducible and valid methods according to professional standards, with a duration of longer than three months or a frequency of more than three times during the past year and will probably reoccur [1]. In recent years, the incidence of chronic diseases in children has significantly increased, which has become a major global public health issue affecting children's physical and mental health [2,3]. Research showed that approximately 10% to 20% of children currently endure chronic diseases, including asthma, epilepsy, diabetes, chronic kidney disease, etc [4]. Due to development of the advanced medical technologies that support disease management and longer life spans, more children with chronic diseases are now living into adolescence and young adulthood [5]. To receive age-appropriate care, adolescents and young adults need to transition from pediatric to adult health care, which involves a significant shift in the care model from chronic disease supervision and management to self-management [6,7]. Pediatric providers tend to prefer a family-centered model of care, with parental involvement in disease management and daily care for children; while adult care, is patient-centered, emphasizing patient independence and personalized care [5,8]. Therefore, the health care transition from child-centered to adult-oriented health care systems is a period of increased risk and vulnerability.

Transition readiness, as a key element of health care transition, refers to the abilities to prepare, enter, continue, and complete the transition. Such preparation mainly reflects self-management ability, including the knowledge, skills, and experience required for disease management [9]. Self-management is an important skill for adolescent chronic disease patients to take over responsibility for their own health care and ensure good disease control [10,11]. Research has shown that inadequate self-management can lead to reduced adherence to treatment, decreased follow-up, and increased risk of mortality among adolescents and young adults with chronic diseases, thus placing a heavy economic burden on adolescents and young adults, their families, and even society [12-14]. Therefore, it is crucial to encourage and promote the improvement of adolescents' and young adults' self-management behaviors to ensure the continuity of their health management. However, changes in self-management behavior entail dynamic, continuous, and complex processes that are influenced by multiple factors, and these changes are often difficult to maintain [15]. Therefore, it is very important to find suitable methods to use in adolescent development to encourage and promote the improvement of self-management behavior among young people.

An innovative way to improve the current care model and foster self-management skills could take the form of digital health interventions, especially in light of the changing landscape of internet use and increasing ownership rates of digital technology among young people. Digital health intervention refers to the use of the internet, smartphones, social media, and mobile apps to provide health care information support and treatment to a target population [16]. Such interventions have come to represent an emerging mode of chronic disease health management due to the advantages of convenience, interactivity, accessibility, and low cost, which can effectively provide personalized and continuous web-based medical assurance [17]. Digital media have come to represent an indispensable aspect of young people's lives, as young people rely on digital media to seek different forms of mental health support, relaxation, and distraction [18-20]. The use of eHealth or mHealth (mobile health) interventions to promote self-disease management among adolescents and young adults is an emerging field that is worthy of exploration, and this approach represents an innovative way to develop adolescents and young adults' self-management skills and prepare them for the health care transition process [21,22].

To date, there has been little evidence produced about the effectiveness and usability of eHealth and mHealth interventions aimed at supporting self-management and health care transition in adolescents and young adults with chronic diseases. A systematic review conducted by Pérez et al [23] examined the literature published from 2015 to 2018 with the goal of evaluating the utility and effectiveness of mobile and web-based health apps that support self-management and transition among young people with chronic diseases, and 6 studies were ultimately included in that review. However, only a limited number of studies on this topic have been published, and it is difficult to draw comprehensive conclusions concerning the effectiveness and usage of mobile and web-based apps. Low and Manias [24] examined the usability of technology-based tools for supporting adolescents and young adults with chronic diseases by searching for data published between 1967 and 2019 in 3 electronic databases. However, among the included studies, most were published before 2015. Due to the rapid development of mobile medical technology and the increasingly mature forms of chronic disease management developed for adolescents and young adults, new interventions that are more suitable for self-management skill development may have emerged. In addition, although a meta-analysis was conducted in this study, the quantitative data was insufficient as it focused on the

XSL•FO
RenderX

experiences and perspectives of adolescents and young adults [24].

Objective

In summary, the effectiveness and usability of eHealth and mHealth interventions among adolescents and young adults with chronic diseases remain unclear. The aims of this systematic review are to provide more up-to-date evidence for the features and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic diseases. In addition, to understand and optimize the users' experience, we attempted to evaluate the acceptability, perceived level of usefulness, and user engagement of mHealth and eHealth interventions.

Methods

Study Design

The protocol for this systematic review was registered on PROSPERO (International Prospective Register of Systematic Review; registration number CRD42023378355). This study was reported per the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines [25].

Search Strategy

We searched five electronic databases, namely MEDLINE complete, Embase, Cochrane Library, CINAHL complete, Web of Science, and ProQuest Health & Medical complete. The end date for article searching was January 2024. We identified studies published in or since 2019 in these databases. In addition to the mentioned search strategies, we also manually searched reference lists of included studies to identify any additional studies that fit the inclusion criteria. The search strategy (Multimedia Appendix 1) was developed in consultation with an information scientist and used standardized indexed search terms and free-text terms about the following four key concepts: (1) adolescents or young adults, (2) chronic conditions, (3) transition or disease management, and (4) technology.

Inclusion and Exclusion Criteria

The studies included in this review met the inclusion criteria outlined in Textbox 1. The year of publication was limited between 2019 and 2024, as this review aims to update the 2 previous systematic reviews cited in the introductory section. The language restriction was attributed to the reviewers' proficiency in English. Exclusion criteria included (1) eHealth and mHealth interventions about the management of mental health, pain, acute cancer, or health risk behaviors; (2) interventions used only for testing equipment, such as a Bluetooth spirometer and blood glucose monitors, or interventions focused on remote health monitoring, such as patient portals and symptom reporting platforms; and (3) less than 50% of the sample involved adolescents or young adults aged 10 to 24 years.

Textbox 1. Eligibility criteria.

Population: Adolescents and young adults aged 10-24 years [26] who had been diagnosed with chronic conditions and who were either transitioning or had already transitioned to adult health care services.

Interventions: Any eHealth or mobile health interventions, such as digital tools, devices, systems, and resources, delivered through a web-based device to support the health care transition and aid adolescents and young adults in self-management.

Comparisons: Intervention versus ordinary care, pre-post or no comparator.

Outcomes: Any physiological, psychological, behavioral, attitudinal, or knowledge outcomes.

Study Design: Original quantitative or qualitative studies or mixed methods studies published in peer-reviewed journals in English.

Year of publication: Studies published in or since 2019.

Study Selection

A PRISMA diagram illustrating the detailed study selection process is shown in the Results section (PRISMA checklist is provided in Multimedia Appendix 2). The search results were collated using Endnote X9 (Clarivate) software and filtered to eliminate duplicates. One reviewer (ZL) was responsible for the first stage of screening, which focused on titles and abstracts in light of the research questions and the inclusion and exclusion criteria. Then, 2 reviewers (JW and YR) screened abstracts for possible inclusion in the full-text screening. In all cases, the decision to include or exclude a single study was approved by both reviewers. If these 2 reviewers could not agree on the decision, a final decision was made by the third reviewer (FL).

Data Extraction

A standardized table was used to extract the following data from each study: author information (name of the first author, year, and country), participant information (age, gender, and chronic condition), study characteristics (study design, sample size, duration, intervention media, intervention components, and quality assessment), and study outcomes. If any discrepancies between the reviewers emerged, they were resolved through discussion with the wider research team.

Quality Assessment

The Downs and Black checklist for randomized and nonrandomized studies was used to appraise the quality of the intervention efficacy trials [27]. This checklist takes into account 5 main assessment areas, namely, reporting, external validity, internal validity based on bias, internal validity based on cofounding and selection bias, and power, and it assigns trials an overall score, with the highest possible score being 28. A score of 24 to 28 was graded excellent, 19 to 23 was graded good, 14 to 18 was graded fair, and <14 was graded poor. Additionally, we assessed the explicitness of reporting of each

XSL•FO
RenderX

qualitative or quantitative questionnaire study to provide contextual details for readers to assess the transferability of this study's findings to their own settings. The 21-item standards for reporting qualitative research (SRQR) checklist was used for qualitative research [28], with a total score of 0-21 (yes=1, partially=0.5, and no=0), and a higher score indicates higher quality. A 16-item tool developed by Tong et al [29] was used for questionnaire research, with the highest possible score being 16, and a higher score indicates higher quality. Two reviewers (RB and YY) assessed the quality of the included studies independently. Any disagreements between these two authors were resolved through discussion among the research group and re-examination of the studies.

Data Synthesis and Analysis

Due to the heterogeneity exhibited by different studies and the different stages of intervention development, meta-analysis was deemed unsuitable for this systematic review. Instead, we used a narrative synthesis methodology to organize, explore, and present potential similarities and differences, associations, and patterns of data across different studies. For qualitative research, including exploratory and feasibility studies, mixed methods studies, and studies that involved interviews on focus groups, NVivo software (version 11; QSR International Pty Ltd) was used to extract and organize the data. The 3-stage thematic synthesis outlined by Thomas and Harden [30] was applied to the data synthesis and thematic analysis by one reviewer.

Results

Overview

Among a total of 16,758 studies identified based on the aforementioned search strategy, 1855 were removed due to duplication. Titles and abstracts of the remaining 14,903 articles were screened based on the inclusion criteria. Consequently, 14,507 articles were excluded, leaving 396 full-text papers that were reviewed for eligibility. Among them, 374 studies were excluded due to wrong research content (n=233), wrong research population (n=56), ineligible study designs (n=28), lack of full-text (n=15), protocol (n=36), and published in another language (n=6). Ultimately, 22 studies were included in the review [31-52] (Figure 1).

Figure 1. Study flow diagram.

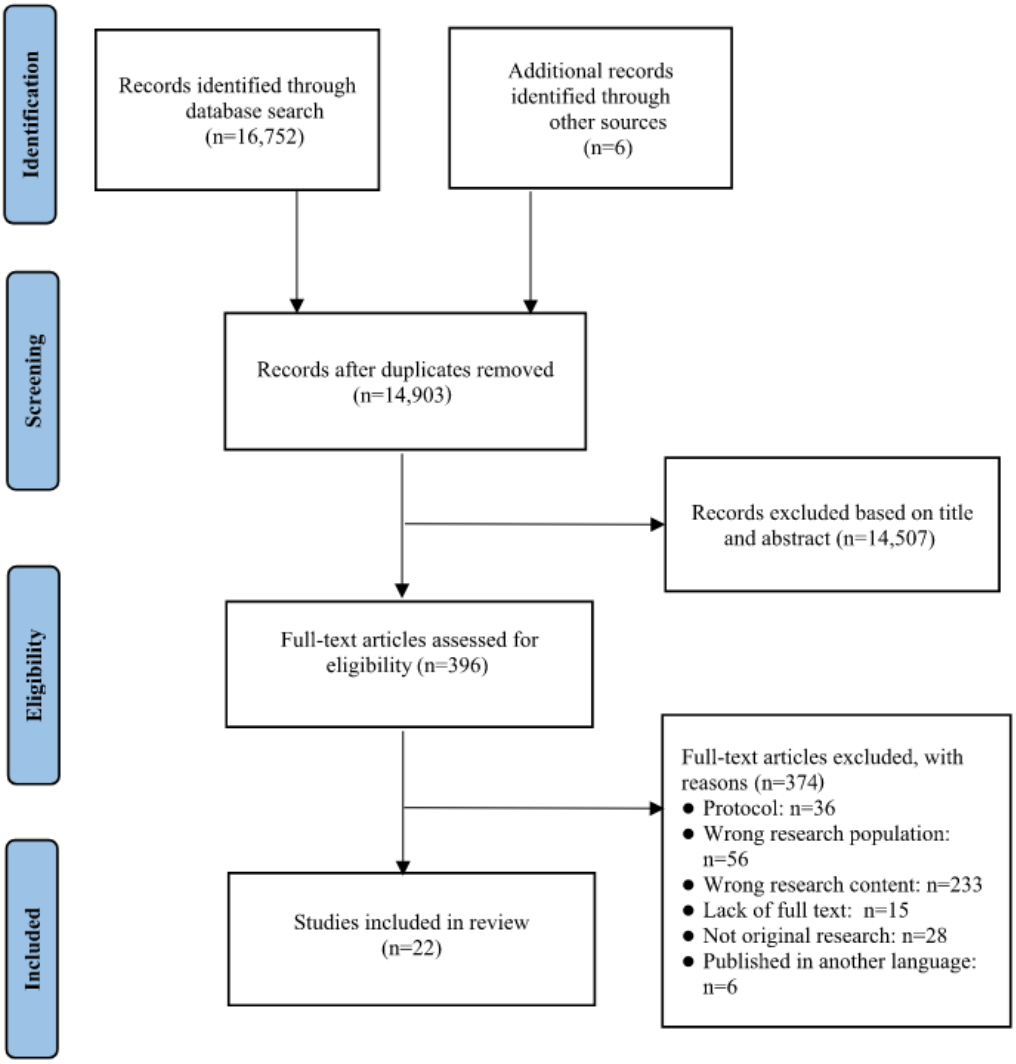

XSL•FO
RenderX

Participants

This review included 1272 participants. Asthma was the most frequently investigated type of chronic disease (n=6) [31,35,36,39,43,51], followed by organ transplantation (n=4) [40,42,44,47], sickle cell disease (n=3) [33,34,37], type 1 diabetes (n=3) [38,41,50], inflammatory bowel disease (n=2) [45,48], cancer (n=1) [32], heart disease (n=1) [52], and HIV (n=1) [49]; 1 study included children with diabetes, those with asthma, those with cerebral palsy, and those with congenital cardiac conditions [46]. Study sample sizes ranged from 4 to 234 participants, who varied in terms of age (range 9-25 years) and gender (female: 562/1214, 46.3%). The characteristics of the included studies are shown in Table 1.

XSL•FO
RenderX

Table 1. Characteristics of the included studies (N=22).

Author (year)	Country	Chronic condition	Study design	Age (years)			Number of participants	Outcome measures	Quality assessment (scoring rate, %)
				Range	Mean	n (%)			
Kosse et al (2019) [31]	Netherlands	Asthma	RCT[a]	• 12-18	15.1	• Male: 111(47.4)	• 234	• Medication adherence[b] • Disease control[c] • Quality of life[d]	• D&B[e]: 71.4
Schwartz et al (2019) [32]	United States	Survivors of Childhood Cancer	Multiphase iterative design and development of an intervention	• Phase 2: 15-21 • Phase 3: 15-29	• Phase 2: 19 • Phase 3: 18	• Phase 2: male: 4(40) • Phase 3: male: 1(12)	• Phase 2: 10 • Phase 3: 8	• Usability and acceptability	• Phase 2: D&B: 50 • Phase 3: SRQR[f]: 57.1
Crosby et al (2020) [33]	United Kingdom	Sickle cell disease	RCT	• 13-21	16	• Male: 25(47.1)	• 53	• Self-efficacy[g] • Self-management skills[h] • SCD knowledge[i] • Health motivation[j] • Quality of life[k]	• D&B: 75
Saulsberry et al (2020) [34]	United States	Sickle cell disease	A retrospective cohort study	• 12-25	14	• Male: 113(62)	• 183	• Disease knowledge[l] • Self-management confidence[m]	• Tong: 42.9
Nichols et al (2020) [35]	United States	Asthma	A qualitative study	• __[n]	• Phase 1: 10.9 • Phase 2: 12	• Phase 1: male: 5(36) • Phase 2: male: 6(75)	• Phase 1: 14 • Phase 2: 8	• Challenges and benefits associated with SAMS • Feasibility, acceptability, and preferences	• SRQR: 83.3
Schneider et al (2020) [36]	United States	Asthma	A qualitative study	• 12-17	14.4	• Male: 9(45)	• 20	• Usability	• SRQR: 78.6
Hood et al (2021) [37]	United Kingdom	Sickle cell disease	RCT	• 13-21	16.7	• Male: 24(46)	• 52	• User engagement[o] • Self-management skills[g] • Self-efficacy skills[p]	• D&B: 53.6

XSL•FO
RenderX

Author (year)	Country	Chronic condition	Study design	Age (years)			Number of participants	Outcome measures	Quality assessment (scoring rate, %)
				Range	Mean	n (%)			
Holtz et al (2021) [38]	United States	Type 1 diabetes	Pilot study (a preclinical or postclinical trial)	• 10-15	12.3	• Male: 13(52)	• 25 adolescents and 25 parents	• Adherence[q] • Quality of life[r] • Family conflict[s] • Satisfaction[t] • Use[u] • HbA1c levels[v]	• D&B: 60.7
Davis et al (2021) [39]	Australia	Asthma	A pilot study	• 15-24	19.7	• Male: 7(58)	• 12	• Asthma control[w] • Asthma quality of life[x] • Usefulness, satisfaction, and acceptability[y] • Self-management[z]	• SRQR: 45.2 • Tong: 50
Brookshire-Gay et al (2021) [40]	United States	Hematopoietic cell transplantation	Quantitative feasibility pilot study using questionnaires	• 11-24	18.1	• Male: 20(67)	• 30	• Use[aa] • Global distress[ab] • Hospital readmission rates	• Tong: 42.9
Butalia et al (2021) [41]	Canada	Type 1 diabetes	Nonrandomized trial	• 17-18	18	• Male: 102(50.2)	• 203	• Routine clinic visits • HbA_{1c} • Albumin and creatinine level ratio • Emergency department visits and hospitalization rates	• D&B: 71.4
Mehta et al (2021) [42]	United States	Liver transplant recipients	A qualitative study	• 12-18	• —	• Male: 3(75)	• 4	• —	• SRQR: 80.9
Fedele et al (2021) [43]	United States	Asthma	A pilot RCT	• 12-15	13.2	• Male: 18(54.5)	• 33		• D&B: 60.7

XSL•FO
RenderX

Author (year)	Country	Chronic condition	Study design	Age (years)						Number of participants	Outcome measures	Quality assessment (scoring rate, %)
				Range		Mean		n (%)				
											• Feasibility and acceptability • Asthma management[ac] • Asthma control[ad] • Quality of life[d] • Self-efficacy[ae] • Family communication[af]	
Sayegh et al (2022) [44]	United States	Liver transplants	A pilot RCT	• 13-21		• 16		• Male: 16(45.7)	• 35	• Medication adherence[ag] • Adherence motivation[ah] • Immunosuppressant level[ai] • Feasibility and acceptability	• D&B: 82.1	
Daraiseh et al (2022) [45]	United States	Inflammatory bowel disease	Multiphase, participatory user research study using individual interviews, and user evaluation	• Design phase: 14-25 • Usability testing: 14-20		—		• Male: 9(27.3)	• Design phase: 14 • Usability testing: 11	• Usability[aj] • System usability scale[ak]	• SRQR: 59.5 • Tong: 57.1	
Miller (2022) [46]	United States	Diabetes, asthma, cerebral palsy, and congenital cardiac disease	A pilot study	• 12-21		• —		• —	• 23	• Knowledge of medical conditions, health care system navigation, and identified health care provider and section	• SRQR: 52.4 • Tong: 28.6	
Kindem et al (2023) [47]	Norway	Solid organ transplantation	A preclinical or postclinical trial	• 14-25		• 14.5		• Male: 8(40)	• 20	• Feasibility • Medication adherenceal • Graft function	• D&B: 67.9	
Hommel et al (2023) [48]	United States	Inflammatory bowel disease		• 9-18		14.1		• Male: 13(62)	• 22		• D&B: 42.9	

XSL•FO
RenderX

Author (year)	Country	Chronic condition	Study design	Age (years)			Number of participants	Outcome measures	Quality assessment (scoring rate, %)
				Range	Mean	n (%)			
			Multiphase, iterative design, and development of an intervention					• Adherence[am] • Feasibility and acceptability • Quality of life[an] • Patient-reported symptoms[ao]	
Fomo et al (2023) [49]	United States	HIV	A qualitative study	• 14-19	16.2	• Male: 9(60)	• 15	• Perceived usefulness • Facilitating conditions	• SRQR: 80.9
Chiang et al (2023) [50]	China	Type 1 diabetes	Multiphase mixed methods design	• 16-25	• —	• —	• 35	• User engagement satisfaction[ap] • Heuristic evaluation[aq] • Think-aloud evaluation[ar]	• SRQR: 50 • Tong: 64.3
Ghozali et al (2023) [51]	Indonesia	Asthma	A preclinical or postclinical trial	• 18-22	• —	• Male: 65(46.4)	• 140	• Asthma control[as]	• D&B: 53.6
Han et al (2023) [52]	Canada	Heart disease	RCT	• 16-18	• —	• Male: 40(59)	• 68	• Transition readiness[at] • Frequency of use • Perceived usefulness	• D&B: 78.6

[a]RCT: randomized controlled trial.

[b]Self-reported medication adherence was measured using Medication Adherence Report Scale.

[c]Disease control was assessed with the Control of Allergic Rhinitis and Asthma Test, which contains ten questions on asthma and allergic rhinitis symptoms.

[d]Asthma related quality of life was assessed with the Pediatric Asthma Quality of Life Questionnaire.

[e]D&B: Downs and Black checklist.

[f]SRQR: standards for reporting qualitative research.

[g]Self-efficacy was assessed using the Patient Activation Measure.

[h]Self-management skills were assessed using the Transition Readiness Assessment Questionnaire and the University of North Carolina (UNC) TRxANSITION Scale.

[i]Sickle cell disease knowledge was assessed using a 25-item disease-specific knowledge questionnaire.

[j]Health motivation was assessed using the Treatment Self-Regulation Questionnaire.

[k]Health-related quality of life was assessed using the PedsQL (Pediatric Quality of Life Inventory Sickle Cell Disease Module).

[l]Knowledge assessment is paper-based and comprises 12 multiple-choice questions of the adolescent sickle cell disease-specific knowledge.

[m]Self-management confidence was assessed with the Self-Management Skills Checklist.

[n]Not available.

[o]Engagement in the mobile app was measured by (1) dividing the number of log-ins by the number of days with access to the app (eg, log-ins or access days) and (2) the calculating number of completed self-management goals.

[p]Self-efficacy skills was measured by the Patient Activation Measure.

[q]Adherence was measured by the Diabetes Behavior Rating Scale.

[r]Quality of life was self-reported by the adolescent using the PedsQL (Pediatric Quality of Life Inventory Sickle Cell Disease Module) generic scale.

XSL•FO
RenderX

[s]Family conflict was measured using the Revised Diabetes Family Conflict Scale.

[t]Satisfaction with the app was measured by using 8 questions from the Post-Study System Usability Questionnaire.

[u]Usage was measured by the number of days used, a day of usage was counted if the child entered the app, and the number of blood glucose entries per day and the number of messages sent per day were also considered.

[v]A repeat venipuncture for HbA_{1c} levels was performed at a laboratory after the intervention.

[w]Asthma control was measured by the Asthma Control Questionnaire.

[x]Asthma quality of life was measured using the Mini Asthma Quality of Life Questionnaire.

[y]Usefulness, satisfaction, and acceptability of the app were measured by the self-made questionnaire.

[z]The contribution of the app to asthma self-management was measured by the self-made questionnaire.

[aa]Use of roadmap was calculated by ratio variables to account for the differential length of access to Roadmap (version 1.0; Pi Network) across participants.

[ab]Global distress was measured using Profile of Mood States Second Edition, a 65-item measure.

[ac]Family asthma management was measured with the Family Asthma Management System Scale.

[ad]Asthma control was assessed using the asthma control test.

[ae]Asthma management self-efficacy was assessed with the Asthma Management Efficacy Questionnaire.

[af]Family communication was measured using the joint decision-making subscale of the Decision-Making Involvement Scale.

[ag]Medication adherence was measured by a 3-item visual analog scale.

[ah]Adherence motivation was measured by the Rollnick Readiness Ruler.

[ai]Immunosuppressant level was measured by laboratory blood draws to measure trough levels of immunosuppressant medications.

[aj]Usability testing using a research protocol.

[ak]System usability scale was measured by a system usability scale that is a simple, reliable, 10-item scale.

[al]Medication adherence was assessed by app-monitored registrations of Tac dose and time of dosing, measurement of trough Tac variability before, during, and after the study period, and patient self-report (BAASIS [Basel Assessment of Adherence to Immunosuppressive Medication Scale]-questionnaire and interview).

[am]Adherence was assessed by pill counts, pill counts of all inflammatory bowel disease medication prescribed to the patient were completed by patients or caregivers via the portal at assessment time points.

[an]Quality of life was assessed by PedsQL (Pediatric Quality of Life Inventory 4.0).

[ao]Patient-reported symptoms was assessed by the partial Harvey Bradshaw Index.

[ap]User interaction satisfaction was assessed by Questionnaire for User Interaction Satisfaction after engaging with the app for 4 weeks.

[aq]Heuristic evaluation included a display of system status; degree of correspondence with the real world; user control freedom; consistency; fool-proof and error-proof mechanisms; app problem resolution using cognition instead of memory; flexibility and efficiency of use; design aesthetics; error detection, debugging, and recovery from errors; and appropriate assistance and instructions.

[ar]Think-aloud evaluation included asking the users to speak loudly about (1) how to operate the app, (2) why the operation is completed this way, and (3) how they feel about it.

[as]Asthma Control Test questionnaire, each answer to a question is scored between 1 and 5, with the total score ranging from 5 to 25.

[at]Transition readiness was assessed by the Transition-Questionnaire.

Study Design

Among the included studies, 12 involved an intervention efficacy trial. A randomized controlled trial was the most commonly used design (n=4) [31,33,37,52], followed by a pilot randomized controlled trial (n=3) [32,43,44] and a pretest-posttest design (n=3) [38,47,51]. The remaining intervention efficacy trials used a single-arm open-label study design [48] or a nonrandomized controlled study design [41].

A total of 9 studies used a qualitative approach: 1 multiphase study that involved the iterative design and development of an intervention [32], 1 feasibility study [49], 1 usability study involving an intervention [35], 2 evaluations of user experiences [36,42], and 4 mixed methods studies [39,45,46,50]. In addition, 6 studies included a questionnaire component in the study design: 1 feasibility study [40], 1 evaluation of user engagement [34], and 4 mixed methods studies [39,45,46,50].

Quality Assessment

An evaluation of the comprehensiveness of reporting was also conducted for all studies. All mixed methods studies were assessed using 2 separate checklists [39,45,46,50]. One study, which was identified by the authors as using a mixed methods approach, involved multiple methods, including a participatory workshop, individual interviews, and user evaluations [45]. Overall, the reporting of questionnaires received considerably lower ratings (n=6; scores were derived using the questionnaire developed by Tong et al [29]: 4/14-9/14, with a scoring rate of 28.6%-64.3%) than the intervention efficacy trials (n=12; Downs and Black scores: 12/28-23/28, with a scoring rate of 42.9%-82.1%) or qualitative studies (n=9; SRQR scores: 9.5/21-17.5/21, with a scoring rate of 45.2%-83.3%).

Among the 12 studies that reported an intervention efficacy trial [31-33,37,38,41,43,44,47,48,51,52], 91.7% (11/12) were fair-to-good-quality studies [31-33,37,38,41,43,44,47,51,52] (Multimedia Appendix 3). Among the 9 qualitative research papers [32,35,36,39,42,45,46,49,50], 4 were missing at least 40% of the items that were identified as important on the SRQR checklist (Multimedia Appendix 4). All 6 studies that included a questionnaire were missing at least 40% of the items that were identified as important with the 16-item checklist of Tong et al

XSL•FO
RenderX

[29] for reporting questionnaire studies [34,39,40,45,46,50] (Multimedia Appendix 5).

Summary of Interventions

Among the 22 studies that described the intervention (Multimedia Appendix 6), 12 were evaluated in the context of an efficacy trial. All 12 evaluations focused on web-based interventions. Most interventions were delivered via a mobile app (n=18) [31-33,35-40,42,43,45-47,49-52]. Other modes of delivery included websites (n=2) [34,48] and telephones (n=2) [41,44]. In addition, only 27.3% (6/22) of the interventions were based upon extant theories [32,35,39,42,46,50], such as goal-setting theory [53], self-determination theory [54], social cognitive theory [55], and transition theory [56]. The interventions were summarized into 6 themes: medication monitoring and reminders, symptom tracking and monitoring, management goal setting, knowledge education and self-management skills training, incentives and reinforcement, and communication.

Theme 1: Medication Monitoring and Reminders

A total of 12 interventions focused on medication monitoring and reminders that targeted multiple aspects of nonadherent behavior, such as medication reminder or alarm [31,32,36,39,42,46], medication usage tracking [35,42,45,47,52], evaluations of medication adherence [31,43,47,48], positive feedback messages [42]. Two good-quality studies described the intervention designed to improve medication adherence. The “ADAPT” (Adolescent Adherence Patient Tool) app was securely connected to a desktop app connected to the patient’s own community pharmacist, which provided functions of medication reminder alarms, chats with the pharmacist, and two questions that were answered once every two weeks to monitor nonadherence [31]. TusenTac (University of Oslo) is a mobile app that features a tailored design for solid organ transplant recipients. In addition to serving as an immunosuppressive medication monitor, it offers the functions of medication reminder alerts and evaluations of medication adherence [47].

Theme 2: Symptom Tracking and Monitoring

The 9 interventions in this review involved symptom tracking and monitoring [31,33,35-37,43,45,50,52]. Specifically, the functions include asthma symptoms assessment and tracking [31,35,36,43,45,52], asthma status feedback [36], blood glucose levels tracking and reminders [50], and daily pain and mood symptoms assessment [33,37]. One of the good-quality interventions focused on tracking and monitoring symptoms based on a real time ecological momentary assessment of asthma symptoms [35]. In another fair-quality study, iManage was used to record progress per the daily pain and mood symptoms exhibited by adolescents with sickle cell disease [37].

Theme 3: Management Goal Setting

The 4 apps included in this review featured self-management goal setting [33,37,39,43]. Two studies focused on asthma patients [33,37], while the other two focused on sickle cell disease patients [39,43]. Two fair-quality studies described the interventions to promote self-management goal setting. AIM2ACT helped adolescents and caregivers identify an asthma management goal after a 1-week period focusing on the assessment of needs. Dyads were then guided through the process of behavioral contracting to outline the specific steps that each person would take to achieve the goal, how long the goal would take to complete, and the reward for accomplishing the goal [43]. iManage is a mobile app that users can create, monitor, and complete self-management goals (eg, exercising, taking medications, or sleeping). They could also link sickle cell disease symptoms to their goals in a visual calendar and see the progress of other users (eg, goals complete or not completed) [37].

Theme 4: Knowledge Education and Self-Management Skills Training

Ten of the interventions provided knowledge education and skills training [31-35,39,43,46,50,51]. The content of information and skills training includes disease-related information [31-33,39,46,50,51], information concerning specific topics of interest to adolescents (such as career development, sex, or pregnancy) [50], transition knowledge and skills [34], asynchronous inhaler use technique [35], and communication trainings with caregivers on asthma management needs [43]. The modes of delivery include movies [31], videos [34,35,43,51], e-books [39,50], phone calls [33], text information [32,51], and websites [46]. One good-quality study described short educational and motivational movies on asthma-related topics [31]. In another fair-quality study, AIM2ACT is a dyadic mHealth intervention providing separate and tailored skills-training videos for adolescents and caregivers (eg, how adolescents can effectively communicate their asthma management needs to caregivers [43].

Theme 5: Incentives and Reinforcement

There are a total of 6 of the interventions that provided rewards to encourage and reinforce self-management behaviors [32,34,38,42,47,50]. The forms of rewards include point rewards [32,38,50], small toys [34], and gamification features [34,42,47]. Two fair-to-good-quality studies described incentive interventions to stimulate continuous use. Users of the “MyT1DHero” app could redeem points for accessories for their “hero” avatar on the app, including capes, boots, logos, masks, and different hair colors and styles [38]. The TusenTac app was age-adapted by including different fun facts that appeared after medication registration; furthermore, it featured a tailored “transplant-designed” gamification system with challenges [47]. A total of 4 of the interventions used cognitive strategies to promote behavioral activation included motivational interviewing [35,37], cognitive behavioral strategies [37,39], and praise text messages [44]. One good-quality study provided 6 weekly, 90-minute group sessions guided by psychologists, which included culturally sensitive motivational interviewing and cognitive behavioral strategies with the goal of enhancing behavioral activation [37]. In another good-quality study, Sayegh et al [44] provided text messages containing praise to adolescents whose laboratory tests indicated immunosuppressant medications within the expected range with the goal of improving their medication adherence.

XSL•FO
RenderX

Theme 6: Communication

A total of 7 mobile apps included the feature to facilitate social media interactions with parents, peers, and health care providers [31,33,38,42,45,50,51]. In one good-quality study, the ADAPT app provided functions of peer chat and pharmacist chat functions to facilitate contact [31]. In another fair-quality study, the MyT1DHero app focused on parent-child interactions to promote positive communication regarding T1D management through 2 separate app interfaces—one for the adolescent and one for the parent. In addition, there are also peer support functions by providing videos of other adolescents with T1D telling their stories and affirming messages [38].

Outcomes

Study outcomes varied based on the stage of development (feasibility, usability, efficiency, and effectiveness) of the intervention in question. The usability outcomes and health outcomes of this research were summarized per the aim of each intervention (Multimedia Appendices 7-10).

Perceived Usability or Acceptability

Perceived usability was explored in the context of 11 interventions [32,35,36,39,40,43,45,48-50,52]. Most of these interventions (n=10) [32,35,36,39,43,45,48-50,52] were evaluated positively by the participants. Most participants felt that the apps were acceptable because they were convenient, easy to use, and easy to access when they needed help managing their health [32,35,36,39,43,45,49,50,52].

User Engagement

The overall levels of user engagement during the intervention were variable [32,35,39,40,43,48,52]. Schwartz et al [32] found that youth engagement with the app ranged from 19.3% to 98.2% of app intervention days, with a median of 63.9% of days actively using the app. Davis et al [39] found that over 6 weeks, without follow-up appointments, telephone calls, or reminders from the research staff, 33% (4/12) of participants used the app one to five times, 25% (3/12) participants used it six to ten times, and 16.7% (2/12) participants used it more than 10 times. Some studies indicated a gradual decline in youth engagement with these apps over time [40,52]. Brookshire-Gay et al [40] showed that, among users (n=22), engagement rates with Roadmap (version 1.0) were 96% and 91% for the first 2 weeks, respectively; however, this percentage declined to 58% (n=12) by week 4. Han et al [52] reported that only 30%-66% of participants either sometimes or frequently used the app at the 3-month mark. Some studies identified barriers preventing user engagement, including time-consuming video uploads [35], noncontinuous access to a phone, literacy levels, limited free time [36], font size, language, and false impressions [49].

Preferred Delivery Method

Adolescents and young adults' preferred delivery method involved visually appealing features [35,36,39,49,50]. For example, colors, backgrounds, graphics, and fun or entertaining elements, such as games, avatars, incentives, design elements, and additional customization options, encouraged them to use these apps more frequently and to be more engaged with the apps. Schneider et al [36] found that many adolescents and young adults appreciated the options to change the wallpaper color and designs on the page and to customize the "emoji" faces they used to express how they felt at the time of data entry.

Preferred Features

Suggestions regarding the inclusion of disease-specific information were made by adolescents and young adults [35,39]. For example, patients with asthma mentioned information concerning strategies to support the self-management of asthma, general information about asthma, and role model (celebrity or athletes) testimonials concerning how they coped with asthma. Communication was mentioned at several levels [35,36], including communication with peers, health care providers, parents, and app teams. Many adolescents and young adults reported that they would appreciate the opportunity to connect with their peers and share their feelings via social media, while some adolescents and young adults appropriated being able to contact or share information with their health care providers [35,36]. Although the use of technology, whether through messaging or video interactions, was strongly preferred by most adolescents and young adults as their primary means of routine communication and interaction, none of the participants wanted to completely replace traditional face-to-face encounters [35].

Health Outcomes

Symptom Control

Three fair-to-good-quality studies reported symptom control assessments, which focused on adolescents and young adults with asthma (aged 12-22 years) [31,43,51]. One fair-quality study conducted asthma control assessments at baseline, postintervention, and 4-month follow-up. Results showed that participants randomized to the intervention group (AIM2ACT) had significant improvements (P=.04) in asthma control scores compared to the control group [43]. Another fair-quality study revealed that there was a significant difference (P<.001) in the pretest and posttest scores of the asthma control score from the intervention group, while no difference was found compared to the control group [51]. In contrast to these above results, one good-quality study [31] found that after six months of access to the ADAPT intervention, no effect was observed on asthma control compared to the control group (P>.05).

Self-Reported Medication Adherence

Four fair-to-good-quality studies reported medication adherence [31,38,44,47] through self-report questionnaires [31,38,47], the medication level variability index [44,47], as well as app-monitored registrations of medication dose and time of dosing [47]. Two good-quality studies reported that mHealth-based intervention had a positive effect on medication adherence compared to the control group (P<.05) [31,44]. Different from the above results, two fair-to-good-quality preclinical or postclinical trials demonstrated that there was no significant change in medication adherence after the short-term intervention (P>.05) [38,47]. In addition, Kindem et al [47] found that four of eleven (36%) who were nonadherent assessed at inclusion turned adherent during the intervention period, and after the intervention, 70% reported improved timing-adherence at the interview.

XSL•FO
RenderX

Quality of Life

Among the five studies to focus on this topic, three (67%) fair-to-good-quality studies reported on quality of life [31,38,43]. Among these 3 studies, 2 (75%) focused on adolescents and young adults with asthma (aged 12-18 years). One fair-quality study [43] found participants randomized to the AIM2ACT cohort had significant improvements in asthma-related quality of life at postintervention (*P*=.002) and at the 4-month follow-up (*P*=.002) compared to control that surpassed the minimally clinically important difference threshold. However, the good-quality study conducted by Kosse et al [31] reported that compared to the control group, no intervention effect was found on asthma-related quality of life at baseline and at the 6-month follow-up (*P*>.05). In a fair-quality preclinical or postclinical trial, significant benefits were demonstrated in quality of life of adolescents and young adults with type 1 diabetes (*P*=.001) after 12 weeks of the MyT1DHero app intervention [38].

Disease Knowledge

Two studies reported on disease knowledge [33,34], which was measured using questionnaires. Saulsberry et al [34] found that there was a positive correlation between participation rate of Sickle Cell Transition E-Learning Program (STEP) intervention (6-module tool) and disease knowledge scores (*P*=.003), and participants who completed ≥3 STEP modules had higher disease knowledge scores compared with those who completed <3 STEP modules (*P*=.007) [34]. However, another good-quality randomized controlled trial showed that although the knowledge scores of participants significantly increased compared to baseline after 6 weeks of SCThrive (sickle cell disease self-management intervention) intervention (*P*<.001), there was no statistically significant difference compared to the control group (*P*>.05) [33].

Self-Management Development

Four studies reported on self-management development [33,34,37,43], which was measured in terms of self-management confidence [34], self-management skills [33,37], and self-management efficacy [33,37,43]. One fair-quality randomized controlled trial [37] found that the number of logins to the mobile app significantly predicted adolescents and young adults-reported self-management skills (*P*<.05), while no correlation was found in the retrospective cohort study [34] between the number of modules completed and the self-management confidence. One fair-quality study [43] indicated that compared to the control group, no intervention effect was found on self-efficacy at baseline and at the 4-month follow-up (*P*>.05), consistent with research findings of the good-quality study conducted by Crosby et al [33]. Moreover, Crosby et al [33] found that although no significant change was found in overall self-management after SCThrive intervention (*P*>.05), there was statistically significant improvement in one self-management skill (tracking health; *P*=.001).

Discussion

Principal Findings

This systematic review examined 22 articles that were published between 2019 and 2023 and focused on adolescents and young adults aged 9 to 25 years. Despite the proliferation of eHealth and mHealth interventions during the past 5 years [57], the overall evidence regarding the efficacy, cost-effectiveness, and long-term health benefits of these interventions is minimal, as most of the interventions identified were in the usability testing stage.

Using the qualitative data, this systematic review revealed most adolescents and young adults felt the interventions delivered by mobile apps or websites were feasible, acceptable, and useful, as they were convenient and easy to use in the context of helping adolescents and young adults manage their health [32,35,36,39,43,45,49,50,52]. Low and Manias [24] confirmed that adolescents and young adults were receptive to receiving health information electronically. In fact, multiple studies have shown that most adolescents and young adults have access to mobile phones that they keep with them at all times and they are interested in using mHealth and apps to manage their health and facilitate the health care transition [58,59]. However, engaging adolescents and young adults in research could be difficult. In this system review, we found that the overall levels of engagement during the intervention were variable [32,35,39,40,43,48,52] and tended to decrease over time [40,52]. Some adolescents and young adults could be defined as “super users” who logged into the app nearly every day, while others engaged with the app infrequently [37,39]. Previous studies have shown that early involvement promotes better engagement [60,61], and when making choices among many apps, users prefer certain features tailored to their needs and preferences [62,63]. This finding highlights the importance of stakeholder involvement in the design process, as adolescents are undergoing a critical period concerning the development of routines and self-management skills, and digital interventions designed for adults may not address the unique developmental and psychosocial barriers faced by adolescents [64]. In addition, a previous study indicated that increasing caregiver involvement can effectively improve the engagement of adolescents and young adults in digital interventions [65]. Considering the high risk that adolescents and young adults will not comply with the intervention, allowing them to designate people other than their caregivers (ie, older siblings or friends) to increase accountability may help them assert their independence from their caregivers.

Adolescents and young adults had different preferred styles of message delivery and functions. Specifically, adolescents and young adults have partiality for more engaging elements (such as visually appealing features, customization options, or fun or entertaining functions) [35,36,39,49,50], disease-specific information [35,39], and opportunities to communicate with their peers, health care providers, and app teams [35,36]. Education is an important aspect of the task of promoting the development of self-management skills [66]. The continuous highlighting of disease-specific educational themes via a

XSL•FO
RenderX

web-based device using video or text can enable patients to easily access reliable information. Interacting with peers and health care providers via social media may promote better outcomes [67]. Adolescents and young adults would prefer to network with their peers, which would enable them to share their experiences and offer each other emotional and informational support. Previous studies have proved that face-to-face support groups and peer guidance were underused but highly anticipated resources among adolescents and young adults with chronic diseases, which should be fully explored in the self-management and health care transition in the future [68]. In addition, some adolescents and young adults preferred to share information or discuss their concerns with their health care providers. Dwyer-Matzky et al [69] found that 59% of adolescents and young adults with chronic diseases would like to establish close connections with medical teams, which has been proven to effectively improve treatment adherence.

From the perspective of intervention effectiveness, although digital health interventions have great potential in the context of adolescent disease management, some results have indicated insufficient evidence of improvement in terms of adolescent self-management behavior and biomedical outcomes. For example, some fair-to-good-quality randomized controlled trials showed that compared to the control group, there were no significant improvements in asthma control scores [31], quality of life [31], disease knowledge [33], and self-efficacy [33,43] at postintervention ($P>.05$). A meta-analysis conducted by Low and Manias [24] did not reveal any differences in the impact of interventions provided through mobile apps and websites on quality of life, self-management, which could mainly be attributed to the lack of high-quality randomized controlled trials. Due to the small number and heterogeneity of the studies, meta-analysis was precluded in this study. Therefore, we cannot draw a conclusion about the efficacy of the eHealth and mHealth interventions on symptom control, self-reported medication adherence, quality of life, disease knowledge, and self-management development from this review. Future large-scale, high-quality randomized controlled trials and investigations of the efficacy and sustainability of eHealth and mHealth interventions should be conducted. In addition, among the included studies, only one study [52] focused on the health care transition process and evaluated transition readiness. In the systematic review conducted by Pérez et al [23], none of the studies measured the transition process as an outcome. Evaluating transition readiness has been widely regarded as an important component of optimizing transition outcomes due to its ability to identify obstacles to transition, individually plan for treatment, and monitor progress over time [70]. More attention should be paid to the transitional outcomes especially transition readiness of adolescents and young adults with chronic diseases to facilitate smooth health care transition and optimize patient health outcomes.

From a cost-effectiveness perspective, eHealth and mHealth technologies may exhibit greater cost-effectiveness than face-to-face consultation or clinical medical care. Due to its ability to overcome spatial limitations, this approach can greatly reduce the time and transportation costs for both users and professional medical personnel [71]. Nevertheless, little is known about the cost-effectiveness of developing and maintaining such technology-based interventions. The World Health Organization strongly recommends that cost analyses of interventions should be conducted [72]. Relevant costs include long-term direct and indirect costs, ranging from digital health intervention development to training and implementation, as well as the ultimate benefits for patients and health care systems, such as improving patient health outcomes and reducing human resource costs. However, no studies included in this review evaluated the economic characteristics and cost-effectiveness of digital interventions, while only one study made a reference to cost in the review by Pérez et al [23]. Two systematic reviews indicated a lack of economic data to support eHealth and mHealth interventions among adolescents and young adults with chronic diseases [22,73]. Butler et al [22] suggested considering costs as early as possible during the process of prototype development to help make strategic decisions, while Badawy et al [73] highlighted the need for a comprehensive economic evaluation of technology-based interventions to facilitate more evidence-based assessments of the scalability, sustainability, and benefits of broader investment of such technology tools among adolescents and young adults with chronic diseases.

Limitations

This review has several limitations. First, our search strategy was limited to an academic context, and we focused on electronically indexed health databases published in peer-reviewed journals rather than apps contained in commercial stores. Second, the studies included in this review used quantitative, qualitative, and mixed methods, which made it difficult to compare their quality. As the heterogeneity was observed across different studies and the different stages of intervention development (feasibility, usability, efficiency, and effectiveness) on which they focused, we used a descriptive synthesis-based methodology. This form of analysis can be subjective, and it faces the risk of reporting bias. To mitigate this risk and improve transparency, all the authors reviewed all the stages of the data analysis. Finally, some relevant studies might have been missed because only studies published in English were included. It was acknowledged that a search including foreign language databases may reveal additional studies published in languages other than English in low- and middle-income countries. Given these limitations, the findings of this systematic review should be interpreted with caution.

Conclusions

This systematic review revealed that adolescents and young adults were receptive to and interested in receiving information and managing their health using a mobile app or website. It should be noted that adolescents and young adults had different preferred styles of message delivery and functions. Therefore, to provide an age-appropriate, reliable condition-specific resource and improve patient engagement during the transition process, the best approach would be to involve adolescents and young adults early in the design process to identify their specific needs and preference. This would be better to be coupled with or followed by obtaining suggestions from health care providers and app teams. As most of the studies remained in the early

XSL•FO
RenderX

stages of exploration, there remain limited data about the effectiveness of eHealth and mHealth interventions facilitating self-management and health care transition of adolescents and young adults with chronic diseases. Large sample, multicenter randomized controlled trials should be conducted in the future to verify the effectiveness of eHealth and mHealth interventions. Moreover, future studies could pay more attention to the transitional outcomes (especially transition readiness), and cost-effectiveness of developing and maintaining eHealth and mHealth interventions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File , 15 KB - jmir_v26i1e56556_app1.docx]

Multimedia Appendix 2
PRISMA checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[DOCX File , 28 KB - jmir_v26i1e56556_app2.docx]

Multimedia Appendix 3
Quality appraisal of intervention efficacy trials using the Down and Black checklist.
[DOCX File , 19 KB - jmir_v26i1e56556_app3.docx]

Multimedia Appendix 4
Quality appraisal of the qualitative component of the studies.
[DOCX File , 17 KB - jmir_v26i1e56556_app4.docx]

Multimedia Appendix 5
Quality appraisal of the questionnaire component of the studies.
[DOCX File , 15 KB - jmir_v26i1e56556_app5.docx]

Multimedia Appendix 6
Characteristics of included interventions.
[DOCX File , 21 KB - jmir_v26i1e56556_app6.docx]

Multimedia Appendix 7
A summary of usability study or users' perceived acceptance of the intervention.
[DOCX File , 17 KB - jmir_v26i1e56556_app7.docx]

Multimedia Appendix 8
Main outcomes.
[DOCX File , 15 KB - jmir_v26i1e56556_app8.docx]

Multimedia Appendix 9
Synthesis of qualitative results.
[DOCX File , 19 KB - jmir_v26i1e56556_app9.docx]

Multimedia Appendix 10
A summary of intervention efficacy evaluations.
[DOCX File , 19 KB - jmir_v26i1e56556_app10.docx]

References

1. Mokkink LB, van der Lee JH, Grootenhuis MA, Offringa M, Heymans HSA, Dutch National Consensus Committee Chronic DiseasesHealth Conditions in Childhood. Defining chronic diseases and health conditions in childhood (0-18 years of age): national consensus in the Netherlands. Eur J Pediatr 2008;167(12):1441-1447. [doi: 10.1007/s00431-008-0697-y] [Medline: 18340463]

XSL•FO
RenderX

2. Perrin JM, Anderson LE, Van Cleave J. The rise in chronic conditions among infants, children, and youth can be met with continued health system innovations. Health Aff (Millwood) 2014;33(12):2099-2105. [doi: 10.1377/hlthaff.2014.0832] [Medline: 25489027]
3. Dall'Oglio I, Gasperini G, Carlin C, Biagioli V, Gawronski O, Spitaletta G, et al. Self-care in pediatric patients with chronic conditions: a systematic review of theoretical models. Int J Environ Res Public Health 2021;18(7):3513 [FREE Full text] [doi: 10.3390/ijerph18073513] [Medline: 33800684]
4. Ferro MA, Boyle MH. The impact of chronic physical illness, maternal depressive symptoms, family functioning, and self-esteem on symptoms of anxiety and depression in children. J Abnorm Child Psychol 2015;43(1):177-187. [doi: 10.1007/s10802-014-9893-6] [Medline: 24938212]
5. Li ZR, Wang H, Lu F, Dong L, Wu J. The experiences of adolescent solid organ transplantation recipients, parents, and healthcare professionals in healthcare transition: a qualitative systematic review. J Pediatr Nurs 2023;71:e154-e164. [doi: 10.1016/j.pedn.2023.05.015] [Medline: 37270387]
6. Gumidyala AP, Greenley RN, Plevinsky JM, Poulopoulos N, Cabrera J, Lerner D, et al. Moving on: transition readiness in adolescents and young adults with IBD. Inflamm Bowel Dis 2018;24(3):482-489. [doi: 10.1093/ibd/izx051] [Medline: 29462383]
7. Longhi D, Brown M, Fromm Reed S. Community-wide resilience mitigates adverse childhood experiences on adult and youth health, school/work, and problem behaviors. Am Psychol 2021;76(2):216-229. [doi: 10.1037/amp0000773] [Medline: 33734790]
8. White PH, Cooley WC, Transitions Clinical Report Authoring Group, American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians. Supporting the health care transition from adolescence to adulthood in the medical home. Pediatrics 2018;142(5):e20182587. [doi: 10.1542/peds.2018-2587] [Medline: 30348754]
9. Uzark K, Afton K, Yu S, Lowery R, Smith C, Norris MD. Transition readiness in adolescents and young adults with heart disease: can we improve quality of life? J Pediatr 2019;212:73-78. [doi: 10.1016/j.jpeds.2019.04.060] [Medline: 31182220]
10. Sawyer SM, Afifi RA, Bearinger LH, Blakemore SJ, Dick B, Ezeh AC, et al. Adolescence: a foundation for future health. Lancet 2012;379(9826):1630-1640. [doi: 10.1016/S0140-6736(12)60072-5] [Medline: 22538178]
11. Sawyer SM, Aroni RA. Self-management in adolescents with chronic illness. What does it mean and how can it be achieved? Med J Aust 2005;183(8):405-409. [doi: 10.5694/j.1326-5377.2005.tb07103.x] [Medline: 16225444]
12. Gold A, Martin K, Breckbill K, Avitzur Y, Kaufman M. Transition to adult care in pediatric solid-organ transplant: development of a practice guideline. Prog Transplant 2015;25(2):131-138. [doi: 10.7182/pit2015833] [Medline: 26107273]
13. Li ZR, Wang HF, Lu FY, Lv LY, Bao RJ, Dong L. Research progress on transition readiness for liver transplant recipients from adolescents to adults. Chin J Nurs 2023;58(7):875-880. [doi: 10.3761/j.issn.0254-1769]
14. Nassetta K, Hussain T, Gambetta K, Le K, O'Dwyer LC, Badawy SM. A systematic review of adherence to immunosuppression among pediatric heart transplant patients. J Cardiovasc Dev Dis 2022;9(5):165 [FREE Full text] [doi: 10.3390/jcdd9050165] [Medline: 35621876]
15. Allegrante JP. Advancing the science of behavioral self-management of chronic disease: the arc of a research Trajectory. Health Educ Behav 2018;45(1):6-13. [doi: 10.1177/1090198117749010] [Medline: 29258348]
16. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of technology-based strategies to promote engagement with digital interventions: a systematic review protocol. JMIR Res Protoc 2015;4(2):e47 [FREE Full text] [doi: 10.2196/resprot.3990] [Medline: 25921274]
17. Vaala SE, Hood KK, Laffel L, Kumah-Crystal YA, Lybarger CK, Mulvaney SA. Use of commonly available technologies for diabetes information and self-management among adolescents with type 1 diabetes and their parents: a web-based survey study. Interact J Med Res 2015;4(4):e24 [FREE Full text] [doi: 10.2196/ijmr.4504] [Medline: 26715191]
18. Government OC, Statistics C. Smartphone use and smartphone habits by gender and age group, inactive. Government of Canada, Statistics Canada. 2021. URL: https://doi.org/10.25318/2210011501-eng [accessed 2022-10-10]
19. Pew RCI, Science. Teens, Social Media & Technology. 2018. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2024-10-30]
20. Smahel D, Machackova H, Mascheroni G. EU kids online 2020: survey results from 19 countries. EU Kids Online. URL: https://www.doi.org/10.21953/lse.47fdeqj01ofo [accessed 2024-10-30]
21. Cafazzo JA, Casselman M, Hamming N, Katzman DK, Palmert MR. Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study. J Med Internet Res 2012;14(3):e70 [FREE Full text] [doi: 10.2196/jmir.2058] [Medline: 22564332]
22. Butler S, Sculley D, Santos DS, Fellas A, Gironès X, Singh-Grewal D, et al. Usability of eHealth and mobile health interventions by young people living with juvenile idiopathic arthritis: systematic review. JMIR Pediatr Parent 2020;3(2):e15833. [doi: 10.2196/15833] [Medline: 33258786]
23. Pérez YIV, Medlow S, Ho J, Steinbeck K. Mobile and web-based apps that support self-management and transition in young people with chronic illness: systematic review. J Med Internet Res 2019;21(11):e13579 [FREE Full text] [doi: 10.2196/13579] [Medline: 31746773]

XSL•FO
RenderX

24. Low JK, Manias E. Use of technology-based tools to support adolescents and young adults with chronic disease: systematic review and meta-analysis. JMIR mHealth uHealth 2019;7(7):e12042 [FREE Full text] [doi: 10.2196/12042] [Medline: 31322129]
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. World Health Organization. Maternal, newborn, child and adolescent health. URL: https://www.who.int/teams/maternal-newborn-child-adolescent-health-and-ageing [accessed 2024-10-30]
27. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52(6):377-384 [FREE Full text] [doi: 10.1136/jech.52.6.377] [Medline: 9764259]
28. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
29. Tong A, Wong G, Hodson E, Walker RG, Tjaden L, Craig JC. Adolescent views on transition in diabetes and nephrology. Eur J Pediatr 2013;172(3):293-304. [doi: 10.1007/s00431-012-1725-5] [Medline: 22576804]
30. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
31. Kosse RC, Bouvy ML, de Vries TW, Koster ES. Effect of a mHealth intervention on adherence in adolescents with asthma: a randomized controlled trial. Respir Med 2019;149:45-51 [FREE Full text] [doi: 10.1016/j.rmed.2019.02.009] [Medline: 30803885]
32. Schwartz LA, Psihogios AM, Henry-Moss D, Daniel LC, Ver Hoeve ES, Velazquez-Martin B, et al. Iterative development of a tailored mHealth intervention for adolescent and young adult survivors of childhood cancer. Clin Pract Pediatr Psychol 2019;7(1):31-43 [FREE Full text] [doi: 10.1037/cpp0000272] [Medline: 38882593]
33. Crosby LE, Hood A, Kidwell K, Nwankwo C, Peugh J, Strong H, et al. Improving self-management in adolescents with sickle cell disease. Pediatr Blood Cancer 2020;67(10):e28492 [FREE Full text] [doi: 10.1002/pbc.28492] [Medline: 32697889]
34. Saulsberry AC, Hodges JR, Cole A, Porter JS, Hankins J. Web-based technology to improve disease knowledge among adolescents with sickle cell disease: pilot study. JMIR Pediatr Parent 2020;3(1):e15093 [FREE Full text] [doi: 10.2196/15093] [Medline: 31909718]
35. Nichols M, Miller S, Treiber F, Ruggiero K, Dawley E, Teufel Ii R. Patient and parent perspectives on improving pediatric asthma self-management through a mobile health intervention: pilot study. JMIR Form Res 2020;4(7):e15295 [FREE Full text] [doi: 10.2196/15295] [Medline: 32442127]
36. Schneider T, Baum L, Amy A, Marisa C. I have most of my asthma under control and I know how my asthma acts: users' perceptions of asthma self-management mobile app tailored for adolescents. Health Informatics J 2020;26(1):342-353 [FREE Full text] [doi: 10.1177/1460458218824734] [Medline: 30732520]
37. Hood AM, Nwankwo C, Walton A, McTate E, Joffe N, Quinn CT, et al. Mobile health use predicts self-efficacy and self-management in adolescents with sickle cell disease. Transl Behav Med 2021;11(10):1823-1831 [FREE Full text] [doi: 10.1093/tbm/ibab041] [Medline: 33949674]
38. Holtz B, Mitchell KM, Holmstrom AJ, Cotten SR, Dunneback JK, Jimenez-Vega J, et al. An mHealth-based intervention for adolescents with type 1 diabetes and their parents: pilot feasibility and efficacy single-arm study. JMIR mHealth uHealth 2021;9(9):e23916 [FREE Full text] [doi: 10.2196/23916] [Medline: 34519670]
39. Davis SR, Peters D, Calvo RA, Sawyer SM, Foster JM, Smith LD. A consumer designed smartphone app for young people with asthma: pilot of engagement and acceptability. J Asthma 2021;58(2):253-261. [doi: 10.1080/02770903.2019.1680997] [Medline: 31691618]
40. Brookshire-Gay K, LaLonde L, Byrd M, Neenan A, Seyedsalehi S, Hanauer DA, et al. Health information technology utilization by adolescent and young adult aged inpatients undergoing hematopoietic cell transplantation. J Adolesc Young Adult Oncol 2021;10(1):100-104 [FREE Full text] [doi: 10.1089/jayao.2020.0079] [Medline: 32716668]
41. Butalia S, Crawford SG, McGuire KA, Dyjur DK, Mercer JR, Pacaud D. Improved transition to adult care in youth with type 1 diabetes: a pragmatic clinical trial. Diabetologia 2021;64(4):758-766. [doi: 10.1007/s00125-020-05368-1] [Medline: 33439284]
42. Mehta P, Moore SL, Bull S, Kwan BM. Building—A mobile health application to improve adolescent medication adherence—using a multidisciplinary approach and academic-industry collaboration. Digit Health 2021;7:20552076211019877 [FREE Full text] [doi: 10.1177/20552076211019877] [Medline: 34104467]
43. Fedele DA, Thomas JG, McConville A, McQuaid EL, Voorhees S, Janicke DM, et al. Using mobile health to improve asthma self-management in early adolescence: a pilot randomized controlled Trial. J Adolesc Health 2021;69(6):1032-1040. [doi: 10.1016/j.jadohealth.2021.06.011] [Medline: 34274211]
44. Sayegh C, Im D, Moss IK, Urquiza R, Patel S, Thomas DW. Randomized pilot trial of praise text messages to improve medication adherence among adolescents and young adults with liver transplants. Pediatr Transplant 2022;26(7):e14361 [FREE Full text] [doi: 10.1111/petr.14361] [Medline: 35854423]

XSL•FO
RenderX

45. Daraiseh NM, Black A, Minar P, Meisman A, Saxe M, Lipstein EA. iBDecide: a web-based tool to promote engagement in shared decision-making among adolescents with ulcerative colitis. Patient Educ Couns 2022;105(6):1628-1633. [doi: 10.1016/j.pec.2021.10.023] [Medline: 34756638]
46. Miller C. Using technology to bridge the transition from pediatric to adult health care. J Am Assoc Nurse Pract 2022;34(6):850-858. [doi: 10.1097/JXX.0000000000000700] [Medline: 35383652]
47. Kindem IA, Åsberg A, Midtvedt K, Bjerre A. Optimizing medication adherence with home-monitoring—a feasibility study using capillary microsampling and mHealth in solid organ-transplanted adolescents. Pediatr Transplant 2023;27(7):e14590. [doi: 10.1111/petr.14590] [Medline: 37543722]
48. Hommel KA, Ramsey RR, Gray WN, Denson LA. Digital therapeutic self-management intervention in adolescents with inflammatory bowel disease. J Pediatr Gastroenterol Nutr 2023;76(1):38-42. [doi: 10.1097/MPG.0000000000003623] [Medline: 36123761]
49. Fomo MF, Newman J, Sibaya T, Ndlela N, Hussen S, Archary M, et al. A qualitative assessment of the perceived acceptability and feasibility of eHARTS, a mobile application for transition readiness assessment for adolescents living with HIV in South Africa. PLOS Digit Health 2023;2(6):e0000272 [FREE Full text] [doi: 10.1371/journal.pdig.0000272] [Medline: 37327198]
50. Chiang YT, Chang CW, Yu HY, Tsay PK, Lo FS, Chen CW, et al. Developing the "Healthcare CEO App" for patients with type 1 diabetes transitioning from adolescence to young adulthood: a mixed-methods study. Nurs Open 2023;10(3):1755-1766 [FREE Full text] [doi: 10.1002/nop2.1432] [Medline: 36289575]
51. Ghozali MT, Mohany M, Milošević M, Satibi, Kurniawan M. Impact of a mobile-app assisted self-management educational intervention on the scores of asthma control test (ACT) questionnaire among young asthmatic patients. Res Social Adm Pharm 2023;19(10):1354-1359. [doi: 10.1016/j.sapharm.2023.06.001] [Medline: 37353396]
52. Han A, Gingrich J, Yaskina M, Rankin K, San Martin-Feeney D, Mackie AS. Just TRAC it! Transitioning responsibly to adult care using smart phone technology. J Adolesc Health 2023;73(3):561-566. [doi: 10.1016/j.jadohealth.2023.04.023] [Medline: 37306646]
53. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 2000;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]
54. Flannery M. Self-determination theory: intrinsic motivation and behavioral change. Oncol Nurs Forum 2017;44(2):155-156. [doi: 10.1188/17.ONF.155-156] [Medline: 28222078]
55. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
56. Meleis AI, Sawyer LM, Im EO, Hilfinger Messias DK, Schumacher K. Experiencing transitions: an emerging middle-range theory. ANS Adv Nurs Sci 2000;23(1):12-28. [doi: 10.1097/00012272-200009000-00006] [Medline: 10970036]
57. Obro LF, Heiselberg K, Krogh PG, Handberg C, Ammentorp J, Pihl GT, et al. Combining mHealth and health-coaching for improving self-management in chronic care. a scoping review. Patient Educ Couns 2021;104(4):680-688 [FREE Full text] [doi: 10.1016/j.pec.2020.10.026] [Medline: 33143907]
58. Lopez KN, O'Connor M, King J, Alexander J, Challman M, Lovick DK, et al. Improving transitions of care for young adults with congenital heart disease: mobile app development using formative research. JMIR Form Res 2018;2(2):e16 [FREE Full text] [doi: 10.2196/formative.9963] [Medline: 30574573]
59. Bendixen RM, Fairman AD, Karavolis M, Sullivan C, Parmanto B. A user-centered approach: understanding client and caregiver needs and preferences in the development of mHealth apps for self-management. JMIR mHealth uHealth 2017;5(9):e141 [FREE Full text] [doi: 10.2196/mhealth.7136] [Medline: 28951378]
60. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018;4:2055207618770325 [FREE Full text] [doi: 10.1177/2055207618770325] [Medline: 29942629]
61. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
62. Majeed-Ariss R, Baildam E, Campbell M, Chieng A, Fallon D, Hall A, et al. Apps and adolescents: a systematic review of adolescents' use of mobile phone and tablet apps that support personal management of their chronic or long-term physical conditions. J Med Internet Res 2015;17(12):e287 [FREE Full text] [doi: 10.2196/jmir.5043] [Medline: 26701961]
63. Mosnaim G, Li H, Martin M, Richardson D, Belice PJ, Avery E, et al. A tailored mobile health intervention to improve adherence and asthma control in minority adolescents. J Allergy Clin Immunol Pract 2015;3(2):288-290.e1 [FREE Full text] [doi: 10.1016/j.jaip.2014.10.011] [Medline: 25609351]
64. Badawy SM, Thompson AA, Kuhns LM. Medication adherence and technology-based interventions for adolescents with chronic health conditions: a few key considerations. JMIR mHealth uHealth 2017;5(12):e202 [FREE Full text] [doi: 10.2196/mhealth.8310] [Medline: 29273573]
65. Fedele DA, Cushing CC, Fritz A, Amaro CM, Ortega A. Mobile health interventions for improving health outcomes in youth: a meta-analysis. JAMA Pediatr 2017;171(5):461-469 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.0042] [Medline: 28319239]

XSL•FO
RenderX

66. Coster S, Li Y, Norman IJ. Cochrane reviews of educational and self-management interventions to guide nursing practice: a review. Int J Nurs Stud 2020;110:103698. [doi: 10.1016/j.ijnurstu.2020.103698] [Medline: 32726709]
67. Otanga H, Semujju B, Mwaniki L, Aungo J. Peer support and social networking interventions in diabetes self-management in Kenya and Uganda: a scoping review. PLoS One 2022;17(9):e0273722 [FREE Full text] [doi: 10.1371/journal.pone.0273722] [Medline: 36155494]
68. Zimmerman C, Garland BH, Enzler CJ, Hergenroeder AC, Wiemann CM. The roles of quality of life and family and peer support in feelings about transition to adult care in adolescents with gastroenterology, renal, and rheumatology diseases. J Pediatr Nurs 2022;62:193-199. [doi: 10.1016/j.pedn.2021.04.032] [Medline: 34116868]
69. Dwyer-Matzky K, Blatt A, Asselin BL, Wood DL. Lack of preparedness for pediatric to adult-oriented health care transition in hospitalized adolescents and young adults. Acad Pediatr 2018;18(1):102-110. [doi: 10.1016/j.acap.2017.07.008] [Medline: 28778828]
70. Marchak JG, Reed-Knight B, Amaral S, Mee L, Blount RL. Providers' assessment of transition readiness among adolescent and young adult kidney transplant recipients. Pediatr Transplant 2015;19(8):849-857. [doi: 10.1111/petr.12615] [Medline: 26508553]
71. Iribarren SJ, Cato K, Falzon L, Stone PW. What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. PLoS One 2017;12(2):e0170581 [FREE Full text] [doi: 10.1371/journal.pone.0170581] [Medline: 28152012]
72. Monitoring and Evaluating Digital Health Interventions: a Practical Guide to Conducting Research and Assessment.: World Health Organization; 2016. URL: https://iris.who.int/bitstream/handle/10665/252183/9789241511766-eng.pdf?sequence=1 [accessed 2024-10-30]
73. Badawy SM, Kuhns LM. Economic evaluation of text-messaging and smartphone-based interventions to improve medication adherence in adolescents with chronic health conditions: a systematic review. JMIR mHealth uHealth 2016;4(4):e121 [FREE Full text] [doi: 10.2196/mhealth.6425] [Medline: 27780795]

Abbreviations

ADAPT: Adolescent Adherence Patient Tool
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Review
SRQR: standards for reporting qualitative research
STEP: Sickle Cell Transition E-Learning Program

Edited by N Cahill; submitted 19.01.24; peer-reviewed by S Butler, L Crosby, A Fellas, R Cruz Martínez; comments to author 12.03.24; revised version received 22.03.24; accepted 21.10.24; published 26.11.24.

Please cite as:
Li Z, Lu F, Wu J, Bao R, Rao Y, Yang Y, Wang H
Usability and Effectiveness of eHealth and mHealth Interventions That Support Self-Management and Health Care Transition in Adolescents and Young Adults With Chronic Disease: Systematic Review
J Med Internet Res 2024;26:e56556
URL: https://www.jmir.org/2024/1/e56556
doi:10.2196/56556
PMID:39589770

©ZhiRu Li, FangYan Lu, JingYun Wu, RuiJie Bao, YuXin Rao, Yun Yang, Huafen Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Personalized, Transdiagnostic Smartphone App (Mello) Targeting Repetitive Negative Thinking for Depression and Anxiety: Qualitative Analysis of Young People's Experience

Lee Valentine[1,2], BA, MSW, PhD; Chelsea Arnold[2], BSci (Hons), PhD; Jennifer Nicholas[1,2], BA, BSc (Hons), PhD; Emily Castagnini[1,2], BS, MPH; Jessi Malouf[2], BDes; Mario Alvarez-Jimenez[1,2*], DClinPsych, MRes, PhD; Imogen H Bell[1,2*], BSci (Hons), PhD

[1]Orygen, Melbourne, Australia
[2]Orygen, Parkville, Australia
[*]these authors contributed equally

Corresponding Author:
Lee Valentine, BA, MSW, PhD
Orygen
35 Poplar Rd
Melbourne, 3052
Australia
Phone: 61 9966 9100
Email: lee.valentine@orygen.org.au

Abstract

Background: The increasing rates of mental health challenges among young people highlight an urgent need for accessible and effective treatment. However, current mental health systems face unprecedented demand, leaving most young people globally with unmet mental health needs. Smartphones present a promising solution to this issue by offering in-the-moment support through innovative just-in-time adaptive interventions, which provide support based on real-time data.

Objective: This study explores young people's experiences with Mello, a just-in-time adaptive intervention that focuses on the transdiagnostic mechanism of repetitive negative thinking (RNT), a significant factor contributing to youth depression and anxiety.

Methods: Semistructured qualitative interviews were conducted with 15 participants aged 16 to 25 years, all of whom had previously participated in a pilot randomized controlled trial of Mello. Of the 15 participants, 9 (60%) identified as women, 4 (27%) as men (including 1 transgender man), and 2 (13%) as nonbinary. Interviews focused on participants' experiences with the Mello app, factors influencing engagement, perceived benefits and limitations, and suggestions for future improvements. Thematic analysis was used to analyze the data.

Results: The analysis identified three superordinate themes: Mello as a tool for intentional reflection; doing therapy your own way; and barriers to engagement during low mood, anxiety, and RNT. Theme 1 explored young people's experiences of how the app facilitated active management of negative thoughts and supported the development of reflective habits, contrasting with their typical strategies of avoidance or resistance. Theme 2 highlighted the value of the app's self-guided nature, with a particular focus on its flexibility and accessibility, particularly when compared to traditional face-to-face therapy. Finally, theme 3 addressed barriers to engagement, particularly during emotionally difficult times, with participants reporting feeling "stuck" in their negative thoughts. To mitigate these challenges, participants suggested incorporating gamification elements, such as progress-tracking visuals, to enhance motivation and increase engagement with the app.

Conclusions: Our findings underscored the value of Mello in promoting intentional engagement and reflection with RNT, consistent with prior research that emphasizes the effectiveness of tailored interventions. Although some users valued the self-guided nature of the application, others encountered difficulties with motivation. Future research should explore strategies to enhance engagement for young people with low mood and motivation, such as co-design methodologies, advanced personalization features, and gamification techniques.

(J Med Internet Res 2024;26:e63732) doi:10.2196/63732

XSL•FO
RenderX

KEYWORDS

repetitive negative thinking; rumination; anxiety; depression; mobile app; just-in-time adaptive interventions; youth mental health; adolescent; mobile phone

Introduction

Background

Young people are facing a crisis of mental ill-health, with estimates suggesting that 1 in 5 have a mental health disorder [1], making them the most disproportionately affected age group for mental health burden [2], with upward of 50% affected by the age of 25 years [3]. The magnitude of this problem has increased over the course of the global pandemic [4], and rates of mental ill-health have steadily escalated over the last decade, placing unprecedented demands on services that exceed available resources [3,5]. For example, an Australian survey of young people found 44.1% of the general population met the criteria for depression and 48.5% for anxiety [6]. Critically, estimates suggest that only 50% of young people with mental ill-health seek help [7]. Among those who do seek care, many encounter long waiting periods [8] and disengage prematurely, resulting in only a small proportion of people who experience long-term recovery [9,10]. Due to these barriers, as many as 55% of young people worldwide are experiencing unmet needs for mental health care [11]. Smartphone technologies present an important opportunity to address the growing burden of mental ill-health among young people by providing real-time support without reliance on traditional face-to-face services [12,13].

High rates of smartphone ownership, coupled with a strong interest in using digital devices for mental health support among young people and clinicians, provide important opportunities to overcome barriers to mental health care for youth [14]. Smartphones possess a critical capability for the self-management of mental ill-health, enabling the detection and response to daily life experiences through personalized interventions [15]. “Just-in-time adaptive interventions” (JITAIs) represent a novel approach wherein smartphones or other mobile devices are leveraged to tailor momentary intervention strategies based on an individual’s specific context and needs [16]. While research evidence is still developing, a meta-analysis of 33 studies spanning physical and mental health conditions found evidence supporting the effectiveness of JITAIs compared to controls [17].

A unique opportunity to improve treatments for depression and anxiety in young people may involve the use of JITAIs to deliver momentary interventions targeting transdiagnostic mechanisms in real-time, real-world scenarios. Transdiagnostic mechanisms refer to underlying psychological or behavioral processes that contribute to multiple mental health conditions irrespective of their specific diagnosis [18]. These mechanisms potentially offer a more personalized and precise intervention target, applicable across a diverse array of mental health disorders, thereby enhancing their relevance and efficacy [19]. JITAIs vary in their complexity, from basic static algorithms that select a limited number of interventions to more sophisticated systems that are contextually responsive and capable of learning from user data and individual behavior. However, many JITAIs currently in use rely on static algorithms [20]. Repetitive negative thinking (RNT; [21]) is a putative transdiagnostic mechanism with considerable evidence implicating a causal role in driving depression and anxiety symptoms, making it a highly promising intervention target [22]. RNT involves frequently occurring negative thoughts that are passive or hard to control, such as worry (concern for future events) and rumination (concern for past or current events or depressive symptoms; [21]). High levels of RNT are commonly observed among clinical populations of youth with clinical depression and anxiety and are highly correlated with the severity of depression and anxiety symptoms [23]. Inducing RNT in experimental studies results in worsening mood states [24], and heightened RNT predicts the later onset of emotional disorders in youth [25,26]. Meta-analyses have further shown that improvements in RNT are associated with reductions in depression and anxiety following psychological treatment in adults [27,28] and youth [14]. Therefore, targeting a reduction in RNT in treatment appears to be an important component of psychological treatment.

Various psychological treatment approaches, such as rumination-focused cognitive behavioral therapy, metacognitive therapy, and attention training, have been found to improve RNT outcomes [5]. However, traditional treatment delivery methods involving infrequent face-to-face sessions with a clinician suffer not only limitations in accessibility but are also poorly timed to target the process of RNT as it occurs in daily life. RNT is a dynamic process that is influenced by external and internal events over time [29]. Therefore, intervention strategies should be appropriately timed and sensitive to contextual variables that might influence how this process unfolds over time [30]. JITAIs are potentially well suited to targeting RNT in real time by detecting and delivering highly personalized and context-sensitive interventions designed to disrupt the process as it is happening. To test this theoretical intervention model, a JITAI called Mello was developed.

Mello is a self-guided smartphone intervention designed to reduce RNT (referred to as “stuck thinking” within the app) among young people experiencing depression and anxiety [31]. By using momentary assessments of stuck thinking intensity, mood states, location, and activity, users receive tailored recommendations for a microintervention strategy aimed at disrupting RNT in real time. In a pilot randomized controlled trial (RCT) [31] involving young people who used Mello for 6 weeks compared to a nonactive control group (n=55), the intervention was found to be acceptable, feasible, and potentially effective at reducing RNT, leading to improvements in depression and anxiety with moderate to large effect sizes. Furthermore, 96% of young people would recommend the app to others with similar experiences, with high satisfaction ratings across all domains. Young people also showed sustained use of Mello over time, with average use on 52% of total days and 59% of young people engaging with the app for the full 6-week

XSL•FO
RenderX

period. These results demonstrated the highly promising potential of the Mello app and the broader theoretical concept of harnessing JITAs to target transdiagnostic mechanisms using personalized, context-sensitive interventions within real-time, real-world environments.

Objectives

Building on the results of the RCT evaluating the feasibility and preliminary treatment effects of the Mello app [31], this study aimed to qualitatively explore young people's experiences with Mello. This qualitative perspective can provide rich insights into how young people engage with and make meaning of their use of Mello, perceived benefits and limitations, and suggestions for refining and enhancing future iterations of the app. A better understanding of young people's experiences with Mello not only informs potential benefits and areas for improvement but also contributes to the broader understanding of designing mobile real-time digital mental health interventions to enhance outcomes for young people experiencing mental health challenges.

Methods

Setting

This qualitative study was conducted as part of a broader research project examining Mello—a transdiagnostic smartphone app targeting RNT for young people experiencing anxiety or depression. The feasibility, acceptability, and preliminary clinical outcomes and mechanisms of Mello were examined in a pilot RCT [31]. During the pilot, participants were randomly assigned to either Mello or a nonactive control group for a 6-week period. This qualitative study was conducted with a subset of participants of the Mello intervention group willing to be interviewed following the completion of the 6-week intervention period.

Mello's use of the JITAI model involved identifying and intervening in moments of need by using real-time assessments to tailor momentary intervention strategies aimed at disrupting RNT (Figure 1). In the pilot RCT, participants allocated to the Mello group received randomized notifications 3 times a day prompting them to complete "check-ins": a brief 4-item questionnaire completed via the chat interface assessing the level of RNT, mood, current activities, and location. On the basis of the response provided during the check-ins, the app recommended 1 of 12 cognitive behavioral activities (Multimedia Appendix 1). Each activity took approximately 2 to 12 minutes to complete and was delivered through guided text or audio instructions using evidence-based strategies for targeting RNT. The selection of therapeutic activities was informed by a meta-analysis of psychological treatment approaches for RNT [5]. These activities included exercises such as mindfulness, problem-solving, cognitive defusion, worry time, thought challenging, gratitude, and self-compassion [31].

Figure 1. Mello app features.

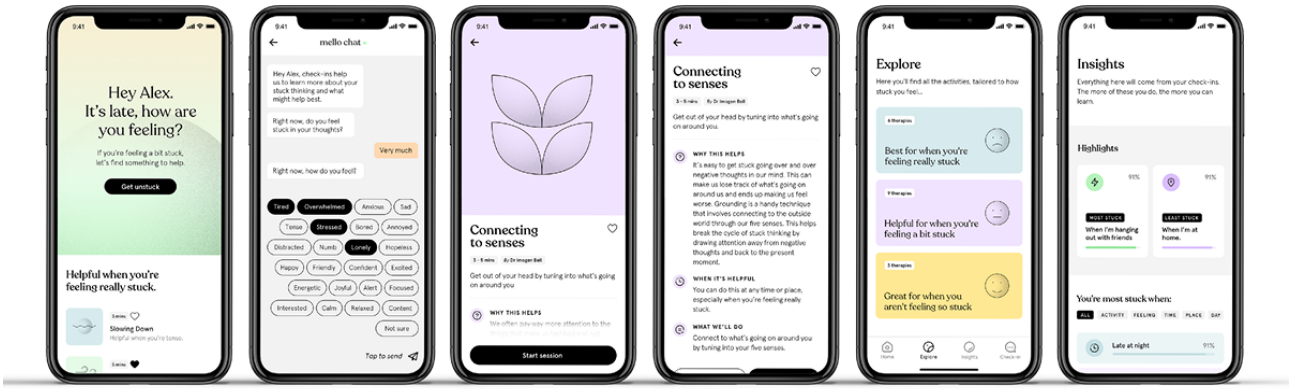

Sampling and Recruitment

Participants in the pilot RCT were aged 16 to 25 years, experiencing clinical levels of depression and anxiety (≥10 on the Patient Health Questionnaire 8 [32] or Generalized Anxiety Disorder Scale 7 [33]), and elevated RNT levels (≥37) on the Perseverative Thinking Questionnaire [21]. Additional inclusion criteria involved the ability to provide informed consent, sufficient command of the English language, and ownership of a smartphone capable of running the Mello app (Android [Google LLC] or iPhone [Apple Inc]). Exclusion criteria comprised current psychiatric inpatient status or receiving treatment from a mental health crisis management team. There were no additional exclusion or inclusion criteria, beyond being randomized to the Mello arm of the RCT, for participating in the present qualitative study.

All participants assigned to the Mello intervention arm of the pilot RCT (N=29) consented to follow-up after the 6-week intervention period and were invited to participate in a qualitative interview exploring their experience with Mello. As such, at the conclusion of the intervention period, emails were sent to all participants in the Mello intervention arm, inviting them to participate in a qualitative interview. Of the 29 young people invited, 15 (52%) responded and agreed to participate. The gender distribution, mean age, and country of birth within this subgroup were consistent with the overall sample from the pilot RCT.

There was variability in app engagement among the sample: with active use ranging from 6 to 38 days (out of a possible 42) during the 6-week intervention period (mean 21.9, SD 11.1) The engagement levels among this sample were reflective of the broader use patterns observed in the original trial, where

XSL•FO
RenderX

participants used the app on 49.2% of days and completed an average of 29.9 check-ins [31]. This variation in engagement is valuable in qualitative research, as it allows a more nuanced understanding of user experiences across the spectrum of engagement, providing insights into both facilitators and barriers to sustained app use across different engagement levels. By including participants with varying degrees of interaction with Mello, this study aimed to identify the factors that either support or hinder ongoing engagement with the app, offering a more nuanced perspective on its efficacy and usability.

Procedure

The qualitative interviews were conducted between February 8, 2022, and May 2, 2022. On average, participants were interviewed 1.9 weeks (range 0-4.7 wk) following the final date they had access to the Mello app. All interviews were conducted via Zoom (Zoom Video Communications, Inc) by authors CA and EC. CA, a female clinical psychologist with a PhD in clinical psychology, specializes in digital mental health and has extensive experience working clinically with people experiencing mental health challenges. She has led and contributed to multiple qualitative studies from study design to data collection and analysis. EC, also a female researcher, holds a master's degree in public health and has substantial demonstrated experience in qualitative research, particularly in co-designing and interviewing young people about their experiences with digital mental health products. Both CA and EC had prior relationships with participants, as they were responsible for conducting pre and post-interviews with participants as part of the Mello RCT.

The interview schedule was designed to better understand young people's experiences and perspectives of the Mello app. Multimedia Appendix 2 provides the full interview schedule. All interviews were conducted and audio recorded via Zoom to facilitate transcription and analysis. The recordings were transcribed verbatim by a third-party service. Both the audio files and transcripts were stored securely on a password-protected server. Identifying information, such as participant consent forms, was stored separately to ensure confidentiality. Access to the recorded interviews and transcripts was restricted to the immediate research team responsible for conducting the analysis.

Data Analysis

Thematic analysis, guided by the framework proposed by Braun and Clarke [34], was used for data analysis. To begin, a single transcript was independently coded by authors JN, LV, CA, and IB. Following this, the 4 authors participated in collaborative discussions to review and compare their initially identified codes and coding process. This iterative process allowed for the refinement of an initial coding dictionary, resulting in the development of a robust coding framework.

To further ensure consistency, authors LV and CA independently coded and discussed 2 additional identical transcripts. This additional coding was compared to the established framework to further solidify alignment and consistency in the coding process. With confidence in the coding alignment, author LV proceeded to independently code the remaining 12 interviews. The coded data were then systematically grouped based on similarities, and through an iterative process among authors, overarching themes and subordinate themes were developed.

The systematic thematic analysis approach used in this study established a strong framework for comprehensively capturing and understanding the nuanced experiences and perspectives of participants using Mello.

Ethical Considerations

The study was approved by the University of Melbourne Human Research ethics committee (project 2021-22524-23553-5). Participants provided written informed consent through the REDCap (Research Electronic Data Capture; Vanderbilt University) web application. Before providing consent, participants were given a thorough explanation of the study in simple language, including the study's aims, procedures, and any potential risks or benefits. Participants were also reminded that they could withdraw from the study at any time, skip questions, or take a break from the interview without any negative consequences. Participants were given a copy of the participant information and consent form and received Aus $45 (US $29.5) as reimbursement for their participation in the research interview. All data were deidentified.

Reflexivity

All research questions were intentionally designed to be open-ended to prevent leading participants in predetermined or unconsciously biased directions. In the interview consent process, young people were explicitly assured that their honest opinions were valued, and the research team welcomed negative and constructive feedback as much as positive input.

Throughout the analysis phase, the research team held regular meetings to critically review codes and thematic patterns. These sessions were instrumental in considering how the research teams' personal experiences, viewpoints, beliefs, and biases may be shaping the interpretation of results and influencing data analysis. The team acknowledges the potential impact of their own beliefs and experiences on the interpretation process.

Ultimately, participants' experiences were deeply considered, with the research team composing rich and detailed paragraphs linked to illustrative quotes. This approach aims to ensure a transparent and traceable lineage between the reported findings and the actual words and experiences shared by the young people themselves.

Results

Overview

The study included 15 participants, with ages ranging from 16 to 25 (mean 20.9, SD 2.9) years. Overall, 60% (9/15) of participants identified as women, 27% (4/15) as men (including 1 transgender man), and 13% (2/15) as nonbinary. Furthermore, 73% (11/15) of the participants were born in Australia, while 27% (4/15) were born overseas. In addition, 53% (9/15) of the sample were receiving mental health treatment while using Mello. The average interview duration was 31.9 minutes, with interview times ranging from 17.0 to 67.2 minutes. Additional demographic details are presented in Table 1.

XSL•FO
RenderX

The analysis identified three superordinate themes reflecting distinct aspects of participants' experiences with the Mello app: (1) Mello as a tool for intentional reflection; (2) doing therapy your own way; and (3) barriers to engagement during moments of low mood, anxiety, and RNT. These themes, along with representative quotes from participants, are presented in the *Results* section.

Superordinate theme 1 highlighted how Mello served as a tool for intentional engagement with RNT. Participants described how the app facilitated active self-management of negative thoughts, contrasting with their usual strategies of either resisting support or avoiding difficult emotions. Subordinate themes include habit formation, where automated check-ins helped participants build consistent reflective habits, and enhanced self-awareness, allowing them to recognize emotional patterns through personalized reflection.

Superordinate theme 2 focused on the value of Mello's self-guided nature, enabling participants to manage their mental health autonomously. Many participants found that the app's flexibility and accessibility alleviated the anxiety that can be associated with traditional face-to-face therapy, and they appreciated its availability during late hours or when other support systems were inaccessible.

Superordinate theme 3 explored the barriers to engagement, particularly during moments of low mood, anxiety, and RNT. Some participants struggled to engage with the app during emotionally difficult times, feeling "stuck" in their negative thoughts. To address these challenges, participants suggested incorporating gamification elements, such as progress-tracking visuals, to enhance motivation and make the app more engaging and rewarding. Textbox 1 presents a visual representation of these themes.

Table 1. Demographic characteristics of participants.

Participant ID	Age (years)	Gender	Country of birth	Currently receiving treatment?	Total use days on Mello, n
1	20	Woman	Australia	Yes	32
2	20	Man	Australia	Yes	27
3	25	Man, transgender	Australia	No	28
4	20	Man	Australia	No	9
5	16	Woman	Australia	Yes	8
6	24	Woman	Australia	No	14
7	17	Nonbinary	Australia	Yes	33
8	20	Nonbinary	Australia	Yes	17
9	24	Woman	China	Yes	12
10	21	Woman	Bangladesh	No	32
11	24	Woman	Australia	No	6
12	25	Woman	Australia	Yes	38
13	17	Man	Australia	Yes	37
14	20	Woman	India	No	19
15	20	Woman	India	Yes	16

Textbox 1. Superordinate and subordinate themes.

Mello as a tool for intentional engagement with repetitive negative thinking (RNT)

- Habit formation
- Enhanced self-awareness through personalized reflection and pattern recognition

Doing therapy your own way

- The value of self-guided therapy
- Accessibility and dependability

Barriers to engagement during moments of low mood, anxiety, and RNT

- Avoiding difficult feelings and getting "stuck" in RNT
- Leveraging gamification to enhance self-guided motivation

XSL•FO
RenderX

Superordinate Theme 1: Mello as a Tool for Intentional Engagement With RNT

Overview

Participants commonly characterized Mello as a tool that facilitated intentional engagement with their thoughts and feelings. A participant noted the following:

> *It was really nice to just take a moment and figure it out and maybe work through a little bit of anxiety or something that I wasn't quite recognising in the moment.* [Participant 8]

Several participants identified that this deliberate approach to *thinking*, sharply contrasted with their typical responses to rumination. Young people generally described these established responses in two distinct ways: resisting support and persisting in a state of rumination, or attempting to avoid engaging with their thoughts altogether. One participant described the impulse to resist support, stating:

> *I feel an urge to remain in that state and to continue ruminating or feeling bad and therefore a resistance to access any kind of assistance that could help me.* [Participant 2]

In contrast, another participant shared the belief that avoiding reflection on RNT might make distressing thoughts disappear, although this often had the opposite effect:

> *Sometimes you feel like you should just avoid it entirely. Maybe if I just avoid reflecting on it, then it'll go away or something, and it obviously doesn't, it sometimes makes it worse.* [Participant 10]

In contrast to these typical responses to rumination, young people described using Mello as a means for actively managing negative thoughts through deliberate reflection. This approach helped them to diverge from habitual responses, like avoidance, and empowered them to actively self-manage negative thoughts in the moment. One participant explained:

> *Once the negative thoughts come on, I would immediately go to the app, and then that way I would kind of help myself sort it out through that.* [Participant 3]

Beyond seeking relief during "stuck" moments, some young people expressed a proactive desire to practice skills during moments when they were not feeling "stuck," using their clearer frame of mind, aiming to better equip themselves for difficult moments in the future. One participant shared:

> *And one of the options [is] that people use it or practice the activities that are useful when you are stuck specifically when they are not so that they gradually feel more comfortable trying those activities when being least stuck to most stuck as they feel more and more stuck. Now they feel comfortable doing those activities. As familiarity increases, the range they feel comfortable doing them increases in moments of stuckness.* [Participant 2]

This proactive approach demonstrates young people's creative use of Mello to practice skills and build resilience during moments of calm, a valuable way for young people to prepare for and manage future challenges with RNT more effectively.

Subordinate Theme 1.1: Habit Formation

Participants commented frequently on Mello's automated check-ins, emphasizing their role in prompting reflection on thoughts, feelings, and surroundings. For some, these *check-ins* felt "difficult," particularly when: (1) the young person was experiencing low mood or motivation and was concerned that engaging with Mello would exacerbate these feelings, or (2) the young person felt that their "lack of energy" from low mood, stress, or anxiety was too limited to engage with the process.

For others, the check-ins served as a functional "trigger," or practical reminder, helping to establish a daily "habit" of self-reflection. Participant 9 framed these check-ins as an instructional cue to "review feelings." Several participants noted that regularly using Mello helped them build a greater sense of self-awareness of their thoughts and emotions. A participant shared:

> *It was really helpful in building my awareness of how I felt at different times during the day when I did do it consistently and tying that to different activities, or people I was around.* [Participant 2]

For some young people, the check-ins were experienced as thoughtful gestures, conveying a sense that someone was thinking of them and cared about them. A participant expressed the following:

> *Every time I see the notice from Mello...[it] gives me a kind of feeling, like friends, they really care about you.[Participant 9]*

This reminder encouraged some young people to remember to extend that sense of care toward themselves. One participant described the experience as follows:

> *So, it's actually quite wholesome to be like, "Hey, how you going?" It's just nice to have that little reminder. It's like, "Hey." And it reminds you, "Oh, I need to like think about myself as well."[Participant 3]*

These insights highlight the meaningful impact of Mello's check-ins, serving not only as a reminder prompting self-reflection but also as a source of perceived care and support. This contributing to a positive and "wholesome" experience akin to the thoughtful gesture of friends checking in on each other.

However, while some identified the check-ins as helpful stepping stones toward building a positive habit of reflection, others acknowledged the difficulty they had in forming such habits. To better support this process, participants suggested that Mello could make the check-in process more compelling and "attractive" thereby increasing the likelihood that a young person would engage and build the habit over time. Participant 2 recommended leveraging the "habit loop of cue and craving routine reward," suggesting that making the cue obvious, the craving appealing, the routine easy, and the reward more satisfying could increase the likelihood of engagement. Furthermore, Participant 2 proposed that notifications could

XSL•FO
RenderX

prompt the user to "imagine" the positive feelings they would experience in the future after checking in or completing a therapy activity, effectively linking the cue to the satisfaction of the craving.

Despite young people's general support for the notifications and regular check-ins, participants emphasized the need for customizable notification schedules to align with various life commitments, such as work, social engagements, and family obligations. This customization also enables users to safeguard their own personal free time without interruption. These examples highlight the importance of a more personalized experience with Mello, ensuring a more seamless integration into the intricacies of daily life.

Subordinate Theme 1.2: Enhanced Self-Awareness Through Personalized Reflection and Pattern Recognition

While some participants found checking in with Mello to be a valuable act of self-reflection of its own accord, others described it as a tool for gaining a deeper understanding of the connection between their "feelings and patterns of behaviour," thereby enhancing their sense of self-awareness. By creating moments for self-reflection, Mello supported young people to recognize emotions, such as anxiety or anger, that might have otherwise gone unnoticed and potentially intensified. One likened this experience to mindfulness:

> *It almost felt like mindfulness, acknowledging that I'm in my room, alone, and feeling angry. I found that really positive.* [Participant 2]

Some users reported that regularly using the Mello app to track their thoughts, emotions, and surroundings helped them to identify recurring emotional and behavioral patterns. This data-driven approach motivated them to stay engaged with the app throughout the intervention period. Recognizing personal patterns, particularly in relation to daily activities and specific times of day, was considered valuable. One participant highlighted the importance of immediate access to personal data via Mello:

> *Being able to notice my own patterns, especially when it comes to the activities I'm doing and times of day, that was very important to me. The idea of having the statistics right there was motivating.* [Participant 7]

Participants commonly requested the ability to personalize the information recorded, including details about their environment, to use the app more effectively. They emphasized the need for this feature in order to better make informed choices about their mental health, based on accurately collected data, stressing the importance of clear and informative data that visually illustrated progress and discernible patterns over time.

Superordinate Theme 2: Doing Therapy Your Own Way

Overview

A subgroup of young people consistently underscored two key aspects of using Mello: it's self-guided nature and its accessibility and dependability. For some, Mello's self-guided approach alleviated the anxiety typically associated with traditional face-to-face therapy. For others, the apps consistent support and constant availability was empowering, contributing positively to their mental health journey.

Subordinate Theme 2.1: The Value of Self-Guided Therapy

Some young people described experiencing the self-guided nature of Mello, which did not include a real-time face-to-face therapist, to be a feature that alleviated the anxiety typically associated with traditional face-to-face therapy dynamic, while also enhancing their sense of agency and control over their mental health treatment:

> *I think it was an interesting experience for me because I've never used an app before to manage my anxiety. And I think it was good to have that new technology experience that I hadn't had before, because I'm so used to talking to people in person. And sometimes that has increased my anxiety and not made it even better so having an app was a nice change and a good experience for me.* [Participant 7]

This quote reflects the perspective of a subgroup of participants who saw Mello as an intervention providing valuable control and autonomy over their mental health experience. Participant 11 further described how they could lose sight of their own priorities in face-to-face therapy: "Sometimes other people's input makes those anxious thoughts spiral, because I'm trying to fit someone else's ideas and someone else's intentions into my life and then I'm not prioritizing my own." In this case, Mello offered a therapeutic avenue for identifying and addressing personal mental health priorities without inadvertently conforming to others' expectations.

Furthermore, participant 11 appreciated the ability to work at their own pace, free from the perceived pressure of being "on someone else's time." They explained:

> *I liked that I could do it at my own pace. What I found before in doing psychology sessions in person is that I feel rushed and I feel like they're expecting, even though it's not the truth, but they're expecting me to say something in a certain way. So having an app was a good change.* [Participant 11]

Participants described valuing the self-guided nature of Mello as it provided a nonjudgmental space to express their feelings. Participant 11 reflected on the positive experience of "venting of the feelings" by recording notes via microphone on the Mello app:

> *Yeah. Yeah. It kind of just made it feel like I have that freedom to say whatever I want and no one's going to judge me for it or force me to act on something I don't want to act on. Sometimes it's just getting that information out there and that's all you need to do rather than have to put together a whole strategy or plan out your entire week or whatever. It's just getting it out there in that moment and then it's done.* [Participant 11]

XSL•FO
RenderX

For some, Mello was experienced as a positive alternative to the traditional face-to-face therapeutic experience, allowing them to communicate their thoughts and emotions without the anxiety brought on by in-person interactions.

> *It was convenient in an app for me and that I could put in my answers and learn more about the different techniques to manage my anxiety. And yeah, it was just good that I had something respond to me and not feeling embarrassed or ashamed or just awkward.* [Participant 11]

The self-guided and autonomous nature of Mello empowered some participants to engage with the intervention components that most resonated with them. This flexibility enabled them to explore therapeutic content without being confined to a specific approach. One participant describes appreciating this approach:

> *I think the variety, there was lots of good variety in the different activities. And again, I really liked how you could both read it or listen to it. I really liked there was that option.* [Participant 1]

While preferences for therapeutic activities varied among participants, some highlighted the app's capacity for positive distraction. For example, participant 9 shared, "I pay attention to the using, so the sadness feeling decreased while I use it." Gratitude practices emerged as a gentle and affirming way for young people to engage with therapy, with one participant noting the following:

> *For example, the gratitude exercise where I had to put in effort into thinking about what was something positive in my day. Even if it was small, I realised that not everything in my day was crappy, and that's great. Maybe it's a small thing now, but tomorrow it's going to be great, a bigger thing.* [Participant 10]

Overall, these diverse experiences highlighted how Mello's self-guided approach, offering a variety of activities, allowed for individual choice and meaning-making.

Subordinate Theme 2.2: Accessibility and Dependability

Young people emphasized the benefit of having access to reliable, "24/7" mental health support through Mello. They particularly highlighted the app's "availability," and the perceived "trustworthiness" of the content. Participants commonly described turning to Mello when friends and family were unavailable, using it before bed, or in the later hours of the night and early morning when it was challenging to find support elsewhere. For example, participant 3 explained:

> *You can't always go to your family members, and I mean you can, but I felt like I couldn't, because I didn't want to bring up anything bad for them. So, that's where Mello was quite helpful, where I could just rant.* [Participant 3]

This high level of accessibility ensured that young people could rely on Mello for assistance whenever needed, solidifying the app's value as a dependable self-guided resource.

Moreover, Mello's continuous availability and support empowered young people to maintain autonomy in self-managing challenging experiences. One participant shared the following:

> *I think it just ultimately comes down to having an accessible resource so I don't have to keep searching for anxiety relieving gifs online, it's right there in the app for me, and I can just reach out for it. And yeah, I just really liked that.* [Participant 10]

Superordinate Theme 3: Barriers to Engagement During Moments of Low Mood, Anxiety, and RNT

Overview

Some young people described experiences that hindered their use of Mello during moments of RNT. These experiences revolved around the perception that engaging with the app during moments of stuck thinking introduced an additional layer of difficulty to an already emotionally challenging time.

For some, feelings of distress or depression left them with limited energy, motivation, or desire to engage with Mello. Participant 4 noted, "But I think that just came down to general procrastination, forgetfulness, and fatigue because I'd feel tired a lot."

Others hesitated to engage with their RNT altogether, fearing that doing so would worsen their emotions and exacerbate their feelings of distress. In this context, it was perceived that using the app had the potential to intensify their negative thinking patterns instead of providing relief, thereby contributing to the very feelings they were seeking to alleviate. Despite the potential for a positive outcome through engagement with the app and therapeutic guidance, this process felt too burdensome for some young people to endure.

Subordinate Theme 3.1: Avoiding Difficult Feelings and Getting "Stuck" in RNT

Several young people expressed a desire to avoid intentionally engaging with their difficult feelings and thoughts, believing that ignoring them would cause the discomfort to dissipate and "go away" of its own accord.

For some, purposefully engaging with their negative thoughts contradicted their established coping mechanisms, which typically involved attempting to avoid them. From this perspective, paying greater attention to their negative thoughts risked amplifying the negative feelings or prolonging the duration of their experience, both of which were outcomes they sought to avoid. One participant explained the following:

> *I see a notification telling me it's time to check in with Mello. And I would dismiss that because it seems difficult, or I don't want to confront how I'm feeling, because that could be painful if I'm already avoiding emotional and mental pain.* [Participant 2]

Thus, when prompted by Mello to check in, participant 2 dismissed the notification because engaging with their feelings seemed daunting and potentially distressing. This experience underscores a reluctance to confront emotional and mental pain, preferring instead to evade such experiences.

XSL•FO
RenderX

The concept of feeling "stuck" was a reoccurring theme among young people when describing their experiences of RNT. This term was also commonly used by young people when describing the impediments they experienced when trying to engage with Mello during moments of worry or anxiety. Many participants expressed that during moments of rumination, feeling "stuck" extended beyond their thought patterns and interfered with their behavior. This created a catch-22 situation, wherein the more entrenched in their thinking they became, the more challenging it became to actively participate in the intervention. One participant articulated this challenge, stating the following:

> *For me, if I am depression and the situation is really serious, I didn't want to do anything else and I didn't care about the app. I didn't care anybody say anything. I just cannot get away that kind of feeling.* [Participant 9]

Consequently, struggling to engage with Mello while caught in cycles of RNT posed a formidable challenge to some young people. In general, young people reported that experiences of depression, distress, stress, and various life events, combined with the perceived emotional effort required to engage with Mello during moments of rumination, acted as barriers to use. Participant 2 explains, "There were times when it would pop up on my notifications, but I just didn't have the energy to even express what I was feeling."

Subordinate Theme 3.2: Leveraging Gamification to Enhance Self-Guided Motivation

Among participants, a distinction emerged between those who valued the autonomy inherent in the self-directed nature of the Mello app and those who found the need for self-motivation to be a barrier to sustained engagement. While some participants appreciated the flexibility to engage with the app at their own pace, others expressed difficulty in maintaining regular use without the external prompts or active follow-up typically provided by a service or clinician. This challenge of self-motivation was a common theme, with several participants emphasizing that without consistent guidance or reminders, it would be harder to prioritize the app as part of their daily life.

To address this, some users suggested that incorporating gamification elements, such as points, rewards, or progress-tracking visuals, could enhance their motivation to use the app. One participant highlighted the potential of integrating a growth-based visual representation, similar to an app they had used in high school, where a tree grows as the user studies or engages in tasks. Participant 2 explained the following:

> *I know that can be very scary for some people—they don't want to be, oh, why is this app demanding me to use it? I don't like that. But, I think when I used to be in high school, there was an app where you put in for how long you want to study for, and then it sets the timer and in that timer, a tree grows. And so, if you sort of kind of procrastinate within that period, that tree dies. Obviously, that wouldn't be the case here, but I think if you sort of have a section where you're, the more you use the app, the more the tree grows, it's sort of you get to see a nice visual of you growing your skills and growing in your mental health journey and progress. I think that would've been great to see because it's great to see these insights and all, but I want to see how am I also growing as a person who's using these activities.*

This suggestion underscores the importance of tangible feedback for young people, particularly those who struggle with self-motivation. For these participants, the ability to visually track their progress or achievements over time could enhance engagement by adding a layer of reward and satisfaction to the process of working on their mental health. Such features could not only enhance motivation but also provide users with a sense of accomplishment and progression, making the self-directed nature of the intervention feel more interactive and supportive.

Discussion

Principal Findings

This study aimed to explore young people's experience of using Mello, a transdiagnostic intervention designed to alleviate RNT associated with depression and anxiety. A total of 15 participants took part in the study following a 6-week intervention period.

The analysis identified three key themes from participants' experiences with the Mello app: (1) Mello as a tool for intentional reflection; (2) doing therapy your own way; and (3) barriers to engagement during low mood, anxiety, and RNT. In theme 1, participants described how the app could facilitate the active management of negative thoughts and supported the development of reflective habits, contrasting with their typical strategies of avoidance or resistance. Theme 2 highlighted the value of the app's self-guided nature, with a particular focus on its flexibility and accessibility, particularly when compared to traditional face-to-face therapy. Finally, theme 3 addressed the barriers to engagement, particularly during emotionally difficult times, with participants reporting feeling "stuck" in their negative thoughts. To mitigate these challenges, participants suggested incorporating gamification elements, such as progress-tracking visuals, to enhance motivation and increase engagement with the intervention. Young people's qualitative experiences spoke to the effectiveness of the JITAI framework [16] in fostering intentional engagement and reflection with thoughts and emotions in times of need, as presented in the first overarching theme. For some, this framework facilitated positive behavioral changes by encouraging active engagement with and reflection on their RNT. It offered a constructive alternative to habitual responses such as resisting support or attempting to avoid engaging with their rumination, with the hope of achieving quicker resolution from difficult feelings. Instead, it encouraged intentional engagement in real time, facilitated meaning-making, and, for some, led to a reduction in RNT. These results may provide explanations for the findings of the pilot clinical trial of Mello [31], which demonstrated moderate to large improvements in depression, anxiety, and RNT in Mello users versus the control group. This suggests that providing brief, smartphone-delivered interventions during moments of RNT offers a promising pathway for disrupting key processes in real time in a way that intentionally engages conscious behaviors and counteracts habitual maladaptive responses. This

XSL•FO
RenderX

opportunity is uniquely suited to the JITAI framework and opens opportunities for future research to explore the application of this approach for targeting other transdiagnostic processes with temporal and contextual qualities, for example, emotion regulation, avoidance, and cognitive biases. Future research may benefit from identifying and refining the timing of intervention prompts to optimally disrupt the RNT process, as well as determining which intervention strategy may work best for the individual in any given moment. This can be achieved using artificial intelligence (AI) approaches, such as reinforcement learning [35], which can learn from individual user data to adapt both the timing and type of intervention over time. By understanding the variety of activities and intervention types chosen by the user, these AI systems can better enhance personalized recommendations and potentially improve outcomes. Our data support the efficacy of the JITAI approach in addressing RNT. We aim to advance from a basic JITAI 1.0 model to a more sophisticated intervention. This next-generation JITAI will incorporate a greater variety of activities and use contextual information and inputs, leveraging AI to tailor the timing and delivery of interventions. This enhancement is expected to significantly improve personalization and therapeutic outcomes.

These findings are consistent with meta-analytic review by Wang and Miller [17], which found that JITAIs saw an increased additive effect when tailored to the individuals' behavioral patterns and current needs. Wang and Miller [17] also found that the effectiveness of the JITAIs was further enhanced by the integration of both a human agent and an algorithm to generate personalized feedback. This mirrors the approach implemented in the Mello intervention in which a static algorithm tailored intervention content specific to the needs and behaviors of the participants. These findings provide support for the effectiveness of tailoring interventions to individuals' needs in the moment and underscore the importance of combining human and algorithmic support in digital interventions. While Mello is a self-guided app without human support, the current findings suggest future iterations should emphasize personalization. By incorporating personalized notifications and content recommendations, the app can customize the user experience, creating a more engaging and tailored environment that mimics the experience of human support [36].

While some young people acknowledged the benefits of using Mello during difficult moments of rumination, others had difficulty engaging with the app due to low mood, stress, the challenge of RNT, or lack of active follow-up due to the self-guided nature of the intervention, as detailed in theme 3. Challenges with motivation are a qualitatively well-documented barrier to engagement with mental health interventions, which may be particularly notable during periods of low mood, heightened anxiety, or with self-guided interventions [37-41]. This theme aligns with broader findings in the digital mental health literature that demonstrates maintaining engagement, particularly in real-world contexts, remains a considerable challenge, with many interventions experiencing low rates of sustained use [42-45]. This highlights the importance of JITAI's and future iterations of Mello, which are designed to provide support in the moment and are well-positioned to engage people with personalized reminders and content that is immediately responsive to the thoughts and feelings people are experiencing in the moment.

Advancing digital therapeutic interventions to support mental health likely necessitates further research into the experience of people with low motivation and the development of strategies tailored for this group. One avenue for investigation could involve identifying effective techniques or features within digital interventions specifically designed to engage low- users with low motivation, such as personalized support and encouragement and reminders or motivational messages. Qualitative research by Berry et al [46] with mental health service users emphasized the importance of interventions being perceived as "positive, fun, practical and interactive," underscoring the necessity for prioritizing engaging and user-centered design in digital mental health interventions to more effectively engage people, particularly those with low motivation, in managing their mental health.

Furthermore, co-design in the mental health context and the involvement of end users in the design of interventions have been associated with a generally enhanced effectiveness [47]. Therefore, actively engaging users with low motivation in the design and implementation process holds great promise for gaining valuable insights into their unique experience of low motivation and for effective solutions. Personalization, in particular, tailors the app's features and content to align with the individual preferences and needs of each user, creating a more relevant, autonomous, and engaging experience [36,48]. Gamification introduces game-like elements and incentives to make interacting with the app more compelling and motivating [36]. In a 2024 study, Lieder et al [49] examined the optimization of gamification as a method to facilitate positive behavior change; for instance, in an app aimed at increasing water intake, users who received feedback based on optimized gamification were more likely to enact the desired behavior more frequently compared to control groups. Overall, further research in this area can help identify strategies to motivate individuals more effectively during times of emotional difficulty, ultimately maximizing the potential of digital therapeutic interventions to support young people's mental health.

The self-guided aspect of Mello was recognized for offering a nonjudgmental environment, empowering users to address their mental health needs without the pressure that can be felt in face-to-face therapy sessions. Beyond this, it offered users the flexibility to access support and manage their mental health beyond the constraints of traditional clinical hours, particularly when seeking support from family and friends was not feasible or preferred. This adaptability is further explored in theme 2, where participants described Mello as a self-guided tool that offers a sense of control and autonomy over their therapy experience. It provides a nonjudgmental outlet for expression, offers flexibility in therapeutic approaches, and ensures continuous accessibility, ultimately empowering individuals to independently and effectively manage their mental health challenges. This flexibility allows young people to engage, adjust, and switch therapeutic activities in real time, tailoring their experience to their needs and preferences. Thus, although

XSL•FO
RenderX

research suggests that dropout rates for self-guided interventions can be as high as 80% [50], the experiences of participants in the Mello trial highlight the significant benefits that self-guided mental health interventions can offer. These include promoting autonomy, flexibility, and reducing anxieties related to therapy. Therefore, understanding how to sustain engagement while maintaining the positive impact of self-guided interventions is crucial for the ongoing development of digital mental health interventions for young people.

Strengths and Limitations

While all young people assigned to the Mello intervention group were offered the opportunity to participate, it is plausible that those with lower motivation may have been less inclined to participate, potentially skewing our findings. While our themes do capture insights from participants who expressed having low motivation, the results may not fully encapsulate the experiences of this subgroup. This study was also conducted in the context of using Mello within a clinical trial; therefore, the experience of real-world users may be different.

Although members of the youth advisory committee of Orygen provided feedback on the interview questions, the analysis phase lacked direct involvement from a young person with lived experience of mental health challenges. Moving forward, incorporating such perspectives would enrich our research endeavors. Young people with lived experience may identify nuances within experiences that those without such background might overlook, underscoring the importance of their inclusion in the analysis process.

Conclusions

Through qualitative exploration, our findings identified the value of Mello's JITAI framework in fostering intentional engagement and reflection on RNT during times of need. This framework reportedly facilitated positive behavioral changes, steering young people away from habitual responses such as resisting support and toward active engagement and reflection on their RNT. These findings align with prior research by Wang and Miller [17], emphasizing the value of tailoring interventions to individuals' needs and combining human and algorithmic support in digital interventions.

However, some participants faced challenges engaging with Mello due to low mood, stress, and avoidance related to the experience on RNT, which highlights the importance of addressing motivation barriers and avoidance in JITAI. Looking ahead, further research into strategies tailored for individuals with low motivation is imperative. This could involve exploring techniques such as co-design to identify effective personalized support and encouragement as well as furthering the incorporation of user-centered design principles such as personalization and gamification. Such efforts aim to enhance engagement and address the dropout rates associated with self-guided interventions.

Overall, while Mello's self-guided nature provided a nonjudgmental space and flexibility for users to manage their mental health, sustaining engagement remains a significant challenge. As such, understanding and mitigating these challenges is crucial for maximizing the potential of digital therapeutic interventions to support young people's mental health in the future.

Acknowledgments

The authors would like to acknowledge the young people who dedicated their time to participating in this study and the young people and clinicians who provided input into the design of Mello, including the Moderated Online Social Therapy (MOST) team at Orygen Digital. Finally, the authors would like to acknowledge the significant role of Sally Coldham and James Gallichio and Paul Hoskins and Nick Davis from Everyhow, in the design and development of Mello, as well as Telstra Foundation (particularly Natalie Falzon) for their generous support and guidance.

The development of the Mello app and this research trial were funded by the Telstra Foundation Tech-4-Good Challenge, in support of Everyhow. This work was also supported by a Wellcome Trust Active Ingredients project on Reducing Repetitive Negative Thinking awarded to IB. JN is supported by a National Health and Medical Research Council (NHMRC) investigator grant (APP2009782) and IB is supported by an NHMRC investigator grant (2017876). MA-J is supported by an NHMRC investigator grant (1177235) and a Dame Kate Campbell Fellowship from the University of Melbourne.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LV, CA, and IB conceptualized and designed the study. CA led the project administration. CA and EC conducted the interviews. LV, CA, IB, and JN conducted the analysis. LV wrote the first draft of the manuscript with inputs from IB, JN, CA, and JM. All authors revised the manuscript and approved the final version of the submitted manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Description of therapy activities in Mello.
[DOCX File , 24 KB - jmir_v26i1e63732_app1.docx]

Multimedia Appendix 2
Mello qualitative interview schedule.
[DOCX File , 15 KB - jmir_v26i1e63732_app2.docx]

References

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602 [FREE Full text] [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
2. McGorry PD, Mei C. Early intervention in youth mental health: progress and future directions. Evid Based Ment Health 2018 Nov;21(4):182-184 [FREE Full text] [doi: 10.1136/ebmental-2018-300060] [Medline: 30352884]
3. McGorry PD. The reality of mental health care for young people, and the urgent need for solutions. Med J Aust 2022 Feb 07;216(2):78-79 [FREE Full text] [doi: 10.5694/mja2.51327] [Medline: 34719048]
4. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
5. Bell IH, Marx W, Nguyen K, Grace S, Gleeson J, Alvarez-Jimenez M. The effect of psychological treatment on repetitive negative thinking in youth depression and anxiety: a meta-analysis and meta-regression. Psychol Med 2023 Jan;53(1):6-16 [FREE Full text] [doi: 10.1017/S0033291722003373] [Medline: 36373473]
6. Bell I, Nicholas J, Broomhall A, Bailey E, Bendall S, Boland A, et al. The impact of COVID-19 on youth mental health: a mixed methods survey. Psychiatry Res 2023 Mar;321:115082 [FREE Full text] [doi: 10.1016/j.psychres.2023.115082] [Medline: 36738592]
7. National study of mental health and wellbeing. Australian Bureau of Statistics. 2023 Oct 5. URL: https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release [accessed 2024-11-13]
8. MacDonald K, Ferrari M, Fainman-Adelman N, Iyer SN. Experiences of pathways to mental health services for young people and their carers: a qualitative meta-synthesis review. Soc Psychiatry Psychiatr Epidemiol 2021 Mar;56(3):339-361 [FREE Full text] [doi: 10.1007/s00127-020-01976-9] [Medline: 33206200]
9. de Haan AM, Boon AE, de Jong JT, Hoeve M, Vermeiren RR. A meta-analytic review on treatment dropout in child and adolescent outpatient mental health care. Clin Psychol Rev 2013 Jul;33(5):698-711 [FREE Full text] [doi: 10.1016/j.cpr.2013.04.005] [Medline: 23742782]
10. Holmes EA, Ghaderi A, Harmer CJ, Ramchandani PG, Cuijpers P, Morrison AP, et al. The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. The Lancet Psychiatry 2018 Mar;5(3):237-286 [FREE Full text] [doi: 10.1016/s2215-0366(17)30513-8]
11. Ghafari M, Nadi T, Bahadivand-Chegini S, Doosti-Irani A. Global prevalence of unmet need for mental health care among adolescents: a systematic review and meta-analysis. Arch Psychiatr Nurs 2022 Feb;36:1-6 [FREE Full text] [doi: 10.1016/j.apnu.2021.10.008] [Medline: 35094819]
12. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: a systematic review of randomised controlled trials. Child Youth Serv Rev 2021 Aug;127:106073 [FREE Full text] [doi: 10.1016/j.childyouth.2021.106073]
13. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019 Dec 02;2:118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
14. Bell IH, Thompson A, Valentine L, Adams S, Alvarez-Jimenez M, Nicholas J. Ownership, use of, and interest in digital mental health technologies among clinicians and young people across a spectrum of clinical care needs: cross-sectional survey. JMIR Ment Health 2022 May 11;9(5):e30716 [FREE Full text] [Medline: 35544295]
15. Valentine L, D'Alfonso S, Lederman R. Recommender systems for mental health apps: advantages and ethical challenges. AI Soc 2022 Jan 17:1-12 (forthcoming) [FREE Full text] [doi: 10.1007/s00146-021-01322-w] [Medline: 35068708]
16. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
17. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov;35(12):1531-1544 [FREE Full text] [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
18. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020 Mar;88(3):179-195 [FREE Full text] [doi: 10.1037/ccp0000482] [Medline: 32068421]

XSL•FO
RenderX

19. Sauer-Zavala S, Gutner CA, Farchione TJ, Boettcher HT, Bullis JR, Barlow DH. Current definitions of "transdiagnostic" in treatment development: a search for consensus. Behav Ther 2017 Jan;48(1):128-138 [FREE Full text] [doi: 10.1016/j.beth.2016.09.004] [Medline: 28077216]
20. Murray C, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2197-2223 [FREE Full text] [doi: 10.1016/S0140-6736(12)61689-4] [Medline: 23245608]
21. Ehring T, Watkins ER. Repetitive negative thinking as a transdiagnostic process. Int J Cognit Ther 2008 Sep;1(3):192-205 [FREE Full text] [doi: 10.1521/ijct.2008.1.3.192]
22. Topper M, Emmelkamp PM, Ehring T. Improving prevention of depression and anxiety disorders: repetitive negative thinking as a promising target. Appl Prevent Psychol 2010 Jun;14(1-4):57-71 [FREE Full text] [doi: 10.1016/j.appsy.2012.03.001]
23. McEvoy PM, Salmon K, Hyett MP, Jose PE, Gutenbrunner C, Bryson K, et al. Repetitive negative thinking as a transdiagnostic predictor of depression and anxiety symptoms in adolescents. Assessment 2019 Mar;26(2):324-335 [FREE Full text] [doi: 10.1177/1073191117693923] [Medline: 29214855]
24. Segerstrom SC, Tsao JC, Alden LE, Craske MG. Worry and rumination: repetitive thought as a concomitant and predictor of negative mood. Cognit Therapy Res 2000;24:671-688 [FREE Full text] [doi: 10.1023/A:1005587311498]
25. Spinhoven P, van Hemert AM, Penninx BW. Repetitive negative thinking as a predictor of depression and anxiety: a longitudinal cohort study. J Affect Disord 2018 Dec 01;241:216-225 [FREE Full text] [doi: 10.1016/j.jad.2018.08.037] [Medline: 30138805]
26. Wilkinson PO, Croudace TJ, Goodyer IM. Rumination, anxiety, depressive symptoms and subsequent depression in adolescents at risk for psychopathology: a longitudinal cohort study. BMC Psychiatry 2013 Oct 8;13(1):250 [FREE Full text] [doi: 10.1186/1471-244x-13-250]
27. Monteregge S, Tsagkalidou A, Cuijpers P, Spinhoven P. The effects of different types of treatment for anxiety on repetitive negative thinking: a meta - analysis. Clin Psychol Sci Pract 2020 Jun;27(2) [FREE Full text] [doi: 10.1037/h0101777]
28. Spinhoven P, Klein N, Kennis M, Cramer AO, Siegle G, Cuijpers P, et al. The effects of cognitive-behavior therapy for depression on repetitive negative thinking: a meta-analysis. Behav Res Ther 2018 Jul;106:71-85 [FREE Full text] [doi: 10.1016/j.brat.2018.04.002] [Medline: 29699700]
29. Rosenkranz T, Takano K, Watkins ER, Ehring T. Assessing repetitive negative thinking in daily life: development of an ecological momentary assessment paradigm. PLoS One 2020;15(4):e0231783 [FREE Full text] [doi: 10.1371/journal.pone.0231783] [Medline: 32310979]
30. Roefs A, Fried EI, Kindt M, Martijn C, Elzinga B, Evers AW, et al. A new science of mental disorders: using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. Behav Res Ther 2022 Jun;153:104096 [FREE Full text] [doi: 10.1016/j.brat.2022.104096] [Medline: 35500541]
31. Bell I, Arnold C, Gilbertson T, D'Alfonso S, Castagnini E, Chen N, et al. A personalized, transdiagnostic smartphone intervention (Mello) targeting repetitive negative thinking in young people with depression and anxiety: pilot randomized controlled trial. J Med Internet Res 2023 Dec 13;25:e47860 [FREE Full text] [doi: 10.2196/47860] [Medline: 38090786]
32. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173 [FREE Full text] [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
33. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
35. Gönül S, Namlı T, Coşar A, Toroslu IH. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. Artif Intell Med 2021 May;115:102062 [FREE Full text] [doi: 10.1016/j.artmed.2021.102062] [Medline: 34001322]
36. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009 Mar;24(1):485-500 [FREE Full text] [doi: 10.17705/1CAIS.02428]
37. Kitchen CE, Lewis S, Ekers D, Gega L, Tiffin PA. Barriers and enablers for young people, parents and therapists undertaking behavioural activation for depression: a qualitative evaluation within a randomised controlled trial. Psychol Psychother 2023 Jun;96(2):504-524 [FREE Full text] [doi: 10.1111/papt.12452] [Medline: 36808208]
38. McCann TV, Lubman DI, Clark E. The experience of young people with depression: a qualitative study. J Psychiatr Ment Health Nurs 2012 May 22;19(4):334-340 [FREE Full text] [doi: 10.1111/j.1365-2850.2011.01783.x] [Medline: 22070382]
39. Wilhelmsen M, Lillevoll K, Risør MB, Høifødt R, Johansen ML, Waterloo K, et al. Motivation to persist with internet-based cognitive behavioural treatment using blended care: a qualitative study. BMC Psychiatry 2013 Nov 7;13(1):296 [FREE Full text] [doi: 10.1186/1471-244x-13-296]

XSL•FO
RenderX

40. Watson R, Harvey K, Pass L, McCabe C, Reynolds S. A qualitative study exploring adolescents' experience of brief behavioural activation for depression and its impact on the symptom of anhedonia. Psychol Psychother 2021 Jun;94(2):266-288 [FREE Full text] [doi: 10.1111/papt.12307] [Medline: 32918843]
41. Valentine L, McEnery C, O'Sullivan S, Gleeson J, Bendall S, Alvarez-Jimenez M. Young people's experience of a long-term social media-based intervention for first-episode psychosis: qualitative analysis. J Med Internet Res 2020 Jun 26;22(6):e17570 [FREE Full text] [doi: 10.2196/17570] [Medline: 32384056]
42. de Beurs D, van Bruinessen I, Noordman J, Friele R, van Dulmen S. Active involvement of end users when developing web-based mental health interventions. Front Psychiatry 2017;8:72 [FREE Full text] [doi: 10.3389/fpsyt.2017.00072] [Medline: 28515699]
43. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
44. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
45. Lipschitz JM, Van Boxtel R, Torous J, Firth J, Lebovitz JG, Burdick KE, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022 Oct 14;24(10):e39204 [FREE Full text] [doi: 10.2196/39204] [Medline: 36240001]
46. Berry N, Lobban F, Bucci S. A qualitative exploration of service user views about using digital health interventions for self-management in severe mental health problems. BMC Psychiatry 2019 Jan 21;19(1):35 [FREE Full text] [doi: 10.1186/s12888-018-1979-1] [Medline: 30665384]
47. Greene Barker T, O'Higgins A, Fonagy P, Gardner F. A systematic review and meta-analysis of the effectiveness of co-designed, in-person, mental health interventions for reducing anxiety and depression symptoms. J Affect Disord 2024 Apr 01;350:955-973 [FREE Full text] [doi: 10.1016/j.jad.2023.12.080] [Medline: 38199405]
48. Deci EL, Ryan RM. Self-determination theory: a macrotheory of human motivation, development, and health. Can Psychol 2008 Aug;49(3):182-185 [FREE Full text] [doi: 10.1037/a0012801]
49. Lieder F, Chen PZ, Prentice M, Amo V, Tošić M. Gamification of behavior change: mathematical principle and proof-of-concept study. JMIR Serious Games 2024 Mar 22;12:e43078 [FREE Full text] [doi: 10.2196/43078] [Medline: 38517466]
50. Saad A, Bruno D, Camara B, D'Agostino J, Bolea-Alamanac B. Self-directed technology-based therapeutic methods for adult patients receiving mental health services: systematic review. JMIR Ment Health 2021 Nov 26;8(11):e27404 [FREE Full text] [doi: 10.2196/27404] [Medline: 34842556]

Abbreviations

AI: artificial intelligence
JITAI: just-in-time adaptive intervention
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
RNT: repetitive negative thinking

Edited by T de Azevedo Cardoso; submitted 30.06.24; peer-reviewed by A Zhou, S Patel; comments to author 18.10.24; revised version received 28.10.24; accepted 29.10.24; published 27.11.24.

Please cite as:
Valentine L, Arnold C, Nicholas J, Castagnini E, Malouf J, Alvarez-Jimenez M, Bell IH
A Personalized, Transdiagnostic Smartphone App (Mello) Targeting Repetitive Negative Thinking for Depression and Anxiety: Qualitative Analysis of Young People's Experience
J Med Internet Res 2024;26:e63732
URL: https://www.jmir.org/2024/1/e63732
doi:10.2196/63732
PMID:

©Lee Valentine, Chelsea Arnold, Jennifer Nicholas, Emily Castagnini, Jessi Malouf, Mario Alvarez-Jimenez, Imogen H Bell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Digital Divide and Cognitive Disparities Among Older Adults: Community-Based Cohort Study in China

Yumeng Li[1], BD; Chen Liu[1], PhD; Jiaqing Sun[2], PhD; Junying Zhang[3], PhD; Xin Li[1], PhD; Zhanjun Zhang[1], PhD

[1]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China

[2]Department of Management, London School of Economics and Political Science, London, United Kingdom

[3]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China

Corresponding Author:
Xin Li, PhD
State Key Laboratory of Cognitive Neuroscience and Learning
Beijing Normal University
Xinjiekouwai Street No 19
Beijing, 100875
China
Phone: 86 010 58802882
Fax: 86 010 58803882
Email: lixin99@bnu.edu.cn

Abstract

Background: The widespread adoption of information and communication technologies (ICTs) further deepens disparities in resource access, particularly among the aging population. However, the relationship between these factors and their resulting impact on cognitive abilities remains uncertain.

Objective: This study aims to investigate the potential impact of the digital divide on individuals' cognitive function and its association with the development and reversion of mild cognitive impairment (MCI).

Methods: This cohort study used data from Beijing Aging Brain Rejuvenation (BABRI) study applying a multistage cluster sampling design between 2008 and 2020. The digital divide was quantified by the frequency of using ICTs. Analysis of covariance (ANCOVA), mixed linear models, and Cox proportional hazards models were used to model the association of digital divide and multidomain cognition.

Results: Among the 10098 participants, nearly half (n=4941, 48.9%) faced the digital divide, which was associated with a worse performance in processing speed (F_{10096}=10.67; P<.001; effect size r=0.42), rather than memory, executive function, and language. The model indicated that individuals' physical and mental health, combined with their educational and occupational prestige, influenced the resources they attained, which ultimately caused the digital divide. Moreover, longitudinal data revealed that older adults who successfully crossed the digital divide during the tracking process and those who had already done so prior to tracking showed significantly slower rates of decline in processing speed (B=–1.98, P<.05; B=–2.62, P<.01) and general cognitive function (B=3.50, P<.001; B=3.13, P<.01). Additionally, overcoming the digital divide was also associated with a lower risk of developing MCI (hazard ratio [HR] 0.5, 95% CI 0.34-0.74; HR 0.43, 95% CI 0.29-0.62) and a greater probability of reversion from MCI to normal cognition (HR 6, 95% CI 3.77-9.56; HR 9.22, 95% CI 5.63-15.11).

Conclusions: Overcoming the digital divide was significantly associated with improved cognitive function, a slower aging rate in cognitive performance, a reduced risk of developing MCI, and a higher likelihood of reverting from MCI to normal cognition.

(J Med Internet Res 2024;26:e59684) doi:10.2196/59684

KEYWORDS

digital divide; internet use; cognitive aging; mild cognitive impairment; socioeconomic status; resource inequality; cognitive disparities; cognitive function; elderly; older adult; aging; community-based; cohort study; China; ANCOVA; mixed linear model; Cox proportional hazards; mental health; internet; digital health; information and communication technologies; ICT; mHealth

XSL•FO
RenderX

Introduction

In today's modern society, digital information and communication technologies (ICTs) have become indispensable in our daily lives. The "digital divide" has become a vivid metaphor to describe the perceived inequality of those who either are unable to make use of ICTs [1]. The digital divide has a profound impact on the older population(between 50 and 90 years), which is experiencing rapid growth in numerous countries, including China, where both absolute numbers and relative proportions are continuously increasing. This demographic has significantly lower rates of acceptance and utilization of ICTs, leading to an increasing trend of exclusion among older adults in the rapidly advancing information society [2,3].

This phenomenon can be better understood through the lens of the resources and appropriation theory of the digital divide [1,4,5]. The core hypothesis of this theory posits that disparities in personal characteristics (such as health and habits) and positional characteristics (such as education and occupation) contribute to unequal access to diverse resources (such as mental well-being and social relationships), thereby resulting in the digital divide. However, there is currently a scarcity of empirical studies that comprehensively incorporate various influential factors and investigate the relationship pathways underlying the digital divide.

The digital divide can potentially contribute to the adoption of a less engaging lifestyle throughout older adults' lifespans, which may diminish their "cognitive reserve" (the resource of the brain can get to tolerate the age-related changes or pathology related to dementia) and ultimately hasten the deterioration of cognitive function [6]. Evidence suggests that daily internet use is associated with a reduced risk of mild cognitive impairment (MCI) [7]. Moreover, longitudinal research indicates that overcoming the digital divide is associated with a decline in the incidence of MCI among older adults after 10 years [8]. Nonetheless, the existing literature focuses too narrowly on cognitive domains and fails to consider the specificity of cognitive differences caused by the digital divide. Similarly, the longitudinal changes in multidomain cognitive function influenced by the digital divide have not been adequately considered. Additionally, while many studies have investigated the progression of MCI, some found that a considerable proportion of individuals with MCI, ranging from 24% to 50%, may revert to a normal cognition (NC) status [9,10]. Thus, it is crucial to consider how the digital divide influences both the development and reversion of MCI.

It can be foreseen that the digital divide will put the older population in a state of resource inequality, impacting their future cognitive development. Based on the large-scale Beijing Aging Brain Rejuvenation (BABRI) study database, this study will investigate the following aspects: (1) the potential relationship among various influential factors contributing to unequal resource distribution related to the digital divide; (2) the cross-sectional differences and longitudinal aging rate alternations in multidomain cognitive function caused by the digital divide; and (3) the associations between the digital divide and the development and reversion of MCI, while controlling for normal risk factors.

Methods

Study Design

The BABRI cohort study is based on the registry of a large community population in Beijing, which collects comprehensive information on aging and tracks changes in cognitive function over the years [11]. All the participants were aged 50 years and above at the time of baseline enrollment, could live independently, did not have nervous system diseases or psychiatric disorders, and had 6 or more years of formal education, which was required for the cognitive assessments. The study had a multistage cluster sampling design, and participants were mainly recruited from communities in Beijing between 2008 and 2019. A total of 10098 qualified participants were recruited from the communities of these districts. All participants registered in BABRI were visited every 2 or 3 years over the total duration of the 20-year project.

This study will exclude longitudinal participants in the next phase of analysis based on the following criteria: (1) clinical diagnoses of neurodegenerative diseases (eg, Alzheimer disease and Parkinson disease), serious brain diseases (eg, severe cerebrovascular diseases, brain tumors, and brain trauma) or psychiatric disorders (eg, severe major depression disease, bipolar disorder, and schizophrenia); (2) history of substance or alcohol abuse/dependence; and (3) missing assessment indicators for the digital divide during the tracking process.

Measurements

The Resources and Appropriation Theory of the digital divide was developed over a 10-year period, culminating in its full and mature presentation [5]. The core hypothesis of this theory posits that individuals' characteristics and status lead to unequal access to resources, ultimately contributing to the emergence of the digital divide. In this study, the influencing factors of the digital divide were classified into 3 main parts. The independent variables included personal and positional categories, while the mediation variable included the resources.

Personal Categories

Overview

Personal categories encompass various factors, such as physical health, which includes subjective health status, BMI, and chronic disease; mental health, including the Geriatric Depression Scale (GDS) and Facial Affect Coding System (FACE); and other factors like smoking and alcohol consumption.

Physical Health: Subjective Health Status, BMI, and Chronic Diseases

Subjective health status refers to a question asking participants to evaluate their overall health status with the following options: "good," "fair," "poor," or "very poor." BMI is calculated by dividing a person's weight in kilograms by their height in meters squared. The participants' medical histories were collected regarding hypertension, diabetes, and hyperlipidemia. This

XSL•FO
RenderX

information was validated using the diagnosis and management records from community health service institutions.

Mental Health

The GDS is a standardized depression scale for older adults [12]. The FACE scale was used to measure subjective well-being, with higher scores indicating lower levels of well-being among older adults.

Daily Unhealthy Habits

Smoke and alcohol consumption were binary variables indicating whether the participants smoked or consumed alcohol every day.

Positional Categories

Positional categories included the variables reflecting participants' social status, such as educational status, subjective social status, and occupational prestige. Educational status refers to an individual's specific duration of education. Subjective social status was assessed through the question, "How would you evaluate your current socioeconomic status?" The responses were rated on a 4-point scale: rich, upper-middle, lower-middle, and poor. Occupational prestige was measured using the Occupational Prestige Scale based on data from the Chinese national survey [13].

Resources

The resources included economic, mental, and social resources. Economic resources refer to an individual's current monthly income, categorized into 13 levels based on increments of 500. Mental and social resources were obtained through a weighted summation based on relevant items derived from the Leisure Activity Scale (LAS), which is a 5-point scale that consists of 23 subitems used to assess cognition and social activity engagements in daily life [14].

Quantifying the Digital Divide

The quantification indicator of the digital divide includes an item from the LAS about the frequency of using ICT: "How often do you use computer and mobile phone?" [15-17]. Participants with scores of 0 (never), 1 (≥ once a year), and 2 (≥ once a month) were classified into the Digital Divide (DD) group, while those with scores of 4 (≥once a week) and 5 (every day) into the Overcoming the Digital Divide (ODD) group. The DD group indicated individuals facing the digital divide, while the ODD group represented those who overcame it.

For cross-sectional data, participants were classified only into DD and ODD groups. To investigate the impact of ODD on cognitive aging using longitudinal data, participants were categorized into 3 groups. Two of these groups represented participants who remained in the DD or ODD group, while the third group, known as the Transition (Trans) group, included participants who moved from the DD group to the ODD group during the tracking process. Participants who transitioned from the ODD to the DD group were excluded, as crossing the digital divide is considered a relatively stable state. A transition from ODD to DD may indicate uncontrollable external factors, which are not the focus of this study.

Cognitive Assessments

All participants underwent a battery of neuropsychological tests at the baseline recruitment [11]. The assessment involved general cognitive ability and cognitive function across 5 domains: memory, language, attention, visuospatial abilities, and executive function. General cognitive ability was tested using the Chinese version of the Mini-Mental State Examination (MMSE). Memory was tested using the Auditory Verbal Learning Test (AVLT) and the Rey-Osterrich Complex Figure (ROCF) test [18]. Executive function was tested using the Stroop Color Word Test (SCWT) and the Trail Making Test (TMT) [19]. Spatial processing was assessed using the Clock Drawing Test (CDT) [20] and the Rey-Osterrieth Complex Figure-Copy (R-Ocopy) test [18]. Attention was evaluated using the Symbol Digit Modification Test (SDMT) [21] and the Trail Making Test Part A (TMT-A) [19]. Language was tested using the Boston Naming Test (BNT) and the Verbal Fluency Test (VFT).

MCI Diagnostic Criteria

The diagnostic criteria for MCI included the following [22]: individuals had to be without dementia, exhibit subjective cognitive decline, and present with at least 1 cognitive domain displaying 2 test scores that fell below 1.5 SD from the mean score of the same age and education level group in objective tests. Additionally, their general cognitive abilities and daily functioning had to be typically unaffected.

Demographic Variables

Demographic variables included age, gender marital status, and residential status. Marital status included 3 options: married, divorced, and widowed. Residential status involved 3 options: living alone, living with a spouse, and living with children.

Statistical Analysis

The participants' baseline characteristics were presented as mean (SD) for continuous variables, while categorical variables were represented as frequencies and proportions.

The Z refers to the Z value obtained from performing a Mann-Whitney test. Intergroup variations in influential factors between the DD and ODD groups were assessed through independent sample *t* tests, chi-square tests, and nonparametric tests. We also used Cohen *d* and Cramer *V* to calculate effect sizes [23,24].

An intermediary model was constructed to elucidate the impact mechanism of the digital divide, using personal and positional characteristics as independent variables, resource acquisition as a mediating variable, and the digital divide as the dependent variable.

Intergroup differences in cognitive abilities were examined using analysis of covariance (ANCOVA) controlling for age, education level, gender, and chronic diseases. Binary logistic regression was employed to investigate the predictive impact of the digital divide variable on the occurrence of MCI.

A linear Bayesian change-point regression was performed on the age group–averaged data to compare the decline trajectories of multidomain cognition among the older population.

XSL•FO
RenderX

To effectively compare the impact of cognitive abilities among the DD, Trans, and ODD groups, dummy variable coding was applied to the longitudinal data to incorporate the effects of Trans-DD and ODD-DD in the model.

The mixed linear model (MLM) was used to examine the influence of the digital divide variable on the rate of cognitive aging at an individual level. Initially, we established the null model and unconditional growth model. The null model was used to determine the hierarchical structure of the longitudinal data for different cognitive functions, which was suitable for MLM analysis. The unconditional growth model was used to identify significant aging patterns in various cognitive functions over time. After selecting these 2 models, we constructed the full model that encompassed the following: (1) Level 1, which described individual cognitive level aging patterns, and (2) Level 2, which investigated the influence of the digital divide variable on aging patterns across multiple cognitive abilities in individuals.

To investigate the impact of the digital divide variable on individual MCI development and outcomes, we employed the Cox proportional hazards model. Two models were developed, the first focusing on the progression from NC to MCI and the second focusing on the transition from MCI to NC.

Ethical Considerations

The study was conducted in accordance with the institutional review board at the Imaging Center for Brain Research at Beijing Normal University (ICBIR_A_0041_002_02) and was approved in March 2015. We used STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) as our reporting framework. Written informed consent followed by sociodemographic information was obtained from the participants before initiating the neuropsychological tests. All participants were reimbursed with daily necessities valued at 20 RMB (approximately $10 US) and provided with a free screening report covering multiple domains of cognition as a token of appreciation.

Results

Overview

This study included 10,098 participants in the cross-sectional analysis, with a mean age of 66.7 (SD 7.9) years. Among them, 6095 (60.4%) were female and 4003 (39.6%) were male. The mean educational level was 10.7 (SD 3.5) years), with 4941 (48.9%) in the DD group and 5157 (51.1%) in the ODD group. The longitudinal data analysis included 2092 participants, categorized into the DD, Trans, and ODD groups.

Among these participants, tracking outcomes fell into several categories: stable NC (NC→NC: N=1473), progression from NC to MCI (NC→MCI, N=190), stable MCI (MCI→MCI: N=193), and improvement from MCI to NC (MCI→NC, N=201).

Characteristic Differences Among Older the Population Facing the Digital Divide

As shown in Table 1, the personal characteristics of the ODD group include hyperlipidemia, low depression scores, high subjective well-being, and nonsmoking. Their positional characteristics included higher education levels and higher self-perceived socioeconomic status. Additionally, the ODD group benefitted from greater access to economic, mental, and social resources.

XSL•FO
RenderX

Table 1. Characteristics and the digital divide among the older population.

Characteristics	DD[a] (n=4941)	ODD[b] (n=5157)	$t/\chi^2/Z$	FDR[c]-adjusted *P* value	Cohen *d*
Demographic information					
Age, mean (SD)	67.65 (7.88)	64.60 (7.74)	19.66	<.001	0.39
Female, n (%)	2908 (58.9)	3187 (61.8)	9	.01	0.03
Married, n (%)	4060 (84.1)	4502 (89.6)	143.51	<.001	0.12
Divorced, n (%)	659 (13.6)	336 (6.7)	—[d]	—	—
Widowed, n (%)	109 (2.3)	185(3.7)	—	—	—
Live alone, n (%)	458 (10.8)	372 (8)	61.53	<.001	0.08
Live with spouse, n (%)	3170 (74.9)	3811(81.7)	—	—	—
Live with children, n (%)	605(14.3)	480 (10.3)	—	—	—
Personal characteristics					
Physical health					
Subjective health, median (IQR)	1876.5 (2222.0)	1889.5 (2037.5)	–2.135	.05	0.04
BMI, mean (SD)	36.19 (820.14)	25.70 (120.96)	0.91	.36	—
Hypertension, n (%)	2342 (49.5)	2414 (48.6)	0.75	.39	0.01
Diabetes, n (%)	1024 (21.7)	4948 (21.6)	0.93	.93	0.001
Hyperlipidemia, n (%)	1328 (31.4)	1969 (42.5)	114.97	<.001	0.11
Mental health					
SWB[e], mean (SD)	2.63 (1.19)	2.51 (1.29)	4.25	<.001	0.1
GDS[f], mean (SD)	8.18 (5.96)	7.61 (5.63)	4.93	.004	0.1
Lifestyle					
Smoking, n (%)	1413 (29.6)	1242 (24.5)	31.74	<.001	0.06
Drinking, n (%)	1183 (27.8)	1320 (27.5)	0.18	.68	0.004
Positional characteristics					
Education, mean (SD)	9.35 (3.43)	11.76 (3.16)	–36.73	<.001	0.73
Self-perceived status, median (IQR)	4451.5 (2450.0)	4926.5 (2363.5)	–7.36	<.001	0.15
Occupation, mean (SD)	0.61 (0.11)	0.61 (0.11)	0.84	.40	0
Resources					
Economic resources, median (IQR)	3828.5 (3437.5)	6268.0 (4087)	–36.66	<.001	0.78
Mental resources, mean (SD)	47.36 (18.66	72.09 (23.04)	–59.37	<.001	1.18
Social resources, mean (SD)	24.66 (13.49	35.83 (15.40)	–38.80	<.001	0.77

[a]DD: Digital Divide group.
[b]ODD: Overcoming the Digital Divide group.
[c]FDR: false discovery rate.
[d]–: not applicable.
[e]SWB: subjective well-being
[f]GDS: Geriatric Depression Scale.

Resources as Mediators in the Digital Divide

The mediation model in Figure 1 illustrates the potential mechanisms behind the digital divide formation among older adults. It suggests that inequalities in individual characteristics and social status contribute to unequal access to economical, mental, and social resources, ultimately leading to the digital divide.

XSL•FO
RenderX

Figure 1. Mediation model Elucidating the Mechanism of the Digital Divide. The model is adjusted for the demographic variables (age, gender, marital status and residential status). Abbreviations: BMI: body mass index; SHS: subjective health score; GDS: Geriatric Depression Scale; SWB: subjective well-being; SSES: subjective socioeconomic scores.

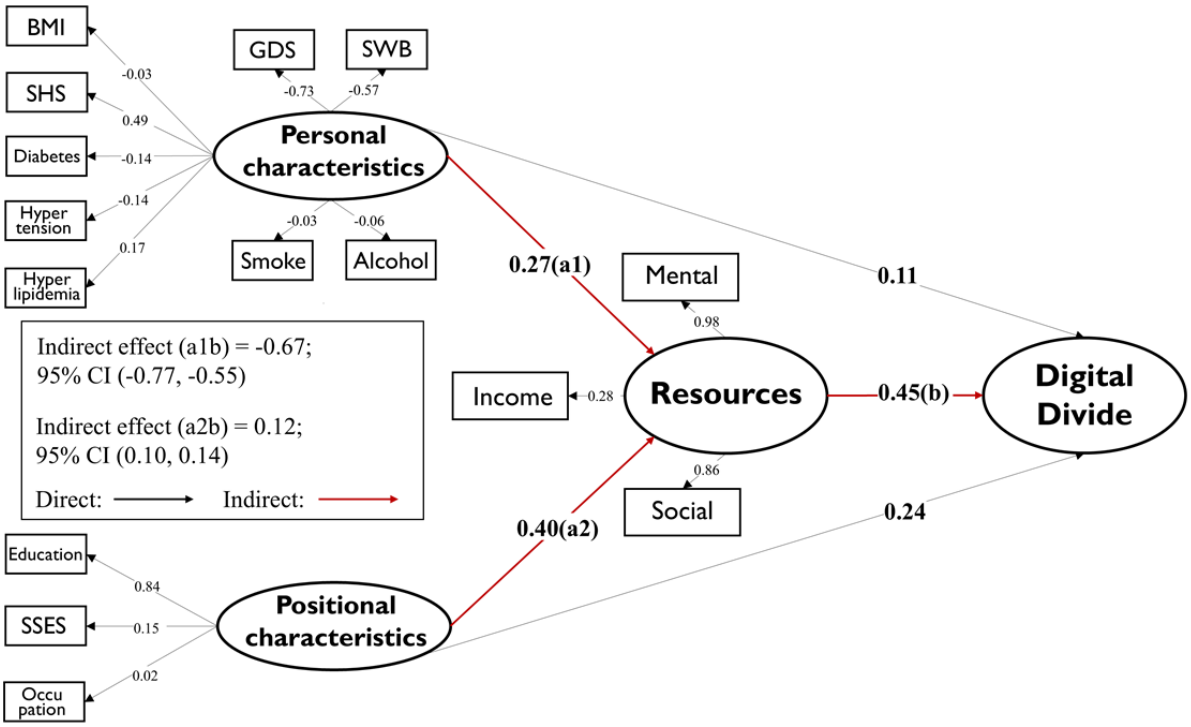

Cognitive Differences Among Older Adults in the Digital Divide

Table 2 presents the differences in multidomain cognitive function between the DD and ODD groups. The results indicate that only the SDMT score of processing speed (F_{10096}=10.67; *P*<.001) and working memory showed significant differences between the 2 groups. However, the effect size of processing speed was the largest (Cohen *d*=0.92).

XSL•FO
RenderX

Table 2. Digital divide and cognitive performance (corrected for age, sex, education, and chronic disease).

Cognitive domain	DD[a] (n=4941), mean (SD)	ODD[b] (n=5157), mean (SD)	*F* statistic	FDR[c]-adjusted *P* value	Cohen *d*
General mental status					
MMSE[d]	26.12 (3.15)	27.51 (2.13)	1.44	.43	0.52
Episodic memory					
AVLT[e]	23.48 (9.59)	27.75 (8.79)	0.12	.78	0.46
R-Odelay[f]	10.52 (6.71)	13.70 (7.11)	2.15	.30	0.46
Spatial processing					
CDT[g]	22.33 (5.87	24.29 (5.05)	0.97	.46	0.36
R-Ocopy[h]	31.66 (6.63	33.96 (3.91)	0.48	.64	0.42
Processing speed					
SDMT[i]	25.83 (10.98	35.92 (11.01)	10.67	<.001	0.92
TMT-A[j]	74.33 (36.10	58.34 (22.93)	0.26	.69	0.53
Executive function					
SCWT[k]	94.12 (35.84	79.81 (25.73)	1.02	.46	0.46
TMT-B[l]	205.4 (92.57	157.4 (62.20)	0.36	.66	0.61
Language					
VFT[m]	39.45 (9.57	45.48 (9.03)	0.01	.99	0.65
BNT[n]	20.86 (4.37	24 (3.13)	5.33	.07	0.83
Working memory					
DSTtotal[o]	11.91 (4.68)	12.65 (4.65)	7.93	0.03	0.16

[a]DD: Digital Divide group.
[b]ODD: Overcoming the Digital Divide group.
[c]MMSE: Mini-Mental State Examination.
[d]FDR: false discovery rate.
[e]AVLT: Auditory Verbal Learning Test
[f]R-Odelay: Rey-Osterrieth Complex Figure Test delayed recall score.
[g]CDT: Clock Drawing Test.
[h]R-Ocopy: Rey-Osterrieth Complex Figure Test copy score.
[i]SDMT: Symbol Digit Modification Test.
[j]TMT-A: Trail-Making Test part A.
[k]SCWT: Stroop Color Word Test.
[l]TMT-B: Trail-Making Test part B.
[m]VFT: Verbal Fluency Test.
[n]BNT: Boston Naming Test.
[o]DSTtotal: Digital Span Test total score.

Cross-sectional Association Between Overcoming the Digital Divide and Risk of MCI Incidence

The risk of developing MCI was significantly higher in the DD group compared to the ODD group (Table S1 in Multimedia Appendix 1), with an odds ratio (OR) of 3.06 (95% CI 2.74-3.42). Among other factors influencing MCI, such as age (OR 1.38, 95% CI 1.24-1.54), diabetes (OR 1.23, 95% CI 2.74-3.42], and hyperlipidemia (OR 0.74, 95% CI 0.66-0.83), DD has the highest OR value.

Longitudinal Evidence of the Digital Divide's Influence on Cognitive Aging

Significant interindividual variations were observed for all cognitive abilities in the null model, indicating the potential for constructing the subsequent mixed linear model (Table S2 in Multimedia Appendix 1). However, in the unconditional growth

XSL•FO
RenderX

model, the age-related change trend of CDT did not reach statistical significance, which prevented the construction of the full model.

Table 3 shows that the digital divide significantly influenced the aging rate of multidomain cognitive function. Compared to the ODD and Trans groups, the DD group exhibited a faster cognitive aging rate, including MMSE (B_{13}=3.50, P<.001; B_{14}=3.13, P=.002), processing speed (TMTA: B_{13}=–1.98, P=.04; B_{14} =–2.62, P<.001), and language (BNT: B_{13}=3.16, P=.002).

Table 3. Longitudinal evidence of the digital divide's influence on cognitive aging[a].

Category	Aging rate (B1j)					Initial state (B0j)				
	Age (B_{11})	Gender (B_{11})	Edu[b] (B_{12})	Trans-DD[c] (B_{13})	ODD-DD[d] (B_{14})	Age (B_{01})	Gender (B_{02})	Edu (B_{03})	Trans-DD(B_{03})	ODD-DD (B_{04})
General mental status										
MMSE[e]	–5.92***	–1.41	1.73	3.50***	3.13**	–6.63***	0.90	1.90	–1.95	–1.11
Episodic memory										
AVLT[f]	–4.74***	–1.88	1.07	0.27	–0.30	1.07	–1.12	2.79**	1.18	2.13*
RO-delay[g]	–3.15**	–1.79	1.57	1.94	1.47	–2.41*	3.83***	1.85	0.05	1.30
Spatial processing										
RO-copy[h]	–2.66**	0.18	2.25*	1.44	0.72	1.42	0.27	1.44	0.32	1.36
Processing speed										
SDMT[i]	–2.48*	2.39*	0.91	1.69	1.61	–3.54***	–4.07***	5.29***	2.44*	1.05
TMT-A[j]	3.71***	–0.33	–1.91	–1.98*	–2.62**	–3.90***	–0.98	–2.821*	0.21	–0.26
Executive function										
StroopC[k]	3.42**	0.95	0.34	-0.94	–0.84	–1.63	1.80	–2.04*	–0.13	–1.04
TMT-B[l]	4.74***	–0.18	–2.28*	–1.23	–1.71	–3.06**	–0.56	–2.80**	–1.13	–2.11*
Working Memory										
DST[m]	–4.16***	–2.02*	1.31	-0.84	–1.42	0.51	2.66**	3.77***	2.73**	4.94***
Language										
VFT[n]	–5.61***	–0.67	0.93	1.61	1.15	5.58***	0.07	2.78	0.66	2.20*
BNT[o]	–3.83***	–4.78***	2.08*	3.16**	1.34	0.73	8.10***	2.67**	–0.94	2.09*

[a]The coefficients of MLM (B1j, B0j) refer to the function of Level 1 and Level 2 (see the Statistical Analysis subsection for more information).

[b]Edu: education level.

[c]Trans-DD: participants who overcame the digital divide only during the tracking process compared to participants who remained facing the digital divide.

[d]ODD-DD: participants who consistently overcame the digital divide throughout the entire tracking process compared to those who remained facing the digital divide.

[e]MMSE: Mini-Mental State Examination.

[f]AVLT: Auditory Verbal Learning Test

[g]RO-delay: Rey-Osterrieth Complex Figure Test delayed recall score.

[h]R-Ocopy: Rey-Osterrieth Complex Figure Test copy score.

[i]SDMT: Symbol Digit Modification Test.

[j]TMT-A: Trail-Making Test part A.

[k]StroopC: Stroop Color test.

[l]TMT-B: Trail-Making Test part B.

[m]DST: Digital Span Test.

[n]VFT: Verbal Fluency Test.

[o]BNT: Boston Naming Test.

***P<.001; **P<.01; * P<.05.

XSL•FO
RenderX

Longitudinal Evidence of the Digital Divide's Influence on the Development and Reversion of MCI

As Table 4 shows, compared to the DD group, both the Trans and ODD groups had a significantly lower probability of developing MCI (Trans-DD: hazard ratio [HR] 0.50, 95% CI 0.34-0.74; ODD-DD: HR 0.43, 95% CI 0.29-0.62), while they have a significantly greater probability of reversion from MCI to the healthy (Trans-DD: HR 6, 95% CI, 3.77-9.56; ODD-DD: HR, 9.22, 95% CI, 5.63-15.11).

Table 4. Longitudinal evidence of the digital divide's influence on the development and reversion of MCI[a].

Independent variables	MCI development (NC[b]→MCI) (n=190), HR[c] (95% CI)	MCI reversion (MCI→NC) (n=201), HR (95% CI)
trans-DD[d]	0.5 (0.34-0.74)	6 (3.77-9.56)
ODD-DD[e]	0.43 (0.29-0.62)	9.22 (5.63-15.11)
Gender	0.76 (0.57-1.02)	1.06 (0.78-1.45)
Age	1.08 (1.06-1.10)	1.01 (0.99-1.03)
Edu[f]	0.93 (0.90-0.97)	0.99 (0.95-1.04)
Hypertension	1.36 (1.01-1.82)	0.97 (0.72-1.31)
Diabetes	1.43 (1.03-1.98)	1.2 (0.89-1.63)
Hyperlipidemia	1.05 (0.79-1.99)	0.88 (0.67-1.17)

[a]MCI, mild cognitive impairment.

[b]NC: normal cognition.

[c]HR: hazard ratio.

[d]Trans-DD: participants who overcame the digital divide only during the tracking process compared to participants who remained facing the digital divide.

[e]ODD-DD: participants who consistently overcame the digital divide throughout the entire tracking process compared to those who consistently remained facing the digital divide.

[f]Edu: education level.

Sensitivity Analyses

To verify that the cross-group differences between the DD group and the ODD group were not due to differences in age and education level, we randomly selected 50 participants from the DD group. Using propensity score matching, we matched these participants with those from the ODD group based on age, education level, and gender. We conducted the same statistical analysis using identical procedures, and the results were consistent, largely replicating the initial findings (Tables S3-S5 in Multimedia Appendix 1).

Discussion

Principal Findings

This population-based cohort study verified the relationship pathway of the digital divide driven by resource inequality influenced by various factors. It also revealed a significant association between the digital divide and both cross-sectional and longitudinal differences in cognition, as well as the development and reversion of MCI. The population that came the digital divide exhibited higher scores in processing speed compared to those who suffered from the digital divide. Additionally, overcoming the digital divide was associated with a slower aging rate in MMSE, processing speed, and language skills, along with a reduced probability of developing MCI and an increased probability of transitioning from MCI into a healthy state. This study examines, for the first time, the impact of the digital divide on multiple cognitive domains and the reversion of MCI in the aging population. Additionally, a model of the digital divide among older adults was developed. Due to its correlation with the use of social media, previous studies have often associated the digital divide with mental health issues [15,25]. Nonetheless, our work aims to shift public attention to the cognitive abilities essential for supporting the daily functioning of the aging population, in response to the potential disability that may arise during their aging process.

As our results have demonstrated, the digital divide is significantly associated with the development and reversion of MCI. Given the numerous clinical treatment failures observed in Alzheimer disease, MCI has increasingly been recognized as a crucial opportunity for disease treatment and intervention [26]. Our study's innovative findings clearly indicate that crossing the digital divide plays a significant role in the conversion from MCI to NC. Despite a tracking period of only 2 to 3 years, the conversion rate was more than 6 times higher in the DD group. This provides valuable insights for early intervention in the MCI stage of Alzheimer disease, as the utilization of ICTs is cost-effective and has a significant impact on neuroplasticity. The underlying reason for its considerable benefits is that ICTs reshape individuals' interactive environments, exposing users to a wealth of environmental stimuli and information [27,28]. This is especially applicable to older adults because bridging the digital divide exposes them to a new lifestyle characterized by abundant information, necessitating a learning process.

Consistent with previous studies on the purpose of online cognitive training to improve cognitive function in older adults, our findings provide further epidemiological evidence for neural

XSL•FO
RenderX

plasticity and cognitive improvement during the aging process. We recommend that cognitive training methods be developed and integrated with digital platforms and portable devices to delay age-related cognitive decline. We also recommend exploring novel approaches for early intervention in pathological aging.

Additionally, our findings indicate that the impact of the digital divide on different cognitive functions is selective. Overcoming the digital divide, both at the cross-sectional group level and the longitudinal developmental level, has a significant positive effect on processing speed. The continuous influx of online information and multimedia streams with multiple presentation modes during the use of ICTs encourages older adults to engage in attention-switching and simultaneous attention to multiple tasks, thereby enhancing the processing speed of cognitive resources [29]. Furthermore, studies have shown that many cognitive-related variables are considered to reflect processing speed, especially in the aging process [30]. Therefore, the influence of the digital divide on processing speed is also evident in overall cognitive abilities (MMSE), according to our findings.

Limitations

This study has a few limitations. First, during the third wave of longitudinal data collection, the sample size of the data was lower than expected due to the COVID-19 pandemic. Second, relatively few variables were used to measure mental health, which may not fully capture the psychological factors affecting the participants. Third, regarding the quantification of the digital divide, this study did not distinguish the specific uses of ICTs. It is possible that some participants passively received phone calls and messages without actively engaging with ICTs to receive more stimuli. They may have been misclassified and placed into the ODD group.

Conclusion

In this cohort study, the underlying mechanism of the digital divide among older adults was influenced by their personal physiological and psychological characteristics, as well as their positional status, impacting their access to economic, cognitive, and social resources. These factors ultimately determined their ability to overcome the digital divide. Based on the cross-sectional association, processing speed shows the strongest effect size in relation to the digital divide. Furthermore, overcoming the digital divide could delay the decline of general cognition, processing speed, and certain language functions. It also decreases the development of MCI and increases the likelihood of reversion from MCI. These findings can help inform strategies for dementia prevention and cognitive reserve strengthening in later life, particularly in the context of modifiable daily routines.

Acknowledgments

We are grateful to all the participants and their relatives as well as our cooperating community for helping us complete this trial.

Data Availability

The data sets analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

LYM and LX conceived the study. LYM and LC participated in the data collection. LYM carried out statistical analysis. LYM and LX analyzed and interpreted the data. LYM wrote the manuscript. LX and SJQ revised the manuscript. ZZJ and ZJY supervised and coordinated the study. All authors contributed to the improvement of this manuscript and approved the final version for submission.

Author ZZ is co-corresponding author. Their email address is zhang_rzs@bnu.edu.cn.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 34 KB - jmir_v26i1e59684_app1.docx]

References

1. van Dijk J, Hacker K. The digital divide as a complex and dynamic phenomenon. Inf Soc 2003 Sep;19(4):315-326. [doi: 10.1080/01972240309487]
2. Anderson M, Perrin A. Tech adoption climbs among older adults. Pew Research Center. URL: https://www.pewinternet.org/wp-content/uploads/sites/9/2017/05/PI_2017.05.17_Older-Americans-Tech_FINAL.pdf [accessed 2017-05-17]
3. Horrigan J. Digital Readiness: Nearly one-third of Americans lack the skills to use next-generation “Internet of things” applications. JBHorrigan.com. 2014. URL: https://jbhorrigan.weebly.com/uploads/3/0/8/0/30809311/digital_readiness.horrigan.june2014.pdf [accessed 2014-06-13]

XSL•FO
RenderX

4. Pick J, Sarkar A. Theories of the digital divide: critical comparison. 2016 Presented at: 49th Hawaii International Conference on System Sciences (HICSS); January 5-8; Maui, HI p. 3888-3897. [doi: 10.1109/hicss.2016.484]
5. Van DJ. The Digital Divide. Canbridge, UK and Medford, MA: John Wiley & Sons; Jan 25, 2020.
6. Stern Y. Cognitive reserve in ageing and Alzheimer's disease. Lancet Neurol 2012 Nov;11(11):1006-1012. [doi: 10.1016/s1474-4422(12)70191-6]
7. Hamer M, Stamatakis E. Prospective study of sedentary behavior, risk of depression, and cognitive impairment. Med Sci Sports Exerc 2014 Apr;46(4):718-723 [FREE Full text] [doi: 10.1249/MSS.0000000000000156] [Medline: 24121248]
8. Quialheiro A, Figueiró TH, Rech CR, Marques LP, Paiva KMD, Xavier AJ, et al. Can internet use reduce the incidence of cognitive impairment? Analysis of the EpiFloripa Aging Cohort Study (2009-2019). Prev Med 2022 Jan;154:106904. [doi: 10.1016/j.ypmed.2021.106904] [Medline: 34863810]
9. Malek-Ahmadi M. Reversion from mild cognitive impairment to normal cognition. Alzheimer Dis Assoc Disord 2016;30(4):324-330. [doi: 10.1097/wad.0000000000000145]
10. Chung J, Yoon H, Kim H, Choi KY, Lee JJ, Lee KH, et al. Reversion from mild cognitive impairment to normal cognition: false-positive error or true restoration thanks to cognitive control ability? Neuropsychiatr Dis Treat 2019;15:3021-3032 [FREE Full text] [doi: 10.2147/NDT.S223958] [Medline: 31749620]
11. Yang C, Li X, Zhang J, Chen Y, Li H, Wei D, Beijing Aging Brain Rejuvenation Initiative workgroup. Early prevention of cognitive impairment in the community population: The Beijing Aging Brain Rejuvenation Initiative. Alzheimers Dement 2021 Oct;17(10):1610-1618. [doi: 10.1002/alz.12326] [Medline: 33792187]
12. Yesavage J. Geriatric depression scale. Psychopharmacol Bull 1988;24(4):709-711. [doi: 10.1037/t00930-000]
13. Li C. Prestige stratification in the contemporary China: occupational prestige measures and socio-economic index. Sociological Research 2005;2(74):74-102. [doi: 10.1515/9780773576681-015]
14. Karp A, Paillard-Borg S, Wang H, Silverstein M, Winblad B, Fratiglioni L. Mental, physical and social components in leisure activities equally contribute to decrease dementia risk. Dement Geriatr Cogn Disord 2006;21(2):65-73. [doi: 10.1159/000089919] [Medline: 16319455]
15. Cotten SR, Ford G, Ford S, Hale TM. Internet use and depression among retired older adults in the United States: a longitudinal analysis. J Gerontol B Psychol Sci Soc Sci 2014 Sep;69(5):763-771. [doi: 10.1093/geronb/gbu018] [Medline: 24671896]
16. Duplaga M. The association between Internet use and health-related outcomes in older adults and the elderly: a cross-sectional study. BMC Med Inform Decis Mak 2021 May 06;21(1):150 [FREE Full text] [doi: 10.1186/s12911-021-01500-2] [Medline: 33957909]
17. Sun X, Yan W, Zhou H, Wang Z, Zhang X, Huang S, et al. Internet use and need for digital health technology among the elderly: a cross-sectional survey in China. BMC Public Health 2020 Oct 11;20(1):1386 [FREE Full text] [doi: 10.1186/s12889-020-09448-0] [Medline: 32917171]
18. Rey A. L'examen psychologique dans les cas d'encephalopathie traumatique. Archives de psychologie 1941;28:286-340.
19. Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. Percept Mot Skills 1958 Dec;8(7):271. [doi: 10.2466/pms.8.7.271-276]
20. Rouleau I, Salmon DP, Butters N, Kennedy C, McGuire K. Quantitative and qualitative analyses of clock drawings in Alzheimer's and Huntington's disease. Brain Cogn 1992 Jan;18(1):70-87. [doi: 10.1016/0278-2626(92)90112-y] [Medline: 1543577]
21. Sheridan LK, Fitzgerald HE, Adams KM, Nigg JT, Martel MM, Puttler LI, et al. Normative Symbol Digit Modalities Test performance in a community-based sample. Arch Clin Neuropsychol 2006 Jan;21(1):23-28. [doi: 10.1016/j.acn.2005.07.003] [Medline: 16139470]
22. Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. Lancet 2006 Apr;367(9518):1262-1270. [doi: 10.1016/s0140-6736(06)68542-5]
23. Cramér H. Mathematical Methods of Statistics. Uppsala, Sweden: Princeton University Press; 1999.
24. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends in Sports Science 2014;1:19-25.
25. Choi NG, Dinitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. J Med Internet Res 2013;15(5):e97 [FREE Full text] [doi: 10.2196/jmir.2333] [Medline: 23681083]
26. Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. N Engl J Med 2005 Jun 09;352(23):2379-2388. [doi: 10.1056/nejmoa050151]
27. Scholz J, Klein MC, Behrens TEJ, Johansen-Berg H. Training induces changes in white-matter architecture. Nat Neurosci 2009 Nov;12(11):1370-1371 [FREE Full text] [doi: 10.1038/nn.2412] [Medline: 19820707]
28. Osterhout L, Poliakov A, Inoue K, McLaughlin J, Valentine G, Pitkanen I, et al. Second-language learning and changes in the brain. J Neurolinguistics 2008 Nov;21(6):509-521 [FREE Full text] [doi: 10.1016/j.jneuroling.2008.01.001] [Medline: 19079740]
29. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the Internet may be changing our cognition. World Psychiatry 2019 Jul 06;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]

XSL•FO
RenderX

30. Salthouse TA. Aging and measures of processing speed. Biol Psychol 2000 Oct;54(1-3):35-54. [doi: 10.1016/s0301-0511(00)00052-1] [Medline: 11035219]

Abbreviations

ANCOVA: analysis of covariance
AVLT: Auditory Verbal Learning Test
BABRI: Beijing Aging Brain Rejuvenation
BNT: Boston Naming Test
CDT: Clock Drawing Test
DD: Digital Divide (participant group)
FACE: Facial Affect Coding System
GDS: Geriatric Depression Scale
LAS: Leisure Activity Scale
MCI: mild cognitive impairment
MLM: mixed linear model
MMSE: Mini-Mental State Examination
NC: normal cognition
ODD: Overcoming the Digital Divide (participant group)
OR: odds ratio
ROCF: Rey-Osterrieth Complex Figure test
R-Ocopy: Rey-Osterrieth Complex Figure-Copy test
SCWT: Stroop Color Word Test
SDMT: Symbol Digit Modification Test
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TMT: Trail Making Test
TMT_A: Trail Making Test Part A
Trans: Transition (participant group)
VFT: Verbal Fluency Test

Edited by A Mavragani; submitted 03.05.24; peer-reviewed by H Chen, J Zhang; comments to author 08.10.24; revised version received 11.10.24; accepted 28.10.24; published 27.11.24.

Please cite as:
Li Y, Liu C, Sun J, Zhang J, Li X, Zhang Z
The Digital Divide and Cognitive Disparities Among Older Adults: Community-Based Cohort Study in China
J Med Internet Res 2024;26:e59684
URL: https://www.jmir.org/2024/1/e59684
doi:10.2196/59684
PMID:

©Yumeng Li, Chen Liu, Jiaqing Sun, Junying Zhang, Xin Li, Zhanjun Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predicting Early Dropout in a Digital Tobacco Cessation Intervention: Replication and Extension Study

Linda Q Yu[1], PhD; Michael S Amato[1,2], PhD; George D Papandonatos[3], PhD; Sarah Cha[1], MSPH; Amanda L Graham[1,2,4], PhD

[1]Innovations Center, Truth Initiative, Washington, DC, United States

[2]Department of Medicine, Mayo Clinic College of Medicine and Science, Rochester, MN, United States

[3]Center for Statistical Sciences, School of Public Health, Brown University, Providence, RI, United States

[4]Department of Oncology, Georgetown University Medical Center, Washington, DC, United States

Corresponding Author:
Amanda L Graham, PhD
Innovations Center
Truth Initiative
900 G St NW Fourth Floor
Washington, DC, 20001
United States
Phone: 1 2024545938
Email: agraham@truthinitiative.org

Abstract

Background: Detecting early dropout from digital interventions is crucial for developing strategies to enhance user retention and improve health-related behavioral outcomes. Bricker and colleagues proposed a single metric that accurately predicted early dropout from 4 digital tobacco cessation interventions based on log-in data in the initial week after registration. Generalization of this method to additional interventions and modalities would strengthen confidence in the approach and facilitate additional research drawing on it to increase user retention.

Objective: This study had two research questions (RQ): RQ1—can the study by Bricker and colleagues be replicated using data from a large-scale observational, multimodal intervention to predict early dropout? and RQ2—can first-week engagement patterns identify users at the greatest risk for early dropout, to inform development of potential “rescue” interventions?

Methods: Data from web users were drawn from EX, a freely available, multimodal digital intervention for tobacco cessation (N=70,265). First-week engagement was operationalized as any website page views or SMS text message responses within 1 week after registration. Early dropout was defined as having no subsequent engagement after that initial week through 1 year. First, a multivariate regression model was used to predict early dropout. Model predictors were dichotomous measures of engagement in each of the initial 6 days (days 2-7) following registration (day 1). Next, 6 univariate regression models were compared in terms of their discrimination ability to predict early dropout. The sole predictor of each model was a dichotomous measure of whether users had reengaged with the intervention by a particular day of the first week (calculated separately for each of 2-7 days).

Results: For RQ1, the area under the receiver operating characteristic curve (AUC) of the multivariate model in predicting dropout after 1 week was 0.72 (95% CI 0.71-0.73), which was within the range of AUC metrics found in the study by Bricker and colleagues. For RQ2, the AUCs of the univariate models increased with each successive day until day 4 (0.66, 95% CI 0.65-0.67). The sensitivity of the models decreased (range 0.79-0.59) and the specificity increased (range 0.48-0.73) with each successive day.

Conclusions: This study provides independent validation of the use of first-week engagement to predict early dropout, demonstrating that the method generalizes across intervention modalities and engagement metrics. As digital intervention researchers continue to address the challenges of low engagement and early dropout, these results suggest that first-week engagement is a useful construct with predictive validity that is robust across interventions and definitions. Future research should explore the applicability and efficiency of this model to develop interventions to increase retention and improve health behavioral outcomes.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e54248) doi:10.2196/54248

KEYWORDS

digital interventions; attrition; engagement; dropout; tobacco; smoking; cessation; mobile health; internet; mobile phone

Introduction

A well-established challenge for digital health behavior change interventions is that many users disengage before achieving a meaningful dose of treatment necessary for desired outcomes [1-5]. A recent paper by Bricker et al [6] aimed to identify early markers of dropout that could generalize across platforms and aid the design of rescue interventions to mitigate dropout. Using the data from 2 clinical trials of 4 web- or app-based tobacco cessation platforms, they compared a variety of classification models to predict early dropout. They considered predictors including baseline and demographic variables, as well as daily log-in data from the first 7 days after registration. The best-performing model that predicted whether a user dropped out after the initial week, across all platforms, was a logistic regression model using daily log-in count data from the first 7 days. This method has the advantage of being straightforward and easy to implement for researchers and intervention designers. The finding that baseline and demographic variables did not contribute to greater predictive accuracy means the same approach can be used for both robust clinical trial datasets and observational datasets that may contain less information. Bricker et al [6] noted that it remains to be seen whether the high predictive power of this model can translate to other platforms and study designs like real-world datasets, which are much more heterogeneous compared to clinical trial data.

This study aimed to conceptually replicate and extend Bricker et al [6] by applying the best-performing model from their analyses to a large, real-world dataset from a tobacco cessation intervention that includes both web and SMS text messaging components. While Bricker et al [6] conceived of the log-in measure as a single, unifying measure across digital platforms, it is exclusively web or app based and does not account for the multimodal nature of many tobacco cessation interventions. In particular, SMS text messaging interventions have been shown to be effective at increasing abstinence outcomes in smoking cessation [7-9] and are now commonly deployed alongside or integrated with web interventions. The log-in measure may also not be practical for web-based platforms in some cases. Given these technical considerations, this study aimed to extend the work of Bricker et al [6] to a multimodal intervention with a broader set of engagement measures beyond log-in data.

The value of predicting early dropout is the opportunity to prevent it. To guide such targeted intervention efforts, we sought to characterize the time course of early engagement and its relationship with early dropout. In particular, we measured how rapidly the proportion of users who remain engaged decreases after registration and then compared the classification performance of early dropout models using engagement data at different time points. Performance metrics of those models can inform decisions about the most efficient timing to intervene for users at risk of early dropout. Such decisions are not universally conclusive but rather require consideration of trade-offs between the costs of a potential rescue intervention and the pool of potential recipients. For example, it might be more efficient to deploy a low-cost notification to a larger group of users with lower average risk, whereas a more resource-intensive appeal might be more efficiently deployed to a smaller group of users that are at high risk of dropping out.

To examine the practical implications of predicting and preventing early dropout, we leveraged a sizable, multiyear dataset from a real-world digital tobacco cessation intervention that incorporates web and SMS text messaging components. The research questions (RQs) addressed by this study were twofold: RQ1—can the model in Bricker et al [6] be replicated in a large-scale, observational, and multimodal dataset to predict early dropout? and RQ2—can patterns of engagement in the first week be used to identify which users are at the greatest risk of early dropout, for the development of potential “rescue” interventions? By answering this pair of RQs, this study aimed to demonstrate a simple, generalizable approach that can be used to address the persistent issue of early dropout across a broad range of digital interventions.

Methods

Study Design

This study uses observational analysis of data from EX, a free, multimodal digital tobacco cessation intervention developed by Truth Initiative in collaboration with the Mayo Clinic. Results are reported according to STROBE (Strengthening the Reporting of Observational studies in Epidemiology) recommendations [10].

Setting

EX is a multimodal, evidence-based tobacco cessation program designed around the US Public Health Service tobacco dependence treatment clinical practice guidelines [11], Social Cognitive Theory [12], and the Mayo Clinic model for engaging tobacco users in cessation treatment [13]. As a real-world intervention, EX has evolved throughout the study period, although the theoretical approach and behavior change techniques have remained consistent. EX educates tobacco users about the behavioral and physiological aspects of nicotine addiction and helps them build coping skills and strategies for quitting through individually tailored content, hands-on exercises, and videos. It also includes the longest-running, internet-based social network dedicated to tobacco cessation where users can connect with current and former tobacco users for real-time support (EX Community).

The website is fully integrated with a bidirectional and dynamically tailored SMS text message program that users can opt into during website registration. The 12-week SMS text message program is tailored around a number of characteristics, including a user’s quit date, their primary tobacco product, and patterns of engagement with the website. Content includes tips

XSL•FO
RenderX

and encouragement, quotes excerpted from the EX Community, links to the EX website, as well as prompts for the user to text back a quit date, their motivations for quitting, the name of a supportive friend or family member, and on-demand keywords for further assistance. Email support is also tailored to each individual, designed to drive ongoing engagement.

Roughly 3000 tobacco users register on EX each month. Most users come through Google ads or organic searches and are motivated to quit or actively engaged in the quitting process. A nominal advertising spend has been consistent over the past decade.

Ethical Considerations

This study was reviewed and determined to be exempt by the Advarra institutional review board (Pro00082535). To register on EX, individuals must agree to its terms of use and privacy policy, which grants permission for their data to be used in observational research to advance the science of nicotine dependence treatment. As such, no additional screening information was obtained, and no separate informed consent was solicited. Users were not compensated for this research. All study data were deidentified prior to analysis. Truth Initiative has stringent privacy and security policies, which include a comprehensive privacy policy, strong encryption used in the transmission of all health-related information, electronic access controls on all personally identifiable information, and industry-standard network protection and intrusion detection.

Participants

We examined users who registered between September 1, 2016, and April 13, 2022 (N=143,121). In this study, adult tobacco users aged 18 years or older who opted into the SMS text message program for whom website page view data were available (unavailable for ~2.5% of users who likely used do-not-track software) were included in the analyses. These criteria yielded 70,265 participants.

Data Sources

Baseline Characteristics

Available measures from website registration used to characterize the study sample included gender, age, tobacco use frequency (every day, some days, and not at all), and zip code (mapped to Rural Area Continuum Codes per Amato and Graham [14]).

Engagement Data

User activity was tracked for 1 year after registration. Website page views were recorded with Adobe Analytics, and SMS text message responses were recorded in EX's custom SMS text messaging application database.

The engagement was operationalized as any website page views or SMS text message responses. Early dropout was defined as having no engagement after day 7 after registration. Under this definition, 38.6% (27,118/70,265) of users were categorized as dropouts. For comparison, the proportion of early dropouts ranged between 49.27% and 65.32% in the interventions analyzed by Bricker et al [6].

Statistical Analyses

First, descriptive statistics were used to determine the risk of early dropout and to characterize the sample using data drawn from website registration. To address RQ1, daily engagement on days 2-7 (1 binary variable for each day) was examined as a predictor of early dropout in a modified Poisson regression model with a logarithmic link and robust error variance [15]. All 6 engagement variables were entered into the model simultaneously. A value of 1 was assigned to each day if the engagement count was positive and 0 otherwise. Of note, all study participants had at least 1 page view on the day they registered on the website. Therefore, day 1 in this model is confounded with the intercept and represents the risk of early dropout among study participants who failed to reengage with the intervention in the first week after their initial enrollment. Regression coefficients for days 2-7 represent the relative risk of early dropout among participants engaging with the study on a particular day relative to those who failed to reengage in the first week after enrollment, controlling for their engagement pattern on the remaining days of that first week. Our statistical model thus provides a modified replication of that used by Bricker et al [6], using a dichotomized, multimodal definition of engagement for the predictors.

To address RQ2, we used modified Poisson regression with robust SEs to examine 6 distinct patterns of engagement as predictors of early dropout, with the goal of informing potential "rescue" interventions. In each univariate regression model, the sole predictor of dropout was "return engagement," a dichotomous measure of whether someone had returned to the website by a particular day of the first week (days 2-7). For example, an individual who reengaged on day 3 but then disengaged from the intervention again from days 4 to 7 would have a positive value starting from the day 3 model through day 7, because the models consider such a person a "re-engager by Day 3." Running a series of day-specific models against each other allowed us to obtain accuracy and performance metrics for the effect of delaying a hypothetical rescue intervention by each additional day and to compare the models against each other. By definition, the intercept in these models corresponds to subjects that were coded as 0 on that particular day, that is, those failing to reengage by that day of the week.

Both models used 80% (56,238/70,265) of the data for training purposes and 20% (14,027/70,265) of the data to test performance, with a single 80:20 split stratified by month to evenly represent users from across the study period. Predictive performance metrics included area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value, and Brier score (mean squared prediction error) as calculated on the training set and evaluated on the test set. For comparison under different intervention cost trade-off scenarios, metrics were evaluated under classification cutoffs of 0.5 and 0.3. AUCs were compared against each other with the *roc.test* function in the R library *pROC* [16].

XSL•FO
RenderX

Results

Participants

The analytic sample included 70,265 participants. The baseline characteristics of the sample are presented in Table 1.

Table 1. Baseline characteristics of the analytic sample (N=70,625).

	Values, n (%)
Age group (years)	
18-24	8060 (11.5)
25-30	15,650 (22.3)
31-44	24,611 (35.0)
45-64	19,736 (28.1)
Older than 65	2208 (3.1)
Sex	
Male	21,695 (30.9)
Female	48,084 (68.4)
Other or not reported	486 (0.7)
Tobacco use	
Every day	68,617 (97.7)
Some days	1586 (2.2)
Metro areas	
Large metro	29,154 (41.5)
Small metro	23,442 (33.3)
Rural	12,405 (17.6)

Patterns of Engagement

Engagement rates by day during the first week after registration are shown in Table 2. The daily engagement metrics (dichotomized as any engagement vs none) are relevant to RQ1, whereas the cumulative reengagement metrics are relevant to RQ2. The latter declined more rapidly than the former, with daily reengagement rates diminishing to single digits by day 4. Overall, 36.7% (27,174/70,265) of EX users did not reengage with the intervention during the initial week after registration. The first day of return engagement was most commonly day 2 (26,796/70,265, 38.1%) followed by a smaller proportion on day 3 (9613/70,265, 13.7%). Day 4 and onward saw single-digit percentages of those returning for the first time following registration. The cumulative return numbers on each day reflect the sum of the first day of return numbers over previous days.

Table 2. Number and proportion of engaged users in the initial week after registration under different definitions of daily intervention engagement (N=70,265).

	Day 1[a], n (%)	Day 2, n (%)	Day 3, n (%)	Day 4, n (%)	Day 5, n (%)	Day 6, n (%)	Day 7, n (%)
Daily engagement	70,265 (100)	26,796 (38.1)	20,838 (29.7)	15,947 (22.7)	12,138 (17.3)	10,844 (15.4)	10,073 (14.3)
First day of return engagement	N/A[b]	26,796 (38.1)	9613 (13.7)	3146 (4.5)	1498 (2.1)	1154 (1.6)	884 (1.3)
Cumulative return engagement	N/A	26,796 (38.1)	36,409 (51.8)	39,555 (56.3)	41,053 (58.4)	42,207 (60.1)	43,091 (61.3)

[a]Day 1=program enrollment (all users engage by definition).

[b]N/A: not applicable.

RQ1: Can the Model in Bricker et al [6] be Replicated in a Large-Scale Observational, Multimodal Dataset to Predict Early Dropout?

An examination of the risk ratios for early dropout based on a daily engagement model showed that those who failed to reengage with the intervention during the first week after registration had a 60% risk of not reengaging with the intervention at any further point during the study period (ie, “dropped out”; see Table 3). For those who returned for even a single day during that week, the early dropout risk was reduced monotonically with time, with the magnitude of reduction ranging from 24% (relative risk=0.76, 95% CI 0.74-0.78) for those who returned on day 3 to 55% (relative risk=0.45, 95% CI 0.42-0.57) for those who returned on day 7.

XSL•FO
RenderX

Table 3. Relative risk (95% CI) of early dropout under a daily engagement model.

Predictor	Relative risk coefficient (95% CI)
Intercept[a]	0.60 (0.59-0.61)
Day 2	0.60 (0.59-0.62)
Day 3	0.76 (0.74-0.78)
Day 4	0.59 (0.57-0.61)
Day 5	0.64 (0.61-0.67)
Day 6	0.57 (0.54-0.60)
Day 7	0.45 (0.42-0.57)

[a]The intercept corresponds to the risk of early dropout for study participants who failed to reengage with the intervention on days 2-7 of the initial week.

The predictive performance metrics for the daily engagement model were evaluated in an unbiased manner in our 20% test set and are outlined in Table 4. The discrimination ability of the model (AUC=0.72, 95% CI 0.71-0.73) considerably exceeded the 0.50 value expected by chance and was comparable to that of similar models reported in Bricker et al [6] (AUC range=0.60-0.94). As expected, the sensitivity and negative predictive value of the model decreased, while the specificity and positive predictive value increased, when the classification threshold was changed from 0.3 to 0.5. The Brier score of the model was 0.20, below the 0.25 value expected by chance, indicating adequate calibration.

Table 4. Test set performance metrics of first-week daily engagement model in predicting early dropout (no engagement after day 7).

Model performance metric	Classification threshold=0.30	Classification threshold=0.50
Sensitivity	0.86	0.59
Specificity	0.49	0.73
PPV[a]	0.50	0.57
NPV[b]	0.84	0.74
True positive, n	4566	3123
True negative, n	4195	6342
False positive, n	4495	2348
False negative, n	771	2214

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

RQ2: Can Patterns of First-Week Engagement be Used to Identify Which Users are at the Greatest Risk of Early Dropout to Inform Potential "Rescue" Interventions?

The 6 cumulative return engagement models examined the effect of delayed return by each additional day in the first week on dropout. A comparison of the model AUCs shows that the prediction accuracy increases until day 4 (AUC=0.66), and levels off thereafter (Table 5). Two other prevalence-free model performance metrics (calculated under a classification probability of 0.3) give a more detailed picture—as the number of days that it takes a user to reengage with the intervention increases, specificity increases but sensitivity decreases. In other words, with each successive day we wait until making a judgment call about the likelihood of early dropout, there is a reduction in the number of false positives (ie, users identified as likely to drop out early who do not actually do so), at the cost of an increase in the number of false negatives (ie, users identified as likely to remain engaged with the intervention who actually drop out early).

XSL•FO
RenderX

Table 5. Test set performance comparisons of the 6 cumulative return models in predicting early dropout (no engagement after day 7) at classification threshold=0.3.

	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
AUC[a] test (95% CI)	0.63 (0.62-0.64)	0.65 (0.64-0.66)	0.66 (0.65-0.67)	0.66 (0.65-0.67)	0.66 (0.65-0.67)	0.66 (0.65-0.67)
Sensitivity	0.79	0.67	0.64	0.62	0.60	0.59
Specificity	0.48	0.63	0.68	0.70	0.72	0.73
PPV[b]	0.48	0.53	0.55	0.56	0.57	0.57
NPV[c]	0.78	0.76	0.75	0.75	0.75	0.74
True positive, n	4197	3602	3408	3292	3204	3123
True negative, n	4135	5481	5941	6117	6249	6342
False positive, n	4555	3209	2749	2573	2441	2348
False negative, n	1140	1735	2045	2601	2133	2214

[a]AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

Discussion

Principal Findings

This study examined the relationship between first-week engagement and early dropout in a large observational dataset from a real-world digital tobacco cessation program. We drew on the methods in Bricker et al [6] to address two RQs: RQ1—does first-week engagement predict early dropout in a large-scale, observational dataset of a multimodal (web and SMS text message) digital tobacco cessation intervention? and RQ2—can first-week engagement be used to inform rescue interventions to prevent dropout? With regard to RQ1, we found that a dichotomized measure of engagement that encompassed web and SMS text message use predicted early dropout with acceptable accuracy [17]. The ability of first-week engagement to predict early dropout was considerably above chance (AUC=0.72) and within the range of AUC values described for the interventions in Bricker et al [6], providing independent validation of their method. With regard to RQ2, reengagement with the intervention any day in the first-week after registration predicted a lower likelihood of early dropout than no reengagement at all, with model predictive performance metrics suggesting that different parts of the week (days 2-3 vs days 4-7) traded off in terms of accurately identifying the users who would drop out.

There were several methodological differences between this study and Bricker et al [6]. First, this study used an observational dataset drawn from a real-world intervention over a 5.5-year time period instead of study participants in a time-limited randomized controlled trial. These characteristics strengthen the external validity of our study and may result in greater heterogeneity in participant characteristics, behavior, and the intervention itself as it evolved throughout the years. Another difference is that 1-week dropout outcomes in this study were defined as users who had no further activity (across both web and SMS text message) after this time period, whereas the previous paper had used groups defined by functional clustering methods that allowed some activity after the first week [18]. Our definition is relatively simpler to implement without the need for sophisticated methods and yielded comparable model performance. Finally, EX is a web and SMS text message intervention, which may differ in reengagement and notification strategies from the app- and web-based platforms examined in Bricker et al [6]. Taken together, the models in this study performed with satisfactory accuracy and within the range of the model performances in Bricker et al [6], suggesting that this method is generalizable across platforms, populations, and measures of engagement.

A salient question during the design of a potential rescue intervention to curb dropout is when to intervene. Ideally, one would like to identify users at risk of dropping out as early as possible. However, it is important to consider the trade-offs, taking into consideration the performance of each additional day in the week to predict dropouts with the cost of a possible intervention. For example, if an intervention is a very low cost (eg, sending out an email or push notification), then a model that includes the days with higher sensitivity is desirable in order to target anyone identified as a predicted dropout. A greater number of false positives (ie, users identified as early dropouts who did not do so) is acceptable since the downsides are few, although there may be opportunity costs to delivering a rescue intervention versus other treatment content. However, if the intervention is more costly (eg, proactive human outreach), then a model including the days with higher specificity (lower false positives) would be more suitable. A holistic approach balancing the trade-offs, alongside factors such as user tolerance for notifications, can inform intervention development. As the model comparison we outlined here is easily implementable and generalizable, researchers using other types of digital interventions can apply it to their platform to characterize their own patterns of user reengagement.

Strengths and Limitations

The strengths of this study included the large sample size of users over multiple years from a freely available intervention

XSL•FO
RenderX

and the 1 year of follow-up data. We used rigorous methods for classification, including 80:20 performance testing and 2 classification thresholds. A limitation of the study is the limited amount of demographic data, although Bricker et al [6] showed that the inclusion of baseline characteristics did not improve model prediction performance. The AUC in the first-week daily engagement model was lower than the best model predictions for 3 of the 4 interventions in the original study (0.94-0.85), but higher than 1 (0.60). The lower performance metrics in this study are likely due to the more heterogenous sample in our study compared to clinical trial datasets but were still within an acceptable range. For the cumulative return models, the performance was notably lower (AUCs=0.63-0.66). These models should be interpreted with caution—performance metrics, such as sensitivity and specificity, should be taken into account with other considerations when designing a rescue intervention, as discussed.

Future Directions

An important next step in this research is to examine factors that may mediate the relationship between continued engagement and health outcomes. Bricker et al [18] showed that users who continued to engage for a month or more were more likely to quit smoking. However, self-selection to continue engagement in that study and other observational studies challenges causal attribution. Future research using alternative designs should further investigate the causal impact of continued engagement on abstinence accounting for potentially confounding factors, such as motivation or confidence.

Previous studies, including those based on the EX intervention [8], have demonstrated the effectiveness of sending tailored reminders in increasing engagement [19-21]. In general, rescue interventions should be in the service of increasing the user's intrinsic motivation and self-efficacy, which have been shown to be associated with retention [22-25]. For example, a reminder notification could point the user toward community support or information about quitting medications [26-29]. Any single strategy likely is not sufficient. People who use digital interventions for behavior change drop out for many reasons [30], not the least of which is limited time and competing priorities from other aspects of their lives [5,31]. Thus, the intervention must be seen as useful, relevant, and important in furthering their goals for users to make accessing the intervention a priority [32].

Conclusions

In conclusion, we independently validated the method outlined in Bricker et al [6] using first-week engagement to predict dropout in a large, observational dataset from a multimodal digital intervention for tobacco cessation. Additionally, a model based on the rate of cumulative return engagement over the first week can serve as a promising and easily implementable method for comparing and developing interventions to prevent dropout as early as practically possible.

Acknowledgments

The study was funded by Truth Initiative. ALG is supported by the US National Cancer Institute (R01CA275521; principal investigator [PI]: Jonathan Shuter), the US National Institute on Drug Abuse (UG3DA057032, PI: Jaimee Heffner; 1K08DA058058-01A1, PI: Brian Williams), the US National Institute of Mental Health (R34MH120142; multiple PIs: Melanie Bennett, Faith Dickerson, and Amanda Graham), and St. Jude or Washington University St. Louis Implementation Sciences Collaborative (multiple PIs: Rachel Tillery or Bernard Fuemmeler). The funders had no role in the study design, collection, analysis or interpretation of the data, writing the paper, or the decision to submit the paper for publication.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request with a data sharing agreement with Truth Initiative.

Authors' Contributions

All authors were involved in conceptualizing the study, drafting the paper, and interpreting the results. LQY, MSA, and GDP designed the analyses. LQY conducted the analyses.

Conflicts of Interest

LQY, MSA, SC, and ALG are employees of Truth Initiative, a nonprofit public health foundation that sells digital tobacco cessation programs to support its mission-driven work.

References

1. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022;24(4):e35120 [FREE Full text] [doi: 10.2196/35120] [Medline: 35471414]
2. Couper MP, Alexander GL, Zhang N, Little RJA, Maddy N, Nowak MA, et al. Engagement and retention: measuring breadth and depth of participant use of an online intervention. J Med Internet Res 2010;12(4):e52 [FREE Full text] [doi: 10.2196/jmir.1430] [Medline: 21087922]

XSL•FO
RenderX

3. Kelders SM, Bohlmeijer ET, van Gemert-Pijnen JE. Participants, usage, and use patterns of a web-based intervention for the prevention of depression within a randomized controlled trial. J Med Internet Res 2013;15(8):e172 [FREE Full text] [doi: 10.2196/jmir.2258] [Medline: 23963284]
4. Krebs P, Sherman SE, Wilson H, El-Shahawy O, Abroms LL, Zhao X, et al. Text2Connect: a health system approach to engage tobacco users in quitline cessation services via text messaging. Transl Behav Med 2020;10(1):292-301. [doi: 10.1093/tbm/ibz033] [Medline: 32011721]
5. Postel MG, de Haan HA, ter Huurne ED, van der Palen J, Becker ES, de Jong CAJ. Attrition in web-based treatment for problem drinkers. J Med Internet Res 2011;13(4):e117 [FREE Full text] [doi: 10.2196/jmir.1811] [Medline: 22201703]
6. Bricker J, Miao Z, Mull K, Santiago-Torres M, Vock DM. Can a single variable predict early dropout from digital health interventions? Comparison of predictive models from two large randomized trials. J Med Internet Res 2023;25:e43629 [FREE Full text] [doi: 10.2196/43629] [Medline: 36662550]
7. Fang YE, Zhang Z, Wang R, Yang B, Chen C, Nisa C, et al. Effectiveness of eHealth smoking cessation interventions: systematic review and meta-analysis. J Med Internet Res 2023;25:e45111 [FREE Full text] [doi: 10.2196/45111] [Medline: 37505802]
8. Graham AL, Papandonatos GD, Cha S, Amato MS, Jacobs MA, Cohn AM, et al. Effectiveness of an optimized text message and internet intervention for smoking cessation: a randomized controlled trial. Addiction 2022;117(4):1035-1046 [FREE Full text] [doi: 10.1111/add.15677] [Medline: 34472676]
9. Zbikowski SM, Hapgood J, Smucker Barnwell S, McAfee T. Phone and web-based tobacco cessation treatment: real-world utilization patterns and outcomes for 11,000 tobacco users. J Med Internet Res 2008;10(5):e41 [FREE Full text] [doi: 10.2196/jmir.999] [Medline: 19017583]
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
11. Fiore M. Treating Tobacco Use and Dependence: 2008 Update: Clinical Practice Guideline. Collingdale, PA: DIANE Publishing; 2009.
12. Bandura A. Social foundations of thought and action?: a social cognitive theory. CiNii. 1986. URL: https://cir.nii.ac.jp/crid/1570854174839355904 [accessed 2023-08-09]
13. Wellness coach training course details and application. Mayo Clinic. 2023. URL: https://wellnesscoachtraining.mayo.edu/why-us/ [accessed 2023-08-09]
14. Amato MS, Graham AL. Geographic representativeness of a web-based smoking cessation intervention: reach equity analysis. J Med Internet Res 2018;20(10):e11668 [FREE Full text] [doi: 10.2196/11668] [Medline: 30355557]
15. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004;159(7):702-706. [doi: 10.1093/aje/kwh090] [Medline: 15033648]
16. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
17. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
18. Bricker JB, Mull KE, Santiago-Torres M, Miao Z, Perski O, Di C. Smoking cessation smartphone app use over time: predicting 12-month cessation outcomes in a 2-arm randomized trial. J Med Internet Res 2022;24(8):e39208 [FREE Full text] [doi: 10.2196/39208] [Medline: 35831180]
19. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of technology-based strategies to promote engagement with digital interventions: a systematic review protocol. JMIR Res Protoc 2015;4(2):e47 [FREE Full text] [doi: 10.2196/resprot.3990] [Medline: 25921274]
20. Brouwer W, Kroeze W, Crutzen R, de Nooijer J, de Vries NK, Brug J, et al. Which intervention characteristics are related to more exposure to internet-delivered healthy lifestyle promotion interventions? A systematic review. J Med Internet Res 2011;13(1):e2 [FREE Full text] [doi: 10.2196/jmir.1639] [Medline: 21212045]
21. Crutzen R, de Nooijer J, Brouwer W, Oenema A, Brug J, de Vries NK. Strategies to facilitate exposure to internet-delivered health behavior change interventions aimed at adolescents or young adults: a systematic review. Health Educ Behav 2011;38(1):49-62. [doi: 10.1177/1090198110372878] [Medline: 21189422]
22. Coa K, Patrick H. Baseline motivation type as a predictor of dropout in a healthy eating text messaging program. JMIR Mhealth Uhealth 2016;4(3):e114 [FREE Full text] [doi: 10.2196/mhealth.5992] [Medline: 27688034]
23. Danaher BG, Smolkowski K, Seeley JR, Severson HH. Mediators of a successful web-based smokeless tobacco cessation program. Addiction 2008;103(10):1706-1712 [FREE Full text] [doi: 10.1111/j.1360-0443.2008.02295.x] [Medline: 18715238]
24. Strecher VJ, McClure J, Alexander G, Chakraborty B, Nair V, Konkel J, et al. The role of engagement in a tailored web-based smoking cessation program: randomized controlled trial. J Med Internet Res 2008;10(5):e36 [FREE Full text] [doi: 10.2196/jmir.1002] [Medline: 18984557]

XSL•FO
RenderX

25. Wojtowicz M, Day V, McGrath PJ. Predictors of participant retention in a guided online self-help program for university students: prospective cohort study. J Med Internet Res 2013;15(5):e96 [FREE Full text] [doi: 10.2196/jmir.2323] [Medline: 23697614]
26. Graham AL, Papandonatos GD, Cha S, Erar B, Amato MS, Cobb NK, et al. Improving adherence to smoking cessation treatment: intervention effects in a web-based randomized trial. Nicotine Tob Res 2017;19(3):324-332 [FREE Full text] [doi: 10.1093/ntr/ntw282] [Medline: 27836982]
27. Poirier J, Cobb NK. Social influence as a driver of engagement in a web-based health intervention. J Med Internet Res 2012;14(1):e36 [FREE Full text] [doi: 10.2196/jmir.1957] [Medline: 22356829]
28. Richardson CR, Buis LR, Janney AW, Goodrich DE, Sen A, Hess ML, et al. An online community improves adherence in an internet-mediated walking program. Part 1: results of a randomized controlled trial. J Med Internet Res 2010;12(4):e71 [FREE Full text] [doi: 10.2196/jmir.1338] [Medline: 21169160]
29. Simon GE, Ludman EJ, Goodale LC, Dykstra DM, Stone E, Cutsogeorge D, et al. An online recovery plan program: can peer coaching increase participation? Psychiatr Serv 2011;62(6):666-669 [FREE Full text] [doi: 10.1176/ps.62.6.pss6206_0666] [Medline: 21632737]
30. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
31. Donkin L, Glozier N. Motivators and motivations to persist with online psychological interventions: a qualitative study of treatment completers. J Med Internet Res 2012;14(3):e91 [FREE Full text] [doi: 10.2196/jmir.2100] [Medline: 22743581]
32. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]

Abbreviations

AUC: area under the receiver operating characteristic curve
RQ: research question
STROBE: Strengthening the Reporting of Observational studies in Epidemiology

Edited by A Mavragani; submitted 06.11.23; peer-reviewed by J Bricker, Y Fang; comments to author 30.12.23; revised version received 15.04.24; accepted 14.10.24; published 27.11.24.

Please cite as:
Yu LQ, Amato MS, Papandonatos GD, Cha S, Graham AL
Predicting Early Dropout in a Digital Tobacco Cessation Intervention: Replication and Extension Study
J Med Internet Res 2024;26:e54248
URL: https://www.jmir.org/2024/1/e54248
doi:10.2196/54248
PMID:

©Linda Q Yu, Michael S Amato, George D Papandonatos, Sarah Cha, Amanda L Graham. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Internet-Delivered Psychoeducation (SCOPE) for Transition-Aged Autistic Youth: Pragmatic Randomized Controlled Trial

Anna Backman[1], PhD; Lise Roll-Pettersson[2], PhD; Are Mellblom[3], BSc; Elisabet Norman-Claesson[3], BSc; Emma Sundqvist[3], MSc; Eric Zander[4,5], PhD; Sarah Vigerland[1,5], PhD; Tatja Hirvikoski[3,4,5], PhD

[1]Child and Adolescent Psychiatry, Stockholm Healthcare Services, Region Stockholm, Stockholm, Sweden

[2]Department of Special Education, Stockholm University, Stockholm, Sweden

[3]Habilitation & Health, Stockholm Healthcare Services, Region Stockholm, Stockholm, Sweden

[4]Karolinska Institutet Center of Neurodevelopmental Disorders, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

[5]Centre for Psychiatry Research, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Anna Backman, PhD
Child and Adolescent Psychiatry
Stockholm Healthcare Services
Region Stockholm
Stockholm
Sweden
Phone: 46 732445857
Email: anna.n.backman@hotmail.com

Abstract

Background: Psychoeducation is a recommended first-line intervention for transition-aged autistic youth, but it has not been previously evaluated in an internet-delivered format. SCOPE (Spectrum Computerized Psychoeducation) is an 8-week individual, internet-delivered, therapist-supported psychoeducative intervention.

Objective: This study aimed to investigate the effectiveness of SCOPE through a 3-armed randomized controlled trial. The intervention aims to increase participants' understanding of autism and, in doing so, increase their quality of life (QoL).

Methods: SCOPE was codeveloped with clinicians and autistic young adults. It contains 8 autism-related modules, each with (1) text describing the module topic, (2) four video vignettes with recurring characters who describe their lives and perspectives on the module topic, (3) a list of neurotypical characteristics related to the module's topic, and (4) self-reflection using 3 or 4 questions about the module topic, answered by multiple-choice bullets and voluntary open-ended written comments. Participants were randomized (2:1:1) to SCOPE, an active control (web-based self-study), or treatment as usual (TAU). The primary outcome was participants' autism knowledge, assessed using the Autism Spectrum Disorder Quiz, and secondary outcomes included acceptance of diagnosis, QoL, and symptoms of mental health problems. All outcomes were assessed at the baseline, postintervention, and 3-month follow-up time points, using mixed-effects models to assess change in outcome measures across time points.

Results: Between 2014 and 2020, a total of 141 participants were randomized to 1 of the 3 treatment arms. The SCOPE participants had significantly greater autism knowledge gains at the posttreatment time point compared to TAU participants with a moderate effect size (d=0.47; P=.05); gains were maintained at the 3-month follow-up (d=0.46; P=.05). The self-study participants also had increased knowledge gains compared to TAU participants at the posttreatment time point with a moderate effect size (d=0.60; P=.03) but did not maintain these gains at the 3-month follow-up, and their autism knowledge scores returned to baseline (mean change score: –0.13, 95% CI –1.20 to 0.94; P=.81). In addition, SCOPE participants reported improved QoL at the postintervention (d=0.37, P=.02) and 3-month follow-up time points (d=0.60; P=.001), compared to the combined controls. The gained autism knowledge was not mirrored by changes in symptoms of anxiety or depression.

Conclusions: Effective internet-delivered interventions may facilitate first-line service access to individuals who are unable or unwilling to use traditional health care interventions or who live in geographically remote locations. Additionally, an intervention such as SCOPE could impart and sustain the knowledge gained through psychoeducation in transition-aged autistic youth. For

XSL•FO
RenderX

future research, qualitative studies could further our understanding of the lived experiences of intervention participation and outcomes after internet-delivered psychoeducation.

Trial Registration: ClinicalTrials.gov NCT03665363; https://clinicaltrials.gov/study/NCT03665363

(J Med Internet Res 2024;26:e49305) doi:10.2196/49305

KEYWORDS

autism; internet based; young adult; intervention; digital communication; life satisfaction; codeveloped; ASD; autism spectrum disorder; autistic; RCT; randomized controlled trial; randomized; psychoeducation; patient education

Introduction

Background

The transition to adulthood is challenging for many autistic individuals [1-4], with the transitioning period often coinciding with a loss of services, sometimes referred to as "dropping off a cliff," rendering autistic individuals and their families without services [5,6]. To achieve the milestones of early adulthood as a transition-aged autistic youth necessitates skills for independence, which are reliant on self-knowledge about their own strengths and difficulties [1,7,8]. Psychoeducation is a well-established, evidence-based intervention for many psychiatric disorders, typically used to increase knowledge about their health condition, self-knowledge, and strategies in everyday life and to improve quality of life (QoL) [9,10]. Furthermore, psychoeducation aims to empower patients and actively engage them in their treatments as well as further service usage [10-12].

Indeed, psychoeducation is the recommended first-line intervention after a confirmed autism diagnosis [13-15]. However, the research field of psychoeducation for autistic individuals is still new and growing. In clinical settings, stepped-care models are recommended [14,16], by first providing a low-intensity intervention to most patients and later moving on to more intensive interventions if needed. In clinical practice in the United Kingdom [17] and in Sweden [13], the first step often entails studying psychoeducative materials, for example, reading an informative website or a pamphlet. The effects of self-study of informative websites about autism have not been evaluated; therefore, it is unknown whether it is an acceptable and effective intervention for autistic individuals. In addition, only a limited number of previous studies have described structured psychoeducational interventions for autistic individuals. In 2015, Gordon et al [18] conducted a randomized controlled trial (RCT) for autistic adolescents and their parents, with low attrition and high satisfaction among participants. Hidalgo et al [19] conducted an open trial for autistic adults and their close relations using group-based psychoeducational intervention, and Crane et al [20] used autistic-led peer groups; both were able to show low attrition as well as high satisfaction with program contents. However, in an open trial of a psychoeducational intervention for older autistic adults, Lenders et al [21] found a high level of attrition. Although all the abovementioned trials reported preliminary positive results of improved autism knowledge [18,19,21], QoL [19], and self-awareness [18], some studies had limitations such as small sample sizes [20,21] or no control group [19,21]. To further corroborate the evidence base of psychoeducation for autistic individuals, there is still a need for studies with rigorous trial designs, especially targeting older adolescents and emerging adults.

Psychoeducation using internet delivery could possibly contribute to a broader selection of intervention modalities and may also facilitate cognitive and communicative accessibility according to individual needs and preferences [22,23]. For example, autistic individuals explicitly ask for digital communication to improve comprehension, feel in control of the communication, and be able to express themselves freely and with humor [24,25]. Furthermore, recent studies of internet-delivered interventions aiming to promote health or behavioral changes for autistic adults showed positive preliminary results [26,27], and a continued use of technology-based interventions has been recommended in a review of technology-based interventions for autistic individuals [28]. Regarding internet-delivered psychoeducational interventions specifically, only 1 uncontrolled feasibility study has been published, showing good feasibility and acceptability for the internet-delivered psychoeducation SCOPE (Spectrum Computerized Psychoeducation) for transition-aged autistic youth [29].

In conjunction with examining the effectiveness of internet-delivered psychoeducation, it is necessary to consider the possible adverse effects of psychological interventions [30,31]. The tentative adverse effects of psychoeducation include lowered self-confidence and increased self-stigma, which are associated with reduced mental well-being [18,32]. Therefore, it is necessary to examine potentially harmful effects in terms of adverse changes in symptoms of anxiety, depression, and reduced self-acceptance following psychoeducation.

In summary, combining the asked-for, internet-based delivery with the recommended first-line psychoeducational intervention could potentially provide autism knowledge and empower transition-aged autistic youth. Therefore, promising interventions showing high acceptance and feasibility [29] need further evaluation in an RCT.

Objectives

The objectives were to evaluate the effectiveness of SCOPE—a therapist-supported, internet-delivered psychoeducative program—using a 3-armed RCT design. We tested the hypotheses in Textbox 1.

XSL•FO
RenderX

Textbox 1. Hypotheses.

- The effectiveness of internet-delivered psychoeducation (SCOPE [Spectrum Computerized Psychoeducation]) on autism knowledge (measured with an autism quiz) is superior to treatment-as-usual (TAU) or active web-based self-study comparators in a clinical outpatient setting.
- Acceptance of autism diagnosis and quality of life will increase at follow-up in SCOPE participants compared to TAU or active web-based self-study comparators.
- Symptoms of depression and anxiety will not increase following SCOPE participation.

Methods

Trial Design and Clinical Setting

This study is a 3-armed pragmatic RCT [33] of the internet-delivered psychoeducational intervention SCOPE for transition-aged autistic youth, conducted at Habilitation & Health (Stockholm Region outpatient disability services) in Stockholm, Sweden. The trial was registered at ClinicalTrials.gov (NCT03665363). Participants were enrolled between 2016 and 2021. Therapists were existing Habilitation and Health clinical staff trained in the SCOPE method.

Ethical Considerations

In accordance with the ethical approval by the Swedish Ethical Review Authority (reference 2014/1868-31/5), prior to data collection, informed consent was obtained from all participants, and they were informed of their right to withdraw at any time. Participants were also informed about the purpose of the study and the methods used, and potential risks and benefits were discussed. Brief explanations about primary and potential secondary analyses of collected data were provided. Additionally, as approved by the Swedish Ethical Review Authority, all participants received a gift certificate of SEK 200 (US $22) after completing follow-up outcome measures.

Participants' Eligibility

Participants (n=141) were older adolescents and young adults, aged 16-25 years, with an autism diagnosis. Participants were recruited nationally through outpatient habilitation services, publicly funded psychiatric services, and private health care service providers. Swedish health care legislation allows nationwide service access, that is, individuals can seek health care in regions other than their listed address. All interested individuals were directed to a study recruitment website, provided by Habilitation & Health, to sign up and submit written consent. Professional staff and parents could assist participants in registering and accessing the study but could not apply on behalf of participants. Criteria for inclusion were (1) diagnosed autism (in accordance with the *Diagnostic Statistical Manual of Mental Disorders* or *International Classification of Diseases, Tenth Revision*) through community diagnosis with a documented IQ in the broad average range (ie, full-scale IQ>70); (2) aged 16-25 years; (3) sufficient Swedish language proficiency; and (4) the ability to log on and use the internet-based platform. Exclusion criteria were (1) diagnosed with acquired brain injury; (2) current problematic substance use; (3) severe suicidal ideation; (4) other severe psychiatric disorders (eg, psychosis); or (5) adverse personal circumstances (eg, being homeless) that could hinder participation. Ongoing pharmacological or other treatment was not considered a reason for exclusion.

Procedure

Overview

An experienced licensed clinical psychologist working at Habilitation & Health (authors AB and ES) conducted the eligibility assessment using a video link. Community-based autism diagnoses were obtained through access to Region Stockholm medical records and signed diagnostic assessment reports. Autism diagnoses were validated using the Ohio State University Global Severity Scale for Autism (OARS-4) autism symptom checklist as an interview [34,35], and functional difficulties were validated using the World Health Organization Disability Assessment Schedule (version 2.0; WHODAS 2.0) interview [36]. In addition, self-rated autism traits (Social Responsiveness Scale, Second Edition [SRS-2]) [37] and symptoms of depression (Montgomery-Åsberg Depression Rating Scale—Revised [MADRS-R]) [38] were collected on the internet platform prior to the intake interview. SRS-2 was used to screen for the extent of autism traits (no cutoff was applicable), and MADRS-R was used to screen for depression symptoms (cutoff >35 was used as an exclusion criteria). The SCOPE intervention and active control condition of self-study consisted of weekly modules presented through the Swedish national online platform for telehealth services (*Stöd och Behandling* [SoB]) over 8 weeks. In contrast, treatment as usual (TAU) only involved the administration of data collection from the preintervention to follow-up time points. At follow-up, the active controls for self-study and TAU were asked if they wished to take part in SCOPE during the following term.

Randomization

Using the online tool randomizer.org, participants were randomized after baseline assessments individually to 1 of 3 conditions, with a 2:1:1 ratio to SCOPE, TAU, and active controls of self-study through 16 blocks of a minimum of 4 and a maximum of 13 participants. Baseline assessments were taken before the allocation of randomization was presented. Randomization sequences were conducted by author TH using a consecutively numbered list of the current block. The randomization results were given to author AB, who informed author AM of the allocation and in turn informed participants of their allocation via the SoB platform. Due to planning purposes and therapist availability, many participants had to wait until there were enough participants to be randomized to the next scheduled SCOPE block, ranging from a few days up to 70 days. Participants and therapists were blind to allocation at the preintervention assessments. When allocation was presented, no masking was achievable as therapists and

XSL•FO
RenderX

participants were informed of the allocations and what they entailed.

Intervention

The SCOPE psychoeducative program was codeveloped with clinicians and autistic young adults, as described in more detail by Backman et al [29]. In addition, autistic designers created the visual presentation of some of the content. The contents have been carefully worded, containing information about autism and explicitly intending to empower the autistic participant. Following the feasibility study [29], SCOPE content was revised based on the participant feedback to amend language-related issues; the stringency of theme content; and, most saliently, include more information about autism-related strengths. SCOPE is an individual program that participants complete whenever and wherever they want. SCOPE has 8 autism-related themes (Introduction to Autism; Social Interaction; Behaviors & Interests; Theory of Mind; Central Coherence; Executive Functions; Intelligence & Memory; and Perception & Motor Skills). Each module focuses on one theme, containing (1) text describing the module topic; (2) four video vignettes with recurring transition-aged autistic characters who speak about their lives and perspectives on the module topic; (3) a bullet-point list of nonautistic characteristics related to the module's topic; and (4) self-reflection using 3 or 4 questions about the module topic, answered by multiple-choice bullets and voluntary open-ended written comments. In addition, the modules include evaluations of the content. Participants could adjust the pace at which they used the program contents but were asked to complete each module within a week.

The conception of the SCOPE treatment manual, therapist training, and supervision were conducted by the program developers (authors AM and ENC). The therapists were trained in the SCOPE manual, using a digital training program. They also received clinical supervision with their first 2 participants, and treatment fidelity was followed up. The manual included instructions for SCOPE's technical aspects and provided the aims and perspectives relevant to SCOPE.

Therapists were health professionals, including clinical psychologists, special educators, social workers, and occupational therapists working in specialized outpatient habilitation service units for older autistic adolescents and autistic adults. SCOPE participants had weekly written contact with a therapist who supported them in interpreting the program content, providing brief motivational comments, dealing with practical issues that arose, and responding to participant comments using a phrasebook included in the SCOPE treatment manual for consistency in style and content of responses. In addition, the therapists reminded participants to log on by sending reminders and, on occasion, separate emails or SMS text messages if the system reminders went unopened.

The *self-study* condition entailed a standardized weekly message in SoB with links to 8 informative websites (1 link per week). The websites included professional, governmental, or stakeholder organization websites covering autism and related conditions (eg, the Swedish association for neurodevelopmental conditions Riksförbundet Attention or the Gillberg Neuropsychiatry Centre's website). Self-study participants received initial information about allocation and what to expect from the self-study concept (ie, to open the weekly messages containing a link to an informative website) but no further personalized contact.

TAU entailed any ongoing support or intervention provided by health care services (eg, mental health or habilitation services). Participants reported on their ongoing health care contacts with a self-report form at baseline and 3-month follow-up (see Multimedia Appendix 1 for details). No restrictions regarding treatments were imposed.

Outcomes

All data were collected online using the Swedish national platform for internet-delivered interventions (SoB) at preintervention (baseline), postintervention, and 3-month follow-up time points. Data were then entered into an anonymized database.

Background and Demographic Data

Diagnostic assessment reports were extracted from the clinical patient records. Participants also completed a questionnaire covering sociodemographic and background information [29,39]. Finally, the participants' use of health care and other services was recorded using the background and demographic information questionnaire, collected at baseline and 3-month follow-up (Multimedia Appendix 1).

Outcome Measures

Effectiveness

The primary outcome measure was autism knowledge. All participants completed the Autism Spectrum Disorder Quiz (ASD Quiz), which had shown acceptable reliability in a previous study [29], modified from a corresponding knowledge quiz [39,40]. The ASD Quiz consists of 16 questions related to facts about autism spectrum disorders; items are rated true, false, or unsure and subsequently assessed as either correct or incorrect, with correct answers summed for a total score out of 16. Higher scores indicate more autism knowledge. To evaluate the reliability of the ASD Quiz items, we applied the average item difficulty analysis, with 0.65 considered acceptable [41]. The average item difficulty at the preintervention time point was 0.62, while at the postintervention and follow-up time points, it was 0.69 and 0.69, respectively. We applied Kuder-Richardson 20 to estimate the internal consistency of the quiz; the coefficient α was 0.57 (n=141) at the preintervention time point, 0.63 (n=108) at the postintervention time point, and 0.59 (n=98) at follow-up.

Secondary effectiveness measures were self-rating questionnaires about the acceptance of diagnosis, QoL, as well as symptoms of mental ill-health. Acceptance of diagnosis was estimated using a modified version of the Acceptance and Action Questionnaire-II (AAQ-II) [29,42], rated from 0 to 7 on a Likert scale, with high scores indicating a low acceptance of the diagnosis. The modifications were done to enable the measurement of attitude (such as acceptance) toward one's autism diagnosis through a questionnaire that had previously operationalized the measurement of "acceptance," which in its original form meant to measure psychological flexibility. The

XSL•FO
RenderX

modified items were as follows: (1) “My diagnosis makes it difficult for me to lead a life I could value”; (2) “I am afraid of my diagnosis”; (3) “I worry about not being able to control my worries and my feelings regarding my diagnosis”; (4) “My diagnosis prevents me from leading a fulfilling life”; (5) “My diagnosis creates problems in my life”; (6) “I feel uncomfortable with my diagnosis”; and (7) “My diagnosis gets in the way of my success.” Cronbach α at baseline for the acceptance of diagnosis was 0.85. Next, QoL was measured using The Brunnsviken Brief Quality of Life Scale (BBQ) [43], rated from 0 to 5 on a Likert scale for 2 aspects (satisfaction and importance) per QoL area, which was multiplied and then summed, with high scores indicating higher QoL. Cronbach α at baseline for the BBQ was 0.75. Finally, questionnaires for symptoms of mental health problems using the Swedish version of the Hospital Anxiety and Depression Scale (HADS) [44-46], rated from 0 to 3 on a Likert scale, with higher scores indicating more symptoms. Cronbach α at baseline for the HADS was 0.85.

Feasibility and Intervention Credibility

Adequate completion rate was defined a priori as >75% of participants completing SCOPE. Intervention completion was defined as having completed at least two-thirds of the modules (ie, 5/8, 62% of modules).

A modified version of the Treatment Credibility Scale (TCS) [47,48] was used to measure expectations of improvement and intervention credibility in SCOPE and self-study participants. The modified TCS is a visual analog scale rated from “Low credibility” or “Not at all” (0) to “High credibility” or “Very much” (10), with the mean of all items as the final score. The TCS was administered following the first module of SCOPE and self-study.

Adverse Events and Serious Adverse Events.

Adverse events (AEs) were defined in the individual Case Report Forms as “any inconvenience that participants spontaneously reported,” and serious AEs (SAEs) were defined as “anything that has required inpatient hospitalization.”

Sample Size

A sample size of 126 participants was calculated to provide 80% power (1 – β) at a 2-sided 5% α level, with expectations of a moderate effect size. The effect size was estimated based on the results from the feasibility trial of SCOPE [29] and on a similar trial about psychoeducation for autistic adolescents [18]. To retain power with the expected attrition of 15%, we aimed to include 150 participants.

Statistical Analyses

All statistical analyses were planned a priori; α was set at P=.05. The 2:1:1 ratio of randomization was chosen to enable a sufficient number of participants within the project time.

All variables were screened for violations of normality and sphericity; outcome data were approximately normally distributed. Statistical outliers for all outcome measures were screened using boxplots in SPSS (IBM Corp), and the few (n=7) extreme outliers (defined as outside the first or third quartile [SD 1.5]) were corrected by moving them to the closest whisker (ie, on the first or third quartile). Missingness was deemed completely at random for all outcome variables and time points using the Little missing completely at random test; we used allocation, age, gender, SRS-2, baseline BBQ, baseline HADS anxiety, and baseline HADS depression as assessment variables.

To determine changes between the intervention conditions (SCOPE vs TAU vs self-study) from the preintervention to postintervention time points and from the preintervention to follow-up time points, linear mixed models [49] used all available data without excluding cases with missing data. The models included random intercept, random time (slope), and other covariance structures; the model with the best fit was established using the log-likelihood ratio test. For outcomes where the slopes between self-study and TAU did not differ significantly (all outcomes apart from the ASD Quiz), the 2 control conditions were collapsed into one to increase power. For the ASD Quiz where the slopes of self-study and TAU differed significantly, post hoc tests were conducted to analyze the slopes of SCOPE with self-study and TAU, respectively. Between-group effect sizes, presented as Cohen d, were calculated by dividing the above-described coefficients of time×group interaction effect by the observed preintervention pooled SD [50]. The benchmark for Cohen d was based on the guidelines proposed by Cohen [51], where d values of 0.2, 0.5, and 0.8 are considered small, medium, and large effects, respectively.

Community Involvement

The SCOPE intervention has been codeveloped and pilot-tested with autistic young adults. Several pilot iterations of the program were carried out online with the explicit purpose of codeveloping the program delivery (ie therapist involvement and feedback) and the contents evaluating the internet delivery, different aspects of the content (ie, module themes, phrasing, and video-vignette content), and therapist involvement and feedback. This work was carried out before the research projects commenced.

Results

Overview

A total of 166 participants were assessed for eligibility; see Figure 1 for the CONSORT (Consolidated Standards of Reporting Trials) flowchart. In total, 4 participants either did not meet inclusion or met exclusion criteria, and another 17 declined participation. The remaining (n=145) participants were enrolled in the study; they were asked to complete baseline measures and subsequently randomized. Baseline characteristics for SCOPE, self-study active controls, and TAU controls are presented in Table 1. Recruitment to the study spanned 5 years to achieve the final number of participants. The lengthy recruitment process was mainly due to slow rates of interested participants.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart of participants. SCOPE: Spectrum Computerized Psychoeducation; SS: self-study; TAU: treatment as usual.

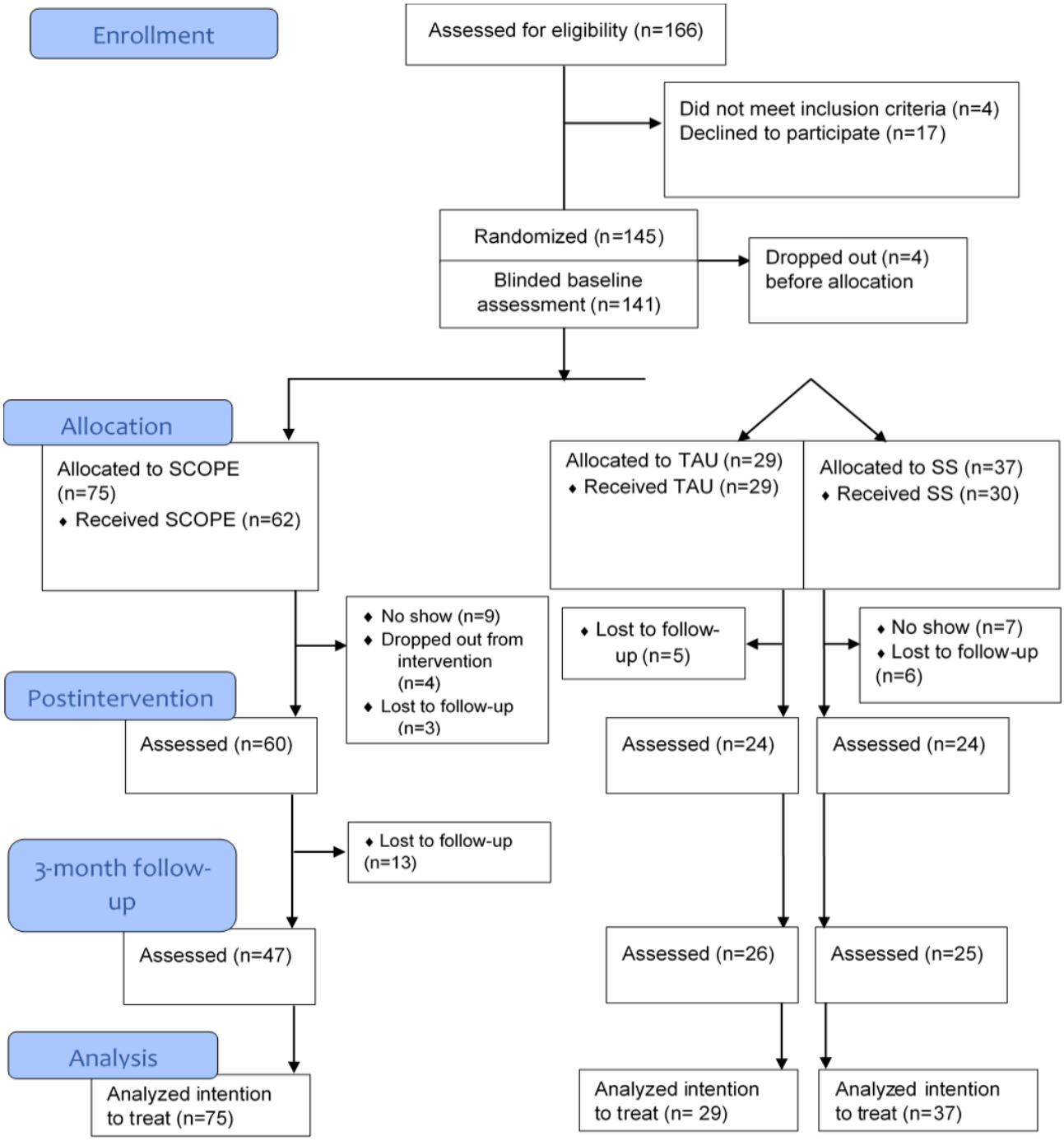

XSL•FO
RenderX

Table 1. Study participants' sociodemographic and clinical characteristics at baseline, presented for the entire study group as well as by intervention condition.

Characteristics	All participants (N=141)	SCOPE[a] (n=75)	Treatment as usual (n=29)	Self-study (n=37)
Age (y), mean (SD)	20.46 (2.96)	20.48 (3.07)	20.66 (3.10)	20.27 (2.67)
Age at diagnosis (y), mean (SD)	17.14 (4.38)	17.38 (3.93)	15.5 (4.76)	17.85 (4.82)
WHODAS 2.0[b] summary score, mean (SD)	34.42 (11.23)	34.82 (11.15)	32.93 (12.09)	34.79 (10.9)
SRS-2[c] *t*-score, mean (SD)	62.99 (7.67)	62.72 (7.68)	63.72 (7.43)	62.94 (7.86)
OARS-4[d] score, mean (SD)	1.43 (0.39)	1.42 (0.37)	1.40 (0.43)	1.45 (0.41)
Gender, female, n (%)	92 (65.2)	54 (72)	18 (62.1)	20 (54.1)
Psychiatric co-occurrence, n (%)				
At least 1 psychiatric condition[e]	89 (63.1)	53 (70.7)	13 (44.8)	23 (62.2)
At least 2 psychiatric conditions	35 (24.8)	20 (26.7)	7 (24.1)	8 (21.6)
ADHD[f] diagnosis	48 (34.0)	28 (37.3)	7 (24.1)	13 (35.1)
Other NDDs[g,h]	15 (10.6)	9 (12)	1 (3.4)	5 (13.5)
Educational attainment[i], n (%)				
Upper secondary	72 (54.1)	35 (46.7)	14 (50)	23 (62.2)
College or university	5 (3.6)	3 (4)	1 (3.6)	1 (2.7)
Living situation, n (%)				
Living independently	42 (29.8)	22 (29.3)	9 (31)	11 (29.7)
Living with partner	11 (7.8)	8 (10.7)	1 (3.4)	2 (5.4)
Living with parent	88 (62.4)	45 (60)	19 (65.5)	24 (64.9)
Occupation, n (%)				
Independent employment or student[j]	94 (66.7)	46 (62.3)	22 (75.9)	26 (70.3)
Supported employment	21 (14.9)	15 (20)	2 (6.9)	4 (10.8)
Not currently working or studying	8 (5.7)	5 (6.7)	2 (6.9)	1 (2.7)

[a]SCOPE: Spectrum Computerized Psychoeducation.

[b]WHODAS 2.0: World Health Organization Disability Assessment Schedule (version 2.0).

[c]SRS-2: Social Responsiveness Scale, Second Edition.

[d]OARS-4: Ohio State University Global Severity Scale for Autism.

[e]At least 1 psychiatric condition entails *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* psychiatric disorders other than the neurodevelopmental disorders.

[f]ADHD: attention-deficit/hyperactivity disorder.

[g]NDD: neurodevelopmental disorder.

[h]Other NDDs include communication disorders, specific learning disorders, tic disorders, and motor disorders.

[i]Educational attainment: ongoing educational attainment.

[j]Independent occupation or student: any work or education that is held without explicit support.

Outcomes

Preintervention, postintervention, and follow-up assessments for all conditions based on the observed means and SDs are presented in Table 2, and interaction effects of between-group effects for preintervention to postintervention and preintervention to follow-up differences for all outcome measures are presented in Table 3.

XSL•FO
RenderX

Table 2. Observed means and SDs for SCOPE[a], self-study active controls, and TAU[b] at the preintervention, postintervention, and 3-month follow-up time points for primary (ASD Quiz[c]) and secondary outcomes (HADS[d] anxiety, HADS depression, modified AAQ-II[e], and BBQ[f]).

Outcome and treatment arm	Preintervention time point (baseline)		Postintervention time point		Follow-up	
	Participants, n	Score, mean (SD)	Participants, n	Score, mean (SD)	Participants, n	Score, mean (SD)
ASD Quiz						
SCOPE	75	10 (2.7)	60	11.28 (2.17)	47	11.66 (2.59)
Self-study	37	10.24 (1.85)	24	11.96 (1.46)	26	10.68 (2.04)
TAU	29	9.62 (2.57)	24	10.13 (2.21)	25	10.27 (2.36)
HADS anxiety						
SCOPE	75	10 (4.39)	60	9.33 (4.06)	47	9.17 (4.19)
Self-study	37	9.70 (4.49)	24	10 (4.02)	26	9.52 (4.38)
TAU	29	9.45 (5.06)	24	10.04 (4.33)	25	9.23 (4.02)
HADS depression						
SCOPE	75	7.17 (4.16)	60	6.97 (4.45)	47	6.47 (4.03)
Self-study	37	7.27 (3.89)	24	7.13 (4.09)	26	7.60 (3.45)
TAU	29	6.00 (3.20)	24	6.13 (4.39)	25	5.62 (3.81)
Acceptance of diagnosis (modified AAQ-II)						
SCOPE	75	24.83 (8.43)	60	22.63 (8.65)	47	21.38 (8.62)
Self-study	37	22.81 (8.62)	24	22.21 (8.71)	26	22.60 (8.80)
TAU	29	26.03 (8.32)	24	23.67 (8.40)	25	22.92 (8.26)
BBQ						
SCOPE	75	41.77 (20.20)	60	45.51 (20.69)	47	49.30 (19.06)
Self-study	37	44.84 (19.32)	24	45 (19.24)	26	43.32 (19.17)
TAU	29	45.03 (17.50)	24	42.33 (17.40)	25	40.24 (18.32)

[a]SCOPE: Spectrum Computerized Psychoeducation.
[b]TAU: treatment as usual.
[c]ASD Quiz: Autism Spectrum Disorder Quiz.
[d]HADS: Hospital Anxiety and Depression Scale.
[e]AAQ-II: Acceptance and Action Questionnaire-II.
[f]BBQ: Brunnsviken Brief Quality of Life Scale.

XSL•FO
RenderX

Table 3. Between-group effect sizes (Cohen d) based on the estimates obtained in the hierarchical linear mixed models, presented together with the coefficient for mean change (least-squares means) and 95% CI for the interaction effect between time and intervention condition.

Measure	Comparison	Preintervention to postintervention difference				Preintervention to follow-up difference			
		Mean change scores	95% CI	Effect size[a], *d*	95% CI	Mean change scores	95% CI	Effect size[a], *d*	95% CI
ASD Quiz[b]	SCOPE[c] vs TAU[d]	1.03	0.04 to 2.02	0.47	–0.01 to 0.96	1.15	0.15 to 2.15	0.46	–0.04 to 0.95
ASD Quiz	SCOPE vs self-study	–0.18	–1.18 to 0.83	–0.09	–0.57 to 0.39	1.27	0.25 to 2.30	0.53	0.02 to 1.03
HADS ANX[e]	SCOPE vs controls[f]	–0.95	–2.03 to 0.13	–0.23	–0.62 to 0.15	–.90	2.01 to 0.21	–0.22	–0.62 to 0.19
HADS DEP[g]	SCOPE vs controls	–.36	–1.57 to 0.86	–0.08	–0.47 to 0.30	–0.06	–1.82 to 0.68	–0.15	–0.55 to 0.25
Acceptance of diagnosis	SCOPE vs controls	–1.03	–3.04 to 0.98	0.12[h]	–0.26 to 0.50	–1.99	–4.07 to 0.09	0.23[h]	-0.15 to 0.61
BBQ[i]	SCOPE vs controls	7.21	1.79 to 12.62	0.37	–0.02 to 0.76	11.33	5.70 to 16.97	0.60	0.21 to 0.99

[a]Effect size measured by Cohen *d* (0.2 = small; 0.5 = moderate; 0.8 = large).

[b]ASD Quiz: Autism Spectrum Disorder Quiz.

[c]SCOPE: Spectrum Computerized Psychoeducation.

[d]TAU: treatment as usual.

[e]HADS ANX: Hospital Anxiety and Depression Scale—Anxiety subscale.

[f]Controls: self-study and TAU collapsed into one group for increased power.

[g]HADS DEP: Hospital Anxiety and Depression—Depression subscale.

[h]Sign of effect sizes have been changed so that positive scores indicate improvement.

[i]BBQ: Brunnsviken Brief Quality of Life Scale.

Primary Outcome

The primary outcome was autism knowledge, estimated using the ASD Quiz, which significantly increased for SCOPE and self-study participants from the baseline to postintervention time point compared to TAU, with a small effect size of *d*=0.47 (*P*=.04) between SCOPE and TAU at the postintervention time point. In addition, self-study participants had significantly higher autism knowledge compared to TAU at the postintervention time point, yielding a medium effect size of *d*=0.60 (mean change score: 1.20, 95% CI 0.11-2.30; *P*=.03). However, self-study participants' scores returned to baseline levels and did not differ significantly from TAU at follow-up (mean change score: –0.13, 95% CI –1.20 to 0.94; *P*=.81). In contrast, the ASD Quiz scores remained stable for SCOPE participants compared to self-study participants (*d*=0.53, *P*=.02) and TAU participants (*d*=0.48, *P*=.02) at 3-month follow-up. Figure 2 shows a visual representation of these interaction effects.

Figure 2. Multiple line graph of observed autism knowledge means at the preintervention, postintervention, and 3-month follow-up time points by intervention condition, with 95% CI error bars. SCOPE: Spectrum Computerized Psychoeducation; TAU: treatment as usual.

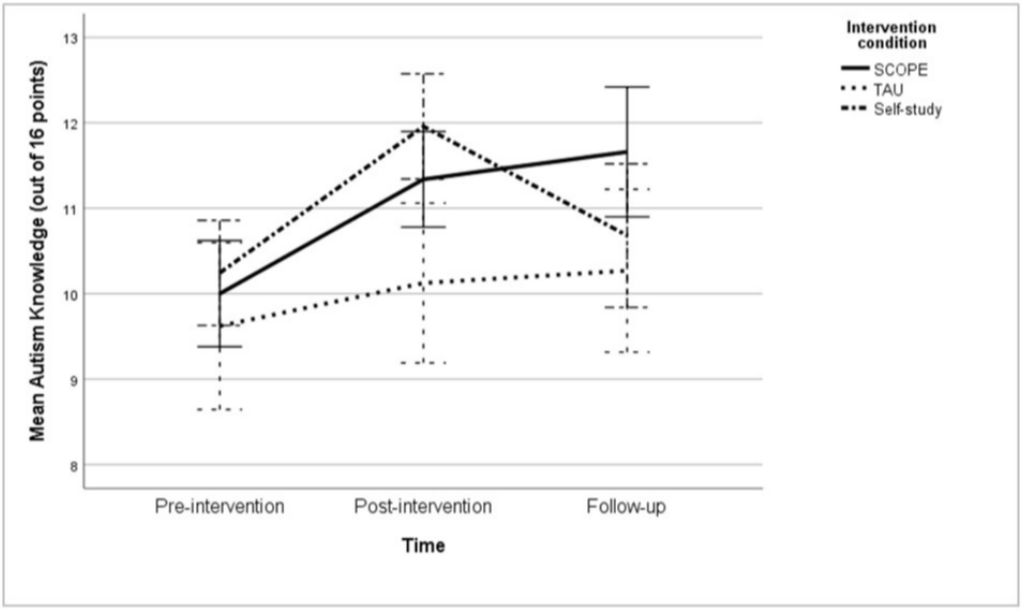

XSL•FO
RenderX

Secondary Outcomes

The self-study active control and TAU conditions did not differ significantly for any secondary outcomes and were collapsed into one control group (see Table 3 for details).

The preintervention to postintervention increase in perceived QoL (the BBQ) was significantly larger in the SCOPE condition than in the combined controls. The between-group effect size was small at the postintervention time point (d=0.37, P=.02). However, QoL increased significantly at follow-up for SCOPE participants compared to combined controls, with a medium effect size of d=0.60 (P=.001).

Neither HADS Anxiety nor HADS Depression scores increased significantly between time points for SCOPE participants or the combined control condition. In addition, acceptance of diagnosis (modified AAQ-II) did not increase significantly between time points for SCOPE participants compared to the combined control condition.

Feasibility and Intervention Credibility

Reported intervention credibility was similar for SCOPE participants (n=60; mean 7.69, SD (1.53) and self-study participants (n=30; mean 7.32, SD 1.90), measured after the first module.

At the postintervention time point, 2 (4%) of 62 SCOPE participants had not completed the intervention, whereas 6 (20%) out of 30 did not complete the self-study. In total, 5 (8%) out of 29 participants allocated to TAU dropped out of the study.

AEs and SAEs

Although no SAEs were reported by SCOPE participants, there were 2 AEs spontaneously reported at the postintervention time point, where one participant dropped out of the study following module 6, referring to mental health problems, with no further contact after therapist attempts. However, it is unclear if the problems were related to SCOPE participation. The other participant completed SCOPE and all measurements but gave several spontaneous comments about feeling angry and commented that she or he felt some examples of autistic individuals were unfair and misrepresentative. The comments were given as part of the weekly written communication with the assigned therapist.

Discussion

Principal Findings

To the best of our knowledge, this is the first RCT to compare a structured, therapist-supported, internet-delivered psychoeducative intervention for transition-aged autistic youth against an active control condition (self-study) and TAU. In this trial, SCOPE was superior to self-study and TAU conditions at follow-up in providing autism knowledge and increasing QoL without increasing mental health problems. The findings of this RCT align with a previous open feasibility trial of SCOPE [29], in that the completion rate was high and the intervention was perceived as credible.

Our findings offer further support for the SCOPE intervention. In terms of our first hypothesis, we observed that SCOPE and self-study active control participants increased their autism knowledge, with small- to medium-sized effects, from the baseline to postintervention time points, whereas TAU did not. However, the increase in knowledge was only retained by SCOPE participants, showing a medium-sized effect at the 3-month follow-up, where self-study active controls participants' knowledge returned to baseline rates. Furthermore, autism knowledge gains at follow-up were not related to preintervention treatment credibility, which was rated equally by the SCOPE and self-study groups. Instead, the maintained autism knowledge may be related to the carefully selected information, the codeveloper's (ie, autistic emerging adults) influence on content and language, and the self-reflection questions possibly facilitating a more thorough cognitive processing of the learned information. Moreover, therapist guidance, through weekly written contact, is included in the SCOPE program but not in the self-study condition. However, this study cannot differentiate the active components in the interventions. Nevertheless, participants' knowledge gains align with previous studies reporting similar results following group-based, structured psychoeducation for autistic adolescents and older adults [18,19]. In addition, a unique aspect of this study is the addition of an active comparator conducting self-studies, which has not previously been reported for similar interventions. Therefore, the results may have important clinical implications, as it is common for mental health care providers to give reading tips for self-study in the postdiagnostic phase. Based on the results of this study, self-study of internet-based reading materials without therapist support provided only temporary knowledge gains.

In addition to measuring autism knowledge, it is important to consider QoL as an important outcome following psychological interventions [31]. Improving QoL aligns well with the aims of psychoeducation—to empower and increase self-esteem. In salient theories on QoL, self-esteem is an integral part of QoL [52,53]. Moreover, Schalock [52] and the autism-focused review of QoL [54] see empowerment as a component of self-determination, which in turn is a component of QoL. As such, our second hypothesis about increased QoL following SCOPE was confirmed by the results, in that SCOPE was superior to controls in increasing QoL at the postintervention time point and even more so at follow-up. This was in line with the preliminary results of Hidalgo et al [19], showing a small increase in QoL in autistic adults following a structured psychoeducational group. Furthermore, enabling an increased sense of QoL is in line with up-to-date goals of treatment guidelines for autism [16] and calls to action in efficient postdiagnostic care [15], explicitly recommending psychoeducation as a first-line intervention after a confirmed autism diagnosis. Attempting to reframe and guide an individual's perception of what autism is for them aligns with modern goals for psychoeducation [55]. A few studies on psychoeducation for autistic individuals have seen tentative effects on self-awareness and self-esteem [18,20], also described in the qualitative responses from the feasibility trial of SCOPE [29]. Increased knowledge about autism could act as a mediator of changes in self-perception and recognition of how autism impacts themselves and their environment. Although we were

XSL•FO
RenderX

not able to conduct that kind of mediator analysis, this could be addressed in future studies.

Moreover, related to the third hypothesis, the increased knowledge among SCOPE and self-study participants was not coupled with increased symptoms of anxiety or depression. Although the primary goals of psychoeducation are not expected to have direct effects on specific symptoms such as anxiety and depression [9,55], and consequently, such effects have not previously been consistently found in the literature on psychoeducational interventions about autism [18,19,21,29,56]. Therefore, positive effects on symptoms of anxiety and depression were not expected to be improved in this study. Notably, the stability of anxiety and depression symptoms over time is an important finding, as psychoeducation may be associated with initial adverse psychological effects [17,32].

Furthermore, this study also partly supports the open feasibility trial of SCOPE [29], demonstrating high completion rates in the SCOPE. However, the recruitment of study participants was slow, with approximately 30 participants included every year (2016-2020). Given the apparent willingness of transition-aged autistic participants to receive interventions [6,57,58], we expected a higher flow of participants. The slow recruitment could suggest that we did not reach the intended target audience; perhaps, the intervention was perceived as too time-consuming or unappealing to transition-aged autistic youth, or the mode of study entry (application through the study website using e-identification) was cumbersome. Unfortunately, this study was unable to clearly identify the complex reasons for participant recruitment difficulties in this trial. A qualitative analysis of participants' written responses and evaluations could deepen our understanding of the lived experiences of SCOPE participation and thus inform future implementation of similar programs in clinical practice.

Limitations

This study has several limitations. First, the primary outcome measure (ASD Quiz) was based on previous similar measures [19,40,59], but it has not been psychometrically evaluated. However, no consensus regarding autism knowledge assessment exists, with considerable heterogeneity in the field [59], and the ASD Quiz used in this study has previously achieved acceptable reliability [29]. Although the knowledge gains were modest regarding absolute numbers, the effect size was medium at 3-month follow-up. However, a risk for ceiling effects in the primary outcome could have limited our ability to measure the treatment effects. Second, identifying secondary outcome measures available in Swedish for autistic individuals was challenging, especially at the inception of this study. For future studies, using outcome measures for self-esteem and empowerment would be valuable to evaluate the core aims of psychoeducation. Failing to measure these intervention outcomes misses the mark of understanding the potential effects of SCOPE within these domains. Third, detection bias may have been introduced due to assessors (the autistic individuals) being aware of intervention allocation at postintervention assessments [60]. Finally, the study's generalizability may have been limited due to potential sampling issues. As noted in Table 1, the sample consists mainly of newly diagnosed female participants. The female majority may be reflective of findings maintaining that female individuals obtain their diagnosis later than male individuals [61,62], which may explain the oversampling of female participants in this type of intervention study. Moreover, autistic female individuals also report higher rates of mental health problems [63], perhaps contributing to more treatment-seeking behavior in general. Additionally, our participants reported many health care contacts (see Multimedia Appendix 1). Overall, our sample could represent a treatment-seeking *and* receiving subset of transition-age autistic youth that is not representative of those that have "fallen off a cliff" [6].

Conclusions

This study shows that structured internet-delivered psychoeducation is effective in an outpatient context. The SCOPE program was superior to self-study active controls and TAU in imparting autism knowledge and increasing QoL. These results were maintained or further improved at the 3-month follow-up. In addition, providing internet-delivered services enables wide geographical accessibility and provides digital communication sought by transition-aged autistic youth.

Acknowledgments

This study was supported by Region Stockholm funding for medical training and research (PPG/ALF) and the following foundations: Promobilia, Jerringfonden, Stiftelsen L.J. Boëthius, Sällskapet Barnavård, and Sunnerdahls Handikappfond. We want to thank all adolescents and young adults who have participated in the study. We would also like to thank research assistant Maj Frostvittra for data management. The authors attest that there was no use of generative AI technology in the generation of text, figures, or other informational content in this manuscript.

Conflicts of Interest

AB, SV, AM, ENC, ES, and TH are employed by Region Stockholm. AB receives royalties from the publisher Natur och Kultur for a popular science book unrelated to this study. TH receives royalties from Hogrefe for treatment manuals unrelated to this study. EZ receives royalties from the Hogrefe Publishing Group for the Swedish version of the Social Responsiveness Scale, Second Edition (SRS-2). LRP discloses no conflicts of interest related to this study.

Multimedia Appendix 1
Self-reported current interventions and psychotropic medications at baseline and 3-month follow-up, by treatment arm.

XSL•FO
RenderX

[DOCX File , 27 KB - jmir_v26i1e49305_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 996 KB - jmir_v26i1e49305_app2.pdf]

References

1. Thompson C, Bölte S, Falkmer T, Girdler S. To be understood: transitioning to adult life for people with autism spectrum disorder. PLoS One 2018;13(3):e0194758 [FREE Full text] [doi: 10.1371/journal.pone.0194758] [Medline: 29579089]
2. Matthews NL, Smith CJ, Pollard E, Ober-Reynolds S, Kirwan J, Malligo A. Adaptive functioning in autism spectrum disorder during the transition to adulthood. J Autism Dev Disord 2015;45(8):2349-2360. [doi: 10.1007/s10803-015-2400-2] [Medline: 25733158]
3. Howlin P. Adults with autism: changes in understanding since DSM-111. J Autism Dev Disord 2021;51(12):4291-4308 [FREE Full text] [doi: 10.1007/s10803-020-04847-z] [Medline: 33474661]
4. McCauley JB, Pickles A, Huerta M, Lord C. Defining positive outcomes in more and less cognitively able autistic adults. Autism Res 2020;13(9):1548-1560. [doi: 10.1002/aur.2359] [Medline: 32851813]
5. Taylor JL, Henninger NA. Frequency and correlates of service access among youth with autism transitioning to adulthood. J Autism Dev Disord 2015;45(1):179-191 [FREE Full text] [doi: 10.1007/s10803-014-2203-x] [Medline: 25081594]
6. Bennett AE, Miller JS, Stollon N, Prasad R, Blum NJ. Autism spectrum disorder and transition-aged youth. Curr Psychiatry Rep 2018;20(11):103. [doi: 10.1007/s11920-018-0967-y] [Medline: 30238166]
7. Lee EAL, Black MH, Falkmer M, Tan T, Sheehy L, Bölte S, et al. "We Can See a Bright Future": parents' perceptions of the outcomes of participating in a strengths-based program for adolescents with autism spectrum disorder. J Autism Dev Disord 2020;50(9):3179-3194. [doi: 10.1007/s10803-020-04411-9] [Medline: 32076957]
8. Cribb S, Kenny L, Pellicano E. 'I definitely feel more in control of my life': the perspectives of young autistic people and their parents on emerging adulthood. Autism 2019;23(7):1765-1781 [FREE Full text] [doi: 10.1177/1362361319830029] [Medline: 30818981]
9. Motlova LB, Balon R, Beresin EV, Brenner AM, Coverdale JH, Guerrero APS, et al. Psychoeducation as an opportunity for patients, psychiatrists, and psychiatric educators: why do we ignore it? Acad Psychiatry 2017;41(4):447-451. [doi: 10.1007/s40596-017-0728-y] [Medline: 28536990]
10. Lyman DR, Braude L, George P, Dougherty RH, Daniels AS, Ghose SS, et al. Consumer and family psychoeducation: assessing the evidence. Psychiatr Serv 2014;65(4):416-428. [doi: 10.1176/appi.ps.201300266] [Medline: 24445678]
11. Zhao S, Sampson S, Xia J, Jayaram MB. Psychoeducation (brief) for people with serious mental illness. Cochrane Database Syst Rev 2015;2015(4):CD010823 [FREE Full text] [doi: 10.1002/14651858.CD010823.pub2] [Medline: 25854522]
12. Sarkhel S, Singh OP, Arora M. Clinical practice guidelines for psychoeducation in psychiatric disorders general principles of psychoeducation. Indian J Psychiatry 2020;62(Suppl 2):S319-S323 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_780_19] [Medline: 32055073]
13. Axén M, Brar A, Huslid E, Nordin V, Nylander L, Walch M. Regionalt Vårdprogram: ADHD, Lindrig Utvecklingsstörning Och Autismspektrumtillstånd Hos Barn, Ungdomar Och Vuxna. Stockholms, Sweden: Landsting SL; 2010.
14. Wilson EC, Roberts G, Gillan N, Ohlsen C, Robertson D, Zinkstok J. The NICE guideline on recognition, referral, diagnosis and management of adults on the autism spectrum. Adv Mental Health Intell Disabilities 2014;8(1):3-14. [doi: 10.1108/AMHID-05-2013-0035]
15. Mandy W. Six ideas about how to address the autism mental health crisis. Autism 2022;26(2):289-292. [doi: 10.1177/13623613211067928] [Medline: 35109701]
16. NICE. Autism Spectrum Disorder in Under 19s: Support and Management. London, United Kingdom: National Institute of Clinical Excellence; 2013.
17. Beresford B, Mukherjee S, Mayhew E, Heavey E, Park AL, Stuttard L, et al. Evaluating specialist autism teams' provision of care and support for autistic adults without learning disabilities: the SHAPE mixed-methods study. Health Serv Deliv Res 2020;8(48). [doi: 10.3310/hsdr08480] [Medline: 33320485]
18. Gordon K, Murin M, Baykaner O, Roughan L, Livermore-Hardy V, Skuse D, et al. A randomised controlled trial of PEGASUS, a psychoeducational programme for young people with high-functioning autism spectrum disorder. J Child Psychol Psychiatry 2015;56(4):468-476 [FREE Full text] [doi: 10.1111/jcpp.12304] [Medline: 25132516]
19. Hidalgo N, Sjöwall D, Agius H, Byström C, Brar A, Borg J, et al. Psychoeducational group intervention for intellectually able adults with autism and their close relations (Prisma) - an open feasibility study. BMC Psychiatry 2022;22(1):556 [FREE Full text] [doi: 10.1186/s12888-022-04134-4] [Medline: 35986348]
20. Crane L, Hearst C, Ashworth M, Davies J, Hill EL. Supporting newly identified or diagnosed autistic adults: an initial evaluation of an autistic-led programme. J Autism Dev Disord 2021;51(3):892-905 [FREE Full text] [doi: 10.1007/s10803-020-04486-4] [Medline: 32266684]

XSL•FO
RenderX

21. Lenders M, Ouwens MA, Wilting RMHJ, Videler AC. First things first: an exploration of the effects of psychoeducation for older autistic adults. Autism 2024;28(8):2028-2039 [FREE Full text] [doi: 10.1177/13623613231219745] [Medline: 38197411]
22. Barak A, Hen L, Boniel-Nissim M, Shapira N. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. J Technol Hum Serv 2008;26(2-4):109-160. [doi: 10.1080/15228830802094429]
23. Kazdin AE, Rabbitt SM. Novel models for delivering mental health services and reducing the burdens of mental illness. Clin Psychol Sci 2013;1(2):170-191. [doi: 10.1177/2167702612463566]
24. Gillespie-Lynch K, Brooks PJ, Someki F, Obeid R, Shane-Simpson C, Kapp SK, et al. Changing college students' conceptions of autism: an online training to increase knowledge and decrease stigma. J Autism Dev Disord 2015;45(8):2553-2566. [doi: 10.1007/s10803-015-2422-9] [Medline: 25796194]
25. Gillespie-Lynch K, Kapp SK, Shane-Simpson C, Smith DS, Hutman T. Intersections between the autism spectrum and the internet: perceived benefits and preferred functions of computer-mediated communication. Intellect Dev Disabil 2014;52(6):456-469. [doi: 10.1352/1934-9556-52.6.456] [Medline: 25409132]
26. Sehlin H, Hedman Ahlström B, Bertilsson I, Andersson G, Wentz E. Internet-based support and coaching with complementary clinic visits for young people with attention-deficit/hyperactivity disorder and autism: controlled feasibility study. J Med Internet Res 2020;22(12):e19658 [FREE Full text] [doi: 10.2196/19658] [Medline: 33382381]
27. Westerberg B, Bäärnhielm S, Giles C, Hylén U, Holländare F, Bejerot S. An internet based intervention for adults with autism spectrum disorder-a qualitative study of participants experiences. Front Psychiatry 2021;12:789663 [FREE Full text] [doi: 10.3389/fpsyt.2021.789663] [Medline: 35002808]
28. Lamash L, Little L, Hen-Herbst L. Telehealth interventions to promote health and behavior-related outcomes in adolescents with autism spectrum disorder. J Autism Dev Disord 2023;53(1):405-423 [FREE Full text] [doi: 10.1007/s10803-022-05440-2] [Medline: 35061162]
29. Backman A, Mellblom A, Norman-Claesson E, Keith-Bodros G, Frostvittra M, Bölte S, et al. Internet-delivered psychoeducation for older adolescents and young adults with autism spectrum disorder (SCOPE): an open feasibility study. Res Autism Spectrum Disorders 2018;54:51-64. [doi: 10.1016/j.rasd.2018.07.001]
30. Rozental A, Castonguay L, Dimidjian S, Lambert M, Shafran R, Andersson G, et al. Negative effects in psychotherapy: commentary and recommendations for future research and clinical practice. BJPsych Open 2018;4(4):307-312 [FREE Full text] [doi: 10.1192/bjo.2018.42] [Medline: 30083384]
31. Cuijpers P. Targets and outcomes of psychotherapies for mental disorders: an overview. World Psychiatry 2019;18(3):276-285 [FREE Full text] [doi: 10.1002/wps.20661] [Medline: 31496102]
32. Jutel A, Nettleton S. Towards a sociology of diagnosis: reflections and opportunities. Soc Sci Med 2011;73(6):793-800. [doi: 10.1016/j.socscimed.2011.07.014] [Medline: 21868144]
33. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci 2011;13(2):217-224 [FREE Full text] [doi: 10.31887/DCNS.2011.13.2/npatsopoulos] [Medline: 21842619]
34. Choque Olsson N, Bölte S. Brief report: "Quick and (not so) dirty" assessment of change in autism: cross-cultural reliability of the developmental disabilities CGAS and the OSU autism CGI. J Autism Dev Disord 2014;44(7):1773-1778. [doi: 10.1007/s10803-013-2029-y] [Medline: 24379174]
35. OSU ORU. OSU Autism CGI. Colombus, OH: OSU Research Unit on Pediatric Psychopharmacology; 2005.
36. Üstün TB, Kostanjesek N, Chatterji N, Rehm J. Measuring Health and Disability: Manual for WHO Disability Assessment Schedule (WHODAS 2.0). Geneva, Switzerland: WHO Press; 2010.
37. Constantino JN, Gruber C. Social Responsiveness Scale, Second Edition (SRS-2). Torrance, CA: Western Psychological Services; 2012.
38. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134(4):382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
39. Hirvikoski T, Waaler E, Alfredsson J, Pihlgren C, Holmström A, Johnson A, et al. Reduced ADHD symptoms in adults with ADHD after structured skills training group: results from a randomized controlled trial. Behav Res Ther 2011;49(3):175-185. [doi: 10.1016/j.brat.2011.01.001] [Medline: 21295767]
40. Bramham J, Young S, Bickerdike A, Spain D, McCartan D, Xenitidis K. Evaluation of group cognitive behavioral therapy for adults with ADHD. J Atten Disord 2009;12(5):434-441. [doi: 10.1177/1087054708314596] [Medline: 18310557]
41. Cohen RJ, Swerdlik ME. Psychological Testing and Assessment: An Introduction to Tests and Measurement. New York City, NY: MacGraw-Hill; 2005.
42. Bond FW, Hayes SC, Baer RA, Carpenter KM, Guenole N, Orcutt HK, et al. Preliminary psychometric properties of the acceptance and action questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. Behav Ther 2011;42(4):676-688. [doi: 10.1016/j.beth.2011.03.007] [Medline: 22035996]
43. Lindner P, Frykheden O, Forsström D, Andersson E, Ljótsson B, Hedman E, et al. The Brunnsviken Brief Quality of Life Scale (BBQ): development and psychometric evaluation. Cogn Behav Ther 2016;45(3):182-195 [FREE Full text] [doi: 10.1080/16506073.2016.1143526] [Medline: 26886248]

XSL•FO
RenderX

44. Snaith RP. The hospital anxiety and depression scale. Health Qual Life Outcomes 2003;1:29 [FREE Full text] [doi: 10.1186/1477-7525-1-29] [Medline: 12914662]
45. Lisspers J, Nygren A, Söderman E. Hospital anxiety and depression scale (HAD): some psychometric data for a swedish sample. Acta Psychiatr Scand 1997;96(4):281-286. [doi: 10.1111/j.1600-0447.1997.tb10164.x] [Medline: 9350957]
46. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
47. Borkovec TD, Nau SD. Credibility of analogue therapy rationales. J Behav Ther Exp Psychiatry 1972;3(4):257-260. [doi: 10.1016/0005-7916(72)90045-6]
48. Ost LG. Fading vs systematic desensitization in the treatment of snake and spider phobia. Behav Res Ther 1978;16(6):379-389. [doi: 10.1016/0005-7967(78)90009-8] [Medline: 32870]
49. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. Arch Gen Psychiatry 2004;61(3):310-317. [doi: 10.1001/archpsyc.61.3.310] [Medline: 14993119]
50. Feingold A. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. Psychol Methods 2009;14(1):43-53 [FREE Full text] [doi: 10.1037/a0014699] [Medline: 19271847]
51. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge; 1988.
52. Schalock RL. The concept of quality of life: what we know and do not know. J Intellect Disabil Res 2004;48(Pt 3):203-216. [doi: 10.1111/j.1365-2788.2003.00558.x] [Medline: 15025663]
53. WHO WG. Development of the world health organization WHOQOL-BREF quality of life assessment. The WHOQOL group. Psychol Med 1998;28(3):551-558. [doi: 10.1017/s0033291798006667] [Medline: 9626712]
54. White K, Flanagan TD, Nadig A. Examining the relationship between self-determination and quality of life in young adults with autism spectrum disorder. J Dev Phys Disabil 2018;30(6):735-754. [doi: 10.1007/s10882-018-9616-y]
55. Higgins A, Murphy R, Barry J, Eustace-Cook J, Monahan M, Kroll T, et al. Scoping review of factors influencing the implementation of group psychoeducational initiatives for people experiencing mental health difficulties and their families. J Ment Health 2022;31(6):859-872. [doi: 10.1080/09638237.2020.1714002] [Medline: 31994955]
56. Chancel R, Miot S, Dellapiazza F, Baghdadli A. Group-based educational interventions in adolescents and young adults with ASD without ID: a systematic review focusing on the transition to adulthood. Eur Child Adolesc Psychiatry 2022;31(7):1-21. [doi: 10.1007/s00787-020-01609-1] [Medline: 32889578]
57. Burke MM, Waitz-Kudla SN, Rabideau C, Taylor JL, Hodapp RM. Pulling back the curtain: issues in conducting an intervention study with transition-aged youth with autism spectrum disorder and their families. Autism 2019;23(2):514-523 [FREE Full text] [doi: 10.1177/1362361317753016] [Medline: 29439586]
58. Volkmar FR, Jackson SLJ, Hart L. Transition issues and challenges for youth with autism spectrum disorders. Pediatr Ann 2017;46(6):e219-e223. [doi: 10.3928/19382359-20170519-03] [Medline: 28599026]
59. Harrison AJ, Slane MM, Hoang L, Campbell JM. An international review of autism knowledge assessment measures. Autism 2017;21(3):262-275. [doi: 10.1177/1362361316638786] [Medline: 27154908]
60. Sandbank M, Bottema-Beutel K, Crowley S, Cassidy M, Dunham K, Feldman JI, et al. Project AIM: autism intervention meta-analysis for studies of young children. Psychol Bull 2020;146(1):1-29 [FREE Full text] [doi: 10.1037/bul0000215] [Medline: 31763860]
61. Bargiela S, Steward R, Mandy W. The experiences of late-diagnosed women with autism spectrum conditions: an investigation of the female autism phenotype. J Autism Dev Disord 2016;46(10):3281-3294 [FREE Full text] [doi: 10.1007/s10803-016-2872-8] [Medline: 27457364]
62. Rutherford M, McKenzie K, Johnson T, Catchpole C, O'Hare A, McClure I, et al. Gender ratio in a clinical population sample, age of diagnosis and duration of assessment in children and adults with autism spectrum disorder. Autism 2016;20(5):628-634. [doi: 10.1177/1362361315617879] [Medline: 26825959]
63. Sedgewick F, Leppanen J, Tchanturia K. Gender differences in mental health prevalence in autism. Adv Autism 2020;7(3):208-224. [doi: 10.1108/aia-01-2020-0007]

Abbreviations

AAQ-II: Acceptance and Action Questionnaire-II
AE: adverse event
ASD Quiz: Autism Spectrum Disorder Quiz
BBQ: Brunnsviken Brief Quality of Life Scale
CONSORT: Consolidated Standards of Reporting Trials
HADS: Hospital Anxiety and Depression Scale
MADRS-R: Montgomery-Åsberg Depression Rating Scale—Revised
OARS-4: Ohio State University Autism Rating Scale
QoL: quality of life
RCT: randomized controlled trial

XSL•FO
RenderX

SAE: serious adverse event
SCOPE: Spectrum Computerized Psychoeducation
SoB: Stöd och Behandling (Swedish national online platform for telehealth services)
SRS-2: Social Responsiveness Scale, Second Edition
TAU: treatment as usual
TCS: Treatment Credibility Scale
WHODAS 2.0: World Health Organization Disability Assessment Schedule (version 2.0)

Edited by A Mavragani; submitted 24.05.23; peer-reviewed by A Videler; comments to author 24.07.24; revised version received 10.10.24; accepted 14.10.24; published 28.11.24.

Please cite as:
Backman A, Roll-Pettersson L, Mellblom A, Norman-Claesson E, Sundqvist E, Zander E, Vigerland S, Hirvikoski T
Internet-Delivered Psychoeducation (SCOPE) for Transition-Aged Autistic Youth: Pragmatic Randomized Controlled Trial
J Med Internet Res 2024;26:e49305
URL: https://www.jmir.org/2024/1/e49305
doi:10.2196/49305
PMID:

©Anna Backman, Lise Roll-Pettersson, Are Mellblom, Elisabet Norman-Claesson, Emma Sundqvist, Eric Zander, Sarah Vigerland, Tatja Hirvikoski. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Messenger App–Based Information Provision for Promoting Social Participation to Enhance Well-Being Among Community-Dwelling Adults: Randomized Controlled Trial

Atsushi Nakagomi[1], MD, PhD; Noriyuki Abe[1], PhD; Yu-Ru Chen[1], MMSc; Kazushige Ide[1], PhD; Shuhei Kobayashi[1], MMS; Masamichi Hanazato[1], PhD; Katsunori Kondo[1], MD, PhD

Center for Preventive Medical Sciences, Chiba University, Chiba, Japan

Corresponding Author:
Atsushi Nakagomi, MD, PhD
Center for Preventive Medical Sciences
Chiba University
1-33 Yayoicho
Inage Ward
Chiba, 263-8522
Japan
Phone: 81 43 251 1111
Email: anakagomi0211@gmail.com

Abstract

Background: Engaging in social activities, interacting with peers, and participating in community events may promote health and well-being. Recently, interventions leveraging information and communications technology have emerged as potent tools for promoting social connections and well-being. Particularly, messenger apps have become an integral part of our daily lives, facilitating communication, information dissemination, and social interaction. However, there remains a gap in the literature regarding the utilization of widely adopted messenger apps for this purpose.

Objective: This study aimed to evaluate the impact of messenger app–based information provision aimed at promoting social participation on the enhancement of subjective well-being among Japanese community-dwelling adults.

Methods: A 2-arm, parallel-group randomized controlled trial was conducted from October 2022 to January 2023 in the Kashiwa-no-ha campus area, Japan—an urban community with active local events. A total of 358 community-dwelling adults who use messenger apps daily were recruited for the study. Of these, 235 (65.6%) participants completed the follow-up survey. Participants were randomly assigned to either the intervention group, receiving the health benefits of social participation and information about local events or spots via a messenger app, or the control group, receiving general health information. The primary outcome was subjective happiness after the intervention, measured on an 11-point scale ranging from 0 (Unhappy) to 10 (Happy). Secondary outcomes included life satisfaction, meaning of life, purpose in life, and participation in local events. The outcomes were analyzed with *t* tests (2-tailed) and multivariable regression based on the intention-to-treat method.

Results: After the intervention, the intervention group reported a mean happiness score of 7.7 (SD 1.7), while the control group reported a score of 7.5 (SD 2.0), with no statistically significant difference (P=.40). Multivariable linear regression analysis adjusted for baseline outcome values and covariates showed that the coefficient of the intervention for life satisfaction was 0.30 (95% CI –0.07 to 0.68; P=.12), while that for meaning of life was 0.33 (95% CI –0.03 to 0.70; P=.07). There was no significant difference in event participation rates between the 2 groups during the study period (P=.22). However, 82.2% (102/124) of the intervention group acknowledged the utility of the event information provided.

Conclusions: Messenger app–based information provision did not yield a significant increase in subjective happiness, while there was a positive but not significant trend in life satisfaction. The findings underscore the need for more intensive intervention in future studies to harness the potential of digital interventions.

Trial Registration: UMIN Clinical Trials Registry UMIN000049047; https://tinyurl.com/2zzrrae8

(J Med Internet Res 2024;26:e57205) doi:10.2196/57205

XSL•FO
RenderX

KEYWORDS

messenger app; event information; happiness; social participation; messenger; app; well-being; adults; aging; randomized controlled trial; RCT; information technology; social activity; Japan; urban community; health information; control group; multivariable regression; life satisfaction; digital intervention; community-dwelling

Introduction

Happiness, an aspect of subjective well-being, has become a major topic in public health over the past decades [1]. It includes not only momentary joy but also the overall sense of positive effect that individuals experience in their lives. Recent research has underscored the influence of well-being, including happiness, on both mental and physical health outcomes [2,3], making it a public health concern. An example of this is the World Happiness Report—an annual publication by the Sustainable Development Solutions Network that ranks countries based on their citizens' self-reported well-being [4]. This report not only emphasizes the importance of happiness as a global metric but also encourages nations to prioritize well-being in their policies and agendas. The prominence of such a report on the global stage underscores the growing recognition of happiness as a crucial component in assessing a nation's health and overall success.

Engaging in social activities, interacting with peers, and participating in community events may promote multidimensional health and well-being [5-8], including subjective happiness [9,10]. Thus, the promotion of social connections can be a strategy to enhance the subjective happiness of populations.

Recently, interventions leveraging information and communications technology have emerged as potent tools to promote social connections [11-13]. Mobile apps, in particular, are widely accessible and commonly tool, making them a potential candidate for information and communications technology–based interventions. For example, a mobile app designed and implemented in the municipality of Enschede, the Netherlands, aimed to promote social participation among older adults within the community [14]. The app provided older individuals with access to a home page comprising 6 distinct features: an inbox for messaging, news updates, activity listings, informative resources, helpful tips, and friend connections. Notably, the app demonstrated favorable usability among the older population and positively changed the mental component of their quality of life. However, it is important to acknowledge the limitations inherent in the aforementioned study, as it was observational rather than a randomized controlled trial (RCT), which may have affected the strength of the conclusions drawn.

Messenger apps also have potential in the abovementioned context [15,16]. They have become an integral part of our daily lives, facilitating communication, information dissemination, and social interaction. These platforms, with their widespread reach and user-friendly interfaces, present a unique opportunity for public health interventions. LINE (LY Corporation), for example, is a messaging app originally developed in Japan, offering a range of communication tools, including SMS text messaging, voice and video calls, and group chats. Beyond its primary function as a communication platform, LINE has expanded its services to include features tailored to businesses. One notable feature is the "LINE Business Account," which allows businesses to disseminate information, promote products or services, and engage with users directly through the app. Given LINE's immense popularity in Japan [17], with over 95 million users, numerous businesses use the platform for promotional activities, while local governments harness its reach for public service announcements, community engagement, and other civic initiatives.

Despite the potential of messenger app interventions in promoting social connections and, in turn, subjective happiness, there remains a gap in the literature regarding the utilization of widely adopted platforms, such as LINE, for happiness promotion in the Japanese context. Although LINE's vast user base and multifunctionality make it a promising tool for public health initiatives, its potential in disseminating information regarding local events and fostering social connections to promote subjective happiness has not been extensively explored. The aim of this study was to investigate whether information provision via LINE, pertaining to local events and recommended spots emphasizing the importance of social connections, can promote subjective happiness. We conducted an RCT to examine whether leveraging a popular messaging platform can enhance subjective happiness among urban Japanese residents.

Methods

Trial Design

This trial was an RCT with parallel groups, conducted in the Kashiwa-no-ha campus area, Kashiwa City, Japan, from October 14, 2022, to January 31, 2023. Eligible participants were randomly assigned (allocation ratio 1:1) to the intervention group or control group. Results from the trial are reported in line with the CONSORT-EHEALTH guidelines [18] (Multimedia Appendix 1).

Study Area

The Kashiwa-no-ha campus area is located within an hour of Tokyo. This area represents a blend of modern urban planning and community-centric development, frequently hosting diverse local events that attract residents and tourists alike. However, this area faces challenges in terms of information dissemination. Multiple sectors, from educational institutions to local governments, independently release event details, leading to fragmented communication. This decentralization complicates navigation for residents, highlighting the need for a unified platform to centralize event and community information.

Sample Size

To determine the sample size, power analysis was conducted using GPower (version 3.1; Heinrich Heine University Düsseldorf), and the estimated effect size was based on the results of a similar study [19]. The required sample size, with an effect size of 0.40 on happiness and an α level of 0.05, was

XSL•FO
RenderX

determined to be 200. Accounting for a 20% dropout rate, the plan was to achieve a final sample comprising 240 participants.

Participants and Procedure

The LINE messaging app, launched in 2011, is of Japanese origin. Beyond SMS text messaging, it also offers voice calls, video calls, and official accounts. We created the official account "Shiawase (ie, happiness) City Kashiwa-no-ha" to navigate study participation and provide information to the intervention and control groups (Figure S1 in Multimedia Appendix 2).

Participants were eligible for the study if they lived or commuted to work in the Kashiwa-no-ha campus area and (1) were men and women aged over 18 years old, (2) were "LINE" mobile app users, and (3) had given consent for this study. Eligible participants were recruited via the posting of flyers seeking their participation in the project. We sent email invitations to various membership groups, in addition to holding study information sessions (both web based and in person) to explain the study protocols and procedures. The registration period was 1 month from the start of recruitment, and prospective participants were asked to register a "Shiawase (ie, happiness) City Kashiwa-no-ha" account for the study on the "LINE" mobile app through their smartphones. After registration of the account, participants were asked to tap a panel on their smartphone screen to access an external website, on which they read the explanations of the study, gave informed consent, and responded to a baseline survey on Google Forms.

Of the 376 eligible participants, 358 (95.2%) completed the baseline questionnaire on Google Forms, and 18 (4.8%) were excluded, mainly as a result of declining to participate. Following baseline evaluation, simple randomization was conducted, before, subsequently, the participants received information according to their assigned group for the 2-month period. After the intervention period, the participants were invited to respond to a follow-up survey on Google Forms.

Randomization and Blinding

Computer-generated randomization was conducted on the LINE messaging platform after participants completed the baseline assessment. Participants were then automatically allocated to chat areas for the intervention and control groups, respectively (see Figure S1C in Multimedia Appendix 2 for the chat area for the intervention group). The study investigators were blinded to group allocation.

Intervention Group

Participants who had been allocated to the intervention group received, for 2 months, information regarding local events (eg, hobby groups) and spots (eg, shops, parks) in the target area (Kashiwa-no-ha), as well as short articles describing the health benefits of social participation on LINE (Figure S1 in Multimedia Appendix 2). Participants also received event information once a week and short articles or recommended places to go 1-2 times a week. Event information was aggregated from the websites of commercial facilities and community websites in the Kashiwa-no-ha campus area and was updated every Monday. The events were categorized into (1) hobbies and volunteer activities, (2) health and sports activities, and (3) childcare and pets. The recommended spot information was based on the results of our pilot survey with Kashiwa-no-ha campus area residents, in which the residents provided their recommendations for indoor or outdoor spots (eg, parks and commercial facilities, etc). Short articles disseminated the findings on the importance of social participation in Japan [20,21]. To allow access to the details of this information, an external link was attached, navigating the participants to a web page that was available only to the intervention group. They could also visit the website by tapping a panel on their smartphone screens. The information was updated by NA.

Control Group

Participants allocated to the control group received, for 2 months, general health information that did not include information about social participation. Participants received this information 2-3 times per week. The health information was presented in the form of headings from the Japanese Ministry of Health, Labour and Welfare's website (e-health net [22]), with external links attached. The external links were on a web page that was accessible only to the control group. They could also visit the website by tapping a panel on their smartphone's screen. The information was updated by NA.

Measures

Primary Outcome

The primary outcome was subjective happiness (hedonic well-being), measured on an 11-point scale ranging from 0 (Unhappy) to 10 (Happy), during the follow-up assessment [1]. We used the following question: "In general, how happy do you usually feel?"

Secondary Outcomes

Our secondary outcomes were life satisfaction (evaluative well-being), meaning of life, and purpose in life (eudaimonic well-being) at the follow-up assessment [1]. We used the following questions or statements: "Overall, how satisfied are you with life as a whole these days? (0=Not satisfied at all, 10=Completely satisfied)," "Overall, to what extent do you feel the things you do in your life are worthwhile? (0=Not at all worthwhile, 10=Completely worthwhile)," and "I understand my purpose in life (0=Strongly disagree, 10=Strongly agree)." Changes in well-being measures between the baseline and follow-up assessments were also secondary outcomes.

In addition, we assessed the effectiveness of the intervention in promoting participation in local events. Participants were asked, as part of the follow-up assessment, about their attendance at events in Kashiwa-no-ha during the study period.

Covariates

We assessed age, gender, self-rated physical health, self-rated mental health, financial stability, and material stability as covariates in the additional analysis [23]. All variables were based on Google Forms questionnaires. We asked the following questions: "In general, how would you rate your physical health? (0=Poor, 10=Excellent)," "How would you rate your overall mental health? (0=Poor, 10=Excellent)," "How often do you worry about being able to meet normal monthly living expenses? (0=Worry all of the time, 10=Do not ever worry)," and "How

XSL•FO
RenderX

often do you worry about safety, food, or housing? (0=Worry all of the time, 10=Do not ever worry).”

Feedback on Information Utility

After the intervention period, participants in the intervention group were asked to provide feedback on the utility of the information they received via LINE. Specifically, they were asked to choose the useful information, if any existed: local events, recommended spots, and the short articles about social connection and health.

Incentives

The participants received an incentive of JP \1000 in e-money (approximately US $7) to complete both the baseline and follow-up questionnaires. Participants could choose to receive their e-money from Amazon, LINE Pay, PayPay Point, Google Pay, and so on.

Statistical Analysis

Analyses were performed on an intention-to-treat basis among people who responded to the follow-up questionnaire. Means and SDs for the continuous variables and frequencies for the categorical variables were presented. We used a Student *t* test to compare the postintervention outcomes and changes in the outcomes from baseline to the postintervention time point between the control and intervention groups (significance level: $P \leq .05$). In addition, we conducted multivariable linear regression in which the outcomes were postintervention values. In model 1, we adjusted for each baseline outcome (subjective happiness, life satisfaction, meaning of life, and purpose in life). In model 2, age, gender, self-rated physical health, self-rated mental health, financial stability, and material stability were added. All analyses were performed using STATA MP (version 18.0; Stata Corp).

Ethical Considerations

Ethical approval for this study was obtained from the Ethics Committee at Chiba University (approval M10428). Informed consent was obtained from each participant. The protocol was registered in the UMIN Clinical Trials Registry with the identifier 000049047. Informed consent was obtained individually online before study participation, adhering to the principles of the Declaration of Helsinki. All data were deidentified to protect the privacy and confidentiality of participants. Participants received JP \1000 in e-money (approximately US $7) as compensation to complete the baseline and follow-up questionnaires.

Results

A total of 235 participants responded to the follow-up survey: 124 (52.8%) from the intervention group (56/180, 31.1% dropout rate) and 111 (47.2%) from the control group (67/178, 37.6% dropout rate; Figure 1). Table 1 summarizes the baseline characteristics of the participants. The mean age was 47.8 (SD 13.8) years, with 63.8% (150/235) of the sample comprising women. There were no significant differences in baseline characteristics and outcomes (all $P>.05$).

There was no difference in subjective happiness between the intervention group (mean 7.7, SD 1.7) and control group (mean 7.5, SD 2.0) after the 2-month intervention period ($P=.40$; Table 2). We also found no significant differences in life satisfaction, meaning of life, and purpose in life between the 2 groups (all $P>.05$).

Multivariable linear regression analysis, adjusted for baseline outcome values and covariates, showed that the coefficient of the intervention for life satisfaction was 0.30 (95% CI –0.07 to 0.68; $P=.12$) and that for meaning of life was 0.33 (95% CI –0.03 to 0.70; $P=.07$; Table S1 in the Multimedia Appendix 2). Our additional analysis, in which outcomes represented changes in well-being between baseline and follow-up, revealed similar trends (life satisfaction: intervention, mean 0.13, SD 1.54 vs control, mean –0.24, SD 1.77, $P=.09$; meaning of life: intervention, mean –0.20, SD 1.46 vs control, mean –0.51, SD 1.58, $P=.12$; Table S2 in the Multimedia Appendix 2).

There was no difference regarding participation in events in Kashiwa-no-ha during the study period between the intervention group (48/124, 38.7%) and the control group (52/111, 46.8%; $P=.21$). Most of the intervention group (102/124, 82.2%) found the event information useful, while only 11.3% (14/124) considered the short articles related to social connection and health useful (Table S3 in the Multimedia Appendix 2).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for study participants.

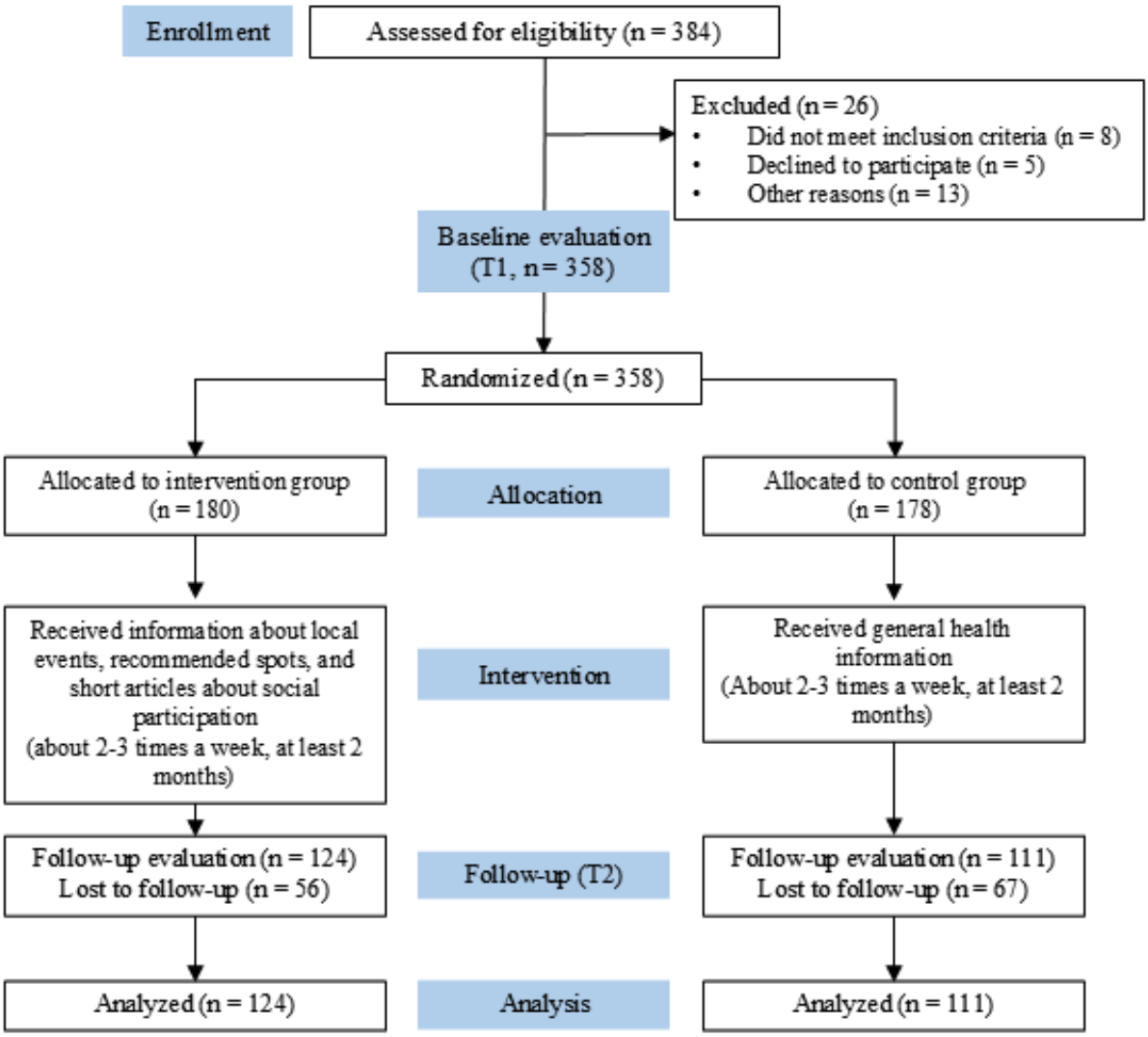

Table 1. Participants' characteristics at baseline.

Characteristics	Total (N=235)	Intervention (n=124)	Control (n=111)	*P* value
Age (years), mean (SD)	47.8 (13.8)	47.8 (13.7)	47.9 (14.0)	.96
Gender (women), n (%)	150 (63.8)	79 (63.7)	71 (64.0)	.89
Physical health (0-10), mean (SD)	7.3 (1.8)	7.3 (1.8)	7.2 (1.8)	.62
Mental health (0-10), mean (SD)	7.3 (1.8)	7.2 (1.8)	7.3 (1.8)	.85
Financial stability (0-10), mean (SD)	6.2 (2)	6.1 (2.2)	6.3 (1.7)	.54
Material stability (0-10), mean (SD)	6 (2)	6.1 (2.2)	5.8 (1.8)	.33
Happiness (0-10), mean (SD)	7.8 (1.5)	7.9 (1.5)	7.7 (1.5)	.48
Life satisfaction (0-10), mean (SD)	7.6 (1.6)	7.5 (1.6)	7.6 (1.5)	.52
Meaning of life (0-10), mean (SD)	7.7 (1.7)	7.7 (1.8)	7.6 (1.7)	.81
Purpose in life (0-10), mean (SD)	7 (2.5)	6.9 (2.8)	7.1 (2.3)	.63

XSL•FO
RenderX

Table 2. Well-being at follow-up and Cohen d for between-group effect sizes.

	Intervention, mean (SD)	Control, mean (SD)	Cohen *d*	*P* value
Happiness	7.7 (1.7)	7.5 (2)	0.11	.40
Life satisfaction	7.6 (1.7)	7.4 (1.8)	0.14	.29
Meaning of life	7.5 (1.7)	7.1 (2)	0.20	.13
Purpose in life	6.4 (2.3)	6.3 (2.2)	0.03	.79

Discussion

Principal Findings

This study explored the efficacy of providing aimed at promoting social participation on the LINE messaging platform, emphasizing the importance of social connections for the promotion of happiness. Our findings suggest that the intervention did not significantly enhance subjective happiness and local events participation. While care should be taken in interpretation, there was a slight positive trend in life satisfaction and meaning of life among the intervention group compared with the control group, but this was not statistically significant.

The absence of a significant difference in subjective happiness between the intervention and control groups may be due to the lack of notable changes in event attendance between the 2 groups. This finding did not align with some studies showing the effectiveness of social participation apps [14,24]. Several factors could explain this finding. First, merely providing information might not have been sufficient to motivate individuals to participate in community events. Previous studies used some other factors, such as message boards and matching, which may promote communication and participation [14,24]. Second, the study did not comprehensively cover events and spots, which could have diluted its potential impact. Third, it is plausible that participants, given their proactive stance on improving their communities, as evidenced by their participation in this study, had already attended local events. The high local event participation rates (41.7% in the intervention group and 50% in the control group) further support this idea. The aforementioned pre-existing engagement may have attenuated the potential impact of the intervention.

Changing health behavior is a complex endeavor, often requiring deep-rooted shifts in mindset and habits. Many studies have harnessed theories, such as self-determination theory [25], to bolster motivation and drive behavioral change. However, promoting social participation presents unique challenges distinct from those seen with traditional health behaviors because the pro-health impacts of social participation are not common in the general population. Additionally, this study revealed that its participants did not highly value short articles linking social participation with health, thus indicating a potential disconnect in recognizing the health benefits of social engagement. Although effective strategies in health behavior change often employ features such as goal setting and feedback [26], these may not be directly translatable to the promotion of social participation. Specifically, the utility of goal setting and feedback diminishes if individuals do not first acknowledge and value the importance of social participation for their health. Additional motivational factors or behavioral nudges may be necessary to promote social engagement as a health behavior [27].

The observed trend of improvement in life satisfaction and the meaning of life, but not in happiness, among members of the intervention group offers an intriguing insight into the multifaceted nature of subjective well-being [1]. While happiness captures momentary emotional states influenced by daily experiences, life satisfaction and the meaning of life reflect more enduring aspects of well-being. Despite the intervention's emphasis on local events and the importance of social connections, it did not lead to an immediate increase in event attendance. The aforementioned finding suggests that the intervention may have influenced participants' cognitive perceptions and evaluations without necessarily prompting an actual increase in social events attendance. This aligns with social contagion, demonstrating that attitudes, behaviors, and emotions spread in a network through the diffusion of information or the transmission of behavioral norms [28]. The consistent exposure to community opportunities, even if not directly acted upon, may have nurtured, within participants, a heightened sense of belonging or a feeling of being more integrated and informed about their community. This could, in turn, have contributed to a more favorable overall judgment of their lives, enhancing life satisfaction and the meaning of life, without altering momentary feelings of happiness. However, given these less robust findings, further research is needed to delve deeper into the differential impacts on various facets of subjective well-being and to understand the underlying mechanisms.

The fact that the majority of the intervention group members found the event information useful suggests the potential impact of tailored community information delivered through familiar platforms such as LINE in Japan. This positive reception indicates that the participants valued being informed about local events and opportunities, suggesting a latent desire for community engagement and connection. Leveraging popular communication tools to bridge the gap between residents and their local communities may drive more active participation in future initiatives and enhance well-being.

Limitation

This study has several limitations that should be considered when interpreting the findings. First, the intervention's duration (2 months) might have been too short to observe significant behavioral changes, such as increased event attendance. Second, the sample might have been skewed toward individuals who are already proactive regarding community engagement, given their residence in the area and willingness to participate in the study, thus potentially limiting the generalizability of the

XSL•FO
RenderX

findings to a broader population. However, it is worth noting that those who did not participate in the study, potentially being less proactive, might have more room for improvement in terms of community engagement and well-being. The aforementioned suggests that future interventions could target this less-engaged demographic, as they might benefit more substantially from initiatives promoting community involvement and social connections. Third, the participants were relatively young and expected to be currently employed. Some of them might not have been able to attend the events. Fourth, we could not identify the actual event participation, merely relying on the responses to the questionnaire at follow-up assessment. Fifth, the coverage of the local information provided might not have been comprehensive, giving rise to the possibility of missing out on events or opportunities that could have been of interest to participants. It is essential to recognize that individuals are more likely to use and trust an information resource if they perceive it to be comprehensive and all-encompassing. Incomplete information might deter consistent engagement, emphasizing the need for future interventions to ensure thoroughness in information dissemination to maximize user trust and utilization. Sixth, we were unable to prevent control group participants from independently browsing information on the health benefits of social participation or local events. Seventh, we did not consider employment and membership with organizations in the analysis. Although this study is an RCT, these factors can affect the results and should be included in multivariable analysis. Finally, this study had a notable dropout rate. If participants in the intervention group who found the intervention beneficial were more likely to complete the follow-up survey, this could have led to an overestimation of the intervention's effectiveness.

Conclusions

In summary, while our study did not find that messenger app–based information provision impacted subjective happiness, the potential positive trend in life satisfaction and meaning of life warrants further exploration. Future research should consider longer intervention durations, the incorporation of behavioral change techniques, and comprehensive information on local events and spots to maximize the benefits of such digital interventions in promoting social participation and well-being.

Acknowledgments

NTT DOCOMO, INC developed the platform of the LINE official account "Shiawase (ie, happiness) City Kashiwa-no-ha." This study was supported by the Open Innovation Platform with Enterprises, Research Institute and Academia (OPERA, JPMJOP1831) from Japan Science and Technology (JST) and by Mitsui Fudosan Co., Ltd. OPERA had no role in the design of the study; in the recruitment; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The Mitsui Fudosan Co., Ltd. assisted in the recruitment of study participants by supplying local information and distributing brochures at various local venues such as events, residential areas, and banks. However, the sponsor had no involvement in the design of the study; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AN and NA initiated and conceptualized the study and drafted the manuscript with initial feedback from all authors. NA, YC, and KI collected the data, with supervision from AN and KK. NA and AN contributed to data preparation and analyzed the data. All authors contributed to the review and editing process of this manuscript.

Conflicts of Interest

KK has received research funding from Mitsui Fudosan Co., Ltd.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 48074 KB - jmir_v26i1e57205_app1.pdf]

Multimedia Appendix 2
Supplementary data.
[DOCX File , 645 KB - jmir_v26i1e57205_app2.docx]

References

1. Lee MT, Kubzansky LD, VanderWeele TJ. Measuring Well-Being: Interdisciplinary Perspectives From the Social Sciences and the Humanities. Oxford, United Kingdom: Oxford University Press; 2021.

XSL•FO
RenderX

2. Diener E, Chan M. Happy people live longerubjective well-being contributes to health and longevity. Appl Psychol Health Well-Being 2011;3(1):1-43. [doi: 10.1111/j.1758-0854.2010.01045.x]
3. Steptoe A. Happiness and Health. Annu Rev Public Health 2019 May 01;40:339-359. [doi: 10.1146/annurev-publhealth-040218-044150] [Medline: 30601719]
4. Helliwell JF, Layard R, Sachs JD, Aknin LB, de Neve JE, Wang S. World Happiness Report 2023. New York, NY: Sustainable Development Solutions Network; 2023. URL: https://worldhappiness.report/ [accessed 2023-09-23]
5. Ngamaba KH. Determinants of subjective well-being in representative samples of nations. Eur J Public Health 2017 May 01;27(2):377-382. [doi: 10.1093/eurpub/ckw103] [Medline: 27426951]
6. Dawson-Townsend K. Social participation patterns and their associations with health and well-being for older adults. SSM Popul Health 2019 Aug;8:100424 [FREE Full text] [doi: 10.1016/j.ssmph.2019.100424] [Medline: 31312713]
7. Nakagomi A, Tsuji T, Saito M, Ide K, Kondo K, Shiba K. Social isolation and subsequent health and well-being in older adults: a longitudinal outcome-wide analysis. Soc Sci Med 2023 Jul;327:115937. [doi: 10.1016/j.socscimed.2023.115937] [Medline: 37167920]
8. National Academies of Sciences, Engineering, and Medicine. Social Isolation and Loneliness in Older Adults Opportunities for the Health Care System. Washington, DC: The National Academies Press; 2020.
9. Golden J, Conroy RM, Bruce I, Denihan A, Greene E, Kirby M, et al. Loneliness, social support networks, mood and wellbeing in community-dwelling elderly. Int J Geriatr Psychiatry 2009 Jul;24(7):694-700. [doi: 10.1002/gps.2181] [Medline: 19274642]
10. Ide K, Jeong S, Tsuji T, Watanabe R, Miyaguni Y, Nakamura H, et al. Suggesting indicators of age-friendly city: social participation and happiness, an ecological study from the JAGES. Int J Environ Res Public Health 2022;19(9):5096 [FREE Full text] [doi: 10.3390/ijerph19095096] [Medline: 35564490]
11. Nakagomi A, Shiba K, Kawachi I, Ide K, Nagamine Y, Kondo N, et al. Internet use and subsequent health and well-being in older adults: an outcome-wide analysis. Comput Hum Behav 2022 May;130:107156 [FREE Full text] [doi: 10.1016/j.chb.2021.107156]
12. Judges RA, Laanemets C, Stern A, Baecker RM. “InTouch” with seniors: exploring adoption of a simplified interface for social communication and related socioemotional outcomes. Comput Hum Behav 2017 Oct;75:912-921. [doi: 10.1016/j.chb.2017.07.004]
13. Thiel C, Günther L, Osterhoff A, Sommer S, Grüneberg C. Feasibility of smartphone-supported, combined physical and cognitive activities in the neighbourhood for stimulating social participation of the elderly. BMC Geriatr 2022 Jul 30;22(1):629 [FREE Full text] [doi: 10.1186/s12877-022-03303-0] [Medline: 35907804]
14. Jansen-Kosterink SM, Bergsma J, Francissen A, Naafs A. The first evaluation of a mobile application to encourage social participation for community-dwelling older adults. Health Technol 2020 May 02;10(5):1107-1113. [doi: 10.1007/s12553-020-00430-9]
15. Ota Y, Nomura K, Fujita N, Suzuki T, Kamatsuka M, Sakata N, et al. Influence of LINE-assisted provision of information about human papillomavirus and cervical cancer prevention on HPV vaccine intention: a randomized controlled trial. Vaccines (Basel) 2022 Dec 24;10(12):2005 [FREE Full text] [doi: 10.3390/vaccines10122005] [Medline: 36560414]
16. Chen YRR, Schulz PJ. The effect of information communication technology interventions on reducing social isolation in the elderly: a systematic review. J Med Internet Res 2016 Jan 28;18(1):e18 [FREE Full text] [doi: 10.2196/jmir.4596] [Medline: 26822073]
17. Why is LINE the most popular social media app in Japan? Digital Marketing for Asia. URL: https://www.digitalmarketingforasia.com/why-line-is-the-most-popular-social-media-app-in-japan/ [accessed 2024-10-18]
18. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
19. Schotanus-Dijkstra M, Drossaert CHC, Pieterse ME, Boon B, Walburg JA, Bohlmeijer ET. An early intervention to promote well-being and flourishing and reduce anxiety and depression: a randomized controlled trial. Internet Interv 2017 Oct;9:15-24 [FREE Full text] [doi: 10.1016/j.invent.2017.04.002] [Medline: 30135833]
20. Yazawa A, Inoue Y, Fujiwara T, Stickley A, Shirai K, Amemiya A, et al. Association between social participation and hypertension among older people in Japan: the JAGES Study. Hypertens Res 2016 Dec;39(11):818-824. [doi: 10.1038/hr.2016.78] [Medline: 27383510]
21. Nakagomi A, Tsuji T, Hanazato M, Kobayashi Y, Kondo K. Association between community-level social participation and self-reported hypertension in older Japanese: a JAGES multilevel cross-sectional study. Am J Hypertens 2019 May 22;32(5):503-514. [doi: 10.1093/ajh/hpz028] [Medline: 30793745]
22. Japanese Ministry of Health, Labour and Welfare. URL: https://www.e-healthnet.mhlw.go.jp/ [accessed 2024-10-18]
23. VanderWeele TJ. On the promotion of human flourishing. Proc Natl Acad Sci USA 2017;114(31):8148-8156 [FREE Full text] [doi: 10.1073/pnas.1702996114] [Medline: 28705870]
24. Hsu PT, Ho CS, Ho YF, Chen JJ, Chen IJ. The effects of a social participation app on seniors. J Nurs Res 2021 Aug 31;29(5):e168. [doi: 10.1097/JNR.0000000000000451] [Medline: 34469399]

XSL•FO
RenderX

25. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
26. Xu L, Shi H, Shen M, Ni Y, Zhang X, Pang Y, et al. The effects of mHealth-based gamification interventions on participation in physical activity: systematic review. JMIR Mhealth Uhealth 2022 Mar 03;10(2):e27794 [FREE Full text] [doi: 10.2196/27794] [Medline: 35113034]
27. Arno A, Thomas S. The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis. BMC Public Health 2016 Jul 30;16:676 [FREE Full text] [doi: 10.1186/s12889-016-3272-x] [Medline: 27475752]
28. Berkman LF, Kawachi I, Glymour MM. Social Epidemiology. 2nd ed. New York, NY: Oxford University Press; 2014.

Abbreviations

RCT: randomized controlled study

Edited by T de Azevedo Cardoso; submitted 07.02.24; peer-reviewed by A Garcia, AKC Wong; comments to author 17.07.24; revised version received 02.08.24; accepted 07.10.24; published 29.11.24.

Please cite as:
Nakagomi A, Abe N, Chen YR, Ide K, Kobayashi S, Hanazato M, Kondo K
Messenger App–Based Information Provision for Promoting Social Participation to Enhance Well-Being Among Community-Dwelling Adults: Randomized Controlled Trial
J Med Internet Res 2024;26:e57205
URL: https://www.jmir.org/2024/1/e57205
doi:10.2196/57205
PMID:39612209

©Atsushi Nakagomi, Noriyuki Abe, Yu-Ru Chen, Kazushige Ide, Shuhei Kobayashi, Masamichi Hanazato, Katsunori Kondo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Biomarkers of Response to Internet-Based Psychological Interventions: Systematic Review

Giulia Gotti[1*], MS; Chiara Gabelli[1*], MS; Sophia Russotto[2], MD; Fabio Madeddu[1], MD, PhD; Philippe Courtet[3,4,5], MD, PhD; Jorge Lopez-Castroman[6,7,8], MD, PhD; Patrizia Zeppegno[9,10], MD, PhD; Carla Maria Gramaglia[9,10*], MD, PhD; Raffaella Calati[1,6*], PsyD, PhD

[1]Department of Psychology, University of Milan-Bicocca, Milan, Italy
[2]Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy
[3]PSNREC, University of Montpellier, INSERM, Centre Hospitalier Universitaire de Montpellier, Montpellier, France
[4]Department of Emergency Psychiatry and Acute Care, Lapeyronie Hospital, Centre Hospitalier Universitaire de Montpellier, Montpellier, France
[5]FondaMental Fondation, Créteil, France
[6]Centro de Investigación Biomédica en Red de Salud Mental, Madrid, Spain
[7]Department of Adult Psychiatry, Nimes University Hospital, Nimes, France
[8]Department of Psychiatry, Radiology, Public Health, Nursing and Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain
[9]Institute of Psychiatry, University of Piemonte Orientale, Novara, Italy
[10]Struttura Complessa Psichiatria, Azienda Ospedaliero-Universitaria Maggiore della Carità, Novara, Italy
[*]these authors contributed equally

Corresponding Author:
Raffaella Calati, PsyD, PhD
Department of Psychology
University of Milan-Bicocca
Piazza dell'Ateneo Nuovo 1
U6 Building, Room 3129
Milan, 20126
Italy
Phone: 39 3477752087
Email: raffaella.calati@unimib.it

Abstract

Background: Internet-based psychological interventions provide accessible care to a wide range of users, overcoming some obstacles—such as distance, costs, and safety—that might discourage seeking help for mental issues. It is well known that psychological treatments and programs affect the body, as well as the mind, producing physiological changes that ought to be considered when assessing the efficacy of the intervention. However, the literature investigating changes in biomarkers specifically after internet-based psychological and mental health interventions has not yet extensively inquired into this topic.

Objective: This systematic review aims to provide a synthesis of literature examining the effects of internet-based psychological interventions—targeting both clinical (mental and physical) and nonclinical conditions—on biomarkers. A secondary aim was to evaluate whether the biomarkers' variations were related to a complementary modification of the psychological or physical symptoms or to a general improvement of the participants' well-being.

Methods: This review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement. A literature search was performed through 3 databases (PubMed, PsycINFO, and Scopus). Studies examining changes in biomarkers before and after internet-based psychological interventions or programs targeting both clinical and nonclinical samples were included, with no exclusion criteria concerning mental or physical conditions.

Results: A total of 24 studies fulfilled the inclusion criteria. These studies involved individuals with psychiatric or psychological problems (n=6, 25%), those with organic or medical diseases (n=10, 42%), and nonclinical populations (n=8, 33%). Concerning psychiatric or psychological problems, cognitive behavioral therapy (CBT) and CBT-informed interventions showed partial effectiveness in decreasing glycated hemoglobin blood glucose level (n=1) and chemokines (n=1) and in increasing connectivity between the default-mode network and the premotor or dorsolateral prefrontal cortex (n=1). Among individuals with organic or medical diseases, studies reported a significant change in cardiac or cardiovascular (n=3), inflammatory (n=2), cortisol (n=2),

XSL•FO
RenderX

glycated hemoglobin (n=2), and immune response (n=1) biomarkers after CBT and CBT-informed interventions, and mindfulness and stress management interventions. Lastly, mindfulness, CBT and CBT-informed interventions, and music therapy succeeded in modifying immune response (n=2), cortisol (n=1), α amylase (n=1), posterior cingulate cortex reactivity to smoking cues (n=1), and carbon monoxide (n=1) levels in nonclinical populations. In some of the included studies (n=5), the psychological intervention or program also produced an improvement of the mental or physical condition of the participants or of their general well-being, alongside significant variations in biomarkers; CBT and CBT-informed interventions proved effective in reducing both psychological (n=2) and physical symptoms (n=2), while a mindfulness program successfully lowered cigarette consumption in a nonclinical sample (n=1).

Conclusions: Although further evidence is required, we hope to raise awareness on the potential impact of internet-based interventions on biomarkers related to mental and physical health.

(J Med Internet Res 2024;26:e55736) doi:10.2196/55736

KEYWORDS

biomarker; cognitive behavioral therapy; internet-based intervention; systematic review; psychological intervention; mental health intervention; meta analysis; psychiatric; blood glucose; mindfulness; stress management; immune response; smoking; cortisol

Introduction

Internet-Based Interventions

In the psychological and mental health field, a growing body of evidence has demonstrated telepsychiatry's ability to provide accessible, wide-ranging, and high-quality services. This allows us to overcome obstacles, such as distance, costs, and safety, making these methods preferable for users who live afar, have transportation difficulties, or have time limitations. It also fosters a sense of empowerment [1]. This explains that online approaches to psychological therapy might be more easily accepted than standard psychotherapy. Additionally, an online setting might create in the users an apparent sense of anonymity, promoting feelings of safety and self-disclosure and reducing the stigma typically associated with seeking mental health support [2-4]. Internet-based programs exhibit remarkable versatility, providing a diverse array of services to a broad spectrum of users. This includes both patients and healthy individuals across various age groups, as well as specific populations such as health care professionals [1].

On the other hand, online interventions might also carry some negative issues: one of the main disadvantages is the risk of potential breaches of privacy, confidentiality, and data security, likely due to the use of unsecured websites or easily hackable software [4]. They may also entail some inconveniences for both users and health care professionals and institutions, such as the need for adequate technological equipment and systems compatibility. Accordingly, cost-effectiveness is a relevant issue, and results from research on this topic are mixed, bringing evidence of both methodological flaws and similar—or in some cases, better—cost-effectiveness compared to traditional in-person therapy [1,4]. Furthermore, the remote nature of online interventions themselves leads to a higher risk of low adherence and greater dropout rates [3,4].

Biomarkers

According to the National Institutes of Health (NIH), a biomarker is "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions. Biomarkers may include molecular, histologic, radiographic, or physiologic characteristics" [5]. The US Food and Drug Administration (FDA)–NIH Biomarker Working Group defined a taxonomy of biomarkers, distinguishing diagnostic, monitoring, response, predictive, prognostic, safety, and susceptibility or risk biomarkers [6]. Response biomarkers, in particular, are used to evaluate change in a biological response, which can potentially be beneficial or harmful, in an individual after being exposed to a medical condition, a clinical intervention (including drug treatments), or an environmental agent.

Within therapeutic interventions, examining potential variations in biomarker levels before and after treatment proves to be an effective method for objectively assessing the treatment's effectiveness. Additionally, it facilitates a deeper understanding of whether these biological modifications align with concurrent changes in the symptoms targeted by the intervention. Mental and physical health are strictly linked and they profoundly impact one another [7]. Thus, psychological treatments and programs can also affect the body, as well as the mind, with beneficial effects on stress, the immune system, and brain activity [8].

A vast body of literature shows changes in biomarkers following psychological interventions. For example, psychological interventions, such as cognitive behavioral therapy (CBT), consistently proved to be efficient in improving both the immune system [9] and neurological functioning biomarkers [10,11] across various mental and organic health conditions. Moreover, Claudino et al [10] found biomarker variations to be associated with symptoms reduction, allowing authors to consider them as a reliable way to assess treatment response [9]. Finally, similar results were reported not only for standard psychological interventions but also for various psychosocial programs, such as music-based psychosocial intervention, together with reductions in cortisol and inflammatory activity [12].

Nonetheless, there is scarce evidence of the efficacy of internet-based psychological interventions in changing biomarkers across different health conditions. Considering the benefits of online treatments and the rapid spreading of technologies in health care—especially following the COVID-19 pandemic [13]—a systematic review is needed to

XSL•FO
RenderX

comprehensively synthesize the current state of research, and it will help assess the efficacy of internet-based programs, compared to in-person ones.

Aim

The main objective of this systematic review was to investigate the effects of internet-based psychological interventions targeting every type of condition (mental or physical) on various biomarkers, assessed before and after the intervention itself. Additionally, we aimed to determine whether the biomarkers' variations were related to a complementary modification of the underlying mental or medical health condition, if present, or of the nonclinical issue addressed by the intervention or program.

The literature investigating changes in biomarkers after internet-based psychological and mental health interventions has not yet extensively inquired into this topic, and to the best of our knowledge, this is the first systematic review on this topic.

Methods

Procedure

The review was not registered. This systematic review was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement [14]. A literature search (on articles published until May 25, 2023) was independently performed by GG, CG, SR, and CMG using the PubMed, PsycINFO, and Scopus databases, with the search terms reported in Multimedia Appendix 1. The reference lists of the identified meta-analysis and reviews were also checked to find additional relevant articles. GG, CG, and SR independently evaluated the suitability of all the identified papers; in case of conflicting opinions regarding the inclusion or exclusion of a paper, GG, CG, and SR consulted with each other and with RC to find an agreement.

We included studies that investigated internet-based or app-based, remote, psychological or psychoeducational interventions or programs and that measured biomarkers before and after the intervention or program. Studies were included if:

1. They administered a psychological or psychoeducational intervention, or a multicomponent intervention based on a psychological approach.
2. At least half of the intervention or program had to be provided online or through a mobile phone app so that it could take place at the participant's home.
3. Biomarkers were assessed both before and after the intervention or program.
4. They targeted psychiatric, psychological, or physical conditions or nonclinical individuals (we considered "nonclinical" as those participants who were not recruited based on a specific psychiatric or organic illness and those interventions that did not specifically target a certain disease—in opposition to "psychiatric or psychological" and "physical" samples, whose participants had a mental or medical diagnosis).
5. They were written in English, French, or Italian.

Studies were excluded if:

1. They assessed biomarkers only before or only after the intervention or program.
2. They provided an internet-based psychological intervention or program without assessing any biomarker (eg, the study by ter Huurne et al [15]).
3. The intervention or program took place entirely in a prearranged setting, specifically ideated for this study, such as in-person, virtual reality (eg, the study by Yang et al [16]), and computer-assisted interventions.
4. They involved lifestyle interventions or programs (eg, physical activity or nutritional habits), without a psychological component (eg, the study by Carolan-Olah and Sayakhot [17]).
5. They conducted only cognitive trainings or tasks focused on executive functions, such as cognitive remediation therapy (CRT; eg, the study by Brockmeyer et al [18]). As CRT is fundamentally distinct in nature from psychological and psychoeducational interventions, adding CRT would have made it challenging to compare the treatments meaningfully. However, we included studies with cognitive or executive tasks whenever these were only a part of the intervention, along with psychotherapy sessions (eg, the study by Lansing et al [19]).
6. They were reviews, meta-analyses, protocols, and letters to the editor.

See the figure in the *Results* section for this study's flowchart.

We used Rayyan software to identify and exclude duplicates among the 3 databases.

From each included article, GG, CG, and SR independently extracted the main characteristics: study author, study title, study design, diagnosis, targeted population, medication, sample size, gender, age, intervention, duration of intervention, follow-up duration, and main results (Table S1 in Multimedia Appendix 2). GG, CG, and SR independently evaluated the strength of reporting with the Quality Assessment Tool for Before-After (Pre-Post) Studies with No Control Group, the Quality Assessment Tool for Case Series Studies, and the Quality Assessment of Controlled Intervention Studies [20].

Quality of Reporting

The strength of reporting was assessed with the Quality Assessment Tools developed by the National Heart, Lung, and Blood Institute tools [20]. The questions in these tools are designed to help the reviewer focus on the key concepts for assessing the internal validity of a study and concern the objective of this study, the population studied, how it was recruited and measured, the response rate, and the statistical analyses conducted (Tables S2-S4 in Multimedia Appendix 2 for details).

Quality ratings ("good," "fair," or "poor") were assigned based on the following calculation:

- Quality assessment of controlled intervention studies: 0-5 yeses: poor, 6-10 yeses: fair, and 11-14 yeses: good (Table S2 in Multimedia Appendix 2);
- Quality assessment tool for before-after (pre-post) studies with no control group: 0-4 yeses: poor, 5-8 yeses: fair, and

XSL•FO
RenderX

9-12 yeses: good (Table S3 in Multimedia Appendix 2); and

- Quality assessment tool for case series studies: 0-3 yeses: poor, 4-6 yeses: fair, and 7-9 yeses: good (Table S4 in Multimedia Appendix 2).

Results

General Description

After screening for titles, abstracts, and full texts, we eventually included 24 papers of the 2934 studies identified through the databases and the 21 studies identified through other sources (Figure 1).

The retained studies were either controlled trial intervention studies or randomized controlled trials (n=19, 79%) [21-39], pre-post studies with no control group (n=4, 17%) [19,40-42], or case studies (n=1, 4%) [43].

The interventions or programs were addressed to individuals with a psychiatric or psychological problem (n=6, 25%), individuals with an organic or medical disease (n=10, 42%), and nonclinical populations (n=8, 33%).

The psychiatric or psychological problems included depression (n=3, 13%) [21,22,41], posttraumatic stress disorder (n=2, 8%) [23,24], and sleep problems (n=1, 4%) [42].

The organic or medical diseases included type 1 diabetes (n=3, 13%) [19,25,43], inflammatory bowel disease (n=2, 8%) [26,27], hypertension (n=2, 8%) [28,29], prostate cancer (n=1, 4%) [30], preterm labor (n=1, 4%) [31], and heart diseases (n=1, 4%) [32].

The nonclinical populations included health care professionals (n=3, 13%) [33,34,40], smokers (n=2, 8%) [35,36], university students (n=2, 8%) [37,38], and healthy adult men (n=1, 4%) [39].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) study flowchart.

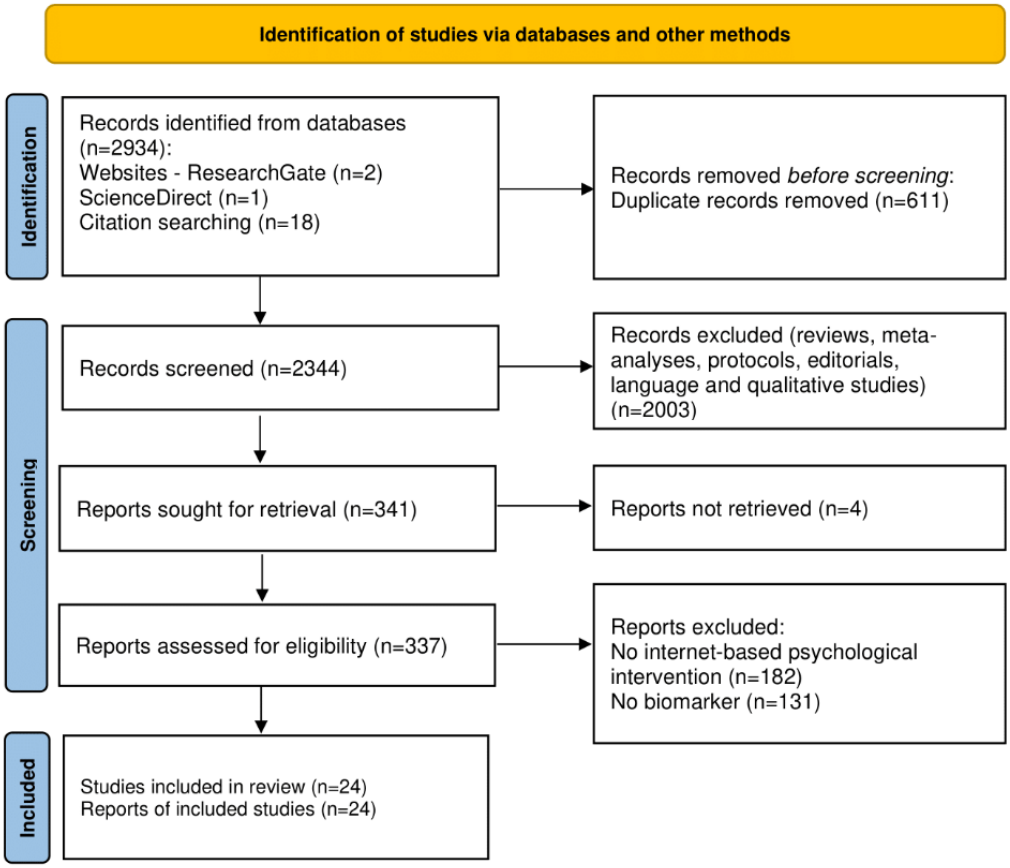

Two studies included patients with both a psychiatric or psychological problem and an organic or medical disease [21,42], and in this case, we included the studies under psychiatric or psychological problems.

The psychotherapeutic interventions or programs were CBT or CBT-informed interventions (we included in this category different types of CBTs, such as trauma-focused CBT, CBT for insomnia, behavioral activation treatment, and motivational enhancement therapy CBT, as well as programs and e-counseling based on CBT techniques; n=14, 59%) [19,21-24,28-31,36,39,41-43], mindfulness-based treatments (n=6, 25%) [26,32,34,35,37,40], stress management interventions (n=2, 8%) [27,33], coping skills training (n=1, 4%) [25], and music therapy (n=1, 4%) [38].

The most frequently measured biomarkers were salivary cortisol (sCort; n=7, 29%) [24,30,34,37,39-41], tumor necrosis factor-α (n=5, 21%) [27,30,34,39,42], α amylase (AA; n=4, 17%) [24,39-41], hair cortisol concentrations (HCC; n=4, 17%) [26,32,38,41], interleukin-6 (n=4, 17%) [27,30,34,42], glycated hemoglobin (HbA_{1c}) blood glucose level (n=4, 17%) [19,21,25,43], and C-reactive protein (CRP; n=4, 17%) [26,30,34,39]. We grouped the biomarkers measured on their

XSL•FO
RenderX

physiological function or meaning, namely stress-related biomarkers, immune response biomarkers, cardiac or cardiovascular biomarkers, electrophysiological biomarkers, inflammatory biomarkers, electrolytes, renal function biomarkers, neurological function biomarkers, and others. Table S5 in Multimedia Appendix 2 provides more details.

Results of the included studies are hereunder presented, divided by the underlying health condition of the targeted population (psychiatric or psychological illnesses, organic or medical diseases, and nonclinical samples), with a primary focus on biomarker variations between groups or in one group after intervention or program.

Interventions Targeting Psychiatric, Psychological Conditions, or Issues

Two studies found significant improvements in patients with depression after CBT and CBT-informed interventions, showing both reduced chemokines [22] and a decrease in HbA_{1c} blood glucose levels in older adults with type 2 diabetes [21]. In Romero-Sanchiz et al [22], all chemokines decreased in concentration at the study's end point; however, these differences were observed within participants with mild and moderate depression (mostly those who were medication-free) rather than between patients with severe depression and healthy controls [22]. Concurrently to immune response biomarkers variations, depression scores significantly improved after the internet CBT intervention, showing a relation between chemokine concentrations and depressive symptoms. In Egede et al [21], participants in the telehealth behavioral activation treatment arm showed stable levels of HbA_{1c}, as opposed to it increasing in the same-room arm [21]. Moreover, both the telehealth and the same-room treatments equally contributed to lower depression symptoms, with no significant differences between the two study arms.

CBT for insomnia led to significantly increased functional connectivity in bilateral posterior cingulate cortices (PCCs) and right premotor or dorsolateral prefrontal cortex in patients with insomnia [42]. Changes in cytokines levels post interventions were not significant. Participants also reported significant improvements in several sleep quality indices, which correlated with biomarker variations.

The remaining studies providing a CBT intervention to psychiatric samples, targeting patients with depression [41] and posttraumatic stress disorder [23,24], showed no significant changes in biomarkers, particularly stress-related biomarkers [24,41] and oxytocin and vasopressin concentrations [23].

Interventions Targeting Organic, Medical Conditions, or Issues

Blood pressure (BP) significantly decreased after CBT-informed e-counseling in patients with hypertension [28,29] and in patients with heart disease after a mindfulness-based intervention [32]. Specifically, in Liu et al [28], BP decreased equally in both the e-counseling group and the control group; at 12 months, the e-counseling group, compared to controls, showed a greater reduction in systolic BP (SBP), along with a significant reduction of urinary sodium in females. Similarly, Nolan et al [29] found a significant reduction of BP in both the e-counseling group and the control group, with the variation of SBP being significantly greater in the e-counseling group, compared to controls, at 12 months. Pulse pressure also lessened after 4 and 12 months, while non–high-density lipoprotein cholesterol lowered at 4 months in the e-counseling group. In Gotink et al [32], in patients with heart disease, heart rate (HR), BP, and HCC levels decreased over time, but these changes showed no significant difference between groups, although SBP reduction in the mindfulness group had a greater effect size in the mindfulness group compared to usual care.

HbA_{1c} blood glucose level significantly decreased in patients with type 1 diabetes in 2 studies after CBT interventions [19,43]. However, Kern et al [43] was a case series study on 2 patients and Lansing et al [19] had a sample of 15 patients only. According to Lansing et al [19], a motivational enhancement CBT-informed intervention would benefit glycemic control and HbA_{1c} in teenagers with diabetes [19]. Following the active treatment phase, the HbA_{1c} was indeed reduced compared to its pretreatment levels, whereas during the subsequent maintenance phase, no additional enhancements were observed. Nevertheless, the progress was preserved in most participants. On the contrary, HbA_{1c} slightly increased in youth with type 1 diabetes in one study [25], both in the coping skills training group and in the control one (with no difference between the two), but this variation was only borderline significant (P=.05). Although this does not reflect an improvement in metabolic control, the fact that HbA_{1c} did not notably rise in teenagers with diabetes was considered a positive outcome by the authors.

Indoleamine 2,3-dioxygenase significantly decreased in patients with inflammatory bowel disease after a stress management intervention [27]. However, the authors failed to find any change in the many other biomarkers investigated, including brain-derived neurotrophic factor and CRP.

Furthermore, reduced inflammation biomarkers were found in patients with inflammatory bowel disease after mindfulness-based interventions [26]. Participants who received both the blended mindfulness-based intervention and the standard medical therapy presented a significant decrease in fecal calprotectin and CRP compared to participants who received standard medical therapy alone, but there was no significant difference in HCC emerged between the two groups. Immune response biomarkers significantly decreased in men with advanced prostate cancer after 6 months, both in the cognitive behavioral stress management intervention and the health promotion condition, with no significant difference between groups [30]. However, this improvement faced a rebound increase at 6 months. Furthermore, contrary to the authors' hypothesis, diurnal cortisol diminished only for participants in the health promotion condition after 6 months, while it was higher in men in the cognitive behavioral stress management group. Cortisol levels remained relatively stable over the subsequent 6 months.

Lastly, no significant difference concerning the reduction of sCort levels was found between pregnant women with preterm labor assigned to either an internet-based cognitive behavioral

XSL•FO
RenderX

stress management program for preterm labor or a control group based on distraction [31].

Programs Targeting Nonclinical Populations

A reduction in certain stress-related biomarkers was found after mindfulness [37], but not after music therapy [38]. Specifically, Beerse et al [37] observed a significant reduction in sCort in both the mindfulness-based art therapy group and the neutral clay task control group, as well as a significant decrease in anxiety and perceived stress only in the experimental group. However, no significant correlations between sCort and perceived stress or between sCort and anxiety were detected. Finnerty et al [38] revealed a slight—yet nonsignificant—decrease in the participants' HCC levels after music therapy, compared to a significant increase in the control group, inferring that the program may enhance stress management skills rather than soothing anxiety [38]. On this behalf, stress and anxiety scores did not change after the 6 weeks program across all kinds of programs and the control group.

Two studies reported heterogeneous changes in immune biomarkers [39,40]. In Heckenberg et al [40], secretory immunoglobulin A increased after a mindfulness-based stress reduction program on direct-care workers, suggesting an improved level of mucosal immunity. The authors also found that changes in salivary AA concentration only approached significance [40].

Schakel et al [39] measured several immune and psychophysiological outcomes in participants who underwent an internet CBT program. After treatment, authors reported a significant increase in immunoglobulin G, a decrease in interleukin-1β and tumor necrosis factor-α, as well as significant variations in a diverse number of chemokines and cytokines. Moreover, participants reported improvements in HR variables over time, but with no significant differences between conditions. Ultimately, HR variability, sCort, AA, and skin conductance showed neither a significant main effect of time nor an interaction between time and condition [39].

Two of the included studies focused on the smoking population and they both reported significant findings [35,36]. In Janes et al [35], posterior cingulate cortex reactivity to smoking cues decreased after both the mindfulness training and the control program, with no significant difference between groups [35]. Participants also reported a decrease in cigarette consumption after the program. On the other hand, in Webb et al [36], 7-day point prevalence abstinence (PPA), defined as self-reported abstinence from smoking for 7 days, displayed higher rates in the program condition compared to the controls at 26 weeks, but this effect did not last up until the 52 weeks' time point. The abstinence was biochemically verified through breath CO levels, which emerged to be consistent with participants' self-evaluation, to the advantage of the program condition versus the control group after 26 weeks, but not after 52 weeks. Additionally, the consecutive 7-day PPA, defined as 2 or more consecutive 7-day PPA observations, resulted in significant enhancement in the program arm rather than in the control group both at 26 and 52 weeks.

Lastly, the remaining studies did not report significant changes in biomarkers. Specifically, stress-related biomarkers were found to be substantially the same before and after both a stress management program [33] and mindfulness-based intervention [34].

Additional Health Outcomes Associated With Biomarker Variations

Although a significant variation in biomarkers can be useful in assessing treatment response, the efficacy of the intervention can be further reinforced if the changes in the biomarkers are associated with a relevant improvement in participants' health outcomes. Some of the included studies reported, along with a significant variation in biomarkers, a reduction of patients' symptoms (in the clinical samples) or an overall improvement of participants' well-being (in the nonclinical samples), reflecting a fairly solid response to the intervention. Hereafter, we summarize the main outcomes of the above-mentioned studies.

Regarding the psychiatric or psychological samples, in Romero-Sanchiz et al [22], the significant decrease in chemokines in the CBT intervention in patients with depression matched a reduction in their symptoms. In Park et al [42], CBT for insomnia successfully improved sleep quality in patients on dialysis with sleeping problems. This achievement correlated with the enhancement, between the pre- and postintervention time points, of the resting-state brain connectivity between the default mode network and the premotor or dorsolateral prefrontal cortex, cortical areas crucially involved in sleep regulation. Among the medical samples, in Liu et al [28], cardiovascular outcomes were associated with improvements in physical activity: specifically, a greater number of daily steps was related to a lower BP. Furthermore, Lansing et al [19], in addition to a decreasing HbA_{1c} after the intervention, reported improvements in teens' performance on executive functions, specifically working memory and inhibitory control tasks. Fewer errors in inhibitory control tasks significantly correlated with the decreased HbA_{1c} (as well as with the increased frequency of glycemic control). Lastly, regarding nonclinical samples, Janes et al [35] reported a significant correlation between the reduction of PCC reactivity to smoking cues and the decrease in cigarette consumption: although this biomarker decreased in both groups, only in the mindfulness training condition did participants who exhibited a reduction in PCC cue reactivity also showed a greater reduction of cigarette use. Nevertheless, in the clinical samples, the patients' underlying condition (whether psychological or medical) could cause biological alterations. This makes it difficult to ascertain whether—and to what extent—the significant variations in biomarkers measured in the studies are due to the intervention, rather than the disorder itself.

Predictors of Response

Regarding the issue of prediction of response, some of the studies investigated whether changes in biomarkers or their baseline values predicted an improvement in symptoms (typically with linear regressions). Janes et al [35] investigated whether variations in PCC reactivity to smoking cues predicted

XSL•FO
RenderX

individual-level smoking reduction. Linear regressions were performed, with ΔPCC (the change in PCC reactivity from baseline to post intervention) significantly predicting changes in cigarettes smoked in the mindfulness group. This predictor was not significant in the control group. However, this model was only significant at the individual level, not at the group level. Laufer et al [41] assessed the predictive value of salivary AA, sCort, and hair cortisol before intervention on treatment response to identifying prescriptive biological markers. No correlation was found between changes in biomarkers and changes in depressive symptoms. However, exploratory analysis found that baseline AA predicted treatment response. Nolan et al [29] reported that total lipoprotein cholesterol and SBP were predictors of cardiovascular risk index after the intervention [29]. Finally, Graham et al [34] failed to report associations between changes in perceived stress and changes in CRP and interleukin-6.

Quality of Reporting

The quality of reporting evaluation is reported in Tables S2-S4 in Multimedia Appendix 2. Four studies were rated as "good" per quality of reporting, 16 studies as "fair," and 4 studies as "poor."

Some response patterns that emerged depending on the type of study should be underscored. In the 4 "before-after" studies, 75% (n=3) did not enroll all eligible participants (item 4) and 75% (n=3) did not have a large enough sample size (item 5); additionally, it was not possible to apply a few items to assess: the blindness of people assessing outcomes in 100% (n=4) of the studies (item 8), the percentage of dropout at follow-up in 75% (n=3) of the studies (item 9) and the individual-level outcome efforts in group-level interventions in 100% (n=3) of the studies (item 12). In 19 controlled intervention studies, 74% (n=14) were not double-blind (item 4), 58% (n=11) did not report intervention adherence measures (item 9), 58% (n=11) had a dropout rate higher than 20% (item 7), 47% (n=9) did not have a large enough sample size (item 12), and 47% (n=9) did not use an intention-to-treat (ITT) analysis (item 14).

Discussion

Principal Findings

The main goal of this systematic review was to investigate the effects of internet-based psychological interventions on various biomarkers. In addition, we aimed to explore whether variations in such biomarkers reflected alterations of the underlying health condition (whether psychiatric or psychological, organic or medical, or nonclinical). After the screening process, we included 24 studies, which were mainly randomized controlled trials. The most frequently targeted populations were patients who are depressed (for psychological or psychiatric samples), patients who are diabetic (for organic or medical samples), and health care professionals (for nonclinical samples); the most frequently administered programs were internet-based CBT or CBT-informed interventions; and the most frequently measured biomarkers were stress-related biomarkers.

Effect of Intervention or Program Duration on Biomarker Variations

Internet- and app-based interventions can extensively differ per intervention duration and sessions' frequency. This produces a variety of conditions that may affect outcomes in different ways. The 24 studies included in this review were rather heterogeneous: except for 4 (17%) studies that provided 12-month interventions or programs [21,28,29,36] and 1 (4%) study with a 25-week-long program [19], in the remaining papers (n=19, 79%) [22-27,30-35,37-43], the treatments lasted no longer than 3 months. Furthermore, among the included studies, the most frequently administered interventions were CBT and mindfulness-based therapy, which are typically short-term treatments.

A brief intervention or program may represent an obstacle to significant findings, considering that some biomarkers might need a longer amount of time to significantly vary, perhaps longer than a brief treatment may allow. All the abovementioned studies that deliver interventions or programs that last 6 to 12 months report significant biomarkers' changes, as opposed to the 8 (42%) out of 19 treatments [22,26,32,37-40,43] that lasted no longer than 3 months. This may suggest that the longer the intervention or program is, the more likely it is that biomarkers will change.

Nonetheless, we should acknowledge that this does not always occur; in fact, some shorter interventions and programs still report biomarkers' changes: for example, Park et al [42], Janes et al [35], Beerse et al [37], Schumacher et al [24], and Schakel et al [39], whose treatments lasted 3, 4, 5, 5, and 6 weeks, respectively.

Potential methodological limitations should also be discussed. Follow-up may impact the data collected by researchers in various ways: biological samples (such as saliva or blood samples) might deteriorate with time and become unusable, if not forthwith analyzed; similarly, data might be lost due to misplacement. A further issue is the possible dropout of participants, who may not be reachable or might refuse to partake in the follow-up data collection, leading to a loss of statistical power in the conducted analysis. Therefore, variability of biomarkers at follow-up might stem from different sources, and methodological issues should be considered when interpreting results.

Effect of Follow-Up on Biomarker Variations

In light of the abovementioned topic, it should also be taken into consideration whether the studies provide a follow-up assessment to evaluate the possibility of a further alteration of biomarkers after a certain amount of time from the intervention end point. Eight (33%) [23-26,30,32,34,39] of the 24 included papers re-evaluated participants at follow-up, yet only 4 (17%) [26,30,32,39] also reported biomarkers' changes.

Unfortunately, the scarcity of follow-up biomarkers' evaluations in the included studies and the gaps in measurements in those that provided a follow-up do not allow us to examine in depth the influence of time on the variations of several biological indicators. It is reasonable to think that changes in biomarker levels might depend on a multitude of factors, such as the type

XSL•FO
RenderX

of intervention or program, its duration and frequency, and the possible presence of a psychiatric or medical condition, besides the nature of the biomarker itself and the length of follow-up. We need a more homogeneous group of studies to derive more definitive and reliable conclusions.

Augmentation

According to the reviewed literature, there is growing evidence suggesting that internet-based interventions and programs may be effective in improving a variety of health conditions, even producing significant alterations of some biological factors. Among the several biomarkers included in this systematic review, those that more frequently varied as a consequence of an online treatment or program were cardiac or cardiovascular biomarkers (especially BP and HR), inflammatory biomarkers, neurological function biomarkers, and HbA_{1c}. Considering the type of internet-based interventions, CBT or CBT-informed and mindfulness-based interventions proved to be most efficient in improving both biomarkers and symptoms. However, this result might be biased, since these two psychological interventions were also the most represented among the included studies. These effects should not be underestimated in scientific research: when investigating the efficacy or effectiveness of a treatment or program, authors should take into consideration whether participants are currently following an online psychosocial intervention or program, seeing as this might represent a confounding variable. On the same note, internet-based interventions and programs also have implications in clinical practice and could be combined with in-person or pharmacological therapy in the treatment of both psychiatric or psychological and medical conditions [44]. Long-term interventions and programs may be useful to maintain symptom reduction in time and prevent relapses, promoting a more holistic approach to patient care. Furthermore, online or app-based interventions and programs could also be beneficial per prevention, for example, increasing motivation in smokers to quit smoking or teaching health care providers effective strategies to manage stress and anxiety. Given their promising qualities, future research should invest more in studying the long-term effects of internet-based interventions and programs: ideally, including larger and more diverse populations would further enrich our preliminary findings, by extending the generalizability of our results.

Strengths and Limitations

To the best of our knowledge, this is the first systematic review exploring how biomarkers change because of internet-based psychological interventions and programs. One of the strongest assets of this research regards the fact that it was conducted by following a meticulous and defined methodology according to the PRISMA statement [14]. By the employment of 3 databases and the workforce of 3 independent evaluators, who executed the screening taking into consideration three languages, we managed to reach a satisfying number of studies (n=24). This should be deemed as a major strength, especially in light of internet-based interventions and programs being a relatively new approach.

Nevertheless, conducting this systematic review presented some challenges. The first and main limitation of this study is that the included papers were greatly heterogeneous per targeted populations, administered interventions or programs, and measured biomarkers. Regarding online interventions, we included a variety of psychological, psychosocial, and motivational programs, although the most frequent were CBT or CBT-informed treatments. Such programs also varied per type of material (eg, videos, reading material, messages, etc) and methods of administration (eg, computer, smartphone app, etc). Another source of heterogeneity stems from the vastly different clinical and nonclinical conditions targeted by the interventions. To avoid possible confusion in reporting such variability, we attempted to mitigate this inconvenience by reporting our results separately for psychiatric or psychological, medical, and nonclinical populations: this clustering method allowed us to better describe and appreciate the different effects that various interventions and programs have on the different examined populations. This heterogeneity becomes even more conspicuous when it comes to biomarkers assessment (Table S5 in Multimedia Appendix 2), making it highly difficult to draw comprehensive conclusions. However, we believe that having an entirely homogeneous sample per biomarker (eg, only stress-related biomarkers or only immune response biomarkers) would provide a narrow perspective of the effect of the intervention, for the scope of our review. A second limitation of this review concerns a methodological fault, namely our inability to retrieve information regarding the agreement between authors in case of conflict when screening articles and assessing their suitability for inclusion (Cohen κ).

Other limitations that should be discussed are not problems of this review itself, but they mainly concern inherent weaknesses of the included studies. First, many studies (n=10, 42%) [30,31,33,34,36-38,40,41,43] relied at least on 1 self-measured biomarker assessment, whether it was by collecting saliva samples (eg, the studies by Graham et al [34] and Beerse et al [37]), app and sensor monitoring (eg, the studies by Bauman et al [33] and Finnerty et al [38]), or by accounting for their self-measured biomarkers values (eg, the study by Kern et al [43]). Self-reporting, especially when it comes to biomarkers, is much more likely to affect the reliability of the collected data. This consideration may not apply to app and sensor monitoring, which were deemed as accurate and sensitive methods by the studies that used them (n=3, 13%) [33,36,38], and therefore, we can safely assume that the results were not gravely affected by them. However, it is undeniable that other self-report methods can potentially interfere with the reliability of data, because participants may not be entirely truthful about their reported values or incorrectly execute the sample collection.

Second, another flaw of most of the included studies is the absence of a follow-up evaluation. As previously stated, depending on the underlying condition and the type of outcome measures, some biomarkers may take longer to show a significant alteration. Therefore, some of the studies that did not find significant changes in biomarkers might have shown different results in the longer term. In the wake of this, in studies in which biomarkers vary significantly, a follow-up would allow us to examine how such improvements develop through time. Long-lasting effects of internet-based interventions and programs could help to avoid relapses and delay medical

XSL•FO
RenderX

deterioration. Maintenance therapy, if permitted by the theoretical framework of the intervention or program, could be useful to preserve the treatment's or program's beneficial effects for a more extended period. Supporting this idea, in Lansing et al [19], the active treatment phase was followed by a maintenance period, which helped sustain the obtained results for 14 additional weeks after treatment ended [19].

Finally, the high dropout rate is another weakness of the included studies that needs to be addressed: as already mentioned in the quality assessment and visible in Tables S2-S4 in Multimedia Appendix 2, as far as it concerns the controlled intervention studies (n=19), more than half of the studies (n=10, 53%) have an overall dropout rate higher than 20%. Higher dropout rates can affect internal validity since they subtract potential additional data. In this review, the high dropout rate may be due to the internet-based nature of the interventions and programs. In face-to-face interventions or programs, the relevance in dropout prevention of a strong therapeutic alliance that encourages compliance has been well documented [45]. Perceiving their therapist as more understanding, involved, and agreeable about their goals may be a protective factor for patients against dropout [46]. Internet-based interventions and programs, especially if self-guided, often lack a relational component, which undoubtedly plays a fundamental role in face-to-face interventions and programs. This may lead to loss of motivation in individuals who find the treatment particularly challenging from an emotional standpoint. Karyotaki et al [3] found that therapist-assisted online interventions or programs have lower dropout rates than self-guided ones. In light of this knowledge, clinicians using internet-based interventions or programs should commit to creating better opportunities to establish a solid alliance with patients, build up strong motivation, and discourage attrition.

Future Research Directions

Even though our review allowed us to reach some notable findings, this research field is quite recent and deserves further development.

As previously stated, internet-based interventions and programs represent a fairly new answer to those seeking mental health or psychological support for a variety of conditions. The COVID-19 pandemic has ultimately increased this need, by both compelling social distance and worsening pre-existent mental health issues in the general population [47]. This, among many others, is one of the reasons why we managed to include only a moderate number of studies. Due to the continuous development of new technologies and the growth of demand for long-distance services, a great deal of research is expected to come along in the next few years, which will hopefully enrich this field of investigation.

Another research domain that should be further explored, especially in psychological terms, concerns biomarkers. Biological indicators are often used to study organic and medical pathologies but quite rarely when it comes to the psychological and mental health field. A stronger body of scientific research on the biology of mental illnesses may allow us, from a clinical point of view, to reach both a better understanding of their underlying mechanisms and a more suitable way of treating them by adding fundamental new information to the pharmacology area. For these reasons, it is our best hope that future researchers will invest more in using biomarkers as a valid and reliable outcome measure for mental health issues as well.

Conclusion

Our review provides a synthesis of literature examining the effects of internet-based psychological treatments and programs on several biomarkers. The results suggest that such interventions may have some beneficial effects on a variety of health conditions, in some cases even presenting significant physiological changes. Nonetheless, some studies show little to no positive effects, and in general, the existing evidence is scarce and somewhat conflicting. Although further evidence is required, we hope to raise awareness on the potential impact of internet-based interventions on biomarkers related to mental and physical health.

Authors' Contributions

RC and CMG conceptualized this study. GG, CG, and SR did the data curation. GG and CG analyzed the data. GG, CG, SR, and RC handled the methodology. RC administrated this study, while RC, FM, PC, JL-C, and PZ supervised it. GG and CG wrote the original draft. GG, CG, RC, and CMG reviewed and edited the writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms.
[DOCX File , 23 KB - jmir_v26i1e55736_app1.docx]

Multimedia Appendix 2
Supplementary tables S1-S5.
[DOCX File , 2990 KB - jmir_v26i1e55736_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e55736_app3.docx]

References

1. Chakrabarti S. Usefulness of telepsychiatry: a critical evaluation of videoconferencing-based approaches. World J Psychiatry 2015;5(3):286-304. [doi: 10.5498/wjp.v5.i3.286] [Medline: 26425443]
2. Richards D, Viganó N. Online counseling: a narrative and critical review of the literature. J Clin Psychol 2013;69(9):994-1011. [doi: 10.1002/jclp.21974] [Medline: 23630010]
3. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015;45(13):2717-2726. [doi: 10.1017/S0033291715000665] [Medline: 25881626]
4. Stoll J, Müller JA, Trachsel M. Ethical issues in online psychotherapy: a narrative review. Front Psychiatry 2020;10:993. [doi: 10.3389/fpsyt.2019.00993] [Medline: 32116819]
5. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69(3):89-95. [doi: 10.1067/mcp.2001.113989] [Medline: 11240971]
6. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. Silver Spring, MD: US Food and Drug Administration; 2016.
7. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: a mediation analysis. Soc Sci Med 2017;195:42-49. [doi: 10.1016/j.socscimed.2017.11.008] [Medline: 29132081]
8. Orosz A, Federspiel A, Eckert A, Seeher C, Dierks T, Tschitsaz A, et al. Exploring the effectiveness of a specialized therapy programme for burnout using subjective report and biomarkers of stress. Clin Psychol Psychother 2021;28(4):852-861. [doi: 10.1002/cpp.2539] [Medline: 33283948]
9. Shields GS, Spahr CM, Slavich GM. Psychosocial interventions and immune system function: a systematic review and meta-analysis of randomized clinical trials. JAMA Psychiatry 2020;77(10):1031-1043 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.0431] [Medline: 32492090]
10. Claudino FCDA, Gonçalves L, Schuch FB, Martins HRS, da Rocha NS. The effects of individual psychotherapy in BDNF levels of patients with mental disorders: a systematic review. Front Psychiatry 2020;11:445. [doi: 10.3389/fpsyt.2020.00445] [Medline: 32508690]
11. Marceau EM, Meuldijk D, Townsend ML, Solowij N, Grenyer BFS. Biomarker correlates of psychotherapy outcomes in borderline personality disorder: a systematic review. Neurosci Biobehav Rev 2018;94:166-178. [doi: 10.1016/j.neubiorev.2018.09.001] [Medline: 30208302]
12. Fancourt D, Perkins R, Ascenso S, Atkins L, Kilfeather S, Carvalho L, et al. Group drumming modulates cytokine response in mental health services users: a preliminary study. Psychother Psychosom 2016;85(1):53-55. [doi: 10.1159/000431257] [Medline: 26610172]
13. Dash SP. The impact of IoT in healthcare: global technological change & the roadmap to a networked architecture in India. J Indian Inst Sci 2020;100(4):773-785. [doi: 10.1007/s41745-020-00208-y] [Medline: 33162693]
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
15. ter Huurne ED, Postel MG, de Haan HA, Drossaert CH, DeJong CA. Web-based treatment program using intensive therapeutic contact for patients with eating disorders: before-after study. J Med Internet Res 2013;15(2):e12. [doi: 10.2196/jmir.2211] [Medline: 23380291]
16. Yang YJD, Allen T, Abdullahi SM, Pelphrey KA, Volkmar FR, Chapman SB. Neural mechanisms of behavioral change in young adults with high-functioning autism receiving virtual reality social cognition training: a pilot study. Autism Res 2018;11(5):713-725 [FREE Full text] [doi: 10.1002/aur.1941] [Medline: 29517857]
17. Carolan-Olah M, Sayakhot P. A randomized controlled trial of a web-based education intervention for women with gestational diabetes mellitus. Midwifery 2019;68:39-47. [doi: 10.1016/j.midw.2018.08.019] [Medline: 30343264]
18. Brockmeyer T, Walther S, Ingenerf K, Wild B, Hartmann M, Weisbrod M, et al. Brain effects of computer-assisted cognitive remediation therapy in anorexia nervosa: a pilot fMRI study. Psychiatry Res Neuroimaging 2016;249:52-56. [doi: 10.1016/j.pscychresns.2016.02.007] [Medline: 27000307]
19. Lansing AH, Stanger C, Budney A, Christiano AS, Casella SJ. Pilot study of a web-delivered multicomponent intervention for rural teens with poorly controlled type 1 diabetes. J Diabetes Res 2016;2016:7485613. [doi: 10.1155/2016/7485613] [Medline: 27610391]
20. Study quality assessment tools. National Heart, Lung, and Blood Institute (NHLBI). URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2023-09-27]
21. Egede LE, Walker RJ, Payne EH, Knapp RG, Acierno R, Frueh BC. Effect of psychotherapy for depression via home telehealth on glycemic control in adults with type 2 diabetes: subgroup analysis of a randomized clinical trial. J Telemed Telecare 2018;24(9):596-602. [doi: 10.1177/1357633X17730419] [Medline: 28945160]

XSL•FO
RenderX

22. Romero-Sanchiz P, Nogueira-Arjona R, Araos P, Serrano A, Barrios V, Argente J, et al. Variation in chemokines plasma concentrations in primary care depressed patients associated with internet-based cognitive-behavioral therapy. Sci Rep 2020;10(1):1078. [doi: 10.1038/s41598-020-57967-y] [Medline: 31974503]
23. Engel S, Schumacher S, Niemeyer H, Kuester A, Burchert S, Klusmann H, et al. Associations between oxytocin and vasopressin concentrations, traumatic event exposure and posttraumatic stress disorder symptoms: group comparisons, correlations, and courses during an internet-based cognitive-behavioural treatment. Eur J Psychotraumatol 2021;12(1):1886499. [doi: 10.1080/20008198.2021.1886499] [Medline: 33968321]
24. Schumacher S, Engel S, Niemeyer H, Küster A, Burchert S, Skoluda N, et al. Salivary cortisol and alpha-amylase in posttraumatic stress disorder and their potential role in the evaluation of cognitive behavioral treatment outcomes. J Trauma Stress 2022;35(1):78-89. [doi: 10.1002/jts.22683] [Medline: 34022094]
25. Jaser SS, Whittemore R, Chao A, Jeon S, Faulkner MS, Grey M. Mediators of 12-month outcomes of two internet interventions for youth with type 1 diabetes. J Pediatr Psychol 2014;39(3):306-315. [doi: 10.1093/jpepsy/jst081] [Medline: 24163439]
26. González-Moret R, Cebolla A, Cortés X, Baños RM, Navarrete J, de la Rubia JE, et al. The effect of a mindfulness-based therapy on different biomarkers among patients with inflammatory bowel disease: a randomised controlled trial. Sci Rep 2020;10(1):6071. [doi: 10.1038/s41598-020-63168-4] [Medline: 32269278]
27. Peerani F, Watt M, Ismond KP, Whitlock R, Ambrosio L, Hotte N, et al. A randomized controlled trial of a multicomponent online stress reduction intervention in inflammatory bowel disease. Therap Adv Gastroenterol 2022;15:17562848221127238. [doi: 10.1177/17562848221127238] [Medline: 36187365]
28. Liu S, Tanaka R, Barr S, Nolan RP. Effects of self-guided e-counseling on health behaviors and blood pressure: results of a randomized trial. Patient Educ Couns 2020;103(3):635-641. [doi: 10.1016/j.pec.2019.10.007] [Medline: 31669047]
29. Nolan RP, Feldman R, Dawes M, Kaczorowski J, Lynn H, Barr SI, et al. Randomized controlled trial of e-counseling for hypertension: REACH. Circ Cardiovasc Qual Outcomes 2018;11(7):e004420. [doi: 10.1161/CIRCOUTCOMES.117.004420] [Medline: 30006474]
30. Penedo FJ, Fox RS, Walsh EA, Yanez B, Miller GE, Oswald LB, et al. Effects of web-based cognitive behavioral stress management and health promotion interventions on neuroendocrine and inflammatory markers in men with advanced prostate cancer: a randomized controlled trial. Brain Behav Immun 2021;95:168-177 [FREE Full text] [doi: 10.1016/j.bbi.2021.03.014] [Medline: 33737170]
31. Urech C, Scherer S, Emmenegger M, Gaab J, Tschudin S, Hoesli I, et al. Efficacy of an internet-based cognitive behavioral stress management training in women with idiopathic preterm labor: a randomized controlled intervention study. J Psychosom Res 2017;103:140-146. [doi: 10.1016/j.jpsychores.2017.10.014] [Medline: 29167041]
32. Gotink RA, Younge JO, Wery MF, Utens EMWJ, Michels M, Rizopoulos D, et al. Online mindfulness as a promising method to improve exercise capacity in heart disease: 12-month follow-up of a randomized controlled trial. PLoS One 2017;12(5):e0175923. [doi: 10.1371/journal.pone.0175923] [Medline: 28486559]
33. Baumann H, Heuel L, Bischoff LL, Wollesen B. Efficacy of individualized sensory-based mHealth interventions to improve distress coping in healthcare professionals: a multi-arm parallel-group randomized controlled trial. Sensors (Basel) 2023;23(4):2322 [FREE Full text] [doi: 10.3390/s23042322] [Medline: 36850920]
34. Graham B, Jin Y, Bazeley P, Husni E, Calabrese LH. Online, low-volume meditation does not alter immune-related biomarkers. Brain Behav Immun Health 2022 Dec;26:100531 [FREE Full text] [doi: 10.1016/j.bbih.2022.100531] [Medline: 36267832]
35. Janes AC, Datko M, Roy A, Barton B, Druker S, Neal C, et al. Quitting starts in the brain: a randomized controlled trial of app-based mindfulness shows decreases in neural responses to smoking cues that predict reductions in smoking. Neuropsychopharmacology 2019 Aug;44(9):1631-1638 [FREE Full text] [doi: 10.1038/s41386-019-0403-y] [Medline: 31039580]
36. Webb J, Peerbux S, Ang A, Siddiqui S, Sherwani Y, Ahmed M, et al. Long-term effectiveness of a clinician-assisted digital cognitive behavioral therapy intervention for smoking cessation: secondary outcomes from a randomized controlled trial. Nicotine Tob Res 2022 Oct 26;24(11):1763-1772 [FREE Full text] [doi: 10.1093/ntr/ntac113] [Medline: 35470860]
37. Beerse ME, van Lith T, Stanwood G. Therapeutic psychological and biological responses to mindfulness-based art therapy. Stress Health 2020 Oct;36(4):419-432. [doi: 10.1002/smi.2937] [Medline: 32073202]
38. Finnerty R, McWeeny S, Trainor L. Online group music therapy: proactive management of undergraduate students' stress and anxiety. Front Psychiatry 2023;14:1183311 [FREE Full text] [doi: 10.3389/fpsyt.2023.1183311] [Medline: 37151974]
39. Schakel L, Veldhuijzen DS, van Middendorp H, Prins C, Drittij AMHF, Vrieling F, et al. An internet-based psychological intervention with a serious game to improve vitality, psychological and physical condition, and immune function in healthy male adults: randomized controlled trial. J Med Internet Res 2020 Jul 24;22(7):e14861 [FREE Full text] [doi: 10.2196/14861] [Medline: 32706667]
40. Heckenberg RA, Hale MW, Kent S, Wright BJ. An online mindfulness-based program is effective in improving affect, over-commitment, optimism and mucosal immunity. Physiol Behav 2019 Feb 01;199:20-27. [doi: 10.1016/j.physbeh.2018.11.001] [Medline: 30395806]

XSL•FO
RenderX

41. Laufer S, Schulze L, Engel S, Klusmann H, Skoluda N, Nater UM, et al. The effect of an internet-based intervention for depression on cortisol and alpha-amylase. Psychoneuroendocrinology 2023 Jun;152:106082. [doi: 10.1016/j.psyneuen.2023.106082] [Medline: 36989562]
42. Park HY, Lee H, Jhee JH, Park KM, Choi EC, An SK, et al. Changes in resting-state brain connectivity following computerized cognitive behavioral therapy for insomnia in dialysis patients: A pilot study. Gen Hosp Psychiatry 2020;66:24-29 [FREE Full text] [doi: 10.1016/j.genhosppsych.2020.05.013] [Medline: 32615333]
43. Kern D, Ljótsson B, Bonnert M, Lindefors N, Kraepelien M. Online guided self-help cognitive behavioral therapy with exposure to anxiety and problem solving in type 1 diabetes mellitus: case study. JMIR Form Res 2022 Jul 13;6(7):e32950 [FREE Full text] [doi: 10.2196/32950] [Medline: 35830220]
44. Calati R. Augmentation/combination strategies in psychotherapy: the need of clear guidelines and future research. Neuropsychobiology 2021;80(6):516-517. [doi: 10.1159/000515932] [Medline: 33857950]
45. Meier PS, Donmall MC, Barrowclough C, McElduff P, Heller RF. Predicting the early therapeutic alliance in the treatment of drug misuse. Addiction 2005 Apr;100(4):500-511 [FREE Full text] [doi: 10.1111/j.1360-0443.2005.01031.x] [Medline: 15784065]
46. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
47. Kaufman KR, Petkova E, Bhui KS, Schulze TG. A global needs assessment in times of a global crisis: world psychiatry response to the COVID-19 pandemic. BJPsych Open 2020 Apr 06;6(3):e48 [FREE Full text] [doi: 10.1192/bjo.2020.25] [Medline: 32250235]

Abbreviations

AA: α amylase
BP: blood pressure
CBT: cognitive behavioral therapy
CRP: C-reactive protein
CRT: cognitive remediation therapy
FDA: US Food and Drug Administration
HbA1c: glycated hemoglobin
HCC: hair cortisol concentration
HR: heart rate
ITT: intention-to-treat
NIH: National Institutes of Health
PCC: posterior cingulate cortex
PPA: point prevalence abstinence
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
SBP: systolic blood pressure
sCort: salivary cortisol

Edited by A Coristine; submitted 22.12.23; peer-reviewed by L Guo, A Hassan, N Cera; comments to author 11.05.24; revised version received 29.06.24; accepted 24.09.24; published 29.11.24.

Please cite as:
Gotti G, Gabelli C, Russotto S, Madeddu F, Courtet P, Lopez-Castroman J, Zeppegno P, Gramaglia CM, Calati R
Biomarkers of Response to Internet-Based Psychological Interventions: Systematic Review
J Med Internet Res 2024;26:e55736
URL: https://www.jmir.org/2024/1/e55736
doi:10.2196/55736
PMID:

©Giulia Gotti, Chiara Gabelli, Sophia Russotto, Fabio Madeddu, Philippe Courtet, Jorge Lopez-Castroman, Patrizia Zeppegno, Carla Maria Gramaglia, Raffaella Calati. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Smart Glass Telemedicine Application for Prehospital Communication: User-Centered Design Study

Zhan Zhang[1], PhD; Enze Bai[1], MS; Yincao Xu[1], MSc; Aram Stepanian[1], MSc; Jared M Kutzin[2,3], DNP, MS, MPH, RN; Kathleen Adelgais[4,5], MD, MPH; Mustafa Ozkaynak[6], PhD

[1]School of Computer Science and Information Systems, Pace University, New York, NY, United States
[2]Icahn School of Medicine, Mount Sinai, New York, NY, United States
[3]Mount Sinai Hospital, New York, NY, United States
[4]School of Medicine, University of Colorado, Aurora, CO, United States
[5]Children's Hospital Colorado, Aurora, CO, United States
[6]College of Nursing, University of Colorado, Aurora, CO, United States

Corresponding Author:
Zhan Zhang, PhD
School of Computer Science and Information Systems
Pace University
1 Pace Plaza
New York, NY, 10038
United States
Phone: 1 3153992627
Email: zzhang@pace.edu

Abstract

Background: Smart glasses have emerged as a promising solution for enhancing communication and care coordination among distributed medical teams. While prior research has explored the feasibility of using smart glasses to improve prehospital communication between emergency medical service (EMS) providers and remote physicians, a research gap remains in understanding the specific requirements and needs of EMS providers for smart glass implementation.

Objective: This study aims to iteratively design and evaluate a smart glass application tailored for prehospital communication by actively involving prospective users in the system design process.

Methods: Grounded in participatory design, the study consisted of 2 phases of design requirement gathering, rapid prototyping, usability testing, and prototype refinement. In total, 43 distinct EMS providers with diverse backgrounds participated in this 2-year long iterative design process. All qualitative data (eg, transcribed interviews and discussions) were iteratively coded and analyzed by at least 2 researchers using thematic analysis. Quantitative data, such as System Usability Scale (SUS) scores and feature ratings, were analyzed using statistical methods.

Results: Our research identified challenges in 2 essential prehospital communication activities: contacting online medical control (OLMC) physicians for medical guidance and notifying receiving hospital teams of incoming patients. The iterative design process led to the identification of 5 key features that could potentially address the identified challenges: video call functionality with OLMC physicians, call priority indication for expedited OLMC contact, direct communication with receiving hospitals, multimedia patient information sharing, and touchless interaction methods for operating the smart glasses. The SUS score for our system design improved from a mean of 74.3 (SD 11.3) in the first phase (classified as *good* usability) to 80.3 (SD 13.1) in the second phase (classified as *excellent* usability). This improvement, along with consistently high ratings for other aspects (eg, willingness to use and feature design), demonstrated continuous enhancement of the system's design across the 2 phases. Additionally, significant differences in SUS scores were observed between EMS providers in urban areas (median 85, IQR 76-94) and rural areas (median 72.5, IQR 66-83; Mann-Whitney U=43; P=.17), as well as between paramedics (median 72.5, IQR 70-80) and emergency medical technicians (median 85, IQR: 74-98; Mann-Whitney U=44.5; P=.13), suggesting that EMS providers in urban settings and those with less training in treating patients in critical conditions perceived the smart glass application as more useful and user-friendly. Finally, the study also identified several concerns regarding the adoption of the smart glass application, including technical limitations, environmental constraints, and potential barriers to workflow integration.

XSL•FO
RenderX

Conclusions: Using a participatory design approach, this study provided insights into designing user-friendly smart glasses that address the current challenges EMS providers face in dynamic prehospital settings.

(J Med Internet Res 2024;26:e53157) doi:10.2196/53157

KEYWORDS

smart glass; telemedicine; participatory design; emergency medical service; health care; prehospital care; mobile health; mHealth; augmented reality

Introduction

Background

Emergency medical services (EMSs) are a specialized medical domain dedicated to providing urgent medical care to patients who are critically ill or the place where an incident occurred. It involves dispatching certified emergency care clinicians, such as paramedics and emergency medical technicians (EMTs), to the scene of an emergency. The primary objective of EMS providers is to promptly stabilize and treat patients with complex and constantly changing conditions. Furthermore, EMS providers must dedicate time to collecting, sharing, and discussing crucial patient information with remote care teamsa process known as *prehospital communication*.

Effective and accurate prehospital communication is critical for ensuring that remote care teams (eg, emergency department [ED] physicians and nurses at the receiving facility) understand the patients' conditions and are thus able to offer more accurate guidance, better prepare for the patient's arrival, and mobilize necessary resources [1]. In the current practice, EMS providers rely on a radio or a phone to communicate with remote care teams. However, these conventional communication methods pose challenges not only for EMS providers trying to verbally convey complex patient situations but also for remote care teams attempting to comprehend the patients' status. Therefore, previous research has called for more advanced technology solutions to address the inherent limitations of conventional communication methods [2,3]. Recognizing this critical need, researchers have explored the potential of telemedicine systems to facilitate prehospital communication [4-11]. Notable examples include ambulance-based telemedicine systems, where computers and cameras are installed inside ambulances to capture video footage of designated areas (eg, the patient body), which can be streamed to remote physicians for medical guidance or decision support. Although these systems can significantly enhance communication and information sharing during EMS care and transport, their adoption rate remains low due to various limitations (eg, lack of portability, limited usability, and reliance on manual input and control) [5,11,12]. For example, Cho et al [5] reported that the size and weight of the telemedicine unit made it cumbersome or even impossible to use outside the ambulance, where a great portion of patient care takes place. In addition, manually handling these systems increases the risk of cross contamination and patient infections [13]. These limitations of ambulance-based telemedicine systems hinder their effective use in the dynamic, mobile, and hands-on EMS environment.

Given the limitations of ambulance-based telemedicine systems, there is a need for more portable telemedicine technologies that can accommodate the hands-busy and mobile nature of EMS work. Smart glasses have emerged as a promising solution due to their potential advantages, such as high portability and hands-free operation [14,15]. For example, several studies have tested the affordances and feasibility of using smart glasses to enable real-time sharing of visual medical information from the field with remote emergency physicians [14,16-18]. In this existing body of research, most studies have used off-the-shelf smart glass devices and teleconferencing software for testing purposes, such as evaluating the technical feasibility or usability of these devices in EMS work [19]. Yet, these studies did not engage with EMS providers to identify their specific needs or to gather their perspectives on how the smart glasses should be designed to align with the unique characteristics of EMS work [20,21]. Therefore, a critical research question remains unanswered: How can smart glasses be designed as a telemedicine platform to enhance prehospital communication while considering the hands-busy and time-critical nature of EMS work?

To address this research question, we adopted a user-centered, participatory design (PD) approach to ensure the involvement of prospective users in the smart glass design process. This work is part of a larger research effort aimed at designing and developing an integrated telemedicine system to enhance prehospital communication, which includes smart glasses worn by EMS providers to connect with a separate desktop application used by remote emergency physicians or experts. Our design goals are not to replace current communication systems (eg, radio) but rather to supplement them to address some of the long-lasting challenges in the prehospital communication process. In this paper, we specifically focused on presenting a detailed account of our 2-year long user-centered design process for the smart glass application.

This work made the following contributions to the fields of medical and health care informatics: (1) an empirical understanding of the challenges and barriers in prehospital communication; (2) design insights for implementing smart glass–based telemedicine systems tailored for fast-paced, hands-busy medical teams such as EMS; and (3) user perceptions and potential barriers about the adoption and use of smart glasses in time-critical medical scenarios.

Related Work

PD in Medical Work

The PD methodology facilitates the rapid development and evaluation of design concepts by directly involving intended users in the development of information systems [22,23]. This approach has proven effective in designing and developing health information technologies (HITs) [24], ensuring their

XSL•FO
RenderX

efficiency, user-friendliness, and seamless integration into clinical workflow [25-28].

Several prior studies have used PD approaches to design and develop HITs for emergency care providers. For example, Kusunoki et al [29,30] engaged ED providers in PD workshops to co-design information displays that support awareness and enhance ED teamwork. Østervang et al [31] conducted PD workshops with patients, family members, health care professionals, and IT specialists to co-design a patient health information system. Another highly relevant work by Kristensen et al [22] used PD to understand the nuanced practices in prehospital care workflow and generate technology ideas and concepts for future EMS practice. Building on this body of work, we used the PD approach in our study to explore how smart glasses should be designed specifically for the prehospital care setting, an area previously unexplored using the PD method.

Smart Glasses in Health Care

The hands-free capabilities of smart glasses allow health care providers to use both hands for patient care while accessing, viewing, and sharing patient information at the point of care. Studies have shown how smart glasses can enhance collaboration among health care providers in different locations, such as when a local surgeon uses smart glasses to receive real-time guidance from a remote expert [32,33], or a nursing student can get support from an expert during cardiopulmonary resuscitation [34]. Particularly relevant to our work are studies investigating the feasibility and potential of using smart glasses in prehospital care settings [19]. In these studies, EMS providers used smart glasses to share live visual medical data from the field with remote emergency physicians, facilitating immediate and informed medical decisions [14,16-18]. These studies highlight the valuable role that smart glasses can play in enhancing communication and coordination between prehospital teams and remote emergency physicians or experts.

However, most of this research has relied on commercially available devices without incorporating feedback from end users in the design process. This gap could result in a mismatch between the technology design and the actual needs and workflows of medical professionals, potentially hindering the adoption of new HITs in complex medical settings [35]. Therefore, it is essential to understand how to effectively integrate smart glasses into the prehospital communication workflow while considering their dynamic and hands-on care practices. To address this gap, this study adopted a PD methodology, engaging EMS providers in an iterative design and evaluation process to create a smart glass application that meets their specific requirements and workflow.

Methods

Study Design

In this study, we used an iterative user-centered design approach, combining PD [22,23] with usability evaluation [36,37]. This study comprised 2 phases, each including design requirement gathering, rapid prototyping, usability evaluation, and prototype refinement (Figure 1). The entire study lasted from November 2021 to April 2023.

Figure 1. An illustration depicting the 2-year user-centered design study process.

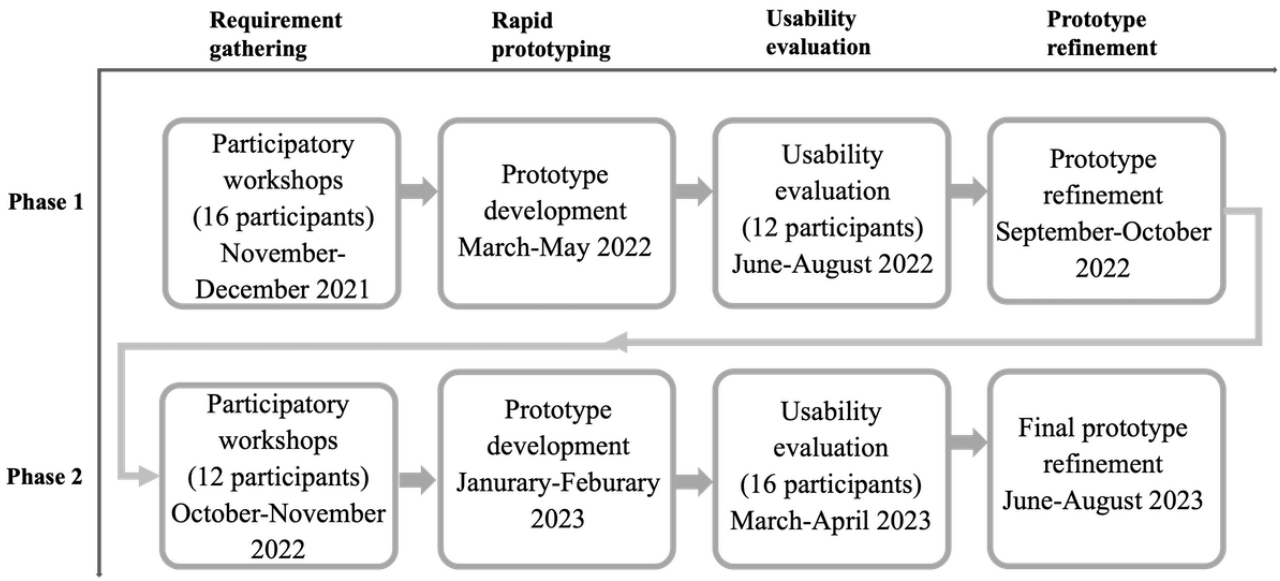

We recruited study participants from 4 EMS agencies: 1 fire-based agency located in the rural mountain region of the United States and 3 hospital-based EMS agencies in an urban area on the East Coast of the United States. The diverse characteristics of these participating EMS agencies (eg, fire vs hospital-based agencies and urban vs rural settings) enhance the generalizability of our research. Directors at each EMS agency disseminated a recruitment email to their teams, instructing interested members to contact the researcher directly. The research objective and eligibility criteria for participation were explained in the recruitment email. The only eligibility criterion for participation was being a licensed EMS provider in those agencies, regardless of gender, ethnicity, years of experience, or other factors. Upon receiving responses from EMS providers, the researchers coordinated with each participant to schedule a convenient time for their involvement in the study. In addition, researchers answered any questions

XSL•FO
RenderX

that providers had about the project and offered further details when requested.

In total, 43 unique EMS providers participated in our study, with 8 (19%) of them taking part in >1 session. The recruited participants held different roles (eg, paramedic vs EMT) and had a wide range of experience (from <1 year to >40 years). More specifically, 20 (47%) participants were paramedics, while the rest (23/43, 53%) were EMTs; 13 (30%) were recruited from the rural area-based agency, with the rest (30/43, 70%) from the urban-based agencies. Participant demographics and the sessions they attended are detailed in Multimedia Appendix 1.

Data Collection

PD Workshop

We conducted 4 PD workshops in phases 1 and 3 PD workshops in phase 2, with each workshop including 4 participants. The number of participants for each phase and the timing of the workshops in each phase can be found in Figure 1.

During phase 1, our primary focus was to understand the challenges of the prehospital communication process, identify critical system features catering to the needs of EMS providers, and determine the most preferred interaction methods for using the smart glass device. In phase 2, the research team shifted its focus to refining system designs and features and identifying various factors and barriers that could affect the practical application of smart glasses in real-world scenarios.

Each workshop lasted up to 2 hours and included the following activities: (1) group discussions, where participants engaged in group discussions to share their requirements and suggestions for creating a smart glass–based telemedicine system; (2) individual sketching, where each participant had the opportunity to either create different design concepts through paper sketching and explain their design rationale or critique an existing prototype created in a previous phase, (3) group-based design, where participants collectively refined and developed design concepts that everybody agreed upon [22,29,30,38,39], and (4) wrap-up discussion, where participants shared their insights regarding sociotechnical challenges in the effective adoption and use of smart glasses in their workflow. We have included the study protocol for the phase 1 workshop in Multimedia Appendix 2 to illustrate how we structured the workshops.

Prototyping and Usability Evaluation

In each phase, the research team started the prototyping process by iteratively creating low-fidelity prototypes using paper and medium-fidelity prototypes using Figma (Figma Inc), based on the findings and user insights collected from preceding design or evaluation studies. Once the design (eg, system features) was finalized, we implemented functional prototypes using the Vuzix M400 platform (Vuzix Corporation), as depicted in Figure 2A. This device features a see-through, near-eye display that presents information without obstructing the wearer's vision. It is also equipped with a camera adjacent to the near-eye display, enabling the capturing of still images and video streaming from a first-person point of view. The embedded GPS allows for real-time location tracking and sharing. By default, the device is operated through tangible buttons or a touchpad. The software development kit provided by Vuzix facilitates application development and enables the implementation of additional interaction methods (eg, voice commands). We used the Zoom application programming interface (Zoom Video Communications Inc) to develop video and audio call features and integrate them into our application. The prototype development in phase 1 took approximately 6 months due to several technical challenges. For example, the embedded Android system in our device as well as the gesture sensing software for implementing hands-free operation were using an old Native Developer Kit (NDK), whereas our teleconference software provider (Zoom) used a new NDK version. The use of different NDK versions of these software and hardware caused integration issues during compiling in a lower-level virtual machine. We worked closely with the technical support teams of all the companies, involving numerous communications and meetings, to resolve the issues of integrating different software packages.

Figure 2. (A) The smart glass device (Vuzix M400) used in the project. (B) A study participant interacting with the application using hand gestures.

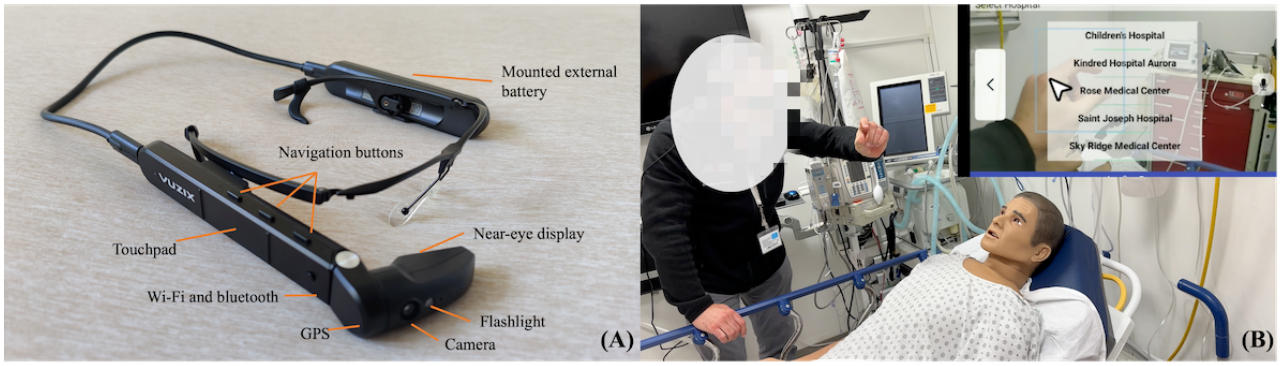

After we successfully developed a functional system prototype, we conducted individual usability evaluations with EMS providers in a controlled environment, such as an office or a simulation laboratory (Figure 2B). The number of participants for each phase is depicted in Figure 1. We have included the study protocol for the usability evaluation in Multimedia Appendix 3 to illustrate how we structured the usability evaluation.

The purpose of this activity was to both gather design requirements and evaluate the designs. In particular, we were interested in finding out whether the application's designs and features could effectively address the needs of EMS providers

XSL•FO
RenderX

and current challenges in the prehospital communication process. At the beginning of each testing session, we provided training to the participants about system features and how to use the application. Once participants were confident enough to use the application independently, they were guided to complete various tasks corresponding to the major features of our application. This step involved participants using different interaction methods (such as tangible buttons and touchless interaction methods like voice commands) in a randomized order to perform an identical set of tasks. This approach helped us determine which interaction method was more user-friendly and practical for EMS providers when using the smart glass application to contact remote physicians.

We used both objective and subjective measures to evaluate the system's usability. Objective measures included task completion time, task success rate, and errors encountered [40]. Subjective measures were administered through a survey, which consisted of a modified version of the System Usability Scale (SUS) [41] and a section of Likert-scale questions for participants to rate the usefulness of different system features and interaction methods on a scale of 1 to 5 (1 representing not useful at all and 5 representing very useful) [42-44], as well as a poststudy interview to elicit participants' perceived benefits and challenges associated with using the application in practice. Feedback and responses from participants informed the prioritization of design modifications and prototype refinements in subsequent steps.

Data Analysis

To facilitate data analysis, all research activities were both audio-recorded and videotaped. Photographs of the created artifacts (eg, design sketches or drawings) were also taken for data analysis purposes. The design sketches collected in design workshops and the annotations on paper prototypes gathered throughout the usability evaluations were used to understand how EMS providers envisioned the functionality and appearance of the application. Specifically, we combined these sketches and annotations with discussions on system features to understand not only the features participants wanted or envisioned but also how they would like to design the user flow and structure the screen layout for easy operation.

All discussions occurred in design workshops, and usability evaluations were transcribed, iteratively coded, and analyzed by at least 2 researchers using descriptive, thematic analysis [45-47]. The transcripts were managed and analyzed using NVivo (version 12; Lumivero). Our analysis focused on user requirements, feedback on system features, perceived benefits and concerns related to using smart glasses in the field, and the sociotechnical considerations of integrating smart glasses into EMS work. Two researchers began the qualitative data analysis by independently analyzing a small subset of transcripts and generating codebooks to standardize the data analysis process. We then used the Cohen κ coefficient to test interrater reliability by having the same 2 researchers independently code another small set of transcripts using the developed codebook and compare their codes. After achieving a substantial intercoder agreement, the 2 researchers separately coded the remaining transcripts. Any new codes that emerged through this process were added to the codebook. Disagreements in the analysis were discussed and resolved during weekly group meetings among all researchers. Once the coding process was completed, the generated codes were organized into high-level categories to identify overarching themes. The researchers also conducted member-checking to ensure the validity of data analysis. For instance, major findings from each study were presented to a subset of study participants and EMS agency directors to confirm that the researchers' interpretations of the data accurately reflected providers' opinions and real-world practices.

Descriptive statistical approaches were used to analyze quantitative data, such as SUS questionnaire responses and feature ratings by EMS providers. For instance, we followed the instructions as outlined by Brooke et al [41], to calculate the SUS score for each major design version and calculated the average rating given by participants for each feature. We also conducted the Mann-Whitney *U* test with Bonferroni corrected *P* values to assess differences in user ratings among participants with varying characteristics. More specifically, we compared SUS scores and user ratings of system features between EMTs and paramedics, EMS providers in rural areas versus urban areas, and those with <10 years of experience against those with >10 years of experience.

To evaluate the usability difference between tangible buttons and touchless interaction methods (eg, voice commands), we used both objective measurements (eg, task completion time, errors encountered, and the time taken to recover from errors) and subjective measurements (eg, user experiences collected through a questionnaire and posttest interviews). We applied the Friedman test to determine if significant differences in measurements (eg, task completion time and errors) existed across all interaction methods. If a significant difference was identified, we then conducted the Wilcoxon signed rank test with Bonferroni corrected *P* values for post hoc pairwise comparisons. More details about the assessment of usability differences between tangible buttons and touchless interaction methods can be found in the study by Zhang et al [48].

Ethical Considerations

This research was approved by the Pace University institutional review board (1708685). All participants provided informed consent by reviewing and signing a consent form, which detailed their participation rights, the risks involved, and the fact that the study would be recorded. The research team also provided a verbal explanation of these rights, ensuring participants understood that their input would solely be used for research purposes and that their identities would be protected. To ensure anonymity and protect the privacy and confidentiality of the participants, all data collected were deidentified, with participant identities removed from transcripts or other data (eg, survey responses) and replaced by unique participant IDs. Finally, participants were compensated for their time at a rate of US $60 per hour. All participants took part in the research during their off-duty hours, ensuring that their normal patient care activities were not disrupted.The individual in Figure 2 provided written informed consent to allow their image to be published.

XSL•FO
RenderX

Results

Work Practices and Challenges in Prehospital Communication

Overview

In the first phase of PD workshops, we examined the current work practices and technology use of EMS providers as well as the challenges they frequently encounter during the prehospital communication process. Understanding these aspects is crucial for designing a smart glass application that effectively tackles the challenges in prehospital communication and aligns with the needs of EMS providers. We subsequently describe the 2 prominent activities in the prehospital communication process and the challenges associated with them.

Call Online Medical Control Physicians for Medical Guidance

A crucial prehospital communication activity for EMS providers is contacting an online medical control (OLMC) physician to seek medical guidance and oversight (Figure 3A), a process guided by established protocols. Reasons for contacting OLMC physicians include seeking approval for medication administration and obtaining decision support regarding patient destination. The OLMC system can be either centralized, with physicians operating from a centralized office, or decentralized, allowing physicians to operate from any location using a dedicated phone or application. In our study, the urban area on the East Coast used a centralized OLMC system design, while the rural area in the mountain region adopted a decentralized design.

Figure 3. A typical prehospital communication process. (A) Emergency medical service (EMS) contacts online medical control physicians to receive medical guidance. (B) EMS calls ambulance dispatchers to initiate prehospital communication. (C) Ambulance dispatchers route the EMS call to the designated hospital. (D) The hospital's dedicated emergency communication center receives and relays prehospital information to the emergency department. (E) EMS providers may also call the hospital directly to expedite the notification process.

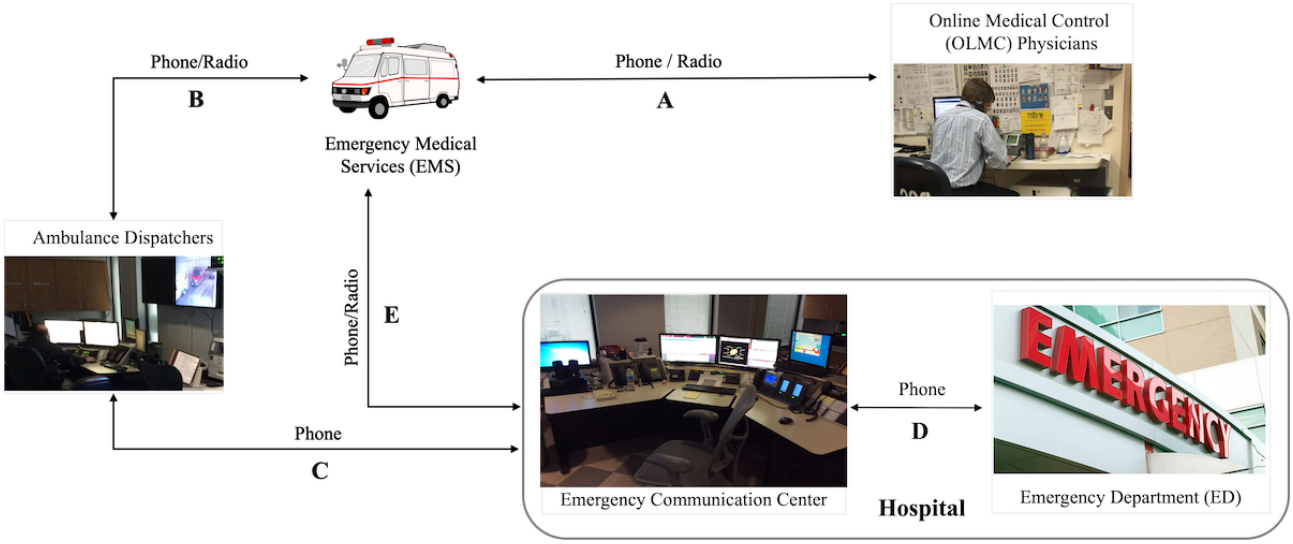

The most significant challenge encountered when contacting the OLMC physician was the prolonged waiting time due to the physicians' limited capacity to handle a large volume of EMS calls. In the urban area where 3 of our study sites were located, only 2 physicians were on call to handle incoming calls from EMS teams serving millions of people. This situation imposed an immense workload on OLMC physicians, causing significant delays in responding to EMS calls. Nearly all our study participants expressed frustration with the current OLMC system:

> *In our city, all five boroughs call one physician. So if that one physician is handling five, six, or seven calls, you could be on hold. Depending on the priority of your patient and what you need, I've been on hold [for] 45 minutes to an hour at times.* [Participant 8]

Another participant shared a similar experience, emphasizing that prolonged waiting times to connect with an OLMC physician are not uncommon, even in critical scenarios:

> *Even in a cardiac arrest, maybe we want orders for extra medications, or just to report the time of death, or maybe we want to transfer the patient out of there, we continue with our CPR and administer medicine while waiting on the phone to determine the next steps.* [Participant 5]

The challenge of communicating with OLMC physicians is exacerbated by the use of radios, as it complicates the sharing of contextual information (eg, patient status and symptoms), leading to miscommunications between EMS providers and OLMC physicians:

> *Miscommunication is a constant issue. It's just an outdated form of communication, it's what we feel about the radio we're using.* [Participant 9]

Notify Hospital for Incoming Patients

In addition to contacting OLMC physicians for medical guidance, EMS providers must also inform the receiving hospital about the patient being transported to their facility. This notification is typically succinct but includes crucial details such as the patient's age, symptoms, initial impression, treatments administered, and estimated time of arrival (ETA). This vital information allows the receiving hospital team,

XSL•FO
RenderX

including physicians and nurses in the ED, to anticipate the patient's needs and allocate necessary resources promptly.

Similar to contacting OLMC physicians, the primary issue with hospital notification is the ineffective communication arising from radio use, as highlighted in prior studies [2,3]. Our participants expressed similar concerns about relying exclusively on radios for hospital notification:

> *So when EMS providers [are] in the field, they find that it's very time consuming to tell the hospital team about the [patient's] stats. Miscommunicated patient stats are a problem that's faced [by care teams]. There's a lot of contextual information that gets lost because the radio is an old piece of technology.* [Participant 9]

Furthermore, as depicted in Figure 3, EMS providers must navigate several communication layers when attempting to notify the receiving ED team. Typically, EMS providers begin the communication process by contacting their dispatcher (Figure 3B), who then transfers the call to the designated hospital (Figure 3C). Some hospitals have an emergency communication center specifically for managing calls from EMS providers and conveying patient information to the ED team (Figure 3D). This multilayer communication approach often results in miscommunication or delays in hospital notification [1]. For instance, in urban cities where transport time is generally shorter compared to rural areas, the ambulance might arrive at the hospital before the ED team receives a notification, as 1 participant explained:

> *I think the biggest problem we have is that, at least here in our city, we have very short timeframes. We're talking about five minutes to get to the closest trauma center or thrombectomy center. And sometimes when we get to the hospital, and they [ED providers] don't even know why we are there because the notification hasn't been sent through yet. That's, you know, because you have to go through different people to get the message along. Also, because we are using radios and it takes several times for the radio transmission to go through.* [Participant 10]

To circumvent these problems, several participants revealed a work-around they have been using, that is, if they transport the patient to their affiliated hospital, they call the ED directly using their personal phones to expedite the notification process (Figure 3E).

Iterative Design of a Smart Glass Application for Addressing Challenges in the Prehospital Communication Process

Overview

Our iterative, user-centered design and evaluation process resulted in the creation of >20 design versions, with 2 major iterations evaluated in the first and second phases of usability testing. In this section, we focus on detailing the most necessary system features, as well as those explored but deemed unnecessary or unhelpful by EMS providers. Figure 4 presents the 2 major versions evaluated through usability testing and the final design of our application, with added notes in the figure to highlight significant design changes in each iteration.

XSL•FO
RenderX

Figure 4. Iterative design of the smart glass application. (A) The first major application version created and tested in phase 1. (B) The second major application version created and tested in phase 2. (C) The final design of the smart glass application. ED: emergency department; ETA: expected time of arrival; OLMC: online medical control.

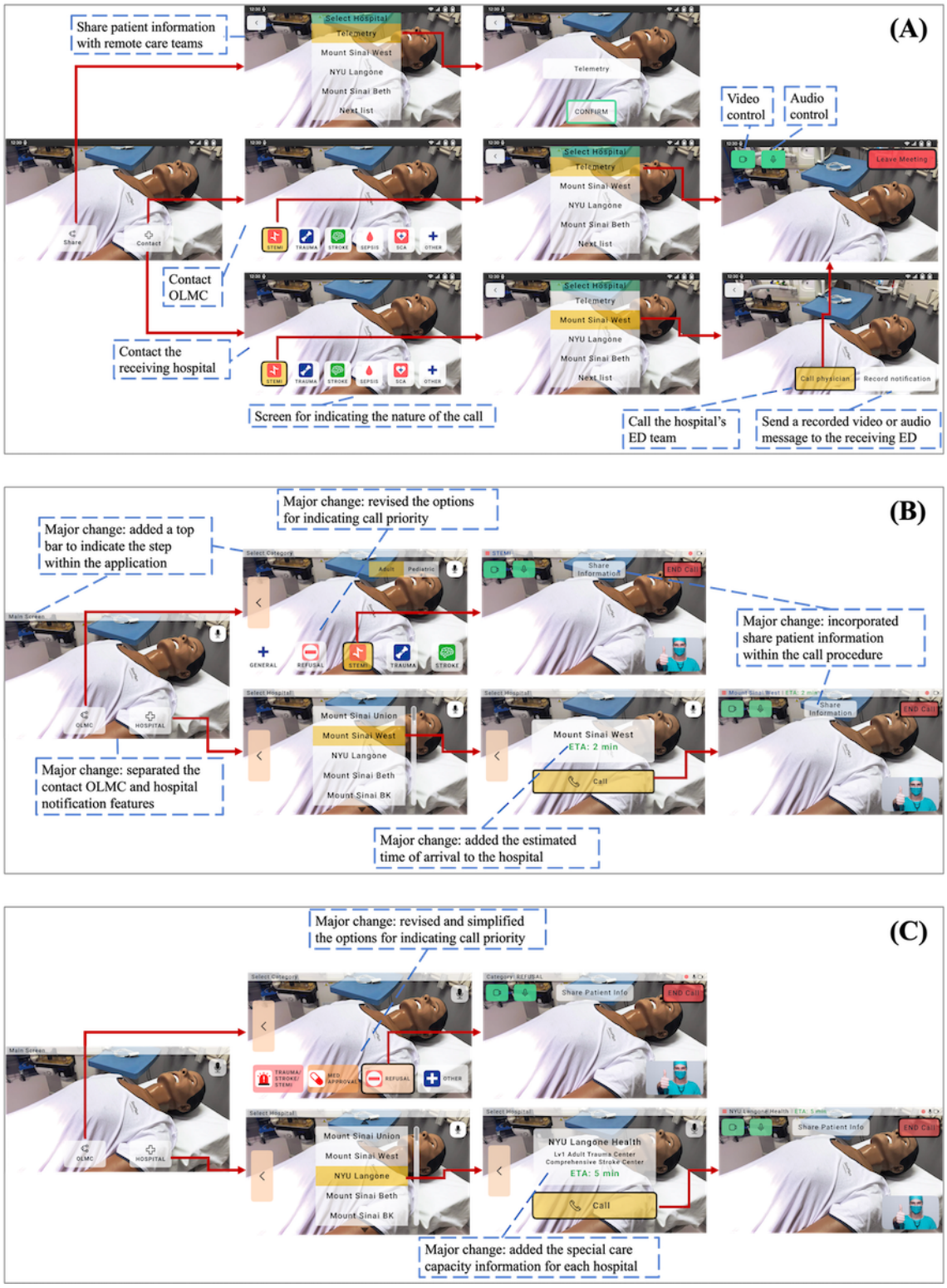

Most Needed System Features by EMS Providers

Overview

Through various design and evaluation studies, EMS providers highlighted five crucial system features that could improve the current prehospital communication process: (1) establishing video calls with OLMC physicians, (2) indicating call priority for expeditious OLMC contact, (3) enabling direct EMS-ED communication, (4) sharing multimedia patient information with remote care teams, and (5) enabling touchless device operation.

Establishing Video Calls With OLMC Physicians

Across all study sessions, there was a consensus among EMS providers that smart glasses could be invaluable for consulting

XSL•FO
RenderX

with OLMC physicians through video calls, enabling physicians to see exactly what EMS providers see. This feature could significantly enhance physicians' understanding of the patient's condition and needs, leading to more informed decision-making. Participants consistently expressed high satisfaction with this feature, rating it an average of 4.45 (out of 5) in phase 1 usability studies and an even higher 4.75 in phase 2 studies. One participant explained as follows:

> *So being able to get on the scene and use smart glasses and being able to say, 'well, this person really doesn't need an ambulance, or they could go to a clinic instead, or there's something very serious,' the doctor can, you know, can better help out.* [Participant 1]

On the basis of user feedback, we incorporated an intuitive, easy-to-operate video call feature for contacting OLMC, allowing EMS providers to turn the camera and audio on or off and leave the call.

Our initial design combined the options to contact OLMC and nearby hospitals into 1 list, with the OLMC option always focused and placed at the top (Figure 4A). This design intended to consolidate both OLMC contact and hospital notification on 1 screen for convenience and easy access. However, subsequent studies (eg, usability testing) prompted a significant design modification, that is, separating "Call OLMC" and "Notify Hospital" from the same list into different screens (Figure 4B). This change was driven by 2 primary considerations. First, EMS providers typically contact hospitals more frequently than OLMC physicians; positioning OLMC as the first option in the list could lead to extraneous device operations as EMS providers would always have to click multiple times to navigate down the list to find the hospital. Second, as described subsequently, our application includes a screen allowing EMS providers to indicate call priority. Participants noted that specifying call urgency was not necessary before connecting with the receiving hospital. Unlike contacting OLMC physicians, who need to prioritize which call to take first, the receiving hospital typically does not need to know the call urgency before picking up the EMS call. Participants in phase 2 usability testing confirmed that it is not necessary to indicate call priority when contacting the hospital, leading us to keep this revised design in the final version (Figure 4C).

Indicating Call Priority to Expedite OLMC Connection

Given the persistent issue of lengthy waiting times to connect with OLMC physicians, we designed a feature that allows EMS providers to indicate the urgency of the call. One participant highlighted the usefulness of this feature by stating the following:

> *It [smart glasses] can be used to solve a bigger problem in terms of the congestion of actually triaging patients with online medical control, which I would really like to highlight here. You could really solve those problems using this technology.* [Participant 3]

This feature underwent significant design changes across the 2 phases. As shown in Figure 4A, the initial design included categories for critical patient conditions such as "STEMI" (ST elevation myocardial infarction), "trauma," "stroke," "sepsis," and "SCA" (Sudden Cardiac Arrest). However, participants felt that limiting the call nature option to only critical scenarios could render it less useful, as all such scenarios have comparable urgency:

> *Prioritize. I think having just a critical or uncritical option to pick between is probably going to be more useful than having many critical options.* [Participant 19]

In addition, a few participants also saw the value in distinguishing between adult and pediatric patients, leading to a major design revision introducing a toggle to specify the patient type. We then refined the options within each type to reflect call priority. For instance, in the adult patient category, we included categories such as "general" and "refusal" for nonurgent consultations and categories such as "trauma," "stroke," and "STEMI" for critical conditions (Figure 4B). The "general" category represents nonurgent reasons for contacting OLMC, such as requests for permission for additional medication doses, and "refusal" pertains to situations where physician approval is needed when patients refuse medical attention. These revisions aimed to enhance distinctions between critical and noncritical scenarios and between adult and pediatric cases.

During phase 2 usability testing, feedback was received that the process of selecting between adult and pediatric patients and then choosing the nature of the call was overly complex. As a result, we refined the design to simplify user interaction by removing the required step to select between adult and pediatric patients. Participants also suggested revisions to the options for indicating call priority:

> *What I would suggest under the medical control would be a general alert like for medication approval. You know, you're gonna have a lot of stuff that's gonna fit under there. And then a separate refusal category. And lastly, a category combining STEMI, traumas, and stroke, like those big ones. So yeah, I think these pretty much have most cases covered.* [Participant 40]

We then followed their suggestions to include 4 options adaptable to different patient types: "medication approval" and "refusal" for common reasons for contacting OLMC; "trauma/stroke/STEMI" for conditions requiring immediate attention and prioritization by OLMC; and an "other" category for other types of or nonurgent issues (Figure 4C). To visually delineate the urgency of calls, different colors were used, ranging from red for critical conditions to white for the "other" category.

Facilitating Direct EMS-ED Communication

Through our research, it became evident that smart glasses can act as a valuable tool to facilitate direct communication between EMS and ED providers via video calls. This feature received consistently high usefulness ratings from participants (4.57 out of 5 in phase 1 of usability testing and 4.56 out of 5 in phase 2 of usability testing). One participant underscored that compared to relying solely on audio, having a direct video call with the

XSL•FO
RenderX

ED physician or nurse could make it easier to describe the patient's status, thus mitigating the risks of miscommunication and helping the receiving ED care team better understand the nature and needs of the incoming patients:

> *I really, really enjoyed how you can use this technology to report to the hospital. And you can actually do that much quicker [with smart glasses] than you would in real life. When we got on the scene, we can actually use smart glasses to easily describe what the actual scene looks like, if it's going to be an actual trauma, how many patients they are going to have, etc. That could help the [emergency department of the receiving] hospital get prepared better.* [Participant 3]

As shown in Figure 4, EMS providers can conveniently select a hospital from a list organized by proximity to the current location of the EMS team. Using embedded GPS, the smart glasses can automatically calculate and display the ETA to the chosen hospital, a feature universally found useful by participants (Figures 4B and 4C):

> *I think the ETA to the hospital is very helpful. So I don't have to look out the window of the ambulance and figure out my ETA to the hospital. It automatically tells that for me. So that way I can stay focused on my patient, instead of trying to figure out where my landmarks are.* [Participant 12]

Furthermore, during phase 2, participants suggested the inclusion of additional information for each hospital, such as the availability of trauma centers or pediatric care facilities, to aid EMS providers in choosing the most appropriate destination for patient needs:

> *Maybe you can put little icons next to each hospital name. So, who is a primary stroke center? Who is a primary trauma center? Who is a primary heart center? So, then they [EMS providers] can better determine which hospital to contact.* [Participant 22]

This addition is vital, as the nearest hospital might not always be equipped with specialized facilities (eg, a trauma center) to treat certain critical conditions. Following this suggestion, information regarding the specialized services available at each hospital was added to the final design (Figure 4C).

Sharing Multimedia Patient Information With Remote Care Teams

In phase 1 PD workshops, many of our participants emphasized the importance of sharing pivotal, textual, and visual patient information (eg, vital signs and photos) with remote care teams to facilitate their conversations. As 1 participant explained:

> *I don't know if I can transmit the information directly to the hospital or online medical control through this type of technology. If so, that would be the absolute best thing that you could ever imagine.* [Participant 4]

To address this crucial user need, we initially designed a unique feature allowing information sharing independently of initiating contact with hospitals or OLMC (Figure 4A). This feature was deemed very useful and received an average rating of 4.6 (out of 5) in phase 1 usability testing.

Despite the positive view of this feature during phase 1, we discovered that sharing information with the hospital or OLMC before establishing the call could create a problematic workflow. The primary reason is that the remote care team may simply ignore the shared information if it comes before the call is established, as 1 participant explained:

> *If hospitals don't have a dedicated person managing the shared information, then the usefulness of transmitting the information is limited. How would you transmit separately if you are not talking to the hospital?* [Participant 1]

On the basis of the received feedback, we made major changes to the design in phase 2. We integrated the feature of sharing patient information directly into the call feature, allowing EMS providers to share patient information only after the call is established (Figure 4B). This revised design received an average rating of 4.75 out of 5 during phase 2 usability testing, surpassing the rating received in phase 1. Given the overall positive feedback on this new design, we kept this design with minor wording modifications in our final version (Figure 4C).

Enabling Touchless Device Operation

Given the nature of EMS work, which often requires the use of hands and entails a risk of cross contamination, participants expressed a strong preference for touchless interaction methods to operate the smart glass device over using the default tangible buttons, as 1 participant explained:

> *So all hands being gloved up while we are taking care of a patient who's bleeding. You can't take your hand off when you're doing compressions to activate the glasses. If you have blood on your gloves, you're going to add it to the glasses. It will help so much if I don't need to touch the glasses.* [Participant 4]

In phase 1 PD workshops, we explored the preferred interaction methods of EMS providers. Among a set of touchless interaction methods [49], voice commands and hand gestures emerged as the 2 most preferred choices. Of the 16 participants (1 participant did not submit the ranking), 10 (62%) indicated voice commands as their top choice, while 5 (31%) participants chose hand gestures as their most preferred method, with 10 (62%) considering it their second preference.

On the basis of these preferences, we integrated both voice commands and hand gestures into our system prototype. More specifically, we used the Vuzix Software Development Kit to program a set of simple voice commands corresponding to text labels on the virtual buttons. To implement the hand gesture–based interactions, we used third-party gesture-sensing software (CrunchFish). This interaction method allows users to perform a pinch gesture by tapping their thumb and index finger together to choose an option on the interface (Figure 2B).

In the usability testing studies conducted during both phases, we assessed the user experience and usability of voice commands and hand gestures compared to the default interaction method—tangible buttons (detailed results are reported in the

XSL•FO
RenderX

study by Zhang et al [48]). A consistent finding across both phases of usability testing was that, compared to tangible buttons, voice commands and hand gestures had suboptimal task performance in all measured aspects (eg, task completion time and errors, etc). Despite this, our participants emphasized the importance and benefits of offering touchless interaction methods to minimize the need for physical operation of the smart glasses:

> *Because we're always using our hands, I think voice commands in parallel with what we're doing in the field are probably the best. And that's been our biggest request for certain applications we use in the field because we really wish there was a voice command functionality where we can continue to interact with patient but using voice command for the computing device. So yes, that I think voice demand is probably the best.* [Participant 43]

> *I liked the hand gestures, because it's very, it's very intuitive and the connection between me and the interface felt the strongest.* [Participant 19]

Unnecessary System Features

Overview

Throughout the design and evaluation process, we explored various features and design concepts, some of which were deemed unnecessary. In this section, we report 2 features that were extensively discussed and tested but were considered unnecessary or not useful by our participants.

Recording Short Audio and Video Messages for Hospital Notification

In phase 1 PD workshops, a few EMS providers expressed interest in a feature allowing them to record and send a short audio or video message to notify the receiving hospital about incoming patients. This feature was believed to be especially useful when a direct connection with the receiving hospital could not be established. One participant highlighted its potential time-saving benefits:

> *I'm more concerned about notifications to the ED with an unstable or critical patient. And I'm not really consulting, because I'm not asking for any real orders. I'm giving them very short, very brief information. So, what could be helpful is maybe just record a message and then maybe just click "Send," and then you could get a confirmation that the notification is delivered to your destination.* [Participant 1]

This user requirement informed our initial design to include a feature to deliver a prerecorded audio or video notification to the hospital ED team (Figure 4A).

However, this feature received mixed feedback during the usability testing studies in phase 1 and the following PD workshops in phase 2. While recognizing the potential time-saving benefits, EMS providers also raised concerns about new issues, such as providers opting not to call the hospital when they have the option to simply send a prerecorded message. There were also concerns about the receiving ED team overlooking the shared audio or video messages, potentially causing workflow disruptions and delays in preparation for patient arrival. For example, one participant explained his concern:

> *I don't see this being very useful; I see it as more frustrating than anything else. Because if I'm doing the recording for notification, this person is probably actively dying in front of me. So if I want to give a notification [to the hospital], I just want to call them and I want to tell them this is what's happening.* [Participant 19]

Given these concerns and the potential impact on current workflows, we decided to remove this feature from the prototype (Figures 4B and 4C).

Augmented Reality–Enabled Annotations

As the smart glass is powered by augmented reality technology, we explored the usefulness of enabling users to annotate real or virtual objects within their field of view. This would allow, for example, drawing a circle over a specific area on a patient's body to draw the attention of remote physicians during teleconsultation. It could also enable the remote consultant to annotate images captured from the live stream and project them back into the visual field of the smart glass wearer for real-time guidance. This innovative feature has been implemented and tested in prior work [50,51].

However, most (12/16, 75%) of our participants in phase 1 workshops deemed the annotation feature unnecessary for 2 main reasons. First, it is very rare for EMS providers to seek guidance from remote physicians on performing a procedure; therefore, the need for annotations is very limited, as 1 participant explained:

> *That's [using annotation] not likely going to happen. Because in order for us to be out in the field, we have to already have that training and those skills to be able to do that, whether it's BLS [basic life support] or ALS [advanced life support] level. So doctors are not going to tell you, "Oh, start an IV on the left or on the right," you know, paramedics already know. You know, if somebody is bleeding from a certain side, you don't need a doctor to tell you "Hey, patch that," you know that. I mean, I understand where you're going with that, like giving more concrete instruction to a junior or not very experienced provider. But you can probably just verbalize anyway.* [Participant 1]

Second, using the annotation feature could become a distraction, as it requires extra operations with the device, such as selecting an annotation tool from a list of options and then closing the annotation toolbox:

> *As far as the annotation tools, honestly, I think it's too much. I don't think that function would really be helpful. It might be more of a distraction, or it might not even get used.* [Participant 12]

XSL•FO
RenderX

User Experience and Potential Barriers to Using Smart Glasses

Overall User Experience

Overall, participants praised the user-friendly design of our application, finding it intuitive and easy to navigate. They expressed confidence in using the application with minimal training, as 1 participant stated:

> *Overall, it was pretty intuitive for me to use. I could see myself using it as is.* [Participant 24]

In addition, there is a consensus among almost all the participants that the smart glass application has great potential to improve communication efficiency and care coordination between distributed emergency care teams, as illustrated in the following 2 quotes:

> *It might be helpful because [when I] turn on the camera, they [physicians] would be able to see what we're seeing, instead of us just having to, you know, dictate that over radio. So I think that's kind of a big feature that would be useful there. I think more information would be passed along, compared to just how we do it right now.* [Participant 41]

> *It helps being able to give a heads up to the trauma team, so they know what to expect. Or even something like, let's say it was a stroke. And you've got the neuro team waiting on standby to see if it's really a stroke or not. It just gives so much more [information] and a clearer picture because you can actually show it [the patient] to them. I think that's really cool.* [Participant 12]

The SUS rating improved from a mean of 74.3 (SD 11.3) in the first phase of usability evaluation (classified as *good* usability [52]) to a mean of 80.3 (SD 13.1) in the second phase of usability evaluation (classified as *excellent* rating [52]). Moreover, users' willingness to use the system in the future increased from a mean of 4.42 out of 5 (SD 0.65) in phase 1 to a mean of 4.63 out of 5 (SD 0.62) in phase 2. Other user ratings, such as the application's fit with the EMS workflow and its ease of navigation, remained consistently high across both phases. This notable positive user feedback demonstrated the continuous enhancement of our system's design and usability through 2 phases.

The statistical analysis showed a significant difference in SUS ratings between EMS providers operating in urban areas (median 85, IQR 76-94) and those in rural areas (median 72.5, IQR 66-83; Mann-Whitney U=43; P=.17). Moreover, we observed a significant difference in SUS ratings between paramedics (median 72.5, IQR 70-80) and EMTs (median 85, IQR 74-98; Mann-Whitney U=44.5; P=.13). However, no significant difference was found in the ratings of different system features among participants with varying characteristics. Altogether, these findings imply that providers in urban areas and those with comparatively less training in treating patients in critical conditions perceived the smart glass application as more useful and user-friendly.

Potential Barriers to Adopting Smart Glasses in EMS Work

Overview

Despite the promising benefits of smart glasses in enhancing prehospital communication (eg, reducing the risk of miscommunication), our participants raised several potential challenges and concerns about adopting this technology in their work practice. As general issues related to the use of smart glasses—such as comfort, device durability, and ergonomic concerns—have been extensively reported in our prior studies [19,53], this paper specifically focuses on issues related to using smart glasses for prehospital communication. To better organize these challenges, we have categorized them into 3 groups: technical factors, environmental factors, and workflow integration considerations.

Technical Factors

A common concern about smart glasses is their reliance on high-speed network connectivity to establish video calls with remote care teams. Participants emphasized the necessity of ensuring stable and high-bandwidth internet connections, especially in areas with limited network coverage such as rural areas or subways. When the internet is unavailable, using smart glasses becomes impractical, as 1 participant explained:

> *I feel like it may be an issue with a lot of places we go around, such as some buildings and rural areas. So, it [internet] can be spotty. That even seems to be an issue with the current phones. It [stable internet connection] is critical for using smart glasses in the field.* [Participant 13]

Furthermore, EMS providers may be dispatched for search and rescue tasks that can extend over long durations. In such scenarios, the device's battery life becomes crucial to ensuring uninterrupted communication between EMS providers and remote experts. One participant voiced the following concern:

> *Another concern for me is battery life. Especially since I can see it being useful in long-term rescue scenarios that can take time.* [Participant 6]

Environmental Factors

The dynamic and noisy setting of EMS work poses challenges for effective communication. Our participants expressed concerns about their ability to hear remote physicians clearly through the speakers of the smart glasses. One participant explained as follows:

> *In the back of an ambulance, it can be very loud, with sirens, horns, driving through town, hitting bumps, and everything rattling back there. Additionally, the monitor is beeping at you. So, I just want to make sure I'd be able to hear it.* [Participant 42]

Workflow Integration Considerations

Participants also discussed the workflow-related considerations associated with implementing the system. To fully use smart glasses, a counterpart system needs to be deployed in both OLMC offices and hospitals' ED departments, requiring significant integration efforts:

XSL•FO
RenderX

> *You guys would have to integrate [the application] into every single notification system. Because right now, the notification system is kind of just like a red phone that's in the emergency room.* [Participant 35]

Moreover, implementing this new system in receiving care teams may impact their workflow, potentially necessitating the assignment of a dedicated person to answer video calls. One participant highlighted this barrier:

> *You need internal systems in the hospital to function properly. You need your hospital to know that they're going to be receiving this [video call and have someone ready to handle that one as it happens.* [Participant 19]

Finally, the EMS system in the United States is complex, involving multiple stakeholders beyond EMS agencies and hospitals, such as state and city regulators, insurance companies, etc. This complexity may present considerable challenges to the widespread adoption of smart glass technology, as 1 participant noted:

> *It's a huge task to get this integrated with the billing service, ambulance services, software vendors, and many other things. It would also be quite an effort for every one of your partner hospitals to have an integrated video platform for communication. I think it's going to be interesting to see how that plays out.* [Participant 36]

Discussion

Principal Findings

To the best of our knowledge, this study is the first of its kind that uses a user-centered design approach to explore the appropriate design of smart glasses in facilitating communication and care coordination between prehospital and remote care providers. Our approach is deeply rooted in PD, which promises to create user-friendly clinical systems and thoroughly addresses user needs [22,29,30,38]. In addition, our iterative design process involved 43 distinct EMS providers with diverse characteristics (eg, fire-based vs hospital-based agencies and urban vs rural areas). This approach is beneficial, as it can significantly increase the generalizability of our findings, effectively identify and address various user needs, and ensure that providers with different roles and occupations (eg, paramedics vs EMTs) can equally voice their needs and concerns.

Our work revealed that the prehospital communication process encompasses 2 critical activities: EMS providers reaching out to OLMC physicians for guidance and medical approval (eg, approving medication administration and determining patient destination) and notifying the receiving hospital about patient arrival. A significant challenge in both activities arises from the limitations of current communication mechanisms (eg, radio or phone) in conveying contextual patient information, often resulting in lengthy verbal descriptions or even miscommunication. Moreover, the shortage of OLMC physicians frequently causes delays in connecting with EMS providers, adversely affecting the efficiency and promptness of patient care in the field. Finally, the communication between EMS and receiving ED teams involves multiple layers, leading to unnecessary complexity in the process. Establishing a direct EMS-ED communication link was recognized as an essential need, as it holds the potential to minimize the chances of miscommunication during critical situations.

These findings, as well as specific user requirements elicited through design workshops and usability testing, informed our system design. The final version of our application incorporates several key features: (1) a video call functionality to contact OLMC physicians, allowing EMS providers to convey contextual patient information; (2) an option to indicate the call severity and priority to expedite the connection with OLMC physicians; (3) a direct video-based communication link with the ED team at the receiving hospital, supplemented with essential information such as ETA and the hospital's special care capabilities, facilitating effective communication and aiding in patient destination decision-making; (4) an option to share critical, multimedia patient information (eg, photos, demographics, treatments, and vital signs) with remote care teams, augmenting the video call and providing comprehensive details for more informed medical decisions; (5) integration of touchless interaction methods, enabling hands-free operation and reducing the likelihood of cross contamination, a crucial aspect given the nature of EMS work.

The ratings given by the participants for our application showed a notable increase from phase 1 to phase 2, indicating continuous improvement in our system design and increased user acceptance of our application. Throughout the study, the study participants consistently highlighted the great potential and benefits of using smart glasses in prehospital communication, such as enhancing communication efficiency and care coordination between distributed emergency care teams. Despite this, the participants expressed a few concerns related to technical factors, environmental challenges, and workflow integration issues. These valuable insights provide important considerations for the further refinement and implementation of smart glasses in real-world settings.

Study Implications and Future Directions

The effective use of smart glasses by EMS providers hinges on their access to a high-bandwidth cellular network to establish quality video and audio calls with remote care teams. However, EMS providers often find themselves in areas with limited signal coverage, presenting challenges for the effective use of smart glasses. Not surprisingly, the issue of reliable internet access for the field deployment of smart glasses was raised as a major concern by many participants and has been reported in previous studies [19,53]. A potential solution lies in harnessing the capabilities of 5G technology to enhance the network connectivity of smart glasses. In addition, ongoing initiatives to establish a dedicated broadband network for first responders [54] promise to address the connectivity issues encountered by EMS providers.

In addition, EMS providers mentioned several other potential barriers (eg, battery life and environmental constraints) to adopting smart glass technology in real practice. Compared to the first generation of smart glasses released almost a decade

XSL•FO
RenderX

ago, recent advancements in smart glass hardware have the potential to address some of the users' concerns. For instance, the device we used—the Vuzix M400—is waterproof, has up to 12 hours of battery life, and can operate in a wide range of temperatures (from −20 °C to 45 °C). These hardware advancements can make the smart glass device adaptable to the EMS setting.

The diverse characteristics of our EMS participants allowed us to explore the disparities in their perceptions of smart glasses. This study revealed that providers in urban areas, especially those with less experience in treating patients in critical conditions, perceived the smart glass application as more beneficial and user-friendly. These insights underscore the fact that user perspectives and preferences concerning smart glasses can differ based on location and occupation. Essentially, the notion of a universal solution (eg, "one size fits all") may not be applicable when designing smart glass applications for EMS providers. Thus, a smart glass–based telemedicine application must be adaptable to meet the distinct needs of EMS agencies in various locations.

An integral aspect of our smart glass application is its capability to share patient information with remote care teams. To realize this feature, the smart glass application needs to be integrated with the existing systems used by EMS providers, such as vital signs monitors and electronic health record systems. However, implementing these integrations could face real-world challenges, primarily due to persistent issues in health system interoperability [55]. Future work could explore viable technical and organizational strategies to effectively integrate smart glasses with medical devices and systems (eg, collaborating with electronic health record vendors and adopting the Fast Healthcare Interoperability Resources standard [56] to enhance their interoperability). In addition, it is critical to ensure that data transmission between smart glasses and medical systems complies with the rules and regulations of the Health Insurance Portability and Accountability Act [57].

This study primarily focused on the design of a smart glass application tailored for EMS providers; however, there are other stakeholders that are critical for the successful implementation of this technology. For example, fully realizing the potential of smart glasses also requires developing and deploying corresponding telemedicine applications for remote care teams, such as OLMC physicians and hospital ED teams. Our forthcoming work involves engaging with these key stakeholders to determine the optimal design and implementation strategy for their respective telemedicine systems. In addition, as emphasized by our participants, other stakeholders beyond care providers, such as EMS regulators and insurance companies, play a critical role in regulating EMS work and affecting technology adoption. Therefore, future research should explore the viewpoints of these additional stakeholders. Attaining "buy-in" from the leadership of these key stakeholders is a critical determining factor in adopting smart glasses [53].

Finally, it is worth noting that technology solutions such as smart glasses or other telemedicine systems can only partially address the longstanding challenges in the prehospital communication process. Some challenges are systemic and must be tackled at the organizational and policy levels. For example, the shortage of OLMC physicians is a major reason for delayed responses to EMS calls and requests. Although our smart glass design includes features to ensure that critical calls are answered promptly by OLMC physicians, it cannot completely resolve issues stemming from the lack of OLMC physicians.

Limitations

This study has certain limitations that should be acknowledged. First, we did not evaluate the application in a natural working environment, meaning we have not fully assessed how smart glasses are used in real practice. To address this, we plan to conduct system testing in near–real-life simulations in the future to investigate the effectiveness of smart glasses in enhancing patient care in the field. Second, although we identified and designed a feature for expedited connection with OLMC physicians, we did not elicit the perspectives of OLMC physicians on its usefulness. To address this gap, we plan to seek feedback and insights from OLMC physicians in future studies, with a specific focus on the call priority and video call features. Third, our study was conducted in the United States, and the attributes and dynamics of prehospital communication here may differ from those in other countries. It is likely that the smart glass application we developed is not applicable to the EMS contexts in different nations. Nonetheless, the design insights and user perspectives on smart glasses could provide valuable guidance for researchers in other regions developing smart glass solutions for emergency care teams. Finally, the evaluation of the smart glass application focused primarily on usability and user perceptions, without assessing the impact of our application on clinical outcomes. In our future work, we will conduct simulations to test the effectiveness of smart glasses in improving the prehospital communication process and whether using smart glasses could help make better decisions and reduce medical errors.

Conclusions

This paper presents a 2-year long, user-centered design research project on a smart glass application aimed at enhancing prehospital communication. Our work illustrated critical features deemed useful by EMS providers, such as video call functionality for expedited OLMC contact, direct video-based communication with ED teams, multimedia patient information sharing, and touchless interaction mechanisms. Rooted in PD, the study has yielded invaluable insights into leveraging smart glasses for enhancing communication and care coordination in the dynamic prehospital environment. Our work reaffirms the critical value of user-centered design in health care technology innovation. The findings and lessons learned from our study are poised to guide future work on the implementation and integration of smart glasses into prehospital care.

XSL•FO
RenderX

Acknowledgments

This study was supported by the National Science Foundation (1948292) and the Agency for Healthcare Research and Quality (1R21HS028104-01A1). The authors extend their heartfelt gratitude to all the emergency medical service providers who participated in the research. The authors would also like to express their sincere appreciation to Jack Finkelstein at Interfaith Medical Center (New York City, New York), Humaira Ali at Maimonides Medical Center (New York City, New York), and Brittany Olvera at Colorado Springs Fire (Colorado Springs, Colorado) for their support in coordinating the studies and recruiting participants.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to institutional review board restrictions and the need to protect human participants' confidentiality, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participant characteristics.
[DOCX File , 16 KB - jmir_v26i1e53157_app1.docx]

Multimedia Appendix 2
Participatory design workshop protocol.
[PDF File (Adobe PDF File), 1752 KB - jmir_v26i1e53157_app2.pdf]

Multimedia Appendix 3
Usability testing protocol.
[PDF File (Adobe PDF File), 129 KB - jmir_v26i1e53157_app3.pdf]

References

1. Zhang Z, Sarcevic A, Bossen C. Constructing common information spaces across distributed emergency medical teams. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017 Presented at: CSCW '17; February 25-March 1, 2017; Portland, OR. [doi: 10.1145/2998181.2998328]
2. Zhang Z, Sarcevic A, Joy K, Ozkaynak M, Adelgais K. User needs and challenges in information sharing between pre-hospital and hospital emergency care providers. AMIA Annu Symp Proc 2021;2021:1254-1263 [FREE Full text] [Medline: 35308923]
3. Reddy MC, Paul SA, Abraham J, McNeese M, DeFlitch C, Yen J. Challenges to effective crisis management: using information and communication technologies to coordinate emergency medical services and emergency department teams. Int J Med Inform 2009 Apr;78(4):259-269. [doi: 10.1016/j.ijmedinf.2008.08.003] [Medline: 18835211]
4. Bergrath S, Rörtgen D, Rossaint R, Beckers SK, Fischermann H, Brokmann JC, et al. Technical and organisational feasibility of a multifunctional telemedicine system in an emergency medical service - an observational study. J Telemed Telecare 2011 Sep 20;17(7):371-377. [doi: 10.1258/jtt.2011.110203] [Medline: 21933897]
5. Cho SJ, Kwon IH, Jeong J. Application of telemedicine system to prehospital medical control. Healthc Inform Res 2015 Jul;21(3):196-200 [FREE Full text] [doi: 10.4258/hir.2015.21.3.196] [Medline: 26279957]
6. Xiao Y, Gagliano D, LaMonte M, Hu P, Gaasch W, Gunawadane R, et al. Design and evaluation of a real-time mobile telemedicine system for ambulance transport. Proc AMIA Symp 1998:1102 [FREE Full text] [doi: 10.1109/iembs.1999.803862]
7. Johansson A, Esbjörnsson M, Nordqvist P, Wiinberg S, Andersson R, Ivarsson B, et al. Technical feasibility and ambulance nurses' view of a digital telemedicine system in pre-hospital stroke care - a pilot study. Int Emerg Nurs 2019 May;44:35-40. [doi: 10.1016/j.ienj.2019.03.008] [Medline: 31047854]
8. Gilligan P, Bennett A, Houlihan A, Padki A, Owen N, Morris D, et al. The doctor can see you now: a key stakeholder study into the acceptability of ambulance based telemedicine. Ir Med J 2018 Jun 07;111(6):769. [Medline: 30518784]
9. Yperzeele L, Van Hooff RJ, De Smedt A, Valenzuela Espinoza A, Van Dyck R, Van de Casseye R, et al. Feasibility of AmbulanCe-Based Telemedicine (FACT) study: safety, feasibility and reliability of third generation in-ambulance telemedicine. PLoS One 2014 Oct 24;9(10):e110043 [FREE Full text] [doi: 10.1371/journal.pone.0110043] [Medline: 25343246]
10. Geisler F, Kunz A, Winter B, Rozanski M, Waldschmidt C, Weber JE, Stroke Emergency Mobile (STEMO) Consortium. Telemedicine in prehospital acute stroke care. J Am Heart Assoc 2019 Mar 19;8(6):e011729 [FREE Full text] [doi: 10.1161/JAHA.118.011729] [Medline: 30879372]

XSL•FO
RenderX

11. Chapman Smith SN, Brown PC, Waits KH, Wong JS, Bhatti MS, Toqeer Q, et al. Development and evaluation of a user-centered mobile telestroke platform. Telemed J E Health 2019 Jul;25(7):638-648 [FREE Full text] [doi: 10.1089/tmj.2018.0044] [Medline: 30207927]
12. Rogers H, Madathil KC, Agnisarman S, Narasimha S, Ashok A, Nair A, et al. A systematic review of the implementation challenges of telemedicine systems in ambulances. Telemed J E Health 2017 Sep;23(9):707-717. [doi: 10.1089/tmj.2016.0248] [Medline: 28294704]
13. Momen KS. Identifying nursing activities to estimate the risk of cross-contamination. University of Toronto. 2013 Jan 7. URL: https://tspace.library.utoronto.ca/handle/1807/34914 [accessed 2024-11-02]
14. Schaer R, Melly T, Muller H, Widmer A. Using smart glasses in medical emergency situations, a qualitative pilot study. In: Proceedings of the IEEE Wireless Health. 2016 Presented at: WH 2016; October 25-27, 2016; Bethesda, MD. [doi: 10.1109/wh.2016.7764556]
15. Romare C, Skär L. Smart glasses for caring situations in complex care environments: scoping review. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e16055 [FREE Full text] [doi: 10.2196/16055] [Medline: 32310144]
16. Broach J, Hart A, Griswold M, Lai J, Boyer EW, Skolnik AB, et al. Usability and reliability of smart glasses for secondary triage during mass casualty incidents. Proc Annu Hawaii Int Conf Syst Sci 2018 Jan 03;2018:1416-1422 [FREE Full text] [Medline: 29398976]
17. Cicero MX, Walsh B, Solad Y, Whitfill T, Paesano G, Kim K, et al. Do you see what I see? Insights from using Google Glass for disaster telemedicine triage. Prehosp Disaster Med 2015 Feb;30(1):4-8. [doi: 10.1017/S1049023X1400140X] [Medline: 25571779]
18. Follmann A, Ohligs M, Hochhausen N, Beckers SK, Rossaint R, Czaplik M. Technical support by smart glasses during a mass casualty incident: a randomized controlled simulation trial on technically assisted triage and telemedical app use in disaster medicine. J Med Internet Res 2019 Jan 03;21(1):e11939 [FREE Full text] [doi: 10.2196/11939] [Medline: 30609988]
19. Zhang Z, Bai E, Joy K, Ghelaa PN, Adelgais K, Ozkaynak M. Smart glasses for supporting distributed care work: systematic review. JMIR Med Inform 2023 Feb 28;11:e44161 [FREE Full text] [doi: 10.2196/44161] [Medline: 36853760]
20. Schlosser P, Matthews B, Salisbury I, Sanderson P, Hayes S. Head-worn displays for emergency medical services staff: properties of prehospital work, use cases, and design considerations. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445614]
21. Zhang Z, Joy K, Harris R, Ozkaynak M, Adelgais K, Munjal K. Applications and user perceptions of smart glasses in emergency medical services: semistructured interview study. JMIR Hum Factors 2022 Feb 28;9(1):e30883 [FREE Full text] [doi: 10.2196/30883] [Medline: 35225816]
22. Kristensen M, Kyng M, Palen L. Participatory design in emergency medical service: designing for future practice. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2006 Presented at: CHI '06; April 22-27, 2006; Montréal, Québec. [doi: 10.1145/1124772.1124798]
23. Schuler D, Namioka A. Participatory Design: Principles and Practices. Boca Raton, FL: CRC Press; 1993.
24. Ozkaynak M, Sircar CM, Frye O, Valdez RS. A systematic review of design workshops for health information technologies. Informatics 2021 May 14;8(2):34. [doi: 10.3390/informatics8020034]
25. Viitanen J, Hyppönen H, Lääveri T, Vänskä J, Reponen J, Winblad I. National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int J Med Inform 2011 Oct;80(10):708-725. [doi: 10.1016/j.ijmedinf.2011.06.010] [Medline: 21784701]
26. Viitanen J, Kuusisto A, Nykänen P. Usability of electronic nursing record systems: definition and results from an evaluation study in Finland. Stud Health Technol Inform 2011;164:333-338. [Medline: 21335733]
27. Militello L, Patterson ES, Tripp-Reimer T, Asch SM, Fung CH, Glassman P, et al. Clinical reminders: why don't they use them? Proc Hum Factors Ergon Soc Annu Meet 2004 Sep 01;48(15):1651-1655. [doi: 10.1177/154193120404801505]
28. Sarcevic A, Ferraro N. On the use of electronic documentation systems in fast-paced, time-critical medical settings. Interact Comput 2017 Mar;29(2):203-219. [doi: 10.1093/iwc/iww021]
29. Kusunoki DS, Sarcevic A, Weibel N, Marsic I, Zhang Z, Tuveson G, et al. Balancing design tensions: iterative display design to support ad hoc and multidisciplinary medical teamwork. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON. [doi: 10.1145/2556288.2557301]
30. Kusunoki D, Sarcevic A, Zhang Z, Yala M. Sketching awareness: a participatory study to elicit designs for supporting ad hoc emergency medical teamwork. Comput Support Coop Work 2015 Feb 01;24(1):1-38 [FREE Full text] [doi: 10.1007/s10606-014-9210-5] [Medline: 25870498]
31. Østervang C, Lassen A, Schmidt T, Coyne E, Dieperink KB, Jensen CM. Development of a health information system to promote emergency care pathways: a participatory design study. Digit Health 2022 Dec 26;8:20552076221145856 [FREE Full text] [doi: 10.1177/20552076221145856] [Medline: 36601282]
32. Mitrasinovic S, Camacho E, Trivedi N, Logan J, Campbell C, Zilinyi R, et al. Clinical and surgical applications of smart glasses. Technol Health Care 2015;23(4):381-401. [doi: 10.3233/THC-150910] [Medline: 26409906]

XSL•FO
RenderX

33. Gupta S, Boehme J, Manser K, Dewar J, Miller A, Siddiqui G, et al. Does wearable medical technology with video recording capability add value to on-call surgical evaluations? Surg Innov 2016 Oct 09;23(5):498-504. [doi: 10.1177/1553350616656278] [Medline: 27335083]
34. Ozkaynak M, Kebapci A, Ponicsan H, Zhang Z, Bai E, Bowler F. Evaluating smart glasses for cardiopulmonary resuscitation. Stud Health Technol Inform 2024 Jul 24;315:19-24. [doi: 10.3233/SHTI240099] [Medline: 39049219]
35. Ozkaynak M, Reeder B, Park SY, Huh-Yoo J. Design for improved workflow. In: Sethumadhavan A, Sasangohar F, editors. Design for Health: Applications of Human Factors. Cambridge, MA: Academic Press; 2020.
36. Nielsen J, Molich R. Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1990 Presented at: CHI '90; April 1-5, 1990; Seattle, WA. [doi: 10.1145/97243.97281]
37. Nielsen J. Usability inspection methods. In: Proceedings of the Conference Companion on Human Factors in Computing Systems. 1994 Presented at: CHI '94; April 24-28, 1994; Boston, MA. [doi: 10.1145/259963.260531]
38. Pollack A, Miller A, Mishra SR, Pratt W. PD-atricians: leveraging physicians and participatory design to develop novel clinical information tools. AMIA Annu Symp Proc 2016;2016:1030-1039 [FREE Full text] [Medline: 28269900]
39. Kusunoki DS, Sarcevic A. Designing for temporal awareness: the role of temporality in time-critical medical teamwork. CSCW Conf Comput Support Coop Work 2015 Mar;2015:1465-1476 [FREE Full text] [doi: 10.1145/2675133.2675279] [Medline: 27478880]
40. Terhoeven J, Wischniewski S. How to evaluate the usability of smart devices as conceivable work assistance: a systematic review. In: Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society. 2016 Presented at: GfA 2016; March 2-4, 2016; Aachen, Germany. [doi: 10.1007/978-3-662-53305-5_19]
41. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
42. Klinker K, Wiesche M, Krcmar H. Digital transformation in health care: augmented reality for hands-free service innovation. Inf Syst Front 2019 Jun 25;22(6):1419-1431. [doi: 10.1007/s10796-019-09937-7]
43. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
44. Nistor N, Lerche T, Weinberger A, Ceobanu C, Heymann O. Towards the integration of culture into the Unified Theory of Acceptance and Use of Technology. Br J Educ Technol 2012 Nov 19;45(1):36-55. [doi: 10.1111/j.1467-8535.2012.01383.x]
45. Charmaz K. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Thousand Oaks, CA: SAGE Publications; 2006.
46. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 2006 Mar 01;5(1):80-92. [doi: 10.1177/160940690600500107]
47. Dourish P. Reading and interpreting ethnography. In: Olson J, Kellogg W, editors. Ways of Knowing in HCI. New York, NY: Springer; 2014.
48. Zhang Z, Bai E, Stepanian A, Jagannath S, Park SY. Touchless interaction for smart glasses in emergency medical services: user needs and experiences. Int J Hum Comput Interact 2024 Mar 19:1-20. [doi: 10.1080/10447318.2024.2328917]
49. Lee LH, Hui P. Interaction methods for smart glasses: a survey. IEEE Access 2018;6:28712-28732. [doi: 10.1109/access.2018.2831081]
50. McCullough M, Kulber L, Sammons P, Santos P, Kulber DA. Google glass for remote surgical tele-proctoring in low- and middle-income countries: a feasibility study from Mozambique. Plast Reconstr Surg Glob Open 2018 Dec;6(12):e1999 [FREE Full text] [doi: 10.1097/GOX.0000000000001999] [Medline: 30656104]
51. Ponce BA, Menendez ME, Oladeji LO, Fryberger CT, Dantuluri PK. Emerging technology in surgical education: combining real-time augmented reality and wearable computing devices. Orthopedics 2014 Nov;37(11):751-757. [doi: 10.3928/01477447-20141023-05] [Medline: 25361359]
52. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J User Exp 2009;4(3):114-123 [FREE Full text]
53. Zhang Z, Ramiya Ramesh Babu NA, Adelgais K, Ozkaynak M. Designing and implementing smart glass technology for emergency medical services: a sociotechnical perspective. JAMIA Open 2022 Dec;5(4):ooac113 [FREE Full text] [doi: 10.1093/jamiaopen/ooac113] [Medline: 36601367]
54. FirstNet Authority. URL: https://firstnet.gov/ [accessed 2024-11-02]
55. M S, Chacko A. Interoperability issues in EHR systems: research directions. In: Lee KC, Samui P, Roy SS, Kumar V, editors. Data Analytics in Biomedical Engineering and Healthcare. Cambridge, MA: Academic Press; 2020:13-28.
56. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
57. Chai PR, Wu RY, Ranney ML, Porter PS, Babu KM, Boyer EW. The virtual toxicology service: wearable head-mounted devices for medical toxicology. J Med Toxicol 2014 Dec 19;10(4):382-387 [FREE Full text] [doi: 10.1007/s13181-014-0420-5] [Medline: 25135308]

XSL•FO
RenderX

Abbreviations

ED: emergency department
EMS: emergency medical service
EMT: emergency medical technician
ETA: estimated time of arrival
HIT: health information technology
NDK: Native Developer Kit
OLMC: online medical control
PD: participatory design
STEMI: ST elevation myocardial infarction
SUS: System Usability Scale

Edited by A Mavragani; submitted 27.09.23; peer-reviewed by A Gangadhara Rao, D Singh, M Fassl; comments to author 06.05.24; revised version received 13.08.24; accepted 16.10.24; published 29.11.24.

Please cite as:
Zhang Z, Bai E, Xu Y, Stepanian A, Kutzin JM, Adelgais K, Ozkaynak M
A Smart Glass Telemedicine Application for Prehospital Communication: User-Centered Design Study
J Med Internet Res 2024;26:e53157
URL: https://www.jmir.org/2024/1/e53157
doi:10.2196/53157
PMID:

©Zhan Zhang, Enze Bai, Yincao Xu, Aram Stepanian, Jared M Kutzin, Kathleen Adelgais, Mustafa Ozkaynak. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patients' and Clinicians' Experiences Using a Real-Time Remote Monitoring System for Chemotherapy Symptom Management (ASyMS): Qualitative Study

Lisa McCann[1], BSc (Hons), MSc, PhD; Liane Lewis[2], BSc, PhD; Olubukola Oduntan, PhD; Jenny Harris[3], BSc, MSc, PhD; Andrew Darley[4], BA, MSc, PhD; Geir V Berg[5], MSc, PhD; Simone Lubowitzki[6], MA, PhD; Katy Cheevers[7], MSc; Morven Miller[1], BA, RGN, MSc, PhD; Jo Armes[3], RGN, BSc, MSc, PhD; Emma Ream[3], BSc, MSc, PhD; Patricia Fox[4], BNS, MA, PhD; Eileen Patricia Furlong[4], MMed Sc, PhD; Alexander Gaiger[6], MD; Grigorios Kotronoulas[8], BSN, MSc, PhD; Elisabeth Patiraki[9], RN, BSc, PhD; Stylianos Katsaragakis[9], BA, MSc, PhD; Paul McCrone[10], BA (Hons), MSc, PhD; Christine Miaskowski[11,12], RN, FAAN, PhD; Antonella Cardone[13], MS, MBA; Dawn Orr[14], BSc, MSc, PhD; Adrian Flowerday[15], BSc, CEng; Simon Skene[16], MSc, PhD; Margaret Moore, BSc (Hons); Nicosha De Souza[17], PhD; Peter Donnan[18], BA, MSc, PhD; Roma Maguire[1], BN, MSci, PhD

[1]Digital Health and Wellness Group (DHaWG), Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom
[2]University Hospital Jena, Jena, Germany
[3]University of Surrey, School of Health Sciences, Guildford, United Kingdom
[4]School of Nursing, Midwifery and Health Systems, University College Dublin, Dublin, Ireland
[5]Section for Nursing, Faculty of Social and Health Sciences, Inland Norway University of Applied Sciences, Elverum, Norway
[6]Department of Internal Medicine, Division of Hematology and Hemostaseology, Medical University of Vienna, Vienna, Austria
[7]Hastings and Rother Healthcare, Hastings, United Kingdom
[8]School of Medicine, Dentistry & Nursing, University of Glasgow, Glasgow, United Kingdom
[9]School of Health Sciences, National and Kapodistrian University of Athens, Athens, Greece
[10]Institute for Lifecourse Development, Faculty of Education, Health & Human Sciences, University of Greenwich, London, United Kingdom
[11]Department of Anesthesia and Perioperative Care, University of California, San Francisco, CA, United States
[12]Department of Physiological Nursing, University of California, San Francisco, CA, United States
[13]Cancer Patients Europe, Brussels, Belgium
[14]NHS 24, Glasgow, United Kingdom
[15]Docobo Limited, Leatherhead, United Kingdom
[16]Surrey Clinical Trials Unit, University of Surrey, Guildford, United Kingdom
[17]Biomathematics and Statistics Scotland (BioSS), The Rowett Institute, Aberdeen, United Kingdom
[18]Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, United Kingdom

Corresponding Author:
Lisa McCann, BSc (Hons), MSc, PhD
Digital Health and Wellness Group (DHaWG)
Department of Computer and Information Sciences
University of Strathclyde
Livingstone Tower
26 Richmond Street
Glasgow, G1 1XQ
United Kingdom
Phone: 44 +441415483587
Email: lisa.mccann@strath.ac.uk

Abstract

Background: Patients receiving chemotherapy require ongoing symptom monitoring and management to optimize their outcomes. In recent years, digital remote monitoring interventions have emerged to provide enhanced cancer care delivery experiences to

XSL•FO
RenderX

patients and clinicians. However, patient and clinician experiential evaluations of these technologies are rare. Therefore, we explored user experiences and perceptions of one such intervention—Advanced Symptom Management System (ASyMS)—after its scaled deployment in the context of the Electronic Symptom Management System Remote Technology (eSMART) trial. The eSMART trial was a large, multicenter randomized controlled trial to evaluate the efficacy of ASyMS in 12 clinical sites in 5 European countries.

Objective: In this qualitative study, both patients' and clinicians' experiences of using ASyMS for up to 6 cycles of chemotherapy were explored to understand the impact of ASyMS on patients' experiences, clinical practice, and supportive care delivery.

Methods: For this analysis, individual, semistructured, one-to-one interviews with 29 patients with breast, colorectal, and hematological cancers and 18 clinicians from Austria, Greece, Ireland, Norway, and the United Kingdom were conducted. Interviews focused on patients' and clinicians' experiences of using ASyMS, care organization and changes in practice following the introduction of ASyMS, perceived changes in care associated with the use of ASyMS, and its potential for future integration into routine chemotherapy care pathways.

Results: Thematic analysis identified several themes that describe patients' and clinicians' experiences using ASyMS. One central orienting theme—ASyMS as a facilitator of change—was supported by 5 key themes associated with human and technology monitoring: reassurance, enhanced communications and relationships, knowing what is "normal" and what is to be expected, enhancing cancer care experiences, and informing future cancer care.

Conclusions: This study is the first to evaluate both patients' and clinicians' experiences of using a digital health intervention to remotely monitor chemotherapy symptoms across 5 countries. Experiences with ASyMS were positive from both patients' and clinicians' perspectives, although some improvements to support the wider-scale rollout and sustained implementation were identified. Overall, this study demonstrates that real-time remote monitoring systems can help patients feel more reassured during their chemotherapy treatments and can help clinicians provide the right care, at the right time, and in the right place.

Trial Registration: ClinicalTrials.gov NCT02356081; https://clinicaltrials.gov/study/NCT02356081

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2016-015016

(J Med Internet Res 2024;26:e53834) doi:10.2196/53834

KEYWORDS

cancer; clinician experiences; digital interventions; patient experiences; remote monitoring; qualitative methods

Introduction

Background

Patients receiving chemotherapy experience a range of adverse effects. Several of these effects can be severe or life-threatening, particularly if not reported promptly and managed effectively [1]. Traditionally, chemotherapy symptom management relies on patient recall, but this approach is limited because it is prone to recall bias and inaccuracy [2]. The lack of real-time, accurate reporting of symptoms impedes timely interventions [3]. Therefore, a growing momentum exists within cancer care services to introduce electronic patient-reported outcome measures (PROMs) to better support real-time clinical practice [4]. Remote patient monitoring systems that facilitate regular assessment of symptoms [5] can influence changes to clinical practice by assisting rapid clinical decision-making and interventions and supporting improvements in patients' experiences of cancer care [6] and outcomes [7].

Digital interventions for symptom monitoring in cancer care are critical due to the increasing global burden of cancer [8], the increasing complexity of cancer treatments [9], and the shortage of cancer care health professionals worldwide [10]. Implementing innovative models and services such as remote monitoring technologies can influence positive changes to deliver safe and optimal cancer care [6,11,12].

Therefore, learning from and reflecting on the experiences of users is key to understanding the potential for sustainable real-world implementation of digitally enabled systems into clinical care. To support changes in clinical practice associated with new digitally driven models of cancer care, we need an increased understanding of patients' and clinicians' experiences of using such digital remote patient monitoring systems. We also need to consider how these systems can be implemented into different health care systems. Thus, in this study conducted across 5 European countries, we evaluated the use of Advanced Symptom Management System (ASyMS) within a large-scale randomized controlled trial (RCT) called Electronic Symptom Management System Remote Technology (eSMART) [13] and explored patients' and clinicians' experiences and perceptions of using ASyMS.

The Digital Intervention and Deployment Context: ASyMS and the eSMART Study

The Digital Intervention

ASyMS is a mobile phone–based, real-time, remote patient monitoring system used to assess and manage chemotherapy-related toxicities experienced by adult patients with cancer. It is a stand-alone, nurse-led, purpose-built, 24-hour anticipatory care system with integrated evidence-based clinical algorithms and alerts that provide rapid access to specialist staff in the patient's cancer hospital and timely initiation of appropriate interventions [13-15]. In eSMART, patients were provided with a mobile phone with the ASyMS app installed to complete an app-based electronic Daily Chemotherapy Toxicity Assessment Questionnaire (DCTAQ) [16]. The

XSL•FO
RenderX

DCTAQ assesses 10 common chemotherapy-related symptoms (nausea, vomiting, diarrhea, constipation, hand-foot syndrome, mucositis, paresthesia, flu-like symptoms, fatigue, and pain). An opportunity exists to enter a free-text report for up to 6 additional symptoms [13-16]. In eSMART, patients used ASyMS for a maximum of 6 cycles of chemotherapy [15].

The integrated clinical risk algorithms and embedded alerts within ASyMS are informed by the Common Terminology Criteria for Adverse Events (CTCAE V4.0) [17]. It uses a green, amber, and red traffic light system to indicate the severity, distress, and bother of symptoms reported, respectively. Alerts are generated in real time from the patient's device to a clinician's device at the local cancer center through a secure server. Clinicians access patient reports via a dedicated website. This approach facilitates time-bound clinician responses according to the alert level (amber alerts=8 hours and red alerts=30 minutes). Additional details on ASyMS and the RCT are described elsewhere [13-15].

All participants received training from researchers or clinicians at each clinical site to use ASyMS before commencing in the RCT. Manuals and videos were also provided for all participants for any ongoing training needs during the study.

Deployment Context

We conducted an RCT to evaluate the efficacy of ASyMS in 12 clinical sites in 5 European countries with public and private health care systems: Austria, Greece, Ireland, Norway, and the United Kingdom [13-15]. Patients with breast, colorectal, or hematological cancers were included in this RCT. Fuller details on patient inclusion criteria, exclusion criteria, and characteristics for the overall RCT are published elsewhere [15]. The findings from the RCT indicate that ASyMS use led to significant and sustained reductions in patients' symptom burden during chemotherapy as well as significant improvements in anxiety, quality of life, self-efficacy, and reductions in unmet supportive care needs [15].

Study Aims

The aim of the study is to explore patients' and clinicians' experiences of using ASyMS during chemotherapy and understand its impact on patients' experiences, clinical practice, and supportive care delivery.

Methods

Study Setting

Participants were recruited from cancer centers in the 5 partner countries: Austria, Greece, Ireland, Norway, and the United Kingdom. Patient interview participants were sampled purposively from the intervention population from the 3 diagnostic groups and were recruited either during their intervention use or within 1 year of completing the RCT. Clinicians were approached directly by members of the local country research teams and were sampled purposively to participate based on their knowledge, exposure, and use of ASyMS in the eSMART study.

Participant Recruitment

Patients randomized to the intervention were sampled purposively across the 3 cancer diagnoses, 5 countries, and at various times (during, at completion, or within 1 year of completion of chemotherapy). Patients were approached by telephone or face-to-face and invited to participate by members of the research team. Of the 415 patients assigned to the RCT intervention, 29 were interviewed. Across the 5 countries, clinicians who used ASyMS were approached by email or face-to-face by members of the research team and invited to participate in interviews. In total, 18 consented to participate. All interviews were conducted face-to-face or by telephone between September 2018 and January 2019.

Procedures

In total, 9 members of the in-country research team at the clinical site (7 female and 2 male members) with academic or nursing backgrounds and with native language skills conducted one-to-one patient and clinician interviews. Semistructured interviews were conducted in clinic and hospital environments using topic guides developed for the study (Multimedia Appendices 1 and 2). Open-ended questions were designed to explore various broad areas, including patients' and clinicians' experiences of using ASyMS, care organization and changes in practice following the introduction of ASyMS, perceived changes in care associated with the use of ASyMS, and its potential for future integration into routine chemotherapy care pathways.

Prior to the interviews, researchers participated in interview training sessions delivered via videoconference in small groups and received training resources before and after these sessions to standardize the interview processes. To ensure quality control of the data collection procedures, LM and LL listened to initial interviews conducted by each interviewer to provide reflective and constructive feedback prior to subsequent interviews. The length of patient and clinician interviews varied; the shortest were 6 and 5 minutes, respectively, and the longest were 28 and 57 minutes, respectively.

Individual face-to-face or telephone interviews were conducted over a 5-month period and were recorded on secure, password-protected, and encrypted devices. Anonymized interview files were stored in a secured cloud-based folder with tiered access managed by LL. All interviews were transcribed verbatim. Interviews conducted in Austrian, Greek, and Norwegian were translated and transcribed simultaneously and then checked by a member of the research team with native language skills. In the event of any errors or inaccuracies, corrections were made to transcribed text prior to the commencement of data analyses.

Ethical Considerations

The eSMART study was registered on ClinicalTrials.gov (NCT02356081) and was granted ethics approval from the National Health Service Lothian Southeast Scotland Research Ethics Committee 02 (14/SS/1062). The study received National Health Service Research and Development approvals and local clinical site ethics approvals in each partner country prior to data collection commencement. All participants provided written

XSL•FO
RenderX

informed consent prior to participating in an interview. All interview transcripts were anonymized prior to analysis; participants are referred to by an allocated ID number only. No compensation of any kind was provided to participants.

Analysis

Thematic analysis was used to analyze the data [18] and was conducted by a team of researchers (LM, LL, and OO). Our a priori focus was on exploring changes in practice arising from digitally enabled chemotherapy supportive care services. This meant a central orienting theme of "ASyMS as a facilitator of change" was somewhat deductively identified prior to open coding of data to ensure codes and themes identified in the data reflected perceptions, experiences, and meanings reported by participants [18].

A codebook was developed to ensure consistency in analytic approaches, given the team-based approach to analysis. Brief descriptions of each theme and subtheme were described in the codebook as a reference for each member of the analysis team. One team member (LM) assumed responsibility to update, revise, and maintain the codebook to ensure consistency and minimize the risk of any ambiguities in its evolution over time. Team meetings were held on a regular basis throughout the qualitative analyses to ensure consistency in the approach and discuss the emergent data to inform decisions on our reaching data saturation. In addition, intercoder reliability checks mitigated against potential inconsistencies and disparities in data analysis and interpretation across team members.

A coding comparison scheme compared the coding conducted by members of the analysis team. Of the 47 interviews conducted, 20% (n=9) of the transcripts were randomly selected for coding comparison to capture any variations. Transcripts were coded by 2 members of the team, as interviews were completed, translated, and transcribed. In addition, quality assurance within the random sample was ensured with the inclusion of interview transcripts conducted by different persons, different diagnoses, and across different stages of chemotherapy. As part of the intercoder reliability assessments, the team discussed problems with coding definitions and clarified these definitions to improve consistency. The coding framework is available (Multimedia Appendix 3). The qualitative software analysis software NVivo (QSR) was used throughout the analysis processes to organize and manage the data.

Results

Patient Characteristics

In total, 29 patients were interviewed across the 5 countries. Patient characteristics are shown in Table 1. Patients' mean age was 48.1 (SD 14.71; range 19-78) years; two-thirds of patients (n=19, 66%) were diagnosed with breast cancer, and most were female (n=22, 76%).

XSL•FO
RenderX

Table 1. Demographic and clinical characteristics of patients participating in an interview.

Characteristics	Patients (n=29)
Cancer diagnosis, n (%)	
Breast cancer	19 (66)
Colorectal cancer	3 (10)
Hematological cancer	7 (24)
Sex, n (%)	
Female	22 (76)
Male	7 (24)
Age (years)	
Mean (SD)	48.1 (14.71)
Range	19-78
Country, n (%) [a]	
Austria	9 (31)
Greece	7 (24)
Ireland	5 (17)
Norway	3 (10)
United Kingdom	5 (17)
Timing of participation in the interview	
During RCT [b]	
Patients, n (%)	9 (31)
Chemotherapy cycles, mean (SD)	4.6 (1.1)
Chemotherapy cycles, range	3-6
After RCT	
Patients, n (%)	20 (69)
Months since RCT completion, mean (SD)	3.9 (2.36)
Months since RCT completion, range	1-7

[a]Some sites were able to recruit more participants from >1 diagnostic group so recruited more patients.

[b]RCT: randomized controlled trial.

Clinician Participant Characteristics

In total, 18 clinicians were interviewed across all the 5 countries (Table 2). Most were female (n=13, 72%) and nurses (n=15, 83%) likely due to ASyMS being a nurse-led intervention.

XSL•FO
RenderX

Table 2. Demographic characteristics of clinicians participating in an interview.

Characteristics	Clinicians (n=18)
Sex, n (%)	
Female	13 (72)
Male	5 (28)
Age (years) (n=11)	
Mean (SD)	39.8 (12.62)
Range	27-60
Country, n (%)	
Austria	2 (11)
Greece	3 (17)
Ireland	4 (22)
Norway	3 (17)
United Kingdom	6 (33)
Roles, n (%)	
Nurses	15 (83)
Medical director	1 (6)
Clinical nurse manager	2 (11)

Central Orienting Theme: ASyMS as a Facilitator of Change

Overview

ASyMS was perceived as a facilitator of change across the different health care settings for either patients, clinicians, or both. Patients highlighted the ease of reporting symptoms, and clinicians reflected on the ease of responding to flagged symptoms, together creating targeted and timely care in what is usually a complex treatment process.

The combined reassurance, enhanced communications, knowledge, and improved cancer care experiences emerged from ASyMS' interactions between human and technology monitoring. The patterns identified in the data that increased our understanding of how and why participants saw ASyMS as a facilitator of change are summarized in Figure 1. Key quotations that support these themes are embedded within the text.

Figure 1. Central orienting theme and supporting themes identified from analysis of patient and clinician interviews. ASyMS: Advanced Symptom Management System.

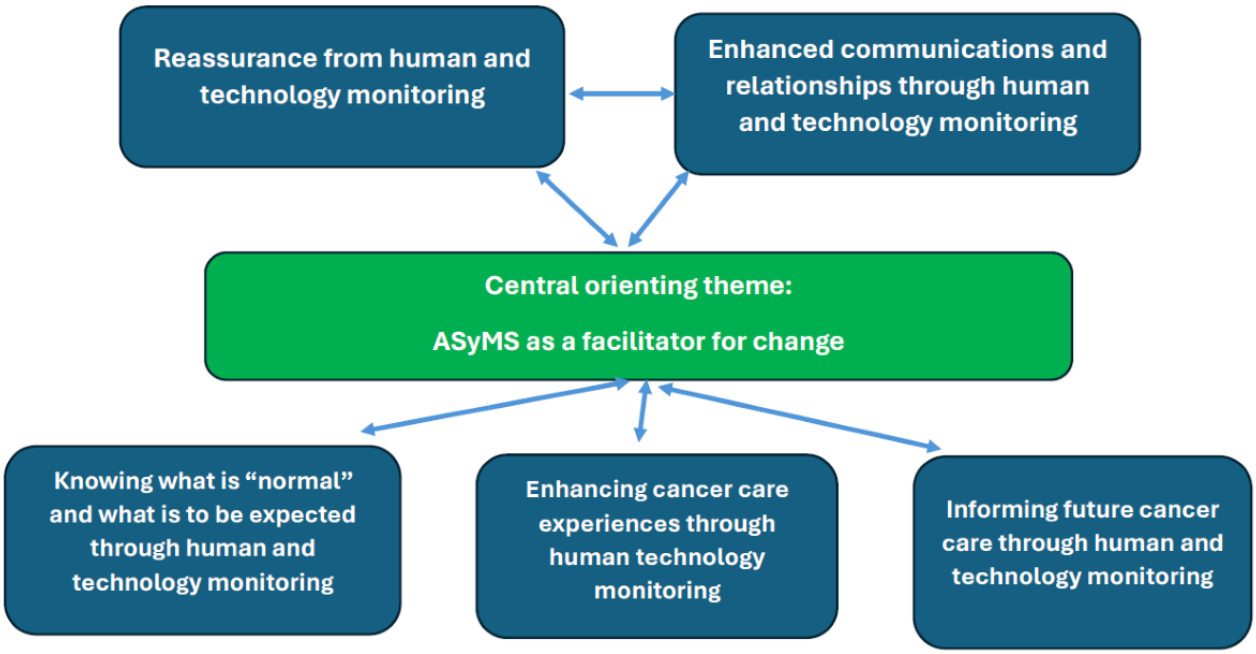

XSL•FO
RenderX

Theme 1: Reassurance From Human and Technology Monitoring

ASyMS was a mechanism that provided reassurance to patients during their chemotherapy treatments. Patients spoke positively about the human-centric proposition of ASyMS. They commented that while they were using technology in a new way to monitor symptoms, it was real humans, not machines, who reacted and responded to their clinical needs, and they found this reassuring:

> *What I think was incredibly good, is this real-life support, so that as soon as something [that] is outside of the norms and crosses certain values, that you then get real feedback and are called and can talk to someone, especially during the weekends.* [Austria, Patient ID 4]
>
> *It felt safe, I answered every day, and if anything at all showed up you called and we talked about it. That's a great reassurance.* [Norway, Patient ID 3]
>
> *And it was an incredible tool in a truly difficult journey and I felt that there was an invisible companion for me during this process and this was very nice.* [Greece, Patient ID 2]

Clinicians were cognizant of the greater reassurance that this approach afforded patients in the intervention group.

> *I think that the eSMART system meant that they were feeling a lot safer at home. Because it can be scary to go out of the hospital and you know that there are quite a lot of severe symptoms you can receive. And the patients always told us they were feeling a lot safer at home knowing that they had the device and could answer and then receive help...* [Norway, Clinician ID 2]
>
> *And in regard to their symptoms. For starters, the patient felt a form of security. That I have someone that I can talk to at any time, because they had their cell phone during the entire 24 hours. And they could contact, while they also had an immediate answer to their issue, it helped them a lot, in regard to their specific symptoms.* [Greece, Clinician ID 3]

However, for some patients, the daily PROM entries became too much as their treatment progressed, and their symptoms stabilized. The constant connection with the clinical setting was inadvertently fatiguing; and therefore, some patient perceptions of the intervention were more negative.

> *Patient ID 3 [Austria]: In the beginning I tried to do it daily. But at one point it got a bit too exhausting...Near the end I knew how to treat something. And it was exhausting to receive a call, despite not having any problems.*
>
> *Interviewer: Can you roughly estimate the point in time when it got exhausting?*
>
> *Patient ID 3: Yes, what was happening then...September, October, November...so after 3-4 months. Roughly. I didn't need any more help then.*

Theme 2: Enhanced Communications and Relationships Through Human and Technology Monitoring

Both patients and clinicians frequently referred to ASyMS' contribution to enhanced dyadic communication and relationships. The technology facilitated clinicians' knowledge of patients and therefore their provision of more personalized and individualized care. ASyMS was perceived as a relationship enabler by clinicians.

> *I think the most useful [thing] was getting to know the patients...So when you did go into the system, if by some chance you couldn't get hold of them there and then, you knew that the patient was, you know, clued up enough to ring you if there was something really wrong.* [United Kingdom, Clinician ID 3]
>
> *Interviewer: Do you do something now that you did not do before?*
>
> *Clinician ID 1 [Austria]: What I did not do before, regarding...I think that I am simply listening better, and I am trying to ask many questions also, listen and not only give suggestions and advice but really focus on the patient and what they really can do. Because we, as caregivers, tend a lot to give suggestions, suggest A, B, C or D and see what the patient does of that or not. Now I look why they might not be able to do something that I have suggested. So that I try to approach the patient in a more individual way and work towards a solution.*

Poignantly, from a patient's perspective, it was evident that the technology enhanced their relationship with their clinicians.

> *I felt that I had some people who cared about me.* [Greece, Patient ID 3]

Theme 3: Knowing What Is "Normal" and What Is to Be Expected Through Human and Technology Monitoring

A sense of uncertainty was apparent within many patients' narratives. Uncertainty about symptoms they expected to experience prior to commencing chemotherapy was coupled with uncertainty about what was to be considered "acceptable" or "normal" symptom experiences. For some patients, this uncertainty was tempered by ASyMS because the symptom algorithm "figures that out for you" (United Kingdom, Patient ID 1) and provided patients with clear indications when their symptoms were beyond what was expected:

> *Because it is really hard to know what you're supposed to be worried about and what you're not supposed to be worried about. So, you answer the questions and just kind of say it how it is, then yeah, it kind of figures that out for you. And, you know that if you put anything in then someone will call you...even...I know it did get irritating, but sometimes it is nice to just be able to talk to someone and go is this normal? Do people get this or is this something that needs looking into?* [United Kingdom, Patient ID 1]

XSL•FO
RenderX

> *I used it to aid my psyche, when you are not used to being in pain it was helpful to check myself. You get worried, but it gave me information that this might be natural. Getting tired, headaches, feeling unwell. It is a natural part of the treatment. It gave me answers, and I took that in account in my daily life. I did not have to worry about it.* [Norway, Patient ID 1]

However, for patients who were experiencing symptoms prior to their chemotherapy ASyMS were not sophisticated enough to record these experiences and integrate them into a more personalized alerting algorithm. For some patients, this impacted their interactions with the daily PROMs, and this may be seen to be a limitation of ASyMS.

> *There was no distinction between symptoms that you experience for the first time and others that you experience as part of your condition. Nowhere did it ask whether you had the symptoms before the treatments or not. But I asked both the doctor and nurse about it. I told them "I had the symptoms from before, so should I write them down?" and they told me "No, do not include them because they continue on without a change."* [Greece, Patient ID 3]

Theme 4: Enhancing Cancer Care Experiences Through Human and Technology Monitoring

Patients incorporated ASyMS into their daily lives and routines so that it became an integral source of support; many found ways to habituate ASyMS into their routines. It became a companion to them and served as an important facilitator to support their psychological well-being.

> *It just kind of just became part of my morning routine. I had to get up, have my medication, sit with a glass of water, and do my questionnaire. It didn't take up a huge amount of time.* [United Kingdom, Patient ID 4]

> *And oh my god, for me, when you start to use it, it became part of your life. But, for me it was the safety and that makes me feel so, so good.* [Ireland, Patient ID 1599]

The enhanced supportive care ASyMS facilitated meant that there was a shift from traditional reactive engagement with clinicians to anticipatory modes of communication. As such, patients across all countries recognized the importance of the alerting algorithms to expedite their care pathways when symptomatic, particularly for out-of-hours care.

> *Now in fairness, the alert system on it is excellent. I would say I had not put it in five minutes and the phone rang...and it was the hospital and, now it was not that I was waiting for them to ring. But they were that quick, because I would have rang them myself...within the hour but they were back within five minutes...* [Ireland, Patient ID 1]

> *I think that this support is absolutely important especially at times when the chemo ward is closed. They do close on a Friday at 4 o'clock and open again on Monday in the morning. There is nothing between that, whom should I call?* [Austria, Patient ID 4]

Theme 5: Informing Future Cancer Care Through Human and Technology Monitoring

A common belief was held by both patients and clinicians that ASyMS has a role in future cancer care delivery with some adjustments and advancements. Some patients suggested that ASyMS is adapted for use beyond acute treatment and is also used for the posttreatment period, particularly for psychological support.

> *I would like to have it here [follow-up phase] now, too. If not daily, with a different protocol. For example, weekly or when needed. Because the person feels a lot and wishes to know...And sometimes I feel that I am in a worse mood now, when compared how I was feeling during chemotherapy. I do not know if such a boost is available, if there could be psychological support. I would like being able to tell someone how I feel.* [Greece, Patient ID 5]

However, some patients were frustrated by the static and generic nature of the risk algorithms and so spoke about ways ASyMS could be smarter going forward.

> *Looking in retrospect, I think that it would have been clever to enter some symptoms so that the statistics is right. I would maybe add that in the programming and make it more sensitive, so that it says that even if you had the symptoms yesterday or last week or have permanently entered them, that one should enter them until. So that the questionnaire could be more explicit about that, and it could be added in there.* [Austria, Patient ID 4]

Some frustrations with ASyMS were reported by clinicians who believed ASyMS duplicated elements of their current care provision, and others believed specific staffing would need to be in place to facilitate timely responses to ASyMS alerts in the future.

> *No, it hasn't changed our [nurse] roles, I can see where it may change roles in centres that don't have the same structure, we have here within the acute oncology service handling the 24/7 telephone triage. But, for us it hasn't because we already have a nurse assigned to that and so it is basically just attaching the eSMART role into that existing role.* [United Kingdom, Clinician ID 4]

> *I think ASyMS is very important, and I think it worked. It is more on the receiving end to have the appropriate services in place, to have the staff that is going to be responsible for responding to ASyMS. So yes, something like an ASyMS nurse coordinator, or even maybe for the liaison nurses for the different patients, for them to be the alert handlers.* [Ireland, Clinician ID 3]

XSL•FO
RenderX

Discussion

Overview

This study is the first to evaluate, from both patients' and clinicians' perspectives, the deployment of a large-scale, international, simultaneous, multicenter, digital remote monitoring intervention in cancer care. Through semistructured interviews and robust qualitative analyses, we identified the positive impact and changes this remote monitoring technology had on patients' and clinicians' experiences and clinical practice and areas for improvement for future routine implementations.

Principal Findings

The qualitative data from patients and clinicians in the eSMART study revealed that both user groups found value in remotely monitoring chemotherapy-related symptoms in real time using the ASyMS intervention. Capturing user experiences across multiple countries and health care systems over a sustained period provides useful insights into the need for, and value of, real-time remote symptom monitoring systems and digitally enabled cancer care models. Patients' and clinicians' experiences with ASyMS were largely positive. Most reported value in its purpose and function.

In all the clinical sites, ASyMS was the first implementation of a digitally enabled, patient-driven, nurse-led, remote monitoring model of care. Patients found personal benefits in remotely monitoring their symptoms in real time in their own environments. However, it was the combination of technology and human monitoring that most participants found to be one of the greatest assets of ASyMS. This finding is consistent with the work by Leonardsen et al [19] who interviewed patients with cancer after the rapid implementation of home-based remote monitoring services during the COVID-19 pandemic. In their study, remote monitoring included the use of mobile phones, videoconferences, and daily questionnaire feedback using a tablet or computer. While engagement was positive, it was noted that technologies were used alongside appropriate person-person contact [19]. In a similar way, we demonstrated the value of clinical risk algorithms driving appropriate person-to-person contact through our traffic light triaging system. By alerting clinicians directly, we ensured patients received immediate, effective, tailored, and human symptom management advice.

Compared to previous systems [20-28], the daily reporting, real-time alerting, and integrated feedback components of ASyMS are unique to our remote monitoring system. For example, in a multiclinic RCT evaluation conducted in the United States that evaluated electronic symptom monitoring with patients with cancer [20], patients randomized to the intervention completed patient-reported outcomes once a week through email or automated call prompts. The web-based app eRAPID (Electronic Patient Self-Reporting of Adverse-Events: Patient Information and Advice) used weekly rather than daily symptom reports [21]. The symptom management system [22] filters symptom alerts to members of the research team that can be forwarded to clinicians to respond to within 1-3 days. This delayed response is problematic because of the need for timely clinical interventions. In contrast, our study findings support previous systematic review evidence, which highlights the positive role of daily PROMs in cancer care for patients [23,24]. Such benefits are amplified when the daily PROMs are digitized [23] and when the PROMs facilitate feedback to patients and clinicians [24]. We observed this too in our own study: we mandated response times for amber and red alerts within 8 hours and 30 minutes, respectively, for our digital PROMs, and our qualitative data demonstrate that these parameters and feedback mechanisms were largely acceptable to both patients and clinicians. We do acknowledge, however, that a small number of patients experienced daily PROM completion fatigue and so this may impact on longer-term sustainable adoption. The findings also revealed that some patients and clinicians were frustrated by the static nature of the alerting algorithms. In the future, we plan to use machine learning capabilities to make the algorithms more dynamic, data-responsive, and personalized to adapt to patients' needs and symptom experiences.

ASyMS was perceived by both patients and clinicians to offer reassurance during chemotherapy largely because of the combined human and technology monitoring afforded by ASyMS. Patients spoke about feeling more knowledgeable about and in control of their symptoms. Clinicians valued real-time monitoring because it did not rely on patients' retrospective recall. These findings align with previous work that reported positive experiences from both patients and clinicians with the use of real-time remote monitoring systems [20,25-28].

As health care moves toward digitally enabled and digitally driven services, lessons from the successful implementation of remote monitoring systems like ASyMS have increasing importance. In our study, we successfully scaled up and deployed a digital remote monitoring intervention in 5 different countries with over 400 randomized patients. Our interviews with a subsample of patients and clinicians provide us with some important insights for others invested in the challenge of supporting digitally enabled health and care models of the future.

Our system was designed to be accessible to patients and easily integrated into their daily lives to support its use and sustained adoption. Patients' narratives and reflections highlighted ways in which they had "cognitively habituated" [29] ASyMS into their daily lives and routines. Such habits included keeping the ASyMS device in the same place as a reminder to complete the DCTAQ, identifying a consistent set time each day to complete it, completing it at the same time as taking medications, and setting an alert reminder on another device. These findings support the hypothesis that users of new technologies make changes to their lives to accommodate new digital services in their homes [29]. Therefore, it is important to allow users some flexibility to find their own approaches to "domesticate" the technology to support engagement and sustained adoption.

Shifting health care contexts, particularly following the COVID-19 pandemic, mean that the roles of clinicians are evolving, especially as technology and digital solutions are increasingly integrated into everyday clinical practice [30]. In a recent scoping review [30], the core roles that nurses already have, and will continue to have, in leading digital transformation practices within clinical settings were noted, given their

XSL•FO
RenderX

centralized communication role between patients and other clinicians. However, to ensure health care professionals can evolve their competencies within transformative health care contexts, continued professional development opportunities must evolve to allow their effective delivery of meaningful, personalized, and person-centered care, positively enhanced by digital interventions [31]. Our work supports this notion, as we observed noticeable changes in processes of care and roles and responsibilities of clinical staff in countries and clinical settings specifically where a nurse-led, digital health intervention was introduced for the first time.

However, our findings suggest that they would be receptive to more continuous use of remote monitoring technologies in treatment follow-up, given their unmet needs beyond diagnosis and treatment completion [32-34]. Indeed, in our study, some patients identified the period after treatment completion to be the most psychologically challenging because access to care ceases. Thus, a role exists for future digital remote monitoring systems to evolve to support patients during and after cancer treatments. ASyMS is not yet available to monitor long-term cancer experience, but it is a future research priority and concurs with previous identification of patients' needs for continued use of a weekly symptom reporting system beyond the treatment period [35].

The landscape of health and care service provision is rapidly evolving. An impetus exists to use digital health technologies to influence the ways in which people engage with and receive care. The COVID-19 pandemic necessitated rapid changes in the management of patients with cancer including the increased use of remote patient monitoring technologies to monitor, manage, and engage with patients [36]. However, our work and that of others [37,38] demonstrate that for patients with cancer, a need exists to balance technological remote monitoring interventions and human-human contact to support adoption. Indeed, users of remote monitoring technologies and solutions that were rapidly implemented during the COVID-19 pandemic reported positive perceptions of these technologies, if used in combination with direct person-to-person interactions and support [38]. Therefore, with the unique insights into the experiences of patients and clinicians from 5 very different countries and health care systems that our data afford us, we are confident that technology adoption and acceptability in various health care contexts are most positive when the technology facilitates, rather than replaces, human-human relationships and communications.

Strengths and Limitations

We interviewed a relatively large sample of patients and clinicians for the qualitative study. Because we interviewed participants in their native language, we avoided a sampling bias of only purposively selecting people who could be interviewed in English. Although successful, this strategy was reliant upon the research teams in each country conducting these interviews. The variability inherent in this approach may have affected the quality and depth of some interviews. However, we mitigated this risk as much as possible by providing thorough preinterview training to personnel, ongoing interview support, and reflective and critical feedback during interviews as part of our quality control measures. We adopted a robust approach to translation and transcription of all interviews to ensure a clear and consistent approach for our data analyses. We are also aware that the patients in this study were predominantly younger age females diagnosed with breast cancer so this may limit the generalizability of our results.

We also acknowledge that our decision to only interview patients in the intervention arm about their symptom management experiences may provide an inflated positive view of the role of daily PROMs and of ASyMS. We did not interview patients in the control group so we do not have comparative narratives, and we did not interview intervention group patients who withdrew from the study. These decisions may mean there is some consequent positivity bias in our findings. However, our large overall qualitative sample from different countries but with similar reported experiences means the validity of our results stands. In addition, attrition in the whole intervention arm in the main RCT was low (34/415, 8.2%, and overall adherence to the intervention was high at 76.9%) [15], providing further support for the largely positive participation experience reported by patients.

Implications and Conclusions

Successful implementation of digital health interventions can be challenging. We have demonstrated through our qualitative data that patients and clinicians can positively adopt and integrate remote monitoring systems when available to them. Our data demonstrated that ASyMS was, on balance, perceived positively by both patients and clinicians across the eSMART partner countries. Overall, our technology enhanced communications and person-to-person contact between patients and clinicians. The timely symptom management provided by clinicians because of the real-time reporting meant that patients were reassured; gained knowledge about symptoms they were experiencing; and received the right care, at the right time, and in the right place. Going forward, the focus should be on the routine and sustained implementation of scalable, accessible, personalized, and usable digitally enabled cancer care services to help deliver optimal models of cancer care.

Acknowledgments

This study was funded by the European Commission via the FP7-HEALTH-2013-INNOVATION-1 grant award scheme (grant 602289). The trial was sponsored by the University of Strathclyde. The views expressed in this paper are those of the authors and not necessarily those of the European Commission or the University of Strathclyde. The funders and the sponsors of the trial had no role in the design, undertaking, and reporting of the study. The authors thank the patients and clinicians across the partner hospital sites for taking part and positively contributing to the overall electronic symptom management system remote technology

XSL•FO
RenderX

program of work (Medical University Vienna Comprehensive Cancer Centre, Austria; Agioi Anargiri Cancer Hospital, Greece; Metropolitan Hospital, Greece; Air Force General Hospital, Greece; St James' Hospital, Ireland; St Vincent's Private Hospital, Ireland; St Vincent's University Hospital, Ireland; Innlandet Hospital Trust Division Gjøvik-Lillehammer, Norway; Guy's and St Thomas' National Health Service Foundation Trust, United Kingdom; Mount Vernon Cancer Centre, United Kingdom; St George's Healthcare Trust, United Kingdom; and University College London Hospitals, United Kingdom). Generative artificial intelligence tools were not used in any portions of the manuscript writing.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RM, LM, GK, NK, ER, JA, PF, EP, EPF, AG, PM, GVB, CM, DO, AF, SK, and PD obtained funding for this research. All authors contributed to either the design or conduct of the overall program of work. LM, LL, and OO conducted data analyses. LM produced the first draft of this manuscript, and all authors reviewed, edited, and approved the final version. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria were omitted.

Conflicts of Interest

AF is the managing director of Docobo, the health technology software development company that built the advanced symptom management system platform for the electronic symptom management system remote technology study. Docobo was a named partner in the program of work and received funding (grant 602289) for this development work.

Multimedia Appendix 1
Patients' interviews topic guide.
[DOCX File , 28 KB - jmir_v26i1e53834_app1.docx]

Multimedia Appendix 2
Clinicians' interviews topic guide.
[DOCX File , 34 KB - jmir_v26i1e53834_app2.docx]

Multimedia Appendix 3
Coding framework.
[DOCX File , 28 KB - jmir_v26i1e53834_app3.docx]

References

1. Brianna, Lee SH. Chemotherapy: how to reduce its adverse effects while maintaining the potency? Med Oncol 2023;40(3):88. [doi: 10.1007/s12032-023-01954-6] [Medline: 36735206]
2. Coolbrandt A, Wildiers H, Laenen A, Aertgeerts B, Dierckx de Casterlé B, van Achterberg T, et al. A nursing intervention for reducing symptom burden during chemotherapy? Oncol Nurs Forum 2018;45(1):115-128. [doi: 10.1188/18.ONF.115-128] [Medline: 29251287]
3. McGregor BA, Vidal GA, Shah SA, Mitchell JD, Hendifar AE. Remote oncology care: review of current technology and future directions. Cureus 2020;12(8):e10156 [FREE Full text] [doi: 10.7759/cureus.10156] [Medline: 33014652]
4. Palos GR, Suarez-Almazor ME. Launching an electronic patient-reported outcomes initiative in real-time clinical practice. J Natl Cancer Inst Monogr 2021;2021(57):23-30. [doi: 10.1093/jncimonographs/lgab005] [Medline: 34478509]
5. Girgis A, Bamgboje-Ayodele A, Rincones O, Vinod SK, Avery S, Descallar J, et al. Stepping into the real world: a mixed-methods evaluation of the implementation of electronic patient reported outcomes in routine lung cancer care. J Patient Rep Outcomes 2022;6(1):70 [FREE Full text] [doi: 10.1186/s41687-022-00475-6] [Medline: 35723827]
6. Milch V, Wang R, Der Vartanian C, Austen M, Hector D, Anderiesz C, et al. Cancer Australia consensus statement on COVID-19 and cancer care: embedding high value changes in practice. Med J Aust 2021;215(10):479-484 [FREE Full text] [doi: 10.5694/mja2.51304] [Medline: 34689343]
7. Mir O, Ferrua M, Fourcade A, Mathivon D, Duflot-Boukobza A, Dumont S, et al. Digital remote monitoring plus usual care versus usual care in patients treated with oral anticancer agents: the randomized phase 3 CAPRI trial. Nat Med 2022;28(6):1224-1231. [doi: 10.1038/s41591-022-01788-1] [Medline: 35469070]
8. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
9. Bedard PL, Hyman DM, Davids MS, Siu LL. Small molecules, big impact: 20 years of targeted therapy in oncology. Lancet 2020;395(10229):1078-1088. [doi: 10.1016/S0140-6736(20)30164-1] [Medline: 32222192]

XSL•FO
RenderX

10. Challinor JM, Alqudimat MR, Teixeira TOA, Oldenmenger WH. Oncology nursing workforce: challenges, solutions, and future strategies. Lancet Oncol 2020;21(12):e564-e574. [doi: 10.1016/S1470-2045(20)30605-7] [Medline: 33212044]
11. Rittberg R, Mann A, Desautels D, Earle CC, Navaratnam S, Pitz M. Canadian cancer centre response to COVID-19 pandemic: a national and provincial response. Curr Oncol 2020;28(1):233-251 [FREE Full text] [doi: 10.3390/curroncol28010026] [Medline: 33704191]
12. Barsom EZ, Jansen M, Tanis PJ, van de Ven AWH, Blussé van Oud-Alblas M, Buskens CJ, et al. Video consultation during follow up care: effect on quality of care and patient- and provider attitude in patients with colorectal cancer. Surg Endosc 2021;35(3):1278-1287 [FREE Full text] [doi: 10.1007/s00464-020-07499-3] [Medline: 32198552]
13. Maguire R, Fox PA, McCann L, Miaskowski C, Kotronoulas G, Miller M, et al. The eSMART study protocol: a randomised controlled trial to evaluate electronic symptom management using the advanced symptom management system (ASyMS) remote technology for patients with cancer. BMJ Open 2017;7(5):e015016 [FREE Full text] [doi: 10.1136/bmjopen-2016-015016] [Medline: 28592577]
14. Furlong E, Darley A, Fox P, Buick A, Kotronoulas G, Miller M, et al. Adaptation and implementation of a multinational eHealth intervention for people with cancer: reflections from the field. J Med Int Res 2018:12. [doi: 10.2196/preprints.10813]
15. Maguire R, McCann L, Kotronoulas G, Kearney N, Ream E, Armes J, et al. Real time remote symptom monitoring during chemotherapy for cancer: European multicentre randomised controlled trial (eSMART). BMJ 2021;374:n1647 [FREE Full text] [doi: 10.1136/bmj.n1647] [Medline: 34289996]
16. Maguire R, Kotronoulas G, Donnan PT, Paterson C, McCann L, Connaghan J, et al. Development and preliminary testing of a brief clinical tool to enable daily monitoring of chemotherapy toxicity: the Daily Chemotherapy Toxicity Self-Assessment Questionnaire. Eur J Cancer Care (Engl) 2018;27(6):e12890 [FREE Full text] [doi: 10.1111/ecc.12890] [Medline: 29993150]
17. Common Terminology Criteria for Adverse Events (CTCAE) 4.0. National Institutes of Health. URL: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm#ctc_40 [accessed 2023-10-17]
18. Braun V, Clarke. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications Ltd; Oct 27, 2021.
19. Leonardsen ACL, Helgesen AK, Stensvold A, Magnussen J, Grøndahl VA. Cancer patients' perspectives on remote monitoring at home during the COVID-19 pandemic—a qualitative study in Norway. BMC Health Serv Res 2022;22(1):453 [FREE Full text] [doi: 10.1186/s12913-022-07897-4] [Medline: 35387645]
20. Basch E, Schrag D, Henson S, Jansen J, Ginos B, Stover AM, et al. Effect of electronic symptom monitoring on patient-reported outcomes among patients with metastatic cancer: a randomized clinical trial. JAMA 2022;327(24):2413-2422 [FREE Full text] [doi: 10.1001/jama.2022.9265] [Medline: 35661856]
21. Absolom K, Holch P, Warrington L, Samy F, Hulme C, Hewison J, et al. Electronic Patient Self-Reporting of Adverse-Events: Patient Information and Advice (eRAPID): a randomised controlled trial in systemic cancer treatment. BMC Cancer 2017;17(1):318 [FREE Full text] [doi: 10.1186/s12885-017-3303-8] [Medline: 28482877]
22. Whitehead L, Emery L, Kirk D, Twigg D, Brown D, Dewar J. Evaluation of a remote symptom assessment and management (SAM) system for people receiving adjuvant chemotherapy for breast or colorectal Cancer: mixed methods study. JMIR Cancer 2020;6(2):e22825 [FREE Full text] [doi: 10.2196/22825] [Medline: 33284122]
23. van Egdom LSE, Oemrawsingh A, Verweij LM, Lingsma HF, Koppert LB, Verhoef C, et al. Implementing patient-reported outcome measures in clinical breast cancer care: a systematic review. Value Health 2019;22(10):1197-1226 [FREE Full text] [doi: 10.1016/j.jval.2019.04.1927] [Medline: 31563263]
24. Graupner C, Kimman ML, Mul S, Slok AHM, Claessens D, Kleijnen J, et al. Patient outcomes, patient experiences and process indicators associated with the routine use of patient-reported outcome measures (PROMs) in cancer care: a systematic review. Support Care Cancer 2021;29(2):573-593 [FREE Full text] [doi: 10.1007/s00520-020-05695-4] [Medline: 32875373]
25. Maguire R, McCann L, Miller M, Kearney N. Nurse's perceptions and experiences of using of a mobile-phone-based advanced symptom management system (ASyMS) to monitor and manage chemotherapy-related toxicity. Eur J Oncol Nurs 2008;12(4):380-386. [doi: 10.1016/j.ejon.2008.04.007] [Medline: 18539527]
26. McCann L, Maguire R, Miller M, Kearney N. Patients' perceptions and experiences of using a mobile phone-based advanced symptom management system (ASyMS) to monitor and manage chemotherapy related toxicity. Eur J Cancer Care (Engl) 2009;18(2):156-164. [doi: 10.1111/j.1365-2354.2008.00938.x] [Medline: 19267731]
27. Velikova G, Absolom K, Hewison J, Holch P, Warrington L, Avery K, et al. Electronic self-reporting of adverse events for patients undergoing cancer treatment: the eRAPID research programme including two RCTs. Programme Grants Appl Res 2022;10(1):1-144. [doi: 10.3310/FDDE8516] [Medline: 35138783]
28. Basch E, Leahy AB, Dueck AC. Benefits of digital symptom monitoring with patient-reported outcomes during adjuvant Cancer treatment. J Clin Oncol 2021;39(7):701-703. [doi: 10.1200/JCO.20.03375] [Medline: 33507823]
29. Søraa RA, Nyvoll P, Tøndel G, Fosch-Villaronga E, Serrano JA. The social dimension of domesticating technology: interactions between older adults, caregivers, and robots in the home. Technol Forecast Soc Change 2021;167:120678 [FREE Full text] [doi: 10.1016/j.techfore.2021.120678]
30. Isidori V, Diamanti F, Gios L, Malfatti G, Perini F, Nicolini A, et al. Digital technologies and the role of health care professionals: scoping review exploring nurses' skills in the digital era and in the light of the COVID-19 pandemic. JMIR Nurs 2022;5(1):e37631 [FREE Full text] [doi: 10.2196/37631] [Medline: 36194466]

XSL•FO
RenderX

31. Jarva E, Oikarinen A, Andersson J, Tuomikoski AM, Kääriäinen M, Meriläinen M, et al. Healthcare professionals' perceptions of digital health competence: a qualitative descriptive study. Nurs Open 2022;9(2):1379-1393 [FREE Full text] [doi: 10.1002/nop2.1184] [Medline: 35094493]
32. Parry C, Lomax JB, Morningstar EA, Fairclough DL. Identification and correlates of unmet service needs in adult leukemia and lymphoma survivors after treatment. J Oncol Pract 2012;8(5):e135-e141 [FREE Full text] [doi: 10.1200/JOP.2011.000464] [Medline: 23277776]
33. Vu JVT, Matusko N, Hendren S, Regenbogen SE, Hardiman KM. Patient-reported unmet needs in colorectal cancer survivors after treatment for curative intent. Dis Colon Rectum 2019;62(7):815-822 [FREE Full text] [doi: 10.1097/DCR.0000000000001326] [Medline: 31188182]
34. Miroševič Š, Prins J, Borštnar S, Besić N, Homar V, Selič-Zupančič P, et al. Factors associated with a high level of unmet needs and their prevalence in the breast cancer survivors 1-5 years after post local treatment and (neo)adjuvant chemotherapy during the COVID-19: a cross-sectional study. Front Psychol 2022;13:969918 [FREE Full text] [doi: 10.3389/fpsyg.2022.969918] [Medline: 36262447]
35. Andikyan V, Rezk Y, Einstein MH, Gualtiere G, Leitao MM, Sonoda Y, et al. A prospective study of the feasibility and acceptability of a web-based, electronic patient-reported outcome system in assessing patient recovery after major gynecologic cancer surgery. Gynecol Oncol 2012;127(2):273-277 [FREE Full text] [doi: 10.1016/j.ygyno.2012.07.124] [Medline: 22871467]
36. ESMO management and treatment adapted recommendations in the COVID-19 era: breast cancer. European Society for Medical Oncology. 2020. URL: https://www.esmo.org/guidelines/cancer-patient-management-during-the-covid-19-pandemic/breast-cancer-in-the-covid-19-era [accessed 2023-10-17]
37. Daly B, Lauria TS, Holland JC, Garcia J, Majeed J, Walters CB, et al. Oncology patients' perspectives on remote patient monitoring for COVID-19. JCO Oncol Pract 2021;17(9):e1278-e1285 [FREE Full text] [doi: 10.1200/OP.21.00269] [Medline: 34085536]
38. Grøndahl VA, Helgesen AK, Holm E, Magnussen J, Leonardsen A. Remote monitoring of cancer patients during the Covid-19 pandemic—an interview study of nurses' and physicians' experiences. BMC Nurs 2022;21(1):169 [FREE Full text] [doi: 10.1186/s12912-022-00953-8] [Medline: 35765023]

Abbreviations

ASyMS: Advanced Symptom Management System
DCTAQ: Daily Chemotherapy Toxicity Assessment Questionnaire
eRAPID: Electronic Patient Self-Reporting of Adverse-Events: Patient Information and Advice
eSMART: Electronic Symptom Management System Remote Technology
PROM: patient-reported outcome measure
RCT: randomized controlled trial

Edited by A Mavragani; submitted 25.10.23; peer-reviewed by MT Crafoord, N Fridriksdottir, A Stover; comments to author 12.01.24; revised version received 02.05.24; accepted 12.09.24; published 03.12.24.

Please cite as:
McCann L, Lewis L, Oduntan O, Harris J, Darley A, Berg GV, Lubowitzki S, Cheevers K, Miller M, Armes J, Ream E, Fox P, Furlong EP, Gaiger A, Kotronoulas G, Patiraki E, Katsaragakis S, McCrone P, Miaskowski C, Cardone A, Orr D, Flowerday A, Skene S, Moore M, De Souza N, Donnan P, Maguire R
Patients' and Clinicians' Experiences Using a Real-Time Remote Monitoring System for Chemotherapy Symptom Management (ASyMS): Qualitative Study
J Med Internet Res 2024;26:e53834
URL: https://www.jmir.org/2024/1/e53834
doi:10.2196/53834
PMID:

©Lisa McCann, Liane Lewis, Olubukola Oduntan, Jenny Harris, Andrew Darley, Geir V Berg, Simone Lubowitzki, Katy Cheevers, Morven Miller, Jo Armes, Emma Ream, Patricia Fox, Eileen Patricia Furlong, Alexander Gaiger, Grigorios Kotronoulas, Elisabeth Patiraki, Stylianos Katsaragakis, Paul McCrone, Christine Miaskowski, Antonella Cardone, Dawn Orr, Adrian Flowerday, Simon Skene, Margaret Moore, Nicosha De Souza, Peter Donnan, Roma Maguire. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers

Anna Jolliff[1], MS, LMHC; Richard J Holden[1], PhD; Rupa Valdez[2,3], PhD; Ryan J Coller[4], MPH, MD; Himalaya Patel[1,5], PhD; Matthew Zuraw[6], MBA; Anna Linden[7], PhD; Aaron Ganci[8], PhD; Christian Elliott[6]; Nicole E Werner[1], PhD

[1]Department of Health & Wellness Design, School of Public Health - Bloomington, Indiana University Bloomington, Bloomington, IN, United States

[2]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, United States

[3]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, United States

[4]Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin - Madison, Madison, WI, United States

[5]Health Systems Research Center for Health Information and Communication, Richard L Roudebush Veterans Affairs Medical Center, United States Department of Veterans Affairs, Indianapolis, IN, United States

[6]CareVirtue Technologies, San Diego, CA, United States

[7]Department of Industrial & Systems Engineering, University of Wisconsin - Madison, Madison, WI, United States

[8]Department of Visual Communication Design, Herron School of Art & Design, Indiana University Indianapolis, Indianapolis, IN, United States

Corresponding Author:
Anna Jolliff, MS, LMHC
Department of Health & Wellness Design
School of Public Health - Bloomington
Indiana University Bloomington
1025 E. 7th Street
Bloomington, IN, 47405
United States
Phone: 1 812 855 1561
Email: annjoll@iu.edu

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e69714

Abstract

Background: Digital health interventions are a promising method for delivering timely support to underresourced family caregivers. The uptake of digital health interventions among caregivers may be improved by engaging caregivers in participatory design (PD). In recent years, there has been a shift toward conducting PD remotely, which may enable participation by previously hard-to-reach groups. However, little is known regarding how best to facilitate engagement in remote PD among family caregivers.

Objective: This study aims to (1) understand the context, quality, and outcomes of family caregivers' engagement experiences in remote PD and (2) learn which aspects of the observed PD approach facilitated engagement or need to be improved.

Methods: We analyzed qualitative and quantitative data from evaluation and reflection surveys and interviews completed by research and community partners (family caregivers) across 4 remote PD studies. Studies focused on building digital health interventions for family caregivers. For each study, community partners met with research partners for 4 to 5 design sessions across 6 months. After each session, partners completed an evaluation survey. In 1 of the 4 studies, research and community partners completed a reflection survey and interview. Descriptive statistics were used to summarize quantitative evaluation and reflection survey data, while reflexive thematic analysis was used to understand qualitative data.

Results: In 62.9% (83/132) of evaluations across projects 1-3, participants described the session as "very effective." In 74% (28/38) of evaluations for project 4, participants described feeling "extremely satisfied" with the session. Qualitative data relating to the engagement context identified that the identities of partners, the technological context of remote PD, and partners'

XSL•FO
RenderX

understanding of the project and their role all influenced engagement. Within the domain of engagement quality, relationship-building and co-learning; satisfaction with prework, design activities, time allotted, and the final prototype; and inclusivity and the distribution of influence contributed to partners' experience of engagement. Outcomes of engagement included partners feeling an ongoing interest in the project after its conclusion, gratitude for participation, and a sense of meaning and self-esteem.

Conclusions: These results indicate high satisfaction with remote PD processes and few losses specific to remote PD. The results also demonstrate specific ways in which processes can be changed to improve partner engagement and outcomes. Community partners should be involved from study inception in defining the problem to be solved, the approach used, and their roles within the project. Throughout the design process, online tools may be used to check partners' satisfaction with design processes and perceptions of inclusivity and power-sharing. Emphasis should be placed on increasing the psychosocial benefits of engagement (eg, sense of community and purpose) and increasing opportunities to participate in disseminating findings and in future studies.

(J Med Internet Res 2024;26:e60353) doi:10.2196/60353

KEYWORDS

user-centered design; family caregivers; mobile health; digital health; web-based intervention; stakeholder engagement; patient engagement; community-based participatory action research; community participation; qualitative evaluation

Introduction

Digital Health Interventions for Family Caregivers

The United States is home to 44 million family caregivers of adults with chronic illness, and 11 million of these family caregivers care for someone with Alzheimer disease and related dementias (ADRD) [1,2]. A total of 1% of children in the United States are medically complex, meaning they have multiple chronic illnesses, have functional limitations, and depend on medical technologies for survival [3]. Caregivers of these children and people living with ADRD are often undersupported and underresourced in their role of providing care at home and in their communities, which can lead to caregiver burnout and increased risk for chronic health disorders [4-7]. Digital health interventions that deliver support and education for caregivers are a promising area of research. Digital interventions, by virtue of their remote nature and powerful functionality, have the potential to reduce costs and time associated with traveling to community resources, quickly synthesize resource availability, provide training to caregivers in rural and underresourced areas, and enable connection between isolated caregivers [8]. However, among caregivers of adults, uptake has been hampered when interventions are inaccessible, unhelpful, or hard to use by their target population [8]. Among parent caregivers, ease of use, customizability, and cultural appropriateness have been key to driving uptake [9]. There is a persistent need to ensure that the design of digital health interventions is tailored to these key target populations. To this end, this study aimed to understand how to best facilitate engagement in remote participatory design (PD) so that high-risk populations can be involved in designing the digital health interventions intended to meet their needs.

PD to Meet Caregiver Needs

The uptake of digital health interventions among caregivers may be improved by engaging caregivers in the design of these interventions. PD is an established strategy for engaging community and academic partners in intervention design [10]. PD is a design process in which academic researchers and software developers partner with those who perform a certain type of work to design a tool, technology, or workflow to support that work [10]. Family caregiving can be understood as work, insofar as multiple people coordinate to perform complex tasks (eg, medication management and symptom tracking) using specialized tools in a particular context [11]. By performing their work, caregivers accrue implicit and tacit knowledge, or work-related knowledge that is highly practical, mostly invisible, and that can be difficult to articulate [12,13]. Because caregivers may not themselves have the skills required to design health intervention technologies, the partnership between caregivers and experts in software development, human-computer interaction, and systems design is key to designing digital interventions that are useful and effective [10,14,15].

Evaluating Caregiver Engagement in PD

Previous research has involved family caregivers in PD of digital health interventions [16,17]. However, beyond providing a description of the design activities used and the final prototype, the engagement of family caregivers and subject matter experts collaborating in the PD of health information technology has rarely been analyzed, either qualitatively or quantitatively [18,19]. Thus, while we know family caregivers can be incorporated into design processes, we do not know whether the processes used are perceived as effective, inclusive, positive, and impactful by the caregivers involved.

One approach to studying PD for caregivers is to investigate their engagement in PD using the Patient-Centered Outcomes Research Institute (PCORI) theory of action [20]. This framework describes how domains, such as engagement context, engagement activities, and engagement quality, shape outcomes for not just the research but the partners (eg, caregivers) involved. Engagement context is multifaceted, referring to the characteristics of the research team (referred to as *research partners* within the framework) and community members (referred to as *community partners*). Engagement context may include the demographic characteristics, lived experiences, sociohistorical context, feelings of trust, and sense of preparedness to engage in research that partners bring to the design process. Engagement context also includes properties of the design environment, whether remote or in-person. Engagement activities describe the actions taken by research

XSL•FO
RenderX

and community partners during the research process, while engagement quality captures research and community partners' *perceptions* of the research process, including their satisfaction with it. Research outcomes include short- and long-term effects of engagement, such as the release and uptake of findings, and partner outcomes describe how participating in research affected partners, including costs of and benefits to participation. The theory of action can be used to understand ways in which the engagement of caregivers in PD embodies (or fails to embody) engagement ideals.

Evaluating PD in Remote Contexts

PD has historically occurred in the context of in-person workshops, which may include activities such as affinity diagramming, paper prototyping, and working in breakout groups [18]. However, in recent years, there has been a shift toward conducting PD in online or remote environments [21-24]. For family caregivers, the benefits of this shift are clear: caregivers frequently care for someone who cannot be left alone, have limited time to engage in research, and may not always live close to universities or other research institutions [23,25]. The benefits were magnified during the COVID-19 pandemic, when contagion was of particular concern for caregivers of high-risk populations, such as people living with dementia and children with medical complexity [26]. Challenges to using a remote over in-person PD approach include the unequal distribution of and familiarity with technology across subgroups, the need to adapt traditionally in-person design activities to the remote context, and the difficulty of trust- and rapport-building in the remote setting [21,27,28]. There is a need to learn how to best facilitate engagement in remote PD so that high-risk populations can be involved in the design of the digital health interventions intended to meet their needs.

Study Objective

In this study, we analyzed qualitative and quantitative data from evaluation and reflection surveys across research and community partners in 4 separate remote PD studies, all of which focused on building digital health interventions that addressed critical issues for caregivers. Community partners were family caregivers of high-risk populations, including caregivers of persons living with dementia and caregivers of children with medical complexity; members of community organizations who serve caregivers; and medical experts. Research partners were members of the research team who assisted with the facilitation of PD sessions. This analysis aimed to (1) understand the context, quality, and outcomes of partners' engagement experiences in remote PD and (2) learn which aspects of the observed PD approach facilitated engagement or need to be improved going forward.

Methods

Overview of Projects

Evaluation and reflection data came from community and research partners in a convenience sample of 4 PD projects conducted remotely between 2021 and 2023. Each project had a distinct purpose and included different research and community partners from the others. The goal of the CareVirtue Planner (Whiplash Technologies; project 1) was to design a website that could assist caregivers of people living with ADRD in the process of financial and legal planning [22]. The goal of Alzheimer's Family Connection (project 2) was to design a website that could provide resources and connections for rural caregivers of people living with ADRD. The goal of MedS@HOME (project 3) was to design a mobile app that could promote safe medication management among caregivers of children with medical complexity [29]. Finally, the goal of Helping the Helpers (project 4) was to design a technology that could assist caregivers of people living with dementia with medication management at home [30]. This study is a primary analysis of evaluation and reflection data from these 4 projects that have never been published.

Ethical Considerations

Each of the 4 projects, including their evaluation and reflection components, was considered minimal risk human subjects research and was subject to an expedited ethics review by the institutional review board at the relevant institution (Indiana University Bloomington or University of Wisconsin-Madison; project 1: 16227; project 2: 16214; project 3: 16293 project 4: 16063). Each of these studies received a waiver of signed consent from the institutional review board and instead used a study information sheet that outlined all known risks and protections against risks. Potential participants were then provided with the opportunity to ask questions and to provide verbal consent before participation. All data were deidentified. In projects 1, 2, and 4, community partners were compensated with a US $50 e-gift card after each session. For project 3, participants received US $25 per session and were mailed a check after the conclusion of all sessions.

Research Partners

Research partners included principal investigators as well as research coordinators, research specialists, and software engineers working on investigators' grants. Areas of academic and professional expertise included human factors engineering, pediatric medicine, gerontology, counseling psychology, software development, visual communication design, and business development. Project 1 and 2 design sessions were facilitated by author MZ (caregiver support officer, caregiver lived experience); project 3 by NEW (cognitive psychology; associate professor); and project 4 by AG (visual communication design; associate professor) and HP (health informatics). Community and research partners were largely unknown to each other before recruitment and enrollment. In the first design session for each project, the facilitator shared briefly about their personal and professional interests in the research.

Recruitment of Community Partners

For projects 1, 2, and 4, community partners (caregivers of people living with ADRD or organizations that serve these caregivers) were recruited via email through our research team's caregiver registry, community agencies (such as aging and disability resource centers and local chapters of the Alzheimer's Association), and our strategic advisory boards. For project 3, community partners (primary or secondary caregivers of children with medical complexity and medical experts) were recruited

XSL•FO
RenderX

from a local pediatric complex care program via email and postal mail. Recruitment included convenience and snowball sampling methods. For projects 1, 2, and 4, no community partners dropped out after enrolling; for project 3, information about dropouts had been destroyed at the time of analysis to meet confidentiality requirements and is, therefore, unavailable.

Remote Context

All sessions were recorded (with participants' consent) and took place via Zoom (Zoom Video Communications) or Webex (Cisco Systems). All partners were able to unmute themselves and use the chat feature freely. Partners were encouraged, but not required, to keep their cameras turned on. All research partners were granted cohost capabilities to assist with the technical aspects of the meeting.

Study Procedure

For all projects, community partners met with the research partners for 4 to 5-hour-long sessions across 6 months. The agenda for each session was guided by our research team's predefined, published 5-stage process for cocreating a prototype [31]. Sessions focused on problem identification, solution generation, convergence around a single solution, prototyping, and initial evaluation. An example facilitator question related to problem identification is "Where can things go wrong with [your caregiving work]?" and a corresponding question related to solution generation is "What have you done to keep things from going wrong?" After each session, community partners completed an evaluation survey. For project 1, after completing the 5-stage process, volunteering community and research partners completed a reflection survey and interview.

Data Collection

Overview

All evaluation and reflection data were deidentified and stored on an encrypted, cloud-based drive. All evaluation and reflection questions were optional and could be skipped. Because there was no incentive to complete surveys and only meeting participants could access the survey links, it was determined that no additional measures were needed to prevent duplicate entries.

Evaluation Data

Immediately after each design session, community partners completed an evaluation survey on Qualtrics (Qualtrics International Inc) software [32]. Survey links were shared via the Zoom or Webex chat feature, and partners could also request that the link be sent via email. In project 4, research partners also completed evaluation surveys. Evaluation surveys were constructed with the intent to understand partners' satisfaction with the sessions, their perceptions of session effectiveness, and any changes that should be implemented in future sessions. Partners were asked to provide a quantitative rating of the session and qualitative responses to open-ended questions. An example quantitative question was, "On a scale from 1 to 5, how would you rate the effectiveness of the design session?" An example qualitative question was, "What would you suggest we do differently for the next sessions?" Table 1 provides all the quantitative and qualitative evaluation questions asked after each design session for all projects. For projects 1 to 3, all questions appeared on a single survey page, whereas in project 4, questions were spread across 3 survey pages and a "back" button was enabled.

XSL•FO
RenderX

Table 1. Quantitative and open-ended questions asked in evaluation surveys.

Participatory design project	Partners completing evaluations	Quantitative evaluation questions	Open-ended evaluation questions
Project 1: Care-Virtue Planner (Whiplash Technologies)	• ADRD[a] caregivers	• On a scale from 1 to 5, how would you rate the effectiveness of the design session?[b]	• What are your thoughts on the effectiveness? • What would you suggest we do differently for the next sessions? • Is there anything else you want to add?
Project 2: Alzheimer's Family Connection	• Rural ADRD caregivers • Community organizations that serve rural ADRD caregivers	• On a scale from 1 to 5, how would you rate the effectiveness of the design session?[b]	• What are your thoughts on the effectiveness? • What would you suggest we do differently for the next sessions? • Is there anything else you want to add?
Project 3: MedS@Home	• CMC[c] caregivers • Secondary CMC caregivers • Medical experts	• On a scale from 1 to 5, how would you rate the effectiveness of the design session?[b]	• What are your thoughts on the effectiveness? • What would you suggest we do differently for the next sessions? • Is there anything else you want to add?
Project 4: Helping the Helpers	• ADRD caregivers • Research partners	• Overall, how satisfied are you with today's design session?[d] • Did this session make you feel worse or better about the remaining sessions?[e]	• What would you need to improve your satisfaction with the design sessions? • What made today's session more effective? • What made today's session less effective? • What could be done to better reach goals or make expectations clearer? • Do you have any comments or suggestions about the session that you did not get a chance to share today?

[a]ADRD: Alzheimer disease and related dementias.

[b]The questions were scaled from 1 to 5: not effective, somewhat effective, neutral, mostly effective, and very effective.

[c]CMC: children with medical complexity.

[d]The questions were scaled from 1 to 5: extremely dissatisfied, somewhat dissatisfied, neither satisfied nor dissatisfied, somewhat satisfied, and extremely satisfied.

[e]The question were scaled from 1 to 5: much worse, somewhat worse, about the same, somewhat better, and much better.

Reflection Data

For project 1, after all the co-design sessions had been completed, research and community partners were offered the opportunity to complete a reflection interview with a staff member who was not a research partner and a reflection survey via Qualtrics. Interviews were recorded, took place on Zoom, and lasted no longer than 30 minutes. Links to the reflection survey were shared in the chat following the interview. The reflection interview guide and survey were coconstructed by research partners on project 1. An example reflection interview question was, "How did your expectations compare/contrast with your actual experience in the design sessions?" In an example quantitative reflection survey question, partners were asked to rate the following statement on a scale of 1 to 100: "My own participation in the design sessions influenced the design of the product." Questions were spread across 3 survey pages and going "back" was not possible. Textbox 1 presents all questions asked in the reflection interview and survey.

XSL•FO
RenderX

Textbox 1. Quantitative and open-ended reflection questions asked after project 1.

Partners completing reflection

- Caregivers of people with Alzheimer disease or related dementias
- Research partners

Quantitative questions

Drag along the slider scale (1-100) to show the extent to which the following statements are true for you:

- My own participation in the design sessions influenced the design of the product.
- The participation of caregivers influenced the design of the product.
- The participation of the research team members influenced the design of the product.

Open-ended questions (interview)

- What expectations did you have for the design process before the first design session?
- How did these expectations compare and contrast with your actual experience in the design sessions?
- What expectations did you have for the legal and financial planning tool before the first design session?
- How did these expectations compare and contrast to the ultimate prototype?
- How do you feel your participation in the design sessions influenced product design?

Open-ended questions (survey)

- Please elaborate on how you think caregiver participation did or did not influence the design of the product.
- Please elaborate on how you think research team member participation did or did not influence the design of the product.

Data Analysis

Descriptive statistics were used to summarize quantitative evaluation and reflection survey data from research and community partners. For the preparation of interview data, 2 research team members (AJ and AL) separately watched each reflection interview and took notes on the content. It was not presumed that research team members' notes could be identical but rather that the use of 2 observers could yield additional and unique observations from which meaning could be derived [33]. The interview notes that did not appear to constitute feedback were excluded from coding. For the preparation of qualitative evaluation and reflection survey data, partner responses that were not believed to constitute feedback (eg, the partner had written "not applicable" or "no opinion") were excluded from coding. These excluded data were reviewed by a senior researcher with more training in reflexive thematic analysis (NEW) to decrease the likelihood that meaning relevant to the research questions was being dismissed. Incomplete responses (eg, if a partner answered only 1 of the 3 open-ended questions) were retained for analysis. Each response was a single unit of analysis, and responses could be and often were coded to multiple themes.

Evaluation and reflection survey data were combined with interview notes and exported into Excel (Microsoft Corporation). One research team member (AJ) reviewed all the data line-by-line and grouped data by 3 PCORI domains, including engagement context, engagement quality, and partner outcomes. This grouping was completed with constant reference to PCORI's verbatim definitions of each domain. Following this grouping, we adopted the "Big Q" approach of reflexive thematic analysis by Braun and Clarke [34]. Thus, an objective or single "correct" interpretation of the data was not presumed to be possible. Rather, it was expected that coding would result in an interpretive story about each PCORI domain. Themes were created, iterated upon, and constantly examined for apparent fit to previously coded data.

The primary coder (AJ) was a licensed mental health counselor with training in interpersonal process groups. Facilitators of these types of groups are trained to be mindful of certain risks to group work, including insufficient preparation and imbalances of power, as well as certain benefits, including self-exploration and a sense of connectedness to other group members [35]. The primary coder was present for many, but not all, of the PD sessions and held a junior status on the research team and a nonfacilitative role in sessions. These aspects of the primary coder's positioning inevitably shaped their interpretation of data. It is not expected that a different coder in a different context would obtain identical results; however, the thick descriptions of our data collection and analytic processes provided in this manuscript may enable others to critically analyze how we arrived at these results and understand the applicability of these results in other contexts.

The final codebook was presented to senior researchers with training in reflexive thematic analysis (NEW and RV) and research partners on projects 1 to 4 to elicit their feedback. The goal of this process was not to rid the codebook of "bias" or the primary coder's perspective. Rather, this team-wide review enabled crystallization, an alternative to triangulation that seeks comprehensiveness and depth over convergence or consensus [36,37]. Because many of the authors of this manuscript were also research partners in the projects described, the cowriting of results allowed many opportunities for member reflections,

XSL•FO
RenderX

which is an alternative to member checking in which participants offer additional, unique insight rather than confirming or denying that the data have been interpreted "correctly" [33].

Results

Quantitative Results

Overview

A total of 43 community partners and 19 research partners were involved in 31 PD sessions. Within these samples, a subset of 43 (100%) community partners and 2 (10%) research partners completed 170 evaluation surveys. Partners in project 1 completed 32.9% (56/170) of the total surveys, project 2 completed 21.8% (37/170), project 3 completed 22.9% (39/170), and project 4 completed 22.4% (38/170). A subset of 14 partners (7/43, 16% of community partners and 7/19, 37% of research partners) completed a reflection interview and survey. Table 2 presents the demographic information for community and research partners.

Table 2. Demographic information for community and research partners.

Characteristics	Community partners (n=43)	Research partners (n=19)
Age (y), mean (SD)	52.3 (15.9)	36.6 (10.1)
Gender, n (%)		
Women	31 (72)	11 (58)
Men	12 (28)	7 (27)
Nonbinary	0 (0)	1 (5)
Race and ethnicity, n (%)		
Asian	1 (2)	1 (5)
Black	2 (5)	0 (0)
Hispanic	1 (2)	0 (0)
Multiracial	1 (2)	0 (0)
White	35 (81)	18 (95)
Other	3 (7)	0 (0)

Quantitative Evaluation and Reflection Data

In 62.9% (83/132) of evaluations across projects 1-3, participants described the session as "very effective." In 74% (28/38) of evaluations for project 4, participants described feeling "extremely satisfied" with the session. The mean effectiveness and satisfaction ratings ranged between 4 and 5 for all design sessions. Table 3 shows the mean effectiveness, satisfaction, and feeling across sessions.

Community partners rated their own influence and the influence of other co-designers on the final design at an average of 88.8 (SD 14.7) and rated the research team's influence at 82.4 (SD 24.5) on a scale of 1 to 100. Research partners rated community partner's influence at an average of 82.6 (SD 21.7) and their own influence and that of other research partners at 64.3 (SD 37.5) on the same scale of 1 to 100.

Table 3. Quantitative evaluations of design sessions with time.

Evaluation question	Session 1 (n=7), mean (SD)	Session 2 (n=7), mean (SD)	Session 3 (n=6), mean (SD)	Session 4 (n=6), mean (SD)	Session 5 (n=5), mean (SD)
On a scale from 1 to 5, how would you rate the effectiveness of the design session?[a] (projects 1-3)	4.4 (0.8)	4.7 (0.6)	4.9 (0.3)	4.8 (0.4)	4.9 (0.2)
Overall, how satisfied are you with today's design session?[b] (project 4)	4.4 (0.8)	4.9 (0.4)	4.6 (0.5)	4.4 (1.4)	4.9 (0.4)
Did this session make you feel worse or better about the remaining sessions?[c] (project 4)	3.75 (1.0)	4.6 (0.5)	4.1 (1.0)	4.4 (0.9)	__[d]

[a]The questions were scaled from 1 to 5: not effective, somewhat effective, neutral, mostly effective, and very effective.

[b]The questions were scaled from 1 to 5: extremely dissatisfied, somewhat dissatisfied, neither satisfied nor dissatisfied, somewhat satisfied, and extremely satisfied.

[c]The questions were scaled from 1 to 5: much worse, somewhat worse, about the same, somewhat better, and much better.

[d]Not applicable.

XSL•FO
RenderX

Qualitative Results

Overview

After subtracting 56 noncoded responses, partners provided 676 free-text responses to evaluation surveys, reflection surveys, and interview notes. Within these responses, 13.5% (91/676) were categorized as nonspecific positive feedback and were not believed to fit an additional theme. In total, 17 themes were created across the 3 PCORI domains of engagement context, engagement quality, and partner outcomes.

Nonspecific Positive Feedback

In response to the free-text evaluation questions, partners often responded with brief, positive answers (nonspecific positive feedback). When asked their thoughts on the effectiveness, some partners provided responses such as “great,” “wonderful,” and “all good so far.” When asked what improvement they would like to see in future sessions, some partners responded with “nothing,” and when asked whether they had any remaining comments or suggestions, they often said “no.”

Engagement Context

We constructed 4 themes within engagement context: identity influence, technological context, project understanding, and role understanding. Table 4 presents the themes, definitions, and example quotations relating to engagement context.

Partners perceived some aspect of their identity (eg, the status of being a sexual or gender minority individual, the generational identity of being a “millennial,” or a career as a publisher) as influencing their engagement (identity influence). Partners also described the remote context of PD as shaping their expected or actual experience (eg, feeling wary of using videoconferencing software and limitations of remote PD in contrast with in-person PD; technological context). One research partner noted that the process was “less hands on” than traditional PD and wondered how to achieve this going forward in a remote context.

Two context themes pertained to whether partners felt prepared to engage in research, including project understanding and role understanding. Research and community partners alike described the project as different from or consistent with their expectations, reflecting the presence or absence of project understanding before engaging in co-design. Partners also expressed confusion with aspects of the projects once they began, such as what they were designing and how they would design it, reflecting the absence of project understanding once the project was underway. Partners expressed either confusion or understanding pertaining to their role (eg, 1 research partner noted being “less involved than I expected to be”) and expressed perceptions of their role that were sometimes contradictory and sometimes consistent with those of their peers (role understanding).

Table 4. Engagement context[a] themes, definitions, and examples.

Theme	Definition	Examples
Identity influence	A partner perceives one of their identities or traits as influencing their engagement in the project.	• “As a sexual and gender minority, some legal and financial processes as a caregiver are more complicated. I was able to bring that perspective.” [Research partner’s interview note] • “Background in publishing, experience brainstorming designs and content” [Community partner’s interview note]
Technological context	A partner perceives the remote context of co-design as a facilitator or barrier to engagement.	• “I suffer from ‘technostress’ - I thought it would be boring, people wouldn’t share.” [Community partner’s interview note] • “One loss of digital platforms- can't have two cooks in the kitchen (can't have side conversations, less potential for multiple people to get involved).” [Research partner’s interview note]
Project understanding	A partner describes the project as consistent with or different from their expectations or describes having no expectations for the project, and a partner expresses understanding or confusion regarding some aspect of the project.	• “I think the goals and expectations are clear!” [Community partner’s evaluation survey] • “I expected to pitch ideas, make sketches, draw on the whiteboard. It was less hands on than that.” [Research partner’s interview note] • “I think that there’s a disconnect between those working on developing the app and us who see the app. For those developing it, many things might seem intuitive as they will understand what’s coming next, but for us, who see the app and can only use the functions highlighted in blue at this point, it’s less intuitive.” [Community partner’s evaluation survey]
Role understanding	A partner expresses understanding or confusion about their role, and partners express contradictory perceptions regarding their role.	• “Thought research team was going to be much more responsible for keeping things going- taking notes, making agendas- basically leading it. and we weren’t.” [Research partner’s interview note] • “In a traditional co-design process, I would have expected that I was as much a co-designer as anyone else. It felt like if I was getting involved in this process, I would have felt like I was taking space away from the participants.” [Research partner’s interview note]

[a]Engagement context: “Resources and circumstances surrounding the practice of engagement in research that may affect how engagement occurs and its impact” [20].

XSL•FO
RenderX

Engagement Quality

We constructed 8 themes within engagement quality: relationship-building, co-learning, desire for prework, satisfaction with design activities, satisfaction with the time allotted, satisfaction with the final tool, influence, and inclusivity. Table 5 presents the themes, definitions, and example quotations relating to engagement quality.

Community partners commented on their appreciation of other group members (eg, their contributions and demeanor; relationship-building) and reported learning from other group members (eg, about a caregiver resource; co-learning). Partners also described feeling satisfied or dissatisfied with various aspects, including prework (eg, appreciating the work assigned before sessions or wanting more of it), the design activities (eg, commenting that activity was effective or suggesting a different activity), the time allotted to complete activities (eg, saying that there was not enough time), and the prototype (eg, appreciating its functionality or wishing it were more complete).

Partners shared both positive and negative perceptions of decision-making processes, including the extent to which decisions were guided by research partners versus community partners (influence). Furthermore, partners commented on the extent to which they felt empowered to contribute, actually did contribute, or perceived contributions as equal across partners (inclusivity).

XSL•FO
RenderX

Table 5. Engagement quality[a] themes, definitions, and examples.

Theme	Definition	Examples
Relationship-building	A partner perceives relating to others to be a key aspect of engagement.	• "We created an environment that was like a support group. It was incredible and I did not expect it." [Research partner's interview note] • "Nice to meet other people in a similar spot." [Community partner's interview note]
Co-learning	A partner perceives the engagement experience as educational (education may come from research partners or from each other).	• "Other members mentioned [resources] in examples that I never heard of before." [Community partner's evaluation survey] • "It is nice to share ideas with people with similar experiences, to understand what has happened to a loved one (or is currently happening)." [Community partner's evaluation survey]
Satisfaction with prework	A partner expresses the desire for more engagement between design sessions.	• "Wished diagrams/materials had been sent out a few days before so he could digest and review." [Community partner's interview note] • "I like the information that is sent prior to our sessions." [Community partner's evaluation survey]
Satisfaction with design activities	A partner evaluates the activities used in sessions, and a partner suggests different activities.	• "I did not like the expectation of drawing something, I will never draw for anyone" [Community partner's interview note] • "I would have liked more time to think about what problem I wanted to solve and how I could envision a solution." [Community partner's evaluation survey] • "I'm excited. I enjoyed the once a month meetings and seeing the development of the app." [Community partner's evaluation survey]
Satisfaction with the time allotted	A partner evaluates the time allotted for sessions.	• "I wouldn't mind if they were scheduled for a longer time -- maybe 1.25 or 1.5 hours. It would be nice to not to feel a bit rushed at the end." [Community partner's evaluation survey] • "I think a couple more sessions would have been beneficial." [Community partner's interview note]
Satisfaction with the prototype	A partner evaluates the prototype seen in the session, and a partner makes a suggestion for the prototype.	• "Oh my gosh, this [product] would have been so helpful... I'm looking forward to having that go live." [Community partner's evaluation survey] • "Not having a robust app to see each of the features wasn't as effective as having a fully functional app." [Community partner's evaluation survey]
Influence	This is the partners' perception of who shaped the final design and how they shaped it.	• "Every time we met, the team introduced us to improvements to the modules that were implemented based on what they heard us say." [Community partner's reflection survey] • "Co-designers directly influenced the design of the product by responding to lead researcher prompts and questions. Those answers were then actualized in design. However, I do feel that the final product would have been more co-designer driven if co-designers had been involved in identifying what type of tool would have been most helpful to them... and identifying what topics should have been discussed at all in design sessions" [Research partner's reflection survey]
Inclusivity	This is the extent to which partners felt empowered to contribute or actually contributed	• "Just like me, every other participant in the design sessions was able to provide ideas, recommendations, and their questions were welcomed." [Community partner's reflection survey] • "The identities of those involved in the co-design teams played a big role in whose voices were heard... There were a couple of White men who spoke up a lot. Our research team was all White; how did that affect who shared and what they shared?" [Research partner's reflection survey]

[a]Engagement quality: "The perceptions, assessments and feelings of partners and researchers about the process of engagement" [20].

Partner Outcomes

We constructed 4 themes within the domain of partner outcomes: ongoing project interest, gratitude, self-esteem, and sense of meaning. Table 6 presents the themes, definitions, and example quotations relating to partner outcomes.

Community partners expressed interest in the future of the project even after the conclusion of design sessions (eg, asking how they could further contribute; ongoing project interest). Furthermore, partners expressed gratitude for the research partners, for their fellow community partners, or for the opportunity to engage in research (gratitude). Some community partners noted more positive self-appraisal as a result of engagement (self-esteem), while others found meaning in having made a positive contribution to science and to other caregivers (sense of meaning).

XSL•FO
RenderX

Table 6. Partner outcomes[a] themes, definitions, and examples.

Theme	Definition	Examples
Ongoing project interest	Partner expresses interest in continued participation in the project even after the design sessions have ended and partners suggest next steps for the project.	• "What happens next? How can we further contribute? Can we see a newer version of the prototype 6 months from now?" [Community partner's interview note] • "I'm looking forward to seeing future progress and if you want any feedback, I'm always willing to participate - even if the funding for gift cards has run out." [Community partner's evaluation survey]
Gratitude	Partner describes that engagement in the design sessions inspires gratitude.	• "Thank you!! Thank you for including [community partner's state] in this effort." [Community partner's evaluation survey]
Self-esteem	Partner's self-evaluation becomes more positive as a result of engagement.	• "I thought I was valuable, I'm proud of myself, I've been a caregiver for years." [Community partner's interview note] • "Sharing information and realizing we struggled with the same issues. You don't feel so alone." [Community partner's evaluation survey]
Sense of meaning	Partner feels that they have made a meaningful contribution through their engagement.	• "The opportunity for participation and to express what we were going through was really powerful." [Community partner's interview note] • "I learned that I am very fortunate in my situation... Eager to help where needed and to simplify others needs when and where ever possible." [Community partner's evaluation survey]

[a]Partner outcomes: "Impact of engagement on the individuals, organizations, and communities partnering in research" [20].

Discussion

Principal Findings

This study aimed to (1) understand the context, quality, and outcomes of partners' engagement experiences in remote PD and (2) learn what aspects of the observed PD approaches facilitated engagement or need to be improved. Much of the evaluation and reflection data were nonspecific but positive, and the effectiveness and satisfaction ratings for each session ranged between 4 and 5 on a 5-point scale, suggesting that many partners had a generally positive experience. Factors associated with engagement context were found to influence engagement, such as the context brought by partners (eg, their identities and their understanding of the project) as well as the context created during the project (eg, the online setting). Within the domain of engagement quality, partners reported greater or lesser degrees of satisfaction with session activities, relationship-building, influence, and inclusivity. Partners who commented on personal outcomes of participation were uniformly positive, noting continued desire to participate in research and a range of psychosocial benefits.

Preparing Partners for the Remote Context

Although analysis of remote PD is nascent, research on how remote settings affect community-based participatory research (CBPR) is more advanced, and learnings may be applicable to PD [23,27,38-41]. In CBPR, scientific researchers partner with community members, often from high-risk populations, to identify a health concern of interest to community members; build a deep, contextualized understanding of this concern; investigate and cocreate potential interventions to address the concern; and, finally, disseminate and try to sustain these interventions within the community [42]. Previous CBPR notes that the remote approach comes with technological challenges, including differing levels of familiarity with teleconferencing software and the development of "Zoom fatigue" [27]. In this study, research and community partners did express some wariness about using teleconferencing software for PD. Research partners noted challenges with translating traditionally in-person design processes to the remote space, noting fewer opportunities for potentially generative side conversations between partners. However, the remote setting did not appear to prohibit colearning or relationship development, although experimental study designs are needed to confirm this finding.

Existing literature in the fields of CBPR and patient engagement highlights the necessity of adequately preparing community partners to engage in research [28,43]. In remote PD, it is not commonplace to provide partners with this thorough background information. In this study, for projects 1 to 4, community partners were provided with a study information sheet, a consent and screening call, and an introductory session in which the description of the project lasted for up to 20 minutes. However, our findings related to project and role understanding suggest that, for some partners, this preparation was insufficient. For some partners, poor understanding may have led to reduced participation. Future remote PD studies should integrate community partners from the project inception, including defining the problem of greatest interest to community partners, coauthoring the approach, and brainstorming how to best use the remote space. Furthermore, during and between design sessions, researchers may use an approachable, predefined process to clarify partners' understanding of the project's goals and partner roles. To this end, the remote context allows researchers to easily distribute links to surveys, conduct polls, and receive public and private chats. Furthermore, researchers should prepare materials in multiple formats (eg, written, auditory, and image-based materials) and provide alternative ways of phrasing concepts to maximize community partner comprehension [44].

XSL•FO
RenderX

Activities to Promote Remote Engagement

Community and research partners were mixed in their evaluation of inclusivity and influence. In contrast to previous research, community partners in this study reported seeing their contributions reflected in the design and rated themselves as having had a strong influence on the design [45]. However, echoing previous research, some research partners in this study felt that the design process was not sufficiently led by the community partners, perceiving community partners as providing feedback on the design but not leading design efforts [46]. For PD products to be the best possible fit for their intended populations and for PD to hold positive meaning for community and research partners alike, inclusivity and equality of influence must be prioritized in PD as they are in CBPR [43,47]. It is recommended that research and community partners alike speak candidly about the potential for power differentials present in the broader culture (eg, among those of different genders, races, disabilities, levels of education, and project roles) to manifest in remote spaces as they do in physical spaces [43,48-51]. Partners should speak frequently about how to create and maintain group structures and norms that combat power imbalances. Conversations such as these require trust. Thus, the concepts of inclusivity and influence should be introduced to partners before the design process begins, senior research partners should model these conversations early, and community partners should be encouraged to contribute as they feel it is safe to do so. For community and research partners who do not feel safe voicing these concerns aloud, the remote context allows for the distribution of anonymous surveys.

Community partners ranged in their satisfaction with the design sessions, including with the presession work, the design activities that occurred in the session, the time allotted, and the status of the prototype in the final session. While the activities used in these remote PD study are consistent with those of in-person studies, it is recommended that future studies use more nuanced evaluation methods to determine which remote activities were perceived as most effective and enjoyable by partners [18]. When designing the structure of the study, research partners sought to make modest requests of community partners to avoid fatigue or study attrition, which are problems that have been documented in other projects conducted remotely [27]. However, qualitative analyses revealed that community partners enjoyed the presession work, often requesting that more materials and reflection questions be sent out in advance. Community partners often commented that 1-hour sessions felt too short for the task at hand. This suggests that community partners are prepared for more intensive engagement in remote design. In future remote studies, online process checks may be used to gather real-time feedback about the workload and time commitment, and this feedback should be reflected in changes to the study timeline.

Maximizing Community Partner Outcomes

Finally, our results concur with CBPR literature, suggesting that remote PD has the potential to create positive outcomes for partners, including empowerment for partners and positive impacts on their health [47]. Community partners in this study expressed gratitude for their involvement, greater self-esteem, and a sense of purpose. Furthermore, community partners expressed interest in continued remote involvement with the design and implementation of the tool. When designing this study, research partners underestimated the extent to which involvement could provide a meaningful experience for community partners. Going forward, it is recommended to conceptualize remote PD as a process that may help community partners (especially those who are isolated due to geographic or physical limitations) achieve developmental tasks, such as intimacy (closeness with others), generativity (benefiting future generations), and integrity (having lived a meaningful life) [52]. Involving community partners to this extent requires compensation for their participation. This goes beyond financial compensation to include remote training relevant to the research topic and methods, facilitating remote networking between partners, distance collaboration on manuscripts, and copresentation at conferences [28,53-55]. Furthermore, for designs that reach the commercial marketplace, community partners should share any resulting profits.

Recommendations for Remote PD Research

Overall, these results point to high satisfaction with our remote PD processes, as well as specific ways in which processes can be changed to improve partner engagement and avenues for maximizing positive partner outcomes. Recommendations for engaging community and research partners in remote PD are listed in Table 7.

XSL•FO
RenderX

Table 7. Integrated results and strategies for engaging research and community partners in remote participatory design (PD) research.

Qualitative theme or quantitative finding	Strategies for engagement in remote PD
Engagement context: identity influence; project understanding; role understanding	• From project inception, include partners in defining the problem to be solved and the approach to be used (to facilitate project understanding) and in defining their role (to facilitate role understanding). • Include partners of different gender identities, sexual orientations, races, disabilities, levels of education, and levels of academic seniority. • During the PD process, use live discussion, surveys, polls, and the chat feature to perform remote "process checks" to ascertain partners' project and role understanding.
Engagement context: technological context	• From project inception, involve partners in defining how to use the remote space, including specifying the technologies (eg, teleconferencing and online whiteboards) and technological functions (eg, chat and poll) with which they are comfortable. • Outside of PD sessions, assist interested partners in gaining technological proficiency through 1:1 or small group meetings.
Engagement quality: satisfaction with prework; satisfaction with design activities; satisfaction with the time allotted; high quantitative satisfaction ratings with design sessions	• From project inception, involve partners in defining the amount of prework desired or needed, imagining design activities, and designating the time allotted for activities. • During the PD process, use live discussion, surveys, polls, and the chat feature to solicit feedback on changes that should be made to the design process. • After the PD process, ask sensitive and specific evaluation and reflection questions to understand their satisfaction with different aspects of the design process.
Engagement quality: satisfaction with the prototype; influence; inclusivity; quantitative difference between community and research partners' estimations of who influenced the final product	• Encourage dual roles, that is, equipping community partners to facilitate the research process or hiring research partners who are members of the community of interest. • From project inception, emphasize and define processes for sharing power, especially in remote spaces (eg, raising hands and inviting quieter partners to share). • During the PD process, define processes for intervening when power is not being shared (ie, what to say, how to say it, whom to say it to, and through what medium). • During the PD process, make available a link to an anonymous survey in which partners can submit feedback and ideas on inclusion and power-sharing.
Engagement quality: relationship-building; co-learning	• From project inception, emphasize and expand the potential for building connections between partners (eg, with consent, facilitating the exchange of contact information). • Elicit caregivers' areas of expertise or experience within caregiving and facilitate knowledge exchange within and outside of sessions.
Partner outcomes: ongoing project interest; gratitude; self-esteem; sense of meaning	• At project conclusion, invite community partners to participate in future aims of the project or related projects. • Budget for involving community partners in the dissemination of findings, including conference travel, coauthorship of manuscripts, and bringing findings to communities.

Limitations and Future Research

This analysis had some limitations. First, across design sessions, research and community partners were disproportionately White and cisgender individuals. This lack of diversity may have led to an incomplete understanding of how concepts such as trust, inclusivity, and influence manifest in remote PD [56]. Identity-based barriers to engagement were seldom mentioned by partners in this study. However, future research should investigate how to meaningfully engage partners with physical disabilities, as certain partners with disabilities may uniquely benefit from the remote setting (remote PD requires less mobility) while others may find it more complicated (remote PD is often dependent on visual media) [57]. Second, evaluation and reflection data are only as accurate as partners felt safe to provide. It is vital to establish and reinforce trust between research and community partners and to provide anonymous remote formats for providing feedback so that partners do not feel pressured to provide socially desirable feedback. Third, although community partner evaluation data were available for all studies, research partners only completed evaluations in 1 study (project 4), and reflection data were only gathered from research and community partners in 1 study (project 1). It is not yet commonplace to gather evaluation and reflection data from research and community partners in PD studies [18]. However, to facilitate growth in the science of improving remote PD engagement, future studies should more rigorously gather evaluation and reflection data from all partners across domains, such as engagement context, engagement quality, and partner outcomes. Fourth, the design of this study did not allow researchers to directly compare in-person PD to remote PD on measures of engagement; future research should adopt experimental designs to this end. Finally, the evaluation and reflection data reflect only short-term partner outcomes of engagement. Future work should evaluate whether heightened senses of meaning, gratitude, and community persist long term as a result of partnering in remote PD.

Conclusions

This analysis of 4 remote PD studies points to ways in which remote PD processes must be more thoroughly evaluated and, where indicated, changed to enhance context, quality, and

XSL•FO
RenderX

outcomes for partners. While the remote context was sometimes a barrier to collaboration, it did not seem to prohibit colearning or the development of relationships. Indeed, many of our findings echo previous work on in-person PD, indicating few losses specific to remote PD. It may be that the gains associated with remote PD, including the reduced expense, reduced travel time, and greater potential for engaging underserved populations, surpass the losses of the remote setting. Future work should not just evaluate the adaptation of in-person activities to remote settings but rather more thoroughly reimagine PD as a remote process with the unique affordances of increasingly sophisticated remote environments. Finally, our results suggest that community partners are open to a more time-intensive commitment to remote PD. Future research should systematically examine whether more intensive involvement of community and research partners in remote PD yields better outcomes for research, partners, and the communities that the research aims to serve.

Acknowledgments

The authors would like to thank all the community partners who participated in and provided feedback on remote participatory co-design. The authors would also like to thank Priya Loganathar for conducting interviews with research partners. Project 1 was supported by the National Institutes of Health's National Institute on Aging (grant 1R44AG074128). Project 2 was supported by the National Institutes of Health's National Institute on Aging (grant R43AG080849). Project 3 was supported by the Agency for Healthcare Research and Quality (grant R18HS028409). Project 4 was supported by the National Institutes of Health's National Institute on Aging (grant 1R21AG072418). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Agency for Healthcare Research and Quality.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

The study was conceptualized by AJ, NEW, RV, and RJH. The funding for the study was acquired by NEW, RC, RJH, MZ, and CE. AJ curated the data of the study. AJ, AL, NEW, and RV were involved in the formal analysis of the study, and NEW supervised the study. AJ and NEW wrote the original draft, and AJ, RJH, RV, NEW, HP, MZ, AL, and RC reviewed and edited the manuscript.

Conflicts of Interest

The results of projects 1 and 2 will be integrated into CareVirtue, which is a digital portal for caregivers. CE is the chief executive officer and cofounder of CareVirtue Technologies and developed CareVirtue. MZ is the caregiver support officer and cofounder of CareVirtue Technologies and developed CareVirtue; an associate for Hilarity For Charity, which is a 501c3 with a mission of bringing light to patients with Alzheimer disease; and a member of the Alzheimer's Impact Movement, which is the advocacy affiliate of the Alzheimer's Association.

References

1. Caregiving. Family Caregiver Alliance. URL: https://www.caregiver.org/resource/caregiving/ [accessed 2024-09-02]
2. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023 Apr;19(4):1598-1695. [doi: 10.1002/alz.13016] [Medline: 36918389]
3. Kuo DZ, Houtrow AJ, Council on Children With Disabilities. Recognition and management of medical complexity. Pediatrics 2016 Dec;138(6):e20163021 [FREE Full text] [doi: 10.1542/peds.2016-3021] [Medline: 27940731]
4. Berry JG, Hall M, Neff J, Goodman D, Cohen E, Agrawal R, et al. Children with medical complexity and Medicaid: spending and cost savings. Health Aff (Millwood) 2014 Dec;33(12):2199-2206 [FREE Full text] [doi: 10.1377/hlthaff.2014.0828] [Medline: 25489039]
5. Miller AR, Condin CJ, McKellin WH, Shaw N, Klassen AF, Sheps S. Continuity of care for children with complex chronic health conditions: parents' perspectives. BMC Health Serv Res 2009 Dec 21;9:242 [FREE Full text] [doi: 10.1186/1472-6963-9-242] [Medline: 20025770]
6. Hughes TB, Black BS, Albert M, Gitlin LN, Johnson DM, Lyketsos CG, et al. Correlates of objective and subjective measures of caregiver burden among dementia caregivers: influence of unmet patient and caregiver dementia-related care needs. Int Psychogeriatr 2014 Aug 08;26(11):1875-1883 [FREE Full text] [doi: 10.1017/s1041610214001240]
7. Cotton QD, Kind AJ, Kim AJ, Block LM, Thyrian JR, Monsees J, et al. Dementia caregivers' experiences engaging supportive services while residing in under-resourced areas. J Alzheimers Dis 2021;84(1):169-177 [FREE Full text] [doi: 10.3233/JAD-210609] [Medline: 34487046]
8. Lindeman DA, Kim KK, Gladstone C, Apesoa-Varano EC. Technology and caregiving: emerging interventions and directions for research. Gerontologist 2020 Feb 14;60(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1093/geront/gnz178] [Medline: 32057082]

XSL•FO
RenderX

9. d'Halluin A, Costa M, Morgiève M, Sebbane D. Attitudes of children, adolescents, and their parents toward digital health interventions: scoping review. J Med Internet Res 2023 May 02;25:e43102 [FREE Full text] [doi: 10.2196/43102] [Medline: 37129931]
10. Spinuzzi C. The methodology of participatory design. Tech Commun 2005;52(2):163-174.
11. Holden RJ, Valdez RS. The Patient Factor: Theories and Methods for Patient Ergonomics. Boca Raton, FL: CRC Press; 2021.
12. Hoogsteyns M, Zaal-Schuller I, Huisman S, Nieuwenhuijse AM, van Etten-Jamaludi F, Willems D, et al. Tacit knowledge in dyads of persons with profound intellectual and multiple disabilities and their caregivers: an interpretative literature study. J Appl Res Intellect Disabil 2023 Sep;36(5):966-977 [FREE Full text] [doi: 10.1111/jar.13134] [Medline: 37339925]
13. Gorman RK, Wellbeloved-Stone CA, Valdez RS. Uncovering the invisible patient work system through a case study of breast cancer self-management. Ergonomics 2018 Dec;61(12):1575-1590 [FREE Full text] [doi: 10.1080/00140139.2018.1503339] [Medline: 30044709]
14. Cornet VP, Toscos T, Bolchini D, Rohani Ghahari R, Ahmed R, Daley C, et al. Untold stories in user-centered design of mobile health: practical challenges and strategies learned from the design and evaluation of an app for older adults with heart failure. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17703 [FREE Full text] [doi: 10.2196/17703] [Medline: 32706745]
15. Cornet VP, Daley C, Bolchini D, Toscos T, Mirro MJ, Holden RJ. Patient-centered design grounded in user and clinical realities: towards valid digital health. Proc Int Symp Human Factors Ergon Health Care 2019 Sep 15;8(1):100-104 [FREE Full text] [doi: 10.1177/2327857919081023]
16. Sin J, Henderson C, Woodham LA, Sesé Hernández A, Gillard S. A multicomponent eHealth intervention for family carers for people affected by psychosis: a coproduced design and build study. J Med Internet Res 2019 Aug 06;21(8):e14374 [FREE Full text] [doi: 10.2196/14374] [Medline: 31389333]
17. Yuwen W, Duran M, Tan M, Ward TM, Cheng SC, Ramirez M. Self-care needs and technology preferences among parents in marginalized communities: participatory design study. JMIR Pediatr Parent 2021 Jun 22;4(2):e27542 [FREE Full text] [doi: 10.2196/27542] [Medline: 34156343]
18. Ozkaynak M, Sircar CM, Frye O, Valdez RS. A systematic review of design workshops for health information technologies. Informatics 2021 May 14;8(2):34 [FREE Full text] [doi: 10.3390/informatics8020034]
19. Leslie M, Khayatzadeh-Mahani A, MacKean G. Recruitment of caregivers into health services research: lessons from a user-centred design study. Res Involv Engagem 2019;5:17-19 [FREE Full text] [doi: 10.1186/s40900-019-0150-6] [Medline: 31139432]
20. Measuring what matters for advancing the science and practice of engagement. Patient-Centered Outcomes Research Institute. URL: https://www.pcori.org/resources/measuring-what-matters-advancing-science-and-practice-engagement [accessed 2024-11-18]
21. Lee KJ, Roldan W, Zhu TQ, Saluja HK, Na S, Chin B, et al. The show must go on: a conceptual model of conducting synchronous participatory design with children online. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan URL: https://doi.org/10.1145/3411764.3445715 [doi: 10.1145/3411764.3445715]
22. Jolliff A, Dudek A, Zuraw M, Parks R, Linden A, Elliott C, et al. Co-design of a financial and legal planning tool for caregivers of people living with Alzheimer's disease and related dementias. Innov Aging 2023;7(5):igad046 [FREE Full text] [doi: 10.1093/geroni/igad046] [Medline: 37360216]
23. Valdez ES, Gubrium A. Shifting to virtual CBPR protocols in the time of Corona Virus/COVID-19. Int J Qual Methods 2020 Dec 07;19 [FREE Full text] [doi: 10.1177/1609406920977315]
24. Hill JR, Abebe E, Holden RJ, McLaughlin AC, Rogers CC, Werner NE. Reaching people where they are: remote macroergonomics research. Proc Human Factors Ergon Soc Annual Meeting 2022 Oct 27;66(1):1617-1621 [FREE Full text] [doi: 10.1177/1071181322661012]
25. Villanova P. Where baby boomers are moving – 2023 study. SmartAsset. 2023 Nov 28. URL: https://smartasset.com/data-studies/where-boomers-moved-2023 [accessed 2024-11-18]
26. Hill JR, Brown JC, Campbell NL, Holden RJ. Usability-in-place-remote usability testing methods for homebound older adults: rapid literature review. JMIR Form Res 2021 Nov 02;5(11):e26181 [FREE Full text] [doi: 10.2196/26181] [Medline: 34726604]
27. Tariq S, Grewal EK, Booth R, Nat B, Ka-Caleni T, Larsen M, et al. Lessons learned from a virtual community-based participatory research project: prioritizing needs of people who have diabetes and experiences of homelessness to co-design a participatory action project. Res Involv Engagem 2023 Jul 04;9(1):46 [FREE Full text] [doi: 10.1186/s40900-023-00456-z] [Medline: 37403184]
28. Hamilton C, Hoens A, Backman C, McKinnon AM, McQuitty S, English K, et al. An empirically based conceptual framework for fostering meaningful patient engagement in research. Health Expect 2018 Feb;21(1):396-406 [FREE Full text] [doi: 10.1111/hex.12635] [Medline: 28984405]
29. Jolliff A, Coller RJ, Kearney H, Warner G, Feinstein JA, Chui MA, et al. An mHealth design to promote medication safety in children with medical complexity. Appl Clin Inform 2024 Jan;15(1):45-54 [FREE Full text] [doi: 10.1055/a-2214-8000] [Medline: 37989249]

XSL•FO
RenderX

30. Werner NE, Campbell NL, Boustani M, Ganci A, Holden RJ. Helping the helpers - a research protocol for user-centered technology to aid caregiver management of medications for people with Alzheimer's disease and related dementias. Res Social Adm Pharm 2022 Sep;18(9):3680-3686 [FREE Full text] [doi: 10.1016/j.sapharm.2022.03.014] [Medline: 35396151]
31. Reddy A, Lester CA, Stone JA, Holden RJ, Phelan CH, Chui MA. Applying participatory design to a pharmacy system intervention. Res Social Adm Pharm 2019 Nov;15(11):1358-1367 [FREE Full text] [doi: 10.1016/j.sapharm.2018.11.012] [Medline: 30509852]
32. Qualtrics XM. URL: https://www.qualtrics.com [accessed 2024-11-18]
33. Tracy SJ. Qualitative quality: eight “big-tent” criteria for excellent qualitative research. Qual Inq 2010 Oct 01;16(10):837-851 [FREE Full text] [doi: 10.1177/1077800410383121]
34. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023 Oct 25;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
35. Yalom ID, Leszcz M. Theory and Practice of Group Psychotherapy Fifth Edition. New York, NY: Basic Books; 2005.
36. Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. Med Educ 2017 Jan;51(1):40-50 [FREE Full text] [doi: 10.1111/medu.13124] [Medline: 27981658]
37. Richardson L, St. Pierre EA. Writing: a method of inquiry. In: Denzin NK, Lincoln YS, editors. The Sage Handbook of Qualitative Research. Thousand Oaks, CA: SAGE Publications; 2005.
38. Tamí-Maury I, Brown L, Lapham H, Chang S. Community-based participatory research through virtual communities. J Commun Healthc 2017;10(3):188-194 [FREE Full text] [doi: 10.1080/17538068.2017.1337604] [Medline: 30386424]
39. Black KZ, Faustin YF. How community-based participatory research can thrive in virtual spaces: connecting through photovoice. Human Org 2022 Aug 22;81(3):240-247 [FREE Full text] [doi: 10.17730/1938-3525-81.3.240]
40. Farmer N, Tuason RT, Kazmi N, Flynn S, Mitchell V, Middleton K, et al. Going virtual during the COVID-19 pandemic: adaptation of a mixed-methods dietary behavior study within a community-based participatory research study of African-American adults at risk for cardiovascular disease. BMC Med Res Methodol 2022 Dec 22;22(1):330 [FREE Full text] [doi: 10.1186/s12874-022-01806-3] [Medline: 36550396]
41. LeGreco M, Leonard D, Ferrier M. Virtual vines: using participatory methods to connect virtual work with community-based practice. In: Virtual Work and Human Interaction Research. Hershey, PA: IGI Global; 2012.
42. Viswanathan M, Ammerman A, Eng E, Garlehner G, Lohr KN, Griffith D, et al. Community-based participatory research: assessing the evidence. Evid Rep Technol Assess (Summ) 2004 Aug(99):1-8. [Medline: 15460504]
43. Sagen JS, Smedslund G, Simonsen AE, Habberstad A, Kjeken I, Dagfinrud H, et al. Patient engagement in the development and delivery of healthcare services: a systematic scoping review. BMJ Open Qual 2023 Jun;12(2):e002309 [FREE Full text] [doi: 10.1136/bmjoq-2023-002309] [Medline: 37369560]
44. Valdez RS, Holden RJ. Health care human factors/ergonomics fieldwork in home and community settings. Ergon Des 2016 Oct;24(4):4-9 [FREE Full text] [doi: 10.1177/1064804615622111] [Medline: 28781512]
45. Bowen S, McSeveny K, Lockley E, Wolstenholme D, Cobb M, Dearden A. How was it for you? Experiences of participatory design in the UK health service. CoDesign 2013 Dec;9(4):230-246 [FREE Full text] [doi: 10.1080/15710882.2013.846384]
46. Porche MV, Folk JB, Tolou-Shams M, Fortuna LR. Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. Front Psychiatry 2022;13:867460 [FREE Full text] [doi: 10.3389/fpsyt.2022.867460] [Medline: 35530032]
47. Wallerstein N, Oetzel J, Duran B, Tafoya G, Belone L, Rae R. What predicts outcomes in CBPR? In: Minkler M, Wallerstein N, editors. Community Based Participatory Research For Health: Process to Outcomes Second Edition. San Francisco, CA: Jossey-Bass Publishers; 2008:371-392.
48. Freeman J. The tyranny of structurelessness. Berkeley J Sociol 1972:151-164 [FREE Full text]
49. Babatunde S, Ahmed S, Santana MJ, Nielssen I, Zelinsky S, Ambasta A. Working together in health research: a mixed-methods patient engagement evaluation. Res Involv Engagem 2023 Aug 01;9(1):62 [FREE Full text] [doi: 10.1186/s40900-023-00475-w] [Medline: 37528438]
50. Examining Community-Institutional Partnerships for Prevention Research Group. Building and sustaining community-institutional partnerships for prevention research: findings from a national collaborative. J Urban Health 2006 Nov;83(6):989-1003 [FREE Full text] [doi: 10.1007/s11524-006-9113-y] [Medline: 17082993]
51. Manalili K, Siad F, Antonio M, Lashewicz B, Santana M. Codesigning person-centred quality indicators with diverse communities: a qualitative patient engagement study. Health Expect 2022 Oct;25(5):2188-2202 [FREE Full text] [doi: 10.1111/hex.13388] [Medline: 34854190]
52. Orenstein GA, Lewis L. Erikson's stages of psychosocial development. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2022.
53. Merker VL, Hyde JK, Herbst A, Solch AK, Mohr DC, Gaj L, et al. Evaluating the impacts of patient engagement on health services research teams: lessons from the veteran consulting network. J Gen Intern Med 2022 Apr;37(Suppl 1):33-41 [FREE Full text] [doi: 10.1007/s11606-021-06987-z] [Medline: 35349028]

XSL•FO
RenderX

54. Caldwell WB, Reyes AG, Rowe Z, Weinert J, Israel BA. Community partner perspectives on benefits, challenges, facilitating factors, and lessons learned from community-based participatory research partnerships in Detroit. Prog Community Health Partnersh 2015;9(2):299-311 [FREE Full text] [doi: 10.1353/cpr.2015.0031] [Medline: 26412771]
55. Sheehan L, Ballentine S, Washington L, Canser M, Connor J, Jones R, et al. Implementing community-based participatory research among African Americans with serious and persistent mental illness: a qualitative study. Gateways 2021;14(1) [FREE Full text] [doi: 10.5130/ijcre.v14i1.6894]
56. Muhammad M, Wallerstein N, Sussman AL, Avila M, Belone L, Duran B. Reflections on researcher identity and power: the impact of positionality on community based participatory research (CBPR) processes and outcomes. Crit Sociol (Eugene) 2015 Nov;41(7-8):1045-1063 [FREE Full text] [doi: 10.1177/0896920513516025] [Medline: 27429512]
57. Valdez RS, Lyon SE, Wellbeloved-Stone C, Collins M, Rogers CC, Cantin-Garside KD, et al. Engaging the disability community in informatics research: rationales and practical steps. J Am Med Inform Assoc 2022 Oct 07;29(11):1989-1995 [FREE Full text] [doi: 10.1093/jamia/ocac136] [Medline: 35972753]

Abbreviations

ADRD: Alzheimer disease and related dementias
CBPR: community-based participatory research
PCORI: Patient-Centered Outcomes Research Institute
PD: participatory design

Edited by A Coristine; submitted 14.05.24; peer-reviewed by KJ Lee, V Astha, S Anderson; comments to author 30.09.24; revised version received 20.10.24; accepted 21.10.24; published 03.12.24.

Please cite as:
Jolliff A, Holden RJ, Valdez R, Coller RJ, Patel H, Zuraw M, Linden A, Ganci A, Elliott C, Werner NE
Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers
J Med Internet Res 2024;26:e60353
URL: https://www.jmir.org/2024/1/e60353
doi:10.2196/60353
PMID:

©Anna Jolliff, Richard J Holden, Rupa Valdez, Ryan J Coller, Himalaya Patel, Matthew Zuraw, Anna Linden, Aaron Ganci, Christian Elliott, Nicole E Werner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Technology-Based Interventions for Substance Use Treatment Among People Who Identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native: Scoping Review

Sarah K Moore[1], PhD; Jesse S Boggis[1,2], MPH; Phoebe R Gauthier[1], MPH; Chantal A Lambert-Harris[1], MA; Emily G Hichborn[1], MA; Kathleen D Bell[1], MS; Elizabeth C Saunders[1], PhD; LaTrice Montgomery[3], PhD; Eilis I Murphy[1], BA; Avery M Turner[1], MS; Nico Agosti[1], BS; Bethany M McLeman[1], BA; Lisa A Marsch[1], PhD

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[2]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[3]College of Medicine, University of Cincinnati, Cincinatti, OH, United States

Corresponding Author:
Sarah K Moore, PhD
Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College
46 Centerra Parkway
EverGreen Center, Suite 315
Lebanon, NH, 03766
United States
Phone: 1 19176789280
Email: sarah.k.moore@dartmouth.edu

Abstract

Background: In the United States, racial and ethnic disparities in substance use treatment outcomes are persistent, especially among underrepresented minority (URM) populations. Technology-based interventions (TBIs) for substance use treatment show promise in reducing barriers to evidence-based treatment, yet no studies have described how TBIs may impact racial or ethnic health equity.

Objective: This study explored whether TBIs in substance use treatment research promote health equity among people who identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native through their inclusion in research. We explored whether research that includes the aforementioned groups consciously considers race and/or ethnicity beyond including these populations as participants.

Methods: We conducted a scoping review of 5 electronic databases to identify TBIs in substance use treatment studies published in English between January 2000 and March 2021. Studies were included if ≥50% of participants identified as African American or Black, Hispanic or Latino, or American Indian or Alaska Native *when combined*. Included studies were evaluated for conscious consideration of race and ethnicity in at least one manuscript section. Finally, we conducted a critical appraisal of each study's potential to facilitate insights into the impact of a TBI for members of specific URM groups.

Results: Of 6897 titles and abstracts screened and 1158 full-text articles assessed for eligibility, nearly half (532/1158, 45.9%) of the full-text articles were excluded due to the absence of data on race, ethnicity, or not meeting the aforementioned demographic eligibility criteria. Overall, 110 studies met the inclusion criteria. Study designs included 39.1% (43/110) randomized trials, and 35.5% (39/110) feasibility studies. In total, 47.3% (52/110) of studies used computer-based interventions, including electronic screening, brief interventions, and referrals to treatment, whereas 33.6% (37/110) used interactive voice response, ecological momentary assessment or interventions, or SMS text messaging via mobile phones. Studies focused on the following substances: alcohol or drugs (45/110, 40.9%), alcohol alone (26/110, 23.6%), opioids (8/110, 7.3%), cannabis (6/110, 5.5%), cocaine (4/110, 3.6%), and methamphetamines (3/110, 2.7%). Of the studies that consciously considered race or ethnicity (29/110, 26.4%), 6.4% (7/110) explicitly considered race or ethnicity in all manuscript sections. Overall, 28.2% (31/110) of the studies were critically appraised as having a high confidence in the interpretability of the findings for specific URM groups.

Conclusions: While the prevalence of TBIs in substance use treatment has increased recently, studies that include and consciously consider URM groups are rare, especially for American Indian or Alaska Native and Hispanic or Latino groups. This review

XSL•FO
RenderX

highlights the limited research on TBIs in substance use treatment that promotes racial and ethnic health equity and provides context, insights, and direction for researchers working to develop and evaluate digital technology substance use interventions while promoting health equity.

(J Med Internet Res 2024;26:e53685) doi:10.2196/53685

KEYWORDS

technology-based interventions; substance use; treatment; health equity; underrepresented; scoping review; mobile phone

Introduction

Background

In the United States, underrepresented minority (URM) groups, including African American or Black, Hispanic or Latino, and American Indian or Alaska Native populations, have experienced more barriers to substance use treatment engagement than their White counterparts [1]. These URM groups comprised only 15% of admissions to publicly funded substance use treatment programs in 2020, whereas White populations comprised 77% of admissions [2]. While the treatment gap is large for all, these gaps are greater for URM groups [3]. Among African American or Black and Hispanic or Latino populations diagnosed with a substance use disorder, 93% and 95%, respectively, did not receive treatment in 2021 [4]. Of the few URM groups who do access treatment, African American or Black and Hispanic or Latino populations are the largest groups and are less likely to complete treatment compared to White individuals [5-8].

American Indian or Alaska Native populations are overall at an increased risk of substance use disorder, particularly for alcohol use disorder [9]. These populations have historically been associated with excessive alcohol consumption, and yet a recent study using national survey data to compare alcohol use between American Indian or Alaska Native and non-Hispanic White individuals found limited differences for race and ethnicity, but the odds of binge drinking in male individuals were over twice those of their female counterparts [10,11]. It is notable that the number of American Indian or Alaska Native populations has consistently been too low to assess the treatment gap for those with a substance use disorder [4]. This statistical problem is likely due to the common research analytic strategy of collapsing demographic data into “more than one race,” thereby losing the potential insights into differences among those in this general category. Awareness of environmental risk factors (eg, for American Indian or Alaska Native populations, related to forced assimilation, such as living in geographically isolated communities without economic opportunities and the cross-generational transmission of the cumulative emotional and psychological trauma of cultural genocide) are essential to both prevention and treatment of substance use for all 3 URM groups that are the focus of this scoping review [10].

Disparities in treatment outcomes remain for these URM groups due in part to differences in social determinants and systemic racism as risk factors [1,12-14]. These differences are preventable as they are likely largely due to structural disadvantages that result in limited opportunities in health care, education, social context, economic stability, and the environment (ie, social determinants of health). To comprehensively address health equity—a state in which all people have a fair opportunity to attain their full health potential and well-being and no one is disadvantaged from doing so because of social position or any other socially defined circumstance [15]—avoidable inequalities and historical and contemporary injustices must be addressed to eliminate health care disparities.

Treatment intervention research, including substance use treatment research, has infrequently been conducted with people who identify as belonging to racial, ethnic, or both minority groups [16]. Indeed, a recent systematic review found that only 9 randomized controlled trials examined differences in substance use treatment outcomes by race or ethnicity for African American or Black and Hispanic or Latino populations, with even fewer reporting on differences in the baseline social determinants of health [17]. Thus, the state of treatment intervention research—particularly substance use treatment research—is a glaring example of limitations of opportunity for URM groups with respect to health care.

Advances in technology-based interventions (TBIs) show promise in reducing barriers and expanding access to evidence-based treatments [18]. However, there is a risk that TBIs may further exacerbate inequities by disproportionately benefiting people from higher socioeconomic status (SES) groups, who are largely White individuals [19]. Apprehension of this risk is warranted; throughout history in the United States, these URM groups have experienced violence and displacement due to settler colonialism (eg, for African American or Black individuals, capture from Africa and forced slavery and the US government’s social construction of Blackness as innately diseased due to presumed biological inferiority [15]; for Hispanic or Latino individuals, pre- and postmigration factors increasing the risk of intergenerational trauma, precarity due to legal status, discrimination, and anti-immigrant policies [20]), leading to the unequal distribution of resources and an overall lower SES [21,22]. However, in recent years, TBIs are increasingly being culturally adapted with the aim of increasing intervention effectiveness for specific URM groups [23-25]. In a systematic review of general technology use by URM groups, Montague and Perchonok [26] found that, while technology can be used to positively impact these communities, the technology must be culturally tailored to the specific URM group community to create positive behavior change. However, within substance use treatment, there has been limited research as to which specific characteristics of TBIs are most effective for URM groups [16].

XSL•FO
RenderX

Objectives

This scoping review aimed to answer one primary question: does the use of TBIs in substance use treatment research promote health equity among people who identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native individuals? We operationally defined health equity as the inclusion of members of underrepresented groups (previously specified) in research on TBIs for substance use, as well as the extent to which the research is race and ethnicity conscious. Thus, the following questions helped establish an effective search strategy and results that responded to the primary research question: does research on TBIs for substance use treatment include people who identify as African American or Black, Hispanic or Latino, or American Indian or Alaska Native individuals? If the substance use research on TBIs does include people who identify as African American or Black, Hispanic or Latino, or American Indian or Alaska Native individuals, is it race and ethnicity conscious?

Methods

Protocol

This scoping review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [27]. The PRISMA-ScR checklist can be found in Multimedia Appendix 1. A summary of the methods is provided in this section, with a published protocol publicly available [28].

Study Eligibility Criteria

Overview

The Joanna Briggs Institute—an organization that offers guidance for conducting systematic scoping reviews—recommends the use of the "PCC mnemonic" (population, concept, and context) for defining the focus of a review. The context for our review was substance use and disorder treatment research; the concept was TBIs; and the population were studies with samples that included at least 50% of participants who identified as African American or Black, Hispanic or Latino, or American Indian or Alaska Native individuals when combined [29].

Inclusion Criteria

The study inclusion criteria were US-based, English-language, peer-reviewed studies of TBIs for substance use treatment with a sample of at least 50% of participants identifying as African American or Black, Hispanic or Latino, or American Indian or Alaska Native individuals (combined) published between January 2000 and March 2021. The decision for the threshold of 50% was based on a review of evidence-based treatments for youth from ethnic minority groups. After the evidence-based criteria were met, the interventions were considered well established, probably efficacious, or possibly efficacious for youth from ethnic minority groups if supporting studies met one or more of several conditions, including whether at least 75% of participants in the study identified as being from ethnic minority groups [30]. By choosing a 50% threshold, we were more inclusive and likely netted more studies than we would have if we had chosen a 75% cutoff. We focused on this period due to a rising trend in the number of published studies starting in 2001 involving the design and development of technology-based behavior change interventions [31]. Study design criteria included qualitative, quantitative, and mixed methods studies.

Exclusion Criteria

We excluded the following manuscript types: protocol papers, descriptions of future work, reviews, commentaries, editorials, opinion pieces, student theses, conference abstracts, book chapters, and guidelines. Tobacco intervention studies were also excluded from this review. (see the work by Hichborn et al [32] for a separate scoping review on tobacco-focused TBI studies).

Data Sources and Search Strategy

Before enlisting the help of research librarians, study team members conducted preliminary independent literature searches in PubMed and Google Scholar using search terms associated with the 3 domains of interest—TBIs, substance use treatment, and sample inclusion of members of URM groups—to identify whether there was indeed a gap in the literature to be filled by this scoping review. This step netted dozens of research studies that appeared to meet the inclusion criteria, confirming the gap in the literature. Following these exploratory efforts, a comprehensive search strategy was implemented using 5 electronic databases (MEDLINE, Scopus, Cochrane Library, CINAHL, and PsycINFO) to identify research that included substance use and disorder treatment interventions using TBIs [29]. The literature searches were conducted by 2 Dartmouth College reference librarians in consultation with the research team. The final electronic search strategy for MEDLINE can be found in Multimedia Appendix 2.

Screening and Selection Procedure

Duplicate studies were removed in EndNote X9 (Clarivate Analytics), and the remaining studies were uploaded into the web-based review tool Rayyan (Rayyan Systems Inc) [33] for screening and selection. Studies were first screened at the title and abstract level and then selected for inclusion after a full-text review by a group of 8 reviewers (SKM, CAL-H, PRG, NA, ECS, EGH, KDB, and AMT) working in 3 small teams. Throughout the independent selection process, individual reviewers were blinded to each other's decisions. The reviewers met regularly to discuss discrepancies and reach a consensus both within teams and as a full group.

Data Extraction

The review team developed a standardized data extraction form in Microsoft Excel (Microsoft Corp) with predefined variables (Textbox 1). A standard operating procedure provided general extraction guidelines, variable definitions, and expected location of variables within the manuscripts. Before data collection, the reviewers piloted the extraction form with 2 studies and reviewed for reliability across team members. At this stage 2 more team members were added, and a 9-person review team continued to work in 3 small groups, with each team member individually performing extraction for a sample of articles within the pool assigned to their team. All studies had a second team

XSL•FO
RenderX

member perform a quality assurance data validation check for accuracy, consistency, and thoroughness in the data extraction fields, mainly focusing on fields with the greater likelihood for inconsistency (eg, TBI).

Textbox 1. Data extraction variables.

Variable and description of variable

- Author: last name of first author
- Year: year when the study was published
- Aim or aims: purpose of the technology-based intervention (TBI)
- Design: the study design
- Participant population (for whom): description of the target population that the study intended to enroll
- End-user involvement (with whom): description of methods used to involve or engage end users in the development, design, or adaptation of TBIs and opportunities for end users to provide feedback on TBIs to inform acceptability, among other things (intervention-focused outcomes)
- N: number of participants in the study
- Sample summary: racial and ethnic profile of the study participants
- Recruitment plan: description of how candidates were informed of and introduced to the study
- Retention plan: description of any efforts to retain participants in the research study
- Primary substance use target: substance use or substance use disorder targeted by the TBIs
- Context: description of intervention delivery setting and other relevant details (eg, urban, rural, remote, or in person)
- TBI: description of TBI and name (if named)
- Device: description of the device used to deliver the TBI
- Delivery mode: description of how the TBI was delivered to participants, including the frequency and duration
- Intervention components: description of interventional components combined with the TBI (eg, counseling and contingency management)
- Comparators: description of other interventions that the TBI was compared with
- Behavior change theory: description of behavior change theory or theories underlying the TBI
- Behavior change techniques: description of behavior change techniques that were used by the TBI
- Race and ethnicity consciousness: yes or no—specific race or ethnicity references in one or more sections of the manuscript (ie, introduction, methods, results, or discussion; see the Race and Ethnicity Consciousness section)

Data Synthesis and Analysis

Organization

We organized netted studies by programs of research, which were defined as ≥2 studies that involved coauthorship networks and referenced the same TBI. Studies were classified as race or ethnicity conscious (yes or no) if the article specifically described research practices informed by at least one URM group in at least one section of the article (ie, title, introduction, method, results, or discussion; Textbox 2). Tables pertaining to race or ethnicity consciousness were populated with data relevant to the studies' specific reference to the 3 URM groups across manuscript sections. The remaining extracted data served to represent the range of the studies and the nature of TBIs in their respective tables, as well as access and inclusion variables.

XSL•FO
RenderX

Textbox 2. Examples of race- and ethnicity-conscious practices in research.

Section of the article and example

- Title: Tailoring an Alcohol Intervention for American Indian Alaska Native Women of Childbearing Age: Listening to the Community
- Introduction: providing epidemiological or other relevant information about one or more racial or ethnic groups in the literature review or using a theory that was described as one that may help address health disparities among racial and ethnic underrepresented groups. An example of this includes sociological trust theory—a bridge between a broad lens of culturally informed design and attention to trust or distrust.
- Methods: race-conscious methods may include references to recruitment or retention efforts aimed at racial or ethnic groups, such as cultural tailoring of materials, consideration of matching staff race and ethnicity to that of the sample participants (in telemedicine appointments and in the animations seen in virtual reality and computer games), or assessment for measurement equivalence. Race-conscious analytic plans may include conducting separate analyses for each race and ethnicity, focusing on within-group differences rather than race comparisons, using stratification methods that balance each race and ethnicity across treatment arms, or considering race and ethnicity in some other way in the plan for analyzing the data.
- Results: presenting findings in a way that highlights differences and similarities for members of different racial or ethnic groups.
- Discussion: interpreting findings for members of racial or ethnic groups by locating results in the context of other development or treatment literature.

Analysis

Included sources of evidence for scoping reviews are not critically appraised regarding methodological quality or risk of bias, making scoping reviews significantly different from systematic reviews. However, one of the PRISMA-ScR items includes an optional “critical appraisal of individual data sources” [27]. In this study, we aimed to critically appraise the research in terms of each study’s *potential* to help us better understand the impact of a TBI for members of URM groups. Our rationale for the critical appraisal was founded on the notion that promoting health equity includes highlighting all netted research and not solely focusing on the TBIs that have been found to be effective. To conduct the critical appraisal, we organized the studies by levels of inclusion of a single URM and race or ethnicity consciousness. A higher appraisal or confidence was assigned to studies explicit in presenting results clarifying study implications with samples that included ≥90% of members of a single URM group or included race-conscious results such that the findings were interpretable due to subgroup analyses based on race or ethnicity (eg, 30% African American or Black participants). A lower appraisal or confidence was assigned to those studies that included <90% of members of a single URM group *and* lacked race-conscious results.

The greater the representation of a single URM group in a sample, the more likely the findings were to have direct implications for that group regardless of whether the results were race conscious.

Results

Sources of Evidence

The systematic searches identified 6897 references (Figure 1 [27]). Title and abstract screening excluded 83.21% (5739/6897) of the references, and a further 90.5% (1048/1158) of the papers were excluded at full-text review, resulting in 110 studies meeting the inclusion criteria for this scoping review. Full-text review excluded 50.67% (531/1048) of the studies due to ineligibility based on demographic criteria or the absence of race or ethnicity data. Most of the studies (81/110, 73.6%) were published from 2014 onward.

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram [27]. SU/D: substance use or disorder; TBI: technology-based intervention; URM: underrepresented minority.

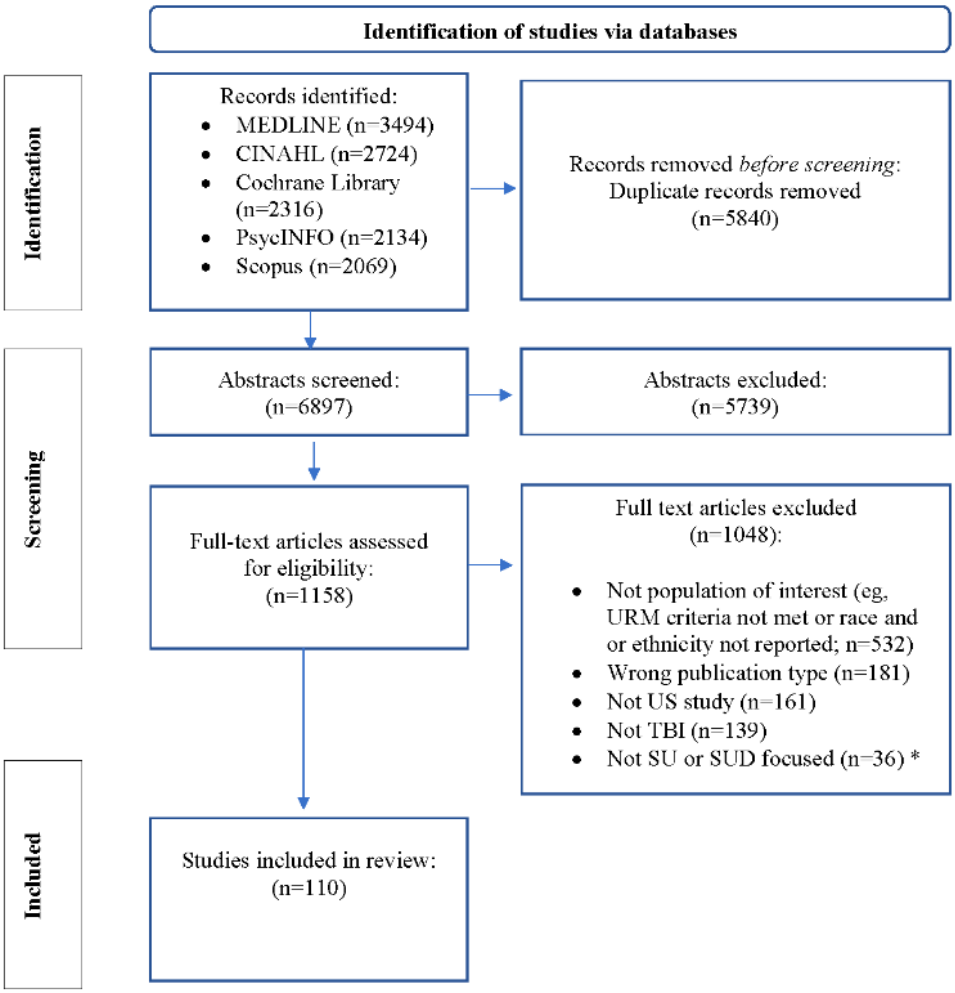

Scoping Review Dataset

The dataset was organized by programs of research (76/110, 69.1% of the studies) and independent studies, defined as studies without companion publications in the dataset (34/110, 30.9%; Table 1).

XSL•FO
RenderX

Table 1. Programs of research and independent studies (n=110).

Programs of research	Studies, n (%)
CEMA[a]	2 (1.8)
MOMENT[b]	2 (1.8)
e-SBI[c] DrinkWise	2 (1.8)
TEXT[d]	2 (1.8)
Web-based SBIRT[e]	3 (2.7)
Web-based MI[f]	3 (2.7)
Tailored SMS text messaging	4 (3.6)
eCheckup	3 (2.7)
TES[g]	3 (2.7)
Theory-based SMS text messaging	3 (2.7)
EMA[h]	4 (3.6)
A-CHESS[i]	5 (4.5)
SaferTeens	6 (5.4)
Health Call	7 (6.3)
CBT4CBT[j]	10 (9)
MES[k] or SBIRT	17 (15.5)
Independent studies	34 (30.9)

[a]CEMA: cellular ecological momentary assessment.
[b]MOMENT: Momentary Self-Monitoring and Feedback + Motivational Enhancement Therapy.
[c]e-SBI: electronic screening and brief intervention.
[d]TEXT: treatment extension by text.
[e]SBIRT: Screening, Brief Intervention and Referral to Treatment.
[f]MI: motivational interviewing.
[g]TES: Therapeutic Education System.
[h]EMA: ecological momentary assessment.
[i]A-CHESS: Addiction Comprehensive Health Enhancement Support System.
[j]CBT4CBT: Computer Based Training for Cognitive Behavioral Therapy.
[k]MES: Motivational Enhancement System.

Study Characteristics

Study designs included 39.1% (43/110) randomized trials, 36.4% (40/110) feasibility studies, 7.3% (8/110) development studies, 6.4% (7/110) secondary analyses, 5.5% (6/110) assessments (ie, interactive voice response, ecological momentary assessment [EMA] solely, with TBI development planned), and 5.5% (6/110) miscellaneous study types. Of the eligible studies, 31.8% (35/110) examined TBIs combined with or integrated into treatment as usual (TAU), whereas 24.5% (27/110) examined TBIs combined with treatment or treatments that were novel to the existing treatment program (eg, motivational interviewing, contingency management [CM], and supportive messages). Of the 27 studies examining TBIs combined with novel treatments, 6 (22%) combined more than one novel treatment with the TBI (eg, naloxone training, HIV, and hepatitis C virus testing; motivational interviewing or transtheoretical model; reading scripture followed by song or relation breathing exercises; or cognitive behavioral therapy [CBT] or tai chi). Of the 110 included studies, 59 (53.6%) reported the use of a comparator to the TBI. Comparators included other TBIs (31/59, 53%); in-person, clinician-delivered versions of the TBI (10/59, 17%) and briefer versions of the targeted TBI (6/59, 10%); TAU (16/59, 27%); nonactive interventions (eg, reading a pamphlet; excluding TAU; 30/59, 51%); motivational interviewing (7/59, 12%); CBT or counseling (4/59, 7%); SMS text messaging or EMA (5/59, 8%); and others (16/59, 27%; eg, DVDs or videos, brochures and website resources, and assessment-only groups).

Beyond the sources of evidence, resultant scoping review dataset and study characteristics typically reported in scoping reviews, the following result subsections (ie, Access or Inclusion, Range and Nature of TBIs, Race and Ethnicity Consciousness, and the critical appraisal of the dataset) all derive from the primary research question and related subquestions developed based on the operationalization of health equity promotion.

XSL•FO
RenderX

Access and Inclusion

Racial or Ethnic Sample Profiles (Analyzed Data)

In 78.2% (86/110) of the included studies, at least 50% of the participants were members of a single URM group: American Indian or Alaska Native (4/86, 5% of the studies) [34-37], Hispanic or Latino (15/86, 17% of the studies) [38-52], and African American or Black (67/86, 78% of the studies; Tables 2 and 3). While most of the studies on African American or Black individuals also included people who identified as Hispanic or Latino and vice versa, only 12.3% (13/106) of the studies with a majority of African American or Black or Hispanic or Latino individuals included people who identified as American Indian or Alaska Native individuals [46,47,49,53-62].

XSL•FO
RenderX

Table 2. Programs of research.

TBI[a] program of research, study, and year	Substance use target	Inclusion
Motivation Enhancement System[b]		
Ondersma et al [63], 2005	Drug use	97% African American or Black[c]
Ondersma et al [64], 2007	Drug use	97% African American or Black[c]
Ondersma et al [65], 2011	Drug use	100% African American or Black
Ondersma et al [66], 2014	Drug use	90.6% African American or Black, 5.8% White, 3.6% other, and 1.5% Hispanic or Latino
Tzilos et al [67], 2011	Alcohol	82% African American or Black, 16% White, and 3% Hispanic or Latino
Ondersma et al [68], 2015	Alcohol	81.3% African American or Black[c]
Pollick et al [69], 2015	Alcohol	100% African American or Black
Ondersma et al [70], 2016	Alcohol	87% African American or Black
Martino et al [71], 2018	ATOD[d]	66.7% African American or Black, 13.2% White, 14.8% Hispanic or Latino, and 5.2% other or multiple races
Forray et al [72], 2019	ATOD	66.7% African American or Black, 13.2% White, 14.8% Hispanic or Latino, and 5.2% other or multiple races
Loree et al [73], 2019	ATOD	66.7% African American or Black, 13.2% White, 14.8% Hispanic or Latino, and 5.2% other or multiple races
Yonkers et al [74], 2020	ATOD	66.7% African American or Black, 13.2% White, 14.8% Hispanic or Latino, and 5.2% other or multiple races
Ondersma et al [75], 2018	Drug use	73.2% African American or Black, 24.2% other or multiple races, and 2.6% White
Braciszewski et al [53], 2018	AOD[e]	29% Hispanic or Latino, 24% African American or Black, 41% White, 18% American Indian or Alaska Native, 6% Hawaiian or Pacific Islander, 6% Asian or Pacific Islander, and 6% other
Braciszewski et al [76], 2018	AOD	52% African American or Black, 27% White, 18% multiple races, and 3% other
Tzilos Wernette et al [54], 2018	AOD	40% Hispanic or Latino, 33% African American or Black, 5% American Indian or Alaska Native, and 50% other or multiple races
Ondersma et al [77], 2019	Cannabis	84.4% African American or Black and 1.2% Hispanic or Latino[c]
CBT4CBT[f]		
Carroll et al [55], 2008	AOD	46% African American or Black, 34% White, 12.3% Hispanic or Latino, and 5.5% American Indian or Alaska Native
Sugarman et al [56], 2010	AOD	46% African American or Black, 34% White, 12.3% Hispanic or Latino, and 5.5% American Indian or Alaska Native
Kiluk et al [57], 2010	AOD	50% African American or Black, 36.5% White, 9.6% Hispanic or Latino, and 3.8% American Indian or Alaska Native
Kiluk et al [58], 2011	AOD	50% African American or Black, 36.5% White, 9.6% Hispanic or Latino, and 3.8% American Indian or Alaska Native
Kiluk et al [78], 2016	Alcohol	54.4% African American or Black, 11.8% Hispanic or Latino, 33.8% White, 7.4% Hispanic or Latino only, and 4.4% other or multiple races
Kiluk et al [79], 2018	Alcohol	54.4% African American or Black, 11.8% Hispanic or Latino, 33.8% White, 7.4% Hispanic or Latino only, and 4.4% other or multiple races
Kiluk et al [80], 2018	AOD	48.9% African American or Black, 34.3% White, 16.1% Hispanic or Latino (ethnicity), 8% Hispanic or Latino (only), and 8.7% other or multiple races
Paris et al [38], 2018	AOD	100% Hispanic or Latino
Roos et al [81], 2020	AOD	54.4% African American or Black, 11.8% Hispanic or Latino, 33.8% White, 7.4% Hispanic or Latino only, and 4.4% other or multiple races
Jordan et al [82], 2021	AOD	100% African American or Black
Health Call[g]		

XSL•FO
RenderX

TBI[a] program of research, study, and year	Substance use target	Inclusion
Aharonovich et al [39], 2006	Alcohol	51.6% Hispanic or Latino, 37.8% African American or Black, 6.5% White, and 3.2% other
Hasin et al [83], 2013	Alcohol	49% African American or Black, 45% Hispanic or Latino, and 6% other
Hasin et al [84], 2014	Alcohol	61.5% African American or Black, 25.6% Hispanic or Latino, and 12.8% other
Aharonovich et al [85], 2012	AOD	63.6% African American or Black, 21.2% Hispanic or Latino, and 6.1% White
Aharonovich et al [86], 2017	AOD	55% African American or Black, 25.8% Hispanic or Latino, and 19.17% other
Aharonovich et al [87], 2017	AOD	78.72% African American or Black and 17.02% Hispanic or Latino
Thompson et al [59], 2020	Alcohol and cannabis	65% African American or Black, 47.5% Hispanic or Latino, 15% American Indian or Alaska Native, 10% White, 5% Native Hawaiian or Pacific Islander, and 2.5% Asian
SaferTeen[h]		
Cunningham et al [88], 2009	Alcohol	55% African American or Black, 37.2% White, 6.2% Hispanic or Latino, and 7.8% other
Walton et al [89], 2010	Alcohol	55.9% African American or Black, 39.1% White, and 6.5% Hispanic or Latino
Walton et al [90], 2013	Cannabis and AOD	60.7% African American or Black and 11% Hispanic or Latino
Blow et al [91], 2017	AOD	52.2% African American or Black, 39.2% White, 8.6% other, and 6% Hispanic or Latino
Waller et al [92], 2019	Cannabis	54% African American or Black[c]
Drislane et al [93], 2020	Alcohol and cannabis	52.2% African American or Black, 39.2% White, 8.6% other, and 6% Hispanic or Latino
A-CHESS and CASA-CHESS[i]		
Dennis et al [94], 2015	AOD	48% African American or Black, 21% White, 7% Hispanic or Latino, and 24% multiple races
Scott et al [95], 2013	__j	85% African American or Black, 6% Hispanic or Latino, 6% White, and 3% other
Muroff et al [40], 2017	AOD	96.2% Hispanic or Latino and 3.8% African American or Black (Spanish speaking)
Muroff et al [41], 2019	AOD	96.2% Hispanic or Latino and 3.8% African American or Black (Spanish speaking)
Scott et al [96], 2020	AOD	70% African American or Black, 20% White, 6% Hispanic or Latino, and 4% other or multiple races
EMA or electronic diary[k]		
Preston et al [97], 2016	AOD	68% African American or Black and 43% White
Preston et al [98], 2017	Opioids and cocaine	64.8% African American or Black and 34% White
Preston et al [99], 2018	Opioids and cocaine	64.8% African American or Black and 34% White
Preston et al [100], 2018	Opioids and cocaine	64.8% African American or Black and 34% White
Theory-based SMS text messaging[l]		
Reback et al [101], 2012	Methamphetamines	38.5% Hispanic or Latino, 21.2% African American or Black, 34.6% White, and 5.8% other or multiple races
Reback et al [102], 2015	Methamphetamines	38.5% Hispanic or Latino, 21.2% African American or Black, 34.6% White, and 5.8% other or multiple races
Reback et al [103], 2019	Methamphetamines	43.7% African American or Black, 25.2% Hispanic or Latino, 19.6% White, and 11.5% other or multiple races
TES[m]		
Brooks et al [104], 2010	AOD	100% African American or Black
Campbell et al [34], 2015	AOD	100% American Indian or Alaska Native

XSL•FO
RenderX

TBI[a] program of research, study, and year	Substance use target	Inclusion
Budney et al [105], 2015	Cannabis	59% African American or Black and 41% White
Women, Infant, and Children eCheck-up[n]		
Delrahim-Howlett et al [106], 2011	Alcohol	44% Hispanic or Latino, 34% White, 8% African American or Black, 8%, other and 14% other or multiple races
Gorman et al [35], 2013	Alcohol	100% American Indian or Alaska Native
Montag et al [36], 2017	Alcohol	84.2% American Indian or Alaska Native[c]
Interactive Mobile Text Message Reminder System[o]		
Muench et al [60], 2013	AOD	50% African American or Black, 36.5% White, 9.6% Hispanic or Latino, and 3.8% American Indian or Alaska Native
Tofighi et al [107], 2016	Opioids	39% African American or Black, 36% White, 22% Hispanic or Latino, and 3% other
Tofighi et al [108], 2017	Opioids	40% African American or Black, 18% Hispanic or Latino, 41% White, and 1% Asian
Telemedicine- or telephone-based opioid treatment—COVID-19		
Tofighi et al [109], 2022	Opioids	34.6% Hispanic or Latino, 34.2% White (non–Hispanic or Latino), 19.2% African American or Black (non–Hispanic or Latino), and 15.4% other
Culturally relevant, web-based motivational interviewing[p]		
Osilla et al [42], 2012	Alcohol	100% Hispanic or Latino
Osilla et al [110], 2015	Alcohol	40% Hispanic or Latino, 39% African American or Black, 35% White, and 7% Asian or Pacific Islander
Kennedy et al [111], 2018	AOD	56% African American or Black, 12% White (non–Hispanic or Latino), 15% Hispanic or Latino, and 17% other or multiple races
Unnamed SBI[q]		
Dawson Rose et al [112], 2015	ATOD	42.7% African American or Black and 15.6% Hispanic or Latino
Dawson Rose et al [113], 2017	ATOD	40% African American or Black, 30% White, 17% Hispanic or Latino, and 13% other
Saberi et al [114], 2020	ATOD	35.6% Hispanic or Latino, 29.7% African American or Black, 13.9% White, 17.8% other or non–Hispanic or Latino
Video+ and Treatment Extension by Text[r]		
Ingersoll et al [61], 2011	Cocaine	81.5% African American or Black, 13% White, 3.7% other, 1.9% American Indian or Alaska Native, and 1.9% Hispanic or Latino
Ingersoll et al [115], 2015	Cocaine	65.1% African American or Black, 28.6% White, and 6.4% other
e-SBI[s] DrinkWise		
Nayak et al [43], 2014	Alcohol	62% Hispanic or Latino, 16% African American or Black, 17% other or multiple races, and 5% White
Nayak et al [44], 2019	Alcohol	75% Hispanic or Latino, 17% White, 5% African American or Black, and 3% other
MOMENT-MET+motivation and feedback[t]		
Shrier et al [116], 2014	Cannabis	44% African American or Black and 37% Hispanic or Latino[c]
Shrier et al [117], 2018	Cannabis	47% African American or Black (non–Hispanic or Latino), 31% Hispanic or Latino, 11.4% White (non–Hispanic or Latino), and 10% multiple races or other (non−Hispanic or Latino)
Cellular ecological momentary assessment[u]		
Comulada et al [45], 2015	AOD	93% Hispanic or Latino, 3.5% Asian or Pacific Islander, and 3.5% other

XSL•FO
RenderX

TBI[a] program of research, study, and year	Substance use target	Inclusion
Swendeman et al [118], 2021	AOD	55% African American or Black, 30% Hispanic or Latino, 5% White, and 5% Asian or Pacific Islander

[a]TBI: technology-based intervention.

[b]Motivation Enhancement System (electronic screening, brief intervention, and referral to treatment [eSBIRT]): a single-session eSBIRT. Assessment—questions presented one at a time using relevant graphics and a 3D cartoon character (program narrator and guide) capable of ≥50 specific animated actions (eg, smiling and expressing concern). Headphone use assures privacy. Intervention—three components based on motivational interviewing (MI) and brief intervention principles: (1) feedback regarding the negative consequences of reported drug use, self-reported readiness to change, and normative feedback (eg, drug use compared to drug use of all American adult women); (2) pros and cons of drug use and related change (participant chooses from a list of options); and (3) a summary and query regarding the participant's interest in change, followed by optional goal setting regarding drug use. Throughout the intervention, the narrator reflects back the participant's answers and helps establish an atmosphere that is as similar as possible to an in-person MI session.

[c]No further demographic information included on the racial and ethnic makeup of the sample.

[d]ATOD: alcohol, tobacco, or drug use.

[e]AOD: alcohol or drug use.

[f]CBT4CBT: Computer-Based Training for Cognitive Behavioral Therapy (computer-based intervention [CBI]). A computer-based version of cognitive behavioral therapy used in conjunction with standard clinical care for current substance users. Six modules and follow-up assignments focus on key concepts in substance use: (1) understanding and changing patterns of substance use, (2) coping with craving, (3) refusing offers of drugs and alcohol, (4) problem-solving skills, (5) identifying and changing thoughts about drugs and alcohol, and (6) improving decision-making skills. The first module provided a brief explanation of how to use and navigate the program; following completion of the first module, participants could choose to access the modules in any order that they preferred and repeat any section or module as many times as they wished. The multimedia presentation (graphic illustrations, videotaped examples, verbal instructions and audio voice-overs, interactive assessments, and practice exercises), which are based on elementary-level computer learning games, requires no previous experience with computers.

[g]Health Call (interactive voice response [IVR] and ecological momentary assessment intervention or SMS text messaging): automated daily telephone self-monitoring system for drinking based on IVR. Generically, IVR is a telephone-based procedure that allows individuals to interact with recorded questions and statements.

[h]SaferTeen (CBI): interactive multimedia computer program viewed on tablet laptops, with touch screens and audio delivered through headphones to ensure privacy. It was narrated and cartoon style, and participants could choose a gender-, race-, and age-appropriate "buddy" to "hang out" with throughout the session. The buddy guides participants through the intervention, including review of tailored feedback based on survey responses, identifying reasons to stay away from drinking and fighting, and role-play scenarios chosen by the computer based on reported risk behaviors. During the scenarios, participants had to interact with peers and make behavioral choices. Feedback was provided on these behavioral choices by the buddy, with possible consequences highlighted and the best possible outcome demonstrated by the characters.

[i]A-CHESS and CASA-CHESS: Addiction Comprehensive Health Enhancement Support System and its translation or adaptation for Spanish-speaking Latinos (IVR or ecological momentary assessment intervention or SMS text messaging). Ecological momentary assessments and recovery support ecological momentary interventions (EMIs) via smartphones. A-CHESS recorded the date, time, location, and responses to all ecological momentary assessments. The EMIs that adolescents could access at any time included (1) recovery support (discussion groups, support team, reaching out to others via SMS text message, listening to recovery stories, web-based self-help meetings, in-person self-help meeting locator, and linking to a sponsor), (2) relaxation (guided relaxation, playing games, listening to music, learning or reading, and physical exercise), (3) recovery motivation (motivational SMS text messaging, recovery words, recovery profiles, and pictures), and (4) social networking (Facebook and contacting friends).

[j]Not applicable.

[k]EMA: ecological momentary assessment. IVR or EMA intervention or SMS text messaging. Transactional electronic diary (EMA on PDA). Smartphones programmed with electronic diary software to prompt participants 3 times per day to answer a series of questions. The timing of the prompts was random but constrained by the participant's self-reported typical waking hours for each day of the week. Participants also initiated entries whenever they experienced a stressful event (SE). Participants initiated an SE entry any time they felt "more stressed, overwhelmed, or anxious than usual" and checked whichever applied when they initiated an entry (multiple responses were possible). Participants categorized the severity of each SE, rated the associated feeling, and indicated the cause of the feeling by selecting from a list (a "fill in the blank" option was also available). In each of these SE entries, participants rated cravings for opioids and cocaine and reported whom they were with, what they were doing, and the type of EMA entry (SE entry vs randomly prompted entry; 2017-2018).

[l]Theory-based SMS text messaging (IVR or EMA intervention or SMS text messaging): real-time SMS text messaging intervention (400 SMS text messages). Prewritten SMS text messages were categorized by their theoretical base and by participant profile (eg, HIV status or whether a participant went online to "hook up"). On the basis of their responses to the 5 brief questions regarding their methamphetamine use administered at baseline as well as their SMS text messaging conversations, participants received messages that fit their profile.

[m]TES: Therapeutic Education Systems (CBI). Interactive, web-based program theoretically grounded in the evidence-based Community Reinforcement Approach to behavior therapy. It comprised interactive multimedia modules, including those focused on cognitive behavioral skill training (eg, effective strategies for refusing drugs and managing thoughts about drug use). TES also included modules to prevent HIV, hepatitis, and sexually transmitted infections. Additional modules taught skills to improve psychosocial functioning (eg, family or social relationships and managing negative moods). TES is a self-directed program that includes a module teaching patients how to use the system and a "customization program" to build an individualized treatment plan for patients.

[n]Women, Infant, and Children (WIC) eCheck-up (eSBIRT): adapted from eCHECKUP TO GO, a brief assessment with motivational feedback tailored to college students. Tailored reading or comprehension level for WIC clients; visual graphics modified for this population. Changes to measurement components incorporated established and validated methods of assessing for alcohol use in women of childbearing potential and feedback tailored to include either general information about fetal alcohol syndrome or personalized information about participants' alcohol use, associated health risks, and health risks associated with alcohol use during pregnancy.

XSL•FO
RenderX

[o]Interactive Mobile Text Message Reminder System (IVR or EMA intervention or SMS text messaging): the SMS text message reminder software notified patients 7, 4, and 1 days before their scheduled office based opioid treatment program appointments. Following enrollment, 1 message was sent to inform participants about terminating SMS text message reminders. Patient privacy was ensured by excluding any patient health information and clinic identifiers, such as "buprenorphine clinic," within the SMS text message content.

[p]Culturally relevant, web-based MI (CBI): web-based MI with feedback tailored to baseline survey responses about drinking behavior and perceptions. Baseline survey responses populated the web-based MI content: (1) how their drinking and guesses of others' drinking compared to those of other men or women their age in the United States, (2) their positive beliefs about drinking, (3) their negative consequences from drinking, and (4) strategies for avoiding consequences in the future.

[q]Unnamed screening and brief intervention (eSBIRT): self-administered SBI for substance use embedded in a web-based personal health record linked to the clinic's electronic medical record (in English and Spanish) accessible via the internet. The SBI was tailored to patients' substance dependence risk scores on the Alcohol, Smoking, and Substance Involvement Screening Test. No- or low-risk participants received positive feedback and behavior maintenance support. Moderate-risk participants received a brief interactive web-based intervention and links to substance use websites and patient resources. High-risk participants were referred to treatment (Saberi et al [114], development for ART adherence among youth living with HIV videoconferencing [video]).

[r]Video+ (video): 4 accurate videos with at least some personal narrative in addition to didactic information presented by diverse peer role models and medical experts. The videos addressed either crack cocaine use or drug use generally, HIV treatment, or both (only the study by Ingersoll et al [61]); Treatment Extension by Text (IVR or EMA or SMS text messaging): bidirectional SMS text messaging system using EMA that sends messages to and receives and interprets messages from participants, enabling the system to send an appropriate intervention response. The automated system sent daily queries of medication dosing, mood, and substance use (only the study by Ingersoll et al [115]; platform changes within program of research).

[s]e-SBI: electronic screening and brief intervention.

[t]MOMENT-MET+motivation and feedback (IVR or EMA intervention or SMS text messaging): brief motivational therapy administered by a counselor in the clinic with mobile self-monitoring and responsive messaging via an electronic mobile device to extend treatment into the natural environment and everyday life.

[u]Cellular ecological momentary assessment (IVR or EMA intervention or SMS text messaging): SMS text messaging–based EMA focused on (1) alcohol or drug use, (2) quantity used, (3) antecedents, (4) intensity of craving, (5) engagement in risky behaviors, and (6) research participants' activities and location.

XSL•FO
RenderX

Table 3. Independent studies.

Platform and study	Substance use targets	Inclusion	TBI[a]
Video, videoconferencing, or telepsychiatry			
Millery et al [46], 2002	Opioids and cocaine	58% Hispanic or Latino, 33% African American or Black, 5% White, 3% Asian, American Indian or Alaska Native, and 1% other	Video resource intervention
Frueh et al [119], 2005	Alcohol	83% African American or Black and 17% White	Telepsychiatry or videoconference
Wechsberg et al [120], 2011	AOD[b]	100% African American or Black	Pregnant women's WHC[c] intervention; Microsoft PowerPoint (Microsoft Corp) presentation via video
Brusoski and Rosen [121], 2015	Opioids	100% African American or Black	Healthy Living Intervention—video chat
Welsh [122], 2016	Opioids	55% African American or Black and 45% White	Cellphone or computer pictures or video for pill counts for OUD[d]
Jaconis et al [123], 2017	AOD	100% African American or Black	Telehealth or videoconference—"COPE"[e]
Palfai et al [124], 2019	Alcohol	80% African American or Black and 20% Hispanic or Latino	Development study plan to use videoconferencing
Castillo et al [125], 2020	Opioids	38% Hispanic or Latino, 43% White, and 19% African American or Black (non–Hispanic or Latino)	Tele-MOUD[f]
Legha et al [37], 2020	AOD	100% American Indian or Alaska Native	Telepsychiatry or video
Mehtani et al [126], 2021	Opioids	67% African American or Black, 8% Hispanic or Latino, and 33% White	ATP[g]
O'Gurek [127], 2021	Opioids	40% African American or Black, 36% White, 22% Hispanic or Latino, and 1.9% other	Telehealth or video—COVID-19
VR[h] or avatar			
Saladin et al [128], 2006	Cocaine ("crack")	91% African American or Black	VR environment—cocaine ("crack house" simulation)
Gordon et al [129], 2017	AOD	59.3% African American or Black, 39% White, and 1.7% other	AAT[i]
EMA[j]			
Freedman et al [130], 2006	Cocaine ("crack")	93% African American or Black and 7% White	EMA via cell phone with homeless adults addicted to crack cocaine in an IOP[k]
Linas et al [131], 2015	Opioids and cocaine	90% African American or Black[l]	Unnamed EMA for drug use
Yang et al [132], 2015	Alcohol	100% African American or Black	EMA for AA[l] men who have sex with men
Sanjuan et al [47], 2019	AOD	66.7% Hispanic or Latino, 78.8% White, 21.2% American Indian or Alaska Native, and 3% African American or Black	EMA for treatment-engaged pregnant women with SUDs[m]
SMS text messaging			
Trudeau et al [48], 2012	AOD	62.5% Hispanic or Latino (White or African American or Black), 29.2% White, and 8.3% African American or Black (non–Hispanic or Latino)	Web-based relapse prevention for adolescents (SMS text messaging)
Buu et al [133], 2017	AOD	60% African American or Black and 20% White	IVR[n] or SMS text messaging—daily or weekly
Suffoletto et al [134], 2018	Alcohol	54% African American or Black and 46% White	TRAC2[o]

XSL•FO
RenderX

Platform and study	Substance use targets	Inclusion	TBI[a]
Tolou-Shams et al [49], 2019	AOD	60% Hispanic or Latino, 33.33% African American or Black, 7% American Indian or Alaska Native, 13.33% Asian, 13.33% multiple races, and 27% other	SMS text message development for youth on probation and their caregivers
Moore et al [135], 2019	AOD	53% African American or Black (non–Hispanic or Latino), 31% Hispanic or Latino, and 16% other (non–Hispanic or Latino)	CBT[p] or tai chi protocol reinforced with SMS text messaging
Glasner et al [136], 2020	AOD	60% African American or Black, 25.7% Hispanic or Latino, and 14.3% White	SMS text messaging or CBT intervention (ALC-TXT-CBT[q])
eSBIRT[r] or eSBI[s]			
Mullen et al [50], 2015	Alcohol	75% Hispanic or Latino, 94% White, 2% African American or Black, and 4% multiple races or other	MATTERS[t]—eSBIRT with DUI[u] offenders
Burner et al [51], 2020	Alcohol	88% Hispanic or Latino	MROAD[v]—SMS text messaging adaptation in Spanish or SBIRT[w]
CBI[x]			
Spohr et al [137], 2015	AOD	65.8% African American or Black and 27.6% White	MAPIT[y]—CBI for probationers+voluntary SMS text message or email reminders
Steers et al [138], 2016	Alcohol	42% Hispanic or Latino, 46% White, 14% African American or Black, 20% other, and 4% multiethnic	PNF[z]—web-based intervention
Aronson et al [52], 2017	Drug use	55% Hispanic or Latino (White), 10% White (non–Hispanic or Latino), 23% African American or Black (non–Hispanic or Latino), and 6% African American or Black (Hispanic or Latino)	MIK[aa]—people with disabilities or people with HIV (web-based CBI)
Ditre et al [139], 2019	Opioids and tobacco	47.1% African American or Black, 41.2% White, 11.8% Hispanic or Latino, and 11.8% Asian	PFI[ab] for people living with HIV—CBI
Gryczynski et al [140], 2021	AOD	84% African American or Black, 13% multiple races, 6% Hispanic or Latino (any race), and 2% White	Computer-delivered behavioral intervention (CBI)
Pill dispenser+CM[ac]			
Moore et al [141], 2015	AOD	60% African American or Black, 40% White, 20% Hispanic or Latino, and 80% non–Hispanic or Latino	CARE[ad] electronic pill dispenser (ART[ae]), CM via electronically loaded study debit cards, and CBT for HIV (phone based)
mCM[af]			
Beckham et al [142], 2018	Cannabis and tobacco	100% African American or Black	ART—ART combines mCM, telephone CBT, and a telehealth clinic for NRT[ag]
Mobile app			
Babson et al [143], 2015	Cannabis	50% African American or Black, 50% White, and 25% Hispanic or Latino	CBT-I[ah] iOS Coach (mobile app)
Email			

XSL•FO
RenderX

Platform and study	Substance use targets	Inclusion	TBI[a]
Alemi et al [62], 2010	AOD	67% African American or Black (non–Hispanic or Latino), 11% White (non–Hispanic or Latino), and 15% American Indian or Alaska Native	MI[ai] counseling (via email)

[a]TBI: technology-based intervention.
[b]AOD: alcohol or drug use.
[c]WHC: Women's Health CoOp.
[d]OUD: opioid use disorder.
[e]COPE: Concurrent Treatment of Posttraumatic Stress Disorder and Substance Use Exposure Using Prolonged Exposure.
[f]MOUD: medication for opioid use disorder.
[g]ATP: Addiction Telehealth Program.
[h]VR: virtual reality.
[i]AAT: avatar-assisted therapy.
[j]EMA: ecological momentary assessment.
[k]IOP: intensive outpatient program.
[l]AA: African American.
[m]SUD: substance use disorder.
[n]IVR: interactive voice response.
[o]TRAC2: Texting to Reduce Alcohol Consumption.
[p]CBT: cognitive behavioral therapy.
[q]ALC-TXT-CBT: cognitive behavioral therapeutic texting intervention for targeting alcohol use.
[r]eSBIRT: electronic screening, brief intervention, and referral to treatment.
[s]eSBI: electronic screening brief intervention.
[t]MATTERS: Motivational Alcohol Treatments to Enhance Roadway Safety.
[u]DUI: driving under the influence.
[v]MROAD: Mobilizing to Reduce Overuse of Alcohol in the emergency department.
[w]SBIRT: screening, brief intervention, and referral to treatment.
[x]CBI: computer-based intervention.
[y]MAPIT: Motivational Assessment Program to Initiate Treatment.
[z]PNF: personalized normative feedback.
[aa]MIK: Mobile Intervention Kit.
[ab]PFI: personalized feedback intervention.
[ac]CM: contingency management.
[ad]CARE: Centralized Off-Site Adherence Enhancement.
[ae]ART: antiretroviral therapy.
[af]mCM: mobile contingency management.
[ag]NRT: nicotine replacement therapy.
[ah]CBT-I: cognitive behavioral therapy for insomnia.
[ai]MI: motivational interviewing.

Recruitment and Retention Efforts

A total of 65.5% (72/110) of the included studies recruited participants from medical centers (43/72, 60%), substance use treatment settings (26/72, 36%), or combined and integrated medical and substance use treatment programs in academic settings (3/72, 4%). A handful of studies recruited from social service programs (6/110, 5.5%; eg, Special Supplemental Nutrition Program for Women, Infants, and Children; foster care transition services; and crisis shelters), justice settings (5/110, 4.5%; eg, courts and programs for driving under the influence), and community settings (2/110, 1.8%; eg, church and syringe service programs). A total of 16.4% (18/110) of the studies recruited from more than one setting, and in other studies, recruitment was not described (2/110, 1.8%) or did not apply (4/110, 3.6%; eg, secondary analyses). Retention efforts included compensation for study participation, the device delivering the TBI being given to participants to keep after study completion, and a combination of these strategies. Some studies did not describe retention efforts, and in many instances, they were not applicable (ie, single-session TBI).

Device or Internet Access Requirements

Most studies (64/110, 58.2%) did not require access to the internet or a device for the TBIs. Among the studies that required internet or a device, 24.5% (27/110) provided devices either to all study participants or to those in need, and 18.2% (20/110) required participants to own or have access to a device to be eligible to take part in the study. The TBIs in 39.1% (43/110) of the studies were delivered on-site and required internet access, whereas in 25.5% (28/110) of the studies, they were delivered in remote settings and also required internet

XSL•FO
RenderX

access. Geographically, the TBIs in 51.8% (57/110) of the studies were delivered in urban settings; however, 44.5% (49/110) of the studies did not clearly report the geographic setting. A total of 49.1% (54/110) of the studies described the involvement of potential end users, defined as individuals or groups for which the technology was developed, in the formative development or adaptation of a TBI using one or more of the following: interviews (19/54, 35%) [34,35,48,49,52,54,59,69,76,82,84,87,108,114,118,119,121,124,132], focus groups (15/54, 28%) [35,36,42,45,53,60,82,84,88,110,120,135,136], feedback sessions or gathering (11/54, 20%) [60,61,63-65,67,68,70,75,128,130], surveys (13/54, 24%) [48,50,53,70,76,82,94,107,116-118,122,141,143], usability testing (6/54, 11%) [47,54,58,108,110,112], and review by native speakers of the new adaptation (3/54, 6%) [40,41,115]. Many studies detailed the use of more than one method to involve potential end users.

End-User Groups

The dataset can be categorized into 11 end-user groups: childbearing-aged, pregnant, or perinatal women (22/110, 20%) [35,36,43,44,47,54,63-75,77,106,120], people living with HIV or AIDS (18/110, 16.4%) [39,61,83-87,112-115,118,122,124,131,135,136,139,141], people in substance use treatment (17/110, 15.5%) [34,37,60,96-100,104,107,108,119,121,122,127], people who inject drugs (2/110, 1.8%) [52,125], people seeking substance use treatment (15/110, 13.6%) [38,40,41,55-58,78-82,105,109,129], adolescents or youth (12/110, 10.9%) [45,48,88-90,94,116,117,133,140], youth exiting foster care (2/110, 1.8%) [53,75], people who were justice or court involved (6/110, 5.5%) [42,49,50,95,110,137], people being treated in hospital (detoxification) or emergency departments (6/110, 5.5%) [46,91-93,134], people who were unhoused (4/110, 3.6%) [59,111,126,130], men who have sex with men (5/110, 4.5%) [101-103,118,132], veterans (2/110, 1.8%) [123,143], non–treatment-seeking adults (1/110, 0.9%) [128], community recruits (2/110, 1.8%) [62,142], and college students (1/110, 0.9%) [138].

Social Determinants of Health

When conducting a parallel scoping review that focused on tobacco use [32], we observed patterns of reported environmental conditions indicative of low SES (eg, being unemployed, having a low or poverty-level income, and being uninsured) among study participants. On the basis of this observation, we explored the dataset for this review and found similar patterns of low income (20/110, 18.2%; eg, 81% of the participants had a barely or totally inadequate income) [116], unemployment (31/110, 28.2%; eg, “18 subjects and only one was employed”) [118], lack of health insurance or state insurance (6/110, 5.5%; eg, 86% of the participants were on Medicaid or uninsured) [108], homelessness (4/110, 3.6%; eg, homeless adult patients in treatment [126]), or low educational attainment (8/110, 7.3%; eg, average of 12 years of education) [143]. Another group of studies in this review included those on adolescents that reported no SES (9/110, 8.2%). Notably, many of the people in these samples fit multiple low-SES categories, but the studies were categorized without duplication based on the most salient SES variable in the study demographics (eg, ≥50% of the sample). Finally, 15.5% (17/110) of the studies, which were focused on adults, did not report SES, and another 10.9% (12/110) of the studies appeared to have higher average levels of SES (eg, more than half of the sample were employed). Thus, aside from studies on adolescents and other studies that did not report SES, as well as a small group that had participants of a slightly higher SES, the study samples of the remaining 65% of the studies represented people with a very low SES.

Range and Nature of TBIs

Platform

Almost half (58/110, 52.7%) of the studies used computer-based intervention platforms, including electronic screening, brief interventions, and referrals to treatment. More specifically, 22.7% (25/110) of the studies used electronic screening, brief interventions, and referrals to treatment. Computer-based interventions were delivered via tablets and computers both with and without internet. One-third (37/110, 33.6%) of the dataset represented interactive voice response, EMAs or recovery support EMA interventions, or SMS text messaging intervention platforms delivered both with and without internet via mobile phones. The last large platform category was video, videoconferencing, or telepsychiatry (14/110, 12.7%). The remaining platforms included virtual reality or avatars (2/110, 1.8%) delivered via a computer, a pill dispenser+CM (1/110, 0.9%), a mobile app (1/110, 0.9%), mobile CM (1/110, 0.9%), and an email intervention delivered via a computer (1/110, 0.9%).

Substance Use Categories

Study-labeled substance use categories included alcohol or drug use or alcohol, tobacco, or drug use (45/110, 40.9%); alcohol only (27/110, 24.5%); and drug use only (6/110, 5.5%). Of the 110 studies, there were 8 (7.3%) opioid-focused studies [107-109,121,122,125-127], 5 (4.5%) that focused on opioids and cocaine [46,98-100,131], and 1 (0.9%) on opioids and tobacco [139]. A total of 5.5% (6/110) of the studies were cannabis focused [77,92,105,116,117,143], and an additional 4 were cannabis or alcohol focused (n=2, 50%) [59,93], cannabis or alcohol or drug use focused (n=1, 25%) [90], and cannabis or tobacco focused (n=1, 25%) [142]. Finally, 3.6% (4/110) of the studies focused on cocaine exclusively [61,115,128,130], and 2.7% (3/110) of the studies focused on methamphetamines [101-103]. It is important to note that the use of the study authors’ substance use labels resulted in an overlap of categories (eg, alcohol or drug use overlaps with drug use and cannabis, opioids, cocaine, and methamphetamines; Tables 2 and 3).

Behavior Change Theories and Techniques

One of the extraction variables for the review was behavior change theories and techniques. Of the 110 analyzed studies, 81 (73.6%) were based on *at least one* underlying behavior change theory, whereas 29 (26.4%) did not report an underlying theory or it was not applicable (ie, EMA studies). The top 6 most frequently referenced underlying theories were CBT (41/81, 51%); (20/81, 25%) screening, brief interventions, and referrals to treatment (10/81, 12%); self-determination theory (5/81, 6%); social cognitive theory (4/81, 5%); and the health

XSL•FO
RenderX

belief model (4/81, 5%). A total of 78.2% (86/110) of the studies reported using *at least one* behavior change technique in their TBIs. Studies reported including the following behavior change techniques: feedback (33/86, 38%), goal setting (19/86, 23%), skill training (18/86, 21%), reminders (17/86, 20%), self-monitoring (16/86, 19%), EMA (14/86, 16%), and social support (7/86, 8%).

Access or Delivery Mode and Frequency

The TBIs were most commonly accessed or delivered independently, with the interventions in 65 studies either being self-initiated (n=45, 69%) or using software-automated prompts (n=20, 31%). A subset of the TBIs were delivered with the support of another individual, including service providers (15/110, 13.6% of the studies) and research team members (7/110, 6.4% of the studies). The interventions in 11.8% (13/110) of the studies were delivered via a hybrid model, combining self-initiated with personnel support. Delivery frequency of the included TBIs was predominantly in a single session (with or without booster sessions; 30/110, 27.3% of the studies) or daily for a limited period (25/110, 22.7% of the studies).

Programs of Research

Programs of research (16/16, 100%) ranged from a minimum of 2 to a maximum of 17 studies and, in some cases, spanned decades. Given space constraints, we feature the 2 largest programs (Motivation Enhancement System [MES] and Computer-Based Training for Cognitive Behavioral Therapy [CBT4CBT]) and encourage readers to explore other programs and independent studies (Tables 2 and 3) to obtain a sense for how a program of research might evolve to promote health equity. The largest program was MES, a computer-delivered, single-session, self-initiated brief intervention originated in a publication in 2005 [63]. This early study provided the material for an additional 14.5% (16/110) of the studies in the dataset, all of which explored MES's utility primarily among perinatal or childbearing-aged women who identified as African American or Black individuals. MES was explored for utility among young adults who had recently exited foster care (ie, interactive health lifestyle preparation) and for utility with perinatal or childbearing-aged women who had different substance use profiles. The single-session nature of the tool also evolved to be combined with SMS text messaging within the program of research.

The second largest program (10/110, 9.1% of the studies) was CBT4CBT, a self-initiated, multi-session computer- or tablet-based version of CBT used in conjunction with standard clinical care for people currently using substances. The original, small-efficacy study comparing CBT4CBT to TAU [55] was referenced as source material in an additional 8.2% (9/110) of the studies in the dataset, including exploration of the TBI in small randomized trials exploring mechanisms as well as a culminating randomized controlled trial evaluating the efficacy and safety of CBT4CBT as a stand-alone treatment. The largest proportion of the samples in the first 70% (7/10) of the studies within this program comprised people who identified as African American or Black; however, each sample also included people who identified as Hispanic or Latino and some who identified as American Indian or Alaska Native. A total of 20% (2/10) of the most recent studies in this program of research included clear examples of work to promote health equity. Paris et al [38] described the primary outcomes of a randomized trial comparing a culturally adapted version of CBT4CBT (CBT4CBT-Spanish) to standard mental health and addiction treatment in treatment-seeking Latino adults, and Jordan et al [82] assessed the feasibility outcomes of a clinical trial for African American or Black adults with substance use disorder in a novel setting—a Black church, where modifications to the TBI were made by church-based health advisors.

Race and Ethnicity Consciousness

A total of 26.4% (29/110) of the studies met our race or ethnicity consciousness criteria that at least one manuscript section be explicit about a particular URM group. In total, 24% (7/29) of these studies included samples that comprised 100% members of a single URM group and were race or ethnicity conscious across all manuscript sections—acknowledging a racial or ethnic group in the title, introduction, methods, results, and discussion sections (African American or Black individuals [82,120], American Indian or Alaska Native individuals [34-36], and Hispanic or Latino individuals [40,41]). An additional 17% (5/29) of the studies [38,42,44,51,132] were race or ethnicity conscious across all sections except for the manuscript title. Among the remaining 59% (17/29) of the studies, evidence of race or ethnicity consciousness was present yet inconsistent (Table 4).

XSL•FO
RenderX

Table 4. Race- and ethnicity-conscious studies by manuscript section.

Study	Title	Introduction	Methods	Results	Discussion
Wechsberg et al [120]	✓	✓	✓	✓	✓
Jordan et al [82]	✓	✓	✓	✓	✓
Gorman et al [35]	✓	✓	✓	✓	✓
Campbell et al [34]	✓	✓	✓	✓	✓
Montag et al [36]	✓	✓	✓	✓	✓
Muroff et al [40]	✓	✓	✓	✓	✓
Muroff et al [41]	✓	✓	✓	✓	✓
Osilla et al [42]		✓	✓	✓	✓
Yang et al [132]		✓	✓	✓	✓
Paris et al [38]		✓	✓	✓	✓
Nayak et al [43]		✓	✓	✓	✓
Nayak et al [44]		✓	✓	✓	✓
Burner et al [51]		✓	✓	✓	✓
Comulada et al [45]	✓	✓	✓		✓
Brusoski and Rosen [121]	✓	✓	✓		
Preston et al [97]		✓		✓	✓
Dawson Rose et al [112]				✓	✓
Forray et al [72]				✓	✓
Saberi et al [114]			✓		✓
Ondersma et al [68]			✓		✓
Reback et al [103]		✓			
Tzilos et al [67]		✓			
Tzilos Wernette et al [54]		✓			
Ondersma et al [70]		✓			
Ondersma et al [75]			✓		
Ondersma et al [77]			✓		
Waller et al [92]			✓		
Suffoletto et al [134]				✓	
Gryczynski et al [140]					✓

Critical Appraisal

Higher appraisal or confidence logically comes from studies with race- or ethnicity-conscious results with samples that included ≥90% of members of a single URM group (10/110, 9.1%) [34,35,38,40-42,82,120,121,132]. Despite not having race- or ethnicity-conscious results, the samples of 11.8% (13/110) of the studies included ≥90% of members of a single URM group, making the results intrinsically applicable to that group (ie, functionally equivalent to being race conscious) [37,45,63-66,69,104,128,130,131,142]. A small group of studies (8/110, 7.3%) [36,43,44,51,72,97,112,134] had samples comprising <90% of members of a single URM group; however, due to their race- or ethnicity-conscious results, there was higher confidence in interpretable findings. A lower appraisal or confidence rating was assigned to those studies that included <90% of members of a single URM group *and* lacked race-conscious results (79/110, 71.8%). As samples increase in racial or ethnic diversity, without subgroup analyses, it becomes challenging to clearly interpret outcomes for different groups. Table 5 shows the critical appraisal of all the studies in the dataset.

XSL•FO
RenderX

Table 5. Critical appraisal of the dataset (N=110)[a].

Race-conscious results	<50% of participants from a single URM[b] group	50%-59% of participants from a single URM group	60%-69% of participants from a single URM group	70%-79% of participants from a single URM group	80%-89% of participants from a single URM group	≥90%-100% of participants from a single URM group
Yes	• Dawson Rose et al [112]	• Suffoletto et al [134]	• Nayak et al [43] • Forray et al [72] • Preston et al [97]	• Nayak et al [44]	• Montag et al [36] • Burner et al [51]	• Campbell et al [34][c] • Gorman et al [35][c] • Paris et al [38] • Muroff et al [40,41] • Osilla et al [42][c] • Jordan et al [82][c] • Brusoski and Rosen [121][c] • Yang et al [132][c] • Wechsberg et al [120][c]
No	• Braciszewski et al [53] • Tzilos Wernette et al [54] • Carroll et al [55] • Sugarman et al [56] • Tofighi et al [108] • Saberi et al [114] • Osilla et al [110] • Dennis et al [94] • Tofighi et al [107] • Shrier et al [116] • Shrier et al [117] • Delrahim-Howlett et al [106] • Hasin et al [83] • Dawson Rose et al [113] • Ditre et al [139] • O'Gurek [127] • Castillo et al [125] • Kiluk et al [80] • Tofighi et al [109] • Reback et al [101,102] • Reback et al [103] • Steers et al [138]	• Aharonovich et al [39] • Comulada et al [46] • Aronson et al [52] • Kiluk et al [57] • Kiluk et al [58] • Muench et al [60] • Braciszewski et al [76] • Saberi et al [114] • Cunningham et al [88] • Moore et al [135] • Welsh [122] • Babson et al [143] • Aharonovich et al [86] • Kiluk et al [78] • Kiluk et al [79] • Roos et al [81] • Budney et al [105] • Gordon et al [129] • Walton et al [89] • Blow et al [91] • Waller et al [92] • Drislane et al [93] • Kennedy et al [111]	• Sanjuan et al [47] • Trudeau et al [48] • Tolou-Shams et al [49] • Thompson et al [59] • Alemi et al [62] • Hasin et al [84] • Glasner et al [136] • Moore et al [141] • Ingersoll et al [115] • Martino et al [71] • Loree et al [73] • Yonkers et al [74] • Aharonovich et al [85] • Preston et al [98-100] • Walton et al [90] • Buu et al [133] • Spohr et al [137] • Mehtani et al [126]	• Mullen et al [50] • Aharonovich et al [87] • Ondersma et al [75] • Scott et al [96]	• Ingersoll et al [61] • Frueh et al [119] • Palfai et al [124] • Tzilos et al [67] • Ondersma et al [70,77] • Gryczynski et al [140] • Scott et al [95]	• Legha et al [37][c] • Comulada et al [45] • Pollick et al [69][c] • Ondersma et al [63-65][c] • Saladin et al [128] • Freedman et al [130] • Ondersma et al [66] • Linas et al [131] • Brooks et al [104] • Jaconis et al [123][c] • Beckham et al [142][c]

[a]Lower appraisal: <90%-100% of participants from a single underrepresented minority group AND NO race-conscious results; higher appraisal: ≥90%-100% of participants from a single underrepresented minority group OR race-conscious results.

[b]URM: underrepresented minority.

[c]100% of participants are from a single underrepresented minority group.

XSL•FO
RenderX

Discussion

Health Equity Promotion

Overview

To our knowledge, this scoping review is the first systematic literature exploration of health equity promotion in TBIs for substance use treatment among URM groups at heightened risk of disparities in substance use treatment and related outcomes in the United States [6,144-146]. This review offers insights into the extent to which the evaluated studies promote health equity and identifies opportunities for growth. Productive health equity research focuses on strengths and solutions in lieu of redocumenting problems [24,147,148]. All the identified areas of opportunity for growth will include examples for how to promote equity in these spaces.

Access or Inclusion

An analysis of US trials in ClinicalTrials.gov [149] revealed that, over the past 2 decades (2000-2020), most registered studies did not report race and ethnicity enrollment data and, among those with demographic data, racial or ethnic minority groups were underrepresented. It is noteworthy that federal requirements to report have only been in place since 2017. In our review, nearly half (532/1158, 45.9%) of the full-text articles assessed for eligibility were excluded due to either ineligibility based on demographic criteria or the absence of reporting of race or ethnicity data, substantiating this historical pattern of lack of access or inclusion that perpetuates health inequity for some of the most vulnerable members of our society. Of those studies that did include URM groups, approximately half (57/110, 51.8%) described engagement of end users in some stage of the TBI research. The scientific community is increasingly focusing on translation and implementation of scientific discoveries with a parallel increase in attention to end-user engagement as a key component in the process of tailoring best practices for specific populations who are hardly reached by research and, thus, disproportionately impacted by chronic health conditions [150,151]. The promotion of representation and equity in research are by-products of these foci [152]. As will be discussed more extensively in the following sections, community-engaged and culturally informed interventions have been demonstrated to improve the health outcomes of underserved groups facing health disparities [26,153,154]. Research collaboration with the communities we seek to serve and partner with to promote health equity through the development of TBIs for substance use is essential.

Range and Nature of TBIs

Most of the netted studies (76/110, 69.1%) were programs of research, highlighting the importance of conducting various types of research studies focused on different stages of science or scientific questions (eg, intervention development, efficacy studies, and effectiveness studies), consistent with the National Institutes of Health (NIH) Stage Model for behavioral intervention development, with the ultimate goal of producing highly potent and maximally implementable interventions that improve health and well-being [155]. This model is also consistent with the path to understanding the replicability and reproducibility of effects in various contexts and populations [24]. The independent studies netted by this review are disproportionately more recent publications, pointing to future research programs with commitments to URM group health equity promotion. Although expected given the rapid proliferation in TBIs in health research generally, the increased volume in recent years, as well as the range of substances targeted by the studies in this dataset, are encouraging signs.

The prevalence of independently delivered or accessed TBIs is notable given the potential for many underserved people to live in contexts in which access to care may be limited. In addition, many people who identify as members of URM groups may be reluctant to engage in traditional models of care due to mistrust of the research and medical communities [156-158]. Furthermore, the prevalence of brief evidence-based interventions accessed or delivered independently reduces the threshold to receive an effective dose of the intervention [159]. Efforts to simplify technological complexity are likely to disproportionately benefit less advantaged groups [19,160], possibly due to the effort involved in negotiating additional barriers to health behaviors that emerge in underserved contexts [161].

Race and Ethnicity Consciousness

Abandoning color-blind approaches to research potentiates exploration of treatment impacts that values the perspectives and voices of others [162]. Approximately one-quarter of the studies (29/110, 26.4%) were explicitly conscious of members of underrepresented racial and ethnic groups in their research. Most of the race- or ethnicity-conscious studies (23/29, 79%) also engaged end users and community members in the design or conduct of the research, further amplifying their voices. These exemplars stand in contradistinction to acontextually developed innovations that may largely benefit health outcomes in one sector of society while potentially creating, sustaining, or increasing health disparities in another [163].

Critical Appraisal

The trend toward an increased focus on including URM groups in TBI research, as well as the increased race and ethnicity consciousness of this research, is unquestionably a positive finding. However, most studies (79/110, 71.8%) were limited in the extent to which the outcomes could be interpreted to have implications for people who identify as African American or Black, Hispanic or Latino, or American Indian or Alaska Native individuals. In our critical appraisal, 28.2% (31/110) of the studies were highly appraised in terms of the interpretability of findings for one or more URM groups. Through this appraisal of the dataset, as well as the scoping review at large, we hope to encourage researchers to identify existing TBIs for potential adaptation, potential gaps in the knowledge base, or insights for new tool development.

Opportunities for Growth

Careful attention to patterns in the review data illuminate several opportunities for promoting health equity in substance use treatment research with URM groups: lowering access thresholds to increase participation in research, especially among Hispanic or Latino and American Indian or Alaska Native individuals;

XSL•FO
RenderX

expanding TBI substance use research focusing on opioids, stimulants, and cannabis; using culturally informed theoretical models or equity approaches to knowledge production; planning and conducting racial and ethnic subgroup analyses; adapting TBIs with demonstrated effectiveness for examination with URM groups; and considering intersectionality.

Lowering Access Thresholds to Increase Participation, Especially Among Hispanic or Latino and American Indian or Alaska Native Individuals

To enhance inclusion, low-threshold, community-based recruitment strategies and settings are needed. Most participants in our dataset were recruited from medical centers, substance use treatment settings, or integrated medical and substance use treatment programs in academic settings. These settings are considered high threshold due to access requirements of transportation [164], insurance [165], trust in the medical community [156,157], and time to participate. The increased challenges of conducting research with people who are incarcerated compared with people in the community [166] and the disproportionate burden of incarceration on URM groups [167] suggest reconsideration of barriers to the conduct of research with people who are incarcerated. This is a significantly understudied population and doubly marginalized by URM group and prisoner status (see the Intersectionality section).

Given the prevalence of high-threshold recruitment settings in our dataset, the dearth of research on TBIs for substance use treatment among people who identify as Hispanic or Latino or American Indian or Alaska Native individuals is unsurprising and significant. Both American Indian or Alaska Native and Hispanic or Latino individuals experience some of the worst health disparities in the United States and are more likely than White individuals to be poor and unemployed and have lower educational attainment [145,146,168]. American Indian or Alaska Native individuals are also more likely than White individuals and other URM groups to live in rural frontier settings, compounding the issue of access and inclusion [169]. Recruitment of sufficient numbers of racial and ethnic minority group participants is imperative to understanding how research affects underrepresented groups [25], with underrepresentation limiting the extent to which members of these groups benefit from advances in substance use treatment research [170]. Community-engaged recruitment strategies are needed to increase access and inclusion of URM groups [171].

Expanding TBI Substance Use Research Targeting Opioids, Stimulants, and Cannabis

Few netted studies (30/110, 27.3%) targeted opioids, stimulants (eg, methamphetamines and cocaine), and cannabis specifically. The US opioid and stimulant epidemics [172], as well as the increasing legalization of cannabis [173], create an expanding need for treatment for all people navigating problematic or disordered use [174]. However, this is particularly true for URM groups due to their marginalization in *both* treatment and research [175]. TBIs offer the potential to mitigate health-related disparities in care by addressing the needs of members of diverse groups and increasing access. While there is limited work focusing on people who use opioids, stimulants, and cannabis in the United States generally, our review identified the lack of research on TBIs for these drug classes that includes URM groups. The gaps in the literature on TBIs for these 3 drug classes offer enormous opportunities for promoting health equity. Several reviews have summarized the use of TBIs related to substance use [18,176-182]. However, most interventions have focused exclusively on impacting alcohol use behaviors [183].

Expanding TBI development and research beyond a heavy focus on alcohol is especially critical due to the opioid overdose epidemic. Opioid-related overdose deaths among people identifying as African American or Black and American Indian or Alaska Native individuals now outpace those of White individuals in the United States, in part because the science and resources brought to bear to fight the opioid epidemic disproportionately exclude people who identify as belonging to URM groups, especially during the early phase of the epidemic [184,185]. A total of 3.6% (4/110) of the identified studies in this scoping review focused on TBIs for opioid use disorder (OUD); all 4 were telehealth focused [186-189]. While regulations expanding telehealth care for OUD issued during the COVID-19 pandemic appear to remain in place [190], further enabling the exploration of the benefits of telehealth for OUD treatment and developing creative, effective, and culturally tailored TBIs for OUD in addition to telehealth interventions is important to expand access to OUD care.

Using Culturally Informed Theoretical Models or Equity Approaches to Knowledge Production

The theoretical models and perspectives represented in the dataset were predominantly generalizability approaches, with negligible use of culturally informed or sensitive theories or equity approaches to knowledge production (eg, sociological trust theory, community-based participatory research, public health critical race praxis, and social determinants of health) [19,163,191,192]. While the identification of the commonalities of human experience is useful, to promote health equity, researchers must explore the differential effectiveness of evidence-based interventions for URM groups.

Several netted studies (4/110, 3.6%) [34,40,41,82] used equity-promoting theories or frameworks based on community-engaged, participatory research practices in adaptations [40,41,82] or explorations of the feasibility of the adaptation [34] of evidence-based TBIs. An understanding of the power of culturally or contextually informed theories and frameworks to disrupt social mechanisms that fuel health disparities may be instructive. For example, the theoretical underpinnings for the Addiction Comprehensive Health Enhancement Support System [40,41] are based on self-determination theory [193]. One of the defining features of self-determination theory is its treatment of both the person and the social context in motivated behavior. By adapting and translating the Addiction Comprehensive Health Enhancement Support System to be linguistically and culturally relevant to Spanish-speaking Latinos in recovery, the authors support competence and autonomy by eliciting and acknowledging patients' context and perspectives [194].

Community-based participatory research methods incorporate community member input at all phases of the research process

XSL•FO
RenderX

[163]. Offering an evidence-based treatment (CBT4CBT) in an alternative setting, such as a Black church, Jordan et al [82] explored the role of context to improve access to and the efficacy of evidence-based treatments. Through engagement with the community, as well as training church staff to deliver evidence-based treatments, the authors assessed feasibility and probed necessary adaptations to the TBI to improve acceptability. Campbell et al [34] used community-engaged research methods to reveal preferences regarding delivery and presentation, including using American Indian or Alaska Native actors for voice and video components, use of stories as opposed to "academic" presentations, and inclusion of American Indian or Alaska Native cultural representation. There is substantial evidence-based work in the field of TBIs for substance use treatment. Using theories and frameworks that are sensitive to context and the diversity of human experiences, community-engaged research methods, and adaptations is possible, and efficacy is likely to be strengthened for members of underrepresented groups.

Planning to Conduct Racial and Ethnic Subgroup Analyses

Limited subpopulation analyses pose significant problems for reviews [159]. Of the 110 studies in this review, 81 (73.6%) missed an opportunity to better understand the race or ethnicity subgroup–specific implications of their findings. Thus, while it may appear that the studies identified in our review focused on TBIs with a strong evidence base, more nuanced assessments of how differentially effective these standard-bearers might be is lacking. Given the limited inclusion of members of underrepresented populations in biomedical research [195], when a sample does include sufficient representation, it is a prime opportunity for interested researchers to review data to improve our understanding of the implications for these groups. Burlew et al [25] further recommend that researchers consider within-race differences and the potential disadvantages of combining racial or ethnic minority groups in the analyses (eg, Weaver et al [196] revealed differential odds of meeting the criteria for major depressive disorder between rural Black women [lower] and urban Black women [higher]).

There is opportunity to explore clinical trials with diverse samples or large datasets that are a part of the public record to help identify moderators to improve our understanding of the differential impacts of our policies and interventions for URM groups and the multiply disadvantaged among us [197]. The NIH recently (2021) recommended that investigators should not control for the contextual variables (eg, neighborhoods and employment opportunities) that are the drivers of inequity but, rather, that these factors be explicitly studied. Systems of oppression are inherently bound together—the intersection of multiple identities (eg, gender and race) within social systems of power may compound and synergistically exacerbate experiences of health disparity [198].

Adapting TBIs With Demonstrated Effectiveness for Examination With URM Groups

Related to the aforementioned opportunity for growth regarding planned subgroup analyses, another opportunity for interested researchers is to adapt TBIs with demonstrated effectiveness for examination among URM groups. Many of the programs of research netted in this review underscore this opportunity. For example, Paris et al [38] described the primary outcomes of a randomized clinical trial comparing a culturally adapted version of CBT4CBT-Spanish to standard outpatient mental health and addiction treatment in a heterogeneous population of treatment-seeking Latino adults. Their rationale for undertaking this research included consideration of previous findings on the efficacy, safety, and durability of the existing English-language versions of CBT4CBT when added to standard outpatient treatment. There is no shortage of TBIs for substance use disorders; however, as this review reveals, there is a shortage of understanding which among the many might be efficacious for URM groups.

Considering Intersectionality

While seeking to characterize the range and nature of TBIs for substance use treatment among people who identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native individuals, we discovered that most of the studies (85/110, 77.3%) included historically underserved populations, such as pregnant or perinatal women, people living with HIV or AIDS, adolescents or youth, youth exiting foster care, people who are justice or court involved, people who are unhoused, and veterans. In addition, most of the participants across these studies were impacted by multiple socioeconomic challenges, notably income, housing, education, and health insurance. A recent literature review on multiple disadvantages in relation to telehealth [199] identified intersectionality theory [200] as an approach that the authors felt might help account for digital disparities. Intersectionality theory posits that considering a single axis of inequality is limited and that considering disadvantage on multiple axes simultaneously is needed [201]. Applied to digital health disparities, this theory suggests the hypothesis that our individual identities and lived experiences are unique and multidimensional and that we will or will not use TBIs based on our unique identities and experiences rather than due to our identification with single categories such as race, age, gender, or disability. Per intersectionality theory, addressing health inequities will entail an increased understanding of how various structural and social determinants of health interact. To achieve this increased understanding, we must work with individuals, families, groups, and communities to understand what shapes their experiences of advantage and disadvantage. The import of this theory cannot be understated for the data that we collected. To capture intersectionality, a first challenge for researchers working in this space and beyond is to take a step back, digest the large-scale studies that appear to have uniform messages about broad factors associated with efficacy or effectiveness applicable to all [202], and commit to collecting and reporting data related to multiple categories at once (eg, race, sex, and income). A second challenge is to combine this knowledge with tailored implementation strategies, acknowledging how different aspects of disadvantage related to social determinants of health interact in individual lives, impacting access, adoption, and effectiveness of TBIs for substance use. Dr Nora Volkow, the director of the National Institute on Drug Abuse, recently blogged that "measuring social determinants of health can help researchers

XSL•FO
RenderX

better design treatment interventions and services, as well as make addiction care more equitable. Research in other areas of medicine has already revealed the distorting effects of failure to take that step" [203]. We must strive to conduct research that promotes health equity for all.

Future Research

We acknowledge researchers who have been working in the health equity field that dates back more than a century [148]. While most of our team members identify as White, we acknowledge the historical damage posed by many White allies and aspire beyond "health equity tourism" [148,204]. We aim to be solution focused, collaborating with deeply knowledgeable equity researchers to build on the knowledge presented herein. In this spirit, we seek to partner with scholars from underrepresented groups on future research exploring optimal practices for research with diverse and underrepresented groups; race and ethnicity consciousness; and effectiveness of TBIs for people who identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native individuals based on this scoping review. We encourage readers to take note of the research opportunities implied in the Opportunities for Growth section as well.

Limitations

Several limitations warrant attention. First, our eligibility criteria of samples of ≥50% of URM group members may have precluded exploration of literature with the potential to provide further insights to promote health equity in the TBI for substance use treatment arena. Second, the threshold for designation of higher appraisal or confidence versus lower appraisal confidence of 90% of members of a single URM group in the study samples was an arbitrary cutoff and, as such, highlights the guidance nature of critical appraisals. Per a recent publication on critical appraisals, "decisions on whether a study is considered weak, moderate, or strong are based on arbitrary cut-off scores" [205]. In theory, any cutoff of >50% would show that most participants were from a single URM group, although 90% could be considered a conservative cutoff to ensure that the results have direct implications for that group. The appraisal is intended as guidance or a gauge of the confidence that a reader should reasonably have that the study in question may have interpretable findings for a particular URM group. Third, to avoid intervention-generated inequalities and ensure that an equity lens is applied in the conduct, reporting, and use of research, an oft-cited recommendation is for researchers to report relevant sociodemographic characteristics of the samples. These include the PROGRESS-Plus (place of residence, race, ethnicity, language, occupation, gender or sex, religion, educational level, SES, social capital) factors of place of residence, race, ethnicity, language, occupation, gender or sex, religion, educational level, SES, age, disability, and sexual orientation [19,206]. This framework functions as a reminder that understanding health equity encompasses numerous factors beyond race and ethnicity that impact intervention access, adherence, adoption, and effectiveness. While our review focused on race or ethnicity, we acknowledge the importance of attending to a broader array of demographic variables to ensure that TBIs do not fall into the trap of inadvertently contributing to increasing health inequities, increasing the gap between the most and least advantaged. Fourth, a general limitation of the TBI research included in this review was the low SES among participants of the netted samples, as well as the high-threshold venues for recruitment. Both are filters that limit a more thorough understanding of the potential of TBIs to promote health equity.

Conclusions

Through this review, we sought to provide context, insights, and direction for researchers working to develop and evaluate digital technology substance use interventions while promoting health equity. An NIH recommendation on enhancing health disparity research suggests that we must go beyond "ticking the box" when including racial and ethnic minority groups in research [158]. Instead, underrepresented populations must be *the focus* of studies, or at the very least, the studies must be race and ethnicity sensitive [197]. As we do not all experience equity in health, we must resist the temptation to use universal instruments to attain universal ends; one size has never fit all [207]. This scoping review was conducted in the spirit of this recommendation. We sought to focus attention on those studies that included members of underrepresented groups and, further, those that gave voice to these people through race- and ethnicity-conscious research or end-user engagement. A noted sociologist who writes about technology and its potential to exacerbate inequities suggests that "the ways we engineer the material world reflects and reinforces (but could also be used to subvert) social hierarchies" [208]. It is our fervent hope that insights into the engineering of TBIs for substance use treatment for URM groups can and will be used to erode structural racism in health care.

Acknowledgments

The authors thank research and education librarians Heather Blunt and Paige Scudder at the Biomedical Libraries of Dartmouth College for their expertise and support in honing the research questions, developing the search strategy, and running the search through the selected databases, creating the dataset used in this scoping review. The authors would also like to acknowledge Dr Kathy Burlew for her knowledge regarding best practices when researching diverse groups, as well as her personal encouragement of this scoping review.

Conflicts of Interest

LAM is affiliated with Square2 Systems, Click Therapeutics, and Pear Therapeutics. These relationships are extensively managed by her academic institution, Dartmouth College. All other authors report no other conflicts of interest.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 108 KB - jmir_v26i1e53685_app1.docx]

Multimedia Appendix 2
MEDLINE electronic search strategy.
[DOCX File , 14 KB - jmir_v26i1e53685_app2.docx]

References

1. Matsuzaka S, Knapp M. Anti-racism and substance use treatment: addiction does not discriminate, but do we? J Ethn Subst Abuse 2019 Jan 14;19(4):567-593. [doi: 10.1080/15332640.2018.1548323]
2. Treatment Episode Data Set (TEDS): 2020 Admissions To And Discharges From Publicly Funded Substance Use Treatment Facilities. Substance Abuse and Mental Health Services Administration. URL: https://www.samhsa.gov/data/report/treatment-episode-data-set-teds-2020-admissions [accessed 2024-11-13]
3. Volkow N. Making addiction treatment more realistic and pragmatic: the perfect should not be the enemy of the good. National Institute on Drug Abuse. 2022. URL: https://nida.nih.gov/about-nida/noras-blog/2022/01/making-addiction-treatment-more-realistic-pragmatic-perfect-should-not-be-enemy-good [accessed 2024-04-29]
4. Highlighted population slides. SAMSHA. URL: https://www.samhsa.gov/data/report/2021-nsduh-highlighted-population-slides [accessed 2023-06-08]
5. Mennis J, Stahler GJ. Racial and ethnic disparities in outpatient substance use disorder treatment episode completion for different substances. J Subst Abuse Treat 2016 Apr;63:25-33. [doi: 10.1016/j.jsat.2015.12.007] [Medline: 26818489]
6. Mennis J, Stahler GJ, El Magd SA, Baron DA. How long does it take to complete outpatient substance use disorder treatment? Disparities among Blacks, Hispanics, and Whites in the US. Addict Behav 2019 Jun;93:158-165. [doi: 10.1016/j.addbeh.2019.01.041]
7. Pinedo M. A current re-examination of racial/ethnic disparities in the use of substance abuse treatment: do disparities persist? Drug Alcohol Depend 2019 Sep 01;202:162-167 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.05.017] [Medline: 31352305]
8. Stahler GJ, Mennis J. Treatment outcome disparities for opioid users: are there racial and ethnic differences in treatment completion across large US metropolitan areas? Drug Alcohol Depend 2018 Sep 01;190:170-178. [doi: 10.1016/j.drugalcdep.2018.06.006] [Medline: 30041092]
9. Whitesell NR, Beals J, Crow CB, Mitchell CM, Novins DK. Epidemiology and etiology of substance use among American Indians and Alaska Natives: risk, protection, and implications for prevention. Am J Drug Alcohol Abuse 2012 Sep;38(5):376-382. [doi: 10.3109/00952990.2012.694527] [Medline: 22931069]
10. Enoch MA, Albaugh BJ. Review: genetic and environmental risk factors for alcohol use disorders in American Indians and Alaskan Natives. Am J Addict 2017 Aug 06;26(5):461-468 [FREE Full text] [doi: 10.1111/ajad.12420] [Medline: 27599369]
11. Emert MB, Giano Z. A comparison of alcohol and commercial tobacco use in populous American Indian/Alaska Native states. Drug Alcohol Depend 2023 Oct 01;251:110941. [doi: 10.1016/j.drugalcdep.2023.110941] [Medline: 37683451]
12. Burger J, Kapron M. Drug policy and indigenous peoples. Health Hum Rights 2017 Jun;19(1):269-278 [FREE Full text] [Medline: 28630559]
13. Farahmand P, Arshed A, Bradley MV. Systemic racism and substance use disorders. Psychiatr Ann 2020 Nov;50(11):494-498. [doi: 10.3928/00485713-20201008-01]
14. Saloner B, Lê Cook B. Blacks and Hispanics are less likely than whites to complete addiction treatment, largely due to socioeconomic factors. Health Aff (Millwood) 2013 Jan;32(1):135-145 [FREE Full text] [doi: 10.1377/hlthaff.2011.0983] [Medline: 23297281]
15. National Academies of Sciences, Engineering, and Medicine, Board on Population Health and Public Health Practice, Committee on the Review of Federal Policies that Contribute to Racial and Ethnic Health Inequities, Geller AB, Polsky DE, Burke SP. Federal Policy to Advance Racial, Ethnic, and Tribal Health Equity. Washington, DC: National Academies Press; 2023.
16. Blume AW. Advances in substance abuse prevention and treatment interventions among racial, ethnic, and sexual minority populations. Alcohol Res 2016;38(1):47-54 [FREE Full text] [Medline: 27159811]
17. Jordan A, Quainoo S, Nich C, Babuscio TA, Funaro MC, Carroll KM. Racial and ethnic differences in alcohol, cannabis, and illicit substance use treatment: a systematic review and narrative synthesis of studies done in the USA. Lancet Psychiatry 2022 Aug;9(8):660-675. [doi: 10.1016/s2215-0366(22)00160-2]
18. Marsch LA, Guarino H, Acosta M, Aponte-Melendez Y, Cleland C, Grabinski M, et al. Web-based behavioral treatment for substance use disorders as a partial replacement of standard methadone maintenance treatment. J Subst Abuse Treat 2014 Jan;46(1):43-51 [FREE Full text] [doi: 10.1016/j.jsat.2013.08.012] [Medline: 24060350]
19. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 01;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]

XSL•FO
RenderX

20. Cerdeña JP, Rivera LM, Spak JM. Intergenerational trauma in Latinxs: a scoping review. Soc Sci Med 2021 Feb;270:113662. [doi: 10.1016/j.socscimed.2020.113662] [Medline: 33476987]
21. Hamed S, Bradby H, Ahlberg BM, Thapar-Björkert S. Racism in healthcare: a scoping review. BMC Public Health 2022 May 16;22(1):988 [FREE Full text] [doi: 10.1186/s12889-022-13122-y] [Medline: 35578322]
22. Metzl JM, Hansen H. Structural competency: theorizing a new medical engagement with stigma and inequality. Soc Sci Med 2014 Feb;103:126-133 [FREE Full text] [doi: 10.1016/j.socscimed.2013.06.032] [Medline: 24507917]
23. Burlew AK, Copeland VC, Ahuama-Jonas C, Calsyn DA. Does cultural adaptation have a role in substance abuse treatment? Soc Work Public Health 2013 May;28(3-4):440-460 [FREE Full text] [doi: 10.1080/19371918.2013.774811] [Medline: 23731430]
24. Burlew AK, Peteet BJ, McCuistian C, Miller-Roenigk BD. Best practices for researching diverse groups. Am J Orthopsychiatry 2019;89(3):354-368. [doi: 10.1037/ort0000350] [Medline: 31070421]
25. Burlew K, McCuistian C, Szapocznik J. Racial/ethnic equity in substance use treatment research: the way forward. Addict Sci Clin Pract 2021 Aug 05;16(1):50 [FREE Full text] [doi: 10.1186/s13722-021-00256-4] [Medline: 34353373]
26. Montague E, Perchonok J. Health and wellness technology use by historically underserved health consumers: systematic review. J Med Internet Res 2012 May 31;14(3):e78 [FREE Full text] [doi: 10.2196/jmir.2095] [Medline: 22652979]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Hichborn EG, Moore SK, Gauthier PR, Agosti NO, Bell KD, Boggis JS, et al. Technology-based interventions in substance use treatment to promote health equity among people who identify as African American/black, Hispanic/Latinx, and American Indian/Alaskan native: protocol for a scoping review. JMIR Res Protoc 2022 May 17;11(5):e34508 [FREE Full text] [doi: 10.2196/34508] [Medline: 35579930]
29. Peters MG, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
30. Huey Jr SJ, Polo AJ. Evidence-based psychosocial treatments for ethnic minority youth. J Clin Child Adolesc Psychol 2008 Jan 03;37(1):262-301 [FREE Full text] [doi: 10.1080/15374410701820174] [Medline: 18444061]
31. Taj F, Klein MC, van Halteren A. Digital health behavior change technology: bibliometric and scoping review of two decades of research. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
32. Hichborn E, Turner A, Moore S, Gauthier P, Bell K, Montgomery L, et al. Technology-based interventions in tobacco use treatment among people who identify as African American/Black, Hispanic/Latina/o, and American Indian/Alaska Native: scoping review. J Med Internet Res 2024 Oct 10;26:e50748 [FREE Full text] [doi: 10.2196/50748] [Medline: 39388699]
33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
34. Campbell AN, Turrigiano E, Moore M, Miele GM, Rieckmann T, Hu M, et al. Acceptability of a web-based community reinforcement approach for substance use disorders with treatment-seeking American Indians/Alaska Natives. Community Ment Health J 2015 May 15;51(4):393-403 [FREE Full text] [doi: 10.1007/s10597-014-9764-1] [Medline: 25022913]
35. Gorman JR, Clapp JD, Calac D, Kolander C, Nyquist C, Chambers CD. Creating a culturally appropriate web-based behavioral intervention for American Indian/Alaska Native women in Southern California: the healthy women healthy Native nation study. Am Indian Alsk Native Ment Health Res 2013;20(1):1-15. [doi: 10.5820/aian.2001.2013.1] [Medline: 23529767]
36. Montag AC, Dusek ML, Ortega ML, Camp-Mazzetti A, Calac DJ, Chambers CD. Tailoring an alcohol intervention for American Indian Alaska native women of childbearing age: listening to the community. Alcohol Clin Exp Res 2017 Nov 26;41(11):1938-1945. [doi: 10.1111/acer.13485] [Medline: 28833270]
37. Legha RK, Moore L, Ling R, Novins D, Shore J. Telepsychiatry in an Alaska Native residential substance abuse treatment program. Telemed J E Health 2020 Jul 01;26(7):905-911. [doi: 10.1089/tmj.2019.0131] [Medline: 31804905]
38. Paris M, Silva M, Añez-Nava L, Jaramillo Y, Kiluk BD, Gordon MA, et al. Culturally adapted, web-based cognitive behavioral therapy for Spanish-speaking individuals with substance use disorders: a randomized clinical trial. Am J Public Health 2018 Nov;108(11):1535-1542. [doi: 10.2105/ajph.2018.304571]
39. Aharonovich E, Hatzenbuehler ML, Johnston B, O'Leary A, Morgenstern J, Wainberg ML, et al. A low-cost, sustainable intervention for drinking reduction in the HIV primary care setting. AIDS Care 2006 Aug 18;18(6):561-568. [doi: 10.1080/09540120500264134] [Medline: 16831783]
40. Muroff J, Robinson W, Chassler D, López LM, Gaitan E, Lundgren L, et al. Use of a smartphone recovery tool for Latinos with co-occurring alcohol and other drug disorders and mental disorders. J Dual Diagn 2017 Jul 10;13(4):280-290. [doi: 10.1080/15504263.2017.1348649] [Medline: 28692420]
41. Muroff J, Robinson W, Chassler D, López LM, Lundgren L, Guauque C, et al. An outcome study of the CASA-CHESS smartphone relapse prevention tool for Latinx Spanish-speakers with substance use disorders. Subst Use Misuse 2019 Mar 31;54(9):1438-1449. [doi: 10.1080/10826084.2019.1585457] [Medline: 30931681]

XSL•FO
RenderX

42. Osilla KC, D'Amico EJ, Díaz-Fuentes CM, Lara M, Watkins KE. Multicultural web-based motivational interviewing for clients with a first-time DUI offense. Cultur Divers Ethnic Minor Psychol 2012 Apr;18(2):192-202 [FREE Full text] [doi: 10.1037/a0027751] [Medline: 22506822]
43. Nayak MB, Korcha RA, Kaskutas LA, Avalos LA. Erratum: feasibility and acceptability of a novel, computerized screening and brief intervention (SBI) for alcohol and sweetened beverage use in pregnancy. BMC Pregnancy Childbirth 2015 Aug 18;15(1):178 [FREE Full text] [doi: 10.1186/s12884-015-0582-4] [Medline: 26285570]
44. Nayak MB, Kaskutas LA, Mericle AA. Randomized trial of an innovative electronic screening and brief intervention for reducing drinking among women of childbearing age. J Addict Med 2019;13(6):450-459 [FREE Full text] [doi: 10.1097/ADM.0000000000000518] [Medline: 30882553]
45. Comulada WS, Lightfoot M, Swendeman D, Grella C, Wu N. Compliance to cell phone-based EMA among Latino youth in outpatient treatment. J Ethn Subst Abuse 2015 Jun 26;14(3):232-250 [FREE Full text] [doi: 10.1080/15332640.2014.986354] [Medline: 26114764]
46. Millery M, Kleinman BP, Polissar NL, Millman RB, Scimeca M. Detoxification as a gateway to long-term treatment: assessing two interventions. J Subst Abuse Treat 2002 Oct;23(3):183-190. [doi: 10.1016/s0740-5472(02)00246-5] [Medline: 12392804]
47. Sanjuan PM, Pearson MR, Poremba C, Amaro HD, Leeman L. An ecological momentary assessment study examining posttraumatic stress disorder symptoms, prenatal bonding, and substance use among pregnant women. Drug Alcohol Depend 2019 Feb 01;195:33-39 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.11.019] [Medline: 30572290]
48. Trudeau KJ, Ainscough J, Charity S. Technology in treatment: are adolescents and counselors interested in online relapse prevention? Child Youth Care Forum 2012 Feb 1;41(1):57-71 [FREE Full text] [doi: 10.1007/s10566-011-9154-6] [Medline: 23930049]
49. Tolou-Shams M, Yonek J, Galbraith K, Bath E. Text messaging to enhance behavioral health treatment engagement among justice-involved youth: qualitative and user testing study. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e10904 [FREE Full text] [doi: 10.2196/10904] [Medline: 30950808]
50. Mullen J, Ryan SR, Mathias CW, Dougherty DM. Feasibility of a computer-assisted alcohol screening, brief intervention and referral to treatment program for DWI offenders. Addict Sci Clin Pract 2015 Nov 09;10(1):25 [FREE Full text] [doi: 10.1186/s13722-015-0046-1] [Medline: 26553284]
51. Burner E, Zhang M, Terp S, Ford Bench K, Lee J, Lam CN, et al. Feasibility and acceptability of a text message-based intervention to reduce overuse of alcohol in emergency department patients: controlled proof-of-concept trial. JMIR Mhealth Uhealth 2020 Jun 04;8(6):e17557 [FREE Full text] [doi: 10.2196/17557] [Medline: 32496203]
52. Aronson ID, Bennett A, Marsch LA, Bania TC. Mobile technology to increase HIV/HCV testing and overdose prevention/response among people who inject drugs. Front Public Health 2017 Aug 23;5:217 [FREE Full text] [doi: 10.3389/fpubh.2017.00217] [Medline: 28879174]
53. Braciszewski JM, Tzilos Wernette GK, Moore RS, Tran TB, Bock BC, Stout RL, et al. Developing a tailored substance use intervention for youth exiting foster care. Child Abuse Negl 2018 Mar;77:211-221 [FREE Full text] [doi: 10.1016/j.chiabu.2018.01.013] [Medline: 29367098]
54. Tzilos Wernette G, Plegue M, Kahler CW, Sen A, Zlotnick C. A pilot randomized controlled trial of a computer-delivered brief intervention for substance use and risky sex during pregnancy. J Womens Health (Larchmt) 2018 Jan;27(1):83-92 [FREE Full text] [doi: 10.1089/jwh.2017.6408] [Medline: 28981379]
55. Carroll KM, Ball SA, Martino S, Nich C, Babuscio TA, Nuro KF, et al. Computer-assisted delivery of cognitive-behavioral therapy for addiction: a randomized trial of CBT4CBT. Am J Psychiatry 2008 Jul;165(7):881-888 [FREE Full text] [doi: 10.1176/appi.ajp.2008.07111835] [Medline: 18450927]
56. Sugarman DE, Nich C, Carroll KM. Coping strategy use following computerized cognitive-behavioral therapy for substance use disorders. Psychol Addict Behav 2010 Dec;24(4):689-695 [FREE Full text] [doi: 10.1037/a0021584] [Medline: 21198228]
57. Kiluk BD, Nich C, Babuscio T, Carroll KM. Quality versus quantity: acquisition of coping skills following computerized cognitive-behavioral therapy for substance use disorders. Addiction 2010 Dec 20;105(12):2120-2127 [FREE Full text] [doi: 10.1111/j.1360-0443.2010.03076.x] [Medline: 20854334]
58. Kiluk BD, Nich C, Carroll KM. Relationship of cognitive function and the acquisition of coping skills in computer assisted treatment for substance use disorders. Drug Alcohol Depend 2011 Apr 01;114(2-3):169-176 [FREE Full text] [doi: 10.1016/j.drugalcdep.2010.09.019] [Medline: 21050679]
59. Thompson RG, Aivadyan C, Stohl M, Aharonovich E, Hasin DS. Smartphone application plus brief motivational intervention reduces substance use and sexual risk behaviors among homeless young adults: results from a randomized controlled trial. Psychol Addict Behav 2020 Sep;34(6):641-649 [FREE Full text] [doi: 10.1037/adb0000570] [Medline: 32175752]
60. Muench F, Weiss RA, Kuerbis A, Morgenstern J. Developing a theory driven text messaging intervention for addiction care with user driven content. Psychol Addict Behav 2013 Mar;27(1):315-321 [FREE Full text] [doi: 10.1037/a0029963] [Medline: 22963375]

XSL•FO
RenderX

61. Ingersoll KS, Farrell-Carnahan L, Cohen-Filipic J, Heckman CJ, Ceperich SD, Hettema J, et al. A pilot randomized clinical trial of two medication adherence and drug use interventions for HIV+ crack cocaine users. Drug Alcohol Depend 2011 Jul 01;116(1-3):177-187 [FREE Full text] [doi: 10.1016/j.drugalcdep.2010.12.016] [Medline: 21306837]
62. Alemi F, Haack M, Nemes S, Harge A, Baghi H. Impact of online counseling on drug use: a pilot study. Qual Manag Health Care 2010;19(1):62-69. [doi: 10.1097/QMH.0b013e3181d1394c] [Medline: 20042934]
63. Ondersma SJ, Chase SK, Svikis DS, Schuster CR. Computer-based brief motivational intervention for perinatal drug use. J Subst Abuse Treat 2005 Jun;28(4):305-312 [FREE Full text] [doi: 10.1016/j.jsat.2005.02.004] [Medline: 15925264]
64. Ondersma SJ, Svikis DS, Schuster CR. Computer-based brief intervention a randomized trial with postpartum women. Am J Prev Med 2007 Mar;32(3):231-238 [FREE Full text] [doi: 10.1016/j.amepre.2006.11.003] [Medline: 17236741]
65. Ondersma SJ, Grekin ER, Svikis D. The potential for technology in brief interventions for substance use, and during-session prediction of computer-delivered brief intervention response. Subst Use Misuse 2011 Dec 29;46(1):77-86 [FREE Full text] [doi: 10.3109/10826084.2011.521372] [Medline: 21190408]
66. Ondersma SJ, Svikis DS, Thacker LR, Beatty JR, Lockhart N. Computer-delivered screening and brief intervention (e-SBI) for postpartum drug use: a randomized trial. J Subst Abuse Treat 2014 Jan;46(1):52-59 [FREE Full text] [doi: 10.1016/j.jsat.2013.07.013] [Medline: 24051077]
67. Tzilos GK, Sokol RJ, Ondersma SJ. A randomized phase I trial of a brief computer-delivered intervention for alcohol use during pregnancy. J Womens Health (Larchmt) 2011 Oct;20(10):1517-1524 [FREE Full text] [doi: 10.1089/jwh.2011.2732] [Medline: 21823917]
68. Ondersma SJ, Beatty JR, Svikis DS, Strickler RC, Tzilos GK, Chang G, et al. Computer-delivered screening and brief intervention for alcohol use in pregnancy: a pilot randomized trial. Alcohol Clin Exp Res 2015 Jul 22;39(7):1219-1226 [FREE Full text] [doi: 10.1111/acer.12747] [Medline: 26010235]
69. Pollick SA, Beatty JR, Sokol RJ, Strickler RC, Chang G, Svikis DS, et al. Acceptability of a computerized brief intervention for alcohol among abstinent but at-risk pregnant women. Subst Abus 2015 Jan 01;36(1):13-20 [FREE Full text] [doi: 10.1080/08897077.2013.857631] [Medline: 24266770]
70. Ondersma SJ, Svikis DS, Thacker LR, Beatty JR, Lockhart N. A randomised trial of a computer-delivered screening and brief intervention for postpartum alcohol use. Drug Alcohol Rev 2016 Nov 23;35(6):710-718 [FREE Full text] [doi: 10.1111/dar.12389] [Medline: 27004474]
71. Martino S, Ondersma SJ, Forray A, Olmstead TA, Gilstad-Hayden K, Howell HB, et al. A randomized controlled trial of screening and brief interventions for substance misuse in reproductive health. Am J Obstet Gynecol 2018 Mar;218(3):322.e1-322.12 [FREE Full text] [doi: 10.1016/j.ajog.2017.12.005] [Medline: 29247636]
72. Forray A, Martino S, Gilstad-Hayden K, Kershaw T, Ondersma S, Olmstead T, et al. Assessment of an electronic and clinician-delivered brief intervention on cigarette, alcohol and illicit drug use among women in a reproductive healthcare clinic. Addict Behav 2019 Sep;96:156-163 [FREE Full text] [doi: 10.1016/j.addbeh.2019.05.007] [Medline: 31100713]
73. Loree AM, Yonkers KA, Ondersma SJ, Gilstad-Hayden K, Martino S. Comparing satisfaction, alliance and intervention components in electronically delivered and in-person brief interventions for substance use among childbearing-aged women. J Subst Abuse Treat 2019 Apr;99:1-7 [FREE Full text] [doi: 10.1016/j.jsat.2019.01.007] [Medline: 30797381]
74. Yonkers KA, Dailey JI, Gilstad-Hayden K, Ondersma SJ, Forray A, Olmstead TA, et al. Abstinence outcomes among women in reproductive health centers administered clinician or electronic brief interventions. J Subst Abuse Treat 2020 Jun;113:107995 [FREE Full text] [doi: 10.1016/j.jsat.2020.02.012] [Medline: 32359666]
75. Ondersma SJ, Svikis DS, Thacker C, Resnicow K, Beatty JR, Janisse J, et al. Computer-delivered indirect screening and brief intervention for drug use in the perinatal period: a randomized trial. Drug Alcohol Depend 2018 Apr 01;185:271-277 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.12.022] [Medline: 29482051]
76. Braciszewski JM, Tzilos Wernette GK, Moore RS, Bock BC, Stout RL, Chamberlain P. A pilot randomized controlled trial of a technology-based substance use intervention for youth exiting foster care. Child Youth Serv Rev 2018 Nov;94:466-476 [FREE Full text] [doi: 10.1016/j.childyouth.2018.08.011] [Medline: 31435121]
77. Ondersma SJ, Beatty JR, Puder KS, Janisse J, Svikis DS. Feasibility and acceptability of e-screening and brief intervention and tailored text messaging for marijuana use in pregnancy. J Womens Health (Larchmt) 2019 Sep 01;28(9):1295-1301 [FREE Full text] [doi: 10.1089/jwh.2018.7169] [Medline: 30657434]
78. Kiluk BD, Devore KA, Buck MB, Nich C, Frankforter TL, LaPaglia DM, et al. Randomized trial of computerized cognitive behavioral therapy for alcohol use disorders: efficacy as a virtual stand-alone and treatment add-on compared with standard outpatient treatment. Alcohol Clin Exp Res 2016 Sep 04;40(9):1991-2000 [FREE Full text] [doi: 10.1111/acer.13162] [Medline: 27488212]
79. Kiluk BD, Frankforter TL, Cusumano M, Nich C, Carroll KM. Change in DSM-5 alcohol use disorder criteria count and severity level as a treatment outcome indicator: results from a randomized trial. Alcohol Clin Exp Res 2018 Jun 05;42(8):1556-1563 [FREE Full text] [doi: 10.1111/acer.13807] [Medline: 29870051]
80. Kiluk BD, Nich C, Buck MB, Devore KA, Frankforter TL, LaPaglia DM, et al. Randomized clinical trial of computerized and clinician-delivered CBT in comparison with standard outpatient treatment for substance use disorders: primary within-treatment and follow-up outcomes. Am J Psychiatry 2018 Sep 01;175(9):853-863 [FREE Full text] [doi: 10.1176/appi.ajp.2018.17090978] [Medline: 29792052]

XSL•FO
RenderX

81. Roos CR, Carroll KM, Nich C, Frankforter T, Kiluk BD. Short- and long-term changes in substance-related coping as mediators of in-person and computerized CBT for alcohol and drug use disorders. Drug Alcohol Depend 2020 Jul 01;212:108044 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108044] [Medline: 32422538]
82. Jordan A, Babuscio T, Nich C, Carroll KM. A feasibility study providing substance use treatment in the Black church. J Subst Abuse Treat 2021 May;124:108218. [doi: 10.1016/j.jsat.2020.108218] [Medline: 33771290]
83. Hasin DS, Aharonovich E, O'Leary A, Greenstein E, Pavlicova M, Arunajadai S, et al. Reducing heavy drinking in HIV primary care: a randomized trial of brief intervention, with and without technological enhancement. Addiction 2013 Jul 17;108(7):1230-1240 [FREE Full text] [doi: 10.1111/add.12127] [Medline: 23432593]
84. Hasin DS, Aharonovich E, Greenstein E. HealthCall for the smartphone: technology enhancement of brief intervention in HIV alcohol dependent patients. Addict Sci Clin Pract 2014 Feb 17;9(1):5 [FREE Full text] [doi: 10.1186/1940-0640-9-5] [Medline: 24533631]
85. Aharonovich E, Greenstein E, O'Leary A, Johnston B, Seol SG, Hasin DS. HealthCall: technology-based extension of motivational interviewing to reduce non-injection drug use in HIV primary care patients - a pilot study. AIDS Care 2012 Mar 20;24(12):1461-1469 [FREE Full text] [doi: 10.1080/09540121.2012.663882] [Medline: 22428809]
86. Aharonovich E, Sarvet A, Stohl M, DesJarlais D, Tross S, Hurst T, et al. Reducing non-injection drug use in HIV primary care: a randomized trial of brief motivational interviewing, with and without HealthCall, a technology-based enhancement. J Subst Abuse Treat 2017 Mar;74:71-79 [FREE Full text] [doi: 10.1016/j.jsat.2016.12.009] [Medline: 28132704]
87. Aharonovich E, Stohl M, Cannizzaro D, Hasin D. HealthCall delivered via smartphone to reduce co-occurring drug and alcohol use in HIV-infected adults: a randomized pilot trial. J Subst Abuse Treat 2017 Dec;83:15-26 [FREE Full text] [doi: 10.1016/j.jsat.2017.09.013] [Medline: 29129192]
88. Cunningham RM, Walton MA, Goldstein A, Chermack ST, Shope JT, Bingham CR, et al. Three-month follow-up of brief computerized and therapist interventions for alcohol and violence among teens. Acad Emerg Med 2009 Nov 04;16(11):1193-1207 [FREE Full text] [doi: 10.1111/j.1553-2712.2009.00513.x] [Medline: 20053240]
89. Walton MA, Chermack ST, Shope JT, Bingham CR, Zimmerman MA, Blow FC, et al. Effects of a brief intervention for reducing violence and alcohol misuse among adolescents: a randomized controlled trial. JAMA 2010 Aug 04;304(5):527-535 [FREE Full text] [doi: 10.1001/jama.2010.1066] [Medline: 20682932]
90. Walton MA, Bohnert K, Resko S, Barry KL, Chermack ST, Zucker RA, et al. Computer and therapist based brief interventions among cannabis-using adolescents presenting to primary care: one year outcomes. Drug Alcohol Depend 2013 Oct 01;132(3):646-653 [FREE Full text] [doi: 10.1016/j.drugalcdep.2013.04.020] [Medline: 23711998]
91. Blow FC, Walton MA, Bohnert AS, Ignacio RV, Chermack S, Cunningham RM, et al. A randomized controlled trial of brief interventions to reduce drug use among adults in a low-income urban emergency department: the HealthiER You study. Addiction 2017 Aug 07;112(8):1395-1405. [doi: 10.1111/add.13773] [Medline: 28127808]
92. Waller R, Bonar EE, Fernandez AC, Walton MA, Chermack ST, Cunningham RM, et al. Exploring the components of an efficacious computer brief intervention for reducing marijuana use among adults in the emergency department. J Subst Abuse Treat 2019 Apr;99:67-72 [FREE Full text] [doi: 10.1016/j.jsat.2019.01.014] [Medline: 30797396]
93. Drislane LE, Waller R, Martz ME, Bonar EE, Walton MA, Chermack ST, et al. Therapist and computer-based brief interventions for drug use within a randomized controlled trial: effects on parallel trajectories of alcohol use, cannabis use and anxiety symptoms. Addiction 2020 Jan 24;115(1):158-169 [FREE Full text] [doi: 10.1111/add.14781] [Medline: 31400240]
94. Dennis ML, Scott CK, Funk RR, Nicholson L. A pilot study to examine the feasibility and potential effectiveness of using smartphones to provide recovery support for adolescents. Subst Abus 2015 Oct 01;36(4):486-492 [FREE Full text] [doi: 10.1080/08897077.2014.970323] [Medline: 25310057]
95. Scott CK, Johnson K, Dennis ML. Using mobile phone technology to provide recovery support for women offenders. Telemed J E Health 2013 Oct;19(10):767-771 [FREE Full text] [doi: 10.1089/tmj.2012.0309] [Medline: 23931730]
96. Scott CK, Dennis ML, Johnson KA, Grella CE. A randomized clinical trial of smartphone self-managed recovery support services. J Subst Abuse Treat 2020 Oct;117:108089 [FREE Full text] [doi: 10.1016/j.jsat.2020.108089] [Medline: 32811628]
97. Preston KL, Jobes ML, Phillips KA, Epstein DH. Real-time assessment of alcohol drinking and drug use in opioid-dependent polydrug users. Behav Pharmacol 2016 Oct;27(7):579-584 [FREE Full text] [doi: 10.1097/FBP.0000000000000250] [Medline: 27579810]
98. Preston KL, Kowalczyk WJ, Phillips KA, Jobes ML, Vahabzadeh M, Lin JL, et al. Context and craving during stressful events in the daily lives of drug-dependent patients. Psychopharmacology (Berl) 2017 Sep 8;234(17):2631-2642 [FREE Full text] [doi: 10.1007/s00213-017-4663-0] [Medline: 28593441]
99. Preston KL, Kowalczyk WJ, Phillips KA, Jobes ML, Vahabzadeh M, Lin JL, et al. Before and after: craving, mood, and background stress in the hours surrounding drug use and stressful events in patients with opioid-use disorder. Psychopharmacology (Berl) 2018 Sep 7;235(9):2713-2723 [FREE Full text] [doi: 10.1007/s00213-018-4966-9] [Medline: 29980821]
100. Preston KL, Schroeder JR, Kowalczyk WJ, Phillips KA, Jobes ML, Dwyer M, et al. End-of-day reports of daily hassles and stress in men and women with opioid-use disorder: relationship to momentary reports of opioid and cocaine use and

XSL•FO
RenderX

stress. Drug Alcohol Depend 2018 Dec 01;193:21-28 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.08.023] [Medline: 30336389]
101. Reback CJ, Grant DL, Fletcher JB, Branson CM, Shoptaw S, Bowers JR, et al. Text messaging reduces HIV risk behaviors among methamphetamine-using men who have sex with men. AIDS Behav 2012 Oct 19;16(7):1993-2002 [FREE Full text] [doi: 10.1007/s10461-012-0200-7] [Medline: 22610370]
102. Reback CJ, Fletcher JB, Shoptaw S, Mansergh G. Exposure to theory-driven text messages is associated with HIV risk reduction among methamphetamine-using men who have sex with men. AIDS Behav 2015 Jun 7;19 Suppl 2(S2):130-141. [doi: 10.1007/s10461-014-0985-7] [Medline: 25563501]
103. Reback CJ, Fletcher JB, Swendeman DA, Metzner M. Theory-based text-messaging to reduce methamphetamine use and HIV sexual risk behaviors among men who have sex with men: automated unidirectional delivery outperforms bidirectional peer interactive delivery. AIDS Behav 2019 Jan 13;23(1):37-47 [FREE Full text] [doi: 10.1007/s10461-018-2225-z] [Medline: 30006792]
104. Brooks AC, Ryder D, Carise D, Kirby KC. Feasibility and effectiveness of computer-based therapy in community treatment. J Subst Abuse Treat 2010 Oct;39(3):227-235. [doi: 10.1016/j.jsat.2010.06.003] [Medline: 20667682]
105. Budney AJ, Stanger C, Tilford JM, Scherer EB, Brown PC, Li Z, et al. Computer-assisted behavioral therapy and contingency management for cannabis use disorder. Psychol Addict Behav 2015 Sep;29(3):501-511 [FREE Full text] [doi: 10.1037/adb0000078] [Medline: 25938629]
106. Delrahim-Howlett K, Chambers CD, Clapp JD, Xu R, Duke K, Moyer 3rd RJ, et al. Web-based assessment and brief intervention for alcohol use in women of childbearing potential: a report of the primary findings. Alcohol Clin Exp Res 2011 Jul;35(7):1331-1338. [doi: 10.1111/j.1530-0277.2011.01469.x] [Medline: 21410488]
107. Tofighi B, Grossman E, Bereket S, D Lee J. Text message content preferences to improve buprenorphine maintenance treatment in primary care. J Addict Dis 2016 Dec 15;35(2):92-100. [doi: 10.1080/10550887.2015.1127716] [Medline: 26670868]
108. Tofighi B, Grazioli F, Bereket S, Grossman E, Aphinyanaphongs Y, Lee JD. Text message reminders for improving patient appointment adherence in an office-based buprenorphine program: a feasibility study. Am J Addict 2017 Sep 11;26(6):581-586. [doi: 10.1111/ajad.12557] [Medline: 28799677]
109. Tofighi B, McNeely J, Walzer D, Fansiwala K, Demner A, Chaudhury CS, et al. A telemedicine buprenorphine clinic to serve New York City: initial evaluation of the NYC public hospital system's initiative to expand treatment access during the COVID-19 pandemic. J Addict Med 2022;16(1):e40-e43 [FREE Full text] [doi: 10.1097/ADM.0000000000000809] [Medline: 33560696]
110. Osilla KC, Paddock SM, Leininger TJ, D'Amico EJ, Ewing BA, Watkins KE. A pilot study comparing in-person and web-based motivational interviewing among adults with a first-time DUI offense. Addict Sci Clin Pract 2015 Sep 03;10(1):18 [FREE Full text] [doi: 10.1186/s13722-015-0039-0] [Medline: 26334629]
111. Kennedy DP, Osilla KC, Hunter SB, Golinelli D, Maksabedian Hernandez E, Tucker JS. A pilot test of a motivational interviewing social network intervention to reduce substance use among housing first residents. J Subst Abuse Treat 2018 Mar;86:36-44 [FREE Full text] [doi: 10.1016/j.jsat.2017.12.005] [Medline: 29415849]
112. Dawson Rose C, Cuca YP, Kamitani E, Eng S, Zepf R, Draughon J, et al. Using interactive web-based screening, brief intervention and referral to treatment in an urban, safety-net HIV clinic. AIDS Behav 2015 Jun 12;19 Suppl 2(0 2):186-193 [FREE Full text] [doi: 10.1007/s10461-015-1078-y] [Medline: 25963770]
113. Dawson-Rose C, Draughon JE, Cuca Y, Zepf R, Huang E, Cooper BA, et al. Changes in specific substance involvement scores among SBIRT recipients in an HIV primary care setting. Addict Sci Clin Pract 2017 Dec 12;12(1):34 [FREE Full text] [doi: 10.1186/s13722-017-0101-1] [Medline: 29229000]
114. Saberi P, Dawson Rose C, Wootton AR, Ming K, Legnitto D, Jeske M, et al. Use of technology for delivery of mental health and substance use services to youth living with HIV: a mixed-methods perspective. AIDS Care 2020 Aug 28;32(8):931-939 [FREE Full text] [doi: 10.1080/09540121.2019.1622637] [Medline: 31132864]
115. Ingersoll KS, Dillingham RA, Hettema JE, Conaway M, Freeman J, Reynolds G, et al. Pilot RCT of bidirectional text messaging for ART adherence among nonurban substance users with HIV. Health Psychol 2015 Dec;34S:1305-1315 [FREE Full text] [doi: 10.1037/hea0000295] [Medline: 26651472]
116. Shrier LA, Rhoads A, Burke P, Walls C, Blood EA. Real-time, contextual intervention using mobile technology to reduce marijuana use among youth: a pilot study. Addict Behav 2014 Jan;39(1):173-180. [doi: 10.1016/j.addbeh.2013.09.028] [Medline: 24139665]
117. Shrier LA, Burke PJ, Kells M, Scherer EA, Sarda V, Jonestrask C, et al. Pilot randomized trial of MOMENT, a motivational counseling-plus-ecological momentary intervention to reduce marijuana use in youth. Mhealth 2018 Jul;4:29 [FREE Full text] [doi: 10.21037/mhealth.2018.07.04] [Medline: 30148142]
118. Swendeman D, Sumstine S, Aguilar E, Gorbach PM, Comulada WS, Gelberg L. Feasibility and acceptability of mobile phone self-monitoring and automated feedback to enhance telephone coaching for people with risky substance use: the QUIT-mobile pilot study. J Addict Med 2021 Apr 01;15(2):120-129 [FREE Full text] [doi: 10.1097/ADM.0000000000000707] [Medline: 32732684]

XSL•FO
RenderX

119. Frueh BC, Henderson S, Myrick H. Telehealth service delivery for persons with alcoholism. J Telemed Telecare 2005 Oct 01;11(7):372-375. [doi: 10.1177/1357633X0501100701]
120. Wechsberg WM, Browne FA, Poulton W, Ellerson RM, Simons-Rudolph A, Haller D. Adapting an evidence-based HIV prevention intervention for pregnant African-American women in substance abuse treatment. Subst Abuse Rehabil 2011;2:35-42 [FREE Full text] [doi: 10.2147/SAR.S16370] [Medline: 24474853]
121. Brusoski M, Rosen D. Health promotion using tablet technology with older adult African American methadone clients: a case study. J Technol Hum Serv 2015 May 26;33(2):119-132. [doi: 10.1080/15228835.2014.989297]
122. Welsh C. Acceptability of the use of cellular telephone and computer pictures/video for "pill counts" in buprenorphine maintenance treatment. J Opioid Manag 2016 May 01;12(3):217-220. [doi: 10.5055/jom.2016.0334] [Medline: 27435442]
123. Jaconis M, Santa Ana EJ, Killeen TK, Badour CL, Back SE. Concurrent treatment of PTSD and alcohol use disorder via telehealth in a female Iraq veteran. Am J Addict 2017 Mar 24;26(2):112-114 [FREE Full text] [doi: 10.1111/ajad.12481] [Medline: 28118514]
124. Palfai TP, Taylor JL, Saitz R, Kratzer MP, Otis JD, Bernstein JA. Development of a tailored, telehealth intervention to address chronic pain and heavy drinking among people with HIV infection: integrating perspectives of patients in HIV care. Addict Sci Clin Pract 2019 Aug 29;14(1):35 [FREE Full text] [doi: 10.1186/s13722-019-0165-1] [Medline: 31464645]
125. Castillo M, Conte B, Hinkes S, Mathew M, Na CJ, Norindr A, et al. Implementation of a medical student-run telemedicine program for medications for opioid use disorder during the COVID-19 pandemic. Harm Reduct J 2020 Nov 17;17(1):88 [FREE Full text] [doi: 10.1186/s12954-020-00438-4] [Medline: 33203460]
126. Mehtani NJ, Ristau JT, Snyder H, Surlyn C, Eveland J, Smith-Bernardin S, et al. COVID-19: a catalyst for change in telehealth service delivery for opioid use disorder management. Subst Abus 2021 Apr 01;42(2):205-212. [doi: 10.1080/08897077.2021.1890676] [Medline: 33684331]
127. O'Gurek DT. Designing and evaluating COVID-19 protocols for an office-based opioid treatment program in an urban underserved setting. J Am Board Fam Med 2021 Feb;34(Suppl):S136-S140 [FREE Full text] [doi: 10.3122/jabfm.2021.S1.200207] [Medline: 33622828]
128. Saladin ME, Brady KT, Graap K, Rothbaum BO. A preliminary report on the use of virtual reality technology to elicit craving and cue reactivity in cocaine dependent individuals. Addict Behav 2006 Oct;31(10):1881-1894. [doi: 10.1016/j.addbeh.2006.01.004] [Medline: 16516397]
129. Gordon MS, Carswell SB, Schadegg M, Mangen K, Merkel K, Tangires S, et al. Avatar-assisted therapy: a proof-of-concept pilot study of a novel technology-based intervention to treat substance use disorders. Am J Drug Alcohol Abuse 2017 Sep 17;43(5):518-524. [doi: 10.1080/00952990.2017.1280816] [Medline: 28635345]
130. Freedman MJ, Lester KM, McNamara C, Milby JB, Schumacher JE. Cell phones for ecological momentary assessment with cocaine-addicted homeless patients in treatment. J Subst Abuse Treat 2006 Mar;30(2):105-111. [doi: 10.1016/j.jsat.2005.10.005] [Medline: 16490673]
131. Linas BS, Latkin C, Genz A, Westergaard RP, Chang LW, Bollinger RC, et al. Utilizing mHealth methods to identify patterns of high risk illicit drug use. Drug Alcohol Depend 2015 Jun 01;151:250-257 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.03.031] [Medline: 25920799]
132. Yang C, Linas B, Kirk G, Bollinger R, Chang L, Chander G, et al. Feasibility and acceptability of smartphone-based ecological momentary assessment of alcohol use among African American men who have sex with men in Baltimore. JMIR Mhealth Uhealth 2015 Jun 17;3(2):e67 [FREE Full text] [doi: 10.2196/mhealth.4344] [Medline: 26085078]
133. Buu A, Massey LS, Walton MA, Cranford JA, Zimmerman MA, Cunningham RM. Assessment methods and schedules for collecting daily process data on substance use related health behaviors: a randomized control study. Drug Alcohol Depend 2017 Sep 01;178:159-164 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.05.003] [Medline: 28651151]
134. Suffoletto B, Chung T, Muench F, Monti P, Clark DB. A text message intervention with adaptive goal support to reduce alcohol consumption among non-treatment-seeking young adults: non-randomized clinical trial with voluntary length of enrollment. JMIR Mhealth Uhealth 2018 Feb 16;6(2):e35 [FREE Full text] [doi: 10.2196/mhealth.8530] [Medline: 29453191]
135. Moore AA, Lake JE, Glasner S, Karlamangla A, Kuerbis A, Preciado D, et al. Establishing the feasibility, acceptability and preliminary efficacy of a multi-component behavioral intervention to reduce pain and substance use and improve physical performance in older persons living with HIV. J Subst Abuse Treat 2019 May;100:29-38 [FREE Full text] [doi: 10.1016/j.jsat.2019.02.003] [Medline: 30898325]
136. Glasner S, Chokron Garneau H, Ang A, Ray L, Venegas A, Rawson R, et al. Preliminary efficacy of a cognitive behavioral therapy text messaging intervention targeting alcohol use and antiretroviral therapy adherence: a randomized clinical trial. PLoS One 2020 Mar 12;15(3):e0229557 [FREE Full text] [doi: 10.1371/journal.pone.0229557] [Medline: 32163431]
137. Spohr SA, Taxman FS, Walters ST. The relationship between electronic goal reminders and subsequent drug use and treatment initiation in a criminal justice setting. Addict Behav 2015 Dec;51:51-56 [FREE Full text] [doi: 10.1016/j.addbeh.2015.07.005] [Medline: 26217929]
138. Steers ML, Coffman AD, Wickham RE, Bryan JL, Caraway L, Neighbors C. Evaluation of alcohol-related personalized normative feedback with and without an injunctive message. J Stud Alcohol Drugs 2016 Mar;77(2):337-342 [FREE Full text] [doi: 10.15288/jsad.2016.77.337] [Medline: 26997192]

XSL•FO
RenderX

139. Ditre JW, LaRowe LR, Vanable PA, De Vita MJ, Zvolensky MJ. Computer-based personalized feedback intervention for cigarette smoking and prescription analgesic misuse among persons living with HIV (PLWH). Behav Res Ther 2019 Apr;115:83-89 [FREE Full text] [doi: 10.1016/j.brat.2018.10.013] [Medline: 30389092]
140. Gryczynski J, Mitchell SG, Schwartz RP, Dusek K, O'Grady KE, Cowell AJ, et al. Computer- vs. nurse practitioner-delivered brief intervention for adolescent marijuana, alcohol, and sex risk behaviors in school-based health centers. Drug Alcohol Depend 2021 Jan 01;218:108423 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108423] [Medline: 33307377]
141. Moore BA, Rosen MI, Wang Y, Shen J, Ablondi K, Sullivan A, et al. A remotely-delivered CBT and contingency management therapy for substance using people with HIV. AIDS Behav 2015 Jun 3;19 Suppl 2(0 2):156-162 [FREE Full text] [doi: 10.1007/s10461-014-0990-x] [Medline: 25645326]
142. Beckham JC, Adkisson KA, Hertzberg J, Kimbrel NA, Budney AJ, Stephens RS, et al. Mobile contingency management as an adjunctive treatment for co-morbid cannabis use disorder and cigarette smoking. Addict Behav 2018 Apr;79:86-92 [FREE Full text] [doi: 10.1016/j.addbeh.2017.12.007] [Medline: 29268146]
143. Babson KA, Ramo DE, Baldini L, Vandrey R, Bonn-Miller MO. Mobile app-delivered cognitive behavioral therapy for insomnia: feasibility and initial efficacy among veterans with cannabis use disorders. JMIR Res Protoc 2015 Jul 17;4(3):e87 [FREE Full text] [doi: 10.2196/resprot.3852] [Medline: 26187404]
144. Mays VM, Jones AL, Delany-Brumsey A, Coles C, Cochran SD. Perceived discrimination in health care and mental health/substance abuse treatment among Blacks, Latinos, and Whites. Med Care 2017 Feb;55(2):173-181 [FREE Full text] [doi: 10.1097/MLR.0000000000000638] [Medline: 27753743]
145. Guerrero EG, Marsh JC, Khachikian T, Amaro H, Vega WA. Disparities in Latino substance use, service use, and treatment: implications for culturally and evidence-based interventions under health care reform. Drug Alcohol Depend 2013 Dec 15;133(3):805-813. [doi: 10.1016/j.drugalcdep.2013.07.027] [Medline: 23953657]
146. Espey DK, Jim MA, Cobb N, Bartholomew M, Becker T, Haverkamp D, et al. Leading causes of death and all-cause mortality in American Indians and Alaska Natives. Am J Public Health 2014 Jun;104(S3):S303-S311. [doi: 10.2105/ajph.2013.301798]
147. Srinivasan S, Williams SD. Transitioning from health disparities to a health equity research agenda: the time is now. Public Health Rep 2014 Jan 01;129(1_suppl2):71-76. [doi: 10.1177/00333549141291s213]
148. McFarling UL. 'Health equity tourists': how white scholars are colonizing research on health disparities. STAT. 2021. URL: https://www.statnews.com/2021/09/23/health-equity-tourists-white-scholars-colonizing-health-disparities-research/#:~:text=While%20structural%20racism%20is%20starting,factors%20like%20%E2%80%9Clack%20of%20trust%E2%80%9D [accessed 2024-04-29]
149. Turner BE, Steinberg JR, Weeks BT, Rodriguez F, Cullen MR. Race/ethnicity reporting and representation in US clinical trials: a cohort study. Lancet Reg Health Am 2022 Jul;11:100252 [FREE Full text] [doi: 10.1016/j.lana.2022.100252] [Medline: 35875251]
150. Sanders Thompson VL, Ackermann N, Bauer KL, Bowen DJ, Goodman MS. Strategies of community engagement in research: definitions and classifications. Transl Behav Med 2021 Mar 16;11(2):441-451 [FREE Full text] [doi: 10.1093/tbm/ibaa042] [Medline: 32421173]
151. Sprague Martinez L, Chassler D, Lobb R, Hakim D, Pamphile J, Battaglia TA. A discussion among deans on advancing community engaged research. Clin Transl Sci 2023 Apr 31;16(4):557-563 [FREE Full text] [doi: 10.1111/cts.13478] [Medline: 36707736]
152. Johnson DJ, Brunette MF, Goodman DJ, Adams M, Bryer C, Doherty JR, et al. Promoting community stakeholder engagement in research on treatment for pregnant women with opioid use disorder. J Comp Eff Res 2022 Oct;11(15):1085-1094 [FREE Full text] [doi: 10.2217/cer-2022-0090] [Medline: 36047333]
153. Yarborough M, Edwards K, Espinoza P, Geller G, Sarwal A, Sharp R, et al. Relationships hold the key to trustworthy and productive translational science: recommendations for expanding community engagement in biomedical research. Clin Transl Sci 2013 Aug 14;6(4):310-313 [FREE Full text] [doi: 10.1111/cts.12022] [Medline: 23919367]
154. Ahmed SM, Palermo AG. Community engagement in research: frameworks for education and peer review. Am J Public Health 2010 Aug;100(8):1380-1387. [doi: 10.2105/ajph.2009.178137]
155. Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science: unifying the discipline to improve the public health. Clin Psychol Sci 2014 Jan 01;2(1):22-34 [FREE Full text] [doi: 10.1177/2167702613497932] [Medline: 25821658]
156. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897 [FREE Full text] [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
157. Pacheco CM, Daley SM, Brown T, Filippi M, Greiner KA, Daley CM. Moving forward: breaking the cycle of mistrust between American Indians and researchers. Am J Public Health 2013 Dec;103(12):2152-2159. [doi: 10.2105/ajph.2013.301480]
158. Enhancing health disparities research related to substance use and addiction: research gaps and opportunities. National Institutes of Health (NIH). 2021. URL: https://ctnlibrary.org/NIDAenhancinghealthdisparities.pdf [accessed 2024-04-29]

XSL•FO
RenderX

159. DiClemente CC, Corno CM, Graydon MM, Wiprovnick AE, Knoblach DJ. Motivational interviewing, enhancement, and brief interventions over the last decade: a review of reviews of efficacy and effectiveness. Psychol Addict Behav 2017 Dec;31(8):862-887. [doi: 10.1037/adb0000318] [Medline: 29199843]
160. Goldman DL. A theory of health disparities and medical technology. Contribut Econ Anal Policy 2005;4(1):23 [FREE Full text] [doi: 10.2202/1538-0645.1395]
161. Senteio C, Veinot T. Trying to make things right: adherence work in high-poverty, African American neighborhoods. Qual Health Res 2014 Dec 11;24(12):1745-1756. [doi: 10.1177/1049732314549027] [Medline: 25212857]
162. Brown CE, Curtis JR, Doll KM. A race-conscious approach toward research on racial inequities in palliative care. J Pain Symptom Manage 2022 May;63(5):e465-e471 [FREE Full text] [doi: 10.1016/j.jpainsymman.2021.11.012] [Medline: 34856335]
163. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
164. Ramos GG, West AE, Begay C, Telles VM, D'Isabella J, Antony V, et al. Substance use disorder and homelessness among American Indians and Alaska Natives in California. J Ethn Subst Abuse 2023 Aug 02;22(2):350-371. [doi: 10.1080/15332640.2021.1952125] [Medline: 34339341]
165. Roman PM, Ducharme LJ, Knudsen HK. Patterns of organization and management in private and public substance abuse treatment programs. J Subst Abuse Treat 2006 Oct;31(3):235-243. [doi: 10.1016/j.jsat.2006.06.017] [Medline: 16996386]
166. Cislo AM, Trestman R. Challenges and solutions for conducting research in correctional settings: the U.S. experience. Int J Law Psychiatry 2013 May;36(3-4):304-310. [doi: 10.1016/j.ijlp.2013.04.002] [Medline: 23683885]
167. Ferguson WJ, Cloud D, Spaulding AC, Shelton D, Trestman RL, Altice FL, et al. A call to action: a blueprint for academic health sciences in the era of mass incarceration. J Health Care Poor Underserved 2016 May;27(2A):5-17. [doi: 10.1353/hpu.2016.0051] [Medline: 27133508]
168. USDHHS Indian Health Service. 2014. The Trends in Indian health: 2014 Edition.: Indian Health Service, US Department of Health and Human Services; 2015 Mar. URL: https://www.ihs.gov/dps/publications/trends2014/ [accessed 2023-08-25]
169. Rural America at a glance, 2021 edition. Economic Research Service, US Department of Agriculture. URL: https://www.ers.usda.gov/webdocs/publications/102576/eib-230.pdf?v=9165.8 [accessed 2023-08-25]
170. Burlew K, Larios S, Suarez-Morales L, Holmes B, Venner K, Chavez R. Increasing ethnic minority participation in substance abuse clinical trials: lessons learned in the National Institute on Drug Abuse's Clinical Trials Network. Cultur Divers Ethnic Minor Psychol 2011 Oct;17(4):345-356 [FREE Full text] [doi: 10.1037/a0025668] [Medline: 21988575]
171. Sankaré IC, Bross R, Brown AF, Del Pino HE, Jones LF, Morris DM, et al. Strategies to build trust and recruit African American and Latino community residents for health research: a cohort study. Clin Transl Sci 2015 Oct 21;8(5):412-420 [FREE Full text] [doi: 10.1111/cts.12273] [Medline: 26094679]
172. Das LT, Kutscher E. Stimulant use disorders in the United States- is another epidemic on the horizon? JAMA Health Forum 2020 Dec 01;1(12):e201486. [doi: 10.1001/jamahealthforum.2020.1486] [Medline: 36218481]
173. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice. The Health Effects of Cannabis and Cannabinoids: The Current State of Evidence and Recommendations for Research. Washington, DC: National Academies Press; 2017.
174. Sweeney MM, Holtyn AF, Stitzer ML, Gastfriend DR. Practical technology for expanding and improving substance use disorder treatment: telehealth, remote monitoring, and digital health interventions. Psychiatr Clin North Am 2022 Sep;45(3):515-528 [FREE Full text] [doi: 10.1016/j.psc.2022.05.006] [Medline: 36055736]
175. Williams EC, Samet JH. Shifts at the Helm: gratitude, re-commitment to our work, and a call for addictions disparities research. Addict Sci Clin Pract 2022 Feb 18;17(1):12 [FREE Full text] [doi: 10.1186/s13722-022-00290-w] [Medline: 35180895]
176. Biroscak BJ, Pantalon MV, Dziura JD, Hersey DP, Vaca FE. Use of non-face-to-face modalities for emergency department screening, brief intervention, and referral to treatment (ED-SBIRT) for high-risk alcohol use: a scoping review. Subst Abus 2019 Jan 01;40(1):20-32. [doi: 10.1080/08897077.2018.1550465] [Medline: 30829126]
177. Colbert S, Thornton L, Richmond R. Smartphone apps for managing alcohol consumption: a literature review. Addict Sci Clin Pract 2020 May 07;15(1):17. [doi: 10.1186/s13722-020-00190-x] [Medline: 32381062]
178. Colonna R, Alvarez L. Characteristics of mobile-based brief interventions targeting substance use among youth: a rapid review. J Subst Use 2022 Mar 13;28(3):293-304. [doi: 10.1080/14659891.2022.2051622]
179. Crane D, Garnett C, Brown J, West R, Michie S. Behavior change techniques in popular alcohol reduction apps: content analysis. J Med Internet Res 2015 May 14;17(5):e118 [FREE Full text] [doi: 10.2196/jmir.4060] [Medline: 25977135]
180. McGinnes RA, Hutton JE, Weiland TJ, Fatovich DM, Egerton-Warburton D. Review article: effectiveness of ultra-brief interventions in the emergency department to reduce alcohol consumption: a systematic review. Emerg Med Australas 2016 Dec 26;28(6):629-640. [doi: 10.1111/1742-6723.12624] [Medline: 27459669]
181. Tofighi B, Chemi C, Ruiz-Valcarcel J, Hein P, Hu L. Smartphone apps targeting alcohol and illicit substance use: systematic search in in commercial app stores and critical content analysis. JMIR Mhealth Uhealth 2019 Apr 22;7(4):e11831 [FREE Full text] [doi: 10.2196/11831] [Medline: 31008713]

XSL•FO
RenderX

182. Wilson H, Stoyanov SR, Gandabhai S, Baldwin A. The quality and accuracy of mobile apps to prevent driving after drinking alcohol. JMIR Mhealth Uhealth 2016 Aug 08;4(3):e98 [FREE Full text] [doi: 10.2196/mhealth.5961] [Medline: 27502956]
183. Colonna R, Tucker P, Holmes J, Wilson J, Alvarez L. Mobile-based brief interventions targeting cannabis-impaired driving among youth: a Delphi study. J Subst Abuse Treat 2022 Oct;141:108802. [doi: 10.1016/j.jsat.2022.108802] [Medline: 35599094]
184. Furr-Holden D, Milam AJ, Wang L, Sadler R. African Americans now outpace whites in opioid-involved overdose deaths: a comparison of temporal trends from 1999 to 2018. Addiction 2021 Mar 15;116(3):677-683. [doi: 10.1111/add.15233] [Medline: 32852864]
185. Barnett ML, Meara E, Lewinson T, Hardy B, Chyn D, Onsando M, et al. Racial inequality in receipt of medications for opioid use disorder. N Engl J Med 2023 May 11;388(19):1779-1789. [doi: 10.1056/nejmsa2212412]
186. Iheanacho T, Payne K, Tsai J. Mobile, community-based buprenorphine treatment for veterans experiencing homelessness with opioid use disorder: a pilot, feasibility study. Am J Addict 2020 Nov 04;29(6):485-491. [doi: 10.1111/ajad.13055] [Medline: 32367557]
187. Saraiya TC, Swarbrick M, Franklin L, Kass S, Campbell AN, Hien DA. Perspectives on trauma and the design of a technology-based trauma-informed intervention for women receiving medications for addiction treatment in community-based settings. J Subst Abuse Treat 2020 May;112:92-101 [FREE Full text] [doi: 10.1016/j.jsat.2020.01.011] [Medline: 32199551]
188. Holtyn AF, Toegel F, Novak MD, Leoutsakos JM, Fingerhood M, Silverman K. Remotely delivered incentives to promote buprenorphine treatment engagement in out-of-treatment adults with opioid use disorder. Drug Alcohol Depend 2021 Aug 01;225:108786 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108786] [Medline: 34087746]
189. Lynch MJ, Houck P, Meyers J, Schuster J, Yealy DM. Use of a telemedicine bridge clinic to engage patients in opioid use disorder treatment. J Addict Med 2022;16(5):584-587. [doi: 10.1097/ADM.0000000000000967] [Medline: 35258040]
190. Second temporary extension of COVID-19 telemedicine flexibilities for prescription of controlled medications. Federal Register. 2023 Oct 10. URL: https://www.federalregister.gov/documents/2023/10/10/2023-22406/second-temporary-extension-of-covid-19-telemedicine-flexibilities-for-prescription-of-controlled [accessed 2024-04-29]
191. Ford CL, Airhihenbuwa CO. The public health critical race methodology: praxis for antiracism research. Soc Sci Med 2010 Oct;71(8):1390-1398. [doi: 10.1016/j.socscimed.2010.07.030] [Medline: 20822840]
192. Dean HD, Williams KM, Fenton KA. From theory to action: applying social determinants of health to public health practice. Public Health Rep 2013 Nov 01;128(6_suppl3):1-4. [doi: 10.1177/00333549131286s301]
193. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
194. Patrick H, Williams GC. Self-determination theory: its application to health behavior and complementarity with motivational interviewing. Int J Behav Nutr Phys Act 2012 Mar 02;9(1):18 [FREE Full text] [doi: 10.1186/1479-5868-9-18] [Medline: 22385676]
195. Hamideh D, Nebeker C. The digital health landscape in addiction and substance use research: will digital health exacerbate or mitigate health inequities in vulnerable populations? Curr Addict Rep 2020 Sep 01;7(3):317-332. [doi: 10.1007/S40429-020-00325-9]
196. Weaver A, Himle JA, Taylor RJ, Matusko NN, Abelson JM. Urban vs rural residence and the prevalence of depression and mood disorder among African American women and non-Hispanic White women. JAMA Psychiatry 2015 Jun 01;72(6):576-583 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.10] [Medline: 25853939]
197. Harris RA, Mandell DS. Years of life lost to drug overdose among Black female individuals in the US, 2015-2021. JAMA Psychiatry 2023 Mar 01;80(3):276-277 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.4194] [Medline: 36515939]
198. Heard E, Fitzgerald L, Wigginton B, Mutch A. Applying intersectionality theory in health promotion research and practice. Health Promot Int 2020 Aug 01;35(4):866-876. [doi: 10.1093/heapro/daz080] [Medline: 31390472]
199. Husain L, Greenhalgh T, Hughes G, Finlay T, Wherton J. Desperately seeking intersectionality in digital health disparity research: narrative review to inform a richer theorization of multiple disadvantage. J Med Internet Res 2022 Dec 07;24(12):e42358 [FREE Full text] [doi: 10.2196/42358] [Medline: 36383632]
200. Crenshaw K. Mapping the margins: intersectionality, identity politics, and violence against women of color. Stan Law Rev 1991 Jul;43(6):1241. [doi: 10.2307/1229039]
201. Tinner L, Holman D, Ejegi-Memeh S, Laverty AA. Use of intersectionality theory in interventional health research in high-income countries: a scoping review. Int J Environ Res Public Health 2023 Jul 15;20(14):6370 [FREE Full text] [doi: 10.3390/ijerph20146370] [Medline: 37510601]
202. Science of behavior change. National Institues of Health OoSC-TCF. 2023. URL: https://commonfund.nih.gov/behaviorchange [accessed 2024-04-29]
203. Nora's blog: social determinants of health can't be extricated from addiction science. National Institute on Drug Abuse. URL: https://nida.nih.gov/about-nida/noras-blog/2023/04/social-determinants-health-cant-be-extricated-addiction-science [accessed 2024-04-29]
204. Lett E, Adekunle D, McMurray P, Asabor EN, Irie W, Simon MA, et al. Health equity tourism: ravaging the justice landscape. J Med Syst 2022 Feb 12;46(3):17. [doi: 10.1007/s10916-022-01803-5] [Medline: 35150324]

XSL•FO
RenderX

205. Tod D, Booth A, Smith B. Critical appraisal. Int Rev Sport Exerc Psychol 2021 Jul 12;15(1):52-72. [doi: 10.1080/1750984x.2021.1952471]
206. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014 Jan;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]
207. McGhee H. The Sum of Us: What Racism Costs Everyone and How We Can Prosper Together. New York City, NY: Penguin Random House; 2021.
208. Benjamin R. Race After Technology: Abolitionist Tools for the New Jim Code. New York, NY: Polity; 2019.

Abbreviations

CBT: cognitive behavioral therapy
CBT4CBT: Computer-Based Training for Cognitive Behavioral Therapy
CM: contingency management
EMA: ecological momentary assessment
MES: Motivation Enhancement System
NIH: National Institutes of Health
OUD: opioid use disorder
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SES: socioeconomic status
TAU: treatment as usual
TBI: technology-based intervention
URM: underrepresented minority

Edited by KJ Craig; submitted 18.10.23; peer-reviewed by A Bennett, N Reed; comments to author 26.02.24; revised version received 21.04.24; accepted 11.10.24; published 03.12.24.

Please cite as:
Moore SK, Boggis JS, Gauthier PR, Lambert-Harris CA, Hichborn EG, Bell KD, Saunders EC, Montgomery L, Murphy EI, Turner AM, Agosti N, McLeman BM, Marsch LA
Technology-Based Interventions for Substance Use Treatment Among People Who Identify as African American or Black, Hispanic or Latino, and American Indian or Alaska Native: Scoping Review
J Med Internet Res 2024;26:e53685
URL: https://www.jmir.org/2024/1/e53685
doi:10.2196/53685
PMID:

©Sarah K Moore, Jesse S Boggis, Phoebe R Gauthier, Chantal A Lambert-Harris, Emily G Hichborn, Kathleen D Bell, Elizabeth C Saunders, LaTrice Montgomery, Eilis I Murphy, Avery M Turner, Nico Agosti, Bethany M McLeman, Lisa A Marsch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of 10 Weeks of Walking With Mobile Step-Tracking Apps on Body Composition, Fitness, and Psychological State in Adolescents Who Are Overweight and Obese: Randomized Controlled Trial

Adrián Mateo-Orcajada[1], PhD; Cristina M Ponce-Ramírez[2], MS; Lucía Abenza-Cano[1], PhD; Raquel Vaquero-Cristóbal[2], PhD

[1]Facultad de Deporte, Universidad Católica de Murcia (UCAM), Murcia, Spain

[2]Research Group Movement Sciences and Sport (MS&SPORT), Department of Physical Activity and Sport, Faculty of Sport Sciences, University of Murcia, Murcia, Spain

Corresponding Author:
Lucía Abenza-Cano, PhD
Facultad de Deporte
Universidad Católica de Murcia (UCAM)
Avenida de los Jerónimos
Murcia, 30107
Spain
Phone: 34 968 278 824
Email: labenza@ucam.edu

Abstract

Background: In recent decades, physical activity intervention programs have been developed to reduce overweight and obesity in adolescents. However, this population is considered hard to reach in physical activity programs due to lack of adherence and poor results. Interventions with mobile phones in the adolescent population with normal weight have shown benefits, so this line of research may provide benefits in adolescents with overweight or obesity, although it has not yet been explored in the scientific literature.

Objective: This study aims to determine the changes produced by a 10-week intervention promoted during school lessons on physical education using step tracker mobile apps in out-of-school hours on physical activity, adherence to the Mediterranean diet, body composition, and the physical condition of adolescents who are overweight and obese, and to analyze the changes achieved by the 10-week intervention on the psychological state of adolescents who are overweight and obese.

Methods: The study was based on a randomized controlled trial with an initial sample of 50 adolescents aged between 12 and 16 years (from the first to the fourth years of compulsory secondary education), whose body composition, physical activity level, physical condition, and psychological state were measured. Participants were divided into an experimental group (EG) and a control group (CG), where the EG performed a series of walking steps with a mobile app in their free time outside physical education classes. Adolescents in the CG continued to perform their physical activities as normal but did not use any mobile apps. Inclusion in the EG and CG was randomized, and the researchers were blinded.

Results: An increase was found in the EG in corrected arm girth (mean difference –0.46; P=.05), curl-up repetitions (mean difference –6.35; P=.02) and push-up repetitions (mean difference –2.27; P=.04) after the intervention. In the CG, there was a significant increase in hip girth (mean difference –1.37; P=.05), corrected thigh girth (mean difference –1.28; P=.04), and muscle mass (mean difference –0.87; P=.04), as well as a significant decrease in competence (mean difference 3.08; P=.03). The covariates gender and age showed an effect on corrected arm girth (gender: P=.04), curl-up repetitions (gender: P=.04) and push-up repetitions (gender: P=.04) in the EG; while in the CG it affected corrected thigh girth (gender: P=.04), adherence to the Mediterranean diet (gender: P=.04 and age: P=.047) competence (gender: P=.04 and age: P=.04) and relatedness (gender: P=.05 and age: P=.04). No significant differences were found when comparing changes in the CG and EG.

Conclusions: A 10-week program of mobile app use by adolescents who are overweight and obese for physical activity outside of school hours does not appear effective in producing improvements in body composition, physical fitness, or adequate psychological state as it does not appear to significantly increase physical activity.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT06089876; http://clinicaltrials.gov/ct2/show/NCT06089876

(J Med Internet Res 2024;26:e55243) doi:10.2196/55243

KEYWORDS

adolescents; obesity; physical activity; overweight; mobile app; physical education

Introduction

Development and Prevalence of Overweight and Obesity in Childhood and Adolescents

Overweight and obesity, understood as an accumulation of excessive fat, are the second leading cause of preventable and avoidable mortality in countries without limited resources [1], with >340 million children and adolescents affected by this disease worldwide [1]. This pathology increases the risk of experiencing other chronic diseases at an early age and during adulthood, such as arterial hypertension [2], diabetes [3], or various types of cancer [4]. This pathology increases the probability of experiencing psychological disorders, such as stress, depression, and anxiety [5], and it decreases life expectancy [6]. Because of the reasons mentioned earlier, overweightness and obesity at an early age are already considered by the World Health Organization (WHO) as a pandemic [7]. This is because it has a high incidence in countries without limited resources and especially in Spain, a country that has been at the top of the rankings of childhood obesity in the last decade, leading them on some occasions [8-10].

Factors Influencing the Development of Overweight and Obesity in Childhood and Adolescents

This worrying situation was compounded by the COVID-19 pandemic. The period of confinement that took place in most countries brought with it a series of changes in the lifestyle habits of adolescents, such as the impossibility of engaging in physical-sports activities [11] and the increased use of technological devices [12,13]. Although the restrictions caused by the COVID-19 pandemic have been lifted, the lifestyle habits generated during the pandemic have become established in the postpandemic adolescent population. More specifically, the level of physical activity of prepandemic adolescents has not recovered [14]. More than 80% of adolescents aged between 12 and 16 years old do not meet daily physical activity recommendations [14], performing <60 minutes per day of aerobic physical activity at moderate or vigorous intensities [15]. This becomes even more relevant when considering that adolescence is a critical stage with respect to the practice of physical activity, with a high rate of sports abandonment during this period [16]. This results in an increase in the percentage of adolescents whose weekly physical activity is limited to that performed in physical education sessions within the school context [17].

In addition, the abuse of new technologies during the COVID-19 pandemic has remained at worrying levels in the post–COVID-19 period [18], favoring an increase in the adolescent population that is considered sedentary [19]. Sedentary behavior is defined as a waking behavior characterized by energy expenditure of <1.5 metabolic task equivalents while maintaining a sitting or reclining posture [20]. The main problem with sedentary behavior lies in the negative health consequences for adolescents [21]. This becomes even more relevant when considering that even if adolescents comply with the physical activity recommendations of 60 minutes per day, they can still experience the negative effects of sedentary behavior if they accumulate many hours per day in sedentary activities [22]. In this sense, male participants spend 507 minutes a day on average in sedentary activities, while female participants spend 523 minutes a day on average, and this is even more pronounced on weekend days [23]. Furthermore, it has also been found that up to 93.8% of adolescent boys and 87.2% of adolescent girls spend >2 hours a day in front of a screen [13]. Mobile phone use has been shown to be a determining factor, with 10% to 16% of adolescents showing problematic use of this device. This is understood as a dependence or need to be connected to the device, as well as difficulties in disconnecting or setting limits on the time dedicated to its use, which interferes with health, well-being, and leisure time behaviors [24,25].

Impact of Unhealthy Habits on Adolescent Health

The decline in physical activity is all the more significant because it is related to the worsening of other healthy lifestyle habits, such as diet, with adolescents who were less physically active having poorer adherence to patterns, such as the adherence to the Mediterranean diet (AMD) [26,27]. This becomes relevant when it is known that AMD is associated with a reduction in cardiovascular risk of up to 38% [28] and a reduction in the risk of type 2 diabetes mellitus of 19% [29]. Moreover, increased adherence significantly reduces all-cause mortality, as well as the incidence of cancer, Parkinson disease, and Alzheimer disease [30]. Therefore, AMD, as well as the practice of physical activity, plays a decisive role in reducing the prevalence of noncommunicable diseases [31].

However, in recent years, there has been an increase in sedentary time among adolescents and a decrease in physical activity, which, in combination with other unhealthy habits, has led to a greater accumulation of body fat [32], increasing the likelihood of being overweight or obese [33]. Thus, following COVID-19, an increase in adolescents who are overweight and obese has been observed, with increases in some countries of up to 30% and 40%, respectively, making obesity 15% more prevalent in this population than before the COVID-19 pandemic [34]. During the pandemic, there was also a notable increase in psychological disorders affecting the adolescent population [35] mainly due to the restriction of sports practice, which is very beneficial for maintaining an adequate state of mental health [36]. In addition, the absence of social relationships with peers [37], the uncertainty generated regarding the near future, and the abusive use of new technologies [35], also contributed to this. Although the adolescent population in general was affected,

XSL•FO
RenderX

adolescents who are overweight and obese were particularly affected, increasing the prevalence of disorders such as depression and anxiety [5].

Physical Activity Practice and Interventions in Overweight and Obese Children and Adolescents

It should be added that in Spain, only two 1-hour sessions of curricular physical education classes per week are provided [38]. This means that on many occasions the real practice time during the sessions is 45 minutes, in which barely 13% of students who are overweight and obese reach a moderate or vigorous intensity of practice during the sessions [39], making physical and psychological benefits more difficult to achieve. This is true even though schools are a place where adolescents spend a large part of their day [40]. In this context, some studies have sought to encourage the practice of physical activity within the school context in the population that is overweight and obese. More specifically, interventions have been carried out during school hours using gymnasium exercises, and improvements have been observed in cardiovascular fitness [41], as well as in the reduction of body fat and insulin levels in adolescents who are obese [41].

In addition, more recent studies have used high-intensity interval training (HIIT) 3 times a week within the school setting, achieving beneficial effects on adolescent girls who are overweight [40]. In this regard, these school-based interventions have shown benefits in body weight, waist-hip ratio, fat percentage, and aerobic capacity in adolescents who are overweight [42,43]. Therefore, research supports that school-based interventions are highly effective in improving body composition in the populations that are overweight and obese [40,41].

However, these interventions also have their detractors, who argue that physical education hours are too limited to devote so much time to a single content and that the pedagogical component is also overlooked during implementation [44,45]. As an alternative to this, it has been suggested that physical education classes could promote interventions that encourage adolescents to practice physical activity in their free time [19]. This way, they would be encouraged to walk a minimum weekly distance by means of mobile apps that allow them to monitor the activity carried out, which is also an element that generates adherence and enjoyment in this population [46]. In this regard, previous studies have shown that the use of a physical activity program controlled by mobile step tracker apps in out-of-school hours has provided benefits in increasing the level of physical activity and improving body composition in adolescents [19,47], as well as in the psychological state of this population [35].

Mobile Apps as Promoters of Physical Activity in Overweight and Obese Children and Adolescents

Despite the benefits obtained with the use of mobile apps in the general adolescent population [19], the use of mobile apps by adolescents who are overweight and obese shows contradictory results, and no accurate conclusions can be firmly drawn [48]. This is because the scientific literature on the use of mobile devices to increase physical activity is scarce in this population, as they have been historically considered a “hard-to-reach population” [49]. Perhaps this is the reason why no previous research has been found on adolescents who are overweight and obese using mobile apps to promote compulsory leisure-time physical activity in physical education classes [19]. However, recent research has shown that there are no differences in how adolescents rate these physical activity apps based on their weight status, nor are there differences in their problematic use of these apps [50], so they could be an effective resource for adolescents who are overweight and obese as well.

Therefore, the objectives of the present research were (1) to determine the changes produced by a 10-week intervention promoted by the physical education school subject using step tracker mobile apps in out-of-school hours on physical activity, AMD, body composition, and physical condition of adolescents who are overweight and obese; and (2) to analyze the changes achieved by the 10-week intervention on the psychological state of adolescents who are overweight and obese.

Given the objectives of this research and the scarce previous research conducted with mobile apps for physical activity in adolescents who are overweight or obese, it is not possible to posit consistent research hypotheses, but it is expected that the promotion of the use of mobile apps from the subject of physical education will lead to significant benefits in the level of physical activity, favoring changes in body composition, greater AMD, and improved physical condition in this population (hypothesis 1); and that increased physical activity has a positive effect on the psychological state of adolescents who are overweight and obese (hypothesis 2).

Methods

Design

A randomized controlled trial was conducted in which adolescents were divided into a control group (CG) and an experimental group (EG). Adolescents in the EG used the STRAVA (Strava, Inc) mobile app for physical activity for 10 weeks, a minimum of 3 times per week, to walk the indicated distance, which was incrementally increased as the weeks passed.

The CONSORT (Consolidated Standards of Reporting Trials) guidelines were followed for the research design, and the study was registered before commencement on ClinicalTrials.gov (NCT06089876). The institutional ethics committee of the Catholic University of Murcia approved the study design in accordance with the World Medical Association and following the Declaration of Helsinki (CE022102).

A school in the region of Murcia with the highest number of adolescents in compulsory secondary education of an urban locality was selected. The school with the largest number of adolescents was contacted, and if the management was not interested in participating, the next school with the largest sample was contacted. Once the school was chosen, contact was made with the school’s management team. Once approval was obtained, data collection was coordinated with those in charge of the physical education department. Subsequently, a meeting was held with the adolescents and their parents to explain the purpose and procedure of the study, emphasizing confidentiality

XSL•FO
RenderX

in the treatment of the data obtained. Those who agreed to participate provided an informed consent signed by them and their parents before the start of the research.

Participants

The sample size calculation was performed using the statistical software Rstudio (version 3.15.0; Rstudio Inc). For this purpose, the SDs of previous research on the use of mobile apps by adolescents to increase physical activity were used (SD 0.68) [19]. Thus, for an error (d) of 0.27 and a 95% CI, the minimum sample needed for each group was 25 adolescents [51].

A total of 456 adolescents were enrolled in the school, of whom 83 (18.2%) were overweight or obese. The initial participation in the research was 60 adolescents, of whom 50 (83%) completed the research (n=26, 52% EG and n=24, 48% CG; Figure 1). A total of 24 male participants (n=11, 46% in the EG and n=13, 54% in the CG; mean age at peak height velocity [APHV] 13.09, SD 0.79 years) and a total of 26 female participants (n=15, 58% in the EG and n=11, 42% in the CG; mean APHV 12.45, SD 0.69 years) participated in the research. The age range was between 12 and 16 years (mean 14.11, SD 1.24 years), and all had a BMI ≥25 (mean BMI 27.66, SD 2.30 kg/m^2; total overweight: n=40; total overweight in the EG: n=21; total overweight in the CG: n=19; total obese: n=10; total obese in the EG: n=5; and total obese in the CG: n=5).

Figure 1. Flow diagram of the sample selection.

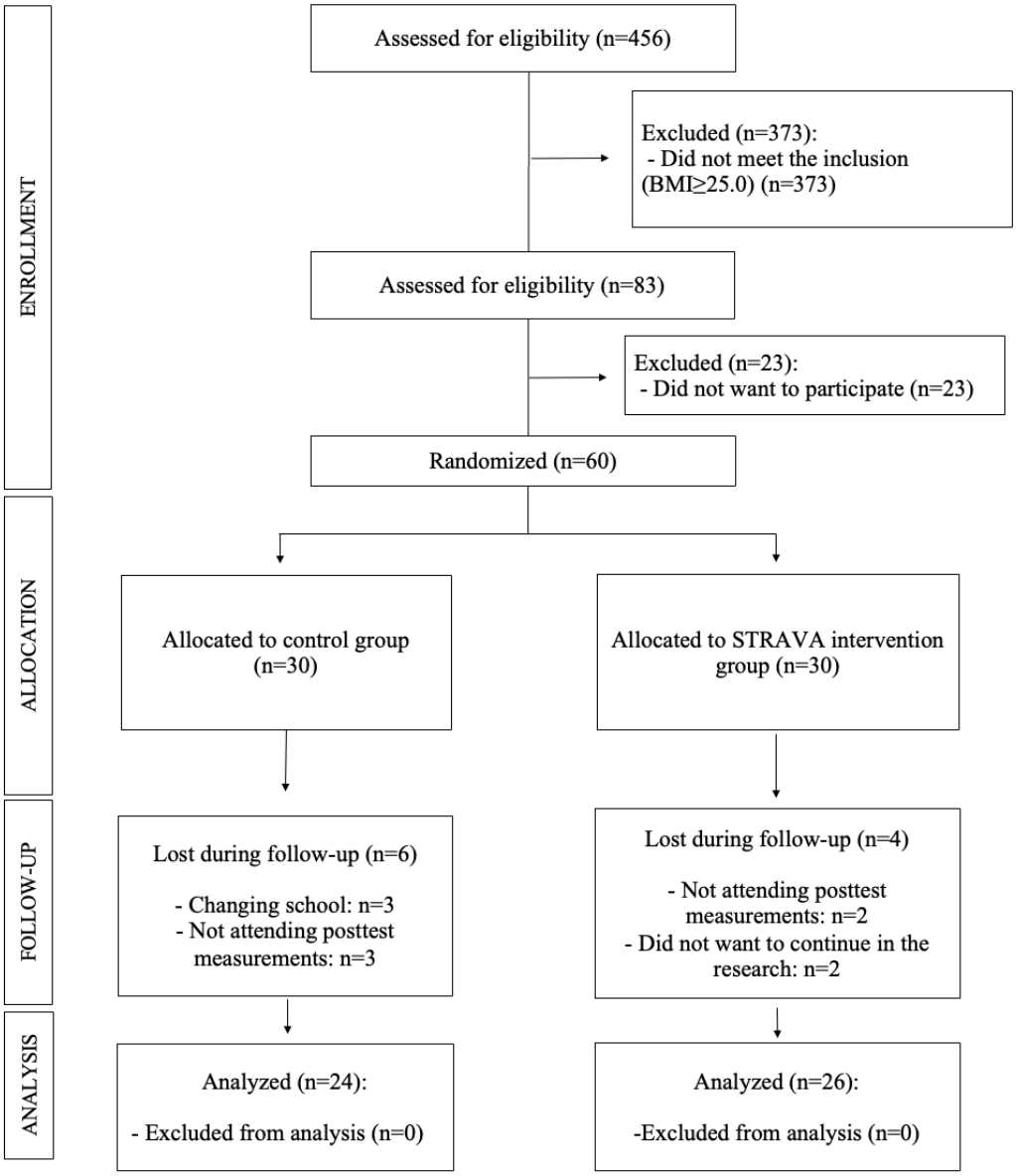

Although BMI has limitations in its use, such as lack of sensitivity in the analysis of fat and fat-free mass, as well as in distinguishing between overweight and obesity in children [52-54], it also has positive factors that make it used ahead of other methods. Simplicity of use is one of its main characteristics [55,56]; when measured accurately and in accordance with growth charts, it is a reliable indicator of overweight and obesity [52,55]; it is a predictive factor that is strongly associated with adult obesity and overweight [53,57]; and the other alternatives have also shown no greater sensitivity than BMI for detecting overweight and obesity in child and adolescent populations [55,57,58].

XSL•FO
RenderX

The inclusion criteria for individuals were as follows: (1) age between 12 and 16 years, (2) compulsory secondary education, and (3) a BMI ≥25 kg/m^2. The exclusion criteria were (1) adolescents in the intervention group not having their own mobile phone; and (2) starting any regular physical activity that was not practiced before the start of the research, understood as going to the gymnasium or starting a specific sport, in both CG and EG.

Randomization and Blinding

The randomization was performed by the principal investigator using a computer-generated random number table in the presence of other investigators not participating in the research. The randomization was carried out considering the BMI of the adolescents. Students were randomly assigned to the CG and EG.

Preintervention measurements were performed after the randomization process. The researchers in charge of the measurements were blinded to the group to which each adolescent belonged. In the postintervention measurements, the researchers were not aware of the scores obtained by each adolescent in the preintervention measurement. In addition, the researchers who oversaw the monitoring of the distance traveled during the intervention by each adolescent were not aware of the scores obtained by each adolescent during the preintervention measurement.

Instruments

Physical Activity Level, AMD, and Psychological State

To assess the level of physical activity of adolescents who are overweight and obese, the Physical Activity Questionnaire for Adolescents (PAQ-A) [59] was used. This is a validated questionnaire in Spanish with moderate validity and reliability (intraclass correlation coefficient of 0.71 for the overall score of the same questionnaire) [60]. The questionnaire consists of 9 items, of which the first 8 are completed using a Likert scale of 1 to 5 points (1: no physical activity and 5: high physical activity), and the last item is answered dichotomously (yes or no). The final score is calculated by the arithmetic mean of the first 8 items (with a minimum score of 1 and a maximum of 5). A higher score reflects a greater engagement in physical activity [59].

AMD was assessed using the Mediterranean Diet Quality Index (KIDMED) [61]. This questionnaire consists of 16 items that are scored as 1 (meets the criteria) or 0 (does not meet the criteria). Out of the 16 items, 12 have a positive connotation (meeting the criterion contributes +1), while the other 4 have a negative connotation (meeting the criterion contributes –1). The final score ranges from 0 to 12 points, with a higher score indicating a greater AMD. The KIDMED has moderate validity and reproducibility for use in adolescents (α=0.79; κ=0.66) [61].

The Basic Psychological Needs Scale (BPNS) [62] and the Satisfaction With Life Scale (SWLS) [63] were used to assess the psychological state of adolescents. The BPNS consists of 18 items distributed in 3 dimensions (competence, autonomy, and social relatedness). Each dimension is composed by 6 items rated on a Likert scale from 1 to 6 points, so the final score for each dimension ranges from 6 to 36 points, with a higher score indicating higher satisfaction; the SWLS has 5 items rated on a Likert scale from 1 to 5 points, with the final score ranging from 5 to 25 points [64]. A higher score on the SWLS shows a higher life satisfaction. Both scales have been previously validated for use in adolescents [65], showing adequate validity and internal consistency (α=0.80 for competence, α=0.69 for autonomy, and α=0.73 for social relatedness) [64,65].

Kinanthropometric and Body Composition Measurements

The body composition analysis was performed following the protocol established by the International Society for the Advancement of Kinanthropometry (ISAK) [66]. Two accredited anthropometrists (levels 3 to 4) measured adolescents who are overweight and obese. The measurement consisted of 2 basic measurements (body mass and height), skinfold in 3 areas (triceps, thigh, and calf) and girth in 5 areas (arm relaxed, waist, hips, thigh, and calf).

Body mass was measured using a TANITA BC 418-MA Segmental scale (TANITA) with an accuracy of 100 g; height was measured with a SECA 213 stadiometer (SECA); girth was measured using Lufkin W606PM inextensible tape (Lufkin) with an accuracy of 0.1 cm; and skinfold was measured using a caliper with an accuracy of 0.2 mm (Harpenden). Before the pre- and postintervention measurements, all instruments were calibrated.

In total, 2 measurements were made for each variable, and if the difference between the 2 measurements was >5% in skinfold or 1% in all other measurements, a third measurement was made. When 2 measurements were taken, the final value corresponded to the average of the 2 measurements. However, when a third measurement was included, the final value was the median of the 3 measurements [66].

The intra- and interrater technical errors of measurement (TEMs) were calculated on a subsample. The intrarater TEM was 0.04% for basic measurements, 1.58% for skinfold, and 0.06% for girth. The interrater TEM was 0.06% for basic measurements, 2.01% for skinfold, and 0.08% for girth.

With the final values of each of the anthropometric measurements, the following variables were calculated: BMI [67], fat mass (%) [67], muscle mass (%) [68], sum of skinfold in 3 areas ($\sum$3; triceps, thigh, and calf) [63], waist-height ratio (waist girth/height), and corrected girth of the arm (arm relaxed girth–[π×triceps skinfold]), thigh (middle thigh girth–[π×thigh skinfold]), and calf (calf girth–[π×calf skinfold]).

The maturity offset was estimated using the sex-specific formula by Mirwald et al [69]. This method has shown validity for estimating maturity offset against the gold standard (radiograph of the left wrist), with R^2 values of 0.92 to 0.89 for male participants and 0.91 to 0.88 for female participants. The result was used to calculate the APHV by means of the following formula: APHV = chronological age – maturity offset. APHV is defined as the age at which there is a dramatic increase in the rate of growth in height and body mass in adolescents [70].

XSL•FO
RenderX

Therefore, a positive value in the maturity offset indicated how many years ago that participant had passed his or her APHV, while a negative value indicated the years remaining before that adolescent reached APHV [69].

Physical Fitness Test

The physical condition of the adolescents was assessed according to previous research carried out in this population [19,71]. The physical capacities assessed were cardiorespiratory capacity, hamstring flexibility, upper limb strength, lower body explosive power, speed, and abdominal strength endurance.

Cardiorespiratory capacity was measured using the 20-m shuttle run test [72]. This is a maximal incremental test where 20 m must be run as many times as possible before the beep sounds. The test is terminated when the participant is exhausted or is unable to complete the distance before the beep sounds on 2 consecutive occasions [19]. At the end of the test, the maximal oxygen uptake can be calculated using the formula by Léger et al [72].

For hamstring flexibility, the sit-and-reach test was used [73]. For its correct execution, the adolescents start seated with their legs extended and ankles flexed at 90°, allowing their feet to rest against an Acuflex Tester III box (Novel Products). From this position, the participant must reach the maximum possible distance by moving the palms of the hands, one on top of the other, along the box by pushing a bar, keeping the legs straight and the feet fully supported against the box at all times [74].

For upper limb strength, the handgrip strength test and the push-ups test were used. Handgrip strength consisted of applying the maximum possible force on a Takei TKK5401 portable digital dynamometer (Take Scientific Instruments) with the elbow fully extended [75]. This is because this position has been shown to be the most valid for applying maximal force. This test has been previously validated for use in this population [76] and was performed on both the right and left hand [77]. For the push-up test, the participants had to start in the prone position with only their feet and hands in contact with the ground. The hands were placed at the sides of the body with the elbows bent at 90°, while the tip of the feet was the area in contact with the ground. Once in the starting position, the adolescents had to push up from the floor with their backs and legs fully straight, achieving a full extension of the arms. The test ended when the adolescents could not fully extend their arms or when 1 minute was exceeded. The final value corresponded to the maximum number of correctly executed repetitions [78].

Lower limb explosive power was assessed using the countermovement jump (CMJ). The adolescents started standing on a force platform (MuscleLab), with hands on hips and feet hip-width apart. From this position, the maximum possible height was to be reached by performing a vertical jump. To do this, the adolescents performed a 90° knee flexion followed by a maximum knee extension without stopping between the 2 phases, keeping their hands on their hips and as vertical as possible during the flight phase. The knees and ankles had to be fully extended during the flight phase [79].

The 20-m sprint test was used to measure the speed of the adolescents. During the test, the adolescents had to cover 20 m in the shortest possible time [80]. Single-beam photocells (Polifemo Light Microgate), located at hip height, were used to measure the time it took the adolescents to cover the distance [81], as this arrangement has been shown to give the greatest validity and reliability to the test. In this position, there is only a 4% chance that the arms will cut the photocell before the rest of the body, whereas at chest height the probability of this happening increases to 60% [82].

Abdominal strength endurance was measured using the curl-up test. The adolescents were placed in supine position on a mat with their arms supported and crossed over their chest, their knees bent at 90°, and their feet fully supported on the floor. From this position, the participant had to perform the maximum number of trunk flexions, lifting the upper back off the ground, until exhaustion was reached or 1 minute of time had elapsed [83].

Procedure

First, the PAQ-A, KIDMED, BPNS, and SWLS questionnaires were completed. Subsequently, the anthropometric assessment was carried out by the ISAK-accredited anthropometrists. Once this was completed, the physical fitness assessment tests were carried out. Before warming up, all the adolescents performed the sit-and-reach test, as previous research has shown that the warm-up influences the performance obtained in this test [84]. Once this test was completed, the adolescents were given an explanation about the handgrip strength, push-ups, CMJ, 20-m sprint, and curl-up tests to become familiarized with them. The adolescents underwent a 5-minute warm-up consisting of progressive running and mobility of the joints involved in the fitness tests (ankles, knees, hips, wrists, and shoulders). The participants randomly performed 2 repetitions of each of the fitness tests, allowing 2 minutes of rest between repetitions of each test and 5 minutes between the different tests. The best repetition was selected as the final value of the test. Once these physical fitness tests were completed, the adolescents performed a single repetition of the 20-m shuttle run test, preventing the fatigue of this maximum test from influencing the execution of the rest of the tests.

It should be noted that all the measurements were carried out during school hours corresponding to physical education classes. The questionnaires were completed in a classroom at the school, avoiding any distractions and maintaining a calm atmosphere for their correct completion. The anthropometric assessment was carried out in the changing rooms of the sports hall, maintaining a stable temperature and as much privacy as possible during the measurements. The pre- and postintervention measurements were carried out at the same time in all groups, as it was the school time provided for the school physical-education classes, thus avoiding that the changes that occur in the body composition variables during the day could affect the results [85]. The physical fitness tests were carried out in the sports hall, maintaining a stable temperature and preventing polluting atmospheric variables from influencing the results.

XSL•FO
RenderX

The warm-up and execution of the physical condition tests were supervised by researchers with previous experience in the execution of these tests, avoiding possible errors in the execution and collection of data. The order of the tests was determined according to the National Strength and Conditioning Association, which bases its suggestions on the fatigue generated during the tests as well as on the metabolic pathways required for each of them [86].

Mobile App Intervention

The initial sample of participants (n=60) was divided into the EG (n=30) and CG (n=30), with the EG adolescents using the STRAVA mobile app for 10 weeks after school. The CG did not use any mobile app in out-of-school hours. Both EG and CG continued to attend physical education classes as normal. Pre- and postintervention physical activity level, body composition, psychological state, and physical condition of all participants were measured.

The assignment given to each group was randomized. Before starting the intervention, the physical activity level, body composition, psychological state, and physical fitness of all adolescents were assessed. This was followed by the 10-week intervention. The EG adolescents were required to use STRAVA after school at least 3 times per week, completing a minimum of 4600 steps per day the first week, ending with 10,000 steps per day, as this is the minimum value for the physical activity performed to be considered moderate-vigorous [55,87]. So, this distance was defined as the final target (week 10: 10,000 steps per day), which was reached by increasing the distance by 600 steps per week (week 1: 4600 steps per day), following indications from previous research with adolescents using mobile apps [88].

To facilitate the compliance of the adolescents with the recommendations, the steps were converted into kilometers, as STRAVA records the distance traveled in kilometers. In this respect, 4600 steps correspond to 2.94 km and 10,000 steps to 6.40 km, considering that 1 km represents approximately 1565 steps in the adolescent population [89]. This mobile app was selected because it has been proven to increase physical activity [90] and because it includes numerous techniques for behavioral change [91].

At the end of the 10 weeks, the physical activity level, body composition, psychological state, and physical condition of all adolescents were reevaluated (postintervention measurements). Those adolescents in the EG who did not complete the minimum weekly distance were not excluded from the study, nor were those who did not complete the intervention.

The final study sample consisted of 50 adolescents (26 EG and 24 CG). The attrition rate of the research was 13% (4/26) in the EG and 20% (6/24) in the CG. The mean reasons for dropping out of the intervention were as follows: changing schools (n=3), not attending the postintervention measurements (n=5), and not wanting to continue in the research (n=2).

Data Analysis

The normality of the variables was analyzed using the Shapiro-Wilk test, as well as kurtosis and skewness. As the variables followed a normal distribution, parametric tests were used for the analysis. The Levene test was used to assess the homogeneity between the EG and CG in the study variables. A mixed model ANOVA was carried out to analyze intragroup differences in physical activity, kinanthropometric and body composition, psychological state, and physical condition. Subsequently, 3 analyses of covariance were carried out to analyze the differences in the study variables when gender, age, and distance covered with the app were included. For the analysis of change, an ANOVA was performed comparing the difference between pre and post EG with respect to the difference between pre and post CG. Two analyses of covariance were subsequently performed with the covariate gender and age. Partial eta squared (η^2) was used to calculate the effect size and was defined as small: ES≥0.10; moderate: ES≥0.30; large: ES≥1.2; or very large: ES≥2.0, with an error of $P<.05$ [92]. A value of $P<.05$ was set to determine statistical significance. The statistical analysis was performed with the SPSS statistical package (version 25.0; SPSS Inc).

Ethical Considerations

Before the start of the study, the institutional ethics committee of the Catholic University of Murcia approved the research design in accordance with the World Medical Association (CE022102). Informed consent was obtained from all individual participants included in the study.

Results

The sample flow diagram showing the final participants in the CG and EG can be found in Figure 1.

Normality and Homogeneity Tests

The results of the normality test are presented in Table 1.

Levene test for homogeneity showed that there was homogeneity between the EG and CG in physical activity (P=.54), body mass (P=.11), BMI (P=.93), waist girth (P=.34), hip girth (P=.97), waist-height ratio (P=.06), corrected arm girth (P=.54), corrected thigh girth (P=.43), corrected calf girth (P=.78), fat mass (P=.17), muscle mass (P=.41), sum of skinfold in 3 areas (P=.30), AMD (P=.55), life satisfaction (P=.07), competence (P=.91), autonomy (P=.57), relatedness (P=.65), maximal oxygen uptake (P=.46), handgrip right hand (P=.32), handgrip left hand (P=.57), sit and reach (P=.15), CMJ (P=.09), 20-m sprint (P=.35), curl up (P=.39), and push up (P=.05).

XSL•FO
RenderX

Table 1. Results of the normality analysis for control and experimental group variables.

	Experimental group, *P* value	Control group, *P* value
Physical activity	.78	.46
Body mass	.10	.20
BMI	.11	.06
Waist girth	.08	.53
Hip girth	.10	.18
Waist-height ratio	.07	.18
Corrected arm girth	.22	.06
Corrected thigh girth	.97	.62
Corrected calf girth	.30	.95
Fat mass	.91	.98
Muscle mass	.08	.08
Sum of skinfold in 3 areas	.69	.80
Adherence to the Mediterranean diet	.29	.52
Life satisfaction	.20	.21
Competence	.52	.21
Autonomy	.11	.81
Relatedness	.34	.75
Maximal oxygen uptake	.42	.05
Handgrip right hand	.07	.19
Handgrip left hand	.12	.75
Sit and reach	.61	.15
Countermovement jump	.99	.08
20-m sprint	.09	.59
Curl up	.35	.06
Push up	.30	.38

Distance Covered With the App

The distance traveled by the EG adolescents using the apps was recorded weekly. The average distance traveled was 76.44 (SD 39.10) km over the 10 weeks of intervention, with the maximum distance traveled by one of the adolescents being 186.2 km, while the minimum was 14.43 km. It is worth noting that the final distance to be completed with the program was 139.97 km over the 10 weeks of intervention, and only 12% (3/26) adolescents reached or exceeded this distance. In fact, 54% (14/26) of the adolescents did not continue to use the app systematically after the sixth week.

Physical Activity Level, AMD, and Psychological State

Table 2 shows the pre- and postintervention differences in the EG and CG for physical activity level, AMD, and psychological variables. The results showed only a significant decrease in competence in the adolescents in the CG (mean difference: 3.08; *P*=.03), with no significant changes in the other variables.

The influence of the covariates gender, age, and distance covered with the app on the level of physical activity, AMD, and psychological variables can be found in Table 3. Results showed influence of gender and age on AMD (gender: *P*=.04; age: *P*=.05), competence (gender: *P*=.04; age: *P*=.04) and relatedness (gender: *P*=.05; age: *P*=.04) of CG adolescents. No significant differences were found for the covariate distance covered with the app.

XSL•FO
RenderX

Table 2. Differences between the pre- and postintervention measurements in the experimental group (EG) and control group (CG) groups of adolescents who are overweight or obese in physical activity, adherence to the Mediterranean diet (AMD), and psychological state.

Variables and time point	Before intervention, mean (SD)	After intervention, mean (SD)	Prepost difference, mean (SD)	*F* test (*df*)	*P* value	95% CI difference	η^2
Physical activity							
EG	2.51 (0.74)	2.62 (0.68)	–0.11 (0.07)	2.430 (1)	.13	–0.248 to 0.031	0.048
CG	2.63 (0.64)	2.69 (0.55)	–0.06 (0.07)	0.706 (1)	.41	–0.207 to 0.085	0.015
AMD							
EG	6.73 (2.16)	6.46 (2.45)	0.27 (0.39)	0.489 (1)	.49	–0.505 to 1.044	0.010
CG	7.21 (2.43)	6.42 (2.62)	0.79 (0.40)	3.901 (1)	.05	–0.014 to 1.598	0.075
Life satisfaction							
EG	17.00 (3.14)	16.88 (2.90)	0.12 (0.70)	0.027 (1)	.87	–1.296 to 1.527	0.001
CG	18.46 (4.40)	17.21 (5.21)	1.25 (0.73)	2.926 (1)	.09	–0.219 to 2.719	0.094
Competence							
EG	24.54 (96.18)	23.85 (5.05)	0.69 (1.31)	0.281 (1)	.60	–1.933 to 3.317	0.006
CG	26.83 (5.98)	23.75 (9.40)	3.08 (1.36)	5.148 (1)	.03	0.351 to 5.816	0.097
Autonomy							
EG	23.96 (5.71)	23.88 (5.64)	0.07 (1.70)	0.002 (1)	.96	–3.338 to 3.491	0.001
CG	25.75 (5.03)	24.08 (8.67)	1.67 (1.77)	0.889 (1)	.35	–1.887 to 5.221	0.018
Relatedness							
EG	24.15 (6.70)	24.54 (6.33)	–0.39 (1.35)	0.082 (1)	.78	–0.385 to 1.347	0.002
CG	23.13 (5.98)	20.42 (8.58)	2.71 (1.40)	3.734 (1)	.06	2.708 to 1.402	0.072

XSL•FO
RenderX

Table 3. Influence of the covariates gender, age, and distance covered with app in the physical activity level, adherence to the Mediterranean diet (AMD), and psychological state.

Variables and time point	App use × gender				App use × age				App use × distance covered with app			
	F test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI diff	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2
Physical activity												
EG[a]	1.913 (1)	.17	–0.247 to 0.046	0.039	1.006 (1)	.32	–0.211 to 0.070	0.021	0.157 (1)	.69	–0.183 to 0.123	0.003
CG[b]	0.842 (1)	.36	–0.222 to 0.083	0.018	1.973 (1)	.17	–0.249 to 0.044	0.040	__[c]	—	—	—
AMD												
EG	0.196 (1)	.66	–0.626 to 0.980	0.004	0.273 (1)	.60	–0.599to 1.020	0.006	3.120 (1)	.08	–0.115 to 1.766	0.062
CG	4.576 (1)	.04	0.053 to 1.730	0.089	4.148 (1)	.05	0.010 to 1.701	0.081	—	—	—	—
Life satisfaction												
EG	0.010 (1)	.92	–1.535 to 1.388	<0.001	0.012 (1)	.91	–1.545 to 1.385	0.000	0.120 (1)	.73	–1.879 to 1.327	0.003
CG	3.682 (1)	.06	–0.071 to 2.980	0.073	3.699 (1)	.06	–0.067 to 2.991	0.073	—	—	—	—
Competence												
EG	0.295 (1)	.59	–2.004 to 3.488	0.006	0.241 (1)	.63	–2.082 to 3.426	0.005	0.564 (1)	.46	–1.882 to 4.123	0.12
CG	4.521 (1)	.04	0.163 to 5.896	0.088	4.723 (1)	.04	0.231 to 5.980	0.091	—	—	—	—
Autonomy												
EG	0.190 (1)	.89	–3.321 to 3.815	<0.001	0.000 (1)	.99	–3.609 to 3.551	0.000	0.035 (1)	.85	–3.550 to 4283	0.001
CG	0.642 (1)	.43	–2.241 to 5.206	0.013	0.920 (1)	.34	–1.956 to 5.519	0.019	—	—	—	—
Relatedness												
EG	0.210 (1)	.65	–3.460 to 2.176	0.004	0.273 (1)	.60	–3.546 to 2.083	0.006	0.137 (1)	.71	–3.678 to 2.535	0.003
CG	4.176 (1)	.047	0.046 to 5.929	0.082	4.459 (1)	.04	0.146 to 6.022	0.087	—	—	—	—

[a]EG: experimental group.

[b]CG: control group.

[c]Not applicable.

Kinanthropometric and Body Composition Measurements

Table 4 shows the differences in kinanthropometric and body composition variables between pre- and postintervention measurements in the CG and EG. The differences were significant in the EG, with an increase in corrected arm girth (mean difference: –0.46; *P*=.05), as well as in the CG, with an increase in hip girth (mean difference: –1.37; *P*=.05), corrected thigh girth (mean difference: –1.28; *P*=.04), and muscle mass (mean difference –0.87; *P*=.04).

The covariate gender showed an effect on the variables corrected arm girth (*P*=.04) of the EG, as well as the corrected thigh girth (*P*=.04) of the CG. No significant differences were found for the covariate age nor for the covariate distance covered (Table 5).

XSL•FO
RenderX

Table 4. Differences between the pre- and postintervention measurements in the experimental group (EG) and control group (CG) of adolescents who are overweight or obese in the kinanthropometric and body composition measurements.

Variables and time point	Before intervention, mean (SD)	After intervention, mean (SD)	Prepost difference, mean (SD)	*F* test (*df*)	*P* value	95% CI difference	η^2
Body mass (kg)							
EG	75.78 (16.83)	75.60 (17.62)	0.18 (0.43)	0.168 (1)	.68	–0.691 to 1.045	0.003
CG	70.21 (10.78)	70.74 (11.08)	–0.53 (0.45)	1.386 (1)	.25	–1.433 to 0.375	0.028
BMI (kg/m^2)							
EG	27.76 (3.51)	27.67 (3.79)	0.09 (0.19)	0.235 (1)	.63	–0.291 to 0.475	0.005
CG	27.71 (3.18)	27.52 (2.89)	0.18 (0.20)	0.855 (1)	.36	–0.215 to 0.582	0.018
Waist girth (cm)							
EG	82.11 (10.39)	81.55 (11.09)	0.57 (0.68)	0.701 (1)	.41	–0.796 to 1.929	0.015
CG	82.71 (8.66)	83.19 (7.50)	–0.48 (0.74)	0.419 (1)	.52	–1.957 to 1.005	0.009
Hip girth (cm)							
EG	104.93 (8.57)	104.96 (9.15)	–0.02 (0.62)	0.001 (1)	.97	–1.269 to 1.224	0.001
CG	99.65 (7.92)	101.02 (7.23)	–1.37 (0.67)	4.157 (1)	.05	–2.728 to –0.017	0.083
Waist-height							
EG	0.50 (0.04)	0.49 (0.05)	0.00 (0.01)	1.070 (1)	.31	–0.004 to 0.013	0.023
CG	0.52 (0.07)	0.52 (0.06)	0.00 (0.01)	0.007 (1)	.94	–0.009 to 0.009	0.001
Corrected arm girth (cm)							
EG	23.55 (2.29)	24.01 (2.53)	–0.46 (0.22)	4.185 (1)	.05	–0.911 to –0.007	0.083
CG	23.55 (2.75)	23.86 (2.60)	–0.31 (0.24)	1.635 (1)	.21	–0.803 to 0.179	0.034
Corrected thigh girth (cm)							
EG	43.91 (5.68)	44.82 (5.47)	–0.91 (0.57)	2.607 (1)	.11	–2.052 to 0.225	0.054
CG	42.69 (4.81)	43.97 (4.00)	–1.28 (0.62)	4.321 (1)	.04	–2.517 to –0.040	0.054
Corrected calf girth (cm)							
EG	30.77 (6.62)	30.25 (3.67)	0.52 (0.94)	0.312 (1)	.58	–1.364 to 2.411	0.007
CG	29.21 (3.73)	29.77 (3.30)	–0.56 (1.02)	0.301 (1)	.59	–2.611 to 1.493	0.007
Fat mass (%)							
EG	39.14 (10.28)	37.09 (8.85)	2.04 (1.25)	2.679 (1)	.11	–0.470 to 4.557	0.055
CG	36.64 (13.74)	37.33 (10.96)	–0.69 (1.36)	0.258 (1)	.61	–3.422 to 2.043	0.006
Muscle mass (kg)							
EG	21.59 (6.01)	22.18 (6.43)	–0.59 (0.37)	2.578 (1)	.12	–1.327 to 0.149	0.053
CG	20.95 (5.07)	21.82 (4.78)	–0.87 (0.40)	4.737 (1)	.04	–1.671 to –0.065	0.093
Summary of skinfold in 3 areas							
EG	93.27 (28.71)	87.29 (23.97)	5.98 (3.50)	2.922 (1)	.09	–1.062 to 13.023	0.060
CG	85.56 (35.56)	84.53 (26.47)	1.03 (3.80)	0.073 (1)	.79	–6.631 to 8.682	0.002

XSL•FO
RenderX

Table 5. Influence of the covariates gender, age, and distance covered with app in the kinanthropometric and body composition measurements.

Variables and time point	App use × gender				App use × age				App use × distance covered with app			
	F test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2
Body mass (kg)												
EG[a]	0.051 (1)	.82	–0.803 to 1.006	0.001	0.028 (1)	.87	–0.829 to 0.979	0.001	0.392 (1)	.54	–0.684 to 1.302	0.008
CG[b]	0.910 (1)	.35	–1.391 to 0.496	0.019	0.798 (1)	.38	–1.362 to 0.525	0.017	__[c]	—	—	—
BMI (kg/m^2)												
EG	0.036 (1)	.85	–0.358 to 0.433	0.001	0.153 (1)	.70	–0.323 to 0.480	0.003	1.231 (1)	.27	–0.193 to 0.668	0.026
CG	1.400 (1)	.24	–0.170 to 0.656	0.029	0.910 (1)	.35	–0.220 to 0.618	0.019	—	—	—	—
Waist girth (cm)												
EG	0.356 (1)	.55	–0.998 to 1.838	0.008	1.331 (1)	.26	–0.595 to 2.192	0.029	0.843 (1)	.36	–0.841 to 2.251	0.018
CG	0.155 (1)	.70	–1.854 to 1.249	0.003	0.984 (1)	.33	–2.274 to 0.773	0.021	—	—	—	—
Hip girth (cm)												
EG	0.045 (1)	.83	–1.438 to 1.164	0.001	0.084 (1)	.77	–1.094 to 1.461	0.002	0.014 (1)	.91	–1.332 to 1.500	0.001
CG	3.070 (1)	.09	–2.660 to 0.185	0.064	5.439 (1)	.05	–3.012 to –0.220	0.088	—	—	—	—
Waist-height												
EG	0.638 (1)	.43	–0.005 to 0.012	0.014	1.961 (1)	.17	–0.003 to 0.014	0.042	1.768 (1)	.19	–0.004 to 0.017	0.038
CG	0.083 (1)	.77	–0.008 to 0.011	0.002	0.111 (1)	.74	–0.011 to 0.008	0.002	—	—	—	—
Corrected arm girth (cm)												
EG	4.378 (1)	.04	–0.963 to –0.180	0.089	2.768 (1)	.10	–0.844 to 0.080	0.058	2.914 (1)	.10	–0.948 to 0.078	0.061
CG	1.147 (1)	.29	–0.791 to 0.242	0.025	2.588 (1)	.12	–0.908 to 0.102	0.054	—	—	—	—
Corrected thigh girth (cm)												
EG	1.962 (1)	.17	–2.018 to 0.362	0.042	3.130 (1)	.08	–2.216 to 0.143	0.065	2.471 (1)	.12	–2.303 to 0.284	0.052
CG	4.558 (1)	.04	–2.682 to –0.078	0.092	3.135 (1)	.08	–2.423 to 0.156	0.065	—	—	—	—
Corrected calf girth (cm)												
EG	0.117 (1)	.73	–1.633 to 2.302	0.003	0.205 (1)	.65	–1.526 to 2.410	0.005	0.373 (1)	.55	–1.495 to 2.795	0.008
CG	0.099 (1)	.76	–2.487 to 1.817	0.002	0.188 (1)	.67	–2.614 to 1.689	0.004	—	—	—	—
Fat mass (%)												
EG	2.105 (1)	.15	–0.736 to 4.524	0.045	2.658 (1)	.11	–0.500 to 4.745	0.056	2.196 (1)	.15	–0.755 to 4.960	0.047
CG	0.129 (1)	.72	–3.390 to 2.364	0.003	0.303 (1)	.59	–3.649 to 2.084	0.007	—	—	—	—

XSL•FO
RenderX

Variables and time point	App use × gender				App use × age				App use × distance covered with app			
	F test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2
Muscle mass (kg)												
EG	2.857 (1)	.10	–1.419 to 0.124	0.060	2.707 (1)	.11	–1.398 to 0.141	0.057	1.602 (1)	.21	–1.366 to 0.312	0.034
CG	3.639 (1)	.06	–1.643 to 0.045	0.075	3.869 (1)	.06	–1.662 to 0.020	0.079	—	—	—	—
Summary of skinfold in 3 areas												
EG	2.092 (1)	.16	–2.069 to 12.610	0.044	3.789 (1)	.06	–0.243 to 14.249	0.078	2.774 (1)	.10	–1.385 to 14.611	0.058
CG	0.219 (1)	.64	–6.164 to 9.893	0.005	0.002 (1)	.96	–8.104 to 7.738	0	—	—	—	—

[a]EG: experimental group.

[b]CG: control group.

[c]Not applicable.

Physical Fitness Tests

Table 6 shows the differences in the fitness variables between pre- and postintervention measurements. In the EG an increase in curl-up (mean difference: –6.35; *P*=.02) and push-up (mean difference: –2.27; *P*=.04) was found after the intervention. No significant differences were found in the CG.

The influence of the covariates on the fitness variables is shown in Table 7. The covariate gender showed an effect on the variables curl-up (*P*=.04) and push-up (*P*=.04) of the EG. No significant differences were found for the covariate age, nor for the covariate distance covered.

XSL•FO
RenderX

Table 6. Differences between the pre- and postintervention measurements in the experimental group (EG) and control group (CG) of adolescents who are overweight or obese in the physical fitness variables.

Variables and time point	Before intervention, mean (SD)	After intervention, mean (SD)	Pre-post difference, mean (SD)	*F* test (*df*)	*P* value	95% CI difference	η^2
VO_2 max [a] (mL/kg/min)							
EG	34.55 (3.43)	34.40 (4.08)	0.15 (0.40)	0.135 (1)	.72	–0.662 to 0.957	0.003
CG	36.56 (4.70)	36.59 (4.92)	–0.03 (0.41)	0.006 (1)	.94	–0.859 to 0.798	0.001
Handgrip right hand (kg)							
EG	26.93 (6.46)	27.55 (6.46)	–0.62 (0.99)	0.391 (1)	.54	–2.612 to 1.373	0.008
CG	26.03 (9.28)	25.42 (11.35)	0.61 (1.03)	0.353 (1)	.56	–1.461 to 2.686	0.007
Handgrip left hand (kg)							
EG	25.32 (6.79)	25.29 (6.80)	0.03 (0.60)	0.003 (1)	.96	–1.175 to 1.237	0.001
CG	24.72 (7.52)	25.15 (8.86)	–0.44 (0.62)	0.491 (1)	.49	–1.693 to 0.818	0.010
Sit-and-reach (cm)							
EG	16.29 (9.84)	17.67 (10.59)	–1.39 (0.82)	2.879 (1)	.10	–3.026 to 0.257	0.058
CG	13.22 (7.53)	14.78 (7.58)	–1.57 (0.87)	3.254 (1)	.08	–3.311 to 0.180	0.065
Countermovement jump (cm)							
EG	17.93 (5.32)	18.49 (6.58)	–0.56 (1.08)	0.271 (1)	.61	–2.727 to 1.605	0.006
CG	18.27 (7.60)	19.45 (7.93)	–1.19 (1.12)	1.122 (1)	.30	–3.442 to 1.067	0.023
20 m sprint							
EG	4.05 (0.96)	3.78 (1.20)	0.27 (0.16)	2.880 (1)	.10	–0.049 to 0.585	0.057
CG	4.17 (1.10)	4.25 (0.63)	–0.08 (0.16)	0.228 (1)	.64	–0.408 to 0.252	0.005
Curl-up (repetitions)							
EG	17.23 (11.28)	23.58 (9.29)	–6.35 (2.51)	6.382 (1)	.02	–11.397 to –1.295	0.117
CG	18.17 (12.93)	21.71 (12.86)	–3.54 (2.62)	1.835 (1)	.18	–8.799 to 1.716	0.037
Push-up (repetitions)							
EG	1.88 (4.48)	4.15 (5.81)	–2.27 (1.05)	4.698 (1)	.04	–4.377 to –0.162	0.093
CG	4.86 (6.47)	4.77 (6.24)	0.09 (1.14)	0.006 (1)	.94	–2.200 to 2.382	0.001

[a]VO_2max: maximal oxygen uptake.

XSL•FO
RenderX

Table 7. Influence of the covariates gender, age, and distance covered with app in the physical fitness variables.

Variables and time point	App use × gender				App use × age				App use × distance covered with app			
	F test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2	*F* test (*df*)	*P* value	95% CI difference	η^2
VO_2 max[a] (mL/kg/min)												
EG[b]	0.199 (1)	.66	–0.917 to 0.585	0.005	0.012 (1)	.91	–0.793 to 0.886	0.000	1.441 (1)	.24	–0.359 to 1.408	0.035
CG[c]	0.611 (1)	.44	–0.472 to 1.068	0.015	0.031 (1)	.86	–0.786 to 0.936	0.001	__[d]	—	—	—
Handgrip right hand (kg)												
EG	0.638 (1)	.43	–2.895 to 1.249	0.013	0.045 (1)	.83	–2.259 to 1.826	0.001	0.400 (1)	.53	–3.006 to 1.568	0.008
CG	0.601 (1)	.44	–1.329 to 2.995	0.013	0.028 (1)	.87	–1.956 to 2.308	0.001	—	—	—	—
Handgrip left hand (kg)												
EG	0.003 (1)	.96	–1.229 to 1.295	0.001	0.067 (1)	.80	–1.095 to 1.418	0.001	0.153 (1)	.70	–1.639 to 1.106	0.003
CG	0.452 (1)	.51	–1.757 to 0.877	0.010	0.788 (1)	.38	–1.891 to 0.733	0.016	—	—	—	—
Sit-and-reach (cm)												
EG	1.977 (1)	.17	–2.915 to 0.517	0.041	2.133 (1)	.15	–2.947 to 0.469	0.044	2.806 (1)	.10	–3.432 to 0.314	0.058
CG	3.797 (1)	.06	–3.609 to 0.059	0.076	3.647 (1)	.06	–3.553 to 0.094	0.073	—	—	—	—
Countermovement jump (cm)												
EG	0.120 (1)	.73	–2.646 to 1.869	0.003	0.273 (1)	.60	–2.862 to 1.682	0.006	0.179 (1)	.67	–3.011 to 1.963	0.004
CG	1.377 (1)	.25	–3.731 to 0.982	0.028	0.961 (1)	.33	–3.528 to 1.216	0.020	—	—	—	—
20-m sprint												
EG	3.260 (1)	.08	–0.034 to 0.627	0.065	1.807 (1)	.19	–0.109 to 0.548	0.037	1.274 (1)	.27	–0.159 to 0.565	0.026
CG	0.408 (1)	.53	–0.454 to 0.235	0.009	0.023 (1)	.88	–0.369 to 0.317	0.000	—	—	—	—
Curl-up (repetitions)												
EG	4.492 (1)	.04	–10.657 to –0.278	0.087	3.692 (1)	.06	–9.801 to 0.225	0.073	3.319 (1)	.08	–10.972 to 0.543	0.066
CG	2.785 (1)	.10	–9.910 to 0.923	0.056	4.042 (1)	.05	–10.463 to 0.004	0.079	—	—	—	—
Push-up (repetitions)												
EG	4.292 (1)	.04	–4.482 to –0.063	0.087	3.764 (1)	.06	–4.359 to 0.082	0.077	3.326 (1)	.08	–4.569 to 0.227	0.069
CG	0.006 (1)	.94	–2.322 to 2.512	0.001	0.003 (1)	.96	–2.495 to 2.368	0.000	—	—	—	—

[a]VO_2max: maximal oxygen consumption.

[b]EG: experimental group.

[c]CG: control group.

[d]Not applicable.

XSL•FO
RenderX

Analysis of Change Between CG and EG

Multimedia Appendix 1 shows an analysis of pre-post differences found in the EG and CG. It should be noted that no significant differences were found with the use of the app, nor after the inclusion of the covariates gender and age.

Discussion

Principal Findings

The main results of the research showed that a large proportion of the adolescents did not continue using the app after the first 6 weeks. Some changes were observed in the study variables, including a decrease in competence in the CG but not in the EG; an increase in corrected arm girth in the EG and in hip girth, corrected thigh girth, and muscle mass in the CG; and an improvement in curl-up repetitions and push-up repetitions in the EG. The covariates also showed an effect on the study variables, mainly gender and age, but not the distance covered with the app. However, despite the results obtained, the changes produced in the EG were not significantly different from those produced in the CG.

The present investigation originated from previous research that had been carried out in the school setting with adolescents who are overweight and obese and that had been effective, mainly on body composition and physical condition. In this regard, school-based HIIT interventions had shown effectiveness on body composition and cardiorespiratory fitness [42,43,93]. These interventions were characterized by being carried out in short periods of time and at high intensity, even allowing to obtain benefits in cardiac and blood parameters, being very beneficial for the health of adolescents [93,94]. In addition to interventions based on HIIT, other interventions based on the accumulation of minutes of physical activity have also been carried out in the school environment, restructuring the organization of classes, increasing the possible physical activity during class, or using breaks between classes. These have also reported benefits on fitness, physical activity level, body composition, and psychological state [95,96]. However, given the limited number of hours of physical education in the school context and the large amount of content that needs to be addressed in the classroom, such interventions have limited effects as they cannot be sustained in the long term [44,45]. So, resources are needed to promote the practice of physical activity outside of school hours. In this context, previous literature has shown mixed results in promoting out-of-school physical activity, with interventions that have led to increased physical activity among adolescents but with others that do not perceive benefits in physical activity levels or psychological variables, such as motivation or competence [97,98]. One of the main reasons for the lack of conclusive results is that after-school physical activity has the problem of physical education teachers' lack of control over what is being done [99]. In this sense, mobile devices allow tracking of the physical activity performed in the out-of-school environment, so these apps have proven to be a useful alternative to promote physical activity in out-of-school hours [100,101].

The first objective of this research was to determine the changes produced by a 10-week intervention promoted by the physical education school subject using step tracker mobile apps in out-of-school hours on physical activity, AMD, body composition, and physical condition of adolescents who are overweight and obese. The results showed that the EG improved their performance in the curl-up and push-up tests, while the CG showed no significant differences in any fitness test. These results are in line with previous research, in which the use of mobile apps for walking improved the physical fitness of adolescents aged between 12 and 16 years [19,102]. Following previous research in this area, a possible explanation for the improvement obtained in the curl-up test would be that the abdominal muscles are activated when walking, increasing their resistance [102]. Regarding the significant improvement in the push-up test, this could be because the initial physical fitness level of the participants was very low, with an average of 2 repetitions performed in this test, so the improvement achieved with the training program could be sufficient to increase performance in certain physical fitness tests [88,103]. However, in the specific case of the present research, no significant improvements in physical activity were observed, nor was there a direct influence of the distance traveled with the use of the app in these variables, so it would be difficult to assume that the changes are due to the intervention. Therefore, another possible explanation would be the maturational process in which the adolescents are immersed since, as the characteristics of the sample show, most of them had already passed the APHV [104]. Around and behind the APHV, hormonal and physical changes occur that lead to changes in body composition and athletic performance [105], with boys increasing their strength and production of hormones, such as testosterone, to a greater extent [105]. This becomes even more relevant when considering that the covariate gender did have an influence on performance in both physical tests, with boys performing better, as was the case in previous research [106]. However, the small sample size makes it difficult to analyze differences according to gender and maturational status, so future studies should address both questions. This would allow us to discover whether this type of intervention program is effective in improving the physical fitness of adolescents who are overweight or obese, or whether it is the gender differences or the effect of the maturational process that led to the changes.

It is important to highlight that the EG did not show significant improvements in the performance of the other fitness tests (20-m shuttle run test, sit-and-reach, handgrip strength, CMJ, and 20-m sprint), which could be because improvements in these tests require a moderate to vigorous intensity of physical practice, as shown in previous scientific literature [107]. The intervention program used included only walking activities, so it is likely that the intensity at which they performed the training was a determining factor in the lack of benefits [108]. In addition, adherence to the training program was low, with >50% (n=14) of the adolescents dropping out of regular training after the sixth week and <12% (n=3) meeting targets at the end of 10 weeks. This makes the intervention program too short, which may influence the benefits obtained, as previous interventions show that programs with a minimum of 10 weeks are the ones that truly lead to significant changes [19]. Therefore, future research should place more emphasis on training parameters, such as volume and intensity, as these are relevant aspects in obtaining

XSL•FO
RenderX

improvements with aerobic training, as shown in previous scientific literature [109,110]. Regarding body composition, the results showed no significant changes in the EG in the variables related to fat mass. The absence of changes in the present investigation could be because the adolescents were not instructed on the intensity at which the walking should be performed, so it is likely that a moderate to vigorous intensity that would produce changes in body composition was not reached [108,110]. This is similar to previous research that compared changes in body composition in adolescents after participating in high-, medium-, and low-intensity aerobic exercise programs, with adolescents in the low-intensity program having the smallest gains in body composition [111].

However, in the variables related to muscle mass, significant changes were observed in both the EG and CG, as an increased corrected arm girth in the EG and corrected thigh girth and muscle mass in the CG. Given that the intervention did not include strength exercises and that those improvements occurred in both groups, one explanation for the increase in muscle mass-related variables would be the increase in the production of steroid hormones that occurs during adolescence [105]. This favors the development of muscle mass and the gain in strength [71]. AMD showed no significant differences between pre- and postintervention measurements in any of the groups analyzed. These results are in line with previous research on adolescents, in which the use of mobile physical activity apps had no influence on AMD [19]. The rationale for including AMD in a physical activity intervention lay in the fact that previous research had shown that those who were more physically active had greater AMD [26,27]. However, in this research, no significant improvement in physical activity level was achieved, which would hinder the improvement of other healthy habits, such as AMD. This could be because the apps currently available on the market for physical activity do not include nutritional content, making it difficult to achieve benefits in this area, as the interventions carried out specifically with mobile nutrition apps do show significant improvements [112]. It is worth noting that the covariates gender and age significantly affected the AMD of the CG adolescents. This is consistent with previous research that found significant differences in AMD between adolescent boys and girls [112], as well as a decrease in AMD adherence with age [27]. Future research should consider mobile apps that include physical activity and nutrition together, as the benefits obtained could be greater for adolescents, as shown in physical exercise programs combined with nutrition [107,112].

The results obtained allow us to reject the first research hypothesis, as it was expected that the promotion of the use of mobile apps from the school subject of physical education would lead to significant benefits in the level of physical activity, favoring changes in body composition, greater AMD, and an improvement in the physical condition of this population. However, the results show that the adolescents of the EG did not achieve significant improvements in the level of physical activity, so no changes were found in most of the variables of body composition, AMD, or physical fitness. Therefore, the proposed program may not be effective, probably due to a lack of adherence, as most of the adolescents stopped regular physical activity after the sixth week. Future research is needed to draw conclusions in this area.

The second aim of this research was to analyze the changes achieved by the 10-week intervention on the psychological state of adolescents who are overweight and obese. The results showed a significant decrease in the competence variable in the CG. Previous research has shown that adolescents with less physical activity showed lower scores on the competence variable [107]. Furthermore, it is worth noting that no changes in the perception of competence were found in the adolescents belonging to the EG. This could be because the intervention was not sufficient for adolescents to perceive that they had increased their physical activity. This may be due to a lack of compliance by participants. Therefore, given these preliminary results, future research is needed to analyze the relationship between the use of mobile apps and the psychological state of the adolescents who are overweight and obese, as well as the relationship with other variables, such as motivation or engagement.

The inclusion of the covariates gender and age showed significant effects on the competence and relatedness of CG adolescents. This is an aspect to be considered, as previous research has shown that physical activity practice is lower in adolescent girls [113], as well as in older adolescents [114]. These groups place greater importance on physical activity to promote social interaction compared to boys and younger adolescents [113,115]. In addition, during adolescence, body image becomes particularly relevant, even altering healthy behaviors, such as nutrition and physical activity [116], with women being particularly affected in this respect [116]. Despite the results obtained, future research with a larger sample is needed to perform a separate analysis according to gender and age to know the real effect of mobile apps on the psychological state of adolescents who are overweight and obese.

The results obtained allow us to reject the second research hypothesis, which stated that increased physical activity has a positive effect on the psychological state of adolescents who are overweight and obese, because only significant decreases were observed in the scores of the psychological variables of the CG, with no changes in the EG. Furthermore, future research should give special relevance to the analysis of differences in psychological state between men and women of different ages who are overweight and obese and who use mobile apps for physical activity.

Although the present research is pioneering in the promotion of physical activity in the adolescent population that is overweight or obese during out-of-school hours by means of mobile apps, it is not without limitations. First, the sample size was small, which made it impossible to divide it into gender or maturity groups. In addition, the small sample of adolescents who completed the intervention in its entirety precludes comparison of actual effects with the CG. Although AMD was assessed, caloric intake was not monitored, which could be a determining factor in the changes achieved in the body composition of the adolescents. The level of physical activity was assessed subjectively by the PAQ-A questionnaire but was not assessed objectively by means of accelerometry. Similarly,

XSL•FO
RenderX

psychological state was assessed by validated questionnaires, but the inclusion of other techniques, such as interviews, in future research could provide more information. Although the total distance traveled with the use of the app was measured, as well as the steps taken by each adolescent, the information that the adolescents walked before starting the intervention was not recorded, so it is not known whether there was an improvement in the daily steps taken. Finally, BMI ≥25 kg/m^2 has been used to define adolescents who are overweight and obese instead of WHO or National Center for Health Statistics growth charts, as it is a simple method [55,56] and is consistent with growth charts when measured accurately [52,55], as was done in the present investigation. However, future research could replicate this study using WHO or National Center for Health Statistics growth charts, especially if applied to a population of children rather than adolescents. All the above limitations affect the generalizability of the study results to the population of adolescents who are overweight and obese.

As for the practical applications derived from this study, it is observed that in adolescents who are overweight and obese, going for a walk in their free time offers benefits on physical condition and psychological state, or at least prevents a worsening of these variables. However, the proposal put forward does not seem to be sufficient to obtain improvements in body composition in this population, which could be because adherence to the intervention with mobile apps was reduced, and from the sixth week onward, few adolescents complied with the training plan. Therefore, the results obtained are relevant for physical education teachers and for the health field because mobile physical activity apps could be a useful element for the promotion of active time in adolescents who are overweight or obese, although future research is needed to provide more scientific evidence in this regard and to increase adherence to these programs. This may require the use of other mobile apps in which gamification is a major part of it, as in previous research, it has been observed that this aspect favors adherence [19] and controls aspects, such as intensity or training volume, which may be relevant to achieve benefits.

Conclusions

The findings of this study affirm that a 10-week program consisting of the use of mobile apps by adolescents who are overweight and obese to practice physical activity outside school hours is not effective in improving the level of physical activity in this population, and therefore no improvements in body composition, fitness, or psychological state were found. The lack of benefits could be because adherence to the program was very low, with >50% (14/26) of the adolescents failing to meet the objectives set after the sixth week of the intervention and <12% (3/26) meeting targets at the end of 10 weeks. Therefore, future research is needed to provide more evidence in this area and to select the optimal mobile apps to achieve benefits in this population.

Acknowledgments

This research has been financed by grants from the Ministry of Science and Innovation of the Spanish Government in the program knowledge generation (titled Promotion of Daily Physical Activity recommended for the Health of Adolescents through Mobile Apps, Wearables and a Gamified TAC; Call: 2022/reference code: PID2022-140245OA-I00). This work is part of the postgraduate studies of CMP-R. The authors would like to thank the schools, management teams, teachers, students, and research assistants who made it possible to carry out the measurements for this project.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request. The measurement protocol was registered before the start of the study at ClinicalTrials.gov (NCT06089876).

Authors' Contributions

AM-O, LA-C, and RV-C participated in conceptualization, data curation, formal analysis, and methodology. AM-O was responsible for investigation and validation. AM-O and CMP-R were responsible for writing the original draft. CMP-R, LA-C, and RV-C were responsible for project administration and supervision. CMP-R contributed to methodology. LA-C was responsible for visualization and writing (review and editing). RV-C was responsible for validation and writing (review and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Changes between the intervention and control groups.
[DOCX File , 301 KB - jmir_v26i1e55243_app1.docx]

Multimedia Appendix 2
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 83 KB - jmir_v26i1e55243_app2.pdf]

XSL•FO
RenderX

References

1. Obesity and overweight. World Health Organization. URL: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2023-09-13]
2. Wellman RJ, Sylvestre MP, Abi Nader P, Chiolero A, Mesidor M, Dugas EN, et al. Intensity and frequency of physical activity and high blood pressure in adolescents: a longitudinal study. J Clin Hypertens (Greenwich) 2020 Feb;22(2):283-290 [FREE Full text] [doi: 10.1111/jch.13806] [Medline: 31955514]
3. Fernandes RA, Zanesco A. Early physical activity promotes lower prevalence of chronic diseases in adulthood. Hypertens Res 2010 Sep;33(9):926-931. [doi: 10.1038/hr.2010.106] [Medline: 20574424]
4. Barbiellini Amidei C, Trevisan C, Dotto M, Ferroni E, Noale M, Maggi S, et al. Association of physical activity trajectories with major cardiovascular diseases in elderly people. Heart 2022 Mar;108(5):360-366. [doi: 10.1136/heartjnl-2021-320013] [Medline: 35165168]
5. Quazi A, Tankha G. A study of personality and body esteem of healthy, overweight, and obese adolescents. J Pharm Negat Results 2023:1473. [doi: 10.47750/pnr.2023.14.02.185]
6. Fontaine KR, Redden DT, Wang C, Westfall AO, Allison DB. Years of life lost due to obesity. JAMA 2003 Jan 08;289(2):187-193 [FREE Full text] [doi: 10.1001/jama.289.2.187] [Medline: 12517229]
7. Salas-Salvadó J, Rubio MA, Barbany M, Moreno B, Grupo Colaborativo de la SEEDO. [SEEDO 2007 consensus for the evaluation of overweight and obesity and the establishment of therapeutic intervention criteria]. Med Clin (Barc) 2007 Feb 10;128(5):184-200. [doi: 10.1016/s0025-7753(07)72531-9] [Medline: 17298782]
8. Gulías-González R, Martínez-Vizcaíno V, García-Prieto JC, Díez-Fernández A, Olivas-Bravo A, Sánchez-López M. Excess of weight, but not underweight, is associated with poor physical fitness in children and adolescents from Castilla-La Mancha, Spain. Eur J Pediatr 2014 Jun;173(6):727-735. [doi: 10.1007/s00431-013-2233-y] [Medline: 24326383]
9. Aranceta-Bartrina J, Serra-Majem L, Foz-Sala M, Moreno-Esteban B, Grupo Colaborativo SEEDO. Prevalencia de obesidad en España. Med Clin (Barc) 2005 Oct 08;125(12):460-466 [FREE Full text] [doi: 10.1157/13079612] [Medline: 16216203]
10. Valdés Pizarro J, Royo-Bordonada M. Prevalence of childhood obesity in Spain: National Health Survey 2006-2007. Nutr Hosp 2012;27(1):154-160. [doi: 10.1590/S0212-16112012000100018] [Medline: 22566315]
11. Frömel K, Groffik D, Valach P, ŠafáŘ M, MitáŠ J. The impact of distance education during the COVID-19 pandemic on physical activity and well-being of Czech and Polish adolescents. J Sch Health 2022 Dec;92(12):1137-1147 [FREE Full text] [doi: 10.1111/josh.13232] [Medline: 35916129]
12. Al Hourani H, Alkhatib B, Abdullah M. Impact of COVID-19 lockdown on body weight, eating habits, and physical activity of Jordanian children and adolescents. Disaster Med Public Health Prep 2022 Oct;16(5):1855-1863 [FREE Full text] [doi: 10.1017/dmp.2021.48] [Medline: 33588981]
13. Christofaro DG, De Andrade SM, Mesas AE, Fernandes RA, Farias Júnior JC. Higher screen time is associated with overweight, poor dietary habits and physical inactivity in Brazilian adolescents, mainly among girls. Eur J Sport Sci 2016;16(4):498-506 [FREE Full text] [doi: 10.1080/17461391.2015.1068868] [Medline: 26239965]
14. Guthold R, Stevens GA, Riley LM, Bull FC. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1·6 million participants. Lancet Child Adolesc Health 2020 Jan;4(1):23-35 [FREE Full text] [doi: 10.1016/S2352-4642(19)30323-2] [Medline: 31761562]
15. Merlo CL, Jones SE, Michael SL, Chen TJ, Sliwa SA, Lee SH, et al. Dietary and physical activity behaviors among high school students - youth risk behavior survey, United States, 2019. MMWR Suppl 2020 Aug 21;69(1):64-76 [FREE Full text] [doi: 10.15585/mmwr.su6901a8] [Medline: 32817612]
16. Lunn PD. The sports and exercise life-course: a survival analysis of recall data from Ireland. Soc Sci Med 2010 Mar;70(5):711-719. [doi: 10.1016/j.socscimed.2009.11.006] [Medline: 19962809]
17. da Costa BG, da Silva KS, da Silva JA, Minatto G, de Lima LR, Petroski EL. Sociodemographic, biological, and psychosocial correlates of light- and moderate-to-vigorous-intensity physical activity during school time, recesses, and physical education classes. J Sport Health Sci 2019 Mar;8(2):177-182 [FREE Full text] [doi: 10.1016/j.jshs.2017.05.002] [Medline: 30997264]
18. Guazzini A, Pesce A, Gino F, Duradoni M. How the COVID-19 pandemic changed adolescents' use of technologies, sense of community, and loneliness: a retrospective perception analysis. Behav Sci (Basel) 2022 Jul 13;12(7):228 [FREE Full text] [doi: 10.3390/bs12070228] [Medline: 35877298]
19. Mateo-Orcajada A, Abenza-Cano L, Albaladejo-Saura MD, Vaquero-Cristóbal R. Mandatory after-school use of step tracker apps improves physical activity, body composition and fitness of adolescents. Educ Inf Technol (Dordr) (Forthcoming) 2023 Jan 24:1-32 [FREE Full text] [doi: 10.1007/s10639-023-11584-0] [Medline: 36714445]
20. Carson V, Hunter S, Kuzik N, Gray CE, Poitras VJ, Chaput J, et al. Systematic review of sedentary behaviour and health indicators in school-aged children and youth: an update. Appl Physiol Nutr Metab 2016 Jun;41(6 Suppl 3):S240-S265 [FREE Full text] [doi: 10.1139/apnm-2015-0630] [Medline: 27306432]
21. Tremblay MS, Colley RC, Saunders TJ, Healy GN, Owen N. Physiological and health implications of a sedentary lifestyle. Appl Physiol Nutr Metab 2010 Dec;35(6):725-740. [doi: 10.1139/H10-079] [Medline: 21164543]
22. Chaput JP, Carson V, Gray C, Tremblay M. Importance of all movement behaviors in a 24 hour period for overall health. Int J Environ Res Public Health 2014 Dec 04;11(12):12575-12581 [FREE Full text] [doi: 10.3390/ijerph111212575] [Medline: 25485978]

XSL•FO
RenderX

23. Colley RC, Garriguet D, Janssen I, Wong SL, Saunders TJ, Carson V, et al. The association between accelerometer-measured patterns of sedentary time and health risk in children and youth: results from the Canadian Health Measures Survey. BMC Public Health 2013 Mar 07;13:200 [FREE Full text] [doi: 10.1186/1471-2458-13-200] [Medline: 23497190]
24. Sahu M, Gandhi S, Sharma MK. Mobile phone addiction among children and adolescents: a systematic review. J Addict Nurs 2019;30(4):261-268. [doi: 10.1097/JAN.0000000000000309] [Medline: 31800517]
25. Billieux J, Philippot P, Schmid C, Maurage P, De Mol J, Van der Linden M. Is dysfunctional use of the mobile phone a behavioural addiction? Confronting symptom-based versus process-based approaches. Clin Psychol Psychother 2015;22(5):460-468 [FREE Full text] [doi: 10.1002/cpp.1910] [Medline: 24947201]
26. García-Hermoso A, Ezzatvar Y, López-Gil JF, Ramírez-Vélez R, Olloquequi J, Izquierdo M. Is adherence to the Mediterranean diet associated with healthy habits and physical fitness? A systematic review and meta-analysis including 565 421 youths. Br J Nutr 2022 Oct 14;128(7):1433-1444. [doi: 10.1017/S0007114520004894] [Medline: 33292901]
27. Iaccarino Idelson P, Scalfi L, Valerio G. Adherence to the Mediterranean diet in children and adolescents: a systematic review. Nutr Metab Cardiovasc Dis 2017 Apr;27(4):283-299. [doi: 10.1016/j.numecd.2017.01.002] [Medline: 28254269]
28. Martinez-Gonzalez MA, Bes-Rastrollo M. Dietary patterns, Mediterranean diet, and cardiovascular disease. Curr Opin Lipidol 2014 Feb;25(1):20-26. [doi: 10.1097/MOL.0000000000000044] [Medline: 24370845]
29. Schwingshackl L, Missbach B, König J, Hoffmann G. Adherence to a Mediterranean diet and risk of diabetes: a systematic review and meta-analysis. Public Health Nutr 2015 May;18(7):1292-1299 [FREE Full text] [doi: 10.1017/S1368980014001542] [Medline: 25145972]
30. Sofi F, Cesari F, Abbate R, Gensini GF, Casini A. Adherence to Mediterranean diet and health status: meta-analysis. BMJ 2008 Sep 11;337:a1344 [FREE Full text] [doi: 10.1136/bmj.a1344] [Medline: 18786971]
31. Martinez-Lacoba R, Pardo-Garcia I, Amo-Saus E, Escribano-Sotos F. Mediterranean diet and health outcomes: a systematic meta-review. Eur J Public Health 2018 Oct 01;28(5):955-961. [doi: 10.1093/eurpub/cky113] [Medline: 29992229]
32. Mateo-Orcajada A, Vaquero-Cristóbal R, Esparza-Ros F, Abenza-Cano L. Physical, psychological, and body composition differences between active and sedentary adolescents according to the "Fat but Fit" paradigm. Int J Environ Res Public Health 2022 Aug 30;19(17):10797 [FREE Full text] [doi: 10.3390/ijerph191710797] [Medline: 36078538]
33. García-Solano M, Gutiérrez-González E, López-Sobaler AM, Dal Re Saavedra M, Robledo de Dios T, Villar-Villalba C, et al. Weight status in the 6 to 9 year-old school population in Spain: results of the ALADINO 2015 study. An Pediatr (Engl Ed) 2021 Jun;94(6):366-376 [FREE Full text] [doi: 10.1016/j.anpede.2020.05.010] [Medline: 34090633]
34. Stavridou A, Kapsali E, Panagouli E, Thirios A, Polychronis K, Bacopoulou F, et al. Obesity in children and adolescents during COVID-19 pandemic. Children (Basel) 2021 Feb 12;8(2):135 [FREE Full text] [doi: 10.3390/children8020135] [Medline: 33673078]
35. Mateo-Orcajada A, Vaquero-Cristóbal R, Albaladejo-Saura MD, Abenza-Cano L. The degree of problematic technology use negatively affects physical activity level, adherence to Mediterranean diet and psychological state of adolescents. Healthcare (Basel) 2023 Jun 11;11(12):1706 [FREE Full text] [doi: 10.3390/healthcare11121706] [Medline: 37372824]
36. Hallal PC, Victora CG, Azevedo MR, Wells JC. Adolescent physical activity and health: a systematic review. Sports Med 2006;36(12):1019-1030. [doi: 10.2165/00007256-200636120-00003] [Medline: 17123326]
37. Gómez-Baya D, Grasmeijer AJ, López-Bermúdez E, Gaspar de Matos M, Mendoza R. Positive youth development and internet use in a sample of Spanish adolescents. Front Pediatr 2022 May 23;10:842928 [FREE Full text] [doi: 10.3389/fped.2022.842928] [Medline: 35676903]
38. Real Decreto 217/2022, de 29 de marzo, por el que se establece la ordenación y las enseñanzas mínimas de la Educación Secundaria Obligatoria. Boletín Oficial del Estado. 2018. URL: https://www.boe.es/eli/es/rd/2022/03/29/217/con [accessed 2024-12-02]
39. Costa M, Oliveira T, Mota J, Santos MP, Ribeiro JC. Objectively measured physical activity levels in physical education classes and body mass index (niveles de actividad física medida objetivamente en las clases de educación física y el índice de masa grasa). Retos 2016 Nov 21(31):271-274. [doi: 10.47197/retos.v0i31.53502]
40. Bogataj Š, Trajković N, Cadenas-Sanchez C, Sember V. Effects of school-based exercise and nutrition intervention on body composition and physical fitness in overweight adolescent girls. Nutrients 2021 Jan 15;13(1):238 [FREE Full text] [doi: 10.3390/nu13010238] [Medline: 33467641]
41. Carrel AL, Clark RR, Peterson SE, Nemeth BA, Sullivan J, Allen DB. Improvement of fitness, body composition, and insulin sensitivity in overweight children in a school-based exercise program: a randomized, controlled study. Arch Pediatr Adolesc Med 2005 Oct 01;159(10):963-968. [doi: 10.1001/archpedi.159.10.963] [Medline: 16203942]
42. Liu Y, Wadey CA, Barker AR, Williams CA. Process evaluation of school-based high-intensity interval training interventions for children and adolescents: a systematic review and meta-analysis of randomized controlled trials. BMC Public Health 2024 Feb 02;24(1):348 [FREE Full text] [doi: 10.1186/s12889-024-17786-6] [Medline: 38308213]
43. Domaradzki J, Cichy I, Rokita A, Popowczak M. Effects of tabata training during physical education classes on body composition, aerobic capacity, and anaerobic performance of under-, normal- and overweight adolescents. Int J Environ Res Public Health 2020 Jan 30;17(3):876 [FREE Full text] [doi: 10.3390/ijerph17030876] [Medline: 32019253]
44. Dauenhauer B, Keating X, Stoepker P, Knipe R. State physical education policy changes from 2001 to 2016. J Sch Health 2019 Jun;89(6):485-493. [doi: 10.1111/josh.12757] [Medline: 30919961]

XSL•FO
RenderX

45. Thorburn M, Stolz S. Embodied learning and school-based physical culture: implications for professionalism and practice in physical education. Sport Educ Soc 2015 Jul 13;22(6):721-731. [doi: 10.1080/13573322.2015.1063993]
46. Mateo-Orcajada A, Vaquero-Cristóbal R, Abenza-Cano L. Mobile application interventions to increase physical activity and their effect on kinanthropometrics, body composition and fitness variables in adolescent aged 12-16 years old: an umbrella review. Child Care Health Dev 2024 Jan 30;50(1):e13146. [doi: 10.1111/cch.13146] [Medline: 37387258]
47. Gil-Espinosa FJ, Merino-Marbán R, Mayorga-Vega D. Endomondo smartphone app to promote physical activity in high school students. Cult Cienc Depote 2020;15(46):463-471 [FREE Full text]
48. Quelly SB, Norris AE, DiPietro JL. Impact of mobile apps to combat obesity in children and adolescents: a systematic literature review. J Spec Pediatr Nurs 2016 Jan;21(1):5-17. [doi: 10.1111/jspn.12134] [Medline: 26494019]
49. Vajravelu ME, Arslanian S. Mobile health and telehealth interventions to increase physical activity in adolescents with obesity: a promising approach to engaging a hard-to-reach population. Curr Obes Rep 2021 Dec;10(4):444-452 [FREE Full text] [doi: 10.1007/s13679-021-00456-8] [Medline: 34596867]
50. Ponce-Ramírez CM, Mateo-Orcajada A, Abenza-Cano L, Vaquero-Cristóbal R. Assessment of step tracker mobile applications for the promotion of physical activity by adolescents based on their weight status. Int J Telemed Appl 2024 Sep 30;2024(1):8038334 [FREE Full text] [doi: 10.1155/2024/8038334] [Medline: 39381367]
51. Adams MA, Caparosa S, Thompson S, Norman GJ. Translating physical activity recommendations for overweight adolescents to steps per day. Am J Prev Med 2009 Aug;37(2):137-140 [FREE Full text] [doi: 10.1016/j.amepre.2009.03.016] [Medline: 19524391]
52. Krebs NF, Himes JH, Jacobson D, Nicklas TA, Guilday P, Styne D. Assessment of child and adolescent overweight and obesity. Pediatrics 2007;120(Supplement_4):S193-S228. [doi: 10.1542/peds.2007-2329d]
53. Freedman DS, Sherry B. The validity of BMI as an indicator of body fatness and risk among children. Pediatrics 2009;124(Supplement_1):S23-S34. [doi: 10.1542/peds.2008-3586e]
54. Wright CM, Cole TJ, Fewtrell M, Williams JE, Eaton S, Wells JC. Body composition data show that high BMI centiles overdiagnose obesity in children aged under 6 years. Am J Clin Nutr 2022 Jul 06;116(1):122-131 [FREE Full text] [doi: 10.1093/ajcn/nqab421] [Medline: 34967839]
55. Himes JH. Challenges of accurately measuring and using BMI and other indicators of obesity in children. Pediatrics 2009;124(Supplement_1):S3-22. [doi: 10.1542/peds.2008-3586d]
56. Daniels SR. The use of BMI in the clinical setting. Pediatrics 2009;124(Supplement_1):S35-S41. [doi: 10.1542/peds.2008-3586f]
57. Simmonds M, Burch J, Llewellyn A, Griffiths C, Yang H, Owen C, et al. The use of measures of obesity in childhood for predicting obesity and the development of obesity-related diseases in adulthood: a systematic review and meta-analysis. Health Technol Assess 2015 Jun;19(43):1-336 [FREE Full text] [doi: 10.3310/hta19430] [Medline: 26108433]
58. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ 2007 Sep 01;85(9):660-667 [FREE Full text] [doi: 10.2471/blt.07.043497] [Medline: 18026621]
59. Benítez-Porres J, Alvero-Cruz JR, Sardinha LB, López-Fernández I, Carnero EA. Valores de corte para clasificar niños y adolescentes activos utilizando el Cuestionario de Actividad Física: PAQ-C y PAQ-A. Nutr Hosp 2016 Sep 20;33(5):564 [FREE Full text] [doi: 10.20960/nh.564] [Medline: 27759968]
60. Martínez-Gómez D, Martínez-de-Haro V, Pozo T, Welk GJ, Villagra A, Calle ME, et al. Fiabilidad y validez del cuestionario de actividad física PAQ-A en adolescentes españoles. Rev Esp Salud Publica 2009 Jun;83(3):427-439 [FREE Full text] [doi: 10.1590/s1135-57272009000300008] [Medline: 19701574]
61. Serra-Majem L, Ribas L, Ngo J, Ortega RM, García A, Pérez-Rodrigo C, et al. Food, youth and the Mediterranean diet in Spain. Development of KIDMED, Mediterranean Diet Quality Index in children and adolescents. Public Health Nutr 2004 Oct;7(7):931-935. [doi: 10.1079/phn2004556] [Medline: 15482620]
62. Wilson PM, Rogers WT, Rodgers WM, Wild T. The psychological need satisfaction in exercise scale. J Sport Exerc Psychol 2006;28(3):15. [doi: 10.1123/jsep.28.3.231]
63. Diener ED, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985 Feb 10;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
64. Atienza FL, Pons D, Balaguer I, García-Merita M. Propiedades psicométricas de la Escala de Satisfacción con la Vida en adolescentes. Psicothema 2000;12:319.
65. Moreno-Murcia JA, Marzo JC, Martínez-Galindo C, Conte L. Validación de la Escala de “Satisfacción de las Necesidades Psicológicas Básicas” y del Cuestionario de la “Regulación Conductual en el Deporte” al contexto español. (Validation of Psychological Need Satisfaction in Exercise Scale and the Behavioural Regulation in Sport Questionnaire to the Spanish context). Rev Int Cienc Deporte 2011 Oct 01;7(26):355-369. [doi: 10.5232/ricyde2011.02602]
66. Esparza-Ros F, Vaquero-Cristóbal R, Marfell-Jones M. International Standars for Anthropometric Assessment. Glasgow, UK: International Society for the Advancement of Kinanthropometry; 2019.
67. Slaughter MH, Lohman TG, Boileau R, Horswill CA, Stillman RJ, Van Loan MD, et al. Skinfold equations for estimation of body fatness in children and youth. Hum Biol 1988 Oct;60(5):709-723. [Medline: 3224965]

XSL•FO
RenderX

68. Poortmans JR, Boisseau N, Moraine JJ, Moreno-Reyes R, Goldman S. Estimation of total-body skeletal muscle mass in children and adolescents. Med Sci Sports Exerc 2005 Feb;37(2):316-322. [doi: 10.1249/01.mss.0000152804.93039.ce] [Medline: 15692329]
69. Mirwald RL, Baxter-Jones AD, Bailey DA, Beunen GP. An assessment of maturity from anthropometric measurements. Med Sci Sports Exerc 2002 Apr;34(4):689-694. [doi: 10.1097/00005768-200204000-00020] [Medline: 11932580]
70. Malina RM, Bouchard C. Growth, maturation, and physical activity. Med Sci Sports Exerc 1992;24(7):841. [doi: 10.1249/00005768-199207000-00018]
71. Albaladejo-Saura M, Vaquero-Cristóbal R, García-Roca JA, Esparza-Ros F. The effect of age, biological maturation and birth quartile in the kinanthropometric and physical fitness differences between male and female adolescent volleyball players. Children (Basel) 2022 Jan 04;9(1):58 [FREE Full text] [doi: 10.3390/children9010058] [Medline: 35053683]
72. Léger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. J Sports Sci 1988;6(2):93-101. [doi: 10.1080/02640418808729800] [Medline: 3184250]
73. Mayorga-Vega D, Merino-Marban R, Viciana J. Criterion-related validity of sit-and-reach tests for estimating hamstring and lumbar extensibility: a meta-analysis. J Sports Sci Med 2014 Jan;13(1):1-14 [FREE Full text] [Medline: 24570599]
74. López-Miñarro P, Vaquero-Cristóbal R, Muyor JM, Alacid F, Isorna M. Validez de criterio del test sit-and-reach como medida de la extensibilidad isquiosural en piragüistas. (Criterion-related validity of the sit-and-reach test as a measure of hamstring extensibility in paddlers). Cult Cienc Depote 2012 Jul 01;7(20):95-101. [doi: 10.12800/ccd.v7i20.55]
75. España-Romero V, Ortega FB, Vicente-Rodríguez G, Artero EG, Rey JP, Ruiz JR. Elbow position affects handgrip strength in adolescents: validity and reliability of Jamar, DynEx, and TKK dynamometers. J Strength Cond Res 2010 Jan;24(1):272-277. [doi: 10.1519/JSC.0b013e3181b296a5] [Medline: 19966590]
76. Matsudo VK, Matsudo SM, de Rezende LF, Raso V. Handgrip strength as a predictor of physical fitness in children and adolescents. Rev Bras Cineantropom Desempenho Hum 2014 Dec 28;17(1):1. [doi: 10.5007/1980-0037.2015v17n1p1]
77. España-Romero V, Ortega FB, Vicente-Rodriguez G, Artero E, Pablo-Rey J, Ruiz J. Elbow position afects handgrip strength in adolescents: validity and reliability of Jamar, Dynex, and TKK dynamometers. J Strength Cond Res 2010;24(1):277. [doi: 10.1519/jsc.0b013e3181b296a5]
78. Castro-Piñero J, Ortega F, Artero E, Girela-Rejón MJ, Mora J, Sjöström M, et al. Assessing muscular strength in youth: usefulness of standing long jump as a general index of muscular fitness. J Strength Cond Res 2010 Jul;24(7):1810-1817. [doi: 10.1519/JSC.0b013e3181ddb03d] [Medline: 20555277]
79. Barker LA, Harry JR, Mercer JA. Relationships between countermovement jump ground reaction forces and jump height, reactive strength index, and jump time. J Strength Cond Res 2018 Jan;32(1):248-254. [doi: 10.1519/JSC.0000000000002160] [Medline: 28746248]
80. Léger LA, Lambert J. A maximal multistage 20-m shuttle run test to predict VO2 max. Eur J Appl Physiol Occup Physiol 1982;49(1):1-12. [doi: 10.1007/BF00428958] [Medline: 7201922]
81. Bastida Castillo A, Gómez Carmona CD, Pino Ortega J, de la Cruz Sánchez E. Validity of an inertial system to measure sprint time and sport task time: a proposal for the integration of photocells in an inertial system. Int J Perform Anal Sport 2017 Sep 13;17(4):600-608. [doi: 10.1080/24748668.2017.1374633]
82. Altmann S, Spielmann M, Engel FA, Neumann R, Ringhof S, Oriwol D, et al. Validity of single-beam timing lights at different heights. J Strength Cond Res 2017 Jul;31(7):1994-1999. [doi: 10.1519/JSC.0000000000001889] [Medline: 28277431]
83. Garcia-Pastor T, Salinero JJ, Sanz-Frias D, Pertusa G, Del Coso J. Body fat percentage is more associated with low physical fitness than with sedentarism and diet in male and female adolescents. Physiol Behav 2016 Oct 15;165:166-172. [doi: 10.1016/j.physbeh.2016.07.016] [Medline: 27474415]
84. Díaz-Soler MA, Vaquero-Cristóbal R, Espejo-Antúnez L, López-Miñarrofecto PA. Efecto de un protocolo de calentamiento en la distancia alcanzada en el test sit-and-reach en alumnos adolescentes. Nutr Hosp 2015;31(6):2623 [FREE Full text]
85. Siklar Z, Sanli E, Dallar Y, Tanyer G. Diurnal variation of height in children. Pediatr Int 2005 Dec 21;47(6):645-648. [doi: 10.1111/j.1442-200x.2005.02144.x] [Medline: 16354217]
86. Coburn JW, Malek MH. Manual NSCA: Fundamentos del Entrenamiento Personal. Barcelona, Spain: Paidotribo; 2014.
87. Lubans DR, Plotnikoff RC, Miller A, Scott JJ, Thompson D, Tudor-Locke C. Using pedometers for measuring and increasing physical activity in children and adolescents. Am J Lifestyle Med 2014 Jul 10;9(6):418-427. [doi: 10.1177/1559827614537774]
88. Mateo-Orcajada A, Vaquero-Cristóbal R, Abenza-Cano L. Importance of training volume through the use of step trackers apps promoted from the subject of physical education to change body composition, physical fitness and physical activity in adolescents and the influence of gender. Physiol Behav 2024 Jan 01;273:114402 [FREE Full text] [doi: 10.1016/j.physbeh.2023.114402] [Medline: 37935300]
89. Morency C, Demers M, Lapierre L. How many steps do you have in reserve?: thoughts and measures about a healthier way to travel. Transp Res Rec 2007 Jan 01;2002(1):1-6. [doi: 10.3141/2002-01]
90. Couture J. Reflections from the 'Strava-Sphere': kudos, community, and (self-)surveillance on a social network for athletes. Qual Res Sport Exerc Health 2020 Dec 07;13(1):184-200. [doi: 10.1080/2159676x.2020.1836514]

XSL•FO
RenderX

91. Bondaronek P, Alkhaldi G, Slee A, Hamilton FL, Murray E. Quality of publicly available physical activity apps: review and content analysis. JMIR Mhealth Uhealth 2018 Mar 21;6(3):e53 [FREE Full text] [doi: 10.2196/mhealth.9069] [Medline: 29563080]
92. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc 2009 Jan;41(1):3-13. [doi: 10.1249/MSS.0b013e31818cb278] [Medline: 19092709]
93. Costigan SA, Eather N, Plotnikoff RC, Taaffe DR, Lubans DR. High-intensity interval training for improving health-related fitness in adolescents: a systematic review and meta-analysis. Br J Sports Med 2015 Oct 18;49(19):1253-1261. [doi: 10.1136/bjsports-2014-094490] [Medline: 26089322]
94. Domaradzki J, Koźlenia D, Popowczak M. Prevalence of positive effects on body fat percentage, cardiovascular parameters, and cardiorespiratory fitness after 10-week high-intensity interval training in adolescents. Biology (Basel) 2022 Mar 10;11(3):424 [FREE Full text] [doi: 10.3390/biology11030424] [Medline: 35336798]
95. Yuksel HS, Şahin FN, Maksimovic N, Drid P, Bianco A. School-based intervention programs for preventing obesity and promoting physical activity and fitness: a systematic review. Int J Environ Res Public Health 2020 Jan 03;17(1):347 [FREE Full text] [doi: 10.3390/ijerph17010347] [Medline: 31947891]
96. Andermo S, Hallgren M, Nguyen TT, Jonsson S, Petersen S, Friberg M, et al. School-related physical activity interventions and mental health among children: a systematic review and meta-analysis. Sports Med Open 2020 Jun 16;6(1):25 [FREE Full text] [doi: 10.1186/s40798-020-00254-x] [Medline: 32548792]
97. Pate RR, O'Neill JR. After-school interventions to increase physical activity among youth. Br J Sports Med 2009 Jan 19;43(1):14-18. [doi: 10.1136/bjsm.2008.055517] [Medline: 19019903]
98. Tapia-Serrano M, López-Gajardo MA, Sánchez-Miguel PA, González-Ponce I, García-Calvo T, Pulido JJ, et al. Effects of out-of-school physical activity interventions based on self-determination theory in children and adolescents: a systematic review and meta-analysis. Scand J Med Sci Sports 2023 Oct 28;33(10):1929-1947. [doi: 10.1111/sms.14436] [Medline: 37381660]
99. Cale L, Harris J, Duncombe R. Promoting physical activity in secondary schools. Eur Phys Educ Rev 2016 Jul 24;22(4):526-544. [doi: 10.1177/1356336x15623774]
100. Alonso-Fernández D, Gutiérrez-Sánchez Á, Portela-Pino I, Taboada-Iglesias Y. Evaluation of applications for mobile devices on the practice of physical exercise in adolescents. Appl Sci 2022 Mar 08;12(6):2784. [doi: 10.3390/app12062784]
101. Mateo-Orcajada A, Vaquero-Cristóbal R, Mota J, Abenza-Cano L. Physical activity, body composition, and fitness variables in adolescents after periods of mandatory, promoted or nonmandatory, nonpromoted use of step tracker mobile apps: randomized controlled trial. JMIR Mhealth Uhealth 2024 Jul 30;12:e51206 [FREE Full text] [doi: 10.2196/51206] [Medline: 39079110]
102. Shnayderman I, Katz-Leurer M. An aerobic walking programme versus muscle strengthening programme for chronic low back pain: a randomized controlled trial. Clin Rehabil 2013 Mar 31;27(3):207-214. [doi: 10.1177/0269215512453353] [Medline: 22850802]
103. Liao T, Duhig SJ, Du G, Luo B, Wang YT. The effect of a functional strength training intervention on movement quality and physical fitness in adolescents. Percept Mot Skills 2022 Feb;129(1):176-194. [doi: 10.1177/00315125211056865] [Medline: 34784820]
104. Baxter-Jones AD, Barbour-Tuck EN, Dale D, Sherar LB, Knight CJ, Cumming SP, et al. The role of growth and maturation during adolescence on team-selection and short-term sports participation. Ann Hum Biol 2020 Jun;47(4):316-323 [FREE Full text] [doi: 10.1080/03014460.2019.1707870] [Medline: 31960720]
105. Rico-Lara T, Mateo-Orcajada A, Abenza-Cano L, Esparza-Ros F, Vaquero-Cristóbal R. Differences in psychological variables and the performance of female futsal players according to playing position, match result and menstruation. Int J Environ Res Public Health 2023 Apr 07;20(8):5429 [FREE Full text] [doi: 10.3390/ijerph20085429] [Medline: 37107711]
106. Steene-Johannessen J, Hansen BH, Dalene KE, Kolle E, Northstone K, Møller NC, Determinants of Diet and Physical Activity knowledge hub (DEDIPAC); International Children's Accelerometry Database (ICAD) Collaborators, IDEFICS Consortium and HELENA Consortium. Variations in accelerometry measured physical activity and sedentary time across Europe - harmonized analyses of 47,497 children and adolescents. Int J Behav Nutr Phys Act 2020 Mar 18;17(1):38 [FREE Full text] [doi: 10.1186/s12966-020-00930-x] [Medline: 32183834]
107. Mateo-Orcajada A, Abenza-Cano L, Cano-Martínez A, Vaquero-Cristóbal R. The importance of healthy habits to compensate for differences between adolescent males and females in anthropometric, psychological and physical fitness variables. Children (Basel) 2022 Dec 08;9(12):1926 [FREE Full text] [doi: 10.3390/children9121926] [Medline: 36553369]
108. Thivel D, Masurier J, Baquet G, Timmons BW, Pereira B, Berthoin S, et al. High-intensity interval training in overweight and obese children and adolescents: systematic review and meta-analysis. J Sports Med Phys Fitness 2019 Feb;59(2):310-324. [doi: 10.23736/S0022-4707.18.08075-1] [Medline: 29589408]
109. Batacan RB, Duncan MJ, Dalbo VJ, Tucker PS, Fenning AS. Effects of high-intensity interval training on cardiometabolic health: a systematic review and meta-analysis of intervention studies. Br J Sports Med 2017 Mar;51(6):494-503. [doi: 10.1136/bjsports-2015-095841] [Medline: 27797726]
110. Cao M, Tang Y, Li S, Zou Y. Effects of high-intensity interval training and moderate-intensity continuous training on cardiometabolic risk factors in overweight and obesity children and adolescents: a meta-analysis of randomized controlled

XSL•FO
RenderX

trials. Int J Environ Res Public Health 2021 Nov 12;18(22):11905 [FREE Full text] [doi: 10.3390/ijerph182211905] [Medline: 34831659]
111. Chiu CH, Ko MC, Wu LS, Yeh DP, Kan NW, Lee PF, et al. Benefits of different intensity of aerobic exercise in modulating body composition among obese young adults: a pilot randomized controlled trial. Health Qual Life Outcomes 2017 Aug 24;15(1):168 [FREE Full text] [doi: 10.1186/s12955-017-0743-4] [Medline: 28836987]
112. De Cock N, Vangeel J, Lachat C, Beullens K, Vervoort L, Goossens L, et al. Use of fitness and nutrition apps: associations with body mass index, snacking, and drinking habits in adolescents. JMIR Mhealth Uhealth 2017 Apr 25;5(4):e58 [FREE Full text] [doi: 10.2196/mhealth.6005] [Medline: 28442455]
113. Peral-Suárez Á, Cuadrado-Soto E, Perea JM, Navia B, López-Sobaler AM, Ortega RM. Physical activity practice and sports preferences in a group of Spanish schoolchildren depending on sex and parental care: a gender perspective. BMC Pediatr 2020 Jul 07;20(1):337 [FREE Full text] [doi: 10.1186/s12887-020-02229-z] [Medline: 32635918]
114. Butt J, Weinberg RS, Breckon JD, Claytor RP. Adolescent physical activity participation and motivational determinants across gender, age, and race. J Phys Act Health 2011 Nov;8(8):1074-1083. [doi: 10.1123/jpah.8.8.1074] [Medline: 22039125]
115. Craggs C, Corder K, van Sluijs EM, Griffin SJ. Determinants of change in physical activity in children and adolescents: a systematic review. Am J Prev Med 2011 Jun;40(6):645-658 [FREE Full text] [doi: 10.1016/j.amepre.2011.02.025] [Medline: 21565658]
116. Vaquero-Cristóbal R, Alacid F, Muyor JM, López-Miñarro P. Imagen corporal; revisión bibliográfica. Nutr Hosp 2013;28(1):27-35 [FREE Full text] [doi: 10.3305/nh.2013.28.1.6016] [Medline: 23808427]

Abbreviations

AMD: adherence to the Mediterranean diet
APHV: age at peak height velocity
BPNS: Basic Psychological Needs Scale
CG: control group
CMJ: countermovement jump
CONSORT: Consolidated Standards of Reporting Trials
EG: experimental group
HIIT: high-intensity interval training
ISAK: International Society for the Advancement of Kinanthropometry
KIDMED: Mediterranean Diet Quality Index
PAQ-A: Physical Activity Questionnaire for Adolescents
SWLS: Satisfaction With Life Scale
TEM: technical errors of measurement
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 06.12.23; peer-reviewed by P Marcos-Pardo, J Domaradzki, AS Neto; comments to author 12.09.24; revised version received 10.10.24; accepted 29.10.24; published 10.12.24.

Please cite as:
Mateo-Orcajada A, Ponce-Ramírez CM, Abenza-Cano L, Vaquero-Cristóbal R
Effects of 10 Weeks of Walking With Mobile Step-Tracking Apps on Body Composition, Fitness, and Psychological State in Adolescents Who Are Overweight and Obese: Randomized Controlled Trial
J Med Internet Res 2024;26:e55243
URL: https://www.jmir.org/2024/1/e55243
doi:10.2196/55243
PMID:39656510

©Adrián Mateo-Orcajada, Cristina M Ponce-Ramírez, Lucía Abenza-Cano, Raquel Vaquero-Cristóbal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Personalized Mobile App–Based Program for Preparation and Recovery After Radical Prostatectomy: Initial Evidence for Improved Outcomes From a Prospective Nonrandomized Study

Alberto Martini[1], MD; Claudia Kesch[1], MD; Alae Touzani[1], MD; Giorgio Calleris[1], MD; Bogdan Buhas[1], MD; Rawad Abou-Zahr[1], MD; Razvan-George Rahota[1], MD; Benjamin Pradère[1], MD; Christophe Tollon[1], MD; Jean-Baptiste Beauval[1], MD; Guillaume Ploussard[1], MD

Department of Urology, La Croix du Sud Hospital, Quit Fonsegrives, France

Corresponding Author:
Alberto Martini, MD
Department of Urology
La Croix du Sud Hospital
52 Chemin De Ribaute
Quint-Fonsegrives, 31130
France
Phone: 33 56340913241
Email: a.martini.md@gmail.com

Abstract

Background: eHealth can help replicate the benefits of conventional surgical prehabilitation programs and overcome organizational constraints related to human resources and health care–related costs.

Objective: We aimed to assess the impact of an optimized perioperative program using a personalized mobile app designed for preparation and recovery after radical prostatectomy (RP).

Methods: We report on a series of 122 consecutive robot-assisted RP before and after the implementation of the *betty.care* app (cohort A: standard of care, n=60; cohort B: optimized program, n=62). The primary end point was continence recovery, defined as “0 or 1 safety pad per day” at 6 weeks after surgery. Secondary end points were length of stay, same-day discharge, complications, readmissions, and number of days alive and out of hospital within 30 days from surgery.

Results: Both cohorts were comparable in terms of age, prostate-specific antigen, prostate volume, and disease aggressiveness. Intraoperative parameters (lymph node dissection, operative time, and bilateral nerve-sparing surgery) were comparable in both groups, except for blood loss, which was significantly higher in cohort B (182 vs 125 cc; *P*=.008). The 6-week continence rate was improved in cohort B in both univariable and multivariable analyses (92% vs 75%; *P*=.01). There were trends favoring cohort B for all secondary end points with a minimal 30% benefit compared with cohort A. Grade 2 or more complications occurred less frequently in cohort B (13% vs 3.2%; *P*=.042). Same-day discharge and readmission rates were 35% and 53% (*P*=.043), and 3.3% and 1.6% (*P*=.54) in cohorts A and B, respectively. Mean length of stay was reduced by 0.2 days in cohort B (0.58 vs 0.78 days; *P*=.10). The main limitation was the absence of randomization.

Conclusions: The implementation of a mobile app that provides a holistic approach to the perioperative period, integrating prehabilitation, rehabilitation, and remote monitoring, could lead to the improvement of important functional outcomes after RP and could replicate an on-site prehabilitation program. Multicenter validation is needed.

(J Med Internet Res 2024;26:e55429) doi:10.2196/55429

KEYWORDS

prehabilitation; radical prostatectomy; robot, outcomes; continence; mobile app; app; electronic health; eHealth; surgical; human resource; health care; single-surgeon; implementation; betty.care app; cohort; perioperative; rehabilitation; mobile health; mHealth

XSL•FO
RenderX

Introduction

Despite the adoption of minimally invasive techniques, recovery after abdominal surgeries is still characterized by nonnegligible morbidity and mortality and significant health care expenditure [1,2]. Optimized perioperative pathways including enhanced recovery after surgery programs have demonstrated benefits in the oncologic surgery field by reducing hospitalization costs and perioperative complications, while maintaining suitable oncologic and functional outcomes [3-6]. Improved patient preparation and awareness to surgery, thanks to prehabilitation (PreHab) programs, helps to minimize side effects, improve patients' postoperative psychological and physiological status, and ease postoperative recovery [7,8]. Several studies assessed the impact of these programs in patients undergoing radical prostatectomy (RP); some evidence exists on the benefits of integrating physical exercise, optimizing nutrition, and preoperative patient counseling in improving postoperative outcomes [9,10]. Previous studies on RP patients highlighted that additional efforts should focus on patient education prior to surgery in order to promote early mobilization, return to work, and to improve patients' experience [11].

Our group demonstrated the benefits of optimizing the aforementioned pathways in the perioperative RP setting. Specifically, we found significant improvements in terms of postoperative metrics and health-related costs when a standard perioperative PreHab pathway was implemented [8,12-14]. The preoperative PreHab pathway included face-to-face workshops (pain management, bladder catheter, compression stockings, and postoperative care) and group-based seminars led by specialized nurses on diet, exercise, and physiotherapy (to reinforce pelvic floor muscles).

Such an implementation also led to meaningful improvements for patients' functional outcomes, with a faster return to urinary continence after surgery and a shorter length of stay [13]. However, physicians may face some issues in adopting and maintaining a dedicated PreHab program, mainly because of organization constraints, financial support, and lack of health care resources. On this matter, eHealth might represent a potential solution.

Over the past decade, eHealth has emerged as a promising tool to enhance (1) physician-patient communication, (2) adherence to educational programs, and (3) treatment outcomes. eHealth can be used throughout the whole perioperative setting, that is, before and after surgery, without the need for excessive human effort and financial support [15].

Therefore, eHealth, by means of a smartphone app, could promote the worldwide spread of optimized pathways and at-home recovery by overcoming the lack of health care resources that are not uncommon in many scenarios [16].

In this study, we aimed to assess the implementation of a mobile app–based program that was made available to every patient who underwent RP at our institution. Patients were asked to join the study by their treating physician, upon urological consultation. Our ultimate goal was to reproduce and spread the benefits originating from our former on-site PreHab program which determined better postoperative outcomes in terms of continence recovery, complications, and length of stay.

Methods

Overview

We designed a prospective study aimed at evaluating the implementation of eHealth in the perioperative RP setting: the BETTY-RP study. Patients were offered the free of charge possibility to enroll in the study. BETTY-RP study was approved by the local ethical committee.

Patients with prostate cancer who were candidates for robotic prostatectomy were consecutively enrolled in the study between January 2023 and November 2023. The study was carried out at La Croix du Sud Hospital—a private hospital located in a suburb of Toulouse, France. The minimal follow-up for data analysis was 6 weeks after surgery. Data for cohort A were collected from January 2023 to early July 2023, when the intervention with the *betty.care* app was introduced. Data for cohort B were collected from July 2023 to November 2023.

The overall cohort included all consecutive RP patients before and after the implementation of the mobile app program.

Surgeons enrolling patients in the study complete 3 of 4 robotic operations per week. Postoperatively, all patients were managed according to the enhanced recovery after surgery protocol, as part of our standard of care [8].

In the interventional arm (cohort B), all consecutive patients had access to eHealth. The mobile app was freely downloadable on App stores or through the website: https://betty.care. The app gave patients access to checklists, remote monitoring, and multiple prehabilitation and rehabilitation educational materials including a specific module for RP (*BETTY coaching*).

The patient app included checklists before key moments (anesthesiologist visit, admission, discharge, and postoperative visit), alerts for starting or stopping activities (medications, physical activity, compressive stockings, diet, and work), and educational materials (podcasts, videos on physical activities, and articles) for improving patient information and his condition before surgery including physiotherapy exercises tailored to RP patients. Advice on a home-based, moderate-intensity exercise regimen before surgery was provided, and patients were asked to perform presurgical pelvic floor exercises. Advice on walking programs, aerobic training, and cardiorespiratory fitness was also provided in order to improve preoperative patient condition. Dietetic counseling on weight loss and oral nutrition was also given. ePatient-Reported Outcomes and Experience Measurements were prospectively collected to record patient satisfaction, experience, and specific outcomes related to RP, such as continence and sexual function (through validated questionnaires).

The patient app communicated with a surgeon app (Betty pro App). Thereby, the surgeon could check at any time within his app the patient characteristics (medical history, identity, medications, and allergies) and could follow him daily via remote monitoring with a possibility of activating thresholds alerts.

XSL•FO
RenderX

The surgeon could also use an in-app easy-to-use risk stratification tool that assessed the individual risk of complication for his patients and allowed them to better adapt the postoperative surveillance phase [17]. Postoperative course was comparable and standardized for both arms. The RP technique and the patient workup and follow-up did not change during the study period. Thromboprophylaxis was given for 2 weeks. Analgesic (acetaminophen) was prescribed on demand. Bladder catheter was removed at day 7 without previous cystogram. Postoperative visits were scheduled at 6 weeks and 6 months after RP. In case of persistent urinary incontinence defined as >1 daily pad at 6 weeks after surgery, 20 sessions (2 per week) of physiotherapist-guided pelvic floor muscle training were prescribed.

Data were collected prospectively by a research assistant (LC) for both study cohorts. The primary end point was the continence recovery, strictly defined as no need for, or 1 safety pad per day at 6 weeks after RP. At 6 weeks patients were also asked to answer a questionnaire regarding their satisfaction with our Pre-Hab App. We also assessed outcomes at 3 months among patients who completed their 3-month follow-up. According to our previous on-site program study, we aimed at reproducing a 20% benefit in absolute value [14]. On this matter, the study was designed to demonstrate an improvement in the continence rate from 70% to 90% (a=5%; b=80%), which yielded 59 patients per group. A few patients more per group were allowed to account for potential dropout.

Secondary end points were length of stay, same-day discharge, complication, readmission, and number of days alive and out of hospital within 30 days after surgery. Perioperative complications were reported according to the Clavien-Dindo classification [18]. Comparisons between cohorts A and B were made according to the use of the optimized program in univariable and multivariable models. Parameters were compared using 2-tailed tests as appropriate. Univariate and multivariate logistic regression analyses were performed to compare outcomes. The limit of statistical significance was defined as $P<.05$. SPSS (version 22.0; IBM Corp) software was used for analysis.

Ethical Considerations

The study was approved by the institutional review board (IRB; National Ethics Committee “Ramsay santé Recherche et Enseignement” [IORG number: IORG0009085]; IRB number: IRB00010835/2023-11-004). This study was also approved by the local ethics committee of the La Croix du Sud Hospital and was conducted in line with the principles of the Declaration of Helsinki. All patients provided informed consent to participate to the study. Data were collected in a deidentified manner. No financial incentive was provided to patients to take part to the study.

Results

Overall, 62 consecutive patients (cohort B) were enrolled in the BETTY-RP study. Their data were compared with those of 60 patients (cohort A) who were operated prior to the implementation of our eHealth program. Mean age, preoperative prostate-specific antigen, prostate volume, estimated intraoperative blood loss, and operating time were comparable between the 2 groups, as shown in Table 1. The proportion of more advanced disease, in terms of ISUP3 and pT3 disease, did not differ between both groups.

XSL•FO
RenderX

Table 1. Baseline characteristics of the overall cohort (comparisons according to the standard of care vs optimized pathway status).

	Cohort A	Cohort B	*P* value
	SOC[a] (n=60)	Optimized pathway (n=62)	
Age (years)	65	64	.60
PSA[b] (ng/mL)	8.5	9.4	.40
Prostate volume (cc)	41	46	.30
MRI[c] lesion diameter (mm)	14	13	.80
Grade group: ≥3, n (%)	23 (38)	27 (44)	.80
pT[d] stage, n (%)			.90
pT3a	19 (32)	19 (31)	
pT3b	4 (7)	3 (5)	
LND[e], n (%)	31 (52)	33 (53)	.90
EBL[f] (mL)	125	182	.008
Operative time (minutes)	80	86	.10
LOS[g] (days)	0.78	0.58	.10
Bilateral NSS[h], n (%)	52 (87)	53 (86)	.90
Readmission, n (%)	2 (3.3)	1 (1.6)	.50
SDD[i], n (%)	21 (35)	33 (53)	.043
Complications, n (%)	8 (13)	2 (3.2)	.042
6-week continence, n (%)	45 (75)	57 (92)	.01

[a]SOC: standard of care.
[b]PSA: prostate-specific antigen.
[c]MRI: magnetic resonance imaging.
[d]pT: pathologic T stage.
[e]LND: lymph node dissection.
[f]EBL: estimated blood loss.
[g]LOS: length of stay.
[h]NSS: nerve-sparing surgery.
[i]SDD: same day discharge.

No difference in terms of intraoperative parameters was observed, except for blood loss which was significantly higher in cohort B (182 vs 125 cc; *P*=.008). Lymph node dissection was performed in 31 (52%) and 33 (53%) of cohorts A and B, respectively (*P*=.90). Operative time was comparable in both cohorts (80 and 86 minutes in cohorts A and B, respectively; *P*=.10). Bilateral nerve-sparing surgery (NSS) was performed in 52 (87%) and 53 (86%) of cohorts A and B patients, respectively (*P*=.90).

Regarding the primary end point, the 6-week continence rate was significantly improved in cohort B relative to cohort A (57, 92% vs 45, 75%; *P*=.01). For the subgroup of patients with a >3-month follow-up, this difference persisted over time (100% vs 95.0% in cohorts B and A; *P*=.08).

There were trends favoring cohort B for all secondary end points. Grade ≥2 complications occurred less frequently in cohort B (8, 13% vs 2, 3.2%; *P*=.042). Same-day discharge and readmission rates were 35% and 53% (*P*=.043), and 3.3% and 1.6% (*P*=.50) in cohorts A and B, respectively. A minimal 30% benefit was seen in cohort B patients for all these end points. Mean length of stay was lower for cohort B (0.58 vs 0.78 days; *P*=.10).

Predictive factors for continence recovery at 6 weeks were also assessed in multivariable analyses taking into account age, optimized pathway, and NSS (Table 2). The optimized Betty pathway was found to be an independent predictive factor for better continence rates at 6 weeks (odds ratio 3.8, 95% CI 1.3-11.6; *P*=.02). Age also significantly influenced this outcome (odds ratio 0.90; 95% CI 0.83-0.99; *P*=.04).

XSL•FO
RenderX

Table 2. Multivariable analysis of predictors for 6-week continence rate.

	OR[a]	95% CI	*P* value
Age (years)	0.9	0.83-0.99	.04
Optimized pathway	3.8	1.26-11.56	.02
Bilateral NSS[b]	1.1	0.27-4.47	.9

[a]OR: odds ratio.

[b]NSS: nerve-sparing surgery.

The preliminary results on the use of our mobile app demonstrated a high usability and satisfaction rate (>80% at the end of follow-up, evaluated through the app).

Discussion

Principal Findings

In this study, we demonstrated that an optimized perioperative pathway, through a mobile app, is associated with better postoperative outcomes and could provide similar advantages that of an on-site, structured program. In fact, we found similar results, in terms of postoperative outcomes, to the ones of our earlier series where an on-site PreHab program was implemented.

Overall, our eHealth program improved surgical outcomes after RP, independently of patient-related factors and surgery refinements. Notably, ours is an expert center that already demonstrated better surgery outcomes than those reported at a nationwide level [19].

The mobile app was able to reproduce the benefits observed in our previous on-site experience in terms of length of stay, complications, readmission, and short-term functional recovery [8,14]. While previous series suggested that prehabilitation could improve patient experience and satisfaction in the perioperative setting, our study evaluated a holistic approach to surgery, integrating prehabilitation, rehabilitation, and remote monitoring. We demonstrated that our eHealth-based pathway was associated with meaningful end points such as urinary continence after RP [5-7,13,14,20]. Interestingly, our optimized pathway also translated into a wider acceptance of same day discharge which results in cost abutments at the hospital and health care levels [12,20,21].

Several factors may concur to our findings. Interventions dedicated to improving patient information may help anticipating potential issues associated with oncologic surgery or surgery in general. The preoperative time is of great importance as patients may be more receptive to modify their physiological and psychological perceptions and to capitalize on advice and physical condition improvements before surgery. In the specific setting of RP, patients also highlighted the importance of early mobilization, fast return to work, compliance in pelvic floor exercises, and the benefit of a better education, which facilitated realistic expectations of the postoperative recovery pathway [11]. Moreover, several studies have demonstrated that pretreatment patient-centered interventions improved satisfaction and reduced regrets in comparison with usual care for patients with newly diagnosed localized prostate cancer [22,23].

We have previously noted significant improvements in patient outcomes and satisfaction within our on-site program [10]. Despite these advantages for patients, surgeons, and hospitals, our experience taught us that the maintenance of this program and its diffusion to other centers could be challenging, mainly due to the motivation of care teams, the absence of financial and incentive support for such programs, and the lack of human resources across different health care systems. Indeed, we were not able to offer an optimized on-site pathway for all our patients due to organization constraints (only 26% of RP patients benefited from our program which was active from 2018 to 2022). Moreover, the COVID-19 outbreak led to the temporary discontinuation of this program for several months and resulted in a trend toward worse annual outcomes [19].

Such pitfalls could be overcome by eHealth, accessible to almost every patient in the whole perioperative setting, without the need for large human resources and financial support [16]. As an example, a recent study has proven that a perioperative combined eHealth care program delivering personalized care by use of goal attainment scaling reduced the time required to return to normal activities after major abdominal surgeries [15]. Other mobile apps also showed potential in terms of usability and changing risk behavior prior to major surgery. Nevertheless, the proven benefits in terms of functional recovery remain unclear [24,25]. The preliminary results on the use of our mobile app demonstrated a high usability and satisfaction rate (>80% at the end of follow-up through the app). These trends should be evaluated in a larger cohort.

Limitations

First, the impact of the surgeon's experience could not be precisely assessed in this series. However, the surgeon was beyond the learning curve at the beginning of the study, and no modification of surgical technique or patient selection was done during the study period. No end point difference was noted in this subanalysis suggesting that the optimized pathway was the main factor explaining outcomes improvements, and that there was no bias related to increased experience over time. A multivariable analysis was also performed to control for potential patient- and surgery-related confounding factors. Second, no randomization was performed, and the impact of this program was not assessed in a prospective, multisurgeon, multicenter study. This paper mainly represents a proof-of-concept study that will lead to further prospective evaluation. In further research, it could be also relevant to assess new patient-centered end points such as return to work, return to active life, patient satisfaction, and overall well-being. Finally, data were not collected on the proportion of eligible patients who were enrolled in the study, nor the proportion of those who

XSL•FO
RenderX

completed the 6-week follow-up; therefore, response rate and attrition rates are unknown.

Conclusions

The implementation of a mobile app that provides a holistic approach to the perioperative period, integrating prehabilitation, rehabilitation, and remote monitoring, could lead to the improvement of important functional outcomes after RP. The use of such an eHealth pathway is effortless and patient could potentially replace and improve an on-site PreHab program. Further studies as well as external series are necessary to confirm our findings.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to manuscript writing, editing, and critical revision. Statistical analysis was done by GP. Supervision was done by GP and JBB.

Conflicts of Interest

AM and GP have own shares of Oltre Medical Consulting, LLC. JBB and GP have own shares of AIMED2, LLC.

References

1. Baltaxe E, Hsieh HW, Roca J, Cano I. The assessment of medical device software supporting health care services for chronic patients in a tertiary hospital: overarching study. J Med Internet Res 2023;25:e40976. [doi: 10.2196/40976] [Medline: 36598817]
2. Wang T, Stanforth PR, Fleming RYD, Wolf JS, Stanforth D, Tanaka H. A mobile app with multimodality prehabilitation programs for patients awaiting elective surgery: development and usability study. JMIR Perioper Med 2021;4(2):e32575 [FREE Full text] [doi: 10.2196/32575] [Medline: 34967752]
3. Azhar RA, Bochner B, Catto J, Goh AC, Kelly J, Patel HD, et al. Enhanced recovery after urological surgery: a contemporary systematic review of outcomes, key elements, and research needs. Eur Urol 2016;70(1):176-187 [FREE Full text] [doi: 10.1016/j.eururo.2016.02.051] [Medline: 26970912]
4. Kehlet H, Wilmore DW. Multimodal strategies to improve surgical outcome. Am J Surg 2002;183(6):630-641. [doi: 10.1016/s0002-9610(02)00866-8] [Medline: 12095591]
5. Singh F, Newton RU, Galvão DA, Spry N, Baker MK. A systematic review of pre-surgical exercise intervention studies with cancer patients. Surg Oncol 2013;22(2):92-104. [doi: 10.1016/j.suronc.2013.01.004] [Medline: 23434347]
6. Santa Mina D, Hilton WJ, Matthew AG, Awasthi R, Bousquet-Dion G, Alibhai SM, et al. Prehabilitation for radical prostatectomy: a multicentre randomized controlled trial. Surg Oncol 2018;27(2):289-298. [doi: 10.1016/j.suronc.2018.05.010] [Medline: 29937184]
7. Minnella EM, Awasthi R, Bousquet-Dion G, Ferreira V, Austin B, Audi C, et al. Multimodal prehabilitation to enhance functional capacity following radical cystectomy: a randomized controlled trial. Eur Urol Focus 2021;7(1):132-138. [doi: 10.1016/j.euf.2019.05.016] [Medline: 31186173]
8. Ploussard G, Almeras C, Beauval J, Gautier J, Garnault V, Frémont N, et al. A combination of enhanced recovery after surgery and prehabilitation pathways improves perioperative outcomes and costs for robotic radical prostatectomy. Cancer 2020;126(18):4148-4155 [FREE Full text] [doi: 10.1002/cncr.33061] [Medline: 32639601]
9. Minnella EM, Carli F, Kassouf W. Role of prehabilitation following major uro-oncologic surgery: a narrative review. World J Urol 2022;40(6):1289-1298. [doi: 10.1007/s00345-020-03505-4] [Medline: 33128596]
10. Paterson C, Primeau C, Pullar I, Nabi G. Development of a prehabilitation multimodal supportive care interventions for men and their partners before radical prostatectomy for localized prostate cancer. Cancer Nurs 2019;42(4):E47-E53. [doi: 10.1097/NCC.0000000000000618] [Medline: 29933304]
11. Reynolds BR, Bulsara C, Zeps N, Codde J, Lawrentschuk N, Bolton D, et al. Exploring pathways towards improving patient experience of robot-assisted radical prostatectomy (RARP): assessing patient satisfaction and attitudes. BJU Int 2018;121(Suppl 3):33-39. [doi: 10.1111/bju.14226] [Medline: 29603580]
12. Ploussard G, Loison G, Almeras C, Gautier J, Cazali P, Tollon C, et al. One-day prehabilitation program before robotic radical prostatectomy in daily practice: routine feasibility and benefits for patients and hospitals. Eur Urol Open Sci 2020;21:14-16 [FREE Full text] [doi: 10.1016/j.euros.2020.06.009] [Medline: 34337463]
13. Ploussard G, Almeras C, Beauval J, Gautier J, Loison G, Salin A, et al. Same-day discharge surgery for robot-assisted radical prostatectomy in the era of ERAS and prehabilitation pathways: a contemporary, comparative, feasibility study. World J Urol 2022;40(6):1359-1365. [doi: 10.1007/s00345-020-03119-w] [Medline: 32065277]

XSL•FO
RenderX

14. Rahota RG, Salin A, Gautier J, Almeras C, Garnault V, Tollon C, et al. A prehabilitation programme implemented before robot-assisted radical prostatectomy improves peri-operative outcomes and continence recovery. BJU Int 2022;130(3):357-363. [doi: 10.1111/bju.15666] [Medline: 34854212]
15. den Bakker CM, Schaafsma FG, Consten ECJ, Schraffordt Koops SE, van der Meij E, van de Ven PM, et al. Personalised electronic health programme for recovery after major abdominal surgery: a multicentre, single-blind, randomised, placebo-controlled trial. Lancet Digit Health 2023;5(8):e485-e494 [FREE Full text] [doi: 10.1016/S2589-7500(23)00084-5] [Medline: 37419843]
16. Dedding C, van Doorn R, Winkler L, Reis R. How will e-health affect patient participation in the clinic? A review of e-health studies and the current evidence for changes in the relationship between medical professionals and patients. Soc Sci Med 2011;72(1):49-53. [doi: 10.1016/j.socscimed.2010.10.017] [Medline: 21129832]
17. Baboudjian M, Abou-Zahr R, Buhas B, Touzani A, Beauval J, Ploussard G. The BETTY score to predict perioperative outcomes in surgical patients. Cancers (Basel) 2023;15(11):3050 [FREE Full text] [doi: 10.3390/cancers15113050] [Medline: 37297012]
18. Clavien P, Barkun J, de Oliveira ML, Vauthey JN, Dindo D, Schulick RD, et al. The Clavien-Dindo classification of surgical complications: five-year experience. Ann Surg 2009;250(2):187-196. [doi: 10.1097/SLA.0b013e3181b13ca2] [Medline: 19638912]
19. Ploussard G, Kesch C, Calleris G, Buhas B, Salin A, Almeras C, et al. From an on-site program to a mobile app for prehabilitation and rehabilitation for robotic radical prostatectomy: lessons learned from 5 years of experience, the COVID-19 outbreak, and comparison with nationwide data. Eur Urol Oncol 2024;7(2):297-299 [FREE Full text] [doi: 10.1016/j.euo.2023.10.007] [Medline: 37865569]
20. Abaza R, Martinez O, Ferroni MC, Bsatee A, Gerhard RS. Same day discharge after robotic radical prostatectomy. J Urol 2019;202(5):959-963. [doi: 10.1097/JU.0000000000000353] [Medline: 31112102]
21. Ploussard G, Grabia A, Barret E, Beauval J, Brureau L, Créhange G, Cancerology Committee of Association Française d'Urologie. Same-day-discharge robot-assisted radical prostatectomy: an annual countrywide analysis. Eur Urol Open Sci 2021;36:23-25 [FREE Full text] [doi: 10.1016/j.euros.2021.12.002] [Medline: 35005649]
22. Tsimopoulou I, Pasquali S, Howard R, Desai A, Gourevitch D, Tolosa I, et al. Psychological prehabilitation before cancer surgery: a systematic review. Ann Surg Oncol 2015;22(13):4117-4123. [doi: 10.1245/s10434-015-4550-z] [Medline: 25869228]
23. Jayadevappa R, Chhatre S, Gallo JJ, Wittink M, Morales KH, Lee DI, et al. Patient-centered preference assessment to improve satisfaction with care among patients with localized prostate cancer: a randomized controlled trial. J Clin Oncol 2019;37(12):964-973. [doi: 10.1200/JCO.18.01091] [Medline: 30860943]
24. van der Velde M, Valkenet K, Geleijn E, Kruisselbrink M, Marsman M, Janssen LM, et al. Usability and preliminary effectiveness of a preoperative mHealth app for people undergoing major surgery: pilot randomized controlled trial. JMIR Mhealth Uhealth 2021;9(1):e23402 [FREE Full text] [doi: 10.2196/23402] [Medline: 33410758]
25. Mata J, Pecorelli N, Kaneva P, Moldoveanu D, Gosselin-Tardiff A, Alhashemi M, et al. A mobile device application (app) to improve adherence to an enhanced recovery program for colorectal surgery: a randomized controlled trial. Surg Endosc 2020;34(2):742-751. [doi: 10.1007/s00464-019-06823-w] [Medline: 31087175]

Abbreviations

IRB: institutional review board
NSS: nerve-sparing surgery
PreHab: prehabilitation
RP: radical prostatectomy

Edited by N Cahill, T Leung; submitted 12.12.23; peer-reviewed by G Ploussard, Y Chu, D Dagate, M Baboudjian, D Charles; comments to author 18.04.24; revised version received 13.05.24; accepted 15.08.24; published 13.12.24.

Please cite as:
Martini A, Kesch C, Touzani A, Calleris G, Buhas B, Abou-Zahr R, Rahota RG, Pradère B, Tollon C, Beauval JB, Ploussard G Personalized Mobile App–Based Program for Preparation and Recovery After Radical Prostatectomy: Initial Evidence for Improved Outcomes From a Prospective Nonrandomized Study
J Med Internet Res 2024;26:e55429
URL: https://www.jmir.org/2024/1/e55429
doi:10.2196/55429
PMID:

XSL•FO
RenderX

©Alberto Martini, Claudia Kesch, Alae Touzani, Giorgio Calleris, Bogdan Buhas, Rawad Abou-Zahr, Razvan-George Rahota, Benjamin Pradère, Christophe Tollon, Jean-Baptiste Beauval, Guillaume Ploussard. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of a Smartphone-Based Breastfeeding Coparenting Intervention Program on Breastfeeding-Related Outcomes in Couples During First Pregnancy: Randomized Controlled Trial

Yi-Yan Huang[1*], RNC; Rong Wang[2*], RNC; Wei-Peng Huang[3], MD; Tian Wu[1,4], RNC; Shi-Yun Wang[1], RNC; Sharon R. Redding[5], RNC; Yan-Qiong Ouyang[1], MD

[1]School of Nursing, Wuhan University, Wuhan, China

[2]Department of Nursing, Renmin Hospital of Wuhan University, Wuhan, China

[3]Department of Critical Care Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

[4]Tongji Medical College, Huazhong University of Science and Technology, Wuhan Central Hospital, Department of Obstetrics, Wuhan, China

[5]Global Health of Project HOPE, Washington, MD, United States

[*]these authors contributed equally

Corresponding Author:
Yan-Qiong Ouyang, MD
School of Nursing
Wuhan University
115 Donghu Rd
Wuchang District
Wuhan, 430072
China
Phone: 86 27 68759710
Email: ouyangyq@whu.edu.cn

Abstract

Background: A low breastfeeding rate causes an increased health care burden and negative health outcomes for individuals and society. Coparenting is an essential tactic for encouraging breastfeeding when raising a child. The efficacy of the coparenting interventions in enhancing breastfeeding-related outcomes is controversial.

Objective: This study aimed to examine the effects of coparenting interventions on exclusive breastfeeding rates, exclusive breastfeeding duration, breastfeeding knowledge, parenting sense of competence, coparenting relationships, depressive symptoms in new couples at 1 and 6 months post partum, and the BMI of infants 42 days post partum.

Methods: This was a randomized, single-blinded controlled clinical trial. Eligible couples in late pregnancy in a hospital in central China were randomly assigned to 2 groups. While couples in the control group received general care, couples in the intervention group had access to parenting classes, a fathers' support group, and individual counseling. Data were collected at baseline (T0), 1 month post partum (T1), and 6 months post partum (T2). Data on exclusive breastfeeding rate and exclusive breastfeeding duration were analyzed using the chi-square, Fisher exact, or Mann-Whitney *U* tests; coparenting relationships and the infant's BMI were analyzed using an independent samples *t* test; and breastfeeding knowledge, parenting sense of competence, and depressive symptoms were analyzed using a generalized estimation equation.

Results: A total of 96 couples were recruited, and 79 couples completed the study. The intervention group exhibited significantly higher exclusive breastfeeding rates at T1 (90% vs 65%, *P*=.02) and T2 (43.6% vs 22.5%, *P*=.02), compared with the control group. Exclusive breastfeeding duration was extended in the intervention group than in the control group at T1 (30, range 30-30 days vs 30, range 26.5-30 days; *P*=.01) and T2 (108, range 60-180 days vs 89, range 28-149.3 days; *P*<.05). The intervention group exhibited greater improvements in maternal breastfeeding knowledge (β=.07, 95% CI 0.006-0.13; *P*=.03) and maternal parenting sense of competence (β=5.49, 95% CI 2.09-8.87; *P*<.01) at T1, enhanced coparenting relationships at T1 (*P*<.001) and T2 (*P*=.02), paternal breastfeeding knowledge at T2 (β=.25, 95% CI 0.15-0.35, *P*<.001), paternal parenting sense of competence at T1 (β=5.35, 95% CI 2.23-8.47, *P*<.01), and reduced paternal depressive symptoms at T2 (β=.25, 95% CI 0.15-0.35, *P*<.001), and there was a rise in infants' BMI at 42 days post partum (β=.33, 95% CI 0.01-0.64, *P*=.04).

XSL•FO
RenderX

Conclusions: An evidence-based breastfeeding coparenting intervention is effective in improving exclusive breastfeeding rate, prolonging exclusive breastfeeding duration within the initial 6 months post partum, enhancing parental breastfeeding knowledge, levels of parenting sense of competence and coparenting relationship, infant's BMI, and reducing paternal depressive symptoms.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300069648; https://tinyurl.com/2p8st2p8

(J Med Internet Res 2024;26:e51566) doi:10.2196/51566

KEYWORDS

breastfeeding; co-parenting; randomized controlled; child; efficacy; depressive symptoms

Introduction

Breastfeeding is a crucial public health issue that has shown health benefits for the mother and infant, providing optimal nutrition for infant growth, promoting immune substances that prevent respiratory infections, diarrhea, and otitis media [1,2]. Benefits to mothers include increased lactogen and oxytocin secretion, reduced risk of maternal diseases, and promoted emotional bonding with infants [3]. Exclusive breastfeeding (EBF) for infants up to 6 months of age, followed by continued breastfeeding for 2 years is recommended by the World Health Organization [4].

However, the current global breastfeeding landscape is bleak, with only 40% of infants under 6 months are exclusively breastfed across 194 nations. In China, the EBF rate stands at 34.1%, and in the United States, it is even lower at 25.8%, falling short of the World Health Organization's target of 50% [5,6]. Low breastfeeding rates lead to increased health care costs and negative health outcomes. The cost of not breastfeeding in the United States was US $13 billion per year in direct medical costs and US $18 billion per year in indirect costs [7]. By enhancing breastfeeding practice, it is possible to prevent the death of 820,000 children each year, of which 87% are infants under 6 months of age [8].

The low breastfeeding rate is attributed to a lack of paternal involvement in childcare [9]. Cultural norms and traditional gender roles frequently discourage fathers from actively participating in breastfeeding. It is crucial to promote, support, and protect breastfeeding worldwide, as it embodies society's shared responsibility [10]. There is an urgent demand for improved breastfeeding intervention programs that address the needs of both parents.

Bich et al [11] discovered that offering breastfeeding education and counseling services to fathers in the intervention group resulted in notably higher EBF rates at 4 and 6 months compared with the control group. Similarly, Su and Ouyang [12] illustrated that coparenting, accompanied by guidance on emotional and practical support, led to significantly increased rates of exclusive breastfeeding at 4 and 6 months, along with a reduction in infant formula usage. However, contrasting findings from other studies suggest that such interventions do not necessarily lead to improved EBF rates [13-15]. Therefore, more vigorously designed studies are required to fully comprehend the possible benefits and drawbacks of incorporating fathers into coparenting methods targeted at encouraging effective breastfeeding practices.

This randomized controlled study aimed to investigate the effects of a breastfeeding coparenting intervention program to breastfeeding-related outcomes in couples during their first pregnancy, including breastfeeding rate, breastfeeding duration, parental breastfeeding knowledge, parental parenting sense of competence, coparenting relationship, parental depressive symptoms, and infant BMI.

Methods

Study Design and Participants

This study was a randomized, single-blinded controlled clinical trial, including measurements of pretest at baseline (T0), posttest after intervention (1 month post partum) (T1), and at 6 months post partum (T2).

The inclusion criteria consisted of couples (1) in the first pregnancy ≥ 28 weeks, (2) planning to breastfeed their infant, and (3) able to complete the related questionnaire independently. Exclusion criteria were couples or one of the couples (1) having or had participated in previous breastfeeding-related research, (2) who are health care workers, and (3) unable to access the internet.

Randomization and Blinding

Participants were couples recruited from an outpatient obstetrics department of a large hospital in central China using convenience sampling. Researchers provided a brief introduction of the study to the couples, who were then directed to a WeChat official account (a popular social platform). Informed consent was obtained, and participants were then given access to the account.

Eligible couples were randomly allocated into either the intervention group or the control group. A randomization sequence was generated by a random service. Participants couples in either the intervention or control group were strangers, therefore they were blinded to reduce contamination bias. Researchers who compiled and analyzed the research data didn't take part in the intervention process. The outcome evaluators were blinded as they did not know the group assignments.

Intervention

Based on the Breastfeeding Co-parenting Framework [16], a breastfeeding coparenting intervention program was developed through a literature review [17] and Delphi method. A total of 14 studies, conducted in 9 different countries between January 1995 and February 2022, were included. The coparenting breastfeeding interventions, including start and stop dates,

XSL•FO
RenderX

duration, program components, and content were outlined. In total, 7 experts were invited to participate as consultants from April 21 to June 23, 2022. After 2 rounds of Delphi consultation involving 6 experts, the intervention program had good validity and reliability with a coefficient judgment basis of 0.93, a familiarity coefficient of 0.87, an authority coefficient of 0.90 and Kendall W of 0.62. The intervention program included a 7-session parenting course, a father's support group and individual counseling from late pregnancy to 6 months post partum (Multimedia Appendix 1).

Couples in the control group received generally available care. Before childbirth, regular prenatal examinations were conducted in the hospital, during which time couples were guided on maternal nutrition, weight management, and breastfeeding. After delivery, regular home visits were performed by the community nurses, who assessed the growth and development of the baby, as well as the uterus recovery status, breastfeeding condition, and pelvic floor function and provided health education. Researchers contacted these female participants by phone each month.

In addition to the interventions administered to the control group, participants in the intervention group were requested to add the researchers to their personal WeChat accounts, join the WeChat intervention group, and engage with a WeChat public account named "Guardian of Maternal and Infant Health" (Multimedia Appendix 2). This account was developed by the research group.

The classes were provided from late pregnancy to 4 weeks post partum, covering a total of 7 topics with a new topic. Between 28 and 37 weeks of gestation, educational sessions focusing on breastfeeding significance and techniques were provided to the intervention group participants by WeChat video on the WeChat account. From 37 weeks of gestation until delivery, participants received information on breastfeeding in specific circumstances through textual and video materials delivered by the research team, supplemented by community nurses. In addition, on the 14th day post partum, researchers and community nurses engaged in face-to-face discussions with couples, offering guidance on infant growth characteristics and distributing informational leaflets. At 4 weeks post partum, screenings for postpartum depression were conducted during personal meetings with couples by researchers and community nurses, along with the provision of resources on postnatal mental well-being, including textual, visual, and video materials. On the 42nd day post partum, intervention participants and their newborns underwent health checkups conducted by community nurses. Researchers sent reminders to participants about these classes by WeChat or phone calls.

Besides, male participants in the intervention group were invited to join the WeChat father's support group, providing them with a platform to freely express their feelings and concerns, which was not available to male participants in the control group. Individual counseling was also provided to the intervention group couples through direct meetings with content experts, WeChat calls or phone calls, when needed. In addition, couples could seek help through individual counseling (contact experts by WeChat or telephone call when needed). The intervention was conducted by the research team consisting of an international board-certified lactation consultant, a national second-level psychological counselor, and a postgraduate student major in maternal and child health.

Outcomes

Primary Outcomes

The primary outcomes were EBF rate and duration. Exclusive breastfeeding is defined as feeding infants with breast milk only from birth, with no supplementation of any other fluids or food, except vitamin, mineral drops, or medication [18]. The EBF duration is the total time of EBF from initiation to cessation. Data on EBF rate and EBF duration were obtained by asking the question: "What is your current feeding mode (exclusive breastfeeding, mixed feeding, or exclusive formula feeding)?" and "How long have you been exclusively breastfeeding your infant (in days)?"

Secondary Outcomes

Breastfeeding Knowledge Questionnaire

Designed by Su and Ouyang [12], this 18-item questionnaire asks about the benefits and skills of breastfeeding. One point is awarded for a correct answer, for a total possible score of 18, with a higher score indicating more enhanced breastfeeding knowledge. The questionnaire had a high level of internal reliability (Cronbach α=0.82) [19].

Parenting Sense of Competence Scale

Originally developed by Johnston to assess parenting satisfaction and efficacy [20], Ngai et al [21] translated the scale into Chinese and validated it with good internal consistency (Cronbach α=0.85) and test-retest reliability (correlation coefficient=0.87). The scale consists of 17 items, each of which adopts a 6-point Likert scale, from "strongly disagree" to "strongly agree." Items 2, 3, 4, 5, 8, 9, 12, 14, and 16 are inversely scored for a total possible score of the scale from 17 to 102 points. The higher the score, the higher the level of perceived competence of the parents.

Brief Co-Parenting Relationship Scale

Developed by Feinberg et al [22] in 2012, the Brief Co-parenting Relationship Scale measures the quality of coparenting relationships between parents when raising a child together. It is composed of 14 items and 7 subscales. The Chinese version was developed by Min Wu and Zhao [23], including 5 dimensions with 14 items. All items are scored on a 7-point Likert scale, with 3 items being reverse scored, for a total score of 0 to 84. A higher score indicates a better coparenting relationship. The split-half reliability coefficient for the total score was 0.516, and Cronbach α coefficient for the total score was 0.613.

Edinburgh Postnatal Depression Scale

Compiled by Cox and translated into Chinese by Lee et al [24], this scale measures postpartum depressive symptoms. There are 10 items, and each item is divided into 4 levels from never (0 points) to always (3 points) for a possible total score of 30 points. If the total score is greater than or equal to 9, the mother has a high risk of experiencing depression and should be referred

XSL•FO
RenderX

to a health care provider for further evaluation. The Chinese version of the scale had good validity. The split-half reliability of the Edinburgh Postnatal Depression Scale (EPDS) was 0.74, Cronbach α was 0.78 and test-retest reliability was 0.90 [25].

BMI of Infant at 42 Days

BMI was calculated by weight and body length data extracted from the newborn's community checkup.

Both female and male participants were required to complete the questionnaire at 3 time points. At T0, questionnaires were collected by trained researchers through a paper survey face to face in an outpatient of obstetrics department. A researcher-designed questionnaire was used for data collection, including sociodemographic characteristics of the participants, Breastfeeding Knowledge Questionnaire (BKQ), and EPDS. At T1, data on breastfeeding rates, breastfeeding duration, BKQ, Parenting Sense of Competence Scale (PSOC), EPDS, and Brief Co-Parenting Relationship Scale were collected in community visits through a paper survey. At T2, in addition to the data collected in T1, weight and body length data of the infants at 42 days post partum were obtained to calculate the BMI. At this time, data were collected through Wenjuanxing (an easy-to-use survey platform).

Sample Size

Sample size was calculated based on the results of Bich and Cuong [26]. According to the calculation formula of PASS15.0, we calculated the sample size was 72 couples (power=0.80, α=.10). With a 20% attrition rate, the sample size of 90 participant couples was more adequate (total sample size). As a result, 90 participant couples were involved (45 couples in each group).

Statistical Analysis

Statistical analysis was undertaken using SPSS (version 24.0; IBM Corp). The per-protocol approach was used to analyze data. Kolmogorov-Smirnov test was used for judgment of data normality. Continuous data are presented as median (IQR) or means (SD); classified data are presented as n (%). The Mann-Whitney *U* test, student *t* test, chi-square test, and the Fisher exact tests were used for data comparison at baseline. Data on EBF rates and EBF duration were analyzed using the chi-square test, Fisher exact tests, or Mann-Whitney *U* test. Data on coparenting relationship were analyzed by independent samples *t* test. To examine the influence of the intervention over time, data on breastfeeding knowledge, parenting sense of competence, coparenting relationship, and depressive symptoms were analyzed by a generalized estimation equation. The significance level is set at .05 (2-tailed).

Ethical Considerations

The study was approved by the Ethics Committee of the Department of Medicine, Wuhan University (IRB2022015) and was registered at ChiCTR.org.cn (ChiCTR2300069648). We followed the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Multimedia Appendix 3) [27].

Results

Participant Characteristics

Sociodemographic characteristics of participants are presented in Table 1. From August 2022 to March 2023, 96 couples were recruited while 79 couples completed the research (39 couples in the intervention group, 40 couples in the control group). The attrition rate was 21.5% (17/79). A total of 17 participants dropped out because of discontinued courses or refusion to complete the survey (Figure 1).

XSL•FO
RenderX

Table 1. Baseline characteristics of participant couples (N=79).

Variables	Total (N=79)	Intervention group (n=39)	Control group (n=40)	*P* value
Maternal				
Age (years), median (IQR)	30 (28-32)	30 (29-33)	30 (28-32)	.10
Urban residence, n (%)	76 (96.2)	38 (97.4)	38 (95)	>.99
Educational level, n (%)				.24
High school or below	11 (13.9)	3 (7.7)	8 (20)	__[a]
Bachelor's degree	60 (76)	31 (79.5)	29 (72.5)	__[a]
Postgraduate or above	8 (10.1)	5 (12.8)	3 (7.5)	__[a]
Employment, n (%)				.71
Worker	1 (1.3)	1 (2.6)	0 (0)	__[a]
Administrator	33 (41.8)	17 (43.6)	16 (40)	__[a]
Teacher	8 (10.1)	5 (12.8)	3 (7.5)	__[a]
Private enterprise	25 (31.6)	12 (30.8)	13 (32.5)	__[a]
Free work	12 (15.2)	4 (10.3)	8 (20)	__[a]
Average monthly household income (Yuan; 1 Yuan=US $0.14), n (%)				.20
≤3000	1 (1.3)	0 (0)	1 (2.5)	__[a]
3000-5000	12 (15.2)	3 (7.7)	9 (22.5)	__[a]
5001-10,000	34 (43)	18 (46.1)	16 (40)	__[a]
≥10,000	32 (40.5)	39 (49.4)	14 (35)	__[a]
Maternal secondary outcomes at baseline, points				
BKQ[b]	0.70 (0.12)	0.71 (0.13)	0.69 (0.12)	.37
PSOC[c]	61.82 (5.70)	61.79 (5.84)	61.85 (5.63)	.97
EPDS[d]	8.0 (4-12)	8.0 (4-12)	7.5 (4-11)	.98
Paternal				
Age (years), median (IQR)	32 (30-35)	32 (30-35)	32 (30-34)	.64
Educational level, n (%)				.78
High school or below	6 (7.6)	2 (5.1)	4 (10)	__[a]
Bachelor's degree	63 (79.8)	32 (82.1)	31 (77.5)	__[a]
Postgraduate or above	10 (12.7)	5 (12.8)	5 (12.5)	__[a]
Employment, n (%)				.68[e]
Worker	22 (27.9)	9 (23.1)	13 (32.5)	__[a]
Administrator	28 (35.4)	16 (41.0)	12 (30)	__[a]
Teacher	12 (15.2)	5 (12.8)	7 (17.5)	__[a]
Private enterprise	16 (20.3)	13 (33.3)	3 (7.5)	__[a]
Free work	1 (1.3)	1 (2.6)	0 (0)	__[a]
Paternal secondary outcomes at baseline, points				
BKQ[b]	0.57 (0.13)	0.58 (0.14)	0.57 (0.12)	.77
PSOC[c]	60.27 (5.83)	60.72 (6.54)	59.83 (5.09)	.5

XSL•FO
RenderX

Variables	Total (N=79)	Intervention group (n=39)	Control group (n=40)	*P* value
EPDS[d]	7.67 (3.28)	7.23 (3.89)	8.10 (2.53)	.24
Infant				
Gestational age (weeks), median (IQR)	39 (38-39)	38.5 (38-39)	39 (38-39)	.53
Delivery mode, vaginal, n (%)	38 (48.1)	21 (53.9)	17 (42.5)	.37
Infants' gender, female, n (%)	43 (54.4)	23 (59)	20 (50)	.5
Infants' birth weight (kg), mean (SD)	3.16 (0.26)	3.18 (0.29)	3.15 (0.24)	.69
Infants' birth length (cm), median (IQR)	50 (49-50)	50 (49-51)	50 (49-50)	.21
Infants' BMI (kg/m^2), mean (SD)	12.89 (1.3)	12.78 (1.29)	13.00 (1.32)	.46

[a]—: not available.
[b]BKQ: Breastfeeding Knowledge Questionnaire.
[c]PSOC: Parenting Sense of Competence Scale.
[d]EPDS: Edinburgh Postnatal Depression Scale.
[e]Fisher exact test.

Figure 1. Flow diagram of participants' inclusion process. T1: 1 month postpartum; T2: 6 months postpartum.

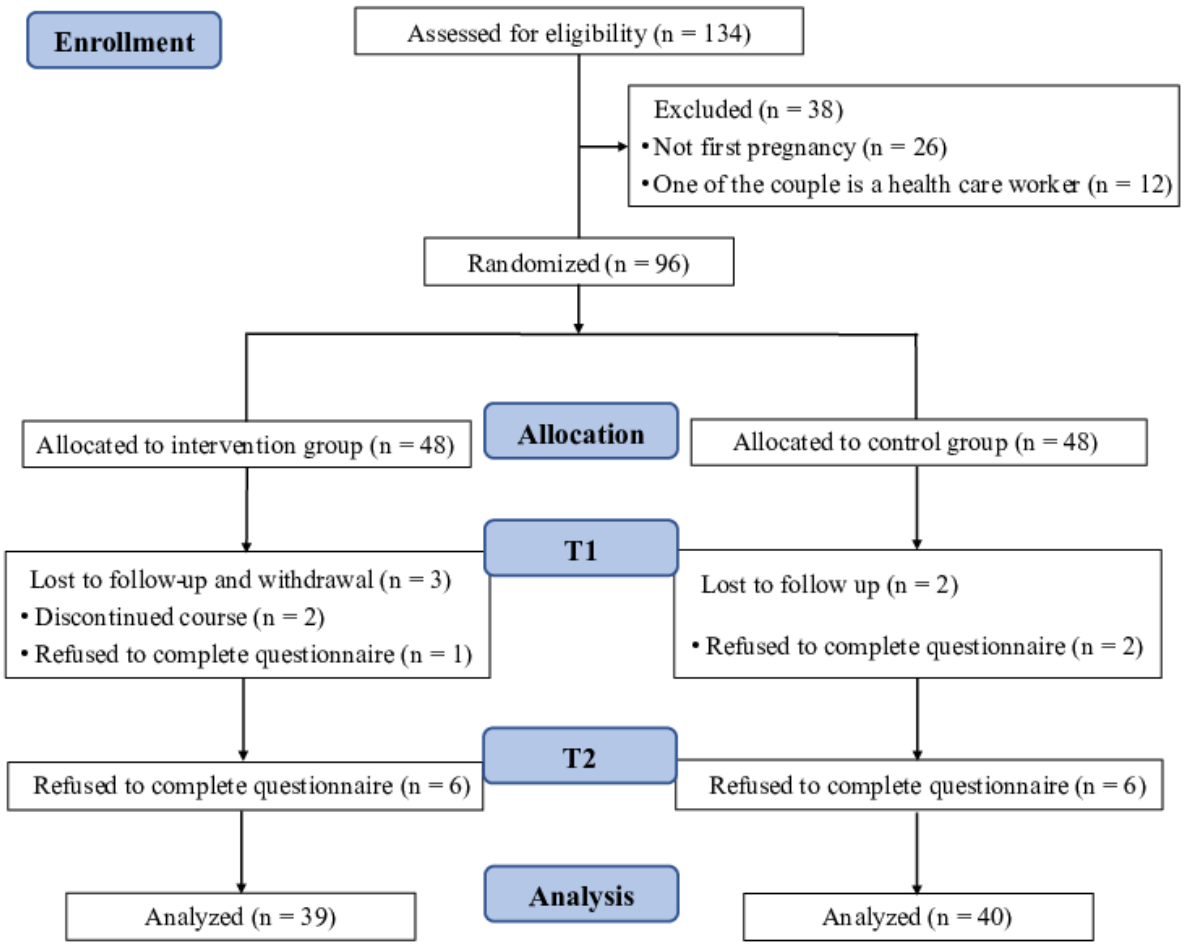

The median maternal age was 30 (IQR 28-32) years, and the median paternal age was 32 (IQR 30-35) years. The median gestational age of infants was 39 (IQR 38-39) weeks with 4 infants being premature. A total of 48.1% (38/79) of the infants were delivered vaginally, and 45.6% (36/79) of the infants were male. Among the maternal, paternal and infant data, no statistical differences existed regarding the baseline data (*P*>.05; Table 1).

XSL•FO
RenderX

Primary Outcomes: Exclusive Breastfeeding Rate and Exclusive Breastfeeding Duration

Details of the primary outcomes are presented in Table 2. At T1, the intervention group exhibited a significantly higher EBF rate compared with the control group (89.7% vs 65%, *P*=.02). At T2, the EBF rate remained higher in the intervention group (43.6% vs 22.5%, *P*=.02). The intervention group had an extended EBF duration than the control group at T1 (*P*=.01) and T2 (*P*<.05).

Table 2. Effects of breastfeeding coparenting intervention program on breastfeeding rates and breastfeeding duration (N=79).

Time and variables	Total	Control group (N=40)	Intervention group (N=39)	*P* value
T1[a] Feeding mode, n (%)				.02
EBF[b]	61 (77.2)	26 (65)	35 (89.8)	__[c]
Mixed feeding	17 (21.5)	13 (32.5)	4 (10.2)	__[c]
Artificial feeding	1 (1.3)	1 (2.5)	0 (0)	__[c]
EBF duration (days)	30 (28-30)	30 (26.5-30)	30 (30-30)	.01
T2[d] Feeding mode, n (%)				.02
EBF	26 (32.9)	9 (22.5)	17 (43.6)	__[c]
Mixed feeding	39 (49.4)	20 (50)	19 (48.7)	__[c]
Artificial feeding	14 (17.7)	11 (27.5)	3 (7.7)	__[c]
EBF duration (days)	102 (32-180)	89 (28-149.3)	108 (60-180)	.045

[a]T1: 1 month post partum.

[b]EBF: exclusive breastfeeding.

[c]—: not available.

[d]T2: 6 months post partum.

Maternal Secondary Outcomes

Secondary outcomes of couples are displayed in Table 3. The generalized estimation equation model revealed a significant interaction term concerning maternal BKQ correct rates (group×time) at T1 (β=0.07, 95% CI 0.006-0.13; *P*=.03), indicating that the intervention effectively improved maternal breastfeeding knowledge at T1. In addition, the model showed a significant interaction term concerning maternal PSOC scores (group×time) at T1 (β=5.48, 95% CI 2.09-8.87; *P*<.01), demonstrating that the intervention effectively improved maternal parenting sense of competence at T1.

XSL•FO
RenderX

Table 3. Effects of breastfeeding co-parenting intervention program on secondary outcomes (N=79).

Parental outcomes	Control group	Intervention group	Group effect		Time effect		Group*time effect	
			β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Maternal								
BKQ[a] correct rates, mean (SD)	—[b]	—	0.03 (−0.03 to 0.08)	.03	—	—	—	—
T0[c]	0.69 (0.12)	0.71 (0.13)	—	—	—	—	—	—
T1[d]	0.68 (0.07)	0.77 (0.06)	—	—	0.01 (0.05 to 0.04)	.71	0.07 (0.006-0.13)	.03
T2[e]	0.78 (0.09)	0.81 (0.12)	—	—	0.05 (0.01-0.09)	.01	−0.05 (−0.02 to 0.11)	.18
PSOC[f], mean (SD)	—	—	0.18 (−2.30 to 2.64)	.89	—	—	—	—
T0[c]	61.85 (5.63)	61.79 (5.84)	—	—	—	—	—	—
T1[d]	60.73 (4.83)	66.72 (3.55)	—	—	−1.13 (−3.56 to 1.31)	.37	5.48 (2.09-8.87)	<.01
T2[e]	63.15 (4.53)	65.05 (6.44)	—	—	1.30 (−1.07 to 3.67)	.28	1.39 (−2.40 to 5.19)	.47
EPDS[g], mean (SD)	—	—	−0.40 (−2.85 to 2.06)	.75	—	—	—	—
T0[b]	8.35 (5.99)	8.18 (5.15)	—	—	—	—	—	—
T1[c]	5.65 (2.80)	2.67 (2.60)	—	—	−2.85 (−4.90 to −0.80)	<.01	−2.66 (−5.48 to 0.15)	.58
T2[d]	7.85 (3.76)	6.69 (4.08)	—	—	−0.63 (−2.78 to 1.53)	.57	−0.86 (−3.90 to 2.17)	.06
Paternal								
BKQ[a] correct rates, mean (SD)	—	—	0.03 (−0.05 to 0.10)	.50	—	—	—	—
T0[b]	0.54 (0.19)	0.58 (0.14)	—	—	—	—	—	—
T1[c]	0.67 (0.12)	0.76 (0.13)	—	—	0.13 (0.07-0.19)	<.001	0.08 (−0.003 to 0.17)	.06
T2[d]	0.63 (0.08)	0.80 (0.08)	—	—	−0.02 (−0.09 to 0.06)	.65	0.25 (0.15-0.35)	<.001
PSOC[f], mean (SD)	—	—	0.99 (−1.55 to 3.55)	.44	—	—	—	—
T0[b]	59.83 (5.09)	60.82 (6.51)	—	—	—	—	—	—
T1[c]	61.55 (5.09)	67.90 (4.04)	—	—	1.73 (−0.37 to 3.82)	.11	5.35 (2.23-8.47)	<.01
T2[d]	60.10 (6.39)	63.85 (4.38)	—	—	0.28 (−2.15 to 2.70)	.82	2.75 (−0.96 to 6.46)	.15
EPDS[g], mean (SD)	—	—	0.03 (−0.05 to 0.10)	.50	—	—	—	—
T0[b]	9.82 (3.86)	8.21 (5.82)	—	—	—	—	—	—
T1[c]	4.90 (2.38)	2.00 (0.95)	—	—	0.13 (0.07-0.19)	<.001	0.08 (−0.003 to 0.17)	.06
T2[d]	11.75 (2.95)	5.03 (2.99)	—	—	−0.02 (−0.09 to 0.06)	.65	0.25 (0.15-0.35)	<.001
Infant								

XSL•FO
RenderX

Parental outcomes	Control group	Intervention group	Group effect		Time effect		Group*time effect	
			β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Weight, mean (SD)	—	—	0.07 (−0.06 to 0.20)	.29	1.07 (0.10-1.14)	<.001	0.09 (−0.05 to 0.24)	.21
At birth	3.15 (0.24)	3.18 (0.29)	—	—	—	—	—	—
42 days post partum	4.17 (0.40)	4.29 (0.42)	—	—	—	—	—	—
Body length, median (IQR)	—	—	0.62 (−0.10 to 1.35)	.09	5.06 (4.64-5.49)	<.001	0.10 (−0.75 to 0.95)	.81
At birth	50 (49-50)	50 (49-51)	—	—	—	—	—	—
42 days post partum	54.25 (53-56)	55 (53-56)	—	—	—	—	—	—
BMI, mean (SD)	—	—	−0.22 (−0.78 to 0.35)	.46	—	—	—	—
At birth	14.20 (1.31)	12.78 (1.29)	—	—	—	—	—	—
42 days post partum	14.17 (1.33)	13.00 (1.32)	—	—	1.18 (1.01-1.34)	<.001	0.33 (0.01-0.64)	.04

[a]BKQ: Breastfeeding knowledge questionnaire.

[b]Not available.

[c]T0: at baseline.

[d]T1: 1 month post partum.

[e]T2: 6 months post partum.

[f]PSOC: Parenting Sense of Competence Scale.

[g]EPDS: Edinburgh Postnatal Depression Scale.

Since coparenting relationships were measured only twice, independent-sample *t* tests were conducted to compare the data. The intervention group exhibited superior coparenting relationships compared with the control group at both T1 (53.67, SD 5.80 vs 47.63, SD 6.21; *P*<.001) and T2 (53.46, SD 6.28 vs 49.70, SD 7.32; *P*=.02). Also, a significant reduction in depressive symptoms was observed at T1 (*P*< .01). To provide a clearer understanding of the study findings, Figure 2 illustrates the changes in secondary outcomes between the groups.

XSL•FO
RenderX

Figure 2. Comparison of the changes in secondary outcomes between groups. EPDS: Edinburgh Postnatal Depression Scale; PSOC: Parenting Sense of Competence Scale; T0: at baseline; T1: 1 month post partum; T2: 6 months post partum.

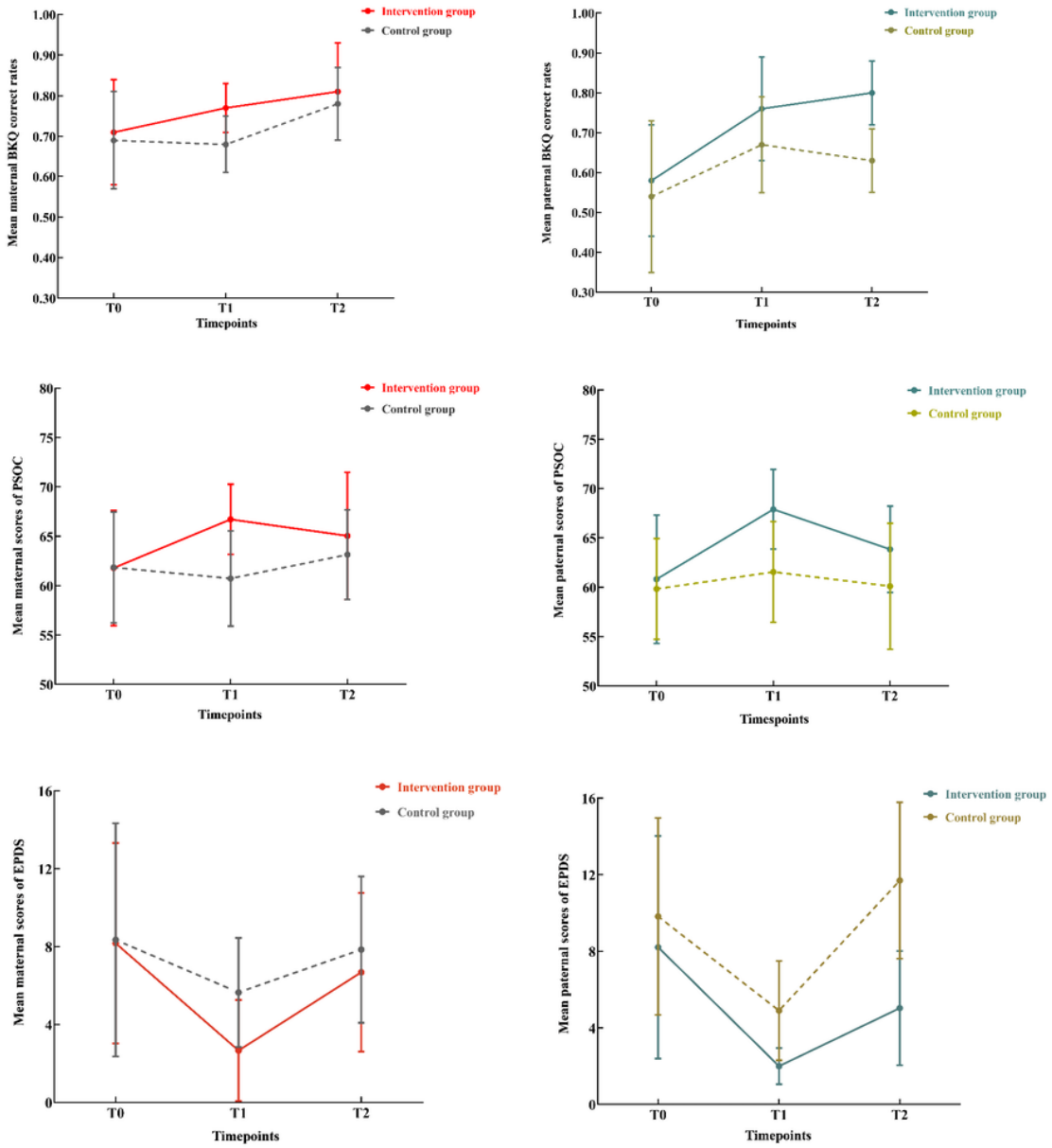

Paternal Secondary Outcomes

The generalized estimation equation model showed significant interaction term in relation to paternal BKQ correct rates (group×time) at T2 (β=0.25, 95% CI 0.15-0.35; *P*<.001), verifying the intervention program's efficacy in enhancing paternal breastfeeding knowledge at T2. Similarly, the model displayed significant interaction terms concerning paternal PSOC scores (group×time) at T1 (β=5.35, 95% CI 2.23-8.47; *P*<.01), affirming the intervention program's effectiveness in improving paternal parenting sense of competence at T1. In addition, there were significant interaction terms observed regarding paternal EPDS scores (group×time) at T2 (β=0.25, 95% CI 0.15-0.35; *P*<.001), validating the intervention program's success in reducing paternal depressive symptoms at T2. Furthermore, the model identified significant interaction terms regarding infants' BMI (group×time; β=0.33, 95% CI 0.01-0.64; *P*=.04), confirming the intervention program's positive impact on improving infants' BMI at 42 days post partum.

XSL•FO
RenderX

Discussion

Principal Results

The main finding of this study was that the breastfeeding coparenting intervention program could effectively improve EBF rates. This is consistent with the study Bich et al [11], who conducted an intervention that provided breastfeeding education materials, counseling services at commune health centers, and household visits to fathers, resulting in a significant increase in exclusive breastfeeding rates at 4 and 6 months. Active participation of fathers in breastfeeding education and support sessions can enhance their understanding of breastfeeding's importance, fostering a supportive home environment [28,29]. Fathers' emotional, practical, and physical support is crucial for successful breastfeeding, easing the mothers' burden and promoting coparenting collaboration [30].

Although higher than the 40% reported by Su and Ouyang [12], EBF rates fell short of the 55.56% reported by Abbass-Dick et al [14]. This discrepancy may be due to Canada's higher breastfeeding prevalence rate compared with China. Nevertheless, the rates still did not meet the World Health Organization's target of 50%. This study highlights the critical role of spousal support in promoting breastfeeding practices. Breastfeeding is a social responsibility that involves individuals, families, and society as a whole. In addition, workplace breastfeeding support and public mother-and-baby rooms are also essential factors in driving change [3,31]. Further work is needed to address these systemic factors. At the same time, the intervention program successfully extended EBF duration by providing comprehensive education and support, enabling participants to overcome breastfeeding-related challenges and maintain EBF for a long period, which is crucial for infant growth and development [3]. The findings highlight the importance of providing adequate support and resources to new parents to promote breastfeeding practices.

This study laid a firm foundation for enhancing breastfeeding knowledge throughout the first 6 months post partum, which is congruent with the findings of Bich and Cuong [26]. After the breastfeeding intervention program, maternal and paternal breastfeeding knowledge had improved significantly and experienced a slight decline over time, albeit the total improvement was maintained. Initially, paternal breastfeeding knowledge was substantially lower than maternal knowledge, but the gap was narrowed following the intervention, which encouraged both parents to participate and become more engaged in breastfeeding. By the end of the intervention, the knowledge of both mothers and fathers had improved significantly and tended to decline over time, although knowledge levels remained high.

The intervention effectively enhanced the levels of parenting sense of competence in both parents, which is similar to the findings of Chu et al [32], who reported that participants who received a mobile phone intervention were more likely to feel confident in adopting effective parenting skills and demonstrated a higher level of parenting sense of competence. Greater learned resourcefulness and social support were directly related to maternal role competence and pleasure [33]. High levels of learned resourcefulness enable parents to effectively deal with obstacles. Highly resourceful parents might use learned resourcefulness abilities like problem-solving tactics to reduce or eliminate troubling thoughts and sensations. Consequently, they are more effective, experience less emotional disturbance, and feel more capable and content when faced with the challenges of new parenthood.

The intervention had no significant effect on alleviating maternal depressive symptoms. The sample of parents reported low scores at baseline on the EPDS, providing a limited possibility for decreasing. In addition, it is also possible that the amount of information conveyed by the intervention was insufficient to elicit a significant reduction in depressive symptoms. Professional therapies, including interpersonal counseling and cognitive behavioral therapy provided by a trained psychotherapist, are necessary for effectively treating perinatal depression [34,35]. In this study, fathers in the control group were at risk of experiencing depressive symptoms at 6 months post partum, which was consistent with the findings of Cameron et al [36]. However, this intervention proved to be effective in alleviating paternal depression at T2. The most common help-seeking barrier reported by parents can refer to a lack of information when needed. The study approach provided content involving mental health and individual counseling, which offered some information support to parents.

This intervention had a notable impact on enhancing coparenting relationships, which differs from the findings of Abbass-Dick et al [13]. The enhancement observed in coparenting relationships may be related to the lessening of conflict. In this study, the researchers included components in which parents jointly developed breastfeeding objectives, accepted responsibility, provided feeding assistance, and developed strong communication and problem-solving skills, thereby minimizing parental conflicts.

Previous studies have examined the effects of breastfeeding coparenting interventions on perception of infants using a 5-point scale [37] and infantile development status using Developmental Milestones Checklist-II [38]. This is the first study to examine the impact of coparenting interventions on an infant's BMI. To collect data on infants of participants in this study during the follow-up period, the researchers referred to health records established in the community. Typically, mothers take their newborns to for wellness checkups at 42 days following delivery. However, data collected at only 42 days post partum are insufficient for determining long-term changes. Therefore, it is necessary to monitor the growth and development of newborns for longer periods.

In total, 17 couples dropped out, although strategies were adopted to reduce the number of participants who dropped out, including reminding participants to engage in the study and gathering feedback from participants through telephone calls. More effort is needed to successfully engage parents in the intervention programs, such as offering tangible rewards or benefits to participants to complete the intervention successfully.

The coparenting intervention program for breastfeeding in this study was comprehensive and scientifically informed, drawing on a literature review and expert consultation. The study used

XSL•FO
RenderX

both e-health technologies and face-to-face interactions to provide health information. As one of the few randomized controlled studies examining the impact of a breastfeeding coparenting intervention program on breastfeeding-related outcomes in couples having a first pregnancy, the results suggest that delivering health interventions using a video presentation and chatroom format may be an effective approach to improving outcomes.

Strengths and Limitations

Dedicated to providing a holistic approach that addressed various aspects of postpartum well-being for both mothers and fathers, this study used a comprehensive intervention approach that included spouses as participants. Study outcomes were assessed 6 months post partum using a range of indicators. The interventions were designed based on theoretical frameworks and implemented under the guidance of multidisciplinary experts.

Nevertheless, this study is subject to several limitations. First, the relatively short follow-up period of 6 months post partum may not fully capture the enduring effects of the interventions on breastfeeding practices. Extending the follow-up duration to 2 years or more in future investigations would provide a more comprehensive understanding of the maintained impact on breastfeeding-related outcomes over time. Second, the sample predominantly comprised participants from urban areas with access to extensive medical services, potentially limiting the generalizability of the findings. This demographic skew potentially constrains the generalizability of the findings to populations with differing socioeconomic backgrounds or geographic locations. To enhance external validity, future studies should strive to recruit a more diverse sample, including couples from rural or remote areas. Finally, while the study effectively evaluated outcome measures, it notably lacked an assessment of process evaluation aspects, such as parental satisfaction with the intervention strategies used. Integrating process evaluation components into future research initiatives would not only provide valuable insights into the acceptability and feasibility of intervention approaches but also facilitate iterative improvements and refinements to enhance program effectiveness.

Conclusions

This evidence-based breastfeeding coparenting intervention effectively enhanced EBF rates and prolonged breastfeeding duration at 1 and 6 months post partum. It also improved maternal and paternal breastfeeding knowledge, parenting sense of competence, coparenting relationship, alleviated paternal depression, and increased infants' BMI at 42 days post partum. Further research is required to evaluate the long-term effects of the intervention across diverse geographic locations and integrate both outcome and process evaluations for a comprehensive understanding of its effectiveness and implementation.

Acknowledgments

The authors disclosed that the research has been funded by the Health Bureau of Songzi City (project HBSSZSHLZX0003).

Authors' Contributions

Y-YH contributed to conceptualization, investigation, formal analysis, methodology, writing-original draft, and writing—review and editing. RW contributed to conceptualization, investigation, methodology, writing-original draft, and writing—review and editing. W-PH managed conceptualization, methodology, and writing-original draft. TW handled investigation and writing-review and editing. S-YW conducted formal analysis and writing—review and editing, SRR contributed to conceptualization and writing-review and editing. Y-QO managed conceptualization, project administration, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structure of the breastfeeding co-parenting intervention program.
[DOCX File , 16 KB - jmir_v26i1e51566_app1.docx]

Multimedia Appendix 2
Screenshots of the WeChat account.
[PDF File (Adobe PDF File), 1446 KB - jmir_v26i1e51566_app2.pdf]

Multimedia Appendix 3
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.
[PDF File (Adobe PDF File), 3394 KB - jmir_v26i1e51566_app3.pdf]

References

XSL•FO
RenderX

1. Arora A, Manohar N, Hayen A, Bhole S, Eastwood J, Levy S, et al. Determinants of breastfeeding initiation among mothers in Sydney, Australia: findings from a birth cohort study. Int Breastfeed J 2017;12(1):39 [FREE Full text] [doi: 10.1186/s13006-017-0130-0] [Medline: 28932255]
2. Ghozy S, Tran L, Naveed S, Quynh TTH, Helmy Zayan A, Waqas A, et al. Association of breastfeeding status with risk of autism spectrum disorder: a systematic review, dose-response analysis and meta-analysis. Asian J Psychiatr 2020;48:101916. [doi: 10.1016/j.ajp.2019.101916] [Medline: 31923810]
3. Rollins NC, Bhandari N, Hajeebhoy N, Horton S, Lutter CK, Martines JC, Lancet Breastfeeding Series Group. Why invest, and what it will take to improve breastfeeding practices? Lancet 2016;387(10017):491-504. [doi: 10.1016/S0140-6736(15)01044-2] [Medline: 26869576]
4. Global targets 2025.: World Health Organization; 2015. URL: https://apps.who.int/nutrition/global-target-2025/en/index.html [accessed 2024-11-05]
5. Global breastfeeding scorecard.: UNICEF; 2022. URL: https://www.globalbreastfeedingcollective.org/global-breastfeeding-scorecard [accessed 2022-04-24]
6. Babies and mothers worldwide failed by lack of investment in breastfeeding.: United Nations International Children's Emergency Fund; 2017. URL: https://tinyurl.com/5n7cm39c [accessed 2023-11-09]
7. Bartick MC, Stuebe AM, Schwarz EB, Luongo C, Reinhold AG, Foster EM. Cost analysis of maternal disease associated with suboptimal breastfeeding. Obstet Gynecol 2013;122(1):111-119. [doi: 10.1097/AOG.0b013e318297a047] [Medline: 23743465]
8. Victora CG, Bahl R, Barros AJD, França GVA, Horton S, Krasevec J, Lancet Breastfeeding Series Group. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. Lancet 2016;387(10017):475-490. [doi: 10.1016/S0140-6736(15)01024-7] [Medline: 26869575]
9. Bennett AE, Kearney JM. Factors associated with maternal wellbeing at four months post-partum in Ireland. Nutrients 2018;10(5):609 [FREE Full text] [doi: 10.3390/nu10050609] [Medline: 29757937]
10. Pérez-Escamilla R, Tomori C, Hernández-Cordero S, Baker P, Barros AJD, Bégin F, 2023 Lancet Breastfeeding Series Group. Breastfeeding: crucially important, but increasingly challenged in a market-driven world. Lancet 2023;401(10375):472-485 [FREE Full text] [doi: 10.1016/S0140-6736(22)01932-8] [Medline: 36764313]
11. Bich TH, Hoa DTP, Målqvist M. Fathers as supporters for improved exclusive breastfeeding in viet nam. Matern Child Health J 2014;18(6):1444-1453. [doi: 10.1007/s10995-013-1384-9] [Medline: 24162509]
12. Su M, Ouyang YQ. Father's role in breastfeeding promotion: lessons from a quasi-experimental trial in China. Breastfeed Med 2016;11:144-149. [doi: 10.1089/bfm.2015.0144] [Medline: 26836960]
13. Abbass-Dick J, Stern SB, Nelson LE, Watson W, Dennis CL. Coparenting breastfeeding support and exclusive breastfeeding: a randomized controlled trial. Pediatrics 2015;135(1):102-110. [doi: 10.1542/peds.2014-1416] [Medline: 25452653]
14. Abbass-Dick J, Sun W, Newport A, Xie F, Godfrey D, Goodman WM. The comparison of access to an eHealth resource to current practice on mother and co-parent teamwork and breastfeeding rates: a randomized controlled trial. Midwifery 2020;90:102812. [doi: 10.1016/j.midw.2020.102812] [Medline: 32739716]
15. Scott JA, Burns SK, Hauck YL, Giglia RC, Jorgensen AM, White BK, et al. Impact of a face-to-face versus smartphone app versus combined breastfeeding intervention targeting fathers: randomized controlled trial. JMIR Pediatr Parent 2021;4(2):e24579 [FREE Full text] [doi: 10.2196/24579] [Medline: 33843604]
16. Abbass-Dick J, Dennis CL. Breast-feeding coparenting framework: a new framework to improve breast-feeding duration and exclusivity. Fam Community Health 2017;40(1):28-31. [doi: 10.1097/FCH.0000000000000137] [Medline: 27870751]
17. Zhao ZH, Huang YY, Qiao J, Huang WP, Redding SR, Wang R, et al. Co-parenting impact on breastfeeding: systematic review and meta-analysis. Breastfeed Med 2023;18(6):431-448. [doi: 10.1089/bfm.2022.0237] [Medline: 37285199]
18. Global nutrition targets 2025: breastfeeding policy brief.: World Health Organization; 2014. URL: https://www.who.int/publications/i/item/WHO-NMH-NHD-14.7 [accessed 2023-11-09]
19. Ouyang YQ, Xu YX, Zhang Q. Survey on breastfeeding among Chinese female physicians and nurses. Nurs Health Sci 2012;14(3):298-303. [doi: 10.1111/j.1442-2018.2012.00699.x] [Medline: 22827770]
20. Johnston C, Mash EJ. A measure of parenting satisfaction and efficacy. Journal of Clinical Child Psychology 1989;18(2):167-175. [doi: 10.1207/s15374424jccp1802_8] [Medline: 5637250]
21. Ngai FW, Wai-Chi Chan S, Holroyd E. Translation and validation of a chinese version of the parenting sense of competence scale in chinese mothers. Nurs Res 2007;56(5):348-354. [doi: 10.1097/01.NNR.0000289499.99542.94] [Medline: 17846556]
22. Feinberg ME, Brown LD, Kan ML. A multi-domain self-report measure of coparenting. Parent Sci Pract 2012;12(1):1-21. [doi: 10.1080/15295192.2012.638870] [Medline: 23166477]
23. Min Wu GJL J, Zhao H. Reliability and validity of Chinese version of Brief Coparenting Relationship Scale. Chinese Journal of Woman and Child Health Research 2017;28(4):1673-5293.
24. Lee DT, Yip SK, Chiu HF, Leung TY, Chan KP, Chau IO, et al. Detecting postnatal depression in Chinese women. Validation of the Chinese version of the edinburgh postnatal depression scale. Br J Psychiatry 1998;172:433-437. [doi: 10.1192/bjp.172.5.433] [Medline: 9747407]
25. Lau Y, Wang Y, Yin L, Chan KS, Guo X. Validation of the mainland Chinese version of the edinburgh postnatal depression scale in chengdu mothers. Int J Nurs Stud 2010;47(9):1139-1151. [doi: 10.1016/j.ijnurstu.2010.02.005] [Medline: 20219196]

XSL•FO
RenderX

26. Bich TH, Cuong NM. Changes in knowledge, attitude and involvement of fathers in supporting exclusive breastfeeding: a community-based intervention study in a rural area of Vietnam. Int J Public Health 2017;62(Suppl 1):17-26. [doi: 10.1007/s00038-016-0882-0] [Medline: 27586035]
27. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
28. Nickerson LE, Sykes AC, Fung TT. Mothers' experience of fathers' support for breast-feeding. Public Health Nutr 2012;15(9):1780-1787. [doi: 10.1017/S1368980011003636] [Medline: 22281278]
29. Baldwin S, Bick D, Spiro A. Translating fathers' support for breastfeeding into practice. Prim Health Care Res Dev 2021;22:e60 [FREE Full text] [doi: 10.1017/S1463423621000682] [Medline: 34728005]
30. Tohotoa J, Maycock B, Hauck YL, Howat P, Burns S, Binns CW. Dads make a difference: an exploratory study of paternal support for breastfeeding in perth, western australia. Int Breastfeed J 2009;4:15. [doi: 10.1186/1746-4358-4-15] [Medline: 19943958]
31. Yang Z, Ding Y, Song S, Zhang Y, Li A, Su M, et al. Factors affecting the breastfeeding duration of infants and young children in China: a cross-sectional study. Nutrients 2023;15(6):1353 [FREE Full text] [doi: 10.3390/nu15061353] [Medline: 36986082]
32. Chu JTW, Wadham A, Jiang Y, Whittaker R, Stasiak K, Shepherd M, et al. Effect of myTeen SMS-based mobile intervention for parents of adolescents: a randomized clinical trial. JAMA Netw Open 2019;2(9):e1911120 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.11120] [Medline: 31509210]
33. Ngai FW, Chan SWC. Psychosocial factors and maternal wellbeing: an exploratory path analysis. Int J Nurs Stud 2011;48(6):725-731. [doi: 10.1016/j.ijnurstu.2010.11.002] [Medline: 21144517]
34. Hankin BL, Demers CH, Hennessey EMP, Perzow SED, Curran MC, Gallop RJ, et al. Effect of brief interpersonal therapy on depression during pregnancy: a randomized clinical trial. JAMA Psychiatry 2023;80(6):539-547. [doi: 10.1001/jamapsychiatry.2023.0702] [Medline: 37074698]
35. Van Lieshout RJ, Layton H, Savoy CD, Xie F, Brown JSL, Huh K, et al. In-person 1-day cognitive behavioral therapy-based workshops for postpartum depression: a randomized controlled trial. Psychol Med 2023;53(14):1-11. [doi: 10.1017/S0033291723000454] [Medline: 36878891]
36. Cameron EE, Sedov ID, Tomfohr-Madsen LM. Prevalence of paternal depression in pregnancy and the postpartum: an updated meta-analysis. J Affect Disord 2016;206:189-203. [doi: 10.1016/j.jad.2016.07.044] [Medline: 27475890]
37. Salonen AH, Kaunonen M, Astedt-Kurki P, Järvenpää AL, Tarkka MT. Development of an internet-based intervention for parents of infants. J Adv Nurs 2008;64(1):60-72. [doi: 10.1111/j.1365-2648.2008.04759.x] [Medline: 18808593]
38. Rempel JK, Rempel LA, Hoa DT, Vui LT, Long TK. Parenting teamwork: the impact of a fathering intervention on mothers and infants in vietnam. Child Dev 2020;91(2):e345-e364. [doi: 10.1111/cdev.13244] [Medline: 30937897]

Abbreviations

BKQ: Breastfeeding Knowledge Questionnaire
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EBF: Exclusive breastfeeding.
EPDS: Edinburgh Postnatal Depression Scale.
PSOC: Parenting Sense of Competence Scale.

Edited by A Coristine; submitted 05.08.23; peer-reviewed by J Abbass-Dick, R Kelishadi; comments to author 22.02.24; revised version received 10.06.24; accepted 23.08.24; published 17.12.24.

Please cite as:
Huang YY, Wang R, Huang WP, Wu T, Wang SY, R. Redding S, Ouyang YQ
Effects of a Smartphone-Based Breastfeeding Coparenting Intervention Program on Breastfeeding-Related Outcomes in Couples During First Pregnancy: Randomized Controlled Trial
J Med Internet Res 2024;26:e51566
URL: https://www.jmir.org/2024/1/e51566
doi:10.2196/51566
PMID:

©Yi-Yan Huang, Rong Wang, Wei-Peng Huang, Tian Wu, Shi-Yun Wang, Sharon R. Redding, Yan-Qiong Ouyang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association Between the Composite Cardiovascular Risk and mHealth Use Among Adults in the 2017-2020 Health Information National Trends Survey: Cross-Sectional Study

Yuling Chen[1], PhD; Ruth-Alma Turkson-Ocran[2], PhD; Binu Koirala[1], PhD; Patricia M Davidson[1,3], PhD; Yvonne Commodore-Mensah[1,4], PhD; Cheryl Dennison Himmelfarb[1,4,5], PhD

[1]Johns Hopkins University School of Nursing, Baltimore, MD, United States
[2]Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA, United States
[3]University of Wollongong, New South Wales, Australia
[4]Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, United States
[5]Johns Hopkins University School of Medicine, Baltimore, MD, United States

Corresponding Author:
Cheryl Dennison Himmelfarb, PhD
Johns Hopkins University School of Nursing
525 North Wolfe Street
Baltimore, MD, 21205
United States
Phone: 1 443 514 7323
Email: chimmelfarb@jhu.edu

Abstract

Background: Numerous studies have suggested that the relationship between cardiovascular disease (CVD) risk and the usage of mobile health (mHealth) technology may vary depending on the total number of CVD risk factors present. However, whether higher CVD risk is associated with a greater likelihood of engaging in specific mHealth use among US adults is currently unknown.

Objective: We aim to assess the associations between the composite CVD risk and each component of mHealth use among US adults regardless of whether they have a history of CVD or not.

Methods: This study used cross-sectional data from the 2017 to 2020 Health Information National Trends Survey. The exposure was CVD risk (diabetes, hypertension, smoking, physical inactivity, and overweight or obesity). We defined low, moderate, and high CVD risk as having 0-1, 2-3, and 4-5 CVD risk factors, respectively. The outcome variables of interest were each component of mHealth use, including using mHealth to make health decisions, track health progress, share health information, and discuss health decisions with health providers. We used multivariable logistic regression models to examine the association between CVD risk and mHealth use adjusted for demographic factors.

Results: We included 10,531 adults, with a mean age of 54 (SD 16.2) years. Among the included participants, 50.2% were men, 65.4% were non-Hispanic White, 41.9% used mHealth to make health decisions, 50.8% used mHealth to track health progress toward a health-related goal, 18.3% used mHealth to share health information with health providers, and 37.7% used mHealth to discuss health decisions with health providers (all are weighted percentages). Adults with moderate CVD risk were more likely to use mHealth to share health information with health providers (adjusted odds ratio 1.49, 95% CI 1.24-1.80) and discuss health decisions with health providers (1.22, 95% CI 1.04-1.44) compared to those with low CVD risk. Similarly, having high CVD risk was associated with higher odds of using mHealth to share health information with health providers (2.61, 95% CI 1.93-3.54) and discuss health decisions with health providers (1.56, 95% CI 1.17-2.10) compared to those with low CVD risk. Upon stratifying by age and gender, we observed age and gender disparities in the relationship between CVD risk and the usage of mHealth to discuss health decisions with health providers.

Conclusions: Adults with a greater number of CVD risk factors were more likely to use mHealth to share health information with health providers and discuss health decisions with health providers. These findings suggest a promising avenue for enhancing health care communication and advancing both primary and secondary prevention efforts related to managing CVD risk factors through the effective usage of mHealth technology.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e46277) doi:10.2196/46277

KEYWORDS

mobile health; usage; cardiovascular risk; association; mobile phone

Introduction

Cardiovascular disease (CVD) remains the leading cause of mortality in the United States and worldwide [1]. CVD-related mortality can be reduced by controlling modifiable major CVD risk factors, including diabetes mellitus, hypertension, physical inactivity, smoking, and overweight or obesity [2]. Despite the development and implementation of interventions to manage CVD risk factors [3-9], the high prevalence of CVD risk factors persists among both US adults with and without a history of CVD [10].

Mobile health (mHealth) technology is becoming an increasingly important part of the self-management of CVD risk factors facilitated by the widespread adoption of smartphones and tablets [11]. In 2018, a total of 73% of US adults with or at risk for CVD owned a smartphone or tablet [12]. Evidence showed that mHealth technology is beneficial for both primary and secondary prevention of CVD [13,14]. Some systematic reviews, meta-analyses, and clinical trials indicate that mHealth interventions may promote physician-patient communications, control modifiable CVD risk factors, and support cardiovascular health, and the effectiveness of mHealth intervention varied by the frequency of mHealth use [12,15-22].

Several studies have examined disparities in mHealth use among individuals with and without CVD risk factors, which indicated that the connection between CVD risk and mHealth use might exhibit variations depending on the total number of CVD risk factors present. For example, a nationally representative cross-sectional study of 3248 US adults from 2018 Health Information National Trends Survey (HINTS) found that individuals diagnosed with CVD or having CVD risk factors (including diabetes mellitus, hypertension, and current smoking) were more likely to use mHealth to share health information compared with those without CVD or the above CVD risk factors [12]. Nonetheless, it is significant to acknowledge that the study [12] omitted specific CVD risk factors, including physical activity and obesity, which could substantially influence mHealth use [23]. Additionally, the extent to which mHealth use differs based on the combination of CVD risk factors (including diabetes mellitus, hypertension, current smoking, physical inactivity, and obesity) was not explored in that study [12]. Another large cross-sectional study of 256,117 US adults from 2011 to 2018 National Health Interview Survey indicated that health information technologies use among adults with a history of CVD was lower than that of health information technologies use among adults without a history of CVD [15]. Similarly, a cross-sectional study of 28,948 US adults from 2012 to 2018 National Health Interview Survey found that health information technologies use was highest among adults with no CVD risk factors, followed by adults with 1 risk factor, and was lowest among adults with multiple CVD risk factors (including diabetes, hypertension, and obesity) [16]. Of note, these 2 studies mainly focus on the use of health information technologies, such as looking up health information on the internet, filling a web-based prescription, scheduling a medical appointment on the internet, communicating with health care providers through email, or using web-based group chats to learn about health topics [15,16]. Whether higher CVD risk is associated with a greater likelihood of engaging in specific mHealth use (such as sharing health information with health providers) among US adults, regardless of whether they have a history of CVD or not while controlling for demographic variables, is currently unknown.

Moreover, past studies have shown age and gender disparities in mHealth use, where older adults and women were less likely to use mHealth than their younger and male comparators, respectively [10,15,17,19,24-26]. However, how the associations between CVD risk and mHealth use differed by age and gender is not fully elucidated. Thus, we sought to assess the associations between the composite CVD risk and mHealth use among US adults using national data. We also aimed to compare the relationship of the composite CVD risk with mHealth use among younger (<65 years) and older adults (≥65 years) and between women and men to better understand the complex relationship between CVD risk and mHealth use.

Methods

Study Design and Setting

The HINTS is a nationally representative cross-sectional study that has been administered every few years by the National Cancer Institute. The HINTS target population is civilian, noninstitutionalized adults aged 18 years or older living in the United States. HINTS has been administered every few years since 2003, including HINTS 1 (2003), HINTS 2 (2005), HINTS 3 (2008), HINTS 4 (Cycle 1-4, 2011-2014), and HINTS 5 (Cycle 1-4, 2017-2020). Our analysis used HINTS 5 Cycle 1 (2017), HINTS 5 Cycle 2 (2018), HINTS 5 Cycle 3 (2019), and HINTS 5 Cycle 4 (2020) data sets. Cycles 1, 2, and 3 data were collected from January through May in 2017, 2018, and 2019; Cycle 4 data were collected from February through June 2020. The survey (HINTS 5 Cycle 1-4) was conducted exclusively by mail with a US $2 prepaid monetary incentive to encourage participation. In Cycles 1, 2, and 4, the survey was conducted using a paper-based format. In Cycle 3, the survey encompassed both a paper-based version and a web-based survey (mailed contact materials containing a link to the web-based survey). Detailed mailing protocol and full description of the HINTS methodologies can be found elsewhere [27,28]. To increase the precision of estimates for minority subpopulations, the sampling frame of addresses was grouped into high- and low-minority strata and included oversampling of the high-minority stratums. We pooled the data sets from the HINTS 5 Cycle 1-4 to increase the precision of the estimates [27]. Our study adheres to the reporting guidelines outlined in the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology).

XSL•FO
RenderX

Study Population

The target population of this analysis comprises adults aged 18 years and older regardless of whether they have a history of CVD or not (n=16,092). That is, both adults with and without CVD were included in our analysis. CVD status was determined based on the question: "Has a doctor or other health professional ever informed you that you had a heart condition such as heart attack, angina, or congestive heart failure?"

Figure 1 illustrates the flowchart outlining the participant selection process. As the outcomes of interest are each component of mHealth use, we initially excluding those adults who lacked data on all mHealth use components (see definition in the *mHealth Use Assessment* section below). Then, we excluded adults with missing data on any of the CVD risk factors (including type 2 diabetes mellitus, hypertension, current smoking, physical inactivity, and overweight or obesity; see definition in the *Cardiovascular Risk Measurements* section below) to ensure the variable of exposure, the composite CVD risk, is complete. By excluding adults lacking data on all mHealth use components (n=701) and those with missing data on any of the CVD risk factors (n=4860), this study's sample comprised 10,531 adults, including 806 with CVD and 9725 without CVD (refer to Table S1 in Multimedia Appendix 1 for details). Among the 10,531 participants, 9342 responded to the use of mHealth to make a health decision, 9368 responded to the use of mHealth to track progress toward a health-related goal, 9828 responded to the use of mHealth to share health information with health providers, and 9337 responded to the use of mHealth to engage in discussions of health decisions with health providers (Table S2 in Multimedia Appendix 1).

Figure 1. Flowchart of participant selection process. CVD: cardiovascular disease; HINTS: Health Information National Trends Survey; mHealth: mobile health.

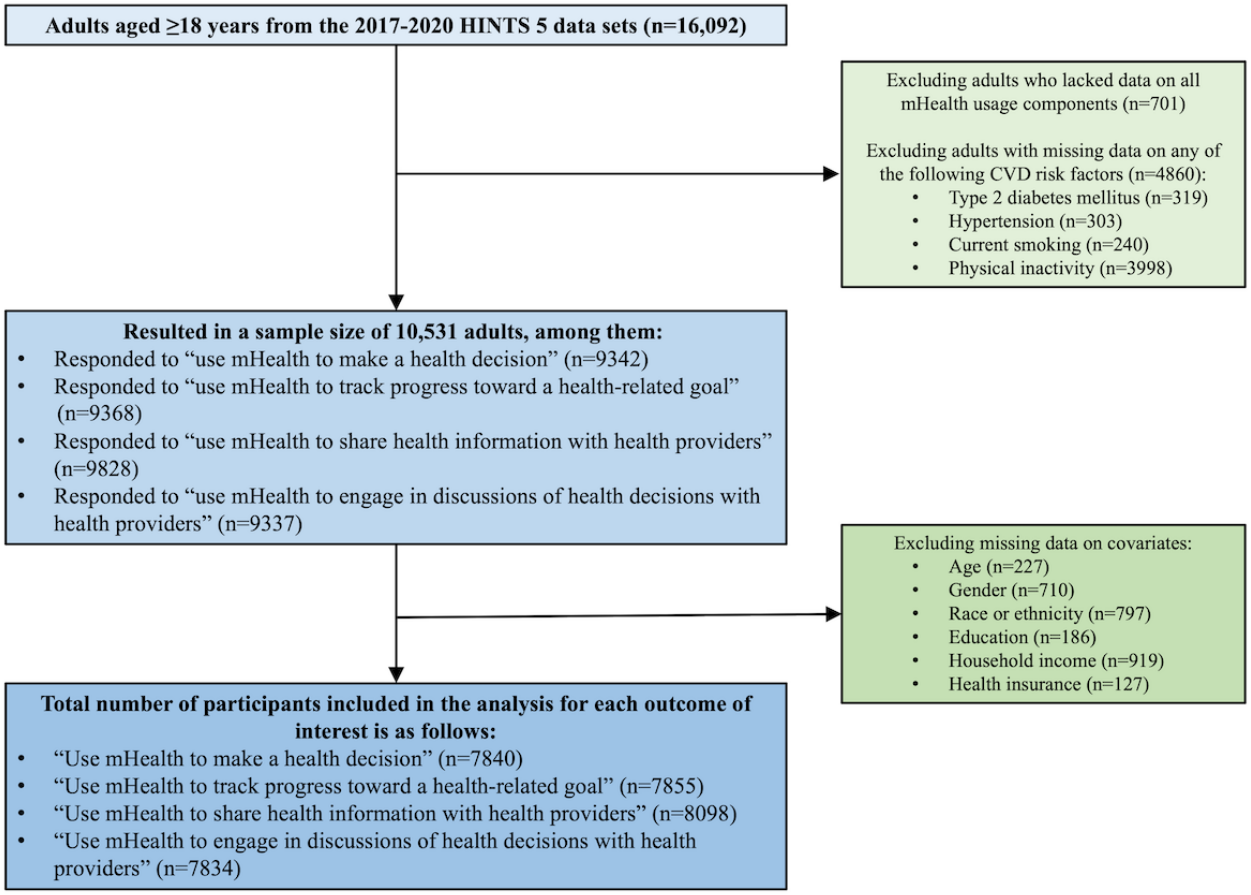

After excluding adults with missing data on covariates (see definition in the *Covariates* section below), the total number of participants included in the analysis for each outcome of interest was as follows: 7840 participants for the outcome of using mHealth to make a health decision, 7855 participants for the outcome of using mHealth to track progress toward a health-related goal, 8098 participants for the outcome of using mHealth to share health information with health providers, and 7834 participants for the outcome of using mHealth to engage in discussions of health decisions with health providers.

Ethical Considerations

HINTS has received approval from the Westat Institutional Review Board and has been designated as exempt by the US National Institutes of Health Office of Human Subjects Research Protections due to the deidentification of the data. Analyses using the HINTS database met the criteria for research involving nonhuman subjects, as determined by the Johns Hopkins University School of Medicine Institutional Review Board, and this analysis did not require review. An expedited approval of

XSL•FO
RenderX

HINTS was obtained for project number 6048.14 (FWA 00005551)

mHealth Use Assessment

The outcome variables of interest consist of each component of mHealth use, which was measured by four self-reported "yes/no" questions as follows: (1) "Has your tablet or smartphone helped you track progress toward a health-related goal, such as quitting smoking, losing weight, or increasing physical activity?" (2) "Has your tablet or smartphone helped you make a decision about how to treat an illness or condition?" (3) "Have you shared health information from either an electronic monitoring device or smartphone with a healthcare professional within the last 12 months?" and (4) "Has your tablet or smartphone helped you in discussions of health decisions with your health providers?"

Cardiovascular Risk Measurements

The exposure of interest in this study was CVD risk, defined as the presence of CVD risk factors. The CVD risk factors assessed in this analysis included type 2 diabetes mellitus, hypertension, current smoking, physical inactivity, and overweight or obesity. Each CVD risk factor was recorded as "0" or "1." According to a previous study [29], we defined low, moderate, and high CVD risk as having 0-1, 2-3, and 4-5 CVD risk factors, respectively.

Diabetes mellitus and hypertension were assessed by 2 self-reported questions: "Has a doctor or other healthcare professional ever told you that you had diabetes or high blood sugar?" and "Has a doctor or other healthcare professional ever told you that you had high blood pressure or hypertension?" Having diabetes or hypertension was recorded as "1."

Smoking status was classified as "current smoking" or "current non-smoking" at the time of the survey. Current smoking was defined as the current use of cigarettes, determined through the combination of the following 2 questions: "Have you smoked at least 100 cigarettes in your entire life?" and "How often do you now smoke cigarettes? (Responses: every day, some days, or not at all)." Current smoking was recorded as "1."

Physical inactivity was defined as <150 minutes of at least moderate-intensity activity each week. Physical activity was derived from the combination of two questions: (1) "How many days do you do any at least moderate physical activity or exercise? (Responses: none, 1 day per week, 2 days per week, 3 days per week, 4 days per week, 5 days per week, 6 days per week, and 7 days per week)" and (2) "How long are you typically doing these activities (minutes and hours)?" Having physical inactivity was recorded as "1."

BMI was initially divided into 6 categories according to World Health Organization (WHO) recommendations: 15.0-18.5 kg/m^2 (underweight), 18.5-24.9 kg/m^2 (normal weight), 25.0-29.9 kg/m^2 (overweight), 30.0-34.9 kg/m^2 (obesity level I), 35.0-39.9 kg/m^2 (obesity level II), and 40-59.9 kg/m^2 (obesity level III) [29]. Having overweight or obesity (level I, II, or III) was recorded as "1."

Covariates

Other covariates examined included baseline age, gender, race, educational levels, household income, location, and health insurance [12]. Age at the time of the interview was divided into 5 categories as follows: 18-34, 35-49, 50-64, 65-74, and 75 years or older. Race was categorized as Hispanic, non-Hispanic Asian, non-Hispanic Black or African American, non-Hispanic others, and non-Hispanic White. Educational levels were classified as less than high school, 12 years or completed high school, some college, and college graduate or higher. Household income was categorized as less than US $20,000; US $20,000-$35,000; US $35,000-$50,000; US $50,000-$75,000; and US $75,000 or more. HINTS comprises 8 questions that inquire about the different types of health insurance held by participants, including insurance obtained through a current or former employer or union, insurance purchased directly from an insurance company, and similar categories. The responses to these questions were compiled into a derived measure of whether or not any health insurance covered the respondent. As a result, health insurance was categorized into 2 groups: "yes" or "no." The location (whether rural or urban) was determined using a single variable available in the HINTS data set, which was classified into categories including metropolitan: large metro, metropolitan: large fringe metro, metropolitan: medium metro, metropolitan: small metro, nonmetropolitan: micropolitan, and nonmetropolitan: noncore. These categories were based on the 2013 National Center for Health Statistics urban-rural classification scheme for counties. We recategorized this variable into 2 groups: rural (nonmetropolitan) and urban (metropolitan) for our analysis.

Statistical Analysis

We described the demographic characteristics and mHealth use using both unweighted and weighted percentages by CVD risk (low, moderate, and high CVD risk). Based on the analytic methods suggested by the *HINTS Users Data Handbook*, we used the survey weighting and Taylor series variance estimation to calculate the prevalence estimated and SEs [27]. Using survey-weighted Pearson chi-square tests by performing "svy:tabulate" command in Stata/SE (version: 17.0; Stata Corp LLC) software, we compared the demographic variables and each component of mHealth use by CVD risk [27,30,31]. We initially ran a full multivariable logistic regression model with survey-weighting to examine the association between the CVD risk category and each component of mHealth use, adjusting for age, gender, race, educational levels, household income, location, and health insurance. Then, we conducted multivariable logistic regression models stratifying by age category (<65 and ≥65 years) and gender.

We also conducted sensitivity analyses excluding individuals previously diagnosed with CVD and individuals aged ≥85 years. In addition, to explore potential variations in the associations between CVD risk factors and mHealth use, we performed 4 distinct weighted multivariate logistic regression models, with each component of mHealth use serving as the dependent variable. In each model, the independent variables included the 5 CVD risk factors (type 2 diabetes mellitus, hypertension,

XSL•FO
RenderX

current smoking, physical inactivity, and overweight or obesity), and was adjusted for the previously mentioned covariates.

We performed all statistical analyses using the "svy" command in Stata/SE (version 17.0). In multivariable logistic regression analyses, we excluded cases with missing data on any of the outcome variables or covariates as the default approach. Adjusted odds ratio and 95% CI were calculated for multivariable logistic regression models. A 2-sided *P* value of <.05 was considered statistically significant for all analyses.

Results

We included 10,531 adults, with a mean age of 54 (SD 16.2) years (Table S2 in Multimedia Appendix 1). Table 1 displays the survey-weighted demographic characteristics of the included participants. Of these, 50.2% (weighted percentage) were men, 65.4% (weighted percentage) were non-Hispanic White, 36.9% (weighted percentage) had a bachelor's degree, and 45.3% (weighted percentage) had an annual income US $75,000. Unweighted participants' proportions of demographic characteristics by CVD risk can be found in Table S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Weighted demographic characteristics by cardiovascular risk among adults with or at risk for CVD[a] (N=10,531).

Characteristics	All (N=10,531; weighted percentage, %)	Low CVD risk[b] (n=4332; weighted percentage, %)	Moderate CVD risk[c] (n=5316; weighted percentage, %)	High CVD risk[d] (n=883; weighted percentage, %)	*P* value
Age group (y)					<.001
18-34	26.91	37.74	19.89	3.45	
35-49	27.78	27.31	28.88	23.25	
50-64	29.76	23.82	33.10	46.18	
65-74	10.57	7.57	12.32	18.47	
≥75	4.98	3.56	5.81	8.65	
Gender					.006
Women	49.80	52.83	47.06	48.59	
Men	50.20	47.17	52.94	51.41	
Race and ethnicity					<.001
Hispanic	15.93	16.27	15.37	17.60	
Non-Hispanic Asian	5.55	6.49	4.90	3.69	
Non-Hispanic Black	10.02	7.07	11.87	17.20	
Non-Hispanic others	3.08	3.10	3.18	2.22	
Non-Hispanic White	65.42	67.07	64.68	59.29	
Education					<.001
Less than high school	5.68	3.24	6.91	13.40	
High school graduate	18.91	15.45	20.83	28.68	
Some college	38.50	36.61	40.15	39.71	
Bachelor's degree	36.91	44.70	32.11	18.21	
Household income (US $)					<.001
<20,000	13.58	11.15	14.41	24.02	
20,000 to <35,000	9.21	7.96	9.60	14.86	
35,000 to <50,000	13.10	11.61	13.91	17.42	
50,000 to <75,000	18.79	17.43	20.43	16.41	
≥75,000	45.33	51.85	41.66	27.29	
Insurance					.01
No	7.42	6.16	8.93	5.27	
Yes	92.58	93.84	91.07	94.73	
Location					.011
Rural	12.26	10.64	13.55	14.10	
Urban	87.74	89.36	86.45	85.90	

[a]CVD: cardiovascular disease.

[b]Low CVD risk: 0-1 risk factors.

[c]Moderate CVD risk: 2-3 risk factors.

[d]High CVD risk: 4-5 risk factors.

Among the included participants, 41.9% (weighted percentage) used mHealth to make health decisions, 50.8% (weighted percentage) used mHealth to track health progress toward a health-related goal, 18.3% (weighted percentage) used mHealth to share health information with health providers, and 37.7% (weighted percentage) used mHealth to discuss health decisions with health providers (Table 2). Significant differences were observed in the proportion of participants using mHealth to track health progress and share health information with health providers among the 3 CVD risk categories (all *P*s<.001).

XSL•FO
RenderX

Table 2. Weighted percentages and SEs of mHealth[a] usage by cardiovascular risk among adults with or at risk for CVD[b].

mHealth use	All, weighted percentage (SE)	Low CVD risk[c], weighted percentage (SE)	Moderate CVD risk[d], weighted percentage (SE)	High CVD risk[e], weighted percentage (SE)	*P* value
Make health decisions	41.90 (0.9)	41.14 (1.5)	42.59 (1.3)	42.51 (3.1)	.76
Track health progress toward a health-related goal	50.78 (0.9)	53.81 (1.5)	49.53 (1.2)	38.02 (3.1)	<.001
Share health information with health providers	18.29 (0.6)	14.97 (0.9)	19.91 (0.9)	28.89 (1.9)	<.001
Discuss health decisions with health providers	37.72 (0.9)	34.94 (1.2)	38.10 (1.4)	40.52 (3.0)	.50

[a]mHealth, mobile health.

[b]CVD: cardiovascular disease.

[c]Low CVD risk: 0-1 risk factors.

[d]Moderate CVD risk: 2-3 risk factors.

[e]High CVD risk: 4-5 risk factors.

After adjusting for age, gender, race, education, household income, location, and health insurance, adults with moderate CVD risk were more likely to use mHealth to share health information with health providers (adjusted odds ratio 1.49, 95% CI 1.24-1.80) and to discuss health decisions with health providers (1.22, 95% CI 1.04-1.44) compared to individuals with low CVD risk (Figure 2). Adults with high CVD risk was associated with higher odds of using mHealth to share health information with health providers (2.61, 95% CI 1.93-3.54) and to discuss health decisions with health providers (1.56, 95% CI 1.17-2.10) when compared to those with low CVD risk (Figure 2). There were no significant associations found between moderate or high CVD risk and the use of mHealth for both making health decisions and tracking progress on health-related goals, as compared to adults with low CVD risk (Figure 2).

Figure 2. Associations between cardiovascular risk and mHealth use among adults with or at risk for CVD. The results are from weighted multivariable logistic regression models. Each logistic regression model was adjusted for age, sex, race, education, household income, location, and health insurance. Low CVD risk: 0-1 risk factors; moderate CVD risk: 2-3 risk factors; high CVD risk: 4-5 risk factors. CVD: cardiovascular disease; mHealth: mobile health; Ref: reference.

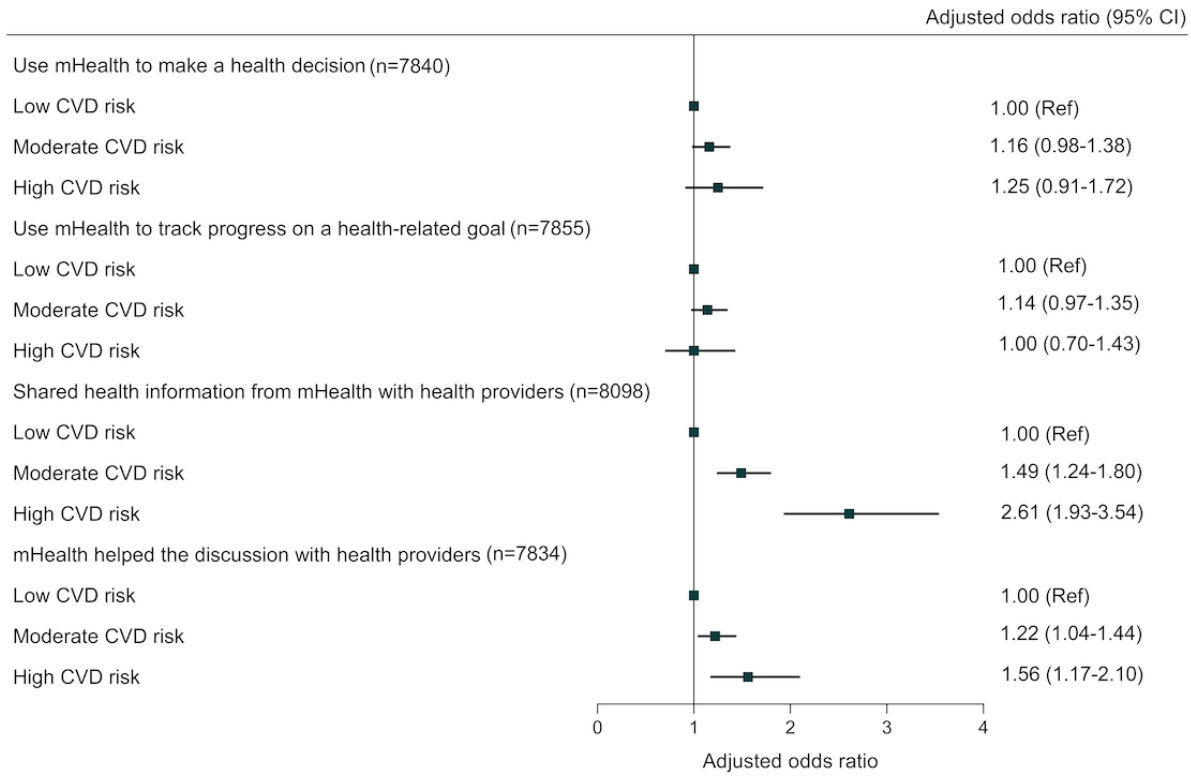

Upon stratifying by age, we observed age disparities in the relationship between CVD risk and mHealth use (Figure 3). Among adults aged <65 years, adults with moderate or high CVD risk were more likely to use mHealth to share health

XSL•FO
RenderX

information and discuss health decisions with health providers, compared to those with low CVD risk. Among adults aged ≥65 years, those with moderate CVD risk (1.59, 95% CI 1.14-2.23) and high CVD risk (3.10, 95% CI 1.95-4.93) were more inclined to use mHealth to share information with health providers, compared to individuals with low CVD risk. However, neither moderate nor high CVD risk were statistically significant factors associated with the use of mHealth for discussing health decisions with health providers in this age group.

Figure 3. Associations between cardiovascular risk and mHealth use among older and younger adults with or at risk for CVD. Low CVD risk: 0-1 risk factors; moderate CVD risk: 2-3 risk factors; high CVD risk: 4-5 risk factors. Each weighted multivariable logistic regression model was adjusted for sex, race, education, household income, location, and health insurance. CVD: cardiovascular disease; mHealth: mobile health; Ref: reference.

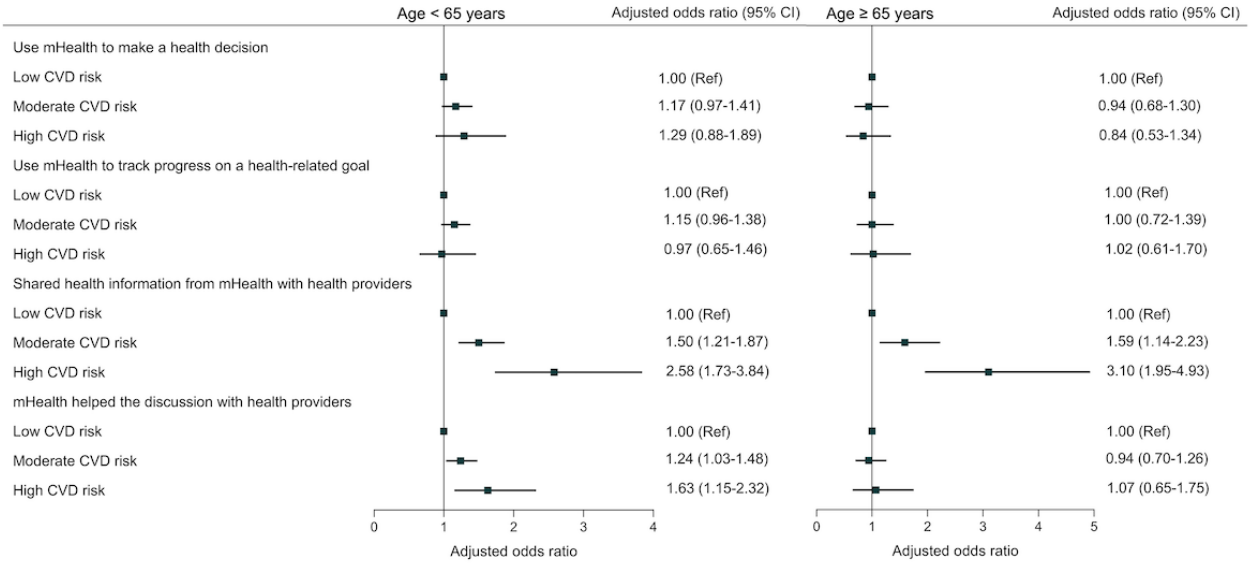

After stratifying by gender, gender disparities in the odds of CVD risk associated with mHealth use were evident (Figure 4). For both women and men, having moderate and high CVD risk were significantly associated with increased odds of using mHealth to share health information with health providers when compared to adults with low CVD risk. Among women, having high CVD risk was associated with higher odds of using mHealth to discuss health decisions with health providers (1.89, 95% CI 1.28-2.80) compared with adults with low CVD risk; however, this significant association was not observed among men (1.24, 95% CI 0.75-2.03).

Figure 4. Associations between cardiovascular risk and mHealth use among women and men with or at risk for CVD. Low CVD risk: 0-1 risk factors; moderate CVD risk: 2-3 risk factors; high CVD risk: 4-5 risk factors. Each weighted multivariable logistic regression model was adjusted for age, race, education, household income, location, and health insurance. CVD: cardiovascular disease; mHealth: mobile health; Ref: reference.

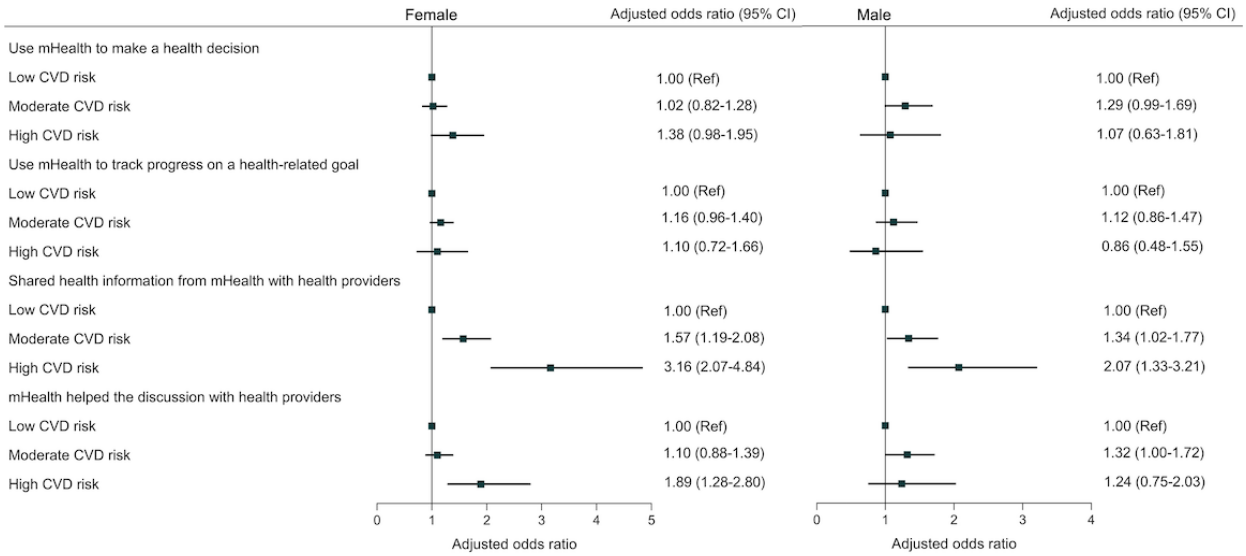

Our findings remained robust with the exclusion of individuals previously diagnosed as having heart disease (n=848) and with the exclusion of individuals aged ≥85 years (n=489; Table S3 in Multimedia Appendix 1).

In the assessment of the associations between each CVD risk factor and mHealth use (Table S4 in Multimedia Appendix 1), we observed that adults with diabetes (1.62, 95% CI 1.26-2.07), hypertension (1.88, 95% CI 1.54-2.30), and overweight or

XSL•FO
RenderX

obesity (1.28, 95% CI 1.50-1.57) were more inclined to share health information through mHealth technology with health care providers compared to their counterparts. Additionally, adults who were overweight or obese (1.39, 95% CI 1.14-1.70) were more likely to use mHealth to track progress toward health-related goals than those with a normal BMI. Furthermore, adults with hypertension (1.38, 95% CI 1.19-1.61) were more inclined to use mHealth for discussions with health care providers compared to those without a hypertension diagnosis.

Discussion

Principal Findings

In this nationally representative sample, we observed that adults with moderate CVD risk (2-3 CVD risk factors) and high CVD risk (4-5 CVD risk factors) were more likely to use mHealth to share health information with health providers than those with low CVD risk (0-1 CVD risk factors). However, no significant associations were detected between moderate or high CVD risk and the usage of mHealth for both making health decisions and tracking progress toward health-related goals. These associations were consistent across various age groups and genders. Age and gender disparities in the likelihood of CVD risk being associated with the use of mHealth for discussing health decisions with health providers were evident. Specifically, among adults aged <65 years and women, individuals with high CVD risk were more inclined to use mHealth for discussing health decisions with health providers compared to those with low CVD risk.

Comparison With Prior Work

Research has demonstrated that the usage of mHealth varies significantly by the presence of chronic conditions or CVD risk factors and demographics (including age, gender, race, educational levels, insurance, and region of residence) [10,15,17,19,24-26]. Numerous studies have explored disparities in mHealth use among individuals both with and without CVD risk factors [12,15]. These investigations have suggested that the relationship between CVD risk and mHealth use could potentially exhibit variations based on the total number of CVD risk factors present. To better explain the associations between CVD risk and mHealth use, we used multivariable logistic regression models, adjusting for multiple demographics identified in the previous studies [12,15]. To gauge the strength of the associations between CVD risk and mHealth use, we categorized individuals into 3 groups: low, moderate, and high CVD risk. These categories allowed us to assess the magnitude of the relationships between CVD risk and mHealth use effectively.

Our findings expand prior reports regarding mHealth use linked to CVD risk with additional patients and analyses. Shan and colleagues [12] used 2018 HINTS data to examine the association between mHealth use among individuals with or at risk for CVD. In their report using 2018 HINTS data [12], compared with adults with no history or risk factors for CVD, adults with CVD risk (defined as reporting a heart condition, diabetes mellitus, hypertension, or current smoking) were more likely to use mHealth to share health information with health providers [12]. In our study, which encompassed a larger and more extensive participant pool spanning 4 years, we likewise found that adults with moderate and high CVD risk (including type 2 diabetes mellitus, hypertension, current smoking, physical inactivity, and overweight or obesity) were more likely to use mHealth to share health information with their health providers compared to those with low CVD risk, corroborating the earlier research findings [12]. This trend suggests that adults with greater CVD risk may perceive mHealth as a valuable tool for enhancing communication with health care professionals, possibly due to a heightened awareness of their health status and the need for frequent monitoring and consultation [15]. This finding highlights the potential of mHealth to facilitate communication between patients and health care professionals, especially in populations at greater risk of CVD [12]. It is crucial to highlight that our study's definition of CVD risk differs from the prior report. Unlike the previous study, our analysis includes factors such as physical inactivity and overweight or obesity but excludes heart conditions. Our study thoroughly examines the relationships between the composite CVD risk and mHealth use across various age groups and genders. To facilitate this study of the dose-response relationship, we categorized CVD risk into low, moderate, and high levels. The results revealed that adults with a moderate CVD risk (2-3 risk factors) were 1.5 times more likely and those with a high CVD risk (4-5 risk factors) were 2.6 times more likely to use mHealth for sharing health information with their health care providers, compared to those with a low CVD risk.

Furthermore, it is notable that we found adults with moderate and high CVD risk displayed a greater likelihood to use mHealth for engaging in discussions with health providers. This observation contradicts the results of a previous study, which reported no difference in the odds of using mHealth to facilitate discussions with health providers [12]. These discrepancies can potentially be attributed to the prior study's exclusion of 2 crucial CVD risk factors: physical inactivity and obesity [12]. Additionally, that study did not investigate how mHealth use varies in relation to the various levels of multiple CVD risk factors [12]. However, our study revealed age and gender disparities in the likelihood of mHealth use for discussing health decisions with health providers. Specifically, among adults aged 65 years or younger, those with moderate or high CVD risk were more inclined to use mHealth for discussing health decisions with health providers compared to those with low CVD risk. These findings align with previous studies that older adults tend to be less likely to use the mHealth for making health-related decisions or discussions with health providers than younger adults [17,19,24,25]. This finding is not surprising, given that older adults may have lower levels of digital health literacy and be less capable of using smartphones, tablets, and complex health-related apps [32,33]. A prior study indicated that older adults might experience the first-level digital divide (lack of access to technologies) or the second-level digital divide (lack of use or skill in data input and analysis) [32]. Other barriers for older adults in mHealth use may include lack of knowledge in using mHealth, decreased sensory perception, poorly designed interface, absence of professional involvement, and the high cost of using mHealth [33,34]. With the rapidly increasing number of older adults globally [35], future researchers should address the barriers to mHealth use or access

XSL•FO
RenderX

among older adults and develop person-centered mHealth-based interventions for CVD risk reduction [32].

In terms of gender, women with high CVD risk were more inclined to use mHealth for discussing health decisions with health providers compared to those with low CVD risk. These findings are similar to a previous study showing women with multiple chronic conditions had higher odds of using their mHealth technologies to discuss with their providers and manage their care [36]. The gender disparities may also reflect variations in health care–seeking behavior, with women tending to exhibit a stronger social motivation for engaging in health-related information searches, and they derive greater enjoyment from these activities [37]. These gender-specific digital inequalities suggest that gender differences should be acknowledged in developing mHealth interventions to empower CVD risk factors management [38].

While the primary focus of our study was to investigate the relationship between the composite CVD risk and mHealth use, we also explored the connection between various individual CVD risk factors and mHealth use. The findings revealed that adults with diabetes, hypertension, and overweight or obesity were more inclined to share health information using mHealth with health care providers. Moreover, overweight or obese individuals demonstrated a higher propensity to use mHealth for tracking health-related goals compared to those with a normal BMI. Additionally, adults with hypertension were more likely to engage in discussions with health care providers through mHealth compared to those without a hypertension diagnosis. These findings indicate that adults with multiple CVD health conditions, specifically diabetes and hypertension, might be more likely to use mHealth technology [12]. This raises the question of whether this pattern—higher CVD risk being associated with increased mHealth technology usage—varies based on the total number of CVD health conditions (including diabetes, hypertension, and hyperlipidemia) and CVD health-related behaviors (such as physical activity, smoking, and diet). Future research could expand its analyses by integrating additional CVD health conditions (eg, hyperlipidemia) and health-related behaviors (eg, diet) to attain a more comprehensive understanding of the relationship between the presence of CVD health conditions, CVD health behaviors, and mHealth use.

Strengths and Limitations

This study used the recent 4 years of HINTS data, a nationally representative survey, to assess the associations between CVD risk and several specific mHealth use behaviors. There were several limitations in our study. First, this study was limited by the design nature (cross-sectional study design). Therefore, it is impossible to establish a cause-and-effect relationship between the composite CVD risk factor control and mHealth use. This study is also unable to assess the chronological relationship between the development of CVD risk factors and mHealth use. Second, there may be response bias due to the self-reported data. However, our analyses used the survey weighting method suggested by the HINTS statistical methodology. Third, hyperlipidemia is also a major CVD risk factor that is not included in this study due to the lack of such questions regarding hyperlipidemia in the 2017-2020 HINTS data set. Fourth, an important limitation of our study lies in the absence of adjustments for health conditions or illnesses, such as cancer, which could potentially impact the relationship between CVD risk and mHealth use. Future research should consider including health conditions or illnesses as covariates to further explore this association. Finally, a substantial proportion of missing data in the variables of interest raises concerns about potential bias in our findings. To mitigate this issue, future studies should employ suitable methods, such as multiple imputations, based on the underlying mechanisms causing the missing data.

Conclusions

In this nationally representative sample of US adults regardless of whether they have a history of CVD or not, adults with a greater number of CVD risk factors displayed a heightened propensity to actively use mHealth technology for sharing health information and engaging in health-related discussions with their health providers. This trend was particularly pronounced among younger adults and women when compared to those with a low CVD risk. This suggests a promising avenue for improving health care communication and advancing both primary and secondary prevention efforts related to managing CVD risk factors through the effective usage of mHealth technology.

Acknowledgments

We thank the investigators of the Health Information National Trends Survey (HINTS) study included in this analysis for their hard work and this study's participants who were included in this analysis. We used the generative artificial intelligence tool ChatGPT by OpenAI to assist in revising certain aspects of this paper's language, which was further reviewed and revised by the coauthors. The original ChatGPT transcripts are made available in Multimedia Appendix 2.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

YC developed the research concepts, acquired the data, conducted all statistical analyses, and wrote this paper. RAT-O, BK, PMD, YC-M, and CDH contributed to writing and reviewing this paper. CDH provided statistical expertise and supervision. All authors reviewed the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Unweighted frequencies and weighted multivariable logistic regression model results.
[DOCX File , 40 KB - jmir_v26i1e46277_app1.docx]

Multimedia Appendix 2
Original ChatGPT transcripts.
[DOCX File , 18 KB - jmir_v26i1e46277_app2.docx]

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020;76(25):2982-3021 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
2. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Circulation 2019;140(11):e596-e646 [FREE Full text] [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
3. Lewey J, Murphy S, Zhang D, Putt ME, Elovitz MA, Riis V, et al. Effectiveness of a text-based gamification intervention to improve physical activity among postpartum individuals with hypertensive disorders of pregnancy: a randomized clinical trial. JAMA Cardiol 2022;7(6):591-599 [FREE Full text] [doi: 10.1001/jamacardio.2022.0553] [Medline: 35442393]
4. Stephen C, Halcomb E, Fernandez R, McInnes S, Batterham M, Zwar N. Nurse-led interventions to manage hypertension in general practice: a systematic review and meta-analysis. J Adv Nurs 2022;78(5):1281-1293. [doi: 10.1111/jan.15159] [Medline: 35244944]
5. Mehta S, Zhao J, Poppe K, Kerr AJ, Wells S, Exeter DJ, et al. Cardiovascular risk management requires a combination of cardiovascular preventive pharmacotherapy and non-pharmacological interventions. Eur J Prev Cardiol 2022 07;29(12):e312-e313. [doi: 10.1093/eurjpc/zwac082] [Medline: 35512449]
6. Jordan KP, Rathod-Mistry T, Bailey J, Chen Y, Clarson L, Denaxas S, et al. Long-term cardiovascular risk and management of patients recorded in primary care with unattributed chest pain: an electronic health record study. J Am Heart Assoc 2022;11(7):e023146 [FREE Full text] [doi: 10.1161/JAHA.121.023146] [Medline: 35301875]
7. Teo KK, Rafiq T. Cardiovascular risk factors and prevention: a perspective from developing countries. Can J Cardiol 2021;37(5):733-743 [FREE Full text] [doi: 10.1016/j.cjca.2021.02.009] [Medline: 33610690]
8. Kris-Etherton PM, Petersen KS, Després JP, Anderson CAM, Deedwania P, Furie KL, et al. Strategies for promotion of a healthy lifestyle in clinical settings: pillars of ideal cardiovascular health: a science advisory from the American Heart Association. Circulation 2021;144(24):e495-e514 [FREE Full text] [doi: 10.1161/CIR.0000000000001018] [Medline: 34689589]
9. Laddu D, Ma J, Kaar J, Ozemek C, Durant RW, Campbell T, et al. Health behavior change programs in primary care and community practices for cardiovascular disease prevention and risk factor management among midlife and older adults: a scientific statement from the American Heart Association. Circulation 2021;144(24):e533-e549 [FREE Full text] [doi: 10.1161/CIR.0000000000001026] [Medline: 34732063]
10. He J, Zhu Z, Bundy JD, Dorans KS, Chen J, Hamm LL. Trends in cardiovascular risk factors in US adults by race and ethnicity and socioeconomic status, 1999-2018. JAMA 2021;326(13):1286-1298 [FREE Full text] [doi: 10.1001/jama.2021.15187] [Medline: 34609450]
11. Khoong EC, Olazo K, Rivadeneira NA, Thatipelli S, Barr-Walker J, Fontil V, et al. Mobile health strategies for blood pressure self-management in urban populations with digital barriers: systematic review and meta-analyses. npj Digit Med 2021;4(1):114 [FREE Full text] [doi: 10.1038/s41746-021-00486-5] [Medline: 34294852]
12. Shan R, Ding J, Plante TB, Martin SS. Mobile health access and use among individuals with or at risk for cardiovascular disease: 2018 Health Information National Trends Survey (HINTS). J Am Heart Assoc 2019;8(24):e014390 [FREE Full text] [doi: 10.1161/JAHA.119.014390] [Medline: 31818220]
13. Burke LE, Ma J, Azar KM, Bennett GG, Peterson ED, Zheng Y, et al. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. Circulation 2015;132(12):1157-1213 [FREE Full text] [doi: 10.1161/CIR.0000000000000232] [Medline: 26271892]

XSL•FO
RenderX

14. Schorr EN, Gepner AD, Dolansky MA, Forman DE, Park LG, Petersen KS, et al. Harnessing mobile health technology for secondary cardiovascular disease prevention in older adults: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2021;14(5):e000103 [FREE Full text] [doi: 10.1161/HCQ.0000000000000103] [Medline: 33793309]
15. Nwokeji U, Spaulding EM, Shan R, Turkson-Ocran RA, Baptiste D, Koirala B, et al. Health information technology use smong persons with self-reported atherosclerotic cardiovascular disease: analysis of the 2011-2018 National Health Interview Survey. J Med Internet Res 2021;23(8):e23765 [FREE Full text] [doi: 10.2196/23765] [Medline: 34397391]
16. Gandrakota N, Ali MK, Shah MK. Trends in health information technology use among the US population with and without cardiovascular risk factors, 2012-2018: evidence from the National Health Interview Survey. JMIR Public Health Surveill 2021;7(9):e29990 [FREE Full text] [doi: 10.2196/29990] [Medline: 34591026]
17. Asan O, Cooper Ii F, Nagavally S, Walker RJ, Williams JS, Ozieh MN, et al. Preferences for health information technologies among US Adults: analysis of the Health Information National Trends Survey. J Med Internet Res 2018;20(10):e277 [FREE Full text] [doi: 10.2196/jmir.9436] [Medline: 30341048]
18. Akinosun AS, Polson R, Diaz-Skeete Y, De Kock JH, Carragher L, Leslie S, et al. Digital technology interventions for risk factor modification in patients with cardiovascular disease: systematic review and meta-analysis. JMIR mHealth uHealth 2021;9(3):e21061 [FREE Full text] [doi: 10.2196/21061] [Medline: 33656444]
19. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the National Cancer Institute's Health Information National Trends Survey Data. J Med Internet Res 2020;22(5):e16299 [FREE Full text] [doi: 10.2196/16299] [Medline: 32406865]
20. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. URL: https://apps.who.int/iris/handle/10665/44607 [accessed 2023-12-12]
21. Bae JW, Woo SI, Lee J, Park SD, Kwon SW, Choi SH, et al. mHealth interventions for lifestyle and risk factor modification in coronary heart disease: randomized controlled trial. JMIR mHealth uHealth 2021;9(9):e29928 [FREE Full text] [doi: 10.2196/29928] [Medline: 34559058]
22. Cruz-Cobo C, Bernal-Jiménez MÁ, Vázquez-García R, Santi-Cano MJ. Effectiveness of mHealth interventions in the control of lifestyle and cardiovascular risk factors in patients after a coronary event: systematic review and meta-analysis. JMIR mHealth uHealth 2022;10(12):e39593 [FREE Full text] [doi: 10.2196/39593] [Medline: 36459396]
23. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
24. Yang X, Yang N, Lewis D, Parton J, Hudnall M. Patterns and influencing factors of eHealth tools adoption among Medicaid and non-Medicaid populations from the Health Information National Trends Survey (HINTS) 2017-2019: questionnaire study. J Med Internet Res 2021;23(2):e25809 [FREE Full text] [doi: 10.2196/25809] [Medline: 33599619]
25. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the Health Information National Trends Survey. Int J Environ Res Public Health 2020;17(18):6856 [FREE Full text] [doi: 10.3390/ijerph17186856] [Medline: 32961766]
26. Seckin G, Hughes S. Patient-reported outcomes in a nationally representative sample of older internet users: cross-sectional survey. JMIR Aging 2021;4(4):e16006 [FREE Full text] [doi: 10.2196/16006] [Medline: 34822340]
27. Health Information National Trends Survey 5 (HINTS 5) Cycle 1 methodology report. Health Information National Trends Survey. URL: https://hints.cancer.gov/docs/methodologyreports/HINTS5_Cycle_1_Methodology_Rpt.pdf [accessed 2023-09-23]
28. Health Information National Trends Survey 5 (HINTS 5) Cycle 3 methodology report. Health Information National Trends Survey. URL: https://hints.cancer.gov/docs/Instruments/HINTS5_Cycle3_MethodologyReport.pdf [accessed 2023-09-23]
29. Yahya T, Acquah I, Taha MB, Valero-Elizondo J, Al-Mallah MH, Chamsi-Pasha MA, et al. Cardiovascular risk profile of Middle Eastern immigrants living in the United States-the National Health Interview Survey. Am J Prev Cardiol 2022;9:100312 [FREE Full text] [doi: 10.1016/j.ajpc.2021.100312] [Medline: 35024678]
30. Okobi OE, Sobayo TO, Arisoyin AE, Adeyemo DA, Olaleye KT, Nelson CO, et al. Association between the use of wearable devices and physical activity among US adults with depression and anxiety: evidence from the 2019 and 2020 Health Information National Trends Survey. Cureus 2023;15(5):e39521 [FREE Full text] [doi: 10.7759/cureus.39521] [Medline: 37366453]
31. Stata Survey Data Reference Manual. URL: https://www.stata.com/manuals/svy.pdf [accessed 2023-09-17]
32. Cosco TD, Firth J, Vahia I, Sixsmith A, Torous J. Mobilizing mHealth data collection in older adults: challenges and opportunities. JMIR Aging 2019;2(1):e10019 [FREE Full text] [doi: 10.2196/10019] [Medline: 31518253]
33. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han HR. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018;36(8):376-382 [FREE Full text] [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
34. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022;12(3):e054561 [FREE Full text] [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]

XSL•FO
RenderX

35. He W, Goodkind D, Kowal P. An Aging World. 2015. URL: https://www.census.gov/library/publications/2016/demo/P95-16-1.html [accessed 2022-08-02]
36. Ajayi KV, Wachira E, Onyeaka HK, Montour T, Olowolaju S, Garney W. The use of digital health tools for health promotion among women with and without chronic diseases: insights from the 2017-2020 Health Information National Trends Survey. JMIR mHealth uHealth 2022;10(8):e39520 [FREE Full text] [doi: 10.2196/39520] [Medline: 35984680]
37. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015;17(6):e156 [FREE Full text] [doi: 10.2196/jmir.4127] [Medline: 26099325]
38. van Elburg FRT, Klaver NS, Nieboer AP, Askari M. Gender differences regarding intention to use mHealth applications in the Dutch elderly population: a cross-sectional study. BMC Geriatr 2022;22(1):449 [FREE Full text] [doi: 10.1186/s12877-022-03130-3] [Medline: 35610577]

Abbreviations

CVD: cardiovascular disease
HINTS: Health Information National Trends Survey
mHealth: mobile health
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by A Mavragani; submitted 05.02.23; peer-reviewed by V Buss, C Yang; comments to author 19.08.23; revised version received 24.09.23; accepted 11.10.23; published 04.01.24.

Please cite as:
Chen Y, Turkson-Ocran RA, Koirala B, Davidson PM, Commodore-Mensah Y, Himmelfarb CD
Association Between the Composite Cardiovascular Risk and mHealth Use Among Adults in the 2017-2020 Health Information National Trends Survey: Cross-Sectional Study
J Med Internet Res 2024;26:e46277
URL: https://www.jmir.org/2024/1/e46277
doi:10.2196/46277
PMID:38175685

©Yuling Chen, Ruth-Alma Turkson-Ocran, Binu Koirala, Patricia M Davidson, Yvonne Commodore-Mensah, Cheryl Dennison Himmelfarb. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Preliminary Efficacy of a Digital Intervention for Adolescent Depression: Randomized Controlled Trial

Emily Peake[1,2*], MS; Ian Miller[1,3*], BA; Jessica Flannery[1,4], PhD; Lang Chen[5], PhD; Jessica Lake[1,2], PhD; Aarthi Padmanabhan[1,2], PhD

[1]Limbix Health Inc, San Francisco, CA, United States
[2]Big Health Inc, San Francisco, CA, United States
[3]Digital Medicine Society, Boston, MA, United States
[4]Akili, Boston, MA, United States
[5]Neuroscience Unit, Santa Clara University, Santa Clara, CA, United States
[*]these authors contributed equally

Corresponding Author:
Aarthi Padmanabhan, PhD
Limbix Health Inc
548 Market St, PMB 91609
San Francisco, CA, 94104
United States
Phone: 1 888 546 2495
Email: aarthi.padmanabhan@bighealth.com

Abstract

Background: Adolescent depression is a significant public health concern; however, access to effective mental health care is limited. Digital therapeutics (DTx) can improve access to evidence-based interventions; however, their efficacy in adolescents is sparsely documented.

Objective: This study aims to examine the efficacy of a mobile app DTx versus an active control as an adjunct treatment for adolescent depression symptoms.

Methods: An internet-based open-label randomized control trial was conducted nationwide with a partial crossover design, and 168 adolescents aged 13 to 21 years with symptoms of depression were recruited between November 2020 and September 2021. Participants were randomized (1:1) to the cognitive behavioral therapy–based treatment app (Spark) or to a psychoeducational control app (control), which they would use for a duration of 5 weeks. The primary outcome was a between-group (Spark vs control) difference in the change in depression symptoms from baseline to postintervention, as measured by the Patient Health Questionnaire-8 (PHQ-8) using a linear mixed-effects analysis. The PHQ-8 ranges from 0 to 24, with scores of 5 to 9 indicating mild depression symptoms, scores of 10 to 14 indicating moderate symptoms, scores of 15 to 19 indicating moderately severe symptoms, and scores of 20 to 24 indicating severe symptoms. A minimal clinically important difference (5-point reduction between baseline and postintervention) in the Spark arm and group differences in remission and treatment response rates based on the PHQ-8 at postintervention were also investigated.

Results: A total of 160 participants were randomized, 80 in the Spark arm (mean age 16.89, SD 2.5 y) and 80 in the control arm (mean age 16.79, SD 2.59 y). Data from 121 participants (Spark: n=63; control: n=58) with moderate to severe (PHQ-8≥10) symptoms at baseline were included in the primary analyses following a modified intention-to-treat principle. A linear mixed-effect analysis revealed a nonsignificant difference between the study arms in depression symptom change over the intervention period. The Spark arm met a minimal clinically important difference threshold (mean −5.08, 95% CI −6.72 to −3.42). The remission rate in the Spark arm was significantly higher than that in the control arm (11/63, 17% vs 2/58, 3%; χ^2_1=6.2; *P*=.01; false discovery rate–adjusted *P*=.03). The treatment response rates were not significantly different between the study arms (*P*=.07; false discovery rate–adjusted *P*=.16). Post hoc analyses including participants with mild to severe (PHQ-8 score ≥5) symptoms at baseline revealed promising evidence that Spark is effective in those with mild to severe symptoms.

Conclusions: There is initial evidence that a self-guided, cognitive behavioral therapy–based DTx intervention may effectively treat mild to severe depression symptoms in adolescents. DTx may improve access to mental health care for adolescents or serve as an important adjunct to the standard of care.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT04524598; https://clinicaltrials.gov/study/NCT04524598

(J Med Internet Res 2024;26:e48467) doi:10.2196/48467

KEYWORDS

adolescent; depression; randomized controlled trial; mobile phone; digital therapeutics; mobile app; cognitive behavioral therapy; behavioral activation; mobile health

Introduction

Background

Depression is a pervasive mental health disorder affecting approximately 4 million adolescents in the United States, and this number is expected to grow annually [1-4]. Depression can be lifelong, often emerging in adolescence and increasing the risk of poor physical health, social and relational problems, academic and employment difficulties, and reduced overall well-being [5-7]. Despite a growing need for effective treatment options, adolescents access [3,6] and adhere to [8,9] mental health treatment at alarmingly low rates owing to a number of factors. Stigma and a general difficulty in identifying symptoms of a mental illness inhibit adolescents' willingness or awareness of seeking treatment [10,11]. When they do seek help, adolescents face months-long waitlists [12]. Other external forces drive the unmet clinical needs of adolescents experiencing depression, including costs [2,13], a widening divide in the accessibility of treatment [14,15], and an outstripped supply of mental health professionals [14,16]. Barriers in access to mental health treatment and depression incidence rates in adolescents have been exacerbated by the COVID-19 pandemic, necessitating action and new solutions [17-21].

Digital health programs, accessible through smartphones and computers, offer potential avenues for delivering treatment options for mental health conditions, such as depression [22-29]. However, it is essential to acknowledge their limitations, which include the absence of evidence-based protocols and clinical validation [24,25,27]. Digital therapeutics (DTx), which are evidence-based software programs designed to prevent, treat, or manage health conditions, serve to address these limitations [30]. Smartphone-based DTx are particularly well suited for adolescents given their technological literacy, ubiquitous cell phone use, and privacy preferences [22,29,31,32]. DTx are also easily scalable, can serve as adjunct to standard care, improve treatment fidelity [33-35], are cost-effective [36]. These technologies have been shown to be preliminarily effective in treating adolescents with a variety of mental health concerns including attention-deficit/hyperactivity disorder [37], sleep disorders, eating disorders, and anxiety [38]. Ultimately, DTx fill critical gaps in non-DTx digital treatments and can help patients receive effective, data-driven treatment with the added benefit of home access at any time [35,39,40].

Cognitive behavioral therapy (CBT) is a widely recognized and recommended treatment for depression and is considered the standard of care for adolescents [41-43]. Behavioral activation (BA), an important component of CBT [44-47], involves engaging in value-based behaviors that are rewarding or elicit a sense of mastery to help reduce depression symptoms using a combination of motivational strategies, reward seeking, and natural reinforcers, alongside reducing maladaptive and avoidant behaviors. BA has been shown to be an effective stand-alone treatment for adolescent depression [44-48]. Digital CBT applications have demonstrated promise in treating depression [23,26,49,50], and a BA-focused digital therapeutic may represent an effective and accessible treatment option for adolescents.

Limitations such as low user engagement and insufficient existing clinical validation (eg, randomized controlled trials [RCTs]) may limit the widespread adoption of DTx [24,25,27,49]. To our knowledge, there are no DTx for depression tailored specifically for adolescents, including BA-based DTx, and limited evidence exists regarding the efficacy of a smartphone-based DTx for adolescent depression [51-53].

Objectives

This study evaluates the efficacy of Spark, a 5-week self-guided mobile app intervention based on validated BA treatment protocols for adolescents with depression [44,45]. Adolescents aged 13 to 21 years were recruited nationwide into an internet-based RCT comparing Spark with an active control. Our main goal was to assess the preliminary efficacy of Spark in treating moderate to severe depression symptoms as an adjunct to usual care by measuring group differences in (1) changes in depression symptoms over time, including whether each group had a MCID in depression symptoms; (2) the proportion of participants who achieved remission; and (3) the proportion of participants who demonstrated treatment response. As this RCT was conducted between November 2020 and September 2021, another goal was to provide mental health resources to adolescents experiencing symptoms of depression during the COVID-19 pandemic.

Methods

Study Design

This open-label, partial crossover RCT compared Spark with an app with psychoeducational content (control app) in adolescents aged 13 to 21 years with self-identified depression symptoms. The study was conducted in 2 phases: an evaluation of feasibility (phase 1) and a preliminary evaluation of efficacy (phase 2). The results from phase 1 (Spark version 2.0) have been reported elsewhere [54]. Here, we focus on the results from phase 2, which combined data collected using Spark version 2.1 and version 2.2 (refer to the Spark section for more information on the versions). Updates to the Spark app to version 2.2 were deployed approximately halfway during the study; version 2.1 was used between September 2020 and March 2021, and version 2.2 was used between June 2021 and September 2021. The study design, randomization, control app, and

XSL•FO
RenderX

outcome measures were the same across both versions. There was an a priori plan to combine the data from the 2 Spark versions for analysis.

Ethical Considerations

The study was approved by the Institutional Review Board of the Western Copernicus Group (20201686) under a nonsignificant risk investigational device exemption and overseen by an independent data and safety monitoring board. The trial was also registered on ClinicalTrials.gov (NCT04524598).

Participants

Participants were recruited nationwide via internet-based advertising and word of mouth. The following inclusion criteria were applied in phase 2: (1) aged 13 to 21 years, (2) self-reported symptoms of depression, (3) residing in the United States for the duration of the 5-week study, (4) under the care of a US-based primary care or licensed mental health care provider and willing and able to provide the name and contact information of the provider during consent appointment, (5) English fluency and literacy of adolescents and consenting legal guardians if aged <18 years, (6) access to a smartphone (iPhone 5s or later or running Android 4.4 KitKat or later) and regular internet access, (7) willingness to provide informed e-consent or assent and have legal guardian willing to provide informed e-consent if aged >18 years, and (8) stable for at least 2 months on any treatment (including medication or psychotherapy) for a mental health disorder.

Participants were excluded if they self-reported any of the following: (1) self-reported lifetime suicide attempt or active self-harm or active suicidal ideation with intent; (2) diagnosed by a clinician with bipolar disorder, substance use disorder, or any psychotic disorder including schizophrenia; (3) incapable of understanding or completing study procedures and digital intervention as determined by the participant, patient or legal guardian, health care provider, or clinical research team, and (4) previously participated in the user testing or clinical testing of the Spark app.

The legal guardians provided consent for participants aged <18 years. Adolescent participants provided consent unless they were under 18. If they were under 18, their legal guardians provided consent and they provided assent.

Procedure

During a virtual consent and enrollment session over video conferencing, participants and their legal guardians (if aged <18 y) were provided with study details, provided electronic informed consent or assent, were assessed for eligibility, and completed web-based baseline questionnaires. Participants downloaded an app on their smartphone and were provided with a safety plan template with instructions [55] to complete it on their own as a personal resource. Participants were randomized 1:1 to the Spark or control arm using a block randomization approach ranging in multiples of 2 from 6 to 12 [56].

All participants had access to their assigned app for the 5-week intervention period. During the intervention period, participants completed an in-app Patient Health Questionnaire-8 (PHQ-8) [57], a self-report questionnaire to assess symptoms of depression, and an internally developed symptom check questionnaire weekly, which they could complete anytime over a 7-day period. Legal guardians completed a weekly symptom check about their child on the web. Following the 5-week intervention period, app access was restricted, and participants and legal guardians completed web-based postintervention questionnaires. Participants randomized to the control arm were also offered access to the Spark app after the 5-week intervention period and completed postintervention questionnaires on the web again following this second intervention period (partial crossover data not reported here). Baseline and postintervention measures assessed participant characteristics, concurrent treatments, depression and anxiety symptoms, resilience, app feedback, and impacts of COVID-19 (refer to Table 1 for schedule and full descriptions) [57-63].

XSL•FO
RenderX

Table 1. Schedule of assessments and descriptions.

Assessment name	Baseline	Week 1	Week 2	Week 3	Week 4	Week 5 postintervention	Description
Baseline questionnaire[a,b]	✓						An internally developed self-report and legal guardian–report questionnaire to collect gender, age, ethnicity, race, education and parent or legal guardian education, medication use, previous and current treatment for mental health disorders, depression diagnosis status, other mental or neurological disorder diagnosis status, and positive and negative impacts of COVID-19.
Brief Resilience Scale[a] [58]	✓						A standardized 6-item self-report measure for assessing the ability to recover from stress. Each item is rated on a frequency scale from 1 (strongly disagree) to 5 (strongly agree).
Patient Health Questionnaire-8[a] [57]	✓	✓	✓	✓	✓	✓	A standardized 8-item self-report measure of depression symptom severity over the past 2 weeks. Each item is rated on a frequency scale from 0 (never) to 3 (nearly every day).
Weekly symptom check[a,b]	✓	✓	✓	✓	✓	✓	An internally developed self-report and legal guardian–report questionnaire to collect information on any negative symptoms or side effects experienced during the intervention period, the severity of each event, and its relatedness to the mobile app.
Generalized Anxiety Disorder-7[a] [59]	✓					✓	A standardized 7-item self-report measure of anxiety symptom severity over the past 2 weeks. Each item is rated on a frequency scale from 0 (never) to 3 (nearly every day).
PROMIS[c] Pediatric Global 7+2[a] [60]	✓					✓	A standardized 7+2–item self-report measure that produces a unidimensional measure of global health perception or well-being and separate measures of fatigue and pain. Each item is rated on a frequency scale from 1 to 5.
PROMIS Pediatric Global 7+2 Parent Proxy[b] [60]	✓					✓	A standardized 7+2–item legal guardian–report measure that produces a unidimensional measure of global health perception or well-being and separate measures of fatigue and pain. Each item is rated on a frequency scale from 1 to 5.
Mood and Feelings Questionnaire Parent Form[b] [61]	✓					✓	A standardized 13-item legal guardian–report measure to assess how the participant has been feeling or acting over the past 2 weeks. Each item is rated on a frequency scale from 0 (not true) to 2 (true).
System usability scale[a] [62]						✓	A standardized 10-item self-report measure on the usability of systems such as computer programs or mobile apps, including the perceived ease of use, complexity, and consistency of systems. Each item is rated on a frequency scale of 1 (strongly disagree) to 5 (strongly agree).
User engagement scale-short form[a] [63]						✓	A standardized 12-item self-report measure on 4 specific engagement factors, including focused attention, perceived usability, esthetic appeal, and reward. Each item is rated on a frequency scale of 1 (strongly disagree) to 5 (strongly agree).
Postintervention questionnaire[a]						✓	An internally developed self-report and legal guardian–report questionnaire to collect information regarding any change in concurrent treatment for depression during the study, mood or depression symptom improvement and enjoyment associated with the app, general feedback about the app or research participation, and positive and negative impacts of COVID-19 since baseline.

[a]Completed by adolescent.

[b]Completed by caregiver.

[c]PROMIS: Patient-Reported Outcomes Measurement Information System.

Interventions

Both the Spark and control programs were divided into 5 modules recommended to be completed at a pace of 1 module per week over the 5-week intervention period, but the participants were able to progress at their own pace. Content for a given week was not expected to take >60 minutes to complete. All participants were prompted to complete a weekly PHQ-8 and a clinical concerns questionnaire on the mobile app. Automated app notification reminders to complete these

XSL•FO
RenderX

questionnaires were sent. If users had not opened the app in 3 days, an automated app notification encouraging participants to use the app was sent. Automated reminder notifications were sent 7 days before the end of the intervention period to remind participants that the intervention period would be ending in 7 days. Crisis resources were available for participants to access anytime in each app.

Spark (Versions 2.1 and 2.2)

The Spark app is based on CBT, which implements BA [44,45]. A character called "Limbot" is used as a therapeutic guide to encourage users to complete activities and model examples of activities for users. In the app, participants read text, answered questions, inputted text, and completed interactive activities. Participants were encouraged to schedule activities to be completed outside of the app and reflect on the impact on their mood. Tasks in the mobile app progress in a linear fashion, that is, each task must be completed to progress to the next task.

Version 2.1 of Spark included a 5-level program focused on providing psychoeducational content and delivering the BA model of depression by teaching 2 core skills (mood activity logging and activity scheduling). Version 2.2 was also divided into 5 levels and expanded on the content of version 2.1, adding problem-solving, mindfulness, and relapse prevention content. Version 2.2 also included a reward system and animations for completing certain activities, some design changes to the user interface, and an increased number of in-app notifications. The 5 levels of Spark version 2.1 were as follows:

- Level 1 (Start Your Journey): Program introduction and learning about the BA model of depression
- Level 2 (Making Choices): Mood tracking and up and down activities
- Level 3 (Solving Problems): Learn about activity scheduling and complete 3 activations,
- Level 4 (Staying Active): Complete 4 activations and
- Level 5 (Journey's End): Complete 5 activations.

Version 2.2 expanded on version 2.1 with the addition of new features, UI elements, and content. The 5 levels of Spark version 2.2 were as follows:

- Level 1: Onboarding and Introduction to BA (no differences)
- Level 2: Mood tracking (no differences)
- Level 3: Mindfulness and Activity Scheduling (addition of 2 psychoeducational tasks teaching and reinforcing mindfulness skills)
- Level 4: Problem-Solving and Activity Scheduling (addition of 2 psychoeducational tasks teaching and reinforcing problem-solving skills), and
- Level 5: Relapse Prevention and Activity Scheduling (addition of 6 psychoeducational tasks teaching relapse prevention skills).

Psychoeducational Control

The control app contained 5 modules of age-appropriate psychoeducational content related to the neurobiology of depression and did not contain any active CBT or BA components. In this app, participants read the text on screen related to the brain and behavior, the adolescent brain, depression in the brain, neurobiological factors influencing depression, and personality. The same version of the control app was used for the entire study. The 5 levels of the control app included Lesson 1: Understanding Behavior, Lesson 2: Exploring the Brain, Lesson 3: Mastering Messengers, Lesson 4: Riding the Wave, and Lesson 5: People and Personality.

Safety Monitoring

A rigorous safety monitoring and classification procedure was followed. Potential safety-related events (ie, clinical concerns), which included worsening or persistently high depressive symptoms, any imminent risk events, suicidal ideation, hospitalizations, injury, illness, nonsuicidal self-injury, and direct or indirect indications of abuse, were identified in the following ways: (1) any concerning information provided to study staff during the onboarding session, in email, text, or phone correspondence during participation; (2) symptom deterioration based on weekly PHQ-8 scores was defined as (a) PHQ-8 score that was ≥15 and ≥5 points higher than the baseline score or (b) PHQ-8 scores ≥20 for 2 weeks in a row during the intervention, and (3) during the 5 week intervention period, all participants and their legal guardians (for those aged >18 y) were asked to complete a weekly internally developed symptom check questionnaire that asked them to report any negative symptoms or side effects that they experienced over the past week, to rate how negative each experience was on a scale of 0 (not at all) to 4 (extremely), and whether they believe each experience was caused by the app; (4) for participants randomized into the Spark arm, there were opportunities within the app that allowed for freeform text input. This freeform text was reviewed daily by study staff.

Data and incoming correspondence were reviewed daily and any identified potential clinical concerns were recorded by the study staff and verified by the study investigators. All logged clinical concerns were reviewed daily by a study investigator who determined whether the clinical concern required escalation to the study clinician (Dr Raph Rose), an independent licensed clinical psychologist not otherwise associated with the study or study sponsor, for clinical input or follow-up with the participant. The study clinician would then determine whether the participant was safe and eligible to continue with the study based on the information provided or based on contact with the participant or legal guardian. If a clinical concern related to suicidality was endorsed during the onboarding session, a study investigator administered the Ask Suicide-Screening Questions Toolkit [64] to the participant to determine the imminent risk level, recommend emergency resources if required, and escalate to the study clinician for follow-up.

Participants were withdrawn from the study if clinical concerns met the following criteria: (1) if the study clinician determined that the participant was no longer eligible to continue with the study, (2) if the clinician could not monitor safety because of not being able to reach the participant or other listed contacts, and (3) if the clinician could not monitor safety because participants did not complete the weekly symptom check questionnaire for 2 consecutive weeks. In any of these

XSL•FO
RenderX

circumstances, the participant was informed, withdrawn from the study, and sent a list of mental health resources via email.

Safety Classification

Safety classification was based on the following definitions set forth by the US Food and Drug Administration (FDA) and FDA-recognized consensus standards, *ISO 14155:2020 Clinical Investigation of Medical Devices for Human Subjects-Good Clinical Practice* [65-67]. After study completion, a clinician who was not otherwise involved in the study (JF) reviewed all clinical concern data and provided preliminary event classifications. Any events deemed to be potential adverse events (AEs) by this clinician were sent to the study clinician (Dr Rose) for external classification. Dr Rose was blinded to identifying participant information, participant ID, and JF's ratings to make final determinations. Information was presented to Dr Rose in a manner that concealed the group condition. Nonetheless, it is possible that Dr Rose could discern the group identification for certain classifications based on the content provided by a participant. To reduce the potential for bias in classification, if Dr Rose provided a stricter classification rating than JF, that rating was maintained.

An AE was defined as an "untoward medical occurrence, unintended disease or injury, or untoward clinical signs (including abnormal laboratory findings) in subjects, users or other persons, whether or not related to the investigational medical device [67] and whether anticipated or unanticipated."

An adverse device effect (ADE) was defined as an AE "related to the use of an investigational medical device" [67]. This includes any AE "resulting from insufficiencies or inadequacies in the instructions for use, the deployment, the implantation, the installation, the operation, or any malfunction of the investigational medical device." This also includes "any event that is a result of a user error or intentional misuse" [67]. For this study, ADEs could have occurred in either the Spark or control arms.

A serious AE (SAE) or serious ADE was defined as an AE or ADE that met more than one of the following criteria: Resulted in fatality, posed a life-threatening risk or immediate risk of death at the time of occurrence, led to persistent or significant disability or incapacity, required prolonged inpatient hospitalization, represented an important medical event, as determined by appropriate medical judgment, that could jeopardize the participant's well-being, or where medical or surgical intervention might be necessary to prevent one of the aforementioned outcomes. This did not include planned hospitalization for a preexisting condition [65].

Unanticipated ADEs (UADEs), as defined in the FDA regulation 21 Code of Federal Regulation 812.3 [65], also referred to as "unanticipated problems," included any serious adverse effect on health or safety or any life-threatening problem or death caused by, or associated with, a device, if that effect, problem, or death was not previously identified in nature, severity, or degree of incidence in the investigational plan or application; or any other unanticipated serious problem associated with a device that relates to the rights, safety, or welfare of participants.

Outcomes and Statistical Analysis

An a priori statistical analysis plan was restricted to participants with moderate to severe symptoms at baseline (PHQ-8≥10; moderate-to-severe cohort), unless otherwise specified. The α level was set to P=.05, and false discovery rate (FDR) correction for multiple comparisons was applied to the specified analyses.

Depression Symptoms: PHQ-8

The primary outcome was a group difference in the change in depression symptoms from baseline to 5 weeks, as measured by the PHQ-8. A modified intention-to-treat (mITT) approach was used for primary analyses by including all participants randomized within the moderate-to-severe cohort. A per protocol (PP) analysis included data from participants in this cohort who had completed all weekly PHQ-8 questionnaires. Post hoc mITT and PP analyses were also conducted for participants with a baseline PHQ-8≥5 (mild-to-severe cohort) to evaluate the efficacy of Spark in participants with mild to severe symptoms of depression. Statistical analyses were performed using R version 4.1.2 (R Foundation for Statistical Computing) by an independent external statistician (LC) [68].

Missing Data Analysis

Little's test [69] was used to determine whether group differences existed in the proportion of missing PHQ-8 data across weeks for both moderate-to-severe and mild-to-severe cohorts. For any resulting significant results, the effects of known factors on missing data including Spark version, treatment group, week, baseline PHQ-8 severity, and age group were evaluated with follow-up χ^2 tests to determine whether data could be missing at random [70].

Multiple Imputation Procedure

Multiple imputation was implemented on missing data from participants who had completed at least the baseline PHQ-8 assessment. Information on Spark version (2.1 and 2.2), treatment group, baseline PHQ-8 score, age group, week, and individual PHQ-8 item was included to impute 100 data sets with the Pan [71] method using the R *mice* package version 3.14.0 [68,72,73]. Missing PHQ-8 item–level scores and assessment completion days from baseline were imputed.

Statistical Analysis

A linear mixed-effects model (LMM) analysis was implemented on an averaged imputed data set to evaluate the main effects of study arm (Spark and control) and week (0-5), and the study arm×week interaction using the R lme4 package (version 1.1.28 [74]). Four models were implemented in total, including the mITT and PP analyses for both the moderate-to-severe and mild-to-severe cohorts. In each model, group and week were entered as fixed factors. Spark version and assessment completion days from baseline were included as fixed factors to control for the effects of app version and differences in time between the completion of successive weekly assessments. PHQ-8 item–level scores and participants were included as random factors for the intercept, and the time from the baseline assessment was included as a random factor for the slope of participants. FDR correction was applied to the 4 P values for the study arm×week interaction effects. The effect size of the

XSL•FO
RenderX

interaction (Cohen f^2) was computed using pseudo R^2 ($f^2=R^2/\{1-R^2\}$) [75,76] with the R MuMIn package (version 1.46.0 [77]). Dfs were estimated using the Satterthwaite method provided in the R lmerTest package (version 3.1-3) [78,79]. Follow-up analyses were performed to evaluate the effect of week within each group.

We also evaluated whether there was an average minimal clinically important difference (MCID) in symptom severity within each group, defined as a ≥5 point average decrease in PHQ-8 score between baseline and postintervention [80].

To assess the robustness of the study arm×week interaction, a generalized LMM (GLMM) with multiple imputation with the same model specification as the LMM analysis was implemented on item-level PHQ-8 data across the 100 imputed data sets using a 2-level Pan method [71], and the resulting 4 *P* values were FDR adjusted.

Group differences in remission rates, defined as PHQ-8 score <5 at postintervention [81], and treatment response rates, defined as a 50% reduction in PHQ-8 score between baseline and postintervention [50], were tested using χ^2 tests, and the resulting 4 *P* values were FDR adjusted.

Secondary Clinical Outcomes

Secondary outcomes included anxiety symptoms (Generalized Anxiety Disorder Scale [59]), legal guardian–rated depression symptoms of their child (Mood and Feelings Questionnaire 61]), and participant-rated and legal guardian–rated (of their child) global functioning (Patient-Reported Outcomes Measurement Information System [60]) assessed at baseline and postintervention. Means, SDs, and 95% CIs for the average difference between baseline and postintervention were computed for these outcomes. A questionnaire was developed internally to assess the positive and negative effects of COVID-19, and the proportion of participants in each group that endorsed each item was computed.

Safety

The total numbers of AE, ADE, SAEs, and UADE were computed per group for all randomized participants.

Other Outcomes

Other outcomes measured included program adherence, engagement, and acceptability. Program adherence was assessed as the proportion of participants completing all sessions by postintervention and percent completion per module. Engagement for both study arms was assessed using the User Engagement Scale-short form [63], minutes spent in the app per week, and total app sessions. Usability was assessed using the system usability scale [62]. App acceptability was assessed with internally developed self-reported and legal guardian–reported ratings of the app with a 10-point Likert scale, asking how much the app improved mood or symptoms of depression and how enjoyable it was. Means, SDs, and mean differences between study arms with 95% CIs were calculated for these measures.

Results

Participants

A total of 168 adolescents consented to participate in this study between November 2020 and June 2021. The postintervention and follow-up data were collected between January 2021 and September 2021, when planned data collection had completed. Eight participants did not meet the inclusion criteria. A total of 160 patients were randomized (80 in the Spark arm; refer to Figure 1 for the CONSORT (Consolidated Standards of Reporting Trials) diagram; refer to Table 2 for the characteristics). In total, 41 participants received Spark version 2.1 and 39 participants received Spark version 2.2. Data from participants with a baseline PHQ-8 score ≥10 (moderate-to-severe cohort; Table 2) were included in all planned analyses, including mITT (n=121) and PP (n=86). Post hoc analyses included data from participants with a baseline PHQ-8 score≥5 (mild-to-severe cohort; Table 2) for mITT (n=153) and PP (n=109). Three participants skipped a question on the PHQ-8 assessment and were included in the PP analyses. The Spark and control arms did not differ significantly on any of the baseline demographic characteristics (all *P*>.05).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. The diagram excludes the control extension arm. Patients were considered lost to follow-up if they did not complete the postintervention questionnaire, considered to have dropped out if they missed 2 weekly safety checks, and considered to have withdrawn from the study if they asked to be removed. Patients were considered as missing week 5 if they did not complete the week 5 questionnaires but moved on to the control extension phase. mITT: modified intention-to-treat; PHQ-8: Patient Health Questionnaire-8; PP: per protocol.

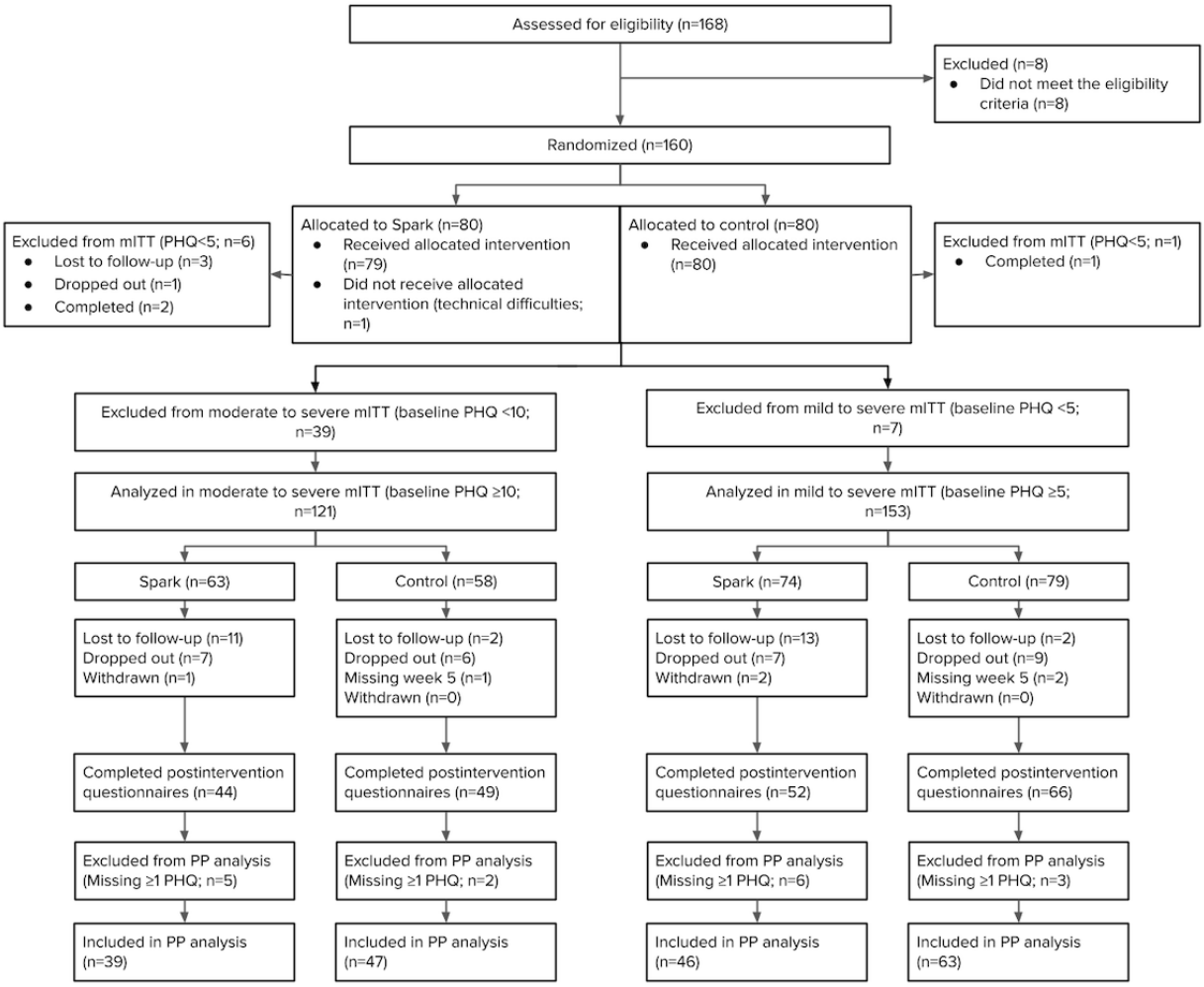

XSL•FO
RenderX

Table 2. Participant characteristics (N=160).

	Spark participants (n=80)	Control participants (n=80)
Full sample, n (%)	80 (100)	80 (100)
Age (y), mean (SD)	16.89 (2.50)	16.79 (2.59)
Gender, n (%)		
Female	50 (63)	51 (64)
Male	24 (30)	19 (24)
Nonbinary	6 (8)	10 (13)
Race, n (%)		
American Indian or Alaska native	2 (3)	2 (3)
Asian	6 (8)	1 (1)
Black or African American	5 (6)	6 (8)
Pacific islander	0 (0)	0 (0)
Unknown	0 (0)	1 (1)
White	59 (74)	59 (74)
>1 race	8 (10)	11 (14)
Ethnicity, n (%)		
Hispanic or Latinx	13 (16)	10 (13)
Not Hispanic or Latinx	61 (76)	64 (80)
Unknown	6 (8)	6 (8)
Population density, n (%)		
Rural	2 (3)	2 (3)
Suburban	6 (8)	10 (13)
Urban	72 (90)	68 (85)
Depression diagnosis, n (%)	61 (76)	69 (86)
Concurrent treatment, n (%)		
Psychotherapy	36 (34)	36 (34)
Medication	39 (37)	46 (43)
Other	4 (4)	5 (5)
None	26 (25)	19 (18)
Past treatment, n (%)		
Psychotherapy	54 (44)	57 (44)
Medication	44 (36)	52 (40)
Other	11 (9)	9 (7)
None	15 (12)	11 (9)
Moderate-severe cohort (n=121), n (%)	63 (100)	58 (100)
Age (y), mean (SD)	17.06 (2.49)	16.46 (2.43)
Gender, n (%)		
Female	42 (67)	39 (67)
Male	16 (25)	10 (17)
Nonbinary	5 (8)	9 (16)
Race, n (%)		
American Indian or Alaska native	1 (2)	2 (3)
Asian	4 (6)	0 (0)

XSL•FO
RenderX

	Spark participants (n=80)	Control participants (n=80)
Black or African American	3 (5)	2 (3)
Pacific islander	0 (0)	0 (0)
Unknown	0 (0)	1 (2)
White	47 (75)	45 (78)
>1 race	8 (13)	8 (14)
Ethnicity, n (%)		
Hispanic or Latinx	11 (18)	8 (14)
Not Hispanic or Latinx	46 (73)	47 (81)
Unknown	6 (10)	3 (5)
Population density, n (%)		
Rural	1 (2)	1 (2)
Suburban	5 (8)	6 (10)
Urban	57 (90)	51 (88)
Depression diagnosis, n (%)	51 (81)	51 (88)
Concurrent treatment, n (%)		
Psychotherapy	28 (34)	23 (31)
Medication	33 (40)	35 (47)
Other	4 (5)	3 (4)
None	18 (22)	14 (19)
Past treatment, n (%)		
Psychotherapy	46 (46)	41 (43)
Medication	37 (37)	40 (42)
Other	10 (10)	6 (6)
None	8 (8)	8 (8)
Mild-to-severe cohort (n=153), n (%)	74 (100)	79 (100)
Age (y), mean (SD)	17.00 (2.48)	8 (13.8)
Gender, n (%)		
Female	48 (65)	50 (63)
Male	20 (27)	19 (24)
Nonbinary	6 (8)	10 (13)
Race, n (%)		
American Indian or Alaska native	2 (3)	2 (3)
Asian	4 (5)	1 (1)
Black or African American	4 (5)	5 (6)
Pacific islander	0 (0)	0 (0)
Unknown	0 (0)	1 (1)
White	56 (76)	59 (75)
>1 race	8 (11)	11 (14)
Ethnicity, n (%)		
Hispanic or Latinx	13 (18)	10 (13)
Not Hispanic or Latinx	55 (74)	63 (80)
Unknown	6 (8)	6 (8)
Population density, n (%)		

XSL•FO
RenderX

	Spark participants (n=80)	Control participants (n=80)
Rural	2 (3)	2 (3)
Suburban	6 (8)	10 (13)
Urban	67 (91)	67 (85)
Depression diagnosis, n (%)	58 (78)	68 (86)
Concurrent treatment, n (%)		
Psychotherapy	33 (34)	35 (34)
Medication	38 (38)	45 (43)
Other	4 (4)	5 (5)
None	23 (24)	19 (18)
Past treatment, n (%)		
Psychotherapy	51 (44)	56 (44)
Medication	43 (37)	51 (40)
Other	11 (9)	9 (7)
None	12 (10)	11 (9)

Treatment Outcomes: PHQ-8

Weekly means, SD, and 95% CIs for baseline to postintervention change scores per group; remission and treatment response rates per group are reported in Table 3.

Table 3. Patient Health Questionnaire-8 descriptives.

Cohort and group	Baseline score, mean (SD)	Week 1 score, mean (SD)	Week 2 score, mean (SD)	Week 3 score, mean (SD)	Week 4 score, mean (SD)	Postintervention score, mean (SD)	Postbaseline score, mean difference (95% CI)	Remission, n (%)	Treatment response, n (%)
Moderate to severe (n=121)									
Spark (n=63)	15.78 (3.62)	13.04 (5.06)	11.70 (5.56)	11.80 (5.55)	10.75 (6.13)	10.70 (5.54)	−5.08 (−6.72 to −3.42)	11 (17)	15 (24)
Control (n=58)	15.31 (3.42)	13.06 (4.37)	12.30 (4.77)	11.68 (4.88)	11.70 (5.69)	11.80 (4.99)	−3.51 (−5.09 to −1.93)	2 (3)	8 (14)
Mild to severe (n=153)									
Spark (n=74)	14.36 (4.78)	11.99 (5.39)	10.73 (5.70)	10.65 (5.88)	9.56 (6.38)	9.51 (5.91)	−4.85 (−6.59 to −3.09)	20 (27)	22 (30)
Control (n=79)	13.29 (4.51)	11.59 (4.73)	10.88 (5.09)	10.25 (5.09)	10.56 (5.47)	10.43 (5.39)	−2.86 (−4.43 to −1.30)	9 (11)	12 (15)

Moderate-to-Severe Cohort

Modified Intention to Treat (n=121)

The LMM revealed a nonsignificant study arm x week interaction ($t_{127.66}$=−1.911; *P*=.06; FDR adjusted *P*=.06; f^2=0.0012; power: 0.468-0.494; Figure 2A). There was a nonsignificant effect of study arm ($t_{177.54}$=.573; *P*=.57; f^2=0.00034; power: 0.069-0.079) and a significant effect of week ($t_{1977.06}$=−4.395; *P*<.001; f^2=0.00215; power: 0.985-0.993). There was a significant effect of week in the Spark arm ($t_{1759.08}$=−3.508; *P*<.001; change score=−5.08, 95% CI −6.72 to −3.42) and control arm ($t_{561.98}$=−3.244; *P*=.001; change score=−3.51, 95% CI −5.09 to −1.93). The GLMM also produced a nonsignificant study arm x week interaction ($F_{1,24186.41}$=3.223; *P*=.07; FDR adjusted *P*=.07).

At the end of the intervention period, the Spark arm showed significantly higher remission rates compared with the control arm (Spark, 11/63, 17% and control, 2/58, 3%; χ^2_1=6.183; *P*=.01; FDR adjusted *P*=.03). Treatment response rates between the study arms (Spark, 15/63, 24% and control, 8/58, 14%) were not statistically significant (χ^2_1=1.968; *P*=.07; FDR adjusted *P*=.16).

XSL•FO
RenderX

Figure 2. Depression severity by week: (A) modified intention-to-treat (mITT) moderate to severe cohort, (B) per protocol (PP) moderate to severe cohort, (C) mITT mild to moderate cohort, and (D) PP mild to moderate cohort. Note that the graphs depict average Patient Health Questionnaire-8 (PHQ-8) scores at each time point based on observed data.

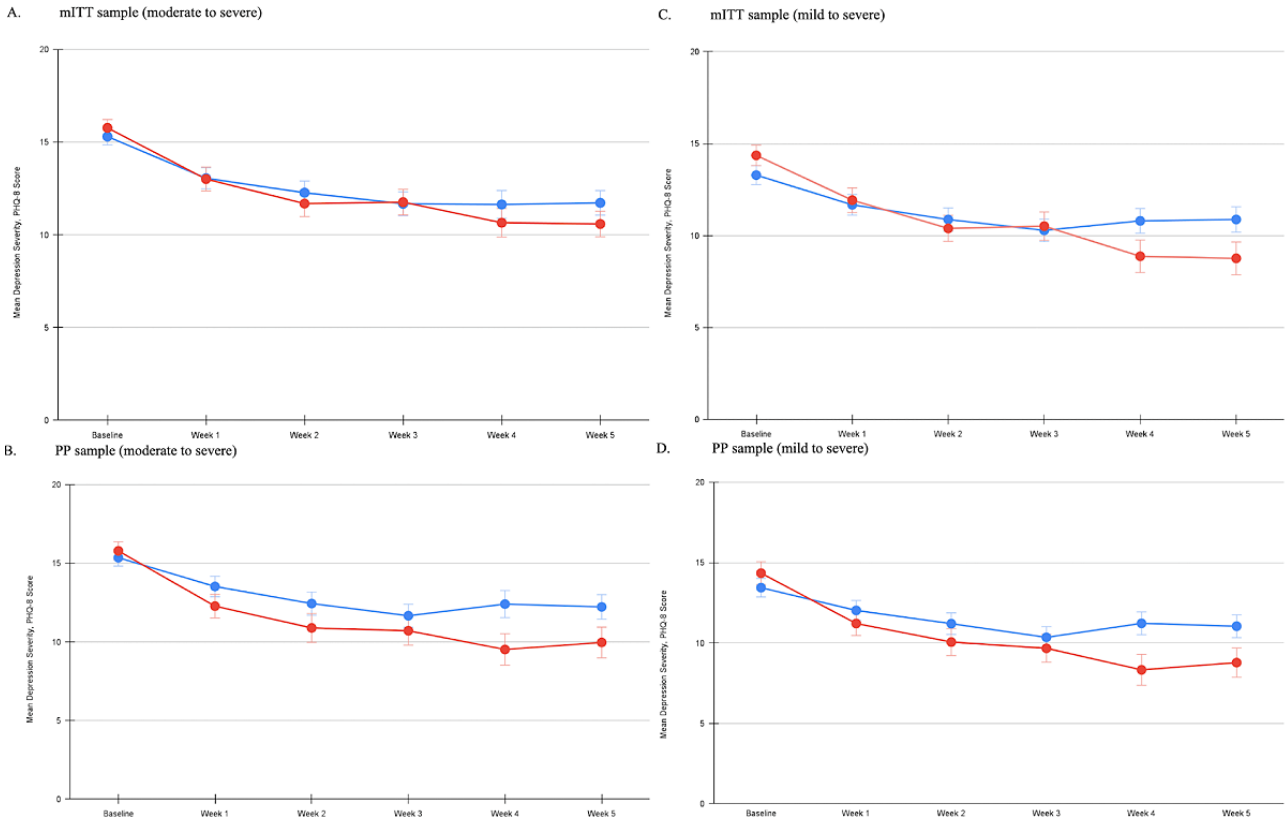

Per Protocol (n=86)

The LMM revealed a significant study arm x week interaction ($t_{91.59}$=−2.546; *P*=.01; FDR adjusted *P*=.02; f^2=0.0107; power: 0.699-0.699; Figure 2B). There was a nonsignificant effect of study arm ($t_{82.43}$=−0.072; *P*=.94; f^2=.00031; power: 0.041-0.056) and a significant effect of week (t_{1511}=−4.876; *P*<.001; f^2=0.00425; power: 0.995-0.997). There was a significant effect of week in the Spark arm ($t_{1183.54}$=−4.242; *P*<.001; change score=−5.82, 95% CI −7.54 to −4.10) and the control arm ($t_{532.3}$=−3.482; *P*<.001; change score=−3.13, 95% CI −4.49 to −1.76). The GLMM also produced a significant study arm x week interaction ($F_{1,6.569e+08}$=6.624; *P*=.01; FDR adjusted *P*=.01).

Mild-Severe Cohort

Modified Intention to Treat (n=153)

The LMM revealed a significant study arm x week interaction ($t_{160.20}$=−3.054; *P*=.003; FDR adjusted *P*=.005; f^2=0.00036; power: 0.846-0.884; Figure 2C). There was a nonsignificant effect of study arm ($t_{149.68}$=1.272; *P*=.21; f^2=.00157; power: 0.233-0.275) and a significant effect of week ($t_{2539.94}$=−3.06; *P*=.002; f^2=0.00058; power: 0.861-0.868). There was a significant effect of week in the Spark arm ($t_{1697.40}$=−3.32; *P*<.001; change score=−4.85, 95% CI −6.59 to −3.09) but not in the control arm ($t_{1177.40}$=−1.838; *P*=.07; change score=−2.86, 95% CI −4.43 to −1.30). The GLMM also produced a significant study arm x week interaction ($F_{1,19992.155}$=8.21; *P*=.004; FDR adjusted *P*=.008).

At the end of the intervention period, the Spark arm showed significantly higher remission (Spark, 20/74, 27% and control, 9/79, 11%; χ^2_1=6.1; *P*=.01; FDR adjusted *P*=.03) and treatment response (Spark, 22/74, 30% and control,12/79, 15%; χ^2_1=4.7; *P*=.03; FDR adjusted *P*=.04) rates compared with the control arm.

Per Protocol (n=109)

The LMM revealed a significant study arm x week interaction ($t_{115.09}$=−3.603; *P*<.001; FDR adjusted *P*=.002; f^2=0.009498; power: 0.941-0.949; Figure 2D). There was a nonsignificant effect of study arm ($t_{105.69}$=0.544; *P*=.59; f^2=.00006; power: 0.078-0.086) and a significant effect of week ($t_{1955.64}$=−3.6; *P*<.001; f^2=0.00169; power: 0.936~0.939). There was a significant effect of week in the Spark arm ($t_{1178.37}$=−4.049; *P*<.001; change score=−5.57, 95% CI −7.04 to −4.09) and the control arm ($t_{1011.58}$=−2.231; *P*=.03; change scores=−2.39, 95% CI −3.60 to −1.19). The GLMM also produced a significant study arm x week interaction ($F_{1,1.026e+09}$=13.205; *P*<.001; FDR adjusted *P*<.001).

Notably, all analyses that showed an uncorrected α threshold of *P*<.05 remained significant following FDR correction, and study arm×week interaction effects were consistent across LMM and GLMM analyses.

XSL•FO
RenderX

Little's test was significant for moderate-to-severe (χ^2_{34}=72.5; P<.001) and mild-to-severe cohorts (χ^2_{39}=106.94; P<.001), suggesting that data were not missing completely at random. Missing data patterns were associated with observed variables such as treatment group, week, and age group (Multimedia Appendix 1), suggesting that data are missing at random. No analysis showed a significant effect of Spark app version (significance P>.05).

Secondary Clinical Outcomes

Means, SDs, and 95% CIs of the difference between baseline and postintervention scores for all secondary participant-reported and legal guardian–reported measures are reported for each group in Table 4. The results for the effects of COVID-19 in the moderate-to-severe cohort are presented in Table 5. Notably, 49.2% (59/120) of the participants reported reduced access to mental health resources during the pandemic at baseline.

Table 4. Secondary clinical measures: participant and legal guardian (moderate-to-severe cohort; n=121).

	Spark participant scores			Control participant scores		
	Baseline, mean (SD)	Post, mean (SD)	Postbaseline, mean difference (95% CI)	Baseline, mean (SD)	Post, mean (SD)	Postbaseline, mean difference (95% CI)
Participant measures: Spark (n=43) and control (n=49)						
GAD-7[a]	12.26 (4.94)	9.05 (5.57)	−3.21 (−4.61 to −1.81)	12.31 (4.54)	10.86 (4.83)	−1.45 (−2.41 to −0.49)
PROMIS[b] 7+2– General Health	19.51 (3.13)	20.86 (4.01)	1.35 (0.30 to 2.40)	19.61 (3.34)	19.27 (3.64)	−0.35 (−0.88 to 0.19)
PROMIS 7+2–Fatigue	4.35 (0.90)	4.00 (1.11)	−0.35 (−0.66 to −0.03)	4.39 (0.84)	4.29 (0.91)	−0.10 (−0.34 to 0.13)
PROMIS 7+2–Pain	3.00 (1.21)	2.72 (1.39)	−0.28 (−0.74 to 0.18)	2.90 (1.29)	2.98 (1.31)	0.08 (−0.34 to 0.50)
BRS[c]	2.56 (0.66)	__[d]	—	2.52 (0.73)	—	—
Legal guardian measures: Spark (n=24) and control (n=28)						
MFQ-SF[e,f]	13.61 (4.71)	8.78 (4.99)	−4.83 (−6.69 to −2.96)	12.89 (5.62)	10.04 (4.43)	−2.85 (−4.49 to −1.21)
PROMIS 7+2 Parent Proxy–General Health	21.50 (2.28)	22.71 (3.93)	1.21 (−0.16 to 2.57)	21.04 (3.89)	22.07 (3.92)	1.04 (0.08 to 1.99)
PROMIS 7+2 Parent Proxy–Fatigue	3.96 (1.00)	3.63 (0.97)	−0.33 (−0.79 to 0.13)	3.89 (0.92)	3.21 (0.99)	−0.69 (−1.10 to −0.26)
PROMIS 7+2 Parent Proxy–Pain	2.96 (1.00)	2.88 (0.99)	−0.08 (−0.67 to 0.50)	2.61 (1.13)	2.39 (1.20)	−0.21 (−0.61 to 0.19)

[a]GAD-7: Generalized Anxiety Disorder-7.

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

[c]BRS: Brief Resilience Scale.

[d]Not available; data were not collected.

[e]MFQ-SF: Mood and Feelings Questionnaire-Short Form.

[f]A total of 2 legal guardians (1 in the Spark arm and 1 in the control arm) completed the PROMIS but not the Mood and Feelings Questionnaire.

XSL•FO
RenderX

Table 5. Positive and negative impacts of COVID-19—participant and legal guardian (moderate-to-severe cohort).

	Baseline: Spark, n (%)	Baseline: control, n (%)	Post: Spark, n (%)	Post: control, n (%)
Participant (moderate-to-severe cohort)				
Negative impacts of COVID-19				
Difficulty with schoolwork	49/62 (79)	48/58 (83)	29/43 (67)	36/49 (74)
Job loss or worry about employment for you or someone you care about	21/62 (34)	15/58 (26)	18/43 (42)	15/49 (31)
Loss of family member or someone you care about	8/62 (12)	10/58 (17)	5/43 (12)	8/49 (16)
Missed milestones (eg, prom or graduation) or meaningful holiday celebrations	45/62 (73)	36/58 (62)	28/43 (65)	31/49 (63)
None	1/62 (2)	1/58 (2)	3/43 (7)	0/49 (0)
Other	3/62 (5)	3/58 (55)	1/43 (25)	2/49 (45)
Personal illness	7/62 (115)	11/58 (19)	7/43 (16)	6/49 (12)
Reduced access to mental health resources or support	33/62 (53)	26/58 (45)	17/43 (40)	23/49 (47)
Serious illness of family member or someone you care about	10/62 (16)	10/58 (17)	11/43 (26)	10/49 (20)
Social isolation/loneliness	58/62 (94)	51/58 (88)	36/43 (84)	44/49 (90)
Worry about having enough food	6/62 (10)	5/58 (9)	5/43 (12)	4/49 (8)
Worry about housing	6/62 (10)	3/58 (5)	4/43 (9)	8/49 (16)
Worry about personal or family finances	27/62 (44)	23/58 (40)	19/43 (44)	25/49 (51)
Worry about physical violence in the home or local community	6/62 (10)	4/58 (7)	3/43 (7)	8/49 (16)
Positive impacts of COVID-19				
Increased in-person time with family or household members	21/63 (33)	26/58 (45)	22/44 (50)	24/49 (49)
Increased time for self-care (eg, exercise)	22/63 (35)	24/58 (41)	26/44 (59)	24/49 (49)
Increased virtual time with family or friends	13/63 (21)	13/58 (22)	12/44 (27)	9/49 (18)
None	18/63 (29)	12/58 (21)	5/44 (11)	8/49 (16)
Other	1/68 (2)	3/58 (5)	1/44 (2)	0/49 (0)
Reduced academic stress	8/63 (13)	9/58 (16)	12/44 (27)	13/49 (27)
Reduced social stress	23/63 (37)	20/58 (34)	22/44 (50)	17/49 (35)
Legal guardian (moderate-to-severe cohort)				
Negative impacts of COVID-19				
Difficulty with schoolwork	23/33 (70)	24/35 (69)	16/24 (67)	19/28 (68)
Job loss or worry about employment for your child or someone your child cares about	3/33 (9)	3/35 (9)	3/24 (13)	3/28 (111)
Loss of family member or someone your child cares about	3/33 (9)	2/35 (6)	2/24 (8)	2/28 (7)
Missed milestones (eg, prom or graduation) or meaningful holiday celebrations	18/33 (55)	23/35 (66)	7/24 (29)	13/28 (46)
None	0/33 (0)	0/35 (0)	3/24 (13)	1/28 (4)
Other	6/33 (18)	8/35 (23)	2/24 (8)	1/28 (4)
Personal illness	2/33 (6)	4/35 (11)	2/24 (8)	2/28 (7)
Reduced access to mental health resources or support	14/33 (42)	16/35 (46)	9/24 (38)	10/28 (36)
Serious illness of family member or someone your child cares about	4/33 (12)	6/35 (17)	4/24 (17)	2/28 (7)
Social isolation, loneliness, or both	33/33 (100)	35/35 (100)	20/24 (83)	24/28 (86)
Worry about having enough food	2/33 (6)	1/35 (3)	0/24 (0)	2/28 (7)
Worry about housing	1/33 (3)	1/35 (3)	1/24 (4)	2/28 (7)
Worry about personal or family finances	8/33 (24)	14/35 (40)	6/24 (25)	10/28 (36)
Worry about physical violence in the home or local community	2/33 (6)	2/35 (6)	1/24 (4)	3/28 (11)
Positive impacts of COVID-19				

XSL•FO
RenderX

	Baseline: Spark, n (%)	Baseline: control, n (%)	Post: Spark, n (%)	Post: control, n (%)
Increased in-person time with family or household members	24/31 (77)	22/35 (63)	16/24 (67)	18/28 (64)
Increased time for self-care (eg, exercise)	9/31 (29)	13/35 (37)	9/24 (38)	6/28 (21)
Increased internet-based time with family or friends	6/31 (19)	14/35 (40)	6/24 (25)	6/38 (21)
None	2/31 (6)	5/35 (14)	6/24 (25)	5/28 (18)
Other	1/31 (3)	1/35 (3)	0/24 (0)	1/28 (4)
Reduced academic stress	8/31 (26)	6/35 (17)	7/24 (29)	8/28 (29)
Reduced social stress	9/31 (29)	15/35 (43)	4/24 (17)	8/28 (29)

Safety Outcomes

There were no SAEs, ADEs, or UADEs in either group. There were a total of 2 AEs in the control arm and 4 in the Spark arm. Both AEs in the control arm were clinical deteriorations in PHQ-8 scores for 2 consecutive weeks. In the Spark arm, there was 1 clinical deterioration in the PHQ-8 score for 2 consecutive weeks, 2 instances of participant-reported suicidal ideation, and 1 instance of a legal guardian reporting that the participant was experiencing nonsuicidal self-injury ideation.

Adherence, Acceptability, and Engagement

In the Spark arm, 87% (55/63) of participants completed module 1, 79% (50/63) completed module 2, 64% (40/64) completed module 3, 48% (30/63) completed module 4, and 40% (25/63) completed module 5. App acceptability metrics are reported in Table 6. Engagement with the app as defined as the average number of sessions are reported in Table 7.

Table 6. Spark and control app acceptability ratings: participant and legal guardian (moderate-to-severe cohort)

	Participant ratings			Legal guardian ratings		
	Spark (n=44), mean (SD)	Control (n=49), mean (SD)	Mean difference (95% CI)	Spark (n=28), mean (SD)	Control (n=24), mean (SD)	Mean difference (95% CI)
Mood or depression symptom improvement	5.20 (2.04)	2.51 (2.41)	2.69 (1.76 to 3.62)	4.00 (2.17)	3.75 (2.47)	0.25 (−1.04 to 1.54)
Enjoyable	7.75 (2.01)	5.24 (2.72)	2.51 (1.52 to 3.50)	5.63 (2.18)	4.86 (2.09)	0.77 (−0.43 to.197)
SUS[a]	82.85 (12.42)	76.12 (13.47)	6.73 (1.02 to 11.74)	N/A[b]	N/A	N/A
UES-SF[c]	3.81 (0.53)	3.22 (0.65)	0.59 (0.34 to 0.84)	N/A	N/A	N/A
UES-SF: focused attention	2.77 (0.89)	2.29 (0.78)	0.48 (0.14 to 0.82)	N/A	N/A	N/A
UES-SF: perceived usability	4.19 (0.77)	4.14 (0.89)	0.05 (−0.29 to.39)	N/A	N/A	N/A
UES-SF: esthetic appeal	4.30 (0.63)	3.03 (0.89)	1.27 (0.95 to 1.59)	N/A	N/A	N/A
UES-SF: reward	3.99 (0.84)	3.43 (0.95)	0.56 (0.19 to 0.93)	N/A	N/A	N/A

[a]SUS: systems usability scale.

[b]N/A: not applicable.

[c]UES-SF: User Engagement Scale-short form.

Table 7. Weekly Spark and control number of app sessions.

	Spark participant app sessions, mean (SD)	Control participant app sessions, mean (SD)
Week 1	7.57 (4.71)	3.53 (2.22)
Week 2	4.78 (3.84)	2.64 (1.62)
Week 3	4.19 (2.97)	2.40 (1.50)
Week 4	3.60 (2.94)	2.40 (1.63)
Week 5	3.46 (3.42)	2.33 (1.82)
Total	23.60 (13.75)	13.29 (6.74)

XSL•FO
RenderX

Discussion

Principal Findings

This RCT showed that, in planned and exploratory analyses, Spark was efficacious in treating mild to severe symptoms of depression, including a clinically meaningful reduction in depression symptoms and a significantly larger reduction in depression symptoms as well as higher remission and treatment response rates at postintervention compared with an active control. Spark did not negatively impact participant safety, as demonstrated by having no ADEs reported A mITT analysis, restricted to participants with moderate to severe symptoms at baseline, showed no group differences in symptom change from baseline to postintervention. The treatment response and remission rates for the Spark arm were consistent with those found in other digital CBT apps [50], and the Spark arm showed significantly higher remission rates compared with the control arm. A follow-up PP analysis restricted to participants who consistently engaged with their assigned app per the study protocol revealed that the Spark arm showed a significantly larger decrease in depression symptoms compared with the control arm. This finding supports the idea that engagement and adherence are important drivers of treatment outcomes [82,83]. Indeed, engagement with and adherence to Spark were also high, compared with other digital therapeutic studies, and higher than most popular wellness apps [82-86]. Participants also rated Spark as moderately enjoyable, usable, and successful in improving mood and depression symptoms [87]. Although out of scope for this study, future research should directly investigate engagement in clinical outcomes, especially in the context of specific therapeutic components of the intervention, such as mood logs and BAs, and how they may mediate symptom change.

Notably, exploratory analyses including a larger cohort of participants with mild to severe symptoms at baseline showed significant group differences in symptom reduction, remission rate, and treatment response rate. The Spark arm within this cohort also showed an MCID within a 95% CI. Statistical tests were robust to correction for multiple comparisons and suggest that Spark may be an effective treatment for a wide range of depression severity, including those with mild symptoms, consistent with a recent meta-analysis of digital CBT [25]. As mild symptoms are often overlooked and, when untreated, can worsen or have long-term consequences [88], self-guided DTx may be a particularly attractive treatment option for this severity group.

The results also suggest that Spark may be promising as an adjunct treatment and part of a stepped care model for depression treatment in adolescents [89]. Participants were able to continue their usual care while enrolled in the study, and more than three quarters (130/160, 81.3%) of the participants reported having been diagnosed with depression and about half were receiving some form of concurrent treatment for depression (medication, 85/160, 53.1% or psychotherapy, 72/160, 45%). Adjunct digital treatments can reduce the time and cost of standard care [84,90], and as a self-contained treatment option, Spark can be easily integrated into a wide range of treatment plans.

These results are also promising in light of issues with access to mental health care for adolescents, many of which were magnified during the COVID-19 pandemic along with an increase in prevalence rates [1,17-19,21]. Approximately 48.8% (59/121) of the participants identified reduced access to mental health resources during the pandemic, underscoring the treatment potential of self-guided evidence-based DTx, such as Spark, which can be made available immediately, widely, and equitably, when other forms of care are limited.

Limitations

The study was not statistically powered to detect group differences in PHQ-8 symptom change; therefore, these results should be replicated in a fully powered sample. Powered analyses are also warranted to directly investigate the impacts of engagement, baseline symptom severity, and concurrent treatments on outcomes. Although an active control was used, the open-label design of the study is a limitation, as participants were not blinded to treatment assignment and the primary outcome was a self-report PHQ-8; blinded studies with clinician-reported outcomes will provide stronger evidence of efficacy. The exploratory results including adolescents with mild to severe symptoms at baseline, although promising, should be interpreted with caution and replicated. Similarly, although our PP analysis indicates that engagement with Spark may drive symptom improvement, these observations may be owing to survivorship bias. The eligibility criteria for this study were intentionally designed to allow adolescents who were seeking resources during the pandemic to participate and did not require stability on concurrent treatment, an assessment for treatment resistance, or stratification of study arms based on concurrent treatment, which may have introduced noise into the data.

Generalizability

Spark was designed as a brief (5 weeks), adjunct to standard of care intervention to reduce symptoms of depression in adolescents (aged 13-21 years as defined by the FDA for medical devices). As such, the results should be interpreted for an adolescent sample; we would not anticipate these results to generalize to the broader adult and child population. We recruited participants directly (ie, through social media advertising rather than through a health care provider), which could have created a self-selection bias. To evaluate the generalizability of these results in a naturalistic setting, future studies should evaluate the efficacy of Spark as a treatment for depression in patients who are introduced to Spark by a health care provider. Future studies should evaluate long-term outcomes and future product developments (eg, evergreen features).

Conclusions

DTx such as Spark could help overcome known barriers to access to care [14,16,91,92] by providing an immediate treatment option and serve as an effective adjunct form of care. DTx can also serve to reduce stigma around mental health treatment, as interventions can be delivered privately, and can improve treatment adherence by providing an engaging

XSL•FO
RenderX

experience. This study provides preliminary evidence that a brief, self-guided, CBT-based DTx can be used to effectively and safely treat mild to severe depression symptoms in adolescents.

Acknowledgments

This study was supported with funding from Limbix Health and a Small Business Innovation Research grant (R44MH12563) from the National Institutes of Health. Limbix Health designed and conducted the study, conducted some data analysis, and prepared and submitted the manuscript for publication. This study was sponsored by Limbix Health. Limbix Health was acquired by Big Health in July 2023. EP, JL, and AP were affiliated with Limbix Health at the time of the study and are currently affiliated with Big Health. IM and JF were affiliated with Limbix Health at the time of the study. The authors thank Vera Kulikov, BA, Phoebe Crosthwaite, BA, and Lauren Smith, BS, for conducting participant recruitment, enrollment, data acquisition, monitoring, and analysis; Judy Liu, BS, Josh Klag, MS, Jason Brownstein, BA, Matthew Frisby, BS, and Brodie Polusny Thiede, BA for database development, technical support, and data monitoring; and Shana Hall, PhD for feedback on the manuscript. The authors also thank Raph Rose, PhD for serving as an independent study clinician for safety monitoring and adverse event classification.

Authors' Contributions

JL and AP contributed to the concept and design of the study. AP, JL, JF, EP, IM, and LC contributed to the acquisition, analysis, or interpretation of data. AP, EP, and IM drafted the manuscript. JL and JF provided critical revision of the manuscript for important intellectual content. LC and EP conducted statistical analysis. AP provided supervision of all study and manuscript completion.

Conflicts of Interest

All authors except LC are salaried employees and stakeholders in Limbix Health.

Multimedia Appendix 1
Missing data analysis by cohort.
[DOCX File , 8 KB - jmir_v26i1e48467_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1174 KB - jmir_v26i1e48467_app2.pdf]

References

1. Data and statistics on children's mental health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/childrensmentalhealth/data.html [accessed 2022-11-18]
2. Mojtabai R, Olfson M, Han B. National trends in the prevalence and treatment of depression in adolescents and young adults. Pediatrics 2016 Dec;138(6):e20161878 [FREE Full text] [doi: 10.1542/peds.2016-1878] [Medline: 27940701]
3. Keyes KM, Gary D, O'Malley PM, Hamilton A, Schulenberg J. Recent increases in depressive symptoms among US adolescents: trends from 1991 to 2018. Soc Psychiatry Psychiatr Epidemiol 2019 Aug;54(8):987-996 [FREE Full text] [doi: 10.1007/s00127-019-01697-8] [Medline: 30929042]
4. Mental illness. National Institute of Mental Health (NIMH). URL: https://www.nimh.nih.gov/health/statistics/mental-illness [accessed 2022-11-18]
5. Zisook S, Lesser I, Stewart JW, Wisniewski SR, Balasubramani GK, Fava M, et al. Effect of age at onset on the course of major depressive disorder. Am J Psychiatry 2007 Oct;164(10):1539-1546. [doi: 10.1176/appi.ajp.2007.06101757] [Medline: 17898345]
6. Alegría M, Green JG, Mclaughlin KA, Loder S. Disparities in child and adolescent mental health and mental health services in the U.S. William T. Grant Foundation. 2015. URL: http://cfs.cbcs.usf.edu/projects-research/_docs/Disparities_in_child_and_adolescent_health.pdf [accessed 2023-11-02]
7. Das JK, Salam RA, Lassi ZS, Khan MN, Mahmood W, Patel V, et al. Interventions for adolescent mental health: an overview of systematic reviews. J Adolesc Health 2016 Oct;59(4S):S49-S60 [FREE Full text] [doi: 10.1016/j.jadohealth.2016.06.020] [Medline: 27664596]
8. Edgcomb JB, Zima B. Medication adherence among children and adolescents with severe mental illness: a systematic review and meta-analysis. J Child Adolesc Psychopharmacol 2018 Oct;28(8):508-520. [doi: 10.1089/cap.2018.0040] [Medline: 30040434]
9. Gearing RE, Schwalbe CS, Short KD. Adolescent adherence to psychosocial treatment: mental health clinicians' perspectives on barriers and promoters. Psychother Res 2012;22(3):317-326. [doi: 10.1080/10503307.2011.653996] [Medline: 22313513]

XSL•FO
RenderX

10. Rickwood D, Deane FP, Wilson CJ, Ciarrochi J. Young people's help-seeking for mental health problems. Adv Ment Health 2005 Dec;4(3):218-251 [FREE Full text] [doi: 10.5172/jamh.4.3.218]
11. Mehra K, Hawke LD, Watson P, Sheikhan NY, Leroux E, Henderson J. Youth perspectives on seeking psychotherapy: a concurrent mixed methods study. J Can Acad Child Adolesc Psychiatry 2021 Aug;30(3):165-176 [FREE Full text] [Medline: 34381509]
12. Westin AM, Barksdale CL, Stephan SH. The effect of waiting time on youth engagement to evidence based treatments. Community Ment Health J 2014 Feb;50(2):221-228. [doi: 10.1007/s10597-012-9585-z] [Medline: 23283487]
13. Bodden DH, Stikkelbroek Y, Dirksen CD. Societal burden of adolescent depression, an overview and cost-of-illness study. J Affect Disord 2018 Dec 01;241:256-262. [doi: 10.1016/j.jad.2018.06.015] [Medline: 30138810]
14. Sheppard R, Deane FP, Ciarrochi J. Unmet need for professional mental health care among adolescents with high psychological distress. Aust N Z J Psychiatry 2018 Jan;52(1):59-67. [doi: 10.1177/0004867417707818] [Medline: 28486819]
15. El-Amin T, Anderson BL, Leider JP, Satorius J, Knudson A. Enhancing mental health literacy in rural America: growth of mental health first aid program in rural communities in the united states from 2008–2016. J Rural Ment Health 2018 Jan;42(1):20-31 [FREE Full text] [doi: 10.1037/rmh0000088]
16. Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. Health Aff (Millwood) 2009 May;28(3):w490-w501. [doi: 10.1377/hlthaff.28.3.w490] [Medline: 19366722]
17. Murthy VH. Protecting youth mental health: the U.S. surgeon general's advisory. US Department of Health and Human Services. 2021. URL: https://www.hhs.gov/sites/default/files/surgeon-general-youth-mental-health-advisory.pdf [accessed 2023-11-02]
18. Racine N, Cooke JE, Eirich R, Korczak DJ, McArthur B, Madigan S. Child and adolescent mental illness during COVID-19: a rapid review. Psychiatry Res 2020 Oct;292:113307 [FREE Full text] [doi: 10.1016/j.psychres.2020.113307] [Medline: 32707216]
19. Zhang L, Zhang D, Fang J, Wan Y, Tao F, Sun Y. Assessment of mental health of Chinese primary school students before and after school closing and opening during the COVID-19 pandemic. JAMA Netw Open 2020 Sep 01;3(9):e2021482 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.21482] [Medline: 32915233]
20. Stinson EA, Sullivan RM, Peteet BJ, Tapert SF, Baker FC, Breslin FJ, et al. Longitudinal impact of childhood adversity on early adolescent mental health during the COVID-19 pandemic in the ABCD study cohort: does race or ethnicity moderate findings? Biol Psychiatry Glob Open Sci 2021 Dec;1(4):324-335 [FREE Full text] [doi: 10.1016/j.bpsgos.2021.08.007] [Medline: 34608463]
21. Panchal U, Salazar de Pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. Eur Child Adolesc Psychiatry 2023 Jul;32(7):1151-1177 [FREE Full text] [doi: 10.1007/s00787-021-01856-w] [Medline: 34406494]
22. Vigerland S, Lenhard F, Bonnert M, Lalouni M, Hedman E, Ahlen J, et al. Internet-delivered cognitive behavior therapy for children and adolescents: a systematic review and meta-analysis. Clin Psychol Rev 2016 Dec;50:1-10 [FREE Full text] [doi: 10.1016/j.cpr.2016.09.005] [Medline: 27668988]
23. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
24. Martinengo L, Stona AC, Griva K, Dazzan P, Pariante CM, von Wangenheim F, et al. Self-guided cognitive behavioral therapy apps for depression: systematic assessment of features, functionality, and congruence with evidence. J Med Internet Res 2021 Jul 30;23(7):e27619 [FREE Full text] [doi: 10.2196/27619] [Medline: 34328431]
25. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
26. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002 [FREE Full text] [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
27. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: a systematic review of randomised controlled trials. Child Youth Serv Rev 2021 Aug;127:106073 [FREE Full text] [doi: 10.1016/j.childyouth.2021.106073]
28. Gräfe V, Greiner W. Internet based treatment of depressive symptoms – a health economic evaluation of costs and benefits. Value Health 2017 Oct;20(9):A714 [FREE Full text] [doi: 10.1016/j.jval.2017.08.1893]
29. Hedman E, Ljótsson B, Lindefors N. Cognitive behavior therapy via the Internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Rev Pharmacoecon Outcomes Res 2012 Dec;12(6):745-764. [doi: 10.1586/erp.12.67] [Medline: 23252357]
30. DTA DTx definition and core principles. Digital Therapeutics Alliance. URL: https://dtxalliance.org/wp-content/uploads/2021/01/DTA_DTx-Definition-and-Core-Principles.pdf [accessed 2023-11-02]

XSL•FO
RenderX

31. Topooco N, Berg M, Johansson S, Liljethörn L, Radvogin E, Vlaescu G, et al. Chat- and internet-based cognitive-behavioural therapy in treatment of adolescent depression: randomised controlled trial. BJPsych Open 2018 Jul;4(4):199-207 [FREE Full text] [doi: 10.1192/bjo.2018.18] [Medline: 29988969]
32. Wannachaiyakul S, Thapinta D, Sethabouppha H, Thungjaroenkul P, Likhitsathian S. Randomized controlled trial of computerized cognitive behavioral therapy program for adolescent offenders with depression. Pac Rim Int J Nurs Res Thail 2017:32-43 [FREE Full text]
33. MacDonell KW, Prinz RJ. A review of technology-based youth and family-focused interventions. Clin Child Fam Psychol Rev 2017 Jun;20(2):185-200 [FREE Full text] [doi: 10.1007/s10567-016-0218-x] [Medline: 27787701]
34. Keles S, Bringedal G, Idsoe T. Assessing fidelity to and satisfaction with the “Adolescent Coping with Depression Course” (ACDC) intervention in a randomized controlled trial. J Ration Emot Cogn Behav Ther 2021 Nov 02;40(3):583-602 [FREE Full text] [doi: 10.1007/s10942-021-00427-x]
35. Brezing CA, Brixner DI. The rise of prescription digital therapeutics in behavioral health. Adv Ther 2022 Dec;39(12):5301-5306 [FREE Full text] [doi: 10.1007/s12325-022-02320-0] [Medline: 36242730]
36. Thase ME, McCrone P, Barrett MS, Eells TD, Wisniewski SR, Balasubramani GK, et al. Improving cost-effectiveness and access to cognitive behavior therapy for depression: providing remote-ready, computer-assisted psychotherapy in times of crisis and beyond. Psychother Psychosom 2020;89(5):307-313 [FREE Full text] [doi: 10.1159/000508143] [Medline: 32396917]
37. Oh S, Choi J, Han DH, Kim E. Effects of game-based digital therapeutics on attention deficit hyperactivity disorder in children and adolescents as assessed by parents or teachers: a systematic review and meta-analysis. Eur Child Adolesc Psychiatry (Forthcoming) 2023 Mar 02. [doi: 10.1007/s00787-023-02174-z] [Medline: 36862162]
38. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual research review: digital health interventions for children and young people with mental health problems - a systematic and meta-review. J Child Psychol Psychiatry 2017 Apr;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
39. Kollins SH, Childress A, Heusser AC, Lutz J. Effectiveness of a digital therapeutic as adjunct to treatment with medication in pediatric ADHD. NPJ Digit Med 2021 Mar 26;4(1):58 [FREE Full text] [doi: 10.1038/s41746-021-00429-0] [Medline: 33772095]
40. Maricich YA, Bickel WK, Marsch LA, Gatchalian K, Botbyl J, Luderer HF. Safety and efficacy of a prescription digital therapeutic as an adjunct to buprenorphine for treatment of opioid use disorder. Curr Med Res Opin 2021 Feb;37(2):167-173 [FREE Full text] [doi: 10.1080/03007995.2020.1846022] [Medline: 33140994]
41. Emslie G, Kratochvil C, Vitiello B, Silva S, Mayes T, McNulty S, Columbia Suicidality Classification Group, TADS Team. Treatment for Adolescents with Depression Study (TADS): safety results. J Am Acad Child Adolesc Psychiatry 2006 Dec;45(12):1440-1455 [FREE Full text] [doi: 10.1097/01.chi.0000240840.63737.1d] [Medline: 17135989]
42. Treatment for Adolescents With Depression Study Team. Treatment for Adolescents With Depression Study (TADS): rationale, design, and methods. J Am Acad Child Adolesc Psychiatry 2003 May;42(5):531-542 [FREE Full text] [doi: 10.1097/01.CHI.0000046839.90931.0D] [Medline: 12707557]
43. Cheung AH, Zuckerbrot RA, Jensen PS, Laraque D, Stein RE, GLAD-PC Steering Group. Guidelines for adolescent depression in primary care (GLAD-PC): part II. Treatment and ongoing management. Pediatrics 2018 Mar;141(3):e20174082. [doi: 10.1542/peds.2017-4082] [Medline: 29483201]
44. McCauley E, Gudmundsen G, Schloredt K, Martell C, Rhew I, Hubley S, et al. The adolescent behavioral activation program: adapting behavioral activation as a treatment for depression in adolescence. J Clin Child Adolesc Psychol 2016;45(3):291-304 [FREE Full text] [doi: 10.1080/15374416.2014.979933] [Medline: 25602170]
45. Pass L, Lejuez CW, Reynolds S. Brief behavioural activation (brief BA) for adolescent depression: a pilot study. Behav Cogn Psychother 2018 Mar;46(2):182-194. [doi: 10.1017/S1352465817000443] [Medline: 28756787]
46. Lejuez CW, Hopko DR, Acierno R, Daughters SB, Pagoto SL. Ten year revision of the brief behavioral activation treatment for depression: revised treatment manual. Behav Modif 2011 Mar;35(2):111-161. [doi: 10.1177/0145445510390929] [Medline: 21324944]
47. Kanter JW, Manos RC, Bowe WM, Baruch DE, Busch AM, Rusch LC. What is behavioral activation? A review of the empirical literature. Clin Psychol Rev 2010 Aug;30(6):608-620. [doi: 10.1016/j.cpr.2010.04.001] [Medline: 20677369]
48. Clarke G, DeBar LL, Pearson JA, Dickerson JF, Lynch FL, Gullion CM, et al. Cognitive behavioral therapy in primary care for youth declining antidepressants: a randomized trial. Pediatrics 2016 May;137(5):e20151851 [FREE Full text] [doi: 10.1542/peds.2015-1851] [Medline: 27244782]
49. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022 Sep 15;5(1):144 [FREE Full text] [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
50. Forman-Hoffman VL, Nelson BW, Ranta K, Nazander A, Hilgert O, de Quevedo J. Significant reduction in depressive symptoms among patients with moderately-severe to severe depressive symptoms after participation in a therapist-supported, evidence-based mobile health program delivered via a smartphone app. Internet Interv 2021 Jun 17;25:100408 [FREE Full text] [doi: 10.1016/j.invent.2021.100408] [Medline: 34401367]

XSL•FO
RenderX

51. Whittaker R, Stasiak K, McDowell H, Doherty I, Shepherd M, Chua S, et al. MEMO: an mHealth intervention to prevent the onset of depression in adolescents: a double-blind, randomised, placebo-controlled trial. J Child Psychol Psychiatry 2017 Sep;58(9):1014-1022. [doi: 10.1111/jcpp.12753] [Medline: 28573672]
52. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
53. Kauer SD, Reid SC, Crooke AH, Khor A, Hearps SJ, Jorm AF, et al. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. J Med Internet Res 2012 Jun 25;14(3):e67 [FREE Full text] [doi: 10.2196/jmir.1858] [Medline: 22732135]
54. Kulikov VN, Crosthwaite PC, Hall SA, Flannery JE, Strauss GS, Vierra EM, et al. A CBT-based mobile intervention as an adjunct treatment for adolescents with symptoms of depression: a virtual randomized controlled feasibility trial. Front Digit Health 2023 May 23;5:1062471 [FREE Full text] [doi: 10.3389/fdgth.2023.1062471] [Medline: 37323125]
55. Home page. Stanley-Brown: Safety Planning Intervention. 2018. URL: https://suicidesafetyplan.com/ [accessed 2023-11-02]
56. Glen S. Permuted block randomization. Statistics How To. 2016. URL: https://www.statisticshowto.com/permuted-block-randomization/ [accessed 2023-11-02]
57. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
58. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
59. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
60. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880 [FREE Full text] [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
61. Eg J, Bilenberg N, Costello EJ, Wesselhoeft R. Self- and parent-reported depressive symptoms rated by the mood and feelings questionnaire. Psychiatry Res 2018 Oct;268:419-425. [doi: 10.1016/j.psychres.2018.07.016] [Medline: 30130708]
62. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
63. O'Brien HL, Cairns P, Hall M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. Int J Hum Comput Stud 2018 Apr;112:28-39 [FREE Full text] [doi: 10.1016/j.ijhcs.2018.01.004]
64. ASQ Screening Tool. National Institute of Mental Health (NIMH). URL: https://www.nimh.nih.gov/research/research-conducted-at-nimh/asq-toolkit-materials/asq-tool/asq-screening-tool [accessed 2023-01-30]
65. CFR - code of federal regulations. US Department of Food & Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=312.32 [accessed 2022-09-21]
66. Office of the Commissioner. What is a serious adverse event? US Food and Drug Administration. URL: https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event [accessed 2022-09-21]
67. Clinical investigation of medical devices for human subjects — good clinical practice. International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:14155:ed-3:v1:en [accessed 2022-09-21]
68. R Core Team. The R project for statistical computing. The R Foundation. URL: https://www.r-project.org/ [accessed 2023-01-23]
69. Little RJ. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202 [FREE Full text] [doi: 10.1080/01621459.1988.10478722]
70. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393 [FREE Full text] [doi: 10.1136/bmj.b2393] [Medline: 19564179]
71. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. J Comput Graph Stat 2002 Jun;11(2):437-457 [FREE Full text] [doi: 10.1198/106186002760180608]
72. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1-67 [FREE Full text] [doi: 10.18637/jss.v045.i03]
73. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011 Mar;20(1):40-49 [FREE Full text] [doi: 10.1002/mpr.329] [Medline: 21499542]
74. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48 [FREE Full text] [doi: 10.18637/jss.v067.i01]
75. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol 2012 Dec 03;4(2):133-142 [FREE Full text] [doi: 10.1111/j.2041-210x.2012.00261.x]
76. Nakagawa S, Johnson PC, Schielzeth H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J R Soc Interface 2017 Sep;14(134):20170213 [FREE Full text] [doi: 10.1098/rsif.2017.0213] [Medline: 28904005]

XSL•FO
RenderX

77. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York, NY: Springer; 2002.
78. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26 [FREE Full text] [doi: 10.18637/jss.v082.i13]
79. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods 2017 Aug;49(4):1494-1502. [doi: 10.3758/s13428-016-0809-y] [Medline: 27620283]
80. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989 Dec;10(4):407-415. [doi: 10.1016/0197-2456(89)90005-6] [Medline: 2691207]
81. Coley RY, Boggs JM, Beck A, Hartzler AL, Simon GE. Defining success in measurement-based care for depression: a comparison of common metrics. Psychiatr Serv 2020 Apr 01;71(4):312-318. [doi: 10.1176/appi.ps.201900295] [Medline: 31847739]
82. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
83. Wu A, Scult MA, Barnes ED, Betancourt JA, Falk A, Gunning FM. Smartphone apps for depression and anxiety: a systematic review and meta-analysis of techniques to increase engagement. NPJ Digit Med 2021 Feb 11;4(1):20 [FREE Full text] [doi: 10.1038/s41746-021-00386-8] [Medline: 33574573]
84. Richards D, Enrique A, Eilert N, Franklin M, Palacios J, Duffy D, et al. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. NPJ Digit Med 2020 Jun 15;3:85 [FREE Full text] [doi: 10.1038/s41746-020-0293-8] [Medline: 32566763]
85. Ritterband LM, Thorndike FP, Ingersoll KS, Lord HR, Gonder-Frederick L, Frederick C, et al. Effect of a web-based cognitive behavior therapy for insomnia intervention with 1-year follow-up: a randomized clinical trial. JAMA Psychiatry 2017 Jan 01;74(1):68-75. [doi: 10.1001/jamapsychiatry.2016.3249] [Medline: 27902836]
86. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
87. Sauro J. Practical guide to the system usability scale background, benchmarks, and best practices. Scientific Research Publishing. 2011. URL: https://scirp.org/reference/referencespapers.aspx?referenceid=3273925 [accessed 2022-12-20]
88. Noyes BK, Munoz DP, Khalid-Khan S, Brietzke E, Booij L. Is subthreshold depression in adolescence clinically relevant? J Affect Disord 2022 Jul 15;309:123-130. [doi: 10.1016/j.jad.2022.04.067] [Medline: 35429521]
89. van Straten A, Seekles W, van 't Veer-Tazelaar NJ, Beekman AT, Cuijpers P. Stepped care for depression in primary care: what should be offered and how? Med J Aust 2010 Jun 07;192(S11):S36-S39. [doi: 10.5694/j.1326-5377.2010.tb03691.x] [Medline: 20528706]
90. Thase ME, Wright JH, Eells TD, Barrett MS, Wisniewski SR, Balasubramani GK, et al. Improving the efficiency of psychotherapy for depression: computer-assisted versus standard CBT. Am J Psychiatry 2018 Mar 01;175(3):242-250 [FREE Full text] [doi: 10.1176/appi.ajp.2017.17010089] [Medline: 28969439]
91. Williams ME, Latta J, Conversano P. Eliminating the wait for mental health services. J Behav Health Serv Res 2008 Jan;35(1):107-114. [doi: 10.1007/s11414-007-9091-1] [Medline: 17975730]
92. Kowalewski K, McLennan JD, McGrath PJ. A preliminary investigation of wait times for child and adolescent mental health services in Canada. J Can Acad Child Adolesc Psychiatry 2011 May;20(2):112-119 [FREE Full text] [Medline: 21541100]

Abbreviations

ADE: adverse device effect
AE: adverse event
BA: behavioral activation
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
DTx: digital therapeutics
FDA: US Food and Drug Administration
FDR: false discovery rate
GLMM: generalized linear mixed-effects model
LMM: linear mixed-effects model
MCID: minimal clinically important difference
mITT: modified intention-to-treat
PHQ-8: Patient Health Questionnaire-8
PP: per protocol
RCT: randomized controlled trial
SAE: serious adverse event

XSL•FO
RenderX

UADE: unanticipated adverse device effect

Edited by A Mavragani; submitted 26.04.23; peer-reviewed by N Mungoli, S Zhao; comments to author 20.05.23; revised version received 28.06.23; accepted 25.10.23; published 07.02.24.

Please cite as:
Peake E, Miller I, Flannery J, Chen L, Lake J, Padmanabhan A
Preliminary Efficacy of a Digital Intervention for Adolescent Depression: Randomized Controlled Trial
J Med Internet Res 2024;26:e48467
URL: https://www.jmir.org/2024/1/e48467
doi:10.2196/48467
PMID:38324367

©Emily Peake, Ian Miller, Jessica Flannery, Lang Chen, Jessica Lake, Aarthi Padmanabhan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Influence of Patient Empowerment Gained Through Mental Health Apps on Patient Trust in the Health Care Provider and Patient Compliance With the Recommended Treatment: Cross-sectional Study

Julien François[1], MSc; Anne-Françoise Audrain-Pontevia[1], PhD; Sana Boudhraâ[2,3], PhD; Stéphane Vial[2,3], PhD

[1]École des Sciences de la Gestion, Université du Québec à Montréal, Montréal, QC, Canada

[2]Centre de Recherche de l'Institut Universitaire en Santé Mentale de Montréal, Montréal, QC, Canada

[3]École de Design, Université du Québec à Montréal, Montréal, QC, Canada

Corresponding Author:
Julien François, MSc
École des Sciences de la Gestion
Université du Québec à Montréal
315 rue Sainte Catherine Est
Montréal, QC, H2X 3X2
Canada
Phone: 1 514 987 3000 ext 3656
Email: francois.julien@uqam.ca

Abstract

Background: In chronic mental illness, noncompliance with treatment significantly worsens the illness course and outcomes for patients. Considering that nearly 1 billion people worldwide experience mental health issues, including 1 of 5 Canadians in any given year, finding tools to lower noncompliance in these populations is critical for health care systems. A promising avenue is apps that make mental health services more accessible to patients. However, little is known regarding the impact of the empowerment gained from mental health apps on patient compliance with recommended treatment.

Objective: This study aimed to investigate the impact of patient empowerment gained through mental health apps on patient trust in the health care provider and patient compliance with the recommended treatment.

Methods: A cross-sectional web-based survey was conducted in Canada. Eligible participants were Canadian adults diagnosed with chronic mental health disorders who were using at least one of the following apps: Dialogue, MindBeacon, Deprexis, Ginger, Talkspace, BetterHelp, MindStrong, Mindshift, Bloom, Headspace, and Calm. A total of 347 valid questionnaires were collected and analyzed using partial least-squares structural equation modeling. Trust in the health care provider and patient compliance were measured with multiple-item scales adapted from existing scales. Patient empowerment was conceived and measured as a higher-order construct encompassing the following 2 dimensions: patient process and patient outcome. All the items contributing to the constructs in the model were measured with 7-point Likert scales. The reliability and validity of the measurement model were assessed, and the path coefficients of the structural model were estimated.

Results: The results clearly show that patient empowerment gained through mental health apps positively influenced patient trust in the health care provider (β=.306; P<.001). Patient trust in the health care provider had a positive effect on patient compliance (β=.725; P<.001). The direct relationship between patient empowerment and patient compliance was not significant (β=.061, P=.23). Interestingly, the data highlight that the effect of patient empowerment on patient compliance was fully mediated by trust in the health care provider (β=.222; P<.001). The results show that patient empowerment gained through the mental health app involves 2 dimensions: a process and an outcome.

Conclusions: This study shows that for individuals living with mental health disorders, empowerment gained through mental health apps enhances trust in the health care provider. It reveals that patient empowerment impacts patient compliance but only through the full mediating effect of patient trust in the health care provider, indicating that patient trust is a critical variable to enhance patient compliance. Hence, our results confirm that health care systems could encourage the use of mental health apps to favor a climate that facilitates patients' trust in health care provider recommendations, possibly leading to better compliance with the recommended treatment.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48182) doi:10.2196/48182

KEYWORDS

patient empowerment; patient compliance; patient trust; mental health app; mental health

Introduction

Background

Defined as the extent to which a patient's behavior coincides with the medical or health advice given by a health care provider (eg, psychiatrists, psychologists, nurses, clinical social workers, or case managers [1]), patient compliance plays a vital role in health care provider–patient relationships and health care systems [2]. However, noncompliance to treatment is still one of the greatest challenges in mental health care services, and how to improve compliance remains an issue [3]. The prevalence of noncompliance in mental health disorders is high; studies estimate that approximately 44% of patients with depression and anxiety [4], 50% of those with schizophrenia [5], and 68% of those with opioid use disorder [6] are not fully compliant with recommendations. The literature emphasizes that a lack of patient compliance leads to poor outcomes, which increase health care service use and overall health care costs [7]. For instance, for 2020, US $290 billion of avoidable health care costs have been attributed to noncompliance in the United States [8], and the cost was US $4 billion in Canada [9]. The global economic burden of mental health disorders is expected to rise to US $8.5 trillion in 2030 [10]. Additionally, mental health disorders are a real public health problem [11]. Today, nearly 1 billion people worldwide have mental health issues [12], including 1 of 5 Canadians in any given year [13]. Therefore, it has become a priority to engage patients with mental health disorders to actively participate in their care with the goal of improving the cost-effectiveness of care delivery [14]. This can be achieved with the help of IT, such as mental health apps [7], which improve access to care for patients at the time and place of user convenience at low cost [8]. Mental health apps have been found to empower patients with mental health disorders by actively engaging them in their treatment and in collaborating with their health care provider in the decision-making process [7]. This study examined the relationships between 2 predictors theoretically related to patient compliance among individuals living with mental health disorders using mental health apps. These predictors were patient empowerment and patient trust in health care providers, which have been identified as critical predictors of patient compliance [15-17]. Hence, the objective of this research was to identify the successive impact of empowerment on trust in the health care provider and patient compliance for mental health app users. By addressing this research objective, this study proposed corresponding theoretical and practical implications.

Mental Health App Use

With the proliferation of digital devices and smartphones, health care systems are now implementing digital approaches to deliver mental health at scale [18]. e–Mental health refers to the set of information and communication technologies dedicated to supporting and enhancing the delivery of mental health care [19]. According to the Mental Health Commission of Canada [20], e–mental health services are an effective and complementary solution to traditional care. Today, e–mental health services are diverse and may include, but are not limited to, smartphone apps, web-based information portals, teleconsultation, and virtual reality devices.

In 2021, there were more than 20,000 mental health apps worldwide, and the global market for mental health apps had an annual growth rate of 20% [21]. Mental health apps are smartphone-delivered platforms that provide self-directed or remotely facilitated mental health services in the areas of communication, self-monitoring, diagnosis, and treatment [22]. These apps are designed and used for patients with major depressive disorders [23], bipolar disorders [24], anxiety disorders [17], substance abuse disorders [25], schizophrenia [26], and psychotic disorders [27]. Mental health apps enable their users to better understand their health conditions and allow them to engage with practices that would benefit their health [28]. These devices allow individuals to monitor their symptoms, track their treatment, learn about their health condition, or exchange information with their health care provider. While improving patient activation and facilitating communication with the health care provider, these apps facilitate patient adherence to the recommended treatment. Yet, the literature reveals that the effectiveness of mental health apps is difficult to assess and remains questioned [29]. Besides, research shows that the use of mental health apps dramatically drops over time, with a sustained use of these apps after 6 weeks varying from 0.5% to 28.6% [30]. In addition, the effectiveness of mental health apps is difficult to assess. As an example, there were over 10,000 mental health mobile apps available in the market in 2017; the effectiveness and quality of these services were questioned [26].

Patient Empowerment

In the health care literature, the concept of patient empowerment was introduced in the early 1990s [31]. Since then, it has been the object of increasing attention from both scholars and practitioners [32]. Still, the health care literature reveals a lack of consensual definition [33] and highlights that patient empowerment can be defined through the lens of 2 distinct conceptual approaches.

A first conceptualization considers patient empowerment as a process of behavior change [34,35]. This approach finds its roots in the self-determination theory [36], which proposes that patients are self-determining agents who have the capacity for autonomy. In the context of mental health, the need for autonomy is achieved through patients' involvement in managing their own health [37]. Patient involvement refers here to the importance patients attribute to their health in general and to the acts of maintaining their health [38]. It is manifested in the form of involvement in decision-making and active participation in consultation with health care providers [39]. In turn, active participation facilitates patients' knowledge

XSL•FO
RenderX

development [35]. This is consistent with Funnel et al [31], who emphasized that patients are empowered when they have sufficient knowledge to make rational decisions and have sufficient experience to evaluate the effectiveness of their decisions.

A second conceptualization of patient empowerment considers this construct as an outcome. Empowerment as an outcome refers to a sense of self-efficacy that occurs as a result of the process [34]. In this view, self-efficacy refers to a function of an individual's abilities, knowledge, and learned skills to achieve a desired goal related to their health [40]. This results in an increased feeling of control for patients over their health condition [41]. This is in accordance with the findings of McAllister et al [14], who posited that patients are empowered when they have behavioral control, that is, when they are able to take action to reduce harm or improve their lives.

Considering the 2 abovementioned conceptualizations, patient empowerment is operationalized as a second-order construct formed by 2 first-order independent constructs: patient process and patient outcome. Specifically, we argue that the 2 patient empowerment dimensions are unique first-order constructs because they are independent. This is why patient empowerment is conceptualized as a reflective-formative second-order construct. Consistent with this, we hypothesize that (H1) patient process has a positive effect on empowerment and (H2) patient outcome has a positive effect on empowerment.

Patient Trust in Health Care Providers

Mental health care providers are health care professionals, including psychiatrists, psychologists, nurses, clinical social workers, or case managers [1]. Research shows that empowerment can encourage patient trust in the health care provider [16]. Patient trust in the health care provider refers to the willingness of a patient to be vulnerable to the actions of a health care provider, based on the expectation that the health care provider will perform a particular action that is important to the patient, irrespective of the patient's ability to monitor or control that health care provider [42]. Among the most commonly described dimensions of physician behavior on which patients are believed to base their trust are competence, compassion, privacy and confidentiality, reliability and dependability, and communication [43]. Patients' comfort and confidence in taking control of their health care strengthen the level of trust and commitment they have with their health care provider [44]. Specifically, patient perceived control and participative communication were found to have direct positive effects on trust [45]. Consistent with this view, scholars have demonstrated that patient control over the chronic illness condition and patient participation during medical consultations positively impact patient trust in the health care provider [16]. On the basis of the rationale above, we believe that empowerment gained through mental health apps leads to greater trust in the health care provider. We therefore propose the following hypothesis: (H3) empowerment gained through mental health apps has a positive effect on trust in the health care provider.

Patient Compliance

Patient compliance can be seen as an outcome of the relationship between the patient and the health care provider [46]. Patient compliance refers to the patient's adherence to treatment recommendations and prescriptions targeted to his or her particular disease [47]. The literature reveals that patient compliance is alternatively conceptualized either as a process or as an outcome. Patient compliance as a process relates to the extent to which a patient's behavior coincides with medical advice [48]. Compliance as an outcome is defined as the number of doses taken correctly, which supports the therapeutic outcome [48]. As patients engage in diverse behaviors to increase their well-being and health, and because health care providers give them the means to do so by offering recommendations, treatments, and guidance, it is expected that patient empowerment will enhance patient compliance. Therefore, we propose that (H4) empowerment has a positive effect on patient compliance. Lastly, research suggests different relationships between the patient and the health care provider. Specifically, scholars highlight that trust in the health care provider facilitates the patient to accept the actions of the health care provider [43], and that trust leads to better commitment to the recommended treatment [49]. Given that compliance requires the health care provider to give patients recommendations and prescriptions [48], it is expected that trust leads to greater patient compliance. On the basis of these considerations, we propose the following hypothesis: (H5) trust in the health care provider has a positive effect on patient compliance.

In summary, based on existing literature, the aim of the study is to investigate the impact of patient empowerment gained through mental health apps on patient trust in the health care provider and patient compliance with the recommended treatment.

Methods

Ethical Considerations

This study was approved by the ethics committee on research involving humans of the University of Québec in Montréal (2022-3559). Several specific ethical considerations were considered, and the following was implemented: First, the participants were informed of the length of the survey, the number of questions, the purpose of the study, and the university organizing data collection. Second, the survey did not include identifiable information, and all participants were anonymous. Third, respondents received an incentive after completing the survey, such as cash, airline miles, gift cards, and vouchers.

Study Design

A cross-sectional web-based survey was conducted with Canadian adults (aged 18 years and older) living with a mental health disorder from May 2022 to July 2022. The self-reported survey was administered on the web-based survey platform Qualtrics. The survey is described according to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Multimedia Appendix 1) [50].

XSL•FO
RenderX

Data Collection

A total of 6 experts reviewed the clarity, writing style, and flow of the questionnaire and assessed the understandability of the items. This led to some minor adjustments in the wording of 3 items. The recruitment of the participants was done by Qualtrics, which used niche panels and randomly selected respondents according to inclusion criteria. Three inclusion criteria were defined while ensuring a representative sample of the sociodemographic characteristics of the Canadian population regarding age, sex, income, education, and race. First, we targeted Canadian adults who were using or had used at least one of the following mental health apps: Dialogue, MindBeacon, Deprexis, Ginger, Talkspace, BetterHelp, MindStrong, Mindshift, Bloom, Headspace, and Calm. This selection is based on the judgment of 2 mental health experts who recommended 2 selection criteria: estimated number of users and availability to the patients on the market. Second, we selected participants diagnosed with a mental health disorder. We offered participants the opportunity to indicate the mental health disorder they were experiencing from the following list, which included the most prevalent mental health disorders in Canada: anxiety, bipolarity, depression, schizophrenia, and posttraumatic stress. Additionally, we gave participants the opportunity to specify their mental health disorder in case it was not identified in the abovementioned list. Third, we chose participants who spoke either French or English. To ensure that the participants met the abovementioned inclusion criteria, self-reported questions were included in the questionnaire.

To determine the minimum sample size, the "10 times the number of structural paths directed at a particular construct in the structural model" rule of thumb was assessed [51]. Accordingly, the minimum sample size for the model of this research is 50 participants because the number of structural paths in the research model is 5. A total of 347 surveys were collected, which is high above the minimum required. Table 1 presents the statistics that describe the participants and their characteristics. Of the participants, 46.1% (160/347) were male, 49.6% (172/347) were female, and 4.3% (15/347) were gender-fluid, nonbinary, or two-spirit. In terms of age distribution, the age groups 26-35 years (112/347, 32.3%) and 36-49 years (94/347, 27.1%) were the most represented. Education level was relatively high, with university level accounting for 46.1% (160/347). The annual income distribution was relatively even. Participants had been diagnosed with a mental health disorder (ie, depression, anxiety, schizophrenia, and posttraumatic stress) and were using or had used at least 1 mental health app. The users in our study mainly use the apps Calm (82/347, 23.6%) and Headspace (60/347, 17.2%). Participants were also or had been under the care of a health care provider.

Table 1. The demographics of the sample (N=347).

Category	Value
Sex, n (%)	
Male	160 (46.1)
Female	172 (49.6)
Gender-fluid, nonbinary, or two-spirit	15 (4.3)
Age (years), n (%)	
18-25	89 (25.6)
26-35	112 (32.3)
36-49	94 (27.1)
50-65	48 (13.8)
>66	4 (1.2)
Educational background	
Postsecondary	114 (32.9)
Secondary	68 (19.6)
University	160 (46.1)
Prefer not to answer	5 (1.4)
Income (US $)	
<39,999	89 (25.6)
40,000-79,999	114 (32.9)
80,000-119,999	72 (20.7)
120,000-159,999	32 (9.2)
160,000-199,999	27 (7.8)
>200,000	13 (3.7)

XSL•FO
RenderX

Scale Development

The scales used in the measurement model were designed by the researchers based on the existing literature. The scale of patient process included 1 item to measure health involvement adapted from Oh and Lee's [52] scale and 3 items to measure knowledge development adapted from the scale by Prigge et al [15]. The scale for patient outcome included 1 item to measure the patient's feeling of control adapted from Oh and Lee's scale [52] and 3 items to measure self-efficacy adapted from the scale by Prigge et al [15]. Items used to measure patient process and patient outcome were adapted to the context of mental health app use behavior. Patient empowerment is conceptualized as a construct constituted by 2 independent dimensions: patient process and patient outcome. Therefore, we operationalized patient empowerment as a second-order construct. We chose the repeated indicators approach to measure this construct as the approach is suitable for measuring a higher-order construct [53]. We used this approach instead of the 2-stage approach or the hybrid approach [53] because patient empowerment is measured as a formative-reflective second-order construct [54] and was an endogenous construct in the path model [53]. Consequently, we used the manifest variables of the first-order latent variables (patient process and patient outcome) to measure the second-order latent variable (patient empowerment). The trust scale included 5 items based on the scale by Anderson and Dedrick [55]. Finally, the patient compliance scale consisted of 2 items based on the scale by Prigge et al [15]. We used age, sex, education, and income as control variables; they were each measured with a single item. The questionnaire used 7-point Likert scales anchored from "completely disagree" to "completely agree" (Multimedia Appendix 2 [15,52,53,55]). As the Likert scales we used included more than 5 categories, these scales were treated as continuous measures. This follows the literature recommendations [56-58] and meets the partial least-squares (PLS) structural requirements [53,59].

Data Analysis

The data were analyzed by PLS structural equation modeling (PLS-SEM). SEM is considered a second-generation multivariate analysis technique that allows researchers to incorporate unobservable or latent variables measured indirectly by observable variables [60]. PLS-SEM was chosen instead of covariance-based SEM as it does not assume that data are distributed normally, as was the case with our data [51]. The measurement model (the inner model) was first assessed; we evaluated the reliability and validity of the estimates for the latent variables. The structural model (the outer model) was then calculated to assess the direction and significance of the relationships between the latent variables [60,61]. To do so, we used the SmartPLS 3 software (SmartPLS GmbH) for the PLS-SEM analysis. We chose the standard bootstrap procedure on 5000 bootstrapping samples to examine the structural model's path coefficients and the corresponding significance levels [62].

Results

Measurement Model (Inner Model)

Because all the constructs were reflectively measured in our model, assessing the measurement model was required to evaluate the reliability, the convergent validity, and the discriminant validity of the measurement scales [51]. Table 2 shows the composite reliability, Cronbach α, and average variance extracted (AVE). The factor indicators, known as the outer loadings or reflexive indicator loadings, were superior to the 0.5 threshold recommended for each indicator, demonstrating that the chosen indicators adequately measure the latent variables [51]. The Cronbach α coefficient values ranged from .762 to .863, and the composite reliability values were between 0.770 and 0.868, indicating that constructs had good reliability. All the average variances extracted were greater than 0.5, suggesting a good convergent validity for the 5 constructs. Altogether, these findings demonstrated that the 5 constructs had good convergent validity.

To establish the discriminant validity, the square root of the AVE of each construct had to be larger than its correlation with other constructs. Additionally, because empowerment was operationalized as a second-order construct, we do not consider the discriminant validity between both patient outcome and patient process and their higher-order component patient empowerment. The correlation between patient outcome (patient process) and patient empowerment is expected to be greater than the square root of the AVE of patient outcome (patient process) because the measurement model of the patient empowerment construct repeats the indicators of its 2 lower-order components (patient process and patient outcome) [53]. Table 3 shows that the square root of each factor's AVE value is greater than the other factor correlation coefficients, indicating a good discriminant validity for the 5 constructs.

XSL•FO
RenderX

Table 2. Construct reliability and convergent validity.

Construct and item	Indicator loading	Cronbach α	Composite reliability	Average variance extracted
PC[a]		.851	0.851	0.870
PC1	0.934	N/A[b]	N/A	N/A
PC2	0.932	N/A	N/A	N/A
PP[c]		.816	0.834	0.651
PP1	0.826	N/A	N/A	N/A
PP2	0.871	N/A	N/A	N/A
PP3	0.870	N/A	N/A	N/A
PP4	0.637	N/A	N/A	N/A
PO[d]		.762	0.770	0.651
PO1	0.872	N/A	N/A	N/A
PO2	0.829	N/A	N/A	N/A
PO3	0.767	N/A	N/A	N/A
PT[e]		.863	0.868	0.646
PT1	0.866	N/A	N/A	N/A
PT2	0.797	N/A	N/A	N/A
PT3	0.773	N/A	N/A	N/A
PT4	0.767	N/A	N/A	N/A
PT5	0.812	N/A	N/A	N/A
Patient empowerment		.843	0.850	0.520
PP1	0.750	N/A	N/A	N/A
PP2	0.816	N/A	N/A	N/A
PP3	0.780	N/A	N/A	N/A
PP4	0.597	N/A	N/A	N/A
PO1	0.760	N/A	N/A	N/A
PO2	0.663	N/A	N/A	N/A
PO3	0.656	N/A	N/A	N/A

[a]PC: patient compliance.
[b]N/A: not applicable.
[c]PP: patient process.
[d]PO: patient outcome.
[e]PT: patient trust in the health care provider.

XSL•FO
RenderX

Table 3. Construct discriminant validity.

Construct	PC[a]	PE[b]	PO[c]	PP[d]	PT[e]
PC	0.933	N/A[f]	N/A	N/A	N/A
PE	0.283	0.721	N/A	N/A	N/A
PO	0.278	0.844	0.824	N/A	N/A
PP	0.230	0.918	0.561	0.807	N/A
PT	0.744	0.306	0.278	0.265	0.804

[a]PC: patient compliance.
[b]PE: patient empowerment.
[c]PO: patient outcome.
[d]PP: patient process.
[e]PT: patient trust in the health care provider.
[f]N/A: not applicable.

Structural Model (Outer Model)

To assess the structural model for collinearity issues, we considered the variance inflation factor in the predictor constructs as indicative of collinearity [60]. In this study, all variance inflation factor values were below 5, which indicates no violation of the multicollinearity assumption. The model's goodness of fit was our criterion to assess the overall fit of the model. The model's goodness of fit for this study was 0.607, allowing us to conclude that our model performed well [63]. The effects of the control variables—age, sex, education, and income—on patient trust in the health care provider have been found to be insignificant.

The results showed that all the hypothesized anticipated relationships are supported except for the impact of empowerment on patient compliance (Figure 1). Four (H1, H2, H3, and H5) of our 5 hypotheses were validated by our data (Table 4). Patient process was found to be a first-order construct of empowerment as patient process positively impacted patient empowerment (β=.648; *P*<.001). Similarly, patient outcome was found to be a first-order construct of empowerment as patient outcome positively impacted patient empowerment (β=.480; *P*<.001). Besides, patient empowerment was found to have a positive effect on patient trust in the health care provider (β=.306; *P*<.001). Patient trust in the health care provider also had a positive effect on patient compliance (β=.725; *P*<.001). Interestingly, the relationship between empowerment and patient compliance was not significant (β=.061, *P*=.23).

We conducted the indirect effect (mediation) of trust to further understand the relationship between empowerment and patient compliance. The indirect effect via the path empowerment→trust→patient compliance significantly mediated the effect of empowerment on patient compliance (β=.222; *P*<.001), indicating that trust fully mediated the relationship between empowerment and patient compliance (Table 5).

Figure 1. Structural model showing path coefficients (β). ***P*<.001. NS: not significant.

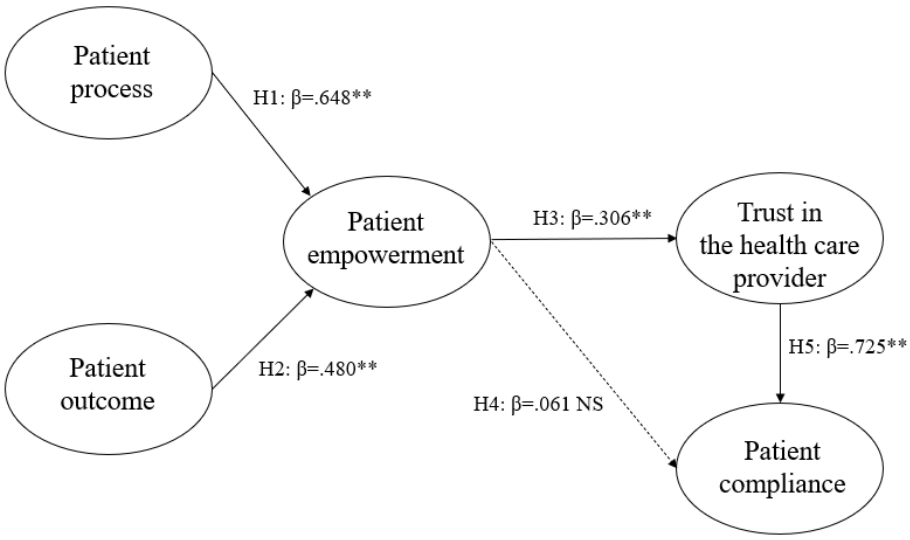

XSL•FO
RenderX

Table 4. Hypothesis test results.

Hypothesis	Path	Path coefficient (SD)	*P* value	Result
H1	PP[a]→PE[b]	0.648 (0.023)	<.001	Supported
H2	PO[c]→PE	0.480(0.021)	<.001	Supported
H3	PE→PT[d]	0.306 (0.062)	<.001	Supported
H4	PE→PC[e]	0.061 (0.051)	.23	Rejected but fully mediated by trust in the health care provider
H5	PT→PC	0.725 (0.042)	<.001	Supported

[a]PP: patient process.
[b]PE: patient empowerment.
[c]PO: patient outcome.
[d]PT: patient trust in the health care provider.
[e]PC: patient compliance.

Table 5. The mediation effect of trust in the health care provider.

Path	Direct effect		Indirect effect	
	β value	*P* value	β value	*P* value
PE[a]→PT[b]→PC[c]	.061	.23	.222	<.001

[a]PE: patient empowerment.
[b]PT: patient trust in the health care provider.
[c]PC: patient compliance.

Discussion

Principal Findings

The objective of this study was to investigate the impact of patient empowerment gained through mental health apps on patient trust in the health care provider and patient compliance. We examined the 2 dimensions of patient empowerment and its effects on trust in the health care provider and patient compliance. Data were collected from participants with mental health disorders using the survey method and analyzed with the PLS-SEM method. Findings reveal that patient empowerment is a second-order construct composed of 2 dimensions: patient process and patient outcome. The data revealed that patient empowerment directly and positively impacts patient trust in the health care provider. Patient trust in the health care provider also directly and positively impacts patient compliance. However, no direct effect of patient empowerment on patient compliance was supported, indicating that trust is a full mediator between patient empowerment and patient compliance.

Theoretical Contributions

This study contributes to the research field in 4 ways. First, this research clearly shows that patient empowerment gained through mental health apps enhances patient compliance. Previous research that has shown a positive relationship between patient empowerment and patient compliance has been conducted on patients living with chronic physical diseases [15,64]. While focusing on individuals living with mental health disorders, this study is the first to demonstrate the effect of patient empowerment on compliance with the recommended treatment in mental health app users.

Second and importantly, this study reveals that the relationship between the patient empowerment gained through the mental health app and patient compliance is fully mediated by patient trust in the health care provider, indicating that patient trust is a critical variable to enhance patient compliance. Past research has found that patients in a hospital setting who were empowered through education and doctor support demonstrated more trust in their caregivers [44], and that patient trust in the health care provider enhances patient compliance [2,65]. Our study supports these findings in the field of mental health. It shows the existence of a positive relationship between patient empowerment gained through an app dedicated to mental health and patient trust in the health care provider. In turn, our findings highlight that patient trust significantly enhances patient compliance with the health care provider recommendations. While showing that patient empowerment gained using a mental health app has no direct effect on compliance with the health care provider's recommended treatment, but that it increases patient trust, this study underlines that, thanks to mental health apps, patient empowerment and patient trust are intertwined to enhance patient compliance. Hence, this research provides a better understanding of the cognitive mechanisms and paths through which mental health apps enhance user compliance with the caregiver's recommendations.

Third, incidentally, this study provides an interesting theoretical view, operationalization, and validation of the construct of patient empowerment. We propose a thorough conceptualization and measurement of this construct. The literature in health care remains inconclusive and reveals a lack of consensus in defining patient empowerment [33,35,66,67]. We propose that the patient empowerment gained using mental health apps be a multidimensional construct, built around 2 dimensions.

XSL•FO
RenderX

Precisely, we define patient empowerment as a patient's sense of control in managing their health status, resulting from (1) an increase in patient knowledge and (2) a voluntary commitment to improving their health. In line with previous studies [30,34], this definition conceives that patient empowerment is a second-order construct, consisting of both a process and an outcome. The outcome dimension refers to the feeling of control and the patient's ability to self-manage, whereas the process dimension refers to knowledge development and self-involvement in managing their health.

Practical Contributions

These findings have several implications for practice. First, the results suggest that health care providers should encourage and educate individuals living with mental health disorders to adopt and use mental health apps. The use of mental health apps would be beneficial given that the number of individuals with mental health disorders is increasing, especially young people. This research demonstrated that empowerment gained by a mental health app has a positive effect on patient trust in the health care provider. Hence, health care providers should play a crucial role in the adoption of mental health apps for their users. They should educate patients, engage in discussions with them, and answer their questions and concerns to facilitate the adoption process of mental health apps. This education process is important because the misuse of such devices can harm patients who would conduct incorrect self-diagnosis due to poor eHealth literacy and misinterpreted information [68]. Health care providers sometimes lack knowledge and skills regarding health app use [69]; as a result, this process also implies that health care providers should be trained to recommend and support mental health apps. Health authorities and mental health app designers or specialists appear to be the right entities to train health care providers.

Second, human-centered designers, alongside software developers, should further develop features that empower the patients, for example, functionalities that improve the patients' self-efficacy or knowledge development or that support and stimulate self-management strategies [70]. Designers and developers could, for instance, develop mental health apps that feature alarms, reminder scheduling, behavioral tracking tools, symptoms manager, care pathway timeline, a system of conversations with peers, automated routines, and healthy lifestyle daily advice. This should be done through a co-design process rooted in a human-centered design [71] that allows end users (patients) to meaningfully contribute to the definition of these features as closely as possible to their lived experience.

Third, this study suggests that, because patients are compliant when they trust their health care provider, these providers should create a climate that facilitates patients' trust in their practices. For instance, health care providers should consider patients' needs and concerns, answer their questions with respect for their feelings, and demonstrate empathy.

Limitations

Although this study enriches the knowledge of patient empowerment using mental health apps and its effects on patient-health care provider relationships, it has several limitations. The first limitation regards the use of a cross-sectional survey, which reflects data from respondents at one point in time. Hence, it is recommended to carry out a longitudinal survey in the future to extend the validity of our findings. Second, our model focused on 3 key variables: patient empowerment gained using mental health apps, patient trust in the health care provider, and patient compliance. Further research investigating the effect of patient empowerment on other variables, such as patient commitment, patient satisfaction, or patient social support, should improve our knowledge of the effects of the use of mental health apps. Third, participants in our sample were individuals living with mental health disorders in Canada, which offers a universal health care system to its citizens. It would be of interest to test the proposed theoretical framework on other countries having different health care systems, such as private health care systems. This would increase the external validity of our findings. Fourth, investigating the effect of mental health apps for each disorder pathology in isolation would enrich the present findings. After this study, we will conduct a second research project to examine case by case the effects of mental health apps across various mental health disorders such as depression, anxiety, or bipolarity. The fifth limitation of this study relies in the self-report assessment of the mental health diagnosis. Yet, self-report assessment for mental health diagnosis was found to be highly and significantly correlated to clinical assessments [72]. Conducting an observational study in the future would complement these findings and thus extend the validity of our results.

Conclusions

This study clearly shows that mental health apps enhance patient empowerment for an individual living with mental health disorders, which in turn leads to patient trust in the health care provider. Interestingly, patient empowerment gained using mental health apps is found to impact patient compliance, but only through the full mediating effect of patient trust in the health care provider. This research suggests that health care providers should encourage patients to use mental health apps to enhance their feeling of empowerment. We believe that this research stream is and shall continue to be of considerable interest as empowering patients has become a priority for policymakers, with the goal of improving the quality and efficacy the health care delivery system.

Acknowledgments

This work was supported in Canada by the Social Sciences and Humanities Research Council and the Chaire Diament lab at the Université du Québec à Montréal.

XSL•FO
RenderX

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JF, A-FA-P, SB, and SV conceptualized the study. JF and A-FA-P conceived the study design and undertook the data analysis. JF, AFAP, SB, and SV drafted the manuscript. JF, A-FA-P, SB, and SV revised the manuscript. All the authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 18 KB - jmir_v26i1e48182_app1.docx]

Multimedia Appendix 2
Measurement instruments.
[DOC File , 84 KB - jmir_v26i1e48182_app2.doc]

References

1. Hatch A, Hoffman JE, Ross R, Docherty JP. Expert consensus survey on digital health tools for patients with serious mental illness: optimizing for user characteristics and user support. JMIR Ment Health 2018 Jun 12;5(2):e46 [FREE Full text] [doi: 10.2196/mental.9777] [Medline: 29895514]
2. Lu X, Zhang R, Wu W, Shang X, Liu M. Relationship between Internet health information and patient compliance based on trust: empirical study. J Med Internet Res 2018;20(8):e253 [FREE Full text] [doi: 10.2196/jmir.9364] [Medline: 30120087]
3. Serobatse MB, Du Plessis E, Koen MP. Interventions to promote psychiatric patients' compliance to mental health treatment: a systematic review. Health SA Gesondheid 2014;19(1):a799 [FREE Full text] [doi: 10.4088/JCP.14m09001]
4. Bet PM, Penninx BWJH, van Laer SD, Hoogendijk WJG, Hugtenburg JG. Current and remitted depression and anxiety disorders as risk factors for medication nonadherence. J Clin Psychiatry 2015;76(9):e1114-e1121. [doi: 10.4088/JCP.14m09001] [Medline: 26455675]
5. Acosta FJ, Hernández JL, Pereira J, Herrera J, Rodríguez CJ. Medication adherence in schizophrenia. World J Psychiatry 2012;2(5):74-82 [FREE Full text] [doi: 10.5498/wjp.v2.i5.74] [Medline: 24175171]
6. Tkacz J, Volpicelli J, Un H, Ruetsch C. Relationship between buprenorphine adherence and health service utilization and costs among opioid dependent patients. J Subst Abuse Treat 2014;46(4):456-462 [FREE Full text] [doi: 10.1016/j.jsat.2013.10.014] [Medline: 24332511]
7. Qudah B, Luetsch K. The influence of mobile health applications on patient—healthcare provider relationships: a systematic, narrative review. Patient Educ Couns 2019;102(6):1080-1089. [doi: 10.1016/j.pec.2019.01.021] [Medline: 30745178]
8. Medication non-adherence: a $290 billion unnecessary expenditure. TripleTree. 2020. URL: https://www.triple-tree.com/strategic-insights/2015/april/medication-non-adherence-a-$290-billion-unnecessa/ [accessed 2024-01-17]
9. Medication non-adherence: exploring the tension between patient safety and patient autonomy. Canada Health Infoway. 2018. URL: https://www.infoway-inforoute.ca/en/component/content/article?id=7874:medication-non-adherence-exploring-the-tension-between-patient-safety-and-patient-autonomy [accessed 2024-01-17]
10. Trautmann S, Rehm J, Wittchen HU. The economic costs of mental disorders: do our societies react appropriately to the burden of mental disorders? EMBO Rep 2016;17(9):1245-1249 [FREE Full text] [doi: 10.15252/embr.201642951] [Medline: 27491723]
11. Reijneveld S. Mental health as a public health issue. Eur J Public Health 2005;15(2):111 [FREE Full text] [doi: 10.1093/eurpub/cki088]
12. World mental health report: transforming mental health for all. UN Health Agency. 2022. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-01-10]
13. Mental illness and addiction: facts and statistics. Centre for Addiction and Mental Health. 2023. URL: https://www.camh.ca/en/driving-change/the-crisis-is-real/mental-health-statistics [accessed 2023-10-06]
14. McAllister M, Dunn G, Payne K, Davies L, Todd C. Patient empowerment: the need to consider it as a measurable patient-reported outcome for chronic conditions. BMC Health Serv Res 2012;12:157 [FREE Full text] [doi: 10.1186/1472-6963-12-157] [Medline: 22694747]
15. Prigge JK, Dietz B, Homburg C, Hoyer WD, Burton JL. Patient empowerment: a cross-disease exploration of antecedents and consequences. Int J Res Mark 2015;32(4):375-386. [doi: 10.1016/j.ijresmar.2015.05.009]

XSL•FO
RenderX

16. Ouschan R, Sweeney J, Johnson L. Customer empowerment and relationship outcomes in healthcare consultations. Eur J Mark 2006;40(9/10):1068-1086. [doi: 10.1108/03090560610681014]
17. Hojat M, Louis DZ, Maxwell K, Markham F, Wender R, Gonnella JS. Patient perceptions of physician empathy, satisfaction with physician, interpersonal trust, and compliance. Int J Med Educ 2010;1:83-87 [FREE Full text] [doi: 10.5116/ijme.4d00.b701]
18. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
19. Riper H, Andersson G, Christensen H, Cuijpers P, Lange A, Eysenbach G. Theme issue on e-mental health: a growing field in internet research. J Med Internet Res 2010;12(5):e74 [FREE Full text] [doi: 10.2196/jmir.1713] [Medline: 21169177]
20. Changing directions, changing lives: the mental health strategy for Canada. Mental Health Commission of Canada. 2012. URL: https://www.mentalhealthcommission.ca/wp-content/uploads/drupal/MHStrategy_Strategy_ENG.pdf [accessed 2024-01-10]
21. Deloitte Insights. Mental health goes mobile: the mental health app market will keep on growing. Deloitte Insights. 2021. URL: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/mental-health-app-market.html [accessed 2022-11-10]
22. Koh J, Tng GYQ, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022;12(9):1376 [FREE Full text] [doi: 10.3390/jpm12091376] [Medline: 36143161]
23. Watts S, Mackenzie A, Thomas C, Griskaitis A, Mewton L, Williams A, et al. CBT for depression: a pilot RCT comparing mobile phone vs. computer. BMC Psychiatry 2013;13:49 [FREE Full text] [doi: 10.1186/1471-244X-13-49] [Medline: 23391304]
24. Faurholt-Jepsen M, Frost M, Vinberg M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as objective measures of bipolar disorder symptoms. Psychiatry Res 2014;217(1-2):124-127 [FREE Full text] [doi: 10.1016/j.psychres.2014.03.009] [Medline: 24679993]
25. Gustafson DH, McTavish FM, Chih MY, Atwood AK, Johnson RA, Boyle MG, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA Psychiatry 2014;71(5):566-572 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.4642] [Medline: 24671165]
26. Ben-Zeev D, Scherer EA, Gottlieb JD, Rotondi AJ, Brunette MF, Achtyes ED, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016;3(3):e34 [FREE Full text] [doi: 10.2196/mental.6348] [Medline: 27465803]
27. Palmier-Claus JE, Ainsworth J, Machin M, Barrowclough C, Dunn G, Barkus E, et al. The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application. BMC Psychiatry 2012;12:172 [FREE Full text] [doi: 10.1186/1471-244X-12-172] [Medline: 23075387]
28. Veer E, Dobele A. Big boys don't cry [Offline]: the phygital disconnect between online and offline mental wellness engagement. J Strategic Mark 2021:1-21. [doi: 10.1080/0965254x.2021.1965190]
29. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021;147:103990 [FREE Full text] [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
30. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
31. Funnell MM, Anderson RM, Arnold MS, Barr PA, Donnelly M, Johnson PD, et al. Empowerment: an idea whose time has come in diabetes education. Diabetes Educ 1991;17(1):37-41 [FREE Full text] [doi: 10.1177/014572179101700108] [Medline: 1986902]
32. Gibson CH. A concept analysis of empowerment. J Adv Nurs 1991;16(3):354-361. [doi: 10.1111/j.1365-2648.1991.tb01660.x] [Medline: 2037742]
33. Pekonen A, Eloranta S, Stolt M, Virolainen P, Leino-Kilpi H. Measuring patient empowerment—a systematic review. Patient Educ Couns 2020;103(4):777-787. [doi: 10.1016/j.pec.2019.10.019] [Medline: 31767243]
34. Anderson RM, Funnell MM. Patient empowerment: myths and misconceptions. Patient Educ Couns 2010;79(3):277-282 [FREE Full text] [doi: 10.1016/j.pec.2009.07.025] [Medline: 19682830]
35. Aujoulat I, d'Hoore W, Deccache A. Patient empowerment in theory and practice: polysemy or cacophony? Patient Educ Couns 2007;66(1):13-20. [doi: 10.1016/j.pec.2006.09.008] [Medline: 17084059]
36. Deci EL, Ryan RM. The general causality orientations scale: self-determination in personality. J Res Pers 1985;19(2):109-134. [doi: 10.1016/0092-6566(85)90023-6]
37. Rice T. Individual autonomy and state involvement in health care. J Med Ethics 2001;27(4):240-244 [FREE Full text] [doi: 10.1136/jme.27.4.240] [Medline: 11479354]
38. Zaichkowsky JL. Measuring the involvement construct. J Consum Res 1985;12(3):341-352. [doi: 10.1086/208520]
39. Vahdat S, Hamzehgardeshi L, Hessam S, Hamzehgardeshi Z. Patient involvement in health care decision making: a review. Iran Red Crescent Med J 2014;16(1):e12454 [FREE Full text] [doi: 10.5812/ircmj.12454] [Medline: 24719703]

XSL•FO
RenderX

40. Loukanova S, Molnar R, Bridges JF. Promoting patient empowerment in the healthcare system: highlighting the need for patient-centered drug policy. Expert Rev Pharmacoecon Outcomes Res 2007;7(3):281-289. [doi: 10.1586/14737167.7.3.281] [Medline: 20528314]
41. François J, Audrain‐Pontevia AP, Menvielle L, Chevalier N. Empowering health care consumers in the era of internet of things. Int J Consumer Studies 2022;47(3):1060-1075. [doi: 10.1111/ijcs.12887]
42. Leisen B, Hyman MR. An improved scale for assessing patients' trust in their physician. Health Mark Q 2001;19(1):23-42. [doi: 10.1300/J026v19n01_03] [Medline: 11727290]
43. Pearson SD, Raeke LH. Patients' trust in physicians: many theories, few measures, and little data. J Gen Intern Med 2000;15(7):509-513 [FREE Full text] [doi: 10.1046/j.1525-1497.2000.11002.x] [Medline: 10940139]
44. Mohd‐Any AA, Sundramohana M, Sarker M. Does patient empowerment matter in building loyalty? Int J Consumer Studies 2021;46(2):653-675. [doi: 10.1111/ijcs.12718]
45. Gabay G. Perceived control over health, communication and patient-physician trust. Patient Educ Couns 2015;98(12):1550-1557. [doi: 10.1016/j.pec.2015.06.019] [Medline: 26187177]
46. Náfrádi L, Nakamoto K, Schulz PJ. Is patient empowerment the key to promote adherence? A systematic review of the relationship between self-efficacy, health locus of control and medication adherence. PLoS One 2017;12(10):e0186458 [FREE Full text] [doi: 10.1371/journal.pone.0186458] [Medline: 29040335]
47. Hausman A. Modeling the patient-physician service encounter: improving patient outcomes. J Acad Mark Sci 2004;32(4):403-417. [doi: 10.1177/0092070304265627]
48. Morris LS, Schulz RM. Patient compliance—an overview. J Clin Pharm Ther 1992;17(5):283-295. [doi: 10.1111/j.1365-2710.1992.tb01306.x] [Medline: 1464632]
49. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
50. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
51. Hair JF, editor. A Primer on Partial Least Squares Structural Equations Modeling (PLS-SEM). Los Angeles, CA: SAGE; 2014.
52. Oh HJ, Lee B. The effect of computer-mediated social support in online communities on patient empowerment and doctor-patient communication. Health Commun 2012;27(1):30-41. [doi: 10.1080/10410236.2011.567449] [Medline: 21797714]
53. Sarstedt M, Hair JF, Cheah JH, Becker JM, Ringle CM. How to specify, estimate, and validate higher-order constructs in PLS-SEM. Australas Mark J 2021;27(3):197-211. [doi: 10.1016/j.ausmj.2019.05.003]
54. Becker JM, Klein K, Wetzels M. Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models. Long Range Planning 2012;45(5-6):359-394. [doi: 10.1016/j.lrp.2012.10.001]
55. Anderson LA, Dedrick RF. Development of the trust in physician scale: a measure to assess interpersonal trust in patient-physician relationships. Psychol Rep 1990;67(3 Pt 2):1091-1100. [doi: 10.2466/pr0.1990.67.3f.1091] [Medline: 2084735]
56. Sullivan GM, Artino AR. Analyzing and interpreting data from Likert-type scales. J Grad Med Educ 2013;5(4):541-542 [FREE Full text] [doi: 10.4300/JGME-5-4-18] [Medline: 24454995]
57. Norman G. Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ Theory Pract 2010;15(5):625-632. [doi: 10.1007/s10459-010-9222-y] [Medline: 20146096]
58. Allen IE, Seaman CA. Likert scales and data analyses. Qual Prog 2007;40(7):64-654 [FREE Full text]
59. Sarstedt M, Ringle CM, Cheah JH, Ting H, Moisescu OI, Radomir L. Structural model robustness checks in PLS-SEM. Tourism Economics 2019 Jan 23;26(4):531-554. [doi: 10.1177/1354816618823921]
60. Sarstedt M, Ringle CM, Hair JF. Partial least squares structural equation modeling. In: Homburg C, Klarmann M, Vomberg AE, editors. Handbook of Market Research. Cham, Switzerland: Springer International Publishing; 2020:1-47.
61. Henseler J, Ringle CM, Sinkovics RR. The use of partial least squares path modeling in international marketing. In: Sinkovics RR, Ghauri PN, editors. New Challenges to International Marketing. Bingley, UK: Emerald Group Publishing; 2009:277-319.
62. Streukens S, Leroi-Werelds S. Bootstrapping and PLS-SEM: a step-by-step guide to get more out of your bootstrap results. Eur Manag J 2016;34(6):618-632. [doi: 10.1016/j.emj.2016.06.003]
63. Wetzels M, Odekerken-Schröder G, van Oppen C. Using PLS path modeling for assessing hierarchical construct models: guidelines and empirical illustration. MIS Quarterly 2009;33(1):177-195. [doi: 10.2307/20650284]
64. Audrain-Pontevia AF, Menvielle L, Ertz M. Effects of three antecedents of patient compliance for users of peer-to-peer online health communities: cross-sectional study. J Med Internet Res 2019;21(11):e14006 [FREE Full text] [doi: 10.2196/14006] [Medline: 31710295]
65. Kerse N, Buetow S, Mainous AG, Young G, Coster G, Arroll B. Physician-patient relationship and medication compliance: a primary care investigation. Ann Fam Med 2004;2(5):455-461 [FREE Full text] [doi: 10.1370/afm.139] [Medline: 15506581]

XSL•FO
RenderX

66. Markwart H, Bomba F, Menrath I, Brenk-Franz K, Ernst G, Thyen U, et al. Assessing empowerment as multidimensional outcome of a patient education program for adolescents with chronic conditions: a latent difference score model. PLoS One 2020;15(4):e0230659 [FREE Full text] [doi: 10.1371/journal.pone.0230659] [Medline: 32315371]
67. Risling T, Martinez J, Young J, Thorp-Froslie N. Evaluating patient empowerment in association with eHealth technology: scoping review. J Med Internet Res 2017;19(9):e329 [FREE Full text] [doi: 10.2196/jmir.7809] [Medline: 28963090]
68. Cline RJ, Haynes KM. Consumer health information seeking on the internet: the state of the art. Health Educ Res 2001;16(6):671-692 [FREE Full text] [doi: 10.1093/her/16.6.671] [Medline: 11780707]
69. Hennemann S, Beutel ME, Zwerenz R. Ready for eHealth? Health professionals' acceptance and adoption of eHealth interventions in inpatient routine care. J Health Commun 2017;22(3):274-284. [doi: 10.1080/10810730.2017.1284286] [Medline: 28248626]
70. Villaggi B, Provencher H, Coulombe S, Meunier S, Radziszewski S, Hudon C, et al. Self-management strategies in recovery from mood and anxiety disorders. Glob Qual Nurs Res 2015;2:2333393615606092 SAGE Publications Inc [FREE Full text] [doi: 10.1177/2333393615606092] [Medline: 28462317]
71. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022;9(6):e35591 [FREE Full text] [doi: 10.2196/35591] [Medline: 35671081]
72. Arrow K, Resnik P, Michel H, Kitchen C, Mo C, Chen S, et al. Evaluating the use of online self-report questionnaires as clinically valid mental health monitoring tools in the clinical whitespace. Psychiatr Q 2023;94(2):221-231 [FREE Full text] [doi: 10.1007/s11126-023-10022-1] [Medline: 37145257]

Abbreviations

AVE: average variance extracted
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
PLS-SEM: partial least-squares structural equation modeling

Edited by T de Azevedo Cardoso, T Leung; submitted 14.04.23; peer-reviewed by L Menvielle, R Cochran, X Lu; comments to author 29.08.23; revised version received 12.10.23; accepted 18.12.23; published 12.02.24.

Please cite as:
François J, Audrain-Pontevia AF, Boudhraâ S, Vial S
Assessing the Influence of Patient Empowerment Gained Through Mental Health Apps on Patient Trust in the Health Care Provider and Patient Compliance With the Recommended Treatment: Cross-sectional Study
J Med Internet Res 2024;26:e48182
URL: https://www.jmir.org/2024/1/e48182
doi:10.2196/48182
PMID:38345851

©Julien François, Anne-Françoise Audrain-Pontevia, Sana Boudhraâ, Stéphane Vial. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Media Use Behavior Mediates the Association Between Family Health and Intention to Use Mobile Health Devices Among Older Adults: Cross-Sectional Study

Jinghui Chang[1*], PhD; Yanshan Mai[2*]; Dayi Zhang[2]; Xixi Yang[1]; Anqi Li[1], MSc; Wende Yan[2]; Yibo Wu[3], PhD; Jiangyun Chen[1], PhD

[1]School of Health Management, Southern Medical University, Guangzhou, China
[2]School of Public Health, Southern Medical University, Guangzhou, China
[3]School of Public Health, Peking University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Jiangyun Chen, PhD
School of Health Management
Southern Medical University
Number 1023, South Shatai Road
Baiyun District
Guangzhou, 510515
China
Phone: 86 1 858 822 0304
Email: cjy112@i.smu.edu.cn

Abstract

Background: With the advent of a new era for health and medical treatment, characterized by the integration of mobile technology, a significant digital divide has surfaced, particularly in the engagement of older individuals with mobile health (mHealth). The health of a family is intricately connected to the well-being of its members, and the use of media plays a crucial role in facilitating mHealth care. Therefore, it is important to examine the mediating role of media use behavior in the connection between the family health of older individuals and their inclination to use mHealth devices.

Objective: This study aims to investigate the impact of family health and media use behavior on the intention of older individuals to use mHealth devices in China. The study aims to delve into the intricate dynamics to determine whether media use behavior serves as a mediator in the relationship between family health and the intention to use mHealth devices among older adults. The ultimate goal is to offer well-founded and practical recommendations to assist older individuals in overcoming the digital divide.

Methods: The study used data from 3712 individuals aged 60 and above, sourced from the 2022 Psychology and Behavior Investigation of Chinese Residents study. Linear regression models were used to assess the relationships between family health, media use behavior, and the intention to use mHealth devices. To investigate the mediating role of media use behavior, we used the Sobel-Goodman Mediation Test. This analysis focused on the connection between 4 dimensions of family health and the intention to use mHealth devices.

Results: A positive correlation was observed among family health, media use behavior, and the intention to use mHealth devices (r=0.077-0.178, P<.001). Notably, media use behavior was identified as a partial mediator in the relationship between the overall score of family health and the intention to use mHealth devices, as indicated by the Sobel test (z=5.451, P<.001). Subgroup analysis further indicated that a complete mediating effect was observed specifically between family health resources and the intention to use mHealth devices in older individuals with varying education levels.

Conclusions: The study revealed the significance of family health and media use behavior in motivating older adults to adopt mHealth devices. Media use behavior was identified as a mediator in the connection between family health and the intention to use mHealth devices, with more intricate dynamics observed among older adults with lower education levels. Going forward, the critical role of home health resources must be maximized, such as initiatives to develop digital education tailored for older adults and the creation of media products specifically designed for them. These measures aim to alleviate technological challenges associated with using media devices among older adults, ultimately bolstering their inclination to adopt mHealth devices.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50012) doi:10.2196/50012

KEYWORDS

older adults; family health; media use; intention to use mobile health devices; China

Introduction

Background

The 2022 United Nations report on "World Population Prospects" predicted that by 2050, the global population will reach 9.7 billion. Within this demographic shift, 1.5 billion individuals aged 65 and above are anticipated, constituting 16% of the total population [1]. Notably, the trend of population aging is intensifying. In the context of population dynamics, China, as a heavily populated nation, is undergoing significant and intricate transformations. The Seventh National Population Census of China revealed that there are 264 million individuals aged 60 or older in the country, comprising 18.7% of the overall population [2]. This underscores the profound changes in China's demographic landscape. The rapidly increasing aging rate in China poses substantial challenges for the future development of the country's medical services. Over 180 million older adults in China grapple with chronic diseases, and a staggering 75% of them contend with multiple chronic illnesses [3]. This places older individuals in a high-risk and vulnerable category, imposing considerable financial and operational burdens on China's medical and health sector.

Mobile health (mHealth) devices typically encompass mHealth programs and wearable devices [4]. Functioning as portable tools leveraging internet communication technology, these devices continuously monitor diverse physiological conditions. They have the capability to track and record users' daily lifestyle and health status data in real-time [5]. These real-time data are instrumental for users to make informed adjustments to their health behaviors, facilitated by prompt feedback on health information [6]. The utilization of mHealth devices addresses the emerging need for self-monitoring and self-management within the expanding medical service market, aligning with heightened health awareness among consumers. These devices play a pivotal role in enabling early diagnosis, intervention, clinical treatment, and monitoring of various diseases by continuously supervising vital signs in real-time. However, it is noteworthy that despite the potential benefits, mHealth devices are not widely embraced by older individuals [7]. Consequently, the robust functionalities and inherent advantages of these devices remain underutilized within this demographic group. Emerging as an inevitable outcome of the internet era and the aging society, mHealth holds substantial potential to offer a promising solution to meet the escalating demands for medical services in developing countries [8]. Recognizing that older individuals constitute the most frequent and substantial users of health services [9], it becomes imperative to cultivate a new social trend, encouraging the integration of older individuals with mHealth [10].

Prior research has demonstrated that mHealth can significantly enhance the health, well-being, and longevity of older individuals in the digital era. However, it also introduces a new social governance challenge—the digital divide among older individuals [11,12]. This divide arises from challenges in accessing or utilizing information infrastructure coupled with a lower level of digital education, resulting in difficulties for older individuals to stay abreast of social, economic, and technological advancements [13]. As outlined in the 50th Statistical Report on the Development of the Internet in China by the China Internet Network Information Center, individuals aged 60 and above constitute the predominant group of non-netizens, comprising 41.6% of this demographic [14]. A confluence of personal, family, social, and technological factors collectively contributes to the estrangement of older individuals from engaging with new media, such as the internet [15]. Research indicates that the motivation for older individuals to actively seek health information on the internet is closely tied to their interactions with family or friends [16]. Older adults primarily rely on their families for social support, and the cohesion within the family unit significantly influences their overall health status [17,18].

Family health represents a collective resource that emerges from the interconnected well-being of each family member, encompassing their health, interactions, capacities, and the family's overall physical, social, emotional, economic, and medical resources [19]. As an interdisciplinary concept, evaluating family health necessitates a thorough examination of various factors, including but not limited to family functioning, emotional support, financial resources, and access to external services [20]. Existing literature demonstrates that family support plays a pivotal role in motivating older individuals to seek medical services [21]. Additionally, family function and overall health serve as crucial indicators for assessing the mental well-being of older individuals [22]. Communication within the family, involving interactions with children, grandchildren, and peer groups, influences older individuals' inclination to adopt smart senior care solutions [23]. While numerous articles predominantly explore family health from a singular dimension [24-26], there exists a research gap concerning the specific influence of family health on older individuals' intention to adopt mHealth devices.

The evolution of mHealth is intricately linked to the technical backing of media. Media technology plays a dual role—it not only generates visual data representing health conditions detected by mHealth devices [27] but also serves as a platform for the public to exchange and share medical information. In the case of older adults, their acceptance of new health services and access to health information are influenced in distinct ways by the utilization of media devices [28,29]. A Chinese empirical analysis revealed a fundamental correlation between media use and the health level of older adults [30]. Social media communication is considered an intervention measure to alleviate the loneliness experienced by older adults, achieved by enhancing social support and contact levels, thereby fostering positive responses to emerging technologies [31,32]. Furthermore, the utilization of mobile phones and other media

XSL•FO
RenderX

significantly influences disparities in medical care. Increasing the frequency of contact and sustained use of media by older individuals can contribute to unlocking the considerable potential of mobile medical technology in the health care of older individuals [33].

In summary, there is an immediate and practical need to reduce the digital divide among older adults. The willingness of older individuals to embrace mHealth devices, as reflected in surveys, signifies their acceptance of new health technologies and, to a certain extent, their integration into the era of mHealth. Previous research on factors influencing the intention to use mHealth devices among older adults has predominantly centered on understanding the behavioral motivations and mechanisms behind users' intentions to use, emphasizing the impact of technical and social aspects on actual usage behavior [34]. Research on influencing factors has primarily delved into age, gender, education level, BMI, income, and health status, among other individual aspects [35-37]. However, there is a paucity of studies examining external environmental factors, notably the influence of family and social dynamics, particularly among the older adult population in China. A previous study indicated that family internet access enhances older adults' cognitive function and increases the frequency of media use [38]. Moreover, family support has been identified as a crucial factor aiding older adults in overcoming barriers to the utilization of mHealth services [39]. Considering the substantial impact of family factors on the proactive health information-seeking behavior of older individuals [40-43], it becomes imperative to delve deeper into the relationship between family health, media use behavior, and the older individual's intention to use mHealth devices. Additionally, exploring the mediating role of media use behavior between family health and the older individual's intention to use mHealth devices is crucial. This comprehensive investigation aims to facilitate the integration of older individuals into the "digital age" starting from the family level, foster the adoption of mHealth in the health care sector, enhance societal healthy aging, and contribute to the realization of the objectives outlined in the "Healthy China 2030 Plan."

Objectives

In this study, information pertaining to family health, media use behavior, and the intention to use mHealth devices among older adults was gathered from the Psychology and Behavior Investigation of Chinese Residents (PBICR) study. The primary objective of this study was to examine the impact of family health and media use behavior on the intention of older individuals to use mHealth devices in China. Furthermore, the study aimed to assess whether media use behavior acts as a mediating factor in the relationship between family health and the intention to use mHealth devices among older adults. Drawing upon the insights gained from the literature review, the following hypotheses were formulated: (1) family health has a direct impact on the intention to use mHealth devices among older adults; (2) family health exerts an indirect influence on the intention to use mHealth devices through the mediating factor of media use behavior; in other words, media use behavior serves as a mediator in the relationship between family health and the intention to use mHealth devices.

Methods

Study Design and Setting

The data for this study were sourced from the PBICR survey, a comprehensive cross-sectional survey initiated by the Peking University School of Public Health in 2022. The survey encompasses 148 cities spanning 23 provinces, 5 autonomous regions, and 4 municipalities directly under the central government in China. Using a multistage sampling approach, the survey uses a stratified sampling method in cities, districts, counties, and communities, and uses a quota sampling method from the community level down to the individual level.

The survey was carried out by adeptly trained investigators. Electronic questionnaires (developed previously [44]) were distributed directly to the public through one-on-one, face-to-face interactions on-site. Respondents could access the questionnaire by scanning the provided QR code. In situations where face-to-face investigations were impeded due to the constraints of the COVID-19 epidemic, investigators distributed the electronic questionnaire on a one-on-one basis through instant communication tools such as WeChat (Tencent Holdings Ltd.). Additionally, online video investigations were conducted through platforms such as Tencent Meeting (Tencent Holdings Ltd.)and WeChat video [45].

Within the PBICR survey, investigators underwent comprehensive training in sampling methods, research tools, and quality control. Only those investigators who strictly adhered to the trained survey procedures were deemed qualified and eligible to participate in the study. Furthermore, during the data processing phase, 2 researchers were designated to perform logical checks. Questionnaires that did not meet the predetermined screening criteria were excluded, ensuring the quality and reliability of the data. Additionally, in this study, further screening was implemented to eliminate questionnaires completed in an excessively short time, those containing outliers, or those with missing values.

In the 2022 PBICR survey, a total of 23,414 questionnaires were collected. Following logical checks and the elimination of outliers, 21,916 questionnaires were deemed valid. For the purposes of this study, the focus will be confined to the age group of 60 years and above. Consequently, the final sample size included 3712 older adults after sorting.

Participants

A total of 21,916 questionnaires were collected, with the screening criterion being individuals aged 60 years and above, ensuring the absence of missing data and logic errors. Following a meticulous summary and screening process, 3712 valid survey responses were obtained for analysis in this study.

The inclusion criteria for participants in this study were as follows: (1) age between 18 and 60 years old; (2) possession of the nationality of the People's Republic of China; (3) status as a Chinese permanent resident with an annual travel time of 1 month or less; (4) willing participation in the study and voluntary completion of the informed consent form; (5) ability to independently complete the questionnaire survey or do so

XSL•FO
RenderX

with the assistance of investigators; (6) capacity to comprehend the meaning of each item in the questionnaire.

The exclusion criteria for participants in this study were as follows: (1) individuals with unconsciousness or mental disorders; (2) individuals with cognitive impairment; (3) those currently participating in other similar research projects; and (4) individuals unwilling to collaborate or reluctant to participate in the study.

Ethics Approval

The study adhered to the principles outlined in the Declaration of Helsinki. Ethical approval for all experimental protocols was granted by the ethics research committees of the Health Culture Research Center of Shaanxi (approval number JKWH-2022-02) and Second Xiangya Hospital of Central South University (approval number 2022-K050). The cover page of the questionnaire provided a clear explanation of the study's purpose and assured participants of anonymity, confidentiality, and the right to refuse participation. Informed consent was obtained from all participants involved in the study.

The questionnaire cover used in this study provided a detailed explanation of the study's purpose and ensured participants of anonymity, confidentiality, and the right to refuse participation. All participants were required to voluntarily sign an informed consent form before engaging in the study. While respondents did not directly benefit from the survey, their input contributed to a more comprehensive and systematic understanding of the physical and mental health status of the public. The data from this study will be strictly managed and used in accordance with the Statistics Law of the People's Republic of China. The research data are intended for academic purposes only, and when the research findings are published, no information about individual participants will be disclosed or adversely affected.

Measurements

General Situation Survey Information

The basic demographic information of the older individuals included gender, age rank, nationality, religion, BMI rank, political status, status of occupation, education level, chronic diseases, and family type (conjugal family, core family, backbone family, and other family).

Family types were defined as follows:

- Conjugal family: a family consisting of only husband and wife.
- Core family: a family consisting of parents and unmarried children.
- Backbone family: a family consisting of parents and married children.
- Other family: other families including joint families, single-parent families, DINK (dual income, no kids) families, and single families.

Short-Form of the Family Health Scale

The assessment of family health in this study used the Chinese version of The Short-Form of the Family Health Scale (FHS-SF), developed by Crandall et al [20]. Wang et al [46] introduced the FHS-SF cross-culturally to create a Chinese version as a quantitative tool for evaluating family health issues in China. The scale comprises 10 items, encompassing 4 dimensions: family social and emotional health processes, family health lifestyle, family health resources, and family external social supports. A 5-point Likert scale was used for each item of the FHS-SF, with response options ranging from 1=strongly disagree to 5=strongly agree. Items with negative wording were scored in reverse. The final score on the scale ranged from 10 to 50, where higher scores indicated higher levels of family health. Wang et al [46] reported that the Cronbach α for the FHS-SF was .83. Additionally, the Cronbach α for the 4 subscales ranged from .70 to .90, and the retest reliability of the scale was 0.75.

In our study, the composite reliability values for the 4 dimensions were 0.912, 0.848, 0.781, and 0.806, respectively. All these values surpass the reliability threshold of 0.7. The average variance extracted values for the dimensions were 0.775, 0.736, 0.553, and 0.677, respectively, all of which exceed the threshold of 0.5. The Cronbach α of the FHS-SF was .90, and the factor loadings ranged from 0.73 to 0.90, all within an acceptable range.

Media Use Behavior Scale

The frequency of media use in this study was gauged using the Media Use Behavior Scale developed by the PBICR survey of Peking University. The scale encompasses various media channels such as newspapers, radio, television, the internet, and mobile phones. Comprising 6 items related to social contact, self-presentation, social behavior, leisure and entertainment, access to information, and business transactions, the scale uses options that signify the degree of media use frequency, ranging from "1=infrequent" to "5=frequent." The total score on the scale ranges from 6 to 30, with higher scores indicative of more frequent use of the media [45].

In this study, the composite reliability for the Media Use Behavior Scale was 0.894, and the average variance extracted was 0.585. The Cronbach α for the Media Use Behavior Scale was .89, indicating strong internal consistency. Additionally, the standardized factor loadings obtained from the validation factor analysis were above 0.50, all falling within acceptable limits.

Intention to Use mHealth Devices

The intention to use mHealth devices in this study was assessed through subjective evaluations. Participants were required to provide a numerical response ranging from 0 to 100 based on their individual subjective awareness. This formed a continuous variable, where a higher numerical value indicated a stronger intention to use mHealth devices.

Data Analysis

Continuous variables were assessed for normality using the Kolmogorov-Smirnov test and presented as the median and IQR. Categorical variables were reported in terms of frequency and percentage. Nonparametric methods were used to test the differences in characteristics related to the total score of the intention to use mHealth devices. Specifically, the Mann-Whitney *U* test was used for dichotomous variables,

XSL•FO
RenderX

while the Kruskal-Wallis *H* test was used for multicategorical variables. The partial correlation coefficient between family health scores, media use behavior scores, and intention to use mHealth devices scores was calculated using a regression model. Linear regression models were used to assess the association between family health scores and media use behavior/intention to use mHealth devices scores, both with and without adjustment for covariates. The associations between media use behavior and intention to use mHealth devices scores were also examined. The results are reported as coefficients along with 95% CIs. Covariates, determined based on previous studies and general knowledge, were included in the models for adjustment. To examine the mediating role of media use behavior scores in the association between family health scores and intention to use mHealth devices scores, we conducted a Sobel-Goodman Mediation Test. This analysis was performed while controlling for all selected covariates. The significance of the indirect effect, direct effect, and the total effect was determined using the bootstrap algorithm.

All *P* values were 2-sided, with a significance level (α) of .05 used to define statistical significance. The data were analyzed using IBM SPSS Statistics 26 and R version 4.1.3 (R Foundation).

Subgroup Analysis

Indeed, empirical studies have consistently indicated a positive association between education and health. Individuals with higher levels of education often exhibit a tendency to adopt healthier lifestyles, and their increased income may lead to greater investment in health-related expenses [47]. Furthermore, education is closely linked to varying levels of internet participation. Generally, individuals with higher educational attainment are more likely to use online platforms for accessing health-related information [48]. In diverse educational and cultural backgrounds, individuals may exhibit varying levels of concern regarding health risks, subsequently influencing their acceptance of health care technology [49]. Additionally, preliminary analysis in our study revealed significant differences in the total score of family health across different education levels (*P*<.001). Building on the established influence of education on health behavior and media use, as outlined in the existing literature and supported by our results, this paper intends to analyze education level as a subgroup. The aim is to comprehensively explore the mediating role of media use behavior among older adults with different education levels in the relationship between family health and their intention to use mHealth devices.

Results

General Characteristics

A total of 3712 older individuals aged 60 and above participated in this study, with an average age of 69.23 (SD 6.13) years. The majority of older adults (3036/3712, 81.79%) fell within the age range of 60-74 years. Basic demographic data for the 3712 older adult participants are detailed in Table 1. Among them, 1839 were males (49.54%) and 1873 were females (50.46%). The majority identified as Han nationality (3370/3712, 90.79%) and nonreligious (3416/3712, 92.03%), with the majority expressing mass political views (3151/3712, 84.89%). There were noteworthy differences in the willingness to use mHealth devices among older adults with varying political statuses, occupational statuses, and chronic disease conditions (*P*<.001). However, no significant differences were observed in the willingness to use mHealth devices among older adults with different family types (*P*=.97; Table 1).

XSL•FO
RenderX

Table 1. Characteristics of respondents (n=3712)[a].

Grouping of characteristics	All (n=3712), n (%)	The total score of the intention to use mobile health devices, median (IQR)	*P* value
Gender			.31
Male	1839 (49.54)	66.00 (36.00)	
Female	1873 (50.46)	64.00 (35.00)	
Age rank (years)			.45
60≤age≤74	3036 (81.79)	65.00 (34.00)	
75≤age≤89	662 (17.83)	64.50 (37.00)	
Age≥90	14 (0.38)	61.00 (60.00)	
Nationality			.08
Han	3370 (90.79)	64.00 (35.00)	
Minority	342 (9.21)	68.00 (34.00)	
Religion			.01
No	3416 (92.03)	65.00 (35.00)	
Yes	296 (7.97)	61.00 (35.00)	
BMI rank (kg/m^2)			.001
Normal (18.5≤BMI<24)	2355 (63.44)	66.00 (34.00)	
Underweight (BMI<18.5)	373 (10.05)	60.00 (39.00)	
Overweight (24≤BMI<28)	841 (22.66)	66.00 (36.00)	
Obese (BMI≥28)	143 (3.85)	60.00 (38.00)	
Political status			<.001
The masses	3151 (84.89)	64.00 (35.00)	
Communist Party members	512 (13.79)	69.00 (39.00)	
Probationary Party members	10 (0.27)	59.00 (36.00)	
Other parties	39 (1.05)	60.00 (38.00)	
Occupation status			<.001
On the job	98 (2.64)	70.00 (44.00)	
Retired	2071 (55.79)	67.00 (33.00)	
Rolling stone	392 (10.56)	68.00 (34.00)	
Unemployed	74 (1.99)	55.00 (41.00)	
Job waiting/jobless	1077 (29.01)	60.00 (38.00)	
Education level			.004
Primary school and below	1817 (48.95)	63.00 (37.00)	
Middle school/vocational school/high school	1426 (38.42)	65.00 (34.00)	
College and above	469 (12.63)	70.00 (35.00)	
Chronic disease			<.001
No	1552 (41.81)	69.00 (35.00)	
Yes	2160 (58.19)	62.00 (36.00)	
Family type			.97
Conjugal family	887 (23.90)	64.00 (37.00)	
Core family	325 (8.76)	64.00 (40.00)	
Backbone family	1854 (49.95)	65.50 (33.00)	

XSL•FO
RenderX

Grouping of characteristics	All (n=3712), n (%)	The total score of the intention to use mobile health devices, median (IQR)	*P* value
Other family	646 (17.40)	65.00 (37.00)	

[a]Median (IQR) was used to describe the continuous variable, whereas n (%) was used to describe the categorical variable.

Association Analysis

After adjusting for covariates, the intention to use mHealth devices exhibited a positive correlation with the total score of family health (r=0.077, P<.001) and the media use behavior score (r=0.178, P<.001). Additionally, the total score of family health was positively correlated with the media use behavior score (r=0.079, P<.001; Table 2).

Table 2. Partial correlation coefficients (r) among family health, media use behavior, and intention to use mobile health devices[a].

Correlations	Family health total score, *r* (*P* value)	Media use behavior scores, *r* (*P* value)	Intention to use mobile health devices, *r* (*P* value)
Family health total score	N/A[b]	N/A	N/A
Media use behavior scores	0.079 (<.001)	N/A	N/A
Intention to use mobile health devices	0.077 (<.001)	0.178 (<.001)	N/A

[a]The model was adjusted for various covariates, including religion, BMI rank, political status, occupational status, education degree, and chronic diseases. Variables achieved statistical significance at P≤.05.

[b]N/A: not applicable.

Relationship Between Family Health and Media Use Behavior Score/Intention to Use mHealth Devices

In the linear regression models before adjustment, the 4 dimensions of family health (ie, family socialization, family healthy lifestyle, family health resources, and family external social support) and the total score were significantly (P<.001) associated with media use behavior. Moreover, they were significantly (P<.001) related to the intention to use mHealth devices, except for family health resources (P=.15). After adjusting for gender and age rank, as well as political status, nationality, religion, BMI rank, occupation status, education level, family type, and chronic diseases, all dimensions remained statistically significant (P<.001) except for family health resources (P=.29; Table 3).

Table 3. Linear regression analysis for media use behavior and intention to use mobile health devices associated with family health[a].

Analysis element	Media use behavior scores				Intention to use mobile health devices			
	Unadjusted		Adjusted		Unadjusted		Adjusted	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Family social	.42 (0.35 to 0.49)	<.001	.32 (0.25 to 0.39)	<.001	.88 (0.56 to 1.20)	<.001	.84 (0.52 to 1.16)	<.001
Family healthy lifestyle	.55 (0.45 to 0.66)	<.001	.45 (0.36 to 0.55)	<.001	1.28 (0.81 to 1.74)	<.001	1.28 (0.81 to 1.75)	<.001
Family health resources	–.26 (–0.32 to –0.20)	<.001	–.25 (–0.31 to –0.20)	<.001	–.19 (–0.46 to 0.07)	.15	–.15 (–0.41 to 0.12)	.29
Family external social supports	.61 (0.15 to 0.71)	<.001	.53 (0.43 to 0.62)	<.001	1.43 (0.98 to 1.88)	<.001	1.45 (1.00 to 1.89)	<.001
Total	.09 (0.07 to 0.12)	<.001	.07 (0.04 to 1.00)	<.001	.30 (0.17 to 0.42)	<.001	.30 (0.18 to 0.43)	<.001

[a]Data were adjusted for gender and age rank, political status, nation, religion, BMI rank, status of occupation, education degree, family type, and chronic diseases.

Relationship Between Media Use Behavior Score and Intention to Use mHealth Devices

In the linear regression models before adjustment, media use behavior was significantly (P<.001) associated with the intention to use mHealth devices. After adjusting for gender and age rank, as well as political status, nationality, religion, BMI rank, occupation status, education level, family type, and chronic diseases, the association remained statistically significant (P<.001; Table 4).

XSL•FO
RenderX

Table 4. Linear regression analysis for intention to use mobile health devices associated with media use behavior[a].

Analysis element	Intention to use mobile health devices			
	Unadjusted		Adjusted	
	β (95% CI)	*P* value	β (95% CI)	*P* value
Media use behavior score	.94 (0.80-1.08)	<.001	.84 (0.69-0.99)	<.001

[a]Data were adjusted for gender and age rank, political status, nation, religion, BMI rank, status of occupation, education degree, family type, and chronic diseases.

Mediation Analysis

The family health total score demonstrated a positive association with the intention to use mHealth devices among older adults. Mediation analysis, including media use behavior, revealed that the relationship between the total score of family health and the intention to use mHealth devices was mediated through media use behavior. In this study, media use behavior partially mediated the association between family health and the intention to use mHealth devices. The mediating variable accounted for nearly a quarter (22.46/100) of the association when adjusting for covariates. The total score of family health was associated with media use behavior (β=.088, *P*<.001) and intention to use mHealth devices (β=.244, *P*<.001). Additionally, media use behavior was linked to the intention to use mHealth devices (β=.810, *P*<.001). The final mediation models depicting the independent variable (total score of family health), the mediating variable (media usage behavior), and the dependent variable (intention to use mHealth devices) are illustrated in Figure 1.

Figure 1. Mediation analysis.

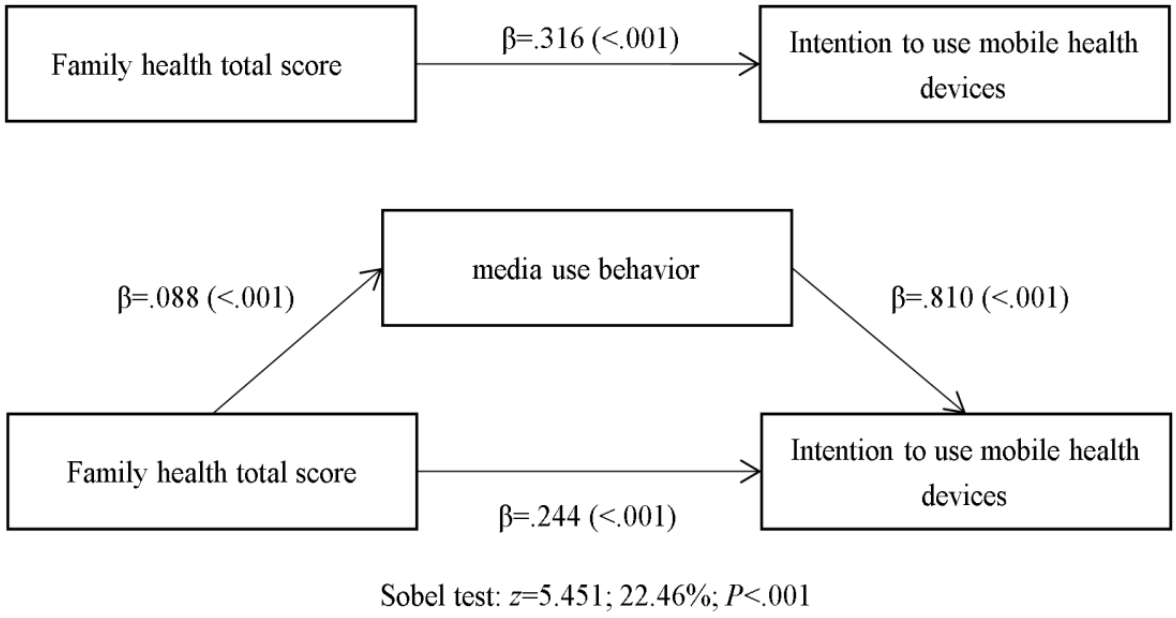

The 4 dimensions of family health were positively associated with the use of mHealth devices among older adults, except for the dimension of family health resources, which had a nonsignificant association (*P*=.72). The mediation analysis involving media use behavior indicated that the direct and total effects of family health resources were not significant (*P*=.72 and *P*=.20, respectively). Media use behavior acted as a full mediator when adjusting for covariates. Media use behavior partially mediated the relationship between family social, family healthy lifestyle, family external social support, and the intention to use mHealth devices, with mediating effects of 35.18/100, 31.78/100, and 31.33/100, respectively, under adjusted covariates (Table 5).

XSL•FO
RenderX

Table 5. Mediation analysis.

Analysis	Family social, β (*P* value)	Family healthy lifestyle, β (*P* value)	Family health resources, β (*P* value)	Family external social supports, β (*P* value)	Total, β (*P* value)
α coefficient	.401 (<.001)	.545 (<.001)	–.284 (<.001)	.616 (<.001)	.088 (<.001)
β coefficient	.790 (<.001)	.789 (<.001)	.844 (<.001)	.770 (<.001)	.810 (<.001)
Indirect effect	.317 (<.001)	.431 (.008)	–.239 (<.001)	.475 (<.001)	.071 (<.001)
Direct effect	.584 (<.001)	.925 (<.001)	.048 (.72)	1.042 (<.001)	.244 (<.001)
Total effect	.901 (<.001)	1.356 (<.001)	–.191(.20)	1.516 (<.001)	.316 (<.001)
The proportion of total effect that is mediated, n/N	35.18/100	31.78/100	125.13/100	31.33/100	22.46/100
Sobel test[a,b]	7.754 (<.001)	7.497 (<.001)	–7.284 (<.001)	–7.284 (<.001)	5.451 (<.001)

[a]The Sobel-Goodman Mediation Test was applied in adjusted models for religion, BMI rank, political status, occupation status, education level, and chronic diseases.

[b]The Sobel test was used to assess the hypothesis that the indirect role was equal to 0, adjusting for covariates such as religion, BMI rank, political status, occupation status, education level, and chronic diseases. Values reach statistical significance at $P \leq .05$.

Subgroup Analysis

Subgroup analyses based on education degrees are presented in Table 6. Among the older adult population with primary school education and below, media use behavior showed no mediating effect between the total score of family health and the intention to use mHealth devices (z=–0.942; indirect effect=–0.019, P=.35; direct effect=0.252, P=.007). Additionally, the mediating effect of media use behavior between family healthy lifestyles and the intention to use mHealth devices was not significant (z=1.953, P=.052). Media use behavior fully mediated the association between family health resources scores and intention to use mHealth devices scores in different education degrees among the older adult population: primary school and below degree older adult population (z=–5.832; indirect effect=–0.331, P<.001; direct effect=0.218, P=.29), middle school/vocational school/high school degree older adult population (z=–3.439; indirect effect=–0.136, P<.001; direct effect=–0.066, P=.76), and college and above degree older adult population (z=–2.516; indirect effect=–0.212, P=.01; direct effect=0.026, P=.93).

XSL•FO
RenderX

Table 6. Subgroup analysis of education degree of mediation models for family health associated with intention to use mobile health devices mediated by media use in older adults.

Education and parameters	Family social, β (*P* value)	Family healthy lifestyle, β (*P* value)	Family health resources, β (*P* value)	Family external social supports, β (*P* value)	Total, β (*P* value)
Primary school and below					
α coefficient	.099 (.04)	.140 (.046)	–.302 (<.001)	.246 (<.001)	–.018 (.34)
β coefficient	1.059 (<.001)	1.057 (<.001)	1.093 (<.001)	1.046 (<.001)	1.078 (<.001)
Indirect effect	.105 (.04)	.148 (.05)	–.331 (<.001)	.257 (<.001)	–.019 (.35)
Direct effect	.489 (.04)	.904 (.01)	.218 (.29)	.808 (<.05)	.252 (.007)
Total effect	.593 (.01)	1.052 (.003)	–.112 (.59)	1.065 (<.01)	.232 (.015)
Proportion of total effect that is mediated, n/N	17.71/100	14.07/100	295.54/100	24.13/100	–8.19/100
Sobel test[a]	2.052 (.04)	1.953 (.052)	–5.832 (<.001)	3.485 (<.001)	–.942 (.34)
Middle school/vocational school/high school					
α coefficient	.508 (<.001)	.718 (<.001)	–.250 (<.001)	.706 (<.001)	.123 (<.001)
β coefficient	.484 (<.001)	.477 (<.001)	.541 (<.001)	.458 (<.001)	.508 (<.001)
Indirect effect	.246 (<.001)	.343 (<.001)	–.136 (<.001)	.323 (<.001)	.063 (<.001)
Direct effect	.598 (.03)	.992 (.01)	–.066 (.76)	1.200 (.002)	.227 (.03)
Total effect	.845 (.002)	1.334 (<.001)	–.201 (.35)	1.524 (<.001)	.290 (.005)
Proportion of total effect that is mediated, n/N	29.11/100	25.71/100	67.66/100	21.19/100	21.72/100
Sobel test[a]	3.624 (<.001)	3.563 (<.001)	–3.439 (<.001)	3.455 (<.001)	3.375 (<.001)
College and above					
α coefficient	.589 (<.001)	.713 (<.001)	–.218 (.004)	1.057 (<.001)	.175 (<.001)
β coefficient	.808 (<.001)	.874 (<.001)	.974 (<.001)	.737 (<.001)	.864 (<.001)
Indirect effect	.476 (<.001)	.624 (<.001)	–.212 (.01)	.779 (<.001)	.151 (<.001)
Direct effect	1.260 (.003)	1.183 (.04)	.026 (.93)	1.926 (<.001)	.425 (.009)
Total effect	1.736 (<.001)	1.807 (.002)	–.186 (.56)	2.704 (<.001)	.576 (<.001)
Proportion of total effect that is mediated, n/N	27.42/100	34.53/100	1139.78/100	28.81/100	26.22/100
Sobel test[a]	3.409 (<.001)	3.428 (<.001)	–2.516 (.01)	3.310 (<.001)	3.193 (.001)

[a]The Sobel-Goodman Mediation Test was applied in adjusted models for religion, BMI rank, political status, status of occupation, and chronic diseases.

Discussion

Principal Findings

Previous studies have consistently demonstrated that family factors play a crucial role in influencing the frequency of media use and the acceptance of mHealth among older adults [50]. The findings of our study further confirm that family health positively contributes to increasing the willingness of older adults to use mHealth devices. Additionally, a high frequency of media use behavior emerges as a significant driver for the utilization of mHealth devices, a behavior that is profoundly influenced by the state of family health. The results align with previous research on the digital divide among older adults, indicating that those with higher family health scores tend to engage in more frequent media contact behaviors. This heightened connectivity to the internet makes them more adaptable to a big data–based mHealth environment, fostering a greater willingness to use mHealth devices. Before conducting the mediation analysis, the study also observed, through univariate analysis, that older individuals over 90 years and those who were unemployed exhibited a lower willingness to use mobile medical devices. The results confirm the existence of differences in the digital divide among age groups, especially with older age groups experiencing inequalities in social and economic support [51,52]. These disparities may further impact their access to and utilization of media devices.

In addition to the descriptive findings, this study delves into the intricate relationship between family health and the willingness to use mHealth devices, uncovering the mediating role of media use behavior. Primarily, the study supports the positive impact of media use behavior, which partially mediates the influence of overall family health levels on the intention to use mHealth devices. Furthermore, the results indicate that media use

XSL•FO
RenderX

behavior serves as a fully mediating variable in the dimension of family health resources. In essence, the findings suggest that older adults lacking family health resources completely lose their willingness to use mHealth devices, primarily due to their challenges in accessing or using media. This underscores the crucial role of family health resources in integrating older adults into the internet sphere and enabling them to benefit from mHealth technology. The study emphasizes the practical importance of addressing resource-related health inequities, with financial support from the family being identified as a critical factor in the daily lives of seniors [52]. To address the imbalance in the distribution of resources among families in different regions at the societal level, it is crucial for the government to assist socioeconomically disadvantaged older adults in gaining greater access to various devices. This can be achieved through economic empowerment initiatives and the development of policies aimed at bridging the digital divide [53].

Building upon the crucial role of media contacts in linking family health resources and the willingness to use mHealth devices among the older population, there is an opportunity to further motivate the desire for mHealth device usage. Leveraging the positive influence of family health resources to increase the frequency of media exposure can enhance the motivation of older individuals. Effective communication within the family emerges as a catalyst for improving the technology literacy and information-seeking skills of older adults [16]. Family members play a crucial role in supporting seniors to build confidence in using internet technology while alleviating their anxiety and fear of new technologies. Encouraging older adults to adapt and learn information technology, such as WeChat and health-related mobile apps, through straightforward and repeated demonstrations can be an effective strategy [54]. Additionally, family support may help mitigate the economic challenges associated with using health care services by influencing older adults' subjective perceptions of financial accessibility [55]. To address financial challenges and enhance older adults' access to technology, a comprehensive approach can be adopted. This involves leveraging both the financial support within the family and external economic resources. Encouraging family members to provide suitable financial assistance to each other, coupled with ensuring stable financial security for older individuals, can be achieved by gradually increasing pensions for retirees. This approach aims to augment the purchasing power of older adults, enabling them to acquire media devices and enhancing their ability to use technological devices in the health care sector to a greater extent.

The subgroup analysis further indicated that media use behavior did not mediate the relationship between the total family health score and the intention to use mHealth devices among older adults with primary school education or below. However, it did partially mediate the association among those with primary school education and above, aligning with the study hypothesis. Given that the older adult population with low education levels may experience relatively weak cognitive function and lack personal health literacy [56,57], the mechanisms by which they are influenced by family, social, and economic environments in the acceptance of new health technologies become more intricate. Conversely, older adults with a high school education or higher often perceive themselves as having an above-average ability to learn, making them less uncomfortable with the changing social environment brought about by technological developments [58]. Moreover, older individuals with limited education often lack access to information technology education or the ability to operate mobile devices [59]. For these individuals, exposure to media devices or mHealth devices is relatively homogeneous. Consequently, they may lack a progressive transition from regular media contact behaviors to the use of mHealth devices.

Disparities in internet participation levels due to education constitute a significant barrier hindering older adults from using media devices to access the mHealth era. To bridge the "digital divide" and enhance the effective use of mHealth devices among older individuals, it is imperative to consider implementing relevant education measures. These measures can focus on improving their ability to use smart technology, thus empowering them to navigate and benefit from the advancements in health care technology. In alignment with the comprehensive "Smart Senior Care" action plan in China [60], communities can implement health education initiatives through a blend of technology-supported learning and traditional lectures. For instance, using touchscreen tablets for courses on healthy diet and nutrition guidance can enhance the older individual's interest in the internet while imparting essential health and hygiene knowledge [61]. This approach serves to bridge the transition from traditional modes of access to mobile health care. Adopting adaptive behaviors and learning strategies can further enhance the efficiency and effectiveness of mobile health care apps [62]. In the mHealth era, the design of mHealth devices should be tailored to the cognitive abilities and mindset of older individuals. Full consideration should be given to their eHealth literacy, incorporating improvements in usability, emphasizing the responsiveness of operations, and integrating monitoring functions that align with the physical activities of older individuals [63]. Such considerations aim to enhance the overall satisfaction of older individuals with mobile health care apps [64]. Moreover, due to prevailing stereotypes about older people, digital platforms often harbor ageist mechanisms that categorize them as users uninterested in technology [65]. This results in an unfavorable digital environment for older individuals. In general, the development and application of internet technology must not overlook the realistic capacity and objective demands of older individuals [66]. Digital platforms should strive to create more inclusive algorithms and use statistical models of social digital media practices that cater to all literacy levels [65]. This may involve reducing complex and lengthy text that is difficult to understand, avoiding in-depth and complex hierarchical options, and adopting simple page designs [67] to mitigate the impact of technological differences on the accessibility of digital health care for older adults.

Strength and Limitations

This study contributes significantly to the existing literature by evaluating the connection between family health, media use behavior, and the intention to use mHealth devices among older adults, using cross-sectional data from the PBICR survey. The findings of this study support our hypothesis that media use

XSL•FO
RenderX

behavior serves as a mediator between family health status and the intention to use mHealth devices among older adults. Furthermore, a subgroup analysis based on education level revealed that the impact of family health on the willingness to use mHealth devices through media use behavior was not significant among older adults with lower education levels, indicating a nuanced mechanism at play. All of the aforementioned studies contribute to the body of research on the digital divide among older individuals.

Despite comprehensive consideration, the results of this study have several limitations. First, due to the exploratory cross-sectional design, no causal inferences can be drawn. Second, the majority of seniors included in this study were in the young-old age group (60 to 74 years old), lacking representation of the entire age spectrum of older adults and potentially neglecting variations in social background associated with age factors. Third, the results obtained in this study may be influenced by economic factors and psychological variables. As mHealth devices represent an evolving component of the health system, their development trajectory is still undergoing exploration. It is possible that various latent factors influencing the relationship between family health, media use behavior, and the intention to use mHealth devices are yet to be uncovered.

Conclusions

In conclusion, this study highlights the substantial impact of family health and media use behavior on the intention of older adults to use mHealth devices. Media use behavior acts as a mediator in the relationship between family health and the intention to use mHealth devices, with more intricate dynamics observed among older adults with lower educational levels. These findings emphasize that robust family health, particularly sufficient family health resources, plays a crucial role in enhancing the media engagement of older individuals, ultimately fostering their interest in embracing mHealth devices. The insights from this work provide valuable recommendations for bridging the gap in digital health adoption among older adults. Furthermore, encouraging teaching by family members can create a supportive environment for seniors to embrace mobile technology, while financial support can enhance their accessibility to health-related mobile devices. Additionally, developing age-specific digital education programs and media products tailored to the needs and preferences of older individuals can contribute to overcoming technological barriers and fostering a positive digital experience for older adults in the realm of mobile health care. These strategies align with the goal of promoting inclusive and user-friendly digital solutions for seniors, ensuring they can benefit from advancements in health technology.

Acknowledgments

This study was conducted with the support of data from the Psychology and Behavior Investigation of Chinese Residents (PBICR). We appreciate all the participants who showed great patience in answering the questionnaires. None of the portions of this article used generative artificial intelligence. This work was supported by the 2023 Guangdong Province Education Science Planning Project (Specialized in Higher Education; 2023GXJK252), the Science and Technology Program of Guangzhou (grant numbers 2023A04J2267 and 2024A04J02668), the Guangdong Basic and Applied Basic Research Foundation (grant number 2021A1515110743), the Health Economics Association of Guangdong Province (grant number 2023-WJMZ-51), the Student Innovation and Entrepreneurship Training Program of Guangdong Province (grant number S202312121283), the Key Laboratory of Philosophy and Social Sciences of Guangdong Higher Education Institutions for Health Policies Research and Evaluation (grant number 2015WSY0010), and the Research Base for Development of Public Health Service System of Guangzhou.

Data Availability

The data sets generated and analyzed during this study are not publicly available because the data still need to be used for other research but are available from the corresponding author on reasonable request.

Authors' Contributions

JHC, YBW, and JYC designed and conducted this study. YBW collected data. YSM, AQL, and XXY participated in the data screening. DYZ and WDY conducted data analysis. JHC and YSM wrote the first draft of the paper. JYC contributed to supervising data analysis and developing the manuscript. All authors made contributions to the critical revision of the manuscript. The authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. World population porospects 2022: summary of results. United Nations Department of Economic and Social Affairs, Population Division. 2022. URL: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf [accessed 2022-07-26]
2. Communique of the Seventh National Population Census (No.5). National Bureau of Statistics. 2021 May 11. URL: https://www.stats.gov.cn/sj/zxfb/202302/t20230203_1901085.html [accessed 2021-05-11]

XSL•FO
RenderX

3. Guiding opinions on establishing and improving the health service system for the elderly. National Health Commission/State Council Information Office. 2019 Nov 01. URL: http://www.scio.gov.cn/xwfb/bwxwfb/gbwfbh/wsjkwyh/202307/t20230703_721062.html [accessed 2024-02-11]
4. Wang X, Wu Y, Meng Z, Li J, Xu L, Sun X, et al. Willingness to use mobile health devices in the post-COVID-19 era: nationwide cross-sectional study in China. J Med Internet Res 2023 Feb 17;25:e44225 [FREE Full text] [doi: 10.2196/44225] [Medline: 36719823]
5. Tran V, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Digit Med 2019;2:53 [FREE Full text] [doi: 10.1038/s41746-019-0132-y] [Medline: 31304399]
6. Piwek L, Ellis DA, Andrews S, Joinson A. The rise of consumer health wearables: promises and barriers. PLoS Med 2016 Feb;13(2):e1001953 [FREE Full text] [doi: 10.1371/journal.pmed.1001953] [Medline: 26836780]
7. Kekade S, Hseieh C, Islam MM, Atique S, Mohammed Khalfan A, Li Y, et al. The usefulness and actual use of wearable devices among the elderly population. Comput Methods Programs Biomed 2018 Jan;153:137-159. [doi: 10.1016/j.cmpb.2017.10.008] [Medline: 29157447]
8. Li H, Zhang T, Chi H, Chen Y, Li Y, Wang J. Mobile health in China: current status and future development. Asian J Psychiatr 2014 Aug;10:101-104. [doi: 10.1016/j.ajp.2014.06.003] [Medline: 25042961]
9. Choi N. Relationship between health service use and health information technology use among older adults: analysis of the US National Health Interview Survey. J Med Internet Res 2011 Apr 20;13(2):e33 [FREE Full text] [doi: 10.2196/jmir.1753] [Medline: 21752784]
10. Mo XT, Deng ZH. An empirical study on mobile health service acceptance behavior of middle-aged and elderly users in Wuhan. Chinese Journal of Health Statistics 2015;32(02):324-327.
11. Yang JH, Liu YF. Longevity bonus of digital era: the capability and endogenous motivation of old people's digital life. Administration Reform 2022;1(1):26-36. [doi: 10.3969/j.issn.1674-7453.2022.01.005]
12. Lu JH, Wei XD. Analysis framework, concept, and pathways of digital divide governance for older adults: from the perspective of digital divide and knowledge gap theory. Population Research 2021;45(03):17-30.
13. Huang CX. Status, challenges and countermeasures of the digital divide in older adults. People's Tribune 2020;29:126-128.
14. The 50th Statistical Report on Internet Development in China. China Internet Network Information Center(CNNIC). 2022. URL: https://www.cnnic.net.cn/n4/2022/0914/c88-10226.html [accessed 2022-08-31]
15. Yang B, Jin DC. The manifestation, motivation and solution of elderly digital divide. Academic Journal of Zhongzhou 2021(12):74-80.
16. Magsamen-Conrad K, Dillon JM, Billotte Verhoff C, Faulkner SL. Online health-information seeking among older populations: family influences and the role of the medical professional. Health Commun 2019 Jul;34(8):859-871 [FREE Full text] [doi: 10.1080/10410236.2018.1439265] [Medline: 29474125]
17. Kim S, Sok SR. Relationships among the perceived health status, family support and life satisfaction of older Korean adults. Int J Nurs Pract 2012 Aug;18(4):325-331. [doi: 10.1111/j.1440-172X.2012.02050.x] [Medline: 22845631]
18. Phillips DR, Feng Z. Challenges for the aging family in the People's Republic of China. Can J Aging 2015 Sep;34(3):290-304. [doi: 10.1017/S0714980815000203] [Medline: 26300189]
19. Weiss-Laxer NS, Crandall A, Okano L, Riley AW. Building a foundation for family health measurement in national surveys: a modified Delphi expert process. Matern Child Health J 2020 Mar;24(3):259-266. [doi: 10.1007/s10995-019-02870-w] [Medline: 31912378]
20. Crandall A, Weiss-Laxer NS, Broadbent E, Holmes EK, Magnusson BM, Okano L, et al. The Family Health Scale: reliability and validity of a short- and long-form. Front Public Health 2020;8:587125 [FREE Full text] [doi: 10.3389/fpubh.2020.587125] [Medline: 33330329]
21. Yuan B, Zhang T, Li J. Family support and transport cost: understanding health service among older people from the perspective of social-ecological model. Arch Public Health 2022 Jul 19;80(1):173 [FREE Full text] [doi: 10.1186/s13690-022-00923-1] [Medline: 35850919]
22. Leung K, Chen C, Lue B, Hsu S. Social support and family functioning on psychological symptoms in elderly Chinese. Arch Gerontol Geriatr 2007;44(2):203-213. [doi: 10.1016/j.archger.2006.05.001] [Medline: 16854478]
23. Zhang Z, Mao YH, Hu YC. A study on willingness to use intelligent elderly care services from the perspective of elderly digital divide. Northwest Population Journal 2023;21:1-12.
24. Molina-Mula J, Gallo-Estrada J, González-Trujillo A. Self-perceptions and behavior of older people living alone. Int J Environ Res Public Health 2020 Nov 24;17(23):8739 [FREE Full text] [doi: 10.3390/ijerph17238739] [Medline: 33255482]
25. Nguyen T, Irizarry C, Garrett R, Downing A. Access to mobile communications by older people. Australas J Ageing 2015 Jun;34(2):E7-E12. [doi: 10.1111/ajag.12149] [Medline: 24750499]
26. Bisschop MI, Kriegsman DMW, van Tilburg TG, Penninx BWJH, van Eijk JTM, Deeg DJH. The influence of differing social ties on decline in physical functioning among older people with and without chronic diseases: the Longitudinal Aging Study Amsterdam. Aging Clin Exp Res 2003 Apr;15(2):164-173. [doi: 10.1007/BF03324496] [Medline: 12889849]
27. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. J Med Internet Res 2013 Jun 14;15(6):e120 [FREE Full text] [doi: 10.2196/jmir.2600] [Medline: 23770578]

XSL•FO
RenderX

28. Li C, Neugroschl J, Zhu CW, Aloysi A, Schimming CA, Cai D, et al. Design considerations for mobile health applications targeting older adults. J Alzheimers Dis 2021;79(1):1-8 [FREE Full text] [doi: 10.3233/JAD-200485] [Medline: 33216024]
29. Devos P, Min Jou A, De Waele G, Petrovic M. Design for personalized mobile health applications for enhanced older people participation. European Geriatric Medicine 2015 Dec;6(6):593-597. [doi: 10.1016/j.eurger.2015.10.004]
30. Wang H, Sun X, Wang R, Yang Y, Wang Y. The impact of media use on disparities in physical and mental health among the older people: an empirical analysis from China. Front Public Health 2022;10:949062 [FREE Full text] [doi: 10.3389/fpubh.2022.949062] [Medline: 36225780]
31. Zhang K, Kim K, Silverstein NM, Song Q, Burr JA. Social media communication and loneliness among older adults: the mediating roles of social support and social contact. Gerontologist 2021 Aug 13;61(6):888-896. [doi: 10.1093/geront/gnaa197] [Medline: 33284972]
32. Ma Y, Liang C, Gu D, Zhao S, Yang X, Wang X. How social media use at work affects improvement of older people's willingness to delay retirement during transfer from demographic bonus to health bonus: causal relationship empirical study. J Med Internet Res 2021 Feb 10;23(2):e18264 [FREE Full text] [doi: 10.2196/18264] [Medline: 33565983]
33. Zhang C. Smartphones and telemedicine for older people in China: opportunities and challenges. Digit Health 2022;8:20552076221133695 [FREE Full text] [doi: 10.1177/20552076221133695] [Medline: 36353695]
34. Lee SM, Lee D. Healthcare wearable devices: an analysis of key factors for continuous use intention. Serv Bus 2020 Oct 15;14(4):503-531. [doi: 10.1007/s11628-020-00428-3]
35. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 04;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]
36. Li J, Ma Q, Chan AH, Man S. Health monitoring through wearable technologies for older adults: smart wearables acceptance model. Appl Ergon 2019 Feb;75:162-169. [doi: 10.1016/j.apergo.2018.10.006] [Medline: 30509522]
37. Stühmann LM, Paprott R, Heidemann C, Baumert J, Hansen S, Zahn D, et al. Health app use and its correlates among individuals with and without type 2 diabetes: nationwide population-based survey. JMIR Diabetes 2020 May 20;5(2):e14396 [FREE Full text] [doi: 10.2196/14396] [Medline: 32432555]
38. Li Y, Han W, Hu M. Does internet access make a difference for older adults' cognition in urban China? The moderating role of living arrangements. Health Soc Care Community 2022 Jul;30(4):e909-e920. [doi: 10.1111/hsc.13493] [Medline: 34245201]
39. Mizrachi Y, Shahrabani S, Nachmani M, Hornik A. Obstacles to using online health services among adults age 50 and up and the role of family support in overcoming them. Isr J Health Policy Res 2020 Aug 21;9(1):42 [FREE Full text] [doi: 10.1186/s13584-020-00398-x] [Medline: 32825840]
40. Tu J, Shen M, Zhong J, Yuan G, Chen M. The perceptions and experiences of mobile health technology by older people in Guangzhou, China: a qualitative study. Front Public Health 2021;9:683712 [FREE Full text] [doi: 10.3389/fpubh.2021.683712] [Medline: 34249848]
41. Navabi N, Ghaffari F, Jannat-Alipoor Z. Older adults' attitudes and barriers toward the use of mobile phones. Clin Interv Aging 2016;11:1371-1378 [FREE Full text] [doi: 10.2147/CIA.S112893] [Medline: 27757025]
42. Harris T, Cook DG, Victor CR, Beighton C, Dewilde S, Carey IM. Linking survey data with computerised records to predict consulting by older people. Br J Gen Pract 2004 Dec;54(509):928-931 [FREE Full text] [Medline: 15588539]
43. Gao M, Li Y, Zhang S, Gu L, Zhang J, Li Z, et al. Does an empty nest affect elders' health? Empirical evidence from China. Int J Environ Res Public Health 2017 Apr 27;14(5):463 [FREE Full text] [doi: 10.3390/ijerph14050463] [Medline: 28448458]
44. Wenjuanxing. URL: https://www.wjx.cn/ [accessed 2024-02-07]
45. Wang Y, Kaierdebieke A, Fan S, Zhang R, Huang M, Li H, et al. Study protocol: a cross-sectional study on psychology and behavior investigation of Chinese residents, PBICR. Psychosom Med Res 2022;4(3):19. [doi: 10.53388/202219]
46. Wang F, Wu Y, Sun X, Wang D, Ming W, Sun X, et al. Reliability and validity of the Chinese version of a short form of the family health scale. BMC Prim Care 2022 May 06;23(1):108 [FREE Full text] [doi: 10.1186/s12875-022-01702-1] [Medline: 35524178]
47. Wang W, Dong Y, Liu X, Zhang L, Bai Y, Hagist S. The more educated, the healthier: evidence from rural China. Int J Environ Res Public Health 2018 Dec 13;15(12):2848 [FREE Full text] [doi: 10.3390/ijerph15122848] [Medline: 30551642]
48. Oh YS, Choi EY, Kim YS. Predictors of smartphone uses for health information seeking in the Korean elderly. Soc Work Public Health 2018;33(1):43-54. [doi: 10.1080/19371918.2017.1391150] [Medline: 29257932]
49. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 06;22(7):e17256 [FREE Full text] [doi: 10.2196/17256] [Medline: 32628122]
50. Chew TH, Chin CP, Leau Y. Untangling factors influencing social networking sites use among older adults: a literature review. Univ Access Inf Soc 2022 Mar 17;22(3):687-698. [doi: 10.1007/s10209-022-00874-y]
51. Hunsaker A, Hargittai E. A review of internet use among older adults. New Media & Society 2018 Jul 16;20(10):3937-3954. [doi: 10.1177/1461444818787348]
52. García MF, Hessel P, Rodríguez-Lesmes P. Wealth and inequality gradients for the detection and control of hypertension in older individuals in middle-income economies around 2007-2015. PLoS One 2022;17(7):e0269118 [FREE Full text] [doi: 10.1371/journal.pone.0269118] [Medline: 35802577]

XSL•FO
RenderX

53. Guo W, Chen L, Perez C. Economic status, family dependence, and health outcomes of older people in western rural China. J Gerontol Soc Work 2019 Oct;62(7):762-775. [doi: 10.1080/01634372.2019.1676860] [Medline: 31594489]
54. Shi Y, Ma D, Zhang J, Chen B. In the digital age: a systematic literature review of the e-health literacy and influencing factors among Chinese older adults. Z Gesundh Wiss 2023;31(5):679-687 [FREE Full text] [doi: 10.1007/s10389-021-01604-z] [Medline: 34104627]
55. Di X, Wang L, Yang L, Dai X. Impact of economic accessibility on realized utilization of home-based healthcare services for the older adults in China. Healthcare (Basel) 2021 Feb 17;9(2):218 [FREE Full text] [doi: 10.3390/healthcare9020218] [Medline: 33671377]
56. Liu H, Byles JE, Xu X, Zhang M, Wu X, Hall JJ. Evaluation of successful aging among older people in China: results from China health and retirement longitudinal study. Geriatr Gerontol Int 2017 Aug;17(8):1183-1190. [doi: 10.1111/ggi.12848] [Medline: 27506327]
57. Quenzel G, Vogt D, Schaeffer D. Differences in health literacy of adolescents with lower educational attainment, older people and migrants. Gesundheitswesen 2016 Nov;78(11):708-710. [doi: 10.1055/s-0042-113605] [Medline: 27649558]
58. Berkowsky RW, Sharit J, Czaja SJ. Factors predicting decisions about technology adoption among older adults. Innov Aging 2018 Jan;2(1):igy002 [FREE Full text] [doi: 10.1093/geroni/igy002] [Medline: 30480129]
59. Jun W. A study on cause analysis of digital divide among older people in Korea. Int J Environ Res Public Health 2021 Aug 14;18(16):8586 [FREE Full text] [doi: 10.3390/ijerph18168586] [Medline: 34444334]
60. National Health Commission, National Council on the Aging. Notice on the in-depth implementation of the “Smart Help for the Elderly” Action in 2022. National Health Commission. 2022. URL: http://www.nhc.gov.cn/lljks/zcwj2/202206/24a5b60b8789409c9053b38e4aab19e7.shtml [accessed 2022-06-16]
61. Chiu C, Kuo S, Lin D. Technology-embedded health education on nutrition for middle-aged and older adults living in the community. Glob Health Promot 2019 Sep;26(3):80-87. [doi: 10.1177/1757975917732351] [Medline: 29099337]
62. Yu Y, Yan XD, Z X, Zhou SL. What they gain depends on what they do: an exploratory empirical research on effective use of mobile healthcare applications. 2019 Jan Presented at: Hawaii International Conference on System Sciences; January 8-11, 2019; Maui, HI URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1477&context=hicss-52 [doi: 10.24251/HICSS.2019.483]
63. Ma Z, Gao Q, Yang M. Adoption of wearable devices by older people: changes in use behaviors and user experiences. International Journal of Human–Computer Interaction 2022 Aug 31;39(5):964-987. [doi: 10.1080/10447318.2022.2083573]
64. Zhang X, Yan X, Cao X, Sun Y, Chen H, She J. The role of perceived e-health literacy in users’ continuance intention to use mobile healthcare applications: an exploratory empirical study in China. Information Technology for Development 2017 Mar 09;24(2):198-223. [doi: 10.1080/02681102.2017.1283286]
65. Rosales A, Fernández-Ardèvol M. Ageism in the era of digital platforms. Convergence (Lond) 2020 Dec;26(5-6):1074-1087 [FREE Full text] [doi: 10.1177/1354856520930905] [Medline: 33239961]
66. Zhou X, Chen L. Digital health care in China and access for older people. The Lancet Public Health 2021 Dec;6(12):e873-e874. [doi: 10.1016/s2468-2667(21)00051-7]
67. Gao Q, Ebert D, Chen X, Ding Y. Design of a mobile social community platform for older Chinese people in Urban areas. Hum Factors Man 2012 Jun 27;25(1):66-89. [doi: 10.1002/hfm.20523]

Abbreviations

DINK: dual income, no kids
FHS-SF: Short-Form of the Family Health Scale
mHealth: mobile health
PBICR: Psychology and Behavior Investigation of Chinese Residents

Edited by T de Azevedo Cardoso; submitted 18.06.23; peer-reviewed by R Sun, X Zhang; comments to author 08.08.23; revised version received 29.08.23; accepted 28.01.24; published 19.02.24.

Please cite as:
Chang J, Mai Y, Zhang D, Yang X, Li A, Yan W, Wu Y, Chen J
Media Use Behavior Mediates the Association Between Family Health and Intention to Use Mobile Health Devices Among Older Adults: Cross-Sectional Study
J Med Internet Res 2024;26:e50012
URL: https://www.jmir.org/2024/1/e50012
doi:10.2196/50012
PMID:38373031

XSL•FO
RenderX

©Jinghui Chang, Yanshan Mai, Dayi Zhang, Xixi Yang, Anqi Li, Wende Yan, Yibo Wu, Jiangyun Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Living Lab Data of Patient Needs and Expectations for eHealth-Based Cardiac Rehabilitation in Germany and Spain From the TIMELY Study: Cross-Sectional Analysis

Boris Schmitz[1,2], PhD; Svenja Wirtz[1,2], MSc; Manuela Sestayo-Fernández[3], BSc; Hendrik Schäfer[1,2], MSc; Emma R Douma[4], MSc; Marta Alonso Vazquez[3], MSc; Violeta González-Salvado[5], MD; Mirela Habibovic[4], PhD; Dimitris Gatsios[6], PhD; Willem Johan Kop[4], PhD; Carlos Peña-Gil[5], MD; Frank Mooren[1,2], MD

[1]Department of Rehabilitation Sciences, Faculty of Health, University of Witten/Herdecke, Witten, Germany
[2]Center for Medical Rehabilitation, DRV Clinic Königsfeld, Ennepetal, Germany
[3]Health Research Institute of Santiago de Compostela, Santiago de Compostela, Spain
[4]Center of Research on Psychological Disorders and Somatic Diseases, Tilburg University, Tilburg, Netherlands
[5]Cardiology and Coronary Care Department, IDIS, CIBER CV, University Hospital of Santiago de Compostela, Santiago de Compostela, Spain
[6]Capemed, Ioannina, Greece

Corresponding Author:
Boris Schmitz, PhD
Department of Rehabilitation Sciences
Faculty of Health
University of Witten/Herdecke
Alfred-Herrhausen-Straße 50
Witten, 58455
Germany
Phone: 49 23339888 ext 156
Email: Boris.Schmitz@uni-wh.de

Abstract

Background: The use of eHealth technology in cardiac rehabilitation (CR) is a promising approach to enhance patient outcomes since adherence to healthy lifestyles and risk factor management during phase III CR maintenance is often poorly supported. However, patients' needs and expectations have not been extensively analyzed to inform the design of such eHealth solutions.

Objective: The goal of this study was to provide a detailed patient perspective on the most important functionalities to include in an eHealth solution to assist them in phase III CR maintenance.

Methods: A guided survey as part of a Living Lab approach was conducted in Germany (n=49) and Spain (n=30) involving women (16/79, 20%) and men (63/79, 80%) with coronary artery disease (mean age 57 years, SD 9 years) participating in a structured center-based CR program. The survey covered patients' perceived importance of different CR components in general, current usage of technology/technical devices, and helpfulness of the potential features of eHealth in CR. Questionnaires were used to identify personality traits (psychological flexibility, optimism/pessimism, positive/negative affect), potentially predisposing patients to acceptance of an app/monitoring devices.

Results: All the patients in this study owned a smartphone, while 30%-40% used smartwatches and fitness trackers. Patients expressed the need for an eHealth platform that is user-friendly, personalized, and easily accessible, and 71% (56/79) of the patients believed that technology could help them to maintain health goals after CR. Among the offered components, support for regular physical exercise, including updated schedules and progress documentation, was rated the highest. In addition, patients rated the availability of information on diagnosis, current medication, test results, and risk scores as (very) useful. Of note, for each item, except smoking cessation, 35%-50% of the patients indicated a high need for support to achieve their long-term health goals, suggesting the need for individualized care. No major differences were detected between Spanish and German patients (all $P>.05$) and only younger age ($P=.03$) but not sex, education level, or personality traits (all $P>.05$) were associated with the acceptance of eHealth components.

Conclusions: The patient perspectives collected in this study indicate high acceptance of personalized user-friendly eHealth platforms with remote monitoring to improve adherence to healthy lifestyles among patients with coronary artery disease during

XSL•FO
RenderX

phase III CR maintenance. The identified patient needs comprise support in physical exercise, including regular updates on personalized training recommendations. Availability of diagnoses, laboratory results, and medications, as part of a mobile electronic health record were also rated as very useful.

Trial Registration: ClinicalTrials.gov NCT05461729; https://clinicaltrials.gov/study/NCT05461729

(J Med Internet Res 2024;26:e53991) doi:10.2196/53991

KEYWORDS

eHealth; coronary artery disease; rehabilitation; mHealth; mobile health; mobile phone; app; apps; applications; personalized medicine; patient empowerment; living lab; coronary; cardiac; cardiology; heart; telehealth; telemedicine; monitoring; survey; surveys; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; acceptance; technology use; usage

Introduction

The application of eHealth technology in cardiac rehabilitation (CR) is being increasingly adopted to enhance patient outcomes. eHealth, which involves the use of digital health technologies, has the potential to facilitate CR programs to offer better, more efficient, and cost-effective care. CR is a crucial aspect of the recovery process after a cardiac event, aiming to reduce the risk of future events and improve the quality of life of patients [1,2]. The European Society of Cardiology defines CR as a multifactorial intervention with core components in patient assessment, physical activity, diet/nutritional counselling, risk factor control, patient education, psychosocial management, vocational advice, and lifestyle behavior change, including patients' adherence and self-management [3]. The CR process is typically divided into 3 stages. During phase I, patients discuss their cardiovascular risk factors and health situation in the acute clinic after a coronary intervention or surgery with their treating physician or a CR nurse. This brief phase lasts only a few days and aims to get patients moving as soon as possible, encouraging mild levels of physical activity [4]. Phase II, the reconditioning phase, occurs at inpatient or outpatient CR centers or even in the home environment with various levels of support. This multidisciplinary phase includes education on risk factors, supervised exercise training, and psychological support, with the goal of improving patients' exercise capacity, functional mobility, and self-management skills [5]. In phase III, also referred to as the maintenance phase, patients continue their care in a community or home-based setting. Phase III is the longest and least structured phase of CR, aiming at lifelong self-care with continuous risk factor management and regular physical activity to maintain the achievements made during phase II [4,6]. However, adherence to a healthy lifestyle, including regular physical activity and risk factor management, during phase III maintenance is challenging and often poorly supported [7,8]. The main reasons for suboptimal adherence to phase III CR include patient-related factors (eg, motivation) and unsustainable costs for lifelong patient support in addition to usual care by general practitioners or cardiologists [9,10]. In addition, patient barriers such as time and travel burden may add to lower adherence and uptake of maintenance programs.

Information and communication technology in the form of eHealth applications has undergone recent developments by targeting reduction of possible barriers of initiation and continued engagement in CR [11]. The advantages of eHealth include less time investment and constraints due to the absence of travel, option of continuous monitoring, and possibility for patients to manage their disease independently [12,13]. The use of eHealth technologies allows for personalization and tailoring of CR programs to individual needs, leading to higher effectiveness and improved outcomes for patients. Furthermore, eHealth applications allow for different CR aspects to be targeted independently or in a combined and synergistic manner and may have positive effects on physical activity, medication adherence, mood states, anxiety, and depression in cardiac patients [14]. However, there is no uniform eHealth platform available combining all aspects of CR for patients with cardiovascular disease over the continuum of care, including phase III maintenance. Although challenging on a technological level, user acceptance and applicability in day-to-day setting are key for implementation and success of such a solution. In addition, factors such as technological skills, trustworthiness, and overall individual attitude toward eHealth need to be considered [15-17].

Based on this background, the goal of this study was to provide a detailed description of the patient perspective on the most important aspects to be included in an eHealth solution to assist phase III CR maintenance. This report is part of the multistakeholder project TIMELY, which aims at developing a personalized eHealth platform to assist patients over the continuum of the disease according to recent coronary artery disease (CAD) guidelines [18]. TIMELY employs artificial intelligence–powered CR components in a patient app connected with a patient management platform and decision support tools for case managers and clinicians. Additionally, artificial intelligence–powered conversational agents (chatbots) will be provided to engage in motivational conversations with patients based on behavior change techniques with the goal of optimizing program and exercise adherence. The development of the TIMELY eHealth solution is guided by a Living Lab approach that allows researchers to co-design innovations such as TIMELY with patients in a real-life context to increase acceptance [19]. Multiple feedback loops are included at pivotal developing stages, incorporating patients and clinicians in a modified Delphi approach [20,21]. Within the TIMELY prospective study, patients are equipped with different devices as part of the envisioned solution, including a long-term 3-channel electrocardiogram (ECG) patch, a hemodynamic monitor for blood pressure measurement and pulse wave analysis, and a wrist-worn activity tracker. This report describes

XSL•FO
RenderX

patients' needs and expectations for eHealth-based CR collected within the TIMELY Living Lab in CR centers from Germany and Spain.

Methods

Approach and Participants

To characterize patients' needs and expectations for an eHealth-based phase III CR maintenance system, a guided survey was conducted at medical rehabilitation centers Clinic Königsfeld, Germany, and University Hospital of Santiago de Compostela, Spain, between July 2021 and March 2022, aiming at a representative sample of ~80 participants. Patients were asked to participate during their inpatient (Germany) or outpatient (Spain) CR program, and participants were recruited consecutively without further selection. Patients diagnosed with CAD were eligible while participating in a structured center-based CR program.

Ethics Approval

This study complied with the Helsinki Declaration "Ethical Principles for Medical Research Involving Human Subjects" and was approved by the ethics committee of University Witten/Herdecke (115/2020) and Servizo Galego de Saúde (2021/190). All participants gave their written informed consent before participating in this study. This study is part of the TIMELY observational trial (ClinicalTrials.gov: NCT05461729), which aims to characterize the progress of patients with CAD during phase II and phase III CR.

Patients' Characteristics

Patients' anthropometric and clinical data, including severity of CAD, type of intervention, and comorbidities (rated using the D'Hoore comorbidity index [22]) were extracted from electronic health records by clinical personnel. Patients' highest level of education was documented and specified by country. Hauptschule and Educación primaria were defined as primary, Realschule and Educación secundaria obligatoria or vocational training as secondary, and Abitur or Bachillerato as tertiary education in Germany (DE) and Spain (ES), respectively. A university degree was classified as the highest educational category. For comparability and due to differing educational systems in Germany and Spain, the level of education was categorized as "lower/equal to high school" (first two levels) or "higher than high school" (all other higher levels).

Interview-Based Survey

This survey was developed with experts from a clinical and theoretical perspective by using the Delphi method until consensus was reached. The survey (20 items) was composed of 3 parts: (1) importance of different CR components in general, (2) digital literacy and current usage of technology/technical devices, and (3) helpfulness of the potential features of eHealth in CR (Multimedia Appendix 1). Closed questions were used with a list of provided answers rated on a 5-point Likert scale (1=unimportant/not useful; 5=very important/very useful). A filter question was used, which optionally exempted participants who indicated that they would never use an eHealth platform linked to devices. These participants were asked for their reasons for refusing to use an eHealth platform. The survey was pretested with selected patients in Clinic Königsfeld, and adaptations for wordings were made, where necessary. The final version of the survey was translated to German (SW and BS) and Spanish (MSF and MA) by at least 2 researchers for each translation. The survey was conducted by researchers of the local rehabilitation center. Questions were read to the patients, and further explanation was provided if needed. Investigators documented the answers by using a paper-pencil version or an electronic version of the survey (Multimedia Appendix 1).

Questionnaires

In a subset of 40 German patients with CAD, questionnaires were used to identify personal traits potentially predisposing patients for acceptance of an app or monitoring devices to document the progress of CR (ie, questions Q12 and Q13 of the survey). Psychological flexibility was assessed using the Acceptance and Action Questionnaire version 2 (AAQ-2) [23], and the Revised Life Orientation Test (LOT-R) [24] was used to identify patients' optimism/pessimism. The Type D scale for social inhibition (DS-14) [25] was used to assess negative affectivity, social inhibition, and type D personality. In addition, the Positive and Negative Affect Schedule (PANAS) was applied [26].

Statistical Analysis

Statistical analyses were performed using the open access program Jamovi (version 2.2.2, The Jamovi project) and SPSS (version 29, IBM Corp). Data are presented as mean and standard deviation, median and range for the Likert rating scales, or n (%) as indicated. Normality was tested using Shapiro-Wilks test. Between group differences were tested using independent 2-sided *t*-test or analysis of variance. Nonparametric tests were used to investigate group differences in Likert scale data (Mann Whitney *U* and Kruskal Wallis test). The associations of sex, age, education level as well as different psychological constructs with openness to using eHealth were analyzed between groups (general willingness [yes/maybe] and patients not willing to use eHealth [no]) by using chi-square test or Mann Whitney *U* test as indicated. To analyze the combined predictive values of multiple patients' characteristics on eHealth acceptance, we used multivariate linear regression and naïve Bayes classification. The statistical significance level was set at $P<.05$.

Results

Patients' Characteristics

Seventy-nine patients participated in the guided survey (Germany, n=49; Spain, n=30; 16/79, 20% females). The mean age (in years) of the patients was 57 (SD 7; range 37-79) (Table 1). In Germany, our sample population was comparable in terms of sex and age to patients with CAD in general (registry data) [27] and to patients with CAD undergoing CR in particular (mean 54.9, SD 7.0 years, in-house data). Further comparison of the study sample to German patients with CAD undergoing CR showed considerable similarity also in terms of ST-elevation myocardial infarction/non–ST-elevation myocardial infarction (~75%), number of affected vessels (1 vessel disease, ~30%-40%), and performed intervention (bypass, ~20%; all

XSL•FO
RenderX

in-house data). For Spain, our study sample was comparable to patients with CAD undergoing CR in terms of age (~61 years), ST-elevation myocardial infarction/non–ST-elevation myocardial infarction (~85%), number of affected vessels (1 vessel disease, ~60%), and performed intervention (bypass, ~5%; all in-house data, region Galicia). Overall, in terms of the education level, 87% (69/79) of the participants were ≤high school and 13% (10/79) were >high school (Table 1). Comparisons between countries suggested good comparability even though the age (in years) of the Spanish participants (mean 62, SD 10) was higher than that of the German participants (mean 56, SD 6; $P<.001$), which was associated with a significantly higher burden of comorbidities (median ES 2.3, IQR 1-8; median DE 1.6, IQR 0-7; $P=.03$). The percentage of former smokers among patients with CAD in Germany was significantly higher than that in Spain (27/49, 55% vs 7/30, 24%; $P<.001$). Overall, 30% (24/79) of the included participants were active smokers. Of the 79 participants, >85% (67/79) indicated that they (highly) appreciated being involved in the planning of a future eHealth solution.

XSL•FO
RenderX

Table 1. Characteristics of the patients.

	Overall (N=79)	Germany (n=49)	Spain (n=30)	*P* value[a]
Age (years), mean (SD)	57 (9)	56 (6)	62 (10)	.001
Sex, n (%)				>.99
Female	16 (20)	10 (20)	6 (20)	
Male	63 (80)	39 (80)	24 (80)	
Height (cm), mean (SD)	172.3 (10.0)	175.9 (8.8)	166.7 (9.4)	<.001
Weight (kg), mean (SD)	86.9 (14.6)	88.3 (13.6)	83.2 (18.0)	.25
BMI (kg/m^2), mean (SD)	29.2 (4.9)	28.6 (4.5)	30.2 (5.4)	.18
Coronary artery disease, n (%)				
One-vessel disease	33 (42)	14 (29)	19 (63)	.02
Two-vessel disease	26 (33)	18 (37)	8 (27)[b]	>.99
Three-vessel disease	20 (25)	17 (13)	3 (10)[b]	.36
ST-elevation myocardial infarction/non–ST-elevation myocardial infarction	63 (80)	36 (73)	27 (90)	.09
Treatment, n (%)				
Percutaneous coronary intervention performed	67 (85)	39 (80)	28 (93)	.12
Bypass performed	13 (16)	11 (22)	2 (7)	.12
Left ventricular ejection fraction, n (%)				
Normal (>50%)	56 (71)	35 (71)	21 (70)	>.99
Slightly reduced (41%-50%)	16 (20)	10 (21)[b]	6 (20)[b]	>.99
Moderately reduced (31%-40%)	7 (9)	4 (8)[b]	3 (10)[b]	>.99
Severely reduced (≤30%)	0 (0)	0 (0)[b]	0 (0)[b]	>.99
Comorbidity index, median (range)[c]	1.9 (0-8)	1.6 (0-7)	2.3 (1-8)	.03
Education level, n (%)				
≤High school	69 (87)	45 (92)	24 (80)	>.99
>High school	10 (13)	4 (8)	6 (20)	>.99
Primary[d]	28 (37)	18 (39)	10 (33)	>.99
Secondary[e]	25 (33)	16 (35)	9 (30)	>.99
Abitur/Bachillerato[f]	13 (17)	8 (17)	5 (17)	>.99
University	10 (13)	4 (9)	6 (20)	>.99
Smoking status, n (%)				
Never	21 (27)	12 (24)	9 (30)	>.99
Former	34 (44)	27 (55)[b]	7 (24)	.001
Active	24 (30)	10 (20)	14 (47)	.57

[a]*P* values were calculated using independent 2-sided *t* test (nonnormally distributed data were analyzed by Mann Whitney *U* test) and analysis of variance (nonnormally distributed variables were analyzed by Kruskal-Wallis rank sum test).

[b]*P*<.05 for within-group comparison.

[c]Comorbidity index was calculated according to the modified D'Hoore comorbidity index.

[d]Primary education is known as Hauptschule in Germany (DE) and educación primaria in Spain (ES).

[e]Secondary education is known as Realschule in Germany (DE) and educación secundaria obligatoria or vocational training in Spain (ES).

[f]Tertiary education is known as Abitur in Germany and Bachillerato in Spain.

XSL•FO
RenderX

Digital Literacy and Current Usage of Technology

For the assessment of the use of technology among patients and their associated digital literacy, participants were asked what devices they owned, for which purpose the devices were used, and how experienced they were with health/fitness apps. All patients owned a smartphone, while a significantly lower proportion of Spanish patients owned a tablet (ES: 11/30, 37%; DE: 34/49, 69%; *P*=.005) (Figure 1). The majority of patients also owned a notebook or PC (ES: 18/30, 60%; DE: 25/30, 84%). Smartwatches (ES: 10/30, 33%; DE: 16/49, 33%) and fitness trackers (ES: 9/30, 30%; DE: 21/49, 43%) were used by a significant proportion of the participants with no differences between centers. Although smartphone, tablet, and notebook/PC were predominantly used for communication and information by the patients, a difference for smartwatch/fitness trackers was recorded in that up to 40% (12/30) of the Spanish patients used those devices also for entertainment. This was only reported by 6% (3/49) of the German patients (*P*=.06). Instead, 50% (25/49) of the German patients used wearables and associated apps for documentation (including physical activity), which was only reported by 20% (6/30) of the Spanish patients (*P*>.05). In terms of experience with automatic blood pressure monitors, 62% (49/79) of the patients reported their level of experience as "experienced" to "very experienced," and 29% (23/79) and 13% (10/79) reported this level of experience for fitness trackers and health apps, respectively (Multimedia Appendix 1). Of note, more than 40% (32/79) of the patients reported at least some experience with health or fitness apps.

Figure 1. Current usage of technology on information and communication devices for different purposes, including health and fitness applications, by patients with coronary artery disease undergoing cardiac rehabilitation. DE: Germany; ES: Spain.

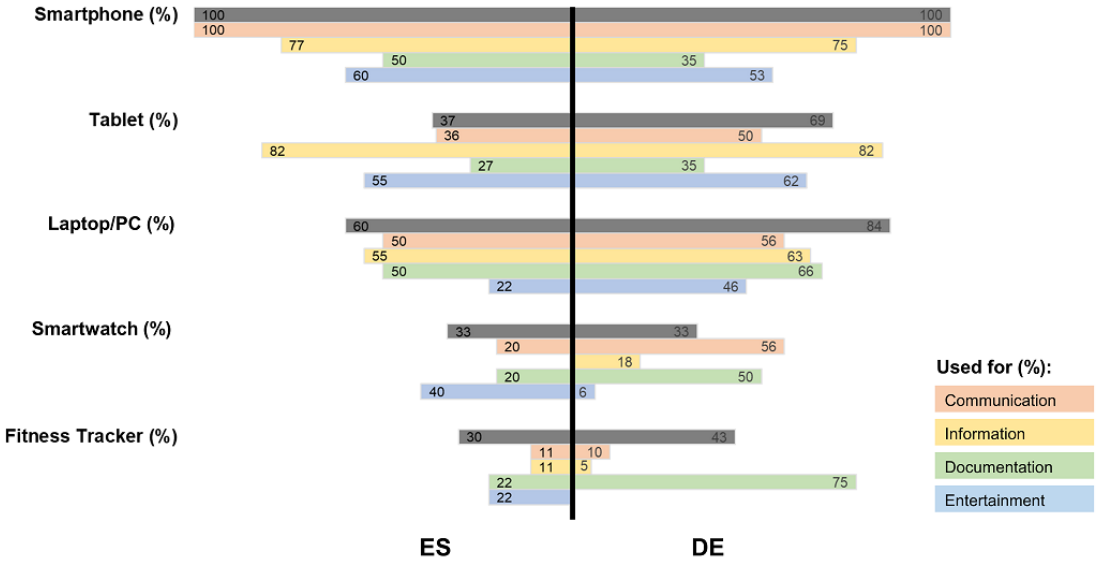

Rating of CR Components

To assess how patients rated the importance of different CR components for disease management, we recorded their feedback on separate aspects of CR (using 5-point rating scales). Patients' overall rating of the importance of CR components along the continuum of care for risk reduction was very high, including regular physical exercise (median 5, IQR 3-5), healthy diet (median 5, IQR 3-5), stress management (median 5, IQR 1-5), smoking cessation (median 5, IQR 1-5), optimal medication (median 5, IQR 3-5), motivation for lifestyle changes (median 5, IQR 3-5), and overall risk factor management (median 5, IQR 2-5), with no significant difference between the 2 centers. Patients also rated their individual need for support during phase III CR maintenance in the beforementioned areas, revealing large interindividual differences with all items ranging from 1 to 5. In general, patients expressed a high need for support for regular physical exercise (median 4, range 1-5), less need for support for smoking cessation (median 1, range 1-5; only active smokers were asked), and less support for healthy diet (median 3, range 1-5), stress management (median 3, range 1-5), medication (median 3, range 1-5), motivation for lifestyle changes (median 3, range 1-5), and risk factor management (median 3, range 1-5). Of note, for each item except from smoking cessation, 35%-50% of the patients indicated a high need for support (≥4) to achieve their long-term health goals, suggesting a need for individualized care. The subgroup of patients expressing low perceived smoking cessation support needs was analyzed further to investigate if it includes patients with high-risk phenotypes. However, this analysis did not suggest an elevated risk for these patients, as age, sex, BMI, disease severity (bypass performed [yes/no]), and comorbidity index were similar to those of the group of smokers indicating need for smoking cessation support.

Rating of eHealth Components to Assist in Phase III CR Maintenance

Overall, 71% (56/79) of the patients reported that they considered technology, including mobile apps, to be helpful in maintaining health goals after phase II CR. To investigate the specific needs and expectations for an eHealth system to assist in phase III CR maintenance, we asked patients about the features that would be the most helpful for reaching their individual health goals if they were free to choose from a predefined set of options. The presented features were selected by the TIMELY investigators involving cardiologists, rehabilitation experts, behavioral change experts, sports scientists, and by considering recent literature on eHealth in CR

XSL•FO
RenderX

[6]. Selected features were grouped into 3 categories for the presentation of results, including exercise-related features, clinical/medical components, and motivational/other features (Figure 2) and were analyzed for differences between nationality, age groups, and men versus women. No significant differences between nationalities were detected for exercise-related features or medical-related entities. In the domain of other CR components, overall progression documentation was significantly rated as more useful/more needed by German patients (median 5, range 1-5) than by Spanish patients (median 4, range 1-5; $P<.001$). German patients also rated "individual feedback of a real person" more useful than Spanish patients (median 5, range 1-5 vs median 4, range 3-5; $P=.005$, respectively). With respect to motivational features, Spanish patients rated the possibility to "share progress with friends and family" as more useful than German patients (median 4, range 1-5 vs median 2, range 1-5]; $P=.02$, respectively). When asked about the preferred frequency for motivational messages, only 5% of the patients answered "several times a day." Approximately 27% (21/79) preferred to receive messages once a day, 26% (20/79) every other day, and 9% (7/79) did not want to receive messages. Approximately 32% (25/79) indicated that they would prefer a flexible schedule for messages. Of note, no differences in preference for any suggested features were detected between women and men or among age groups. However, the score for most items ranged from 1 to 5, highlighting that perceived usefulness of potential eHealth features differs substantially between individuals.

XSL•FO
RenderX

Figure 2. Results of the patient survey: rating of cardiac rehabilitation components. Patient-reported perceived importance of different cardiac rehabilitation components to assist during phase III cardiac rehabilitation maintenance. (A) Components to support regular physical activity, (B) components informing on diagnosis, clinical/laboratory parameters, and risk score, (C) other cardiac rehabilitation–related topics/functions. 5-point Likert scale (1=unimportant/not useful; 5=very important/very useful). DE: Germany; Doc.: documentation; ECG: electrocardiogram; ES: Spain.

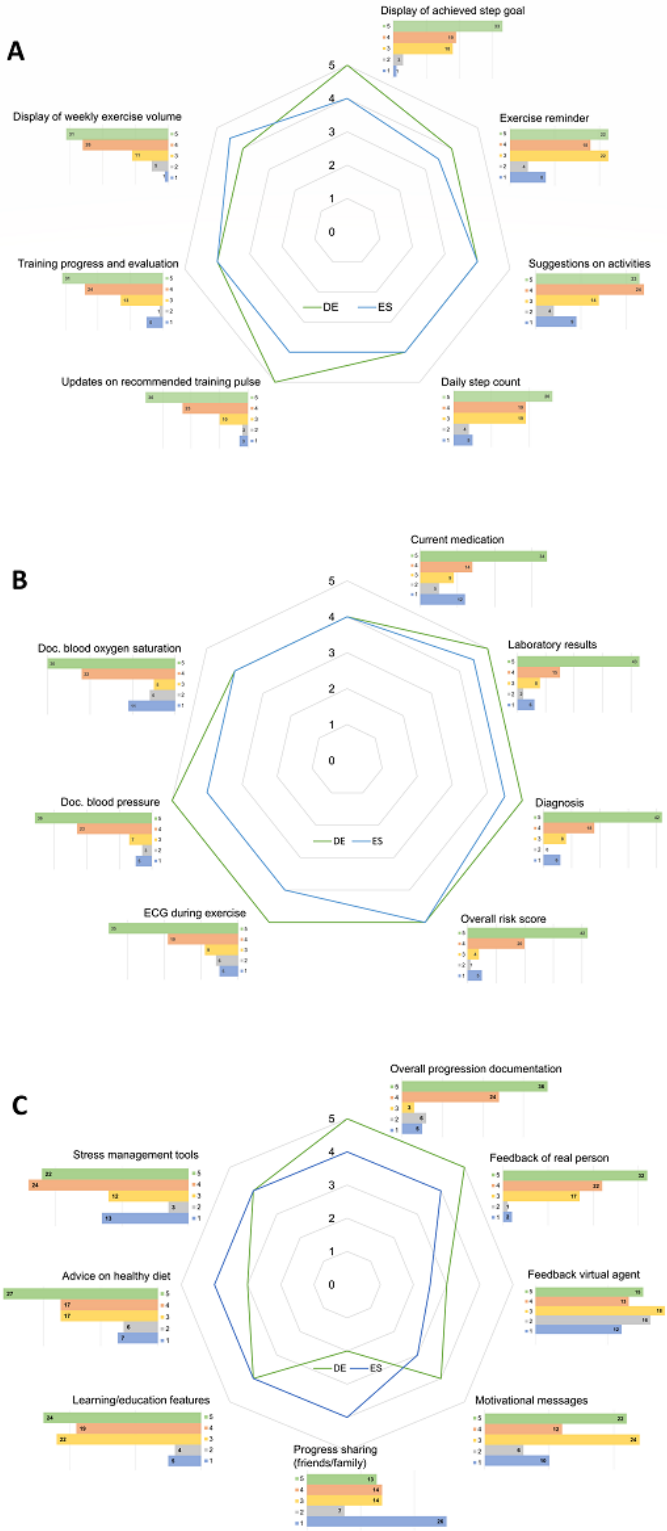

Factors Associated With Acceptance of eHealth in CR Maintenance

To investigate the factors associated with the acceptance of eHealth, we used questionnaires to analyze factors such as sex, age, clinical data, educational as well as psychological factors. Questionnaires involved LOT-R for optimism/pessimism, AAQ-2 for psychological flexibility, DS-14 for social inhibition, and PANAS for positive/negative affectivity. Education level was not associated with the acceptance of eHealth components (Table 2). No differences were observed with regard to acceptance between women and men, but younger age was significantly associated with more acceptance of monitoring devices (P=.03), while only a tendency was seen for willingness

XSL•FO
RenderX

to use a mobile app (*P*=.11). Of note, only 6% (3/49) of the patients who accepted eHealth indicated they would likely not use eHealth components because of privacy concerns, and 8% (4/49) of the patients did not like the idea of being monitored. Although multivariate linear regression analysis did not identify a combination of factors associated with eHealth acceptance, naïve Bayes classification suggested that eHealth acceptance may potentially be predicted based on younger age, a lower AAQ-2 score indicating psychological flexibility, and the index event (having experienced myocardial infarction). Willingness to use a mobile app was predicted with an overall accuracy of 97.9% (using age and AAQ-2), and the acceptance of monitoring devices was predicted with an overall accuracy of 91.7% (using age, AAQ-2, and myocardial infarction). However, validation in an independent data set was not performed.

Table 2. Predictors of eHealth acceptance among patients with coronary artery disease[a].

	Would use mobile app for support			Would use devices for monitoring		
	Yes/maybe (n=42)	No (n=7)	*P* value	Yes/maybe (n=40)	No (n=9)	*P* value
Sex, n (%)			.62			>.99
Female	8 (80)	2 (20)		8 (80)	2 (20)	
Male	34 (87)	5 (13)		32 (82)	7 (18)	
Age (years), n (%)			.11			.03
≥57 years	21 (78)	6 (22)		19 (70)	8 (30)	
<57 years	21 (95)	1 (5)		19 (70)	1 (5)	
Education[b], n (%)			>.99			.66
≥High school	11 (92)	1 (8)		11 (92)	1 (8)	
<High school	29 (85)	5 (15)		28 (76)	6 (24)	
LOT-R[c], median (IQR)						
Optimism	4 (0-9)	4 (2-6)	.79	4 (0-9)	4 (2-6)	.61
Pessimism	7 (2-12)	4 (3-7)	.11	6 (2-12)	7 (3-9)	.76
AAQ-2[d] flexibility, median (IQR)	14 (7-36)	19 (7-27)	.92	10 (7-36)	15 (7-27)	.23
DS-14[e], median (IQR)						
Negative affectivity	10 (1-28)	13 (4-19)	.69	11 (1-28)	10 (2-21)	.80
Social inhibition	12 (0-20)	11 (7-16)	.81	12 (0-20)	8 (7-18)	.23
PANAS[f], median (IQR)						
Positive affect	3 (1-4)	3 (2-5)	.69	3 (1-4)	3 (2-5)	.77
Negative affect	2 (1-4)	2 (1-4)	.47	2 (1-4)	2 (1-4)	.47

[a]Data are given as n (%) and median and range. Patients were asked if they would use a mobile app for their cardiac rehabilitation maintenance support and if they would use monitoring devices (eg, blood pressure monitor, electrocardiogram, activity tracker) during maintenance. Options provided were yes/maybe or no. Between-group comparison was performed using chi-square test or Mann-Whitney *U* test.

[b]Three missing. Only German patients (n=40) were involved.

[c]LOT-R: Revised Life Orientation Test; 2 dimensions; range 0-12 (higher = larger optimism/pessimism).

[d]AAQ-2: Acceptance and Action Questionnaire version 2; range 7-49 (higher = greater psychological inflexibility).

[e]DS-14: Type D scale for social inhibition; 2 dimensions; range 0-28 (higher = larger negative affectivity/social inhibition).

[f]PANAS: Positive and Negative Affect Schedule; 2 dimensions; range 0-10 (higher = larger affect).

Discussion

Principal Findings

This study aimed to define patients' needs and expectations for eHealth-based CR to assist them during the lifelong maintenance phase. A Living Lab approach was used for German and Spanish patients with CAD to characterize their use of technology, their preferences and rating of importance for different components of a future eHealth solution for CR maintenance, as well as their general willingness to use eHealth. In brief, our main findings are (1) patients with CAD appreciated being involved in the planning of a future eHealth system, and they had sufficient levels of digital literacy, (2) patients rated the importance of CR components along the continuum of care for risk reduction as very high, (3) 71% (56/79) of the patients expected that technology could help them to maintain health goals after center-based CR, and (4) a large intraindividual heterogeneity was detected in terms of reported needs and perceived usefulness for different eHealth components.

XSL•FO
RenderX

CAD is a chronic disease, necessitating innovative approaches for effective management and support over the lifelong maintenance phase after successful intervention and rehabilitation [1-3]. In recent years, telemedicine and eHealth solutions have emerged as promising tools for improving the care of patients with CAD [6]. In this regard, eHealth has already been shown to be an effective alternative to phase II CR, and a recent meta-analysis suggested that telehealth-based phase II CR may be even superior to center-based programs at least for enhancing physical activity levels [28-30]. In addition, eHealth may have the potential to involve a large number of patients since it may also be an option for patients who cannot or do not want to attend a center-based CR. In terms of cost efficiency, Frederix et al [30] estimated that a 6-month internet-based program consisting of exercise training with telemonitoring support, text messages, and web service can be cost-efficient for up to 2 years after the end of the intervention [30]. However, the development of eHealth solutions tailored for patients with CAD requires a dynamic and patient-centered approach since low user acceptance is one of the largest barriers for success of these solutions. The European Society of Cardiology e-Cardiology Working Group reported that digital health developments are often technically driven and not based on the needs and expectations of patients, thereby calling for cocreation with patient involvement in the design [15]. The European Society of Cardiology position paper strongly emphasized that patient-related barriers and user characteristics may hinder the large-scale deployment of eHealth services. Thus, the TIMELY project includes a Living Lab as means to involve patients and patient organizations, and our analyses reflect part of this patient-centered approach.

Per definition, Living Labs represent open innovation ecosystems to cocreate, assess, and refine innovative (technical) solutions [19]. To achieve a user-centric design, Living Labs prioritize the engagement of patients together with health care professionals to ensure that the resulting applications align with the needs, preferences, and challenges faced by the specific needs of a patient group. It is however important to place Living Labs in authentic settings, as implemented in this study, where patients with CAD undergoing center-based phase II CR are involved. These patients had received comprehensive information on the etiology and treatment of their disease as well as lifestyle factors that modify CAD. The majority of the involved patients indicated that they liked the approach and appreciated being involved in the conception and development of an eHealth solution to assist them during the maintenance phase even though some indicated that too much effort might keep them from using such a solution. In terms of predictors of eHealth use, previous research on sociodemographic factors among US adult internet users suggested that patients with lower education levels had lower odds of using certain features, including web-based tracking of personal health information, using a website to support physical activity, or downloading health information to a mobile device [31]. That study also indicated that being female was a predictor of eHealth use across health care and user-generated content, while age influenced health information–seeking [31]. In comparison, our data also suggest that younger age was associated with the indicated acceptance of technology, but women were as likely as men to accept eHealth for managing their disease, and the education level was not identified as a predictor. These findings might be based on the fact that smartphones, device hardware, and mobile apps are rapidly advancing, and daily exposure lowers the barriers for patients to use technology [32]. Although our study was performed among a selected group of patients with CAD participating in a prospective study, it is interesting to compare our cohort also in terms of the necessary hardware availability, that is, smartphone ownership in this patient group in general. Between 2019 and 2020, a large cross-sectional study among cardiac inpatients in Australia reported a high frequency of smartphone ownership (85%-89%) among patients aged 50-69 years and lower ownership (~60%) in patients aged 70-79 years [33]. In our sample (mean age 57 years, SD 9 years), every patient owned a smartphone and one-third also used activity trackers/smartwatches, which might also be explained by the differences between countries (Australia vs Germany/Spain). Percentage of technology ownership as well as usage and expectations for eHealth were not different between Germany and Spain, even though the Spanish population was significantly older (P=.001) and clinical characteristics differed to some extent. Further, CR in Spain is based on outpatient care, which, while equally effective in terms of reaching the main CR outcomes, could have affected the estimated need for eHealth in this population. Of the analyzed psychological factors, only psychological flexibility showed some predictive value for eHealth acceptance. This result partly contradicts previous findings among older (>60 years) residents of Hong Kong, wherein optimism was significantly related to perceived eHealth usefulness [34]. To what extent these differences are caused by differences in age or cultural background warrant further investigations.

State-of-the-art digital health care programs face numerous technical and interoperability hurdles that make implementation difficult. This includes transmitting physiological measurements from ECGs and blood pressure monitors as well as data from activity trackers and other wearables to a centralized platform. Respective solutions rely on wireless networks; different hardware, software, and algorithms for capturing and processing data; as well as connected dashboards. Challenges include system reliability, data quality, interoperability, and overall, the highest level of data security. We have not asked the involved patients about their opinions on system availability and stability, as these aspects as well as data security and privacy need to meet the highest standards as conditio sine qua non when providing eHealth to patients. However, information regarding these aspects needs to be provided to patients in sufficient detail, since privacy-related concerns represent considerable barriers [15,35]. These technical requirements and interdependencies result in high costs for any eHealth solution targeting to improve patients' self-care. Foreseen functionalities should thus not only be based on current guidelines but should be aligned with patient needs and expectations. This study shows that patients with CAD expected considerable merit in the documentation and availability of their diagnosis, laboratory results, and current medication—all details that would be part of an electronic health record. Patients also showed interest in their overall risk score, which TIMELY will base on a biomarker score to predict the 10-year mortality risk [36,37]. The majority of patients rated

XSL•FO
RenderX

the usefulness of blood pressure and ECG monitors as high or very high. Functionalities related to support daily physical activities and physical exercise were perceived as (very) useful, with most patients indicating a high need for progress documentation and regular updates on personalized training recommendations. This observation is relevant since commercial activity trackers have been reported to significantly increase the daily step count and aerobic capacity in patients undergoing CR [38,39], and a considerable number of patients were already relying on commercial solutions, which, however, do not always provide the necessary level of data protection and have not been tested sufficiently in patient populations. Functionalities related to other important parts of CR, including smoking cessation, stress management, advice on heart-healthy eating, as well as self-education, were perceived as less useful or rated neutral, likely depending on the individual perceived needs of the patients. This aspect was pronounced for smoking cessation, which was perceived as an important part of CR, but 50% of the smokers indicated that they did not want support with this health-related aspect.

Limitations

Although reporting on 2 samples of participants undergoing CR from Germany and Spain with cultural and socioeconomic differences is a strength of this study, this report may be affected by the potential study selection bias since patients participating in scientific research studies differ in terms of motivational aspects. However, our sample population did not differ with respect to the sociodemographic characteristics of the samples of patients with CAD undergoing CR who were analyzed in previous reports [22]. It should be noted that health literacy, a central factor in eHealth usage and a pivotal determinant of health in general, is a complex construct and was not assessed in all dimensions in our study population. The results of naïve Bayes classification should be interpreted with care since validation in an independent data set was not performed. The timepoint and situation of this survey may also have affected the results since patients may answer differently when asked in their home environment or with greater time interval after an acute event. Focus groups may allow for more and detailed information on the reasoning underlying the reported answers to this guided survey, and the results of focus groups within TIMELY will be reported elsewhere.

Conclusion

This survey involving patients undergoing CR in Germany and Spain revealed that eHealth for CR maintenance should emphasize on support for regular physical activity and physical exercise, including patient feedback on achievements and renewal of training recommendations. Devices for physiological measurements, including blood pressure and ECG monitors, were considered useful, and most patients expressed a need for the documentation of diagnosis, medication, and laboratory results in terms of an electronic health record. In general, the patients who took part in this project showed a sufficient level of digital literacy and current usage of technology to make good use of even more advanced eHealth solutions. Although only minor differences were observed among Spanish and German patients as well as between female and male patients and educational status did not appear to be a contributing factor, it is crucial to note substantial variability in patients' individual needs and expectations. Consequently, eHealth solutions should prioritize personalization to enhance user acceptance. Next steps of the TIMELY Living Lab will involve analyses of details on the implementation of the individual CR functionalities and feedback on the mobile app design.

Acknowledgments

We thank all the patients involved in this study for participating and appreciate the help of our colleagues in answering the Delphi questions to develop the survey used in this project. BS, FM, MH, CP-G, and WJK received funding from the European Commission within the H2020 framework (project TIMELY, grant agreement number 101017424).

Data Availability

The data generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

BS, SW, and FM designed this study. SW, MSF, HS, MAV, and VG-S performed the survey and collected the data. SW, MSF, and BS analyzed the data. BS, WJK, and MH interpreted the results. BS, SW, and ERD wrote the manuscript. FM, WJK, MH, CP-G, and DG provided important intellectual content. All authors contributed to the revision of the manuscript and approved the final version of the manuscript.

Conflicts of Interest

BS is the Associate Editor of *JMIR Rehabilitation and Assistive Technologies*. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Details of the survey.
[DOCX File , 55 KB - jmir_v26i1e53991_app1.docx]

References

XSL•FO
RenderX

1. Graham I, Atar D, Borch-Johnsen K, Boysen G, Burell G, Cifkova R, European Society of Cardiology (ESC) Committee for Practice Guidelines (CPG). European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts). Eur Heart J 2007 Oct;28(19):2375-2414. [doi: 10.1093/eurheartj/ehm316] [Medline: 17726041]
2. Piepoli MF, Corrà U, Adamopoulos S, Benzer W, Bjarnason-Wehrens B, Cupples M, et al. Secondary prevention in the clinical management of patients with cardiovascular diseases. Core components, standards and outcome measures for referral and delivery: A Policy Statement from the Cardiac Rehabilitation Section of the European Association for Cardiovascular Prevention & Rehabilitation. Endorsed by the Committee for Practice Guidelines of the European Society of Cardiology. Eur J Prev Cardiol 2014 Jun;21(6):664-681. [doi: 10.1177/2047487312449597] [Medline: 22718797]
3. Ambrosetti M, Abreu A, Corrà U, Davos CH, Hansen D, Frederix I, et al. Secondary prevention through comprehensive cardiovascular rehabilitation: From knowledge to implementation. 2020 update. A position paper from the Secondary Prevention and Rehabilitation Section of the European Association of Preventive Cardiology. Eur J Prev Cardiol 2021 May 14;28(5):460-495. [doi: 10.1177/2047487320913379] [Medline: 33611446]
4. Simon M, Korn K, Cho L, Blackburn GG, Raymond C. Cardiac rehabilitation: A class 1 recommendation. Cleve Clin J Med 2018 Jul;85(7):551-558 [FREE Full text] [doi: 10.3949/ccjm.85a.17037] [Medline: 30004380]
5. Leon AS, Franklin BA, Costa F, Balady GJ, Berra KA, Stewart KJ, American Heart Association, Council on Clinical Cardiology (Subcommittee on Exercise, Cardiac Rehabilitation,Prevention), Council on Nutrition, Physical Activity,Metabolism (Subcommittee on Physical Activity), American association of CardiovascularPulmonary Rehabilitation. Cardiac rehabilitation and secondary prevention of coronary heart disease: an American Heart Association scientific statement from the Council on Clinical Cardiology (Subcommittee on Exercise, Cardiac Rehabilitation, and Prevention) and the Council on Nutrition, Physical Activity, and Metabolism (Subcommittee on Physical Activity), in collaboration with the American association of Cardiovascular and Pulmonary Rehabilitation. Circulation 2005 Jan 25;111(3):369-376. [doi: 10.1161/01.CIR.0000151788.08740.5C] [Medline: 15668354]
6. Heimer M, Schmitz S, Teschler M, Schäfer H, Douma ER, Habibovic M, et al. eHealth for maintenance cardiovascular rehabilitation: a systematic review and meta-analysis. Eur J Prev Cardiol 2023 Oct 26;30(15):1634-1651. [doi: 10.1093/eurjpc/zwad145] [Medline: 37154363]
7. Denton F, Waddell A, Kite C, Hesketh K, Atkinson L, Cocks M, et al. Remote maintenance cardiac rehabilitation (MAINTAIN): A protocol for a randomised feasibility study. Digit Health 2023;9:20552076231152176 [FREE Full text] [doi: 10.1177/20552076231152176] [Medline: 36818155]
8. Andersen RM, Skou ST, Clausen MB, Jäger M, Zangger G, Grøntved A, et al. Maintenance of physical activity after cardiac rehabilitation (FAIR): study protocol for a feasibility trial. BMJ Open 2022 Apr 05;12(4):e060157 [FREE Full text] [doi: 10.1136/bmjopen-2021-060157] [Medline: 35383088]
9. Bjarnason-Wehrens B, Grande G, Loewel H, Völler H, Mittag O. Gender-specific issues in cardiac rehabilitation: do women with ischaemic heart disease need specially tailored programmes? Eur J Cardiovasc Prev Rehabil 2007 Apr;14(2):163-171. [doi: 10.1097/HJR.0b013e3280128bce] [Medline: 17446793]
10. Russell MW, Huse DM, Drowns S, Hamel EC, Hartz SC. Direct medical costs of coronary artery disease in the United States. Am J Cardiol 1998 May 01;81(9):1110-1115. [doi: 10.1016/s0002-9149(98)00136-2] [Medline: 9605051]
11. Brørs G, Pettersen TR, Hansen TB, Fridlund B, Hølvold LB, Lund H, et al. Modes of e-Health delivery in secondary prevention programmes for patients with coronary artery disease: a systematic review. BMC Health Serv Res 2019 Jun 10;19(1):364 [FREE Full text] [doi: 10.1186/s12913-019-4106-1] [Medline: 31182100]
12. Walsh DMJ, Moran K, Cornelissen V, Buys R, Claes J, Zampognaro P, et al. The development and codesign of the PATHway intervention: a theory-driven eHealth platform for the self-management of cardiovascular disease. Transl Behav Med 2019 Jan 01;9(1):76-98. [doi: 10.1093/tbm/iby017] [Medline: 29554380]
13. Silva-Cardoso J, Juanatey JRG, Comin-Colet J, Sousa JM, Cavalheiro A, Moreira E. The future of telemedicine in the management of heart failure patients. Card Fail Rev 2021 Mar;7:e11 [FREE Full text] [doi: 10.15420/cfr.2020.32] [Medline: 34136277]
14. Verburg A, Selder JL, Schalij MJ, Schuuring MJ, Treskes RW. eHealth to improve patient outcome in rehabilitating myocardial infarction patients. Expert Rev Cardiovasc Ther 2019 Mar;17(3):185-192. [doi: 10.1080/14779072.2019.1580570] [Medline: 30732481]
15. Frederix I, Caiani EG, Dendale P, Anker S, Bax J, Böhm A, et al. ESC e-Cardiology Working Group Position Paper: Overcoming challenges in digital health implementation in cardiovascular medicine. Eur J Prev Cardiol 2019 Jul;26(11):1166-1177 [FREE Full text] [doi: 10.1177/2047487319832394] [Medline: 30917695]
16. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of ehealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
17. Herrera S, Salazar A, Nazar G. Barriers and supports in eHealth implementation among people with chronic cardiovascular ailments: integrative review. Int J Environ Res Public Health 2022 Jul 07;19(14):8296 [FREE Full text] [doi: 10.3390/ijerph19148296] [Medline: 35886149]

XSL•FO
RenderX

18. Kop W, Schmitz B, Gatsios D, Peña-Gil C, Gonzalez Juanatey J, Cantarero Prieto D, et al. Patient-centered lifestyle intervention using artificial intelligence methodologies: The TIMELY project for cardiac rehabilitation. Journal of Psychosomatic Research 2022 Jun;157:110872. [doi: 10.1016/j.jpsychores.2022.110872]
19. Zipfel N, Horreh B, Hulshof CTJ, de Boer AGEM, van der Burg-Vermeulen SJ. The relationship between the living lab approach and successful implementation of healthcare innovations: an integrative review. BMJ Open 2022 Jun 28;12(6):e058630 [FREE Full text] [doi: 10.1136/bmjopen-2021-058630] [Medline: 35768105]
20. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technol Assess 1998;2(3):i-iv, 1 [FREE Full text] [Medline: 9561895]
21. Schmitz B, De Maria R, Gatsios D, Chrysanthakopoulou T, Landolina M, Gasparini M, et al. Identification of genetic markers for treatment success in heart failure patients: insight from cardiac resynchronization therapy. Circ Cardiovasc Genet 2014 Dec;7(6):760-770. [doi: 10.1161/CIRCGENETICS.113.000384] [Medline: 25210049]
22. Teschler M, Heimer M, Schmitz B, Kemmler W, Mooren FC. Four weeks of electromyostimulation improves muscle function and strength in sarcopenic patients: a three-arm parallel randomized trial. J Cachexia Sarcopenia Muscle 2021 Aug;12(4):843-854 [FREE Full text] [doi: 10.1002/jcsm.12717] [Medline: 34105256]
23. Bond FW, Hayes SC, Baer RA, Carpenter KM, Guenole N, Orcutt HK, et al. Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. Behav Ther 2011 Dec;42(4):676-688. [doi: 10.1016/j.beth.2011.03.007] [Medline: 22035996]
24. Scheier MF, Carver CS, Bridges MW. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. J Pers Soc Psychol 1994 Dec;67(6):1063-1078. [doi: 10.1037//0022-3514.67.6.1063] [Medline: 7815302]
25. Denollet J. DS14: standard assessment of negative affectivity, social inhibition, and Type D personality. Psychosom Med 2005;67(1):89-97. [doi: 10.1097/01.psy.0000149256.81953.49] [Medline: 15673629]
26. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 1988 Jun;54(6):1063-1070. [doi: 10.1037//0022-3514.54.6.1063] [Medline: 3397865]
27. Reinecke H, Breithardt G, Engelbertz C, Schmieder RE, Fobker M, Pinnschmidt HO, et al. Baseline characteristics and prescription patterns of standard drugs in patients with angiographically determined coronary artery disease and renal failure (CAD-REF registry). PLoS One 2016;11(2):e0148057 [FREE Full text] [doi: 10.1371/journal.pone.0148057] [Medline: 26859890]
28. Kraal JJ, Van den Akker-Van Marle ME, Abu-Hanna A, Stut W, Peek N, Kemps HM. Clinical and cost-effectiveness of home-based cardiac rehabilitation compared to conventional, centre-based cardiac rehabilitation: Results of the FIT@Home study. Eur J Prev Cardiol 2017 Aug;24(12):1260-1273 [FREE Full text] [doi: 10.1177/2047487317710803] [Medline: 28534417]
29. Rawstorn JC, Gant N, Direito A, Beckmann C, Maddison R. Telehealth exercise-based cardiac rehabilitation: a systematic review and meta-analysis. Heart 2016 Aug 01;102(15):1183-1192. [doi: 10.1136/heartjnl-2015-308966] [Medline: 26936337]
30. Frederix I, Solmi F, Piepoli MF, Dendale P. Cardiac telerehabilitation: A novel cost-efficient care delivery strategy that can induce long-term health benefits. Eur J Prev Cardiol 2017 Nov;24(16):1708-1717. [doi: 10.1177/2047487317732274] [Medline: 28925749]
31. Kontos E, Blake KD, Chou WS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
32. Gray R, Indraratna P, Lovell N, Ooi S. Digital health technology in the prevention of heart failure and coronary artery disease. Cardiovasc Digit Health J 2022 Dec;3(6 Suppl):S9-S16 [FREE Full text] [doi: 10.1016/j.cvdhj.2022.09.002] [Medline: 36589760]
33. Indraratna P, Magdy J, Li J, et al. Patterns and predictors of smartphone ownership in a cardiology inpatient population. Eur Heart J 2021 Oct 14:3110-3111. [doi: 10.1093/eurheartj/ehab724.3111]
34. Kim S, Chow BC, Park S, Liu H. The usage of digital health technology among older adults in Hong Kong and the role of technology readiness and eHealth literacy: path analysis. J Med Internet Res 2023 Apr 12;25:e41915 [FREE Full text] [doi: 10.2196/41915] [Medline: 37043274]
35. Coorey GM, Neubeck L, Mulley J, Redfern J. Effectiveness, acceptability and usefulness of mobile applications for cardiovascular disease self-management: Systematic review with meta-synthesis of quantitative and qualitative data. Eur J Prev Cardiol 2018 Mar;25(5):505-521. [doi: 10.1177/2047487317750913] [Medline: 29313363]
36. Goliasch G, Kleber ME, Richter B, Plischke M, Hoke M, Haschemi A, et al. Routinely available biomarkers improve prediction of long-term mortality in stable coronary artery disease: the Vienna and Ludwigshafen Coronary Artery Disease (VILCAD) risk score. Eur Heart J 2012 Sep;33(18):2282-2289. [doi: 10.1093/eurheartj/ehs164] [Medline: 22745355]
37. Kleber ME, Goliasch G, Grammer TB, Pilz S, Tomaschitz A, Silbernagel G, et al. Evolving biomarkers improve prediction of long-term mortality in patients with stable coronary artery disease: the BIO-VILCAD score. J Intern Med 2014 Aug;276(2):184-194 [FREE Full text] [doi: 10.1111/joim.12189] [Medline: 25184160]

XSL•FO
RenderX

38. Ashur C, Cascino TM, Lewis C, Townsend W, Sen A, Pekmezi D, et al. Do wearable activity trackers increase physical activity among cardiac rehabilitation participants? A systematic review and meta-analysis. J Cardiopulm Rehabil Prev 2021 Jul 01;41(4):249-256. [doi: 10.1097/HCR.0000000000000592] [Medline: 33828045]
39. Su JJ, Yu DSF, Paguio JT. Effect of eHealth cardiac rehabilitation on health outcomes of coronary heart disease patients: A systematic review and meta-analysis. J Adv Nurs 2020 Mar;76(3):754-772. [doi: 10.1111/jan.14272] [Medline: 31769527]

Abbreviations

AAQ-2: Acceptance and Action Questionnaire version 2
CAD: coronary artery disease
CR: cardiac rehabilitation
DE: country code for Germany
DS-14: Type D scale for social inhibition
ECG: electrocardiogram
ES: country code for Spain
LOT-R: Revised Life Orientation Test
PANAS: Positive and Negative Affect Schedule

Edited by T de Azevedo Cardoso, S He; submitted 26.10.23; peer-reviewed by J Su, D Liu, P Dilaveris; comments to author 20.12.23; revised version received 28.12.23; accepted 30.01.24; published 22.02.24.

Please cite as:
Schmitz B, Wirtz S, Sestayo-Fernández M, Schäfer H, Douma ER, Alonso Vazquez M, González-Salvado V, Habibovic M, Gatsios D, Kop WJ, Peña-Gil C, Mooren F
Living Lab Data of Patient Needs and Expectations for eHealth-Based Cardiac Rehabilitation in Germany and Spain From the TIMELY Study: Cross-Sectional Analysis
J Med Internet Res 2024;26:e53991
URL: https://www.jmir.org/2024/1/e53991
doi:10.2196/53991
PMID:38386376

©Boris Schmitz, Svenja Wirtz, Manuela Sestayo-Fernández, Hendrik Schäfer, Emma R Douma, Marta Alonso Vazquez, Violeta González-Salvado, Mirela Habibovic, Dimitris Gatsios, Willem Johan Kop, Carlos Peña-Gil, Frank Mooren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Effectiveness of STAPP@Work, a Self-Management Mobile App, in Reducing Work Stress and Preventing Burnout: Single-Case Experimental Design Study

Sevda Demirel[1], MSc; Yvette Roke[1], MD, PhD; Adriaan W Hoogendoorn[2,3], PhD; Jamie Hoefakker[4], MSc; Kirsten Hoeberichts[1], MSc, MD; Peter N van Harten[5,6], MD, PhD

[1]Expertise Center for Autism Spectrum Disorder, GGz Centraal, Almere, Netherlands
[2]Department of Psychiatry, Amsterdam Universitair Medische Centra, Amsterdam, Netherlands
[3]Mental Health Program, Amsterdam Public Health Research Institute, Amsterdam, Netherlands
[4]Praktijk InTeam, Den Haag, Netherlands
[5]Department of Psychiatry, GGz Centraal, Amersfoort, Netherlands
[6]Department of Psychiatry and Neuropsychology, School of Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands

Corresponding Author:
Sevda Demirel, MSc
Expertise Center for Autism Spectrum Disorder
GGz Centraal
Boomgaardweg 12
Almere, 1326 AD
Netherlands
Phone: 31 41807329
Email: S.demirel@ggzcentraal.nl

Abstract

Background: Work-related stress and burnout remain common problems among employees, leading to impaired health and higher absenteeism. The use of mobile health apps to promote well-being has grown substantially; however, the impact of such apps on reducing stress and preventing burnout is limited.

Objective: This study aims to assess the effectiveness of STAPP@Work, a mobile-based stress management intervention, on perceived stress, coping self-efficacy, and the level of burnout among mental health employees.

Methods: The study used a single-case experimental design to examine the use of STAPP@Work among mental health employees without a known diagnosis of burnout (N=63). Participants used the app for 1 week per month repeatedly for a period of 6 months. Using a reversal design, the participants used the app 6 times to assess replicated immediate (1 week after use) and lasting (3 weeks after use) effects. The Perceived Stress Scale, the Coping Self-Efficacy Scale, and the Burnout Assessment Tool were used to measure the outcomes. Linear mixed models were used to analyze the data.

Results: After 6 months of app use for 1 week per month, the participants showed a statistically significant decrease in perceived stress (b=–0.38, 95% CI –0.67 to –0.09; *P*=.01; Cohen *d*=0.50) and burnout symptoms (b=–0.31, 95% CI –0.51 to –0.12; *P*=.002; Cohen *d*=0.63) as well as a statistically significant improvement in problem-focused coping self-efficacy (b=0.42, 95% CI 0-0.85; *P*=.049; Cohen *d*=0.42). Long-term use of the app provided consistent reductions in burnout symptoms over time, including in the level of exhaustion and emotional impairment.

Conclusions: The use of an app-based stress management intervention has been shown to reduce burnout symptoms and enhance coping self-efficacy among mental health workers. Prevention of burnout and minimization of work-related stress are of utmost importance to protect employee health and reduce absenteeism.

(J Med Internet Res 2024;26:e48883) doi:10.2196/48883

KEYWORDS

mental health; stress; coping; burnout; stress management; digital intervention; health promotion; mobile apps; mobile health; mHealth; mental health professionals

XSL•FO
RenderX

Introduction

Background

Work-related stress remains a persistent challenge for both employees and organizations, adversely affecting health outcomes and reducing job performance [1]. Stress is often defined as a condition in which an environmental demand exceeds the natural regulatory capacity of an organism, particularly in situations that are unpredictable and uncontrollable [2]. Numerous studies over the last several decades have demonstrated that psychosocial stress adversely affects hormonal processes and organ systems in the body [3-6]. Excessive stress has a detrimental effect on mental health by causing emotional distress and increasing the risk of mental illnesses such as depression, anxiety, and burnout [7].

Burnout is defined as a psychological syndrome primarily caused by a prolonged period of stress and the inability or unwillingness to cope effectively, resulting in both physical and mental exhaustion [8]. The term burnout is often used in the occupational context, where stress originates from the work environment. Such job stressors include heavy workload, long working hours, job insecurity, conflicts, low support, role ambiguity, and a lack of autonomy [9]. When work stress is of a long duration without applying effective coping strategies, job demands exceed capacity, leading to the depletion of one's emotional, physical, and cognitive resources [9]. Individual consequences further include job dissatisfaction, depersonalization, cynicism, and low fulfillment [10]. Organizational consequences include absenteeism, reduced productivity levels, a higher turnover rate, and an increase in health care costs [9,11]. Organizations may incur financial costs not only from losing and replacing experienced staff but also by providing employee burnout support, implementing employee assistance programs, and facilitating reintegration processes [11-14]. Burnout causes significant distress owing to symptoms such as chronic fatigue, trouble sleeping, depressive symptoms, and difficulties with concentration and memory [9]. In 2021, approximately 17% of Dutch employees reported being mentally fatigued by their work, 16% were emotionally exhausted, and 15% were completely exhausted after work [15]. Moreover, the Dutch health care sector had one of the highest proportions of employees experiencing work-related mental exhaustion, at 20% [16].

Effective interventions for preventing work-related stress and burnout are of high importance. The use of digital health, including mobile apps, as behavioral interventions to promote positive mental health is rapidly growing. The benefits of mobile mental health interventions include, but are not limited to, high accessibility, low costs, less time consumption, no clinician involvement, timely support, and the promotion of autonomy [17]. A recent systematic review found that mindfulness was the most commonly applied theory-based stress management strategy in apps [18]. They further concluded that such apps implemented stress prevention strategies less often than stress management strategies for stress recovery [18]. However, stressor identification and reduction, as primary prevention, is a crucial element in effective stress management [19]. Moreover, another review found that only a few stress management apps combined both stress monitoring and intervention based on detecting users' stress levels [20].

Previous studies examining the effectiveness of web- and app-based interventions among employees provided evidence of their effectiveness in stress reduction [21]. Improvements regarding emotional exhaustion and perceived stress in health care professionals using digital interventions were reported earlier as well [22,23]. However, the few available studies on apps for health care workers yielded mixed findings on stress and were often based on mindfulness practices or psychoeducation. For instance, 2 randomized controlled trials (RCTs) assessing an app among health care workers found no significant improvements in stress [24,25]. Although other mindfulness-based apps did find improvements in stress among physicians, nurses, and therapists [26]; hospital nurses [27]; and employees within the general practitioner practice [28], little is known regarding the effectiveness of app-based interventions aimed at stress management and burnout symptoms among employees, specifically health care professionals. Moreover, there is a lack of evidence for the sustained effects of long-term use of such app-based interventions [29].

Mobile-based ecological momentary interventions (EMIs) comprise an emerging area of research. These interventions are based on ecological momentary assessment (EMA), a widely used tool in psychiatry to collect repeated daily assessments of individuals' behaviors, emotions, experiences, or thoughts as they occur in their own "real-world" settings [30]. EMA is seen as a promising method that provides insight into variations in symptoms and behavior over time and their underlying person-environment interactions. EMIs offer interventions in the moment of an individual's daily life, usually based on the assessment [31]. Previous studies on the effects of EMIs showed a positive overall medium effect on stress in clinical populations [32]. To date, no study has assessed the effectiveness of a mobile stress management intervention based on stress monitoring among health care workers.

Objectives

To bridge this gap, our group developed a self-management stress-signaling app using a participatory design with mental health employees called STAPP@Work. The app is developed to monitor daily stress levels and track the daily activities in employees' natural settings, including work settings, based on the principles of EMA. STAPP@Work is a tool to manage stress at work by providing the employee feedback regarding their tracked stress levels and associated activities and creating a visual overview to make stress patterns during workdays and weeks visible. Furthermore, the app provides real-time suggestions for coping strategies specified for both the work environment and at home. The aim of this study was to evaluate the effectiveness of the STAPP@Work app by examining perceived stress, coping self-efficacy, and burnout symptoms. On the basis of previous literature outlined in the *Background* section, we hypothesized that after 6 months of using the STAPP@Work app for 1 week per month, perceived stress and coping self-efficacy would be improved and burnout symptoms would be reduced in mental health workers. In addition, we

XSL•FO
RenderX

explored the progression of perceived stress, coping self-efficacy, and burnout scores over time to examine their lasting effects.

Methods

Study Design

This study adopted a single-case experimental design (SCED), a within-person design that measures the effectiveness of interventions [33]. Using an SCED, it became possible to focus on individual changes in outcomes (perceived stress, coping self-efficacy, and burnout symptoms) over time. This design was considered a good fit, as it enabled the assessment of behavior change over time in "real-world" settings using heterogeneous samples [33]. Moreover, SCEDs are often used to capture small but meaningful differences in behavior and to examine behavior that may change gradually over time, which is not feasible to assess in RCTs with a high number of successive measurements [34]. In addition, SCEDs are suitable for being able to evaluate the effectiveness of novel innovative interventions before they are evaluated in time-consuming and costly RCTs [35]. The results of this SCED study will provide an indication of whether it is worthwhile to conduct a large-scale RCT to further assess the effectiveness of the app [36]. Given that STAPP@Work is a novel behavioral intervention, using an SCED constitutes a useful first step to measure its effectiveness.

Furthermore, an A1-B1-A2-B2 reversal design was used, in which A1 represents the baseline period, B1 the introduction of an intervention, A2 the withdrawal of the intervention, and B2 the reintroduction of the intervention [36]. In this study, this translated to the participants having used the STAPP@Work app for 1 week after the baseline period, followed by 3 weeks of nonuse, after which they used the app again for the second time for a weeklong period. Withdrawing the intervention allowed for an assessment of whether the effects of the app were lasting.

Sample Size

Owing to the complexity and practical unfeasibility of the sample size calculation for this SCED, the sample size for this study was based on the traditional paired sample 2-tailed *t* test design. Similar studies on stress management apps among health care employees are limited. Instead, we relied on a meta-analysis of the effectiveness of web- and app-based interventions on stress among employees, which found moderate-to-large effect sizes for mindfulness-based interventions and small-to-moderate effect sizes for stress management interventions [37]. A medium effect size (Cohen d=0.5) was considered in the sample size calculation for this study. To conduct a paired sample 2-tailed *t* test and detect a medium effect size (Cohen d=0.5) with a low internal correlation (r=0.1), a sample size of 59 participants was required. Considering the potential dropout rate of 7.5%, the required sample size was set at 63 participants.

Participants

Participants were eligible to participate in the study when they were employed within GGz Centraal, the fourth largest mental health care institution in the Netherlands. No further inclusion or exclusion criteria were applied. Recruitment strategies included the promotion of the study through flyers, digital media, and approaching teams or staff in various departments.

The STAPP@Work App

App Overview

STAPP@Work is a self-management app designed to manage stress and improve well-being by (1) measuring daily stress levels, (2) providing personal coping advice, and (3) visualizing stress patterns. The app is based on a stress-signaling plan, a widely used tool in Dutch mental health institutions to monitor clients' stress levels in relation to their mental health [38]. This allows for recognizing and dealing with early warning signs of distress. A stress-signaling plan often consists of 4 phases: green, yellow, orange, and red. Each phase characterizes how a person is feeling and what symptoms are associated, with the colors increasing from green (no stress) to red (very high stress) [38]. Furthermore, the plan describes both stress-inducing and protective factors as well as personal coping strategies to encourage the clients' self-management [38]. STAPP@Work includes the concept of a stress-signaling plan and the EMA method by assessing stress levels in daily life and ensuring that these are linked to specific moments and activities in a day. The app was designed together with mental health employees and was based on their needs, ideas, and experiences that were explored during focus group discussions. The ongoing iterative development of the prototypes continued until the desired version of the STAPP@Work app was achieved.

Measuring Daily Stress

The app calculated daily stress levels by preparing a questionnaire. The employee could specify whether to complete 2, 3, or 4 questionnaires per day via the app settings. The app required a minimum of 2 questionnaires, as there were 4 hours between each questionnaire, and filling in 2 questionnaires would cover a working day of 8 hours. If the employee chose to fill in 3 or 4 questionnaires a day, the app would cover a 12-hour or 16-hour period out of 1 day. On the home page, the user could see if a questionnaire was available to fill in as well as the results of the previous stress measurement, as shown in Figure 1.

XSL•FO
RenderX

Figure 1. Screenshots of the home page and the visual overview page of the STAPP@Work app (translated from Dutch).

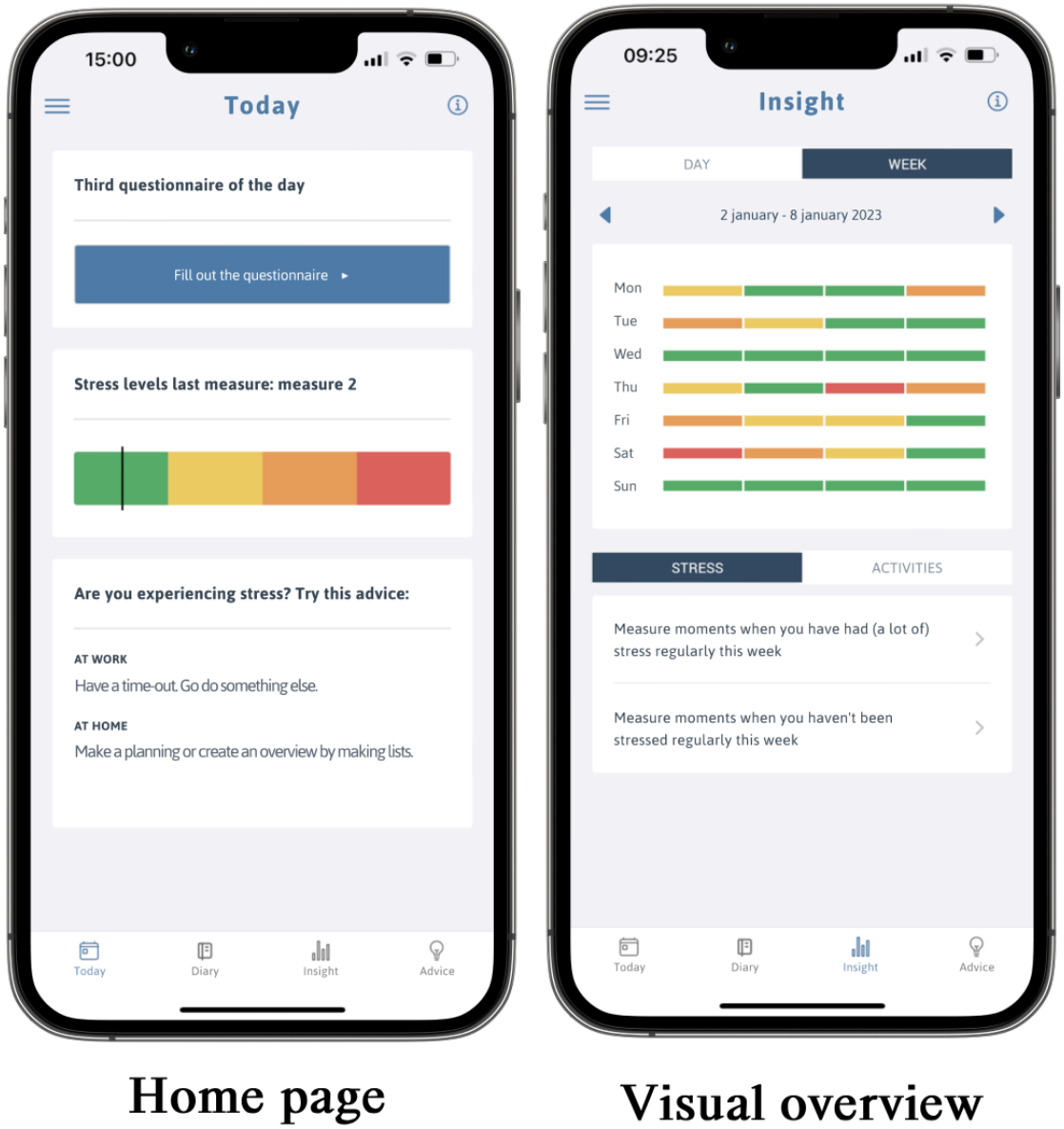

Each questionnaire became available for the user to complete within a 1-hour time frame, probing activities that were performed in the past 4 hours and how one felt while doing so. Then, it followed with 7 stress-signaling questions (eg, “Did you feel irritable?” and “Were you able to take good care of yourself?”) to derive a stress level. The app calculates stress levels based on the answers to 7 stress questions. The stress questions were formulated and validated by focus group discussions and individual interviews with employees, following the design thinking method. The questions were not derived from an existing stress scale. The answer options for each stress question are as follows: “no”; “yes, but not more than normal”; “yes, more than normal”; and “yes, much more than normal.” A score is assigned to each answer option, and from there, the app computes a total stress score based on the answers to all 7 stress questions. The app then assigns the user’s perceived stress to 1 of 4 categories: no stress, little stress, stress, or a lot of stress, based on the total score from the stress-related questions. The app also asked for the user’s feedback on the assigned stress categories and gave personal coping advice.

Visual Overview

The app includes a visual overview of the assigned stress levels at both the day and week levels, helping the users to recognize their own stress patterns (Figure 1). The overview also showed which activities were associated with (a lot of) stress and no stress during the week.

Personal Coping Advice

During the exploratory focus groups and interviews, employees were also asked how they cope with stress at work and at home. Together with desk research and practitioner experience, a list of different effective approaches to coping with stress was formed, which was referred to as coping strategies. This list forms the general list of predefined suggestions for coping in the app. Suggestions for coping at work included the following: contact or discuss the situation with a colleague, get moving, go walking, plan or make an overview, do breathing exercises, take a break, do something else, and take care of yourself such as eating and drinking enough. However, the participants could

XSL•FO
RenderX

add their own suggestions for coping to the list and remove existing suggestions. After each stress assessment, the app suggests a coping strategy displayed in text, which the app randomly selects from the personalized coping list of the user.

Personalization

The app was highly customizable, as the users could choose how many questionnaires they wanted to fill out in a day and at what times, to fit their working days. The app consisted of a list of activities created by mental health employees that fit within a workday, but users could add or remove activities as they wished. The same was true for coping advice, in which the app provided a list of coping strategies that the user could change at any time.

Measures

Perceived Stress

The Perceived Stress Scale (PSS) is a 10-item self-report questionnaire that was used to measure how stressful certain situations were perceived in the past month (eg, "How often have you felt that you were in control of things?"), with items rated on a 5-point scale ranging from 0 (never) to 4 (very often) [39]. The PSS has shown good internal consistency (10 items; Cronbach α=0.84) [40].

Coping Self-Efficacy

The 13-item version of the Coping Self-Efficacy Scale (CSES) measures one's confidence in various coping behaviors [41]. The CSES consists of 3 subscales: problem-focused coping (6 items; Cronbach α=0.91), emotion-focused coping (4 items; Cronbach α=0.91), and social support coping (3 items; Cronbach α=0.80) [41]. Items have a 10-point scale ranging from 0 (cannot do at all) to 10 (certainly can do).

Burnout Symptoms

The shorter version of the Burnout Assessment Tool (BAT), which consists of a 12-item self-report questionnaire, determines the extent to which burnout symptoms are experienced (eg, "I do not recognize myself in the way I react emotionally at work"), rated from 0 (never) to 4 (always) [42]. Cronbach α has been shown to be similar to the full version of the BAT (23 items) [43]. The BAT (23 items; Cronbach α=0.90) is subdivided into 4 validated core dimensions: exhaustion (8 items; Cronbach α=0.92), mental distance (5 items; Cronbach α=0.91), cognitive impairment (5 items; Cronbach α=0.92), and emotional impairment (5 items; Cronbach α=0.90) [42].

Procedure

A measurement involved filling out the questionnaire that was composed of the PSS, CSES, and BAT items. The researchers (KH and JH) sent the web-based questionnaire to all participants at each time of measurement, regardless of whether they had completed the previous questionnaire or not. A total of 17 measurements were obtained from the participants over a period of 9 months. Of these, 24% (4/17) were conducted during the baseline period and 71% (12/17) during the intervention phase, ending with 6% (1/17) follow-up measurement. We chose to include 4 baseline measurements to account for possible learning effects from merely participating in the study. Being introduced to the study and responding to the questionnaires on stress could have possibly contributed to an increased awareness of one's own stress. During the 7-week baseline phase, the participants completed the questionnaire every 2 weeks. After the baseline measurement, the research team provided individual explanatory sessions on installing, setting up, and using the STAPP@Work app as well as the proceedings of the experiment. The participants could share their comments and questions about the study, as well as inquire about app functionalities, with the help desk that was available.

The intervention concerned the daily use of the STAPP@Work app for 1 week, called the intervention week. After the withdrawal of the intervention, a questionnaire was completed in the week after use (1 week after use) and 3 weeks after use to examine the immediate and lasting effects, respectively. The introduction and withdrawal of the intervention were repeated 6 times throughout the intervention phase. A follow-up measurement took place 6 weeks after the last time the app was used. Figure 2 shows the order of the events.

Figure 2. Schematic overview of the study design, including the timeline.

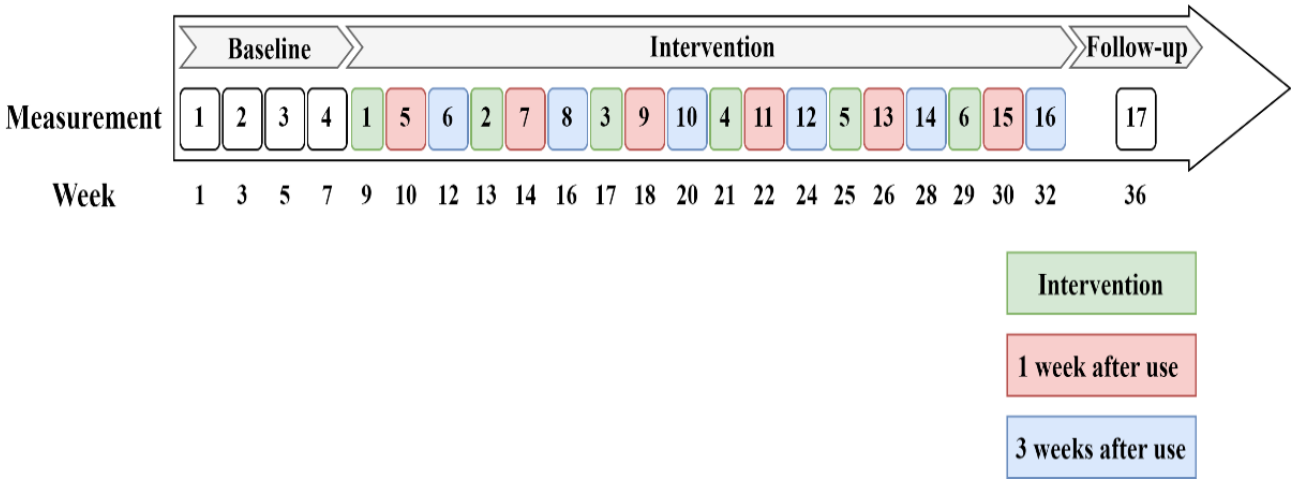

XSL•FO
RenderX

Statistical Analysis

All analyses were performed using SPSS (version 28; IBM Corp). For data analysis, mean scores were calculated from the total scores of the PSS, CSES, and BAT, including the subscales. The subscales of the CSES and BAT were included as separate outcome variables to gain more insight into the subdomains. Descriptive statistics were used to describe participant characteristics and determine response rates over time. We categorized participants as early dropouts if they failed to use the app at least 3 times and definitively ceased participation. Data analysis did not include early dropouts. Linear mixed model (LMM) analysis was used to analyze the data [44]. LMMs are well suited for longitudinal intervention studies, as they allow the measurement of change in the response variables across units of time and within participants [45,46] and account for missing values [47]. Change scores between baseline and all consecutive intervention measurements were modeled for all outcome variables. We chose the fourth and last baseline measurement (week 7) as a reference to correct for learning effects during the baseline period [36].

When conducting the LMM, assumptions of normality and equal variance of residuals were checked using graphical representations. As the data were collected over time, the presence of autocorrelation was evaluated both visually by the partial autocorrelation function plot and by the Durbin-Watson statistic test. Accordingly, the appropriate covariance structure was selected for the models for each outcome variable. Time was represented by (binary) week indicators that were included in the LMM as fixed effects, enabling the assessment of mean scores for each week of app use. For each outcome, Cohen *d* effect sizes were calculated for the mean change score from baseline to 6 months, standardized by the baseline SD. The LMM results were used to create graphs, visually presenting the estimated mean scores of the outcomes and illustrating the progression over time. The significance level was set to .05.

Ethical Considerations

To ensure data confidentiality, consideration has been given to secure data collection and storage. Data were collected through web-based questionnaires and stored in a folder. Only researchers (JH, KH, YR, and SD) within the team with access could view the responses. Data were then anonymized and stored in a protected folder on a secure server of GGz Centraal, where access for a single researcher (SD) was protected by a password and a soft token-based authentication. This study was conducted in accordance with the Declaration of Helsinki and was approved by the Medical Ethics Committee of Isala Zwolle (211006211006) on November 11, 2021. Informed consent was obtained from all participants involved in the study. The informed consent form included extensive information regarding the purpose of the study, procedures, data security, and the participant's rights. Participation in the study was completely voluntary and without any compensation. Furthermore, the use of the STAPP@Work app is in accordance with the General Data Protection Regulation of the European Union.

Results

Sample Characteristics

A total of 63 mental health employees were included in the study, with 10 early dropouts, resulting in 53 (84%) participants. Of these, 74% (39/53) were female participants, and the average age was 41 (SD 11.7) years (Table 1).

Various mental health professions were included, with the majority represented by nurses (22/53, 42%). The response rates of the participants ranged from 45% (24/53) to 100% (53/53), as presented in Table 2.

Table 1. Characteristics of the participants (n=53).

Characteristic	Value
Age (years), mean (SD)	41 (11.7)
Sex n (%)	
Female	39 (74)
Male	14 (26)
Profession, n (%)[a]	
Nurse	22 (42)
Psychologist	7 (13)
Management function	7 (13)
Counselor	7 (13)
Staff member or adviser	4 (8)
Administration or IT function	3 (6)
Psychiatrist	2 (4)
Medical doctor in training	1 (2)

[a]Due to rounding, percentages may not add up to 100%.

XSL•FO
RenderX

Table 2. Response rates among participants (n=53).

	Questionnaire	Value, response rate (%)[a]
Baseline	Questionnaire 1	53 (100)
Baseline	Questionnaire 2	52 (98)
Baseline	Questionnaire 3	52 (98)
Baseline	Questionnaire 4	51 (96)
Intervention	Questionnaire 5	46 (87)
Intervention	Questionnaire 6	42 (79)
Intervention	Questionnaire 7	41 (77)
Intervention	Questionnaire 8	41 (77)
Intervention	Questionnaire 9	37 (70)
Intervention	Questionnaire 10	38 (72)
Intervention	Questionnaire 11	35 (66)
Intervention	Questionnaire 12	29 (55)
Intervention	Questionnaire 13	24 (45)
Intervention	Questionnaire 14	27 (51)
Intervention	Questionnaire 15	28 (53)
Intervention	Questionnaire 16	26 (49)
Follow-up	Questionnaire 17	30 (57)

[a]Response rate refers to the number of participants who completed the questionnaire.

Perceived Stress

The perceived stress score significantly improved after 6 months (b=–0.38, 95% CI –0.67 to –.09; *P*=.01) compared with baseline, with a medium effect size (Cohen *d*=0.50), as presented in Table 3.

Figure 3 shows a lower stress score after each intervention week for 1 week after use, except for the fifth intervention week. A significant decrease in stress score was not found at follow-up.

Table 3. Results of linear mixed model for all outcome variables: perceived stress, coping self-efficacy with subscales, and burnout symptoms with subscales after 6 months of monthly 1-week app use.

Variable	Unstandardized beta coefficient, b (SE: 95% CI)	*P* value	Cohen *d*
Perceived stress	–0.38 (–0.15; 0.67 to –0.09)	.01	0.50
Total coping self-efficacy[a]	0.40 (0.22; –0.03 to 0.83)	.07	0.37
Problem-focused coping	0.42 (0.21; 0 to 0.85)	.049	0.42
Emotion-focused coping	0.52 (0.30; –0.08 to 1.11)	.09	0.17
Social support coping	0.24 (0.33; –0.41 to 0.89)	.47	0.29
Total burnout symptoms[b]	–0.31 (0.10; –0.51 to –0.12)	.002	0.63
Exhaustion	–0.38 (0.15; –0.68 to –0.08)	.02	0.50
Mental distance	–0.22 (0.14; –0.48 to 0.05)	.11	0.34
Emotional impairment	–0.42 (0.13; –0.68 to –0.16)	.002	0.69
Cognitive impairment	–0.24 (0.13; –0.49 to 0.01)	.06	0.41

[a]Total score of all 13 items on the Coping Self-Efficacy Scale.

[b]Total scores of all 12 items on the Burnout Assessment Tool.

XSL•FO
RenderX

Figure 3. Graphs showing the progression of the estimated means from linear mixed models of perceived stress, total coping self-efficacy, problem-focused coping self-efficacy, emotion-focused coping self-efficacy, and social support coping self-efficacy over a 36-week period. The baseline, intervention, and follow-up phases are presented. Each intervention week is indicated by the corresponding number. Error bars represent CIs. *P<.05. **P<.01 significant change from baseline (wk 7).

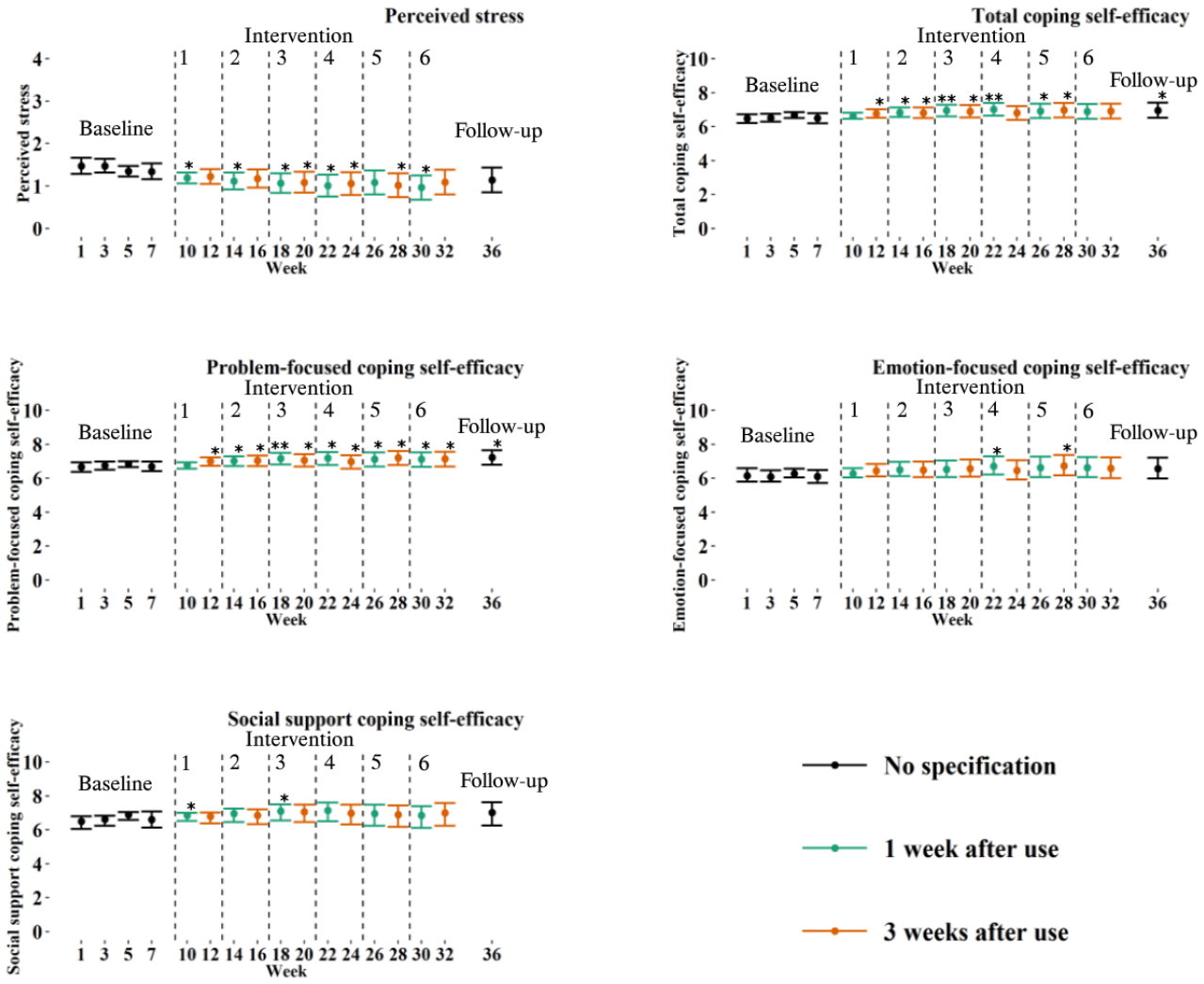

Coping Self-Efficacy

Coping self-efficacy did not significantly improve after 6 months of app use for 1 week per month (b=0.40, 95% CI –0.03 to 0.83; P=.07), as shown in Table 3. Focusing on the subscales, problem-focused coping abilities significantly improved after 6 months (b=0.42, 95% CI 0-0.85; P=.049) with a small to medium effect size (Cohen d=0.42), and emotion-focused coping (b=0.52, 95% CI –0.08 to 1.11; P=.09) and social support coping (b=0.24, 95% CI –0.41 to 0.89; P=.47) were not significantly improved. Figure 3 shows significant, consistent improvements in problem-focused coping starting from the second intervention week for both 1 week and 3 weeks after use until follow-up. No significant, consistent improvements were found regarding emotion-focused and social support coping self-efficacy scores (Figure 3).

Burnout Symptoms

Table 3 shows a significant improvement in the measure of burnout symptoms after 6 months of app use for 1 week per month (b=–0.31, 95% CI –0.51 to –0.12; P=.002) compared with baseline, with a medium effect size (Cohen d=0.63). Regarding the subscales, a significant decrease was found for exhaustion (b=–0.38, 95% CI –0.68 to –0.08; P=.02) and emotional impairment (b=–0.42, 95% CI –0.68 to –0.16; P=.002). The effect sizes were found to be medium for both exhaustion (Cohen d=0.50) and emotional impairment (Cohen d=0.69).

When examining the progression over time, a significant decrease in emotional impairment was found after the first week of app use and was sustained for both 1 week post use and 3 weeks after use until follow-up (Figure 4). Consistent improvements following initial app use were found for exhaustion as well, with no significant change at follow-up (Figure 4). No consistent and sustained improvements were found for the level of mental distance and cognitive impairment (Figure 4).

XSL•FO
RenderX

Figure 4. Graphs showing the progression of the estimated means from linear mixed models of total burnout symptoms, exhaustion, mental distance, cognitive impairment, and emotional impairment over a 36-week period. The baseline, intervention, and follow-up phases are presented. Each intervention week is indicated by the corresponding number. Error bars represent CIs. *P<.05, **P<.01 significant change from baseline (wk 7).

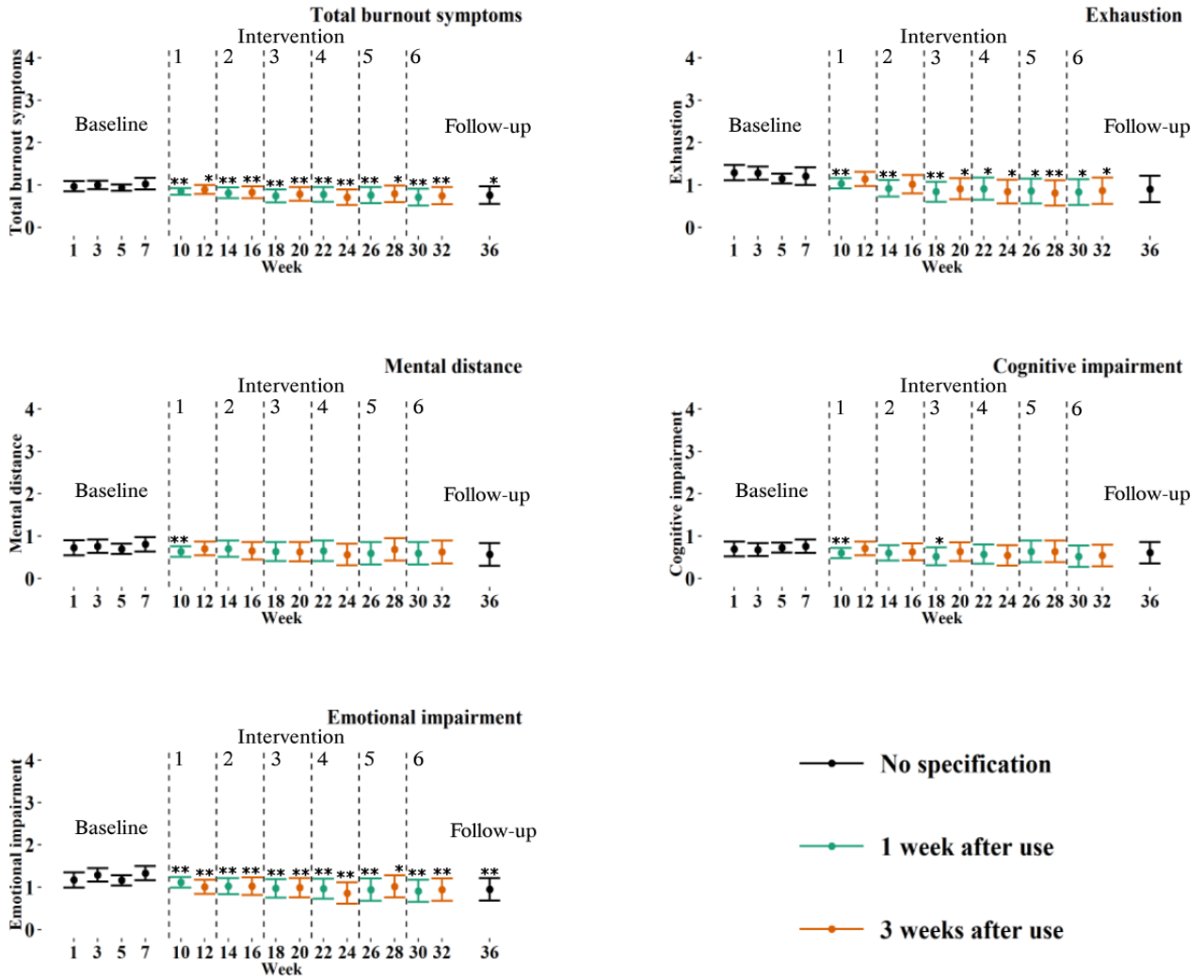

Discussion

Principal Findings

This effectiveness study showed that the participants using STAPP@Work were able to consistently reduce their measure of burnout symptoms over time. Specifically, the findings consistently showed improvements in exhaustion and emotional impairment, with sustained improvement observed in emotional impairment. Simultaneously, the participants showed lasting improvements in their confidence in problem-focused coping abilities. Perceived stress levels were reduced after 6 months of monthly 1-week app use. Following the findings, the app provides a means for improving problem-focused coping strategies to cope effectively with job stressors and reduce burnout complaints.

This study found lasting improvements in burnout symptoms. Although each participant scored a measure of burnout symptoms, it is important to emphasize that they did not have a known burnout diagnosis. All participants were working during the time of this study. From the findings, it became apparent that particularly 2 core dimensions—exhaustion and emotional impairment—were consistently reduced. Improvements in exhaustion indicate that the employees felt less tired and experienced higher physical and mental energy levels at work [42]. A decrease in emotional impairment implies that the participants were able to improve their abilities to adequately control their own emotions at work [42]. The app providing insights into one's stress patterns, and work-related stressors may have contributed to higher awareness and control over one's stress levels at work. The results of this study showed reductions in perceived stress among the employees. The employees may have been able to reduce their stress as the app assessed their perceived stress several times a day, allowing them to intervene early when stress levels were rising to prevent worsening, similar to how a stress-signaling plan functions [38].

Moreover, they improved their problem-focused coping self-efficacy, indicating increased confidence in coping effectively by resolving the stressful situation or altering the source of the stress [41]. This suggests that the participants were more able to cope effectively with work-related stressors and demands, resulting in lower exhaustion and higher control of emotional processes. Stress assessment and surveillance are crucial for determining the causal and contributing factors of

XSL•FO
RenderX

stress according to the theory of preventive stress management (PSM) [19]. The PSM framework further encourages individual stress assessment by assuming a transactional relationship between an individual and work-related stress because of individual differences [19]. STAPP@Work calculated daily stress levels and created a personal overview in which one could see the moments and activities that were connected to high stress. This allowed the employee to recognize moments of high stress during the day and week as well as the associated activities to identify stress-inducing factors. This may have empowered employees to lessen, modify, or manage recognized work stressors, aiming to reduce or prevent the stress response as part of primary prevention [19]. Addressing work stressors reduces high job demands, leading to less fatigue and more energy, as the job demands–resources model assumes that high job demands and low resources are important predictors of the exhaustion component of burnout [48].

The PSM theory argues that chronic stress that remains unaddressed and builds over time is manageable by providing interventions focused on both primary and secondary prevention [19]. The app addresses secondary prevention by providing real-time suggestions for coping strategies to manage work stress. Stressful situations become more manageable and less threatening because of such effective coping strategies, which act as a protective factor for the development of burnout [9]. Improvements in problem-focused coping self-efficacy were both immediate and lasting, suggesting that app use has led to the learning of problem-focused coping skills. In contrast, emotion-focused and social support coping abilities did not significantly improve; therefore, our hypothesis only supports problem-focused coping self-efficacy. The features of the app may have been especially profitable for problem-focused coping. Insights into stress patterns and daily stress levels through app use may have caused participants to be more proactive in identifying and resolving the cause of stress. Previous research has shown that individuals engaged more in problem-focused coping instead of emotion-focused coping when they were clearer and more communicative about their own emotions [49]. Therefore, participants gaining insights and self-reflection regarding their stress level may have little to gain by engaging in emotion-focused coping.

Indeed, the perceived stress level significantly reduced after 6 months; however, when examining the effects over time, these were often to be found immediate instead of lasting with no significant change at follow-up. This implies that the app is most effective when users actively engage with it, leading to increased awareness and control. Conversely, when the app is not in use and does not provide feedback on stress levels, the effects diminish until the app is used again.

The few studies examining mental health apps in relation to burnout have shown similar results in the improvement of burnout symptoms among employees [50] and specifically in reducing emotional exhaustion among health care professionals [23,51]. Furthermore, findings from an RCT assessing a self-management app among employees showed medium effects for reduction in perceived stress, lower emotional exhaustion, and higher emotion regulation, which are consistent with our findings [52]. Most of the scientifically studied mental health apps were based on the theories of mindfulness, psychoeducation, or cognitive behavioral therapy and were studied in RCTs [22,53]. The findings of this study add to the existing body of research by providing evidence that a stress management app, aligned to the work environment and assessed in a naturalistic context, can reduce symptoms that indicate burnout and, therefore, has the potential to preventively avert the risk of burnout. Furthermore, this study provides insights into the sustained effects of mental health apps, which were often lacking in previous studies [29].

Strengths and Limitations

One of the strengths of this study includes the use of an SCED, which made it possible to offer the intervention to all participants [20] and enabled the measurement of the changes of each participant in “real-life” conditions rather than comparing them between the 2 groups in a highly controlled setting [54]. Furthermore, the relatively long length of the experiment allowed for close observation of how stress, coping, and burnout symptoms progressed separately over a 9-month period and whether these were learned over time, such as coping [33]. In addition, STAPP@Work involves an innovative design in which an app incorporates a stress-signaling intervention and is of potential value in terms of reducing absenteeism.

There are several limitations to consider when interpreting this study. First, the participants may have induced a social desirability bias. However, it is unlikely that such a bias persisted for several months after the intervention. Second, we observed that the number of missing values increased as the study progressed. Participants may have failed to adhere because of a loss of interest or other external circumstances. However, LMMs can deal with dropout if the missing data mechanism is missing at random, which is the case if dropout depends on observed characteristics such as stress in the baseline period [47].

Third, each participant received the intervention at the same time following the baseline period, as it was not feasible to measure baseline stability at the individual level. Such stability is valuable to ensure that results are accountable to the intervention rather than to contextual influences [55]. This study did apply a reversal design, contributing to a higher degree of experimental control by repeatedly introducing and withdrawing the intervention, leading to a more robust demonstration of replicated effects and resulting in higher internal validity [56]. Fourth, the small sample size and disproportionate representation of female participants limited the generalizability of the findings to a broader and more diverse population. However, the distribution of sex in this study population corresponds to the distributions reported for Dutch mental health institutions [57]. Moreover, potential knowledge on stress management by mental health professionals restricts the generalizability to other industries or work settings as well. Finally, as this study followed a repeated measures design, it became more difficult to assess if changes occurring over time were a result of repeated use of the app or rather the cumulative effects of previous app use [36]. However, this did not limit the interpretation of whether such previous effects of the app were lasting.

XSL•FO
RenderX

Implications and Future Research

The preliminary results provided promising evidence that STAPP@Work is effective in reducing burnout symptoms and improving problem-focused coping. These findings build upon existing research by showing that a mobile health app can lead to sustainable effects and is able to address burnout-related symptoms. This is particularly relevant for preventive implementation in organizations to prevent burnout. This self-management app may have the potential to offer support in occupational contexts to promote less absenteeism and a healthier workforce, which are important for well-functioning and successful organizations. In addition, the app has the potential to integrate into daily life to support employees during their workdays or challenging periods. As the app helps to identify job stressors and provides coping advice suited for work environments, it can be implemented in any workplace to protect the health of employees and enhance job satisfaction.

Because of their high accessibility and autonomy, mobile interventions can be a more convenient and resource-saving way to prevent the onset of aggravated stress symptoms from within the workplace. This is especially relevant as recent challenges in the health care workforce include high-pressure working conditions and a shortage of workers, although the number of employees has increased in recent years [58]. It is of utmost importance to focus on the health of these employees, who are putting in the effort to provide the best care for clients.

An SCED study is typically characterized by a within-person design. Hence, this design does not include a control group but instead uses longer baseline periods and multiple measures over time. This impeded the ability to draw firm conclusions on the effects of the app, especially for a larger group, and therefore highlights the need for future experimentally controlled studies investigating larger and more heterogeneous samples. Further qualitative research is necessary to complement the findings of this study, as it remains unclear how the app contributes to perceived stress, coping skills, and burnout symptoms. By exploring this context and examining the underlying mechanisms and motivations for employees' behavior, a richer understanding of users' experiences and more insight into the meaning of the app in everyday work life are created [59,60].

Although the intervention resulted in a significant reduction in perceived stress immediately after using STAPP@Work, this study found no significant change in stress levels at follow-up. This suggests that the effectiveness of the app in reducing stress in the short term did not persist in the long term. This underscores the importance of conducting future research studies to investigate this specific aspect further. In addition, it would be beneficial to conduct further research on longer-term follow-up assessments to gain more insight into the lasting impact of the app. As the app shows promise in reducing burnout symptoms, it would be clinically relevant to explore the effect of the app on specific groups of individuals, for instance, those with a history of burnout or absenteeism. Because of the accessibility of the app, it can offer solutions to those individuals who are in immediate need of support in managing their stress, such as clients who are on the waiting list for psychological help, which prompts further research ideas.

Conclusions

This intervention study showed that STAPP@Work, as a stress management app, is effective in reducing burnout symptoms and improving problem-focused coping skills among mental health workers. The preliminary findings specifically showed reductions in exhaustion levels and emotional impairment. Moreover, the effects of the app emerged consistently and sustainably for >6 months. This offers great opportunities for implementation in the workplace, as the app has the potential to prevent burnout and reduce distress from work-related stress as well as decrease absenteeism.

Acknowledgments

The authors thank GGz Centraal for making this research possible with their support. The authors thank the software company, Eaglescience, for developing the app software. The authors gratefully thank all the study participants and the members of the SAM-STAPP (Stress Autism Mate-Stresssignaleringsapplicatie) team who were closely involved in the design and conduct of the study.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JH, KH, and YR contributed to conceptualization, investigation, and project administration. AWH and SD contributed to data curation, formal analysis, methodology, validation, and visualization. SD contributed to the writing of the original draft and the reviewing and editing of the manuscript. AWH contributed to the reviewing and editing of the manuscript. YR and PNvH contributed to supervision and the reviewing and editing of the manuscript.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Niedhammer I, Bertrais S, Witt K. Psychosocial work exposures and health outcomes: a meta-review of 72 literature reviews with meta-analysis. Scand J Work Environ Health 2021 Oct 01;47(7):489-508 [FREE Full text] [doi: 10.5271/sjweh.3968] [Medline: 34042163]
2. Koolhaas JM, Bartolomucci A, Buwalda B, de Boer SF, Flügge G, Korte SM, et al. Stress revisited: a critical evaluation of the stress concept. Neurosci Biobehav Rev 2011 Apr;35(5):1291-1301. [doi: 10.1016/j.neubiorev.2011.02.003] [Medline: 21316391]
3. Block JP, He Y, Zaslavsky AM, Ding L, Ayanian JZ. Psychosocial stress and change in weight among US adults. Am J Epidemiol 2009 Jul 15;170(2):181-192 [FREE Full text] [doi: 10.1093/aje/kwp104] [Medline: 19465744]
4. Jansson C, Wallander MA, Johansson S, Johnsen R, Hveem K. Stressful psychosocial factors and symptoms of gastroesophageal reflux disease: a population-based study in Norway. Scand J Gastroenterol 2010;45(1):21-29. [doi: 10.3109/00365520903401967] [Medline: 19961344]
5. Kemeny ME, Schedlowski M. Understanding the interaction between psychosocial stress and immune-related diseases: a stepwise progression. Brain Behav Immun 2007 Nov;21(8):1009-1018. [doi: 10.1016/j.bbi.2007.07.010] [Medline: 17889502]
6. Yonas MA, Lange NE, Celedón JC. Psychosocial stress and asthma morbidity. Curr Opin Allergy Clin Immunol 2012 Apr;12(2):202-210 [FREE Full text] [doi: 10.1097/ACI.0b013e32835090c9] [Medline: 22266773]
7. Tafet GE, Nemeroff CB. The links between stress and depression: psychoneuroendocrinological, genetic, and environmental interactions. J Neuropsychiatry Clin Neurosci 2016;28(2):77-88. [doi: 10.1176/appi.neuropsych.15030053] [Medline: 26548654]
8. Guseva Canu I, Marca SC, Dell'Oro F, Balázs Á, Bergamaschi E, Besse C, et al. Harmonized definition of occupational burnout: a systematic review, semantic analysis, and Delphi consensus in 29 countries. Scand J Work Environ Health 2021 Mar 01;47(2):95-107 [FREE Full text] [doi: 10.5271/sjweh.3935] [Medline: 33258478]
9. Edú-Valsania S, Laguía A, Moriano JA. Burnout: a review of theory and measurement. Int J Environ Res Public Health 2022 Mar 04;19(3):1780 [FREE Full text] [doi: 10.3390/ijerph19031780] [Medline: 35162802]
10. Salvagioni DA, Melanda FN, Mesas AE, González AD, Gabani FL, Andrade SM. Physical, psychological and occupational consequences of job burnout: a systematic review of prospective studies. PLoS One 2017 Oct 04;12(10):e0185781 [FREE Full text] [doi: 10.1371/journal.pone.0185781] [Medline: 28977041]
11. Lutz N, Taeymans J, Ballmer C, Verhaeghe N, Clarys P, Deliens T. Cost-effectiveness and cost-benefit of worksite health promotion programs in Europe: a systematic review. Eur J Public Health 2019 Jun 01;29(3):540-546. [doi: 10.1093/eurpub/cky269] [Medline: 30608540]
12. Song Z, Baicker K. Effect of a workplace wellness program on employee health and economic outcomes: a randomized clinical trial. JAMA 2019 Apr 16;321(15):1491-1501 [FREE Full text] [doi: 10.1001/jama.2019.3307] [Medline: 30990549]
13. Lokman S, Volker D, Zijlstra-Vlasveld MC, Brouwers EM, Boon B, Beekman AT, et al. Return-to-work intervention versus usual care for sick-listed employees: health-economic investment appraisal alongside a cluster randomised trial. BMJ Open 2017 Oct 05;7(10):e016348 [FREE Full text] [doi: 10.1136/bmjopen-2017-016348] [Medline: 28982815]
14. Van der Ploeg K, Van der Pal S, de Vroome E, Van den Bossche S. De kosten van ziekteverzuim voor werkgevers in Nederland. TNO. Leiden: TNO; 2014. URL: http://tinyurl.com/mss9hw2j [accessed 2023-11-18]
15. Psychosociale arbeidsbelasting (PSA) werknemers; geslacht en leeftijd. CBS.: Centraal Bureau voor de Statistiek URL: http://tinyurl.com/arc73k6t [accessed 2023-03-19]
16. Psycho-sociale arbeidsbelasting werknemers. CBS Statline.: Centraal Bureau voor de Statistiek URL: http://tinyurl.com/2z56sjz2 [accessed 2023-03-19]
17. Koh J, Tng GY, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022 Aug 25;12(9):1376 [FREE Full text] [doi: 10.3390/jpm12091376] [Medline: 36143161]
18. Paganini S, Meier E, Terhorst Y, Wurst R, Hohberg V, Schultchen D, et al. Stress management apps: systematic search and multidimensional assessment of quality and characteristics. JMIR Mhealth Uhealth 2023 Aug 29;11:e42415 [FREE Full text] [doi: 10.2196/42415] [Medline: 37642999]
19. Quick JC, Wright TA, Adkins JA, Nelson DL, Quick JC. Preventive Stress Management in Organizations. 2nd edition. Washington, DC: American Psychological Association; 2013.
20. Alhasani M, Mulchandani D, Oyebode O, Baghaei N, Orji R. A systematic and comparative review of behavior change strategies in stress management apps: opportunities for improvement. Front Public Health 2022 Feb 24;10:777567 [FREE Full text] [doi: 10.3389/fpubh.2022.777567] [Medline: 35284368]
21. Egger SM, Frey S, Sauerzopf L, Meidert U. A literature review to identify effective web- and app-based mHealth interventions for stress management at work. Workplace Health Saf 2023 Oct 30;71(10):452-463 [FREE Full text] [doi: 10.1177/21650799231170872] [Medline: 37254448]
22. Pospos S, Young IT, Downs N, Iglewicz A, Depp C, Chen JY, et al. Web-based tools and mobile applications to mitigate burnout, depression, and suicidality among healthcare students and professionals: a systematic review. Acad Psychiatry 2018 Feb;42(1):109-120 [FREE Full text] [doi: 10.1007/s40596-017-0868-0] [Medline: 29256033]
23. Salvado M, Marques DL, Pires IM, Silva NM. Mindfulness-based interventions to reduce burnout in primary healthcare professionals: a systematic review and meta-analysis. Healthcare (Basel) 2021 Oct 09;9(10):1342 [FREE Full text] [doi: 10.3390/healthcare9101342] [Medline: 34683022]

XSL•FO
RenderX

24. Mistretta EG, Davis MC, Temkit M, Lorenz C, Darby B, Stonnington CM. Resilience training for work-related stress among health care workers: results of a randomized clinical trial comparing in-person and smartphone-delivered interventions. J Occup Environ Med 2018 Jun;60(6):559-568. [doi: 10.1097/JOM.0000000000001285] [Medline: 29370014]
25. Fiol-DeRoque MA, Serrano-Ripoll MJ, Jiménez R, Zamanillo-Campos R, Yáñez-Juan AM, Bennasar-Veny M, et al. A mobile phone-based intervention to reduce mental health problems in health care workers during the COVID-19 pandemic (PsyCovidApp): randomized controlled trial. JMIR Mhealth Uhealth 2021 May 18;9(5):e27039 [FREE Full text] [doi: 10.2196/27039] [Medline: 33909587]
26. Prado K, Robinson A, Chao YY. The impact of mindful meditation on health care workers during the COVID-19 pandemic. J Nurse Pract 2023 Mar;19(3):104513 [FREE Full text] [doi: 10.1016/j.nurpra.2022.11.023] [Medline: 36748057]
27. Hwang WJ, Jo HH. Evaluation of the effectiveness of mobile app-based stress-management program: a randomized controlled trial. Int J Environ Res Public Health 2019 Nov 03;16(21):4270 [FREE Full text] [doi: 10.3390/ijerph16214270] [Medline: 31684169]
28. Taylor H, Cavanagh K, Field AP, Strauss C. Health care workers' need for headspace: findings from a multisite definitive randomized controlled trial of an unguided digital mindfulness-based self-help app to reduce healthcare worker stress. JMIR Mhealth Uhealth 2022 Aug 25;10(8):e31744 [FREE Full text] [doi: 10.2196/31744] [Medline: 36006668]
29. Hwang WJ, Ha JS, Kim MJ. Research trends on mobile mental health application for general population: a scoping review. Int J Environ Res Public Health 2021 Mar 02;18(5):2459 [FREE Full text] [doi: 10.3390/ijerph18052459] [Medline: 33801537]
30. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr;4:1-32 [FREE Full text] [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
31. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(1):1-39 [FREE Full text] [doi: 10.1348/135910709X466063] [Medline: 19646331]
32. Versluis A, Verkuil B, Spinhoven P, van der Ploeg MM, Brosschot JF. Changing mental health and positive psychological well-being using ecological momentary interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jun 27;18(6):e152 [FREE Full text] [doi: 10.2196/jmir.5642] [Medline: 27349305]
33. Dallery J, Cassidy RN, Raiff BR. Single-case experimental designs to evaluate novel technology-based health interventions. J Med Internet Res 2013 Feb 08;15(2):e22 [FREE Full text] [doi: 10.2196/jmir.2227] [Medline: 23399668]
34. Krasny-Pacini A, Evans J. Single-case experimental designs to assess intervention effectiveness in rehabilitation: a practical guide. Ann Phys Rehabil Med 2018 May;61(3):164-179 [FREE Full text] [doi: 10.1016/j.rehab.2017.12.002] [Medline: 29253607]
35. Epstein LH, Bickel WK, Czajkowski SM, Paluch RA, Moeyaert M, Davidson KW. Single case designs for early phase behavioral translational research in health psychology. Health Psychol 2021 Dec;40(12):858-874 [FREE Full text] [doi: 10.1037/hea0001055] [Medline: 34370494]
36. Epstein LH, Dallery J. The family of single-case experimental designs. Harv Data Sci Rev 2022 Sep 08;4(SI3) [FREE Full text] [doi: 10.1162/99608f92.ff9300a8] [Medline: 36926648]
37. Stratton E, Lampit A, Choi I, Calvo RA, Harvey SB, Glozier N. Effectiveness of eHealth interventions for reducing mental health conditions in employees: a systematic review and meta-analysis. PLoS One 2017 Dec 21;12(12):e0189904 [FREE Full text] [doi: 10.1371/journal.pone.0189904] [Medline: 29267334]
38. Olivier-Pijpers V, Vugts B. Signaleringsplan: Zegen of Vloek?. Australiëlaan, The Netherlands: Centrum voor Consultatie en Expertise; Mar 2016.
39. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
40. Klein EM, Brähler E, Dreier M, Reinecke L, Müller KW, Schmutzer G, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16:159 [FREE Full text] [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
41. Chesney MA, Neilands TB, Chambers DB, Taylor JM, Folkman S. A validity and reliability study of the coping self-efficacy scale. Br J Health Psychol 2006 Sep;11(3):421-437 [FREE Full text] [doi: 10.1348/135910705X53155] [Medline: 16870053]
42. Schaufeli WB, Desart S, De Witte H. Burnout Assessment Tool (BAT)-development, validity, and reliability. Int J Environ Res Public Health 2020 Dec 18;17(24):9495 [FREE Full text] [doi: 10.3390/ijerph17249495] [Medline: 33352940]
43. Hadžibajramović E, Schaufeli W, De Witte H. Shortening of the Burnout Assessment Tool (BAT)-from 23 to 12 items using content and Rasch analysis. BMC Public Health 2022 Mar 22;22:560 [FREE Full text] [doi: 10.1186/s12889-022-12946-y] [Medline: 35313849]
44. Declercq L, Jamshidi L, Fernández-Castilla B, Beretvas SN, Moeyaert M, Ferron JM, et al. Analysis of single-case experimental count data using the linear mixed effects model: a simulation study. Behav Res Methods 2019 Dec;51(6):2477-2497. [doi: 10.3758/s13428-018-1091-y] [Medline: 30105444]
45. DeHart WB, Kaplan BA. Applying mixed-effects modeling to single-subject designs: an introduction. J Exp Anal Behav 2019 Mar;111(2):192-206. [doi: 10.1002/jeab.507] [Medline: 30758051]

XSL•FO
RenderX

46. Hoffman L, Rovine MJ. Multilevel models for the experimental psychologist: foundations and illustrative examples. Behav Res Methods 2007 Mar;39(1):101-117. [doi: 10.3758/bf03192848] [Medline: 17552476]
47. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. Arch Gen Psychiatry 2004 Mar;61(3):310-317 [FREE Full text] [doi: 10.1001/archpsyc.61.3.310] [Medline: 14993119]
48. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. The job demands-resources model of burnout. J Appl Psychol 2001 Jun;86(3):499-512 [FREE Full text] [Medline: 11419809]
49. Baker JP, Berenbaum H. Emotional approach and problem-focused coping: a comparison of potentially adaptive strategies. Cogn Emot 2007 Feb 3;21(1):95-118. [doi: 10.1080/02699930600562276]
50. Phillips EA, Gordeev VS, Schreyögg J. Effectiveness of occupational e-mental health interventions: a systematic review and meta-analysis of randomized controlled trials. Scand J Work Environ Health 2019 Nov 01;45(6):560-576 [FREE Full text] [doi: 10.5271/sjweh.3839] [Medline: 31184758]
51. Osman I, Hamid S, Singaram VS. Efficacy of a brief online mindfulness-based intervention on the psychological well-being of health care professionals and trainees during the COVID-19 pandemic: a mixed method design. Health SA 2021 Sep 30;26 [FREE Full text] [doi: 10.4102/hsag.v26i0.1682] [Medline: 34667655]
52. Ebert DD, Heber E, Berking M, Riper H, Cuijpers P, Funk B, et al. Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 2016 May;73(5):315-323 [FREE Full text] [doi: 10.1136/oemed-2015-103269] [Medline: 26884049]
53. López-Del-Hoyo Y, Fernández-Martínez S, Pérez-Aranda A, Barceló-Soler A, Bani M, Russo S, et al. Effects of eHealth interventions on stress reduction and mental health promotion in healthcare professionals: a systematic review. J Clin Nurs 2023 Sep;32(17-18):5514-5533 [FREE Full text] [doi: 10.1111/jocn.16634] [Medline: 36703266]
54. Lobo MA, Moeyaert M, Baraldi Cunha A, Babik I. Single-case design, analysis, and quality assessment for intervention research. J Neurol Phys Ther 2017 Jul;41(3):187-197 [FREE Full text] [doi: 10.1097/NPT.0000000000000187] [Medline: 28628553]
55. Smith JD. Single-case experimental designs: a systematic review of published research and current standards. Psychol Methods 2012 Dec;17(4):510-550 [FREE Full text] [doi: 10.1037/a0029312] [Medline: 22845874]
56. Dallery J, Raiff BR. Optimizing behavioral health interventions with single-case designs: from development to dissemination. Transl Behav Med 2014 Sep;4(3):290-303 [FREE Full text] [doi: 10.1007/s13142-014-0258-z] [Medline: 25264468]
57. van Wijk M. Arbeidsmarktprofiel van zorg en welzijn. Centraal Bureau voor de Statistiek.: CBS; 2020. URL: http://tinyurl.com/upt93ynn [accessed 2023-11-18]
58. Health and care workforce in Europe: time to act. World Health Organization. Copenhagen: The WHO Regional Office for Europe; 2022. URL: https://www.who.int/europe/publications/i/item/9789289058339 [accessed 2023-03-04]
59. Burney V, Arnold-Saritepe A, McCann CM. Rethinking the place of qualitative methods in behavior analysis. Perspect Behav Sci 2023 Mar;46(1):185-200 [FREE Full text] [doi: 10.1007/s40614-022-00362-x] [Medline: 37006599]
60. Bonell C, Warren E, Melendez-Torres GJ. Methodological reflections on using qualitative research to explore the causal mechanisms of complex health interventions. Evaluation 2022 Apr;28(2):166-181. [doi: 10.1177/13563890221086309]

Abbreviations

BAT: Burnout Assessment Tool
CSES: Coping Self-Efficacy Scale
EMA: ecological momentary assessment
EMI: ecological momentary intervention
LMM: linear mixed model
PSM: preventive stress management
PSS: Perceived Stress Scale
RCT: randomized controlled trial
SCED: single-case experimental design

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 13.05.23; peer-reviewed by E Phillips, A Ferrario; comments to author 18.10.23; revised version received 06.12.23; accepted 17.01.24; published 29.02.24.

Please cite as:
Demirel S, Roke Y, Hoogendoorn AW, Hoefakker J, Hoeberichts K, van Harten PN
Assessing the Effectiveness of STAPP@Work, a Self-Management Mobile App, in Reducing Work Stress and Preventing Burnout: Single-Case Experimental Design Study
J Med Internet Res 2024;26:e48883
URL: https://www.jmir.org/2024/1/e48883
doi:10.2196/48883
PMID:38275128

©Sevda Demirel, Yvette Roke, Adriaan W Hoogendoorn, Jamie Hoefakker, Kirsten Hoeberichts, Peter N van Harten. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile-Delivered Mindfulness Intervention on Anxiety Level Among College Athletes: Randomized Controlled Trial

Yu Gao[1*], MA; Lu Shi[2*], PhD; Ning Fu[3], PhD; Nan Yang[1], PhD; Tracy Weeks-Gariepy[4], PhD; Yuping Mao[5], PhD

[1]Shanghai University of Finance and Economics, Shanghai, China

[2]Department of Health Science, Pace University, New York, NY, United States

[3]School of Public Administration and Emergency Management, Jinan University, Guangzhou, China

[4]Department of Public Health Sciences, Clemson University, Clemson, SC, United States

[5]Department of Communication Studies, California State University Long Beach, Long Beach, CA, United States

[*]these authors contributed equally

Corresponding Author:
Lu Shi, PhD
Department of Health Science
Pace University
One Pace Plaza
New York, NY, 10038
United States
Phone: 1 (866) 722 3338
Email: lshi@pace.edu

Abstract

Background: College athletes are a group often affected by anxiety. Few interventional studies have been conducted to address the anxiety issues in this population.

Objective: We conducted a mobile-delivered mindfulness intervention among college athletes to study its feasibility and efficacy in lowering their anxiety level and improving their mindfulness (measured by the Five Facet Mindfulness Questionnaire [FFMQ]).

Methods: In April 2019, we recruited 290 college athletes from a public university in Shanghai, China, and 288 of them were randomized into an intervention group and a control group (closed trial), with the former (n=150) receiving a therapist-guided, smartphone-delivered mindfulness-based intervention and the latter receiving mental health promotion messages (n=138). We offered in-person instructions during the orientation session for the intervention group in a classroom, with the therapist interacting with the participants on the smartphone platform later during the intervention. We used generalized linear modeling and the intent-to-treat approach to compare the 2 groups' outcomes in dispositional anxiety, precompetition anxiety, and anxiety during competition, plus the 5 dimensions of mindfulness (measured by the FFMQ).

Results: Our intent-to-treat analysis and generalized linear modeling found no significant difference in dispositional anxiety, precompetition anxiety, or anxiety during competition. Only the “observation” facet of mindfulness measures had a notable difference between the changes experienced by the 2 groups, whereby the intervention group had a net gain of .214 yet fell short of reaching statistical significance (P=.09). Participants who specialized in group sports had a higher level of anxiety (β=.19; SE=.08), a lower level of “nonjudgemental inner experience” in FFMQ (β=–.07; SE=.03), and a lower level of “nonreactivity” (β=–.138; SE=.052) than those specializing in individual sports.

Conclusions: No significant reduction in anxiety was detected in this study. Based on the participant feedback, the time availability for mindfulness practice and session attendance for these student athletes in an elite college could have compromised the intervention’s effectiveness. Future interventions among this population could explore a more student-friendly time schedule (eg, avoid final exam time) or attempt to improve cognitive and scholastic outcomes.

Trial Registration: Chinese Clinical Trial Registry ChiCTR1900024449; https://www.chictr.org.cn/showproj.html?proj=40865

(J Med Internet Res 2024;26:e40406) doi:10.2196/40406

KEYWORDS

anxiety; athletes; body; calmness; cognition; college students; college; feasibility; feedback; intervention; meditation; mHealth; mindfulness; mobile; participant; positive; program; relaxation; sleep; students

XSL•FO
RenderX

Introduction

It has been observed that anxiety disorders are not uncommon among college athletes [1], a group of young individuals who typically need to balance daily sports training, major competitive tournaments, and college coursework. These workloads could mean substantial stress and anxiety for these young students, many of whom have just left home for a new location without local support from family or social networks. Unique challenges such as the long-term impact of concussions (traumatic brain injury) on the neural system could also elevate the mental health risk for the college athlete population [2,3]. Therefore, more academic attention is needed to address the mental health challenges among college athletes during and after their college years.

Clinical resources such as mental health counseling, psychotherapy, and student services centers are common health care resources on campus that college athletes could turn to for their mental health issues. Common challenges for college athletes attempting to use mental health resources include perceived stigma associated with seeking mental health care [4] and service availability [5], barriers that are unlikely to disappear soon. As such, lifestyle interventions, which could be sustained by the college athletes themselves after active intervention programs, could play complementary roles to clinical interventions such as counseling and outpatient visits.

Mindfulness-based interventions (MBIs), a type of lifestyle intervention aimed at developing a nonjudgmental awareness and experiential acceptance of present experiences (often through meditation practice) [6], have been shown to be effective in reducing stress and anxiety among adolescents and college-age young adults [7-9]. Its nonpharmacological, nonclinical approach makes it feasible for nonclinical populations, such as college athletes. Recent evidence has shown MBIs' feasibility and acceptability when delivered through mobile phone–based apps such as Headspace [10,11]. On the other hand, the traditional 8-week time frame of standardized MBIs such as mindfulness-based stress reduction (MBSR) and mindfulness behavioral cognitive therapy (MBCT) [12] might be too challenging for college students, who reported time availability as the most common barrier to meditation practice [13,14]. For college athletes who have to maintain rigorous sports training and academic coursework simultaneously, the time constraint might be even more salient than for other college students. Therefore, a smartphone-based MBI with more flexible session scheduling (eg, the content is delivered with the option of asynchronous attendance) could be more accessible for college athletes. However, while smartphone-based MBIs have been assessed among the adult population [15,16] as well as university students [17], very few such interventions have been tested among college athletes. To fill this research gap, we conducted a therapist-guided, asynchronously delivered mobile health MBI among college athletes to study its feasibility and efficacy in lowering their anxiety level and improving their mindfulness, as measured by the Five Facet Mindfulness Questionnaire (FFMQ) [18].

Methods

Design

We designed and conducted a randomized controlled trial whereby college athletes were randomized to receive a smartphone-based, therapist-led mindfulness-based intervention, with anxiety measures as our primary outcome of interest (Figure 1).

Figure 1. Study design flowchart.

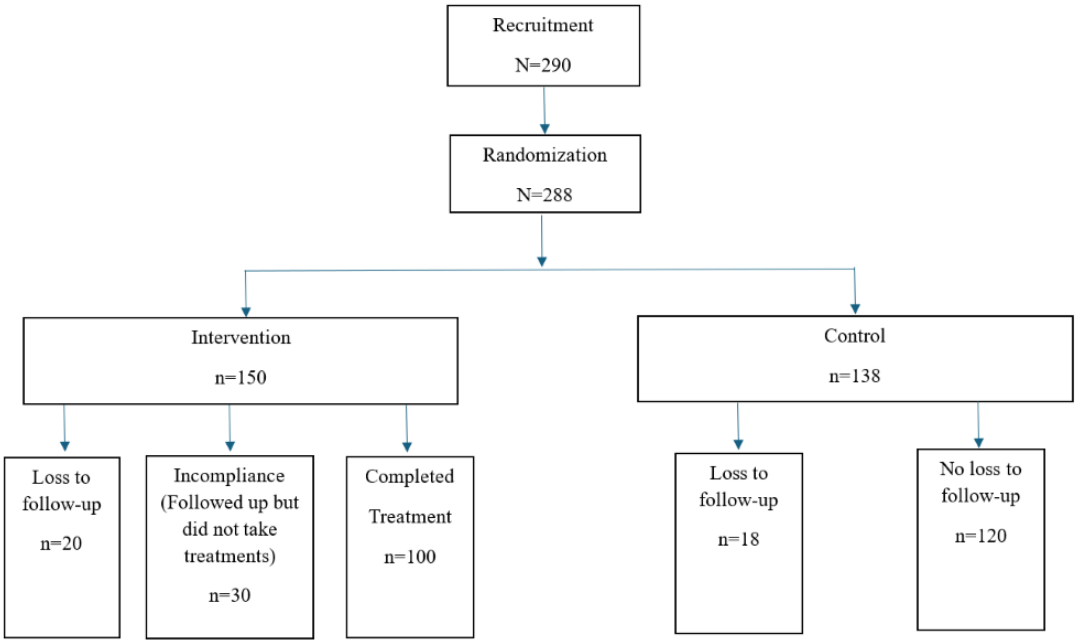

XSL•FO
RenderX

Setting

In April 2019 (Spring semester in Chinese universities), we offered in-person instructions during the orientation session for the intervention group in a classroom at Shanghai University of Finance and Economics in Shanghai, China. After the orientation session, a therapist interacted with the participants on a smartphone platform to address their day-to-day feedback about the video-delivered MBSR sessions.

Participants

Using the G*Power 3 [19] software and previous meta-analyses of the effect size of mindfulness interventions on anxiety [20], we estimated that we would need a sample size of 290 cases for the study.

The inclusion criteria are the following: (1) currently enrolled as college student athletes; (2) willing to download a WeChat app on their cellphone, conduct 15 minutes of mindfulness meditation following the courses in the app, and report their feelings subsequently on a daily basis; and (3) able to participate in routine school activities including study, exams, athletic training, and competitions. The second and third criteria here ensure computer and internet literacy, given the selectivity of the university.

Randomization and Intervention

A total of 290 college student athletes were enrolled (closed trial) by Professor Yu Gao in April 2019. Before our random allocation, participants were asked to fill out the baseline survey about their mental health status, health behavior, and social demographic characteristics. A total of 2 (0.69%) students chose to leave the study at the point of the baseline survey without giving a reason. After the baseline phase, Dr Shi randomized the 288 participants into control and intervention groups using the permuted block randomization method [21], with Professor Gao's research team assigning 150 to the intervention group and 138 to the control group. The allocation list was stored on a password-protected computer at Clemson University in Clemson, South Carolina, United States.

On April 9, 2019, the intervention group met and was provided with an in-person orientation session about MBSR [22]. The session provided knowledge about the origin and development of meditation practice, and the whole group practiced a 20-minute mindfulness meditation under instructions. During this session, participants downloaded a mindfulness app on their cellphones that included a series of instructional videos about mindfulness practice in daily life and guidance from certified mindfulness therapists. This app was developed on the social media platform of WeChat, the most popular social media platform in mainland China [23,24]. Following the orientation session, the intervention group practiced daily mindfulness meditation for 15 minutes following the daily video update in the app, while documenting their reflections about the meditation experience in the shared web-based space. The intervention lasted 43 days between April 9 and May 21, 2019. It was implemented by 2 health care professionals (a primary care physician and a licensed counselor) who had extensive experience with mindfulness training. We adapted the standardized MBSR protocol for this smartphone-based intervention to fit the mobile platform.

For the 6-week duration of the intervention, the experimental group listened to 15-minute audio clips of guided meditation (mindful yoga, an integral part of the MBSR protocol [25,26]) while providing their questions and feedback about the intervention to the mobile platform. The participants were expected to submit a daily check-in message to briefly update their mindfulness practice experience. Common feedback from the student participants included the difficulty for the mind to stay quiet and the inability to control their own breath. The licensed mindfulness therapist then provided her real-time answers and advice on questions and feedback during a daily 15-minute live session. The therapist reminded student participants that it was important not to let the intervention program interfere with their routine curriculum and daily lives.

We provided the control group with informational pamphlets about college students' mental and physical health, but no active behavioral intervention was conducted. They were not required to meet during the intervention. We only contacted this group once at the end of the study to fill out the end point survey.

The third day after the treatment finished, our research assistants administered the posttreatment survey among both the intervention and control groups. A total of 2 research assistants independently entered the participants' questionnaire responses to build Excel (Microsoft Corporation)-format data sets and then cross-checked the accuracy of the data entry to ensure the quality of the survey data. By the end of the experiment, 20 participants and 18 participants were lost to follow-up from the intervention and control groups, respectively.

Blinding

While participants and instructors were not blinded to intervention, statistical data analysts were blinded to group assignment when conducting the analyses using STATA/MP (version 14.0; StataCorp) [27].

Outcome Measures

The primary outcome of interest in this study is assessed by 3 anxiety measures: participants' level of dispositional anxiety, precompetition anxiety, and anxiety during competition, through online questionnaires. These were measured by a Chinese version of 3 validated psychometric scales among the athlete population [28]: "athlete trait anxiety scale," "athlete state anxiety scale (before competition)," and "athlete state anxiety scale (in competition)." These 3 instruments comprised a total of 28 items rated using a Likert scale with 5 options ranging from 1 (strongly disagree) to 5 (strongly agree). Multimedia Appendix 1 contains the questionnaire questions. These scales were based on Spielberger's State-Trait Anxiety Inventory [29] and Sport Competition Anxiety Test [30] and were adopted for the Chinese athlete population. They were chosen based on validity and reliability. All scales had adequate construct validity and reliability to assess mental health in sports. Wu and Lin [28] reported Cronbach α values of each of the 3 scales: athlete trait anxiety (α=0.868), precompetition anxiety (α=0.833), and anxiety during competition (α=0.848).

XSL•FO
RenderX

The secondary outcomes in this study include the self-assessed FFMQ: observation, description, mindful actions, nonjudgmental inner experience, and nonreactivity [18] (from the Chinese version of the FFMQ [31]).

The student participants' reflections and feedback were submitted in simplified Mandarin Chinese, the official language on mainland China. Their comments were selected and translated into English to be included in this study. To ensure the English translation accurately captures the meaning in the statements in simplified Mandarin Chinese, the translation followed by the back-translation procedure was applied by Chinese-born US-based researchers who are fluent in both English and Chinese and have rich experience conducting transnational research.

Statistical Analysis

Baseline characteristics were described by the intervention groups and compared for homogeneity using the 2-tailed Student *t* test for continuous variables and the chi-square test for noncontinuous variables. An intent-to-treat approach was used for all effectiveness analyses. All participants were analyzed according to the groups to which they were originally randomized, accounting for the loss to follow-up. Descriptive summary statistics of outcome variables were presented for the pre-, post-, and between-group differences. We also used generalized linear mixed models, adjusting for baseline characteristics [32]. Treatment condition, time, and the interaction of treatment-by-time were included, with treatment-by-time as fixed factors and time as a random factor. The treatment*time interaction term captured the effectiveness of the treatment. The treatment analysis was conducted by excluding untreated participants who were assigned to the treatment group. In addition, we performed sensitivity analyses to study if the treatment was more effective for participants who reported high anxiety levels at baseline by running 2 separate analyses for participants with high versus low anxiety levels.

The statistically significant level was set at $P<.05$ based on 2-tailed *t* tests. The data were analyzed using STATA/MP (version 14.0).

Qualitative Analysis

There are 3446 records of qualitative comments made by 102 unique participants. In analyzing the qualitative comments provided by participants, we adopted an interdisciplinary approach. A total of 3 researchers, who specialize in sports education, public health, and communication, read the comments independently and identified key themes. They discussed disagreements and reached consensus on major themes. NVivo (version 12) [33] was used to perform qualitative analysis based on the frequency of keywords. Researchers drew key themes based on frequency and relevance. After the key themes were agreed upon, representative comments were selected to demonstrate each theme.

Ethical Considerations

Upon enrollment, we provided the participants with an overview of the mindfulness anxiety reduction experiment program and a copy of the consent form for every participant to sign. We obtained informed consent before randomization.

The study was approved by the Office of Scientific Research at Shanghai University of Finance and Economics (review number HIRB SOP/06/3.0) and was retrospectively registered at the Chinese Clinical Trial Registry (ChiCTR1900024449). Due to the college schedule of the final exam week and summer break (when the students would be unavailable for intervention), the approval of our clinical trial registration occurred after the initiation of participant enrollment, since we were unable to logistically defer the enrollment to accommodate an unexpected delay in trial registration approval. This study's data were deidentified after data entry. Every participant received a participation incentive of a Shanghai Disneyland ticket worth US $59.46.

Results

Overview

At baseline, our total sample (N=288) included 62.8% (181/288) male participants and 37.2% (107/288) female participants, with an age range between 17 and 25 (mean 19.4, SD 1.5) years (Table 1). From the 288 participants, 83 (28.8%) participated in group or team sports including soccer, volleyball, cheerleading, and basketball; 205 (71.2%) participated in individual sports including badminton, bridge, chess, fencing, golf, pingpong or table tennis, skipping rope, swimming, taekwondo, and tennis. In our data, the Cronbach α was 0.923 for anxiety of trait, 0.906 for precompetition anxiety, and 0.861 for anxiety during the competition. The reliability coefficient for each scale was high, indicating the questionnaires were reliable. We used independent 2-tailed *t* tests and chi-square tests to compare the baseline characteristics, anxiety outcomes, and FFMQ outcomes between the intervention and control groups and found that for these variables there was no statistically significant difference between the 2 groups (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics.

Characteristics	Total (N=288)	Intervention group (n=150)	Control group (n=138)	*P* value
Age (years), mean (SD)	19.4 (1.5)	19.5 (1.6)	19.3 (1.4)	.26
Sex, n (%)				.76
Female	181 (62.8)	93 (62.0)	88 (63.8)	
Male	107 (37.2)	57 (38.0)	50 (36.2)	
Residency, n (%)				.99
Urban	238 (82.6)	124 (82.7)	114 (82.6)	
Rural	50 (17.4)	26 (17.3)	24 (17.4)	
Group sports, n (%)				.65
Group sports	83 (28.8)	45 (30.0)	38 (27.5)	
Individual sports	205 (71.2)	105 (70.0)	100 (72.5)	
Dispositional anxiety, mean (SD)	2.2 (0.8)	2.7 (0.8)	2.2 (0.8)	.35
Precompetition anxiety, mean (SD)	1.9(0.7)	1.9 (0.7)	1.8 (0.7)	.75
Anxiety during competition, mean (SD)	2.1 (0.6)	2.1 (0.6)	2.1 (0.6)	.96
FFMQ[a]-observe, mean (SD)	3.1 (0.7)	3.1 (0.7)	3.1 (0.7)	.62
FFMQ-describe, mean (SD)	3.3 (0.6)	3.3 (0.6)	3.4 (0.6)	.59
FFMQ-aware, mean (SD)	3.6 (0.6)	3.6 (0.6)	3.6 (0.6)	.76
FFMQ-nonjudgemental, mean (SD)	3.0 (0.6)	3.0 (0.6)	3.0 (0.6)	.72
FFMQ-nonreactivity, mean (SD)	2.9 (0.6)	2.9 (0.6)	2.90 (0.6)	.81

[a]FFMQ: Five Facet Mindfulness Questionnaire.

Social Media Messages

Qualitative feedback from participants, saved as social media messages, showed positive feedback from the participants on the content of the experimental intervention. Overall, these messages showed that it was feasible to promote the meditation intervention through the mobile phone–based platform among these college athletes. At the end of the course, many expressed their willingness to repeat the course.

> *I hope I can repeat this course, I feel used to listening to the content of the positive meditation for a period of time every day, it can make me feel very relaxed.*

While those in the intervention group were told to leave a daily check-in message to give a 1-sentence note about their meditation experience of the day (as a way of keeping the participants engaged as well as a way of providing instant feedback for the therapist), on average, a participant in the intervention group only left his or her message for 53.4% (6613 check-in messages out of 12,384 person days) of the time throughout the 43-day intervention.

Among the 3446 social media messages, around 57.8% (1991/3446) are simple “check-in” messages without additional comments. The qualitative analysis of the social media messages with additional comments shows participants reported 5 main effects of the meditation: sleepiness, calmness, relaxation, body sensation, and decentering. Although these effects are distinctively different from each other, they are interconnected. Therefore, it is common for some participants to share more than 1 effect in 1 social media message.

Sleepiness emerged as a major effect that participants experienced and shared. A total of 41 comments indicated a “drowsy” or “sleepy” feeling during or after the meditation session. For example, participant F commented, “My spine feels relaxed, my headache is soothed, and my anxiety is reduced. I have the feeling of falling sleep.” Participants also believe falling asleep is likely to be a common experience, as participant E expressed: “I fell asleep during meditation...I should not be the only one though...”

Sleepiness and calmness are sometimes experienced simultaneously in meditation, as participant W reflected: “Ah, it is close to exam time again. The meditation makes me calm and sleepy, then I don’t feel so nervous.” Participants also reported that the calmness that the meditation brings to them helps their performance in competitive and high-anxiety situations. Participant E provided the following personal example:

> *My participation in the meditation program helps a lot. In the championship competition of the university’s basketball teams, our rope skipping team was invited to perform before the competition. In the rehearsal, I felt extremely nervous, then I suddenly remembered the techniques to control breath in the meditation, and this helped me to calm down gradually. In the formal performance, although I still felt a bit nervous, it was much better than before, and I had few mistakes.*

XSL•FO
RenderX

Participant A described the following: "feel extremely calm, sense the blood flowing in my body, and most importantly, forget the annoying things."

Relaxation also emerged as a common experience among participants, with 51 comments including the keyword "relaxation." Participant J shared:

> *After meditation, I feel much relaxed, and all the pressure disappeared suddenly. I have been very busy with reviewing for exams, and squeezing some time for meditation is very helpful.*

Some participants identify relaxation as one of the major benefits they get from the multiple meditation sessions. Participant S's social media message well summarizes the short- and long-term effects of relaxation:

> *1. After this morning's meditation exercise, I felt much relaxed. I have so many different thoughts lately and want to do different things well. However, the more thoughts, the more I feel my ability falls short of my wishes. Today's meditation made me feel more relaxed.*
>
> *2. After 20-day meditation, I feel good. I feel more relaxed every morning and my emotions are more stable.*

Decentering is defined as one's ability to observe his or her thoughts and feelings in a detached manner [34]. Relaxation sometimes helps participants to decenter themselves, as participant J shared:

> *I felt very relaxed, like after a deep sleep, my body also felt more relaxed. I learned how to experience things around me such as sound, light, and texture...I don't think too much, but focus on the experience in the moment, feel immersed in this atmosphere.*

Body sensations reported by the participants included sleepiness, relaxation, and calmness. Participant M's message shows the process of how body sensation connects with relaxation and calmness:

> *My body feelings changed from tense to relaxed, then relaxed to tense, then tense to relaxed. My brain thought about the upcoming exam, and I noticed the audio in the meditation, and I thought about things happened today and in the past. My emotion shifted from nervous to relaxed and calm, then from calm to anxious, and then from anxious to calm.*

Although participants mainly commented on the positive effects of the meditation sessions, they also shared insights on factors that could compromise the effects of the meditation. In particular, they identified the following factors: distraction in the environment, lack of time, lack of an incentive and supervision system, and lack of consistent long-term participation. Participant M elaborated on the concerns:

> *This meditation activity takes away my time for reviewing study materials, so I don't think I will be able to keep participating. There are so many exams and other things during this period. I don't think many people will be able to finish the entire meditation program if there isn't a good system of incentives and supervision. Everyone has so many things, and it is not a simple task to spend more than ten minutes per day to watch the meditation video. Even for those who check in everyday, it is hard to tell whether they listened with attention.*

Statistical Analysis Results

As Table 2 shows, for the 3 anxiety scales and the 5 facets of mindfulness, none of the within-group changes from the baseline were statistically significant. Additionally, the between-group difference-in-difference analyses did not reveal significant between-group differences in their changes from the baseline. As a sensitivity analysis, our intent-to-treatment analysis revealed similar patterns whereby no significant effect of the intervention was revealed.

In the generalized linear modeling analysis (Table 3), indicator variables are included for group or individual sports and urban or rural residency. The overall patterns of interventional effectiveness remain statistically insignificant (dispositional anxiety β=–.043; *P*=.71; pretournament anxiety β=.013; *P*=.89; and midtournament anxiety β=–.01; *P*=.91). There is no significant role for urban or rural residency in anxiety levels or FFMQ outcomes. Only the "observation" facet of mindfulness measures had some notable difference between the changes experienced by the 2 groups, whereby the intervention group had a net gain of .214 without reaching the common threshold of statistical significance (*P*=.09).

As for the covariates used in the generalized linear modeling model, our results show that participants who specialized in group sports had a higher level of anxiety (β=.19, SE=.08; *P*=.02), a lower level of "nonjudgemental inner experience" in FFMQ (β=–.07, SE=.03; *P*=.03), and a lower level of "nonreactivity" (β=–.138, SE=.052; *P*=.01).

Stratified generalized linear modeling analyses by baseline dispositional anxiety (Table 4) do not identify notable between-group differences in changes from baseline (for participants with a high level of dispositional anxiety at baseline, dispositional anxiety β=.05; *P*=.73; precompetition anxiety β=.03; *P*=.69; and anxiety during competition β=.02; *P*=.82; while for participants with a low level of dispositional anxiety at baseline, dispositional anxiety β=.003; *P*=.97; precompetition anxiety β=.02; *P*=.80; and anxiety during competition β=.004; *P*=.94).

XSL•FO
RenderX

Table 2. Outcome measures at baseline and posttreatment by groups.

Outcome	Total, mean (SE)	Intervention group, mean (SE)	Control group, mean (SE)	Between-treatment differences in improvement from baseline, mean (SE)	*P* value
Dispositional anxiety					
Baseline	2.22 (0.04)	2.26 (0.06)	2.18 (0.07)	–0.08 (0.09)	.35
Posttreatment	2.11 (0.04)	2.13 (0.07)	2.09 (0.07)	–0.04 (0.09)	.64
Difference in difference	N/A[a]	N/A	N/A	–0.04 (0.13)	.77
Precompetition anxiety					
Baseline	1.85 (0.04)	1.86 (0.05)	1.83 (0.06)	0.02 (0.08)	.79
Posttreatment	1.96 (0.04)	1.98 (0.06)	1.94 (0.06)	0.04 (0.09)	.64
Difference in difference	N/A	N/A	N/A	0.02 (0.12)	.87
Anxiety during competition					
Baseline	2.10 (0.03)	2.10 (0.05)	2.10 (0.05)	–0.01 (0.07)	.86
Posttreatment	2.19 (0.04)	2.19 (0.05)	2.20 (0.06)	–0.01 (0.08)	.94
Difference in difference	N/A	N/A	N/A	–0.01 (0.10)	.93
FFMQ[b]-observe					
Baseline	3.08 (0.04)	3.06 (0.06)	3.11 (0.06)	–0.05 (0.08)	.56
Posttreatment	3.04 (0.04)	3.10 (0.06)	2.99 (0.06)	0.11 (0.09)	.23
Difference in difference	N/A	N/A	N/A	0.16 (0.12)	.20
FFMQ-describe					
Baseline	3.34 (0.04)	3.32 (0.05)	3.36 (0.05)	–0.04 (0.07)	.55
Posttreatment	3.32 (0.04)	3.33 (0.05)	3.30 (0.05)	0.03 (0.07)	.74
Difference in difference	N/A	N/A	N/A	0.07 (0.11)	.52
FFMQ-aware					
Baseline	3.60 (0.04)	3.60 (0.05)	3.60 (0.05)	–0.001 (0.07)	.89
Posttreatment	3.50 (0.04)	3.51 (0.06)	3.49 (0.06)	0.03 (0.08)	.76
Difference in difference	N/A	N/A	N/A	0.04 (0.11)	.75
FFMQ-nonjudgmental					
Baseline	3.02 (0.04)	3.01 (0.05)	3.04 (0.05)	–0.03 (0.07)	.71
Posttreatment	3.04 (0.04)	3.01 (0.05)	3.07 (0.05)	–0.06 (0.07)	.41
Difference in difference	N/A	N/A	N/A	–0.03 (0.10)	.75
FFMQ-nonreactivity					
Baseline	2.91 (0.03)	2.91 (0.05)	2.90 (0.05)	0.016 (0.07)	.81
Posttreatment	3.08 (0.04)	3.10 (0.05)	3.05 (0.05)	0.05 (0.71)	.51
Difference in difference	N/A	N/A	N/A	0.03 (0.10)	.75

[a]N/A: not applicable.

[b]FFMQ: Five Facet Mindfulness Questionnaire.

XSL•FO
RenderX

Table 3. Generalized linear mixed models about the intervention effectiveness.

Outcome	Time effect		Group effect		Time*group effect	
	β	*P* value	β	*P* value	β	*P* value
Dispositional anxiety	–.09	.36	.08	.37	–.04	.71
Precompetition anxiety	.11	.10	.03	.70	.01	.89
Anxiety during competition	.11	.09	–.01	.90	–.01	.91
FFMQ[a]-observe	–.18	.07	–.04	.08	.21	.10
FFMQ-describe	–.06	.44	–.03	.66	.06	.37
FFMQ-aware	–.13	.11	–.02	.75	.05	.67
FFMQ-nonjudgemental	.03	.62	–.04	.57	–.03	.69
FFMQ-nonreactivity	.15	.01	.02	.73	.03	.78

[a]FFMQ: Five Facet Mindfulness Questionnaire.

Table 4. Generalized linear mixed model analysis examining the intervention effectiveness, as stratified by baseline anxiety.

Outcome	Time effect		Group effect		Time*group	
	Regression slope	*P* value	Regression slope	*P* value	Regression slope	*P* value
High levels of anxiety at baseline						
Dispositional anxiety	–.06	.53	–.01	.89	.05	.73
Precompetition anxiety	.01	.92	–.01	.95	.03	.69
Anxiety during competition	–.01	.95	–.13	.05	.02	.82
Low levels of anxiety at baseline						
Dispositional anxiety	.03	.49	.01	.86	.003	.97
Precompetition anxiety	.002	.97	.02	.62	.02	.80
Anxiety during competition	.07	.27	–.03	.51	–.004	.94

Harms

Given the qualitative feedback we checked, no harm was detected as a result of this intervention.

Discussion

Principal Findings

Our 43-day intervention did not achieve statistically significant effect sizes on reducing students' anxiety levels. Additionally, we did not find a significant improvement in the 5 mindfulness facets we measured (observe, describe, aware, nonreactivity, and nonjudgement) as compared with the control group. In other words, our implementation of a mindfulness intervention on a mobile phone–based platform for athletes, a population among whom mental health and disorders were common [35], did not produce significant intervention effects.

Comparison With Other Studies

One possible reason for our insignificant results could be that our sample was recruited neither from clinically diagnosed anxiety patients nor from those who self-report high levels of anxiety. Previous evidence of MBIs' effect has been documented either among populations with elevated levels of anxiety [36-38] or during a time when anxiety tended to be high (eg, during the initial phase of the COVID-19 pandemic [39]). The potential benefit of anxiety reductions from a MBI might not be manifest for a nonclinical young adult population, while future interventions targeting nonclinical populations could have more salient benefits if the intervention program is targeted at those who are screened as having high risk for anxiety disorders upon enrollment. For example, a mobile health mindfulness intervention among university students who wished to lose weight found significant benefits in stress reduction [40], suggesting that the mindfulness app could provide mental health benefits for students who worry about their weight. Our analyses identified those in group sports as having higher levels of anxiety, which indicates that future mindfulness interventions among student athletes might target these students as a high-priority subgroup.

It is important to note that a standardized eight-week MBSR achieved significant improvements in psychological well-being, subjective and objective sleep quality, athletic coping skills, and rowing performance among college athletes [7]. Similarly, a 10-session variant of MBSR (lasting approximately 1 hour for 2 sessions a week) achieved a reduction in perceived stress and an improvement in mindfulness among both athletes and recreationally active people [41]. Compared with these studies, it is possible that the relatively short duration of this study's intervention sessions (as compared to the traditional stress reduction intervention session that could last up to 120 minutes [42]) may explain the lack of measurable short-term effects,

XSL•FO
RenderX

since researchers in the past have noted the possible link between the shorter duration of mindfulness sessions and the lack of significant mental health benefits in mindfulness interventions [43]. While this study demonstrates the feasibility of mobile phone–based mindfulness interventions in a nonclinical college student population, a protocol with a longer duration of active intervention might be needed for better outcomes.

The effect of MBIs has been observed to increase with the age of the participants [44]. Although there are previous examples of significant effects of MBIs in younger populations (eg, prehypertensive adolescents [45]), the magnitude of blood pressure reduction examined was significantly less than for similar interventions in older populations [44]. Here, in addition to the factor of higher baseline blood pressure in the older population with more room for improvement, there is also the possibility that older adults have more time and more leisure to complete the daily exercises provided by the orthostatic instructor. Thus, the statistical insignificance of the intervention benefits documented in this study might be due to the fact that the college athletes did not have much time to practice mindfulness meditation given their busy schedule of sports training, tournament competitions, and academic curriculum. This indicates that it would be advisable for future interventions to be implemented when the young participants have a more flexible schedule.

Limitations

One limitation of this study is that the participants were from a nonprobability college student sample from the economically prosperous city of Shanghai, which might not be very representative of college athletes from other parts of China. Additionally, college athletes in different sports might have different competition seasons and training schedules, and therefore having a uniform intervention and evaluation schedule might not be optimal in terms of timing the intervention delivery and outcome assessment.

We understand that it would have been better to register this trial prospectively as a standard research practice. A longer-term plan (eg, starting trial registration 6 months earlier than the planned enrollment initiation) would have accommodated the unexpected delay in clinical trial registration approval and made this project a stronger study.

Conclusions

MBSR has been conducted by various China-based research groups to address mental health issues among different adult populations [46-50], including young adults such as military recruits [49] and university students [46]. The MBSR study among Chinese college students [46] found out that the reduction of fear of positive emotion was smaller in magnitude than decreases in fear of anger, depression, and anxiety, which may be related to the “Eastern” dialectical cultural script: a hedonic emotion regulation pattern (savoring positive emotion more while dampening it less) has been observed to be less pronounced for Easterners than it was for Westerners [51]. Hence, future MBSR interventions conducted in an East Asian context might find it helpful to emphasize the savoring component in its current curriculum to reduce the participants’ potential fear of positive emotion, a possible predictor of depression, anxiety, and stress [52].

An area to explore for future interventions among college athletes is the cognitive and scholastic outcomes of MBIs. Completing college in a timely manner while maintaining an athletic training schedule has been a common challenge for this student population [53]. MBIs have documented success in cognitive outcomes such as attention [54], working memory [55], and scholastic performance [56]. The feasibility and acceptance of our mobile phone–based MBI among this college athlete population, therefore, lays groundwork for future interventions that target these additional outcomes beyond anxiety and stress.

Acknowledgments

This study was funded by the Shanghai Municipal Education Commission.

Conflicts of Interest

None declared.

Editorial notice: This randomized study was only retrospectively registered, explained by authors regarding the timing of participant availability for recruitment to the study. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials, because the risk of bias appears low and the study was considered formative, guiding the evaluation of the application studied. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
Questionnaire for participants (in Chinese).
[DOCX File , 51 KB - jmir_v26i1e40406_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1124 KB - jmir_v26i1e40406_app2.pdf]

References

1. Wilson GS, Raglin JS, Pritchard ME. Optimism, pessimism, and precompetition anxiety in college athletes. Pers Individ Differ 2002;32(5):893-902. [doi: 10.1016/s0191-8869(01)00094-0]
2. Kerr ZY, Evenson KR, Rosamond WD, Mihalik JP, Guskiewicz KM, Marshall SW. Association between concussion and mental health in former collegiate athletes. Inj Epidemiol 2014;1(1):28 [FREE Full text] [doi: 10.1186/s40621-014-0028-x] [Medline: 27747661]
3. Howlett JR, Nelson LD, Stein MB. Mental health consequences of traumatic brain injury. Biol Psychiatry 2022;91(5):413-420 [FREE Full text] [doi: 10.1016/j.biopsych.2021.09.024] [Medline: 34893317]
4. Kaier E, Cromer LD, Johnson MD, Strunk K, Davis JL. Perceptions of mental illness stigma: comparisons of athletes to nonathlete peers. J Coll Stud Dev 2015;56(7):735-739. [doi: 10.1353/csd.2015.0079]
5. Moreland JJ, Coxe KA, Yang J. Collegiate athletes' mental health services utilization: a systematic review of conceptualizations, operationalizations, facilitators, and barriers. J Sport Health Sci 2018;7(1):58-69 [FREE Full text] [doi: 10.1016/j.jshs.2017.04.009] [Medline: 30356496]
6. Ludwig DS, Kabat-Zinn J. Mindfulness in medicine. JAMA 2008;300(11):1350-1352. [doi: 10.1001/jama.300.11.1350]
7. Jones BJ, Kaur S, Miller M, Spencer RMC. Mindfulness-based stress reduction benefits psychological well-being, sleep quality, and athletic performance in female collegiate rowers. Front Psychol 2020;11:572980 [FREE Full text] [doi: 10.3389/fpsyg.2020.572980] [Medline: 33071908]
8. Petterson H, Olson BL. Effects of mindfulness-based interventions in high school and college athletes for reducing stress and injury, and improving quality of life. J Sport Rehabil 2017;26(6):578-587. [doi: 10.1123/jsr.2016-0047] [Medline: 27632857]
9. Gu Y, Xu G, Zhu Y. A randomized controlled trial of mindfulness-based cognitive therapy for college students with ADHD. J Atten Disord 2018;22(4):388-399. [doi: 10.1177/1087054716686183] [Medline: 28038496]
10. Rung AL, Oral E, Berghammer L, Peters ES. Feasibility and acceptability of a mobile mindfulness meditation intervention among women: intervention study. JMIR Mhealth Uhealth 2020 Jun 02;8(6):e15943 [FREE Full text] [doi: 10.2196/15943] [Medline: 32442147]
11. Taylor H, Cavanagh K, Field AP, Strauss C. Health Care workers' need for headspace: findings from a multisite definitive randomized controlled trial of an unguided digital mindfulness-based self-help app to reduce healthcare worker stress. JMIR Mhealth Uhealth 2022 Aug 25;10(8):e31744 [FREE Full text] [doi: 10.2196/31744] [Medline: 36006668]
12. Fjorback LO, Arendt M, Ornbøl E, Fink P, Walach H. Mindfulness-based stress reduction and mindfulness-based cognitive therapy: a systematic review of randomized controlled trials. Acta Psychiatr Scand 2011;124(2):102-119. [doi: 10.1111/j.1600-0447.2011.01704.x] [Medline: 21534932]
13. Gryffin P, Chen W, Erenguc N. Knowledge, attitudes and beliefs of meditation in college students: barriers and opportunities. Am J Educ Res 2014;2(4):189-192 [FREE Full text] [doi: 10.12691/education-2-4-2]
14. tefan CA, Căpraru C, Szilágyi M. Investigating effects and mechanisms of a mindfulness-based stress reduction intervention in a sample of college students at risk for social anxiety. Mindfulness 2018;9(5):1509-1521. [doi: 10.1007/s12671-018-0899-y]
15. Howells A, Ivtzan I, Eiroa-Orosa FJ. Putting the 'app' in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. J Happiness Stud 2014;17(1):163-185. [doi: 10.1007/s10902-014-9589-1]
16. Wylde CM, Mahrer NE, Meyer RML, Gold JI. Mindfulness for novice pediatric nurses: smartphone application versus traditional intervention. J Pediatr Nurs 2017;36:205-212. [doi: 10.1016/j.pedn.2017.06.008] [Medline: 28888505]
17. Lan Y, Ding JE, Li W, Li J, Zhang Y, Liu M, et al. A pilot study of a group mindfulness-based cognitive-behavioral intervention for smartphone addiction among university students. J Behav Addict 2018 Dec 01;7(4):1171-1176 [FREE Full text] [doi: 10.1556/2006.7.2018.103] [Medline: 30418075]
18. de Bruin EI, Topper M, Muskens JGAM, Bögels SM, Kamphuis JH. Psychometric properties of the Five Facets Mindfulness Questionnaire (FFMQ) in a meditating and a non-meditating sample. Assessment 2012;19(2):187-197. [doi: 10.1177/1073191112446654] [Medline: 22589426]
19. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191 [FREE Full text] [doi: 10.3758/bf03193146] [Medline: 17695343]
20. Hofmann SG, Sawyer AT, Witt AA, Oh D. The effect of mindfulness-based therapy on anxiety and depression: a meta-analytic review. J Consult Clin Psychol 2010;78(2):169-183 [FREE Full text] [doi: 10.1037/a0018555] [Medline: 20350028]
21. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. Control Clin Trials 1988;9(4):327-344. [doi: 10.1016/0197-2456(88)90047-5] [Medline: 3203524]
22. Chiesa A, Serretti A. Mindfulness-based stress reduction for stress management in healthy people: a review and meta-analysis. J Altern Complement Med 2009;15(5):593-600. [doi: 10.1089/acm.2008.0495] [Medline: 19432513]

XSL•FO
RenderX

23. Yang M, Jia G, Sun S, Ye C, Zhang R, Yu X. Effects of an online mindfulness intervention focusing on attention monitoring and acceptance in pregnant women: a randomized controlled trial. J Midwifery Womens Health 2019;64(1):68-77. [doi: 10.1111/jmwh.12944] [Medline: 30695166]
24. Montag C, Becker B, Gan C. The multipurpose application WeChat: a review on recent research. Front Psychol 2018;9:2247 [FREE Full text] [doi: 10.3389/fpsyg.2018.02247] [Medline: 30618894]
25. Carmody J, Baer RA. Relationships between mindfulness practice and levels of mindfulness, medical and psychological symptoms and well-being in a mindfulness-based stress reduction program. J Behav Med 2007;31(1):23-33. [doi: 10.1007/s10865-007-9130-7]
26. Haghshenas M, Assarian F, Omidi A, Razaghof M, Rahimi H. Efficacy of mindfulness-based stress reduction in hemodialysis patients with anxiety and depression: a randomized, double-blind, parallel-group trial. Electron physician 2019;11(1):7370-7377 [FREE Full text] [doi: 10.19082/7370]
27. Baum CF, Schaffer ME, Stillman S. Using STATA for applied research: reviewing its capabilities. J Econ Surv 2011;25(2):380-394. [doi: 10.1111/j.1467-6419.2010.00678.x]
28. Wu S, Lin J. Developing anxiety scales for athletes. Tamsui Oxford J Inf Math Sci 2009;6:13.
29. Spielberger CD. State-trait anxiety inventory. In: The Corsini Encyclopedia of Psychology. Hoboken: Wiley; 2010:1.
30. Dunn JG, Dunn JC. Relationships among the sport competition anxiety test, the sport anxiety scale, and the collegiate hockey worry scale. J Appl Sport Psychol 2010;13(4):411-429. [doi: 10.1080/104132001753226274]
31. Deng YQ, Liu XH, Rodriguez MA, Xia CY. The five facet mindfulness questionnaire: psychometric properties of the Chinese version. Mindfulness 2011;2(2):123-128. [doi: 10.1007/s12671-011-0050-9]
32. Stroup WW. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Boca Raton, FL: CRC Press; 2012.
33. Edhlund B, McDougall A. NVivo 12 Essentials. Morrisville, NC: Lulu.com; 2018.
34. Soler J, Franquesa A, Feliu-Soler A, Cebolla A, García-Campayo J, Tejedor R, et al. Assessing decentering: validation, psychometric properties, and clinical usefulness of the experiences questionnaire in a Spanish sample. Behav Ther 2014;45(6):863-871. [doi: 10.1016/j.beth.2014.05.004] [Medline: 25311294]
35. Reardon CL, Hainline B, Aron CM, Baron D, Baum AL, Bindra A, et al. Mental health in elite athletes: International Olympic Committee consensus statement (2019). Br J Sports Med 2019;53(11):667-699 [FREE Full text] [doi: 10.1136/bjsports-2019-100715] [Medline: 31097450]
36. Hölzel BK, Hoge EA, Greve DN, Gard T, Creswell JD, Brown KW, et al. Neural mechanisms of symptom improvements in generalized anxiety disorder following mindfulness training. Neuroimage Clin 2013;2:448-458 [FREE Full text] [doi: 10.1016/j.nicl.2013.03.011] [Medline: 24179799]
37. Hoge EA, Bui E, Marques L, Metcalf CA, Morris LK, Robinaugh DJ, et al. Randomized controlled trial of mindfulness meditation for generalized anxiety disorder: effects on anxiety and stress reactivity. J Clin Psychiatry 2013;74(8):786-792 [FREE Full text] [doi: 10.4088/JCP.12m08083] [Medline: 23541163]
38. Kim YW, Lee SH, Choi TK, Suh SY, Kim B, Kim CM, et al. Effectiveness of mindfulness-based cognitive therapy as an adjuvant to pharmacotherapy in patients with panic disorder or generalized anxiety disorder. Depress Anxiety 2009;26(7):601-606. [doi: 10.1002/da.20552] [Medline: 19242985]
39. Sanilevici M, Reuveni O, Lev-Ari S, Golland Y, Levit-Binnun N. Mindfulness-based stress reduction increases mental wellbeing and emotion regulation during the first wave of the COVID-19 pandemic: a synchronous online intervention study. Front Psychol 2021;12:720965 [FREE Full text] [doi: 10.3389/fpsyg.2021.720965] [Medline: 34858260]
40. Lyzwinski LN, Caffery L, Bambling M, Edirippulige S. The mindfulness app trial for weight, weight-related behaviors, and stress in university students: randomized controlled trial. JMIR Mhealth Uhealth 2019;7(4):e12210 [FREE Full text] [doi: 10.2196/12210] [Medline: 30969174]
41. di Fronso S, Robazza C, Bondár RZ, Bertollo M. The effects of mindfulness-based strategies on perceived stress and psychobiosocial states in athletes and recreationally active people. Int J Environ Res Public Health 2022;19(12):7152 [FREE Full text] [doi: 10.3390/ijerph19127152] [Medline: 35742394]
42. Omidi A, Zargar F. Effect of mindfulness-based stress reduction on pain severity and mindful awareness in patients with tension headache: a randomized controlled clinical trial. Nurs Midwifery Stud 2014;3(3):e21136 [FREE Full text] [doi: 10.17795/nmsjournal21136] [Medline: 25699282]
43. Song Y, Lindquist R. Effects of mindfulness-based stress reduction on depression, anxiety, stress and mindfulness in Korean nursing students. Nurse Educ Today 2015;35(1):86-90. [doi: 10.1016/j.nedt.2014.06.010] [Medline: 25066651]
44. Shi L, Zhang D, Wang L, Zhuang J, Cook R, Chen L. Meditation and blood pressure: a meta-analysis of randomized clinical trials. J Hypertens 2017;35(4):696-706. [doi: 10.1097/HJH.0000000000001217] [Medline: 28033127]
45. Wright LB, Gregoski MJ, Tingen MS, Barnes VA, Treiber FA. Impact of stress reduction interventions on hostility and ambulatory systolic blood pressure in African American adolescents. J Black Psychol 2011;37(2):210-233 [FREE Full text] [doi: 10.1177/0095798410380203] [Medline: 22485058]
46. Li J, Qin X. Efficacy of mindfulness-based stress reduction on fear of emotions and related cognitive behavioral processes in Chinese university students: a randomized controlled trial. Psychol Sch 2021;58(10):2068-2084. [doi: 10.1002/pits.22578]

XSL•FO
RenderX

47. Wang M, Zhang H, Zhang X, Zhao Q, Chen J, Hu C, et al. Effects of a online brief modified mindfulness-based stress reduction therapy for anxiety among Chinese adults: a randomized clinical trial. J Psychiatr Res 2023;161:27-33 [FREE Full text] [doi: 10.1016/j.jpsychires.2023.03.009] [Medline: 36893668]
48. Zhang JY, Cui YX, Zhou YQ, Li YL. Effects of mindfulness-based stress reduction on prenatal stress, anxiety and depression. Psychol Health Med 2019;24(1):51-58. [doi: 10.1080/13548506.2018.1468028] [Medline: 29695175]
49. Guo D, Sun L, Yu X, Liu T, Wu L, Sun Z, et al. Mindfulness-based stress reduction improves the general health and stress of Chinese military recruits: a pilot study. Psychiatry Res 2019;281:112571. [doi: 10.1016/j.psychres.2019.112571] [Medline: 31590104]
50. Zhang JY, Zhou YQ, Feng ZW, Fan YN, Zeng GC, Wei L. Randomized controlled trial of Mindfulness-Based Stress Reduction (MBSR) on posttraumatic growth of Chinese breast cancer survivors. Psychol Health Med 2017;22(1):94-109. [doi: 10.1080/13548506.2016.1146405] [Medline: 26853191]
51. Miyamoto Y, Ma X. Dampening or savoring positive emotions: a dialectical cultural script guides emotion regulation. Emotion 2011;11(6):1346-1357. [doi: 10.1037/a0025135] [Medline: 21910543]
52. Gilbert P, McEwan K, Catarino F, Baião R, Palmeira L. Fears of happiness and compassion in relationship with depression, alexithymia, and attachment security in a depressed sample. Br J Clin Psychol 2014;53(2):228-244. [doi: 10.1111/bjc.12037] [Medline: 24283291]
53. Eckard EW. NCAA athlete graduation rates: less than meets the eye. J Sport Manag 2010;24(1):45-59. [doi: 10.1123/jsr.2022-0027]
54. Zeidan F, Johnson SK, Diamond BJ, David Z, Goolkasian P. Mindfulness meditation improves cognition: evidence of brief mental training. Conscious Cogn 2010;19(2):597-605. [doi: 10.1016/j.concog.2010.03.014] [Medline: 20363650]
55. Jha AP, Denkova E, Zanesco AP, Witkin JE, Rooks J, Rogers SL. Does mindfulness training help working memory 'work' better? Curr Opin Psychol 2019;28:273-278. [doi: 10.1016/j.copsyc.2019.02.012] [Medline: 30999122]
56. Franco C, Mañas I, Cangas AJ, Gallego J. Exploring the effects of a mindfulness program for students of secondary school. Int J Knowl Soc Res 2011;2(1):14-28 [FREE Full text] [doi: 10.4018/jksr.2011010102]

Abbreviations

FFMQ: Five Facet Mindfulness Questionnaire
MBCT: mindfulness behavioral cognitive therapy
MBI: mindfulness-based intervention
MBSR: mindfulness-based stress reduction

Edited by G Eysenbach; submitted 20.06.22; peer-reviewed by J Geller, L Guo, S Norouzi; comments to author 10.11.22; revised version received 29.12.22; accepted 31.01.24; published 08.03.24.

Please cite as:
Gao Y, Shi L, Fu N, Yang N, Weeks-Gariepy T, Mao Y
Mobile-Delivered Mindfulness Intervention on Anxiety Level Among College Athletes: Randomized Controlled Trial
J Med Internet Res 2024;26:e40406
URL: https://www.jmir.org/2024/1/e40406
doi:10.2196/40406
PMID:38457201

©Yu Gao, Lu Shi, Ning Fu, Nan Yang, Tracy Weeks-Gariepy, Yuping Mao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding Heterogeneity in Individual Responses to Digital Lifestyle Intervention Through Self-Monitoring Adherence Trajectories in Adults With Overweight or Obesity: Secondary Analysis of a 6-Month Randomized Controlled Trial

Shiyu Li[1], MSc, PhD; Yan Du[2], MPH, PhD; Hongyu Miao[3], PhD; Kumar Sharma[4], MD; Chengdong Li[3], PhD; Zenong Yin[5], PhD; Bradley Brimhall[6], MD; Jing Wang[3], MPH, PhD

[1]Department of Kinesiology, The Pennsylvania State University, University Park, PA, United States

[2]School of Nursing, The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

[3]College of Nursing, Florida State University, Tallahassee, FL, United States

[4]Center for Precision Medicine, Long School of Medicine, The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

[5]Department of Public Health, The University of Texas at San Antonio, San Antonio, TX, United States

[6]Department of Pathology and Laboratory Medicine, Long School of Medicine, The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

Corresponding Author:
Jing Wang, MPH, PhD
College of Nursing
Florida State University
Vivian M Duxbury Hall
98 Varsity Way
Tallahassee, FL, 31306
United States
Phone: 1 850 644 3299
Email: jingwang@nursing.fsu.edu

Abstract

Background: Achieving clinically significant weight loss through lifestyle interventions for obesity management is challenging for most individuals. Improving intervention effectiveness involves early identification of intervention nonresponders and providing them with timely, tailored interventions. Early and frequent self-monitoring (SM) adherence predicts later weight loss success, making it a potential indicator for identifying nonresponders in the initial phase.

Objective: This study aims to identify clinically meaningful participant subgroups based on longitudinal adherence to SM of diet, activity, and weight over 6 months as well as psychological predictors of participant subgroups from a self-determination theory (SDT) perspective.

Methods: This was a secondary data analysis of a 6-month digital lifestyle intervention for adults with overweight or obesity. The participants were instructed to perform daily SM on 3 targets: diet, activity, and weight. Data from 50 participants (mean age: 53.0, SD 12.6 y) were analyzed. Group-based multitrajectory modeling was performed to identify subgroups with distinct trajectories of SM adherence across the 3 SM targets. Differences between subgroups were examined for changes in clinical outcomes (ie, body weight, hemoglobin A_{1c}) and SDT constructs (ie, eating-related autonomous motivation and perceived competence for diet) over 6 months using linear mixed models.

Results: Two distinct SM trajectory subgroups emerged: the *Lower SM group* (21/50, 42%), characterized by *all-around low and rapidly declining SM*, and the *Higher SM group* (29/50, 58%), characterized by *moderate and declining diet and weight SM with high activity SM*. Since week 2, participants in the Lower SM group exhibited significantly lower levels of diet (P=.003), activity (P=.002), and weight SM (P=.02) compared with the Higher SM group. In terms of clinical outcomes, the Higher SM group achieved a significant reduction in body weight (estimate: −6.06, SD 0.87 kg; P<.001) and hemoglobin A_{1c} (estimate: −0.38, SD 0.11%; P=.02), whereas the Lower SM group exhibited no improvements. For SDT constructs, both groups maintained high levels of autonomous motivation for over 6 months. However, the Lower SM group experienced a significant decline in

XSL•FO
RenderX

perceived competence (*P*=.005) compared with the Higher SM group, which maintained a high level of perceived competence throughout the intervention (*P*=.09).

Conclusions: The presence of the Lower SM group highlights the value of using longitudinal SM adherence trajectories as an intervention response indicator. Future adaptive trials should identify nonresponders within the initial 2 weeks based on their SM adherence and integrate intervention strategies to enhance perceived competence in diet to benefit nonresponders.

Trial Registration: ClinicalTrials.gov NCT05071287; https://clinicaltrials.gov/study/NCT05071287

International Registered Report Identifier (IRRID): RR2-10.1016/j.cct.2022.106845

(J Med Internet Res 2024;26:e53294) doi:10.2196/53294

KEYWORDS

self-monitoring; adherence; weight loss; digital technology; behavior change; group-based trajectory modeling; precision health; mobile phone

Introduction

Background

The prevalence of obesity is increasing at an alarming rate in the United States, with approximately 42% of US adults aged ≥20 years considered obese [1]. An unhealthy lifestyle is a major risk factor for obesity in adults in the United States [2]. The United States Preventative Task Force recommends behavioral lifestyle interventions that include healthy diet and physical activity for weight management to overcome the burden of obesity [3]. However, conventional in-person lifestyle interventions have often reported reduced effectiveness for weight loss owing to poor attendance and adherence rates [4]. Therefore, it is necessary to improve the effectiveness of lifestyle interventions to address the obesity epidemic.

The surge of digital health technologies has resulted in significant progress in digital lifestyle interventions for obesity management [5]. The literature consistently reports the comparable efficacy of digital lifestyle interventions to in-person interventions in promoting short-term weight loss and improving health behaviors [6,7]. However, emerging evidence suggests that digital interventions may not be as effective as in-person interventions in maintaining weight loss and preventing weight regain [7]. It is arguable that the efficacy of current digital lifestyle interventions is hindered by a static, *one-size-fits-all* approach, in which all participants receive the same intervention without considering their personal characteristics, needs, capabilities, and preferences [8,9]. Adaptive interventions, characterized by repeated adjustment of intervention types or doses based on personal responses, represent a promising approach for optimizing digital lifestyle interventions for obesity management [10].

Developing adaptive interventions requires the early identification of participant subgroups that respond differently to the intervention. During an adaptive intervention, interventionists decide which intervention option is suited for those based on responses in intermediate tailoring variables. Intermediate tailoring variables may either lie on the causal pathway of the intervention (eg, intervention mediators) or be early predictors of long-term outcomes of interest (eg, improved clinical outcomes) [11]. Previous adaptive lifestyle interventions have used initial weight loss as an intermediate tailoring variable because of its predictive role in weight loss success, yielding mixed findings [9,11,12]. Self-monitoring (SM) adherence could be a more advantageous alternative in 2 ways. First, SM is a key component of the causal pathway of lifestyle interventions. The primary mechanism driving weight loss in lifestyle interventions is self-regulation, which involves individuals consistently monitoring their diet, physical activity, and weight. This ongoing SM enables them to reflect on and adjust their behaviors in accordance with their behavioral and weight loss goals [13,14]. Second, evidence suggests that SM is an important predictor of weight loss. This dual role of SM adherence, acting both as a predictor and causal factor for weight loss, potentially makes it a more impactful and actionable intermediate tailoring variable in adaptive lifestyle interventions for weight loss.

Investigating “how” to individualize interventions requires the identification of modifiable factors that can differentiate participants with different levels of intervention response. Self-determination theory (SDT) offers a unique theoretical perspective by focusing on the processes by which individuals initiate and maintain behaviors over time [15]. According to SDT, meaningful and lasting health behavior change requires that an individual is autonomously motivated for the change (autonomous motivation) and perceives themselves to be competent in making the change (perceived competence) [16]. Increasing autonomous motivation and perceived competence through intervention strategies has consistently been linked to successful changes in health behaviors [17-19]. However, studies have rarely examined whether lifestyle intervention participants differ in within-individual changes over time in terms of autonomous motivation and perceived competence. Notably, Webber et al [20] found that autonomous motivation levels at week 4, rather than baseline levels, predicted adherence to SM. Thus, comparing longitudinal changes in autonomous motivation and perceived competence among participants with different levels of intervention responses may provide valuable insights for developing effective adaptive interventions.

Contrary to the straightforward use of initial weight loss as an intermediate tailoring variable, employing SM adherence in this role requires a comprehensive consideration. In digital lifestyle interventions, participants are typically instructed to perform daily SM of diet, physical activity, and weight [21]. However, defining SM adherence is complicated by the presence of multiple SM targets and arbitrary selection of adherence thresholds for each [22]. This hinders the straightforward

XSL•FO
RenderX

grouping of participants as responders or nonresponders based on a single SM adherence cut-off point, an approach frequently used in previous studies where initial weight loss was the intermediate tailoring variable [12]. Moreover, empirical evidence has consistently demonstrated a decline in SM adherence over time across various targets, underscoring its dynamic nature [21]. Consequently, a key challenge is to effectively distinguish participant responsiveness based on adherence to multiple SM adherence, considering its dynamic nature. Specifically, this study aimed to simultaneously capture the between-individual and within-individual changes in multiple SM adherence over time during a lifestyle intervention.

Group-based multitrajectory modeling (GBMM) is a promising approach for detecting both between-individual and within-individual variations in SM adherence [23]. GBMM assumes that (1) the population is composed of a finite number of homogenous subpopulations, each defined by its unique developmental trajectories across various indicators [24], and (2) significant heterogeneity exists in developmental trajectories among different population subgroups [24]. Applying the GBMM to longitudinal SM adherence data may uncover participant response subgroups that may not be detected by conventional demographic or clinical characteristics. For example, Yang et al [25] applied GBMM and identified 3 distinct participant subgroups based on their adherence to SM of physical activity, weight, and blood glucose during a digital lifestyle intervention in adults with type 2 diabetes [25]. Among these subgroups, a "low and waning" category emerged, characterized by consistently lower and declining adherence to weight and blood glucose SM. Moreover, this subgroup differed significantly in multiple baseline demographic characteristics and had higher baseline A_{1c} levels compared with the other 2 subgroups. However, this study did not associate SM trajectory subgroups with changes in clinical outcomes.

Another significant advantage of GBMM is its capacity to compare longitudinal SM adherence trajectories across participant response subgroups, allowing interventionists to pinpoint "when" participants began exhibiting varying response levels. This can inform the selection of a "Decision Point" for future adaptive interventions, an essential time point for assessing individual responsiveness, and for adapting intervention strategies [26]. Selecting decision points requires careful consideration of an individual's dynamic state and the moments when meaningful changes occur in tailoring variables [27,28]. Previous research has often failed to justify predetermined decision points. For example, when using "initial weight loss" as the intermediate tailoring variable, several studies have used various time points as the "Decision Point," with durations ranging from 3 to 7 weeks [9,11,12]. However, evidence suggests that the choice of decision point might indeed impact weight loss outcomes. Taken together, comparing SM adherence trajectories across participant subgroups identified by GBMM might aid in establishing evidence-based decision points.

Objectives

Taken together, this study aimed to use GBMM to identify participant subgroups based on their longitudinal adherence patterns to SM of diet, physical activity, and weight over 6 months in a behavioral lifestyle intervention enhanced by digital SM of multiple behaviors. We also aimed to explore whether the intervention responses differed among the identified SM subgroups with respect to weight loss, glycemic control, and blood pressure. Finally, we explored whether SDT accounted for differences in SM trajectories.

Methods

Study Design and Participants

This was a secondary study based on a subset of data obtained from a pilot randomized controlled trial. Full details of the parent trial protocol have been published previously [29]. In brief, 60 participants (50 of whom were included in this secondary analysis) participated in a technology-assisted lifestyle intervention between June 2021 and October 2022 and were randomly assigned to receive either a ketogenic diet or a low-fat low-calorie diet.

The key inclusion criteria were (1) being 18 years of age or older, (2) being overweight or obese (BMI≥25 kg/m^2), (3) with or without self-reported type 2 diabetes, (4) with or without evidence of chronic kidney disease, (5) being able to read and write in English, and (6) owning a smartphone. Key exclusion criteria included (1) being pregnant or breastfeeding, (2) under sodium-glucose cotransporter-2 treatment, (3) inability to walk without assistance, (4) triglyceride ≥500 mg/dL or low-density lipoprotein ≥129 mg/dL, (5) having severe psychiatric disorders or chronic conditions that may interfere with study procedures, (6) planning to leave the city for over 2 weeks within the intervention duration, (7) enrollment in other diet or weight loss programs, and (8) being unable or willing to consent.

Participants completed assessments of clinical and psychosocial outcomes at baseline and at 3 and 6 months of the parent trial. Participants were instructed to perform daily SM of diet, physical activity, and weight using the Fitbit food log (Fitbit Inc, United States), Fitbit Inspire 2 (Fitbit Inc), and Withing Body Scale (Withings Inc). All SM devices were provided by the study team.

The participants received a weekly personalized feedback message via email. These messages were tailored to each individual, focusing on their performance in SM and how well they adhered to diet and physical activity goals. If a participant did not provide any SM data for a month, a research assistant would reach out via phone to remind them about the importance of SM and assist them with any issues. In addition, individual web sessions with interventionists were scheduled at the start of the study and then at the 3- and 5-month marks. These meetings aimed to address and resolve any challenges participants faced in adhering to the study's procedures, including SM.

In this secondary analysis, we excluded participants (n=5) who withdrew from the parental trial. This was necessary because adherence in clinical trials is defined as "the degree to which the behavior of trial participants corresponds to the intervention assigned to them" [30]. Those who withdrew from the study no longer received the intervention, including SM. We also

XSL•FO
RenderX

excluded participants who did not finish setting up the SM devices (n=5) because we were unable to assess their SM behaviors.

The parent trial was approved by the Institutional Review Board of UT Health San Antonio. All participants provided written informed consent and received financial compensation.

Measures

Adherence to SM

Adherence to SM of diet, physical activity, and weight were measured daily and operationalized as binary variables (adherence=1, nonadherence=0) based on whether a participant logged at least 800 kcal of total caloric intake (adherent: logged ≥800 kcal), at least 500 steps (adherent: logged ≥500 steps), and body weight (adherent: with a body weight value entry). In cases where participants did not provide SM data for a specific SM target on a given day, such instances were coded as *0*, reflecting nonadherence to the prescribed SM activities.

We chose an 800 kcal/d threshold for defining adherence to diet SM, as it was found to be the best predictor of weight loss among other diet SM metrics from food tracking apps [31]. In addition, this threshold was deemed to be a plausible daily caloric intake and has been commonly used in previous studies as a threshold for dietary SM [32]. We used 500 steps/d as the threshold for defining adherence to physical activity SM, as it is considered a realistic daily step count and has been commonly used as a threshold in previous studies [33].

Clinical Outcomes

All measurements were collected in a fasting state by trained staff members using standardized protocols at the UT Health San Antonio School of Nursing at baseline and at 3 and 6 months. Weight (in kg) was measured using the mean of two measurements by a professional-grade digital scale, while participants wore light clothing with bare feet on the scale's footpads. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were recorded as the average of 2 measurements using an Omron professional blood pressure machine (OMRON Group, Japan) while participants were in a relaxed sitting position. Fasting blood samples were stored at −80 °C. Hemoglobin A_{1c} (HbA_{1c}) levels were analyzed using Quest Diagnostics (Dallas, TX).

Autonomous Motivation

Autonomous motivation for diet was assessed using the *autonomous regulation* subscale from the Treatment Self-Regulation Questionnaire for Diet [34]. The questionnaire includes various statements that represent different types of motivation for following a healthy diet based on SDT. Participants expressed their level of agreement with each statement using a 7-point Likert scale, where a score of 1 indicated "not at all true" and a score of 7 signified "very true." Our analysis focused on 6 statements that measured autonomous motivation. The autonomous motivation score at each time point was determined by calculating the average score of the 6 items related to autonomous motivation. Cronbach α for autonomous motivation was 0.86 at baseline, 0.92 at 3 months, and 0.91 at 6 months, respectively.

Perceived Competence

Perceived competence in diet was assessed using the Perceived Competence Scale for Diet [35]. The scale presented participants with 4 statements assessing their belief in their ability to maintain a healthy diet. Participants rated their agreement with each statement on a 7-point Likert scale ranging from 1 (not at all true) to 7 (very true). The total perceived competence score was calculated as the average of the 4 items at each time point. Cronbach α for perceived competence for diet were 0.94 at baseline, 0.97 at 3 months, and 0.95 at 6 months, respectively.

Demographic Characteristics

Participant demographics were obtained from self-reported questionnaires and surveys at baseline. The demographic variables included age (years), sex, race and ethnicity, education, and annual household income.

Statistical Analysis

Identifying Distinct Trajectories of Adherence to SM of Diet, Activity, and Weight

Overview

We established 3 daily binary variables to measure *adherence* versus *nonadherence* to SM of diet, activity, and weight, respectively. We then calculated the number of days of adherence to each SM target at 2-week intervals over 6 months.

We used the STATA PROC TRAJ package to fit GBMM using censored regressions as the link function to model the longitudinal trajectories of adherence to SM of diet, activity, and weight. The purpose of GBMM was to (1) identify the number of SM trajectory subgroups and (2) describe how these subgroups are distributed across populations.

Model Fitting

GBMM models were estimated using a 2-step approach as described in the study by Nagin et al [23]. In step 1, we estimated models with 2 to 5 subgroups for each of the SM targets (diet, physical activity, weight) separately with trajectories modeled as linear or quadratic functions of time within each model. The objective was to identify the types of distinctive trajectories that were important for representation in the GBMM model. In this step, we observed that the trajectories generally followed a linear time function, as evidenced by the significance of all the linear terms (Multimedia Appendix 1 [23]).

In step 2, we proceeded to estimate SM for all 3 SM targets together, with a linear time function and considering 2-4 subgroups. Our findings indicate that all estimated GBMM models satisfied our predefined criteria for model selection. Notably, the models with a greater number of subgroups demonstrated improved Bayesian Information Criterion scores.

After a closer examination of SM adherence trajectories for 2-4 subgroups, we observed that the diet SM adherence trajectories in models with 3 and 4 groups did not show significant differences compared with the 2-group model. In addition, the 2-group model effectively captured the distinct trajectories for each SM target, as revealed in step 1 of the analysis. Thus, our final preferred model for multitrajectory analysis consisted of

XSL•FO
RenderX

2 distinct groups for simplicity and ease of interpretation in subsequent multivariate analyses.

Model Selection

Given that the Bayesian Information Criteria and Bayes factor increased with the increasing number of subgroups, we selected the final GBMM model based on the following criteria: (1) model value of $P \leq .05$, (2) trajectory shape similarity, (3) average posterior probability ≥0.7, (4) odds of correct classification ≥5.0, (5) minimum sample size for each subgroup ≥10%, (5) the proportion of a sample assigned to a certain group is close to the proportion estimated from the model, and (6) 99% CIs of the estimated proportion are reasonably narrow. Thus, a 2-group model with linear terms was considered to be the optimal model, and both groups had posterior probabilities of 0.97 or higher. The details of the model selection procedures are described in the Multimedia Appendix 1.

Examining Differences between SM Trajectory Subgroups on SM Adherence, Metabolic Outcomes, and Motivation

Continuous variables were first checked for normality using the Shapiro-Wilk test. To examine differences between SM trajectory subgroups on baseline characteristics, we performed an independent *t* test or Mann-Whitney *U* test for continuous variables, depending on data normality, and a chi-square test for categorical variables.

Linear mixed models (LMMs) were used to examine the differences between SM trajectory subgroups on SM adherence over time, changes in metabolic outcomes, changes in eating-related autonomous motivation, and changes in perceived competence for diet.

All LMMs included a random intercept, accounting for an individual's deviation from the ground mean. Fixed effects for the LMMs included time (2-week intervals for SM adherence; baseline, 3 months, and 6 months for clinical outcomes and SDT constructs), allocated SM trajectory subgroup (Lower SM vs Higher SM), and interaction between time and SM trajectory subgroup. Given the primary focus of the parent trial on dietary interventions, we have consistently included diet type (ketogenic diet vs low-fat low-calorie diet) as a fixed effect in all LMMs to control for the potential influences of diet type on specific outcomes.

Considering the small sample size in this study, confounders such as age, sex, Hispanic ethnicity, annual household income, and education level were included in the full model and then selected using backward stepwise variable selection to identify the parsimonious model structure with insignificant ($P > .05$) confounders gradually removed. All statistical and data visualization procedures were performed using R version 4.2.1.

Ethical Considerations

All study procedures were approved by the Institutional Review Board of The University of Texas Health Science Center at San Antonio (#20190528HU).

Results

Baseline Characteristics of Study Participants by SM Trajectory Groups

Two SM trajectory subgroups emerged: the *Lower SM group* (21/50, 42%), characterized by *all-around low and rapidly declining SM*, and the *Higher SM group* (29/50, 58%), characterized by *moderate and declining diet and weight SM with high activity SM*.

The baseline characteristics of the SM trajectory groups are presented in Table 1. Participants were middle-aged (age: 53.0 SD 2.6 y), mostly of Hispanic or Latino origin (28/50, 56%), and at a generally higher socioeconomic status and education level. We found that the 2 SM trajectory groups were significantly different at baseline; those in the Higher SM group were older compared with those in the Lower SM group. We also found that participants in the Higher SM group had higher baseline SBP and DBP levels.

XSL•FO
RenderX

Table 1. Baseline characteristics by self-monitoring (SM) trajectory groups.

	Total (n=50)	Lower SM (n=21)	Higher SM (n=29)	Between-group *P* value
Age (years), mean (SD)	53. 0 (12.6)	47.1 (13.9)	57.2 (9.8)	.004
Female, n (%)	32 (64)	16 (76.2)	16 (55.2)	.13
Hispanic or Latino, n (%)	28 (56)	13 (61.9)	15 (51.7)	.08
Type 2 diabetes, n (%)	25 (50)	8 (38.1)	17 (58.6)	.15
Chronic kidney disease, n (%)	17 (34)	7 (33.3)	10 (34.5)	.93
Education, n (%)				.86
High school	6 (12)	3 (14.3)	3 (10.3)	
Some college	19 (38)	7 (33.3)	12 (41.4)	
College degree	13 (26)	5 (23.8)	8 (27.6)	
Graduate degree	12 (24)	6 (28.6)	6 (20.7)	
Annual household income (US $), n (%)				.77
<40,000	10 (20)	4 (19.1)	6 (20.7)	
40,000-79,999	14 (28)	5 (23.8)	9 (31)	
≥80,000	26 (52)	12 (57.1)	14 (48.3)	
Treatment group, n (%)				.54
Ketogenic diet	26 (52)	12 (57.1)	14 (48.3)	
Low-fat low-calorie diet	24 (48)	9 (42.9)	15 (51.7)	
Body weight (kg), mean (SD)	102.5 (22.1)	103.2 (26.2)	102.0 (19.1)	.86
HbA_{1c}[a,b], mean (SD)	6.1 (1.1)	5.9 (1.2)	6.2 (1.1)	.47
SBP[a,c] (mm Hg), mean (SD)	132.0 (13.7)	126.0 (10.8)	136.4 (14.0)	.007
DBP[d] (mm Hg), mean (SD)	72.0 (9.2)	68.8 (9. 6)	74.3 (8.3)	.04
Autonomous motivation[a], mean (SD)	6.0 (0.8)	6.1 (0.9)	6.0 (0.8)	.75
Perceived competence, mean (SD)	5.4 (1.0)	5.7 (0.9)	5.3 (1.0)	.29

[a]Nonparametric Mann-Whitney *U* test was performed.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]SBP: systolic blood pressure.

[d]DBP: diastolic blood pressure.

Adherence to Multiple SM by SM Trajectory Groups

Figure 1 presents the adherence to multiple SM by SM trajectory group. Overall, participants in the Lower SM group (circle) subgroup exhibited a low-to-moderate level of adherence to diet and weight SM. Levels of physical activity SM were slightly higher than diet and weight SM at the beginning of the intervention but decreased rapidly after approximately 1 month of intervention. The Higher SM group (triangle) exhibited a moderate level of diet and weight SM, whereas physical activity SM remained high and stable over 6 months.

At week 2, participants in the Lower SM group started having a significantly lower SM adherence for diet (mean difference: 30.36, SD 9.56%; *P*=.003), physical activity (mean difference: 25.26, SD7.73%; *P*=.002), and weight (mean difference: 23.96, SD 10.1%; *P*=.02) compared with the Higher SM group. The adjusted LMM further revealed a significant main effect of time and SM trajectory group on diet SM (estimate: −2.68, SD 0.2, *P*<.001), activity SM (estimate: −2.24, SD 0.21, *P*<.001), and weight SM (estimate: −2.68, SD 0.2, *P*<.001). In addition, there was a significant time and SM trajectory group interaction on activity SM (estimate: −1.85, SD 0.21, *P*<.001) and weight SM (estimate: −0.62, SD 0.22, *P*=.005), suggesting that participants in the Lower SM group decreased their adherence to activity and weight SM at a faster rate.

XSL•FO
RenderX

Figure 1. Adherence to multiple self-monitoring (SM) by trajectory subgroup.

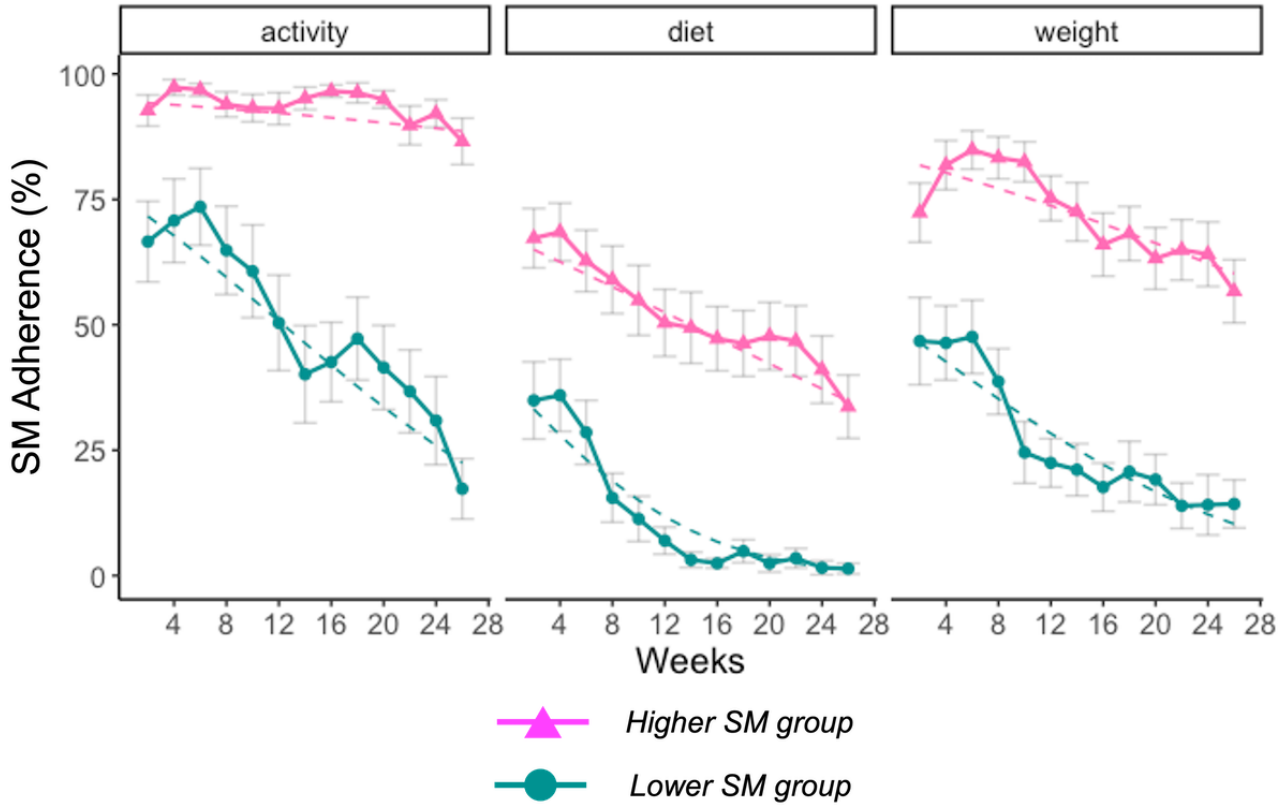

Changes in Clinical Outcomes Between SM Trajectory Groups

Table 2 shows estimated mean changes in clinical outcomes over 6 months. Adjusted LMMs revealed significant interaction between time and SM trajectory group on weight, HbA_{1c}, SBP, and DBP over 6 months. Specifically, participants in the Higher SM group significantly reduced body weight (estimate: −5.98, SE 0.87 kg, *P*<.001), HbA_{1c} (estimate: −0.46, SE 0.11%, *P*=.002), SBP (estimate: −9.71, SE 2.25 mmHg, *P*<.001), and DBP (estimate: −5.39, SE 1.49 mmHg, *P*=.008) from baseline to 3 months (Table 2, Figure 2). In addition, they maintained significant weight loss (estimate: −6.06, SE 0.87 kg, *P*<.001) and HbA_{1c} control at 6 months (estimate: −0.38, SE 0.11%; *P*=.02).

Table 2. Estimate mean changes and SE in clinical outcomes by self-monitoring (SM) trajectory group from baseline to 3 and 6 months.

	Baseline to 3 months			Baseline to 6 months		
	Lower SM group	Higher SM group	Between-group *P* value	Lower SM group	Higher SM group	Between-group *P* value
Weight (kg), mean (SE)	−2.86 (1.23)	−5.98 (0.87)[a]	.02	−3.48 (1.4)	−6.06 (0.87)[a]	.20
HbA_{1c} (%), mean (SE)	−0.04 (0.16)	−0.46 (0.11)[b]	.02	−0.03 (0.18)	−0.38 (0.11)[c]	.03
SBP[d] (mm Hg), mean (SE)	−0.01 (3.05)	−9.71 (2.26)[e]	.048	−0.42 (3.42)	−6.44 (2.26)	.23
DBP[f] (mm Hg), mean (SE)	2.15 (2.03)	−5.39 (1.49)[b]	.04	1.69 (2.29)	−2.59 (1.49)	.08

[a]Within-group: *P*<.001.
[b]Within-group: *P*<.01.
[c]Within-group: *P*<.05.
[d]SBP: systolic blood pressure.
[e]Within-group: *P*<.001.
[f]DBP: diastolic blood pressure.

XSL•FO
RenderX

Figure 2. Estimated mean changes in clinical outcomes over time by self-monitoring (SM) trajectory group. DBP: diastolic blood pressure; HbA_{1c}: hemoglobin A_{1c}; SBP: systolic blood pressure.

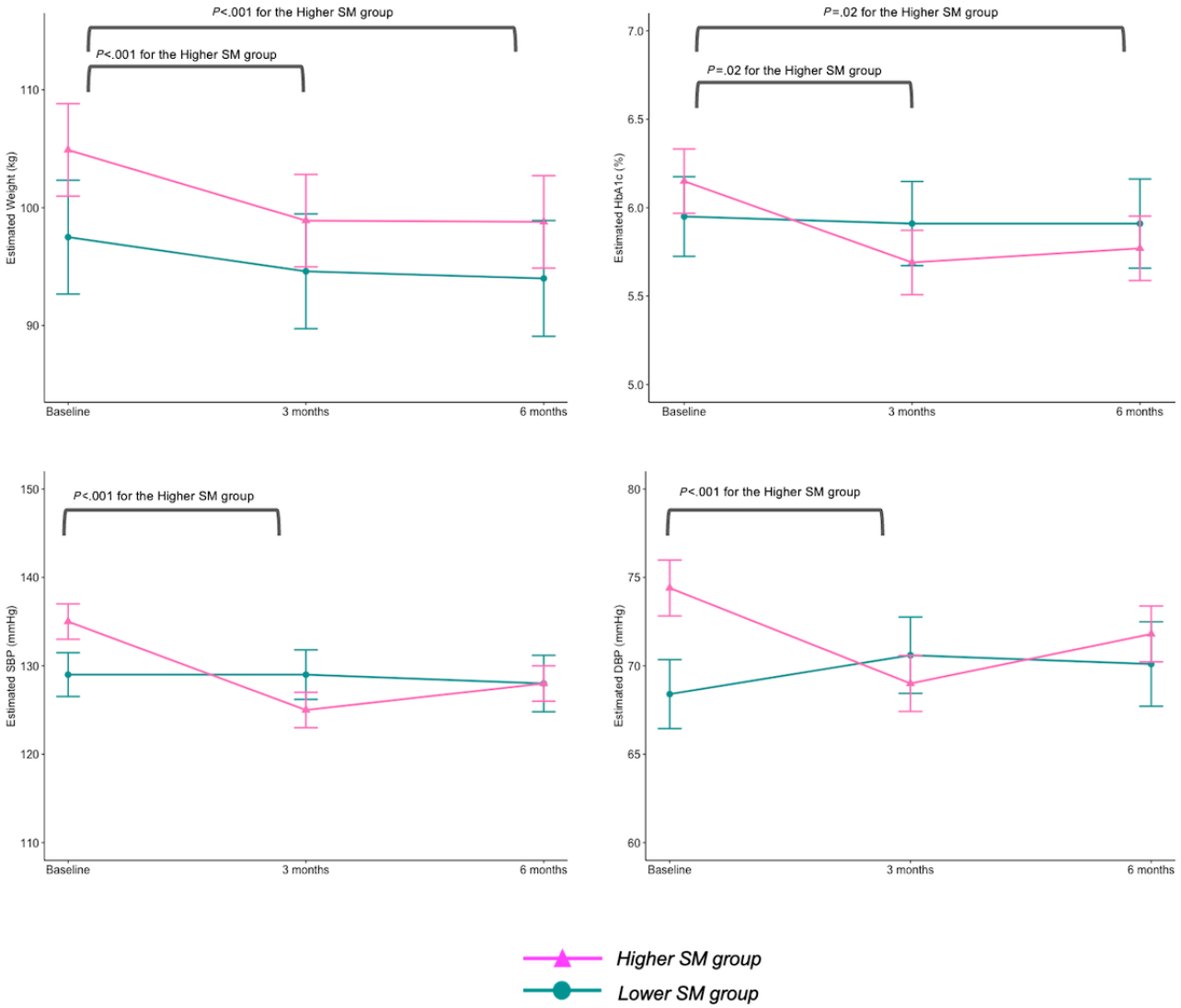

Changes in Autonomous Motivation and Perceived Competence by SM Trajectory Groups

Longitudinal changes in autonomous motivation and perceived competence in diet are shown in Figure 3. Findings from the LMM showed that both SM trajectory groups maintained a high level of autonomous motivation for over 6 months.

The LMM reported a significant interaction effect of the SM trajectory group and time (P<.001), as well as a main effect for the SM trajectory group (Lower SM group vs Higher SM group). Posthoc analyses revealed that although the responders maintained a high level of perceived competence for diet throughout the intervention period, nonresponders experienced a significant decrease in perceived competence from baseline to 3 months (estimate: –1.42, SE 0.37, P=.003) and 6 months (estimate: −1.31, SE 0.4, P=.02). Furthermore, the Higher SM group showed a significantly higher level of perceived competence in diet at 3 months than the Lower SM group (P=.02).

XSL•FO
RenderX

Figure 3. Estimated mean changes in autonomous motivation (left) and perceived competence (right) for diet over 6 months. SM: self-monitoring.

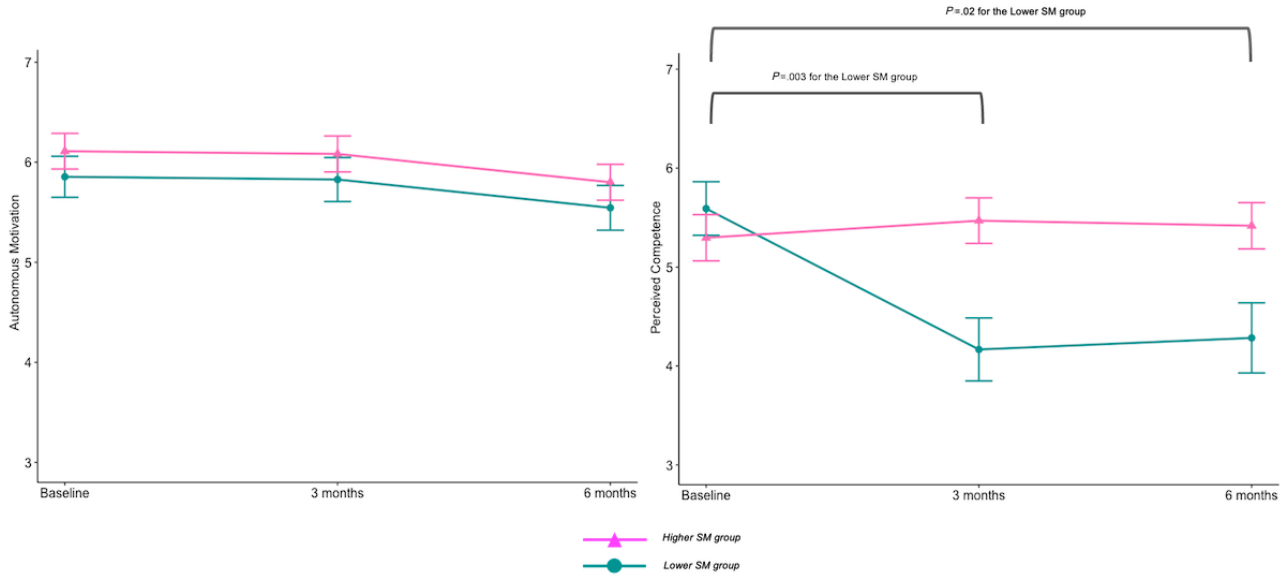

Attrition and Individual Session Attendance by SM Trajectory Groups

Table 3 presents the differences in attrition rates between subgroups at the 3- and 6-month data collection points as well as differences in attendance at individual sessions during months 3 and 5. Notably, no attrition was observed in the Higher SM group at 3 or 6 months. In addition, this group exhibited approximately 90% attendance in individual sessions.

Table 3. Attrition rate and individual session attendance by self-monitoring (SM) trajectory subgroup.

	Lower SM group, n (%)	Higher SM group, n (%)	Total, n (%)	*P* value
Attrition				
3 months	6 (28.6)	0 (0)	6 (12)	.002
6 months	9 (42.9)	0 (0)	9 (18)	<.001
Individual session attendance				
3 months	14 (66.7)	27 (93.1)	41 (82)	.02
5 months	10 (47.6)	26 (89.7)	36 (72)	<.001

Discussion

Principal Findings

In this study, we used GBMM and identified 2 participant subgroups exhibiting distinct SM trajectories during a 6-month lifestyle intervention in adults with overweight or obesity. Notably, we observed that the Lower SM group started displaying early low and rapidly declining levels of SM within the first 2 weeks. At 6 months, the Lower SM group did not show significant improvement in clinical outcomes and experienced a significant decline in perceived competence for diet compared with the Higher SM group. Our findings suggest that adherence to multiple SM may be a salient intermediate tailoring variable for identifying clinically meaningful participant subgroups. Moreover, the presence of the Lower SM group highlights the importance of early intervention, particularly within the initial 2 weeks, to positively impact clinical health outcomes. In addition, our study implies the potential of adopting early intervention strategies aimed at enhancing perceived competence in diet to benefit participants at risk of not responding well to digital lifestyle interventions.

Comparison With Prior Work

Similar to previous studies, we found that older adults were more likely to be in the Higher SM group [25,36]. Consistent evidence suggests that age is among the most prominent predictors of adherence to behavioral lifestyle interventions for weight loss [4]. Older adults are willing to learn digital health technologies. When they have an internal desire to make lifestyle changes, they are likely to engage and explore various digital health technologies [37]. Thus, tailored intervention strategies are needed to improve adherence to digital SM among younger middle-aged participants [38].

Participants in the Lower SM group consistently exhibited lower levels of multitarget SM, even during the initial 2 weeks, and did not experience significant health benefits at 6 months. Frequent SM adherence during the early intervention phase is an important predictor of later weight loss [39]. Therefore, early SM is important for successful weight loss. However, individuals' attitudes toward SM can change rapidly over a few days. As reported by Fausset et al [40] all participants initially exhibited strong motivation and a positive attitude toward SM activity, but after 2 weeks, a significant proportion experienced

XSL•FO
RenderX

a shift in attitude, perceiving SM as ineffective, and subsequently discontinued its use. In line with previous research, our study highlights the first 2 weeks of the intervention as a critical time window for interventionists to identify *nonresponsive* individuals and provide tailored strategies. To optimize interventions and promote better long-term health outcomes, future studies should investigate the feasibility and effectiveness of providing additional support during this initial phase for at-risk individuals, and compare it with conventional *one-size-fits-all* interventions.

We observed that although SM subgroups displayed differing levels of SM adherence and health improvement, they also showed similarly high levels of autonomous motivation. This pattern highlights an intention-behavior gap in health behavior change, aligning with the Health Action Process Approach theory [41]. Health Action Process Approach conceptualizes behavior change as a process with 2 distinct phases: the motivational phase, where intentions are formed, and the volitional phase, where these intentions are put into action and maintained [41]. The intention-behavior gap emerges when there is a failure to translate intentions into actions. The consistently high levels of autonomous motivation observed in both subgroups indicate that participants were intending for health behavior change. However, the disparity in actual adherence and health improvements, particularly in the Lower SM group, suggests a lack of self-regulatory skills or strategies to actualize these intentions into sustainable action [42]. This finding highlights the necessity of identifying the specific phase of behavioral change that an individual is in when designing and implementing adaptive interventions. Accordingly, it is crucial that these interventions are adapted to either enhance motivation or facilitate the translation of intentions into health behaviors.

Although both SM trajectory groups initially had similar levels of motivation and perceived competence in diet, individuals in the Lower SM group experienced a significant decline in perceived competence in diet over time. This highlights the potential role of SM in maintaining perceived competence in a healthy lifestyle and sustaining long-term health behaviors. For the Higher SM group, regular SM and ongoing assessment of progress toward behavioral and weight loss goals could assist participants in gaining mastery in making lifestyle behavioral adjustments to achieve goals, which could explain the higher level of perceived competence in this group. Perceived competence has been identified as an effective target for achieving long-term health behavior changes [19]. A previous study on adults with type 2 diabetes found that targeting perceived competence in a short period (3 days) can lead to significant improvements in physical activity and glycemic control at 3 months [43]. Therefore, intervention strategies focusing on enhancing perceived competence for a healthy diet among low-SM adherents during the early phase of the intervention may be effective for long-term health outcomes.

Participants in the Higher SM group achieved clinically significant weight loss (≥5%) at 6 months, despite only monitoring their diet and weight at a moderate level within the initial 2 weeks of the intervention. This aligns with findings from Xu et al [36] and others that a moderate level of SM diet is associated with clinically significant weight loss. Expanding on their research, our study further demonstrated that a moderate level of both diet and weight SM during the early phase of intervention could be associated with significant health improvements. In contrast, we observed that participants in the Higher SM group not only demonstrated better adherence to SM but also higher engagement in study procedures other than SM and had no attrition. Given that attendance was also a significant predictor of weight loss [44], it becomes clear that future research should focus on disentangling the complex interplay between SM, intervention engagement, and health outcomes.

Previous studies have often associated reduced SM adherence with a failure of behavioral self-regulation and have focused on developing strategies to enhance SM adherence [29,45]. However, our study provides a different perspective, revealing that participants in the higher SM group achieved continued weight loss and maintained a high level of autonomous motivation and perceived competence in their diet, despite experiencing a decline in SM adherence between 3 and 6 months. This indicates the possibility of habit formation for healthy lifestyles and a voluntary *happy abandonment* of SM within the higher SM group. Similarly, Attig and Franke [46] reported that individuals might stop using activity SM devices once they have formed habitual exercise behaviors [46]. To optimize interventions, researchers must recognize that the role of SM lies in supporting behavioral self-regulation and sustaining motivation for a healthy lifestyle. If participants have established long-lasting motivation and formed habits for a healthy lifestyle, additional SM strategies may be unnecessary and wasteful of resources. Taken together, to optimize the SM component in lifestyle interventions, future studies need to use intensive longitudinal data collection and analytic strategies (ie, ecological momentary assessments) to unfold the dynamics between motivation for SM, motivation for a healthy lifestyle, and real-time lifestyle behavior change [47].

Limitations

This study had several limitations. First, we focused on SM of diet, activity, and weight, as they were predominantly recommended for obesity management; however, adults living with overweight or obesity with comorbidities might also perform other types of SM (ie, blood glucose, blood pressure, etc), which could impact their changes in metabolic health during the intervention.

Second, although we defined SM adherence based on the current literature, the definitions are still arbitrary. Future studies are needed to examine the impact of various SM adherence criteria on the identification of SM trajectory subgroups to ensure replicability of our findings.

Third, the GBMM approach assigned participants to subgroups based on their closest SM trajectory for each target, which could obscure within-group variability in their SM trajectories for each target.

Fourth, owing to the small sample size, we had limited statistical power to identify additional SM trajectory subgroups.

XSL•FO
RenderX

Fifth, we assigned 0s to indicate nonadherence on days when participants did not provide SM data for a specific SM target. However, this approach could not differentiate between participants who completely withdrew from all the study procedures and those who merely neglected SM while adhering to other aspects of the study. These participants could appear indistinguishable when solely analyzing the SM adherence data. It is important to note that different strategies may be required for tailoring interventions. Therefore, future research should examine SM trajectory patterns more closely to explore methods that can effectively differentiate these participants.

Sixth, we chose 2 group model as the final GBMM model to balance the sample size of each subgroup and increase the power of the comparison between participant subgroups. However, there could have been unexplored SM trajectories in this study. Thus, future studies with larger sample sizes are warranted.

Seventh, the participants in our study were mostly from higher socioeconomic backgrounds with higher levels of education, which could limit the generalizability of our findings to a broader population. Furthermore, we excluded participants who did not finish setting up SM technologies or who withdrew from the parent trial. However, these participants might represent a distinct subgroup that requires different intervention strategies to improve their lifestyle and metabolic health.

Finally, the relatively short duration of our intervention limits our ability to thoroughly examine the long-term effects of different SM trajectory groups on sustaining health improvements.

Conclusions

By modeling the longitudinal SM adherence trajectories using the GBMM, we identified distinct participant subgroups exhibiting varied responses to a digital lifestyle intervention in terms of SM adherence, clinical outcomes, and motivation factors. Our findings support the use of SM adherence as a salient proximal outcome for the early identification of participants at risk for *nonresponding*. Moreover, our research highlights the potential of effective intervention strategies aimed at improving perceived competence in the diet. The findings of our study hold the promise of informing the development of future adaptive lifestyle interventions for obesity management.

Acknowledgments

The parent trial was funded by the 80/20 Foundation (KS). JW was supported by the University of Florida–Florida State University Clinical and Translational Sciences Award project, which is supported in part by the National Center for Advancing Translational Sciences under the award number UL1TR001427.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Group-based multitrajectory model procedures.
[DOCX File , 730 KB - jmir_v26i1e53294_app1.docx]

References

1. Stierman B, Afful J, Carroll MD, Chen TC, Davy O, Fink S, et al. National Health and Nutrition Examination Survey 2017–March 2020 prepandemic data files development of files and prevalence estimates for selected health outcomes. Centers for Disease Control and Prevention. 2021. URL: https://stacks.cdc.gov/view/cdc/106273 [accessed 2024-01-29]
2. Hruby A, Manson JE, Qi L, Malik VS, Rimm EB, Sun Q, et al. Determinants and consequences of obesity. Am J Public Health 2016 Sep;106(9):1656-1662. [doi: 10.2105/AJPH.2016.303326] [Medline: 27459460]
3. US Preventive Services Task Force, Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, et al. Behavioral weight loss interventions to prevent obesity-related morbidity and mortality in adults: US preventive services task force recommendation statement. JAMA 2018 Sep 18;320(11):1163-1171. [doi: 10.1001/jama.2018.13022] [Medline: 30326502]
4. Burgess E, Hassmén P, Pumpa KL. Determinants of adherence to lifestyle intervention in adults with obesity: a systematic review. Clin Obes 2017 Jun;7(3):123-135. [doi: 10.1111/cob.12183] [Medline: 28296261]
5. Bhardwaj NN, Wodajo B, Gochipathala K, Paul DP3, Coustasse A. Can mHealth revolutionize the way we manage adult obesity? Perspect Health Inf Manag 2017;14(Spring):1a [FREE Full text] [Medline: 28566984]
6. Mamalaki E, Poulimeneas D, Tsiampalis T, Kouvari M, Karipidou M, Bathrellou E, et al. The effectiveness of technology-based interventions for weight loss maintenance: a systematic review of randomized controlled trials with meta-analysis. Obes Rev 2022 Sep;23(9):e13483. [doi: 10.1111/obr.13483] [Medline: 35686875]
7. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MD, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019 Jan 08;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
8. König LM, Attig C, Franke T, Renner B. Barriers to and facilitators for using nutrition apps: systematic review and conceptual framework. JMIR Mhealth Uhealth 2021 Apr 01;9(6):e20037 [FREE Full text] [doi: 10.2196/20037] [Medline: 34254938]

XSL•FO
RenderX

9. Miller CK. Adaptive intervention designs to promote behavioral change in adults: what is the evidence? Curr Diab Rep 2019 Jan 25;19(2):7. [doi: 10.1007/s11892-019-1127-4] [Medline: 30684109]
10. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol 2012;8:21-48 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032511-143152] [Medline: 22224838]
11. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. Transl Behav Med 2014 Sep;4(3):260-274 [FREE Full text] [doi: 10.1007/s13142-014-0265-0] [Medline: 25264466]
12. Sherwood NE, Butryn ML, Forman EM, Almirall D, Seburg EM, Lauren Crain A, et al. The BestFIT trial: a SMART approach to developing individualized weight loss treatments. Contemp Clin Trials 2016 Mar;47:209-216 [FREE Full text] [doi: 10.1016/j.cct.2016.01.011] [Medline: 26825020]
13. Kanfer FH. Self-monitoring: Methodological limitations and clinical applications. J Consult Clin Psychol 1970 Oct;35(2):148-152. [doi: 10.1037/h0029874]
14. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011 Jan;111(1):92-102 [FREE Full text] [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
15. Ryan RM, Patrick H, Deci EL, Williams GC. Facilitating health behaviour change and its maintenance: interventions based on self-determination theory. Eur J Health Psychol 2008;10(1):2-5.
16. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY: The Guilford Press; 2017.
17. Williams GC, McGregor HA, Sharp D, Levesque C, Kouides RW, Ryan RM, et al. Testing a self-determination theory intervention for motivating tobacco cessation: supporting autonomy and competence in a clinical trial. Health Psychol 2006 Jan;25(1):91-101. [doi: 10.1037/0278-6133.25.1.91] [Medline: 16448302]
18. Sheeran P, Wright CE, Avishai A, Villegas ME, Lindemans JW, Klein WM, et al. Self-determination theory interventions for health behavior change: meta-analysis and meta-analytic structural equation modeling of randomized controlled trials. J Consult Clin Psychol 2020 Aug;88(8):726-737. [doi: 10.1037/ccp0000501] [Medline: 32437175]
19. Sheeran P, Wright CE, Avishai A, Villegas ME, Rothman AJ, Klein WM. Does increasing autonomous motivation or perceived competence lead to health behavior change? A meta-analysis. Health Psychol 2021 Oct;40(10):706-716. [doi: 10.1037/hea0001111] [Medline: 34881939]
20. Webber KH, Tate DF, Ward DS, Bowling JM. Motivation and its relationship to adherence to self-monitoring and weight loss in a 16-week Internet behavioral weight loss intervention. J Nutr Educ Behav 2010 May;42(3):161-167. [doi: 10.1016/j.jneb.2009.03.001] [Medline: 20138583]
21. Patel ML, Wakayama LN, Bennett GG. Self-monitoring via digital health in weight loss interventions: a systematic review among adults with overweight or obesity. Obesity (Silver Spring) 2021 Mar 23;29(3):478-499. [doi: 10.1002/oby.23088] [Medline: 33624440]
22. Payne JE, Turk MT, Kalarchian MA, Pellegrini CA. Defining adherence to dietary self-monitoring using a mobile app: a narrative review. J Acad Nutr Diet 2018 Nov;118(11):2094-2119. [doi: 10.1016/j.jand.2018.05.011] [Medline: 30115555]
23. Nagin DS, Jones BL, Passos VL, Tremblay RE. Group-based multi-trajectory modeling. Stat Methods Med Res 2018 Jul 17;27(7):2015-2023. [doi: 10.1177/0962280216673085] [Medline: 29846144]
24. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol 2010;6:109-138. [doi: 10.1146/annurev.clinpsy.121208.131413] [Medline: 20192788]
25. Yang Q, Hatch D, Crowley MJ, Lewinski AA, Vaughn J, Steinberg D, et al. Digital phenotyping self-monitoring behaviors for individuals with type 2 diabetes mellitus: observational study using latent class growth analysis. JMIR Mhealth Uhealth 2020 Jun 11;8(6):e17730 [FREE Full text] [doi: 10.2196/17730] [Medline: 32525492]
26. Hall KL, Nahum-Shani I, August GJ, Patrick ME, Murphy SA, Almirall D. Adaptive intervention designs in substance use prevention. In: Sloboda Z, Petras H, Robertson E, Hingson R, editors. Prevention of Substance Use. Cham, Switzerland: Springer; 2019:263-280.
27. Golbus JR, Dempsey W, Jackson EA, Nallamothu BK, Klasnja P. Microrandomized trial design for evaluating just-in-time adaptive interventions through mobile health technologies for cardiovascular disease. Circ Cardiovasc Qual Outcomes 2021 Feb;14(2):e006760. [doi: 10.1161/circoutcomes.120.006760]
28. Collins LM, Kugler KC. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST). Cham, Switzerland: Springer; 2018.
29. Du Y, Wang J, Li S, Dennis B, Meireles C, Siddiqui N, et al. A technology assisted precision ketogenic diet intervention for cardio-renal-metabolic health in overweight or obese adults: protocol for a randomized controlled trial. Contemp Clin Trials 2022 Aug;119:106845. [doi: 10.1016/j.cct.2022.106845] [Medline: 35809772]
30. Adherence to long-term therapies: evidence for action. World Health Organization. 2003. URL: https://www.paho.org/en/documents/who-adherence-long-term-therapies-evidence-action-2003 [accessed 2024-01-29]
31. Turner-McGrievy GM, Dunn CG, Wilcox S, Boutté AK, Hutto B, Hoover A, et al. Defining adherence to mobile dietary self-monitoring and assessing tracking over time: tracking at least two eating occasions per day is best marker of adherence within two different mobile health randomized weight loss interventions. J Acad Nutr Diet 2019 Sep;119(9):1516-1524 [FREE Full text] [doi: 10.1016/j.jand.2019.03.012] [Medline: 31155473]

XSL•FO
RenderX

32. Schumacher LM, Martinelli MK, Convertino AD, Forman EM, Butryn ML. Weight-related information avoidance prospectively predicts poorer self-monitoring and engagement in a behavioral weight loss intervention. Ann Behav Med 2021 Mar 16;55(2):103-111 [FREE Full text] [doi: 10.1093/abm/kaaa034] [Medline: 32491152]
33. Thomson JL, Landry AS, Zoellner JM, Tudor-Locke C, Webster M, Connell C, et al. Several steps/day indicators predict changes in anthropometric outcomes: HUB City Steps. BMC Public Health 2012 Nov 15;12(1):983 [FREE Full text] [doi: 10.1186/1471-2458-12-983] [Medline: 23153060]
34. Levesque CS, Williams GC, Elliot D, Pickering MA, Bodenhamer B, Finley PJ. Validating the theoretical structure of the Treatment Self-Regulation Questionnaire (TSRQ) across three different health behaviors. Health Educ Res 2007 Oct 23;22(5):691-702. [doi: 10.1093/her/cyl148] [Medline: 17138613]
35. Williams GC, Niemiec CP, Patrick H, Ryan RM, Deci EL. The importance of supporting autonomy and perceived competence in facilitating long-term tobacco abstinence. Ann Behav Med 2009 Jun;37(3):315-324 [FREE Full text] [doi: 10.1007/s12160-009-9090-y] [Medline: 19373517]
36. Xu R, Bannor R, Cardel MI, Foster GD, Pagoto S. How much food tracking during a digital weight-management program is enough to produce clinically significant weight loss? Obesity (Silver Spring) 2023 Jul 04;31(7):1779-1786. [doi: 10.1002/oby.23795] [Medline: 37271576]
37. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
38. Venditti EM, Wylie-Rosett J, Delahanty LM, Mele L, Hoskin MA, Edelstein SL. Short and long-term lifestyle coaching approaches used to address diverse participant barriers to weight loss and physical activity adherence. Int J Behav Nutr Phys Act 2014 Feb 12;11:16 [FREE Full text] [doi: 10.1186/1479-5868-11-16] [Medline: 24521153]
39. Carpenter CA, Eastman A, Ross KM. Consistency with and disengagement from self-monitoring of weight, dietary intake, and physical activity in a technology-based weight loss program: exploratory study. JMIR Form Res 2022 Feb 18;6(2):e33603 [FREE Full text] [doi: 10.2196/33603] [Medline: 35179513]
40. Fausset CB, Mitzner TL, Price CE, Jones BD, Fain BW, Rogers WA. Older adults' use of and attitudes toward activity monitoring technologies. Proc Hum Factors Ergon Soc Annu Meet 2013 Sep 30;57(1):1683-1687 [FREE Full text] [doi: 10.1177/1541931213571374] [Medline: 31263349]
41. Schwarzer R. Health Action Process Approach (HAPA) as a theoretical framework to understand behavior change. Act Psic 2016 Dec 05;30(121):119-130. [doi: 10.15517/ap.v30i121.23458]
42. Sniehotta FF, Scholz U, Schwarzer R. Bridging the intention–behaviour gap: planning, self-efficacy, and action control in the adoption and maintenance of physical exercise. Psychol Health 2005 Apr;20(2):143-160. [doi: 10.1080/08870440512331317670]
43. Trouilloud D, Regnier J. Therapeutic education among adults with type 2 diabetes: effects of a three-day intervention on perceived competence, self-management behaviours and glycaemic control. Glob Health Promot 2013 Jun 15;20(2 Suppl):94-98. [doi: 10.1177/1757975913483331] [Medline: 23678503]
44. Byrne S, Barry D, Petry NM. Predictors of weight loss success. Exercise vs. dietary self-efficacy and treatment attendance. Appetite 2012 Apr;58(2):695-698 [FREE Full text] [doi: 10.1016/j.appet.2012.01.005] [Medline: 22248709]
45. Krukowski RA, Harvey J, Borden J, Stansbury ML, West DS. Expert opinions on reducing dietary self-monitoring burden and maintaining efficacy in weight loss programs: a Delphi study. Obes Sci Pract 2022 Aug;8(4):401-410 [FREE Full text] [doi: 10.1002/osp4.586] [Medline: 35949285]
46. Attig C, Franke T. Abandonment of personal quantification: a review and empirical study investigating reasons for wearable activity tracking attrition. Comput Human Behav 2020 Jan;102:223-237. [doi: 10.1016/j.chb.2019.08.025]
47. Hufford MR, Shiffman S, Paty J, Stone AA. Ecological Momentary Assessment: real-world, real-time measurement of patient experience. In: Fahrenberg J, Myrtek M, editors. Progress in Ambulatory Assessment: Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies. Boston, MA: Hogrefe & Huber; 2001:69-92.

Abbreviations

DBP: diastolic blood pressure
GBMM: group-based multitrajectory modeling
HbA_{1c}: hemoglobin A_{1c}
LMM: linear mixed model
SBP: systolic blood pressure
SDT: self-determination theory
SM: self-monitoring

XSL•FO
RenderX

Edited by G Eysenbach; submitted 02.10.23; peer-reviewed by L Schumacher; comments to author 31.10.23; revised version received 01.01.24; accepted 31.01.24; published 20.03.24.

Please cite as:
Li S, Du Y, Miao H, Sharma K, Li C, Yin Z, Brimhall B, Wang J
Understanding Heterogeneity in Individual Responses to Digital Lifestyle Intervention Through Self-Monitoring Adherence Trajectories in Adults With Overweight or Obesity: Secondary Analysis of a 6-Month Randomized Controlled Trial
J Med Internet Res 2024;26:e53294
URL: https://www.jmir.org/2024/1/e53294
doi:10.2196/53294
PMID:38506903

©Shiyu Li, Yan Du, Hongyu Miao, Kumar Sharma, Chengdong Li, Zenong Yin, Bradley Brimhall, Jing Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation

Hyo-Jeong Ahn[1], MD; Eue-Keun Choi[1,2], MD, PhD; So-Ryoung Lee[1,2], MD, PhD; Soonil Kwon[1], MD; Hee-Seok Song[3], MSc; Young-Shin Lee[3], MSc; Seil Oh[1,2], MD, PhD

[1]Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[2]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[3]Seers Technology Co, Ltd, Seongnam-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Eue-Keun Choi, MD, PhD
Department of Internal Medicine
Seoul National University Hospital
101 Daehak-ro
Seoul, 03080
Republic of Korea
Phone: 82 220720688
Email: choiek417@gmail.com

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e59984

Abstract

Background: Wearable electrocardiogram (ECG) monitoring devices are used worldwide. However, data on the diagnostic yield of an adhesive single-lead ECG patch (SEP) to detect premature ventricular complex (PVC) and the optimal duration of wearing an SEP for PVC burden assessment are limited.

Objective: We aimed to validate the diagnostic yield of an SEP (mobiCARE MC-100, Seers Technology) for PVC detection and evaluate the PVC burden variation recorded by the SEP over a 3-day monitoring period.

Methods: This is a prospective study of patients with documented PVC on a 12-lead ECG. Patients underwent simultaneous ECG monitoring with the 24-hour Holter monitor and SEP on the first day. On the subsequent second and third days, ECG monitoring was continued using only SEP, and a 3-day extended monitoring was completed. The diagnostic yield of SEP for PVC detection was evaluated by comparison with the results obtained on the first day of Holter monitoring. The PVC burden monitored by SEP for 3 days was used to assess the daily and 6-hour PVC burden variations. The number of patients additionally identified to reach PVC thresholds of 10%, 15%, and 20% during the 3-day extended monitoring by SEP and the clinical factors associated with the higher PVC burden variations were explored.

Results: The recruited data of 134 monitored patients (mean age, 54.6 years; males, 45/134, 33.6%) were analyzed. The median daily PVC burden of these patients was 2.4% (IQR 0.2%-10.9%), as measured by the Holter monitor, and 3.3% (IQR 0.3%-11.7%), as measured in the 3-day monitoring by SEP. The daily PVC burden detected on the first day of SEP was in agreement with that of the Holter monitor: the mean difference was –0.07%, with 95% limits of agreement of –1.44% to 1.30%. A higher PVC burden on the first day was correlated with a higher daily (R^2=0.34) and 6-hour burden variation (R^2=0.48). Three-day monitoring by SEP identified 29% (12/42), 18% (10/56), and 7% (4/60) more patients reaching 10%, 15%, and 20% of daily PVC burden, respectively. Younger age was additionally associated with the identification of clinically significant PVC burden during the extended monitoring period (P=.02).

Conclusions: We found that the mobiCARE MC-100 SEP accurately detects PVC with comparable diagnostic yield to the 24-hour Holter monitor. Performing 3-day PVC monitoring with SEP, especially among younger patients, may offer a pragmatic alternative for identifying more individuals exceeding the clinically significant PVC burden threshold.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e46098) doi:10.2196/46098

KEYWORDS

premature ventricular complex; single-lead electrocardiogram; wearable device; extended monitor; electrocardiogram; EKG; ECG; wearable; heart; cardiology; cardiovascular

Introduction

Premature ventricular complexes (PVCs) are commonly observed in individuals who have undergone long-term ambulatory monitoring [1]. Frequent PVCs can result in reversible cardiomyopathy, although patients with PVCs without underlying structural heart disease are usually expected to have a benign clinical prognosis [2-4].

Several risk factors are known to contribute to the overall risk of developing PVC-induced cardiomyopathy (PIC), and the burden of PVC is one of the most important predictors of clinical deterioration [1,5,6]. Although there is no absolute cutoff of PVC burden to identify patients at risk of developing PIC, 16%-26% of PVC burden is reported as a significant threshold to develop PIC [5,7,8]. Of note, at least 10% of the PVC burden is considered sufficient for developing PIC to warrant a regular assessment of structural and functional cardiac change [9].

For the determination of an appropriate treatment strategy for PVC, one of the essential pieces of information that should be accurately assessed is the PVC burden, together with the accompanying symptoms and the presence or absence of structural heart disease. However, defining the exact PVC burden remains a challenge because of its substantial hourly or daily variation, and recent evidence recommends that patients undergo ambulatory ECG monitoring throughout the day to observe its maximum burden [1]. The 24-hour Holter monitoring has been regarded as the gold standard for evaluating PVC burden, but the Holter monitor is uncomfortable to wear, and the 24-hour duration is insufficient to appreciate the substantial burden variation demonstrated among individuals [10,11].

With the advent of widespread use of wearable electrocardiogram (ECG) patch monitoring, studies have suggested that extended monitoring is required to estimate more accurate daily PVC burden [10,12]. Recently, an adhesive single-lead ECG patch (SEP) has been found to enable more convenient extended monitoring to capture fluctuations in the PVC burden, but the diagnostic yield of SEP has not been thoroughly validated. Moreover, the optimal duration of ambulatory ECG monitoring for PVC evaluation has not been determined, and the clinical significance of 3-day SEP monitoring (a shorter period than studied before) [13] is yet to be explored.

In this study, we investigated the (1) diagnostic yield of SEP for PVC detection through a direct comparison with the 24-hour Holter monitoring and (2) variations in the PVC burden recorded during the 3-day extended monitoring using SEP and the associated clinical factors. Ultimately, we aimed to explore the potential clinical utility of SEP as a more convenient diagnostic tool of PVC, with the goal of identifying more patients who have clinically significant PVC burden.

Methods

Ethics Approval

The institutional review board at Seoul National University Hospital authorized this study (H-2103-010-1201). All study participants provided written informed consent. This study is reported according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Multimedia Appendix 1).

Study Design and Patient Selection

This is a single-center and prospective cohort study of patients with PVC, which was performed between May 2021 and June 2022. Patients (≥19 years old) with documented PVC on a 12-lead ECG and scheduled for evaluation of the PVC burden (percentage of number of PVCs over the total QRS complexes) by 24-hour Holter monitoring were included. Since SEP utilizes a smartphone app to transmit ECG data, those unable to operate and manipulate the app were excluded. Three electrophysiologists (Choi EK, Lee SR, and SO) screened and recruited eligible patients at the outpatient clinic. Baseline clinical features, including demographic data, symptoms regarding PVC, comorbidities, lifestyle behaviors, medication history, and echocardiographic parameters were investigated.

Validation of PVC Diagnosis by SEP Through Comparison With the Holter Monitor

The 24-hour Holter monitor (SEER Light, GE HealthCare) and SEP (mobiCARE MC-100, Seers Technology) record the electrical activity of the heart through 3 channels (leads I, V1, and V6) and a single channel (lead II), respectively. The process of validating the ECG algorithm for PVC detection by SEP and its performance are detailed in Multimedia Appendix 2. Patients underwent simultaneous ECG monitoring with the Holter monitor and SEP on the first day of monitoring. After 24 hours of monitoring, the Holter device was detached, and ECG monitoring was continued for the following 48 hours with SEP; the assessment of PVC burden using the Holter monitor is a part of the guideline-adherent practice [14], with patients covering the associated costs through local health insurance coverage. The detailed specifications of the SEP and the attachment methods for both devices are described in our prior report [15]. Information on total wear time, proportion of signal noise, total QRS complexes, and PVC burden was collected. An example strip of PVC detection using SEP is shown in Multimedia Appendix 3. For the evaluation of the diagnostic yield of PVC by SEP, the burdens of PVC assessed by the Holter monitor and SEP on the first day were compared with each other.

XSL•FO
RenderX

Variability of PVC Burden and Clinical Factors Associated With Greater Variability

The fluctuation of PVC burden was assessed by the changes in the daily and 6-hour PVC burden monitored over a 3-day period using SEP. The purpose of evaluating daily and 6-hour PVC burden variations was to verify the fluctuations in the values. This analysis enables us to address the challenge of identifying the accurate overall PVC burden, and consequently, determining the appropriate treatment strategy. We defined daily PVC burden variation as the difference between the maximum and the minimum daily PVC burden measured by the SEP. The 6-hour PVC burden was calculated as the burden of PVC from midnight to 6 AM, 6 AM to noon, noon to 6 PM, and 6 PM to midnight. Similarly, the 6-hour PVC burden variation was defined as the difference between the maximum and minimum 6-hour PVC burden. We estimated the proportion of patients who were additionally identified to reach a clinically significant daily PVC burden (10%, 15%, or 20%) [12] during the 3-day monitoring. Further, clinical factors associated with a greater variation in PVC burden and those additionally influencing the PVC burden to cross the clinically meaningful threshold during the extended monitoring were investigated.

Self-Reported Questionnaire About the Experience of the 3-Day Extended ECG Monitoring

The participants' clinical experience of the 3-day extended ECG monitoring for PVC detection by SEP was evaluated using a self-reported questionnaire. The survey was conducted on a voluntary basis for all the study participants without offering any form of compensation on the day the SEP device was returned. Questions were based on the overall convenience of the Holter monitor and SEP, any instances of unexpected reattachment, and accessibility and user friendliness of mobile apps for SEP. Participants responded to the questions by using a scale ranging from 1 to 5 (1=no, 2=minimally, 3=sometimes, 4=likely, or 5=very likely).

Statistical Analysis

Continuous variables were reported as mean (SD) or median (IQR) and compared by Student 2-sided *t* test or Mann-Whitney test. Categorical variables were presented as n (%), and Pearson chi-square test or Fisher exact test were applied as required. The diagnostic yield of PVC by SEP was compared to that by the Holter monitor with Bland-Altman plots with limits of agreement and scatter plots. The degree of agreement for the burden of PVC monitored by each device was analyzed. The associations between individuals' baseline characteristics and the additional identification of significant PVC thresholds (10%, 15%, or 20%) during the extended monitoring were evaluated by multivariable logistic regression analysis. All analyses were performed using Stata (version 17, StataCorp LLC).

Results

Baseline Characteristics of the Patients

Of the data of 156 enrolled patients with documented PVC on a 12-lead ECG, the data of 13 patients could not be retrieved due to errors in their smartphones, unfamiliarity with app use, or missing telemetry data transmission of SEP. Nine patients withdrew from the examination due to discomfort and skin irritability: 5 patients due to frequent alarms and the requirement of keeping their smartphones in close proximity, 2 due to skin irritability, and 2 due to inexperienced operation of the app. Finally, data from 134 patients with enough wear time of the patch (≥4200 minutes) and whose data were transmitted completely were included in this analysis.

The baseline characteristics of the included patients are described in Table 1. The mean age was 54.6 (SD 13.3) years, and 45 (33.6%) patients were males. The majority of the patients reported symptoms such as palpitation (96/134, 71.6%) or dizziness (15/134, 11.2%). Hypertension (46/134, 34.3%) and diabetes mellitus (21/134, 15.7%) were the common comorbidities. Congestive heart failure and cardiomyopathy were reported in 5.9% (8/134) and 2.9% (4/134) of the patients, respectively. Patients frequently took β-blockers (97/134, 72.4%) or calcium channel blockers (10/134, 7.5%) to relieve symptoms and PVC burden.

XSL•FO
RenderX

Table 1. Baseline characteristics of the patients with premature ventricular complex (N=134).

	Values
Age (years), mean (SD)	54.6 (13.3)
Sex (male), n (%)	45 (33.6)
BMI (kg/m^2), mean (SD)	24.3 (3.4)
Symptoms, n (%)	
Palpitation	96 (71.6)
Syncope	3 (2.2)
Dizziness	15 (11.2)
Comorbidities, n (%)	
Hypertension	46 (34.3)
Diabetes mellitus	21 (15.7)
Congestive heart failure	8 (5.9)
Cardiomyopathy	4 (2.9)
Coronary artery disease	0 (0)
Hypo/hyperthyroidism	5 (3.7)
Stroke/transient ischemic attack/thromboembolism	3 (2.2)
Lifestyle behaviors, n (%)	
Current smoking	8 (5.9)
Alcohol intake	16 (11.9)
Medications, n (%)	
β-blocker	97 (72.4)
Calcium channel blocker	10 (7.5)
Amiodarone	3 (2.2)
Class1c antiarrhythmic drug	3 (2.2)
Renin-angiotensin-aldosterone system blockade	20 (14.9)
Diuretics	5 (3.7)
Statin	29 (21.6)
Antiplatelet	15 (11.2)
Ejection fraction (%), mean (SD)	58.8 (6)

Diagnostic Performance of PVC Detection and Burden by SEP Compared to That by the Holter Monitor

An example of PVC detection by the adhesive SEP is presented in Multimedia Appendix 3. Detailed parameters of monitoring and the overall burden of PVC were compared between the 2 monitoring methods (Holter monitor and SEP) and are summarized in Table 2. The total mean wear time was 1395.1 (SD 33.5) minutes for the Holter monitor and 4318.0 (SD 12.4) minutes for SEP. The transmitted data were retrieved without significant noise for both devices (0.1% for the Holter monitor vs 0.9% for SEP). The median PVC burden of all the enrolled patients was 2.4% (IQR 0.2%-10.9%), as measured by the Holter monitor, and 3.3% (IQR 0.3%-11.7%), as measured by SEP during the 3-day extended monitoring, without significant difference (*P*=.58). During the first 24-hour monitoring by SEP, the burden of PVC among all patients was 7.5% (SD 10%; median 2.8%, IQR 0.3%-11%) without a significant difference from the PVC burden measured by the Holter monitor (*P*=.24). For those with PVC burden ≥5%, the overall PVC burden was detected as 13.3% (IQR 8%-24%) by the Holter monitor and 13.2% (IQR 9.7%-20.5%) by SEP during the 3-day extended monitoring (*P*=.10). During the first 24-hour monitoring by SEP, the burden of PVC among patients with a burden ≥5% was 16.5% (SD 9.9%; median 13.7%, IQR 8.1%-24.4%); the *P* value compared to the PVC burden detected by the Holter monitor was .82 (Table 2).

The overall distribution of the patients with PVC burden is presented in Figure 1A. Most of the patients (97/134, 72.4%) had a PVC burden <10%. PVC burden was reported as 10%-19.9% and ≥20% in 18 (13.4%) and 19 (14.2%) patients, respectively. An individual comparison of patients' PVC detection by the Holter monitor and SEP is shown in Figure 1B. The PVC burden evaluated by the Holter monitor was repeated with almost the same value as the PVC burden estimated by

XSL•FO
RenderX

SEP on the first day, whereas PVC burden variation was observed across days on SEP (Figure 1B). The validation of the diagnostic yield of SEP was performed by Bland-Altman analysis, and high agreement of PVC detection between the Holter monitor and SEP was confirmed with a mean difference of –0.07% and 95% limit of agreement of –1.44% to 1.30% (Figure 2A). PVC burdens detected by the Holter monitor and SEP on the first day of monitoring were highly correlated with each other (Figure 2B, R^2=0.995), indicating the accurate detection of PVC by SEP.

Table 2. Evaluation of premature ventricular complex by the Holter monitor and the adhesive single-lead electrocardiogram patch (N=134).

	Holter monitor	Single-lead ECG[a] patch	*P* value
Wear time (min), mean (SD)	1395.1 (33.5)	4318 (12.4)	N/A[b]
Noise (%), median (IQR)	0.1 (0.1-0.1)	0.9 (0.4-2.3)	<.001
Average heart rate (bpm), mean (SD)	71.1 (8.3)	72 (7.8)[c]	.01
Minimum heart rate (bpm), mean (SD)	52.1 (39.5)	49.1 (23.6)[c]	.44
Maximum heart rate (bpm), mean (SD)	122.4 (18.1)	131.8 (20.8)[c]	<.001
Premature ventricular complex information			
Patients with PVC[d] burden ≥5%, n (%)	56 (41.8)	57 (42.5)	<.001
Patients with PVC burden ≥10%, n (%)	37 (27.6)	38 (28.4)	<.001
Patients with PVC burden ≥15%, n (%)	23 (17.2)	22 (16.4)	<.001
Burden of PVC among all patients (%), mean (SD); median (IQR)	7.4 (10); 2.4 (0.2-10.9)	7.3 (9.4)[e]; 3.3 (0.3-11.7)[e]	.58
Burden of PVC among patients with burden ≥5% (%), mean (SD); median (IQR)	16.4 (9.9); 13.3 (8-24)	15.8 (9.3)[f]; 13.2 (9.7-20.5)[f]	.10

[a]ECG: electrocardiogram.

[b]N/A: not applicable.

[c]During the first 24-hour monitoring with the single-lead electrocardiogram patch, the average, minimum, and maximum heart rates (bpm) were 72.2 (SD 8.3), 48.8 (SD 6.2), and 123.2 (SD 17.7), respectively. The *P* values with the comparison of each value with the Holter monitor were <.001, .04, and .002, respectively.

[d]PVC: premature ventricular complex.

[e]During the first 24-hour monitoring with the single-lead electrocardiogram patch, the burden of premature ventricular complex among all patients was 7.5% (SD 10%; median 2.8%, IQR 0.3%-11%). The *P* value with the comparison of premature ventricular complex burden detected on the Holter monitor was .24.

[f]During the first 24-hour monitoring with the single-lead ECG patch, the burden of premature ventricular complex among patients with burden ≥5% was 16.5% (SD 9.9%; median 13.7%, IQR 8.1%-24.4%). The *P* value with the comparison of premature ventricular complex burden detected on the Holter monitor was .82.

Figure 1. (A) Distribution of the patients with premature ventricular complex. (B) Comparison of the burden of premature ventricular complex evaluated by the Holter monitor and the adhesive single-lead electrocardiogram patch. ECG: electrocardiogram; PVC: premature ventricular complex.

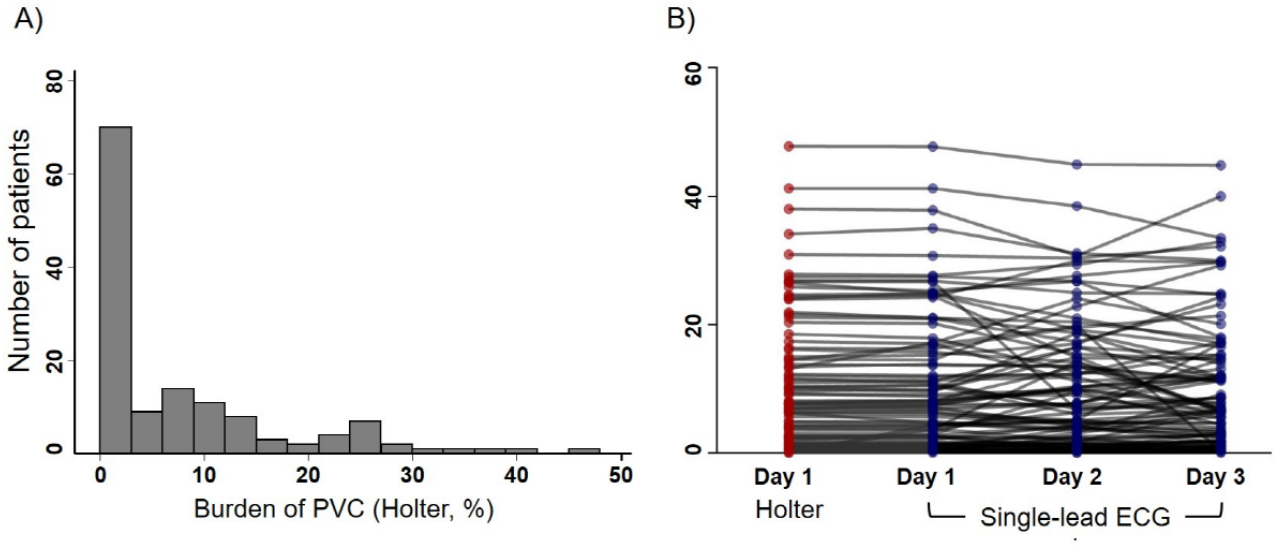

XSL•FO
RenderX

Figure 2. Agreement for the premature ventricular complex burden between Holter monitoring and adhesive single-lead electrocardiogram patch monitoring (mobiCARE MC-100) on day 1 analyzed by (A) Bland-Altman plot with limits of agreement and (B) scatter plot. LOA: limits of agreement; PVC: premature ventricular complex.

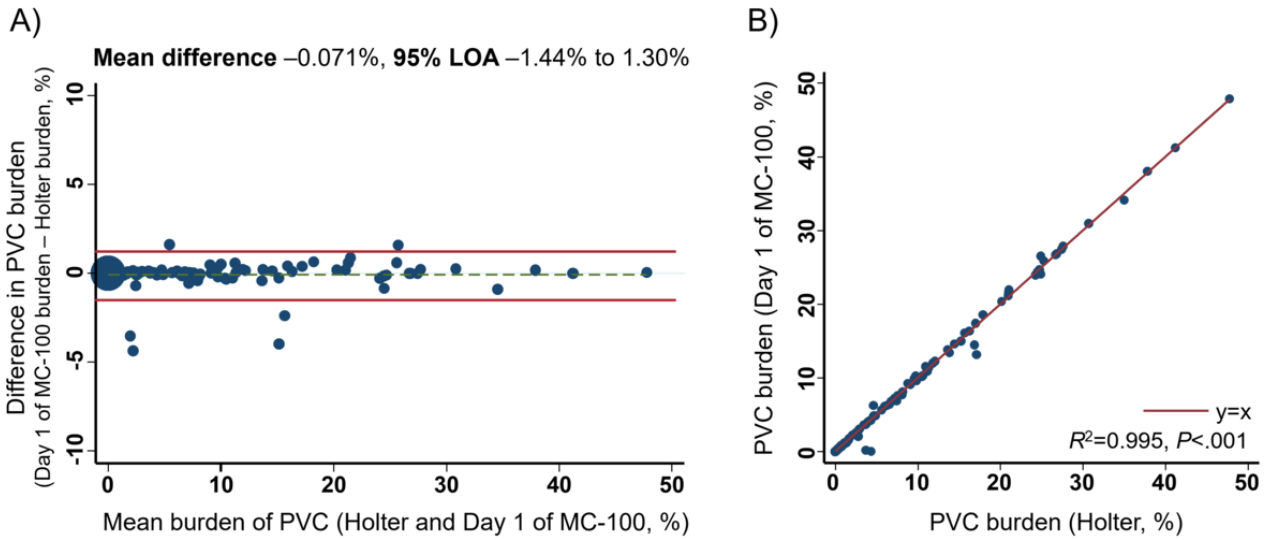

Variability of PVC Burden and Clinical Factors Associated With Greater Variability

PVC burden varied widely from day to day and hour to hour, as shown in Figure 1B. Greater variations in PVC burden were observed when evaluated every 6 hours. In the same patient, the maximum daily PVC burden was 1.68-fold (IQR 1.31-3.02) higher than the minimum daily burden and the maximum 6-hour burden was 12.08-fold (IQR 4.00-57.50) higher than the minimum 6-hour burden; for those with PVC burden ≥1%, the maximum daily and 6-hour PVC burden were 1.52-fold (IQR 1.28-2.07) and 12.17-fold (IQR 5.06-49.46) higher than each minimum burden, respectively. Moreover, the variation was dependent on the individuals' burden value; patients with higher PVC burden exhibited greater daily or hourly burden variation. Daily and 6-hour burden variations according to the value of PVC burden are presented in Figure 3, showing a moderate to high degree of correlation (R^2=0.337 for daily and 0.483 for 6-hour burden variation). An example of an ECG strip of an individual presenting a high variation in daily and 6-hour PVC burden monitored using SEP is shown in Multimedia Appendix 4.

The number of patients with PVC burden lower than the clinically significant thresholds was 42 for 10%, 56 for 15%, and 60 for 20%. However, the 3-day extended monitoring additionally identified that 29% (12/42), 18% (10/56), and 7% (4/60) of the patients reached 10%, 15%, and 20% thresholds of the PVC burden, respectively (Figure 4, Table 3).

The clinical features of those crossing the PVC burden cutoff of 10%, 15%, or 20% during the 3-day extended monitoring were compared with those of the other patients with PVC burden lower than the clinically significant thresholds (Multimedia Appendix 5). Patients who were additionally identified to reach clinically significant thresholds during the 3-day extended monitoring were younger (*P*=.03), had congestive heart failure (*P*=.01) or cardiomyopathy (*P*=.01), and were using amiodarone (*P*=.04). When multivariable logistic regression analysis was performed, young age was independently associated with possible additional detection of PVC burden with clinical significance (Table 4; β=.95, 95% CI .92-.99; *P*=.02 for age).

XSL•FO
RenderX

Figure 3. Correlation of the premature ventricular complex burden with (A) daily burden variation and (B) 6-hour burden variation. PVC: premature ventricular complex.

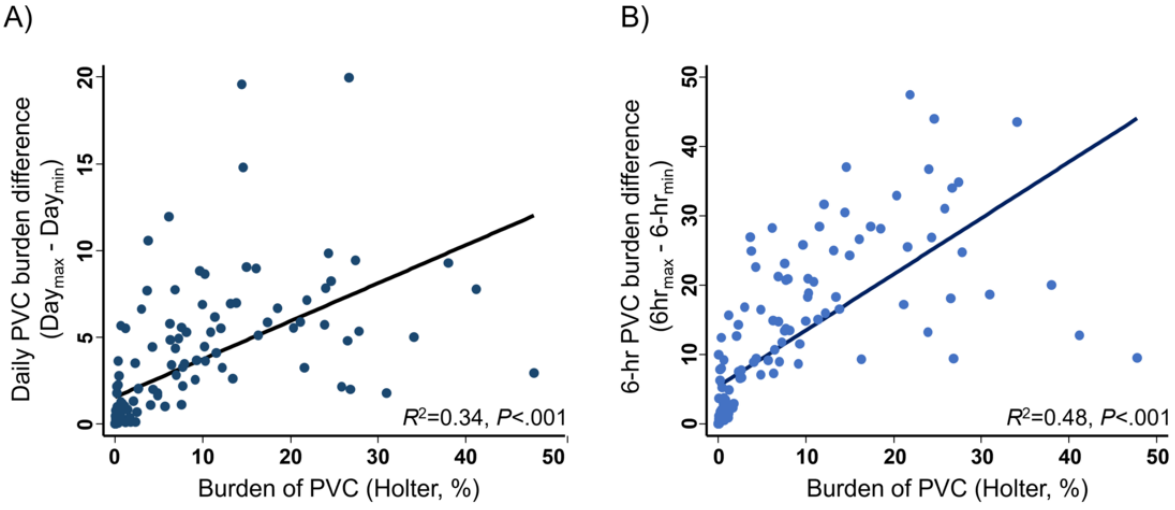

Figure 4. Patients additionally identified to reach the daily premature ventricular complex burden of 10%, 15%, and 20% during the 3-day monitoring with the adhesive single-lead electrocardiogram patch (mobiCARE MC-100). PVC: premature ventricular complex.

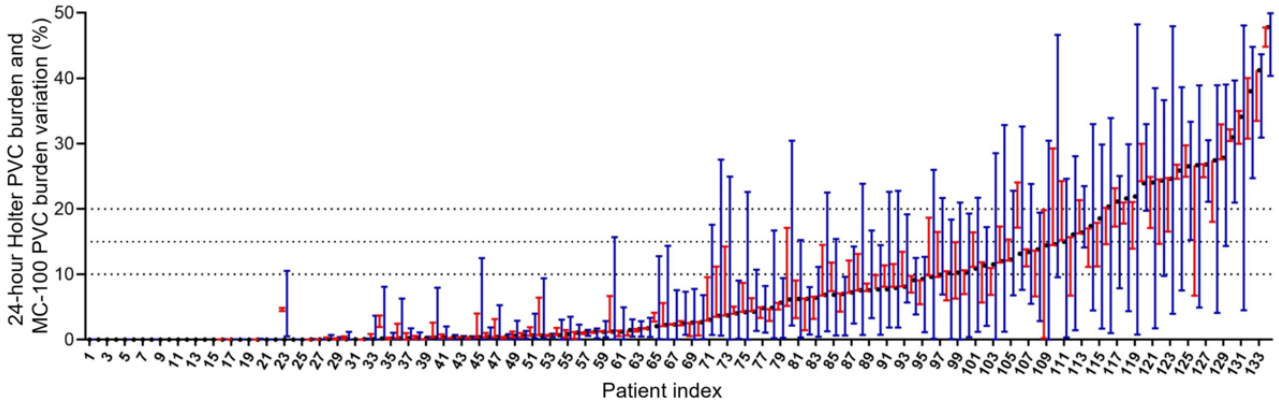

Table 3. Patients additionally identified to reach the daily premature ventricular complex burden of 10%, 15%, and 20% during the 3-day monitoring with the adhesive single-lead electrocardiogram patch.

PVC[a] threshold	Patients with PVC burden lower than the threshold (n)	Patients identified to reach the threshold during the 3-day monitoring with SEP[b], n (%)	
		Daily burden	6-hour burden
10%	42	12 (29)	25 (60)
15%	56	10 (18)	31 (55)
20%	60	4 (7)	24 (40)

[a]PVC: premature ventricular complex.

[b]SEP: single-lead electrocardiogram patch.

XSL•FO
RenderX

Table 4. Clinical factors associated with additional detection of premature ventricular complex burden of thresholds 10%, 15%, and 20%.

Variable	Univariable analysis		Multivariable analysis	
	Odds ratio (95% CI)	*P* value	β (95% CI)	*P* value
Age	0.961 (0.928-0.996)	.03	.95 (.92-.99)	.02
Congestive heart failure	6.875 (1.562-30.256)	.01	N/A[a]	N/A
Cardiomyopathy	19.94 (1.960-202.893)	.01	N/A	N/A
Amiodarone	12.556 (1.082-145.709)	.04	N/A	N/A
Burden of premature ventricular complex	1.024 (0.982-1.069)	.27	N/A	N/A

[a]N/A: not applicable.

Patients' Experience of 3-Day Extended ECG Monitoring

A survey on patients' experience with Holter monitoring and SEP was completed by 124 patients (response rate, 124/134, 92.5%), and the results are detailed in Multimedia Appendix 6. Overall, patients experienced lesser discomfort and lesser skin irritability with SEP compared to that with the Holter monitor ($P<.001$ and $P=.02$, respectively), but they wished to stop monitoring with SEP due to the longer duration of the examination. Indeed, since the patients wished to stop monitoring with SEP due to the longer duration of the examination, they detached the SEP and reattached the SEP later, thereby resulting in more instances of reattachment of the SEP (14/112, 12.5% for the Holter monitor and 66/118, 55.9% for SEP; $P=.02$). Thus, patients were more favorable to using SEP than the Holter monitor; 75.7% (84/111) of the patients responded to choosing a patch monitor for their next examination.

Discussion

Principal Results

Our principal findings are as follows: (1) SEP can diagnose PVC with comparable accuracy to the 24-hour Holter monitor, (2) the wide variation in the PVC burden recorded by SEP was proportional to the overall PVC frequency, (3) 3-day extended monitoring could identify those who reach clinically significant cutoffs of PVC burden, and (4) young age is associated with additional detection of clinically significant burdens of PVC during extended monitoring.

Comparison to Prior Work

Ambulatory ECG monitoring is the principal diagnostic tool for evaluating PVC burden, and recently developed wearable ECG monitoring devices enable extended monitoring for more than 24 hours [1,16,17]. In particular, SEP is considered to be a sufficient and practical method to evaluate the PVC burden accurately and comprehend the wide variations daily [11,18]. Previous studies [19,20] have demonstrated variations in the PVC burden to up to 23% change across days, while a recent study [10] has reported a 2.5-fold median difference between the maximum and minimum 24-hour PVC burden in the same patient over a 14-day monitoring period. Regarding the optimal duration of PVC monitoring, the median time to detect one's maximum PVC burden was reported as 6 (IQR 2-11) days [12], and the ideal duration for accurate PVC burden assessment was suggested as 7 days [13]. Although the monitoring period in our study (3 days) was shorter than the observation period in previous reports [12,13], we still observed a considerable intraindividual median daily PVC burden difference of 1.7-fold. Given that the optimal period for monitoring PVCs has not been established and a long recording length could be burdensome for both patients and clinicians, we suggest a compromise in the monitoring length (3 days), thereby providing an acceptable level of daily PVC burden, which may improve patients' compliance and adherence. Despite the general comfort of SEP, our survey results revealed that patients felt more likely to discontinue monitoring as the monitoring time became longer. Furthermore, an extended monitoring duration could potentially exacerbate patient discomfort and fatigue because of the requirement of carrying their mobile phones for data transmission. The discomfort could also be intensified by the alarm systems and skin irritability. In our study, 9 patients withdrew from the examination due to these practical challenges; thus, optimizing the monitoring duration could potentially decrease the dropout rate associated with these issues.

A wide variation in the daily PVC burden detected during the 3-day monitoring enabled us to additionally identify 6.7%-28.6% of the patients who reach clinically significant thresholds (10%, 15%, or 20%). Compared to a previous study [12] that showed that extended monitoring nearly doubled the identification of those reaching the 10% threshold during 14-day monitoring, the low proportion of identification in our study could be due to the shorter duration of monitoring (3 days). However, this finding is consistent with the fact that extended monitoring brings incremental gain in identifying patients reaching the significant cutoffs of daily PVC burden. We observed that SEP could diagnose PVC with almost the same accuracy as the Holter monitor, and 3-day monitoring could detect patients who might benefit from an advanced treatment modality such as radiofrequency catheter ablation.

Implications for Practice

Several risk factors are known to be associated with the presence of PVC and higher PVC frequency. Increased age, taller height, higher blood pressure, and unhealthy lifestyle behaviors such as smoking and less physical activity are consistently shown to be associated with the presence or higher frequency of PVC [21-23]. However, no study has demonstrated the clinical factors related to the large variations in PVC burden, which might be necessary to stratify patients requiring extended monitoring for

XSL•FO
RenderX

a more accurate assessment of PVC burden. We found that the PVC burden evaluated during the first 24-hour monitoring was correlated not only with the day-to-day but also with the hour-to-hour variation. The higher frequency of PVCs was linked to a wider variation in the overall burden. In addition, we found that younger age, presence of congestive heart failure or cardiomyopathy, and use of amiodarone are clinical factors that may increase the likelihood that PVC burden reaches clinically significant thresholds during the 3-day extended monitoring. Above all, multivariable analysis showed that younger age was consistently associated with the significant variations in the PVC burden (P=.02). Although no explainable mechanism for this association has been reported yet, the fact that older age and a lack of diurnal variation of PVC frequency are related to a higher risk of PIC [24,25] imply that there might be a link between young age and a greater variation in the PVC burden. Our analyses revealed patients who required extended monitoring for a more precise evaluation of PVC burden, thus improving the determination of the treatment strategy. The variable circadian distribution of PVC burden and its associated clinical factors can provide additional value, as this information can be used to guide the pharmacologic induction of PVCs during radiofrequency catheter ablation and predict the outcome in patients [25].

Strengths and Limitations

Our study has several limitations. First, the diagnostic yield of PVC via SEP and the degree of variation monitored for 3 days should be examined in a larger population. Second, since most of the included patients had normal ventricular systolic function, the accurate diagnosis of PVC and changes in PVC frequency monitored by SEP should be further validated in patients with reduced ejection fraction or structural heart diseases. Third, more than two-thirds of the patients had low PVC burden (<10%); thus, the extrapolation of our findings in patients with higher PVC burden (10%, 15%, or 20%) could be limited. Nonetheless, the distribution of the PVC burden in our study may represent the characteristics of patients in real-world settings, suggesting that our findings can be readily applied in a clinical context. Lastly, the external replication of our findings by using other wearable ECG patches in a multicenter setting would be required for the confirmation of generalizability.

Conclusions

In a single-center prospective registry of patients with PVC, we validated that SEP can accurately diagnose PVC with almost the same yield as the 24-hour Holter monitor. We found that during the 3-day extended monitoring with SEP, the significant fluctuations in daily and 6-hour PVC burden were proportional to the overall PVC frequency. Further, 3-day extended monitoring of PVC by using SEP enabled the identification of more patients exceeding the clinically significant burden threshold; thus, SEP monitoring could be a practical method with acceptable patient adherence to enhance the decision on the optimal treatment strategy (ie, catheter ablation) for PVC. Several clinical factors, especially younger age, were associated with a higher variation of daily PVC burden, implying the necessity of extended monitoring for accurate assessments.

Acknowledgments

This work was supported in part by the Korea Medical Device Development Fund grant funded by the Korean government (Ministry of Science and Information and Communication Technology, Ministry of Trade, Industry and Energy, Ministry of Health and Welfare, and Ministry of Food and Drug Safety; project HI20C1662, 1711138358, KMDF_PR_20200901_0173) and a grant funded by Seers Technology.

Data Availability

All data generated or analyzed during this study are included in this paper and Multimedia Appendices 1-6. The data of this study will be shared upon reasonable request to the corresponding author.

Authors' Contributions

HJA conceptualized the study design, conducted data curation, performed formal analysis, visualized the data, and was primarily responsible for writing this paper. EKC conceptualized and reviewed the study design, performed data collection, reviewed the manuscript, supervised the study, and performed project administration with funding acquisition. SK and SO reviewed the methodology, investigation, and the manuscript. SRL collected data and reviewed the methodology, investigation, and manuscript. HSS and YSL conducted data curation, performed formal analysis, prepared the resources and software, and reviewed the manuscript. EKC is the guarantor of this work and takes responsibility for the integrity and accuracy of the data analysis. All authors read and approved the final manuscript.

Conflicts of Interest

HSS and YSL are stockholders of Seers Technology Co, Ltd. EKC received the research grants or speaking fees from Abbott, Bayer, BMS/Pfizer, Biosense Webster, Chong Kun Dang, Daewoong Pharmaceutical Co, Daiichi-Sankyo, DeepQure, Dreamtech Co, Ltd, Jeil Pharmaceutical Co Ltd, Medtronic, Samjinpharm, Seers Technology, and Skylabs. Stock options were received from Seers Technology and Skylab.

Multimedia Appendix 1

XSL•FO
RenderX

STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines.
[DOCX File , 18 KB - jmir_v26i1e46098_app1.docx]

Multimedia Appendix 2
Process of validating the electrocardiogram algorithm for premature ventricular complex detection by single-lead electrocardiogram patch and its performance.
[DOCX File , 538 KB - jmir_v26i1e46098_app2.docx]

Multimedia Appendix 3
An example strip of premature ventricular complex detection by the single-lead electrocardiogram patch.
[DOCX File , 242 KB - jmir_v26i1e46098_app3.docx]

Multimedia Appendix 4
An example of an electrocardiogram strip of an individual presenting a high variation in daily and 6-hour premature ventricular complex burden monitored using a single-lead electrocardiogram patch.
[DOCX File , 211 KB - jmir_v26i1e46098_app4.docx]

Multimedia Appendix 5
Comparison of the clinical features of patients crossing the premature ventricular complex burden cutoffs of 10%, 15%, or 20% during the 3-day extended monitoring.
[DOCX File , 17 KB - jmir_v26i1e46098_app5.docx]

Multimedia Appendix 6
Survey completed by 124 patients on their experience with the Holter monitor and single-lead electrocardiogram patch.
[DOCX File , 13 KB - jmir_v26i1e46098_app6.docx]

References

1. Marcus GM. Evaluation and management of premature ventricular complexes. Circulation 2020 Apr 28;141(17):1404-1418. [doi: 10.1161/CIRCULATIONAHA.119.042434] [Medline: 32339046]
2. El Kadri M, Yokokawa M, Labounty T, Mueller G, Crawford T, Good E, et al. Effect of ablation of frequent premature ventricular complexes on left ventricular function in patients with nonischemic cardiomyopathy. Heart Rhythm 2015 Apr;12(4):706-713. [doi: 10.1016/j.hrthm.2014.12.017] [Medline: 25527251]
3. Kennedy HL, Whitlock JA, Sprague MK, Kennedy LJ, Buckingham TA, Goldberg RJ. Long-term follow-up of asymptomatic healthy subjects with frequent and complex ventricular ectopy. N Engl J Med 1985 Jan 24;312(4):193-197. [doi: 10.1056/NEJM198501243120401] [Medline: 2578212]
4. Bigger JT, Fleiss JL, Kleiger R, Miller JP, Rolnitzky LM. The relationships among ventricular arrhythmias, left ventricular dysfunction, and mortality in the 2 years after myocardial infarction. Circulation 1984 Feb;69(2):250-258. [doi: 10.1161/01.cir.69.2.250] [Medline: 6690098]
5. Baman TS, Lange DC, Ilg KJ, Gupta SK, Liu T, Alguire C, et al. Relationship between burden of premature ventricular complexes and left ventricular function. Heart Rhythm 2010 Jul;7(7):865-869. [doi: 10.1016/j.hrthm.2010.03.036] [Medline: 20348027]
6. Bogun F, Crawford T, Reich S, Koelling TM, Armstrong W, Good E, et al. Radiofrequency ablation of frequent, idiopathic premature ventricular complexes: comparison with a control group without intervention. Heart Rhythm 2007 Jul;4(7):863-867. [doi: 10.1016/j.hrthm.2007.03.003] [Medline: 17599667]
7. Ban J, Park H, Park J, Nagamoto Y, Choi J, Lim H, et al. Electrocardiographic and electrophysiological characteristics of premature ventricular complexes associated with left ventricular dysfunction in patients without structural heart disease. Europace 2013 May;15(5):735-741. [doi: 10.1093/europace/eus371] [Medline: 23194696]
8. Hasdemir C, Ulucan C, Yavuzgil O, Yuksel A, Kartal Y, Simsek E, et al. Tachycardia-induced cardiomyopathy in patients with idiopathic ventricular arrhythmias: the incidence, clinical and electrophysiologic characteristics, and the predictors. J Cardiovasc Electrophysiol 2011 Jun;22(6):663-668. [doi: 10.1111/j.1540-8167.2010.01986.x] [Medline: 21235667]
9. Latchamsetty R, Bogun F. Premature ventricular complex-induced cardiomyopathy. JACC Clin Electrophysiol 2019 May;5(5):537-550 [FREE Full text] [doi: 10.1016/j.jacep.2019.03.013] [Medline: 31122375]
10. Mullis AH, Ayoub K, Shah J, Butt M, Suffredini J, Czarapata M, et al. Fluctuations in premature ventricular contraction burden can affect medical assessment and management. Heart Rhythm 2019 Oct;16(10):1570-1574. [doi: 10.1016/j.hrthm.2019.04.033] [Medline: 31004780]

XSL•FO
RenderX

11. Voskoboinik A, Hadjis A, Alhede C, Im SI, Park H, Moss J, et al. Predictors of adverse outcome in patients with frequent premature ventricular complexes: The ABC-VT risk score. Heart Rhythm 2020 Jul;17(7):1066-1074. [doi: 10.1016/j.hrthm.2020.02.020] [Medline: 32109563]
12. Loring Z, Hanna P, Pellegrini CN. Longer ambulatory ECG monitoring increases identification of clinically significant ectopy. Pacing Clin Electrophysiol 2016 Jun;39(6):592-597. [doi: 10.1111/pace.12852] [Medline: 27005323]
13. Hsia BC, Greige N, Patel SK, Clark RM, Ferrick KJ, Fisher JD, et al. Determining the optimal duration for premature ventricular contraction monitoring. Heart Rhythm 2020 Dec;17(12):2119-2125. [doi: 10.1016/j.hrthm.2020.07.013] [Medline: 32679267]
14. Zeppenfeld K, Tfelt-Hansen J, de Riva M, Winkel BG, Behr ER, Blom NA, ESC Scientific Document Group. 2022 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death. Eur Heart J 2022 Oct 21;43(40):3997-4126. [doi: 10.1093/eurheartj/ehac262] [Medline: 36017572]
15. Kwon S, Lee S, Choi E, Ahn H, Song H, Lee Y, et al. Comparison between the 24-hour Holter test and 72-hour single-lead electrocardiogram monitoring with an adhesive patch-type device for atrial fibrillation detection: prospective cohort study. J Med Internet Res 2022 May 09;24(5):e37970 [FREE Full text] [doi: 10.2196/37970] [Medline: 35532989]
16. Al-Khatib SM, Stevenson WG, Ackerman MJ, Bryant WJ, Callans DJ, Curtis AB, et al. 2017 AHA/ACC/HRS guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Heart Rhythm 2018 Oct;15(10):e73-e189. [doi: 10.1016/j.hrthm.2017.10.036] [Medline: 29097319]
17. Turakhia MP, Hoang DD, Zimetbaum P, Miller JD, Froelicher VF, Kumar UN, et al. Diagnostic utility of a novel leadless arrhythmia monitoring device. Am J Cardiol 2013 Aug 15;112(4):520-524 [FREE Full text] [doi: 10.1016/j.amjcard.2013.04.017] [Medline: 23672988]
18. Heckbert SR, Austin TR, Jensen PN, Floyd JS, Psaty BM, Soliman EZ, et al. Yield and consistency of arrhythmia detection with patch electrocardiographic monitoring: the multi-ethnic study of atherosclerosis. J Electrocardiol 2018;51(6):997-1002 [FREE Full text] [doi: 10.1016/j.jelectrocard.2018.07.027] [Medline: 30497763]
19. Winkle RA. Antiarrhythmic drug effect mimicked by spontaneous variability of ventricular ectopy. Circulation 1978 Jun;57(6):1116-1121. [doi: 10.1161/01.cir.57.6.1116] [Medline: 639231]
20. Morganroth J, Michelson EL, Horowitz LN, Josephson ME, Pearlman AS, Dunkman WB. Limitations of routine long-term electrocardiographic monitoring to assess ventricular ectopic frequency. Circulation 1978 Sep;58(3 Pt 1):408-414. [doi: 10.1161/01.cir.58.3.408] [Medline: 679430]
21. von Rotz M, Aeschbacher S, Bossard M, Schoen T, Blum S, Schneider S, et al. Risk factors for premature ventricular contractions in young and healthy adults. Heart 2017 May;103(9):702-707. [doi: 10.1136/heartjnl-2016-309632] [Medline: 27798051]
22. Kerola T, Dewland TA, Vittinghoff E, Heckbert SR, Stein PK, Marcus GM. Modifiable predictors of ventricular ectopy in the community. J Am Heart Assoc 2018 Nov 20;7(22):e010078 [FREE Full text] [doi: 10.1161/JAHA.118.010078] [Medline: 30571495]
23. Simpson RJ, Cascio WE, Schreiner PJ, Crow RS, Rautaharju PM, Heiss G. Prevalence of premature ventricular contractions in a population of African American and white men and women: the Atherosclerosis Risk in Communities (ARIC) study. Am Heart J 2002 Mar;143(3):535-540. [doi: 10.1067/mhj.2002.120298] [Medline: 11868062]
24. Bas HD, Baser K, Hoyt J, Yokokawa M, LaBounty T, Morady F, et al. Effect of circadian variability in frequency of premature ventricular complexes on left ventricular function. Heart Rhythm 2016 Jan;13(1):98-102. [doi: 10.1016/j.hrthm.2015.07.038] [Medline: 26247319]
25. Hamon D, Abehsira G, Gu K, Liu A, Blaye-Felice Sadron M, Billet S, et al. Circadian variability patterns predict and guide premature ventricular contraction ablation procedural inducibility and outcomes. Heart Rhythm 2018 Jan;15(1):99-106. [doi: 10.1016/j.hrthm.2017.07.034] [Medline: 28765087]

Abbreviations

ECG: electrocardiogram
PIC: premature ventricular complex–induced cardiomyopathy
PVC: premature ventricular complex
SEP: single-lead electrocardiogram patch
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by T Leung, G Eysenbach; submitted 30.01.23; peer-reviewed by Y Guo, M Mekhael; comments to author 16.06.23; revised version received 13.08.23; accepted 12.02.24; published 21.03.24.

Please cite as:
Ahn HJ, Choi EK, Lee SR, Kwon S, Song HS, Lee YS, Oh S
Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation
J Med Internet Res 2024;26:e46098
URL: https://www.jmir.org/2024/1/e46098
doi:10.2196/46098
PMID:38512332

©Hyo-Jeong Ahn, Eue-Keun Choi, So-Ryoung Lee, Soonil Kwon, Hee-Seok Song, Young-Shin Lee, Seil Oh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessment of the Barriers and Enablers of the Use of mHealth Systems in Sub-Saharan Africa According to the Perceptions of Patients, Physicians, and Health Care Executives in Ethiopia: Qualitative Study

Genet Tadese Aboye[1,2], BSc, MSc, PGD; Gizeaddis Lamesgin Simegn[2], BSc, MSc, PhD; Jean-Marie Aerts[1], BSc, MSc, PhD

[1]M3-BIORES (Measure, Model & Manage Bioreponses), Division of Animal and Human Health Engineering, Department of Biosystems, KU Leuven, Leuven, Belgium

[2]School of Biomedical Engineering, Jimma Institute of Technology, Jimma University, Jimma, Ethiopia

Corresponding Author:
Genet Tadese Aboye, BSc, MSc, PGD
M3-BIORES (Measure, Model & Manage Bioreponses)
Division of Animal and Human Health Engineering, Department of Biosystems
KU Leuven
kasteelpark Arenberg 30
Leuven, 3001
Belgium
Phone: 32 489304866
Email: genettadese.aboye@kuleuven.be

Abstract

Background: Digital technologies are increasingly being used to deliver health care services and promote public health. Mobile wireless technologies or mobile health (mHealth) technologies are particularly relevant owing to their ease of use, broad reach, and wide acceptance. Unlike developed countries, Sub-Saharan Africa experiences more challenges and obstacles when it comes to deploying, using, and expanding mHealth systems. In addition to barriers, there are enabling factors that could be exploited for the design, implementation, and scaling up of mHealth systems. Sub-Saharan Africa may require tailored solutions that address the specific challenges facing the region.

Objective: The overall aim of this study was to identify the barriers and enablers for using mHealth systems in Sub-Saharan Africa from the perspectives of patients, physicians, and health care executives.

Methods: Multi-level and multi-actor in-depth semistructured interviews were employed to qualitatively explore the barriers and enablers of the use of mHealth systems. Data were collected from patients, physicians, and health care executives. The interviews were audio recorded, transcribed verbatim, translated, and coded. Thematic analysis methodology was adopted, and NVivo software was used for the data analysis.

Results: Through this rigorous study, a total of 137 determinants were identified. Of these determinants, 68 were identified as barriers and 69 were identified as enablers. Perceived barriers in patients included lack of awareness about mHealth systems and language barriers. Perceived enablers in patients included need for automated tools for health monitoring and an increasing literacy level of the society. According to physicians, barriers included lack of available digital health systems in the local context and concern about patients' mHealth capabilities, while enablers included the perceived usefulness in reducing workload and improving health care service quality, as well as the availability of mobile devices and the internet. As perceived by health care executives, barriers included competing priorities alongside digitalization in the health sector and lack of interoperability and complete digitalization of implemented digital health systems, while enablers included the perceived usefulness of digitalization for the survival of the highly overloaded health care system and the abundance of educated manpower specializing in technology.

Conclusions: mHealth systems in Sub-Saharan Africa are hindered and facilitated by various factors. Common barriers and enablers were identified by patients, physicians, and health care executives. To promote uptake, all relevant stakeholders must actively mitigate the barriers. This study identified a promising outlook for mHealth in Sub-Saharan Africa, despite the present

XSL•FO
RenderX

barriers. Opportunities exist for successful integration into health care systems, and a user-centered design is crucial for maximum uptake.

(J Med Internet Res 2024;26:e50337) doi:10.2196/50337

KEYWORDS

barriers; chronic disease; digital health; eHealth; enablers; health care; mHealth; mobile health; perspectives; Sub-Saharan Africa

Introduction

eHealth, or the secure and cost-effective application of information and communication technology (ICT) to support health and health-related sectors, includes, but is not limited to, the use of mobile wireless technologies for public health or mobile health (mHealth) [1,2]. Digital technologies are increasingly being used to deliver health care services and promote public health. mHealth technologies are particularly relevant due to their ease of use, broad reach, and wide acceptance [3]. Global health service delivery could change as a result of the use of mHealth to support the accomplishment of health objectives. This transformation is being fueled by a potent confluence of different elements. These include the quick development of mobile technologies and apps, the rise of new chances for incorporating mHealth into already existing eHealth services, and the ongoing expansion of mobile cellular network coverage [4,5].

In the literature, it has been shown that mHealth systems in the form of SMS text messages, apps, and telemedicine projects are being used efficiently in the developed world. mHealth solutions are advancing rapidly in these regions [6-8]. The reason for this is the fact that there are a number of enabling factors that facilitate or encourage the use and implementation of such systems. Today, a wide variety of barriers and enablers are present in the developing world that influence the use and implementation of mHealth systems [9-11].

The prevalence and exploitation of mHealth technologies are steadily rising, presenting a significant opportunity for their incorporation into clinical services as a means to enhance the provision of high-quality medical care. Recent data on mobile phone usage in Sub-Saharan Africa revealed that 51% of residents possess a mobile device [12]. The latest statistics on mobile service subscriptions suggest an ongoing upward trajectory, with expectations of further increases. As of the end of 2020, 495 million individuals, equivalent to 46% of the population in Sub-Saharan Africa, held active mobile service subscriptions. Projections anticipate that this figure will rise, with an estimated 50% of the population (equivalent to 619 million individuals) anticipated to subscribe to mobile services by the end of 2025. Presently, there are 303 million internet users in Sub-Saharan Africa, constituting 28% of the population, and this number is projected to increase to 474 million by 2025 [12,13]. These data underscore the widespread penetration of mobile technology in the region and lay a solid foundation for exploring the potential advantages and challenges associated with its integration into health care services.

Even though the mass penetration of mobile phones is a fact and a very substantial condition to adopt mHealth systems in such countries, there are a number of factors that need to be considered when designing and developing mHealth systems in developing countries. Most mHealth systems initiated in such countries remain in the proof of concept or pilot testing stage. The large implementation of such systems is not widely documented [14,15].

mHealth approaches are widely embraced, and the health care system heavily relies on them in developed regions. From patient management systems to individual patient-centered mobile apps, these initiatives have been operational for a considerable period, in contrast to their counterparts in developing nations. A minor portion of Sub-Saharan African countries, less than half of the total [16,17], are currently endeavoring to incorporate mHealth platforms into their health care systems. This underscores the requirement for additional development of mHealth technology in the region. Ethiopia is currently focusing on strengthening its health care system and aligning it with the Sustainable Development Goals [18]. With a largely rural population, the nation encounters difficulties in accessing essential services like health care among others. A potential path for improvement lies in the digitalization of health care services, offering the opportunity to enhance efficiency, accessibility, and overall health care outcomes [19].

The implementation of digital health systems may encounter various challenges, and these challenges are generally context-specific. There is no universal solution that can address these challenges in all circumstances. Given the variation in challenges and the absence of a one-size-fits-all solution, success factors center on designing an mHealth platform that is specific to the context and the target population. This involves identifying these determining factors and incorporating sensitive design considerations to address unique needs [13]. Various studies have been performed for understanding technology adaption factors using various frameworks such as the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) [20-22]. The research conducted by Liu et al [21] using the UTAUT indicated that users' intentions to adopt mHealth systems are positively influenced by factors such as effort expectancy, performance expectancy, subjective norm, and perceived ubiquitousness. Notably, privacy concerns exhibited a significantly negative impact only on perceived ubiquity, with no significant effects observed on effort expectancy, performance expectancy, subjective norm, and intention to adopt [21]. The study by Yang et al [20] examined consumers' intentions and behaviors related to the use of digital applications based on the UTAUT and provided valuable guidance for broadening the use of mHealth apps among consumers.

A systematic research study by Jacob et al [23] comprehending the sociotechnical factors influencing patients' acceptance of

XSL•FO
RenderX

mHealth tools proposed adopting a patient-centric strategy by ensuring that the tools seamlessly integrate into the overall patient journey and treatment plan. This involves giving priority to inclusive design and ensuring thorough patient education and support. Different frameworks have been used to evaluate the implementation of mHealth approaches, but they fall short of addressing all aspects comprehensively. In response, researchers have put forth a consolidated framework to address this limitation by incorporating various factors such as organizational and policy factors, social and personal factors, and technical and material factors [24].

The digital divide, characterized by unequal access to digital technology, is another factor that may be worsened or improved through the adoption of mHealth [25-27]. Despite evidence showing increased access to mobile devices, there is insufficient implementation of mHealth in Sub-Saharan Africa. Accelerating the integration of digitalization into the heavily burdened health care system could help address the challenges associated with inadequate health care and contribute to narrowing the digital divide in the region. Sub-Saharan Africa presents a unique challenge when it comes to the implementation and scaling up of mHealth systems. In contrast to developed countries, the region faces significant barriers that hinder the adoption and effective use of these systems. However, despite the numerous obstacles facing mHealth adoption in Sub-Saharan Africa, there are also opportunities and enabling factors. Consequently, the design and implementation of mHealth interventions in Sub-Saharan Africa must be tailored to address these specific challenges. Ethiopia, a representative country of Sub-Saharan Africa, is considered as one of these nations with limited implementation and consumption of mHealth systems. Thus, there is a pressing need for a comprehensive study of the barriers and enablers for mHealth adoption in Sub-Saharan Africa in order to help guide the development of tailored mHealth interventions that are suited to the local context and can effectively address the unique challenges facing the region. For this, it is vital to investigate the determinants that could affect the use of mHealth approaches. mHealth strategies are tailored to diverse end users, with certain approaches adapted for organizational use, while others target health care professionals or patients, and sometimes both. The factors influencing each of these populations may exhibit overlaps, yet they are varied and complex, necessitating a comprehensive and separate investigation for each category.

The overall aim of this study was to identify the barriers and enablers from the perspectives of patients, physicians, and health care executives for using mHealth systems in Sub-Saharan Africa and to provide recommendations on mHealth system design and policy-making. The study's anticipated outcome is a description of the elements that encourage or inhibit the use of mHealth approaches, along with suggestions for resolving these barriers and exploring the enablers. This study can be used as a primary step in undertaking a user-centered design study of mHealth platforms in the Sub-Saharan African context.

Methods

Study Design

A multi-level and multi-actor in-depth semistructured interview was employed in order to identify the barriers and enablers of the use of mHealth systems.

Study Area and Sampling

The sample areas for this study included 1 city administration and 2 regions in Ethiopia, namely, Addis Ababa city administration, Oromia region, and Harari region. The study locations were selected based on the Human Development Index (HDI), which serves as a composite measure of a region's average achievements in 3 fundamental aspects of human development, namely, health, knowledge, and standard of living [28]. Participants were recruited through nonprobability sampling, with a specific emphasis on purposive sampling techniques. For recruiting individuals, the snowball sampling technique [29] was used by means of colleagues, organizational contacts, and initial participants.

Participants

Data were collected from patients, physicians, and health care executives. For the health care executive group, individuals eligible for participation included decision-making persons and managers of health offices, hospitals, or similar organizations. For the physician group, individuals eligible for participation included health care professionals working at chronic disease outpatient departments (OPDs), who possessed the ability to communicate in Amharic. Lastly, for the patient group, individuals eligible for participation included chronic disease patients aged between 17 and 50 years who could communicate effectively in Amharic. These criteria were carefully defined to ensure that the selection of participants aligned with the specific characteristics and roles of each group within the study.

The participants in the health care executive group had a variety of positions, including office and hospital directors, coordinators, and ICT heads in federal offices, health bureaus, and hospitals. Additionally, in-depth interviews were conducted with physicians working at chronic disease OPDs. Furthermore, patients who visited the selected health facilities for chronic disease follow-ups were involved in the study.

Data Collection and Analysis

Semistructured interviews [30,31] were selected for this study as they provide the required balance between flexibility and structure for our research. This approach offers the necessary room to explore a subject while maintaining sufficient structure to accomplish the objectives of the study. The interview guides for patients, physicians, and executives are presented in Multimedia Appendix 1, Multimedia Appendix 2, and Multimedia Appendix 3, respectively. The interview guides were structured based on the consolidated framework of the factors impacting clinicians' adaptation of mHealth [24]. All interviews were conducted in-person with the participants. Prior to conducting the interviews, participants received an information letter explaining the overall goal of the PhD research and this specific study. An oral explanation was also provided

XSL•FO
RenderX

where necessary. Informed consent was obtained from each participant in written form, and their participation was voluntary. When confidentiality and privacy could be assured, interviews with patients were conducted in the waiting area of the hospital. Interviews with physicians and executives were conducted at their offices. No monetary compensation was provided to any of the participants. All of the interviews were carried out by a female researcher (author GTA), who is a biomedical engineer and a PhD student with a focus on the design and development of mHealth systems for Sub-Saharan Africa. The researcher has experience and training in various research methodologies, including qualitative study. The interviews were conducted in Amharic and were audio recorded. The audio recordings from the interviews were first transcribed verbatim and then translated to English. No automatic tool was used for transcription. Google Translate was used for translation, and the information was checked manually for correctness. The thematic analysis methodology described by Braun et al [32] was employed for this study. The author GTA coded the interview transcripts and revised them with the 2 researchers JMA and GLS. NVivo software (QSR International) was used during the data analysis to code and categorize the data and to create a thematic framework. To report the study, we used the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [33], which has been provided in Multimedia Appendix 4.

Ethical Considerations

Ethical clearance was obtained from the KU Leuven Social and Societal Ethics Committee (SMEC) (G-2022-5491-R3(MIN)) and from the Jimma University Institute of Health Institutional Review Board (IRB) (JUIH/IRB/311/23).

Results

Overview

A total of 48 interviewees participated in the one-to-one interviews. In the patient group, 17 patients (10 men and 7 women) were interviewed. Eight additional patients were approached, but they declined to be interviewed as they were in a hurry, were not willing to be interviewed, or were already frustrated with the system. The mean patient age was 33.9 years, and the mean interview duration was 6.25 minutes.

In the physician group, 19 physicians (12 men and 7 women) were interviewed. Six additional physicians were approached, but they declined to be interviewed as they had hectic schedules. The physicians included in the study were general practitioners who worked in chronic disease OPDs and had an average of 5 years of experience. In this group of individuals, the mean interview duration was 15.7 minutes.

In the health care executive group, 12 health care executives (10 men and 2 women) were interviewed. Owing to a scheduling conflict, 1 additional executive declined to be interviewed. In this group, the mean interview duration was 19.34 minutes. The executives had a variety of positions, including directors in the country's health minister's office, directors in city health administrative offices at the zonal and regional levels, hospital chief clinical officers, and ICT heads at hospitals. Owing to the limited number of key respondents in the health care executive group, the number of participants in our study was restricted. Nonetheless, we ensured that the interview process was comprehensive enough to address all crucial aspects, and we included all key informants while also exploring any new viewpoints that emerged.

The total sample size was established based on data saturation, which occurs when new data no longer provide additional perspectives or insights after a certain point. In this study, data saturation was reached following 48 interviews (17 patients, 19 physicians, and 12 health care executives), indicating that an adequate sample size had been achieved. Table 1 provides an overview of the interview characteristics.

XSL•FO
RenderX

Table 1. Overview of the interview characteristics.

Variable	Patient group (n=17)	Physician group (n=19)	Health care executive group (n=12)
Gender, n (%)			
Male	10 (59)	12 (63)	10 (83)
Female	7 (41)	7 (37)	2 (17)
Age (years), n (%)			
20-25	5 (29)	N/A[a]	N/A
26-30	4 (24)	N/A	N/A
31-45	5 (29)	N/A	N/A
>45	3 (18)	N/A	N/A
Years in practice, n (%)			
<5	N/A	6 (32)	N/A
5-10	N/A	13 (68)	N/A
Position, n (%)			
Director	N/A	N/A	7 (58)
ICT[b] head	N/A	N/A	2 (17)
Coordinator	N/A	N/A	3 (25)
Location, n (%)			
Addis Ababa	6 (35)	7 (37)	5 (42)
Oromia	6 (35)	5 (26)	4 (33)
Harari	5 (30)	7 (37)	3 (25)
Interview duration (min), mean (SD)	6.25 (5.76)	15.70 (6.97)	19.34 (8.15)

[a]N/A: not applicable.

[b]ICT: information and communication technology.

A total of 137 determinants for using mHealth systems in Sub-Saharan Africa were identified from the perspectives of patients, physicians, and health care executives. Of the 137 determinants, 68 were barriers and 69 were enablers. Themes were derived from the data. Further categorization of these aspects in the study resulted in the identification of 3 major categories: organizational and policy, social and personal, and technical and material. These categories were based on the consolidated framework of the factors impacting clinicians' adaptation of mHealth [24], which takes into account the numerous factors influencing adoption of mHealth systems. The organizational and policy factors pertain to health care organizations' internal workings, workflow, policies, regulations, patient-related factors, and user engagement. The technical and material factors pertain to system design, system usefulness, IT capabilities, compatibility, data management, user experiences, monetary factors, and ease of use. The social and personal factors include personal characteristics, social and cultural factors, and moderating factors.

Perspectives of Patients

Among patients, after a detailed examination of the data, we identified 10 unique factors that could impede progress, which were referred to as barriers, and 8 unique factors that could facilitate progress, which were referred to as enablers. Figure 1 provides a comprehensive list of these barriers and enablers. Patients were assigned a code based on the study area as follows: those from Addis Ababa were coded "AP," those from Harari were coded "HP," and those from Jimma were coded "JP."

XSL•FO
RenderX

Figure 1. Barriers and enablers for using mobile health (mHealth) systems: patients' perspectives.

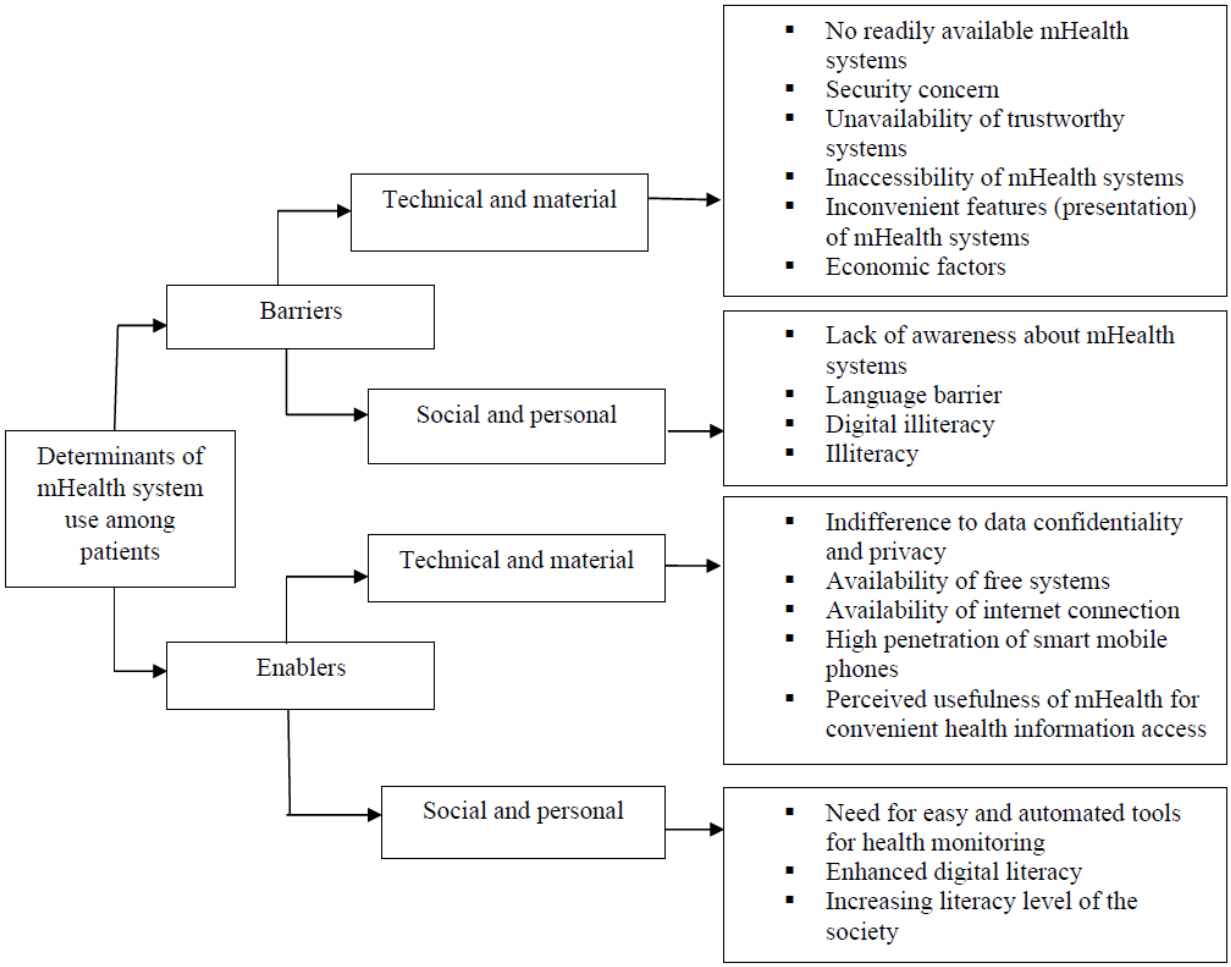

Barriers

Among the 10 barriers hindering the adoption of mHealth systems, 6 were under the technical and material category. These challenges included system design, system usefulness, IT capabilities, compatibility, data management, user experiences, monetary factors, and ease of use. The remaining 4 barriers were under the social and personal category, which included personal characteristics, social and cultural factors, and moderating factors.

The majority of patients reported that they have never used mHealth systems. However, most of them reported that they would use mHealth systems if they had awareness and if they could access these systems in an affordable way.

More than half of the participants had no awareness about mHealth systems and how they could benefit them. However, upon a detailed explanation about mHealth systems and their importance, they reported that they recognize the usefulness of such tools. They reported that they would use these systems if they could get trustable systems and if recommended by their doctors.

Regarding the barriers associated with using mHealth approaches, the most frequently cited theme was a lack of awareness about mHealth systems.

> *I don't have the information about such things either. Apart from seeing the texts that come to me, I never installed the application and tried nothing.* [AP4]
>
> *To be honest... I have never heard or seen health practitioners recommending this to me. I have no idea about such systems.* [AP6]
>
> *I don't know much about that, but I think it is useful for monitoring my diet and sugar levels.* [HP3]
>
> *The knowledge gap… For example, when I see my friends using such things, I don't know the source of it. And I learn from them, but for example, when I am talking about this idea with my friends with whom I have the same interest like me they ask me from where I get it from.* [JP2]
>
> *I don't have the information about this.* [JP6]

Patients also perceived the fact that mHealth systems are not readily available as a barrier to not using them. They reported that they browse Google or YouTube when they need to obtain information related to health.

> *I will download some stuff on YouTube. Video… I see…. I am downloading some video.* [JP4]
>
> *The only thing I've used so far is Walk Exercise. It is by just browsing.* [AP2]

XSL•FO
RenderX

The affordability of smart mobile phones, the requirement for a paid subscription, and the high cost of smart systems were identified as economic factors that can impede the adoption of mHealth systems. These factors are believed to create obstacles for individuals who may not have access to the necessary technology or cannot afford the associated costs.

> *Because this information is from the internet. It is not possible to access the internet if it is not purchased. Again, you can find that if you go to a place with internet. Even if I have a cell phone, it costs money.* [HP1]
>
> *But the affordability... I can't afford that mobile so I won't use it.* [HP1]
>
> *But with the systems, the smart ones are a little expensive, so the cost of the sensors is a barrier.* [AP6]

A lack of trustworthy mHealth systems is a significant obstacle to their adoption. Patients feared that the information and services provided by these systems may not be based on reliable evidence. This concern arose from the potential harm caused by inaccurate information or ineffective treatments.

> *I want it to be something secure. First is the issue of medication. Medical information is not something you can just throw away. Therefore, they have a question of credibility.* [AP4]
>
> *The reason I don't use it is because most of them are not desirable. Because I hear things that are not trustworthy; I won't take advantage of it.* [AP5]
>
> *The application must be supported by evidence.* [JP4]

Inconvenient features (presentation) of mHealth systems were also perceived as an obstacle to the use of these systems.

> *I think that a lot of data frustrates people. When you read more data, it is like an education. It needs your time. It needs your opinion. Anyone who is not in the health field may not have an interest in such things.* [JP2]
>
> *The presentation is a biggest barrier... even if it is translated... there are some words that are cultural and from our community language... there is, isn't it... I think there is a small barrier. Medical terms... because it's a bit difficult to how interpret them.* [AP6]

Digital illiteracy, which refers to the lack of ability to effectively use digital technologies, such as computers, smartphones, and the internet, is a recognized barrier to the adoption of mHealth technologies. This challenge is believed to limit the ability to take advantage of the benefits provided by mHealth tools.

> *Lack of skill in using mobile phones. Configuring systems specially in IOS it is also same for android is difficult. Like Entering Personal Details. And because the app won't start without you doing it.* [AP6]
>
> *I don't know how to use this kind of thing. I don't have the knowledge of how to use such systems. So, I never used it.* [HP4]

Enablers

Among the 8 enablers facilitating the adoption of mHealth systems, 5 were under the technical and material category and the remaining 3 were under the social and personal category.

Patients with chronic diseases had a need for easy and automated tools for health monitoring.

> *It's a matter of health. To control my blood sugar level, there are mandatory things that I have to do. So if there are things installed on my phone to help me with this, it helps.* [AP2]
>
> *I am suffering from hemorrhoids. Some say wash with cold water; Some say to wash with warm water; not to be confused it will be good if there is a tool to use in my phone.* [AP3]
>
> *One of the things that inspired me to use it was to find out how many calories I burned in a day; It reminds me how much I should move, if I don't move, I will be exposed to other related diseases. It means it has health benefit for myself.* [AP6]
>
> *It will be very good. I may not always in need to go to the hospital. I am able to adjust myself from that information; Use the medicine on time. Adjusting the food system; doing activities; There are also recommendations. I have arrived to that conclusion. And it's good. And it's important to have.* [HP1]
>
> *I think it would be helpful to have more of information, especially to monitor my health.* [HP4]
>
> *I think it makes things easier for us. If I use it, it will be beneficial. The usefulness is very high.* [JP1]

Patients recognized the availability of the internet as one of the facilitating factors of the use of mHealth systems as it could help them access such platforms.

> *At home, I have Wi-Fi, so it's easy.* [AP2]
>
> *Now there is telephone, there is Internet, there is Wi-Fi, now there is internet even in a pool house, there is even in a tea house.* [JP2]
>
> *There is a big difference between where there is internet and where there is no internet. Internet is a way to get many new technologies. We have it now more than ever.* [JP4]

Another enabler was the perceived usefulness of mHealth as a convenient tool to access health information.

> *…. And it helps you to be proactive about your health. It means it prevents you from going to the hospital after something happens.* [AP6]

Perspectives of Physicians

Among physicians, we identified a total of 54 factors that were perceived to influence the adoption of mHealth systems. Of the 54 factors, 27 were identified as barriers that hinder the use of mHealth systems and 27 were identified as enablers that facilitate their adoption. The study further classified these factors into 3 main categories, namely, organizational and policy, social and personal, and technical and material. Physicians were assigned a code based on the study area as follows: those from

XSL•FO
RenderX

Addis Ababa were coded “AD,” those from Harari were coded “HD,” and those from Jimma were coded “JD.”

Barriers

Specifically, among the 27 identified barriers, 7 were under the organizational and policy category, 11 were under the social and personal category, and 9 were under the technical and material category. Figure 2 displays the obstacles to the use of mHealth systems from the viewpoint of physicians.

Figure 2. Barriers for using mobile health (mHealth) systems: physicians’ perspectives.

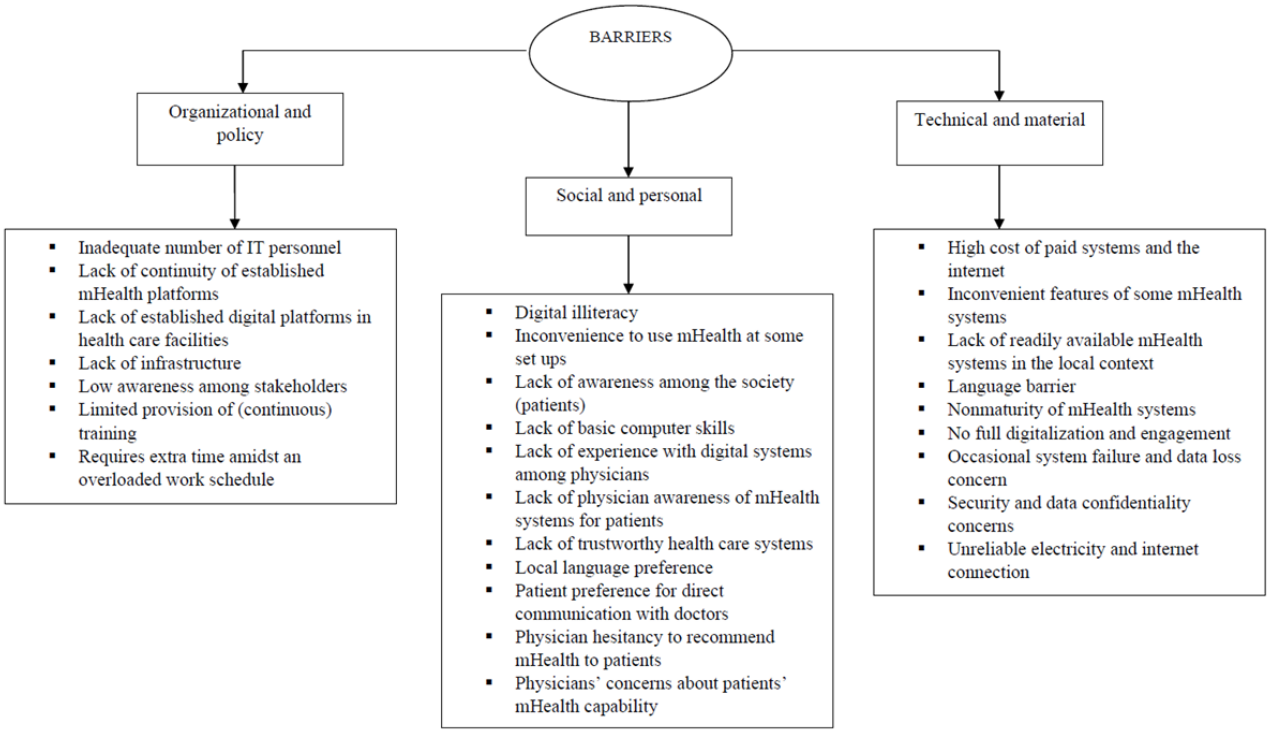

Physicians identified the lack of established mHealth systems as the primary reason for not using them. They acknowledged that their lack of exposure to these systems has prevented them from gaining experience in their use.

> *Actually, there is no such system or application based health care at the organization where I work.* [HD3]
>
> *For example, a patient comes to us and we ask about the patient's problem, then we ask and investigate to do a case diagnosis, and at the end we write an investigation paper to investigate and send it to the patient. And we give them hard paper. And we have not prepared them in a computerized way. There is none.* [HD6]
>
> *The main thing is that this type of system does not exist here.* [JD2]

According to physicians, certain types of mHealth systems have been implemented in some organizations they work for. However, these systems are often interrupted and lack continuity. Physicians considered the discontinuity of established mHealth systems as one of the key barriers to the effective use of mHealth tools.

> *Around two years ago, this program was being popularized, but nobody in the middle was using it. It is not waited until we become experts. It was interrupted in the middle.* [JD5]
>
> *Sometimes, due to the data and various factors, the system encounters some errors. So we will go back to the manual paperwork at that time. Therefore, when we report using a digital system, if there is an interruption, we often go back to paperwork.* [HD7]

Another barrier preventing the use of mHealth systems was the required extra time amidst an overloaded work schedule, especially for recommending such systems to patients.

> *And again, just because there is high patient load here; I don't have that kind of time. It does not allow to say use this or not. As one patient leaves, another enters. There is a load, and it doesn't allow us to have much contact with the patient. We are going to send by diagnosing and treating with what we already have.* [AD1]

A further barrier to the adoption of mHealth systems was the lack of physicians’ awareness about mHealth systems for patients. Some physicians reported that they are not well-informed about these systems, and even for their own personal health and wellness, they rely on browsing the internet rather than using stand-alone health systems. This lack of knowledge and personal use can limit their ability to confidently recommend and implement mHealth tools in their practice, which in turn can hinder patient adoption and engagement. To overcome this barrier, there is a need for greater education and training programs for physicians on the benefits and effective use of mHealth systems, as well as efforts to increase physician

XSL•FO
RenderX

engagement with these tools to help build their confidence and understanding.

> *Now, for example, I don't know anything personally, but for example, I don't know an app for diet. Honestly, now, in terms of exercise, I use YouTube downloads even for myself.* [AD2]

Some physicians claimed that they have never given these systems any consideration.

> *From my point of view, there were no situations where we would recommend using this to my patients. Maybe, I have never thought about it.* [JD4]

> *I don't have the knowledge. I mean, I didn't know there was anything like that about mHealth that you just told me about. To tell you the truth, I only found out about Mobile Health Solution today.* [AD5]

Another barrier was physicians' concerns about patients' mHealth capability. They acknowledged being worried that their patients do not have the necessary skills or access to technology to make the most of mHealth tools.

> *But as I said, it's not that the applications have problems; As I told you before, I don't think anyone will use it. It just doesn't feel like that to me.* [AD1]

> *Most of them are illiterate. It is uneducated and I don't think they will use this app.* [AD5]

> *As I mentioned earlier, the problem is that the demographics of the patients are too old to use such systems. They may have a little trouble.* [AD6]

> *What I would consider being the biggest obstacle... Most of the chronic follow-ups have a low level of education. So they may find it very difficult to use applications.* [AD7]

> *I think they are less literate and less qualified. Because the majority cannot read a book. So I think that will be like a gap.* [JD2]

Physicians mentioned a further obstacle in the adoption of mHealth systems, namely, lack of trustworthy systems, which raises concerns about the potential risks of recommending them to patients.

> *People with chronic conditions are very bitter, so they search for something to escape. Sometimes there is information that takes you to stop taking medicine. So, this is the side effect of modern digital. It's not even called a limitation.* [JD4]

Another barrier was the lack of readily available mHealth systems in the local context.

> *But it is not prepared at the app level in our country. It is better if we use this application prepared like this. We tell patients that they will find it on Google because it is not set up like that.* [AD7]

> *The first and most important thing that we can't do is we see it from our country's perspective, if we see it like Ethiopia, there are no rich digital applications for that. Not being able to provide us with that information is the number one factor.* [HD7]

> *Most of the time, because those data are based on foreign populations, those findings have nothing to do with the population of our country.* [HD1]

Security and data confidentiality concerns were also perceived as barriers to the use of mHealth systems.

> *I usually log in with my Google account. I have already accepted my Google account. And I have doubts. I am not sure. As I don't have details about the system. I don't know much about server security. Some app says you can accept or not about personal issue when you open an app. It may be exposed. I'm not sure about the data security and confidentiality.* [HD4]

Another barrier was the inconvenient features of some mHealth systems. Most mHealth systems that physicians can use do not have an offline option, and many of these systems work online.

> *The biggest problem is that they are all online. They don't have offline. And you absolutely need data.* [HD4]

A lack of reliable internet and electricity was an additional obstacle.

> *After all, if there is no internet, we cannot get this information. The main obstacle is the internet. Most mHealth applications are internet-based, so information cannot be accessed when the internet is down.* [HD7]

Enablers

Among the 27 enablers that facilitate the adoption of mHealth systems, 14 were under the organizational policy category. These factors pertain to health care organizations' internal workings, workflow, policies, regulations, patient-related factors, and user engagement. Moreover, 4 enablers were under the personal and social category involving personal characteristics, cultural factors, and moderating factors. Furthermore, 16 enablers were under the technical and material category encompassing mHealth system design, IT capability, compatibility, user experience, data-related factors, ease of use, and monetary aspects. Physicians' perspectives on factors that facilitate the use of mHealth systems are depicted in Figure 3.

XSL•FO
RenderX

Figure 3. Enablers for using mobile health (mHealth) systems: physicians' perspectives.

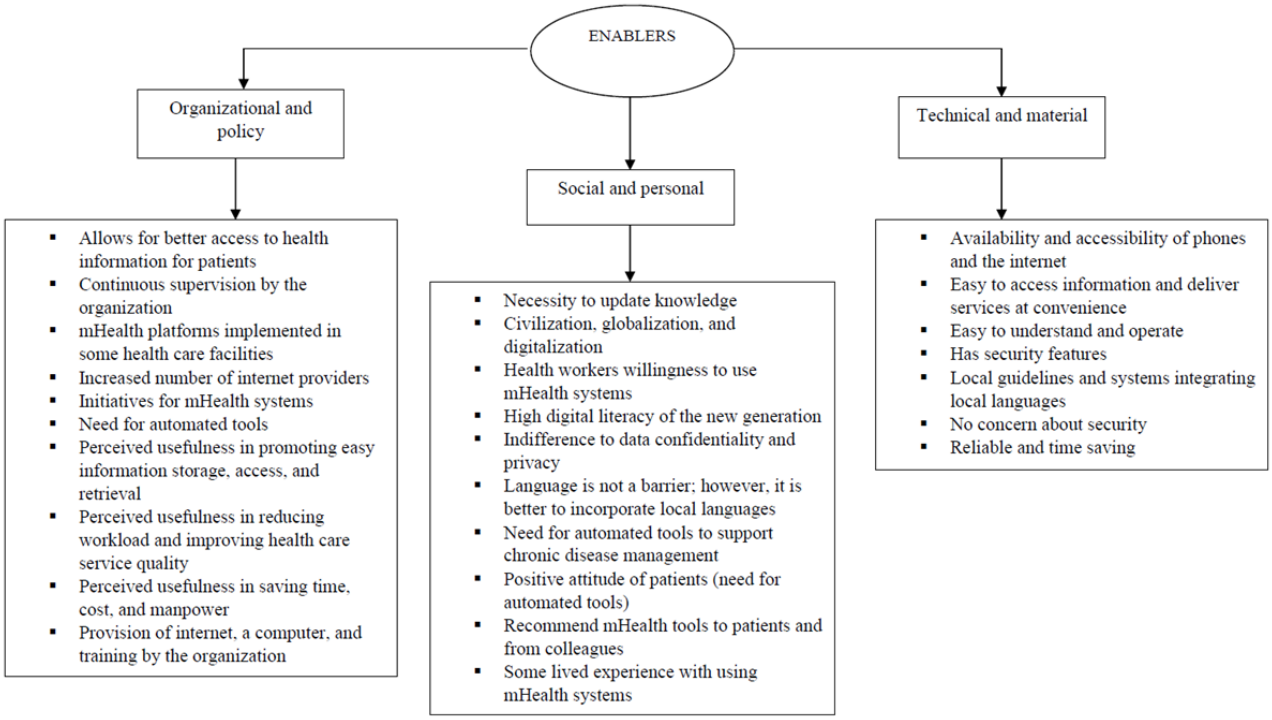

Physicians perceived usefulness in promoting easy information storage, access, and retrieval as one of the enablers of the use of mHealth systems. Despite limited exposure to these systems, physicians believed that they represent a convenient way to manage patient data, with the advantages of avoiding loss of information and enabling reliable information retrieval.

> *On the positive side, there is no such thing as loss of information such as manual cards and paper, papers can be torn and lost. Patients can also be abused. But if we use mHealth systems such as EMR, it will not disappear because we will simply be here and send all the investigations. Whether the data is from this year or before, I can check many things about the patient. It makes it easier for me because I have everything I need there.* [AD1]

> *You will not lose the file, it is well recorded. What did they have before? Or what medication were they taking? Even if the patients don't know what medicine they are taking, how much dose they are prescribed, how many times a day they are taking, so I can get a record from this.* [AD2]

> *Therefore, getting reliable information and then delivering that information to the concerned patient is of great importance. Using mHealth in general allows us to get better information and updated information.* [HD7]

> *You will find information in applications in a very easy and understandable manner. There are books in hard copy. Sometimes when you read those hard copies, you don't understand them, when you go with those applications, they are very encouraging because they are easy to understand.* [HD3]

> *Knowledge is not a big deal as it used to be. Knowledge is cheap. Knowledge is what you can find anywhere by Googling. It is a matter of reading and not reading. But everything is up in the air. And it's not like you used to go to the library to look for a book and get a book. Everything is on your phone. You can open and use it even during operation. Whenever you want. It's very easy.* [HD4]

Another enabler was the perceived usefulness in reducing workload and improving health care quality. According to physicians, using mHealth systems has the potential to lessen the strain of health care personnel while also raising the standard of care.

> *The first reason for not providing quality service is the workload and if something reduces that load, if there are supportive things like this, I think the quality will improve.* [AD1]

> *The timing itself. If you have a current patient, what time did she take her medicine? Did she take it or not? She will be evaluated strictly on time. She takes her medicine on time. A professional can't lie if he skips it knowingly or unknowingly. He cannot say that he gave it to her. When you record what you just did, it will record the time.* [AD4]

> *You can treat your patient in an international standard way.* [HD4]

Physicians mentioned that mHealth systems allow for better access to health information for patients. Some believed that providing patients with access to various health-related information at their fingertips will empower them.

> *Using mHealth in general allows to get better information and updated information. Especially when*

XSL•FO
RenderX

> *it comes to health. It helps to live a healthy life by getting new information every day.* [HD7]
>
> *If the patients have it on their hands, I think it will be a daily experience for them too. It will be easily accessible and it will save our time for the patients as well.* [HD2]
>
> *Our patients, who have been treated, do not come back with any complications. Because they can get health information and monitor theirs case. Because they may be able to get every detail on the system.* [HD6]

Another enabler was the need for a chronic disease management support tool. Physicians recognized the potential benefits of digital automated tools for patients having chronic diseases, as consultations and visits to their office alone may not suffice. To promote healthy lifestyle changes and provide necessary support, it is imperative to equip patients with user-friendly tools that can be integrated into their daily routine.

> *First, when there is a patient like this... there are chronic ones, for example, diabetic, hypertensive, heart failure. Because they need not only medication but also life modification, they may not achieve what we have told them in one day. But everything how to modify their lifestyle detail is there, so when they get access, they can easily remember what we told them and continue their life. Beyond medication, by the way, also cares about their lifestyle.* [HD6]

Availability of phones and the internet was also perceived as a facilitator. According to physicians, availability of phones and the internet has tremendously facilitated the adoption of mHealth systems in a variety of ways. First, regardless of location or the time of day, users may readily access health information and services via their mobile phones. Second, health professionals can use digital tools to gather, preserve, and share patient data, which can improve the quality and efficiency of health care services. The increasing use of mobile phones and the internet in Sub-Saharan Africa has created an opportunity to use these platforms to deliver health information and services to individuals who may have limited access to traditional health care services.

> *This is because there is internet access and broadband, so everyone works connected to the network. Without that, this simple system wouldn't exist.* [AD1]
>
> *Now, for medical and other purposes, because of this technology, because of the advent of smartphones, the spread of Wi-Fi, the advent of the Internet, I think that it has made everything easier even the medical education.* [HD4]
>
> *Having a mobile phone and the internet makes things easier.* [HD5]

Physicians who have experienced mHealth systems recognized that these systems are easy to understand and operate. They declared that for users who are familiar with technology, these systems are easy to use.

> *So you need to know how to access the systems. They are mostly easy. There is nothing difficult unless you are someone who is not familiar with technology to use it. You can search. You can log in and access what you want. I don't have anyone who says it's a challenge.* [AD6]

Perspectives of Health Care Executives

Among health care executives, we identified 65 factors that were perceived to influence the adoption of mHealth systems. Of these 65 factors, 31 were classified as barriers and the remaining 34 were classified as enablers. Further categorization of these aspects in the study resulted in the identification of 3 major categories: organizational and policy, social and personal, and technical and material. Health care executives were assigned a code based on the study area as follows: those from Addis Ababa were coded “AA,” those from Harari were coded “HA,” and those from Jimma were coded “JA.”

Barriers

Among the 31 barriers, 14 were under the organizational and policy category involving elements like the internal working environment of health care organizations, workflow-related challenges, policies and regulations, patient-related factors, and user engagement issues. Moreover, 6 barriers were under the social and personal category involving aspects like social and cultural norms, personal traits, and moderating factors. Furthermore, 11 barriers were under the technical and material category involving difficulties with usefulness, IT capability, compatibility, data-related factors, user experiences, financial factors, and ease of use of mHealth systems. Health care executives’ viewpoints on obstacles to the use of mHealth systems are presented in Figure 4.

XSL•FO
RenderX

Figure 4. Barriers for using mobile health (mHealth) systems: health care executives' perspectives.

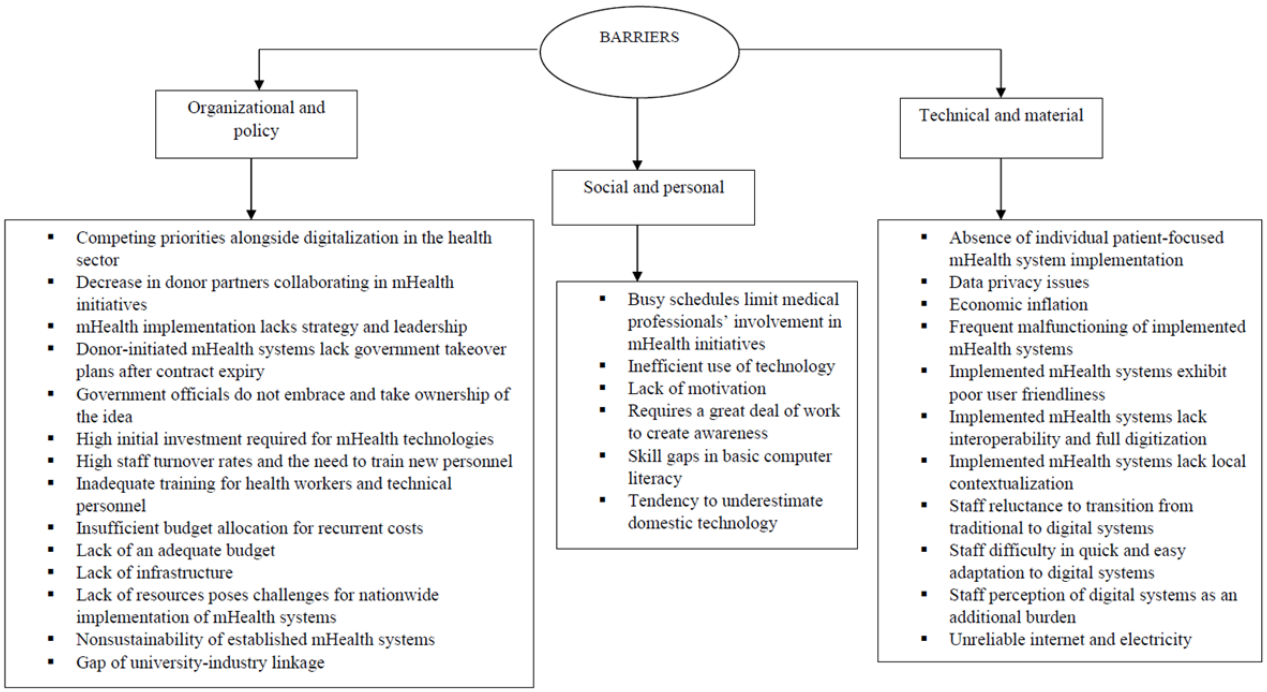

From the perspectives of executives, one of the challenges of implementing mHealth systems in Sub-Saharan Africa is the availability of competing priorities alongside digitalization in the health sector. There are competing demands for resources and attention, such as investing in health care infrastructure, improving health care workforce capacity, and addressing immediate health needs.

> *We are facing multiple health challenges, including infectious diseases as well as non-communicable diseases like diabetes and hypertension. Additionally, the health sector faces critical issues such as inadequate infrastructure, a shortage of skilled healthcare workers, and limited financial resources. Because there are competing priority areas in the sector that require money. From all of them, it is assigned to this sector and these are taken as a challenge not to work completely with hands.* [AA5]

High initial investment required for mHealth technologies was also perceived as a barrier by executives. The cost of implementing mHealth systems includes not only the purchase of hardware and software, but also the cost of training health care workers, adapting existing health care workflows, and establishing new data management and privacy protocols.

> *As a problem in supply, mHealth technology is expensive. Not being able to avail of all of them is an issue.* [HA3]
>
> *These start-ups require financing. Although the return is high, the initial investment is also high. mHealth activities require so much investment.* [JA2]

Another barrier was that some donor-initiated mHealth systems previously implemented in health care facilities lack government takeover plans after contract expiry. These mHealth systems involving donor support do not have clear take over and exit strategies for ensuring sustainability of the projects.

> *A project has a lifecycle. It comes up with a project by donors and be started that way.. It won't be owned when it is done. It will not be owned by the government.* [AA34]

Skill gap in basic computer literacy was also perceived as a barrier. These skill gaps can be partly attributed to limited exposure or access to digital devices such as computers.

> *Computer access is limited. The smaller the access, the less likely it is to use mHealth activities. So there is an obvious skill gap. So does the professional.* [JA2]

Another barrier was inefficient use of technology.

> *Efficiently manipulating those technologies if we had them; Proper use of data; One of the challenges is not using technologies properly. We should use them properly and take the appropriate value from them.* [AA5]

The absence of individual patient–focused mHealth system implementation was perceived as another barrier to the adoption and implementation of mHealth systems in health care.

While some health care facilities have implemented mHealth initiatives, these initiatives are more focused on the institution rather than the individual patient.

> *The one that we are using on ART is not to be used by the individual patients. It is an application more with professionals on the periphery than with patients.* [JA2]

A further barrier was that implemented mHealth systems lack interoperability and full digitalization. Although attempts to implement mHealth are sparse in health care facilities, lessons can be borrowed from implementation challenges in other areas, such as electronic medical records (EMRs). Some facilities have

XSL•FO
RenderX

deployed EMR systems. EMR systems are not fully digitized, with most of them involving a combination of manual and digital systems. Some health care executives also reported that due to mismatch or lack of integration between some medical devices, they are unable to deploy fully digital systems.

> *There is a lot of fragmented stuff.* [AA34]
>
> *It is half and half... It is only for reporting, but to make our work more active, it is very good if the facility is fully digital and the delivery points pass through the system through the network.* [HA1]
>
> *Interoperability is not yet implemented. The two systems are not interoperable. The interoperability is being tested for the country.* [JA1]
>
> *We have implemented about 80% of the Laboratory Information System (LIS) in our hospital.* [JA2]

Another barrier was the difficulty of staff to quickly and easily adapt to mHealth systems.

> *First of all, sometimes when new things come, there are problems of getting used to that digital thing quickly and not keeping up with it.* [JA1]

Another barrier was the lack of awareness among staff members. Going fully digital requires a great deal of work to create awareness, as mHealth technologies are relatively new in the region and many health care workers may not be familiar with them.

> *As a challenge, we still need a lot of attention for technology in general by the government. We are still in the early stages of our own user experience.* [AA2]
>
> *There is an opportunity created by technology. However, creating awareness of the work at all levels, from the leadership to health institutions or to the community should be done. Therefore, one of the biggest tasks is to make the awareness. From top to bottom, the importance needs to be well inculcated. This is the lack of awareness of the decision makers from the community level to the top level.* [AA5]

The above issue is further exacerbated by the fact that the barrier is related with staff's reluctance to transition from traditional to digital systems. This lack of awareness and knowledge can lead to resistance or hesitancy in adopting mHealth systems, even if they have the potential to improve patient outcomes and operational efficiency. Health care staff may be unsure of how to use mHealth tools and systems, or they may not understand the benefits of using them.

> *People who are used to paper work often don't want to be told that we are going to do digital work.* [JA2]
>
> *The biggest thing I have seen is that there is a problem of commitment. It means that it is very challenging for a person to leave what he used to and come to something new.* [JA3]

Enablers

Among the 34 factors that facilitate the adoption of mHealth systems, 14 were under the organizational and policy category. These factors pertain to health care organizations' internal workings, workflow, policies, regulations, patient-related factors, and user engagement. Moreover, 4 enablers were under the personal and social category involving personal characteristics, cultural factors, and moderating factors. Additionally, 16 enablers were under the technical and material category encompassing mHealth system design, IT capability, compatibility, user experiences, data-related factors, ease of use, and monetary aspects. Health care executives' viewpoints on enablers of the use of mHealth systems are presented in Figure 5.

XSL•FO
RenderX

Figure 5. Enablers for using mobile health (mHealth) systems: health care executives' perspectives.

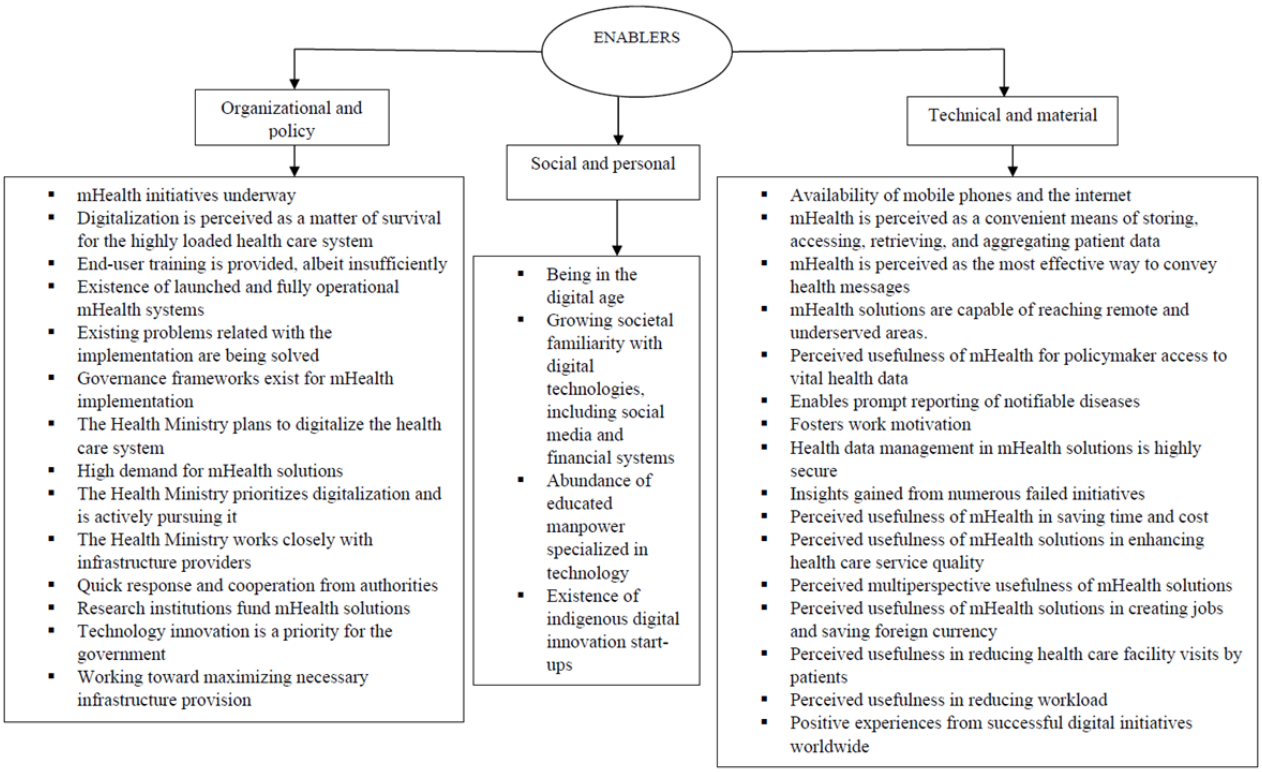

From the perspectives of executives, the health ministry's plan to digitalize the health care system is one of the main facilitators or enablers of the use of mHealth systems.

> *There is a direction according to the government. To use such technologies in most cases; It is for online use. The ministry itself has a strategic plan. To digitalize the healthcare system. It is a priority.* [AA1]

The executives emphasized that the plan is not just a matter of execution but also of giving priority and actively pursuing it.

> *And now technology is making a big contribution to the health sector. And the ministry understands what a great advantage we have; Departments in the Ministry of Health are working with great initiative using mobile technology.* [AA5]

Another factor that enables the implementation of mHealth systems is the existence of governance frameworks. Such frameworks were nonexistent a few years ago. However, now, in order to regulate and promote the adoption of mHealth systems, the health ministry has developed a framework that is currently being deployed. Furthermore, a dedicated directorate office has been established to oversee and manage the adoption and implementation of mHealth systems.

> *Therefore, it is mandatory to create favorable conditions. Facilitating conditions on the ground such as governance frameworks are required. We are working on it. Standards guidelines are required. In addition to this, data standards have also been developed. Now, for example, there are disease codes. A standardized health data access policy has been developed. All these guides the mobile health and govern mHealth as a whole.* [AA5]

> *Now, for example, during the time of Covid, There were mandatory situations to initiate telehealth teleconsultation, but if you ask how it works, there was nothing. But now, there is this direction at the level of the Ministry of Health. They have a directorate office.* [JA2]

Another enabler is the recognition that digitalization is crucial for the survival of the heavily burdened health care system in Sub-Saharan Africa. Hospital executives strongly believed that the failure to digitalize the entire health care system within the facility will ultimately lead to its downfall.

> *But now, considering the complexity of the treatment and the number of patients we see, the staff is now demanding digital solutions. It is understood that we cannot survive if we do not digitize. We have to digitize the entire operation of the hospital. As I said before, it is a matter of survival. It is a hospital where three thousand four thousand people come and go every day. We have 250,000 to 300,000 outpatients per year. We treat about 20000 people. About 170,000 to 180,000 people will come for emergency. We perform surgery on 15000 people. So this process it is huge process. If this process is not digitized, we cannot provide a smooth service. Because it is causing problems for the patient. As a hospital, we have an assessment that it is increasing the number of deaths.* [JA2]

According to the executives, another enabler is efforts made toward maximizing the necessary infrastructure provision required for the implementation of mHealth systems. Executives at all levels within the organization realized the importance of providing the required infrastructure, such as ICT and hardware

XSL•FO
RenderX

or software resources, to support the deployment of mHealth systems.

> *In terms of infrastructure... the main thing for digitalization is the internet. They have been given internet access in all nine districts. In terms of computers, they have computers. All adequate internet access to the DHIS health information system, capable computers, and PCs are available at all health centers in the district.* [HA2]

> *For example, they can be computers. Servers. It is not directly from the office but it is fulfilled by different partners as they work together with us.* [HA3]

> *Along with training, tablets are distributed along with resources such as tablets in areas where they are needed.* [JA1]

> *We have invested a lot in infrastructure. Especially our hospital as a health institute has a large ICT infrastructure. We deployed computers. We are working on servers and databases. We trained people and experts for that purpose... It is itself strategically supported and supported by infrastructure.* [JA2]

Another enabler is the growing societal familiarity with digital technologies, including social media and financial systems. As more individuals become comfortable with using digital technologies in their daily lives, they may also be more open to using mHealth systems.

> *At the same time, the opportunity is also very profitable because the number of users of social media is increasing from time to time.* [HA3]

> *Mobile is now used by many people for money transfer. A lot of things... even if a mHealth system doesn't expand, in the finances... a lot... you sit at home and pay. You can withdraw without an ATM machine. It's simple, a person in finance has experience elsewhere. Mobile is now used by many people for money transfers. It is one exposure to general digital systems.* [JA2]

The abundance of educated manpower specialized in technology was also perceived as an enabling factor for the adoption of mHealth systems in Sub-Saharan Africa. With a growing pool of technology professionals in the region, there is a greater capacity to develop and implement mHealth solutions that are tailored to the specific needs of the region. This includes the development of innovative technologies, such as mHealth systems and telemedicine platforms, which can improve access to health care services and support patient care. In addition, the presence of a skilled technology workforce can support the deployment, maintenance, and troubleshooting of mHealth systems, which can improve the overall reliability and sustainability of these systems.

> *… there are many educated young people who can do a lot of creative work; Universities have expanded significantly.* [AA2]

> *… Because educational institutions at the national level can easily produce graduates who can easily design and develop these emerging technology applications.* [AA5]

> *We also have an ICT director with us. We have teams working on software development. So, as a university, as a health institute, as a medical center, we are doing this work. Second, we are building human capacity. Many students are studying. Despite the quality, I think we have that capacity human power.* [JA2]

Executives perceived that mHealth solutions are capable of reaching remote and underserved areas where traditional health care services are not readily available, which is a significant enabler for the adoption of mHealth systems in Sub-Saharan Africa. Furthermore, mHealth solutions were thought to aid in overcoming the scarcity of health care personnel by providing remote consultation, training, and education to health care workers, as well as assisting in the delivery of health care services in places with limited resources.

> *In terms of accessibility, the technology is generally accessible to all communities. The community of our country is located in a very remote place and access to health facilities is very limited. Therefore, mobile technology can be used to make these sections of society that are accessible at different distances easily accessible. Secondly, as mentioned, we have a shortage of health professionals. It is very important to make the service available remotely even if our experts are very limited. From this perspective, it is very important for a society like Ethiopia where there is a shortage of health professionals and the most accessible health facilities are few.* [AA5]

Executives perceived mHealth solutions as useful tools for enhancing health care service quality, which is a significant enabler for the adoption of mHealth systems in the region. Health care providers can improve the quality and efficiency of health care services by embracing mHealth technologies, such as electronic health records, clinical decision support systems, and telemedicine platforms. Real-time data collection, monitoring, and analysis enabled by mHealth technologies can also influence evidence-based decision-making and assist health care providers in providing fast and accurate diagnoses and treatment plans.

> *For example, the patient can take advice without coming to the facility. Alerts can go be sent to him. You will find a lot of information. It means that they will not be abused by looking for their doctors. And it generally makes things easier.* [AA1]

> *It's easy. It's very simple. It makes it easier. It will be easy for everything in time and for the number of people. So it is very good if we use and serve. It is from two sides. Client side and provider side. It also reduces other errors. Another thing is that it reduces time. So it's very good.* [HA1]

> *It will help us gain efficiency. As mentioned earlier, efficiency gains for the health care system are linked to access, mortality, and quality.* [JA2]

Another enabler perceived by executives is the availability of mobile phones and the internet. The increasing availability of

XSL•FO
RenderX

mobile phones and the internet in Sub-Saharan Africa has created an opportunity to use these platforms to deliver health information and services to individuals who may have limited access to traditional health care services.

> *The fact that the majority of the society is using mobile technology gives the ministry a great opportunity to implement the system. Second, mobile accessibility is expanding. In Ethiopia, it is said that quite a large number of people have mobile phones in their hands. Therefore, the service can be accessed without coming to the health facility.* [AA5]

> *For example, if you visit the OPD in a hospital, you will find very few people without a smartphone. If you want to buy a laptop and buy a tablet, It is easy. It is available.* [JA2]

The overlapping barriers and enablers among the various stakeholders are summarized in Tables 2 and 3, respectively.

Table 2. Overlapping barriers among the various stakeholders.

Barrier	Reported by patients	Reported by physicians	Reported by health care executives
Lack of readily available or established mHealth[a] systems	Yes	Yes	No
Security concern	Yes	Yes	Yes
Unavailability of trustworthy mHealth systems	Yes	Yes	No
Inconvenient features (presentation) of mHealth systems	Yes	Yes	No
Lack of awareness about mHealth systems	Yes	Yes	No
Digital illiteracy	Yes	Yes	No
Lack of continuity of established mHealth platforms	No	Yes	Yes
Lack of infrastructure	No	Yes	Yes
Lack or limited provision of training	No	Yes	Yes
Lack of available mHealth systems in a local context	No	Yes	Yes
Unreliable electricity and internet connection	No	Yes	Yes

[a]mHealth: mobile health.

Table 3. Overlapping enablers among the various stakeholders.

Enabler	Reported by patients	Reported by physicians	Reported by health care executives
Indifference to data confidentiality and privacy	Yes	Yes	No
Internet and mobile phone availability	Yes	Yes	Yes
Need for easy and automated tools	Yes	Yes	No
Perceived usefulness in promoting easy information storage, access, and retrieval	No	Yes	Yes
Perceived usefulness of mHealth[a] systems in saving time, cost, and manpower	No	Yes	Yes
Perceived usefulness of mHealth systems in improving health care service quality	No	Yes	Yes

[a]mHealth: mobile health.

Discussion

Principal Findings

The barriers and enablers of the implementation of mHealth solutions in Sub-Saharan Africa were identified and explored in detail in this study. All interviewees (patients, physicians, and health care executives) recognized the potential benefits of mHealth. Several overlapping barriers and enablers were identified among the 3 participant groups.

Lack of awareness about mHealth solutions was highlighted as a common barrier to implementing mHealth systems by patients and physicians. They claimed that it was preventing them from experiencing and using mHealth services. They emphasized the importance of raising awareness among stakeholders. Developers and advocates of mHealth solutions must give outreach and education efforts top priority in order to overcome these issues and increase public knowledge of the potential advantages of these solutions. This might entail collaborations with health care institutions, neighborhood-based outreach initiatives, and population-specific marketing plans. Additionally, in order for health care professionals to use mHealth solutions efficiently and explain the advantages to their patients, they need to receive proper training and support.

XSL•FO
RenderX

Finally, overcoming cultural barriers will call for sensitivity to the distinctive cultural perspectives and beliefs of various populations, as well as the customization of outreach and education initiatives [11,34].

Another overlapping challenge noted by patients and physicians is digital illiteracy, which refers to a lack of competency or expertise in using digital technologies, such as computers, the internet, and other digital devices or tools. The good news is that an improving literacy level in the society was also viewed as a facilitator by patients and clinicians. According to physicians, the younger generation's strong digital literacy level is a tremendous opportunity to leverage and apply mHealth systems. The perspective of executives that supports this view is their observation that society is becoming increasingly accustomed to digital technologies, including social media and financial applications. A study involving cancer survivors reported that low digital literacy may hinder information acquisition and technology-enabled cancer care. The study recommended that digital interventions should be adaptable to varying levels of digital health literacy. Policymakers in health care should acknowledge digital disparities and create targeted initiatives to narrow the digital divide while also meeting the pressing demand for the digitization of health care services [25]. Using digital health information resources and engaging in digital interactions with health care providers offer significant advantages, holding the potential to enhance the efficiency, quality, and accessibility of health care systems, all while empowering patients [26,35].

Another hindrance mentioned is the lack of incorporation of local languages and contextual factors, such as demography, culture, and population, in mHealth systems. Physicians reported that while they are willing to use mHealth systems developed in foreign languages, they prefer those available in the local language. In contrast, all patients interviewed expressed a preference for mHealth systems that incorporate one or more local languages. Studies have also reported the positive impact of user-centric design and local contextualization of mHealth approaches for improved uptake [36,37].

The challenge in engaging patients for accessing diverse patient populations for education or engagement remains an issue. The study by Martin [38] described barriers for patient engagement, including literacy, access to hard-wired technologies, and understanding of an increasingly complex network of medical care. Developers of mHealth systems need to concentrate more on patient-centered design, involve users in the development process, and work to deliver a customized user experience in order to overcome these difficulties. This can make mHealth systems more usable, effective, and accessible for patients, which will ultimately improve patient outcomes. Additionally, by educating patients on the use of mHealth systems, offering support for their use, and highlighting the advantages of these systems for patient care, health care organizations and providers can encourage patients to use them.

Aside from not being aware of mHealth systems, those who have been exposed to the technology are concerned about the availability of trustworthy systems. Because health is a sensitive issue, patients mentioned that if doctors recommend it during their follow-up, they could use it. However, the majority of doctors said that they had never recommended such systems to their patients. This is due to 2 major reasons. First, physicians have little expertise with mHealth systems, which limits their capacity to use them. Second, because of a lack of understanding of patient-focused mHealth systems, physicians are often unaware of mHealth systems that they could recommend to their patients. According to recent research conducted in developed countries, the findings indicate that health care professionals do not endorse the use of mHealth systems to their patients [39]. Nonetheless, the aforementioned study revealed that while health care professionals do not recommend mHealth systems to their patients, they do inform them about the existence of these systems should they express a desire to use them. In contrast to the notion that health care professionals solely remind their patients about the availability of mHealth systems, the findings of this study suggest that physicians harbor a degree of skepticism regarding the ability of their patients, particularly those who are elderly or possess limited education, to effectively use such technologies. Consequently, these physicians generally do not recommend mHealth systems to individuals having chronic diseases. The other reason, as perceived by physicians, is that they could not find trustworthy health care systems to prescribe to patients confidently. The lack of adequate regulation and control is one of the key reasons for the unreliability of mHealth systems. Numerous systems make efficacy claims without any supporting data from the scientific community [21,40]. This may cause consumers to rely on incorrect or partial information, which could be harmful to their health.

Even though security issues were mentioned as common barriers by all the stakeholders in the study, some respondents also expressed different beliefs regarding data confidentiality and security. Some of the participants in both patient and physician groups were not worried about security. Some believed that the system they are using has security features, and some mentioned that they are not worried at all whether the system is secured. However, some patients mentioned that they are not sure whether data confidentiality will be maintained during the use of such systems and said that they are skeptical about it. They recommended better security features while designing such systems. The potential for data privacy violations is another problem. Some systems collect users' private health information but fail to adequately safeguard it from unauthorized access by outside parties. Users may experience serious repercussions as a result, such as discrimination and identity theft. It is crucial to set precise standards and guidelines for mHealth systems in order to address these problems. Researchers have also suggested ways to underpin a radical rethinking of information privacy, confidentiality, security, and integrity to unlock the potential of mHealth and ensure verified access to often sensitive data [41]. Given the rapid pace of technological development, the protection of personal health information stored in mHealth solutions is an important consideration. In order to protect people's privacy, it is essential to ensure the confidentiality of such data. This is especially important in light of laws like the Global Data Protection Regulation (GDPR) [42,43].

XSL•FO
RenderX

All patients, physicians, and health care executives perceived the widespread availability of the internet, mobile phones, and digital devices as creating a favorable opportunity to integrate mHealth systems. Though the Sub-Saharan African region is way behind other regions in employing mHealth activities, there is now a high level of commitment and a strong strategic plan to benefit from this opportunity. Furthermore, there is a high need for an individualized patient-focused mHealth system for managing chronic conditions such as diabetes and hypertension. A study conducted by Doyle et al [27] concluded that implementing mobile phone–based interventions is viable; however, there is a risk of exacerbating inequities, particularly if these interventions necessitate internet access. Internet-based mHealth approaches should carefully assess potential risks for participants and include skill-building sessions on secure internet and phone usage.

The Sub-Saharan African region's population is growing at an alarming rate, and the region is facing the double burden of communicable and noncommunicable diseases. As a result of these, access to high-quality health care systems may be jeopardized. To make problems worse, social distancing necessitated by the COVID-19 outbreak made face-to-face consulting and care-taking difficult in many cases. According to health care executives, integrating digitalization in health care systems is becoming increasingly important for the health care sector, which is under high pressure. Sub-Saharan Africa faces numerous health care challenges, including limited access to health care services, inadequate health care infrastructure, and a shortage of health care professionals. The COVID-19 pandemic has also highlighted the need for mHealth solutions that can help provide care remotely [44].

Digitalization can help address these challenges by improving access to health care services, enabling remote consultation and telemedicine, and improving the accuracy and completeness of patient records. This can help reduce the burden on health care facilities and make it easier for patients to access care, especially in remote or underserved areas. Physicians believed that employing a mHealth system can help to reduce workload and also help to achieve quality health care services. Other scholarly works have also affirmed that digital technologies contribute to improved efficiency and the streamlining of health care services. The potential of mHealth to provide health information to patients and the accessibility of patient data through EHRs make the situation easier for health care practitioners, reducing administrative tasks and enhancing care coordination [10,45,46].

According to patients, a lack of easily available mHealth systems is also a hindrance. This viewpoint was also shared by both physicians and executives. Physicians reported some experience in using mHealth systems such as reference guidelines, gestational age calculators, etc. These mHealth systems, however, lack local context. Furthermore, physicians reported that they were unable to find standalone systems in the form of apps that would allow them easy access to vital information to update their knowledge and fill knowledge gaps. Despite the fact that the integration of mHealth in health care facilities is not widespread, it is possible to draw on the experiences and obstacles encountered during the implementation of other technologies, such as EMR systems. EMR systems deployed in some health facilities have a lot of challenges for use to the full extent possible. However, owing to discontinuity and nonmaturity of the systems, they are not fully functional and some of them have stopped working. This is also perceived as another hindering factor. Despite efforts to launch mHealth systems, health care executives indicated that most launched mHealth systems so far are for health care personnel. Governments, health care organizations, and technology firms must collaborate to develop infrastructure, raise money, lower regulatory barriers, and encourage user adoption of mHealth systems in order to address these issues. By doing this, we can guarantee that mHealth systems are made more accessible to individuals around the world and contribute to improving access to health care services. A body of literature indicated that health care providers face challenges in improving digital health applications, and collaborating with stakeholders for value creation remains a significant obstacle. Despite these challenges, involving stakeholders and addressing their needs could promote the sustainable development of digital health services [47].

According to reports from executive stakeholders, many attempts were made to implement mHealth systems in various health care facilities, but most of them were unsuccessful due to a range of issues such as inadequate government ownership, insufficient budget for recurring operational costs, lack of training, insufficient infrastructure, and lack of interoperability. A recent systematic review of systematic reviews indicated that infrastructure, lack of equipment, and technology gaps together accounted for barriers to the use of mHealth systems in developing countries [48]. Despite these challenges, there were several valuable lessons, such as the need for allocating a sufficient operational budget, proper takeover of projects, and engaging well-matured systems through learning from failed projects. As a result, there is now a renewed focus on digitalization in health care, and mHealth solutions are viewed as valuable tools for addressing complex health care challenges from multiple angles.

A multitude of analogous studies have examined the obstacles hindering the adoption of mHealth systems within health care facilities. Many studies reported lack of knowledge of mHealth systems, infrastructure, lack of equipment, and technology gaps as barriers, and identified the ubiquity of smartphones and apps as a facilitator, especially for the younger generation [39,48]. Absence of a national policy on mHealth, poor internet connectivity, and shortage of electricity were also highlighted as important inhibiting factors for mHealth adoption in low and middle income countries [49]. The perceived usefulness of mHealth approaches among patients has been reported [50]. Some of the enablers identified in this study were introduced by previous studies, such as perceived usefulness in improving health service quality and perceived ease of use [51].

One of the common barriers as perceived by all the stakeholders is security concerns. The results of this study are consistent with the findings of other studies reporting that mHealth app users have security and privacy concerns. The study reported that lack of security features in mHealth apps was a barrier for adoption [52,53].

XSL•FO
RenderX

In the literature, knowledge and limited literacy were presented as barriers for mHealth adoption. However, in this study, an increasing literacy level of the society was identified as one of the enablers. This same factor was reported as a barrier in another similar study [53]. Consistent with the results in this study, a prior study reported that limited digital literacy and the unreliability of health information from mHealth platforms are barriers for mHealth use [25].

Developing mHealth solutions that are suited to the local environment, raising awareness, offering adequate training, assigning adequate funding, incorporating various security features, and putting in place and implementing simple governance principles are a few ways to tackle the challenges.

This study is the first of its kind to provide a comprehensive exploration of the barriers and enablers of the use of mHealth systems in Sub-Saharan Africa, with a focus on the multi-level and multi-actor perspectives of patients, physicians, and health care executives. The barriers identified in this study highlight the challenges and limitations that must be overcome in order to successfully implement mHealth systems in Sub-Saharan Africa. Understanding these barriers can help inform the design and implementation of mHealth solutions that are tailored to the specific needs and context of the region. Addressing these challenges necessitates increased investment in mHealth infrastructure, health care worker training programs, and financial sustainability methods for mHealth initiatives. If these barriers are effectively addressed, it may become possible to overcome the challenges associated with implementing mHealth systems in the region, thereby unlocking the full potential of mHealth to enhance health care outcomes for patients, health care providers, and policy makers. The identified enablers will be further investigated and considered in the design of future mHealth platforms for studies in Sub-Saharan Africa.

Limitations of the Study

This study is subject to certain limitations, notably the relatively modest sample size, which may restrict the extent to which the findings can be extrapolated to the broader population. Nonetheless, efforts were made to address this concern by selectively recruiting participants from diverse age cohorts and employment backgrounds. Other limitations include a small patient population (the disease may have influenced the patients' perceptions). Future investigations could benefit from expanding the sample size to enhance the representativeness of the outcomes. Another potential limitation of this study is the absence of triangulation of results. However, steps were taken to mitigate this shortcoming by collecting sufficient data to comprehensively comprehend the phenomenon, thereby enhancing the validity and reliability of the findings.

Conclusion

The use of mHealth systems in Sub-Saharan Africa has been hindered by a range of factors and has also been facilitated by various enabling factors. Patients, physicians, and health care executives identified common barriers and enablers to the uptake of mHealth systems. The identified barriers must be actively mitigated through the involvement of all relevant stakeholders. Despite the existing barriers, the findings of this study provide a promising outlook for the implementation of mHealth systems in Sub-Saharan Africa. The study highlights the numerous opportunities that exist for the successful integration of mHealth systems into the region's health care systems. To ensure maximum uptake, it is crucial to adopt a user-centered design approach in mHealth system design and development.

The results of this study have important implications for both mHealth system design and policy-making. The identified barriers and enablers can serve as a guide for the design of mHealth systems that are tailored to meet the needs and preferences of patients, physicians, and health care executives. Additionally, the study findings can inform policy makers on the necessary steps to be taken to facilitate the successful integration of mHealth systems into health care systems in Sub-Saharan Africa.

In conclusion, some of the barriers and enablers of the uptake of mHealth systems in Sub-Saharan Africa are interconnected and require the active involvement of all stakeholders to be addressed. The study provides valuable insights that can inform mHealth system design and policy-making, with the aim of facilitating the successful integration of mHealth systems into health care systems in Sub-Saharan Africa.

Acknowledgments

We extend our gratitude to the Network for Advancement of Sustainable Capacity in Education and Research in Ethiopia (NASCERE). We would also like to thank Eyerusalem Tadese for assisting in interview transcription.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide for patients.
[DOCX File , 16 KB - jmir_v26i1e50337_app1.docx]

Multimedia Appendix 2
Interview guide for physicians.
[DOCX File , 16 KB - jmir_v26i1e50337_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Interview guide for health care executives.
[DOCX File , 15 KB - jmir_v26i1e50337_app3.docx]

Multimedia Appendix 4
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 481 KB - jmir_v26i1e50337_app4.pdf]

References

1. mHealth: use of mobile wireless technologies for public health. World Health Organization. URL: https://tinyurl.com/yc82pnty [accessed 2024-03-05]
2. eHealth: Report by the Secretariat. World Health Organization. URL: https://tinyurl.com/ya3xwjua [accessed 2024-03-05]
3. mHealth: Use of appropriate digital technologies for public health. World Health Organization. URL: https://apps.who.int/gb/ebwha/pdf_files/EB142/B142_20-en.pdf [accessed 2024-03-05]
4. mHealth: New horizons for health through mobile technologies. World Health Organization. URL: https://iris.who.int/bitstream/handle/10665/44607/9789241564250_eng.pdf [accessed 2024-03-05]
5. Bradway M, Carrion C, Vallespin B, Saadatfard O, Puigdomènech E, Espallargues M, et al. mHealth Assessment: Conceptualization of a Global Framework. JMIR Mhealth Uhealth 2017 May 02;5(5):e60 [FREE Full text] [doi: 10.2196/mhealth.7291] [Medline: 28465282]
6. STOA Workshop eHealth in Europe: Reality and challenges ahead. European Parliament. URL: https://www.europarl.europa.eu/cmsdata/149015/Booklet.pdf [accessed 2024-03-05]
7. United States Mhealth App Market Report 2021: Market is Projected to Grow at a CAGR of 26.15% to Reach $5.139 Billion by 2026 from $1.011 Billion in 2019. GlobeNewswire. URL: https://tinyurl.com/36f2yz4v [accessed 2023-04-30]
8. mHealth in an mWorld How mobile technology is transforming health care. Deloitte Center for Health Solutions. URL: https://tinyurl.com/mtpk3npk [accessed 2024-03-05]
9. Alghamdi M, Gashgari H, Househ M. A Systematic Review of Mobile Health Technology Use in Developing Countries. Stud Health Technol Inform 2015;213:223-226. [Medline: 26152999]
10. Medhanyie AA, Little A, Yebyo H, Spigt M, Tadesse K, Blanco R, et al. Health workers' experiences, barriers, preferences and motivating factors in using mHealth forms in Ethiopia. Hum Resour Health 2015 Jan 15;13(1):2 [FREE Full text] [doi: 10.1186/1478-4491-13-2] [Medline: 25588973]
11. O'Connor Y, O'Connor S, Heavin C, Gallagher J, O'Donoghue J. Chapter 10 - Sociocultural and Technological Barriers Across all Phases of Implementation for Mobile Health in Developing Countries. In: Al-Jumeily D, Hussain A, Mallucci C, Oliver C, editors. Applied Computing in Medicine and Health A volume in Emerging Topics in Computer Science and Applied Computing. Burlington, MA: Morgan Kaufmann; 2016:212-230.
12. The Mobile Economy Sub-Saharan Africa 2021. GSM Association. URL: https://tinyurl.com/yf684tsx [accessed 2024-03-05]
13. The mHealth Planning Guide: Key Considerations for Integrating Mobile Technology into Health Programs. K4Health. URL: https://pdf.usaid.gov/pdf_docs/PA00MDPV.pdf [accessed 2024-05-03]
14. Latif S, Rana R, Qadir J, Ali A, Imran MA, Younis MS. Mobile Health in the Developing World: Review of Literature and Lessons From a Case Study. IEEE Access 2017;5:11540-11556. [doi: 10.1109/ACCESS.2017.2710800]
15. Aranda-Jan CB, Mohutsiwa-Dibe N, Loukanova S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC Public Health 2014 Feb 21;14:188 [FREE Full text] [doi: 10.1186/1471-2458-14-188] [Medline: 24555733]
16. Aboye G, Vande Walle M, Simegn G, Aerts J. Current evidence on the use of mHealth approaches in Sub-Saharan Africa: A scoping review. Health Policy and Technology 2023 Dec;12(4):100806. [doi: 10.1016/j.hlpt.2023.100806]
17. Betjeman TJ, Soghoian SE, Foran MP. mHealth in Sub-Saharan Africa. Int J Telemed Appl 2013;2013:482324 [FREE Full text] [doi: 10.1155/2013/482324] [Medline: 24369460]
18. Sustainable Development. United Nations. URL: https://sdgs.un.org/goals [accessed 2024-03-15]
19. Ethiopia - Country Commercial Guide. International Trade Administration. URL: https://www.trade.gov/country-commercial-guides/ethiopia-healthcare [accessed 2024-01-29]
20. Yang M, Al Mamun A, Gao J, Rahman MK, Salameh AA, Alam SS. Predicting m-health acceptance from the perspective of unified theory of acceptance and use of technology. Sci Rep 2024 Jan 03;14(1):339 [FREE Full text] [doi: 10.1038/s41598-023-50436-2] [Medline: 38172184]
21. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: Self-efficacy and privacy concerns. Front Psychol 2022 Aug 11;13:944976. [doi: 10.3389/fpsyg.2022.944976] [Medline: 36033004]
22. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A Systematic Review of the Technology Acceptance Model in Health Informatics. Appl Clin Inform 2018 Jul 15;9(3):604-634 [FREE Full text] [doi: 10.1055/s-0038-1668091] [Medline: 30112741]

XSL•FO
RenderX

23. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical Factors Affecting Patients' Adoption of Mobile Health Tools: Systematic Literature Review and Narrative Synthesis. JMIR Mhealth Uhealth 2022 May 05;10(5):e36284 [FREE Full text] [doi: 10.2196/36284] [Medline: 35318189]
24. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding Clinicians' Adoption of Mobile Health Tools: A Qualitative Review of the Most Used Frameworks. JMIR Mhealth Uhealth 2020 Jul 06;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
25. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital Trends, Digital Literacy, and E-Health Engagement Predictors of Breast and Colorectal Cancer Survivors: A Population-Based Cross-Sectional Survey. Int J Environ Res Public Health 2023 Jan 13;20(2):1472 [FREE Full text] [doi: 10.3390/ijerph20021472] [Medline: 36674237]
26. Arias López M, Ong BA, Borrat Frigola X, Fernández A, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: A scoping review. PLOS Digit Health 2023 Oct;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
27. Doyle AM, Bandason T, Dauya E, McHugh G, Grundy C, Dringus S, et al. Mobile Phone Access and Implications for Digital Health Interventions Among Adolescents and Young Adults in Zimbabwe: Cross-Sectional Survey. JMIR Mhealth Uhealth 2021 Jan 13;9(1):e21244 [FREE Full text] [doi: 10.2196/21244] [Medline: 33439136]
28. Human Development Index (HDI). United Nations Development Program. URL: https://tinyurl.com/4vhpfhnn [accessed 2024-03-05]
29. Nikolopoulou K. What Is Snowball Sampling? Scribbr. URL: https://tinyurl.com/52ndd2p4 [accessed 2023-05-10]
30. Semi-Structured Interviews. KnowFife Hub. 2018 Mar. URL: https://tinyurl.com/ex5prr8s [accessed 2024-03-05]
31. George T. Semi-Structured Interview. Scribbr. URL: https://www.scribbr.com/methodology/semi-structured-interview/ [accessed 2023-05-10]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
33. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. Addotey-Delove M, Scott R, Mars M. Review of patients' perspectives of m-health adoption factors in the developing world. Development of a proposed conceptual framework. Informatics in Medicine Unlocked 2020;21:100460 [FREE Full text] [doi: 10.1016/j.imu.2020.100460]
35. Busse TS, Nitsche J, Kernebeck S, Jux C, Weitz J, Ehlers JP, et al. Approaches to Improvement of Digital Health Literacy (eHL) in the Context of Person-Centered Care. Int J Environ Res Public Health 2022 Jul 07;19(14):8309 [FREE Full text] [doi: 10.3390/ijerph19148309] [Medline: 35886158]
36. Koumpouros Y. User-Centric Design Methodology for mHealth Apps: The PainApp Paradigm for Chronic Pain. Technologies 2022 Jan 31;10(1):25. [doi: 10.3390/technologies10010025]
37. Gong E, Gu W, Luo E, Tan L, Donovan J, Sun C, et al. Development and Local Contextualization of Mobile Health Messages for Enhancing Disease Management Among Community-Dwelling Stroke Patients in Rural China: Multimethod Study. JMIR Mhealth Uhealth 2019 Dec 17;7(12):e15758 [FREE Full text] [doi: 10.2196/15758] [Medline: 31845901]
38. Martin T. Assessing mHealth: opportunities and barriers to patient engagement. J Health Care Poor Underserved 2012 Aug;23(3):935-941. [doi: 10.1353/hpu.2012.0087] [Medline: 24212144]
39. Byambasuren O, Beller E, Hoffmann T, Glasziou P. Barriers to and Facilitators of the Prescription of mHealth Apps in Australian General Practice: Qualitative Study. JMIR Mhealth Uhealth 2020 Jul 30;8(7):e17447 [FREE Full text] [doi: 10.2196/17447] [Medline: 32729839]
40. The Business Case For Rigorous Evaluation Of Mobile Health Apps. Health Affairs. URL: https://tinyurl.com/56e494ny [accessed 2024-03-05]
41. Weiss M, Botha A, Herselman M, Loots G. Blockchain as an enabler for public mHealth solutions in South Africa. 2017 Presented at: 2017 IST-Africa Week Conference (IST-Africa); May 30-June 02, 2017; Windhoek, Namibia. [doi: 10.23919/ISTAFRICA.2017.8102404]
42. Galvin HK, DeMuro PR. Developments in Privacy and Data Ownership in Mobile Health Technologies, 2016-2019. Yearb Med Inform 2020 Aug;29(1):32-43 [FREE Full text] [doi: 10.1055/s-0040-1701987] [Medline: 32823298]
43. Hernández-Álvarez L, de Fuentes JM, González-Manzano L, Hernández Encinas L. Privacy-Preserving Sensor-Based Continuous Authentication and User Profiling: A Review. Sensors (Basel) 2020 Dec 25;21(1):92 [FREE Full text] [doi: 10.3390/s21010092] [Medline: 33375710]
44. Gruson D. Controlling Reliability, Interoperability and Security of Mobile Health Solutions. EJIFCC 2021 Jun;32(2):118-123 [FREE Full text] [Medline: 34421479]
45. Digitalization in Healthcare: Benefits and Challenges. CareCloud. URL: https://tinyurl.com/28zfazyh [accessed 2024-01-30]
46. Can digital healthcare solutions reduce clinician workload? siHealth. URL: https://tinyurl.com/tr6dtf7n [accessed 2024-01-30]
47. Nipa NS, Alam M, Haque MS. Identifying Relevant Stakeholders in Digital Healthcare. In: Mahmud M, Kaiser MS, Kasabov N, Iftekharuddin K, Zhong N, editors. Applied Intelligence and Informatics. AII 2021. Communications in Computer and Information Science, vol 1435. Cham: Springer; 2021:349-357.

XSL•FO
RenderX

48. Kruse C, Betancourt J, Ortiz S, Valdes Luna SM, Bamrah IK, Segovia N. Barriers to the Use of Mobile Health in Improving Health Outcomes in Developing Countries: Systematic Review. J Med Internet Res 2019 Oct 09;21(10):e13263 [FREE Full text] [doi: 10.2196/13263] [Medline: 31593543]
49. Aamir J, Ali SM, Kamel Boulos MN, Anjum N, Ishaq M. Enablers and inhibitors: A review of the situation regarding mHealth adoption in low- and middle-income countries. Health Policy and Technology 2018 Mar;7(1):88-97. [doi: 10.1016/j.hlpt.2017.11.005]
50. Place J, Farret A, Del Favero S, Bruttomesso D, Renard E. Assessment of Patient Perceptions About Web Telemonitoring Applied to Artificial Pancreas Use at Home. J Diabetes Sci Technol 2014 Mar 02;8(2):225-229 [FREE Full text] [doi: 10.1177/1932296814525540] [Medline: 24876571]
51. Said GRE. Factors Affecting mHealth Technology Adoption in Developing Countries: The Case of Egypt. Computers 2022 Dec 28;12(1):9. [doi: 10.3390/computers12010009]
52. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and Facilitators of the Use of Mobile Health Apps From a Security Perspective: Mixed-Methods Study. JMIR Mhealth Uhealth 2019 Apr 16;7(4):e11223 [FREE Full text] [doi: 10.2196/11223] [Medline: 30990458]
53. Zakerabasali S, Ayyoubzadeh SM, Baniasadi T, Yazdani A, Abhari S. Mobile Health Technology and Healthcare Providers: Systemic Barriers to Adoption. Healthc Inform Res 2021 Oct;27(4):267-278 [FREE Full text] [doi: 10.4258/hir.2021.27.4.267] [Medline: 34788907]

Abbreviations

EMR: electronic medical record
ICT: information and communication technology
mHealth: mobile health
OPD: outpatient department
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Mavragani; submitted 28.06.23; peer-reviewed by L Guo, H Pilabré; comments to author 11.10.23; revised version received 01.11.23; accepted 13.02.24; published 27.03.24.

Please cite as:
Aboye GT, Simegn GL, Aerts JM
Assessment of the Barriers and Enablers of the Use of mHealth Systems in Sub-Saharan Africa According to the Perceptions of Patients, Physicians, and Health Care Executives in Ethiopia: Qualitative Study
J Med Internet Res 2024;26:e50337
URL: https://www.jmir.org/2024/1/e50337
doi:10.2196/50337
PMID:38536231

©Genet Tadese Aboye, Gizeaddis Lamesgin Simegn, Jean-Marie Aerts. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Minder Mobile Mental Health and Substance Use Intervention for University Students: Randomized Controlled Trial

Melissa Vereschagin[1], BSc; Angel Y Wang[1], BA, MPhil; Chris G Richardson[2], PhD; Hui Xie[3], PhD; Richard J Munthali[1], PhD; Kristen L Hudec[1], PhD; Calista Leung[1], BA; Katharine D Wojcik[4], PhD; Lonna Munro[1], BSc; Priyanka Halli[1], MPH, MD; Ronald C Kessler[5], PhD; Daniel V Vigo[1,2], LicPs, MD, DrPH

[1]Department of Psychiatry, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[2]School of Population and Public Health, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[3]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

[4]Menninger Department of Psychiatry & Behavioural Sciences, Baylor College of Medicine, Houston, TX, United States

[5]Department of Health Care Policy, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Daniel V Vigo, LicPs, MD, DrPH
Department of Psychiatry
Faculty of Medicine
University of British Columbia
2255 Wesbrook Mall
Vancouver, BC, V6T2A1
Canada
Phone: 1 6048228048
Email: daniel.vigo@ubc.ca

Abstract

Background: University attendance represents a transition period for students that often coincides with the emergence of mental health and substance use challenges. Digital interventions have been identified as a promising means of supporting students due to their scalability, adaptability, and acceptability. *Minder* is a mental health and substance use mobile app that was codeveloped with university students.

Objective: This study aims to examine the effectiveness of the *Minder* mobile app in improving mental health and substance use outcomes in a general population of university students.

Methods: A 2-arm, parallel-assignment, single-blinded, 30-day randomized controlled trial was used to evaluate *Minder* using intention-to-treat analysis. In total, 1489 participants were recruited and randomly assigned to the intervention (n=743, 49.9%) or waitlist control (n=746, 50.1%) condition. The *Minder* app delivers evidence-based content through an automated chatbot and connects participants with services and university social groups. Participants are also assigned a trained peer coach to support them. The primary outcomes were measured through in-app self-assessments and included changes in general anxiety symptomology, depressive symptomology, and alcohol consumption risk measured using the 7-item General Anxiety Disorder scale, 9-item Patient Health Questionnaire, and US Alcohol Use Disorders Identification Test–Consumption Scale, respectively, from baseline to 30-day follow-up. Secondary outcomes included measures related to changes in the frequency of substance use (cannabis, alcohol, opioids, and nonmedical stimulants) and mental well-being. Generalized linear mixed-effects models were used to examine each outcome.

Results: In total, 79.3% (589/743) of participants in the intervention group and 83% (619/746) of participants in the control group completed the follow-up survey. The intervention group had significantly greater average reductions in anxiety symptoms measured using the 7-item General Anxiety Disorder scale (adjusted group mean difference=−0.85, 95% CI −1.27 to −0.42; $P<.001$; Cohen $d=-0.17$) and depressive symptoms measured using the 9-item Patient Health Questionnaire (adjusted group mean difference=−0.63, 95% CI −1.08 to −0.17; $P=.007$; Cohen $d=-0.11$). A reduction in the US Alcohol Use Disorders Identification Test–Consumption Scale score among intervention participants was also observed, but it was not significant ($P=.23$). Statistically significant differences in favor of the intervention group were found for mental well-being and reductions in the frequency of

XSL•FO
RenderX

cannabis use and typical number of drinks consumed. A total of 77.1% (573/743) of participants in the intervention group accessed at least 1 app component during the study period.

Conclusions: In a general population sample of university students, the *Minder* app was effective in reducing symptoms of anxiety and depression, with provisional support for increasing mental well-being and reducing the frequency of cannabis and alcohol use. These findings highlight the potential ability of e-tools focused on prevention and early intervention to be integrated into existing university systems to support students' needs.

Trial Registration: ClinicalTrials.gov NCT05606601; https://clinicaltrials.gov/ct2/show/NCT05606601

International Registered Report Identifier (IRRID): RR2-10.2196/49364

(J Med Internet Res 2024;26:e54287) doi:10.2196/54287

KEYWORDS

mental health; substance use; college students; mobile interventions; digital interventions; randomized controlled trial; mobile phone

Introduction

Background

University attendance is a transitional period in which many students experience novel stressors related to moving away from home, navigating new social environments, and managing increased educational and financial demands in the absence of their traditional support systems [1,2]. The transition to attending university also coincides with the peak period of onset of many mental disorders, including mood, anxiety, and substance use disorders [3,4]. Studies have documented the high rates of mental health and substance use problems experienced by university students [5], with research also indicating that students with preexisting mental health problems can experience a worsening of their conditions following the transition to attending university [6].

Despite this high need, most students experiencing mental health problems do not receive treatment [7]. Research investigating help seeking among university students indicates that, compared to structural barriers, attitudinal barriers are the most important reasons for not seeking help [8]. The most commonly cited reason students give for not seeking help is a preference for handling things on their own [7,8]. One way of adapting interventions to align with this preference is to provide students with tools that are self-guided and allow them autonomy over how and when to use the tools provided. e-Interventions can be accessed by users at any time and have been demonstrated to be effective in improving various mental health [9] and substance use outcomes among university students [10]. These interventions have also been identified as key components in proposed models of care for universities [11]. While much of the literature on e-interventions has been focused on web-based tools, mobile apps have been identified as a promising means of delivering mental health interventions due to not only the increase in smartphone use but also the wide range of interventions that can be delivered through mobile platforms [12,13].

Given the range of challenges faced by university students, including the high rates of disorder-level and subclinical mental health and substance use problems, transdiagnostic approaches to early intervention and prevention may be beneficial for this population [14]. Developing this type of intervention requires the use of a holistic, student-centered design approach to identify evidence-based condition-specific and cross-cutting opportunities for intervention. Furthermore, the intervention needs to be aligned with the perceived needs and preferences of students to ensure meaningful engagement [15]. On the basis of these requirements, we codeveloped a mental health and substance use mobile app called *Minder* for Canadian university students. This participatory codevelopment process involved significant input from students through the creation of a Student Advisory Committee, usability testing via a virtual boot camp (ie, individual user-testing combined with a web-based survey), focus groups, and a pilot feasibility study [16].

Objectives

The objective of this study was to test the effectiveness of the *Minder* mobile app in improving mental health and substance use outcomes in a general population of university students.

Methods

Trial Design

This study was based on a 2-arm, parallel-assignment, single-blinded (the statistician was blinded), 30-day randomized controlled trial with 1 intervention group and 1 waitlist control group. The study was registered at ClinicalTrials.gov (NCT05606601), and a full study protocol has been published [17]. No significant changes to the trial protocol or intervention content were made during the trial period; however, several minor adjustments, along with a description of minor technical issues, can be found in Multimedia Appendix 1.

Ethical Considerations

Ethics approval was obtained from the University of British Columbia (UBC) Behavioural Research Ethics Board on January 6, 2022 (ethics ID: H21-03248). Informed consent was obtained through a web-based self-assessment questionnaire at the beginning of the study. Participants were informed of their ability to opt out at any point within the study by emailing the research team. Identifiable data were stored in data files within the app backend, which were separate from all deidentified app use and survey data. This information can only be linked using a unique study ID number. Participants received a CAD $10 (US $7.4) gift card for completion of the baseline survey and

XSL•FO
RenderX

an additional CAD $10 (US $7.4) gift card for completion of the 30-day follow-up survey.

Participants

The study was conducted at the UBC Point Grey (Vancouver, British Columbia, Canada) campus. Participants needed to confirm their eligibility using a web-based self-assessment questionnaire before registering and consenting to the study. The inclusion criteria were as follows: students currently enrolled at the UBC Vancouver campus, aged ≥17 years, having access to and being able to use a smartphone with Wi-Fi or cellular data, and speaking English. The only exclusion criterion was based on a single screening question assessing suicidality risk ("We want to make sure that this app is appropriate for you at this time. Do you have a current suicidal plan [i.e., a plan to end your life]?"). Anyone endorsing a current suicidal plan (ie, answering "yes") was prevented from registering and was instead provided with a list of local crisis resources. The eligibility criterion of being a current UBC student was confirmed using a unique student log-in checkpoint as part of the registration process. This process also ensured that each student could only enroll once.

Recruitment and Consent

Given the large sample size needed for this study, many different recruitment methods were used. Online recruitment occurred through various social media platforms and a linked ongoing Student E-Mental Health trend study [18]. Recruitment also occurred through in-person and on-campus engagements, such as setting up informational booths at the university, displaying posters about the study, visiting in-person and online classes, having professors share study information with their classes, and contacting student groups to share information with their members. Paid bus and bus stop advertisements at the university were also used. A more detailed description of the recruitment methods can be found in the study protocol for this trial [17].

Participants' consent was obtained using Qualtrics (Qualtrics International Inc), a web-based form, after completing the eligibility screening. The consent form indicated that participants would either gain access to the full app immediately or in 30 days following completion of the final survey. Individual accounts were created for each participant and sent to them with a link to download the app. Upon downloading the app and completing the baseline survey, participants were randomly assigned through the app to the intervention group, which received full access to the *Minder* app, or to the control group, which only had access to a restricted version of the app that included a short introduction video and the baseline and follow-up surveys. Participants received a CAD $10 (US $7.40) gift card for completion of the baseline survey and an additional CAD $10 (US $7.40) gift card for completion of the 30-day follow-up survey; however, the use of the app itself was not remunerated.

Randomization and Intervention

Participants were randomized using a custom-developed automated process incorporated directly into the mobile app following completion of the baseline survey. The system assigned participants to either the intervention or control group using a predetermined block randomization list (1:1 randomization in blocks of 10) stratified for past drug use (any lifetime use of opioids or nonmedical stimulants). Stratification by past drug use was used to account for the low number of students using these substances and the need to ensure that they were evenly distributed across the study groups. The randomization lists (1 for each stratification group) were generated using the web-based stratified block randomization list creator in the clinical trial software Sealed Envelope (Sealed Envelope Ltd) [19]. The intervention and control groups completed the main assessments of the primary and secondary outcomes at baseline and the 30-day follow-up. The intervention group was also prompted to complete a short survey at 2 weeks that consisted of a limited set of questions on anxiety and depression symptoms.

The *Minder* mobile app was codeveloped with university students and professionals with the goal of creating an engaging self-directed tool for students to improve their mental health and manage substance use. The intervention is designed for a general population of students and, thus, addresses a wide range of challenges related to postsecondary student life, including managing emotions, relationships, well-being, and university life. The self-directed nature of the app also allows students to access features when needed. The codevelopment process consisted of ongoing student input through student staff members and volunteers along with several phases of purposeful student engagement and feedback. Further details on the codevelopment process can be found in the study by Vereschagin et al [16].

Participants who were randomized to the intervention group were given full access to the *Minder* app and instructed to use it as they wanted. They were also presented with a tutorial video outlining the different features of the app. The *Minder* intervention consists of 4 main components: Chatbot Activities, Services, Community, and Peer Coaching. The chatbot activities consist of an automated preprogrammed chatbot that delivers evidence-based messages and videos. The content is based primarily on cognitive behavioral therapy and psychoeducation; however, there is also content adapted from dialectical behavioral therapy, mindfulness, metacognitive training, and motivational interviewing. The content sections are presented on a home page map with several islands: *University Life*, *Wellbeing*, *Relationships*, *Sadness*, *Stress & Anxiety*, and *Substance Use* (Figure 1). There is also an *Explore Chat* located on the home page map that guides participants to select an activity that may be relevant to their current needs. A full list of the content included can be found in Multimedia Appendix 2 [16]. Most of the chat activities also contain a summary page that is unlocked upon completion of the chat activity and allows participants to review content they learned at a later time. In addition, several chat activities contain specific practice components that also become unlocked after the main activity is complete.

The Peer Coaching component consists of trained volunteers assigned to each participant. These peer coaches reach out to participants at the beginning of the trial and midway through. They can provide support in navigating the app or nonclinical peer support based on active listening and problem-solving.

XSL•FO
RenderX

Peer coaches can communicate with participants through an in-app chat asynchronously or synchronously through scheduled appointments delivered over in-app chat message or audio or video call. Before engaging with peer coaches, participants must provide a phone number that can be used to contact them in a crisis situation and affirm that they are not currently at risk of self-harm or having suicidal thoughts, are not under the influence of substances, and understand the circumstances in which confidentiality would need to be broken (ie, crisis situations or abuse of a minor; Figure 2A).

Figure 1. Minder home screen and content maps. The home screen contains 6 topic islands that users can select from. Each island leads to a separate map with chatbot activities and summaries. Additional features include the SOS button and the Explore Chat that guides users to activities.

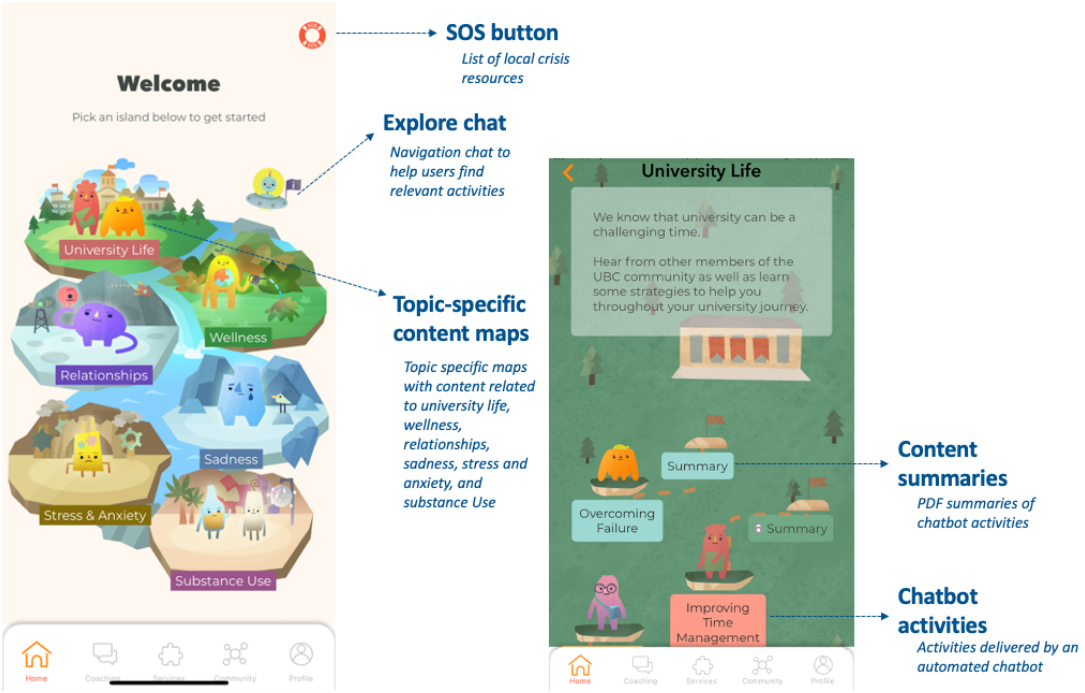

Figure 2. Peer coach, services, and community components. (A) The peer coach component allows users to connect with trained student volunteers, (B) the services component matches users with resources, and (C) the community component lists student groups by interest.

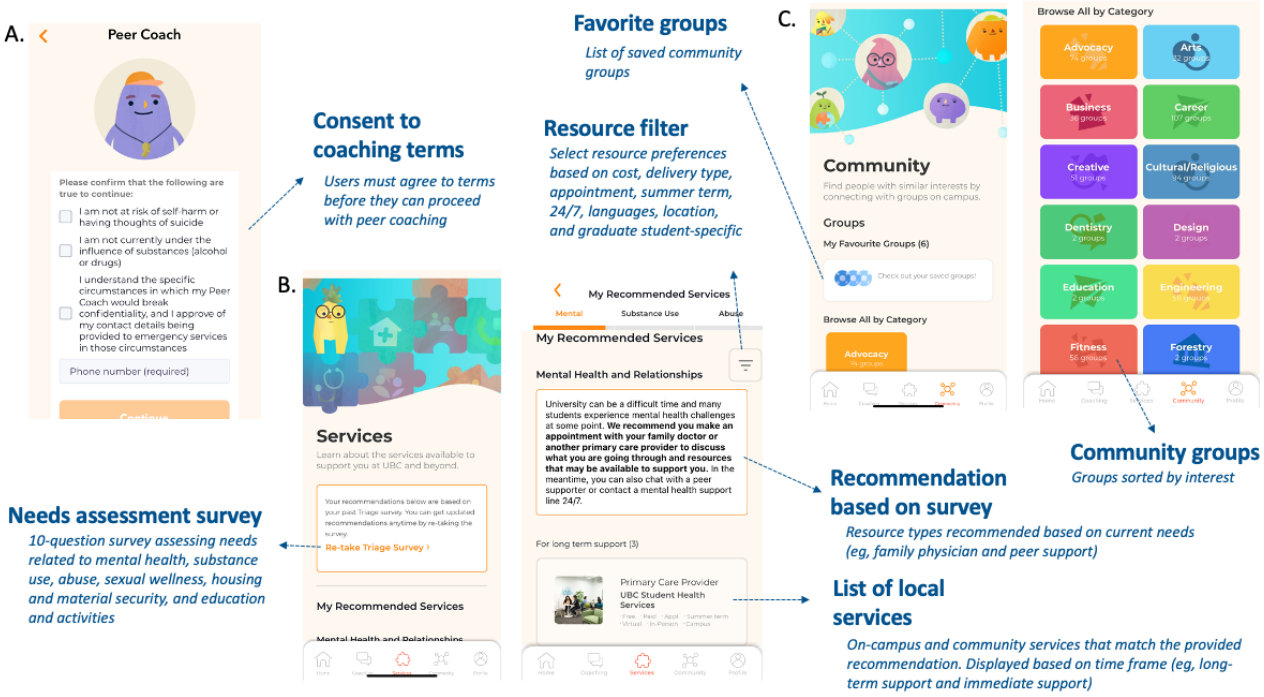

The Services component consists of a 10-question survey tool that provides participants with recommendations for resources based on their current needs and preferences (Figure 2B). The survey tool and recommendations were adapted from a

XSL•FO
RenderX

previously developed tool for university students [20] and can be completed multiple times to receive new recommendations. Recommendations are provided for 6 areas related to student well-being: mental health and relationships, substance use, abuse, sexual wellness, housing, and education and activities. An additional safety component was added so that participants who based on the services survey were considered to be at high risk of suicidality (ie, plan for suicide or recent attempt and thoughts of hurting others) were asked to consent to provide their contact information and receive an expedited appointment with the university counseling services.

The Community component consists of a searchable directory of student groups or clubs at the university that are sorted by interest (eg, volunteering, arts, and advocacy; Figure 2C).

The *Minder* app also contains several other general features. An SOS button appears in the top corner of the home page and provides a list of crisis resources if needed (Figure 1). The settings page allows participants to update their username and password as well as change their avatar. Several types of push notifications were delivered through the app. General notifications were sent on days 4, 18, and 24. Reminders to complete the 2-week and 30-day follow-up surveys were sent as push notifications and via automated email reminders on those dates. Additional email reminders were sent on days 35 and 41 to remind participants to complete the follow-up survey if they had not done so already.

Participants who were randomized to the control group had access to a locked version of the app that only allowed them to complete the baseline survey and view a short introduction video that appears before the log-in screen. Following completion of the baseline survey, they received a pop-up message telling them that they would be notified when it was time to complete the next survey. The app was then locked so that control participants were not able to access any other areas of the app. At 30 days, participants were notified that it was time to complete the 30-day follow-up survey within the app, and this survey became unlocked. The app provided push notifications and automated email reminders to complete the follow-up survey on day 30 as well as additional email reminders on days 35 and 41 if the survey had not yet been completed.

Individuals who consented to participate in the study and received an invitation email to download the app but did not complete the baseline survey received additional email reminders at 7 days, at approximately 20 days, and several months later following the creation of their account.

The participants' use of the app components was recorded through the app back-end system. This included starting each of the chatbot activities, completing the services survey and receiving recommendations, viewing community groups, and communicating with a peer coach.

Measures

The main assessments of the primary and secondary outcomes were collected through self-assessment within the Minder app at baseline and at the 30-day follow-up. The 30-day follow-up survey had to be completed within 44 days of the beginning of the baseline survey (30 days plus 2 weeks to accommodate the use of reminders) for the participants to be included in the analysis.

Primary Outcomes

The primary outcomes assessed in this study were changes in general anxiety symptomology, depressive symptomology, and alcohol consumption risk from baseline to follow-up at 30 days. All outcomes were assessed using self-report questionnaires completed directly in the mobile app.

Anxiety symptoms were assessed using the 7-item General Anxiety Disorder scale (GAD-7) assessment—a commonly used self-report scale that assesses symptoms of generalized anxiety [21]. Each GAD-7 question is scored from 0 (*not at all*) to 3 (*nearly every day*), with total scores ranging from 0 to 21 and higher scores indicating a worse outcome (ie, greater frequency of anxiety symptoms).

Depressive symptoms were assessed using the 9-item Patient Health Questionnaire (PHQ-9) self-report scale [22]. Each of the 9 questions is scored from 0 (*not at all*) to 3 (*nearly every day*). The total scores range from 0 to 29, with higher scores indicating a worse outcome (ie, a greater frequency of depressive symptoms).

Alcohol consumption risk was assessed using the US Alcohol Use Disorders Identification Test–Consumption Scale (USAUDIT-C) [23]. The USAUDIT-C is a 3-item self-report scale adapted from the consumption questions in the Alcohol Use Disorders Identification Test (AUDIT) [24]. Compared to the AUDIT, the USAUDIT-C includes expanded response options for the first 3 AUDIT questions—from 5 to 7 categories—to allow for more precise measurements when accounting for differences in standard drink sizes and cutoff limits. The higher the total score on the USAUDIT-C, the greater the respondent's alcohol consumption and related risk [23].

Secondary Outcomes

A range of secondary outcomes, including reduced use of other substances and additional mental health constructs that the *Minder* app was theorized to affect, were also assessed. For the purposes of this study, we focused on examining changes from baseline to follow-up at 30 days in frequency of substance use (cannabis, alcohol, opioids, and nonmedical stimulants) and mental well-being.

Frequency of cannabis use was assessed using a single self-report question on frequency of cannabis consumption in the previous 30 days. The 3 questions on the USAUDIT-C assessed unique dimensions of alcohol consumption. Frequency of alcohol use was assessed using responses to the first question on the USAUDIT-C, which asks how often participants have a drink containing alcohol. The number of drinks consumed in a typical drinking session was assessed using the second question on the USAUDIT-C, which asks participants how many drinks containing alcohol they have on a typical day when drinking. Frequency of binge drinking was assessed using the third question in the USAUDIT-C, which asks participants how often they have ≥5 (if sex at birth was male) *or* ≥4 (if sex at birth was female) drinks on 1 occasion.

XSL•FO
RenderX

Frequency of any opioid use in the previous 30 days was assessed using self-reported questions that asked about any pharmaceutical opioid (eg, oxycodone; morphine; hydromorphone; meperidine; fentanyl patches; and codeine or codeine-containing products such as Tylenol 1, 2, or 3) with a physician's prescription and taken as prescribed; any pharmaceutical opioid (eg, oxycodone; morphine; hydromorphone; meperidine; fentanyl patches; and codeine or codeine-containing products such as Tylenol 1, 2, or 3) either without a physician's prescription or in larger doses than prescribed to get high, buzzed, or numbed out; and any street opioid (eg, heroin and fentanyl) or any other opioid obtained "on the street." The final opioid use outcome was defined as the most frequent number among any prescribed, nonprescribed, and street opioids.

Frequency of nonmedical stimulant use in the previous 30 days was assessed using self-report questions that asked about the frequency of using any street stimulant (eg, cocaine, crack, methamphetamines, and crystal meth) or prescription stimulant (eg, amphetamine, methylphenidate, and modafinil) either without a physician's prescription or in larger doses than prescribed to get high, buzzed, numbed out, or help them study or for any other reason. The final nonmedical stimulant use outcome was defined as the most frequent number among any prescription stimulant without a prescription or not as prescribed and any street stimulant. The exact wording and response options for the substance use secondary outcome assessments can be found in Multimedia Appendix 3.

Mental well-being was assessed using the Short Warwick-Edinburgh Mental Wellbeing Scale. The Short Warwick-Edinburgh Mental Wellbeing Scale is a 7-item scale that has been widely validated [25], with total scores ranging from 7 to 35 and higher scores indicating a better outcome (ie, higher positive mental well-being).

Sample Size Estimation

The a priori sample size calculation for a small effect assessed using the PHQ-9, GAD-7, and USAUDIT-C was performed using an effect size defined by Cohen $d=\Delta/\sigma$, where Δ is the group mean difference at the completion of the study and σ is the (pooled) within-group SD [17]. For a small effect size (Cohen d=0.2), the sample size required to have 80% power at a P=.02 level of significance (ie, 0.05/3 primary outcomes) was 524 in each group. After incorporating a 30% attrition rate, we anticipated requiring 748 participants in each group for a total of 1496 participants for the trial to be adequately powered.

Statistical Analysis

The study used a single-blinded approach in which only the statistician, who was external to the team, was blinded to the treatment group assignment when examining the primary hypotheses. The primary analysis was intention-to-treat (ITT) including all participants who completed the baseline assessment and were randomized to either the control or intervention group following the analysis plan prespecified in our protocol paper [17]. The analysis considered the following 2 features of the trial design: 3 primary end points and the use of block randomization.

For the 3 primary end points (GAD-7, PHQ-9, and USAUDIT-C scores), a global test for the null hypothesis of no treatment difference in all primary end points between the control and treatment groups was conducted first using a multivariate analysis of covariance for correlated data on GAD-7, PHQ-9, and USAUDIT-C scores at 30 days after the baseline, adjusting for their values at baseline and randomization blocks. Compared with testing each outcome separately, the advantages of the global test from joint modeling included more parsimonious hypothesis tests and mitigated concerns related to multiple testing [26-28] as well as pooling of the information over the correlated outcomes to increase study power, especially with missing outcomes [29]. If the global test rejected the null hypothesis and we concluded that there was an intervention effect on at least 1 of the 3 end points, we would then analyze each end point separately to identify which of the 3 study end points were affected by the intervention. We then used the sequential Hochberg correction method [30] to control the overall familywise error rate at α=.05 when testing the hypothesis for each individual primary end point. For baseline characteristics, 2-sample t tests (2-tailed) were used to compare means, chi-square tests were used to compare proportions, and the Kruskal-Wallis test was used to test for differences in medians.

The general approach for analyzing all the types of individual study outcomes separately (including the 3 primary end points and the secondary end points) was generalized linear mixed-effects models (GLMMs) for clustered measures with the randomization block as the clustering variable. These models can handle a wide range of outcome types, including continuous, binary, ordinal, and count, and can account for the correlations among observations within the same cluster. GLMMs have been widely used for conducting ITT analysis in randomized controlled trials with missing outcome data and can account for data missing at random (MAR) without the need to model why data are missing or to perform explicit imputations of the missing values [31]. Specifically, we analyzed each primary end point using a linear mixed-effects model, a special case of the GLMM.

For secondary outcomes, we used linear mixed-effects models to analyze mental well-being measures and a mixed-effects quasi–Poisson regression model with a log link (a special case of GLMM) for substance use frequency measures, and the zero-inflated Poisson was used for the number of drinks to deal with the excess zero counts. The treatment effect on an outcome at 30 days was assessed in these models with the treatment allocation as the main explanatory variable and with adjustment for baseline outcome assessment value as a fixed effect and the randomization block as a random effect. Robust empirical sandwich SE estimates that are robust to model misspecifications (eg, nonnormal error terms) were used for statistical inference.

The results from these GLMMs for the analysis are reported in the Results section as the model-adjusted differences in the group mean values of the continuous end points between the intervention and control groups. These intervention effect estimates, 95% CIs, and P values were obtained from the aforementioned GLMMs, clustering on randomization block effects and adjusting for the baseline outcome values.

XSL•FO
RenderX

Standardized effect sizes were calculated for continuous scores by dividing the adjusted mean differences by the SDs across all participants at baseline. The incidence rate ratios (IRRs) from Poisson regression through GLMM and zero inflation for the number of drinks are also reported.

To evaluate the robustness of the results to alternative assumptions regarding missing data, sensitivity analyses were conducted on the primary outcomes via (1) using selection models to measure the potential impacts of data missing not at random [32,33] and (2) adjusting analysis for additional baseline covariates potentially predictive of missing data. All the primary analyses were conducted in SAS (version 9.4; SAS Institute), except for the sensitivity analysis of the selection models, which was conducted in the *isni* package in R (version 3.4; R Foundation for Statistical Computing) [34], whereas secondary analyses were conducted in Stata (version 15.1; StataCorp) [35].

Data and Privacy

Many steps were taken to ensure the privacy of participants. Each participant received an individual account that had a unique username and password. They were then able to change this password upon logging in. Only participant emails and phone numbers (if provided through access to peer coaching) were stored within the app; names were used only for consent and were never entered into the app. Instead, participants could create a unique username within the app that they were informed should not include their full name. Identifiable data (email and phone number) were stored in data files within the app back end separate from all the app use and survey data, which were recorded with only a study ID number.

Results

Recruitment

Recruitment initially began on September 4, 2022, during an on-campus student orientation event where interested students were asked to provide contact information to receive a follow-up email; however, participants were not provided with app download information and user accounts to begin the study until September 28, 2022, due to technical delays. Recruitment concluded on June 2, 2023, with the last participant beginning the trial on June 11, 2023.

Participant Flow

In total, 2293 individuals were invited to participate in the trial following eligibility screening and provision of informed consent. Of those 2293 individuals, 1489 (64.9%) participants completed the baseline survey and were randomized into the trial, with 743 (49.9%) of the 1489 participants in the intervention group and 746 (50.1%) in the control group. A total of 79.3% (589/743) of participants in the intervention group and 83% (619/746) of participants in the control group completed the 30-day follow-up survey within the specified 44-day period to be included in the analysis (ie, 279/1489, 18.7% of participants who completed the baseline survey did not complete the 30-day follow-up survey within 44 days and were therefore considered lost to follow-up). Additional information on the participant flow is shown in Figure 3.

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) flow diagram completed with participant information.

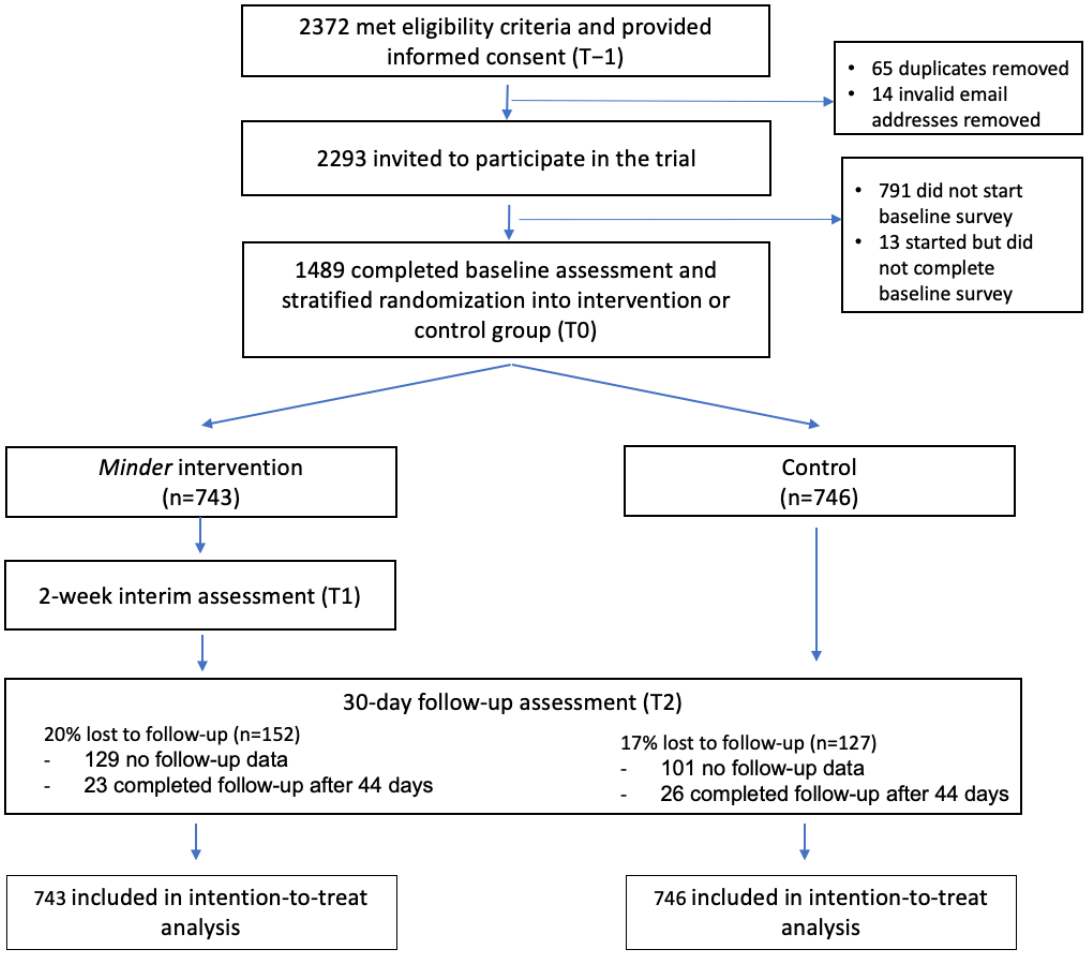

Participant Characteristics

The baseline characteristics of the participants in the intervention and control groups are presented in Table 1. The median age of the participants was 20 years, and 70.3% (1045/1487) self-identified as women. In terms of mental health, 33.8% (455/1347) reported a history of anxiety, and 38.9% (576/1481) reported moderate or greater levels of recent anxiety (ie, total score of ≥10) based on the GAD-7. A history of depression was reported by 28.4% (382/1347) of participants, with 43.8% (645/1474) reporting moderate or greater levels of recent depressive symptomology (ie, total score of ≥10) on the PHQ-9. The intervention group had higher baseline scores on both the GAD-7 (P=.02) and PHQ-9 (P=.02). No other statistically significant differences in baseline characteristics were found between the intervention and control groups (Table 1).

XSL•FO
RenderX

Table 1. Participant characteristics at baseline (N=1489)[a].

Characteristics	Overall	Intervention (n=743)	Control (n=746)	*P* value
Age (y; n=1042), median (IQR)	20 (19-23)	20 (19-23)	20 (19-23)	.54
Student year (n=1487), n (%)				.28
Undergraduate: year 1	384 (25.8)	195 (26.3)	189 (25.3)	
Undergraduate: year 2	280 (18.8)	126 (17)	154 (20.6)	
Undergraduate: year 3	264 (17.8)	136 (18.4)	128 (17.2)	
Undergraduate: year 4 or 5	287 (19.3)	136 (18.4)	151 (20.2)	
Graduate	206 (13.9)	114 (15.4)	92 (12.3)	
Other	66 (4.4)	34 (4.6)	32 (4.3)	
Gender (n=1487), n (%)				.71
Woman	1045 (70.3)	527 (71.1)	518 (69.4)	
Man	364 (24.5)	178 (24)	186 (24.9)	
Nonbinary	78 (5.2)	36 (4.9)	42 (5.6)	
Race or ethnicity (n=1454), n (%)				.66
Black	18 (1.2)	7 (1)	11 (1.5)	
East or Southeast Asian	479 (32.9)	240 (33.1)	239 (32.8)	
Hispanic or Latino	46 (3.2)	24 (3.3)	22 (3)	
Indigenous or First Nations	28 (1.9)	14 (1.9)	14 (1.9)	
South Asian	198 (13.6)	102 (14)	96 (13.2)	
West or Central Asian	58 (4)	30 (4.1)	28 (3.8)	
White	436 (30)	210 (28.9)	226 (31)	
Multiracial	187 (12.9)	95 (13.1)	92 (12.6)	
Other	4 (0.3)	4 (0.6)	0 (0)	
Mental health diagnosis or treatment history (n=1347), n (%)				
History of anxiety (yes)	455 (33.8)	233 (34.5)	222 (33)	.57
History of depression (yes)	382 (28.4)	196 (29)	186 (27.7)	.58
History of substance (alcohol or drug) use (yes)	19 (1.4)	9 (1.3)	10 (1.5)	.81
History of other mental health disorders (yes)	353 (26.2)	178 (26.4)	175 (26)	.89
Current mental health and substance use				
GAD-7[b] (n=1481), mean (SD)	8.7 (5)	9.0 (5)	8.4 (5)	.02
GAD-7≥10 (moderate and above), n (%)	576 (38.9)	297 (40.1)	279 (37.7)	.35
PHQ-9 [c] n=1474), mean (SD)	9.2 (5.8)	9.6 (5.9)	8.9 (5.6)	.02
PHQ-9≥10 (moderate and above), n (%)	645 (43.8)	326 (44.5)	319 (43)	.55
USAUDIT-C[d] (n=1483), mean (SD)	3.5 (3.2)	3.6 (3.2)	3.5 (3.3)	.31
SWEMWS[e] (n=1474), mean (SD)	21.5 (4.2)	21.3 (4.3)	21.6 (4.1)	.19

[a]Variations in the total n values for each characteristic were due to missing responses. We used 2-sample *t* tests (2-tailed) to compare means, chi-square tests to compare proportions, and the Kruskal-Wallis test to test for differences in medians.

[b]GAD-7: 7-item General Anxiety Disorder scale.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]USAUDIT-C: US Alcohol Use Disorders Identification Test–Consumption Scale.

[e]SWEMWS: Short Warwick-Edinburgh Mental Wellbeing Scale.

XSL•FO
RenderX

App Use

Among those in the intervention group (743/1489, 49.9%), 77.1% (573/743) accessed at least 1 app component during the 30-day study period. More specifically, 73.8% (548/743) engaged in 1 or more chatbot activities, 21.8% (162/743) accessed the community component, 27.9% (207/743) completed the services survey, and 17.2% (128/743) accessed a peer coach.

Outcomes

Primary Outcomes

For the 3 primary end points (GAD-7, PHQ-9, and USAUDIT-C scores), a global Wald test with the null hypothesis of no treatment difference in all primary end points between the control and treatment groups was conducted using a marginal multivariate analysis of covariance model for correlated data. This multivariate test of overall group differences was statistically significant (Hotelling test=0.02; *P*=.001; Table 2), indicating that there were group differences in at least 1 of the primary outcomes. We then tested each of the primary outcomes using a GLMM for clustered measures, adjusting for baseline values with the randomization block as the clustering variable, and applied a sequential Hochberg correction method [30] to control the overall familywise error rate at .05 when testing the hypothesis for each individual primary end point. The results of these GLMMs indicated that participants in the intervention group had significantly greater reductions in anxiety (adjusted group mean difference=−0.85, 95% CI −1.27 to −0.42; *P*<.001; Cohen *d*=−0.17, 95% CI −0.26 to −0.09) and depressive (adjusted group mean difference=−0.63, 95% CI −1.08 to −0.17; *P*=.007; Cohen *d*=−0.11, 95% CI −0.19 to −0.03) symptoms than those in the control group (Table 2). Although participants in the intervention group also demonstrated a greater reduction in alcohol risk scores on the USAUDIT-C, this difference was not statistically significant (adjusted group mean difference=−0.13, 95% CI −0.34 to 0.08; *P*=.23).

We conducted the following analysis to quantify the robustness of primary findings to the assumption of data MAR. First, baseline covariates potentially predictive of missing data were added to the GLMM outcome models, and the intervention effect estimates remained similar and yielded qualitatively similar *P* values (Table 2, last 2 columns). Second, selection models were used that permit the missingness probability to depend on the unobserved outcome values after conditioning on the observed data, after which we computed the index of local sensitivity to nonignorability [34]. The index of local sensitivity to nonignorability analysis results are reported in Table 3. The *ISNI/SD* (ISNI divided by SD) column in Table 3 estimates the change in intervention effect estimates for a moderate size of nonrandom missingness, where a 1-SD (SD of the outcome) increase in the outcome is associated with an e^1=2.7–fold increase in the odds of being observed conditioned on the same values of the observed predictors for missingness. For such moderately sized nonrandom missingness, the changes in the intervention effect estimates were small for both GAD-7 and PHQ-9 scores (Table 3). The *MinNI* column in Table 3 computes the minimum magnitude of nonignorable missingness needed for substantial sensitivity such that the selection bias due to data missing not at random is of the same size as the SE. The smaller the value of the minimum nonignorable missingness, the greater the sensitivity. A minimum nonignorable missingness of 1 is suggested as the cutoff value for important sensitivity [32]. The minimum nonignorable missingness values for both the GAD-7 and PHQ-9 far exceeded 1, indicating that no sensitivity to potential missingness not at random was present for the primary findings. Table 3 shows that both the control and intervention groups had moderate and comparable missing data percentages for both the GAD-7 and PHQ-9, which can explain why our analysis results were insensitive to the MAR assumption.

XSL•FO
RenderX

Table 2. Results for the analysis of primary outcomes at baseline and at the 30-day follow-up.

Primary outcome and period	Unadjusted, mean (SD)		Adjusted difference[a] (primary analysis; 95% CI)	*P* value[b]	Adjusted difference[c] (robustness checking; 95% CI)	*P* value[d]
	Intervention (n=743)	Control (n=746)				
GAD-7[e]						
Baseline	9.0 (5.0)[f]	8.4 (5.0)[g]	N/A[h]	N/A	N/A	N/A
30 days	7.7 (5.1)[i]	8.2 (5.1)[j]	−0.85 (−1.27 to −0.42)	<.001	−0.82 (−1.25 to −0.39)	<.001
Effect size (Cohen *d*)	N/A	N/A	−0.17 (−0.26 to −0.09)	<.001	−0.16 (−0.25 to −0.08)	<.001
PHQ-9[k]						
Baseline	9.6 (5.9)[l]	8.9 (5.6)[m]	N/A	N/A	N/A	N/A
30 days	8.6 (6.2)[n]	8.8 (5.8)[o]	−0.63 (−1.08 to −0.17)	.007	−0.58 (−1.04 to −0.11)	.01
Effect size (Cohen *d*)	N/A	N/A	−0.11 (−0.19 to −0.03)	.007	−0.10 (−0.18 to −0.02)	.01
USAUDIT-C[p]						
Baseline	3.6 (3.2)[g]	3.4 (3.3)[q]	N/A	N/A	N/A	N/A
30 days	3.1 (3.0)[r]	3.1 (3.1)[s]	−0.13 (−0.34 to 0.08)	.23	−0.12 (−0.32 to 0.09)	.28

[a]Adjusting for block numbers and baseline outcome values.
[b]Global test: *P*=.001.
[c]Adjusting for age, gender, student status, race, substance use at baseline, block numbers, and baseline outcome values.
[d]Global test: *P*=.003.
[e]GAD-7: 7-item General Anxiety Disorder scale.
[f]n=741.
[g]n=740.
[h]N/A: not applicable.
[i]n=575.
[j]n=589.
[k]PHQ-9: 9-item Patient Health Questionnaire.
[l]n=732.
[m]n=742.
[n]n=563.
[o]n=587.
[p]USAUDIT-C: US Alcohol Use Disorders Identification Test–Consumption Scale.
[q]n=743.
[r]n=559.
[s]n=590.

XSL•FO
RenderX

Table 3. Robustness checking of group comparison of primary end points to the assumption of data missing at random[a].

Primary end points	Missing data at 30 days, n_m (%)		ISNI[b]/SD	MinNI[c]
	Intervention (n=743)	Control (n=746)		
GAD-7[d]	168 (22.6)	157 (21.0)	−0.0004	57.7
PHQ-9[e]	180 (24.2)	159 (21.3)	0.0020	10.3
USAUDIT-C[f]	184 (24.8)	156 (20.9)	0.0125	1.2

[a]In the missing data percentage (n_m/n), n_m is the number of participants not completing the outcome assessment at 30 days, and n is the number of participants in the intention-to-treat sample.

[b]ISNI: index of sensitivity to nonignorability; SD refers to the SD of the outcomes.

[c]MinNI: minimum magnitude of nonignorable missingness. A MinNI of <1 indicates sensitivity to data missing not at random, and a MinNI of >1 indicates no sensitivity to data missing not at random.

[d]GAD-7: 7-item General Anxiety Disorder scale.

[e]PHQ-9: 9-item Patient Health Questionnaire.

[f]USAUDIT-C: US Alcohol Use Disorders Identification Test–Consumption Scale.

Secondary Outcomes

The GLMM for clustered measures with the randomization block as the clustering variable was also used to test for differences between the intervention and control groups on secondary outcomes without any adjustment for multiple testing. The results of these GLMMs indicated that (Table 4), compared to those in the control group, participants in the intervention group had significantly greater improvements in mental well-being (adjusted mean difference=0.73, 95% CI 0.35-1.11; *P*<.001; Cohen *d*=0.17, 95% CI 0.08-0.26) and were significantly associated with a 20% reduction in their frequency of cannabis use (IRR=0.80, 95% CI 0.66-0.96; *P*=.02) and a 13% reduction in the typical number of drinks consumed when drinking (IRR=0.87, 95% CI 0.77-0.98; *P*=.03). No significant differences were found in frequency of binge drinking (IRR=0.98, 95% CI 0.86-1.13; *P*=.83), frequency of drinking (IRR=0.97, 95% CI 0.88-1.06; *P*=.48), or frequency of any opioid use (IRR=0.62, 95% CI 0.16-2.31; *P*=.48). The impact of the intervention on nonmedical stimulant use could not be assessed due to the small number of nonmedical stimulant users at baseline and follow-up.

XSL•FO
RenderX

Table 4. Results for the analysis of secondary outcomes at baseline and at the 30-day follow-up.

Secondary outcome and period	Intervention (n=743)	Control (n=746)	IRR[a] (95% CI)	*P* value
Warwick well-being[b], mean (SD)				
Baseline	21.3 (4.3)[c]	21.6 (4.1)[c]	N/A[d]	N/A
30 days	22.3 (4.5)[e]	21.8 (4.3)[f]	0.73 (0.35-1.11)	<.001
Effect size (Cohen *d*)	N/A	N/A	0.17 (0.08-0.26)	N/A
Cannabis use frequency, median category (IQR)				
Baseline	0 (0-2[g])[h]	0 (0-2[g])[c]	N/A	N/A
30 days	0 (0-2[g])[i]	0 (0-2[g])[j]	0.80 (0.66-0.96)	.02
Binge drinking frequency, median category (IQR)				
Baseline	1 (0-1[g])[k]	1 (0-1[g])[l]	N/A	N/A
30 days	1 (0-1[g])[m]	1 (0-1[g])[n]	0.98 (0.86-1.13)	.83
Drinking frequency, median category (IQR)				
Baseline	2 (0-3[g])[h]	1 (0-2[g])[o]	N/A	N/A
30 days	1 (0-3[g])[p]	1 (0-2[g])[q]	0.97 (0.88-1.06)	.48
Number of alcoholic drinks, median category (IQR)				
Baseline	1 (0-2[g])[h]	1 (0-2[g])[r]	N/A	N/A
30 days	0 (0-2[g])[s]	0 (0-2[g])[t]	0.87 (0.77-0.98)	.03
Stimulant use frequency, median category (IQR)				
Baseline	0 (0-0[g])[u]	0 (0-0[g])[v]	No analyses	N/A
30 days	0 (0-0[g])[w]	0 (0-0[g])[q]	N/A	N/A
Opioid use frequency, median category (IQR)				
Baseline	0 (0-0[g])[v]	0 (0-0[g])[o]	N/A	N/A
30 days	0 (0-0[g])[x]	0 (0-0[g])[y]	0.62 (0.16-2.31)	.48

[a]IRR: incidence rate ratio.
[b]Adjusted difference.
[c]n=737.
[d]N/A: not applicable.
[e]n=567.
[f]n=587.
[g]Refer to Multimedia Appendix 3 for the category definitions for each secondary measure.
[h]n=742.
[i]n=535.
[j]n=577.
[k]n=743.
[l]n=740.
[m]n=581.
[n]n=616.
[o]n=746.
[p]n=611.
[q]n=645.
[r]n=745.
[s]n=594.
[t]n=621.
[u]n=739.

XSL•FO
RenderX

[v]n=741.
[w]n=610.
[x]n=609.
[y]n=643.

Discussion

Principal Findings

Minder was codeveloped with students to provide them with a set of self-guided tools to manage their mental health and substance use. In this study, we found that participants in the intervention group who had access to the *Minder* app reported significantly greater average reductions in symptoms of anxiety (GAD-7) and depression (PHQ-9) than those in the control group. This finding aligns with the presentation of "stress and anxiety" and "sadness" as key topic areas within the *Minder* app and each topic having its own dedicated content island. Although these findings showed small effects on average in our sample, this may be related to the fact that, at baseline, only 38.9% (576/1481) of the participants had moderate or greater levels of anxiety and 43.8% (645/1474) had moderate or greater levels of depressive symptoms. Many participants in our sample reported only mild or no symptoms of anxiety or depression, which may have contributed to the finding of a small effect on average. Recent reviews on the effects of smartphone apps on anxiety and depression have found larger effects than those reported in this study; however, most studies recruit participants with clinical-level problems [12,13]. A meta-analysis examining internet interventions for university students found small effect sizes for anxiety and depression [36]. The *Minder* intervention group also demonstrated significant improvements in mental well-being in our analysis of secondary outcomes, which reflects a more general positive domain of mental health that may be more relevant to students without existing mental health concerns.

One of the key decisions in this trial was to include students with few or no symptoms of anxiety or depression. Providing interventions to nonclinical populations makes them more accessible to the large proportion of the population who may not meet clinical diagnostic criteria but may still experience occasional mental health challenges that can be addressed using existing tools (eg, via app-based cognitive behavioral therapy) [37]. In addition, this study used an ITT analysis that included all participants who were randomized regardless of whether they used the intervention. There was also no minimum amount of required content for participants to complete, nor were they remunerated for their use of the app. This pragmatic approach provided an approximation of the average effect of *Minder* on mental health and substance use outcomes in a university population if it were to be made available to everyone. As outlined in our study protocol, we plan to complement these findings with additional secondary analyses to examine the impact of *Minder* in subgroups of participants defined by the extent of their app use and baseline mental health and sociodemographic characteristics.

Although we did not find evidence of an effect on overall alcohol consumption risk in our primary outcome analyses of the USAUDIT-C, we did find significant reductions in cannabis use frequency and the typical number of alcoholic drinks consumed in a drinking session in our analysis of secondary outcomes. Reduction in number of drinks consumed when drinking and reducing alcohol-related harms were a main focus of the alcohol intervention content in the *Minder* app. For example, there was an activity in the app that encouraged users to set a goal for how many drinks they would consume in a drinking session and then track the number consumed in real time using the app. In addition, psychoeducational and motivational interviewing content focused on reducing harms associated with drinking, including following lower-risk drinking guidelines or cutting back on alcohol use. Similarly, cannabis use content focused on following lower-risk cannabis use guidelines, such as reducing the frequency of use to once a week or weekends [38]. We did not find significant reductions in measures of opioid use or nonmedical stimulant use frequency. However, there were low numbers of participants who used these substances in the study, particularly nonmedical stimulants. This finding may be related to the way in which the current *Minder* app allows users to select whatever content they think is most relevant to them regardless of their current mental health or substance use status. Substance use is often perceived by students to be higher than it actually is among their peers, thus normalizing its use on university campuses [39,40]. As a result, students may not be as motivated to address substance use compared to other aspects of their lives in which they may be experiencing distress. Previous studies have found that attitudes and norms surrounding drinking predict alcohol use behaviors among college students [41,42]. Addressing existing positive attitudes regarding commonly used substances (eg, alcohol and cannabis), as well as variations in norms such as for different genders [43], may be important in improving engagement with this content and tailoring the app to the needs of users. Future versions of the *Minder* app could include more nuanced approaches to address potentially harmful norms by providing tailored recommendations for substance use content within the app.

These findings are promising when considering the potential benefits of the *Minder* app as a tool in more comprehensive approaches to campus mental health that include early intervention and prevention. Given the extensive mental health needs of university students, stepped-care approaches have been identified as an efficient strategy to organize the delivery of campus mental health services in Canada [44]. The *Minder* intervention, which requires few resources (support is limited to online formats and provided by volunteer student coaches), could be readily integrated into such systems to support self-screening (via the existing services component and completion of the PHQ-9 and GAD-7) and the provision of immediate support and resources for students without higher levels of clinical concerns. It is also important to note that the *Minder* app was co-designed with students to connect users with the broader campus mental health systems through the Services component as well as with the greater student body

XSL•FO
RenderX

using the Community component. In this way, it can be used to strengthen the connections between these existing systems and fill gaps in services, particularly in the area of prevention and treatment for mild to moderate symptomology.

Strengths

A major strength of the *Minder* app is that it has been meaningfully codeveloped with university students and campus health care providers. Many mobile apps are not able to retain users after initial download and have low engagement rates overall [45]. Co-design processes have been identified as an effective means of ensuring that e-tools meet the needs of the end users they are trying to help [15]. The *Minder* app used an extensive codevelopment process that allowed for many improvements to be made with direct input from students and the clinicians currently supporting them [16]. The positive impact and low rates of loss to follow-up in this trial provide some support for the general acceptability of *Minder*. Previous studies on internet-based interventions for university students have also demonstrated their effectiveness in reducing mental health outcomes and that these interventions are generally acceptable to students; however, acceptability is often not reported [9].

Another strength of this study is the pragmatic trial design, which included a large nonclinical sample. Remuneration was only provided to participants for completion of the baseline and follow-up surveys, not for use of the intervention itself. Participants were also not told to use the app in any specific way or for any given amount of time when starting the study. Although this approach may have contributed to the finding of small average effects, it does increase the generalizability of our findings to other populations and provides an estimate of the impact of the app if it were made available to all students.

Limitations

Several limitations should be considered when interpreting the results of our study. As with many trials for online interventions, the participants were not blinded to what condition they were in, leading to a potential for placebo effects. However, control group participants did download the app and completed the surveys within it, which may have helped mitigate these effects on the trial. There were also several minor technical issues throughout the trial that may have impacted the participants' experience with the app; however, these were resolved quickly by the research team. An explanation of these issues can be found in Multimedia Appendix 1. There were some participants in the intervention group who did not use the app at all apart from completing the surveys. Given the ITT trial design, all randomized participants were included in the analysis; however, future analyses are planned to examine the effects of the intervention for those who actually engaged with the app content, along with the identification of subgroups that may benefit the most from this type of intervention.

Implications for Future Research

Further research will be conducted to better understand how to optimize the *Minder* intervention for different needs and additional populations. By better understanding who benefitted from the intervention and what content they used, we plan to make the app more personalized (including recommendations and features). In addition, future codevelopment processes will be needed to further improve app use and incorporation into existing systems of care. There were some participants who did not use the app outside the surveys, so trying to make the intervention more appealing for these users will be important. This may involve gamification strategies, refinement of content and enhancement of the chatbot using artificial intelligence tools, and the development of new features.

Conclusions

The *Minder* app was effective in reducing symptoms of anxiety and depression, with provisional support for increasing mental well-being and reducing the frequency of cannabis and alcohol use in a general population of university students. These findings support our use of a codevelopment approach and provide evidence of the potential of digital intervention tools such as *Minder* to support prevention and early intervention efforts for university students.

Acknowledgments

This work was supported by the Health Canada Substance Use and Addictions Program (arrangement: 1920-HQ-000069; University of British Columbia ID: F19-02914).

Conflicts of Interest

For the past 3 years, RCK has been a consultant for the Cambridge Health Alliance; Canandaigua Veterans Affairs Medical Center; Holmusk; Partners Healthcare, Inc; RallyPoint Networks, Inc; and Sage Therapeutics. He has stock options in Cerebral Inc, Mirah, PYM, Roga Sciences, and Verisense Health.

Multimedia Appendix 1
Technical issues and protocol changes.
[DOCX File , 17 KB - jmir_v26i1e54287_app1.docx]

Multimedia Appendix 2
Chatbot activity content overview.
[DOCX File , 20 KB - jmir_v26i1e54287_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Identified secondary outcomes with scoring.
[DOCX File , 17 KB - jmir_v26i1e54287_app3.docx]

Multimedia Appendix 4
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1731 KB - jmir_v26i1e54287_app4.pdf]

References

1. Cunningham S, Duffy A. Investing in our future: importance of postsecondary student mental health research. Can J Psychiatry 2019 Feb;64(2):79-81 [FREE Full text] [doi: 10.1177/0706743718819491] [Medline: 30819000]
2. Chan V, Moore J, Derenne J, Fuchs DC. Transitional age youth and college mental health. Child Adolesc Psychiatr Clin N Am 2019 Jul;28(3):363-375 [FREE Full text] [doi: 10.1016/j.chc.2019.02.008] [Medline: 31076114]
3. de Girolamo G, Dagani J, Purcell R, Cocchi A, McGorry PD. Age of onset of mental disorders and use of mental health services: needs, opportunities and obstacles. Epidemiol Psychiatr Sci 2012 Mar;21(1):47-57. [doi: 10.1017/s2045796011000746] [Medline: 22670412]
4. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustün TB. Age of onset of mental disorders: a review of recent literature. Curr Opin Psychiatry 2007 Jul;20(4):359-364 [FREE Full text] [doi: 10.1097/YCO.0b013e32816ebc8c] [Medline: 17551351]
5. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. WHO World Mental Health Surveys International College Student Project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
6. Duffy A, Keown-Stoneman C, Goodday S, Horrocks J, Lowe M, King N, et al. Predictors of mental health and academic outcomes in first-year university students: identifying prevention and early-intervention targets. BJPsych Open 2020 May 08;6(3):e46 [FREE Full text] [doi: 10.1192/bjo.2020.24] [Medline: 32381150]
7. Eisenberg D, Hunt J, Speer N, Zivin K. Mental health service utilization among college students in the United States. J Nerv Ment Dis 2011 May;199(5):301-308. [doi: 10.1097/NMD.0b013e3182175123] [Medline: 21543948]
8. Ebert DD, Mortier P, Kaehlke F, Bruffaerts R, Baumeister H, Auerbach RP, et al. Barriers of mental health treatment utilization among first-year college students: first cross-national results from the WHO World Mental Health International College Student Initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1782 [FREE Full text] [doi: 10.1002/mpr.1782] [Medline: 31069905]
9. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
10. Carey KB, Scott-Sheldon LA, Elliott JC, Bolles JR, Carey MP. Computer-delivered interventions to reduce college student drinking: a meta-analysis. Addiction 2009 Nov;104(11):1807-1819 [FREE Full text] [doi: 10.1111/j.1360-0443.2009.02691.x] [Medline: 19744139]
11. Newfoundland and Labrador stepped care 2.0© e-mental health demonstration project final report. Mental Health Commission of Canada. URL: https://www.mentalhealthcommission.ca/wp-content/uploads/drupal/2019-09/emental_health_report_eng_0.pdf [accessed 2023-09-28]
12. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
13. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? a meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
14. Mei C, Nelson B, Hartmann J, Spooner R, McGorry PD. Transdiagnostic early intervention, prevention, and prediction in psychiatry. In: Personalized Psychiatry. London, UK: Academic Press; 2020:27-37.
15. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
16. Vereschagin M, Wang AY, Leung C, Richardson CG, Hudec KL, Doan Q, et al. Co-developing tools to support student mental health and substance use: minder app development from conceptualization to realization. J Behav Cogn Ther 2023 Mar;33(1):35-49. [doi: 10.1016/j.jbct.2023.02.002]
17. Wang AY, Vereschagin M, Richardson CG, Xie H, Hudec KL, Munthali RJ, et al. Evaluating the effectiveness of a codeveloped e-mental health intervention for university students: protocol for a randomized controlled trial. JMIR Res Protoc 2023 Aug 30;12:e49364 [FREE Full text] [doi: 10.2196/49364] [Medline: 37647105]

XSL•FO
RenderX

18. Jones LB, Judkowicz C, Hudec KL, Munthali RJ, Prescivalli AP, Wang AY, et al. The World Mental Health International College Student Survey in Canada: protocol for a mental health and substance use trend study. JMIR Res Protoc 2022 Jul 29;11(7):e35168 [FREE Full text] [doi: 10.2196/35168] [Medline: 35696337]
19. Randomisation and online databases for clinical trials. Sealed Envelope. URL: https://www.sealedenvelope.com/ [accessed 2023-09-28]
20. Virk P, Arora R, Burt H, Gadermann A, Barbic S, Nelson M, et al. HEARTSMAP-U: adapting a psychosocial self-screening and resource navigation support tool for use by post-secondary students. Front Psychiatry 2022 Feb 22;13:812965 [FREE Full text] [doi: 10.3389/fpsyt.2022.812965] [Medline: 35280181]
21. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
22. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
23. Higgins-Biddle JC, Babor TF. A review of the Alcohol Use Disorders Identification Test (AUDIT), AUDIT-C, and USAUDIT for screening in the United States: past issues and future directions. Am J Drug Alcohol Abuse 2018;44(6):578-586 [FREE Full text] [doi: 10.1080/00952990.2018.1456545] [Medline: 29723083]
24. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction 1993 Jun;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
25. Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health education population survey. Health Qual Life Outcomes 2009 Feb 19;7:15 [FREE Full text] [doi: 10.1186/1477-7525-7-15] [Medline: 19228398]
26. Guo L, Qian Y, Xie H. Assessing complier average causal effects from longitudinal trials with multiple endpoints and treatment noncompliance: an application to a study of Arthritis Health Journal. Stat Med 2022 Jun 15;41(13):2448-2465 [FREE Full text] [doi: 10.1002/sim.9364] [Medline: 35274333]
27. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984 Dec;40(4):1079-1087. [doi: 10.2307/2531158]
28. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics 1987 Sep;43(3):487-498. [doi: 10.2307/2531989]
29. Teixeira-Pinto A, Siddique J, Gibbons R, Normand SL. Statistical approaches to modeling multiple outcomes in psychiatric studies. Psychiatr Ann 2009 Jul 01;39(7):729-735 [FREE Full text] [doi: 10.3928/00485713-20090625-08] [Medline: 20161512]
30. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988 Dec;75(4):800-802. [doi: 10.1093/biomet/75.4.800]
31. Diggle P, Heagerty P, Liang KY, Zeger S. Analysis of Longitudinal Data. Oxford, UK: Oxford University Press; 2013.
32. Jing B, Qian Y, Heitjan DF, Xie H. Tutorial: assessing the impact of nonignorable missingness on regression analysis using index of local sensitivity to nonignorability. Psychol Methods 2023 Nov 16 (forthcoming). [doi: 10.1037/met0000616] [Medline: 37971833]
33. Xie H, Qian Y. Measuring the impact of nonignorability in panel data with non-monotone nonresponse. J Appl Econom 2012 Jan 18;27(1):129-159. [doi: 10.1002/jae.1157]
34. Xie H, Gao W, Xing B, Heitjan DF, Hedeker D, Yuan C. Measuring the impact of nonignorable missingness using the R package isni. Comput Methods Programs Biomed 2018 Oct;164:207-220 [FREE Full text] [doi: 10.1016/j.cmpb.2018.06.014] [Medline: 30195428]
35. Stata Corp. Stata statistical software: release 15.1. StataCorp LLC. 2017. URL: https://www.stata.com/stata15/ [accessed 2024-03-11]
36. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]
37. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 01;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]
38. Canada's lower-risk cannabis use guidelines. Government of Canada. URL: https://www.canada.ca/en/health-canada/services/drugs-medication/cannabis/resources/lower-risk-cannabis-use-guidelines.html [accessed 2023-09-28]
39. Baer JS, Stacy A, Larimer M. Biases in the perception of drinking norms among college students. J Stud Alcohol 1991 Nov;52(6):580-586. [doi: 10.15288/jsa.1991.52.580] [Medline: 1758185]
40. Lewis MA, Neighbors C. Gender-specific misperceptions of college student drinking norms. Psychol Addict Behav 2004 Dec;18(4):334-339. [doi: 10.1037/0893-164X.18.4.334] [Medline: 15631605]
41. DiBello AM, Miller MB, Neighbors C, Reid A, Carey KB. The relative strength of attitudes versus perceived drinking norms as predictors of alcohol use. Addict Behav 2018 May;80:39-46 [FREE Full text] [doi: 10.1016/j.addbeh.2017.12.022] [Medline: 29329007]

XSL•FO
RenderX

42. Neighbors C, Lee CM, Lewis MA, Fossos N, Larimer ME. Are social norms the best predictor of outcomes among heavy-drinking college students? J Stud Alcohol Drugs 2007 Jul;68(4):556-565 [FREE Full text] [doi: 10.15288/jsad.2007.68.556] [Medline: 17568961]
43. Hathaway A, Mostaghim A, Kolar K, Erickson PG, Osborne G. A nuanced view of normalisation: attitudes of cannabis non-users in a study of undergraduate students at three Canadian universities. Drugs Educ Prev Policy 2016;23(3):238-246. [doi: 10.3109/09687637.2015.1112362]
44. Duffy A, Saunders KE, Malhi GS, Patten S, Cipriani A, McNevin SH, et al. Mental health care for university students: a way forward? Lancet Psychiatry 2019 Nov;6(11):885-887. [doi: 10.1016/S2215-0366(19)30275-5] [Medline: 31324561]
45. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]

Abbreviations

AUDIT: Alcohol Use Disorders Identification Test
GAD-7: 7-item General Anxiety Disorder scale
GLMM: generalized linear mixed-effects model
IRR: incidence rate ratio
ITT: intention-to-treat
MAR: missing at random
PHQ-9: 9-item Patient Health Questionnaire
UBC: University of British Columbia
USAUDIT-C: US Alcohol Use Disorders Identification Test–Consumption Scale

Edited by G Eysenbach; submitted 05.11.23; peer-reviewed by S Gordon; comments to author 30.11.23; revised version received 05.12.23; accepted 17.02.24; published 27.03.24.

Please cite as:
Vereschagin M, Wang AY, Richardson CG, Xie H, Munthali RJ, Hudec KL, Leung C, Wojcik KD, Munro L, Halli P, Kessler RC, Vigo DV
Effectiveness of the Minder Mobile Mental Health and Substance Use Intervention for University Students: Randomized Controlled Trial
J Med Internet Res 2024;26:e54287
URL: https://www.jmir.org/2024/1/e54287
doi:10.2196/54287
PMID:38536225

©Melissa Vereschagin, Angel Y Wang, Chris G Richardson, Hui Xie, Richard J Munthali, Kristen L Hudec, Calista Leung, Katharine D Wojcik, Lonna Munro, Priyanka Halli, Ronald C Kessler, Daniel V Vigo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Digital Self-Monitoring of Weight on Improving Diabetes Clinical Outcomes: Quasi-Randomized Study

Yifat Fundoiano-Hershcovitz[1], PhD; Marilyn D Ritholz[2], PhD; David L Horwitz[3], MD, PhD; Ephraim Behar[1], MSc; Omar Manejwala[1], MD; Pavel Goldstein[4], PhD

[1]Dario Health, Caesarea, Israel
[2]Joslin Diabetes Center, Harvard Medical School, Boston, MA, United States
[3]DLH Biomedical Consulting, Las Vegas, NV, United States
[4]School of Public Health, University of Haifa, Haifa, Israel

Corresponding Author:
Yifat Fundoiano-Hershcovitz, PhD
Dario Health
Tarshish 5
Caesarea, 3079821
Israel
Phone: 972 525296979
Email: yifat@dariohealth.com

Abstract

Background: The management of type 2 diabetes (T2D) and obesity, particularly in the context of self-monitoring, remains a critical challenge in health care. As nearly 80% to 90% of patients with T2D have overweight or obesity, there is a compelling need for interventions that can effectively manage both conditions simultaneously. One of the goals in managing chronic conditions is to increase awareness and generate behavioral change to improve outcomes in diabetes and related comorbidities, such as overweight or obesity. There is a lack of real-life evidence to test the impact of self-monitoring of weight on glycemic outcomes and its underlying mechanisms.

Objective: This study aims to assess the efficacy of digital self-monitoring of weight on blood glucose (BG) levels during diabetes management, investigating whether the weight changes may drive glucose fluctuations.

Methods: In this retrospective, real-world quasi-randomized study, 50% of the individuals who regularly used the weight monitoring (WM) feature were propensity score matched with 50% of the users who did not use the weight monitoring feature (NWM) based on demographic and clinical characteristics. All the patients were diagnosed with T2D and tracked their BG levels. We analyzed monthly aggregated data 6 months before and after starting their weight monitoring. A piecewise mixed model was used for analyzing the time trajectories of BG and weight as well as exploring the disaggregation effect of between- and within-patient lagged effects of weight on BG.

Results: The WM group exhibited a significant reduction in BG levels post intervention ($P<.001$), whereas the nonmonitoring group showed no significant changes ($P=.59$), and both groups showed no differences in BG pattern before the intervention ($P=.59$). Furthermore, the WM group achieved a meaningful decrease in BMI ($P<.001$). Finally, both within-patient ($P<.001$) and between-patient ($P=.008$) weight variability was positively associated with BG levels. However, 1-month lagged back BMI was not associated with BG levels ($P=.36$).

Conclusions: This study highlights the substantial benefits of self-monitoring of weight in managing BG levels in patients with diabetes, facilitated by a digital health platform, and advocates for the integration of digital self-monitoring tools in chronic disease management. We also provide initial evidence of testing the underlying mechanisms associated with BG management, underscoring the potential role of patient empowerment.

(J Med Internet Res 2024;26:e54940) doi:10.2196/54940

KEYWORDS

obesity; diabetes management; weight monitoring; digital health platform; self-monitoring; clinical outcome; type 2 diabetes; weight changes; blood glucose; patient empowerment

XSL•FO
RenderX

Introduction

Background

People with type 2 diabetes (T2D) face challenging self-management regimens to improve glycemia and decrease morbidity and mortality while often dealing with high costs of care [1]. Obesity is one of the most common, serious, and costly medical condition in the United States, with a prevalence of 41.9% from 2017 to 2020 [2]. After a dramatic increase in its prevalence over several decades, obesity has become a major public health crisis in the United States [3]. Obesity has become one of the leading causes of death, as it is known to be the main risk factor for several noncommunicable diseases, particularly T2D [4]. It is crucial to take effective and decisive actions to hinder both the rise in the prevalence of obesity and the prevention and treatment of obesity and other obesity-related comorbidities. Approximately 80% to 90% of patients with T2D have overweight or obesity, which imposes a considerable burden on individuals, families, communities, and the health system [5,6].

Obesity and overweight are considered the primary *accelerators* for the T2D inflammatory component inducing progressive loss of beta cell insulin secretion with coexisting insulin resistance [7-9]. In addition, the expansion of white adipose tissue is related to a changed microenvironment in obesity, which impairs insulin signaling, reduces insulin-stimulated glucose transport activity, and accelerates beta cell dysfunction [10].

Previous studies have shown the beneficial effect of weight-lowering treatment on diabetes outcomes [11].

Healthful weight reduction in patients with obesity can improve glucose metabolism [12]. Weight reduction via carbohydrate-restricted nutritional intervention in patients with preobesity or obesity and prediabetes or T2D may contribute to improvement or remission in diabetes mellitus [13].

Antiobesity therapies for the treatment of patients with obesity and T2D include those that reduce body weight and improve glucose levels and other metabolic parameters. Considering the prevalence of obesity-related conditions such as adiposopathy and the fact that a significant portion of patients in cardiovascular outcomes trials for T2D had overweight or obesity, there is support for the “treat obesity first” therapeutic approach [13]. It is recommended in the guidelines for obesity that appropriate 5% to 10% weight loss can achieve significant metabolic improvement [14]. For the prevention of T2D, even modest weight reduction as little as 5%, can significantly reduce diabetes-associated complications [15,16]. Previous studies have shown that changes in various indexes such as blood lipid, blood glucose (BG), and insulin improved when weight loss reached 15% [14]. Furthermore, long-term tight weight control resulted in significant glycemic improvement, particularly demonstrated in the overweight population with T2D [17,18].

One of the goals of chronic condition management is to increase awareness and generate behavioral change to improve clinical outcomes. Behavior change for effective self-management was proven to improve health outcomes and quality of life in people living with chronic conditions such as obesity, T2D, and heart disease [19]. Underlying well-intentioned lifestyle messages is the assumption that if people deem health important, are aware of exercise and nutrition guidelines, and have access to healthy options to maintain proper levels of nutrition, diet, and exercise, then they will make healthier choices [20]. The American Diabetes Association guidelines state that lifestyle management should be intensive and involve frequent follow-ups [21].

Despite these recommendations, data from the National Health and Nutrition Examination Survey indicate that only 54.6% of patients reported receiving any diabetes education and only 13.4% had received an educational visit of any kind [16]. Earlier studies showed that helping participants with goal setting and self-monitoring of behavior, for instance, using a logbook and receiving feedback on the outcome of behavior, was associated with better intervention effects [22].

Facilitating behavior change involves using a series of strategies aimed at empowering patients, enabling them to take increasing control of their condition. This includes setting clear, achievable, and personalized goals, as well as enhancing self-efficacy [23-25]. The timing of health information and feedback focuses on when health behavior messages are delivered to people with diabetes. As diabetes care visits usually take place every 3 months, there can be a significant gap between these appointments and the daily engagement in desired behaviors. This gap makes it challenging to offer timely behavioral prompts or reinforcement [26].

In fact, patients’ mindset may modulate health outcomes, including glucose levels, in patients with diabetes [27]. Indeed, increasing perceived self-monitoring would be expected to result in subsequent health benefits [28], including glucose control in diabetes [29]. Individuals possess significant psychological influence over their health [30].

Currently, traditional health care models are being revamped with digital technologies. Digital platforms have the potential to improve our ability to enhance the delivery of health care for individual patients as well as empower patients to have more control over, and make better-informed decisions about, their health. Treatment optimization through digital health could enhance users’ alertness to their health condition through real-time monitoring, leading to effective treatments that build awareness of their daily health-related behaviors and promote increased engagement with those behaviors [31-34]. Technology-driven solutions can help people with diabetes build awareness of their daily health-related behaviors and promote increased engagement with those behaviors [32-34].

Communication of test results has been shown to be highly desired by people who have overweight, and lifestyle-focused educational messages providing advice, motivational reminders, and support have also been shown to be effective in improving chronic conditions [35]. Using a mobile platform for self-management purposes could facilitate individuals with chronic conditions in gaining insight into and controlling their BG and weight levels. Self-monitoring is a core component of behavioral obesity treatment; however, it is unknown how digital health has been used for self-monitoring and what engagement rates are achieved in these interventions [36].

XSL•FO
RenderX

Mobile apps have been shown to improve diabetes outcomes via education and support for adherence to evidence-based recommendations [37-40]. Mobile technology has emerged as a potentially useful platform to facilitate weight management [41]. Mobile apps for weight management typically offer similar features, including self-monitoring of diet and physical activity. Users can set goals within specified time frames and input data into the app, often receiving reminders or text messages. These apps have shown promising results [41]. Numerous digital health technologies have been developed to support the self-management of single chronic diseases, primarily diabetes. These technologies provide timely feedback, enhance patient education, and support the behavioral changes necessary for effective weight management. Recent research has indicated that digital self-monitoring tools can significantly influence health behaviors in patients with T2D, leading to better management of their condition [42,43]. However, given the rise in the number of people managing multiple chronic conditions, it is imperative to design and implement digital health technologies to deal with the additional complexities of multiple chronic conditions, such as the management of multiple symptoms and self-management tasks, avoiding further burden or inconvenience to the user [44-46]. Integrating the management of multiple conditions onto a single platform, where users can monitor their measurements and relevant lifestyle parameters, interact with all their data, share their data, and receive educational support, could help to minimize the known burden of multimorbidity self-management [47-49].

However, there is limited research on platforms that have been implemented to tackle multimorbidity or evaluated over longitudinal periods [45]. Specifically, the current literature is missing rigorous real-life studies to test the role of a simple self-monitoring of weight and diabetes management platform to better understand the direct association between weight monitoring and glycemic outcomes. Mainly, data are lacking on whether more frequent self-monitoring of those 2 conditions (weight and glycemia) has any impact on body weight and glycemic control in real-world clinical practice among patients with T2D and obesity [50]. In addition, many of these exclusively weight loss programs are time consuming and costly [51].

Objectives

Our study seeks to address this gap by exploring the efficacy of digital self-monitoring of weight in managing BG levels in patients with T2D who are also managing their weight. We used a retrospective analysis of a home-use digital platform containing a diabetes BG meter and weight monitoring system with full longitudinal data capture using a supportive mobile platform among people with T2D and overweight levels. We followed users for 6 months before and 6 months after using the app for self-monitoring of weight and compared them with a matched control group that never used weight monitoring on the platform. We hypothesized that self-monitoring of weight would result in a significant improvement in BG levels. Moreover, weight monitoring (WM) will be associated with a reduction in weight levels. We also hypothesized a linkage between the changes in weight levels and the reduction in BG levels.

Methods

Platform

This study used the Dario Health digital therapeutics solution for chronic conditions to support the self-management of BG and weight levels. The platform combines an innovative meter with a phone app that is available for both Android and iOS devices. The glucose meter consists of a small pocket-sized holder for strips, a lancet, and the meter. The meter is removed from the holder and plugged directly into a cell phone, effectively converting the cell phone into the display screen for the meter. Weight level monitoring data are logged manually into the app on a special data entry screen (Figure 1).

Connecting the BG meter directly to the phone and adding weight levels improves the quality of data collection. Additional information for weight measurement includes an informative color scale of weight ranges reflecting the Centers for Disease Control and Prevention definitions for BMI (kg/m^2) interpretation: <18.5 is underweight, 18.5 to 24.9 is healthy weight, 25.0 to 29.9 is overweight, and ≥30.0 is obese [52]. All information is stored in the users' logbook in the app "attached" to the specific BG or weight reading. Data are uploaded to the cloud for backup and further analysis. Digital platform functions include interface design elements as well as specific educational content, wording, or digital interventions that affect the users' choices in the digital environment; these functions provide personal health information and prompt feedback.

XSL•FO
RenderX

Figure 1. Dario mobile app platform: (A) the data entry screen allows adding the weight and height measurements to monitor where it falls within the BMI range and (B) the logbook screen presenting the blood glucose and weight measurements.

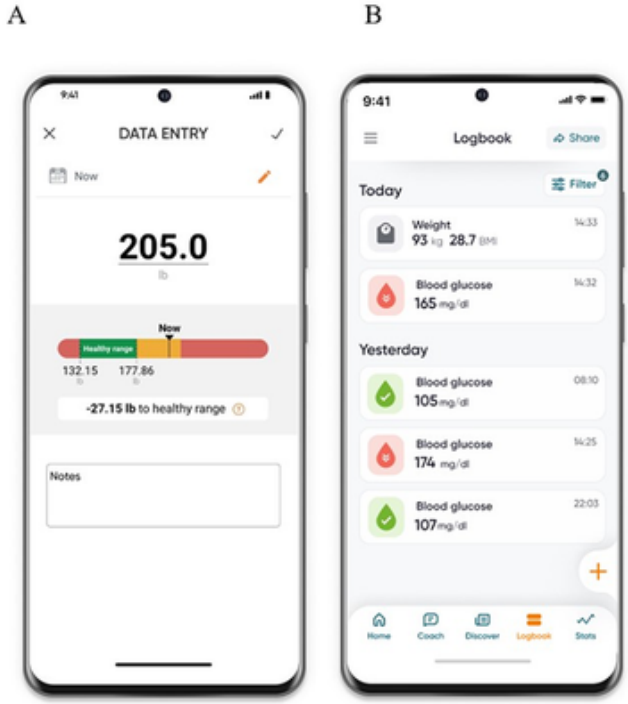

Measures

The monthly average weight level, which was defined as the means of all of a user's weight measurements taken over a 30-day interval, was used as the core outcome metric. The monthly average BG level, which was defined as the meaning of all of a user's BG measurements taken over a 30-day interval, was used as another outcome metric. The mobile platform collected the following medical and sociodemographic information (by self-report) for each user: sex, age, BMI during registration, physical activity level, stress level (0=no stress and 10=very stressed), alcohol consumption (number of drinks per week), smoking (0=never and 3=yes), and added comorbidities (such as high lipids, chronic kidney disease, cardiovascular disease, sleep disorder, cancer, or stress and depression). Socioeconomic status (median household income) was matched by applying zip code data to Census [53] and app engagement (number of app sessions per month). All data were transferred and stored in compliance with the Health Insurance Portability and Accountability Act requirements using Amazon Web Services database solutions. All data were anonymized before extraction for this study.

Study Population

A retrospective data study was performed on the Dario database on individuals who used the Dario platform between 2017 and 2023. The users purchased the device via a direct-to-consumer channel. The inclusion criteria were as follows: individuals who reported in the Dario app as diagnosed with T2D with a BG level of >140 mg/100 mL and BMI >25 kg/m^2 in their first month on the platform (baseline) and weight monitoring system (WM group) and used the weight monitoring system (WM monitoring). The resulting group of users was matched through the propensity scores procedure with users with similar clinical parameters but who have not been using the weight monitoring (non–weight monitoring; [NWM] group).

Study Design

The aim of our study design was to evaluate the impact of weight monitoring on BG levels. For the WM group, it was crucial to establish a clear start point for weight monitoring to assess its effects accurately. This start point is a defined intervention onset, marking when participants began actively monitoring their weight using the digital platform.

Conversely, for the NWM group, such a start point for "nonintervention" does not inherently exist, as these participants did not engage in weight monitoring. Hence, selecting a random start point for this group was a methodological necessity. This approach ensures that any observed differences in outcomes are attributable to the act of weight monitoring itself, rather than temporal factors or external influences. Therefore, the comparison between the groups hinges on the presence or absence of weight monitoring behavior. Using this approach, we enhanced the internal validity of the study. This allowed us to isolate the effect of weight monitoring from that of other variables and assess its impact on BG levels more accurately.

Propensity Scores: Causal Inference

Propensity score matching was used in this study to address potential confounding factors and enhance the comparability of the WM and NWM groups. The rationale behind using propensity score matching lies in its ability to reduce bias and mimic the randomization process, thereby facilitating causal inference in observational studies [54].

In originally nonrandomized studies, it is common for treatment assignment (in this case, use of the weight monitoring system) to be influenced by patient characteristics and other confounding

XSL•FO
RenderX

variables. These factors may introduce bias and affect the estimation of treatment effects. Propensity score matching offers a systematic approach to account for such biases and create comparable treatment and control groups [55].

The propensity score, defined as the conditional probability of receiving the treatment given a set of observed covariates, summarizes the individual's likelihood of being assigned to the WM group. By incorporating a comprehensive set of covariates that are potential confounders, such as age, sex, initial BG and BMI levels, smoking status, alcohol consumption, stress level, comorbidities, median household income, and platform engagement, the propensity score attempts to balance the distribution of these covariates between the WM and NWM groups.

Matching participants based on their propensity scores allows a comparison between similar individuals who only differ in terms of the treatment received. This strategy helps to reduce selection bias and confounding effects, enabling a more valid estimation of the causal effect of weight monitoring on glycemic control.

The use of propensity score matching aligns with the principle of exchangeability, as it creates groups that are comparable in terms of observed characteristics. By achieving a balance on observed covariates, the propensity score matching enhances the internal validity of the study and strengthens the plausibility of causal inference from the observed associations [56].

In this study, the propensity scores were calculated for each participant using the "matchit()" function from the R package *matchit*, which followed a nearest-neighbor approach, and the distance metric used was based on logistic regression using a 1:1 ratio between the 2 study groups [57].

To achieve balanced groups, nearest-neighbor matching with a caliper width of 0.1 SDs of the propensity score was applied. The matching procedure aimed to identify, for each WM user, a corresponding NWM participant with the closest propensity score. Participants without suitable matches were excluded from the analysis. Figure 2 presents the efficacy of the matching procedure for balancing the groups. A caliper width of 0.05 SDs was reached for all the parameters except alcohol consumption which remained within 0.1 SDs.

Figure 2. The plot presents the efficacy of the matching procedure for balancing the groups. A caliper width of 0.05 SDs was reached for all the parameters except alcohol consumption, which remained within 0.1 SDs. The variable labels are as follows: monthly average BG value=monthly average blood glucose value, number BG measurements=monthly number of blood glucose measurements, baseline BMI=starting BMI value, age=reported age, sex=reported sex of the users, comorbidities=sum of reported comorbidities (hypertension, high blood lipids, kidney disease, cardiovascular disease, sleep disorder, cancer, depression, and no comorbidity), smoking=reported smoking, alcohol consumption=reported alcohol consumption, activity level=reported level of daily activity, stress level=reported level of stress, median household income=median household income matched based on zip code, and app engagement=number of monthly app sessions. BG: blood glucose.

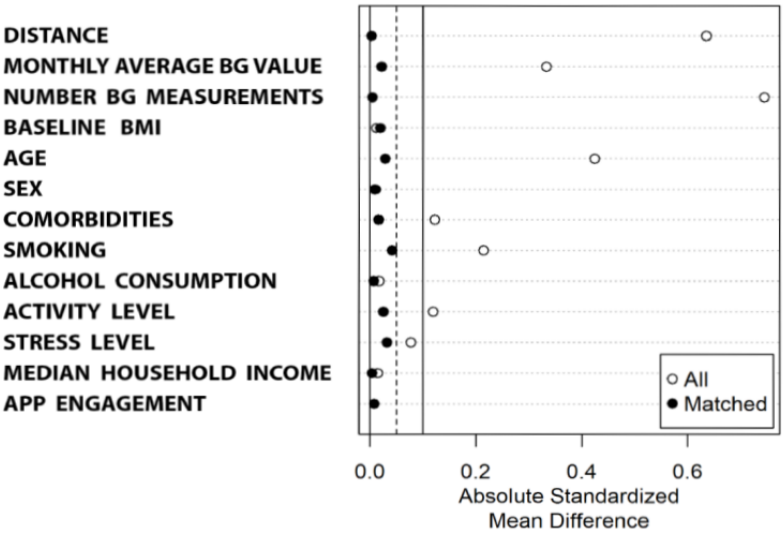

Ethical Considerations

All data used for the analysis were anonymized before extraction for this study. The study received an exemption from the institutional review board under the Ethical and Independent Review Services, a professional review board, which issued the institutional review board exemption for this study (18032-06#) [58]. The users who participated in the study were provided with a Terms of Use document mentioning the legally valid consent of the end user for the company to collect and access their information. The use of the app, site, or services shall be deemed to constitute user consent to be legally bound by these Terms and the Privacy Policy. Please refer to the current "Terms-of-use" at the link [59].

Analytic Approach

Traditionally, a classical linear longitudinal model assumes a single-slope growth pattern for changes in an outcome variable over time. However, empirical data often exhibit more complex patterns that cannot be adequately captured by such a simple model. In our study, we used piecewise-based mixed-effects models to allow for greater flexibility in modeling trajectories over time [60].

The piecewise model approach offers the advantage of accommodating different linear trends in the data across distinct regions. Specifically, we used a mixed piecewise model to assess differential changes in the WM and NWM groups in the monthly average BG level in 2 segments: before and after weight

XSL•FO
RenderX

monitoring use. This approach enabled us to capture the potential changes in BG trajectories associated with the introduction of the monitoring system. Using a statistical model that accounted for curvilinear changes, we were able to test the complex effects and capture the dynamics of the associated BG fluctuations.

For the analysis, the time data were centered around the beginning of the weight monitoring period. We included a 6-month timeframe before and after this point to capture the potential impact of weight monitoring use. In the NWM group, we randomly selected a cutoff point and included data collected during the 6 months before and after the simulated cutoff point. To model the temporal changes in the monthly average weight level between the WM and NWM groups, we fitted a piecewise-based mixed-effects model. The piecewise cutoff point was set at the initiation of weight monitoring, assuming a change in the time-related trajectory of the monthly average BG level between the 2 groups. We incorporated interaction terms between the time trajectories and groups to capture this differential effect. Thus, 2 time parameters (pre and postintervention) were used as covariates, the groups (WM and NWM) were considered as a factor, and the monthly number of BG measurement served as a potential confounding variable. All the tests were 2-tailed and the type 1 error was set to 5%. The model included random intercepts and random slopes for the time trajectory after the piecewise cutoff, accounting for individual variability in BG changes.

In addition, we used mixed model analysis to examine the time trajectory of BMI changes (covariate) for the initial 6 months of weight monitoring in the WM group, controlling for baseline BMI and the number of monthly BMI measures as confounding variables. These models included random intercepts and random slopes of the time trajectory to capture individual variations in weight changes over time. Unstandardized regression weights (B), test statistics (*t*), and associated significance (*P* values) were reported.

Finally, the monthly BMI levels were disaggregated to separate within- and between-person variabilities using person-level centering and person-level aggregation [61]. In addition, a 1-month lagged within-person BMI was calculated. Thereafter, a mixed model was applied to test the 1-month lagged and simultaneous association of monthly within-person BMI changes and between-person BMI with the monthly average BG level. All the model predictors were defined as covariates.

Results

Users

In total, 1932 users were included in the study. The WM group included 50% users, and the NWM group, matched through the propensity scores procedure, included 50% users. The study cohort comprised 51.6% (997/1932) of men, and 60.82% (1175/1932) of the participants had comorbidities. The average age of the participants was 62.8 (SD 12.5) years, with an average BMI of 35.4 (SD 7.3). The median household income for the participants was US $68,200 (SD US $25,100). The distribution of the other parameters is presented in Table 1 by study group.

No differences were found between the WM and NWM groups. The study included individuals with diabetes who monitored their BG levels and weight using the Dario platform.

The distribution of various sample characteristics overall and by WM and NWM groups is presented in Table 1, and any significant differences were shown.

Table 1. Distribution of various sample characteristics overall and by WM[a] and NWM[b] groups.

	NWM (n=966)	WM (n=966)	Total (N=1932)	*P* value
Monthly BG[c], mean (SD)	170 (40)	180 (41)	175 (40)	.62
Number of BG measurements, mean (SD)	39 (33)	39 (29)	39 (31)	.92
Baseline BMI, mean (SD)	36 (7.6)	35 (7.0)	35 (7.3)	.68
Age (years), mean (SD)	63 (12)	63 (13)	63 (12)	.52
Sex, n (%)				.86
Male	496 (51.3)	501 (51.9)	997 (51.60)	
Female	470 (48.7)	465 (48.1)	935 (48.39)	
Number of comorbidities, mean (SD)	1.1 (1.1)	1.1 (1.2)	1.1 (1.2)	.71
Smoking, mean (SD)	0.25 (0.44)	0.27 (0.45)	0.26 (0.44)	.35
Alcohol consumption, mean (SD)	1.4 (3.3)	1.3 (3.4)	1.3 (3.4)	.88
Physical activity level, mean (SD)	4.1 (2.4)	4.2 (2.2)	4.1 (2.3)	.59
Stress level, mean (SD)	5.1 (2.5)	5.1 (2.5)	5.1 (2.5)	.48
Median household income, mean (SD)	68,000 (25,000)	68,000 (25,000)	68,000 (25,000)	.94
App engagement, mean (SD)	56 (53)	56 (52)	56 (52)	.86

[a]WM: weight monitoring.

[b]NWM: non–weight monitoring.

[c]BG: blood glucose.

XSL•FO
RenderX

Weight Monitoring Is Associated With BG Levels

The results from the piecewise mixed model analysis indicated a significant interaction between the time trajectory, starting weight monitoring and the group (B=3.02; *t*=6.03; *P*<.001) on BG levels (Table 2). Specifically, the WM group demonstrated a significant reduction in the BG levels (B=−2.81; *t*=−8.88; *P*<.001), whereas the NWM group did not exhibit a significant time trend (B=0.21; *t*=0.55; *P*=.59; Figure 3). Before weight monitoring, there was no significant difference observed in BG time trends between the 2 groups (B=0.69; *t*=1.06; *P*=.29). Furthermore, we investigated the proportion of users who achieved a BG level reduction in their last month of measurement less than the average BG levels of 154 mg/100 mL, 183 mg/100 mL, and 212 mg/100 mL, equivalent to estimated glycated hemoglobin (HbA_{1c}) of 7.0, 8.0, and 9.0, respectively [62]. Remarkably, of the 966 users examined per group, 45% (435/966) versus 36% (348/966), 71% (686/966) versus 59% (570/966), and 85% (821/966) versus 76% (734/966) of the WM versus NWM individuals demonstrated substantial reductions in HbA_{1c} levels of <154 mg/100 mL, 183 mg/100 mL, and 212 mg/100 mL, respectively (*P*<.001 for all).

Table 2. Piecewise mixed model analysis of the weight monitoring affecting the monthly BG[a] levels[b].

	Outcome: monthly BG levels		
	Estimates (B, 95% CI)	Statistic *t*	*P* value
Predictors			
Intercept	192.63 (189.05 to 196.20)	105.62	<.001
#BG measurements[c]	−0.22 (−0.25 to −0.19)	−14.11	<.001
time1[d]	−0.17 (−1.34 to 1.00)	−0.29	.77
group (NWM[e])	−2.19 (−6.93 to 2.55)	−0.91	.37
time2[f]	−2.81 (−3.43 to −2.19)	−8.88	<.001
time1×group (NWM)	0.69 (−0.58 to 1.96)	1.06	.29
time2×group (NWM)	3.02 (2.04 to 4.00)	6.03	<.001

[a]BG: blood glucose.

[b]σ2 residual variability=843.15; τ00 UID random intercept=2099.24; τ11 UID.time2 random slope of the second slope=53.68; ρ01 UID: covariance between the random intercept and slope=−0.27; intraclass correlation=0.72.

[c]#BG measurements=number of BG measurements per month.

[d]time1represents the piecewise slopes before the weight monitoring intervention.

[e]NWM: non–weight monitoring.

[f]time2 represents the piecewise slopes after the weight monitoring intervention.

Figure 3. Monthly average blood glucose (BG) fluctuation for the weight monitoring (WM) group and the non-WM (NWM) group. Zero on the x-axis means the start of WM. Vertical lines represent a 95% CI over time.

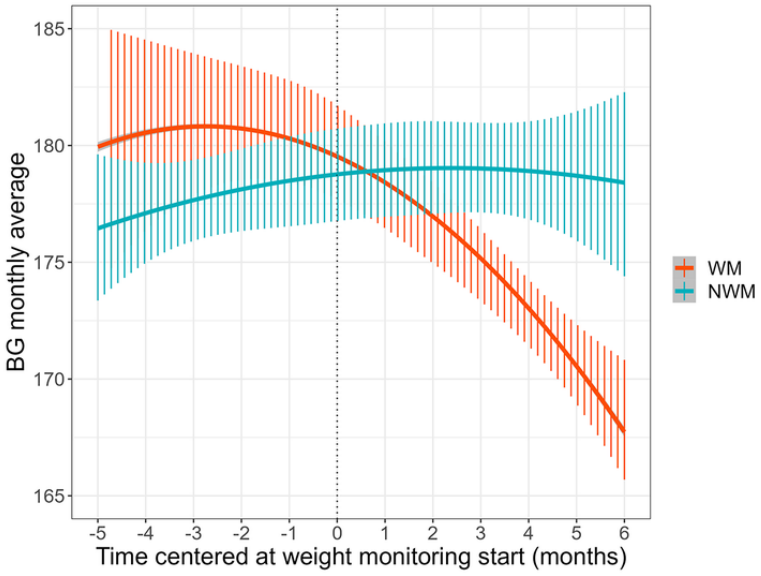

XSL•FO
RenderX

BMI Fluctuations and the Link to BG

During the weight monitoring period (Table 3) of the WM group, a significant decrease in BMI was observed (B=−0.13; *t*=−9.35; *P*<.001).

Interestingly, the number of monthly measurements was negatively associated with BMI (B=−0.003; *t*=−2.22; *P*=.03). Furthermore, the findings of the lagged analysis disaggregating within- and between-person variabilities shed light on BMI as a potential mechanism driving BG (Table 4). Specifically, the analysis demonstrated that an increase in within-person BMI was associated with elevated BG levels (B=4.67; *t*=3.47; *P*<.001). Similarly, an increase in between-person BMI was found to be associated with higher BG levels (B=0.61; *t*=2.65; *P*=.008). However, the 1-month lagged back BMI was not associated with BG levels (B=−0.77; *t*=−0.91; *P*=.36).

Table 3. BMI fluctuations over time in the WM[a] group[b].

	Outcome: BMI		
	Estimates (B, 95% CI)	Statistic *t*	*P* value
Predictors			
Intercept	0.88 (0.49 to 1.26)	4.48	<.001
Monthly number of weight measurements	−0.006 (−0.01 to −0.002)	−2.22	.03
Baseline BMI	0.98 (0.97 to 0.99)	180.41	<.001
Time[c]	−0.13 (−0.16 to −0.10)	−9.35	<.001

[a]WM: weight monitoring.

[b]σ2 residual variability=0.24; τ00 UID random intercept=1.29; τ11 UID.time random slope of the time=0.17; ρ01 UID covariance between the random intercept and slope=−0.25; intraclass correlation=0.93.

[c]time represents the slope over 6 months after the intervention.

Table 4. Within- and between-person associations between BMI and BG[a] in the WM[b] group.

	BG		
	Estimates (B, 95% CI)	Statistic *t*	*P* value
Predictors[c]			
Intercept	147.47 (131.49 to 163.45)	18.09	<.001
BMI within-person	4.67 (2.03 to 7.30)	3.47	.001
BMI between-person	0.61 (0.16 to 1.07)	2.65	.008
BMI within-person lag (−1 month)	−0.77 (−2.43 to 0.89)	−0.91	.36

[a]BG: blood glucose.

[b]WM: weight monitoring.

[c]σ2 residual variability=464.96; τ00 UID random intercept=1684.51; τ11 UID.time random slope of the within-person BMI=325.17; ρ01 UID covariance between the random intercept and slope=−0.01; intraclass correlation=0.80.

Discussion

Principal Findings

This study examined the ability of people with diabetes to regulate BG levels through simple weight monitoring. It used propensity score matching for the control group and used a piecewise mixed model as a statistical framework to describe the nonlinear behavior in BG levels, comparing 2 user cohorts over time. Our analysis indicated that before the weight monitoring phase, both groups demonstrated flat trajectories in BG levels. However, after starting the use of the self-monitoring of weight, the WM group experienced a significant reduction in BG levels, whereas the NWM group's BG levels remained flat.

In addition, by disaggregating within- and between-person BMI variabilities, we showed an association between both BMI sources and BG levels, suggesting that general BMI levels and BMI fluctuations can potentially contribute to BG modulation. However, a lagged analysis did not find an association between within-person BMI fluctuations and next-month BG levels, which does not support the claim of BMI as a potential mechanism of BG changes.

This study demonstrates that the use of digital tools for self-monitoring of weight can significantly affect BG levels in patients with T2D. This finding offers a practical approach to enhancing T2D management, especially for the majority of patients who are also dealing with overweight or obesity issues. Given that weight loss has been consistently shown to improve glycemic control in patients with T2D, as highlighted in previous studies [12-16,25], our findings reinforce the importance of weight management as an integral part of diabetes care. Self-monitoring can enhance patient awareness and engagement in their health management, leading to better outcomes. This

XSL•FO
RenderX

aligns with the growing body of evidence suggesting that patient engagement and empowerment are critical in managing chronic conditions such as T2D [34,40,63,64].

Self-monitoring is the centerpiece of behavioral weight loss intervention programs. A significant association between self-monitoring and weight loss was consistently reported for various health conditions; however, the level of evidence was weak due to methodological limitations [65]. The use of self-monitoring in behavioral changes has a strong theoretical foundation. Self-management was defined as "the personal application of behavior-change tactics that produces a desired change in behavior" [66]. Through self-management interventions, individuals learn to identify occurrences of their own target responding, accurately self-recording the target response, self-evaluating their behavior, and self-delivering reinforcement as a consequence [67].

Although self-monitoring has been described as the cornerstone of behavioral treatment for weight loss, there is a limited examination conducted in the literature [65]. More recently, self-weighing has been introduced as a monitoring component. Daily weighing is valuable for individuals trying to lose weight or prevent weight gain [68]. Consistent with our findings, frequent self-weighing was associated with a lower fat intake, a greater history of dieting to lose weight, and a lower current BMI [68,69].

Previous systematic reviews provided extensive evidence that self-monitoring via digital health, including weight, diet, and physical activity, is associated with superior weight loss [36]. It was specifically shown how distinct features of a digital therapeutic app have the potential to deliver equitable person-centric care and how digital engagement can play a key role in enhancing a person's chronic condition self-management [63,64,70,71].

Self-monitoring has been shown repeatedly to be an important feature of behavioral weight loss digital programs [71]. Self-monitoring of weight and diet were positively correlated with weight loss, and the more consistently the monitoring occurred, the better the weight loss [72,73]. Self-monitoring is also a core component of behavioral obesity treatment, but there is limited knowledge about the efficacy of digital self-monitoring of weight in diabetes [36]. We had previously demonstrated how digital engagement and digital blood pressure monitoring may improve diabetes management [34,74]. Prominently, in this study, the WM and NWM groups were not different in their digital engagement. In addition, the median household income distribution of users in both groups was comparable, suggesting that the digital solution is desired and affordable across lower-, middle-, and high-income levels to enhance glycemic and weight loss outcomes. Mobile apps can successfully help patients lose weight and represent a cost-effective and accessible alternative to intensive in-person weight loss programs [51].

From a psychological perspective, it is assumed that individuals using a digital platform may develop more active roles in managing their health, and self-monitoring affects health in part or in whole via the placebo effect, initiated by mindset modulations [75]. In the realm of physical exercise, a compelling body of research highlights the remarkable impact of mindset on various health parameters. It has been demonstrated that individuals' mindsets about stress could profoundly alter their cortisol levels and influence various hormonal and cardiovascular functions when confronted with stressful situations [76]. Levy et al [77] conducted a noteworthy study that revealed a significant association between individuals' mindsets about aging and their cardiovascular function as well as their actual longevity. A notable example of this phenomenon is evident in the study conducted by Crum and Langer [28], who investigated the effects of mindset on hotel room attendants. These workers, upon adopting the mindset that their daily work constituted a form of beneficial exercise, experienced substantial improvements in several critical health indicators, including weight, BMI, and systolic blood pressure. The potential mechanism may include beliefs and feelings of control people have over their health [28,78]. Collectively, these studies illuminate the potent role of the mindset in shaping various aspects of physical health, providing a background for the potential effect of weight monitoring on BG levels through mindset changes, considering the absence of a quasi-causal association between BMI and BG levels. There is evidence supporting the idea that the placebo effect plays a role in prompting the psychological benefits associated with health-related outcomes [79]. Treatments are delivered in a context that includes social and physical signals, verbal suggestions, and clinical history. This context is actively interpreted by the brain and can elicit expectations, memories, and emotions, which in turn can influence health-related outcomes in the body [79]. Considering the absence of a quasi-causal association between BMI and BG levels, one may consider the effect of self-weighing on BG levels to be mediated by perceptional processes, including mindset modulation.

In agreement with previous studies, we found that the WM group, which monitored their weight, also improved their BMI levels over time [80]. There is strong and consistent evidence that obesity management can delay the progression from prediabetes to T2D and is highly beneficial in treating T2D [80-83]. A significant overlap between T2D and overweight or obesity in etiology and disease mechanisms was broadly investigated. Previous studies have shown a significant improvement in all diabetes-related outcomes, including weight reduction, in patients with T2D and those who have overweight [17]. Controlling both diseases through weight management requires an intensive multidisciplinary approach [84,85]. As body weight increases, patients become more insulin resistant [86], which further drives the need for higher doses of antihyperglycemic medications to keep BG levels in the target range, which then may increase weight gain [17]. It was observed that in obese individuals, adipose tissue releases increased amounts of nonesterified fatty acids, glycerol, hormones, proinflammatory cytokines, and other factors that are involved in the development of insulin resistance. Insulin resistance associated with dysfunction of pancreatic islet beta cells results in the failure to maintain BG levels within the target range [86]. Losing weight can alleviate many of these issues. While losing weight, the pancreas is better able to keep up with the body's need for insulin. In some cases, weight loss is enough to restore BG to a normal level, which eliminates diabetes or

XSL•FO
RenderX

even lowers the need for insulin therapy or other medications to control diabetes [87]. However, other important components may also play a role in weight and BG levels, and other laboratory tests may need to be performed by health care providers.

Monitoring several chronic conditions may have the potential to offer a greater means for helping people with diabetes who have overweight or obesity effectively modulate their glycemia and weight than managing each condition separately. The findings of this study suggest the need for further exploration of how digital health platforms can be effectively integrated into routine clinical practice. Future research should focus on how these tools can be optimized for individual patient needs and how they can be incorporated into broader diabetes management programs. In addition, exploring the long-term impacts of such interventions on patient outcomes and health care use will be valuable. We expect that our analytical framework will be useful for examining other chronic conditions and metabolic syndrome outcomes (eg, lipid profile).

Limitations

As in all studies involving retrospective real-world data, groups were not randomly assigned, and treatment protocols were not prescribed. The propensity score matching approach, while comprehensive, introduces potential biases due to selection methods, which may not be fully mitigated. Unobserved confounders, measurement errors, and other limitations inherent to observational studies may still impact the validity of causal inferences. Nevertheless, propensity score matching represents a rigorous analytical technique that is widely accepted and applied in observational studies to approximate the design and control of randomized controlled trials, making it a valuable tool in the pursuit of causal inference. In this study, we relied on self-reported data for measurements, including weight and BMI. Although self-reporting is a common and practical method in many observational and digital health studies, it can be subject to inaccuracies due to factors such as recall bias or the desire to present oneself in a certain light.

In this real-world data analysis, we designed our timescale to capture changes over a 6-month period both before and after the initiation of weight monitoring. Nevertheless, it is worth noting that the research question of interest in this study could potentially be explored at various temporal scales, including daily, weekly, or monthly intervals. Given the practical challenges associated with monitoring daily changes in real-world settings, most studies in this domain tend to emphasize monthly fluctuations. Moreover, while BG levels offer real-time data and are sensitive to immediate changes in diabetes management, they do not provide a comprehensive view of long-term glycemic control. In addition, longer-term weight monitoring can provide a long-term perspective on glycemic control. Monitoring HbA_{1c} levels over a longer period would provide additional insights into the long-term effects of self-monitoring of weight on glycemic control. However, our study was designed to assess the short-term impacts of digital self-monitoring of weight on BG levels with increased resolution compared to HbA_{1c}. The 6-month period was chosen as it provides a sufficient window to observe significant changes in BG levels in response to weight management, without extending to long-term effects where other variables might confound the results. In addition, the average BMI of the participants was 35.0 (SD 7.3), which is considered a unique population of obesity. Furthermore, longer-term follow-up with these participants is needed because it is not clear how long this weight loss and glycemic improvement will last and what needs to be added for increased sustainability.

Conclusions

In summary, our study underscored the tangible benefits of self-monitoring of weight in the modulation of BG levels among people with diabetes. By leveraging an innovative analytical framework, we found that self-monitoring of weight led to significant reductions in BG levels in the WM group, despite the lack of a direct causal link between BMI fluctuations and BG changes. Drawing from extensive evidence, both historical and from our study, the act of self-monitoring seems to foster a heightened sense of agency and potentially influence health outcomes through mindset modulations.

From a practical standpoint, these findings reinforce the importance of digital health tools in chronic disease management, especially in the realms of diabetes and obesity. Digital self-monitoring platforms not only offer scalable and affordable solutions but also empower individuals to take a proactive role in their health journey. Moreover, the convergence of digital health tools with robust psychological mechanisms, such as the placebo effect and mindset modulation, paves the way for a holistic approach to health care.

This research may also open the door to a myriad of possibilities. While we have illuminated the potential effects of self-monitoring of weight on diabetes management, similar methods could be deployed to investigate the impact on other chronic conditions and metabolic syndrome outcomes. Furthermore, with advancing technology, more granular, real-time data can be leveraged to delve deeper into the daily or even hourly impacts of such interventions. Future studies should also focus on investigating the mechanisms underlying the comorbidity of diabetes and obesity and their management, identifying, and applying mediation models that drive behavioral change that goes beyond multiple chronic conditions.

In light of the significant overlap between T2D and obesity, there is an imperative need to conceptualize and design multifaceted interventions. Blending digital innovation, behavioral science, and clinical knowledge, we can usher in a new era of person-centric health care that is not only responsive but also preemptive. The journey has just begun, and the road ahead promises transformative potential for patients and health care systems alike.

XSL•FO
RenderX

Conflicts of Interest

YFH, EB, and OM are employees of Dario Health. MDR and DLH serve as Dario Health scientific advisory board members, and PG has received a consulting fee to assist with analyses but otherwise has no conflicts of interest.

References

1. Amante DJ, Harlan DM, Lemon SC, McManus DD, Olaitan OO, Pagoto SL, et al. Evaluation of a diabetes remote monitoring program facilitated by connected glucose meters for patients with poorly controlled type 2 diabetes: randomized crossover trial. JMIR Diabetes 2021 Mar 11;6(1):e25574 [FREE Full text] [doi: 10.2196/25574] [Medline: 33704077]
2. Bryan S, Afful J, Carroll M, Te-Ching C, Orlando D, Fink S, et al. National health and nutrition examination survey 2017–March 2020 prepandemic data files—development of files and prevalence estimates for selected health outcomes. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/nchs/data/nhsr/nhsr158-508.pdf [accessed 2024-03-16]
3. Cawley J, Biener A, Meyerhoefer C, Ding Y, Zvenyach T, Smolarz BG, et al. Direct medical costs of obesity in the United States and the most populous states. J Manag Care Spec Pharm 2021 Mar;27(3):354-366 [FREE Full text] [doi: 10.18553/jmcp.2021.20410] [Medline: 33470881]
4. Leitner DR, Frühbeck G, Yumuk V, Schindler K, Micic D, Woodward E, et al. Obesity and type 2 diabetes: two diseases with a need for combined treatment strategies - EASO can lead the way. Obes Facts 2017;10(5):483-492 [FREE Full text] [doi: 10.1159/000480525] [Medline: 29020674]
5. Nianogo RA, Arah OA. Forecasting obesity and type 2 diabetes incidence and burden: the ViLA-obesity simulation model. Front Public Health 2022;10:818816 [FREE Full text] [doi: 10.3389/fpubh.2022.818816] [Medline: 35450123]
6. Ruze R, Liu T, Zou X, Song J, Chen Y, Xu R, et al. Obesity and type 2 diabetes mellitus: connections in epidemiology, pathogenesis, and treatments. Front Endocrinol (Lausanne) 2023;14:1161521 [FREE Full text] [doi: 10.3389/fendo.2023.1161521] [Medline: 37152942]
7. Eizirik DL, Pasquali L, Cnop M. Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. Nat Rev Endocrinol 2020 Jul;16(7):349-362. [doi: 10.1038/s41574-020-0355-7] [Medline: 32398822]
8. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2019. Diabetes Care 2019 Jan;42(Suppl 1):S13-S28. [doi: 10.2337/dc19-S002] [Medline: 30559228]
9. Versini M, Jeandel PY, Rosenthal E, Shoenfeld Y. Obesity in autoimmune diseases: not a passive bystander. Autoimmun Rev 2014 Sep;13(9):981-1000. [doi: 10.1016/j.autrev.2014.07.001] [Medline: 25092612]
10. Tuomi T, Santoro N, Caprio S, Cai M, Weng J, Groop L. The many faces of diabetes: a disease with increasing heterogeneity. Lancet 2014 Mar 22;383(9922):1084-1094. [doi: 10.1016/S0140-6736(13)62219-9] [Medline: 24315621]
11. Thom G, Messow CM, Leslie WS, Barnes AC, Brosnahan N, McCombie L, et al. Predictors of type 2 diabetes remission in the Diabetes Remission Clinical Trial (DiRECT). Diabet Med 2021 Aug;38(8):e14395. [doi: 10.1111/dme.14395] [Medline: 32870520]
12. Bays HE. Why does type 2 diabetes mellitus impair weight reduction in patients with obesity? A review. Obes Pillars 2023 Sep;7:100076 [FREE Full text] [doi: 10.1016/j.obpill.2023.100076] [Medline: 37990681]
13. Bays HE, Bindlish S, Clayton TL. Obesity, diabetes mellitus, and cardiometabolic risk: an obesity medicine association (OMA) clinical practice statement (CPS) 2023. Obes Pillars 2023 Mar;5:100056 [FREE Full text] [doi: 10.1016/j.obpill.2023.100056] [Medline: 37990743]
14. Kong DX, Xiao Y, Zhang ZX, Liu YB. Study on the correlation between metabolism, insulin sensitivity and progressive weight loss change in type-2 diabetes. Pak J Med Sci 2020;36(7):1523-1528 [FREE Full text] [doi: 10.12669/pjms.36.7.3027] [Medline: 33235568]
15. Scheen AJ, Van Gaal LF. Combating the dual burden: therapeutic targeting of common pathways in obesity and type 2 diabetes. Lancet Diabetes Endocrinol 2014 Nov;2(11):911-922 [FREE Full text] [doi: 10.1016/S2213-8587(14)70004-X] [Medline: 24731666]
16. Apovian CM, Okemah J, O'Neil PM. Body weight considerations in the management of type 2 diabetes. Adv Ther 2019 Jan;36(1):44-58 [FREE Full text] [doi: 10.1007/s12325-018-0824-8] [Medline: 30465123]
17. Hamdy O, Ashrafzadeh S, Mottalib A. Weight management in patients with type 2 diabetes: a multidisciplinary real-world approach. Curr Diab Rep 2018 Jul 17;18(9):66 [FREE Full text] [doi: 10.1007/s11892-018-1030-4] [Medline: 30019229]
18. Bolignano D, Zoccali C. Effects of weight loss on renal function in obese CKD patients: a systematic review. Nephrol Dial Transplant 2013 Nov;28 Suppl 4:iv82-iv98. [doi: 10.1093/ndt/gft302] [Medline: 24092846]
19. Araújo-Soares V, Hankonen N, Presseau J, Rodrigues A, Sniehotta FF. Developing behavior change interventions for self-management in chronic illness: an integrative overview. Eur Psychol 2019;24(1):7-25 [FREE Full text] [doi: 10.1027/1016-9040/a000330] [Medline: 31496632]
20. Grady PA, Gough LL. Self-management: a comprehensive approach to management of chronic conditions. Am J Public Health 2014 Aug;104(8):e25-e31. [doi: 10.2105/AJPH.2014.302041] [Medline: 24922170]
21. Davies MJ, D'Alessio DA, Fradkin J, Kernan WN, Mathieu C, Mingrone G, et al. Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the

XSL•FO
RenderX

Study of Diabetes (EASD). Diabetes Care 2018 Dec;41(12):2669-2701 [FREE Full text] [doi: 10.2337/dci18-0033] [Medline: 30291106]
22. Samdal GB, Eide GE, Barth T, Williams G, Meland E. Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. Int J Behav Nutr Phys Act 2017 Mar 28;14(1):42 [FREE Full text] [doi: 10.1186/s12966-017-0494-y] [Medline: 28351367]
23. Magkos F, Yannakoulia M, Chan JL, Mantzoros CS. Management of the metabolic syndrome and type 2 diabetes through lifestyle modification. Annu Rev Nutr 2009;29:223-256 [FREE Full text] [doi: 10.1146/annurev-nutr-080508-141200] [Medline: 19400751]
24. Estabrooks PA, Nelson CC, Xu S, King D, Bayliss EA, Gaglio B, et al. The frequency and behavioral outcomes of goal choices in the self-management of diabetes. Diabetes Educ 2005;31(3):391-400. [doi: 10.1177/0145721705276578] [Medline: 15919639]
25. Reynolds LR, Anderson JW. Practical office strategies for weight management of the obese diabetic individual. Endocr Pract 2004;10(2):153-159. [doi: 10.4158/EP.10.2.153] [Medline: 15256334]
26. Hood KK, Hilliard M, Piatt G, Ievers-Landis CE. Effective strategies for encouraging behavior change in people with diabetes. Diabetes Manag (Lond) 2015;5(6):499-510 [FREE Full text] [Medline: 30100925]
27. Raffali A, Asraf RM. Fostering growth mindset principles in the prevention of type 2 diabetes through a narrative game. Int J Learn Teach Educ Res 2023;22(2):244-261 [FREE Full text] [doi: 10.26803/ijlter.22.2.14]
28. Crum AJ, Langer EJ. Mind-set matters: exercise and the placebo effect. Psychol Sci 2007 Feb;18(2):165-171. [doi: 10.1111/j.1467-9280.2007.01867.x] [Medline: 17425538]
29. Park C, Pagnini F, Reece A, Phillips D, Langer E. Blood sugar level follows perceived time rather than actual time in people with type 2 diabetes. Proc Natl Acad Sci U S A 2016 Jul 19;113(29):8168-8170 [FREE Full text] [doi: 10.1073/pnas.1603444113] [Medline: 27382161]
30. Lee KS, Feltner FJ, Bailey AL, Lennie TA, Chung ML, Smalls BL, et al. The relationship between psychological states and health perception in individuals at risk for cardiovascular disease. Psychol Res Behav Manag 2019;12:317-324 [FREE Full text] [doi: 10.2147/PRBM.S198280] [Medline: 31191053]
31. Awad A, Trenfield SJ, Pollard TD, Ong JJ, Elbadawi M, McCoubrey LE, et al. Connected healthcare: improving patient care using digital health technologies. Adv Drug Deliv Rev 2021 Nov;178:113958. [doi: 10.1016/j.addr.2021.113958] [Medline: 34478781]
32. Cahn A, Akirov A, Raz I. Digital health technology and diabetes management. J Diabetes 2018 Jan;10(1):10-17. [doi: 10.1111/1753-0407.12606] [Medline: 28872765]
33. Han J, King F, Klonoff D, Drincic A, Crosby KP, Robinson T, et al. Digital diabetes congress 2019. J Diabetes Sci Technol 2019 Sep;13(5):979-989 [FREE Full text] [doi: 10.1177/1932296819872107] [Medline: 31466480]
34. Fundoiano-Hershcovitz Y, Hirsch A, Dar S, Feniger E, Goldstein P. Role of digital engagement in diabetes care beyond measurement: retrospective cohort study. JMIR Diabetes 2021 Feb 18;6(1):e24030 [FREE Full text] [doi: 10.2196/24030] [Medline: 33599618]
35. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
36. Patel ML, Wakayama LN, Bennett GG. Self-monitoring via digital health in weight loss interventions: a systematic review among adults with overweight or obesity. Obesity (Silver Spring) 2021 Mar;29(3):478-499. [doi: 10.1002/oby.23088] [Medline: 33624440]
37. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
38. Offringa R, Sheng T, Parks L, Clements M, Kerr D, Greenfield MS. Digital diabetes management application improves glycemic outcomes in people with type 1 and type 2 diabetes. J Diabetes Sci Technol 2018 May 25;12(3):701-708 [FREE Full text] [doi: 10.1177/1932296817747291] [Medline: 29277103]
39. Bollyky JB, Melton ST, Xu T, Painter SL, Knox B. The effect of a cellular-enabled glucose meter on glucose control for patients with diabetes: prospective pre-post study. JMIR Diabetes 2019 Oct 07;4(4):e14799 [FREE Full text] [doi: 10.2196/14799] [Medline: 31593545]
40. Osborn CY, van Ginkel JR, Rodbard D, Heyman M, Marrero DG, Huddleston B, et al. One drop | mobile: an evaluation of hemoglobin A1c improvement linked to app engagement. JMIR Diabetes 2017 Aug 24;2(2):e21 [FREE Full text] [doi: 10.2196/diabetes.8039] [Medline: 30291059]
41. Ghelani DP, Moran LJ, Johnson C, Mousa A, Naderpoor N. Mobile apps for weight management: a review of the latest evidence to inform practice. Front Endocrinol (Lausanne) 2020;11:412 [FREE Full text] [doi: 10.3389/fendo.2020.00412] [Medline: 32670197]
42. MacPherson MM, Merry KJ, Locke SR, Jung ME. mHealth prompts within diabetes prevention programs: a scoping review. Mhealth 2022;8:20 [FREE Full text] [doi: 10.21037/mhealth-21-22] [Medline: 35449504]

XSL•FO
RenderX

43. MacPherson MM, Merry KJ, Locke SR, Jung ME. Effects of mobile health prompts on self-monitoring and exercise behaviors following a diabetes prevention program: secondary analysis from a randomized controlled trial. JMIR Mhealth Uhealth 2019 Sep 05;7(9):e12956 [FREE Full text] [doi: 10.2196/12956] [Medline: 31489842]
44. Stellefson M, Chaney B, Barry AE, Chavarria E, Tennant B, Walsh-Childers K, et al. Web 2.0 chronic disease self-management for older adults: a systematic review. J Med Internet Res 2013 Feb 14;15(2):e35 [FREE Full text] [doi: 10.2196/jmir.2439] [Medline: 23410671]
45. Doyle J, Murphy E, Gavin S, Pascale A, Deparis S, Tommasi P, et al. A digital platform to support self-management of multiple chronic conditions (ProACT): findings in relation to engagement during a one-year proof-of-concept trial. J Med Internet Res 2021 Dec 15;23(12):e22672 [FREE Full text] [doi: 10.2196/22672] [Medline: 34914612]
46. Ancker JS, Witteman HO, Hafeez B, Provencher T, Van de Graaf M, Wei E. "You get reminded you're a sick person": personal data tracking and patients with multiple chronic conditions. J Med Internet Res 2015 Aug 19;17(8):e202 [FREE Full text] [doi: 10.2196/jmir.4209] [Medline: 26290186]
47. Doyle J, Murphy E, Kuiper J, Smith S, Hannigan C, Jacobs A, et al. Managing multimorbidity: identifying design requirements for a digital self-management tool to support older adults with multiple chronic conditions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, Scotland p. 1-14 URL: https://dl.acm.org/doi/10.1145/3290605.3300629
48. Lim CY, Berry AB, Hartzler A, Hirsch T, Carrell DS, Bermet ZA, et al. Facilitating self-reflection about values and self-care among individuals with chronic conditions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, Scotland p. 1-12 URL: https://dl.acm.org/doi/10.1145/3290605.3300885 [doi: 10.1145/3290605.3300885]
49. Caldeira C, Gui X, Reynolds TL, Bietz M, Chen Y. Managing healthcare conflicts when living with multiple chronic conditions. Int J Hum Comput Stud 2021 Jan;145:102494 [FREE Full text] [doi: 10.1016/j.ijhcs.2020.102494]
50. Tomah S, Mahmoud N, Mottalib A, Pober DM, Tasabehji MW, Ashrafzadeh S, et al. Frequency of self-monitoring of blood glucose in relation to weight loss and A1C during intensive multidisciplinary weight management in patients with type 2 diabetes and obesity. BMJ Open Diabetes Res Care 2019;7(1):e000659 [FREE Full text] [doi: 10.1136/bmjdrc-2019-000659] [Medline: 31413841]
51. Ufholz K, Werner J. The efficacy of mobile applications for weight loss. Curr Cardiovasc Risk Rep 2023 Mar 23;17(4):83-90 [FREE Full text] [doi: 10.1007/s12170-023-00717-2] [Medline: 36974130]
52. About adult BMI. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html [accessed 2022-05-03]
53. Measuring America's people, places, and economy. The United States Census Bureau. URL: https://www.census.gov/ [accessed 2024-03-16]
54. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011 May;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
55. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 2012 Mar 12;39(1):33-38. [doi: 10.1080/00031305.1985.10479383]
56. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010 Feb 01;25(1):1-21 [FREE Full text] [doi: 10.1214/09-STS313] [Medline: 20871802]
57. Ho D, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42(8):1-28. [doi: 10.18637/jss.v042.i08]
58. Home page. Ethical & Independent Review Services (E&I). URL: https://eandireview.com/ [accessed 2023-06-13]
59. Terms of use. DarioHealth Corp. URL: https://www.dariohealth.com/terms-of-use/ [accessed 2024-03-16]
60. Kohli N, Peralta Y, Zopluoglu C, Davison ML. A note on estimating single-class piecewise mixed-effects models with unknown change points. Int J Behav Dev 2018 Feb 15;42(5):518-524. [doi: 10.1177/0165025418759237]
61. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
62. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ, A1c-Derived Average Glucose Study Group. Translating the A1C assay into estimated average glucose values. Diabetes Care 2008 Aug;31(8):1473-1478 [FREE Full text] [doi: 10.2337/dc08-0545] [Medline: 18540046]
63. Frias J, Virdi N, Raja P, Kim Y, Savage G, Osterberg L. Effectiveness of digital medicines to improve clinical outcomes in patients with uncontrolled hypertension and type 2 diabetes: prospective, open-label, cluster-randomized pilot clinical trial. J Med Internet Res 2017 Jul 11;19(7):e246 [FREE Full text] [doi: 10.2196/jmir.7833] [Medline: 28698169]
64. Shan R, Sarkar S, Martin SS. Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. Diabetologia 2019 Jun 8;62(6):877-887. [doi: 10.1007/s00125-019-4864-7] [Medline: 30963188]
65. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011 Jan;111(1):92-102 [FREE Full text] [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
66. Cooper JO, Heron TE, Heward WL. Applied Behavior Analysis. 3rd edition. Old Bridge, NJ: Pearson Prentice Hall; 2020.

XSL•FO
RenderX

67. Erhard P, Wong T, Barnett M, Falcomata T, Lang R. Self-management skills and applied behavior analysis. In: Matson JL, Sturmey P, editors. Handbook of Autism and Pervasive Developmental Disorder: Assessment, Diagnosis, and Treatment. Cham, Switzerland: Springer; 2022:957-973.
68. Linde JA, Jeffery RW, French SA, Pronk NP, Boyle RG. Self-weighing in weight gain prevention and weight loss trials. Ann Behav Med 2005 Dec;30(3):210-216. [doi: 10.1207/s15324796abm3003_5] [Medline: 16336072]
69. Patel ML, Hopkins CM, Brooks TL, Bennett GG. Comparing self-monitoring strategies for weight loss in a smartphone app: randomized controlled trial. JMIR Mhealth Uhealth 2019 Feb 28;7(2):e12209 [FREE Full text] [doi: 10.2196/12209] [Medline: 30816851]
70. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 1. improving care and promoting health in populations: standards of care in diabetes-2023. Diabetes Care 2023 Jan 01;46(Supple 1):S10-S18 [FREE Full text] [doi: 10.2337/dc23-S001] [Medline: 36507639]
71. Lin PH, Grambow S, Intille S, Gallis JA, Lazenka T, Bosworth H, et al. The association between engagement and weight loss through personal coaching and cell phone interventions in young adults: randomized controlled trial. JMIR Mhealth Uhealth 2018 Oct 18;6(10):e10471 [FREE Full text] [doi: 10.2196/10471] [Medline: 30341051]
72. Burke LE, Warziski M, Starrett T, Choo J, Music E, Sereika S, et al. Self-monitoring dietary intake: current and future practices. J Ren Nutr 2005 Jul;15(3):281-290. [doi: 10.1016/j.jrn.2005.04.002] [Medline: 16007557]
73. Baker RC, Kirschenbaum DS. Self-monitoring may be necessary for successful weight control. Behav Ther 1993 Jun;24(3):377-394. [doi: 10.1016/S0005-7894(05)80212-6]
74. Gershoni T, Ritholz MD, Horwitz D, Manejwala O, Donaldson-Pitter T, Fundoiano-Hershcovitz Y. Glycemic management by a digital therapeutic platform across racial/ethnic groups: a retrospective cohort study. Applied Sciences 2022 Dec 29;13(1):431 [FREE Full text] [doi: 10.3390/app13010431] [Medline: 33599618]
75. Tighe SA, Ball K, Kensing F, Kayser L, Rawstorn JC, Maddison R. Toward a digital platform for the self-management of noncommunicable disease: systematic review of platform-like interventions. J Med Internet Res 2020 Oct 28;22(10):e16774 [FREE Full text] [doi: 10.2196/16774] [Medline: 33112239]
76. Crum AJ, Salovey P, Achor S. Rethinking stress: the role of mindsets in determining the stress response. J Pers Soc Psychol 2013 Apr;104(4):716-733. [doi: 10.1037/a0031201] [Medline: 23437923]
77. Levy BR, Slade MD, Kunkel SR, Kasl SV. Longevity increased by positive self-perceptions of aging. J Pers Soc Psychol 2002 Aug;83(2):261-270. [doi: 10.1037//0022-3514.83.2.261] [Medline: 12150226]
78. Salvia MG, Ritholz MD, Craigen KL, Quatromoni PA. Women's perceptions of weight stigma and experiences of weight-neutral treatment for binge eating disorder: a qualitative study. EClinicalMedicine 2023 Feb;56:101811 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101811] [Medline: 36618893]
79. Wager TD, Atlas LY. The neuroscience of placebo effects: connecting context, learning and health. Nat Rev Neurosci 2015 Jul;16(7):403-418 [FREE Full text] [doi: 10.1038/nrn3976] [Medline: 26087681]
80. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, on behalf of the American Diabetes Association. 8. Obesity and weight management for the prevention and treatment of type 2 diabetes: standards of care in diabetes-2023. Diabetes Care 2023 Jan 01;46(Suppl 1):S128-S139 [FREE Full text] [doi: 10.2337/dc23-S008] [Medline: 36507637]
81. Rothberg AE, McEwen LN, Kraftson AT, Fowler CE, Herman WH. Very-low-energy diet for type 2 diabetes: an underutilized therapy? J Diabetes Complications 2014;28(4):506-510 [FREE Full text] [doi: 10.1016/j.jdiacomp.2014.03.014] [Medline: 24849710]
82. Garvey WT, Ryan DH, Bohannon NJ, Kushner RF, Rueger M, Dvorak RV, et al. Weight-loss therapy in type 2 diabetes: effects of phentermine and topiramate extended release. Diabetes Care 2014 Dec;37(12):3309-3316. [doi: 10.2337/dc14-0930] [Medline: 25249652]
83. Davies MJ, Bergenstal R, Bode B, Kushner RF, Lewin A, Skjøth TV, NN8022-1922 Study Group. Efficacy of liraglutide for weight loss among patients with type 2 diabetes: the SCALE diabetes randomized clinical trial. JAMA 2015 Aug 18;314(7):687-699. [doi: 10.1001/jama.2015.9676] [Medline: 26284720]
84. Garber AJ, Abrahamson MJ, Barzilay JI, Blonde L, Bloomgarden ZT, Bush MA, et al. Consensus statement by the American association of clinical endocrinologists and American college of endocrinology on the comprehensive type 2 diabetes management algorithm - 2018 executive summary. Endocr Pract 2018 Jan;24(1):91-120. [doi: 10.4158/CS-2017-0153] [Medline: 29368965]
85. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, on behalf of the American Diabetes Association. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes-2023. Diabetes Care 2023 Jan 01;46(Supple 1):S68-S96 [FREE Full text] [doi: 10.2337/dc23-S005] [Medline: 36507648]
86. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature 2006 Dec 14;444(7121):840-846. [doi: 10.1038/nature05482] [Medline: 17167471]
87. Bishop S. Losing weight can have big impact on those with diabetes. Mayo Clinic News Network. URL: https://newsnetwork.mayoclinic.org/discussion/losing-weight-can-have-big-impact-on-those-with-diabetes/ [accessed 2023-08-06]

XSL•FO
RenderX

Abbreviations

BG: blood glucose
NWM: non–weight monitoring (group)
T2D: type 2 diabetes
WM: weight monitoring (group)

Edited by G Eysenbach, T Leung; submitted 28.11.23; peer-reviewed by Y Ding, T de Azevedo Cardoso; comments to author 02.01.24; revised version received 18.02.24; accepted 22.02.24; published 02.04.24.

Please cite as:
Fundoiano-Hershcovitz , Ritholz MD, Horwitz DL, Behar E, Manejwala O, Goldstein P
The Impact of Digital Self-Monitoring of Weight on Improving Diabetes Clinical Outcomes: Quasi-Randomized Study
J Med Internet Res 2024;26:e54940
URL: https://www.jmir.org/2024/1/e54940
doi:10.2196/54940
PMID:38564266

© Yifat Fundoiano-Hershcovitz, Marilyn D Ritholz, David L Horwitz, Ephraim Behar, Omar Manejwala, Pavel Goldstein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Regulatory Standards and Guidance for the Use of Health Apps for Self-Management in Sub-Saharan Africa: Scoping Review

Benard Ayaka Bene[1,2], MBBS, MPH; Sunny Ibeneme[3], MD, PhD; Kayode Philip Fadahunsi[1], MBBS, MPH; Bala Isa Harri[4], MBBS, MPH, MSc; Nkiruka Ukor[5], MSc; Nikolaos Mastellos[1], BSc, PhD; Azeem Majeed[1], MD; Josip Car[1,6], MSc, MD, PhD

[1]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

[2]Department of Public Health, Federal Ministry of Health, Abuja, Nigeria

[3]Digital Health Specialist, UNICEF East Asia Pacific Regional Office, Bangkok, Thailand

[4]Department of Health Planning, Research and Statistics, Federal Ministry of Health, Abuja, Nigeria

[5]Strategic Health Information Cluster, World Health Organization, Abuja, Nigeria

[6]School of Life Course & Population Sciences, King's College London, London, United Kingdom

Corresponding Author:
Benard Ayaka Bene, MBBS, MPH
Department of Primary Care and Public Health
School of Public Health
Imperial College London
The Reynolds Building
St Dunstan's Road
London, W6 8RP
United Kingdom
Phone: 44 7598439185
Email: b.bene15@imperial.ac.uk

Abstract

Background: Health apps are increasingly recognized as crucial tools for enhancing health care delivery. Many countries, particularly those in sub-Saharan Africa, can substantially benefit from using health apps to support self-management and thus help to achieve universal health coverage and the third sustainable development goal. However, most health apps published in app stores are of unknown or poor quality, which poses a risk to patient safety. Regulatory standards and guidance can help address this risk and promote patient safety.

Objective: This review aims to assess the regulatory standards and guidance for health apps supporting evidence-based best practices in sub-Saharan Africa with a focus on self-management.

Methods: A methodological framework for scoping reviews was applied. A search strategy was built and applied across the following databases, gray literature sources, and institutional websites: PubMed, Scopus, World Health Organization (WHO) African Index Medicus, OpenGrey, WHO Regional Office for Africa Library, ICTworks, WHO Directory of eHealth policies, HIS Strengthening Resource Center, International Telecommunication Union, Ministry of Health websites, and Google. The search covered the period between January 2005 and January 2024. The findings were analyzed using a deductive descriptive content analysis. The policy analysis framework was adapted and used to organize the findings. The Reporting Items for Stakeholder Analysis tool guided the identification and mapping of key stakeholders based on their roles in regulating health apps for self-management.

Results: The study included 49 documents from 31 sub-Saharan African countries. While all the documents were relevant for stakeholder identification and mapping, only 3 regulatory standards and guidance contained relevant information on regulation of health apps. These standards and guidance primarily aimed to build mutual trust; promote integration, inclusion, and equitable access to services; and address implementation issues and poor coordination. They provided guidance on systems quality, software acquisition and maintenance, security measures, data exchange, interoperability and integration, involvement of relevant stakeholders, and equitable access to services. To enhance implementation, the standards highlight that legal authority, coordination of activities, building capacity, and monitoring and evaluation are required. A number of stakeholders, including governments,

XSL•FO
RenderX

regulatory bodies, funders, intergovernmental and nongovernmental organizations, academia, and the health care community, were identified to play key roles in regulating health apps.

Conclusions: Health apps have huge potential to support self-management in sub-Saharan Africa, but the lack of regulatory standards and guidance constitutes a major barrier. Hence, for these apps to be safely and effectively integrated into health care, more attention should be given to regulation. Learning from countries with effective regulations can help sub-Saharan Africa build a more robust and responsive regulatory system, ensuring the safe and beneficial use of health apps across the region.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2018-025714

(J Med Internet Res 2024;26:e49163) doi:10.2196/49163

KEYWORDS

regulation; standard; framework; guidance; guideline; health app; self-management; World Health Organization; WHO African Region; sub-Saharan Africa

Introduction

Background

Health apps are the most widely used digital health products globally [1,2]. Harnessing the potential of health apps creates a huge opportunity in providing support for health care delivery, including patient communication, patient education, and decision support for self-management [3-8]. Health apps can be an effective tool to strengthen health systems worldwide, especially in low- and middle-income countries including those in sub-Saharan Africa [4,5,9]. As a result, the attainment of universal health coverage (UHC) and sustainable development goal (SDG) 3, good health and well-being, can be accelerated [8,10].

Many health apps fall below the expected quality threshold [11]. Several studies have found that widely used health apps are often technically unreliable and clinically unsafe [12-14] and do not comply with ethical standards and the principles of confidentiality of information and data privacy [15,16]. In addition, many commercially available health apps were not developed using interoperability standards that are widely accepted in sub-Saharan Africa (eg, Fast Healthcare Interoperability Resources [FHIR]) [17-20]. Consequently, it becomes difficult to integrate these apps into a clinical workflow.

Hence, regulation through robust mechanisms is crucial to enhance the development, implementation, and adoption of health apps. Regulatory standards and guidance are essential for the safety of patients as they ensure quality assurance of any new technology in health care and contribute to building mutual trust while promoting the optimal use of the technology [21-23]. Therefore, to ensure that health apps that are used to support the self-management of patients are technically reliable and clinically safe, interoperable across systems, and compliant with the principles of confidentiality of information and data privacy, there is a need for effective regulatory standards. Furthermore, effective regulation can help ensure that health apps for self-management are culturally functional and competent and are accessible to those who need them regardless of gender, ethnicity, geographical location, or financial status [24-31].

Since 2005, there have been ongoing efforts to strengthen digital health governance at both the national and international levels [32,33]. In 2018, the World Health Organization (WHO) member states renewed their commitment to using digital health technologies (DHTs) to advance UHC and SDG 3 [33]. However, to date, the extent to which the use of health apps for self-management is regulated across countries within the WHO African Region (also known as sub-Saharan Africa) remains unclear. Therefore, this review was conducted to identify available regulatory standards and guidance and assess the extent to which they regulate health apps for self-management in sub-Saharan Africa. The review also mapped out the key stakeholders and their roles in regulating health apps for self-management across sub-Saharan Africa.

Review Questions

The review attempted to answer the following questions: (1) What regulatory standards and guidance are available for regulating health apps for self-management across sub-Saharan Africa? (2) To what extent do regulatory standards and guidance regulate health apps for self-management in terms of what aspects are regulated; why, how, and for whom; and what aspects are not regulated? (3) Who are the key stakeholders and what are their roles in regulating health apps for self-management?

Methods

Study Design

The process of this scoping review followed the methodological framework for conducting a scoping study originally described by Arksey and O'Malley [34] and the updated methodological guidance for conducting a Joanna Briggs Institute scoping review [34-37]. The reporting of the review was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [38]. A completed PRISMA-ScR checklist is provided in Multimedia Appendix 1. The protocol of this scoping review was published in *BMJ Open* [30].

Identifying Relevant Documents

Two reviewers (BAB and SI) developed the search strategy with the assistance of a librarian and in consultation with other research team members (KPF, BIH, NU, NM, AM, and JC). The following key terms were included: policy, legislation, strategy, regulation, standard, criterion, framework, guidance, guideline, digital health, eHealth, app, WHO African Region,

XSL•FO
RenderX

and sub-Saharan Africa, and the names of all sub-Saharan African countries.

Owing to the absence of regulatory standards and guidance in scientific databases, the search focus was narrowed down to gray literature sources and institutional websites, including OpenGrey, WHO Regional Office for Africa (AFRO) Library, repositories for digital health policies (ICTworks, WHO's Directory of eHealth Policies, and Health Information System Strengthening Resource Center), as well as the websites of WHO, International Telecommunication Union (ITU), and Ministries of Health (MOHs). The only scientific databases searched were PubMed, Scopus, and WHO AIM. PubMed was not included in the protocol. We also conducted a systematic search on Google. We used truncation to increase the yield of the results. The search strategy was then applied across PubMed, Scopus, and WHO AIM databases using Boolean terms (mainly *OR* and *AND*) to combine search results. Gray literature sources and institutional websites were searched using phrases containing ≥2 keywords such as "eHealth regulation," "digital health regulatory standard," "eHealth regulatory standard," "digital health regulation," "digital health policy," "eHealth policy," "digital health strategy," and "eHealth strategy." For Google search, we added the names of the country to the phrases (eg, "digital health regulation Nigeria"). The reference lists of the included documents were also searched, and key individuals at the MOHs, WHO Country Offices, and the WHO AFRO were contacted for related documents. When our search was conducted, the WHO Directory of eHealth policies website was unavailable, and the WHO AFRO Library was undergoing reconstruction. The search strategies for PubMed, Scopus, and WHO AIM are provided in Multimedia Appendix 2. The search was conducted between 2005 and January 2024.

Study Selection

The search results obtained from PubMed, Scopus, and WHO AIM were imported into Mendeley (Elsevier) [39] to remove duplicates. The search conducted on OpenGrey did not yield any results, whereas relevant records obtained from institutional websites, repositories, and Google were downloaded as PDF copies and uploaded to Mendeley. After removing duplicates, the remaining results were imported into Covidence (Veritas Health Innovation) [40] for screening. Two reviewers (BAB and SI) applied the predefined eligibility criteria (Textbox 1) to screen the documents in 2 stages (title and abstract or executive summary). All discrepancies were discussed until the reviewers reached agreement.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Type of document: Regulatory standards, guidance, policies, strategies, and committee or government reports that address regulatory issues related to the use of health apps for self-management
- Location: Documents developed and implemented in countries within sub-Saharan Africa
- Date of publication: Documents developed since 2005; the global efforts toward promoting standards to minimize variability and potential harms that could arise from poorly regulated use of digital health began in 2005 [33]
- Language: Documents written in English language and other official languages of sub-Saharan African countries (Portuguese and French)

Exclusion criteria

- Type of document: Standards, guidance, policies, strategies, and reports not related to regulation of health apps
- Location: Documents from countries outside sub-Saharan Africa
- Date of publication: Documents developed before 2005
- Language: None

Data Charting (Extraction)

Two reviewers (BAB and SI), in consultation with the other members of the research team, developed the data extraction forms using an iterative process that included piloting data extraction and refinement until a consensus was reached.

We proposed in the study protocol [30] that data extraction would be conducted by the 2 reviewers independently. However, owing to the approach adopted for data extraction (deductive qualitative content analysis), 1 reviewer, rather than 2, initially extracted data from the included documents, and any concerns were discussed with a second reviewer [41]. Unresolved issues were then discussed and resolved with a third reviewer in a steering group meeting.

Collating, Summarizing, and Reporting Results

To address the research questions (particularly question 2), we adopted a deductive descriptive qualitative content analysis method to analyze and report the key findings. The policy analysis framework by Walt and Gilson [42] was adapted and applied to ensure that there was a consistent way of organizing the key findings: (1) Content (which aspects are regulated and which aspects are not?)—these are the components that directly or indirectly address regulatory issues related to the use of health apps for self-management, including areas that have not been addressed. (2) Context (why are those aspects regulated?)—this characterizes the rationale indicated for addressing regulatory issues related to the use of health apps for self-management. (3) Process (how are the regulatory standards developed and implemented?)—this describes the methods or approaches used to develop and implement regulatory standards. (4) Actors (who

XSL•FO
RenderX

are the regulatory standards targeted toward?)—these are the key actors targeted by the standards.

Using a deductive descriptive qualitative content analysis approach, we examined each included document to systematically identify texts for concepts, patterns, and other relevant information. We then categorized them under *content*, *context*, *process*, or *actors* in relation to regulating health apps for self-management. The findings under content and context were further organized based on 4 predefined regulatory categories or themes as documented in the literature, namely (1) *technical and clinical safety* [12-14], (2) *data protection and security* [15,16], (3) *standards and interoperability* [28,31], and (4) *inclusion and equitable access* [24-29].

To address the third research question, the Reporting Items for Stakeholder Analysis (RISA) tool [41] was used as a guide to group key stakeholders based on role categorization as recognized globally by the WHO, the ITU, and UNESCO [32,33,43].

Ethical Considerations

Primary data were not collected in this study. Therefore, no ethics approval was required.

Results

Search Results

A total of 2900 records were obtained after removing duplicates. Although the literature search was conducted in English, the search also yielded documents written in French and Portuguese from the ICTworks repository [44]. Following the initial screening of the title and abstract (or executive summaries), 73 documents were retrieved for full-text assessment. After applying the inclusion criteria for the full-text assessment, 49 documents were found eligible for inclusion in the review.

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram [45] showing the study selection process is presented in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the study selection process. The descriptive qualitative content analysis included only 3 of the 49 (6%) documents used for stakeholder mapping. WHO: World Health Organization.

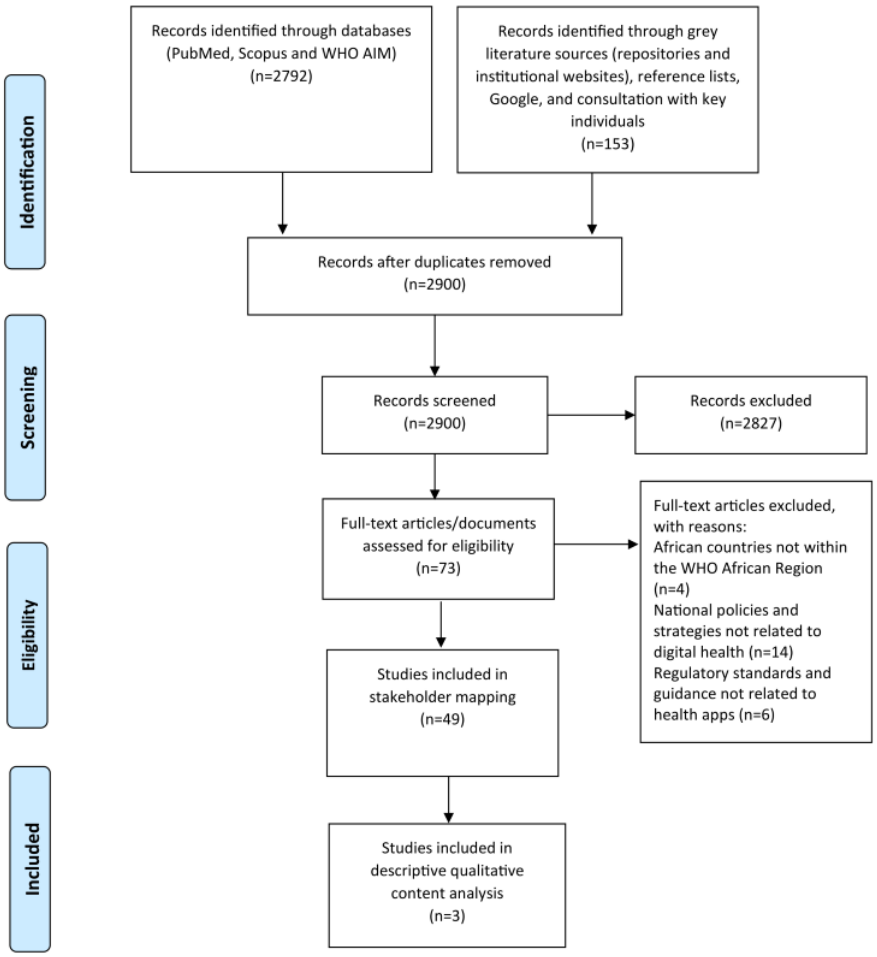

Types of Documents

On the basis of the inclusion criteria, 3 categories of documents were considered for this review, namely "stand-alone regulatory standards and guidance that potentially regulate health apps for self-management," "national policies and strategies on digital health," and "other national documents that relate to the regulation of health apps for self-management." Table 1 presents the types of documents obtained for each country within sub-Saharan Africa.

XSL•FO
RenderX

Table 1. Types of documents obtained for each sub-Saharan African country.

Country	Type of document		
	Stand-alone regulatory standards and guidance	National policies and strategies on digital health	Other related national documents
Algeria			
Angola			
Benin		✓	
Botswana		✓	
Burkina Faso		✓	
Burundi		✓	
Cameroon		✓	
Cape Verde			
Central African Republic			
Chad			
Comoros		✓	
Côte d'Ivoire (Ivory Coast)		✓	
Democratic Republic of the Congo		✓	
Equatorial Guinea			
Eritrea			
Eswatini		✓	
Ethiopia	✓	✓	✓
Gabon		✓	
Gambia		✓	
Ghana		✓	
Guinea			
Guinea-Bissau			
Kenya	✓	✓	✓
Lesotho		✓	
Liberia		✓	✓
Madagascar		✓	
Malawi		✓	
Mali		✓	
Mauritania			
Mauritius		✓	
Mozambique		✓	
Namibia			
Niger		✓	
Nigeria	✓	✓	✓
Republic of the Congo (Congo Brazzaville)			
Rwanda		✓	
São Tomé and Príncipe			
Senegal		✓	
Seychelles			
Sierra Leone			

XSL•FO
RenderX

Country	Type of document		
	Stand-alone regulatory standards and guidance	National policies and strategies on digital health	Other related national documents
South Africa		✓	✓
South Sudan			
Tanzania		✓	✓
Togo		✓	
Uganda		✓	
Zambia		✓	
Zimbabwe		✓	

Characteristics of the Included Documents

Stand-Alone Regulatory Standards and Guidance

We identified and included 6 stand-alone regulatory standards [18,19,46-49] from 3 countries (Ethiopia, Kenya, and Nigeria). All 6 documents were written in English. The years of development ranged between 2013 and 2021, as indicated in Multimedia Appendix 3. The years of implementation were not specifically stated.

Although none of the included regulatory standards were exclusively developed to regulate health apps for self-management, 3 of them (Kenya Standards and Guidelines for mHealth Systems [18], Kenya Standards and Guidelines for E-Health Systems Interoperability [47], and Health Sector Information and Communications Technology Standards and Guidelines [48]) provided concept and information relevant to the regulation of health apps and were included in the qualitative content analysis. The Kenya Standards and Guidelines for mHealth Systems [18] provides standards and guidelines on the design, development, and implementation of mobile health (mHealth) solutions to ensure they are interoperable, scalable, and sustainable. The Kenya Standards and Guidelines for E-Health Systems Interoperability [47] outlines the principles, requirements, and standards for eHealth systems interoperability in Kenya. The Health Sector Information and Communications Technology Standards and Guidelines [48] provide guidance and a consistent approach across the health sector in Kenya for establishing, acquiring, and maintaining current and future information systems and information and communications technology (ICT) infrastructure that foster interoperability across systems. These 3 documents are a good combination of regulatory standards and guidance that provide content and context relevant to the regulation of health apps in sub-Saharan Africa.

The remaining 3 standards (standard for electronic health record [EHR] system in Ethiopia [19], standards and guidelines for electronic medical record systems in Kenya [46], and the health information exchange standard operating procedure and guideline [49]) were exclusively developed for EHRs or electronic medical records. However, they contain information relevant for mapping stakeholders with potential roles in regulating health apps for supporting self-management.

National Policies and Strategies on Digital Health

This review includes 35 national policies and strategies that are related to digital health (potentially covering health apps) [50-84] from 31 countries written in English, French, and Portuguese (Benin, Botswana, Burkina Faso, Burundi, Cameroon, Comoros, Côte d'Ivoire [Ivory Coast], Democratic Republic of the Congo, Eswatini, Ethiopia, Gabon, Ghana, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, South Africa, Tanzania, Togo, Uganda, Zambia, and Zimbabwe). Although the literature search was conducted in English, it also yielded documents written in French and Portuguese from the ICTworks repository. The years of development and implementation range between 2005 and 2030. Policies and strategies written in French and Portuguese were translated into English using Google Translate. Documents labeled as national development plans, strategic plans, and strategic development plans were considered as national strategies.

National policies and strategies do not offer specific standards or guidance, but rather outline the country's vision, policy directions, and strategies for using digital technologies in health care. They provide useful information for identifying digital health stakeholders who can play a role in regulating health apps for self-management. For example, Nigeria has a separate National Digital Health Policy [72] and a National Digital Health Strategy [71]. Both documents were developed by building on the lessons learned from the end-term evaluation of the previous National Health ICT Strategic Framework [85]. They describe Nigeria's renewed vision, mission, goals, objectives, and strategies for the development and implementation of digital health with the aim to improve the quality, efficiency, and effectiveness of health service delivery and health outcomes.

It is worth noting that for countries with >1 policy or strategy, we included only the most recent versions. For instance, as mentioned earlier, Nigeria now has both a national digital health policy and a national digital health strategy. These 2 documents supersede and thus replace the old National Health ICT Strategic Framework [86]. Details of included documents are presented in Multimedia Appendix 3.

Other Related National Documents

We included 8 other documents [20,85,87-92] from 6 countries (Ethiopia, Kenya, Liberia, Nigeria, South Africa, and Tanzania)

XSL•FO
RenderX

that did not fall under either stand-alone regulatory standards and guidance or national policies and strategies. These were mostly frameworks, road maps, and reports that potentially provide information relevant to the use of health apps. The years of development and implementation range from 2016 to 2025. These documents do not provide standards or guidance, but they contain information that can help map the digital health stakeholders that potentially play a role in regulating health apps for self-management. When multiple versions of a document exist, only the latest version was taken into consideration. Multimedia Appendix 3 provides details of the included documents.

Content: Aspects That Are Regulated and Aspects That Are Not

Technical and Clinical Safety

Technical and clinical safety standards are required to prevent or minimize the harm that may arise from the use of the health ICT systems (including mHealth systems) as well as to improve the health outcomes and user satisfaction. As shown in Figure 2, two subthemes were generated from included standards [18,47,48] as content under technical and clinical safety: v(1) guidance on system quality and (2) guidance on software or app development, acquisition, support, and maintenance.

Figure 2. Summary of themes and subthemes covering content, context, process, and actors of the regulatory standards. The process is divided into development and implementation process. The subthemes relating to content and context are further categorized under technical and clinical safety, data protection and security, standards and interoperability, and inclusion and equitable access to services. ICT: information and communications technology.

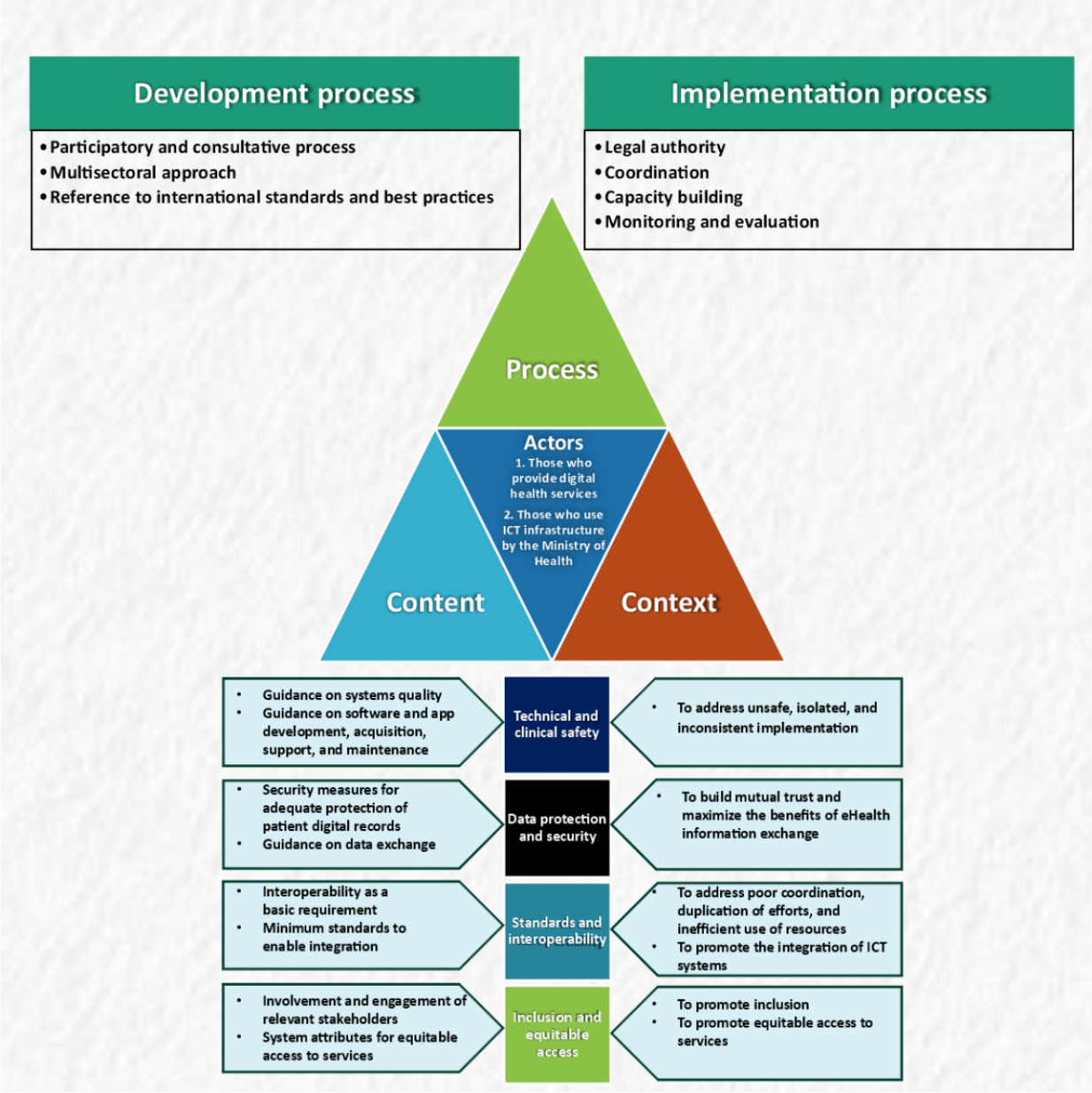

Notably, 2 of the included standards [18,47] provide guidance on system quality to ensure the quality, security, reliability, performance, and maintenance of eHealth and mHealth systems. The Kenya Standards and Guidelines for E-Health Systems Interoperability [47] recommend the implementation of a data quality protocol to ensure that the data collection, collation,

XSL•FO
RenderX

analysis, interpretation, dissemination, and use are managed in accordance with the quality standards. Similarly, the Kenya Standards and Guidelines for mHealth Systems [18] recommends the inclusion of the following requirements in the technical manual: (1) minimum hardware requirements that should incorporate the preferred hardware architecture, (2) minimum software requirements that should include the minimum version of the underlying operating system as well as acceptable versions of related software, and (3) a detailed list of software dependencies (external libraries) necessary for the system to function properly.

The included standards [18,48] cover guidance on software or app development, acquisition, support, and maintenance, which aim to ensure the efficiency and effectiveness of eHealth and mHealth systems. The Kenya Standards and Guidelines for mHealth Systems [18] recommends a technical manual to provide a detailed description of the system's installation and maintenance processes for system administrators and implementers; a developer's guide for software developers and programmers to provide them with an overview of the system, description of the software design methodologies, description of the system architecture, and technical design diagrams; and a user manual to aid users in understanding how the system works and how each feature operates; in addition, the technical manual contains instructions for operating the software; entering and updating data; and generating, saving, and printing reports.

Although the contents generated here provide guidance that is relevant to health apps, they are not specific to health apps. Moreover, there are no clear measures to enable individuals or organizations that use health apps to manage clinical risk appropriately.

Data Protection and Security

Data protection and security are crucial aspects of managing patient information, thus ensuring the confidentiality, integrity, and availability of data as well as the rights and interests of the patient. Two subthemes related to data protection and security are (1) security measures for adequate protection of patients' digital records and (2) guidance on data exchange.

The included standards [18,48] provide security measures for eHealth or mHealth systems to ensure the adequate protection of digitally accessible patient records. These measures include authentication, accountability, identification, authorization, integrity, confidentiality, availability, security, administration, and audit. This will help to achieve confidentiality, integrity, availability, and nonrepudiation of patient data or health records. Additional levels of security such as data encryption are required when there is a need to store sensitive information on removable devices or media or outside the MOH premises.

The Kenya Standards and Guidelines for mHealth Systems [18] provide the following guidance on data exchange to ensure privacy: (1) anonymize client data as much as possible before they can be shared; (2) where possible, use pseudonyms for the client data before they can be shared; (3) aggregate client data before they can be shared to reduce possibilities of tracing the data back to the client; and (4) minimize data so that access is available only to the data set required for that particular use. With regard to privacy rules, the Kenya Standards and Guidelines for E-Health Systems Interoperability [47] propose that a notice of privacy practices should be given to patients describing how their information may be used or shared while also specifying their legal rights.

Standards and Interoperability

Standards and interoperability are essential concepts in the field of IT, especially for systems that need to communicate and exchange data, as seen in the use of health apps for self-management. Two subthemes related to standards and interoperability are (1) interoperability as a basic requirement and (2) minimum standards to enable integration.

All the regulatory standards [18,47,48] highlight the importance of having interoperability as a basic requirement when selecting software products or services for use within the health system. This facilitates interaction across systems. For instance, to facilitate seamless interaction between mHealth systems and primary information systems for data capture, reporting, and decision support in various domains of the health system, the Kenya Standards and Guidelines for mHealth Systems [18] recommends the incorporation of at least 3 types of interoperability, namely, technical interoperability, semantic interoperability, and process interoperability.

Furthermore, 2 regulatory standards [18,47] proposed minimum interoperability standards to enable the integration of services and data exchange between various systems in health care. For instance, the Kenya Standards and Guidelines for mHealth Systems [18] suggests standards (for interoperability) for mHealth systems that are consistent with the recommendations in internationally accepted standards. They include the following: (1) clinical messaging—ensuring mHealth systems conform to Health Level 7 (HL7) version 3 standards and corresponding implementation guideline; (2) clinical terminology—ensuring terminologies and classifications for clinical concepts (eg, International Classification of Diseases, tenth revision—for diseases; Systemized Nomenclature of Medicine—for clinical data coding; Logical Observation Identifiers Names and Codes—for laboratories; and RxNorm—for Pharmacies); (3) the mHealth system must use the latest versions of international standards, such as HL7 Clinical Document Architecture for electronic sharing of clinical documents; (4) concepts—mHealth systems will use the idea of "concepts" so that information can be transmitted between systems without losing meaning or context, and HL7 Reference Implementation Model or other appropriate standards are recommended for implementing concepts; (5) architecture—to develop mHealth systems, developers should define the system architecture that should include data elements and business logic. Furthermore, to define how mHealth systems interact with other systems, developers of mHealth solutions must provide application programming interfaces. FHIR is the preferred application programming interface interoperability standard.

Inclusion and Equitable Access

Inclusion and equitable access are essential principles to ensure that health apps are culturally appropriate and relevant and

XSL•FO
RenderX

accessible to everyone, regardless of gender, ethnicity, location, or economic status.

All the included regulatory standards [18,47,48] indicate that they were developed based on a combination of participatory and consultative approaches involving multiple actors or stakeholders, thus promoting inclusion. However, there are no specific measures or guidance to ensure adequate engagement and representation of all the relevant stakeholders and to sustain that engagement.

The Kenya Standards and Guidelines for mHealth Systems [18] proposes the following systems attributes to ensure equitable access to mHealth services at all times and from anywhere: (1) allocation of adequate storage and bandwidth capacity; (2) fast response time; (3) fast recovery capabilities; (4) performance monitoring; (5) business continuity processes, for example, backups; and (6) redundant sites and links. Furthermore, the Kenya Standards and Guidelines for mHealth Systems [18] prescribes the following metrics for measuring system availability: (1) downtime per year, (2) mean time between failure, (3) mean time to repair, and (4) failure in time.

Although the abovementioned systems attributes and metrics for measuring system availability are important, the included standards do not offer any concrete guidance or model for achieving a sustainable funding mechanism for health apps to ensure that they are readily available and accessible to those who need them.

Context: Reasons Why Those Aspects Are Regulated

Technical and Clinical Safety

The 3 standards [18,47,48] were developed to address unsafe, isolated, and inconsistent implementation. The Health Sector ICT Standards and Guidelines [48] suggest that although there has been a lot of ICT investment in the health sector leading to improvement in service delivery and information exchange, there remains the challenge of inconsistency in ICT implementation and harmonization of the health sector system requirements. Hence, there is a need to adopt global best practices for software development, acquisition, support, and maintenance by the MOH. In addition, the Kenya Standards and Guidelines for mHealth Systems [18] indicates that standards and guidelines are necessary to ensure a consistent approach to the development of ICT systems. Similarly, the Kenya Standards and Guidelines for E-Health Systems Interoperability [47] recognize the need to ensure that the processes of collecting, collating, analyzing, interpreting, disseminating, and using data are consistent with data quality standards.

Data Protection and Security

To build mutual trust and maximize the benefits of eHealth information exchange, the Kenya Standards and Guidelines for E-Health Systems Interoperability [47] reiterate that as health data are constantly being exchanged across health information systems, robust security standards are required to maintain their integrity and confidentiality. This will build the trust of service users and consequently help to maximize the benefits of eHealth information exchange such as in self-management.

Standards and Interoperability

Two of the included regulatory standards [47,48] indicate that the context for standards and interoperability was (1) to address poor coordination, duplication of efforts, and inefficient use of resources and (2) to promote the integration of ICT systems.

The Kenya Standards and Guidelines for E-Health Systems Interoperability [47] acknowledge that the absence of interoperability standards over the years has led to the duplication of efforts and the inefficient use of ICT resources in health care. Now that ICT has become increasingly relevant in improving efficiency in health service delivery, the Kenya MOH recognizes the need to adopt a standardized approach, hence the development of interoperability standards for eHealth systems. In addition, the Health Sector ICT Standards and Guidelines [48] emphasize the relevance of interoperability as a requirement for addressing the inconsistency in implementing ICT in the health sector.

The Health Sector ICT Standards and Guidelines [48] consider "integration of ICT systems" as one of its key guiding principles, acknowledging the lack of information systems integration as a challenge experienced by ICT services across Kenya.

Inclusion and Equitable Access

The contexts for inclusion and equitable access as generated from included standards [18,47,48] were (1) to promote inclusion and (2) to promote equitable access to services.

To promote inclusion, the standards [18,47,48] highlight the importance of involving and engaging multiple actors and stakeholders during the development process. However, no emphasis was placed on the need to sustain stakeholder engagement during the implementation process.

Pertaining to equitable access, the Kenya Standards and Guidelines for mHealth Systems [18] acknowledges that the public health care system is largely unavailable to most of the population in many developing countries because of geographical location, resource constraints, inefficiencies, and lack of awareness. Hence, it recognizes the importance of ensuring that mHealth services are always accessible by users and from anywhere as well as the need to put in place mechanisms to make this happen.

Process: How the Regulations Are Developed and Implemented

Two themes were generated from the included standards: development and implementation processes [18,47,48].

Development Process

All the included standards [18,47,48] indicate that they were developed through a participatory process and in consultation with a range of subject experts and interest groups. In addition, the standards [18,47,48] adopted a multisectoral approach to engage health-related stakeholders from government ministries or agencies and development partners and a range of subject experts and interest groups. It has also been reported that these standards [18,47,48] were developed based on international best practices and with reference to international standards. However, there is no indication that a stakeholder engagement strategy

XSL•FO
RenderX

was adopted to sustain the engagement of stakeholders through the entire development and implementation process.

Implementation Process

The 3 regulatory standards [18,47,48] identify the key requirements to ensure effective implementation of IT services in the health sector. These are (1) legal authority, (2) coordination, (3) building capacity, and (4) monitoring and evaluation.

The included standards [18,47,48] were established based on the legal provisions enshrined in the health and other related acts and laws of Kenya as well as the relevant policies and strategies. Hence, it is expected that their implementation will comply with and be backed by those legal provisions. For example, the Health Sector ICT Standards and Guidelines [48] indicate that its implementation will be supported by the authority from the Kenya Communications Act 2009, E-Government Strategy, and National ICT Policy. Similarly, the Kenya Standards and Guidelines for mHealth Systems [18] asserts that it will be implemented by complying with existing and relevant national policies, legal frameworks, strategies, and standards, including the Health Information Policy, ICT Standards, and System Interoperability Principles.

The included standards [18,47,48] report that the implementation of regulations will require robust coordination mechanisms. For instance, the Health Sector ICT Standards and Guidelines [48] indicate that, as the Ministry's ICT resource manager, the principal secretary (also the head of ICT), in collaboration with the ICT Governance Committee, is responsible for coordinating the implementation of the standard. The ICT Governance Committee comprises representatives from the heads of departments and ICT development partners in the health sector. The committee's responsibilities include overseeing, enforcing, and reviewing standards as well as initiating ICT projects.

The Health Sector ICT Standards and Guidelines [48] highlight the need for capacity building or training of the MOH staff and stakeholders who are the primary users of the Ministry's ICT services. This will enhance their capacity to implement the guidelines provided in the document in line with the ministry's human resource development policies, regulations, and rules. However, it is acknowledged that building capacity for health ICT is a challenge given that there is low adoption of ICT among health providers, and ICT is not routinely included in the course content of most training programs. The Kenya Standards and Guidelines for mHealth Systems [18] listed the "number of mHealth practitioners trained on the standards and guidelines" as one of the indicators for monitoring and evaluating mHealth interventions.

The Health Sector ICT Standards and Guidelines [48] assert that monitoring and evaluation is an essential role of the MOH to ensure efficiency, accountability, and transparency throughout the implementation period. It further stresses that all those who use the Ministry's ICT services are required to adhere to the provisions in the standard as the MOH will carry out quarterly monitoring exercises on the use of the standard to ensure compliance based on clear indicators. Furthermore, the ICT Governance Committee will periodically review and amend the standard to keep it relevant and effective. Similarly, the Kenya Standards and Guidelines for mHealth Systems [18] establishes the following key indicators for effectively monitoring and evaluating the implementation of the standards and guidelines: (1) the number of counties in which the MOH has disseminated the standards and guidelines, (2) the number of counties successfully implementing the standards and guidelines, (3) the number of mHealth practitioners trained on the standards and guidelines, (4) the number of mHealth practitioners accessing the standards and guidelines, (5) the number of mHealth practitioners who correctly understand the standards and guidelines, (6) the number of stakeholders who adhere to the standards and guidelines, (7) the number of mHealth systems that follow the required development steps, and (8) the number of mHealth practitioners who have implemented their systems by using the standards and guidelines. In addition, the Kenya Standards and Guidelines for mHealth Systems [18] indicates that the outlined standards will be reviewed every 3 years to ensure they are up to date with new changes including the changes in policies and systems upgrades.

Although all the abovementioned indicators are relevant, the implementation process is not explicit on the approach for regulating health apps and ensuring compliance with regulatory standards and guidance.

Actors: Those the Regulations Are Targeted at

The included standards [18,47,48] identified 2 main groups of actors for whom the regulations and guidance were targeted. They included (1) those who provide digital health services and (2) those who use the ICT infrastructure of the MOH.

Two of the standards [47,48] indicated that the regulations should be implemented by all individuals and organizations that provide ICT-related health care services to the public. Similarly, the Health Sector ICT Standards and Guidelines [48] state that all those who access or use the MOH ICT infrastructure are expected to adhere to the guidelines outlined in the document.

Mapping of Stakeholders

To address the third research question, we conducted a stakeholder mapping guided by the RISA tool [41].

A total of 11 categories of key stakeholders were identified from all 49 included documents (6 stand-alone regulatory standards and guidance, 35 national policies or strategies, and 8 other related documents). These categories are consistent with the digital health stakeholders recognized by the WHO, ITU, and UNESCO [32,33,43]. Table 2 presents the mapping of stakeholders according to their role categorization. A more detailed table with a potential role description with regard to regulating health apps for self-management is presented in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Mapping of stakeholders according to their potential role with regard to regulating health apps for self-management.

Stakeholder category	List of stakeholders	Role categorization
A1: Government (health sector)	• Ministry of Health • Relevant departments and agencies, including the National Medicines Regulatory Authority	Coordination and provision of an enabling environment
A2: Government (non–health sector)	• Ministry of Power or Energy • Ministry of Information and Communications Technology or Telecommunication • Ministry of Education • Ministry of Science and Technology • Ministry of Finance • Ministry of Justice	Coordination and provision of an enabling environment
B: Regulatory bodies	• Relevant health regulatory agencies • Ministry of Justice • Law enforcement agencies	Compliance
C1: Funding bodies	• Donors and aid agencies • Foundations and development banks • The private sector • Other health care funders	Funding and insurance
C2: Insurance	• The insurance industry	Funding and insurance
D1: Intergovernmental, international, and continental organizations	• African Union • WHO[a] or WHO Regional Office for Africa • International Telecommunication Union • World Bank • United Nations Children's Fund	Strategic support
D2: Nonstate actors	• Nongovernmental organizations • Civil society organizations • Faith-based organizations	Strategic support
E1: Industries and businesses that influence the use of health apps	• App developers • Network or internet providers • App evaluators	Resources and skills
E2: Academia and research bodies and institutions	• Universities • Teaching hospitals • Research institutes	Resources and skills
E3: Professionals in research and practice or	• Subject matter experts	Resources and skills
F1: The health care community (providers)	• Health care providers (eg, hospitals, clinics, and primary health cares) • Health care professionals	Service delivery and use
F2: The health care community (users)	• Patients • Caregivers • Families • Community groups	Service delivery and use

[a]WHO: World Health Organization.

Discussion

Overview

This paper presents the findings of a scoping review of regulatory standards and guidance for the use of health apps for self-management in sub-Saharan Africa. To the best of our knowledge, this is the first study that attempted to identify and assess the extent to which regulatory standards and guidance regulate and guide the use of health apps for self-management in sub-Saharan Africa as well as map out the key stakeholders and their potential roles.

Our findings reveal that only 1 country (Kenya) in sub-Saharan Africa currently has national regulatory standards that could potentially regulate the use of health apps for self-management.

XSL•FO
RenderX

The included standards failed to adequately address adequate attention to inclusion and equitable access. This is concerning given the growing need to promote the adoption of culturally appropriate and relevant health apps and to ensure that they are available to those who need them regardless of gender, ethnicity, geographical location, or financial status [24-29]. Consequently, this review provides insights into the regulation of health apps for self-management in sub-Saharan Africa, which needs to be given more attention if the potential of these apps is to be harnessed in the region.

Principal Findings

We identified 49 documents from 31 countries in sub-Saharan Africa. Although none of the included standards provided a specific set of regulations on health apps for self-management, we identified 3 standards [18,47,48] that provided relevant information regarding the regulation of health apps. The included national policies and strategies, in contrast, only outline the goals and commitments made by national governments to promote the adoption of digital technologies in the health sector and the plans and paths set forth to achieve these goals. However, the information they provided was relevant for identifying and mapping digital health stakeholders who potentially have vital roles in regulating the use of health apps for self-management.

The policy analysis framework (content, context, process, and actors) [42] was adapted and applied to organize the key findings. The content covered the following areas: guidance on systems quality; guidance on software and app development, acquisition, support, and maintenance; security measures for adequate protection of patients' digital records; guidance on data exchange; interoperability as a basic requirement; minimum standards to enable integration; involvement and engagement of relevant stakeholders; and system attributes for equitable access to services. Meanwhile, the context was to address unsafe, isolated, and inconsistent implementation; to build mutual trust and maximize the benefits of eHealth information exchange; to address poor coordination, duplication of efforts, and inefficient use of resources; to promote the integration of ICT systems; and to promote inclusion and equitable access to services. The process involved the development process (which covers participatory and consultative processes and multisectoral approach, with reference to international standards and best practices) and the implementation process (which covers legal authority, coordination, capacity building, and monitoring and evaluation). The targeted actors were those who provided digital health services and those who used the ICT infrastructure of the MOH.

Furthermore, key stakeholders with potential roles in regulating health apps for self-management were identified. They include the government, regulatory bodies, funders, intergovernmental and nongovernmental organizations, academia, and the health care community.

Implications of the Study Findings for Practice

Overview

Regulatory standards and guidance act as a bridge between technological innovation and its safe and effective use in health care. They ensure that while technology continues to advance, the safety and trust of patients are never compromised. Among the plethora of health apps on the market, the over-the-counter, nonregulated apps such as wellness and fitness apps are the most mainstream [93-95]. On the other side of the spectrum, there are regulated health apps that are classified under medical devices or software as medical device products [94,95]. Some of these are prescription-only apps, such as digital therapeutics (DTx) apps for managing substance dependence [95,96].

Although some high-income countries have made significant strides in ensuring the safety, effectiveness, and accessibility of health apps, the journey has indeed not been without challenges and hurdles. Sub-Saharan Africa, although dealing with its own unique set of challenges, has the opportunity to learn from the experiences of these high-income countries. This could potentially allow the region to bypass some of the hurdles encountered by high-income countries in their journeys.

Technical and Clinical Safety

Technical and clinical safety are essential requirements that health apps must meet before they can be considered for use for self-management to minimize the risk of harm to patients. It is well documented that health apps that function poorly pose a serious threat to the safety of patients. An example illustrating how health apps used for self-management can threaten patient safety is evident in a study [12]. This study [12] revealed that widely used health apps designed to calculate and estimate insulin doses could endanger patients by providing incorrect or inappropriate dose recommendations. Similarly, 2 successive studies that assessed the contents and tools of apps for asthma discovered that none of the apps in the first study offered comprehensive information or adequate tools for asthma self-management, whereas the follow-up study, which was conducted 2 years later, showed a 2-fold increase in the number of asthma apps, yet there was no improvement in the content and tools offered by the newer apps. In fact, many apps recommended self-management procedures that were not supported by evidence [13,14]. Accordingly, some health apps that support the self-management of long-term conditions do not adhere to evidence-based guidelines and are unresponsive to the evolving health needs of patients.

Although the context of included regulatory standards with regard to technical and clinical safety was to address unsafe, isolated, and inconsistent implementation, the guidance provided by these regulatory standards is not specific to health apps, and they do not provide appropriate guidance and standards for health organizations and other key stakeholders to establish a framework for managing the clinical risks associated with deploying and implementing self-management health apps. Considering the rapid advancements in digital health (including artificial intelligence [AI] or machine learning and big data), health apps will increasingly play a crucial role in supporting self-management through digitally enabled care pathways that will improve personalized care and health outcomes [97,98]. Therefore, it is imperative to ensure the technical reliability and clinical safety of health apps for self-management through robust regulatory standards and guidance. For instance, a guide on the criteria for health app assessment, developed by the UK

XSL•FO
RenderX

government, includes technical stability and clinical safety as criteria for deciding whether health apps should be considered for use in the National Health Service (NHS) [99]. In addition, medical device apps are required to conform to the NHS clinical risk management standards as part of the clinical safety requirements [99,100]. In the event of any concerns regarding the safety of a medical device app, the Yellow Card reporting system can be used by a responsible clinical safety officer or any other individual to notify the Medicines and Healthcare products Regulatory Agency (MHRA) [101,102].

Data Protection and Security

To adequately manage patient information when health apps are used for self-management, data protection and security standards and guidance are required. They guarantee that data are kept and handled safely and responsibly within the provisions of the law and that patients' rights and interests are respected.

There have been ongoing concerns about compliance with ethical standards, the principles of confidentiality of information, and data privacy. For example, an assessment of apps that had previously been endorsed by the former UK NHS Apps Library revealed substantial gaps in compliance with data protection principles regarding the collection, storage, and transmission of personal information. This has raised a fundamental concern about the credibility of developer disclosures and whether these disclosures can be trusted by certification programs [15]. A study assessed the privacy practices of the 36 most popular apps for depression and smoking cessation for Android and iOS in the United States and Australia [16]. The findings revealed that although only 69% (25/36) of the apps included a privacy policy, 92% (33/36) of the apps shared data with a third party, and only 92% (23/25 with privacy policy) of the apps disclosed sharing data with a third party in their policy. Although 81% (29/36) of the apps shared data with Google and Facebook for the purposes of advertising, marketing, or analytics, only 43% (12/28) of the apps that shared data with Google and 50% (6/12) of the apps that shared data with Facebook disclosed this in their policy [16].

In this regard, health app developers and providers in the United Kingdom are required to conduct a data protection risk assessment before they launch or update their apps to ensure compliance with the United Kingdom General Data Protection Regulation (GDPR) and other relevant regulations, including the Data Protection Act 2018 [103]. By conducting a data protection risk assessment, health app developers and providers can demonstrate that they are accountable; they respect the privacy and dignity of their users; and that they deliver safe, effective, and ethical solutions [104].

Health apps are expected to play an increasingly important role in supporting self-management. However, this ambition can only be achieved if citizens trust that these apps are collecting and analyzing data safely and in accordance with robust regulatory standards and guidance. It is also crucial that these apps provide reliable information that clinicians can act on [98]. The context of the standards included in this study regarding data protection and security was to build mutual trust and maximize the benefits of eHealth information exchange. Trust is a key factor in the successful adoption and use of health apps, and transparency in data handling and clinical decision-making is essential to build and maintain that trust. This is also paramount for the widespread acceptance and impact of health apps on health care outcomes in sub-Saharan Africa.

We acknowledge the existence of numerous national laws related to data protection and security outside the health sector. Hence, guidelines that link these legislations together must be provided to ensure compliance with all relevant laws and guidance when using patient data. An example of how to achieve this is the United Kingdome's guide to good practice for digital and data-driven health technologies that provides guidelines on how to abide by the laws and principles that govern data security and protection in the United Kingdom, including the GDPR, Data Protection Act 2018, and Caldicott Principles [105].

Standards and Interoperability

Standards and interoperability are essential for effectively developing, deploying, and implementing health apps to support self-management in sub-Saharan Africa. Interoperability is the ability of different systems, devices, or applications to communicate and exchange data with each other in a coordinated manner, thus providing timely and seamless portable information across organizational, regional, and national boundaries and optimizing both individual and population health [106]. In the same vein, standards enable interoperability between systems or devices through a common language and a common set of expectations [106].

Interoperability is crucial in improving the quality, safety, and efficiency of care delivery as well as empowering patients and providers with access to relevant and timely information [99]. One of the most widely used and accepted interoperability standards for health care data exchange is FHIR [106,107]. FHIR is a global industry standard developed by HL7 International. FHIR is designed to be quick to learn and implement and to support a variety of use cases, including self-management [108]. By using apps that are based on an FHIR standard, patients can benefit from data analytics that show how their health data relate to their chronic conditions or wellness goals [109]. They could also access all their health information from one place, even if they visit different health professionals who use different electronic medical records or EHR, thus promoting integrated care [28,31,33,109-115]. As a result, patient care can easily be coordinated.

The context of the included regulatory standards with regard to standards and interoperability was to address poor coordination, duplication of efforts, and inefficient use of resources and to promote the integration of ICT systems. However, in sub-Saharan Africa, there are many challenges and barriers to the adoption and implementation of interoperability standards, such as the lack of awareness or knowledge of the benefits and requirements of interoperability standards among stakeholders; lack of incentives or regulations to encourage or enforce the adoption of interoperability standards by app developers and vendors; lack of resources or capacity to implement interoperability standards, including technical expertise, infrastructure, funding, or governance; and lack of alignment or coordination among the different actors and initiatives

XSL•FO
RenderX

involved in developing, deploying, and implementing the digital health interventions [30,116-119]. To address these challenges, some possible solutions may include raising awareness and education on the importance and value of interoperability standards for health apps among all relevant actors; developing and implementing policies and guidelines that promote or mandate the use of interoperability standards by app developers and vendors; providing technical assistance and support for app developers and vendors to adopt and implement interoperability standards, such as tools, frameworks, testing, certification, or accreditation; and establishing and strengthening collaboration and coordination among the different stakeholders and initiatives involved in health app development, deployment, and implementation in sub-Saharan Africa. In addition, the Digital Health Platform Handbook, a toolkit developed by the collaborative efforts of the WHO and ITU [120], can help countries in sub-Saharan Africa to develop and implement digital health platforms as the underlying infrastructure for interoperable and integrated national digital health systems. The digital health platform is a system-wide approach to developing digital health solutions with the aim to overcome the problems of siloed, vertical, and isolated applications and systems that hamper data management, innovation, efficiency, and impact in the health sector.

Inclusion and Equitable Access

Inclusion and equitable access are crucial to ensuring that health apps and related services are culturally appropriate and relevant as well as accessible to all who need them, regardless of gender, ethnicity, geographical location, ability, or financial status [24-29]. This is the key to promoting a “sense of belonging” and “ownership” and thus underscoring the importance of stakeholder mapping and involvement or engagement through the development and implementation process [22].

In this study, the included regulatory standards demonstrate the importance of inclusion by adopting both a participatory and consultative approach involving multiple stakeholders from different sectors. However, the standards do not provide clear guidance to ensure the adequate participation and sustained engagement of all relevant stakeholders. The lack of concise guidance to ensure the adequate participation and engagement of all relevant stakeholders, especially the susceptible and disadvantaged groups, can increase the risk of tokenistic tendencies, which can undermine the cultural appropriateness of health apps [25,121]. Some susceptible groups, such as women and people with low socioeconomic status, may face additional barriers to accessing and using health apps, such as lack of digital literacy, privacy concerns, cultural norms, or stigma [25]. Similarly, the cost of developing, maintaining, and updating health apps may not be covered by public or private health insurance schemes, which could limit their affordability and availability for low-income or uninsured populations [95]. However, there is no specific guidance or model for an effective funding mechanism for health apps in the included regulatory standards.

To address these challenges and ensure equitable access to health apps for self-management in sub-Saharan Africa, possible measures may include developing policies and regulations that support integrating health app interventions into existing health systems and financing mechanisms and engaging with stakeholders from different sectors and backgrounds (including health professionals, patients, communities, governments, civil society, academia, and industry) to co-develop and co-implement frameworks or models that promote the use of health apps for self-management in ways that are responsive to the local context and needs. Moreover, establishing regulations that provide appropriate financing or reimbursement options will reduce the risk of developers of good quality health apps turning to data mining for revenue, thus increasing privacy concerns [95]. For instance, in Germany, the reimbursement of health apps classified as medical devices (Digitale Gesundheitsanwendungen) was introduced in 2021 under the statutory health insurance [122,123]. When a medical device is prescribed by a physician or a physiotherapist, the manufacturer must submit an application to the German Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte) for approval [123]. The Federal Association of the Statutory Health Insurance Funds (Spitzenverband Bund der Krankenkassen) determines and negotiates the reimbursement thresholds following approval. However, the manufacturer must demonstrate that the app is safe, functional, and of good quality; complies with data protection requirements; and benefits patient care [123].

Process

The process of regulating health apps essentially involves the development and implementation of regulatory standards and guidance. According to our study, the development process comprises a participatory and consultative process, a multisectoral approach, and a reference to international standards and best practices. In contrast, the implementation process is ongoing and requires appropriate legal authority, coordination, capacity building, and monitoring and evaluation.

We recognize that health apps can be accessed and used by patients from different parts of the world, and this means that countries need to carefully consider whether health apps that are accessed and used by their citizens meet the national or regional legal and ethical requirements, including their cultural and linguistic needs [23]. For countries in sub-Saharan Africa, a cross-border or regional collaboration between national legal authorities through the coordination of agencies such as the African Medicines Regulatory Harmonization (AMRH) may help to ensure that health apps built for the region are safe, effective, and user-friendly for everyone, considering the contextual differences of the countries [23]. For instance, all medical device companies that want to sell their products in the European market must obtain a Conformité Européenne (CE) mark for their devices, which indicates that they meet the legal requirements and can be freely circulated within the European Union [124]. Although the European Union member states regulate medical devices, the European Medicines Agency is involved in the regulatory process.

The regulation of health apps is extremely complex and involves a wide range of stakeholders. Therefore, a robust coordination mechanism is essential to reduce the risk of fragmentation and duplication of efforts and to promote the efficient use of

XSL•FO
RenderX

resources. Most countries in sub-Saharan Africa have units in health ministries that coordinate and oversee the regulation of medical products. These units should be autonomous, full-fledged departments with legal authority (boards or commissions) to ensure independent, transparent, and accountable decision-making, but this is often not the case [125]. These units are recognized by the national authorities as regulators (eg, the National Medicines Regulatory Authority [NMRA]) [126]. Such organizational structures hinder the effectiveness of the national regulatory authorities in fulfilling their mandate and prevent the establishment of quality management systems to ensure transparent and accountable decision-making [125].

Furthermore, Essén et al [23] analyzed health app policy or regulation in 9 high-income countries (Sweden, Norway, Denmark, Netherlands, Belgium, Germany, England, the United States, and Singapore) and found that most of these countries adopted centralized approaches to app evaluation. Although centralized approaches might have advantages over self-evaluation, they may create bottlenecks and limit the availability of high-quality health apps for users. As suggested by Essén et al [23], a decentralized approach, such as the accreditation of evaluation agencies, maybe a worthwhile solution. However, this will require adequate coordination to ensure the consistency and reliability of the evaluation criteria and methods across different agencies as well as the transparency and accountability of the accreditation process. A possible way to achieve this is to adopt a common framework that can guide the evaluation and accreditation of health apps.

Similarly, the postmarket surveillance (PMS) system, which is a new regulation for medical devices in Europe, is a process of collecting and analyzing data on medical devices after they have been launched into the market to ensure their safety and performance and to identify any problems or need for improvements [127,128]. The PMS system is important because premarket data, which are obtained from testing a medical device before it is launched, have limitations in capturing the long-term performance and risks of the device [128]. Currently, the PMS system does not cover fitness and wellness apps, which are commonly used in self-management. Hence, Yu [93] proposed that the PMS system should also be applied to DHTs, such as fitness and wellness apps. They argue that the postmarket data would help regulators periodically review and adjust the regulatory standards for these groups of health apps based on their risks and benefits.

Drawing on the experience of the United Kingdom, it can be clearly demonstrated that the regulation of health apps is a complex, a multifaceted, and an evolving process that involves different regulators and criteria depending on the nature and function of the app. For instance, a centralized NHS Apps Library was launched as a beta site in April 2017 to provide patients with a collection of trusted and easy-to-use digital health tools [129]. The library provided access to a range of health apps that were reviewed and approved by the NHS, including apps that could help patients manage conditions such as diabetes, mental health, and chronic obstructive pulmonary disease [130]. However, the library was closed in December 2021 [131]. Although no reason for the closure was provided on the website, it is likely because of persistent concerns regarding the safety of patients and data privacy involving multiple apps including those listed in the library [12,14-16,131,132]. The NHS App was introduced in January 2019 before the closure of the NHS Apps Library to serve as the gateway for accessing NHS services including ordering repeat prescriptions and booking or managing appointments [133].

Furthermore, the United Kingdom Health Security Agency, formerly known as Public Health England, issued a guidance on criteria for health app assessment in October 2017 [99]. The purpose of this guidance was to ensure that all health apps built for the UK population work well and provide clear information about their functions, benefits, and intended outcomes for patients and health care professionals. On the basis of this guidance, those intending to build an app are required to conform to certain regulations before being considered for the app assessment process. The 2 main regulations are the medical device regulation and the Care Quality Commission (CQC) registration. Apps that are considered as medical devices must register with the MHRA and have a CE mark. Apps providing health or social care that fit into 1 of 14 regulated activities are required to register with the CQC before they can be assessed [134]. CQC is an independent regulator of health and social care services in England.

Similarly, the Organisation for the Review of Care and Health Apps (ORCHA) is a UK-based organization that independently evaluates and distributes health apps. It provides services such as app review, accreditation, curation, and recommendation within the United Kingdom and across the world [135]. ORCHA also enables organizations (including the NHS) to build a decentralized web-based digital health library of consumer-friendly over-the-counter apps [135-137]. These apps are continuously assessed by ORCHA against the standards and regulations in clinical and professional assurance, data quality and privacy, and usability and accessibility [137].

In addition, the Digital Technology Assessment Criteria (DTAC) were introduced in beta in October 2020, and its first official version was subsequently launched in February 2021 [138]. The DTAC plays a crucial role in ensuring that digital health tools meet the necessary standards in areas such as clinical safety, data protection, technical security, interoperability, usability, and accessibility. By serving as the national baseline criteria for DHTs in the NHS and social care, it provides a valuable framework for health care organizations during procurement. It also offers guidance for developers on the expectations for their digital technologies within the NHS and social care. This is an example of how a harmonized framework can help ensure the quality and safety of DHTs, including health apps.

In addition, the National Institute for Health and Care Excellence Evidence Standards Framework is a set of evidence standards for a wide range of DHTs designed to help evaluators and decision makers in the health care system to consistently identify DHTs that are likely to offer benefits to the users and the health care system [139]. The Evidence Standards Framework was first published in March 2019 and is ideally used before DHTs (including health apps) are considered for commissioning or

XSL•FO
RenderX

procurement by the NHS [140]. It is a crucial tool for ensuring that DHTs are clinically effective and offer value to the health and care system in the United Kingdom. In August 2022, the framework was updated to include AI and data-driven technologies with adaptive algorithms [140].

Furthermore, DTx apps, which are a type of medical device, are not allowed into the UK market unless they comply with the UK GDPR and meet the requirements of DTAC. In addition, they must bear the CE or UK Conformity Assessed marks [141]. This means that DTx apps must demonstrate their safety and efficacy through clinical trials and comply with the relevant regulations for data protection and quality standards as regulated by the MHRA. DTx products are also recognized as DHTs under the National Institute for Health and Care Excellence Evidence Standards Framework [142]. DTx incorporates software to treat, prevent, or manage specific diseases or conditions [143,144]. The fact that DTx products typically focus on a narrow clinical indication and generate evidence of clinical efficacy underscores their potential to make a substantial contribution to self-management and health care delivery in general. The increasing recognition of the role of DTx in patient care by regulators is also noteworthy, and the creation of regulatory and reimbursement pathways for approved apps further enables DTx products to continue to play an important role in impacting health care delivery [1,143]. This is a testament to the potential of regulated health apps to revolutionize health care and improve patient outcomes.

Among the many lessons to learn from the experience of the United Kingdom is that the regulation of health apps must evolve to keep pace with advances in DHTs and adapt to the changing needs and demands of digital health. Moreover, efforts are being made to streamline the multifaceted approaches to simplify app regulation and access in the United Kingdom [23]. Therefore, a robust and dynamic coordination mechanism, along with political will, skilled personnel, reliable funding, and a robust framework for monitoring and evaluating progress and aligning key performance indicators, is essential for countries in sub-Saharan Africa to keep pace with the advancement in the regulation of health apps. There is also a need to strengthen collaboration and ensure regulatory harmonization among national regulatory authorities and continental bodies such as the regional economic communities, AMRH, and the WHO AFRO [126].

Capacity building and monitoring and evaluation are important factors for ensuring effective regulation of health apps given the complex nature of the process. The regulation of medical products (including health apps) in sub-Saharan Africa generally includes licensing and accreditation, evaluation, inspection, quality control, information dissemination and promotion, and monitoring of adverse events [125]. Therefore, high-level skills as well as effective monitoring and evaluation will be required to ensure the success of the process. For most countries in sub-Saharan Africa, the NMRA is responsible for coordinating and overseeing the regulatory system of medical products [125,126]. However, in most cases, NMRAs are unable to perform the core regulatory functions expected of them [145]. More than 90% of African countries have limited or no capacity to regulate medical products, with only 7% having moderately developed capabilities [145]. The lack of effective NMRAs in Africa exposes the citizens to potential harm by allowing unsafe, low-quality, and fake medical products to circulate and be used [145].

Although it is the responsibility of governments to establish functional regulatory systems and ensure effective monitoring and evaluation of the regulatory process, the involvement of international and continental organizations to support sub-Saharan African countries improve the regulatory capacity of their national regulatory agencies would be extremely beneficial. For instance, the African Medicines Agency (AMA) was established in November 2019 as a treaty adopted by the African Union Member States to help address the concerns arising from weak regulatory systems on the continent. At present, 37 countries have signed the AMA treaty, including 26 countries that have ratified it [146]. The main objective of the AMA is to enhance the capacity of States Parties and regional economic communities to regulate medical products to improve the quality, safety, and efficacy of medical products on the continent [147]. The AMA, in collaboration with other existing capacity building initiatives or organizations, such as the WHO Global Initiative on Digital Health, ITU, AMRH, WHO AFRO, and United Nations Children's Fund, can assist sub-Saharan African countries in aligning their regulatory requirements with available resources and support them to acquire the necessary tools and skills to build effective and sustainable regulatory systems for health apps. This can be achieved by adopting a decentralized approach to engage a network of technical experts across the African Union similar to the model of the European Medicines Agency [148].

Actors or Stakeholders

The regulation of health apps often requires working with a wide range of actors or stakeholders. However, in this review, we identified only 2 main actor groups (those who provide digital health services and those who use the ICT infrastructure of the health ministry). These are the groups that are targeted by the included regulatory standards.

From a broader perspective, 12 categories of stakeholders according to their potential role in regulating health apps for the self-management were mapped in this study. The potential contribution of these stakeholders to the regulation of health apps for self-management in sub-Saharan Africa not only depends on their roles and responsibilities but also on their interests, needs, expectations, and influence [41,149-151]. Thus, a robust stakeholder analysis is paramount as it can help define the scope of the regulatory process, prioritize the requirements, manage the expectations, and ensure the engagement and participation of stakeholders throughout the regulatory process [41,152-156]. Our stakeholder mapping, as presented in Table 2 (refer to Multimedia Appendix 4 for more details), lays the foundation for national governments to conduct a robust stakeholder analysis and to adopt an all-inclusive stakeholder engagement strategy to manage and sustain the engagement and participation of all relevant stakeholders [157,158].

XSL•FO
RenderX

Recommendations

Our review found that the regulation of health apps in sub-Saharan Africa is especially poor and almost nonexistent, as only Kenya has national standards that could address some of the regulatory issues related to health apps. Therefore, we recommend the following actions to help sub-Saharan African countries improve the regulation of health apps to support self-management:

- Establish a clear and consistent definition of what constitutes a health app (considering AI or machine learning) and what level of regulation is required for different types of apps.
- Develop and implement criteria and guidelines that ensure the quality, safety, and usability of health apps.
- Engage with independent app evaluators, such as ORCHA, to adopt a common framework that can guide the evaluation and accreditation of health apps and use the framework to create and maintain decentralized and transparent platforms that showcase and evaluate health apps for users and health care professionals.
- Develop and implement policies and regulations that enable sustainable funding for health apps such as integrating the use of health apps for self-management into existing health systems and financing pathways or mechanisms.
- Support and facilitate innovation and collaboration across the sub-Saharan Africa region, especially in areas including but not limited to data security and privacy, interoperability standards, usability, accessibility, funding, capacity building, and monitoring and evaluation of the regulatory process.
- Manage and sustain the engagement, involvement, and participation of all relevant stakeholders in the regulatory process by conducting a robust stakeholder analysis and adopting an all-inclusive stakeholder engagement strategy.

Strengths and Limitations of the Study

This study has several strengths, which include an extensive search of gray literature and repositories, contact with key individuals, and the use of a systematic approach. Given that regulatory standards and guidance are unavailable in scientific databases, a wide range of gray literature and repositories were searched. In addition, contact was made with key staff members to obtain relevant documents, including those at the MOHs, the WHO country offices, and the WHO AFRO. Second, to enhance the strength of the study, a policy analysis framework was adapted and used to systematically organize the key study findings, whereas a deductive descriptive qualitative content analysis approach was used to identify and analyze texts that contained relevant concepts and other related information based on the 4 predefined themes. Third, the RISA tool was used to guide the mapping of key stakeholders. This has further increased the robustness of the study findings.

The limitations of this study include the fact that our literature search was conducted in English. Although the literature search was conducted in English, it yielded documents written in French and Portuguese from the ICTworks repository. Second, regulatory standards and guidance are not readily available on scientific databases; hence, it is possible that some relevant documents might have been missed. However, efforts were made to obtain these documents by contacting key stakeholders including key contact persons at the WHO AFRO, WHO country offices, and MOHs. In addition, contacting key individuals only for the purposes of requesting documents rather than conducting direct interviews was one of the limitations of this study. Interviewing key contact persons and stakeholders to obtain additional information could have strengthened the review; however, we did not interview any key individuals or stakeholders because it was beyond the scope of this review. Nonetheless, we recommend that future studies consider incorporating interviews to explore the perspectives of key stakeholders.

Conclusions

Health apps are increasingly being used by patients to manage their health, and sub-Saharan African countries can leverage these apps to advance their progress toward achieving SDG 3 (good health and well-being) and UHC, especially given the rapid advancement of AI and big data. However, our study has established that the regulation of health apps in sub-Saharan Africa is inadequate to ensure that health apps are technically reliable and clinically safe; interoperable across systems; compliant with the principles of confidentiality of information and data privacy; culturally appropriate and relevant; and accessible to everyone regardless of gender, ethnicity, location, or income. Therefore, the region can learn from the experiences of some high-income countries such as the United Kingdom and Germany to develop and implement a robust and responsive regulatory system that supports the widespread adoption of safe, effective, and beneficial health apps for its population.

Following the publication of this review, a summary of the findings will be disseminated to the relevant organizations. In addition, the key findings will be summarized and presented at national, regional, and international conferences.

Acknowledgments

The authors would like to thank Rebecca Jones, the Library Manager and Liaison Librarian at Charing Cross Library, who advised and assisted with the search strategy for this study. This work is part of the PhD research of BAB, which is sponsored by the government of Nigeria. AM and JC were supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Northwest London (NIHR200180). The views expressed in this publication are those of the authors and not necessarily those of the government of Nigeria or the NIHR or the Department of Health and Social Care. In the Results and Discussion sections, Microsoft Copilot in Bing [159] was used to help summarize and modify a few texts as well as suggest some citations.

XSL•FO
RenderX

Data Availability

The search strategy for PubMed, Scopus, and the World Health Organization AIM is presented in Multimedia Appendix 1. All data generated or analyzed during this study are included in this published article (and its supplementary information files). The documents analyzed are available directly from the relevant institutional websites, ICTworks repository [44] or upon request from the relevant government departments in each country. Additionally, documents in the list of references that are not accessible on the web can be solicited from the corresponding author on reasonable request.

Authors' Contributions

BAB and JC conceived the study. BAB designed the study with contributions from JC and NM. BAB drafted the manuscript, and JC, NM, AM, SI, KPF, BIH, and NU read and contributed to it. AM was the clinical lead, and JC acted as a guarantor for this study. The final manuscript was read and approved by all the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 108 KB - jmir_v26i1e49163_app1.docx]

Multimedia Appendix 2
Database search strategy.
[DOCX File , 30 KB - jmir_v26i1e49163_app2.docx]

Multimedia Appendix 3
Details of included documents.
[DOCX File , 21 KB - jmir_v26i1e49163_app3.docx]

Multimedia Appendix 4
Mapping of the stakeholders according to their potential role in regulating health apps for self-management.
[DOCX File , 22 KB - jmir_v26i1e49163_app4.docx]

References

1. Aitken M, Nass D. Digital health trends 2021: innovation, evidence, regulation, and adoption. IQVIA Institute for Human Data Science. 2021. URL: https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/digital-health-trends-2021/iqvia-institute-digital-health-trends-2021.pdf?&_=1669449368070 [accessed 2022-11-26]
2. Mobile app threat landscape report. RiskIQ. 2020. URL: https://www.riskiq.com/2020-mobile-threat-landscape-report-thank-you/ [accessed 2021-07-19]
3. El-Sappagh S, Ali F, Hendawi A, Jang JH, Kwak KS. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97 [FREE Full text] [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
4. Labrique AB, Vasudevan L, Kochi E, Fabricant R, Mehl G. mHealth innovations as health system strengthening tools: 12 common applications and a visual framework. Glob Health Sci Pract 2013 Aug 06;1(2):160-171 [FREE Full text] [doi: 10.9745/ghsp-d-13-00031]
5. Adepoju IOO, Albersen BJA, De Brouwere V, van Roosmalen J, Zweekhorst M. mHealth for clinical decision-making in sub-Saharan Africa: a scoping review. JMIR Mhealth Uhealth 2017 Mar 23;5(3):e38 [FREE Full text] [doi: 10.2196/mhealth.7185] [Medline: 28336504]
6. Vegesna A, Tran M, Angelaccio M, Arcona S. Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed J E Health 2017 Jan;23(1):3-17 [FREE Full text] [doi: 10.1089/tmj.2016.0051] [Medline: 27116181]
7. Use of appropriate digital technologies for public health: Report by the Director-General. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/274134 [accessed 2023-05-06]
8. El-Osta A, Rowe C, Majeed A. Developing a shared definition of self-driven healthcare to enhance the current healthcare delivery paradigm. J R Soc Med 2022 Nov;115(11):424-428 [FREE Full text] [doi: 10.1177/01410768221136245] [Medline: 36455623]
9. Hussein R. A review of realizing the universal health coverage (UHC) goals by 2030: part 2- what is the role of eHealth and technology? J Med Syst 2015 Jul;39(7):72. [doi: 10.1007/s10916-015-0255-x] [Medline: 26044851]
10. Sustainable development goal 3: Ensure healthy lives and promote well-being for all at all ages. United Nations. URL: https://sdgs.un.org/goals/goal3 [accessed 2023-05-07]

XSL•FO
RenderX

11. Coronavirus: apps to help the elderly. Organisation for the Review of Care and Health Apps. 2020. URL: https://orchahealth.com/coronavirus-apps-to-help-the-elderly/ [accessed 2021-07-19]
12. Huckvale K, Adomaviciute S, Prieto JT, Leow MKS, Car J. Smartphone apps for calculating insulin dose: a systematic assessment. BMC Med 2015 May 06;13:106 [FREE Full text] [doi: 10.1186/s12916-015-0314-7] [Medline: 25943590]
13. Huckvale K, Car M, Morrison C, Car J. Apps for asthma self-management: a systematic assessment of content and tools. BMC Med 2012 Nov 22;10:144 [FREE Full text] [doi: 10.1186/1741-7015-10-144] [Medline: 23171675]
14. Huckvale K, Morrison C, Ouyang J, Ghaghda A, Car J. The evolution of mobile apps for asthma: an updated systematic assessment of content and tools. BMC Med 2015 Mar 23;13:58 [FREE Full text] [doi: 10.1186/s12916-015-0303-x] [Medline: 25857569]
15. Huckvale K, Prieto JT, Tilney M, Benghozi PJ, Car J. Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment. BMC Med 2015 Sep 07;13:214 [FREE Full text] [doi: 10.1186/s12916-015-0444-y] [Medline: 26404673]
16. Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw Open 2019 Apr 05;2(4):e192542 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2542] [Medline: 31002321]
17. Ndlovu K, Mars M, Scott RE. Interoperability frameworks linking mHealth applications to electronic record systems. BMC Health Serv Res 2021 May 13;21(1):459 [FREE Full text] [doi: 10.1186/s12913-021-06473-6] [Medline: 33985495]
18. Kenya standards and guidelines for mHealth systems. Kenya Ministry of Health. 2017. URL: https://www.health.go.ke/wp-content/uploads/2020/02/Revised-Guidelines-For-Mhealth-Systems-May-Version.pdf [accessed 2023-03-21]
19. Standard for electronic health record system (EHRs) in Ethiopia. Ethiopia Minister of Health. 2021. URL: https://registry.betterehealth.eu/ehealth-policies/standard-electronic-health-record-system-ehrs-ethiopia [accessed 2023-04-21]
20. National health normative standards framework for digital health interoperability in South Africa. South Africa Department of Health. 2021. URL: https://www.health.gov.za/wp-content/uploads/2022/10/HNSF_Gazette_21_October_2022.pdf [accessed 2023-05-15]
21. Ferretti A, Ronchi E, Vayena E. From principles to practice: benchmarking government guidance on health apps. Lancet Digit Health 2019 Jun;1(2):e55-e57 [FREE Full text] [doi: 10.1016/S2589-7500(19)30027-5] [Medline: 33323230]
22. Diao JA, Venkatesh KP, Raza MM, Kvedar JC. Multinational landscape of health app policy: toward regulatory consensus on digital health. NPJ Digit Med 2022 May 11;5(1):61 [FREE Full text] [doi: 10.1038/s41746-022-00604-x] [Medline: 35545663]
23. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):31 [FREE Full text] [doi: 10.1038/s41746-022-00573-1] [Medline: 35304561]
24. Brown SA, Garcia AA, Kouzekanani K, Hanis CL. Culturally competent diabetes self-management education for Mexican Americans: the Starr County border health initiative. Diabetes Care 2002 Feb;25(2):259-268 [FREE Full text] [doi: 10.2337/diacare.25.2.259] [Medline: 11815493]
25. Chaney SC, Mechael P. Self-Care Trailblazer Group. 2020. URL: https://media.psi.org/wp-content/uploads/2020/09/31000510/Digital-Self-Care-Final.pdf [accessed 2021-05-20]
26. Kanzaveli T. Healthcare: shiftingfrom “one size fits all” to “one size fits one”. Medium. 2017. URL: https://tkanzaveli.medium.com/healthcare-shifting-from-one-size-fits-all-to-one-size-fits-one-d56136ded705 [accessed 2022-03-04]
27. Myth 1 – one app will fit all!. Organisation for the Review of Care and Health Apps. URL: https://orchahealth.com/myth-1-one-app-will-fit-all/ [accessed 2022-03-04]
28. Aitken M, Lyle J. Patient adoption of mHealth: use, evidence and remaining barriers to mainstream acceptance. IQVIA Institute for Human Data Science. 2015 Sep. URL: https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/patient-adoption-of-mhealth.pdf [accessed 2021-05-21]
29. Mechael P, Batavia H, Kaonga N. Barriers and gaps affecting mhealth in low and middle income countries: policy white paper. Center for Global Health and Economic Development Earth Institute, Columbia University. 2010. URL: http://www.globalproblems-globalsolutions-files.org/pdfs/mHealth_Barriers_White_Paper.pdf [accessed 2021-03-24]
30. Bene BA, Ibeneme S, Fadahunsi KP, Harri BI, Ukor N, Mastellos N, et al. Regulatory standards and guidance for the use of health applications for self-management in Africa: scoping review protocol. BMJ Open 2022 Feb 11;12(2):e058067 [FREE Full text] [doi: 10.1136/bmjopen-2021-058067] [Medline: 35149576]
31. Aitken M, Gauntlett C. Patient apps for improved healthcare: from novelty to mainstream. IMS Institute for Healthcare Informatics. 2013. URL: https://ignacioriesgo.es/wp-content/uploads/2014/03/iihi_patient_apps_report_editora_39_2_1.pdf [accessed 2024-03-10]
32. National eHealth Strategy Toolkit. World Health Organization, International Telecommunication Union. 2012. URL: https://www.itu.int/pub/D-STR-E_HEALTH.05-2012 [accessed 2021-06-28]
33. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2021-06-23]
34. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]

XSL•FO
RenderX

35. Anderson S, Allen P, Peckham S, Goodwin N. Asking the right questions: scoping studies in the commissioning of research on the organisation and delivery of health services. Health Res Policy Syst 2008 Jul 09;6:7 [FREE Full text] [doi: 10.1186/1478-4505-6-7] [Medline: 18613961]
36. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
37. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126 [FREE Full text] [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
38. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
39. Leitner C, Potenziani D. Mendeley reference manager. Mendeley. 2022. URL: https://www.mendeley.com/reference-management/reference-manager [accessed 2022-08-03]
40. Better systematic review management. Covidence. URL: https://www.covidence.org/ [accessed 2023-02-13]
41. Franco-Trigo L, Fernandez-Llimos F, Martínez-Martínez F, Benrimoj SI, Sabater-Hernández D. Stakeholder analysis in health innovation planning processes: A systematic scoping review. Health Policy 2020 Oct;124(10):1083-1099 [FREE Full text] [doi: 10.1016/j.healthpol.2020.06.012] [Medline: 32829927]
42. Walt G, Gilson L. Reforming the health sector in developing countries: the central role of policy analysis. Health Policy Plan 1994 Dec;9(4):353-370 [FREE Full text] [doi: 10.1093/heapol/9.4.353] [Medline: 10139469]
43. Digital health: a call for government leadership and cooperation between ICT and health. Broadband Commission. 2017. URL: https://broadbandcommission.org/wp-content/uploads/2021/09/WGHealth_Report2017-.pdf [accessed 2021-06-28]
44. Vota W. Every African country's national eHealth strategy or digital health policy. ICT works. 2019. URL: https://www.ictworks.org/african-national-ehealth-strategy-policy/ [accessed 2023-12-10]
45. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
46. Standards and guidelines for electronic medical record systems in Kenya. Kenya Ministry of Medical Services, Kenya Ministry of Public Health and Sanitation. 2010. URL: http://guidelines.health.go.ke:8000/media/Standards_and_Guidelines_for_EMR_Systems.pdf [accessed 2023-04-21]
47. Kenya standards and guidelines for E-health systems interoperability. Kenya Ministry of Health, AfyaInfo Project. 2014. URL: https://pdf.usaid.gov/pdf_docs/PA00TB2K.pdf [accessed 2023-03-21]
48. Health sector ICT standards and guidelines. Kenya Ministry of Health. 2013. URL: https://www.medbox.org/pdf/5e148832db60a2044c2d2895 [accessed 2023-03-21]
49. Health information exchange standard operating procedure (SOP) and guideline. Nigeria Federal Ministry of Health. 2020 Jul.
50. National eHealth strategy 2018-2022. Benin Ministry of Health. 2017.
51. The eHealth strategy of botswana 2020-2024. Botswana Ministry of Health. URL: https://ehealth.ub.bw/bhdc/Docs/MOH%20ehealth%20Strategy%20Book%20A4.pdf [accessed 2023-04-22]
52. Health sector digital strategy 2016-2020. Burkina Faso Ministry of Health.
53. National health informatics development plan of Burundi. Burundi Ministry of Public Health. 2015.
54. The 2020-2024 national digital health strategic plan. Cameroon Ministry of Public Health. 2020.
55. National eHealth strategy 2017-2021. Comoros Ministry of Health. 2016.
56. eHealth strategic plan. Cote d'Ivoire Minister of Health and Public Hygiene. 2011.
57. National development plan for health informatics. Democratic Republic of Congo Ministry of Public Health. 2014.
58. Kingdom of Swaziland eHealth strategy 2016 - 2020. Kingdom of Swaziland Ministry Of Health. 2016.
59. Information revolution strategic plan (2018-2025). Ethiopia Ministry of Health. 2018.
60. Strategic master plan of the health information system of the Gabon. Gabon Ministry of Public Health and Population. 2017.
61. National e-Health strategy. Ghana Ministry of Health. 2010.
62. Kenya national e-Health strategy. Kenya Ministry of Medical Services, Kenya Ministry of Public Health & Sanitation. 2011.
63. Kenya national eHealth policy 2016-2030. Kenya Ministry of Health. 2016.
64. National strategy - Liberia - 2016-2021. Liberia Ministry of Health. 2016.
65. Strategic plan for strengthening the health information system of Madagascar 2018–2022. Madagascar Ministry of Public Health. 2017.
66. National digital health strategy 2020-2025. Malawi Ministry of Health. 2020.
67. National eHealth policy in Mali. Mali Ministry of Health and Public Hygiene. 2013.
68. Health 2015: seamless continuity of care. Mauritius Ministry of Health and Quality of Life. 2015.
69. Strategic plan of information system for health 2009-2014. Mozambique Ministry of Health. 2009.

XSL•FO
RenderX

70. National eHealth strategy 2019-2023. Niger Ministry of Public Health. 2018.
71. National digital health strategy 2021-2025. Nigeria Federal Ministry of Health. 2021.
72. National digital health policy. Nigeria Federal Ministry of Health. 2021.
73. National digital health strategic plan 2018-2023. Rwanda Ministry of Health. 2018. URL: https://elearning.helinanet.org/mod/resource/view.php?id=890 [accessed 2023-05-09]
74. Strategic plan for digital health 2018-2023. Senegal Ministry of Health and Social Action. 2018.
75. National digital health strategy 2018-2023. Sierra Leone Ministry of Health and Sanitation, Sierra Leone Ministry of Information and Communication. 2018.
76. The national digital health strategy 2019 – 2024. Tanzania Ministry of Health, Community Development, Gender, Elderly and Children. 2019.
77. National digital health strategy for South Africa 2019 - 2024. South Africa Department of Health. 2019.
78. Strategic plan for the development of eHealth in Togo 2013-2015. Togo Ministry of Health. 2012.
79. Uganda national eHealth policy. Uganda Ministry of Health. 2016.
80. Uganda national eHealth strategy 2017 - 2021. Uganda Ministry of Health. URL: https://health.go.ug/sites/default/files/National%20e_Health%20Strategy_0.pdf [accessed 2023-05-16]
81. National eHealth strategy 2017-2021. Zambia Ministry of Health. 2017.
82. Zimbabwe's E-Health strategy 2012-2017. Ministry of Health and Child Welfare. 2012.
83. National eHealth strategy 2021-2025. Namibia Ministry of Health & Social Services. 2021. URL: https://www.scribd.com/document/639371316/eHealth-Strategy-Namibia-2021# [accessed 2023-05-13]
84. Health sector ICT policy and strategy. Ghana Ministry of Health. 2005. URL: https://www.moh.gov.gh/wp-content/uploads/2016/02/Health-Sector-ICT-Policy-and-Strategy.pdf [accessed 2023-05-08]
85. Adebola OJ. Beyond national digital health strategy: final report of end term evaluation for the National Health ICT Strategic Framework 2015-2020. Nigeria Federal Ministry of Health. 2021 May.
86. National Health ICT Strategic Framework 2015 - 2020. Nigeria Federal Ministry of Health. 2016. URL: https://www.health.gov.ng/doc/HealthICTStrategicFramework.pdf [accessed 2023-05-16]
87. Digital health blueprint. Ethiopia Ministry of Health. 2021. URL: http://repository.iifphc.org/bitstream/handle/123456789/1658/Ethiopian-Digital-Health-Blueprint.pdf?sequence=1&isAllowed=y [accessed 2023-05-16]
88. Kenya health information systems interoperability framework. Kenya Ministry of Health. 2020. URL: https://www.data4sdgs.org/sites/default/files/services_files/Kenya%20Health%20Information%20Systems%20Interoperability%20Framework.pdf [accessed 2023-05-16]
89. National community health digitization strategy 2020-2025. Kenya Ministry of Health, Division of Community Health Services. 2021. URL: https://www.eahealth.org/sites/www.eahealth.org/files/content/attachments/2021-08-02/eCHIS-Strategy-2020-2025.pdf [accessed 2023-05-16]
90. Leitner C, Potenziani D. Health information systems interoperability in Liberia. IntraHealth International. 2016. URL: https://elearning.helinanet.org/mod/resource/view.php?id=938 [accessed 2023-05-16]
91. Narrative for 2022 national digital health annual operational plan (AOP). Nigeria Federal Ministry of Health. 2022.
92. Tanzania digital health investment road map 2017-2023. Tanzania Ministry of Health, Community Development, Gender, Elderly and Children, President's Office Regional Administration and Local Government. 2017.
93. Yu H. Regulation of digital health technologies in the European Union: intended versus actual use. In: Cohen GI, Minssen T, Price II NW, Robertson C, Shachar C, editors. The Future of Medical Device Regulation: Innovation and Protection. Cambridge: Cambridge University Press; Mar 31, 2022:103-114.
94. Policy for device software functions and mobile medical applications: guidance for industry and Food and Drug Administration staff. U.S. Food and Drug Administration. 2022. URL: https://www.fda.gov/media/80958/download [accessed 2023-10-10]
95. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14 [FREE Full text] [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
96. FDA clears mobile medical app to help those with opioid use disorder stay in recovery programs. U.S. Food and Drug Administration. 2018. URL: https://www.fda.gov/news-events/press-announcements/fda-clears-mobile-medical-app-help-those-opioid-use-disorder-stay-recovery-programs [accessed 2021-01-27]
97. Digital maturity model: achieving digital maturity to drive growth. Deloitte. 2018. URL: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Technology-Media-Telecommunications/deloitte-digital-maturity-model.pdf [accessed 2021-10-20]
98. May E. How digital health apps are empowering patients. Deloitte. 2021. URL: https://www2.deloitte.com/us/en/blog/health-care-blog/2021/how-digital-health-apps-are-empowering-patients.html [accessed 2023-10-06]
99. Guidance: criteria for health app assessment. Public Health England. 2017. URL: https://www.gov.uk/government/publications/health-app-assessment-criteria/criteria-for-health-app-assessment [accessed 2023-10-16]
100. Clinical risk management standards. National Health Service Digital. 2020. URL: https://digital.nhs.uk/services/clinical-safety/clinical-risk-management-standards [accessed 2023-10-28]

XSL•FO
RenderX

101. Report a problem with a medicine or medical device. Gov.uk. URL: https://www.gov.uk/report-problem-medicine-medical-device [accessed 2023-11-07]
102. Digital technology assessment criteria for health and social care (DTAC) - Version 1.0. National Health Service X. 2021. URL: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Ftransform.england.nhs.uk%2Fmedia%2Fdocuments%2FDTAC_version_1.0_FINAL_updated_16.04.odt&wdOrigin=BROWSELINK [accessed 2023-11-07]
103. Data protection impact assessment: NHS login - formerly Citizen Identity. National Health Service Digital. 2022. URL: https://digital.nhs.uk/services/nhs-login/data-protection-impact-assessment [accessed 2023-11-07]
104. Risks and data protection impact assessments (DPIAs). Information Commissioner's Office. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/accountability-framework/risks-and-data-protection-impact-assessments-dpias/ [accessed 2023-11-07]
105. A guide to good practice for digital and data-driven health technologies. Department of Health and Social Care. 2021. URL: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology [accessed 2023-10-30]
106. Interoperability in healthcare. Healthcare Information and Management Systems Society (HIMSS). 2023. URL: https://www.himss.org/resources/interoperability-healthcare [accessed 2023-10-17]
107. DAPB4020: UK core Fast Healthcare Interoperability Resources (FHIR) release 4 (R4) governance. National Health Service Digital. 2022. URL: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dapb4020-uk-core-fhir-r4-governance [accessed 2023-10-17]
108. Fast Healthcare Interoperability Resources (FHIR). National Health Service Digital. 2022. URL: https://digital.nhs.uk/services/fhir-apis [accessed 2023-10-17]
109. FHIR Interoperability Basics: 4 things to know. Health IT Analytics. 2022. URL: https://healthitanalytics.com/news/4-basics-to-know-about-the-role-of-fhir-in-interoperability [accessed 2023-11-07]
110. Giordanengo A, Bradway M, Pedersen R, Grøttland A, Hartvigsen G, Årsand E. Integrating data from apps, wearables and personal electronic health record (pEHR) systems with clinicians' electronic health records (EHR) systems. Int J Integr Care 2016 Nov 09;16(5):16 [FREE Full text] [doi: 10.5334/IJIC.2565]
111. A plan for digital health and social care. Department of Health & Social Care. 2022. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2022-12-01]
112. Ryu B, Kim N, Heo E, Yoo S, Lee K, Hwang H, et al. Impact of an electronic health record-integrated personal health record on patient participation in health care: development and randomized controlled trial of MyHealthKeeper. J Med Internet Res 2017 Dec 07;19(12):e401 [FREE Full text] [doi: 10.2196/jmir.8867] [Medline: 29217503]
113. Winter A, Takabayashi K, Jahn F, Kimura E, Engelbrecht R, Haux R, et al. Quality requirements for electronic health record systems*. A Japanese-German information management perspective. Methods Inf Med 2017 Aug 07;56(7):e92-e104 [FREE Full text] [doi: 10.3414/ME17-05-0002] [Medline: 28925415]
114. Wachter RM. Making IT work: harnessing the power of health information technology to improve care in England. National Advisory Group on Health Information Technology. 2016. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/550866/Wachter_Review_Accessible.pdf [accessed 2021-07-22]
115. Framework on integrated people-centred health services (IPCHS). World Health Organisation. 2023. URL: https://www.who.int/teams/integrated-health-services/clinical-services-and-systems/service-organizations-and-integration [accessed 2023-06-05]
116. Ibeneme S, Karamagi H, Muneene D, Goswami K, Chisaka N, Okeibunor J. Strengthening health systems using innovative digital health technologies in Africa. Front Digit Health 2022;4:854339 [FREE Full text] [doi: 10.3389/fdgth.2022.854339] [Medline: 35434700]
117. Ibeneme S, Ukor N, Ongom M, Dasa T, Muneene D, Okeibunor J. Strengthening capacities among digital health leaders for the development and implementation of national digital health programs in Nigeria. BMC Proc 2020;14(Suppl 10):9 [FREE Full text] [doi: 10.1186/s12919-020-00193-1] [Medline: 32714444]
118. Delivering safe digital health. Organisation for the Review of Care and Health Apps. URL: https://orchahealth.com/ [accessed 2023-10-22]
119. Mamuye AL, Yilma TM, Abdulwahab A, Broomhead S, Zondo P, Kyeng M, et al. Health information exchange policy and standards for digital health systems in Africa: a systematic review. PLOS Digit Health 2022 Oct;1(10):e0000118 [FREE Full text] [doi: 10.1371/journal.pdig.0000118] [Medline: 36812615]
120. Digital health platform handbook: Building a digital information infrastructure (infostructure) for health. World Health Organization, International Telecommunication Union. 2022. URL: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-E_HEALTH.10-2020-PDF-E.pdf [accessed 2021-05-22]
121. Framework for involving patients in patient safety 2021. National Health Service England. 2021. URL: https://www.england.nhs.uk/patient-safety/framework-for-involving-patients-in-patient-safety/ [accessed 2023-03-23]

XSL•FO
RenderX

122. Olesch A. Towards harmonised EU landscape for digital health: summary of the roundtable discussions in selected EIT Health InnoStars countries. EIT Health InnoStars. 2023 Jan. URL: https://eithealth.eu/wp-content/uploads/2023/02/EIT_Health_DiGA_report_Jan2023.pdf [accessed 2023-10-10]
123. Grieb J, Tschammler D, Färber C, Woitz S. Digital health laws and regulations germany. Global Legal Group. 2023. URL: https://iclg.com/practice-areas/digital-health-laws-and-regulations/germany [accessed 2023-11-03]
124. Human regulatory: medical devices. European Medicines Agency. URL: https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices [accessed 2023-10-12]
125. Strengthening the capacity for regulation of medical products in the African region. World Health Organization Regional Office for Africa. 2013. URL: https://iris.who.int/bitstream/handle/10665/94308/AFR_RC63_7.pdf?sequence=1 [accessed 2023-10-17]
126. Ncube BM, Dube A, Ward K. Establishment of the African Medicines Agency: progress, challenges and regulatory readiness. J Pharm Policy Pract 2021 Mar 08;14(1):29 [FREE Full text] [doi: 10.1186/s40545-020-00281-9] [Medline: 33685518]
127. Post market surveillance system. European Union Medical Device Regulation. 2023. URL: https://eumdr.com/post-market-surveillance-system/ [accessed 2023-10-31]
128. Dayal R. Effective post-market surveillance for medical devices: An essential part of medical devices regulation (MDR). Capgemini. 2020. URL: https://www.capgemini.com/insights/expert-perspectives/effective-post-market-surveillance-for-medical-devices-an-essential-part-of-mdr/ [accessed 2023-10-31]
129. NHS app library reaches 70 apps in honour of the NHS birthday. Northampton General Hospital NHS Trust. 2018. URL: https://www.northamptongeneral.nhs.uk/News/News-Archive/2018/NHS-App-Library-reaches-70-apps-in-honour-of-the-NHS-birthday.aspx [accessed 2023-09-21]
130. Developers invited to add to NHS apps library. National Health Service Digital. 2018. URL: https://digital.nhs.uk/news/2018/developers-invited-to-add-to-nhs-apps-library [accessed 2023-09-22]
131. The NHS apps library has closed. National Health Service Digital. 2021. URL: https://digital.nhs.uk/services/nhs-apps-library#:~:text=The%20NHS%20Apps%20Library%20was%20decommissioned%20in%20December%202021.&text=Further%20information%20can%20be%20found%20on%20the%20NHS.UK%20website [accessed 2023-09-22]
132. Larsen ME, Huckvale K, Nicholas J, Torous J, Birrell L, Li E, et al. Using science to sell apps: evaluation of mental health app store quality claims. NPJ Digit Med 2019;2:18 [FREE Full text] [doi: 10.1038/s41746-019-0093-1] [Medline: 31304366]
133. About the NHS app. National Health Service. 2023 Dec 4. URL: https://www.nhs.uk/nhs-app/about-the-nhs-app/ [accessed 2023-09-22]
134. Scope of registration: regulated activities. Care Quality Commission. 2022. URL: https://www.cqc.org.uk/guidance-providers/scope-registration-regulated-activities [accessed 2023-11-05]
135. Distributing great apps into health and care services across the world. Organisation for the Review of Care and Health Apps. 2020. URL: https://orchahealth.com/wp-content/uploads/2020/12/Health-and-Care-1.pdf [accessed 2023-10-09]
136. Our founder, our story and our values: we exist to make digital health healthy. Organisation for the Review of Care and Health Apps. URL: https://orchahealth.com/about-us/ [accessed 2023-10-09]
137. Health app library: empower your community with safe access to health apps and digital health products. Organisation for the Review of Care and Health Apps. URL: https://orchahealth.com/our-products/health-app-library/#:~:text=A%20Health%20App%20Library%20is,on%20the%20Health%20App%20Library [accessed 2023-10-09]
138. Digital technology assessment criteria (DTAC). National Health Service X. URL: https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/ [accessed 2023-10-09]
139. Evidence standards framework (ESF) for digital health technologies. National Institute for Health and Care Excellence. 2023. URL: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [accessed 2023-10-08]
140. Tsang L, Kerr-Peterson H. UK NICE updates its evidence standards framework for data-driven digital health technologies. Ropes & Gray. 2022. URL: https://www.ropesgray.com/en/insights/alerts/2022/10/uk-nice-updates-its-evidence-standards-framework-for-data-driven-digital-health-technologies [accessed 2023-10-09]
141. Guidance: medical device stand-alone software including apps (including IVDMDs). Medicines and healthcare products regulatory agency. 2023. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1168485/Medical_device_stand-alone_software_including_apps__including_IVDMDs_.pdf [accessed 2023-10-09]
142. Digital therapeutics in the United Kingdom. Digital Therapeutics Alliance. 2021. URL: https://dtxalliance.org/wp-content/uploads/2021/06/DTA_DTx-Overview_UK.pdf [accessed 2023-10-09]
143. Transforming global healthcare by advancing digital therapeutics. Digital Therapeutics Alliance. 2023. URL: https://dtxalliance.org/ [accessed 2023-10-10]
144. International Organization for Standardization (ISO) digital therapeutic definition. Digital Therapeutic Alliance. 2023 Jun. URL: https://dtxalliance.org/wp-content/uploads/2023/06/DTA_FS_ISO-Definition.pdf [accessed 2023-10-09]
145. Ndomondo-Sigonda M, Miot J, Naidoo S, Dodoo A, Kaale E. Medicines regulation in Africa: current state and opportunities. Pharmaceut Med 2017;31(6):383-397 [FREE Full text] [doi: 10.1007/s40290-017-0210-x] [Medline: 29200865]
146. Chinele J. East Africa shows solid support for African Medicines Agency treaty. Health Policy Watch. 2023 Aug 16. URL: https://healthpolicy-watch.news/east-africa-shows-solid-support-for-african-medicines-agency-treaty/ [accessed 2023-10-09]

XSL•FO
RenderX

147. Treaty for the establishment of the African Medicines Agency 2019. African Union. 2019. URL: https://au.int/sites/default/files/treaties/36892-treaty-0069_-_ama_treaty_e.pdf [accessed 2023-10-17]
148. European Medicines Agency: about us. European Medicines Agency. 2023 Mar 1. URL: https://www.ema.europa.eu/en/documents/other/about-us-european-medicines-agency-ema_en.pdf [accessed 2023-10-18]
149. Bryson JM. What to do when stakeholders matter. Public Adm Rev 2004 Mar;6(1):21-53 [FREE Full text] [doi: 10.1080/14719030410001675722]
150. Iyawa G, Herselman M, Botha A. Potential stakeholders and perceived benefits of a digital health innovation ecosystem for the Namibian context. Procedia Computer Science 2017;121:431-438. [doi: 10.1016/j.procs.2017.11.058]
151. Ferretti V. From stakeholders analysis to cognitive mapping and multi-attribute value theory: an integrated approach for policy support. European Journal of Operational Research 2016 Sep;253(2):524-541. [doi: 10.1016/j.ejor.2016.02.054]
152. Brugha R, Varvasovszky Z. Stakeholder analysis: a review. Health Policy Plan 2000 Sep;15(3):239-246. [doi: 10.1093/heapol/15.3.239] [Medline: 11012397]
153. Schmeer K. Guidelines for conducting a stakeholder analysis 1999. Partnerships for Health Reform, Abt Associates. 1999. URL: https://www.ktecop.ca/wordpress/wp-content/uploads/guidelines-stakeholder-analysis-PHR-1999.pdf [accessed 2023-10-17]
154. Gilmour J, Beilin R. Stakeholder mapping for effective risk assessment and communication. Australian Centre of Excellence for Risk Analysis, University of Melbourne. 2007 Apr. URL: https://cebra.unimelb.edu.au/__data/assets/pdf_file/0006/2220990/gilmour0609.pdf [accessed 2023-10-17]
155. Quality, service improvement and redesign tools: stakeholder analysis. National Health Service England, National Health Service Improvement. 2022. URL: https://www.england.nhs.uk/wp-content/uploads/2022/02/qsir-stakeholder-analysis.pdf [accessed 2023-10-20]
156. Craven MP, Lang AR, Martin JL. Developing mHealth apps with researchers: multi-stakeholder design considerations. : Springer; 2014 Presented at: Third International Conference, DUXU 2014, held as a part of HCI International; June 22-27, 2014; Heraklion, Greece p. 15-24 URL: https://doi.org/10.1007/978-3-319-07635-5_2 [doi: https://doi.org/10.1007/978-3-319-07635-5_2]
157. How to encourage stakeholder participation. SustaiNet Software International. URL: https://sustainet.com/how-to-encourage-stakeholder-participation/ [accessed 2023-10-20]
158. Stakeholder engagement. Organisation for Economic Cooperation and Development. URL: https://www.oecd.org/governance/better-international-rulemaking/compendium/keyprinciples/stakeholderengagement.htm [accessed 2023-10-20]
159. Microsoft Copilot in Bing. Microsoft. URL: https://www.bing.com/chat?form=NTPCHB [accessed 2023-03-15]

Abbreviations

AFRO: Regional Office for Africa
AI: artificial intelligence
AMA: African Medicines Agency
AMRH: African Medicines Regulatory Harmonization
CE: Conformité Européenne
CQC: Care Quality Commission
DHT: digital health technology
DTAC: Digital Technology Assessment Criteria
DTx: digital therapeutics
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
GDPR: General Data Protection Regulation
HL7: Health Level 7
ICT: information and communications technology
ITU: International Telecommunication Union
mHealth: mobile health
MHRA: Medicines and Healthcare products Regulatory Agency
MOH: Ministry of Health
NHS: National Health Service
NMRA: National Medicines Regulatory Authority
ORCHA: Organisation for the Review of Care and Health Apps
PMS: postmarket surveillance
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RISA: Reporting Items for Stakeholder Analysis

XSL•FO
RenderX

SDG: sustainable development goal
UHC: universal health coverage
WHO: World Health Organization

Edited by A Mavragani; submitted 19.05.23; peer-reviewed by N O'Brien, A Essén; comments to author 07.09.23; revised version received 08.12.23; accepted 23.02.24; published 11.04.24.

Please cite as:
Bene BA, Ibeneme S, Fadahunsi KP, Harri BI, Ukor N, Mastellos N, Majeed A, Car J
Regulatory Standards and Guidance for the Use of Health Apps for Self-Management in Sub-Saharan Africa: Scoping Review
J Med Internet Res 2024;26:e49163
URL: https://www.jmir.org/2024/1/e49163
doi:10.2196/49163
PMID:38602718

©Benard Ayaka Bene, Sunny Ibeneme, Kayode Philip Fadahunsi, Bala Isa Harri, Nkiruka Ukor, Nikolaos Mastellos, Azeem Majeed, Josip Car. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effectiveness of a Digital App for Reduction of Clinical Symptoms in Individuals With Panic Disorder: Randomized Controlled Trial

KunJung Kim[1], MD; Hyunchan Hwang[1], MD, PhD; Sujin Bae[1], PhD; Sun Mi Kim[1], MD, PhD; Doug Hyun Han[1], MD, PhD

Chung Ang University Hospital, Seoul, Republic of Korea

Corresponding Author:
Doug Hyun Han, MD, PhD
Chung Ang University Hospital
102 Heucsock ro
Seoul, 06973
Republic of Korea
Phone: 82 2 6299 3132
Fax: 82 2 6299 3100
Email: hduk70@gmail.com

Abstract

Background: Panic disorder is a common and important disease in clinical practice that decreases individual productivity and increases health care use. Treatments comprise medication and cognitive behavioral therapy. However, adverse medication effects and poor treatment compliance mean new therapeutic models are needed.

Objective: We hypothesized that digital therapy for panic disorder may improve panic disorder symptoms and that treatment response would be associated with brain activity changes assessed with functional near-infrared spectroscopy (fNIRS).

Methods: Individuals (n=50) with a history of panic attacks were recruited. Symptoms were assessed before and after the use of an app for panic disorder, which in this study was a smartphone-based app for treating the clinical symptoms of panic disorder, panic symptoms, depressive symptoms, and anxiety. The hemodynamics in the frontal cortex during the resting state were measured via fNIRS. The app had 4 parts: diary, education, quest, and serious games. The study trial was approved by the institutional review board of Chung-Ang University Hospital (1041078-202112-HR-349-01) and written informed consent was obtained from all participants.

Results: The number of participants with improved panic symptoms in the app use group (20/25, 80%) was greater than that in the control group (6/21, 29%; χ^2_1=12.3; P=.005). During treatment, the improvement in the Panic Disorder Severity Scale (PDSS) score in the app use group was greater than that in the control group ($F_{1,44}$=7.03; P=.01). In the app use group, the total PDSS score declined by 42.5% (mean score 14.3, SD 6.5 at baseline and mean score 7.2, SD 3.6 after the intervention), whereas the PDSS score declined by 14.6% in the control group (mean score 12.4, SD 5.2 at baseline and mean score 9.8, SD 7.9 after the intervention). There were no significant differences in accumulated oxygenated hemoglobin ($accHbO_2$) at baseline between the app use and control groups. During treatment, the reduction in $accHbO_2$ in the right ventrolateral prefrontal cortex (VLPFC; $F_{1,44}$=8.22; P=.006) and the right orbitofrontal cortex (OFC; $F_{1,44}$=8.88; P=.005) was greater in the app use than the control group.

Conclusions: Apps for panic disorder should effectively reduce symptoms and VLPFC and OFC brain activity in patients with panic disorder. The improvement of panic disorder symptoms was positively correlated with decreased VLPFC and OFC brain activity in the resting state.

Trial Registration: Clinical Research Information Service KCT0007280; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=21448

(J Med Internet Res 2024;26:e51428) doi:10.2196/51428

XSL•FO
RenderX

KEYWORDS

digital app; mHealth; mobile health; app; apps; application; applications; functional near-infrared spectroscopy; hemodynamic; hemodynamics; panic disorder; anxiety; panic; mental; fear; spectroscopy; digital therapy; fNIRS; brain; imaging; neurology; neuroscience; cortex; cortices

Introduction

Panic disorder is a common and important disease in clinical practice that leads to a reduction of individual productivity and increased use of health care [1]. The lifetime prevalence of panic disorder in the general population is 4.8%, and 22.7% of people experience panic attacks [2]. The most common symptoms of panic disorder include palpitations, shortness of breath, chest pain, numbness of the hands and feet, and cardiorespiratory-type symptoms, in addition to fear of dying, sweating, tremors, dizziness, nausea, and chills [3]. The US Food and Drug Administration has currently only approved selective serotonin reuptake inhibitors (SSRIs) for the treatment of panic disorder [4]. However, it is clinically difficult to expect an improvement in symptoms using SSRIs alone in the acute phase; thus we treat patients with benzodiazepine, which can lead to dependence and withdrawal symptoms [5,6]. The most common side effects of SSRIs reported by patients are reduced sexual function, drowsiness, and weight gain [7], and clinicians may hesitate to use benzodiazepines due to dependence and withdrawal symptoms [8]. Cognitive behavioral therapy (CBT) is the most widely used nonpharmaceutical treatment for anxiety disorders [9]. Additional nonpharmaceutical treatments, such as group therapy and supportive psychotherapy, are also available for patients with panic disorder [10,11]. However, these treatments have the disadvantage of requiring face-to-face contact; therefore, other therapeutic alternatives should be offered to patients during pandemics such as COVID-19.

The definition of a digital therapeutic (DTx) is a therapeutic that delivers evidence-based interventions to prevent, manage, or treat a medical disorder or disease; DTxs are currently used in many areas [12]. This kind of medical and public health use of smartphones and digital technologies is also known as mobile health (mHealth). DTxs related to mental health medicine are actively used in various psychiatric disorders, such as insomnia, substance abuse, attention-deficit/hyperactivity disorder, and anxiety and depression, among others [13]. In particular, the use of Freespira, a panic disorder DTx, reduced panic symptoms, avoidance behaviors, and treatment costs in patients with panic disorder [14].

As brain imaging technology advances, a great deal of functional mapping information on the human brain has been accumulated from positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS). Among these technologies, fNIRS can measure brain activity in a noninvasive and safe manner through measuring changes in the hemoglobin oxygenation state of the human brain [15]. Various studies have been conducted using fNIRS and fMRI to reveal correlations between panic disorder and brain regions. For example, patients with panic disorder show increased activity in the inferior frontal cortex, hippocampus, cingulate (both anterior and posterior), and orbitofrontal cortex (OFC) [16]. Previously, we confirmed that patients with panic disorder during rest periods showed increased activity in the OFC [17].

In this study, we determined whether an app for panic disorder would improve panic disorder symptoms. In addition, we used fNIRS to confirm the association between changes in panic disorder symptoms and changes in activity in specific brain regions.

Methods

Participants

Patients who had experiences of panic attacks were recruited between March 1 and July 30, 2022, through billboard advertisements at our hospital. The inclusion criteria for the study were as follows: (1) age between 20 and 65 years, (2) diagnosis of panic disorder based on the *Diagnostic and Statistical Manual of Mental Disorders*, *Fifth Edition* and (3) ability to use apps without problems. The exclusion criteria were as follows: (1) a history of other psychiatric disorders, except for anxiety disorder, or substance dependence, except for habitual alcohol and tobacco use; and (2) a history of head trauma and chronic medical conditions. The research clinician assessed whether patients fulfilled the inclusion or exclusion criteria. Written informed consent was acquired from all participants at the first visit. This study has been registered with the Clinical Research Information Service (KCT0007280).

Assessment Scales for Anxiety Symptoms

The severity of panic symptoms was assessed using the Panic Disorder Severity Scale (PDSS). The PDSS was developed by Shear et al [18] in 1997. It is a 7-item instrument used to rate the overall severity of panic disorder and was validated in Korea by Lim et al [19] in 2001.

The anxiety symptoms of all participants were assessed using the clinician-based Hamilton Anxiety Scale (HAM-A) questionnaire and the participant-based Generalized Anxiety Disorder-7 (GAD-7) questionnaire. The HAM-A was developed by Hamilton in 1969 [20]. The 14-item version remains the most used outcome measure in clinical trials of treatments for anxiety disorders and was validated in Korea by Kim [21] in 2001.

The GAD-7 questionnaire, developed by Spitzer et al [22], is a 7-item self-report anxiety questionnaire designed to assess the patient's health status during the previous 2 weeks. The GAD-7 was translated into the Korean language and is freely downloadable on the Patient Health Questionnaire website [23].

Hemodynamic Response of the Prefrontal Cortex

The hemodynamics in the frontal cortex during the resting state were measured using the fNIRS device (NIRSIT; OBELAB Inc). The NIRSIT has 24 laser diodes (sources) emitting light at 2 wavelengths (780 nm and 850 nm) and 32 photodetectors with a sampling rate of 8.138 Hz [24]. The distance between

XSL•FO
RenderX

the source and photodetector is 15 mm. Based on the suggested suitable sensor-detector separation distance for measuring cortical hemodynamic changes, only 30-mm channels were analyzed in this study [25].

For our study, we used the 48-channel configuration (Figure 1). The detected light signals in each wavelength were filtered with a band-pass filter (0.00 Hz-0.1 Hz) to reduce the effect of environmental noise–related light and body movements. In addition, channels with low-quality information (signal-to-noise ratio <30 dB) were removed from the hemodynamic analysis. The accumulated oxygenated hemoglobin ($accHbO_2$) values in the resting state represent the activation of the prefrontal cortex. In accordance with the theory that oxygenated hemoglobin has superior sensitivity and signal-to-noise ratio compared to deoxygenated hemoglobin data, only oxygenated hemoglobin were used for this analysis [26-28].

Figure 1. Illustration of the functional near-infrared spectroscopy channels and their respective regions, with the change in accumulated oxygenated hemoglobin ($accHbO_2$) in each region. (A) Comparison of changes in $accHbO_2$ in the right ventrolateral prefrontal cortex in the app use and control groups (F1,44=8.22; *P*=.006). (B) Comparison of changes in $accHbO_2$ values in the right orbitofrontal cortex in the app use and control groups (F1,44=8.88, *P*=.005).

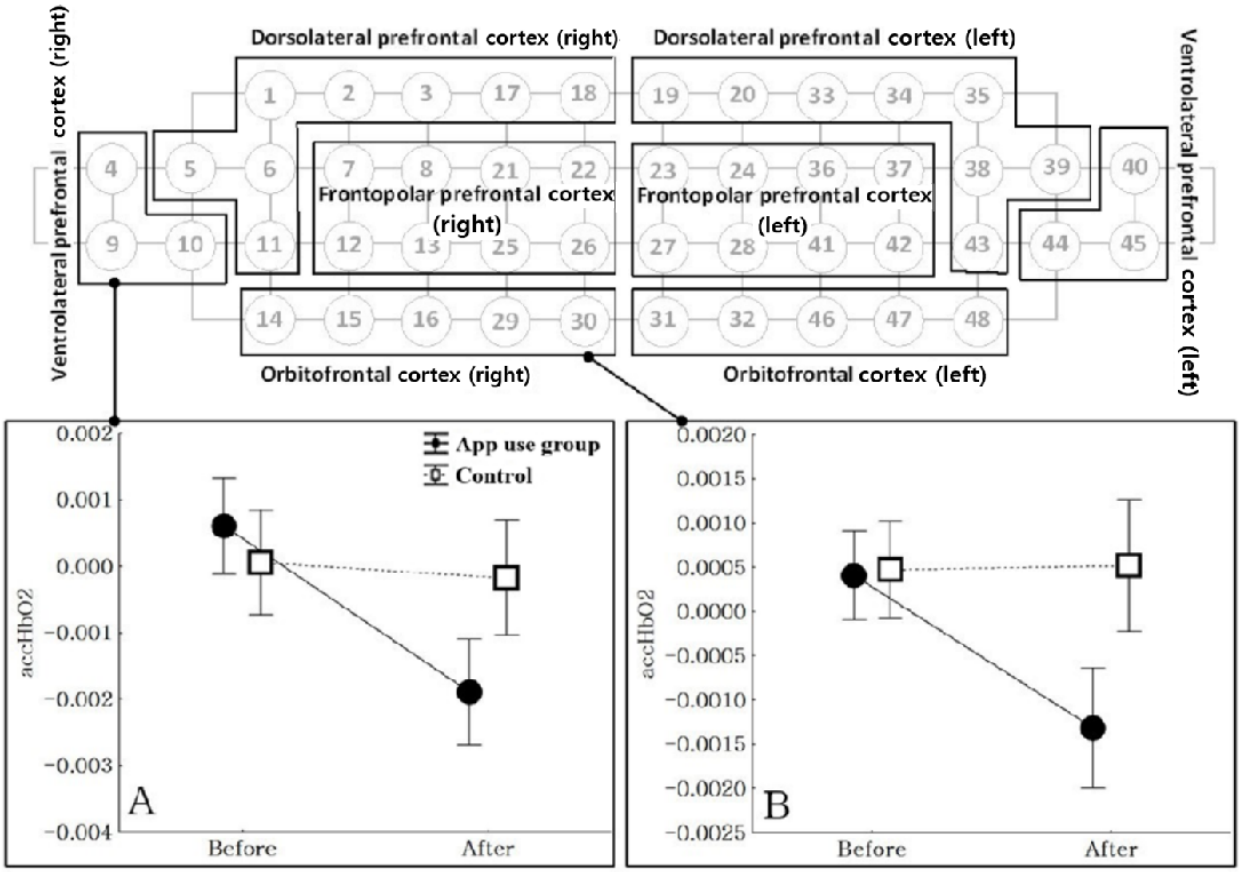

The means and SDs for $accHbO_2$ were calculated from regions of interest (ROIs) in the right and left dorsolateral prefrontal cortices (DLPFCs), right and left ventrolateral prefrontal cortices (VLPFCs), right and left frontopolar cortices (FPCs), and right and left orbitofrontal cortices (OFCs), based on Brodmann area 46. The right and left DLPFCs comprise channels 1, 2, 3, 5, 6, 11, 17, and 18 and channels 19, 20, 33, 34, 35, 38, 39, and 43, respectively. The right and left VLPFCs comprise channels 4, 9, and 10 and channels 40, 44, and 45, respectively. The right and left FPCs comprise channels 7, 8, 12, 13, 21, 22, 25, and 26 and channels 23, 24, 27, 28, 36, 37, 41, and 42, respectively. The right and left OFCs comprise channels 14, 15, 16, 29, and 30 and channels 31, 32, 46, 47, and 48, respectively (Figure 1).

Digital App for Panic Disorder

The app for panic disorder is a smartphone-based app for treatment of clinical symptoms of panic disorder. The mobile app has 4 categories: diary, education, quest, and serious games. The diary category has three items: (1) assessment of daily psychological status, including mood and anxiety; (2) assessment of panic symptoms, including frequency and severity; and (3) consumption of medication, including regular medication and pro re nata medications. The education category has three items: (1) knowledge about panic disorders, (2) knowledge about medications for panic disorder, and (3) knowledge about panic disorder treatment, including CBT, breathing therapy, and positive thinking therapy. The quests include two treatments: (1) eye movement desensitization and reprocessing therapy and (2) positive thinking therapy. The serious games include two games: (1) a breathing game and (2) an exposure therapy game.

The diary, education, and serious games (ie, the breathing game and exposure therapy game) are important parts of CBT for panic disorder [29-32]. The efficacy of CBT for panic disorder

XSL•FO
RenderX

has been examined in various randomized controlled trials [33,34]. Eye movement desensitization and reprocessing therapy are also known to help reduce panic symptoms [35,36]. We confirmed that the replacement of worry with different forms of positive ideation shows beneficial effects [37], so a similar type of positive thinking therapy can also be expected to show benefits. Multimedia Appendix 1 provides additional information on the app.

Ethical Considerations

The study trial was approved by the institutional review board of Chung-Ang University Hospital (1041078-202112-HR-349-01) and written informed consent was obtained from all participants. Participants received an explanation from the researchers that included an overview of the study and a description of the methodology and purpose before deciding to participate. Additionally, they were informed that participation was voluntary, informed about our confidentiality measures, given the option to withdraw, and informed about potential side effects and compensation. Participants in this study received ₩100,000 (US $75.50) as transportation reimbursement. Additionally, the various scales and fNIRS assessments were offered at no cost to the participants. The participants received the results of the tests in the form of a report via postal mail or email after the conclusion of the study. They also receive an explanatory document and consent form from the researchers that included contact information for any inquiries. If the participant agreed to take part in the study after understanding the consent form, the research proceeded. The participants' personal information was not collected. Instead, a unique identifier was assigned to the collected data for the sole purpose of research management.

Study Procedure

A randomized and treatment-as-usual–controlled design was applied in this study. After screening, all participants with panic disorder were randomly assigned to the app use group or the control group. The randomization sequence in our design was generated using SPSS (version 24.0; IBM Corp), with a 1:1 allocation between groups. At baseline and after intervention, all patients with panic disorder were assessed with the PDSS for panic symptoms, the HAM-A for objective anxiety symptoms, and the GAD-7 for subjective anxiety symptoms. At baseline and after intervention, the hemodynamic response in all patients with panic disorder was assessed using NIRSIT. The app use group was asked to use the app for panic disorder 20 minutes per day, 5 times per week, for 4 weeks. The control group was asked to read short educational letters that were delivered via a social network service 5 times per week for 4 weeks. The short letters contained information about panic disorder and its treatment.

Results

Demographic and Clinical Characteristics

After recruitment, 56 patients underwent eligibility assessments. A total of 6 individuals were excluded because they did not meet the inclusion criteria. The remaining patients were divided into 2 groups: 25 were assigned to the app use group and 21 to the control group, as 4 patients were excluded; contact was suddenly lost with 1 patient contact and 1 dropped out for personal reasons. In addition, 2 patients in the control group quit the study after reporting poor benefits from the short educational letters. Therefore, 25 people in the app use group and 21 people in the control group were analyzed. Figure 2 shows the Consolidated Standards of Reporting Trials (CONSORT) flowchart for participant flow through the trial.

XSL•FO
RenderX

Figure 2. Consolidated Standards of Reporting Trials (CONSORT) flowchart.

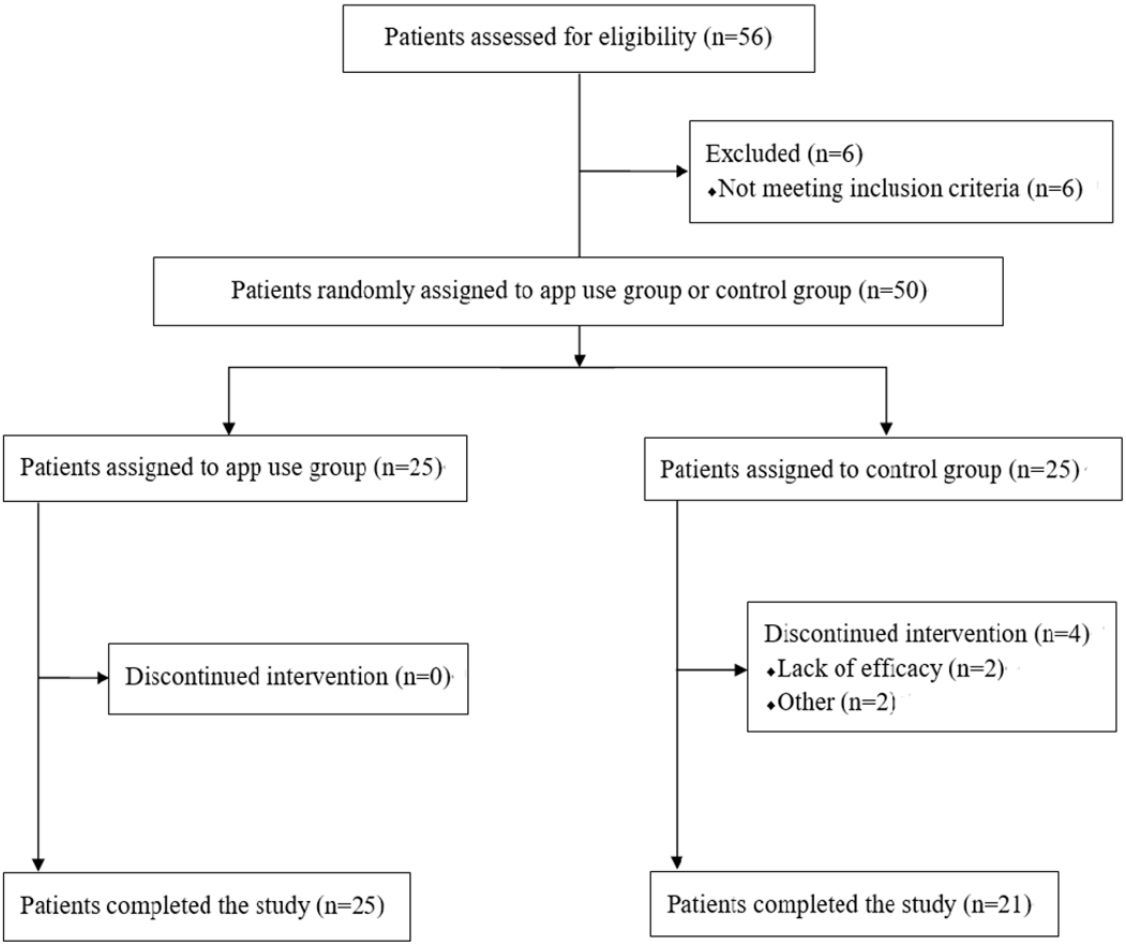

There were no significant differences in age, sex ratio, years of education, marital status, employment status, or substance habits, including smoking and alcohol use, between the app use group and the control group (Table 1).

XSL•FO
RenderX

Table 1. Demographic characteristics. There were no significant differences in demographic characteristics between the app and control groups.

	App use group (n=25)	Control group (n=21)	*t* test (*df*) or chi-square (*df*)	*P* value
Age (years), mean (SD)	36.5 (12)	33.1 (12)	0.97 (45)[a]	.33
Sex, n (%)				
Male	4 (16)	8 (38)		
Female	21 (84	13 (62)	2.89 (1)[b]	.11
Education, n (%)				
High school	5 (20)	9 (43)	5.50 (1)[b]	.06
University	17 (68)	7 (33)		
Undergraduate	3 (12)	5 (24)		
Marital status, n (%)				
Married	10 (40)	9 (43)	0.07 (1)[b]	.97
Not married	14 (56)	11 (52)		
Divorced	1 (4)	1 (5)		
Employment status, n (%)				
Unemployed	2 (8)	5 (24)	3.07 (1)[b]	.38
Student	5 (20)	4 (19)		
Part-time employment	6 (24)	6 (29)		
Regular employment	12 (48)	6 (29)		
Smoking, n (%)				
<10 cigarettes/month	23 (92)	17 (81)	2.59 (1)[b]	.46
10-19 cigarettes/month	1 (4)	2 (10)		
>20 cigarettes/month	1 (4)	2 (10)		
Alcohol, n (%)				
≤1 time/month	15 (60)	14 (67)		
2-4 times/month	9 (36)	7 (33)	1.30 (1)[b]	.73
≥4 times/month	1 (4)	0 (0)		
Clinical scores, mean (SD)				
Hamilton Anxiety Scale	22.7 (4.1)	21.8 (5.6)	0.66 (45)[a]	.51
Generalized Anxiety Disorder–7	10.4 (4.7)	8.9 (5.4)	1.00 (45)[a]	.32
Panic Disorder Severity Scale	14.3 (6.7)	12.4 (6.2)	1.05 (45)[a]	.29

[a]*t* test.

[b]Chi-square.

There were no significant differences in HAM-A score, GAD-7 score, or PDSS score at baseline between the app use group and control group (Table 1).

Comparison of Changes in Clinical Scales Between App Use Group and Control Group

The number of participants with improved panic symptoms in the app use group (20/25, 80%) was greater than in the control group (6/21, 29%; χ^2_1=12.3; *P*=.005).

During the treatment period, the app use group showed greater improvement in PDSS score than the control group ($F_{1,44}$=7.03; *P*=.01). In the app use group, the PDSS score decreased by 42.5% (mean score 14.3, SD 6.5 at baseline and mean score 7.2, SD 3.6 after the intervention), while the score decreased by 14.6% in the control group (mean score 12.4, SD 5.2 at baseline and mean score 9.8, SD 7.9 after intervention) (Figure 3).

XSL•FO
RenderX

Figure 3. Illustration of changes in the (A) Panic Disorder Severity Scale (PDSS), (B) Hamilton Anxiety Scale (HAM-A), and (C) Generalized Anxiety Disorder–7 (GAD-7) scores before and after app use. The comparison used a repeated measures ANOVA. *P<.05.

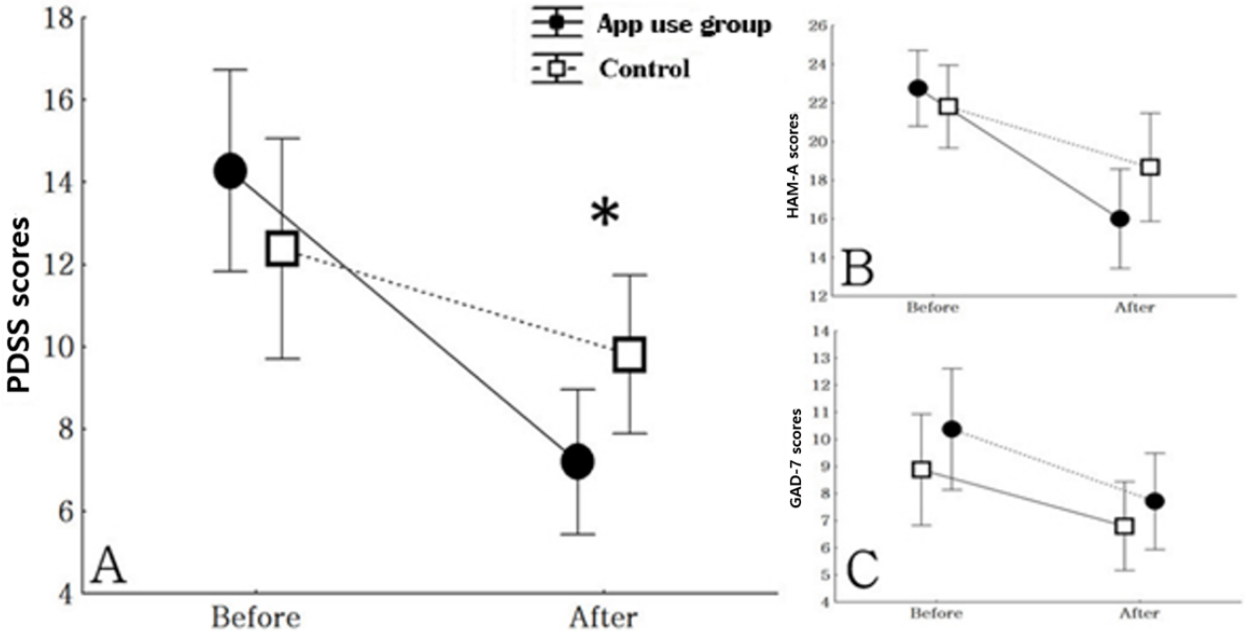

During the treatment period, there were no significant differences in the change in HAM-A scores ($F_{1,44}$=2.83; P=.09) and GAD-7 scores ($F_{1,44}$=0.22; P=.64) between the app use group and control group (Figure 3).

Comparison of Changes in accHbO$_2$ Values Between App Use Group and Control Group

There were no significant differences in accHbO$_2$ in the right (t_{45}=0.84; P=.40) or left (t_{45}=0.73; P=.46) DLPFCs, right (t_{45}=1.04; P=.31) or left (t_{45}=0.88; P=.39) VLPFCs, right (t_{45}=-0.18; P=.86) or left (t_{45}=1.85; P=.07) FPCs, or right (t_{45}=0.33; P=.74) or left (t_{45}=1.89; P=.07) OFCs in the app use and control groups at baseline.

During the treatment period, the app use group showed a greater reduction in accHbO$_2$ in the right VLPFC ($F_{1,44}$=8.22; P=.006) and right OFC ($F_{1,44}$=8.88; P=.005) compared to the control group (Figure 1). During the treatment period, there were no significant differences in the change in accHbO$_2$ in the other ROIs between the app use and control groups.

Correlations Between the Changes in PDSS Scores and the Changes in accHbO$_2$

In all participants (ie, the app use group plus the control group), there was a positive correlation between the change in PDSS score and the change in accHbO$_2$ in the right VLPFC (r=0.44; P=.002). In the app use group, there was a positive correlation between the change in PDSS score and the changes in accHbO$_2$ in the right VLPFC (r=0.42; P=.04). However, in the control group, there was no significant correlation between the change in PDSS score and the change in accHbO$_2$ in the right VLPFC (r=0.22; P=.16).

In all participants, there was a positive correlation between the change in PDSS score and the change in accHbO$_2$ in the right OFC (r=0.44; P=.002). In both the app use group (r=0.34; P=.09) and control group (r=0.33; P=.13), there was no significant correlation between the change in PDSS score and the change in accHbO$_2$ in the right OFC (Figure 4).

XSL•FO
RenderX

Figure 4. Illustration of the correlation between changes in (A) the Panic Disorder Severity Scale (PDSS) score and (B) accumulated glycated hemoglobin ($accHbO_2$) variations across different brain regions.

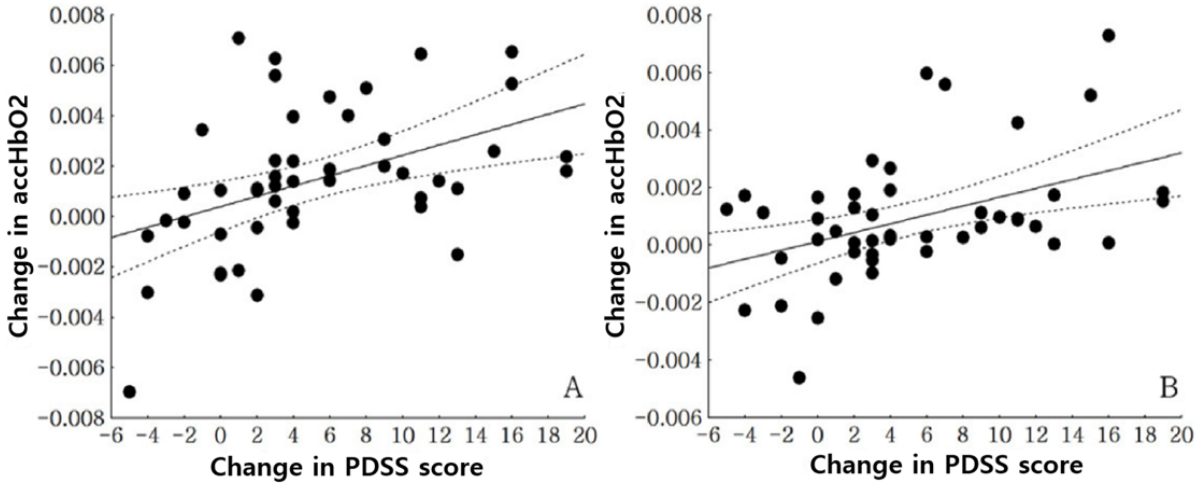

Discussion

Principal Findings

This study showed that a digital app was effective for symptom reduction, as well as decreasing brain activity in the VLPFCs and OFCs, in patients with panic disorder. In addition, the panic disorder symptom improvement was positively correlated with decreased brain activity in the VLPFCs and OFCs in the resting state.

The digital app used in this trial proved to be effective in reducing panic symptoms when compared to the control group, as demonstrated by the reduction in the PDSS score. We believe that this is due to the combined effect of the 4 parts of the program, namely the diary, education, quest, and serious games. The diary component helps identify and correct faulty perceptions and enables cognitive reconstruction. The education component provides information about the nature and physiology of panic disorder. The breathing game helps the participant return to a relaxed condition, while the exposure therapy game allows the participant to experience agoraphobic situations in a safe environment, which helps cognitive restructuring. These are the important parts of CBT for panic disorder and have shown efficacy, as reported earlier [29-32]. The control group also received educational data, including the importance of keeping a diary of one's panic symptoms and how to do it, as well as self-guided direction on breathing exercises, but failed to show a significant reduction of symptoms compared to the app use group. We think this is due to lack of proper feedback in the control group. The app shows real-time feedback on breathing exercises using breathing sounds, and a message was sent if the user of the program failed to use the program for more than 2 days. We know that the therapeutic effect is better when immediate feedback is provided to patients undergoing CBT treatment [38]. Therefore, we think that the decrease in PDSS score was smaller because the control group did not receive feedback from the app.

The control group also received educational data on diary recording, panic disorder information, and how to execute breathing therapy and exposure therapy. We measured their reduction in the PDSS score, but we found it was less than in the app use group due to a lack of proper daily management.

However, the app failed to lead to a difference in the reduction in anxiety, as defined by the HAM-A and GAD-7 scales, between the 2 groups. This is most likely due to a lack of power, as the trial was conducted as a pilot study. Other studies using CBT techniques or serious games have demonstrated reductions in anxiety symptoms in patients with panic disorder [14]. Likewise, this study showed a trend toward a reduction in anxiety symptoms, although this was not statistically significant, and future research with more participants may show that these kinds of programs are also effective in controlling anxiety.

Two major changes in brain activity were noted in the app use group, namely reductions in VLPFC and OFC activation. The functions of the OFC are varied and include control of inappropriate behavior and emotional responses, decision-making, and solving problems [39,40]. Abnormalities in the function of the OFC can cause problems in dealing with anxiety and show that it is deeply involved in the increasing the sense of fear in the fear response [17]. The results of this study confirm that OFC activity decreases as treatment progresses. This reinforces the results of a previous study, which showed that patients with panic disorder had increased OFC activity and that when the panic disorder was treated, the activity of the OFC was reduced, as indicated by decreased cerebral glucose metabolic rates [17,41].

The VLPFC is known to be associated with the amygdala and to maintain flexible attention and responses to environmental threats [42,43]. The amygdala is the backbone of the fear network, and the VLPFC is also known to be deeply involved in the processing of fear [43-45]. Several studies have shown increased activity in patients with panic disorder in the inferior frontal gyrus, which envelops the VLPFC, and other related regions, including the prefrontal cortex, hippocampus, and OFC [16,46,47]. After panic disorder treatment, such as with CBT, decreased amygdala and inferior frontal gyrus activation in fear situations was confirmed [48,49]. Through panic disorder treatment, inferior frontal gyrus activation decreased to a normal level; this happened because the treatment reduced fear cognition related to harm expectancy or attention to threats

XSL•FO
RenderX

[49-51]. We consider that VLPFC activation increases to modulate the amygdala and decreases with treatment for panic disorder.

We believe that these reductions of brain activity in the VLPFC and OFC reflect how the app affected the patients. We know that overprediction of fear or panic is an important feature of anxiety disorders [52]. The app for panic disorder, including diary, education, quest, and serious game components, allowed users to correct their faulty perceptions about fear. As mentioned earlier, the VLPFC and OFC are related to fear management, so we can expect that activity of the VLPFC and OFC will be reduced through repeated app use as users learn how to deal with fear.

Limitations

This study has the following limitations: Most of the patients were effectively treated with alprazolam or other anxiolytics, such as SSRIs. Thus, treatment with antianxiety drugs may have influenced our results. Moreover, this study assessed changes immediately after app use. A long-term follow-up to evaluate the sustainability of the observed improvements would provide valuable insights into the effectiveness of the intervention over time. App use time could be easily tracked for the app use group; however, it was challenging to independently monitor the time the control group spent reading educational materials. Due to the limitations of available research tools, no investigation has been conducted on deep brain structures such as the amygdala, which is most closely related to panic disorders.

Conclusions

We believe that this app for panic disorder effectively reduces symptoms and noticeably impacts brain activity in specific areas. We observed a positive link between improvement in panic symptoms and decreased brain activity in the VLPFCs and OFCs in a resting state. These findings support the use of targeted interventions to determine the brain's contribution to symptom relief. Further research should explore the duration of these positive effects and make digital therapy accessible to more individuals, thus unlocking its full potential in mental health care.

Data Availability

The data sets generated and analyzed during this study are not publicly available as they contain information that could compromise the privacy and consent of the research participants. However, the transformed data are available upon reasonable request from the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital app for panic disorder.
[PDF File (Adobe PDF File), 496 KB - jmir_v26i1e51428_app1.pdf]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1237 KB - jmir_v26i1e51428_app2.pdf]

References

1. Roy-Byrne PP, Craske MG, Stein MB. Panic disorder. Lancet 2006 Sep 16;368(9540):1023-1032. [doi: 10.1016/S0140-6736(06)69418-X] [Medline: 16980119]
2. Kessler RC, Chiu WT, Jin R, Ruscio AM, Shear K, Walters EE. The epidemiology of panic attacks, panic disorder, and agoraphobia in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2006 Apr;63(4):415-424 [FREE Full text] [doi: 10.1001/archpsyc.63.4.415] [Medline: 16585471]
3. Meuret AE, White KS, Ritz T, Roth WT, Hofmann SG, Brown TA. Panic attack symptom dimensions and their relationship to illness characteristics in panic disorder. J Psychiatr Res 2006 Sep;40(6):520-527. [doi: 10.1016/j.jpsychires.2005.09.006] [Medline: 16293263]
4. Locke A, Kirst N, Shultz CG. Diagnosis and management of generalized anxiety disorder and panic disorder in adults. Am Fam Physician 2015 May 01;91(9):617-624 [FREE Full text] [doi: 10.35594/KATA.2022.29.1.006] [Medline: 25955736]
5. Bruce SE, Vasile RG, Goisman RM, Salzman C, Spencer M, Machan JT, et al. Are benzodiazepines still the medication of choice for patients with panic disorder with or without agoraphobia? Am J Psychiatry 2003 Aug;160(8):1432-1438. [doi: 10.1176/appi.ajp.160.8.1432] [Medline: 12900305]
6. Owen R, Tyrer P. Benzodiazepine dependence. A review of the evidence. Drugs 1983 Apr;25(4):385-398. [doi: 10.2165/00003495-198325040-00003] [Medline: 6133736]
7. Cascade E, Kalali AH, Kennedy SH. Real-world data on SSRI antidepressant side effects. Psychiatry (Edgmont) 2009 Feb;6(2):16-18 [FREE Full text] [Medline: 19724743]

XSL•FO
RenderX

8. Breilmann J, Girlanda F, Guaiana G, Barbui C, Cipriani A, Castellazzi M, et al. Benzodiazepines versus placebo for panic disorder in adults. Cochrane Database Syst Rev 2019 Mar 28;3(3):CD010677 [FREE Full text] [doi: 10.1002/14651858.CD010677.pub2] [Medline: 30921478]
9. Schuurmans J, Comijs H, Emmelkamp PMG, Weijnen IJC, van den Hout M, van Dyck R. Long-term effectiveness and prediction of treatment outcome in cognitive behavioral therapy and sertraline for late-life anxiety disorders. Int Psychogeriatr 2009 Dec;21(6):1148-1159. [doi: 10.1017/S1041610209990536] [Medline: 19860993]
10. Rufer M, Albrecht R, Schmidt O, Zaum J, Schnyder U, Hand I, et al. Changes in quality of life following cognitive-behavioral group therapy for panic disorder. Eur Psychiatry 2010 Jan;25(1):8-14. [doi: 10.1016/j.eurpsy.2009.05.003] [Medline: 19545979]
11. Shear MK, Houck P, Greeno C, Masters S. Emotion-focused psychotherapy for patients with panic disorder. Am J Psychiatry 2001 Dec;158(12):1993-1998. [doi: 10.1176/appi.ajp.158.12.1993] [Medline: 11729015]
12. Digital health, digital medicine, digital therapeutics (DTx): What's the difference. HealthXL. 2023 Oct 23. URL: https://www.healthxl.com/blog/digital-health-digital-medicine-digital-therapeutics-dtx-whats-the-difference [accessed 2023-03-19]
13. Cho C, Lee H. Could digital therapeutics be a game changer in psychiatry? Psychiatry Investig 2019 Feb;16(2):97-98 [FREE Full text] [doi: 10.30773/pi.2019.01.20] [Medline: 30808114]
14. Kaplan A, Mannarino AP, Nickell PV. Evaluating the impact of Freespira on panic disorder patients' health outcomes and healthcare costs within the Allegheny Health Network. Appl Psychophysiol Biofeedback 2020 Sep;45(3):175-181. [doi: 10.1007/s10484-020-09465-0] [Medline: 32342249]
15. Hoshi Y. Functional near-infrared optical imaging: utility and limitations in human brain mapping. Psychophysiology 2003 Jul;40(4):511-520. [doi: 10.1111/1469-8986.00053] [Medline: 14570159]
16. Bystritsky A, Pontillo D, Powers M, Sabb FW, Craske MG, Bookheimer SY. Functional MRI changes during panic anticipation and imagery exposure. Neuroreport 2001 Dec 21;12(18):3953-3957. [doi: 10.1097/00001756-200112210-00020] [Medline: 11742218]
17. Milad MR, Rauch SL. The role of the orbitofrontal cortex in anxiety disorders. Ann N Y Acad Sci 2007 Dec;1121(1):546-561. [doi: 10.1196/annals.1401.006] [Medline: 17698998]
18. Shear MK, Brown TA, Barlow DH, Money R, Sholomskas DE, Woods SW, et al. Multicenter collaborative Panic Disorder Severity Scale. Am J Psychiatry 1997 Nov;154(11):1571-1575. [doi: 10.1176/ajp.154.11.1571] [Medline: 9356566]
19. Lim YJ, Yu BH, Kim JH. Korean Panic Disorder Severity Scale: construct validity by confirmatory factor analysis. Depress Anxiety 2007;24(2):95-102. [doi: 10.1002/da.20206] [Medline: 16845647]
20. Maier W, Buller R, Philipp M, Heuser I. The Hamilton Anxiety Scale: reliability, validity and sensitivity to change in anxiety and depressive disorders. J Affect Disord 1988;14(1):61-68. [doi: 10.1016/0165-0327(88)90072-9] [Medline: 2963053]
21. Kim CY. Psychiatric Assessment Instruments. Seoul, South Korea: Hana Psychiatric Publishing; 2001:99-101.
22. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
23. Screener overview. PHQ Screeners. URL: https://www.phqscreeners.com/select-screener [accessed 2024-03-27]
24. Choi J, Kim J, Hwang G, Yang J, Choi M, Bae H. Time-divided spread-spectrum code-based 400 fw-detectable multichannel fNIRS IC for portable functional brain imaging. IEEE J Solid-State Circuits 2016 Feb;51(2):484-495 [FREE Full text] [doi: 10.1109/jssc.2015.2504412]
25. Yamamoto T, Maki A, Kadoya T, Tanikawa Y, Yamad Y, Okada E, et al. Arranging optical fibres for the spatial resolution improvement of topographical images. Phys Med Biol 2002 Sep 21;47(18):3429-3440. [doi: 10.1088/0031-9155/47/18/311] [Medline: 12375830]
26. Cui X, Bray S, Bryant DM, Glover GH, Reiss AL. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. Neuroimage 2011 Feb 14;54(4):2808-2821 [FREE Full text] [doi: 10.1016/j.neuroimage.2010.10.069] [Medline: 21047559]
27. Miyai I, Tanabe HC, Sase I, Eda H, Oda I, Konishi I, et al. Cortical mapping of gait in humans: a near-infrared spectroscopic topography study. Neuroimage 2001 Nov;14(5):1186-1192. [doi: 10.1006/nimg.2001.0905] [Medline: 11697950]
28. Strangman G, Culver JP, Thompson JH, Boas DA. A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. Neuroimage 2002 Oct;17(2):719-731. [Medline: 12377147]
29. Barlow DH, Craske MG, Cerny JA, Klosko JS. Behavioral treatment of panic disorder. Behav Ther 1989;20(2):261-282. [doi: 10.1016/s0005-7894(89)80073-5]
30. Clark D, Salkovskis PM, Hackmann A, Wells A, Ludgate J, Gelder M. Brief cognitive therapy for panic disorder: a randomized controlled trial. J Consult Clin Psychol 1999 Aug;67(4):583-589. [doi: 10.1037//0022-006x.67.4.583] [Medline: 10450630]
31. Telch MJ, Lucas JA, Schmidt NB, Hanna HH, LaNae Jaimez T, Lucas RA. Group cognitive-behavioral treatment of panic disorder. Behav Res Ther 1993 Mar;31(3):279-287. [doi: 10.1016/0005-7967(93)90026-q] [Medline: 8476402]
32. Bouchard S, Gauthier J, Laberge B, French D, Pelletier M, Godbout C. Exposure versus cognitive restructuring in the treatment of panic disorder with agoraphobia. Behav Res Ther 1996 Mar;34(3):213-224. [doi: 10.1016/0005-7967(95)00077-1] [Medline: 8881091]

XSL•FO
RenderX

33. Barlow D, Gorman JM, Shear MK, Woods SW. Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. JAMA 2000 May 17;283(19):2529-2536. [doi: 10.1001/jama.283.19.2529] [Medline: 10815116]
34. Gould RA, Ott MW, Pollack MH. A meta-analysis of treatment outcome for panic disorder. Clin Psychol Rev 1995 Jan;15(8):819-844. [doi: 10.1016/0272-7358(95)00048-8]
35. Faretta E. EMDR and cognitive behavioral therapy in the treatment of panic disorder: a comparison. J EMDR Prac Res 2013 Jan 01;7(3):121-133. [doi: 10.1891/1933-3196.7.3.121]
36. Goldstein AJ, Feske U. Eye movement desensitization and reprocessing for panic disorder: a case series. J Anxiety Disord 1994;8(4):351-362. [doi: 10.1016/0887-6185(94)00023-9]
37. Eagleson C, Hayes S, Mathews A, Perman G, Hirsch CR. The power of positive thinking: pathological worry is reduced by thought replacement in generalized anxiety disorder. Behav Res Ther 2016 Mar;78:13-18 [FREE Full text] [doi: 10.1016/j.brat.2015.12.017] [Medline: 26802793]
38. Shapiro JR, Bauer S. Use of short message service (SMS)-based interventions to enhance low intensity CBT. Oxford guide to low intensity CBT interventions. In: Bennett-Levy J, editor. Oxford Guide to Low Intensity CBT Interventions. Oxford, UK: Oxford University Press; 2010:281-286.
39. Schoenbaum G, Roesch M. Orbitofrontal cortex, associative learning, and expectancies. Neuron 2005 Sep 01;47(5):633-636 [FREE Full text] [doi: 10.1016/j.neuron.2005.07.018] [Medline: 16129393]
40. O'Doherty J. Can't learn without you: predictive value coding in orbitofrontal cortex requires the basolateral amygdala. Neuron 2003 Aug 28;39(5):731-733 [FREE Full text] [doi: 10.1016/s0896-6273(03)00525-7] [Medline: 12948440]
41. Nordahl TE, Stein MB, Benkelfat C, Semple WE, Andreason P, Zametkin A, et al. Regional cerebral metabolic asymmetries replicated in an independent group of patients with panic disorders. Biol Psychiatry 1998 Nov 15;44(10):998-1006. [doi: 10.1016/s0006-3223(98)00026-2] [Medline: 9821564]
42. Monk CS, Nelson EE, McClure EB, Mogg K, Bradley BP, Leibenluft E, et al. Ventrolateral prefrontal cortex activation and attentional bias in response to angry faces in adolescents with generalized anxiety disorder. Am J Psychiatry 2006 Jun;163(6):1091-1097. [doi: 10.1176/ajp.2006.163.6.1091] [Medline: 16741211]
43. Hariri AR, Mattay VS, Tessitore A, Fera F, Weinberger DR. Neocortical modulation of the amygdala response to fearful stimuli. Biol Psychiatry 2003 Mar 15;53(6):494-501. [doi: 10.1016/s0006-3223(02)01786-9] [Medline: 12644354]
44. Gorman J, Kent JM, Sullivan GM, Coplan JD. Neuroanatomical hypothesis of panic disorder, revised. Am J Psychiatry 2000 Apr;157(4):493-505. [doi: 10.1176/appi.ajp.157.4.493] [Medline: 10739407]
45. Tian S, Huang F, Gao J, Li P, Ouyang X, Zhou S, et al. Ventrolateral prefrontal cortex is required for fear extinction in a modified delay conditioning paradigm in rats. Neuroscience 2011 Aug 25;189:258-268. [doi: 10.1016/j.neuroscience.2011.05.002] [Medline: 21596104]
46. Ball TM, Ramsawh HJ, Campbell-Sills L, Paulus MP, Stein MB. Prefrontal dysfunction during emotion regulation in generalized anxiety and panic disorders. Psychol Med 2013 Jul;43(7):1475-1486 [FREE Full text] [doi: 10.1017/S0033291712002383] [Medline: 23111120]
47. Dresler T, Hindi Attar C, Spitzer C, Löwe B, Deckert J, Büchel C, et al. Neural correlates of the emotional Stroop task in panic disorder patients: an event-related fMRI study. J Psychiatr Res 2012 Dec;46(12):1627-1634. [doi: 10.1016/j.jpsychires.2012.09.004] [Medline: 23058446]
48. Beutel M, Stark R, Pan H, Silbersweig D, Dietrich S. Changes of brain activation pre-post short-term psychodynamic inpatient psychotherapy: an fMRI study of panic disorder patients. Psychiatry Res 2010 Nov 30;184(2):96-104. [doi: 10.1016/j.pscychresns.2010.06.005] [Medline: 20933374]
49. Kircher T, Arolt V, Jansen A, Pyka M, Reinhardt I, Kellermann T, et al. Effect of cognitive-behavioral therapy on neural correlates of fear conditioning in panic disorder. Biol Psychiatry 2013 Jan 01;73(1):93-101. [doi: 10.1016/j.biopsych.2012.07.026] [Medline: 22921454]
50. Hofmann S. Cognitive processes during fear acquisition and extinction in animals and humans: implications for exposure therapy of anxiety disorders. Clin Psychol Rev 2008 Feb;28(2):199-210 [FREE Full text] [doi: 10.1016/j.cpr.2007.04.009] [Medline: 17532105]
51. Bishop SJ. Neural mechanisms underlying selective attention to threat. Ann N Y Acad Sci 2008;1129(1):141-152. [doi: 10.1196/annals.1417.016] [Medline: 18591476]
52. Rachman S. The overprediction of fear: a review. Behav Res Ther 1994 Sep;32(7):683-690. [doi: 10.1016/0005-7967(94)90025-6] [Medline: 7980354]

Abbreviations

accHbO$_2$: accumulated oxygenated hemoglobin
CBT: cognitive behavioral therapy
DLPFC: dorsolateral prefrontal cortex
DTx: digital therapeutic
fNIRS: functional near-infrared spectroscopy

XSL•FO
RenderX

FPC: frontopolar cortex
GAD-7: Generalized Anxiety Disorder–7
HAM-A: Hamilton Anxiety Scale
mHealth: mobile health
OFC: orbitofrontal cortex
PDSS: Panic Disorder Severity Scale
ROI: region of interest
SSRI: selective serotonin reuptake inhibitor
VLPFC: ventrolateral prefrontal cortex

Edited by A Mavragani; submitted 03.08.23; peer-reviewed by M Aksoy; comments to author 01.09.23; revised version received 11.09.23; accepted 08.03.24; published 12.04.24.

Please cite as:
Kim K, Hwang H, Bae S, Kim SM, Han DH
The Effectiveness of a Digital App for Reduction of Clinical Symptoms in Individuals With Panic Disorder: Randomized Controlled Trial
J Med Internet Res 2024;26:e51428
URL: https://www.jmir.org/2024/1/e51428
doi:10.2196/51428
PMID:38608270

©KunJung Kim, Hyunchan Hwang, Sujin Bae, Sun Mi Kim, Doug Hyun Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Problems and Barriers Related to the Use of mHealth Apps From the Perspective of Patients: Focus Group and Interview Study

Godwin Denk Giebel[1], MSc; Carina Abels[1], Dr; Felix Plescher[1], MA; Christian Speckemeier[1], MSc; Nils Frederik Schrader[1], MSc; Kirstin Börchers[2], Prof Dr; Jürgen Wasem[1], Prof Dr; Silke Neusser[1], Dr; Nikola Blase[1], Dr

[1]Institute for Health Care Management and Research, Universität Duisburg-Essen, Essen, Germany
[2]QM BÖRCHERS CONSULTING+, Herne, Germany

Corresponding Author:
Godwin Denk Giebel, MSc
Institute for Health Care Management and Research
Universität Duisburg-Essen
Weststadt-Carree
Thea-Leymann-Straße 9
Essen, 45127
Germany
Phone: 49 20118 ext 33180
Email: godwin.giebel@medman.uni-due.de

Abstract

Background: Since fall 2020, mobile health (mHealth) apps have become an integral part of the German health care system. The belief that mHealth apps have the potential to make the health care system more efficient, close gaps in care, and improve the economic outcomes related to health is unwavering and already partially confirmed. Nevertheless, problems and barriers in the context of mHealth apps usually remain unconsidered.

Objective: The focus groups and interviews conducted in this study aim to shed light on problems and barriers in the context of mHealth apps from the perspective of patients.

Methods: Guided focus groups and individual interviews were conducted with patients with a disease for which an approved mHealth app was available at the time of the interviews. Participants were recruited via self-help groups. The interviews were recorded, transcribed, and subjected to a qualitative content analysis. The content analysis was based on 10 problem categories ("validity," "usability," "technology," "use and adherence," "data privacy and security," "patient-physician relationship," "knowledge and skills," "individuality," "implementation," and "costs") identified in a previously conducted scoping review. Participants were asked to fill out an additional questionnaire about their sociodemographic data and about their use of technology.

Results: A total of 38 patients were interviewed in 5 focus groups (3 onsite and 2 web-based) and 5 individual web-based interviews. The additional questionnaire was completed by 32 of the participants. Patients presented with a variety of different diseases, such as arthrosis, tinnitus, depression, or lung cancer. Overall, 16% (5/32) of the participants had already been prescribed an app. During the interviews, all 10 problem categories were discussed and considered important by patients. A myriad of problem manifestations could be identified for each category. This study shows that there are relevant problems and barriers in the context of mHealth apps from the perspective of patients, which warrant further attention.

Conclusions: There are essentially 3 different areas of problems in the context of mHealth apps that could be addressed to improve care: quality of the respective mHealth app, its integration into health care, and the expandable digital literacy of patients.

(J Med Internet Res 2024;26:e49982) doi:10.2196/49982

KEYWORDS

mobile health; mHealth; mHealth app; patient; problems; barriers; digital health applications; DiGA; app; barrier; mobile app; health care system; economic outcome; context; mobile phone

XSL•FO
RenderX

Introduction

Background

Worldwide, patients can access an ever-growing number of mobile health (mHealth) apps. Six years ago, there were 325,000 mHealth apps available, which were created by 84,000 different developers [1]. Regardless of the high number of mHealth apps, there is no fundamental quality control by app stores as distribution channels. Thus, unsafe mHealth apps are a potential threat for patient health [2,3].

An indicator for safe mHealth apps is the approval as a medical device. This is achieved by the European conformity or Food and Drug Administration certification in Europe and the United States, respectively [4,5]. Although medical device certification focuses particularly on safety and medical-technical performance with respect to the intended purpose specified by the manufacturer, there are also other areas where problems can arise.

A more extensive certification program is the "Fast-Track Process for Digital Health Applications (DiGA)" established in Germany. In this process, in addition to the European conformity certification as proof of "safety" and "suitability," other requirements, such as "data protection," "information security," "interoperability," and "user-friendliness" are reviewed [6]. Despite the relatively broad testing approach in the fast-track process, it has not been conclusively clarified whether it is sufficient to face all potential problems in the context of mHealth apps. This is partly due to the novelty of the field of care.

In addition to governmental guidelines, there is also a variety of approaches in science to evaluate the quality of mHealth apps. Well-known assessment scales include the Mobile App Rating Scale [7], the User Version of the Mobile Application Rating Scale [8], Enlight [9], and the System Usability Scale [10]. On an aggregated level, a review by Nouri et al [11], which was published in 2018, analyzed such rating tools. The review identified "design," "information/content," "usability," "functionality," "ethical issues," "security and privacy," and "user-perceived value" as relevant quality dimensions [11]. However, similar to the presented governmental certification processes, it is also questionable whether these quality assessment procedures and identified quality dimensions in the scientific field are sufficient to address potential problems in the context of mHealth.

Knowing the existing problems and barriers related to mHealth apps and their use is a fundamental requirement to guarantee the comprehensive quality assurance of mHealth apps. However, only few studies have aimed to identify weaknesses and problems of mHealth apps [12-14]. A scoping review on this topic revealed 10 problem categories: "validity," "usability," "technology," "use and adherence," "data privacy and security," "patient-physician relationship," "knowledge and skills," "individuality," "implementation," and "costs" [14].

However, it remains uncertain whether these categories are relevant from the perspective of users or patients. A systematic review of qualitative studies identified 25 studies focusing on barriers of and facilitators to the use of lifestyle apps. However, the authors emphasized that the included studies have considered only healthy individuals, and thus, the obtained results are not transferable to patients with diseases [15]. In another qualitative study from Australia, patients were recruited in a general practice setting and interviewed about barriers and enablers to the use of mHealth apps. Nevertheless, the authors emphasized that the sample size was skewed toward relatively healthy patients, and future studies should target patients with long-term medical conditions [16].

Objectives

As there is insufficient evidence base regarding the problems and barriers in the context of mHealth apps, qualitative research methods could provide an explorative insight in this research field. Perspectives of both mHealth app users as well as nonusers should be taken into account. This is necessary to reveal existing problems with the use of mHealth apps as well as concerns and barriers preventing patients from using these apps. Therefore, this study surveyed both app using and nonusing patients in focus groups and individual interviews to identify patient-relevant problems and barriers related to the use of mHealth apps.

Methods

We followed the standards of the study by O'Brien et al [17] and subsequently checked the manuscript by applying the 32-item COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [18] to ensure transparency in all aspects of our qualitative research.

Theoretical Framework

Focus groups and interviews were conducted with patients who had a condition for which an approved digital health applications (DiGA) existed at the time of the study. Patients were either actual users of DiGA and other mHealth apps or did not use any mHealth apps. The aim of the research was to shed light on the perspective of patients on mHealth apps. Questions covered "problems and barriers," "facilitating factors," "reasons for attrition," and "important properties" that mHealth apps should meet. This paper focuses especially on problems and barriers identified in the underlying conversations.

The decision for a qualitative research design was made to take an exploratory step into the still poorly researched field of problems and barriers in the context of mHealth apps. Conducting focus groups and interviews allowed the collection of individual opinions. Those can subsequently be used to develop new research hypotheses.

Participant Selection

Patients were recruited via self-help groups and included both users and nonusers of mHealth apps. While patients with experience in mHealth app use could mainly contribute to the existing problems with use, patients without experience could mainly contribute to the barriers to use. Saturation was reached when all relevant indications for which there were approved DiGA at that time were covered.

XSL•FO
RenderX

Before conducting the interviews, participants received an informational letter about the topic "quality assurance of DiGA" and the proceeding of the interviews by email. Furthermore, patients were informed in the cover letter that the focus groups or interviews were part of a project ("QuaSiApps") funded by the German Federal Joint Committee, which aims to develop a continuous quality assurance concept for DiGA. More detailed information was not provided beforehand.

Ethical Considerations

Upon request and presentation of the project schedule to the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen, it was confirmed that ethics approval was not required. This was because, as part of the patient survey, neither personal nor disease-related questions were to be asked. Written informed consent was obtained from all participants before the survey. The participants were informed and asked to consent to the recording of their conversations for subsequent transcription. At the beginning of the interviews, the participants were informed that they could end the conversations at any time without giving a reason and without facing disadvantages of any type.

Setting

The settings of focus groups and interviews were heterogeneous. Conversations were either conducted web-based or in person. While the focus groups each included >3 participants, the interviews were conducted with 1 interviewee at a time. The onsite focus groups were conducted with self-help groups in their respective familiar settings. The web-based interviews were conducted via the videoconferencing platform Zoom (Zoom Video Communications, Inc). The choice for the web-based setting arose from the preference of participants.

No abnormalities were observed in the study setting that could influence patient statements. All patients were in a familiar environment during the interviews. No other people except 2 relatives of participants in 1 focus group were present besides participants and researchers. In 2 web-based focus groups, people did not know each other, which could lead to patients being more reluctant to participate.

There were no known relationships with participants. Notable characteristics of researchers, which could influence the participants, were not identified. There were no conflicts of interest, and the moderators did not take a firm stance for or against the use of mHealth apps.

Data Collection

Each focus group or interview was conducted by 1 of 3 moderators (GDG, CA, and KB). Of the 3 moderators, 2 were female (CA and KB) and 1 was male (GDG). For quality assurance reasons, at least 2 of the moderators participated in each focus group or interview conducted. While 2 moderators hold degrees with health economics background (GDG and CA), 1 moderator originally stems from the field of medicine and holds an MD degree (KB). GDG and CA are researchers at the Institute for Healthcare Management and Research, University of Duisburg-Essen. KB is the managing director of BÖRCHERS CONSULTING+ and an honorary professor at the University of Duisburg-Essen. While CA and KB are very experienced in conducting focus groups, GDG was less experienced but received a detailed briefing.

A prior scoping review [14] on problems and barriers related to the use of mHealth apps similar to the German concept of DiGA served as a basis to develop a noninfluencing interview guideline. This provided a consistent interview framework (Multimedia Appendix 1). The guideline followed a uniform structure. Each topic started with open questions followed by some concrete questions, which were shown if the participants did not find a spontaneous answer. Personal data were not addressed in the guideline.

Focus groups or interviews were recorded either onsite with microphones or web-based via the integrated recording function in Zoom. Data collection took place in the period between the end of September 2021 and the beginning of December 2021. Subsequently, to record transcription, data analysis started in February 2022 and was completed at the end of October 2022. The conduct of the interviews worked without problems, and the guide proved to be understandable. Therefore, no adjustments were made to the method of data generation during the study.

The recordings of the interviews and focus groups were transcribed and pseudonymized by project assistants or employed students. A second person, also a project assistant or student assistant, performed quality assurance of the transcripts. Afterward, the transcripts were loaded into MAXQDA (VERBI Software GmbH) for data analysis, and the audio recordings were deleted. After data extraction, it was no longer possible to trace back individual statements to individual participants.

Qualitative data analysis was based on Mayring [19] and performed by GDG and CA. NB supervised the process and served as the decisive authority in the event of disagreements. As recommended by Mayring [19], deductive codes used in the data analysis were defined before the data analysis. Deductive codes were taken from the results of a scoping review [14]. Inductive subcodes were developed iteratively at the time of data extraction.

Data analysis was conducted in 4 steps to enhance trustworthiness. In the first step, GDG coded each transcript with deductive codes. In the second step, CA checked the coding made in each transcript, and in case of disagreement, problems were solved by discussion. In the third step, GDG developed inductive subcodes, which were discussed in group (GDG, CA, NB, and FP). Finally, subcodes were applied by GDG and checked by CA.

Results

Overview

In total, 38 patients and 2 relatives participated in 5 focus groups and 5 interviews between September and December 2021. A total of 7 conversations took place on the web via Zoom, and 3 conversations took place in person. A structured overview of the focus groups and interviews conducted is provided in Table 1.

XSL•FO
RenderX

Table 1. Overview of the focus groups and interviews conducted (N=40).

Number	Indication	Setting	Format	Patients, n	Date
1	Migraine	Focus group	Web-based	5	September 23, 2021
2	Tinnitus	Interview	Web-based	1	October 20, 2021
3	Obesity	Focus group	On site	10	October 21, 2021
4	Depression	Focus group	Web-based	3	October 26, 2021
5	Tinnitus	Interview	Web-based	1	October 27, 2021
6	Arthrosis	Focus group	On site	7	October 28, 2021
7	Multiple mental health disorders (depression, panic disorder, eating disorder, suicidal thoughts, sleep disorder, and agoraphobia)	Interview	Web-based	1	November 4, 2021
8	Lung cancer	Focus group	On site	8 (+2 relatives)	November 23, 2021
9	Depression	Interview	Web-based	1	November 25, 2021
10	Alcohol addiction and cancer	Interview	Web-based	1	December 7, 2021

Of the 40 included participants, 32 (80%) consented to give information about sociodemographical data (Table 2). Further details on the study participants can be found in Multimedia Appendix 2. The intrinsic motivation and participation of the participants were high. Nevertheless, the moderators made sure that everybody had their say and had the opportunity to express themselves.

Table 2. Core characteristics of the surveyed participants (N=32).

Characteristic	Participants, n (%)
Sex	
Male	11 (34)
Female	21 (66)
Age (y)	
21-45	10 (31)
46-65	13 (41)
≥66	9 (28)
Device available to run apps[a]	
Any device	32 (100)
Smartphone	29 (91)
Computer or laptop	24 (75)
Tablet	16 (50)
Smartwatch	5 (16)
App already prescribed once	
Yes	5 (16)
No	19 (60)
N/A[b]	8 (25)

[a]Multiple answers possible.

[b]N/A: not applicable.

As recruitment was conducted with the help of self-help groups, no binding statement can be made about how many participants declined to participate. None of the included patients dropped out of the study after recruitment.

The findings describe the perspectives of patients on problems and barriers in the context of mHealth apps. During the interviews and focus groups, all the 10 problem categories that served as deductive codes were addressed. The 10 mentioned

XSL•FO
RenderX

problem categories, which served as deductive codes, include "validity," "usability," "technology," "use and adherence," "data privacy and security," "patient-physician relationship," "knowledge and skills," "individuality," "implementation," and "costs." The respective definitions of the categories can be found in the published scoping review [14]. The corresponding inductive subcodes are listed in Multimedia Appendix 3. Multimedia Appendix 4 includes all relevant statements of the patients systematized into the problem categories.

Validity

Problems with "validity" were particularly mentioned in the 4 areas of "poor content and quality of information," "lack of (added) value," "lack of therapeutic setting," and "patient safety." The first problem area, "poor content and quality of information," was observed by patients, especially in the lack of empirical evidence, inappropriate content, nonfunctioning links, and deficient exercise instructions:

> *[...] that's why I'm very demanding when it comes to, for example, progressive muscle relaxation or something like that [...] and I find the [instructions] that are in the app-. That doesn't work at all, so for me.* [Patient with migraine]

"Lack of validity, reliability, and accuracy of app-collected data" was identified as a second problem area. For example, in the area of obesity, patients reported the overrating of physical activity within the app. Another concern was that results could be skewed due to, for example, tattoos, heavy arm hair, or athletic activity, or false output could be generated by interconnectivity with other apps:

> *The problem is that another app somewhere is still feeding me data, even in the GoogleFit, so that I have twice the amount of steps every day. And that has blown things up, of course. I was well into the minus range of calories that I could take in. That's not quite mature yet [...]* [Patient with obesity]

The third problem area of "validity" and the possible reason for the discontinuation of use is "lack of (added) value." Patients reported that lack of health improvement would lead them to try something else:

> *Failure to improve health [...] "doesn't help me, I have to try something else."* [Patient with tinnitus]

The "lack of therapeutic setting" was identified as the fourth problem area of "validity." It was explained that the app alone would not be sufficient, and a real person would have to pay attention to what the patient was doing:

> *With me, the problem is more. There has to be someone behind it to watch what I'm doing.* [Patient with obesity]

The last area to be addressed was "patient safety" under the problem area of "validity." It was explicitly stated that the lack of human control and correction posed a direct safety issue to patients. Thus, in the context of guided practice, errors could creep in, or worse, in an emergency, such as the occurrence of suicidal thoughts, no appropriate response could occur:

> *Or at some point I had indicated that I was having suicidal thoughts. And only then did the app unlock that you can call the telephone counselling service directly from the app.* [Patient with depression]

The adverse effects that could occur through the app were also described as problematic for patient safety. For example, patients described a constant preoccupation with symptoms, symptom amplification, or the activation of triggers by the app as harmful. A final issue was the risk of social withdrawal:

> *Many people don't want to talk about [the disease] and withdraw. And I think that if you only have this app in front of you and nothing personal at all, that's very difficult for cancer patients. For all seriously ill people.* [Patient with cancer]

Usability

The problems mentioned in the category of "usability" can basically be divided into "problems with the instructions" on the one hand and "difficulties with the usage" on the other. Thus, patients rejected a time-consuming introduction to the functions of the app and were dismissive toward independently searching through functions in the app:

> *I would say, from my experience, I really don't have the nerve to read it at that moment. [...] But you could do it quite simply, like a how-to with screenshots. So rather with pictures where you have to click on them.* [Patient with migraine]

Further difficulties in the category of "usability" were found, particularly in the lack of an easy overview and nonintuitive app design. A deficient menu navigation was explicitly criticized:

> *But then there's kind of no main menu after that. So when I finish it, I would like to have a main menu like that.* [Patient with migraine]

Technology

In addition to the fundamental problem expressed that technical complexities are particularly disturbing in the case of illness, 3 areas of technological problems were identified: "problems with the software," "problems with the hardware," and "problems with interoperability and network connection."

In the area of software, it was criticized that some apps were only available as web apps but not as mobile apps. Furthermore, patients mentioned the danger of viruses as well as problems, errors, and bugs in the software itself. Problems, errors, and bugs in the software could show up in app crashes and were seen because of poor programming:

> *And it was programmed in a very very clumsy and annoying way.* [Patient with tinnitus]

In addition to software issues, problems also arose concerning hardware, specifically with the devices in use and their updates. One point raised was related to an existing concern that updates for the smartphone would impair the functionality of the app.

Furthermore, besides updates, some technical features, such as digital displays, were criticized. Thus, for example, in the case

XSL•FO
RenderX

of migraine, working on the screen was described to be problematic due to light sensitivity:

> *I think that I would then use it even easier if I could sort of just talk to the phone and not have to sit down and look at the display big time.* [Patient with migraine]

Further problems were observed with running and connected devices. Specifically, the limitation of using the app on only 1 device and the absence of the option to use it across multiple devices (such as smartphones or tablets) were considered problematic. On the other hand, additional devices, such as smartwatches and virtual reality glasses, were partially dismissed:

> *If you think about possible accessories, for example. I'm kind of getting out of the game with any more tech accessories. So, I don't use a smartwatch, nor do I have VR goggles, and I don't really want to.* [Patient with cancer]

In the area of compatibility and network connection, an example was cited in which data imported from other apps did not work properly. Furthermore, dependence on an internet connection was criticized:

> *And that I then-, we also had a bad Internet connection there, then I couldn't use that at all like that in some cases.* [Patient with tinnitus]

Use and Adherence

The problem area of "use and adherence" included, in particular, "problems due to the attitude of users"; "problems that occurred in the context of usage"; "inadequate, unappealing content design"; and "limited time resources of the patients."

A fundamental problem stated by the patients was the attitude of users. They often have a lack of motivation and engagement. It was expressed that it may be unnecessary to prescribe apps to those who have a negative attitude in advance. On the one hand, lack of motivation may be fundamental, but on the other hand, it may be due to an aversion to the digital format:

> *Yes, maybe that you don't have the motivation to do it alone. Many need the group to do the exercises.* [Patient with osteoarthritis]

> *The second that I don't want to do therapy then on digital devices because as a journalist I'm on the go all day with tablet, cell phone and laptop.* [Patient with tinnitus]

If patients actually use the apps, problems can also arise in the (social) context of use. These include, for example, a low prioritization of use over other obligations such as work or private commitments, a predetermined daily schedule or paternalism exercised by the app, or distraction by other people:

> *And I can't do that now in front of the running TV, when the family is sitting around, of course that doesn't work. And if someone says: "I don't have the time, because I have small children or we have a lot going on in the apartment or something-." So you do need a bit of a place to retreat.* [Patient with tinnitus]

Furthermore, use can also be inhibited by the content design. This happens when the content is not very varied and boring or when it lacks human or emotional aspects:

> *For me, it was so that the app-, I didn't find it very-, so it had become somehow boring for me. So I didn't enjoy entering my things there anymore, because it was always the same. It always just recorded the app and then calculated it for me and that was it.* [Patient with obesity]

Another problem and possible reason for discontinuation of use across many indications was that apps were too extensive and time consuming. This problem was exacerbated by a lack of interruption or pause options, among other things:

> *It's just very time-consuming, and that's the thing that bothered me a little bit before.* [Patient with migraine]

Data Privacy and Security

In the context of data privacy and security, fears about the loss of personal data was noted on the patients' side. Besides insufficient data protection within the app, another problem was the lack of trust in running operating systems. Furthermore, patients rejected indiscriminate data provision to third parties:

> *So if I have the feeling that the data is not in good hands [...] I do turn my innermost thoughts inside out with these apps. Especially when it comes to mental illness.* [Patient with depression]

> *[...] if I now have to prove my sporting activities to the health insurance company, then they just get the sporting activities, but they are not supposed to know, how I felt [...]* [Patient with obesity]

Patient-Physician Relationship

There were mainly 2 different types of problems regarding the physician-patient relationship. On the one hand, there were "problems with usage not accompanied by a doctor or therapist," and on the other hand, there were "problems in the patient-physician relationship triggered by the app."

A fundamental problem in the use of the app without medical or therapeutic accompaniment was seen in the limited scope of support that apps can offer. Therefore, care must be taken to ensure that they do not lead to the neglect of personal therapeutic contact:

> *And also to that, that an app has limits, I think that's important and that should be apps... and that should also, when the doctors then eventually know what that is, also tell the patients [laughs], "Yes, that's just an app and it has its limits."* [Patient with depression]

Furthermore, patients expressed that app use can have an impact on the patient-physician relationship for various reasons. This occurs, for example, when patients need to report their app use to physicians or when health care providers are dismissive of the use, perceiving the app as competition to their abilities:

> *Because, of course, it must be assumed that the prescribing physician or the attending physician is*

XSL•FO
RenderX

> *also behind it. If he now inwardly rejects it, then it can of course be that the relationship with him suffers.* [Patient with tinnitus]

Finally, a particularly problematic situation for patients can arise when there are inconsistent opinions between the physician and the app. Patients then find themselves in the situation of having to determine which information is correct:

> *Do I trust my doctor more and the statement he says? Or do I also additionally trust what an app suggests I might do? That's where I stand in between.* [Patient with cancer]

Knowledge and Skills

Problems in the area of "knowledge and skills" were seen on the "patient side" but also on the "side of physicians and therapists." On the part of the patients, a distinction was made here between actual "lack of skills, knowledge and experience" on the one hand and "perception" on the other hand.

The respondents saw potential for improvement particularly in the areas of patients' technical skills and media competence. Although in the case of media competence, the assessment was seen as particularly problematic, in the technical area, the focus was on the lack of skills:

> *I can't operate at all, I can't operate at all. I'm not a person who can handle digital things at all.* [Patient with cancer]

Participants expressed that older patients do not always consider themselves as a target group. Older adults often had a preference for analog methods:

> *[...] I've already handed over the 60, we just didn't grow up with these things in our hands. So maybe we just don't want to be digital for a change, and the grip on paper is just more familiar.* [Patient with migraine]

In some cases, patients did not use apps because there was a lack of confidence in the therapeutic effect, the apps were perceived as complicated, or patients were fundamentally opposed to the technology:

> *I haven't used an app until now, for the reason that I always imagined it to be very complicated, I have to read a lot, scroll.* [Patient with obesity]

Issues regarding "knowledge and skills" were also observed on the practitioners' and health insurers' side, particularly stemming from a poor level of knowledge:

> *I think it would make a lot of sense. So you would just have to somehow bring it to the doctors a bit, because I think most of them don't really know how to deal with it yet.* [Patient with depression]

Individuality

A further group of problems and a possible reason for discontinuing use was observed in the lack of "individuality." In particular, "inadequate adaption to individual user abilities and needs" as well as the problem of "too generalized approaches" were addressed. It was emphasized that individual abilities are different, older users have special requirements for app use, and preferences exist for different (physical and mental) exercises:

> *Is so individual, of course. Everyone has their own problems of course, everyone perhaps does their own exercises.* [Patient with osteoarthritis]

Approaches that are too generalized were considered a further problem in this category. Depending on the stage of the disease and the individual, different configurations or designs were desirable:

> *Tinnitus symptoms are very individual, both in terms of the triggers and the expression. And dealing with it and then the question is always, if such an app comes as a therapy concept, how fine-tuned is that [...]"* [Patient with tinnitus]

Implementation

The field of "implementation" included many different problems. The focus was particularly on the "barriers to access" and "additional burden for patients" as well as "additional burden for health care providers" and "low acceptance by health care providers." Other points mentioned were "difficult transfer into clinical practice," "too many options to use," and "fear of consequences due to app usage."

Patients encountered barriers to access, especially due to language barriers and lack of accessibility. The individual degree of illness and age should also be taken into account in the context of access. Further barriers to access were seen in complicated access to the app or complicated acquisition via the health insurance company and use limited to 1 device:

> *So, yes, it has to be low-threshold and if it's not, I think it's also difficult from a certain age then to deal with it or also from a certain degree of illness.* [Patient with depression]

The interviewees stated that the use of apps was associated with additional effort, both for themselves and for the physicians involved. In particular, the time required was emphasized critically:

> *Because that's another new task for the doctors. First of all, they have so much to do with their patients, who are all individuals, focused on their specialty. So it becomes difficult to bring it all together afterwards. Because there won't be just one patient who wants to use such an app, but five or even 100.* [Patient with cancer]

Patients stated that problems also arise when service providers only have a low level of acceptance toward the apps. This can be due to a lack of interest, competition between the app and health care providers, or a fundamental rejection. Physicians might perceive apps as interfering with their expertise, might express disapproval, or might advise patients against using apps. Divergent opinions between different physicians were mentioned as a particular problem:

> *Do I go to my urologist, do I go to my thoracic surgeon, do I go to my family doctor, or do I go to*

XSL•FO
RenderX

> *my dermatologist? And if one is in favor, but the others are against, what do I do?* [Patient with cancer]

"Difficult transfer into clinical practice" was identified as a further problem. The transfer might be impeded by a lack of presence of the apps. It was criticized that the apps are still rarely recommended by health care providers and that the social awareness is still low:

> *Because just no one has recommended this app yet. So I've really been through a bunch of psychologists and I-, several clinics [...] But the topic of the app is not in the waiting room with a flyer, nor at the doctor's office, nor in the support group [...]* [Patient with depression]

The confusing app market was also described as complicated. Especially for patients who are already limited by diseases, the search and selection process can be complicated and time consuming:

> *I'm struggling with my illness or I'm struggling with my various illnesses, and then now there's this additional psychological pressure: "I might have the wrong thing I'm using after all. Maybe there's something significantly better in the meantime." These digital health apps, they're totally confusing to me as a user.* [Patient with cancer]

A final, but not specified, problem was seen in the consequences, especially for the health insurance of patients. It was questioned whether using the apps might have any consequences on their insurance coverage:

> *Then what does that do to my health insurance coverage afterward?* [Patient with cancer]

Costs

Patients stated that costs might result in different problems. Patients, health insurers, and physicians were named as the groups of people affected by costs. Problems were observed in "loss of revenue," "low willingness to pay," "alternative financing methods," and "waste of money."

While a loss of revenue was considered problematic on the side of the service providers, a low willingness to pay was observed on the side of the patients. The lack of willingness to pay was related, on the one hand, to the apps themselves and, on the other hand, to the accessories required:

> *So for me, the reason was that it's paid now and that-, so this app was super good, except for some initial difficulties, but just, I didn't want to pay for it.* [Patient with obesity]

Alternative financing methods, especially through advertising, were another problem mentioned. Both advertising for products within the app and advertising about the app in other media were discussed:

> *If, for example, advertising were to come on all at once. If now from different drug manufacturers there is always something inserted there.* [Patient with cancer]

The last point mentioned in the category of costs was that under certain circumstances, mHealth apps could lead to a waste of money. Two such situations were discussed. First, a lack of control over whether the app is used was considered critical; second, payment, despite a lack of evidence, was criticized:

> *And it was, that was also point of attack, of those who do not think anything at all of e-health, that the scientific proof is missing and now the insurance already pays.* [Patient with tinnitus]

Discussion

Principal Findings

Our study showed that patients face a multitude of problems and barriers in the context of mHealth apps. While the problem categories were determined based on a previously conducted scoping review [14], many new expressions of problems emerged in each category. These expressions provide a deeper insight into the thinking and attitudes of patients related to mHealth apps.

There were mainly 3 different factors related to problems and barriers in the context of mHealth app use. First, the mHealth apps themselves could lead to problems. Such problems were observed, for example, in defective design, technical aspects, or low or no added value for patients.

Second, the integration into the health care system was considered partially problematic. Thus, problems were found, for instance, in remuneration and lack of time of health care providers, influences on the relation between health care providers and patients, and a lack of presence of the topic in the health care system.

Third, the users themselves as well as their attitude and knowledge were found to be a barrier for mHealth app use. Several participants expressed a lack of interest in using mHealth apps and explained that they sometimes lack knowledge and skills to use the digital technology. This fact was especially pronounced for older patients.

In our study, we included patients with different diseases (migraine, tinnitus, obesity, depression, arthrosis, cancer, and alcohol addiction) to receive broad feedback from people with diverse needs. While most of the problems were overarching and concerned the general use of mHealth apps, some problems were disease specific. The latter were mainly found in the category of individuality but also included technological problems such as light sensitivity leading to an avoidance of screen use in the case of patients with migraine or other problems such as the risk of increased social withdrawal of individuals with cancer because of app use. For other indications, there are additional disease-specific problems that could not be completely covered here.

Implications

To guarantee a sustainable, safe, and effective use of mHealth apps, it is necessary to deal with all three areas where problems can arise or barriers exist: (1) the mHealth apps, (2) the integration into the health care system, and (3) the users.

XSL•FO
RenderX

mHealth Apps

Problems with the mHealth apps themselves were found in the categories of "validity," "usability," "technology," "individuality," "data privacy and security," and "costs." One approach to identify these is to use various quality assessment tools, which aim to guarantee a high quality of mHealth apps [11,20,21]. Quality assessment tools provide a practicable approach to distinguish high-quality mHealth apps from low-quality mHealth apps and thereby indicate if patients might have more or less problems with app use.

Furthermore, patients should be involved during the whole development process of mHealth apps. This includes (1) the need assessment, (2) the design and development, (3) the laboratory evaluation, and (4) the field evaluation. Eligible methods include qualitative data collection methods such as interviews, focus groups, observations, and think-aloud techniques as well as quantitative approaches such as self-report questionnaires [22].

An eligible way to do this is the participatory design [23]. By this, patients would have the possibility to comment on problems and barriers as soon as they become apparent. By including both experienced and nonexperienced mHealth apps users during the development process of mHealth apps, problems and barriers can be minimized for all users. A good example in which a 3-phase user-centered design approach has been successfully implemented is the study by Newton et al [24].

In addition, problems and barriers become apparent through user feedback in app stores. Therefore, user feedback should not only be included during the development but also in the evaluation and rework of mHealth apps. One innovative approach to incorporate user text review is the ACCU^{3}RATE rating scale [25].

Integration of mHealth Apps Into the Health Care System

The results of this study show that problems and barriers are not restricted to the mHealth apps themselves. Thus, it is necessary to obtain more knowledge about the contextual problems and barriers that patients encounter when they want to or actually use mHealth apps.

The focus groups and interviews conducted indicate that, in the perception of patients, health care practitioners are one of the most important stakeholder groups facing problems with mHealth apps. Therefore, it is a fundamental requirement that they are convinced of the positive effects and low risk of mHealth apps for their patients and of the simple and sustainable integration in their daily routine.

However, this is not always the case, as shown by a study on the attitudes of physicians toward mHealth apps falling under the German concept of DiGA [26]. In principle, general practitioners, other outpatient care physicians, and psychotherapists were in favor of prescribing DiGA but faced significant barriers, such as insufficient information, insufficient reimbursement for DiGA-related medical services, lack of medical evidence, and legal and technological uncertainties. Therefore, such problems and barriers faced by health care providers should also be further investigated.

Germany was the first country in the world to incorporate certain mHealth apps (ie, DiGA) as a fixed part into the benefit package of the health care system [26]. Although the German concept is an innovative and commendable approach, it still faces several challenges. Further investigations into problems and barriers related to integrating mHealth apps could provide insights into improving regulatory systems. This would aim to make app use easier while ensuring safety and sustainability.

Another problem, depending on the respective health care system, could be the costs of using mHealth apps for patients. In the German statutory health insurance system, however, the costs are covered and thus represent only an indirect problem for users. Therefore, prices must be negotiated between health insurers and manufacturers, in particular.

mHealth App Users

Besides the mHealth apps themselves and their integration into health care, problems on the patients' side were also reported. In our study as well as in the literature [14], it was found that low digital literacy of patients was seen as a problem in the context of mHealth app use.

According to the technology acceptance model by Davis et al [27], an actual system use is the result of perceived ease of use as well as perceived usefulness. Therefore, another approach to optimize the use of mHealth apps should focus on reinforcing the digital literacy of potential users, especially older patients. This involves supporting them with app use and effectively communicating the benefits of mHealth apps [28,29].

Both our study and other studies [14,16,30] found that age is a problem in the context of mHealth app use. Interestingly, however, the reason for this is predominantly seen in the lack of technological affinity and digital literacy among older people. The extent to which cognitive and psychological decline plays a role should be investigated further.

Limitations

Given the scarce evidence regarding problems and barriers in the context of mHealth app use, qualitative research seems to be an eligible first step to gain further evidence [14]. Nevertheless, qualitative research is always accompanied by uncertainty. The statements made by participants are not necessarily representative for all patients, especially as participants were recruited from self-help groups and presumably had a high motivation to actively shape their to actively shape their experience of living with the disease. Thus, the problems and barriers identified in this study should serve as the first evidence to conduct further qualitative and quantitative studies.

Two points concerning the methods must still be made. First, we were not able to determine the number of people who refused to take part in our study. As we recruited our participants via self-help groups, we do not exactly know how each group distributed the information about our study to their respective members. Second, we did not calculate an agreement rate or other measurements regarding the coding discrepancies. In

XSL•FO
RenderX

retrospect, however, there were very few disagreements and minimal need for discussion in this regard.

To make the research comprehensible and as free from arbitrariness as possible, we described the methods precisely. Therefore, we followed the standards by O'Brien et al [17] and checked the manuscript against the 32-items of the COREQ checklist [18]. Thereby, the strengths and weaknesses of our study became very transparent.

In Germany, the use of mHealth apps in the context of diseases is not yet very common [31]. This also became obvious in our study. Only 16% (5/32) of the participants in our study reported ≥1 prescription of an mHealth app. Thus, most participants could not contribute with experience in this field. Nevertheless, we included opinions, fears, and concerns of nonusers in our research. These should be taken into account when developing and integrating mHealth apps in health care.

A limitation of the article, but not of the study or the results, is that the patient statements were originally made in German and subsequently translated into English. We have taken care not to change the meaning of the statements. The original citations can be requested from the authors of the study.

Conclusions

Problems and barriers in the context of mHealth apps should be considered to guarantee their sustainable, safe, and effective use. Such problems can be categorized into problems originating from the app, problems with the integration of the mHealth app into the health care system, and problems and barriers on the users' side. While problems on the level of the app and the health care system should be taken into account when developing mHealth apps and corresponding assessment tools, problems on the patients' side should be solved by increasing the digital literacy of potential users. On the basis of our findings, further research should be conducted to generate more evidence on problems and barriers as well as how to counteract them.

Acknowledgments

This study is part of a larger research project (Continuous quality assurance of Digital Health Applications [DiGA] ["QuaSiApps"]). The project is funded by the German Federal Joint Committee. The funders had no influence on the study design, the conduct of the study, or the decision to publish or prepare the manuscript. The authors acknowledge the support by the Open Access Publication Fund of the University of Duisburg-Essen.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. In addition to the English patient statements included in the text, the original statements made in German can be requested from the authors.

Authors' Contributions

All authors participated in the conception of the study. Data collection was performed by GDG, CA, KB, and NB. GDG and CA conducted data analysis and interpreted the data together with FP and NB. The manuscript was drafted by GDG with the aid of NB. Feedback from the entire consortium was incorporated.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guidelines.
[DOCX File , 15 KB - jmir_v26i1e49982_app1.docx]

Multimedia Appendix 2
Participant characteristics.
[DOCX File , 16 KB - jmir_v26i1e49982_app2.docx]

Multimedia Appendix 3
The coding system including problem categories and subthemes.
[DOCX File , 16 KB - jmir_v26i1e49982_app3.docx]

Multimedia Appendix 4
Systematized statements of the patients.
[DOCX File , 32 KB - jmir_v26i1e49982_app4.docx]

XSL•FO
RenderX

References

1. mHealth economics 2017/2018 – connectivity in digital health. Research2Guidance. URL: https://research2guidance.com/product/connectivity-in-digital-health/ [accessed 2024-04-10]
2. van Velthoven MH, Wyatt JC, Meinert E, Brindley D, Wells G. How standards and user involvement can improve app quality: a lifecycle approach. Int J Med Inform 2018 Oct;118:54-57 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.08.003] [Medline: 30153922]
3. Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. J Am Med Inform Assoc 2020 Feb 01;27(2):330-340 [FREE Full text] [doi: 10.1093/jamia/ocz175] [Medline: 31599936]
4. CE marking. European Union. URL: https://tinyurl.com/yfk2mjtx [accessed 2024-04-10]
5. Overview of device regulation. U.S. Food & Drug Administration. URL: https://tinyurl.com/464fzbtu [accessed 2023-01-29]
6. The fast-track process for digital health applications (DiGA) according to Section 139e SGB V. A guide for manufacturers, service providers and users. Federal Institute for Drugs and Medical Devices. URL: https://tinyurl.com/k4efvfzx [accessed 2023-01-29]
7. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
8. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
9. Baumel A, Faber K, Mathur N, Kane JM, Muench F. Enlight: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based eHealth interventions. J Med Internet Res 2017 Mar 21;19(3):e82 [FREE Full text] [doi: 10.2196/jmir.7270] [Medline: 28325712]
10. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
11. Nouri R, R Niakan Kalhori S, Ghazisaeedi M, Marchand G, Yasini M. Criteria for assessing the quality of mHealth apps: a systematic review. J Am Med Inform Assoc 2018 Aug 01;25(8):1089-1098 [FREE Full text] [doi: 10.1093/jamia/ocy050] [Medline: 29788283]
12. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study. JMIR Mhealth Uhealth 2019 Apr 16;7(4):e11223 [FREE Full text] [doi: 10.2196/11223] [Medline: 30990458]
13. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
14. Giebel GD, Speckemeier C, Abels C, Plescher F, Börchers K, Wasem J, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808 [FREE Full text] [doi: 10.2196/43808] [Medline: 37171838]
15. Shabir H, D'Costa M, Mohiaddin Z, Moti Z, Rashid H, Sadowska D, et al. The barriers and facilitators to the use of lifestyle apps: a systematic review of qualitative studies. Eur J Investig Health Psychol Educ 2022 Jan 27;12(2):144-165 [FREE Full text] [doi: 10.3390/ejihpe12020012] [Medline: 35200235]
16. Byambasuren O, Beller E, Hoffmann T, Glasziou P. Barriers to and facilitators of the prescription of mHealth apps in Australian general practice: qualitative study. JMIR Mhealth Uhealth 2020 Jul 30;8(7):e17447 [FREE Full text] [doi: 10.2196/17447] [Medline: 32729839]
17. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
18. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
19. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken. Weinheim, Germany: Beltz Publishing; Feb 2, 2015.
20. Azad-Khaneghah P, Neubauer N, Miguel Cruz A, Liu L. Mobile health app usability and quality rating scales: a systematic review. Disabil Rehabil Assist Technol 2021 Oct 08;16(7):712-721 [FREE Full text] [doi: 10.1080/17483107.2019.1701103] [Medline: 31910687]
21. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open 2021 Mar 19;11(3):e047001 [FREE Full text] [doi: 10.1136/bmjopen-2020-047001] [Medline: 33741674]
22. An Q, Kelley MM, Hanners A, Yen PY. Sustainable development for mobile health apps using the human-centered design process. JMIR Form Res 2023 Aug 25;7:e45694 [FREE Full text] [doi: 10.2196/45694] [Medline: 37624639]

XSL•FO
RenderX

23. Clemensen J, Larsen SB, Kyng M, Kirkevold M. Participatory design in health sciences: using cooperative experimental methods in developing health services and computer technology. Qual Health Res 2007 Jan;17(1):122-130 [FREE Full text] [doi: 10.1177/1049732306293664] [Medline: 17170250]
24. Newton A, Bagnell A, Rosychuk R, Duguay J, Wozney L, Huguet A, et al. A mobile phone-based app for use during cognitive behavioral therapy for adolescents with anxiety (MindClimb): user-centered design and usability study. JMIR Mhealth Uhealth 2020 Dec 08;8(12):e18439 [FREE Full text] [doi: 10.2196/18439] [Medline: 33289671]
25. Biswas M, Tania MH, Kaiser MS, Kabir R, Mahmud M, Kemal AA. ACCU3RATE: a mobile health application rating scale based on user reviews. PLoS One 2021 Dec 16;16(12):e0258050 [FREE Full text] [doi: 10.1371/journal.pone.0258050] [Medline: 34914718]
26. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians' attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021 Nov 23;9(11):e33012 [FREE Full text] [doi: 10.2196/33012] [Medline: 34817385]
27. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 1989 Aug;35(8):982-1003 [FREE Full text] [doi: 10.1287/mnsc.35.8.982]
28. Lee M, Kang D, Yoon J, Shim S, Kim IR, Oh D, et al. The difference in knowledge and attitudes of using mobile health applications between actual user and non-user among adults aged 50 and older. PLoS One 2020;15(10):e0241350 [FREE Full text] [doi: 10.1371/journal.pone.0241350] [Medline: 33108792]
29. Pang NQ, Lau J, Fong SY, Wong CY, Tan KK. Telemedicine acceptance among older adult patients with cancer: scoping review. J Med Internet Res 2022 Mar 29;24(3):e28724 [FREE Full text] [doi: 10.2196/28724] [Medline: 35348462]
30. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review protocol. BMJ Open 2020 Mar 16;10(3):e033870 [FREE Full text] [doi: 10.1136/bmjopen-2019-033870] [Medline: 32184309]
31. Bericht des GKV-spitzenverbandes über die inanspruchnahme und entwicklung der versorgung mit digitalen gesundheitsanwendungen. Deutscher Bundestag. URL: https://tinyurl.com/3h6bub6b [accessed 2023-03-26]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DiGA: digital health applications
mHealth: mobile health

Edited by T Leung; submitted 15.06.23; peer-reviewed by R Sun, R Eckhoff; comments to author 11.10.23; revised version received 24.10.23; accepted 31.01.24; published 23.04.24.

Please cite as:
Giebel GD, Abels C, Plescher F, Speckemeier C, Schrader NF, Börchers K, Wasem J, Neusser S, Blase N
Problems and Barriers Related to the Use of mHealth Apps From the Perspective of Patients: Focus Group and Interview Study
J Med Internet Res 2024;26:e49982
URL: https://www.jmir.org/2024/1/e49982
doi:10.2196/49982
PMID:38652508

©Godwin Denk Giebel, Carina Abels, Felix Plescher, Christian Speckemeier, Nils Frederik Schrader, Kirstin Börchers, Jürgen Wasem, Silke Neusser, Nikola Blase. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Smartphone-Based Speech Therapy for Poststroke Dysarthria: Pilot Randomized Controlled Trial Evaluating Efficacy and Feasibility

Yuyoung Kim[1], MSc; Minjung Kim[1,2], MS; Jinwoo Kim[1,2], PhD; Tae-Jin Song[3], MD, PhD

[1]Human Computer Interaction Lab, Graduate Program in Cognitive Science, Yonsei University, Seoul, Republic of Korea

[2]HAII Corporation, Seoul, Republic of Korea

[3]Department of Neurology, Seoul Hospital, Ewha Womans University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Tae-Jin Song, MD, PhD
Department of Neurology
Seoul Hospital
Ewha Womans University College of Medicine
22, Ewhayeodae 1an-gil, Seodaemun-gu
Seoul, 03766
Republic of Korea
Phone: 82 10 8919 8764
Email: knstar@ewha.ac.kr

Abstract

Background: Dysarthria is a common poststroke speech disorder affecting communication and psychological well-being. Traditional speech therapy is effective but often poses challenges in terms of accessibility and patient adherence. Emerging smartphone-based therapies may offer promising alternatives for the treatment of poststroke dysarthria.

Objective: This study aimed to assess the efficacy and feasibility of smartphone-based speech therapy for improving speech intelligibility in patients with acute and early subacute poststroke dysarthria. This study also explored the impact of the intervention on psychological well-being, user experience, and overall feasibility in a clinical setting.

Methods: Participants were divided into 2 groups for this randomized, evaluator-blinded trial. The intervention group used a smartphone-based speech therapy app for 1 hour per day, 5 days per week, for 4 weeks, with guideline-based standard stroke care. The control group received standard guideline-based stroke care and rehabilitation. Speech intelligibility, psychological well-being, quality of life, and user acceptance were assessed using repeated measures ANOVA.

Results: In this study, 40 patients with poststroke dysarthria were enrolled, 32 of whom completed the trial (16 in each group). The intervention group showed significant improvements in speech intelligibility compared with the control group. This was evidenced by improvements from baseline ($F_{1,30}$=34.35; P<.001), between-group differences ($F_{1,30}$=6.18; P=.02), and notable time-by-group interactions ($F_{1,30}$=6.91; P=.01). Regarding secondary outcomes, the intervention led to improvements in the percentage of correct consonants over time ($F_{1,30}$=5.57; P=.03). In addition, significant reductions were noted in the severity of dysarthria in the intervention group over time ($F_{1,30}$=21.18; P<.001), with a pronounced group effect ($F_{1,30}$=5.52; P=.03) and time-by-group interaction ($F_{1,30}$=5.29; P=.03). Regarding quality of life, significant improvements were observed as measured by the EQ-5D-3L questionnaire ($F_{1,30}$=13.25; P<.001) and EQ-VAS ($F_{1,30}$=7.74; P=.009) over time. The adherence rate to the smartphone-based app was 64%, with over half of the participants completing all the sessions. The usability of the app was rated high (system usability score 80.78). In addition, the intervention group reported increased self-efficacy in using the app compared with the control group ($F_{1,30}$=10.81; P=.003).

Conclusions: The smartphone-based speech therapy app significantly improved speech intelligibility, articulation, and quality of life in patients with poststroke dysarthria. These findings indicate that smartphone-based speech therapy can be a useful assistant device in the management of poststroke dysarthria, particularly in the acute and early subacute stroke stages.

Trial Registration: ClinicalTrials.gov NCT05146765; https://clinicaltrials.gov/ct2/show/NCT05146765

(J Med Internet Res 2024;26:e56417) doi:10.2196/56417

XSL•FO
RenderX

KEYWORDS

dysarthria; stroke; smartphone; speech therapy; app; acute and early subacute; feasibility; mobile phone

Introduction

Stroke is a leading cause of mortality and morbidity worldwide [1]. Approximately 40% of people who had survived a stroke experience disabilities [2,3], and over half of the patients with acute stroke develop motor speech disorders, particularly dysarthria [4]. Poststroke dysarthria results from weakened, slow, or impaired speech production muscles caused by cranial nerve damage [5]. Poststroke dysarthria can cause abnormalities in vocal quality, pace, strength, and volume, ultimately leading to reduced speech intelligibility. Consequently, decreased speech intelligibility may trigger communication problems, impaired social interactions, anxiety, depression, and decreased quality of life [6,7].

Starting speech therapy immediately after a stroke can enhance recovery [8-10]. Evidence indicates that early, consistent, intensive treatment yields significantly better outcomes [11,12]. However, despite the recognized importance of early intervention, there is a notable lack of clinical studies that specifically address poststroke dysarthria, particularly in the early stages of stroke. The lack of evidence underscores the need for further studies. In animal studies, neuroplastic changes after an ischemic stroke have been shown to aid neural recovery. However, the direct applicability of these findings in human patients remains uncertain [13,14]. Therefore, further research is needed to define the benefits and risks of early interventions after stroke [10].

Unfortunately, treatment adherence is negatively affected by the perception that current speech treatments are tedious and repetitive [15]. Furthermore, patients may face restrictions regarding therapeutic resources because speech therapy requires substantial time and effort by clinicians or speech-language pathologists (SLPs) [16]. Approximately one-third of the patients received sufficient speech therapy. Additionally, the amount and frequency of therapy received varies among patients [17].

Digital speech therapy apps may offer significant advancements over traditional approaches [18,19]. They also enhance therapeutic accessibility and patient engagement. Additionally, they deliver effective therapeutic dosages and offer tailored feedback to patients [6]. Most importantly, smartphone-based speech therapy apps offer flexibility and ease of access. This is particularly beneficial for patients with stroke who find clinic visits challenging. In addition, smartphone-based speech therapy apps can reduce time and economic burden [20]. Smartphone-based speech therapy can play a crucial role in increasing therapy intensity. High-intensity practice leads to better outcomes in poststroke dysarthria treatment [5,21]. Smartphone-based speech therapy can be delivered using multimedia resources. This approach enhances patient engagement through repetitive practice. Finally, smartphone-based speech therapy enables patients to practice speech independently by measuring various vocal parameters and providing tailored feedback [22]. Real-time feedback can assist patients in recognizing and correcting inappropriate speech patterns [23,24]. This system can enhance the effectiveness of speech therapy by providing valuable insights and improving motivation. Additionally, such feedback is crucial to enhance patient self-efficacy and promote positive behavioral changes [25].

Our primary aim was to evaluate the effect of smartphone-based speech therapy on speech intelligibility, particularly in patients with poststroke dysarthria in the acute and early subacute stroke stages. Additionally, we focused on articulation function, dysarthria severity, and psychological well-being, including depression, anxiety, and quality of life. This study also assessed the feasibility of the trial by examining the adherence, recruitment, and dropout rates. Furthermore, we evaluated the usability and self-efficacy of the app experienced by the participants. This study aimed to explore the efficacy of early intervention and assess how digital tools can enhance speech therapy outcomes in patients with poststroke dysarthria.

Methods

Study Design

This was a prospective, randomized, evaluator-blinded trial study. Participants were allocated to intervention and control groups. They were recruited from 2 stroke centers in South Korea: Ewha Womans University Seoul Hospital and Mokdong Hospital. The trial was registered at ClinicalTrials.gov (NCT05146765).

The participants were screened for eligibility and randomly allocated to the intervention or control groups. Demographic and clinical characteristics were recorded, and a detailed baseline assessment of poststroke dysarthria was conducted. After 4 weeks, the participants underwent a postevaluation to reassess their condition and measure the efficacy of the intervention. The trial was designed according to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Multimedia Appendix 1).

Ethical Considerations

In adherence to our commitment to ethical research standards, we observed several vital considerations throughout this study. Our adherence to these ethical principles was fundamental to ensuring all participants' dignity, rights, safety, and well-being. Upon receiving ethics approval from the Ewha Womans University Seoul Hospital Institutional Review Board (approval SEUMC 2021-12-011), we ensured that all research procedures strictly adhered to the guidelines outlined in the Declaration of Helsinki [26]. Before participating, participants identified as neurologically stable and survived a stroke in the acute and early subacute stages received detailed information about the study's goals, procedures, and potential benefits and risks. Each participant provided written informed consent to affirm their voluntary participation and understanding of the study. This consent process was necessary to ensure participants were fully

XSL•FO
RenderX

informed and their autonomy respected. Next, strict data protection measures were implemented to safeguard our participants' privacy and confidentiality. All collected data were anonymized throughout the research process to preserve participants' privacy. We offered participants a monetary compensation of ₩50,000 (US $40) for their involvement in the study, which amounts to ₩25,000 (US $20) per visit. This compensation was offered as a token of appreciation for their valuable contribution to our research and to acknowledge the personal investment each participant made by dedicating their time to our study.

Participants

A principal investigator (TJS), specializing in stroke, screened and enrolled the eligible participants. The inclusion criteria were as follows: (1) patients diagnosed with dysarthria by a stroke specialist [27], (2) patients who are neurologically stable, (3) first-time patients with stroke, (4) patients who are in the acute or early subacute phase of stroke defined as having experienced their initial stroke event within the past 1 month, (5) patients with sufficient cognitive abilities to use a smartphone-based speech therapy app (Mini-Mental State Examination score ≥26) [28], and (6) patients with adequate vision [29], hearing [30], communication skills, and motor skills [27]. The exclusion criteria were as follows: (1) coexisting language disorders (eg, aphasia) or neurological disorders (eg, dementia, Pick disease, Huntington disease, Parkinson disease, or Parkinsonism) that could influence dysarthria, (2) history of severe mental disorders (eg, depression, schizophrenia, alcohol addiction, or drug addiction), (3) inability to use or access smartphone technology, (4) illiteracy, and (5) inability to communicate in Korean, the primary language of the study location.

Randomization and Masking

An independent researcher managed the randomization. A computerized system with permuted block sizes of 2 and 4 was used to ensure a balanced and unpredictable group distribution [31]. The block sizes were disclosed to the participants or researchers at the end of the trial to ensure randomization.

Given the intervention's interactive nature, it was impossible to blind the participants to their group assignments [32]. However, independent evaluators and those not involved in the treatment process were blinded to the group allocation to minimize potential bias. This masking was crucial to maintain the integrity of the assessment. To preserve the integrity of the blinded assessment, participants were instructed not to disclose any intervention-related details during the evaluation sessions.

Intervention

Intervention Group

Participants in the intervention group received a smartphone-based speech therapy app and standard guideline-based stroke care. This app allowed participants to achieve speech therapy independently without the support of caregivers or therapists. The participants were instructed to use the app for 1 hour daily for at least 5 days per week for 4 consecutive weeks. The intervention could be completed in a single session or distributed across multiple daily sessions.

The app was tailored for older adults with poststroke dysarthria and optimized for users facing age-related challenges [33]. The interface was designed to minimize unintentional interactions for participants with motor impairments. Intuitive design elements, such as sequential tabs and text-labeled buttons, were included to enhance usability for older adults [34]. Moreover, button size and spacing were adjusted to facilitate ease of use and reduce inadvertent presses.

The app provided 6 components of speech exercises for 1 day based on established behavioral therapies [5,10]. These included oro-motor exercises, sustained sound, pitch variation, velopharyngeal closure, reading practice, and syllable repetition (Table 1). The primary goal of these exercises was to improve overall speech intelligibility and enhance articulation.

Speech exercises such as sustaining sounds, repeating syllables, and reading provided real-time auditory and visual feedback. Real-time feedback was provided throughout the sessions to promote attention and self-awareness during speech therapy. Pronunciations and speech signals were transmitted during speech exercises. Our engine analyzed the speech parameters and provided feedback. The feedback results were displayed on the participants' devices. For instance, the "sustaining sound" task required participants to sustain a vowel sound, such as /ah/, for 5-15 seconds. Subsequently, real-time feedback on the loudness, sound length, and pitch was provided. Participants could address speech errors through insights gained from the feedback (eg, "Speak more loudly!" in Figure 1B).

The treatment results are presented in 2 formats as shown in Figure 2. First, a summary of each therapy session focused on speech outcomes, including pronunciation accuracy, loudness, and pitch. The participants understood these outcomes better through voice- and text-guided interactions. After the exercise, they listened to their recorded voices and provided feedback. This feedback helped them assess their progress (Figure 2A). Second, the app provided cumulative analysis. The analysis included the daily treatment results, weekly and monthly progress, and speech outcome scores (Figure 2B).

The app automatically logged all the results in a database. The researchers could access these results using a data-logging system. Researchers monitored the participants' adherence to the intervention and offered coaching for lapses or technical issues. The researchers evaluated the app use every evening to monitor participants' adherence. If reduced adherence was observed, the researchers contacted the participants the following day via phone call or SMS text message to encourage therapeutic engagement. The participants were encouraged to report any app-related issues, which were promptly addressed by the researchers.

XSL•FO
RenderX

Table 1. Procedure for each speech exercise.

Exercise	Procedure	Outcome monitoring
Oro-motor exercise	Oro-motor exercises were designed to relax oral muscles and actively engage them. These exercises improved muscle coordination and strength. A 5-minute instructional video guided users through specific exercises targeting the lips, tongue, cheeks, and jaw. Participants could use a camera to mirror these actions for real-time comparison with the guide in the video.	Compliance rate
Sustaining sound	Sustaining a vowel sound was designed to improve vocal control and strength. Participants practiced holding a vowel sound clearly and loudly for 5 to 15 seconds. They received real-time visual feedback on pitch and duration to help maintain a steady pitch and extend the sound. This approach promoted effective and precise vocal training.	Loudness, pitch, and length
Pitch variation	Pitch variation exercises improved speech intonation and expressiveness. Participants practiced raising and lowering vocal pitch with “do-re-mi” exercises. The app provided visual feedback for these pitch changes. This feedback allowed participants to adjust their pitch with precision in real time. Participants could enhance the efficiency and effectiveness of the exercise by ensuring they achieve the correct pitch.	Length, intonation, and pitch
Velopharyngeal closure	The velopharyngeal closure exercise focused on improving clear speech articulation. This exercise involved practicing words that require the closure of the velopharynx. Participants started by blocking the velopharynx at the beginning of each word. They held this block for 5 seconds to build muscle strength and then released the breath on the following syllable. The app supported this process by showing anatomical animations. These animations aided users in comprehending muscle tension and the timing for its application and release. Emphasis was placed on maintaining a 5-second hold and correctly timing the exhalation.	Loudness, pronunciation accuracy, and breath-holding time
Reading	Reading exercises were structured to improve articulation accuracy and speech intelligibility. They began with single words and then progressed to sentences and paragraphs. This repeated practice enhanced speech precision. Participants were encouraged to read aloud slowly and clearly. The app initially guided pronunciations to demonstrate how to read. It also provided visual cues for participants to slowly read along. Participants practiced reading with the provided material at a steady pace.	Loudness, rhythmic accuracy, and speech rate
Repeating syllables	The repeating syllables exercise was designed to improve articulation and control over speech rate. Participants were instructed to repeat specific syllables, such as /pa/, in sync with a given rhythm. The exercise started slowly and gradually sped up the repetition speed. Visual rhythm cues were provided to help participants maintain the appropriate speed. This method enhanced speech fluidity, precision, articulation, and rhythm control.	Loudness, pronunciation accuracy, and rhythmic accuracy

Figure 1. Two components of the smartphone-based speech therapy app. (A) Oro-motor exercise videos for specific muscle stretches with camera mirroring. (B) Speech exercises (sustaining sound) with real-time feedback to improve articulation.

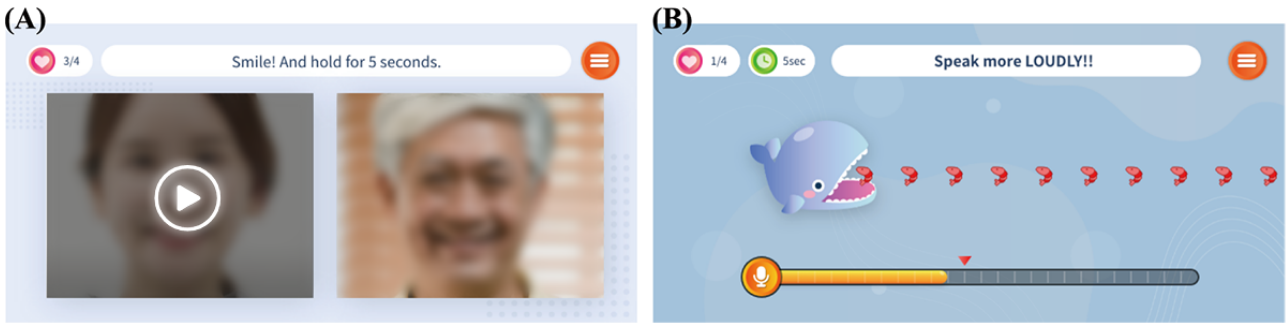

XSL•FO
RenderX

Figure 2. Treatment results in the smartphone-based speech therapy app. (A) Session summary with scores based on speech exercise performance. (B) Cumulative analysis presenting daily and weekly or monthly speech outcome scores.

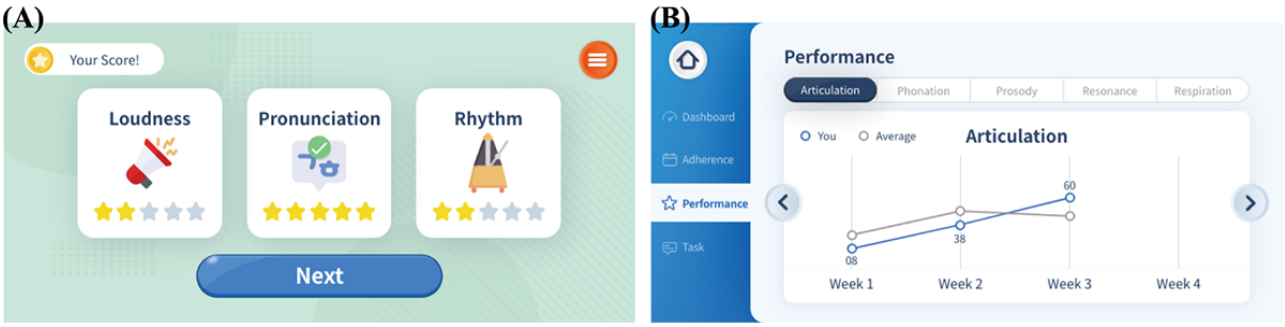

Control Group

Participants in the control group received standard stroke care for 4 weeks. Standard stroke care includes medical treatment, routine stroke therapy, and rehabilitative exercises, as outlined in the basic guidelines [35,36]. This care encompassed the conventional speech treatment recommended in standard protocols, such as vocal and articulation exercises. Care was provided by clinicians and SLPs who adhered to the conventional stroke therapy methods. Treatment was tailored to each participant's clinical needs, established through a collaborative agreement between clinicians and participants, and modified to reflect their progress. Additionally, after the 4-week study period, participants in the control group were allowed to use a smartphone-based speech therapy app.

Outcome Measures

Assessments were conducted at 2 time points: at baseline and then immediately after the 4-week intervention period.

Primary Outcome: Speech Intelligibility

The primary outcome of this study was a change in speech intelligibility. To evaluate speech intelligibility, participants were asked to read the "Gaeul" passage, a standardized tool used in Korean paragraph reading tests for speakers with motor speech disorders, developed by Kim [37]. This passage consists of 369 syllables representing the frequency of occurrence of Korean vowels and consonants.

Participants were instructed to read the passage aloud at their natural pitch and loudness. Recordings were made using a high-quality digital recorder (Sony ICD-UX560F) positioned 30 cm from the participants in a quiet room. The evaluation was carried out in an environment free from noise, which ensured that the conditions were consistent for every assessment [38]. Participants were seated close to the evaluator to ensure optimal sound quality. The primary SLP evaluator conducted the assessment in the room during the recording. Subsequently, experienced SLPs, who were blinded to the participants' group allocation, listened to each recording and assessed the speech intelligibility. All 3 SLPs who conducted the assessment possessed over 6 years of clinical experience, specialized in poststroke dysarthria, and held certifications in Korean speech-language pathology. Additionally, they had experienced specialized training in poststroke dysarthria. Speech intelligibility was rated on a scale ranging from 0 (intelligible, can be understood without difficulty) to 6 (unintelligible, cannot be understood at all) [39]. The other 2 evaluators assessed speech intelligibility based on the recorded audio. The average score from the 3 SLPs was used to determine each participant's final speech intelligibility score.

Secondary Outcomes

Secondary outcomes were measured to assess factors related to dysarthria and psychological well-being. First, the Urimal Test of Articulation and Phonology 2 (U-TAP2) was used [40]. This measurement was used to identify the percentage of consonants correct for detecting articulation anomalies [41]. Participants were asked to read 30 words from U-TAP2 in a quiet room. The SLPs then recorded these readings and calculated the percentage of consonants correct by marking misarticulated consonants (94 in total) and converting them into a percentage score.

Stroke-related neurological deficits were measured using the National Institute of Health Stroke Scale [27], with a specific focus on components related to dysarthria. Stroke specialists quickly evaluated the severity of dysarthric speech. As the participants spoke specific words, the severity was rated on a 3-point scale: 0=normal, 1=mild to moderate, and 2=severe. This assessment was conducted by a seasoned neurologist with over 20 years of experience in stroke specialization and certified in the Korean National Institute of Health Stroke Scale.

Finally, participants' psychological well-being was measured using self-reported questionnaires. The Patient Health Questionnaire-9 [42,43] and the Generalized Anxiety Disorder 7-Item Scale [44,45] were used to evaluate depressive and anxiety symptoms. Furthermore, the EQ-5D-3L questionnaire [46] was used to assess the participants' quality of life across 5 different areas: their ability to move around, care for themselves, perform their usual activities, levels of pain or discomfort, and mood. To assign specific values to these quality-of-life measures, we applied weights based on the preferences of the South Korean population. These weights were calculated using the time trade-off method and scores from a visual analog scale [47].

Feasibility and User Acceptance

Feasibility was assessed based on several aspects. The participant recruitment rates were documented to reflect the level of engagement. Adherence to the intervention was evaluated by tracking the completion rates of the prescribed speech therapy sessions within the app, the frequency of app use, and the average duration of each session. These data, which were transmitted to a dedicated web system, allowed for a

XSL•FO
RenderX

detailed analysis of adherence. Potential adverse events and safety concerns were continuously monitored. Any reported issues with app use or challenges faced by the participants were investigated by analyzing the app's log data.

The usability and acceptance of smartphone-based speech therapy apps were measured using 2 surveys: the System Usability Scale (SUS) [48] and the Modified Computer Self-Efficacy Scale (mCSES) [49]. The usability of the app was evaluated using a 10-item, 5-point Likert scale that measured effectiveness, efficiency, and satisfaction. The mCSES was used to gauge participants' confidence in using the new technology, especially tailored for older patients and those with disabilities.

Statistical Analysis

Power analysis focused on measuring the changes in speech intelligibility. We initially calculated that 32 participants were required to achieve 80% power [50] to detect a moderate effect size of 0.29 [51] with a significance level set at .02. However, we aimed to enroll 8 more participants to account for an anticipated dropout rate of 25%. Therefore, our goal was to recruit 40 participants with 20 participants per group [52].

Descriptive statistics (mean, SD, and percentage) were used to summarize the clinical and demographic characteristics of the participants. To ensure homogeneity between the intervention and control groups, a 2-tailed independent sample *t* test was conducted for continuous variables, whereas a chi-square test was used for categorical variables. Following the intention-to-treat principle, repeated measures ANOVA was applied to detect changes in outcome measures between and within groups. This analysis incorporated fixed effects for time, group, and time-by-group interactions, with measures taken at baseline and 4 weeks after the intervention. All analyses were performed using SPSS (version 27.0; IBM Corp). Statistical significance was set at $P<.05$ and was considered statistically significant.

Data Management

All data were encrypted to ensure privacy. After encryption, the system was securely transmitted to a dedicated web system. This process maintained the confidentiality and safety of the data. Real-time data such as app use frequency, session duration, and speech performance metrics are necessary for monitoring therapeutic progress and adapting the intervention as needed. Our research team used proactive measures to ensure consistent participation. For instance, reduced adherence to the app triggered alerts, which prompted our team to reconnect with the participants to understand and address their concerns. While participants could withdraw from the study at any time, the research team reserved the right to exclude those who required immediate medical attention for reasons that were not limited to the study parameters.

Results

Participants

We recruited 129 patients with acute to early subacute cerebral infarction between January 18, 2022, and May 31, 2022. These patients were screened based on the eligibility criteria. Of these, 81 patients exhibited symptoms of dysarthria. During the screening process, 14 patients were excluded due to coexisting aphasia, 10 due to psychological problems or medication, 11 due to dementia or cognitive dysfunction, 3 due to inability to use or access smartphone technology, and 3 due to visual or hearing impairment. Finally, 40 participants were enrolled.

The 40 participants were randomized into 2 study groups, as shown in Figure 3. We excluded 7 participants who could not complete the study for personal reasons: 5 in the intervention group and 2 in the control group. Additionally, 1 participant in the control group was excluded because of another speech disorder, apraxia. The final analysis included 32 participants (16 each in the treatment and control groups).

Table 2 presents the baseline characteristics of the participants. Chi-square and independent 2-tailed *t* tests revealed no significant differences between the 2 study groups. Among the 32 participants, 25 were male and 7 were female, with a mean age of 65.25 (SD 12.97; treatment group: mean 60.44, SD 11.94 and control group: mean 70.06, SD 12.47) years. All the participants were in the acute and early subacute phases of poststroke dysarthria. The treatment group participants were observed for an average of 7.06 (SD 3.66) days after stroke. In contrast, the control group participants were assessed on an average of 7.88 (SD 6.45) days after stroke.

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

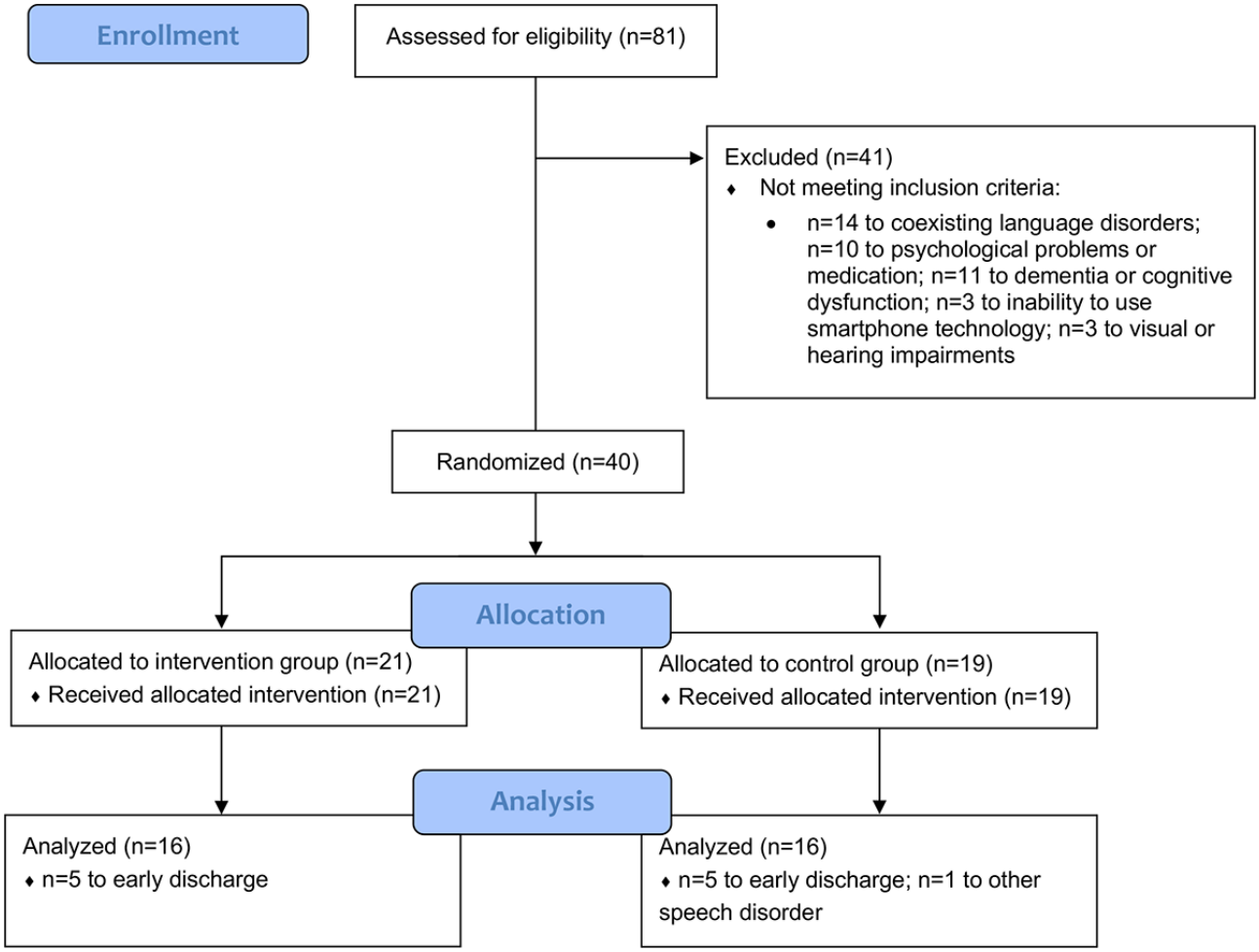

XSL•FO
RenderX

Table 2. Baseline demographics and clinical characteristics of participants.

Participants characteristics	Intervention (n=16)	Control (n=16)	Total (n=32)	Chi-square (*df*)	*P* value
Sex, n (%)				1.6 (1)	.20
Male	14 (44)	11 (34)	25 (78)		
Female	2 (6)	5 (16)	7 (22)		
Age (years), mean (SD)	60.44 (11.94)	70.06 (12.47)	65.25 (12.97)	26.0 (21)	.21
Type of stroke, n (%)					
Cerebral infarction	16 (50)	15 (47)	31 (97)	N/A[a]	N/A
Transient ischemic attack	0 (0)	1 (3)	1 (3)	N/A	N/A
Duration after stroke (days), mean (SD)	7.06 (3.66)	7.88 (6.45)	7.47 (5.17)	8.7 (12)	.73
Speech ability, mean (SD)					
Speech intelligibility	1.56 (0.89)	2.31 (1.30)	1.94 (1.16)	4.3 (4)	.37
U-TAP2[b] (%)	95.67 (3.33)	91.41 (7.34)	93.54 (6.01)	9.4 (12)	.47
NIHSS[c] dysarthria	1.00 (0.00)	1.06 (0.25)	1.03 (0.18)	1.0 (1)	.31
Psychological well-being, mean (SD)					
PHQ-9[d]	5.88 (5.26)	12.13 (7.07)	9.00 (6.90)	18.0 (17)	.39
GAD-7[e]	5.25 (5.25)	7.69 (6.89)	6.47 (6.15)	11.5 (12)	.48
EQ-5D-3L	0.73 (0.28)	0.57 (0.30)	0.65 (0.30)	19.6 (18)	.36
EQ-VAS	69.88 (15.24)	52.19 (28.28)	61.03 (24.09)	19.5 (11)	.04
Feasibility, mean (SD)					
mCSES[f]	66.50 (33.28)	40.56 (26.48)	53.53 (32.38)	26.0 (21)	.21

[a]N/A: not applicable.
[b]U-TAP2: Urimal Test of Articulation and Phonology 2.
[c]NIHSS: National Institute of Health Stroke Scale.
[d]PHQ-9: Patient Health Questionnaire-9.
[e]GAD-7: Generalized Anxiety Disorder 7-Item Scale.
[f]mCSES: Modified Computer Self-Efficacy Scale.

Primary Outcome

During the baseline assessment, none of the participants were rated as 0=completely understandable or 6=completely unintelligible. Of the total 32 participants, 16 had a rating of 1, indicating slight difficulties in speech intelligibility. Another 8 participants had a rating of 2, demonstrating mild dysarthria. A range of speech intelligibility issues was observed: 5 participants had a rating of 3, which indicated moderate dysarthria; and 2 participants had a rating of 4, which suggested more severe difficulties. Only 1 participant had a rating of 5, which indicated they were close to being unintelligible.

Repeated measures ANOVA was conducted to assess the impact of time, group, and time-by-group interactions on speech intelligibility. The results revealed a significant effect of time ($F_{1,30}$=34.35; P<.001). This finding indicated that there were significant changes in speech intelligibility between baseline and 4 weeks after the intervention. The mean speech intelligibility score in the intervention group improved from 1.56 (SD 0.89) at baseline to 0.69 (SD 1.09) after intervention. Additionally, a significant group effect was observed ($F_{1,30}$=6.18; P=.02). This analysis suggested significant differences in speech intelligibility between the treatment and control groups. Furthermore, the interaction effect between time and group was also significant ($F_{1,30}$=6.91; P=.01), which indicates that the changes in speech intelligibility over time varied significantly between the groups.

Secondary Outcomes

The intervention group demonstrated notable improvements in secondary outcomes compared with the control group after intervention (Table 3).

XSL•FO
RenderX

Table 3. Repeated measures ANOVA of outcome scores.

Variables	Outcome scores, mean (SD)		Time		Group		Time×group	
	Intervention	Control	*F* test (*df*=1)	*P* value	*F* test (*df*=1)	*P* value	*F* test (*df*=1)	*P* value
Speech intelligibility			34.35	<.001	6.18	.02	6.91	.01
Baseline	1.56 (0.89)	2.31 (1.30)						
Postintervention	0.69 (1.09)	1.98 (1.43)						
U-TAP2[a] (%)			5.57	.03	3.52	.07	4.13	.05
Baseline	95.67 (3.33)	91.41 (7.34)						
Postintervention	95.88 (2.93)	94.22 (4.38)						
NIHSS[b] dysarthria			21.18	<.001	5.52	.03	5.29	.03
Baseline	1.00 (0.00)	1.06 (0.25)						
Postintervention	0.44 (0.51)	0.87 (0.50)						
PHQ-9[c]			1.42	.24	8.33	.007	0.66	.42
Baseline	5.88 (5.26)	12.13 (7.07)						
Postintervention	5.38 (4.10)	9.50 (7.98)						
GAD-7[d]			2.09	.16	2.15	.15	0.13	.91
Baseline	5.25 (5.25)	7.69 (6.87)						
Postintervention	3.56 (4.87)	6.25 (6.05)						
EQ-5D-3L			13.25	<.001	3.64	.07	0.76	.79
Baseline	0.73 (0.28)	0.57 (0.30)						
Postintervention	0.87 (0.12)	0.74 (0.25)						
EQ-VAS			7.74	.009	6.06	.02	0.15	.70
Baseline	69.88 (15.24)	52.19 (28.28)						
Postintervention	79.06 (14.17)	64.31 (25.10)						
mCSES[e]			2.99	.09	10.81	.003	0.97	.33
Baseline	66.50 (33.28)	40.56 (26.48)						
Postintervention	77.88 (25.01)	43.69 (28.38)						

[a]U-TAP2: Urimal Test of Articulation and Phonology 2.
[b]NIHSS: National Institute of Health Stroke Scale.
[c]PHQ-9: Patient Health Questionnaire-9.
[d]GAD-7: Generalized Anxiety Disorder 7-Item Scale.
[e]mCSES: Modified Computer Self-Efficacy Scale.

First, the percentage of correct consonants measured by the U-TAP2 showed a significant time effect ($F_{1,30}$=5.57; P=.03) compared to the change between baseline and 4 weeks after the intervention. However, the group effect ($F_{1,30}$=3.52; P=.07) and time-by-group interaction ($F_{1,30}$=4.13; P=.05) were not statistically significant.

Second, significant findings emerged from the assessment of the severity of poststroke dysarthria. The time effect was significant ($F_{1,30}$=2.21; P≤.001). This highlights a notable improvement in the severity over 4 weeks. Furthermore, a significant group effect ($F_{1,30}$=5.52; P=.03) indicated differences in severity between the treatment and control groups. Most importantly, the significant time-by-group interaction ($F_{1,30}$=5.29; P=.03) suggests that the groups experienced different trajectories of severity over time.

Third, no significant benefits were observed for depression or anxiety. For depression, as measured by the Patient Health Questionnaire-9, there was no significant time effect ($F_{1,30}$=1.42; P=.24), and the time-by-group interaction was also not significant ($F_{1,30}$=0.66; P=.42). However, a significant group effect was observed ($F_{1,30}$=8.33; P=.007). In terms of anxiety levels, as assessed by the Generalized Anxiety Disorder 7-Item Scale, no significant effects were found for time ($F_{1,30}$=2.09; P=.16; group: $F_{1,30}$=2.15; P=.15; or time-by-group interaction: $F_{1,30}$=0.13; P=.91).

XSL•FO
RenderX

Finally, a significant time effect was noted for the overall quality of life measured by the EQ-5D-3L ($F_{1,30}$=13.25; P≤.001). No significant effects were observed for group ($F_{1,30}$=3.64; P=.07) or time-by-group interactions ($F_{1,30}$=0.76; P=.79). In addition, the EQ-VAS scores showed a significant time effect ($F_{1,30}$=7.74; P=.009) and group effect ($F_{1,30}$=6.06; P=.02). However, there was no significant time-by-group interaction ($F_{1,30}$=0.15; P=.70).

Feasibility

We met our recruitment goal by successfully enrolling 40 participants during the study period. The final assessment completion rate was 80%. Regarding adherence, 64% (n=20) of participants in the intervention group consistently used the smartphone-based speech therapy app throughout the designated period. More than 51% (n=16) of the participants completed the prescribed sessions.

System usability was considered excellent, as measured by the mean SUS score of 80.78 (SD 16.27). Concerning self-efficacy, measured by the mCSES, the intervention group had a substantial group effect ($F_{1,30}$=10.81; P=.003), but there were no significant changes over time ($F_{1,30}$=2.99; P=.09) or in the time-by-group interaction ($F_{1,30}$=0.97; P=.33). No significant adverse events were observed during the study period.

Discussion

Principal Findings

Despite its significant impact on communication and psychosocial well-being, poststroke dysarthria remains underresearched. In particular, there is a lack of evidence on poststroke dysarthria interventions, highlighting the urgent need for more comprehensive research [53]. Understanding the prognosis of speech therapy in the critical initial months after stroke is vital because early intervention can hasten recovery [9]. Unfortunately, there is a knowledge gap in the evidence regarding poststroke dysarthria during the acute and early subacute phases [54]. Our trial findings provide evidence of the efficacy of smartphone-based speech therapy in the treatment of poststroke dysarthria.

In this study, participants experienced significant improvements in speech intelligibility and articulation after 4 weeks of using the smartphone-based speech therapy app compared to those receiving standard stroke care. This intervention was effective in several ways. It showed the potential for reducing the severity of dysarthria. It also helped alleviate depression and improve the quality of life of the participants. Consistent with prior studies, these results underscore the reliability of smartphone-based interventions [55,56].

The efficacy of traditional behavioral speech therapy has been proven in the chronic phase; however, studies on patients with acute and early subacute strokes are limited. Prior studies have shown encouraging results for behavioral speech therapy such as breathing exercises, nonspeech oro-motor exercises, and Lee Silverman Voice Treatment for the chronic poststroke phase [57]. One study used the Lee Silverman Voice Treatment that focuses on high phonatory effort and reading exercises [58]. This study showed promising results in a small group of 4 individuals who have survived a stroke with dysarthria for 9 months. According to another study, repetitive speech therapy had a positive effect on patients with stroke for at least 6 months [59].

Our study expands traditional behavioral speech therapy into a digital format using a smartphone-based app [58-61]. This approach overcomes the limitations of traditional methods by offering more accessible, engaging, and cost-effective speech therapy that enables self-management [62-65]. Patients can perform various speech exercises at home. Home-based treatment reduces the need for frequent clinical visits and reduces expenses [66,67]. Moreover, the app provides uninterrupted therapy sessions, even during the COVID-19 pandemic. This serves as a reliable alternative to clinical treatment [68].

Patients with poststroke dysarthria also commonly experience adverse psychological effects [6,7]. Previous studies focusing on speech therapy in participants with poststroke aphasia have demonstrated improvements in depression [69], anxiety [70], and quality of life [71]. However, specific evidence for poststroke dysarthria remains limited. Although we observed a significant decrease in depressive symptoms, no significant changes in anxiety levels were observed. Notably, the EQ-5D-3L and EQ-VAS scores indicated a substantial improvement in quality of life over time and a positive effect of the intervention. However, the lack of significant group differences in these scores suggests that improvements in quality of life were not solely attributable to the intervention. This divergence in findings highlights the complexity of assessing the full effect of speech therapy interventions on psychological well-being. Due to the significant impact of psychological well-being deterioration in patients with poststroke dysarthria, cognitive behavioral therapy should also be considered as a potential treatment [72]. Since this study is primarily focused on speech intelligibility, it may not have fully captured the broader impact of speech therapy on psychological well-being. Given these findings, there is a clear need for further research with larger sample sizes to provide a better understanding of the benefits of speech therapy interventions on the psychological well-being of patients with poststroke dysarthria. This can help develop effective treatment strategies, specifically in the areas of speech and psychological well-being.

Meanwhile, the average SUS score of 80.78 (SD 16.27) signifies excellent usability, which indicates that the participants found the app user-friendly and efficient. Participants also noted increased self-efficacy in app use compared with before treatment. These results suggest that the app helped overcome apprehensions about using the technology, particularly among older users. This increased system feasibility is a promising sign of active participation in therapy.

However, the treatment adherence was lower than expected. Notably, measuring adherence was challenging because of variable internet connectivity among the participants. Due to low-specification phones or unstable home internet connections, many participants, especially older users, experienced frequent internet disconnections. These challenges hindered the proper

XSL•FO
RenderX

storage of log data, which may have led to inaccuracies in adherence measurements. Our app includes features, such as progress graphs and feedback, to address adherence-related issues and encourage self-monitoring [12]. Although these features are standard in health apps and are crucial for self-therapy, they have limited long-term effectiveness [73,74]. This limitation is particularly relevant for older adults who are unfamiliar with digital devices [75,76]. Given these challenges, future research should focus on improving adherence to therapy and making it more accessible to diverse patient groups. Including more subjects and a broader range of variables could enhance our understanding of how digital interventions can be most effectively used in poststroke care. Regarding home therapies, various factors, such as the patient's social context and home environment, can affect the treatment effectiveness. For example, providing an admin system to monitor and control patient performance data is recommended. This would allow clinicians or family caregivers to remotely track adherence and performance and address potential issues arising from the lack of face-to-face interactions. This could help older adults maintain adherence and maximize the therapeutic effects of treatment [77].

Limitations

This study has several limitations. First, even as a pilot trial, this study included a small number of participants. Additionally, there was a gender imbalance with a significantly higher number of male participants. Future studies should aim for larger sample sizes and consider recruitment from multiple centers to improve the feasibility and generalizability.

Second, this study focused only on patients with poststroke dysarthria in the acute and early subacute stages. However, dysarthria affects patients in both the acute and chronic stages of stroke. To validate the effectiveness of the intervention across diverse patient profiles, future research should include a broader range of patients with stroke and consider the onset period and severity of dysarthria. Additionally, this study only recruited patients in the acute and early subacute stages of joint impairment after stroke, which may have resulted in the exclusion of patients with severe joint impairment. These selection criteria may have influenced the observed effects of smartphone-based speech therapy. In future studies, it would be beneficial to include participants with varying degrees of dysarthria to understand better the efficacy of this therapy across a spectrum of severity. A more detailed analysis, which may include secondary assessments, can be carried out to evaluate the therapy's efficacy in addressing speech impairments of varying severity. This approach would enable a deeper understanding of the therapy's applicability to a broader range of dysarthria cases after stroke.

Third, regarding the measurement of consonant accuracy using U-TAP2 at the word level, we recognize that this approach has limitations, particularly in adult poststroke dysarthria. While U-TAP2 is extensively used to assess articulatory precision in Korean children with developmental articulation disorders, its application is limited [40]. When measuring speech intelligibility in adults with poststroke dysarthria, particularly in continuous speech, U-TAP2 may not fully capture all the complexities. This tool needs to be equipped to grasp the full range of speech intelligibility challenges this adult population faces. Specifically, this method may overlook critical aspects of speech, such as rhythm, prosody, and coarticulation effects, which are essential for understanding overall speech severity. The choice of U-TAP2 was influenced by the absence of standardized assessment tools for adult poststroke dysarthria in the Korean clinical environment. However, we acknowledge that future research should explore more comprehensive tools like the Frenchay Dysarthria Assessment to analyze the various influencing factors of dysarthria more thoroughly [78].

Finally, the smartphone-based speech therapy app used in this study was developed in Korean. Future research should aim to create multilingual versions of the app. Studying multilingual versions would enable researchers to assess their effectiveness across different nationalities and broaden their reach.

Conclusions

This study emphasized the importance of digital speech therapy in the treatment of poststroke dysarthria. Smartphone apps designed for speech therapy can be used alongside traditional speech therapies and have shown promising results in improving speech outcomes and the overall quality of life. Our findings provide encouraging evidence for the integration of these apps into existing treatment plans. However, more extensive and comprehensive studies are needed to fully understand the impact of digital speech therapy and optimize its use in treating poststroke dysarthria.

Acknowledgments

This research was supported by the Technology Development Program (S3301230) and funded by the Ministry of SMEs and Startups (Korea). We sincerely thank all the participants for their valuable time and commitment to this study. This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant RS-2023-00262087 to TJS). This research was supported by the Institute of Information & Communications Technology Planning & Evaluation grant funded by the Korean government (MSIT 2022-0-00621) for the development of artificial intelligence technology that provides dialog-based multimodal explainability. This paper presents the original work of the authors and is not under consideration for publication elsewhere.

Authors' Contributions

All authors contributed significantly to this study. YK developed app content conceived, designed the study, analyzed and interpreted the data, and drafted and revised the paper. MK contributed to the app design and development and paper revision.

XSL•FO
RenderX

As the principal investigator, TJS recruited participants and critically reviewed the paper. JK and TJS secured the funding for this study. All authors have reviewed and approved the final version of the paper for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) V 1.6.1 checklist.
[PDF File (Adobe PDF File), 1735 KB - jmir_v26i1e56417_app1.pdf]

References

1. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke 2022;17(1):18-29 [FREE Full text] [doi: 10.1177/17474930211065917] [Medline: 34986727]
2. Hackett ML, Duncan JR, Anderson CS, Broad JB, Bonita R. Health-related quality of life among long-term survivors of stroke: results from the Auckland Stroke Study, 1991-1992. Stroke 2000;31(2):440-447 [FREE Full text] [doi: 10.1161/01.str.31.2.440] [Medline: 10657420]
3. Hankey GJ, Jamrozik K, Broadhurst RJ, Forbes S, Anderson CS. Long-term disability after first-ever stroke and related prognostic factors in the Perth Community Stroke Study, 1989-1990. Stroke 2002;33(4):1034-1040 [FREE Full text] [doi: 10.1161/01.str.0000012515.66889.24] [Medline: 11935057]
4. Flowers HL, Silver FL, Fang J, Rochon E, Martino R. The incidence, co-occurrence, and predictors of dysphagia, dysarthria, and aphasia after first-ever acute ischemic stroke. J Commun Disord 2013;46(3):238-248. [doi: 10.1016/j.jcomdis.2013.04.001] [Medline: 23642855]
5. Duffy JR. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. 4th Edition. St. Louis, MO: Elsevier Mosby; 2020.
6. Dickson S, Barbour RS, Brady M, Clark AM, Paton G. Patients' experiences of disruptions associated with post-stroke dysarthria. Int J Lang Commun Disord 2008;43(2):135-153. [doi: 10.1080/13682820701862228] [Medline: 18283594]
7. Mackenzie C, Kelly S, Paton G, Brady M, Muir M. The Living with Dysarthria group for post-stroke dysarthria: the participant voice. Int J Lang Commun Disord 2013;48(4):402-420. [doi: 10.1111/1460-6984.12017] [Medline: 23889836]
8. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004;10(2):307-312. [doi: 10.1111/j..2002.384.doc.x] [Medline: 15189396]
9. Coleman ER, Moudgal R, Lang K, Hyacinth HI, Awosika OO, Kissela BM, et al. Early rehabilitation after stroke: a narrative review. Curr Atheroscler Rep 2017;19(12):59 [FREE Full text] [doi: 10.1007/s11883-017-0686-6] [Medline: 29116473]
10. Godecke E, Armstrong E, Rai T, Ciccone N, Rose ML, Middleton S, et al. A randomized control trial of intensive aphasia therapy after acute stroke: the Very Early Rehabilitation for SpEech (VERSE) study. Int J Stroke 2021;16(5):556-572 [FREE Full text] [doi: 10.1177/1747493020961926] [Medline: 33019888]
11. Kiran S, Thompson CK. Neuroplasticity of language networks in aphasia: advances, updates, and future challenges. Front Neurol 2019;10:295 [FREE Full text] [doi: 10.3389/fneur.2019.00295] [Medline: 31001187]
12. Robertson SJ, Thomson F. Speech therapy in Parkinson's disease: a study of the efficacy ad long term effects of intensive treatment. Br J Disord Commun 1984;19(3):213-224. [doi: 10.3109/13682828409029837] [Medline: 6508992]
13. Krakauer JW, Carmichael ST, Corbett D, Wittenberg GF. Getting neurorehabilitation right: what can be learned from animal models? Neurorehabil Neural Repair 2012;26(8):923-931 [FREE Full text] [doi: 10.1177/1545968312440745] [Medline: 22466792]
14. Cramer SC. Repairing the human brain after stroke: I. Mechanisms of spontaneous recovery. Ann Neurol 2008;63(3):272-287. [doi: 10.1002/ana.21393] [Medline: 18383072]
15. Tamayo-Serrano P, Garbaya S, Bouakaz S, Blazevic P. A game-based rehabilitation therapy for post-stroke patients: an approach for improving patient motivation and engagement. IEEE Syst Man Cybern Mag 2020;6(4):54-62. [doi: 10.1109/msmc.2020.3002519]
16. Avan A, Digaleh H, Di Napoli M, Stranges S, Behrouz R, Shojaeianbabaei G, et al. Socioeconomic status and stroke incidence, prevalence, mortality, and worldwide burden: an ecological analysis from the Global Burden of Disease Study 2017. BMC Med 2019;17(1):191 [FREE Full text] [doi: 10.1186/s12916-019-1397-3] [Medline: 31647003]
17. Dobkin BH, Dorsch A. New evidence for therapies in stroke rehabilitation. Curr Atheroscler Rep 2013;15(6):331 [FREE Full text] [doi: 10.1007/s11883-013-0331-y] [Medline: 23591673]
18. Valentin LSS. Can digital games be a way of improving the neuroplasticity in stroke damage? Can the adult brain grow new cells or rewire itself in response to a new experience? Open J Med Psychol 2017;6(2):153-165 [FREE Full text] [doi: 10.4236/ojmp.2017.62013]
19. Suchak P, Ameer K, Gaylard J, Newman J, Ali K. Improving patients experience in a neuro-rehabilitation unit using tablet computers. Eur J Pers Cent Healthc 2016;4(2):289. [doi: 10.5750/ejpch.v4i2.1082]

XSL•FO
RenderX

20. Macoir J, Lavoie M, Routhier S, Bier N. Key factors for the success of self-administered treatments of poststroke aphasia using technologies. Telemed J E Health 2019;25(8):663-670. [doi: 10.1089/tmj.2018.0116] [Medline: 30142032]
21. Fox CM, Ramig LO, Ciucci MR, Sapir S, McFarland DH, Farley BG. The science and practice of LSVT/LOUD: neural plasticity-principled approach to treating individuals with Parkinson disease and other neurological disorders. Semin Speech Lang 2006;27(4):283-299. [doi: 10.1055/s-2006-955118] [Medline: 17117354]
22. Chen YPP, Johnson C, Lalbakhsh P, Caelli T, Deng G, Tay D, et al. Systematic review of virtual speech therapists for speech disorders. Comput Speech Lang 2016;37:98-128. [doi: 10.1016/j.csl.2015.08.005]
23. Bakker M, Beijer L, Rietveld T. Considerations on effective feedback in computerized speech training for dysarthric speakers. Telemed J E Health 2019;25(5):351-358. [doi: 10.1089/tmj.2018.0050] [Medline: 30074851]
24. Muñoz ML, Hoffman LM, Brimo D. Be smarter than your phone: a framework for using apps in clinical practice. CICSD 2013;40(Fall):138-150 [FREE Full text] [doi: 10.1044/cicsd_40_f_138]
25. Chiviacowsky S, Lessa HT. Choices over feedback enhance motor learning in older adults. J Mot Learn Dev 2017;5(2):304-318. [doi: 10.1123/jmld.2016-0031]
26. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194 [FREE Full text] [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
27. Oh MS, Yu KH, Lee JH, Jung S, Ko IS, Shin JH, et al. Validity and reliability of a Korean version of the National Institutes of Health Stroke Scale. J Clin Neurol 2012;8(3):177-183 [FREE Full text] [doi: 10.3988/jcn.2012.8.3.177] [Medline: 23091526]
28. Kim J, Shin I, Yoon J, Lee H. Comparison of diagnostic validities between MMSE-K and K-MMSE for screening of dementia. J Korean Neuropsychiatr Assoc 2003;42(1):124-130 [FREE Full text]
29. Rowe F, Brand D, Jackson CA, Price A, Walker L, Harrison S, et al. Visual impairment following stroke: do stroke patients require vision assessment? Age Ageing 2009;38(2):188-193 [FREE Full text] [doi: 10.1093/ageing/afn230] [Medline: 19029069]
30. Isaacson JE, Vora NM. Differential diagnosis and treatment of hearing loss. Am Fam Physician 2003;68(6):1125-1132 [FREE Full text] [Medline: 14524400]
31. Broglio K. Randomization in clinical trials: permuted blocks and stratification. JAMA 2018;319(21):2223-2224. [doi: 10.1001/jama.2018.6360] [Medline: 29872845]
32. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002;359(9307):696-700. [doi: 10.1016/S0140-6736(02)07816-9] [Medline: 11879884]
33. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: using an aging barriers framework to classify usability problems. Int J Med Inform 2019;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
34. de Barros AC, Leitão R, Ribeiro J. Design and evaluation of a mobile user interface for older adults: navigation, interaction and visual design recommendations. Procedia Comput Sci 2014;27:369-378 [FREE Full text] [doi: 10.1016/j.procs.2014.02.041]
35. Ko SB, Park HY, Kim BM, Heo JH, Rha JH, Kwon SU, et al. 2019 update of the Korean Clinical Practice Guidelines of Stroke for endovascular recanalization therapy in patients with acute ischemic stroke. J Stroke 2019;21(2):231-240 [FREE Full text] [doi: 10.5853/jos.2019.00024] [Medline: 30991800]
36. Kleindorfer DO, Towfighi A, Chaturvedi S, Cockroft KM, Gutierrez J, Lombardi-Hill D, et al. 2021 guideline for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline from the American Heart Association/American Stroke Association. Stroke 2021;52(7):e364-e467 [FREE Full text] [doi: 10.1161/STR.0000000000000375] [Medline: 34024117]
37. Kim HH. Perceptual, acoustical, and physiological tools in ataxic dysarthria management: a case report. 1996 Presented at: Proceedings of the KSPS Conference; 1996; South Korea p. 9-22 URL: https://koreascience.kr/article/CFKO199613842057217.page
38. Patel RR, Awan SN, Barkmeier-Kraemer J, Courey M, Deliyski D, Eadie T, et al. Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. Am J Speech Lang Pathol 2018;27(3):887-905 [FREE Full text] [doi: 10.1044/2018_AJSLP-17-0009] [Medline: 29955816]
39. Lousada M, Jesus LMT, Hall A, Joffe V. Intelligibility as a clinical outcome measure following intervention with children with phonologically based speech-sound disorders. Int J Lang Commun Disord 2014;49(5):584-601. [doi: 10.1111/1460-6984.12095] [Medline: 24861159]
40. Kim YT, Park H, Kang JK, Kim JA, Shin MJ, Kim SJ, et al. Validity and reliability analyses for the development of Urimal Test of Articulation and Phonology-2. Commun Sci Disord 2018;23(4):959-970 [FREE Full text] [doi: 10.12963/csd.18545]
41. Shriberg LD, Austin D, Lewis BA, McSweeny JL, Wilson DL. The percentage of consonants correct (PCC) metric: extensions and reliability data. J Speech Lang Hear Res 1997;40(4):708-722. [doi: 10.1044/jslhr.4004.708] [Medline: 9263938]

XSL•FO
RenderX

42. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 1999;282(18):1737-1744 [FREE Full text] [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
43. Choi HS, Choi JH, Park KH, Joo KJ, Ga H, Ko HJ, et al. Standardization of the Korean version of Patient Health Questionnaire-9 as a screening instrument for major depressive disorder. J Korean Acad Fam Med 2007;28(2):114-119 [FREE Full text]
44. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
45. Ahn JK, Kim Y, Choi KH. The psychometric properties and clinical utility of the Korean version of GAD-7 and GAD-2. Front Psychiatry 2019;10:127 [FREE Full text] [doi: 10.3389/fpsyt.2019.00127] [Medline: 30936840]
46. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001;33(5):337-343. [doi: 10.3109/07853890109002087] [Medline: 11491192]
47. Lee YK, Nam HS, Chuang LH, Kim KY, Yang HK, Kwon IS, et al. South Korean time trade-off values for EQ-5D health states: modeling with observed values for 101 health states. Value Health 2009;12(8):1187-1193 [FREE Full text] [doi: 10.1111/j.1524-4733.2009.00579.x] [Medline: 19659703]
48. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, McClelland IL, Thomas B, Weerdmeester BA, editors. Usability Evaluation in Industry. Boca Raton, FL: CRC Press; 1996:189-194.
49. Laver K, George S, Ratcliffe J, Crotty M. Measuring technology self efficacy: reliability and construct validity of a modified computer self efficacy scale in a clinical rehabilitation setting. Disabil Rehabil 2012;34(3):220-227. [doi: 10.3109/09638288.2011.593682] [Medline: 21958357]
50. Muller KE, Lavange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. J Am Stat Assoc 1992;87(420):1209-1226 [FREE Full text] [doi: 10.1080/01621459.1992.10476281] [Medline: 24790282]
51. Levy ES, Moya-Galé G, Chang YHM, Freeman K, Forrest K, Brin MF, et al. The effects of intensive speech treatment on intelligibility in Parkinson's disease: a randomised controlled trial. EClinicalMedicine 2020;24:100429 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100429] [Medline: 32639484]
52. Mitchell C, Bowen A, Tyson S, Conroy P. ReaDySpeech for people with dysarthria after stroke: protocol for a feasibility randomised controlled trial. Pilot Feasibility Stud 2018;4:25 [FREE Full text] [doi: 10.1186/s40814-017-0169-0] [Medline: 28748108]
53. Sellars C, Hughes T, Langhorne P. Speech and language therapy for dysarthria due to non-progressive brain damage. Cochrane Database Syst Rev 2005(3):CD002088. [doi: 10.1002/14651858.CD002088.pub2] [Medline: 16034872]
54. Spencer KA, Brown KA. Dysarthria following stroke. Semin Speech Lang 2018;39(1):15-24. [doi: 10.1055/s-0037-1608852] [Medline: 29359302]
55. Sapir S, Spielman JL, Ramig LO, Story BH, Fox C. Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings. J Speech Lang Hear Res 2007;50(4):899-912 [FREE Full text] [doi: 10.1044/1092-4388(2007/064)] [Medline: 17675595]
56. Tickle-Degnen L, Ellis T, Saint-Hilaire MH, Thomas CA, Wagenaar RC. Self-management rehabilitation and health-related quality of life in Parkinson's disease: a randomized controlled trial. Mov Disord 2010;25(2):194-204 [FREE Full text] [doi: 10.1002/mds.22940] [Medline: 20077478]
57. Chiaramonte R, Pavone P, Vecchio M. Speech rehabilitation in dysarthria after stroke: a systematic review of the studies. Eur J Phys Rehabil Med 2020;56(5):547-562 [FREE Full text] [doi: 10.23736/S1973-9087.20.06185-7] [Medline: 32434313]
58. Mahler LA, Ramig LO. Intensive treatment of dysarthria secondary to stroke. Clin Linguist Phon 2012;26(8):681-694. [doi: 10.3109/02699206.2012.696173] [Medline: 22774928]
59. Park S, Theodoros D, Finch E, Cardell E. Be Clear: a new intensive speech treatment for adults with nonprogressive dysarthria. Am J Speech Lang Pathol 2016;25(1):97-110. [doi: 10.1044/2015_AJSLP-14-0113] [Medline: 26882004]
60. Wenke RJ, Cornwell P, Theodoros DG. Changes to articulation following LSVT® and traditional dysarthria therapy in non-progressive dysarthria. Int J Speech Lang Pathol 2010;12(3):203-220. [doi: 10.3109/17549500903568468] [Medline: 20433339]
61. Beijer LJ, Rietveld ACM, Ruiter MB, Geurts ACH. Preparing an E-learning-based Speech Therapy (EST) efficacy study: identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers. Clin Linguist Phon 2014;28(12):927-950. [doi: 10.3109/02699206.2014.936627] [Medline: 25025268]
62. Abbadessa G, Brigo F, Clerico M, De Mercanti S, Trojsi F, Tedeschi G, et al. Digital therapeutics in neurology. J Neurol 2022;269(3):1209-1224 [FREE Full text] [doi: 10.1007/s00415-021-10608-4] [Medline: 34018047]
63. Kamoen O, Maqueda V, Yperzeele L, Pottel H, Cras P, Vanhooren G, et al. Stroke coach: a pilot study of a personal digital coaching program for patients after ischemic stroke. Acta Neurol Belg 2020;120(1):91-97. [doi: 10.1007/s13760-019-01218-z] [Medline: 31701472]
64. Karlsen C, Ludvigsen MS, Moe CE, Haraldstad K, Thygesen E. Experiences of community-dwelling older adults with the use of telecare in home care services: a qualitative systematic review. JBI Database System Rev Implement Rep 2017;15(12):2913-2980. [doi: 10.11124/JBISRIR-2017-003345] [Medline: 29219874]

XSL•FO
RenderX

65. Griffin M, Bentley J, Shanks J, Wood C. The effectiveness of Lee Silverman Voice Treatment therapy issued interactively through an iPad device: a non-inferiority study. J Telemed Telecare 2018;24(3):209-215. [doi: 10.1177/1357633X17691865] [Medline: 28147896]
66. Abelson JS, Kaufman E, Symer M, Peters A, Charlson M, Yeo H. Barriers and benefits to using mobile health technology after operation: a qualitative study. Surgery 2017;162(3):605-611. [doi: 10.1016/j.surg.2017.05.007] [Medline: 28651777]
67. Bowser DM, Shepard DS, Nandakumar A, Okunogbe A, Morrill T, Halasa-Rappell Y, et al. Cost effectiveness of mobile health for antenatal care and facility births in Nigeria. Ann Glob Health 2018;84(4):592-602 [FREE Full text] [doi: 10.29024/aogh.2364]
68. Zhou X, Snoswell CL, Harding LE, Bambling M, Edirippulige S, Bai X, et al. The role of telehealth in reducing the mental health burden from COVID-19. Telemed J E Health 2020;26(4):377-379. [doi: 10.1089/tmj.2020.0068] [Medline: 32202977]
69. Baker C, Worrall L, Rose M, Hudson K, Ryan B, O'Byrne L. A systematic review of rehabilitation interventions to prevent and treat depression in post-stroke aphasia. Disabil Rehabil 2018;40(16):1870-1892. [doi: 10.1080/09638288.2017.1315181] [Medline: 28420284]
70. Kneebone I, Walker-Samuel N, Swanston J, Otto E. Relaxation training after stroke: potential to reduce anxiety. Disabil Rehabil 2014;36(9):771-774. [doi: 10.3109/09638288.2013.808275] [Medline: 23841562]
71. Aprile I, Di Stasio E, Romitelli F, Lancellotti S, Caliandro P, Tonali P, et al. Effects of rehabilitation on quality of life in patients with chronic stroke. Brain Inj 2008;22(6):451-456. [doi: 10.1080/02699050802060639] [Medline: 18465386]
72. Ahrens J, Shao R, Blackport D, Macaluso S, Viana R, Teasell R, et al. Cognitive-behavioral therapy for managing depressive and anxiety symptoms after stroke: a systematic review and meta-analysis. Top Stroke Rehabil 2023;30(4):368-383. [doi: 10.1080/10749357.2022.2049505] [Medline: 35352629]
73. Damschroder LJ, Buis LR, McCant FA, Kim HM, Evans R, Oddone EZ, et al. Effect of adding telephone-based brief coaching to an mHealth App (Stay Strong) for promoting physical activity among veterans: randomized controlled trial. J Med Internet Res 2020;22(8):e19216 [FREE Full text] [doi: 10.2196/19216] [Medline: 32687474]
74. Ly K, Janni E, Wrede R, Sedem M, Donker T, Carlbring P, et al. Experiences of a guided smartphone-based behavioral activation therapy for depression: a qualitative study. Internet Interv 2015;2(1):60-68 [FREE Full text] [doi: 10.1016/j.invent.2014.12.002]
75. Kim H, Lee SH, Cho NB, You H, Choi T, Kim J. User-dependent usability and feasibility of a swallowing training mHealth app for older adults: mixed methods pilot study. JMIR Mhealth Uhealth 2020;8(7):e19585 [FREE Full text] [doi: 10.2196/19585] [Medline: 32663161]
76. Lu SC, Wen TN, Chang PL. The study of smartphone usage competency assessment and training for the elderly. Stud Health Technol Inform 2017;245:161-165. [Medline: 29295073]
77. Chen Y, Abel KT, Janecek JT, Chen Y, Zheng K, Cramer SC. Home-based technologies for stroke rehabilitation: a systematic review. Int J Med Inform 2019;123:11-22 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.12.001] [Medline: 30654899]
78. Enderby P. Frenchay Dysarthria Assessment. Int J Lang Commun Disord 1980;15(3):165-173. [doi: 10.3109/13682828009112541]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
mCSES: Modified Computer Self-Efficacy Scale
SLP: speech-language pathologist
SUS: System Usability Scale
U-TAP2: Urimal Test of Articulation and Phonology 2

Edited by G Eysenbach; submitted 16.01.24; peer-reviewed by SJ Lee; comments to author 13.02.24; revised version received 21.02.24; accepted 20.03.24; published 25.04.24.

Please cite as:
Kim Y, Kim M, Kim J, Song TJ
Smartphone-Based Speech Therapy for Poststroke Dysarthria: Pilot Randomized Controlled Trial Evaluating Efficacy and Feasibility
J Med Internet Res 2024;26:e56417
URL: https://www.jmir.org/2024/1/e56417
doi:10.2196/56417
PMID:38509662

XSL•FO
RenderX

©Yuyoung Kim, Minjung Kim, Jinwoo Kim, Tae-Jin Song. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of a Smartphone App–Based Intervention With Bluetooth-Connected Monitoring Devices and a Feedback System in Heart Failure (SMART-HF Trial): Randomized Controlled Trial

Minjae Yoon[1], MD; Seonhwa Lee[2], MD; Jah Yeon Choi[3], MD, PhD; Mi-Hyang Jung[4], MD, PhD; Jong-Chan Youn[4], MD, PhD; Chi Young Shim[5], MD, PhD; Jin-Oh Choi[6], MD, PhD; Eung Ju Kim[3], MD, PhD; Hyungseop Kim[2], MD, PhD; Byung-Su Yoo[7], MD, PhD; Yeon Joo Son[8], BS; Dong-Ju Choi[1], MD, PhD

[1]Division of Cardiology, Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seognam, Republic of Korea

[2]Division of Cardiology, Department of Internal Medicine, Keimyung University Dongsan Hospital, Daegu, Republic of Korea

[3]Cardiovascular Center, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[4]Division of Cardiology, Department of Internal Medicine, Seoul St. Mary's Hospital, Catholic Research Institute for Intractable Cardiovascular Disease, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[5]Division of Cardiology, Department of Internal Medicine, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

[6]Cardiac and Vascular Center, Samsung Medical Center, Sungkyunkwan University, Seoul, Republic of Korea

[7]Department of Internal Medicine, Wonju Severance Christian Hospital, Yonsei University Wonju College of Medicine, Woonju, Republic of Korea

[8]Healthcare Business Department, AI/DX Convergence Business Group, KT, Seoul, Republic of Korea

Corresponding Author:
Dong-Ju Choi, MD, PhD
Division of Cardiology, Department of Internal Medicine
Seoul National University Bundang Hospital, Seoul National University College of Medicine
82, Gumi-ro 173 Beon-gil, Bundang-gu
Seognam, 13620
Republic of Korea
Phone: 82 317877007
Fax: 82 317877041
Email: djchoi@snubh.org

Abstract

Background: Current heart failure (HF) guidelines recommend a multidisciplinary approach, discharge education, and self-management for HF. However, the recommendations are challenging to implement in real-world clinical settings.

Objective: We developed a mobile health (mHealth) platform for HF self-care to evaluate whether a smartphone app–based intervention with Bluetooth-connected monitoring devices and a feedback system can help improve HF symptoms.

Methods: In this prospective, randomized, multicenter study, we enrolled patients 20 years of age and older, hospitalized for acute HF, and who could use a smartphone from 7 tertiary hospitals in South Korea. In the intervention group (n=39), the apps were automatically paired with Bluetooth-connected monitoring devices. The patients could enter information on vital signs, HF symptoms, diet, medications, and exercise regimen into the app daily and receive feedback or alerts on their input. In the control group (n=38), patients could only enter their blood pressure, heart rate, and weight using conventional, non-Bluetooth devices and could not receive any feedback or alerts from the app. The primary end point was the change in dyspnea symptom scores from baseline to 4 weeks, assessed using a questionnaire.

Results: At 4 weeks, the change in dyspnea symptom score from baseline was significantly greater in the intervention group than in the control group (mean –1.3, SD 2.1 vs mean –0.3, SD 2.3; *P*=.048). A significant reduction was found in body water composition from baseline to the final measurement in the intervention group (baseline level mean 7.4, SD 2.5 vs final level mean 6.6, SD 2.5; *P*=.003). App adherence, which was assessed based on log-in or the percentage of days when symptoms were first observed, was higher in the intervention group than in the control group. Composite end points, including death, rehospitalization, and urgent HF visits, were not significantly different between the 2 groups.

XSL•FO
RenderX

Conclusions: The mobile-based health platform with Bluetooth-connected monitoring devices and a feedback system demonstrated improvement in dyspnea symptoms in patients with HF. This study provides evidence and rationale for implementing mobile app–based self-care strategies and feedback for patients with HF.

Trial Registration: ClinicalTrials.gov NCT05668000; https://clinicaltrials.gov/study/NCT05668000

(J Med Internet Res 2024;26:e52075) doi:10.2196/52075

KEYWORDS

heart failure; mobile applications; mobile health; self-care; vital sign monitoring; mobile phone

Introduction

Heart failure (HF) is a rapidly growing public health problem, with an estimated prevalence of 64 million people globally [1-5]. Although outcomes of HF have recently improved with the development of medications, a high rate of readmission remains after initial hospitalization for acute HF [6-10]. Recent studies have shown that 18%-23% of patients with acute HF were readmitted within 1 month, and these numbers have not decreased recently [6,11,12]. Therefore, the postdischarge management of patients with HF is important. In particular, multidisciplinary interventions, in addition to pharmacotherapy and discharge education, are known to improve quality of life and reduce hospitalizations [13-16]. A recent meta-analysis has shown that self-management interventions, including symptom and sign monitoring, education, and enhancement of drug adherence, improve outcomes of HF-related hospitalization and all-cause death, despite heterogeneity in the interventions [17]. Thus, current HF guidelines recommend multidisciplinary management, specific discharge education, and support to facilitate HF self-care [18-20].

Although encouraging self-care and providing feedback from health care providers are the most effective means, they are difficult to implement in a real-world clinical setting. These are also associated with higher costs and require more infrastructure and manpower. Smartphones are currently available to most of the general population at an affordable cost. The advent of mobile health (mHealth) technology and advances in artificial intelligence (AI) have enabled patient self-care, symptom, and sign monitoring, as well as mobile-based feedback via smartphone apps. Recent studies using smartphone apps for the self-management of cardiovascular disease demonstrated that this intervention led to better outcomes than the control groups [21-23]. Additionally, some recent studies and meta-analyses regarding mHealth apps have supported self-care among patients with HF [24-30]. However, these previous studies had a small number of participants and limited app functionality. Self-monitoring and feedback are particularly important, and the lack of these critical features limits the interpretation of the effectiveness of HF apps. Furthermore, functions provided in previous mobile-based interventions for HF were heterogeneous and varied between studies, which may lead to inconsistent results.

Given the uncertainty regarding the benefits of mobile apps for HF and their diverse functionality, we developed a mHealth platform to provide self-management interventions for patients with HF. This study aimed to evaluate whether this smartphone app–based intervention with Bluetooth-connected monitoring devices and a feedback system could improve the symptoms and clinical outcomes of HF.

Methods

Study Design and Population

The self-monitoring using a mobile app to improve symptoms and reduce rehospitalization and mortality in heart failure (SMART-HF) study is a prospective, multicenter, randomized, open-label trial to evaluate the efficacy of smartphone apps in improving the symptoms and outcomes of HF. A total of 7 tertiary university hospitals in South Korea participated in this study. The study design was registered at ClinicalTrials.gov (NCT05668000).

We enrolled patients aged older than 20 years who were hospitalized for acute HF with obvious symptoms or signs of HF at admission. Patients with N-terminal pro-B-type natriuretic peptide (NT-proBNP) levels ≥400 pg/mL or brain natriuretic peptide levels ≥100 pg/mL were enrolled. Since smartphone use was essential in this study, participants should be able to use Android smartphones well and be capable of following instructions on how to use apps. Patients with a baseline systolic blood pressure (BP) of <90 mm Hg or a resting heart rate (HR) of <50 beats per minute were excluded from this study. Additionally, patients were excluded if they had a cardiac [30] implantable electronic device that could interfere with body water analysis. Patients who were expected to have a prolonged hospital stay owing to medical problems other than HF were also excluded. The detailed inclusion and exclusion criteria are presented in Table S1 in Multimedia Appendix 1.

Ethical Considerations

This clinical trial was approved by the institutional review boards of Seoul National University Bundang Hospital (B-2211-795-304) and other hospitals. The study was conducted in accordance with the Declaration of Helsinki, and all study data were deidentified. All patients provided written informed consent upon enrollment.

Patient Recruitment and Randomization

After a comprehensive interview, eligible participants were asked to provide written informed consent before being discharged from the hospital. At baseline (visit 1), all participants were assessed for sex, age, demographics, vital signs, comorbidities, laboratory data, and medication use. Eligible participants were randomized 1:1 to either the intervention group (app with Bluetooth-connected monitoring devices and feedback) or the control group (app only) using a web-based central randomization service (Figure 1A). Due to

XSL•FO
RenderX

the nature of the intervention, the study participants and investigators interacting with the patients were not blinded to the group allocation.

Figure 1. Study design of the (A) SMART-HF trial and (B) system of app and platform. (A) Patients with acute HF were randomized into the intervention (app with monitoring devices and feedback) or control (app only) groups. (B) The app and platform are comprised of three parts (1) a smartphone app for patients, (2) a dashboard system for physicians, and (3) a clinical decision support system on the platform. PC: personal computer; SMART-HF: self-monitoring using a mobile app to improve symptoms and reduce rehospitalization and mortality in heart failure.

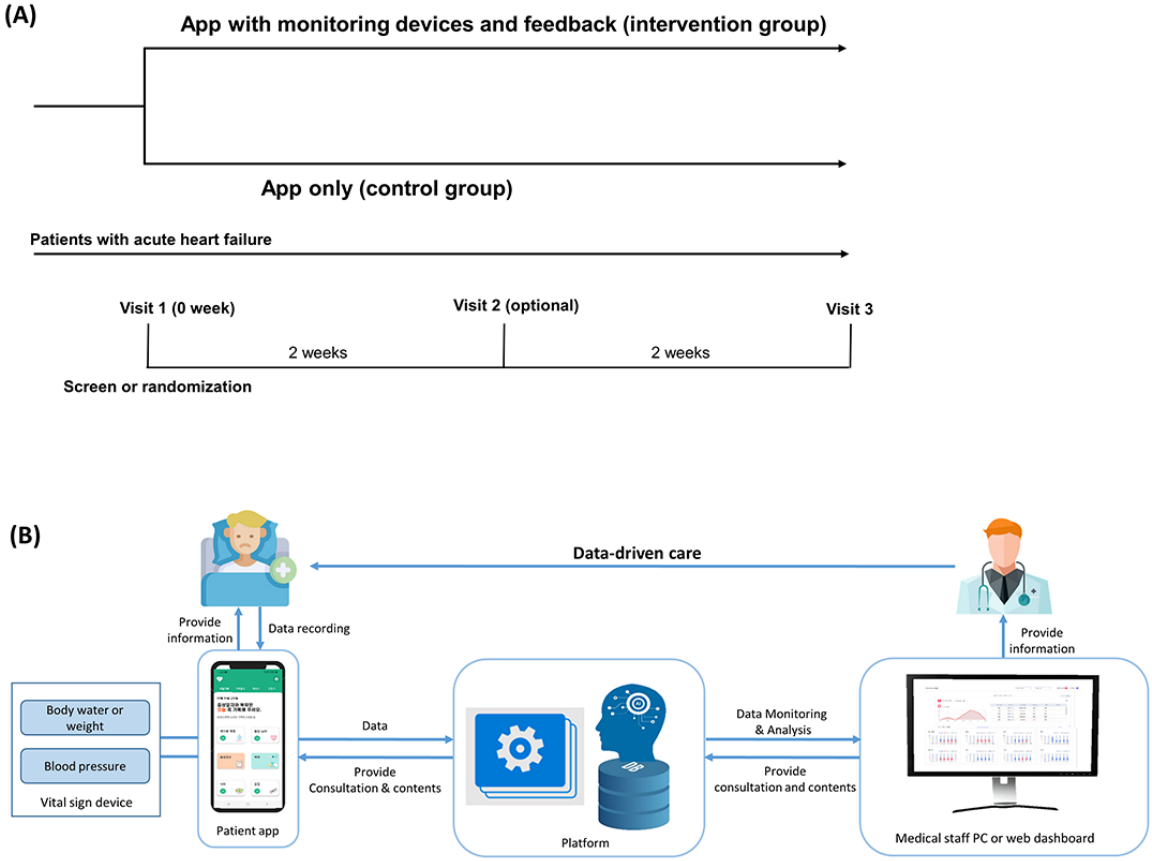

Intervention and Follow-Up

We developed a mobile app and operating platform to improve the symptoms and outcomes of patients with HF. The app and platform comprised three parts: (1) a smartphone app for patients, (2) a dashboard system for physicians, and (3) a clinical decision support system (CDSS) on the platform (Figure 1B and Figures S1 and S2 in Multimedia Appendix 1).

The smartphone app comprised 4 menu screens (Figure 2), and the detailed functions of the app are described in Table S2 in Multimedia Appendix 1. In the "Today" menu, patients can enter their information, including body water, weight, BP, HR, symptom diary, medications, diet, and exercise. Patients can input their symptoms of HF, including dyspnea, fatigue, ankle edema, and palpitations, daily. Symptoms were scored as 0 (no symptoms), 1 (mild symptoms), 2 (moderate symptoms), and 3 (severe symptoms; Table S3 in Multimedia Appendix 1). Moreover, they could enter their daily diet or physical activity into the apps, and the amount of sodium in their diet was assessed through a camera image using AI (Figure S3 in Multimedia Appendix 1). Additionally, the BP, HR, weight, and body water could be entered into the app by the patients (intervention + control group) or automatically paired using Bluetooth-connected monitoring devices (intervention group only). In the "Records" menu, patients can access and review their historical data entered by date, as well as receive feedback regarding medication adherence, dietary habits, and exercise. Furthermore, the app can analyze and evaluate a patient's symptom scores or measured vital signs and subsequently generate alerts based on a predefined algorithm (Tables S3 and S4 in Multimedia Appendix 1) using the CDSS. In the event of rapid changes in a patient's symptoms or vital signs, the app prompts the patient to provide information regarding medication adherence, symptom severity, and any other concerns. Upon perceiving significant changes, it sends a message to the patients, advising them to contact their health care provider. In the "Contents" menu, the app can provide patients with information about HF, including symptoms, treatment, dietary guidelines, and exercise recommendations using chatbots.

XSL•FO
RenderX

Figure 2. Application menu screens and functions. (A) Menu screens and (B) application functions. BP: blood pressure; DBP: diastolic blood pressure; HR: heart rate; SBP: systolic blood pressure.

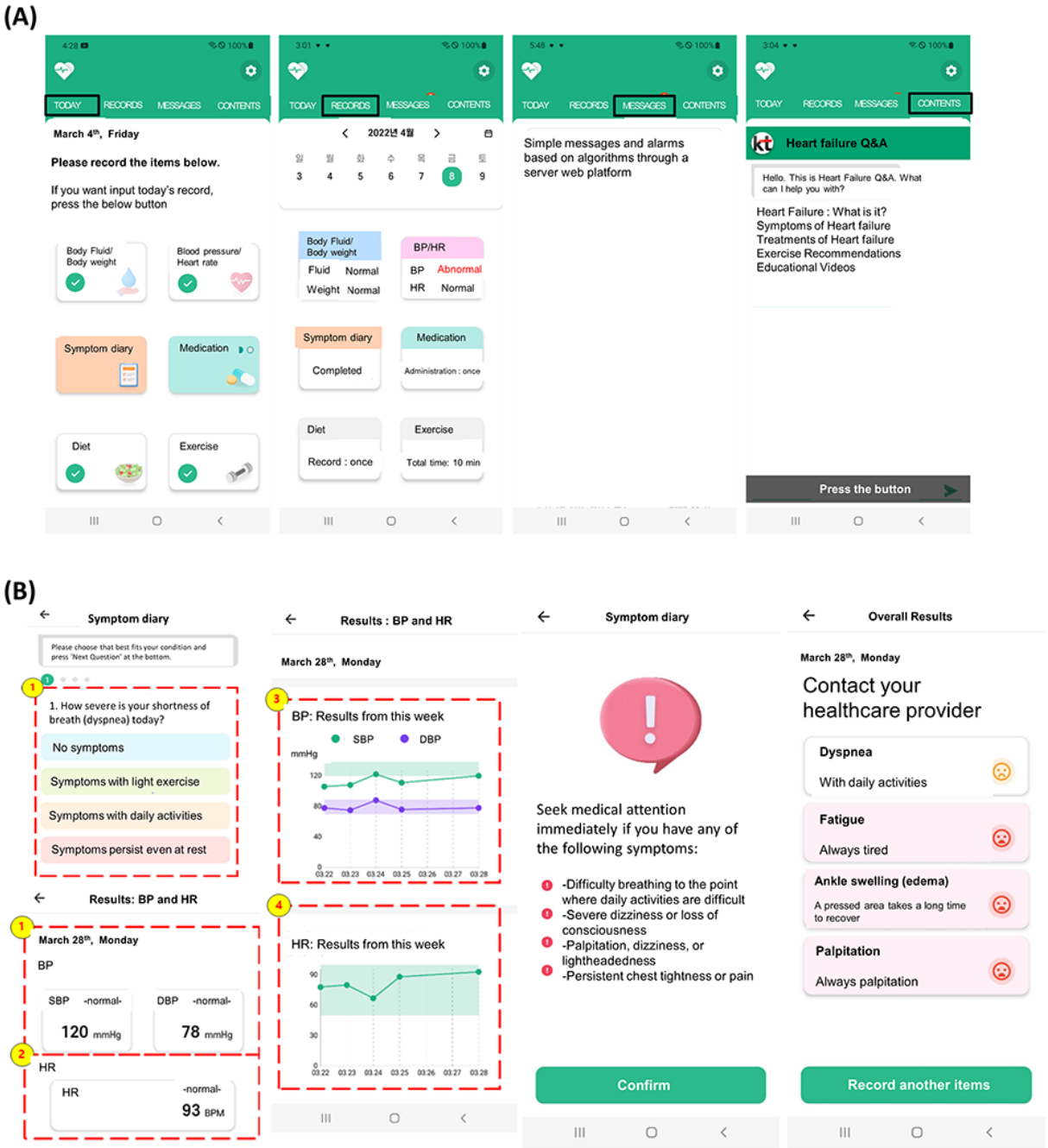

The dashboard system was designed to present stored information to physicians (Figure S4 in Multimedia Appendix 1). This information included BP, HR, body weight, body water composition, HF symptom score, medications, diet, exercise, and any data entered by the patient. Physicians could easily observe the vital sign trends (daily, weekly, and monthly) of the participants. The CDSS analyzes the input data and generates recommendations for users using AI and algorithms. For example, it aggregates data and information from the patient and sends necessary messaging alerts to the patient's app and, if needed, to the health care provider's dashboard during office visits.

After randomization, both groups installed the study app on their Android smartphones, which the research nurse provided. Only the intervention group had full access to all functions of the app, and the apps were automatically paired with Bluetooth-connected monitoring devices, including a BP monitor, weight scale, and body water analyzer. The intervention group received feedback and alerts from the app. In the control group, patients could enter their BP, HR, and weight into the app using conventional non-Bluetooth devices only. Additionally, they received no feedback or alerts from the app. The "Contents" menu providing information about HF was available in both groups. The study aimed to evaluate the improvement in HF symptoms based on the functionality of the

XSL•FO
RenderX

mobile app between the intervention and control groups, with a particular focus on Bluetooth-connected monitoring devices and the feedback function of the app.

We used a Bluetooth-enabled BP monitor BP170 (InBody Co). Bioimpedance analysis was performed using a portable multifrequency bioimpedance device (BWA ON; InBody Co) to analyze body water composition. Bioimpedance analysis has recently proven its efficacy in patients with acute HF with dyspnea [31], and the device can estimate body fluid status using extracellular water divided by total body water of the 4 limbs and trunk. The parameters of extracellular water divided by total body water <0.390, 0.390-0.400, and >0.400 were considered to be "normal," "slightly over," and "over," respectively (Table S5 in Multimedia Appendix 1). The device classifies body water levels from levels 1 to 16.

Follow-up visits were scheduled at 2 (visit 2, optional) and 4 (visit 3) weeks after randomization. Both groups received standard care according to the current HF guidelines [20,32], and HF medication could be modified at the discretion of the treating physicians.

Study Outcomes

The primary end point of the study was the change in dyspnea symptom scores from baseline to 4 weeks, which was assessed using a questionnaire. We used a visual analog scale and a numerical rating scale for dyspnea ranging from 0 to 10 based on previous studies [33,34] (Figure S5 in Multimedia Appendix 1).

There were two secondary end points: (1) a composite outcome, including death, rehospitalization, and urgent visit for HF and (2) the change in body water composition from baseline to the last measurement in the intervention group. The exploratory clinical outcome was adherence to the app, which was calculated using either log-in access logs or the percentage of days when symptoms were entered into the app. App adherence was defined as follows:

We also compared app satisfaction scores at 4 weeks through a questionnaire, ranging from 0 to 10, between the intervention and control groups. If patients could not attend the scheduled study date, the outcome measures were assessed telephonically.

Sample Size and Statistical Analysis

Due to the absence of previous research and the pioneering nature of our intervention, our study adopts an exploratory approach. The primary objective was not to test a specific hypothesis regarding efficacy but rather to collect preliminary data and assess feasibility in preparation for a larger-scale study. Consequently, a sample size calculation was not conducted at this preliminary stage. Since this was a pilot study to explore the benefits of mobile apps for HF, we enrolled a total of 84 participants across 7 institutes, considering the duration of the study and the number of participating hospitals.

Categorical variables are reported as frequencies (percentages), and continuous variables are expressed as means (SD) or medians with IQR. Categorical variables were compared using the Pearson chi-square test or Fisher exact test, and continuous variables were compared using the Student *t* test or the Mann-Whitney *U* test. The change in body water composition from baseline to the last measurement after randomization was analyzed using the paired *t* test.

The intention-to-treat analysis included all randomized patients. The efficacy end points were primarily analyzed using the full analysis set, which included randomized participants who used the mobile app at least once. We also performed a per-protocol sensitivity analysis, including patients who completed the study protocol. For patients who dropped out before the end of the trial or had missing data, we used the latest available records for analysis, namely, the last observation carried forward method.

All tests were 2-tailed, and a P value <.05 was considered statistically significant. Statistical analyses were performed by using R (version 4.2.0; R Core Team).

Results

Patient Enrollment and Clinical Characteristics

From October 2022 to January 2023, 132 patients from 7 centers were screened for eligibility, and 84 were randomly assigned to the intervention (n=43) or control (n=41) group (Figure 3 and CONSORT [Consolidated Standards of Reporting Trials] checklist in Multimedia Appendix 2). After allocation, 7 patients (4 and 3 in the intervention and control groups, respectively) were excluded because they did not use the app at least once; thus, 77 patients (39 and 38 in the intervention and control groups, respectively) were included in the full analysis set.

XSL•FO
RenderX

Figure 3. CONSORT flow diagram. CONSORT: Consolidated Standards of Reporting Trials; ICU: intensive care unit.

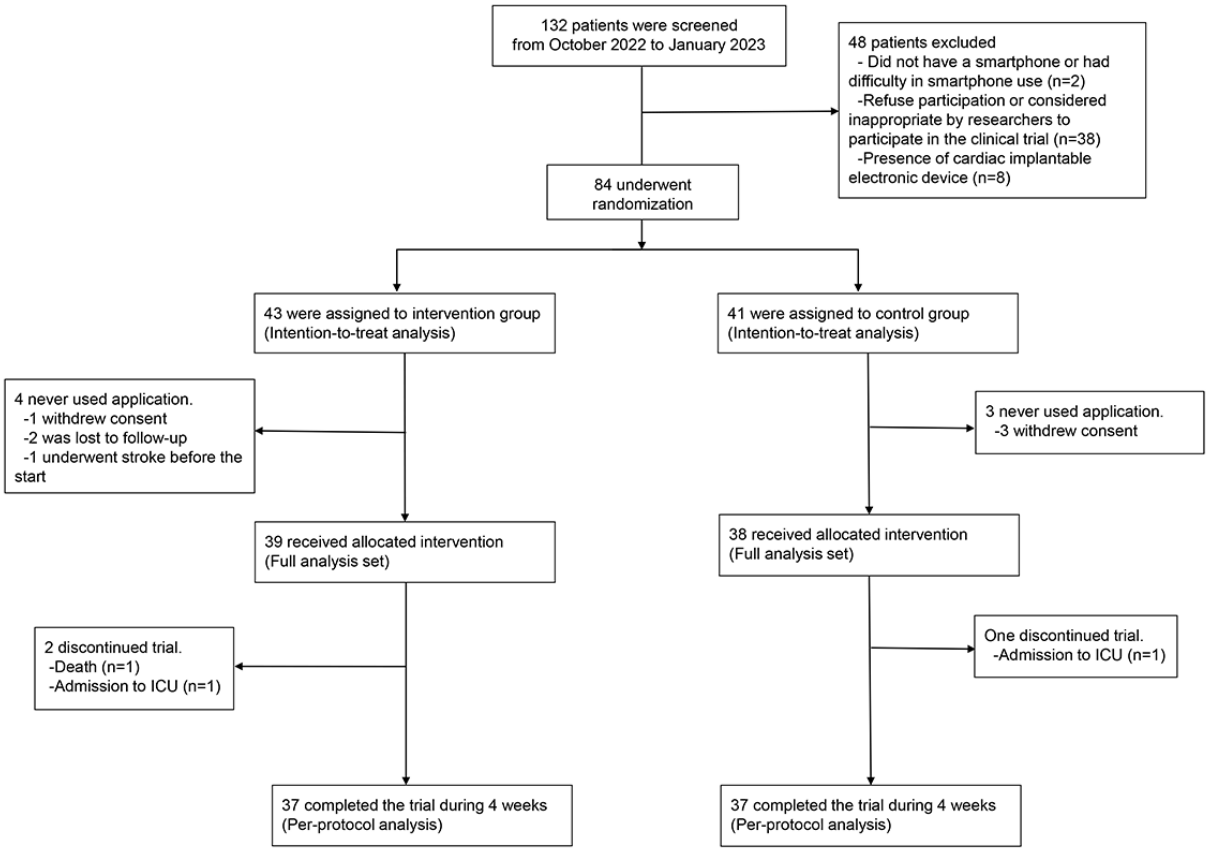

The baseline characteristics of the groups are presented in Table 1. The mean age of the total population was 62.1 (SD 14.7) years; 62% were male, 61% had de novo HF, and 30% had HF of an ischemic etiology. The median left ventricular ejection fraction was 38.7% (IQR 28%-58%) and the median NT-proBN*P* value was 2572 (IQR 1211-6651) pg/mL. The baseline characteristics, including age, sex, demographic data, laboratory values, and comorbidities, were well balanced between the 2 groups, except for diastolic BP. Additionally, the medication use at discharge was not significantly different between the 2 groups.

XSL•FO
RenderX

Table 1. Baseline characteristics of the patients according to the treatment groups (full analysis set).

Variables	Intervention (n=39)	Control (n=38)	*P* value
Clinical characteristics			
Age (years), mean (SD)	58.4 (16.0)	65.7 (12.6)	.06
Age ≥65 years, n (%)	16 (41	19 (50)	.57
Male, n (%)	25 (64)	23 (61)	.93
Height (cm), mean (SD)	164.5 (8.8)	163.3 (9.6)	.59
Weight (kg), median (IQR)	63.0 (57.0-75.1)	61.0 (52.6-71.4)	.16
BMI (kg/m²), median (IQR)	24.1 (21.4-27.3)	23.9 (20.6-25.5)	.25
Systolic blood pressure (mm Hg), mean (SD)	124.1 (16.7)	117.8 (17.6)	.11
Diastolic blood pressure (mm Hg), mean (SD)	76.4 (10.9)	69.1 (11.7)	.006
De novo HF[a], n (%)	23 (59)	24 (63)	.89
Ischemic etiology, n (%)	12 (31)	11 (29)	>.99
Hypertension, n (%)	18 (46)	12 (32)	.28
Diabetes mellitus, n (%)	14 (36)	15 (40)	.93
Chronic kidney disease, n (%)	9 (23)	5 (13)	.44
Atrial fibrillation, n (%)	13 (33)	16 (42)	.58
NYHA[b] functional class (n=36), n (%)			.78
I	3 (8)	4 (11)	
II	20 (56)	23 (64)	
III	11 (31)	8 (22)	
IV	2 (6)	1 (3)	
LVEF[c], median (IQR)	39.7 (30.1-58.5)	34.7 (25.0-56.0)	.27
Laboratory test			
Hemoglobin (g/dL), mean (SD)	12.5 (2.5)	12.7 (2.6)	.63
eGFR[d] (mL/minute/1.73 m^2), median (IQR)	78.0 (51.5-94.0)	74.0 (53.8-91.0)	.84
NT-proBNP[e] (pg/mL; n=74), median (IQR)	2866 (1333-8198)	2058 (911-6651)	.30
Medication use at discharge, n (%)			
RAS[f] inhibitor	26 (67)	27 (71)	.87
Beta-blocker	29 (74)	26 (68)	.75
Mineralocorticoid receptor antagonist	20 (51)	26 (68)	.19
SGLT2[g] inhibitor	17 (44)	14 (36)	.71
Loop diuretic	23 (59)	31 (82)	.06

[a]HF: heart failure.
[b]NYHA: New York Heart Association.
[c]LVEF: left ventricular ejection fraction.
[d]eGFR: estimated glomerular filtration rate.
[e]NT-proBNP: N-terminal pro-B-type natriuretic peptide.
[f]RAS: renin-angiotensin system.
[g]SGLT2: sodium-glucose cotransporter 2.

Dyspnea Symptom Score

The dyspnea symptom scores assessed using the questionnaire are presented in Table 2 and Figure 4. The baseline dyspnea symptom scores were not significantly different between the 2 groups (intervention vs control: mean 2.4, SD 2.8 vs mean 2.0, SD 2.0; *P*=.46). Regarding the primary end point, change in dyspnea symptom score from baseline to 4 weeks was significantly greater in the intervention group than in the control group (mean –1.3, SD 2.1 vs mean –0.3, SD 2.3; *P*=.048).

XSL•FO
RenderX

Table 2. Dyspnea symptom scores, obtained using the questionnaire, according to the treatment groups (full analysis set).

Variables	Intervention (n=39), mean (SD)	Control (n=38), mean (SD)	Absolute difference (95% CI)	*P* value
Dyspnea symptom scores at baseline	2.4 (2.8)	2.0 (2.0)	0.5 (–0.7 to 1.5)	.46
Dyspnea symptom scores at 4 weeks	1.1 (1.7)	1.7 (2.1)	–0.6 (–1.5 to 0.3)	.19
Change in dyspnea symptom scores from baseline to 4 weeks	–1.3 (2.1)	–0.3 (2.3)	–1.0 (–2.0 to 0.0)	.048

Figure 4. Change in dyspnea symptom score in the intervention and control groups. (A) Dyspnea score during the trial follow-up. (B) Change in dyspnea score from baseline to 4 weeks. The error bars represent the 95% CI.

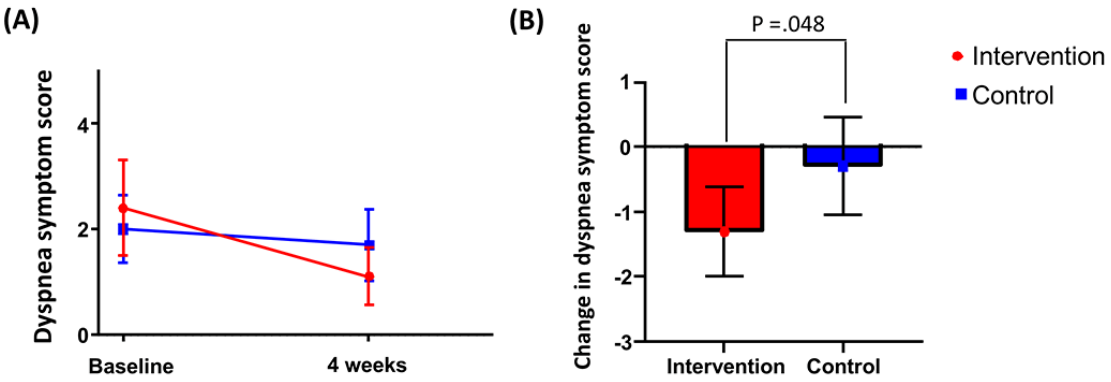

Secondary End Points

Secondary composite end points, including death, rehospitalization, and urgent visit for HF, were not significantly different between the 2 groups (3/39 vs 3/38; *P*>.99). A significant reduction was found in the level of body water composition from baseline to the last measurement in the intervention group (baseline level mean 7.4, SD 2.5 vs final level mean 6.6, SD 2.5; *P*=.003; Figure 5).

Figure 5. Change in body water composition level from baseline to last measurement in the intervention group.

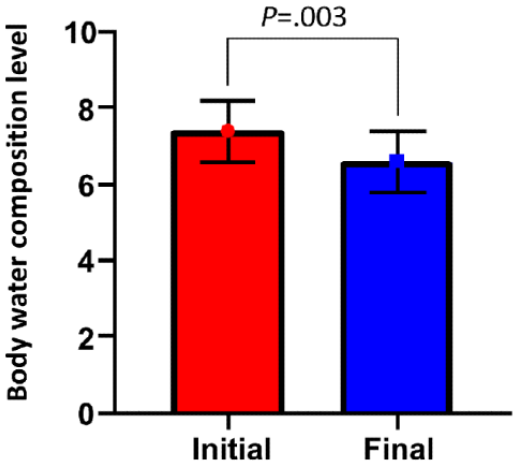

App Adherence and Patient Satisfaction

The distribution of app adherence among the treatment and control groups is presented in Figure 6. App adherence, which was assessed based on log-in access logs was higher in the intervention group than in the control group (80.0%, IQR 45.0%-96.8% vs 39.3%, IQR 20.6%-77.4%; *P*=.003). Additionally, the app adherence evaluated based on the percentage of days when symptom was entered into the app was significantly higher in the intervention group than in the control group (58.6%, IQR 19.2%-79.5% vs 30.6%, IQR 9.1%-64.3%; *P*=.046).

XSL•FO
RenderX

Figure 6. Distribution of app adherence according to the treatment groups. App adherence was assessed based on log in in (A) the intervention and (B) control groups. App adherence evaluated based on the percentage of days when symptoms were entered into the app in (C) the intervention and (D) control groups. IQR, interquartile range.

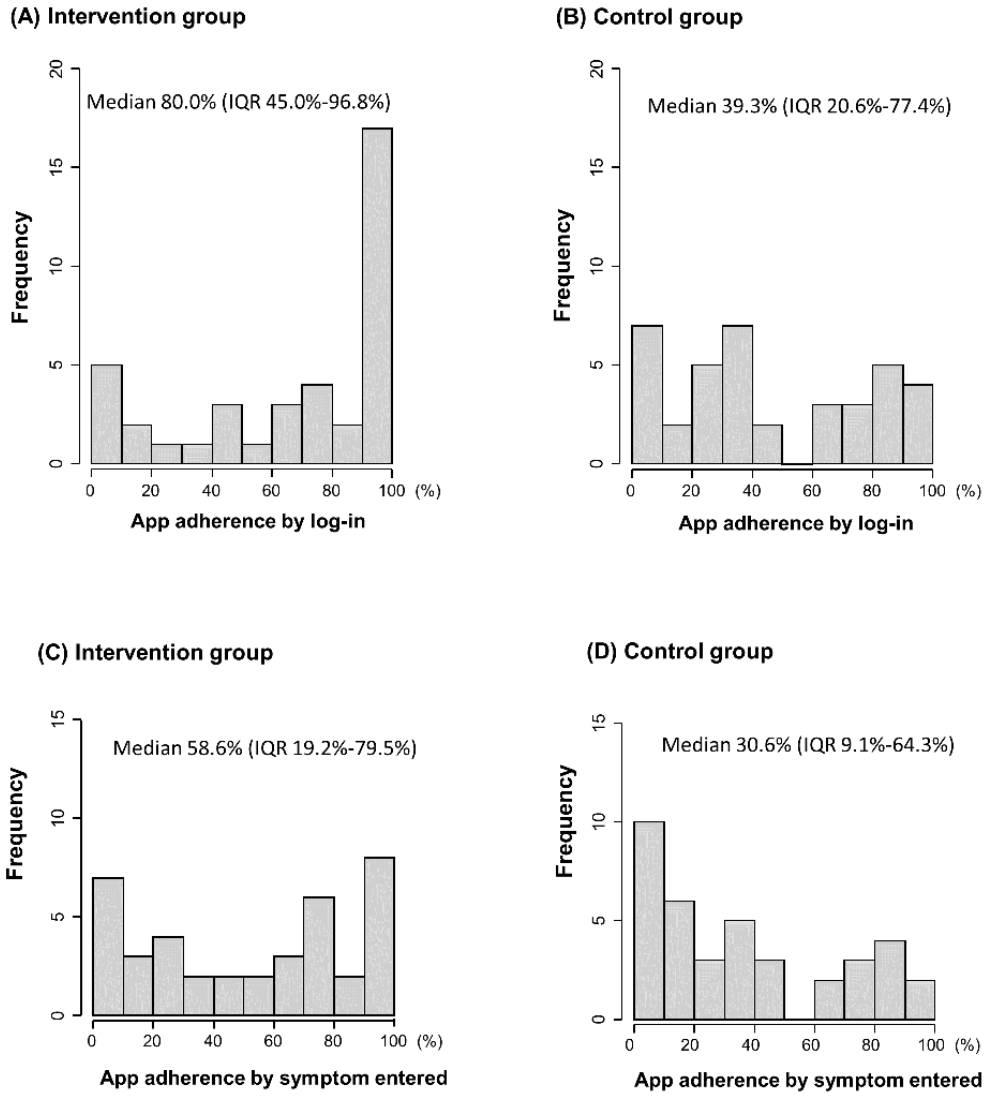

We also analyzed differences in app adherence according to age. Patients aged 65 years and older showed lower app adherence, as indicated by log-in access logs than those aged younger than 65 years in the overall population (45.9%, IQR 8.9%-79.2% vs 74.9%, IQR 35.7%-95.5%; *P*=.02). The intervention group showed similar results (61.4%, IQR 13.5%-89.5% vs 92.9%, IQR 64.2%-98.5%; *P*=.04); however, in the control group, no significant difference was found in app adherence according to the age cutoff of 65 years (39.4%, IQR 5.7%-76.7% vs 39.3%, IQR 27.8%-74.6%; *P*=.40).

App satisfaction scores at 4 weeks tended to be higher in the intervention group than in the control group; however, the difference was not statistically significant (mean 7.1, SD 2.6 vs mean 6.2, SD 2.9; *P*=.15).

Sensitivity Analysis

Overall, 74 patients (37 in the intervention group and 37 in the control group) completed the trial over 4 weeks (per-protocol analysis; Figure 3). Table S6 and Figure S6 in Multimedia Appendix 1 show the results of sensitivity analyses of the per-protocol population. There was a trend toward a greater reduction in dyspnea symptom scores from baseline to 4 weeks in the intervention group than in the control group, similar to the primary analyses (mean –1.4, SD 2.1 vs mean –0.5, SD 2.1; *P*=.08). Regarding body water composition, the intervention group showed a greater reduction in the level of body water composition from baseline to last measurement, consistent with the main analysis (baseline level mean 7.4, SD 2.5 vs final level mean 6.6, SD 2.6; *P*=.006).

Discussion

Principal Results

We developed a smartphone app and mHealth platform for patients with HF. The main findings of this study are (1) a mobile-based health platform with Bluetooth-connected monitoring devices and feedback improved dyspnea symptoms compared with the control group and (2) this app-based intervention led to an improvement in the level of body water composition and higher adherence to the app. We believe that

XSL•FO
RenderX

the findings of this prospective, randomized study will help in formulating public health strategies to improve HF outcomes using smartphones.

Support for facilitating self-care is crucial for patients with HF [18-20]. However, proper patient self-care and delivery of health care provider feedback are challenging to implement in a real-world clinical setting because of the required time, financial, and human resources. It is challenging for patients to know how to respond to changes in their symptoms or vital signs. It is also impossible for medical staff to quickly check patient information or provide feedback 24 hours a day, 7 days a week. These disadvantages can be overcome by self-management interventions via smartphones, including self-care, symptom and sign monitoring, and mobile-based feedback, which are easily accessible anytime, anywhere. Considering these advantages of a smartphone app–based intervention, including cost-effective and accessible methods, our group has previously developed mobile app platforms for patients with atrial fibrillation [35,36] or hypertension [37]. Therefore, we planned to develop an app for patients with HF.

Our mobile app for HF provides diverse functions using the concept of mHealth platforms for personalized interventions. Vital signs and body water can be paired with Bluetooth-connected devices, and patients can enter their HF symptoms into the apps. Patients can also receive feedback or alerts from the app based on the information they have entered. Additionally, patients can obtain information about HF using chatbots. Although direct interaction between patients and health care providers may be an ideal situation, it may not be feasible owing to rising costs and a shortage of medical staff. Our app-based monitoring and feedback system had the advantage of being accessible and quick for patients to receive feedback promptly on their changes without high personnel or cost. Although patients in the control group had access to the smartphone app for HF self-care, the improvement in dyspnea symptoms was more pronounced in the intervention group than in the control group. Patients in the control group could only use the limited functionality of the app, were unable to use Bluetooth devices, and could not receive any feedback or alerts. Therefore, our study suggests that the comprehensive functionality of apps, especially in the monitoring and feedback system, is particularly important for their role.

Comparison With Prior Work

Some previous studies have focused on mHealth apps for HF [24-30]. However, these studies were limited by the small number of participants and limited app functionality. Additionally, previous mobile apps have heterogeneous functions and platforms [24,25], and this diversity might have influenced the benefits of mobile apps in HF. Considering the inconsistent benefits of mobile apps in HF [25], we developed a mobile app with multiple functions in collaboration with 1 of the largest high-tech information technology companies in Korea (KT). In our opinion, this app has a considerably wider range of features and is more efficiently organized than other existing apps [28-30]. Our app has the advantage of keeping patients engaged in their condition and self-management after leaving the hospital. Not only does this provide psychological reassurance, but it also serves the patient's best interests by keeping a close eye on any changes in their clinical presentation or symptoms. Patients can assess their symptoms more frequently, make lifestyle changes, or contact health care providers earlier. The dashboard system also enables health care professionals to identify changes in a patient's condition quickly. Furthermore, the app contains elements that may be of interest to patients, including sodium analysis in food using AI and various HF information. Overall, we assumed that the monitoring and feedback function and app adherence are especially important for the mobile app's role in HF and believe that the benefits of our app could improve the symptoms of HF.

Increasing app adherence has been a common challenge in many studies. For example, a recent study showed that only approximately 30% of patients actively used the HF app, underscoring the importance of caution regarding the enrollment of critically ill, postacute, and older patients [30]. In our study, median adherence to apps based on log-in access logs was 80% and 39% in the intervention and control groups, respectively, which we believe should be further improved. In the intervention group, we believe that a more comprehensive functionality of apps, including Bluetooth-connected monitoring devices and a feedback system, may lead to higher adherence to apps than in the control group. Additionally, this disparity in app functionality may be associated with a higher trend in app satisfaction scores within the intervention group.

Particularly, as a large proportion of patients with HF tend to be older and app adherence appears to decline in those aged 65 years and older in our study, increasing adherence to apps for HF is crucial; this can be achieved by improving the platform to make it more user-friendly for older patients. Interestingly, at the cutoff age of 65 years, a difference was observed in app adherence according to age category in the intervention group but not in the control group, probably due to overall higher app adherence in the intervention group. In the subgroup analysis, no difference was found in the change of dyspnea symptom scores between the intervention and control groups according to an app adherence cutoff of 50%. This is probably due to the small sample size. Therefore, further research with larger study populations is warranted to confirm whether the frequency of app usage directly affects the outcomes.

Limitations and Strengths

This study had some limitations. First, our results are limited by the small sample size of the study. Second, we only enrolled patients who had smartphones and could use them; therefore, the results of our trial may not be applicable to patients who cannot use smartphones, such as very old adults. Since a large proportion of patients with HF are older and have difficulty using smartphones, excluding them from participation may lead to bias and study limitations. Third, the true engagement and use frequency of the apps and the real responses of participants to feedback could not be analyzed. There is a lack of clarity on exactly how the frequency of app use affects clinical outcomes. Additionally, there was a relatively large number of older patients, and adherence to the app declined toward the end of the study. These limitations are commonly observed in studies that use mobile apps. Fourth, we evaluated the dyspnea symptom

XSL•FO
RenderX

score using a visual analog scale and a numerical rating scale ranging from 0 to 10. However, at baseline, the average of these scores was <3 points, suggesting that this study was conducted on patients who already had achieved significant improvement in their dyspnea, which limits surveying on a scale of 0-10 over 4 weeks. Fifth, the HF medications could be modified at the discretion of each physician, which could be a confounding factor. Sixth, baseline diastolic BP was higher in the intervention group than in the control group, while patients in the intervention group tended to be younger. This difference in baseline characteristics may limit the generalizability of our conclusions. Seventh, the follow-up period in our studies was only 4 weeks, which was shorter than that in a similar study [29]; therefore, the effect of mobile apps on patients' symptoms for a long-term follow-up period was unknown. Furthermore, secondary clinical outcomes, including death, rehospitalization, and urgent visits for HF, were not significantly different between the 2 groups in this study, although a similar study using apps [28] for HF showed reduced hospitalization for HF by mobile app. Finally, we tested a single app, and our results may not be generalizable to other smartphone-based apps for patients with HF.

Despite these limitations, the major strength of this study is the use of a mobile app with various functions developed in collaboration with a cutting-edge technology company. We believe that the features of our app can be effective solutions for improving self-care in patients with HF.

Conclusions

The smartphone app–based intervention with Bluetooth-connected monitoring devices and feedback improved dyspnea symptoms among patients with HF compared with the control group. Considering the high cost of classical patient-provider interventions, self-care, and feedback through mobile apps are promising alternatives. Therefore, our study provides evidence and rationale for mobile app–based self-care and feedback for patients with HF.

Acknowledgments

The app used in this study was developed in collaboration between the Digital Health Council (DHC) from the Korean Society of Heart Failure and KT Corporation. This research was supported by KT's Internalization of Digital Therapeutics Business project. This research was supported by a grant from the Korea Health Technology Research & Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI21C1074).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
App characteristics and functionalities, and participants' clinical characteristics and outcomes.
[DOCX File , 970 KB - jmir_v26i1e52075_app1.docx]

Multimedia Appendix 2
CONSORT-EHEALTH checklist.
[PDF File (Adobe PDF File), 1066 KB - jmir_v26i1e52075_app2.pdf]

References

1. Shahim B, Kapelios CJ, Savarese G, Lund LH. Global public health burden of heart failure: an updated review. Card Fail Rev 2023;9:e11 [FREE Full text] [doi: 10.15420/cfr.2023.05] [Medline: 37547123]
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
3. Park JJ, Lee CJ, Park SJ, Choi JO, Choi S, Park SM, et al. Heart failure statistics in Korea, 2020: a report from the Korean Society of Heart Failure. Int J Heart Fail 2021;3(4):224-236 [FREE Full text] [doi: 10.36628/ijhf.2021.0023] [Medline: 36262554]
4. Conrad N, Judge A, Tran J, Mohseni H, Hedgecott D, Crespillo AP, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. Lancet 2018;391(10120):572-580 [FREE Full text] [doi: 10.1016/S0140-6736(17)32520-5] [Medline: 29174292]

XSL•FO
RenderX

5. Cho JY, Cho DH, Youn JC, Kim D, Park SM, Jung MH, et al. Korean Society of Heart Failure guidelines for the management of heart failure: definition and diagnosis. Int J Heart Fail 2023;5(2):51-65 [FREE Full text] [doi: 10.36628/ijhf.2023.0009] [Medline: 37180563]
6. Kimmoun A, Takagi K, Gall E, Ishihara S, Hammoum P, El Bèze N, et al. Temporal trends in mortality and readmission after acute heart failure: a systematic review and meta-regression in the past four decades. Eur J Heart Fail 2021;23(3):420-431 [FREE Full text] [doi: 10.1002/ejhf.2103] [Medline: 33443295]
7. Bueno H, Ross JS, Wang Y, Chen J, Vidán MT, Normand SLT, et al. Trends in length of stay and short-term outcomes among Medicare patients hospitalized for heart failure, 1993-2006. JAMA 2010;303(21):2141-2147 [FREE Full text] [doi: 10.1001/jama.2010.748] [Medline: 20516414]
8. Bozkurt B, Savarese G, Eryd SA, Bodegård J, Cleland JGF, Khordoc C, et al. Mortality, outcomes, costs, and use of medicines following a first heart failure hospitalization: EVOLUTION HF. JACC Heart Fail 2023;11(10):1320-1332 [FREE Full text] [doi: 10.1016/j.jchf.2023.04.017] [Medline: 37354145]
9. Maggioni AP, Dahlström U, Filippatos G, Chioncel O, Leiro MC, Drozdz J, et al. EURObservational Research Programme: regional differences and 1-year follow-up results of the Heart Failure Pilot Survey (ESC-HF Pilot). Eur J Heart Fail 2013;15(7):808-817 [FREE Full text] [doi: 10.1093/eurjhf/hft050] [Medline: 23537547]
10. Lee SE, Lee HY, Cho HJ, Choe WS, Kim H, Choi JO, et al. Clinical characteristics and outcome of acute heart failure in Korea: results from the Korean Acute Heart Failure Registry (KorAHF). Korean Circ J 2017;47(3):341-353 [FREE Full text] [doi: 10.4070/kcj.2016.0419] [Medline: 28567084]
11. Khan MS, Sreenivasan J, Lateef N, Abougergi MS, Greene SJ, Ahmad T, et al. Trends in 30- and 90-day readmission rates for heart failure. Circ Heart Fail 2021;14(4):e008335 [FREE Full text] [doi: 10.1161/CIRCHEARTFAILURE.121.008335] [Medline: 33866827]
12. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. Circulation 2015;131(4):e29-322 [FREE Full text] [doi: 10.1161/CIR.0000000000000152] [Medline: 25520374]
13. Rich MW, Beckham V, Wittenberg C, Leven CL, Freedland KE, Carney RM. A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure. N Engl J Med 1995;333(18):1190-1195 [FREE Full text] [doi: 10.1056/NEJM199511023331806] [Medline: 7565975]
14. Koelling TM, Johnson ML, Cody RJ, Aaronson KD. Discharge education improves clinical outcomes in patients with chronic heart failure. Circulation 2005;111(2):179-185 [FREE Full text] [doi: 10.1161/01.CIR.0000151811.53450.B8] [Medline: 15642765]
15. McAlister FA, Stewart S, Ferrua S, McMurray JJJV. Multidisciplinary strategies for the management of heart failure patients at high risk for admission: a systematic review of randomized trials. J Am Coll Cardiol 2004;44(4):810-819 [FREE Full text] [doi: 10.1016/j.jacc.2004.05.055] [Medline: 15312864]
16. Bhatia A, Maddox TM. Remote patient monitoring in heart failure: factors for clinical efficacy. Int J Heart Fail 2021;3(1):31-50 [FREE Full text] [doi: 10.36628/ijhf.2020.0023] [Medline: 36263114]
17. Jonkman NH, Westland H, Groenwold RHH, Ågren S, Atienza F, Blue L, et al. Do self-management interventions work in patients with heart failure? An individual patient data meta-analysis. Circulation 2016;133(12):1189-1198 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.018006] [Medline: 26873943]
18. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. Corrigendum to: 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 2021;42(48):4901 [FREE Full text] [doi: 10.1093/eurheartj/ehab670] [Medline: 34649282]
19. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: a report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. J Am Coll Cardiol 2022;79(17):e263-e421 [FREE Full text] [doi: 10.1016/j.jacc.2021.12.012] [Medline: 35379503]
20. Youn JC, Kim D, Cho JY, Cho DH, Park SM, Jung MH, et al. Korean Society of Heart Failure guidelines for the management of heart failure: treatment. Int J Heart Fail 2023;5(2):66-81 [FREE Full text] [doi: 10.36628/ijhf.2023.0011] [Medline: 37180564]
21. Cruz-Ramos NA, Alor-Hernández G, Colombo-Mendoza LO, Sánchez-Cervantes JL, Rodríguez-Mazahua L, Guarneros-Nolasco LR. mHealth apps for self-management of cardiovascular diseases: a scoping review. Healthcare (Basel) 2022;10(2):322 [FREE Full text] [doi: 10.3390/healthcare10020322] [Medline: 35206936]
22. Kalagara R, Chennareddy S, Scaggiante J, Matsoukas S, Bhimani A, Smith C, et al. Blood pressure management through application-based telehealth platforms: a systematic review and meta-analysis. J Hypertens 2022;40(7):1249-1256. [doi: 10.1097/HJH.0000000000003164] [Medline: 35762467]
23. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR Mhealth Uhealth 2018;6(7):e10723 [FREE Full text] [doi: 10.2196/10723] [Medline: 30037787]

XSL•FO
RenderX

24. Wali S, Demers C, Shah H, Wali H, Lim D, Naik N, et al. Evaluation of heart failure apps to promote self-care: systematic app search. JMIR Mhealth Uhealth 2019;7(11):e13173 [FREE Full text] [doi: 10.2196/13173] [Medline: 31710298]
25. Athilingam P, Jenkins B. Mobile phone apps to support heart failure self-care management: integrative review. JMIR Cardio 2018;2(1):e10057 [FREE Full text] [doi: 10.2196/10057] [Medline: 31758762]
26. Mortara A, Vaira L, Palmieri V, Iacoviello M, Battistoni I, Iacovoni A, et al. Would you prescribe mobile health apps for heart failure self-care? An integrated review of commercially available mobile technology for heart failure patients. Card Fail Rev 2020;6:e13 [FREE Full text] [doi: 10.15420/cfr.2019.11] [Medline: 32537246]
27. Giordan LB, Tong HL, Atherton JJ, Ronto R, Chau J, Kaye D, et al. The use of mobile apps for heart failure self-management: systematic review of experimental and qualitative studies. JMIR Cardio 2022;6(1):e33839 [FREE Full text] [doi: 10.2196/33839] [Medline: 35357311]
28. Sahlin D, Rezanezad B, Edvinsson ML, Bachus E, Melander O, Gerward S. Self-care management intervention in heart failure (SMART-HF): a multicenter randomized controlled trial. J Card Fail 2022;28(1):3-12 [FREE Full text] [doi: 10.1016/j.cardfail.2021.06.009] [Medline: 34161807]
29. Dorsch MP, Farris KB, Rowell BE, Hummel SL, Koelling TM. The effects of the ManageHF4Life mobile app on patients with chronic heart failure: randomized controlled trial. JMIR Mhealth Uhealth 2021;9(12):e26185 [FREE Full text] [doi: 10.2196/26185] [Medline: 34878990]
30. Zisis G, Carrington MJ, Oldenburg B, Whitmore K, Lay M, Huynh Q, et al. An m-Health intervention to improve education, self-management, and outcomes in patients admitted for acute decompensated heart failure: barriers to effective implementation. Eur Heart J Digit Health 2021;2(4):649-657 [FREE Full text] [doi: 10.1093/ehjdh/ztab085] [Medline: 36713108]
31. Park CS, Lee SE, Cho HJ, Kim YJ, Kang HJ, Oh BH, et al. Body fluid status assessment by bio-impedance analysis in patients presenting to the emergency department with dyspnea. Korean J Intern Med 2018;33(5):911-921 [FREE Full text] [doi: 10.3904/kjim.2016.358] [Medline: 29241303]
32. Park SM, Lee SY, Jung MH, Youn JC, Kim D, Cho JY, et al. Korean Society of Heart Failure guidelines for the management of heart failure: management of the underlying etiologies and comorbidities of heart failure. Korean Circ J 2023;53(7):425-451 [FREE Full text] [doi: 10.4070/kcj.2023.0114] [Medline: 37525389]
33. Zhang X, Zhao C, Zhang H, Liu W, Zhang J, Chen Z, et al. Dyspnea measurement in acute heart failure: a systematic review and evidence map of randomized controlled trials. Front Med (Lausanne) 2021;8:728772 [FREE Full text] [doi: 10.3389/fmed.2021.728772] [Medline: 34692723]
34. Placido R, Gigaud C, Gayat E, Ferry A, Cohen-Solal A, Plaisance P, et al. Assessment of dyspnoea in the emergency department by numeric and visual scales: a pilot study. Anaesth Crit Care Pain Med 2015;34(2):95-99. [doi: 10.1016/j.accpm.2014.09.001] [Medline: 25858616]
35. Kim IC, Lee JH, Choi DJ, Park SJ, Lee JH, Park SM, et al. Rationale design and efficacy of a smartphone application for improving self-awareness of adherence to edoxaban treatment: study protocol for a randomised controlled trial (adhere app). BMJ Open 2022;12(4):e048777 [FREE Full text] [doi: 10.1136/bmjopen-2021-048777] [Medline: 35477876]
36. Yoon M, Park JJ, Hur T, Hua CH, Shim CY, Yoo BS, et al. The ReInforcement of adherence via self-monitoring app orchestrating biosignals and medication of RivaroXaban in patients with atrial fibrillation and co-morbidities: a study protocol for a randomized controlled trial (RIVOX-AF). Front Cardiovasc Med 2023;10:1130216 [FREE Full text] [doi: 10.3389/fcvm.2023.1130216] [Medline: 37324622]
37. Choi DJ, Park JJ, Yoon M, Park SJ, Jo SH, Kim EJ, et al. Self-monitoring of blood pressure and feed-back using APP in TReatment of UnconTrolled Hypertension (SMART-BP): a randomized clinical trial. Korean Circ J 2022;52(10):785-794 [FREE Full text] [doi: 10.4070/kcj.2022.0133] [Medline: 36217600]

Abbreviations

AI: artificial intelligence
BP: blood pressure
CDSS: clinical decision support system
CONSORT: Consolidated Standards of Reporting Trials
HF: heart failure
HR: heart rate
mHealth: mobile health
NT-proBNP: N-terminal pro-B-type natriuretic peptide
SMART-HF: self-monitoring using a mobile app to improve symptoms and reduce rehospitalization and mortality in heart failure

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.08.23; peer-reviewed by O El-Gayar, P Athilingam; comments to author 18.10.23; revised version received 07.11.23; accepted 12.03.24; published 29.04.24.

Please cite as:
Yoon M, Lee S, Choi JY, Jung MH, Youn JC, Shim CY, Choi JO, Kim EJ, Kim H, Yoo BS, Son YJ, Choi DJ
Effectiveness of a Smartphone App–Based Intervention With Bluetooth-Connected Monitoring Devices and a Feedback System in Heart Failure (SMART-HF Trial): Randomized Controlled Trial
J Med Internet Res 2024;26:e52075
URL: https://www.jmir.org/2024/1/e52075
doi:10.2196/52075
PMID:38683665

©Minjae Yoon, Seonhwa Lee, Jah Yeon Choi, Mi-Hyang Jung, Jong-Chan Youn, Chi Young Shim, Jin-Oh Choi, Eung Ju Kim, Hyungseop Kim, Byung-Su Yoo, Yeon Joo Son, Dong-Ju Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Self-Management Platform for Adult Asthma: Randomized Attention-Placebo Controlled Trial

Aaron Kandola[1,2], BSc, MSc, PhD; Kyra Edwards[3], BSc, MSc; Joris Straatman[2], BA, MSc; Bettina Dührkoop[2], DipKFM; Bettina Hein[2], MA, MSc, Ref.iur; Joseph Hayes[2,3,4], MBChB, MSc, MRCPsych, PhD

[1]Medical Research Council Unit of Lifelong Health and Aging, University College London, London, United Kingdom
[2]juli Health, Hull, MA, United States
[3]Division of Psychiatry, University College London, London, United Kingdom
[4]Camden and Islington NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Joseph Hayes, MBChB, MSc, MRCPsych, PhD
Division of Psychiatry
University College London
Maple House
149 Tottenham Court Road
London, W1T 7BN
United Kingdom
Phone: 44 2089288300
Email: joseph.hayes@ucl.ac.uk

Abstract

Background: Asthma is one of the most common chronic conditions worldwide, with a substantial individual and health care burden. Digital apps hold promise as a highly accessible, low-cost method of enhancing self-management in asthma, which is critical to effective asthma control.

Objective: We conducted a fully remote randomized controlled trial (RCT) to assess the efficacy of juli, a commercially available smartphone self-management platform for asthma.

Methods: We conducted a pragmatic single-blind, RCT of juli for asthma management. Our study included participants aged 18 years and older who self-identified as having asthma and had an Asthma Control Test (ACT) score of 19 or lower (indicating uncontrolled asthma) at the beginning of the trial. Participants were randomized (1:1 ratio) to receive juli for 8 weeks or a limited attention-placebo control version of the app. The primary outcome measure was the difference in ACT scores after 8 weeks. Secondary outcomes included remission (ACT score greater than 19), minimal clinically important difference (an improvement of 3 or more points on the ACT), worsening of asthma, and health-related quality of life. The primary analysis included participants using the app for 8 weeks (per-protocol analysis), and the secondary analysis used a modified intention-to-treat (ITT) analysis.

Results: We randomized 411 participants between May 2021 and April 2023: a total of 152 (37%) participants engaged with the app for 8 weeks and were included in the per-protocol analysis, and 262 (63.7%) participants completed the week-2 outcome assessment and were included in the modified ITT analysis. Total attrition between baseline and week 8 was 259 (63%) individuals. In the per-protocol analysis, the intervention group had a higher mean ACT score (17.93, SD 4.72) than the control group (16.24, SD 5.78) by week 8 (baseline adjusted coefficient 1.91, 95% CI 0.31-3.51; P=.02). Participants using juli had greater odds of achieving or exceeding the minimal clinically important difference at 8 weeks (adjusted odds ratio 2.38, 95% CI 1.20-4.70; P=.01). There were no between group differences in the other secondary outcomes at 8 weeks. The results from the modified ITT analyses were similar.

Conclusions: Users of juli had improved asthma symptom control over 8 weeks compared with users of a version of the app with limited functionality. These findings suggest that juli is an effective digital self-management platform that could augment existing care pathways for asthma. The retention of patients in RCTs and real-world use of digital health care apps is a major challenge.

Trial Registration: International Standard Randomised Controlled Trial Number (ISRCTN) registry ISRCTN87679686; https://www.isrctn.com/ISRCTN87679686

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50855) doi:10.2196/50855

KEYWORDS

asthma; mobile health; self-management; randomized controlled trial; randomized; controlled trial; controlled trials; RCT; RCTs; respiratory; pulmonary; smartphone; platform; digital health; chronic; breathing; disease management; mHealth; mobile health; app; apps; application; applications; mobile phone

Introduction

Asthma is one of the most common chronic conditions worldwide, with an increasing prevalence that currently affects 1 in 10 people at some time [1-3]. The inflammatory disease causes mild-to-severe respiratory symptoms, including shortness of breath, chest tightness, wheezing, and cough. It significantly burdens patients and health care services, including the need for long-term treatment, emergency care, and hospitalizations that will cost the US economy an estimated US $300 billion over the next 20 years in direct health care expenditure [4]. Effective asthma control is necessary to reduce these costs and improve the quality of life for people with the condition.

Asthma management is based on achieving symptom control and reducing the frequency and severity of exacerbations [5]. This involves the use of inhaled anti-inflammatory medications and the avoidance of asthma triggers. Symptom control is associated with improved quality of life, reduced health care costs, and better work performance [6]. However, a significant proportion of individuals with asthma have suboptimal control because of poor adherence to medication, insufficient recognition of triggers, comorbidities (such as rhinitis or obesity), health behaviors (such as smoking), and inadequate information about treatment [7]. Mobile apps may address some of these treatment challenges by enabling people with asthma to more easily and consistently self-manage their condition compared to existing treatment plans. For example, digital apps can offer timely reminders to improve medication adherence or real-time feedback to identify and adapt to possible triggers and health behaviors [8,9].

A 2017 systematic review and meta-analysis of randomized controlled trials (RCTs) of mobile, web-based, and messaging service apps to support asthma self-management [9] concluded that these interventions could improve asthma control, but that effectiveness and important features of the apps varied. The majority of these apps included combinations of medication prompts, patient education, digital diaries, action plans, and professional support facilitation [9]. A similar 2018 review of RCTs and observational studies concluded that, in adults with asthma, mobile apps were more effective than other types of digital interventions, such as web-based interventions [10]. Studies of app-based interventions published since these reviews have generally been feasibility trials or small underpowered RCTs [11-14]. A 2022 Cochrane review examined the effect of digital apps on asthma medication adherence, concluding they were likely to be useful in poorly adherent populations, but again highlighting heterogeneity among mobile or web-based interventions [8]. Despite the mixed evidence for effectiveness, several apps are publicly available. These apps frequently incorporate behavior change techniques and gamification. Reviews of these apps have highlighted that they vary considerably in quality, use a range of behavior change techniques, struggle with adequate engagement and retention, and lack clinical validation of efficacy [15-17]. The *Global Strategy for Asthma Management and Prevention* (Global Initiative for Asthma [GINA]) highlights that, despite the use of digital technologies rapidly increasing in patients with asthma, “high-quality studies are needed to evaluate their utility and effectiveness” [3].

We aimed to address the fundamental issue that commercially available apps require sufficient evaluation of their effectiveness, by conducting an RCT of juli. This is a digital health app that aims to support people with asthma by combining numerous approaches that have been shown as effective in research-grade apps for asthma, including symptom tracking; medication reminders; trigger identification (including geolocated weather, pollen, and air pollution data); data visualization of respiratory symptoms, mood, exercise, activity, sleep, and heart rate variability; and behavioral activation recommendations about how to improve these parameters [18,19]. Our RCT was fully remote, increasing time efficiency, cost-effectiveness, and reach. We hypothesized that participants randomized to juli would have a greater reduction in asthma symptoms at 8 weeks than those randomized to the attention-placebo control.

Methods

Study Design and Participants

We conducted a fully remote pragmatic single-blind, placebo-control RCT to test the efficacy of juli in adults with asthma. The trial was open to individuals from anywhere in the world, provided they were aged 18 to 65 years, English-speaking, had access to a smartphone, and self-identified as having asthma. We also only included people with asthma symptoms that were uncontrolled according to a score of 19 or lower on the Asthma Control Test (ACT) at baseline. An ACT score of 19 is consistent with GINA-defined uncontrolled asthma [20].

Recruitment

Recruitment ran from May 2021 until April 2023. We recruited via self-help groups for asthma, online adverts, and social media posts. For the duration of the RCT, we modified the onboarding so that recruitment was automated with study information provided to participants within the app. The support for potential participants interested in the RCT was provided by this study’s team via email.

Ethical Considerations

The University College London Ethics Committee gave full ethical approval (19413/001). All participants supplied written informed consent within the app, with additional information on a dedicated web page. Data required for the RCT were stored

XSL•FO
RenderX

separately in an anonymized format. The juli app is Health Insurance Portability and Accountability Act, Service Organization Control Type 2, and General Data Protection Regulation compliant. Participants in both arms of the RCT were entered into a prize draw at 2, 4, 6, and 8 weeks with the possibility of winning US $20 at each time point. The trial was entered on the International Standard Randomised Controlled Trial Number (ISRCTN) registry (ISRCTN87679686). At the same time, we were running an RCT of the juli app for depression. This RCT had a similar design and analysis [21].

Randomization and Masking

We assigned participants in a 1:1 ratio to either an attention-placebo control or the full version of juli. We automated and conducted randomization within the app, using random block sizes ranging from 4 to 8. To ensure data integrity, the treatment allocation was concealed from both the research team and independent statisticians until the analysis was finalized.

Intervention

The juli app was developed by gamification experts in collaboration with patients, a psychiatrist, and a pulmonologist. A patient with asthma and a psychiatrist with expertise in mental health and physical health interface are the chief technical officer and chief medical officer of juli, respectively. We held development and user testing interviews with 10 patients with asthma (5 female individuals, aged 18-65 years). The app underwent multiple iterations following feedback from these patient panel interviews and discussions with a pulmonologist. Our trial used a full version of the juli app for the intervention group and a limited version in the attention-placebo control group. Participants with the complete juli app received automatic prompts to open the app each day at a user-inputted time. The app asked participants about how their asthma was affecting them on a 5-face emoji scale, their emergency inhaler use that day, how often they had a shortness of breath episode, and whether they woke in the night due to shortness of breath. Individuals could also track various factors they regarded as relevant to their asthma symptoms, such as tobacco smoke exposure [19]. The app connects to smart peak flow meters (such as Smart Peak Flow [Smart Respiratory Products Ltd] or MIR Smart One [Smart One]) through Google Fit or Apple HealthKit, or participants could enter this information manually.

The app presented participants with regular, geolocated weather, pollen, and air pollution data relevant to their asthma [22]. All participants could also access passively gathered smartphone data on relevant health-related factors, including, activity, menstrual cycle, and sleep. Participants could check this information daily and see associations with their asthma [23-25]. If they had them, participants with wearables that they chose to connect to the app would see additional data on workouts and heart rate variability, as well as improved data on activity and sleep. However, the lack of access to a wearable was not an exclusion criterion.

The app also uses behavioral activation techniques to provide personalized recommendations about these factors to encourage healthy behaviors. The app includes customizable medication reminders to improve medication adherence [26]. The juli app also encouraged participants to use the positive affect journaling function [27]. The design of the juli app guides participants toward all elements of the app but allows them to flexibly choose where they want to engage.

Attention-Placebo Control

Participants in the control arm had a limited version of the app. The app prompted participants to open it each day and rate how they were feeling on the 5 emoji scale, but they did not have access to any further functionality or intervention. There was no change to usual care in either arm.

Assessment Tools

Participants in both arms completed baseline assessments and follow-up assessments at 2, 4, 6, and 8 weeks remotely from within the app. Assessments included the ACT for asthma symptoms and the 12-Item Short Form Health Survey (SF-12) for health-related quality of life. The ACT is a widely used, self-completed asthma symptom scale that is responsive to change with scores ranging from 5 to 25 [28]. A cutoff score of 19 or lower identifies patients with uncontrolled asthma. The SF-12 is a self-reported measure assessing the impact of health on an individual's everyday life. Scores ranging from 0 to 100 with higher scores indicate a better quality of life [29].

Outcomes

The total ACT score at 8 weeks was our primary outcome. Secondary outcomes were continuous ACT score at 2, 4, 6, and 8 weeks in a repeated measures analysis using mixed-effect models; remission, defined as a score of >19 at 8 weeks; remission at 2, 4, 6, and 8 weeks in a repeated measures analysis; SF-12 physical and mental component scores at 8 weeks; and SF-12 physical and mental component scores at 4 and 8 weeks in a repeated measures analysis.

We added achieving a minimal clinically important difference (MCID) at 8 weeks (a 3-point increase on the ACT) [30] and a worsening of asthma symptoms (ie, a decrease in ACT scores from baseline) as post hoc outcomes.

Sample Size Estimation

The best MCID estimate for the ACT is between 2.2 and 3.0 (SD 3.1 to 4.7) [30]. A 2-sided 5% significance level at 80% power requires a total sample size of 146 for an MCID of 3. We aimed to recruit 90 participants per arm, allowing for 23% attrition [31].

Statistical Analyses

We preprinted the analysis plan on UCL Discovery [32] and preregistered the RCT on the ISRCTN registry with a description of the primary and secondary outcomes before the trial started. In reporting and analyzing our data we followed the CONSORT (Consolidated Standards of Reporting Trials) guidelines [33].

Our primary outcome was the difference in total ACT score at 8 weeks between the control and intervention groups in a per-protocol analysis. We estimated this difference with a linear regression model adjusted for baseline ACT and any imbalanced baseline covariates. We tested how robust the result was to model specification by also using a Poisson model and adjusting

XSL•FO
RenderX

for any variables not balanced at baseline. We used logistic regression to calculate the odds ratio (OR) of remission at 8 weeks (ACT>19), achieving MCID (≥3 point ACT improvement), and worsening of asthma, adjusting for baseline ACT. We completed the repeat measures analyses using linear or logistic mixed-effect models adjusting for ACT at baseline.

We repeated the analysis of all outcomes in a modified intention-to-treat (ITT) analysis. This analysis included all randomized participants with a complete baseline and week 2 ACT score, dropping participants who were randomized but never used the app (see Figure 1). We imputed the missing ACT scores first using multiple imputation models and then using the last observation carried forward [34]. The multiple imputation models included predictive mean matching with 5 nearest neighbors and 50 iterations. This method means that only plausible values are imputed and is more robust to model misspecification than fully parametric imputation [35].

An independent statistician with no conflicts of interest with the company providing juli completed the analyses. All analyses we conducted using Stata (version 17; StataCorp) and R (version 4.3.1 for Windows; R Foundation for Statistical Computing).

Figure 1. CONSORT diagram. ACT: Asthma Control Test; CONSORT: Consolidated Standards of Reporting Trials.

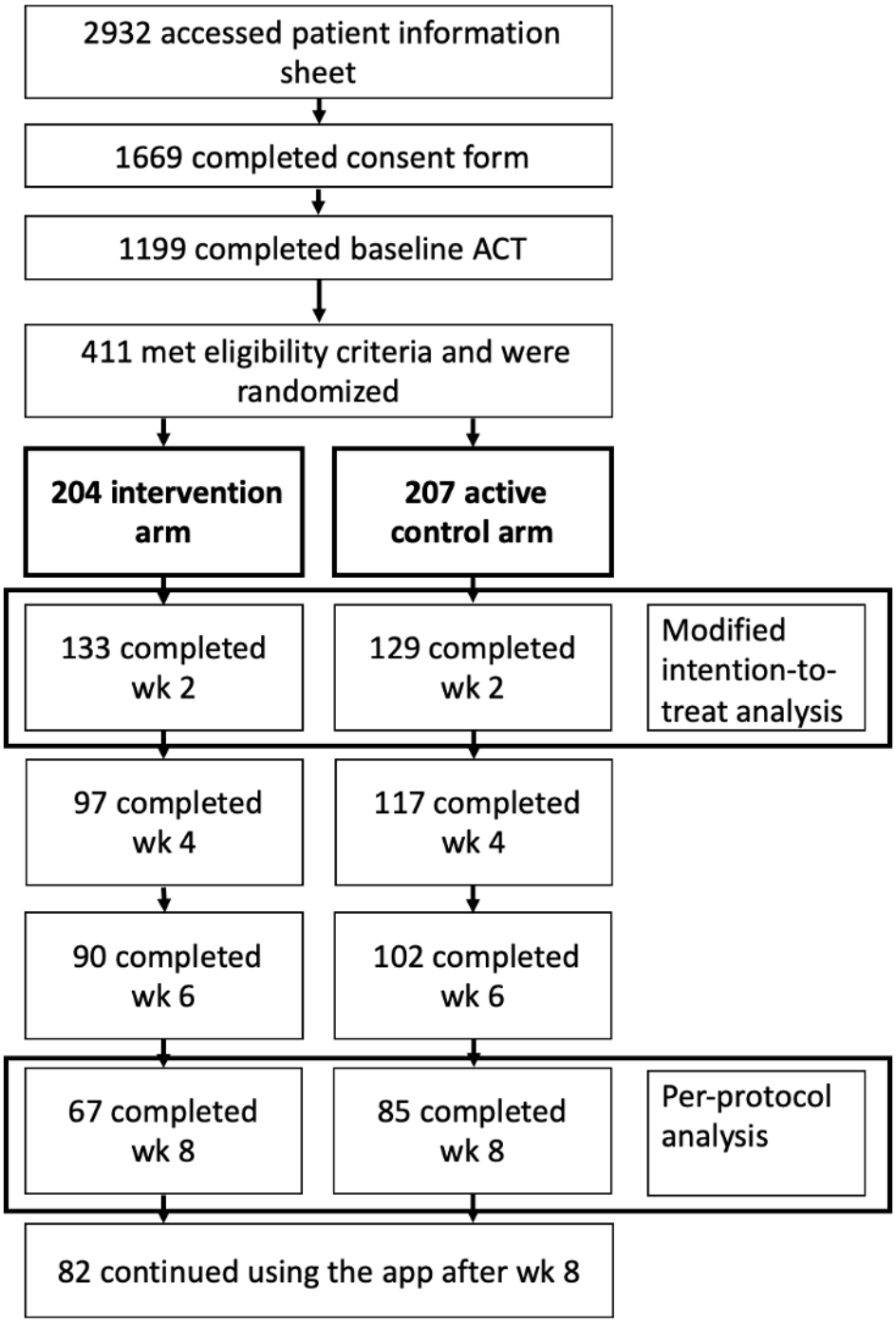

XSL•FO
RenderX

Results

Participants

Of 1199 participants who completed the baseline ACT, 411 (34.3%) participants met eligibility criteria. The 411 participants were randomized: 204 (49.6%) to the intervention arm and 207 (50.4%) to the active control arm. Of the 411 participants randomized, 325 (79.1%) were from the United States. Attrition was similar in both arms: 71 (34.8%) out of 204 participants in the intervention arm and 78 (37.7%) out of 207 participants in the active control arm left this study before the week-2 ACT. The remaining 262 participants contributed to our modified ITT analysis (Figure 1). Further attrition occurred between week 2 and week 8: a total of 66 (49.6%) out of 133 remaining participants left the intervention group, and 44 (34.1%) out of 129 remaining participants left the active control group. The remaining 152 participants contributed to our per-protocol analysis (Figure 1). Participants included in the modified ITT and per-protocol analyses were similar in terms of baseline characteristics (see Table 1 and Multimedia Appendix 1).

Table 1. Baseline characteristics of participants completing 8 weeks of the trial, by randomization group. Data used in the per-protocol analysis of individuals completing the week-8 ACT.

Characteristics	Intervention (n=67)	Control (n=85)	All (n=152)
Age (years), mean (SD)	35.73 (11.48)	36.62 (13.23)	36.23 (12.45)
Sex, n (%)			
Female	52 (77.6)	70 (82.4)	122 (80.3)
Male	13 (19.4)	13 (15.3)	26 (17.1)
Other	2 (3)	2 (2.4)	4 (2.6)
Asthma duration, n (%)			
<1 month	1 (1.5)	1 (1.2)	2 (1.3)
1 to <3 months	2 (3)	2 (2.4)	4 (2.6)
3 months to <1 year	2 (3)	4 (4.7)	6 (4)
1 year to <2 years	4 (6)	5 (5.9)	9 (5.9)
2 years to <5 years	11 (16.4)	5 (5.9)	16 (10.5)
>5 years	47 (70.2)	68 (80)	115 (75.7)
Physician contact, n (%)			
Regular	22 (32.8)	30 (35.3)	52 (34.2)
Occasional	42 (62.7)	40 (47.1)	82 (54)
Not anymore	3 (4.5)	10 (11.9)	13 (8.6)
Never	0 (0)	5 (5.9)	5 (3.3)
Diagnosed by a physician, n (%)			
Yes	67 (100)	81 (95.3)	148 (97.4)
No	0 (0)	4 (4.7)	4 (2.6)
ACT[a] total score, mean (SD)	12.60 (4.10)	13.04 (3.93)	12.84 (4.00)
SF-12 physical health subscale[b], mean (SD)	39.28 (8.95)	39.86 (9.04)	39.61 (8.98)
SF-12 mental health subscale[c], mean (SD)	38.35 (9.84)	37.74 (10.74)	38.01 (10.32)

[a]ACT: Asthma Control Test (possible range 5-25).

[b]SF-12 physical health subscale: Short-Form Health Survey-12 physical health subscale (possible range 0-100).

[c]SF-12 mental health subscale: Short-Form Health Survey-12 mental health subscale (possible range 0-100).

Per-Protocol Analysis

Of the 152 participants in the per-protocol analysis, they were mostly female individuals (n=122, 80.3%) who had been diagnosed by a physician more than 5 years ago (n=115, 75.7%) and had ongoing contact with a doctor about their asthma (n=134, 88.2%; Table 1). Participants had a mean baseline ACT score of 12.84 (SD 4.00).

Intervention group participants had a mean ACT score of 17.93 (SD 4.72) compared with 16.24 (SD 5.78) in the control group after 8 weeks (see Figure 2). After adjusting for baseline ACT score, the intervention group showed a greater improvement in symptom scores at 8 weeks than those in the control group (adjusted coefficient 1.91, 95% CI 0.31-3.51; *P*=.02; Table 2). After adjusting for imbalanced baseline characteristics, the improvement was 2.01 (95% CI 0.48-3.53; *P*=.01) points on

XSL•FO
RenderX

the ACT. Using Poisson regression rather than linear regression did not alter our results.

The chance of being in remission by week 8 did not differ between the intervention and control groups after accounting for baseline asthma. However, participants in the intervention group were more likely to experience an MCID (adjusted OR 2.38, 95% CI 1.20-4.70; *P*=.01) than those in the control group. This effect was consistent across the 2-, 4-, 6-, and 8-week assessments (Table 2). The odds of worsening symptoms were similar in both arms (adjusted OR 0.55, 95% CI 0.23-1.32, *P*=.18). There were no between-group differences in SF-12 mental or physical component scores.

Figure 2. Mean change in ACT score over 8 weeks. ACT: Asthma Control Test.

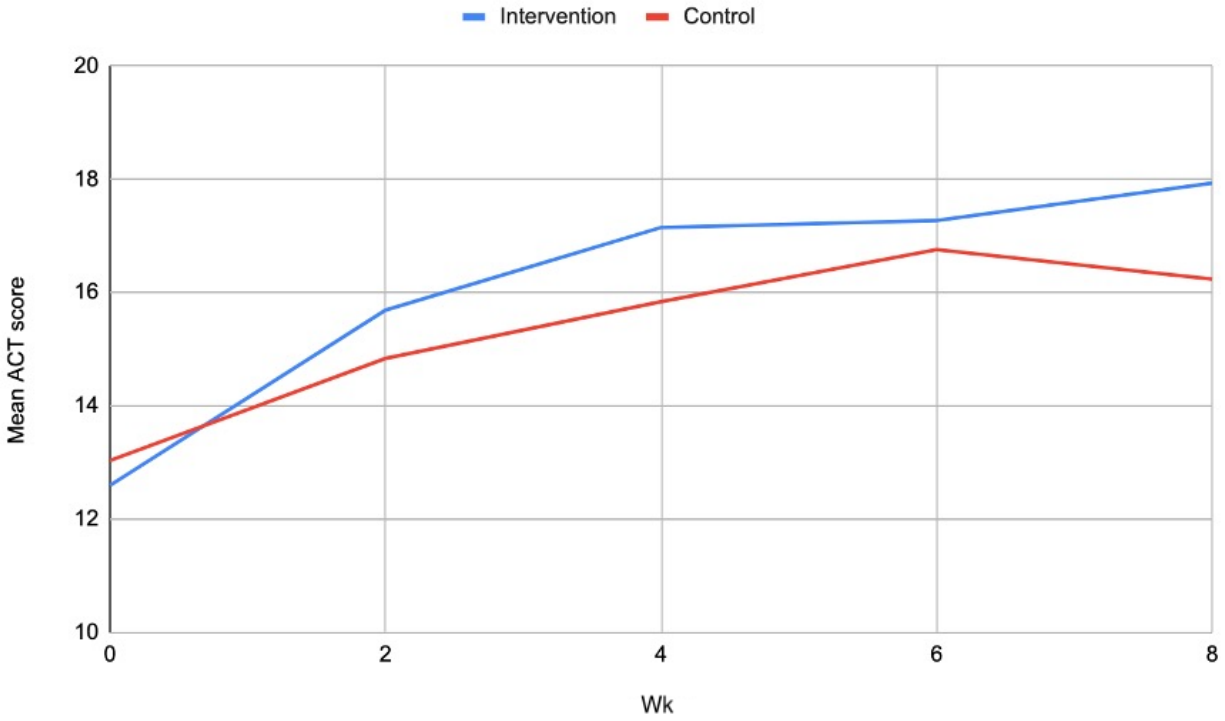

XSL•FO
RenderX

Table 2. Outcomes of the per-protocol and modified intention-to-treat analyses.

Characteristics	Effect estimate (95% CI)	*P* value
Per-protocol (n=152)		
ACT[a] at 8 weeks[b]	1.91 (0.31 to 3.51)	.02
Remission at 8 weeks[c]	1.47 (0.75 to 2.90)	.26
ACT repeated measures (2, 4, 6, and 8 weeks)[b]	1.34 (0.15 to 2.53)	.03
Remission repeated measures (2, 4, 6, and 8 weeks)[c]	1.67 (0.62 to 4.52)	.31
SF-12[d] physical component score at 8 weeks[b]	0.81 (–1.49 to 3.10)	.49
SF-12 mental component score at 8 weeks[b]	0.84 (–1.97 to 3.65)	.56
SF-12 physical component score repeated measures (4 and 8 weeks)[b]	0.611 (–1.32 to 2.54)	.53
SF-12 mental component score repeated measures (4 and 8 weeks)[b]	0.91 (–1.21 to 3.03)	.40
MCID[e] at 8 weeks[c]	2.38 (1.20 to 4.70)	.01
Worse at 8 weeks[c]	0.55 (0.23 to 1.32)	.18
ITT[f], multiple imputation for missing outcomes (n=262)		
ACT at 8 weeks[b]	1.56 (0.32 to 2.79)	.01
Remission at 8 weeks[c]	1.45 (0.80 to 2.63)	.22
ACT repeated measures (2, 4, 6, and 8 weeks)[b]	1.23 (0.33 to 2.12)	.007
Remission repeated measures (2, 4, 6, and 8 weeks)[c]	1.40 (0.85 to 2.32)	.19
SF-12 physical component score at 8 weeks[b]	0.58 (–1.45 to 2.60)	.58
SF-12 mental component score at 8 weeks[b]	0.73 (–1.75 to 3.22)	.56
SF-12 physical component score repeated measures (4 and 8 weeks)[b]	0.75 (–1.15 to 2.64)	.44
SF-12 mental component score repeated measures (4 and 8 weeks)[b]	0.23 (–1.93 to 2.40)	.83
MCID at 8 weeks[c]	2.17 (1.25 to 3.78)	.006
Worse at 8 weeks[c]	0.76 (0.39 to 1.56)	.45
ITT, last observation carried forward for missing outcomes (N=262)		
ACT at 8 weeks[b]	1.17 (0.02 to 2.31)	.046
Remission at 8 weeks[c]	0.92 (0.53 to 1.57)	.75
ACT repeated measures (2, 4, 6, and 8 weeks)[b]	1.03 (0.12 to 1.93)	.03
Remission repeated measures (2, 4, 6, and 8 weeks)[c]	0.88 (0.41 to 1.93)	.76
SF-12 physical component score at 8 weeks[b]	0.17 (–1.67 to 2.02)	.85
SF-12 mental component score at 8 weeks[b]	0.61 (–1.63 to 2.85)	.59
SF-12 physical component score repeated measures (4 and 8 weeks)[b]	–0.13 (–2.21 to 1.95)	.90
SF-12 mental component score repeated measures (4 and 8 weeks)[b]	0.76 (–1.69 to 3.21)	.54
MCID at 8 weeks[c]	1.95 (1.17 to 3.24)	.01
Worse at 8 weeks[c]	0.65 (0.35 to 1.21)	.17

[a]ACT: Asthma Control Test.

[b] coefficient.

[c]Odds ratio.

[d]SF-12: 12-Item Short Form Health Survey.

XSL•FO
RenderX

[e]MCID: minimal clinically important difference.

[f]ITT: intention-to-treat.

ITT Analysis

The baseline characteristics of participants in the intervention and control groups were similar to the per-protocol analysis. Following multiple imputations of missing outcomes, there was a greater improvement in ACT scores in the intervention group than in the active control group (adjusted coefficient 1.56, 95% CI 0.32-2.79; *P*=.01; Table 2). MCID was more common in the intervention group than the control group (adjusted OR 2.17, 95% CI 1.25-3.78, *P*=.006). Both arms had similar odds of remission, worsening of symptoms, and SF-12 scores. The results from the last observation carried forward analyses were consistent with the per-protocol and multiply imputed results.

Discussion

Principal Findings

Our primary analysis showed that juli users had a greater improvement in asthma symptoms at 8 weeks compared to an attention-placebo control. The mean improvement in the intervention group was 5.33 (SD 5.33) compared with 3.20 (SD 5.26) in the control group. This total improvement and the difference between arms are consistent with a clinically important effect of juli on asthma control [30]. Participants assigned to juli had more than twice the odds of a 3-point (MCID) or greater improvement on the ACT. However, the mean ACT score at 8 weeks in both arms fell below the established cut point for "well-controlled" asthma, and there was no difference between arms in terms of odds of remission. The results from our multilevel models covering outcomes from 2 to 8 weeks and the modified ITT analysis with all individuals who were randomized and used the app for at least 2 weeks were consistent with these primary findings.

Participants entering our trial had a mean baseline ACT score of 12.84 (SD 4.00), indicating they fulfilled the GINA definition of "very poorly controlled" (score of 13) as uncontrolled is scores of 19 or lower [20], and most reported having asthma for several years with routine physician contact, suggesting difficulties with long-term asthma control. The results of this trial indicate that juli can augment the treatment of uncontrolled asthma as indicated by improved ACT scores over 8 weeks. There is consistent evidence that low ACT scores are associated with rescue medication use, asthma exacerbations, reduced lung function, and reduced asthma-specific quality of life, sleep, work, and productivity [6]. Increases in ACT scores are associated with decreased health care usage and health care costs [6].

It is unclear which component of juli resulted in improved ACT scores, but participants likely chose elements that suited them, which is a strength of juli's design, allowing for a degree of self-personalization. Previous research into asthma app functionality has highlighted symptom tracking, clinical questionnaires, goal setting, performance feedback, medication reminders, and tracking as valuable to patients [17]. Gamification and contingent rewards are also important features incorporated into juli [17]. Positive affect journaling is a novel, evidence-based addition to juli's functionality [36]. Other commercially available apps for adult asthma self-management use similar behavior change techniques, health education, symptom recording, environmental data, medication reminders, and data presentation. A recent review identified over 500 asthma-related mobile and inhaler-based monitoring apps [37]. However, only a small number of these had any degree of scientific evaluation; with positive fully powered trials being rare [37]. An additional problem for patients is the high rate of failure of companies providing these apps, with only a small number with evidence being available currently. These include AsthmaMD (AsthmaMD Inc) [14], Kiss myAsthma (University of Sydney, the Woolcock Institute of Medical Research, and The University of Melbourne) [38], ASTHMAXcel (ASTHMAXcel) [39], and eAMS (EAPOC [Evidence at the Point-of-Care]) [40], each having positive pilot data.

The juli app is available in Android and iOS formats globally. It is a highly accessible platform for people with asthma, and our trial provides methodologically robust evidence of its efficacy in managing asthma. Additional research is required to understand the most cost-effective support procedures to improve adherence to digital self-management tools and how best to integrate them into clinical practice. The majority of the early attrition in our RCT was in participants who never began to use the app. To reduce this, future RCTs of digital interventions may benefit from a run-in period, in which participants become familiarized with the app before randomization [41].

Strengths and Limitations

There were several strengths and limitations to this RCT. We successfully and remotely recruited, screened, randomized, treated, and assessed participants worldwide. People could easily participate in the trial as our modified version of the juli app allowed consent, randomization, and assessments to occur within the platform. This facilitated a low-cost global recruitment strategy and a pragmatic trial design with good external validity. However, our focus on reducing participant burden limited the types and richness of the data we were able to collect at baseline. For example, we lack relevant information on income, education, and other social determinants of health. Despite this, we did achieve a postrandomization balance in recorded characteristics at baseline, indicating successful randomization. Most of the participants were female, reflecting established differences in sex-specific rates of asthma [42], health behaviors, and health care use in adults [43].

Participants completed the ACT, which is a recommended primary end point in clinical trials for asthma [6]. We also preregistered our primary and secondary outcome measures along with a full analysis plan, which we adhered to. However, we lacked a broader battery of outcome measures that could have further contextualized our findings and identified possible mechanisms of action.

XSL•FO
RenderX

Attrition was greater than we predicted. The attrition in our trial follows a similar pattern to other digital RCTs, including for asthma apps, where it mostly occurs between randomization and week 2. Dropout rates in previous RCTs have ranged from 20% to 60% [10]. However, studies recruiting via social media have had low retention at 30 days (<20%) [44], and a similar, all-remote RCT of mobile health support for asthma had an attrition of 62% at 9 weeks [45]. To manage attrition, we continued recruiting and randomizing participants until we had a sufficient number of participants completing the week 8 outcome measures to meet our sample size calculation. We examined differences in completers versus noncompleters (see Table 1 and Multimedia Appendix 1). There were unlikely to be differences between those who dropped out of this study and those who completed it, based on their baseline characteristics, including asthma severity. Our modified ITT and primary analysis findings were similar, suggesting the intervention would have had a similar effect on those who dropped out. The ITT analysis used 2 imputation methods that make different assumptions [34], and results were consistent using both methods. Despite this, it is impossible to rule out attrition bias and our results should be seen as reflecting the effect on people motivated and able to use juli.

Conclusions

The juli app has been demonstrated to decrease asthma symptoms within 8 weeks, with an increased chance of achieving MCID, but no difference in terms of odds of remission. As such, juli represents a low-risk and low-cost adjunct to the care regimen of individuals with asthma.

Acknowledgments

AK is supported by the UK Research and Innovation (UKRI) Digital Youth program award (MRC project reference MR/W002450/1), which is part of the Arts and Humanities Research Council/Economic and Social Research Council/Medical Research Council (AHRC/ESRC/MRC) Adolescence, Mental Health and the Developing Mind program. JH is supported by the UK Research and Innovation grant MR/V023373/1, the University College London Hospitals NIHR Biomedical Research Centre, and the NIHR North Thames Applied Research Collaboration. This study was funded by juli Health.

Data Availability

All data produced in this study are available upon reasonable request to the authors.

Authors' Contributions

JH conceived this study. JH, AK, BD, BH, and JS designed this study. AK, BD, and BH collected the data. KE analyzed the data. JH wrote the initial draft. All authors edited and approved the final paper.

Conflicts of Interest

AK, BD, BH, JS, and JH are shareholders in juli Health. AK has received consultancy fees from juli Health and Wellcome Trust. BD, BH, JS, and JH are cofounders of juli Health. JH has received consultancy fees from juli Health and Wellcome Trust. KE has no conflicts of interest. The funders played no part in the analysis of the data.

Multimedia Appendix 1
Baseline characteristics of individuals in the modified intention-to-treat analysis.
[DOCX File , 17 KB - jmir_v26i1e50855_app1.docx]

Multimedia Appendix 2
CONSORT-eHealth checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1252 KB - jmir_v26i1e50855_app2.pdf]

References

1. Nunes C, Pereira AM, Morais-Almeida M. Asthma costs and social impact. Asthma Res Pract 2017;3:1 [FREE Full text] [doi: 10.1186/s40733-016-0029-3] [Medline: 28078100]
2. Song P, Adeloye D, Salim H, Dos Santos JP, Campbell H, Sheikh A, et al. Global, regional, and national prevalence of asthma in 2019: a systematic analysis and modelling study. J Glob Health 2022;12:04052 [FREE Full text] [doi: 10.7189/jogh.12.04052] [Medline: 35765786]
3. Global strategy for asthma management and prevention. Global Initiative for Asthma. 2023. URL: https://ginasthma.org/2023-gina-main-report/ [accessed 2024-03-29]
4. Yaghoubi M, Adibi A, Safari A, FitzGerald JM, Sadatsafavi M. The projected economic and health burden of uncontrolled asthma in the United States. Am J Respir Crit Care Med 2019;200(9):1102-1112 [FREE Full text] [doi: 10.1164/rccm.201901-0016OC] [Medline: 31166782]

XSL•FO
RenderX

5. Global Initiative for Asthma Executive Committee. Global strategy for asthma management and prevention (revised 2002). NHLBI/WHO Workshop Report. 2006. URL: https://cir.nii.ac.jp/crid/1572543025541173888 [accessed 2024-03-29]
6. van Dijk BCP, Svedsater H, Heddini A, Nelsen L, Balradj JS, Alleman C. Relationship between the Asthma Control Test (ACT) and other outcomes: a targeted literature review. BMC Pulm Med 2020;20(1):79 [FREE Full text] [doi: 10.1186/s12890-020-1090-5] [Medline: 32245451]
7. Larsson K, Kankaanranta H, Janson C, Lehtimäki L, Ställberg B, Løkke A, et al. Bringing asthma care into the twenty-first century. NPJ Prim Care Respir Med 2020;30(1):25 [FREE Full text] [doi: 10.1038/s41533-020-0182-2] [Medline: 32503985]
8. Chan A, De Simoni A, Wileman V, Holliday L, Newby CJ, Chisari C, et al. Digital interventions to improve adherence to maintenance medication in asthma. Cochrane Database Syst Rev 2022;6(6):CD013030 [FREE Full text] [doi: 10.1002/14651858.CD013030.pub2] [Medline: 35691614]
9. Hui CY, Walton R, McKinstry B, Jackson T, Parker R, Pinnock H. The use of mobile applications to support self-management for people with asthma: a systematic review of controlled studies to identify features associated with clinical effectiveness and adherence. J Am Med Inform Assoc 2017;24(3):619-632 [FREE Full text] [doi: 10.1093/jamia/ocw143] [Medline: 27694279]
10. Unni E, Gabriel S, Ariely R. A review of the use and effectiveness of digital health technologies in patients with asthma. Ann Allergy Asthma Immunol 2018;121(6):680-691.e1 [FREE Full text] [doi: 10.1016/j.anai.2018.10.016] [Medline: 30352288]
11. Ainsworth B, Greenwell K, Stuart B, Raftery J, Mair F, Bruton A, et al. Feasibility trial of a digital self-management intervention 'my breathing matters' to improve asthma-related quality of life for UK primary care patients with asthma. BMJ Open 2019;9(11):e032465 [FREE Full text] [doi: 10.1136/bmjopen-2019-032465] [Medline: 31722952]
12. Ljungberg H, Carleborg A, Gerber H, Öfverström C, Wolodarski J, Menshi F, et al. Clinical effect on uncontrolled asthma using a novel digital automated self-management solution: a physician-blinded randomised controlled crossover trial. Eur Respir J 2019;54(5):1900983 [FREE Full text] [doi: 10.1183/13993003.00983-2019] [Medline: 31481605]
13. Newhouse N, Martin A, Jawad S, Yu LM, Davoudianfar M, Locock L, et al. Randomised feasibility study of a novel experience-based internet intervention to support self-management in chronic asthma. BMJ Open 2016;6(12):e013401 [FREE Full text] [doi: 10.1136/bmjopen-2016-013401] [Medline: 28031210]
14. Murphy J, McSharry J, Hynes L, Molloy GJ. A smartphone app to support adherence to inhaled corticosteroids in young adults with asthma: multi-methods feasibility study. JMIR Form Res 2021;5(9):e28784 [FREE Full text] [doi: 10.2196/28784] [Medline: 34468325]
15. Tinschert P, Jakob R, Barata F, Kramer JN, Kowatsch T. The potential of mobile apps for improving asthma self-management: a review of publicly available and well-adopted asthma apps. JMIR mHealth uHealth 2017;5(8):e113 [FREE Full text] [doi: 10.2196/mhealth.7177] [Medline: 28768606]
16. Ramsey RR, Caromody JK, Voorhees SE, Warning A, Cushing CC, Guilbert TW, et al. A systematic evaluation of asthma management apps examining behavior change techniques. J Allergy Clin Immunol Pract 2019;7(8):2583-2591 [FREE Full text] [doi: 10.1016/j.jaip.2019.03.041] [Medline: 30954644]
17. Camacho-Rivera M, Vo H, Huang X, Lau J, Lawal A, Kawaguchi A. Evaluating asthma mobile apps to improve asthma self-management: user ratings and sentiment analysis of publicly available apps. JMIR mHealth uHealth 2020;8(10):e15076 [FREE Full text] [doi: 10.2196/15076] [Medline: 33118944]
18. Van Lieshout RJ, Macqueen G. Psychological factors in asthma. Allergy Asthma Clin Immunol 2008;4(1):12 [FREE Full text] [doi: 10.1186/1710-1492-4-1-12] [Medline: 20525122]
19. Gibson PG, Powell H, Wilson A, Abramson MJ, Haywood P, Bauman A, et al. Self-management education and regular practitioner review for adults with asthma. Cochrane Database Syst Rev 2002;2010(3):CD001117.
20. Thomas M, Kay S, Pike J, Williams A, Rosenzweig JRC, Hillyer EV, et al. The Asthma Control Test (ACT) as a predictor of GINA guideline-defined asthma control: analysis of a multinational cross-sectional survey. Prim Care Respir J 2009;18(1):41-49 [FREE Full text] [doi: 10.4104/pcrj.2009.00010] [Medline: 19240948]
21. Kandola A, Edwards E, Muller MAE, Dührkoop B, Hein B, Straatman J, et al. Digitally managing depression: a fully remote randomized attention-placebo controlled trial. medRxiv Preprint posted online on April 11 2023 [FREE Full text] [doi: 10.1101/2023.04.05.23288184]
22. Lee SW, Yon DK, James CC, Lee S, Koh HY, Sheen YH, et al. Short-term effects of multiple outdoor environmental factors on risk of asthma exacerbations: age-stratified time-series analysis. J Allergy Clin Immunol 2019;144(6):1542-1550.e1 [FREE Full text] [doi: 10.1016/j.jaci.2019.08.037] [Medline: 31536730]
23. Lehrer PM, Irvin CG, Lu SE, Scardella A, Roehmheld-Hamm B, Aviles-Velez M, et al. Heart rate variability biofeedback does not substitute for asthma steroid controller medication. Appl Psychophysiol Biofeedback 2018;43(1):57-73 [FREE Full text] [doi: 10.1007/s10484-017-9382-0] [Medline: 29124506]
24. Scheer FAJL, Hilton MF, Evoniuk HL, Shiels SA, Malhotra A, Sugarbaker R, et al. The endogenous circadian system worsens asthma at night independent of sleep and other daily behavioral or environmental cycles. Proc Natl Acad Sci U S A 2021;118(37):e2018486118 [FREE Full text] [doi: 10.1073/pnas.2018486118] [Medline: 34493686]

XSL•FO
RenderX

25. Sánchez-Ramos JL, Pereira-Vega AR, Alvarado-Gómez F, Maldonado-Pérez JA, Svanes C, Gómez-Real F. Risk factors for premenstrual asthma: a systematic review and meta-analysis. Expert Rev Respir Med 2017;11(1):57-72. [doi: 10.1080/17476348.2017.1270762] [Medline: 27935742]
26. Engelkes M, Janssens HM, de Jongste JC, Sturkenboom MCJM, Verhamme KMC. Medication adherence and the risk of severe asthma exacerbations: a systematic review. Eur Respir J 2015;45(2):396-407 [FREE Full text] [doi: 10.1183/09031936.00075614] [Medline: 25323234]
27. Boggiss AL, Consedine NS, Brenton-Peters JM, Hofman PL, Serlachius AS. A systematic review of gratitude interventions: effects on physical health and health behaviors. J Psychosom Res 2020;135:110165. [doi: 10.1016/j.jpsychores.2020.110165] [Medline: 32590219]
28. Schatz M, Sorkness CA, Li JT, Marcus P, Murray JJ, Nathan RA, et al. Asthma control test: reliability, validity, and responsiveness in patients not previously followed by asthma specialists. J Allergy Clin Immunol 2006;117(3):549-556 [FREE Full text] [doi: 10.1016/j.jaci.2006.01.011] [Medline: 16522452]
29. Jenkinson C, Layte R, Jenkinson D, Lawrence K, Petersen S, Paice C, et al. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? J Public Health Med 1997;19(2):179-186 [FREE Full text] [doi: 10.1093/oxfordjournals.pubmed.a024606] [Medline: 9243433]
30. Schatz M, Kosinski M, Yarlas AS, Hanlon J, Watson ME, Jhingran P. The minimally important difference of the Asthma Control Test. J Allergy Clin Immunol 2009;124(4):719-723.e1 [FREE Full text] [doi: 10.1016/j.jaci.2009.06.053] [Medline: 19767070]
31. McLean G, Murray E, Band R, Moffat KR, Hanlon P, Bruton A, et al. Interactive digital interventions to promote self-management in adults with asthma: systematic review and meta-analysis. BMC Pulm Med 2016;16(1):83 [FREE Full text] [doi: 10.1186/s12890-016-0248-7] [Medline: 27215329]
32. Hayes J. Randomised control trial of a digital asthma management application: juli: protocol and analysis plan. UCL Discovery. 2021. URL: https://discovery.ucl.ac.uk/id/eprint/10129351/ [accessed 2023-08-22]
33. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. J Pharmacol Pharmacother 2010;1(2):100-107 [FREE Full text] [doi: 10.4103/0976-500X.72352] [Medline: 21350618]
34. Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. Stat Med 2020;39(21):2815-2842 [FREE Full text] [doi: 10.1002/sim.8569] [Medline: 32419182]
35. Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables. Stat Neerl 2014;68(1):61-90. [doi: 10.1111/stan.12023]
36. Cook KA, Woessner KM, White AA. Happy asthma: improved asthma control with a gratitude journal. J Allergy Clin Immunol Pract 2018;6(6):2154-2156. [doi: 10.1016/j.jaip.2018.04.021] [Medline: 29730452]
37. Himes BE, Leszinsky L, Walsh R, Hepner H, Wu AC. Mobile health and inhaler-based monitoring devices for asthma management. J Allergy Clin Immunol Pract 2019;7(8):2535-2543 [FREE Full text] [doi: 10.1016/j.jaip.2019.08.034] [Medline: 31706485]
38. Davis SR, Peters D, Calvo RA, Sawyer SM, Foster JM, Smith LD. A consumer designed smartphone app for young people with asthma: pilot of engagement and acceptability. J Asthma 2021;58(2):253-261. [doi: 10.1080/02770903.2019.1680997] [Medline: 31691618]
39. Hsia B, Mowrey W, Keskin T, Wu S, Aita R, Kwak L, et al. Developing and pilot testing ASTHMAXcel, a mobile app for adults with asthma. J Asthma 2021;58(6):834-847. [doi: 10.1080/02770903.2020.1728770] [Medline: 32046564]
40. Gupta S, Price C, Agarwal G, Chan D, Goel S, Boulet LP, et al. The electronic Asthma Management System (eAMS) improves primary care asthma management. Eur Respir J 2019;53(4):1802241 [FREE Full text] [doi: 10.1183/13993003.02241-2018] [Medline: 30765503]
41. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
42. Chowdhury NU, Guntur VP, Newcomb DC, Wechsler ME. Sex and gender in asthma. Eur Respir Rev 2021;30(162):210067 [FREE Full text] [doi: 10.1183/16000617.0067-2021] [Medline: 34789462]
43. Boulet LP, Lavoie KL, Raherison-Semjen C, Kaplan A, Singh D, Jenkins CR. Addressing sex and gender to improve asthma management. NPJ Prim Care Respir Med 2022;32(1):56 [FREE Full text] [doi: 10.1038/s41533-022-00306-7] [Medline: 36539451]
44. Hui CY, McKinstry B, Walton R, Pinnock H. Strategies to promote adoption and usage of an application to support asthma self-management: a qualitative observational study. J Innov Health Inform 2018;25(4):243-253 [FREE Full text] [doi: 10.14236/jhi.v25i4.1056] [Medline: 30672405]
45. Koufopoulos JT, Conner MT, Gardner PH, Kellar I. A web-based and mobile health social support intervention to promote adherence to inhaled asthma medications: randomized controlled trial. J Med Internet Res 2016;18(6):e122 [FREE Full text] [doi: 10.2196/jmir.4963] [Medline: 27298211]

XSL•FO
RenderX

Abbreviations

ACT: Asthma Control Test
CONSORT: Consolidated Standards of Reporting Trials
GINA: Global Initiative for Asthma
ISRCTN: International Standard Randomised Controlled Trial Number
ITT: intention-to-treat
MCID: minimal clinically important difference
OR: odds ratio
RCT: randomized controlled trial
SF-12: 12-Item Short Form Health Survey

Edited by T de Azevedo Cardoso; submitted 19.07.23; peer-reviewed by Anonymous; comments to author 17.08.23; revised version received 23.08.23; accepted 26.03.24; published 29.04.24.

Please cite as:
Kandola A, Edwards K, Straatman J, Dührkoop B, Hein B, Hayes J
Digital Self-Management Platform for Adult Asthma: Randomized Attention-Placebo Controlled Trial
J Med Internet Res 2024;26:e50855
URL: https://www.jmir.org/2024/1/e50855
doi:10.2196/50855
PMID:38684084

©Aaron Kandola, Kyra Edwards, Joris Straatman, Bettina Dührkoop, Bettina Hein, Joseph Hayes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the World Health Organization's SkinNTDs App as a Training Tool for Skin Neglected Tropical Diseases in Ghana and Kenya: Cross-Sectional Study

Mireia Cano[1,2], MSc; José A Ruiz-Postigo[3], PhD; Paul Macharia[4], PhD; Yaw Ampem Amoako[5,6], MSc; Richard Odame Phillips[5,6], PhD; Esther Kinyeru[7], MSc; Carme Carrion[1,8,9], PhD

[1]eHealth Lab Research Group, eHealth Center, School of Health Sciences, Universitat de Catalunya, Barcelona, Spain
[2]Innovation, Digital Transformation and Health Economics Research Group, Research Institut Germans Trias i Pujol, Badalona, Spain
[3]Prevention, Treatment and Care Unit, Department of Control of Neglected Tropical Diseases, World Health Organization, Geneva, Switzerland
[4]Strathmore University, Nairobi, Kenya
[5]Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[6]School of Medicine and Dentistry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[7]Ministry of Health, Nakuru, Kenya
[8]School of Health Sciences, Universitat de Girona, Girona, Spain
[9]Network for Research on Chronicity, Primary Care, and Health Promotion, Barcelona, Spain

Corresponding Author:
Carme Carrion, PhD
eHealth Lab Research Group
eHealth Center, School of Health Sciences
Universitat de Catalunya
Rambla del Poblenou, 156
Barcelona, 08018
Spain
Phone: 34 630983328
Email: mcarrionr@uoc.edu

Abstract

Background: Neglected tropical diseases (NTDs) affect over 1.5 billion people worldwide, primarily impoverished populations in low- and middle-income countries. Skin NTDs, a significant subgroup, manifest primarily as skin lesions and require extensive diagnosis and treatment resources, including trained personnel and financial backing. The World Health Organization has introduced the SkinNTDs app, a mobile health tool designed to train and be used as a decision support tool for frontline health care workers. As most digital health guidelines prioritize the thorough evaluation of mobile health interventions, it is essential to conduct a rigorous and validated assessment of this app.

Objective: This study aims to assess the usability and user experience of World Health Organization SkinNTDs app (version 3) as a capacity-building tool and decision-support tool for frontline health care workers.

Methods: A cross-sectional study was conducted in Ghana and Kenya. Frontline health care workers dealing with skin NTDs were recruited through snowball sampling. They used the SkinNTDs app for at least 5 days before completing a web-based survey containing demographic variables and the user version of the Mobile Application Rating Scale (uMARS), a validated scale for assessing health apps. A smaller group of participants took part in semistructured interviews and one focus group. Quantitative data were analyzed using SPSS with a 95% CI and $P \leq .05$ for statistical significance and qualitative data using ATLAS.ti to identify attributes, cluster themes, and code various dimensions that were explored.

Results: Overall, 60 participants participated in the quantitative phase and 17 in the qualitative phase. The SkinNTDs app scored highly on the uMARS questionnaire, with an app quality mean score of 4.02 (SD 0.47) of 5, a subjective quality score of 3.82 (SD 0.61) of 5, and a perceived impact of 4.47 (SD 0.56) of 5. There was no significant association between the app quality mean score and any of the categorical variables examined, according to Pearson correlation analysis; app quality mean score vs age (P=.37), sex (P=.70), type of health worker (P=.35), country (P=.94), work context (P=.17), frequency of dealing with skin NTDs (P=.09), and dermatology experience (P=.63). Qualitative results echoed the quantitative outcomes, highlighting the ease of use,

XSL•FO
RenderX

the offline functionality, and the potential utility for frontline health care workers in remote and resource-constrained settings. Areas for improvement were identified, such as enhancing the signs and symptoms section.

Conclusions: The SkinNTDs app demonstrates notable usability and user-friendliness. The results indicate that the app could play a crucial role in improving capacity building of frontline health care workers dealing with skin NTDs. It could be improved in the future by including new features such as epidemiological context and direct contact with experts. The possibility of using the app as a diagnostic tool should be considered.

International Registered Report Identifier (IRRID): RR2-10.2196/39393

(J Med Internet Res 2024;26:e51628) doi:10.2196/51628

KEYWORDS

SkinNTDs app; mHealth; mobile health; neglected tropical diseases; NTDs; skin neglected tropical diseases; low- and middle-income countries; tropical disease; app; training tool; digital health; intervention; decision support tool; mobile phone

Introduction

Background

The World Health Organization (WHO) defines mobile health (mHealth) as the application of wireless technology, such as mobile phones, to the provision of health care. It is regarded as a component of eHealth, which also refers to the secure and cost-effective use of information and communication technology to assist with health system and disciplines related with health [1]. In fact, the WHO itself has recognized that eHealth has the potential to significantly contribute to sustainability and accessibility in the health system. In this regard, the number of mHealth interventions worldwide has been steadily increasing over the last decade.

This is even more relevant in regions such as Africa where the health challenges are much greater. In 2022, a systematic review of digital health interventions in sub-Saharan Africa identified 738 digital health interventions, highlighting multiple overlapping solutions with limited focus and scalability [2]. The WHO is therefore working to address these challenges, and already in 2010, it published a continent-specific digital health strategy [3].

Among the diseases that seem likely to benefit from such interventions are the neglected tropical diseases (NTDs) [4], a group of 20 diseases and conditions identified by the WHO. As their name implies, NTDs are diseases that have not historically received any kind of priority attention from international health organizations, despite affecting more than 1.5 billion people [5]. This is directly related to the fact that most of these illnesses are only found in low- and middle-income countries (LMICs) in tropical and subtropical regions, and they primarily affect women and children [5].

If undetected or untreated, some of these diseases can either be fatal or become chronic and irreversible, leading not only to lifelong disabilities but also compromised mental well-being; stigmatization; social exclusion; and, in certain countries, even experiences of racism [6-8]. This in turn perpetuates a cycle of poverty that has a direct impact on the development and economic productivity of LMICs [6].

In 2005, the WHO finally decided that a comprehensive strategy was needed to address the complexity surrounding these diseases. It therefore shifted from tackling them individually to tackling them as a single group, the NTDs [5]. It later decided to create subgroups depending on their management: those that are potentially preventable through large-scale chemotherapy interventions and those that require individual case management [9]. Within the latter category lie skin NTDs, a group of NTDs that manifest primarily as skin lesions, such as edema, patches, and ulceration, which can be detected through visual screening [10]. This group consists of Buruli ulcer, cutaneous leishmaniasis, deep fungal infections, post–kala-azar dermal leishmaniasis, leprosy, lymphatic filariasis, mycetoma, onchocerciasis, scabies and other ectoparasites, and yaws. The management of these diseases hinges on early detection and treatment, which demands significant resources, including skilled personnel and financial support [11,12]. However, early detection of these diseases is often difficult due to various factors. For instance, they are frequently painless, which can prevent them from seeking medical attention at an early stage. In addition, there is low awareness of these diseases among the population at highest risk, and, as mentioned before, stigma and discrimination associated with these diseases can also discourage people from seeking medical help [11,12].

Therefore, a key player that emerges in the strategic framework for successfully managing skin NTDs are the *frontline health workers*. This term refers to any health worker who directly provides service to a community. Although they frequently lack specialized medical training, knowledge of data collection techniques, and peer contact, they are real key players [11,12]. Indeed, they often are the initial point of contact for disease control for most skin NTDs, given that they can be visually identified and that clinical diagnosis is the most accessible diagnostic tool available [11]. Hence, it is important to enhance their education and training to facilitate their role in diagnosing, in treating, in referring patients to another level of the health system [10], and even in mitigating the stigma associated with skin NTDs [13]. In order to achieve this last-mentioned goal, numerous training initiatives for frontline health workers have been undertaken to date with good results and have been well received by them [11,13-17]. However, the predominance of in-person small group formats, with the program often targeting only 1 condition, limits the reach and increases the expense of these initiatives. Identifying alternatives to enhance the efficiency and expandability of these programs is crucial [11]. mHealth stands out as a promising, practical, and extendable approach to support frontline health worker training [18].

XSL•FO
RenderX

Although its adoption for this specific purpose has been limited, reviews of existing literature have highlighted its potential effectiveness [18]. Furthermore, specifically for skin NTDs, the WHO promotes the use of teledermatology and web-based training courses and materials whenever possible. That is the main reason why the WHO's Department of Control of Neglected Tropical Diseases has developed the SkinNTDs app [19], a mobile version of the training guide they published in 2018 [20]. By using an algorithm based on identifying signs and symptoms and providing more information about these diseases, this app assists frontline health workers in the diagnosis and management of skin NTDs.

However, although the third version has already been released, the app's usability, efficacy, and effectiveness have not yet been evaluated. This is a critical step in guaranteeing quality and empowering end users not to rely entirely on popularity or "star"-rating systems, which have already been proved insufficient [21]. Evaluation of mHealth interventions is a key component of most digital health technology frameworks [22-24], even more so when app-based mHealth interventions are scarce in Africa, as most mHealth interventions in this region are SMS text messaging based [25].

The intervention maturity life cycle schematic proposed by WHO [24] is a practical guide explaining the goals at each stage, the number of participants required, and the measurement targets. Considering these guidelines and the stage of maturity of the SkinNTDs app (the first and second stage), it now seems appropriate to evaluate its feasibility and usability. In this stage, questions regarding how the app is used by end users, how it fits into their workflow, and how easy the learning curve of use is should be answered before moving on to the next stage (efficacy).

One reliable tool available to assess usability is the user version of the Mobile Application Rating Scale (uMARS) [26], a simple tool for classifying and rating mHealth apps based on 4 objective subdomains (engagement, functionality, aesthetics, and information quality) and 1 subjective quality subdomain.

Objective

Given the WHO's commitment to mHealth as a tool to achieve a range of sustainable goals, it is essential to evaluate interventions that have the potential to be highly scalable and cost-effective. This paper summarizes the results of a cross-sectional study assessing the engagement, functionality, aesthetics, and information quality of version 3 of the SkinNTDs app for the real end user in their actual context according to a validated tool. In addition, a secondary objective was to check whether the demographic information gathered influenced the final uMARS score, given that the developers of the SkinNTD app had no plans to customize the app for different settings upon its implementation.

Methods

Study Design

A cross-sectional study was conducted between December 2022 and April 2023 in Ghana and Kenya. The study design and methods are described in detail in the published protocol [27]. Deviations from this protocol are summarized in Multimedia Appendix 1.

Ethical Considerations

This study was conducted according to the ethical principles established by the World Medical Association in the Declaration of Helsinki of Ethical Principles for Medical Research Involving Human Subjects [28]. To guarantee that the protocol complied with the ethical standards of all 3 countries involved, it was approved by their respective ethics committees. The protocol was authorized by the Ethics Committee of Universitat Oberta de Catalunya (Spain) (20201127_mcarrion_NTDs), the Ethical Committee of Kwame Nkrumah University of Science and Technology (Ghana) (CHRPE/AP/576/22), and the Ethical Committee of Coast General Teaching and Referral Hospital (Kenya) (ERC-CGH/MSc/VOL.I).

Participants gave informed consent to participate in the study before taking part. To ensure confidentiality and anonymity, each participant was assigned a unique identification number (eg, P01) for coding all collected information and data. Moreover, participants did not receive any economic compensation for their participation in this study.

Participant Recruitment and Eligibility Criteria

Nonprobabilistic snowball sampling was used due to the difficulty in making direct contact with frontline health care workers in these 2 countries. Snowball sampling is a recruitment technique in which chosen participants are asked to find and contact other potential participants from among their acquaintances. Although this approach is nonrandomized, it appeared to be the best strategy to locate the participants of this study, as the main researchers were based on a different continent.

A diagram representation of this snowball sampling recruitment is shown in Figure 1. More detailed information can be found in the published protocol [27].

Anyone who worked or had worked with skin NTDs on a regular basis in the 2 selected countries, who were or had been in charge of their diagnosis and treatment, who had a smartphone (Apple or Android), and who had downloaded and used the SkinNTDs app on at least 5 different days was eligible to take part. In addition, participants had to use WhatsApp (Meta Platforms) or email to send the informed consent. Limited understanding of the English language and refusal to sign informed consent form were factors for exclusion.

The same population was the target for both parts of the study. Before signing the informed consent form, participants were asked to read the information page.

XSL•FO
RenderX

Figure 1. Diagrammatic representation of snowball sampling recruitment. MARS: Mobile Application Rating Scale; WHO: World Health Organization.

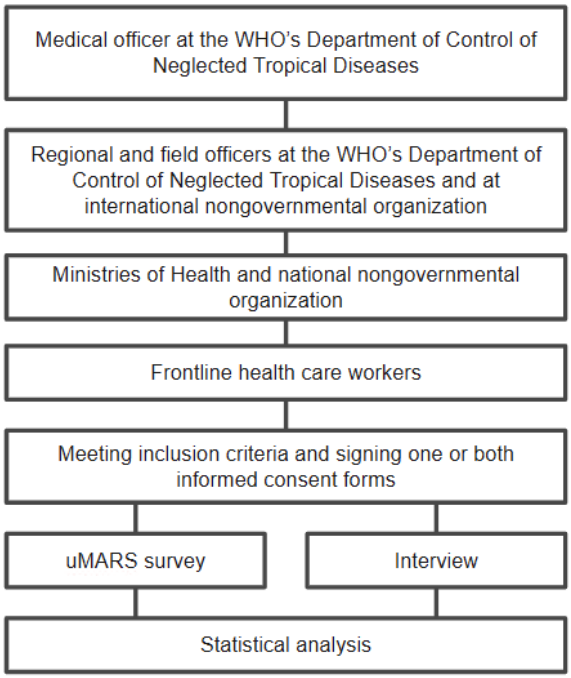

Sample Size Calculation

As already mentioned, this study followed the steps proposed in the intervention maturity life cycle schematic described by the WHO in *Monitoring and Evaluating Digital Health Interventions* [24]. As a result, sample size was determined using this guide.

The SkinNTDs app is now at the prototype stage, which, in accordance with WHO recommendations, corresponds to stages 1 and 2. The WHO suggests evaluating the intervention with a sample size between 10 and 100 people. Given that a sample of 100 participants is the maximum number advised and that this study was conducted on the web, we assumed a 50% dropout rate. These calculations resulted in a 50-person final sample size.

Outcomes

Demographic Variables

Participants were asked to complete an anonymous survey of various demographic data that included age, sex, country of residence, type of frontline health care worker, frequency of dealing with skin NTDs, experience and training in dermatology, work environment, working institution, knowledge of mobile technology, and languages spoken.

Quantitative Method: uMARS Questionnaire

The uMARS was used to assess the quality of the SkinNTDs app [26]. This tool assesses 20 items clustered into 4 objective subscales (engagement, functionality, aesthetics, and information quality) and 1 subjective subscale. Participants rated each item using a Likert scale from 1 to 5; higher numbers meant a better rating. A "not applicable" option was available in case an item could not be assessed. In the extra "app-specific" category, only the relevant questions for this study were chosen. Finally, 9 self-created questions were added to complete the survey.

A total of 2 scores were obtained based on the original recommendation from the authors of this scale: app quality score and app subjective quality score. Questions rated as not applicable were excluded from the score.

In our study, we adapted the evaluation process based on the original article's recommendations, which suggested a minimum application use time of 10 minutes before completing the uMARS. However, we considered that this time was not enough to complete a comprehensive evaluation, and we extended to 5 days.

Qualitative Methods: Semistructured Interviews and Focus Groups

Due to the different realities of the context of the 2 countries, 2 qualitative methods were used.

Semistructured interviews were conducted as a qualitative method for exploring the perspectives, perceptions, and opinions of participants, combining prepared questions with others that arose during the interview [29]. Each participant was asked 7 questions based on the key findings obtained from the uMARS questionnaire in the same order and in the same words as a standardization that facilitated comparison [30]. In addition, interviewers could ask unforeseen questions based on a participant's answer. The questions were unbiased, open-ended, and well-written and used simple terms.

Focus groups were also used as another qualitative research method for the same purpose as before. This method involves a moderator posing targeted questions to participants in a group setting, with the advantage of capturing valuable interactions among them [31]. Focus group sessions also incorporated the same set of 7 questions used in the semistructured interviews. Moreover, the moderator had the same flexibility to ask unforeseen questions based on the participants' interactions and responses.

XSL•FO
RenderX

Our study design incorporated a predetermined objective to include a minimum of 10% of the overall sample size or continue data collection until reaching a point of information saturation using both methods.

More details of the instruments can be found in the study protocol [27].

Data Collection and Study Procedure

The principal researcher delivered the consent form via email or WhatsApp when frontline health care workers contacted her. However, participants enrolled at Coast General Teaching and Referral Hospital (Kenya), due to ethical requirements, signed a paper-based consent form delivered to them by the local investigator. After it was signed, participants received a follow-up email with more details about the study and a link to the uMARS survey, which they had to answer after using the app for at least 5 days.

Participants were able to express their willingness to take part in the semistructured interviews in the same questionnaire.

After reading the detailed information sheet for this second phase of the study, participants who had agreed were required to sign a second consent form. Then, the interview was scheduled by mutual agreement.

The interviews were conducted via Google Meet (Google) and had a duration ranging from 25 to 40 minutes. The focus groups were conducted face-to-face and moderated by a local researcher for 60 minutes. All sessions were recorded regardless of the qualitative method used.

During this time, the interviewer and moderator could take notes and reiterate any questions or sentences for clarification.

Interviews and focus groups were all transcribed using Otter.ai software (Sam Liang) for analysis. The research protocol provides more information on the procedure for data collection [27].

Statistical Analysis

All quantitative data analyses were conducted using SPSS (version 25; IBM Corp) for Windows, with the statistical significance set at $P \leq .05$. A descriptive analysis was conducted to describe the demographic variables and data obtained from the uMARS scores. Categorical variables were described in terms of frequencies and continuous variables in terms of central tendency and dispersion measurements. All data were shown in tables. Data normality was evaluated using Kolmogorov-Smirnov test ($P=.05$). The t test (2-tailed) was used for comparing means of bivariate variables and ANOVA for comparing means of multivariate variables. A logistic regression analysis was performed to add the covariates that could skew the main association under analysis. A CI of 95% was assumed, and $P \leq .05$ was considered a significant difference.

Selected quotes were returned to participants for approval. Qualitative data derived from the semistructured interviews and focus groups were analyzed using ATLAS.ti (ATLAS.ti Scientific Software Development GmbH). We identified attributes, clustered them into different themes, and then coded these themes to analyze the various dimensions explored during the interviews.

Results

Participants

In total, 60 participants took part in this study. Demographic characteristics are shown in Table 1.

Of the 60 participants, 57% (n=34) were male individuals. Participants were divided quite homogeneously in the different age groups, <36 years (n=22, 37%), 36 to 45 years (n=22, 37%), and 46 to 65 years (n=16, 27%). Participation was higher in Kenya with 36 (60%) participants, compared with the 24 (40%) from Ghana. Up to 93% (n=56) of them referred to working in a public health care institution. From all these participants, 72% (n=43) identified themselves as frontline health care workers, and 62% (n=37) mentioned working in a rural area. The vast majority of the participants were not dermatology specialists but said they had some experience with the topic (n=44, 73%). The participants were distributed similarly in terms of how frequently they dealt with skin. More than half of the participants (n=41, 68%) considered they had extensive knowledge of mobile technology, and 87% (n=52) had no additional training before using the app. Speaking English was a requirement for participation, so everyone who participated did. However, there were up to 19 more languages registered, among which Swahili stood out as the second most spoken language by the participants in Kenya (25/36, 69%) and Twi in Ghana (20/24, 83%).

XSL•FO
RenderX

Table 1. Participant demographics in the World Health Organization's SkinNTDs app usability and user experience assessment through the user version of the Mobile Application Rating Scale survey (n=60) shown by frequencies (%).

Variable	Participants, n (%)
Age (y)	
<35	22 (37)
36-45	22 (37)
46-65	16 (27)
Sex	
Female	26 (43)
Male	34 (57)
Country	
Kenya	36 (60)
Ghana	24 (40)
Type of health worker	
Frontline	43 (72)
Nonfrontline	17 (28)
Work context	
Rural	37 (62)
Urban	23 (38)
Type of working institution	
Public health care setting	56 (93)
Private health care setting	2 (3)
Nongovernmental organization	2 (3)
Dermatology experience	
Not trained, no experience in dermatology	13 (22)
Not trained, but some experience in dermatology	44 (73)
Trained and experienced in dermatology	3 (5)
Frequency of dealing with skin neglected tropical diseases (cases/month)	
Rarely (<1)	13 (22)
Occasionally (1-3)	13 (22)
Frequently (4-6)	21 (35)
Usually (>6)	13 (22)
Knowledge of mobile technology	
High knowledge	41 (68)
Medium knowledge	19 (32)
Extra training in app use	
Yes	8 (13)
No	52 (87)
Language	
Kenyan participants (n=36)	
English	36 (100)
Swahili	25 (69)
Kamba	3 (8)
Kikuyu	2 (6)

XSL•FO
RenderX

Variable	Participants, n (%)
Marakwet	2 (6)
Dhulo	1 (3)
Kisii	1 (3)
Kalenjin	1 (3)
Oromo	1 (3)
Somali	1 (3)
Ghanaian participants (n=24)	
English	24 (100)
Twi	20 (83)
Asanti	3 (12)
Fante	2 (8)
Akan	1 (4)
French	1 (4)
Chinese	1 (4)
Ga	1 (4)
Ewe	1 (4)

uMARS Score

Tables 2 and 3 show the overall results from the 3 sections of the uMARS questionnaire and specific results for each subdomain. They appear separately according to the recommendations of the uMARS authors, in order to strengthen the objectivity of the final result.

The SkinNTDs app received an overall score of 4.02 (0.47) out of 5 in the app quality mean score, 3.82 (0.61) out of 5 in the subjective mean score, and 4.47 (0.56) out of 5 in the perceived impact section. Thus, the app received more than 75% of the maximum score in all 3 sections.

These overall scores can be broken down as follows: starting with the most objective scoring, the app quality mean score is the average obtained in each of the 4 uMARS domains (engagement, functionality, aesthetics, and information quality), which in turn are divided into 16 subdomains. The scores range from 3.65 to 4.20, with engagement being the lowest rated and information quality the highest rated. Moreover, SDs are between 0.47 and 0.65, indicating that the dispersion of the data tends to be medium-low. Notably, the lowest-rated subdomain was customization (mean 3.18, SD 1.02, out of 5), which refers to whether the app provides all necessary settings and preferences for app features (eg, sound, content, and notifications). In contrast, the credibility of the source subdomain was the highest rated of the 16 (mean 4.55 (SD) out of 5), referring to whether the app comes from a legitimate source.

At a subjective level, the app obtained a mean score of 3.82 (SD) out of 5. The findings in this section indicate that individuals are not very eager to pay for this app (mean 2.57, SD 1.33); however, the subvariable measuring whether they would recommend the app had a very high score (mean 4.65, SD 0.70). Finally, when the participants were asked for the overall rating, the average score was 4.08 (SD), which is extremely close to the quality mean score of 4.02 (SD). In this case, the SDs are slightly higher than in the previous section, indicating more variability in the responses.

Finally, the participants rated the perceived impact score (including awareness and knowledge) highly, with mean values of 4.40 (SD) and 4.55 (SD) out of 5 correspondingly.

Table 3 shows more detailed information regarding the specific results for each subdomain. In addition, graph 1 presents the scores of all objective domains assessed in a visual way.

Table 3 specifies the detailed results for each subdomain. In addition, Figure 2 shows the results obtained in each subdomain visually by country.

The study was completed with the comparison of the mean, which was based on the comparison between the global app quality mean score versus the main domains. The Kolmogorov-Smirnov test demonstrates a normal distribution of the data derived from the app quality mean score. Thus, a *t* test was used for comparing means of bivariate variables and ANOVA for comparing means of multivariate variables.

XSL•FO
RenderX

Table 2. Results of the validated user version of the Mobile Application Rating Scale domains according to the cross-sectional study.

Domain	Mean (SD)	Minimum	Maximum
App quality mean score	4.02 (0.47)	2.89	4.83
Engagement mean score	3.65 (0.57)	2.20	4.80
Functionality mean score	4.18 (0.65)	2.50	5
Aesthetics mean score	4.02 (0.53)	3	5
Information mean score	4.20 (0.49)	3	5
App subjective mean score	3.82 (0.61)	2.25	5
App perceived impact	4.47 (0.56)	2.50	5

Table 3. Results of each validated user version of the Mobile Application Rating Scale subscale domain according to the cross-sectional study.

Subdomains	Mean (SD)	Minimum	Maximum
Engagement score			
Entertainment	4.03 (0.69)	3	5
Interest	4.08 (0.72)	2	5
Customization	3.18 (1.02)	1	5
Interactivity	3.22 (0.92)	1	5
Target group	3.98 (0.97)	1	5
Functionality score			
Performance	4.17 (0.93)	2	5
Ease of use	4.20 (0.71)	3	5
Navigation	4.30 (0.83)	2	5
Gestural design	4.15 (0.71)	2	5
Aesthetics score			
Layout	4.20 (0.73)	3	5
Graphics	3.93 (0.68)	3	5
Visual appeal	3.90 (0.54)	3	5
Information score			
Quality of information	3.93 (0.63)	2	5
Quantity of information	4.32 (0.75)	2	5
Visual information	4.18 (0.65)	3	5
Credibility of source	4.55 (0.80)	2	5
Subjective items score			
Would you recommend?	4.65 (0.75)	2	5
How many times?	3.98 (0.83)	2	5
Would you pay?	2.57 (1.33)	1	5
Overall (star) rating	4.08 (0.70)	2	5
Perceived impact score			
Awareness	4.40 (0.59)	3	5
Knowledge	4.55 (0.62)	2	5

XSL•FO
RenderX

Figure 2. Radar chart evaluation of each app's objective quality domain according to the user version of the Mobile Application Rating Scale (uMARS), with results divided by country.

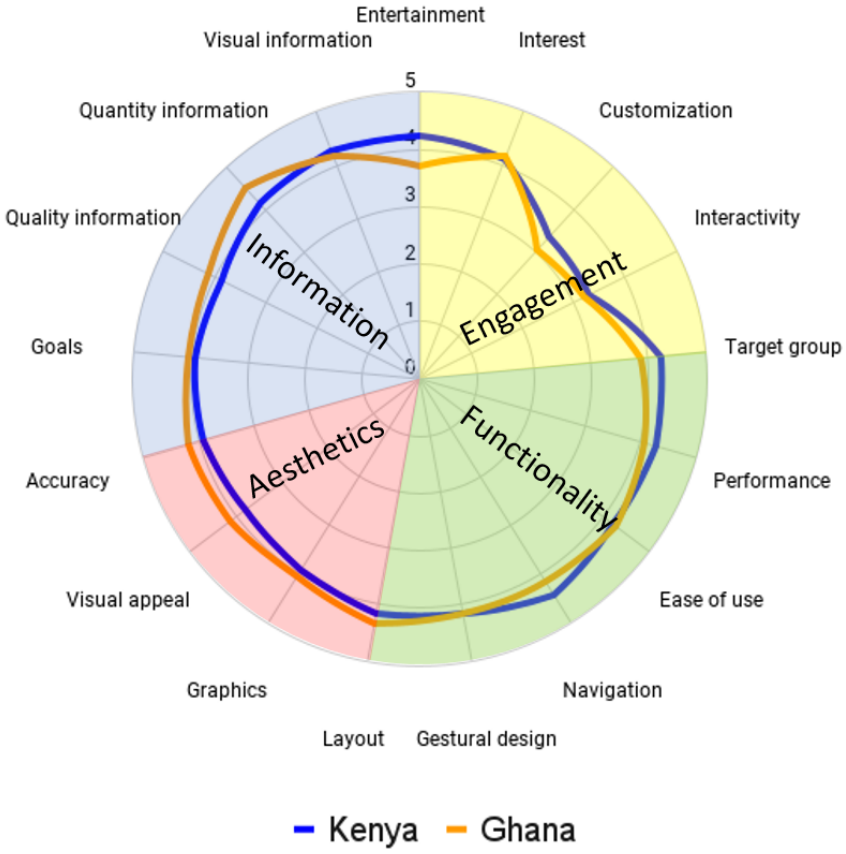

The results (Table 4) revealed that there was no statistically significant difference in any of the variables.

Similarly, linear regression showed that there was no significant association between the app quality mean score and any of the categorical variables examined above (F=0.60; P=.76), with R^2=0.07.

The linear regression model analyzing the main demographic variables of the participants and their correlation with the user version of the Mobile Application Rating Scale (uMARS) app quality mean score in the cross-sectional study is as follows: The model shows an R value of 0.27, an R^2 of 0.07, and an adjusted R^2 of –0.05. The F test for this model (df=0.60) resulted in a P value of .76. The dependent variable for this analysis was the app quality mean score, and the predictors included age, sex, type of health worker, country, work context, frequency of dealing with skin NTDs, and dermatology experience. Meaning that there was no significant association between the app quality mean score and any of the categorical variables examined. Finally, as mentioned in the Methods section, 7 questions were added to complete the app evaluation, which are shown in Table 5.

It is worth noting that 64% (36/60) of participants used the app more than the minimum number of times required before answering the survey, and the majority believed that using the app would reduce the time it takes to diagnose skin NTDs. In terms of connectivity, while up to 67% (40/60) claimed to have a strong connection, a substantial minority group of 33% (20/60) still had limited or no access to the internet.

Almost all the participants favored the inclusion of skin NTDs surveillance, the ability to include patient records, and the availability of a computer version of the app.

Slightly more than half of the participants (32/60, 53%) considered it necessary for the app to be available in other languages, with Twi and Swahili being the first on the list.

XSL•FO
RenderX

Table 4. Comparison of means between user version of the Mobile Application Rating Scale app quality mean score versus the main demographic variables of participants in the cross-sectional study.

Variables compared and statistical tests performed	Value	*P* value
App quality mean score vs age		
Levene statistic	1.42	.25
ANOVA	1.01	.37
App quality mean score vs sex		
Levene statistic	2.24	.14
t test (*58*)	0.40	.70
App quality mean score vs type of health worker		
Levene statistic	0.04	.84
t test (*58*)	–0.94	.35
App quality mean score vs country		
Levene statistic	5.09	.03
T-Welch	0.08	.94
App quality mean score vs work context		
Levene statistic	2.95	.09
t test (*df*)	–1.40 (58)	.17
App quality mean score vs frequency of dealing with skin neglected tropical diseases		
Levene statistic	4.13	.01
Kruskal-Wallis test	6.40	.09
App quality mean score vs dermatology experience		
Levene statistic	0.88	.50
ANOVA	0.70	.63

XSL•FO
RenderX

Table 5. Other relevant questions added to the user version of the Mobile Application Rating Scale survey (n=60).

Questions	Responses, n (%)
Duration of use	
<7 days	21 (35)
1-4 weeks	26 (43)
1-2 months	11 (18)
2-4 months	2 (3)
Internet connection	
I hardly have internet access	3 (5)
Most of the time I do not have internet access	8 (13)
Sometimes I have internet access	9 (15)
Most of the time I have internet access	31 (52)
I always have internet access	9 (15)
Translation	
Yes	32 (53)
No	28 (47)
Surveillance	
Yes	100 (100)
Patient records	
Yes	59 (98)
No	1 (2)
Desktop version	
Yes	51 (85)
No	9 (15)

Semistructured Interviews

Overview

Regarding the second part of the study, 55% (33/60) of participants showed an interest in participating. However, only 13 men and 4 women were interviewed (5 from Ghana and 12 from Kenya). Upon interviewing 17 people, we observed that the responses started to repeat, indicating data saturation. The data collection comprised 10 semistructured interviews, with 5 participants from Ghana and 4 from Kenya, along with a single focus group conducted in Kenya involving 8 participants.

Most participants (10/17, 59%) were aged between 46 and 65 years. Approximately, 76% (13/17) of the participants were frontline health care workers, and 47% (8/17) said that they had occasional contact with skin NTDs. Demographic characteristics are shown in Table 6.

Despite being asked 7 questions, not all participants were able to provide detailed answers that allowed attributes to be identified. This was mainly due to technological and language barriers.

Both the semistructured interview and the focus group data were transcribed by 2 researchers (MC and PM) and thoroughly reviewed by 1 researcher (MC) to gain an initial understanding of the data. Subsequently, expressions aligning with the study objectives were extracted from the text and transformed into attributes (n=95) using ATLAS.ti software. These codes represented condensed versions of the professionals' thoughts. On the basis of the content similarity, the codes were further grouped into themes (n=18), which were then associated with each question asked. At this stage, another researcher (CC) examined the coding, subcategories, and higher-level categories. Discussions were held to reach a consensus on the content and names assigned to categories.

Table 7 presents the description and frequency of the identified themes and selected participant quotes.

XSL•FO
RenderX

Table 6. Participant demographics in the World Health Organization's SkinNTDs app usability and user experience assessment through semistructured interviews and focus group (n=17).

Variable	Participants, n (%)
Age (y)	
<35	4 (23.5)
36-45	3 (17.6)
46-65	10 (58.8)
Sex	
Female	4 (23.5)
Male	13 (76.5)
Country	
Kenya	12 (70.6)
Ghana	5 (29.4)
Type of health worker	
Frontline	13 (76.5)
Nonfrontline	4 (23.5)
Frequency of dealing with skin neglected tropical diseases (cases/month)	
Rarely (<1)	4 (23.5)
Occasionally (1-3)	8 (47.1)
Frequently (4-6)	3 (17.6)
Usually (>6)	2 (11.8)
App assessment duration (5 days)	
Yes	14 (82.4)
No	3 (17.6)

XSL•FO
RenderX

Table 7. Description and frequency of the identified themes and selected participant quotes.

Theme	Participants[a], n (%)	Best points
Best points of the SkinNTDs app (n=10)		
Good for people without SkinNTDs experience	8 (80)	• "The App gives you a wide understanding about SkinNTDs without being an expert in this topic." (P1) • "The App helps people without knowledge in SkinNTDs and leads them to the right directions to make diagnosis." (P3)
Ease of use	3 (30)	• "What I really appreciate from the App is that is really easy to navigate." (P3)
Friendly interface	4 (40)	• "I really like the interface." (P4)
Offline functionality	4 (40)	• "Is nice to have this App always available with you, even when you don't have access to internet." (P2)
Easy accessibility	4 (40)	• "The fact that this App works offline, it means it is portable. And even in the most remote places it is still accessible. Making in a huge plus for its spread use." (P4) • "It is nice that the App can be accessed anytime." (P9) • "The app is always accessible." (P8)
Simple language	2 (20)	"It is easy to understand because uses simple English."(P9)
Scope	2 (20)	• "So it is good because, many people can be able to get the app and use it." (P14) • "I think the app can be used especially in those areas, remote areas that are hard to reach areas where we have no doctors." (P16)
Aspect to be improved of the SkinNTDs app (n=7)		
Signs and symptoms section	4 (57)	• "It is important to understand the sign and symptoms to don't make wrong decisions. It may be not enough for nonexperts in Skin NTDs with the current information in the app. Hence, make wrong diagnosis." (P1) • "I think it is needed to add more images to correct identify correctly the signs/symptoms." (P5)
Customization aspect	1 (14)	• "I think some holistic approach is needed to capture all skin NTDs." (P6) • "The app should provide full information about a condition and its management according to different levels of care." (P9) • "There is not an option to customize some preferences related to the App." (P4)
Issues to be solved	1 (14)	• "There are still small issues to be solved, such as hyperlinks which do not work, specially related to the images. Moreover, there is not a direct connections with the developers to make them aware of it." (P3)
Lack of an expert panel	2 (29)	• "I would like to have an option to upload photos, and ask a panel of experts to give their opinion in real-time." (P2)
Lack of African photos	2 (29)	• "Most of the photos are referred to outside African countries." (P9)
How to improve customization aspect (n=4)		
Notifications	4 (100)	• "In terms of notifications it may help to receive notifications when something is updated. I think this is the most significant part from customization, the notifications part." (P1) • "I would like to have the option to programm sounds and notifications." (P10)
Aesthetics	1 (25)	• "I would like to have more options to customize some aspects, such as letter size or colors of the App." (P4)
How to improve interactivity aspect (n=4)		

XSL•FO
RenderX

Theme	Participants[a], n (%)	Best points
Feedback	3 (75)	• "Maybe it could improve by adding more feedback and reminders." (P6) • "The feedback and inputs should be reviewed on monthly basis." (P10)
Frequently asked question section	1 (25)	• "Put a FAQ on the Skin NTDs APP could be beneficial." (P7)
Extra aspects (n=3)		
Dissemination strategy	3 (100)	• "Liaise with the professional body, for example with professional association. This will be good way to disseminate." (P13) • "I think the first thing is creating awareness of the existence of the app through some kind of a seminar." (P15) • "It can also be introduced through the training institutions. You know can introduce it to them, they use it in training their students and then once they graduate, they go out well informed and then they can use it right away." (P15) • "The Ministry of Health takes it up and then it becomes as a standard of the way of managing patients with the skin diseases." (P16)

[a]Not every participant provided sufficient elaboration to allow for identification of themes, so the frequency of individual themes does not add up to the total number of participants who participated in the semistructured interviews.

Best Features of the SkinNTDs App (n=10)

Most of the participants (n=8, 80%) considered that the best feature of the SkinNTDs app is how good it is for people without skin NTDs experience. They stated that the app gives them a wider understanding and helpful directions to obtain further knowledge about these conditions. In addition, 40% (n=4) of participants identified the app's user-friendly interface and offline functionality as significant aspects. Participants also mentioned the use of simple language in the app (n=2, 20%) and highlighted its potential scope (n=2, 20%), particularly as a tool suitable for use in rural areas.

Points to Improve in the SkinNTDs App (n=7)

The signs and symptoms section was the most frequently mentioned area for improvement (n=4, 57%), especially with regard to understanding the definition of signs and symptoms, which could lead to misdiagnosis. Furthermore, 29% (n=2) of participants expressed concerns about the absence of an expert panel for addressing doubts and queries. In addition, they highlighted the limited availability of African photographs (2/7, 29%), which posed challenges in accurately identifying symptoms on dark skin. Finally, 1 (14%) participant suggested enhancing the app's personalization aspect, while another pointed out unresolved issues that negatively impacted his experience with the SkinNTDs app.

How to Improve the Customization Domain of the SkinNTDs App (n=4)

In terms of customization, the worst-rated domain in the uMARS, all participants (n=4, 100%) unanimously identified the notifications as a key point to improve the customization of the app. Specifically, participants emphasized the need for features such as sound alerts or notifications to indicate app updates. In addition, 1 (25%) participant expressed the importance of having the option to customize the font size within the app.

How to Improve the Interactivity Domain of the SkinNTDs App (n=4)

The 2 requests from participants to improve the interactivity of the SkinNTDs app were to add feedback (n=3, 75%) and a frequently asked question section (n=1, 25%).

App Assessment Duration (n=11)

Of the 11 participants who responded to the question on whether 5 days for testing the app was sufficient time to apply the appropriate evaluation criteria, 8 (73%) participants stated that it was sufficient, while 3 (27%) participants indicated that they felt the testing period should be increased to between 10 and 15 days.

Medical Divide Feasibility (n=6)

When asked whether the SkinNTDs app could be used as a medical device in the future, 5 (83%) participants responded affirmatively. They cited the app's ability to be easily used almost anywhere and its regular updates as the main reasons for their positive outlook. However, 1 (17%) participant believed that the current state of the app was not sufficiently developed to be considered a medical device. They suggested that further development and clinical validation would be necessary to consider that as a viable option.

Use of Other Health Apps (n=15)

The vast majority of participants (n=13, 86%) denied using any kind of health care app. In contrast, 2 (13%) participants admitted to using some form of health app. One participant reported using an app specifically designed for managing patients with tuberculosis; the other participant mentioned having used an app but could not recall its specific purpose.

Dissemination Strategy (n=3)

After addressing all the questions, 3 (100%) participants emphasized the importance of discussing the dissemination strategy for the SkinNTDs app. They unanimously agreed that official institutions, such as regional or national associations, universities, and the Ministry of Health, should be involved in

XSL•FO
RenderX

the dissemination process. They stressed that involving these associations would help establish the app as a standard of care and ensure that future health care professionals are trained in using technology effectively in their daily work practices.

Discussion

Principal Findings

Skin NTDs are a subgroup of 13 NTDs that present primarily as skin lesions that can be detected through visual screening. Tackling these diseases depends on prompt diagnosis and treatment; hence, the resources used tend to be substantial. If they are not identified or treated, they could become chronic and irreversible, favoring a cycle of declining social, economic, and health conditions, which has a negative impact on the overall growth of LMICs [5,9].

Efforts to improve the diagnosis and management of skin NTDs are crucial, yet many existing strategies, despite showing effectiveness, are constrained by their face-to-face nature, focus on single conditions, and lack of sustainability and scalability [11,13-15]. Furthermore, rural areas still tend not to benefit from these interventions because they are often excluded [32]. Recognizing these challenges, the WHO highlights the significance of mHealth solutions, since it could offer a versatile and scalable approach to overcoming the limitations of traditional methods and more sustainable management of skin NTDs across diverse settings [19].

A study conducted by Krah and de Kruijf [25] revealed that in Africa, effective mHealth interventions are characterized by straightforward design and modest objectives, favoring simplicity over complicated interventions. This could explain why, according to the report, most mHealth interventions in the region are SMS text messaging based rather than app based and are primarily focused on disorders such as HIV and malaria as well as sexual and reproductive health. Along the same lines, a comprehensive evaluation of mHealth strategies for addressing skin NTDs indicated that most interventions are SMS text messaging based, and further effort is needed to homogenize therapies and eliminate methodological limitations [33]. The review by Carrion et al [33] identified only 2 interventions which were app based [16,17]: Skin App, developed by Netherlands Leprosy Relief and currently integrated into the SkinNTDs app evaluated in this study and Guaral App, developed by Universidad Icesi Grupo I2T. [2,11,13-15,25].

In order to increase the likelihood of success, it is vital to understand whether the proposed solutions are feasible in the real context and whether they are suitable for the end users [25]. It is therefore crucial to conduct a thorough evaluation of mHealth interventions in their different stages of development. Guaral App is designed to assist community health professionals in rural Colombia with the detection and referral of patients with cutaneous leishmaniasis, and as such, it is the only app found that is comparable to the SkinNTDs app. Furthermore, its usability has already been evaluated with positive findings, and sensitivity was found to be greater than 95%. However, because the researchers had recruited only 9 participants and did not use a validated measure to assess the usability, the results should be interpreted with caution [34].

The recently released eSkinHealth is another app that aims to enable onsite and remote diagnosis, monitoring, and clinical decision support for skin diseases, including NTDs, specifically designed for LMICs and darker skin types. Researchers plan to assess its usability with the System Usability Scale [32].

The SkinNTDs app developed by the WHO's Department of Control of Neglected Tropical Diseases is currently in the early stages of its life cycle. Therefore, before moving on, it is crucial to assess at this stage whether it has been properly designed to fulfill its purpose, which is to be an educational and decision support tool to help frontline health care workers improve the management of NTDs in their daily practice. According to WHO guidelines, it is at the stage when its feasibility and usability should be evaluated [24].

The uMARS tool was chosen to assess version 3 of the SkinNTDs app, as it has already been widely used to evaluate other health apps, and it has proven its simplicity and reliability [26]. In addition, semistructured interviews were undertaken to complete the study from all perspectives.

Before discussing the results in more detail, it is important to note that the sample includes both frontline health care workers and non–frontline health care workers. The link to participate in the study also reached other health professionals who were not classified as frontline health care workers. Although it was initially thought that the sample should be restricted to actual end users, responses of non–frontline health care workers were also recorded, allowing for a more detailed analysis. Since there were no differences between these 2 groups in any of the variables, it was decided to include them in the study and ensure greater consistency in the findings.

The results of this study show that the SkinNTDs app has a high level of quality as measured by the uMARS questionnaire, obtaining an app quality mean score of 4 (SD 0.47) of 5, a subjective quality score of 3.83 (SD 0.61) of 5, and a perceived impact of 4.5 (SD 0.56) of 5. These results seem to indicate that the app meets the users' needs.

Perhaps the most important result obtained in this study is that there is neither difference nor correlation between any variables with the app quality mean score. This may suggest the potential feasibility of the app being widely distributed to countries affected by skin NTDs without requiring any customization other than making new languages available. Some articles have highlighted the importance of the effective use of innovative technologies in health care and the development of sustainable strategies [35]. The SkinNTDs app is obviously an example of a case that appears to have great potential for dissemination without a significant financial investment. Another important finding is that 84.1% (50/60) of participants rated the app with a 4 or 5 for ease of use, although most (47/50, 94%) of them did not receive any training before using it. This fits well for the design of any future dissemination campaign, as the app could be offered without prior training, saving money on implementation. The worst-rated domains could be enhanced by the app developers with some changes. In particular,

XSL•FO
RenderX

improving the customization subdomain (the worst rated) could help the app to adapt to the priorities of each user.

Participants also expressed interest in adding other capabilities to the app, such as patient records, surveillance information for skin NTDs, or a desktop version. Something important to bear in mind when building these potential new features is that 30.7% (16/60) of participants reported having a poor internet connection.

Regarding the second part of the study, in which the semistructured interviews and focus group were conducted, 56% (34/60) of participants showed interest in participating and were contacted by email. Despite sending numerous reminders to all of them, only 17 people were finally interviewed, thanks to the efforts of 2 key actors and researchers from this study (EK and PM), without whom the second part of this study could not have been conducted. However, the web-based interviews were challenging due to unreliable internet connections, which hindered the smooth flow of conversation. As a result, not all participants were able to provide responses to all the questions. These facts will be further discussed in the following Limitations section. Nevertheless, most of the participants' answers corroborated the results obtained in the first part of the study with the uMARS questionnaire.

Once again, the best features of the app were considered to be the target group (frontline health care workers), the user-friendly interface, the offline functionality, the use of simple language, and its potential scope. The target user group is especially important to mention, since the goal of this app is to train workers who deal with these conditions but do not have in-depth knowledge of them.

In contrast, the participants also identified some points that should be reviewed in order to improve the whole experience with the app. The sign and symptoms section was where the participants found the most issues that could be improved, mainly because understanding them is vital for a correct diagnosis. This is directly related to the group of people who are expected to use the app and lack the necessary knowledge or experience to correctly identify the signs and symptoms if these are not clearly explained. Thus, this is a major point to consider. Another possible weakness in the signs and symptoms section is the use of photos that are not specific to the region where the app is used. In this case, participants mentioned the lack of African photos, which posed challenges in accurately identifying symptoms on dark skin.

Participants reached a consensus regarding the potential improvements to the 2 weakest domains, customization and interactivity. They suggested that incorporating notifications, sounds, and a frequently asked question section could address many of their concerns. In addition, they emphasized the importance of seeking feedback from experts. However, it is crucial to carefully consider this suggestion, as implementing web-based functionalities may compromise one of the key advantages of this app, as highlighted by the participants: the offline functionality, which enables effective dissemination in areas with limited connectivity.

Regarding the classification of the app as a medical device, the consensus among participants was that achieving this status could be feasible with future updates, further development, and rigorous validation. This would mean that the SkinNTDs App could expand its utility beyond merely training frontline health care workers, serving additionally as a diagnostic tool. However, this step could be very relevant, considering the positive impact that some projects incorporating artificial intelligence tools in LMIC have had. However, there are still numerous challenges and limitations that need to be taken into account beforehand, such as reliability issues, impacts on workflows, ease of use, and the importance of local context in the effectiveness of artificial intelligence tools [36].

Finally, participants deemed it important to discuss the most effective dissemination strategy for the app. Given that the WHO views this app as a valuable approach to addressing skin NTDs in LMICs, it is worth noting that only 2 participants in the study reported using other health apps in their work. Consensus was reached among all participants that the most favorable approach would be to collaborate with official institutions for dissemination purposes. However, while a marketing strategy could also be considered, studies affirm that health care marketing requires its own approach due to the unique features of the industry [37].

Considering that the app is still in development and has not yet been widely used in countries where skin NTDs are common, this study presents the opportunity to include end user feedback in future editions of the app. Furthermore, the impact of any subsequent implementation campaigns is likely to be influenced by this action.

Limitations

There were various limitations associated with this study that should be considered.

The first limitation was the decision to perform this study in Ghana and Kenya, based on 3 factors: (1) when the study was carried out, the SkinNTDs app could only be downloaded in English; hence, the choices were limited to countries where English is an official language; (2) t2 countries have a long history of web-based disease tracking and at least 8 skin NTDs are endemic in each; and (3) to ensure adequate sampling, countries were chosen based on the accessibility of key officials and program managers and their ability to contact enough respondents.

Regarding the methodology used, snowball sampling is a nonprobabilistic and nonrandomized method; thus, the sample may not be representative. However, because the study was conducted entirely on the web and the participants and authors were based in different places, this was the only reasonable way we could connect with frontline health care workers —a highly specialized group that can be challenging to reach in LMICs if not contacted remotely. The truth, however, revealed that this recruitment procedure was still challenging and would not have been accomplished without the direct assistance of 2 key actors from these 2 countries.

In addition, it is also important to point out some limitations of uMARS. First, this tool has only been validated in young people

XSL•FO
RenderX

using 2 specific apps related to other health areas [26]. However, its high reliability and the inclusion of both objective and subjective domains make it the only validated and simple questionnaire available for users to assess health apps.

Another requirement is that the frontline health care workers had to use the app for at least 5 days as opposed to the 10 minutes required by the original uMARS publication. Since there was no direct control over this requirement being met, this aspect could not be guaranteed. We asked participants during the semistructured interviews whether 5 days was enough time to test the app before answering the uMARS, and most (8/11, 73%) of them agreed that it was.

Several complications arose during the second semistructured part of the study, mainly due to not being able to carry out the interviews onsite because of COVID-19–related lockdowns and travel restrictions. The first issue was that no matter how many reminders were sent to the participants who agreed to be interviewed, only 17 finally participated. In fact, all those who participated in this part of the study were contacted directly by these 2 key agents. Moreover, during the web-based semistructured interviews, other complications appeared, such as participants constantly dropping in and out of the video call because they lost their internet connection, making it very difficult to create a warm environment for them to be more engaged. Similar difficulties arose in the focus group, and the expected interactions between the participants did not occur. As a result, many of them did not discuss their answers, leading to the exclusion of those responses from the study. It is important to acknowledge that this directly impacted the quality of the app's qualitative evaluation.

Finally, it is important to highlight that this study did not intend to assess the clinical effectiveness of the SkinNTDs app. This would require a randomized clinical trial and exceed the scope of this study. Although this study evaluated version 3, there are still issues to address, including whether information in the app is updated in response to the most recent research.

Conclusions

The SkinNTDs app represents a significant step forward in the use of mHealth technologies for the management of skin NTDs in LMICs. This study's findings that highlight the app's good usability and usefulness suggest its potential to be used as a training and support tool for frontline health care workers, without necessitating extensive customization. However, future development efforts must address identified weaknesses and consider user feedback to enhance the app's effectiveness and user satisfaction. By doing so, the SkinNTDs app can play a pivotal role in the global effort to combat skin NTDs and reduce their burden on affected communities.

In addition, its potential as a diagnostic tool merits exploration, and further research must evaluate this tool's clinical efficacy and determine the optimum method for achieving global diffusion.

Data Availability

The data sets generated during this study are available to be downloaded in the multimedia appendix. Specifically, Multimedia Appendix 2 contains data derived from the user version of the Mobile Application Rating Scale (uMARS), while Multimedia Appendix 3 features key responses from participants gathered through semistructured interviews or focus groups. All data considered confidential have been excluded from these resources.

Authors' Contributions

MC wrote the manuscript and conducted all quantitative and qualitative analyses, MC and CC conducted the semistructured interviews. PM conducted the focus group. EK, PM, ROP, and YAA played a key role in the participants' recruitment. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Deviations from the study protocol.
[DOCX File , 7 KB - jmir_v26i1e51628_app1.docx]

Multimedia Appendix 2
User version of the Mobile Application Rating Scale raw data derived from the cross-sectional study.
[XLSX File (Microsoft Excel File), 17 KB - jmir_v26i1e51628_app2.xlsx]

Multimedia Appendix 3
Data derived from the semistructured interviews and focus groups.
[DOCX File , 10 KB - jmir_v26i1e51628_app3.docx]

References

XSL•FO
RenderX

1. mHealth: use of appropriate digital technologies for public health. World Health Organization. 2018. URL: https://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_20-en.pdf [accessed 2023-01-02]
2. Karamagi HC, Muneene D, Droti B, Jepchumba V, Okeibunor JC, Nabyonga J, et al. eHealth or e-Chaos: the use of digital health interventions for health systems strengthening in sub-Saharan Africa over the last 10 years: a scoping review. J Glob Health 2022 Dec 03;12:04090 [FREE Full text] [doi: 10.7189/jogh.12.04090] [Medline: 36462201]
3. eHealth solutions in the African region: current context and perspectives. World Health Organization. 2010. URL: https://apps.who.int/iris/bitstream/handle/10665/1685/AFR-RC60-5.pdf?sequence=1&isAllowed=y [accessed 2023-03-02]
4. Tilahun B, Gashu KD, Mekonnen ZA, Endehabtu BF, Angaw DA. Mapping the role of digital health technologies in the case detection, management, and treatment outcomes of neglected tropical diseases: a scoping review. Trop Med Health 2021 Feb 22;49(1):17 [FREE Full text] [doi: 10.1186/s41182-021-00307-1] [Medline: 33618757]
5. Hotez PJ. The neglected tropical diseases and the neglected infections of poverty: overview of their common features, global disease burden and distribution, new control tools, and prospects for disease elimination. In: Choffnes ER, Relman DA, editors. The Causes and Impacts of Neglected Tropical and Zoonotic Diseases: Opportunities for Integrated Intervention Strategies. Washington, DC: National Academies Press; 2011.
6. Verrest L, Dorlo TP. Lack of clinical pharmacokinetic studies to optimize the treatment of neglected tropical diseases: a systematic review. Clin Pharmacokinet 2017 Jun 15;56(6):583-606 [FREE Full text] [doi: 10.1007/s40262-016-0467-3] [Medline: 27744580]
7. Adekeye O, Chowdhury S, McRae A, Olorunfemi T, Ozokede E, Dubukumah L, et al. Exploring the well-being of people affected by skin NTDs in Kaduna and Kwara States, Nigeria: a photovoice and scoping review study. Int Health 2023 Mar 24;15(Suppl 1):i100-i109 [FREE Full text] [doi: 10.1093/inthealth/ihad003] [Medline: 36960812]
8. da Conceição JR, Lopes CP, Ferreira EI, Epiphanio S, Giarolla J. Neglected tropical diseases and systemic racism especially in Brazil: from socio-economic aspects to the development of new drugs. Acta Trop 2022 Nov;235:106654. [doi: 10.1016/j.actatropica.2022.106654] [Medline: 35988823]
9. Mitra AK, Mawson AR. Neglected tropical diseases: epidemiology and global burden. Trop Med Infect Dis 2017 Aug 05;2(3):36 [FREE Full text] [doi: 10.3390/tropicalmed2030036] [Medline: 30270893]
10. Mitjà O, Marks M, Bertran L, Kollie K, Argaw D, Fahal AH, et al. Integrated control and management of neglected tropical skin diseases. PLoS Negl Trop Dis 2017 Jan 19;11(1):e0005136 [FREE Full text] [doi: 10.1371/journal.pntd.0005136] [Medline: 28103250]
11. Yotsu RR. Integrated management of skin NTDs-lessons learned from existing practice and field research. Trop Med Infect Dis 2018 Nov 14;3(4):120 [FREE Full text] [doi: 10.3390/tropicalmed3040120] [Medline: 30441754]
12. Tiwari R, Amien A, Visser WI, Chikte U. Counting dermatologists in South Africa: number, distribution and requirement. Br J Dermatol 2022 Aug 03;187(2):248-250 [FREE Full text] [doi: 10.1111/bjd.21036] [Medline: 35084741]
13. Tora A, Mengiste A, Davey G, Semrau M. Community involvement in the care of persons affected by podoconiosis-a lesson for other skin NTDs. Trop Med Infect Dis 2018 Aug 16;3(3):87 [FREE Full text] [doi: 10.3390/tropicalmed3030087] [Medline: 30274483]
14. Chandler DJ, Fuller LC. The Skin-A common pathway for integrating diagnosis and management of NTDs. Trop Med Infect Dis 2018 Sep 10;3(3):101 [FREE Full text] [doi: 10.3390/tropicalmed3030101] [Medline: 30274497]
15. Boateng LA, Ghana Southampton Scabies Research Partnership. Healthcare-seeking behaviour in reporting of scabies and skin infections in Ghana: a review of reported cases. Trans R Soc Trop Med Hyg 2020 Nov 06;114(11):830-837. [doi: 10.1093/trstmh/traa071] [Medline: 32853365]
16. Mieras LF, Taal AT, Post EB, Ndeve AG, van Hees CL. The development of a mobile application to support peripheral health workers to diagnose and treat people with skin diseases in resource-poor settings. Trop Med Infect Dis 2018 Sep 15;3(3):102 [FREE Full text] [doi: 10.3390/tropicalmed3030102] [Medline: 30274498]
17. Navarro A, Rubiano L, Arango JD, Rojas CA, Alexander N, Saravia NG, et al. Developing mobile health applications for neglected tropical disease research. PLoS Negl Trop Dis 2018 Nov;12(11):e0006791 [FREE Full text] [doi: 10.1371/journal.pntd.0006791] [Medline: 30383809]
18. Agarwal S, Perry HB, Long LA, Labrique AB. Evidence on feasibility and effective use of mHealth strategies by frontline health workers in developing countries: systematic review. Trop Med Int Health 2015 Aug 14;20(8):1003-1014 [FREE Full text] [doi: 10.1111/tmi.12525] [Medline: 25881735]
19. Neglected tropical diseases of the skin: WHO launches mobile application to facilitate diagnosis. World Health Organization. URL: https://www.who.int/news/item/16-07-2020-neglected-tropical-diseases-of-the-skin-who-launches-mobile-application-to-facilitate-diagnosis [accessed 2022-03-15]
20. Recognizing neglected tropical diseases through changes on the skin. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241513531 [accessed 2022-03-15]
21. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]

XSL•FO
RenderX

22. Evidence standards framework for digital health technologies. National Institutes for Health and Care Excellence. 2021. URL: https://www.nice.org.uk/corporate/ecd7 [accessed 2023-04-05]
23. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]
24. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://iris.who.int/bitstream/handle/10665/252183/9789241511766-eng.pdf?sequence=1 [accessed 2023-03-05]
25. Krah EF, de Kruijf JG. Exploring the ambivalent evidence base of mobile health (mHealth): a systematic literature review on the use of mobile phones for the improvement of community health in Africa. Digit Health 2016 Nov 24;2:2055207616679264 [FREE Full text] [doi: 10.1177/2055207616679264] [Medline: 29942576]
26. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
27. Frej A, Cano M, Ruiz-Postigo JA, Macharia P, Phillips RO, Amoako YA, et al. Assessing the quality of the world health organization's skin NTDs app as a training tool in Ghana and Kenya: protocol for a cross-sectional study. JMIR Res Protoc 2022 Dec 08;11(12):e39393 [FREE Full text] [doi: 10.2196/39393] [Medline: 36480252]
28. The Helsinki Declaration of the World Medical Association (WMA). Ethical principles of medical research involving human subjects. Pol Merkur Lekarski 2014;36:301. [doi: 10.1515/9783110208856.233]
29. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec 23;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
30. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019 Mar 08;7(2):e000057 [FREE Full text] [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
31. Wong LP. Focus group discussion: a tool for health and medical research. Singapore Med J 2008 Mar;49(3):256-261 [FREE Full text] [Medline: 18363011]
32. Yotsu RR, Itoh S, Yao KA, Kouadio K, Ugai K, Koffi YD, et al. The early detection and case management of skin diseases with an mHealth app (eSkinHealth): protocol for a mixed methods pilot study in Côte d'Ivoire. JMIR Res Protoc 2022 Sep 21;11(9):e39867 [FREE Full text] [doi: 10.2196/39867] [Medline: 35922062]
33. Carrion C, Robles N, Sola-Morales O, Aymerich M, Ruiz Postigo JA. Mobile health strategies to tackle skin neglected tropical diseases with recommendations from innovative experiences: systematic review. JMIR Mhealth Uhealth 2020 Dec 31;8(12):e22478 [FREE Full text] [doi: 10.2196/22478] [Medline: 33382382]
34. Rubiano L, Alexander ND, Castillo RM, Martínez Á, García Luna JA, Arango JD, et al. Adaptation and performance of a mobile application for early detection of cutaneous leishmaniasis. PLoS Negl Trop Dis 2021 Feb 11;15(2):e0008989 [FREE Full text] [doi: 10.1371/journal.pntd.0008989] [Medline: 33571192]
35. Mbunge E, Batani J, Gaobotse G, Muchemwa B. Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: a systematic review. Glob Health J 2022 Jun;6(2):102-113 [FREE Full text] [doi: 10.1016/j.glohj.2022.03.001] [Medline: 35282399]
36. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. NPJ Digit Med 2022 Oct 28;5(1):162 [FREE Full text] [doi: 10.1038/s41746-022-00700-y] [Medline: 36307479]
37. Purcarea VL. The impact of marketing strategies in healthcare systems. J Med Life 2019 Apr;12(2):93-96 [FREE Full text] [doi: 10.25122/jml-2019-1003] [Medline: 31406509]

Abbreviations

LMICs: low- and middle-income countries
mHealth: mobile health
NTD: neglected tropical disease
uMARS: user version of the Mobile Application Rating Scale
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 06.08.23; peer-reviewed by H Wang, C Fuller; comments to author 08.02.24; revised version received 27.02.24; accepted 08.03.24; published 30.04.24.

Please cite as:
Cano M, Ruiz-Postigo JA, Macharia P, Ampem Amoako Y, Odame Phillips R, Kinyeru E, Carrion C
Evaluating the World Health Organization's SkinNTDs App as a Training Tool for Skin Neglected Tropical Diseases in Ghana and Kenya: Cross-Sectional Study
J Med Internet Res 2024;26:e51628
URL: https://www.jmir.org/2024/1/e51628
doi:10.2196/51628
PMID:38687587

©Mireia Cano, José A Ruiz-Postigo, Paul Macharia, Yaw Ampem Amoako, Richard Odame Phillips, Esther Kinyeru, Carme Carrion. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation

Han Shi Jocelyn Chew[1], PhD; Nicholas WS Chew[2], MBBS; Shaun Seh Ern Loong[3], MBBS; Su Lin Lim[4], PhD; Wai San Wilson Tam[1], PhD; Yip Han Chin[3], MBBS; Ariana M Chao[5], PhD; Georgios K Dimitriadis[6], MBBS, MSc; Yujia Gao[7]; Jimmy Bok Yan So[8], MB ChB, FRCS, MPH; Asim Shabbir[8], MBBS, MMed, FRCS; Kee Yuan Ngiam[9], MBBS, FRCS

[1]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2]Department of Cardiology, National University Hospital, Singapore, Singapore
[3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[4]Department of Dietetics, National University Hospital, Singapore, Singapore
[5]School of Nursing, Johns Hopkins University, Baltimore, MD, United States
[6]Department of Endocrinology ASO/EASO COM, King's College Hospital NHS Foundation Trust, London, United Kingdom
[7]Division of Hepatobiliary & Pancreatic Surgery, Department of Surgery, National University Hospital, Singapore, Singapore
[8]Division of General Surgery (Upper Gastrointestinal Surgery), Department of Surgery, National University Hospital, Singapore, Singapore
[9]Division of Thyroid & Endocrine Surgery, Department of Surgery, National University Hospital, Singapore, Singapore

Corresponding Author:
Han Shi Jocelyn Chew, PhD
Alice Lee Centre for Nursing Studies
Yong Loo Lin School of Medicine
National University of Singapore
Level 3, Clinical Research Centre
Block MD11, 10 Medical Drive
Singapore, 117597
Singapore
Phone: 65 65168687
Email: jocelyn.chew.hs@nus.edu.sg

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e62767

Abstract

Background: A plethora of weight management apps are available, but many individuals, especially those living with overweight and obesity, still struggle to achieve adequate weight loss. An emerging area in weight management is the support for one's self-regulation over momentary eating impulses.

Objective: This study aims to examine the feasibility and effectiveness of a novel artificial intelligence–assisted weight management app in improving eating behaviors in a Southeast Asian cohort.

Methods: A single-group pretest-posttest study was conducted. Participants completed the 1-week run-in period of a 12-week app-based weight management program called the Eating Trigger-Response Inhibition Program (eTRIP). This self-monitoring system was built upon 3 main components, namely, (1) chatbot-based check-ins on eating lapse triggers, (2) food-based computer vision image recognition (system built based on local food items), and (3) automated time-based nudges and meal stopwatch. At every mealtime, participants were prompted to take a picture of their food items, which were identified by a computer vision image recognition technology, thereby triggering a set of chatbot-initiated questions on eating triggers such as who the users were eating with. Paired 2-sided *t* tests were used to compare the differences in the psychobehavioral constructs before and after the 7-day program, including overeating habits, snacking habits, consideration of future consequences, self-regulation of eating behaviors, anxiety, depression, and physical activity. Qualitative feedback were analyzed by content analysis according to 4 steps, namely, decontextualization, recontextualization, categorization, and compilation.

XSL•FO
RenderX

Results: The mean age, self-reported BMI, and waist circumference of the participants were 31.25 (SD 9.98) years, 28.86 (SD 7.02) kg/m^2, and 92.60 (SD 18.24) cm, respectively. There were significant improvements in all the 7 psychobehavioral constructs, except for anxiety. After adjusting for multiple comparisons, statistically significant improvements were found for overeating habits (mean –0.32, SD 1.16; P<.001), snacking habits (mean –0.22, SD 1.12; P<.002), self-regulation of eating behavior (mean 0.08, SD 0.49; P=.007), depression (mean –0.12, SD 0.74; P=.007), and physical activity (mean 1288.60, SD 3055.20 metabolic equivalent task-min/day; P<.001). Forty-one participants reported skipping at least 1 meal (ie, breakfast, lunch, or dinner), summing to 578 (67.1%) of the 862 meals skipped. Of the 230 participants, 80 (34.8%) provided textual feedback that indicated satisfactory user experience with eTRIP. Four themes emerged, namely, (1) becoming more mindful of self-monitoring, (2) personalized reminders with prompts and chatbot, (3) food logging with image recognition, and (4) engaging with a simple, easy, and appealing user interface. The attrition rate was 8.4% (21/251).

Conclusions: eTRIP is a feasible and effective weight management program to be tested in a larger population for its effectiveness and sustainability as a personalized weight management program for people with overweight and obesity.

Trial Registration: ClinicalTrials.gov NCT04833803; https://classic.clinicaltrials.gov/ct2/show/NCT04833803

(J Med Internet Res 2024;26:e46036) doi:10.2196/46036

KEYWORDS

artificial intelligence; chatbot; chatbots; weight; overweight; eating; food; weight loss; mHealth; mobile health; app; apps; applications; self-regulation; self-monitoring; anxiety; depression; consideration of future consequences; mental health; conversational agent; conversational agents; eating behavior; healthy eating; food consumption; obese; obesity; diet; dietary

Introduction

Overweight and obesity remain a public health concern that affects slightly more than half of the global adult population [1]. Across 52 Organization for Economic Co-operation and Development, Group of Twenty, and European Union 28 countries, treating conditions related to overweight and obesity costs US $425 billion per year, based on purchasing power parity. Each US dollar used to prevent obesity results in a 6-fold return in economic benefits [2]. Strategies for maintaining a healthy weight range from policy mandates on nutritional food labeling [3] to clinical treatments focused on lifestyle modifications, pharmacotherapy, and bariatric surgery [4]. However, the effectiveness of such strategies is limited by insurance coverage [5] and challenges with weight loss maintenance [6-9]. Some participants have been reported to regain up to 100% of their initial weight loss within 5 years [9,10].

With the rapid digitalization and smartphone penetration worldwide, weight loss apps have been gaining popularity, as they help overcome the temporospatial challenges of in-person weight loss programs [11]. For instance, participants enrolled in conventional weight management programs typically attend multiple face-to-face sessions at designated facilities, which could be burdensome and inconvenient as one needs to schedule appointments and travel to the facility that may be beyond one's usual mobility pattern. Moreover, such programs are resource-intensive, requiring a multidisciplinary team of trained health care professionals (eg, physicians, dietitians, physiotherapists, nurses), infrastructure (eg, counselling room), and equipment (eg, weighing scale, stadiometer) to maintain. Well-known apps that support weight loss in the market include MyFitnessPal [12], MyPlate Calorie Tracker [13], and Fitbit [14]. In Singapore, Healthy 365 [15] is available for the public, while nBuddy [16] is used for the clinical population. These apps mostly focus on calorie tracking, health status tracking, and progress monitoring. Increasingly, apps are enhanced with features that allow intuitive synchronization of health metrics across apps to provide a more holistic progress monitoring experience. With a fee, some apps even match users to a health coach who would provide personalized weight management plans to support weight loss. However, there is a need for apps that include monitoring and support for one's self-regulation over momentary eating impulses, which are often triggered and influenced by dietary lapse triggers such as visual food cues, eating out, negative affect, and sleep deprivation [17-20]. Self-regulation of eating behaviors during weight loss treatment commonly includes portion control, increasing fruit and vegetable consumption, reducing unhealthy food (sugar-sweetened beverages and high-fat food items) consumption, and reducing overall caloric consumption [17]. Therefore, we aimed to examine the feasibility and effectiveness of a novel artificial intelligence (AI)-assisted weight management app on improving eating behaviors and to explore the mechanism by which this app influences eating behaviors, as hypothesized in our earlier work [21].

Methods

Study Design

A single-group pretest-posttest study was conducted and reported according to the TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) checklist (Multimedia Appendix 1) [22]. Despite the limitations of the study design, it was deemed the most appropriate and feasible experimental study design for a preliminary understanding of the usability, acceptability, and effectiveness of the app [23].

Participant Recruitment

Participants older than 21 years with BMI ≥23 kg/m^2 and not undergoing a commercial weight loss program were recruited from January 2022 to October 2022 through social media platforms and physical recruitment at a local tertiary hospital's specialist weight management clinic in Singapore. Using

XSL•FO
RenderX

G*Power (version 3.1.9.7) [24], to detect a small effect size of 0.2 at .05 significance level and 80% power while accounting for an attrition rate of 20%, 248 participants are required. To be conservative, 250 participants were recruited.

Intervention

Immediately after completing the pretest questionnaire, participants were onboarded to the Eating Trigger-Response Inhibition Program (eTRIP) app by a trained research assistant to complete the 1-week run-in of the program. During the onboarding, participants were invited to enter their anthropometric details, desired weight loss goals, and motivation. They were also encouraged to personalize certain app functions such as the timing of the check-in prompts and preferred name for interaction with a chatbot. At every mealtime (at least 3 times a day), the participants were prompted to take a picture of their food, which was immediately recognized by a food-based computer vision image recognition technology, which then triggered a set of chatbot-initiated questions on eating triggers (eg, how they are feeling). These questions were developed based on our past work on eating behaviors [25-30]. Participants were able to view their image-based food log and eating habits on a dashboard, reflect upon their eating habits throughout the day, and set their goals and action plans for the next day. On the 8th day, all participants' user accounts were locked, and they were unable to make any changes but were able to still view their check-in logs. Participants could also provide feedback on the app by filling out the comments section in one of the app pages. Participants were reimbursed SGD 25 (SGD 1=US $0.74) for completing this program.

eTRIP App

The eTRIP app was developed as a 12-week AI-assisted, app-based, self-regulation program targeted at improving weight loss through healthy eating. eTRIP was developed largely based on a modified temporal self-regulation theory [31,32], behavioral change taxonomy [33], and our previous work on healthy eating and weight loss [27-29,34,35]. This includes studies on people with overweight and obesity in the areas of personal motivators, self-regulation facilitators, and barriers [27]; the potential of AI, apps, and chatbots in improving weight loss [6,25,29]; perceptions and needs of AI to increase its adoption in weight management [26]; and the essential elements of a weight loss app [28]. The development of eTRIP was split into 2 phases: (1) development of an AI-assisted self-monitoring system and (2) development of an AI-assisted behavioral nudging system. In this paper, we report the feasibility and effectiveness of an AI-assisted self-monitoring system after a 1-week run-in. The self-monitoring system is built upon 3 main components, namely, (1) chatbot-based check-ins on eating lapse triggers, (2) food-based computer vision image recognition (system built based on local food items), and (3) automated time-based nudges and meal stopwatch.

Outcomes

All participants completed the same self-report questionnaire before and after the 1-week run-in of the app, which reflected their sociodemographic profile, BMI, waist circumference, intention to improve eating behaviors, habits of overeating (Self-Report Habit Index) [36], habits of snacking [36], consideration of future consequences (Consideration of Future Consequences Scale-6 items) [37], self-regulation of eating behavior (Self-Regulation of Eating Behavior Questionnaire) [38], physical activity (International Physical Activity Questionnaire-Short Form) [39], anxiety symptoms (Generalized Anxiety Disorder-2 items) [40], and depressive symptoms (Patient Health Questionnaire-2 items) [41]. Details are reported in Multimedia Appendix 2. The primary outcomes were overeating habits, snacking habits, immediate thinking, self-regulation of eating habits, depression, anxiety, and physical activity. The secondary outcomes were their subscale scores.

Data Analysis

SPSS statistical software (version 27; IBM Corp) [42] was used for the analyses. The baseline characteristics of the participants were presented in mean (SD) and frequency (%). Paired 2-sided *t* tests were used to compare the differences in the psychobehavioral constructs before and after the 7-day program, including overeating habits, snacking habits, consideration of future consequences, self-regulation of eating behaviors, anxiety, depression, and physical activity. To account for the increased risk of a type 1 error due to multiple comparisons [43], the Bonferroni-corrected significant level was set to $P \leq .007$. Qualitative feedback were analyzed using content analysis according to 4 steps, namely, decontextualization, recontextualization, categorization, and compilation [44]. Feedback was first consolidated verbatim and read iteratively by 2 coders (Nagadarshini Nicole Rajasegaran and HSJC). The verbatim feedback was then analyzed independently by 2 reviewers into meaning units. Meaning units were then reconstituted, categorized, and reported as themes and subthemes.

Ethics Approval

This single-group pretest-posttest study was approved by the National Healthcare Group Domain Specific Review Board (ref 2020/01439), registered with the ClinicalTrials.gov (ref NCT04833803) on April 6, 2021.

Results

Baseline Characteristics of the Participants

A total of 251 participants were enrolled in this study (Chew HSJ, unpublished data, 2023); 20 (7.9%) participants dropped out of the 1-week program due to the inability to perform check-ins every day. Among those who completed the program (n=231), 1 participant was removed from the analyses due to ineligibility. The mean age, self-reported BMI, and waist circumference of the participants was 31.25 (SD 9.98) years, 28.86 (SD 7.02) kg/m^2, and 92.6 (SD 18.24) cm, respectively (Table 1). Approximately 47.8% (111/230) of the participants were males, indicating a good mix of participants from both sexes, and most of the participants were single (169/230, 73.6%), Chinese (181/230, 78.7%), and had a university education (148/230, 64.1%).

XSL•FO
RenderX

Table 1. Baseline characteristics of the participants who completed and who dropped out from the 1-week program.

Characteristics	Completed (n=230)	Dropped out (n=20)
Age (years), mean (SD)	31.25 (9.98)	36.25 (10.29)
Young adults (21-35 years), n (%)	168 (73)	11 (55)
Middle-aged adults (36-64 years), n (%)	62 (27)	9 (45)
Sex, n (%)		
Males	110 (48)	9 (45)
Females	120 (52)	11 (55)
Marital status, n (%)		
Single	169 (74)	7 (35)
Married	59 (26)	12 (60)
Divorced	2 (1)	1 (5)
Race, n (%)		
Chinese	181 (79)	7 (35)
Malay	19 (8)	7 (35)
Indian	24 (10)	5 (25)
Others	6 (3)	1 (5)
Religion, n (%)		
Buddhism	75 (33)	2 (10)
Christianity	57 (25)	3 (15)
Hinduism	16 (7)	2 (10)
Islam	25 (11)	10 (50)
Free thinker	43 (19)	3 (15)
Others	14 (6)	0 (0)
Highest education level, n (%)		
Primary school	2 (1)	0 (0)
Secondary school	9 (4)	2 (10)
Institute of Technical Education/polytechnic/junior college	71 (31)	6 (30)
University	148 (64)	12 (60)
Per capita household income (SGD/month)[a], n (%)		
<1000	21 (9)	2 (10)
1000-3000	47 (20)	6 (30)
3001-5000	57 (25)	3 (15)
5001-10,000	79 (34)	6 (30)
>10,000	26 (11)	3 (15)
Residential region, n (%)		
Central	41 (18)	3 (15)
East	17 (7)	3 (15)
North	22 (10)	1 (5)
Northeast	36 (16)	1 (5)
West	76 (33)	5 (25)
Smoking, n (%)		
No	220 (96)	18 (90)
Yes	10 (4)	2 (10)

XSL•FO
RenderX

Characteristics	Completed (n=230)	Dropped out (n=20)
Employment, n (%)		
Part-time	51 (22)	2 (10)
Full-time	171 (74)	18 (90)
Retired	9 (4)	0 (0)
BMI (kg/m^2), mean (SD)	28.86 (7.02)	33.73 (8.21)
Weight status, n (%)		
Overweight (23.0-27.4 kg/m^2, moderate risk)	142 (62)	4 (20)
Class I (27.5-32.4 kg/m^2, high risk)	41 (18)	6 (30)
Class II (32.5-37.4 kg/m^2, very high risk)	16 (7)	4 (20)
Class III (≥37.4 kg/m^2, very high risk)	31 (14)	6 (30)
BMI (*z* score), mean (SD)	–0.05 (0.97)	0.62 (1.14)
Waist circumference (cm), mean (SD)	92.60 (18.24)	102.75 (23.69)
Large waist circumference (male ≥90 cm; female ≥80 cm), n (%)	153 (67)	18 (90)
Waist circumference (*z* score), mean (SD)	–0.04 (0.97)	0.50 (1.26)
Generalized anxiety disorder-2 items score, mean (SD)	1.85 (1.63)	1.10 (0.85)
Potentially at risk cutoff (≥3), n (%)	61 (27)	0 (0)
Patient health questionnaire-2 items score, mean (SD)	1.56 (1.62)	1.3 (1.26)
Potentially at risk cutoff (≥3), n (%)	50 (22)	2 (10)

[a]SGD 1=US $0.74.

Mean Baseline Scores on Each Outcome Variable

The mean baseline scores on each outcome variable of the participants who completed and who dropped out from the 1-week program were calculated (Table 2). As the dropout rate was only 8.4% (20/251), statistical comparisons between those who dropped out and those who completed the program was not necessary.

XSL•FO
RenderX

Table 2. Mean baseline scores on each outcome variable of the participants who completed and who dropped out from the 1-week program.

Characteristics	Completed (n=230)	Dropped out (n=20)
Intention to change eating behaviors, mean (SD)	5.71 (1.16)	6.00 (1.30)
Overeating, mean (SD)		
SRHI[a]	4.28 (1.54)	4.11 (1.34)
Behavioral frequency	4.44 (1.63)	4.32 (1.56)
Automaticity	4.35 (1.60)	4.06 (1.46)
Self-identify	4.05 (1.59)	3.95 (1.3)
Snacking, mean (SD)		
SRHI	4.20 (1.47)	4.20 (1.08)
Behavioral frequency	4.46 (1.59)	4.83 (1.33)
Automaticity	4.03 (1.62)	3.85 (1.24)
Self-identify	4.10 (1.49)	3.91 (1.23)
CFCS-6[b], mean (SD)		
Total	4.34 (1.11)	4.67 (1.08)
Immediate	4.30 (1.56)	4.00 (1.57)
Future	4.98 (1.11)	5.33 (0.86)
SREBQ[c], mean (SD)	2.86 (0.53)	2.89 (0.42)
Low, n (%)	86 (37)	4 (20)
Moderate, n (%)	133 (58)	16 (80)
High, n (%)	11 (5)	0 (0)
IPAQ-SF[d] (MET[e]-min/week), mean (SD)	2230.76 (2503.66)[f]	1678.95 (3114.76)
Low, n (%)	76 (35)	12 (60)
Moderate, n (%)	63 (29)	2 (10)
Vigorous, n (%)	79 (34)	6 (30)
Identifies the following foods as tempting, n (%)		
Chocolate	149 (65)	15 (75)
Crisps	160 (70)	12 (60)
Cake	151 (66)	16 (80)
Ice cream	141 (61)	10 (50)
Bread/toast	100 (44)	9 (45)
Fizzy drinks	54 (24)	5 (25)
Biscuits	58 (25)	2 (10)
Sweets	43 (19)	3 (15)
Popcorn	34 (15)	3 (15)
Pastries	96 (42)	6 (30)
Pizza	91 (40)	9 (45)
Fried foods	138 (60)	11 (55)
Chips	106 (46)	12 (60)
Nil	8 (4)	0 (0)
Others	40 (17)	1 (5)

[a]SRHI: Self-Report Habit Index.

[b]CFCS-6: Consideration of Future Consequences Scale-6 items.

XSL•FO
RenderX

[c]SREBQ: Self-Regulation of Eating Behavior Questionnaire.
[d]IPAQ-SF: International Physical Activity Questionnaire-Short Form.
[e]MET: metabolic equivalent task.
[f]n=218.

Pretest and Posttest Mean Differences

There were significant improvements in all the 7 psychobehavioral constructs, except for anxiety. After adjusting for multiple comparisons, there were only statistically significant improvements in the overeating habit, snacking habit, self-regulation of eating behavior, depression, and physical activity (Table 3). Forty-one participants reported skipping at least 1 meal (ie, breakfast, lunch, or dinner), summing to a total of 578 (67.1%) of the 862 meals skipped.

Table 3. Mean differences in each outcome before and after the 1-week program for completers (n=230).

Outcomes	Pretest, mean (SD)	Posttest, mean (SD)	Mean difference (SD)	*t (df)*	*P* value
Overeating					
SRHI[a]	4.28 (1.54)	3.96 (0.10)	–0.32 (1.16)	–4.21 (458)	<.001[b]
Behavioral frequency	4.44 (1.63)	4.11 (1.61)	–0.33 (1.23)	–4.12 (458)	<.001[b]
Automaticity	4.36 (1.61)	4.00 (1.55)	–0.36 (1.31)	–4.23 (458)	<.001[b]
Self-identity	4.05 (1.59)	3.78 (1.52)	–0.27 (1.24)	–3.29 (458)	<.001[b]
Snacking					
SRHI	4.19 (1.48)	3.97 (1.45)	–0.22 (1.12)	–2.96 (458)	.002[b]
Behavioral frequency	4.46 (1.60)	4.20 (1.60)	–0.26 (1.25)	–3.09 (458)	<.001[b]
Automaticity	4.01 (1.62)	3.77 (1.54)	–0.24 (1.35)	–2.71 (458)	.004[b]
Self-identity	4.09 (1.50)	3.93 (1.47)	–0.16 (1.20)	–1.98 (458)	.02
CFCS-6[c]					
Total	4.34 (1.11)	4.45 (1.08)	0.10 (0.89)	1.75 (458)	.04
Immediate[d]	4.29 (1.57)	4.10 (1.55)	–0.18 (1.27)	–2.18 (458)	.015
Future[e]	4.98 (1.11)	5.00 (1.11)	0.02 (1.09)	0.30 (458)	.38
SREBQ[f]	2.85 (0.53)	2.93 (0.51)	0.08 (0.49)	2.46 (458)	.007[b]
Anxiety	1.92 (0.81)	1.84 (0.87)	–0.08 (0.80)	–1.58 (458)	.06
Depression	1.79 (0.81)	1.67 (0.84)	–0.12 (0.74)	–2.46 (458)	.007[b]
IPAQ-SF[g]					
Total (n=128)	2569.45 (2514.51)	3858.05 (3327.26)	1288.60 (3055.20)	4.77 (254)	<.001[b]
Vigorous (n=170)	939.53 (1279.50)	1121.18 (1314.25)	181.65 (1313.22)	1.80 (338)	.04
Moderate (n=146)	487.12 (855.41)	974.25 (1328.51)	487.12 (1379.22)	4.27 (290)	<.001[b]
Walk	1061.95 (1445.45)	1335.71 (1621.90)	273.76 (1649.22)	2.52 (458)	.006[b]

[a]SRHI: Self-Report Habit Index.
[b]Significant at $P<.007$.
[c]CFCS-6: Consideration of Future Consequences Scale-6 items.
[d]Consideration of Future Consequences Scale-6 immediate subscale.
[e]Consideration of Future Consequences Scale-6 future subscale.
[f]SREBQ: Self-Regulation of Eating Behavior Questionnaire.
[g]IPAQ-SF: International Physical Activity Questionnaire-Short Form.

User Engagement

Among those who completed the program, 97% (46,867/48,316) chatbot-based questions were completed. As participants were given the option to add additional check-ins for snacks, the percentage of completed check-ins could not be accurately computed.

XSL•FO
RenderX

Qualitative Feedback

Of the 230 participants, 80 (34.8%) provided textual feedback that indicated satisfactory experience with eTRIP. Four themes emerged, namely, (1) becoming more mindful of self-monitoring, (2) personalized reminders with prompts and chatbot, (3) food logging with image recognition, and (4) engaging with a simple, easy, and appealing user interface.

Becoming More Mindful of Self-Monitoring

By checking in with the app for every meal, the participants mentioned being more aware of their unhealthy eating habits and more mindful of their next meal. One participant said, "It (eTRIP) incentivizes me to stick to my diet plan because I am reminded of my diet plan daily. Ticking the box that indicates 'I did not meet my diet plan' made me guilty and it motivates me to opt for healthier food choice the next time round" (Female, Chinese, 22 years old). Another participant said, "I really liked the eTRIP app! Has a lot of potential for further expansion and use by more people. I like how it sends prompts during selected times of the day to be careful of what we see on social media. The rating of our mood before meals also helps me know how mood can affect my eating patterns. Lastly, the stopwatch function is great because it reminds me to eat more mindfully" (Male, Chinese, 27 years old).

Personalized Reminders With Prompts and Chatbot

Some participants mentioned the appreciation for reminders to check in with themselves in terms of the triggers of overeating. One participant said, "I like that there's a reminder to check in for every meal and users get to decide what time the app should prompt!" (Female, Indian, 27 years old). Some also suggested to develop the prompting system to prompt based on the user's previous check-in timings to optimize the prediction of mealtimes and prompt the check-in sessions intuitively. One participant suggested, "I think what would make this better is if you could aggregate the time the meals are entered from the past few days and estimate the time the user will normally eat and auto-adjust the timing…" (Female, Chinese, 23 years old). Others suggested to include reminders of how to make their meal options healthier, "Might be good to have reminders that reminds us to eat healthy with some tips on how to choose food" (Male, Chinese, 25 years old).

Food Logging With Image Recognition

Many participants highlighted their appreciation for the image recognition-based food logging, as it was accurate and convenient for food logging. One participant said, "It is very accurate in determining the food I've eaten just from the picture, and this saved me a lot of time from typing out the food I've eaten" (Male, Chinese, 21 years old).

Engaging With a Simple, Easy, and Appealing User Interface

All the participants who commented on the user experience expressed being impressed with the user interface and structure. One participant said, "the flow was smooth, quite clear. Graphics were cute. Very easy to input my info (information) especially from the homepage, I like how there's the ability to skip a meal" (Female, Malay, 25 years old). Another participant said, "The app is very smart, … yes it's very easy to fill and I loss (lost) like 0.5kg?" (Female, Chinese, 25 years old).

Participants' Suggestions

In terms of the areas for improvement, the participants preferred to have (1) more options and rating scales for each domain of eating trigger instead of typing out in the "others" field (although there was a stored text for repeated entries); (2) summary of the instances where one was able to achieve the goal of the day, which the user sets daily for the next day (based on a user preset list of goals); (3) examples of standard portions and frequency of meals; and (4) feedback on how to improve upon the unhealthy meals logged.

Discussion

Overview

Real-time interventions that can effectively address eating lapse triggers and improve eating behavior self-regulation, lapse events, weight loss, and weight maintenance remain unclear [45]. OnTrack, a just-in-time adaptive intervention that has been tested, is a smartphone app that uses machine learning to predict dietary lapses based on the repeated assessments of lapse triggers (ecological momentary assessment). OnTrack is used in conjunction with existing weight loss apps such as WeightWatchers app and provides personalized recommendations to prevent dietary lapses. The compliance rate for completing the lapse trigger survey in OnTrack was 62.9% over 3 months, and the studied sample was mostly females who were Whites [46]. Evidence has shown that factors influencing obesity and overweight are population-specific, influenced by socioeconomic, cultural, and genetic factors among others [45,47]. Singapore is a multiethnic society with a unique food culture influenced by various racial beliefs and traditions [48]. The differences in geographical, social, environmental, and genetic characteristics could define a different set of triggers and response to such weight loss apps.

Principal Findings

In this paper, we report the effectiveness of a weeklong AI-assisted weight loss app for improving overeating habits, snacking habits, immediate thinking, self-regulation of eating habits, depression, and physical activity. Interestingly, there were no significant improvements in the anxiety symptoms before and after using eTRIP, potentially due to the already low level of anxiety in those who completed the program (ie, ceiling effect). We also report corresponding qualitative user feedback on the experience with using eTRIP, where the users appreciated the app for enabling them to become more mindful of self-monitoring; personalized reminders with prompts and chatbot; food logging with image recognition; and engaging with a simple, easy, and appealing user interface. The significant improvements observed among the participants in this study reveal the potential of this app to influence weight loss in the context of a Southeast Asian cohort with overweight and obesity. The qualitative feedback also informs future app development to enhance user engagement and reduce dropout rates.

Eating habits contribute to overweight and obesity [49,50]. Encouraged by an obesogenic environment, overeating is

XSL•FO
RenderX

commonly triggered by situational factors such as food novelty or variety, social company (eg, eating with certain people), affect emotional states (which trigger emotional eating), and distractions (eg, concurrent tasks) [30,50-53]. Other studies have suggested that people at risk for obesity exhibit hyperresponsivity in the neural reward system to calorie-dense foods, which is associated with increased food consumption [54]. Alongside users' feedback that the app made them more mindful of their eating patterns, the significant improvement in overeating and snacking habits could have been due to an increased awareness of one' maladaptive eating habits and subsequently, the motivation to change. This coincides with a review that reported the effectiveness of mindful eating interventions on reducing food consumption in people with overweight and obesity [55]. Our qualitative findings showed that by self-monitoring one's eating behavior through chatbot-initiated check-ins, one could enhance mindful eating and reduce overeating without the need for undergoing mindful eating training. This could eventually lead to a reduction in total food consumption and weight loss. However, more quantitative evidence is needed to support this point.

It is noteworthy that some participants reported skipping meals as planned, which might have led to reduced energy consumption. However, this has to be examined further, as studies have shown that the calories avoided during a skipped meal may be compensated by an increase in snacking or overeating during mealtimes [56]. One additional element that can be explored in future studies is the effectiveness of promoting healthy snacking, which includes snacking on foods rich in proteins, fruits, vegetables, and whole grains, as opposed to nutrient-poor and energy-dense foods [57,58]. These healthier alternatives have been found not only to be associated with earlier satiety but also to be more nutritious, with their contents being more consistent with the established dietary recommendations and guidelines [59,60]. This strategy can be explored in conjunction with the current approach to decrease participants' overall snacking habits.

The improvement in the self-regulation of eating habits could be attributed to several factors, including the app content focused on reminding participants of their weight loss goals and to adopt healthier eating habits of less snacking and overeating during mealtimes. In particular, in commonly stigmatized populations like those with overweight and obesity, personalization of interventions enhances one's feeling of being taken care of, nurtured, and respected, providing them with a sense of confidence [61]. This may have improved individuals' willingness to engage with the eTRIP content, knowing that they would be well-respected and seen as individuals through personalized chatbot conversations and reminders [62]. Other studies have shown that personalized eHealth interventions are more effective than conventional programs in enhancing weight loss maintenance, BMI, waist circumference, and various other metabolic indicators [63].

In conjunction with improvements in eating habits, participants also engaged in greater levels of physical activity by the end of this study. Increased health consciousness and self-education about the impacts and types of physical activity are factors that may explain the observed increased level of physical activity among the participants [64]. Various studies have found that a combination of diet and exercise is superior to diet-only interventions in inducing weight loss [65]. The level of physical activity is also an important factor for improving long-term weight loss [66-68]. Moderate amounts of physical activity were observed to prevent weight regain after weight loss [65,68,69]. In addition, the American College of Sports Medicine recommends 200-300 minutes of moderate physical activity a week to prevent similar weight regain [70]. In our study, low and moderate levels of exercise were seen to increase significantly among the participants. Although the sustainability of the increase in the exercise levels is still unknown, the preliminary data are encouraging to show the potential of the app in impacting physical activity. The amounts of high levels of exercise were, however, not impacted significantly. Additional interventions, including the provision of educational materials about the benefits of and types of exercise, along with personalized reminders for exercise can potentially further increase the success of the app in increasing moderate and high levels of exercise among its participants [71,72].

In addition to improvements in the eating habits and levels of physical activity, there were changes in the psychological factors among the participants. The mean depressive symptoms were significantly decreased at the end of the weeklong program. Various studies have shown that healthy living characterized by various factors such as healthy eating and sufficient levels of physical activity have the potential to positively impact psychological factors such as mood and emotions [73]. Healthy eating with adherence to dietary recommendations has been found to reduce the levels of inflammation, increase the levels of various micronutrients such as vitamins, and regulate the levels of simple sugars, all of which are protective against mental illnesses, especially depression [73-76]. Studies have also shown that the use of chatbots in the app may decrease depressive symptoms among some participants. Chatbots provide individuals the ability to provide self-care in an environment that is neither costly nor stigmatizing [77]. This may enable participants to be more open with their emotions, as well as to have an outlet to gain relief through their interaction with the chatbot as a proxy of human interaction [77]. Studies have shown that improvements in psychological factors such as depressive symptoms have been positively associated with weight loss and maintenance, further increasing the effectiveness of weight loss efforts among participants with overweight and obesity [45].

Strengths and Limitations

This study was the first to characterize the effectiveness of an AI-assisted weight loss app in the context of a Southeast Asian cohort. One strength of this study was the demographics of the participants, which was generally representative of the Singaporean population in terms of sex and ethnicity. The consideration of population-specific determinants of obesity and overweight during the design of the app would have also increased its applicability in this population [78], having considered the various nuances and practical needs of its target demographic. For example, the food image recognition system, which was built based on local food items, reduced the amount of time and effort required for the logging of food, improved

XSL•FO
RenderX

the usability of the intervention, and enhanced user experience. The success of this app thus provides evidence that the consideration of population-specific underpinnings and practical requirements were essential toward the successful design and implementation of a weight loss intervention [79].

Although the app presents significant potential in this weeklong trial, this study is limited due to its short time frame. This presents with difficulties in understanding the midterm to long-term impacts of using the app. However, it is reassuring that despite the limited time frame of this study, the various behavioral and psychological indicators were observed to be significantly improved. Through the feedback gathered from the participants, the app may be improved in specific aspects, including (1) refining choices available for the various survey fields such as the provision of a drop-down menu for the selection of weight loss goals; (2) providing additional feedback and weekly summaries to the participants for knowledge of their progress in various aspects; and (3) providing educational materials to provide participants with the means to improve, especially for what to do after eating lapses and suggestions for healthy snacking. Another limitation was the lack of feedback quotes from older individuals as opposed to those from younger individuals. This could be due to various reasons, of which decreased media literacy among older individuals might present additional obstacles for the provision of feedback [80]. Lastly, we did not collect information on the participants' medical and pharmacological history, where certain diseases and drugs are known to influence weight gain through various metabolic and neural pathways. Weight and waist circumference were also self-reported, and thus, data from these measures should be interpreted cautiously.

Conclusion

This study was the first to characterize the effectiveness of an AI-assisted weight loss app in the context of a Southeast Asian cohort. The positive findings of this study show the feasibility of implementing this app and the large potential it has in impacting weight loss efforts, especially among individuals with overweight and obesity. Efforts should be made to lengthen and upscale this program for a greater understanding of the midterm to long-term effects of this app.

Conflicts of Interest

AMC has served on advisory boards to Eli Lilly and Boehringer Ingelheim and received grant support, on behalf of the University of Pennsylvania, from Eli Lilly and WW (Weight Watchers). No other authors declare conflicts of interest.

Multimedia Appendix 1
TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) checklist.
[DOCX File , 29 KB - jmir_v26i1e46036_app1.docx]

Multimedia Appendix 2
Details on outcome measures.
[DOCX File , 20 KB - jmir_v26i1e46036_app2.docx]

References

1. Abdelaal M, le Roux CW, Docherty NG. Morbidity and mortality associated with obesity. Ann Transl Med 2017 Apr;5(7):161 [FREE Full text] [doi: 10.21037/atm.2017.03.107] [Medline: 28480197]
2. OECD. The heavy burden of obesity: the economics of prevention. OECD Health Policy Studies 2019 Oct 10:1-100 [FREE Full text] [doi: 10.1787/67450d67-en]
3. Hawkes C, Smith TG, Jewell J, Wardle J, Hammond RA, Friel S, et al. Smart food policies for obesity prevention. The Lancet 2015 Jun;385(9985):2410-2421. [doi: 10.1016/s0140-6736(14)61745-1]
4. Johns DJ, Hartmann-Boyce J, Jebb SA, Aveyard P, Behavioural Weight Management Review Group. Diet or exercise interventions vs combined behavioral weight management programs: a systematic review and meta-analysis of direct comparisons. J Acad Nutr Diet 2014 Oct;114(10):1557-1568 [FREE Full text] [doi: 10.1016/j.jand.2014.07.005] [Medline: 25257365]
5. Clarke B, Kwon J, Swinburn B, Sacks G. Understanding the dynamics of obesity prevention policy decision-making using a systems perspective: A case study of Healthy Together Victoria. PLoS One 2021;16(1):e0245535 [FREE Full text] [doi: 10.1371/journal.pone.0245535] [Medline: 33481898]
6. Chew HSJ, Koh WL, Ng JSHY, Tan KK. Sustainability of weight loss through smartphone apps: systematic review and meta-analysis on anthropometric, metabolic, and dietary outcomes. J Med Internet Res 2022 Sep 21;24(9):e40141 [FREE Full text] [doi: 10.2196/40141] [Medline: 36129739]
7. Booth HP, Prevost TA, Wright AJ, Gulliford MC. Effectiveness of behavioural weight loss interventions delivered in a primary care setting: a systematic review and meta-analysis. Fam Pract 2014 Dec;31(6):643-653 [FREE Full text] [doi: 10.1093/fampra/cmu064] [Medline: 25298510]
8. LeBlanc ES, Patnode CD, Webber EM, Redmond N, Rushkin M, O'Connor EA. Behavioral and pharmacotherapy weight loss interventions to prevent obesity-related morbidity and mortality in adults: updated evidence report and systematic

XSL•FO
RenderX

review for the US Preventive Services Task Force. JAMA 2018 Sep 18;320(11):1172-1191. [doi: 10.1001/jama.2018.7777] [Medline: 30326501]
9. MacLean PS, Wing RR, Davidson T, Epstein L, Goodpaster B, Hall KD, et al. NIH working group report: Innovative research to improve maintenance of weight loss. Obesity (Silver Spring) 2015 Jan;23(1):7-15 [FREE Full text] [doi: 10.1002/oby.20967] [Medline: 25469998]
10. Daley A, Jolly K, Madigan C, et al. A brief behavioural intervention to promote regular self-weighing to prevent weight regain after weight loss: a RCT. Public Health Research 2019:7. [doi: 10.3310/phr07070] [Medline: 31042335]
11. Hadžiabdić MO, Mucalo I, Hrabač P, Matić T, Rahelić D, Božikov V. Factors predictive of drop-out and weight loss success in weight management of obese patients. J Hum Nutr Diet 2015 Feb;28 Suppl 2:24-32. [doi: 10.1111/jhn.12270] [Medline: 25220046]
12. Evans D. MyFitnessPal. Br J Sports Med 2016 Jan 27;51(14):1101-1102. [doi: 10.1136/bjsports-2015-095538]
13. Garcia R. Utilization, Integration & Evaluation of LIVESTRONG's MyPlate Telehealth Technology. URL: https://www.researchgate.net/publication/273635099_Utilization_Integration_Evaluation_of_LIVESTRONG's_MyPlate_Telehealth_Technology [accessed 2023-04-01]
14. Hartman SJ, Nelson SH, Weiner LS. Patterns of Fitbit use and activity levels throughout a physical activity intervention: exploratory analysis from a randomized controlled trial. JMIR Mhealth Uhealth 2018 Feb 05;6(2):e29 [FREE Full text] [doi: 10.2196/mhealth.8503] [Medline: 29402761]
15. Yao J, Tan CS, Chen C, Tan J, Lim N, Müller-Riemenschneider F. Bright spots, physical activity investments that work: National Steps Challenge, Singapore: a nationwide mHealth physical activity programme. Br J Sports Med 2020 Sep;54(17):1047-1048. [doi: 10.1136/bjsports-2019-101662] [Medline: 31857340]
16. Lim SL, Johal J, Ong KW, Han CY, Chan YH, Lee YM, et al. Lifestyle intervention enabled by mobile technology on weight loss in patients with nonalcoholic fatty liver disease: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 13;8(4):e14802 [FREE Full text] [doi: 10.2196/14802] [Medline: 32281943]
17. Birkett D. Disinhibition. In: The Psychiatry of Stroke. London, UK: Routledge; 2012:150-161.
18. Carels RA, Hoffman J, Collins A, Raber AC, Cacciapaglia H, O'Brien WH. Ecological momentary assessment of temptation and lapse in dieting. Eat Behav 2001;2(4):307-321. [doi: 10.1016/s1471-0153(01)00037-x] [Medline: 15001025]
19. Kwasnicka D, Dombrowski SU, White M, Sniehotta FF. N-of-1 study of weight loss maintenance assessing predictors of physical activity, adherence to weight loss plan and weight change. Psychol Health 2017 Jun;32(6):686-708 [FREE Full text] [doi: 10.1080/08870446.2017.1293057] [Medline: 28323457]
20. McKee HC, Ntoumanis N, Taylor IM. An ecological momentary assessment of lapse occurrences in dieters. Ann Behav Med 2014 Dec;48(3):300-310. [doi: 10.1007/s12160-014-9594-y] [Medline: 24562984]
21. Chew HSJ, Loong SSE, Lim SL, Tam WSW, Chew NWS, Chin YH, et al. Socio-demographic, behavioral and psychological factors associated with high BMI among adults in a Southeast Asian multi-ethnic society: a structural equation model. Nutrients 2023 Apr 10;15(8):1826 [FREE Full text] [doi: 10.3390/nu15081826] [Medline: 37111045]
22. Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004 Mar;94(3):361-366. [doi: 10.2105/ajph.94.3.361] [Medline: 14998794]
23. Knapp T. Why is the one-group pretest-posttest design still used? Clin Nurs Res 2016 Oct;25(5):467-472. [doi: 10.1177/1054773816666280] [Medline: 27558917]
24. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
25. Chew HSJ. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022 Apr 13;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
26. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. J Med Internet Res 2022 Jan 14;24(1):e32939 [FREE Full text] [doi: 10.2196/32939] [Medline: 35029538]
27. Chew HSJ, Gao Y, Shabbir A, Lim SL, Geetha K, Kim G, et al. Personal motivation, self-regulation barriers and strategies for weight loss in people with overweight and obesity: a thematic framework analysis. Public Health Nutr 2022 Feb 22;25(9):2426-2435. [doi: 10.1017/s136898002200043x]
28. Chew H, Lim S, Kim G, Kayambu G, So BYJ, Shabbir A, et al. Essential elements of weight loss apps for a multi-ethnic population with high BMI: a qualitative study with practical recommendations. Transl Behav Med 2023 Apr 03;13(3):140-148. [doi: 10.1093/tbm/ibac090] [Medline: 36689306]
29. Chew HSJ, Ang WHD, Lau Y. The potential of artificial intelligence in enhancing adult weight loss: a scoping review. Public Health Nutr 2021 Feb 17;24(8):1993-2020. [doi: 10.1017/s1368980021000598]
30. Chew HSJ, Lau ST, Lau Y. Weight-loss interventions for improving emotional eating among adults with high body mass index: A systematic review with meta-analysis and meta-regression. Eur Eat Disord Rev 2022 Jul;30(4):304-327 [FREE Full text] [doi: 10.1002/erv.2906] [Medline: 35460323]

XSL•FO
RenderX

31. Chew HSJ, Sim KLD, Choi KC, Chair SY. Effectiveness of a nurse-led temporal self-regulation theory-based program on heart failure self-care: A randomized controlled trial. Int J Nurs Stud 2021 Mar;115:103872. [doi: 10.1016/j.ijnurstu.2021.103872] [Medline: 33516047]
32. Hall PA, Fong GT. Temporal self-regulation theory: A model for individual health behavior. Health Psychology Review 2007 Mar;1(1):6-52. [doi: 10.1080/17437190701492437]
33. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
34. Chew HSJ, Li J, Chng S. Improving adult eating behaviours by manipulating time perspective: a systematic review and meta-analysis. Psychol Health 2023 Jan 24:1-17. [doi: 10.1080/08870446.2023.2169320] [Medline: 36691788]
35. Chew HSJ, Rajasegaran NN, Chng S. Effectiveness of interactive technology-assisted interventions on promoting healthy food choices: a scoping review and meta-analysis. Br J Nutr 2023 Jan 25;130(7):1250-1259. [doi: 10.1017/s0007114523000193]
36. Verplanken B, Orbell S. Reflections on past behavior: a self‐report index of habit strength. J Applied Social Pyschol 2006 Jul 31;33(6):1313-1330. [doi: 10.1111/j.1559-1816.2003.tb01951.x]
37. Chng S, Chew HSJ, Joireman J. When time is of the essence: Development and validation of brief consideration of future (and immediate) consequences scales. Personality and Individual Differences 2022 Feb;186:111362. [doi: 10.1016/j.paid.2021.111362]
38. Kliemann N, Beeken RJ, Wardle J, Johnson F. Development and validation of the Self-Regulation of Eating Behaviour Questionnaire for adults. Int J Behav Nutr Phys Act 2016 Aug 02;13:87 [FREE Full text] [doi: 10.1186/s12966-016-0414-6] [Medline: 27484457]
39. Craig CL, Marshall AL, Sjöström M, Bauman AE, et al. International Physical Activity Questionnaire: 12-country reliability and validity. Medicine & Science in Sports & Exercise 2003;35(8):1381-1395. [doi: 10.1249/01.mss.0000078924.61453.fb]
40. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 06;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
41. Kroenke K, Spitzer RL, Williams JBW. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003 Nov;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
42. IBM SPSS statistics for Windows. IBM Corp. 2021 May 21. URL: https://www.ibm.com/products/spss-statistics [accessed 2024-03-28]
43. VanderWeele T, Mathur M. Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad? Am J Epidemiol 2019 Mar 01;188(3):617-618 [FREE Full text] [doi: 10.1093/aje/kwy250] [Medline: 30452538]
44. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
45. Varkevisser RDM, van Stralen MM, Kroeze W, Ket JCF, Steenhuis IHM. Determinants of weight loss maintenance: a systematic review. Obes Rev 2019 Feb;20(2):171-211 [FREE Full text] [doi: 10.1111/obr.12772] [Medline: 30324651]
46. Forman EM, Goldstein SP, Crochiere RJ, Butryn ML, Juarascio AS, Zhang F, et al. Randomized controlled trial of OnTrack, a just-in-time adaptive intervention designed to enhance weight loss. Transl Behav Med 2019 Nov 25;9(6):989-1001. [doi: 10.1093/tbm/ibz137] [Medline: 31602471]
47. Qasim A, Turcotte M, de Souza RJ, Samaan MC, Champredon D, Dushoff J, et al. On the origin of obesity: identifying the biological, environmental and cultural drivers of genetic risk among human populations. Obes Rev 2018 Feb;19(2):121-149. [doi: 10.1111/obr.12625] [Medline: 29144594]
48. Executive summary on National Population Health Survey 2016/17. Singapore MoH. URL: https://www.moh.gov.sg/docs/librariesprovider5/resources-statistics/reports/executive-summary-nphs-2016_17.pdf [accessed 2023-04-01]
49. McCrory MA, Suen VM, Roberts SB. Biobehavioral influences on energy intake and adult weight gain. The Journal of Nutrition 2002 Dec;132(12):3830S-3834S. [doi: 10.1093/jn/132.12.3830s]
50. Davis C, Levitan RD, Muglia P, Bewell C, Kennedy JL. Decision-making deficits and overeating: a risk model for obesity. Obes Res 2004 Jun;12(6):929-935 [FREE Full text] [doi: 10.1038/oby.2004.113] [Medline: 15229331]
51. Hetherington MM. Cues to overeat: psychological factors influencing overconsumption. Proc. Nutr. Soc 2007 Feb 28;66(1):113-123. [doi: 10.1017/s0029665107005344]
52. Borer K. Understanding human physiological limitations and societal pressures in favor of overeating helps to avoid obesity. Nutrients 2019 Jan 22;11(2):227. [doi: 10.3390/nu11020227]
53. Borer KT. Why we eat too much, have an easier time gaining than losing weight, and expend too little energy: suggestions for counteracting or mitigating these problems. Nutrients 2021 Oct 26;13(11):3812. [doi: 10.3390/nu13113812]
54. Stice E, Burger K. Neural vulnerability factors for obesity. Clin Psychol Rev 2019 Mar;68:38-53 [FREE Full text] [doi: 10.1016/j.cpr.2018.12.002] [Medline: 30587407]
55. Warren JM, Smith N, Ashwell M. A structured literature review on the role of mindfulness, mindful eating and intuitive eating in changing eating behaviours: effectiveness and associated potential mechanisms. Nutr. Res. Rev 2017 Jul 18;30(2):272-283. [doi: 10.1017/s0954422417000154]

XSL•FO
RenderX

56. Savige G, Macfarlane A, Ball K, Worsley A, Crawford D. Snacking behaviours of adolescents and their association with skipping meals. Int J Behav Nutr Phys Act 2007 Sep 17;4:36 [FREE Full text] [doi: 10.1186/1479-5868-4-36] [Medline: 17868479]
57. Njike VY, Smith TM, Shuval O, Shuval K, Edshteyn I, Kalantari V, et al. Snack food, satiety, and weight. Adv Nutr 2016 Sep;7(5):866-878 [FREE Full text] [doi: 10.3945/an.115.009340] [Medline: 27633103]
58. Larson N, Story M. A review of snacking patterns among children and adolescents: what are the implications of snacking for weight status? Child Obes 2013 Apr;9(2):104-115. [doi: 10.1089/chi.2012.0108] [Medline: 23470091]
59. Lucan SC, Karpyn A, Sherman S. Storing empty calories and chronic disease risk: snack-food products, nutritive content, and manufacturers in Philadelphia corner stores. J Urban Health 2010 May;87(3):394-409 [FREE Full text] [doi: 10.1007/s11524-010-9453-5] [Medline: 20405225]
60. Hess J, Jonnalagadda S, Slavin J. What is a snack, why do we snack, and how can we choose better snacks? The definitions of snacking, motivations to snack, contributions to dietary intake, and recommendations for improvement. The FASEB Journal 2016 Apr;30(S1):466-475. [doi: 10.1096/fasebj.30.1_supplement.895.5]
61. Coulter A, Entwistle V, Eccles A, Ryan S, Shepperd S, Perera R. Personalised care planning for adults with chronic or long-term health conditions. Cochrane Database Syst Rev 2015;2015(3):1-117. [doi: 10.1002/14651858.cd010523.pub2]
62. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367 [FREE Full text] [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
63. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: A systematic review and meta-analysis of randomised controlled trials. Prev Med 2020 Mar;132:106001. [doi: 10.1016/j.ypmed.2020.106001] [Medline: 31991155]
64. Wang Q, Egelandsdal B, Amdam GV, Almli VL, Oostindjer M. Diet and physical activity apps: perceived effectiveness by app users. JMIR Mhealth Uhealth 2016 Apr 07;4(2):e33 [FREE Full text] [doi: 10.2196/mhealth.5114] [Medline: 27056639]
65. Kim B, Choi D, Jung C, Kang S, Mok J, Kim C. Obesity and physical activity. J Obes Metab Syndr 2017 Mar;26(1):15-22 [FREE Full text] [doi: 10.7570/jomes.2017.26.1.15] [Medline: 31089489]
66. Anderson JW, Konz EC, Frederich RC, Wood CL. Long-term weight-loss maintenance: a meta-analysis of US studies. Am J Clin Nutr 2001 Nov;74(5):579-584. [doi: 10.1093/ajcn/74.5.579] [Medline: 11684524]
67. Fogelholm M, Kukkonen-Harjula K. Does physical activity prevent weight gain--a systematic review. Obes Rev 2000 Oct;1(2):95-111. [doi: 10.1046/j.1467-789x.2000.00016.x] [Medline: 12119991]
68. Thomas DM, Bouchard C, Church T, Slentz C, Kraus WE, Redman LM, et al. Why do individuals not lose more weight from an exercise intervention at a defined dose? An energy balance analysis. Obesity Reviews 2012 Jun 11;13(10):835-847. [doi: 10.1111/j.1467-789x.2012.01012.x]
69. Saris WHM, Blair SN, van Baak MA, Eaton SB, Davies PSW, Di Pietro L, et al. How much physical activity is enough to prevent unhealthy weight gain? Outcome of the IASO 1st Stock Conference and consensus statement. Obes Rev 2003 May;4(2):101-114. [doi: 10.1046/j.1467-789x.2003.00101.x] [Medline: 12760445]
70. Haskell W, Lee I, Pate R, Powell K, Blair S, Franklin B. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007;39(8):1423-1434. [doi: 10.1249/mss.0b013e3180616b27]
71. Fredriksson SV, Alley SJ, Rebar AL, Hayman M, Vandelanotte C, Schoeppe S. How are different levels of knowledge about physical activity associated with physical activity behaviour in Australian adults? PLoS One 2018;13(11):e0207003 [FREE Full text] [doi: 10.1371/journal.pone.0207003] [Medline: 30485310]
72. Muntaner-Mas A, Sanchez-Azanza VA, Ortega FB, Vidal-Conti J, Borràs PA, Cantallops J, et al. The effects of a physical activity intervention based on a fatness and fitness smartphone app for University students. Health Informatics J 2021;27(1):1460458220987275 [FREE Full text] [doi: 10.1177/1460458220987275] [Medline: 33446036]
73. Hautekiet P, Saenen ND, Martens DS, Debay M, Van der Heyden J, Nawrot TS, et al. A healthy lifestyle is positively associated with mental health and well-being and core markers in ageing. BMC Med 2022 Sep 29;20(1):328 [FREE Full text] [doi: 10.1186/s12916-022-02524-9] [Medline: 36171556]
74. Lassale C, Batty GD, Baghdadli A, Jacka F, Sánchez-Villegas A, Kivimäki M, et al. Healthy dietary indices and risk of depressive outcomes: a systematic review and meta-analysis of observational studies. Mol Psychiatry 2019 Jul;24(7):965-986 [FREE Full text] [doi: 10.1038/s41380-018-0237-8] [Medline: 30254236]
75. Petridou ET, Kousoulis AA, Michelakos T, Papathoma P, Dessypris N, Papadopoulos FC, et al. Folate and B12 serum levels in association with depression in the aged: a systematic review and meta-analysis. Aging Ment Health 2016 Sep;20(9):965-973. [doi: 10.1080/13607863.2015.1049115] [Medline: 26055921]
76. Wang C, Yang T, Wang G, Zhao Y, Yang L, Bi B. Association between dietary patterns and depressive symptoms among middle-aged adults in China in 2016-2017. Psychiatry Res 2018 Feb;260:123-129. [doi: 10.1016/j.psychres.2017.11.052] [Medline: 29182923]

XSL•FO
RenderX

77. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
78. Population in brief. National Population and Talent Division SG. 2022. URL: https://www.strategygroup.gov.sg/files/media-centre/publications/population-in-brief-2022.pdf [accessed 2023-04-01]
79. Hebden L, Cook A, van der Ploeg HP, King L, Bauman A, Allman-Farinelli M. A mobile health intervention for weight management among young adults: a pilot randomised controlled trial. J Hum Nutr Diet 2014 Aug;27(4):322-332. [doi: 10.1111/jhn.12155] [Medline: 23992038]
80. Rasi P, Vuojärvi H, Rivinen S. Promoting media literacy among older people: a systematic review. Adult Education Quarterly 2020 May 25;71(1):37-54. [doi: 10.1177/0741713620923755]

Abbreviations

AI: artificial intelligence
eTRIP: Eating Trigger-Response Inhibition Program
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs

Edited by A Mavragani; submitted 26.01.23; peer-reviewed by YC Liu, V Jennings; comments to author 07.12.23; revised version received 12.12.23; accepted 12.03.24; published 07.05.24.

Please cite as:
Chew HSJ, Chew NWS, Loong SSE, Lim SL, Tam WSW, Chin YH, Chao AM, Dimitriadis GK, Gao Y, So JBY, Shabbir A, Ngiam KY Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation
J Med Internet Res 2024;26:e46036
URL: https://www.jmir.org/2024/1/e46036
doi:10.2196/46036
PMID:38713909

©Han Shi Jocelyn Chew, Nicholas WS Chew, Shaun Seh Ern Loong, Su Lin Lim, Wai San Wilson Tam, Yip Han Chin, Ariana M Chao, Georgios K Dimitriadis, Yujia Gao, Jimmy Bok Yan So, Asim Shabbir, Kee Yuan Ngiam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characteristic Changes of the Stance-Phase Plantar Pressure Curve When Walking Uphill and Downhill: Cross-Sectional Study

Christian Wolff[1], MSc; Patrick Steinheimer[2], MSc; Elke Warmerdam[3], MSc, PhD; Tim Dahmen[1], MSc, PhD; Philipp Slusallek[1], MSc, PhD; Christian Schlinkmann[1], MSc; Fei Chen[1], MSc; Marcel Orth[2], MD, PhD; Tim Pohlemann[2], MD, PhD; Bergita Ganse[2,3], MD, PhD

[1]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

[2]Department of Trauma, Hand and Reconstructive Surgery, Departments and Institutes of Surgery, Saarland University, Homburg/Saar, Germany

[3]Innovative Implant Development (Fracture Healing), Departments and Institutes of Surgery, Saarland University, Homburg/Saar, Germany

Corresponding Author:
Bergita Ganse, MD, PhD
Innovative Implant Development (Fracture Healing)
Departments and Institutes of Surgery
Saarland University
Kirrberger Straße 1
Building 57
Homburg/Saar, 66421
Germany
Phone: 49 684116 ext 31570
Email: bergita.ganse@uks.eu

Abstract

Background: Monitoring of gait patterns by insoles is popular to study behavior and activity in the daily life of people and throughout the rehabilitation process of patients. Live data analyses may improve personalized prevention and treatment regimens, as well as rehabilitation. The M-shaped plantar pressure curve during the stance phase is mainly defined by the loading and unloading slope, 2 maxima, 1 minimum, as well as the force during defined periods. When monitoring gait continuously, walking uphill or downhill could affect this curve in characteristic ways.

Objective: For walking on a slope, typical changes in the stance phase curve measured by insoles were hypothesized.

Methods: In total, 40 healthy participants of both sexes were fitted with individually calibrated insoles with 16 pressure sensors each and a recording frequency of 100 Hz. Participants walked on a treadmill at 4 km/h for 1 minute in each of the following slopes: −20%, −15%, −10%, −5%, 0%, 5%, 10%, 15%, and 20%. Raw data were exported for analyses. A custom-developed data platform was used for data processing and parameter calculation, including step detection, data transformation, and normalization for time by natural cubic spline interpolation and force (proportion of body weight). To identify the time-axis positions of the desired maxima and minimum among the available extremum candidates in each step, a Gaussian filter was applied (σ=3, kernel size 7). Inconclusive extremum candidates were further processed by screening for time plausibility, maximum or minimum pool filtering, and monotony. Several parameters that describe the curve trajectory were computed for each step. The normal distribution of data was tested by the Kolmogorov-Smirnov and Shapiro-Wilk tests.

Results: Data were normally distributed. An analysis of variance with the gait parameters as dependent and slope as independent variables revealed significant changes related to the slope for the following parameters of the stance phase curve: the mean force during loading and unloading, the 2 maxima and the minimum, as well as the loading and unloading slope (all P<.001). A simultaneous increase in the loading slope, the first maximum and the mean loading force combined with a decrease in the mean unloading force, the second maximum, and the unloading slope is characteristic for downhill walking. The opposite represents uphill walking. The minimum had its peak at horizontal walking and values dropped when walking uphill and downhill alike. It is therefore not a suitable parameter to distinguish between uphill and downhill walking.

Conclusions: While patient-related factors, such as anthropometrics, injury, or disease shape the stance phase curve on a longer-term scale, walking on slopes leads to temporary and characteristic short-term changes in the curve trajectory.

(J Med Internet Res 2024;26:e44948) doi:10.2196/44948

XSL•FO
RenderX

KEYWORDS

podiatry; podiatric medicine; movement analysis; ground reaction forces; wearables; slope; gait analysis; monitoring; gait; rehabilitation; treatment; sensor; injury; postoperative treatment; sensors; personalized medicine; movement; digital health; pedography; baropedography

Introduction

Long-term monitoring of gait patterns and plantar-pressure distributions via insoles are increasingly popular ways to study behavior and activity in the field and in the everyday lives of people and patients, including healing, personalized prevention, and treatment or disease progression [1-3]. In recent years, the usability of instrumented insoles for gait analyses has increased. Several technical issues could be resolved, including calibration, hysteresis and drift, durability, usability, limited energy supply and battery life, data storage capacity, and the restriction to low sample frequencies associated with higher error rates, that is, when force peaks are missed [3-5]. The usability of instrumented insoles is currently still limited by difficulties in data analysis. Advanced algorithms and tools are needed and currently developed to be able to draw meaningful conclusions from such insole gait data [6,7]. When analyzing long-term field data and developing smart health care innovations, automated data annotation is desirable to determine and quantify the activities a person has conducted. Ideally, the activity type can be determined algorithmically from the plantar pressure data alone.

Characteristic gait changes have been reported for walking on slopes, such as changes in the contribution of the ankle joint to leg work [8]. In addition, uphill walking on a treadmill increases hip and knee flexion angles during the stance phase, as well as the forward tilt of the thorax [9]. Furthermore, a decrease in dorsiflexion was observed during downhill walking at initial contact, in midstance, and during the second half of the swing phase [9]. During uphill walking with increasing inclination, more positive joint work was identified for the ankle and hip joint, while negative joint work increased during downhill walking [10]. Older individuals were shown to have a disproportionate recruitment of hip muscles and smaller increases in activity of the medial gastrocnemius muscle with steeper uphill slopes than younger adults, resulting in difficulty walking on steep slopes [11].

The M-shaped curve of ground reaction forces or plantar pressure during the stance phase is mainly defined by the loading and unloading slope, 2 maxima, 1 minimum, as well as the force during defined periods [12]. When monitoring gait continuously via insoles, walking uphill or downhill on a slope could affect the gait cycle curve in characteristic ways. If these typical changes were known, one could correct for such confounders when analyzing insole data. We hypothesized that walking on a slope generates typical changes in the plantar pressure stance phase curve that vary between uphill and downhill walking.

Methods

Study Design

This study is part of the project Smart Implants 2.0—Weight-bearing and Gait Observation for Early Monitoring of Fracture Healing and Individualized Therapy after Trauma, funded by the Werner Siemens Foundation. It was registered in the German Clinical Trials Register (DRKS00025108).

Ethical Considerations

Ethical approval was obtained from the Institutional Review Board of Saarland Medical Board (Ärztekammer des Saarlandes, 30/21).

Data Collection

Inclusion criteria were the ability to walk on a treadmill, and aged 18 years and older. Exclusion criteria were aged under 18 years, use of walking aids, inability to give consent, pregnancy, immobility, and previous injury of the lower legs or pelvis. The aim was to collect data from healthy volunteers.

The healthy participants of both sexes (none of them identified as diverse) were fitted with individually calibrated OpenGO insoles (Moticon GmbH) with 16 pressure sensors in each insole to be used in regular running shoes. Calibration to the individual body weight was conducted using the Moticon OpenGO app by letting the participants walk and shift their body weight in a standardized way. The insole size was selected to fit the individual participant's shoe size. Measurements were conducted with a recording frequency of 100 Hz in the record mode of the device. Raw data were exported for analyses. The participants walked on a treadmill at 4 km/h (Mercury, HP Cosmos) for 1 minute while insole data were collected with 3-minute breaks for recovery. Recordings were obtained for slopes of −20%, −15%, −10%, −5%, 0%, 5%, 10%, 15%, and 20%. The participants were asked to walk for 1 minute straight, and recording was only commenced when the walking was already in progress to avoid bias by including altered steps upon gait initiation.

Data Processing

The pressure readings of the force sensors in the insole device yield a weighted sum as a total vertical ground reaction force reading. To compute the force, every summand is weighted by its sensor area and a respective scaling factor accounting for the sensor's surrounding area, as well as gaps between sensors that depend on the insole size. This process is conducted by the Moticon software as an automated processing step before file export. Insole data were exported as described previously [13,14]. A custom-developed data platform was then used for further processing and parameter calculation, in which step detection was conducted as follows. The stance phases were identified and extracted from the time series data by considering any activity with consecutive force readings above 30 N. A tolerance of up to 3 missing values was implemented to account for possible recording issues. Any activity with a duration of less than 300 milliseconds or more than 2000 milliseconds was discarded. Both the force and time axes were normalized. Force readings were transformed from Newton to a proportion of the body weight of the respective participant. Of note, as plantar

XSL•FO
RenderX

pressure was measured instead of weight, due to acceleration, values regularly exceeded the body weight for peak load-bearing instances. Normalizing the time axis was more complex, as the lack of a fixed cadence resulted in varying step lengths and thus differing numbers of true measurement points for each step. Therefore, a natural cubic spline interpolation was conducted on the original raw data. Based on the resulting curve for each stance phase, 100 equidistant samples were taken, resulting in an interpolated force measurement point for every 1% of the overall stance phase length. This approach accounted for the lower recording frequency and higher sensor noise inherent to the insoles when compared with other gait measuring devices, such as sensor-equipped treadmills or force plates. Parameters that describe the trajectory of the stance phase curve are usually based on or derived from the characteristic local extrema, that is, the first and second force peak and the local minimum in-between force peaks. These maxima and the minimum are used as parameters themselves to describe the curve trajectory [13]. Sensor jitter may lead to the existence of multiple ambiguous candidates for the named extrema. As a solution to this, a Gaussian filter was applied to the original raw data in a repetition of the normalization process. The applied filtering strategy (σ=3, kernel size 7) was chosen to prioritize the elimination of extrema ambiguity at the expense of signal precision. This can result in overcorrection in areas with higher signal volatility, mostly at the start and end of the stance phase. Hence, to avoid loss of high-frequency detail, the filtered and normalized curve was not used for parameter analysis, but only to determine unambiguous time-axis positions (indices) for the extremum candidates. These indices were then reapplied to the nonfiltered, normalized data to identify the corresponding plantar pressure measurement closest to the original raw data. In case the use of the filtered data still led to inconclusive extremum candidates, the following additional detection strategies were applied in the named order: (1) time plausibility: extremum candidates occurring within the first or last 10 indices (first/last 10% of overall time span) were eliminated; (2) maximum or minimum-pool filtering: should multiple extremum candidates occur within a pool size of 5 indices (equals to 5% of overall time span), the candidate with the highest or lowest force value was chosen; (3) monotony-check: in case of multiple remaining extremum candidates, candidates where the curve did not display a strict monotonous decrease or increase in both directions within 5 indices each were eliminated; and (4) monotony grace: in case the monotony check had eliminated too many candidates (less than 2 maximum candidates or less than 1 minimum candidate remaining), the eliminated candidates were reinstated in descending order of their highest achieved monotony distance until the target number of candidates was reached.

After applying these strategies, every stance activity that remained with an irregular amount of unambiguous extremum candidates was removed from the data set. In total, 585 load-bearing events were excluded as not fitting the strict parameter definitions.

Parameters

For each participant, across the minute of walking all stance phase curves were extracted. The parameters illustrated in Figure 1 were calculated for each stance phase and used to analyze changes in the trajectory of the stance phase curve. To do so, data from both feet were pooled. The curve is mainly described by 2 maxima and a minimum in between the maxima, Fz2 (the first maximum), Fz3 (the minimum), and Fz4 (the second maximum). The mean force over the entire stance phase is referred to as $Fmean_{stance}$. The mean force between the start of the loading phase and Fz2 is $Fmean_{load}$. The mean force between Fz2 and Fz4 is $Fmean_{mid}$. The mean force between Fz4 and the end of the unloading phase is $Fmean_{unload}$. All these parameters have the unit percent body weight. In addition, the loading and unloading slope have the units percent body weight or percent stance phase duration. The loading slope was computed as the slope of the line defined by the start of the loading phase and the first force reading equal to or higher than 80% of Fz2. The unloading slope was calculated as the slope of the line defined by the first force reading in the unloading phase below 80% of Fz4 and the end of the stance phase event.

XSL•FO
RenderX

Figure 1. Depiction of the analyzed parameters of the stance phase. %BW: percent body weight; $Fmean_{load}$: the mean force between the start of loading-phase and Fz2; $Fmean_{mid}$: the mean force between Fz2 and Fz4; $Fmean_{stance}$: the mean force over the entire stance phase; $Fmean_{unload}$: the mean force between Fz4 and the end of the unloading-phase; Fz2: the first maximum; Fz3: the minimum; Fz4: the second maximum.

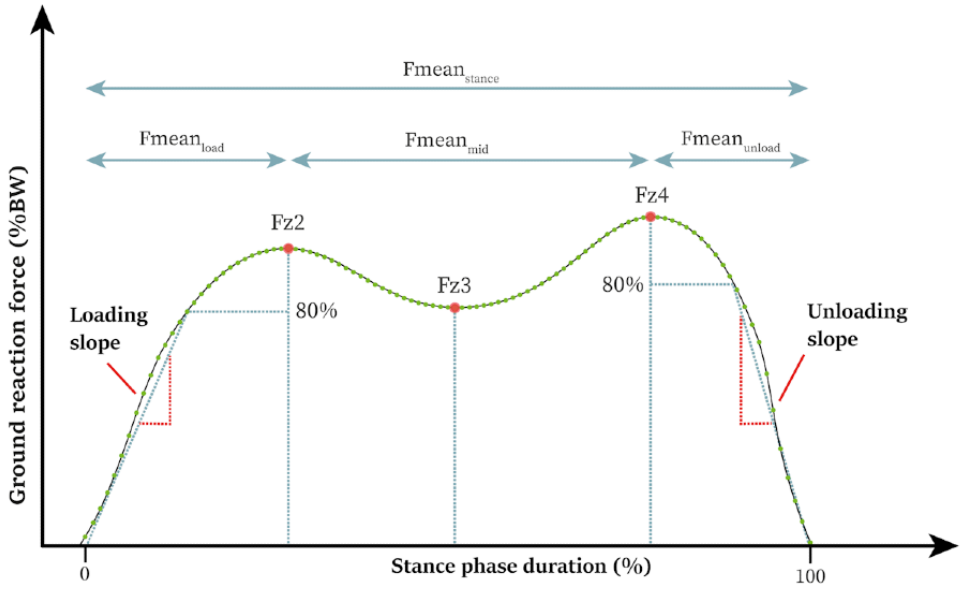

Statistical Analyses

Statistical tests were executed with SPSS Statistics (version 29; IBM Corp). Significance was defined as $P<.05$. The normal distribution of data was tested by the Kolmogorov-Smirnov and Shapiro-Wilk tests. A linear regression analysis of variance was conducted for each of the gait parameters as the dependent variable, with the slope (−20% to 20%) as the independent variable. Mean values and SD are reported. Linear regression slopes are reported for comparability and to allow for correction, even though for some of the parameters other but differing regression types yielded higher R^2 values. The sample size of 40 was an estimate based on what is common in the field, and taking into account the aim to measure a very diverse group of volunteers. An a priori sample size calculation was not conducted due to a lack of comparable data.

Results

Measurements were taken from 40 healthy participants (19 women and 21 men) with an average age of 43.90 (SD 17.30, range 18-87) years. Participant characteristics are summarized in Table 1. Data were successfully recorded for all of the participants and slope levels, resulting in a complete data set (Multimedia Appendix 1).

Table 1. Participant characteristics.

	Total (N=40)	Women (n=19)	Men (n=21)
Age (years), mean (SD)	43.90 (17.30)	39.05 (14.65)	48.29 (18.64)
Height (cm), mean (SD)	174.43 (11.24)	165.79 (6.05)	182.24 (8.85)
Weight (kg), mean (SD)	80.40 (26.85)	66.22 (16.15)	93.24 (28.40)
BMI (kg/m^2), mean (SD)	23.04 (6.83)	20.15 (5.06)	25.65 (7.28)

Data were normally distributed. Figure 2 visualizes the differences between the analyzed slope values on the stance phase curve. Figure 3 shows the normalized changes in the analyzed parameters with the slope of the treadmill. The analysis of variance revealed significant changes with the slope for $Fmean_{load}$, $Fmean_{unload}$, Fz2, Fz3, Fz4, loading and unloading slope (all $P<.001$). There was no significant correlation of the slope with $Fmean_{stance}$ ($P=.98$) and $Fmean_{mid}$ ($P=.13$). Other than the other parameters with significant changes related to slope, Fz3 had its peak at horizontal walking and values dropped when walking uphill and downhill alike. Thus, a simultaneous and short-term increase in loading slope and $Fmean_{load}$ combined with a decrease in $Fmean_{unload}$, Fz2, Fz4, and the unloading slope indicates downhill walking, while the opposite indicates uphill walking. Fz3 is not a suitable parameter to distinguish between uphill and downhill walking, as its value decreases both when walking uphill as well as downhill. Mean values and the SD of the analyzed parameters for each treadmill slope level in absolute values are displayed in Table 2. Table 3 indicates the linear regression slopes and R^2-values for each of the curves shown in Figure 3.

XSL•FO
RenderX

Figure 2. (A) The mean trajectories of the stance phase curve for each of the analyzed slope levels. (B) The mean trajectories and the 95% CI for −20%, 0%, and 20%. %BW: percent body weight.

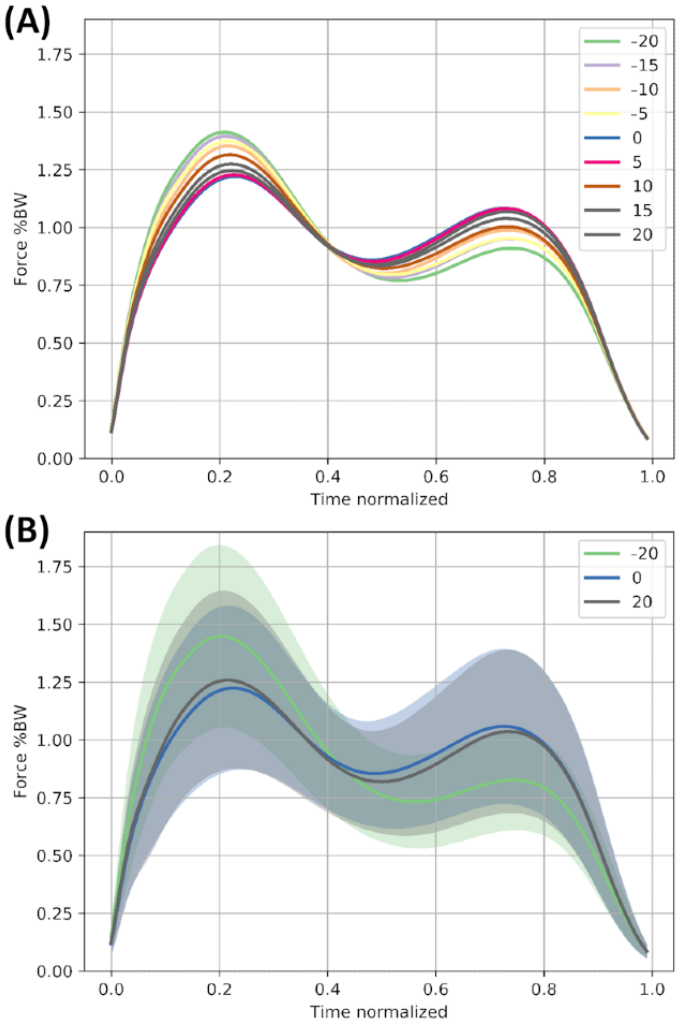

XSL•FO
RenderX

Figure 3. For each slope level, the mean value of each parameter is shown in the percent of horizontal walking (all values averaged over all participants). $Fmean_{load}$: the mean force between the start of loading-phase and Fz2; $Fmean_{mid}$: the mean force between Fz2 and Fz4; $Fmean_{stance}$: the mean force over the entire stance phase; $Fmean_{unload}$: the mean force between Fz4 and the end of the unloading-phase; Fz2: the first maximum; Fz3: the minimum; Fz4: the second maximum.

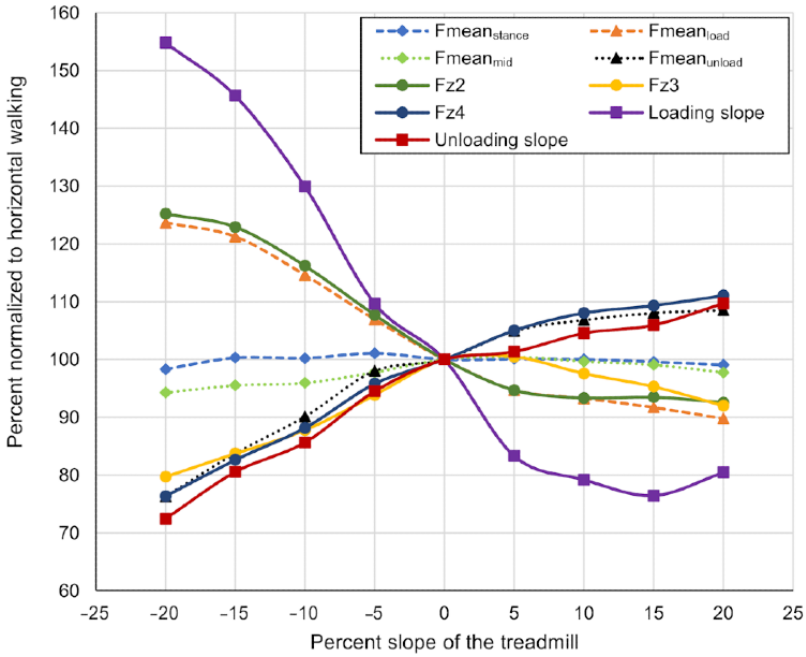

Table 2. Mean values and SD of the analyzed parameters for each slope level (absolute values).

	−20%	−15%	−10%	−5%	0%	5%	10%	15%	20%
$Fmean_{stance}$[a] (% body weight)	0.88 (0.21)	0.90 (0.21)	0.90 (0.21)	0.91 (0.21)	0.90 (0.21)	0.90 (0.21)	0.90 (0.21)	0.89 (0.20)	0.89 (0.20)
$Fmean_{load}$[b] (% body weight)	1.06 (0.30)	1.04 (0.28)	0.98 (0.26)	0.91 (0.23)	0.86 (0.21)	0.81 (0.20)	0.80 (0.20)	0.78 (0.19)	0.77 (0.18)
$Fmean_{mid}$[c] (% body weight)	0.97 (0.24)	0.99 (0.23)	0.99 (0.23)	1.01 (0.23)	1.03 (0.25)	1.04 (0.25)	1.03 (0.25)	1.02 (0.25)	1.01 (0.24)
$Fmean_{unload}$[d] (% body weight)	0.55 (0.14)	0.61 (0.14)	0.65 (0.16)	0.71 (0.16)	0.72 (0.17)	0.76 (0.18)	0.77 (0.18)	0.78 (0.18)	0.79 (0.18)
Fz2[e] (% body weight)	1.50 (0.39)	1.48 (0.38)	1.39 (0.36)	1.29 (0.33)	1.20 (0.30)	1.14 (0.29)	1.12 (0.29)	1.12 (0.29)	1.11 (0.29)
Fz3[f] (% body weight)	0.70 (0.20)	0.74 (0.19)	0.77 (0.18)	0.83 (0.19)	0.88 (0.22)	0.88 (0.21)	0.86 (0.21)	0.84 (0.22)	0.81 (0.19)
Fz4[g] (% body weight)	0.88 (0.24)	0.96 (0.24)	1.02 (0.26)	1.11 (0.28)	1.16 (0.32)	1.22 (0.33)	1.25 (0.33)	1.26 (0.34)	1.29 (0.35)
Loading slope (% body weight/% stance phase duration)	11.00 (4.58)	10.35 (4.01)	9.23 (3.58)	7.79 (2.99)	7.10 (2.91)	5.92 (2.11)	5.63 (2.07)	5.43 (1.85)	5.71 (2.07)
Unloading slope (% body weight/% stance phase duration)	−5.30 (1.66)	−5.89 (1.87)	−6.26 (2.10)	−6.91 (2.34)	−7.31 (2.71)	−7.41 (2.64)	−7.65 (2.79)	−7.75 (2.92)	−8.02 (3.02)

[a]$Fmean_{stance}$: the mean force over the entire stance phase.

[b]$Fmean_{load}$: the mean force between the start of the loading phase and Fz2.

[c]$Fmean_{mid}$: the mean force between Fz2 and Fz4.

[d]$Fmean_{unload}$: the mean force between Fz4 and the end of the unloading phase.

[e]Fz2: the first maximum.

[f]Fz3: the minimum.

[g]Fz4: the second maximum.

XSL•FO
RenderX

Table 3. Linear regression slopes and R2 values for each of the curves shown in Figure 3.

	Linear regression slope (%)	R^2
$Fmean_{stance}$[a] (% body weight)	−0.002	0.001
$Fmean_{load}$[b] (% body weight)	0.930	0.943
$Fmean_{mid}$[c] (% body weight)	0.116	0.527
$Fmean_{unload}$[d] (% body weight)	0.807	0.909
Fz2[e] (% body weight)	−0.926	0.907
Fz3[f] (% body weight)	0.367	0.483
Fz4[g] (% body weight)	0.894	0.952
Loading slope (% body weight/% stance phase duration)	−2.109	0.910
Unloading slope (% body weight/% stance phase duration)	0.900	0.935

[a]$Fmean_{stance}$: the mean force over the entire stance phase.

[b]$Fmean_{load}$: the mean force between the start of the loading phase and Fz2.

[c]$Fmean_{mid}$: the mean force between Fz2 and Fz4.

[d]$Fmean_{unload}$: the mean force between Fz4 and the end of the unloading phase.

[e]Fz2: the first maximum.

[f]Fz3: the minimum.

[g]Fz4: the second maximum.

Discussion

Principal Results

This study identified characteristic changes when walking with an uphill or downhill slope in insole plantar pressure data of healthy participants. The most pronounced changes with treadmill slope were found in the loading slope of the curve. A typical combination of changes in several parameters was reported that defines uphill and downhill walking and may be used for annotation and correction when analyzing such data. These changes in the trajectory of the force curve with different surface slopes relative to the force vector of Earth's gravity are related to changes in plantar load distribution. When walking downhill, Fz2 was found to be higher compared to when walking uphill, which is caused by the more pronounced force transfer through the heel of the foot, followed by a lower second maximum due to the even lower surface at push-off.

While patient-related factors, such as curve characteristics related to body size, muscle power, degenerative disease, etc, would remain constant throughout an insole measurement, fatigue-related changes [15] may increasingly appear and then stay toward the later stages of a recording of a walking bout. Additionally, age, body height, body weight, BMI, and handgrip strength were shown to cause characteristic changes in the plantar pressure force curve, that would usually only change on a long-term scale [16]. In contrast, as shown in the present data set, walking on slopes leads to temporary and characteristic changes in specific properties of the stance-phase curve. Changes over time in the identified parameters should thus be considered and correctly interpreted when studying long-term field gait data collected via insoles. To analyze the healing process, that is, after an injury, slow changes in parameters would be expected, and a trend toward what is considered normal over several weeks [17]. Short-term changes over minutes or hours would thus not be explainable by the healing progress and should have a different cause. In addition, the asymmetry between the legs should slowly decrease throughout healing [18]. When walking on a slope, asymmetry could also be affected, if the injury causes increasing problems such as pain when walking uphill or downhill. It is also recommendable to identify the characteristics of walking with walking aids, such as crutches, to be able to classify the nature of the observed changes and the treatment stage better.

Limitations

Effects of walking speed were not analyzed in this study, even though it is known that lower extremity joint loading is affected by varying step length and cadence during graded uphill and downhill walking [19]. These parameters, however, do not seem to be necessary to successfully annotate gait data obtained by insoles. For participant or patient convenience, it would be desirable if insoles did not need to be combined with further devices or wearables. The present data suggest that at least the identification of walking on slopes does not require further sensors. It is also known that kinematic, kinetic, and electromyographic parameters differ between treadmill walking and overground gait, while spatiotemporal, kinematic, kinetic, electromyographic, and energy consumption outcome measures are largely comparable [20]. Another limitation of this study is that the parameters analyzed here can only be used when a regular gait curve is present. If this is not the case, other methods need to be applied, that is, machine learning for step detection and segmentation or the analysis of further parameters, possibly slopes and averages, or differences between individual sensors [21]. Differences between the 16 sensors embedded in each

XSL•FO
RenderX

insole were not analyzed in this study and could be assessed in the future, for example, to distinguish between ground types (gravel, sand, etc). Another limitation is that the present characteristic changes that were assessed in healthy participants may differ for patients with gait disorders, depending on their disease or injury type. It will therefore be important to collect longitudinal data on different slopes from patients with defined diseases and injuries throughout the healing process or throughout different disease stages. These studies would serve to identify if the reported findings are valid also for patients, and for which patient groups this is true.

Use of Wearables in Patients

The insole technology and present data may be valuable in real-world settings when investigating changes in mechanical properties during walking, that is, in occupational health research, sport and exercise science, for urban planning, and to plan inclusive architecture. For instance, the global average slope of urban areas is about 3.70° [22]. Wearables such as pressure insoles are increasingly used to study gait and movement, as well as for fall detection, fall classification, and fall risk assessment in the daily life of patients, and furthermore for lifestyle and health monitoring [1,3,23-27]. Long-term monitoring, especially if combined with additional sensors, may produce large amounts of data that require advanced strategies for analyses. Apart from regression statistics, among the options is the use of machine learning algorithms trained with annotated data for pattern recognition [24,26]. For longer-term monitoring of patients, it would be desirable if such algorithms were trained to identify various key activities of daily life that might indicate the level of healing progress. For example, when a patient with a tibial fracture is capable of cycling again, this is likely an indication for advances in the healing process. It would also be of interest to identify risky behavior, possibly leading to excessive forces, and to warn the patient by giving, for example, an audible or haptic warning signal. To guarantee meaningful data interpretation, machine learning may be combined with conventional regression-based analyses, such as the ones proposed in this paper to best tackle data complexity. Furthermore, prediction algorithms could be implemented for falls and diseases that enable more refined individual recommendations. Ideally, such interventions would be based on live data analyses. Limitations in the computing power of small wearable devices can increasingly be mitigated by both algorithmic optimization techniques in machine learning, such as dimensionality reduction, reservoir computing, and network pruning, as well as hardware innovations [27,28]. In the near future, such advances will likely allow real-time feedback based on data from various sources combined [29,30]. Alternatively, extracting decision-making systems (symbolic artificial intelligence), such as threshold-based methods, might offer an immediate route to real-time feedback.

Sensors in Orthoses and Implants

Apart from insoles, very similar data might be collected from mechanical sensors embedded in orthoses [31] or implants [32]. Potentially, walking on a slope in these recordings changes the data in similar ways as described here. It would be highly desirable if patients did not need to use separate wearables such as insoles anymore, but if orthoses and implants had sensors embedded, not only to monitor healing progress but also to identify healing problems or complications and the need for surgical revision [33]. If similar load data could be collected by sensors in artificial hip or knee joints, or potentially even by plates or nails that stabilize bone fractures, recovery regimen could be monitored continuously and advice given on time [34]. Alarms could go off if forces exceeded certain thresholds or if live pattern analyses revealed unfavorable patterns known to be associated with exceeding forces or problems. As these developments seem to have a high potential with regard to rehabilitation and postoperative treatment, data analyses of insole data should be further studied and ideally, details such as algorithms and characteristics should be published to enable for the further development and widespread application of the named interventions.

Conclusions

Characteristic changes in the plantar-pressure stance phase curve were identified, which reflect uphill and downhill walking. Automated annotation and continuous analyses of gait data via wearables could enable improved rehabilitation and feedback systems for prevention and treatment. A combination of traditional regression statistics embedded in heuristics combined with artificial intelligence methods may yield the best results.

Acknowledgments

The Werner Siemens Foundation (project Smart Implants 2.0) funded this work. The authors would like to acknowledge the help of Aynur Gökten and Jacqueline Orth during the measurements, as well as the help of Lisa-Marie Jost in designing Figure 1.

Authors' Contributions

CW contributed to the data processing platform, data analysis, methods, and Figure 2. P Steinheimer conducted the measurements. BG contributed to the idea; ran the statistical analyses; interpreted the data; made the tables; and drafted, submitted, and revised this paper. TD, CS, and FC took part in the data platform implementation. EW, TD, P Slusallek, CS, FC, MO, and TP helped with data interpretation. All authors have contributed to this paper's drafting and revision, and read and approved the submitted version of this paper.

Conflicts of Interest

TP is President and Board Member of the AO-Foundation, Switzerland, and Extended Board Member of the German Society of Orthopedic Trauma Surgery (DGU), the German Society of Orthopedic Surgery and Traumatology (DGOU), and the German

XSL•FO
RenderX

Society of Surgery (DGCH). TP is also the speaker of the Medical Advisory Board of the German Ministry of Defence. The other authors do not have a conflict of interest.

Multimedia Appendix 1
Data file.
[XLSX File (Microsoft Excel File), 72 KB - jmir_v26i1e44948_app1.xlsx]

References

1. Braun BJ, Veith NT, Rollmann M, Orth M, Fritz T, Herath SC, et al. Weight-bearing recommendations after operative fracture treatment-fact or fiction? Gait results with and feasibility of a dynamic, continuous pedobarography insole. Int Orthop 2017;41(8):1507-1512. [doi: 10.1007/s00264-017-3481-7] [Medline: 28421239]
2. Ramirez-Bautista JA, Huerta-Ruelas JA, Chaparro-Cardenas SL, Hernandez-Zavala A. A review in detection and monitoring gait disorders using in-shoe plantar measurement systems. IEEE Rev Biomed Eng 2017;10:299-309. [doi: 10.1109/RBME.2017.2747402] [Medline: 28866600]
3. Subramaniam S, Majumder S, Faisal AI, Deen MJ. Insole-based systems for health monitoring: current solutions and research challenges. Sensors (Basel) 2022;22(2):438 [FREE Full text] [doi: 10.3390/s22020438] [Medline: 35062398]
4. Elstub LJ, Grohowski LM, Wolf DN, Owen MK, Noehren B, Zelik KE. Effect of pressure insole sampling frequency on insole-measured peak force accuracy during running. J Biomech 2022;145:111387. [doi: 10.1016/j.jbiomech.2022.111387] [Medline: 36442432]
5. North K, Kubiak EN, Hitchcock RW. Sensor packaging design for continuous underfoot load monitoring. Biomed Microdevices 2012;14(1):217-224. [doi: 10.1007/s10544-011-9599-2] [Medline: 22009581]
6. Anderson W, Choffin Z, Jeong N, Callihan M, Jeong S, Sazonov E. Empirical study on human movement classification using insole footwear sensor system and machine learning. Sensors (Basel) 2022;22(7):2743 [FREE Full text] [doi: 10.3390/s22072743] [Medline: 35408358]
7. Chatzaki C, Skaramagkas V, Kefalopoulou Z, Tachos N, Kostikis N, Kanellos F, et al. Can gait features help in differentiating Parkinson's disease medication states and severity levels? A machine learning approach. Sensors (Basel) 2022;22(24):9937 [FREE Full text] [doi: 10.3390/s22249937] [Medline: 36560313]
8. Montgomery JR, Grabowski AM. The contributions of ankle, knee and hip joint work to individual leg work change during uphill and downhill walking over a range of speeds. R Soc Open Sci 2018;5(8):180550 [FREE Full text] [doi: 10.1098/rsos.180550] [Medline: 30225047]
9. Strutzenberger G, Leutgeb L, Claußen L, Schwameder H. Gait on slopes: differences in temporo-spatial, kinematic and kinetic gait parameters between walking on a ramp and on a treadmill. Gait Posture 2022;91:73-78. [doi: 10.1016/j.gaitpost.2021.09.196] [Medline: 34653877]
10. Alexander N, Strutzenberger G, Ameshofer LM, Schwameder H. Lower limb joint work and joint work contribution during downhill and uphill walking at different inclinations. J Biomech 2017;61:75-80. [doi: 10.1016/j.jbiomech.2017.07.001] [Medline: 28734544]
11. Franz JR, Kram R. How does age affect leg muscle activity/coactivity during uphill and downhill walking? Gait Posture 2013;37(3):378-384 [FREE Full text] [doi: 10.1016/j.gaitpost.2012.08.004] [Medline: 22940542]
12. Larsen AH, Puggaard L, Hämäläinen U, Aagaard P. Comparison of ground reaction forces and antagonist muscle coactivation during stair walking with ageing. J Electromyogr Kinesiol 2008;18(4):568-580. [doi: 10.1016/j.jelekin.2006.12.008] [Medline: 17363275]
13. Braun BJ, Veith NT, Hell R, Döbele S, Roland M, Rollmann M, et al. Validation and reliability testing of a new, fully integrated gait analysis insole. J Foot Ankle Res 2015;8:54 [FREE Full text] [doi: 10.1186/s13047-015-0111-8] [Medline: 26396594]
14. Stöggl T, Martiner A. Validation of Moticon's OpenGo sensor insoles during gait, jumps, balance and cross-country skiing specific imitation movements. J Sports Sci 2017;35(2):196-206 [FREE Full text] [doi: 10.1080/02640414.2016.1161205] [Medline: 27010531]
15. Chardon M, Barbieri FA, Penedo T, Santos PCR, Vuillerme N. The effects of experimentally-induced fatigue on gait parameters during obstacle crossing: a systematic review. Neurosci Biobehav Rev 2022;142:104854. [doi: 10.1016/j.neubiorev.2022.104854] [Medline: 36084846]
16. Wolff C, Steinheimer P, Warmerdam E, Dahmen T, Slusallek P, Schlinkmann C, et al. Effects of age, body height, body weight, body mass index and handgrip strength on the trajectory of the plantar pressure stance-phase curve of the gait cycle. Front Bioeng Biotechnol 2023;11:1110099 [FREE Full text] [doi: 10.3389/fbioe.2023.1110099] [Medline: 36873371]
17. Agres AN, Alves SA, Höntzsch D, El Attal R, Pohlemann T, Schaser KD, et al. Improved weight bearing during gait at 6 weeks post-surgery with an angle stable locking system after distal tibial fracture. Gait Posture 2024;107:169-176. [doi: 10.1016/j.gaitpost.2023.09.013] [Medline: 37845132]

XSL•FO
RenderX

18. Rosenbaum D, Macri F, Lupselo FS, Preis OC. Gait and function as tools for the assessment of fracture repair—the role of movement analysis for the assessment of fracture healing. Injury 2014;45(Suppl 2):S39-S43. [doi: 10.1016/j.injury.2014.04.007] [Medline: 24857027]
19. Schwameder H, Lindenhofer E, Müller E. Effect of walking speed on lower extremity joint loading in graded ramp walking. Sports Biomech 2005;4(2):227-243. [doi: 10.1080/14763140508522865] [Medline: 16138659]
20. Semaan MB, Wallard L, Ruiz V, Gillet C, Leteneur S, Simoneau-Buessinger E. Is treadmill walking biomechanically comparable to overground walking? A systematic review. Gait Posture 2022;92:249-257. [doi: 10.1016/j.gaitpost.2021.11.009] [Medline: 34890914]
21. Blades S, Marriott H, Hundza S, Honert EC, Stellingwerff T, Klimstra M. Evaluation of different pressure-based foot contact event detection algorithms across different slopes and speeds. Sensors (Basel) 2023;23(5):2736 [FREE Full text] [doi: 10.3390/s23052736] [Medline: 36904942]
22. Shi K, Liu G, Zhou L, Cui Y, Liu S, Wu Y. Satellite remote sensing data reveal increased slope climbing of urban land expansion worldwide. Landsc Urban Plan 2023;235:104755. [doi: 10.1016/j.landurbplan.2023.104755]
23. Cates B, Sim T, Heo HM, Kim B, Kim H, Mun JH. A novel detection model and its optimal features to classify falls from low- and high-acceleration activities of daily life using an insole sensor system. Sensors (Basel) 2018;18(4):1227 [FREE Full text] [doi: 10.3390/s18041227] [Medline: 29673165]
24. Kraus M, Saller MM, Baumbach SF, Neuerburg C, Stumpf UC, Böcker W, et al. Prediction of physical frailty in orthogeriatric patients using sensor insole-based gait analysis and machine learning algorithms: cross-sectional study. JMIR Med Inform 2022;10(1):e32724 [FREE Full text] [doi: 10.2196/32724] [Medline: 34989684]
25. Subramaniam S, Faisal AI, Deen MJ. Wearable sensor systems for fall risk assessment: a review. Front Digit Health 2022;4:921506 [FREE Full text] [doi: 10.3389/fdgth.2022.921506] [Medline: 35911615]
26. Harris EJ, Khoo IH, Demircan E. A survey of human gait-based artificial intelligence applications. Front Robot AI 2022;8:749274 [FREE Full text] [doi: 10.3389/frobt.2021.749274] [Medline: 35047564]
27. Hou CKJ, Behdinan K. Dimensionality reduction in surrogate modeling: a review of combined methods. Data Sci Eng 2022;7(4):402-427 [FREE Full text] [doi: 10.1007/s41019-022-00193-5] [Medline: 36345394]
28. Tanaka G, Yamane T, Héroux JB, Nakane R, Kanazawa N, Takeda S, et al. Recent advances in physical reservoir computing: a review. Neural Netw 2019;115:100-123 [FREE Full text] [doi: 10.1016/j.neunet.2019.03.005] [Medline: 30981085]
29. Chiasson-Poirier L, Younesian H, Turcot K, Sylvestre J. Detecting gait events from accelerations using reservoir computing. Sensors (Basel) 2022;22(19):7180 [FREE Full text] [doi: 10.3390/s22197180] [Medline: 36236278]
30. Zhang Q, Jin T, Cai J, Xu L, He T, Wang T, et al. Wearable triboelectric sensors enabled gait analysis and waist motion capture for IoT-Based smart healthcare applications. Adv Sci (Weinh) 2022;9(4):e2103694 [FREE Full text] [doi: 10.1002/advs.202103694] [Medline: 34796695]
31. Moreira L, Figueiredo J, Cerqueira J, Santos CP. A review on locomotion mode recognition and prediction when using active orthoses and exoskeletons. Sensors (Basel) 2022;22(19):7109 [FREE Full text] [doi: 10.3390/s22197109] [Medline: 36236204]
32. Alves SA, Polzehl J, Brisson NM, Bender A, Agres AN, Damm P, et al. Ground reaction forces and external hip joint moments predict in vivo hip contact forces during gait. J Biomech 2022;135:111037 [FREE Full text] [doi: 10.1016/j.jbiomech.2022.111037] [Medline: 35313250]
33. Warmerdam E, Orth M, Pohlemann T, Ganse B. Gait analysis to monitor fracture healing of the lower leg. Bioengineering (Basel) 2023;10(2):255 [FREE Full text] [doi: 10.3390/bioengineering10020255] [Medline: 36829749]
34. Ganse B, Orth M, Roland M, Diebels S, Motzki P, Seelecke S, et al. Concepts and clinical aspects of active implants for the treatment of bone fractures. Acta Biomater 2022;146:1-9 [FREE Full text] [doi: 10.1016/j.actbio.2022.05.001] [Medline: 35537678]

Abbreviations

$Fmean_{load}$: the mean force between the start of the loading phase and Fz2
$Fmean_{mid}$: the mean force between Fz2 and Fz4
$Fmean_{stance}$: the mean force over the entire stance phase
$Fmean_{unload}$: the mean force between Fz4 and the end of the unloading phase
Fz2: the first maximum
Fz3: the minimum
Fz4: the second maximum

XSL•FO
RenderX

Edited by G Eysenbach, T Leung; submitted 24.01.23; peer-reviewed by M Kraus, S Okita; comments to author 21.12.23; revised version received 11.01.24; accepted 17.02.24; published 08.05.24.

Please cite as:
Wolff C, Steinheimer P, Warmerdam E, Dahmen T, Slusallek P, Schlinkmann C, Chen F, Orth M, Pohlemann T, Ganse B
Characteristic Changes of the Stance-Phase Plantar Pressure Curve When Walking Uphill and Downhill: Cross-Sectional Study
J Med Internet Res 2024;26:e44948
URL: https://www.jmir.org/2024/1/e44948
doi:10.2196/44948
PMID:38718385

©Christian Wolff, Patrick Steinheimer, Elke Warmerdam, Tim Dahmen, Philipp Slusallek, Christian Schlinkmann, Fei Chen, Marcel Orth, Tim Pohlemann, Bergita Ganse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Patients' Perspectives on the Data Confidentiality, Privacy, and Security of mHealth Apps: Systematic Review

Nasser Alhammad[1,2], MSc; Mohannad Alajlani[1], PhD; Alaa Abd-alrazaq[3], PhD; Gregory Epiphaniou[1], PhD; Theodoros Arvanitis[4], PhD

[1]Institute of Digital Healthcare, WMG, University of Warwick, Coventry, United Kingdom
[2]Health Informatics, Saudi Electronic University, Jeddah, Saudi Arabia
[3]AI Center for Precision Health, Weill Cornell Medicine, Doha, Qatar
[4]School of Engineering, University of Birmingham, Birmingham, United Kingdom

Corresponding Author:
Nasser Alhammad, MSc
Institute of Digital Healthcare, WMG
University of Warwick
Millburn House
Coventry, CV47AL
United Kingdom
Phone: 66 558885007
Email: N.alhammad@seu.edu.sa

Abstract

Background: Mobile health (mHealth) apps have the potential to enhance health care service delivery. However, concerns regarding patients' confidentiality, privacy, and security consistently affect the adoption of mHealth apps. Despite this, no review has comprehensively summarized the findings of studies on this subject matter.

Objective: This systematic review aims to investigate patients' perspectives and awareness of the confidentiality, privacy, and security of the data collected through mHealth apps.

Methods: Using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a comprehensive literature search was conducted in 3 electronic databases: PubMed, Ovid, and ScienceDirect. All the retrieved articles were screened according to specific inclusion criteria to select relevant articles published between 2014 and 2022.

Results: A total of 33 articles exploring mHealth patients' perspectives and awareness of data privacy, security, and confidentiality issues and the associated factors were included in this systematic review. Thematic analyses of the retrieved data led to the synthesis of 4 themes: concerns about data privacy, confidentiality, and security; awareness; facilitators and enablers; and associated factors. Patients showed discordant and concordant perspectives regarding data privacy, security, and confidentiality, as well as suggesting approaches to improve the use of mHealth apps (facilitators), such as protection of personal data, ensuring that health status or medical conditions are not mentioned, brief training or education on data security, and assuring data confidentiality and privacy. Similarly, awareness of the subject matter differed across the studies, suggesting the need to improve patients' awareness of data security and privacy. Older patients, those with a history of experiencing data breaches, and those belonging to the higher-income class were more likely to raise concerns about the data security and privacy of mHealth apps. These concerns were not frequent among patients with higher satisfaction levels and those who perceived the data type to be less sensitive.

Conclusions: Patients expressed diverse views on mHealth apps' privacy, security, and confidentiality, with some of the issues raised affecting technology use. These findings may assist mHealth app developers and other stakeholders in improving patients' awareness and adjusting current privacy and security features in mHealth apps to enhance their adoption and use.

Trial Registration: PROSPERO CRD42023456658; https://tinyurl.com/ytnjtmca

(J Med Internet Res 2024;26:e50715) doi:10.2196/50715

KEYWORDS

mobile health apps; mHealth apps; mobile health; privacy; confidentiality; security; awareness; perspectives; mobile phone

XSL•FO
RenderX

Introduction

Background

In recent years, the use of mobile health (mHealth) apps by both the public and health care professionals (HCPs) has significantly increased with the introduction of smartphones [1] and growing interest in the health care industry and research field [2]. The COVID-19 pandemic has further accelerated reliance on digital health [3]. mHealth apps are used by patients to manage diseases, self-monitor, gather health information, supervise behavior changes, manage fitness, and remind them of their medication and rehabilitation schedules [4]. From HCPs' point of view, mHealth apps help manage health records, provide easy access to health records, and provide a path to conduct mobile consultations and remote monitoring during and after treatment [5]. In addition, mHealth apps provide easy access to HCPs by connecting them to clinical information system resources such as electronic health records [6]. Although mHealth apps could provide evidence-based and cost-effective health data and 2-way communication between patients and their HCPs in a remote setting, a few barriers have blocked the expansion of mHealth apps in the health care industry. Data confidentiality, privacy, and security and the regulatory supervision of the apps are some known barriers that hinder mHealth adoption in the health care field.

Despite various benefits of mHealth apps, data confidentiality, privacy, and security issues have caused patients or the public to display less interest and low confidence in mHealth app practice [7]. It could be due to the uncertainty about the information gathered or kept in mHealth apps, the function of the stored data, and who can view or access the data [4]. The term "confidentiality" is defined as the responsibility of those who obtain data (app providers) to uphold the concerns of those to whom the information is related (consumers) [8]. The study by Bhuyan et al [9] mentioned that the National Committee on Vital and Health Statistics defined privacy as an individual legal right or freedom to protect or disclose their health information, and security is defined as personal, mechanical, or authority protection tools used to guard health information against unwanted people or access, whereas privacy is defined as the physical, mechanical, or legislative mechanism or tools to shield personal health information from unauthorized disclosure [10].

Confidentiality, privacy, and security act as challenges in boosting mHealth app adoption. Patients' perceptions of these issues may influence their adoption of mHealth apps, but such events are context dependent. While users are more likely to raise concerns about privacy or confidentiality issues when probed about mHealth apps, such perceptions may not ultimately influence their behavior regarding the actual use or adoption of such apps. Thus, it is pertinent to explore whether privacy concerns are prioritized by users when they engage with mHealth apps and whether such concerns affect their decision to use the apps or not.

There are several reasons for data protection in mHealth apps, particularly to address the risk of any unauthorized to keyed-in information and stored data by hackers [9]. In addition, data management and storage, data privacy disclosure, data integration, data encryption, app operability, and authentication are established factors contributing to data breaches [9].

Several studies have highlighted the connection between patients' awareness and the risk of data breaches. End users have an obligation for the security and privacy of their data to be maintained [4]. As the main stakeholders of the health care system, patients have a contractual relationship with health care providers as the latter are expected to ensure the safety and confidentiality of patients' health information. Health care app developers must protect sensitive patient data by complying with data privacy regulations such as the General Data Protection Regulation (GDPR) and HIPAA (Health Insurance Portability and Accountability Act). To ensure data privacy and security, mHealth apps are encouraged to incorporate data encryption, implement secure authentications, and perform regular risk assessments [10]. While HIPAA encompasses physical, administrative, and technical aspects to ensure the security of personal health information, the GDPR requires health centers or organizations to collect detailed consent from users before recording their personal data and giving them the right to access, amend, delete, or restrict the processing of their data. These privacy requirements for app security are known not only by mHealth app providers but also by patients and users [10,11]. This represents another dimension that may influence patients' perspectives on and adoption of mHealth apps.

Objectives

Despite the pivotal role of patients' views and awareness in the successful implementation of mHealth apps, as demonstrated in several studies, the findings are yet to be summarized to elucidate the barriers and facilitators, which may assist clinicians, HCPs, policy makers, and other stakeholders in their decision-making processes. A previous systematic review on the security and privacy of mHealth apps was conducted almost a decade ago and did not focus on any specific stakeholders [12]. Meanwhile, 2 other reviews related to this topic were a narrative and a scoping review [13,14], which are open to bias as the methods used in retrieving the reviewed articles were not succinctly described. This study aimed to fill the research gap by conducting a systematic review to elucidate patients' perspectives and awareness of the privacy, security, and confidentiality of mHealth apps, as well as the associated factors.

Methods

This study was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [15]. This systematic review was retrospectively registered in PROSPERO (CRD42023456658).

Search Strategy

The systematic article search was conducted using 3 electronic databases: PubMed, PsycNet, and ScienceDirect. PsycNet was accessed via Ovid as a search interface. These databases were selected given their suitability and specificity for research in health and medical sciences, thus increasing the chances of retrieving articles relevant to the research topic. The first author of this systematic review performed the literature search from

XSL•FO
RenderX

February 2022 to April 2022. Articles published between 2014 and 2022 and written in English were considered in the literature search. We focused on studies published from 2014 to 2022 given the growing interest in the use of mHealth apps in the last decade [2].

Aligning with the objectives of this review, the search terms were broadly categorized into 3 components or groups of keywords. Alternative keywords were permitted for each component as denoted using the Boolean operator "OR." The separator "AND" was then used to combine each component with other wordings. The search query for each of the databases is presented in Table 1.

Table 1. The search terms used and the total number of publications retrieved from each database.

Database	Search string and search terms	Initial search results (number of articles)	Date of retrieval
Scopus	Main search terms using document title and abstract: ("Mobile health" OR mhealth OR "mobile phone*" OR "Smart phone*" OR Smartphone* OR tablet*) AND ("Perspective OR opinion* OR attitude* OR perception* OR awareness") AND (Privacy OR confidential* OR security)	1277	February 20, 2023
ScienceDirect	("Mobile health" OR mhealth OR "mobile phone") AND ("Perspective OR opinion OR awareness") AND (Privacy OR confidential OR security)	7156	March 21, 2023
APA PsycNet	"Mobile health" OR mhealth OR "mobile phone*" OR "Smart phone*" OR Smartphone* OR tablet* AND "Perspective OR opinion* OR attitude* OR perception* OR awareness AND Privacy OR confidential* OR security"	441	April 19, 2023

Study Eligibility Criteria

The inclusion and exclusion criteria of the review are presented in Table 2. The first inclusion criterion was the document type, whereby only published original articles were considered. Other document types, such as review articles, chapters in books, books, and conference proceedings, were all excluded. The next inclusion criterion was the publication year, whereby only articles published from 2014 to 2022 were selected. The inclusion or exclusion of retrieved articles was based on agreement among the authors. Issues arising during the process were resolved through consensus.

Table 2. Inclusion and exclusion criteria.

Consideration factor	Inclusion criteria	Exclusion criteria
Study design	Empirical studies involving qualitative, quantitative, or mixed methods	Reviews (systematic, scoping, narrative, and rapid reviews and meta-analyses)
Publication type	Peer-reviewed journal articles	Editorial letters, protocols, expert opinions, policy briefs, theses or dissertations, and conference papers
mHealth[a] apps	Studies involving mHealth apps regardless of their aim, target disease, and system type (eg, iOS or Android)	Studies involving apps that are not linked to CISs[b]
Study participants	Studies involving patients with or without HCPs[c], such as physicians, nurses, pharmacists, and care teams, regardless of demographic characteristics (ie, age, gender, and ethnicity)	Studies solely involving app developers, HCPs, and telehealth providers
Outcomes	Experience, perceptions, awareness, and knowledge of patients after practical use of mHealth apps	Description of the impact of mHealth apps on the patient-HCP relationship
Language	English	Other languages
Year	2014 onward	Before 2014

[a]mHealth: mobile health.

[b]CIS: clinical information system.

[c]HCP: health care professional.

Study Selection

In total, 2 researchers performed the study selection independently. Articles retrieved from the primary literature search from each database were sent to the researchers' email repositories and stored for future reference. The screening process was carried out using the filter feature available in all 3 databases. The initial search results were checked for duplicates, which were then identified and removed accordingly. The last screening stage was full-text reading.

Data Extraction

The final articles included in this study were assessed, reviewed, and examined upon completing the eligibility process. An Excel (Microsoft Corp) spreadsheet form was created to use in data extraction. The data extracted from the studies were as follows: first author; year of publication; study location; study design; types and purposes of mHealth apps; issues related to the privacy, security, and confidentiality of mHealth apps; and the main findings. Data extraction was performed by the first author, and thus, the intercoder agreement was not assessed.

XSL•FO
RenderX

Quality Appraisal

Quality appraisal was not performed in the review due to the heterogeneity of the research objectives, designs, and methodology used in the included studies [16].

Data Synthesis

A narrative synthesis was considered in this review due to the heterogeneity of the designs used in the studies. Specifically, thematic analysis was conducted to summarize the findings of the included studies. The data extracted were analyzed thematically. All the authors participated in the discussion to determine the themes that would be synthesized from the analysis. The themes decided on were (1) barriers to and facilitators of using mHealth apps and (2) recommendations to increase the use of mHealth apps by addressing privacy, security, and confidentiality issues. Any further analysis and reassessment of the themes and subthemes was conducted continuously.

Results

Search Outcomes

A total of 1696 articles were retrieved from the initial searches on PubMed (n=659, 38.86%), ScienceDirect (n=172, 10.14%), and Ovid (n=865, 51%). Of the 1696 search results, 425 (25.06%) were removed from the list as duplicates (Figure 1), whereas 1121 (66.1%) were considered ineligible upon screening the titles and abstracts. The 150 remaining articles were then subjected to a full-text review, which led to the final selection of 33 (22%) articles for the systematic review based on the inclusion and exclusion criteria (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

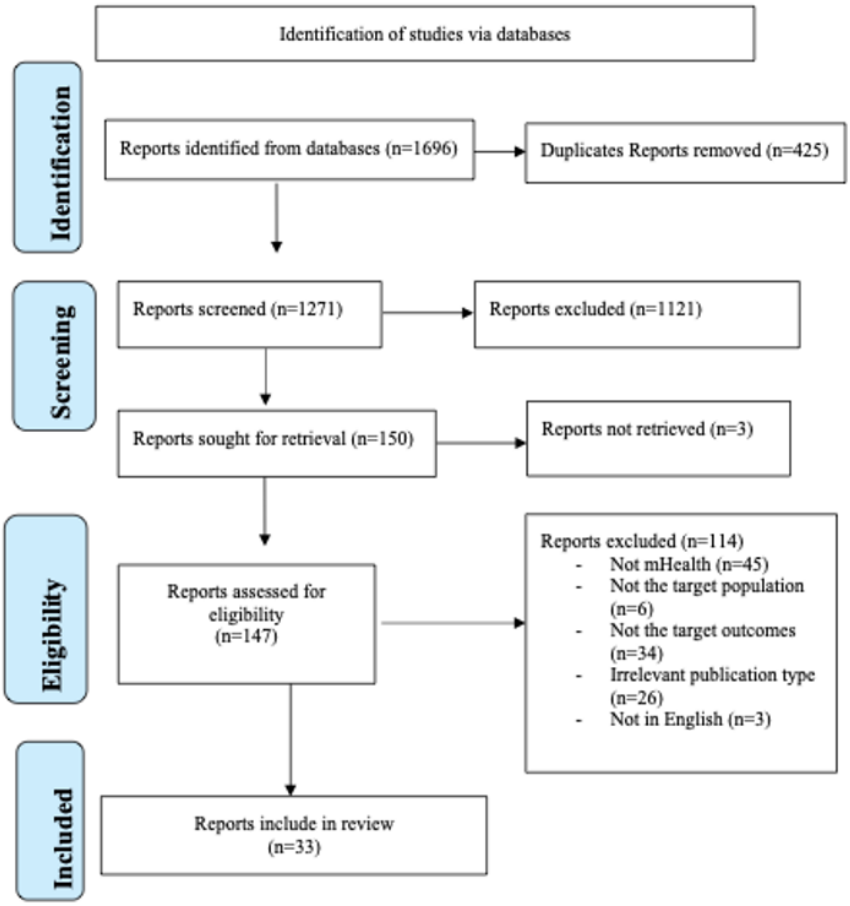

Characteristics of the Studies

As shown in Table 3, the highest number of studies was published in 2021 (9/33, 27%) and 2019 (7/33, 21%). The included studies were quantitative (13/33, 39%), qualitative (11/33, 33%), and mixed methods (9/33, 27%). In terms of study location, most of the studies were conducted in countries with countries with sufficient resources (26/33, 79%) compared to those undertaken in resource-limited countries (7/33, 21%). While 9% (3/33) of the articles emphasized the general use of mHealth apps for routine health records, 36% (12/33) of the articles focused on specific mHealth apps for collecting patients' health data and management of various medical conditions, such as chronic obstructive pulmonary disease [17], cancer [18-21], and diabetes [22], and pregnancy care [23,24] (Table 4).

XSL•FO
RenderX

Table 3. Descriptive information of the articles (N=33).

Variables	Studies, n (%)
Year of publication	
2014	2 (6)
2015	2 (6)
2016	2 (6)
2017	2 (6)
2018	4 (12)
2019	7 (21)
2020	4 (12)
2021	9 (27)
2022	1 (3)
Study design	
Qualitative	13 (39)
Quantitative	11 (33)
Mixed methods	9 (27)
Study location	
High-income countries	26 (79)
Middle-high–income countries	7 (21)
Low-income countries	0 (0)
Purpose of mHealth[a] apps	
Routine electronic health records	3 (9)
Specific apps for patient management	12 (36)
Unspecific	18 (55)

[a]mHealth: mobile health.

XSL•FO
RenderX

Table 4. Details of author name, publication year, type of intervention, study design, setting, instruments, and findings of the 33 reviewed articles.

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Wyatt et al [25]	PhotoExam apps (EHRs[a])	Quantitative	United States	App use and survey	(1) Only 3% (2/71) of patients expressed privacy or confidentiality concerns related to the photos taken; (2) 26% (18/70) of patients mentioned that the security features of the apps were explained by the HCPs[b]; and (3) 16% (11/70) indicated that the security features were not explained, and most of them (41/70, 59%) did not remember and were unsure of whether the security features were explained	Concerns about privacy; concerns about confidentiality
Zhou et al [4]	mHealth[c] apps	Mixed methods	United States	Questionnaire and qualitative analysis (semistructured interview and psychometric analysis)	(1) Participants had some level of concern about the privacy of their personal data and wanted to have some specific protections; (2) Participants believed that a level of privacy protection is currently available in mHealth apps; (3) Participants desired to have informed consent, access control, a privacy policy, and remote wiping features in mHealth apps; (4) two-thirds of the users (66.7%) stated that the cost and lack of security features of mHealth apps were the main barriers to adopting the technology.	Concerns about data privacy; concerns about data security
Natsiavas et al [26]	eHealth data	Quantitative	14 European countries	Questionnaire	(1) 11.96% did not think about possible health data risks, and 36.41% felt informed about these risks; (2) 66.21% of the respondents did not read the "Terms and Conditions," with >30% declaring that they did not feel that it was worth it given the time required to read them and 19.79% declaring indifference toward them; (3) 26.09% of the respondents felt confident regarding their eHealth data privacy, 38.04% felt concerned but helpless, and 16.3% stated that they avoided using eHealth services due to the lack of confidence regarding their data handling; (4) 20% of the respondents felt that their privacy was fully covered in the "Terms and Conditions" of the apps, and 12.5% declared that they did not understand them; (5) 72.46% of respondents were willing to share their personal data for research purposes, at least under anonymization	Awareness of data privacy and confidentiality; concerns about data privacy

XSL•FO
RenderX

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Alaiad et al [27]	mHealth model	Quantitative	Jordan	Questionnaire	(1) Security and privacy risks have a direct negative effect on the patients' intention to use mHealth; (2) mHealth patients in lower-income countries are often using mHealth services at their own risk, which makes them prone to data breaches and misuse by unknown parties	Data security; data privacy
Özkan et al [28]	EHRs	Quantitative	Turkey	Survey	(1) Most of the participants (60.9%) stated that they did not know who had the right to access their medical records, and 7.4% believed to have comprehensive knowledge on the topic; (2) the medical records of 9.7% of the respondents had been used or released without their consent; (3) 15.1% stated that they avoided being tested due to violation risks, and 3.5% asked their physicians to enter a less embarrassing health status in their records; (4) most participants (94%) responded that they should have full access to their medical data, whereas 50% of respondents wanted other parties (children, parents, physicians, spouses, and other hospital staff) to have limited access rights	Awareness of data confidentiality and privacy; data confidentiality; data privacy
Glauser et al [29]	NeuroPath (mHealth app)	Qualitative	United States	App use and telephone interview	14 (63.6%) respondents stated that they waited before trying new technology, and 6 (27.3%) of these respondents were concerned about data security when asked about their readiness to adopt new technologies.	Concerns about data security
Zhou et al [30]	SecSim (security simulator)	Mixed methods, quasi-experimental study and interview session	United States	App use, IBM PSSUQ[d], and open-ended interview questions	(1) The comparison between the pre– and post–security education selection regarding security settings indicated that 21% (14/66) to 32% (21/66) of participants chose a stronger security measure in text encryption, access control, and image encryption; 0% (0/66) to 2% (1/66) of participants chose a weaker measure in these 3 security features; and the remainder kept their original selections; (2) a significant percentage of patients (21%-32%) needed guidance to make an informed selection regarding security settings	Concerns about data security

XSL•FO
RenderX

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Barutçu et al [31]	e-Pulse (mHealth app)	Quantitative	Turkey	Survey	(1) Participants considered mHealth apps as less credible in protecting their personal information; (2) positive relationships were found between user satisfaction level with mHealth apps and ease of use, trust, privacy, usefulness, and information quality of mHealth apps; (3) mHealth apps' user satisfaction was negatively influenced by the privacy of mHealth apps; (4) perceived ease of use, trust, privacy, perceived usefulness, and information quality were the major factors influencing satisfaction with and intention to adopt mHealth apps	Concerns about data privacy; associated factors: satisfaction with mHealth apps
Bauer et al [32]	mHealth platform supporting collaborative care	Mixed methods	United States	App use, surveys, and interviews	(1) Patients felt that the data they submitted were not entirely secure, but they did not believe that the information reported in the apps was highly personal, and therefore, the potential for a data breach was not a major concern; (2) some patients wished to have a better understanding of who else had access to their health information and the ability to control such access (based on qualitative findings)	Data security; awareness of data security and privacy
Goetz et al [33]	PRELAX (eHealth application)	Mixed methods	Germany	Application use, self-administered questionnaire, and semistructured interviews	(1) Hospitalized women (6/30, 20%) were worried about unauthorized third-party access to their stored medical data; (2) several women (8/30, 27%) expressed concerns about data security, especially in the field of mobile apps as many free apps make private data easily accessible; (3) data security and personal data storage in pregnancy apps were general causes for concern	Data confidentiality; data privacy

XSL•FO
RenderX

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Richardson and Ancker [34]	her	Quantitative	United States	Survey	(1) Approximately three-quarters of individuals believed that storing medical information on a phone would threaten privacy and security (74% in 2013 and 75% in 2014); (2) approximately two-thirds thought that sharing data between a mobile device and a physicianherEHR would threaten privacy and security (69% in 2013 and 67% in 2014); (3) participants expressed greater concern about worsened privacy and security with storing data on mobile phones and mobile herne–EHR communication (74% in 2013 and 69% in 2014) than with the health information exchange between EHRs and physicians (41% in 2013 and 47% in 2014)	Data privacy; data security
Sanger et al [35]	Postacute care apps	Qualitative	United States	Semistructured interviews	Participants were most concerned about the collection and transmission of particularly sensitive information, such as photos of the groin area	Data privacy; data security
Dang et al [36]	Not specific	Quantitative survey	China	Survey	Privacy concerns among patients were positively enhanced by perceived health information sensitivity (β path coefficient=0.505; *P*<.001). Patients' health information disclosure intention decreased significantly with higher concern levels (β path coefficient=–0.338; *P*<.001). The relationship between perceived health information sensitivity and privacy concerns was negatively moderated (β path coefficient=–0.17; *P*=.09) by the informational support dimension. A similar moderating effect was observed in the association between privacy concerns and health information disclosure intention (β path coefficient=–0.11; *P*=.09).	Data privacy; data security; associated factors: health information disclosure
Alwashmi et al [17]	Not specific	Mixed methods	Canada	In-depth interviews and survey instrument	The barriers to adoption were technical issues, lack of awareness, potentially limited uptake from older adults, and privacy and confidentiality issues.	Awareness of data privacy and confidentiality; associated factors: demographic factors; data privacy issues and confidentiality issues; facilitators and enablers
Biswas et al [37]	ACCU3RATE (a specific AI[e]-enabled mHealth app rating scale)	Quantitative	Multinational study	Survey instrument	Patients provided positive feedback regarding the apps' features for accessibility, protection, and privacy of patient data.	Data privacy

XSL•FO
RenderX

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Casilang et al [38]	mHealth for the development of an exclusive breastfeeding tool	Qualitative	Dominican Republic	Focus group discussion	Barriers to mHealth use included the cost of internet service, privacy concerns, and perceived credibility of information sources	Facilitators and enablers; data privacy
Zhang et al [39]	Unspecific	Quantitative	United States	Web-based survey	Data type (P=.003), data stage (P<.001), privacy victimization experience (P=.01), and privacy awareness (P=.08) showed positive effects on patients' privacy concerns. Higher privacy concerns were reported for social interaction data (P=.007) and self-reported data (P=.001) than for biometric data. Privacy concerns were also higher for data transmission (P=.01) and data sharing (P<.001) than for data collection. Privacy concerns affected the attitude toward privacy protection (P=.001), thereby affecting continuous use intention.	Privacy concerns; associated factors: data type and stage and victimization experience
Harris et al [40]	Mobile phone health apps	Mixed methods	Pakistan, Tanzania, Kenya, Nigeria, and Bangladesh	Scoping study and analysis of survey data	Stakeholder willingness was high provided challenges regarding technology, infrastructure, data security, confidentiality, acceptability, and health system integration were addressed. Mobile consultations can reduce affordability barriers and facilitate care-seeking practices.	Data security; data confidentiality
Moodley et al [18]	Mobile phone health apps	Mixed methods	South Africa	Survey and focus group discussion	Users were interested in the use of mobile phone apps for health intervention in receiving Papanicolau smear results and appointment reminders. However, concerns were raised regarding the confidentiality of SMS text messages, loss or theft of mobile phones, receiving negative results, and the accessibility or clarity of the language used to convey the messages.	Data confidentiality
Li et al [24]	mHealth technology for monitoring pregnancy care	Qualitative	Australia	In-depth interviews	The clinical and technical challenges regarding the introduction of mHealth for pregnancy care were also identified, whereas usability and data privacy were among the main concerns of the participants.	Data privacy; facilitators and enablers
Gill et al [21]	mHealth technology to support postabortion care	Mixed methods	United States	Survey and semistructured interviews	Qualitative analysis revealed that participants preferred a comprehensive website with secured email or SMS text message notifications to provide tools and resources for emotional well-being, contraceptive decision-making, general sexual health, and postprocedural care.	Data security; facilitators and enablers

XSL•FO
RenderX

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Bradbury et al [19]	Remote real-time videoconferencing for patients with cancer	Quantitative and experimental	United States	Experimental design	Most patients reported that their privacy was respected after the first (56/57, 98%) and second sessions (40/41, 98%), respectively. Meanwhile, some patients reported concerns that RVC[f] might increase the risk of a confidentiality breach of their health information—after V1: 14/57 (25%); after V2: 12/41 (29%).	Data confidentiality
Al-Anezi [41]	Unspecific	Quantitative	Saudi Arabia	Survey	Fear of privacy violations, fear of loss of personal data and information, and lack of technical support were highlighted as the main reasons for the lack of motivation to adopt the mHealth system	Privacy issues
Ermakova et al [42]	Health clouds	Quantitative	Germany	Survey	Confidentiality assurance was vital in influencing individuals' acceptance of health clouds for sensitive medical data, but such an effect was lacking for nonsensitive medical data.	Data confidentiality
Rodrigues et al [43]	Unspecific	Qualitative	South India	In-depth interviews	Participants perceived the risk of unintentional disclosure of their HIV status and the stigma thereof via the intervention and took initiatives to mitigate this risk.	Data confidentiality
de Vries et al [22]	Unspecific	Qualitative	The Netherlands	Focus group discussions and interviews	The use of mobile apps for reporting ADRs[g] was influenced by source of information, app's security, type of feedback, storage pattern of ADR reports, ease of use, and the type of language.	Data security; facilitators and enablers
Al-Mahrouqi et al [44]	Unspecific	Qualitative	Oman	Semistructured qualitative interviews	Although clients acknowledged the positive impact of telehealth in improving mental health care services in Oman, primary concerns were related to privacy, the security of telehealth systems, lack of public tele–mental health services, lack of specified tele–mental health guidelines, shortage of trained therapists, and limited access to high-speed internet and electronic devices.	Concerns about data privacy; concerns about data security
Turcotte et al [45]	Teleconsultation (AYP[h])	Qualitative	Canada	Semistructured interviews	Although both users and nonusers showed positive experiences and perspectives on the AYP platform were mostly positive, concerns were raised regarding patients' safety.	Data security issues

XSL•FO
RenderX

Study	Type of intervention	Study design	Study location	Data collection method	Main findings	Themes identified
Hackett et al [23]	Mobile phone apps	Qualitative	Tanzania	Semistructured interviews	(1) Perceptions of personal privacy and confidentiality were negatively and positively impacted by the use of new technologies to capture health service user data during pregnancy and childbirth; (2) women's concerns regarding privacy aligned closely with a belief that pregnancies and expected delivery dates must be kept secret, reflecting fears that pregnancy renders women vulnerable to witchcraft by jealous neighbors; (3) they were also concerned that health workers' male partners could access their private information.	Concerns about data privacy; concerns about data confidentiality; facilitators and enablers
Morton et al [46]	Mobile phone apps	Mixed methods	Many countries	Survey and in-depth interviews	Concerns regarding data security were prevalent. Data security, content quality or accuracy, ease of use, and cost were among the prioritized mHealth features. The ability to share data with others was described as vital by less than half of the respondents.	Concerns about data security
Cavazos-Rehg et al [47]	mHealth mental health intervention	Quantitative	United States	Survey	The main reasons provided for unwillingness to obtain parental consent to participate in the intervention included the importance of preserving privacy and the feeling that parents lack awareness or understanding of mental health issues.	Privacy concerns; awareness of data privacy
Sangers et al [20]	mHealth apps for skincare screening	Qualitative	The Netherlands	Semistructured interviews	The main barriers to using mHealth apps included privacy concerns, perceived lack of value, perception of untrustworthiness, preference for a physician, a complex user interface, and high costs.	Privacy concerns; facilitators and enablers
Hattingh et al [48]	mHealth in pharmacy settings	Qualitative	Australia	In-depth interviews and focus group discussion	Consumers indicated a desire to receive information in a way that respects their privacy and confidentiality in an appropriate space. Important areas were identified that require improved protection of privacy and confidentiality during pharmacy interactions.	Privacy, security, and confidentiality issues

[a]EHR: electronic health record.

[b]HCP: health care professional.

[c]mHealth: mobile health.

[d]PSSUQ: Post-Study System Usability Questionnaire.

[e]AI: artificial intelligence.

[f]RVC: remote videoconferencing.

[g]ADR: adverse drug reaction.

[h]AYP: Ask Your Pharmacy.

XSL•FO
RenderX

mHealth Apps for Specific Interventions

A total of 36% (12/33) of the studies included in this review reported the use of mHealth apps for patient health data collection and management of specific health conditions. In total, 9% (3/33) of the studies entailed the use of specific mHealth apps for assessing patient treatment progress [25,29,35]. PhotoExam apps [25] entailed the collection of patients' photos, which were then assessed for patient response to the treatment provided. Glauser et al [29] developed an app named "NeuroPath" with the support of Apple, the Institutional Department of Neurosurgery, and the Department of Information Technology. The areas targeted by the app included patient surgical preparation, prevention of perioperative risk, wound care, activity monitoring, and opioid use management. Meanwhile, Sanger et al [35] focused on mHealth apps for postacute care.

Regarding specific medical conditions, 9% (3/33) of the studies focused on mHealth apps for pregnancy management, such as a patient engagement pregnancy app (PRELAX) [33], pregnancy care [24], and support for postabortion care [14]. Other studies involved mHealth apps for real-time videoconferencing for patients with cancer [19], mental health interventions [47], support for collaborative care [32], and teleconsultation for pharmaceutical services [43].

Finally, 6% (2/33) of the studies emphasized the security of mHealth apps without focusing on medical conditions or groups of patients. Zhou et al [30] developed a security simulator named SecSim to reveal the consequences of selecting different security options available in the security settings of mHealth apps. Meanwhile, Biswas et al [37] used a specific artificial intelligence–enabled mHealth app rating scale, ACCU3RATE, to obtain users' feedback on the security features. The influence of these interventions on patient perception on the security, confidentiality, and privacy of data collected via mHealth apps is presented in the thematic analysis.

Results of the Thematic Analysis

The thematic analysis generated four broad themes from the findings of the studies: (1) privacy, confidentiality, and security; (2) awareness of privacy, security, and confidentiality; (3) facilitators and enablers; and (4) associated factors. The following subsections present more detailed information about the synthesized themes.

Theme 1: Privacy, Confidentiality, and Security

As expected, all the included studies (33/33, 100%) investigated privacy, confidentiality, and security issues related to patients' use of mHealth. Different levels of privacy and confidentiality concerns were reflected in the studies. In 15% (5/33) of the studies, less than half of the patients expressed concerns about the privacy or confidentiality of the various data required by mHealth apps [19,25,26,28,33]. Meanwhile, in 6% (2/33) of the studies, >50% of the respondents raised diverse issues regarding the privacy and confidentiality of their data [4,34]. In terms of data security, 9% (3/33) of the studies, which used a quantitative design, revealed that a higher proportion of patients (>50%) acknowledged issues related to data security [29,30,34].

This theme was also depicted in several qualitative and mixed methods studies [4,20,22-24,37,39,40,45,48]. For instance, patients opined that the privacy protection level in mHealth apps needed to be improved [4,29,38]. Meanwhile, patients felt that their data were not completely secure and were concerned about data breaches [21,32]. In 9% (3/33) of the studies, issues related to privacy, confidentiality, and security were identified as barriers to mHealth use [24,38,40]. Meanwhile, Biswas et al [37] found that respondents were satisfied with the apps' features for the protection and privacy of patient data.

Theme 2: Awareness of Privacy, Security, and Confidentiality

Patients' awareness of mHealth apps' privacy, security, and confidentiality was highlighted in 12% (4/33) of the studies [4,25,26,28]. Nevertheless, the awareness level differed across the studies.

Natsiavas et al [26] found that 12% of patients in their study were unaware of the possibility of health data risks, whereas a higher percentage of participants (61%) in the study by Özkan et al [28] did not know who had the right to access their medical records. Thematic analysis of the data gathered by Bauer et al [32] revealed that patients opted for a better understanding of other parties who have access to their health information and their capacity to regulate such access. Meanwhile, Alwashmi et al [17] and Zhang et al [39] identified awareness of privacy and confidentiality issues as barriers to adopting mHealth apps and raised concerns about data privacy, respectively.

Theme 3: Facilitators and Enablers

The third theme synthesized in this systematic literature review is the facilitators of increased use or adoption of mHealth apps among patients based on the perceived benefits of mHealth apps and recommendations to address data privacy, security, and confidentiality issues. In total, 18% (6/33) of the studies reported the perceived benefits of mHealth apps that may reduce patients' concerns about issues related to data privacy and security, thus improving the adoption rate of such apps [17,21-24,30]. Some of the benefits highlighted by patients included improved health status by reducing the rate of hospitalization [17], increased trust, better patient-HCP relationships [22,23], and exchange of information in real time [21]. Overall, patients viewed mHealth apps installed on smartphones as an added value, which assisted in improving the confidentiality of their data, their trust, and their relationship with health care personnel.

As for recommendations and facilitating conditions to address data privacy, security, and confidentiality issues related to mHealth, the consistent points raised in the studies were the protection of personal data, ensuring that health status or medical conditions are not mentioned, brief training or education on data security, and assuring data confidentiality and privacy [21,24,30]. Gill et al [21] found that participants prioritized privacy and confidentiality by preferring discrete mHealth designs that did not mention the specific medical condition that prompted them to visit the clinic.

XSL•FO
RenderX

Theme 4: Associated Factors

The last theme gleaned from this review was the factors associated with patients' concerns regarding issues related to data confidentiality, privacy, and security when using mHealth apps. This theme was synthesized from the findings reported in 15% (5/33) of the articles [4,31,36,39,42], which comprised patients' sociodemographic factors, satisfaction with mHealth, data type and stage, and experience with mHealth apps.

In terms of sociodemographic factors, Zhou et al [4] found that married patients showed higher information security and privacy concerns and desired more stringent security protection compared to single patients. The weakest concerns about privacy and security were exhibited by users with <US $10,000 in annual income compared to patients who earned >US $75,000 annually. Similarly, patients in the older age group (51-65 years) reflected a higher level of concern about privacy in mHealth apps relative to the younger age group (18-28 years). In terms of experience, participants who had previously used mHealth apps had greater concerns about data security and privacy despite still being interested in continuing to use the technology.

Only 3% (1/33) of the studies reported the association between satisfaction levels with mHealth apps and privacy concerns [31]. Specifically, a positive relationship was observed between user satisfaction levels with mHealth apps and privacy concerns. Meanwhile, Dang et al [36] found that a higher perceived health information sensitivity heightened the privacy concerns (β path coefficient=0.505; $P<.001$) raised by patients regarding mHealth apps. The provision of informational support moderated the association between privacy concerns and health information sensitivity.

Zhang et al [39] revealed the diverse levels of privacy concerns depending on data type, data stage, and privacy victimization experience. For instance, privacy concerns were higher for patients' social interaction, self-reported, and biometric data. Users were less concerned about privacy issues during data collection compared to the data transmission and sharing stages. These events had negative impacts on the continuous intention to use mHealth apps [39]. Ermakova et al [42] also found that patients' acceptance of health clouds for nonsensitive medical data was not significantly affected by confidentiality assurance; however, this relationship was significant for sensitive medical data.

Discussion

Principal Findings

This systematic review evaluated patients' perspectives and understanding of the data confidentiality, privacy, and security of mHealth apps connected to clinical information systems. A total of 33 relevant articles were extracted and included in this systematic review using the PRISMA guidelines. Descriptive analyses revealed that most of the studies (26/33, 79%) were conducted in high-income countries compared to those undertaken in middle- to low-income countries (7/33, 21%). These findings reflect the disparity in the implementation and adoption of mHealth apps in line with different countries' economic status and infrastructural capacity. This is evident in the use of mHealth apps for the management of specific medical conditions in the United States, the United Kingdom, Australia, and Turkey, whereas the few studies conducted in lower-income countries focused mainly on either the feasibility or introductory stages of general mHealth apps. Notwithstanding the patients' socioeconomic status, data privacy, confidentiality, and security issues were highlighted in most studies.

The extensive research on mHealth apps was also reflected in the diverse medical conditions in which the technology was explored, such as chronic obstructive pulmonary disease [17], cancer [19,20], postabortion care [21], diabetes [22], and pregnancy care [24]. Descriptive analyses also revealed the use of various research methods (qualitative, quantitative, and mixed methods) in the reviewed articles, which is not surprising given that the research topic can be explored via interviews, focus group discussions, and surveys.

Thematic analyses of the qualitative studies revealed 4 main themes, comprising primary concerns about privacy, confidentiality, and security of data; awareness of privacy, security, and confidentiality issues; facilitators and enablers; and associated factors. Regarding the first theme, patients were concerned about data security and privacy, particularly in terms of the collection and transmission of sensitive information such as identity-revealing data and images of body parts [22,33,35]. Some hospitalized patients were even more concerned about unauthorized third-party access to their medical data given that mHealth apps are mostly free and easily accessible [17,33]. In contrast, some patients were indifferent to these issues in the same studies as they were willing to use mHealth apps and share their health data with HCPs. These diverse views could be linked to patients' consideration of the benefits and risks associated with using mHealth apps for routine health records or managing their health conditions.

We observed concordant perspectives on the research topic as patients and end users consistently raised concerns about data privacy, security, and confidentiality issues that prevented them from using mHealth apps [18,20,35]. This result aligns with those of a previous review by Nurgalieva et al [13] in which low levels of security and privacy were reported as the main reason for low use among patients and end users. Serious issues may arise from mHealth apps with low levels of security or privacy, and such events may have severe consequences for users and organizations.

Given the extensive privacy and security issues raised in the reviewed studies, the findings suggest the need for mHealth environments to improve the security of these apps by exploring advances in cyberspace security [4]. Similarly, the reviews both by Martínez-Pérez et al [14] and Nurgalieva et al [13] highlighted security incidents, including vulnerabilities discovered in widely used mHealth apps and malware attacks. The concerns raised by patients are plausible as the proliferation of mobile devices with location sensors has facilitated access to location-based services [14]. These advanced devices transmit the user's location information to third-party location servers, which are accessible by other service providers. Users aware of this potential data breach may feel that they are continuously tracked. In this review, some patients seemed to be aware of

XSL•FO
RenderX

the risk of third-party and unauthorized access to their medical data [4,25].

For the second theme, both qualitative and empirical findings reflected that patients' awareness of the privacy and security issues of mHealth apps differed across the studies. For example, more than two-thirds of participants expressed their concerns about personal data security and privacy and requested user authentication and data encryption to protect user data [4,34], and <12% were unaware of the risk of health data breaches [19,31]. Meanwhile, more than half of the respondents in the studies conducted by Özkan et al [28] and Wyatt et al [25] were unsure about the privacy and security information on their mHealth apps. Thus, the different awareness levels among patients may influence their diverse perspectives on the privacy and security of mHealth apps.

The third theme emerging from the thematic analyses entailed the facilitators for increased use or adoption of mHealth apps based on patients' perceived benefits of mHealth apps and suggestions to address data privacy, security, and confidentiality issues. Resultantly, improved health status [17], better patient-HCP relationships [22,23], and trust were the main benefits mentioned by patients. On the other hand, personal data protection, ensuring that health status or medical conditions are not mentioned, brief training or education on data security, and assuring data confidentiality and privacy were the consistent recommendations provided by patients [21,24,30]. These findings coincide with the suggestion by Perera [49] regarding the use of an alphanumeric passcode to ensure the protection of mHealth apps rather than using a 4-digit personal identification number, as well as wiping data from the mobile device after a specific number of failed passcode attempts [49,50]. In their review, Nurgalieva et al [13] also emphasized the frequency of notifications and alerts programmed into mHealth apps. Accordingly, discreet or private notifications were advised to prevent any distress to users, particularly in situations in which someone else could accidentally view the app icon. Overall, these recommendations reflect the need for users to have complete control in using their mobile devices for mHealth and avoid any intrusion in their daily life. Training and education as recommended by patients in this review corroborate the report by Lewis and Wyatt [50] as reviewed by Nurgalieva et al [13]. The latter authors suggested that the risk factors for violation of users' privacy and security in mHealth apps can be categorized into external and internal risks. Appropriate regulation can be used to effectively minimize the internal risk factors, whereas proper education and training are pertinent to eliminate the external risk factors.

Concerning the fourth theme, information on the factors influencing patients' views regarding data privacy and confidentiality was obtained from a few of the studies included in this systematic review (5/33, 15%). Examples included sociodemographic characteristics such as age, income level, marital status, previous experience with mHealth apps [30], patients' satisfaction levels [31], perceived health information sensitivity [36,42], data type, data stage, and privacy victimization experience [39,42]. These findings are consistent with the results of several previous studies [51] and the contextual nature of the theory of privacy [52]. More importantly, certain adjustments to the security and privacy features of mHealth apps need to be incorporated by the app developers upon considering patients' and users' demographics. However, given the low number of studies reporting the underlying factors influencing patients' views on data privacy, confidentiality, and security issues in mHealth apps, more research is needed to elucidate the relationships.

Implications of the Findings

The findings of this study have pertinent implications for mHealth app developers, HCPs, and policy makers. Both HCPs and mHealth app developers have a vital role to play in addressing the diverse views exhibited by patients and end users on data privacy, confidentiality, and security issues in mHealth apps.

While mHealth app developers are primarily responsible for designing security measures and features to ensure their apps' data privacy and confidentiality, from the perspective of health care provider and patient relationships, the former play a pivotal role in educating patients or end users regarding such measures. This could be discussed during routine consultations, as well as reiterating the need for collecting sensitive data and what they are going to be used for. Regarding unauthorized third-party access and potential data breaches, users may benefit from information on the privacy requirements and meeting of the standards set by the GDPR and HIPAA, which are designed to ensure that such privacy issues and data breaches are prevented.

Apart from informing patients and end users on the type of data to be collected and the intended use, the aspect of training and educating end users on the available features and measures to ensure data privacy and security cannot be overemphasized. Meanwhile, stakeholders need to gauge the users' level of awareness and knowledge of these issues as well as the underlying reasons for the diverse views on these issues to tailor educational interventions accordingly. As gleaned from the studies included in this review, users' lower concern about their data privacy may stem from being completely ignorant, or they may perceive that the benefits of using mHealth outweigh the potential risks.

This study also has important implications for mHealth app developers given the fact that patients and end users have raised concerns about data privacy, security, and confidentiality issues that end up affecting their use of such apps [18,32]. Nevertheless, accumulated findings reflect that patients' perceptions on privacy issues and the latter's influence on patients' adoption of mHealth apps are context dependent [39,43]. As most of the reviewed studies involved surveys, it is expected that users will raise concerns on privacy or confidentiality issues related to mHealth apps, but further inquiries are required to elucidate whether such concerns influence their behavior and actual use of such apps. These events need to be succinctly explored in future studies, and stakeholders need to understand this gray area to effectively address the issues. In addition, mHealth app developers may evaluate the issues raised by users and strategize on how to improve their current security and privacy measures, particularly ensuring that they meet the requirements of the HIPAA and GDPR. Furthermore, the appropriate actions to be taken by end

XSL•FO
RenderX

users in different circumstances should be clarified in the developed security policy and guidelines. In summary, providing suitable awareness in the security guidelines and policies for end users is as pertinent as developing secure mHealth apps.

For policy makers and researchers, this study highlights the aspects prioritized by patients regarding the adoption of and desired outcomes of implementing mHealth apps, as well as addressing issues related to data privacy and security. Accordingly, better patient-HCP relationships, trust, personal data protection, ensuring that health status or medical conditions are not mentioned, and brief training or education on data security were the consistent recommendations provided by patients [19,23,24]. While most of the aforementioned points can be conveyed to users, trust is a feature that has to be earned, which has to be driven by health care providers and the government. These stakeholders also have to mitigate apprehensions related to patients' privacy concerns to ensure enhanced trust between patients and service providers, which is crucial for the successful delivery of eHealth services.

Certain concrete actions can be taken by stakeholders at the government level by considering HIPAA and the European Union Safe Harbor law, which advocates for strict security measures for the exchange and sharing of health data. Failure to comply with such laws entails severe consequences. From health care providers' perspective, and given the confidentiality and sensitivity of patient data, only authorized users such as medical staff should be given access to stored health data. Nevertheless, confidentiality and availability need to be carefully balanced when structuring this critical security system. Despite the fact that all patients' health data are made available to be exchanged, shared, and monitored to provide robust health care services, certain aspects of the data may be considered confidential and, for security reasons, must be kept restricted or inaccessible. These goals should be rationalized to ensure that patients receive the best possible care.

Strengths and Limitations

This study involved a comprehensive systematic literature search and identification of relevant and recent articles on mHealth apps' privacy, security, and confidentiality published in the last 8 years. Detailed information on patients' perspectives and awareness of the privacy and security of mHealth apps were gleaned from this review, thus bridging the current research gap as no systematic review has been conducted on this topic. Thematic and empirical analyses were also conducted to obtain robust data from the various designs used in the studies and triangulate the findings.

Nevertheless, the limitations of this study are well acknowledged. The literature search was restricted to 3 databases; thus, some important articles on the research topic might have been missed. Only patients and end users were considered as the primary targets in this review, whereas the perspectives of other active stakeholders in mHealth apps, such as HCPs, app developers, and policy makers, were not documented. The issues raised in this study might be better understood if all relevant stakeholders were considered. This also limits the generalizability of the findings as no inference could be made regarding health care personnel and mHealth app developers. The use of 1 reviewer for data extraction is also an important limitation; however, the reviewer was trained on how to perform the data extraction and coding before the study to ensure that the process was reproducible and consistent.

Meanwhile, a general limitation of the retrieved articles is the need for a clear definition of data privacy and security. Most of the reviewed studies considered security and privacy as a single concept, particularly as part of a general assessment of mHealth app design. Although security and privacy may overlap when ensuring patients' confidentiality, the 2 concepts are fundamentally different. In addition, only 3% (1/33) of the studies assessed the relationship between patients' sociodemographic factors and their concerns regarding mHealth apps' privacy, security, and confidentiality. A more robust assessment of patients' demographic characteristics, environmental factors, and patients' antecedents regarding data breaches and leakage to unwanted third parties requires further investigation.

Conclusions

This systematic review elucidated patients' perspectives and awareness regarding mHealth apps' privacy, security, and confidentiality. Patients showed diverse perspectives on the trio of concepts, ranging from users who were satisfied with the privacy and security features of their current mHealth apps to those who raised pertinent issues affecting technology use. Patients also conveyed specific approaches to improve the use of mHealth apps (facilitators), such as protection of personal data, ensuring the confidentiality of health status or medical conditions, and provision of brief training or education on data security and privacy.

The aggregation of the empirical and thematic results reflects that these diverse perspectives might be linked to the awareness of the subject matter, which also differed across the studies and was influenced by patients' sociodemographic characteristics, such as age, income level, and marital status, as well as their experience with mHealth apps, satisfaction levels, data type, and data stage. Thus, the findings of this review may be beneficial to mHealth app developers and other stakeholders in improving patients' awareness and adjusting current privacy and security features to enhance the use and adoption of mHealth apps for routine health monitoring and management of specific health conditions.

Acknowledgments

The authors appreciate the support provided by the Institute of Digital Healthcare, University of Warwick, Coventry, United Kingdom, in developing this manuscript. This study did not receive any funding support. The author is grateful for the support extended by Warwick University in Coventry, United Kingdom, and Saudi Electronic University in Jeddah, Saudi Arabia, in the development of this manuscript.

XSL•FO
RenderX

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request

Authors' Contributions

GE, TA, and NA contributed to conceptulization and project administration, whereas NA and MA performed the data curation and formal analysis. GE, TA, and NA were involved in the methodological design and supervision. The original draft was developed by NA while GE, TA, and NA were involved in writing and reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 21 KB - jmir_v26i1e50715_app1.docx]

References

1. Chen J, Lieffers J, Bauman A, Hanning R, Allman-Farinelli M. The use of smartphone health apps and other mobile health (mHealth) technologies in dietetic practice: a three country study. J Hum Nutr Diet 2017 Aug;30(4):439-452. [doi: 10.1111/jhn.12446] [Medline: 28116773]
2. Hussain M, Zaidan AA, Zidan BB, Iqbal S, Ahmed MM, Albahri OS, et al. Conceptual framework for the security of mobile health applications on Android platform. Telemat Inform 2018 Aug;35(5):1335-1354. [doi: 10.1016/j.tele.2018.03.005]
3. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284 [FREE Full text] [doi: 10.2196/19284] [Medline: 32501804]
4. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study. JMIR Mhealth Uhealth 2019 Apr 16;7(4):e11223 [FREE Full text] [doi: 10.2196/11223] [Medline: 30990458]
5. Kao CK, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM R 2017 May;9(5S):S106-S115. [doi: 10.1016/j.pmrj.2017.02.018] [Medline: 28527495]
6. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
7. Bani Issa W, Al Akour I, Ibrahim A, Almarzouqi A, Abbas S, Hisham F, et al. Privacy, confidentiality, security and patient safety concerns about electronic health records. Int Nurs Rev 2020 Jun 21;67(2):218-230. [doi: 10.1111/inr.12585] [Medline: 32314398]
8. Martin K. Understanding privacy online: development of a social contract approach to privacy. J Bus Ethics 2015 Feb 19;137(3):551-569. [doi: 10.1007/s10551-015-2565-9]
9. Bhuyan SS, Kim H, Isehunwa OO, Kumar N, Bhatt J, Wyant DK, et al. Privacy and security issues in mobile health: current research and future directions. Health Policy Technol 2017 Jun;6(2):188-191. [doi: 10.1016/j.hlpt.2017.01.004]
10. Sampat BH, Prabhakar B. Privacy risks and security threats in mHealth apps. J Int Technol Inf Manag 2017 Dec 01;26(4):126-153. [doi: 10.58729/1941-6679.1353]
11. Esposito C, De Santis A, Tortora G, Chang H, Choo KK. Blockchain: a panacea for healthcare cloud-based data security and privacy? IEEE Cloud Comput 2018 Jan;5(1):31-37. [doi: 10.1109/MCC.2018.011791712]
12. Fernández-Alemán JL, Señor IC, Lozoya P, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013 Jun;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
13. Nurgalieva L, O'Callaghan D, Doherty G. Security and privacy of mHealth applications: a scoping review. IEEE Access 2020;8:104247-104268. [doi: 10.1109/access.2020.2999934]
14. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Privacy and security in mobile health apps: a review and recommendations. J Med Syst 2015 Jan;39(1):181 [FREE Full text] [doi: 10.1007/s10916-014-0181-3] [Medline: 25486895]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009 Jun;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
17. Alwashmi MF, Fitzpatrick B, Farrell J, Gamble JM, Davis E, Nguyen HV, et al. Perceptions of patients regarding mobile health interventions for the management of chronic obstructive pulmonary disease: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 23;8(7):e17409 [FREE Full text] [doi: 10.2196/17409] [Medline: 32706697]

XSL•FO
RenderX

18. Moodley J, Constant D, Botha MH, van der Merwe FH, Edwards A, Momberg M. Exploring the feasibility of using mobile phones to improve the management of clients with cervical cancer precursor lesions. BMC Womens Health 2019 Jan 07;19(1):2 [FREE Full text] [doi: 10.1186/s12905-018-0702-1] [Medline: 30616579]
19. Bradbury A, Patrick-Miller L, Harris D, Stevens E, Egleston B, Smith K, et al. Utilizing remote real-time videoconferencing to expand access to cancer genetic services in community practices: a multicenter feasibility study. J Med Internet Res 2016 Feb 01;18(2):e23 [FREE Full text] [doi: 10.2196/jmir.4564] [Medline: 26831751]
20. Sangers TE, Wakkee M, Kramer-Noels EC, Nijsten T, Lugtenberg M. Views on mobile health apps for skin cancer screening in the general population: an in-depth qualitative exploration of perceived barriers and facilitators. Br J Dermatol 2021 Nov 05;185(5):961-969 [FREE Full text] [doi: 10.1111/bjd.20441] [Medline: 33959945]
21. Gill R, Ogilvie G, Norman WV, Fitzsimmons B, Maher C, Renner R. Feasibility and acceptability of a mobile technology intervention to support postabortion care in British Columbia: phase I. J Med Internet Res 2019 May 29;21(5):e13387 [FREE Full text] [doi: 10.2196/13387] [Medline: 31144668]
22. de Vries ST, Wong L, Sutcliffe A, Houÿez F, Ruiz CL, Mol PG. Factors influencing the use of a mobile app for reporting adverse drug reactions and receiving safety information: a qualitative study. Drug Saf 2017 May 29;40(5):443-455 [FREE Full text] [doi: 10.1007/s40264-016-0494-x] [Medline: 28035492]
23. Hackett KM, Kazemi M, Sellen DW. Keeping secrets in the cloud: mobile phones, data security and privacy within the context of pregnancy and childbirth in Tanzania. Soc Sci Med 2018 Aug;211:190-197 [FREE Full text] [doi: 10.1016/j.socscimed.2018.06.014] [Medline: 29960170]
24. Li J, Silvera-Tawil D, Varnfield M, Hussain MS, Math V. Users' perceptions toward mHealth technologies for health and well-being monitoring in pregnancy care: qualitative interview study. JMIR Form Res 2021 Dec 02;5(12):e28628 [FREE Full text] [doi: 10.2196/28628] [Medline: 34860665]
25. Wyatt KD, Finley A, Uribe R, Pallagi P, Willaert B, Ommen S, et al. Patients' experiences and attitudes of using a secure mobile phone app for medical photography: qualitative survey study. J Med Internet Res 2020 May 12;22(5):e14412 [FREE Full text] [doi: 10.2196/14412] [Medline: 32396127]
26. Natsiavas P, Kakalou C, Votis K, Tzovaras D, Koutkias V. Citizen perspectives on cross-border eHealth data exchange: a European survey. Stud Health Technol Inform 2019 Aug 21;264:719-723. [doi: 10.3233/SHTI190317] [Medline: 31438018]
27. Alaiad A, Alsharo M, Alnsour Y. The determinants of m-health adoption in developing countries: an empirical investigation. Appl Clin Inform 2019 Oct 30;10(5):820-840 [FREE Full text] [doi: 10.1055/s-0039-1697906] [Medline: 31667819]
28. Özkan Ö, Son YA, Aydinoğlu AU. Security and privacy concerns regarding genetic data in mobile health record systems: an empirical study from Turkey. bioRxiv Preprint posted online June 23, 2019. [doi: 10.1101/678912]
29. Glauser G, Ali ZS, Gardiner D, Ramayya AG, Pessoa R, Grady MS, et al. Assessing the utility of an IoS application in the perioperative care of spine surgery patients: the NeuroPath Pilot study. Mhealth 2019;5:40 [FREE Full text] [doi: 10.21037/mhealth.2019.09.01] [Medline: 31620467]
30. Zhou L, Parmanto B, Alfikri Z, Bao J. A mobile app for assisting users to make informed selections in security settings for protecting personal health data: development and feasibility study. JMIR Mhealth Uhealth 2018 Dec 11;6(12):e11210 [FREE Full text] [doi: 10.2196/11210] [Medline: 30538088]
31. Barutçu S, Barutçu E, Adigüzel DU. A technology acceptance analysis for mHealth apps: the case of Turkey. Balkan Near Eastern J Soc Sci 2018;4(4):104-113.
32. Bauer AM, Iles-Shih M, Ghomi RH, Rue T, Grover T, Kincler N, et al. Acceptability of mHealth augmentation of collaborative care: a mixed methods pilot study. Gen Hosp Psychiatry 2018;51:22-29 [FREE Full text] [doi: 10.1016/j.genhosppsych.2017.11.010] [Medline: 29272712]
33. Goetz M, Müller M, Matthies LM, Hansen J, Doster A, Szabo A, et al. Perceptions of patient engagement applications during pregnancy: a qualitative assessment of the patient's perspective. JMIR Mhealth Uhealth 2017 May 26;5(5):e73 [FREE Full text] [doi: 10.2196/mhealth.7040] [Medline: 28550005]
34. Richardson JE, Ancker JS. Public perspectives of mobile phones' effects on healthcare quality and medical data security and privacy: a 2-year nationwide survey. AMIA Annu Symp Proc 2015;2015:1076-1082 [FREE Full text] [Medline: 26958246]
35. Sanger P, Hartzler A, Lober WB, Evans HL, Pratt W. Design considerations for post-acute care mHealth: patient perspectives. AMIA Annu Symp Proc 2014;2014:1920-1929 [FREE Full text] [Medline: 25954465]
36. Dang Y, Guo S, Guo X, Wang M, Xie K. Privacy concerns about health information disclosure in mobile health: questionnaire study investigating the moderation effect of social support. JMIR Mhealth Uhealth 2021 Feb 08;9(2):e19594 [FREE Full text] [doi: 10.2196/19594] [Medline: 33555266]
37. Biswas M, Tania MH, Kaiser MS, Kabir R, Mahmud M, Kemal AA. ACCU3RATE: a mobile health application rating scale based on user reviews. PLoS One 2021 Dec 16;16(12):e0258050 [FREE Full text] [doi: 10.1371/journal.pone.0258050] [Medline: 34914718]
38. Casilang CG, Stonbraker S, Japa I, Halpern M, Messina L, Steenhoff AP, et al. Perceptions and attitudes toward mobile health in development of an exclusive breastfeeding tool: focus group study with caregivers and health promoters in the Dominican Republic. JMIR Pediatr Parent 2020 Aug 21;3(2):e20312 [FREE Full text] [doi: 10.2196/20312] [Medline: 32821063]

XSL•FO
RenderX

39. Zhang D, Lim J, Zhou L, Dahl AA. Breaking the data value-privacy paradox in mobile mental health systems through user-centered privacy protection: a web-based survey study. JMIR Ment Health 2021 Dec 24;8(12):e31633 [FREE Full text] [doi: 10.2196/31633] [Medline: 34951604]
40. Harris B, Ajisola M, Alam RM, Watkins JA, Arvanitis TN, Bakibinga P, et al. Mobile consulting as an option for delivering healthcare services in low-resource settings in low- and middle-income countries: a mixed-methods study. Digit Health 2021 Aug 19;7:20552076211033425 [FREE Full text] [doi: 10.1177/20552076211033425] [Medline: 34777849]
41. Al-Anezi FM. Evaluating the readiness of mobile technology with respect to e-heath for medication in Saudi Arabia: an integrative perspective. J Multidiscip Healthc 2021 Jan;Volume 14:59-66 [FREE Full text] [doi: 10.2147/jmdh.s287321]
42. Ermakova T, Fabian B, Zarnekow R. Improving individual acceptance of health clouds through confidentiality assurance. Appl Clin Inform 2017 Dec 18;07(04):983-993 [FREE Full text] [doi: 10.4338/aci-2016-07-ra-0107]
43. Rodrigues R, Poongulali S, Balaji K, Atkins S, Ashorn P, De Costa A. 'The phone reminder is important, but will others get to know about my illness?' Patient perceptions of an mHealth antiretroviral treatment support intervention in the HIVIND trial in South India. BMJ Open 2015 Nov 02;5(11):e007574 [FREE Full text] [doi: 10.1136/bmjopen-2015-007574] [Medline: 26525717]
44. Al-Mahrouqi T, Al-Alawi K, Al-Alawi M, Al Balushi N, Al Ghailani A, Al Sabti H, et al. A promising future for tele-mental health in Oman: a qualitative exploration of clients and therapists' experiences. SAGE Open Med 2022 Mar 29;10:20503121221086372 [FREE Full text] [doi: 10.1177/20503121221086372] [Medline: 35371483]
45. Turcotte V, Chagnon A, Guénette L. Experience and perspectives of users and non-users of the Ask your pharmacist teleconsultation platform. Explor Res Clin Soc Pharm 2021 Jun;2:100031 [FREE Full text] [doi: 10.1016/j.rcsop.2021.100031] [Medline: 35481130]
46. Morton E, Nicholas J, Lapadat L, O'Brien HL, Barnes SJ, Poh C, et al. Use of smartphone apps in bipolar disorder: an international web-based survey of feature preferences and privacy concerns. J Affect Disord 2021 Dec 01;295:1102-1109. [doi: 10.1016/j.jad.2021.08.132] [Medline: 34706421]
47. Cavazos-Rehg P, Min C, Fitzsimmons-Craft EE, Savoy B, Kaiser N, Riordan R, et al. Parental consent: a potential barrier for underage teens' participation in an mHealth mental health intervention. Internet Interv 2020 Sep;21:100328 [FREE Full text] [doi: 10.1016/j.invent.2020.100328] [Medline: 32528858]
48. Hattingh HL, Knox K, Fejzic J, McConnell D, Fowler JL, Mey A, et al. Privacy and confidentiality: perspectives of mental health consumers and carers in pharmacy settings. Int J Pharm Pract 2015 Feb;23(1):52-60 [FREE Full text] [doi: 10.1111/ijpp.12114] [Medline: 24761999]
49. Perera C. Principles of security for the use of mobile technology in medicine. J Mobile Technol Med 2012 Jun 01;2:5-7. [doi: 10.7309/jmtm.10]
50. Lewis TL, Wyatt JC. mHealth and mobile medical apps: a framework to assess risk and promote safer use. J Med Internet Res 2014 Sep 15;16(9):e210 [FREE Full text] [doi: 10.2196/jmir.3133] [Medline: 25223398]
51. Atienza AA, Zarcadoolas C, Vaughon W, Hughes P, Patel V, Chou WY, et al. Consumer attitudes and perceptions on mHealth privacy and security: findings from a mixed-methods study. J Health Commun 2015;20(6):673-679. [doi: 10.1080/10810730.2015.1018560] [Medline: 25868685]
52. Nissenbaum H. Privacy as contexual integrity. Wash Law Rev 2004;79(1):119.

Abbreviations

GDPR: General Data Protection Regulation
HCP: health care professional
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach, T Leung, T de Azevedo Cardoso; submitted 10.07.23; peer-reviewed by K Wyatt, H Abu Serhan, B Salam; comments to author 02.08.23; revised version received 03.09.23; accepted 25.01.24; published 31.05.24.

Please cite as:
Alhammad N, Alajlani M, Abd-alrazaq A, Epiphaniou G, Arvanitis T
Patients' Perspectives on the Data Confidentiality, Privacy, and Security of mHealth Apps: Systematic Review
J Med Internet Res 2024;26:e50715
URL: https://www.jmir.org/2024/1/e50715
doi:10.2196/50715
PMID:38820572

XSL•FO
RenderX

©Nasser Alhammad, Mohannad Alajlani, Alaa Abd-alrazaq, Gregory Epiphaniou, Theodoros Arvanitis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Just-in-Time Adaptive Intervention for Stabilizing Sleep Hours of Japanese Workers: Microrandomized Trial

Hiroki Takeuchi[1], PhD; Tetsuro Ishizawa[2,3], MD, PhD; Akifumi Kishi[2], PhD; Toru Nakamura[4], PhD; Kazuhiro Yoshiuchi[2], MD, PhD; Yoshiharu Yamamoto[1], PhD

[1]Graduate School of Education, The University of Tokyo, Tokyo, Japan
[2]Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[3]Central Medical Support Co, Tokyo, Japan
[4]Institute for Datability Science, Osaka University, Osaka, Japan

Corresponding Author:
Hiroki Takeuchi, PhD
Graduate School of Education
The University of Tokyo
Bunkyo-ku Hongo 7-3-1
Tokyo, 113-8654
Japan
Phone: 81 03 5841 3981
Email: takeuchi@p.u-tokyo.ac.jp

Abstract

Background: Sleep disturbance is a major contributor to future health and occupational issues. Mobile health can provide interventions that address adverse health behaviors for individuals in a vulnerable health state in real-world settings (just-in-time adaptive intervention).

Objective: This study aims to identify a subpopulation with vulnerable sleep state in daily life (study 1) and, immediately afterward, to test whether providing mobile health intervention improved habitual sleep behaviors and psychological wellness in real-world settings by conducting a microrandomized trial (study 2).

Methods: Japanese workers (n=182) were instructed to collect data on their habitual sleep behaviors and momentary symptoms (including depressive mood, anxiety, and subjective sleep quality) using digital devices in a real-world setting. In study 1, we calculated intraindividual mean and variability of sleep hours, midpoint of sleep, and sleep efficiency to characterize their habitual sleep behaviors. In study 2, we designed and conducted a sleep just-in-time adaptive intervention, which delivered objective push-type sleep feedback messages to improve their sleep hours for a subset of participants in study 1 (n=81). The feedback messages were generated based on their sleep data measured on previous nights and were randomly sent to participants with a 50% chance for each day (microrandomization).

Results: In study 1, we applied hierarchical clustering to dichotomize the population into 2 clusters (group A and group B) and found that group B was characterized by unstable habitual sleep behaviors (large intraindividual variabilities). In addition, linear mixed-effect models showed that the interindividual variability of sleep hours was significantly associated with depressive mood (β=3.83; P=.004), anxiety (β=5.70; P=.03), and subjective sleep quality (β=−3.37; P=.03). In study 2, we found that providing sleep feedback prolonged subsequent sleep hours (increasing up to 40 min; P=.01), and this effect lasted for up to 7 days. Overall, the stability of sleep hours in study 2 was significantly improved among participants in group B compared with the participants in study 1 (P=.001).

Conclusions: This is the first study to demonstrate that providing sleep feedback can benefit the modification of habitual sleep behaviors in a microrandomized trial. The findings of this study encourage the use of digitalized health intervention that uses real-time health monitoring and personalized feedback.

(J Med Internet Res 2024;26:e49669) doi:10.2196/49669

KEYWORDS

objective push-type sleep feedback; stability of habitual sleep behaviors; just-in-time adaptive intervention; microrandomized trial; mobile phone

XSL•FO
RenderX

Introduction

Importance and Necessity of Addressing Vulnerable Sleep States

Sleep plays a key role in preventing future adverse health outcomes, such as coronary heart disease [1,2], type 2 diabetes [2-4], and Alzheimer disease [5,6]. In addition to being relevant to chronic diseases, sleep disturbance is associated with impaired psychological wellness, including psychiatric disorders [7,8] and disruption of daytime moods [9,10], and affects work productivity (presenteeism) [11-13] and cognitive function [14,15]. In 2015, the National Sleep Foundation formulated age-specific sleep duration recommendations to prevent future health problems and recommended 7 to 9 hours of sleep for young adults and adults (aged 18-64 years) [16]. However, a recent meta-analysis of 36 epidemiological studies revealed that approximately 1 in 4 people (24.5%) slept less than the recommended amount for their age groups [17]. This suggests the difficulty of self-management for habitual sleep behaviors and the necessity for support in daily life.

Mobile Health

Mobile health (mHealth) [18,19] is a medical health practice that uses digital devices, including smartphone apps and wearable devices. Major mHealth methodologies include collecting health-related data in real-world settings (real-world data), conducting interventions for adverse health behaviors using digital devices, and generating evidence from the collected real-world data (real-world evidence). Just-in-time adaptive intervention (JITAI) is an appealing mHealth application that supports the modification of health behaviors for individuals in a vulnerable health state [20-22]. JITAI can alleviate current health problems and prevent future disease onset by making individuals aware of their daily health behaviors through timely and adaptive interventions.

Recent Needs for Real-World Evidence

Recently, the methodology of mHealth has received increased attention, even in the general health care field, with growing interest in leveraging real-world data. This has been triggered by the establishment of the 21st Century Cures Act in the United States and the Food and Drug Administration's recommendation for using real-world evidence as well as evidence from gold standard clinical trials (randomized controlled trials) [23,24]. Real-world data are expected to become an essential source for monitoring an individual's health status outside the clinic and evaluating the effectiveness of clinical and health trials. However, this trend calls for novel approaches to generate real-world evidence for the effectiveness of treatments conducted in real-world settings [25]. Microrandomization, or microrandomized trial (MRT), is a trial design used to generate real-world evidence for intervention options in mHealth trials, including JITAI [26,27]. In MRT, a specific intervention option is randomly provided for each individual via their digital device, allowing for the examination of the within-individual proximal effect by comparing subsequent behaviors with and without the intervention. For instance, Klasnja et al [28] randomly provided healthy adults with contextually tailored suggestions that encouraged physical activity during a 6-week trial period. By comparing the subsequent physical activity with and without suggestions, they revealed that providing suggestion messages proximally increased subsequent step counts and that this effect attenuated by day.

mHealth Sleep Intervention

Advancements in mobile and wearable sensing technologies enable the objective measurement of habitual sleep behaviors across multiple days [29-31]. Thus, night-to-night sleep dynamics (eg, stability and transition in sleep data) can also be captured in a real-world setting. Despite the rapid growth in the use of real-world sleep data, a prior systematic review indicated that there are few mHealth trials leveraging objective sleep data recorded by wearable devices [32]. Hence, incorporating wearable-based sleep monitoring into mHealth trials for controlling habitual sleep behaviors is challenging. Furthermore, using push-type feedback based on the objective sleep monitoring may promote the improvement of habitual sleep behaviors. For example, according to the control theory on health behavior change [33], which has been applied to numerous health intervention trials [34-36], individuals monitor their current health state and change their behavior to reduce the discrepancy against the ideal state or set point. However, individuals may not be able to accurately recognize their current sleep state in daily lives without external feedback. Therefore, providing objective push-type sleep feedback may assist self-regulation of habitual sleep behaviors through enhancing the awareness for their current sleep state. Recently, we conducted a formative mHealth trial that provided push-type feedback messages based on objective sleep monitoring for Japanese workers [37]. The stabilized habitual sleep behavior and improved physical symptoms upon waking up were observed in the workers who received sleep feedback every day, suggesting the utility of wearable-based sleep monitoring and objective push-type sleep feedback. As the next step, we aimed to achieve sleep JITAI by identifying individuals in a vulnerable sleep state and designing appropriate intervention options to alleviate their sleep problems. Moreover, by implementing an MRT, it should be examined whether individuals immediately change their sleep behavior in response to the sleep feedback (proximal effect).

Aim of This Paper

In study 1, we aimed to monitor the sleep of Japanese workers to identify a subpopulation that had problems in their habitual sleep behaviors. Considering the close link between sleep and psychological wellness, we measured momentary symptoms using a smartphone-based ecological momentary assessment (EMA) during the study. In study 2, we aimed to conduct an MRT to test whether providing sleep feedback messages improved subsequent sleep behaviors and psychological wellness proximally in real-world settings. In addition to the within-individual proximal effects, we aimed to examine the distal effects of the sleep JITAI using a pretest-posttest comparison design, which may provide a comprehensive understanding of the effectiveness of this trial.

XSL•FO
RenderX

Methods

Study 1

Study Design

Using a field data sampling method, we conducted an observational study of Japanese workers to reveal the sleep demographics of the current Japanese society and identify a subpopulation that needed support to improve their habitual sleep behaviors. Habitual sleep behaviors were objectively measured using a wearable device. Momentary symptoms were recorded on a smartphone app in real time to examine the association between psychological wellness and habitual sleep behavior with ecological validity. Baseline data, including demographic characteristics, pretrial psychological symptoms, and chronotype, were also measured before the survey.

Health Care Internet of Things System

We developed a cloud-based health care Internet of Things (HIT) system that can continuously acquire health-related information, including momentary symptoms, biological signals, and surrounding environmental information, that are recorded as parts of daily living [37,38]. The HIT system consists of a cloud server and a smartphone app for data collection (HIT server and HIT app, respectively). The HIT app is equipped with an EMA, and users can record momentary symptoms in daily life. In addition, the HIT app can connect with various digital devices, including the proprietary activity monitor used in this study (refer to the Activity Monitor section), using Bluetooth Low Energy (BLE). The data were transferred from the devices to the HIT server. The HIT server can store, integrate, and manage data uploaded from the app and send personalized messages (push-type feedback messages) to HIT app users (Multimedia Appendix 1).

Participants

Recruitment was conducted from January 17 to January 31, 2022. Author TI sent digital flyers to health managers of the contracted companies. Subsequently, these health managers forwarded the flyers to their employees. Workers who agreed to participate in the study were asked to access and complete the web-based registration form via QR code provided on the flyers. Inclusion criteria involved participants that are Japanese speakers, aged >18 years, and have a personal smartphone. Workers who had difficulties or dangers while using their smartphones during the day, such as above-ground workers and long-distance drivers, were excluded.

As we needed to identify individuals in a vulnerable sleep state and commence study 2 immediately after completing study 1, it was necessary to minimize variations in survey dates across participants. Hence, the sample size for study 1 was set to the number of workers who responded to the web-based registration form within the recruitment period.

Baseline Questionnaire

Before the survey, the participants completed a baseline questionnaire to assess their demographic characteristics, psychological symptoms, and chronotype. The items included in the questionnaire are listed in the following sections.

Demographic Characteristics

Participants were asked to indicate their age, sex, BMI, clinical history, and employment status (full-time employee, contingent worker, part-time worker, or other) as their demographic information. On the basis of the occupational classification table published by the Ministry of Health, Labour, and Welfare in Japan [39], they were asked to select their occupation from 12 categories: managerial, technical, sales, service, security, agricultural, productive, transportation, machinery, construction, cleaning, and others.

Psychological Symptoms

The pretrial psychological symptoms included the Japanese version of the Beck Depression Inventory second edition (BDI-II) [40], the State-Trait Anxiety Inventory-JZY (STAI) [41], and the Positive and Negative Affect Schedule (PANAS) [42,43]. The BDI-II is a 21-item self-report inventory used to measure the presence and severity of depression (score range 0-63). The BDI-II classifies individuals into 4 categories based on their overall score: minimal or no depression, 0 to 13; mild depression, 14 to 19; moderate depression, 20 to 28; and severe depression, 29 to 63. A score of ≥14 points was used as the clinical cutoff point for depression. The 20-item STAI is a standardized self-report inventory for measuring state and trait anxiety. The state anxiety subscale measures the intensity of the anxiety felt by an individual in the present, whereas the trait anxiety subscale measures how often an individual feels anxious. Scores range from 20 to 80 for each subscale, with higher scores indicating higher anxiety levels. The PANAS is a self-report questionnaire that assesses emotional well-being and measures positive and negative affect with 8 items. Higher scores indicate that individuals show positive and negative affect in daily life. In the Japanese version, the Cronbach α were 0.83 and 0.82 for positive and negative affect, respectively [43].

Chronotype

Chronotype, which represents an individual's preferred timing for engaging in various daytime activities, was measured using the Japanese versions of the Morningness-Eveningness Questionnaire (MEQ) [44,45] and the Munich Chronotype Questionnaire (MCTQ) [46,47]. The MEQ consists of 19 items about morning or nighttime activities and classifies individuals into the morning, intermediate, and evening types based on the total score (score range 16-86). A higher score indicates that the individual prefers to engage in various activities in the morning (morning type). The MCTQ assumes that an individual's sleep timing on work-free days adheres to their endogenous circadian rhythm and accumulated sleep loss rather than their social schedule. Hence, this scale calculates the midpoint between sleep onset and offset on work-free days (midsleep point on free days). Subsequently, sleep-corrected midpoint of sleep on work-free days (MSFsc) was introduced to assess chronotype after correcting the effect of sleep loss accumulated during workdays from midsleep point on free days. The MEQ score and MSFsc have been used as proxies for individual circadian phases because they are associated with physiological biomarkers of the human endogenous circadian rhythm [47,48]. In addition to MSFsc, the MCTQ can be used to calculate the discrepancy between an individual's circadian

XSL•FO
RenderX

phase and actual sleep timing (social jetlag). The absolute value (absolute social jetlag) is often used in chronobiological and sleep studies to measure this discrepancy.

EMA Questionnaire

EMA is a method for recording participants' behavior, psychological state, and physical symptoms in real time and at multiple time points, allowing the collection of self-report and objective data with reliability and ecological validity [49,50]. Using the HIT app, participants answered an EMA questionnaire 5 times daily to assess momentary symptom data (eg, depressive mood, anxiety, and subjective sleep quality) in real time. Depressive mood and anxiety were scored using the Depression and Anxiety Mood Scale [51]. This scale comprises the following 9 adjectives representing mood states: "vigorous," "gloomy," "concerned," "happy," "unpleasant," "anxious," "cheerful," "depressed," and "worried." On the basis of these 9 items, anxious (the average of "concerned," "anxious," and "worried" scores); positive (the average of "vigorous," "happy," and "cheerful" scores); and negative (the average of "gloomy," "unpleasant," and "depressed" scores) moods were calculated. Depressive mood scores were obtained by combining the last 2 mood scores as follows: score = (100 − [positive mood score + negative mood score])/2. In addition, participants were asked to rate their subjective sleep quality when they woke up: "How was your sleep quality last night?" These measurements were rated using a visual analog scale from 0 to 100 displayed on the screen. All scores were transferred to the HIT server immediately after completing each EMA questionnaire.

Activity Monitor

We used a wristband activity monitor (Sciencenet device; Sciencenet Inc) to measure participants' habitual sleep behaviors. This device is equipped with triaxial piezoelectric accelerometers that measure 1-minute zero-crossing count data, which counts the number of times per epoch where the acceleration signal level crosses 0. The Cole-Kripke algorithm [52] with the rescoring rules by Webster et al [53] was applied to the data to identify whether the 1-minute epoch was sleep or wake. Subsequently, we estimated a square wave that maximized the coefficient of determination with the Cole-Kripke identifications [37]. The length and midpoint of the interval were defined as the sleep hours and midpoint of sleep of the day, respectively. Furthermore, sleep efficiency was evaluated by dividing the sleep hours by the duration that was classified as sleep. Details of this procedure can be found in our previous work [37]. We confirmed that the device performed at a level equivalent to the research-grade actigraphy (Ambulatory Monitors, Inc), which is widely used in clinical settings. Results of the comparative analysis are presented in Multimedia Appendix 2.

Study Protocol

We mailed the activity monitor, survey progression guide, informed consent form, and consent withdrawal form to the workers who responded to the web-based registration form. The BDI-II and STAI were also sent to the participants because they should not be replicated. The use of the device and the HIT app was explained in a tutorial video. Participants were instructed to complete other questionnaires on the web, whereas the BDI-II and STAI were answered on paper. After completing the informed consent form, they commenced the observational study.

During the 14-day study period, we instructed them to complete EMA questionnaires at randomly selected times within −10 minutes to +10 minutes of predetermined times (11 AM, 3 PM, and 7 PM) by sending a reminder message on their HIT app. In addition, they were asked to voluntarily complete the EMA when they woke up and went to bed. Furthermore, they were instructed to wear a wristband activity monitor on their nondominant wrist to measure their habitual sleep behaviors, except while bathing, showering, performing rigorous exercises, or engaging in any other activities likely to damage the device.

After the study, participants returned the survey instruments and written informed consent form to the University of Tokyo and received remuneration and a personalized sleep report. This sleep report included summarized information about their sleep (including sleep hours, midpoint of sleep, and sleep efficiency) and the sleep hygiene guidelines proposed by the Ministry of Health, Labour and Welfare in Japan [54]. The survey dates varied among participants, and data collection was completed on February 24, 2022.

Statistical Analysis

We conducted complete case analysis after excluding the data of the participants with <4 days of sleep events recorded. All analyses were performed using R (version 4.0.2; **R** Foundation for Statistical Computing). Statistical significance was defined as $P<.05$. We evaluated the intraindividual mean (IIM) and intraindividual variability (IIV) of measured sleep hours, midpoint of sleep, and sleep efficiency ($IIM^{(SH)}$, $IIM^{(MID)}$, $IIM^{(SE)}$, $IIV^{(SH)}$, $IIV^{(MID)}$, and $IIV^{(SE)}$, respectively). IIM and IIV were quantified by calculating the average and SD for each individual. Circular mean and SD were calculated for each individual using the "circular" package when evaluating $IIM^{(MID)}$ and $IIV^{(MID)}$. We adopted a hierarchical clustering algorithm (the Ward method in 6-dimensional Euclidean space) to the intraindividual statistics using the "hclust" function after standardizing these intraindividual statistics. We dichotomized the population based on the dendrogram (group A and group B), postulating that the overall sample comprised individuals in healthy and vulnerable sleep state. Then, we performed 2-tailed Welch *t* tests for intraindividual statistics between the groups to capture their characteristics of habitual sleep behaviors. When analyzing the IIV statistics, they were log transformed to satisfy the normality assumption. The same analysis was performed on the MEQ score, MSFsc, and absolute social jetlag as additional variables representing their sleep-wake schedules. Furthermore, the effect sizes (Hedges *g*) were calculated using the "effsize" package. We performed linear mixed-effect models for EMA-recorded depressive mood, anxiety, and subjective sleep quality using the "lmerTest" package to examine which aspects of habitual sleep behaviors were associated with momentary symptoms. The 95% CIs of the parameters were estimated by using the "confint" function.

XSL•FO
RenderX

Study 2

Study Design

As shown in the Study 1 section we classified the workers into 2 clusters (groups A and B) and found that group B was characterized by unstable habitual sleep behaviors. On the basis of the results, in study 2, we designed and conducted a sleep JITAI for stabilizing their habitual sleep hours. Similar to study 1, habitual sleep behaviors were objectively measured using a wearable device, and momentary symptoms were assessed using a smartphone app in real time. On the basis of the measured sleep data, we sent objective push-type sleep feedback to the smartphone app with a 50% chance, enabling us to compare the subsequent sleep behaviors with and without the feedback (to examine the within-individual and proximal effects). Subjective sleep quality and emotional well-being were assessed using baseline and follow-up questionnaires to examine the whole effect (distal effect) of the trial. In particular, we tested whether the effect of the sleep JITAI was more pronounced for individuals in vulnerable state (group B) than those in healthy state (group A).

Participants

After completing study 1, we invited a subset of the participants in study 1 to study 2. Due to budgetary constraints, the maximum sample size of study 2 was set to 100. Therefore, we asked 67 workers in group B and 33 workers in group A to participate in the sleep JITAI. The workers in group A were randomly selected by using the "sample" function in R (version 4.0.2) regardless of age and sex. Recruitment was conducted from March 11 to March 14, 2022, by sending digital flyers with a QR code that links to a web-based registration form. As in study 1, those who agreed to participate were asked to complete the web-based form.

Baseline and Follow-Up Questionnaire

Only the possible outcomes for the sleep JITAI were assessed because their demographic characteristics were already assessed in study 1. The Japanese versions of the Pittsburgh Sleep Quality Index (PSQI) [55,56] and the PANAS were assessed before and after the survey. The PSQI is a self-report inventory used to assess sleep quality in the preceding month. The PSQI consists of 19 items on self-reported sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. Scores range from 0 to 21, with a higher total score indicating poorer sleep quality. The strong reliability and validity of this questionnaire have been confirmed in a recent meta-analysis [57]. The explanation for the PANAS is described in the Psychological Symptoms section.

Intervention Protocol

Most protocols were similar to those used in study 1. We mailed the activity monitor, survey progression guide, informed consent form, and consent withdrawal form to the workers who responded to the web-based registration form. The participants were asked to complete the PSQI and PANAS as baseline questionnaires on the web. The participants commenced the survey after completing the baseline questionnaire and informed consent form.

During the 14-day survey period, participants were asked to complete the EMA questionnaires 5 times daily and wear a wristband activity monitor on their nondominant wrist, except while bathing, showering, performing rigorous exercises, or engaging in any other activities likely to damage the device. In this survey, at noon, the physical activity data uploaded to the HIT server were collected and analyzed to estimate the sleep hours for each individual using a local data analysis server. On the basis of the estimated sleep data, relative sleep sufficiency was calculated (difference between the $IIM^{(SH)}$ in study 1 and estimated sleep hour), and personalized sleep feedback messages were generated and sent to the participants with a probability of 0.50 each day of the 14-day study; hence, the total number of messages that a participant could have received ranged between 0 to 14, theoretically. The messages informed the participants about their relative sleep sufficiency and requested them to plan and adjust their daytime activities, with reference to the sleep hygiene guidelines provided in study 1. The message reads: "You slept XX minutes longer (shorter) than your average yesterday. Stabilize your sleep habits with reference to sleep hygiene guidelines." We designed the feedback sentences to adaptively change the meanings depending on the substituted sleep hours. We expected this message to act as negative feedback when sleep hours were relatively long and as an alert when sleep hours were relatively short. If sufficient physical activity data (<720 records/day) had not been uploaded to the server by noon, an alternative message was sent to the participant requesting them to confirm the BLE pairing of the activity monitor with their app: "It seems that your data have not been uploaded successfully. We will analyze the data again at 2 PM. Please check the BLE connection of your HIT app before then." The same computation process was executed at 2 PM for the relevant participants. These processes were executed by local servers and were fully automated.

After the survey, the participants returned the survey instruments and written informed consent form to the University of Tokyo and received the personalized sleep report as in study 1. Finally, participants were requested to answer a web-based follow-up questionnaire, including the PSQI and PANAS. The survey dates varied among participants, and data collection was completed on May 18, 2022.

Statistical Analysis

Overview

We used linear mixed-effect models and hierarchical Bayesian models to exploratorily examine the within-individual proximal effects of the intervention. Mixed-effect ANOVA models were also performed to evaluate the distal effect. We conducted complete case analysis after excluding the data of the participants with <4 days of sleep events recorded. All analyses were performed using R (version 4.0.2; **R** Foundation for Statistical Computing). Statistical significance was defined as $P<.05$ or when the 95% credible interval did not include the null value.

Liner Mixed-Effect Model

We used the linear mixed-effect model to examine the within-individual proximal effect of the feedback message on

XSL•FO
RenderX

subsequent sleep hours using the "lmerTest" package. The 95% CIs of the parameters were estimated by using the "confint" function. With reference to a model proposed in a previous study [58], this model incorporated an indicator of whether feedback was provided or not (feedback), an index representing the elapsed day from the beginning of the survey (day), and their interaction effect (feedback×day). Control variables, including sleep hours before providing feedback, age, and sex, were added to the model. Similar models were used for each group as a subgroup analysis to consider group heterogeneity in the effectiveness of the feedback. Furthermore, we stratified the feedback messages based on whether they indicated that the individual slept longer or shorter than the average for study 1 and tested whether the effects of the feedback differed depending on its content using a similar model. To verify the statistical power of the effect of the feedback, we performed ex-post power analysis in Multimedia Appendix 3.

Hierarchical Bayesian Model

Using hierarchical Bayesian modeling, we performed counterfactual simulations to examine (1) the number of days the feedback message was effective for subsequent sleep and (2) whether the changes in sleep hours were linked to improvements in psychological wellness (depressive mood, anxiety, and subjective sleep quality). In particular, we estimated the cascading effect of feedback on momentary symptoms in the next morning through changed sleep hours. Detailed model equations are provided in Multimedia Appendix 4. The model parameters and their 95% credible intervals were computed using the "rstan" package. The number of chains was set to 4 when performing the Markov chain Monte Carlo sampling. All iterations and burn-in samples were set to 8000 and 4000, respectively. Convergence across the 4 chains was defined as the point at which the value of the Gelman-Rubin statistic (Rhat) [59] for all parameters was <1.10. The model was run after controlling for age and sex.

Mixed-Effect ANOVA Model

Individual sleep statistics (IIM and IIV) were analyzed using 2 studies (studies 1 and 2)×2 groups (groups A and B) mixed-effect ANOVA to examine the distal effect of the intervention. When analyzing the IIV statistics, they were log transformed. The PSQI and PANAS scores, before and after the survey, were also analyzed using 2 time points (before and after)×2 groups (groups A and B) mixed-effect ANOVA. These ANOVA models were run after controlling for age and sex using the "lmerTest" package. The "lsmeans" package was used to perform the pairwise comparison tests with Tukey correction.

Ethical Considerations

This nested study, including studies 1 and 2, underwent an organized single ethics review.

Ethics Approval

The Ethics Committee of the University of Tokyo approved these studies and the informed consent forms (approval number 21-353).

Informed Consent

Participants were explained the aim and procedure of this study through the mailed survey progression guide. If participants had any questions on the studies, they could contact author HT via email. When declining participation, they were asked to return the written consent withdrawal form. In addition, they could withdraw their consent by emailing a scanned consent withdrawal form to author HT.

Privacy and Confidentiality

When analyzing the data, an ID was assigned for each participant, and personal information was excluded. The correspondence table between individuals and IDs is stored by author HT, and the other authors could not access participants' personal information.

Remuneration

In study 1, all participants who did not withdraw their consent were paid with vouchers worth at least ¥5000 (US $1=¥116). The renumeration amount was increased depending on the response rate for their EMA questionnaires. Participants whose response rate was ≥80% received vouchers worth ¥10,000.

In study 2, all participants were paid with a voucher worth ¥10,000 when starting the survey. Participants were able to withdraw their consent even after receiving remuneration.

Results

Study 1

Participant Sampling

In total, 202 workers completed the web-based registration form, but 20 (10%) withdrew from participation before writing the consent form because of problems in installing the HIT app and unexpected work schedules (Figure S1 in Multimedia Appendix 5). Hence, 182 (90%) participants started the survey to measure their habitual sleep data and record momentary symptoms in daily life (Figure 1). Among them, 4 (2%) dropped out during the survey because of discomfort with the survey instruments. A total of 16 (9%) participants who did not return their consent forms were excluded from the analysis. Furthermore, 22 (12%) participants were excluded, as they failed to measure the sleep data, with <4 days of sleep events recorded. Finally, data from 140 (77%) participants were used for subsequent analysis. The mean age of the participants was 39.15 (SD 10.09) years, and 27.7% (29/140) were women. Table S1 in Multimedia Appendix 6 presents other baseline data and sleep variables recorded by the wearable devices.

XSL•FO
RenderX

Figure 1. Overview of study 1. During the 14-day study period, participants wore a wristband activity monitor on their nondominant wrist to measure their sleep data and recorded momentary symptoms using smartphone-based ecological momentary assessment (EMA). Depressive mood and anxiety were assessed 5 times daily, and subjective sleep quality was assessed only when they woke up. "Taking off" refers to the period that the individual took off the device.

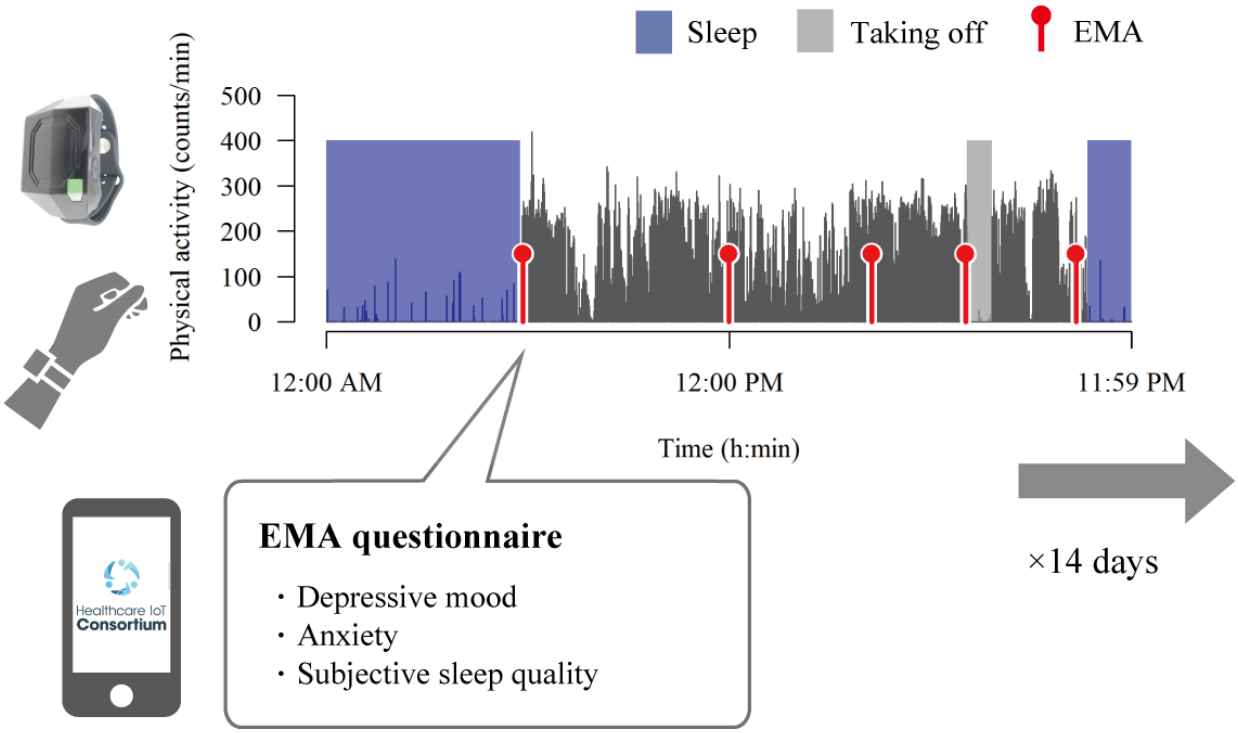

Data-Driven Approach for Identifying Sleep Problems

On the basis of the measured sleep data (1347 nights in total), we calculated the intraindividual statistics characterizing their habitual sleep behaviors (ie, ($IIM^{(SH)}$, $IIM^{(MID)}$, $IIM^{(SE)}$, $IIV^{(SH)}$, $IIV^{(MID)}$, and $IIV^{(SE)}$). Subsequently, we adopted a hierarchical clustering algorithm and divided the participants into 2 clusters (Figure 2). One group (group A) consisted of 73 participants, including 55 (75%) men and 18 (25%) women, with a mean age of 39.26 (SD 9.56) years. The other group (group B) consisted of 67 participants, including 56 (84%) men and 11 (16%) women, with a mean age of 39.03 (SD 10.71) years. When performing 2-tailed Welch *t* tests, the participants in group B showed not only delayed $IIM^{(MID)}$ (P=.01) and lower $IIM^{(SE)}$ (P<.001) but also higher $IIV^{(SH)}$ (P<.001), $IIV^{(MID)}$ (P<.001), and $IIV^{(SE)}$ (P<.001) than those in group A (Figure 3 and Table S2 in Multimedia Appendix 6). Furthermore, large effect sizes (Hedges g >0.8) were observed in the differences in $IIM^{(SE)}$ and all the IIV statistics, implying that the participants in group B were characterized especially by the IIV aspects of their sleep data (sleep instability) rather than their IIM aspects. In addition, MEQ (P<.001), MSFsc (P<.001), and absolute social jetlag (P=.004) also significantly differed between the groups, with moderate effect sizes (Hedges g=|0.51-0.65|; Table S2 in Multimedia Appendix 6). These results suggest the necessity of an mHealth intervention to stabilize habitual sleep behaviors especially for participants in group B.

XSL•FO
RenderX

Figure 2. Summary of feature extraction and clustering procedure. Over the study period, objective sleep data (sleep hour, midpoint of sleep, and sleep efficiency) were collected repeatedly (up to 14 nights/person). Participants' habitual sleep behaviors were characterized by evaluating the intraindividual mean and intraindividual variability of the measurements. After standardizing them, we performed hierarchical clustering algorithm with Ward's method in Euclidean space to identify a subpopulation with sleep problems in daily life.

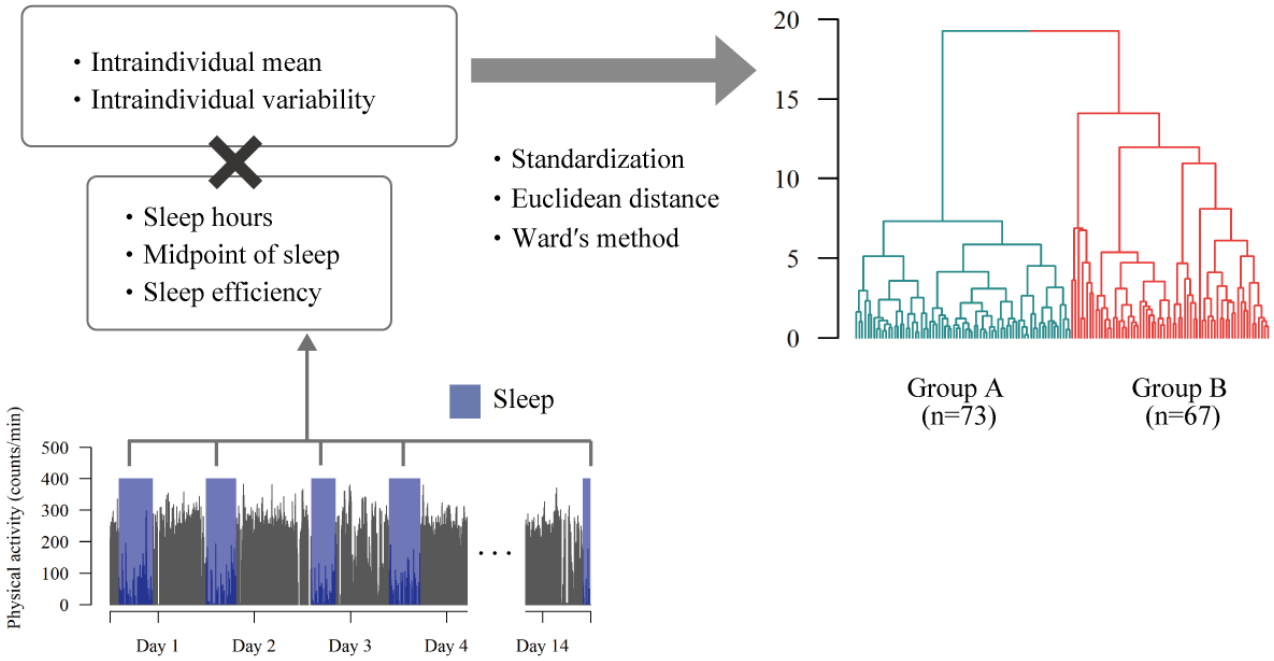

Figure 3. Group comparisons of intraindividual sleep statistics. Group comparison of the intraindividual mean (A-C) and intraindividual variability of sleep hours, midpoint of sleep, and sleep efficiency (D-F). Hedges g was calculated when significant group difference (*P*<.05) was observed. **P*<.05; ***P*<.01.

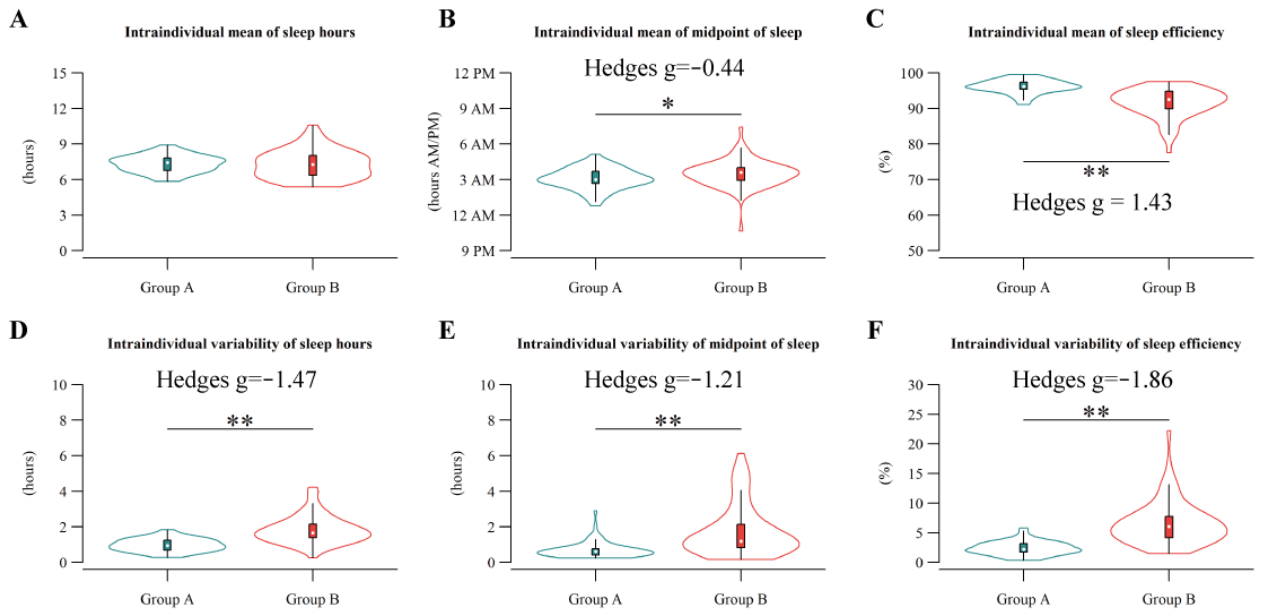

Associations Between IIV and Momentary Symptoms

The EMA procedures yielded 9349 records on depressed mood and anxiety, as well as 1896 records on subjective sleep quality. We performed linear mixed-effect models to identify the aspects of IIV associated with these momentary symptoms. The models indicated a significant association between $IIV^{(SH)}$ and these symptoms (depressive mood, *P*=.004; anxiety, *P*=.03; subjective sleep quality, *P*=.03; Figure 4 and Table S3 in Multimedia Appendix 6), implying that improving $IIV^{(SH)}$ might benefit psychological wellness.

XSL•FO
RenderX

Figure 4. Associations between intraindividual variability of sleep hours and momentary symptoms. (A: depressive mood; B: anxiety; and C: subjective sleep quality) using linear mixed-effect models. The points and error bars represent the individual means and SEs of the momentary symptoms, respectively. The red dotted lines in the figure indicate the regression lines of intraindividual variability of sleep hours.

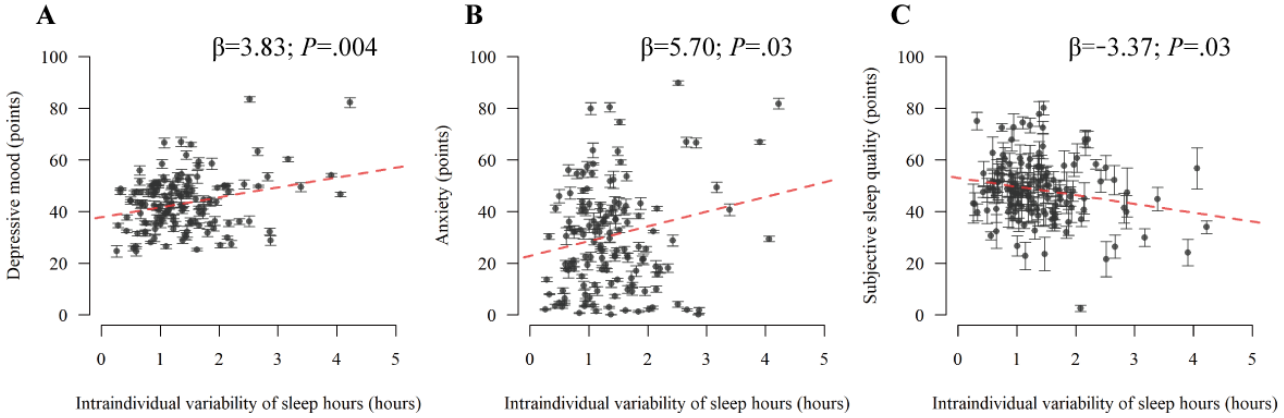

Study 2

Participant Sampling

We invited 100 workers who had completed study 1, of whom 83 (83%) agreed to participate in study 2 (Figure S2 in Multimedia Appendix 5). Among them, 2 (2%) workers withdrew before writing the consent form because of unexpected schedules. Finally, the sleep JITAI trial (Figure 5) included 81 workers. We excluded all data from 14 (17%) participants from the subsequent analysis because 9 (64%) did not complete the follow-up questionnaire and 5 (36%) had failed to measure sleep behaviors. Table S4 in Multimedia Appendix 6 presents the demographic characteristics of participants whose data were used in the following analyses (n=67; group A: n=23, 34%; group B: n=44, 66%). The 2-tailed Welch *t* test and chi-squared test revealed no substantial differences in demographic characteristics between the groups (Multimedia Appendix 6).

Figure 5. Overview of study 2. As in study 1, participants were instructed to collect sleep data and momentary symptoms using digital devices. In study 2, sleep feedback messages were generated based on their sleep data and provided at random with a 50% chance for each participant over the study period. This message was expected to act as negative feedback when sleep hours were relatively long and as an alert when sleep hours were relatively short. We could examine the within-individual proximal effect of the intervention by comparing these data with and without the feedback. "Taking off" refers to the period that the individual took off the device. EMA: ecological momentary assessment.

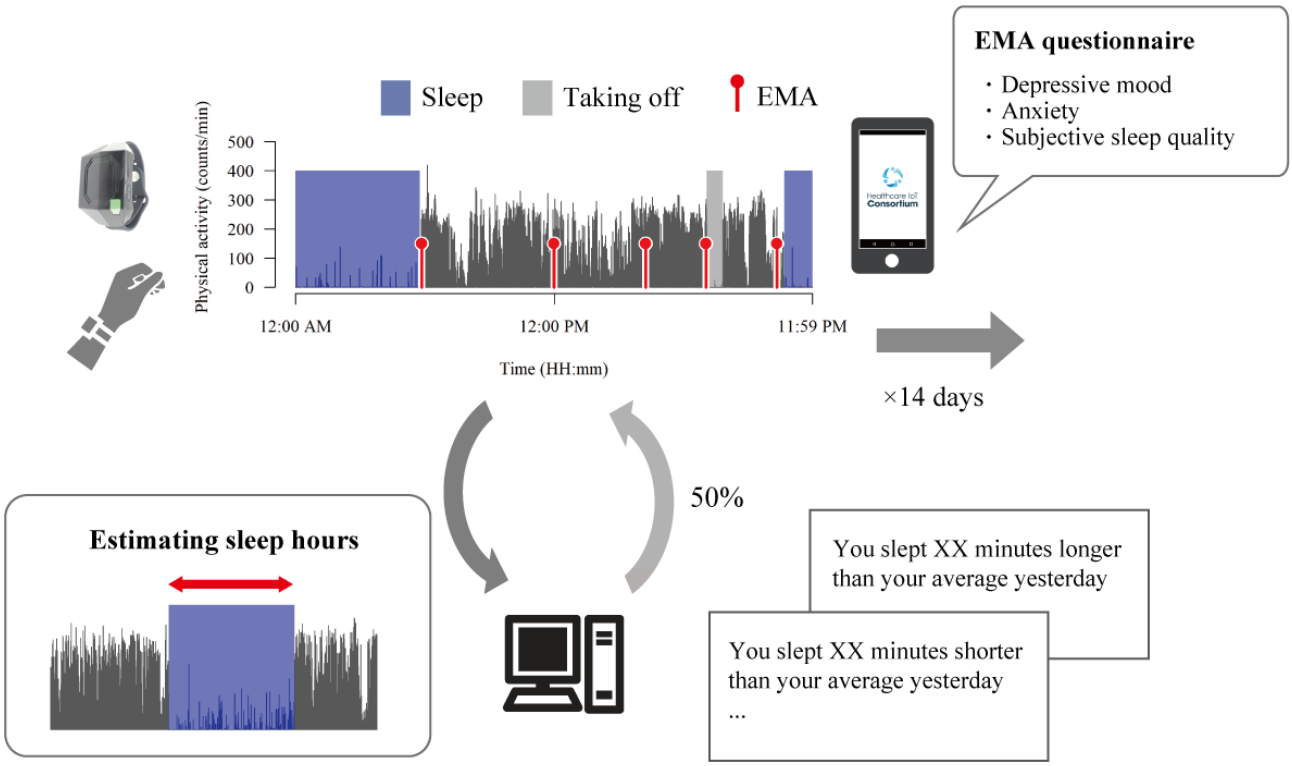

Within-Individual Effect of Feedback Messages

We examined the within-individual effect of feedback messages on subsequent sleep hours using a linear mixed-effect model (494 nights in total; Table 1). The model revealed that providing a feedback message significantly prolonged subsequent sleep hours by 40 minutes compared with not providing feedback messages on the first day of the study (*P*=.01). However, a significant interaction was observed between feedback and day

XSL•FO
RenderX

(*P*=.04). Thus, the effect of feedback messages decreased by 4 minutes per day over the study period. In addition, subgroup analysis (Table 2) revealed that the effect of the feedback message was only significant in group B (an increase of 59 minutes; *P*=.007), suggesting the potential heterogeneity of effectiveness between the groups. Furthermore, when stratified by cases in which the individual slept longer or shorter than the average in study 1, sleep hours were significantly prolonged after receiving the feedback message, indicating that the previous sleep was shorter than the average (an increase of 1 hour 8 minutes; *P*=.01; Table S5 in Multimedia Appendix 6).

Table 1. Within-individual proximal effect of the intervention on subsequent sleep hour.

	Linear mixed-effect model	
	β coefficient (SE; 95% CI)	*P* value
Intercept (no feedback)	6 h 47 min (13 min; 6 h 22 min to 7 h 13 min)	<.001
Age	−2 (1; −4 to −1) min	.001
Sex	35 min (24 min; −11 min to 1 h 20 min)	.15
Sleep hours	2 (3; −3 to 8) min	.46
Feedback	40 min (16 min; 9 min to 1 h 11 min)	.01
Day	2 (1; 0 to 5) min	.09
Feedback×day	−4 (2; −9 to 0 min)	.04

Table 2. Subgroup analysis of the within-individual proximal effect.

	Group A		Group B	
	β coefficient (SE; 95% CI)	*P* value	β coefficient (SE; 95% CI)	*P* value
Intercept (no feedback)	7 h 20 min (17 min; 6 h 41 min to 7 h 44 min)	<.001	6 h 32 min (18 min; 6 h 2 min to 7 h 11 min)	<.001
Age	−3 (1; −5 to −1) min	.005	−3 (1; −5 to −1) min	.02
Sex	8 (22; −33 to 49) min	.72	1 h 4 min (46 min; −25 min to 2 h 33 min)	.17
Sleep hours	−3 (5; −12 to 7) min	.60	3 (4; −3 to 10) min	.35
Feedback	−1 (21; −41 to 39) min	.96	59 min (22 min; 17 min to 1 h 42 min)	.007
Day	1 (2; −2 to 5) min	.44	3 (2; −1 to 7) min	.21
Feedback×day	−2 (3; −7 to 4) min	.57	−6 (3; −11 to 0) min	.07

Time-Variant Cascading Effect of Feedback Messages on Momentary Symptoms

After performing the Markov chain Monte Carlo sampling, we confirmed that the Gelman-Rubin statistic values for all parameters were <1.10, indicating successful convergence of the posterior distributions of the parameters (Table S6 in Multimedia Appendix 6). On the basis of their point estimates (expected a posteriori) and 95% credible intervals, the model suggested that providing feedback messages substantially prolonged subsequent sleep hours during the first 7 days (Figure 6). Furthermore, it was shown that providing feedback improved depressive mood and subjective sleep quality in the following wake-up time during the first 7 and 4 days, respectively. As sleep feedback was randomly provided, it should be noted that participants did not receive it daily. Table S6 in Multimedia Appendix 6 shows all point estimates and corresponding 95% credible intervals.

XSL•FO
RenderX

Figure 6. Time-variant effect of the feedback message on subsequent sleep hours and the cascading effect on depressive mood, anxiety, and subjective sleep quality the next morning among group A (A, B, C, and D, respectively) and group B (E, F, G, and H, respectively). The points represent the expected a posteriori of the coefficient, and the error bars indicate the corresponding 95% credible intervals. The asterisk (*) indicates that the 95% credible interval of the estimate did not include the null value.

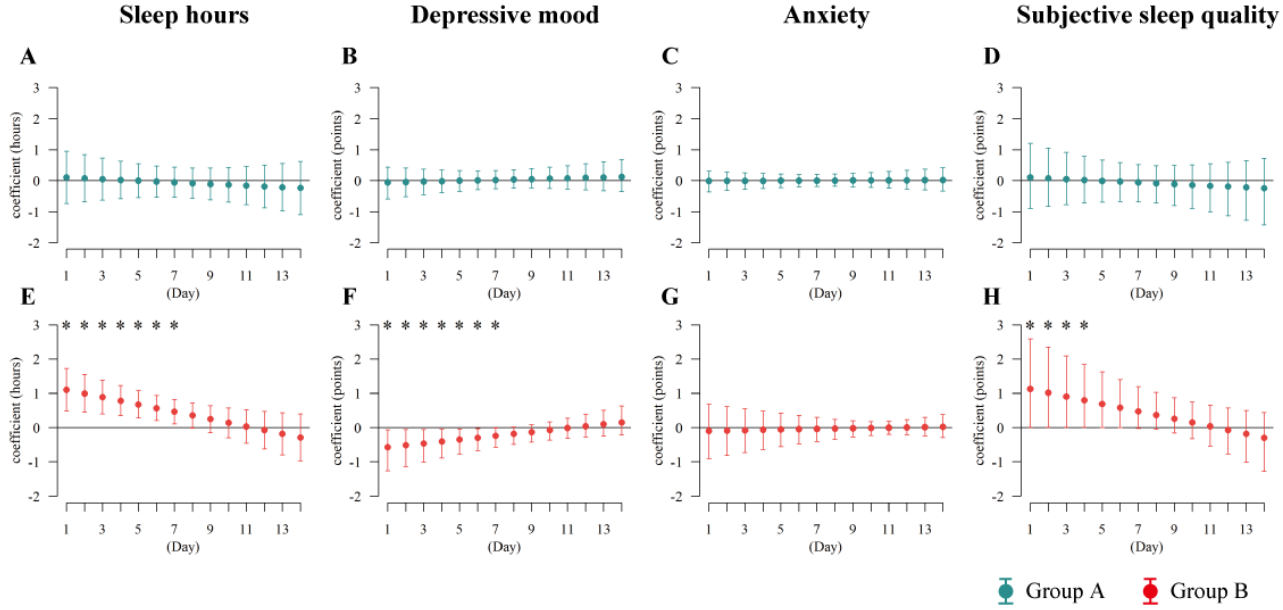

Comparing Sleep Statistics Between Studies 1 and 2

Mixed-effect ANOVA models were performed to examine the distal effect of the trial on their sleep statistics (Tables S7 and S8 in Multimedia Appendix 6). The models revealed a significant interaction effect between them for $IIV^{(SH)}$ ($F_{1,65}$=10.06; P=.002) and $IIV^{(SE)}$ ($F_{1,65}$=6.86; P=.01). Pairwise comparison tests with Tukey correction demonstrated that $IIV^{(SH)}$ significantly decreased only in group B (Figure 7A; P=.001), while significant differences in $IIV^{(SE)}$ were not observed in both groups (group A, P=.06; group B, P=.07). In addition, the models showed a significant main effect of the study on $IIM^{(SH)}$ ($F_{1,65}$=6.07; P=.02), and pairwise comparison tests indicated that $IIM^{(SH)}$ was significantly shortened during the trial (study 2) compared with study 1 in the overall sample (P=.03). Furthermore, a significant main effect in the groups was observed in $IIM^{(MID)}$ ($F_{1,65}$=4.33; P=.04), $IIM^{(SE)}$ ($F_{1,65}$=15.31; P<.001), $IIV^{(SH)}$ ($F_{1,65}$=17.02; P<.001), $IIV^{(MID)}$ ($F_{1,65}$=6.83; P=.01), and $IIV^{(SE)}$ ($F_{1,65}$=30.53; P<.001), although this was evident from study 1 findings.

Figure 7. Distal effects on intraindividual variability of sleep hours and the Pittsburgh Sleep Quality Index (PSQI) total score. (A) Changes in the intraindividual variability of sleep hours between studies 1 and 2, (B) changes in the PSQI total score before and after the trial.

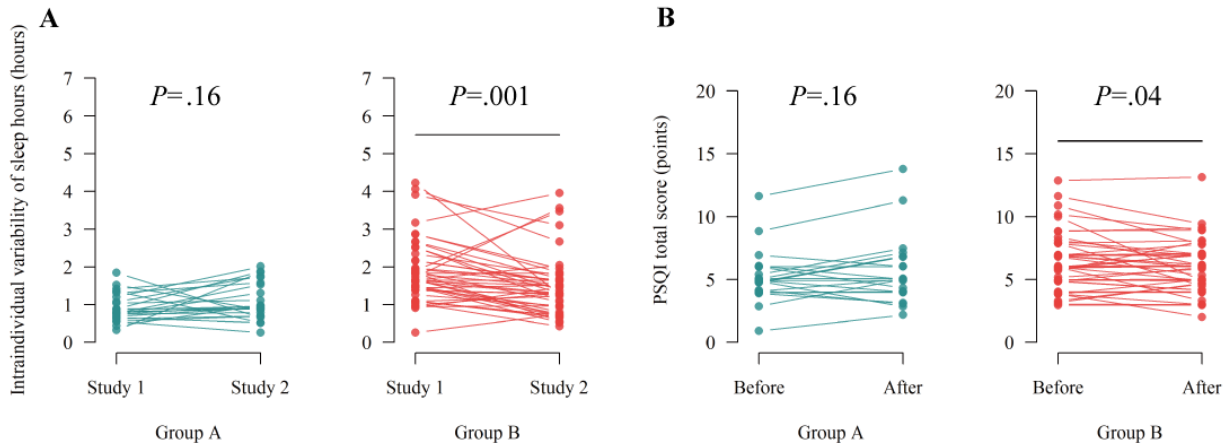

Pretest-Posttest Comparison of PSQI and PANAS Scores

The PSQI total and PANAS scores (positive and negative affect) were assessed before and after the trial, and Table S9 in Multimedia Appendix 6 shows the scores. The mixed-effect ANOVA models (Table S10 in Multimedia Appendix 6) showed a significant main effect of time on PANAS positive ($F_{1,65}$=4.60; P=.04) and negative affect ($F_{1,65}$=4.11; P=.046) and a significant interaction between time and group in the PSQI total score ($F_{1,65}$=5.60; P=.02). The pairwise comparison test revealed that the PSQI total score significantly improved in group B (Figure 7B; P=.04), while no significant differences were observed in the PANAS scores (positive affect, P=.14; negative affect, P=.08).

XSL•FO
RenderX

Discussion

Overview

This study aimed to identify a subpopulation with sleep problems in their daily lives using real-world objective sleep data (study 1) and to test whether the sleep feedback messages proximally and distally improved habitual sleep behaviors (study 2). In addition to sleep behaviors, the cascading effect on psychological wellness was examined using EMA recordings, considering the close link between sleep problems and poor psychological wellness [9,10]. Appropriate sleep patterns are essential for promoting the quality of life of workers; several studies have reported an association between sleep problems and adverse occupational phenomena, such as presenteeism [11-13], burnout syndrome [60], low back pain [60,61], and errors at work [62]. Furthermore, a report by the RAND Corporation estimated that sleep problems cost up to US $680 billion annually in 5 Organisation for Economic Co-operation and Development countries (Canada, Germany, Japan, the United States, and the United Kingdom) [63]. The economic loss in Japan was estimated to be approximately US $148 billion annually (equivalent to 2.92% of its GDP), the highest among the 5 countries. Sleep JITAI may solve these occupational and economic problems and disease onset [1-8] by encouraging daily self-management of habitual sleep behaviors.

Principal Findings: Study 1

In study 1, a hierarchical clustering algorithm was adapted to analyze multidimensional sleep data and identify a subpopulation characterized by larger IIV in sleep data (sleep instability). Recently, the stability of habitual sleep behavior has been considered an additional factor in preventing future health problems [64,65] along with IIMs. Previous studies have indicated that the instability of sleep hours is associated with psychological wellness (including depressive symptoms, anxiety, subjective sleep quality, and psychological well-being) [65-67]. The linear mixed-effect models of this study showed that unstable sleep hours were associated with worse depressive mood, anxiety, and subjective sleep quality in daily life. This result is consistent with the recent findings of the studies that used the EMA method [66], augmenting real-world evidence for this association. Hence, group B, characterized by unstable sleep behaviors, can be regarded as a vulnerable subpopulation that needs sleep management. Along with the instability, individuals in group B showed a delayed sleep-wake schedule (later sleep timing and evening chronotype), larger social jetlag, and lower sleep efficiency, corroborating the necessity for the sleep JITAI for this subpopulation. Further studies are required to elucidate the role of individual circadian rhythms in habitual sleep behavior in real-world settings; however, it has been speculated that their unstable sleep-wake schedule and circadian misalignment caused fragmented sleep during the night, as supported by findings from experimental studies [68,69].

Principal Findings: Study 2

In study 2, by using MRT, we first examined whether the participants changed their habitual sleep behaviors in response to the sleep feedback. The linear mixed-effect model revealed that sleep hours were prolonged after receiving objective push-type sleep feedback, particularly in group B. On the basis of the control theory [33], the individuals are postulated to change their health behavior to reduce the discrepancy against the set point value. In other words, they might have rather reduced sleep hours after being informed that they slept longer than their set point values (ie, the $IIM^{(SH)}$ in study 1). However, the participants in group B changed sleep hours only after receiving the alerting feedback, indicating that they slept shorter than their set point value. Thus, they might have interpreted the messages indicating the relatively long sleep hours as encouraging messages and adopted strategies maintaining their sleep behaviors. In future, it is necessary to investigate how workers change their sleep hours, when altered set point values are used in feedback messages (eg, the recommended sleep hours of the National Sleep Foundation: 7-9 hours) [16]. Moreover, throughout the analyses, we found no evidence suggesting the effectiveness of the sleep feedback for those in group A. The reduced necessity for sleep intervention might result in the null effect because the participants in group A had relatively healthy sleep habits in study 1. Previous MRT studies also suggested that mHealth interventions were more effective, especially when the individual showed adverse health behaviors [70]. Thus, interventions are likely most effective when just in time. In this study, we reproduced similar findings at the interindividual level. In other words, the findings of this study demonstrate the importance of identifying vulnerable individuals before the trial, as in study 1.

We also discovered that the effect of feedback weakened over time, and the hierarchical Bayesian model predicted that the effect was no longer considerable after 7 days. This attenuation has also been reported in prior MRT studies and is attributed to the habituation of the trial [28]. Personalizing intervention options may reduce habituation risk and improve overall efficacy [28,71]. Therefore, a method to infer the intraindividual behavioral factors leading to better sleep should be developed, and an algorithm to automatically optimize the intervention options during the trial period should be implemented. For example, some studies have proposed adapting reinforcement learning algorithms in mHealth trials to sophisticate and personalize intervention policies over the trial period [72,73]. Incorporating artificial intelligence into mHealth (artificial intelligence–powered mHealth) [74] is a possible future direction for sustainable health management as a part of daily life. The hierarchical Bayesian model also predicted that providing feedback improved depressive mood and subjective sleep quality through changes in sleep hours, at least during the first 4 days. These findings provide novel insights into the immediate psychological response to sleep improvement and encourage the statement that improving habitual sleep behavior contributes to psychological wellness. However, the model quantified the indirect effects of the objective sleep feedback on the momentary symptoms of the next morning and implicitly assumed that providing the sleep feedback does not affect these symptoms independently of changes in sleep hours. Therefore, it is noted that the model should not be used to examine the effects of intervention options, such as relaxation and meditation, that may also directly affect psychological wellness.

XSL•FO
RenderX

Corresponding to the findings in the analyses of the proximal effect, whole effect analyses showed stabilized sleep hours and improved PSQI total scores after study 2, especially in group B. Thus, in this study, we demonstrated the distal effects of sleep JITAI on sleep stability and subjective sleep quality. Distal outcomes are usually achieved through the accumulation of the proximal effect and are the goals of mHealth trials [21,22,26]. However, despite their importance, previous MRT studies did not investigate distal outcomes [28,70]. Future MRTs should incorporate protocols to assess distal outcomes, such as pretest-posttest comparison, to provide a comprehensive understanding of the effectiveness of the MRT. Contrary to expectations, study 2 had decreased $IIM^{(SH)}$ compared with study 1. As this effect was globally observed across the groups, it may be irrelevant to the interventions and could be caused by the influence of social factors, such as changes in work schedules or remote work frequency due to the spread of the COVID-19 pandemic. Future studies should measure data representing lifestyle to analyze the influence of social schedules on sleep. The $IIV^{(SE)}$ showed an interaction effect in the ANOVA model, suggesting that the response pattern differed by group. However, pairwise comparison tests did not show substantial changes for each group. Similarly, the main effects of time on PANAS scores (positive and negative affect) were significant, while post hoc tests did not detect substantial changes. Therefore, we must be conservative when demonstrating the distal effect of the intervention on the stability of sleep efficiency and emotional well-being, providing opportunities for future mHealth trials to clarify these points.

Strengths and Future Directions

To our knowledge, a previous MRT study focused on sleep hours, physical activity, and mood among medical interns, who were assumed to be affected by a stressful environment [70]. Their 6-month study examined the effects of the intervention and these moderators, suggesting that an individual's current health state influenced receptivity to interventions. However, in this study, we performed MRT to generate real-world evidence for the effectiveness of the intervention options rather than exploring the factors to consider when constructing a specific JITAI. On the basis of real-world sleep data, we identified individuals in a vulnerable sleep state and specified their sleep problems (sleep instability) before the trial. Furthermore, we used $IIM^{(SH)}$ measured in study 1 to personalize the set point value of the habitual sleep hours and elaborated feedback messages to stabilize their habitual sleep hours. Through these procedures, we made the mHealth intervention just in time and adaptive.

The intervention effects may diminish over time; however, our results indicate that providing push-type sleep feedback can modify sleep instability in the short term. Furthermore, it was suggested that interventions are unnecessary after the successful stabilization of sleep behavior and until sleep instability recurs. As a possible future direction, we can update this sleep JITAI to be more suitable for daily life by continuously tracking individuals' sleep stability and providing treatments only when their sleep becomes unstable. Such a timely intervention strategy may be widely applicable to address various health behaviors in daily life. Moreover, an integrative health care information system, such as the HIT system, that enables the multifaceted and real-time monitoring of health-related data may be a technical basis for implementing JITAI.

Limitations

This study had several limitations. First, considering the scalability of mHealth technologies, the sample size of this study was not large. In fact, a previous mHealth sleep intervention trial had included >1000 poor sleepers (patients with insomnia) [32,75]. However, in this study, the data used for the analyses of the proximal effect of the MRT were repeatedly measured by digital devices. These augmented data could contribute to enhancing the statistical power. In fact, ex-post power analysis (Multimedia Appendix 3) showed that the results of the proximal effect among group B had a 77.2% of observed statistical power, which is close to the typical 80% threshold. Second, the trial period was 2 weeks, which is relatively shorter than that of previous mHealth trials (average duration of 6-7 weeks) [32]. How long the participants would continue to use the device if the trial lasted longer is unknown. Consumer-grade wearable trackers are often abandoned after a few months [76], and daily hassles in data collection and unhelpful suggestions are frequently mentioned reasons [76,77]. We implemented push-type sleep feedback in this trial to minimize the burden of accessing sleep data; however, personalizing the intervention options can also address this issue. Finally, this study included workers who were healthy enough to engage in a typical work schedule in daily life. Therefore, the applicability of this intervention in clinical trials for sleep disorders (eg, insomnia and circadian rhythm sleep disorders) remains unclear. Further studies are warranted to determine whether a similar effect is observed in the clinical population. In addition, whether additive or synergistic effects are observed when incorporating JITAI methodologies into traditional pharmacological or psycho-behavioral sleep interventions should be investigated.

Conclusions

In conclusion, we provide real-world evidence that objective push-type sleep feedback proximally prolongs sleep hours and distally improves sleep stability and subjective sleep quality. Timely interventions powered by real-time health monitoring and personalized feedback may offer opportunities for individuals in vulnerable health states to manage their health behaviors in daily life. Furthermore, incorporating MRT into a traditional trial design can be helpful to comprehensively understand the effectiveness of the interventions and increase the sophistication of the trial design.

XSL•FO
RenderX

Acknowledgments

The authors thank the workers who participated in these studies. This study was supported in part by Japan Society for the Promotion of Science KAKENHI (20H00569), the Japan Science and Technology Agency Mirai Program (21473074), and the Japan Science and Technology Agency Moonshot Program (JPMJMS229B).

Data Availability

The data sets generated during and analyzed during this study are not publicly available because the participants of this study did not agree to their data being shared publicly, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization and methodology were developed by AK, TN, KY, and YY. HT and TI collected the data. HT performed data analysis and wrote the first draft of the manuscript. All authors critically reviewed and approved the final version of the manuscript.

Conflicts of Interest

TI is a delegate of the Central Medical Support Co, and the participants in this work were recruited from his contracted companies. All other authors declare no financial or nonfinancial competing interests.

Multimedia Appendix 1
Screenshots of the Health care Internet of Things app.
[DOC File , 446 KB - jmir_v26i1e49669_app1.doc]

Multimedia Appendix 2
Validity of the Sciencenet activity monitor.
[DOC File , 1336 KB - jmir_v26i1e49669_app2.doc]

Multimedia Appendix 3
Explanation of hierarchical Bayesian model.
[DOCX File , 18 KB - jmir_v26i1e49669_app3.docx]

Multimedia Appendix 4
Flow diagram of studies 1 and 2.
[PPTX File , 44 KB - jmir_v26i1e49669_app4.pptx]

Multimedia Appendix 5
Supplementary data.
[XLSX File (Microsoft Excel File), 45 KB - jmir_v26i1e49669_app5.xlsx]

Multimedia Appendix 6
The ex-post power analysis.
[DOCX File , 207 KB - jmir_v26i1e49669_app6.docx]

References

1. Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. Eur Heart J 2011 Jun;32(12):1484-1492. [doi: 10.1093/eurheartj/ehr007] [Medline: 21300732]
2. Itani O, Jike M, Watanabe N, Kaneita Y. Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression. Sleep Med 2017 Apr;32:246-256. [doi: 10.1016/j.sleep.2016.08.006] [Medline: 27743803]
3. Jike M, Itani O, Watanabe N, Buysse DJ, Kaneita Y. Long sleep duration and health outcomes: a systematic review, meta-analysis and meta-regression. Sleep Med Rev 2018 Jun;39:25-36 [FREE Full text] [doi: 10.1016/j.smrv.2017.06.011] [Medline: 28890167]
4. Anothaisintawee T, Reutrakul S, Van Cauter E, Thakkinstian A. Sleep disturbances compared to traditional risk factors for diabetes development: systematic review and meta-analysis. Sleep Med Rev 2016 Dec;30:11-24. [doi: 10.1016/j.smrv.2015.10.002] [Medline: 26687279]

XSL•FO
RenderX

5. Bubu OM, Brannick M, Mortimer J, Umasabor-Bubu O, Sebastião YV, Wen Y, et al. Sleep, cognitive impairment, and Alzheimer's disease: a systematic review and meta-analysis. Sleep 2017 Jan 01;40(1). [doi: 10.1093/sleep/zsw032] [Medline: 28364458]
6. Shi L, Chen SJ, Ma MY, Bao Y, Han Y, Wang Y, et al. Sleep disturbances increase the risk of dementia: a systematic review and meta-analysis. Sleep Med Rev 2018 Aug;40:4-16. [doi: 10.1016/j.smrv.2017.06.010] [Medline: 28890168]
7. Zhai L, Zhang H, Zhang D. Sleep duration and depression among adults: a meta-analysis of prospective studies. Depress Anxiety 2015 Sep 05;32(9):664-670. [doi: 10.1002/da.22386] [Medline: 26047492]
8. Scott J, Kallestad H, Vedaa O, Sivertsen B, Etain B. Sleep disturbances and first onset of major mental disorders in adolescence and early adulthood: a systematic review and meta-analysis. Sleep Med Rev 2021 Jun;57:101429. [doi: 10.1016/j.smrv.2021.101429] [Medline: 33549912]
9. Tomaso CC, Johnson AB, Nelson TD. The effect of sleep deprivation and restriction on mood, emotion, and emotion regulation: three meta-analyses in one. Sleep 2021 Jun 11;44(6):33367799 [FREE Full text] [doi: 10.1093/sleep/zsaa289] [Medline: 33367799]
10. Short MA, Booth SA, Omar O, Ostlundh L, Arora T. The relationship between sleep duration and mood in adolescents: a systematic review and meta-analysis. Sleep Med Rev 2020 Aug;52:101311. [doi: 10.1016/j.smrv.2020.101311] [Medline: 32240932]
11. Ishibashi Y, Shimura A. Association between work productivity and sleep health: a cross-sectional study in Japan. Sleep Health 2020 Jun;6(3):270-276. [doi: 10.1016/j.sleh.2020.02.016] [Medline: 32360021]
12. Takano Y, Ibata R, Nakano N, Sakano Y. Association between insomnia severity and presenteeism among Japanese daytime workers. J Sleep Res 2023 Feb 25;32(1):e13711. [doi: 10.1111/jsr.13711] [Medline: 36054731]
13. Takano Y, Ibata R, Nakano N, Sakano Y. Impact of sleep debt, social jetlag, and insomnia symptoms on presenteeism and psychological distress of workers in Japan: a cross-sectional study. Biopsychosoc Med 2022 Jun 03;16(1):13 [FREE Full text] [doi: 10.1186/s13030-022-00242-5] [Medline: 35658925]
14. Fortier-Brochu E, Beaulieu-Bonneau S, Ivers H, Morin CM. Insomnia and daytime cognitive performance: a meta-analysis. Sleep Med Rev 2012 Feb;16(1):83-94. [doi: 10.1016/j.smrv.2011.03.008] [Medline: 21636297]
15. Lowe CJ, Safati A, Hall PA. The neurocognitive consequences of sleep restriction: a meta-analytic review. Neurosci Biobehav Rev 2017 Sep;80:586-604. [doi: 10.1016/j.neubiorev.2017.07.010] [Medline: 28757454]
16. Hirshkowitz M, Whiton K, Albert SM, Alessi C, Bruni O, DonCarlos L, et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health 2015 Mar;1(1):40-43. [doi: 10.1016/j.sleh.2014.12.010] [Medline: 29073412]
17. Kocevska D, Lysen TS, Dotinga A, Koopman-Verhoeff ME, Luijk MP, Antypa N, et al. Sleep characteristics across the lifespan in 1.1 million people from the Netherlands, United Kingdom and United States: a systematic review and meta-analysis. Nat Hum Behav 2021 Jan 16;5(1):113-122. [doi: 10.1038/s41562-020-00965-x] [Medline: 33199855]
18. Becker S, Miron-Shatz T, Schumacher N, Krocza J, Diamantidis C, Albrecht UV. mHealth 2.0: experiences, possibilities, and perspectives. JMIR Mhealth Uhealth 2014 May 16;2(2):e24 [FREE Full text] [doi: 10.2196/mhealth.3328] [Medline: 25099752]
19. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
20. Spruijt-Metz D, Nilsen W. Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. IEEE Pervasive Comput 2014 Jul;13(3):13-17. [doi: 10.1109/MPRV.2014.46]
21. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219 [FREE Full text] [doi: 10.1037/hea0000306] [Medline: 26651462]
22. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
23. Khosla S, White R, Medina J, Ouwens M, Emmas C, Koder T, et al. Real world evidence (RWE) - a disruptive innovation or the quiet evolution of medical evidence generation? F1000Res 2018 Aug 29;7:111 [FREE Full text] [doi: 10.12688/f1000research.13585.2] [Medline: 30026923]
24. Milne-Ives M, van Velthoven MH, Meinert E. Mobile apps for real-world evidence in health care. J Am Med Inform Assoc 2020 Jun 01;27(6):976-980 [FREE Full text] [doi: 10.1093/jamia/ocaa036] [Medline: 32374376]
25. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020 Aug 27;3(1):110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
26. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228 [FREE Full text] [doi: 10.1037/hea0000305] [Medline: 26651463]

XSL•FO
RenderX

27. Inan OT, Tenaerts P, Prindiville SA, Reynolds HR, Dizon DS, Cooper-Arnold K, et al. Digitizing clinical trials. NPJ Digit Med 2020 Jul 31;3(1):101 [FREE Full text] [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]
28. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 03;53(6):573-582 [FREE Full text] [doi: 10.1093/abm/kay067] [Medline: 30192907]
29. Li L, Nakamura T, Hayano J, Yamamoto Y. Age and gender differences in objective sleep properties using large-scale body acceleration data in a Japanese population. Sci Rep 2021 May 11;11(1):9970 [FREE Full text] [doi: 10.1038/s41598-021-89341-x] [Medline: 33976280]
30. Kuula L, Gradisar M, Martinmäki K, Richardson C, Bonnar D, Bartel K, et al. Using big data to explore worldwide trends in objective sleep in the transition to adulthood. Sleep Med 2019 Oct;62:69-76. [doi: 10.1016/j.sleep.2019.07.024] [Medline: 31563008]
31. Katori M, Shi S, Ode KL, Tomita Y, Ueda HR. The 103,200-arm acceleration dataset in the UK Biobank revealed a landscape of human sleep phenotypes. Proc Natl Acad Sci U S A 2022 Mar 22;119(12):e2116729119 [FREE Full text] [doi: 10.1073/pnas.2116729119] [Medline: 35302893]
32. Arroyo AC, Zawadzki MJ. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. JMIR Mhealth Uhealth 2022 Apr 04;10(4):e33527 [FREE Full text] [doi: 10.2196/33527] [Medline: 35377327]
33. Carver CS, Scheier MF. Control theory: a useful conceptual framework for personality–social, clinical, and health psychology. Psychol Bull 1982;92(1):111-135. [doi: 10.1037/0033-2909.92.1.111]
34. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep 15;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
35. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
36. Harkin B, Webb TL, Chang BP, Prestwich A, Conner M, Kellar I, et al. Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. Psychol Bull 2016 Feb;142(2):198-229 [FREE Full text] [doi: 10.1037/bul0000025] [Medline: 26479070]
37. Takeuchi H, Suwa K, Kishi A, Nakamura T, Yoshiuchi K, Yamamoto Y. The effects of objective push-type sleep feedback on habitual sleep behavior and momentary symptoms in daily life: mHealth intervention trial using a health care internet of things system. JMIR Mhealth Uhealth 2022 Oct 06;10(10):e39150 [FREE Full text] [doi: 10.2196/39150] [Medline: 36201383]
38. Qian K, Koike T, Nakamura T, Schuller B, Yamamoto Y. Learning multimodal representations for drowsiness detection. IEEE Trans Intell Transport Syst 2022 Aug;23(8):11539-11548. [doi: 10.1109/tits.2021.3105326]
39. The occupational classification table. Ministry of Health, Labor and Welfare in Japan. 2014. URL: https://jsite.mhlw.go.jp/kanagawa-roudoukyoku/var/rev0/0112/9664/06bunruihyou.pdf [accessed 2023-06-07]
40. Beck AT, Steer RA, Brown GK. BDI-II: Beck Depression Inventory, 2nd edition. Psychological Corporation. 1996. URL: https://www.nctsn.org/measures/beck-depression-inventory-second-edition [accessed 2024-04-29]
41. Hidano T, Fukuhara M, Iwawaki S, Soga S, Spielberger CD. New STAI for the State-Trait Anxiety Inventory-Form JYZ. Tokyo, Japan: Jitsumu Kyouiku Press; 2000.
42. Watson D, Wiese D, Vaidya J, Tellegen A. The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence. J Pers Soc Psychol 1999 May;76(5):820-838. [doi: 10.1037/0022-3514.76.5.820]
43. Sato A, Yasuda A. Development of the Japanese version of Positive and Negative Affect Schedule (PANAS) scales. Jpn J Personal 2001;9(2):138-139. [doi: 10.2132/jjpjspp.9.2_138]
44. Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. Int J Chronobiol 1976;4(2):97-110. [Medline: 1027738]
45. Ishihara K, Miyashita A, Inugami M, Fukuda K, Yamazaki K, Miyata Y. The results of investigation by the Japanese version of Morningness-Eveningness Questionnaire. Shinrigaku Kenkyu 1986 Jun;57(2):87-91. [doi: 10.4992/jjpsy.57.87] [Medline: 3784166]
46. Roenneberg T, Wirz-Justice A, Merrow M. Life between clocks: daily temporal patterns of human chronotypes. J Biol Rhythms 2003 Feb;18(1):80-90 [FREE Full text] [doi: 10.1177/0748730402239679] [Medline: 12568247]
47. Kitamura S, Hida A, Aritake S, Higuchi S, Enomoto M, Kato M, et al. Validity of the Japanese version of the Munich ChronoType questionnaire. Chronobiol Int 2014 Aug 14;31(7):845-850. [doi: 10.3109/07420528.2014.914035] [Medline: 24824747]
48. Kantermann T, Sung H, Burgess HJ. Comparing the morningness-eveningness questionnaire and Munich ChronoType questionnaire to the Dim Light Melatonin Onset. J Biol Rhythms 2015 Oct 04;30(5):449-453 [FREE Full text] [doi: 10.1177/0748730415597520] [Medline: 26243627]
49. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16(3):199-202. [doi: 10.1093/abm/16.3.199]

XSL•FO
RenderX

50. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
51. Fukui I. The Depression and Anxiety Mood Scale (DAMS): scale development and validation. Jpn J Behav Ther 1997 Sep;23(2):83-93. [doi: 10.24468/jjbt.23.2_83]
52. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992 Oct;15(5):461-469. [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
53. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. Sleep 1982;5(4):389-399. [doi: 10.1093/sleep/5.4.389] [Medline: 7163726]
54. Sleep guidelines for health promotion 2014. Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/file/06-Seisakujouhou-10900000-Kenkoukyoku/0000047221.pdf [accessed 2023-06-07]
55. Buysse DJ, Reynolds 3rd CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
56. Doi Y, Minowa M, Uchiyama M, Okawa M, Kim K, Shibui K, et al. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh Sleep Quality Index (PSQI-J) in psychiatric disordered and control subjects. Psychiatry Res 2000 Dec 27;97(2-3):165-172. [doi: 10.1016/s0165-1781(00)00232-8] [Medline: 11166088]
57. Mollayeva T, Thurairajah P, Burton K, Mollayeva S, Shapiro CM, Colantonio A. The Pittsburgh Sleep Quality index as a screening tool for sleep dysfunction in clinical and non-clinical samples: a systematic review and meta-analysis. Sleep Med Rev 2016 Feb;25:52-73. [doi: 10.1016/j.smrv.2015.01.009] [Medline: 26163057]
58. Qian T, Klasnja P, Murphy SA. Linear mixed models with endogenous covariates: modeling sequential treatment effects with application to a mobile health study. Stat Sci 2020 Aug 1;35(3):375-390 [FREE Full text] [doi: 10.1214/19-sts720] [Medline: 33132496]
59. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statist Sci 1992 Nov 1;7(4):457-472. [doi: 10.1214/ss/1177011136]
60. Elfering A, Kottwitz MU, Tamcan Ö, Müller U, Mannion AF. Impaired sleep predicts onset of low back pain and burnout symptoms: evidence from a three-wave study. Psychol Health Med 2018 Dec 24;23(10):1196-1210. [doi: 10.1080/13548506.2018.1479038] [Medline: 29792065]
61. Yabe Y, Hagiwara Y, Sekiguchi T, Sugawara Y, Tsuchiya M, Yoshida S, et al. Association between sleep disturbance and low back pain: a 3-year longitudinal study after the great east Japan earthquake. Spine (Phila Pa 1976) 2022 Feb 15;47(4):361-368. [doi: 10.1097/BRS.0000000000004234] [Medline: 34593731]
62. Adams RJ, Appleton SL, Taylor AW, Gill TK, Lang C, McEvoy RD, et al. Sleep health of Australian adults in 2016: results of the 2016 Sleep Health Foundation national survey. Sleep Health 2017 Feb;3(1):35-42. [doi: 10.1016/j.sleh.2016.11.005] [Medline: 28346149]
63. Hafner M, Stepanek M, Taylor J, Troxel WM, van Stolk C. Why sleep matters-the economic costs of insufficient sleep: a cross-country comparative analysis. Rand Health Q 2017 Jan;6(4):11 [FREE Full text] [Medline: 28983434]
64. Bei B, Wiley JF, Trinder J, Manber R. Beyond the mean: a systematic review on the correlates of daily intraindividual variability of sleep/wake patterns. Sleep Med Rev 2016 Aug;28:108-124 [FREE Full text] [doi: 10.1016/j.smrv.2015.06.003] [Medline: 26588182]
65. Slavish DC, Taylor DJ, Lichstein KL. Intraindividual variability in sleep and comorbid medical and mental health conditions. Sleep 2019 Jun 11;42(6):zsz052 [FREE Full text] [doi: 10.1093/sleep/zsz052] [Medline: 30843059]
66. Fang Y, Forger DB, Frank E, Sen S, Goldstein C. Day-to-day variability in sleep parameters and depression risk: a prospective cohort study of training physicians. NPJ Digit Med 2021 Feb 18;4(1):28 [FREE Full text] [doi: 10.1038/s41746-021-00400-z] [Medline: 33603132]
67. Lemola S, Ledermann T, Friedman EM. Variability of sleep duration is related to subjective sleep quality and subjective well-being: an actigraphy study. PLoS One 2013 Aug 14;8(8):e71292 [FREE Full text] [doi: 10.1371/journal.pone.0071292] [Medline: 23967186]
68. Wyatt JK, Ritz-De Cecco A, Czeisler CA, Dijk D. Circadian temperature and melatonin rhythms, sleep, and neurobehavioral function in humans living on a 20-h day. Am J Physiol 1999 Oct 01;277(4 Pt 2):R1152-R1163. [doi: 10.1152/ajpregu.1999.277.4.r1152] [Medline: 10516257]
69. Dijk D, Duffy JF, Riel E, Shanahan TL, Czeisler CA. Ageing and the circadian and homeostatic regulation of human sleep during forced desynchrony of rest, melatonin and temperature rhythms. J Physiol 1999 Apr 15;516 (Pt 2)(Pt 2):611-627 [FREE Full text] [doi: 10.1111/j.1469-7793.1999.0611v.x] [Medline: 10087357]
70. NeCamp T, Sen S, Frank E, Walton MA, Ionides EL, Fang Y, et al. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. J Med Internet Res 2020 Mar 31;22(3):e15033 [FREE Full text] [doi: 10.2196/15033] [Medline: 32229469]
71. Bidargaddi N, Schrader G, Klasnja P, Licinio J, Murphy S. Designing m-Health interventions for precision mental health support. Transl Psychiatry 2020 Jul 07;10(1):222 [FREE Full text] [doi: 10.1038/s41398-020-00895-2] [Medline: 32636358]

XSL•FO
RenderX

72. Wang S, Zhang C, Kröse B, van Hoof H. Optimizing adaptive notifications in mobile health interventions systems: reinforcement learning from a data-driven behavioral simulator. J Med Syst 2021 Oct 18;45(12):102 [FREE Full text] [doi: 10.1007/s10916-021-01773-0] [Medline: 34664120]
73. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Mar 18;4(1):1-22 [FREE Full text] [doi: 10.1145/3381007] [Medline: 34527853]
74. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mHealth: scoping review. JMIR Mhealth Uhealth 2022 Jun 09;10(6):e35053 [FREE Full text] [doi: 10.2196/35053] [Medline: 35679107]
75. Espie CA, Emsley R, Kyle SD, Gordon C, Drake CL, Siriwardena AN, et al. Effect of digital cognitive behavioral therapy for insomnia on health, psychological well-being, and sleep-related quality of life: a randomized clinical trial. JAMA Psychiatry 2019 Jan 01;76(1):21-30 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.2745] [Medline: 30264137]
76. Lazar A, Koehler C, Tanenbaum TJ, Nguyen DH. Why we use and abandon smart devices. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 7-11 URL: https://doi.org/10.1145/2750858.2804288 [doi: 10.1145/2750858.2804288]
77. Epstein DA, Caraway M, Johnston C, Ping A, Fogarty J, Munson SA. Beyond abandonment to next steps: understanding and designing for life after personal informatics tool use. Proc SIGCHI Conf Hum Factor Comput Syst 2016 May;2016:1109-1113 [FREE Full text] [doi: 10.1145/2858036.2858045] [Medline: 28503678]

Abbreviations

BDI-II: Beck Depression Inventory second edition
BLE: Bluetooth Low Energy
EMA: ecological momentary assessment
HIT: health care Internet of Things
IIM: intraindividual mean
IIV: intraindividual variability
JITAI: just-in-time adaptive intervention
MCTQ: Munich Chronotype Questionnaire
MEQ: Morningness-Eveningness Questionnaire
mHealth: mobile health
MRT: microrandomized trial
MSFsc: sleep-corrected midpoint of sleep on work-free days
PANAS: Positive and Negative Affect Schedule
PSQI: Pittsburgh Sleep Quality Index
STAI: State-Trait Anxiety Inventory-JZY

Edited by T de Azevedo Cardoso; submitted 13.06.23; peer-reviewed by J Yap; comments to author 12.07.23; revised version received 21.08.23; accepted 08.05.24; published 11.06.24.

Please cite as:
Takeuchi H, Ishizawa T, Kishi A, Nakamura T, Yoshiuchi K, Yamamoto Y
Just-in-Time Adaptive Intervention for Stabilizing Sleep Hours of Japanese Workers: Microrandomized Trial
J Med Internet Res 2024;26:e49669
URL: https://www.jmir.org/2024/1/e49669
doi:10.2196/49669
PMID:38861313

©Hiroki Takeuchi, Tetsuro Ishizawa, Akifumi Kishi, Toru Nakamura, Kazuhiro Yoshiuchi, Yoshiharu Yamamoto. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association of Smartwatch-Based Heart Rate and Physical Activity With Cardiorespiratory Fitness Measures in the Community: Cohort Study

Yuankai Zhang[1], MA; Xuzhi Wang[1], MS; Chathurangi H Pathiravasan[1], PhD; Nicole L Spartano[2], PhD; Honghuang Lin[3], PhD; Belinda Borrelli[4], PhD; Emelia J Benjamin[5,6], ScM, MD; David D McManus[7,8], ScM, MD; Martin G Larson[1,5], SD; Ramachandran S Vasan[5,6], MD; Ravi V Shah[9], MD; Gregory D Lewis[10,11], MD; Chunyu Liu[1], PhD; Joanne M Murabito[5,12*], ScM, MD; Matthew Nayor[13*], MPH, MD

[1]Department of Biostatistics, School of Public Health, Boston University, Boston, MA, United States

[2]Section of Endocrinology, Diabetes, Nutrition, and Weight Management, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[3]Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

[4]Center for Behavioral Science Research, Department of Health Policy & Health Services Research, Boston University, Henry M. Goldman School of Dental Medicine, Boston, MA, United States

[5]Boston University's and National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, United States

[6]Section of Preventive Medicine and Epidemiology and Cardiovascular Medicine, Departments of Medicine and Epidemiology, Boston University Chobanian & Avedisian School of Medicine and School of Public Health, Boston, MA, United States

[7]Cardiology Division, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

[8]Department of Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[9]Cardiology Division, Vanderbilt Translational and Clinical Research Center, Vanderbilt University Medical Center, Nashville, TN, United States

[10]Cardiology Division, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States

[11]Pulmonary Critical Care Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, United States

[12]Section of General Internal Medicine, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[13]Sections of Cardiology and Preventive Medicine and Epidemiology, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Matthew Nayor, MPH, MD
Sections of Cardiology and Preventive Medicine and Epidemiology
Department of Medicine
Boston University Chobanian & Avedisian School of Medicine
72 E Concord St, Suite L-516
Boston, MA, 02118
United States
Phone: 1 617 638 8771
Email: mnayor@bu.edu

Abstract

Background: Resting heart rate (HR) and routine physical activity are associated with cardiorespiratory fitness levels. Commercial smartwatches permit remote HR monitoring and step count recording in real-world settings over long periods of time, but the relationship between smartwatch-measured HR and daily steps to cardiorespiratory fitness remains incompletely characterized in the community.

Objective: This study aimed to examine the association of nonactive HR and daily steps measured by a smartwatch with a multidimensional fitness assessment via cardiopulmonary exercise testing (CPET) among participants in the electronic Framingham Heart Study.

XSL•FO
RenderX

Methods: Electronic Framingham Heart Study participants were enrolled in a research examination (2016-2019) and provided with a study smartwatch that collected longitudinal HR and physical activity data for up to 3 years. At the same examination, the participants underwent CPET on a cycle ergometer. Multivariable linear models were used to test the association of CPET indices with nonactive HR and daily steps from the smartwatch.

Results: We included 662 participants (mean age 53, SD 9 years; n=391, 59% women, n=599, 91% White; mean nonactive HR 73, SD 6 beats per minute) with a median of 1836 (IQR 889-3559) HR records and a median of 128 (IQR 65-227) watch-wearing days for each individual. In multivariable-adjusted models, lower nonactive HR and higher daily steps were associated with higher peak oxygen uptake (VO_2), % predicted peak VO_2, and VO_2 at the ventilatory anaerobic threshold, with false discovery rate (FDR)–adjusted *P* values <.001 for all. Reductions of 2.4 beats per minute in nonactive HR, or increases of nearly 1000 daily steps, corresponded to a 1.3 mL/kg/min higher peak VO_2. In addition, ventilatory efficiency (V_E/VCO_2; FDR-adjusted *P*=.009), % predicted maximum HR (FDR-adjusted *P*<.001), and systolic blood pressure-to-workload slope (FDR-adjusted *P*=.01) were associated with nonactive HR but not associated with daily steps.

Conclusions: Our findings suggest that smartwatch-based assessments are associated with a broad array of cardiorespiratory fitness responses in the community, including measures of global fitness (peak VO_2), ventilatory efficiency, and blood pressure response to exercise. Metrics captured by wearable devices offer a valuable opportunity to use extensive data on health factors and behaviors to provide a window into individual cardiovascular fitness levels.

(J Med Internet Res 2024;26:e56676) doi:10.2196/56676

KEYWORDS

mobile health; smartwatch; heart rate; physical activity; cardiorespiratory fitness; cardiopulmonary exercise testing

Introduction

Impaired cardiorespiratory fitness (CRF) is a powerful predictor of health outcomes including incident cardiovascular disease (CVD) events and mortality [1-6]. To optimize the prevention of CVD, the American Heart Association supports the clinical assessment of CRF as a “vital sign” in all adults [6]. While routine CRF assessment may be a valuable tool, given the possibility of modifying CRF with greater physical activity and other lifestyle modifications [7,8], frequent CRF assessment at the population level is not currently feasible, especially in resource-limited settings. Additional markers that track CRF are therefore desirable to gauge general CRF and changes with lifestyle modifications or therapeutic interventions.

Metrics derived from wearable devices present a unique opportunity to leverage large amounts of collected data on health factors and behaviors to provide a window into individual CRF levels. With the rise in technology use [9], nearly 70% of US adults report tracking at least 1 health measure [10,11], including millions of individuals who track their steps using wearable devices [12]. In addition to daily steps taken, wearable devices permit continuous monitoring of heart rate (HR) in real-world environments over long periods of time. The American Heart Association recommends that individuals get to know their HR to help self-monitor CRF [13]. Resting HR associated with CRF [14] and lower HR have also been associated with a lower risk for mortality and adverse cardiovascular outcomes [15,16]. In addition, routine physical activity is known to be associated with improvements in CRF in women and men and across cardiovascular risk [7]. However, the relationships between HR and physical activity indices from consumer-grade wearables to gold standard assessments of CRF measures from cardiopulmonary exercise testing (CPET) are sparse.

We investigated the association between the measures of CRF obtained from CPET with smartwatch measures of physical activity and nonactive HR in a community-based sample of middle-aged and older adults from the electronic Framingham Heart Study (eFHS). We investigated both peak VO_2 and novel CPET measures that capture complementary dimensions of CRF. We hypothesized that higher levels of CRF (higher peak VO_2) are associated with lower average HR and greater steps per day over follow-up. The ability of individuals to track real-world physiologic measures that reflect overall CRF in real time may encourage engagement with lifestyle behaviors that result in improvements in not only cardiovascular health but also overall health.

Methods

Study Sample

The Framingham Heart Study (FHS) is a community-based, prospective study initiated in 1948 in the town of Framingham, Massachusetts, to study CVD and its determinants [17-19]. In 2002-2005, the grandchildren of the original participants, the Third Generation cohort (Gen 3), a cohort of multiple ancestries (Omni 2), and a cohort of New Offspring Spouses were enrolled and underwent examinations every 6-8 years [17]. At the time of research examination 3 (2016-2019; n=3521), the participants were invited to the eFHS. The eFHS enrolled 2139 English-speaking FHS participants who owned a smartphone (iPhone 4S or newer iPhones with at least iOS 8.2 or Android Phones) and provided consent. Participants with iPhones were offered a study smartwatch (Apple Watch Series 0) to record step count and HR data for 1 year (or more if they desired) following enrollment. The participants who owned an Apple Watch were allowed to use their own smartwatch. The participants in eFHS were instructed to wear the smartwatch after waking up in the morning and remove the watch at bedtime for charging. The eFHS included a smartphone app for electronic consent, health questionnaires, notifications, and integration with the smartwatch. The eFHS research technician assisted

XSL•FO
RenderX

participants with app download and smartwatch setup and provided participants with a written protocol for remote use that included information on how to download the app, enter registration information, sign the consent forms, enable notifications on the smartphone, and set up the smartwatch. In research examination 3, the participants were also offered participation in CPET [20]. Of the 3521 participants attending the examination, 3117 completed the CPET, of whom 1050 enrolled in eFHS with an iPhone and chose to use a smartwatch (Figure 1). We further excluded 388 participants who did not contribute 30 or more days of HR or step data leaving a final sample for our primary analysis of 662 participants.

Figure 1. Flowchart of analytic sample inclusion. eFHS: electronic Framingham Heart Study; FHS: Framingham Heart Study; HR: heart rate.

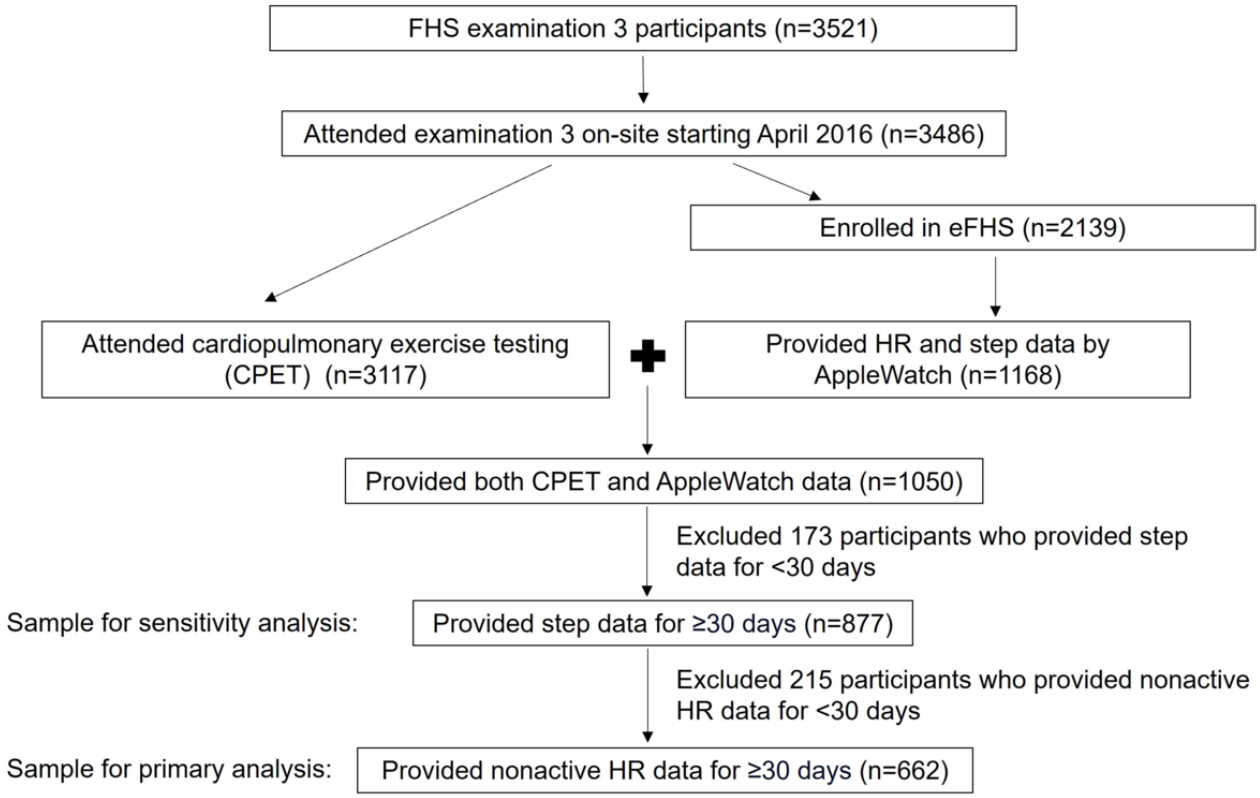

Ethical Considerations

The Boston University Medical Campus institutional review board reviewed and approved the study (H-36586 and H-32132). All participants provided informed consent for their participation in eFHS and for participation in examination 3, which included the CPET. Data were deidentified for analysis. Participants were not compensated for participating in the study.

CPET Measures (Independent Variables)

The CPET protocol used in FHS has been described previously [7,21,22] and is briefly summarized through the subsequent procedures. Maximal effort CPET was conducted on a cycle ergometer (Lode) with breath-by-breath gas exchange data measured (Medgraphics). The protocol consisted of 3 minutes of unloaded exercise and incremental ramp exercise using 1 of 2 ramp protocols (15 and 25 W/min). For this study, the primary CPET variables of interest were the "gold standard" measures of CRF: peak oxygen uptake (VO_2; in mL/kg/min), which was measured as the highest 30-second median value during the final minute of exercise and % predicted peak VO_2 (calculated using the Wasserman-Hansen equations) [23]. The VO_2 was also measured at the ventilatory anaerobic threshold (VAT), which was assessed using the V-slope method [24]. We evaluated ventilatory responses to exercise via ventilatory efficiency (assessed by the ratio of minute ventilation to carbon dioxide production [V_E/VCO_2] and calculated as the lowest 30-second median value observed during exercise) [21], HR responses to exercise (assessed by the % predicted maximum HR achieved, with the predicted maximum HR calculated using the Tanaka formula) [25], and blood pressure/vascular responses to exercise (assessed by systolic blood pressure-to-workload slope [SBP/W slope] and calculated as [peak SBP–rest SBP]/peak workload) [26].

Smartwatch Measures (Dependent Variables)

In this study, participants' HR and step counts were collected from smartwatches for up to 3 years, with a median follow-up of 128 (IQR 65-227) days.

The Apple Watch has built-in sensors including an accelerometer, gyroscope, and optical heart sensor to record the user's HR and determines the motion context, that is, whether the user is in sedentary or active status [27]. The Apple Watch Series 0 captured incomplete information on motion context, with only 31% of HR measures associated with information on motion context. Therefore, we developed a classification algorithm to infer the sedentary state for HR measures without motion context information. The smartwatch typically returns HR data every few minutes, and more frequent HR measures are returned when users are active. Using the HR data with known motion context, we observed that the recording intervals

XSL•FO
RenderX

were less than 1 minute for ~99% of HR measures with active status (Figure S1 in Multimedia Appendix 1). On the other hand, ~99% of HR measures with sedentary status had a recording interval greater than 1 minute. Based on the observed HR data with known motion context status, we inferred the motion context status for HR data with unknown status. We defined an HR measure to be "nonactive" if the HR measure satisfied the following conditions: (1) the recording interval between the adjacent HR measures was greater than 1 minute and (2) hourly step counts were less than 30 steps within the same hour when the HR was recorded. After inferring the motion context status for the 69% of HR measures that lacked motion context information directly returned from Apple Watch Series 0, we calculated the average nonactive HR measures for each participant over the study period.

The Apple Watch uses a built-in accelerometer to track users' wrist motion and estimate step counts. We first excluded any person-days on which the watch-wearing time was less than 5 hours or daily step counts were less than 1000 [28,29]. Subsequently, we derived average daily step counts for all participants over the study period.

Covariates

Demographic variables were assessed during examination 3 at the research center. BMI was determined by dividing the participant's body weight (in kilograms) by the square of their height (in meters). The use of medication was identified based on self-reported information regarding receiving such treatment in the year preceding examination 3. Heart rate–lowering medications were identified using the 2018 American College of Cardiology/American Heart Association/Heart Rhythm Society guideline on the evaluation and the management of patients with bradycardia and included β-blocking agents, calcium channel blocking agents, antiarrhythmic medications, and other medications [30]. Current smoking was defined as individuals who reported smoking at least 1 cigarette per day in the year prior to examination 3. In addition, several laboratory biomarkers, including total cholesterol, high-density lipoprotein, and fasting blood glucose, were measured during examination 3. The presence of diabetes was determined by a fasting blood glucose ≥126 mg/dL, or nonfasting blood glucose ≥200 mg/dL, or the use of any blood glucose–lowering medication. Hypertension was defined as systolic blood pressure (SBP) ≥140 mm Hg, diastolic blood pressure ≥90 mm Hg, or use of any antihypertensive medication. Hyperlipidemia was defined as fasting total cholesterol ≥200 mg/dL, triglyceride ≥150 mg/dL, or use of any lipid-lowering medication. Prevalent CVD was defined as a history of heart failure, myocardial infarction, angina, stroke, intermittent claudication, coronary insufficiency, or transient ischemic attack occurring any time from examination 1 to examination 3 that was reviewed by a panel of senior investigators using all available evidence including hospital records.

Statistical Analyses

In primary analyses, we examined the associations of CPET fitness measures as independent variables with both average nonactive HR and daily step counts as dependent variables. We selected smartwatch data as the dependent variable because temporally the smartwatch data were collected after the CPET. Peak VO_2 and VO_2 at VAT were right skewed and were natural—logarithmically transformed for analysis. The CPET measures were then standardized with a mean of 0, with an SD of 1 prior to analyses. We performed multivariable linear regression models based on different sets of covariates. In model 1, the covariates included age, sex, and self-reported race and ethnicity. In model 2, we additionally adjusted for BMI, current smoking status, total cholesterol, high-density lipoprotein, fasting glucose, diabetes status, resting SBP, prevalent CVD, lipid-lowering treatment status, hypertension treatment status, HR-lowering treatment status (only in analysis of nonactive HR), watch-wearing time (only in the analysis of daily steps), the season of enrollment, and the state of residence.

Furthermore, we conducted a sensitivity analysis to examine the association of CPET measures with nonactive HR with additional adjustment for daily step counts. To investigate whether we had consistent associations between daily steps and CPET fitness measures with a previous study that included 2070 FHS participants and used step data derived from waist-worn accelerometers, we additionally conducted a confirmatory analysis, following their methodology to treat daily step counts as independent variables and CPET fitness measures as dependent variables [7]. Furthermore, to mitigate the impact of insufficient volitional effort, we conducted a sensitivity analysis on a subset of participants who achieved a respiratory exchange ratio (RER) of 1.05 or higher.

For secondary analyses, we evaluated for effect modification by sex, age (older than 53 years or 53 years and younger, the median age in years), and BMI (a 3-level categorical variable: <25, ≥25 but <30, and ≥30 kg/m^2) on the associations of smartwatch variables with CPET measures using multiplicative interaction terms, as well as stratified analyses by these clinical variables.

In addition to the 662 participants who met the inclusion criteria for analyses of both nonactive HR and daily step data, 215 participants who were excluded from analyses of nonactive HR data met the inclusion criteria for analyses of daily step data. Hence, as a sensitivity analysis, we investigated the association between CPET fitness measures and daily steps in a larger sample including these 215 participants (N=877; Figure 1).

All analyses were performed in R (version 4.3.0; R Foundation for Statistical Computing). We used false discovery rate (FDR)–adjusted *P* value (Benjamini-Hochberg) <.05 to determine statistical significance.

Results

Characteristics of Study Participants

The primary study sample consisted of 662 participants, with 59% (n=391) women, mean age 53 (SD 9) years, and 91% (n=599) White (Table 1). The mean peak VO_2 was 27 (SD 7) mL/kg/min for men and 22 (SD 6) mL/kg/min for women and was 99% of the predicted values overall. The participants had a median follow-up of 128 (IQR 65-227) days and a median daily smartwatch wear time of 14 (IQR 13-15) hours. Overall,

XSL•FO
RenderX

we analyzed 2 million HR values over 365,240 person-days. The mean inferred nonactive HR of participants was 73 (SD 6) bpm. On average, the participants took 7269 (SD 2760) steps per day.

Of the 3117 participants who underwent CPET, 1813 were also enrolled in the eFHS. These 1813 participants, on average, were younger (52 vs 56 years) and a larger proportion were women (n=1020, 56% vs n=793, 49%) than those not in eFHS. They also tended to be in superior cardiovascular health with better CRF (Table S1 in Multimedia Appendix 1). Of the 1813 participants engaged in both CPET and eFHS, we excluded 1151 participants who did not use a smartwatch (n=763) or had insufficient smartwatch data (n=388). Those excluded had, on average, lower % predicted peak VO_2 and % predicted maximum HR than the final study sample. However, there was no significant difference in demographics or metabolic characteristics, such as age, sex, and CVD risk factors (Table S2 in Multimedia Appendix 1).

Table 1. Participant characteristics.

Characteristics	Overall (N=662)	Men (n=271)	Women (n=391)
Age (years), mean (SD)	53 (9)	54 (9)	52 (9)
White, n (%)	599 (91)	240 (89)	359 (92)
Hyperlipidemia, n (%)	388 (59)	171 (63.1)	217 (55.5)
Diabetes, n (%)	35 (5)	21 (7.7)	14 (3.6)
Hypertension, n (%)	164 (25)	92 (33.9)	72 (18.4)
Prevalent CVD[a], n (%)	20 (3)	13 (4.8)	7 (1.8)
BMI, mean (SD)	28.0 (5.3)	29.4 (5.0)	27.1 (5.3)
Peak VO_2[b] (mL/kg/min), mean (SD)	23.7 (6.9)	26.6 (6.8)	21.6 (6.1)
% Predicted peak VO_2, mean (SD)	99.0 (20.8)	96.6 (18.7)	100.6 (22.1)
VO_2 at VAT[c] (mL/kg/min), mean (SD)	12.7 (3.7)	13.9 (3.8)	12.0 (3.4)
V_E/VCO_2[d], mean (SD)	27.0 (2.7)	26.3 (2.5)	27.4 (2.7)
% Predicted maximum HR[e], mean (SD)	90.4 (9.7)	90.4 (10.3)	90.5 (9.2)
SBP/W[f] slope (mm Hg/W), mean (SD)	0.34 (0.14)	0.29 (0.1)	0.38 (0.15)
Follow-up days, median (IQR)	128 (65-227)	144 (70-246)	113 (63-205)
Number of HR records, median (IQR)	1836 (889-3559)	2062 (960-4063)	1632 (777-3203)
Watch wear time (hour/day), median (IQR)	14 (13-15)	14 (13-15)	14 (13-15)
Nonactive HR (bpm), mean (SD)	73 (6)	73 (7)	73 (6)
Daily steps, mean (SD)	7269 (2760)	7226 (2866)	7299 (2688)

[a]CVD: cardiovascular disease.

[b]VO_2: oxygen uptake.

[c]VAT: ventilatory anaerobic threshold.

[d]V_E/VCO_2: ventilatory efficiency.

[e]HR: heart rate.

[f]SBP/W: systolic blood pressure-to-workload.

Association of Cardiorespiratory Fitness With Nonactive HR

Adjusting for age, sex, and self-reported race and ethnicity, we observed that each 1-SD higher (favorable) value in $\log_e$ peak VO_2 (estimated β=2.4; FDR-adjusted P<.001), % predicted peak VO_2 (estimated β=1.88; FDR-adjusted P<.001), and $\log_e$ VO_2 at VAT (estimated β=1.98; FDR-adjusted P<.001) was associated with a lower nonactive HR (Table 2 and Figure 2A). Results also showed that a higher nonactive HR was associated with higher (unfavorable) value in V_E/VCO_2 (estimated β=.81; FDR-adjusted P=.002) and SBP/W slope (estimated β=1.11; FDR-adjusted P<.001), as well as higher % predicted maximum HR (estimated β=.87; FDR-adjusted P<.001). In the multivariable-adjusted model (model 2), we observed slight attenuation of the effect estimates, which remained statistically significant. Adjusted R^2 of model 2 ranged from 0.07 to 0.12 (Table S3 in Multimedia Appendix 1). Additional adjustment for smartwatch-based daily steps in sensitivity analyses attenuated the association between CPET fitness measures and nonactive HR by around 10%-20% compared with model 2 (Table S4 in Multimedia Appendix 1). The association of CPET fitness measures with nonactive HR persisted after excluding

XSL•FO
RenderX

21 participants with peak RER <1.05 (Table S5 in Multimedia Appendix 1).

We further assessed whether nonactive HR had different associations with CPET fitness measures between men and women and by age (above and below the median age) and BMI categories. We observed evidence of effect modification by age (FDR-adjusted *P* value of multiplicative interaction <.001; Table S6 in Multimedia Appendix 1) on the association of peak VO_2 with nonactive HR, such that each 1-SD higher $\log_e$ peak VO_2 was associated with a larger decrease in the observed nonactive HR in participants younger than 53 years compared with older participants (53 years and older). In addition, stratified analyses showed that the associations between higher peak VO_2 (both in mL/kg/min and % predicted) and lower nonactive HR remained significant across strata of sex, age, and BMI category (Figure 3A and Figure S2 in Multimedia Appendix 1). Apart from peak VO_2, age also modified the association of % predicted peak VO_2, VO_2 at VAT, and % predicted maximum HR on nonactive HR (FDR-adjusted *P* value of multiplicative interaction <.05). We did not find significant effect modification by sex or BMI of the association between CPET fitness measures and nonactive HR.

Table 2. Associations of smartwatch-based measures with CPET[a] fitness measures.[b]

CPET measure[c]	Nonactive HR[d] (dependent variable)						Daily steps (dependent variable)					
	Model 1 (n=662)			Model 2 (n=654)			Model 1 (n=662)			Model 2 (n=654)		
	Estimated β	SE	FDR[e]-adjusted *P* value	Estimated β	SE	FDR-adjusted *P* value	Estimated β	SE	FDR-*P*	Estimated β	SE	FDR-adjusted *P* value
Peak VO_2[f]	–2.40	0.27	<.001	–2.39	0.35	<.001	1268	110	<.001	916	129	<.001
% Predicted peak VO_2[g]	–1.88	0.24	<.001	–1.72	0.25	<.001	943	100	<.001	691	92	<.001
VO_2 at VAT[h]	–1.98	0.26	<.001	–1.66	0.32	<.001	1099	105	<.001	782	115	<.001
V_E/VCO_2[i]	0.81	0.26	.002	0.71	0.27	.009	206	111	.06	–157	98	.13
% Predicted maximum HR[j]	0.87	0.25	<.001	1.40	0.26	<.001	376	105	<.001	98	96	.31
SBP/W[k] slope	1.11	0.26	<.001	0.71	0.27	.01	367	112	.001	–175	99	.12

[a]CPET: cardiopulmonary exercise testing.

[b]Model 1 included age, sex, and self-reported race and ethnicity as covariates. Model 2 was additionally adjusted for body mass index, smoking status, total cholesterol, high-density lipoprotein, fasting glucose, diabetes status, resting systolic blood pressure, prevalent CVD, lipid-lowering treatment status, hypertension treatment status, HR-lowering treatment status (only in analysis of nonactive HR), watch-wearing time (only in analysis of daily steps), season of enrollment, and state of residence.

[c]CPET measures were standardized with a mean of 0 and an SD of 1 prior to analysis; peak VO_2 and VO_2 at VAT were standardized after natural log transformation.

[d]HR: heart rate.

[e]FDR: false discovery rate

[f]VO_2: oxygen uptake.

[g]% Predicted peak VO_2 was calculated using the Wasserman equation.

[h]VAT: ventilatory anaerobic threshold.

[i]V_E/VCO_2: ventilatory efficiency.

[j]Predicted maximum HR was calculated using the Tanaka formula.

[k]SBP/W: systolic blood pressure-to-workload.

XSL•FO
RenderX

Figure 2. Forest plots of associations between smartwatch-based measures and CPET fitness measures. CPET measures were standardized with a mean of 0 and SD of 1 prior to analysis. Peak VO_2 and VO_2 at VAT were standardized after natural log-transformation. % Predicted peak VO2 was calculated using the Wasserman equation. The predicted maximum HR was calculated using the Tanaka formula. (A) Nonactive HR. (B) Daily steps. CPET: cardiopulmonary exercise testing; HR: heart rate; SBP/W: systolic blood pressure-to-workload; VAT, ventilatory anaerobic threshold; VE/VCO_2: ventilatory efficiency; VO_2: oxygen uptake.

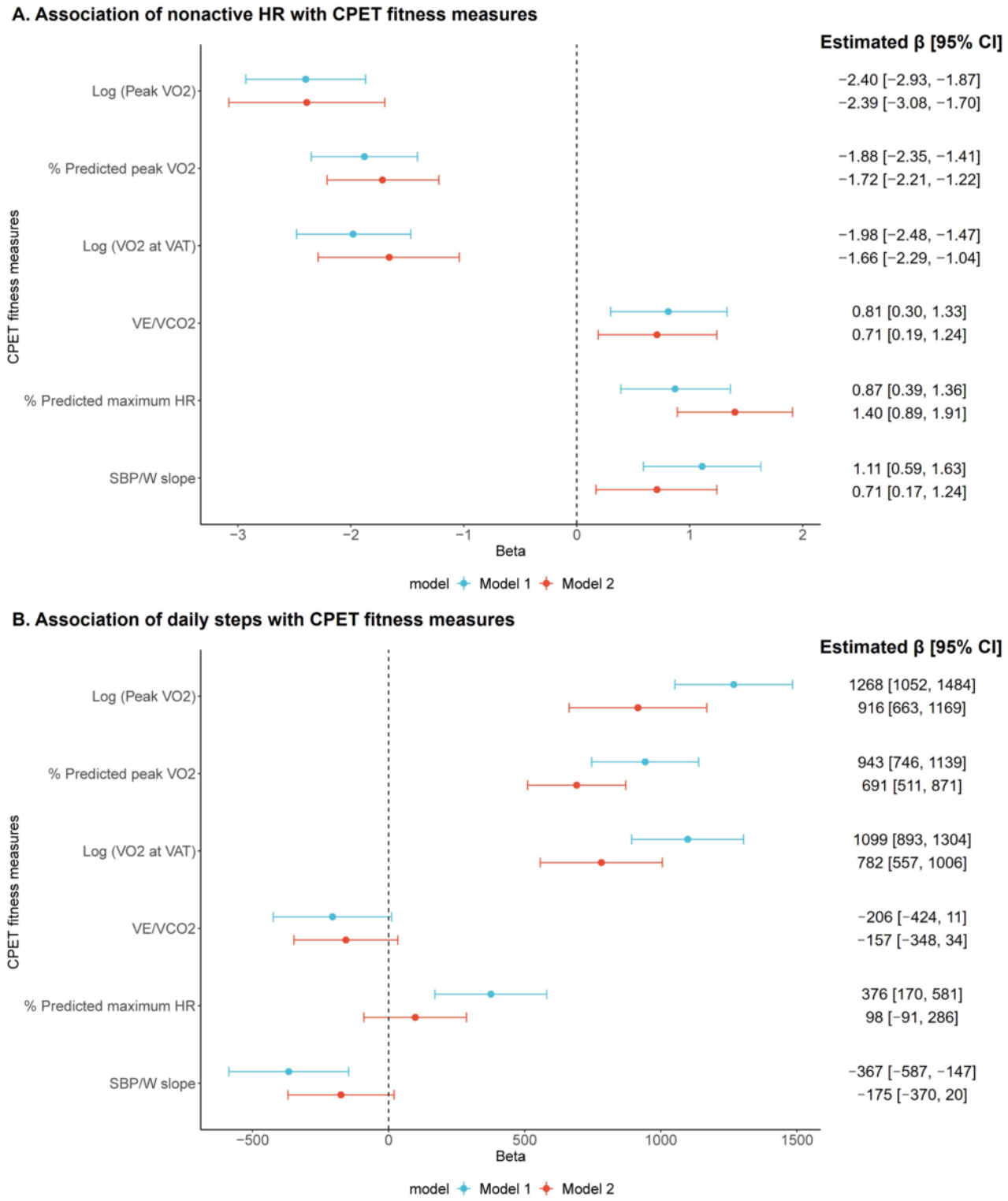

XSL•FO
RenderX

Figure 3. Forest plots of association between smartwatch-based measures and peak VO_2 stratified by sex, age, and BMI. Peak VO_2 was standardized with mean of 0 and SD of 1 after natural log-transformation. (A) Nonactive HR. (B) Daily steps. FDR-P: false discovery rate–adjusted *P* values; HR: heart rate; VO_2: oxygen uptake.

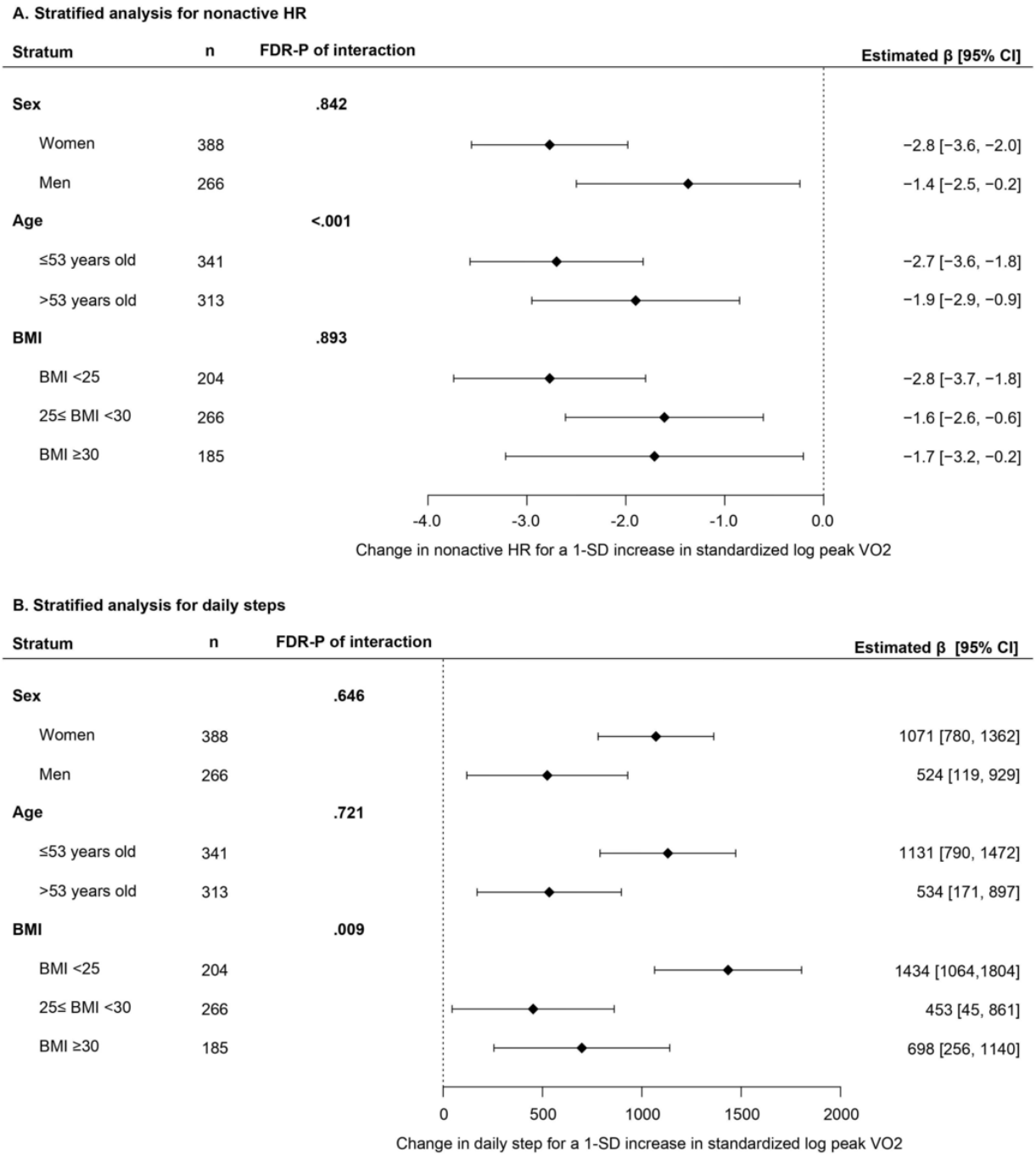

Association of Cardiorespiratory Fitness With Daily Steps

Next, we evaluated the relationship between CPET fitness measures and smartwatch-based daily steps. Greater daily steps were significantly related to a higher peak VO_2, % predicted peak VO_2, and VO_2 at VAT for both model 1 and model 2 (FDR-adjusted *P*<.001; Table 2 and Figure 2B). Each 1-SD higher $\log_e$ peak VO_2 was associated with 1268 more daily steps in model 1 and 916 more daily steps after adjusting for additional covariates in model 2. We observed no association of V_E/VCO_2 with daily steps although the directionality of β estimates was as expected. Both % predicted maximum HR (estimated β=376; FDR-adjusted *P*<.001) and SBP/W slope (estimated β=367; FDR-adjusted *P*=.001) were significantly associated with daily steps in model 1, whereas in model 2, the associations were attenuated (FDR-adjusted *P*>.05). Model 2 adjusted R^2 ranged from 0.3 to 0.36 (Table S3 in Multimedia Appendix 1). A sensitivity analysis restricted to participants with peak RER ≥1.05 (n=641) yielded concordant results (Table S5 in Multimedia Appendix 1). Furthermore, we observed consistent effect sizes compared with a previous study [7] in a

XSL•FO
RenderX

confirmatory analysis for which CPET fitness measures were considered as the dependent variable and daily steps (measured by accelerometers worn for a maximum of 8 days) were the independent variable (Table S7 in Multimedia Appendix 1).

Moreover, we observed significant effect modification by BMI on the associations of daily steps with peak VO_2, % predicted peak VO_2, and VO_2 at VAT (FDR-adjusted *P* value of multiplicative interaction <.05; Table S8 in Multimedia Appendix 1) for which the effect sizes were significantly greater in participants with BMI <25 kg/m^2 compared to higher BMI categories. The effect of % predicted maximum HR was also modified by BMI (FDR-adjusted *P* value of multiplicative interaction=.048), but the greatest effect size was observed among participants with BMI ≥30. Stratified analyses demonstrated that higher peak VO_2 (both in mL/kg/min and % predicted) was significantly associated with more daily steps across strata of sex, age, and BMI category (Figure 3B; Figure S2 in Multimedia Appendix 1). The association of CPET fitness measures on daily steps was largely consistent across sex and age categories (FDR-adjusted *P* value of multiplicative interaction >.05).

In a sensitivity analysis, we examined the association of CPET fitness measures with daily steps using the larger sample (n=877) with 215 additional participants (Figure 1). The results for peak VO_2, % predicted peak VO_2, VO_2 at VAT, % predicted maximum HR, and SBP/W slope remained very similar to the primary analysis with a sample size of 662 participants (Table S9 in Multimedia Appendix 1). Nevertheless, we observed that higher V_E/VCO_2 was significantly associated with lower daily steps in the larger sample (FDR-adjusted *P*<.05), which was not revealed by the primary analysis.

Discussion

Principal Findings

In our community-based sample of middle age and older adults, we observed an association between higher peak VO_2 and lower nonactive HR measured from a consumer-grade smartwatch, with every 1.3 mL/kg/min higher peak VO_2 corresponding to a 2.4-bpm lower nonactive HR. The association persisted after adjusting for CVD risk factors, prevalent CVD, and background physical activity. In addition to peak VO_2 (an assessment of global CRF), lower nonactive HR was associated with complementary exercise response patterns such as higher pre-VAT VO_2, better ventilatory efficiency, and favorable blood pressure response to exercise. Our study also observed an association between higher peak VO_2 and physical activity with 1.3 mL/kg/min higher peak VO_2 associated with nearly 1000 more daily steps. While associations between peak VO_2 and daily steps were observed across strata of age, sex, and BMI, the associations were greater in individuals with normal BMI. Using wearable devices rather than research-grade accelerometers provides the advantage of direct feedback to the participants as well as the ability to incorporate information over a longer time horizon. These findings suggest that HR and daily steps measured from wearable devices may provide one approach to empower and motivate individuals to monitor physiologic measures to improve CRF and overall cardiovascular health.

Our findings are consistent with the UK Fenland Study, a community-based study of adults that noted strong cross-sectional and longitudinal relationships between resting HR and CRF [31]. In that study, resting HR was inversely related to CRF, and within-person change in HR was associated with within-person change in CRF. Furthermore, during the COVID-19 pandemic, a subsample of adults in the UK Fenland Study used a smartphone app for remote HR monitoring and investigators observed that resting HR trajectories differed by prepandemic CRF level. Others have also observed an inverse association between resting HR and CRF [32,33] and importantly between resting HR and adverse outcomes such as mortality independent of physical fitness [32,34]. Our work extends these studies by examining the association of CPET fitness measures with measures of HR obtained remotely from a smartwatch continuously over longer periods of time. Technology now permits remote HR monitoring at the individual and population levels providing a way to track fitness over time and may provide researchers with a population-level biomarker of CRF [31]. Our findings on the association of daily steps with CRF are consistent with a previous study involving 2070 FHS participants and using waist-worn accelerometry step data [7].

A key feature of our study is that we used multiparametric data from a consumer-grade smartwatch to identify HR values measured during nonactive periods. This approach has several potential strengths. First, it derives a meaningful metric of HR regulation from extant smartwatch data rather than requiring special acquisition steps. Second, nonactive HR encompasses many “resting” activities of daily life (eg, working and watching TV) that may lead to differential HR effects when compared with resting HR assessed in a clinical setting. Third, consumer-grade smartwatches conveniently gather a substantial amount of data. In this study, we analyzed 2 million HR values over 365,240 person-days among 662 participants. While it is established that an exaggerated HR increase at the onset of exercise may be a sign of cardiovascular risk [35], whether a lower HR during normal daily nonactive activities is associated with better health outcomes has not been previously evaluated to our knowledge. Our findings suggest that individuals with higher CRF also have lower HR with nonactive daily activities. The observation of effect modification of this association by age such that older individuals demonstrate an attenuated association of CRF and nonactive HR is also intriguing. This finding may be partially attributable to additional comorbidities that may affect HR control (eg, autonomic function) in older individuals and suggests that nonactive HR may be a better correlate of CRF in younger individuals. Finally, not only was global CRF associated with lower nonactive HR but so were lower (better) ventilatory efficiency and BP response to exercise. These observations may suggest that lower nonactive HR reflects multiple components of CRF.

People spend very little time engaged with a physician or other health care provider but spend an estimated 5000 hours each year in activities that can influence their health [36]. National survey data reveal that nearly half of all US adults report using

XSL•FO
RenderX

wearable devices daily, and the majority are willing to share their health data with health care providers [37]. A substantial majority of Americans across a range of demographic groups own a smartphone [9] and nearly 90% (n=1800) of users look at their phone within 1 hour of waking and throughout the day [38], providing opportunities to engage adults with monitoring HR, daily steps, and other health metrics in both urban and rural locations where health care access may be challenging. Moreover, many large research programs such as the *All of Us* research program are using wearables to gather HR and physical activity data with the ability to link the data to health outcomes available in electronic health records [39,40]. By demonstrating relations of CRF to 2 measures that are relatively straightforward to assess via wearable devices, this investigation provides potentially useful metrics to monitor with the goals of maintaining or improving CRF.

Another strength of this study is that the eFHS is embedded in the FHS, a long-running observational cohort study with deeply characterized participants with directly measured CVD risk factors and validated CVD outcomes. This allowed us to account for risk factors and CVD in our association analyses.

Limitations

Our study also has limitations. The eFHS participants were primarily White, well-educated, and healthier than other FHS participants and thus our results may not be generalizable to more diverse samples, other sociodemographic groups, or geographic regions. Our study smartwatch was the Apple Watch version 0. While the Apple Watch has shown good accuracy for HR [41], we did not have motion context information (ie, sedentary or active) for most HR returns. Based on HR return intervals and hourly step counts, we used an imputation method to identify nonactive HR, which was highly correlated with known sedentary HR (Figure S3 in Multimedia Appendix 1). Others have used similar algorithms to identify resting HR from smartwatches or Fitbit data [42]. Our nonactive HR (mean 73, SD 6 bpm) was lower than "known resting HR" (geometric mean 81.8, SD 19.6 bpm) measured at any time by an app that uses a smartphone camera in the Health eHeart cohort [42]. More than 30% (n=2414) of their sample were younger than 30 years, which might explain the higher HR. In contrast, our nonactive HR was higher than resting HR (59-66 bpm) measured in the Michigan Predictive Activity and Clinical Trajectories in Health study [43]. Some of the differences could be due to differences in study samples as HR was observed to differ by demographics (eg, age, sex, race, ethnicity, and BMI).

Conclusions

This study demonstrated that lower nonactive HR and higher daily steps were each associated with better CRF measures. Our findings suggest the potential use of consumer-grade smartwatches in capturing key physiological metrics that correlate with cardiovascular health and CRF. We demonstrate that peak VO_2, an established marker of CRF, is associated with nonactive HR and daily steps measured via smartwatches in a community-based sample of middle-aged and older adults. The associations are robust and persist even after rigorous adjustments and excluding participants who did not expend adequate volitional effort. The consistency of our results with previous studies further supports the credibility of our findings. Our research reveals the advantage of leveraging smartwatches in modern health monitoring. These devices not only offer a convenient method for long-term data capture in real-world environments but also provide direct feedback, providing the potential to motivate users to engage in health management. The smartwatch-based physiological metrics, such as nonactive HR and daily steps, may be a reliable proxy for global CRF as our study suggests. However, the limited generalizability of our findings underscores the necessity for further research in more diverse and representative populations. By understanding the correlations between physiological metrics captured by wearable devices and established health indicators, we can potentially leverage smartwatch-based physiological metrics as a population-level biomarker of fitness in epidemiological settings.

Acknowledgments

The electronic Framingham Heart Study is supported by R01-HL141434 and Robert Wood Johnson award 74624, and the cardiopulmonary exercise testing is supported by R01-HL131029 and by the American Heart Association (grant 15GPSGC24800006). The Framingham Heart Study is supported by the National Heart, Lung, and Blood Institute at the National Institutes of Health (contracts N01-HC-25195, HHSN268201500001I, and 75N92019D00031). EJB is supported in part by R01-HL09257 and R01-HL141434. HL is supported in part by U01AG068221. MN is supported by the National Institutes of Health (grants R01-HL156975 and R01-HL131029). The Apple Watches were provided to Boston University by Apple Inc at no cost to the study. The investigators thank the Framingham Heart Study participants. Generative AI was not used in any portion of the manuscript writing.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request and will be available at public repositories (DbGAP and BioLINNC).

Authors' Contributions

JMM, EJB, and DMM obtained funding and supervised collection of electronic Framingham Heart Study data; CL and YZ created analysis-ready data sets of the eFHS data; RSV and GDL obtained funding for cardiopulmonary exercise testing (CPET); RSV, GDL, RVS, and MN supervised data collection for CPET; MGL cleaned and analyzed CPET data; YZ conducted the statistical

XSL•FO
RenderX

analyses with CL for this study; and YZ, MN, and JMM drafted the original manuscript. All authors reviewed and edited the final manuscript.

Conflicts of Interest

RVS receives grants from the National Institutes of Health and the American Heart Association. In the last year, he has received funds from Amgen (scientific advisory board). RVS is a coinventor on a patent for ex-RNA signatures of cardiac remodeling and has patents pending on the proteomics of fitness and liver and lung disease. The spouse of RVS works for UpToDate (Wolters Kluwer). GDL has research funding from Amgen, Cytokinetics, Applied Therapeutics, AstraZeneca, and SoniVie about projects and clinical trials investigating exercise capacity that are distinct from this work. He has served as a scientific advisor for Pfizer, Merck, Boehringer-Ingelheim, Novartis, American Regent, Relypsa, Cyclerion, Cytokinetics, and Amgen, and receives royalties from UpToDate for scientific content authorship related to exercise physiology. MN has received speaking honoraria from Cytokinetics. DDM receives consulting fees from NAMSA, Avania, Fitbit, and Heart Rhythm Society for work on data safety, advisory, and editorial work. DDM has received research grant support from Bristol Myers Squibb, Pfizer, and Boeringher Ingelheim. DDM has received non-grant research support from Care Evolution, Apple Computer, and Fitbit. The other authors have no conflicts of interest to disclose. Apple was not involved in the study design, analysis, interpretation, or reporting of study results.

Multimedia Appendix 1
Additional figures and tables supporting the analysis presented in the main text.
[DOCX File , 442 KB - jmir_v26i1e56676_app1.docx]

References

1. Shah RV, Murthy VL, Colangelo LA, Reis J, Venkatesh BA, Sharma R, et al. Association of fitness in young adulthood with survival and cardiovascular risk: the Coronary Artery Risk Development in Young Adults (CARDIA) study. JAMA Intern Med 2016 Jan;176(1):87-95 [FREE Full text] [doi: 10.1001/jamainternmed.2015.6309] [Medline: 26618471]
2. Kodama S, Saito K, Tanaka S, Maki M, Yachi Y, Asumi M, et al. Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. JAMA 2009;301(19):2024-2035 [FREE Full text] [doi: 10.1001/jama.2009.681] [Medline: 19454641]
3. Balady GJ, Arena R, Sietsema K, Myers J, Coke L, Fletcher GF, et al. Clinician's guide to cardiopulmonary exercise testing in adults: a scientific statement from the American Heart Association. Circulation 2010;122(2):191-225 [FREE Full text] [doi: 10.1161/CIR.0b013e3181e52e69] [Medline: 20585013]
4. Sandvik L, Erikssen J, Thaulow E, Erikssen G, Mundal R, Rodahl K. Physical fitness as a predictor of mortality among healthy, middle-aged Norwegian men. N Engl J Med 1993;328(8):533-537 [FREE Full text] [doi: 10.1056/NEJM199302253280803] [Medline: 8426620]
5. Wei M, Kampert JB, Barlow CE, Nichaman MZ, Gibbons LW, Paffenbarger RS, et al. Relationship between low cardiorespiratory fitness and mortality in normal-weight, overweight, and obese men. JAMA 1999;282(16):1547-1553 [FREE Full text] [doi: 10.1001/jama.282.16.1547] [Medline: 10546694]
6. Ross R, Blair SN, Arena R, Church TS, Després JP, Franklin BA, et al. Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. Circulation 2016;134(24):e653-e699 [FREE Full text] [doi: 10.1161/CIR.0000000000000461] [Medline: 27881567]
7. Nayor M, Chernofsky A, Spartano NL, Tanguay M, Blodgett JB, Murthy VL, et al. Physical activity and fitness in the community: the Framingham Heart Study. Eur Heart J 2021 Nov 21;42(44):4565-4575 [FREE Full text] [doi: 10.1093/eurheartj/ehab580] [Medline: 34436560]
8. Ross LM, Barber JL, McLain AC, Weaver RG, Sui X, Blair SN, et al. The association of cardiorespiratory fitness and ideal cardiovascular health in the aerobics center longitudinal study. J Phys Act Health 2019;16(11):968-975. [doi: 10.1123/jpah.2018-0220] [Medline: 31553947]
9. Mobile fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/#mobile-phone-ownership-over-time [accessed 2022-01-26]
10. Fox S, Duggan M. Tracking for health. Pew Research Center. 2013. URL: https://www.pewresearch.org/internet/2013/01/28/tracking-for-health/ [accessed 2024-04-23]
11. Abril EP. Tracking myself: assessing the contribution of mobile technologies for self-trackers of weight, diet, or exercise. J Health Commun 2016;21(6):638-646. [doi: 10.1080/10810730.2016.1153756] [Medline: 27168426]
12. Connectivity and mobile trends survey, 2nd edition. Deloitte. 2021. URL: https://www.deloitte.com/an/en/our-thinking/insights/industry/media-telecommunications/connectivity-mobile-trends-survey.html [accessed 2024-04-23]
13. All about heart rate (pulse). American Heart Association. URL: https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse [accessed 2024-04-23]

XSL•FO
RenderX

14. Plaza-Florido A, Amaro-Gahete FJ, Acosta FM, Sacha J, Alcantara JMA. Heart rate rather than heart rate variability is better associated with cardiorespiratory fitness in adults. Eur J Sport Sci 2021;22(6):836-845. [doi: 10.1080/17461391.2021.1892198] [Medline: 33591861]
15. Kannel WB, Kannel C, Paffenbarger RS, Cupples LA. Heart rate and cardiovascular mortality: the Framingham Study. Am Heart J 1987;113(6):1489-1494. [doi: 10.1016/0002-8703(87)90666-1] [Medline: 3591616]
16. Aune D, Sen A, ó'Hartaigh B, Janszky I, Romundstad PR, Tonstad S, et al. Resting heart rate and the risk of cardiovascular disease, total cancer, and all-cause mortality—a systematic review and dose-response meta-analysis of prospective studies. Nutr Metab Cardiovasc Dis 2017;27(6):504-517. [doi: 10.1016/j.numecd.2017.04.004] [Medline: 28552551]
17. Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, et al. The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol 2007;165(11):1328-1335 [FREE Full text] [doi: 10.1093/aje/kwm021] [Medline: 17372189]
18. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham Offspring Study. Design and preliminary data. Prev Med 1975;4(4):518-525. [doi: 10.1016/0091-7435(75)90037-7] [Medline: 1208363]
19. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 1951;41(3):279-281. [doi: 10.2105/ajph.41.3.279] [Medline: 14819398]
20. Nayor M, Shah RV, Tanguay M, Blodgett JB, Chernofsky A, Miller PE, et al. Feasibility, methodology, and interpretation of broad-scale assessment of cardiorespiratory fitness in a large community-based sample. Am J Cardiol 2021;157:56-63. [doi: 10.1016/j.amjcard.2021.07.020] [Medline: 34391575]
21. Nayor M, Xanthakis V, Tanguay M, Blodgett JB, Shah RV, Schoenike M, et al. Clinical and hemodynamic associations and prognostic implications of ventilatory efficiency in patients with preserved left ventricular systolic function. Circ Heart Fail 2020;13(5):e006729 [FREE Full text] [doi: 10.1161/CIRCHEARTFAILURE.119.006729] [Medline: 32362167]
22. Nayor M, Shah RV, Miller PE, Blodgett JB, Tanguay M, Pico AR, et al. Metabolic architecture of acute exercise response in middle-aged adults in the community. Circulation 2020 Nov 17;142(20):1905-1924 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.050281] [Medline: 32927962]
23. Hansen JE, Sue DY, Wasserman K. Predicted values for clinical exercise testing. Am Rev Respir Dis 1984;129(2 Pt 2):S49-S55. [doi: 10.1164/arrd.1984.129.2P2.S49] [Medline: 6421218]
24. Beaver WL, Wasserman K, Whipp BJ. A new method for detecting anaerobic threshold by gas exchange. J Appl Physiol (1985) 1986;60(6):2020-2027 [FREE Full text] [doi: 10.1152/jappl.1986.60.6.2020] [Medline: 3087938]
25. Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. J Am Coll Cardiol 2001 Jan;37(1):153-156 [FREE Full text] [doi: 10.1016/s0735-1097(00)01054-8] [Medline: 11153730]
26. Nayor M, Gajjar P, Murthy VL, Miller PE, Velagaleti RS, Larson MG, et al. Blood pressure responses during exercise: physiological correlates and clinical implications. Arterioscler Thromb Vasc Biol 2023 Jan;43(1):163-173 [FREE Full text] [doi: 10.1161/ATVBAHA.122.318512] [Medline: 36384270]
27. Monitor your heart rate with Apple Watch. Apple Inc. URL: https://support.apple.com/en-us/HT204666 [accessed 2024-04-23]
28. Lin H, Sardana M, Zhang Y, Liu C, Trinquart L, Benjamin EJ, et al. Association of Habitual Physical Activity with Cardiovascular Disease Risk. Circ Res 2020;127(10):1253-1260 [FREE Full text] [doi: 10.1161/CIRCRESAHA.120.317578] [Medline: 32842915]
29. Patel MS, Benjamin EJ, Volpp KG, Fox CS, Small DS, Massaro JM, et al. Effect of a game-based intervention designed to enhance social incentives to increase physical activity among families: the BE FIT randomized clinical trial. JAMA Intern Med 2017 Nov 01;177(11):1586-1593 [FREE Full text] [doi: 10.1001/jamainternmed.2017.3458] [Medline: 28973115]
30. Kusumoto FM, Schoenfeld MH, Barrett C, Edgerton JR, Ellenbogen KA, Gold MR, et al. 2018 ACC/AHA/HRS guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation 2019 Aug 20;140(8):e382-e482 [FREE Full text] [doi: 10.1161/CIR.0000000000000628] [Medline: 30586772]
31. Gonzales TI, Jeon JY, Lindsay T, Westgate K, Perez-Pozuelo I, Hollidge S, et al. Resting heart rate is a population-level biomarker of cardiorespiratory fitness: The Fenland Study. PLoS One 2023;18(5):e0285272 [FREE Full text] [doi: 10.1371/journal.pone.0285272] [Medline: 37167327]
32. Jensen MT, Suadicani P, Hein HO, Gyntelberg F. Elevated resting heart rate, physical fitness and all-cause mortality: a 16-year follow-up in the Copenhagen Male Study. Heart 2013 Jun;99(12):882-887 [FREE Full text] [doi: 10.1136/heartjnl-2012-303375] [Medline: 23595657]
33. Nauman J, Aspenes ST, Nilsen TIL, Vatten LJ, Wisløff U. A prospective population study of resting heart rate and peak oxygen uptake (the HUNT Study, Norway). PLoS One 2012;7(9):e45021 [FREE Full text] [doi: 10.1371/journal.pone.0045021] [Medline: 23028740]
34. Aladin AI, Whelton SP, Al-Mallah MH, Blaha MJ, Keteyian SJ, Juraschek SP, et al. Relation of resting heart rate to risk for all-cause mortality by gender after considering exercise capacity (the Henry Ford exercise testing project). Am J Cardiol 2014;114(11):1701-1706. [doi: 10.1016/j.amjcard.2014.08.042] [Medline: 25439450]

XSL•FO
RenderX

35. Falcone C, Buzzi MP, Klersy C, Schwartz PJ. Rapid heart rate increase at onset of exercise predicts adverse cardiac events in patients with coronary artery disease. Circulation 2005;112(13):1959-1964 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.105.545111] [Medline: 16172270]
36. Asch DA, Muller RW, Volpp KG. Automated hovering in health care--watching over the 5000 hours. N Engl J Med 2012 Jul 05;367(1):1-3 [FREE Full text] [doi: 10.1056/NEJMp1203869] [Medline: 22716935]
37. Dhingra LS, Aminorroaya A, Oikonomou EK, Nargesi AA, Wilson FP, Krumholz HM, et al. Use of wearable devices in individuals with or at risk for cardiovascular disease in the US, 2019 to 2020. JAMA Netw Open 2023 Jun 01;6(6):e2316634 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.16634] [Medline: 37285157]
38. 2017 global mobile consumer survey: US edition: the dawn of the next era in mobile. Deloitte. 2017. URL: https://tinyurl.com/y2fvyyp4 [accessed 2024-04-23]
39. Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, et al. Association of step counts over time with the risk of chronic disease in the All of Us Research Program. Nat Med 2022;28(11):2301-2308 [FREE Full text] [doi: 10.1038/s41591-022-02012-w] [Medline: 36216933]
40. Master H, Kouame A, Marginean K, Basford M, Harris P, Holko M. How Fitbit data are being made available to registered researchers in All of Us Research Program. Pac Symp Biocomput 2023;28:19-30 [FREE Full text] [Medline: 36540961]
41. Wang R, Blackburn G, Desai M, Phelan D, Gillinov L, Houghtaling P, et al. Accuracy of wrist-worn heart rate monitors. JAMA Cardiol 2017;2(1):104-106 [FREE Full text] [doi: 10.1001/jamacardio.2016.3340] [Medline: 27732703]
42. Avram R, Tison GH, Aschbacher K, Kuhar P, Vittinghoff E, Butzner M, et al. Real-world heart rate norms in the health eHeart study. NPJ Digit Med 2019;2:58 [FREE Full text] [doi: 10.1038/s41746-019-0134-9] [Medline: 31304404]
43. Golbus JR, Pescatore NA, Nallamothu BK, Shah N, Kheterpal S. Lancet Digit Health 2021 Nov;3(11):e707-e715 [FREE Full text] [doi: 10.1016/S2589-7500(21)00138-2] [Medline: 34711377]

Abbreviations

CPET: cardiopulmonary exercise testing
CRF: cardiorespiratory fitness
CVD: cardiovascular disease
eFHS: electronic Framingham Heart Study
FDR: false discovery rate
FHS: Framingham Heart Study
HR: heart rate
RER: respiratory exchange ratio
SBP: systolic blood pressure
SBP/W slope: systolic blood pressure-to-workload slope
VAT: ventilatory anaerobic threshold

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 23.01.24; peer-reviewed by V Manea; comments to author 20.02.24; revised version received 11.03.24; accepted 20.03.24; published 13.06.24.

Please cite as:
Zhang Y, Wang X, Pathiravasan CH, Spartano NL, Lin H, Borrelli B, Benjamin EJ, McManus DD, Larson MG, Vasan RS, Shah RV, Lewis GD, Liu C, Murabito JM, Nayor M
Association of Smartwatch-Based Heart Rate and Physical Activity With Cardiorespiratory Fitness Measures in the Community: Cohort Study
J Med Internet Res 2024;26:e56676
URL: https://www.jmir.org/2024/1/e56676
doi:10.2196/56676
PMID:38870519

©Yuankai Zhang, Xuzhi Wang, Chathurangi H Pathiravasan, Nicole L Spartano, Honghuang Lin, Belinda Borrelli, Emelia J Benjamin, David D McManus, Martin G Larson, Ramachandran S Vasan, Ravi V Shah, Gregory D Lewis, Chunyu Liu, Joanne M Murabito, Matthew Nayor. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Assessment of Seasonal Impacts and Depression Associations on Circadian Rhythm Using Multimodal Wearable Sensing: Retrospective Analysis

Yuezhou Zhang[1], PhD; Amos A Folarin[1,2,3,4,5], PhD; Shaoxiong Sun[1,6], PhD; Nicholas Cummins[1], PhD; Yatharth Ranjan[1], MSc; Zulqarnain Rashid[1], PhD; Callum Stewart[1], PhD; Pauline Conde[1], BSc; Heet Sankesara[1], BSc; Petroula Laiou[1], PhD; Faith Matcham[7,8], PhD; Katie M White[7], BSc; Carolin Oetzmann[7], MSc; Femke Lamers[9,10], PhD; Sara Siddi[11,12,13], PhD; Sara Simblett[14], PhD; Srinivasan Vairavan[15], PhD; Inez Myin-Germeys[16], PhD; David C Mohr[17], PhD; Til Wykes[14,18], PhD; Josep Maria Haro[11,12,13], PhD; Peter Annas[19], PhD; Brenda WJH Penninx[9,10], PhD; Vaibhav A Narayan[15,20], PhD; Matthew Hotopf[3,7,18], PhD; Richard JB Dobson[1,2,3,4,5], PhD; RADAR-CNS consortium[21]

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[2]Institute of Health Informatics, University College London, London, United Kingdom
[3]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust, London, United Kingdom
[4]NIHR Biomedical Research Centre at University College London Hospitals, NHS Foundation Trust, London, United Kingdom
[5]Health Data Research UK London, University College London, London, United Kingdom
[6]Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
[7]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[8]School of Psychology, University of Sussex, Falmer, United Kingdom
[9]Department of Psychiatry, Amsterdam UMC, Vrije Universiteit, Amsterdam, Netherlands
[10]Mental Health Program, Amsterdam Public Health Research Institute, Amsterdam, Netherlands
[11]Centro de Investigación Biomédica en Red de Salud Mental, Madrid, Spain
[12]Teaching Research and Innovation Unit, Parc Sanitari Sant Joan de Déu, Fundació Sant Joan de Déu, Barcelona, Spain
[13]Faculty of Medicine and Health Sciences, Universitat de Barcelona, Barcelona, Spain
[14]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[15]Janssen Research and Development LLC, Titusville, NJ, United States
[16]Department of Neurosciences, Center for Contextual Psychiatry, Katholieke Universiteit Leuven, Leuven, Belgium
[17]Center for Behavioral Intervention Technologies, Department of Preventive Medicine, Northwestern University, Chicago, IL, United States
[18]South London and Maudsley NHS Foundation Trust, London, United Kingdom
[19]H Lundbeck A/S, Copenhagen, Denmark
[20]Davos Alzheimer's Collaborative, Geneva, Switzerland
[21]See Acknowledgments

Corresponding Author:
Yuezhou Zhang, PhD
Department of Biostatistics & Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London
SGDP Centre, De Crespigny Park, Denmark Hill
London, SE5 8AF
United Kingdom
Phone: 44 7579856617
Email: yuezhou.zhang@kcl.ac.uk

Abstract

XSL•FO
RenderX

Background: Previous mobile health (mHealth) studies have revealed significant links between depression and circadian rhythm features measured via wearables. However, the comprehensive impact of seasonal variations was not fully considered in these studies, potentially biasing interpretations in real-world settings.

Objective: This study aims to explore the associations between depression severity and wearable-measured circadian rhythms while accounting for seasonal impacts.

Methods: Data were sourced from a large longitudinal mHealth study, wherein participants' depression severity was assessed biweekly using the 8-item Patient Health Questionnaire (PHQ-8), and participants' behaviors, including sleep, step count, and heart rate (HR), were tracked via Fitbit devices for up to 2 years. We extracted 12 circadian rhythm features from the 14-day Fitbit data preceding each PHQ-8 assessment, including cosinor variables, such as HR peak timing (HR acrophase), and nonparametric features, such as the onset of the most active continuous 10-hour period (M10 onset). To investigate the association between depression severity and circadian rhythms while also assessing the seasonal impacts, we used three nested linear mixed-effects models for each circadian rhythm feature: (1) incorporating the PHQ-8 score as an independent variable, (2) adding seasonality, and (3) adding an interaction term between season and the PHQ-8 score.

Results: Analyzing 10,018 PHQ-8 records alongside Fitbit data from 543 participants (n=414, 76.2% female; median age 48, IQR 32-58 years), we found that after adjusting for seasonal effects, higher PHQ-8 scores were associated with reduced daily steps (β=–93.61, *P*<.001), increased sleep variability (β=0.96, *P*<.001), and delayed circadian rhythms (ie, sleep onset: β=0.55, *P*=.001; sleep offset: β=1.12, *P*<.001; M10 onset: β=0.73, *P*=.003; HR acrophase: β=0.71, *P*=.001). Notably, the negative association with daily steps was more pronounced in spring (β of PHQ-8 × spring = –31.51, *P*=.002) and summer (β of PHQ-8 × summer = –42.61, *P*<.001) compared with winter. Additionally, the significant correlation with delayed M10 onset was observed solely in summer (β of PHQ-8 × summer = 1.06, *P*=.008). Moreover, compared with winter, participants experienced a shorter sleep duration by 16.6 minutes, an increase in daily steps by 394.5, a delay in M10 onset by 20.5 minutes, and a delay in HR peak time by 67.9 minutes during summer.

Conclusions: Our findings highlight significant seasonal influences on human circadian rhythms and their associations with depression, underscoring the importance of considering seasonal variations in mHealth research for real-world applications. This study also indicates the potential of wearable-measured circadian rhythms as digital biomarkers for depression.

(J Med Internet Res 2024;26:e55302) doi:10.2196/55302

KEYWORDS

circadian rhythm; biological rhythms; mental health; major depressive disorder; MDD; wearable; mHealth; mobile health; digital health; monitoring

Introduction

Depression is a globally prevalent mental disorder with various negative impacts, including reduced quality of life, disability, premature mortality, and an increased risk of suicide [1-6]. However, current depression diagnostic methods face several critical limitations: (1) reliance on individuals' subjective recall of past behaviors, which introduces recall bias and neglects day-to-day fluctuations [7,8]; (2) dependence on skilled and experienced clinicians [9,10]; and (3) evaluations often delayed until the mental health issues have progressed to a more severe, difficult-to-treat stage [11]. Consequently, these limitations result in the underdiagnosis and delayed treatment for those with depression [12,13], underscoring the critical need for objective and timely methods for early detection.

Previous research has found significant links between circadian rhythms and depression [14,15]. Circadian rhythms are approximately 24-hour endogenous oscillations, controlled by the master clock in the suprachiasmatic nucleus of the hypothalamus, that regulate many aspects of human behavior and physiology, such as sleep-wake cycles, hormone secretion, and body temperature [16-18]. Disturbances in circadian rhythms have been associated with an increased risk of both physical and mental diseases [15,19-21]. Therefore, tracking human circadian rhythms is a potential objective method for early-stage depression identification. Traditional methods for circadian rhythm assessment involve tracking melatonin in bodily fluids (such as saliva, urine, or blood samples) in a constant light environment to prevent external light from altering melatonin production and biasing the assessment of the circadian phase [22,23]. However, this traditional approach is expensive, labor-intensive, and impractical for large cohort studies and long-term monitoring in real-world settings [24,25].

With the development of ubiquitous sensors and mobile technologies, wearable devices provide a convenient and cost-effective way to continuously monitor individuals' daily behaviors and physiological signals in real-world settings [26]. Previous mobile health (mHealth) studies have explored the approximation of human circadian rhythms through wearable-measured patterns, including sleep-wake cycles, rest-activity patterns, and circadian rhythms in heart rate (HR) [24,27-32]. Higher depression severity has been found to be associated with later sleep onset and offset times, higher sleep variability, lower amplitude of activity, lower intradaily stability, and later acrophase of activity and HR [24,27-32].

However, previous mHealth studies have not fully accounted for seasonal variations, possibly due to their short study durations. Seasonal changes in sunlight and temperature are crucial environmental zeitgebers for the internal circadian clock, impacting human circadian rhythms [33-35]. Prior research has

XSL•FO
RenderX

reported significant seasonal effects on sleep patterns and activity levels [36-39]. Ignoring these seasonal impacts may introduce bias into the associations between depression and wearable-measured circadian rhythms in real-world settings. Thus, examining the effects of seasonality on wearable-measured circadian rhythms and their connections to depression in a comprehensive longitudinal data set is needed.

The primary aim of this study was to explore the associations between depression severity and wearable-measured circadian rhythms, accounting for seasonal effects and investigating potential variations across seasons. Our secondary aim was to quantify the seasonal changes in wearable-measured circadian rhythms within a European mHealth study for depression [40].

Methods

Data Set

Participants and Settings

The data analyzed in this study were sourced from the Remote Assessment of Disease and Relapse Major Depressive Disorder (RADAR-MDD) research program, which aimed to investigate the utility of remote technologies for monitoring depression and understanding factors that could help predict relapse in MDD [40]. A total of 623 participants were recruited from 3 study sites across 3 European countries (United Kingdom, Spain, and the Netherlands) and followed for up to 2 years [41]. The first participant was enrolled in November 2017 and the last participant was enrolled in June 2020, and the data collection was completed in April 2021. Due to this rolling enrollment process, participants' involvement in the study varied from 11 months to 24 months [41]. The RADAR-MDD program used the RADAR-base open-source platform to concurrently gather both active (eg, questionnaires) and passive (eg, wearable) data [42]. Comprehensive details of the study protocol and data set have been documented in publications [40] and [41], respectively.

Patient Involvement

The RADAR-MDD protocol was co-developed with a patient advisory board who shared their opinions on several user-facing aspects of the study including the choice and frequency of survey measures, the usability of the study app, participant-facing documents, selection of optimal participation incentives, selection and deployment of wearable devices, and the data analysis plan.

Ethical Considerations

Ethical approvals were obtained from the Camberwell St. Giles Research Ethics Committee (17/LO/1154) in the United Kingdom, the Fundacio Sant Joan de Deu Clinical Research Ethics Committee (CI: PIC-128-17) in Spain, and the Medische Ethische Toetsingscommissie VUmc (2018.012-NL63557.029.17) in the Netherlands. All participants provided their written informed consent. Before the data collection, participants were assured of their privacy protection, informed of their right to withdraw at any time without needing to justify their decision, and allowed to request the deletion of all their collected data. The data from participants were pseudonymized and stored in a research database, adhering to the General Data Protection Regulation. Participants were compensated with £15 (US $19.09)/€20 (US $21.68) for enrollment, £5 (US $6.36)/€10 (US $10.84) for each clinical assessment conducted every 3 months, and an additional £10 (US $12.73)/€10 (US $10.84) for each qualitative interview they completed. No participant is identifiable in any images in the manuscript or the supplementary materials. Further details can be found in the study protocol and data description papers [40,41].

Measures

Depression Symptom Severity

Participants' depression symptom severity was measured using the 8-item Patient Health Questionnaire (PHQ-8) [43] conducted via mobile phones every 2 weeks. The PHQ-8 comprises 8 questions, and the total score of PHQ-8 ranges from 0 to 24, indicating increasing severity [43].

Fitbit Data

Participants were asked to wear a Fitbit Charge 2/3 wrist-worn device throughout the whole study. Participants' sleep, step count, and HR were continuously (24/7) measured and recorded.

- Sleep data: Fitbit provided sleep labels ("awake," "light sleep," "deep sleep," and "rapid eye movement") along with the corresponding local clock times every 30 seconds.
- Step data: Participants' accumulated steps were counted every minute.
- HR data: Fitbit provided an estimate of HR every 5 seconds, using an embedded photoplethysmography sensor. However, technical issues resulted in the absence of some sample points. To obtain the robust HR trend and align with step data, we calculated the average HR over 1 minute.

Season

The seasonal division used in this study was based on European Union astronomical seasons: spring (March 20 to June 20), summer (June 21 to September 22), autumn (September 23 to December 20), and winter (December 21 to March 19).

Covariates

In accordance with findings from previous studies [44-46], we considered several covariates that could potentially influence participants' circadian rhythms, including age, gender, and employment status. Because the COVID-19 pandemic and relevant restrictions had some significant impacts on individuals' behavior [47], we introduced a covariate "lockdown" to indicate the presence of a national lockdown. Furthermore, as the experience of seasons can be different across countries, the study site was also considered as a covariate. These covariates were considered in our statistical analysis.

Feature Extraction

PHQ-8 Interval

To link human circadian rhythms with depression severity, we extracted circadian rhythm features from each 14-day PHQ-8 interval—specifically, 14 days of Fitbit data preceding a completed PHQ-8. This feature window aligns with the PHQ-8's purpose of evaluating depressive symptom severity over the

XSL•FO
RenderX

past 2 weeks [43] and is consistent with both the existing literature in the mHealth field [48-50] and our previous studies [51,52]. Figure 1 shows an example of a participant's processed HR, step, and sleep data in a 14-day PHQ-8 interval.

Figure 1. An example of a participant's processed heart rate (HR), step, and sleep Fitbit data during the preceding 14 days of an 8-item Patient Health Questionnaire assessment collected via the RADAR–base platform.

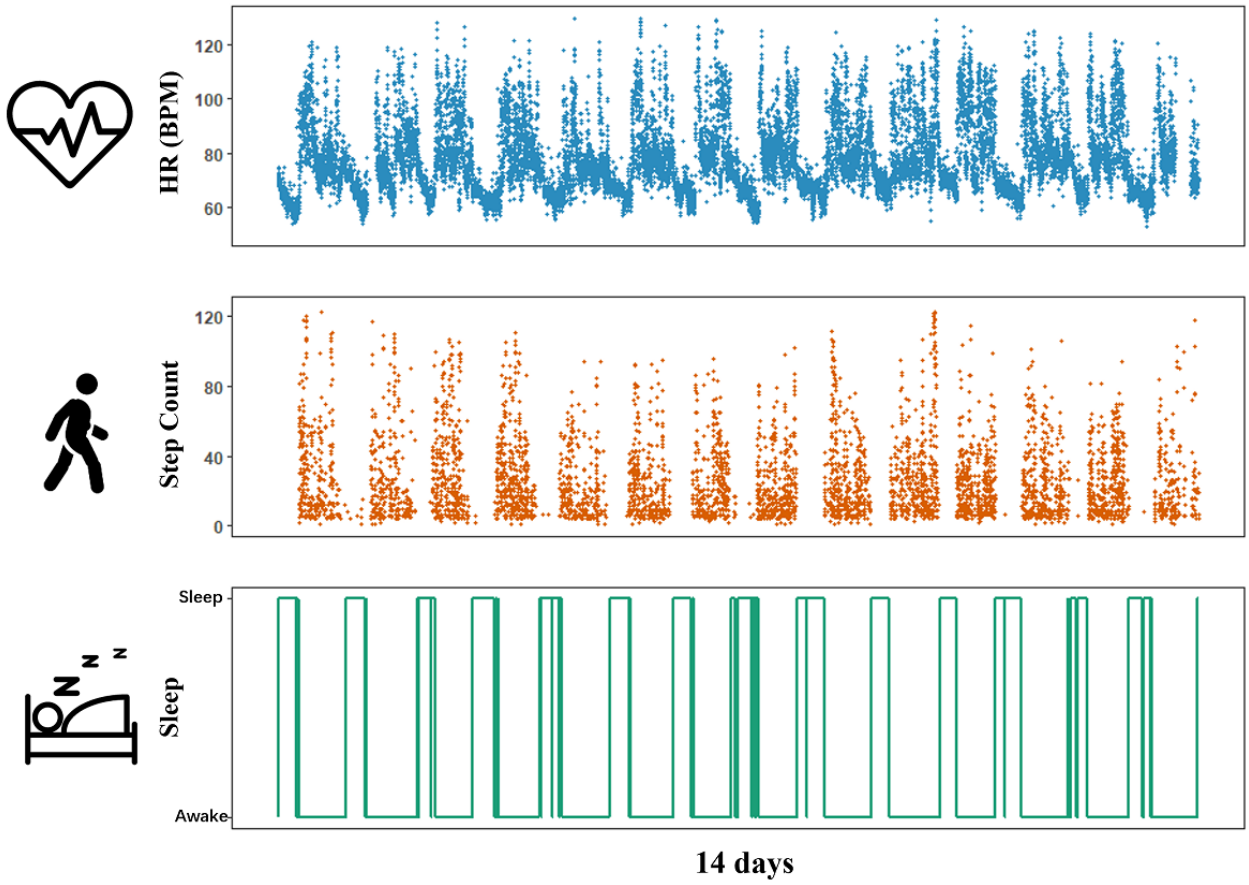

Data Inclusion Criteria

Some Fitbit data were missing in our data set for several reasons, including device damage, low battery level, and not being worn. Building on insights from our prior research [53], which discussed the sufficient days for stable feature calculation, we focused on 14-day PHQ-8 intervals with at least 8 days having more than 80% of step and HR data and a sleep recording in this study. Considering the potential impact of daylight saving time on individuals' behaviors [54], we excluded the first 14-day PHQ-8 interval after the time switching [37]. We then extracted a total of 12 features for reflecting the circadian rhythms from sleep, step count, and HR data.

Sleep-Wake Rhythms

Fitbit shows promise in identifying sleep-wake status [55,56]. Therefore, to reflect the sleep-wake rhythms, we computed four features: (1) sleep duration (the mean total sleep time), (2) sleep variability (the SD of total sleep time), (3) sleep onset (the mean clock time of falling asleep), and (4) sleep offset (the mean clock time of wake-up) [51].

Rest-Activity Rhythms

We extracted five nonparametric features from the Fitbit step count recordings to characterize the stability, fragmentation, timing, and mean activity level of participants' rest-activity rhythms, using the R package *nparACT* (R Foundation for Statistical Computing) [57,58]. These features include (1) intradaily variability of steps (step IV), quantifying the fragmentation in the rest-activity cycle; (2) interdaily stability of steps (step IS), quantifying the stability of rest-activity patterns over a 14-day PHQ-8 interval; (3) L5 onset, representing the onset of least active continuous 5-hour period; (4) M10 onset, representing the onset of the most active continuous 10-hour period; and (5) daily step, representing the mean of daily total steps in a 14-day PHQ-8 interval [57,58].

Circadian Rhythm in HR

For estimating circadian rhythms in HR, we used cosinor analysis—fitting a cosine wave to time series behavioral data through least-squares regression, which has been widely used in previous mHealth studies [59-61]. Using the R package *cosinor*, we performed the cosinor analysis on Fitbit HR data of each 14-day PHQ-8 interval and extracted the following parameters: (1) HR MESOR, the midline estimating statistic of the fitted cosine wave for HR; (2) HR amplitude, the difference between the peak value and MESOR of the fitted cosine wave for HR; and (3) HR acrophase, the timing of the HR peak [59-61].

XSL•FO
RenderX

Statistical Analysis

Given the longitudinal nature of our data set, that is, each participant had repeated measurements, we used the linear mixed-effects model [62] with a participant-specific random intercept in this study, implemented using the R package *lmerTest*. To investigate whether disregarding seasonal effects biases the associations between depression severity and circadian rhythms, we established and compared the following 3 models for each of circadian rhythm features.

- Model 1: A linear mixed-effects model was established to regress each circadian rhythm feature with only the PHQ-8 score as the independent variable.
- Model 2: Season was included as an independent variable in addition to the PHQ-8 score, considering seasonal effects on the circadian rhythms feature.
- Model 3: To further explore potential variations in the association between depression severity and circadian rhythms across seasons, an interaction term between the PHQ-8 score and season was added to the main-effects model (model 2).

All models were adjusted by covariates: age, gender, study site, lockdown, and employment status. The equations of these 3 models are outlined as follows:

- Model 1: Circadian rhythm = β_1 PHQ-8 + COVs
- Model 2: Circadian rhythm = β_1 PHQ-8 + β_2 season + COVs
- Model 3: Circadian rhythm = β_1 PHQ-8 + β_2 season + β_3 PHQ-8 × season + COVs

where COVs represents all covariates mentioned above and circadian rhythm is one of the wearable-measured circadian rhythm features.

Likelihood ratio tests were then performed to examine whether including more variables (season and the interaction term) can significantly improve the fitting of the regression model. The Benjamini-Hochberg method was used for the correction of multiple comparisons [63]. Despite the inclusion of lockdown as a covariate, we also repeated our analysis on the subset of data before the COVID-19 pandemic (before January 31, 2020).

Results

Data Summary

In accordance with our data inclusion criteria (see the Methods section), we analyzed a total of 10,018 PHQ-8 records alongside corresponding Fitbit data from 543 participants, with an average of 16 recordings per participant. The cohort selected for this study had a median age of 48 (IQR 32-58) years, was predominantly female (n=414, 76.2%), and included 230 (42.4%) employed participants. The distribution of PHQ-8 records was approximately uniform across the seasons, with winter accounting for 26.9%, spring 23.6%, summer 25.5%, and autumn 23.9%. We observed that demographics and data collection varied across study sites, with participants at the CIBER (Centro de Investigación Biomédica en Red) site in Spain being older and more likely to be retired, whereas the King's College London site in the United Kingdom had the highest number of participants and PHQ-8 records collected. Table 1 summarizes participant demographics and PHQ-8 records for the entire cohort, with site-specific comparisons. Figure 2 visualizes the variations of wearable-measured circadian rhythm features across a year.

Table 1. A summary of demographics of participants and PHQ-8[a] questionnaires in the present study, with the site-specific comparisons.[b]

Characteristic	Total	CIBER[c] (Spain)	KCL[d] (UK)	VUMC[e] (Netherlands)
Participants, n	543	126	303	114
Age (years), median (IQR)	48.00 (32.00-58.00)	53.00 (47.25-60.00)	44.00 (30.00-56.00)	39.50 (26.00-57.75)
Female, n (%)	414 (76.2)	91 (72.2)	233 (76.9)	90 (78.9)
Employed, n (%)	230 (42.4)	30 (23.8)	164 (54.1)	36 (31.6)
PHQ-8 records, n	10,018	1990	5899	2129
Records in winter, n (%)	2699 (26.9)	541 (27.2)	1519 (25.8)	639 (30.0)
Records in spring, n (%)	2363 (23.6)	450 (22.6)	1504 (25.5)	409 (19.2)
Records in summer, n (%)	2559 (25.5)	520 (26.1)	1508 (25.6)	531 (24.9)
Records in autumn, n (%)	2397 (23.9)	479 (24.1)	1368 (23.2)	550 (25.8)
Participants before the COVID-19 pandemic, n	442	115	259	68
Records before the COVID-19 pandemic, n (%)	4202 (41.9)	999 (50.2)	2650 (44.9)	553 (26.0)

[a]PHQ-8: 8-item Patient Health Questionnaire.

[b]Due to the private issue, participants' specific geographic information was not collected.

[c]CIBER: Centro de Investigación Biomédica en Red.

[d]KCL: King's College London.

[e]VUMC: Vrije Universiteit Medisch Centrum.

XSL•FO
RenderX

Figure 2. Variations of wearable-measured circadian rhythm features across a year. For each participant, circadian rhythm features were normalized to reduce the individual differences. HR: heart rate.

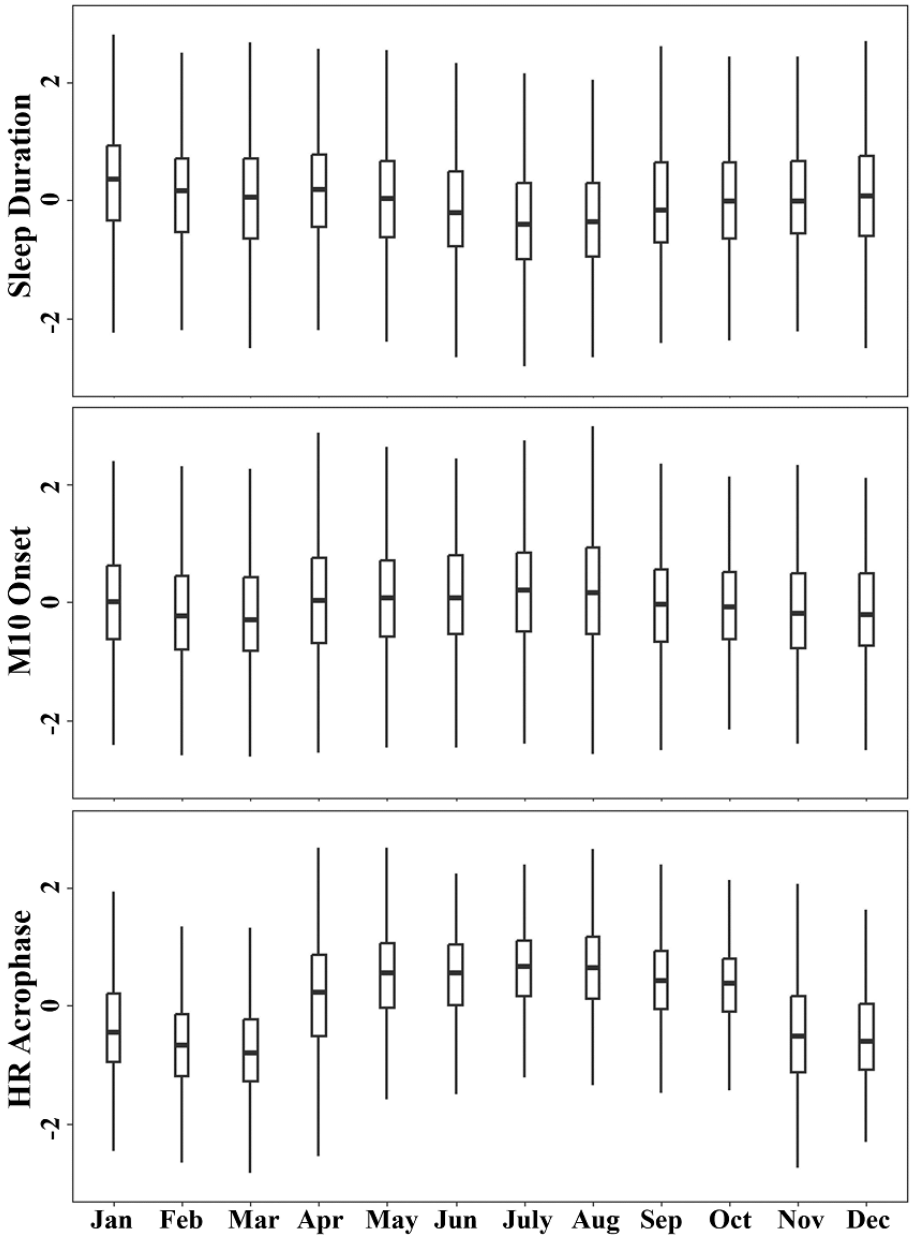

Associations Between the PHQ-8 score and Wearable-Measured Circadian Rhythm Features

The results of models 1 and 2 are displayed in Table 2. The likelihood ratio tests revealed that incorporating seasonal effects (model 2) significantly improved the model fit for all 12 circadian rhythm features. Specifically, model 2 demonstrated significant positive associations of the PHQ-8 score with several features: sleep duration (β_1=0.46, *P*<.001), sleep onset (β_1=0.55, *P*=.001), sleep offset (β_1=1.12, *P*<.001), sleep variability (β_1=0.96, *P*<.001), L5 onset (β_1=0.43, *P*=.04), M10 onset (β_1=0.73, *P*=.003), and HR acrophase (β_1=0.71, *P*=.001). Conversely, significant negative associations were observed with daily step (β_1=–93.61, *P*<.001), step IS (β_1=–0.001, *P*<.001), HR MESOR (β_1=–0.03, *P*<.001), and HR amplitude (β_1=–0.04, *P*<.001). However, the effect sizes of PHQ-8 on step IS, HR MESOR, and HR amplitude are small, according to their practical meanings.

XSL•FO
RenderX

Table 2. Effects of PHQ-8[a] and seasonality on wearable-measured circadian rhythm features estimated by mixed-effects regression models (models 1[b] and 2[c]).[d]

Feature	Model 1	Model 2				Likelihood ratio test (*P* values)
	PHQ-8, β_1 (SE)	PHQ-8, β_1 (SE)	Spring, β_2 (SE)	Summer, β_2 (SE)	Autumn, β_2 (SE)	
Sleep duration	0.51 (0.14)***	0.46 (0.14)***	–6.48 (1.32)***	–16.63 (1.31)***	–6.13 (1.31)***	<.001
Sleep onset	0.52 (0.17)**	0.55 (0.17)**	3.68 (1.61)*	7.16 (1.59)***	–4.64 (1.60)**	<.001
Sleep offset	1.13 (0.15)***	1.12 (0.15)***	–0.25 (1.44)	–6.91 (1.43)***	–8.32 (1.43)***	<.001
Sleep variability	0.97 (0.14)***	0.96 (0.14)***	–8.48 (1.36)***	–5.29 (1.35)***	–2.82 (1.35)*	<.001
Daily step	–94.71 (6.43)***	–93.61 (6.41)***	–19.21 (61.28)	394.46 (60.81)***	186.16 (60.75)**	<.001
Step IV[e]	0.001 (0.0006)	0.0007 (0.0006)	0.001 (0.006)	–0.026 (0.006)***	–0.021 (0.006)***	.006
Step IS[f]	–0.001 (0.0002)***	–0.001 (0.0002)***	0.002 (0.002)	–0.006 (0.002)*	0.004 (0.002)	.007
L5 onset[g]	0.43 (0.21)*	0.43 (0.21)*	–1.10 (2.13)	–3.84 (2.12)	–3.98 (2.12)	.002
M10 onset[h]	0.66 (0.25)**	0.73 (0.25)**	11.94 (2.45)***	20.51 (2.43)***	1.61 (2.44)	<.001
HR[i] MESOR[j]	–0.03 (0.01)***	–0.03 (0.01)***	–0.47 (0.09)***	–0.01 (0.09)	0.04 (0.09)	<.001
HR amplitude	–0.04 (0.01)***	–0.04 (0.01)***	0.19 (0.05)***	0.92 (0.05)***	0.39 (0.05)***	<.001
HR acrophase	0.51 (0.24)*	0.71 (0.22)**	43.35 (2.15)***	67.94 (2.13)***	24.09 (2.13)***	<.001

[a]PHQ-8: 8-item Patient Health Questionnaire.

[b]Model 1: Circadian rhythm = β_1 PHQ-8 + COVs.

[c]Model 2: Circadian rhythm = β_1 PHQ-8 + β_2 season + COVs, where COVs represents covariates mentioned in the Methods section. Note that winter is the reference season.

[d]The exact *P* values are reported in the main text.

[e]IV: intradaily variability.

[f]IS: interdaily stability.

[g]L5 onset: the onset of least active continuous 5-hour period.

[h]M10 onset: the onset of the most active continuous 10-hour period.

[i]HR: heart rate.

[j]HR MESOR: midline estimating statistic of the fitted cosine wave for heart rate.

*P<.05, **P<.01, ***P<.001.

The results of model 3, along with the likelihood ratio comparison with model 2, are summarized in Table 3. We found that the inclusion of the interaction term between PHQ-8 and seasonality significantly improved the model fit for daily step, step IV, step IS, M10 onset, HR MESOR, and HR amplitude. Notably, the negative association between the PHQ-8 score and daily step exhibited greater strength in spring (β_3 of PHQ-8 × spring = –31.51, P=.002) and summer (β_3 of PHQ-8 × summer = –42.61, P<.001) than in winter. Additionally, a significant positive association between M10 onset and PHQ-8 was observed exclusively in summer (β_3 of PHQ-8 × summer = 1.06, P=.008), with no significant associations observed in other seasons.

XSL•FO
RenderX

Table 3. Coefficients, SE, and significance[a] of PHQ-8[b], seasonality, and the interaction term between PHQ-8 and seasons in model 3[c] for the entire data set.

Feature	PHQ-8, β_1 (SE)	Spring, β_2 (SE)	Summer, β_2 (SE)	Autumn, β_2 (SE)	PHQ-8 × spring β_3 (SE)	PHQ-8 × summer, β_3 (SE)	PHQ-8 × autumn, β_3 (SE)	Likelihood ratio test (*P* values)
Sleep duration	0.49 (0.18)**	–4.55 (2.49)	–16.27 (2.45)***	–7.2 (2.46)**	–0.20 (0.22)	–0.04 (0.22)	0.11 (0.21)	.50
Sleep onset	0.79 (0.22)***	4.57 (3.04)	11.14 (2.98)***	0.55 (3)	–0.09 (0.26)	–0.41 (0.26)	–0.53 (0.26)*	.49
Sleep offset	1.01 (0.20)***	–2.88 (2.71)	–7.28 (2.67)**	–9.74 (2.68)***	0.27 (0.24)	0.04 (0.23)	0.15 (0.23)	.53
Sleep variability	1.23 (0.19)***	–4.32 (2.56)	0.19 (2.52)	–1.43 (2.53)	–0.42 (0.22)	–0.57 (0.22)**	–0.14 (0.22)	.26
Daily step	–73.21 (8.55)***	289.04 (115.3)*	804.05 (113.32)***	311.21 (113.92)**	–31.51 (10.01)**	–42.61 (9.95)***	–12.66 (9.87)	<.001
Step IV[d]	–0.001 (0.0008)	–0.02 (0.01)	–0.06 (0.01)***	–0.04 (0.01)***	0.002 (0.001)*	0.004 (0.001)***	0.002 (0.001)	<.001
Step IS[e]	–0.001 (0.0003)**	0.01 (0.005)*	–0.001 (0.004)	0.005 (0.004)	–0.001 (0.0004)*	–0.001 (0.0004)	–0.0001 (0.0004)	<.001
L5 onset[f]	0.51 (0.29)	–4.11 (4.02)	1.26 (3.95)	–2.89 (3.98)	0.31 (0.35)	–0.54 (0.35)	–0.11 (0.34)	.23
M10 onset[g]	0.26 (0.34)	4.86 (4.63)	10.33 (4.54)*	–0.44 (4.57)	0.72 (0.41)	1.06 (0.40)**	0.21 (0.40)	.04
HR[h] MESOR[i]	–0.03 (0.01)*	–0.57 (0.16)***	0.05 (0.16)	0.28 (0.16)	0.01 (0.01)	–0.01 (0.01)	–0.03 (0.01)	.003
HR amplitude	–0.04 (0.01)***	0.36 (0.1)***	0.87 (0.1)***	0.38 (0.1)***	–0.02 (0.01)*	0.01 (0.01)	0.002 (0.008)	.001
HR acrophase	0.50 (0.30)	40.39 (4.05)***	62.99 (3.98)***	23.3 (4)***	0.30 (0.35)	0.52 (0.35)	0.08 (0.35)	.75

[a]The exact *P* values are reported in the main text.

[b]PHQ-8: 8-item Patient Health Questionnaire.

[c]Model 3: Circadian rhythm = β_1 PHQ-8 + β_2 season + β_3 PHQ-8 × season + COVs, where COVs represents covariates mentioned in the Methods section. Note that winter is the reference season.

[d]IV: intradaily variability.

[e]IS: interdaily stability.

[f]L5 onset: the onset of least active continuous 5-hour period.

[g]M10 onset: the onset of the most active continuous 10-hour period.

[h]HR: heart rate.

[i]HR MESOR: midline estimating statistic of the fitted cosine wave for heart rate.

*P<.05, **P<.01, ***P<.001.

Seasonal Effects on Wearable-Measured Circadian Rhythm Features

Our results reveal significant seasonal impacts on various circadian rhythms, as indicated by the seasonal coefficients in model 2 (Table 2). Notably, HR acrophase varied significantly across seasons, delayed by 43.4 minutes in spring (P<.001), 67.9 minutes in summer (P<.001), and 24.1 minutes in autumn (P<.001) compared with winter (the reference season).

In terms of rest-activity rhythms, compared with winter, summer was associated with 394.5 more daily steps (P<.001) and 0.03 lower step IV (P<.001), and M10 onset was 20.5 minutes later (P<.001); autumn was associated with 186.2 more daily steps (P=.002) and 0.02 lower step IV (P<.001); and spring was associated M10 onset being 11.9 minutes later (P<.001).

Regarding sleep-wake rhythms, compared with winter, we found that (1) sleep duration decreased by 6.5 minutes in spring (P<.001), 16.6 minutes in summer (P<.001), and 6.1 minutes in autumn (P<.001); (2) sleep onset was 4.6 minutes earlier in autumn (P=.004), 3.7 minutes later in spring (P=0.02), and 7.2 minutes later in summer (P<.001); (3) sleep offset was 6.9 minutes earlier in summer (P<.001) and 8.3 minutes earlier in autumn (P<.001); and (4) sleep variability was 8.5 minutes lower in spring (P<.001), 5.3 minutes lower in summer (P<.001), and 2.8 minutes lower in autumn (P=.04).

XSL•FO
RenderX

Site-to-Site Analysis

Results of our subset analyses across 3 study sites are shown in Tables S1-S3 in Multimedia Appendix 1. Despite a reduction in sample sizes leading to some nonsignificant coefficients, the overall direction of associations between the PHQ-8 and circadian rhythm features remained consistent with the results in the entire data set. Notably, the seasonal changes in sleep-related features and HR acrophase are larger in the Spain site than the sites in the United Kingdom and Netherlands. For instance, during summer compared with winter, participants from the Spain site experienced a significant decrease in sleep duration by 36.6 minutes, whereas participants from the sites in the United Kingdom and Netherlands experienced reductions in sleep duration by 18.9 and 8.5 minutes, respectively. Furthermore, compared with winter, the HR acrophase for participants in Spain was delayed by 123.2 minutes in summer, significantly more than the delays of 58.2 minutes in the UK site and 64.0 minutes in the Netherlands site during summer.

Pre–COVID-19 Subset Analysis

The pre–COVID-19 subset analysis (Table 4) indicated that the directions and significances of associations between depression severity and circadian rhythm features were consistent with results in the entire data set. The seasonal effects on most circadian rhythm features (except for daily step) were similar between the entire data set and the pre–COVID-19 subset. Notably, daily step exhibited larger seasonal fluctuations in the pre–COVID-19 period. Specifically, compared with winter, participants in the pre–COVID-19 subset exhibited 657.2, 1096.4, and 321.6 more daily steps in spring, summer, and autumn, respectively.

Table 4. Coefficients, SE, and significance[a] of PHQ-8[b], seasonality, and the interaction term between PHQ-8 and seasons in model 3[c] for the pre-COVID subset.

Feature	PHQ-8, β_1 (SE)	Spring, β_2 (SE)	Summer, β_2 (SE)	Autumn, β_2 (SE)	PHQ-8 × spring, β_3 (SE)	PHQ-8 × summer, β_3 (SE)	PHQ-8 × autumn, β_3 (SE)
Sleep duration	0.97 (0.26)***	–3.01 (3.99)	–14.86 (3.67)***	–6.88 (3.45)*	–0.01 (0.34)	–0.05 (0.31)	0.11 (0.29)
Sleep onset	0.87 (0.36)*	–2.1 (5.46)	9.84 (5.02)	0.58 (4.73)	–0.25 (0.46)	–0.51 (0.43)	–0.64 (0.39)
Sleep offset	1.43 (0.31)***	–6.17 (4.65)	–7.6 (4.27)	–10.04 (4.03)*	–0.24 (0.39)	–0.22 (0.36)	0.07 (0.34)
Sleep variability	1.47 (0.26)***	1.94 (4.09)	2.29 (3.76)	0.62 (3.55)	–0.76 (0.35)*	–0.89 (0.32)**	–0.33 (0.3)
Daily step	–71.62 (12.05)***	657.17 (179.07)***	1096.36 (164.67)***	321.63 (155.06)*	–20.45 (15.21)	–48.65 (13.99)***	1.49 (12.91)
Step IV[d]	–0.001 (0.001)	–0.04 (0.02)*	–0.03 (0.02)	–0.01 (0.02)	0.004 (0.002)*	0.003 (0.001)*	0.0007 (0.0013)
Step IS[e]	–0.0003 (0.0005)	–0.004 (0.0072)	0.0021 (0.0066)	0.0035 (0.0063)	–0.0003 (0.0006)	–0.0007 (0.0006)	–0.0002 (0.0005)
L5 onset[f]	0.57 (0.45)	–17.84 (7.16)*	–6.74 (6.58)	–2.26 (6.22)	0.4 (0.61)	–0.2 (0.56)	–0.17 (0.52)
M10 onset[g]	0.09 (0.51)	–7.37 (7.89)	4.83 (7.25)	1.28 (6.84)	1.14 (0.67)	1.22 (0.62)*	0.07 (0.57)
HR[h] MESOR[i]	–0.08 (0.02)***	–0.4 (0.27)	–0.53 (0.25)*	–0.38 (0.23)	0.01 (0.02)	0.04 (0.02)*	0.04 (0.02)*
HR amplitude	–0.05 (0.01)***	0.31 (0.17)	0.58 (0.15)***	0.07 (0.14)	–0.01 (0.01)	0.02 (0.01)	0.02 (0.01)
HR acrophase	0.52 (0.38)	35.11 (5.77)***	68.14 (5.3)***	22.2 (5.0)***	1.15 (0.49)*	0.26 (0.45)	0.19 (0.42)

[a]The exact *P* values are reported in the main text.

[b]PHQ-8: 8-item Patient Health Questionnaire.

[c]Model 3: Circadian rhythm = β_1 PHQ-8 + β_2 season + β_3 PHQ-8 × season + COVs, where COVs represents covariates mentioned in the Methods section.

[d]IV: intradaily variability.

[e]IS: interdaily stability.

[f]L5 onset: the onset of least active continuous 5-hour period.

[g]M10 onset: the onset of the most active continuous 10-hour period.

[h]HR: heart rate.

[i]HR MESOR: midline estimating statistic of the fitted cosine wave for heart rate.

*P<.05, **P<.01, ***P<.001.

XSL•FO
RenderX

Discussion

Primary Findings

Our study, derived from a large European longitudinal mHealth study focused on depression, revealed significant seasonal effects on wearable-measured circadian rhythms and their associations with depression.

One of our key findings is that the associations between depression severity and certain wearable-measured circadian rhythms varied across different seasons. Specifically, we found a stronger negative association between depression and daily step count during the warmer seasons of spring and summer compared with winter. Additionally, a significant positive association was identified between depression severity and the onset of the most active continuous 10-hour period exclusively in summer. This may be attributed to the more favorable weather conditions for outdoor activities in summer [64], which could increase the observable relationship between rest-activity rhythms and depression. Furthermore, we found that the inclusion of seasonal effects can significantly improve the model fit for all wearable-measured circadian rhythms. These findings highlight the critical importance of including seasonal effects in longitudinal mHealth studies.

On adjusting for seasonal effects, we found that higher depression severity was significantly associated with lower levels of physical activity (daily step), more irregular activities (sleep variability and step IS), and later circadian rhythm timings (sleep onset and offset, M10 onset, and HR acrophase). These relationships are consistent with prior research. The linkage between reduced physical activity levels and worsened depression severity was reported in both survey-based and mHealth studies [65,66]. The preventive and therapeutic roles of physical activity against depression have been well-documented [67,68]. Moreover, there is extensive evidence linking depression with more irregular daily behaviors, including increased sleep variability [51,69-71], less stable rest-activity rhythms [72], and irregular patterns in Bluetooth [52] and GPS [73] data. The delayed circadian rhythm timings have been found to be associated with higher depression severity in multiple data streams, such as sleep recordings [51,69-71], activity logs [61], and HR recordings [74]. These consistent findings highlight the close associations between circadian rhythms and depression severity.

This study also revealed significant seasonal impacts on circadian rhythms, most notably between summer and winter. The cohort demonstrated shorter and later sleep patterns, increased daily step counts, reduced intradaily step variability, and delayed circadian rhythm timings during summer compared with winter. Especially, the circadian phase (HR acrophase) was delayed by 67.9 minutes in summer. These findings mostly align with previous laboratory and survey-based studies, which suggested that people tend to sleep longer in winter, possibly due to the effects of light exposure on melatonin production [75-78]. Previous studies have similarly observed higher activity levels [36,38,39], lower intradaily variability [79,80], and delayed hormone secretion rhythm [81,82] in summer compared with winter, further supporting our findings.

Conducted on a large cohort over an extended period using 3 wearable data modalities, this study aligned with some previous survey and laboratory findings, highlighting the precision of mobile technology in tracking behavioral rhythms. Our findings suggest that wearable-measured circadian rhythms have the potential to be digital biomarkers for depression detection. Furthermore, this study could inform the future design of seasonal, context-sensitive mHealth interventions, such as tailored activity recommendations and light exposure therapy.

Limitations

This study has several limitations that may impact its findings and interpretations. First, the presence of missing data may introduce bias, as our previous study indicated data compliance is associated with depression severity and other personal traits (eg, age) [83]. Second, the specificity of our cohort, characterized by a history of depression and a predominantly female composition, may limit the generalizability of our findings, highlighting the need for validation in more general populations. Third, while we accounted for national lockdowns as a covariate in the regression models and conducted analysis on the pre–COVID-19 subset, the complexities and varied impacts of COVID-19 necessitate additional validations in the postpandemic data sets. Fourth, our study relied on PHQ-8 scores as depression labels, which may introduce subjective biases.

Lastly, while we adjusted for study sites (in different countries) as a covariate in our regression models and conducted the site-to-site analyses, the influence of geographical location on circadian rhythms and their relationship is still needed to be explored in future research, ideally through the randomized controlled trials.

Conclusions

Our analysis of longitudinal wearable data from a large cohort reveals the significant seasonal impact on circadian rhythms and their associations with depression, emphasizing the importance of accounting for seasonal variations in longitudinal mHealth research. Additionally, we found that wearable-measured circadian rhythms are significantly linked to depression severity, indicating their potential to be digital biomarkers for depression detection. These findings enrich our understanding of the mechanisms and pathology underlying depression and could inform the future design of mental health monitoring and interventions.

Acknowledgments

The Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement No. 115902. This Joint Undertaking receives

XSL•FO
RenderX

support from the European Union's Horizon 2020 Research and Innovation Program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). This communication reflects the views of the RADAR-CNS consortium, and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding bodies have not been involved in the design of the study, the collection or analysis of data, or the interpretation of data. This study represents independent research partly funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre (BRC) at South London, and Maudsley National Health Service (NHS) Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This research was reviewed by a team with experience of mental health problems and their careers, who have been specially trained to advise on research proposals and documentation through Feasibility and Acceptability Support Team for Researchers (FAST-R), a free, confidential service in England provided by the NIHR Maudsley BRC via King's College London and South London and Maudsley NHS Foundation Trust. We thank all GLAD Study volunteers for their participation and gratefully acknowledge the NIHR BioResource, NIHR BioResource centers, NHS Trusts, and staff for their contribution. We also acknowledge NIHR BRC, King's College London, South London and Maudsley NHS Trust, and King's Health Partners. We thank the NIHR, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Program. Participants in the CIBER site came from the following 4 clinical communities in Spain: Parc Sanitari Sant Joan de Déu Network services, Institut Català de la Salut, Institut Pere Mata, and Hospital Clínico San Carlos. Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl, a Dutch online registry that facilitates participant recruitment for neuroscience studies. Hersenonderzoek.nl is funded by ZonMwMemorabel (project No. 73305095003), a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer's Society in the Netherlands, and Brain Foundation Netherlands. RJBD is supported by the following: (1) NIHR BRC at South London and Maudsley NHS Foundation Trust and King's College London; (2) Health Data Research UK, which is funded by the UK Medical Research Council (MRC), Engineering and Physical Sciences Research Council (EP/Y035216/1), Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) the BigData@Heart Consortium, funded by the IMI 2 Joint Undertaking (which receives support from the European Union's Horizon 2020 research and innovation program and EFPIA, partnering with 20 academic and industry partners and European Society of Cardiology); (4) the NIHR University College London Hospitals BRC; (5) the NIHR BRC at South London and Maudsley (related to attendance at the American Medical Informatics Association) NHS Foundation Trust, and King's College London; (6) the UK Research and Innovation (UKRI) London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare (AI4VBH); and (7) the NIHR Applied Research Collaboration (ARC) South London at King's College Hospital NHS Foundation Trust.

Conflicts of Interest

SV is an employee of Janssen Research and Development LLC. VAN was employed by Janssen Research and Development LLC during this study. PA was employed by the pharmaceutical company H. Lundbeck A/S during this study. DCM has accepted honoraria and consulting fees from Apple Inc, Otsuka Pharmaceuticals, Pear Therapeutics, and the One Mind Foundation; received royalties from Oxford Press; and an ownership interest in Adaptive Health Inc. MH is the principal investigator of the Remote Assessment of Disease and Relapse–Central Nervous System project, a private public precompetitive consortium that receives funding from Janssen, UCB, Lundbeck, MSD, and Biogen. CO is supported by the UK Medical Research Council (MR/N013700/1) and King's College London, member of the MRC Doctoral Training Partnership in Biomedical Sciences. AAS holds shares in Alphabet, owner of FitBit devices. All other authors declare no competing interests.

Multimedia Appendix 1
Results of the subset analyses across 3 study sites.
[DOCX File , 23 KB - jmir_v26i1e55302_app1.docx]

References

1. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS Med 2013;10(11):e1001547 [FREE Full text] [doi: 10.1371/journal.pmed.1001547] [Medline: 24223526]
2. Depression and other common mental disorders: global health estimates. WHO: World Health Organization; 2017.
3. Lenox-Smith A, Macdonald MTB, Reed C, Tylee A, Peveler R, Quail D, et al. Quality of life in depressed patients in UK primary care: the FINDER study. Neurol Ther 2013;2(1-2):25-42 [FREE Full text] [doi: 10.1007/s40120-013-0006-1] [Medline: 26000214]
4. Lerner D, Adler DA, Chang H, Berndt ER, Irish JT, Lapitsky L, et al. The clinical and occupational correlates of work productivity loss among employed patients with depression. J Occup Environ Med 2004;46(6 Suppl):S46-S55 [FREE Full text] [doi: 10.1097/01.jom.0000126684.82825.0a] [Medline: 15194895]

XSL•FO
RenderX

5. Lewinsohn PM, Solomon A, Seeley JR, Zeiss A. Clinical implications of "subthreshold" depressive symptoms. J Abnorm Psychol 2000;109(2):345-351. [Medline: 10895574]
6. Hawton K, Casañas I Comabella C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013;147(1-3):17-28. [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
7. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]
8. Devaux M, Sassi F. Social disparities in hazardous alcohol use: self-report bias may lead to incorrect estimates. Eur J Public Health 2016;26(1):129-134 [FREE Full text] [doi: 10.1093/eurpub/ckv190] [Medline: 26585784]
9. Wells KB, Hays RD, Burnam MA, Rogers W, Greenfield S, Ware JE. Detection of depressive disorder for patients receiving prepaid or fee-for-service care. Results from the Medical Outcomes Study. JAMA 1989;262(23):3298-3302. [Medline: 2585674]
10. Schulberg HC, Saul M, McClelland M, Ganguli M, Christy W, Frank R. Assessing depression in primary medical and psychiatric practices. Arch Gen Psychiatry 1985;42(12):1164-1170. [doi: 10.1001/archpsyc.1985.01790350038008] [Medline: 4074109]
11. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015;38(3):218-226 [FREE Full text] [doi: 10.1037/prj0000130] [Medline: 25844912]
12. Harman JS, Edlund MJ, Fortney JC. Disparities in the adequacy of depression treatment in the United States. Psychiatr Serv 2004;55(12):1379-1385. [doi: 10.1176/appi.ps.55.12.1379] [Medline: 15572565]
13. Young AS, Klap R, Sherbourne CD, Wells KB. The quality of care for depressive and anxiety disorders in the United States. Arch Gen Psychiatry 2001;58(1):55-61. [doi: 10.1001/archpsyc.58.1.55] [Medline: 11146758]
14. Carpenter JS, Crouse JJ, Scott EM, Naismith SL, Wilson C, Scott J, et al. Circadian depression: a mood disorder phenotype. Neurosci Biobehav Rev 2021;126:79-101. [doi: 10.1016/j.neubiorev.2021.02.045] [Medline: 33689801]
15. Crouse JJ, Carpenter JS, Song YJC, Hockey SJ, Naismith SL, Grunstein RR, et al. Circadian rhythm sleep-wake disturbances and depression in young people: implications for prevention and early intervention. Lancet Psychiatry 2021;8(9):813-823. [doi: 10.1016/S2215-0366(21)00034-1] [Medline: 34419186]
16. Partch CL, Green CB, Takahashi JS. Molecular architecture of the mammalian circadian clock. Trends Cell Biol 2014;24(2):90-99 [FREE Full text] [doi: 10.1016/j.tcb.2013.07.002] [Medline: 23916625]
17. Foster RG, Kreitzman L. The rhythms of life: what your body clock means to you!. Exp Physiol 2013;99(4):599-606. [doi: 10.1113/expphysiol.2012.071118] [Medline: 24363383]
18. Refinetti R. Circadian rhythmicity of body temperature and metabolism. Temperature (Austin) 2020;7(4):321-362 [FREE Full text] [doi: 10.1080/23328940.2020.1743605] [Medline: 33251281]
19. Vetter C, Devore EE, Wegrzyn LR, Massa J, Speizer FE, Kawachi I, et al. Association between rotating night shift work and risk of coronary heart disease among women. JAMA 2016;315(16):1726-1734 [FREE Full text] [doi: 10.1001/jama.2016.4454] [Medline: 27115377]
20. Knutsson A, Kempe A. Shift work and diabetes--a systematic review. Chronobiol Int 2014;31(10):1146-1151. [doi: 10.3109/07420528.2014.957308] [Medline: 25290038]
21. Musiek ES, Bhimasani M, Zangrilli MA, Morris JC, Holtzman DM, Ju YES. Circadian rest-activity pattern changes in aging and preclinical Alzheimer disease. JAMA Neurol 2018;75(5):582-590 [FREE Full text] [doi: 10.1001/jamaneurol.2017.4719] [Medline: 29379963]
22. Duffy JF, Dijk DJ. Getting through to circadian oscillators: why use constant routines? J Biol Rhythms 2002;17(1):4-13. [doi: 10.1177/074873002129002294] [Medline: 11837947]
23. Keijzer H, Smits MG, Duffy JF, Curfs LMG. Why the dim light melatonin onset (DLMO) should be measured before treatment of patients with circadian rhythm sleep disorders. Sleep Med Rev 2014;18(4):333-339. [doi: 10.1016/j.smrv.2013.12.001] [Medline: 24388969]
24. Bowman C, Huang Y, Walch OJ, Fang Y, Frank E, Tyler J, et al. A method for characterizing daily physiology from widely used wearables. Cell Rep Methods 2021;1(4):100058 [FREE Full text] [doi: 10.1016/j.crmeth.2021.100058] [Medline: 34568865]
25. Cui S, Lin Q, Gui Y, Zhang Y, Lu H, Zhao H, et al. CARE as a wearable derived feature linking circadian amplitude to human cognitive functions. NPJ Digit Med 2023;6(1):123 [FREE Full text] [doi: 10.1038/s41746-023-00865-0] [Medline: 37433859]
26. Zapata-Lamana R, Lalanza JF, Losilla JM, Parrado E, Capdevila L. mHealth technology for ecological momentary assessment in physical activity research: a systematic review. PeerJ 2020;8:e8848 [FREE Full text] [doi: 10.7717/peerj.8848] [Medline: 32257648]
27. Smagula SF, Krafty RT, Thayer JF, Buysse DJ, Hall MH. Rest-activity rhythm profiles associated with manic-hypomanic and depressive symptoms. J Psychiatr Res 2018;102:238-244 [FREE Full text] [doi: 10.1016/j.jpsychires.2018.04.015] [Medline: 29705489]

XSL•FO
RenderX

28. White KH, Rumble ME, Benca RM. Sex differences in the relationship between depressive symptoms and actigraphic assessments of sleep and rest-activity rhythms in a population-based sample. Psychosom Med 2017;79(4):479-484 [FREE Full text] [doi: 10.1097/PSY.0000000000000434] [Medline: 27922568]
29. Slyepchenko A, Allega OR, Leng X, Minuzzi L, Eltayebani MM, Skelly M, et al. Association of functioning and quality of life with objective and subjective measures of sleep and biological rhythms in major depressive and bipolar disorder. Aust N Z J Psychiatry 2019;53(7):683-696. [doi: 10.1177/0004867419829228] [Medline: 30759998]
30. Robillard R, Hermens DF, Naismith SL, White D, Rogers NL, Ip TKC, et al. Ambulatory sleep-wake patterns and variability in young people with emerging mental disorders. J Psychiatry Neurosc 2015;40(1):28-37. [doi: 10.1503/jpn.130247] [Medline: 25203899]
31. Carr O, Saunders KEA, Bilderbeck AC, Tsanas A, Palmius N, Geddes JR, et al. Desynchronization of diurnal rhythms in bipolar disorder and borderline personality disorder. Transl Psychiatry 2018;8(1):79 [FREE Full text] [doi: 10.1038/s41398-018-0125-7] [Medline: 29643339]
32. Ávila Moraes C, Cambras T, Diez-Noguera A, Schimitt R, Dantas G, Levandovski R, et al. A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters. BMC Psychiatry 2013;13:77 [FREE Full text] [doi: 10.1186/1471-244X-13-77] [Medline: 23510455]
33. Wright KP, McHill AW, Birks BR, Griffin BR, Rusterholz T, Chinoy ED. Entrainment of the human circadian clock to the natural light-dark cycle. Curr Biol 2013;23(16):1554-1558 [FREE Full text] [doi: 10.1016/j.cub.2013.06.039] [Medline: 23910656]
34. Khalsa SBS, Jewett ME, Cajochen C, Czeisler CA. A phase response curve to single bright light pulses in human subjects. J Physiol 2003;549(3):945-952 [FREE Full text] [doi: 10.1113/jphysiol.2003.040477] [Medline: 12717008]
35. Minors DS, Waterhouse JM, Wirz-Justice A. A human phase-response curve to light. Neurosci Lett 1991;133(1):36-40. [doi: 10.1016/0304-3940(91)90051-t] [Medline: 1791996]
36. Tucker P, Gilliland J. The effect of season and weather on physical activity: a systematic review. Public Health 2007;121(12):909-922. [doi: 10.1016/j.puhe.2007.04.009] [Medline: 17920646]
37. Mattingly SM, Grover T, Martinez GJ, Aledavood T, Robles-Granda P, Nies K, et al. The effects of seasons and weather on sleep patterns measured through longitudinal multimodal sensing. NPJ Digit Med 2021;4(1):76 [FREE Full text] [doi: 10.1038/s41746-021-00435-2] [Medline: 33911176]
38. Hamilton SL, Clemes SA, Griffiths PL. UK adults exhibit higher step counts in summer compared to winter months. Ann Hum Biol 2008;35(2):154-169. [doi: 10.1080/03014460801908058] [Medline: 18428010]
39. Clemes SA, Hamilton SL, Griffiths PL. Summer to winter variability in the step counts of normal weight and overweight adults living in the UK. J Phys Act Health 2011;8(1):36-44. [doi: 10.1123/jpah.8.1.36] [Medline: 21297183]
40. Matcham F, Barattieri di San Pietro C, Bulgari V, de Girolamo G, Dobson R, Eriksson H, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019;19(1):1-11. [doi: 10.1186/s12888-019-2049-z]
41. Matcham F, Leightley D, Siddi S, Lamers F, White KM, Annas P, RADAR-CNS consortium. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022;22(1):136 [FREE Full text] [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
42. Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, Hyve, RADAR-CNS Consortium. RADAR-Base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019;7(8):e11734 [FREE Full text] [doi: 10.2196/11734] [Medline: 31373275]
43. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
44. Huang YL, Liu RY, Wang QS, Van Someren EJW, Xu H, Zhou JN. Age-associated difference in circadian sleep-wake and rest-activity rhythms. Physiol Behav 2002;76(4-5):597-603. [doi: 10.1016/s0031-9384(02)00733-3] [Medline: 12126998]
45. Duffy JF, Cain SW, Chang AM, Phillips AJK, Münch MY, Gronfier C, et al. Sex difference in the near-24-hour intrinsic period of the human circadian timing system. Proc Natl Acad Sci USA 2011;108 Suppl 3(Suppl 3):15602-15608 [FREE Full text] [doi: 10.1073/pnas.1010666108] [Medline: 21536890]
46. Jones SH, Hare DJ, Evershed K. Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder. Bipolar Disord 2005;7(2):176-186. [doi: 10.1111/j.1399-5618.2005.00187.x] [Medline: 15762859]
47. Sun S, Folarin AA, Ranjan Y, Rashid Z, Conde P, Stewart C, RADAR-CNS Consortium. Using smartphones and wearable devices to monitor behavioral changes during COVID-19. J Med Internet Res 2020;22(9):e19992 [FREE Full text] [doi: 10.2196/19992] [Medline: 32877352]
48. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
49. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537 [FREE Full text] [doi: 10.7717/peerj.2537] [Medline: 28344895]

XSL•FO
RenderX

50. Farhan AA, Yue C, Morillo R, Ware S, Lu J, Bi J, et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. : IEEE; 2016 Presented at: 2016 IEEE Wireless Health (WH); October 25-27 2016; Bethesda, MD, USA. [doi: 10.1109/wh.2016.7764553]
51. Zhang Y, Folarin AA, Sun S, Cummins N, Bendayan R, Ranjan Y, RADAR-CNS Consortium. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021;9(4):e24604 [FREE Full text] [doi: 10.2196/24604] [Medline: 33843591]
52. Zhang Y, Folarin AA, Sun S, Cummins N, Ranjan Y, Rashid Z, et al. Predicting depressive symptom severity through individuals' nearby Bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021;9(7):e29840 [FREE Full text] [doi: 10.2196/29840] [Medline: 34328441]
53. Sun S, Folarin AA, Zhang Y, Cummins N, Garcia-Dias R, Stewart C, RADAR-CNS Consortium. Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: retrospective analysis. J Med Internet Res 2023;25:e45233 [FREE Full text] [doi: 10.2196/45233] [Medline: 37578823]
54. Harrison Y. The impact of daylight saving time on sleep and related behaviours. Sleep Med Rev 2013;17(4):285-292. [doi: 10.1016/j.smrv.2012.10.001] [Medline: 23477947]
55. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. Chronobiol Int 2017;35(4):465-476. [doi: 10.1080/07420528.2017.1413578] [Medline: 29235907]
56. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis. J Med Internet Res 2019;21(11):e16273 [FREE Full text] [doi: 10.2196/16273] [Medline: 31778122]
57. van Someren EJ, Hagebeuk EE, Lijzenga C, Scheltens P, de Rooij SE, Jonker C, et al. Circadian rest-activity rhythm disturbances in Alzheimer's disease. Biol Psychiatry 1996;40(4):259-270. [doi: 10.1016/0006-3223(95)00370-3] [Medline: 8871772]
58. Van Someren EJW, Swaab DF, Colenda CC, Cohen W, McCall WV, Rosenquist PB. Bright light therapy: improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods. Chronobiol Int 1999;16(4):505-518. [doi: 10.3109/07420529908998724] [Medline: 10442243]
59. Cornelissen G. Cosinor-based rhythmometry. Theor Biol Med Model 2014;11:16 [FREE Full text] [doi: 10.1186/1742-4682-11-16] [Medline: 24725531]
60. Ali FZ, Parsey RV, Lin S, Schwartz J, DeLorenzo C. Circadian rhythm biomarker from wearable device data is related to concurrent antidepressant treatment response. NPJ Digit Med 2023;6(1):81 [FREE Full text] [doi: 10.1038/s41746-023-00827-6] [Medline: 37120493]
61. Esaki Y, Obayashi K, Saeki K, Fujita K, Iwata N, Kitajima T. Association between circadian activity rhythms and mood episode relapse in bipolar disorder: a 12-month prospective cohort study. Transl Psychiatry 2021;11(1):525 [FREE Full text] [doi: 10.1038/s41398-021-01652-9] [Medline: 34645802]
62. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38(4):963-974. [Medline: 7168798]
63. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
64. Chen L, Ng E. Outdoor thermal comfort and outdoor activities: a review of research in the past decade. Cities 2012;29(2):118-125. [doi: 10.1016/j.cities.2011.08.006]
65. Melo MCA, Daher EDF, Albuquerque SGC, de Bruin VMS. Exercise in bipolar patients: a systematic review. J Affect Disord 2016 Jul 01;198:32-38. [doi: 10.1016/j.jad.2016.03.004] [Medline: 26998794]
66. Melo MCA, Garcia RF, de Araújo CFC, Rangel DM, de Bruin PFC, de Bruin VMS. Physical activity as prognostic factor for bipolar disorder: an 18-month prospective study. J Affect Disord 2019 May 15;251:100-106. [doi: 10.1016/j.jad.2019.03.061] [Medline: 30921592]
67. Harvey SB, Øverland S, Hatch SL, Wessely S, Mykletun A, Hotopf M. Exercise and the prevention of depression: results of the HUNT Cohort Study. Am J Psychiatry 2018;175(1):28-36. [doi: 10.1176/appi.ajp.2017.16111223] [Medline: 28969440]
68. Cooney GM, Dwan K, Greig CA, Lawlor DA, Rimer J, Waugh FR, et al. Exercise for depression. Cochrane Database Syst Rev 2013;2013(9):CD004366 [FREE Full text] [doi: 10.1002/14651858.CD004366.pub6] [Medline: 24026850]
69. Fang Y, Forger DB, Frank E, Sen S, Goldstein C. Day-to-day variability in sleep parameters and depression risk: a prospective cohort study of training physicians. NPJ Digit Med 2021 Feb 18;4(1):28. [doi: 10.1038/s41746-021-00400-z] [Medline: 33603132]
70. Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C, et al. Objective assessment of depressive symptoms with machine learning and wearable sensors data. 2017 Presented at: International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII; October 23-26, 2017; San Antonio, TX. [doi: 10.1109/acii.2017.8273620]
71. Matcham F, Carr E, Meyer N, White KM. Wearable-derived sleep features predict relapse in major depressive disorder. ResearchGate. 2023 Mar. URL: https://www.researchgate.net/publication/369514383_Wearable-derived_sleep_features_predict_relapse_in_Major_Depressive_Disorder [accessed 2024-05-22]

XSL•FO
RenderX

72. Luik AI, Zuurbier LA, Direk N, Hofman A, Van Someren EJ, Tiemeier H. 24-hour activity rhythm and sleep disturbances in depression and anxiety: a population-based study of middle-aged and older persons. Depress Anxiety 2015;32(9):684-692. [doi: 10.1002/da.22355] [Medline: 25693731]
73. Zhang Y, Folarin AA, Sun S, Cummins N, Vairavan S, Bendayan R, RADAR-CNS consortium. Longitudinal relationships between depressive symptom severity and phone-measured mobility: dynamic structural equation modeling study. JMIR Ment Health 2022;9(3):e34898 [FREE Full text] [doi: 10.2196/34898] [Medline: 35275087]
74. Otsuka K, Mitsutake G, Yano S. Depression, quality of life, and lifestyle: chronoecological health watch in a community. Biomed Pharmacother 2002;56 Suppl 2:231s-242s. [doi: 10.1016/s0753-3322(02)00297-4] [Medline: 12653175]
75. Monsivais D, Ghosh A, Bhattacharya K, Dunbar RIM, Kaski K. Tracking urban human activity from mobile phone calling patterns. PLoS Comput Biol 2017;13(11):e1005824 [FREE Full text] [doi: 10.1371/journal.pcbi.1005824] [Medline: 29161270]
76. de la Iglesia HO, Fernández-Duque E, Golombek DA, Lanza N, Duffy JF, Czeisler CA, et al. Access to electric light is associated with shorter sleep duration in a Traditionally Hunter-Gatherer Community. J Biol Rhythms 2015;30(4):342-350 [FREE Full text] [doi: 10.1177/0748730415590702] [Medline: 26092820]
77. Yetish G, Kaplan H, Gurven M, Wood B, Pontzer H, Manger PR, et al. Natural sleep and its seasonal variations in three pre-industrial societies. Curr Biol 2015;25(21):2862-2868 [FREE Full text] [doi: 10.1016/j.cub.2015.09.046] [Medline: 26480842]
78. Cepeda M, Koolhaas CM, van Rooij FJA, Tiemeier H, Guxens M, Franco OH, et al. Seasonality of physical activity, sedentary behavior, and sleep in a middle-aged and elderly population: The Rotterdam study. Maturitas 2018;110:41-50. [doi: 10.1016/j.maturitas.2018.01.016] [Medline: 29563034]
79. Kume Y, Makabe S, Singha-Dong N, Vajamun P, Apikomonkon H, Griffiths J. Seasonal effects on the sleep-wake cycle, the rest-activity rhythm and quality of life for Japanese and Thai older people. Chronobiol Int 2017;34(10):1377-1387. [doi: 10.1080/07420528.2017.1372468] [Medline: 29064742]
80. Smagula SF, DuPont CM, Miller MA, Krafty RT, Hasler BP, Franzen PL, et al. Rest-activity rhythms characteristics and seasonal changes in seasonal affective disorder. Chronobiol Int 2018;35(11):1553-1559 [FREE Full text] [doi: 10.1080/07420528.2018.1496094] [Medline: 30024782]
81. Kanikowska D, Roszak M, Rutkowski R, Sato M, Sikorska D, Orzechowska Z, et al. Seasonal differences in rhythmicity of salivary cortisol in healthy adults. J Appl Physiol (1985) 2019;126(3):764-770 [FREE Full text] [doi: 10.1152/japplphysiol.00972.2018] [Medline: 30702977]
82. Hadlow NC, Brown S, Wardrop R, Henley D. The effects of season, daylight saving and time of sunrise on serum cortisol in a large population. Chronobiol Int 2013;31(2):243-251. [doi: 10.3109/07420528.2013.844162] [Medline: 24156521]
83. Zhang Y, Pratap A, Folarin AA, Sun S, Cummins N, Matcham F, RADAR-CNS consortium. Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. NPJ Digit Med 2023;6(1):25 [FREE Full text] [doi: 10.1038/s41746-023-00749-3] [Medline: 36806317]

Abbreviations

BRC: Biomedical Research Centre
EFPIA: European Federation of Pharmaceutical Industries and Associations
HR: heart rate
IMI: Innovative Medicines Initiative
mHealth: mobile health
MRC: Medical Research Council
NHS: National Health Service
NIHR: National Institute for Health Research
PHQ-8: 8-item Patient Health Questionnaire
RADAR-MDD: Remote Assessment of Disease and Relapse Major Depressive Disorder

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, S Ma; submitted 08.12.23; peer-reviewed by A Tacchino, JC Shin, D Singh; comments to author 13.02.24; revised version received 22.02.24; accepted 29.03.24; published 28.06.24.

Please cite as:
Zhang Y, Folarin AA, Sun S, Cummins N, Ranjan Y, Rashid Z, Stewart C, Conde P, Sankesara H, Laiou P, Matcham F, White KM, Oetzmann C, Lamers F, Siddi S, Simblett S, Vairavan S, Myin-Germeys I, Mohr DC, Wykes T, Haro JM, Annas P, Penninx BWJH, Narayan VA, Hotopf M, Dobson RJB, RADAR-CNS consortium
Longitudinal Assessment of Seasonal Impacts and Depression Associations on Circadian Rhythm Using Multimodal Wearable Sensing: Retrospective Analysis
J Med Internet Res 2024;26:e55302
URL: https://www.jmir.org/2024/1/e55302
doi:10.2196/55302
PMID:

©Yuezhou Zhang, Amos A Folarin, Shaoxiong Sun, Nicholas Cummins, Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Pauline Conde, Heet Sankesara, Petroula Laiou, Faith Matcham, Katie M White, Carolin Oetzmann, Femke Lamers, Sara Siddi, Sara Simblett, Srinivasan Vairavan, Inez Myin-Germeys, David C Mohr, Til Wykes, Josep Maria Haro, Peter Annas, Brenda WJH Penninx, Vaibhav A Narayan, Matthew Hotopf, Richard JB Dobson, RADAR-CNS consortium. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Web-Based Warfarin Management (Alfalfa App) Versus Traditional Warfarin Management: Multicenter Prospective Cohort Study

Wenfei Chen[1], BMed; Jiana Chen[1], MMES; Shaojun Jiang[1], MME; Chunhua Wang[2], DEd; Jinhua Zhang[1], DEd

[1]Department of Pharmacy, Fujian Maternity and Child Health Hospital College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, China

[2]Department of Neurosurgery, Fujian Medical University Union Hospital, fuzhou, China

Corresponding Author:
Jinhua Zhang, DEd
Department of Pharmacy
Fujian Maternity and Child Health Hospital College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics
Fujian Medical University
#18 Daoshan Road
Fuzhou, 350001
China
Phone: 86 13306912381
Email: pollyzhang2006@126.com

Abstract

Background: Poor anticoagulation management of warfarin may lead to patient admission, prolonged hospital stays, and even death due to anticoagulation-related adverse events. Traditional non–web-based outpatient clinics struggle to provide ideal anticoagulation management services for patients, and there is a need to explore a safer, more effective, and more convenient mode of warfarin management.

Objective: This study aimed to compare differences in the quality of anticoagulation management and clinical adverse events between a web-based management model (via a smartphone app) and the conventional non–web-based outpatient management model.

Methods: This study is a prospective cohort research that includes multiple national centers. Patients meeting the nadir criteria were split into a web-based management group using the Alfalfa app or a non–web-based management group with traditional outpatient management, and they were then monitored for a 6-month follow-up period to collect coagulation test results and clinical events. The effectiveness and safety of the 2 management models were assessed by the following indicators: time in therapeutic range (TTR), bleeding events, thromboembolic events, all-cause mortality events, cumulative event rates, and the distribution of the international normalized ratio (INR).

Results: This national multicenter cohort study enrolled 522 patients between June 2019 and May 2021, with 519 (99%) patients reaching the follow-up end point, including 260 (50%) in the non–web-based management group and 259 (50%) in the web-based management group. There were no observable differences in baseline characteristics between the 2 patient groups. The web-based management group had a significantly higher TTR than the non–web-based management group (82.4% vs 71.6%, $P<.001$), and a higher proportion of patients received effective anticoagulation management (81.2% vs 63.5%, $P<.001$). The incidence of minor bleeding events in the non–web-based management group was significantly higher than that in the web-based management group (12.1% vs 6.6%, $P=.048$). Between the 2 groups, there was no statistically significant difference in the incidence of severe bleeding and thromboembolic and all-cause death events. In addition, compared with the non–web-based management group, the web-based management group had a lower proportion of INR in the extreme subtreatment range (17.6% vs 21.3%) and severe supertreatment range (0% vs 0.8%) and a higher proportion in the treatment range (50.4% vs 43.1%), with statistical significance.

Conclusions: Compared with traditional non–web-based outpatient management, web-based management via the Alfalfa app may be more beneficial because it can enhance patient anticoagulation management quality, lower the frequency of small bleeding events, and improve INR distribution.

(J Med Internet Res 2024;26:e46319) doi:10.2196/46319

XSL•FO
RenderX

KEYWORDS

warfarin; telemedicine; smart phone application; anti-coagulation management; management; cohort study; application; chronic disease; support; effectiveness; online model; patient management

Introduction

With a usage history spanning over 60 years, warfarin is currently the most widely used anticoagulant prescription drug for preventing thrombosis after heart valve surgery, orthopedic surgery, atrial fibrillation, and varicose veins of the lower extremities, as well as systemic embolism, pulmonary embolism, and venous thromboembolism [1]. Daily management of warfarin is highly correlated with clinical outcomes. High-quality anticoagulation management can effectively reduce the risk of bleeding and thromboembolic events during dosing, while poor anticoagulation management may lead to hospital admission, prolonged hospital stay, and even patient death due to anticoagulation-related adverse events. The anticoagulant intensity of warfarin is measured by the international normalized ratio (INR), with different target ranges of INR for different anticoagulation indications. Warfarin's intensity of anticoagulation can be influenced by numerous variables. Along with the generally constant elements like genotype and gender, other variables may evolve with time, resulting in modifications in the pharmacokinetic and pharmacodynamic properties of warfarin in addition to changes in INR. At the same time, patients who take warfarin need to have INR testing done regularly and receive personalized dose adjustments to maintain the proper anticoagulant intensity owing to the drug's limited therapeutic window, the numerous factors that can influence its anticoagulant effect, and the wide range of individual differences. [2].

Currently, the quality of anticoagulation in Chinese patients receiving warfarin therapy is not satisfactory [3]. On the one hand, due to the relative scarcity of medical and health resources, as well as their unequal geographical distribution in China [4], physicians and pharmacists with specialized anticoagulation training are usually concentrated in urban areas, and patients in most rural areas lack convenient access to high-quality anticoagulation management. On the other hand, the traditional non–web-based management model requires patients to have good compliance, to be able to take their medication on schedule, and to be proactive in visiting the hospital regularly for INR testing and dose adjustment. However, in clinical practice, patients often miss medication and prolong the interval of blood examination by themselves. All these factors increase the risk of warfarin-related anticoagulant complications [5].

Therefore, it is necessary to explore a safer, more effective, and convenient mode of warfarin management for patients. We have previously conducted a retrospective study in a single center [3], but there are often issues with retrospective studies, such as low data completeness and accuracy, which may introduce selective bias. Therefore, we conducted a nationwide multicenter prospective cohort study to collect patient demographic information, coagulation test results, and clinical events and further compare whether there are any differences in the quality of anticoagulation management and cardiovascular event rates between web-based management provided through a smartphone app and traditional non–web-based outpatient management.

Methods

Study Design and Participants

We conducted a national multicenter prospective cohort study to explore the differences in quality of anticoagulation management as well as clinical adverse events between the web-based management model via the Alfalfa app and the traditional non–web-based outpatient management model. The study flow is shown in Figure 1. This study was reported in strict adherence to the RECORD-PE (Reporting of Studies Conducted Using Observational Routinely Collected Health Data Statement for Pharmacoepidemiology) checklist (Multimedia Appendix 1) [6]. Participants were enrolled at 5 central hospitals in China from June 2019 to May 2021, after approval by the ethical committees of each center. These 5 hospitals are in the southwest, southeast, and central parts of China, which are different in terms of geographic location, climate, and diet, and to a certain extent represent most parts of the country and are representative. The study was also registered with the Chinese Clinical Trial Registry (ChiCTR1900021920). At the time of registration, this study was designed as a randomized controlled trial, but during actual grouping, it was difficult to carry out randomized grouping because the patients were in the same ward and their desire to group on their own was so strong that it could only be designed as a prospective cohort study in the end. The associated study design has been previously published [7].

XSL•FO
RenderX

Figure 1. Flowchart of the study.

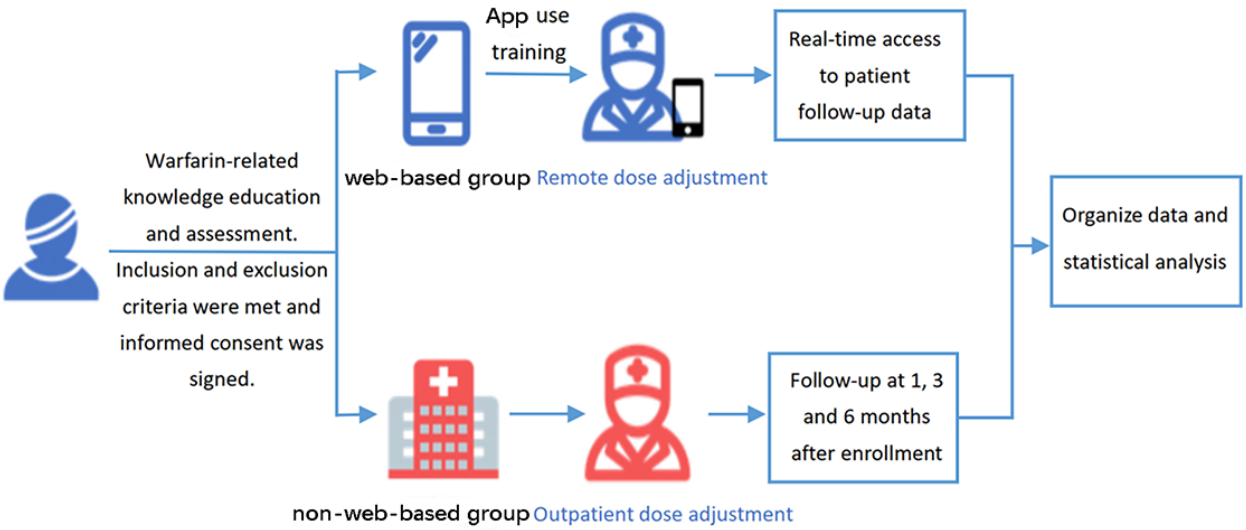

Ethical Considerations

This study was approved by the Ethics Committee of Union Hospital affiliated with Fujian Medical University (2019YF020-01). The local ethical committee of each hospital also provided ethics approval. All patients had to sign an informed consent form to participate.

Inclusion and Exclusion Criteria

The inclusion criteria were: (1) age ≥18 years; (2) patients with the following indications and requiring warfarin therapy for at least 3 months: receiving heart valve replacement or repair, atrial fibrillation confirmed by electrocardiogram or ambulatory electrocardiogram, receiving radio-frequency/cryoballoon ablation, and venous thromboembolism confirmed by venography or Doppler ultrasound; (3) willingness to learn to use and accept web-based or non–web-based warfarin administration; (4) willingness to be followed up; and (5) willing to sign patient consent and signed informed consent form.

To address potential sources of bias, we excluded patients who (1) were pregnant or planning to become pregnant; (2) taking or planning to take other oral anticoagulants; (3) had experienced therapeutic or subtherapeutic bleeding or thrombotic complications in the 6 months prior to enrollment; (4) expected to have an imminent procedure requiring warfarin interruption; (5) had severe renal insufficiency (endogenous creatinine clearance ≤20 mL/min); (6) had severe hepatic insufficiency (Child-Pugh score ≥ 10); (7) had severe heart failure (New York Heart Association heart function class IV and above); and (8) had severe infection and respiratory failure.

Patient Management

Patients in the web-based management group were managed remotely via a smartphone app, where the investigator provided dose adjustment protocols and had immediate access to patient follow-up data. The web-based management tool used in this study was the Alfalfa app (Fuzhou Alfalfa Health Management Co). The non–web-based management group was defined as patients who used the usual outpatient route for dose adjustment after being discharged from the hospital. The investigators plan to follow up with patients in the non–web-based management group at 1, 3, and 6 months after enrollment to collect coagulation test results and data related to clinical events.

Grouping and Pregrouping Education

For all patients who met the inclusion and exclusion criteria, the investigators provided between 3 and 4 warfarin dosing information sessions to the patients and their relatives prior to discharge and distributed electronic dosing education materials. The materials went over the effects of warfarin, medication purpose, medication precautions, indicators to be monitored during administration and their importance, the INR target range, the INR monitoring interval, factors affecting the efficacy of warfarin, and the potential side effects of warfarin.

Patients in the non–web-based management group were instructed by the investigator to attend the nearby hospital or clinic frequently for coagulation testing and to consult with their local physician for dose adjustment options. Patients and their relatives were also instructed to keep track of the results and dates of all previous INR tests, as well as any possible bleeding or thrombotic events that occurred during this period. Finally, patients were notified that they would be followed up at 1, 3, and 6 months with the expectation of patient cooperation.

Patients in the web-based management group were provided with extra training on operating the Alfalfa app, which focused on registering the software, main functions, procedures to upload coagulation results, and instructions on taking medication and checking blood according to the medical staff's response. The investigators ensured that patients or their relatives were able to operate the Alfalfa app independently and accurately.

Anticoagulation Management Program

INR Target Range

The INR target ranges for each anticoagulation indication in this study are shown in Table 1. If the patient had a combination of multiple anticoagulation indications, a higher INR target

XSL•FO
RenderX

range corresponding to the indications was used in the development of the dose adjustment protocol. It is worth noting that the target range of INR after heart valve surgery was based on clinical experience at the main central hospital and is not entirely consistent with the target range recommended by the Chinese Guidelines for the Prevention and Treatment of Thrombotic Diseases. This INR target range has been applied to over 40,000 patients for more than 30 years and can ensure the security and efficacy of anticoagulation therapy.

Table 1. INR[a] target ranges for different indications.

Anticoagulation indication	INR target range
Aortic valve replacement or repair	1.5-2
Mitral valve replacement or repair	1.7-2.5
Tricuspid valve replacement or repair	2-2.5
Atrial fibrillation	2-3
Venous thromboembolism	2-3

[a]INR: international normalized ratio.

INR Monitoring Frequency

Each patient's initial INR was checked weekly after discharge. When 2 consecutive INR values were within the patient's target range, their INR testing interval could be extended by 1 week (ie, once every 2 weeks). With biweekly INR values, if 2 consecutive INR values were within the patient's target range, the blood test interval could be extended by 1 week (ie, every 3 weeks), and so on. The maximum interval between blood tests could not exceed 1 month (ie, the INR value should be tested at least once a month). If the warfarin dose needed adjustment, the process returned to weekly blood checks and was repeated until the dose stabilized again. If a patient reported extreme INR values, the anticoagulation management team planned to empirically adjust the INR monitoring interval based on previous INR records.

Dose Adjustment Scheme

The dose adjustment protocol for warfarin is empirically developed by clinical pharmacists or physicians based on information such as INR results, comorbidities and medications, and clinical events with reference to the Chinese Expert Consensus on Warfarin Anticoagulation Therapy. The general principles of dose adjustment are as follows：(1) if the INR value is within ± 0.2 of the upper and lower limits of the target range, the warfarin dose remains unchanged; (2) if INR <lower limit of the target range, 0.2, warfarin increases by 0.25 tablets; (3) if the upper limit of target range + 0.2 <INR ≤3, warfarin decreases by 0.25 tablets; (4) If INR >3, stop the drug for 1 day and recheck INR on the next day; (5) if INR ≤3 on the next day, reduce 0.25 tablets and repeat the aforementioned process; and (6) if INR >3 on the next day, continue to stop the drug until INR ≤3.

Management Style

Patients in the web-based management group were managed remotely with the Alfalfa app based on the WeChat public platform for anticoagulation (see Multimedia Appendix 2 [8-12] for details on the Alfalfa app). The anticoagulation management team at each central hospital consists of clinical pharmacists and physicians specializing in anticoagulation. They were responsible for providing dose adjustment and medication counseling to patients in the web-based management group. Web-based management was achieved through the following steps. First, the patient patients had to inform the physician or pharmacists of the results of the coagulation report as well as recent physical and dietary conditions, concomitant medications, adverse reactions, and other information. Second, the anticoagulation management team members responded to the patient with the recommended dose and time of the next blood test based on the information reported by the patient. Finally, patients took their medication as prescribed and performed their next coagulation test on time. If a patient reported an extreme INR value or clinical event, the medical staff followed up with the patient by phone to ask about their medical condition and advise on the proper management. Patients in the non–web-based management group underwent dose adjustments at their local hospital, with specific anticoagulation management protocols developed by their local physician or pharmacist.

Outcome Indicators

The time in therapeutic range (TTR), this study's primary outcome metric, was calculated using the linear interpolation method proposed by Professor Rosendaal. The secondary outcome indicators were the occurrence of clinical events and the distribution of INR. The clinical event focused on evaluating the differences in safety and effectiveness between web-based and traditional non–web-based outpatient management. Safety outcomes referred to bleeding events and were classified as minor bleeding events and serious bleeding events. The former included epistaxis, gingival or oral bleeding, skin ecchymosis, fundus hemorrhage, menorrhagia or prolongation, and hematuria. Using the classification criteria of the International Society on Thrombosis and Hemostasis (ISTH) [13], the latter was defined as any hemorrhage requiring hospitalization or blood transfusion, including gastrointestinal bleeding and intracranial hemorrhage. Effective outcomes were thromboembolic events, including transient ischemic attack, venous thromboembolism, valve thrombosis, and ischemic stroke. Other clinical events included hospitalizations and emergencies associated with warfarin. Likewise, the distribution of INR was used as part of the secondary outcome indicators in this study. The patients' INR outcomes were classified as extreme subtherapeutic range, subtherapeutic range, within

XSL•FO
RenderX

therapeutic range, supertherapeutic range, and extreme supertherapeutic range according to their INR target range. Table 2 shows the specific definitions of these 5 ranges.

Table 2. Classification criteria for INR[a].

INR classification	Definition
Extreme subtherapeutic range	INR < lower limit of target range (0.2)
Subtherapeutic range	Lower limit of target range (0.2) ≤ INR < lower limit of target range
Therapeutic range	Lower limit of target range ≤ INR ≤ upper limit of target range
Supratherapeutic range	Upper limit of target range＋0.2< INR ≤4.5
Extreme supratherapeutic range	INR >4.5

[a]INR: international normalized ratio.

Sample Size Calculation

This study used the primary outcome indicator TTR to calculate the sample size. Based on the results of previous studies, the TTR was approximately 73.1% and 66.% for the web-based and non–web-based groups, respectively [14]. When the one-class error (α) was .05 and the degree of certainty (1-β) was 80%, the web-based group was included in a 1:1 ratio to the non–web-based group, and a minimum of 206 patients per group was required as calculated by the statistical software Power Analysis & Sample Size (version 15; NCSS). Assuming a 20% lost-to-review rate, 258 patients were needed in each group, for a total of 516 patients.

Data Collection and Management

All patients' demographic information was collected from the hospital information system. INR records and patients' clinical events in the non–web-based management group were obtained by telephone follow-up, which was conducted at 1, 3, and 6 months after patient enrollment. Patients who did not meet the follow-up end point were excluded. INR records and clinical events of patients in the web-based management group were derived from the Alfalfa app background management system.

Statistical Analysis

The Kolmogorov-Smirnov test was used to verify the normality of the distribution of all continuous variables. Normally distributed continuous variables were expressed as mean (SD), and the Student *t* test was used to compare the differences between the 2 groups of variables. Continuous variables that did not conform to the normal distribution are represented as medians (IQR), and their differences were compared by using the Mann-Whiney U test. Categorical variables were represented as examples (percentages), and chi-square tests or Fisher exact probability methods were used for the statistical analysis of categorical variables. In addition, the difference in the cumulative event rate between the 2 groups was represented by a Kaplan-Meier curve and compared by a log-rank test. In this study, the difference was statistically significant when $P<.05$.

The aforementioned statistical analysis was performed using SPSS software (version 20.0; IBM Corp) and R software (version 4.1.1; R Foundation for Statistical Computing).

Results

Baseline Situation

A total of 522 patients were enrolled in this study between June 2019 and May 2021, with the last follow-up completed on November 12, 2021. Among all patients, 3 (0.57%) in the non–web-based group did not meet the follow-up end point because 1 (0.19%) patient became pregnant during follow-up and 2 (0.38%) patients refused follow-up. The final data analysis included information on a total of 519 patients, including 260 web-based patients (50%) and 259 non–web-based patients (50%). The patients' baseline information is shown in Table 3.

Among the patients in the non–web-based management group, 140 (53.8%) were male. The average age was 49.1 (SD 11.7) years, and the average BMI was 22.1 (SD 4.1) kg/m^2. The number of patients who smoked and drank was 79 (30.4%) and 55 (21.2%), respectively. In the web-based management group, there were 155 (59.9%) male patients, the average age was 48 (SD 14) years, and the average BMI was 22.6 (SD 3.9) kg/m^2. The number of patients who smoked and drank was 69 (26.6%) and 60 (23.2%), respectively. There was no statistically significant difference in demographic information between the 2 groups ($P>.05$). The most common comorbidity, which accounted for more than 15% (n=40) of patients in both groups, was hypertension. The proportion of patients with antiplatelet therapy was 3.9% (n=10) in the web-based management group, slightly higher than the 1.6% (n=3) in the non–web-based management group, but the difference was not statistically significant. Except for antiplatelet therapy, the 2 groups were similar in terms of other comorbidities and medications, and neither group showed a statistically significant difference. In addition, there was no statistically significant difference in the proportion of patients with each anticoagulant indication between the 2 groups.

XSL•FO
RenderX

Table 3. Baseline information of the study population.

Variable name	Non–web-based management group (n= 260)	Web-based management group (n=259)	*P* value
Demographic information			
Male sex, n (%)	140 (53.8)	155 (59.9)	.18
Age (years), mean (SD)	49 (11.7)	48 (14)	.40
BMI (kg/m^2), mean (SD)	22.1 (4.1)	22.6 (3.9)	.12
Smoking history, n (%)	79 (30.4)	69 (26.6)	.38
Drinking history, n (%)	55 (21.2)	60 (23.2)	.60
Concomitant diseases and medications, n (%)			
Hypertension	44 (16.9)	40 (15.4)	.72
Diabetes	12 (4.6)	14 (5.4)	.69
Gout or hyperuricemia	7 (2.7)	11 (4.3)	.35
Combined antiplatelet therapy	3 (1.6)	10 (3.9)	.05
Previous history of thrombosis	7 (2.7)	9 (3.5)	.62
Previous history of bleeding	3 (1.2)	2 (0.8)	.99
Combined tumor	0 (0)	2 (0.8)	.25
Combined atrial fibrillation	1 (0.4)	0 (0)	.99
Anticoagulation indications, n (%)			
Heart valve replacement or repair	251 (96.5)	252 (97.3)	.80
Atrial fibrillation	5 (1.9)	2 (0.8)	.45
VTE[a]	4 (1.6)	5 (1.9)	.75

[a]VTE: venous thromboembolism.

Anticoagulation Management Quality

In this study, a total of 6084 INR records were collected, including 3125 times (51.4%) in the non–web-based management group and 2959 times (48.6%) in the web-based management group. The TTR of the non–web-based management group was significantly lower than in the web-based management group, at 71.6% (SD 25.9%) versus 82.4% (SD 18.6%), respectively (*P*<.001). The distribution of TTR is shown in Figure 2. The proportion of TTR greater than 65% in the non–web-based management group was significantly lower than in the web-based management group, and there was a statistical difference between the 2 groups (63.5% vs 81.2%, *P*<.001).

The distribution of INR values is shown in Table 4. The proportion of INR values in the non–web-based management group in the extreme subtreatment range and extreme overtreatment range was significantly higher than that in the web-based management group, and the proportion in the treatment range was significantly lower than that in the web-based management group (*P*<.001). However, the proportion of INR values in the subtherapeutic range and the supertherapeutic range in the non–web-based management group was also higher than that in the web-based management group, but the difference was not statistically significant (*P*=.18 and *P*=.14, respectively).

XSL•FO
RenderX

Figure 2. Time distribution in the treatment range.

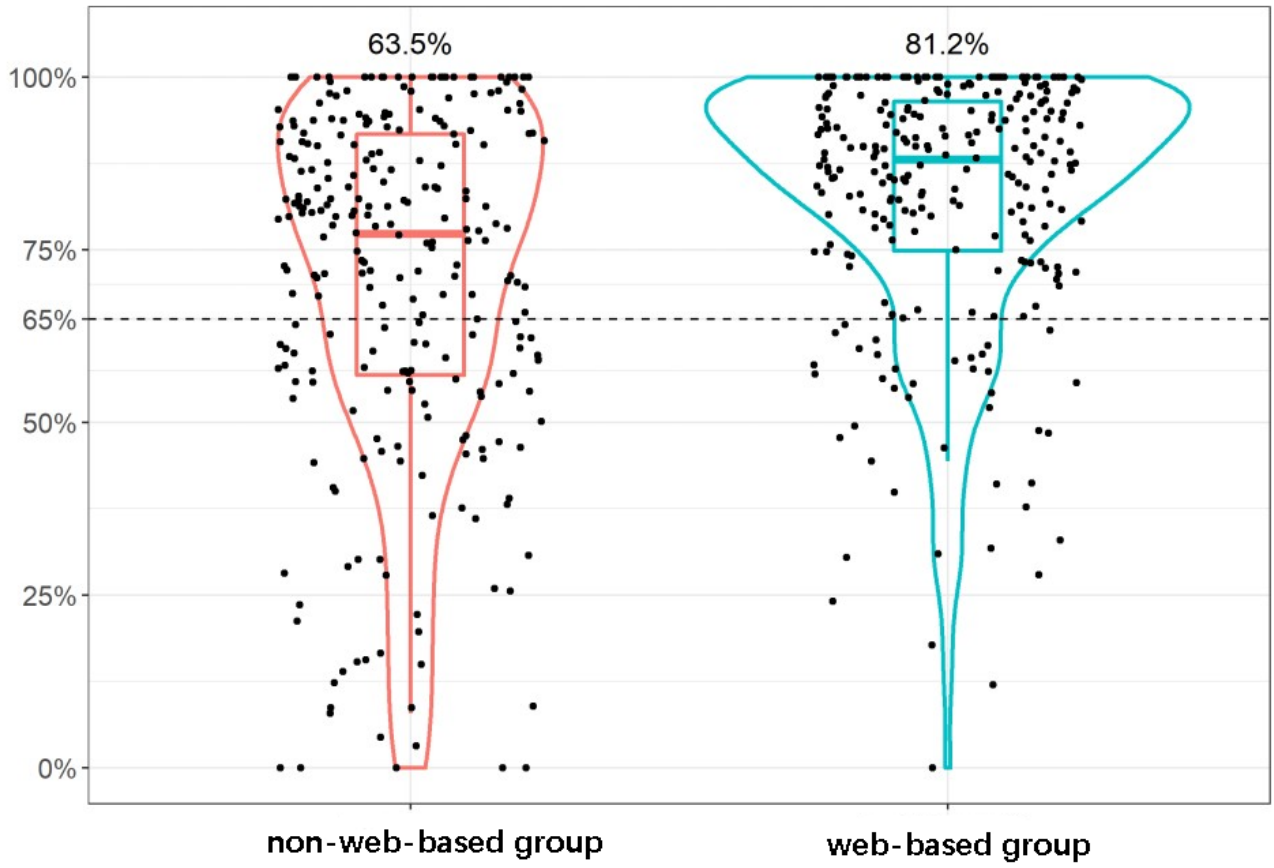

Table 4. The distribution of INR[a] values.

INR value classification	Non–web-based management group (n=3125), n (%)	Web-based management group (n=2959), n (%)	*P* value
Extreme subtherapeutic range	666 (21.3)	521 (17.6)	<.001
Subtherapeutic range	515 (16.5)	450 (15.2)	.18
Therapeutic range	1348 (43.1)	1490 (50.4)	<.001
Supratherapeutic range	572 (18.3)	498 (16.8)	.14
Extreme supratherapeutic range	24 (0.8)	0 (0)	<.001

[a]INR: international normalized ratio.

Clinical Event Outcomes

In terms of clinical events, the non–web-based management group had a significantly higher incidence of minor bleeding events than the web-based management group (12.1% vs 6.6%, respectively, *P*=.048). Furthermore, there was no significant difference in the incidence of major bleeding events and thromboembolic events between the 2 groups. A total of 2 (0.78%) fatalities were observed in patients in the non–web-based management group, with acute prosthetic valve thrombosis and acute ischemic stroke. There were no recorded fatalities in the web-based management group. The specific occurrence of clinical events is shown in Multimedia Appendix 3. Epistaxis and gingival bleeding were the 2 most common types of bleeding, with a total of 16 (6.18%) and 19 (7.33%) cases, respectively. Figure 3 shows the cumulative clinical event rate between the web-based management group and the non–web-based management group during the follow-up period. Compared with the web-based management group, the non–web-based management group had a higher cumulative incidence of minor bleeding time (*P*=.035), but there were no statistical differences in severe bleeding time, thrombotic events, and all-cause mortality between the 2 groups (*P*>.05). Table 5 shows the clinical event outcomes.

XSL•FO
RenderX

Figure 3. Kaplan-Meier curve. A. Minor bleeding; B. Severe bleeding; C. Thromboembolic; D. All-cause deaths.

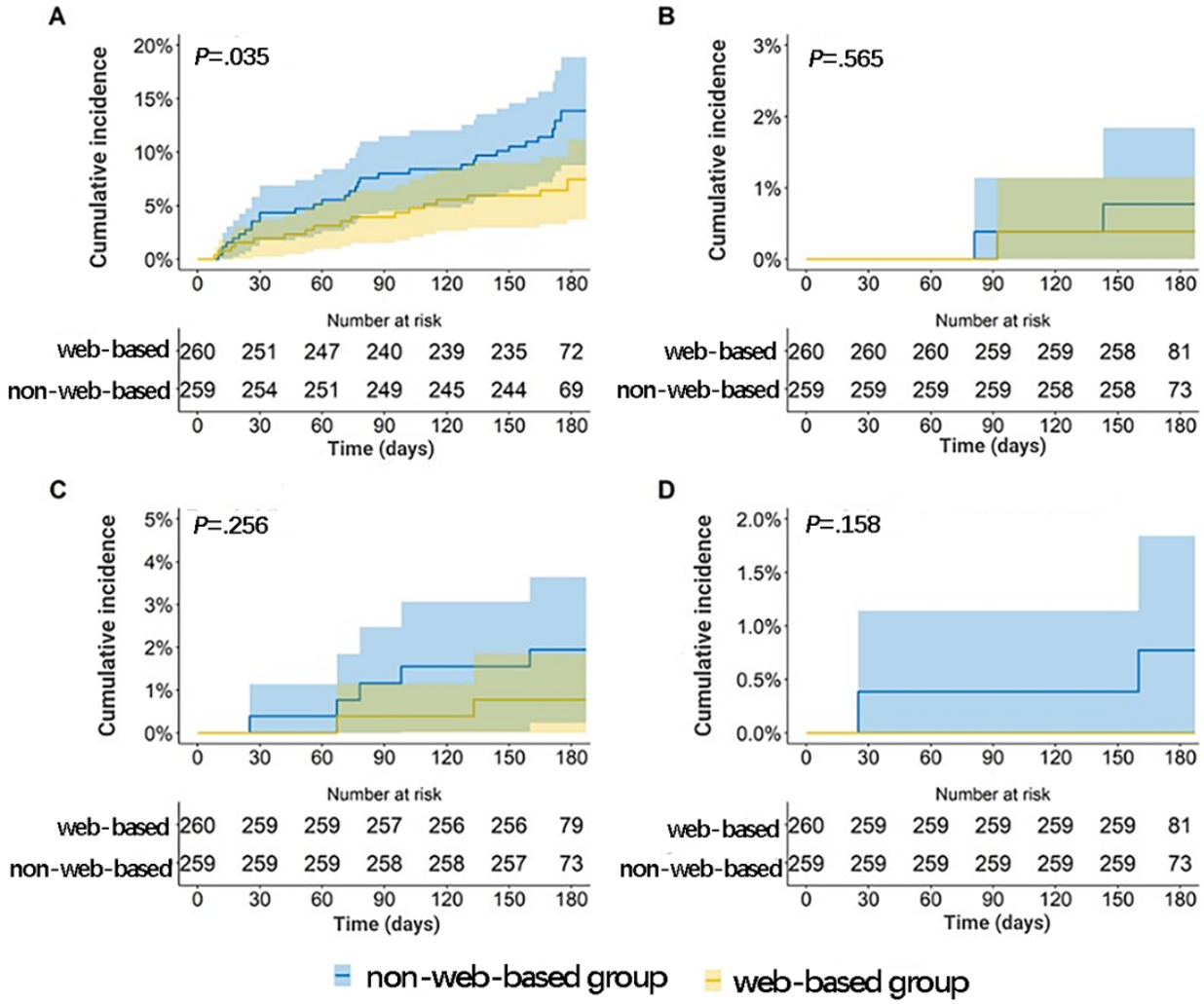

Table 5. Clinical event outcomes.

Clinical events	Non–web-based management group (n= 260), n (%)	Web-based management group (n= 259), n (%)	*P* value
Minor bleeding events	31 (12.1)	17 (6.6)	.048
Severe bleeding events	2 (0.78)	1 (0.39)	.99
Thromboembolic events	5 (1.95)	2 (0.77)	.45
All-cause death events	2 (0.78)	0 (0)	.49

Discussion

Principal Findings

This national multicenter prospective cohort study is the first in China to compare the difference between the quality of anticoagulation management and clinical event outcomes between web-based anticoagulation management carried out via the smartphone Alfalfa app and traditional non–web-based outpatient management. The study was conducted in 5 centers in China, located in the southwest, southeast, and central regions, which are disparate in terms of geographical location, climate, and diet. To some extent, these regions represent most of the country. In general, a TTR greater than 65% is considered the standard for effective anticoagulation management, in which case the benefits of anticoagulation outweigh the bleeding risk associated with anticoagulation [15]. The average TTRs of patients in the non–web-based management group and the web-based management group were 71.6% and 82.6%, respectively. According to the aforementioned criteria, patients in the 2 groups were able to obtain effective anticoagulation overall. However, more patients in the web-based management group (n=210, 81.2%) received effective anticoagulation management than in the non–web-based management group (n=165, 63.5%). The Alfalfa app may improve the quality of anticoagulation through the following ways: (1) the app is equipped with patient education modules to improve patients' medical knowledge regarding anticoagulation; (2) the app is

XSL•FO
RenderX

also equipped with useful medication and blood check reminder modules to prevent the occurrence of missed doses and prolong intervals between INR testing; and (3) the convenient model of anticoagulation management through the app is more acceptable to patients and can improve their medication compliance.

Of note, previous studies have reported a distribution roughly between 17.9% and 56.2% of TTRs under routine management [16] and that effective anticoagulant management is not available for 65% of patients [17]. The results of this study are not consistent with the aforementioned studies. This may be due to the relatively young sample composition of the non–web-based management group in this study, with an average age of 49 years, and their better compliance and medication-related knowledge. Additionally, China's health system construction in recent years has improved the availability of high-quality medical resources, which may also be the reason for this difference. Recently published studies have also reported that the TTR of patients in China under routine management mode exceeds 70% [18]. In terms of clinical events, more patients in the non–web-based management group experienced minor bleeding events during the 6-month follow-up period, but there was no statistically significant difference between the 2 groups for severe bleeding, thromboembolic events, and all-cause mortality. Web-based management can also increase the proportion of INR values in the therapeutic range and reduce the probability of INR values in the extreme subtherapeutic and extreme overtreatment ranges.

Chinese medical practitioners have extensively explored the remote management of warfarin. Numerous studies have demonstrated the benefits of remote management to varying degrees, although the specific approaches and conclusions of these studies are not exactly the same. Min Gu et al [19] conducted a randomized controlled study to evaluate the feasibility and clinical application effect of the “follow-up patient” mobile app in patients with atrial fibrillation taking warfarin, and the results showed that the anticoagulation management model guided by telemedicine technology could reduce the risk of minor bleeding and ischemic stroke and decrease the number of emergency department visits. Jinze Li et al [20] evaluated the security and efficacy of an “internet + smartphone app” in patients with mechanical heart valve replacement, and the TTR of the remote group was significantly higher than that of the non–web-based group, with no significant difference in the incidence of anticoagulation-related complications between the 2 groups during the 12-month follow-up period. Cao et al [3] conducted a retrospective, observational cohort study in 2018 comparing the efficacy and safety of anticoagulation management services provided by pharmacists through hospital anticoagulation clinics versus web-based anticoagulation clinics. In their study, patients in both groups had similar TTR and clinical event outcomes but a higher incidence of out-of-the-treatment range INR values in the web-based anticoagulation outpatient group.

Regarding the Alfalfa app used in this study, previous retrospective cohort studies have also shown that web-based administration improves TTR and reduces the incidence of minor bleeding events, major bleeding events, warfarin-related emergency department events, and warfarin-related hospitalizations [3]. In terms of the quality of anticoagulant management, the results of this study are consistent with the aforementioned study, in that remote administration can improve the quality of anticoagulation management in patients using warfarin. However, results in clinical events are not completely consistent and may be due to several factors. First, the study population came from different regions of China, and there were differences in demographic characteristics, dietary habits, adherence, and other factors, which in turn affected the clinical outcome of the anticoagulation therapy. Second, the existing studies had a follow-up period of 6 months to 1 year, and some studies lacked sufficient sample sizes (less than 100 people) to reflect the effect of remote anticoagulation management on clinical events with low incidences such as intracranial hemorrhage and ischemic cerebral infarction. It is expected that future large studies will demonstrate the long-term clinical outcomes of remote anticoagulation management.

Although new oral anticoagulants such as dabigatran and rivaroxaban have been marketed in China in recent years, they cannot completely replace warfarin clinically due to their specific indications and higher prices. Therefore, the warfarin management model still requires long-term exploration and optimization. As a new management model, web-based warfarin management has been positively evaluated by many medical practitioners. In addition to the clinical benefits, remote warfarin management is also beneficial to public health. First, using internet technology can reduce costs and increase patients' access to medical services [21]. Second, telemedicine can increase the efficacious use of medical and health resources and promote their redistribution in geographical distribution [22]. Moreover, the advantages of warfarin remote management are also evident in the transportation lockdown and social distancing measures taken to prevent the spread of COVID-19, in that patients no longer need to visit the hospital to obtain a dose adjustment regimen, reducing the risk of exposure to the novel coronavirus. Compared with urban areas in China, patients in rural areas face a long-term higher risk of intracranial hemorrhage and ischemic stroke and lack of effective anticoagulation management [23]. Benefiting from the efficient network infrastructure in China, the total number of internet users reached 1.011 billion by June 2021, of which the size of rural internet users reached 297 million [24]. These populations are likely to be beneficiaries of remote warfarin administration.

However, the use of internet technology has improved the quality of anticoagulation management while also raising the barriers to access to care, making the population of beneficiaries limited. Patients usually need to be trained in the use of the software before participating in warfarin telemanagement, which can be challenging for certain populations, such as older adults or those with low literacy levels [25]. For example, the usability evaluation findings of the Alfalfa app showed that some patients had difficulty using it, especially older adults [8]. There are 3 main reasons for this phenomenon. First, the user group was limited by age and literacy level, making it challenging for them to learn and use the telemedicine software. To solve this challenge, the users of the Alfalfa app are not limited to the patients themselves. Medical staff can encourage patients' family members to replace or assist this group with the operational

XSL•FO
RenderX

steps of remote management. Second, the operation and functionality of the software are so complex that it takes a long time for patients to learn how to operate the software. Therefore, in the process of developing remote anticoagulation management software, medical staff need to fully communicate with software technicians to improve ease of use by simplifying the operation and adding suggestive symbols and text, thereby expanding the population that can benefit from remote anticoagulation management. Third, in addition to the basic operation of the app, patients need to acquire additional warfarin-related knowledge. This may not be directly related to the operation of the software, but it is important to ensure that patients are able to manage anticoagulation-related complications promptly and correctly. Striking a balance between the usability and security of telemedicine software is the focus of future research. Furthermore, data management and privacy protection are key issues in the implementation of telemedicine. Imprudent implementation of telemedicine technologies may lead to loss of patient privacy and autonomy, which telemedicine practitioners must carefully address.

Study Limitations

This study has the following limitations. First, the inclusion process of patients was not randomized to groups and may have been subject to selection bias. Second, patients from western and northeastern China were not included in this study, and caution is needed to extrapolate the findings to the national level. Third, the 6-month follow-up period may make it difficult to capture the difference between web-based and non–web-based anticoagulation management in terms of low incidence of clinical events. Future studies with longer follow-ups will be needed to observe the long-term clinical outcomes of both management models.

Conclusion

Compared with traditional non–web-based outpatient management, web-based management via the Alfalfa app may improve the quality of anticoagulation management of patients with respect to the percentage of patients meeting effective anticoagulation criteria. It can also reduce the incidence of minor bleeding events and the probability of INR values in the extreme subtherapeutic and extreme supratherapeutic ranges.

Acknowledgments

We thank JZ for conceiving this study. This work was supported by the Natural Science Foundation of Fujian Province of China (2018Y0037), which we are grateful for.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study's conception and design. WC, JC, and SJ carried out the manuscript preparation, data collection, and analysis. WC wrote the first draft of the manuscript, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
RECORD-PE (Reporting of Studies Conducted Using Observational Routinely Collected Health Data Statement for Pharmacoepidemiology) checklist.
[PDF File (Adobe PDF File), 116 KB - jmir_v26i1e46319_app1.pdf]

Multimedia Appendix 2
Details of the Alfalfa app.
[DOCX File , 541 KB - jmir_v26i1e46319_app2.docx]

Multimedia Appendix 3
Occurrence of clinical events.
[DOCX File , 106 KB - jmir_v26i1e46319_app3.docx]

References

1. Barnes GD, Lucas E, Alexander GC, Goldberger ZD. National trends in ambulatory oral anticoagulant use. Am J Med 2015 Dec;128(12):1300-5.e2 [FREE Full text] [doi: 10.1016/j.amjmed.2015.05.044] [Medline: 26144101]
2. Baumgartner H, Falk V, Bax JJ. 2017 ESC/EACTS guidelines for the management of valvular heart disease. Eur Heart J 2017;38(36):2739-2791. [doi: 10.1007/s00059-017-4643-5]

XSL•FO
RenderX

3. Cao H, Jiang S, Lv M, Wu T, Chen W, Zhang J. Effectiveness of the Alfalfa app in warfarin therapy management for patients undergoing venous thrombosis prevention and treatment: cohort study. JMIR Mhealth Uhealth 2021 Mar 02;9(3):e23332 [FREE Full text] [doi: 10.2196/23332] [Medline: 33650976]
4. Yao H, Zhan C, Sha X. Current situation and distribution equality of public health resource in China. Arch Public Health 2020 Sep 22;78(1):86 [FREE Full text] [doi: 10.1186/s13690-020-00474-3] [Medline: 32983449]
5. Yang X, Li Z, Zhao X, Wang C, Liu L, Wang C, et al. Use of warfarin at discharge among acute ischemic stroke patients with nonvalvular atrial fibrillation in China. Stroke 2016 Feb;47(2):464-470. [doi: 10.1161/strokeaha.115.011833]
6. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ 2018 Nov 14;363:k3532 [FREE Full text] [doi: 10.1136/bmj.k3532] [Medline: 30429167]
7. Xia X, Fu J, Wu T, Chen W, Jinhua Z. Comparison of the outcomes of warfarin therapy and economics by online and offline anticoagulation management models: protocol for a randomised controlled trial. BMJ Open 2019 Sep 05;9(9):e031730 [FREE Full text] [doi: 10.1136/bmjopen-2019-031730] [Medline: 31492796]
8. Jiang S, Lv M, Wu T, Chen W, Zhang J. A smartphone application for remote adjustment of warfarin dose: Development and usability study. Appl Nurs Res 2022 Feb;63:151521. [doi: 10.1016/j.apnr.2021.151521] [Medline: 35034699]
9. McDowell TY, Lawrence J, Florian J, Southworth MR, Grant S, Stockbridge N. Relationship between International Normalized Ratio and Outcomes in Modern Trials with Warfarin Controls. Pharmacotherapy 2018 Sep;38(9):899-906. [doi: 10.1002/phar.2161] [Medline: 29920722]
10. Bowman TS, Gaziano JM, Kase CS, Sesso HD, Kurth T. Blood pressure measures and risk of total, ischemic, and hemorrhagic stroke in men. Neurology 2006 Sep 12;67(5):820-823. [doi: 10.1212/01.wnl.0000233981.26176.e1] [Medline: 16966544]
11. Hilkens NA, Greving JP, Algra A, Klijn CJM. Blood pressure levels and the risk of intracerebral hemorrhage after ischemic stroke. Neurology 2017 Jan 10;88(2):177-181. [doi: 10.1212/WNL.0000000000003489] [Medline: 27927936]
12. Schlunk F, Chang Y, Ayres A, Battey T, Vashkevich A, Raffeld M, et al. Blood pressure burden and outcome in warfarin-related intracerebral hemorrhage. Int J Stroke 2016 Oct;11(8):898-909. [doi: 10.1177/1747493016658300] [Medline: 27462095]
13. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the ScientificStandardization Committee of the International Society on ThrombosisHaemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J Thromb Haemost 2005 Apr;3(4):692-694 [FREE Full text] [doi: 10.1111/j.1538-7836.2005.01204.x] [Medline: 15842354]
14. Zhang J, Liu M, Chen Q, Wu J, Cao H. Outcomes of an online pharmacist-managed anticoagulation clinic for individuals on warfarin therapy living in rural communities. Thromb Res 2017 Sep;157:136-138. [doi: 10.1016/j.thromres.2017.07.016] [Medline: 28746903]
15. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Ann Intern Med 2007 Jun 19;146(12):857-867. [doi: 10.7326/0003-4819-146-12-200706190-00007] [Medline: 17577005]
16. Feng X, Huan Y, Lv Y. Letter by Feng et al regarding article, “Ischemic stroke and intracranial hemorrhage with aspirin, dabigatran, and warfarin: impact of quality of anticoagulation control”. Stroke 2015 Mar;46(3). [doi: 10.1161/strokeaha.114.008578]
17. Li X, Pathadka S, Man KKC, Ng VWS, Siu CW, Wong ICK, et al. Comparative outcomes between direct oral anticoagulants, warfarin, and antiplatelet monotherapy among Chinese patients with atrial fibrillation: a population-based cohort study. Drug Saf 2020 Oct 18;43(10):1023-1033. [doi: 10.1007/s40264-020-00961-0] [Medline: 32557348]
18. Jiang S, He Q, Yan J, Zhao L, Zheng Y, Chen P, et al. Evaluation of a pharmacist-led remote warfarin management model using a smartphone application (Yixing) in improving patients' knowledge and outcomes of anticoagulation therapy. Front Pharmacol 2021 Jul 1;12:677943 [FREE Full text] [doi: 10.3389/fphar.2021.677943] [Medline: 34276368]
19. Gu M. Study of telemedicine application for warfarin anticoagulation management in atrial fibrillation. Thesis. Yangzhou University. 2019. URL: https://authserver-443.webvpn.fjmu.edu.cn/authserver/login?service=https%3A%2F%2Fwebvpn.fjmu.edu.cn%2Fusers%2Fauth%2Fcas%2Fcallback%3Furl [accessed 2022-03-03]
20. Li JZ, Dong Y, Qian Y. Remote management of anticoagulation after cardiac mechanical valve replacement:a prospective cohort studyJ. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery 2019;26(04):343-347.
21. Baksheev A, Turchina Z, Sharova O. Innovations in medicine: features of regulation and prospects for the development of telemedicine. Revista Inclusiones 2020:447-459.
22. Cui F, Ma Q, He X, Zhai Y, Zhao J, Chen B, et al. Implementation and application of telemedicine in China: cross-sectional study. JMIR Mhealth Uhealth 2020 Oct 23;8(10):e18426 [FREE Full text] [doi: 10.2196/18426] [Medline: 33095175]
23. Wang W, Jiang B, Sun H, Ru X, Sun D, Wang L, et al. Prevalence, incidence, and mortality of stroke in China. Circulation 2017 Feb 21;135(8):759-771. [doi: 10.1161/circulationaha.116.025250]
24. The 48th statistical report on the development of the Internet in China. China Internet Information Center. 2021. URL: http://www.bjjubao.org.cn/2021-09/26/content_43536.html [accessed 2022-03-01]
25. Correa T, Pavez I. Digital Inclusion in rural areas: a qualitative exploration of challenges faced by people from isolated communities. J Comput-Mediat Comm 2016 Mar 24;21(3):247-263. [doi: 10.1111/jcc4.12154]

XSL•FO
RenderX

Abbreviations

INR: international normalized ratio
ISTH: International Society on Thrombosis and Hemostasis
RECORD-PE: Reporting of Studies Conducted Using Observational Routinely Collected Health Data Statement for Pharmacoepidemiology
TTR: time in therapeutic range

Edited by A Mavragani; submitted 06.02.23; peer-reviewed by Y Fang, R Gao; comments to author 30.11.23; revised version received 17.12.23; accepted 20.05.24; published 29.07.24.

Please cite as:
Chen W, Chen J, Jiang S, Wang C, Zhang J
Web-Based Warfarin Management (Alfalfa App) Versus Traditional Warfarin Management: Multicenter Prospective Cohort Study
J Med Internet Res 2024;26:e46319
URL: https://www.jmir.org/2024/1/e46319
doi:10.2196/46319
PMID:39073869

©Wenfei Chen, Jiana Chen, Shaojun Jiang, Chunhua Wang, Jinhua Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Effective Deep Learning Framework for Fall Detection: Model Development and Study Design

Jinxi Zhang[1,2], PhD; Zhen Li[3], PhD; Yu Liu[4], MS; Jian Li[4], MS; Hualong Qiu[4], MS; Mohan Li[4], MS; Guohui Hou[5,6], MS; Zhixiong Zhou[2], PhD

[1]Beijing Kupei Sports Culture Corporation Limited, Beijing, China
[2]Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, Beijing, China
[3]School of Physical Education and Sport Science, Fujian Normal University, Fuzhou, China
[4]Institute for Sport Performance and Health Promotion, Capital University of Physical Education and Sports, Beijing, China
[5]Bioelectronics Center of YZW, Shanghai, China
[6]Walt Technology Group Co, Ltd, Jiaxing, China

Corresponding Author:
Zhixiong Zhou, PhD
Institute of Artificial Intelligence in Sports
Capital University of Physical Education and Sports
11 North Third Ring West Road
Haidian District
Beijing, 100191
China
Phone: 86 13552505679
Email: zhouzhixiong@cupes.edu.cn

Abstract

Background: Fall detection is of great significance in safeguarding human health. By monitoring the motion data, a fall detection system (FDS) can detect a fall accident. Recently, wearable sensors–based FDSs have become the mainstream of research, which can be categorized into threshold-based FDSs using experience, machine learning–based FDSs using manual feature extraction, and deep learning (DL)–based FDSs using automatic feature extraction. However, most FDSs focus on the global information of sensor data, neglecting the fact that different segments of the data contribute variably to fall detection. This shortcoming makes it challenging for FDSs to accurately distinguish between similar human motion patterns of actual falls and fall-like actions, leading to a decrease in detection accuracy.

Objective: This study aims to develop and validate a DL framework to accurately detect falls using acceleration and gyroscope data from wearable sensors. We aim to explore the essential contributing features extracted from sensor data to distinguish falls from activities of daily life. The significance of this study lies in reforming the FDS by designing a weighted feature representation using DL methods to effectively differentiate between fall events and fall-like activities.

Methods: Based on the 3-axis acceleration and gyroscope data, we proposed a new DL architecture, the dual-stream convolutional neural network self-attention (DSCS) model. Unlike previous studies, the used architecture can extract global feature information from acceleration and gyroscope data. Additionally, we incorporated a self-attention module to assign different weights to the original feature vector, enabling the model to learn the contribution effect of the sensor data and enhance classification accuracy. The proposed model was trained and tested on 2 public data sets: SisFall and MobiFall. In addition, 10 participants were recruited to carry out practical validation of the DSCS model. A total of 1700 trials were performed to test the generalization ability of the model.

Results: The fall detection accuracy of the DSCS model was 99.32% (recall=99.15%; precision=98.58%) and 99.65% (recall=100%; precision=98.39%) on the test sets of SisFall and MobiFall, respectively. In the ablation experiment, we compared the DSCS model with state-of-the-art machine learning and DL models. On the SisFall data set, the DSCS model achieved the second-best accuracy; on the MobiFall data set, the DSCS model achieved the best accuracy, recall, and precision. In practical validation, the accuracy of the DSCS model was 96.41% (recall=95.12%; specificity=97.55%).

Conclusions: This study demonstrates that the DSCS model can significantly improve the accuracy of fall detection on 2 publicly available data sets and performs robustly in practical validation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56750) doi:10.2196/56750

KEYWORDS

fall detection; deep learning; self-attention; accelerometer; gyroscope; human health; wearable sensors; Sisfall; MobiFall

Introduction

According to the World Health Organization, falls rank as the second leading cause of accidental injury-related deaths on a global scale [1]. Statistics from the Centers for Disease Control and Prevention [2] reveal that a significant proportion, at least a quarter, of US residents aged 65 years and older experience a fall annually. Among the older adult population, accidental falls have been the second leading cause of mortality and injury [3]. To prevent falls from causing severe subsequent harm to individuals, it is essential to develop an accurate and efficient fall recognition system, to identify falls and raise the alarm [4]. In the existing research, 2 main categories of systems have emerged: fall prediction systems [5-7] and fall detection systems (FDSs) [8-21].

Although fall prediction systems can assist users in proactively preventing potential falls, the associated installation and maintenance costs can be prohibitively high, thereby limiting its widespread application. In contrast, FDSs can promptly issue alerts upon detecting a fall event, ensuring swift assistance for the individual who has experienced the fall. In recent years, extensive research has been conducted on FDSs and related solutions, categorized as follows: (1) vision-based FDSs [8,9], which monitor and analyze a person's movements and postures using cameras or imaging devices to detect falls; (2) ambient device–based FDSs [10-12], which use environmental sensors such as Wi-Fi or radar signals to track an individual's movement data within their living space and detect falls; and (3) wearable sensors–based FDSs [13-21], which use sensors attached to the body to monitor a user's movements and postures to detect falls. Among these solutions, wearable sensor–based fall detection technology has garnered significant attention because of its affordability and nonintrusive characteristics [21]. Wearable devices, such as inertial measurement units [13,14], smartwatches [15-17], and smartphones [18-20], use high-precision sensors to collect motion data seamlessly. These devices have been widely applied in fall detection and safety monitoring [22].

In wearable sensor–based FDSs, the design of fall detection algorithms is essential. Presently, algorithms used for fall detection can be classified into 3 categories: threshold-based models, machine learning–based models, and deep learning (DL) models. Threshold-based models and machine learning–based models necessitate the extraction of distinctive features from data sets containing fall incidents. In essence, it primarily entails capturing features such as the intensity of activities (eg, magnitude, energy) and variations in the intensity of activities (eg, frequency and SD) from the input data. Threshold-based models discern falls by comparing feature values against predefined thresholds, while machine learning models, such as k-nearest neighbors (KNN) [23], support vector machines (SVM) [24], and decision trees (DT) [25], categorize falls and activities of daily life (ADL) based on the handcrafted features. In contrast, DL models, such as convolutional neural networks (CNN) [26] and long short-term memory (LSTM) [27,28], automatically extract high-level features and scrutinize the temporal characteristics of the data for fall detection.

However, the aforementioned models solely use acceleration data as the input for their algorithms. A previous work [29] has demonstrated that relying exclusively on acceleration data is inadequate for effectively distinguishing falls from ADL. This insufficiency is primarily due to the sensitivity of acceleration data to sensor placement and its inability to capture spatial rotation information. Consequently, researchers advocate for the integration of acceleration data with gyroscope data to provide a more comprehensive understanding of body movements, which can significantly enhance the accuracy of fall detection [30,31].

In this case, Hussain et al [32], Son et al [33], Liu et al [34], and Koo et al [35] use acceleration data with gyroscope data as the input to the model. Hussain et al [32] and Son et al [33] designed fall detection algorithms using KNN and SVM, both of which require manual feature extraction. Additionally, Liu et al [34] developed a CNN-LSTM–based FDS. However, using CNN, LSTM, or their combination fails to evaluate the significance of each component of the feature vector, thereby hindering the differentiation between actual falls and fall-like activities. Moreover, Koo et al [35] proposed a dual-stream–based algorithm for human activity recognition. While this model effectively differentiates between various daily activities, it struggles with the classification of falls and fall-like activities, which exhibit similar data trends.

To address the above challenges, we incorporate the self-attention (SA) mechanism [36] after the CNN module and propose a dual-stream CNN-SA (DSCS) model for fall detection. The SA mechanism has been widely applied in classification tasks such as sleep apnea [37] and skeleton point–based human activity recognition [38]. However, it has not been applied in FDSs. The SA mechanisms empower FDS models to allocate varying weights to different segments of the input data or extracted features.

Compared with existing methods that use manually generated features, the proposed method can automatically extract features using a dual-stream architecture. Specifically, the DSCS model uses acceleration data along with gyroscope data as the input, and then the 2-stream data will pass through a 3-layer CNN to extract discriminable features. Unlike the models using CNN, LSTM, or their combinations, CNN-SA excels at effectively capturing long-term dependencies within input data, enabling the assignment of diverse weights to features from different phases of the fall process. This, in turn, aids the model in achieving an enhanced understanding of the context and pertinent information associated with fall behavior, thereby elevating the model's capability for contextual modeling.

XSL•FO
RenderX

In this paper, we introduce an accurate and embeddable DSCS model for fall detection. Considering that the handcrafted features rely on expert knowledge and may not yield satisfactory generalization performance, we design a DL framework. In this framework, both accelerometer and gyroscope data are used as input for the model. Then, we introduce a feature-generating method grounded in the CNN-SA architecture. Here, CNN is used to extract features from dual-stream data and capture spatial patterns, to preliminarily identify the pattern differences between fall and ADLs. The feature vector extracted by CNN subsequently passes through an SA layer, which assigns weights to accelerometer and gyroscope features. Finally, the predicted label (fall or ADL) is output by the classification module.

The main contributions of this study are as follows: first, we introduce an accurate and embeddable dual-stream model, DSCS, for fall detection using wearable sensors. The DSCS model comprises a feature extraction module, an SA module, and a classification module. Second, the SA mechanism is applied in the task of fall detection, which assigns different weights to the feature vectors, learning the contribution effect of the sensor data, thereby effectively enhancing classification accuracy. Third, we validate the performance of the DSCS model using publicly available data sets and practical validation. Our model outperforms state-of-the-art machine learning and DL models and demonstrates excellent generalization performance.

The rest of this paper is organized as follows: the *Methods* section illustrates the proposed framework DSCS framework. The *Results* section presents the performance comparison with state-of-the-art algorithms and practical validation. The *Discussion* section covers the results, performance gap, limitations of this study, and conclusions.

Methods

Data Sets for Fall Detection

Overview

In this paper, we validated the performance of the DSCS model using 2 publicly available data sets: SisFall [39] and MobiFall [40]. The 2 data sets consist of the accelerometer and gyroscope data recorded during fall and ADL trials of multiple participants.

SisFall Data Set

The SisFall data set collects accelerometer and gyroscope data during human motion, with a sampling frequency of 200 Hz. This data set comprises data on falls and ADLs collected from 38 volunteers, including 23 young individuals (younger than 30 years) and 15 older individuals (60 years and older). It encompasses 15 falls (such as fall forward, fall backward, and lateral fall) and 19 ADLs (eg, walking, jogging, and jumping).

MobiFall Data Set

The MobiFall data set collects accelerometer, gyroscope, and orientation data from smartphones' built-in inertial measurement units during human motion. The accelerometer has a sampling frequency of 87 Hz, while the gyroscope and orientation sensor have a sampling frequency of 100 Hz. This data set is contributed by 24 volunteers and contains 4 distinct fall activities: fall forward, fall forward with knees on the ground, fall lateral, and fall backward onto a chair. Additionally, it contains 9 various daily activities (eg, standing, walking, and jogging).

Data Processing

The processing of accelerometer and gyroscope data consisted of 2 steps: data filtering and data segmentation.

- Data filtering: Considering the susceptibility of sensor data to external environmental noise when worn on the body, we implemented the filtering on the accelerometer and gyroscope data. The filtering process enhances the stability and accuracy of the resulting signals. Since human movements often occur at a frequency of around 20 Hz, a low-pass filter with a cutoff frequency of 20 Hz was used to filter the data [41].
- Data segmentation: A complete fall event typically lasts 8-12 seconds, encompassing 4 phases: prefall, falling, impact, and postfall. To match the sample length in the SisFall and MobiFall data sets, we used the time windows of 12 and 8 seconds for the data segmentation of the SisFall and MobiFall data sets, respectively. Then, we further reduced the sampling rate to 50 Hz to alleviate computational complexity.

DL Framework

Overview

After data processing, we obtained 1798 fall segments and 4117 ADL segments from the SisFall data set and 288 fall segments and 1137 ADL segments from the MobiFall data set. As input to the model, each segment from the SisFall data set has a length of 12 seconds (600 sampling points), while each segment from the MobiFall data set has a length of 8 seconds (400 sampling points). After passing through a DL architecture, each segment produces a predicted label (fall or ADL).

Network Architecture

The deep neural network architecture proposed in this paper is illustrated in Figure 1, which mainly consists of 3 modules: feature extraction module, SA module, and classification module.

XSL•FO
RenderX

Figure 1. The architecture of the proposed dual-stream convolutional neural network self-attention model. ADL: activities of daily life; BN: batch normalization; BS: batch size; FC: fully connected; LoS: length of segment.

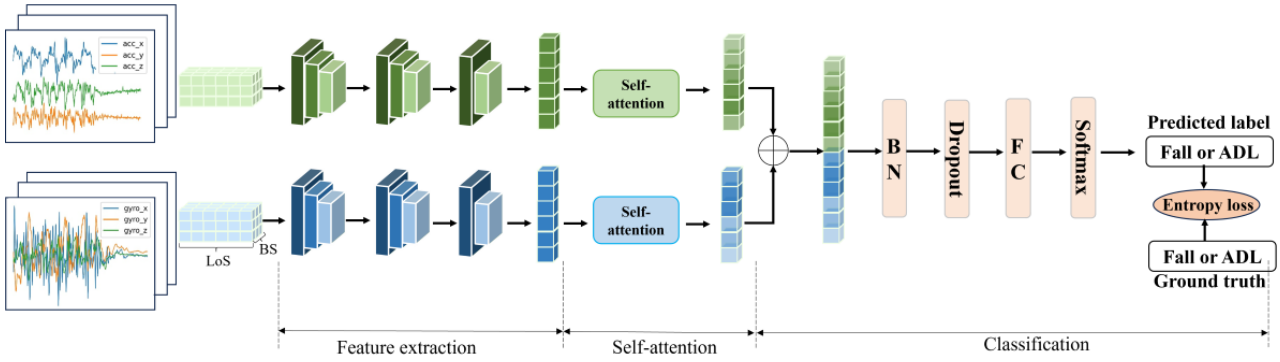

Feature Extraction Module

To extract features, we designed the same architecture for acceleration and gyroscope data stream: a 3-layer CNN to transform the 3-axis data into feature representations. The size of the input data is $N \times LoS \times 3$, where N is the batch size and LoS is the length of a segment. The first part of the 3-layer comprises a 1D convolution with 64 filters, a kernel size of 3, and a stride of 1. The outputs of the first 2 layers are activated using a rectified linear unit (ReLU) and followed by a max-pooling layer with a pooling size of 2. The output of the third convolution layer is directly fed into a global pooling layer, which generates a 64D global feature vector for each element in the batch.

SA Module

The extracted feature representations from the acceleration and gyroscope data are further refined using the SA module. This module enables the acceleration or gyroscope stream network to efficiently discover the local information highly relevant to discriminating falls and ADL, by capturing the attention weights of local feature information. Specifically, during each training iteration, a total of N vectors is fed into the SA module. Within the SA module, these vectors are processed using query, key, and value matrices, each with dimensions of 64×64. Ultimately, a weighted feature vector is generated for each sample in the batch.

Classification Module

The weighted acceleration or gyroscope vectors, which have been processed by the SA module, are concatenated to a 128D vector and fed into the classification module. This module is responsible for mapping the concatenated vector to the corresponding prediction label. Specifically, the 128D vector is first processed through a 128D batch normalization (BN) layer, followed by a dropout layer with a dropout probability of 0.5. Finally, the vector is input into a fully connected layer, which comprises a 256D hidden layer and a softmax output layer.

Next, we will provide an in-depth illustration of the architecture and functionality of the 3 modules.

Feature Extraction Module

Following Koo et al [35], we design a feature extraction module for accelerometer and gyroscope data streams. This module comprises 3 consecutive encoders, with each encoder combining CNNs and pooling layers.

For acceleration or gyroscope stream, the input data are initially processed by the first encoder, where a 1D CNN extracts shallow-level feature representations from the input data. Subsequently, a max-pooling layer is applied to reduce data dimensions and emphasize key features.

In the second encoder, a CNN further extracts features in a local context, enabling the capture of more intricate patterns and relationships within the time series data. Similarly, a max-pooling layer is used to reduce feature map dimensions and highlight critical features.

The third encoder is dedicated to global feature abstraction, allowing it to capture overarching patterns and essential information across the entire time series, rather than only local features. A global pooling layer enables pooling across the entire feature map, producing a global feature vector that encapsulates summarized information from the entire data sequence.

Through the hierarchical processing by the 3 encoders, the DSCS model can extract both local and global features from the original 3D accelerometer or gyroscope data, thereby enhancing the model's prediction performance and generalization capabilities.

SA Module

The feature vectors generated by the feature extraction module pass through a SA module. This module uses an attention mechanism to dynamically adjust the weights across different positions of the 1D feature vectors, capturing the interdependencies and correlations between different positions.

The detailed process of generating a weighted feature vector is illustrated in Figure 2. Specifically, the SA module constructs query, key, and value vectors by multiplying the input feature vectors with respective weight matrices. The module calculates the dot product between the query and key vectors and then applies the softmax function to determine the relationships between the feature values at each position and those at other

XSL•FO
RenderX

positions. The result of the dot product represents the strength of their relationships, commonly referred to as "attention". Finally, the weighted feature representations for each position are obtained by applying the corresponding attention weights to the value vectors. In the SA mechanism, this weighting operation helps describe the correlations and associations between each position and the others, enabling the model to focus on different positions within the information.

Figure 2. The process of generating weighted feature vectors from original feature vectors.

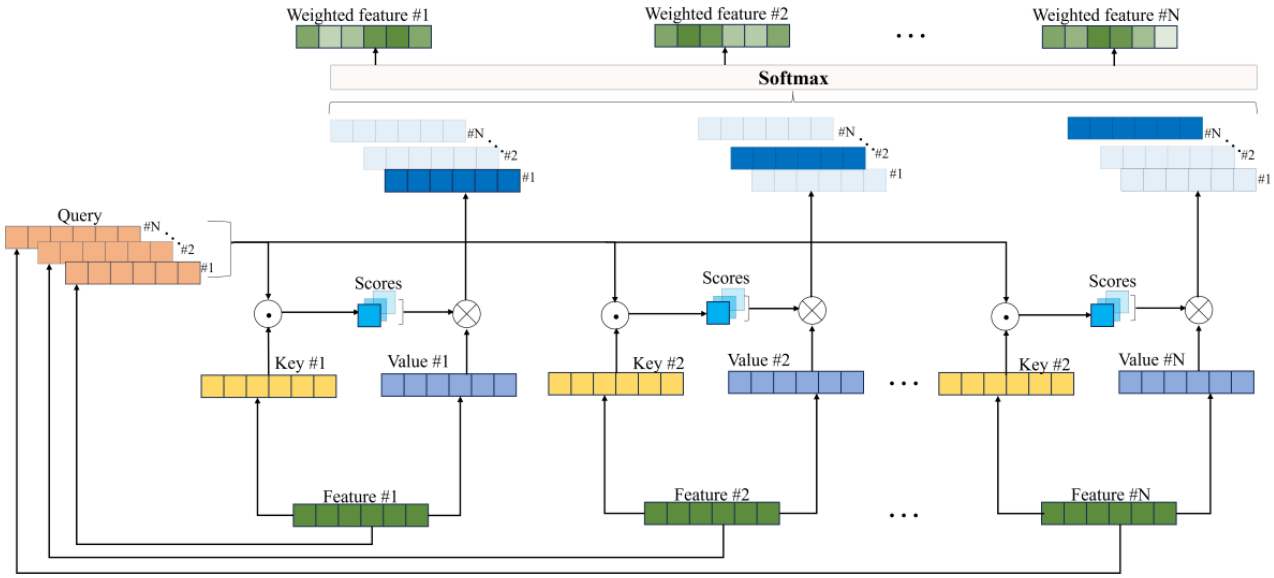

Classification Module

The classification module is used to transform the input-weighted feature vectors into the final prediction labels. Specifically, each vector first passes through the BN module. The main function of the BN module is to normalize input data, accelerate model training, and improve the model's adaptability to different features. Next, the dropout layer is deployed, which randomly sets the output of a portion of neurons to 0, to reduce the risk of overfitting and improve the model's generalization ability. After BN and dropout, the feature vectors pass through a fully connected layer. Finally, the output of the fully connected layer passes through the softmax layer and generates the probability distribution of fall or ADL labels. The final predicted label generated by the model corresponds to the category with the highest probability.

Training Parameters

For each data set, we divided the participants in the data set into training and testing groups with a ratio of 0.8 and 0.2, respectively. The proposed model was deployed in Python (version 3.8.10; Python Software Foundation) using the DL library of *PyTorch* (version 1.7.0). The training process was carried out on a computer server equipped with an Intel Xeon Gold 6330 CPU at 2.00 GHz and an NVIDIA GeForce RTX 3090. We used the cross-entropy loss function and used the Adam optimizer for optimizing the network parameters. In the training phase, we used a maximum of 300 iterations with an initial learning rate set at 0.001, gradually decreasing as the training progressed. The batch size was configured at 128 for optimal model refinement. The hyperparameter selection method for training the model is a random search. Parameters are randomly selected from the hyperparameter space, such as batch size values of 32, 64, and 128, and learning rate values of 0.001, 0.005, and 0.01. By exploring various combinations of batch sizes and learning rates, the optimal hyperparameters are identified based on the performance metrics. This approach ensures that the most effective combination is chosen to achieve the best model performance.

Performance Evaluation Metrics

To evaluate the fall detection performance of the proposed DSCS model, we used accuracy, recall, precision, and F_1-score as the evaluation metrics. Accuracy is the ratio of the number of segments that are correctly predicted (including falls and ADL) to the total number of segments, which is used to measure the overall performance of the model. Recall is the ratio of the number of correctly predicted fall segments to the total number of fall segments, and high recall means that the model has a strong ability to detect falls. Precision is the ratio of the number of correctly predicted fall segments to the total number of predicted fall segments. High precision means that the model can reduce the number of false alarms and accurately distinguish falls from fall-like activities. F_1-score is calculated based on recall and precision: $F_1\text{-}score = 2 \times precision \times recall / (precision + recall)$. In practical validation, we also used specificity to evaluate the model's performance in accurately distinguishing falls from ADLs, including fall-like activities. Specificity is defined as the ratio of correctly predicted ADL segments to the total number of ADL segments.

Ethical Considerations

The studies involving human participants were reviewed and approved by the Ethics Committee of the Capital University of Physical Education and Sports, Beijing, China (approval number 2023A036). The participants provided their written informed consent to participate in this study.

XSL•FO
RenderX

Results

Detection Performance

The performance of the proposed DSCS model was evaluated using the publicly available SisFall and MobiFall data sets. The model demonstrated robust performance on both data sets, achieving an accuracy of 99.32% (recall=99.15%; precision=98.58%; F_1-score=98.86%) on the testing sets of SisFall and an accuracy of 99.65% (recall=100%; precision=98.39%; F_1-score=99.19%) on the testing sets of MobiFall.

Figure 3 illustrates the confusion matrices for the DSCS model across 2 test sets: SisFall and MobiFall. Notably, out of 1183 test samples from the SisFall data set, only 8 samples were misclassified (5 false positives and 3 false negatives). Similarly, among the 285 test samples from the MobiFall data set, only 1 sample was misclassified (1 false positive). The result shows that the model can achieve high classification accuracy on both data sets.

Figure 3. Detection performance of the proposed DSCS model: (A) confusion matrix of SisFall test set and (B) confusion matrix of MobiFall test set. ADL: activities of daily life; DSCS: dual-stream convolutional neural network self-attention.

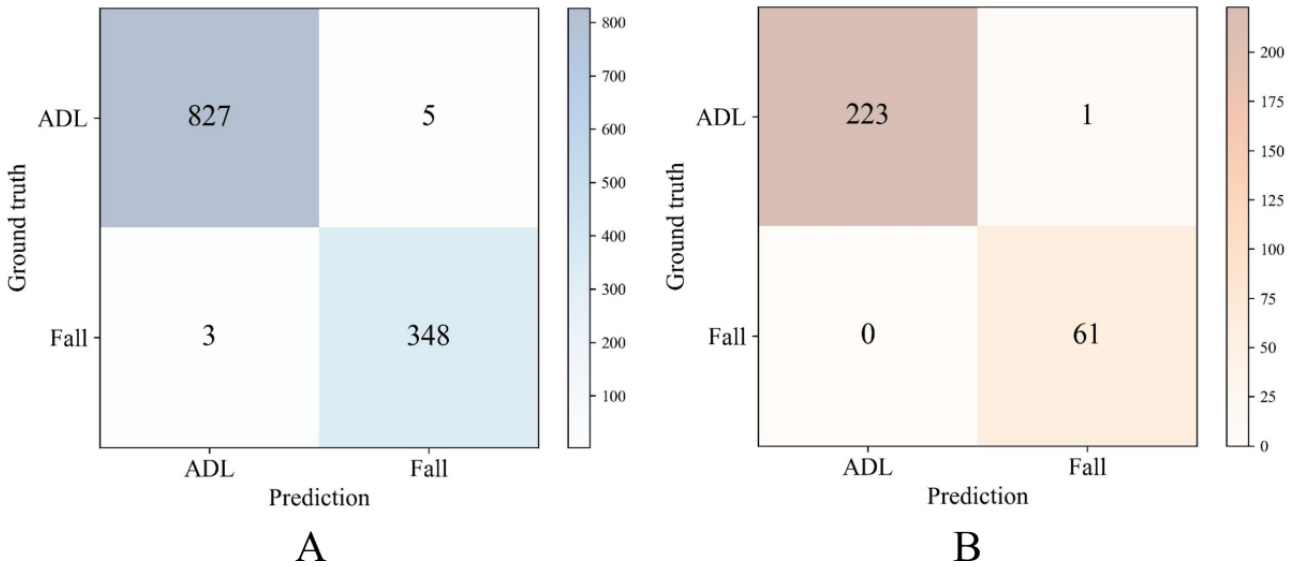

Cross-Validation

To further validate the generalization ability and reliability of the proposed model, we conducted cross-validation to thoroughly assess its performance. The results in Tables 1 and 2 indicate that the proposed model maintains stable performance on both the SisFall and MobiFall data sets. Specifically, the average F_1-scores for 5-fold cross-validation were 99.09% and 98.69%, respectively, while for 10-fold cross-validation, the average F_1-scores were 99.03% and 98.67%, respectively.

Table 1. F1-score of the proposed dual-stream convolutional neural network self-attention model under 5-fold cross-validation.

Data set	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Average (%)
SisFall	99.29	99.01	99.57	98.86	98.72	99.09
MobiFall	99.19	98.36	98.33	98.39	99.17	98.69

Table 2. F1-score of the proposed dual-stream convolutional neural network self-attention model under 10-fold cross-validation.

Data set	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold5 (%)	Fold 6 (%)	Fold 7 (%)	Fold 8 (%)	Fold9 (%)	Fold 10 (%)	Average (%)
SisFall	99.15	99.43	99.43	99.44	99.72	99.15	99.43	98.86	98.58	99.15	99.03
MobiFall	98.36	100	98.31	98.31	100	98.36	96.67	100	98.36	98.31	98.67

Comparison With State-of-the-Art Algorithms

We compared our proposed DSCS model with state-of-the-art fall detection models. Specifically, machine learning models—including KNN [32], SVM [33], DT [25]—and DL models—such as CNN [26], LSTM [27], CNN-LSTM [34], contrastive accelerometer-gyroscope embedding [35], few-shot transfer learning [42], and deep convolutional LSTM (DeepConvLSTM) [43]—were used as benchmark models. The results in Table 3 demonstrate the superior performance of the DSCS model. On the SisFall data set, the DSCS model achieved the second-best performance, only slightly behind KNN, which, however, required manual feature extraction. On the MobiFall data set, the DSCS model outperformed all state-of-the-art machine learning and DL models in terms of accuracy, recall, precision, and F_1-score.

XSL•FO
RenderX

Table 3. Performance comparison of the proposed DSCS[a] model with state-of-the-art fall detection algorithms.

Data set, algorithms, and corresponding models	Accuracy (%)	Recall (%)	Precision (%)	F_1-score (%)
SisFall				
ML[b]				
DT[c] [25]	97.97	96.30	96.85	96.57
KNN[d] [32]	99.41	99.43	98.59	99.01
SVM[e] [33]	99.24	98.86	98.58	98.72
DL[f]				
CNN[g] [26]	97.97	97.44	95.80	96.61
LSTM[h] [27]	97.80	96.01	96.56	96.29
CNN-LSTM [34]	98.73	97.15	98.55	97.85
CAGE[i] [35]	98.90	98.29	98.01	98.15
FSTL[j] [42]	98.65	97.44	97.99	97.71
DeepConvLSTM[k] [43]	98.99	98.58	98.02	98.30
DSCS	99.32	99.15	98.58	98.86
MobiFall				
ML				
DT [25]	94.74	95.08	82.86	88.55
KNN [32]	98.60	96.72	96.72	96.72
SVM [33]	96.14	90.16	91.67	90.91
DL				
CNN [26]	95.79	88.52	91.53	90.00
LSTM [27]	94.74	85.25	89.66	87.39
CNN-LSTM [34]	98.60	96.72	96.72	96.72
CAGE [35]	98.95	98.36	96.77	97.56
FSTL [42]	97.19	95.08	92.06	93.55
DeepConvLSTM [43]	98.25	95.08	96.67	95.87
DSCS	99.65	100	98.39	99.19

[a]DSCS: dual-stream convolutional neural network self-attention.
[b]ML: machine learning.
[c]DT: decision tree.
[d]KNN: k-nearest neighbors.
[e]SVM: support vector machine.
[f]DL: deep learning.
[g]CNN: convolutional neural network.
[h]LSTM: long short-term memory.
[i]CAGE: contrastive accelerometer-gyroscope embedding.
[j]FSTL: few-shot transfer learning.
[k]DeepConvLSTM: deep convolutional LSTM.

Effect of Low Sampling Rate

To validate the performance of the DSCS model under low sampling rates, we reduced the sampling rate of acceleration and gyroscope data to 10 Hz and 5 Hz and tested the performance of the DSCS model. Table 4 presents the overall performance comparison of the DSCS model under the original sampling rate of 50 Hz and the reduced sampling rates of 10 Hz and 5 Hz. The results show that the accuracy decreased under a sampling rate of 10 Hz and further declined under a sampling rate of 5 Hz. These findings indicate that the variation in

XSL•FO
RenderX

sampling rates significantly impacts the accuracy of fall detection.

Table 4. Performance comparison of the proposed DSCS[a] model under different sampling rates.

Data set and model	Accuracy (%)	Recall (%)	Precision (%)	F_1-score (%)
SisFall				
DSCS	99.32	99.15	98.58	98.86
DSCS (10 Hz)	97.63	96.58	95.49	96.03
DSCS (5 Hz)	95.60	92.88	92.35	92.61
MobiFall				
DSCS	99.65	100	98.39	99.19
DSCS (10 Hz)	96.49	95.08	89.23	92.06
DSCS (5 Hz)	94.04	86.89	85.48	86.18

[a]DSCS: dual-stream convolutional neural network self-attention.

Effect of SA Mechanism

To further explore the data distribution of feature vectors before and after passing through the SA module, we used the t-distributed stochastic neighbor embedding (t-SNE) algorithm for visualizing the test sets from SisFall and MobiFall. The t-SNE algorithm uses nonlinear dimensionality reduction to map the original 64D feature vectors to a more intuitive 2D space while preserving the similarity relationships among data points. The resulting visualization of data distribution through t-SNE allows us to assess the clustering and dispersion of different data categories (fall or ADL) in the 2D space, providing a better understanding and comparison of the separation between fall and ADL. As shown in Figure 4, after passing through the SA module, the feature vectors of different categories became more dispersed. This illustrates why the proposed DSCS algorithm performs better in fall detection and further confirms the effectiveness of the introduced SA mechanism.

Figure 4. t-SNE plot of (A) SisFall: before SA, (B) SisFall: after SA, (C) MobiFall: before SA, and (D) MobiFall: after SA. ADL: activities of daily life; SA: self-attention; t-SNE: t-distributed stochastic neighbor embedding.

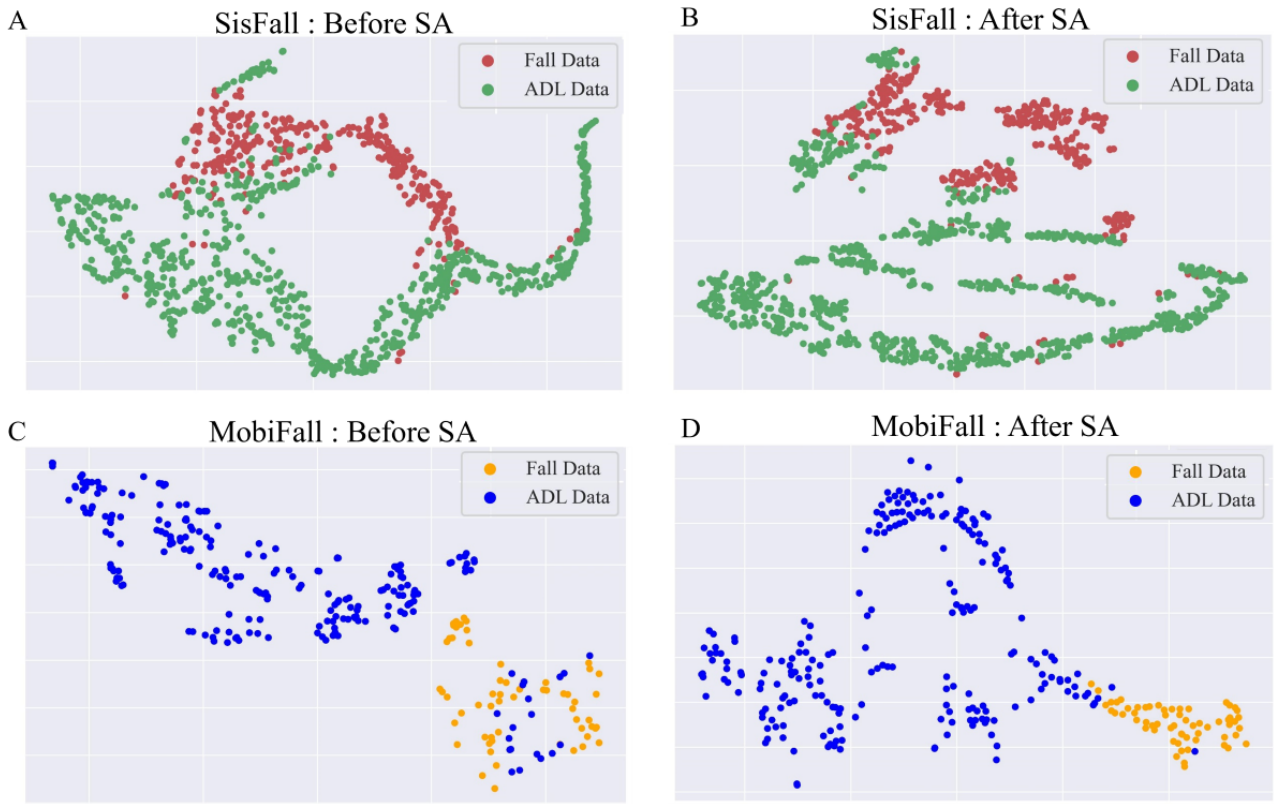

Practical Validation of DSCS

Beyond publicly available data sets, we conducted practical validation on the DSCS model, to assess its generalizability to new data and new users. Specifically, we embedded the DSCS model onto a smartwatch (Huawei Watch 3, Huawei Tech Co, Ltd) equipped with accelerometers and gyroscopes and developed a fall detection alert application. The sensors in the

XSL•FO
RenderX

smartwatch operated at a sampling rate of 50 Hz. As depicted in Figure 5A, participants wore smartwatches during experiments. Every 0.5 seconds, the data from the past 8 seconds (400 sampling points) was fed into the fall detection algorithm. When the algorithm detected a fall, a fall alert page, as shown in Figure 5B, was triggered.

Figure 5. Illustration of practical validation details. (A) Each participant wore a watch. (B) Watch alarm page: prompting detection of fall behavior.

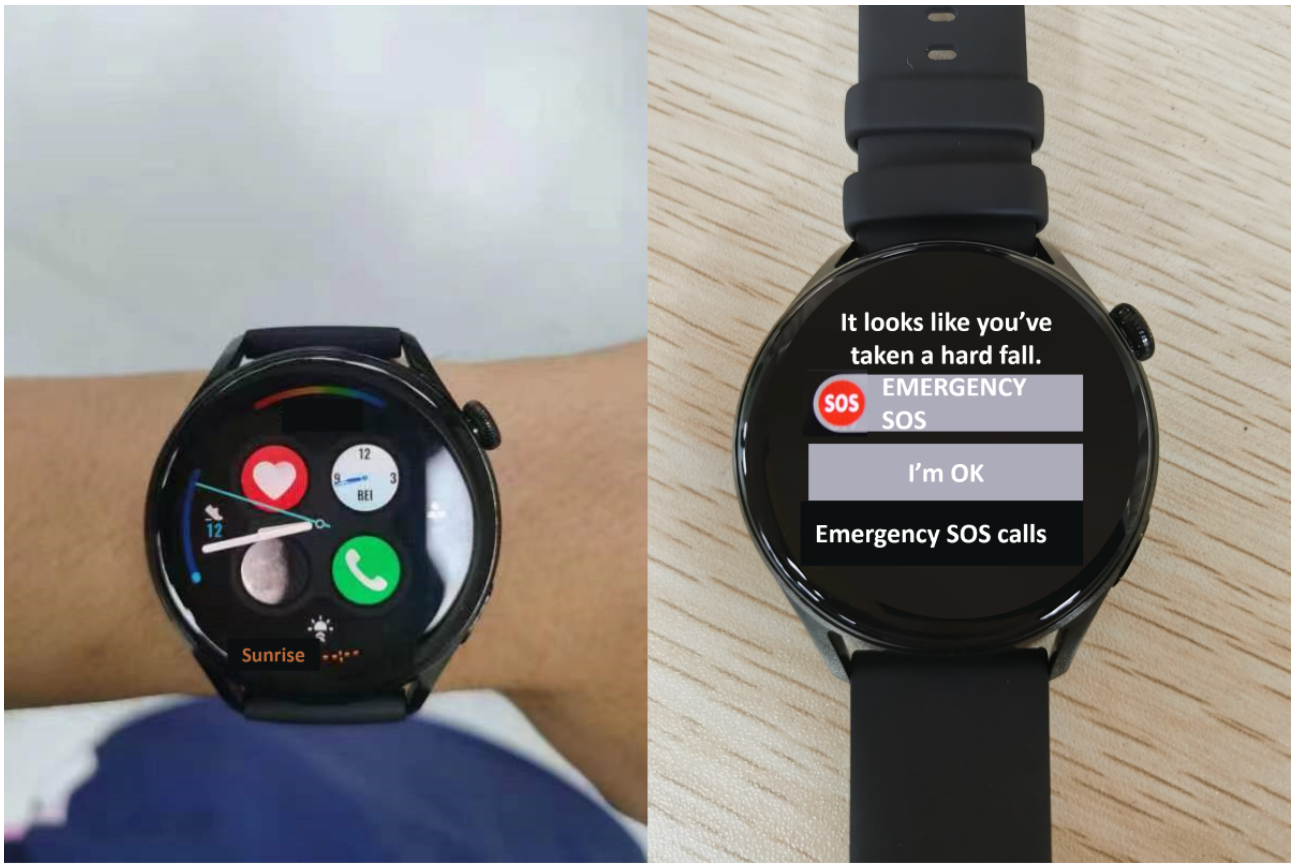

A B

We recruited 10 healthy students from the Capital University of Physical Education and Sports, Beijing, China as participants. The demographics of the participants are detailed in Table 5. Each participant wore a smartwatch and performed 8 fall activities and 9 fall-like activities. The eight fall activities were as follows: (1) fall forward, (2) fall backward, (3) fall to the left, (4) fall to the right, (5) fall from chairs, (6) fall while walking, (7) fall while running, and (8) fall while riding. The nine fall-like activities were as follows: (1) walking, (2) jogging, (3) sprinting, (4) standing and then sitting heavily in a chair, (5) bending down to tie shoelaces, (6) stretching and dropping hands, (7) long jump, (8) descending stairs, and (9) free fall on a trampoline. Figure 6 depicts participants wearing a smartwatch while performing fall activities during the validation process.

Table 5. Basic information of the 10 participants.

Information	Participant									
	1	2	3	4	5	6	7	8	9	10
Age (years)	24	25	24	25	35	26	25	26	27	35
Sex	Female	Male	Male	Female	Male	Male	Male	Female	Male	Male
Height (cm)	170	175	183	168	177	180	179	168	178	182
Weight (kg)	53	65	92	70	75	76	75	55	70	80

XSL•FO
RenderX

Figure 6. The 4 typical fall activities performed by the participants.

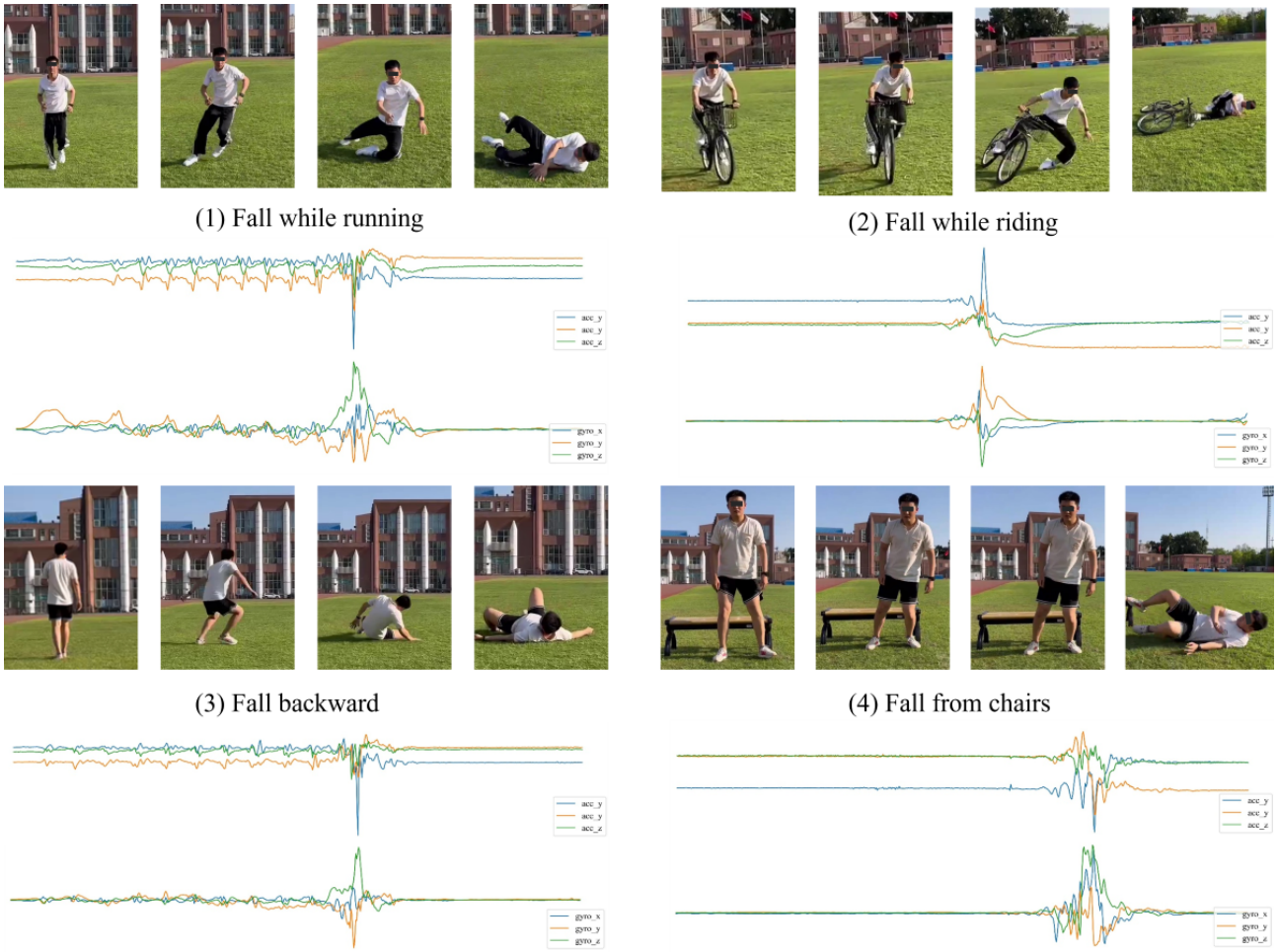

Each activity was repeated in 100 trials, resulting in a total of 1700 trials (800 fall trials and 900 fall-like trials). The DSCS model achieved satisfactory performance in practical validation, with accuracy, recall, and specificity of 96.41%, 95.12%, and 97.55%, respectively. Table 6 illustrates the recall performance for fall activities and the specificity performance for fall-like activities. The experimental results indicate that the DSCS model maintains a robust fall detection performance in practical applications and possesses satisfactory generalization ability.

XSL•FO
RenderX

Table 6. Performance validation of different falls and fall-like activities of daily life.

Performance validation and activity	Recall (%)	Specificity (%)
Recall performance		
Fall forward	97	—[a]
Fall to the left	96	—
Fall to the right	96	—
Fall from chairs	96	—
Fall while running	96	—
Fall while walking	94	—
Fall while riding	93	—
Fall backward	93	—
Specificity performance		
Descending stairs	—	100
Stand and then sit heavily in a chair	—	100
Jogging	—	100
Walking	—	100
Bend down to tie shoelaces	—	99
Stretch and drop hands	—	97
Long jump	—	95
Sprinting	—	95
Free fall on a trampoline	—	92

[a]Not applicable.

Discussion

Principal Findings

In this paper, we proposed a DSCS model for fall detection. The proposed model not only performs well on public data sets: SisFall and MobiFall, but also excels in practical validation. By comparing the performance of fall recognition results, we found that the dual-stream mechanism can effectively improve the accuracy of classification. At the same time, from the visualization results of data distribution, it can be observed that after passing through the SA module, the distance between fall data and ADL data becomes larger, which is beneficial for accurately distinguishing falls and ADLs.

Comparison With Prior Work

Unlike previous studies, this paper integrates the dual-stream data processing architecture and SA module for the first time. Compared with other DL schemes, such as CNN, LSTM, CNN-LSTM, contrastive accelerometer-gyroscope embedding, few-shot transfer learning, and DeepConvLSTM, the proposed scheme in this paper can automatically focus on the features closely related to fall behavior, to achieve better classification performance. In Wang et al [44], the attention mechanism is introduced into the fall detection algorithm. However, in the feature extraction stage, it still uses a machine learning algorithm for manual extraction. Comparative research shows that on the SisFall data set, the proposed model achieved the second-best accuracy (only second to KNN); on the MobiFall data set, the proposed model achieved the best accuracy, recall, and precision.

Performance Gap Between Public Data Sets and Practical Validation

Beyond public data sets, we embedded the proposed model into smartwatches and developed fall detection and alarm software to verify the fall detection algorithm proposed in this paper in practical applications. According to experimental results, we found that the performance of the model in practical validation dropped, but it was still at a good level. Specifically, accuracy degraded to 96.41%, with recall and specificity being 95.12% and 97.55%, respectively. The main reasons for the performance degradation are differences in sensor models and sampling frequencies, differences in participant conditions, and variations in participant movement norms.

Limitations

Our research has a few limitations. First, due to practical constraints in data collection, the training and test data sets used in this study may not encompass all possible fall poses, introducing uncertainty regarding the model's performance in the presence of unique or novel fall scenarios. Second, for safety considerations, the exclusion of older adult data from both public data sets and actual validations during the training and testing of fall detection may compromise the model's accuracy in handling falls among older adults, who constitute the most vulnerable demographic.

XSL•FO
RenderX

Conclusions

In this paper, we present the DSCS model, a DL framework designed for fall detection. The DSCS uses a dual-stream architecture incorporating both acceleration and gyroscope data, followed by a 3-layer CNN, SA mechanism, and classification modules. The DSCS model outperforms state-of-the-art algorithms, achieving detection accuracies of 99.32% and 99.65% on the SisFall and MobiFall data sets, respectively. Furthermore, the model maintains a high practical validation accuracy of 96.41%. These results demonstrate the effectiveness of the CNN-SA architecture for classification tasks. This study also highlights how incorporating SA into FDSs can improve classification accuracy by focusing on critical segments of data that significantly contribute to distinguishing between falls and fall-like activities.

Acknowledgments

This work was funded by the National Key Research and Development Program of China (2020YFC2006200) and the Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES). This work was also supported by the China Postdoctoral Science Foundation under grant number 2024M750191.

Conflicts of Interest

None declared.

References

1. Falls. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/falls [accessed 2024-06-22]
2. Bergen G, Stevens MR, Burns ER. Falls and fall injuries among adults aged ≥65 years–United States, 2014. MMWR Morb Mortal Wkly Rep 2016;65(37):993-998 [FREE Full text] [doi: 10.15585/mmwr.mm6537a2] [Medline: 27656914]
3. Global Report on Aging and Health. USA: World Health Organization; 2015.
4. Liu J, Li X, Huang S, Chao R, Cao Z, Wang S, et al. A review of wearable sensors based fall-related recognition systems. Eng Appl Artif Intell 2023;121:105993. [doi: 10.1016/j.engappai.2023.105993]
5. Ferreira de Sousa FAS, Escriba C, Avina Bravo EG, Brossa V, Fourniols J, Rossi C. Wearable pre-impact fall detection system based on 3D accelerometer and subject's height. IEEE Sensors J 2022;22(2):1738-1745. [doi: 10.1109/jsen.2021.3131037]
6. Jung H, Koo B, Kim J, Kim T, Nam Y, Kim Y. Enhanced algorithm for the detection of preimpact fall for wearable airbags. Sensors 2020;20(5):1277 [FREE Full text] [doi: 10.3390/s20051277] [Medline: 32111090]
7. Jain R, Semwal VB. A novel feature extraction method for preimpact fall detection system using deep learning and wearable sensors. IEEE Sensors J 2022;22(23):22943-22951. [doi: 10.1109/jsen.2022.3213814]
8. Zhong C, Ng WWY, Zhang S, Nugent CD, Shewell C, Medina-Quero J. Multi-occupancy fall detection using non-invasive thermal vision sensor. IEEE Sensors J 2021;21(4):5377-5388. [doi: 10.1109/jsen.2020.3032728]
9. Inturi AR, Manikandan VM, Kumar MN, Wang S, Zhang Y. Synergistic integration of skeletal kinematic features for vision-based fall detection. Sensors 2023;23(14):6283 [FREE Full text] [doi: 10.3390/s23146283] [Medline: 37514578]
10. Ding J, Wang Y. A WiFi-based smart home fall detection system using recurrent neural network. IEEE Trans Consumer Electron 2020;66(4):308-317. [doi: 10.1109/tce.2020.3021398]
11. Sadreazami H, Bolic M, Rajan S. Fall detection using standoff radar-based sensing and deep convolutional neural network. IEEE Trans Circuits Syst II Express Briefs 2020;67(1):197-201. [doi: 10.1109/tcsii.2019.2904498]
12. Adnan SM, Irtaza A, Aziz S, Ullah MO, Javed A, Mahmood MT. Fall detection through acoustic local ternary patterns. Appl Acoust 2018;140:296-300. [doi: 10.1016/j.apacoust.2018.06.013]
13. Ahn S, Kim J, Koo B, Kim Y. Evaluation of inertial sensor–based pre-impact fall detection algorithms using public dataset. Sensors 2019;19(4):774 [FREE Full text] [doi: 10.3390/s19040774] [Medline: 30781886]
14. Siwadamrongpong W, Chinrungrueng J, Hasegawa S, Nantajeewarawat E. Fall detection and prediction based on IMU and EMG sensors for elders. 2022 Jun Presented at: 19th International Joint Conference on Computer Science and Software Engineering (JCSSE); June 22-25, 2022; Bangkok, Thailand. [doi: 10.1109/jcsse54890.2022.9836284]
15. Ngu AH, Metsis V, Coyne S, Srinivas P, Salad T, Mahmud U, et al. Personalized watch-based fall detection using a collaborative edge-cloud framework. Int J Neural Syst 2022;32(12):2250048. [doi: 10.1142/S0129065722500484] [Medline: 35972790]
16. Chen L, Li R, Zhang H, Tian L, Chen N. Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch. Measurement 2019;140:215-226. [doi: 10.1016/j.measurement.2019.03.079]
17. Şengül G, Karakaya M, Misra S, Abayomi-Alli OO, Damaševičius R. Deep learning based fall detection using smartwatches for healthcare applications. Biomed Signal Process Control 2022;71:103242. [doi: 10.1016/j.bspc.2021.103242]
18. Shawen N, Lonini L, Mummidisetty CK, Shparii I, Albert MV, Kording K, et al. Addendum of: fall detection in individuals with lower limb amputations using mobile phones: machine learning enhances robustness for real-world applications. JMIR Mhealth Uhealth 2017;5(12):e167 [FREE Full text] [doi: 10.2196/mhealth.9177] [Medline: 29261509]

XSL•FO
RenderX

19. Shahzad A, Kim K. FallDroid: an automated smart-phone-based fall detection system using multiple kernel learning. IEEE Trans Ind Inf 2019;15(1):35-44. [doi: 10.1109/tii.2018.2839749]
20. Kerdegari H, Mokaram S, Samsudin K, Ramli AR. A pervasive neural network based fall detection system on smart phone. J Ambient Intell Smart Environ 2015;7(2):221-230. [doi: 10.3233/ais-150306]
21. Karar ME, Shehata HI, Reyad O. A survey of IoT-based fall detection for aiding elderly care: sensors, methods, challenges and future trends. Appl Sci 2022;12(7):3276. [doi: 10.3390/app12073276]
22. Ahmad Qureshi S, Hussain L, Rafique M, Sohail H, Aman H, Rahat Abbas S, et al. EML-PSP: a novel ensemble machine learning-based physical security paradigm using cross-domain ultra-fused feature extraction with hybrid data augmentation scheme. Expert Syst Appl 2024;243:122863. [doi: 10.1016/j.eswa.2023.122863]
23. Saadeh W, Butt SA, Altaf MAB. A patient-specific single sensor IoT-based wearable fall prediction and detection system. IEEE Trans Neural Syst Rehabil Eng 2019;27(5):995-1003. [doi: 10.1109/TNSRE.2019.2911602] [Medline: 30998473]
24. Alizadeh J, Bogdan M, Classen J, Fricke C. Support vector machine classifiers show high generalizability in automatic fall detection in older adults. Sensors 2021;21(21):7166 [FREE Full text] [doi: 10.3390/s21217166] [Medline: 34770473]
25. Nahian MJA, Ghosh T, Banna MHA, Aseeri MA, Uddin MN, Ahmed MR, et al. Towards an accelerometer-based elderly fall detection system using cross-disciplinary time series features. IEEE Access 2021;9:39413-39431. [doi: 10.1109/access.2021.3056441]
26. Casilari E, Lora-Rivera R, García-Lagos F. A study on the application of convolutional neural networks to fall detection evaluated with multiple public datasets. Sensors 2020;20(5):1466 [FREE Full text] [doi: 10.3390/s20051466] [Medline: 32155936]
27. Luna-Perejón F, Domínguez-Morales MJ, Civit-Balcells A. Wearable fall detector using recurrent neural networks. Sensors 2019;19(22):4885 [FREE Full text] [doi: 10.3390/s19224885] [Medline: 31717442]
28. Lin Z, Wang Z, Dai H, Xia X. Efficient fall detection in four directions based on smart insoles and RDAE-LSTM model. Expert Syst Appl 2022;205:117661. [doi: 10.1016/j.eswa.2022.117661]
29. Wang X, Ellul J, Azzopardi G. Elderly fall detection systems: a literature survey. Front Robot AI 2020;7:71. [doi: 10.3389/frobt.2020.00071] [Medline: 33501238]
30. Alarifi A, Alwadain A. Killer heuristic optimized convolution neural network-based fall detection with wearable IoT sensor devices. Measurement 2021;167:108258. [doi: 10.1016/j.measurement.2020.108258]
31. Kim TH, Choi A, Heo HM, Kim H, Mun JH. Acceleration magnitude at impact following loss of balance can be estimated using deep learning model. Sensors 2020;20(21):6126 [FREE Full text] [doi: 10.3390/s20216126] [Medline: 33126491]
32. Hussain F, Hussain F, Ehatisham-ul-Haq M, Azam MA. Activity-aware fall detection and recognition based on wearable sensors. IEEE Sensors J 2019;19(12):4528-4536. [doi: 10.1109/jsen.2019.2898891]
33. Son H, Lim JW, Park S, Park B, Han J, Kim HB. A machine learning approach for the classification of falls and activities of daily living in agricultural workers. IEEE Access 2022;10:77418-77431. [doi: 10.1109/ACCESS.2022.3190618]
34. Liu L, Hou Y, He J, Lungu J, Dong R. An energy-efficient fall detection method based on FD-DNN for elderly people. Sensors 2020;20(15):4192 [FREE Full text] [doi: 10.3390/s20154192] [Medline: 32731465]
35. Koo I, Park Y, Jeong M, Kim C. Contrastive accelerometer-gyroscope embedding model for human activity recognition. IEEE Sensors J 2023;23(1):506-513. [doi: 10.1109/jsen.2022.3222825]
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones LA, Gomez AN, et al. Attention is all you need. 2017 Dec Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA, USA p. 5998-6008.
37. Fan X, Chen X, Ma W, Gao W. BAFNet: bottleneck attention based fusion network for sleep apnea detection. IEEE J Biomed Health Inform 2024;28(5):2473-2484. [doi: 10.1109/JBHI.2023.3278657] [Medline: 37216250]
38. Khan IU, Lee JW. PAR-Net: an enhanced dual-stream CNN-ESN architecture for human physical activity recognition. Sensors 2024;24(6):1908 [FREE Full text] [doi: 10.3390/s24061908] [Medline: 38544172]
39. Sucerquia A, López JD, Vargas-Bonilla JF. SisFall: a fall and movement dataset. Sensors 2017;17(1):198 [FREE Full text] [doi: 10.3390/s17010198] [Medline: 28117691]
40. Vavoulas G, Pediaditis M, Spanakis E, Tsiknakis M. The MobiFall dataset: an initial evaluation of fall detection algorithms using smartphones. : GREECE; 2013 Nov Presented at: IEEE 13th International Conference on Bioinformatics and Bioengineering (BIBE); November 10-13, 2013; Chania, Greece. [doi: 10.1109/bibe.2013.6701629]
41. Cho Y, Cho H, Kyung CM. Design and implementation of practical step detection algorithm for wrist-worn devices. IEEE Sensors J 2016;16(21):7720-7730. [doi: 10.1109/jsen.2016.2603163]
42. Ganesha HS, Gupta R, Gupta SH, Rajan S. Few-shot transfer learning for wearable IMU-based human activity recognition. Neural Comput Appl 2024;36(18):10811-10823. [doi: 10.1007/s00521-024-09645-7]
43. Hellmers S, Krey E, Gashi A, Koschate J, Schmidt L, Stuckenschneider T, et al. Comparison of machine learning approaches for near-fall-detection with motion sensors. Front Digit Health 2023;5:1223845 [FREE Full text] [doi: 10.3389/fdgth.2023.1223845] [Medline: 37564882]
44. Wang G, Li Q, Wang L, Zhang Y, Liu Z. CMFALL: a cascade and parallel multi-state fall detection algorithm using waist-mounted tri-axial accelerometer signals. IEEE Trans Consumer Electron 2020;66(3):261-270. [doi: 10.1109/tce.2020.3000338]

XSL•FO
RenderX

Abbreviations

ADL: activities of daily life
BN: batch normalization
CNN: convolutional neural network
DeepConvLSTM: deep convolutional LSTM
DL: deep learning
DSCS: dual-stream convolutional neural network self-attention
DT: decision tree
FDS: fall detection system
KNN: k-nearest neighbor
LSTM: long short-term memory
ReLU: rectified linear unit
SA: self-attention
SVM: support vector machine
t-SNE: t-distributed stochastic neighbor embedding

Edited by Q Jin; submitted 25.01.24; peer-reviewed by MAU Alam, SA Qureshi, S Okita; comments to author 14.05.24; revised version received 07.06.24; accepted 11.06.24; published 05.08.24.

Please cite as:
Zhang J, Li Z, Liu Y, Li J, Qiu H, Li M, Hou G, Zhou Z
An Effective Deep Learning Framework for Fall Detection: Model Development and Study Design
J Med Internet Res 2024;26:e56750
URL: https://www.jmir.org/2024/1/e56750
doi:10.2196/56750
PMID:

©Jinxi Zhang, Zhen Li, Yu Liu, Jian Li, Hualong Qiu, Mohan Li, Guohui Hou, Zhixiong Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating Best Practices for Ecological Momentary Assessment: Nationwide Factorial Experiment

Michael S Businelle[1,2], PhD; Emily T Hébert[3], DPH; Dingjing Shi[4], PhD; Lizbeth Benson[1,5], PhD; Krista M Kezbers[1], PhD; Sarah Tonkin[1], PhD; Megan E Piper[6], PhD; Tianchen Qian[7], PhD

[1]TSET Health Promotion Research Center, Stephenson Cancer Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[2]Department of Family and Preventive Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[3]School of Public Health, University of Texas Health Science Center at Houston, Houtson, TX, United States

[4]Department of Psychology, University of Oklahoma, Norman, OK, United States

[5]Institute for Social Research, University of Michigan, Ann Arbor, MI, United States

[6]Center for Tobacco Research and Intervention, Department of Medicine, University of Wisconsin, Madison, Madison, WI, United States

[7]Department of Statistics, University of California, Irvine, Irvine, CA, United States

Corresponding Author:
Michael S Businelle, PhD
TSET Health Promotion Research Center
Stephenson Cancer Center
University of Oklahoma Health Sciences Center
655 Research Parkway
Suite 400
Oklahoma City, OK, 73104
United States
Phone: 1 405 271 8001 ext 50460
Email: michael-businelle@ouhsc.edu

Abstract

Background: Ecological momentary assessment (EMA) is a measurement methodology that involves the repeated collection of real-time data on participants' behavior and experience in their natural environment. While EMA allows researchers to gain valuable insights into dynamic behavioral processes, the need for frequent self-reporting can be burdensome and disruptive. Compliance with EMA protocols is important for accurate, unbiased sampling; yet, there is no "gold standard" for EMA study design to promote compliance.

Objective: The purpose of this study was to use a factorial design to identify optimal study design factors, or combinations of factors, for achieving the highest completion rates for smartphone-based EMAs.

Methods: Participants recruited from across the United States were randomized to 1 of 2 levels on each of 5 design factors in a 2×2×2×2×2 design (32 conditions): factor 1—number of questions per EMA survey (15 vs 25); factor 2—number of EMAs per day (2 vs 4); factor 3—EMA prompting schedule (random vs fixed times); factor 4—payment type (US $1 paid per EMA vs payment based on the percentage of EMAs completed); and factor 5—EMA response scale type (ie, slider-type response scale vs Likert-type response scale; this is the only within-person factor; each participant was randomized to complete slider- or Likert-type questions for the first 14 days or second 14 days of the study period). All participants were asked to complete prompted EMAs for 28 days. The effect of each factor on EMA completion was examined, as well as the effects of factor interactions on EMA completion. Finally, relations between demographic and socioenvironmental factors and EMA completion were examined.

Results: Participants (N=411) were aged 48.4 (SD 12.1) years; 75.7% (311/411) were female, 72.5% (298/411) were White, 18.0% (74/411) were Black or African American, 2.7% (11/411) were Asian, 1.5% (6/411) were American Indian or Alaska Native, 5.4% (22/411) belonged to more than one race, and 9.6% (38/396) were Hispanic/Latino. On average, participants completed 83.8% (28,948/34,552) of scheduled EMAs, and 96.6% (397/411) of participants completed the follow-up survey. Results indicated that there were no significant main effects of the design factors on compliance and no significant interactions. Analyses also indicated that older adults, those without a history of substance use problems, and those without current depression tended to complete more EMAs than their counterparts. No other demographic or socioenvironmental factors were related to

XSL•FO
RenderX

EMA completion rates. Finally, the app was well liked (ie, system usability scale score=82.7), and there was a statistically significant positive association between liking the app and EMA compliance.

Conclusions: Study results have broad implications for developing best practices guidelines for future studies that use EMA methodologies.

Trial Registration: ClinicalTrials.gov number NCT05194228; https://clinicaltrials.gov/study/NCT05194228

(J Med Internet Res 2024;26:e50275) doi:10.2196/50275

KEYWORDS

ecological momentary assessment; mobile health; smartphone; compliance; ambulatory assessment; adherence; experience sampling; mobile phone; mHealth; real-time data; behavior; dynamic behavioral processes; self-report; factorial design

Introduction

Ecological momentary assessment (EMA) is a measurement methodology that involves the repeated collection of real-time data on participants' behavior and experience in their natural environment [1]. EMA has been used in behavioral science for years [2], including research examining the environmental and psychological antecedents of cigarette smoking, substance use disorders, anxiety, eating, and sleep [3-6]. EMA sampling strategies are typically time-based or event-based [7]. Time-based sampling often involves prompting EMAs on an interval or random schedule; for example, a daily diary prompted every day at the same time, or at a randomized time within a specific window (eg, every 3-4 hours) each day [1]. In contrast, event-based sampling is typically initiated by a participant when specific events occur. For example, a participant may be asked to initiate an assessment every time they smoke a cigarette [1].

The EMA methodology has several advantages over more traditional lab or clinic-based sampling strategies [1], including (1) minimization of recall bias due to frequent and real-time assessments, (2) more granular examination of the contextual associations between behaviors and psychological and environmental variables, and (3) examination of behavioral trends over time and across situations and settings. While EMA allows researchers to gain valuable insights into dynamic behavioral processes, the need for frequent self-reporting can be burdensome and disruptive to participants [1]. As a result, missing data due to poor compliance with EMA protocols can undermine the validity of the data. Although there are no "gold standards" for EMA protocols, the development of balanced procedures to maximize data coverage and minimize participant burden is important for accurate, unbiased sampling [7,8].

Few studies have experimentally examined the impact of EMA design features on participant compliance. Stone et al [9] randomized 91 patients with chronic pain to 1 of 4 EMA monitoring intensities (ie, 0, 3, 6, or 12 EMAs per day) for 14 days. Results indicated that compliance with the EMA protocol was not related to sampling density. Eisele et al [10] assigned a sample of 152 college students to complete 30 or 60 item EMAs 3, 6, or 9 times per day. Results showed that compliance was lower for those assigned to complete EMAs with more items, but sampling frequency was unrelated to compliance. Finally, Hasselhorn et al [11] randomized college students into 2 separate 14-day studies to examine relations between the number of EMAs prompted per day and EMA compliance (ie, 3 or 9 EMAs per day, N=313; study 1) and the number of EMA items per assessment and EMA compliance (ie, 33- or 82-item EMAs 3 times per day, N=282; study 2). Findings indicated that neither the number of EMAs per day nor the number of items per EMA were related to EMA compliance.

A number of meta-analyses and systematic reviews have also been conducted to examine how EMA design features may affect compliance. In a recent meta-analysis of compliance with EMA protocols among people who use substances (N=126 studies), there was no evidence that compliance rates were associated with prompt frequency, length of assessment period, or type of device used to prompt EMAs [8]. Another study using a pooled dataset of 10 EMA studies found compliance declined over time, women were more compliant than men, older adults were more compliant than younger adults, and compliance varied significantly depending on the time of day [12]. Ottenstein and Werner's recent review of 488 EMA studies [13] showed longer study periods and higher total numbers of EMAs were related to lower overall EMA compliance. In a meta-analysis of EMA studies in mental health research, greater compliance was associated with female sex, higher incentives for completing EMAs, having EMA prompts at fixed times, and fewer EMAs per day; however, study duration and average age of the sample were not related to EMA compliance [14]. One of the most comprehensive meta-analyses of EMA studies to date (ie, 105 trials were included) indicated that the median study duration was 7 days, a median of 5 EMAs were prompted per day, EMAs included a median of 8-10 questions, and 81.9% of all EMAs were completed [15]. Findings also indicated that studies that had 3 or fewer daily prompts had greater EMA completion than studies with greater than 3 daily prompts, and studies with less than 27 EMA items had higher completion rates than EMAs with 27 or more items. Finally, effect estimates for predictors of compliance varied widely across studies [15]. While the existing literature can offer some guidance regarding best practices for EMA study design, varied reporting across studies with respect to the definition of compliance, study design elements, and study populations make it difficult to clearly derive best practice guidelines for EMA design. To identify optimal EMA design features, it is necessary to use a randomized controlled trial to support the strongest level of inference.

This study used a factorial design to identify optimal factors, or combinations of factors, for achieving the best compliance rates for smartphone-based EMAs. A factorial design is ideal for exploring these research questions because it is statistically

XSL•FO
RenderX

more efficient, as this type of design needs fewer participants to answer questions about each experimental factor of interest and allows researchers to examine interaction effects. The included study factors were specifically selected to inform future studies that aim to use EMA data to tailor just-in-time adaptive interventions (JITAIs). Our hypotheses were that fewer EMA items (ie, 15 vs 25), greater EMA frequency (ie, 4 vs 2 per day), paying participants based upon the percentage of EMAs completed (versus US $1 per EMA completed), fixed EMA prompt type (versus random prompts), and slider-type response scale (vs Likert-type scales) would result in greater EMA completion. Finally, we examined the relation between participant demographics (eg, age and sex) and other variables (eg, time of day, day of week) and EMA completion across the 28-day study period.

Methods

Procedures

Participants (N=411) were recruited through nationwide Facebook advertisements from November 30, 2021, through September 27, 2022. One example advertisement read: “We’re looking for Android users to complete brief daily surveys on their smartphones. Qualified participants will be compensated up to $152 over 4 weeks. NO in person visits required.” Those interested in participating clicked a link to a secure, REDCap (Research Electronic Data Capture; Vanderbilt University)-based survey to determine their initial eligibility for the study. Upon screening for the study, all participants were scheduled for a 20- to 30-minute screening and enrollment phone call. During the phone screening, all potential participants completed informed consent electronically via REDCap. Individuals were included in the study if they (1) were ≥18 years of age, (2) demonstrated >6th grade English literacy level, (3) possessed an active Android smartphone with a data plan and an operating system (version 6.0) or higher that was compatible with our Insight smartphone app, (4) agreed to install the Insight app onto their personal phone, (5) texted the study team a picture of their identification, including address (to reduce the likelihood of fraudulent attempts to enroll in the study), and (6) agreed to complete EMAs prompted by and completed through the study smartphone app. Finally, participants were required to identify 14 continuous waking hours each day in which the app could prompt EMAs. Participants were then randomized into 1 of 32 groups, downloaded the Insight app onto their personal smartphone, and were given a unique code to initialize the app. Participants were briefly trained to use the intuitive Insight app for data collection purposes. Participants were instructed to click the “Call Staff” button to speak with study research staff in the event of difficulty with the phone app, and to click the “Payment” button to access an up-to-the-moment accounting of credit earned for all EMAs completed (see Figure 1). All participants completed sample EMA items during the screening phone call to ensure that they understood how to use the app to answer study questions. Participants were given 1 week to complete the baseline assessment in the Insight app. Upon completion of the baseline assessment, participants began the 28-day EMA period. After the 28-day EMA period, participants gained access to a follow-up assessment that was embedded in the Insight app. See Figure 2 for the participant flow through the study. The study smartphone app automatically encrypted and uploaded survey data to our server multiple times per day.

All participants were prompted to complete their assigned EMA battery for the entire 28-day EMA period. The study app automatically prompted brief (ie, approximately 1-3 minutes) EMAs during each individual’s normal waking hours (ie, daily wake and sleep times were collected after participants were screened into the study).

XSL•FO
RenderX

Figure 1. Exemplar home screen.

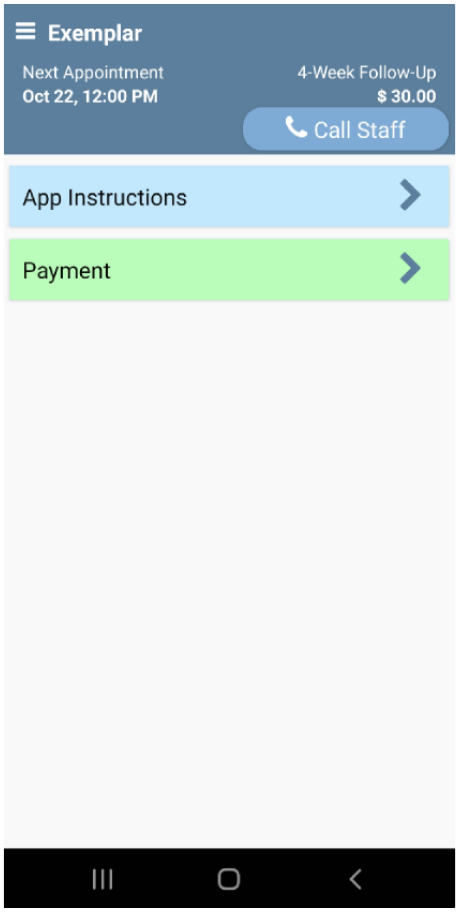

Figure 2. Participant flow.

Study Design

This study was designed to examine the impact of 5 EMA design factors that are relevant and tuned for studies that may use EMA data to inform JITAIs: (1) number of EMA items (15 vs 25 items), (2) EMA frequency (2 vs 4 EMAs per day), (3) payment type (US $1 per completed EMA versus payment based upon the percentage of EMAs completed), (4) EMA prompt type (EMAs prompted at random versus fixed times of the day, assessment times are presented in Table 1), and (5) EMA response scale type (slider-type response scale vs Likert-type response scale). All design factors were specified at the between-person level (ie, each person received 1 level or the other for the entire study period), except for the EMA response scale factor. The EMA response scale factor was specified at the within-person level such that each participant received 1 level (ie, slider scale or Likert scale) for the first 14 days and subsequently received the other level of the factor for the last 14 days of the study. The order of EMA response scale types was counterbalanced across participants, such that approximately half of participants received the slider-type scale for days 1-14 and the Likert-type scale for days 15-28, and the other half of participants received the scale types in the opposite order. The 32 experimental conditions are described in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Random and fixed ecological momentary assessment (EMA) prompt times.

EMA order	Fixed		Random	
	2 per day	4 per day	2 per day	4 per day
First EMA	30 minutes after waking	30 minutes after waking	First half of waking hours	First quarter of waking hours
Second EMA	—[a]	300 minutes after waking	—	Second quarter of waking hours
Third EMA	—	570 minutes after waking	—	Third quarter of waking hours
Fourth EMA	60 minutes before bed	60 minutes before bed	Second half of waking hours	Fourth quarter of waking hours

[a]Not applicable.

Study Assessments and Measures

Screening Assessment (via REDCap)

Upon viewing the study Facebook advertisements, potential participants self-initiated the initial screening using the secure web-based REDCap platform [16]. Those who met the eligibility criteria were scheduled for a 20- to 30-minute call to complete the informed consent procedure, additional screening items, and confirm the compatibility of their personal smartphone with the Insight app.

Baseline and Follow-Up Assessments

The baseline and follow-up assessments were completed via the Insight app. Participants answered questions about their ethnicity (Hispanic or Latino or not), race (more than one race, American Indian, Asian, Black, Native Hawaiian or other Pacific Islander, or White; dichotomized as White versus minoritized race in the current analyses), biological sex, employment status (dichotomized as employed at least part-time or not employed), years of education (continuous), and income (continuous). Other baseline measures included the 5-item Overall Anxiety Severity and Impairment Scale (a measure of current anxiety [17]), the Overall Depression Severity and Impairment Scale (a measure of current depression [18]), and a single item that asked, “Has your use of any of the following substances caused you significant problems?” (check all that apply: alcohol, cannabis, cocaine, K2, opiates, sedatives or hypnotics or anxiolytics, other substances, and none of the above). A single question was used to gauge health literacy (ie, “How confident are you filling out medical forms by yourself?” from 1=Extremely to 5=Not at all [19]). Physical and mental health were also assessed. Specifically, history of mental illness was coded as 1 if a participant indicated ever having been diagnosed with depression, schizophrenia or schizoaffective disorder, bipolar disorder, posttraumatic stress disorder, or an anxiety disorder besides posttraumatic stress disorder (all others were coded as 0). Similarly, those who reported ever receiving a diagnosis of a chronic disease, including cardiovascular diseases, pulmonary diseases, cancers, high blood pressure, diabetes, or arthritis, were coded as 1 (all others were coded as 0).

The 4-week follow-up assessment included many items that were assessed at baseline and some additional items. Participants completed the system usability scale (SUS) [20], a 10-item measure that is commonly used to assess the usability of smartphone applications (scores range from 0 to 100, with higher scores indicating greater usability; scores greater than 68 are considered above average, and scores above 80 are in the top 10th percentile) [21]. In addition, participants were asked several questions to assess other qualities of the study smartphone application. Specifically, participants were asked: “Consider the number of assessments that were automatically pushed by the smartphone application, was the number of assessments: 1=Too high, 2=About right, 3=Not enough”; “Did carrying the phone and answering questions make you more aware of your thoughts, feelings, and behavior?” with answer options ranging from 1=Definitely no to 4=Definitely yes; and “Do you find the smartphone application to be annoying?” with answer options ranging from 1=Not at all to 5=Extremely.

EMA

The EMA methodology used in this study is similar to what was used in our previous studies and by other researchers [22-31]. Throughout the 28 EMA days, the app prompted participants to complete each EMA by ringing and vibrating for 30 seconds. If the participant did not respond after 5 prompts, the EMA was automatically rescheduled, and participants were prompted again up to 2 additional times 15 minutes later. EMAs asked participants about their mood, environmental context, and health behaviors (eg, fruit and vegetable intake, physical activity, soda consumption). In addition, readiness to change unhealthy behaviors was assessed once per week during the first EMA of the day. Participant GPS locations were collected by the app multiple times per day to be used in future studies that examine location-based effects on mood, thoughts, and behaviors.

Additional Derived Variables

Additional variables were derived to determine if they were related to the EMA completion rate. Specifically, a variable indicating whether each EMA was completed on a weekday (=0) or weekend day (=1), time of day the EMA was completed (ie, morning=5 AM-11:59 AM, afternoon=12 PM-4:59 PM, evening=5 PM-8:59 PM, late night=9 PM-4:59 AM), and urban (=0) vs rural (=1) residence (based on address) were calculated and entered into the dataset. We also calculated a time period variable (*time*) corresponding to whether each EMA was completed during the first 14-day period of the study or the second 14-day period of the study.

App Programming

The Insight app was developed by the mHealth Shared Resource at the University of Oklahoma Health Sciences Center and NCI Designated Stephenson Cancer Center. The Insight mHealth platform enables researchers to rapidly build, test, and launch technology-based assessment and intervention tools. The

XSL•FO
RenderX

mHealth resource employs 9.5 staff members, including 5.5 computer scientists and engineers who develop and maintain web and mobile applications and relational databases.

Ethical Considerations

This study was approved by the University of Oklahoma Health Sciences Center Institutional Review Board (13684). All participants provided informed consent and were informed that they could opt out of the study at any time. Study datasets were deidentified. In addition to compensation for the baseline assessment (US $10 payments made via a reloadable Greenphire Clincard) and follow-up assessment (US $30 via Greenphire Clincard), all participants were compensated at study completion for either the percentage of EMAs that they completed or US $1 per EMA completed over the 4-week study period. Participants in the 4 EMAs per day groups could earn up to $112 for completing EMAs, and those in the 2 EMAs per day groups could earn up to US $56 for completing EMAs. Those randomly assigned to the groups that were paid based upon the percentage of EMAs completed received US $112 (4 per day groups) or US $56 (2 per day groups) for completing 95%-100% of the EMAs, US $90 (4 per day groups) or US $45 (2 per day groups) for completing 80%-94% of the EMAs, US $56 (4 per day groups) or US $28 (2 per day groups) for completing 50%-79% of the EMAs, and $0 for completing less than 50% of the EMAs. Thus, considering compensation for the baseline and follow-up assessments and EMAs, participants could earn up to US $96-$152 for the entire study. Study Greenphire cards were mailed to participants upon completion of the baseline assessment, and compensation was loaded onto the card after the participant indicated that they had received the card. A final follow-up and EMA completion payment were loaded onto the card after the follow-up assessment was completed.

Statistical Methods

Descriptive statistics were used to summarize participant demographics and engagement with the smartphone app. For all primary analyses, the percentage of EMAs completed over each of the two 14-day study periods was used as the primary outcome variable. Specifically, for each participant during each of the 2 study periods, the number of completed EMAs was divided by the number of total possible EMAs (ie, either 28 or 56, depending on whether the number of EMAs per day was set to 2 or 4) and multiplied by 100. Two compliance scores per person were calculated (instead of 1 total compliance score) to accommodate the within-person nature of the EMA response scale factor.

To examine how each of the design factors related to EMA compliance, we used multilevel (2-level) linear regression, while adjusting for several hypothesized person-level characteristics, including age, Hispanic ethnicity, race, sex, education level, employment status, income, rural versus urban address, current depression, current anxiety, mental diagnosis history, chronic illness diagnosis history, substance use problem history, health literacy, and SUS score. Effect coding was used for each of the 5 main effects: EMA payment type (–1=percentage-based payment, +1=$1 for each EMA), number of EMAs per day (–1=2 EMAs per day, +1=4 EMAs per day), number of items per EMA (–1=15 items, +1=25 items), prompt schedule (–1=random prompts, +1=fixed prompts), and EMA response scale (–1=Likert-type scale, +1=slider-type scale). All design factors, except for the EMA response scale, were at the between-person level in the model (level 2) and only varied between participants. The EMA response scale was at the within-person level in the model (level 1). In the main model, all within-level and cross-level interactions (main effects, 2-way, 3-way, 4-way, and 5-way) were allowed among the design factors. Additionally, time was added as a covariate at the within-person level (level 1; –1=first 14-day period, +1=second 14-day period). An initial adjusted model additionally accounted for between-person variation in EMA compliance due to the person-level characteristics described earlier. A final adjusted model was calculated by using backwards elimination (threshold set to $P<.10$) to identify which of the demographic and environmental predictors from the above list were independently related to EMA completion. A priori power analyses indicated that a sample size of 416 would allow us to detect meaningful differences in EMA completion rates. For example, these analyses indicated 0.81 power to detect a 12-percentage point difference in EMA completion rates of 70% versus 82% or a 10-percentage point difference in EMA completion rates of 80% versus 90%. These base rates were consistent with prior research [13].

To examine whether weekday and time of day were associated with EMA compliance, we created a dichotomous variable to represent whether or not an EMA was prompted on a weekday versus a weekend day, and we categorized EMAs based on the actual notification time on the participant's phone, with morning defined as 5 AM-11:59 AM, afternoon defined as 12 PM-4:59 PM, evening defined as 5 PM-8:59 PM, and late night defined as 9 PM-4:59 AM. EMAs in which the notification time was missing (eg, due to the phone being off or an app error) were excluded. We used multilevel logistic regression with EMA completion (yes or no) as the outcome, and time of day and weekday versus weekend as the independent variables in 2 separate models.

Finally, to examine whether participant compliance was associated with perceptions of the app (ie, SUS score, answers to whether the number of assessments was too high, whether they felt the app was annoying, and if the app made them more aware of their thoughts, feelings, and behaviors), we used Pearson correlation coefficients and independent samples t-tests. All analyses were performed using R Statistical Software (version 4.1.1; R Foundation for Statistical Computing).

Results

Participants

The study advertisement was clicked 5236 times, and 3547 individuals completed the study screener. A total of 1928 individuals were prescreened eligible to participate in this study, and 485 were enrolled and randomized into 1 of 32 study groups. Since the main purpose of this study was to examine the effects of specific factors on EMA completion rates, the decision was made to not include participants who experienced unusual technical difficulties in which their phones did not prompt 10 or more of the scheduled EMAs (eg, phone specific software,

XSL•FO
RenderX

including app blockers, on participant phones prevented app prompts). A total of 69 participants were excluded due to this issue, and an additional 5 were excluded for not completing any EMAs. See Figure 3 for the study CONSORT table.

Participants (N=411) were mostly female (311/411, 75.7%), White (298/411, 72.5%), and on average 48.4 (SD 12.1) years of age (see Table 2 for participant demographic characteristics). Participants were enrolled from nearly all US states, but most participants were from the East and West coasts (Figure 4). Nearly all participants (397/411, 96.6%) completed the end of study follow-up assessment.

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) table. EMA: ecological momentary assessment.

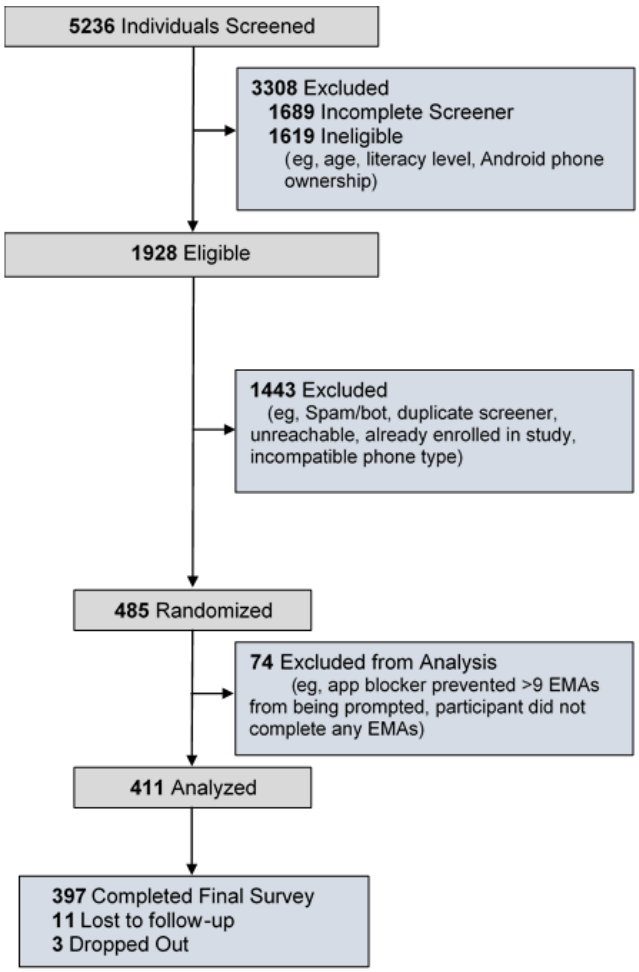

XSL•FO
RenderX

Table 2. Demographic characteristics.

		Payment		Number of EMAs[a]/day		Number of items/EMA		EMA schedule	
	Total sample	Percent-age-based (–1)	US $1/EMA (+1)	2/day (–1)	4/day (+1)	15 Items (–1)	25 Items (+1)	Random (–1)	Fixed (+1)
Sample size, n	411	206	205	205	206	206	205	206	205
Age (years), mean (SD)	48.4 (12.1)	49.2 (11.9)	47.5 (12.1)	48.2 (12.2)	48.5 (12.0)	48.6 (12.4)	48.2 (11.8)	47.4 (11.7)	49.4 (12.4)
Hispanic/Latino, %	9.6	8.0	11.3	10.8	8.3	7.7	11.4	9.7	9.5
Race, %									
American Indian	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
Asian	2.7	1.5	3.9	3.9	1.5	3.9	1.5	1.9	3.4
Black	18.0	21.8	14.1	15.5	20.5	16.6	19.4	19.4	16.6
White	72.5	69.4	75.6	73.8	71.2	72.7	72.3	70.4	74.6
Other/more than 1 race	5.4	5.8	4.9	5.3	5.4	5.4	5.3	6.8	3.9
Female, %	75.7	74.8	76.6	75.7	75.6	76.6	74.8	76.2	75.1
Years of education, mean (SD)	14.5 (2.2)	14.6 (2.2)	14.4 (2.1)	14.6 (2.2)	14.5 (2.1)	14.5 (2.1)	14.5 (2.2)	14.6 (2.1)	14.5 (2.2)
Full- or part-time employed, %	51.8	49.0	50.7	51.9	47.8	49.8	50.0	51.5	48.3
Household income <US $30,000/year, %	29.7	31.1	28.3	29.6	29.8	26.3	33.0	27.2	32.2
Rural residence, %	15.3	15.0	15.6	15.5	15.1	16.6	14.1	15.5	15.1
Probable depression via ODSIS[b], %	34.1	33.8	34.4	34.5	33.7	34.9	33.3	35.4	32.8
Probable anxiety via OASIS[c], %	36.6	35.3	42.1	39.9	37.3	38.5	38.8	41.0	36.3
History of mental illness, %	58.6	60.2	57.1	52.9	64.4	58.0	59.2	60.2	57.1
History of chronic illness, %	67.6	73.8	61.5	63.6	71.7	68.8	66.5	66.0	69.3
History of substance use problem, %	14.4	15.0	13.7	12.6	16.1	13.7	15.0	16.0	12.7
Health literacy, mean (SD)	4.8 (0.6)	4.7 (0.6)	4.7 (0.6)	4.8 (0.7)	4.7 (0.5)	4.8 (0.5)	4.7 (0.6)	4.8 (0.5)	4.7 (0.7)
SUS score, mean (SD)	82.7 (13.2)	79.6 (13.4)	85.8 (12.3)	83.2 (12.8)	82.3 (13.6)	82.4 (12.8)	83.1 (13.6)	84.1 (12.7)	81.3 (13.6)

[a]EMA: ecological momentary assessment.

[b]ODSIS: Overall Depression Severity and Impairment Scale.

[c]OASIS: Overall Anxiety Severity and Impairment Scale.

XSL•FO
RenderX

Figure 4. Participant zip codes by rural and urban residence.

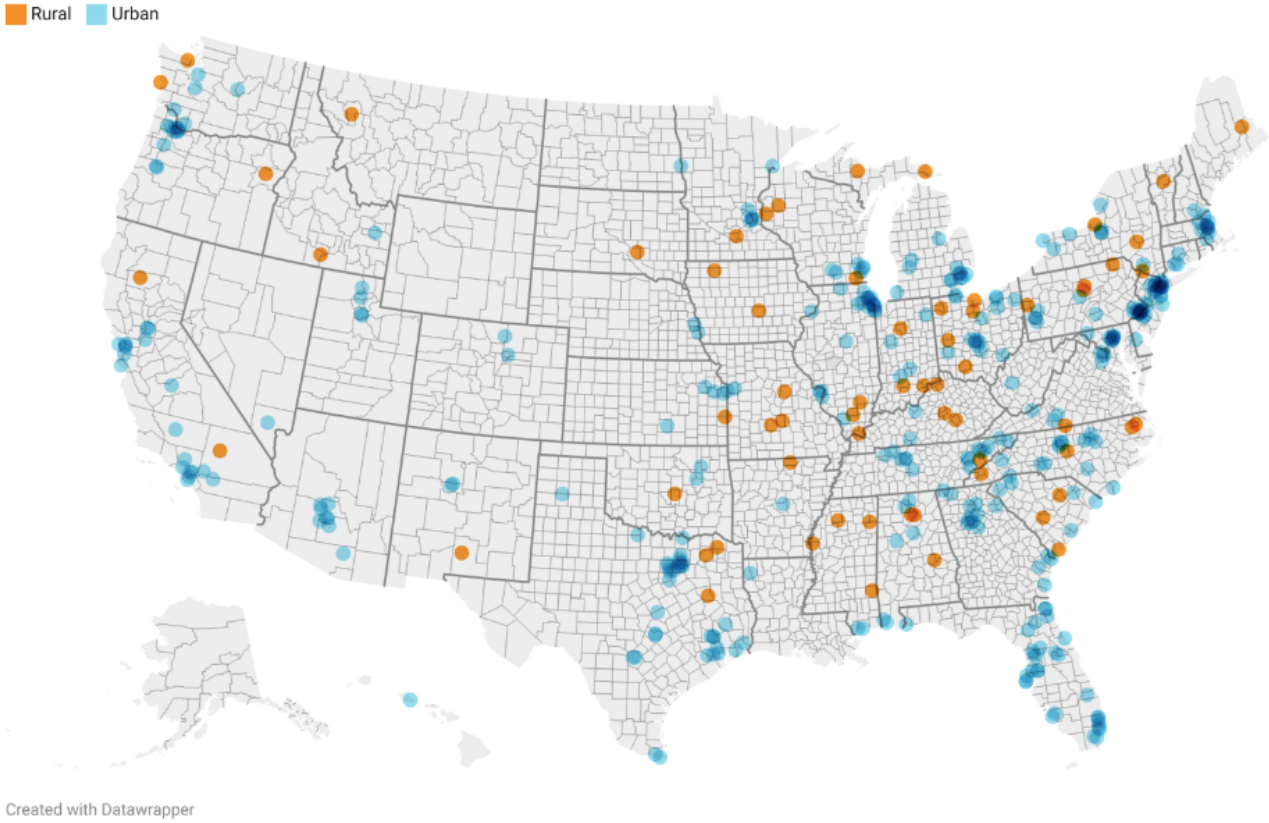

Effects of Experimental Factors on EMA Completion Rates

Over the 28-day study period, participants completed 83.8% of all scheduled EMAs (28,948 completed EMAs/34,552 total possible EMAs). See Table 3 for EMA completion rates by each study factor. Results indicated no significant main effects or interactions (2-way, 3-way, 4-way, or 5-way), as shown in the columns of Table 4 labeled as unadjusted models. On average, participants earned US $102. The 2 EMAs per day group earned an average of US $83, while the 4 EMAs per day group earned an average of US $130.

Table 3. Compliance rates by ecological momentary assessment (EMA) design factor.

Design factor and condition	EMA completion rate, mean (SD)
Payment	
Percentage-based (–1)	83.0 (18.8)
$1 Per EMA (+1)	83.9 (13.3)
Number of EMAs[a]/day	
2/day (–1)	82.4 (16.9)
4/day (+1)	84.5 (15.6)
Number of items/EMA	
15 Items (–1)	84.5 (15.7)
25 Items (+1)	82.4 (16.9)
EMA schedule	
Random (–1)	83.0 (16.0)
Fixed (+1)	83.9 (16.6)
Scale type	
Slider	83.8 (16.8)
Likert	84.3 (15.7)

[a]EMA: ecological momentary assessment.

XSL•FO
RenderX

Table 4. Linear multilevel regression models[a] for compliance rates by EMA[b] design factors (unadjusted model) and EMA design factors plus demographic characteristics (adjusted model).

	Unadjusted model[c]			Initial adjusted model[d]			Final adjusted model[e]		
	Estimate	SE	*P* value	Estimate	SE	*P* value	Estimate	SE	*P* value
Fixed									
Intercept	83.7	0.8	<.001	*87.5* [f]	2.3	<.001	*87.4*	0.8	<.001
Paid per EMA (+1)	0.1	0.8	.80	*–1.3*	0.7	.04	–1.3	0.6	.10
4 EMAs per day (+1)	0.9	0.8	.20	0.4	0.6	.60	0.2	0.6	.70
25 Items per EMA (+1)	–1	0.8	.20	–1.2	0.6	.10	–1.1	0.6	.10
Fixed schedule (+1)	0.5	0.8	.50	0.3	0.6	.70	0.2	0.6	.70
Slider-type response	0.4	0.3	.20	0.3	0.3	.40	0.3	0.3	.40
Time	–0.6	0.3	.10	–0.4	0.3	.20	–0.4	0.3	.20
Paid per EMA×4 EMAs per Day	0.4	0.8	.60	–0.1	0.6	.90	–0.1	0.6	.90
Paid per EMA×25 items per EMA	0.3	0.8	.70	0.3	0.6	.60	0.4	0.6	.50
Paid per EMA×fixed schedule	0.9	0.8	.20	0.4	0.6	.50	0.6	0.6	.40
4 EMAs per day×25 items per EMA	0.1	0.8	.90	0.2	0.6	.70	0.1	0.6	.90
4 EMAs per day×fixed	–0.1	0.8	.90	0.1	0.6	.90	–0.1	0.6	.90
25 Items per EMA×fixed schedule	–0.8	0.8	.30	0	0.6	>.99	–0.2	0.6	.80
Slider-type response×paid per EMA	–0.1	0.3	.80	–0.1	0.3	.80	–0.1	0.3	.80
Slider-type response×4 EMAs per day	–0.2	0.3	.60	–0.1	0.3	.70	–0.1	0.3	.70
Slider-type response×25 Items per EMA	–0.4	0.3	.30	–0.5	0.3	.10	–0.5	0.3	.10
Slider-type response×fixed schedule	0.1	0.3	.70	0.2	0.3	.60	0.2	0.3	.60
Age	—	—	—	0.1	0.1	.10	*0.1*	0.1	.02
Hispanic or Latino	—	—	—	–2.5	2.2	.30	—	—	—
White	—	—	—	1	1.5	.50	—	—	—
Female	—	—	—	1.2	1.6	.50	—	—	—
Years of education	—	—	—	0.6	0.3	.10	—	—	—
Full- or part-time employed	—	—	—	–2.6	1.5	.10	—	—	—
Household income <US $30,000/year	—	—	—	–1.6	1.7	.30	—	—	—
Rural residence	—	—	—	0.7	1.8	.70	—	—	—
Probable depression via ODSIS[g]	—	—	—	*–4.4*	1.8	.01	*–3.9*	1.3	.004
Probable anxiety via OASIS[h]	—	—	—	0.1	1.8	>.99	—	—	—
History of mental illness	—	—	—	1.4	1.6	.40	—	—	—
History of chronic illness	—	—	—	–0.8	1.6	.60	—	—	—
History of substance use problem	—	—	—	–3.2	1.9	.10	*–3.9*	1.8	.03
Health literacy	—	—	—	1.0	1.2	.40	—	—	—
SUS[i] score	—	—	—	*0.2*	0.1	<.001	*0.2*	0.1	<.001

[a]Models included all main effects and interactions (ie, 2-way, 3-way, 4-way, and 5-way).

[b]EMA: ecological momentary assessment.

[c]Random intercept SD estimate 13.5 (95% CI 12.2-14.8); random residual SD estimate 9.9 (95% CI 9.2-10.6).

[d]Random intercept SD estimate 10.5 (95% CI 9.5-11.7); random residual SD estimate 8.7 (95% CI 8.1-9.4).

XSL•FO
RenderX

[e]Random intercept SD estimate 10.6 (95% CI 9.6-11.7); random residual SD estimate 8.7 (95% CI 8.1-9.4).

[f]All significant results are italicized.

[g]ODSIS: Overall Depression Severity and Impairment Scale.

[h]OASIS: Overall Anxiety Severity and Impairment Scale.

[i]SUS: system usability scale.

Relations Among Demographics, Environmental Variables, and EMA Completion Rates

Overview

An initial adjusted linear multilevel regression model was fit to examine which demographic factors were associated with EMA compliance, after accounting for EMA design factors (see Table 4, Initial adjusted model). After backwards elimination was used to identify which demographic factors had the largest association with EMA compliance (Table 4, Final adjusted model), results indicated that age was positively associated with EMA compliance, such that for every year increase in age, EMA compliance was expected to increase by 0.1% (*P*=.02). Both current depression via the Overall Depression Severity and Impairment scale (β=–3.9, *P*=.004) and substance use problem history (β=–3.9, *P*=.03) were associated with lower EMA compliance. Finally, the app usability score was positively associated with EMA compliance (β=.2, *P*<.001).

EMA Compliance by Day of Week

On average, participants completed 84.5% (20,807/24,625) of EMAs that were prompted on weekdays and 82.6% (8137/9852) of EMAs prompted on weekend days. Results showed that the log odds of completing EMAs prompted on weekdays were significantly higher than the log odds of completing an EMA prompted on a weekend (β=.16, 95% CI 0.09-0.23, *P*<.001).

EMA Compliance by Time of Day

Of the prompts that occurred in the morning, 87.3% (10,183/11,669) were completed. Of the prompts that occurred in the afternoon, 88.8% (6415/7227) were completed. Of the prompts that occurred in the evening, 89.9% (6781/7544) were completed. Of the prompts that occurred in the late night, 88.5% (5567/6292) were completed. Results from fitting a multilevel logistic model indicated that the log odds of completing an EMA prompted in the morning (the reference group) were 2.2 (*P*<.001; 95% CI 2.1-2.3). Further, the results showed that the log odds of completing EMAs prompted in the afternoon (β=.2, 95% CI 0.1-0.3, *P*<.001), evening (β=.3, 95% CI 0.2-0.4, *P*<.001), and late night (β=.2, 95% CI 0.1-0.3, *P*<.001) were significantly higher than the log odds of completing an EMA prompted in the morning. Refitting the model with afternoon as the reference group indicated that the log odds of completing EMAs prompted in the evening were also significantly higher than the log odds of completing EMAs prompted in the afternoon (β=.114, 95% CI 0.002-0.227, *P*=.046). In contrast, EMAs prompted in the late night were not statistically different from those prompted in the afternoon (β=–.04, 95% CI –0.16 to 0.08, *P*=.55). Refitting the model once more with evening as the reference group showed that the log odds of completing EMAs prompted in the late night were significantly lower than those prompted in the evening (β=–.15, 95% CI –0.27 to 0.03, *P*=.01).

Participant Perceptions of EMAs

Overall, the Insight app was well liked, as measured by the SUS (mean 82.7, SD 13.2), and the SUS score was positively related to the overall EMA completion rate (*r*=0.14, *P*=.005). A minority of participants (71/397, 17.9%) reported that the number of EMAs that they were prompted to complete was too high. Those assigned 4 EMAs per day (*P*=.001), paid by percentage of EMAs completed (*P*<.001), and on a fixed EMA schedule (*P*=.03) were more likely to indicate that the number of EMAs was too high. A minority of participants (86/397, 21.7%) indicated that the app was moderately to extremely annoying. Those assigned to receive payment by percentage of EMAs completed were more likely to find the smartphone application annoying (*P*=.002) compared to those who were paid US $1 per EMA completed. Most participants (336/397, 84.6%) reported that answering EMA questions made them more aware of their thoughts, feelings, and behavior. Those assigned to receive payment per EMA completed were more likely to indicate that answering EMA questions made them more aware of their thoughts, feelings, and behavior (*P*=.03) compared with those who were paid US $1 per EMA completed.

App Engagement

On average, participants clicked the payment button 66.7 (SD 61.8) times and the app instructions button 9.2 (SD 11.3) times. The average number of Payment button clicks was much higher for those assigned to be paid based upon the percentage of EMAs completed (mean 77.2) than those assigned to receive US $1 per EMA (mean 56.1; *t*=–3.509, *P*<.001).

Discussion

Principal Findings

This study examined the impact of EMA design features on participant compliance in a prospective, experimental design. The results of this study demonstrated that EMA frequency (2 vs 4 EMAs per day), number of EMA items (15 vs 25 questions), EMA prompt type (fixed vs random), payment type (US $1 per EMA vs payment by percentage of EMAs completed), and EMA response type (slider-type responses vs Likert-type responses) were not associated with EMA completion rates. Prior research on the impact of EMA design factors on compliance has been mixed. For example, Jones et al [8] similarly found that EMA compliance was not associated with prompt frequency, while 2 other meta-analyses found that fewer prompts were associated with increased compliance rates [14,15]. Further, while Vachon et al [14] found that fixed EMA prompts were associated with higher rates of compliance, our results suggest that there was no significant difference in compliance between fixed versus random prompt types. These differences may indicate that there are other confounding factors and individual variability that impact EMA compliance beyond design factors such as frequency or length of assessments, such

XSL•FO
RenderX

as variations in participants' daily schedules, or changes in motivation, attention span, or engagement in the study.

We also examined the impact of several demographic and environmental characteristics on the EMA completion rate, and used backwards elimination to identify the factors that had the largest association with EMA completion. Age, history of a substance use problem, probable current depressive disorder, and SUS score were the factors that were most associated with EMA completion, accounting for experimental factors. While prior research has suggested that older adults may experience barriers related to smartphone use such as lack of interest and technological knowledge [32,33], our results suggested that older age was associated with higher levels of EMA completion. Our results are similar to those of Ono et al [34] and Cain et al [35], which found that EMA protocols are generally feasible and acceptable to older adults. Further, we found that adults without a history of a substance use problem completed 5.9% more EMAs on average than those with a history of substance use problems (ie, 84.3% vs 78.4%). This is consistent with a 2019 meta-analysis of compliance with EMA protocols among substance users, which found that substance-dependent samples had lower compliance rates compared to nondependent samples [8]. In addition, individuals with probable current depression had lower EMA completion rates than those without current depression. Although this effect was statistically significant, current probable depression was related to a 2.1% average difference (84.3% vs 82.2%) in EMA completion over the 28-day study period and thus should not be overstated. There were no differences in EMA completion by race, ethnicity, current anxiety, rural versus urban residency, income, years of education, history of chronic illness, history of mental illness, or health literacy. These findings suggest that EMA protocols may be feasible among a diverse group of individuals and that limitations that may impact participation in traditional health behavior interventions, such as living in a rural area or lower health literacy, may not be barriers to smartphone-based studies. However, for studies that use EMA to inform the delivery of health behavior interventions (eg, JITAIs), it may be beneficial to plan EMA prompts based on an individual's actual availability (eg, outside of work hours) to maximize engagement with EMA protocols.

There were no main effects of EMA frequency, number of EMA items, EMA prompt type, payment type, or EMA response scale type on EMA completion rates. These study results and several others have implications for future EMA study designs. First, while payment type did not impact EMA completion rates, participants assigned to receive compensation based upon the percentage of EMAs completed tended to engage with the app "Payment" button more frequently than those assigned to receive $1 per EMA. This may indicate that participants assigned to the percentage condition frequently clicked the payment button to determine if they had missed any EMAs. For studies in which engagement with an app is important (eg, to view intervention content), a payment based on the percentage of EMAs completed may increase time spent in the app or button clicks. In addition, we found that participant perceptions of the Insight app were extremely positive via the SUS. In fact, SUS scores were in the top 10th percentile [21] and SUS scores were positively associated with EMA completion. Therefore, an individual's confidence in using a specific app, its ease of use, and its perceived usability may be important factors for EMA completion rates. Finally, our results demonstrated that EMA compliance was higher on weekdays compared to weekends and that there were differences in compliance by time of day, with the highest level of compliance for evening assessments compared with morning, afternoon, and late-night assessments. This finding suggests that participants may be more engaged during the week, and that evening assessments may be more amenable to participants. Upon study setup, participants were able to specify sleep and wake times for each day of the week so that they would only receive EMAs during waking hours. This type of customization may be particularly important for shift workers, and individuals who work night shifts.

Study Strengths and Limitations

The strengths of this study include its prospective, randomized design and large sample size drawn from the entire United States (ie, participants were recruited from all but 4 states). However, there are several limitations. First, our sample was primarily non-Hispanic White and female, so our findings may not be generalizable to other populations. It should be noted, however, that we did not find any significant difference in EMA completion rates by sex or race. Second, the Insight app, at the time of data collection, was only compatible with phones using the Android operating system. However, it is important to keep in mind that, as of 2022, Android accounted for 71.8% of the worldwide mobile operating system market share [36]. An additional limitation is that individuals who sign up for remote smartphone-based studies and individuals recruited via Facebook may be different from the general population. Specifically, individuals who are more comfortable with technology, have a Facebook account, and do not perceive that smartphone-based surveys are burdensome may be more likely to sign up for an EMA study than individuals who are not comfortable with smartphones. Findings indicated that there was a very high proportion of participants with current anxiety or depression. This finding is consistent with the significant increase in anxiety and depression that was reported during the COVID-19 pandemic [37,38]. Finally, it was unexpected that all of the factors that were examined were not related to EMA compliance. These findings may indicate that the differences between each level of these design factors (eg, 15 vs 25 questions) were not large enough to result in significant differential completion rates of prompted EMAs based on the study sample size. Bigger differences in factors like the number of EMAs within a day [14,15] or the number of questions per EMA [15] could result in greater differentiation in EMA compliance. However, the factors, levels of each factor, and duration of the EMA period (ie, 28 days) were specifically selected and tuned to assess compliance with and acceptability of EMA schedules that could be embedded into future smartphone-based JITAI that use EMA data to tailor intervention content.

Future Work

Future analyses from these data will evaluate whether slider-type and Likert-type response scales can be interchangeably used to predict daily health behaviors, whether there are differences in

XSL•FO
RenderX

the predictive value of data collected during fixed and random EMAs, and if there is additional value in asking participants to complete 4 versus 2 EMAs per day.

Conclusions

Participants were highly compliant with a 28-day EMA design that varied across 5 EMA design dimensions. Studies that incorporate EMA data collection methods must balance the benefits of obtaining granular, real-time data about participants' thoughts, feelings, and behaviors with the participant burden of an intensive survey protocol. The results of this study indicated that all of the factors that were examined (ie, number of EMA items, EMA frequency, EMA payment type, EMA prompt type, and EMA response scale type) did not significantly impact EMA completion rates, nor did multiple demographic characteristics (eg, race or ethnicity or urban vs rural), history of mental illness, or history of other chronic illnesses. However, younger adults, those with current depression, and those with a history of a substance use problem completed fewer EMAs than their counterparts. EMA studies may benefit from customizing protocols to the behaviors, phenomena, or populations being studied, without arbitrary limitations on design factors. For example, the frequency and timing of EMA prompts should match the anticipated frequency and diversity of the behavior being assessed.

Acknowledgments

We would like to acknowledge all individuals who participated in this randomized controlled trial. Programming and technological support were provided through the mobile health shared resource of the Stephenson Cancer Center via an NCI Cancer Center Support Grant (P30CA225520). Data analysis and manuscript preparation were additionally supported through a grant (R00DA046564) and through the Oklahoma Tobacco Settlement Endowment Trust grant (R22-02).

Data Availability

Data from the Exemplar study are not publicly available. Deidentified data are available upon reasonable requests made to the lead author of this paper.

Authors' Contributions

All authors contributed to the conceptualization of the study and assisted in the drafting and editing of the manuscript. MSB was involved in conceptualization, methodology, project administration, writing—original draft, and writing—review and editing. ETH was involved in conceptualization, data curation, formal analyses, writing—original draft, and writing—review and editing. LB was involved in data curation, formal analyses, and writing—review and editing. KMK was involved in project administration and writing—review and editing. ST was involved in writing—review and editing. MEP was involved in conceptualization and writing—review and editing. TQ was involved in conceptualization and writing—review and editing.

Conflicts of Interest

MSB is the primary inventor of the Insight mHealth Platform, which was used to develop the Exemplar app. MSB receives royalties related to the Insight Platform, but he did not receive royalties in this case because he was the study PI.

Multimedia Appendix 1
Ecological momentary assessment conditions.
[DOCX File , 17 KB - jmir_v26i1e50275_app1.docx]

References

1. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
2. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavorial medicine. Ann Behav Med 1994;16(3):199-202.
3. Shiffman S, Dunbar MS, Li X, Scholl SM, Tindle HA, Anderson SJ, et al. Smoking patterns and stimulus control in intermittent and daily smokers. PLoS One 2014;9(3):e89911 [FREE Full text] [doi: 10.1371/journal.pone.0089911] [Medline: 24599056]
4. Gwaltney CJ, Shiffman S, Balabanis MH, Paty JA. Dynamic self-efficacy and outcome expectancies: prediction of smoking lapse and relapse. J Abnorm Psychol 2005;114(4):661-175. [doi: 10.1037/0021-843X.114.4.661] [Medline: 16351387]
5. Forman EM, Schumacher LM, Crosby R, Manasse SM, Goldstein SP, Butryn ML, et al. Ecological momentary assessment of dietary lapses across behavioral weight loss treatment: characteristics, predictors, and relationships with weight change. Ann Behav Med 2017;51(5):741-753 [FREE Full text] [doi: 10.1007/s12160-017-9897-x] [Medline: 28281136]
6. McKee HC, Ntoumanis N, Taylor IM. An ecological momentary assessment of lapse occurrences in dieters. Ann Behav Med 2014;48(3):300-310. [doi: 10.1007/s12160-014-9594-y] [Medline: 24562984]

XSL•FO
RenderX

7. Trull TJ, Ebner-Priemer UW. Ambulatory assessment in psychopathology research: a review of recommended reporting guidelines and current practices. J Abnorm Psychol 2020;129(1):56-63. [doi: 10.1037/abn0000473] [Medline: 31868388]
8. Jones A, Remmerswaal D, Verveer I, Robinson E, Franken IHA, Wen CKF, et al. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. Addiction 2019;114(4):609-619 [FREE Full text] [doi: 10.1111/add.14503] [Medline: 30461120]
9. Stone AA, Broderick JE, Schwartz JE, Shiffman S, Litcher-Kelly L, Calvanese P. Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. Pain 2003;104(1-2):343-351. [doi: 10.1016/s0304-3959(03)00040-x] [Medline: 12855344]
10. Eisele G, Vachon H, Lafit G, Kuppens P, Houben M, Myin-Germeys I, et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment 2022;29(2):136-151. [doi: 10.1177/1073191120957102] [Medline: 32909448]
11. Hasselhorn K, Ottenstein C, Lischetzke T. The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. Behav Res Methods 2022;54(4):1541-1558 [FREE Full text] [doi: 10.3758/s13428-021-01683-6] [Medline: 34505997]
12. Rintala A, Wampers M, Myin-Germeys I, Viechtbauer W. Response compliance and predictors thereof in studies using the experience sampling method. Psychol Assess 2019;31(2):226-235. [doi: 10.1037/pas0000662] [Medline: 30394762]
13. Ottenstein C, Werner L. Compliance in ambulatory assessment studies: investigating study and sample characteristics as predictors. Assessment 2022;29(8):1765-1776 [FREE Full text] [doi: 10.1177/10731911211032718] [Medline: 34282659]
14. Vachon H, Viechtbauer W, Rintala A, Myin-Germeys I. Compliance and retention with the experience sampling method over the continuum of severe mental disorders: meta-analysis and recommendations. J Med Internet Res 2019;21(12):e14475 [FREE Full text] [doi: 10.2196/14475] [Medline: 31808748]
15. Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K. Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. J Med Internet Res 2021;23(3):e17023 [FREE Full text] [doi: 10.2196/17023] [Medline: 33656451]
16. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
17. Campbell-Sills L, Norman SB, Craske MG, Sullivan G, Lang AJ, Chavira DA, et al. Validation of a brief measure of anxiety-related severity and impairment: the overall anxiety severity and impairment scale (OASIS). J Affect Disord 2009;112(1-3):92-101 [FREE Full text] [doi: 10.1016/j.jad.2008.03.014] [Medline: 18486238]
18. Bentley KH, Gallagher MW, Carl JR, Barlow DH. Development and validation of the overall depression severity and impairment scale. Psychol Assess 2014;26(3):815-830. [doi: 10.1037/a0036216] [Medline: 24708078]
19. Wallace LS, Rogers ES, Roskos SE, Holiday DB, Weiss BD. Brief report: screening items to identify patients with limited health literacy skills. J Gen Intern Med 2006;21(8):874-877 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00532.x] [Medline: 16881950]
20. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
21. Sauro J. Measuring Usability with the System Usability Scale (SUS). 2011. URL: https://measuringu.com/sus/ [accessed 2024-06-26]
22. Shiffman S, Hickcox M, Paty JA, Gnys M, Richards T, Kassel JD. Individual differences in the context of smoking lapse episodes. Addict Behav 1997;22(6):797-811. [doi: 10.1016/s0306-4603(97)00063-4] [Medline: 9426798]
23. Wetter DW, McClure JB, Cofta-Woerpel L, Costello TJ, Reitzel LR, Businelle MS, et al. A randomized clinical trial of a palmtop computer-delivered treatment for smoking relapse prevention among women. Psychol Addict Behav 2011;25(2):365-371 [FREE Full text] [doi: 10.1037/a0022797] [Medline: 21500879]
24. Stone AA, Schwartz JE, Neale JM, Shiffman S, Marco CA, Hickcox M, et al. A comparison of coping assessed by ecological momentary assessment and retrospective recall. J Pers Soc Psychol 1998;74(6):1670-1680. [doi: 10.1037//0022-3514.74.6.1670] [Medline: 9654765]
25. Shiffman S, Paty JA, Gnys M, Kassel JA, Hickcox M. First lapses to smoking: within-subjects analysis of real-time reports. J Consult Clin Psychol 1996;64(2):366-379. [doi: 10.1037//0022-006x.64.2.366] [Medline: 8871421]
26. Piasecki TM, Cooper ML, Wood PK, Sher KJ, Shiffman S, Heath AC. Dispositional drinking motives: associations with appraised alcohol effects and alcohol consumption in an ecological momentary assessment investigation. Psychol Assess 2014;26(2):363-369 [FREE Full text] [doi: 10.1037/a0035153] [Medline: 24274049]
27. Simons JS, Dvorak RD, Batien BD, Wray TB. Event-level associations between affect, alcohol intoxication, and acute dependence symptoms: effects of urgency, self-control, and drinking experience. Addict Behav 2010;35(12):1045-1053 [FREE Full text] [doi: 10.1016/j.addbeh.2010.07.001] [Medline: 20685044]
28. Swendsen JD, Tennen H, Carney MA, Affleck G, Willard A, Hromi A. Mood and alcohol consumption: an experience sampling test of the self-medication hypothesis. J Abnorm Psychol 2000;109(2):198-204. [Medline: 10895557]

XSL•FO
RenderX

29. Kendzor DE, Shuval K, Gabriel KP, Businelle MS, Ma P, High RR, et al. Impact of a mobile phone intervention to reduce sedentary behavior in a community sample of adults: a quasi-experimental evaluation. J Med Internet Res 2016;18(1):e19 [FREE Full text] [doi: 10.2196/jmir.5137] [Medline: 26810027]
30. Businelle MS, Ma P, Kendzor DE, Frank SG, Vidrine DJ, Wetter DW. An ecological momentary intervention for smoking cessation: evaluation of feasibility and effectiveness. J Med Internet Res 2016;18(12):e321 [FREE Full text] [doi: 10.2196/jmir.6058] [Medline: 27956375]
31. Businelle MS, Ma P, Kendzor DE, Frank SG, Wetter DW, Vidrine DJ. Using intensive longitudinal data collected via mobile phone to detect imminent lapse in smokers undergoing a scheduled quit attempt. J Med Internet Res 2016;18(10):e275 [FREE Full text] [doi: 10.2196/jmir.6307] [Medline: 27751985]
32. Mohadis HM, Ali NM. A study of smartphone usage and barriers among the elderly. 2014 Presented at: 3rd International Conference on User Science and Engineering (i-USEr); 2014 November 14; Bangi, Malaysia.
33. Paschoarelli L, Fernandes N, Ferro-Marques L. Understanding the barriers and challenges between older users and smartphones: a systematic literature review. In: Perspectives on Design and Digital Communication II: Research, Innovations and Best Practices. Cham, Switzerland: Springer, Cham Press; 2021:53-64.
34. Ono M, Schneider S, Junghaenel DU, Stone AA. What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. J Med Internet Res 2019;21(2):e11398 [FREE Full text] [doi: 10.2196/11398] [Medline: 30720437]
35. Cain AE, Depp CA, Jeste DV. Ecological momentary assessment in aging research: a critical review. J Psychiatr Res 2009;43(11):987-996 [FREE Full text] [doi: 10.1016/j.jpsychires.2009.01.014] [Medline: 19272611]
36. Taylor P. Mobile operating systems' market share worldwide from 1st quarter 2009 to 4th quarter 2022 2023. URL: https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/ [accessed 2024-06-26]
37. Delpino FM, da Silva CN, Jerônimo JS, Mulling ES, da Cunha LL, Weymar MK, et al. Prevalence of anxiety during the COVID-19 pandemic: a systematic review and meta-analysis of over 2 million people. J Affect Disord 2022;318:272-282 [FREE Full text] [doi: 10.1016/j.jad.2022.09.003] [Medline: 36096370]
38. Khubchandani J, Sharma S, Webb FJ, Wiblishauser MJ, Bowman SL. Post-lockdown depression and anxiety in the USA during the COVID-19 pandemic. J Public Health (Oxf) 2021;43(2):246-253 [FREE Full text] [doi: 10.1093/pubmed/fdaa250] [Medline: 33426559]

Abbreviations

EMA: ecological momentary assessment
JITAI: just-in-time adaptive intervention
REDCap: Research Electronic Data Capture
SUS: system usability scale

Edited by S Ma; submitted 11.07.23; peer-reviewed by M Kondo, G Eisele; comments to author 07.12.23; revised version received 08.02.24; accepted 12.06.24; published 12.08.24.

Please cite as:
Businelle MS, Hébert ET, Shi D, Benson L, Kezbers KM, Tonkin S, Piper ME, Qian T
Investigating Best Practices for Ecological Momentary Assessment: Nationwide Factorial Experiment
J Med Internet Res 2024;26:e50275
URL: https://www.jmir.org/2024/1/e50275
doi:10.2196/50275
PMID:

©Michael S Businelle, Emily T Hébert, Dingjing Shi, Lizbeth Benson, Krista M Kezbers, Sarah Tonkin, Megan E Piper, Tianchen Qian. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multicomponent Support Program for Secondary Prevention of Stroke Using Digital Health Technology: Co-Design Study With People Living With Stroke or Transient Ischemic Attack

David Silvera-Tawil[1*], BEng, MEng, PhD; Jan Cameron[2,3*], BN, MHSc, PhD; Jane Li[1], MBBS, PhD; Marlien Varnfield[1], BSc, MSc, PhD; Liam P Allan[1,2], BSc; Mitch Harris[4], BA; Natasha A Lannin[3,5], BSc, PhD; Christian Redd[1], BEng, MEng, PhD; Dominique A Cadilhac[2,6], BN, MPH, PhD

[1]Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Sydney, Australia

[2]Department of Medicine, Monash University, Melbourne, Australia

[3]Australian Centre for Heart Health, Melbourne, Australia

[4]Data 61, Commonwealth Scientific and Industrial Research Organisation, Melbourne, Australia

[5]Department of Neuroscience, Monash University, Melbourne, Australia

[6]Florey Institute of Neuroscience and Mental Health, Melbourne, Australia

[*]these authors contributed equally

Corresponding Author:
David Silvera-Tawil, BEng, MEng, PhD
Australian e-Health Research Centre
Commonwealth Scientific and Industrial Research Organisation
158-164 Hawkesbury Road
Westmead
Sydney, 2145
Australia
Phone: 61 2 9372 4282
Email: david.silvera@csiro.au

Abstract

Background: Few individuals (<2%) who experience a stroke or transient ischemic attack (TIA) participate in secondary prevention lifestyle programs. Novel approaches that leverage digital health technology may provide a viable alternative to traditional interventions that support secondary prevention in people living with stroke or TIA. To be successful, these strategies should focus on user needs and preferences and be acceptable to clinicians and people living with stroke or TIA.

Objective: This study aims to co-design, with people with lived experience of stroke or TIA (referred to as consumers) and clinicians, a multicomponent digital technology support program for secondary prevention of stroke.

Methods: A consumer user needs survey (108 items) was distributed through the Australian Stroke Clinical Registry and the Stroke Association of Victoria. An invitation to a user needs survey (135 items) for clinicians was circulated via web-based professional forums and national organizations (eg, the Stroke Telehealth Community of Practice Microsoft Teams Channel) and the authors' research networks using Twitter (subsequently rebranded X, X Corp) and LinkedIn (LinkedIn Corp). Following the surveys, 2 rounds of user experience workshops (design and usability testing workshops) were completed with representatives from each end user group (consumers and clinicians). Feedback gathered after each round informed the final design of the digital health program.

Results: Overall, 112 consumers (male individuals: n=63, 56.3%) and 54 clinicians (female individuals: n=43, 80%) responded to the survey; all items were completed by 75.8% (n=85) of consumers and 78% (n=42) of clinicians. Most clinicians (46/49, 94%) indicated the importance of monitoring health and lifestyle measures more frequently than current practice, particularly physical activity, weight, and sleep. Most consumers (87/96, 90%) and clinicians (41/49, 84%) agreed that providing alerts about potential deterioration in an individual's condition were important functions for a digital program. Intention to use a digital program for stroke prevention and discussing the data collected during face-to-face consultations was high (consumers: 79/99, 80%; clinicians 36/42, 86%). In addition, 7 consumers (male individuals: n=5, 71%) and 9 clinicians (female individuals: n=6, 67%) took part in the user experience workshops. Participants endorsed using a digital health program to help consumers manage

XSL•FO
RenderX

stroke or TIA and discussed preferred functions and health measures in a digital solution for secondary prevention of stroke. They also noted the need for a mobile app that is easy to use. Clinician feedback highlighted the need for a customizable clinician portal that captures individual consumer goals.

Conclusions: Following an iterative co-design process, supported by evidence from user needs surveys and user experience workshops, a consumer-facing app that integrates wearable activity trackers and a clinician web portal were designed and developed to support secondary prevention of stroke. Feasibility testing is currently in progress to assess acceptability and use.

(J Med Internet Res 2024;26:e54604) doi:10.2196/54604

KEYWORDS

mobile app; stroke; transient ischemic attack; health service delivery; mobile health; mHealth; mobile phone

Introduction

Background

Stroke is the third leading global cause of death and disability, with approximately 12.2 million incident strokes reported in 2019 [1,2]. Recurrent stroke events are frequently reported as 4% within 90 days of the index event [3] and 26% within 5 years [4]. The risk of stroke following a transient ischemic attack (TIA) is also substantial, with a 20% risk 90 days following the index event [5]. Therefore, for people who have experienced a stroke or TIA, prevention of another vascular event (ie, secondary prevention) is a priority [6,7]. Clinical guidelines recommend control of modifiable risk factors via lifestyle behavior changes in conjunction with pharmacotherapeutic measures to reduce the risk of recurrence [8,9]. These include reducing blood pressure, blood cholesterol, blood glucose, and body weight; improving diet and physical activity; and reducing or ceasing alcohol and cigarette consumption [8-10].

Interventions promoting uptake of secondary prevention strategies have been used effectively for controlling blood pressure [11] and improving diet [12], physical activity [13], and medication adherence [14]. Implementation of secondary prevention strategies largely rests in the hands of individuals who must take an active role in self-managing their health. However, health care and patient-level barriers impede uptake and adherence to these strategies. Significant improvements in the acute management of stroke have resulted in shorter hospital admissions of approximately 4 days [15], with little education and time provided for survivors to consider the long-term physical, emotional, and lifestyle behaviors needed for recovery and prevention of future events [16].

Compounding an individual's motivation to address their vascular risk is the poor mental health experienced by many survivors of stroke [17], which often hinders their engagement in secondary prevention programs and motivation for undertaking lifestyle adaptations [18]. For example, 1 in 3 patients stop taking their antihypertension medication within the first year of stroke [19]. Consequently, a fundamental practice gap prevails with the adoption of recommended stroke prevention strategies into real-world settings [20]. Novel secondary prevention programs that are feasible to implement and are broad reaching are needed and could be realized using digital health.

The rising use of consumer devices makes mobile health (mHealth) strategies a potential solution to increase participation in secondary prevention of stroke by providing low-cost, readily accessible, and scalable interventions to promote engagement in health behavior changes [21,22]. mHealth interventions incorporating smartphone apps, wearable activity trackers, clinician portals, and electronic messaging have been used to improve the management of relevant chronic conditions associated with stroke. Findings from a systematic review indicated the use of smartphone apps, phone calls, and combined programs (>2 technologies such as SMS text messages combined with a smartphone app) improved management of hypertension, including blood pressure reduction and adherence to medication [23]. Similarly, smartphone apps have been linked to improved management of diabetes mellitus [24] as well as to enhanced management of chronic heart disease when used alongside SMS text messages, phone calls, and pedometers [25,26]. Their impact includes improving clinical cardiovascular outcomes (eg, blood pressure control and medication adherence) [23-26] as well as psychosocial outcomes [26]. These interventions have also been successfully implemented in the real world. Cardihab (Cardihab Pty Ltd), for example, combines a smartphone app with a clinician-facing web portal to deliver cardiac rehabilitation to patients after a myocardial infarct [27]. When evaluated in a randomized controlled trial, Cardihab improved functional walking capacity, psychosocial health, weight, and quality of life [26].

Despite the potential benefits of mHealth for risk-factor reduction and chronic disease management, evidence-based approaches for secondary prevention of stroke are still nascent, and further research is required to inform the development of platforms that are acceptable and appropriate for people living with stroke or TIA [28]. Furthermore, development of these programs must include considerations of accessibility for those with less access to technology, lower digital literacy, and ongoing impairments after stroke [28].

Objectives

Using a co-design approach, we aimed to develop a novel digital secondary prevention program with people living with stroke or TIA (consumers) and clinicians. The aim of this program is to offer an easily accessible solution that supports secondary stroke prevention, empowers self-management, and enhances health outcomes. The CAPS (Care Assistant and Support Program for People After Stroke or TIA) program draws upon past successes of the cloud-based Mobile Technology–Enabled Rehabilitation platform [26,29] and an electronic interface for scheduling electronic health messages aligned to individual

XSL•FO
RenderX

goals for recovery and secondary prevention of stroke (inspiring Virtual Enabled Resources following Vascular Events) [30].

We report here on the user needs analysis and user experience workshops. Finally, we present the resulting components of the mHealth platform (ie, mobile app and clinician portal) and care management program that uses lifestyle and health data collected through a wearable device to facilitate secondary prevention for individuals after stroke or TIA.

Methods

The design of the CAPS program and digital platform were completed using a mixed methods approach, undertaken in 2 main stages: (1) a user needs analysis to explore the main requirements, priorities, and potential use of digital technology to support secondary prevention of stroke or TIA and (2) user experience workshops to inform the design of the main components of the digital health program.

Stage 1: User Needs Analysis

The requirements and challenges associated with the creation of CAPS were investigated via researcher-developed surveys completed between December 2020 and April 2021. Particular attention was given to the potential use of 2 broad digital technologies: mobile apps and wearable or health monitoring devices.

Participants

User needs surveys were completed by people with lived experience of stroke or TIA (referred here on as consumers) and clinicians who provide health care and education services to individuals after stroke or TIA. These populations have lived experience of stroke (and care for patients after stroke) and are well placed to comment on the needs, preferences, and perceived potential of secondary prevention programs.

Consumer participants were invited to participate via invitation letters posted to 840 registrants of the Australian Stroke Clinical Registry (AuSCR), a national clinical quality registry of patients admitted to participating hospitals with a diagnosis of acute stroke or TIA [31]. The AuSCR consists of >35,000 Australian survivors of stroke who are known to be alive and willing to be contacted to participate in research projects. Once the proposal was reviewed, the AuSCR staff distributed invitations to registrants based on the study eligibility criteria. Invitation letters were also distributed through the Stroke Association (Victoria, Australia), which operates 5 support centers across regional Victoria, offering dedicated community-based support services for people recovering from stroke. These methods were used to improve response rates by directly inviting potentially eligible participants from a cohort that has previously been hard-to-reach [32,33].

Eligibility criteria included a diagnosis of stroke or TIA in the past 6 months to 5 years; aged ≥18 years; proficiency in spoken and written English; and, for AuSCR participants, those who agreed to be contacted by the AuSCR for future research. Individuals residing in a nursing home or those with no access to the internet were ineligible to participate.

Invitations to clinicians were circulated via the authors' research networks, professional forums, and national organizations. We sought responses from an unbiased and representative sample, recruiting through Twitter (subsequently rebranded X, X Corp) and LinkedIn (LinkedIn Corp) on personal research accounts (>1200 followers, many of whom shared the original tweet) and the Stroke Society of Australia (>7000 members), who also shared the invitation via Twitter and email. We also used the Australian Stroke Telehealth Community of Practice Microsoft Teams Channel, with >500 clinicians and researchers that use the group as a forum to exchange ideas and knowledge. Participation in the surveys was anonymous.

Eligibility included all clinicians who provide health care and education services to individuals after stroke or TIA. Participant eligibility was confirmed via the demographics questionnaire in each survey, including self-reported clinical history (for consumers) or self-reported clinician expertise.

Survey Development

Consumer experience and perceptions were explored in three dimensions: (1) the participants' experience and interests in mobile apps and wearable and health monitoring devices, (2) their preferred functions and priorities around health measures required in a digital support program for secondary prevention of stroke, and (3) their attitude toward a digital support program for secondary prevention of stroke. The consumer survey consisted of 28 questions, with a total of 108 items (Multimedia Appendix 1).

Similarly, clinicians' experience and perception were explored in three dimensions: (1) existing health measures commonly collected in the management of people with stroke and TIA, (2) the participants' priorities around health measurements required in a digital support program for secondary prevention of stroke, and (3) their attitude toward a digital support program for secondary prevention of stroke. The clinician survey comprised 16 questions, with a total of 135 items (Multimedia Appendix 2).

In both surveys, the questions related to the participants' experience and interest in mobile apps, wearable and health devices, as well as their preferred functions for a digital support program were adapted from surveys developed and previously used in mHealth research [26,29,34]. The questions related to the priorities around health measures commonly collected and proposed for a digital support program were also informed by previous mHealth research [26] and clinical guidelines for secondary prevention [8,9]. Finally, the constructs of attitude toward a digital support program were adapted from the technology acceptance model, previously validated and commonly used in mHealth studies with both patients and clinicians [35-37].

While the constructs of the consumer and clinician surveys are similar, differences exist due to the roles and experience of each participant group. Therefore, we aimed to focus more attention toward the consumer preferences and engagement with a digital program, while clinician surveys focused on clinical value (eg, relevant data collection) and usability. The final surveys were

XSL•FO
RenderX

tested for clarity by 4 adults (2 of them with clinical experience) before they were shared with participants.

Data Collection

Clinician surveys were collected using a web-based survey. For consumers, a study pack was posted to potential participants, including a participant explanatory statement and consent form, a paper-based survey with a return prepaid envelope, and a link to a web-based survey. Consumer participants were asked to complete only 1 version of the survey, either web- or paper-based, depending on their preference.

The web-based survey was captured using the REDCap (Research Electronic Data Capture; Vanderbilt University), a secure software platform designed to support data capture for research studies [38,39].

Stage 2: User Experience Workshops

Overview

The aim of this stage was to complete the co-design of the CAPS program and digital platform. Workshops were conducted via teleconference (Webex, Cisco Systems Inc) in 2 rounds. The first round (design workshops), completed between April and May 2021, focused on confirming insights from the user needs surveys and gathering design requirements from participants to develop initial prototypes of CAPS, while the second round (usability testing), completed between October and November 2021, focused on collecting usability feedback on the prototypes to inform the final design.

Participants

Purposive sampling was used in recruiting participants. For the design workshops, all survey participants who indicated their willingness for participation in future research were considered. The final selection process was designed to achieve a balanced sample based on participants' age, sex, diagnosis (for consumer participants), and professional roles (for clinician participants). In addition, an invitation was sent to the Stroke Foundation of Australia requesting for a lived experience representative and advocate to attend a workshop. Consumer participants were further screened via a phone call to include participants with different levels of familiarity with mobile apps and those able to connect remotely (via phone or videoconference).

For the usability testing, all participants from the design workshops were invited to participate. Invitations were also extended to a selected number of survey participants following the same criteria used in the design workshops.

Procedure

The user experience workshops were completed in three steps: (1) results from the stage 1 surveys were used to generate discussion topics for the design workshops; (2) findings from the design workshops were used to inform the design of wireframes and a web-based clickable prototype of a user-facing app and clinician portal; and (3) the clickable prototypes were shared with participants during the usability testing workshops to give them the opportunity to explore some of its features and provide feedback.

All workshop sessions were conducted separately with consumer and clinician participants. Workshop sessions lasted approximately 30 to 60 minutes, depending on the number of participants involved in the session. Each session was facilitated by 2 researchers. All sessions were audio recorded using Webex.

Data Analysis

Quantitative survey data were summarized and analyzed with Excel (Microsoft Corp) and Python (version 3.0; Python Software Foundation), using descriptive statistics. Ranking questions, where respondents ranked multiple items in order of preference, were calculated using weighted averages [40-42], as shown in equation 1:

in which w is the ranked position for each answer choice as selected by a participant, x is the response count for each answer choice, and n is the number of choices. Each participant's preferred choice (ranked as 1) has the largest weight n, and their least preferred choice has a weight of 1.

Qualitative data from the workshops were auto transcribed in Webex and cross-checked for accuracy, reviewed, and independently analyzed for themes by 2 researchers using Excel. An inductive approach was used to determine the coding of the data guided by the dimensions explored in the surveys and with a focus on design needs and preferences. There were no significant discrepancies between the reviewers. The final themes and subthemes (eg, design needs and preferences) were critically reviewed and discussed by the research team who addressed any discordance in the coding through consensus.

To avoid a misleading sense of statistical validity given the nonrandom purposive sample and the semistructured nature of workshops [43,44], we refrain from reporting qualitative results using whole numbers or percentages. Instead, to provide a sense of generalizability, we describe findings as emerging from all (100%), (approximately 76%-99%), many (approximately 51%-75%), some (approximately 25%-50%), or a few (<25%) participants.

Ethical Considerations

This project was conducted with ethics approval from the Monash University (Human Research Ethics Committee, HREC 24806) and the Commonwealth Scientific and Industrial Research Organisation (HREC 2020_076_RR). All participants provided written informed consent before their participation. Participants who completed the web-based survey (both consumers and clinicians) were directed to a digital explanatory statement and consent form before the start of the survey.

Participation in the surveys was anonymous. However, after completion of the survey, participants were given the option to provide their contact details if they were willing to be further involved in the research program. This information was stored in a separate REDCap database unlinked from the survey responses. Workshop participants received an AUD $50 (US $33.82) gift card for their participation.

XSL•FO
RenderX

Results

Stage 1: User Needs Analysis

Participant Summary

A total of 112 consumers (male consumers: n=63, 56%) aged between 40 and 98 years completed the survey (Table 1). Of the 112 consumers, 16 (14.3%) responded to the web-based survey and the rest responded via post. A total of 101/112 (90.2%) answered at least 1 item in all questions and 85 (76%) completed all items. Most of them (110/112, 98.2%) were registrants of the AuSCR, and only 1.8% (2/112) of them responded to the invitation distributed through the Stroke Association of Victoria.

Most participants had lived experienced of stroke (72/112, 64.3%) within the last 5 years, were taking medications to manage their symptoms of stroke or reduce the risk of a future stroke (104/112, 92.9%), were not seeing a health professional to support their recovery or ongoing management after stroke (71/112, 63.4%), and were not participating in a stroke or TIA rehabilitation program (108/112, 96.4%).

A total of 54 individuals (female: n=43, 80%) with an average of 17 years of experience since finishing their entry-level clinical training responded to the clinician survey. Most of them (42/54, 78%) answered all questions. A further 24 participants answered only demographic questions and were not included in further analysis (Table 2).

XSL•FO
RenderX

Table 1. Characteristics of the consumer participants who have experienced a stroke or transient ischemic attack (n=112).

Characteristics	Values
Demographics	
Sex, n (%)	
Male	63 (56.3)
Female	45 (40.2)
No answer	4 (3.6)
Age (y), mean (SD)	71.94 (10.25)
Education, n (%)	
≤Year 12	68 (60.7)
Vocational qualification	11 (9.8)
Associate diploma	2 (1.8)
Undergraduate studies	12 (10.7)
Postgraduate studies	15 (13.4)
No answer	4 (3.6)
Living arrangements, n (%)	
Alone	30 (26.8)
With others	78 (69.6)
No answer	4 (3.6)
Medical history	
Years since most resent stroke or TIA[a], mean (SD)	2.65 (0.86)
Stroke or TIA, n (%)	
Stroke	72 (64.3)
TIA	29 (25.9)
Don't know	11 (9.8)
Participating in a stroke rehabilitation program, n (%)	
Yes	4 (3.6)
No	106 (94.6)
No answer	2 (1.8)
Receiving a home rehabilitation program, n (%)	
Yes	2 (1.8)
No	108 (96.4)
No answer	2 (1.8)
Seeing a health professional to support stroke management, n (%)	
Yes	39 (34.8)
No	71 (63.4)
No answer	2 (1.8)
Take medications to manage stroke or TIA, n (%)	
Yes	104 (92.9)
No	6 (5.4)
No answer	2 (1.8)

[a]TIA: transient ischemic attack.

XSL•FO
RenderX

Table 2. Demographics and characteristics of the clinician participants. A total of 78 participants answered all questions, with 24 who answered only demographic questions being excluded from further analysis.

Demographics and characteristics	Participants (n=54)	Removed participants[a] (n=24)
Sex, n (%)		
Male	11 (20)	5 (21)
Female	43 (80)	19 (79)
Age (y), n (%)		
≤29	7 (13)	1 (4)
30-39	18 (33)	9 (38)
40-49	15 (28)	4 (17)
50-59	11 (20)	9 (38)
60-69	3 (6)	0 (0)
70-79	0 (0)	1 (4)
Years of experience, mean (SD)	16.76 (10.13)	17.58 (9.86)
Percentage of time providing stroke or TIA[b] care, mean (SD)	54.77 (34.88)	58.1 (33.59)
People with stroke or TIA treated per week, mean (SD)	9.62 (8.03)	10.75 (7.52)
Profession, n (%)		
Medical doctor[c]	13 (24)	4 (17)
Occupational therapist	12 (22)	0
Physiotherapist	8 (15)	8 (33)
Nurse	12 (22)	6 (25)
Speech pathologist	4 (7)	1 (4)
Social worker	1 (2)	2 (8)
Other[d]	4 (7)	3 (13)
Clinical setting, n (%)		
Acute hospital	25 (46)	9 (38)
Community-based service	5 (9)	5 (21)
Primary health care	1 (2)	1 (4)
Rehabilitation	19 (35)	4 (17)
Other	4 (7)	5 (21)

[a]Participants who only completed the demographic questions and were removed from further analysis.

[b]TIA: transient ischemic attack.

[c]Includes neurology, cardiology, rehabilitation, and general medicine.

[d]Includes stroke coordinator, radiographer, and exercise physiologist.

Experience Using Mobile Apps

A total of 81.2% (91/112) of the consumers reported using smartphones or tablets, with 58 (64%) of them using apps on a regular basis and 22 (38%) of those using apps to manage their health and well-being, particularly to track their exercise and activity levels (n=20, 35%). Many of them reported owning and using a weight scale (74/112, 66.1%) and blood pressure monitor (58/112, 51.8%), and only 26.8% (30/112) reported owning and using a smart watch or fitness tracker (Figure 1). From a clinician perspective, only 37% (20/54) of the clinicians have recommended their patients mobile apps to manage their health and well-being, particularly for physical activity monitoring (14/20, 70%) and mental health support (8/20, 40%).

XSL•FO
RenderX

Figure 1. Total of the current device use by consumer participants (N=112). ECG: electrocardiogram.

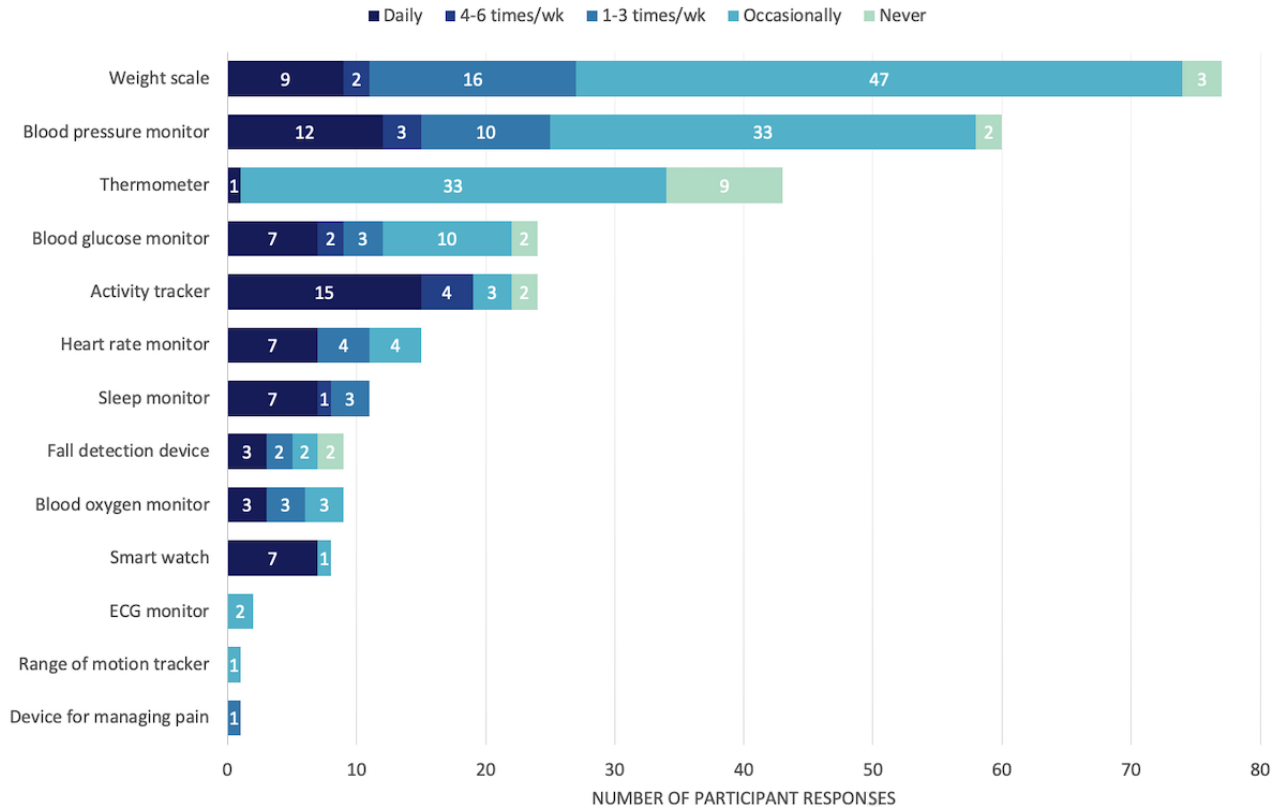

Preferred Functions

When asked about their preferred functions to include in a digital support program for stroke or TIA, consumers and clinicians prioritized two main functions (Table 3): (1) providing alerts to patients about potential deterioration in their condition and (2) collecting and monitoring lifestyle measures (eg, steps, sleep, and exercise). They also prioritized data accuracy and the need for a program that was easy to set up and use (Figures 2 and 3).

Table 3. Top 15 functions to be included in a digital support program for people after stroke or TIA[a], as ranked by participants. Consumers ranked function from 1 (most important) to 3, while clinicians ranked them from 1 (most important) to 5.

Preferred function	Consumer rankings	Clinician rankings
Alerting patients about potential deterioration in their condition	1.72	2.04
Collection and monitoring lifestyle measure (eg, steps and sleep)	0.52	1.76
Tools to manage new lifestyle (eg, exercises, diet, alcohol, and smoke)	0.17	1.67
Educational information in general about stroke or TIA treatment and prevention	0.27	1.55
Alerting clinicians about potential deterioration in a patient's condition	__[b]	1.49
Collection and monitoring of medical measures (eg, pain and blood pressure)	0.58	1.41
Reminders about the warning signs of stroke or TIA and what to do	—	0.96
Receiving encouraging or persuasive messages	0.01	0.94
Tools to manage your medication	0.13	0.65
Reminders (eg, appointments)	0.27	0.53
Tools to manage mental health	0.19	0.51
Supporting clinician-patient communication	0.3	—
Receiving instructions from clinicians (eg, text via app, SMS text message, and emails)	0.29	—
Receiving virtual *awards* as encouragement for achieving goals	0	0.33
Supporting clinician-patient communication (via messages from patients)	—	0.25

[a]TIA: transient ischemic attack.

[b]Data not available for this population.

XSL•FO
RenderX

Figure 2. Consumer participants' ratings on the main considerations for a digital support program for people after stroke or transient ischemic attack (N=112).

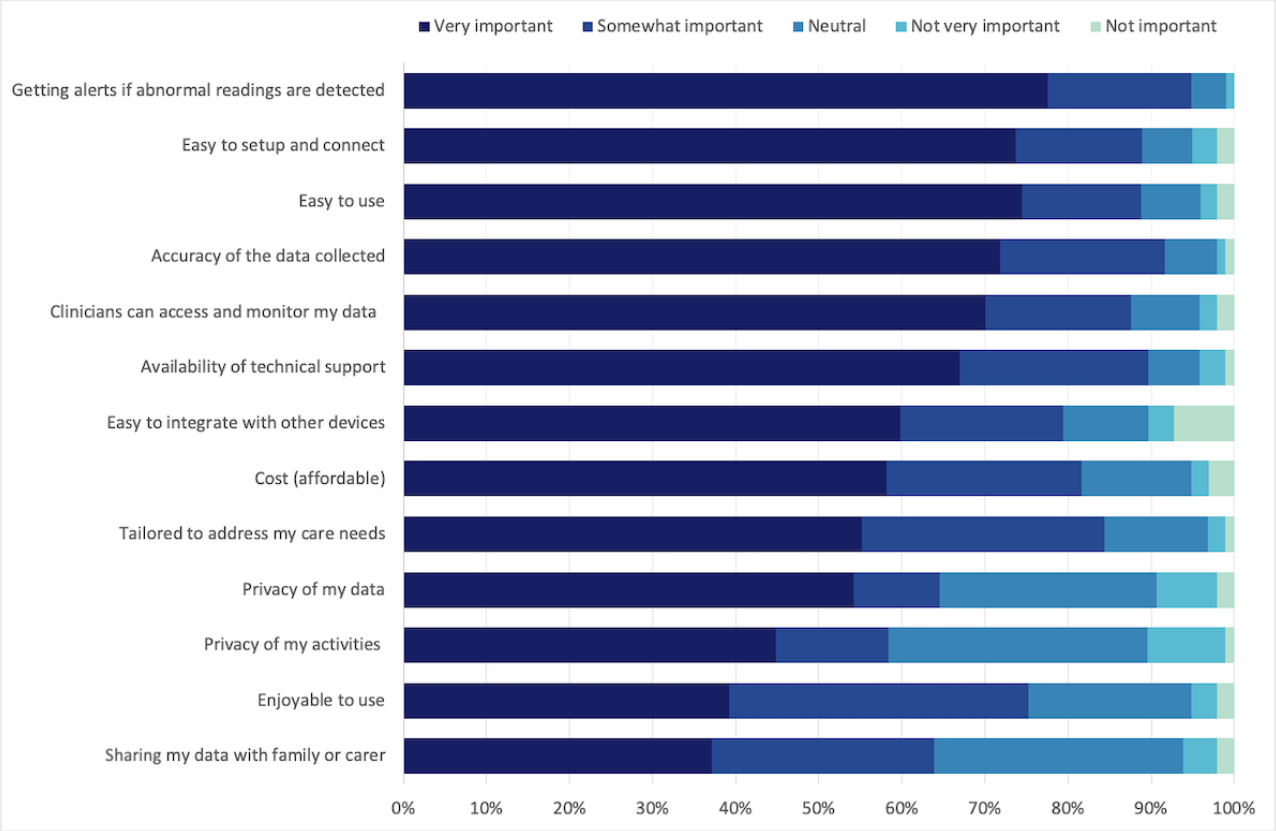

Figure 3. Clinician participants' ratings on the main considerations for a digital support program for people after stroke or transient ischemic attack (n=42).

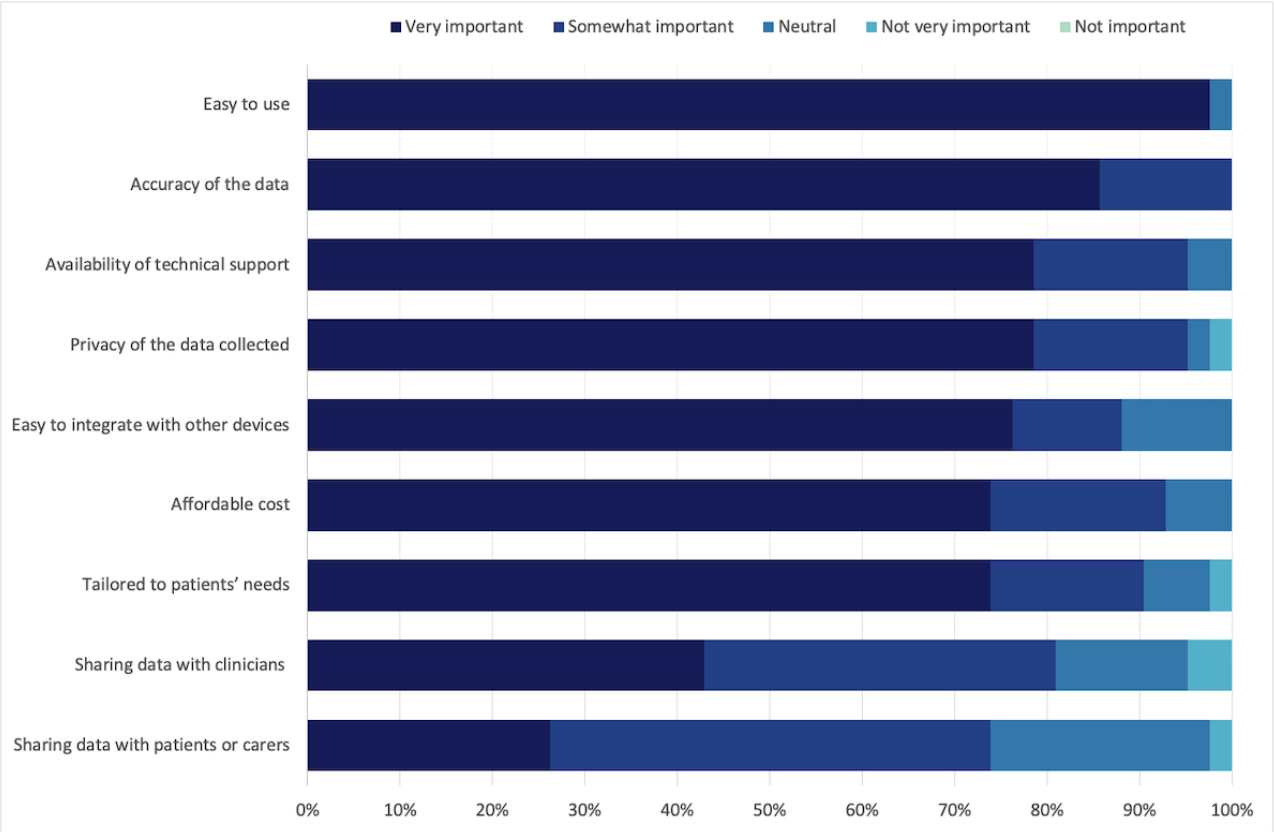

XSL•FO
RenderX

Preferred Measures to Collect and Monitor

Consumers rated physiological measures (eg, blood pressure, heart rate, and sleep quality), followed by mobility (eg, range of motion) and mental health measures (eg, stress and anxiety), as the top health and well-being indicators to monitor. Clinicians agreed on the importance of physiological and mental health measures but rated the collection of lifestyle data (eg, alcohol consumption, tobacco use, physical activity, diet, and nutrition) above all (Figure 4). Blood pressure, exercise, and heart rate were ranked by clinicians as the most important measures to collect via wearable and health monitoring devices (Table 4).

Figure 4. Consumers' (n=96) and clinicians' (n=48) rating of the types of measures that should be collected in a digital support program for people after stroke or transient ischemic attack.

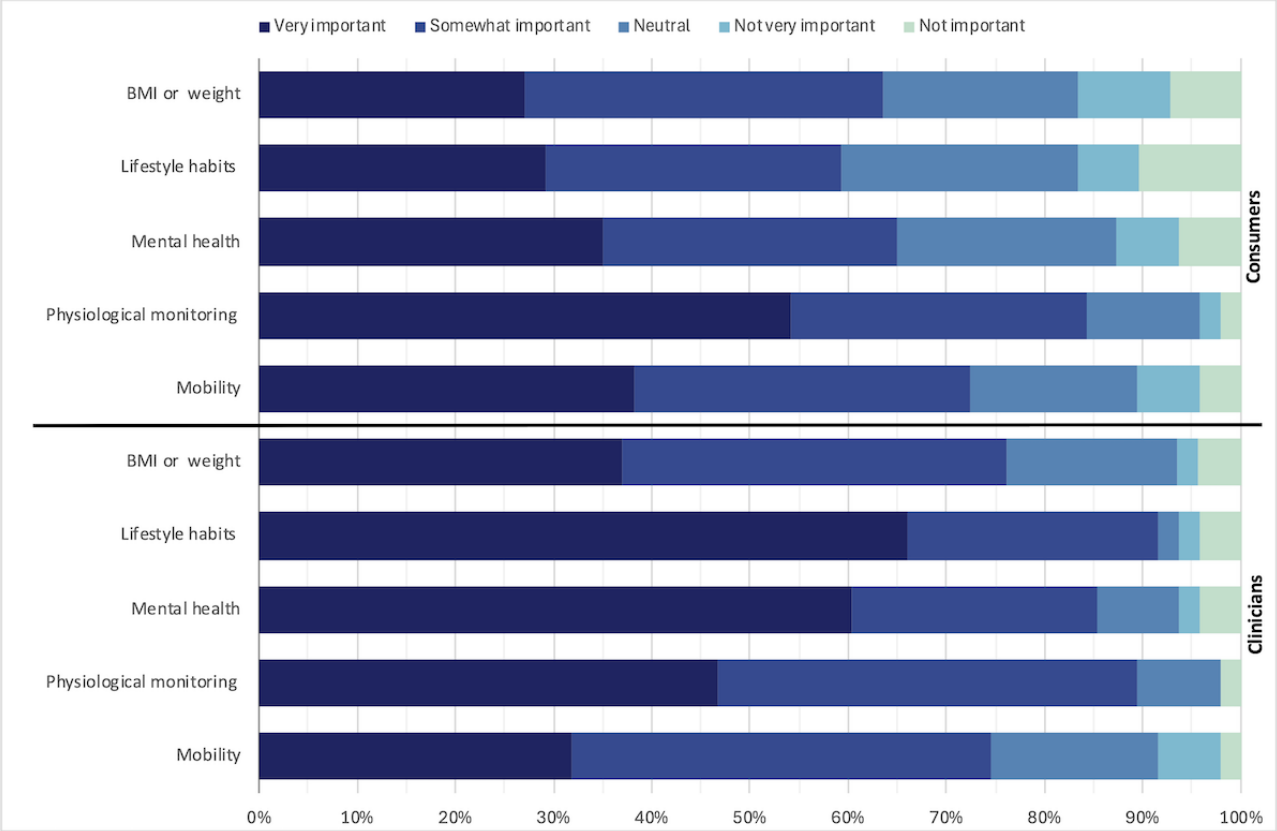

Table 4. Clinician participants' ranking of measures that would provide value and utility in the management and secondary prevention of stroke. Participants ranked each function from 1 (most important) to 3.

Measurement	Ranking score
Blood pressure	1.76
Exercise	1.43
Electrocardiogram/heart rate/rhythm	1.29
Falls	0.60
Blood glucose	0.38
Body weight	0.31
Pain	0.29
Blood oxygen	0.1
Sleep	0.1
Range of motion	0.07
Body temperature	0.05

Most clinicians (46/49, 94%) also indicated the need for data monitoring more frequently than current practice, particularly related to waist circumference, dietary intake, and physical activity (Multimedia Appendix 3). Different preferences existed depending on the profession.

XSL•FO
RenderX

Perception of Digital Technology to Support Their Health

When asked about their perceptions of digital technology to manage and support their health goals after stroke or TIA, most consumers and clinicians agreed that a program that uses digital technology could be useful to monitor the consumers' health and well-being (consumers: 78/97, 80%; clinicians: 40/42, 95%) and help them adapt to a new lifestyle (consumers: 76/99, 77%; clinicians: 38/42, 90%). More than 70% of consumers and clinicians also noted their intention to use the digital program as often as needed (consumers: 71/99, 72%; clinicians: 33/42, 78%) and discuss the data collected with their physicians (patients consumers: 79/99, 80%; clinicians: 36/42, 86%; Multimedia Appendices 4 and 5). Clinicians agreed they would review the data during consultations (26/42, 62%), at least once a week (10/42, 24%), or occasionally (3/42, 7%). Only 3 (all nurses) mentioning that they were too busy to review the data. When asked how they want to access the collected data, >97% (41/42) of clinicians noted their interest to access the data. Moreover, 52% (22/42) would like to access the data via a clinician web portal and a mobile app, 21% (9/42) would like to look at the data via the consumers' app during consultations, 17% (7/42) would like to access the data via a clinician web portal, 5% (2/42) would like to access the data via a clinician app, and 2% (1/42) would like the data to be linked to the patients' medical records. Most clinicians (35/42, 83%) expressed their intention to use the program to care for people after stroke or TIA.

Stage 2: User Experience Workshops

Overview

A total of 7 consumers (male individuals: n=5, 71%) and 9 clinicians (female individuals: n=6, 67%) took part in the user experience workshops. In the design stage, 3 sessions (1 group and 2 individual sessions) were conducted with 5 consumers and 3 sessions (2 group and 1 individual sessions) were conducted with 6 clinicians. In the usability testing, a total of 4 individual sessions were conducted with consumers and 4 with clinicians. A total of 2 consumers and 2 clinicians took part in both the design workshops and usability testing (Table 5).

Table 5. Participant demographics of the user experience workshops.

Participant demographics	Consumers (n=7)	Clinicians (n=9)
Sex, n (%)		
Male	5 (71)	3 (33)
Female	2 (29)	6 (67)
Age (y), range	67-84	<29-49
Workshops, n (%)		
Design	3 (43)	5 (56)
Usability testing	2 (29)	2 (22)
Both	2 (29)	2 (22)
Diagnosis, n (%)		
Stroke	4 (57)	—[a]
Transient ischemic attack	3 (43)	—
Profession or role, n (%)		
Registered nurse	—	2 (22)
Neurologist	—	1 (11)
Physiotherapist	—	2 (22)
Stroke Foundation representative	—	1 (11)
Stroke coordinator	—	2 (22)
Speech pathologist	—	1 (11)

[a]Data not available for this population group.

Design Workshop

Qualitative analysis focused on four practical themes of user experience and preferences: (1) the consumers' experience using mobile apps and wearable devices, (2) consumer and clinician perception of mHealth for secondary stroke prevention, (3) preferred functions, and (4) preferred health indicators for a digital program for secondary prevention of stroke. In these workshops the terms *patient* and *client* were used interchangeably by clinicians when referring to consumers.

Consumers' Experience Using Mobile Apps and Wearable Devices

All consumer participants confirmed that they used digital technology regularly, including smartphones, tablets, and PCs, for activities such as banking, taking photographs, appointment

XSL•FO
RenderX

reminders, phone calls, and taking notes. Most of them also used alarms for medication reminders. Many of them used older devices that were given to them by their children or grandchildren and mentioned that while they knew how to use them, they did not feel as comfortable as the younger populations or as they were before their strokes:

> *I've got my daughter's old Apple phone. I learned it quickly. After stroke I managed it reasonably well still, but I cannot do everything on it, I don't know how to do it.* [Consumer 2]

Most of them used apps to track their exercise, particularly their step counts. Only 1 participant reported using a smart watch (Samsung watch, Samsung Electronics Co Ltd) to track their exercise. No other health-related apps were reported, with consumers noting that they only used these apps when they were suggested by their treating clinicians.

Perception of mHealth for Secondary Stroke Prevention

Clinicians mentioned that regular contact with consumers was only provided while they were actively involved with their service (eg, rehabilitation therapy), with only annual monitoring provided after discharge. They noted that after discharge, self-monitoring was important, with some clinicians expressing their interest in a telehealth model of care where they could monitor consumers remotely (eg, via a clinician portal) and look at a summary of their data during clinical visits. They all agreed that providing tools to educate and empower consumers to self-manage and self-monitor their condition was important:

> *I feel like a lot of our role is around education...they also need to have the tools to enable them to...perform some of that self-monitoring where they are able to do so.* [Clinician 1]

All consumer participants supported the idea of a mobile app to help them manage their stroke or TIA.

Preferred Functions

Consumers noted that receiving alerts or early warnings of potential deteriorations and medication reminders were their main interests:

> *Medication alarm is my favorite...* [Consumer 2]

They also mentioned that the ability to record information on their phone to share with their treating clinicians would be relevant. They expressed particular interest in accessing a graph or a summary of their information to track their own progress. Most clinicians agreed that the collection of regular and accurate data to support self-monitoring and clinical management would be beneficial to track changes over time, particularly during the first weeks to months after discharge:

> *I can see its benefits outside the acute setting in the weeks and month following discharge.* [Clinician 7]

One consumer highlighted that their handwriting was becoming difficult to read and expressed the need for an app to replace handwritten notes to record information that needed to be shared with their physician.

When asked about the need for educational content, consumers agreed that they trusted their clinicians and hospital more than the information available in the web, so they were not interested in alternative sources of information. However, they were interested in having access to the information provided to them during their clinical visits, including the warning signs of stroke and updated information from their clinicians:

> *I trust the specialist who has kept me alive.* [Consumer 1]

Both clinicians and consumers agreed that the use of gamification techniques, such as virtual medals, was inappropriate:

> *Receiving virtual medals may be to a very, very slight few rewarding...but in general really not appropriate.* [Clinician 6]

Preferred Health Indicators to Collect via mHealth for Secondary Prevention

When talking about data collection, consumers agreed that blood pressure was the most relevant data they collected to manage their risk factors after stroke. Some of them collected it regularly, while others only if they felt unwell or if their medication changed. In this vein, clinicians agreed that consumer compliance was high, particularly in the context of the collection of blood pressure. However, they mentioned that lifestyle and exercise data were not routinely collected even if it was considered important because consumers' reports were often unreliable:

> *A lot of this relies on self-reporting, and patients might under-report or overestimate the amount of exercise.* [Clinician 4]

Some clinicians mentioned that well-being and lifestyle data were regularly collected over the phone, with fatigue often coming up as relevant. Consumers echoed their interest to record and track their fatigue levels after stroke using a reliable scale.

Some clinicians also mentioned that social isolation was particularly important for people living in remote areas and noted that this was commonly measured visually (eg, signals of neglect, lack of access to medications, and disheveled appearance) as some consumers preferred not to talk about it:

> *Talking to clients, hoping that they are willing to talk about isolation and support. Some clients willingly and openly talk about it and others won't.* [Clinician 5]

Few clinicians mentioned that electrocardiogram and waist circumference were not particularly important, while 1 participant noted that medication adherence and health care use history were relevant.

Usability Testing

Overview

Following the design workshops, wireframes and prototypes for the CAPS program were developed using 3 main technologies: a consumer-facing mobile *app* accessed via smartphones or tablets, a *smart watch*, and a *clinician web portal*. These prototypes were used to collect additional participant feedback during the user testing workshops. This

XSL•FO
RenderX

section provides the main insights from both consumer and clinician to guide the final design of the CAPS platform.

Mobile App and Wearable Feedback and Insights

The app design needs to be simple, with a shallow architecture (without multiple screen levels) to reduce cognitive load. The use of pictures and icons was preferred over text. Labels and instructions should be clear, with large buttons and scalable font size. Some participants also suggested an option for the app to "speak" instruction and the ability to collect data through voice, for consumers whose reading or writing ability was affected:

> *[L]ove the fact that you've got the icons there, it's really easy to see what each of the options are and represent.* [Clinician 1]

The home screen should prioritize the most important functions, with a daily check in presented as steps and with clear signs to mark its completion. Many participants expressed the need to view historical data points.

A "symptoms" screen that provides educational information about the early signs of stroke was suggested, but there were mixed opinions about the need to include any other educational content within the app.

The ability to add notes into a health journal was noted. Furthermore, some participants mentioned that using a camera to add pictures within the notes might be difficult for some users and expressed privacy concerns.

Participants also supported the inclusion of several data sources, including weight, blood glucose levels, alcohol and tobacco consumption, ECG, measures of social interaction and social support, and an option to add activities (eg, walking and swimming) and step counts using a smartphone if a smart watch was not available.

Finally, participants questioned the practicalities of collecting survey data via the smart watch, given the small form factor, but strongly supported the use of the smart watch for sensor-based data collection and reminders.

Clinician Portal

The clinician portal was designed to be used for onboarding and monitoring of consumers. In this vein, participants noted that the onboarding process should allow clinicians to record consumer's consent and a set of secondary prevention goals. Data collection needs and alerts, when data are out of bounds, should be customizable via the portal, so that they are personalized to consumers. There was also a suggestion for clinicians to manage medication reminders via the clinician portal, providing consumers with the ability to track their compliance via a "tick box" within the app:

> *[P]erhaps have a little reminder to say, okay, it's time to take your medication. This might be something that the clinician sets up.* [Clinician 3]

For consumer monitoring, some participants requested a customizable dashboard view with traffic light system (red, orange, and green) to quickly flag individuals that need prompt follow-up and facilitate monitoring. They also highlighted the importance of an "export function" that would allow them to download (as pdf, csv, etc), print out, and share a summary of the data when needed.

The CAPS Platform

Overview

The final CAPS platform (Figure 5) was designed using the evidence from the co-design stages and consists of a mobile app (available in iOS and Android versions), supported by both smartphones and tablets, that integrates wearable activity trackers and smart devices (Fitbit watches and fitness trackers, Oura ring, or Apple Watch) and a clinician web portal (Portal) that includes an electronic message system (Message system). All data collected by the wearable devices are synchronized via Bluetooth to the app and uploaded to the Portal.

XSL•FO
RenderX

Figure 5. The Care Assistant and support Program for people who have experienced a Stroke or transient ischemic attack (CAPS) platform model (left) and the CAPS mobile app home screen (right).

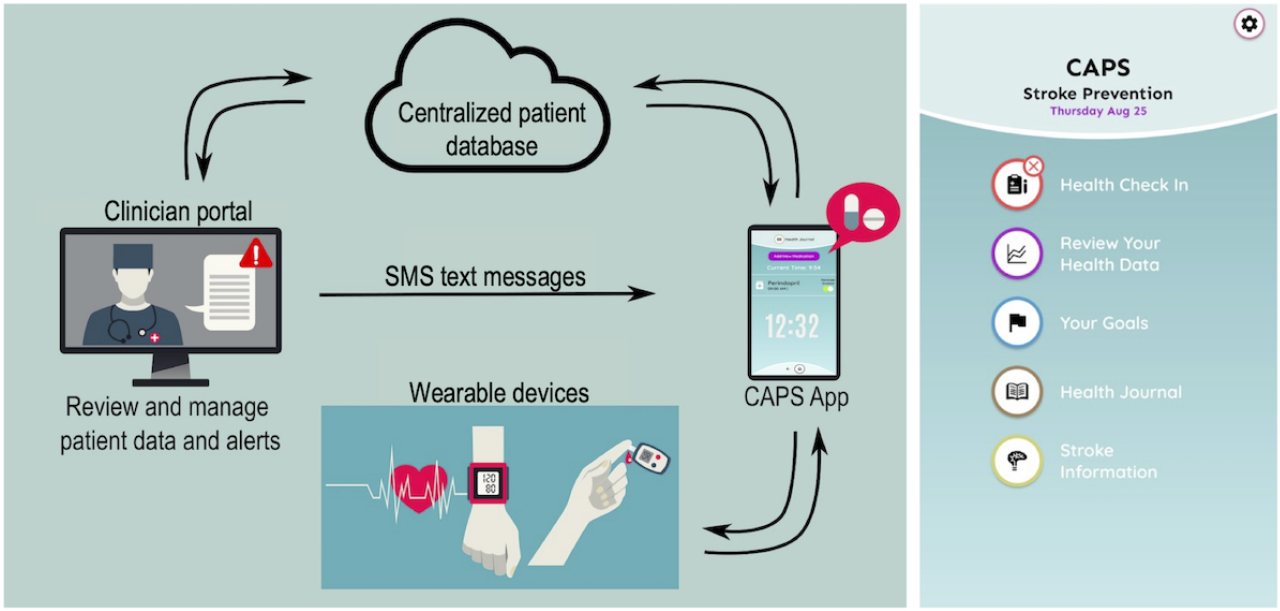

Mobile app

The app is designed for use by consumers daily. Figure 6 highlights the key features.

Figure 6. Key features of the CAPS (Care Assistant and support Program for people who have experienced a Stroke or transient ischemic attack) mobile app. From left to right: health check in (distress level question 7 of 12), review health data, goals, health journal, and stroke information.

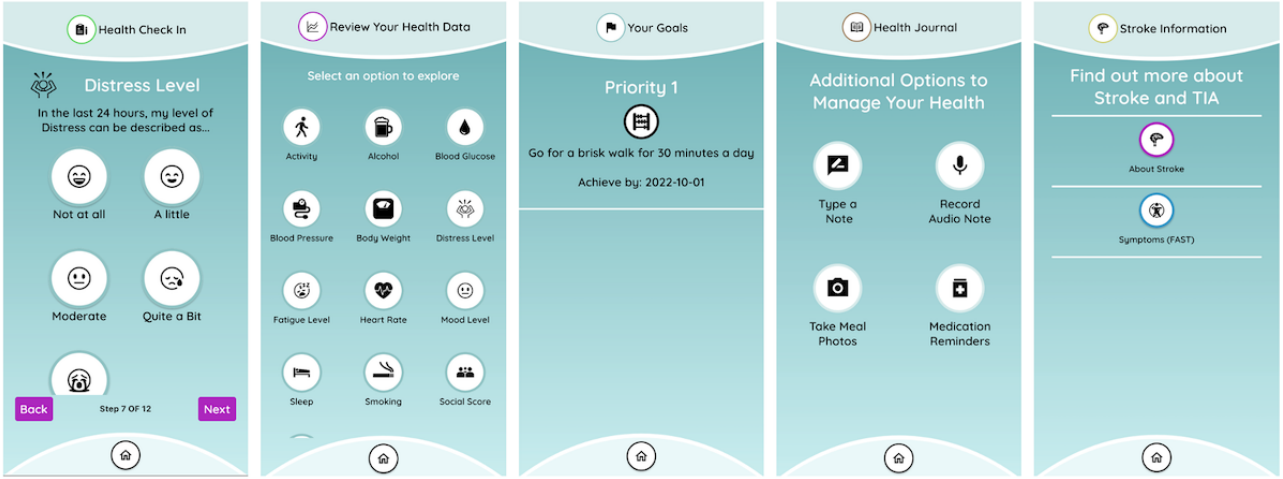

Health Check In

A daily check in of self-recorded data, including activities, blood pressure, distress, and other relevant health measures customized by clinicians during onboarding (Table 6). Most of the questions and answers included in this section were informed by the literature or clinical expert advice, including those related to blood pressure, blood cholesterol, blood glucose, body weight, alcohol and cigarette consumption [8], mood [45,46], distress [46], fatigue [47], social connection [48-50], social support [48-50], and social isolation [51]. The daily health check in was designed to be easy to understand and takes <5 minutes to complete.

XSL•FO
RenderX

Table 6. Health check-in questions available in the CAPS (Care Assistant and support Program for people who have experienced a Stroke or transient ischemic attack) app.

Item	Question	Answer options
Activities	Select activities done in the last day	Walking, swimming, cycling,..., other exercise
Alcohol	Number of alcoholic drinks in the last day	Discrete numbers
Blood glucose	Current blood glucose (mmol)	Discrete numbers
Blood pressure	Systolic and diastolic value	Discrete numbers
Body weight	Current weight (kg)	Discrete numbers
Tobacco	Number of cigarettes consumed in the last day	Discrete numbers
Distress	In the last 24 hours, my level of distress can be described as...	Not at all, a little, moderate, quite a bit, and extremely
Fatigue	My fatigue level is...	Not fatigued at all, a little fatigued, moderately fatigued, very fatigued, total fatigue, and exhaustion
Mood	Today I feel...	Happy, somewhat happy, neutral, somewhat sad, and sad
Social connection	In the last 24 hours, I was satisfied with the amount of time I spent talking or meeting with people who do not live with me	Strongly agree, agree, neutral, disagree, and strongly disagree
Social support	In the last 24 hours, support from family, friends, health workers, or others was available when I needed it	Strongly agree, agree, neutral, disagree, and strongly disagree

Review Health Data

Reviewing and monitoring of historical self-recorded data and data collected by the wearable devices (eg, heart rate, physical activity, and sleep). Data are graphically displayed.

Goals

Secondary prevention SMART (specific, measurable, achievable, relevant, and time-bound) goals [52,53] developed by clinicians with consumers are captured within the app.

Health Journal

Notes (text and audio) and photos can be taken and recorded for anything that is helpful or relevant to consumers. These data are kept private within the mobile device and are not uploaded to the clinician portal. Medication reminders can be set up to prompt consumers to take their medications.

Stroke Information

In addition to a symptom section based on the FAST (facial drooping, arm weakness, speech difficulties, and time to call emergency) acronym, web links related to practical information of stroke prevention and healthy living are provided.

Settings (Top Right Gear)

Allows consumers to manage SMS text messages, notifications, and font size.

Portal

The portal was designed for use by clinicians and their assistants during onboarding and data monitoring. Key features are described below.

Onboarding and Setup

Health measures required to be collected by consumers, thresholds for health measure alerts, and secondary prevention goals can be set and captured based on the characteristics of individual consumers.

Consumer Dashboard

Presents an “at a glance” summary of all consumers, including goals and alert highlights if a consumer’s health measures are out of normal range. Data can be downloaded into CSV or PDF files (Figure 7).

XSL•FO
RenderX

Figure 7. The Care Assistant and support Program for people who have experienced a Stroke or transient ischemic attack clinician portal. From top to bottom: consumer dashboard home screen, participant goals screen, alerts screen, and patient recorded data screen (blood pressure graph).

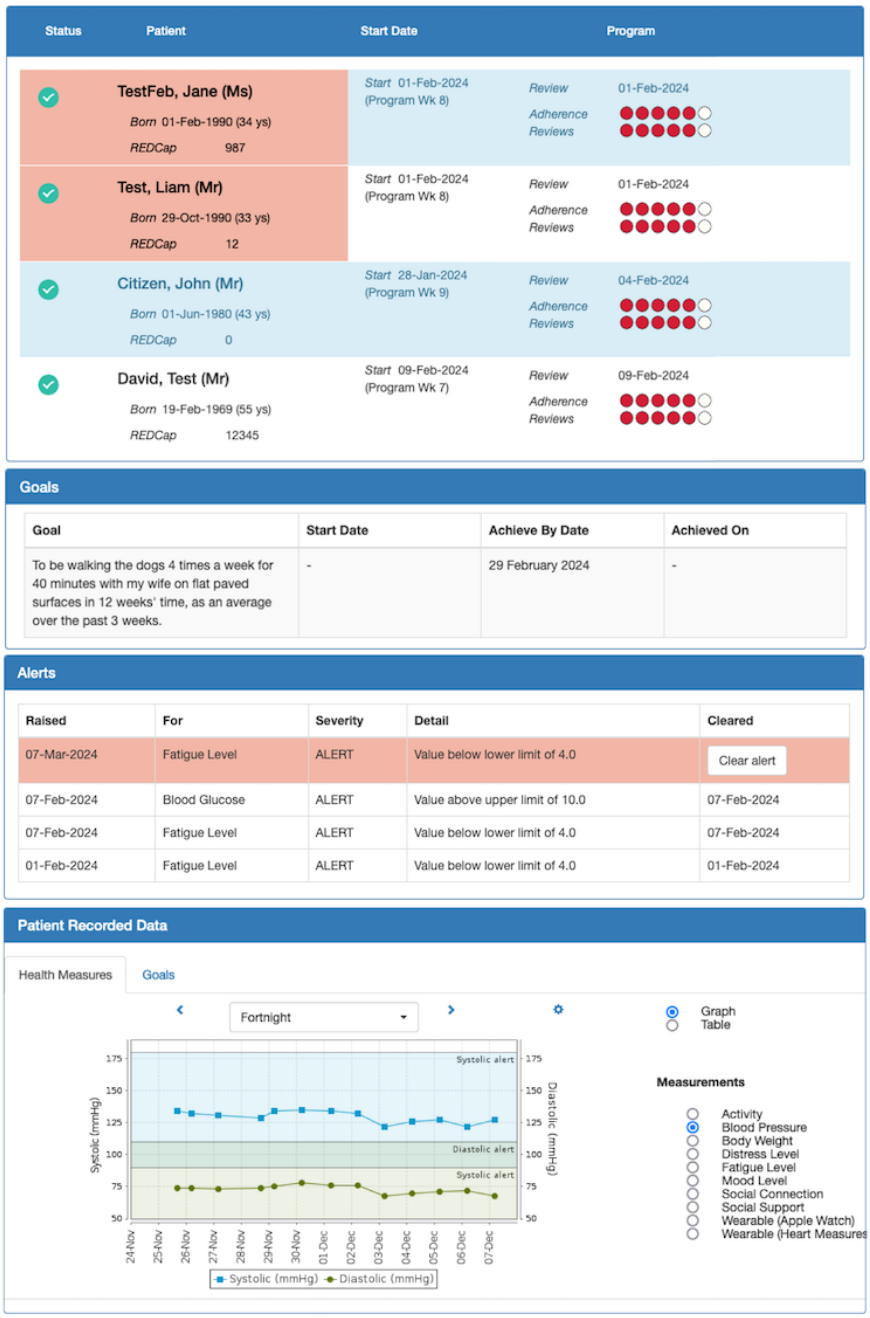

Consumer Record Details

Presents a summary of a consumer's personal details, collected data, and goals. Longitudinal data gathered via the consumer app and wearable can be displayed in graph or table formats.

Message System

Clinicians can assign personalized electronic messages to be automatically sent to consumers, aligned with their secondary prevention goals.

Discussion

Principal Findings

Effective management of long-term lifestyle behavior changes following discharge after stroke or TIA is essential for secondary prevention. This paper presents the co-design of a novel, multicomponent, clinician-facilitated, patient-centered digital program to support secondary prevention of stroke, which is now ready for feasibility testing with clinicians and people with lived experience of stroke or TIA.

The development of CAPS followed a multistaged co-designed process that considered the needs and preferences of both consumers and clinicians involved in stroke management. The user needs surveys and user experience workshops confirmed

XSL•FO
RenderX

the need for a program to support self-management of lifestyle changes after discharge, with most participants managing their lifestyle changes without a rehabilitation program (108/112, 96.4%) or the support of a health professional (71/112, 63.4%).

When considering a digital program for secondary prevention of stroke, participants noted particular interest on supporting an individual's ability to adapt to a new lifestyle (eg, physical activity), the collection and monitor of lifestyle (eg, steps and sleep) and physiological measures (eg, blood pressure and heart rate), and receiving alerts about potential deterioration in their health. This was expected, given that clinical practice guidelines prioritize the control of blood pressure and atrial fibrillation for stroke prevention. In addition, lifestyle modifications to reduce individual risk for cardiovascular disease or stroke are key performance quality indicators of care; however, >25% of patients are not provided discharge education targeting their personal cardiovascular risk factors, indicating the need to address this clinical practice gap [54].

Surveys also highlighted the clinical need for more regular access to consumers' data. These results align with the global trend to incorporate emerging technologies, such as mobile apps, sensors, and wearable devices, in health applications that target early identification and tailored treatment of health conditions [55].

With most consumer participants (91/112, 81.2%) reporting the use of smartphones regularly and most clinicians (33/42, 78%) and consumers (71/99, 72%) reporting their intention to use the program, the potential for an mHealth program to support secondary prevention for stroke was strong. Participants also believed a mobile app would fit well with their routine and has the potential to improve their recovery and quality of life.

When considering the use of a mobile app in secondary prevention of stroke, participants highlighted the importance of accuracy in the collected data. The need to provide clinicians access to the collected data was also rated highly by consumers. This aligns with previous mobile health evaluations that have shown the value of remote monitoring in a patient's engagement and motivation [34]. The inclusion of general information about stroke or TIA treatment and prevention that consumers could access in their own time was suggested by clinicians. Consumers, by contrast, preferred to receive their educational content directly from their hospital or treating clinicians but supported the inclusion of the information provided to them during their clinical visits (including the warning signs of stroke) into the app. The inclusion of a function to provide medication reminders was also requested by some consumers.

The CAPS Platform

On the basis of the information collected during the user needs surveys and the design workshops, a prototype of a consumer-facing mobile app accessed via smartphones or tablets, a smart watch, and a clinician web portal were developed and user tested with both consumer and clinician participants. During this stage, participants noted the need for the app to be easy to use, with large font size, large buttons, and iconography to facilitate user engagement. The ability for the app to "speak" instruction and collect data through voice was also noted for consumers whose reading or writing ability is affected. They also questioned the need to share personal notes or recordings and photographs with clinicians, due to privacy concerns.

Following the user testing stage, a mobile app and clinician portal were developed using a user-centered design that includes a single check-in stage that is clear, simple, and can be tailored (during onboarding) based on consumer needs. While the questions included in the check-in stage were informed by the literature, their correlation with standard tools (eg, the Depression Anxiety Stress Scale-21 item [56] and the Duke Social Support Index [48]) will be evaluated in future research. The collected data can be displayed on a graph (on both the mobile app and the clinician portal) or can be downloaded and printed to be used more effectively by both consumers and clinicians. Informed by the literature [57,58], the program also allows clinicians to send SMS text messages to participants with relevant information to support their secondary prevention goals. One-way SMS text messages were preferred over app notifications due to the increased ability to ensure the delivery of the messages. The inclusion of gamification techniques was discarded, as it was considered inappropriate for this population.

An important consideration was the inclusion of a medication reminder function. While the need for this function was supported by clinicians and consumers, there was no agreement on the safety and practical implementation. After detailed consideration, it was decided to provide consumers with the ability to include their own medication reminders as needed with notifications in both, the app and the wearable device, but not the ability to track compliance due to increased risk on medication errors. The clinician portal was developed to be accessed via a website, as was indicated by clinician preferences (29/42, 69%).

Clinicians and consumers highly rated the importance of collecting data on mental health (eg, stress, anxiety, and depression). Psychological distress, which incorporates elevated symptoms of depression, anxiety, and psychological stress [59], occurs frequently in people living with stroke [60]. Poststroke emotional distress is also associated with poor clinical outcomes, including poor medication adherence, poor quality of life, poor engagement with rehabilitation, and increased mortality [45]. Hence, it was decided, that the daily check in should capture both mood and distress to enable analysis and correlation with validated tools, which will provide evidence for the most appropriate item to remain in the mobile app. Emoji icons to represent mood and distress were selected guided by previous work [45]. There was also significant clinical support for the collection of fatigue and social isolation data, which has become particularly relevant since the COVID-19 pandemic.

The CAPS program aligns with the World Health Organization's recommendations for digital health interventions that support equitable access to quality health services that strengthen and scale up health promotion and disease prevention strategies [61]. A strength of the CAPS programs is the co-design approach and process to ensure this digital health strategy addresses the needs of people with lived experience of stroke or TIA and clinicians. We are now ready to take the program to feasibility testing to ensure it is something that would be of benefit to

XSL•FO
RenderX

implement into clinical practice as a secondary prevention stroke program.

Limitations

While AuSCR facilitated access to registrants across Australia and provided access to a broad range of people with stroke (72/112, 64.3%) and TIA with wide age range (aged 40-98 years), it was difficult to access consumers with a recent experience of stroke or TIA (ie, within the last 6 months). Future work should explore alternative recruitment strategies to access those consumers.

While >50 clinicians answered all survey questions, including 13 medical doctors (eg, a neurologist, a cardiologist, and rehabilitation clinicians), we did not receive any responses from general practitioners. Lack of representation from primary health care needs to be addressed in future stages of the project, including feasibility testing and clinical evaluations. In addition, due to our recruitment methods, the clinicians who opted to participate were likely to be interested in the use of digital health to improve care for patients and may not have been fully representative of the greater stroke health care community.

Comparison With Previous Work

There is currently little information in the literature for evidence-based and available digital health platforms specifically designed to support people living with stroke. However, there is increasing research in how mobile apps can be used to support people living with stroke. They have been used to facilitate the delivery of interventions promoting increased physical activity [62,63], blood pressure control [62,63], medication adherence [64,65], and stroke awareness [66,67]. Additional intervention components that have been used to supplement the mobile app include SMS text messages, gamification, GPS tracking, and pedometers, often with remote monitoring capabilities for clinicians [63,64,68,69].

While existing digital health apps incorporate various digital components, the CAPS program is unique to its comparators due to combining 4 technological components (mobile app, wearable device, clinician portal, and SMS text messages developed for a self-management support package for people who experienced a stroke) with standardized patient-focused secondary prevention goal setting [30]. This promotes congruity in perception of care for patients and clinicians, better supporting lifestyle behavior changes and goal attainment. The technology facilitates both active and passive data collection, with remote access to the data collected and the ability to respond to alerts for measurements outside clinical guidelines. The secondary prevention goals developed with trained personnel support patients to control and reduce risk factors, assisted by SMS text messages developed to promote behavior change.

Conclusions

We have outlined the detailed co-design of the CAPS platform that aims to support secondary prevention of stroke via goal attainment and tailored mHealth strategies. Need, acceptability, and intention to use the digital program were supported by consumers and clinicians. Feasibility testing is currently in process. The password-protected app is available for download via the Android Play Store and the Apple Store.

Acknowledgments

The authors would like to acknowledge the staff from the Australian Stroke Clinical Registry (AuSCR) who coordinated and facilitated the invitation mailout to consumer participants, noting that the primary data collected from the Australian Stroke Data Tool are used to collect patient data for the AuSCR. The authors would also like to acknowledge the Stroke Association of Victoria and the support center coordinators who facilitated the distribution of surveys to people attending their centers. The authors would like to acknowledge the Australian Telehealth for Stroke Community of Practice who facilitated the promotion of the survey to clinicians registered with secondary prevention and global telehealth groups. This work was supported by the Commonwealth Scientific and Industrial Research Organisation and Monash University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Consumer survey.
[PDF File (Adobe PDF File), 71 KB - jmir_v26i1e54604_app1.pdf]

Multimedia Appendix 2
Clinician survey.
[PDF File (Adobe PDF File), 80 KB - jmir_v26i1e54604_app2.pdf]

Multimedia Appendix 3
Participants' rated need to access data more frequently than current practice (n=49).
[PDF File (Adobe PDF File), 64 KB - jmir_v26i1e54604_app3.pdf]

XSL•FO
RenderX

Multimedia Appendix 4
Total consumer participants' perception on the introduction of a new support program that used digital technologies for people who had experienced a stroke or transient ischemic attack (n=105).
[PNG File , 144 KB - jmir_v26i1e54604_app4.png]

Multimedia Appendix 5
Clinician participants' perception on the introduction of a new support program that used digital technologies for people who had experienced a stroke or transient ischemic attack (n=42).
[PNG File , 163 KB - jmir_v26i1e54604_app5.png]

References

1. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021 Oct;20(10):795-820 [FREE Full text] [doi: 10.1016/S1474-4422(21)00252-0] [Medline: 34487721]
2. Feigin VL, Krishnamurthi RV, Parmar P, Norrving B, Mensah GA, Bennett DA, et al. Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: the GBD 2013 study. Neuroepidemiology 2015 Oct 28;45(3):161-176 [FREE Full text] [doi: 10.1159/000441085] [Medline: 26505981]
3. Cadilhac DA, Dalli LL, Morrison J, Lester M, Paice K, Moss K, et al. The Australian stroke clinical registry annual report 2020. The Florey Institute of Neuroscience and Mental Health. 2021 Dec. URL: https://auscr.com.au/wp-content/uploads/2022/12/auscr-2020-annual-report_final_v2.pdf [accessed 2024-06-28]
4. Mohan KM, Wolfe CD, Rudd AG, Heuschmann PU, Kolominsky-Rabas PL, Grieve AP. Risk and cumulative risk of stroke recurrence: a systematic review and meta-analysis. Stroke 2011 May;42(5):1489-1494. [doi: 10.1161/STROKEAHA.110.602615] [Medline: 21454819]
5. Wu CM, McLaughlin K, Lorenzetti DL, Hill MD, Manns BJ, Ghali WA. Early risk of stroke after transient ischemic attack: a systematic review and meta-analysis. Arch Intern Med 2007 Dec 10;167(22):2417-2422. [doi: 10.1001/archinte.167.22.2417] [Medline: 18071162]
6. Hardie K, Hankey GJ, Jamrozik K, Broadhurst RJ, Anderson C. Ten-year risk of first recurrent stroke and disability after first-ever stroke in the Perth Community Stroke Study. Stroke 2004 Mar;35(3):731-735. [doi: 10.1161/01.str.0000116183.50167.d9]
7. Dhamoon MS, Sciacca RR, Rundek T, Sacco RL, Elkind MS. Recurrent stroke and cardiac risks after first ischemic stroke: the Northern Manhattan Study. Neurology 2006 Mar 14;66(5):641-646. [doi: 10.1212/01.wnl.0000201253.93811.f6] [Medline: 16534100]
8. Living clinical guidelines for stroke management. Stroke Foundation. URL: https://informme.org.au/guidelines/living-clinical-guidelines-for-stroke-management [accessed 2024-06-28]
9. Kleindorfer DO, Towfighi A, Chaturvedi S, Cockroft KM, Gutierrez J, Lombardi-Hill D, et al. 2021 Guideline for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline from the American Heart Association/American Stroke Association. Stroke 2021 Jul;52(7):e364-e467 [FREE Full text] [doi: 10.1161/STR.0000000000000375] [Medline: 34024117]
10. Boehme AK, Esenwa C, Elkind MS. Stroke risk factors, genetics, and prevention. Circ Res 2017 Feb 03;120(3):472-495 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.308398] [Medline: 28154098]
11. Liljehult J, Christensen T, Molsted S, Overgaard D, Mesot Liljehult M, Møller T. Effect and efficacy of lifestyle interventions as secondary prevention. Acta Neurol Scand 2020 Oct 15;142(4):299-313 [FREE Full text] [doi: 10.1111/ane.13308] [Medline: 32620044]
12. Parappilly BP, Field TS, Mortenson WB, Sakakibara BM, Eng JJ. Effectiveness of interventions involving nurses in secondary stroke prevention: a systematic review and meta-analysis. Eur J Cardiovasc Nurs 2018 Dec 01;17(8):728-736. [doi: 10.1177/1474515118779732] [Medline: 29856237]
13. Lennon O, Galvin R, Smith K, Doody C, Blake C. Lifestyle interventions for secondary disease prevention in stroke and transient ischaemic attack: a systematic review. Eur J Prev Cardiol 2014 Aug 11;21(8):1026-1039. [doi: 10.1177/2047487313481756] [Medline: 23478741]
14. Sakakibara BM, Kim AJ, Eng JJ. A systematic review and meta-analysis on self-management for improving risk factor control in stroke patients. Int J Behav Med 2017 Feb;24(1):42-53 [FREE Full text] [doi: 10.1007/s12529-016-9582-7] [Medline: 27469998]
15. Cadilhac DA, Dalli LL, Morrison J, Paice K, Carter H, Campbell B, et al. The Australian stroke clinical registry annual report 2022. The Florey Institute of Neuroscience and Mental Health. 2023 Dec. URL: https://auscr.com.au/wp-content/uploads/2024/03/australian-stroke-clinical-registry-annual-report-2022.pdf [accessed 2024-06-28]
16. Andrew NE, Kilkenny M, Naylor R, Purvis T, Lalor E, Moloczij N, et al. Understanding long-term unmet needs in Australian survivors of stroke. Int J Stroke 2014 Oct 01;9 Suppl A100(SA100):106-112. [doi: 10.1111/ijs.12325] [Medline: 25042019]

XSL•FO
RenderX

17. Lincoln NB, Brinkmann N, Cunningham S, Dejaeger E, De Weerdt W, Jenni W, et al. Anxiety and depression after stroke: a 5 year follow-up. Disabil Rehabil 2013 Jan 22;35(2):140-145. [doi: 10.3109/09638288.2012.691939] [Medline: 22725629]
18. Rao A, Zecchin R, Newton PJ, Phillips JL, DiGiacomo M, Denniss AR, et al. The prevalence and impact of depression and anxiety in cardiac rehabilitation: a longitudinal cohort study. Eur J Prev Cardiol 2020 Mar 09;27(5):478-489. [doi: 10.1177/2047487319871716] [Medline: 31597473]
19. Dalli LL, Kim J, Thrift AG, Andrew NE, Sanfilippo FM, Lopez D, et al. Patterns of use and discontinuation of secondary prevention medications after stroke. Neurology 2021 Jan 05;96(1):e30-e41. [doi: 10.1212/wnl.0000000000011083]
20. O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. The Lancet 2016 Aug;388(10046):761-775 [FREE Full text] [doi: 10.1016/s0140-6736(16)30506-2]
21. Gambhir SS, Ge TJ, Vermesh O, Spitler R. Toward achieving precision health. Sci Transl Med 2018 Feb 28;10(430):eaao3612 [FREE Full text] [doi: 10.1126/scitranslmed.aao3612] [Medline: 29491186]
22. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
23. Cavero-Redondo I, Saz-Lara A, Sequí-Dominguez I, Gómez-Guijarro MD, Ruiz-Grao MC, Martinez-Vizcaino V, et al. Comparative effect of eHealth interventions on hypertension management-related outcomes: a network meta-analysis. Int J Nurs Stud 2021 Dec;124:104085 [FREE Full text] [doi: 10.1016/j.ijnurstu.2021.104085] [Medline: 34601205]
24. Eberle C, Löhnert M, Stichling S. Effectiveness of disease-specific mHealth apps in patients with diabetes mellitus: scoping review. JMIR Mhealth Uhealth 2021 Feb 15;9(2):e23477 [FREE Full text] [doi: 10.2196/23477] [Medline: 33587045]
25. Su JJ, Yu DS, Paguio JT. Effect of eHealth cardiac rehabilitation on health outcomes of coronary heart disease patients: a systematic review and meta-analysis. J Adv Nurs 2020 Mar;76(3):754-772. [doi: 10.1111/jan.14272] [Medline: 31769527]
26. Varnfield M, Karunanithi M, Lee CK, Honeyman E, Arnold D, Ding H, et al. Smartphone-based home care model improved use of cardiac rehabilitation in postmyocardial infarction patients: results from a randomised controlled trial. Heart 2014 Nov;100(22):1770-1779 [FREE Full text] [doi: 10.1136/heartjnl-2014-305783] [Medline: 24973083]
27. Cardihab homepage. Cardihab Pty Ltd. URL: https://cardihab.com/ [accessed 2024-06-28]
28. Allan LP, Beilei L, Cameron J, Olaiya MT, Silvera-Tawil D, Adcock AK, et al. A scoping review of mHealth interventions for secondary prevention of stroke: implications for policy and practice. Stroke 2023 Nov;54(11):2935-2945. [doi: 10.1161/strokeaha.123.043794]
29. Hussain MS, Li J, Brindal E, van Kasteren Y, Varnfield M, Reeson A, et al. Supporting the delivery of total knee replacements care for both patients and their clinicians with a mobile app and web-based tool: randomized controlled trial protocol. JMIR Res Protoc 2017 Mar 01;6(3):e32 [FREE Full text] [doi: 10.2196/resprot.6498] [Medline: 28249832]
30. Cadilhac DA, Busingye D, Li JC, Andrew NE, Kilkenny MF, Thrift AG, et al. Development of an electronic health message system to support recovery after stroke: inspiring Virtual Enabled Resources following Vascular Events (iVERVE). Patient Prefer Adherence 2018 Jul;Volume 12:1213-1224. [doi: 10.2147/ppa.s154581]
31. Weickhardt C, Morrison J, Cadilhac D, Lannin NA, Dalli L, Kim J, et al. Australian stroke clinical registry public summary report 2020. The Florey Institute of Neuroscience and Mental Health. 2020 Dec. URL: https://auscr.com.au/wp-content/uploads/2021/12/2020-auscr-public-summary-report_final.pdf [accessed 2024-06-28]
32. Lloyd G, Dean CM, Ada L. Issues in recruiting community-dwelling stroke survivors to clinical trials: the AMBULATE trial. Contemp Clin Trials 2010 Jul;31(4):289-292. [doi: 10.1016/j.cct.2010.04.003] [Medline: 20435166]
33. Ferreira IS, Pinto CB, Saleh Velez FG, Leffa DT, Vulcano de Toledo Piza P, Fregni F. Recruitment challenges in stroke neurorecovery clinical trials. Contemp Clin Trials Commun 2019 Sep;15:100404 [FREE Full text] [doi: 10.1016/j.conctc.2019.100404] [Medline: 31360793]
34. Varnfield M, Redd C, Stoney RM, Higgins L, Scolari N, Warwick R, et al. M♡THer, an mHealth system to support women with gestational diabetes mellitus: feasibility and acceptability study. Diabetes Technol Ther 2021 May 01;23(5):358-366 [FREE Full text] [doi: 10.1089/dia.2020.0509] [Medline: 33210954]
35. Zhang X, Han X, Dang Y, Meng F, Guo X, Lin J. User acceptance of mobile health services from users' perspectives: the role of self-efficacy and response-efficacy in technology acceptance. Inform Health Soc Care 2017 Mar 26;42(2):194-206. [doi: 10.1080/17538157.2016.1200053] [Medline: 27564428]
36. Sezgin E, Özkan-Yildirim S, Yildirim S. Understanding the perception towards using mHealth applications in practice. Inf Dev 2016 Dec 19;34(2):182-200. [doi: 10.1177/0266666916684180]
37. Jenkins C, Burkett NS, Ovbiagele B, Mueller M, Patel S, Brunner-Jackson B, et al. Stroke patients and their attitudes toward mHealth monitoring to support blood pressure control and medication adherence. Mhealth 2016 May 03;2:24 [FREE Full text] [doi: 10.21037/mhealth.2016.05.04] [Medline: 27347490]
38. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]

XSL•FO
RenderX

39. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
40. Selvin S. A Biostatistics Toolbox for Data Analysis. New York, NY: Cambridge University Press; 2015.
41. Agnieszka. How are ranking survey question points calculated? Survicate Help Center. URL: https://help.survicate.com/en/articles/2694441-how-are-ranking-survey-question-points-calculated [accessed 2024-06-28]
42. Kondic M. Average rating calculator. CalCon Calculators. 2023 Mar 1. URL: https://calconcalculator.com/statistics/average-rating-calculator/ [accessed 2024-06-28]
43. Maxwell JA. Using numbers in qualitative research. Qual Inquiry 2010 Apr 15;16(6):475-482. [doi: 10.1177/1077800410364740]
44. Vicsek L. Issues in the analysis of focus groups: generalisability, quantifiability, treatment of context and quotations. Qual Report 2010;15(2):122-141. [doi: 10.46743/2160-3715/2010.1143]
45. Lee AC, Tang SW, Yu GK, Cheung RT. The smiley as a simple screening tool for depression after stroke: a preliminary study. Int J Nurs Stud 2008 Jul;45(7):1081-1089. [doi: 10.1016/j.ijnurstu.2007.05.008] [Medline: 17707824]
46. van Rijsbergen GD, Bockting CL, Berking M, Koeter MW, Schene AH. Can a one-item mood scale do the trick? Predicting relapse over 5.5-years in recurrent depression. PLoS One 2012;7(10):e46796 [FREE Full text] [doi: 10.1371/journal.pone.0046796] [Medline: 23056456]
47. Micklewright D, St Clair Gibson A, Gladwell V, Al Salman A. Development and validity of the rating-of-fatigue scale. Sports Med 2017 Nov 10;47(11):2375-2393 [FREE Full text] [doi: 10.1007/s40279-017-0711-5] [Medline: 28283993]
48. Koenig HG, Westlund RE, George LK, Hughes DC, Blazer DG, Hybels C. Abbreviating the Duke social support index for use in chronically ill elderly individuals. Psychosomatics 1993 Jan;34(1):61-69. [doi: 10.1016/s0033-3182(93)71928-3]
49. Sarason IG, Sarason BR, Shearin EN, Pierce GR. A brief measure of social support: practical and theoretical implications. J Soc Pers Relationsh 2016 Jun 30;4(4):497-510. [doi: 10.1177/0265407587044007]
50. Cohen S, Hoberman HM. Positive events and social supports as buffers of life change stress. J Appl Soc Psychol 2006 Jul 31;13(2):99-125. [doi: 10.1111/j.1559-1816.1983.tb02325.x]
51. Sansoni J, Marosszeky N, Sansoni E, Fleming G. Final report: effective assessment of social isolation. University of Wollongong. 2010. URL: https://ro.uow.edu.au/cgi/viewcontent.cgi?article=1472&context=ahsri [accessed 2024-06-28]
52. Bowman J, Mogensen L, Marsland E, Lannin N. The development, content validity and inter-rater reliability of the SMART-goal evaluation method: a standardised method for evaluating clinical goals. Aust Occup Ther J 2015 Dec 19;62(6):420-427. [doi: 10.1111/1440-1630.12218] [Medline: 26286379]
53. Turner-Stokes L. Goal attainment scaling (GAS) in rehabilitation: a practical guide. Clin Rehabil 2009 Apr;23(4):362-370. [doi: 10.1177/0269215508101742] [Medline: 19179355]
54. National stroke audit – acute services report 2023. Stroke Foundation. 2023. URL: https://informme.org.au/media/uppnqukx/2023-national-acute-services-national-report.pdf [accessed 2024-06-28]
55. Silvera-Tawil D, Hussain MS, Li J. Emerging technologies for precision health: an insight into sensing technologies for health and wellbeing. Smart Health 2020 Mar;15:100100. [doi: 10.1016/j.smhl.2019.100100]
56. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol 2005 Jun;44(Pt 2):227-239. [doi: 10.1348/014466505X29657] [Medline: 16004657]
57. Cadilhac DA, Hoffmann S, Kilkenny M, Lindley R, Lalor E, Osborne RH, et al. A phase II multicentered, single-blind, randomized, controlled trial of the stroke self-management program. Stroke 2011 Jun;42(6):1673-1679. [doi: 10.1161/STROKEAHA.110.601997] [Medline: 21493910]
58. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
59. Gaffey AE, Gathright EC, Fletcher LM, Goldstein CM. Screening for psychological distress and risk of cardiovascular disease and related mortality: a systematized review, meta-analysis, and case for prevention. J Cardiopulm Rehabil Prev 2022 Nov 01;42(6):404-415 [FREE Full text] [doi: 10.1097/HCR.0000000000000751] [Medline: 36342683]
60. Hackett ML, Pickles K. Part I: frequency of depression after stroke: an updated systematic review and meta-analysis of observational studies. Int J Stroke 2014 Dec 12;9(8):1017-1025. [doi: 10.1111/ijs.12357] [Medline: 25117911]
61. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-06-28]
62. Rintala A, Kossi O, Bonnechère B, Evers L, Printemps E, Feys P. Mobile health applications for improving physical function, physical activity, and quality of life in stroke survivors: a systematic review. Disabil Rehabil 2023 Dec 02;45(24):4001-4015. [doi: 10.1080/09638288.2022.2140844] [Medline: 36325613]
63. Paul L, Wyke S, Brewster S, Sattar N, Gill JM, Alexander G, et al. Increasing physical activity in stroke survivors using STARFISH, an interactive mobile phone application: a pilot study. Top Stroke Rehabil 2016 Jun;23(3):170-177 [FREE Full text] [doi: 10.1080/10749357.2015.1122266] [Medline: 27077973]

XSL•FO
RenderX

64. Sarfo FS, Treiber F, Gebregziabher M, Adamu S, Nichols M, Singh A, et al. Phone-based intervention for blood pressure control among Ghanaian stroke survivors: a pilot randomized controlled trial. Int J Stroke 2019 Aug 22;14(6):630-638. [doi: 10.1177/1747493018816423] [Medline: 30465630]
65. Yan LL, Gong E, Gu W, Turner EL, Gallis JA, Zhou Y, et al. Effectiveness of a primary care-based integrated mobile health intervention for stroke management in rural China (SINEMA): a cluster-randomized controlled trial. PLoS Med 2021 Apr 28;18(4):e1003582 [FREE Full text] [doi: 10.1371/journal.pmed.1003582] [Medline: 33909607]
66. Kim DY, Kwon H, Nam KW, Lee Y, Kwon HM, Chung YS. Remote management of poststroke patients with a smartphone-based management system integrated in clinical care: prospective, nonrandomized, interventional study. J Med Internet Res 2020 Feb 27;22(2):e15377 [FREE Full text] [doi: 10.2196/15377] [Medline: 32130140]
67. Wan LH, Zhang XP, You LM, Ruan HF, Chen SX. The efficacy of a comprehensive reminder system to improve health behaviors and blood pressure control in hypertensive ischemic stroke patients: a randomized controlled trial. J Cardiovasc Nurs 2018;33(6):509-517. [doi: 10.1097/JCN.0000000000000496] [Medline: 29901484]
68. Requena M, Montiel E, Baladas M, Muchada M, Boned S, López R, et al. Farmalarm: application for mobile devices improves risk factor control after stroke. Stroke 2019 Jul;50(7):1819-1824. [doi: 10.1161/strokeaha.118.024355]
69. Grau-Pellicer M, Lalanza JF, Jovell-Fernández E, Capdevila L. Impact of mHealth technology on adherence to healthy PA after stroke: a randomized study. Top Stroke Rehabil 2020 Jul 02;27(5):354-368. [doi: 10.1080/10749357.2019.1691816] [Medline: 31790639]

Abbreviations

AuSCR: Australian Stroke Clinical Registry
CAPS: Care Assistant and support Program for people after Stroke or transient ischemic attack
FAST: facial drooping, arm weakness, speech difficulties, and time to call emergency
mHealth: mobile health
REDCap: Research Electronic Data Capture
SMART: specific, measurable, achievable, relevant, and time-bound
TIA: transient ischemic attack

Edited by A Mavragani; submitted 15.11.23; peer-reviewed by J Carter, V Manea; comments to author 29.02.24; revised version received 23.04.24; accepted 14.06.24; published 22.08.24.

Please cite as:
Silvera-Tawil D, Cameron J, Li J, Varnfield M, Allan LP, Harris M, Lannin NA, Redd C, Cadilhac DA
Multicomponent Support Program for Secondary Prevention of Stroke Using Digital Health Technology: Co-Design Study With People Living With Stroke or Transient Ischemic Attack
J Med Internet Res 2024;26:e54604
URL: https://www.jmir.org/2024/1/e54604
doi:10.2196/54604
PMID:39172512

©David Silvera-Tawil, Jan Cameron, Jane Li, Marlien Varnfield, Liam P Allan, Mitch Harris, Natasha A Lannin, Christian Redd, Dominique A Cadilhac. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Establishing an Evaluation Indicator System for User Satisfaction With Hypertension Management Apps: Combining User-Generated Content and Analytic Hierarchy Process

Yunfan He[1], MS; Han Chen[2], MD, PhD; Peng Xiang[3,4], MD, MS; Min Zhao[5,6], MD, PhD; Yingjun Li[7], PhD; Yongcheng Liu[8], MS; Tong Wang[9,10,11], PhD; Jun Liang[1,3,7,12,13], MS; Jianbo Lei[14,15,16], MD, PhD

[1]Center for Health Policy Studies, School of Public Health, Zhejiang University, Hangzhou, China
[2]Department of Cardiology, Second Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China
[3]Department of AI and IT, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[4]Intelligent Medical Research Center, Zhejiang University Institute of Computer Innovation Technology, Hangzhou, China
[5]IT Center, The First Affiliated Hospital of Xiamen University, XiaMen, China
[6]Department of Gynecology, The First Affiliated Hospital of Xiamen University, XiaMen, China
[7]School of Public Health, Hangzhou Medical College, Hangzhou, China
[8]Netease Group, Hangzhou, China
[9]School of Health and Life Sciences, University of Health and Rehabilitation Sciences, Qingdao, China
[10]School of Basic Medical Sciences, Shandong University, Jinan, China
[11]Qingdao Hospital, University of Health and Rehabilitation Sciences (Qingdao Municipal Hospital), Qingdao, China
[12]National Key Laboratory of Transvascular Implantable Devices, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[13]School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, China
[14]Clinical Research Center, Affiliated Hospital of Southwest Medical University, Luzhou, China
[15]The First Affiliated Hospital, Hainan Medical University, Haikou, China
[16]Center for Medical Informatics, Health Science Center, Peking University, Beijing, China

Corresponding Author:
Jun Liang, MS
Department of AI and IT
Second Affiliated Hospital, School of Medicine
Zhejiang University
Number 88, Jiefang Road
Shangcheng District
Hangzhou, 310027
China
Phone: 86 571 87783942
Fax: 86 571 87783526
Email: junl@zju.edu.cn

Abstract

Background: Hypertension management apps (HMAs) can be effective in controlling blood pressure, but their actual impact is often suboptimal. Establishing a user satisfaction evaluation indicator system for HMAs can assist app developers in enhancing app design and functionality, while also helping users identify apps that best meet their needs. This approach aims to improve the overall effectiveness of app usage.

Objective: This study aims to systematically collect data on HMAs and their user reviews in the United States and China. It analyzes app usage patterns and functional characteristics, identifies factors influencing user satisfaction from existing research, and develops a satisfaction evaluation indicator system to provide more accurate recommendations for improving user satisfaction.

Methods: We conducted a descriptive statistical analysis to assess the development status of HMAs in both countries and applied the task-technology fit model to evaluate whether the app functionalities align with business needs. We separately summarized the factors influencing user satisfaction in both countries from previous research, utilized the analytic hierarchy process to develop

XSL•FO
RenderX

an evaluation indicator system for HMA user satisfaction, and calculated satisfaction levels. Based on these findings, we propose improvements to enhance app functionality and user satisfaction.

Results: In terms of current development status, there were fewer HMAs and user reviews in China compared with the United States. Regarding app functional availability, fewer than 5% (4/91) of the apps achieved a demand fulfillment rate exceeding 80% (8/10). Overall, user satisfaction in both countries was low.

Conclusions: In the United States, user satisfaction was lowest for advertising distribution, data synchronization, and reliability. By contrast, Chinese apps need improvements in cost efficiency and compatibility.

(J Med Internet Res 2024;26:e60773) doi:10.2196/60773

KEYWORDS

hypertension management; mobile health; user satisfaction; evaluation indicator system; analytic hierarchy process

Introduction

The global prevalence of hypertension is high, with low control rates. Although hypertension management apps (HMAs) have proven effective in controlled settings for helping patients manage blood pressure, user satisfaction and real-world effectiveness are still not well understood. The World Health Organization's "Global Report on Hypertension: The Race Against a Silent Killer" [1] indicates that the number of patients with hypertension worldwide increased from 648 million in 1990 to 1.278 billion in 2019, with a prevalence rate exceeding 33% and a control rate of only 21%. As a digital health tool [2-4], HMAs can enhance users' awareness of self-health management [5] and improve drug compliance [6,7], thereby aiding in blood pressure management [8,9]. Clinical trials have confirmed their effectiveness [10-12]. However, before 2015, the number of HMAs was relatively small, with fewer than 10 new apps introduced per year. At that time, users needed more robust core functions to effectively manage their health. Gaps in functional design led to poor user satisfaction [13] and limited effectiveness [14,15]. Over 80% of mobile health app users exhibited abandonment behavior [16], with 30% of mobile health apps being abandoned within a month [17], and only 6.6% of users continuing to use HMAs long-term [18].

Understanding the current status and functional characteristics of HMAs and developing a user satisfaction evaluation indicator system (USEIS) can significantly improve HMA design, leading to enhanced user satisfaction and effectiveness. By analyzing user demand and app functionality, developers can identify unmet needs and refine app features accordingly. Comparing USEISs enables targeted improvements on factors with high importance but low user satisfaction. By understanding the product characteristics and overall user satisfaction of various HMAs through the evaluation indicator system, users can select the app that best meets their needs.

Both the United States and China, which face the most severe cases of hypertension worldwide [19,20], have launched numerous HMAs [21,22]. This indicates strong market prospects for HMAs in both countries. Comparing and analyzing the product features and user satisfaction of HMAs are crucial for optimizing their functionality and performance in both markets. Analyzing app demand and functionality in both countries and constructing a USEIS based on previous research [23] are crucial for accurately evaluating and improving user satisfaction.

Current research on HMA user satisfaction can be broadly divided into offline user research and online user data analysis. Most studies in this area concentrate on user satisfaction surveys and the factors influencing them. Offline user research typically involves methods such as interviews [24], questionnaire surveys [25-27], literature reviews [28], and case analyses [29]. For instance, studies such as those by Ramirez et al [30] and Breil et al [31] provide insights into user experiences through these traditional research approaches. Online user data analysis uses techniques such as manual coding [32,33] and text mining [34,35] to extract information about HMAs' main functions from app descriptions and user reviews and to investigate factors influencing user satisfaction (eg, Alnooh et al [36] and He et al [23]). Additionally, Haggag et al [37] utilized user-generated content to preliminarily explore user satisfaction and influencing factors of mobile health apps. Desmal et al [38] and Taylor et al [39] established satisfaction evaluation indicator systems for mobile government portal websites and consumer health information websites through literature reviews and assessed satisfaction through case analyses. Meanwhile, other scholars have developed user satisfaction indicator evaluation systems for products and services across various fields using expert Delphi [40,41] and analytic hierarchy process (AHP) [42,43], a decision analysis method for solving complex decision problems. The process of users posting ratings and reviews is a direct expression of their satisfaction with the app's functionality and services. This user-generated content can reveal the main attributes and genuine emotional tendencies that users focus on while using the app in real time. By analyzing this content, it is possible to gain an accurate understanding of users' health management needs and develop more comprehensive app features to better support users in managing their own health.

However, existing studies have not sufficiently analyzed HMA demands and functions nor established a USEIS based on influencing factors. These studies have primarily summarized HMA functions through manual coding or content analysis [44-46] and have categorized the distribution of app functions. However, they have not assessed the apps' functional availability—the extent to which their functionality meets business needs, specifically the self-management demands for hypertension as outlined in clinical guidelines [22]. This field lacks quantitative research driven by real-world big data [47,48] and evaluation research on app user satisfaction [49-51]. Although we previously used text mining to extract the main

XSL•FO
RenderX

factors influencing HMA user satisfaction in the United States and China [23], we did not explore the importance of these factors or develop a USEIS.

Therefore, this study comprises 2 main parts: functional availability analysis and user satisfaction analysis of HMAs in China and the United States. First, we summarize the main functions and business demands of HMAs in both countries. Using the theory of task-technology fit (TTF; an information technology evaluation method based on task and technology matching), we calculate the compatibility between functions and demands to analyze functional availability and explore how well these HMAs meet user needs. Functional availability analysis helps understand the development status of HMAs and quickly identifies functions that have not yet met user needs. Second, based on the influencing factors of user satisfaction identified in previous studies [23], we determine the specific indicators and their weights affecting user satisfaction in the United States and China using the AHP. User satisfaction analysis helps identify the core elements affecting HMA usage and assists developers in crafting precise optimization strategies. For research significance, this study not only offers a suitable method and process for analyzing functional availability and user satisfaction of mobile health apps but also uncovers the differences in functional availability and user satisfaction between Chinese and American HMAs. It provides suggestions for enhancing the quality of HMAs.

Methods

Theoretical Basis and Methodological Process

This article, based on TTF theory and decision theory, uses descriptive statistics, the TTF model, and AHP to evaluate the functional availability of HMAs in China and the United States. It extracts factors affecting user satisfaction from user reviews and establishes a USEIS. The study compares the differences in results between the 2 countries. The methodology is divided into 4 main parts: analysis of app usage status and functional availability, induction of factors affecting user satisfaction, establishment of a USEIS, and suggestions for app improvement. The detailed method flow is illustrated in Figure 1.

Figure 1. Method flowchart. AHP: analytic hierarchy process; HMA: hypertension management app; TTF: task-technology fit.

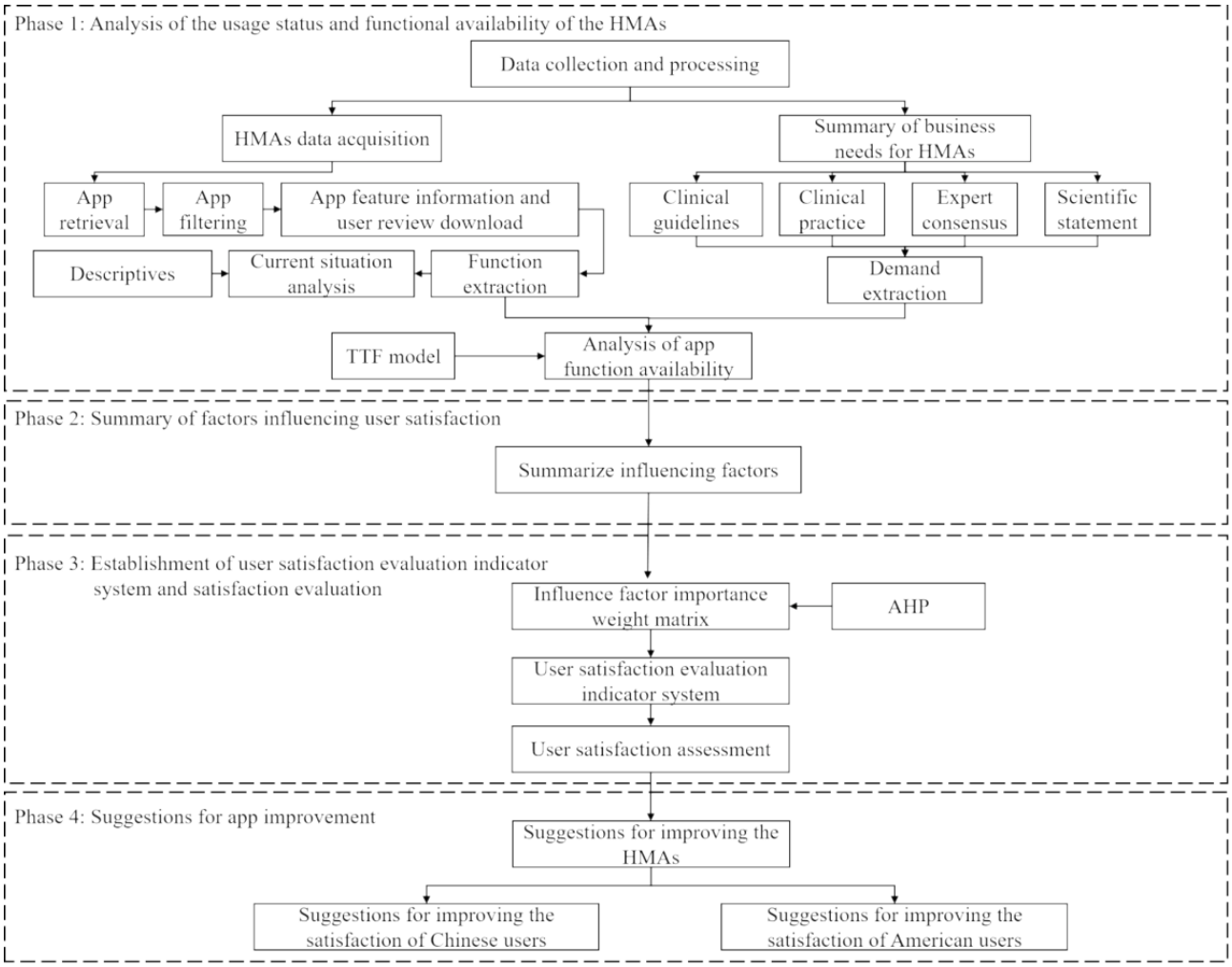

Data Acquisition, Filtering, and Preprocessing

We used HMA user reviews from the United States and China, obtained in a previous study [23]. This includes a brief description of the process for obtaining and screening HMAs, as well as preprocessing user reviews from both countries. In April 2023, we retrieved 5016 related apps using 2 English keywords across 2 American app stores and 10 Chinese keywords across 8 Chinese app stores. Following deduplication, 3591 apps remained (for detailed search terms and results, see

XSL•FO
RenderX

Table S1 in Multimedia Appendix 1). Two researchers (YFH and JL) independently conducted initial screenings, with a third researcher (HC) arbitrating to identify HMAs, resulting in 261 HMAs with user reviews. Of these, 170 apps were exclusive to app stores in the United States, 41 to app stores in China, and 50 apps were available in both (for detailed inclusion and exclusion criteria and screening flowcharts, see Table S2 in Multimedia Appendix 1 and Figure 2, respectively (also see Figure S1 in Multimedia Appendix 1). To gather and preprocess app user reviews, we used the Qimai Mobile Application Data Analysis Platform [52] and self-developed Python (Python Foundation) scripts, accumulating a total of 295,927 user reviews and rating data from both the United States and China, including 250,193 US reviews and 45,734 Chinese reviews. To ensure the authenticity and reliability of these reviews, we performed data preprocessing. This involved removing reviews from robot accounts (identified and deleted using the tweetbotnot package [53]), duplicates, invalid reviews (including garbled characters, emoticons, non-Chinese or non-English content, and meaningless reviews such as advertisements), reviews with mismatched emotional polarity and ratings [54], and blank reviews. Ultimately, this study analyzed 116,686 US and Chinese user reviews to develop a USEIS (for the number of apps and user reviews, see Table S3 in Multimedia Appendix 1; for details on data acquisition, screening, and preprocessing processes, see the "Obtaining, Screening, and Preprocessing of Hypertension Management Apps and Their User Reviews" section in Multimedia Appendix 1).

Figure 2. Flowchart of the hypertension management app screening process.

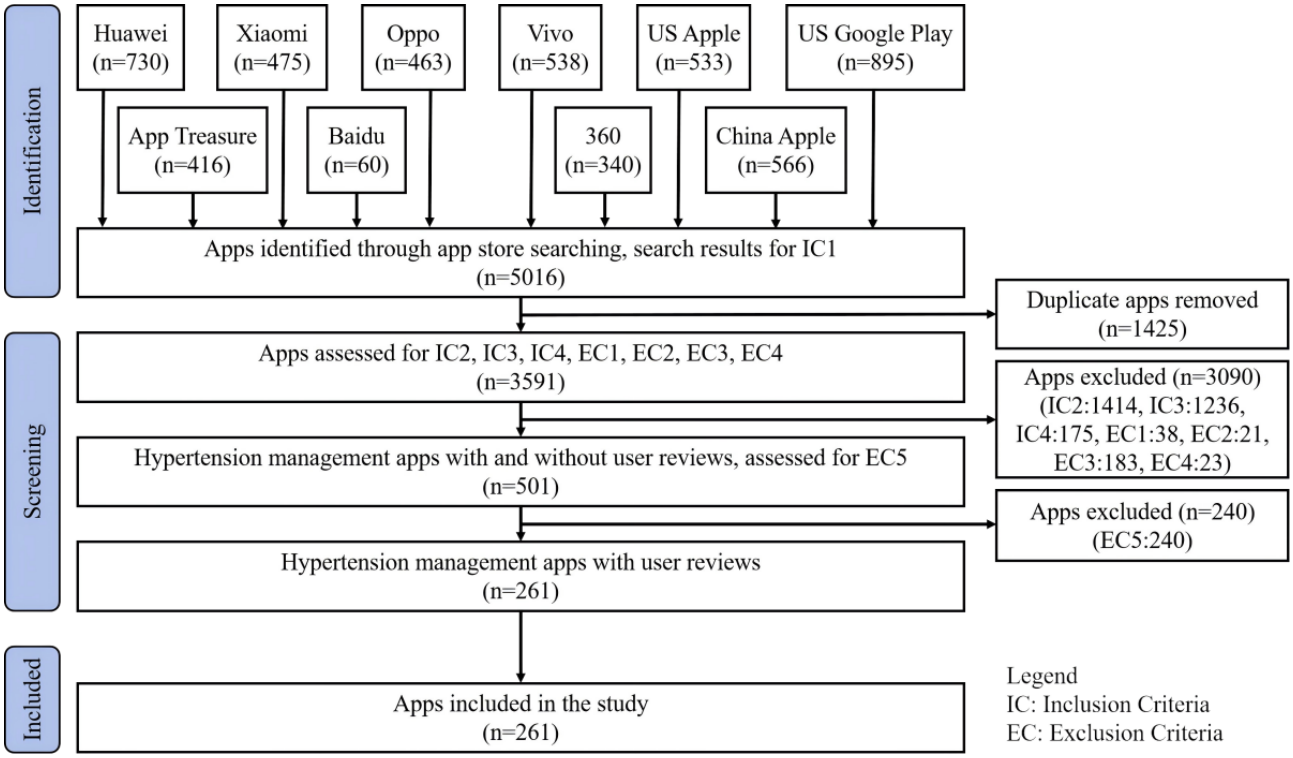

Extraction and Validation of App Product Features

App Product Information Extraction

To study the current trends, usage, and technological attributes of HMAs in the United States and China, this study builds upon previous research [46,55] to develop a comprehensive app information extraction framework. This framework extracts the product features of each HMA, and we used descriptive statistical methods to analyze the distribution of these features. The framework encompasses 5 dimensions: basic information, listing status, user usage, development standardization, and technical completeness. A detailed description of the framework is presented in Table S4 in Multimedia Appendix 1. To ensure consistency in information extraction, 2 researchers (YFH and JL) extracted all product features of the 261 HMAs, with any inconsistencies resolved by a third researcher (HC). Additionally, to assess the consistency of the data extraction between the 2 researchers, we randomly selected 1 field from the 20 extracted fields and calculated the consistency. The kappa consistency coefficient was 0.89, indicating a high level of agreement between the researchers.

App Feature Information Extraction

This study, building upon previous research [22], involved manually extracting and summarizing the functional characteristics of HMAs in the United States and China based on app store descriptions using the induction method. First, comprehensive function descriptions for 261 HMAs—including blood pressure measurement, blood pressure recording, and doctor-patient communication—were manually extracted from app store descriptions, software interface screenshots, and function module introductions provided in the app stores. These descriptions were then consolidated, standardizing various terminologies for identical functions and categorizing them into distinct functional dimensions based on similarity. This process resulted in a standardized functional summary table (Table S5 in Multimedia Appendix 1), which helped ensure consistency

XSL•FO
RenderX

in functional descriptions across all apps. Finally, we separately summarized the specific functions of HMAs in both countries.

Analysis of App Function Availability Based on the Task-Technology Fit Model

Overview

The availability of app functionality is a crucial metric for evaluating its completeness, specifically assessing whether the app provides sufficient features to meet users' needs. This concept pertains to how well an app's existing functions address business demands [22]. The TTF model is an information technology evaluation model that quantifies how well information technology can perform specific tasks for users by assessing the compatibility between information technology functions and task demands. The model suggests that if the alignment between technical functions and task demands is high, technology can enhance task performance. This study drew inspiration from the TTF concept in the TTF model to evaluate the functional availability of HMAs by analyzing the alignment between their functions and the needs of hypertension management tasks. It assessed how well these apps promote users' ability to complete hypertension management tasks. The study systematically reviewed the main functions of HMAs in the United States and China, using the TTF model to calculate the adaptation of specific functions and business demands in each country separately, to evaluate their functional availability. Additionally, the study aimed to understand the apps' effectiveness in helping users in both countries manage their hypertension.

Summary of Business Demands

First, we compiled and reviewed the latest clinical guidance documents on hypertension management and diagnosis from the official websites of hypertension institutions in both countries, as well as from international institutions and English and Chinese literature databases. This review was supplemented by consultations with expert clinicians. Finally, a clinical expert in the cardiovascular field systematically reviewed the clinical guidance documents for hypertension in the United States and China and summarized the business needs of users in both countries for independently managing hypertension. For details on the process of summarizing app business demands, see the "Summary of App Business Requirements" section in Multimedia Appendix 1.

Functional Availability Analysis of the Apps

For overall app functionality availability, we will match all the functions of HMAs in each country with the business needs for hypertension self-management specific to that country. This will help identify any business needs that lack functional support, highlighting areas where the apps have not yet provided adequate functionality. For individual app functional availability, we matched the specific functions of each HMA in each country with the business needs for hypertension management in that country. Based on these matching results, we created function demand adaptation maps for HMAs in the United States and China. We then calculated the average functional utilization rate (see Equation 1) and demand satisfaction rate (see Equation 2) for apps in both countries, analyzing the accuracy and completeness of the app's functional design. The app functional utilization rate refers to the extent to which app functions meet business needs. The app demand satisfaction rate indicates the proportion of hypertension management needs that have been met. Based on the matching results of each app's functions and demands, we summarized the main unmet needs for HMAs in both countries. Finally, we compared the functional differences between HMAs in the United States and China, analyzing the specific functions and functional availability in each country.

App functional utilization rate = (Number of functions that meet the demands)/(Total number of app functions) **(1)**

App demand satisfaction rate = (The number of demands that the app has met)/(Total number of demands) **(2)**

Constructing a User Satisfaction Evaluation Indicator System

System Design

Based on the factors influencing user satisfaction with HMAs in the United States and China identified in a previous study [23], we further used AHP to construct a USEIS. AHP is a decision-making method that establishes a hierarchical structure to convert subjective judgments into pairwise comparisons of the importance of various factors. We innovated the construction of the weight matrix by quantifying the proportion of user reviews to enhance objectivity and reliability in the decision results. Additionally, we referred to multiple relevant studies [56-59] to ensure the reliability of the evaluation indicator system construction process.

Building an Evaluation Indicator System Framework

Based on the AHP decision path model of "target layer-criterion layer-indicator layer," this study used significant influencing factors from previous research [23] to summarize the evaluation indicator system framework for HMA user satisfaction in the United States and China. First, a framework for evaluating user satisfaction in both countries was constructed according to the decision-path model. Second, factors significantly influencing satisfaction among Chinese and American HMA users were used as specific indicators in the indicator layer of the framework. These indicators were then classified into criteria based on attribute similarities. The indicators, indicator weights, and criterion weights for China and the United States are independent and distinct. These dimensions were then used as specific criteria for the criterion layer. Finally, user satisfaction was defined as the specific goal of the target layer.

Determining Indicator Weights

This study followed the analytical process of AHP [60] to determine the weights of each layer of indicators in the 2 evaluation indicator system frameworks. First, using the clustering results from the latent Dirichlet allocation topic model in the previous study [23], the frequency proportion of corresponding reviews for each topic from the United States and China was used as the basis for each indicator. The frequency proportions of each criterion were calculated by summing the frequency proportions of the indicators within the same criterion layer. Next, the importance scale correspondence table was created based on the maximum proportion difference

XSL•FO
RenderX

between different indicators and criteria within the same evaluation indicator system, and the scale of levels 1-9 in the indicator and criterion judgment matrix. Hierarchical single sorting and consistency tests were then performed. Based on the importance scale table and the proportional differences between various criteria and indicators, a corresponding judgment matrix was constructed. This matrix was used to calculate the eigenvalues and eigenvectors through hierarchical single sorting. Eigenvectors corresponding to the maximum eigenvalue were identified, and a consistency coefficient verification was performed. A consistency coefficient of ≤0.1 indicates that the feature matrix is consistent and that the normalized maximum eigenvector can be used as the weight for the corresponding indicator [61]. Finally, hierarchical total ranking and consistency testing were conducted. Using the standardized feature vectors that passed consistency testing as the importance weights for the corresponding indicator and criterion layers, we calculated the comprehensive weights of each indicator in the indicator layer relative to the target layer through a linear weighted sum. Combined consistency testing and overall consistency testing were then performed. If both the combined and overall consistency coefficients are less than or equal to 0.1, consistency is present [61]. The entire indicator weight calculation was performed using a self-developed Python (Python Foundation) script.

User Satisfaction Assessment

We calculated the overall user satisfaction and satisfaction with each evaluation indicator for HMAs in the United States and China based on the 2 established evaluation indicator systems (which were constructed based on user reviews in this study). First, this study is based on both positive (satisfied) and negative (unsatisfied) deviations in user reviews from the 2 countries, as identified in the previous study [23]. The proportion of user reviews with a satisfied attitude toward a particular topic was used as the initial evaluation value of user satisfaction for that topic. Then, based on the indicator weights from the USEISs in the United States and China and the satisfaction levels for each indicator, we calculated the satisfaction for each criterion and the overall user satisfaction using a linear weighted sum method.

Ethical Considerations

All data in this study are sourced from internet mobile app stores, and all app information and user reviews can be publicly accessed. Therefore, this study does not involve any medical ethics issues.

Results

Development and Use Status of the App

Trend of Changes in the Number of Apps and the Frequency of Version Updates

For app development and listing, the 261 HMAs were first introduced in 2010. From 2010 to 2023, the annual number of listed HMAs generally exhibited a stepwise upward trend, with an average annual growth rate of 51.48% (6.692/13). The number peaked in 2021 and has decreased slightly since then. The number of HMAs listed in Chinese app stores each year was relatively small, often not exceeding 10, while the number listed in American app stores showed a fluctuating upward trend, significantly surpassing that of Chinese app stores (Figure 3). It is important to note that because the same app may be listed in both US and Chinese app stores, the total number of HMAs listed each year does not equal the sum of the number of apps listed in each store.

XSL•FO
RenderX

Figure 3. The trend in the number of hypertension management apps.

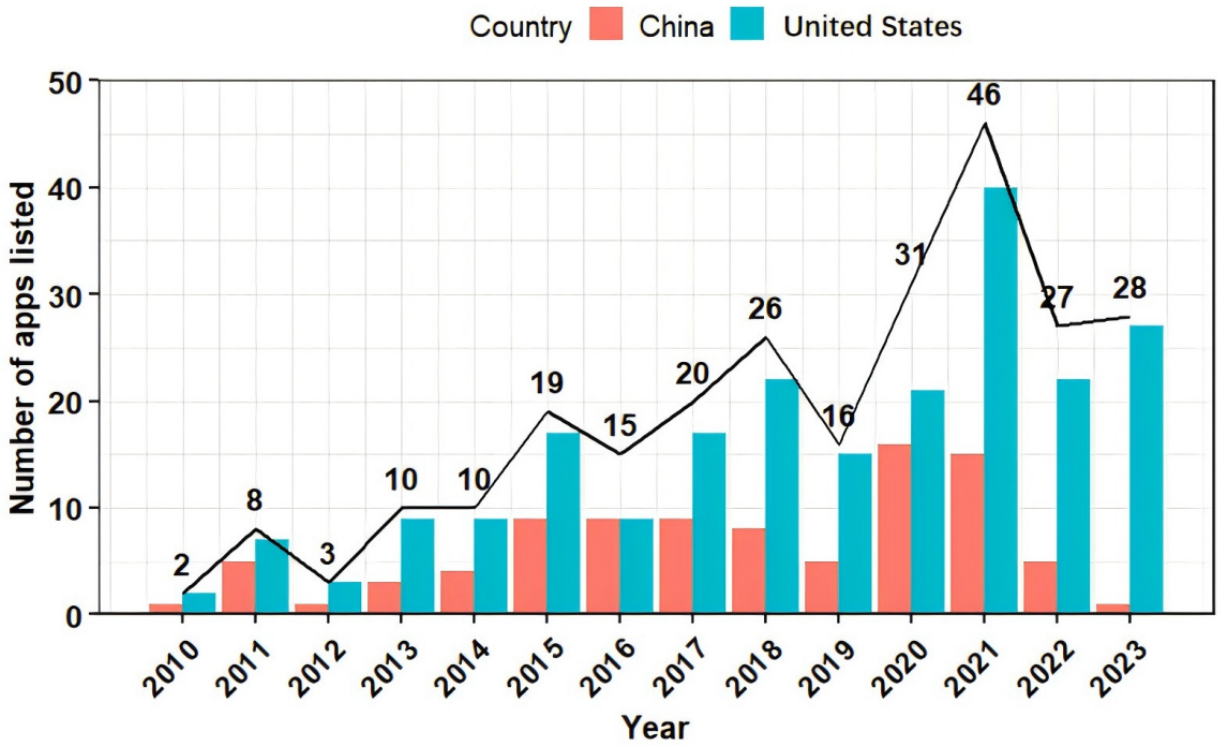

For app version updates, the 261 HMAs underwent a total of 5508 version iterations, averaging 21 updates per app. Each app was updated approximately 6 times a year. Among these, apps listed in Chinese app stores were updated 2719 times (an average of 29 updates per app), with each app being updated about 7 times a year. By contrast, apps in US app stores were updated 4870 times (an average of 22 updates per app), also with an average of 7 updates per year. The annual update frequency of apps is illustrated in Figure 4.

Figure 4. The proportion of annual average update frequency of hypertension management apps.

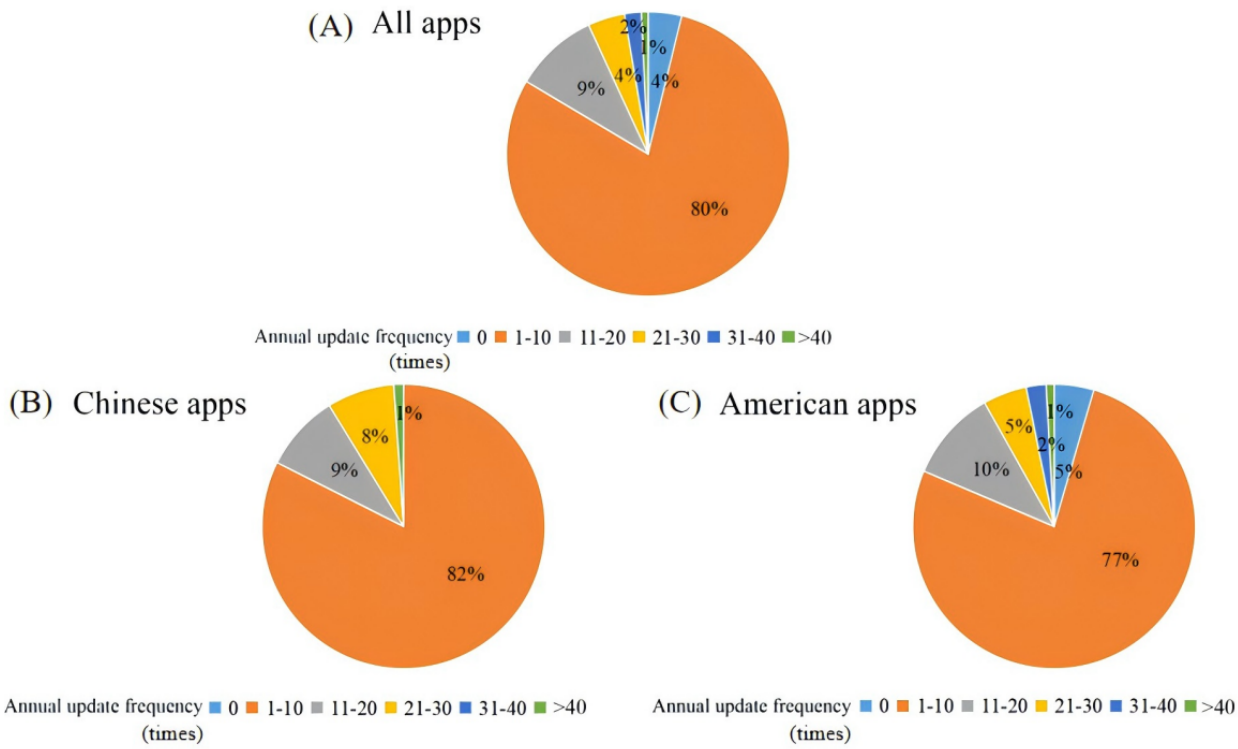

XSL•FO
RenderX

Product Feature Distribution of Apps

Standardization and technical perfection are the 2 main external product characteristics of HMAs. Detailed external feature information for the 261 HMAs and the results summary are presented in Table S6 in Multimedia Appendix 1. Regarding app development standardization, 237 apps (90.8%) did not adhere to clinical guidelines for hypertension, 246 apps (94.3%) did not use specific diagnostic and treatment methods for blood pressure management, 134 apps (51.3%) did not provide disclaimers, and 238 apps (91.2%) disclosed their data privacy policies to users. Regarding the apps' technical completeness, 88 apps (33.7%) used artificial intelligence technology, with 86 (33.0%) using rule-based algorithms. Additionally, 105 apps (40.2%) required connections to specific data acquisition devices, with blood pressure monitors (n=79, 30.3%) and wearable devices (n=26, 10.0%) being the most commonly used. Moreover, 120 apps (46.0%) supported family blood pressure monitoring and data sharing, while only 1 app (0.4%) was classified as digital therapy software [23]. Regarding differences in product features between apps in the United States and China, most HMAs listed in Chinese app stores required specific external devices, and there was a greater emphasis on supporting family blood pressure monitoring and data sharing (Figure 5).

Figure 5. Product feature distribution of hypertension management apps in the US and China.

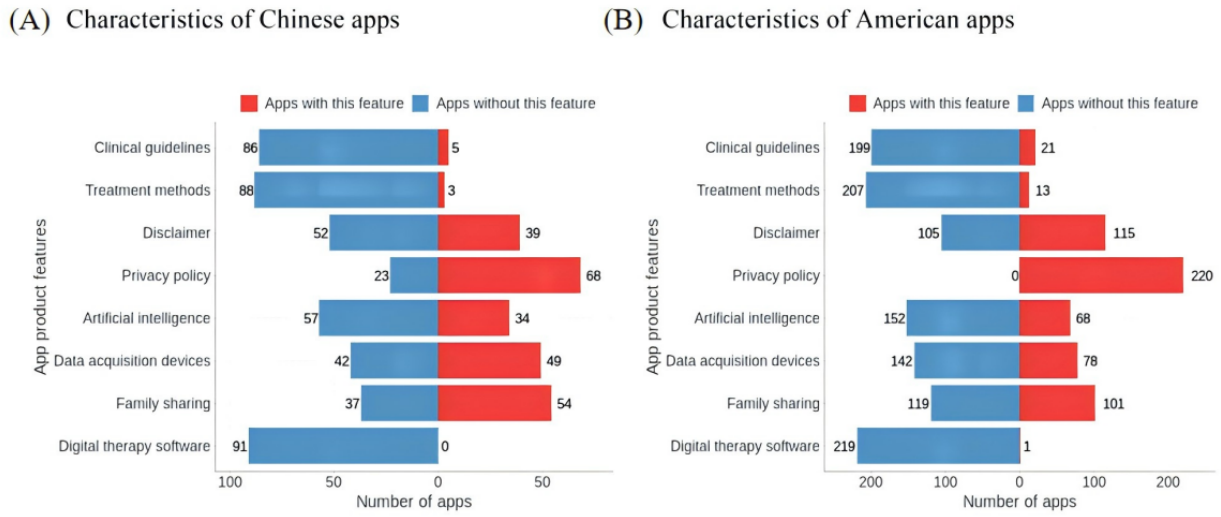

App Downloads and User Ratings

All HMAs were downloaded a total of 171,074,013 times from major app stores. The most downloaded app was the Blood Pressure App developed by QR Code Scanner, with over 50 million downloads. By contrast, the least downloaded app had only a few hundred downloads. On average, each app was downloaded 655,455 times. Most apps had fewer than 10,000 downloads (n=85, 32.6%) or between 10,000 and 100,000 downloads (n=74, 28.4%). For apps listed in Chinese app stores, there were 24,204,832 downloads, with an average of 265,987 downloads per app. For apps listed in US app stores, the total number of downloads was 148,869,181, averaging 676,678 downloads per app. There was no significant difference in the number of downloads between apps listed in the US and Chinese app stores (t_{309}=–1.044, P=.30).

Further, all HMAs received a total of 1,721,889 ratings. Apps listed in Chinese app stores received 104,986 ratings, while those listed in American app stores received 1,616,903 ratings. Regarding user review ratings, 4- and 5-star ratings accounted for 75.22% (87,773/116,686) of all ratings, with Chinese apps receiving 83.72% (13,866/16,561) 4- and 5-star ratings and American apps receiving 73.81% (73,907/100,125).

Functional Availability Evaluation of Hypertension Management Apps

This study systematically summarized clinical guidance documents on hypertension from the United States, China, and relevant international organizations, identifying a total of 26 relevant documents: 9 from the United States, 14 from China, 2 from the World Health Organization, and 1 from the International Society of Hypertension (see Table S7 in Multimedia Appendix 1). We summarized the specific functions of 261 HMAs and identified 9 functional feature dimensions and 50 specific functions (see the "Functional Availability of Hypertension Management Apps" section in Multimedia Appendix 1 for detailed results, Table S8 in Multimedia Appendix 1 for business needs, Table S9 in Multimedia Appendix 1 for functions, and Figure S2 in Multimedia Appendix 1 for functional characteristics). For overall app availability, there are 10 dimensions of business needs for hypertension management in China. The functions of the Chinese HMAs can meet 9 of these demand dimensions proposed in the Chinese Hypertension Clinical Guidance Document, with the exception of the measurement/treatment effect evaluation demand dimension, which remains unmet. The feature functions provided by HMAs meet 40 specific needs, with 15 specific needs remaining unmet. These unmet needs include adverse reaction monitoring, health behavior intervention, and follow-up treatment, such as recording adverse reactions, smoking cessation intervention, drug intervention,

XSL•FO
RenderX

and follow-up for patients with hypertension. The business requirements for hypertension management in the United States encompass 12 dimensions. HMA functions in the United States can meet the 9 demand dimensions outlined in the clinical guidelines for hypertension in the United States. However, 3 demand dimensions—measurement/treatment efficacy evaluation, treatment plan reliability, and software performance—were not addressed. The feature functions effectively meet 39 specific needs, but 12 specific needs remain unmet. These specific unmet needs include personalized treatment for patients, classification of hypertension causes, and evaluation of treatment effects. Examples of these unmet needs are setting blood pressure control goals tailored to different patients, classifying hypertension phenotypes, and evaluating the effects of antihypertensive treatments.

For single-app functionality, the average functional utilization rate of Chinese HMAs was 100% (91/91), but the average demand satisfaction rate was only 55.9%. Additionally, 37% (34/91) of Chinese apps had a demand satisfaction rate between 41% and 60%, while only 4 apps (4%, 4/91) had a demand satisfaction rate exceeding 80% (Figure 6A). Summarizing the satisfaction of each demand dimension in Chinese apps, the vast majority were able to meet the demand dimensions for health indicator recording (91/91, 100%), health indicator trend tracking (68/91, 75%), and health indicator measurement (62/91, 68%). However, the dimensions of measurement/treatment effect evaluation (0/91, 0%), health support (38/91, 42%), and intelligent diagnosis of hypertension (43/91, 47%) were not effectively met (Figure 6B). The average functional utilization rate of HMAs in the United States was 100% (220/220), but the average demand satisfaction rate was only 39.47% (86.83/220). Additionally, 45.4% (100/220) of American apps had a demand satisfaction rate of 41%-60%, and 15% (33/220) had a rate of 0%-20%. No American app had a demand satisfaction rate exceeding 80% (Figure 6C). Regarding the satisfaction status of each demand dimension in American apps, most American apps meet the demand dimensions of health indicator recording (206/220, 93.6%), health indicator trend tracking (172/220, 78.2%), and health intervention measures (123/220, 55.9%). However, the dimensions of measurement/treatment effect evaluation, treatment plan reliability, and software performance were not met (Figure 6D).

Figure 6. Distribution of demand satisfaction rate and features-demand adaptation of hypertension management apps in the US and China.

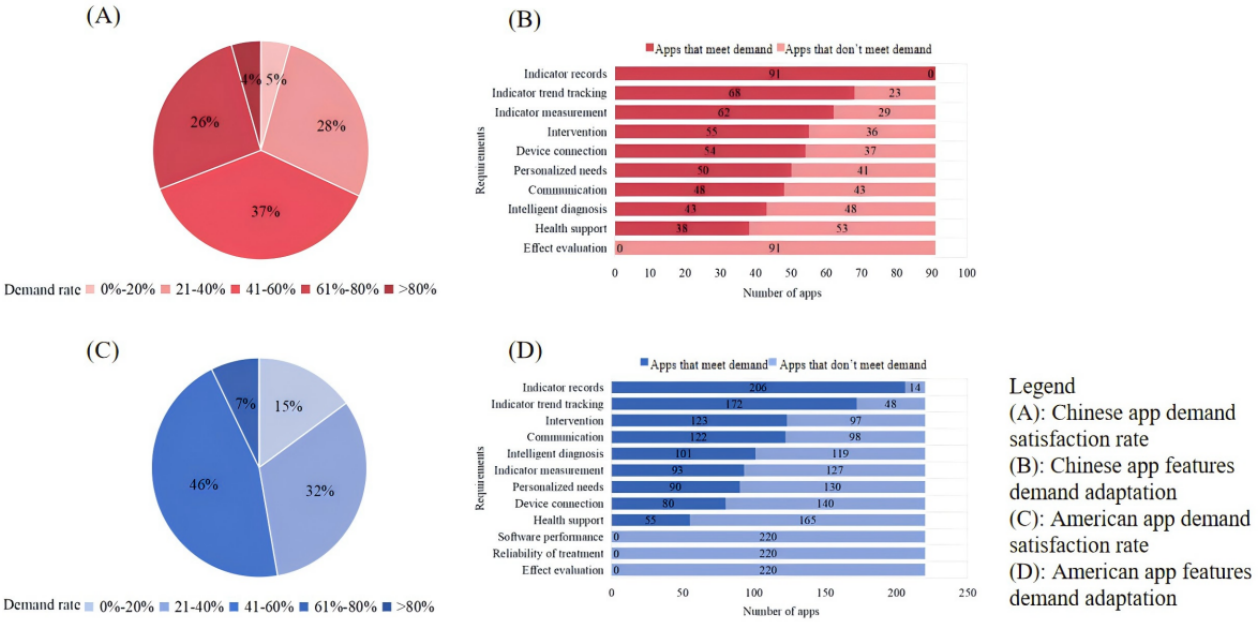

User Satisfaction Evaluation Indicator System for Hypertension Management Apps

In the previous study [23], 10 out of 12 significant factors were found to affect user satisfaction with HMAs in both China and the United States. Based on these factors, this study developed a framework for evaluating user satisfaction in both countries (see Figure S3 in Multimedia Appendix 1). Using the proportions of themes corresponding to various influencing factors in user reviews from the United States and China (see Table S10 in Multimedia Appendix 1), we applied AHP to calculate the weights of each indicator in the user satisfaction framework for both countries. The results are detailed in Tables S11-S16 in Multimedia Appendix 1. Using the indicator system framework and calculated weights, we constructed the USEISs for HMAs in the United States and China (Figure 7). In the US evaluation indicator system, the top 3 comprehensive weighted indicators were convenience (41.99%), blood pressure tracking (14.95%), and data synchronization (7.91%). In the Chinese USEIS, the top 3 indicators were convenience (52.46%), reliability (9.78%), and measurement accuracy (9.00%). The process of developing the framework, calculating indicator weights, and performing consistency checks using the AHP for HMA user USEISs in the United States and China is detailed in the "User Satisfaction Evaluation Indicator System for Hypertension Management App" section in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 7. User satisfaction evaluation indicator systems in the United States and China.

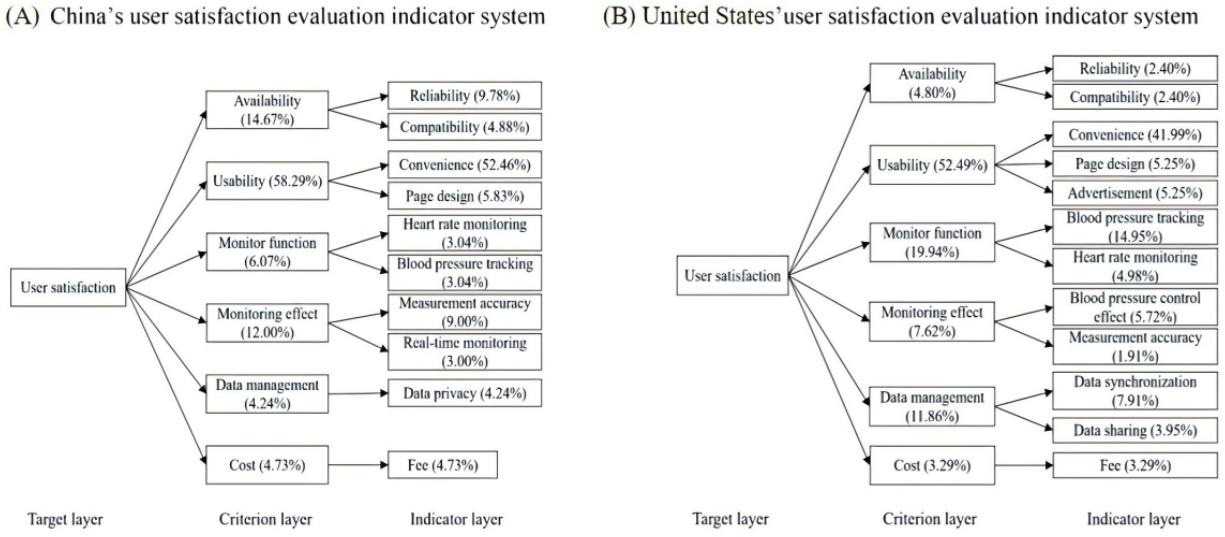

User Satisfaction Evaluation Results of Hypertension Management Apps

This study used the proportion of positive reviews corresponding to each indicator as the initial satisfaction value and assessed overall user satisfaction and specific satisfaction with each criterion for both American and Chinese users, based on the importance weights in the 2 USEISs (see Figures S4-S7 in Multimedia Appendix 1). The evaluation results are presented in Tables 1 and 2. Among them, Chinese users had relatively low overall satisfaction with HMAs (evaluation value = 0.7891 out of 1). At the criteria layer, users expressed higher satisfaction with software usability (0.8721) and software monitoring functions (0.8243), while their satisfaction with cost (0.4303) and software monitoring effectiveness (0.6305) was relatively low. In the indicator layer, American users were relatively satisfied with convenience (0.8909) and heart rate monitoring (0.8358), while their satisfaction with cost (0.4303) and compatibility (0.4419) was relatively low. The overall satisfaction level of American users with HMAs was lower than that of Chinese users (0.6861 vs 0.7891). At the criteria layer, users had higher satisfaction with software monitoring effectiveness (0.8457) and cost (0.8093), but their satisfaction with data management (0.3163) and software availability (0.3647) was relatively low. In the indicator layer, users were relatively satisfied with the effectiveness of blood pressure management (0.8701), page design (0.8552), blood pressure tracking (0.8347), and cost (0.8093). However, their satisfaction with advertising distribution (0.2058), data synchronization (0.2145), and reliability (0.2977) was relatively low.

XSL•FO
RenderX

Table 1. Comprehensive satisfaction evaluation results of users of the US hypertension management apps.

Criterion layer (criteria satisfaction) and indicator layer[a,b]	Indicator satisfaction
Availability (0.3647)	
Reliability	0.2977
Compatibility	0.4316
Usability (0.7353)	
Convenience	0.7865
Page design	0.8552
Advertising distribution	0.2058
Monitoring function (0.7726)	
Blood pressure tracking	0.8347
Heart rate monitoring	0.5863
Monitoring effect (0.8457)	
Effect of blood pressure management	0.8701
Measurement accuracy	0.7723
Data management (0.3163)	
Data synchronization	0.2145
Data sharing	0.5200
Cost (0.8093)	
Fee	0.8093

[a]Target layer: user satisfaction.
[b]Overall satisfaction: 0.6861.

Table 2. Comprehensive satisfaction evaluation results of users of the Chinese hypertension management apps.

Criterion layer (criteria satisfaction) and indicator layer[a,b]	Indicator satisfaction
Availability (0.7044)	
Reliability	0.8356
Compatibility	0.4419
Usability (0.8721)	
Convenience	0.8909
Page design	0.7028
Monitoring function (0.8243)	
Heart rate monitoring	0.8358
Blood pressure tracking	0.8127
Monitoring effect (0.6305)	
Measurement accuracy	0.6188
Real-time monitoring	0.6656
Data management (0.7396)	
Data privacy	0.7396
Cost (0.4303)	
Fee	0.4303

[a]Target layer: user satisfaction.
[b]Overall satisfaction: 0.7891.

XSL•FO
RenderX

Discussion

Product Development and User Usage Status of Hypertension Management Apps

In this study, although the number of HMAs increased in both the United States and China, there were significant differences in the number and frequency of app usage between the 2 countries. The overall number of HMAs launched has been increasing each year. However, the total number of HMAs in China (n=91) and the annual number of listings (n=6) were both lower than those in the United States (n=220 and n=15, respectively). This discrepancy may be related to differences in hypertension management practices between the 2 countries. In China, there is a greater emphasis on hospital visits and follow-ups, whereas the United States promotes self-health management by patients. As a result, American users are more likely to utilize HMAs. Additionally, most HMAs do not adhere to hypertension clinical guidelines, lack disclaimers, exhibit low levels of intelligence, and have inadequate standardized review processes. It is recommended to improve the standardization of app development by integrating standard clinical guidelines, enhancing the app's intelligence with new wearable devices and artificial intelligence technologies, clarifying regulatory approval policies for digital therapy software for chronic diseases, and refining the approval process. Additionally, while the overall usage frequency of HMAs is relatively high, further efforts to address these areas could enhance their effectiveness and user satisfaction. Although the usage frequency of HMAs among Chinese users was significantly lower compared with American users, the average usage frequency for each type of HMA did not differ substantially between the 2 countries. The total downloads of all HMAs exceeded 170 million, with over 140 million downloads in the United States and around 20 million in China. There was no significant difference in the average downloads per app between the 2 countries. This indicates that both American and Chinese users generally have positive attitudes toward HMA use. The lower total download rate among Chinese users may be attributed to the smaller number of HMA users in China.

Availability of Hypertension Management Apps

HMAs have diverse business demands but often exhibit poor functional availability. The summary of clinical guidelines for hypertension in the United States and China reveals that the business needs of HMAs in the United States emphasize the reliability of treatment plans and app performance, while in China, the focus is more on blood pressure monitoring and intelligent diagnosis. This may be related to the differing usage habits of users in the United States and China. US users prioritize ensuring the safety and effectiveness of apps, while Chinese users focus on comprehensive self-management of hypertension. Although HMAs currently offer some basic functions, their functional availability is still inadequate and does not fully address most needs for hypertension management. The summary of app functions reveals that most HMAs can assist users with recording blood pressure and heart rate, dynamically tracking blood pressure trends, and providing health interventions, thus addressing the need for blood pressure recording and health support. However, current apps still lack features such as intelligent hypertension diagnosis, personalized services, and health behavior interventions. HMAs must address business needs including blood pressure classification, blood pressure warnings, personalized treatment for different patient groups, family health management, and health behavior management.

User Satisfaction Evaluation Indicator System for Hypertension Management Apps

The results of HMA USEIS for the United States and China indicate that while the evaluation criteria for app user satisfaction are consistent across both countries, there are significant differences in the weighting of these criteria. This discrepancy may be attributed to variations in app performance attributes and the differing usage habits of HMA users in China and the United States. Although the performance attributes of apps in both countries are consistent, the differing usage purposes and habits lead to variations in the weighting of evaluation criteria. Chinese users prioritize usability, functionality, and monitoring effectiveness, while American users focus more on usability, monitoring functions, and data management. Additionally, there are significant differences in the specific evaluation indicators and their importance weights. This may be related to differences in the demand for hypertension management between users in the United States and China. Clear operating procedures and user-friendly designs are fundamental needs for users in both countries and simple, intuitive apps are appealing. Chinese users primarily use apps for blood pressure measurement and recording, which leads to higher requirements for the app's reliability and measurement accuracy. If an app frequently crashes or has significant measurement errors, it directly impacts its usability. American users primarily use apps for comprehensive blood pressure management, which includes data collection from wearable devices, real-time data synchronization, blood pressure recording, and visualization of blood pressure charts. Therefore, robust blood pressure tracking features and effective data synchronization are crucial for user satisfaction. Additionally, both American and Chinese users reported low overall satisfaction with HMAs (0.6861 vs 0.7891, respectively), which may be attributed to the inadequate functionality of HMAs in meeting user expectations. Most users expect HMAs to support comprehensive self-management of hypertension throughout their entire life cycle, including blood pressure measurement, recording, visual chart display, health behavior intervention, drug intervention, and intelligent diagnosis. However, most HMAs offer only limited functionality, failing to meet all these needs and resulting in low user satisfaction.

Suggestions for Improving User Satisfaction of Hypertension Management Apps

Improve Product Functionality According to Clinical Guidelines

For Chinese HMAs, it is recommended to enhance functions such as health communities, intelligent diagnosis of hypertension, and doctor-patient communication. Building an

XSL•FO
RenderX

online health community for patients with hypertension is crucial. This involves strengthening communication and exchange among users, providing additional emotional support, and enhancing their engagement with the app. Second, using artificial intelligence to develop a diagnostic algorithm for hypertension based on users' personal health data is recommended. This approach can help users quickly assess their hypertension risk or determine its severity. Third, establishing a doctor-patient communication platform is recommended. This platform would allow users to have online consultations, enable doctors to manage patients remotely, strengthen doctor-patient interactions, and reduce unnecessary procedures.

For HMAs in the United States, it is recommended to enhance functions such as family health management, external device connectivity, and personalized patient management. First, developing a family blood pressure management feature can enable real-time monitoring of blood pressure indicators and provide early warnings for abnormalities, particularly for older family members. Second, enhancing the use of smart wearable devices and blood pressure monitors will improve the app's data synchronization capabilities, enabling real-time monitoring of blood pressure indicators. Third, in line with American clinical guidelines for hypertension, it is important to develop personalized management functions. This includes tailored diagnosis and treatment for patients with various types of hypertension or hypertension with different complications.

Strengthen Monitoring and Improvement of User Satisfaction With a Hypertension Management App

Suggestions for strengthening the monitoring and improvement of user satisfaction are based on the HMA satisfaction evaluation indicator system. This involves real-time monitoring and evaluation of user satisfaction according to indicator weights and satisfaction ratings, followed by targeted improvements for factors associated with lower satisfaction. For example, American users have lower satisfaction with HMA advertising distribution, data synchronization, and reliability, with data synchronization being a factor with higher importance weights. This suggests that HMAs in the United States should prioritize improving data synchronization, optimizing advertising distribution, and enhancing app reliability. Chinese users reported poor satisfaction with cost, compatibility, and measurement accuracy, with measurement accuracy being a highly important factor. This indicates that Chinese HMAs need to focus on addressing inaccurate measurement results and, in addition, resolve issues related to high fees and poor app compatibility.

Increase Policy and Financial Support for Hypertension Management Apps

It is imperative to provide companies with entrepreneurial subsidies, establish expedited review channels for digital health software aimed at chronic disease management, and promote collaboration between companies and hospitals. Additionally, increasing funding for chronic disease digital health software and offering financial support for HMA development are recommended.

Limitations

This study had some limitations. First, due to differences in health care policies, diagnoses, and treatment models between the United States and China, the quantity and frequency of HMA use in China are lower than in the United States. To ensure data representativeness, the study did not use specialized sampling but instead included all available apps and user reviews. The volume of data collected was sufficient to accurately reflect the situation of HMAs in the United States and China. Second, while this study identified differences in the USEISs between the 2 countries and provided a preliminary analysis of these differences, a more in-depth analysis is needed. Finally, future research should focus on the following aspects: First, obtaining nonsensitive user data from app developers, such as age, sex, patient status, and whether the app is used personally, could further elucidate how different user characteristics impact satisfaction. Second, exploring how different data acquisition methods (eg, automatic collection from devices such as blood pressure monitors versus manual input by users) affect satisfaction is valuable. Third, further research should investigate the relationship between user satisfaction and continuance intention, focusing on factors influencing continuance intention and their specific pathways. Fourth, as users are increasingly concerned with therapeutic outcomes, future studies should integrate clinical efficacy data with user satisfaction to better understand the relationship between user satisfaction and clinical effectiveness.

Conclusions

Research indicates that both the functional availability and user satisfaction with HMAs in the United States and China are lacking. The current app functions fail to meet all core business needs, and there are notable differences in the USEISs between the 2 countries. The key indicators in the evaluation system for US HMAs are convenience, blood pressure tracking, and data synchronization. By contrast, the most important indicators for Chinese HMAs are convenience, reliability, and measurement accuracy. These findings offer valuable insights for app developers to enhance product performance and boost user satisfaction. Additionally, they assist users in selecting apps that best meet their needs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant number 81871455), the Zhejiang Provincial Natural Science Foundation of China (grant number LY22H180001), the Key Research and Development Program of Zhejiang Province (grant number 2022C03111), the Municipal Natural Science Foundation of Beijing of China (grant number 7222306), and the National TCM innovation team and talent support projects (grant number ZYYCXTD-C-202210).

XSL•FO
RenderX

Data Availability

This study obtained all HMA and user reviews from 8 Chinese app stores (China Apple App Store, Huawei App Store, Xiaomi App Store, Oppo App Store, Vivo App Store, Baidu App Store, 360 App Store, and Application Treasure App Store) and 2 English app stores (US Apple App Store and US Google Play Store) through web scraping technology and the Qimai app data official website. The crawler code can be found in [62]. The code for latent Dirichlet allocation topic modeling and statistical analysis in the author's previous study [23] can be found in [63]. The code for calculating the weight of indicators using the AHP can be found in [64].

Authors' Contributions

JL conceptualized and designed the study, undertook data screening, and made significant revisions to the manuscript. YFH drafted the manuscript, summarized the literature review, and undertook data acquisition, screening, analysis, and interpretation. HC supervised the data filtering and analysis process. JBL and MZ supervised the review method and interpretation of the data and provided valuable improvement suggestions. PX, JRL, WT, and YCL did a critical revision of the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 1372 KB - jmir_v26i1e60773_app1.docx]

References

1. World Health Organization (WHO). Global report on hypertension: the race against a silent killer. WHO. Geneva, Switzerland: WHO; 2023 Sep 19. URL: https://www.who.int/publications/i/item/9789240081062 [accessed 2024-03-28]
2. Intelligent Research Consulting. Development strategy analysis and future prospects planning report on China's mobile healthcare industry. Zhiyan Consulting. URL: https://www.chyxx.com/industry/1163905.html [accessed 2024-03-28]
3. IQVIA Institute For Human Data Science. Digital health app market booming. Medical Device Network. URL: https://www.medicaldevice-network.com/news/digital-health-apps [accessed 2024-03-28]
4. Ai media consulting. Prediction of download and market size of mobile medical apps. Zhihu. URL: https://www.zhihu.com/question/483945110 [accessed 2024-03-28]
5. Falah F, Sajadi SA, Pishgooie AH. Effect of a mobile-based educational app on blood pressure of patients with hypertension. BMJ Mil Health 2023 Apr 10;169(2):e001577. [doi: 10.1136/bmjmilitary-2020-001577] [Medline: 33303692]
6. Alzahrani S, Bin Muammar MF, Bin Muammar AF, Alolah A, Almutawa M. The adoption and acceptance of mHealth interventions for self-management of hypertension among adult patients: a systematic review. Cureus 2022 Nov;14(11):e31584 [FREE Full text] [doi: 10.7759/cureus.31584] [Medline: 36540533]
7. Kassavou A, Wang M, Mirzaei V, Shpendi S, Hasan R. The association between smartphone app-based self-monitoring of hypertension-related behaviors and reductions in high blood pressure: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Jul 12;10(7):e34767 [FREE Full text] [doi: 10.2196/34767] [Medline: 35819830]
8. Sun S, Yi S. Clinical practice of updating the 2023 Chinese guidelines for the prevention and treatment of hypertension. Electrocardiogram and Circulation 2023;42(3):203-206. [doi: 10.1097/CCM.0000000000003299]
9. Zhao P, Jiang L, Li K. A study on the current status and influencing factors of self health management in hypertensive patients. Preventive Medicine in South China 2023;49(1):10-13. [doi: 10.1097/01.hco.0000166597.52335.23]
10. Choi D, Park JJ, Yoon M, Park S, Jo S, Kim EJ, et al. Self-monitoring of blood pressure and feed-back using app in treatment of uncontrolled hypertension (SMART-BP): a randomized clinical trial. Korean Circ J 2022 Oct;52(10):785-794 [FREE Full text] [doi: 10.4070/kcj.2022.0133] [Medline: 36217600]
11. Gong K, Yan Y, Li Y, Du J, Wang J, Han Y, et al. Mobile health applications for the management of primary hypertension: a multicenter, randomized, controlled trial. Medicine (Baltimore) 2020 Apr;99(16):e19715 [FREE Full text] [doi: 10.1097/MD.0000000000019715] [Medline: 32311957]
12. Han H, Guo W, Lu Y, Wang M. Effect of mobile applications on blood pressure control and their development in China: a systematic review and meta-analysis. Public Health 2020 Aug;185:356-363. [doi: 10.1016/j.puhe.2020.05.024] [Medline: 32738577]
13. Kassavou A, Wang M, Mirzaei V, Shpendi S, Hasan R. The association between smartphone app-based self-monitoring of hypertension-related behaviors and reductions in high blood pressure: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Jul 12;10(7):e34767. [doi: 10.2196/34767] [Medline: 35819830]

XSL•FO
RenderX

14. Sharman JE, Ordunez P, Brady T, Parati G, Stergiou G, Whelton PK, et al. The urgency to regulate validation of automated blood pressure measuring devices: a policy statement and call to action from the world hypertension league. J Hum Hypertens 2023 Feb 07;37(2):155-159 [FREE Full text] [doi: 10.1038/s41371-022-00747-0] [Medline: 36476777]
15. Picone DS, Campbell NRC, Schutte AE, Olsen MH, Ordunez P, Whelton PK, et al. Validation status of blood pressure measuring devices sold globally. JAMA 2022 Feb 15;327(7):680-681 [FREE Full text] [doi: 10.1001/jama.2021.24464] [Medline: 35166811]
16. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
17. Khalil AA, Meyliana, Hidayanto AN, Prabowo H. Identification of factor affecting continuance usage intention of mHealth application: a systematic literature review. New York, NY: IEEE; 2020 Presented at: 2020 4th International Conference on Informatics and Computational Sciences (ICICoS); November 10-11, 2020; Semarang, Indonesia p. 2024-2003. [doi: 10.1109/icicos51170.2020.9299038]
18. Robbins R, Krebs P, Jagannathan R, Jean-Louis G, Duncan DT. Health app use among US mobile phone users: analysis of trends by chronic disease status. JMIR Mhealth Uhealth 2017 Dec 19;5(12):e197 [FREE Full text] [doi: 10.2196/mhealth.7832] [Medline: 29258981]
19. Healthy China Action Promotion Committee. Healthy China Action (2019-2030). Government of China. URL: http://www.gov.cn/xinwen/2019-07/15/content_5409694.htm [accessed 2024-03-28]
20. Jaeger BC, Chen L, Foti K, Hardy ST, Bress AP, Kane SP, et al. Hypertension statistics for US adults: an open-source web application for analysis and visualization of National Health and Nutrition Examination Survey data. Hypertension 2023 Jun;80(6):1311-1320. [doi: 10.1161/hypertensionaha.123.20900]
21. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR Mhealth Uhealth 2018 Jul 23;6(7):e10723 [FREE Full text] [doi: 10.2196/10723] [Medline: 30037787]
22. Liang J, He X, Jia Y, Zhu W, Lei J. Chinese mobile health apps for hypertension management: a systematic evaluation of usefulness. J Healthc Eng 2018;2018:7328274 [FREE Full text] [doi: 10.1155/2018/7328274] [Medline: 29744027]
23. He Y, Zhu W, Wang T, Chen H, Xin J, Liu Y, et al. Mining user reviews from hypertension management mobile health apps to explore factors influencing user satisfaction and their asymmetry: comparative study. JMIR Mhealth Uhealth 2024 Mar 28;12:e55199 [FREE Full text] [doi: 10.2196/55199] [Medline: 38547475]
24. Cordoba E, Idnay B, Garofalo R, Kuhns LM, Pearson C, Bruce J, et al. Examining the Information Systems Success (ISS) of a mobile sexual health app (MyPEEPS Mobile) from the perspective of very young men who have sex with men (YMSM). Int J Med Inform 2021 Sep;153:104529 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104529] [Medline: 34385097]
25. Po Harvey Chin Y, Hsin Huang I, Yu Hou Z, Yu Chen P, Bassir F, Han Wang H, et al. User satisfaction with a smartphone-compatible, artificial intelligence-based cutaneous pigmented lesion evaluator. Comput Methods Programs Biomed 2020 Oct;195:105649. [doi: 10.1016/j.cmpb.2020.105649] [Medline: 32750631]
26. Esmaeili Givi M, Keshavarz H, Kargar Azad Z. Quality assessment of E-learning website using asymmetric impact–performance analysis and Kano's customer satisfaction model: a case study based on WebQual 4.0. IDD 2022 May 09;51(1):35-46. [doi: 10.1108/idd-08-2021-0083]
27. Akter S, D'Ambra J, Ray P, Hani U. Modelling the impact of mHealth service quality on satisfaction, continuance and quality of life. Behaviour & Information Technology 2013 Dec;32(12):1225-1241. [doi: 10.1080/0144929x.2012.745606]
28. Kalankesh LR, Nasiry Z, Fein R, Damanabi S. Factors influencing user satisfaction with information systems: a systematic review. Galen Med J 2020 Jun 26;9:e1686 [FREE Full text] [doi: 10.31661/gmj.v9i0.1686] [Medline: 34466567]
29. Yu SY, Park J, Kim J, Lee H, Yoon HJ. A case study of customer satisfaction for information technology solutions. J Cent South Univ 2014 Nov 15;21(11):4279-4285. [doi: 10.1007/s11771-014-2425-2]
30. Ramirez V, Johnson E, Gonzalez C, Ramirez V, Rubino B, Rossetti G. Assessing the use of mobile health technology by patients: an observational study in primary care clinics. JMIR Mhealth Uhealth 2016 Apr 19;4(2):e41 [FREE Full text] [doi: 10.2196/mhealth.4928] [Medline: 27095507]
31. Breil B, Salewski C, Apolinário-Hagen J. Comparing the acceptance of mobile hypertension apps for disease management among patients versus clinical use among physicians: cross-sectional survey. JMIR Cardio 2022 Jan 06;6(1):e31617 [FREE Full text] [doi: 10.2196/31617] [Medline: 34989683]
32. Yuhao W. Research on the influencing factors and improvement strategies of user satisfaction with short videos on Bilibili. PhD thesis. Zhengzhou University. 2022. URL: https://xueshu.baidu.com/usercenter/paper/show?paperid=1p3504b0g50m0t70jp6u0gb0a1393858&site=xueshu_se [accessed 2024-08-22]
33. Yiying Z, Fengyi G, Xiaohong H. A study on the factors influencing user satisfaction of mobile medical app-based on the situation of sudden public health incidents. Health Economics Research 2023;40(7):63-67 [FREE Full text]
34. Kar AK. What affects usage satisfaction in mobile payments? Modelling user generated content to develop the "digital service usage satisfaction model". Inf Syst Front 2021 Jul 18;23(5):1341-1361 [FREE Full text] [doi: 10.1007/s10796-020-10045-0] [Medline: 32837261]

XSL•FO
RenderX

35. Ahn H, Park E. Motivations for user satisfaction of mobile fitness applications: an analysis of user experience based on online review comments. Humanit Soc Sci Commun 2023 Jan 03;10(1):3 [FREE Full text] [doi: 10.1057/s41599-022-01452-6] [Medline: 36619599]
36. Alnooh G, Alessa T, Noorwali E, Albar S, Williams E, de Witte LP, et al. Identification of the most suitable mobile apps to support dietary approaches to stop hypertension (DASH) diet self-management: systematic search of app stores and content analysis. Nutrients 2023 Aug 06;15(15):3476 [FREE Full text] [doi: 10.3390/nu15153476] [Medline: 37571413]
37. Haggag O, Grundy J, Abdelrazek M, Haggag S. A large scale analysis of mHealth app user reviews. Empir Softw Eng 2022;27(7):196 [FREE Full text] [doi: 10.1007/s10664-022-10222-6] [Medline: 36246486]
38. Desmal A, Hamid S, Othman M, Zolait A. A user satisfaction model for mobile government services: a literature review. PeerJ Comput Sci 2022;8:e1074. [doi: 10.7717/peerj-cs.1074] [Medline: 36091981]
39. Taylor HA, Sullivan D, Mullen C, Johnson CM. Implementation of a user-centered framework in the development of a web-based health information database and call center. J Biomed Inform 2011 Oct;44(5):897-908 [FREE Full text] [doi: 10.1016/j.jbi.2011.03.001] [Medline: 21396486]
40. Qiong L, Yunbei N, Guifeng L. Construction of scientific data quality evaluation index system and research on user satisfaction. Introduction to Library and Information 2022;7(12):31-38 [FREE Full text]
41. Na W, Jie L. Research on user satisfaction evaluation of FAQ question answering system based on AHP entropy weight method: taking question answering robot in university library as an example. Intelligence Science 2023;41(9):164-172 [FREE Full text]
42. Jian L, Xiaofei W. Research on the construction of user satisfaction evaluation index system for university library websites based on analytic hierarchy process. Intelligence Science 2022;40(12):80-87 [FREE Full text]
43. Cong Y, Linsong M, Hao W. Analysis of customer satisfaction of catering delivery based on IAHP and fuzzy comprehensive evaluation. Era Economy and Trade 2023;20(3):123-125 [FREE Full text]
44. Kumar N, Khunger M, Gupta A, Garg N. A content analysis of smartphone-based applications for hypertension management. J Am Soc Hypertens 2015 Feb;9(2):130-136. [doi: 10.1016/j.jash.2014.12.001] [Medline: 25660364]
45. Kaur M, Kaur H, Rathi S, Ashwitha M, Joanna J, Reddy S, et al. Apps on Google Play Store to assist in self-management of hypertension in Indian context: features analysis study. Mhealth 2022 Apr;8:14-14 [FREE Full text] [doi: 10.21037/mhealth-21-25] [Medline: 35449503]
46. Hui CY, Creamer E, Pinnock H, McKinstry B. Apps to support self-management for people with hypertension: content analysis. JMIR Mhealth Uhealth 2019 Jun 03;7(6):e13257 [FREE Full text] [doi: 10.2196/13257] [Medline: 31162124]
47. Greer DB, Abel WM. Exploring feasibility of mHealth to manage hypertension in rural black older adults: a convergent parallel mixed method study. Patient Prefer Adherence 2022;16:2135-2148 [FREE Full text] [doi: 10.2147/PPA.S361032] [Medline: 35999840]
48. Oladapo H, Chakraborty J. Blood pressure concerns: findings from a usability study of culturally infused mHealth design. In: Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies. Cham, Switzerland: Springer; 2022:296-305 URL: https://xueshu.baidu.com/usercenter/paper/show?paperid=103w04v09w1t0rd0fd2y0g60gp568289 [accessed 2024-08-22]
49. DiFilippo K, Chapman-Novakofski K. P134 Mobile apps for hypertension management: cross-sectional survey of dietitian app use in patient care. Journal of Nutrition Education and Behavior 2020 Jul;52(7):S79-S80. [doi: 10.1016/j.jneb.2020.04.181]
50. Thomas NA, Drewry A, Racine Passmore S, Assad N, Hoppe KK. Patient perceptions, opinions and satisfaction of telehealth with remote blood pressure monitoring postpartum. BMC Pregnancy Childbirth 2021 Feb 19;21(1):153. [doi: 10.1186/s12884-021-03632-9]
51. Wang S, Lee HS, Choi W. A feature-oriented analysis of developers' descriptions and user reviews of top mHealth applications for diabetes and hypertension. International Journal of Medical Informatics 2021 Dec;156:104598. [doi: 10.1016/j.ijmedinf.2021.104598]
52. Qimai Data (formerly ASO100). Professional mobile product business analysis platform. Qimai. URL: https://www.qimai.cn [accessed 2024-04-02]
53. tweetbotornot. GitHub. URL: https://github.com/mkearney/tweetbotornot [accessed 2024-04-02]
54. Tian H, Gao C, Xiao X. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics; 2020 Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Seattle, WA p. 4067-4076 URL: https://aclanthology.org/2020.acl-main.374.pdf [doi: 10.18653/v1/2020.acl-main.374]
55. Chen Z, Liu L, He C. A study on the evaluation of hypertension mobile medical software and self management related functions. Chinese General Practice Medicine 2018;21(19):2348-2352 [FREE Full text]
56. Tongtong Z. Research on the construction of usability evaluation index system for text based mobile user generated content. PhD thesis. Hebei University. 2019. URL: https://book.douban.com/subject/34817908/ [accessed 2024-08-22]
57. He L, Shiying Z. Research on quality evaluation system of mobile internet user generated content. Information Theory and Practice 2015;38(10):37-11. [doi: 10.16353/j.cnki.1000-7490.2015.10.002]

XSL•FO
RenderX

58. Chan L, Li T, Wende Z. Construction of risk assessment index system for copyright infringement of video knowledge payment content. Intelligence Theory and Practice 2021;44(03):108-190. [doi: 10.16353/j.cnki.1000-7490.2021.03.012]
59. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
60. Saaty TL. The Analytic Hierarchy Process. New York, NY: McGraw Hill; 1980. URL: https://max.book118.com/html/2019/0320/6154141144002015.shtm
61. Armstrong M, Baron A. Performance Management. London, UK: The Cromwell Press; 1998. URL: https://www.hanspub.org/reference/Reference?ReferenceID=23717
62. GitHub. URL: https://github.com/YunfanHe123/Python-crawler [accessed 2024-08-20]
63. GitHub. URL: https://github.com/YunfanHe123/User-satisfaction-of-high-blood-pressure-management-Apps [accessed 2024-08-20]
64. GitHub. URL: https://github.com/YunfanHe123/Analytic-Hierarchy-Process-Index-Weight- Calculation [accessed 2024-08-20]

Abbreviations

AHP: analytic hierarchy process
HMA: hypertension management app
TTF: task-technology fit
USEIS: user satisfaction evaluation indicator system

Edited by G Eysenbach; submitted 21.05.24; peer-reviewed by Y Yu, K Sang; comments to author 14.06.24; revised version received 24.06.24; accepted 09.08.24; published 03.09.24.

Please cite as:
He Y, Chen H, Xiang P, Zhao M, Li Y, Liu Y, Wang T, Liang J, Lei J
Establishing an Evaluation Indicator System for User Satisfaction With Hypertension Management Apps: Combining User-Generated Content and Analytic Hierarchy Process
J Med Internet Res 2024;26:e60773
URL: https://www.jmir.org/2024/1/e60773
doi:10.2196/60773
PMID:39226103

©Yunfan He, Han Chen, Peng Xiang, Min Zhao, Yingjun Li, Yongcheng Liu, Tong Wang, Jun Liang, Jianbo Lei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study

Daun Shin[1,2], MD, PhD; Hyoseung Kim[3], BS; Seunghwan Lee[3], BS; Younhee Cho[2,4], DA; Whanbo Jung[2], MD

[1]Department of Psychiatry, Anam Hospital, Korea University, Seoul, Republic of Korea
[2]Doctorpresso, Seoul, Republic of Korea
[3]VOLTWIN, Seoul, Republic of Korea
[4]Department of Design, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Daun Shin, MD, PhD
Department of Psychiatry
Anam Hospital
Korea University
73 Goryeodae-ro
Seongbuk-gu
Seoul, 02841
Republic of Korea
Phone: 82 1093649735
Fax: 82 2 920 5185
Email: rune1018@gmail.com

Abstract

Background: Depressive disorders have substantial global implications, leading to various social consequences, including decreased occupational productivity and a high disability burden. Early detection and intervention for clinically significant depression have gained attention; however, the existing depression screening tools, such as the Center for Epidemiologic Studies Depression Scale, have limitations in objectivity and accuracy. Therefore, researchers are identifying objective indicators of depression, including image analysis, blood biomarkers, and ecological momentary assessments (EMAs). Among EMAs, user-generated text data, particularly from diary writing, have emerged as a clinically significant and analyzable source for detecting or diagnosing depression, leveraging advancements in large language models such as ChatGPT.

Objective: We aimed to detect depression based on user-generated diary text through an emotional diary writing app using a large language model (LLM). We aimed to validate the value of the semistructured diary text data as an EMA data source.

Methods: Participants were assessed for depression using the Patient Health Questionnaire and suicide risk was evaluated using the Beck Scale for Suicide Ideation before starting and after completing the 2-week diary writing period. The text data from the daily diaries were also used in the analysis. The performance of leading LLMs, such as ChatGPT with GPT-3.5 and GPT-4, was assessed with and without GPT-3.5 fine-tuning on the training data set. The model performance comparison involved the use of chain-of-thought and zero-shot prompting to analyze the text structure and content.

Results: We used 428 diaries from 91 participants; GPT-3.5 fine-tuning demonstrated superior performance in depression detection, achieving an accuracy of 0.902 and a specificity of 0.955. However, the balanced accuracy was the highest (0.844) for GPT-3.5 without fine-tuning and prompt techniques; it displayed a recall of 0.929.

Conclusions: Both GPT-3.5 and GPT-4.0 demonstrated relatively reasonable performance in recognizing the risk of depression based on diaries. Our findings highlight the potential clinical usefulness of user-generated text data for detecting depression. In addition to measurable indicators, such as step count and physical activity, future research should increasingly emphasize qualitative digital expression.

(J Med Internet Res 2024;26:e54617) doi:10.2196/54617

XSL•FO
RenderX

KEYWORDS

depression; screening; artificial intelligence; digital health technology; text data

Introduction

Depressive disorders are globally prevalent mental health conditions that significantly impact social and occupational functioning [1-3]. Major depressive disorder (MDD) and dysthymia are particularly noteworthy, as together, they were the second leading cause of years lived with disability in 2010, with MDD and dysthymia contributing 8.2% and 1.4%, respectively. The global incidence of depression increased from 172 million in 1990 to 258 million in 2017, reflecting a 49.86% rise, and the associated burden increased by 37.5% from 1990 to 2010 [4,5]. Moreover, depression is a key contributor to increased mortality risk, as evidenced by a meta-analysis indicating a hierarchy in the lifetime prevalence of suicide among patients with affective disorders [6,7]. Early detection and intervention for clinically significant depression have garnered increased attention. Prior to the 2000s, there was insufficient evidence supporting the use of screening tools. However, the United States Preventive Services Task Force revised its stance in June 2002, recommending that physicians screen for MDD [8].

Depressive symptoms are typically evaluated through self-reports or clinical assessment, with notable assessment methods including the Center for Epidemiologic Studies Depression (CES-D) scale, Patient Health Questionnaire–9 (PHQ-9), and Beck Depression Inventory [9-13]. However, these methods have limitations. Patients may underreport symptoms, and there can be discrepancies between subjective reports and objective severity [14,15]. Additionally, individuals often seek initial treatment from general practitioners rather than psychiatry specialists, partly due to stigma and lack of awareness, further complicating accurate assessment [16].

To address these limitations, researchers are exploring various biomarkers, genetic markers, and ecological momentary assessments (EMAs) for more objective and accurate screening [17-19]. EMAs involve real-time data collection, either actively by user input or passively through sensors on wearable devices [20]. Quantitative data such as exercise levels, step counts, and sleep cycles are relatively straightforward to collect and analyze in relation to depression scales [21]. However, analyzing qualitative data, such as user-generated text, presents a more complex challenge.

User-generated text data hold significant clinical potential. Advances in artificial intelligence (AI), particularly in natural language processing (NLP), have enabled sophisticated analysis of such data [22,23]. Large language models (LLMs) like ChatGPT (OpenAI) have facilitated various medical applications, including in psychiatry [24,25]. These models can analyze language used by patients in everyday contexts, such as social media posts, speech, or writing, to detect markers associated with depression [26,27]. By examining linguistic patterns, NLP and LLMs can predict depression risk without relying on traditional survey participation [28].

Despite their potential, much of the text data used in depression research has been sourced from electronic medical records or social media, which may not fully represent natural language use [29]. Daily writing is a universal human activity with therapeutic benefits, including improvements in depression and nonsuicidal self-injury [30-34]. Diary writing supports patient introspection, growth, and communication with therapists, making it a promising EMA data source.

In this study, we aimed to develop an algorithm for depression screening based on user-generated diary text data. Using a daily writing app for emotions, we collected semistructured diary texts and analyzed them with an LLM. Our goal was to validate the clinical utility of these texts as an EMA data source for identifying depression risk.

By leveraging the therapeutic and diagnostic potential of daily writing, combined with the analytical power of LLMs, we seek to contribute to the field of digital mental health. Our approach addresses the limitations of traditional screening tools and highlights the importance of qualitative data in mental health assessment and intervention.

Methods

App Process and Diary Log

The text data were obtained through an app named Mind Station. This app prompts users to write daily emotional diaries. Based on diary data, an AI system assesses the risk of depression and provides this information to mental health professionals. In addition, it uses generative AI to create responses to emotional diaries. A psychiatric clinician reviews these responses to modify and enhance them in a cognitive-behavioral or supportive psychotherapeutic manner, then provides replies to the user's diary (Figure 1).

The daily diary log was divided into 4 paragraphs, beginning with a description of the events that occurred that day, followed by reflections on these events and the resulting emotions, and concluding with a free-form diary entry. For each paragraph, the limit was set to a maximum of 500 characters to ensure brevity. Based on the written diaries, a specialist in mental health and medicine provided individualized emotional support. This step involved encouraging exploration, rectifying distorted cognition, guiding antidepressive activities in daily life, and recommending medical assistance (if necessary). Our objective was to enhance mental health through numerous responses tailored to individual needs.

Furthermore, this app was not developed for research purposes but rather as a form of mental self-care. However, we had plans to collect various digital phenotype data through the Mind Station app and to build algorithms for diagnosing and treating depression by analyzing text data. Therefore, users of the app were allowed to write diaries and receive responses from clinical psychiatrists only after agreeing that their anonymized data could be used for future research purposes.

XSL•FO
RenderX

Figure 1. Application process. AI: artificial intelligence.

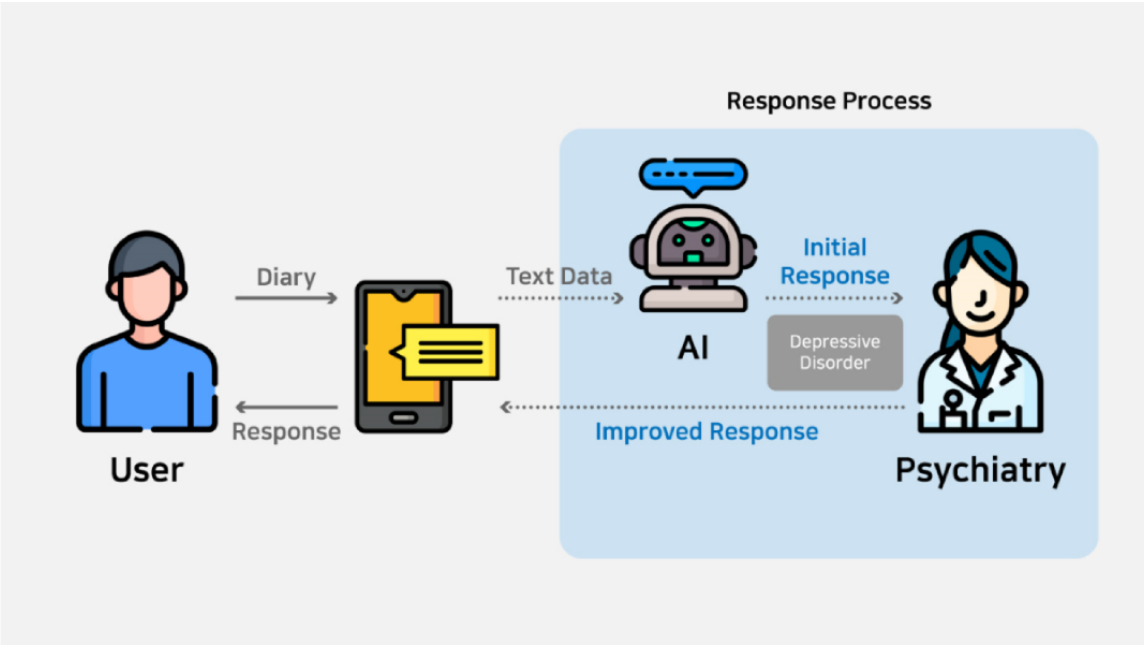

Recruitment

We recruited 91 participants from October 1, 2022, to April 30, 2023. Participants were recruited through promotion within internet communities for beta testing of the app. These promotions informed individuals about an app for writing diaries to care for their mental health where they could receive free responses from clinical psychiatrists. Additionally, they were notified that their anonymized data might be used for future research purposes. The participants received a reward of approximately ₩30,000 (US $22) for completing diaries for at least 1 week, writing daily, and submitting self-reported questionnaires about their mood. Furthermore, they voluntarily agreed to receive information on data collection and use before using the app. In addition, rewards were only disbursed after the completion of the week-long app experience and the submission of feedback regarding technical errors or diary responses encountered in the app. This was done to reduce the likelihood of participants writing diaries arbitrarily. Researchers should collect well-established risk factors, such as sex, family history, and employment status, for depression screening. However, an increase in the amount of input information can lead to lower compliance, potentially diminishing its role as a screening tool. Therefore, we performed an analysis based solely on text data from diaries written by users and from self-reported questionnaires. Diaries that were filled only with words lacking substantive content, such as “lol,” or solely with names of other individuals who were present at a particular location, were excluded from the analysis. This study aimed to retrospectively analyze the collected data.

Ethical Considerations

This study adhered to the ethical principles outlined in the Helsinki Declaration. Approval was obtained from the Institutional Review Board of Korea University Anam Hospital (2023AN0379). Each participant received approximately ₩30,000 (US $22) for completing diaries as long as they fulfilled the criteria described above. Data were anonymized. The participants were informed and consented to the use of their data for research purposes upon accessing the app. As this study involved retrospective analysis of the data, no written informed consent was obtained.

Depression Assessment Scale, Diary Classification, and Statistics

All participants were evaluated for depressive symptoms using the PHQ-9 and for potential crises in suicide situations using the Beck Scale for Suicide Ideation (BSS). The PHQ-9 consists of 9 questions related to mood, sleep, and other factors, with scores ranging from 0 to 3 for each question. The total score ranges from 0 to 27 [35]. The optimal cutoff score for depression screening is 10, with scores ≥11 indicating a risk of depression [36]. The BSS is a self-reported questionnaire for assessing suicide risk, and its potential as a screening tool in emergency rooms and inpatient settings has been validated [37-39]. These scores are well established in clinical research, with the PHQ-9 widely recognized for its validity and reliability in assessing depressive symptoms.

The classification of diaries as depressive was based on a combination of the validated PHQ-9, the BSS, and the clinical psychiatrist’s review. The primary criterion for determining a “true” classification was the PHQ-9 and BSS scores. A participant was classified as “depressed” if their PHQ-9 score was 10 or higher or if their BSS score was 8 or higher at the closest point to the diary-writing day. For diaries without available PHQ-9 and BSS scores, a psychiatrist reviewed the content. Diaries were classified as depressive if they contained direct expressions of depression, such as “I want to die” or “I am depressed,” or if they had clinically recognizable signs of depression, like “I am very stressed.” Additionally, if a participant’s PHQ-9 score was 10 or higher at the time of writing the first diary entry, but subsequent diaries (4-5 days later) included statements indicating an improved mood, such as “I feel better” or “Today was a good day,” the psychiatrist reviewed these diaries and classified them as not exhibiting

XSL•FO
RenderX

signs of depression. This dual approach of using validated depression scales and clinician review ensured a comprehensive and accurate classification of the diaries, balancing quantitative and qualitative assessments.

The statistical analysis method used for comparing depression and suicide scores before and after diary writing was a 2-tailed Student *t* test. Statistical analyses were conducted using SPSS Statistics (version 24.0; IBM Corp).

AI Technology and Data Preprocessing

LLMs have achieved remarkable success in the field of NLP. Initially, we applied the most popular LLMs, such as ChatGPT with GPT-3.5 and GPT-4 [40], to our data set without fine-tuning. In addition, we used a GPT-3.5 fine-tuning model with our training data set. Furthermore, we compared the performance of each model by applying chain-of-thought (CoT) and zero-shot prompting (Figure 2) [41,42]. CoT prompting, zero-shot prompting, and few-shot prompting were defined as follows in this study: CoT prompting is a method used in NLP to guide the model in generating intermediate reasoning steps leading to the final answer. Instead of directly producing an answer, the model generates a sequence of logical thoughts or steps connecting the input to the output. Zero-shot prompting refers to a scenario where an LLM is presented with a task without any prior specific examples or training on that task. The model relies on its preexisting knowledge and understanding to generate an appropriate response based on the provided prompt. Few-shot prompting is a technique in NLP where a model is given a small number of examples (usually ranging from one to a few dozen) of a task to learn from before being asked to perform the task itself. These examples aid the model in better understanding the task and improving its performance. Using the CoT approach, we structured the text and analyzed the words, expressions, and emotions used to classify it. Furthermore, we incorporated additional methods to compare the similarity of familiar texts, such as a compression algorithm based on the GNU Zip (Gzip) algorithm and k-nearest clustering. This step helped us compare the differences between the LLMs and conventional algorithms [43,44]. We adopted stratified 5-fold cross-validation to classify diaries while preserving the percentage of the ratio of labels. We divided the entire data set into training and test data sets at a ratio of 8:2.

In this study, we defined and calculated key metrics to evaluate the performance of our classification models. Recall, formerly known as sensitivity, is the proportion of actual positives correctly identified by the model. It is calculated by dividing the number of true positive predictions by the sum of true positive and false negative predictions. This metric helps in understanding how well the model can identify positive instances. Precision is the proportion of true positive predictions among the total positive predictions. It is calculated by dividing the number of true positive predictions by the sum of true positive and false positive predictions. Precision indicates the accuracy of the model's positive predictions. Specificity is the proportion of actual negatives correctly identified by the model. It is calculated by dividing the number of true negative predictions by the sum of true negative and false positive predictions. This metric is crucial for understanding the model's ability to correctly identify negative instances. Accuracy is the overall proportion of correct predictions, including both true positives and true negatives. It is calculated by dividing the sum of true positive and true negative predictions by the total number of predictions, which includes true positives, false positives, false negatives, and true negatives. Accuracy provides a general measure of how well the model performs across all classes [45].

Figure 2. Chain-of-thought (CoT) prompting. Differences with standard prompting are shown in blue.

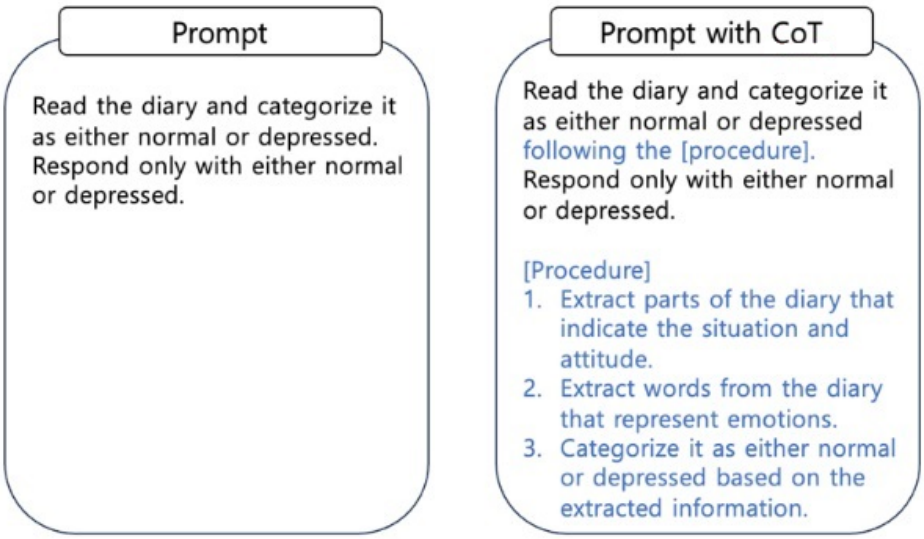

Results

We collected 428 diaries from 91 participants, with an average of 4.7 (SD 4.44; median 7.0) diaries authored by each user. Among the participants, 34 consented to disclosing their sex and 31 consented to disclosing their age. At baseline, 85 participants responded to the PHQ-9, and at the end of the 2-week period, 34 participants responded. Similarly, 85 participants responded to the baseline BSS, and 30 participants responded at the end of the 2-week period. Among the respondents, 81% (25/31) were women, and 19% (6/31) were men. Regarding age, 87% (27/31) of the total participants were

XSL•FO
RenderX

aged 20-39 years, while 13% (4/31) were aged 40-49 years. Every participant underwent an initial assessment using the PHQ-9 and BSS while writing their diaries. At baseline, 85 participants completed the PHQ-9. Their average score was 7.353 (SD 6.849), and 26 participants had scores of 10 or higher. Additionally, 85 participants completed the baseline BSS, and 30 participants completed it again at the end of the 2-week period. The average baseline BSS score was 4.200 (SD 5.708). Among the participants, 2 had PHQ-9 scores of 10 or lower and BSS scores of 8 or higher. As described in the Methods section, even if the initial PHQ-9 and BSS scores exceeded the cutoff points, diaries written 4-5 days later that included direct expressions of improved mood, such as "I feel better," were classified as nondepressive diaries based on the psychiatrist's judgment. Therefore, the total number of depressive diaries was 73. Furthermore, the PHQ-9 and BSS scores when terminating use of the app were lower than the scores at the beginning of use (PHQ-9: P=.32; BSS: P=.40) (Table 1).

Balanced accuracy was calculated as the average of the recall and specificity divided by 2, representing the accuracy for each class. This step is particularly useful for evaluating classes in unbalanced data sets. We used the F_1-score, accuracy, precision, recall, and specificity as evaluation metrics. GPT-3.5 fine-tuning demonstrated the best performance, with an accuracy of 0.902, F_1-score of 0.685, precision of 0.759, and specificity of 0.955. The best fold exhibited higher performance, with an accuracy of 0.942 and an F_1-score of 0.8. However, the recall for GPT-3.5 fine-tuning was moderate, at 0.643. By contrast, GPT-4 demonstrated the best performance, with a recall of 0.972. Nevertheless, GPT-4 lagged behind GPT-3.5 in fine-tuning regarding accuracy (0.743) and F_1-score (0.57) (Table 2). Figures 3 and 4 depict the combined confusion matrix for all folds of each model.

Table 1. Characteristics of participants.

Characteristics	Values
Sex of participants, n (%)	
Male	6 (19)
Female	25 (81)
Age group of participants (years), n (%)	
20-39	27 (87)
40-49	4 (13)
Patient Health Questionnaire–9 score , mean (SD)	
Baseline (n=85)	7.353 (6.849)
End (n=34)	5.735 (6.336)
Beck Scale for Suicide Ideation score , mean (SD)	
Baseline (n=85)	4.200 (5.708)
End (n=30)	2.967 (5.436)

XSL•FO
RenderX

Table 2. Performance of each algorithm with and without chain-of-thought (CoT) prompting. Italics indicate high performance.

Algorithm	Accuracy	Balanced accuracy	F_1-score	Precision	Recall	Specificity
Gpt3.5_ft						
Average	*0.883*	*0.808*[a]	0.670	0.653	0.695	*0.921*[b]
Maximum	0.930	0.853	0.786	0.846	0.733	0.972
Gpt3.5_ft_CoT						
Average	*0.902*[b]	0.799	0.685	0.759	0.643	*0.955*[c]
Maximum	0.942	0.833	0.800	1.000	0.667	1.000
Gpt3.5						
Average	0.789	*0.844*[a]	0.607	0.453	*0.929*[b]	0.761
Maximum	0.860	0.915	0.714	0.556	1.000	0.831
Gpt3.5_CoT						
Average	0.752	*0.818*[a]	0.560	0.405	*0.917*[b]	0.718
Maximum	0.802	0.828	0.605	0.464	0.867	0.789
Gpt4						
Average	0.713	*0.827*[a]	0.546	0.378	*1.000*[c]	0.653
Maximum	0.791	0.873	0.625	0.455	1.000	0.746
Gpt4_CoT						
Average	0.743	*0.834*[a]	0.570	0.406	*0.972*[c]	0.696
Maximum	0.826	0.894	0.667	0.500	1.000	0.789
Compression						
Average	*0.846*[a]	0.651	0.431	0.584	0.356	*0.947*[b]
Maximum	0.919	0.819	0.741	0.833	0.667	0.972

[a]Performance >0.80.
[b]Performance >0.90.
[c]Performance >0.95.

XSL•FO
RenderX

Figure 3. Confusion matrix for GPT 3.5 models. CoT: chain of thought; LLM: large language model.

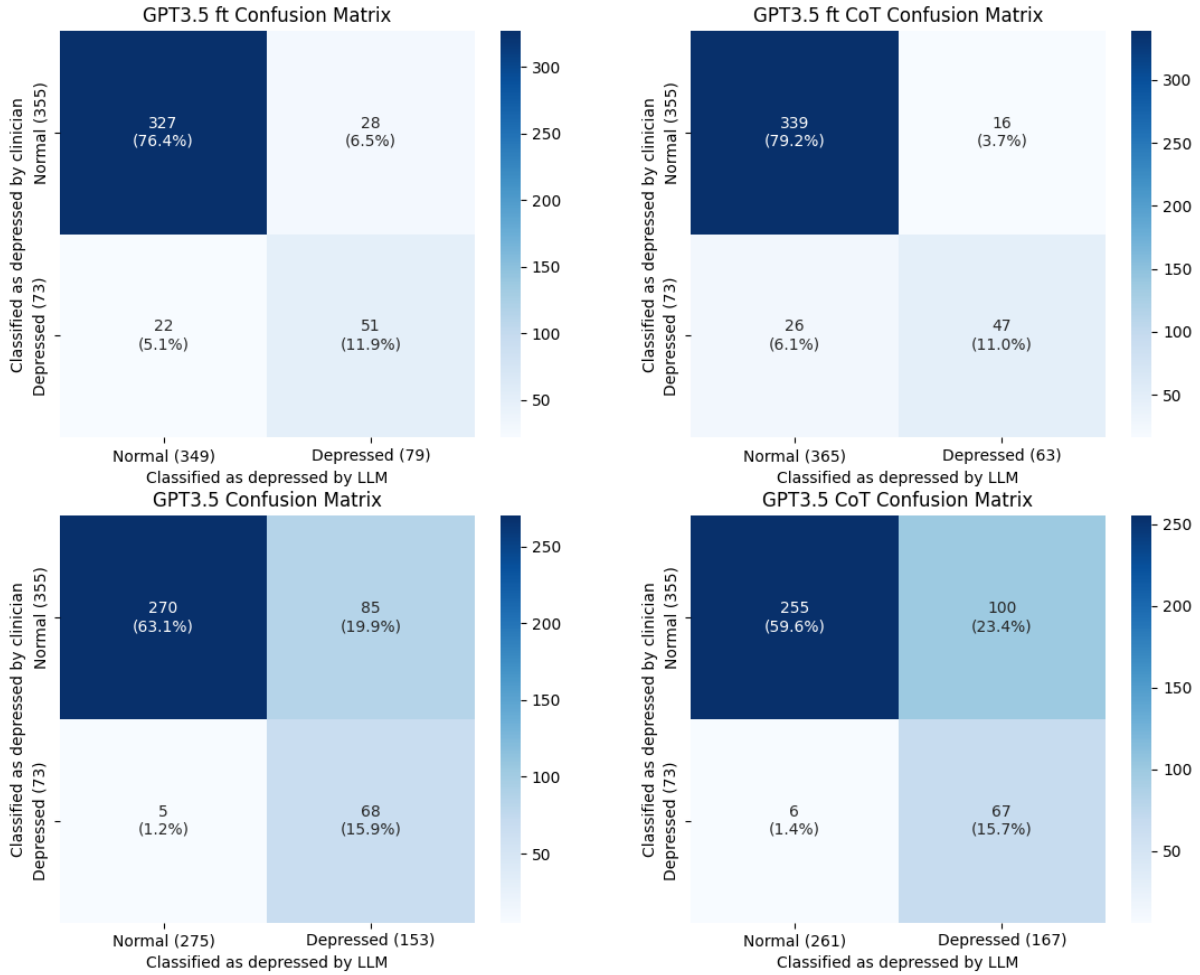

Figure 4. Confusion matrix for other artificial intelligence models. CoT: chain of thought.

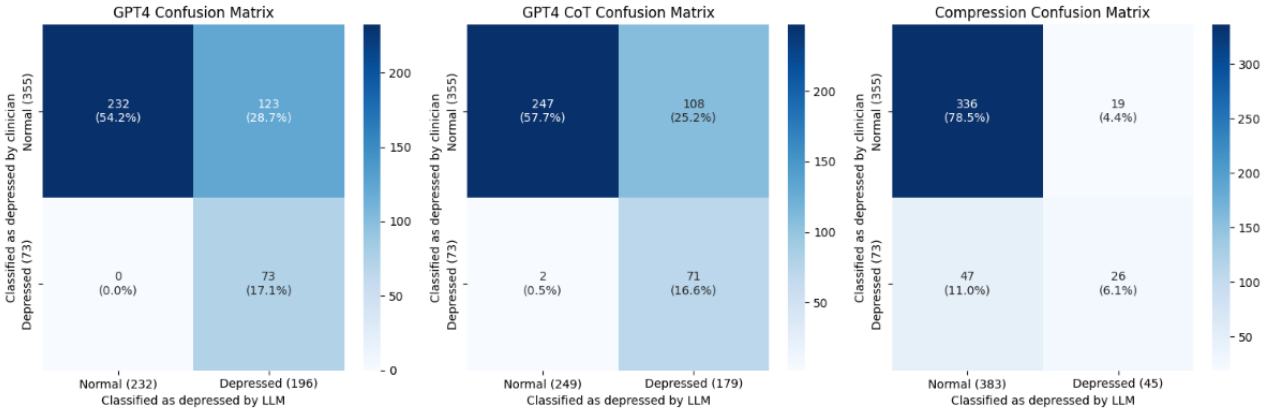

Discussion

Principal Findings

In this study, log data comprising 428 daily diaries were collected from 91 participants. We predicted the risk of depression based solely on text data from the diaries, excluding other clinical data. Approximately 80% of the users were women, and the majority were aged 20-39 years. The participants were prompted to write diaries about their moods and daily experiences. The average initial PHQ-9 score was 5.719, indicating most users did not exhibit a significant risk of depressive symptoms. The GPT-3.5 model, combined with CoT prompting, achieved an accuracy of 0.902 and a specificity of 0.955 in classifying high-risk cases of depression based on diary text data. GPT-4.0 showed superior recall (1.000) compared to the GPT-4.0 CoT model (0.972), although both models demonstrated similar balanced accuracy. The high imbalance in the data set, with 82.9% (355/428) nondepressive and 17.1% (73/428) depressive diaries, underscores the importance of using balanced accuracy as a metric.

XSL•FO
RenderX

The analysis reveals that while the GPT-3.5 fine-tuned model had high accuracy and specificity, suggesting effectiveness in predicting false outcomes, its relatively low balanced accuracy implies potential overfitting during the fine-tuning process and inadequate prediction of true outcomes. On the other hand, GPT-4 achieved perfect recall but may have overlooked errors in false predictions due to its focus on distinguishing between normal and abnormal data. This raises concerns about false predictions with the GPT-4 model. Despite GPT-4's lower performance compared to GPT-3.5, both models had balanced accuracy metrics that remained similar. The similarity in performance can be attributed to factors such as the sentiment task nature of the query and the reinforcement of ethical filters in GPT-4. These findings align with the existing literature and highlight the need for further investigation into AI model performance and ethical considerations in sentiment analysis tasks.

Language, particularly text data, excluding voice, is used by individuals and reflects their values and emotions. It can be utilized for EMA. However, various methodological limitations make it challenging to use EMA for depression screening. Previous research on diagnosing or detecting depressive symptoms using text has been based on data from various sources, including time-based narratives, suicide notes, and publicly available social media content. These studies aimed to detect depression not by analyzing the text itself but by extracting various indicators, such as the counts of positive and negative words. These indicators were used as proxies for depression detection [46-48]. Furthermore, social media posts and poems often contain abbreviations or words that are not typically used in everyday, natural-language conversations. Thus, using such data as effective digital biomarkers poses limitations because they may not represent real-life communication accurately. With the evolution of various text analysis technologies, such as NLP, the scope of text used for screening and diagnosing depression has expanded to include electronic medical records [49,50]. However, the language used in electronic medical records is often influenced by the patient's clinical judgment, which can result in an altered representation. Consequently, text-based depression screening based on electronic medical record data has limitations and poor generalizability.

Fine-tuning is the process of optimizing a pretrained model through additional training for a specific domain or task. Concentrated learning of data related to a specific domain or task can improve performance while conducting tasks related to that domain. CoT is a simple and popular prompting algorithm for improving the performance of language models without fine-tuning. Our results demonstrate the performance of GPT-3.5 and GPT-4 on classification tasks and their improvement upon techniques for additional prompting and fine-tuning. The compression algorithm Gzip is an easy and lightweight nonparametric alternative to neural networks used for text classification. Gzip demonstrates good performance even in out-of-distribution data sets. We determined the results that would emerge from using existing technology instead of deep learning.

Mood diaries exert therapeutic effects. Our results suggest lower depression and suicide scores after completing diary writing than scores after commencing; nonetheless, the difference was statistically insignificant. This finding is partly attributable to the insufficient sample size to prove statistical power. Moreover, the study was conducted in the general population, and few participants demonstrated significantly higher levels of depressive symptoms. The therapeutic effects of diary writing on depression should be confirmed through large-scale studies with a larger sample size.

This study confirmed that emotion diaries among the general population can be used to screen for depression at a level similar to that of passive EMA. The area under the receiver operating characteristic curve of depression through passive sensing depends on the data used; however, the accuracy ranges from approximately 60% to 90% [51]. Text-based depression prediction has a lower accuracy (approximately 0.632); nonetheless, the accuracy of depression diagnosis increases upon using text data from medical interviews along with other data, such as vocal features [52]. Accuracy increases upon analyzing a daily mood diary rather than predicting depression and suicidal thoughts through EMA data obtained through a wearable device [53]. This study was conducted in a relatively large population; however, it has the disadvantage of poor generalizability because the targets were medical interns. A previous study suggested that journaling by older adults can help detect depression, and through this study, we confirmed that depression detection based on journaling can be extended to other age groups [54].

The findings of this research have important clinical implications. If depression could be diagnosed based on daily diaries, it would offer significant clinical benefits. Early detection could identify at-risk individuals, allowing for timely interventions like counseling or therapy to prevent symptom progression. Analyzing diary text data could reveal triggers and patterns, informing personalized therapy approaches such as cognitive behavioral therapy to address negative thought patterns or environmental factors. Regular diary analysis could monitor treatment effectiveness, providing real-time feedback for adjustments and improving patient outcomes. In summary, leveraging daily diaries for depression diagnosis could revolutionize clinical practice by enabling early intervention, personalized therapy, and ongoing treatment monitoring.

Strengths

The strength of this study is that we confirmed the basis for predicting mild levels of depression. This is because the recruitment of participants was not centered on hospitals. Most text data, excluding social media data, are based on medical records. Therefore, the data are supposedly biased toward individuals with severe depression. An existing depression diagnostic algorithm based on the frequency of negative language demonstrated low accuracy upon application to our data [52]. In addition, the text data are meaningful in that they can be used for simultaneous diagnosis and treatment because writing a diary is already therapeutic. Therefore, they can be a good evaluation tool for determining the severe burden placed on patients from obtaining active EMA data [55]. In addition,

XSL•FO
RenderX

we confirmed that prompt engineering alone affected depression detection. Thus, generative AI can facilitate medical judgment, and prompt technology is important.

Limitations

This study has several limitations. First, the complete exclusion of other clinical data, such as sex, age, and occupation, serves as both an advantage and a disadvantage. The advantage is that it confirms the diagnostic accuracy of pure text data; however, the sample size was small, and the predominant users were women aged 20-39 years. Therefore, the data are possibly biased. Second, we used generative AI; thus, we could not confirm the clinical characteristics recorded in the depression diary. Existing studies have demonstrated clinical implications, such as the use of *I*-centered words in suicide diaries and the frequent use of negative language in social media about depression [46,48]. In addition, the small sample size warrants large-scale research to diagnose depression using written diaries and the clinical characteristics of patients with depression. Third, text data may have a high possibility of being used in combination with other biomarkers; however, such data were not collected in this study. Therefore, further research is required on various types of EMA data that can be used in algorithms to differentiate depression in the general population.

Conclusions

We investigated whether an LLM can detect depression based on user-generated emotional diaries. Using a data set comprising 428 diaries from 91 users, the accuracy of predicting depression increased upon applying the CoT prompt technique to both GPT-3.5 and GPT-4.0. GPT-3.5 generated the highest average accuracy (0.902). However, without fine-tuning or prompting techniques, GPT-3.5 exhibited the highest balanced accuracy of 0.844 and recall of 0.929. Despite decreased depressive scores after the participants began writing their diaries, this change was statistically insignificant. Therefore, additional research is warranted to explore the potential antidepressant effects of diary writing. Based on these findings, we identified the potential clinical utility of voluntarily generated text data for detecting depression. Beyond quantifiable indicators, such as step count, daily physical activity, and sleep duration, we consider that there will be a continued and increased focus on qualitative digital expressions in research.

Acknowledgments

We express our gratitude to all the participants who contributed to this paper.

Conflicts of Interest

DS is the chief medical officer at Doctorpresso. While DS does not receive a salary from the company, she owns 25% of its equity. YC is a project manager at Doctorpresso and receives a salary for her work. WJ is the chief executive officer of Doctorpresso and owns 61% of the company's equity. HK and SL receive salaries from VOLTWIN but do not own any company equity. The Mind Station app was created by DS in collaboration with WJ and YC with the aim of providing mental health care based on writing emotion diaries. The app is owned by Doctorpresso. The data collected through this app were analyzed by VOLTWIN, a company specializing in data analysis. It is important to note that the 4 coauthors, aside from the first and corresponding author, DS, did not influence the study's design or analysis. YC contributed to the design of the Mind Station app, and WJ worked on community outreach projects based on the Mind Station app. VOLTWIN was subcontracted to perform artificial intelligence–related analysis, and the analysis plan was carried out in consultation with DS, ensuring that the results were not affected by the company. This work was also supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare, Republic of Korea (grant HI23C1234). The funding source had no role in the study design, data collection, analysis, or interpretation or in the decision to submit the manuscript for publication.

References

1. Ferrari AJ, Charlson FJ, Norman RE, Flaxman AD, Patten SB, Vos T, et al. The epidemiological modelling of major depressive disorder: application for the Global Burden of Disease Study 2010. PLoS One 2013;8(7):e69637 [FREE Full text] [doi: 10.1371/journal.pone.0069637] [Medline: 23922765]
2. Ferrari AJ, Somerville AJ, Baxter AJ, Norman R, Patten SB, Vos T, et al. Global variation in the prevalence and incidence of major depressive disorder: a systematic review of the epidemiological literature. Psychol Med 2013 Mar;43(3):471-481. [doi: 10.1017/S0033291712001511] [Medline: 22831756]
3. Moreno-Agostino D, Wu Y, Daskalopoulou C, Hasan MT, Huisman M, Prina M. Global trends in the prevalence and incidence of depression:a systematic review and meta-analysis. J Affect Disord 2021 Feb 15;281:235-243. [doi: 10.1016/j.jad.2020.12.035] [Medline: 33338841]
4. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, et al. Burden of depressive disorders by country, sex, age, and year: findings from the Global Burden of Disease Study 2010. PLoS Med 2013 Nov;10(11):e1001547 [FREE Full text] [doi: 10.1371/journal.pmed.1001547] [Medline: 24223526]
5. Liu Q, He H, Yang J, Feng X, Zhao F, Lyu J. Changes in the global burden of depression from 1990 to 2017: findings from the Global Burden of Disease study. J Psychiatr Res 2020 Jul;126:134-140 [FREE Full text] [doi: 10.1016/j.jpsychires.2019.08.002] [Medline: 31439359]

XSL•FO
RenderX

6. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat 2011;7(Suppl 1):3-7 [FREE Full text] [doi: 10.2147/NDT.S19617] [Medline: 21750622]
7. Bostwick JM, Pankratz VS. Affective disorders and suicide risk: a reexamination. Am J Psychiatry 2000 Dec;157(12):1925-1932. [doi: 10.1176/appi.ajp.157.12.1925] [Medline: 11097952]
8. U.S. Preventive Services Task Force. Screening for depression: recommendations and rationale. Ann Intern Med 2002 May 21;136(10):760-764 [FREE Full text] [doi: 10.7326/0003-4819-136-10-200205210-00012] [Medline: 12020145]
9. Radloff LS. The CES-D Scale. Appl Psychol Meas 2016 Jul 26;1(3):385-401. [doi: 10.1177/014662167700100306]
10. Vieweg BW, Hedlund JL. The General Health Questionnaire (GHQ): a comprehensive review. J Op Psychiatry 1983;14(2):74-81.
11. Ball K, MacPherson C, Hurowitz G, Settles-Reaves B, DeVeaugh-Geiss J, Weir S. M3 checklist and SF-12 correlation study. Best Pract Ment Health 2015 Mar;11(1):83-89(7) [FREE Full text] [doi: 10.4324/9781315795621-9]
12. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatry 2015;37(1):67-75. [doi: 10.1016/j.genhosppsych.2014.09.009] [Medline: 25439733]
13. Lasa L, Ayuso-Mateos J, Vázquez-Barquero JL, Díez-Manrique FJ, Dowrick C. The use of the Beck Depression Inventory to screen for depression in the general population: a preliminary analysis. J Affect Disord 2000;57(1-3):261-265. [doi: 10.1016/s0165-0327(99)00088-9] [Medline: 10708841]
14. Hunt M, Auriemma J, Cashaw ACA. Self-report bias and underreporting of depression on the BDI-II. J Pers Assess 2003 Feb;80(1):26-30. [doi: 10.1207/s15327752jpa8001_10]
15. Ma S, Kang L, Guo X, Liu H, Yao L, Bai H, et al. Discrepancies between self-rated depression and observed depression severity: The effects of personality and dysfunctional attitudes. Gen Hosp Psychiatry 2021;70:25-30. [doi: 10.1016/j.genhosppsych.2020.11.016] [Medline: 33689981]
16. Parslow RA, Jorm AF. Who uses mental health services in Australia? An analysis of data from the National Survey of Mental Health and Wellbeing. Aust N Z J Psychiatry 2000 Dec 17;34(6):997-1008. [doi: 10.1046/j.1440-1614.2000.00839.x]
17. Le-Niculescu H, Kurian SM, Yehyawi N, Dike C, Patel SD, Edenberg HJ, et al. Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol Psychiatry 2009 Feb;14(2):156-174. [doi: 10.1038/mp.2008.11] [Medline: 18301394]
18. Abi-Dargham A, Horga G. The search for imaging biomarkers in psychiatric disorders. Nat Med 2016 Nov;22(11):1248-1255. [doi: 10.1038/nm.4190] [Medline: 27783066]
19. Rykov Y, Thach T, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. JMIR Mhealth Uhealth 2021 Oct 25;9(10):e24872 [FREE Full text] [doi: 10.2196/24872] [Medline: 34694233]
20. Kamath J, Leon Barriera R, Jain N, Keisari E, Wang B. Digital phenotyping in depression diagnostics: integrating psychiatric and engineering perspectives. World J Psychiatry 2022 Mar 19;12(3):393-409 [FREE Full text] [doi: 10.5498/wjp.v12.i3.393] [Medline: 35433319]
21. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
22. Cohen KB, Hunter L. Getting started in text mining. PLoS Comput Biol 2008 Jan;4(1):e20 [FREE Full text] [doi: 10.1371/journal.pcbi.0040020] [Medline: 18225946]
23. Rzhetsky A, Seringhaus M, Gerstein MB. Getting started in text mining: part two. PLoS Comput Biol 2009 Jul;5(7):e1000411 [FREE Full text] [doi: 10.1371/journal.pcbi.1000411] [Medline: 19649304]
24. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res 2016 Jun;25(2):86-100 [FREE Full text] [doi: 10.1002/mpr.1481] [Medline: 26184780]
25. Cheng S, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci 2023 Nov;77(11):592-596 [FREE Full text] [doi: 10.1111/pcn.13588] [Medline: 37612880]
26. Babu NV, Kanaga EGM. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Comput Sci 2022;3(1):74 [FREE Full text] [doi: 10.1007/s42979-021-00958-1] [Medline: 34816124]
27. Nanomi Arachchige IA, Sandanapitchai P, Weerasinghe R. Investigating machine learning and natural language processing techniques applied for predicting depression disorder from online support forums: a systematic literature review. Information 2021 Oct 27;12(11):444. [doi: 10.3390/info12110444]
28. Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Trans Asian Low-Resour Lang Inf Process 2024 Jan 15;23(1):1-20. [doi: 10.1145/3569580]
29. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomed Health Inform 2021 Aug;25(8):3121-3129. [doi: 10.1109/jbhi.2021.3063721]

XSL•FO
RenderX

30. Bolger N, Davis A, Rafaeli E. Diary methods: capturing life as it is lived. Annu Rev Psychol 2003;54(1):579-616. [doi: 10.1146/annurev.psych.54.101601.145030] [Medline: 12499517]
31. Chen Y, Ishak Z. Gratitude diary: the impact on depression symptoms. Psych 2022 Mar;13(03):443-453. [doi: 10.4236/PSYCH.2022.133030]
32. Hooley JM, Fox KR, Wang SB, Kwashie AND. Novel online daily diary interventions for nonsuicidal self-injury: a randomized controlled trial. BMC Psychiatry 2018 Aug 22;18(1):264 [FREE Full text] [doi: 10.1186/s12888-018-1840-6] [Medline: 30134866]
33. Abdallah CG, Jackowski A, Salas R, Gupta S, Sato JR, Mao X, et al. The nucleus accumbens and ketamine treatment in major depressive disorder. Neuropsychopharmacology 2017 Jul;42(8):1739-1746 [FREE Full text] [doi: 10.1038/npp.2017.49] [Medline: 28272497]
34. Faccio E, Turco F, Iudici A. Self-writing as a tool for change: the effectiveness of a psychotherapy using diary. Res Psychother 2019 Aug 09;22(2):378 [FREE Full text] [doi: 10.4081/ripppo.2019.378] [Medline: 32913803]
35. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
36. Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. Gen Hosp Psychiatry 2015;37(6):567-576. [doi: 10.1016/j.genhosppsych.2015.06.012] [Medline: 26195347]
37. Beck AT, Kovacs M, Weissman A. Assessment of suicidal intention: the Scale for Suicide Ideation. J Consult Clin Psychol 1979 Apr;47(2):343-352. [doi: 10.1037//0022-006x.47.2.343] [Medline: 469082]
38. Healy DJ, Barry K, Blow F, Welsh D, Milner KK. Routine use of the Beck Scale for Suicide Ideation in a psychiatric emergency department. Gen Hosp Psychiatry 2006;28(4):323-329. [doi: 10.1016/j.genhosppsych.2006.04.003] [Medline: 16814632]
39. Pinninti N, Steer R, Rissmiller D, Nelson S, Beck A. Use of the Beck Scale for Suicide Ideation with psychiatric inpatients diagnosed with schizophrenia, schizoaffective, or bipolar disorders. Behav Res Ther 2002 Sep;40(9):1071-1079. [doi: 10.1016/s0005-7967(02)00002-5] [Medline: 12296492]
40. GPT-4. OpenAI. URL: https://openai.com/research/gpt-4 [accessed 2024-09-13]
41. Wei J, Wang X, Schuurmans D, Bosma M. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, editors. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates; 2022:24824-24837.
42. Brown T, Mann B, Ryder N. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, editors. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates; 2020:1877-1901.
43. Jiang Z, Yang M, Tsirlin M. “Low-resource” text classification: a parameter-free classification method with compressors. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA: Association for Computational Linguistics; 2023:6810-6828.
44. Salvador-Meneses J, Ruiz-Chavez Z, Garcia-Rodriguez J. Compressed NN: K-nearest neighbors with data compression. Entropy (Basel) 2019 Feb 28;21(3):234 [FREE Full text] [doi: 10.3390/e21030234] [Medline: 33266949]
45. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022 Apr 08;12(1):5979 [FREE Full text] [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
46. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 2017 Dec;18:43-49. [doi: 10.1016/j.cobeha.2017.07.005]
47. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emot 2004 Dec;18(8):1121-1133. [doi: 10.1080/02699930441000030]
48. Stirman SW, Pennebaker JW. Word use in the poetry of suicidal and nonsuicidal poets. Psychosom Med 2001;63(4):517-522. [doi: 10.1097/00006842-200107000-00001] [Medline: 11485104]
49. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Jan;4(1):ooab011 [FREE Full text] [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
50. Zhong Q, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak 2018 May 29;18(1):30. [doi: 10.1186/s12911-018-0617-7]
51. Colombo D, Fernández-Álvarez J, Patané A, Semonella M, Kwiatkowska M, García-Palacios A, et al. Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: a systematic review. J Clin Med 2019 Apr 05;8(4):465 [FREE Full text] [doi: 10.3390/jcm8040465] [Medline: 30959828]
52. Shin D, Kim K, Lee S, Lee C, Bae YS, Cho WI, et al. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. Front Psychiatry 2022;13:801301 [FREE Full text] [doi: 10.3389/fpsyt.2022.801301] [Medline: 35686182]

XSL•FO
RenderX

53. Horwitz A, Czyz E, Al-Dajani N, Dempsey W, Zhao Z, Nahum-Shani I, et al. Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns. J Affect Disord 2022 Sep 15;313:1-7 [FREE Full text] [doi: 10.1016/j.jad.2022.06.064] [Medline: 35764227]
54. Chepenik LG, Have TT, Oslin D, Datto C, Zubritsky C, Katz IR. A daily diary study of late-life depression. Am J Geriatr Psychiatry 2006 Mar;14(3):270-279. [doi: 10.1097/01.JGP.0000194644.63245.42] [Medline: 16505132]
55. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017 Mar 15;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]

Abbreviations

AI: artificial intelligence
BSS: Beck Scale for Suicide Ideation
CES-D: Center For Epidemiologic Studies Depression
CoT: chain-of-thought
EMA: ecological momentary assessment
Gzip: GNU Zip
LLM: large language model
MDD: major depressive disorder
NLP: natural language processing
PHQ-9: Patient Health Questionnaire–9

Edited by G Eysenbach; submitted 16.11.23; peer-reviewed by Z Su, L Luo; comments to author 30.04.24; revised version received 17.05.24; accepted 11.08.24; published 18.09.24.

Please cite as:
Shin D, Kim H, Lee S, Cho Y, Jung W
Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study
J Med Internet Res 2024;26:e54617
URL: https://www.jmir.org/2024/1/e54617
doi:10.2196/54617
PMID:39292502

©Daun Shin, Hyoseung Kim, Seunghwan Lee, Younhee Cho, Whanbo Jung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

mHealth Physical Activity and Patient-Reported Outcomes in Patients With Inflammatory Bowel Diseases: Cluster Analysis

Ashley C Griffin[1,2], MSPH, PhD; Lucas Mentch[3], PhD; Feng-Chang Lin[4], PhD; Arlene E Chung[5], MD, MMCi, MHA

[1]VA Palo Alto Health Care System, Palo Alto, CA, United States

[2]Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

[3]Department of Statistics, University of Pittsburgh, Pittsburgh, PA, United States

[4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[5]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, United States

Corresponding Author:
Ashley C Griffin, MSPH, PhD
VA Palo Alto Health Care System
795 Willow Road
Palo Alto, CA, 94025
United States
Phone: 1 6504935000
Email: griffina@stanford.edu

Abstract

Background: Regular physical activity is associated with improved quality of life in patients with inflammatory bowel diseases (IBDs), although much of the existing research is based on self-reported data. Wearable devices provide objective data on many rich physical activity dimensions including steps, duration, distance, and intensity. Little is known about how patients with IBDs engage in these varying dimensions of exercise and how it may influence their symptom and disease-specific patient-reported outcomes (PROs).

Objective: This study aims to (1) cluster physical activity patterns from consumer-grade wearable devices and (2) assess the relationship between the clusters and PROs in patients with IBDs.

Methods: We conducted a cross-sectional and longitudinal cohort study among adults with IBDs in the Crohn's and Colitis Foundation IBD Partners cohort. Participants contribute physical activity data through smartphone apps or wearable devices in a bring-your-own-device model. Participants also complete biannual PRO questionnaires from the Patient-Reported Outcomes Measurement Information System short forms and IBD-specific questionnaires. K-means cluster analysis was used to generate physical activity clusters based on 3 key features: number of steps, duration of moderate to vigorous activity (minutes), and distance of activity (miles). Based on the clusters, we conducted a cross-sectional analysis to examine differences in mean questionnaire scores and participant characteristics using one-way ANOVA and chi-square tests. We also conducted a longitudinal analysis to examine individual cluster transitions among participants who completed multiple questionnaires, and mean differences in questionnaire scores were compared using 2-tailed paired sample *t* tests across 6-month periods.

Results: Among 430 participants comprising 1255 six-week physical activity periods, we identified clusters of low (33.7%, n=423), moderate (46%, n=577), and high (20.3%, n=255) physical activity. Scores varied across clusters for depression (P=.004), pain interference (P<.001), fatigue (P<.001), sleep disturbance (P<.001), social satisfaction (P<.001), and short Crohn Disease Activity Index (P<.001), with those in the low activity cluster having the worst scores. Sociodemographic characteristics also differed, and those with low physical activity were older (P=.002), had higher BMIs (P<.001), and had longer disease durations (P=.02) compared to other clusters. Among 246 participants who completed at least 2 consecutive questionnaires consisting of 726 questionnaire periods, 67.8% (n=492) remained in the same cluster, and only 1.2% (n=9) moved to or from the furthest clusters of low and high activity across 6-month periods.

Conclusions: For patients with IBDs, there were positive associations between physical activity and PROs related to disease activity and psychosocial domains. Physical activity patterns mostly did not fluctuate over time, suggesting little variation in exercise levels in the absence of an intervention. The use of real-world data to identify subgroups with similar lifestyle behaviors could be leveraged to develop targeted interventions that provide support for psychosocial symptoms and physical activity for personalized IBD care.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48020) doi:10.2196/48020

KEYWORDS

inflammatory bowel diseases; patient-reported outcome measures; cluster analysis; wearable electronic devices; medical informatics; mHealth; mobile health; physical activity; bowel disease; psychosocial; smartphone; wearables; mobile phone

Introduction

Background

Inflammatory bowel diseases (IBDs), which include Crohn disease (CD) and ulcerative colitis (UC), are chronic intestinal disorders of the gastrointestinal tract. IBDs are characterized by cycles of active and dormant states of inflammatory immune response with symptoms such as abdominal pain, diarrhea, or fatigue [1]. Physical symptoms are frequently accompanied by stress, anxiety, depression, or diminished quality of life and may be exacerbated by common immunosuppressive therapies or corticosteroid treatments [2]. Supporting patients with IBDs to directly report their symptoms or functional status through patient-reported outcome (PRO) measures provides an important clinical end point to understanding the burden of the disease [3,4]. Physical activity may improve these physical and psychosocial symptoms, although there may be variations in inflammation related to the multidimensions of activity such as duration, frequency, or intensity [1,5]. For example, moderate physical activity interventions, including cardiovascular training, strength training, and yoga, have demonstrated generally positive effects on symptoms and inflammation [5-9]. However, high-intensity or extended durations of exercise can lead to inflammation and exacerbate gastrointestinal symptoms [6,10]. Examining these patterns of physical activity relative to PROs has not been well-studied but may be important to inform strategies for physically inactive subgroups of patients with IBDs.

Wearable devices provide objective mobile health data on physical activity, which include many rich dimensions of activity (ie, steps, distance, duration, frequency, intensity, and gait speed). Estimates indicate that up to 45% of the US population owns a wrist-worn wearable device [11,12]. Few studies have used data from wearables to assess physical activity in patients with IBDs, as much of this research is based on self-reported physical activity questionnaires [1]. Early findings using wearable devices suggest an association between physical activity and biomarkers for inflammation and disease activity [13,14]. Digital phenotyping, defined by Torous et al [15], is the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices." Within the context of physical activity, digital phenotyping has been valuable in distinguishing different activity types from high-volume, heterogeneous wearable device data [16]. Prior studies have used clustering methods, such as K-means, to derive physical activity phenotypes, including "weekend warriors," "busy bees," "drivers," "insufficiently active," "cardio active," and "deconditioned" [16-20]. Quantifying these physical activity behaviors can reveal unique lifestyle characteristics and insights into the disease's impact on health status and functioning. Given the complex pathophysiology and heterogeneity of IBDs [21], accounting for individual variability in lifestyle behaviors is critical to inform personalized IBD care.

Little is known about whether there are physical activity phenotypes for patients with IBDs, yet identifying groups with similar physical activity patterns could facilitate tailoring lifestyle recommendations that enhance support for IBD. The use of real-world wearable device data provides opportunities to objectively understand the relationship between physical activity and the burden of disease. The Precision Visualization and Statistical Approaches (VISSTA) for patient-generated health data to enable precision medicine study focuses on developing health recommendations for lifestyle behaviors to improve health outcomes (ie, symptoms and quality of life) [22]. In this substudy of Precision VISSTA, we sought to identify physical activity phenotypes from consumer-grade wearable devices and examine the relationship between these phenotypes and PROs to support precision IBD care.

Objectives

The objectives of this study were to (1) identify clusters of physical activity patterns and (2) evaluate the associations between physical activity and PROs (disease activity and quality of life domains) in patients with IBDs.

Methods

Study Design and Setting

We conducted a cross-sectional and longitudinal cohort study among participants in the Crohn's and Colitis Foundation (CCF) IBD Partners study, which is an internet-based cohort of adults (18 years and older) living with IBD [23]. Participants in the CCF IBD Partners study have access to a portal to sync physical activity tracking smartphone apps or wearables (eg, Fitbit and Garmin) through a bring-your-own-device model [23,24]. Biannual questionnaires are completed for the PROMIS (Patient-Reported Outcomes Measurement Information System) domains of depression, anxiety, pain interference, sleep disturbance, satisfaction with social roles and activities, and fatigue. Disease activity questionnaires specific to the type of IBD (short Crohn Disease Activity Index [SCDAI] and Simple Clinical Colitis Activity Index [SCCAI]) are also completed biannually. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) to report observational studies (Multimedia Appendix 1) [25].

Ethical Considerations

This study was reviewed and approved by the institutional review board of the University of North Carolina at Chapel Hill (#17-2148). All individuals provided informed consent to participate in the IBD Partners parent study, and the data used for this study were considered secondary data analysis as all data were deidentified to protect privacy and ensure confidentiality. Separate consent was not required for this

XSL•FO
RenderX

current study. A description of the data management system and establishment of IBD Partners (previously CCFA Partners) were previously reported.

Participants

Participants were included in this analysis if they were enrolled in the CCF IBD Partners study between 2011 and 2020, completed at least one PROMIS and disease activity questionnaire, and contributed at least 50% of nonconsecutive or consecutive wearable device data within the 6 weeks before questionnaire completion. Participants could be included in multiple time periods if they completed more than one questionnaire. Approximately 20% (n=113) of the 543 participants who completed a questionnaire did not have sufficient activity data and were excluded. Thus, a total of 430 participants were included in our final analytic sample.

Measures

PROMIS short forms were completed for each of the 6 health-related quality of life domains above and scored using standardized T-scores with a population mean of 50 (SE 10) [26]. Higher T-scores represent more of the concept being measured (eg, more fatigue or more satisfaction with social roles and activities). The SCDAI assesses the severity of symptoms (ie, stool frequency, abdominal pain, and well-being) within the past 7 days [27]. SCDAI scores range from 0 to over 600 with scores <150 indicating clinical remission and ≥150 indicating active disease (mild activity: 150-219; moderate activity: 220-450; and severe activity: >450) [27]. UC or indeterminate colitis (IC) disease activity was measured using the SCCAI, which assesses the severity of symptoms (ie, frequency of bowel movement during the day and night, urgency of defecation, blood in stool, well-being, and extracolonic features) within the past 7 days [28]. SCCAI total scores range from 0 to 19. Scores <2.5 correlate with remission, whereas scores ≥2.5 correlate with active disease [29]. We also obtained sociodemographic and clinical characteristics of the cohort.

Wearable Data Preprocessing

Physical activity features varied across device brands (eg, Fitbit, Garmin, and Under Armour), so matrices were created to illustrate the capabilities of each device. Features were standardized into minutes for duration and miles for distance. Tukey's method was used to assess outliers, which were considered observations 1.5 times less or greater than the lower and upper quartile ranges, respectively [30]. Unrealistic values were removed (eg, activity duration of 24 hours). To determine the period of physical activity prior to completing a questionnaire to include in our analysis, Spearman's correlation coefficients were calculated to measure the strength of the relationship between physical activity (steps) and disease activity (SCCAI and SCDAI). Coefficients were calculated at various weeks prior to completing a questionnaire (ie, weeks 1, 2, 4, 6, 8, 12, and 24) with at least 50% of days with data returned during the time period. The period with the highest correlation between steps and disease activity was selected, which was 6 weeks. During this 6-week period, the correlation coefficient for steps and SCDAI was –0.19 (P<.001) and –0.14 (P=.007) for SCCAI.

Features included in the analysis were daily averages for each day that activity was recorded for a participant: number of steps, duration of moderate to vigorous activity (minutes), distance of activity (miles), number of calories burned, number of days the device was used during weekdays, and number of days the device was used during weekends. Features were then averaged during the 6 weeks prior to each questionnaire timepoint per participant. We initially performed featuring scaling (z score standardization and min-max normalization), which had little effect on the quality of clustering. Thus, the raw preprocessed data were used in the clustering.

K-means algorithm or Lloyd's method [31], which is a simple and widely used algorithm that organizes data into nonhierarchical groups, was iteratively run using combinations of these features. The quality of clusters was evaluated using silhouette coefficients, which measure the amount of cohesion and separation within and among clusters [32,33]. Ranging from –1 to 1, higher coefficients indicate better-defined clusters [32,33]. The combination of features that produced the highest average silhouette coefficient was selected. This included 3 features: number of steps, duration of moderate to vigorous activity (minutes), and distance of activity (miles). These features were moderate to highly correlated as demonstrated by Spearman correlation coefficients (ranging between 0.63 and 0.95).

Here, a_x is the average distance from an individual data point x to all data points within the same cluster and b_x is the average distance from an individual data point x to all data points in the nearest cluster.

Cluster Identification and Evaluation

K-means cluster analysis was then used to generate the physical activity clusters. K-means uses an iterative approach to assign each data point (x) to the closest centroid c_i within each cluster C_i by minimizing the average sum of squared Euclidean distance [32]. The equation below represents the K-means algorithm in which the sum of squared errors (SSE) or cluster scatter is minimized. K-means analyses were conducted using the Scikit-learn package in Python 3.7 [34]. The following parameters were used: number of clusters=3; initialization=k-means++; number of initializations=10; maximum iterations=300; tolerance=0.0001; random state=none. The algorithm was initialized 10 times with the centroids initially selected randomly (random state=none), and the initial centroid distances were optimized with k-means++. K-means++ places centroids to be generally distant from each other, which has been shown to improve the speed of convergence [35]. When the difference in the centroids across 2 consecutive iterations was lower than the tolerance level (0.0001), the algorithm converged and the iteration stopped.

Here, K is the number of clusters, x is a data point that belongs to the cluster C_i, C_i is the ith cluster, $dist$ is the standard

XSL•FO
RenderX

Euclidean distance between two data points, and c_i is the mean (centroid) of cluster C_i.

The number of clusters (K) was determined by calculating the SSE for different values of K. The value of K was selected where the change in SSE decreased and was becoming plateau, indicating additional clusters produced little value ("elbow point"). We iteratively evaluated the silhouette coefficients of 2 to 5 clusters and selected the model with the highest average silhouette coefficient, which was 3 clusters (Figure 1). Averages of silhouette coefficients were calculated for each cluster and the average of each of the 3 clusters. Once the clusters were identified, data within each cluster were assessed to determine a label that appropriately represented each cluster's attributes [36].

Figure 1. The SSE plot. SSE: sum of squared errors.

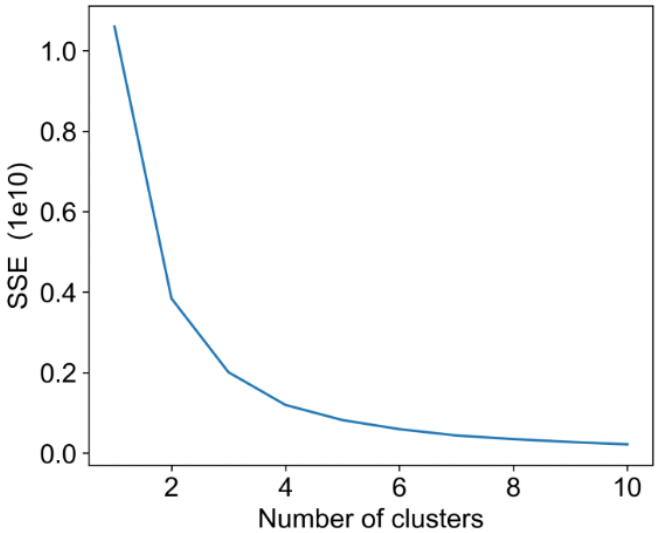

Association Between Physical Activity and PROs

Based on these clusters, we conducted two analyses: (1) a cross-sectional analysis (n=430) to assess differences in sociodemographics and PRO scores among the clusters, and (2) a longitudinal analysis (n=246) to examine mean changes in disease activity scores as participants moved to or from clusters among a subset of those who completed at least 2 consecutive biannual questionnaires. In the cross-sectional analysis, differences in sociodemographics and questionnaire scores among the clusters were assessed using one-way ANOVA tests and chi-square tests. Participants were included at multiple time periods if they had more than one 6-week physical activity period and completed the questionnaire. Including multiple observations per participant allowed us to examine the change in consecutive questionnaire scores as participants moved to or from clusters in our second analysis. In the longitudinal analysis, participants were grouped into categories related to staying in the same activity cluster or moving into another cluster across consecutive questionnaire time periods. For each move between clusters, mean differences in disease activity scores were compared using 2-tailed paired sample *t* tests. For all analyses, a *P* value <.05 was considered statistically significant.

Results

Participant Characteristics

The final analytic sample contained a total of 430 participants, of which 285 (66.3%) participants had CD and 145 (33.7%) participants had ulcerative or indeterminate colitis (Table 1). Participants were primarily women, White, and attained at least a college degree. On average, age was 42.1 (SD 13.6) years, BMI was 25.8 (SD 4.8) kg/m^2, and duration of disease was 15.1 (SD 11.8) years. Nearly all had used corticosteroid medications (93.5%, n=400) or 5-aminosalicylic acid medications (86.3%, n=371) to treat their IBD. Participants in the sample completed an average of 3 (SD 2) questionnaires from 2015 to 2020. During the 6-week period prior to completing questionnaires, participants used their wearable device for an average of 37.3 (SD 6.2) days (88.9% of the 6-week period). This included approximately 89.7% (26.9 of 30) of weekdays and 86.7% (10.4 of 12) of weekends. The majority used a Fitbit device (86.3%, n=371). On average, participants took 7892.9 (SD 2752.7) daily steps, performed moderate to vigorous activity for 40.7 (SD 38.3) minutes, traveled 3.5 (SD 1.3) miles, and burned 520.7 (SD 170) calories during exercise.

XSL•FO
RenderX

Table 1. Sample characteristics[a].

Characteristics	Value (N=430)
Age (years), mean (SD)	42.1 (13.6)
Gender, n (%)	
Woman	318 (74)
Man	112 (26)
Race[b], n (%)	
Asian	5 (1.2)
Black	7 (1.7)
White	394 (95.2)
Other	8 (1.9)
Ethnicity[b], n (%)	
Not Hispanic or Latino	405 (97.1)
Hispanic or Latino	13 (2.9)
Education[b], n (%)	
High school or less	21 (5)
Some college	72 (17.1)
College degree or more	327 (77.9)
BMI (kg/m^2), mean (SD)[b]	25.8 (4.8)
Smoking status, n (%)	
Ever	121 (28.1)
Never	309 (71.9)
Type of IBD[c], n (%)	
Crohn disease	285 (66.3)
Ulcerative or indeterminate colitis	145 (33.7)
Duration of disease (years), mean (SD)	15.1 (11.8)
History of IBD treatments (ever used)[b], n (%)	
5-Aminosalicylic acid medications	371 (86.3)
Biologic therapies	301 (70.3)
Corticosteroid medications	400 (93.5)
Immunomodulators	282 (65.9)
Device brand, n (%)	
Fitbit	371 (86.3)
Garmin	43 (10)
Jawbone	12 (2.8)
Under Armour	4 (0.9)
Questionnaires completed, mean (SD)	3.0 (2.0)
Days device used (maximum of 42 days), mean (SD)	37.3 (6.2)
Days device used on weekdays (maximum of ~30 days), mean (SD)	26.9 (4.4)
Days device used on weekends (maximum of ~12 days), mean (SD)	10.4 (2.1)
Daily steps, mean (SD)	7892.9 (2752.7)
Daily moderate to vigorous duration in minutes, mean (SD)	40.7 (38.3)
Daily distance in miles, mean (SD)	3.5 (1.3)

XSL•FO
RenderX

Characteristics	Value (N=430)
Daily activity-related calories, mean (SD)	520.7 (170)

[a]Demographic and clinical characteristics are from the baseline questionnaire.

[b]Missing data: race=16; ethnicity=12; education=10; BMI=14; history of treatments for inflammatory bowel diseases=2 participants.

[c]IBD: inflammatory bowel disease.

Clusters of Physical Activity Patterns

After examining data within each of the 3 clusters, we labeled the clusters in a way that was interpretable: low physical activity, moderate physical activity, and high physical activity (Figure 2). For the 430 participants, there were 1255 total 6-week time periods of which 423 (33.7%) time periods were classified as low activity, 577 (46%) time periods as moderate activity, and 255 (20.3%) time periods as high activity. A total of 146 of 430 participants were in multiple clusters where they had completed more than 1 questionnaire. Overall, clusters were moderately defined with an average silhouette coefficient=0.54 (Table 2). The quality of the clustering was highest in the low activity cluster (silhouette coefficient=0.60), suggesting that participants had the most similarities for steps, distance, and moderate to vigorous duration within this cluster. Those in the high activity cluster had the most variance in their levels of exercise, as indicated by the lowest silhouette coefficient (0.48) and highest standard deviations for steps, distance, and minutes of moderate to vigorous duration compared to the other clusters. Sociodemographic characteristics varied across clusters, and those in the low activity cluster were older (P=.002), had higher BMIs (P<.001), and had longer disease duration (P=.02) when compared to the other clusters. Participants in the low activity cluster also had higher proportions of those taking corticosteroids (P<.001) but lower proportions of those taking immunomodulators (P=.04).

Figure 2. Physical activity clusters.

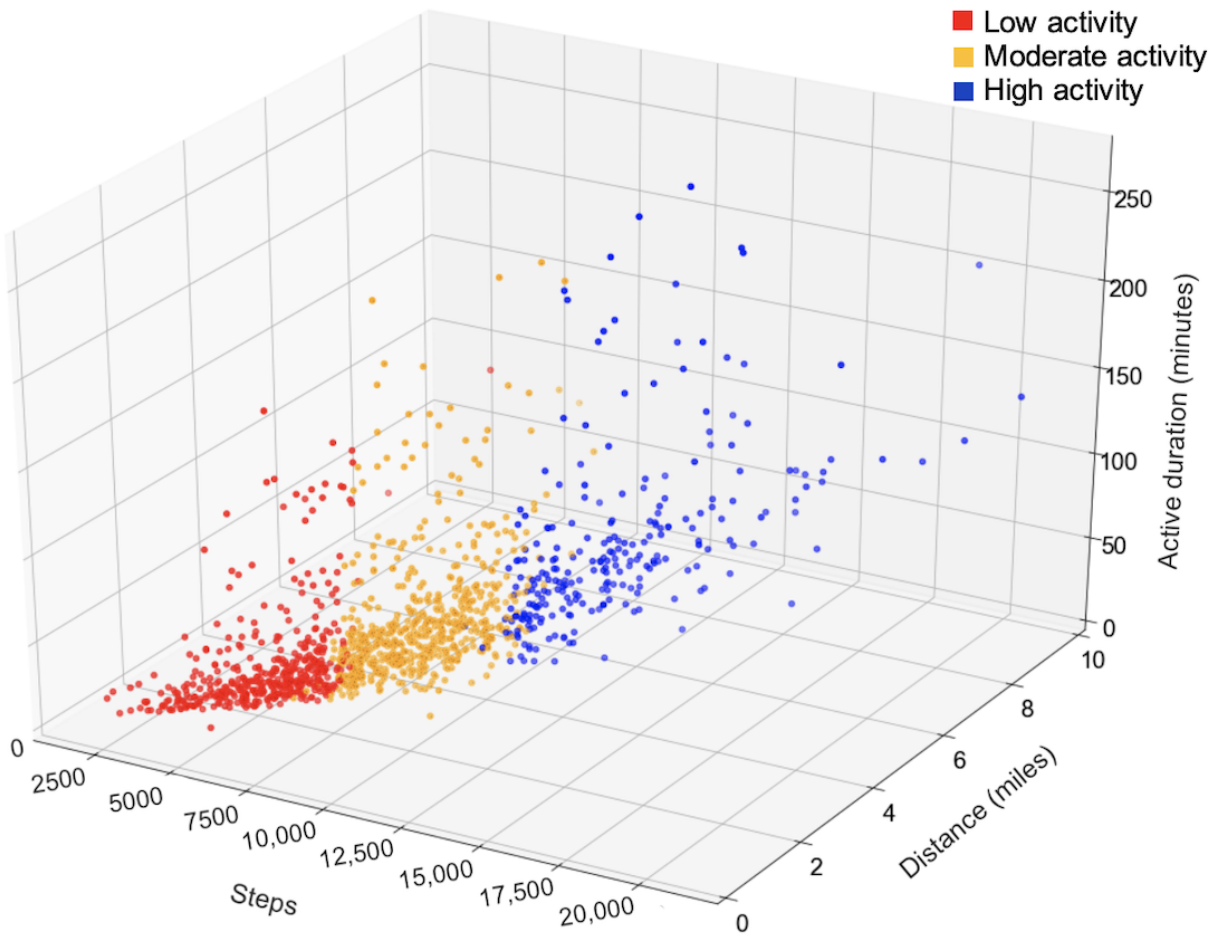

XSL•FO
RenderX

Table 2. Physical activity cluster profiles.

Characteristics	Low activity (n=423)	Moderate activity (n=577)	High activity (n=255)	*P* value
Evaluation				
Silhouette coefficient	0.60	0.52	0.48	__[a]
Features, mean (SD)				
Steps	5000.3 (1062.4)	8229.5 (1033.0)	12,319.4 (1899.6)	<.001
Distance (miles)	2.2 (0.6)	3.7 (0.7)	5.5 (1.1)	<.001
Moderate to vigorous duration (minutes)	21.3 (27.0)	40.6 (34.6)	73.0 (46.4)	<.001
Sociodemographics and clinical characteristics, mean (SD)				
Age (years)	45.8 (13.8)	43.5 (13.7)	42.1 (12.6)	.002
BMI (kg/m^2)	27.8 (3.5)	25.6 (4.5)	23.6 (3.5)	<.001
Duration of disease (years), mean (SD)	17.4 (12.3)	16.9 (12.0)	14.8 (10.0)	.02
Current IBD[b] treatment, n (%)				
5-Aminosalicylic acid medications	160 (37.8)	182 (31.5)	77 (30.2)	.06
Biologic therapies	252 (59.6)	328 (56.8)	141 (55.3)	.51
Corticosteroid medications	48 (11.3)	53 (9.2)	6 (2.4)	<.001
Immunomodulators	103 (24.3)	183 (31.7)	74 (29.0)	.04

[a]Not applicable.

[b]IBD: inflammatory bowel disease.

Association Between Physical Activity Clusters and PROs

Cross-Sectional Analysis

Across all PROs, patients in the low activity cluster had the worst scores, and those in the high activity cluster had the best scores (Table 3). Scores varied significantly across clusters on the level of depression (*P*=.004), pain interference (*P*<.001), fatigue (*P*<.001), sleep disturbance (*P*<.001), social satisfaction (*P*<.001), and SCDAI (*P*<.001). The largest mean differences between the low and high activity clusters for the PROMIS domains were scores for social satisfaction (5.0), fatigue (4.2), and pain interference (4.0). Mean SCCAI scores did not reach significance for patients with UC or IC, and those in the low and moderate activity clusters likely had some active disease activity during the 6-week period as scores ≥2.5 have previously demonstrated correlations with active disease [29]. Mean SCDAI scores were in remission (<150) in all clusters [27], and scores were highest in the low activity cluster.

Table 3. Patient-reported outcome scores across physical activity clusters.

	Low activity (n=423), mean (SD)	Moderate activity (n=577), mean (SD)	High activity (n=255), mean (SD)	*P* value
SCDAI[a,b]	133.3 (80.3)	117.0 (71.9)	102.2 (59.7)	<.001
SCCAI[a,c]	2.9 (2.1)	2.6 (1.8)	2.3 (2.5)	.20
PROMIS[d] anxiety	50.4 (9.2)	49.3 (8.9)	49.2 (8.9)	.16
PROMIS depression	48.6 (8.3)	47.1 (7.8)	46.9 (7.3)	.004
PROMIS pain interference	50.7 (9.5)	48.5 (8.4)	46.7 (7.6)	<.001
PROMIS fatigue	54.9 (10.7)	51.5 (10.9)	50.7 (9.6)	<.001
PROMIS sleep disturbance	51.2 (7.5)	49.4 (7.7)	49.2 (7.8)	<.001
PROMIS satisfaction with social roles and activities	50.3 (9.7)	54.0 (9.1)	55.3 (9.1)	<.001

[a]Missing data: SCDAI (short Crohn Disease Activity Index)=44; SCCAI (Simple Clinical Colitis Activity Index)=78 questionnaires.

[b]SCDAI: short Crohn Disease Activity Index.

[c]SCCAI: Simple Clinical Colitis Activity Index.

[d]PROMIS: Patient-Reported Outcomes Measurement Information System.

XSL•FO
RenderX

Longitudinal Analysis

Among the sample of 430 participants who contributed wearable data and completed at least one questionnaire, 246 (57.2%) participants completed at least 2 consecutive disease activity questionnaires. In this subset of 246 participants (726 total questionnaires), 67.8% (n=492) did not change physical activity clusters during 6-month periods (Table 4). As expected, there were no significant changes in mean disease activity scores among those who did not move between clusters. For those who transitioned into another cluster, 15.8% (n=115) moved between (to and from) moderate and high activity, 15.2% (n=110) moved between low and moderate activity, and only 1.2% (n=9) moved between low and high activity clusters. Proportions of cluster movement were similar for patients with CD and UC or IC. There were significant associations between cluster movement and mean disease activity score for the 3 types of transitions. When patients with UC or IC transitioned from low to moderate activity clusters, disease scores decreased (*P*=.04). For CD, when patients moved from moderate to high activity or high to moderate activity clusters, disease scores decreased (*P*=.04) or increased (*P*=.02), respectively.

Table 4. Movement across clusters for consecutive inflammatory bowel disease activity scores.

Change in physical activity and cluster movement	n	Change in disease activity, mean (SD)	*P* value
Ulcerative or indeterminate colitis (SCCAI[a])			
Improved			
Low → Moderate	21	–1.4 (2.4)	.04
Moderate → High	19	–0.4 (1.7)	.75
Low → High	0	N/A[b]	__[c]
Reduced			
Moderate → Low	25	0.1 (1.2)	.74
High → Moderate	17	0.1 (2.1)	.91
High → Low	4	3.2 (3.8)	__[c]
No change			
Low	51	0.1 (1.8)	.59
Moderate	76	–0.1 (1.3)	.55
High	38	–0.4 (1.8)	.18
Crohn disease (SCDAI[d])			
Improved			
Low → Moderate	26	–0.5 (82.8)	.34
Moderate → High	38	–25.3 (64.5)	.04
Low → High	1	0	__[c]
Reduced			
Moderate → Low	38	7.6 (63.6)	.88
High → Moderate	41	20.6 (55.5)	.02
High → Low	4	78.8 (111.8)	__[c]
No change			
Low	124	0.5 (67.9)	.94
Moderate	140	–6.3 (59.5)	.21
High	63	8.2 (69.8)	.35

[a]SCCAI: Simple Clinical Colitis Activity Index.

[b]Not applicable.

[c]Paired sample *t* tests not calculated for n<5.

[d]SCDAI: Short Crohn Disease Activity Index.

XSL•FO
RenderX

Discussion

Principal Results

This study demonstrates how physical activity clusters can be generated from consumer-based wearable devices in patients with IBDs. Most participants (46%, n=577) were clustered in moderate activity, 33.7% (n=423) as low activity, and 20.3% (n=255) as high activity. Sociodemographic and clinical characteristics varied across clusters, and those with low activity were older, had higher BMIs, and longer disease durations. Our findings indicate positive associations between physical activity and health-related quality of life PROs in accordance with existing research [14,37]. Individuals in the low activity cluster had the worst scores across all PROs, and scores varied meaningfully on levels of depression, pain interference, fatigue, sleep disturbance, social satisfaction, and CD activity across clusters. UC activity indices did not vary significantly across clusters, which may be due to individuals in the low activity and moderate activity clusters who had some degree of active disease. These disease indices were aligned with individual's IBD treatments, as there were higher proportions of those in the low activity cluster taking corticosteroid medications, which can be used as short-term treatment for IBD flares. Whereas, there were lower proportions of patients in the low activity cluster taking immunomodulators, which are usually used to maintain remission of IBD. We also found that those in the low physical activity cluster had the most homogeneity in exercise attributes, which suggests worse health status might hinder variations in exercise. These findings should be validated with other IBD cohorts to assess their reproducibility.

When we longitudinally assessed changes in physical activity and disease activity scores across 6-month periods, exercise patterns mostly did not fluctuate. Approximately 68% (n=492) of patients remained in their original cluster, and only 1% (n=9) of patients transitioned to or from the furthest clusters of low and high activity. This indicates that exercise levels may not vary to extremes over time. As expected, mean disease activity scores among patients who remained in the same cluster did not change over time. Despite prior literature suggesting the benefits of physical activity [1], the long 6-month longitudinal survey timepoints preclude any causal inferences. It is not possible to know, for example, whether increased physical activity levels reduced disease outcomes, or whether patients experiencing reduced symptoms were better able to exercise. Future studies that solicit more time points for symptom data from patients are needed to support this type of investigation.

Implications for Health Care and Research

This research has several implications for the use of wearable devices and PROs for patients with IBDs. The use of real-world data to identify phenotypes with similar activity attributes could be leveraged to develop interventions that promote self-management and coping abilities, as these are important components in managing IBDs [38,39]. Previous studies have used wearables to personalize behavioral coaching strategies, which resulted in improvements in physical activity and clinical biomarkers (eg, lipids, hemoglobin A_{1c} levels) [40,41]. In this study, the low activity group was characterized by short durations of moderate to vigorous activity, steps, and distance. Interventions effective in improving sedentary behavior often use established behavior change techniques, including goal setting, self-monitoring, or social support [42]. We found the greatest difference among PROMIS domains between the low and high activity groups for social satisfaction, and evidence suggests social support may improve psychological symptoms and self-management behaviors in patients with IBDs [38,43]. New models of patient-centered care have been proposed, such as the IBD specialty medical home, which involve multifaceted approaches focusing on social support, behavioral skills, and stress management techniques [44]. Providing outreach to patients with changes in health status or physical activity, which could be indicated by cluster transitions, could be an important aspect of personalized IBD care. Moreover, as new therapies are being developed, the ability to observe distinguishable physical activity phenotypes could reveal insights related to the impact of pharmacotherapies within clinical trials or at the point of care. However, if real-time data are used within clinical care or for just-in-time interventions, necessary protocols and validation strategies are needed to ensure accurate data are presented to patients and care teams in a meaningful and easily interpretable way [45]. Visual analytics and additional machine learning approaches are being developed and evaluated as part of the Precision VISSTA study.

Patient-generated health data also have promising potential to facilitate the detection of inflammation. Evidence suggests certain inflammatory responses may be able to be detected through physiological measurements or lifestyle characteristics from wearables, such as elevated heart rate, elevated skin temperature, or sleep deviations [46-48]. Very few studies have assessed the relationship between wearable physiological measurements and inflammatory responses in patients with IBDs [13,14]. Sossenheimer et al [13] found lower daily steps within the week prior to elevated inflammatory biomarkers (ie, C-reactive protein and fecal calprotectin), but did not find differences in resting heart rate. Wiestler et al [14] also found lower levels of physical activity in patients with active disease compared to those in remission, in addition to lower sleep efficiency. Given the vast amount of data from sensor devices, robust analytical pipelines are necessary to process, analyze, combine with other data streams, and derive actionable information from the data [49]. Unsupervised learning models, which do not require costly labeled data, are useful for partitioning large data sets into smaller groups of related information. These related subgroups have the potential to inform strategies for the detection or mitigation of inflammatory responses given the fluctuating symptoms and disease trajectories that vary by patient.

Limitations

The sample was limited to 430 participants in the CCF IBD Partners internet-based cohort who synced their smartphone app or wearable device, which may not be representative of all patients with IBDs. Future work should include larger and more diverse samples to improve the generalizability of these findings. In addition, the sample was more physically active than the general population, with our sample taking approximately 7900 steps per day. It is estimated that the US population takes 4800

XSL•FO
RenderX

steps per day (~5000 steps worldwide) [50]. There are also several limitations to using consumer-grade wearable device data. Most brands do not make the details of their algorithms or firmware updates available, so there may be differences in hardware or sensors across brands and devices over time. Innate user differences may also exist, such as the location the device was worn (dominant vs nondominant hand), wear time, or accuracy of manually logged exercise, which cannot be verified in the existing data. In addition, our clustering approach was limited by the lack of gradient or differentiation between individuals once clusters were established, so it is possible that individuals with similar activity patterns could be in different clusters. Finally, the sparse nature of the 6-month longitudinal survey timepoints precludes causal relationships.

Conclusions

Recognition of patterns and changes in lifestyle behaviors by leveraging real-world mobile health data supports opportunities to inform interventions. Unsupervised learning techniques facilitate the identification of these patterns from vast, multidimensional wearable data. Patients in the low physical activity cluster reported the worst health-related quality of life and disease activity (ie, depression, pain interference, fatigue, sleep disturbance, social satisfaction, and CD activity) compared to those in moderate and high activity clusters. Additional support for physical and psychosocial symptoms and exercise may be valuable for those low physical activity IBD subgroups. Future research should examine these findings among more diverse cohorts with more frequent PRO measurements and validate reproducibility.

Acknowledgments

Support for this research was provided by the National Institutes of Health (award R01EB025024) (principal investigator: AEC). This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ACG acknowledges funding support by the National Institutes of Health National Library of Medicine training grant (5T15LM012500-04) while at the University of North Carolina at Chapel Hill when this work was completed. ACG is currently supported by a VA Advanced Fellowship in Medical Informatics. The opinions expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or those of the US Government. No use of generative artificial intelligence was used in this manuscript.

FCL acknowledges funding support from the National Institutes of Health National Center for Advancing Translational Sciences (UM1TR004406).

Data Availability

The data sets analyzed during this study are not publicly available to protect the privacy of individuals who participated in the study. Data are available from the corresponding author upon reasonable request. IBD Partners supports research studies, and investigators interested in accessing the data and pursuing a study should register on their website [51].

Authors' Contributions

ACG and AEC are responsible for the overall study concept and design. All authors contributed to the study design, manuscript writing, and revision, and approved the final manuscript. ACG is primarily responsible for data analysis.

Conflicts of Interest

AEC is now employed by Google, LLC and has equity in the company. However, this work was completed prior to employment at Google when AEC was at the University of North Carolina at Chapel Hill School of Medicine. No resources or funding were provided by Google for this research.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[PDF File (Adobe PDF File), 117 KB - jmir_v26i1e48020_app1.pdf]

References

1. Bilski J, Brzozowski B, Mazur-Bialy A, Sliwowski Z, Brzozowski T. The role of physical exercise in inflammatory bowel disease. Biomed Res Int 2014;2014:429031 [FREE Full text] [doi: 10.1155/2014/429031] [Medline: 24877092]
2. Subramanian CR, Triadafilopoulos G. Care of inflammatory bowel disease patients in remission. Gastroenterol Rep (Oxf) 2016;4(4):261-271 [FREE Full text] [doi: 10.1093/gastro/gow032] [Medline: 27899522]
3. Kochar B, Martin C, Kappelman M, Spiegel BM, Chen W, Sandler RS, et al. Evaluation of Gastrointestinal Patient Reported Outcomes Measurement Information System (GI-PROMIS) symptom scales in subjects with inflammatory bowel diseases. Am J Gastroenterol 2018;113(1):72-79. [doi: 10.1038/ajg.2017.240] [Medline: 28853727]

XSL•FO
RenderX

4. Kappelman MD, Long MD, Martin C, DeWalt DA, Kinneer PM, Chen W, et al. Evaluation of the patient-reported outcomes measurement information system in a large cohort of patients with inflammatory bowel diseases. Clin Gastroenterol Hepatol 2014;12(8):1315-1323 [FREE Full text] [doi: 10.1016/j.cgh.2013.10.019] [Medline: 24183956]
5. Eckert KG, Abbasi-Neureither I, Köppel M, Huber G. Structured physical activity interventions as a complementary therapy for patients with inflammatory bowel disease—a scoping review and practical implications. BMC Gastroenterol 2019;19(1):115 [FREE Full text] [doi: 10.1186/s12876-019-1034-9] [Medline: 31266461]
6. Bilski J, Mazur-Bialy A, Brzozowski B, Magierowski M, Zahradnik-Bilska J, Wójcik D, et al. Can exercise affect the course of inflammatory bowel disease? Experimental and clinical evidence. Pharmacol Rep 2016;68(4):827-836. [doi: 10.1016/j.pharep.2016.04.009] [Medline: 27255494]
7. Ng V, Millard W, Lebrun C, Howard J. Low-intensity exercise improves quality of life in patients with Crohn's disease. Clin J Sport Med 2007;17(5):384-388. [doi: 10.1097/JSM.0b013e31802b4fda] [Medline: 17873551]
8. Klare P, Nigg J, Nold J, Haller B, Krug AB, Mair S, et al. The impact of a ten-week physical exercise program on health-related quality of life in patients with inflammatory bowel disease: a prospective randomized controlled trial. Digestion 2015;91(3):239-247. [doi: 10.1159/000371795] [Medline: 25823689]
9. Gatt K, Schembri J, Katsanos KH, Christodoulou D, Karmiris K, Kopylov U, et al. Inflammatory bowel disease [IBD] and physical activity: a study on the impact of diagnosis on the level of exercise amongst patients with IBD. J Crohns Colitis 2019;13(6):686-692. [doi: 10.1093/ecco-jcc/jjy214] [Medline: 30561568]
10. Ho GWK. Lower gastrointestinal distress in endurance athletes. Curr Sports Med Rep 2009;8(2):85-91. [doi: 10.1249/JSR.0b013e31819d6b7b] [Medline: 19276909]
11. The wearable life 2.0: connected living in a wearable world. PWC. 2018. URL: https://www.pwc.com/ee/et/publications/pub/pwc-cis-wearables.pdf [accessed 2020-01-10]
12. About one-in-five Americans use a smart watch or fitness tracker. Pew Research Center. URL: https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/ [accessed 2024-07-24]
13. Sossenheimer PH, Yvellez OV, Andersen M, Pearl T, Jurdi E, Rubin DB, et al. Wearable devices can predict disease activity in inflammatory bowel disease patients. Gastroenterology 2019;156(6):S-111. [doi: 10.1016/s0016-5085(19)37069-6]
14. Wiestler M, Kockelmann F, Kück M, Kerling A, Tegtbur U, Manns MP, et al. Quality of life is associated with wearable-based physical activity in patients with inflammatory bowel disease: a prospective, observational study. Clin Transl Gastroenterol 2019;10(11):e00094 [FREE Full text] [doi: 10.14309/ctg.0000000000000094] [Medline: 31770137]
15. Torous J, Kiang MV, Lorme J, Onnela J. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
16. Marschollek M. A semi-quantitative method to denote generic physical activity phenotypes from long-term accelerometer data—the ATLAS index. PLoS One 2013;8(5):e63522 [FREE Full text] [doi: 10.1371/journal.pone.0063522] [Medline: 23667631]
17. Lee I, Sesso HD, Oguma Y, Paffenbarger RS. The "weekend warrior" and risk of mortality. Am J Epidemiol 2004;160(7):636-641. [doi: 10.1093/aje/kwh274] [Medline: 15383407]
18. Metzger JS, Catellier DJ, Evenson KR, Treuth MS, Rosamond WD, Siega-Riz AM. Patterns of objectively measured physical activity in the United States. Med Sci Sports Exerc 2008;40(4):630-638. [doi: 10.1249/MSS.0b013e3181620ebc] [Medline: 18317384]
19. McConnell MV, Shcherbina A, Pavlovic A, Homburger JR, Goldfeder RL, Waggot D, et al. Feasibility of obtaining measures of lifestyle from a smartphone app: the myheart counts cardiovascular health study. JAMA Cardiol 2017;2(1):67-76. [doi: 10.1001/jamacardio.2016.4395] [Medline: 27973671]
20. Koffman LJ, Crainiceanu CM, Roemmich RT, French MA. Identifying unique subgroups of individuals with stroke using heart rate and steps to characterize physical activity. J Am Heart Assoc 2023;12(18):e030577 [FREE Full text] [doi: 10.1161/JAHA.123.030577] [Medline: 37681556]
21. Verstockt B, Noor N, Marigorta U, Pavlidis P, Deepak P, Ungaro RC, Scientific Workshop Steering Committee. Results of the seventh scientific workshop of ECCO: precision medicine in IBD-disease outcome and response to therapy. J Crohns Colitis 2021;15(9):1431-1442 [FREE Full text] [doi: 10.1093/ecco-jcc/jjab050] [Medline: 33730756]
22. VISSTA: Visualization & statistical approaches for PGHD to enable precision medicine. URL: https://github.com/PrecisionVISSTA/ [accessed 2022-02-01]
23. Long MD, Kappelman MD, Martin CF, Lewis JD, Mayer L, Kinneer PM, et al. Development of an internet-based cohort of patients with inflammatory bowel diseases (CCFA partners): methodology and initial results. Inflamm Bowel Dis 2012;18(11):2099-2106. [doi: 10.1002/ibd.22895] [Medline: 22287300]
24. Chung AE, Sandler RS, Long MD, Ahrens S, Burris JL, Martin CF, et al. Harnessing person-generated health data to accelerate patient-centered outcomes research: the Crohn's and Colitis foundation of America PCORnet patient powered research network (CCFA partners). J Am Med Inform Assoc 2016;23(3):485-490 [FREE Full text] [doi: 10.1093/jamia/ocv191] [Medline: 26911821]

XSL•FO
RenderX

25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
26. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63(11):1179-1194 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
27. Thia K, Faubion WA, Loftus EV, Persson T, Persson A, Sandborn WJ. Short CDAI: development and validation of a shortened and simplified Crohn's disease activity index. Inflamm Bowel Dis 2011;17(1):105-111. [doi: 10.1002/ibd.21400] [Medline: 20629100]
28. Jowett SL, Seal CJ, Phillips E, Gregory W, Barton JR, Welfare MR. Defining relapse of ulcerative colitis using a symptom-based activity index. Scand J Gastroenterol 2003;38(2):164-171. [doi: 10.1080/00365520310000654] [Medline: 12678333]
29. Higgins PDR, Schwartz M, Mapili J, Krokos I, Leung J, Zimmermann EM. Patient defined dichotomous end points for remission and clinical improvement in ulcerative colitis. Gut 2005;54(6):782-788 [FREE Full text] [doi: 10.1136/gut.2004.056358] [Medline: 15888785]
30. Tukey JW. Exploratory Data Analysis (Addison-Wesley Series in Behavioral Science). 1st Edition. London, England: Pearson; 1977.
31. Lloyd S. Least squares quantization in PCM. IEEE Trans Inform Theory 1982;28(2):129-137. [doi: 10.1109/tit.1982.1056489]
32. Tan P, Steinbach M, Kumar V. Introduction to Data Mining. Pearson Addison Wesley: Boston, MA; 2005.
33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
35. Arthur D, Vassilvitskii S. K-Means++: the advantages of careful seeding. 2007 Presented at: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms; January 7-9, 2007; New Orleans, Louisiana, USA p. 1027-1035 URL: http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf
36. Reedy J, Wirfält E, Flood A, Mitrou PN, Krebs-Smith SM, Kipnis V, et al. Comparing 3 dietary pattern methods—cluster analysis, factor analysis, and index analysis—with colorectal cancer risk: The NIH-AARP diet and health study. Am J Epidemiol 2010;171(4):479-487 [FREE Full text] [doi: 10.1093/aje/kwp393] [Medline: 20026579]
37. Jones PD, Kappelman MD, Martin CF, Chen W, Sandler RS, Long MD. Exercise decreases risk of future active disease in patients with inflammatory bowel disease in remission. Inflamm Bowel Dis 2015;21(5):1063-1071. [doi: 10.1097/MIB.0000000000000333] [Medline: 25723616]
38. Kamp KJ, West P, Holmstrom A, Luo Z, Wyatt G, Given B. Systematic review of social support on psychological symptoms and self-management behaviors among adults with inflammatory bowel disease. J Nurs Scholarsh 2019;51(4):380-389. [doi: 10.1111/jnu.12487] [Medline: 31119856]
39. Conley S, Redeker N. A systematic review of self-management interventions for inflammatory bowel disease. J Nurs Scholarsh 2016;48(2):118-127 [FREE Full text] [doi: 10.1111/jnu.12189] [Medline: 26756193]
40. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol 2017;35(8):747-756 [FREE Full text] [doi: 10.1038/nbt.3870] [Medline: 28714965]
41. Van Hoye K, Boen F, Lefevre J. The impact of different degrees of Feedback on physical activity levels: a 4-week intervention study. Int J Environ Res Public Health 2015;12(6):6561-6581 [FREE Full text] [doi: 10.3390/ijerph120606561] [Medline: 26067990]
42. Stephenson A, McDonough SM, Murphy MH, Nugent CD, Mair JL. Using computer, mobile and wearable technology enhanced interventions to reduce sedentary behaviour: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2017;14(1):105 [FREE Full text] [doi: 10.1186/s12966-017-0561-4] [Medline: 28800736]
43. Regueiro M, Greer JB, Szigethy E. Etiology and treatment of pain and psychosocial issues in patients with inflammatory bowel diseases. Gastroenterology 2017;152(2):430-439.e4. [doi: 10.1053/j.gastro.2016.10.036] [Medline: 27816599]
44. Regueiro MD, McAnallen SE, Greer JB, Perkins SE, Ramalingam S, Szigethy E. The inflammatory bowel disease specialty medical home: a new model of patient-centered care. Inflamm Bowel Dis 2016;22(8):1971-1980. [doi: 10.1097/MIB.0000000000000819] [Medline: 27135486]
45. Chung AE, Basch EM. Potential and challenges of patient-generated health data for high-quality cancer care. J Oncol Pract 2015;11(3):195-197 [FREE Full text] [doi: 10.1200/JOP.2015.003715] [Medline: 25852139]
46. Li X, Dunn J, Salins D, Zhou G, Zhou W, Schüssler-Fiorenza Rose SM, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLoS Biol 2017;15(1):e2001402 [FREE Full text] [doi: 10.1371/journal.pbio.2001402] [Medline: 28081144]
47. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health 2020;2(2):e85-e93 [FREE Full text] [doi: 10.1016/S2589-7500(19)30222-5] [Medline: 33334565]

XSL•FO
RenderX

48. Gadaleta M, Radin JM, Baca-Motes K, Ramos E, Kheterpal V, Topol EJ, et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. NPJ Digit Med 2021;4(1):166 [FREE Full text] [doi: 10.1038/s41746-021-00533-1] [Medline: 34880366]
49. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018;15(5):429-448. [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
50. Althoff T, Sosič R, Hicks JL, King AC, Delp SL, Leskovec J. Large-scale physical activity data reveal worldwide activity inequality. Nature 2017;547(7663):336-339 [FREE Full text] [doi: 10.1038/nature23018] [Medline: 28693034]
51. IBD Partners. 2023. URL: https://ccfa.med.unc.edu/ [accessed 2024-08-01]

Abbreviations

CCF: Crohn's and Colitis Foundation
CD: Crohn disease
IBD: inflammatory bowel disease
IC: indeterminate colitis
Precision VISSTA: Precision Visualization and Statistical Approaches
PRO: patient-reported outcome
PROMIS: Patient-Reported Outcomes Measurement Information System
SCCAI: Simple Clinical Colitis Activity Index
SCDAI: Short Crohn Disease Activity Index
SSE: sum of squared errors
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UC: ulcerative colitis

Edited by T de Azevedo Cardoso; submitted 08.04.23; peer-reviewed by B Senst, M Elbattah, T Orlemann; comments to author 11.03.24; revised version received 05.06.24; accepted 03.07.24; published 24.09.24.

Please cite as:
Griffin AC, Mentch L, Lin FC, Chung AE
mHealth Physical Activity and Patient-Reported Outcomes in Patients With Inflammatory Bowel Diseases: Cluster Analysis
J Med Internet Res 2024;26:e48020
URL: https://www.jmir.org/2024/1/e48020
doi:10.2196/48020
PMID:

©Ashley C Griffin, Lucas Mentch, Feng-Chang Lin, Arlene E Chung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mobile Gaming App to Train Teenage Mothers on Appropriate Child Feeding Practices: Development and Validation Study

Mercy Eloho Sosanya[1,2], PhD; Folake Olukemi Samuel[3], PhD; Sadia Bashir[4], MS; Victoria Osariemen Omoera[5], MBA, MEd; Jeanne H Freeland-Graves[1], PhD

[1]Department of Nutritional Sciences, University of Texas at Austin, Austin, TX, United States
[2]Department of Nutrition and Dietetics, The Federal Polytechnic, Bauchi, Bauchi, Nigeria
[3]Department of Human Nutrition and Dietetics, College of Medicine, University of Ibadan, Ibadan, Nigeria
[4]PixelArt Games Academy, Islamabad, Pakistan
[5]Centre for Family Health Initiative, Abuja, Nigeria

Corresponding Author:
Mercy Eloho Sosanya, PhD
Department of Nutritional Sciences
University of Texas at Austin
200 W 24th Street
Austin, TX, 78712-1247
United States
Phone: 1 512 967 2026
Email: mercysosanya@utexas.edu

Abstract

Background: Undernutrition is an underlying factor in nearly 50% of 1 million estimated annual deaths among Nigerian children aged <5 years. Inappropriate maternal infant and young child feeding (IYCF) practices are basic contributors to child undernutrition. Teenage motherhood exacerbates the problem of inadequate child feeding. One possible intervention method to improve IYCF knowledge and practices of teenage mothers is the use of mobile gaming technologies. Despite extreme poverty in low- and middle-income countries, a ubiquity of mobile phone networks exists.

Objective: This study aims to develop and validate a mobile gaming app, called BabyThrive, to train Nigerian teenage mothers on appropriate IYCF practices.

Methods: To identify gaps in current IYCF practices in northern Nigeria, we conducted an extensive search of the literature and held 2 focus group interviews with 16 teenage mothers with low-income status. An initial app content design was then created, and content validity was established by 10 nutrition experts. Next, we developed an app prototype, which was assessed for quality by 7 nutrition and mobile gaming experts and evaluated for usability by 90 teenage mothers from rural areas in Abuja, the country's capital. The final app, BabyThrive, is a 2D mobile game that is fully functional offline and available in English as well as Hausa, which is commonly spoken in northern Nigeria. The efficacy of the BabyThrive app was assessed using IYCF knowledge scores obtained from the administration of the validated Teen Moms Child Feeding Questionnaire for Sub-Saharan Africa. Construct validity was established via crossover design by comparing the total IYCF knowledge scores of the teenage mothers obtained after a verbal training program and BabyThrive app use.

Results: Large proportions of the study participants were married (53/90, 59%) and had no personal income (63/90, 70%). The mean quality rating for the BabyThrive app was 4.3 (SD 0.39) out of 5.0. High levels (>80%) of usability and user satisfaction were documented. Knowledge of exclusive breastfeeding (P<.001) and total knowledge scores (P=.002) were significantly higher in the BabyThrive group than in the verbal training group. The IYCF knowledge scores obtained from both groups showed coherence, with a statistically significant Spearman correlation coefficient of 0.50 (P<.001).

Conclusions: This research developed and validated a novel, offline mobile gaming app. It will be an easy, effective, and acceptable method to disseminate critical knowledge on IYCF practices to teenage mothers in rural Nigeria.

(J Med Internet Res 2024;26:e53560) doi:10.2196/53560

KEYWORDS

mobile health; mHealth; mobile gaming app; validation; infant and young child feeding; teenage mother; Nigeria; mobile phone

XSL•FO
RenderX

Introduction

Background

Adolescent girls (aged 15-19 y) in low- and middle-income countries account for almost 10% of 227 million global pregnancies, resulting in 12 million births each year [1-3]. Globally, sub-Saharan Africa has the highest percentage (46%) of adolescent girls bearing children [4]. Nigeria contributes to this high prevalence of teenage motherhood, with 38.3% of girls aged ≤19 years in rural areas having children [5]. Teenage mothers are generally undernourished and less educated, which often results in severely underweight infants who experience malnutrition [6,7]. In Nigeria, the prevalence of malnutrition in children aged 0 to 5 years is substantial: 37% are diagnosed with stunting, 18% with wasting, and 29% with underweight [8]. Furthermore, undernutrition is an underlying factor in nearly 50% of 1 million estimated annual deaths in Nigerian children aged <5 years [9,10]. In southwestern Nigeria, the situation is even worse among children born to teenage mothers: 73.4% were diagnosed with wasting, stunting, or underweight [11]. In northern Nigeria, children born to teenage mothers were 3 times more likely to die within the first 5 years after birth compared with children born to older mothers [12].

One basic contributor to this child undernutrition is inappropriate maternal infant and young child feeding (IYCF) practices [13]. A multiplicity of interventions have attempted to improve IYCF practices [14], but the results have been dismal. Less than one-third of children (29%) are exclusively breastfed for 6 months, only 42% of newborns are breastfed within 1 hour of birth, and almost half (49%) receive prelacteal feeds [10]. Similarly, less than a quarter of children aged 6 to 23 months receive meals of adequate diversity (at least 5 food groups included), and only 42% are fed the appropriate number of times for their age [10]. Even fewer children (11%) aged 6 to 23 months receive the minimum acceptable diet (a composite indicator of minimum dietary diversity and minimum meal frequency) [10].

The problem of inadequate child feeding is exacerbated by teenage motherhood. Previous studies have shown unacceptably low rates of early initiation, duration, and exclusivity of breastfeeding in this age group [15]. In comparison with adult mothers, poor complementary feeding practices are more frequent. These practices include improper timing of the initiation of complementary foods, less responsive feeding styles, and feeding of diets low in protein and high in fat and sugars [16]. In Nigeria, a 15-year trend analysis has documented consistently lower rates of timely breastfeeding initiation (43.7% vs 55%) and exclusive breastfeeding (24.8% vs 28.6%) among adolescent versus adult mothers, as well as higher rates of the use of prelacteal feeds (59.2% vs 49.7%) [17].

In northern Nigeria, IYCF interventions have been implemented by a collaboration of state governments and international development partners. These behavioral change interventions have been delivered to mothers via group and individualized counseling at health facilities by health workers, volunteer-led community counseling for mothers, and community meetings for fathers and grandmothers [18]. A qualitative evaluation of these IYCF programs identified numerous problems, including heavy work burden for health workers due to personnel shortages, limited financial remuneration for community volunteers, time constraints for personnel and participants, and challenges with transportation [18]. All these issues are potent threats to the viability and effectiveness of IYCF programs [18].

One possible intervention method to improve IYCF knowledge and practices of teenage mothers is the use of mobile health (mHealth) for the achievement of health objectives [19]. mHealth interventions can target a wider client base, with decreased costs, time, need for space, and requirements for personnel training [20]. Despite extreme poverty in low- and middle-income countries, a ubiquity of mobile phone networks exists. In fact, this network surpasses other infrastructure, including paved roads and electricity [19]. Nigeria, with a population of approximately 213 million, has >190 million mobile phone lines and approximately 105 million monthly mobile internet lines [21-23]. In African countries, including Nigeria, mothers who participated in IYCF interventions delivered via mobile voice and SMS text messages were 1.2 to 2.6 times more likely to practice exclusive and timely initiation of breastfeeding [24-26].

The Focus of This Research

This research focuses on mobile games as a means of a nutrition intervention. Mobile games can be effective because they have the capacity to modify the attitudes, beliefs, behaviors, knowledge, and skills of users pertaining to health while providing entertainment [27,28]. These video games are designed to entertain users while eliciting changes in behavior and are often referred to as serious games [29]. Serious mobile games are useful in surmounting health challenges such as limited access to health information and may transcend social barriers, including low literacy and gender biases [30]. These games have shown promising results in a variety of health applications, including the improvement of treatment compliance, the rehabilitation of individuals with disabilities, obesity management, the promotion of healthy lifestyles, training parents on vegetable feeding practices for toddlers, and the improvement of knowledge and intention regarding healthy behaviors among older children [29,31,32]. Although gamification has been extensively used in numerous exercise and fitness apps, few apps have focused on changing maternal IYCF behaviors using this technique [33]. To date, no studies have explored the application of mobile gaming techniques to facilitate IYCF behavior change in Africa. Therefore, this study developed and validated a mobile gaming app, called BabyThrive, to train teenage mothers on appropriate IYCF practices.

Methods

App Development

The mobile gaming app, BabyThrive, was developed through an iterative and incremental process. Figure 1 shows the flowchart for the development and validation of the app. We conducted an extensive literature search and held 2 focus group discussions with 16 teenage mothers with low-income status to identify gaps in current IYCF practices in northern Nigeria and

XSL•FO
RenderX

effective behavior change techniques (BCTs) suitable for the study population. Guided by the findings from the literature search and focus group discussions, an initial app content design was created based on information adapted from the United Nations Children's Fund (UNICEF) IYCF training package for Nigeria [34]. Permission to adapt content and images from the Community Infant and Young Child Feeding Key Messages Booklet (part of the training package) was obtained from the Federal Ministry of Health and Social Welfare in Abuja, the country's capital.

The validity of the nutrition content of the initial app design was evaluated by a panel of 10 nutrition experts from family health, nutrition science, and dietetics departments at the University of Texas at Austin, Texas; the University of Ibadan, Ibadan, Nigeria; the Bauchi State Agricultural Development Program, Nigeria; and Nigeria's Federal Ministry of Health in Abuja. Game art concepts, stories, dialogues, and user interface and user experience design were generated from the initial app content design, along with feedback from the panel of nutrition experts. These were incorporated into an app prototype developed using the Agile methodology [35] solely for the Android platform (the predominant mobile phone platform used in the study area) [36]. The app prototype underwent a series of in-house tests and correction of errors and then was evaluated for quality by 7 nutrition and mobile gaming experts using the Mobile App Rating Scale (MARS) [37]. These nutrition and gaming experts were recruited from the University of Illinois at Urbana-Champaign, Illinois; Federal Polytechnic Bauchi, Bauchi, Nigeria; Nigeria's Federal Ministry of Health in Abuja; Kano State College of Health Sciences and Technology, Nigeria; Endoclan Applications (a game development firm), Port Harcourt, Nigeria; Abubakar Tafawa Balewa University, Bauchi, Nigeria; and GitLab Inc. Each quality component on the MARS was rated on a 5-point Likert scale ranging from 1=*inadequate* to 5=*excellent*, with a total possible score of 5. After modification based on suggestions from the experts, the app was assessed for comprehensibility and ease of use by a focus group of 8 teenage mothers in Nigeria. The feedback was integrated into the final version of the app.

The study has been registered with ClinicalTrials.gov (NCT05181293).

Figure 1. Flowchart for the development and validation of the BabyThrive app.

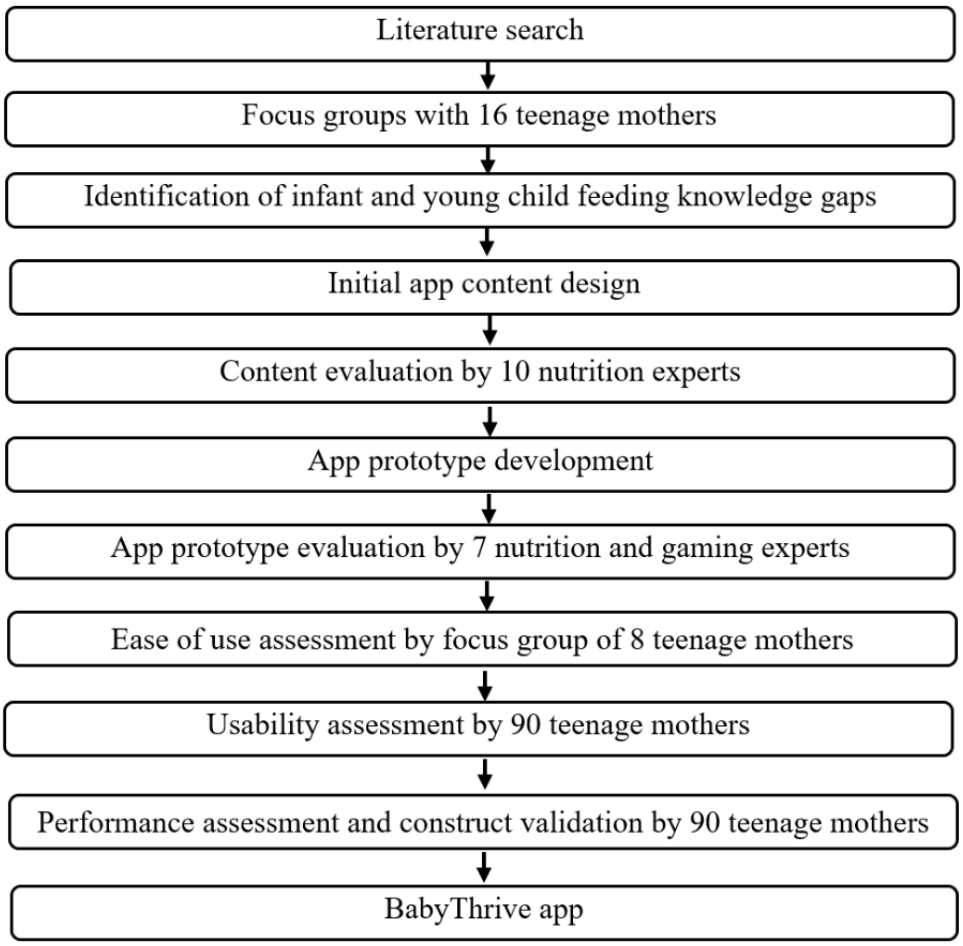

App Description

Overview

The final app, BabyThrive, is a 2D character-animated mobile game featuring voice-acted audio dialogues. Multimedia Appendix 1 shows the app interface. The BabyThrive app is fully functional offline and provides 2 language options: English and Hausa (commonly spoken in northern Nigeria). The app contains 4 modules, which are described in the following subsections.

XSL•FO
RenderX

Interactive Game

This role-play video game with 5 levels, in which the player (a teenage mother) must learn, practice, and execute correct IYCF recommendations and decisions regarding her child's health, provides detailed narratives concerning breastfeeding, including the benefits of exclusive breastfeeding to the child, mother, and community, as well as tips for successful breastfeeding, including techniques, position, attachment, the use of colostrum, concepts of foremilk and hindmilk, and breast milk expression. The game narratives and dialogues also cover complementary feeding concepts, including age-appropriate amounts, frequency and food group diversity, hygiene, responsiveness, and child feeding and care in illness. As young Nigerian mothers play the game, they identify with the game characters and interactively acquire knowledge to improve their practices and attitudes toward IYCF practices. To encourage active engagement in the learning process, BabyThrive uses simple gaming mechanics, including treasure hunts to find animated, informational cards; the dragging and dropping of matching objects; and the fixing of puzzles. These simple mechanics were used because evidence shows that extremely complex digital interventions are less effective [38]. In addition, at the end of each level, an interactive quiz is administered to ensure knowledge retention. Game incentives are provided as game levels and quizzes are completed successfully.

The mechanisms of behavior change for this app were derived from social cognitive theory [39]. The social cognitive theory constructs of social modeling (game characters), mastery experience (in-app practice of IYCF recommendations), verbal persuasion (voice-acted narratives), reinforcement (rewards), facilitation (in-game modeling of support by health personnel and influential family members), and emotional state improvement (relaxing game) were used to influence dynamic and reciprocal relationships between IYCF behaviors, personal factors (knowledge, self-efficacy, and child feeding skills), and environmental factors (social support).

Recipe Guide

This module presents 30 nutritionally adequate recipes derived from locally available and affordable foods [40]. Mothers are prompted to prepare the meals and take photographs of their children's meals to be stored in the app's database for subsequent nutritional evaluation by researchers.

Child Health Planner

This planner includes a calendar with reminders for available health services (birth registration, well-child clinics, immunization, deworming, vitamin A supplementation, and malaria prophylaxis) that are poorly used to date. Mothers were prompted to access these services, document activities, and upload photos of paperwork on the app to evaluate compliance.

Child Growth Monitor

This monitor is based on algorithms from the World Health Organization Anthro software [41]. It documents measurements of the weight and height of the child taken during child welfare clinic visits. The app plots child growth patterns using age and anthropometric indicators. If needed, it recommends corrective actions, such as more frequent meals, recipes with increased nutrient density, or clinic visits.

BCTs Used

Six groups of BCTs [42] were used in the BabyThrive app: (1) shaping knowledge, (2) natural consequences, (3) comparison of behavior, (4) repetition and substitution, (5) reward and threat, and (6) feedback and monitoring; for instance, under shaping knowledge, the BCT of instruction on how to perform the behavior was implemented as detailed steps to successful breastfeeding and complementary feeding. The depiction of BCTs relating to natural consequences includes (1) Information about health consequences (in-app description of the effects of commercial infant feeds or mixed feeding), (2) salience of consequences (the benefits of exclusive breastfeeding, including reduction of risk for postpartum bleeding as well as breast and ovarian cancers), and (3) anticipated regret (the possible effects of nonexclusive breastfeeding on the child, such as increased risks of diarrhea, infections, and future chronic illnesses). A BCT relating to comparison of behaviors was demonstration of the behavior, implemented in the app via animated demonstrations of breastfeeding attachment and positions as well as breast milk expression. Behavioral practice and rehearsal (in-game interactive quizzes) and graded tasks (graded quiz responses that must be repeated until mastery is achieved) were 2 BCTs used under repetition and substitution. Reward and threat (outcome rewards) were operationalized by earning points for correct quiz responses. The BCTs concerning feedback and monitoring were self-monitoring (in-app documentation of the child's anthropometric measurements by the mother) and feedback on the outcome or outcomes of behavior (in-app visual feedback related to the child's anthropometric measurements and suggestions for corrective actions).

App Validation

Previous studies have indicated that evaluations of mobile apps should encompass an analysis of their content, an assessment of usability and user satisfaction, and the testing of efficacy to assess whether there is a reliable relationship between the effects of the app and outcomes from other methods [43-46]. Content evaluation was conducted during the app development phase by nutrition experts. App usability and user satisfaction were assessed by a different group of 90 teenage mothers (aged 14-19 y) using the mHealth App Usability Questionnaire [47]. These mothers had at least 1 child aged <2 years, and they were recruited from rural communities (Sabon-Lugbe, Karmo, Angwan Sayawa, Kagini, Gosa, Iddo Pada, Dape, Kubwa, and Kabusa) in the suburbs of Abuja. The study was publicized in the target area through village heads and local information channels. Potential participants were assessed for eligibility. Eligible individuals were informed about the research procedures, risks, benefits, and freedom to opt out of the study at any time and were enrolled after they provided informed consent. The mHealth App Usability Questionnaire consists of 18 usability statements rated on a 7-point Likert scale from 1=*very strongly disagree* to 7=*very strongly agree*, with a total possible score of 7 per statement. Very few study participants (0/90, 0%-3/90, 3%) chose the *very strongly disagree* response; thus, this option was removed during analysis, and responses

XSL•FO
RenderX

under this option were merged with the *strongly disagree* category. The final version of the BabyThrive app was validated by crossover design in the teenage mothers, who were required to have access to a smartphone. The crossover design was used to achieve a high statistical power, minimize the effect of interparticipant variability, and facilitate a comparison of within-participant effects (because each study participant serves as their own control) [48]. The efficacy of the app in disseminating IYCF knowledge was assessed by IYCF knowledge scores obtained from interviewer administration of the validated Teen Moms Child Feeding Questionnaire for Sub-Saharan Africa [49]. Only the knowledge scale was used; we did not use the attitude scale because the time period was considered too short to be able to observe notable changes in attitude [50,51]. The knowledge scale contains 23 questions under 3 subscales (food type, exclusive breastfeeding, and complementary feeding). The IYCF knowledge of the mothers was tested before the training, and then the 90 teenage mothers were randomly assigned by drawing lots into 2 groups with a 1:1 allocation ratio (verbal IYCF training group: n=45, 50%; BabyThrive group: n=45, 50%). Over the course of 2 days, 1 group participated in a 4-hour session of verbal IYCF training, with content adapted from the UNICEF IYCF training package for Nigeria. Simultaneously, the app was installed by research assistants on the mobile phones of participants in the second group. These respondents were trained briefly on the use of the BabyThrive app and asked to use it for 7 days due to the possibility of power failures. Each participant in both groups then received the same assessment questionnaire. After a washout period of 1 week [52,53], participants were crossed over to alternate groups to receive either the verbal IYCF training or the app. The assessment questionnaire was then administered a third time to each participant. Each participant remained in the study for approximately 3 weeks. App efficacy was determined by comparing IYCF knowledge before the training and after verbal training and BabyThrive app use. Construct validity [54,55] was established for the BabyThrive app by examining the relationship of total IYCF knowledge scores of the teenage mothers obtained after verbal training and BabyThrive app use.

Statistical Analyses

Statistical analysis was performed using R software (version 4.2.2; R Foundation for Statistical Computing) and the RStudio integrated development environment (version 2022.07.2+554; Posit Software, PBC) [56]. Content validity was established by a panel of 10 nutrition experts. App quality rating by experts was analyzed via descriptive statistics (mean and quartiles) of Likert responses to the MARS. App usability by the teenage mothers was described by the frequencies of Likert responses to the mHealth App Usability Questionnaire. App efficacy was assessed by (1) the proportions of correct responses to the IYCF knowledge questions, with differences between the groups determined by a chi-square test; (2) mean IYCF knowledge scores for questionnaire subscales, with mean differences between the training groups determined by 1-way ANOVA and post hoc analysis by the Tukey highest significant difference test; and (3) differences in the effects of the training groups, explored using generalized estimating equations. Generalized estimating equations are an extension of the standard linear regression that takes correlations between repeated measures into consideration in analyzing longitudinal or clustered data [57,58]. Carryover effects were tested by analyzing differences in the effects of the order of group participation. Construct validity was established through Spearman correlation of the total knowledge scores between the verbal training and BabyThrive groups, with a coefficient ranging from 0.40 to 0.80 considered appropriate [54,55], and by a graphical illustration of the agreement between both groups via a Bland-Altman plot. A Bland-Altman plot visually represents the relationship between paired variables on the same measurement scale [59]. The Bland-Altman statistics were computed from the mean differences in knowledge scores as a function of the mean knowledge score from both groups. Three outlying data points were excluded from this analysis. To further explore the agreement between the verbal training and BabyThrive app groups, the mean absolute difference was computed from the differences in knowledge scores between both groups. Statistical significance for all tests was set at $P<.05$.

Ethical Considerations

Ethics approval for the study was obtained from the institutional review board of the University of Texas at Austin (STUDY00001047) and the Federal Capital Territory Health Research Ethics Committee in Abuja (FHREC/2021/01/148/14-12-21). All participants provided informed consent before enrollment into the study, and consent was obtained from the parents or husbands of the teenage mothers in the form of a signature on the informed consent form. Each respondent received an incentive equivalent to US $5 on completing the final questionnaire. All data were deidentified, coded as participant numbers on the data sheets and questionnaires, and stored in a locked cabinet in a locked office for confidentiality purposes.

Results

Sociodemographic Characteristics

Table 1 shows the sociodemographic characteristics of the 90 teenage mothers who participated in the validation study. Figure 2 shows the CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the study. A large proportion (74/90, 82%) had attended secondary school, were married (53/90, 59%), were not working (63/90, 70%), and had no personal income (63/90, 70%). More than half of the children (49/90, 54%) were boys, and a little more than half were aged >6 months (47/90, 52%). First-born children made up the majority (74/90, 82%) of the participants.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of the teenage mother-child dyad from rural areas of northern Nigeria who participated in the validation study (n=90).

Characteristics	Participants, n (%)
Mother	
Age (y)	
15-17	6 (7)
18-19	84 (93)
Education	
None	2 (2)
Primary	11 (12)
Junior secondary	30 (33)
Senior secondary	44 (49)
Tertiary	3 (3)
Marital status	
Single, separated, or cohabiting	37 (41)
Married	53 (59)
Living with	
Parents, relatives, or friends	54 (60)
Husband	36 (40)
Occupation	
None	63 (70)
Services (eg, hairdressing, tailoring)	18 (20)
Other (eg, petty trading, farming, teaching)	9 (10)
Personal monthly income, naira (US $)	
None	63 (70)
≤20,000 (≤44)	25 (28)
>20,000 (>44)	2 (2)
Household monthly income, naira (US $)	
Not known	34 (38)
≤20,000 (≤44)	11 (12)
>20,000 (>44)	45 (50)
Child	
Sex	
Male	49 (54)
Female	41 (46)
Age (mo)	
0-5	43 (48)
6-11	17 (19)
12-17	15 (17)
18-23	15 (17)
Birth order	
First	74 (82)
Second	12 (13)
Third	4 (4)

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the development and validation of the BabyThrive app.

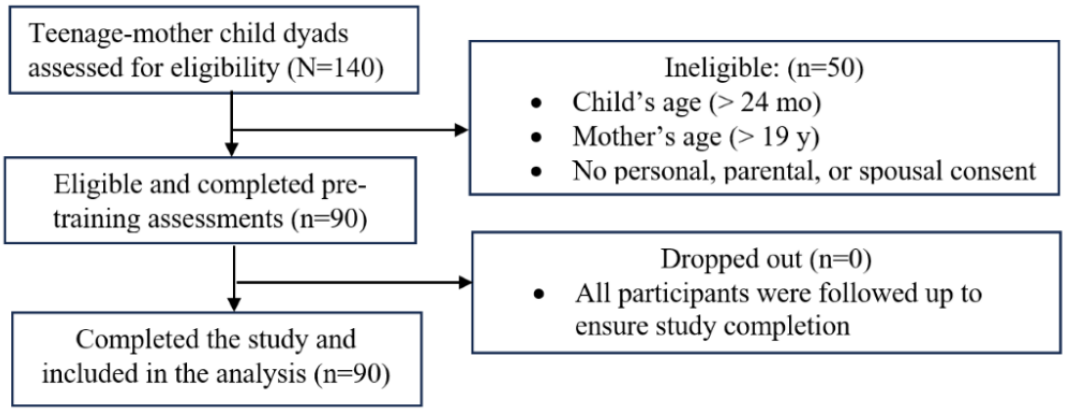

Figure 3 depicts the quality ratings of the BabyThrive app given by experts using the MARS. The highest mean score of 4.6 (SD 0.14) was for perceived app impact, which encompasses the ability of the app to create awareness, disseminate knowledge, influence attitudes and intention, and promote help seeking and behavior change. The mean rating for quality of app information (accuracy of description of the app and its goals as well as quality and quantity of visual and other information) was 4.4 (SD 0.29). User engagement ratings (entertainment value, ability to sustain interest in the target group, ease of customization, and interactivity) ranged widely (from 3.1 to 4.6), with a mean score of 4.0 (SD 0.58). Similar mean scores of 4.1 (SD 0.18) and 4.1 (SD 0.19) were reported for functionality ratings (gestural design, ease of use, navigation, and app performance) and aesthetic value (visual appeal, app attractiveness, and graphical quality). Overall, the mean quality rating for the BabyThrive app was 4.3 (SD 0.39).

Figure 4 displays the usability ratings of the BabyThrive app given by teenage mothers using the mHealth App Usability Questionnaire. Very few teenage mothers strongly disagreed (1/90, 1%-3/90, 3%), disagreed (1/90, 1%-5/90, 6%), or were indifferent (1/90, 1%-10/90, 11%) to the usability statements. Rather, the majority of the participants (77/90, 86%-90/90, 100%) expressed some measure of agreement. A large proportion of participants very strongly agreed that they could use the BabyThrive app without an internet connection (72/90, 80%), would use the app again (70/90, 78%), and were satisfied with the app (61/90, 68%). Fewer participants very strongly agreed that they could easily and quickly recover from mistakes (30/90, 33%), the app informed them of their progress (32/90, 36%), and the interface allowed them to use all app functions (34/90, 38%).

Figure 3. Quality ratings of the BabyThrive app given by experts using the Mobile App Rating Scale.

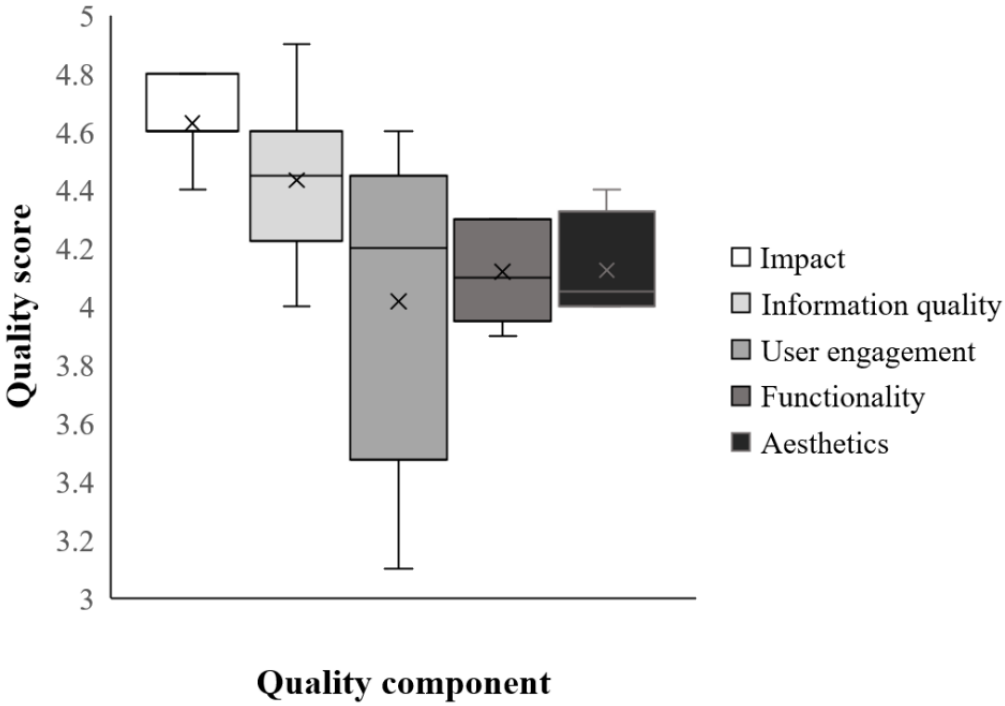

XSL•FO
RenderX

Figure 4. Usability ratings of the BabyThrive app given by teenage mothers using the mHealth App Usability Questionnaire.

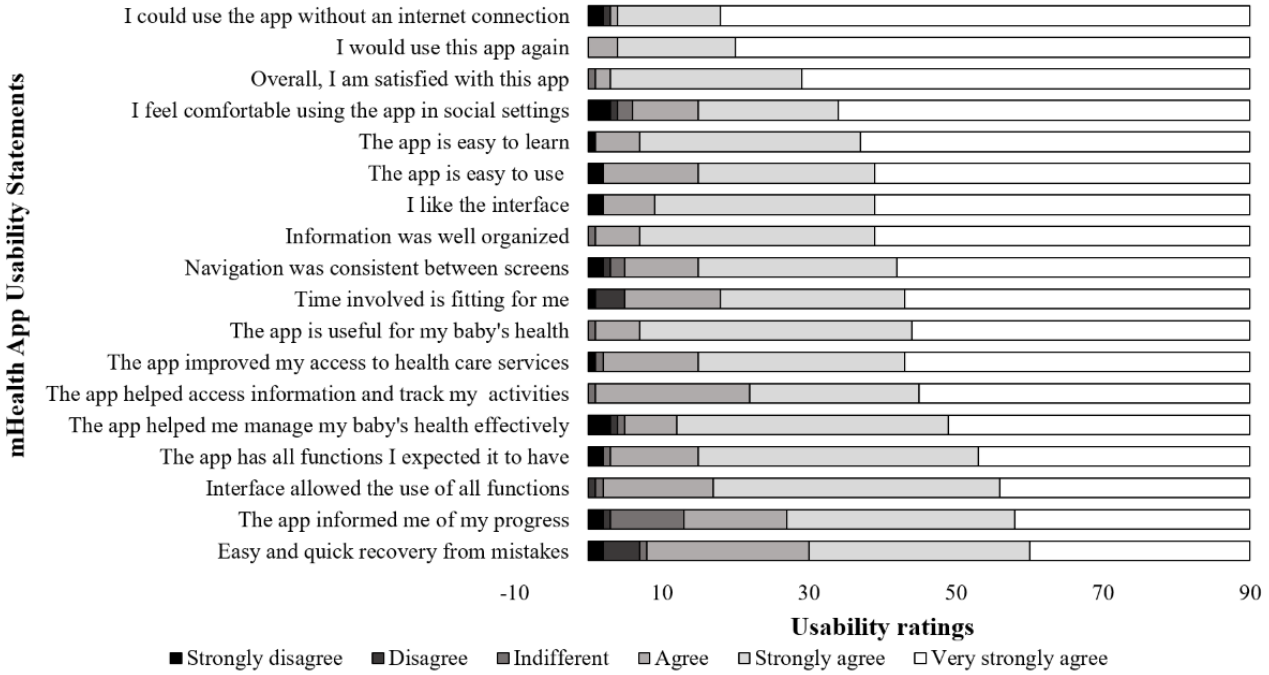

Child Feeding Knowledge

Tables 2 and 3 summarize the child feeding knowledge of the teenage mothers by crossover design before the training and after verbal training and BabyThrive app use. Chi-square tests showed that IYCF knowledge of the teenage mothers (characterized by correct responses to the knowledge questions) was significantly higher in those who received either the verbal training or the BabyThrive app first compared with those before receiving any training. Before the training, knowledge concerning how a mother can keep up her breast milk supply (8/90, 9%), how long breast milk can keep without refrigeration (14/90, 16%), and reasons for allowing the baby to suckle on 1 breast (before changing to the other) for a specific length of time (23/90, 26%) were the lowest. The highest proportions of correct responses before the training were reported for knowledge concerning the first food appropriate for a newborn (74/90, 82%); how often an infant aged <6 months should be breastfed daily (67/90, 74%); and actions to take when an infant is vomiting, stooling, or convulsing (66/90, 73%).

Furthermore, there were no significant differences in knowledge between those who started with either the verbal training or the BabyThrive app, except in the case of 6 questions. Those who started with the app had significantly higher proportions of correct responses. The 6 questions included knowledge concerning reasons why infants should receive either thick or watery porridges (39/45, 98% vs 44/45, 87%; P=.04), how a mother can maintain her breast milk supply (25/45, 78% vs 35/45, 56%; P<.001), the benefits of exclusive breastfeeding for mothers (34/45, 100% vs 45/45, 76%; P=.001), actions to overcome difficulties with breastfeeding (41/45, 100% vs 45/45, 91%; P=.04), how long and why an infant should be allowed to suckle on 1 breast before changing to the other (45/45, 100% vs 36/45, 80%; P=.002; and 45/45, 100% vs 28/45, 62%; P=.02, respectively), and signs that an infant is getting enough milk (43/45, 96% vs 31/45, 69%; P<.001). Figure 5 presents mean differences in IYCF knowledge scores across the training groups. All pretraining mean knowledge scores were significantly lower than the scores in both the verbal training and BabyThrive groups as well as after verbal training and BabyThrive app use. Knowledge of food type and complementary feeding did not differ significantly. Knowledge of exclusive breastfeeding (P<.001) and total knowledge scores (P=.002) were significantly higher in the BabyThrive group than in the verbal training group.

XSL•FO
RenderX

Table 2. Correct responses of child feeding knowledge by crossover design before the training and after verbal training and BabyThrive mobile app use.

	Correct responses	
	Before the training[a] (n=90), n (%)	After verbal training and BabyThrive mobile app use[b] (n=90), n (%)
Food type		
What is the first food for a newborn?	74 (82)	90 (100)
What should infants <6 months be fed?	44 (49)	90 (100)
When a mother resumes work, what should her infant be fed?	34 (38)	90 (100)
Complementary feeding		
At what age should complementary foods be introduced?	62 (69)	85 (94)
Why should complementary foods be introduced at the stated age?	61 (68)	90 (100)
How many times daily should a 6-month-old receive complementary foods?	43 (48)	90 (100)
What quantity of food should an 8-month-old receive per meal?	42 (47)	88 (98)
Should an infant receive thick or watery pap or porridges? Give reasons.	50 (56)	85 (94)
What should a mother do when her infant is vomiting, stooling, and convulsing?	66 (73)	90 (100)
How should you feed an infant who is refusing to eat complementary food?	29 (32)	90 (100)
What kind of diet provides the greatest amount of nutrients?	29 (32)	89 (99)
Exclusive breastfeeding		
How long should infants be fed with only breast milk after birth?	58 (64)	90 (100)
Why should infants be fed with breastmilk alone for some time after birth?	54 (60)	90 (100)
How often should an infant <6 months be breastfed daily?	67 (74)	90 (100)
How can a mother keep up her breast milk supply?	8 (9)	78 (87)
What are the benefits of feeding an infant with breast milk only for mothers?	18 (20)	84 (93)
How long can expressed breast milk keep without refrigeration?	14 (16)	89 (99)
What should a mother do to overcome difficulties with breastfeeding?	27 (30.0)	90 (100)
What are the effects of commercial or mixed feeding in infants <6 months?	57 (63)	90 (100)
What are the signs that an infant needs to be breastfed?	59 (66)	88 (98)
How long should an infant be suckled on 1 breast?	24 (27)	89 (99)
Why should an infant suckle on 1 breast for the stated length of time?	23 (26)	83 (92)
What are the signs that an infant is getting enough breast milk?	25 (28)	88 (98)

[a]All pretraining proportions of correct responses were significantly lower than in both the verbal training and BabyThrive groups ($P<.05$).

[b]All postverbal training and post–BabyThrive app use proportions of correct responses were significantly higher than pretraining proportions ($P<.05$).

XSL•FO
RenderX

Table 3. Correct responses of child feeding knowledge by crossover design in participants who received verbal training initially and the BabyThrive mobile app initially (n=90).

	Correct responses	
	After receiving verbal training initially (n=45), n (%)	After receiving the BabyThrive mobile app initially (n=45), n (%)
Food type		
What is the first food for a newborn?	45 (100)	45 (100)
What should infants <6 months be fed?	43 (96)	44 (98)
When a mother resumes work, what should her infant be fed?	45 (100)	45 (100)
Complementary feeding		
At what age should complementary foods be introduced?	41 (91)	41 (91)
Why should complementary foods be introduced at the stated age?	44 (98)	45 (100)
How many times daily should a 6-month-old receive complementary foods?	45 (100)	45 (100)
What quantity of food should an 8-month-old receive per meal?	43 (96)	43 (96)
Should an infant receive thick or watery pap or porridges? Give reasons.	39 (87)	44 (98)[a]
What should a mother do when her infant is vomiting, stooling, and convulsing?	45 (100)	45 (100)
How should you feed an infant who is refusing to eat complementary food?	44 (98)	45 (100)
What kind of diet provides the greatest amount of nutrients?	40 (89)	43 (96)
Exclusive breastfeeding		
How long should infants be fed with only breast milk after birth?	44 (98)	45 (100)
Why should infants be fed with breastmilk alone for some time after birth?	43 (96)	45 (100)
How often should an infant <6 months be breastfed daily?	45 (98)	45 (100)
How can a mother keep up her breast milk supply?	25 (56)	35 (78)[a]
What are the benefits of feeding an infant with breast milk only for mothers?	34 (76)	45 (100)[a]
How long can expressed breast milk keep without refrigeration?	43 (96)	45 (100)
What should a mother do to overcome difficulties with breastfeeding?	41 (91)	45 (100)[a]
What are the effects of commercial or mixed feeding in infants <6 months?	44 (98)	45 (100)
What are the signs that an infant needs to be breastfed?	43 (96)	45 (100)
How long should an infant be suckled on 1 breast?	36 (80)	45 (100)[a]
Why should an infant suckle on 1 breast for the stated length of time?	28 (62)	45 (100)[a]
What are the signs that an infant is getting enough breast milk?	31 (69)	43 (96)[a]

[a]Proportions of correct responses were significantly higher in the BabyThrive group than in the verbal training group ($P<.05$).

XSL•FO
RenderX

Figure 5. Mean (SD) scores for child feeding knowledge before the training and after verbal training and BabyThrive app use.

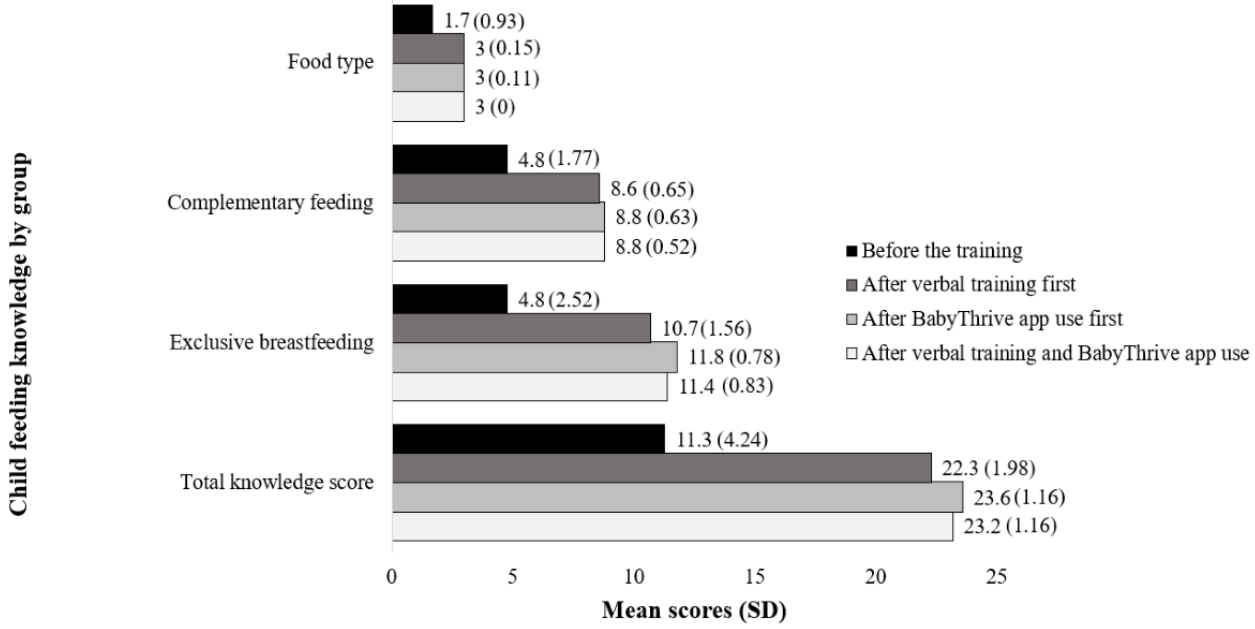

A Comparison of Child Feeding Knowledge

Table 4 presents a comparison of child feeding knowledge using generalized estimating equations. Knowledge of food type was approximately 1.3 times higher after verbal training and BabyThrive app use than before the training (*P*<.001). Similarly, both the verbal training and BabyThrive groups had significantly increased (>3-fold) knowledge concerning complementary feeding in comparison with before the training (*P*<.001). Knowledge of exclusive breastfeeding was approximately 5.91 times greater after verbal training and 6.99 times higher after BabyThrive app use than before the training (*P*<.001). Total knowledge scores after verbal training and BabyThrive app use exceeded pretraining scores by approximately 11-fold and 12-fold, respectively (*P*<.001). In the BabyThrive group, exclusive breastfeeding knowledge (odds ratio [OR] 1.08, 95% CI 0.65 to 1.50) and total knowledge scores (OR 1.29, 95% CI 0.72 to 1.86) were significantly greater than in the verbal training group (*P*<.001). No differences were observed between the verbal training and BabyThrive groups with respect to knowledge of food type (*P*=.83) and complementary feeding (*P*=.63). No carryover effects were recorded because the order of group participation had no significant effects on the outcomes of the training. Knowledge of food type did not differ in participants who received the verbal training first compared with those who received the BabyThrive app first and vice versa (OR 0.033, 95% CI –0.056 to 0.034; *P*=.18). Similarly, knowledge of complementary feeding (OR 0.03, 95% CI –0.29 to 0.22; *P*=.95), exclusive breastfeeding (OR 0.31 95% CI –0.02 to 0.94; *P*=.24), and total knowledge scores (OR 0.44; 95% CI –0.17 to 1.06, *P*=.21), did not differ significantly by order of treatment.

Figure 6 presents a Bland-Altman plot depicting agreement in IYCF knowledge scores between the verbal training and BabyThrive groups. The vertical axis represents the mean differences in IYCF knowledge scores, while the horizontal axis represents the mean IYCF knowledge scores of the verbal training and the BabyThrive groups. The dotted line at the 0 point on the vertical axis indicates an ideal situation where there is zero difference in IYCF knowledge scores from the verbal training and the BabyThrive app. The solid line represents the mean absolute difference in child feeding knowledge between the verbal training and BabyThrive groups. The dashed line above the solid line denotes the upper limit of agreement between both methods, while the lower dashed line shows the lower limit of agreement. Both methods can be said to be in agreement because 80% of the plot points lie within the limits of agreement. Thus, the BabyThrive app is a more effective tool to train teenage mothers on child feeding in place of the verbal IYCF training. The mean absolute difference in knowledge scores between the verbal training and BabyThrive groups was –1.14, with upper and lower limits of agreement of 1.93 and –4.21, respectively. The IYCF knowledge scores obtained from both groups showed coherence, with a statistically significant Spearman correlation coefficient of 0.50 (*P*<.001).

XSL•FO
RenderX

Table 4. Comparison of child feeding knowledge using generalized estimating equations.

Child feeding knowledge across groups	β estimate (regression coefficient, SE; 95% CI)	Z statistic	*P* value
Food type			
Before the training	Reference	Reference	Reference
Verbal training group	1.29 (0.01; 1.06 to 1.52)	12.95	<.001[a]
BabyThrive group	1.30 (0.01; 1.07 to 1.53)	13.14	<.001[a]
BabyThrive group vs verbal training group	0.01 (0.02; –0.06 to 0.03)	0.58	.83
Complementary feeding			
Before the training	Reference	Reference	Reference
Verbal training group	3.43 (0.23; 2.85 to 3.95)	14.56	<.001[a]
BabyThrive group	3.47 (0.23; 2.95 to 4.05)	14.93	<.001[a]
BabyThrive group vs verbal training group	0.10 (0.11; –0.15 to 0.35)	0.92	.63
Exclusive breastfeeding			
Before the training	Reference	Reference	Reference
Verbal training group	5.91 (0.32; 5.17 to 6.65)	18.68	<.001[a]
BabyThrive group	6.99 (0.27; 6.36 to 7.62)	26.00	<.001[a]
BabyThrive group vs verbal training group	1.08 (0.18; 0.65 to 1.50)	5.97	<.001[a]
Total knowledge score			
Before the training	Reference	Reference	Reference
Verbal training group	11.03 (0.50; 9.86 to 12.21)	21.95	<.001[a]
BabyThrive group	12.32 (0.46; 11.24 to 13.40)	26.81	<.001[a]
BabyThrive group vs verbal training group	1.29 (0.24; 0.72 to 1.86)	5.29	<.001[a]

[a]Significant *P* value.

XSL•FO
RenderX

Figure 6. Bland-Altman plot depicting agreement in child feeding knowledge scores after verbal training and BabyThrive app use.

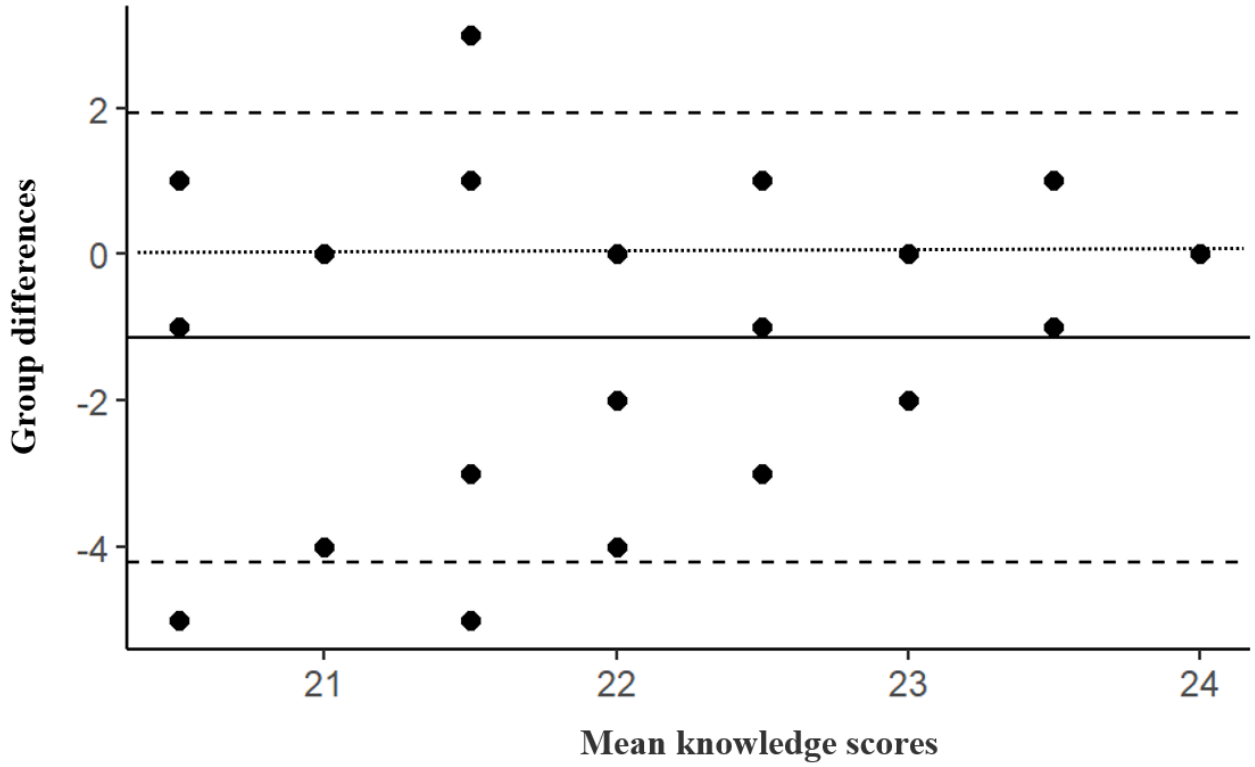

Discussion

Principal Findings

This study developed and validated the BabyThrive app, a mobile gaming app aimed at improving the knowledge of IYCF practices among teenage mothers in Nigeria. The app is an easy, low-cost method to distribute educational content about child nutrition to mothers in rural communities in Nigeria. This interactive mobile app combines active engagement with learning critical nutritional information. Electronic media–based interventions, such as mobile apps and video games, offer distinct advantages over conventional methods of health education because they may be less expensive and more effective in producing behavior change [27,28]. The BabyThrive app distinguishes itself from other apps by focusing on IYCF practices to diminish malnutrition in very young children. It uses gaming technology, interactive videos, and audio techniques adapted to a local Nigerian language as a means of information delivery.

BabyThrive was created using a stepwise process based on existing evidence and resources, as well as qualitative and quantitative inputs and assessments by target users and experts. Quality ratings given by experts (≥4.0 out of 5) and usability ratings given by teenage mothers (>80% in agreement with usability statements) were high. The BabyThrive app improved maternal IYCF knowledge 12-fold in comparison with baseline status. It performed 1.29 times better than the verbal training in transmitting IYCF knowledge to teenage mothers in Nigeria. The IYCF scores from the verbal training were slightly lower but in agreement with scores from the app, as reflected in a small mean absolute difference of 1.14 (which is negligible on the 23-point knowledge scale). Similarly, IYCF knowledge scores obtained from both training methods showed coherence, with a statistically significant, moderate correlation of 0.50. These data suggest that the BabyThrive app is equivalent to, and even surpasses, the verbal training in disseminating IYCF knowledge.

The current research found high levels of user acceptability, but in an evaluation of the use of a breastfeeding tracking app in the United States, user experiences were mixed (helpful, excessive time required, and anxiety inducing) [60]. The high levels (>80%) of usability and user satisfaction in our study are consistent with the findings of >80% satisfaction with the Growing Healthy mHealth app in Australia [61]. In Australian mothers, similarly high levels (>80%) of user agreement with usability statements were reported concerning the Milky Way app for breastfeeding support [62]. In rural India, the breastfeeding support app, Best4Baby, reported high usability values (>95th percentile) [63]. On the basis of assessments conducted using the mHealth App Usability Questionnaire (which was also used in our study), high usability and user satisfaction values were reported for the BabyByte app, which was designed to promote responsive feeding practices in parents of infants and toddlers in the United States [64]. In Malaysia, a wide range of positive responses to usability statements (50%-100%) was reported for the Gigiku Sihat app for caregiver training on child diet and oral health. These results differ from those of our study, which had a narrower range (85%-100%) of agreement with usability statements [65]. In comparison with the values (100% for both ease of use and ease of learning the app) reported for the Gigiku Sihat app [65], similarly high

XSL•FO
RenderX

proportions of participants (97.8% and 98.8%, respectively) in our study had positive responses to statements concerning ease of use and ease of learning the app. In Egypt, an overall mean app quality rating of 3.7 was reported for the Sehhat Tefly app, which was aimed at improving the capacity of caregivers to track the health and well-being of children [66]. By contrast, using the same quality assessment scale as the Sehhat Tefly app [66], our study had a greater overall mean app quality rating of 4.3 (SD 0.39).

Among mothers living in rural Nepal, the serious game MANTRA, which used simple game mechanics such as "drag and drop," recorded an average gain of 6.0 points (on a 100-point scale) in knowledge concerning neonatal health [30]. This mechanism was also used in our BabyThrive app. In comparison with mHealth studies in Pakistan (14.8% increase) and Australia (6.9% increase), our BabyThrive app successfully increased IYCF knowledge of mothers to a much higher degree (12-fold over pretraining levels) [67,68]. In Ethiopia, an mHealth study to promote exclusive breastfeeding knowledge in 121 parent pairs recorded a lower mean difference in IYCF knowledge (>7.0) than that in our investigation (12.3) [69]. Similarly, the knowledge of 720 Sri Lankan mothers concerning child feeding also increased significantly in an mHealth intervention [70].

Numerous studies on mHealth apps aimed at improving maternal knowledge of IYCF practices have compared app efficacy to either baseline or controls with no treatment [30,67-69]. Only a few mHealth studies have compared app efficacy to other methods for achieving the same outcomes. Our study compared app efficacy to baseline (before the training), as well as to outcomes of verbal training (the usual method of IYCF knowledge dissemination). Despite the mixed results on user experience shown on the breastfeeding tracking app in the United States [60], the agreement (intraclass correlation coefficient of 0.97) with the interview method was notably higher than the moderate Spearman correlation coefficient of 0.50 documented in our study. Our lower correlation may be because the app was 1.29-fold more effective than the verbal training in transmitting IYCF knowledge. Similar to our findings, the use of the Khunlook mobile app given to parents in Thailand greatly increased health literacy in comparison to an intervention delivered via a child health handbook [71], and consistent with our findings, a complementary feeding education trial in Iran reported an 18-point difference in maternal nutrition literacy between mHealth app users and controls who received regular well-child services [72].

We used a mobile game as the vehicle for change because gamification has been shown to have great potential for improving knowledge and changing behavior through reinforcement and personalized or tailored approaches [73-77]. Games appeal to broad audiences, can trigger motivation to initiate and sustain desired behaviors, and promote well-being by inducing positive emotions and experiences [74-79]. In particular, mobile games are suitable for health interventions in adolescents, given their high proclivity for playing games and using technology [76,79].

The BabyThrive app was created with the purpose of training teenage mothers considered to be at risk with malnourished infants to gain knowledge and learn practices to improve their child's nutritional health. We believe that this improvement has the potential to reduce child morbidity and mortality, lower health care costs for families, and improve intergenerational family income and overall prosperity of the community. This app intervention will be sustainable because the app can be widely disseminated through web-based app stores, development partners, health facilities, and community volunteers; in addition, app maintenance (periodic updates and app store fees) involves minimal costs. The UNICEF IYCF training package (on which the BabyThrive app is based) is widely used in sub-Saharan Africa and in >90 countries globally [80]. The results of this study can be generalized to adolescent mothers in rural areas of Nigeria and in countries where this training package is used; however, the specific findings may vary across different settings. Furthermore, as the BabyThive app is available in 2 languages, templates exist for the translation of the app into other languages. Thus, with cultural adaptation of the recipes contained in the app to fit local contexts, the BabyThrive app can be easily adapted for use in other countries in sub-Saharan Africa as well as low- and middle-income nations.

Strengths and Limitations

The strengths of this validated app include the use of gaming technologies that promote user engagement, ease of use despite unreliable internet connectivity due to offline functionality, availability in >1 language, and ease and acceptability of use among teenage mothers with low-income status in rural Nigeria. A major strength of this research is the inclusion of populations of teenage mothers who are commonly excluded from technology-driven solutions due to low household income, poor literacy levels, and language barriers. One limitation of this study is the absence of social peer-to-peer connection due to the offline nature of the app. Another limitation is that a large proportion of the participants (84/90, 93%) who participated in the validation study were aged >17 years because very few younger mothers (6/90, 7%) met the eligibility criteria.

Conclusions

This research developed a novel, offline mobile gaming app. This app has been documented to be a valid, easy, effective, and acceptable method to disseminate critical knowledge on IYCF practices to teenage mothers in rural Nigeria. This app will be a useful instrument for professionals and personnel in the nutrition, community health, governmental, nonprofit, and child health fields, as well as other stakeholders, to improve maternal awareness of IYCF practices. Randomized controlled trials should be conducted to test the effectiveness of the BabyThrive app in improving maternal IYCF knowledge, attitudes, and practices. In addition, the ability of the app to ameliorate macro- and micronutrient undernutrition in childhood, particularly among children born to teenage mothers, should be evaluated.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Schlumberger Foundation Faculty for the Future Program. Publication costs were jointly covered by the Department of Nutritional Sciences at the University of Texas at Austin and the Bess Heflin Centennial Professorship at the University of Texas at Austin. The authors would like to thank PixelArt Games Academy in Islamabad, Pakistan for developing the BabyThrive app and the nonprofit Centre for Family Health Initiative in Abuja, Nigeria, for assistance with data collection. In addition, the authors are grateful to the panel of nutrition and gaming experts who assisted with establishing app content validity from the Department of Nutritional Sciences at the University of Texas at Austin, Texas; Department of Human Nutrition and Dietetics at the University of Ibadan, Ibadan, Nigeria; the Bauchi State Agricultural Development Program, Nigeria; Nigeria's Federal Ministry of Health and Social Welfare in Abuja; the Department of Nutritional Sciences at the University of Illinois at Urbana-Champaign, Illinois; the Department of Nutrition and Dietetics at the Federal Polytechnic Bauchi, Bauchi, Nigeria; Kano State College of Health Sciences and Technology, Nigeria; Endoclan Applications (a game development firm), Port Harcourt, Nigeria; the Information and Communication and Technology Centre of the Abubakar Tafawa Balewa University, Bauchi, Nigeria; and GitLab Inc.

Authors' Contributions

All authors made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; drafted the article or revised it critically for important intellectual content; and approved the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the BabyThrive app.
[PDF File (Adobe PDF File), 1154 KB - jmir_v26i1e53560_app1.pdf]

References

1. Sully EA, Biddlecom A, Darroch JE, Riley T, Ashford L, Lince-Deroche N, et al. Adding it up: investing in sexual and reproductive health 2019. Guttmacher Institute. 2019. URL: https://www.guttmacher.org/report/adding-it-up-investing-in-sexual-reproductive-health-2019 [accessed 2023-06-30]
2. Darroch J, Woog V, Bankole A, Ashford LS. Adding it up: costs and benefits of meeting the contraceptive needs of adolescents. Guttmacher Institute. 2016. URL: https://www.guttmacher.org/report/adding-it-meeting-contraceptive-needs-of-adolescents [accessed 2023-06-30]
3. Abortion facts and figures 2021. Population Reference Bureau. URL: https://www.prb.org/wp-content/uploads/2021/03/2021-safe-engage-abortion-facts-and-figures-media-guide.pdf [accessed 2023-06-30]
4. Motherhood in childhood: the untold story. United Nations Fund for Population. URL: https://www.unfpa.org/sites/default/files/pub-pdf/MotherhoodInChildhood_report.pdf [accessed 2023-06-30]
5. Neal S, Channon AA, Chandra-Mouli V, Madise N. Trends in adolescent first births in sub-Saharan Africa: a tale of increasing inequity? Int J Equity Health 2020 Sep 04;19(1):151 [FREE Full text] [doi: 10.1186/s12939-020-01251-y] [Medline: 32887618]
6. Chakole S, Akre S, Sharma K, Wasnik P, Wanjari MB. Unwanted teenage pregnancy and its complications: a narrative review. Cureus 2022 Dec;14(12):e32662 [FREE Full text] [doi: 10.7759/cureus.32662] [Medline: 36686124]
7. Nguyen P, Scott S, Khuong L, Pramanik P, Ahmed A, Afsana K, et al. Why are adolescent mothers more likely to have stunted and underweight children than adult mothers? A path analysis Using Data from 30,000 Bangladeshi mothers, 1996–2014. Curr Dev Nutr 2020 Jun;4:nzaa061_091. [doi: 10.1093/cdn/nzaa061_091]
8. Nigeria demographic and health survey 2018. National Population Commission (NPC), Nigeria. 2019. URL: https://www.dhsprogram.com/pubs/pdf/FR359/FR359.pdf [accessed 2024-04-29]
9. Child mortality, stillbirth, and causes of death estimates. UN Inter-agency Group for Child Mortality Estimation. URL: https://childmortality.org/data/Nigeria [accessed 2023-07-01]
10. Report on the nutrition and health situation of Nigeria: national nutrition and health survey. National Bureau of Statistics, Nigeria. 2018. URL: https://www.unicef.org/nigeria/media/2181/file/Nigeria-NNHS-2018.pdf [accessed 2023-07-01]
11. Olodu MD, Adeyemi AG, Olowookere SA, Esimai OA. Nutritional status of under-five children born to teenage mothers in an urban setting, south-western Nigeria. BMC Res Notes 2019 Mar 04;12(1):116 [FREE Full text] [doi: 10.1186/s13104-019-4147-x] [Medline: 30832719]
12. Oyeyemi AL, Aliyu SU, Sa'ad F, Rufa'i AA, Jajere AR, Oyeyemi AY. Association between adolescent motherhood and maternal and child health indices in Maiduguri, Nigeria: a community-based cross-sectional study. BMJ Open 2019 Mar 08;9(3):e024017 [FREE Full text] [doi: 10.1136/bmjopen-2018-024017] [Medline: 30852533]

XSL•FO
RenderX

13. Muhimbula H, Kinabo J, O'Sullivan A. Determinants of infant nutrition status in rural farming households before and after harvest. Matern Child Nutr 2019 Jul 29;15(3):e12811 [FREE Full text] [doi: 10.1111/mcn.12811] [Medline: 30897661]
14. Ogbo FA, Page A, Idoko J, Claudio F, Agho KE. Have policy responses in Nigeria resulted in improvements in infant and young child feeding practices in Nigeria? Int Breastfeed J 2016 Feb 8;12(1):9 [FREE Full text] [doi: 10.1186/s13006-017-0101-5] [Medline: 28194220]
15. Kullmann MP, Leader D. Knowledge, attitudes, and barriers to breastfeeding in adolescent mothers: a review. Cooper Rowan Med J 2021 Mar 16;2(1):98-115. [doi: 10.31986/issn.2578-3343_vol2iss1.8]
16. Costa KA, Antunes MM, Cabral PC, Silva GA. Feeding style of adolescent mothers and complementary feeding practice of their infants. Rev Nutr 2018 Feb;31(1):49-58. [doi: 10.1590/1678-98652018000100005]
17. Benova L, Siddiqi M, Abejirinde IO, Badejo O. Time trends and determinants of breastfeeding practices among adolescents and young women in Nigeria, 2003-2018. BMJ Glob Health 2020 Aug 06;5(8):e002516 [FREE Full text] [doi: 10.1136/bmjgh-2020-002516] [Medline: 32764127]
18. Adeyemi O, Afolabi WA, Ferguson E, Hoyt J, Lawal H, Okunola RA, et al. How to strengthen the infant and young child feeding (IYCF) programme in Northern Nigeria. Nutrition Research in Northern Nigeria. URL: https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/6906 [accessed 2023-07-01]
19. Kay M, Santos J, Takane M. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://www.afro.who.int/publications/mhealth-new-horizons-health-through-mobile-technologie [accessed 2024-04-29]
20. Kazemi DM, Borsari B, Levine MJ, Lamberson KA, Dooley B. REMIT: development of a mHealth theory-based intervention to decrease heavy episodic drinking among college students. Addict Res Theory 2018 Dec 26;26(5):377-385 [FREE Full text] [doi: 10.1080/16066359.2017.1420783] [Medline: 32694964]
21. Abubakar I, Dalglish SL, Angell B, Sanuade O, Abimbola S, Adamu AL, et al. The Lancet Nigeria commission: investing in health and the future of the nation. Lancet 2022 Mar;399(10330):1155-1200. [doi: 10.1016/s0140-6736(21)02488-0]
22. Statistics and reports. Nigerian Communications Commission. 2020. URL: https://www.ncc.gov.ng/contact-ncc/13-statistics-report [accessed 2023-07-01]
23. Total population of Nigeria. The World Bank. URL: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=NG [accessed 2023-07-08]
24. Sellen D, Mbugua S, Webb‐Girard A, Lou W, Duan W, Kamau‐Mbuthia E. Cell phone based peer counselling can support exclusive breastfeeding: a randomized controlled trial in Kenya (119.5). FASEB J 2014 Apr;28(S1):8. [doi: 10.1096/fasebj.28.1_supplement.119.5]
25. Flax VL, Negerie M, Ibrahim AU, Leatherman S, Daza EJ, Bentley ME. Integrating group counseling, cell phone messaging, and participant-generated songs and dramas into a microcredit program increases Nigerian women's adherence to international breastfeeding recommendations. J Nutr 2014 Jul;144(7):1120-1124 [FREE Full text] [doi: 10.3945/jn.113.190124] [Medline: 24812071]
26. Unger JA, Ronen K, Perrier T, DeRenzi B, Slyker J, Drake A, et al. Short message service communication improves exclusive breastfeeding and early postpartum contraception in a low- to middle-income country setting: a randomised trial. BJOG 2018 Nov 28;125(12):1620-1629 [FREE Full text] [doi: 10.1111/1471-0528.15337] [Medline: 29924912]
27. Pereira CV, Figueiredo G, Esteves MG, de Souza JM. We4Fit: a game with a purpose for behavior change. In: Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design. 2014 Presented at: CSCWD '14; May 21-23, 2014; Hsinchu, Taiwan p. 83-88 URL: https://ieeexplore.ieee.org/document/6846821 [doi: 10.1109/cscwd.2014.6846821]
28. Thompson D. Designing serious video games for health behavior change: current status and future directions. J Diabetes Sci Technol 2012 Jul 01;6(4):807-811 [FREE Full text] [doi: 10.1177/193229681200600411] [Medline: 22920806]
29. Lamas S, Rebelo S, da Costa S, Sousa H, Zagalo N, Pinto E. The influence of serious games in the promotion of healthy diet and physical activity health: a systematic review. Nutrients 2023 Mar 14;15(6):1399 [FREE Full text] [doi: 10.3390/nu15061399] [Medline: 36986129]
30. Mueller S, Soriano D, Boscor A, Saville NM, Arjyal A, Baral S, et al. MANTRA: improving knowledge of maternal health, neonatal health, and geohazards in women in rural Nepal using a mobile serious game. Front Public Health 2020 Dec 11;8:584375 [FREE Full text] [doi: 10.3389/fpubh.2020.584375] [Medline: 33363082]
31. Brand L, Beltran A, Buday R, Hughes S, O'Connor T, Baranowski J, et al. Training vegetable parenting practices through a mobile game: iterative qualitative alpha test. JMIR Serious Games 2015 Jul 24;3(2):e6 [FREE Full text] [doi: 10.2196/games.4081] [Medline: 26208899]
32. Espinosa-Curiel IE, Pozas-Bogarin EE, Hernández-Arvizu M, Navarro-Jiménez ME, Delgado-Pérez EE, Martínez-Miranda J, et al. HelperFriend, a serious game for promoting healthy lifestyle behaviors in children: design and pilot study. JMIR Serious Games 2022 May 06;10(2):e33412 [FREE Full text] [doi: 10.2196/33412] [Medline: 35522474]
33. Edwards EA, Lumsden J, Rivas C, Steed L, Edwards LA, Thiyagarajan A, et al. Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps. BMJ Open 2016 Oct 04;6(10):e012447 [FREE Full text] [doi: 10.1136/bmjopen-2016-012447] [Medline: 27707829]

XSL•FO
RenderX

34. The Nigeria community and facility infant and young child feeding package. Federal Ministry of Health Nigeria. 2012. URL: https://www.spring-nutrition.org/sites/default/files/training_materials/files/c-iycf_participant_materials.pdf [accessed 2023-07-30]
35. Cruz A, Alves AC. Traditional, agile and lean project management - a systematic literature review. J Mod Proj Manag 2022(8):2 [FREE Full text]
36. Market share of mobile operating systems in Nigeria from January 2018 to April 2023. Statista. URL: https://www.statista.com/statistics/1063846/market-share-held-by-mobile-operating-systems-in-nigeria/ [accessed 2024-04-29]
37. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
38. Tang K, Gerling K, Chen W, Geurts L. Information and communication systems to tackle barriers to breastfeeding: systematic search and review. J Med Internet Res 2019 Sep 27;21(9):e13947 [FREE Full text] [doi: 10.2196/13947] [Medline: 31573903]
39. Bandura A. Social Foundations of Thought and Action. Englewood Cliffs, NJ: Prentice-Hall; 1986.
40. Sosanya ME, Freeland-Graves JH, Gbemileke AO, Adeosun FF, Samuel FO, Shokunbi OS. Characterization of traditional foods and diets in rural areas of Bauchi state, Nigeria: analysis of nutrient components. European J Nutr Food Saf 2021 Oct 15;13(7):1-16. [doi: 10.9734/ejnfs/2021/v13i730432]
41. Software for assessing growth and development of the world's children. World Health Organization. 2011. URL: https://cdn.who.int/media/docs/default-source/child-growth/child-growth-standards/software/anthro-pc-manual-v322.pdf?sfvrsn=c4e76522_2 [accessed 2024-04-29]
42. Agbadjé TT, Elidor H, Perin MS, Adekpedjou R, Légaré F. Towards a taxonomy of behavior change techniques for promoting shared decision making. Implement Sci 2020 Aug 20;15(1):67 [FREE Full text] [doi: 10.1186/s13012-020-01015-w] [Medline: 32819410]
43. Jake-Schoffman DE, Silfee VJ, Waring ME, Boudreaux ED, Sadasivam RS, Mullen SP, et al. Methods for evaluating the content, usability, and efficacy of commercial mobile health apps. JMIR Mhealth Uhealth 2017 Dec 18;5(12):e190 [FREE Full text] [doi: 10.2196/mhealth.8758] [Medline: 29254914]
44. Muro-Culebras A, Escriche-Escuder A, Martin-Martin J, Roldán-Jiménez C, De-Torres I, Ruiz-Muñoz M, et al. Tools for evaluating the content, efficacy, and usability of mobile health apps according to the consensus-based standards for the selection of health measurement instruments: systematic review. JMIR Mhealth Uhealth 2021 Dec 01;9(12):e15433 [FREE Full text] [doi: 10.2196/15433] [Medline: 34855618]
45. Questions and answers: qualification of digital technology-based methodologies to support approval of medicinal products. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/other/questions-answers-qualification-digital-technology-based-methodologies-support-approval-medicinal_en.pdf [accessed 2023-07-15]
46. Ahlin R, Sigvardsson I, Skokic V, Landberg R, Steineck G, Hedelin M. development and validation of a mobile phone application developed for measuring dietary fiber intake. Nutrients 2021 Jun 22;13(7):2133 [FREE Full text] [doi: 10.3390/nu13072133] [Medline: 34206639]
47. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): Development and Validation Study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500. [doi: 10.2196/11500] [Medline: 30973342]
48. Lim CY, In J. Considerations for crossover design in clinical study. Korean J Anesthesiol 2021 Aug;74(4):293-299 [FREE Full text] [doi: 10.4097/kja.21165] [Medline: 34344139]
49. Sosanya ME, Beamon I, Muhammad R, Freeland-Graves JH. Development and validation of the Teen Moms Child Feeding Questionnaire for Sub-Saharan Africa. BMC Public Health 2023 Aug 04;23(1):1487 [FREE Full text] [doi: 10.1186/s12889-023-16365-5] [Medline: 37542311]
50. Lorenz HS, Stuebing MD, Nambeye C, Lungu G, Littlefield LM. Substance use treatment using cultural arts and 12 steps: curriculum training and community-led implementation in Zambia. Addict Behav Rep 2022 Jun;15:100424 [FREE Full text] [doi: 10.1016/j.abrep.2022.100424] [Medline: 35541027]
51. Briñol P, Rucker DD, Petty RE. Naïve theories about persuasion: implications for information processing and consumer attitude change. Int J Advert 2015 Jan 19;34(1):85-106. [doi: 10.1080/02650487.2014.997080]
52. Haug S, Paz Castro R, Scholz U, Kowatsch T, Schaub MP, Radtke T. Assessment of the efficacy of a mobile phone-delivered just-in-time planning intervention to reduce alcohol use in adolescents: randomized controlled crossover trial. JMIR Mhealth Uhealth 2020 May 26;8(5):e16937 [FREE Full text] [doi: 10.2196/16937] [Medline: 32452818]
53. Craig Rushing S, Kelley A, Bull S, Stephens D, Wrobel J, Silvasstar J, et al. Efficacy of an mHealth intervention (BRAVE) to promote mental wellness for American Indian and Alaska native teenagers and young adults: randomized controlled trial. JMIR Ment Health 2021 Sep 15;8(9):e26158 [FREE Full text] [doi: 10.2196/26158] [Medline: 34524092]
54. Streiner D, Norman G. Health Measurement Scales: A Practical Guide to Their Development and Us. 4th edition. New York, NY: Oxford University Press; 2008.
55. Alrubaiy L, Hutchings HA, Williams JG. Assessing patient reported outcome measures: a practical guide for gastroenterologists. United European Gastroenterol J 2014 Dec;2(6):463-470 [FREE Full text] [doi: 10.1177/2050640614558345] [Medline: 25452841]

XSL•FO
RenderX

56. RStudio Team. RStudio: integrated development for R. RStudio. URL: http://www.rstudio.com/ [accessed 2023-01-15]
57. Ballinger GA. Using generalized estimating equations for longitudinal data analysis. Organ Res Methods 2016 Jun 29;7(2):127-150. [doi: 10.1177/1094428104263672]
58. Cummins KM, Brumback T, Chung T, Moore RC, Henthorn T, Eberson S, et al. Acceptability, validity, and engagement with a mobile app for frequent, continuous multiyear assessment of youth health behaviors (mNCANDA): mixed methods study. JMIR Mhealth Uhealth 2021 Feb 10;9(2):e24472 [FREE Full text] [doi: 10.2196/24472] [Medline: 33565988]
59. Riffenburgh RH, Gillen DL. 27-techniques to aid analysis. In: Riffenburgh RH, Gillen DL, editors. Statistics in Medicine. 4th edition. New York, NY: Academic Press; 2020:631-649.
60. Demirci JR, Bogen DL. Feasibility and acceptability of a mobile app in an ecological momentary assessment of early breastfeeding. Matern Child Nutr 2017 Jul 14;13(3):e12342 [FREE Full text] [doi: 10.1111/mcn.12342] [Medline: 27411497]
61. Laws RA, Denney-Wilson EA, Taki S, Russell CG, Zheng M, Litterbach E, et al. Key lessons and impact of the growing healthy mHealth program on milk feeding, timing of introduction of solids, and infant growth: quasi-experimental study. JMIR Mhealth Uhealth 2018 Apr 19;6(4):e78 [FREE Full text] [doi: 10.2196/mhealth.9040] [Medline: 29674313]
62. Meedya S, Win K, Yeatman H, Fahy K, Walton K, Burgess L, et al. Developing and testing a mobile application for breastfeeding support: the Milky Way application. Women Birth 2021 Mar;34(2):e196-e203. [doi: 10.1016/j.wombi.2020.02.006] [Medline: 32081557]
63. Ma T, Chang K, Alyusuf A, Bajracharya E, Washio Y, Kelly PJ, et al. Design, development, and testing of BEST4Baby, an mHealth technology to support exclusive breastfeeding in India: pilot study. JMIR Form Res 2022 Sep 08;6(9):e32795 [FREE Full text] [doi: 10.2196/32795] [Medline: 36074546]
64. Mobley A, Jake-Schoffman D, Fedele D, Varela E, Zeldman J. Babybyte: usability testing of a mHealth app prototype to improve responsive feeding practices in parents of infants and toddlers. Curr Dev Nutr 2021 Jun;5:983. [doi: 10.1093/cdn/nzab051_027]
65. Akmal MN, Hasan R, Saddki N, Mohd Arshad MR, Ahmad M. Development and usability testing of mobile application on diet and oral health. PLoS One 2021 Sep 8;16(9):e0257035 [FREE Full text] [doi: 10.1371/journal.pone.0257035] [Medline: 34495979]
66. Ibrahim NM, Ez-Elarab HS, Momen M, Mossad IM, Eletriby SS. A novel wide scale well-baby clinic mobile application: an Egyptian pilot study. BMC Health Serv Res 2023 Jun 24;23(1):687 [FREE Full text] [doi: 10.1186/s12913-023-09720-0] [Medline: 37353827]
67. Akber S, Mahmood H, Fatima R, Wali A, Alam A, Sheraz SY, et al. Effectiveness of a mobile health intervention on infant and young child feeding among children ≤ 24 months of age in rural Islamabad over six months duration. F1000Res 2019 Jul 29;8:551. [doi: 10.12688/f1000research.17037.2]
68. Laws RA, Cheng H, Rossiter C, Kuswara K, Markides BR, Size D, et al. Perinatal support for breastfeeding using mHealth: a mixed methods feasibility study of the My Baby Now app. Matern Child Nutr 2023 Apr;19(2):e13482 [FREE Full text] [doi: 10.1111/mcn.13482] [Medline: 36725007]
69. Gebremariam KT, Mulugeta A, Gallegos D. Theory-based mHealth targeting fathers and mothers to improve exclusive breastfeeding: a quasi-experimental study. Int Breastfeed J 2023 Jan 06;18(1):2 [FREE Full text] [doi: 10.1186/s13006-022-00537-x] [Medline: 36604757]
70. Peiris DR, Wijesinghe MS, Gunawardana BM, Weerasinghe WM, Rajapaksha RM, Rathnayake KM, et al. Mobile phone-based nutrition education targeting pregnant and nursing mothers in Sri Lanka. Int J Environ Res Public Health 2023 Jan 28;20(3):2324 [FREE Full text] [doi: 10.3390/ijerph20032324] [Medline: 36767691]
71. Areemit R, Saengnipanthkul S, Sutra S, Lumbiganon P, Pornprasitsakul P, Paopongsawan P, et al. Effectiveness of a Mobile App (KhunLook) versus the maternal and child health handbook on Thai parents' health literacy, accuracy of health assessments, and convenience of use: randomized controlled trial. J Med Internet Res 2023 May 09;25:e43196 [FREE Full text] [doi: 10.2196/43196] [Medline: 37159258]
72. Seyyedi N, Rahimi B, Eslamlou HR, Afshar HL, Spreco A, Timpka T. Smartphone-based maternal education for the complementary feeding of undernourished children under 3 years of age in food-secure communities: randomised controlled trial in Urmia, Iran. Nutrients 2020 Feb 24;12(2):587 [FREE Full text] [doi: 10.3390/nu12020587] [Medline: 32102310]
73. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
74. Azevedo J, Padrão P, Gregório MJ, Almeida C, Moutinho N, Lien N, et al. A web-based gamification program to improve nutrition literacy in families of 3- to 5-year-old children: the Nutriscience project. J Nutr Educ Behav 2019 Mar;51(3):326-334. [doi: 10.1016/j.jneb.2018.10.008] [Medline: 30579894]
75. Mackert M, Mandell D, Donovan E, Walker L, Henson-García M, Bouchacourt L. Mobile apps as audience-centered health communication platforms. JMIR Mhealth Uhealth 2021 Aug 17;9(8):e25425 [FREE Full text] [doi: 10.2196/25425] [Medline: 34402797]
76. Pernencar C, Sousa P, Frontini R, Martinho R, Runte D, Mendes D, et al. Planning a health promotion program: mobile app gamification as a tool to engage adolescents. Procedia Comput Sci 2018;138:113-118. [doi: 10.1016/j.procs.2018.10.016]

XSL•FO
RenderX

77. Toh W, Kirschner D. Developing social-emotional concepts for learning with video games. Comput Educ 2023 Mar;194:104708. [doi: 10.1016/j.compedu.2022.104708]
78. Hightow-Weidman LB, Muessig KE, Bauermeister J, Zhang C, LeGrand S. Youth, technology, and HIV: recent advances and future directions. Curr HIV/AIDS Rep 2015 Dec 18;12(4):500-515 [FREE Full text] [doi: 10.1007/s11904-015-0280-x] [Medline: 26385582]
79. Pendergrass TM, Hieftje K, Fiellin LE. Improving health outcomes and quality of life for African adolescents: the role of digital and mobile games. In: Eloff I, editor. Handbook of Quality of Life in African SocietiesEloff. Cham, Switzerland: Springer; 2019:149-176.
80. Hromi-Fiedler AJ, Pérez-Escamilla R, Segura-Pérez S, Garg A, Bégin F. Assessing the nurturing care content of UNICEF's community infant and young child feeding counselling package: gaps, best practices, and lessons learned. Curr Dev Nutr 2022 Mar;6(3):nzac018 [FREE Full text] [doi: 10.1093/cdn/nzac018] [Medline: 35368736]

Abbreviations

BCT: behavior change technique
CONSORT: Consolidated Standards of Reporting Trials
IYCF: infant and young child feeding
MARS: Mobile App Rating Scale
mHealth: mobile health
OR: odds ratio
UNICEF: United Nations Children's Fund

Edited by S Ma, T Leung; submitted 10.10.23; peer-reviewed by A Palacios, P Velmovitsky; comments to author 01.04.24; revised version received 20.04.24; accepted 07.06.24; published 26.09.24.

Please cite as:
Sosanya ME, Samuel FO, Bashir S, Omoera VO, Freeland-Graves JH
A Mobile Gaming App to Train Teenage Mothers on Appropriate Child Feeding Practices: Development and Validation Study
J Med Internet Res 2024;26:e53560
URL: https://www.jmir.org/2024/1/e53560
doi:10.2196/53560
PMID:

©Mercy Eloho Sosanya, Folake Olukemi Samuel, Sadia Bashir, Victoria Osariemen Omoera, Jeanne H Freeland-Graves. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile Phones and HIV Testing: Multicountry Instrumental Variable Analysis From Sub-Saharan Africa

Francesco Iacoella[1*], PhD; Nyasha Tirivayi[2*], PhD

[1]UNICEF Evaluation Office, New York, NY, United States

[2]UNICEF Innocenti-Global Office of Research and Foresight, Florence, Italy

[*]all authors contributed equally

Corresponding Author:
Francesco Iacoella, PhD
UNICEF Evaluation Office
3 United Nations Plaza
New York, NY, 10017
United States
Phone: 1 9296459846
Email: iacoella@merit.unu.edu

Abstract

Background: Sub-Saharan Africa has been a technological hothouse when it comes to mobile phone technology adoption. However, evidence on the role played by mobile technology on infectious disease prevention has been mostly limited to experimental studies.

Objective: This observational study investigates the role of mobile phone connectivity on HIV testing in sub-Saharan Africa.

Methods: We make use of the novel and comprehensive OpenCelliD cell tower database and Demographic and Health Survey geocoded information for over 400,000 women in 29 sub-Saharan African countries. We examine, through ordinary least square and instrumental variable regressions, whether women's community distance from the closest cell tower influences knowledge about HIV testing facilities and the likelihood of ever being tested for HIV.

Results: After finding a negative and significant impact of distance to the nearest cell tower on knowledge of HIV testing facility (–0.7 percentage points per unit increase in distance) and HIV testing (–0.5 percentage points per unit increase), we investigate the mechanisms through which such effects might occur. Our analysis shows that distance to a cell tower reduces HIV-related knowledge (–0.4 percentage points per unit increase) as well as reproductive health knowledge (–0.4 percentage points per unit increase). Similar results are observed when the analysis is performed at community level.

Conclusions: Results suggest that the effect of mobile phone connectivity is channeled through increased knowledge of HIV, sexually transmittable infections, and modern contraceptive methods. Further analysis shows that cell phone ownership has an even larger impact on HIV testing and knowledge. This paper adds to the recent literature on the impact of mobile-based HIV prevention schemes by showing through large-scale analysis that better mobile network access is a powerful tool to spread reproductive health knowledge and increase HIV awareness.

(J Med Internet Res 2024;26:e48794) doi:10.2196/48794

KEYWORDS

information and knowledge; communication; health and economic development; public health; technological change; choices and consequences; mobile phone; connectivity; access; HIV; testing; Sub-Saharan Africa; women's health

Introduction

Sub-Saharan Africa carries a disproportionate burden of HIV infections and AIDS-related diseases. In 2016, of the 6000 new HIV infections happening globally every day, 2 out of 3 occurred in sub-Saharan Africa [1]. Global AIDS-related deaths have sharply decreased in the past decade, with sub-Saharan African countries leading this downward trend [2], largely due to the increased access to antiretroviral drugs. AIDS-related mortality has decreased by approximately 40% since 2010 and by 60% since it reached its peak in 2004 [2]. However, the number of new infections has not decreased at the same pace. Although the number of new HIV infections decreased by approximately 23% worldwide since 2010, with Southern and

XSL•FO
RenderX

Eastern Africa showing the largest reduction, in 2019, there were over 1.7 million new infections, of which 1.4 million happened in sub-Saharan Africa (and 59% of whom are women) [2]. More importantly, not every person living with HIV is aware of her condition [3]. Estimates for people aware of their condition in sub-Saharan African countries range between 45% [4] and 65% [3]; hence, the numbers of infected individuals are potentially being underestimated by the hundreds of thousands. To increase the capacity of governments to correctly identify HIV-infected individuals and therefore reduce the spreading of the virus, standard HIV-testing facilities and awareness-raising campaigns are being paired with alternative methods such as community-based testing, self-test kits, and point-of-care testing for early infants [3]. Major barriers to testing (eg, stigma, inaccessibility of services, and low awareness) are being gradually overcome by these new techniques, which are currently being rolled out in numerous sub-Saharan countries [3-5]. Expanding HIV-related knowledge and introducing optimized HIV-testing techniques such as community-based testing and self-testing can potentially increase the returns from national anti-HIV programs [6].

Sub-Saharan Africa has proven to be a technological hothouse when it comes to the expansion of mobile phone coverage and use. The rapid increase in mobile subscribers' numbers and the expansion of network penetration have resulted in sub-Saharan Africa becoming the world's fastest-growing mobile region. Mobile phones facilitate individuals' access to information from sources that would otherwise be unreachable or costly and enable near-instant communication through voice and text (or video, in the case of smartphones). Mobile phone penetration had a pivotal role in the region, influencing economic growth and human development [7,8]. It has been argued that the "communication" component of mobile phones holds the most potential of supporting economic development by creating knowledge-sharing networks [9-11]. Recently, researchers have been investigating the role of mobile phones in public health and found them to be powerful awareness-raising tools and facilitators of individuals' communication with health workers and other patients [12-14]. For mobile phone users, new information is, therefore, acquired mostly through communication with other individuals. However, mobile phones can also facilitate access to institutional resources (eg, by receiving informative SMS text messages from a health clinic, listening to radio, or, for those with access to a smartphone, surfing the internet). Evidence from rural Indonesia has shown that midwives' use of mobile phones increased access to institutional resources and, consequently, fostered better reproductive health knowledge. Further, access to peer resources was associated with higher self-efficacy, which was as well positively associated with health knowledge [15]. Mobile phone ownership has been found to significantly influence contraception use and reproductive health attitudes [16,17]. Phones have also been used in recent years as a new way of reaching individuals with or at risk of HIV with awareness-raising campaigns (see Project Masiluleke in South Africa and Text to Change in Uganda). Finally, recent evidence identified mobile phone ownership as a more relevant determinant of contraceptive use and reproductive health care access than television or radio [18].

Some studies have focused on investigating mobile-based, awareness-raising messaging and adherence to HIV/AIDS treatment programs through randomized controlled trials [19-23]. Lester et al [20], in a seminal study on this topic, examined the association between higher adherence to antiretroviral treatment and SMS text messaging–based support in Kenya. More recent studies have questioned the role of SMS text messaging–based support as a "silver bullet" in resource-limited settings, especially after second-line antiretroviral treatment failure, although it still remains a valuable tool for promoting higher participation in anti-HIV/AIDS programs in general [19]. A study on the role of mobile apps for patients newly diagnosed as HIV-positive found that young individuals in particular are more responsive to mobile-based treatment adherence campaigns [20], while another is planning to test an innovative self-testing service that uses artificial intelligence [23]. However, the impact of mobile technology on HIV and HIV-testing knowledge and uptake might go way beyond the effectiveness of SMS text messaging–based or app-based support.

The lack of literature on the topic makes it difficult to formulate hypotheses on the role played by cellular phones, although it has been shown that innovative, mobile-based, awareness-raising campaigns at community and individual levels represent an economically sensible alternative to traditional awareness-raising campaigns [3]. Moreover, mobile technology–enabled networks facilitate peer-to-peer communication, which has been proven to increase HIV knowledge and treatment adherence while reducing HIV risk [24-26]. Finally, mobile devices may accord privacy in communication, which could facilitate the discussion of a topic such as HIV prevention, which is still stigmatized. We hypothesize that mobile-enabled communication increases HIV- and reproductive health–related knowledge and, through that, it affects HIV-testing knowledge and uptake.

Considering the high number of new infections that happen every year in sub-Saharan African countries and the fact that a large portion of people living with HIV are still unaware of their condition, it is crucial to better understand if and how mobile technology can support HIV prevention and testing. Although there is an emerging body of evidence linking mobile phones with health promotion and awareness, to our knowledge, evidence on the role played by mobile phone ownership or connectivity on HIV-testing knowledge and attitudes, and an investigation of impact pathways, is lacking. Our study seeks to address this knowledge gap. It examines the role mobile connectivity can play in promoting HIV testing through multicountry analysis in sub-Sahara Africa. We focus on mobile connectivity instead of mobile ownership to account for the fact that phone sharing is very common in sub-Saharan African countries [9,13]. Our study is, to our knowledge, the first of its kind to establish a connection between widespread access to mobile phone network and knowledge of HIV-testing facilities and ultimately HIV-testing uptake. We use recent data on cell tower locations and data on women's HIV testing and knowledge in 29 sub-Saharan African countries. In our paper, we also explore the mechanisms through which mobile phone connectivity affects HIV-testing knowledge and uptake. We test if mobile connectivity has a direct effect on levels of HIV

XSL•FO
RenderX

knowledge and whether it influences overall knowledge about reproductive health (ie, knowledge about other sexually transmittable infections [STIs] and about contraceptive methods). We find that indeed mobile connectivity influences general knowledge about HIV and reproductive health.

With our research, we establish a causal link between mobile network connectivity and HIV-testing knowledge and practices using individual-level data from 29 African countries, and we shed light on potential mechanisms for these results. Our work fills a significant gap in HIV-testing literature by providing observational evidence and investigating the role of both mobile connectivity and mobile phone ownership.

Methods

Data

Several data sources have been used to build the dataset used in our analysis.

Cell Connectivity Information

Information on the location of cell towers is derived from the OpenCelliD dataset. OpenCelliD is the largest, daily updated, publicly available database of cell towers. The linear distance between household clusters and the nearest cell tower is calculated in kilometers and serves as a proxy for cell connectivity and signal strength and as our main explanatory variable. OpenCelliD dataset reports the location, operator, and type of signal for over 40 million towers worldwide. Database entries for the African continent (Figure 1) started in 2008. Unfortunately, the date of entry does not unequivocally represent the date of construction of a tower. Because of this, many countries show a total absence of cell towers before 2013, a clearly misleading representation. This is why we calculate the distance from the nearest cell tower including all cell towers in the country regardless of the date in which they have been added to the OpenCelliD dataset. This is done because we assume that towers have most probably been built before their date of entry in the dataset. Since this is a very strong assumption, we test its robustness by running our analysis again, this time limiting the number of cell towers per country to those added to OpenCelliD before the year in which individual survey data were collected. As we will show in the *Results* section, results remain consistent even when the number of cell towers (and, therefore, of countries) is reduced. One more issue with OpenCelliD data is that the tower location is not verified after it is inserted in the database, which might result in a discrepancy between reported and actual location. However, one study comparing independently collected data on cell tower coordinates and OpenCelliD information on the same towers shows that the localization error estimated from OpenCelliD data agrees well with the experimental error distribution [27]. As an extra precaution, however, we consider only towers whose coordinates' maximum error (reported by OpenCelliD) lies within 1 km^2.

Figure 1. Cell phone coverage in sub-Saharan African countries included in our study. Coverage is calculated as a buffer of 35 km around each cell tower for the 29 countries included in the dataset. The image is the authors' elaboration of OpenCelliD data. GSM: Global System for Mobile Communication; LTE: long-term evolution; UMTS: Universal Mobile Telecommunications System.

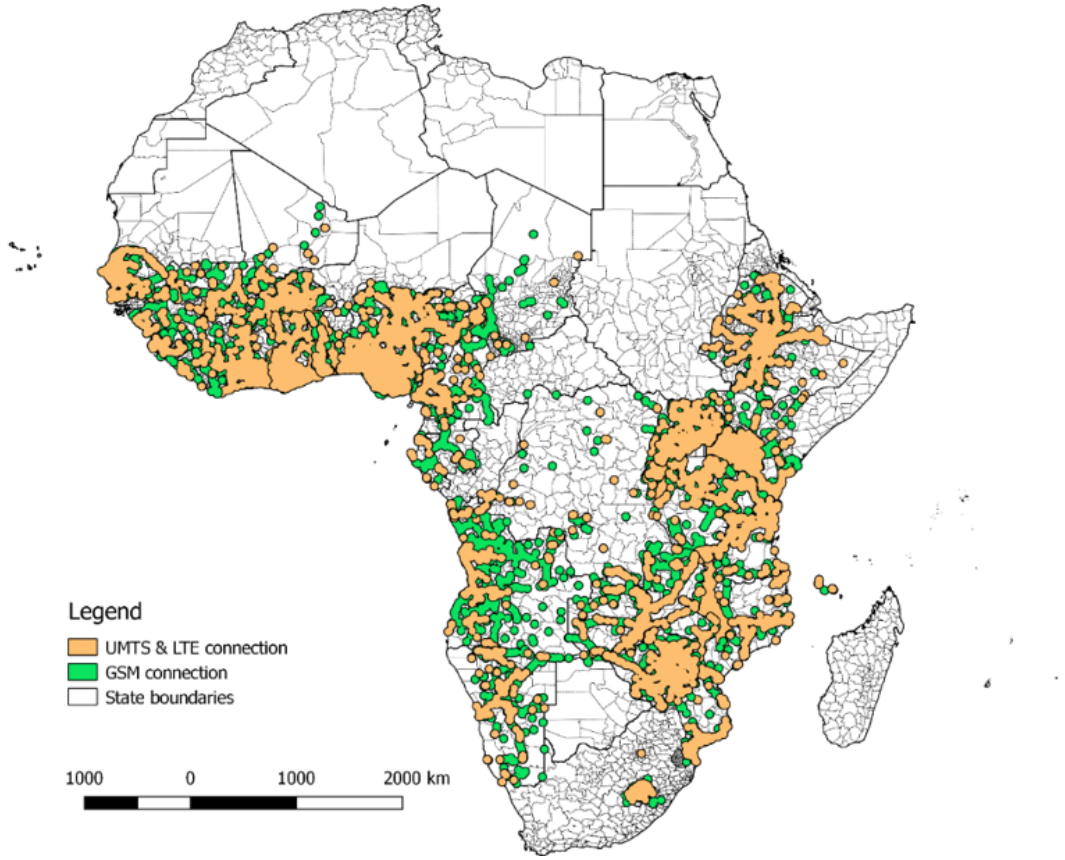

XSL•FO
RenderX

Individual- and Community-Level Data

Individual data are obtained from the Demographic and Health Survey (DHS) program from 29 African countries—Angola, Benin, Burkina Faso, Burundi, Cameroon, Chad, Comoros, Ivory Coast, Democratic Republic of Congo, Ethiopia, Gabon, Ghana, Guinea, Kenya, Lesotho, Liberia, Malawi, Mali, Mozambique, Namibia, Nigeria, Rwanda, Senegal, Sierra Leone, Tanzania, Togo, Uganda, Zambia, and Zimbabwe. DHS provides cross-sectional information on women and household access to health and health status together with demographic and labor statistics. Besides women's demographic and employment information, we take from DHS our main dependent variables: a dummy for knowledge of facilities where women could be tested for HIV and a dummy for ever been tested for HIV. We use additional variables on knowledge of reproductive health to test our hypothesis on possible impact pathways. Individual-level characteristics are added to the analysis to serve as confounders. We include information on the women's age, education, employment, marital status, if ever had children, and exposure to media (ie, newspapers, radio, or television). We also include household-level indicators of gender of the household head, place of living (ie, urban or rural), a wealth index built by DHS and based on selected assets, dwelling construction materials, and types of water access and sanitation facilities. Finally, to address the role played by access to health facilities in facilitating HIV testing, we include in our analysis 2 dummy variables: the first reporting value one for women having visited a health facility in the past 12 months and the second reporting value one for perceiving distance to health facilities as a problem. However, evidence shows that perception of distance might play a bigger role than actual distance when deciding whether to go to a health facility for testing [28]). For each country, DHS included in this study has been conducted between 2010 and 2016. The final sample comprises 407,000 women grouped in 19,603 clusters, of whom, approximately 350,000 are included in the main analysis. As DHS is conducted to be representative at the country level, our study provides estimates relevant for the entire population of the countries involved in it.

Geospatial information of households in DHS datasets is clustered to guarantee respondents' privacy. For these clusters, DHS provides spatial information retrieved from secondary sources within a buffer of 10 km for rural communities and 2 km for urban ones. We include in the analysis community-level (ie, cluster-level) information on average travel time to the nearest settlement (ie, an urban area with at least 50,000 inhabitants) and population density. We also use household-level information on access to electricity to generate a community-level electrification rate indicator. This indicator is simply the percentage of households who have access to electricity within a cluster.

Living a long distance from the nearest cell tower might be one of many indicators of a community's geographical isolation. More isolated communities could have lower level of HIV testing regardless of their use of mobile phones [29]. To better measure community isolation and to control for it in the analysis, we calculate the linear distance of each household's cluster from the nearest primary or secondary road (trails and nonpaved routes have not been considered in the analysis). Information on roads in the African continent is obtained through the Digital Chart of the World (DCW) vector basemap. Although last updated in 1998, DCW remains a very detailed map of Africa's road system. Authors have considered using more recent street maps for the analysis; however, they ultimately decided to make use of DCW due to its reliability. National Aeronautics and Space Administration Global Roads Open Access Dataset, although more recent, does not provide users the chance to adequately distinguish types of roads in the African continent. However, for those countries in which a comparison between DCW and Global Roads Open Access Dataset was possible (eg, Zimbabwe), the authors verified that primary roads reported in the latter were also present in the former (graphical representation in Figure S1 in Multimedia Appendix 1). Distance to the road is used, together with distance to the nearest settlement and community-level electrification rate, as a proxy for community isolation.

Finally, we include information on land gradient and HIV prevalence for each of the clusters. This information is obtained from DHS elaboration of the United States Geological Survey GTOPO30. HIV prevalence (as a number of people who are HIV positive over 100,000) within a 10-km radius from the cluster is obtained from the mapping exercise by Dwyer-Lindgren et al [30], which produced a 5×5-km grid of HIV prevalence among adults for sub-Saharan Africa. Grids are provided for every year between 2000 and 2017. Since 2010 represents the less recent years in which surveys included in our dataset were collected, we use HIV prevalence information from that year. Data are analyzed using QGIS (QGIS Development Team) and Stata (version 16; StataCorp LLC) software.

Sample Characteristics

Table 1 reports descriptive statistics from the sample. The average community distance from the cell tower is 13.06 km. This is considering both rural and urban areas. The average woman's age is 28.46 years, with 5 years of education. About 28.34% of interviewed women live in a female-headed household, 72.05% are exposed to at least 1 type of media (ie, radio, television, and newspaper), 52.34% are in a union, 58.60% are currently working, and 35.40% live in an urban area. Half of the women in the sample report to have visited a health facility in the last 12 months, although distance from said facility is considered a problem by 41.04% of them. A total of 72.66% of sampled women have given birth to at least 1 child. The average population density in the communities was 1290 people per km^2 (this is considering a buffer of 2 km^2 for urban areas and 10 km^2 for rural areas, 2015 population data). Distance from the closest road (trails and nonpaved routes have been excluded from the analysis) is 6.12 km on average, and the average travel time to the closest settlement is about 77.48 minutes. Finally, the prevalence of households with access to electricity in the community is about 31.81% on average.

XSL•FO
RenderX

Table 1. Descriptive statistics of the sample.

Variables	Values	Total, N
Distance from the nearest cell tower (km), mean	13.06	517,290
Woman age, mean	28.46	517,290
Woman years of education, mean	5.18	517,110
Female-headed households (%)	28.34	517,290
Woman exposed to any type of media (%)	72.05	517,257
Woman currently working (%)	58.60	500,001
Ever had a child (%)	72.66	517,290
Number of lifetime sex partners, mean	2.16	407,542
Woman in union (%)	52.34	517,284
Living in urban area (%)	35.40	517,290
Population density in community (2015)[a], mean	1290.64	516,411
Wealth index quintiles (%)		517,290
Poorest	19.42	
Poorer	18.56	
Middle	19.01	
Richer	19.96	
Richest	23.05	
Travel time to closest settlement (minutes)[a], mean	77.48	516,411
Distance from the closest road (km)[a], mean	6.12	517,290
HIV prevalence in 10-km radius, mean	5.05	517,022
Visited a health facility in the past 12 months (%)	50.31	499,767
Distance to a health facility is considered a problem (%)	41.04	458,976
Electrification (cluster level)[a] (%)	31.81	517,290

[a]Community-level variables obtained by averaging household figures at the community level.

Statistical Methods

Using probit and ordinary least square regressions, we design a naive model at first.

$$\text{HIV}t_i=\beta_0+\beta_1 D_i+\beta_2 Z_i+\beta_3 C+u_i \quad \textbf{(1)}$$

Where HIVt_i is a dummy variable indicating if either woman knows where the HIV-testing facility is or has ever been tested for HIV. β_1 is a vector of our main explanatory variable or distance to nearest cell tower D for woman i. β_2 is a vector of covariates Z for woman i. A full list of covariates is present in Table 1. Finally, β_3 is a vector of community-level characteristics and country-fixed effects C. Location information, from which distance from cell towers was calculated, could only be obtained at the community level. Community clusters contain between 1 and 92 households each. To address intracluster correlation, we cluster SEs at the community level in each regression. Common practice is to cluster SEs at the most aggregate level [31]. However, it is important to consider that in our case sampling has been done at the cluster level and, more importantly, the treatment (ie, distance from the cell tower) is measured at the community level. Correlation in SEs at the community level is, therefore, very plausible, and clustering errors at a higher level might be unnecessarily conservative [32]. The authors decided to present results with SEs clustered at the community level to address such correlation. However, robustness checks have been performed by clustering errors at the country and administrative area levels. Both models produced similar and significant results (results available upon request). Denormalized individual sample weights are included. Individual women's weights have been calculated by denormalizing DHS individual weights following Ruilin Ren's note on DHS weight denormalization. As a robustness check, analysis has been run without weights, and results remain consistent (results available upon request).

As Klonner and Nolen [33] observe, the distribution of cell towers is not exogenous over a country's territory. It might follow an established path (ie, roads) and be affected by the power supply and other infrastructural needs. Moreover, network providers might consider certain areas not profitable enough to repay them for the investment of building a cell tower. Figure 2 shows this endogenous distribution in the context of Zimbabwe.

XSL•FO
RenderX

Figure 2. Distribution of cell towers in Zimbabwe. The image is the authors' elaboration on OpenCelliD data.

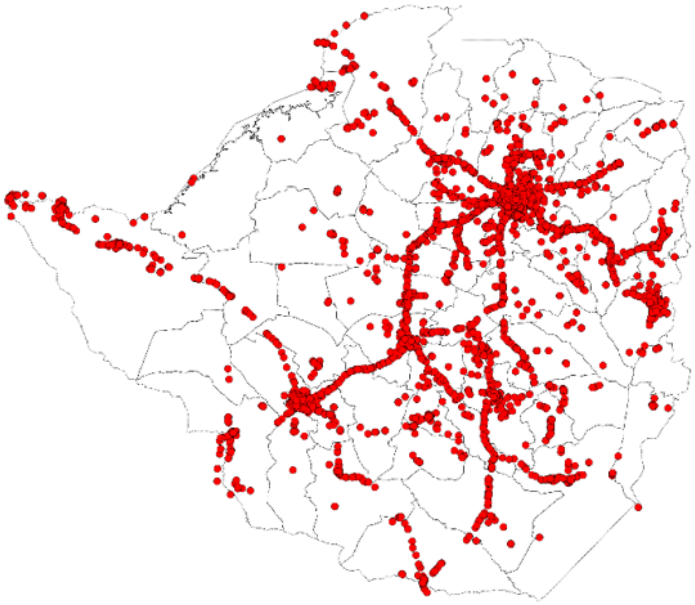

To deal with endogeneity, we look for an instrument that could influence towers' location but is uncorrelated with our dependent variable error terms. Following Dinkelman et al [34], we use the average community land gradient as an instrumental variable (IV). Terrain slope raises the costs of electric line routing and antennas' building [33,35,36], acting as a disincentive for mobile network companies to build cell towers in areas with higher land gradient. On the contrary, more competitive markets in sub-Saharan Africa can lead to a considerable increase in cell phone area coverage [37]. Proxies for market competition have been used in the past in tandem with terrain roughness to explain the positioning of cell towers [38]. We build our IV interacting land gradient and number of phone service providers in each country following Berman et al [39]. We add to the list of covariates information on household electrification. In this way, we control for effects related to access to electricity regardless of whether a cell phone network is available. Our second model, a 2-stage regression is, therefore:

$$D_i = \pi_0 + \pi_1 Slope \times Network\ providers_i + \pi_2 Z_i + \pi_3 Elec_i + \pi_4 C + \varepsilon_i \quad \textbf{(2)}$$

In equation 2, the measure of land gradient *Slope* multiplied by number of *Network providers* is the exogenous regressors. Electrification *Elec* ranges between 0 and 1 and represents the percentage of households who have access to electricity in the community. Equation 3 is similar to equation 1, although here distance *D* has been instrumented by *Slope*×*Network providers*. The land gradient is calculated by DHS based on the United States Geological Survey Global 30 arc-second elevation model (GTOPO30) [40], and we consider it invariant through time (or, at least, within the same geological era). This model is estimated via 2-stage least squares (2SLS).

The identification assumption is that, conditional on individual and community characteristics, our IV does not directly affect knowledge of facilities or decision to test for HIV and only does so through the distance from the nearest cell tower. To check our assumption, we perform several placebo tests. Mobile phone signal spreads unevenly in the area surrounding cell towers due to a variety of factors. Given the geographical extent of our analysis, we are not able to calculate network coverage exactly. However, technical limitations of standard Global System for Mobile Communication (GSM) towers set their maximum signal range to a 35-km radius; therefore, communities located more than 35 km away from the closest cell tower (9.23% of the sample) can be considered not covered by the mobile network. The 35-km maximum range is the one imposed by technical limitations on GSM towers; however, it does not take into account barriers to signal that might be present within the radius of the cell tower. This is why the dummy build with this cutoff point is used only as an addition to our main explanatory variable (ie, distance from the cell tower). See 3GPP for additional information on timing advance and time division multiple access technology and the way they influence GSM tower range. For the sake of standardization, we assume towers do not possess features to extend their range. Using this cutoff point, which theoretically divides the sample into connected and nonconnected communities, we examine whether our IV only affects outcomes when communities are covered by a mobile network. Results (Table S1 in Multimedia Appendix 1) demonstrate that our instrument is only correlated with outcomes in network-covered communities. Our second test investigates whether the IV or instrumented distance from the cell tower is in any way correlated to proxies of communities' remoteness, a factor that could in turn affect knowledge of HIV testing place and test uptake. We select 3 different proxies: distance from the nearest road (excluding trails and unpaved roads), exposure to any type of media as proxies for remoteness, and average population density in the community as a proxy for dispersion. Both the instrument and the instrumented distance from the nearest cell tower show no correlation with any of the proxies, supporting our initial assumption that a longer distance from a cell tower is not necessarily related to the remoteness of a

XSL•FO
RenderX

community (Table S2 in Multimedia Appendix 1 for full results). Finally, we test the sensibility of our IV. Our results might be driven by countries where markets are more accessible for mobile phone providers. Competition might decrease mobile technology consumption costs and facilitate knowledge sharing. We run our models once again, this time dividing the sample based on the number of mobile network providers in the country of residence. Results remain consistently negative throughout the models (Figure 3), confirming that the relationship between distance from cell towers and our outcomes does not change based on mobile network market characteristics.

Figure 3. Placebo test. Number of providers correspond to, from left to right, the 50th percentile of the sample, the 75th percentile, the 90th percentile, and the 100th percentile.

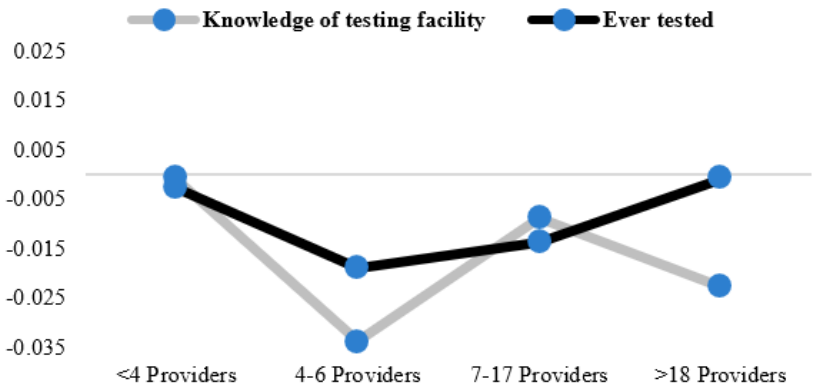

Ethical Considerations

No ethical approval was required for this study as it uses secondary data from DHS. These data are available from DHS on request and received approval from the International Coaching Federation Institutional Review Board. Study data are anonymous.

Results

Mobile Phone Connectivity and Knowledge of HIV-Testing Facilities

Our analysis examines whether better mobile connectivity is correlated with women's knowledge of which facilities offer HIV testing. Results are presented in Table 2. Probit regression results show a strong negative correlation ($P<.001$) between distance from the cell tower and knowledge of an HIV-testing facility. Results from the 2SLS estimation are consistent with our probit results. Due to the nature of the variables, interpreting the magnitude of the impact is not straightforward. We could infer that for each unit increase in distance to the nearest cell tower, the probability of knowing where one can be tested for HIV is 0.7 percentage points lower (0.2 percentage points for probit analysis). To facilitate the interpretation of results and to put them in the right order of magnitude, we report (in Figure S2 in Multimedia Appendix 1) a graphical illustration of our IV model's predicted share of women knowledgeable about testing facilities by deciles of distance to the nearest cell tower. As it can be noticed, 90% or more of women living in the top deciles are knowledgeable about HIV-testing facility locations. The share starts dropping starting from the fourth decile, going below 80% in the bottom 2 deciles (ie, with a distance >15 km), and reaching 40% in the lowest decile (ie, distance >33 km).

The clustering of households' coordinates allows us to calculate the distance from the nearest cell tower at the community level only. To check whether this might affect our results, we perform community-level analysis by averaging women's characteristics at the community level. The impact of distance to cell towers remains highly significant ($P<.001$). The magnitude of the impact also remains consistent, with a 0.8-percentage point decrease in the probability of knowing the testing facility (0.1 percentage points for probit). Underidentification and weak-instrument (Kleibergen-Paap Wald F statistic) tests for all IV regressions show that our instrument is relevant and reasonably strong.

XSL•FO
RenderX

Table 2. Impact of phone connectivity on knowledge of HIV testing facilities.

	Knowledge of HIV testing facility			
	Individual level		Community level	
	Probit	IV[a,b] (2SLS[c])	Probit	IV[d] (2SLS)
Distance from the nearest cell tower (km) (cluster-robust SEs)	–0.002[e] (<0.001)	–0.007[e] (0.002)	–0.001[e] (<0.001)	–0.008[e] (0.002)
Individual-level characteristic	Yes	Yes	__[f]	—
Community-level characteristics	Yes	Yes	Yes	Yes
Country-year-fixed effects	Yes	Yes	Yes	Yes
Constant (cluster-robust SEs)	–0.547[e] (0.058)	0.505[e] (0.026)	0.304[e] (0.021)	0.399[e] (0.038)
Observations	351,613	351,488	18,437	18,431
R^2	0.307	0.183	0.682	0.402

[a]IV: instrumental variable.

[b]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 32.13.

[c]2SLS: 2-stage least squares.

[d]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 28.74.

[e]*P*<.001.

[f]Not available.

From Knowledge to Action—Mobile Connectivity and HIV Testing

We have identified a significantly positive effect of proximity to cell towers on knowledge of HIV-testing facilities. It is important to also determine whether knowledge acquired through mobile phones ultimately translates into a behavioral change. We therefore investigate whether women living closer to cell towers not only acquire knowledge on where HIV testing can be conducted but also get tested.

Newly acquired knowledge affects individuals' health behavior through what has been called the knowledge-attitude-behavior continuum [41]. Studies show that when agents act on the way information is processed by individuals, one could expect desired behavioral changes to ultimately happen [41]. Activities aimed at increasing knowledge about HIV transmission among pregnant women have been shown to change attitudes and behavior toward HIV testing [42,43]. At the same time, access to peer networks can influence individuals' attitudes toward HIV testing. Studies in the past have found that peer-based interventions can increase HIV knowledge and adherence to AIDS treatment [44-46].

We examine whether better mobile connectivity has a direct effect on women's probability of ever been tested for HIV. Our analysis is operationalized in a variation of equations 1 and 3, where having tested for HIV is now the dependent variable, and the explanatory variable is the distance from the nearest cell tower. Table 3 reports the results of our analysis. We find that a unit increase in distance from cell towers significantly reduces women's probability of ever been tested for HIV by 0.5 percentage points in the individual-level model and by 0.4 percentage points in the community-level model. Again, we present in Figure S2 in Multimedia Appendix 1 a graphical illustration of our IV model's predicted share of women tested by deciles of distance to the nearest cell tower. Similarly to the knowledge of testing facilities, the share of tested women is constant at around 70% in the top 4 deciles. At the start of the bottom 2 deciles, the share of tested women is already below 60%, and it reaches approximately 10% in the lowest decile.

XSL•FO
RenderX

Table 3. Impact of phone connectivity on HIV testing uptake.

	Ever tested for HIV			
	Individual level		Community level	
	Probit	IV[a,b]	Probit	IV[c]
Distance from the nearest cell tower (km) (cluster-robust SEs)	–0.005[d] (0.001)	–0.005[e] (0.002)	–0.001[d] (<0.001)	–0.005[d] (0.002)
Individual-level characteristic	Yes	Yes	—[f]	—
Community-level characteristics	Yes	Yes	Yes	Yes
Country-year-fixed effects	Yes	Yes	Yes	Yes
Constant (cluster-robust SEs)	–1.383[d] (0.050)	0.147[d] (0.025)	–0.036[g] (0.019)	0.014 (0.029)
Observations	365,108	364,982	18,480	18,474
R^2	0.323	0.343	0.793	0.734

[a]IV: instrumental variable.

[b]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 34.21.

[c]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 29.08.

[d]$P<.001$.

[e]$P=.02$.

[f]Not available.

[g]$P=.08$.

Robustness Checks

Reduced Number of Cell Towers

As a first robustness check, we examine whether the results remain significant when the distance from the nearest cell tower is calculated using only the cell towers inserted in the OpenCelliD dataset starting from the year in which each DHS was collected. Several countries are dropped from this analysis due to the complete absence of data on cell towers from the year of the survey onward, namely, Cameroon, Comoros, Gabon, Guinea, Mozambique, and Senegal. However, the figures presented in Table 4 show that results remain significant and consistent with previous findings.

XSL•FO
RenderX

Table 4. Impact of phone connectivity on HIV testing (restricted cell towers' sample).

	Knowledge of HIV testing facility				Ever tested for HIV			
	Individual level		Community level		Individual level		Community level	
	OLS[a]	IV[b,c]	OLS	IV[d]	OLS	IV[e]	OLS	IV[f]
Distance from the nearest cell tower (km) (cluster-robust SEs)	–<0.001[g] (<0.001)	–0.003[h] (0.001)	–<0.001[h] (<0.001)	–0.002[h] (<0.001)	–<0.001[h] (<0.001)	–0.002[i] (0.001)	–<0.001[h] (<0.001)	–0.001[j] (<0.001)
Individual-level characteristic	Yes	Yes	—[k]	—	Yes	Yes	—	—
Community-level characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year-fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant (cluster-robust SEs)	–0.636[h] (0.066)	0.475[h] (0.025)	0.335[h] (0.025)	0.369[h] (0.041)	–1.568[h] (0.059)	0.118[h] (0.023)	–0.043[l] (0.022)	–0.019 (0.032)
Observations	247,902	247,777	12,867	12,861	261,802	261,676	12,910	12,904
Pseudo-R^2	0.315	0.0897	0.676	0.234	0.354	0.312	0.814	0.643

[a]OLS: ordinary least square.

[b]IV: instrumental variable.

[c]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 41.26.

[d]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 13.48.

[e]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 43.80.

[f]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 13.41.

[g]P=.84.

[h]P<.001.

[i]P=.004.

[j]P=.007.

[k]Not available.

[l]P=.74.

Impact on Other Health Care Seeking and Use

One question that can arise when looking at our results is whether the impact of mobile connectivity on knowledge of HIV-testing facilities and testing uptake is driven by a general impact of access to mobile phones on health seeking and use. For example, better connectivity may increase demand for health services from the population while at the same time reducing costs to access such services (eg, increased knowledge and better networking may provide individuals with more options for obtaining health care). To test whether this is the case, we repeat our analysis and replace our outcomes of interest with dummy variables for having health insurance and having delivered one's last child in a safe structure ("safe structure" is defined here as either a public or a private health center). The results, presented in Table S3 in Multimedia Appendix 1, are not statistically significant in any of our preferred models (ie, 2SLS estimations). These results suggest that increased mobile connectivity does not have a general impact on health care seeking and use.

Alternative IV

As we have shown in our Statistical Methods section, terrain roughness and market competition are often associated with infrastructural development. However, there are additional characteristics that influence cell tower positioning in developed and developing countries alike. As an additional robustness check, we test the validity of lightning intensity and cost of mobile phone use as alternative instruments.

Frequent electrostatic discharges damage mobile phone infrastructures and negatively affect connectivity. Building cell towers in areas with high lightning intensity represents a risk for mobile phone provides since power surge protection for cell towers is costly and reduces profits from their investments. Being Africa the continent with the highest lightning intensity in the world, evidence has shown that a 1 SD increase in lightning intensity leads to a lower penetration rate of mobile phone technology of approximately 0.43 percentage points per year [47]. Mobile phone use depends on a variety of factors, of which cost is surely one of the most important [48]. Higher costs might be driven by low competition and, at the same time, act as an incentive for new companies to provide network access to a larger share of the population.

We interact the average intensity of lightning strikes in a 10-km radius around each DHS cluster with a country-level monthly cost of running a mobile phone as a percentage of per-capita income. Lightning strikes' intensity data are publicly available at the Global Hydrology Resource Center. The satellite data are collected by the National Aeronautics and Space Administration's Global Hydrology Resource Center. We included in our analysis yearly average lightning strikes for countries included in the analysis for 2010. Lightning strike incidence is proved to be stable through years, making the yearly average for 2010 a good proxy for the subsequent decade [47]. Information on the monthly cost of running mobile phones by country is retrieved from the 2015 International

XSL•FO
RenderX

Telecommunication Union report on Measuring the Information Society. A slight modification to equation 2 produces

$$D_i = \pi_0 + \pi_1 Flash\ intensity \times Mobile\ cost_i + \pi_2 Z_i + \pi_3 Elec_i + \pi_4 C + \varepsilon_i \quad (4)$$

Analysis conducted using the alternative IV produces results consistent with previous findings, which remain statistically significant (results presented in Table S4 in Multimedia Appendix 1).

Mobile Phone Ownership Instead of Connectivity

We have focused our analysis on the impact of mobile network connectivity on women's HIV and health knowledge and practices regardless of their ownership of a cellular phone. We did so because evidence has shown that mobile phone sharing is a common practice in sub-Saharan Africa [9]. It is the intention of the authors to argue for the potential of better phone connectivity regardless of increased ownership, and our results have supported our claims. However, it is logical to think that the easier the access to a mobile phone, the more relevant its impact on our outcomes of interest should be, and ownership of a mobile phone grants a much higher exposure to mobile technology than borrowing. DHS has recently started including questions on mobile phone ownership in its questionnaires as part of individuals' and households' assets. This provides us with a reduced sample of women who reported whether or not they own a cellular phone. Information on phone ownership is present for Angola (2015), Burundi (2016), Ethiopia (2016), Malawi (2015), Tanzania (2015), Uganda (2016), and Zimbabwe (2015).

To analyze the role of phone ownership, we make use of probit models as in equation 1, while we specify a new probit-2SLS model for the IV analysis (in this case, we use the same instrument from our main analysis, the interaction between land gradient and number of network providers). Given the binary nature of the phone ownership variable, a binary probit model for the first stage regression is more adequate. We present the results in Table 5. According to our IV estimation, phone ownership increases the chance of knowing a facility that conducts HIV tests by 8 percentage points, and it increases the chance of ever been tested for HIV by approximately 14 percentage points. Considering the percentage of people knowledgeable about HIV-testing place and ever tested for HIV who do not own a phone, our percentage points translate into a 9% higher chance of knowing a testing facility and a 22% higher chance of having tested for HIV for mobile phone owners.

Table 5. Phone ownership and HIV testing.

	Knowledge of HIV testing facility		Ever tested for HIV	
	Probit	IV[a]	Probit	IV
Mobile phone owner (cluster-robust SEs)	0.220[b] (0.040)	0.078[b] (0.016)	0.240[b] (0.028)	0.142[b] (0.025)
Individual-level characteristic	Yes	Yes	Yes	Yes
Community-level characteristics	Yes	Yes	Yes	Yes
Country-year-fixed effect	Yes	Yes	Yes	Yes
Constant (cluster-robust SEs)	–0.765[b] (0.148)	0.708[b] (0.012)	–1.654[b] (0.101)	0.183[b] (0.015)
Observations	76,241	76,241	81,048	81,048
Pseudo-R^2	0.300	0.183	0.318	0.320

[a]IV: instrumental variable.

[b]$P<.001$.

Mechanisms for Impact

Cell phone connectivity has been found to positively impact the probability of knowing HIV-testing facilities and of ever been tested for HIV. Through which mechanisms such impact is achieved is what we investigate in this section. To do so, we need to understand the barriers to HIV testing that mobile phone connectivity might help overcome. Studies have shown that stigma, both self and social, remains the most significant barrier to HIV testing in sub-Saharan Africa [49,50]. The provision of HIV counseling for patients at high risk has yielded good results among adolescents thanks to increased knowledge and awareness, although cultural change is hard to achieve [51]. Physical barriers to HIV testing are also limiting the effectiveness of awareness-raising programs. Distance from health facilities, inconvenient opening hours, and long waiting lines are among the reasons cited for low test uptake [51].

Mobile connectivity has been proven to be a powerful tool to spread information and increase social cohesion in the Global South [52]. Mobile phones also facilitate knowledge diffusion, which in turn reduces information asymmetry [7,53]. We argue that, while better mobile phone connectivity might do little to reduce physical barriers to HIV testing, it can increase women's knowledge of HIV-related symptoms, help fight HIV misconceptions, and foster reproductive health knowledge.

To test our hypotheses, we perform an analysis on the role of distance from the nearest cell tower on several intermediate outcomes. We use knowledge of HIV and STIs and knowledge of modern contraceptive methods as dependent variables. HIV knowledge is measured, following standard DHS guidelines, as an index ranging from 0 to 1 measuring knowledge of HIV symptoms and preventive measures [54]. Knowledge of STIs and modern contraceptive methods are dummies assigning value

XSL•FO
RenderX

1 to knowledgeable women. The results of the analysis are presented in Table 6.

Distance from cell towers appears to have a significantly negative impact on HIV- and STI-related knowledge. In the same way, it also reduces the chance that a woman would be knowledgeable about modern contraceptive methods. These results seem to confirm our hypothesis that better cellular connectivity facilitates the acquisition of knowledge on reproductive health and sexually transmittable diseases. These findings are supported by previous evidence on the role played by mobile phones in knowledge-sharing and learning [15,55,56].

Table 6. Analysis of mechanisms of cell connectivity and HIV testing correlation.

	Knowledge of HIV		Knowledge of STIs[a]		Knowledge of contraceptive method	
	OLS[b]	IV[c,d]	Probit	IV[e]	Probit	IV[e]
Distance from the nearest cell tower (km) (cluster-robust SEs)	<–0.001[f] (<0.001)	–0.004[g] (0.001)	–0.003[g] (0.001)	–0.002[g] (0.001)	–0.002[g] (0.001)	–0.004[g] (0.001)
Individual-level characteristic	Yes	Yes	Yes	Yes	Yes	Yes
Community-level characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Country-year-fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Constant (cluster-robust SEs)	0.577[g] (0.012)	0.760[g] (0.022)	–0.153[h] (0.077)	0.741[g] (0.016)	–0.660[g] (0.069)	0.640[g] (0.018)
Observations	340,062	339,965	359,431	368,812	369,000	368,874
Pseudo-R^2	0.222	0.114	0.258	0.0505	0.345	0.0784

[a]STI: sexually transmittable infection.

[b]OLS: ordinary least square.

[c]IV: instrumental variable.

[d]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 31.52.

[e]Underidentification (*P* value) test: <.001 and weak-instrument (Kleibergen-Paap rk Wald *F* statistic) test: 34.01.

[f]P=.02.

[g]P<.001.

[h]P=.03.

To complete our analysis on mechanisms, we test whether results remain consistent when considering mobile phone ownership instead of mobile connectivity. Findings are reported in Table 7. Our results show that phone ownership significantly increases knowledge about STIs and modern contraceptive methods, while it seems to have no statistically significant impact on general knowledge of HIV.

Recent evidence has shown that mobile phone ownership might have an impact on contraceptive use and knowledge about family planning [57] and facilitate the spread of reproductive, maternal, newborn, and child health in vulnerable settings [58]. At the same time, some research has argued that better outcomes for phone owners are mostly driven by their economic status [59]. In our analysis, we control for additional factors that could influence women's knowledge, such as media exposure, education, asset-based wealth, and place of residence, and the effect of mobile ownership remains statistically significant.

XSL•FO
RenderX

Table 7. Phone ownership and mechanisms.

	Knowledge of HIV		Knowledge of STIs[a]		Knowledge of contraceptive method	
	Probit	IV[b]	Probit	IV	Probit	IV
Mobile phone owner (cluster-robust SEs)	0.012[c] (0.005)	0.006[d] (0.013)	0.243[c] (0.050)	0.053[c] (0.016)	0.193[c] (0.051)	0.089[c] (0.015)
Individual-level characteristic	Yes	Yes	Yes	Yes	Yes	Yes
Community-level characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Country-year-fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Constant (cluster-robust SEs)	0.651[c] (0.019)	0.702[c] (0.011)	–0.527[c] (0.129)	0.722[c] (0.014)	–0.297[e] (0.137)	0.733[c] (0.014)
Observations	66,344	66,344	81,048	81,048	81,048	81,048
Pseudo-R^2	—[f]	—	0.325	—	0.478	—
R^2	0.296	0.239	—	0.174	—	0.236

[a]STI: sexually transmittable infection.

[b]IV: instrumental variable.

[c]$P<.001$.

[d]$P=.64$.

[e]$P=.03$.

[f]Not available.

Discussion

Principal Findings

This paper analyses the impact of cell phone connectivity on women's HIV-testing knowledge and behavior. We used distance from the nearest cell tower as a proxy for phone connectivity strength and quality, and we estimated how increasing distance from said cell tower affects the probability of knowing where one can be tested for HIV and the probability of having taken such a test at least once in one's life. Mobile phones enhance individuals' access to information thanks primarily to knowledge-sharing through communication, although they can also put people in contact with institutional resources [15,56]. Due to phone-sharing practices in sub-Saharan Africa, we used cell phone connectivity as the primary indicator for analysis, with mobile phone ownership used for robustness check. We attempt to expand on the work that has been done on a smaller scale on the role played by mobile technology on HIV awareness-raising and testing [19,21]. We make use of a pooled cross-sectional dataset from 29 sub-Saharan African countries and different methods of analysis to produce robust inferences on the role played by phone connectivity.

Results show that mobile connectivity has a statistically significant and positive impact on the knowledge of HIV-testing facilities and having been tested for HIV, highlighting its role as an important determinant of these practices. Results are consistent in all models and at both individual and community levels. We test the hypothesis and find evidence that the impact of mobile phone connectivity on HIV-testing knowledge and behavior might be driven by improved knowledge of HIV, STIs, and contraceptive methods. Moreover, we find that the positive effects of mobile technology are even greater when we consider phone ownership as the explanatory variable.

Comparison to Prior Work

Our results are in agreement with previous evidence that shows the role mobile technology can play in raising awareness about HIV and HIV testing [60]. The study also fits well into the new body of literature investigating mobile health and the role it could play in developing countries. Promising evidence shows how informal mobile health and access to mobile phones among young Africans can help bridge health care gaps in countries such as Ghana, Malawi, and South Africa [61,62]. Not only mobile phones but also all digital technology has been found to be beneficial for treating and preventing mental disorders in low-income and middle-income countries [63]. Our results support the idea that stronger mobile connectivity can foster HIV-testing practices in sub-Saharan Africa.

Strength and Limitations

Our study is not exempt from limitations. First, although we are able to establish a link between mobile connectivity and knowledge of HIV-testing facilities and mobile connectivity and HIV-testing uptake, due to data limitations, we cannot dig deeper into how mobile connectivity shapes women's attitudes toward testing, since no questions were asked to interviewed women about their opinion of HIV testing. Second, the OpenCelliD dataset is a relatively new dataset that might be lacking detail especially for remote areas. Missing information on the date in which cell towers are built also poses a limitation. We addressed both these issues by only including cell towers with the most reliable GPS coordinates and by performing robustness checks that aligned the building year of the cell towers with the year of the survey. When and if data will be

XSL•FO
RenderX

available in the future, additional longitudinal analysis could be conducted to investigate the evolution of HIV-testing behavior through time for women with better mobile connectivity. A third limitation is represented by the clustering of individuals' positions under 1 GPS coordinate representing the community. This clustering allows distance from the cell tower to be calculated only at the community level. We show, however, that our results remain significant even if individual figures are averaged at the cluster level, and community-level analysis is performed.

Future Directions

Our study provides useful insights on the potential that new communication technologies represent to foster better reproductive health practices in sub-Saharan Africa and paves the way for future analysis on how mobile phones enhance knowledge-sharing around HIV/AIDS knowledge, behavior, and practices.

Data Availability

All data used in this study are available on the Demographic and Health Survey repository, pending authorization, and the OpenCelliD repository.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional robustness tests.
[DOCX File , 355 KB - jmir_v26i1e48794_app1.docx]

References

1. Kharsany ABM, Karim QA. HIV infection and AIDS in sub-Saharan Africa: current status, challenges and opportunities. Open AIDS J 2016;10:34-48 [FREE Full text] [doi: 10.2174/1874613601610010034] [Medline: 27347270]
2. UNAIDS data 2019. UNAIDS. 2019. URL: https://www.unaids.org/sites/default/files/media_asset/2019-UNAIDS-data_en.pdf [accessed 2020-08-03]
3. Knowledge is power: Know your status, know your viral load. UNAIDS. 2018. URL: https://www.unaids.org/sites/default/files/media_asset/jc2940_knowledge-is-power-report_en.pdf [accessed 2020-08-03]
4. Sweat M, Fonner V. Overcoming resistance to HIV testing in sub-Saharan Africa. Lancet HIV 2016;3(3):e106-e107. [doi: 10.1016/S2352-3018(16)00004-7] [Medline: 26939730]
5. Jani IV, Meggi B, Loquiha O, Tobaiwa O, Mudenyanga C, Zitha A, et al. Effect of point-of-care early infant diagnosis on antiretroviral therapy initiation and retention of patients. AIDS 2018;32(11):1453-1463. [doi: 10.1097/QAD.0000000000001846] [Medline: 29746301]
6. Swann M. Economic strengthening for HIV testing and linkage to care: a review of the evidence. AIDS Care 2018;30(sup3):85-98. [doi: 10.1080/09540121.2018.1476665] [Medline: 29985055]
7. Asongu SA, Boateng A, Akamavi RK. Mobile phone innovation and inclusive human development: evidence from sub-Saharan Africa. African Gov Dev Inst 2016:36. [doi: 10.2139/ssrn.2822103]
8. Lee SH, Levendis J, Gutierrez L. Telecommunications and economic growth: an empirical analysis of sub-Saharan Africa. Appl Econ Routledge 2012;44(4):461-469. [doi: 10.1080/00036846.2010.508730]
9. Aker JC, Mbiti IM. Mobile phones and economic development in Africa. J Econ Perspect 2010;24(3):207-230 [FREE Full text] [doi: 10.1257/jep.24.3.207]
10. Donner J. The use of mobile phones by microentrepreneurs in Kigali, Rwanda: changes to social and business networks. Inf Technol Int Dev 2006;3(2):3-19 [FREE Full text]
11. Nakato GV, Beed F, Bouwmeester H, Ramathani I, Mpiira S, Kubiriba J, et al. Building agricultural networks of farmers and scientists via mobile phones: case study of banana disease surveillance in Uganda. Can J Plant Pathol 2016;38(3):307-316. [doi: 10.1080/07060661.2016.1230149]
12. Anstey Watkins JOT, Goudge J, Gómez-Olivé FX, Griffiths F. Mobile phone use among patients and health workers to enhance primary healthcare: a qualitative study in rural South Africa. Soc Sci Med 2018;198:139-147 [FREE Full text] [doi: 10.1016/j.socscimed.2018.01.011] [Medline: 29335160]
13. Kaplan WA. Can the ubiquitous power of mobile phones be used to improve health outcomes in developing countries? Global Health 2006;2(1):9 [FREE Full text] [doi: 10.1186/1744-8603-2-9] [Medline: 16719925]
14. Chen S, Fang Y, Chan PSF, Kawuki J, Mo P, Wang Z. Counseling supporting HIV self-testing and linkage to care among men who have sex with men: systematic review and meta-analysis. JMIR Public Health Surveill 2024;10(1):e45647 [FREE Full text] [doi: 10.2196/45647] [Medline: 38265866]
15. Lee S, Chib A, Kim JN. Midwives' cell phone use and health knowledge in rural communities. J Health Commun 2011;16(9):1006-1023. [doi: 10.1080/10810730.2011.571344]

XSL•FO
RenderX

16. Greenleaf AR, Ahmed S, Moreau C, Guiella G, Choi Y. Cell phone ownership and modern contraceptive use in Burkina Faso: implications for research and interventions using mobile technology. Contraception 2019;99(3):170-174 [FREE Full text] [doi: 10.1016/j.contraception.2018.11.006] [Medline: 30468721]
17. LeFevre AE, Shah N, Bashingwa JJH, George AS, Mohan D. Does women's mobile phone ownership matter for health? Evidence from 15 countries. BMJ Glob Health 2020;5(5):e002524 [FREE Full text] [doi: 10.1136/bmjgh-2020-002524] [Medline: 32424014]
18. Iacoella F, Gassmann F, Tirivayi N. Which communication technology is effective for promoting reproductive health? Television, radio, and mobile phones in sub-Saharan Africa. PLoS One 2022;17(8):e0272501 [FREE Full text] [doi: 10.1371/journal.pone.0272501] [Medline: 35976900]
19. Gross R, Ritz J, Hughes MD, Salata R, Mugyenyi P, Hogg E, et al. Two-way mobile phone intervention compared with standard-of-care adherence support after second-line antiretroviral therapy failure: a multinational, randomised controlled trial. Lancet Digit Health 2019;1(1):e26-e34 [FREE Full text] [doi: 10.1016/S2589-7500(19)30006-8] [Medline: 31528850]
20. Lester RT, Ritvo P, Mills EJ, Kariri A, Karanja S, Chung MH, et al. Effects of a mobile phone short message service on antiretroviral treatment adherence in Kenya (WelTel Kenya1): a randomised trial. Lancet 2010;376(9755):1838-1845. [doi: 10.1016/S0140-6736(10)61997-6] [Medline: 21071074]
21. Amankwaa I, Boateng D, Quansah DY, Akuoko CP, Evans C. Effectiveness of short message services and voice call interventions for antiretroviral therapy adherence and other outcomes: a systematic review and meta-analysis. PLoS One 2018;13(9):e0204091 [FREE Full text] [doi: 10.1371/journal.pone.0204091] [Medline: 30240417]
22. Kliner M, Knight A, Mamvura C, Wright J, Walley J. Using no-cost mobile phone reminders to improve attendance for HIV test results: a pilot study in rural Swaziland. Infect Dis Poverty 2013;2(1):12 [FREE Full text] [doi: 10.1186/2049-9957-2-12] [Medline: 23849508]
23. Chen S, Zhang Q, Chan CK, Yu FY, Chidgey A, Fang Y, et al. Evaluating an innovative HIV self-testing service with web-based, real-time counseling provided by an artificial intelligence Chatbot (HIVST-Chatbot) in increasing HIV self-testing use among Chinese men who have sex with men: protocol for a noninferiority randomized controlled trial. JMIR Res Protoc 2023;12:e48447 [FREE Full text] [doi: 10.2196/48447] [Medline: 37389935]
24. Kaponda CP, Jere DL, Chimango JL, Chimwaza AF, Crittenden KS, Kachingwe SI, et al. Impacts of a peer-group intervention on HIV-related knowledge, attitudes, and personal behaviors for urban hospital workers in Malawi. J Assoc Nurses AIDS Care 2009;20(3):230-242 [FREE Full text] [doi: 10.1016/j.jana.2008.12.005] [Medline: 19427600]
25. Marino P, Simoni JM, Silverstein LB. Peer support to promote medication adherence among people living with HIV/AIDS: the benefits to peers. Soc Work Health Care 2007;45(1):67-80 [FREE Full text] [doi: 10.1300/J010v45n01_05] [Medline: 17804348]
26. Mbeba MM, Kaponda CPN, Jere DL, Kachingwe SI, Crittenden KS, McCreary LL, et al. Peer group intervention reduces personal HIV risk for Malawian health workers. J Nurs Scholarsh 2011;43(1):72-81 [FREE Full text] [doi: 10.1111/j.1547-5069.2011.01384.x] [Medline: 21342427]
27. Ulm M, Widhalm P, Brandle N. Characterization of mobile phone localization errors with OpenCellID data. 2015 Presented at: 2015 4th International Conference on Advanced Logistics and Transport (ICALT); May 20-22, 2015; Valenciennes, France p. 100-104. [doi: 10.1109/icadlt.2015.7136601]
28. Hawthorne T, Kwan MP. Using GIS and perceived distance to understand the unequal geographies of healthcare in lower-income urban neighbourhoods. Geogr J 2012;178(1):18-30 [FREE Full text] [doi: 10.1111/j.1475-4959.2011.00411.x]
29. Susan A, Karita E, N'gandu N, Tichacek A. The evolution of voluntary testing and counseling as an HIV prevention strategy. In: Preventing HIV in Developing Countries. Boston, MA: Springer; 2002:87-108.
30. Dwyer-Lindgren L, Cork MA, Sligar A, Steuben KM, Wilson KF, Provost NR, et al. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. Nature 2019;570(7760):189-193 [FREE Full text] [doi: 10.1038/s41586-019-1200-9] [Medline: 31092927]
31. Colin Cameron A, Miller DL. A practitioner‘s guide to cluster-robust inference. J Hum Resour 2015;50(2):317-372. [doi: 10.3368/jhr.50.2.317]
32. Abadie A, Athey S, Imbens G, Wooldridge YC. When Should You Adjust Standard Errors for Clustering?. Cambridge, MA: National Bureau of Economic Research; 2017.
33. Klonner S, Nolen P. Does ICT benefit the poor? Evidence from South Africa. 2008. URL: https://pdfs.semanticscholar.org/0ce4/92f33b778dee001f29aff461a970a2bd4d3d.pdf [accessed 2019-12-04]
34. Dinkelman T, Beaman L, Beuermann D, Busso M, Case A, Choi H, et al. The effects of rural electrification on employment: new evidence from South Africa. Am Econ Rev 2011;101(7):3078-3108 [FREE Full text] [doi: 10.1257/aer.101.7.3078]
35. Monteiro C, Ramirez-Rosado I, Miranda V, Zorzano-Santamaria P, Garcia-Garrido E, Fernandez-Jimenez L. GIS spatial analysis applied to electric line routing optimization. IEEE Trans Power Deliv 2005;20(2):934-942. [doi: 10.1109/tpwrd.2004.839724]
36. Batzilis D, Dinkelman T, Oster E, Thornton R, Zanera D. New cellular networks in Malawi: correlates of service rollout and network performance. National Bureau of Economic Research. 2016. URL: https://www.nber.org/system/files/working_papers/w16616/w16616.pdf [accessed 2024-08-20]

XSL•FO
RenderX

37. Buys P, Dasgupta S, Thomas T, Wheeler D. Determinants of a digital divide in sub-Saharan Africa: a spatial econometric analysis of cell phone coverage. World Dev Pergamon 2009:1494-1505. [doi: 10.1016/j.worlddev.2009.01.011]
38. Pierskalla JH, Hollenbach FM. Technology and collective action: the effect of cell phone coverage on political violence in Africa. Am Polit Sci Rev 2013;107(2):207-224. [doi: 10.1017/s0003055413000075]
39. Berman N, Couttenier M, Rohner D, Thoenig M. This mine is mine! How minerals fuel conflicts in Africa. Am Econ Rev 2017;107(6):1564-1610. [doi: 10.1257/aer.20150774]
40. Mayala B, Fish T, Eitelberg D, Dontamsetti T. The DHS Program Geospatial Covariate Datasets Manual. 2018. URL: https://spatialdata.dhsprogram.com/references/DHS%20Covariates%20Extract%20Data%20Description%202.pdf [accessed 2024-08-15]
41. Bettinghaus EP. Health promotion and the knowledge-attitude-behavior continuum. Prev Med 1986;15(5):475-491. [doi: 10.1016/0091-7435(86)90025-3] [Medline: 3774779]
42. Mahmoud MM, Nasr AM, Gassmelseed DEA, Abdalelhafiz MA, Elsheikh MA, Adam I. Knowledge and attitude toward HIV voluntary counseling and testing services among pregnant women attending an antenatal clinic in Sudan. J Med Virol 2007;79(5):469-473. [doi: 10.1002/jmv.20850] [Medline: 17385672]
43. Rahbar T, Garg S, Tripathi R, Gupta V, Singh M. Knowledge, attitude, behavior and practice (KABP) regarding HIV/AIDS among pregnant women attending PPTCT programme in New Delhi. J Commun Dis 2007;39(3):179-184 [FREE Full text] [Medline: 18697582]
44. Broadhead R, Heckathorn D, Altice F, van Hulst Y, Carbone M, Friedland G, et al. Increasing drug users' adherence to HIV treatment: results of a peer-driven intervention feasibility study. Soc Sci Med 2002;55(2):235-246. [doi: 10.1016/s0277-9536(01)00167-8] [Medline: 12144138]
45. Haignere CS, Freudenberg N, Silver DR, Maslanka H, Kelley J. One method for assessing HIV/AIDS peer-education programs. J Adolesc Heal 1997;21(2):76-79 [FREE Full text] [doi: 10.1016/s1054-139x(97)00005-0]
46. Pearlman DN, Camberg L, Wallace LJ, Symons P, Finison L. Tapping youth as agents for change: evaluation of a peer leadership HIV/AIDS intervention. J Adolesc Health 2002;31(1):31-39. [doi: 10.1016/s1054-139x(02)00379-8] [Medline: 12090963]
47. Manacorda M, Tesei A. Liberation technology: mobile phones and political mobilization in Africa. Econometrica 2020;88(2):533-567. [doi: 10.3982/ecta14392]
48. World Development Report 2016: Digital Dividends. World Bank Group. 2016. URL: https://www.worldbank.org/en/publication/wdr2016 [accessed 2024-08-20]
49. Mohlabane N, Tutshana B, Peltzer K, Mwisongo A. Barriers and facilitators associated with HIV testing uptake in South African health facilities offering HIV counselling and testing. Health SA Gesondheid 2016;21:86-95. [doi: 10.1016/j.hsag.2015.11.001]
50. Ayieko J, Brown L, Anthierens S, van Rie A, Getahun M, Charlebois E, et al. "Hurdles on the path to 90-90-90 and beyond": qualitative analysis of barriers to engagement in HIV care among individuals in rural East Africa in the context of test-and-treat. PLoS One 2018;13(8):e0202990 [FREE Full text] [doi: 10.1371/journal.pone.0202990] [Medline: 30161172]
51. Chikwari CD, Dringus S, Ferrand RA. Barriers to, and emerging strategies for, HIV testing among adolescents in sub-Saharan Africa. Curr Opin HIV AIDS 2018;13(3):257-264. [doi: 10.1097/COH.0000000000000452] [Medline: 29401121]
52. Ling R, Horst H. Mobile communication in the global south. New Media Soc 2011;13(3):363-374. [doi: 10.1177/1461444810393899]
53. Asongu SA, Nwachukwu JC. The mobile phone in the diffusion of knowledge for institutional quality in sub-Saharan Africa. World Dev 2016;86:133-147. [doi: 10.1016/j.worlddev.2016.05.012]
54. Wang W, Alva S, Wang S. HIV-related knowledge and behaviors among people living with HIV in eight high HIV prevalence countries in sub-Saharan Africa. The DHS Program. 2012. URL: https://dhsprogram.com/publications/publication-as29-analytical-studies.cfm#:~:text=While%20condom%20use%20at%20last,for%20both%20women%20and%20men [accessed 2019-10-07]
55. Aker JC, Ksoll C, Lybbert TJ. Can mobile phones improve learning? Evidence from a field experiment in Niger. Am Econ J Appl Econ 2012;4(4):94-120. [doi: 10.1257/app.4.4.94]
56. Qureshi S. What is the role of mobile phones in bringing about growth? Inf Technol Dev 2013;19(1):1-4. [doi: 10.1080/02681102.2013.764597]
57. Jadhav A, Weis J. Mobile phone ownership, text messages, and contraceptive use: is there a digital revolution in family planning? Contraception 2020;101(2):97-105. [doi: 10.1016/j.contraception.2019.10.004] [Medline: 31782990]
58. Alami SO, Dulli L, Dal SL, Rastagar SH, Seddiqi S, Hemat S, et al. Getting health information to internally displaced youth in Afghanistan: can mobile phone technology bridge the gap? Gates Open Res F1000 Res Limited 2019;3:1483. [doi: 10.12688/gatesopenres.13008.1]
59. Nie J, Unger JA, Thompson S, Hofstee M, Gu J, Mercer MA. Does mobile phone ownership predict better utilization of maternal and newborn health services? A cross-sectional study in Timor-Leste. BMC Pregnancy Childbirth 2016;16(1):183 [FREE Full text] [doi: 10.1186/s12884-016-0981-1] [Medline: 27448798]

XSL•FO
RenderX

60. Venter W, Coleman J, Chan VL, Shubber Z, Phatsoane M, Gorgens M, et al. Improving linkage to HIV care through mobile phone apps: randomized controlled trial. JMIR Mhealth Uhealth 2018;6(7):e155 [FREE Full text] [doi: 10.2196/mhealth.8376] [Medline: 30021706]
61. Hampshire K, Porter G, Owusu SA, Mariwah S, Abane A, Robson E, et al. Informal m-health: how are young people using mobile phones to bridge healthcare gaps in sub-Saharan Africa? Soc Sci Med 2015;142:90-99 [FREE Full text] [doi: 10.1016/j.socscimed.2015.07.033] [Medline: 26298645]
62. Cilliers L, Viljoen KL, Chinyamurindi WT. A study on students' acceptance of mobile phone use to seek health information in South Africa. Health Inf Manag 2018;47(2):59-69. [doi: 10.1177/1833358317706185] [Medline: 28537211]
63. Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unützer J, Patel V, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. Lancet Psychiatry 2017;4(6):486-500 [FREE Full text] [doi: 10.1016/S2215-0366(17)30096-2] [Medline: 28433615]

Abbreviations

2SLS: 2-stage least squares
DCW: Digital Chart of the World
DHS: Demographic and Health Survey
GSM: Global System for Mobile Communication
IV: instrumental variable
STI: sexually transmittable infection

Edited by A Mavragani; submitted 07.05.23; peer-reviewed by S Chen, J Wang, A Gangadhara Rao; comments to author 15.03.24; revised version received 25.03.24; accepted 24.05.24; published 27.09.24.

Please cite as:
Iacoella F, Tirivayi N
Mobile Phones and HIV Testing: Multicountry Instrumental Variable Analysis From Sub-Saharan Africa
J Med Internet Res 2024;26:e48794
URL: https://www.jmir.org/2024/1/e48794
doi:10.2196/48794
PMID:

©Francesco Iacoella, Nyasha Tirivayi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of a Mobile App to Promote Social Participation on Older Adults: Randomized Controlled Trial

Kenjiro Kawaguchi[1], MD, MPH, PhD; Atsushi Nakagomi[1], MD, PhD; Kazushige Ide[1], PhD; Katsunori Kondo[1,2], MD, PhD

[1]Department of Social Preventive Medical Sciences, Center for Preventive Medical Sciences, Chiba University, Chiba, Japan

[2]Research Department, Institute for Health Economics and Policy, Tokyo, Japan

Corresponding Author:
Kenjiro Kawaguchi, MD, MPH, PhD
Department of Social Preventive Medical Sciences
Center for Preventive Medical Sciences
Chiba University
1-33 Yayoicho, Inage
Chiba, 263-8522
Japan
Phone: 81 90 7596 8742
Email: kawaguchikenjiro@chiba-u.jp

Abstract

Background: Social participation is crucial for healthy aging, improving physical and mental health, cognitive function, and quality of life among older adults. However, social participation tends to decline with age due to factors like loss of social networks and health issues. Mobile health apps show promise in promoting healthy behaviors among older adults, but their effectiveness in increasing social participation remains understudied.

Objective: This randomized controlled trial aimed to evaluate the efficacy of a mobile app called Encouragement of Social Participation (ESP, “Shakai Sanka no Susume;” Hitachi) in promoting social participation and physical activity among community-dwelling older adults.

Methods: The study recruited 181 community-dwelling adults aged 60 years or older from 2 municipalities in Japan and through a web-based research panel. Participants were randomly assigned to either the intervention group (n=87), which used the ESP app for 12 weeks, or the control group (n=94), which used only Google Fit. The ESP app incorporated features such as self-monitoring of social participation, personalized feedback, gamification elements, and educational content. Primary outcomes were changes in social participation frequency over the previous 2 months and changes in step counts, measured at baseline and week 12. Secondary outcomes included changes in specific types of social activities and subjective well-being. Data were analyzed using analysis of covariance and linear mixed-effects models.

Results: The intervention group showed a significantly greater increase in social participation frequency compared with the control group (adjusted difference 3.03; 95% CI 0.17-5.90; *P*=.04). Specifically, the intervention group demonstrated higher frequencies of participation in hobbies (adjusted difference: 0.82; 95% CI 0.01-1.63) and cultural clubs (adjusted difference 0.65; 95% CI 0.07-1.23) compared with the control group. However, there were no significant differences in weekly step counts between the groups. Subgroup analyses suggested potentially larger effects among participants who were older than 70 years, female, had lower educational attainment, and were recruited from community settings, although only females and the lower educational attainment subgroups demonstrated 95% CIs that did not encompass zero.

Conclusions: The ESP mobile app effectively promoted social participation among community-dwelling older adults, particularly in hobbies and cultural club activities. However, it did not significantly impact physical activity levels as measured by step counts. These findings suggest that mobile apps can be valuable tools for encouraging social engagement in older populations, potentially contributing to healthy aging. Future research should focus on optimizing app features to maintain long-term engagement and exploring strategies to enhance physical activity alongside social participation.

Trial Registration: University Medical Information Network Clinical Trial Registry UMIN000049045; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000055781

XSL•FO
RenderX

(J Med Internet Res 2024;26:e64196) doi:10.2196/64196

KEYWORDS

gerontology; geriatrics; older adults; elderly; older people; community dwelling older adult; aging; social participation; walking; mHealth; apps; smartphone; digital health; digital technology; digital interventions; physical activity; exercise

Introduction

Social participation, involvement in activities that facilitate interactions with others in the community [1], is a key determinant of healthy aging [2]. Social participation benefits older adults, improving their physical activities [3,4], physical and mental health [5-8], cognitive function [9,10], and overall quality of life [11]. Despite these well-established benefits, social participation declines with age owing to factors such as the loss of social networks [12] and health issues [13]. Therefore, interventions that can effectively promote and maintain social participation among older adults are needed.

There has been a growing interest in the potential of digital technologies, particularly mobile apps, to promote healthy aging [14,15]. A survey conducted by the Ministry of Internal Affairs and Communications of Japan revealed that more than 60% of Japanese adults aged 70 years and older own smartphones [16], indicating a rapidly rising trend in smartphone ownership among older adults in Japan. Research suggests that mobile health interventions can effectively promote physical activities [17,18], self-management of chronic conditions [19,20], and well-being [21] in this demographic. These findings suggest that smartphones and apps are accessible and scalable interventions that can reach a broader population of older adults and enhance their health and well-being.

The development of apps that integrate evidence-based strategies for behavior change is one promising approach to increasing social participation among older adults [22]. An observational study showed that a mobile app designed to encourage social participation among community-dwelling older adults was acceptable to them, with approximately 60% expressing a willingness to continue using the app [23]. A quasi-experimental study [24] examined the effects of a mobile app on the stimulation of social participation among older adults, showing that it increased the intention for social participation and frequency of social participation among community-dwelling older adults. According to a systematic review [25], incorporating features such as performance feedback and encouragement, self-monitoring capabilities, and educational content may enhance the effectiveness of health behavior promotion apps for older adults. The incorporation of behavior change techniques, such as self-monitoring and social comparison, also improves engagement and outcomes [26,27]. By designing apps that incorporate these techniques, we can create apps that facilitate more successful behavior change.

Despite the potential of apps to encourage social participation and promote healthy aging, limited research examines the effects of mobile health interventions on social participation [24]. This study introduces a novel app that integrates evidence-based features specifically designed to promote behavior change among older adults. In contrast to earlier apps [22-24], our app integrates evidence-based features, such as self-monitoring of social participation, personalized feedback and encouragement, gamification elements, and educational content regarding the benefits of social participation, all based on established behavior change theories. The primary objective of this study was to evaluate the efficacy of the app in promoting social participation and physical activity (as measured by step count) among community-dwelling older adults. We hypothesized that the use of the app would lead to a significant increase in the frequency of social participation compared with the control group. The secondary objective was to assess the effects of the app on specific types of social activity and subjective well-being.

Methods

Study Design

We conducted a randomized controlled trial (allocation ratio of 1:1) to determine the efficacy of an app that promoted social participation among community-dwelling older adults. In the intervention group, we provided participants with an investigational app for 12 weeks.

Participants

The participants were community-dwelling older adults from 2 different sources. First, we invited older adults residing in 2 municipalities to participate in this study. The first municipality is a small town with a population of 10,000, while the second is a city with a population of about 500,000 in two prefectures near Tokyo. We distributed information about the study through posters, flyers, and municipal public relations magazines with the cooperation of the staff at the municipal office's Department of Senior Services. We held briefing sessions for those interested in participating and explained the research content in detail. After explaining the research objectives comprehensively, we selected participants who met the inclusion criteria and expressed a willingness to participate. Second, we recruited a web-based research panel from the Internet survey research company ASMARQ Co, Ltd. The registrants for the web-based research panel who met the inclusion criteria received an email invitation to participate in this study. The company continued sending email invitations until 100 participants were obtained.

The inclusion criteria were as follows: (1) community-dwelling older adults aged 60 years or older, who (2) used a smartphone daily, (3) spoke Japanese, (4) usually used LINE (LY Corporation; one of the most popular apps for social communication in Japan [16]), and (5) were not certified as requiring support or nursing care under long-term care insurance in Japan. We excluded participants who used a senior smartphone (featuring simplified menus, enlarged text, and icons) from this study because of the inappropriate functioning of the social participation app.

XSL•FO
RenderX

Procedure

We used the messenger app LINE to communicate with the participants in the intervention and control groups. In Japan, more than 90% of individuals use LINE as their primary social media platform [16]. Users can create their accounts using their phone numbers or email addresses and connect with friends through LINE ID. The participants were asked to connect their LINE ID with our research account. We sent a welcome message and research description to the participants' accounts and obtained their informed consent through checkboxes on the app. The participants could send messages to our research team through LINE to ask questions or withdraw their consent. To maintain their motivation and commitment to our study, we implemented a stamp card system to track their progress (Figure 1). The participants earned 1 stamp per week, for a total of 12 stamps over 12 weeks. We sent weekly reminder messages to prompt participants to press the stamp tile and use the app.

We instructed the participants in the intervention group to install and use the Encouragement of Social Participation (ESP) app on a weekly basis, in addition to installing Google Fit to obtain step count data. By contrast, the control group only installed Google Fit and occasionally used it. We asked the participants to fill in the app-based questionnaires at the beginning and end of the study to obtain baseline information and assess outcomes. We sent invitations to complete the questionnaires to the participants in both groups through LINE at the beginning and after 85 days of the study.

Figure 1. Example of the LINE app interface customized for this study, including the main screen, stamp card, and questionnaire.

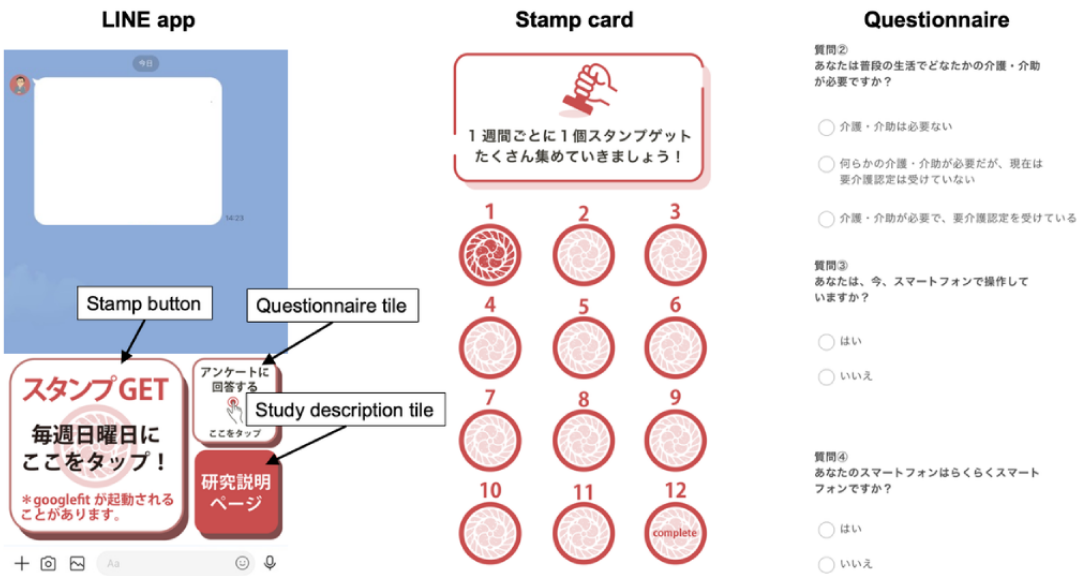

Randomization

After providing informed consent, participants were promptly assigned to either the intervention or control group based on a random number generated by a computer-based randomization system at the backend of the LINE app. The study investigators were blinded to the group allocation. This random allocation process ensured that the group assignments were not predictable in real time.

Intervention: Encouragement of a Social Participation Smartphone App

The ESP app ("Shakai Sanka no Susume") was developed by Hitachi. This free app is compatible with both the Android (Google) and iOS (Apple Inc) operating systems. It incorporates several key features designed to promote healthy behaviors (Figure 2). It facilitates self-monitoring of social participation through features such as feedback and encouragement (based on social participation levels) and educational content that emphasizes the benefits of social participation. Furthermore, the app uses smartphone data such as location information, step counts, and other relevant data to measure outdoor activities. It tracks the user's movements using GPS and displays the walking distance, step counts, walked route, and number of locations visited on a map along with the step counts. This enables self-monitoring of social participation and physical activity. The app uses a specific classification algorithm based on the user's average step count, number of places visited, and number of days spent outside. This algorithm categorizes the user's social participation level into 4 categories, that are "Expert," "Master," "Skilled," and "Beginner." This gamification element aims to increase user motivation and engagement. Furthermore, the ESP app delivers weekly digestible columns to users. These columns inform users about the importance and value of social participation based on previous research findings and serve as an educational component to raise awareness and reinforce the importance of staying socially active. In the preliminary testing phase, the pilot app, featuring only GPS tracking and column delivery, was evaluated among a small group of older adults in a rural area, indicating an increase in the distance traveled. We were not involved in the development of this app.

XSL•FO
RenderX

Figure 2. Example of the Encouragement of Social Participation smartphone app interface, including the step counts, walking distance, walked route, number of days out, ranking report, and educational columns.

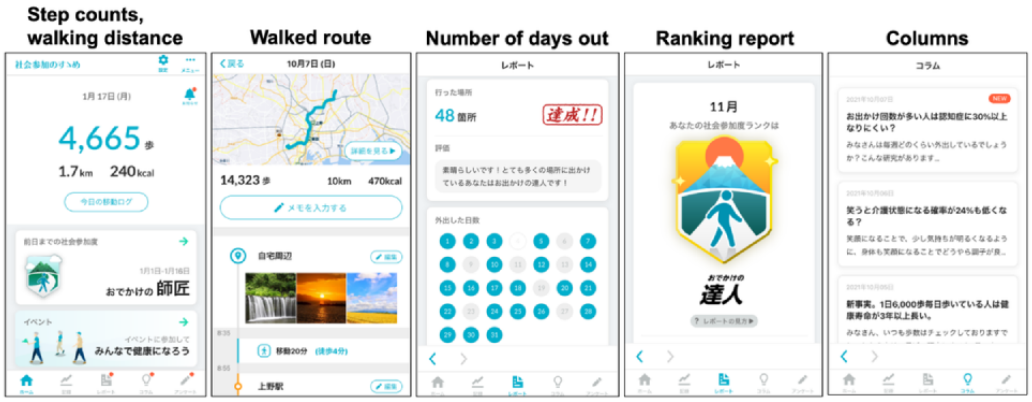

Control: Google Fit App

The control group used the Google Fit app, a widely available free health and fitness tracking app developed by Google for Android and iOS. The app allows users to monitor their physical activities, such as walking, running, and cycling and track various metrics, such as distance, time, and calories burned. We instructed the participants in the control group to occasionally use Google Fit to track their physical activity levels and monitor their progress. We did not inform the control participants of the details of the ESP app, including the product name.

Outcome Measures

The study's primary outcomes were (1) changes in social participation frequency over the previous 2 months and (2) changes in step counts, both measured at baseline and week 12. Participants in both groups completed app-based baseline and follow-up questionnaires to assess their frequency of social participation. The survey asked, "Over the last two months, how many times did you participate in the following groups or activities?" The participants reported their participation frequency in 6 types of group activities, such as volunteering, sports, hobbies, cultural clubs, salons, and others. We selected these activities based on the Long-term Care Needs Survey template provided by the Ministry of Health, Labor and Welfare. We summed the frequency in the 6 activities for each participant, and the difference during the 12 weeks represented the change in the frequency of social participation over the previous 2 months.

In addition, we used the participants' step counts as another primary outcome to evaluate the effect of the ESP app on walking. Step counts were measured using Google Fit for both the intervention and control groups, regardless of the smartphone operating system. We calculated the weekly average of daily step counts for each participant by dividing the total number of steps recorded in a given week by the number of days with valid step count data in that week as tracked by Google Fit. We excluded daily step counts with fewer than 100 steps [28].

The secondary outcomes included (1) changes in the social participation frequency for each activity (volunteering, sports, hobbies, cultural clubs, salons, and other activities) over the previous 2 months; (2) changes in the frequency of going out weekly over the previous month; and (3) changes in subjective well-being, including happiness, life satisfaction, and purpose in life (rated on a scale of 0-10, with higher scores indicating greater well-being).

In the follow-up questionnaires, we asked the participants how often they used the ESP app to measure their intensity of use.

Sample Size

We based the sample size on a previous study [24]; we assumed that social participation was 19.5 times per 2 months in the intervention group and 15 times per 2 months in the control group, with an SD of 14 and a dropout rate of 12.5%. Therefore, we required 350 participants for a 2-tailed α level of 5% and a statistical power of 80%.

Statistical Analysis

We described the baseline characteristics of the study participants, including sex, age, educational attainment, marriage, instrumental activities of daily living (IADLs), self-reported health, social participation frequency, frequency of going out weekly, and subjective well-being. All continuous values were expressed as mean (SD), and categorical variables were reported as numbers (percentages). We assessed IADLs using the 5-item Tokyo Metropolitan Institute of Gerontology Index of Competence [29], which includes questions on public transportation, shopping, meal preparation, bill payment, and banking (yes or no). Responses (0-5) were summed, with higher scores indicating greater independence. We considered the mean and SD for continuous variables and the number and proportion for categorical variables. We conducted all analyses on an intention-to-treat basis.

The primary analysis used analysis of covariance (ANCOVA) to compare the change in social participation frequency between the intervention and control groups, considering the effects of baseline variables, including sex, age, IADLs, and social participation frequency as covariates [30]. This study reported the adjusted mean difference in the change in social participation frequency between the 2 groups, along with a 95% CI. In addition, we compared weekly changes in step counts from weeks 1 to 12 between groups, using covariate-adjusted linear mixed models [31]. We considered the average weekly step counts to be nested within individuals with random intercepts. This approach enabled the analysis of longitudinal data with repeated measurements while accounting for the correlation

XSL•FO
RenderX

between observations from the same individual. The model included fixed effects for the group (intervention or control), time (weeks), an interaction between group and time, and covariates of sex, age, and IADLs. The interaction term indicates the difference in the weekly steps between the groups in a specific week. The normal distribution of each primary outcome variable was confirmed using the Shapiro-Wilk test.

We compared the secondary outcomes between the groups using ANCOVA adjusted for baseline variables, including sex, age, IADLs, and outcomes. In addition, we conducted subgroup analyses, stratified by sex, age (dichotomized as ≤70 or >70 years based on the median age of the sample), educational attainment (below high school, high school, or above high school), and the recruitment source (community or web-based research panel), to investigate potential differences in the app's effects among several demographic subgroups [32]. We calculated the *P* values for the interaction of the group with each baseline variable using the primary ANCOVA model.

Further, we conducted an ad hoc analysis to compare changes in weekly step counts from week 1 to each subsequent week (weeks 2-12) between groups. This analysis aimed to investigate the short-term effects of the ESP app on physical activity levels, which may not have been captured in the primary analysis. We performed an ad hoc analysis using the same linear mixed model approach as in the primary analysis.

All tests were 2-sided; we considered significance at a *P* value of less than .05. We used R version 4.2.1 (R Foundation for Statistical Computing).

Ethical Considerations

The ethics committees of the Japan Agency for Gerontological Evaluation Study (approval number 2022-01) and Chiba University (approval number M10635) reviewed and approved the study protocol. All participants provided informed consent to participate in this study through checkboxes on LINE app. No personally identifiable information such as name or address was collected. Participants were offered an electronic gift certificate for JPY ¥1500 (US $10.38) as an honorarium.

Results

Participant Characteristics

Figure 3 shows a flowchart of the participant selection between October 2022 and August 2023; 181 participants were recruited (81 from the community and 100 from internet survey registrants). In October 2023, the app's developer introduced an update that removed the feature providing digestible columns on social participation. To preserve the consistency and integrity of the intervention, we decided not to recruit additional participants, even before reaching the preplanned sample size of 350. A post hoc power analysis revealed that our sample size provided a power of 63.5% to detect the observed effects with an α level of 0.05. Of the 181 participants, we assigned 87 and 94 to the intervention and control groups, respectively. In addition, 2 participants in the intervention group withdrew their consent because of health concerns, leaving 179. All the participants completed the baseline questionnaires, whereas 80 and 85 in the intervention and control groups, respectively, completed the follow-up questionnaires, resulting in a follow-up rate of 92.2% (165/179). The baseline characteristics of the participants in the intervention and control groups were balanced (Table 1). The mean age of the study participants was approximately 70 years, 45% (80/179) were male, 70% (126/179) were educated beyond high school, 80% (143/179) were married (their mean IADL score was 4.6), and 90% (159/179) reported good health. The mean overall frequency of participation in any form of social activity was 12.28. Sports had the highest mean frequency of participation in specific activities (3.99), followed by volunteering (2.21), hobbies (21.98), salons (1.85), other activities (1.23), and cultural clubs (1.02). The mean study period was 86.9 days. Regarding the use of the ESP app, 25% (20/80) of participants in the intervention group reported minimal or no usage, 6% (5/80) reported using it several times a month, and 69% (55/80) used it at least once per week (Table 2).

Of the 179 participants, 155 (86.6%) recorded their daily step count at least once during the 12 weeks. On average, each participant recorded valid step counts (ie, daily step counts ≥100) for 40 days. The average daily step count was 5474 (SD 4743). Between weeks 3 and 7, participants in the intervention group took more steps daily than those in the control group (differences in step counts between the 2 groups ranged from 330 to 934); however, after week 8, almost no difference was observed, or slightly more steps were observed in the control group (Figure 4 and Table 3).

XSL•FO
RenderX

Figure 3. Flowchart of participant selection for analysis. ESP: Encouragement of Social Participation.

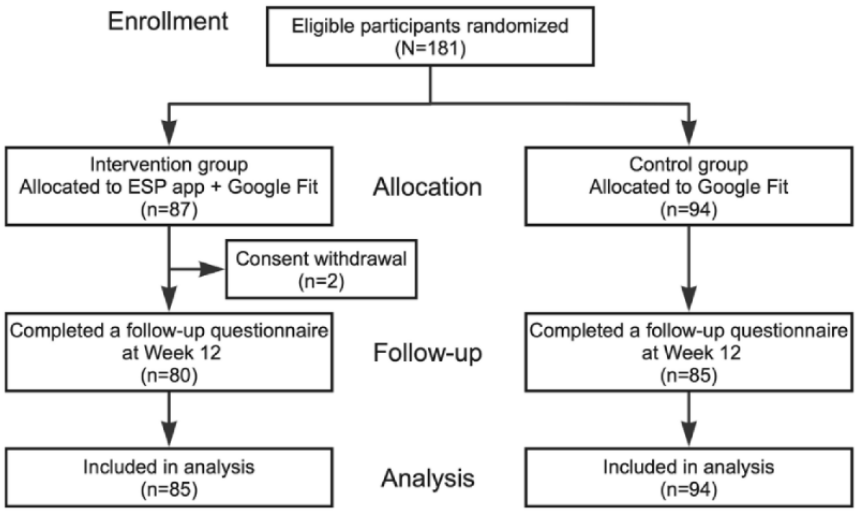

Table 1. Baseline characteristics of the study participants.

Characteristics	Total (N=179)	Intervention (n=85)	Control (n=94)
Age (years), mean (SD)	70 (6.3)	69.9 (6)	70.1 (6.7)
Male, n (%)	80 (44.7)	43 (51)	37 (39)
Educational attainment, n (%)			
High school or less	53 (29.6)	20 (26)	33 (35)
Above high school	126 (70.4)	65 (77)	61 (65)
Married, n (%)	143 (79.9)	71 (84)	72 (77)
IADL[a] (0-5, higher scores indicating more independence), mean (SD)	4.6 (0.7)	4.7 (0.7)	4.7 (0.7)
Self-reported health, n (%)			
Poor	5 (2.8)	4 (5)	1 (1)
Fair	15 (8.4)	5 (56)	10 (11)
Good	137 (76.5)	67 (79)	70 (74)
Very good	22 (12.3)	9 (11)	13 (14)
Social participation frequency over the previous 2 months, mean (SD)			
Any	12.28 (13.06)	11.16 (10.50)	13.30 (14.99)
Volunteering	2.21 (4.40)	1.61 (3.19)	2.76 (5.22)
Sports	3.99 (5.82)	4.02 (5.63)	3.96 (6.01)
Hobbies	1.98 (2.95)	1.78 (2.71)	2.17 (3.15)
Cultural clubs	1.02 (2.06)	0.79 (1.54)	1.23 (2.42)
Salons	1.85 (3.49)	1.79 (2.95)	1.90 (3.93)
Other activities	1.23 (3.06)	1.18 (3.61)	1.28 (2.49)
Frequency of going out weekly over the previous month, mean (SD)	6.41 (4.10)	6.07 (3.49)	6.72 (4.58)
Subjective well-being, mean (SD)			
Happiness	7.52 (1.75)	7.54 (1.80)	7.50 (1.71)
Life satisfaction	7.44 (1.82)	7.45 (1.80)	7.44 (1.85)
Purpose in life	7.40 (1.88)	7.34 (1.80)	7.45 (1.95)

[a]IADL: instrumental activities of daily living.

XSL•FO
RenderX

Table 2. Frequency of use of the Encouragement of Social Participation app in the intervention group.

Frequency	Intervention (n=80), n (%)
Not used	8 (10)
Rarely used	12 (15)
Several times a month	5 (6)
Once a week	10 (13)
2-3 days a week	14 (18)
4 times or more a week	31 (39)

Figure 4. Average weekly step counts (error bars denote the SE of the mean).

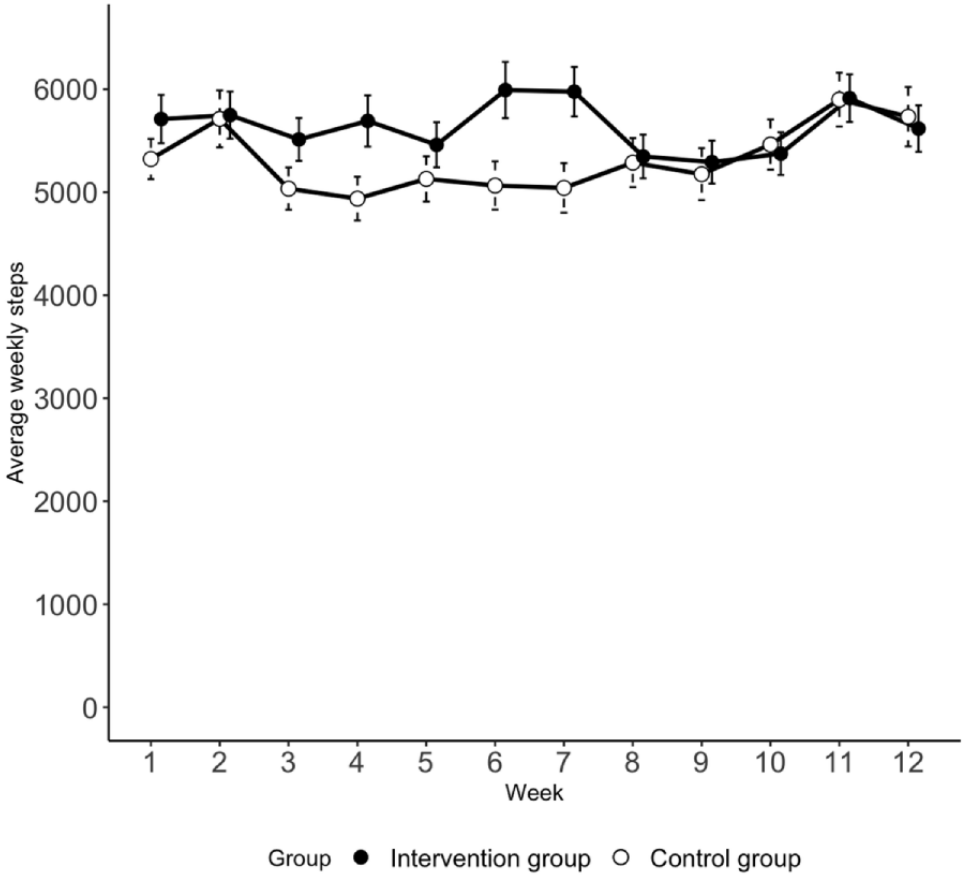

Table 3. Average weekly step counts.

Group	Week											
	1	2	3	4	5	6	7	8	9	10	11	12
Intervention, mean (SE)	5710 (234)	5748 (229)	5512 (209)	5692 (249)	5460 (219)	5992 (272)	5976 (241)	5347 (212)	5292 (209)	5375 (207)	5913 (231)	5618 (226)
Control, mean (SE)	5322 (195)	5712 (276)	5034 (205)	4938 (212)	5129 (219)	5065 (234)	5042 (241)	5287 (238)	5175 (252)	5463 (243)	5900 (262)	5733 (288)
Group × Week[a]	—[b]	–565 (.15)	63 (.87)	74 (.85)	–333 (.41)	245 (.54)	396 (.33)	–220 (.59)	–208 (.61)	–330 (.42)	–266 (.52)	–340 (.42)

[a]Data are presented as values of interaction terms between groups and weeks (*P* value).

[b]Not applicable.

XSL•FO
RenderX

Primary Outcomes

Social Participation

The frequency of social participation over the 2 months changed by +1.69 in the intervention group and –2.45 in the control group. The intervention group had a significantly greater increase in social participation frequency compared with the control group, with an adjusted difference of 3.03 (95% CI 0.17-5.90, *P*=.04).

Step Counts

The results of the linear mixed-effects model demonstrated that the interaction between the group and week 12 was insignificant (*P*=.53), indicating no significant difference in weekly steps between the groups.

Secondary Outcomes

Table 4 presents the secondary outcomes. The intervention group showed higher mean changes compared with the control group for several activities, with notable differences observed for hobbies and cultural clubs (95% CIs not including zero). For other activities, including volunteering, sports, salons, and going out, the intervention group showed trends toward increased participation, but the 95% CIs included zero. Regarding subjective well-being measures (happiness, life satisfaction, and purpose in life), the intervention group showed small increases compared to the control group, though the 95% CIs for these measures also included zero.

Table 4. Effects of promotion of social participation by the Encouragement of Social Participation app.

Outcome	Mean changes during 12 weeks		
	Intervention	Controls	Adjusted difference[a] (95% CI)
Social participation			
Volunteering	0.15	–0.66	0.29 (–0.51 to 1.09)
Sports	–0.18	–1.11	0.89 (–0.43 to 2.22)
Hobbies	0.75	–0.26	0.82 (0.01 to 1.63)
Cultural clubs	0.68	–0.28	0.65 (0.07 to 1.23)
Salons	0.05	0.00	0.20 (–0.80 to 1.20)
Other activities	0.24	–0.14	0.36 (–0.41 to 1.13)
Going out	0.71	–0.65	1.11 (–0.03 to 2.25)
Subjective well-being			
Happiness	0.04	0.00	0.09 (–0.28 to 0.46)
Life satisfaction	–0.01	–0.08	0.12 (–0.30 to 0.54)
Purpose in life	0.18	0.00	0.22 (–0.25 to 0.70)

[a]Adjusted for sex, age, instrumental activities of daily living, and baseline outcome variables using analysis of covariance.

Subgroup Analysis

We conducted subgroup analyses to investigate whether the effects of the ESP app on promoting social participation varied across participant characteristics. Subgroups were defined based on age (≤70 and >70 years), gender, education level (high school or less and above high school), and recruitment source (community and web-based research panel).

Table 5 presents the adjusted differences and 95% CIs for each subgroup, along with *P* values for interactions. While some subgroups showed trends toward increased social participation in the intervention group, most 95% CIs included zero. Notable exceptions were the female subgroup and the high school or less education subgroup, where the 95% CIs did not include zero.

XSL•FO
RenderX

Table 5. Effects of the Encouragement of Social Participation app on social participation by subgroups.

Subgroup	Mean changes during 12 weeks			
	Intervention	Controls	Adjusted difference[a] (95% CI)	*P* value for interaction
Age (years)				.24
≤70	1.42	–1.48	1.53 (–1.33 to 4.41)	
>70	2.00	–3.59	4.68 (–0.68 to 10.04)	
Sex				.67
Male	0.12	–4.00	2.34 (–2.57 to 7.25)	
Female	3.33	–1.33	3.62 (0.15 to 7.43)	
Educational attainment				.38
High school or less	2.79	–0.93	3.59 (0.39 to 6.79)	
Above high school	–1.59	–5.38	–0.23 (–7.30 to 7.01)	
Recruitment source				.35
Community	1.49	–4.29	4.50 (–0.12 to 9.13)	
Web-based research panel	1.88	–0.65	2.01 (–1.21 to 5.24)	

[a]Adjusted for sex, age, instrumental activities in daily living, and the baseline outcome variable by an analysis of covariance.

Ad Hoc Analysis: Short-Term Effects on Physical Activity

We conducted an ad hoc analysis to investigate the short-term effects of the ESP app on physical activity levels. We compared the weekly step counts from weeks 1 to 8 between the groups using the same linear mixed model approach as in the primary analysis, with fixed effects for group, week, and their interaction, and random intercepts for individuals. Table 3 shows that some of the interaction terms between the group and weeks (weeks 2-8) yielded positive results whereas no interaction terms did not reach statistical significance at *P* value of <.05. This indicates that the differences in step counts might not differ clearly between the intervention and control groups in any week.

Discussion

Principal Findings

This randomized controlled trial evaluated the efficacy of the ESP app in promoting social participation and physical activity among community-dwelling older adults. The results indicated that participants who used the ESP app during the 12-week study significantly increased the frequency of social participation compared to the control group. Furthermore, the intervention group showed higher frequencies of participation in hobbies and cultural clubs than the control group. Conversely, the intervention group did not show an increase in weekly step counts.

Several key features in the app may explain the positive effect of the ESP app on social participation. First, the app's self-monitoring capabilities enable users to easily track and visualize their social participation levels, providing feedback that may increase their awareness and motivation to engage in more social activities. Previous interventional research reported that the goal-setting and self-monitoring features of a tablet app aimed at improving balance and strength helped motivate two-thirds of older participants [33]. The inclusion of educational content on the benefits of social participation or other health topics may also have helped to increase knowledge, motivation, and skills among this demographic, as several studies have demonstrated [34,35]. Finally, a gamification element that displays achievement levels of social participation may enhance engagement with the app.

The observed step counts were higher for ESP app users than for the control group from week 3 to 7, suggesting that interventions aimed at promoting social participation can have spillover effects on physical activity [3,36]. However, these effects did not lead to continued increases in step count over the 12-week period. These temporary effects may be attributed to the decrease in novelty of the app wearing off over time [37,38]. When participants first started using the app, the excitement of trying the new technology and the motivation to engage in social activities may have led to an initial increase in physical activity levels. However, the participants may have become less interested in exploring the app's features or less responsive to the feedback and encouragement provided by the app as the study progressed. Furthermore, the app's gamification elements (achievement levels and rewards) may have diminished as users became accustomed to the mechanism [39]. Although extrinsic rewards can promote engagement, relying on these alone may not be effective in the long term.

Future development of mobile health apps like the ESP app should consider a multifaceted approach to maintain long-term user engagement and promote lasting behavior changes. Research suggests that cultivating intrinsic motivation maintains a user's interest and involvement over an extended period [40]. According to self-determination theory, intrinsic motivation can be fostered by ensuring the fulfillment of 3 basic psychological needs, that are autonomy, competence, and relatedness [41]. Autonomy, the need for self-control over behavior and goals [41], can be supported by allowing users to set personal goals and choose methods, with customizable

XSL•FO
RenderX

features [42]. Competence, the need to feel effective and capable [41], can be enhanced by offering challenges suited to the user's skill level and constructive feedback on progress, similar to gamification elements. Relatedness, the need for social connection [41], can be promoted through social features that enable progress sharing, mutual support, and community engagement [43]. Addressing these needs may foster intrinsic motivation and maintaining long-term engagement in app users, leading to more effective interventions. In addition, developing a hybrid model that combines digital tools with human interaction could enhance intervention effectiveness and user engagement, as suggested by human-computer interaction literature [44]. Furthermore, incorporating artificial intelligence–powered support features, such as chatbots or AI-assisted personal guides, could provide scalable human-like interaction, potentially increasing user engagement as shown in recent studies [45]. Combining these elements, such as intrinsic motivation cultivation, human support integration, and artificial intelligence–powered features, could promote continued engagement with mobile health apps aiming at behavior change.

We did not observe the differences in the step counts between the intervention and control groups. The study's sample size was likely not sufficiently large to detect a significant difference in step counts between the groups [46], given the considerable variability in daily step counts among the participants (SD 4743). It is also possible that the use of Google Fit as a step counter in both the intervention and control groups contributed to the lack of significant differences. Google Fit is a widely used and well-established app that provides users with step-tracking and other physical activity-monitoring features. Its presence in the control group may have inadvertently led to an increased awareness and motivation to engage in physical activity among the control participants [47], thereby reducing the potential differences between the 2 groups.

Notably, the effect sizes were larger among participants who were 70 years or older, female, had lower educational attainment, and recruited from community settings. This suggests that these subgroups may benefit more from the ESP app. According to a previous study that examined the usage patterns of apps and wearables to support physical activity in a sample of Japanese-speaking individuals [48], the app users were more likely to be younger, male, and have a higher level of education. The participants in these subgroups in the current study may have had less previous experience with physical activity apps before joining the study. Therefore, the ESP app may have had a greater effect on social participation among the older, female, and lower educational subgroups. In addition, the larger effect sizes observed among female participants may be due to gender-based variations in preferences for app designs and features. A previous study that investigated design preferences for an electronic mental health program, aimed at preventing depression [49], revealed that women preferred user-responsive and guided formats. Conversely, men preferred video game–type designs. These findings suggest that tailoring the design, features, and content of an ESP app to the preferences and needs of specific subgroups can enhance its impact on social participation.

Limitations

This study had several limitations. First, the ESP app was updated during this study and the company removed the feature that delivered digestible columns for social participation, which may have led to smaller effects of the app on social participation. In addition, due to the update, we halted the recruitment of study participants before we reached the planned sample size, which may have reduced the statistical power to detect significant effects. Second, the participants were selected from the local community and through a web-based research company. The use of the research panel may have excluded older adults who are less comfortable with information and communication technology or have lower digital literacy. These specific recruitment sources and methods may have resulted in a sample that was not representative of the diverse Japanese older adult population. This recruitment process may limit the generalizability of the findings to a broader older adult population. Third, this study included a 12-week follow-up period, which did not allow the assessment of long-term effects. Further research is needed to evaluate the long-term sustainability of these effects and explore strategies for optimizing engagement with apps over time. Fourth, a relatively low adherence to step-count tracking was observed. On average, each participant recorded valid step counts (ie, daily step counts ≥100) for only 40 days out of the 84-day study period. The incomplete data may have reduced the accuracy and representativeness of the step count findings. Therefore, the step count results should be interpreted with caution. Fifth, the study relied on self-reported measures of social participation, which may have been subject to recall bias. Further studies should incorporate objective measures of social activity, such as GPS tracking or check-in systems, to complement self-reported data. Finally, a quarter of the participants in the intervention group reported minimal or no usage of the app. This relatively low level of engagement may have led to an underestimation of the app's effect. However, it may also reflect real-world conditions, where not all users fully engage with health interventions.

Conclusion

The findings of this study suggest that the ESP app promotes social participation among community-dwelling older adults. The intervention group showed a significant increase in the self-reported frequency of social participation compared with the control group, indicating the app's efficacy. This increase in social participation highlights the app's potential as an affordable and expandable method for promoting social participation and healthy aging. Thus, the ESP can promote social participation among older adults and may contribute to healthy aging. Further research and modifications are warranted to optimize its effectiveness and maintain its effects.

XSL•FO
RenderX

Acknowledgments

We thank all study participants and staff members at the municipal office's Department of Senior Services in the 2 cities. We also express our gratitude to Drs Yoshinori Fujiwara, Satoshi Seino, Yuri Yokoyama, Daisuke Kumazawa, and Motoki Tamura for their support in recruiting participants, and to Medical Logue, Inc (Tokyo, Japan) for developing the LINE research platform. We would also like to thank Editage (Editage website) for the English language editing. This research was supported by JST RISTEX Japan (grant JPMJRS22B1).

Conflicts of Interest

K Kondo received research funding from Hitachi (Tokyo, Japan). This funding aimed to work on a project to promote health and well-being using smartphone data. The funding contract did not impose any limitations on the study design, analysis, interpretation, or writing of the article, or on its subsequent publication.

Multimedia Appendix 1
CONSORT checklist.
[PDF File (Adobe PDF File), 386 KB - jmir_v26i1e64196_app1.pdf]

References

1. Levasseur M, Lussier-Therrien M, Biron ML, Raymond É, Castonguay J, Naud D, et al. Scoping study of definitions of social participation: update and co-construction of an interdisciplinary consensual definition. Age Ageing 2022;51(2):afab215 [FREE Full text] [doi: 10.1093/ageing/afab215] [Medline: 35134843]
2. Levasseur M, Richard L, Gauvin L, Raymond E. Inventory and analysis of definitions of social participation found in the aging literature: proposed taxonomy of social activities. Soc Sci Med 2010;71(12):2141-2149 [FREE Full text] [doi: 10.1016/j.socscimed.2010.09.041] [Medline: 21044812]
3. Ihara S, Ide K, Kanamori S, Tsuji T, Kondo K, Iizuka G. Social participation and change in walking time among older adults: a 3-year longitudinal study from the JAGES. BMC Geriatr 2022;22(1):238 [FREE Full text] [doi: 10.1186/s12877-022-02874-2] [Medline: 35317737]
4. Kikuchi H, Inoue S, Fukushima N, Takamiya T, Odagiri Y, Ohya Y, et al. Social participation among older adults not engaged in full- or part-time work is associated with more physical activity and less sedentary time. Geriatr Gerontol Int 2017;17(11):1921-1927. [doi: 10.1111/ggi.12995] [Medline: 28230301]
5. Douglas H, Georgiou A, Westbrook J. Social participation as an indicator of successful aging: an overview of concepts and their associations with health. Aust Health Rev 2017;41(4):455-462. [doi: 10.1071/AH16038] [Medline: 27712611]
6. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. PLoS Med 2010;7(7):e1000316 [FREE Full text] [doi: 10.1371/journal.pmed.1000316] [Medline: 20668659]
7. Stuck AE, Walthert JM, Nikolaus T, Büla CJ, Hohmann C, Beck JC. Risk factors for functional status decline in community-living elderly people: a systematic literature review. Soc Sci Med 1999;48(4):445-469. [doi: 10.1016/s0277-9536(98)00370-0] [Medline: 10075171]
8. Webber M, Fendt-Newlin M. A review of social participation interventions for people with mental health problems. Soc Psychiatry Psychiatr Epidemiol 2017;52(4):369-380 [FREE Full text] [doi: 10.1007/s00127-017-1372-2] [Medline: 28286914]
9. Bourassa KJ, Memel M, Woolverton C, Sbarra DA. Social participation predicts cognitive functioning in aging adults over time: comparisons with physical health, depression, and physical activity. Aging Ment Health 2017;21(2):133-146. [doi: 10.1080/13607863.2015.1081152] [Medline: 26327492]
10. Kelly ME, Duff H, Kelly S, McHugh Power JE, Brennan S, Lawlor BA, et al. The impact of social activities, social networks, social support and social relationships on the cognitive functioning of healthy older adults: a systematic review. Syst Rev 2017;6(1):259 [FREE Full text] [doi: 10.1186/s13643-017-0632-2] [Medline: 29258596]
11. Lestari SK, de Luna X, Eriksson M, Malmberg G, Ng N. A longitudinal study on social support, social participation, and older Europeans' quality of life. SSM Popul Health 2021;13:100747 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100747] [Medline: 33644292]
12. Ajrouch KJ, Blandon AY, Antonucci TC. Social networks among men and women: the effects of age and socioeconomic status. J Gerontol B Psychol Sci Soc Sci 2005;60(6):S311-S317. [doi: 10.1093/geronb/60.6.s311] [Medline: 16260713]
13. Zhang C, Zhao Y, Chen X, Li X, Liu Q, Peng R, et al. Trajectories of social participation and its predictors in older adults: based on the CLHLS cohorts from 2002 to 2018. Int J Environ Res Public Health 2023;20(5):4260 [FREE Full text] [doi: 10.3390/ijerph20054260] [Medline: 36901270]
14. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023;15(23):16503 [FREE Full text] [doi: 10.3390/su152316503]

XSL•FO
RenderX

15. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023;25:e43542 [FREE Full text] [doi: 10.2196/43542] [Medline: 36951896]
16. Ministry of Internal Affairs and Communications of Japan. Reiwa yo-nendo Joho Tsushin Media no Riyo Jikan to Joho Kodo ni Kansuru Chosa [A Report on the Usage time and Activities Regarding Telecommunications Media in the Fiscal Year 2023]. 2023. URL: https://www.soumu.go.jp/main_content/000887660.pdf [accessed 2024-09-20]
17. McGarrigle L, Todd C. Promotion of physical activity in older people using mHealth and eHealth technologies: rapid review of reviews. J Med Internet Res 2020;22(12):e22201 [FREE Full text] [doi: 10.2196/22201] [Medline: 33372894]
18. Sohaib Aslam A, van Luenen S, Aslam S, van Bodegom D, Chavannes NH. A systematic review on the use of mHealth to increase physical activity in older people. Clinical eHealth 2020;3:31-39. [doi: 10.1016/j.ceh.2020.04.002]
19. Doyle J, Murphy E, Gavin S, Pascale A, Deparis S, Tommasi P, et al. A digital platform to support self-management of multiple chronic conditions (ProACT): findings in relation to engagement during a one-year proof-of-concept trial. J Med Internet Res 2021;23(12):e22672 [FREE Full text] [doi: 10.2196/22672] [Medline: 34914612]
20. Sheng Y, Doyle J, Bond R, Jaiswal R, Gavin S, Dinsmore J. Home-based digital health technologies for older adults to self-manage multiple chronic conditions: a data-informed analysis of user engagement from a longitudinal trial. Digit Health 2022;8:20552076221125957 [FREE Full text] [doi: 10.1177/20552076221125957] [Medline: 36171962]
21. Ren P, Klausen SH. Smartphone use, intergenerational support and older adults' wellbeing. Curr Psychol 2023;43(1):407-424 [FREE Full text] [doi: 10.1007/s12144-023-04309-6]
22. Choi HK, Lee SH. Trends and effectiveness of ICT interventions for the elderly to reduce loneliness: a systematic review. Healthcare (Basel) 2021;9(3):293 [FREE Full text] [doi: 10.3390/healthcare9030293] [Medline: 33800099]
23. Jansen-Kosterink SM, Bergsma J, Francissen A, Naafs A. The first evaluation of a mobile application to encourage social participation for community-dwelling older adults. Health Technol 2020;10(5):1107-1113. [doi: 10.1007/s12553-020-00430-9]
24. Hsu PT, Ho CS, Ho YF, Chen JJ, Chen IJ. The effects of a social participation app on seniors. J Nurs Res 2021;29(5):e168. [doi: 10.1097/JNR.0000000000000451] [Medline: 34469399]
25. Liu N, Yin J, Tan SS, Ngiam KY, Teo HH. Mobile health applications for older adults: a systematic review of interface and persuasive feature design. J Am Med Inform Assoc 2021;28(11):2483-2501 [FREE Full text] [doi: 10.1093/jamia/ocab151] [Medline: 34472601]
26. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016;10(3):277-296 [FREE Full text] [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
27. Greaves CJ, Sheppard KE, Abraham C, Hardeman W, Roden M, Evans PH, IMAGE Study Group. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. BMC Public Health 2011;11:119 [FREE Full text] [doi: 10.1186/1471-2458-11-119] [Medline: 21333011]
28. Dang K, Ritvo P, Katz J, Gratzer D, Knyahnytska Y, Ortiz A, et al. The role of daily steps in the treatment of major depressive disorder: secondary analysis of a randomized controlled trial of a 6-month internet-based, mindfulness-based cognitive behavioral therapy intervention for youth. Interact J Med Res 2023;12:e46419 [FREE Full text] [doi: 10.2196/46419] [Medline: 38064262]
29. Koyano W, Shibata H, Nakazato K, Haga H, Suyama Y. Measurement of competence: reliability and validity of the TMIG Index of Competence. Arch Gerontol Geriatr 1991;13(2):103-116. [doi: 10.1016/0167-4943(91)90053-s] [Medline: 15374421]
30. Holmberg MJ, Andersen LW. Adjustment for baseline characteristics in randomized clinical trials. JAMA 2022;328(21):2155-2156. [doi: 10.1001/jama.2022.21506] [Medline: 36394881]
31. Nguyen DV, Sentürk D, Carroll RJ. Covariate-adjusted linear mixed effects model with an application to longitudinal data. J Nonparametr Stat 2008;20(6):459-481 [FREE Full text] [doi: 10.1080/10485250802226435] [Medline: 19266053]
32. Patel RJS, Ding J, Marvel FA, Shan R, Plante TB, Blaha MJ, et al. Associations of demographic, socioeconomic, and cognitive characteristics with mobile health access: MESA (multi-ethnic study of atherosclerosis). J Am Heart Assoc 2022;11(17):e024885 [FREE Full text] [doi: 10.1161/JAHA.121.024885] [Medline: 36056720]
33. van Het Reve E, Silveira P, Daniel F, Casati F, de Bruin ED. Tablet-based strength-balance training to motivate and improve adherence to exercise in independently living older people: part 2 of a phase II preclinical exploratory trial. J Med Internet Res 2014;16(6):e159 [FREE Full text] [doi: 10.2196/jmir.3055] [Medline: 24966165]
34. Lee KH, Choi YY, Jung ES. Effectiveness of an oral health education programme using a mobile application for older adults: a randomised clinical trial. Gerodontology 2023;40(1):47-55. [doi: 10.1111/ger.12616] [Medline: 34951060]
35. Sun C, Sun L, Xi S, Zhang H, Wang H, Feng Y, et al. Mobile phone-based telemedicine practice in older Chinese patients with type 2 diabetes mellitus: randomized controlled trial. JMIR Mhealth Uhealth 2019;7(1):e10664 [FREE Full text] [doi: 10.2196/10664] [Medline: 30609983]
36. Tsuji T, Miyaguni Y, Kanamori S, Hanazato M, Kondo K. Community-level sports group participation and older individuals' depressive symptoms. Med Sci Sports Exerc 2018;50(6):1199-1205 [FREE Full text] [doi: 10.1249/MSS.0000000000001541] [Medline: 29298218]

XSL•FO
RenderX

37. Mair JL, Castro O, Salamanca-Sanabria A, Frese BF, von Wangenheim F, Tai ES, et al. Exploring the potential of mobile health interventions to address behavioural risk factors for the prevention of non-communicable diseases in Asian populations: a qualitative study. BMC Public Health 2023;23(1):753 [FREE Full text] [doi: 10.1186/s12889-023-15598-8] [Medline: 37095486]
38. Subramaniam A, Hensley E, Stojancic R, Vaughn J, Shah N. Careful considerations for mHealth app development: lessons learned from QuestExplore. Mhealth 2022;8:24 [FREE Full text] [doi: 10.21037/mhealth-21-51] [Medline: 35928515]
39. Tizuka MM, Clua EWG, de Castro Salgado LC. Investigating m-health gamification rewards elements for adults 50+. : IEEE; 2020 Presented at: 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH); 2020 Aug 12-14; BC, Canada. [doi: 10.1109/SeGAH49190.2020.9201658]
40. Hamari J, Koivisto J, Sarsa H. Does gamification work? -- a literature review of empirical studies on gamification. : IEEE; 2014 Presented at: 2014 47th Hawaii International Conference on System Sciences; 2014; Tampere, Finland p. 3025-3034. [doi: 10.1109/hicss.2014.377]
41. Ryan RM, Patrick H, Deci EL, Williams GC. Facilitating health behaviour change and its maintenance: interventions based on self-determination theory.: The European Health Psychologist; 2008. URL: https://selfdeterminationtheory.org/wp-content/uploads/2014/04/2008_RyanPatrickDeciWilliams_EHP.pdf [accessed 2024-09-20]
42. Bonvanie A, Broekhuis M, Janssen O, Maeckelberghe E, Wortmann JHC. Health self-management applications in the work environment: the effects on employee autonomy. Front Digit Health 2020;2:9 [FREE Full text] [doi: 10.3389/fdgth.2020.00009] [Medline: 34713022]
43. Zhan ES, Molina MD, Rheu M, Peng W. What is there to fear? Understanding multi-dimensional fear of AI from a technological affordance perspective. International Journal of Human–Computer Interaction 2023:1-18. [doi: 10.1080/10447318.2023.2261731]
44. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
45. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
46. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
47. Chaudhry UAR, Wahlich C, Fortescue R, Cook DG, Knightly R, Harris T. The effects of step-count monitoring interventions on physical activity: systematic review and meta-analysis of community-based randomised controlled trials in adults. Int J Behav Nutr Phys Act 2020;17(1):129 [FREE Full text] [doi: 10.1186/s12966-020-01020-8] [Medline: 33036635]
48. Oba T, Takano K, Katahira K, Kimura K. Use patterns of smartphone apps and wearable devices supporting physical activity and exercise: large-scale cross-sectional survey. JMIR Mhealth Uhealth 2023;11:e49148 [FREE Full text] [doi: 10.2196/49148] [Medline: 37997790]
49. Smail-Crevier R, Powers G, Noel C, Wang J. Health-related internet usage and design feature preference for e-mental health programs among men and women. J Med Internet Res 2019;21(3):e11224 [FREE Full text] [doi: 10.2196/11224] [Medline: 30882361]

Abbreviations

ANCOVA: analysis of covariance
ESP: Encouragement of Social Participation
IADL: instrumental activity of daily living

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 11.07.24; peer-reviewed by Y Wu, T Moe-Byrne; comments to author 06.08.24; revised version received 24.08.24; accepted 04.09.24; published 30.09.24.

Please cite as:
Kawaguchi K, Nakagomi A, Ide K, Kondo K
Effects of a Mobile App to Promote Social Participation on Older Adults: Randomized Controlled Trial
J Med Internet Res 2024;26:e64196
URL: https://www.jmir.org/2024/1/e64196
doi:10.2196/64196
PMID:39348180

XSL•FO
RenderX

©Kenjiro Kawaguchi, Atsushi Nakagomi, Kazushige Ide, Katsunori Kondo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Health Maintenance Organization–mHealth Versus Face-to-Face Interaction for Health Care in Israel: Cross-Sectional Web-Based Survey Study

Avi Zigdon[1], MHA, PhD; Moti Zwilling[2], MBA, MSc, MA, PhD; Ofek Zigdon[3], BSc, BMedSc, MBA; Orna Reges[1], MPH, PhD

[1]Department of Health Systems Management, School of Health Sciences, Ariel University, Ariel, Israel

[2]Department of Economics and Business Administration, Ariel University, Ariel, Israel

[3]Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

Corresponding Author:
Avi Zigdon, MHA, PhD
Department of Health Systems Management
School of Health Sciences
Ariel University
Science Park, POB 3
Ariel, 40700
Israel
Phone: 972 74 7288070
Email: aviz@ariel.ac.il

Abstract

Background: Health maintenance organization–mobile health (HMO-mHealth) services have a direct impact on patients' daily lives, and HMOs regularly expand their range of mHealth services. HMO-mHealth apps are saving HMOs time and money, as services are becoming more accessible to patients. However, the willingness to use mHealth apps depends on user perception. Although mHealth apps can change the relationship dynamic between HMOs and patients, patients prefer to use them to facilitate face-to-face interactions rather than replace them.

Objective: This study aims to examine the extent to which Israeli adults prefer adopting health care services using HMO-mHealth as a replacement for face-to-face interaction.

Methods: Israeli adults aged ≥18 years completed an electronic questionnaire. Data were collected from December 2020 to February 2021. All services in the main HMO-mHealth apps of the 4 Israeli HMOs were mapped. The 29 health care services used in this study were identical in all 4 HMO-mHealth apps in Israel. The association between sociodemographic characteristics and health condition with preference for HMO-mHealth or face-to-face interaction was analyzed separately for each health service by using a logistic model.

Results: A total of 6321 respondents completed the questionnaire (female: 4296/6321, 68%; male: 2025/6321, 32%). Approximately 80.9% (5115/6321) to 88.2% (5578/6321) of the respondents preferred using HMO-mHealth apps for administrative matters. However, 55.3% (3498/6321), 52.2% (3301/6321), and 46.9% (2969/6321) preferred face-to-face meetings for the initial medical diagnosis, medical treatment, and medical diagnosis results, respectively. Seven main variables were found to be associated with HMO-mHealth adoption, including gender, age, education, marital status, religious affiliation, and subjective health condition. Female respondents were more likely than male respondents to prefer HMO-mHealth apps for administrative matters and face-to-face interaction for personal medical diagnosis and treatment (odds ratio [OR] 0.74, 95% CI 0.67-0.83; $P<.001$ and OR 0.82, 95% CI 0.74-0.92; $P<.001$, respectively). Married individuals preferred using HMO-mHealth apps over face-to-face meetings for a new medical diagnosis (OR 1.31, 95% CI 1.15-1.49; $P<.001$) or treatment (OR 1.34, 95% CI 1.18-1.52; $P<.001$). Improved health perception was associated with higher preference for HMO-mHealth apps across all health care services in this study (OR 1.11, 95% CI 1.02-1.22; $P<.02$ to OR 1.38, 95% CI 1.25-1.53; $P<.001$). No significant association was found between the presence of a chronic disease and the preferred mode of interaction for most services.

Conclusions: HMO-mHealth is proving to be a robust and efficient tool for health care service delivery. However, there are barriers that affect vulnerable populations when adopting HMO-mHealth. Therefore, it is important to tailor HMO-mHealth apps

XSL•FO
RenderX

for older adults, the chronically ill, and minorities in society, as these groups have a greater need for these services. Future studies should focus on identifying the barriers that affect the utilization of HMO-mHealth in these groups.

(J Med Internet Res 2024;26:e55350) doi:10.2196/55350

KEYWORDS

HMO-mHealth; mHealth; face-to-face; digital health; digital health apps; eHealth; HMO-mHealth adoption; health care; mHealth adoption; mobile phone; HMO; health maintenance organization

Introduction

The rapid development of mobile health (mHealth) directly impacts patients' daily lives, relationships, and communication with health maintenance organizations (HMOs). In the absence of an epidemic and its consequences, the use of mHealth apps is at the discretion of the patient [1,2], and only a few use them frequently [3,4]. Typically, mHealth apps are used to independently manage medical interventions and personalize treatments, aiming to reduce the demand on health care providers and eliminate geographic barriers [5]. Nevertheless, the potential of mHealth is vast [6]. The medical information and services provided to patients by HMOs can be personalized to their specific needs anytime and anywhere [7,8]. Utilization of mHealth is beneficial in many ways. It enables self-management of disease, evaluation of personal medical information, and easier and more accessible contact with medical professionals [9-13]. mHealth can also help support patients' lifestyles [13], improve patients' health status [14], improve their quality of life and health outcomes, and reduce the incidence of disease [12]. Sometimes, they even help with hospitalization [2] and take over the roles of general practitioners [15]. In addition, mHealth influences patients' sense of autonomy and their need for well-being [2].

HMOs are regularly expanding the services offered in mHealth. The use of information and communication technologies to meet the needs of health systems around the world is increasing [16]. In England, for example, the National Health Service has developed mHealth apps that allow users to make appointments with doctors, request prescription renewals and order medications, receive medical advice, view personal medical records, declare their willingness to donate organs, and with the COVID-19 pandemic outbreak, obtain authorizations and information for outbreak durations [17].

In Germany, the Digital Healthcare Act, which was introduced in 2019, allows health insurers to promote the development of apps that include the ability to renew a prescription, provide medical advice via video calls, and access a secure data network from anywhere in the country. In addition, the use of mHealth apps can help promote digital health literacy among patients to achieve equal access and high involvement on the patients' part [18]. Like the health systems in England and Germany, that in Israel is primarily a public health system. According to the Israeli health insurance law [19], every resident in Israel is entitled to health insurance that includes a standard basket of services, which is termed the Basic Healthcare Basket, established by law. In addition to this system, residents can expand the basic services basket by purchasing an additional basket of services from HMOs. Health care services are provided to all Israeli residents by 4 health care providers (Clalit Health Services, Maccabi Healthcare Services, Leumit Health Services, and Kupat Holim Meuhedet) to which the residents are registered [19]. In May 2011, mHealth apps launched by 2 HMO funds were launched: Maccabi Healthcare Services [20] and Leumit Health Services [21]. In August 2011, the mHealth app of Kupat Holim Meuhedet was launched, and in February 2012, the mHealth app of Clalit Health Services was launched [22,23]. Over the years, the services offered by HMOs through mHealth apps have evolved. Currently, they offer a wide range of services both in the main health care services app and in additional apps (eg, childcare support, pregnancy support). During the COVID-19 pandemic, the use of telehealth services increased [24], and there was an increase in the use of HMO mHealth services, especially among patients with chronic diseases [25]. There was also an increase in the use of COVID-19 management apps (monitoring exposure to diagnosed patients and diagnosing symptoms) [24]. The use of remote medicine services has also increased in Israel [26,27], wherein some of these health care services were provided by the HMO-mHealth apps [20-23]. However, the willingness to use mHealth apps is influenced by the user's perception of perceived usefulness, perceived ease of use, subjective norms [28], app quality, and user health literacy [29,30]. mHealth apps are promising platforms that can be used to provide efficient and convenient access to therapeutic services [31]. The use of mobile platforms is highly recommended in combination with face-to-face interventions to support people with their daily routines [6]. However, mHealth appointments were found to be more expensive than face-to-face health care appointments [32]. A meta-ethnographic review of qualitative studies by Vo et al [6] sheds light on the dynamics of patient engagement and opportunities to help patients become more empowered through the use of mHealth apps. Vo et al [6] reviewed 43 papers on qualitative studies that addressed patient evaluation, expectations, and perceptions of mHealth apps. Patients described mHealth apps as tools that enable discussions with their health care providers that can improve adherence to care and their health care experience. mHealth apps were described as a tool for patients to engage in a 2-way dialogue. Although mHealth apps can change the dynamics of patient-provider relationships by providing relevant information for conducting assessments, diagnoses, registering for treatments, etc, patients prefer to use them simply as a tool to facilitate rather than replace these relationships [6]. The aim of this study was to examine the extent to which adults in Israel prefer to adopt HMO-mHealth apps and receive health care services via HMO-mHealth as a substitute for face-to-face interaction and to identify factors associated with these preferences.

XSL•FO
RenderX

Methods

Study Design

We designed a cross-sectional web-based survey to examine the preferences of adults in Israel for adopting HMO-mHealth apps and receiving health care services versus HMO-mHealth as a substitute for face-to-face interaction. Data for this study were collected from December 2020 to February 2021. Respondents were not required to provide any identifying information, and any participant could stop the survey at any point.

Participant Recruitment

The questionnaire was designed on the Qualtrics platform and distributed by research assistants via social media (WhatsApp groups, Facebook, etc). Our target population was Israeli adults aged ≥18 years, who were sampled using convenience sampling. Responders were allowed to fill in the questionnaire electronically only once. To estimate the required sample size, we relied on data from the Central Bureau of Statistics in Israel. The population of the State of Israel in January 2024 was estimated to be around 9.855 million [33]. Since the proportion of mHealth use in the population is unknown and is assumed to be a maximum of 50%, with a significance level of 5%, a sample size of 384 residents was calculated for this study. In practice, there was a high response rate, and 6321 Israeli adults completed the research questionnaire. It was decided to use all the data to obtain greater validity.

Survey Development and Definition of Variables

All services available in the main HMO-mHealth apps of the 4 Israeli HMOs were mapped. A total of 29 health care services, which were identical across all 4 HMO-mHealth apps in Israel, were selected for this study and included in the HMO-mHealth/face-to-face interaction preference questionnaire. These services included administrative matters such as setting a visit or changing a visit date as well as medical services such as receiving a medical diagnosis or treatment. For each of the 29 items, participants were asked whether they preferred mHealth or face-to-face interaction to receive the health service. Each item was tested as a separate outcome. In addition, the following information was self-reported by the participants: gender, age (years), marital status (married/single/divorced/widowed), number of children, birthplace (Israel/other), religion (Jewish/non-Jewish), religiosity (secular/religious), education (uneducated/elementary-middle school/high school/vocational training/BA/MA/PhD), residence (Center, North, or South Jerusalem), subjective health condition (from 1: worst to 5: very good), and previous diagnosis of chronic illness by a licensed physician (yes/no).

Data Analysis

The main characteristics of the study population were described using percentages for categorical variables and means with standard deviations for continuous variables. Chi-square tests or independent-sample 2-sided *t* tests were employed to evaluate the association of categorical or continuous predictors, respectively, with each of the 29 health services included in this study. Variables that were significant ($P<.05$) in the univariable tests for at least one of the health care services were included in the final multivariable logistic models. The association between significant personal characteristics with the preference of mHealth or face-to-face interaction was assessed separately for each health care service by using multivariable logistic regression analysis. In each model, a particular health service was evaluated as the dependent variable (0=face-to-face; 1=mHealth). The predictor variables were entered simultaneously into each model and included sociodemographic characteristics (gender, age, marital status, birthplace, number of children, education, religion, religiosity, residence), subjective health condition, and previous diagnosis of chronic illness. To mitigate the potential for spurious statistical significance arising from multiple tests, Benjamini-Hochberg adjusted *P* values were computed using WinPepi version 11.65 [34]. *P* values less than .05 were considered statistically significant.

Ethics Approval

This study was approved by the ethics committee of Ariel University (AU-HEA-AZ-20201217). All participants were informed of the aims of the study in an introductory section, and they gave their informed consent to participate in the study. No conditioning questions were asked to conduct the survey, as health care services are provided to all residents of Israel by 4 health care providers to which the residents are registered.

Results

Between December 2020 and February 2021, 6321 Israeli adults completed the study questionnaire. The main characteristics of the study population are described in Table 1. The mean age of the participants was 35.42 (SD 13.09) years, and the majority were female (4296/6321, 68%), Jewish (5234/6321, 82.8%), and born in Israel (5283/6321, 83.6%). About half were married (3299/6321, 52.2%), lived in the center of Israel (3104/6321, 49.1%), had no children (3253/6321, 51.5%), and had an academic degree (BA, MA, or PhD: 3152/6321, 49.9%). More than half described themselves as secular or traditional (3546/6321, 56.1%). Most participants had no chronic diseases (5235/6321, 82.8%) and declared that they were in very good health condition (3775/6321, 59.7%).

XSL•FO
RenderX

Table 1. General characteristics of the participants (N=6321).

Demographic variables	Values
Gender, n (%)	
Female	4296 (68)
Male	2025 (32)
Age (years)	
≤39, n (%)	4383 (69.3)
40-64, n (%)	1723 (27.3)
65-74, n (%)	164 (2.6)
≥75, n (%)	51 (0.8)
Mean (SD)	35.4 (13.1)
Marital status, n (%)	
Married	3299 (52.2)
Single	2708 (42.8)
Divorced/widowed	313 (5)
Children, n (%)	
0	3253 (51.5)
1	557 (8.8)
2	803 (12.7)
3	787 (12.4)
4	493 (7.8)
5	219 (3.5)
≥6	209 (3.3)
Birthplace, n (%)	
Israel	5283 (83.6)
Other	1038 (16.4)
Religious affiliation, n (%)	
Jewish	5234 (82.8)
Non-Jewish	1087 (17.2)
Religiosity, n (%)	
Secular	3546 (56.1)
Religious	2775 (43.9)
Education, n (%)	
Uneducated	50 (0.8)
Elementary/middle school	135 (2.1)
High school	2245 (35.5)
Vocational training	739 (11.7)
Bachelor of Arts	2385 (37.7)
Master of Arts/Doctor of Philosophy	767 (12.2)
Subjective health condition , n (%)	
Very good	3775 (59.7)
Good	2184 (34.6)
Moderate	299 (4.7)
Not good	54 (0.9)

XSL•FO
RenderX

Demographic variables	Values
Bad	9 (0.1)
Chronic illness, n (%)	
Yes	1086 (17.2)
No	5235 (82.8)

Figure 1 shows participants' preferences for receiving health care services via HMO-mHealth apps or face-to-face interaction for 29 health services offered by all HMO-mHealth apps in Israel. The vast majority of the participants (5115/6321, 80.9% to 5578/6321, 88.2%) preferred to use HMO-mHealth apps for administrative services such as changing visit dates, scheduling a visit for vaccination, submitting a prescription renewal request, and generation of a sick-day certificate. When the primary purpose of the visit was for medical services involving treatment or diagnostics, more individuals preferred face-to-face meetings over HMO-mHealth services. Specifically, 55.3% (3498/6321) of the participants reported that they preferred a face-to-face meeting to receive the initial medical diagnosis, and 52.2% (3301/6321) reported that they preferred a face-to-face meeting to receive medical treatment. Approximately 46.9% (2969/6321) preferred a face-to-face meeting for receiving medical diagnosis results.

XSL•FO
RenderX

Figure 1. Participants' preferences for receiving health care services via health maintenance organization–mobile health apps or face-to-face interaction for 29 health services included in this study. mHealth: mobile health.

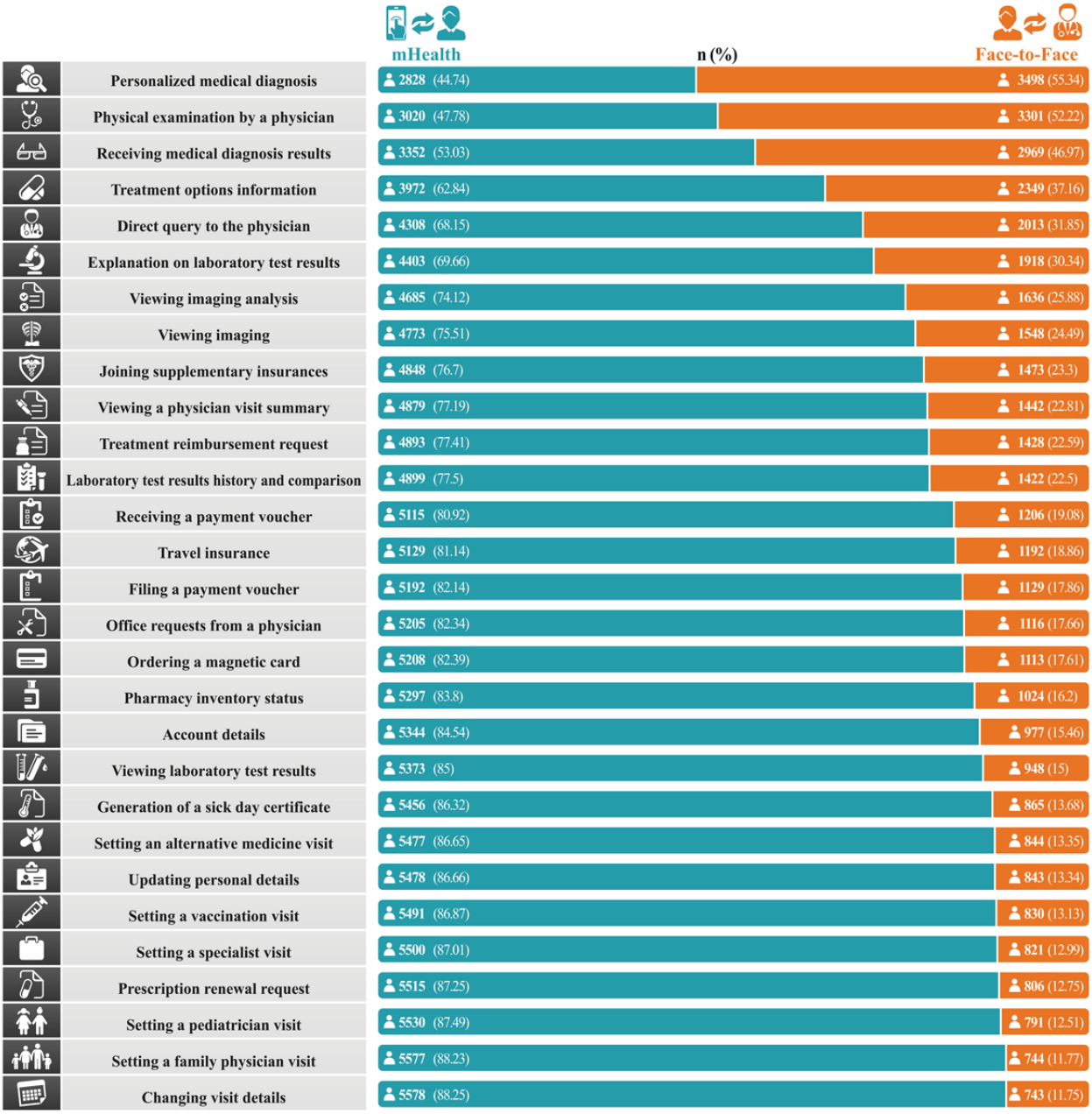

Figure 2 summarizes the association of personal characteristics with the preference for mHealth or face-to-face interaction. The multivariable logistic regression models revealed differences in the factors associated with preferences for use of health services via HMO-mHealth apps or face-to-face interaction. Among sociodemographic characteristics, sex, age, marital status, education level, ethnicity, and religiosity were significantly associated with the preference of mHealth/face-to-face interaction for most health care services. Specifically, compared to unmarried individuals, married individuals preferred using HMO-mHealth apps over face-to-face meetings to obtain a new medical diagnosis (odds ratio [OR] 1.31, 95% CI 1.15-1.49; *P*<.001) or treatment (OR 1.34, 95% CI 1.18-1.52; *P*<.001) and for changing visit details (OR 1.35, 95% CI 1.10-1.66; *P*=.004) or scheduling a new visit with a primary care physician (OR 1.39, 95% CI 1.13-1.71; *P*<.001), pediatrician (OR 1.54, 95% CI 1.26-1.89; *P*<.001), or specialist (OR 1.36, 95% CI 1.12-1.66; *P*<.001). Individuals with an academic background were more likely to prefer HMO-mHealth apps over face-to-face meetings when administrative issues were involved, such as making an appointment with the family physician (OR 1.81, 95% CI 1.53-2.14; *P*<.001) or changing visit details (OR 1.89, 95% CI 1.60-2.24; *P*<.001) and for medical information such as direct query for the physician (OR 1.32, 95% CI 1.18-2.48; *P*<.001), viewing imaging analysis (OR 1.43, 95% CI 1.27-1.61; *P*<.001), or a visit summary (OR 1.64, 95% CI 1.45-1.87; *P*<.001). However, no association was found between academic status and preferred mode of interaction for medical diagnosis and treatment. Non-Jewish and religious individuals were less likely to use HMO-mHealth apps and preferred face-to-face interaction for both administrative and medical purposes. Female

XSL•FO
RenderX

respondents were more likely than male respondents to prefer HMO-mHealth apps for administrative services and face-to-face interaction for personal medical diagnosis or treatment (OR 0.74, 95% CI 0.67-0.83; *P*<.001 and OR 0.82, 95% CI 0.74-0.92; *P*<.001, respectively).

A better subjective perception of health condition was associated with a higher preference for using HMO-mHealth apps. However, no significant association was found between the presence of chronic disease and the preferred mode of interaction for most services.

Figure 2. Multivariable logistic regression models of the factors associated with the adoption of health maintenance organization–mobile health versus face-to-face interaction (please also refer to Multimedia Appendix 1). HMO: health maintenance organization; mHealth: mobile health; OR: odds ratio.

Health care services \ Main characteristics	Female OR (*P* value)	Age OR (*P* value)	Married OR (*P* value)	Born in Israel OR (*P* value)	Children OR (*P* value)	Academic OR (*P* value)	Non-Jewish OR (*P* value)	Religious OR (*P* value)	Living in the north OR (*P* value)	Living in the south OR (*P* value)	Living in Jerusalem area OR (*P* value)	Chronic disease OR (*P* value)	Subjective health condition OR (*P* value)
Personalized medical diagnosis	0.74 (<.001)	1.00 (.19)	1.31 (<.001)	1.21 (.04)	0.98 (.64)	1.00 (.99)	0.79 (<.001)	0.86 (.01)	0.99 (.85)	0.87 (.29)	0.89 (.25)	0.95 (.75)	1.16 (<.001)
Physical examination by a physician	0.82 (<.001)	0.99 (.05)	1.34 (<.001)	1.14 (.12)	0.99 (.81)	1.10 (.07)	0.70 (<.001)	0.76 (<.001)	1.05 (.55)	0.79 (.07)	0.80 (<.001)	1.00 (.95)	1.16 (<.001)
Receiving medical diagnosis results	0.70 (<.001)	1.00 (.25)	1.28 (<.001)	1.17 (.07)	1.02 (.64)	1.14 (.011)	0.82 (<.001)	0.86 (<.001)	0.93 (.41)	0.83 (.15)	0.88 (.23)	0.92 (.52)	1.13 (<.001)
Treatment options information	0.79 (<.001)	0.99 (<.001)	1.26 (<.001)	1.02 (.77)	1.03 (.64)	1.22 (<.001)	0.64 (<.001)	0.76 (<.001)	0.95 (.56)	0.76 (<.001)	0.95 (.57)	0.99 (.95)	1.23 (<.001)
Direct query to the physician	0.96 (.57)	0.99 (.012)	1.58 (<.001)	1.28 (<.001)	1.06 (.39)	1.32 (<.001)	0.49 (<.001)	0.71 (<.001)	0.91 (.33)	0.75 (<.001)	0.81 (<.001)	1.14 (.44)	1.11 (.02)
Explanation on laboratory test results	1.01 (.91)	0.99 (.02)	1.64 (<.001)	1.15 (.14)	1.00 (.95)	1.40 (<.001)	0.55 (<.001)	0.82 (<.001)	0.91 (.34)	0.92 (.57)	0.95 (.57)	1.21 (.19)	1.21 (<.001)
Viewing imaging analysis	1.02 (.87)	0.98 (<.001)	1.53 (<.001)	1.08 (.40)	1.04 (.64)	1.43 (<.001)	0.46 (<.001)	0.73 (<.001)	0.77 (<.001)	0.79 (.12)	0.94 (.57)	1.30 (<.001)	1.22 (<.001)
Viewing imaging	1.06 (.45)	0.98 (<.001)	1.43 (<.001)	1.13 (.18)	1.06 (.39)	1.44 (<.001)	0.39 (<.001)	0.75 (<.001)	0.77 (<.001)	0.80 (.15)	0.91 (.45)	1.18 (.24)	1.15 (.01)
Joining supplementary insurances	0.99 (.91)	0.99 (<.001)	1.24 (.01)	1.09 (.38)	0.98 (.64)	1.46 (<.001)	0.40 (<.001)	0.69 (<.001)	0.76 (<.001)	0.80 (.15)	0.94 (.57)	0.99 (.95)	1.28 (<.001)
Viewing a physician visit summary	1.10 (.21)	0.98 (<.001)	1.55 (<.001)	1.21 (.07)	1.03 (.64)	1.64 (<.001)	0.35 (<.001)	0.67 (<.001)	0.80 (=.02)	0.71 (<.001)	0.87 (.26)	1.12 (.48)	1.22 (<.001)
Treatment reimbursement request	0.95 (.52)	0.98 (<.001)	1.37 (<.001)	1.26 (.04)	1.06 (.39)	1.76 (<.001)	0.40 (<.001)	0.67 (<.001)	0.89 (.32)	0.83 (.22)	1.11 (.41)	1.03 (.91)	1.34 (<.001)
Laboratory test results history and comparison	1.10 (.21)	0.99 (<.001)	1.63 (<.001)	1.15 (.16)	0.98 (.70)	1.51 (<.001)	0.39 (<.001)	0.83 (<.001)	0.83 (.09)	0.90 (.51)	1.01 (.91)	1.34 (<.001)	1.27 (<.001)
Receiving a payment voucher	1.30 (<.001)	0.98 (<.001)	1.70 (<.001)	1.44 (<.001)	1.03 (.64)	1.62 (<.001)	0.28 (<.001)	0.68 (<.001)	0.75 (<.001)	0.80 (.15)	1.10 (.51)	1.22 (.23)	1.24 (<.001)
Travel insurance	1.05 (.57)	0.99 (<.001)	1.30 (<.001)	1.23 (.06)	0.98 (.64)	1.60 (<.001)	0.41 (<.001)	0.70 (<.001)	0.73 (<.001)	0.80 (.15)	1.13 (.36)	1.13 (.48)	1.38 (<.001)
Filing a payment voucher	1.17 (.05)	0.99 (<.001)	1.46 (<.001)	1.32 (<.001)	1.03 (.64)	1.73 (<.001)	0.30 (<.001)	0.68 (<.001)	0.79 (.02)	0.89 (.51)	1.19 (.25)	1.15 (.48)	1.31 (<.001)
Office requests from a physician	1.35 (<.001)	0.98 (<.001)	1.58 (<.001)	1.17 (.17)	1.05 (.64)	1.82 (<.001)	0.29 (<.001)	0.66 (<.001)	0.79 (.02)	0.84 (.31)	0.97 (.80)	1.12 (.50)	1.26 (<.001)
Ordering a magnetic card	1.08 (.41)	0.98 (<.001)	1.25 (.01)	1.19 (.13)	1.04 (.64)	1.69 (<.001)	0.36 (<.001)	0.64 (<.001)	0.77 (.02)	0.80 (.15)	1.07 (.57)	1.07 (.75)	1.34 (<.001)
Pharmacy inventory status	1.18 (.05)	0.98 (<.001)	1.39 (<.001)	1.12 (.33)	1.01 (.90)	1.80 (<.001)	0.29 (<.001)	0.61 (<.001)	0.77 (.02)	0.83 (.29)	0.95 (.66)	1.03 (.92)	1.25 (<.001)
Account details	1.23 (.02)	0.99 (.14)	1.29 (.01)	1.30 (.04)	0.99 (.90)	1.74 (<.001)	0.34 (<.001)	0.69 (<.001)	0.69 (<.001)	0.92 (.62)	1.07 (.57)	1.03 (.92)	1.35 (<.001)
Viewing laboratory test results	1.33 (<.001)	0.98 (<.001)	1.61 (<.001)	1.17 (.18)	1.04 (.64)	1.73 (<.001)	0.33 (<.001)	0.75 (<.001)	0.81 (.10)	0.78 (.15)	1.03 (.82)	1.24 (.23)	1.28 (<.001)
Generation of a sick day certificate	1.39 (<.001)	0.98 (<.001)	1.56 (<.001)	1.31 (.04)	1.05 (.64)	1.80 (<.001)	0.25 (<.001)	0.62 (<.001)	0.86 (.24)	0.87 (.49)	1.23 (.24)	1.14 (.50)	1.37 (<.001)
Setting an alternative medicine visit	1.45 (<.001)	0.98 (<.001)	1.36 (<.001)	1.13 (.33)	1.00 (.95)	1.82 (<.001)	0.31 (<.001)	0.64 (<.001)	0.83 (.15)	1.00 (.99)	1.04 (.80)	1.04 (.91)	1.34 (<.001)
Updating personal details	1.38 (<.001)	0.98 (<.001)	1.58 (<.001)	1.31 (.04)	1.00 (.95)	1.88 (<.001)	0.29 (<.001)	0.65 (<.001)	0.76 (.02)	0.86 (.48)	1.06 (.66)	1.18 (.44)	1.32 (<.001)
Setting a vaccination visit	1.23 (.02)	0.98 (<.001)	1.43 (<.001)	1.13 (.33)	0.97 (.64)	2.01 (<.001)	0.30 (<.001)	0.64 (<.001)	0.89 (.33)	0.94 (.76)	1.34 (.07)	1.05 (.90)	1.35 (<.001)
Setting a specialist visit	1.33 (<.001)	0.98 (<.001)	1.36 (<.001)	1.27 (.06)	1.02 (.70)	1.86 (<.001)	0.31 (<.001)	0.57 (<.001)	0.88 (.33)	0.99 (.99)	1.13 (.45)	1.15 (.48)	1.28 (<.001)
Prescription renewal request	1.49 (<.001)	0.98 (<.001)	1.67 (<.001)	1.28 (.07)	1.04 (.64)	1.78 (<.001)	0.22 (<.001)	0.57 (<.001)	0.85 (.21)	0.95 (.81)	1.41 (<.001)	1.08 (.75)	1.27 (<.001)
Setting a pediatrician visit	1.42 (<.001)	0.97 (<.001)	1.54 (<.001)	1.25 (.08)	1.01 (.81)	1.94 (<.001)	0.30 (<.001)	0.60 (<.001)	0.82 (.14)	0.93 (.72)	1.18 (.32)	1.03 (.92)	1.21 (<.001)
Setting a family physician visit	1.44 (<.001)	0.98 (<.001)	1.39 (<.001)	1.04 (.77)	1.02 (.70)	1.81 (<.001)	0.29 (<.001)	0.59 (<.001)	0.83 (.17)	0.88 (.51)	1.17 (.36)	1.06 (.84)	1.31 (<.001)
Changing visit details	1.40 (<.001)	0.97 (<.001)	1.35 (<.001)	1.08 (.52)	1.03 (.64)	1.89 (<.001)	0.30 (<.001)	0.61 (<.001)	0.88 (.33)	0.86 (.48)	1.19 (.32)	1.11 (.65)	1.25 (<.001)

Face-to-face preference | HMO-mHealth preference | No significant preference

Discussion

Principal Findings

Adoption of mHealth by patients is a complex and multidimensional process with its own advantages, disadvantages, and barriers. This study was the first to examine the extent to which adults in Israel prefer to receive health care services through HMO-mHealth as a substitute for face-to-face interaction and to identify factors associated with these preferences. Approximately 46.9% (2969/6321) of our study participants preferred face-to-face interaction for the initial treatment or diagnosis rather than through HMO-mHealth apps (represented by the items, namely, personalized medical diagnosis, physical examination by a physician, and receiving medical diagnosis results). However, 80.9% (5115/6321) to 88.2% (5578/6321) of our study participants were interested in receiving administrative services through HMO-mHealth apps (demonstrated in items such as changing visit details, setting a family physician visit, and prescription renewal request). Nevertheless, 11.7% (743/6321) to 19.1% (1206/6321) of the

XSL•FO
RenderX

participants were still interested in receiving administrative and information services through face-to-face meetings. Our axiom is that the respondents have adequate digital literacy since the questionnaire was distributed and filled out using a digital tool. However, good digital literacy is not necessarily a guarantee of the ability to use digital health services [1], and the validity of information in mHealth apps that focus on emotional content may influence the adoption of HMO-mHealth services, even among those with high digital literacy [6,35]. The high usage rate of HMO-mHealth in this study can be explained by the Unified Theory of Acceptance and Use of Technology (UTAUT), which states that effort expectancy (defined as the degree of ease of use) affects the acceptance and use of HMO-mHealth [36]. Seven main variables were found to be associated with HMO-mHealth adoption, including gender, age, education, marital status, religious affiliation, and subjective health condition.

Our results show that women are more likely to adopt HMO-mHealth apps for medical information or administrative services (eg, prescription renewal request, setting a family physician visit, account details inquiry). However, they prefer face-to-face interactions with physicians for receiving medical services with treatment or diagnostic significance (eg, personal medical diagnosis, medical examination, receiving the results of medical diagnoses, information on treatment options). This result is interesting because one might assume that gender has a similar effect on the adoption of all HMO-mHealth services and that there are no differences between the different types of services. Previous studies suggest that there are generally no clear trends in gender differences in the willingness to adopt mHealth technologies [3,14,25,27,36-38]; gender is mentioned as a moderating variable in mHealth studies dealing with UTAUT, and it plays a moderating role with threat appraisal and coping appraisal factors in mHealth adoption behavior [39,40]. However, traditional face-to-face health care is frequently preferred over digital health care [1], and women, in particular, tend to be more skeptical about relying solely on mHealth for treatment and place higher value on in-person doctor visits [41]. This inclination is likely influenced by greater health anxiety and risk aversion among women [42-45].

Consistent with previous studies [11,14,27,37], our study also shows that age, education, religious affiliation, and ethnicity were barriers to HMO-mHealth adoption. On the one hand, people with a higher level of education and younger individuals had a higher tendency to adopt HMO-mHealth services compared to older individuals and those with a lower level of education who preferred face-to-face meetings. On the other hand, societal minorities (non-Jews) and religious individuals tended to prefer face-to-face meetings over the adoption of HMO-mHealth services. It is possible that internet use is perceived as more accessible among young people, and those with higher education are perceived as having better digital literacy [4,46]. However, our study shows that people with a higher level of education and younger individuals who have a poor subjective health condition generally prefer face-to-face meetings. Based on this finding, it can be assumed that despite the effectiveness of HMO-mHealth services, young and educated people still find it necessary to have human contact with a health provider for diagnosis and treatment.

This study also shows that ethnic minorities (non-Jewish) and religious people would prefer face-to-face meetings over HMO-mHealth for all health care provider services examined in this study. There is evidence of the effectiveness of internet-based interventions that improve the health of ethnic minorities [47]. These ethnic minority groups reported greater interest in using mHealth technologies than the nonminority population [48]. A recent study found that there are significant ethnic differences in the use of telemedicine between different ethnic groups [27]. However, the association of religion with the adoption of health technologies has, to our knowledge, not yet been sufficiently investigated. The adoption of technologies by religious people is considered at higher-than-average risk, and previous studies have shown that religious leaders can influence the opinions and behaviors of religious people when adopting medical technologies [49]. We link these 2 facts because we believe that barriers to HMO-mHealth adoption can be associated with ethnicity and religion. Both cultural and community characteristics, language barriers, emotional state, and health and digital literacy can be assumed to play an important role in HMO-mHealth adoption [46,50]. A study examining the influence of cultural aspects on the adoption of mHealth in 3 different countries, based on UTAUT, found that cultural differences have a decisive effect on the adoption behavior model and concluded that mHealth services must be adapted to the culture of the population for which the services are developed [51].

Greater involvement of minorities in the development of HMO-mHealth will lead to the adoption of these tools by minorities [52], and the involvement of leaders from these communities in the development of technological medical products will enable broader adoption of HMO-mHealth in these populations. Another interesting finding of this study is that married people would prefer to adopt HMO-mHealth over face-to-face meetings. A systematic literature review [53] and narrative synthesis showed that the effects of marital status on mHealth uptake are inconsistent across studies. Some studies suggest that marital status has no significant effect on the use of mHealth services, while others propose possible differences in the intention to use depending on marital status [53]. In our opinion, addressing family health needs and marital responsibilities can increase interest in technologies that improve health monitoring and save time. Since there is a gender dynamic in most marital relationships regarding domestic responsibilities and health care decisions, this is likely to have an impact on the adoption of HMO-mHealth.

This study raises an interesting question: for whom are HMO-mHealth services really intended for? In this study, it was found that patients with chronic diseases do not prefer to use HMO-mHealth services. Conversely, people who describe their subjective health as good are more likely to use HMO-mHealth services. It is generally assumed that patients with chronic diseases will adopt more mHealth services [54], as mHealth is seen as part of medical practice and a factor that can support patients with chronic diseases [37]. There is evidence in the literature that subjective health condition is a

XSL•FO
RenderX

factor that influences the use of digital means to access health care services [55]. However, studies show that the presence of a chronic disease does not predict demand for medical services [54,56-58]. It can be assumed that the support provided by the family of a patient with chronic diseases will reduce the need for HMO-mHealth services [58].

Limitations

Our study has several limitations. First, due to the cross-sectional nature of our study, it is not possible to establish causation or demonstrate cause-and-effect relationships. Second, the data were self-reported. The researchers were not able to assess whether the questionnaire was completed during work or leisure time or whether health limitations influenced the completion of the questionnaire. Third, the services in the HMO-mHealth apps are similar, but the interaction between the patient and the health care provider in each app may be different. It is possible that a better response from the health care provider for certain services in the HMO-mHealth apps could influence the information collected in the questionnaire. Fourth, health literacy was not examined in this study. It is likely that the level of health literacy may also influence patients' interactions with HMO-mHealth services.

Conclusions and Future Directions

HMO-mHealth proves to be a robust and efficient tool for health care service delivery when compared to face-to-face health care interaction. However, barriers that affect vulnerable populations in HMO-mHealth adoption still exist. HMO-mHealth services will not be able to completely replace face-to-face interactions with health care providers and will be a complementary tool for face-to-face meetings with therapists. The utilization of health applications within health services constitutes an increasingly substantial component of the communication interface between the health care system and its patients, accompanied by a rise in the array of services offered through HMO-mHealth. On the one hand, this study's outcomes furnish insights for policy makers engaged in the development of HMO-mHealth services, facilitating the formulation of culturally sensitive HMO-mHealth services. On the other hand, there exists a compelling necessity for policy makers to institute a comprehensive training regimen aimed at equipping patients with the requisite skills for utilizing HMO-mHealth. This is particularly imperative for aiding demographic segments such as the older adults, chronically ill patients, and societal minorities, who are in greater need of these services. Health care providers need to develop intervention plans for accessibility, adaptation, and implementation of HMO-mHealth in this population. Such adaptations are essential for mitigating health care inequalities in Western societies. To provide an optimal response to these populations, future studies should focus on identifying the barriers that affect the utilization of HMO-mHealth in these groups. Future studies should also use longitudinal designs to better establish cause-and-effect relationships between variables; assess health literacy, which is a potential factor in the adoption of HMO-mHealth; and include comparisons with mHealth adoption in other countries. These future studies could significantly enhance our understanding of the cultural and systemic variations.

Acknowledgments

The authors would like to thank Ariel Nahmias and Danit Menachem (both are PhD candidates) for their help in mapping the health maintenance organization–mobile health service health care provider apps and creating the questionnaire. We would also like to thank Yifat Narinsky of Studio N for her help in creating the infographics in this paper.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request, subject to approval by the ethics committee of Ariel University.

Authors' Contributions

AZ contributed to data curation, investigation, methodology, project administration, visualization, and original draft preparation. OR contributed to the formal analysis, methodology, visualization, and supervision. MZ and OZ contributed to the investigation, methodology, and resources. All authors reviewed and commented on the subsequent drafts of this paper and approved the final version of this paper for publication. No generative artificial intelligence was used in any portion of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data.
[XLSX File (Microsoft Excel File), 16 KB - jmir_v26i1e55350_app1.xlsx]

References

XSL•FO
RenderX

1. Kaihlanen A, Virtanen L, Buchert U, Safarov N, Valkonen P, Hietapakka L, et al. Towards digital health equity - a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188 [FREE Full text] [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
2. Li J, Zhang C, Li X, Zhang C. Patients' emotional bonding with mHealth apps: An attachment perspective on patients' use of MHealth applications. International Journal of Information Management 2020 Apr;51:102054. [doi: 10.1016/j.ijinfomgt.2019.102054]
3. Kayyali R, Peletidi A, Ismail M, Hashim Z, Bandeira P, Bonnah J. Awareness and use of mHealth apps: A study from England. Pharmacy (Basel) 2017 Jun 14;5(2):33 [FREE Full text] [doi: 10.3390/pharmacy5020033] [Medline: 28970445]
4. Shahrabani S, Mizrachi Y. Factors affecting compliance with use of online healthcare services among adults in Israel. Isr J Health Policy Res 2016;5:15 [FREE Full text] [doi: 10.1186/s13584-016-0073-8] [Medline: 27307983]
5. Saghafian S, Murphy SA. Innovative health care delivery: the scientific and regulatory challenges in designing mHealth interventions. NAM Perspect 2021;2021:10.31478/202108b [FREE Full text] [doi: 10.31478/202108b] [Medline: 34611601]
6. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019 Jul 10;7(7):e13817 [FREE Full text] [doi: 10.2196/13817] [Medline: 31293246]
7. Alam MZ, Hoque MR, Hu W, Barua Z. Factors influencing the adoption of mHealth services in a developing country: A patient-centric study. International Journal of Information Management 2020 Feb;50:128-143. [doi: 10.1016/j.ijinfomgt.2019.04.016]
8. Schiavone F. User Innovation in Health Care: How Patients and Caregivers React Creatively to Illness. Switzerland: Springer Nature; May 06, 2020:229.
9. Zigdon A, Zigdon T, Moran DS. Attitudes of nurses towards searching online for medical information for personal health needs: cross-sectional questionnaire study. J Med Internet Res 2020 Mar 16;22(3):e16133 [FREE Full text] [doi: 10.2196/16133] [Medline: 32175910]
10. Giunti G, Guisado Fernández E, Dorronzoro Zubiete E, Rivera Romero O. Supply and demand in mHealth apps for persons with multiple sclerosis: systematic search in app stores and scoping literature review. JMIR Mhealth Uhealth 2018 May 23;6(5):e10512 [FREE Full text] [doi: 10.2196/10512] [Medline: 29792295]
11. Lin TTC, Bautista JR. Understanding the relationships between mHealth apps' characteristics, trialability, and mHealth literacy. J Health Commun 2017 Apr;22(4):346-354. [doi: 10.1080/10810730.2017.1296508] [Medline: 28323546]
12. Drincic A, Prahalad P, Greenwood D, Klonoff DC. Evidence-based mobile medical applications in diabetes. Endocrinol Metab Clin North Am 2016 Dec;45(4):943-965 [FREE Full text] [doi: 10.1016/j.ecl.2016.06.001] [Medline: 27823614]
13. Pfaeffli Dale L, Dobson R, Whittaker R, Maddison R. The effectiveness of mobile-health behaviour change interventions for cardiovascular disease self-management: A systematic review. Eur J Prev Cardiol 2016 May;23(8):801-817. [doi: 10.1177/2047487315613462] [Medline: 26490093]
14. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? a secondary data analytics approach. J Med Internet Res 2017 Apr 19;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
15. Wattanapisit A, Teo CH, Wattanapisit S, Teoh E, Woo WJ, Ng CJ. Can mobile health apps replace GPs? A scoping review of comparisons between mobile apps and GP tasks. BMC Med Inform Decis Mak 2020 Jan 06;20(1):5 [FREE Full text] [doi: 10.1186/s12911-019-1016-4] [Medline: 31906985]
16. Catan G, Espanha R, Mendes RV, Toren O, Chinitz D. Health information technology implementation - impacts and policy considerations: a comparison between Israel and Portugal. Isr J Health Policy Res 2015;4:41 [FREE Full text] [doi: 10.1186/s13584-015-0040-9] [Medline: 26269740]
17. About the NHS app. NHS. URL: https://www.nhs.uk/nhs-app/about-the-nhs-app/ [accessed 2024-09-06]
18. Driving the digital transformation of Germany's healthcare system for the good of patients. Federal Ministry of Health. URL: https://www.bundesgesundheitsministerium.de/en/digital-healthcare-act [accessed 2024-09-06]
19. National Health Insurance Law, Israel book of laws, no 1469. State of Israel. URL: https://fs.knesset.gov.il/ /13/law/13_lsr_211132.PDF [accessed 2024-09-06]
20. Maccabi health services app. Maccabi Health Services. URL: https://www.maccabi4u.co.il/5271-he/Maccabi.aspx?TabId=23127_5467 [accessed 2023-11-02]
21. Leumit health services app. Leumit Health Services. URL: https://www.leumit.co.il/heb/Useful/App [accessed 2023-11-02]
22. Kupat Holim Meuhedet app. Kupat Holim Meuhedet. URL: https://play.google.com/store/apps/details?id=org.meuhedet.android&hl=he&gl=US [accessed 2023-11-02]
23. Clalit health service app. Clalit Health Service. URL: https://www.clalit.co.il/he/info/services/Pages/new_app.aspx [accessed 2023-11-02]
24. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
25. Camacho-Rivera M, Islam JY, Rivera A, Vidot DC. Attitudes toward using COVID-19 mHealth tools among adults with chronic health conditions: secondary data analysis of the COVID-19 impact survey. JMIR Mhealth Uhealth 2020 Dec 17;8(12):e24693 [FREE Full text] [doi: 10.2196/24693] [Medline: 33301415]

XSL•FO
RenderX

26. Special report: The State of Israel's coping with the COVID-19 pandemic. State of Israel. 2021. URL: https://www.mevaker.gov.il/sites/DigitalLibrary/Documents/2021/COVID-19/2021-COVID-19-001-EN.pdf [accessed 2023-11-26]
27. Reges O, Feldhamer I, Wolff Sagy Y, Lavie G. Factors associated with using telemedicine in the primary care clinics during the COVID-19 pandemic in Israel. Int J Environ Res Public Health 2022 Oct 14;19(20):13207 [FREE Full text] [doi: 10.3390/ijerph192013207] [Medline: 36293788]
28. Yee TS, Seong LC, Chin WS. Patient's intention to use mobile health app. J Manag Res 2019 May 28;11(3):18. [doi: 10.5296/jmr.v11i3.14776]
29. Fu S, Chen X, Zheng H, Ou M. Understanding health information literacy of mHealth app users from digital wellbeing perspective: evidence from regression analysis and fsQCA. Library & Information Science Research 2021 Jul;43(3):101108. [doi: 10.1016/j.lisr.2021.101108]
30. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52. [doi: 10.2196/jmir.3951] [Medline: 25803266]
31. Kim H, Xie B. Health literacy and internet‐ and mobile app‐based health services: A systematic review of the literature. Proc Assoc Info Sci Tech 2016 Feb 24;52(1):1-4. [doi: 10.1002/pra2.2015.145052010075]
32. Tully L, Sorensen J, O'Malley G. Pediatric weight management through mHealth compared to face-to-face care: cost analysis of a randomized control trial. JMIR Mhealth Uhealth 2021 Sep 14;9(9):e31621. [doi: 10.2196/31621] [Medline: 34519665]
33. Israeli Central Bureau of Statistics. URL: https://www.cbs.gov.il/en/Pages/default.aspx [accessed 2024-03-18]
34. Abramson JH. WINPEPI updated: computer programs for epidemiologists, and their teaching potential. Epidemiol Perspect Innov 2011 Feb 02;8(1):1 [FREE Full text] [doi: 10.1186/1742-5573-8-1] [Medline: 21288353]
35. Roth-Cohen O, Levy S, Zigdon A. The mediated role of credibility on information sources and patient awareness toward patient rights. IJERPH 2021 Aug 15;18(16):8628 [FREE Full text] [doi: 10.3390/ijerph18168628] [Medline: 34444377]
36. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
37. Drake C, Cannady M, Howley K, Shea C, Snyderman R. An evaluation of mHealth adoption and health self-management in emerging adulthood. AMIA Annu Symp Proc 2019;2019:1021-1030 [FREE Full text] [Medline: 32308899]
38. van Elburg FRT, Klaver NS, Nieboer AP, Askari M. Gender differences regarding intention to use mHealth applications in the Dutch elderly population: a cross-sectional study. BMC Geriatr 2022 May 24;22(1):449 [FREE Full text] [doi: 10.1186/s12877-022-03130-3] [Medline: 35610577]
39. Calegari LP, Tortorella GL, Fettermann DC. Getting connected to mHealth technologies through a meta-analysis. IJERPH 2023 Feb 28;20(5):4369 [FREE Full text] [doi: 10.3390/ijerph20054369] [Medline: 36901379]
40. Nunes A, Limpo T, Castro SL. Acceptance of mobile health applications: examining key determinants and moderators. Front Psychol 2019 Dec 10;10:2791 [FREE Full text] [doi: 10.3389/fpsyg.2019.02791] [Medline: 31920836]
41. Srinivasulu S, Manze MG, Jones HE. "I totally didn't need to be there in person": New York women's preferences for telehealth consultations for sexual and reproductive healthcare in primary care. Fam Pract 2023 Mar 28;40(2):402-406. [doi: 10.1093/fampra/cmac102] [Medline: 36124892]
42. Abelson JS, Kaufman E, Symer M, Peters A, Charlson M, Yeo H. Barriers and benefits to using mobile health technology after operation: A qualitative study. Surgery 2017 Sep;162(3):605-611. [doi: 10.1016/j.surg.2017.05.007] [Medline: 28651777]
43. Piggott T, Duran A, Jalnapurkar I, Kimm T, Linscheid S, Allen M. Sex differences in anxiety disorders. In: The Oxford Handbook of Evolutionary Psychology and Behavioral Endocrinology. Oxford, UK: Oxford University Press; 2019:405-455.
44. Fernandez Puente AC, Sanchez-Sanchez N. How do gender based disparities affect women's self-rated health and anxiety in the European Union? Journal of Gender Studies 2021 Mar 15;30(7):786-796. [doi: 10.1080/09589236.2021.1899904]
45. Rajak M, Shaw K. An extension of technology acceptance model for mHealth user adoption. Technology Soc 2021 Nov;67:101800. [doi: 10.1016/j.techsoc.2021.101800]
46. Jiang J, Zhu Q, Zheng Y, Zhu Y, Li Y, Huo Y. Perceptions and acceptance of mHealth in patients with cardiovascular diseases: a cross-sectional study. JMIR Mhealth Uhealth 2019 Feb 04;7(2):e10117 [FREE Full text] [doi: 10.2196/10117] [Medline: 30714942]
47. Armaou M, Araviaki E, Musikanski L. eHealth and mHealth interventions for ethnic minority and historically underserved populations in developed countries: an umbrella review. Int J Com WB 2019 Dec 30;3(2):193-221. [doi: 10.1007/s42413-019-00055-5]
48. Schrauben SJ, Appel L, Rivera E, Lora CM, Lash JP, Chen J, CRIC Study Investigators. Mobile health (mHealth) technology: assessment of availability, acceptability, and use in CKD. Am J Kidney Dis 2021 Jun;77(6):941-950.e1 [FREE Full text] [doi: 10.1053/j.ajkd.2020.10.013] [Medline: 33309860]
49. Orlandi LB, Febo V, Perdichizzi S. The role of religiosity in product and technology acceptance: evidence from COVID-19 vaccines. Technol Forecast Soc Change 2022 Sep 12:122032 [FREE Full text] [doi: 10.1016/j.techfore.2022.122032] [Medline: 36117496]

XSL•FO
RenderX

50. Alsswey A, Al-Samarraie H. Elderly users' acceptance of mHealth user interface (UI) design-based culture: the moderator role of age. J Multimodal User Interfaces 2019 Jul 20;14(1):49-59. [doi: 10.1007/s12193-019-00307-w]
51. Dwivedi YK, Shareef MA, Simintiras AC, Lal B, Weerakkody V. A generalised adoption model for services: A cross-country comparison of mobile health (m-health). Government Information Quarterly 2016 Jan;33(1):174-187. [doi: 10.1016/j.giq.2015.06.003]
52. Radu I, Scheermesser M, Spiess MR, Schulze C, Händler-Schuster D, Pehlke-Milde J. Digital health for migrants, ethnic and cultural minorities and the role of participatory development: a scoping review. Int J Environ Res Public Health 2023 Oct 23;20(20):6962 [FREE Full text] [doi: 10.3390/ijerph20206962] [Medline: 37887700]
53. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical factors affecting patients' adoption of mobile health tools: systematic literature review and narrative synthesis. JMIR Mhealth Uhealth 2022 May 05;10(5):e36284 [FREE Full text] [doi: 10.2196/36284] [Medline: 35318189]
54. Stampe K, Kishik S, Müller SD. Mobile health in chronic disease management and patient empowerment: exploratory qualitative investigation into patient-physician consultations. J Med Internet Res 2021 Jun 15;23(6):e26991 [FREE Full text] [doi: 10.2196/26991] [Medline: 34128817]
55. Reges O, Vilchinsky N, Leibowitz M, Manor D, Mosseri M, Kark JD. Perceptions of cause of illness in acute myocardial infarction patients: a longitudinal study. Patient Educ Couns 2011 Nov;85(2):e155-e161. [doi: 10.1016/j.pec.2010.12.022] [Medline: 21310582]
56. Breckner A, Litke N, Göbl L, Wiezorreck L, Miksch A, Szecsenyi J, et al. Effects and processes of an mHealth intervention for the management of chronic diseases: prospective observational study. JMIR Form Res 2022 Aug 25;6(8):e34786 [FREE Full text] [doi: 10.2196/34786] [Medline: 36006666]
57. Mendez KJW, Budhathoki C, Labrique AB, Sadak T, Tanner EK, Han HR. Factors associated with intention to adopt mHealth apps among dementia caregivers with a chronic condition: cross-sectional, correlational study. JMIR Mhealth Uhealth 2021 Aug 31;9(8):e27926 [FREE Full text] [doi: 10.2196/27926] [Medline: 34463637]
58. Shaked O, Korn L, Shapiro Y, Zwilling M, Zigdon A. Medical and social factors influencing the utilization of healthcare services among older adults in Israel during the COVID-19 lockdown. Front Public Health 2023;11:1218507 [FREE Full text] [doi: 10.3389/fpubh.2023.1218507] [Medline: 37829095]

Abbreviations

HMO: health maintenance organization
mHealth: mobile health
OR: odds ratio
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by T de Azevedo Cardoso; submitted 10.12.23; peer-reviewed by TAR Sure, D Singh; comments to author 13.02.24; revised version received 29.03.24; accepted 31.07.24; published 30.09.24.

Please cite as:
Zigdon A, Zwilling M, Zigdon O, Reges O
Health Maintenance Organization–mHealth Versus Face-to-Face Interaction for Health Care in Israel: Cross-Sectional Web-Based Survey Study
J Med Internet Res 2024;26:e55350
URL: https://www.jmir.org/2024/1/e55350
doi:10.2196/55350
PMID:

©Avi Zigdon, Moti Zwilling, Ofek Zigdon, Orna Reges. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Challenges and Alternatives to Evaluation Methods and Regulation Approaches for Medical Apps as Mobile Medical Devices: International and Multidisciplinary Focus Group Discussion

Laura Maaß[1,2,3], PhD; Robert Hrynyschyn[2,4], MSc; Martin Lange[5], Prof Dr; Alexandra Löwe[5], Dipl; Kathrin Burdenski[6], MA; Kaley Butten[7], PhD; Sebastian Vorberg[8,9], LLM; Mariam Hachem[10,11], MPH; Aldo Gorga[12], MD; Vittorio Grieco[13], MD; Vincenzo Restivo[14], Prof Dr; Giuseppe Vella[15], MD; Marlien Varnfield[7], PhD; Felix Holl[2,16], PhD

[1]University of Bremen, SOCIUM - Research Center on Inequality and Social Policy, Department of Health, Long Term Care and Pensions, Bremen, Germany

[2]Leibniz ScienceCampus Digital Public Health, Bremen, Germany

[3]Digital Health Section, European Public Health Association (EUPHA), Utrecht, Netherlands

[4]Charite´ – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Health and Nursing Science, Institute of Health and Nursing Science, Berlin, Germany

[5]Department of Fitness & Health, IST University of Applied Sciences, Düsseldorf, Germany

[6]School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, United Kingdom

[7]Australian eHealth Research Centre (CSIRO), Brisbane, Australia

[8]QuR.digital - Vorberg.law, Hamburg, Germany

[9]Bundesverband Internetmedizin eV, Hamburg, Germany

[10]Department of Medicine, Austin Health, Faculty of Dentistry, Medicine and Health Sciences, University of Melbourne, Melbourne, Australia

[11]Australian Centre for Accelerating Diabetes Innovations, Faculty of Dentistry, Medicine and Health Sciences, University of Melbourne, Melbourne, Australia

[12]Department of Sciences of Public Health and Pediatrics, University of Turin, Turin, Italy

[13]Department of Medical, Surgical Sciences and Advanced Technologies “GF Ingrassia”, University of Catania, Catania, Italy

[14]School of Medicine, University of Enna, Enna, Italy

[15]Department of Health Promotion, Maternal and Infant Care, Internal Medicine and Medical Specialties (PROMISE) “G. D’Alessandro”, University of Palermo, Palermo, Italy

[16]DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany

Corresponding Author:
Laura Maaß, PhD
University of Bremen, SOCIUM - Research Center on Inequality and Social Policy
Department of Health, Long Term Care and Pensions
Mary-Somerville-Straße 3
Bremen, 28359
Germany
Phone: 49 1719202704
Fax: 49 421 218 586
Email: laura.maass@uni-bremen.de

Abstract

Background: The rapid proliferation of medical apps has transformed the health care landscape by giving patients and health care providers unprecedented access to personalized health information and services. However, concerns regarding the effectiveness and safety of medical apps have raised questions regarding the efficacy of randomized controlled trials (RCTs) in the evaluation of such apps and as a requirement for their regulation as mobile medical devices.

Objective: This study aims to address this issue by investigating alternative methods, apart from RCTs, for evaluating and regulating medical apps.

Methods: Using a qualitative approach, a focus group study with 46 international and multidisciplinary public health experts was conducted at the 17th World Congress on Public Health in May 2023 in Rome, Italy. The group was split into 3 subgroups

XSL•FO
RenderX

to gather in-depth insights into alternative approaches for evaluating and regulating medical apps. We conducted a policy analysis on the current regulation of medical apps as mobile medical devices for the 4 most represented countries in the workshop: Italy, Germany, Canada, and Australia. We developed a logic model that combines the evaluation and regulation domains on the basis of these findings.

Results: The focus group discussions explored the strengths and limitations of the current evaluation and regulation methods and identified potential alternatives that could enhance the quality and safety of medical apps. Although RCTs were only explicitly mentioned in the German regulatory system as one of many options, an analysis of chosen evaluation methods for German apps on prescription pointed toward a "scientific reflex" where RCTs are always the chosen evaluation method. However, this method has substantial limitations when used to evaluate digital interventions such as medical apps. Comparable results were observed during the focus group discussions, where participants expressed similar experiences with their own evaluation approaches. In addition, the participants highlighted numerous alternatives to RCTs. These alternatives can be used at different points during the life cycle of a digital intervention to assess its efficacy and potential harm to users.

Conclusions: It is crucial to recognize that unlike analog tools, digital interventions constantly evolve, posing challenges to inflexible evaluation methods such as RCTs. Potential risks include high dropout rates, decreased adherence, and nonsignificant results. However, existing regulations do not explicitly advocate for other evaluation methodologies. Our research highlighted the necessity of overcoming the gap between regulatory demands to demonstrate safety and efficacy of medical apps and evolving scientific practices, ensuring that digital health innovation is evaluated and regulated in a way that considers the unique characteristics of mobile medical devices.

(J Med Internet Res 2024;26:e54814) doi:10.2196/54814

KEYWORDS

medical apps; mobile medical devices; evaluation methods; mobile medical device regulation; focus group study; alternative approaches; logic model; mobile phone

Introduction

Background

Medical apps have emerged as a subgroup of health apps. They are considered as powerful tools with the potential to revolutionize health care by supporting, managing, and enhancing individual and population health [1]. These software programs can run on mobile devices and offer new possibilities for processing health-related data [2,3]. According to the General Data Protection Regulation, data processing includes any operation performed on personal (health) data, including collecting, organizing, storing, adapting, visualizing, retrieving, disseminating, restricting, or erasing the data [4,5]. Unlike most types of health apps, medical apps do not primarily target healthy or health-conscious individuals but patients, health professionals, or family caregivers. This stems from their primary use as treatment supplements or medical devices in clinical contexts such as for diagnostics, treatment, or rehabilitation. Figure 1 summarizes the distinction between medical apps as a subgroup and other health apps (such as wellness apps) from a legal point of view [6].

Medical apps promise personalized health care, improving access to medical resources and fostering active patient participation [7]. However, this thematic orientation of medical apps introduces a range of potential risks. Health and patient information may be inaccurate or flawed [8]; sensitive data can be disclosed to the wrong recipients or lost [9]; applications frequently target broad user adoption via the internet, increasing the risk of rapid dissemination of potential dangers. Due to the sensitive nature of health data, medical apps must adhere to rigorous data security standards. As many apps are used across borders, different jurisdictions and regulatory standards may need to be met [10,11].

While considering covering the costs of these products through various entities, such as through health insurance providers, demonstrating their clinical and medical benefits becomes essential for them to be regulated as mobile medical devices [12-15]. This advantage must be substantiated and evaluated to align costs adequately within the insurance community [16]. In addition, high data protection requirements, such as General Data Protection Regulation, impact the possibilities of data collection, analysis, and, therefore, the evaluation of medical apps [17-19]. Software is an increasingly critical area of health care product development; there is a need to establish a standard and converged understanding of clinical evaluation globally [20].

XSL•FO
RenderX

Figure 1. Medical apps as a subgroup of health apps; adapted from Maaß et al [6].

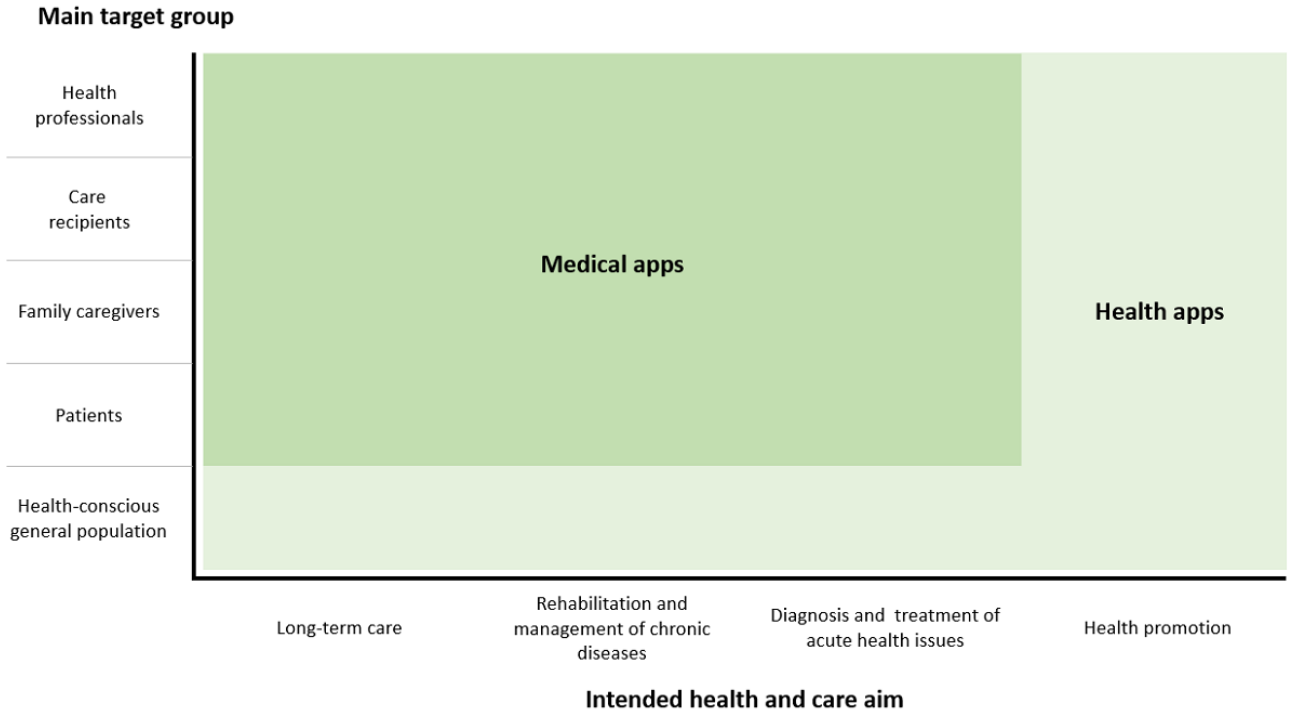

Challenges in the Evaluation of Medical Apps

With the increase in the use of medical apps in health care, it is crucial to validate their effectiveness. Randomized controlled trials (RCTs) are commonly considered the gold standard for establishing the cause-and-effect relationships because of the rigorous methodology used for such trials [21-23]. However, there is an apparent disparity between the design of RCTs and the need to evaluate medical apps: RCTs are structured and rigid, whereas medical apps are versatile, updated frequently, and depend on specific contexts [24]. A critical challenge arises from the discrepancy between the thoroughness of evaluations and the pace of generating evidence, as depicted in Figure 2 [25]. This misalignment underscores the need for developing more adaptable evaluation strategies in the rapidly evolving field of medical technology.

Figure 2. The balancing act between increasing evidence levels and the time needed to generate them; adapted from Guo et al [26].

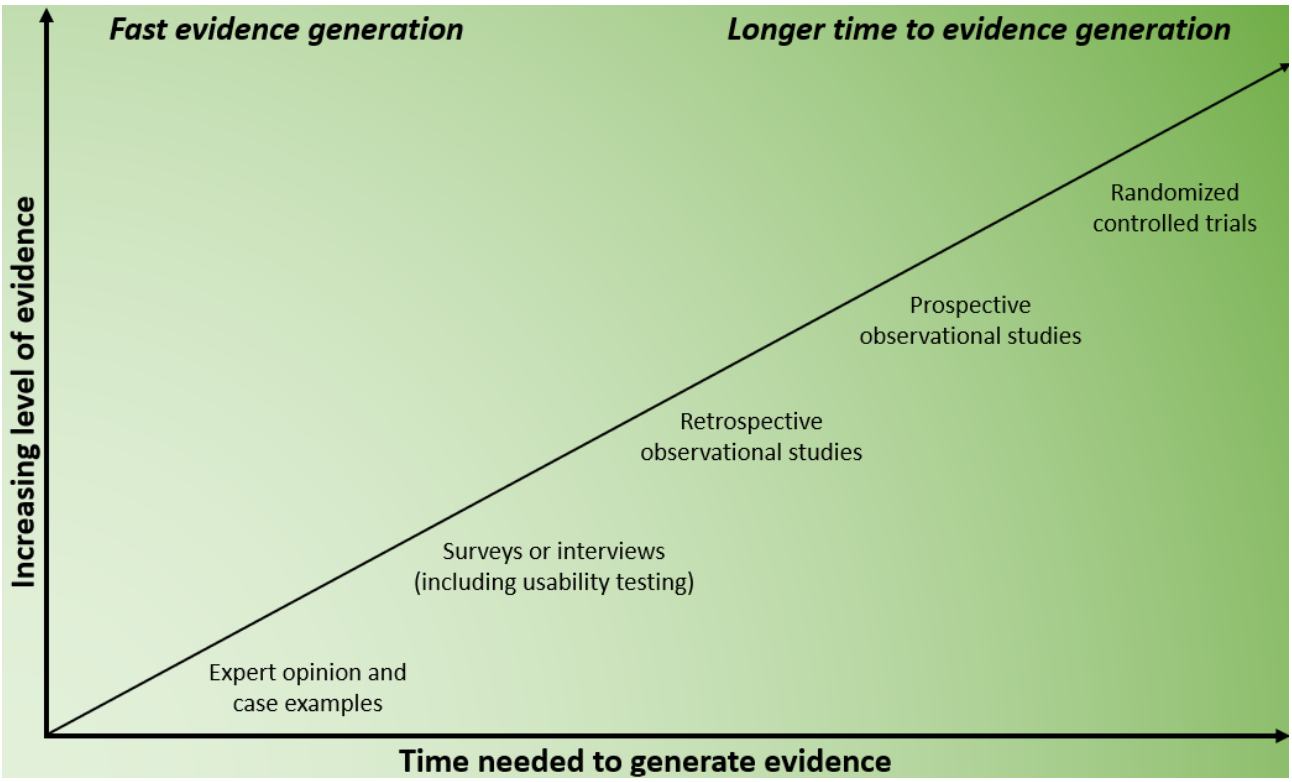

XSL•FO
RenderX

Furthermore, traditional trials such as RCTs do not necessarily align with the unique characteristics of medical apps [27,28]. The evaluations of medical apps is made on the basis of short-term studies [29-31], limited patient samples [32], feasibility trials, and user preference surveys [33]. While these methods can swiftly provide valuable insights during the development phase, they only offer a narrow view of the true efficacy of the apps [34].

Another challenge is that in contrast to the intergroup design of RCTs, medical apps are often customized to individual needs. Continuous updates and tweaks to medical apps aimed toward improving their performance can sometimes shift their fundamental functionality. It is essential to recognize that these updates might influence patient outcomes and call for continuous reassessment [24]. Given users' diverse health needs, preferences, and technological expertise, a universal evaluation method seems impractical. Moreover, blinding participants and researchers, a cornerstone of RCTs to reduce bias, can be difficult with medical apps. This is because users are often aware of their app use, which will lead to developers struggling to create "placebo" apps with identical functionalities.

In addition, app evaluation is multifaceted, demanding attention to factors beyond the health metrics typically assessed in RCTs [34]. Aspects, such as user engagement and sustained use, are vital for long-lasting behavioral shifts [35-37]. During the developmental phase of a medical app, alternative study designs, including feasibility studies and user preference surveys, might be more suitable [33], even if they offer limited efficacy data [32]. They can promptly provide essential evidence regarding developmental and application areas. The multifaceted demands of medical app evaluation call for innovative evaluation methods that form the foundation of medical app regulations.

The Complexity and Struggles of Overregulation

Regulatory and quality standards that are legally mandated and government-reviewed ensure medical the safety, efficacy, and data privacy of medical apps [26]. Historically, pharmaceuticals and medical devices have faced increased regulation only after substantial scandals and incidents have emerged. "Fueled" by these scandals, medical app regulation seeks to prevent scandals and incidents through rigorous guidelines [38-40]. However, there is a risk that anticipatory regulation may overemphasize securing against hypothetical risks [41]. It is crucial to weigh hypothetical risks against the effort required for regulatory compliance and the expected benefits of maintaining a balanced regulatory approach [42]. Relevant criteria for regulating medical apps are displayed in Textbox 1 (adapted from Torous et al [43] and Meskó and deBronkart [44]). Regulatory solutions should adopt the mildest measures possible to minimize resistance to technological advancement. Consequently, it is essential to incorporate the potential severity of harm into regulatory assessments and, if necessary, accept manageable errors as learning opportunities. Postmarket surveillance is a critical component of quality control and assurance in the regulation of innovative medical developments.

Textbox 1. Relevant criteria for regulating mobile apps as mobile medical devices.

Criteria

- Effectiveness and safety: Regulatory frameworks should be adapted to the specific characteristics of digital health applications (DiGAs), necessitating clinical studies and testing procedures tailored for these apps.
- Data privacy and security: Regulation must ensure that the applications implement appropriate data protection and security measures.
- Interoperability: DiGAs should be integrable into existing health care systems to ensure seamless care delivery.
- Provider qualifications: Regulations should mandate that qualified professionals develop and operate applications to ensure patient safety and quality.
- Monitoring and traceability: Regulatory mechanisms should enable the ongoing monitoring and traceability of DiGAs to identify and address adverse events promptly.
- Patient involvement: Patients should be actively engaged in the development and evaluation of DiGAs to ensure that their needs and expectations are met.

Need for Alternative Evaluation Approaches and Regulation

Although the results are prone to reduction and distortion, regulatory requirements homogenize research conditions. At the same time, medical apps are, by nature, complex interventions due to their structure, a large number of interactions, and various intervention components. The efficacy of medical apps depends on context and leads to distinct effects on individuals. This requires a broader evaluation framework that does not focus solely on specific intervention outcomes but combines different perspectives from feasibility testing to intervention, implementation, and impact evaluation [26,45,46]. Current regulatory approaches to evaluating medical apps do not appropriately reflect the complex nature of such apps.

Regulation of Medical Apps in Selected Countries

To illustrate similarities and differences, 4 countries have been selected to conduct case studies about the existing regulation methods for the use of medical apps as mobile medical devices. The analysis for Germany, Italy, Australia, and Canada followed an adapted version of the policy benchmarking framework by Essén et al [38] (Table 1).

XSL•FO
RenderX

Table 1. Overview of medical app regulation for selected countries.

	Germany	Italy	Australia	Canada
Actors developing national policy regulation	Health Innovation Hub, an external, interdisciplinary expert think tank to the Federal Ministry of Health and the Federal Institute for Drugs and Medical Devices	Ministry of Health overseeing conformity following the European Medical Device Regulation and classification of SaMD[a]	TGA[b]	Health Canada (federal government agency)
Intended use of the framework	It helps in determining the eligibility of an app for reimbursement through the public insurance scheme either permanently or on a preliminary basis.	—[c]	It provides national guidance and a helpful reference tool for app developers working on mobile health apps for release in Australia. It provides a solid basis for further research and analysis.	It helps in determining whether the app meets the legal definition of a SaMD.
Key regulations underpinning the policy framework	DiGAs[d] under the Digital Care Act	—	The Privacy Act 1988, overseen by the Office of the Australian Information Commissioner; the Competition and Consumer Act 2010 administered by the Australian Competition and Consumer Commission	Risk classification as an SaMD
Risk classification framework for medical apps	European Medical Device Regulation (Risk classes I, IIa, IIb and, III)	European Medical Device Regulation (risk classes I, IIa, IIb, and III)	TGA (Medical Devices) Regulation 2002 (risk classes I, IIa, IIb, and III)	Health Canada Medical Device Classification (risk classes I, II, III, and IV)
Reimbursement approval policy regulation developed?	Applications get approved through the DiGA fast-track procedure. Approved DiGAs are reimbursed through health insurance	—	—	—
End-user interface to clinical practice and patients, which lists approved medical apps	DiGA directory	—	—	None. However, Health Canada does engage with the Canadian Agency for Drugs and Technologies to provide evidence and publicly available health technology assessments.
Are RCTs explicitly mentioned in the regulation?	Mentioned as an example of a methodology designed to demonstrate a positive care effect	—	—	—

[a]SaMD: software as a medical device.

[b]TGA: Therapeutic Goods Administration.

[c]Not applicable.

[d]DiGAs: digital health applications.

Regulations in Germany

Since December 2019, digital health applications (DiGAs) have become eligible for reimbursement through statutory health insurance in Germany. This mechanism is included in the Digital Care Act (DVG) [47,48]. Reimbursable DiGAs are listed in a directory administered by the Federal Institute for Drugs and Medical Devices, DiGA directory. Applications can be added to the directory through a fast-track procedure [49]. Applications can either be permanently included in the list if evidence of a positive care effect is provided at the time of application or a provisional reimbursability is granted when apps meet a set of essential requirements with evidence of the positive care effect is still missing. For permanent inclusion, evidence of the positive care effect must be provided within 12 months. Otherwise, the app will be removed from the list. The term "positive care effect" introduced by the DVG can either be a "medical benefit" or "patient-relevant improvement of structure and processes." For instance, the app can help improve health literacy and improve coordination of treatment processes or can reduce therapy costs. Currently, most DiGAs provide evidence for the positive care effect using data from RCTs [50]. As of October 13, 2023, the DiGA directory included 55 applications of which 23 were provisionally included, 26 were permanently included, and 6 were removed. Although the DVG does not explicitly call for an RCT [51], all the apps were assessed through an RCT. A limitation of the DVG is the restriction on low-risk devices (risk classes I and IIa) under the Medical Devices Act.

XSL•FO
RenderX

All applications outside the Medical Devices Act, such as preventive applications, and high-risk medical devices do not fall under the DVG. The fast-track procedure applies only to applications classified and certified as medical devices in risk classes I and IIa; therefore, many potentially helpful applications do not fall within the scope of the DVG [52].

Regulations in Italy

In Italy, medical app, provided they meet legal requirements, can be promptly marketed and are classified as software as a medical device (SaMD), and obtain a CE marking [53,54]. On May 25, 2021, the Italian Ministry of Health issued a newsletter to offer interested stakeholders with recommendations for clinical investigations of medical devices in line with current regulations [55]. The Ministry is currently responsible for the approval process. However, there is no national framework for market access or reimbursement approval; a law for the health technology assessment of medical apps is also lacking. While an app can be sold, it cannot be prescribed or reimbursed, hindering equitable health care access. Discussions are ongoing in legislative bodies to address this issue; a Parliamentary Intergroup for Digital Health and Digital Therapies was established in May 2023 [56]. This initiative aims to develop regulations that align with those of other European countries, ideally, during the current legislative term.

Regulations in Australia

In Australia, medical devices, including software-based medical devices, are classified according to the medical device classification rules of the Therapeutic Goods (Medical Devices) Regulations 2002 [57]. This act oversees the quality, safety, efficacy, and availability of therapeutic goods in Australia. Medical devices are classified and regulated according to the level of harm they may pose to the users or patients. This includes medical device software and medical devices that incorporate software. For the manufacturer to determine the inclusion of an app in the Australian Register of Therapeutic Goods [58], a 4-tiered classification system is used. This is done to determine the minimum conformity assessment procedures or evidence requirements for comparable overseas regulators. The classification rules are applied according to the manufacturer's intended purpose and the device's functionality. Where more than 1 rule applies, the device must be classified at the highest applicable level [57,59]. However, the Australian regulatory system lacks refinement regarding emerging technologies such as artificial intelligence in medical decision-making and wearable technology. This generates problems in deciding which tools and devices must be registered and to which classification standard [60].

Regulations in Canada

Similar to Australia, within Canada, mobile health apps are classified as SaMD and regulated according to a 4-tier risk-based classification system used for medical devices. The manufacturer's intended use determines the classification of potential SaMD and corresponding regulation. It also determines whether the federal department, Health Canada, agrees with the use case [61]. The clinical and evaluation evidence required to support the use case varies depending on the tiered classification and intended use. However, the guidance on SaMD, released by Health Canada in 2019, suggests broad exclusion criteria, eliminating many health apps from requiring SaMD status, thereby leaving many apps unregulated and without evaluation [62].

Study Aim

Due to the described challenges in regulating and evaluating medical apps, this study pursued 2 central objectives. Building on the previously described examples of country-specific uses for the current regulation and evaluation practices of medical apps, this study first aimed to determine whether there are international challenges in regulation and evaluation. Second, this study aimed to gather the opinions and perspectives of digital (public) health researchers on current practices and future developments through a focus group approach. The results from both the objectives were then used to develop a logic model proposing a differentiated approach to regulating and evaluating medical apps. Therefore, this study attempts to initiate a discussion about current regulation and evaluation practices of medical apps and accelerate the adoption of innovations in health care systems.

Methods

Quality Assessment

The reporting quality of this study is assessed through COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 1).

Focus Group Discussion Workshop at the 17th World Congress on Public Health

Participants for the focus groups were recruited through a convenient sampling method. Interested attendees who are members of the 17th World Congress on Public Health (WCPH) attended the workshop in person. The study was advertised to participants through the conference program. Due to the nature of the context, information on nonparticipation was not recorded. In total, 46 WCPH delegates from 14 countries joined the workshop. Besides the moderators and workshop participants, 2 conference technical-team members were in the room to oversee the technical equipment. They did not participate in the workshop. At the beginning of the session, the participants were introduced to the general concepts of health and medical apps and the problems that RCTs pose for evaluating these types of digital health interventions.

To collect demographic data and information regarding the participants' knowledge, we used a QR code leading to a LimeSurvey questionnaire (LimeSurvey GmbH) hosted by the University of Bremen (Germany). All questions were optional. The 46 participants were equally distributed among 3 focus groups, with each group led and protocolled by one of the authors. Furthermore, 1 female and 2 male authors introduced themselves as researchers interested in evaluating digital apps and explained the scope of the workshop. They discussed the 2 questions presented in the aim section of the study. The 2 questions were not pilot tested. The 3 authors have a background in public health and medical informatics and have conducted previous research on the topic; one author is pursuing his PhD

XSL•FO
RenderX

on alternative evaluation approaches [63]; the second author has published a review on the definitions for health and medical apps [6]; the third author has conducted an assessment of evaluation methods for mobile health interventions [35]. The authors had the underlying assumption that most participants are from Italy and that the discussions will focus more on the first question due to the conference's target group of public health researchers and practitioners.

Focus groups are used in qualitative research to gain in-depth knowledge of specific issues, which, in our case, is the evaluation and regulation of medical apps. Unlike in surveys or individual interviews, a strength of focus groups is that participants can interact with other participants and the moderator rather than merely sharing their views [64]. In our case, we followed the methodology proposed by Traynor [64] with a minor adaptation to avoid audio recording the discussions due to the noise in the room. Instead, we asked 1 participant per group to assist the moderator in taking notes on a flip chart visible to the respective focus group. The discussion lasted 40 minutes, with the first half of the session focusing on question 1 and the second half on question 2. The moderators asked the 2 guiding questions and follow-up collective questions, encouraged quieter participants to express their opinions, and quietened more dominant speakers if necessary. After the workshop, all participants were invited to contribute to the project as coauthors; following this, 9 participants contributed to the project. The workshop transcripts were only provided to the 9 coauthoring participants (Multimedia Appendix 2).

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki; the protocol was approved by the Joint Ethics Committee of the Universities of Applied Sciences of Bavaria (GEHBa-202304-V-105, dated April 28, 2023). After the introductory presentations described above, the researchers handed out written participant information sheets. This was done to educate the participants on the study and to determine which data would be collected, stored, and analyzed. All data were anonymized. This means that the researchers did not collect information that would allow the identification of participants. All participants were informed that their study participation was voluntary. This means that they could withdraw from the study at any time without providing reasons. Before the data collection process, participants were informed that they would not receive any financial compensation for participating in this study. However, they were invited as coauthors if they expressed written interest. In this case, the participants' email addresses were collected on a paper sheet. These were not linked to the web survey through which their sociodemographic data were collected. Finally, all participants provided their written informed consent before data collection.

Qualitative Analysis and Data Synthesis

For the thematic analysis (TA) methodology of Braun and Clarke [65,66], data analysis and synthesis followed the conceptual and design thinking. TA aims to indicate and analyze patterns (so-called themes) in minimally organized data sets. TA was deemed the most suitable framework because it can be applied without preexisting theoretical frameworks. This makes it an ideal methodology for multidisciplinary and interdisciplinary research questions such as the current topics currently discussed in the study [65,66]. In our case, 2 authors individually clustered the workshop results inductively, without predefined themes, in 2 separate Microsoft Excel (Microsoft Corp) sheets. Both authors were provided with transcripts of all the flip charts for the focus groups. They then merged individual statements that were based on the same typology and meaning. For example, when 2 groups mentioned the same evaluation designs or barriers, these were merged together. The results were then grouped according to a shared umbrella theme. The grouping was performed separately for the evaluation and regulation of medical apps. The final themes were approved during a discourse among 3 authors, with 1 author having the ultimate decision-making power.

On the basis of the workshop findings and policy analysis, we developed a logic model to capture the complexity of evaluating and regulating medical apps. Logic models can be helpful to display the life cycle of a medical app based on overarching themes, for instance, problem, target, intervention and content, moderating and mediating factors, outcomes, and impact [67]. These models have been used to evaluate digital health interventions [68,69]. The models can also be adapted to the specific contexts. It enables the display of the interaction and feedback loops. It also distinguishes between outcomes and impacts in different domains [67]. Furthermore, adjusted logic models can provide a framework to display how complex interventions work across multiple domains in a single setting, with interlinking actions producing a range of outputs and outcomes. In our logic model, we used the example of a generic medical app to display substantial findings and assign them to specific phases of the medical app life cycle such as development, implementation, and impact.

Results

Overview

In total, 46 experts participated in the 1-hour international workshop at the 17th WCPH [70]. The participants were divided into 3 focus groups of 15 to 16 people each. The participant demographics are displayed in Table 2.

We asked the participants about their experiences and opinions regarding evaluating and regulating medical mobile apps. The TA of the contributions in the 3 groups resulted in 4 clusters for the evaluation and 3 for the regulation of medical mobile apps. These are presented in more detail in Textboxes 2 and 3.

XSL•FO
RenderX

Table 2. World Congress on Public Health workshop: participant demographics.

Criteria	Participants (N=46), n (%)
Highest qualification	
Bachelor's degree	2 (4.3)
Master's degree	13 (28.3)
Diploma	2 (4.3)
Medical doctor	18 (39.1)
PhD	11 (23.9)
Primary discipline	
Computer science	5 (10.9)
Medicine	3 (6.5)
Psychology	3 (6.5)
Public health	34 (73.9)
Sociology	1 (2.2)
Gender	
Women	26 (56.5)
Men	20 (43.5)
Country of residence	
Australia	5 (10.9)
Austria	1 (2.2)
United States	4 (8.7)
Canada	1 (2.2)
Finland	8 (17.4)
Germany	1 (2.2)
Hungary	1 (2.2)
Indonesia	17 (37)
Italy	1 (2.2)
Netherlands	2 (4.3)
New Zealand	1 (2.2)
Philippines	1 (2.2)
Portugal	1 (2.2)
Switzerland	1 (2.2)
Taiwan	1 (2.2)
Age	
Mean (SD; range)	35.1 (9.6; 25-73)
Median (IQR)	31.5 (8.3)
Years of general work experience in the primary field	
Mean (SD; range)	8.3 (8.5; 1-40)
Median (IQR)	5 (7.0)
Years of experience in developing, implementing, regulating, or evaluating health or medical apps	
Mean (SD; range)	1.8 (2,7; 0-15)
Median (IQR)	1 (2.0)

XSL•FO
RenderX

Textbox 2. Results of thematic data analysis on alternative evaluation methods.

Traditional randomized controlled trials (RCTs)

- These determine the strengths of RCTs (eg, internal validity and the strength of causal relationships).
- These determine the weaknesses of RCTs to represent reality, resource consumption, and low generalizability.
- Randomization might cause dissatisfaction and increase in dropouts.
- Effectiveness trials tend to produce too little engagement.
- The speed of app development does not align with RCT duration.

General evaluation aspects

- Different types of outcomes need to be acknowledged by evaluation goals and methods.
- Power (ie, a priori sample size and bias) and data validity (ie, level of causality and purity of data) are evaluated.
- Differentiation of traditional (eg, effectiveness and satisfaction) and implementation outcomes (eg, acceptability, costs, and feasibility) is conducted.
- Qualitative prestudy generates fast evidence.
- Levels of engagement and interaction (eg, user frequency and duration of use) are determined.
- Importance of multiple and standardized outcomes is determined.
- A suitable comparison group (ie, new, existing, or nontreatment intervention) is identified.
- Intervention effects (eg, Hawthorne) are considered.
- Outcomes for stakeholder and target groups differ and need to be considered.
- Relevance of the component's effectiveness is determined.
- Relevance of output evaluation is determined.

Alternatives for recruiting and grouping in RCTs

- Preference-based controlled trials
- Evaluation of specific implementation stages rather than considering implementation as a whole
- Pragmatic RCTs or real-world RCTs
- Ensuring data quality through volunteer participation
- Quasi-controlled trial
- Best-choice experiment after randomization
- Stepped wedge trials
- Waitlist-control-group design
- Evaluation as treated

Alternatives to RCTs

- Collecting qualitative empirical data (eg, user interviews) to understand engagement
- Improving user experience to promote engagement
- Creating a prototype evaluation checklist for researchers with less experience in app development.
- Using real-world evidence (eg, hospital data and cohort data set)
- Using shortcut evaluation if a known aspect works (eg, McCarthy Evaluation)
- Evaluating the practice or rollout phase
- Using model apps through existing, similar data sets
- Applying user analytics (eg, Google Analytics)
- Using of additional data through smartphone sensors

XSL•FO
RenderX

Textbox 3. Results of thematic data analysis on alternative regulation approaches.

Functionality of medical apps

- Regulation should be proportional to app functionality.
- No discrimination and not doing any harm need to be considered as the minimum requirements for the analysis.
- Privacy of data needs to be ensured.
- Accountability and informed consent are considered as the developer's responsibility.
- Efficacy should be measured.

Tools for regulation

- Using labels that rate various aspects of the app (eg, accessibility and health benefits)
- Standardizing regulation across geographical areas (eg, in Europe, create minimal requirements)

Barriers to regulation

- There should be no overregulation.
- Silo apps: if a chronically person needs 20 apps to manage their symptoms, this is unfeasible and expensive.
- It is equally important to consider the device as the app.

Alternative Evaluation Methods

The following general question was presented to the participants at the beginning of the focus group discussions: What could be the alternative evaluation methods for assessing the effectiveness of medical applications? According to the participants, 4 overarching themes were crystallized as pivotal (Textbox 2).

Traditional Evaluation Practice

Participants emphasized the advantages of RCTs in evaluating medical apps, acknowledging their ability to reduce bias and establish causal relationships. However, concerns were raised about participant dissatisfaction and high attrition rates due to random assignment, questioning sustainable implementation and user centricity. Incongruence between the temporal requirements of an RCT and the rapid evolution of medical apps necessitated frequent updates, prompting the consideration of alternative evaluation methods.

Universal Considerations for All Evaluation Types

The importance of engagement and user data in app evaluations was highlighted; participants advocated qualitative assessment in prestudy designs to reduce dropouts. In addition, the participants stated that standard outcome parameters were essential for effectiveness evaluation and could be supplemented by output data to demonstrate efficacy. The participants further suggested segmental app evaluations and various comparative interventions. Regardless of the study design, the importance of enhancing power and data quality through a priori sample calculations and suitable study designs was highlighted.

Alternative Approaches for Participant Recruitment and Grouping in an RCT

Alternative recruitment and grouping methods were explored to address dissatisfaction and attrition among the participants of RCTs. These included preference clinical trials, pragmatic RCTs, partially randomized patient-preference trials, and best-choice experiments. By increasing flexibility and allowing participants to choose their interventions, these approaches might reduce withdrawal, shorten study duration, reduce bias, and facilitate participant recruitment.

Alternative Research Designs to RCTs

Other research designs, such as qualitative and participatory approaches, can help to understand engagement barriers and enhance user experience. Prototype checklists were proposed for researchers unfamiliar with alternative evaluation designs. Alternative data sources, including real-world evidence, smartphone sensors, and commercially available user analytics, can complement traditional evaluation methods.

A glossary explaining the alternative evaluation methods in more detail is available in Multimedia Appendix 3.

Alternative Regulation Approaches

Overview

Participants worked on the regulative part during the second half of the focus group discussions. The guiding question was as follows: What could be alternative approaches to regulating medical applications as mobile medical devices? Three themes emerged from these discussions (Textbox 3).

Minimal Standards for Medical Apps

Participants emphasized essential minimum standards that medical apps should adhere to ensure safety and harmlessness, including data privacy, user accountability, informed consent, and prevention of biases or discrimination. Experts acknowledged that regulatory requirements should be proportional to the app's function and data processing. Central to the regulation process is prioritizing the efficacy of the medical app; the efficacy is established through rigorous evaluation procedures.

Tools for Regulation

Labels are commonly used in public health and can offer a concise way to communicate information and evaluate app

XSL•FO
RenderX

content and user friendliness. To establish a foundational standard for further country-specific development, participants stressed the need for standardized guidelines across geographic regions, exemplified by the medical device regulation in Europe.

Barriers to Regulation

The primary concern regarding the law was potential overregulation. Complex approval processes could stifle app innovations. Overregulation might lead to an abundance of specialized "silo apps." Patients managing multiple health conditions require numerous apps, hindering efficiency and user friendliness. In addition, participants recognized that the app's usability was contingent on the device, a factor beyond the scope of app regulation.

Logic Model for the Evaluation and Regulation of Medical Apps

On the basis of country-based assessments and the focus group discussions, we developed a logic model highlighting the complexity of medical app evaluation and regulation. A logic model depicts the relationship between a program's activities and its intended effects [71]. In this study, the logic model illustrates the different stages of app development (read from left to right) along with the various types of evidence and regulation required at these stages [72,73]. The model consists of 3 independent horizontal blocks (Figure 3). The first horizontal block provides an example of a generic medical app. The second horizontal block focuses on evaluation designs for the effectiveness and usability of such apps; the third contains regulatory aspects. Each block is divided into "columns" defined as the general phases of logic models (ie, problem, target, intervention & content, moderating & mediating factors, outcomes, and impact). The boxes on evaluation and regulation are additionally influenced by the conceptual, implementation, and evaluation quality. Apart from the mere analysis of outcomes (as measured through RCTs), it is essential to consider these factors while determining evaluation and regulation approaches for medical apps.

Figure 3. Logic model for evaluating and regulating medical apps through the intervention's life cycle phases.

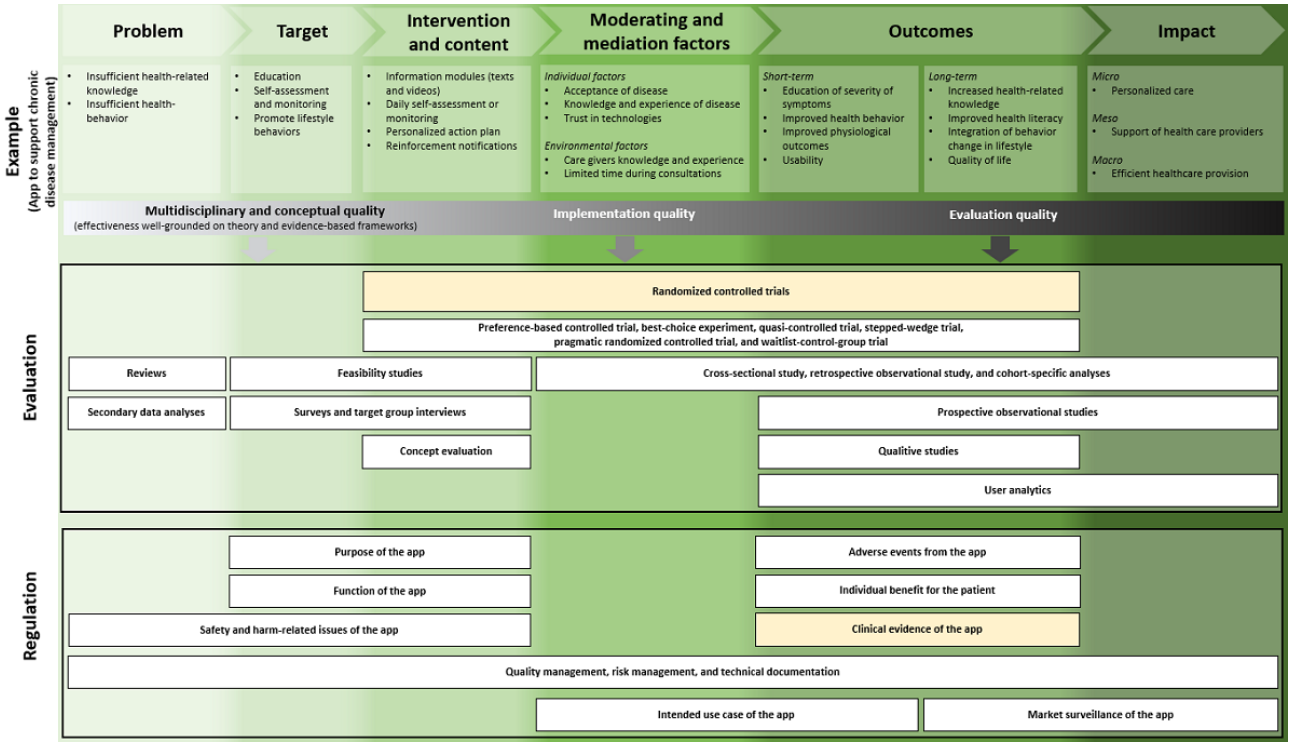

Feasibility studies can be applied to evaluate the medical apps' target and content. They can include multidisciplinary and participatory designs or theoretical models based on evidence-based frameworks. Machine-learning algorithms and app use modeling from similar data sets can provide additional feasibility analyses [74]. Clinical simulations can be run to test digital interventions in a safe, cost effective, and efficient manner [26]. Alternative trial and study designs that integrate qualitative studies and user analytics (to evaluate the outcomes of the medical app) allow for integrating multiple perspectives and offer efficient approaches to evidence generation. Similarly, the regulation of medical apps should incorporate the app's life cycle as well. An a priori regulatory concept should be preapplied during the target and content phase of the app development based on the purpose (eg, disease-related) and functionality (eg, documenting, storing, monitoring, data analyzing, or transferring data) of the app. Its potential harm concerning data protection and safety of use (eg, informed consent and appropriateness of data processing) should be best regulated even earlier, that is, during the problem-identification phase. During the outcome phase, relevant regulation is needed in case of adverse events, for patient benefits, and to obtain clinical evidence. The quality and risk management as well as technical documentation should be regulated continuously. Together with the market surveillance after the implementation of the medical app, these procedures will ensure that the application adheres to regulations and fulfills its intended benefits.

XSL•FO
RenderX

"Randomized controlled trials" and "clinical evidence of apps" are highlighted in yellow. This is because most medical apps focus on these methods and outcomes [51]. This is despite the fact that the relevant regulatory framework requires a sufficient evaluation, which does not necessarily have to be an RCT [12]. In addition, our logic model underlines that starting points for RCTs are typically placed too late in the medical app development cycle given that they do not include the two starting phases *problem* and *target*, but instead are placed first in *intervention and content*–defining phases. As such, they are unsuitable for including and assessing all developmental stages.

Discussion

Principal Findings

This study explored alternative methods of evaluating and regulating medical apps in an international focus group of public health professionals. The focus group debate highlighted that medical app development could benefit from flexible study designs that can evaluate multiple components simultaneously and independently. While acknowledging the value of RCTs in establishing causal relationships, participants were dissatisfied with RCTs as the "go-to" evaluation method. Instead, they perceived them as a barrier to medical app development. Although participants were aware of various alternative evaluation methods, they found that their application in practice is lacking because the default method remains RCTs.

According to participants, medical app regulation should align with the rapid technological development of medical apps through an agile and iterative evaluation process, which could more appropriately capture the breadth of efficacy and safety outcomes associated with medical apps and potentially reduce the time taken to provide evidence and time taken to launch the app in the market.

The focus group enabled a difficult discussion to challenge the medical dogma of RCTs. Following the global thalidomide tragedy in 1962, an RCT became established as the best practice and is considered a requirement to demonstrate safety and efficacy for drugs entering the drug market [75]. At the time, the scientific community successfully lobbied for stricter evaluation and regulation. In the context of medical apps, regulation is in its infancy. In 2008, the launch of a diagnostic radiology app (MIMvista) at Apple's World Wide Developers Conference caused a regulatory concern and was arguably the catalyst for medical app regulation worldwide [76]. As scientists, we know the benefits of evaluation. Nevertheless, we now see that RCTs are not apt in the context of the agile nature of digital interventions; viewing RCTs as the status quo stifles innovation in the digital health sector.

The focus group provided an opportunity to share our collective desire for change and collate alternative methods of evaluation and regulation, which are presented in the logic model. This model offers a more dynamic approach to medical app evaluation and regulation. Its structure is designed to incorporate the need for detailed evaluation and regulation of medical apps. The model also encourages researchers to combine alternative study designs to meet their research goals. Alternative study designs mentioned in the focus group include pragmatic RCTs [77,78] such as cohort multiple control randomized studies [79], regression discontinuity designs [80], and registry-based randomized trials [81], which were incorporated into our logic model. To enable sustainable evidence generation, a paradigm shift in the approach to evaluation is needed. Thus, a shift from RCTs as the gold standard to a multimethod, flexible approach that reflects the agility of the medical app market [28] is required. Selected methods should aim to reduce time-to-market access, provide robust evidence on the desired (health) outcomes, and be compatible with the rapid technological development of medical apps.

Finally, it is essential to recognize that digital products constantly evolve, necessitating the ongoing efforts to optimize them retrospectively. Undertaking resource-intensive RCTs to obtain updates on medical apps, such as the German DiGA, may pose significant problems in adherence, increase dropout rate. This may lead to statistically nonsignificant results when analysis is conducted for the whole study population. The high costs and risks of re-evaluating apps using RCTs increase developmental costs, thereby increasing the risk for app developers. This, in turn, increases the probability that apps will not be launched into the market as medical apps but as health and well-being apps, circumventing regulatory approaches. If this happens, overregulation becomes dysregulation.

Strengths and Limitations

Focus groups have become a frequently used method in health care research. This is due to their usefulness in identifying and analyzing collective opinions and experiences [82,83]. Our study included multinational and multidisciplinary experts, allowing us to collect global perspectives from practitioners and researchers. The participants provided valuable insights on the challenges of using RCTs for medical app evaluation and how alternatives could be applied and integrated into current regulations. Real-world examples were shared as the group included participants who had used RCTs in medical app evaluation or refrained from developing medical apps owing to the challenges discussed in this paper. Finally, our analysis of the 4 most prominently represented countries (ie, Germany, Italy, Canada, and Australia) followed a predefined framework to guarantee comparability between countries [38]. This analysis highlights the lack of a legal obligation to use RCTs to evaluate and regulate medical apps. It suggests that there may be a reflex to use RCTs for evaluation.

Our study comes with certain limitations. Conducting the workshop at an international conference meant that we had no control on recruitment and participant demographics. Despite the global attendance at the conference, this workshop was primarily attended by people living in European and Western countries, with an underrepresentation of participants from the Global South and low- and middle-income countries. Given the international nature of digital health interventions and, in particular, its potential in low- and middle-income countries [84,85], the focus group was not likely to adequately capture the full spectrum of experiences and viewpoints. This limits the generalizability of the findings, particularly given the general underrepresentation of researchers from low- and middle-income

XSL•FO
RenderX

countries in discussions about the evaluation criteria of medical apps [86]. With a more diverse participant pool, future research could provide a more comprehensive understanding of challenges and opportunities in medical app regulation. Furthermore, participants differed in their age, academic background, and previous knowledge of this topic. Selection bias, defined as participants choosing to participate based on personal interest, limits the study's external validity.

Large focus groups, such as our study with 15 participants per group, tend to be challenging to moderate and may not achieve in-depth discussions [82,83]. This was addressed by having experienced focus group moderators who were able to facilitate an open discussion atmosphere. Although approximately 8 participants per focus group are recommended [82,83], larger groups allow for capturing more views in a shorter period. As the goal of the focus group was to gain a broad overview of current views on app evaluation and regulation, large focus groups seemed appropriate.

In addition, the feedback quality was mixed, potentially originating from the self-reported low average experiences with medical app development, evaluation, or regulation (average 1.8, SD 2.7 y; median 1.0, range 0-15 y). Another reason might be a differing understanding of medical apps compared to that of health app. This due to global variations in the definitions [6] and regulation practices associated with health apps. This might have led to an inaccurate participant understanding of medical apps. A 20-minute introductory workshop on RCTs, medical apps, and the regulation approaches for medical apps as mobile medical devices, conceivably cleared up some of the misconceptions.

Further Research and a First Look at the Desired Future

Medical apps do not necessarily have borders, making this a topic of global interest. Thus, the same challenge occurs in various countries. A focus group discussion with researchers can only be the first step in assessing this topic's relevance and diverse aspects. Further research needs to develop an evaluation framework that allows different methodological approaches for specific stages during the evaluation of the life span (from development, implementation, and creating broader impact among all stakeholders such as developers, patients, or app prescribers) of a medical app. Simultaneously, this evaluation framework should include regulatory aspects beyond the clinical evidence generated by RCTs.

Establishing the use of evaluation frameworks, such as the proposed logic model, can provide developers and regulators with a scientifically sound set of evaluation criteria that can address the agile and time-efficient development of medical apps while gradually generating evidence. In turn, this will help overcome the misconception of using RCTs as the cultural gold standard for evaluating DiGAs. Furthermore, future research should also focus on other regulatory aspects, such as patient safety and effectiveness, which are vital for the market approval of medical apps. These aspects, however, are not necessarily covered by RCTs. Therefore, future evaluation frameworks must enhance the acceptability of alternative evaluation methods for stakeholders despite the traditional research evidence provided by RCTs.

Medical apps developed by professionals without medical backgrounds can lead to several challenges, including ineffective, unsafe, and difficult-to-use interventions. Therefore, the development of medical applications should be conducted in multiprofessional teams to balance the priority of health outcomes over profitable motives.

Establishing clear regulatory criteria centered on safety, efficacy, and data protection is essential to minimize errors and harm while fully harnessing the potential of medical applications to improve health care delivery. By carefully evaluating hypothetical risks, avoiding unnecessary regulatory burdens (overregulation), and incorporating the potential for learning from manageable errors, innovation, and safety can be balanced in regulating DiGAs. This approach ensures that digital medical apps continue to evolve and improve the quality of health care services while safeguarding the well-being of users.

Conclusions

While regulations, such as the German DVG or the regulations in Canada and Australia, emphasize the need for comprehensive clinical data, their lack of strict RCT mandates allows for flexibility in evaluating digital health interventions such as medical apps. This flexibility, reinforced by consensus in focus groups, highlights the need for integrating adaptable evaluation techniques into regulatory frameworks. Policy recommendations should include specific guidance regarding the unique considerations of mobile medical devices to bridge the gap between regulatory requirements and evolving scientific methods. Focusing on ongoing assessment, adaptation of evaluation strategies, and balancing patient safety and innovation will create a robust evaluation system for medical apps. This will ultimately empower patients and health care providers while ensuring the safe and effective advancement of digital health technology. Regulatory frameworks should be updated to include guidelines for incorporating real-world data analysis, user engagement studies, and other adaptable evaluation techniques alongside traditional clinical trials.

Acknowledgments

The authors thank the workshop participants for joining the focus group discussions and for their valuable input. Without their contribution, this study would not have been possible. The authors gratefully acknowledge the support of the Leibniz ScienceCampus Bremen Digital Public Health, which is jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen, and the Leibniz Institute for Prevention Research and Epidemiology-BIPS GmbH.

XSL•FO
RenderX

The authors declare that they did not receive funding for this project. However, the State and University Library Bremen covered the open-access publication fee.

Authors' Contributions

LM initiated the project, led the workshop with RH and FH, structured the manuscript, conducted the qualitative content analysis of workshop results, modified the logic model, wrote sections in the manuscript, and edited the manuscript. ML participated in the workshop, wrote sections in the manuscript, developed and modified the logic model, and edited the manuscript. AL participated in the workshop, conducted the qualitative content analysis of workshop results, modified the logic model, and wrote sections in the manuscript. K Burdenski participated in the workshop, modified the logic model, and wrote sections in the manuscript. K Butten participated in the workshop, wrote sections in the manuscript, and edited the manuscript. SV wrote sections in the manuscript, modified the logic model, and conducted the qualitative content analysis of workshop results. MH participated in the workshop, wrote sections in the manuscript, and edited the manuscript. AG participated in the workshop, wrote sections in the manuscript, and edited the manuscript. VR participated in the workshop and wrote sections in the manuscript. GV participated in the workshop and wrote sections in the manuscript. MV wrote sections in the manuscript. FH led the workshop with LM and RH, obtained institutional review board approval, structured the manuscript, modified the logic model, wrote sections in the manuscript, and edited the manuscript. VG participated in the workshop, conducted the qualitative content analysis of workshop results, and wrote sections in the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 157 KB - jmir_v26i1e54814_app1.pdf]

Multimedia Appendix 2
Transcribed flip charts of the focus group discussions.
[PDF File (Adobe PDF File), 123 KB - jmir_v26i1e54814_app2.pdf]

Multimedia Appendix 3
Glossary of technical terms given during the focus group discussions.
[PDF File (Adobe PDF File), 101 KB - jmir_v26i1e54814_app3.pdf]

References

1. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. NPJ Digit Med 2023 Mar 10;6(1):38 [FREE Full text] [doi: 10.1038/s41746-023-00777-z] [Medline: 36899073]
2. Hicks JL, Althoff T, Sosic R, Kuhar P, Bostjancic B, King AC, et al. Best practices for analyzing large-scale health data from wearables and smartphone apps. NPJ Digit Med 2019;2:45 [FREE Full text] [doi: 10.1038/s41746-019-0121-1] [Medline: 31304391]
3. Van Ameringen M, Turna J, Khalesi Z, Pullia K, Patterson B. There is an app for that! The current state of mobile applications (apps) for DSM-5 obsessive-compulsive disorder, posttraumatic stress disorder, anxiety and mood disorders. Depress Anxiety 2017 Jun;34(6):526-539 [FREE Full text] [doi: 10.1002/da.22657] [Medline: 28569409]
4. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation) (text with EEA relevance). European Union. URL: https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1 [accessed 2024-08-21]
5. Mulder T. Health apps, their privacy policies and the GDPR. Eur J Law Technol 2019;10(1):1-20 [FREE Full text]
6. Maaß L, Freye M, Pan CC, Dassow H, Niess J, Jahnel T. The definitions of health apps and medical apps from the perspective of public health and law: qualitative analysis of an interdisciplinary literature overview. JMIR Mhealth Uhealth 2022 Oct 31;10(10):e37980 [FREE Full text] [doi: 10.2196/37980] [Medline: 36315221]
7. Timmers T, Janssen L, Pronk Y, van der Zwaard BC, Koëter S, van Oostveen D, et al. Assessing the efficacy of an educational smartphone or tablet app with subdivided and interactive content to increase patients' medical knowledge: randomized controlled trial. JMIR Mhealth Uhealth 2018 Dec 21;6(12):e10742 [FREE Full text] [doi: 10.2196/10742] [Medline: 30578185]
8. Lewis TL, Wyatt JC. mHealth and mobile medical apps: a framework to assess risk and promote safer use. J Med Internet Res 2014 Sep 15;16(9):e210 [FREE Full text] [doi: 10.2196/jmir.3133] [Medline: 25223398]

XSL•FO
RenderX

9. LaMonica HM, Roberts AE, Lee GY, Davenport TA, Hickie IB. Privacy practices of health information technologies: privacy policy risk assessment study and proposed guidelines. J Med Internet Res 2021 Sep 16;23(9):e26317 [FREE Full text] [doi: 10.2196/26317] [Medline: 34528895]
10. Rowland SP, Fitzgerald JE, Lungren M, Lee E, Harned Z, McGregor A. Digital health technology-specific risks for medical malpractice liability. NPJ Digit Med 2022 Oct 20;5(1):157 [FREE Full text] [doi: 10.1038/s41746-022-00698-3] [Medline: 36261469]
11. van der Storm SL, Jansen M, Meijer H, Barsom EA, Schijven MP. Apps in healthcare and medical research; European legislation and practical tips every healthcare provider should know. Int J Med Inform 2023 Sep;177:105141 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105141] [Medline: 37419042]
12. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. European Union. 2017. URL: https://op.europa.eu/en/publication-detail/-/publication/83bdc18f-315d-11e7-9412-01aa75ed71a1/language-en/format-PDF/source-58036705 [accessed 2024-08-21]
13. An act to prohibit the movement in interstate commerce of adulterated and misbranded food, drugs, devices, and cosmetics, and for other purposes. Rule No. 201(h). Federal Food Drug, and Cosmetic Act. URL: https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title21-section321&num=0&edition=prelim [accessed 2024-08-21]
14. Medical devices: differentiation and classification. Federal Institute for Drugs and Medical Devices. URL: https://www.bfarm.de/EN/Medical-devices/Tasks/Differentiation-and-classification/_node.html%3Bjsessionid=41CB9196DC8D464C7991BE2CB09567D3.intranet382 [accessed 2024-09-27]
15. Policy for device software functions and mobile medical applications: guidance for industry and food and drug administration staff. U.S. Food & Drug Administration (FDA). 2022. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-device-software-functions-and-mobile-medical-applications [accessed 2024-04-29]
16. Wyatt JC. How can clinicians, specialty societies and others evaluate and improve the quality of apps for patient use? BMC Med 2018 Dec 03;16(1):225 [FREE Full text] [doi: 10.1186/s12916-018-1211-7] [Medline: 30501638]
17. Chassang G. The impact of the EU general data protection regulation on scientific research. Ecancermedicalscience 2017;11:709 [FREE Full text] [doi: 10.3332/ecancer.2017.709] [Medline: 28144283]
18. Chico V. The impact of the general data protection regulation on health research. Br Med Bull 2018 Dec 01;128(1):109-118 [FREE Full text] [doi: 10.1093/bmb/ldy038] [Medline: 30445448]
19. Rumbold JM, Pierscionek B. The effect of the general data protection regulation on medical research. J Med Internet Res 2017 Feb 24;19(2):e47 [FREE Full text] [doi: 10.2196/jmir.7108] [Medline: 28235748]
20. Stern AD, Brönneke J, Debatin JF, Hagen J, Matthies H, Patel S, et al. Advancing digital health applications: priorities for innovation in real-world evidence generation. Lancet Digit Health 2022 Mar;4(3):e200-e206 [FREE Full text] [doi: 10.1016/S2589-7500(21)00292-2] [Medline: 35216754]
21. Gamble T, Haley D, Buck R, Sista N. Designing randomized controlled trials (RCTs). In: Guest G, Namey EE, editors. Public Health Research Methods. London, UK: SAGE Publications; 2015:223-250.
22. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. BJOG 2018 Dec;125(13):1716 [FREE Full text] [doi: 10.1111/1471-0528.15199] [Medline: 29916205]
23. White BK, Burns SK, Giglia RC, Scott JA. Designing evaluation plans for health promotion mHealth interventions: a case study of the Milk Man mobile app. Health Promot J Austr 2016 Feb;27(3):198-203 [FREE Full text] [doi: 10.1071/HE16041] [Medline: 27784506]
24. Mohr DC, Schueller SM, Riley WT, Brown CH, Cuijpers P, Duan N, et al. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. J Med Internet Res 2015;17(7):e166 [FREE Full text] [doi: 10.2196/jmir.4391] [Medline: 26155878]
25. Ahmed I, Ahmad NS, Ali S, Ali S, George A, Saleem DH, et al. Medication adherence apps: review and content analysis. JMIR Mhealth Uhealth 2018 Mar 16;6(3):e62 [FREE Full text] [doi: 10.2196/mhealth.6432] [Medline: 29549075]
26. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020;3:110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
27. Bonten TN, Rauwerdink A, Wyatt JC, Kasteleyn MJ, Witkamp L, Riper H, EHealth Evaluation Research Group. Online guide for electronic health evaluation approaches: systematic scoping review and concept mapping study. J Med Internet Res 2020 Aug 12;22(8):e17774 [FREE Full text] [doi: 10.2196/17774] [Medline: 32784173]
28. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023 Jan;29(1):49-58. [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
29. Carlo AD, Hosseini Ghomi R, Renn BN, Areán PA. By the numbers: ratings and utilization of behavioral health mobile applications. NPJ Digit Med 2019;2:54 [FREE Full text] [doi: 10.1038/s41746-019-0129-6] [Medline: 31304400]
30. Larsen ME, Nicholas J, Christensen H. Quantifying app store dynamics: longitudinal tracking of mental health apps. JMIR Mhealth Uhealth 2016 Aug 09;4(3):e96 [FREE Full text] [doi: 10.2196/mhealth.6020] [Medline: 27507641]

XSL•FO
RenderX

31. Singh K, Drouin K, Newmark LP, Filkins M, Silvers E, Bain PA, et al. Patient-facing mobile apps to treat high-need, high-cost populations: a scoping review. JMIR Mhealth Uhealth 2016 Dec 19;4(4):e136 [FREE Full text] [doi: 10.2196/mhealth.6445] [Medline: 27993761]
32. Eis S, Solà-Morales O, Duarte-Díaz A, Vidal-Alaball J, Perestelo-Pérez L, Robles N, et al. Mobile applications in mood disorders and mental health: systematic search in apple app store and google play store and review of the literature. Int J Environ Res Public Health 2022 Feb 15;19(4):2186 [FREE Full text] [doi: 10.3390/ijerph19042186] [Medline: 35206373]
33. Agarwal P, Gordon D, Griffith J, Kithulegoda N, Witteman HO, Sacha Bhatia R, et al. Assessing the quality of mobile applications in chronic disease management: a scoping review. NPJ Digit Med 2021 Mar 10;4(1):46 [FREE Full text] [doi: 10.1038/s41746-021-00410-x] [Medline: 33692488]
34. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 09;4(3):e107 [FREE Full text] [doi: 10.2196/mhealth.5720] [Medline: 27613084]
35. Holl F, Kircher J, Swoboda WJ, Schobel J. Methods used to evaluate mHealth applications for cardiovascular disease: a quasi-systematic scoping review. Int J Environ Res Public Health 2021 Nov 23;18(23):12315 [FREE Full text] [doi: 10.3390/ijerph182312315] [Medline: 34886039]
36. Jake-Schoffman DE, Silfee VJ, Waring ME, Boudreaux ED, Sadasivam RS, Mullen SP, et al. Methods for evaluating the content, usability, and efficacy of commercial mobile health apps. JMIR Mhealth Uhealth 2017 Dec 18;5(12):e190 [FREE Full text] [doi: 10.2196/mhealth.8758] [Medline: 29254914]
37. Liew MS, Zhang J, See J, Ong YL. Usability challenges for health and wellness mobile apps: mixed-methods study among mHealth experts and consumers. JMIR Mhealth Uhealth 2019 Jan 30;7(1):e12160 [FREE Full text] [doi: 10.2196/12160] [Medline: 30698528]
38. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):31 [FREE Full text] [doi: 10.1038/s41746-022-00573-1] [Medline: 35304561]
39. Shuren J, Patel B, Gottlieb S. FDA regulation of mobile medical apps. JAMA 2018 Jul 24;320(4):337-338 [FREE Full text] [doi: 10.1001/jama.2018.8832] [Medline: 29971339]
40. Digital health center of excellence: empowering digital health stakeholders to advance health care. U.S. Food & Drug Administration (FDA). URL: https://www.fda.gov/medical-devices/digital-health-center-excellence [accessed 2024-09-27]
41. Kao CK, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM R 2017 May 18;9(5S):S106-S115 [FREE Full text] [doi: 10.1016/j.pmrj.2017.02.018] [Medline: 28527495]
42. Bates DW, Landman A, Levine DM. Health apps and health policy: what is needed? JAMA 2018 Nov 20;320(19):1975-1976 [FREE Full text] [doi: 10.1001/jama.2018.14378] [Medline: 30326025]
43. Torous J, Stern AD, Bourgeois FT. Regulatory considerations to keep pace with innovation in digital health products. NPJ Digit Med 2022 Aug 19;5(1):121 [FREE Full text] [doi: 10.1038/s41746-022-00668-9] [Medline: 35986056]
44. Meskó B, deBronkart D. Patient design: the importance of including patients in designing health care. J Med Internet Res 2022 Aug 31;24(8):e39178 [FREE Full text] [doi: 10.2196/39178] [Medline: 36044250]
45. Silberman J, Wicks P, Patel S, Sarlati S, Park S, Korolev I, Evidence DEFINED Workgroup, et al. Rigorous and rapid evidence assessment in digital health with the evidence DEFINED framework. NPJ Digit Med 2023 May 31;6(1):101 [FREE Full text] [doi: 10.1038/s41746-023-00836-5] [Medline: 37258851]
46. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
47. Driving the digital transformation of Germany’s healthcare system for the good of patients: the act to improve healthcare provision through digitalisation and innovation (Digital Healthcare Act – DVG). Federal Ministry of Health Germany. 2019. URL: https://www.bundesgesundheitsministerium.de/en/digital-healthcare-act.html [accessed 2024-04-29]
48. Gesetz für eine bessere Versorgung durch digitalisierung und innovation. Federal Ministry of Health Germany. URL: https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&start=//*[@attr_id=%27bgbl119s0400.pdf%27]#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl119s2562.pdf%27%5D__1697794668989 [accessed 2024-04-29]
49. The fast-track process for Digital Health Applications (DiGA) according to section 139e SGB V. A guide for manufacturers, service providers and users. Federal Institute for Drugs and Medical Devices. URL: https://www.bfarm.de/SharedDocs/Downloads/EN/MedicalDevices/DiGA_Guide.html [accessed 2024-04-29]
50. Lauer W, Löbker W, Höfgen B. Digitale Gesundheitsanwendungen (DiGA): Bewertung der Erstattungsfähigkeit mittels DiGA-Fast-Track-Verfahrens im Bundesinstitut für Arzneimittel und Medizinprodukte. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct;64(10):1232-1240 [FREE Full text] [doi: 10.1007/s00103-021-03409-7] [Medline: 34529095]
51. DiGA-Verzeichnis. Federal Institute for Drugs and Medical Devices. URL: https://diga.bfarm.de/de/verzeichnis [accessed 2024-09-27]
52. Gerke S, Stern AD, Minssen T. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. NPJ Digit Med 2020;3:94 [FREE Full text] [doi: 10.1038/s41746-020-0306-7] [Medline: 32685700]

XSL•FO
RenderX

53. Guidelines on the ualification and classification of stand alone software used in healthcare within the regulatory framework of medical devices. European Commission. 2016. URL: https://ec.europa.eu/docsroom/documents/17921 [accessed 2024-10-20]
54. MDCG 2019-11: guidance on qualification and classification of software in regulation (EU) 2017/745 – MDR and regulation (EU) 2017/746 – IVDR. Medical Device Coordination Group. 2017. URL: https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2019_11_guidance_qualification_classification_software_en_0.pdf [accessed 2024-10-20]
55. Applicazione del Regolamento UE 2017/745 del Parlamento europeo e del Consiglio, del 5 aprile 2017, nel settore delle indagini cliniche relative ai dispositivi medici. Ministero della Salute. URL: https://www.trovanorme.salute.gov.it/norme/renderNormsanPdf?anno=2021&codLeg=80614&parte=1%20&serie=null [accessed 2024-11-05]
56. Patto di Legislatura: Sanità digitale e terapie digitali. Intergruppo Parlamentare Sanità Digitale e Terapie Digitali. URL: https://www.panoramasanita.it/wp-content/uploads/2023/05/Patto-di-legislatura.pdf [accessed 2024-11-05]
57. Therapeutic goods (medical devices) regulations 2022. Statutory rules 236, 2002. Compilation No. 32. Federal Register of Legislation, Australian Government. URL: https://www.legislation.gov.au/Details/F2017C00534 [accessed 2024-10-20]
58. Australian register of therapeutic goods (ARTG). Department of Health and Aged Care, Government of Australia. URL: https://www.tga.gov.au/resources/artg [accessed 2024-10-20]
59. Assessment framework for mHealth apps. Australian Digital Health Agency. URL: https://www.digitalhealth.gov.au/about-us/strategies-and-plans/assessment-framework-for-mhealth-apps [accessed 2024-04-29]
60. O'Shea B, Sridhar R. Digital health laws and regulations Australia 2024. The International Comparative Legal Guides. URL: https://iclg.com/practice-areas/digital-health-laws-and-regulations/australia [accessed 2024-08-21]
61. Jogova M, Shaw J, Jamieson T. The regulatory challenge of mobile health: lessons for Canada. Healthc Policy 2019 Feb 28;14(3):19-28 [FREE Full text] [doi: 10.12927/hcpol.2019.25795] [Medline: 31017863]
62. Health C. Guidance document: software as a medical device (SaMD): definition and classification. Health Canada. 2019. URL: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/software-medical-device-guidance-document.html [accessed 2024-04-29]
63. Hrynyschyn R, Prediger C, Stock C, Helmer SM. Evaluation methods applied to digital health interventions: what is being used beyond randomised controlled trials?: a scoping review. Int J Environ Res Public Health 2022 Apr 25;19(9):5221 [FREE Full text] [doi: 10.3390/ijerph19095221] [Medline: 35564616]
64. Traynor M. Focus group research. Nurs Stand 2015 May 13;29(37):44-48 [FREE Full text] [doi: 10.7748/ns.29.37.44.e8822] [Medline: 25967446]
65. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
66. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022 Feb;9(1):3-26 [FREE Full text] [doi: 10.1037/qup0000196]
67. Mills T, Lawton R, Sheard L. Advancing complexity science in healthcare research: the logic of logic models. BMC Med Res Methodol 2019 Mar 12;19(1):55 [FREE Full text] [doi: 10.1186/s12874-019-0701-4] [Medline: 30871474]
68. Berry A, McClellan C, Wanless B, Walsh N. A tailored app for the self-management of musculoskeletal conditions: evidencing a logic model of behavior change. JMIR Form Res 2022 Mar 08;6(3):e32669 [FREE Full text] [doi: 10.2196/32669] [Medline: 35258462]
69. Davis JA, Ohan JL, Gibson LY, Prescott SL, Finlay-Jones AL. Understanding engagement in digital mental health and well-being programs for women in the perinatal period: systematic review without meta-analysis. J Med Internet Res 2022 Aug 09;24(8):e36620 [FREE Full text] [doi: 10.2196/36620] [Medline: 35943773]
70. About the Congress. World Federation of Public Health Associations (WFPHA). URL: https://wcph.org/about-the-congress/ [accessed 2024-08-21]
71. Logic models. CDC approach to evaluation. Center for Disease Control and Prevention. URL: https://www.cdc.gov/evaluation/logicmodels/index.htm [accessed 2024-05-06]
72. Fernandez ME, Ruiter RA, Markham CM, Kok G. Intervention mapping: theory- and evidence-based health promotion program planning: perspective and examples. Front Public Health 2019 Aug 14;7:209 [FREE Full text] [doi: 10.3389/fpubh.2019.00209] [Medline: 31475126]
73. O'Cathain A, Croot L, Sworn K, Duncan E, Rousseau N, Turner K, et al. Taxonomy of approaches to developing interventions to improve health: a systematic methods overview. Pilot Feasibility Stud 2019 Mar 12;5(1):41 [FREE Full text] [doi: 10.1186/s40814-019-0425-6] [Medline: 30923626]
74. Wu X, Guo X, Zhang Z. The efficacy of mobile phone apps for lifestyle modification in diabetes: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Jan 15;7(1):e12297 [FREE Full text] [doi: 10.2196/12297] [Medline: 30664494]
75. Bhatt A. Evolution of clinical research: a history before and beyond james lind. Perspect Clin Res 2010 Jan;1(1):6-10 [FREE Full text] [Medline: 21829774]
76. Lievevrouw E, Marelli L, Van Hoyweghen I. The FDA's standard-making process for medical digital health technologies: co-producing technological and organizational innovation. Biosocieties 2022 May 13;17(3):549-576 [FREE Full text] [doi: 10.1057/s41292-021-00232-w] [Medline: 34002115]

XSL•FO
RenderX

77. Gamerman V, Cai T, Elsäßer A. Pragmatic randomized clinical trials: best practices and statistical guidance. Health Serv Outcomes Res Method 2018 Dec 1;19(1):23-35 [FREE Full text] [doi: 10.1007/s10742-018-0192-5]
78. Sherman RE, Anderson SA, Dal PG, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
79. Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. BMJ 2010 Mar 19;340(mar19 1):c1066 [FREE Full text] [doi: 10.1136/bmj.c1066] [Medline: 20304934]
80. Berk RA, Rauma D. Capitalizing on nonrandom assignment to treatments: a regression-discontinuity evaluation of a crime-control program. J Am Stat Assoc 1983 Mar;78(381):21-27 [FREE Full text] [doi: 10.1080/01621459.1983.10477917]
81. Colli A, Pagliaro L, Duca P. The ethical problem of randomization. Intern Emerg Med 2014 Oct;9(7):799-804 [FREE Full text] [doi: 10.1007/s11739-014-1118-z] [Medline: 25194693]
82. Gill P, Baillie J. Interviews and focus groups in qualitative research: an update for the digital age. Br Dent J 2018 Oct 05;225(7):668-672 [FREE Full text] [doi: 10.1038/sj.bdj.2018.815] [Medline: 30287965]
83. Goodman C, Evans C. Focus groups. In: Gerrish K, Lathlean J, Cormack D, editors. The Research Process in Nursing, 7th edition. Oxford, UK: John Wiley & Sons; 2015:401-412.
84. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2024-04-29]
85. Osei E, Mashamba-Thompson TP. Mobile health applications for disease screening and treatment support in low-and middle-income countries: a narrative review. Heliyon 2021 Mar;7(3):e06639 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e06639] [Medline: 33869857]
86. Woulfe F, Fadahunsi KP, Smith S, Chirambo GB, Larsson E, Henn P, et al. Identification and evaluation of methodologies to assess the quality of mobile health apps in high-, low-, and middle-income countries: rapid review. JMIR Mhealth Uhealth 2021 Oct 12;9(10):e28384 [FREE Full text] [doi: 10.2196/28384] [Medline: 34636737]

Abbreviations

DiGA: digital health applications
DVG: Digital Care Act
RCT: randomized controlled trial
SaMD: software as a medical device
TA: thematic analysis
WCPH: World Congress on Public Health

Edited by A Mavragani; submitted 28.11.23; peer-reviewed by CC Pan, TAR Sure, B Ekman; comments to author 27.03.24; revised version received 15.05.24; accepted 24.07.24; published 30.09.24.

Please cite as:
Maaß L, Hrynyschyn R, Lange M, Löwe A, Burdenski K, Butten K, Vorberg S, Hachem M, Gorga A, Grieco V, Restivo V, Vella G, Varnfield M, Holl F
Challenges and Alternatives to Evaluation Methods and Regulation Approaches for Medical Apps as Mobile Medical Devices: International and Multidisciplinary Focus Group Discussion
J Med Internet Res 2024;26:e54814
URL: https://www.jmir.org/2024/1/e54814
doi:10.2196/54814
PMID:

©Laura Maaß, Robert Hrynyschyn, Martin Lange, Alexandra Löwe, Kathrin Burdenski, Kaley Butten, Sebastian Vorberg, Mariam Hachem, Aldo Gorga, Vittorio Grieco, Vincenzo Restivo, Giuseppe Vella, Marlien Varnfield, Felix Holl. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Brief Peer-Supported Web-Based Skills Training in Affective and Interpersonal Regulation (BPS webSTAIR) for Trauma-Exposed Veterans in the Community: Randomized Controlled Trial

Laura E Ong[1], AB; Sarah Speicher[2], MS; Diana Villasenor[2], BS; Jamie Kim[2], MA; Adam Jacobs[2], BS; Kathryn S Macia[2,3,4], PhD; Marylene Cloitre[2], PhD

[1]Department of Psychology, Northern Illinois University, DeKalb, IL, United States

[2]National Center for Posttraumatic Stress Disorder, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, United States

[3]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, United States

[4]Veterans Affairs Health Systems Research, Center for Innovation to Implementation, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, United States

Corresponding Author:
Marylene Cloitre, PhD
National Center for Posttraumatic Stress Disorder
Veterans Affairs Palo Alto Health Care System
425 First Street
Unit 3004
Palo Alto, CA, 94105
United States
Phone: 1 6504935000
Email: Marylene.Cloitre@va.gov

Abstract

Background: Peer-supported mobile health (mHealth) programs hold the promise of providing a low-burden approach to increasing access to care and improving mental health. While peer support has been shown to improve engagement in care, there is limited investigation into the impact of peers on symptom outcomes. Trauma-exposed populations frequently endure co-occurring posttraumatic stress and depressive symptoms as well as difficulties in day-to-day functioning. This study evaluated the potential benefits of a peer-supported, transdiagnostic mHealth program on symptom outcomes and functioning.

Objective: This randomized controlled trial tested the effectiveness of Brief Peer-Supported (BPS) web-based Skills Training in Affective and Interpersonal Regulation (webSTAIR), a 6-module transdiagnostic digital program derived from Skills Training in Affective and Interpersonal Regulation and compared to waitlist control in a community sample of veterans who screened positive for either posttraumatic stress disorder (PTSD) or depression.

Methods: A total of 178 veterans were enrolled in this study using a 2:1 randomization scheme with 117 assigned to BPS webSTAIR and 61 assigned to waitlist control. PTSD and depressive symptoms as well as emotion regulation and psychosocial functioning were assessed at pretreatment, posttreatment, and 8-week follow-up time points. Mixed-effects models were used to assess change in outcome measures across time points. Exploratory analyses were conducted to determine whether the type and number of peer interactions influenced outcomes.

Results: Significant interaction effects were observed for all outcomes such that participants randomized to BPS webSTAIR reported significantly greater improvement at the posttreatment time point compared to waitlist control with moderate effect sizes for PTSD (d=0.48), depression (d=0.64), emotion regulation (d=0.61), and functional impairment (d=0.61); gains were maintained at 8-week follow-up. An initial cohort of participants who were required to engage with a peer coach to progress through the modules interacted more frequently with peers but completed fewer modules compared to a later cohort for whom peer engagement was optional. Overall, those who completed more modules reported greater improvement in all outcomes.

Conclusions: BPS webSTAIR was effective in improving PTSD and depression symptoms, emotion regulation, and psychosocial functioning in community veterans. Peer-supported, transdiagnostic mHealth programs may be a particularly efficient, effective, and low-burden approach to improving mental health among trauma-exposed populations. Investigation of peer-supported programs among other populations is necessary to evaluate the generalizability of the findings. Analyses comparing peer support that was required versus optional indicated that some veterans may not need or want peer support. Future research should evaluate

XSL•FO
RenderX

how best to deliver peer support and for whom it is most beneficial. If successful, peer-supported tech programs may increase the Veteran Affairs workforce as well as improve veteran mental health services and outcomes.

Trial Registration: ClinicalTrials.gov NCT04286165; https://clinicaltrials.gov/study/NCT04286165

(J Med Internet Res 2024;26:e52130) doi:10.2196/52130

KEYWORDS

posttraumatic stress disorder; PTSD; depression; depressive symptoms; veterans; veterans health; mHealth; mobile health; peer support; peer-to-peer; transdiagnostic; mental health; mental health services; community; emotion regulation; interpersonal regulation; mHealth program

Introduction

Background

Among US veterans, the prevalence of comorbid posttraumatic stress disorder (PTSD) and major depressive disorder is estimated to be roughly twice the rate of PTSD alone [1], and the symptoms of each of these disorders have been associated with deficits in emotion regulation [2,3] and psychosocial functioning [4,5]. Meanwhile, veterans experience numerous practical and cultural barriers to accessing evidence-based mental health treatment. These include the need to travel long distances to facilities, financial constraints, health conditions, inflexible work schedules, and caregiver responsibilities [6-8], as well as veterans' desire to be self-reliant and the perceived stigma of receiving mental health care [9,10]. There are, in addition, systemic barriers including a shortage of mental health providers in many parts of the United States [11]. Mobile health (mHealth) programs, including web-based and app programs, address these difficulties and have been found to be effective in the treatment of symptoms of PTSD and depression [12-15]. However, it has been shown that mHealth programs with human support consistently provide greater benefits than entirely self-guided programs [16,17]. Given the burdens on clinicians and other licensed professionals in effectively serving patient needs, the introduction of peer-supported mHealth programs holds the promise of reducing the burden on the health care system while maintaining and potentially increasing access to care and supporting patients in obtaining optimal outcomes.

Peer specialists are persons with lived experience of recovery from mental health difficulties who are trained to support patients with similar problems in engaging in and completing mental health programs [18]. Peer specialists bring particular strengths to their roles, including their ability to develop strong rapports with participants and reduce stigma [19,20]. Literature reviews have indicated that the presence of peers increases engagement in treatment across persons with a variety of mental health difficulties and disorders, treatment modalities, and delivery strategies [18,21]. In particular, the use of peer support in mHealth has been found to increase engagement among veterans with unmet mental health needs [22]. However, randomized controlled trials assessing the impact of peers on symptom change are relatively few in number [23] including for mHealth interventions [21].

Studies integrating peer specialists into Veteran Affairs (VA) mHealth programs have strongly supported their acceptability and feasibility [22,24-26]. Randomized controlled trials of peer-supported, web-based programs have reported moderate to modest symptom change outcomes on PTSD and depression when compared to waitlist [27] and found that utilization of peer support was associated with greater reductions in depression when compared to self-directed program usage [22]. Additional trials are needed to assess mental health outcomes, particularly symptom change and psychosocial functioning to better understand the optimal contribution of peer-supported mHealth programs to mental health care. This study evaluates a peer-supported, transdiagnostic mHealth program for trauma-exposed veterans. Given the frequency of comorbid symptomatology among trauma-exposed populations [1,28], a transdiagnostic approach, if successful, provides the opportunity for broad use across various trauma-exposed populations with diverse needs. Moreover, different types of symptoms share issues with emotion regulation and psychosocial functioning, making these difficulties important treatment targets.

The program evaluated, web-based Skills Training in Affective and Interpersonal Regulation (webSTAIR), is a web-based program adapted from Skills Training in Affective and Interpersonal Regulation (STAIR), a manualized evidence-supported cognitive behavioral intervention for trauma-exposed persons with symptoms of PTSD or depression that focuses on improving psychosocial functioning by building emotion regulation and interpersonal skills. Investigations of 10-12 session STAIR have reported large effect sizes in the reduction of PTSD and depression as well as in emotion regulation and psychosocial impairment [29,30]. Similar outcomes have been obtained in a briefer 6-session version, including in a randomized controlled trial in primary care [31] and in an open trial delivered by peers in a low-income, primary care community service [32]. In parallel to the STAIR studies, investigations of webSTAIR assessed a 10-module version with weekly coaching sessions from licensed mental health providers. These studies have included 2 open trials [33,34] and 1 comparison trial [35]. The open trials found significant improvements with moderate to large effect sizes for PTSD, depression, emotion regulation difficulties, and psychosocial impairment, and the comparison study, with a noninferiority design, found that reducing coaching support to biweekly (5 coaching sessions) from weekly (10 coaching sessions) resulted in noninferior benefits on all outcomes.

Given the evidence supporting the success of brief, peer-supported STAIR in reducing a range of symptoms, along with growing support for the effectiveness of web-based mental health programs, a 6-session peer-supported version of webSTAIR was developed. This abbreviated version of the

XSL•FO
RenderX

webSTAIR protocol can be completed in a shorter time frame than the 10-session webSTAIR and integrates support from trained veteran peers at the end of each module. The potential benefits of this program, if successful, is the availability of a low-burden, brief mHealth program that is effective in reducing PTSD and depression symptoms as well as improving emotion regulation and psychosocial functioning.

Objectives

This study aimed to assess the benefits of Brief Peer-Supported (BPS) webSTAIR compared to a waitlist control condition among US veterans with symptoms of PTSD, depression, or both. We hypothesized that participants assigned to the BPS webSTAIR condition would experience significantly greater improvement in PTSD symptoms and depressive symptoms (primary outcomes) as well as in emotion regulation and psychosocial functioning (secondary outcomes) compared to the waitlist, and that these gains would be maintained at an 8-week follow-up. Exploratory analyses were conducted to determine whether the number and type of peer interactions influenced outcomes. This study was funded by the National Center for PTSD Dissemination and Training Division within the VA Palo Alto Health Care System.

Methods

Procedures

Candidates for study participation were recruited via social media advertisements and directed to complete a short set of online screening questions. Following the screening, study candidates were contacted via phone and scheduled for a 30- to 60-minute phone assessment with this study's coordinator to assess inclusion and exclusion criteria, receive an explanation of the program, and complete (verbal) informed consent. If eligible for this study, the participants were provided access to the program. Randomization was allocated at a 2:1 ratio (webSTAIR to waitlist) in blocks of 12 and was computer generated via an algorithm developed by staff otherwise not involved in this study. Participant randomization assignment was indicated upon opening the program URL. Participants in the BPS webSTAIR condition received instructions to start the program within 2 weeks following the assessment and to complete the 6-module program within 10 weeks. Posttreatment assessment (10-week mark) and an 8-week follow-up were conducted via phone by study staff who were unaware of the condition and assessment period. Participants in the waitlist condition were given the information that they would have access to the webSTAIR program in 10 weeks and completed assessments at the pretreatment and posttreatment time points (10-week mark). Waitlist participants were then given access to webSTAIR. Recruitment began on October 16, 2020, and follow-up ended on September 28, 2022. This study is registered at ClinicalTrials.gov (NCT04286165).

Participants

Persons were eligible if they were veteran or military personnel; aged older than 21 years; reported experiencing at least 1 traumatic event; reported symptoms of PTSD or depression, or both, as indicated by a score of ≥3 on the Primary Care Posttraumatic Stress Disorder Screen for *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [36] or ≥2 the Patient Health Questionnaire-2 item (PHQ-2) [37]; were able to read and write in English; had a PC and stable internet connection; and resided in the United States. Exclusion criteria were determined during the phone assessment and were (1) the presence of significant suicidality as indicated by the presence of a plan and means, (2) the presence of cognitive difficulties or active psychosis that would indicate a low likelihood of benefiting from treatment, and (3) current participation in trauma-focused treatment.

Measures

Overview

The pretreatment assessment included a series of sociodemographic questions and the Life Events Checklist (LEC) [38], which in this study was used to measure only those potentially traumatic events that directly happened to the person, in addition to the measures below. The posttreatment assessment was scheduled for week 10 of this study and follow-up at 8 weeks after treatment.

PTSD Symptoms

PTSD symptoms were measured using the Posttraumatic Stress Disorder Checklist for *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (PCL-5) [39]. This 20-item self-report measure asks participants to indicate how much they have been bothered by each symptom in the past month, on a scale from 0 ("not at all") to 4 ("extremely"). A total score is calculated by summing all items, where higher scores correspond to greater PTSD symptom severity. A cutoff score of 31-33 points or more has been established as indicating probable PTSD [40]. The PCL-5 has demonstrated excellent internal consistency, good test-retest reliability, and excellent convergent validity among veterans [40]. The baseline Cronbach α of the PCL-5 in the current sample was 0.92.

Depressive Symptoms

Depressive symptoms were measured using the Patient Health Questionnaire-8 item (PHQ-8) [41]. The PHQ-8 is a self-report measure that comprises the first 8 items of the Patient Health Questionnaire-9 item (PHQ-9). It asks participants to rate how often they have been bothered by each problem over the past month, on a scale from 0 ("not at all") to 3 ("nearly every day"). A total score is calculated by summing all items, where higher scores correspond to greater depressive symptom severity. The PHQ-8 has demonstrated good internal consistency and convergent validity among both psychiatric outpatients and the general population [41,42]. The baseline Cronbach α of the PHQ-8 in the current sample was 0.83.

Emotion Regulation

Emotion regulation was measured using the Difficulties in Emotion Regulation Scale-16 item (DERS-16) [43]. This 16-item self-report measure asks participants to rate how often each of a series of statements applies to them on a scale from 1 ("almost never") to 5 ("almost always"). A total score is calculated by summing all items, where higher scores correspond to greater difficulties with emotion regulation. The DERS-16

XSL•FO
RenderX

has demonstrated excellent internal consistency, good test-retest reliability, and good convergent and discriminant validity [44]. The baseline Cronbach α of the DERS-16 in the current sample was 0.93.

Psychosocial Functioning

Psychosocial functioning was measured using the Work and Social Adjustment Scale (WSAS) [45]. This 5-item self-report measure asks participants to rate their degree of impairment in carrying out each of a series of activities on a scale from 0 ("not at all") to 8 ("very severely"). A total score is calculated by summing all items, where higher scores correspond to greater functional impairment. The WSAS has demonstrated strong psychometric properties among psychiatric samples [45,46]. The baseline Cronbach α of the WSAS in the current sample was 0.82.

Engagement and Completion Data

Metadata on chat sessions, including the number of messages exchanged with a peer coach, was collected at the participant level. Measures of program engagement within the BPS webSTAIR condition included the number of modules completed, the total number of messages exchanged, and the average number of messages exchanged per module.

Treatment

The BPS webSTAIR intervention is a shortened, 6-module version of webSTAIR adapted from brief STAIR [31]. The 6 modules of BPS webSTAIR focus on emotional awareness; emotion management through the body, thought, and behavior channels; distress tolerance; and self-compassion. The program integrates text, audio, and video delivery of psychoeducation with interactive exercises and worksheets to help patients understand and internalize the concepts introduced.

At the end of each module, participants were prompted to complete a chat session with a veteran peer coach to help clarify program content and answer any questions they might have about how to implement the skills presented in the module. Interaction with the peer occurred via secure instant messaging on the program platform. Participants could message peers as often as they liked while they were on the platform. It was not expected that the participant would receive a response from the same peer. The peer program was committed to responding to text messages within 24 hours. Due to changes in the availability of veteran peer coaches over this study's period, 3 distinct cohorts were identified based on participant start date. In cohort 1 (from October 16, 2020, to July 9, 2021), coaches were available for 24/7 support, and participants were required to complete a chat session with a coach at the end of each module to move forward to the next module. In cohort 2 (from July 10, 2021, to November 22, 2021), coach availability was reduced to 8 hours a week, with 4 hours at the beginning of the day; 8:30 AM-12:30 PM) and 4 hours at the end of the day and into the evening (4:30 PM). In cohort 3 (from November 23, 2021, to July 20, 2022), coach availability remained reduced, but participants were able to independently skip any chat sessions they chose. For all 3 cohorts, the research coordinator was available as backup for logistical and content support Monday through Friday by phone call or secure text message during an 8-hour workday (8 AM to 4 PM Pacific) with responses completed within 24 hours or less.

Veteran Peer Coaches

All veteran peer coaches were VA-certified peer counselors, which includes a working knowledge of VA guidelines for assessing suicidal risk and referral to the veteran suicide hotline. In addition, for this project, the peers received training from the developers of the web-based platform about navigation of the system (Vets Prevail) and received a manual and a workshop consisting of two 2-hour meetings taught by webSTAIR staff (MC and SS) which described the interventions and provided suggestions about how to connect the program material to ongoing stressors in the life of the veteran client and to recommend the practice of skills between sessions.

Statistical Analysis Plan

Analyses were conducted using R (version 4.2.2; R Foundation for Statistical Computing). All participants who met eligibility requirements and completed the pretreatment assessment were included in the final dataset, regardless of whether they completed the program or participated in subsequent assessments. Consequently, all analyses were conducted using the intention-to-treat (ITT) sample of randomized participants. Categorical pretreatment demographic variables and responses to individual LEC items were compared between the treatment and control groups using chi-squared tests of independence. The mean LEC scores for the 2 groups were compared using independent samples 2-tailed *t* tests. Rates of missingness were 0% (0/178) at baseline, 38.2% (68/178) at posttreatment assessment with 47% (55/117) missing for webSTAIR and 21% (13/61) missing for waitlist, and 41.9% (49/117) at follow-up (webSTAIR condition only). Missing data were addressed using multiple imputation. Joint multiple imputation of missing values at posttreatment and follow-up assessments was carried out using the *panImpute* function in package *mitml*. Due to (1) some of the presently reported analyses being specific to webSTAIR participants, while other analyses were relevant to the full sample, (2) differences in the assessment schedule between the treatment and control conditions (ie, only webSTAIR participants were assessed at follow-up), (3) the unbalanced randomization design, and (4) differences in rates of missingness between conditions, multiple imputation procedures were carried out separately for the data that would be used to examine within-group change for webSTAIR participants across the 3 time points and between-group differences in change from pre- to posttreatment assessments. In other words, the between-group imputed datasets consisted of baseline and posttreatment data for all participants, whereas the within-group imputed datasets included baseline, posttreatment, and follow-up data for only the participants randomized to webSTAIR. In each case, 100 datasets were generated with missing values imputed. Before imputing the data, chi-square tests and 2-tailed *t* tests were used to identify predictors of missingness and baseline characteristics related to nonmissing values on outcome variables. The missing data correlates identified from these preliminary analyses (employment status, education level, service era, and gender) were included as auxiliary variables in the imputation models. To ensure that any interactions between time and condition were

XSL•FO
RenderX

appropriately preserved in the between-group imputed datasets, this interaction term was included as a predictor in the between-group imputation model. The results were pooled across imputed datasets using the test estimates function in the *mitml* package.

For the main analyses of our primary and secondary outcomes, linear mixed-effects models were fitted using restricted maximum likelihood estimation with the *lmer* function in the package *lme4*. These models included main effects for time and condition, a time-by-condition interaction, as well as fixed effects to adjust for potential differences between the cohorts that may have resulted from changes in peer accessibility and chat requirements during this study. To control for potential cohort effects, 2 dummy variables were created to identify the participants in cohorts 2 and 3, with cohort 1 as the reference group. A random intercept was included in all models to account for repeated measurements within participants. The time by condition interaction term provided a test of the difference in change from pre- to posttreatment assessments between the webSTAIR and waitlist conditions. Cohen *d* between-group effect sizes for each measure were calculated by dividing the estimated interaction term by the SD of pretreatment scores in the total sample. Similar models were constructed to measure within-group change among webSTAIR participants across all 3 time points. The time predictor in these models was specified so that posttreatment assessment was the reference category, which enabled us to evaluate the magnitude and statistical significance of changes that occurred from pre- to posttreatment assessments, as well as whether any improvements during treatment were retained during the follow-up period (ie, as indicated by a nonsignificant estimate for change between the posttreatment and follow-up assessments). In the within-group analyses, Cohen *d* effect sizes for change from pre- to posttreatment assessment and from posttreatment to follow-up assessment in the BPS webSTAIR group were calculated by dividing the estimated coefficients for time by the SD of pretreatment scores for the BPS webSTAIR condition.

To evaluate the impact of program completion on outcomes, module completion was categorized as “none” (did not complete any web-based content, including the welcome module), “some” (completed the welcome module and no more than module 2), or “moderate to complete” (completed module 3 or higher). Within each completion group, linear mixed-effects models were constructed for each outcome measure to evaluate within-group change across time. Time predictor dummies were specified such that posttreatment assessment was the reference category, and a random intercept was included in each model. Cohen *d* effect sizes for pre- to posttreatment change within each completion group were calculated by dividing the estimated coefficient for time by the SD of pretreatment scores for BPS webSTAIR participants in the respective completion group (eg, the SD of pretreatment PCL-5 scores among BPS webSTAIR participants in the “moderate to complete” group).

Exploratory analyses were conducted to better understand the impact of the type (required versus optional) and number of peer chat sessions in the different cohorts. For each participant, we calculated the total number of chat messages exchanged with a peer support coach and the average number of chat messages exchanged per module of the program completed. Descriptive statistics for chat messages were examined by cohort. Comparisons were drawn between cohort 1, in which chat sessions were required to advance to the next module, and cohort 3, in which chat sessions were not required for module advancement, to assess whether chat engagement and completion rates differed in these 2 contexts. We did not include cohort 2 in these analyses, given that this was a transitional phase when participants’ experiences with the chat feature may have varied in ways that are not quantifiable (eg, experienced highly variable wait times in connecting with a peer).

Finally, the relationship of chat messages with treatment outcomes was explored by fitting linear regression models that included average messages exchanged per module and pretreatment scores for the outcome as predictors of pre- to posttreatment change in the outcome. Due to differences between cohorts in the availability of peer coaches and program instructions for module completion, the relationships between messages exchanged per module and changes in outcomes could not be interpreted in the same way across cohorts. Given these differences in context, separate models were estimated for each cohort so that relationships between chat engagement and outcomes could be interpreted according to the specific circumstances of each cohort. Additionally, differences in the number of modules completed by participants were taken into account by examining the relationships between messages exchanged *per module* (vs total messages exchanged) and change in outcomes.

Ethical Considerations

Institutional review board approval for the study was obtained from Stanford University (Protocol 51076). All participants provided verbal informed consent. Participants were paid US $40 for the pretreatment assessment and US $60 for the posttreatment and follow-up assessments. Data were collected online and stored in a secure VA-approved environment.

Results

A total of 178 eligible veterans were enrolled in this study using a 2:1 randomization scheme with 117 assigned to BPS webSTAIR and 61 assigned to the waitlist control. As indicated in the CONSORT chart (Figure 1), 201 participants were initially enrolled but 23 participants (16 in webSTAIR and 7 on waitlist) were excluded from this study predominantly due to low credibility regarding their veteran status (eg, did not know their rank or were uncertain about which branch of the armed forces they served in). Sample sociodemographic characteristics are shown in Table 1. Similar to other veteran samples, most participants were male (125/178, 70.2%). Roughly two-thirds (120/178, 67.4%) were White and the average age was 48.08 (SD 9.04) years. A total of 57.9% (103/178) were approved for VA service connection status for PTSD (at any disability rating), 27.0% (48/178) had never applied, 8.4% (15/178) had applied or had an application under review, and 6.7% (12/178) had an application denied. A total of 32.0% (57/178) were currently engaged in other (non–trauma-focused) counseling, and 6.7% (12/178) had received evidence-based treatment for PTSD or related

XSL•FO
RenderX

conditions within the past year. No significant differences by condition were found for any sociodemographic item (all $P>.05$). At the pretreatment assessment, the 2 conditions did not significantly differ on any outcome measures except for the PHQ-8 ($P=.03$).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram.

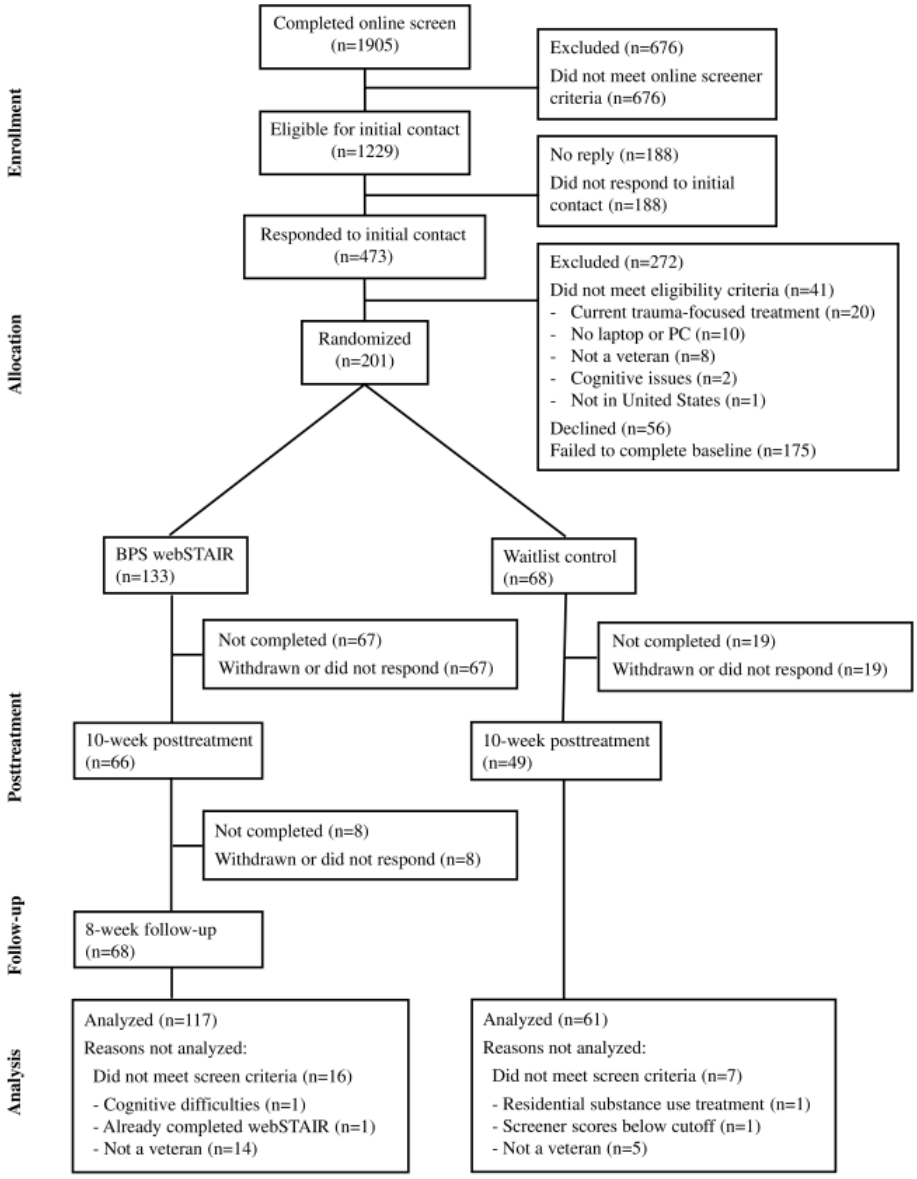

XSL•FO
RenderX

Table 1. Sociodemographic and clinical characteristics of participants at baseline.

Baseline characteristic	BPS[a] webSTAIR[b] (n=117)	Waitlist (n=61)	*P* value
Age[c] (years), mean (SD)	47.66 (9.31)	48.90 (8.49)	.38
Race or ethnic background, n (%)			.49
White	80 (68.4)	40 (66)	
Black or African American	17 (14.5)	9 (15)	
Hispanic, Latino/a, or Mexican	9 (7.7)	7 (12)	
American Indian or Alaskan Native	4 (3.4)	3 (5)	
Native Hawaiian or Pacific Islander	2 (1.7)	1 (1.6)	
Asian	0 (0)	1 (1.6)	
Other	5 (4.3)	0 (0)	
Self-identified gender, n (%)			.32
Male	81 (69.2)	44 (72)	
Female	36 (30.8)	16 (26)	
Transgender	0 (0)	1 (2)	
Lifetime trauma (LEC[d]), mean (SD)	10.92 (2.7)	10.84 (3)	.84
Baseline outcomes, mean (SD)			
PCL-5[e]	48.41 (14.55)	43.93 (16.84)	.08
PHQ-8[f]	15.41 (5.23)	13.48 (5.66)	.03
DERS-16[g]	52.44 (13.95)	48.18 (15.26)	.07
WSAS[h]	27.05 (8.70)	25.25 (9.14)	.21

[a]BPS: Brief Peer-Supported.

[b]webSTAIR: web-based Skills Training in Affective and Interpersonal Regulation.

[c]Age data were unavailable for 1 participant in the webSTAIR condition and 2 participants in the waitlist condition.

[d]LEC: Life Events Checklist.

[e]PCL-5: Posttraumatic Stress Disorder Checklist for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

[f]PHQ-8: Patient Health Questionnaire-8 item.

[g]DERS-16: Difficulties in Emotion Regulation Scale-16 item.

[h]WSAS: Work and Social Adjustment Scale.

Table 2 presents the imputed means and SDs of scores on outcome measures by condition across time points, along with the model-implied Cohen *d* effect size measures for between-group differences in change and overall change within the webSTAIR group, derived from the linear mixed-effects model results. Significant condition×treatment interaction effects were observed for the PCL-5 (*P*=.02), PHQ-8 (*P*=.01), DERS-16 (*P*=.003), and WSAS (*P*=.045), indicating that webSTAIR participants experienced significantly greater improvements from pre- to posttreatment assessment than waitlist participants. The results of the mixed-effects models examining change in outcomes across the 3 time points at which webSTAIR participants were assessed similarly indicated that webSTAIR participants experienced significant improvements on all measures between pre- and posttreatment assessment: PCL-5 (*P*=.001), PHQ-8 (*P*<.001), DERS-16 (*P*<.001), and WSAS (*P*<.001). Improvements experienced during treatment on all outcome measures were maintained through follow-up with no significant changes observed from posttreatment to follow-up assessment for any of the outcome measures (all *P*>.05). Male veterans did not differ from veterans who identified as other genders on any of the outcome measures at posttreatment (PCL-5: *P*=.50; PHQ-8: *P*=.31; DERS-16: *P*=.22; WSAS: *P*=.42) or follow-up assessment (PCL-5: *P*=.47; PHQ-8: *P*=.18; DERS-16: *P*=.21; WSAS; *P*=.15).

XSL•FO
RenderX

Table 2. Imputed means, SD, and Cohen d effect sizes of outcomes[a].

Measure and condition	Pretreatment assessment	Posttreatment assessment	Follow-up assessment	Time (pre- to posttreatment assessment) × condition interaction		Pre- to posttreatment change for BPS[b] webSTAIR[c]		Posttreatment to FU[d] change for BPS webSTAIR	
				Cohen *d* (95% CI)	*P* value	Cohen *d* (95% CI)	*P* value	Cohen *d* (95% CI)	*P* value
PCL-5[e]				–0.43 (–0.78 to –0.07)	.02	–0.48 (–0.76 to –0.21)	.001	0.05 (–0.23 to 0.32)	.75
BPS webSTAIR	48.41 (14.55)	41.28 (17.56)	42.09 (18.28)						
Waitlist	43.93 (16.84)	43.39 (17.86)	__[f]						
PHQ-8[g]				–0.51 (–0.90 to –0.12)	.01	–0.64 (–0.87 to –0.40)	<.001	0.12 (–0.15 to 0.39)	.38
BPS webSTAIR	15.41 (5.23)	12.05 (5.77)	12.71 (6.13)						
Waitlist	13.48 (5.66)	12.88 (5.77)	—						
DERS-16[h]				–0.60 (–0.99 to –0.21	.003	–0.61 (–0.85 to –0.36)	<.001	0.07 (–0.19 to 0.34)	.60
BPS webSTAIR	52.44 (13.95)	43.04 (15.10)	44.95 (15.97)						
Waitlist	48.18 (15.26)	47.50 (14.82)	—						
WSAS[i]				–0.38 (–0.74 to –0.01)	.045	–0.61 (–0.87 to –0.35)	<.001	0.13 (–0.12 to 0.38)	.32
BPS webSTAIR	27.05 (8.70)	21.74 (10.44)	22.83 (10.72)						
Waitlist	25.25 (9.14)	23.27 (10.40)	—						

[a]All models controlled for potential cohort effects.

[b]BPS: Brief Peer-Supported.

[c]webSTAIR: web-based Skills Training in Affective and Interpersonal Regulation.

[d]FU: follow-up.

[e]PCL-5: Posttraumatic Stress Disorder Checklist for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

[f]Not applicable.

[g]PHQ-8: Patient Health Questionnaire-8 item.

[h]DERS-16: Difficulties in Emotion Regulation Scale-16 item.

[i]WSAS: Work and Social Adjustment Scale.

Table 3 shows the estimated coefficients and effect sizes for pre- to posttreatment change in the outcome measures by module completion group in the BPS webSTAIR condition. Participants who completed no modules (n=25; coded as “None”) did not experience a significant change in PCL-5 score from pretreatment to posttreatment assessment (estimate=–1.88, SE 4.72; *P*=.69; *d*=–0.13). Those who completed at least the welcome module but did not move beyond the second full module of BPS webSTAIR (n=48; coded as “Some”) experienced a mean PCL-5 total score reduction of 6.77 (SE 3.16; *P*=.03; *d*=–0.50). Participants who completed 3 or more (out of 6 total) modules of BPS webSTAIR (n=44; coded as “Moderate to Complete”), experienced an estimated PCL-5 total score reduction of 10.13 (SE 2.49) points from pretreatment to posttreatment assessment (*P*<.001; *d*=–0.64). The PHQ-8, DERS-16, and WSAS followed a similar pattern, wherein significant pre- to posttreatment improvements were observed for the “some” and “moderate to complete” groups, but not the “none” group.

XSL•FO
RenderX

Table 3. Results from mixed models examining pre- to posttreatment changes in outcome by completion group.

Measure	Module completion category								
	None (n=25)[a]			Some (n=48)[b]			Moderate to complete (n=44)[c]		
	Estimate (SE)	*P* value	Cohen *d* (95% CI)	Estimate (SE)	*P* value	Cohen *d* (95% CI)	Estimate (SE)	*P* value	Cohen *d* (95% CI)
PCL-5[d]	–1.88 (4.72)	.69	–0.13 (–0.79 to 0.53)	–6.77 (3.16)	.03	–0.50 (–0.95 to –0.04)	–10.13 (2.49)	<.001	–0.64 (–0.95 to –0.33)
PHQ-8[e]	–1.15 (1.46)	.43	–0.23 (–0.82 to 0.35)	–3.09 (0.96)	.001	–0.68 (–1.09 to –0.27)	–4.82 (0.82)	<.001	–0.81 (–1.08 to –0.54)
DERS-16[f]	–2.18 (4.26)	.61	–0.13 (–0.65 to 0.38)	–8.44 (2.69)	.002	–0.70 (–1.13 to –0.26)	–12.16 (2.07)	<.001	–0.83 (–1.11 to –0.56)
WSAS[g]	–2.33 (2.62)	.37	–0.25 (–0.79 to 0.30)	–5.04 (1.74)	.004	–0.64 (–1.07 to –0.20)	–7.31 (1.32)	<.001	–0.81 (–1.10 to –0.52)

[a]Completed no modules.

[b]Completed through the welcome module, module 1, or module 2.

[c]Completed through module 3 or beyond.

[d]PCL-5: Posttraumatic Stress Disorder Checklist for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

[e]PHQ-8: Patient Health Questionnaire-8 item.

[f]DERS-16: Difficulties in Emotion Regulation Scale-16 item.

[g]WSAS: Work and Social Adjustment Scale.

Chat usage data are displayed for all participants in the BPS webSTAIR condition and by cohort in Table 4. Participants in cohort 1 completed a significantly higher number of chat sessions compared with participants in cohort 3 (*P*=.004). The average number of messages exchanged per module was 11.92 (SD 11.35) in cohort 1 and 4.44 (SD 9.24) in cohort 3 (*P*=.004), and participants in cohort 3 were significantly more likely to exchange no messages at all (χ^2_1=10.00, N=71, *P*=.002). However, participants in cohort 3 completed more modules, an average of 4.05 (SD 2.75) modules (including the welcome module) compared to an average of 2.45 (SD 2.68) modules completed in cohort 1 ($t_{65.4}$=–2.47, *P*=.02). Additionally, 16 (40%) of the 40 participants in cohort 3 completed the program compared with only 6 (19%) of the 31 participants in cohort 1 (χ^2_1=3.48, N=71, *P*=.06). Regarding the relationships of chat usage with outcomes, the number of messages exchanged per module in cohort 1 was not significantly associated with pre- to posttreatment change in any of this study's outcomes (PCL-5: *P*=.87; PHQ-8: *P*=.37; DERS-16: *P*=.84; WSAS: *P*=.49). The results were similar for cohort 3 in that the number of messages exchanged per module was not significantly associated with pre- to posttreatment change in any outcome (PCL-5: *P*=.89; PHQ-8: *P*=.67; DERS-16: *P*=.82; WSAS: *P*=.66).

Table 4. Peer chat usage metrics by cohort.

	Cohort 1 (n=31)	Cohort 2 (n=46)	Cohort 3 (n=40)	All cohorts (N=117)
No messages exchanged, n (%)	10 (32)	17 (37)	28 (70)	55 (47)
Total messages exchanged, mean (SD)	48.87 (62.24)	21.46 (36.09)	23.58 (53.04)	29.44 (50.90)
Messages exchanged per module[a], mean (SD)	11.92 (11.35)	5.97 (7.79)	4.44 (9.24)	7.02 (9.73)
No modules completed, n (%)	11 (36)	13 (28)	1 (3)	25 (21.4)
Modules completed[a], mean (SD)	2.45 (2.68)	2.43 (2.53)	4.05 (2.75)	2.99 (2.73)
Completed program, n (%)	6 (19)	8 (17)	16 (40)	30 (25.6)

[a]BPS webSTAIR (Brief Peer-Supported web-based Skills Training in Affective and Interpersonal Regulation) includes a welcome module and 6 intervention modules. Module completion is presented on a scale of 0 to 7, where 7 indicates completion of the welcome module and 6 intervention modules.

Discussion

Principal Findings

The results of this study indicated that as compared to the waitlist, peer-supported webSTAIR is effective in improving symptoms of PTSD and depression, as well as emotion regulation and psychosocial functioning, with gains maintained at 8-week follow-up. The within-group pre- to posttreatment effect sizes for the webSTAIR participants ranged from 0.48 to 0.64, somewhat larger than those found for other peer-supported mHealth programs [18,21]. BPS webSTAIR offers a convenient and potentially more scalable option for veterans who have limited time available, lack access to a mental health provider, or prefer to work through the content in a self-guided manner.

XSL•FO
RenderX

The outcome varied depending on the number of modules completed. Among those who completed 3 or more modules, the mean reduction in PCL-5 score was over 10 points, and for the PHQ-8, it was nearly 5 points, reductions that are typically considered clinically meaningful [47,48]. Furthermore, our findings suggest that even partial completion of BPS webSTAIR (completing at least the welcome module but not past the second module) is associated with improved PTSD and depressive symptoms, emotion regulation, and psychosocial functioning relative to the waitlist. Therefore, the program may be of some benefit to patients who are unable to fully complete it due to low functional status or other demands on their time.

Overall, only about half (62/117, 53.0%) of the veterans engaged with peers, with the highest level of nonengagement (28/40, 70%) occurring in cohort 3. Analyses of chat data may provide insights about the role of peer engagement in mHealth. Based on our analyses of chat data from cohort 1, requiring a peer chat to move to the next module of the program was associated with a greater number of chat sessions and more messages exchanged, per module and in total. However, it did not contribute to program completion. In fact, the average number of modules completed in cohort 1 and the proportion of program completers was significantly lower than that in cohort 3. The introduction of a required engagement with the peer was intended to prompt the participants into peer engagement that otherwise might not have happened and provide an experience of contact that would lead to more engagement with the program and greater completion rates; however, this did not happen. Rather, the higher module completion rate within cohort 3 suggests that participants were motivated to work through the content even without support. Similar to this study, Possemato and colleagues [22] reported that 55% of veterans provided with the option to engage with a peer chose not to and suggested that some veterans might not need or want peer support. We speculate that a flexible, personalized approach to peer use leads to better program use and potentially better outcomes. The analyses indicated that while greater module completion was associated with the cohort with flexible peer contact, and increasing module completion was associated with a better outcome, the number of contacts (peer chats) was not associated with a better outcome in either cohort. This may be due to the potential moderating effects of patient preference regarding the use of peer support. For example, only patients who need and want peer support may benefit from it while others might do equally well or better without it. It would be valuable to identify the baseline characteristics of clients who might substantially benefit from peer support versus those who will not. In addition, it may be that clients who will be helped by peer support will experience greater benefits if the peer is proactive in reaching out to the client. In qualitative interviews, mental health coaches have pointed to the value of including proactive contact initiated by coaches rather than placing the onus on participants to engage [49]. In this study, the peers were available to the veterans to chat, but the veterans initiated the contact. Future efforts to enhance the impact of peers may involve modification regarding how proactive they are.

Limitations

This study has limitations. First, we relied fully on self-report measures administered by phone. Second, for ethical reasons, we did not maintain the waitlist group during the follow-up period and so were unable to conduct between-group analyses of change from pretreatment to follow-up. Third, this study design did not intentionally include an implementation variation regarding required versus optional engagement peers. As a result, the cohort analyses are post hoc and we have limited information about the veterans' experience. We do not know, for example, the average wait time for a chat, the range of wait time, or the frequency of contact with the research coordinator. Differences in these factors across the cohorts could have contributed to the observed differences in retention rates. Furthermore, chat transcripts were not accessible and thus we were limited in our ability to explain why more chat messages were not associated with better outcomes. It could be that the conversations were routine and pragmatic (eg, resolving technical issues) versus meaningful guidance about the content of webSTAIR. Lastly, participants were veterans recruited via social media advertisements which may limit generalization of findings beyond for veterans or persons who do not interact with social media.

Conclusions

This study adds to existing literature on mHealth by demonstrating that a brief, transdiagnostic, peer-supported, web-based program can provide moderate to large improvements in PTSD and depression symptoms as well as in emotion regulation and psychosocial functioning. Given the potential of peer-supported mHealth programs to expand access to evidence-based care for symptoms of PTSD and depression, future research should evaluate how best to deliver peer support, how much to deliver, and for whom the support benefits most. Future studies may also explore alternative formats for the delivery of peer support; for instance, proactive outreach from peer coaches or supplementary channels of communication (eg, phone contacts) to increase flexibility in ways in which a peer relationship can be built. The particular strengths of peer support for specific and varied aspects of mental health care such as treatment engagement, program completion, and symptom reduction need to be determined to develop maximally efficient and effective mental health services.

Acknowledgments

This study was funded by the National Center for Posttraumatic Stress Disorder.

XSL•FO
RenderX

Authors' Contributions

LEO was the data analyst and handled this paper's development. SS and AJ were project coordinators and worked on this paper's review and edits. DV was the project assessor and also worked on this paper's review and edits. JK carried out this paper's review and edits. KSM aided in the data analysis plan, and this paper's review and edits. MC was responsible for project conceptualization, data analysis plan, and this paper's review and edits.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 98 KB - jmir_v26i1e52130_app1.pdf]

References

1. Nichter B, Norman S, Haller M, Pietrzak RH. Psychological burden of PTSD, depression, and their comorbidity in the U.S. veteran population: suicidality, functioning, and service utilization. J Affect Disord 2019;256:633-640. [doi: 10.1016/j.jad.2019.06.072] [Medline: 31299445]
2. Ehring T, Quack D. Emotion regulation difficulties in trauma survivors: the role of trauma type and PTSD symptom severity. Behav Ther 2010;41(4):587-598. [doi: 10.1016/j.beth.2010.04.004] [Medline: 21035621]
3. Joormann J, Stanton CH. Examining emotion regulation in depression: a review and future directions. Behav Res Ther 2016;86:35-49. [doi: 10.1016/j.brat.2016.07.007] [Medline: 27492851]
4. Fried EI, Nesse RM. The impact of individual depressive symptoms on impairment of psychosocial functioning. PLoS One 2014;9(2):e90311 [FREE Full text] [doi: 10.1371/journal.pone.0090311] [Medline: 24587318]
5. Pietrzak RH, Johnson DC, Goldstein MB. Psychosocial buffers of traumatic stress, depressive symptoms, and psychosocial difficulties in veterans of operations enduring freedom and Iraqi freedom: the role of resilience, unit support, and postdeployment social support. J Affect Disord 2010;120(1-3):188-192. [doi: 10.1037/e717702011-019]
6. Buzza C, Ono SS, Turvey C, Wittrock S, Noble M, Reddy G, et al. Distance is relative: unpacking a principal barrier in rural healthcare. J Gen Intern Med 2011;26 Suppl 2(Suppl 2):648-654 [FREE Full text] [doi: 10.1007/s11606-011-1762-1] [Medline: 21989617]
7. Brooks E, Dailey NK, Bair BD, Shore JH. Listening to the patient: women veterans' insights about health care needs, access, and quality in rural areas. Mil Med 2016;181(9):976-981. [doi: 10.7205/MILMED-D-15-00367] [Medline: 27612340]
8. Cheney AM, Koenig CJ, Miller CJ, Zamora K, Wright P, Stanley R, et al. Veteran-centered barriers to VA mental healthcare services use. BMC Health Serv Res 2018;18(1):591 [FREE Full text] [doi: 10.1186/s12913-018-3346-9] [Medline: 30064427]
9. Brown NB, Bruce SE. Stigma, career worry, and mental illness symptomatology: factors influencing treatment-seeking for operation enduring freedom and operation iraqi freedom soldiers and veterans. Psychol Trauma: Theory, Res, Pract, Policy 2016;8(3):276-283. [doi: 10.1037/tra0000082] [Medline: 26390109]
10. Goode J, Swift JK. An empirical examination of stigma toward mental health problems and psychotherapy use in veterans and active duty service members. Mil Psychol 2019;31(4):335-345. [doi: 10.1080/08995605.2019.1630231]
11. Andrilla CHA, Patterson DG, Garberson LA, Coulthard C, Larson EH. Geographic variation in the supply of selected behavioral health providers. Am J Prev Med 2018 Jun;54(6 Suppl 3):S199-S207 [FREE Full text] [doi: 10.1016/j.amepre.2018.01.004] [Medline: 29779543]
12. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR mHealth uHealth 2020;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]
13. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
14. Simblett S, Birch J, Matcham F, Yaguez L, Morris R. A systematic review and meta-analysis of e-mental health interventions to treat symptoms of posttraumatic stress. JMIR Ment Health 2017;4(2):e14 [FREE Full text] [doi: 10.2196/mental.5558] [Medline: 28526672]
15. Wickersham A, Petrides PM, Williamson V, Leightley D. Efficacy of mobile application interventions for the treatment of post-traumatic stress disorder: a systematic review. Digit Health 2019;5:2055207619842986 [FREE Full text] [doi: 10.1177/2055207619842986] [Medline: 31019722]
16. Gratzer D, Khalid-Khan F. Internet-delivered cognitive behavioural therapy in the treatment of psychiatric illness. CMAJ 2016 Mar 01;188(4):263-272 [FREE Full text] [doi: 10.1503/cmaj.150007] [Medline: 26527829]
17. Lewis C, Roberts NP, Simon N, Bethell A, Bisson JI. Internet-delivered cognitive behavioural therapy for post-traumatic stress disorder: systematic review and meta-analysis. Acta Psychiatr Scand 2019;140(6):508-521. [doi: 10.1111/acps.13079] [Medline: 31359407]

XSL•FO
RenderX

18. Chinman M, George P, Dougherty RH, Daniels AS, Ghose SS, Swift A, et al. Peer support services for individuals with serious mental illnesses: assessing the evidence. Psychiatr Serv 2014;65(4):429-441. [doi: 10.1176/appi.ps.201300244] [Medline: 24549400]
19. Walker G, Bryant W. Peer support in adult mental health services: a metasynthesis of qualitative findings. Psychiatr Rehabil J 2013;36(1):28-34. [doi: 10.1037/h0094744] [Medline: 23477647]
20. Shalaby RAH, Agyapong VIO. Peer support in mental health: literature review. JMIR Ment Health 2020;7(6):e15572 [FREE Full text] [doi: 10.2196/15572] [Medline: 32357127]
21. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
22. Possemato K, Wu J, Greene C, MacQueen R, Blonigen D, Wade M, et al. Web-based problem-solving training with and without peer support in veterans with unmet mental health needs: pilot study of feasibility, user acceptability, and participant engagement. J Med Internet Res 2022;24(1):e29559 [FREE Full text] [doi: 10.2196/29559] [Medline: 35023846]
23. Lloyd-Evans B, Mayo-Wilson E, Harrison B, Istead H, Brown E, Pilling S, et al. A systematic review and meta-analysis of randomised controlled trials of peer support for people with severe mental illness. BMC Psychiatry 2014 Feb 14;14:39 [FREE Full text] [doi: 10.1186/1471-244X-14-39] [Medline: 24528545]
24. Blonigen DM, Harris-Olenak B, Kuhn E, Timko C, Humphreys K, Smith JS, et al. Using peers to increase veterans' engagement in a smartphone application for unhealthy alcohol use: a pilot study of acceptability and utility. Psychol Addict Behav 2021;35(7):829-839 [FREE Full text] [doi: 10.1037/adb0000598] [Medline: 32597665]
25. Pfeiffer PN, King C, Ilgen M, Ganoczy D, Clive R, Garlick J, et al. Development and pilot study of a suicide prevention intervention delivered by peer support specialists. Psychol Serv 2019;16(3):360-371 [FREE Full text] [doi: 10.1037/ser0000257] [Medline: 30382743]
26. Possemato K, Johnson EM, Emery JB, Wade M, Acosta MC, Marsch LA, et al. A pilot study comparing peer supported web-based CBT to self-managed web CBT for primary care veterans with PTSD and hazardous alcohol use. Psychiatr Rehabil J 2019;42(3):305-313 [FREE Full text] [doi: 10.1037/prj0000334] [Medline: 30489140]
27. Hobfoll SE, Blais RK, Stevens NR, Walt L, Gengler R. Vets prevail online intervention reduces PTSD and depression in veterans with mild-to-moderate symptoms. J Consult Clin Psychol 2016;84(1):31-42. [doi: 10.1037/ccp0000041] [Medline: 26322788]
28. Brady KT, Killeen TK, Brewerton T, Lucerini S. Comorbidity of psychiatric disorders and posttraumatic stress disorder. J Clin Psychiatry 2000;61 Suppl 7:22-32. [Medline: 10795606]
29. Cloitre M, Morabito D, Macia K, Speicher S, Froelich J, Webster K, et al. A home-based telehealth randomized controlled trial of skills training in affective and interpersonal regulation versus present-centered therapy for women veterans who have experienced military sexual trauma. J Consult Clin Psychol 2024;92(5):261-274. [doi: 10.1037/ccp0000872] [Medline: 38227462]
30. Weiss BJ, Azevedo K, Webb K, Gimeno J, Cloitre M. Telemental health delivery of skills training in affective and interpersonal regulation (STAIR) for rural women veterans who have experienced military sexual trauma. J Trauma Stress 2018;31(4):620-625. [doi: 10.1002/jts.22305] [Medline: 30070399]
31. Jain S, Ortigo K, Gimeno J, Baldor DA, Weiss BJ, Cloitre M. A randomized controlled trial of brief skills training in affective and interpersonal regulation (STAIR) for veterans in primary care. J Trauma Stress 2020;33(4):401-409. [doi: 10.1002/jts.22523] [Medline: 32506563]
32. Smith AM, Stewart K, Baul T, Valentine SE. Peer delivery of a brief cognitive-behavioral treatment for posttraumatic stress disorder: a hybrid effectiveness-implementation pilot study. J Clin Psychol 2020;76(12):2133-2154 [FREE Full text] [doi: 10.1002/jclp.23020] [Medline: 32632945]
33. Bauer A, Amspoker AB, Fletcher TL, Jackson C, Jacobs A, Hogan J, et al. A resource building virtual care programme: improving symptoms and social functioning among female and male rural veterans. Eur J Psychotraumatol 2021;12(1):1860357 [FREE Full text] [doi: 10.1080/20008198.2020.1860357] [Medline: 34025913]
34. Fletcher TL, Amspoker AB, Wassef M, Hogan JB, Helm A, Jackson C, et al. Increasing access to care for trauma-exposed rural veterans: a mixed methods outcome evaluation of a web-based skills training program with telehealth-delivered coaching. J Rural Health 2022;38(4):740-747. [doi: 10.1111/jrh.12628] [Medline: 34648188]
35. Cloitre M, Amspoker AB, Fletcher TL, Hogan JB, Jackson C, Jacobs A, et al. Comparing the ratio of therapist support to internet sessions in a blended therapy delivered to trauma-exposed veterans: quasi-experimental comparison study. JMIR Ment Health 2022;9(4):e33080 [FREE Full text] [doi: 10.2196/33080] [Medline: 35475777]
36. Prins A, Bovin MJ, Smolenski DJ, Marx BP, Kimerling R, Jenkins-Guarnieri MA, et al. The primary care PTSD screen for DSM-5 (PC-PTSD-5): development and evaluation within a veteran primary care sample. J Gen Intern Med 2016;31(10):1206-1211 [FREE Full text] [doi: 10.1007/s11606-016-3703-5] [Medline: 27170304]
37. Kroenke K, Spitzer RL, Williams JBW. The patient health questionnaire-2: validity of a two-item depression screener. Med Care 2003;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
38. Gray MJ, Litz BT, Hsu JL, Lombardo TW. Psychometric properties of the life events checklist. Assessment 2004;11(4):330-341. [doi: 10.1177/1073191104269954] [Medline: 15486169]

XSL•FO
RenderX

39. Weathers FW, Litz BT, Keane TM, Palmieri PA, Marx BP, Schnurr PP. The PTSD Checklist for DSM-5 (PCL-5). US Department of Veterans Affairs. 2013. URL: https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp [accessed 2024-08-08]
40. Bovin MJ, Marx BP, Weathers FW, Gallagher MW, Rodriguez P, Schnurr PP, et al. Psychometric properties of the PTSD checklist for diagnostic and statistical manual of mental disorders-fifth edition (PCL-5) in veterans. Psychol Assess 2016;28(11):1379-1391. [doi: 10.1037/pas0000254] [Medline: 26653052]
41. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
42. Shin C, Lee SH, Han KM, Yoon HK, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. Psychiatry Investig 2019;16(4):300-305 [FREE Full text] [doi: 10.30773/pi.2019.02.01] [Medline: 31042692]
43. Bjureberg J, Ljótsson B, Tull M. Difficulties in emotion regulation scale-16. PsycTESTS Dataset 2016. [doi: 10.1037/t57299-000]
44. Bjureberg J, Ljótsson B, Tull MT, Hedman E, Sahlin H, Lundh L, et al. Development and validation of a brief version of the difficulties in emotion regulation scale: the DERS-16. J Psychopathol Behav Assess 2016;38(2):284-296 [FREE Full text] [doi: 10.1007/s10862-015-9514-x] [Medline: 27239096]
45. Mundt JC, Marks IM, Shear MK, Greist JH. The work and social adjustment scale: a simple measure of impairment in functioning. Br J Psychiatry 2002;180:461-464. [doi: 10.1192/bjp.180.5.461] [Medline: 11983645]
46. Pedersen G, Kvarstein EH, Wilberg T. The work and social adjustment scale: psychometric properties and validity among males and females, and outpatients with and without personality disorders. Personal Ment Health 2017;11(4):215-228. [doi: 10.1002/pmh.1382] [Medline: 28681505]
47. Kuhn E, Kanuri N, Hoffman JE, Garvert DW, Ruzek JI, Taylor CB. A randomized controlled trial of a smartphone app for posttraumatic stress disorder symptoms. J Consult Clin Psychol 2017;85(3):267-273. [doi: 10.1037/ccp0000163] [Medline: 28221061]
48. Kroenke K. Enhancing the clinical utility of depression screening. CMAJ 2012;184(3):281-282 [FREE Full text] [doi: 10.1503/cmaj.112004] [Medline: 22231681]
49. Bradley T, Hansen V, Wye P, Campbell E, Bartlem K, Reid K, et al. Telephone-delivered health behaviour change support for people with a mental health condition: the coaches' perspective. BMC Health Serv Res 2021;21(1):1130 [FREE Full text] [doi: 10.1186/s12913-021-07126-4] [Medline: 34670561]

Abbreviations

BPS: Brief Peer-Supported
DERS-16: Difficulties in Emotion Regulation Scale-16 item
ITT: intention-to-treat
LEC: Life Events Checklist
mHealth: mobile health
PCL-5: Posttraumatic Stress Disorder Checklist for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
PHQ-2: Patient Health Questionnaire-2 item
PHQ-8: Patient Health Questionnaire-8 item
PHQ-9: Patient Health Questionnaire-9 item
PTSD: posttraumatic stress disorder
STAIR: Skills Training in Affective and Interpersonal Regulation
VA: Veteran Affairs
webSTAIR: web-based Skills Training in Affective and Interpersonal Regulation
WSAS: Work and Social Adjustment Scale

Edited by K Williams; submitted 24.08.23; peer-reviewed by G Reger, A AL-Asadi, S Li, D Liu; comments to author 05.02.24; revised version received 15.04.24; accepted 09.07.24; published 02.10.24.

Please cite as:
Ong LE, Speicher S, Villasenor D, Kim J, Jacobs A, Macia KS, Cloitre M
Brief Peer-Supported Web-Based Skills Training in Affective and Interpersonal Regulation (BPS webSTAIR) for Trauma-Exposed Veterans in the Community: Randomized Controlled Trial
J Med Internet Res 2024;26:e52130
URL: https://www.jmir.org/2024/1/e52130
doi:10.2196/52130
PMID:39012722

XSL•FO
RenderX

©Laura E Ong, Sarah Speicher, Diana Villasenor, Jamie Kim, Adam Jacobs, Kathryn S Macia, Marylene Cloitre. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Behavioral Change–Based Mobile Intervention for Promoting Regular Physical Activity in Medical Rehabilitation Maintenance of Patients With Coronary Artery Disease: Controlled Trial

Melina Waranski[1,2], MA; René Garbsch[1,2], MSc; Mona Kotewitsch[1,2], MSc; Marc Teschler[1,2], Dr Rer Biol Hum; Boris Schmitz[1,2], Dr Rer Nat; Frank C Mooren[1,2], Prof Dr Med

[1]Department of Rehabilitation Sciences, Faculty of Health, University of Witten/Herdecke, Witten, Germany
[2]DRV Clinic Königsfeld, Center for Medical Rehabilitation, Ennepetal, Germany

Corresponding Author:
Boris Schmitz, Dr Rer Nat
Department of Rehabilitation Sciences
Faculty of Health
University of Witten/Herdecke
Alfred-Herrhausen-Straße 50
Witten, 58455
Germany
Phone: 49 23339888156
Email: boris.schmitz@uni-wh.de

Abstract

Background: Cardiac rehabilitation is known to reduce coronary artery disease (CAD) severity and symptoms, but adoption of a healthy postrehabilitation lifestyle remains challenging. Innovative eHealth solutions could help, but behavioral change–based eHealth maintenance programs for patients with CAD are scarce. RehaPlus+ aims to improve postrehabilitation outcomes with a personalized eHealth intervention built on behavioral change concepts emphasizing healthy lifestyle changes, especially regular physical activity (PA).

Objective: This study aims to evaluate the effectiveness of the personalized eHealth program RehaPlus+ for promoting regular PA against usual care.

Methods: A total of 169 patients with CAD who had undergone stent implantation or bypass surgery were recruited after completing center-based phase II rehabilitation. They were then divided, without blinding, into 2 groups using a quasi-experimental approach: a case manager–assisted 24-week eHealth program (RehaPlus+; n=84) and a conventional physician-assisted outpatient program (usual care; n=85). The study was designed as a noninferiority trial. RehaPlus+ participants received motivational messages twice weekly for 6 months, and the usual care group engaged in a 6-month outpatient program (twenty-four 90-minute strength and endurance training sessions). The primary outcomes, evaluated using the self-assessed Bewegungs- und Sportaktivität questionnaire, were regular PA (≥150 min/wk) and weekly activities of daily living (ADLs) 6 months after rehabilitation. Secondary outcomes involved PA during work and floors climbed weekly (measured by Bewegungs- und Sportaktivität questionnaire), psychological well-being (assessed by the 5-item World Health Organization Well-Being Index), cardiac self-efficacy, health-related quality of life (measured by the 36-Item Short Form Survey), and work ability (using the Work Ability Index).

Results: Data of 105 patients (RehaPlus+: n=44, 41.9%; usual care: n=61, 58.1%; male patients: n=80, 76.2%; female patients: n=25, 23.8%; mean age 56.0, SD 7.3 years) were available at the 6-month follow-up. At 6 months after discharge from phase II cardiac rehabilitation, the RehaPlus+ group exhibited 182 (SD 208) minutes per week of PA and the usual care group exhibited 119 (SD 175) minutes per week of PA (P=.15), with no interaction effect (P=.12). The RehaPlus+ group showed an ADL level of 443 (SD 538) minutes per week compared to the usual care group with 308 (SD 412) minutes per week at the 6-month follow-up, with no interaction effect (P=.84). The differences observed in PA and ADL levels between the RehaPlus+ and usual care groups were within the predefined 1-sided noninferiority margin, indicating that the RehaPlus+ intervention is not inferior to usual care based on these outcomes. There were no differences between the groups for all secondary outcomes (P>.05).

Conclusions: RehaPlus+ is not inferior to the usual care program, as both groups improved PA and ADLs to a similar extent. These findings emphasize the potential of eHealth interventions to assist in maintaining healthy lifestyles after rehabilitation.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT06162793; https://clinicaltrials.gov/study/NCT06162793

(J Med Internet Res 2024;26:e56480) doi:10.2196/56480

KEYWORDS

rehabilitation; eHealth; mobile health; mHealth; telemedicine; cardiovascular disease; behavioral change; mobile phone

Introduction

Background

Coronary artery disease (CAD) is a chronic disease without a definitive cure; however, its consequences in terms of restricted physical performance, health-related quality of life (QoL), and perceived well-being can be minimized. Cardiac rehabilitation (CR) is an evidence-based and class IA–recommended therapy [1] provided to minimize the physiological and psychological impact of CAD, decrease morbidity and mortality rates, and enhance physical performance. CR constitutes a complex approach with a strong focus on physical exercise [2] and cardiovascular risk factor reduction, which have been shown to exert the most significant influence on the effectiveness and success of CR [3,4].

CR can be categorized into 3 distinct stages (Figure 1) [5]. The initial stage, phase I CR, typically takes place in acute care clinics, often immediately following a coronary intervention or surgery. Patients receive education about their health condition and risk factors, with an emphasis on early mobilization and moderate physical activity (PA) [6]. Phase II CR, the reconditioning phase, takes place in inpatient or outpatient CR centers [7], focusing on patient education, supervised exercise training, diet, smoking cessation, and psychological support [8] to reduce cardiovascular risk, enhance exercise capacity, and support personal health management. Successful completion of phase II CR reduces mortality and morbidity risks [9] and restores the ability to work and engage in social activities [10]. Phase III CR, the maintenance phase, emphasizes lifelong self-care, risk factor management, and regular (self-organized) PA [6]. The successful transition from phase II to phase III CR, including the implementation of healthy lifestyle habits are of paramount significance for the long-term health of patients with CAD [5].

Figure 1. Stages of cardiac rehabilitation (CR). In phase III CR, the usual care program can be prescribed as an outpatient program to patients in Germany. Nearby (<30 km) usual care centers are not always available, leading to many individuals without postrehabilitation program alternatives. eHealth programs such as RehaPlus+ may bridge this gap by providing a readily available alternative for patients with coronary artery disease (CAD).

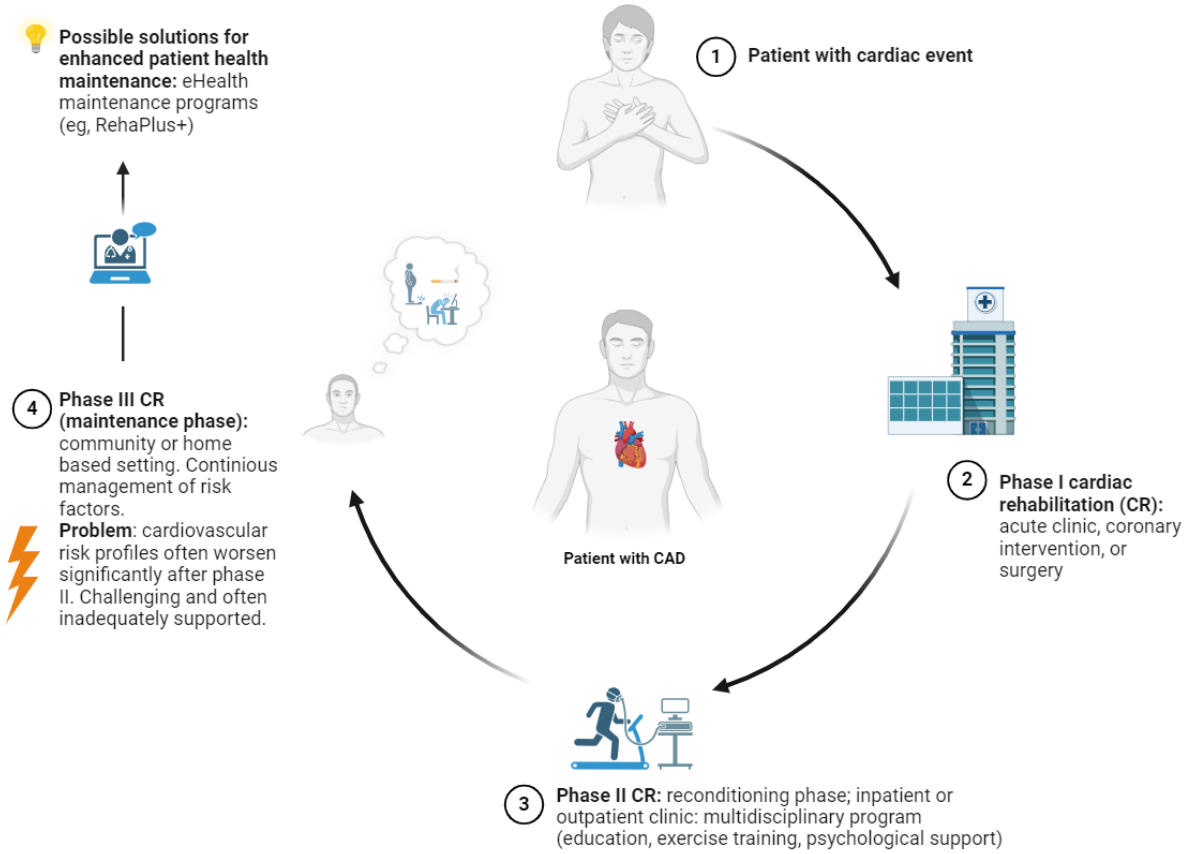

However, and despite the general effectiveness of phase II CR [11], cardiovascular risk profiles often deteriorate significantly thereafter [12]. This effect has been attributed to the fact that maintaining a healthy lifestyle, including regular PA, is

XSL•FO
RenderX

challenging for most patients, and adequate support is often not available [13]. Existing maintenance programs are largely affected by high monetary expenses for lifelong support if provided by general practitioners and local cardiologists [14,15]. On the patient's side, time constraints and travel limitations, impeding participation due to work and family commitments, and general lack of individualization may reduce adherence to center-based care. In addition, during the COVID-19 pandemic, on-site maintenance programs were largely unavailable [16].

Current eHealth solutions for CR maintenance are spurred by innovative technologies and the growing prevalence of mobile devices among patients. The general momentum toward digitalization, including the expansion of mobile data transfer infrastructure, aligns with the concept of mobile health care. This evolution presents significant prospects for enhanced patient health maintenance, as eHealth applications possess the capacity to amplify rehabilitation effectiveness and to sustain patient support postdischarge [17,18]. Prior studies have already provided evidence that electronic communication and health information technology in the form of eHealth may represents an effective alternative to phase II CR [19-21], and a recent meta-analysis on the use of eHealth in phase III CR maintenance suggested that eHealth based on behavior change techniques (BCTs) may assist patients with CAD to achieve improved health outcomes [22]. Most previous eHealth studies in the recent meta-analysis have predominantly used SMS text messages and have not explicitly focused on incorporating BCTs. Hence, this study represents a novelty in this regard, as it is based on behavioral change theories and BCTs, emphasizing the use of individualized messages tailored to the patient's life situation. A recent study found that a telehealth program may effectively support both behavioral and emotional recovery following a cardiac event [23].

Objectives

As increasing PA is one of the major goals in phase III CR that has been associated with a 27% to 35% decrease in cardiovascular mortality [24], this study aimed to evaluate the effectiveness of an individualized, message-based, eHealth maintenance concept (RehaPlus+) for the motivation of patients with CAD toward increased PA. Our hypothesis was that RehaPlus+ would be equally effective as the German usual care in form of a center-based maintenance program (Individualized Rehabilitation Aftercare in Post-Acute Treatment; IRENA) in terms of supporting regular PA and activities of daily living (ADLs) as well as \ improvement in psychological well-being, cardiac self-efficacy (CSE), health-related QoL, and work ability assessed 6 months after discharge from phase II CR.

Methods

Study Design and Participants

Overview

To compare the effects of the eHealth maintenance program RehaPlus+ to the German usual care in the form of a center-based maintenance program (IRENA, provided by the German pension fund), a quasi-experimental study with a comparison at 6-month follow-up (post phase II CR) was performed (Clinical Trials NCT06162793) in accordance with the eHealth CONSORT (Consolidated Standards of Reporting Trials) guidelines (Multimedia Appendix 1). RehaPlus+ has been developed by the Clinic Königsfeld to provide a multimodal eHealth-based aftercare program to patients with CAD as an alternative to the standard rehabilitation aftercare concept. Main outcomes assessed were PA and ADLs. Data for both groups were collected at 2 time points: at the beginning of inpatient CR (baseline; defined as T_0) and at follow-up (24 weeks after discharge; defined as T_1). Both groups transitioned to phase III CR within 1 or 2 weeks after discharge from inpatient CR.

Eligibility Criteria

The study enrolled patients with documented CAD during inpatient CR (Figure 2). The patients were recruited on-site at Clinic Königsfeld during phase II CR. Enrollment window was within 6 weeks of a cardiac event or intervention, including ST-elevation myocardial infarction and non–ST-elevation myocardial infraction, and stent implantation, and bypass surgery and their combinations. Patients had to declare readiness for behavioral change during phase II CR (see Group Allocation section for details). The patients were required to have the necessary computer and internet skills (assessed in a one-on-one interview) and possess a smartphone. Patients with significant language barriers were not eligible to participate.

XSL•FO
RenderX

Figure 2. Study design. Patients were recruited at the beginning of the 3-week inpatient cardiac rehabilitation (baseline, T0) and participated in the RehaPlus+ or usual care program, depending on local availability after discharge (24-week maintenance phase). Only patients indicating willingness to change were included. The RehaPlus+ and usual care group attended 2 group seminars and 1 face-to-face individual counseling session in phase II cardiac rehabilitation (CR). In phase III CR, RehaPlus+ recipients received 3 customized motivational text messages weekly for 6 months, while the usual care group engaged in a 6-month outpatient program involving 24 sessions of 90-minute strength and endurance training. Follow-up examinations were conducted 24 weeks after discharge (T1). CAD: coronary artery disease.

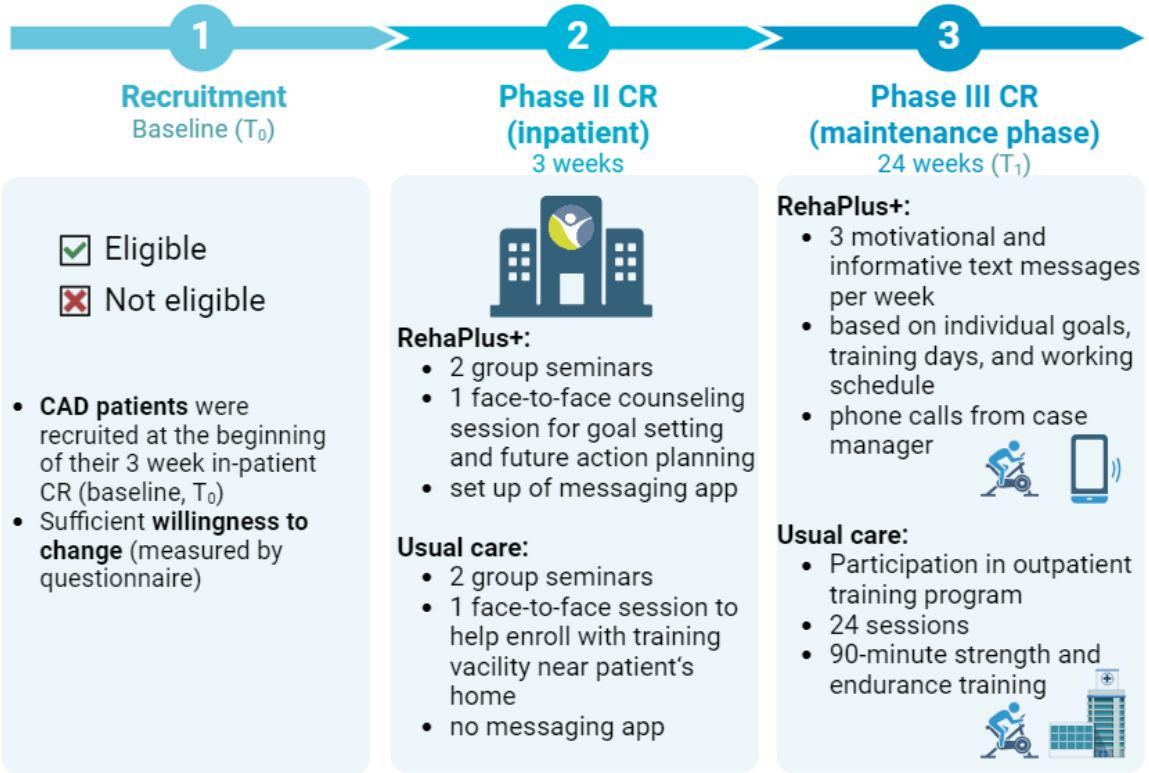

Ethical Considerations

The study complied with the Helsinki Declaration on Ethical Principles for Medical Research Involving Human Subjects, received approval by the ethics committee of University Witten/Herdecke (#91/2018), and was performed at the medical rehabilitation center Clinic Königsfeld in Germany. Before enrollment, all participants provided written informed consent. The patients had the option to withdraw from participation at any time. The data were deidentified, with patients being assigned numerical identifiers. No compensation was provided to the patients for participation.

Group Allocation

A quasi-experimental design, as described in the study by Axelrod and Hayward [25], was chosen due to practical and ethical considerations, given that IRENA is regarded as an effective aftercare concept. This approach enables the examination within the constraints of a real-world setting, where full random allocation is not viable.

Both groups (RehaPlus+ and usual care) participated in 2 group seminars during phase II CR (Figure 2). The patients were invited to the initial group seminars based on the transtheoretical model (TTM) of behavior change [26-28], where intentional health-relevant behavior change is categorized into different (subsequent) stages (precontemplation, contemplation, preparation, action, maintenance, and termination). As action-oriented programs are unsuitable for individuals in stage 1 or 2 [28], only patients in at least stage 3 were included. As RehaPlus+ was an experimental program that could not be actively requested by patients, patients with CAD eligible to participate were screened between September 2021 and May 2023 using a questionnaire based on the TTM to determine their readiness for behavioral change [29]. Readiness for behavioral change was congruently accepted by participation in the usual care program, as patients needed to actively request participation by making an appointment with an assistant and connecting with the providing center.

For the individual counseling sessions, both patient groups registered voluntarily. Patients were then assigned to either usual care or the RehaPlus+ program based on the availability of the program in their residential area (30 km radius). Patients with no access to a local center or those who were unable to participate in the program due to occupational reasons were assigned to the RehaPlus+ program. The key difference between the groups was that RehaPlus+ participants received additional support through an app and phone calls in phase III CR, which included personalized messages and action planning, while usual care took part in the usual care program.

Intervention (RehaPlus+)

Overview

The group seminars were rooted in various psychological theories of health behavior change [27,30,31] and concepts of self-efficacy for promoting healthy habits [32]. The initial 60-minute seminar (≤15 participants) focused on a health

XSL•FO
RenderX

psychological intervention and aimed at elevating participant's self-efficacy expectation through action planning, including setting of goals and creating strategies to achieve them. The importance of health-promoting lifestyle changes, including regular PA and ADLs, was explained. Participants outlined their health goals and motivations for altering their lifestyle into a structured action plan. During a subsequent 30-minute seminar, coping strategies tailored to distinct circumstances of each participant were discussed. The health psychology group seminars aimed to empower patients to take independent action and enable them to autonomously pursue and achieve their health-related goals. During the individual face-to-face counseling session, participants received their individual access to the application and were guided through the process of installing the messenger app on their smartphone, and a comprehensive overview of the associated procedures was provided. For individualization of prompts and messages, habits, identified problems or challenges, work commitments, shift schedules, and aspirations were documented and aligned with the patient's action plan, emphasizing on its feasibility and practicality.

At the end of phase II CR, the eHealth intervention started for a period of 6 months. The RehaPlus+ participants began receiving motivational and informational messages (3 per week using a mobile phone app; described in the App and Messaging section) tailored to their individual objectives, training schedule, and working commitments. Participants were contacted by their designated case manager at 2 and 5 months via telephone to address any technical challenges and barriers encountered in incorporating physical exercise into their daily routine.

App and Messaging

An app was developed for the unidirectional delivery of informative, educational, and motivational messages (Figure 3). The messages were sent using end-to-end encryption technology, and patients used a pseudonym to register. No personal data were exchanged via the app. Messages were delivered 3 times per week, with a total of 72 messages within 24 weeks.

Figure 3. Sample message. Each RehaPlus+ patient received a total of 72 messages over a period of 24 weeks.

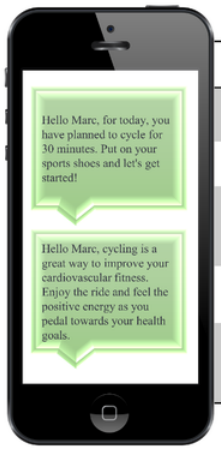

Application (24 weeks utilization)	
Content	Informative, educational, motivational
Frequency	3 x/week
Message tailoring	Tailored to the patient based on their profile, rehabilitation goals, and predetermined activities.
Message timing	Timing was customized to align with the patient's work schedule.
Purpose	Ongoing support and guidance during the rehabilitation journey.

The creation process of the messages followed a Delphi approach, involving iterative feedback loops with patients (n=56). Initially, the case manager developed message templates based on common activities. Subsequently, through rounds of anonymous feedback from a panel of patients (n=5), adjustments were made to ensure the messages resonated with their preferences and needs.

Building upon insights gained through the Delphi method, it was identified where individualization would be most beneficial. Messages were then customized accordingly, tailoring them to each patient's specific needs, defined activities, scheduled timing, and considerations such as working hours. Each message was crafted to focus on specific rehabilitation objectives and suitable PAs. This personalized approach ensured that messages were delivered at optimal times to align with the patient's schedule, providing continuous encouragement and guidance throughout the rehabilitation process. Messages were intended to provide continuous encouragement and guidance throughout, delivering nudges and support. As an example, a message reads as follows: "Hello [pseudonym], for today, you have planned to pedal for 30 minutes. Put on your sports shoes and let's get started! Remember, cycling is a great way to improve your cardiovascular fitness and strengthen your leg muscles. Keep up the fantastic work!" The back-end of the system offered a multimessaging functionality, allowing case managers to prepare predefined messaging queues, which were then triggered automatically.

Usual Care (IRENA)

In Germany, the phase II CR center assists the individual patient in finding suitable postrehabilitation facilities offering the usual care program. It is essential that these recommended postrehabilitation services are located within a convenient distance from the patient's residence (<30 km) [33]. The usual care program IRENA is a multimodal postrehabilitation program that comprises 24 sessions conducted in small groups with a maximum of 10 participants and begins the end of phase II CR for a period of 6 months. Each session includes a 90-minute combination of strength and endurance training as well as information, motivation, and education, as described in the specifications of the German pension fund [34]. The program includes both an initial assessment and a concluding discussion.

XSL•FO
RenderX

Blinding

Due to the nature of the intervention, blinding was not possible.

Outcomes

Primary Study Outcomes

Primary study outcomes were defined as self-reported PA and ADLs at 24 weeks (6 months) after discharge, assessed by questionnaire at 4 weeks before rehabilitation (baseline) and 6-month follow-up (see the Questionnaires section). Secondary study outcomes included PA at work and floors climbed per week at 6-month follow-up as well as change in psychological well-being, CSE, health-related QoL, and work ability from 4 weeks before rehabilitation to 6-month follow-up, all assessed by a questionnaire.

Questionnaires

PA and ADLs (including PA at work and floors climbed per week) were assessed by the validated German Bewegungs- und Sportaktivität questionnaire [35]. Health-related QoL was assessed through the validated German version of the 36-item Short Form Survey (SF-36; 4-week version; Cronbach α=0.87-0.89) following the guidelines of the RAND Corporation. The SF-36 physical and mental component subscores (physical component summary [PCS] and mental component summary [MCS]) were calculated according to the study by Ware et al [36]. Subjective psychological well-being was assessed using the validated 5-item World Health Organization Well-Being Index (Cronbach α=0.92) [37]. Work ability was evaluated using the validated [38] Work Ability Index (Cronbach α=0.58-0.77) [39]. CSE was measured using the validated CSE Scale (Cronbach α=0.87-0.90) [40]. All outcomes were self-assessed through web-based questionnaires.

Statistical Analyses

Statistical analyses were performed using SPSS (version 28; IBM) and GraphPad Prism (version 10; GraphPad Software). Constant variables are expressed as mean (SD), median (range), or 95% CI as indicated. Categorical variables are presented as n (%). The normal distribution was statistically and graphically tested via Kolmogorov - Smirnov test. Differences between groups over time (RehaPlus+ vs usual care) were analyzed using 2-way ANOVA. Multivariate ANOVA with 2 measurement time points, with gender and group as between-subject factors, was conducted to assess the difference between women and men. Categorical variables (gender, education, occupation, and marital status) were analyzed using chi-square test. Data were analyzed for outliers using the robust regression followed by outlier identification option under GraphPad Prism. Data points falling outside the 1% threshold were considered outliers and potentially excluded from the analysis. Cohen *d* was used to express effect sizes. Power calculation was performed using G*Power (version 3.1.9; G-Power). Sample size was calculated based on unpublished data on PA in patients with CAD assessed by Bewegungs- und Sportaktivität questionnaire with an effect size of d=0.4 and a noninferiority limit of half the SD, suggesting that a sample size of 100 participants (50 per group) would result in a statistical power of $1-\beta=0.80$ at $\alpha=.05$. A noninferiority comparison between the RehaPlus+ group and the usual care group was conducted using a predefined margin, assessed by calculating the 95% CI around the primary end point. The noninferiority margin was set at half the SD of 87.5 of the primary end point to represent a clinically acceptable difference between the new intervention and the established standard therapy. According to this approach, a difference less than this noninferiority limit between the groups is considered irrelevant (not meaningful) [41]. The 1-sided interpretation of results was based on the meaningful differences. Responder analysis was performed for PA, ADLs, PA at work, and floors climbed per week, as described using the typical error (TE) method and the following equation: $TE=SD_{diff}/\sqrt{2}$, where SD_{diff} is calculated as the difference between the variance (SD) of 2 repeated measures [42]. Responders were defined as participants who demonstrated an increase greater than 2×TE away from 0. Statistical significance was accepted at $P<.05$.

Results

General

A total of 1258 eligible patients with CAD were screened and 169 (13%) were included (refer to the flowchart in Figure 4). Participants' baseline characteristics are presented in Table 1. Following a 3-week period of inpatient CR, patients transitioned to the long-term follow-up phase. After 6 months, 41% (31/75) of the RehaPlus+ group and 27% (22/83) of the usual care group were lost to follow-up (P=.07) mainly due to contacting problems for follow-up examinations. The final analysis included 105 patients with assessments at both time points (RehaPlus+, n=44; usual care, n=61). During the study, a total of 72 individual messages was sent to each patient in the RehaPlus+ group as planned (n=72, 100%). In general, patients indicated that the messages were motivating and that they served as helpful reminders for performing regular PA. In addition, the handling of the messaging system was found to be straightforward and user-friendly by the case managers, further contributing to its efficiency in supporting patient engagement.

XSL•FO
RenderX

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) flowchart. A total of 169 patients referred to either a 24-week eHealth group (RehaPlus+, n=84) or a conventional center-based program (usual care, n=85). A total of 11 patients (RehaPlus+, n=9; usual care, n=2) did not complete the baseline assessment due to premature discharge. A total of 31 patients and 22 patients were lost to follow-up due to contacting problems for RehaPlus+ and for usual care, respectively. The final analysis included 105 patients (RehaPlus+, n=44; usual care, n=61). IRENA: Individualized Rehabilitation Aftercare in Post-Acute Treatment.

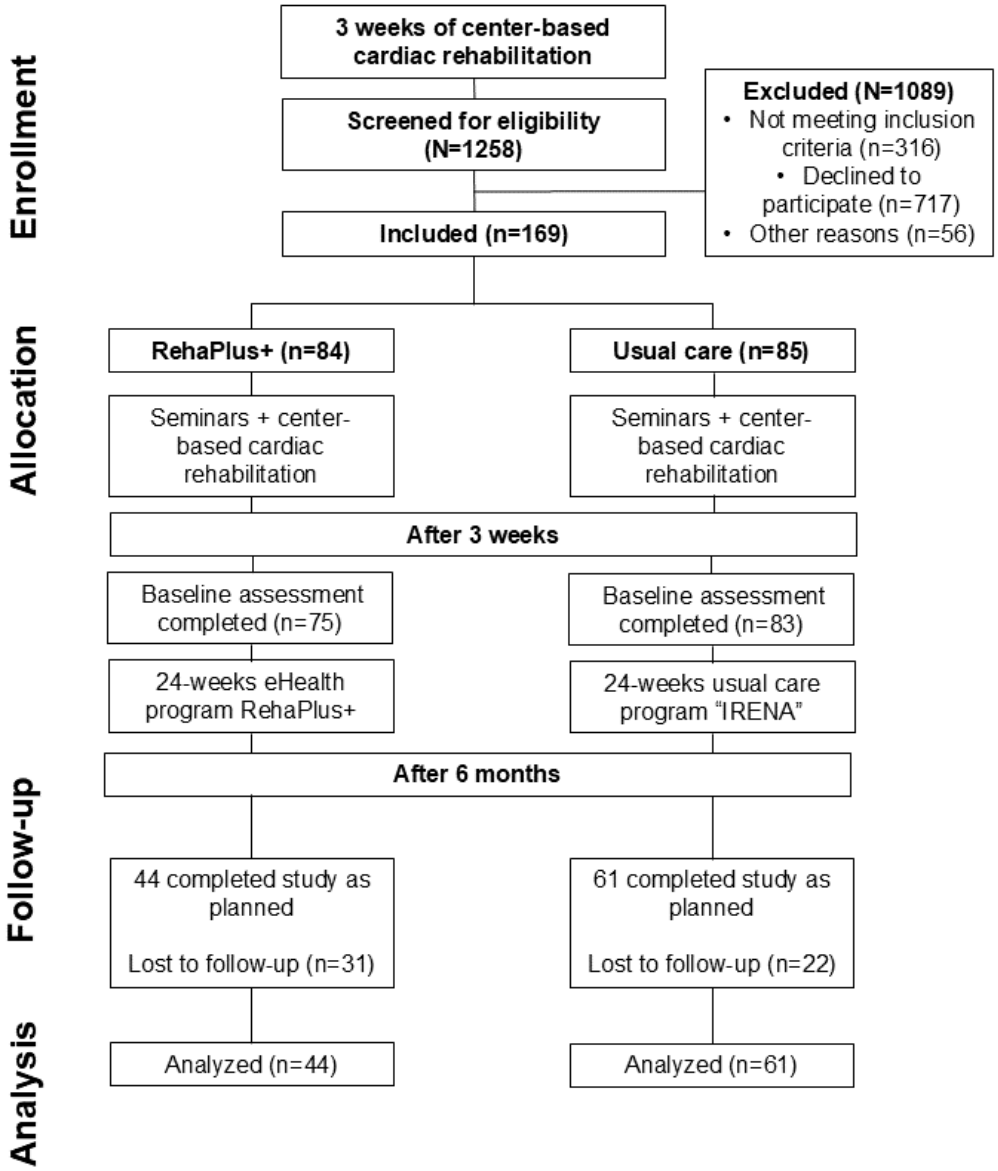

XSL•FO
RenderX

Table 1. Patient characteristics[a].

Characteristics	Overall (n=105)	RehaPlus+ (n=44)	Usual care (n=61)	*P* value
Anthropometric data				
Age (y), mean (SD)	54.9 (7.2)	53.6 (8.1)	56.2 (6.2)	.06
Sex, n (%)				
Female	25 (24)	12 (27)	13 (21)	.47
Male	80 (76)	32 (73)	48 (79)	__[b]
Height (cm), mean (SD)	177.5 (9.1)	179.3 (9.1)	176.0 (8.9)	.06
Weight (kg), mean (SD)	94.1 (18.2)	98.2 (16.6)	90.7 (18.9)	*.03* [c]
BMI (kg/m^2), mean (SD)	29.8 (5.3)	30.6 (5.2)	29.2 (5.4)	.14
Education[d], n (%)				
Less than high school	73 (94)	22 (67)	51 (94)	*.002*
Greater than or equal to high school	13 (6)	10 (33)	3 (6)	—
Occupation[e], n(%)				
Worker	22 (22)	8 (19)	14 (25)	.53
Employee	71 (71)	33 (77)	38 (67)	.27
Self-employed	5 (5)	1 (2)	4 (7)	.15
Retiree or unemployed	2 (2) 0 (0)	1 (2) 0 (0)	1 (1) 0 (0)	.84
Marital status[f], n (%)				
Single	24 (24)	9 (21)	15 (26)	.52
Married	63 (62)	27 (61)	36 (63)	.37
Divorced	12 (12)	7 (16)	5 (9)	.08
Widowed	2 (2)	1 (2)	1 (2)	.85
Clinical data				
Diseases of the circulatory system				
Coronary artery disease, n (%)				
1-vessel disease	31 (30)	13 (30)	18 (30)	.84
2-vessel disease	39 (37)	18 (41)	21 (34)	.81
3-vessel disease	35 (33)	13 (30)	22 (36)	.62
ST-segment elevation myocardial infarction or non–ST-segment elevation myocardial infarction, n (%)	59 (56)	26 (59)	33 (54)	.63
Cardiac arrhythmia, n (%)	6 (6)	4 (9)	2 (3)	.29
Arterial hypertension, n (%)	78 (74)	32 (72)	46 (75)	.55
Pulmonary embolism, n (%)	6 (5)	3 (7)	3 (5)	.53
Endocrine, nutritional, or metabolic diseases, n (%)				
Obesity	28 (24)	13 (25)	15 (24)	.94
Type 2 diabetes mellitus	16 (15)	4 (9)	12 (19)	.18
Other	75 (71)	33 (75)	42 (68)	.59
(Hypo- and hyperthyroidism, n (%)	10 (9)	6 (11)	4 (6)	.36
Diseases of the musculoskeletal system and connective tissue, n (%)	20 (19)	9 (20)	11 (18)	.98
Depressive and adjustment disorders, n (%)	18 (17)	8 (18)	10 (16)	.99
Medication, n (%)				

XSL•FO
RenderX

Characteristics	Overall (n=105)	RehaPlus+ (n=44)	Usual care (n=61)	*P* value
Angiotensin converting enzyme inhibitor	54 (47)	22 (42)	32 (51)	.32
Statin	102 (97)	42 (95)	60 (98)	.15
β-Blocker	94 (89)	36 (82)	56 (91)	.36
Angiotensin-II receptor blocker	37 (32)	18 (34)	19 (30)	.67
Calcium channel blocker	22 (19)	8 (15)	14 (22)	.33
Anticoagulant	103 (98)	42 (95)	61 (100)	.45
Antiarrhythmic	1 (1)	1 (2)	0 (0)	.27
Diuretic	41 (39)	22 (50)	19 (31)	*.03*
Analgesic	13 (11)	6 (11)	7 (11)	.97
Antidepressant	5 (5)	2 (5)	3 (5)	.53
Diabetes medication	18 (17)	5 (11)	13 (21)	.12

[a]Between-group comparison was performed using unpaired 2-sided *t* test or chi-square test.

[b]Not applicable.

[c]Italicization denotes statistical significance.

[d]A total of 19 participants did not provide their educational level (RehaPlus+ =12; usual care=7).

[e]A total of 5 participants did not provide their occupation level (RehaPlus+ =1; usual care=4).

[f]A total of 4 participants did not provide their marital status (RehaPlus+ =0; usual care=4).

Primary Outcome

Physical Activity

At 6 months after discharge from phase II CR, the RehaPlus+ group exhibited 182 (SD 208) minutes of PA per week, while the usual care group exhibited 119 (SD 175) minutes of PA per week (*P*=.12). To ensure statistical robustness and impartiality, outliers were removed (n=2), without any effect on the observed results (*P*=.10). Over time (T_0 to T_1) both groups combined showed a significant increase in PA per week, with an average increase of 92 (SD 211) minutes per week from prerehabilitation to 6-month follow-up (Figure 5A; *P*=.001). There were no significant effects of gender over time in terms of change in PA levels (*P*=.10). Notably, a responder analysis for PA revealed that 70% (31/44) of the RehaPlus+ participants were responders compared to 48% (29/61) of the usual care participants.

XSL•FO
RenderX

Figure 5. Changes in primary and secondary outcome measures over 6 months. All variables were assessed by questionnaire at T0 (start of phase II cardiac rehabilitation) and after 24 weeks. Overall, a significant increase at 6-month follow-up compared to baseline was detected in physical activity (*P*=.001), activities of daily living (*P*=.006), physical activity during work (*P*=.03), floors climbed weekly (*P*=.02), well-being (*P*=.001), and cardiac self-efficacy (*P*=.001) with no significant difference between groups. Data are presented as mean (SD). Differences between groups over time (RehaPlus+ vs usual care) were analyzed using 2-way ANOVA.

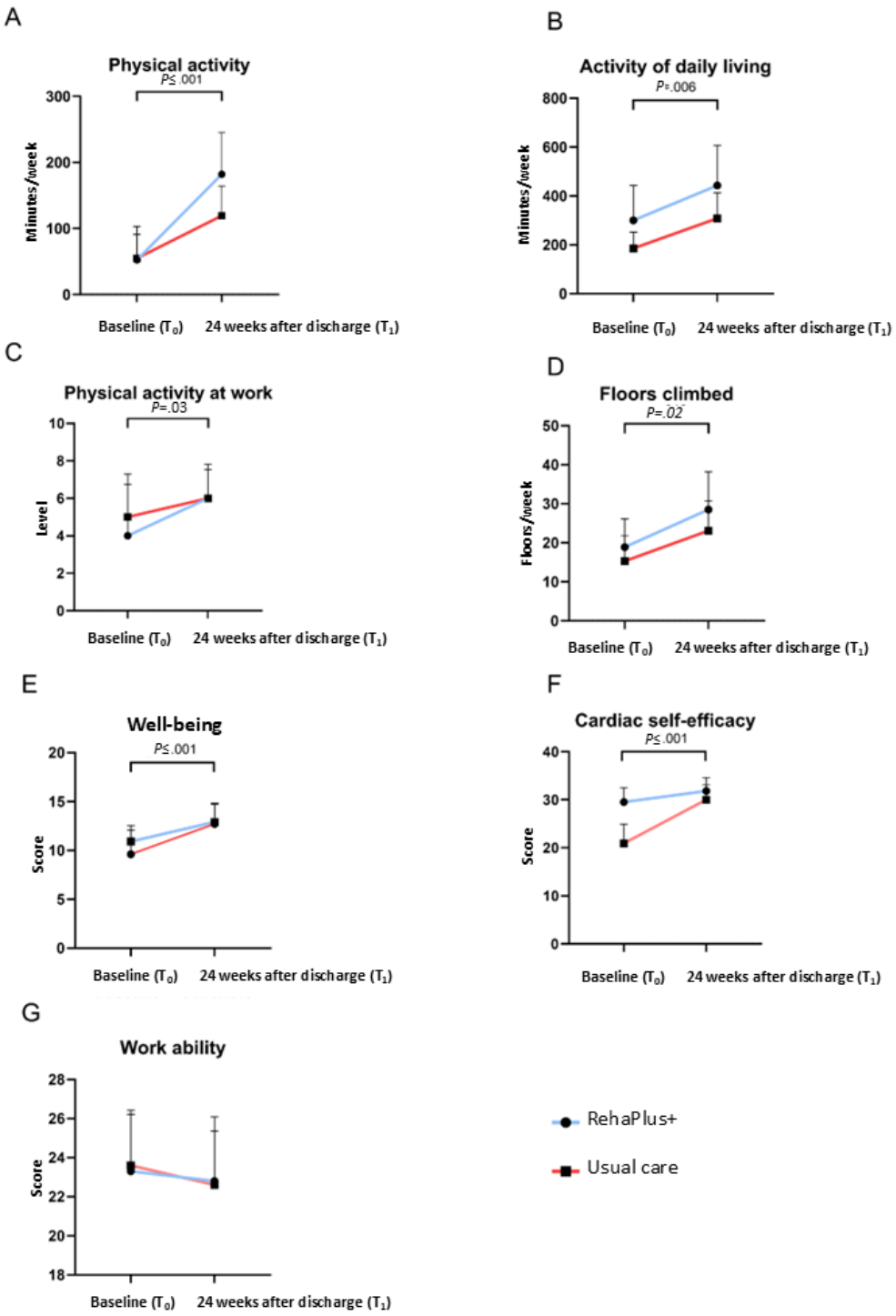

Activities of Daily Living

The RehaPlus+ group showed an ADL level of 443 (SD 538) minutes per week compared to the usual care group with 308 (SD 412) minutes per week at 6-month follow-up without significance (Table 2; *P*=.84). Of note, ADL levels were already higher in the RehaPlus+ group at baseline assessment. Removal of outliers (n=7) had no effect on the observed results (*P*=.37). Over time (T_0 to T_1), both groups combined showed a significant increase in ADLs per week from prerehabilitation to 6-month follow-up compared to T_0 (Figure 5B; mean 131, SD 472

XSL•FO
RenderX

minutes; *P*=.006) without significant gender effects (*P*=.38). A responder analysis for ADLs revealed that 45% (20/44) of the RehaPlus+ participants were responders compared to 33% (20/61) of the usual care participants.

Table 2. Physical activity at 6-month follow-up by group[a].

	Overall (N=105)	*P* value for time	RehaPlus+ (n=44)	Usual care (n=61)	*P* value for group	95% CI interaction
Physical activity		*.001* [b]				–127.2 to 11.2
T0, mean (SD)	53.8 (164.8)		52.4 (127.1)	54.7 (188.3)	.99	
T1, mean (SD)	145.5 (191.0)		181.85 (207.84)	119.20 (174.89)	.15	
Δ, mean (SD)	91.7 (210.8)		129.5 (220.6)	64.5 (200.9)	.12	
95% CI	–132.5 to –50.9		–196.5 to –62.4	–115.9 to –13.0		
Activities of daily living		*.006*				–200.3 to 75.9
T0, mean (SD)	233.7 (365.9)		300.5 (469.1)	185.43 (262.2)	.31	
T1, mean (SD)	364.6 (471.0)		442.79 (537.93)	308.27 (411.52)	.20	
Δ, mean (SD)	131.0 (472.0)		142.3 (574.6)	122.8 (386.5)	.84	
95% CI	–210.2 to –30.6		–317.0 to 32.4	–197.8 to –11.0		
Levels of physical activity at work		*.03*				–1.8 to 0.1
T0, mean (SEM)	5 (9)		4 (9)	5 (9)	.08	
T1, mean (SEM)	6 (6)		6 (6)	6 (6)	.96	
Δ, mean (SEM)	1 (2)		1 (11)	0 (12)	.07	
95% CI	–1.0 to –0.1		–1.8 to –0.2	–0.7 to 0.4		
Floors climbed per week		*.02*				–16.1 to 12.7
T0, mean (SD)	16.8 (24.5)		18.9 (23.7)	15.3 (25.2)	.76	
T1, mean (SD)	25.4 (30.7)		28.5 (32.0)	23.1 (29.8)	.56	
Δ, mean (SD)	8.6 (36.4)		9.6 (36.4)	7.9 (36.8)	.82	
95% CI	–15.6 to –1.5		–20.6 to 1.5	–17.3 to 1.6		

[a]Variables were assessed using the Bewegungs- und Sportaktivität questionnaire after 24 weeks. Differences between groups over time (RehaPlus+ vs usual care) were analyzed using 2-way ANOVA. Levels of physical activity at work: range 0 to 9 (higher=greater activity). T0: end of rehabilitation, T1: after 6 months, Δ, change from T0 to T1.

[b]Italicization denotes statistical significance.

Secondary Outcomes

PA During Work

At 6-month follow-up (T_1), the RehaPlus+ group and usual care group both reported an average level of PA at work of 6 (SEM 6) (; Table 2; *P*=.96), with no significant time×group interaction effect (*P*=.07). Compared with prerehabilitation (T_0), an overall change in the level of PA during work (+ 1; *P*=.02) was observed at 6 months (Figure 5C).

Floors Climbed Weekly

At T_1, the RehaPlus+ group showed an average of 28.5 (SD 32.0) floors climbed per week, while the usual care group had an average of 23.1 (SD 29.8) floors climbed per week (Table 2; *P*=.82). Overall, (both groups combined), there was a significant change in floors climbed weekly at 6-month follow-up compared to T_0, with an average increase of 8.6 (SD 36.4) floors climbed per week (Figure 5D; *P*=.02).

Psychological Well-Being

The preintervention scores for the RehaPlus+ and the usual care group showed an increase in psychological well-being, with the RehaPlus+ group changing from 9.6 (SD 8.1) to 12.7 (SD 7.0) postintervention and the usual care group changing from 10.9 (SD 6.4) to 12.9 (SD 7.2), without a significant time by group interaction (Table 3; *P*=.53). Compared to T_0, an overall change in well-being (mean 2.5, SD 7.6; *P*=.001) was observed at 6 months (Figure 5E).

XSL•FO
RenderX

Table 3. Change in well-being, cardiac self-efficacy, and work ability.

	Overall (N=105)	*P* value time	RehaPlus+ (n=44)	Usual care (n=61)	*P* value group	95% CI interaction
Well-being		*.001* [b]				–3.9 to 2.0
T0, mean (SD)	10.3 (7.1)		9.6 (8.1)	10.9 (6.4)	.61	
T1, mean (SD)	12.9 (7.1)		12.7 (7.0)	12.9 (7.2)	.97	
Δ, mean (SD)	2.5 (7.6)		3.1 (8.7)	2.1 (6.7)	.53	
95% CI	–4.0 to –1.0		–5.7 to –0.4	–3.8 to –0.4		
Cardiac self-efficacy		*.001*				–0.5 to 13.1
T0, mean (SD)	24.5 (14.1)		29.5 (9.8)	20.9 (15.6)	*.001*	
T1, mean (SD)	30.8 (10.9)		31.8 (9.2)	30.0 (12.0)	.71	
Δ, mean (SD)	6.3 (16.3)		2.3 (9.6)	9.1 (19.3)	*.03*	
95% CI	–9.4 to –3.1		–5.2 to 0.6	–14.1 to –4.2		
Work ability		.66				–5.2 to 3.3
T0, mean (SD)	23.3 (10.3)		23.0 (10.6)	23.6 (10.2)	.97	
T1, mean (SD)	22.8 (10.8)		23.0 (11.0)	22.6 (10.8)	.98	
Δ, mean (SD)	–1.0 (9.5)		0.0 (7.7)	–1.0 (10.6)	.66	
95% CI	–1.5 to 2.6		−2.7 to 2.7	–2.1 to 4.0		

[a]The 5-item World Health Organization Well-Being Index (WHO-5), cardiac self-efficacy (CSE), and Work Ability Index (WAI) were assessed by a questionnaire at baseline (start of phase II cardiac rehabilitation, T0) and after 24 weeks (T1). Δ indicates the mean individual difference between T0 and T1. Differences between groups over time (RehaPlus+ vs usual care) were analyzed using 2-way ANOVA. WHO-5: range 0 to 25 (higher=greater well-being), WAI: range 7 to 49 (higher=improved work ability), and CSE: range 0 to 100 (higher=greater well-being).

[b]Italicization denotes statistical significance.

CSE Overview

The baseline values of CSE were significantly different, with the RehaPlus+ group starting with a significantly higher value (*P*=.001). The preintervention scores for the RehaPlus+ and the usual care group showed an increase in CSE, with the RehaPlus+ group changing from 29.5 (SD 9.8) to 31.8 (SD 9.2) postintervention and the usual care group changing from 20.9 (SD 15.6) to 30.2 (SD 12.0), with a significant time by group interaction (Table 3; *P*=.03). At the 6-month follow-up, there was a significant improvement in CSE across all patients compared to baseline (T0), with an average increase of 6.3 points (SD 16.3; Figure 5F; *P*=.001).

Work Ability

At T_0, the RehaPlus+ group had an average Work Ability Index score of 23.0 (SD 10.6) at baseline and an average score of 23.0 (SD 11.0) at T_1, while the usual care group decreased from 23.6 (SD 10.2) to 22.6 (SD 10.8); however, there was no significant difference between the 2 groups over time (Table 3; *P*=.66).

Health-Related QoL

The SF-36 scores were comparable between both groups at $T_{0;}$ however, the usual care group started with a higher MCS score (*P*=.03). Over time (T_0-T_1), health-related QoL changed significantly during phase III CR over both groups combined (RehaPlus+: MCS+14% and PCS+18%; usual care: MCS+12% and PCS+17%; *P*=.05), with no difference between the groups (*P*=.36; Figure 6).

XSL•FO
RenderX

Figure 6. Health-related quality of life assessed by 36-item Short Form Survey (SF-36) questionnaire. At baseline, SF-36 scores were comparable between both groups, although the usual care group had a higher mental component summary (MCS) score (*P*=.03). Both groups increased their preintervention score, with no difference observed between the 2 groups (P≥.362). Data are presented as 95% CI; changes over time between groups (time×group interaction) were calculated using general linear model. SF-36 score range: 0-100 (higher=greater well-being). BP: bodily pain; GH: general health; MH: mental health; PCS: physical component summary; PF: physical functioning; RE: role emotional; RP: role physical; SF: social functioning; Vit: vitality.

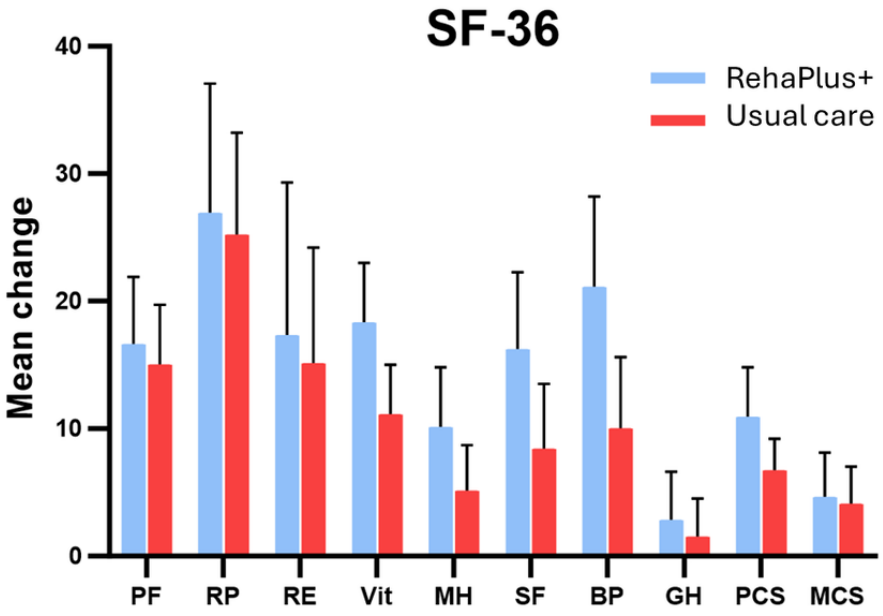

Discussion

Principal Findings

The aim of this study was to evaluate the efficacy of the multimodal, behavioral change–based maintenance program RehaPlus+ and to determine its noninferiority to the German usual care. In brief, the main findings of the study are that (1) the RehaPlus+ program, which uses BCTs and theories delivered partly via eHealth, was found to be noninferior to the usual care program; (2) both interventions induced a comparable and significant increase in PA during work, floors climbed weekly, psychological well-being, CSE, work ability, and health-related QoL; and (3) the implementation of the RehaPlus+ messaging system was perceived as user-friendly by the involved case managers. The analysis revealed that there were no significant differences between the RehaPlus+ and usual care group across outcome measures.

RehaPlus+ uses a range of established BCTs such as goal setting, action planning, and motivational and coping strategies to empower patients with cardiac conditions to make positive lifestyle changes. BCTs are designed to target and modify a specific behavior that plays a role in overall health and well-being. Goal setting as a central component likely empowered participants to establish clear and attainable objectives related to their desired outcomes, as research highlights the effectiveness of these techniques in promoting PA in patients with cardiac conditions [43]. Action planning is crucial in the initial phases of rehabilitation, while coping strategies are hypothesized to facilitate the long-term maintenance of behavior change [44]. Coping strategies may have helped RehaPlus+ participants to identify and address barriers, enabling them to develop strategies for overcoming challenges and sustain their progress. This is supported by the observation that individuals with higher levels of coping planning postdischarge are more likely to engage in increased levels of exercise [45]. Coping planning is a self-regulation strategy designed to anticipate and counteract personal risk situations that threaten intended behavior. It creates a mental connection between potential risks and appropriate coping responses [45]. It has been demonstrated that people who effectively plan for high-risk situations, such as exercise relapse, are less likely to give in to these challenges [46]. This approach is aligned with techniques used in cognitive behavioral therapy, where anticipating and preparing for potential challenges helps manage unwanted behaviors. While action planning facilitates task execution, coping planning focuses on avoiding distractions. The effectiveness of coping planning increases over time with experience, as it relies on an individuals' understanding of their personal risk situations and responses. This strategy is crucial for maintaining focus on long-term goals and preventing unwanted behaviors [45]. Coping planning interventions are found to be effective, especially when participants receive support in developing their coping plans [47]. This approach is similar to the one used in the RehaPlus+ program, where guided assistance in coping plan formulation is a key component.

In RehaPlus+, the implementation of BCTs may have contributed to the development of healthy exercise habits and enhanced self-efficacy for lifestyle changes. However, it is crucial to assess the respective contributions of the

XSL•FO
RenderX

technology-driven message delivery system and the support provided by the case managers to the overall success of the intervention. The case manager met the patient in a face-to-face session and helped to create an action plan and set goals, considering potential barriers. The case manager then developed a tailored message chain for the patient, considering the discussed goals and planned actions. In addition, the case manager contacted the individuals with cardiac conditions during 2 phone calls to adjust the timing or content of the messages based on any changes in their activities or work schedules. This aspect represents a common challenge in studies of this nature, and the individual impact of each component on the outcomes cannot be differentiated.

According to the Health Action Process Approach [31], the maintenance of health behavior requires specific input, including: (1) action planning, to specify situation parameters ("when" and "where") and a sequence of action ("how") to implement intended behavior; (2) action control, to help sustaining the behavioral change; and (3) coping planning as a self-regulatory strategy or alternative behavior to overcome barriers. Therefore, RehaPlus+ case managers suggested practical solutions to everyday barriers and motivated participants to incorporate "what-if" sentences to their individual action plans. These sentences focus on potential obstacles and simultaneously identified solutions, such as "If it's raining during my scheduled outdoor workout, I will move indoors and complete my exercise routine." RehaPlus+ aimed to motivate patients with CAD toward increased, self-organized PA as well as increased ADLs through targeted lifestyle adjustments and health-promoting behavior in everyday life. In contrast, the usual care program focuses on promoting PA through regular appointments for guided group-based exercise sessions. However, despite its targeted focus on PA, usual care did not demonstrate superiority over RehaPlus+. Both programs were effective in promoting PA, suggesting that different methodologies in rehabilitation can be equally successful in achieving similar outcomes in PA enhancement.

With regard to ADLs, a recent meta-analysis reported significant effects of different eHealth interventions in increasing walking steps in older adults [48], suggesting that eHealth has the potential to effectively enhance everyday activities with minimal barriers and versatility in practice. Our data support these findings in that the RehaPlus+ group showed an increase in ADLs, likely caused by the fact that RehaPlus+ was designed to stimulate a broader range of daily activities. This observation is consistent with previous research [49] emphasizing the significant impact of BCTs on habit formation, which play a crucial role in maintaining regular exercise routines for cardiovascular health promotion. The effects of RehaPlus+ may also relate to the personalized approach, which considered individual barriers. Tailoring the intervention to individual characteristics and preferences is a crucial aspect of ensuring the success of the rehabilitation program. By focusing also on implementation, RehaPlus+ sought to maximize the effectiveness of the BCTs, aligning them with the unique needs of each patient. This approach is in line with the principle that the success of behavior change interventions is significantly influenced by how well they are adapted to individual patient profiles [50], ensuring that the strategies used are not only theoretically sound but also practically relevant and effective in the specific context of each patients' rehabilitation journey. However, it is possible that the combination of human intervention and personal contact during the maintenance period together with the use of technology may have contributed to the observed treatment effect synergistically. RehaPlus+ integrates a personalized eHealth intervention with behavioral change concepts, emphasizing healthy lifestyle changes also supported by case managers. It is likely that the personal contact enhances the value and trustworthiness of the messages, further enhancing the effectiveness of the intervention. In addition, technology enhances accessibility, convenience, and scalability, allowing for continuous engagement at home beyond traditional center visits. Further studies are warranted to delineate the relative contributions of human interaction and technology in driving treatment outcomes. For instance, future investigations could explore the impact of reducing the duration of personal contact or modifying the intensity of the message component.

The effectiveness of BCTs is assumed to depend on their appropriate implementation, which should take into account individual characteristics and preferences. The improvements in health-related QoL, CSE, and overall well-being may have likely played a role in adherence to the respective programs.

With respect to the participants' baseline characteristics, it seems important to note that 94% (99/105) of the participants did not have qualifications extending beyond high school level. Of note, the RehaPlus+ group had a higher percentage of participants (15/44, 33%) with an educational level exceeding high school. While it is unlikely that these differences may have influenced the outcomes, the finding underscores that eHealth interventions should be adaptable to the individual participant and potential hurdles for patients with lower literacy should be reduced. In general, it is crucial that eHealth solutions, such as mobile apps, should be easily understandable and accessible to all patients, regardless of their educational background. In the context of gender-sensitive care, caregiving can exert distinct mental and physical health effects depending on gender roles and societal expectations. A gender-sensitive approach allows for tailored interventions and support systems, acknowledging and mitigating these unique health impacts experienced by caregivers of different genders.

The use of eHealth programs such as RehaPlus+ provides several advantages, including cost-effectiveness [51]. In this case, the outpatient usual care program IRENA incurs costs of approximately €770 (US $850) per patient for 24 weeks. In contrast, the eHealth program RehaPlus+ costs approximately €540 (US $600) per patient for the same duration [52]. However, RehaPlus+ is scalable, and the costs in other countries will depend on local conditions and circumstances. In addition, similar outcomes can be achieved with lower costs, reduced resource use, and less staff burden.

Recent systematic reviews have highlighted the cost-effectiveness of cardiac telerehabilitation in a broader context, showing that eHealth interventions hold the capacity to reach a broader and more diverse audience at a lower cost [53,54]. Such programs can overcome geographical barriers,

XSL•FO
RenderX

enhancing accessibility for individuals situated in remote or underserved areas. Furthermore, eHealth programs can be structured for self-administration, reducing the demand for extensive human resources and subsequently lowering operational expenses. However, for a comprehensive assessment, a detailed cost-effectiveness analysis of RehaPlus+ would be required to compare the expenses associated with implementing eHealth programs against the benefits in terms of health outcomes and reduced health care use.

Limitations

The self-reported nature of PA is one of the major limitations of this study. This method relies on subjective assessments and memory recall, which could lead to biases in reporting participants' actual activity levels, even if likely comparable between both groups. The study experienced considerable loss to follow-up rates, primarily due to difficulties in obtaining questionnaire-based outcome data 6 months after the intervention. Although the observed rates were comparable between both groups and equal to previous eHealth and non–technology-based interventions for behavior maintenance [55-57], it is crucial to thoroughly examine that motivation is a relevant aspect in eHealth approaches, and there is a risk that eHealth interventions may ultimately leave only highly compliant patients. Patients who drop out of eHealth interventions may require different incentives and motivations to stay engaged and achieve their health goals in the long term. To effectively support individuals in earlier stages of change, it may be essential to first implement other interventions aimed at promoting intrinsic motivation for change. These interventions may include motivational interviewing, psychoeducation about the risks and benefits of behavior change, and building self-efficacy through goal setting and reinforcement. By addressing these factors, individuals in stage 1 or 2 (according to the TTM) can be supported in developing the internal drive necessary for sustained behavior change [28]. In addition, it is important to acknowledge that baseline data from 11 participants were unavailable for analysis. These participants were prematurely discharged from their rehabilitation due to COVID-19 infections, hindering the completion of baseline assessments. Furthermore, outcomes were measured by questionnaires, and no objective measurements were used.

Conclusions

This study demonstrates that the behavioral change–based eHealth maintenance program RehaPlus+ was equally effective as the standard German usual care in promoting regular ADLs and PA in patients with CAD. This achievement can be attributed, in part, to the successful use of BCTs within the RehaPlus+ program. This study underscores the importance of evidence-based BCTs in eHealth interventions and highlights the need for personalized eHealth intervention strategies. Future research should refine BCT implementation in eHealth programs to improve health behaviors across diverse populations.

Acknowledgments

The authors greatly acknowledge the cooperation of all participants, without whom this study would not have been possible. The authors acknowledge the support of clinicians and assistants at Clinic Königsfeld. The assistance of Ms Julia Müller, Mr Dennis Köstler, and Ms Julia Büchner is also gratefully acknowledged. The translation of 36-Item Short Form Survey is based on the 36-Item Short Form Survey developed and owned by the RAND Corporation. MW was supported by the internal grant program (project IFF 2022-13) of the Faculty of Health at Witten/Herdecke University, Germany, for conducting the study. FCM and BS are supported by the European Commission within the Horizon 2020 framework program (grant 101017424).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MW, FCM, and BS designed the study. MW coordinated the study. MW and MT enrolled participants. MW supervised group seminars and individual counseling. MW, BS, RG, and MK analyzed the data and interpreted the results. MW and BS drafted the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) version 1.6.
[PDF File (Adobe PDF File), 1112 KB - jmir_v26i1e56480_app1.pdf]

References

1. Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, et al. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J 2020 Jan 14;41(3):407-477 [FREE Full text] [doi: 10.1093/eurheartj/ehz425] [Medline: 31504439]

XSL•FO
RenderX

2. Goel K, Lennon RJ, Tilbury RT, Squires RW, Thomas RJ. Impact of cardiac rehabilitation on mortality and cardiovascular events after percutaneous coronary intervention in the community. Circulation 2011 May 31;123(21):2344-2352. [doi: 10.1161/CIRCULATIONAHA.110.983536] [Medline: 21576654]
3. Balady GJ, Williams MA, Ades PA, Bittner V, Comoss P, Foody JM, et al. Core components of cardiac rehabilitation/secondary prevention programs: 2007 update: a scientific statement from the American heart association exercise, cardiac rehabilitation, and prevention committee, the council on clinical cardiology; the councils on cardiovascular nursing, epidemiology and prevention, and nutrition, physical activity, and metabolism; and the American association of cardiovascular and pulmonary rehabilitation. Circulation 2007 May 22;115(20):2675-2682. [doi: 10.1161/CIRCULATIONAHA.106.180945] [Medline: 17513578]
4. Heran BS, Chen JM, Ebrahim S, Moxham T, Oldridge N, Rees K, et al. Exercise-based cardiac rehabilitation for coronary heart disease. Cochrane Database Syst Rev 2011 Jul 06(7):CD001800 [FREE Full text] [doi: 10.1002/14651858.CD001800.pub2] [Medline: 21735386]
5. Giuliano C, Parmenter BJ, Baker MK, Mitchell BL, Williams AD, Lyndon K, et al. Cardiac rehabilitation for patients with coronary artery disease: a practical guide to enhance patient outcomes through continuity of care. Clin Med Insights Cardiol 2017 Jun 12;11:1179546817710028 [FREE Full text] [doi: 10.1177/1179546817710028] [Medline: 28638244]
6. Simon M, Korn K, Cho L, Blackburn GG, Raymond C. Cardiac rehabilitation: a class 1 recommendation. Cleve Clin J Med 2018 Jul;85(7):551-558 [FREE Full text] [doi: 10.3949/ccjm.85a.17037] [Medline: 30004380]
7. Bjarnason-Wehrens B, McGee H, Zwisler AD, Piepoli MF, Benzer W, Schmid JP, et al. Cardiac rehabilitation in Europe: results from the European cardiac rehabilitation inventory survey. Eur J Cardiovasc Prev Rehabil 2010 Aug;17(4):410-418. [doi: 10.1097/HJR.0b013e328334f42d] [Medline: 20300001]
8. Walsh DM, Moran K, Cornelissen V, Buys R, Claes J, Zampognaro P, et al. The development and codesign of the PATHway intervention: a theory-driven eHealth platform for the self-management of cardiovascular disease. Transl Behav Med 2019 Jan 01;9(1):76-98. [doi: 10.1093/tbm/iby017] [Medline: 29554380]
9. Dalal HM, Doherty P, Taylor RS. Cardiac rehabilitation. BMJ 2015 Sep 29;351:h5000 [FREE Full text] [doi: 10.1136/bmj.h5000] [Medline: 26419744]
10. Guidelines for Cardiac Rehabilitation and Secondary Prevention Programs. Champaign, IL: American Association of Cardiovascular and Pulmonary Rehabilitation; 2013.
11. Janssen V, De Gucht V, van Exel H, Maes S. A self-regulation lifestyle program for post-cardiac rehabilitation patients has long-term effects on exercise adherence. J Behav Med 2014 Apr;37(2):308-321. [doi: 10.1007/s10865-012-9489-y] [Medline: 23334387]
12. Hansen D, Dendale P, Raskin A, Schoonis A, Berger J, Vlassak I, et al. Long-term effect of rehabilitation in coronary artery disease patients: randomized clinical trial of the impact of exercise volume. Clin Rehabil 2010 Apr;24(4):319-327. [doi: 10.1177/0269215509353262] [Medline: 20176771]
13. Bjarnason-Wehrens B, Grande G, Loewel H, Völler H, Mittag O. Gender-specific issues in cardiac rehabilitation: do women with ischaemic heart disease need specially tailored programmes? Eur J Cardiovasc Prev Rehabil 2007 Apr;14(2):163-171. [doi: 10.1097/HJR.0b013e3280128bce] [Medline: 17446793]
14. Russell MW, Huse DM, Drowns S, Hamel EC, Hartz SC. Direct medical costs of coronary artery disease in the United States. Am J Cardiol 1998 May 01;81(9):1110-1115. [doi: 10.1016/s0002-9149(98)00136-2] [Medline: 9605051]
15. Tarride JE, Lim M, DesMeules M, Luo W, Burke N, O'Reilly D, et al. A review of the cost of cardiovascular disease. Can J Cardiol 2009 Jun;25(6):e195-e202 [FREE Full text] [doi: 10.1016/s0828-282x(09)70098-4] [Medline: 19536390]
16. Drwal KR, Forman DE, Wakefield BJ, El Accaoui RN. Cardiac rehabilitation during COVID-19 pandemic: highlighting the value of home-based programs. Telemed J E Health 2020 Nov;26(11):1322-1324. [doi: 10.1089/tmj.2020.0213] [Medline: 32552412]
17. Silva-Cardoso J, Juanatey JR, Comin-Colet J, Sousa JM, Cavalheiro A, Moreira E. The future of telemedicine in the management of heart failure patients. Card Fail Rev 2021 May 28;7:e11 [FREE Full text] [doi: 10.15420/cfr.2020.32] [Medline: 34136277]
18. Schmitz B, Gatsios D, Peña-Gil C, Juanatey JR, Prieto DC, Tsakanikas V, et al. Patient-centered cardiac rehabilitation by AI-powered lifestyle intervention – the timely approach. Atherosclerosis 2022 Aug;355:251. [doi: 10.1016/j.atherosclerosis.2022.06.959]
19. Chen J, Sun D, Zhang S, Shi Y, Qiao F, Zhou Y, et al. Effects of home-based telerehabilitation in patients with stroke: a randomized controlled trial. Neurology 2020 Oct 27;95(17):e2318-e2330. [doi: 10.1212/WNL.0000000000010821] [Medline: 32999058]
20. Pakrad F, Ahmadi F, Grace SL, Oshvandi K, Kazemnejad A. Traditional vs extended hybrid cardiac rehabilitation based on the continuous care model for patients who have undergone coronary artery bypass surgery in a middle-income country: a randomized controlled trial. Arch Phys Med Rehabil 2021 Nov;102(11):2091-101.e3. [doi: 10.1016/j.apmr.2021.04.026] [Medline: 34175270]
21. Rawstorn JC, Gant N, Direito A, Beckmann C, Maddison R. Telehealth exercise-based cardiac rehabilitation: a systematic review and meta-analysis. Heart 2016 Aug 01;102(15):1183-1192. [doi: 10.1136/heartjnl-2015-308966] [Medline: 26936337]

XSL•FO
RenderX

22. Heimer M, Schmitz S, Teschler M, Schäfer H, Douma ER, Habibovic M, et al. eHealth for maintenance cardiovascular rehabilitation: a systematic review and meta-analysis. Eur J Prev Cardiol 2023 Oct 26;30(15):1634-1651. [doi: 10.1093/eurjpc/zwad145] [Medline: 37154363]
23. Rogerson MC, Jackson AC, Navaratnam HS, Le Grande MR, Higgins RO, Clarke J, et al. Behavioural and psychological telehealth support for people with cardiac conditions: randomized trial of the 'back on track' self-management programme. Eur J Cardiovasc Nurs 2024 Jan 12;23(1):42-54. [doi: 10.1093/eurjcn/zvad034] [Medline: 36989400]
24. Löllgen H, Löllgen D. [Risk reduction in cardiovascular diseases by physical activity]. Internist (Berl) 2012 Jan;53(1):20-29. [doi: 10.1007/s00108-011-2889-1] [Medline: 22086105]
25. Axelrod DA, Hayward R. Nonrandomized interventional study designs (quasi-experimental designs). In: Penson DF, Wei JT, editors. Clinical Research Methods for Surgeons. Humana Press. Totowa, NJ: Humana Press; 2006.
26. Prochaska JO, Diclemente CC. Toward a comprehensive model of change. In: Miller WR, Heather N, editors. Treating Addictive Behaviors. Applied Clinical Psychology. Boston, MA: Springer; 1986.
27. Prochaska JO, DiClemente CC. The transtheoretical approach. In: Handbook of Psychotherapy Integration. Oxford, UK: Oxford Academic Press; Apr 2019.
28. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
29. Marshall SJ, Biddle SJ. The transtheoretical model of behavior change: a meta-analysis of applications to physical activity and exercise. Ann Behav Med 2001;23(4):229-246. [doi: 10.1207/S15324796ABM2304_2] [Medline: 11761340]
30. Schwarzer R. Modeling health behavior change: how to predict and modify the adoption and maintenance of health behaviors. Appl Psychol 2008 Jan 30;57(1):1-29. [doi: 10.1111/j.1464-0597.2007.00325.x]
31. Schwarzer R. Self-efficacy in the adoption and maintenance of health behaviors: theoretical approaches and a new model. In: Self-Efficacy: Thought Control Of Action. London, UK: Hemisphere Publishing Corp; 1992.
32. Bandura A. Self-Efficacy: The Exercise of Control. New York City, NY: W H Freeman; 1997.
33. Deutsche R. Rahmenkonzept zur Nachsorge für medizinische Rehabilitation nach § 15 SGB VI. - 2024:2015.
34. Fachkonzept „IRENA®“. Deutsche Rentenversicherung. 2021 Oct. URL: https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Experten/infos_reha_einrichtungen/nachsorge/fachkonzept_irena.pdf?__blob=publicationFile&v=1 [accessed 2024-09-05]
35. Fuchs R, Klaperski S, Gerber M, Seelig H. Messung der bewegungs- und sportaktivität mit dem BSA-fragebogen. Zeitschrift für Gesundheitspsychologie 2015 Jun 30;23(2):60-76. [doi: 10.1026/0943-8149/a000137]
36. Ware JEJ, Keller SD, Kosinski M. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab, New England Medical Center; Dec 1994.
37. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 well-being index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
38. Martus P, Jakob O, Rose U, Seibt R, Freude G. A comparative analysis of the work ability index. Occup Med (Lond) 2010 Oct;60(7):517-524. [doi: 10.1093/occmed/kqq093] [Medline: 20573840]
39. Ilmarinen J, Tempel J. Erhaltung, förderung und entwicklung der arbeitsfähigkeit — konzepte und forschungsergebnisse aus Finnland. In: Badura B, Schellschmidt H, Vetter C, editors. Demographischer Wandel: Herausforderung für die Betriebliche Personal- und Gesundheitspolitik. Berlin, Germany: Springer; 2003.
40. Sullivan MD, LaCroix AZ, Russo J, Katon WJ. Self-efficacy and self-reported functional status in coronary heart disease: a six-month prospective study. Psychosom Med 1998;60(4):473-478. [doi: 10.1097/00006842-199807000-00014] [Medline: 9710293]
41. Kishore K, Mahajan R. Understanding superiority, noninferiority, and equivalence for clinical trials. Indian Dermatol Online J 2020 Sep 19;11(6):890-894 [FREE Full text] [doi: 10.4103/idoj.IDOJ_130_20] [Medline: 33344335]
42. Schmitz B, Niehues H, Thorwesten L, Klose A, Krüger M, Brand SM. Sex differences in high-intensity interval training-are HIIT protocols interchangeable between females and males? Front Physiol 2020 Jan 29;11:38 [FREE Full text] [doi: 10.3389/fphys.2020.00038] [Medline: 32063866]
43. Luszczynska A, Scholz U, Schwarzer R. The general self-efficacy scale: multicultural validation studies. J Psychol 2005 Sep;139(5):439-457. [doi: 10.3200/JRLP.139.5.439-457] [Medline: 16285214]
44. Baker A, Cornwell P, Gustafsson L, Stewart C, Lannin NA. Developing tailored theoretically informed goal-setting interventions for rehabilitation services: a co-design approach. BMC Health Serv Res 2022 Jun 22;22(1):811 [FREE Full text] [doi: 10.1186/s12913-022-08047-6] [Medline: 35733190]
45. Sniehotta FF, Schwarzer R, Scholz U, Schüz B. Action planning and coping planning for long-term lifestyle change: theory and assessment. Eur J Soc Psychol 2005 Jul 15;35(4):565-576. [doi: 10.1002/ejsp.258]
46. Simkin LR, Gross AM. Assessment of coping with high-risk situations for exercise relapse among healthy women. Health Psychol 1994 May;13(3):274-277. [doi: 10.1037//0278-6133.13.3.274] [Medline: 8055862]
47. Kwasnicka D, Presseau J, White M, Sniehotta FF. Does planning how to cope with anticipated barriers facilitate health-related behaviour change? A systematic review. Health Psychol Rev 2013 Feb 19;7(2):129-145. [doi: 10.1080/17437199.2013.766832]

XSL•FO
RenderX

48. Kwan RY, Salihu D, Lee PH, Tse M, Cheung DS, Roopsawang I, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020 Apr 21;17:7 [FREE Full text] [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
49. Moore SM, Charvat JM, Gordon NH, Pashkow F, Ribisl P, Roberts BL, et al. Effects of a CHANGE intervention to increase exercise maintenance following cardiac events. Ann Behav Med 2006 Feb;31(1):53-62. [doi: 10.1207/s15324796abm3101_9] [Medline: 16472039]
50. Hardcastle SJ, Hagger MS. Psychographic profiling for effective health behavior change interventions. Front Psychol 2016 Jan 06;6:1988 [FREE Full text] [doi: 10.3389/fpsyg.2015.01988] [Medline: 26779094]
51. Frederix I, Solmi F, Piepoli MF, Dendale P. Cardiac telerehabilitation: a novel cost-efficient care delivery strategy that can induce long-term health benefits. Eur J Prev Cardiol 2017 Nov;24(16):1708-1717. [doi: 10.1177/2047487317732274] [Medline: 28925749]
52. Kernangebote der reha-nachsorge. Deutsche Rentenversicherung. URL: https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Experten/infos_reha_einrichtungen/konzepte_systemfragen/konzepte/beschreibung_kernangebote_reha_nachsorge.html [accessed 2024-09-05]
53. Shields GE, Wells A, Doherty P, Heagerty A, Buck D, Davies LM. Cost-effectiveness of cardiac rehabilitation: a systematic review. Heart 2018 Sep;104(17):1403-1410 [FREE Full text] [doi: 10.1136/heartjnl-2017-312809] [Medline: 29654096]
54. Scherrenberg M, Falter M, Dendale P. Cost-effectiveness of cardiac telerehabilitation in coronary artery disease and heart failure patients: systematic review of randomized controlled trials. Eur Heart J Digit Health 2020 Nov 23;1(1):20-29 [FREE Full text] [doi: 10.1093/ehjdh/ztaa005] [Medline: 37056294]
55. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
56. Wangberg SC, Bergmo TS, Johnsen JA. Adherence in internet-based interventions. Patient Prefer Adherence 2008 Feb 02;2:57-65 [FREE Full text] [Medline: 19920945]
57. Fjeldsoe B, Neuhaus M, Winkler E, Eakin E. Systematic review of maintenance of behavior change following physical activity and dietary interventions. Health Psychol 2011 Jan;30(1):99-109. [doi: 10.1037/a0021974] [Medline: 21299298]

Abbreviations

ADL: activity of daily living
BCT: behavior change technique
CAD: coronary artery disease
CONSORT: Consolidated Standards of Reporting Trials
CR: cardiac rehabilitation
CSE: cardiac self-efficacy
IRENA: Individualized Rehabilitation Aftercare in Post-Acute Treatment
MCS: mental component summary
PA: physical activity
PCS: physical component summary
QoL: quality of life
TE: typical error
TTM: transtheoretical model

Edited by T de Azevedo Cardoso; submitted 17.01.24; peer-reviewed by S Meister, M Hollings; comments to author 30.04.24; revised version received 20.06.24; accepted 22.07.24; published 08.10.24.

Please cite as:
Waranski M, Garbsch R, Kotewitsch M, Teschler M, Schmitz B, Mooren FC
A Behavioral Change–Based Mobile Intervention for Promoting Regular Physical Activity in Medical Rehabilitation Maintenance of Patients With Coronary Artery Disease: Controlled Trial
J Med Internet Res 2024;26:e56480
URL: https://www.jmir.org/2024/1/e56480
doi:10.2196/56480
PMID:

©Melina Waranski, René Garbsch, Mona Kotewitsch, Marc Teschler, Boris Schmitz, Frank C Mooren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Menstrual Cycle Management and Period Tracker App Use in Millennial and Generation Z Individuals: Mixed Methods Study

Minji Hong[1], MA; Vasuki Rajaguru[2], PhD; KyungYi Kim[2], PhD; Suk-Yong Jang[2], MD, JD, PhD; Sang Gyu Lee[3], MD, PhD, MBA

[1]Department of Medical Device Engineering and Management, College of Medicine, Yonsei University, Seoul, Republic of Korea

[2]Department of Healthcare Management, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea

[3]Department of Preventive Medicine, College of Medicine, Yonsei University, Seoul, Republic of Korea

Corresponding Author:
Sang Gyu Lee, MD, PhD, MBA
Department of Preventive Medicine
College of Medicine
Yonsei University
50-1, Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 02 2228 1500
Email: Leevan@yuhs.ac

Abstract

Background: Menstruation is a physical symptom that occurs in women of reproductive age. It has a significant impact on the daily life and health of women when their academic and social activities are most active. Since many women experience difficulties in daily life because of premenstrual syndrome and dysmenorrhea, it is important to identify, prepare for, and manage the menstrual cycle in advance.

Objective: This study aimed to investigate the relationship between menstruation-related experiences and the use of mobile-based period tracker apps by millennial and generation Z (gen Z) individuals. The objectives of this study are to investigate (1) menstrual cycle management, (2) factors affecting app usage (3) factors affecting cycle management, and (4) motivators and barriers to using period tracker apps, in millennial and gen Z women.

Methods: A mixed methods design was used for this study. The participants were young women aged 20-39 years and recruited via the Ovey application. Data were collected through surveys and focus group interviews. The survey was conducted among 700 women, and 8 of them participated in the focus group interviews.

Results: In total, 431 (62.3%) participants used period tracker apps primarily to predict their next menstrual cycle. Factors affecting app usage included childbirth experience (odds ratio [OR] 0.475, $P<.05$), number of dysmenorrhea symptoms (OR 1.136, $P<.05$), and cycle management level (OR 2.279, $P<.001$). Additionally, education level (OR 1.122, $P<.05$ [university level compared high school level]) and the number of dysmenorrhea symptoms (OR 1.024, $P<.05$) showed a positive association with the cycle management level. However, childbirth experience (OR 0.902, $P<.05$) and period irregularity (OR 0.929, $P<.001$) were negatively associated with the cycle management level.

Conclusions: Period tracker apps are becoming the new normal among millennials and gen Z individuals for managing their menstrual cycles. The use of a period tracker app empowers women by helping them gain a better understanding of their bodies, ultimately enhancing their social, academic, and health-related lives. Improving the accuracy and literacy of the app is an ongoing task for period-tracking apps, and it is important to provide added value tailored to users' needs. Therefore, the findings of this study should be considered when designing or upgrading period tracker apps to facilitate the adoption of digital technology among young women, thereby promoting their overall well-being and reproductive health.

(J Med Internet Res 2024;26:e53146) doi:10.2196/53146

XSL•FO
RenderX

KEYWORDS

menstruation; dysmenorrhea; period tracker app, menstrual cycle management, health care application, millennial; Gen Z; mobile phone

Introduction

Menstruation is a physical symptom that occurs in women of reproductive age and can bring pain or cramps during the menstrual period (hereafter dysmenorrhea) and premenstrual syndrome (PMS), which includes symptoms experienced before the menstrual period. Menstruation significantly impacts women's daily lives when their academic and social activities are most active [1-5], with 85.4% experiencing dysmenorrhea, and 38.4% reporting reduced activity due to menstrual pain [6].

Period tracker apps (PTA), introduced in 2013, offer a new way to manage menstrual cycles, helping over 50 million women globally to track their menstrual cycles, mood changes, and symptoms [7]. PTA users may better understand these patterns, control their moods, and improve their general well-being with the help of PTA.

The main purpose of PTA is to provide information on the number of days left until the next cycle and ovulation. Most apps provide additional features such as cycle-related mood changes, pain and sleep pattern records, and statistical information about average cycles [8]. These apps provide valuable data on cycle patterns, which can help women manage their menstruation more effectively. In the United Kingdom, 38% of women use PTAs mainly for cycle monitoring (72%), pregnancy planning (34%), and obtaining information on infertility treatment (12%) [9].

While PTAs are not specifically designed to treat PMS or dysmenorrhea, studies suggest that they may help alleviate these symptoms [10]. Studies have been focused on the experience of using PTAs or PTA data to understand the experience and app use [7,8,11], feasibility of apps [10-13], effectiveness of period prediction [9], and evaluate the apps from various aspects, such as the information and functions they provide [14-23]. In addition, the association between dysmenorrhea and PMS compared the characteristics of PTA users and nonusers was conducted [9,24].

While numerous studies have focused on PTAs, there remains a significant gap in understanding how these apps are integrated into women's daily lives and how user characteristics influence app usage.

This study aims to explore the relationship between menstruation-related experiences and PTA use among millennials and generation Z (gen Z), who are adept with smartphones and digital devices. This study proposes future implications for the development of women's health mobile apps by providing basic research data on menstrual experiences and management in millennial and gen Z women. The research addresses the following research questions: (1) What is the menstrual cycle management and app usage status of millennials and gen Z? (2) Which social factors and menstrual-related experiences affect PTA use in millennials and gen Z? (3) Which social factors and menstrual-related experiences influence their cycle management level in millennials and gen Z? (4) What are the motivators and barriers to using PTAs in millennials and gen Z?

Methods

Study Design

A mixed methods approach was used in this study. A quantitative survey was first conducted, and qualitative research data were collected through interviews with some of the survey respondents. This explanatory sequential design has the advantage of detailing the quantitative results using a qualitative method [25,26]. The detailed research design is shown in Figure 1. It includes a first phase to collect and analyze quantitative data, a connecting procedure between the first and second phases, and a second phase to collect and analyze qualitative data. Quantitative and qualitative data were linked using panel codes. The study design was planned by referring to the mixed methods research guideline published by the National Institutes of Health [27].

Figure 1. Study design.

XSL•FO
RenderX

Ethical Considerations

The study protocol was reviewed and approved by the institutional review board of Severance Hospital (4-2023-0097). This approval ensures that the study complies with ethical standards and regulations for research involving human subjects.

Participants responded to the survey only after reading the study's purpose and agreement details provided through the mobile app. Given the nature of the survey, which involved minimal risk and was conducted via a mobile application, the requirement for written consent was waived by the institutional review board. The comprehensive survey information sheet for participants is presented in Table S1 in Multimedia Appendix 1.

To protect the privacy and confidentiality of participants, all collected data were anonymized. Identifiable information was not collected, ensuring that individual responses could not be traced back to any participant. Data were securely stored and only accessible to the research team. Additionally, any published results do not include information that could potentially identify participants.

Survey participants received 500 credits, equivalent to 500 Korean won, as compensation for their participation. Additionally, participants in the focus group interviews (FGIs) received 60,000 credits, equivalent to 60,000 Korean won. These incentives were provided to recognize the time and effort contributed by the participants while ensuring transparency and fairness in the compensation process.

Participants

The target population of this study was women aged 20-39 years. Participants were recruited from “Open survey” (Ovey), which is a platform that can provide surveys and pay compensation and has 830,000 voluntarily registered panels. Among those who met the criteria, participants were randomly recruited through an app push, and only those who read and agreed with the explanation participated in the survey. The target number of participants was 700, including 350 females aged 20-29 years and 350 females aged 30-39 years, respectively. We defined gen Z as those aged ≤27 years and millennials as those aged 28-39 years at the time of response.

Survey

The survey was conducted on the web in March 2023, using the Ovey application. The web-based survey comprised 29 multiple choices, Likert, open-ended questions. Questions were related to respondents’ menstrual experience, menstrual cycle management behavior, usage of PTAs, and thoughts about PTAs. Depending on the preceding response, the number of questions to be answered could change; the participants responded to 15-26 questions.

The questions in the survey were developed based on previous research and the specific objectives of this study. We conducted pilot testing with a sample of 20 women. Based on the feedback and comments received during the pilot test, we revised the questionnaire to improve clarity and construction. On average, it took participants approximately 5 minutes to complete the questionnaire. We designed the survey to be concise to minimize the burden on respondents while still collecting essential information. The reporting of the survey was guided by the CHERRIES (Checklist for Reporting Results of Internet E-Surveys; Table S2 in Multimedia Appendix 2).

Interviews

Among the participants who responded to the survey, those who were available for FGIs were recruited. A total of 8 people were interviewed by selecting 4 from PTA users and 4 from nonusers. In May 2023, FGIs were conducted twice on the web, and each interview lasted approximately 60 minutes. The first and second interviews were conducted among 4 app users and 4 nonusers, respectively. FGIs were recorded on Zoom and transcribed automatically using Clova Note from Naver. The questionnaire for the FGI can be found in Table S3 in Multimedia Appendix 3 of the FGI guide. The reporting of the focus groups was guided by the COREQ (consolidated criteria for reporting qualitative research; Table S4 in Multimedia Appendix 4)

Data Analyses

Data analysis was performed in 3 parts. First, we gathered data from app users’ information. Second, a survey was conducted to determine women’s menstrual experiences, app use, and tracking practices. Cycle management level is an indicator used to represent the degree to which women manage their menstrual cycles, and it is divided into 6 levels. Level 0=No record; 1=Record sometimes; 2=Record regularly; 3=Record and predict next cycle; 4=Record, predict, and reference plans; and 5=Record, predict, and change plans. Logistic regression and a generalized linear model (with log link and gamma distribution) were used to analyze the survey data. Logistic regression was used to determine the association between user characteristics and app usage, and gamma regression was used to identify factors influencing cycle management level. We examined collinearity of variables and the results of the collinearity test indicated that the VIF values for all variables were confirmed to be 2.7 or below. Third, the FGI results of the verbalized transcripts were coded, and the transcriptions were analyzed using inductive thematic analysis. Two authors independently coded each record after reviewing the encoded transcripts, implying a significant user perspective. The team then repeatedly incorporated and altered the codes to create a set of themes. Finally, the themes were identified and described using illustrative participant comments. The collected data were quantitatively analyzed using STATA (version 17.0; StataCorp), and qualitative data were analyzed using NVivo software (version 14; QSR International).

Results

General Characteristics

A total of 700 responses were collected, and 8 responses which gave inconsistent answers were excluded from the screening process. Table 1 presents the demographics of the survey respondents. Most of these individuals were at the university level (n=450, 73.9%) in education, were employed (n=403, 58.2%), and were unmarried (n=475, 68.6%). Millennials made up 64% (n=443) of the total respondents, with the majority aged 30-34 years (n=190, 42.9%).

XSL•FO
RenderX

Table 1. Demographics of the total 692 survey respondents.

Characteristics	Overall	Generation Z	Millennials
Total, n (%)	692 (100.0)	249 (36.0)	443 (64.0)
Age in years, mean (95% CI)	29.9 (29.6-30.3)	24.3 (24.1-24.6)	33.1 (32.8-33.4)
Age groups (years), n (%)			
20-24	119 (17.2)	119 (47.8)	—[a]
25-29	226 (32.7)	130 (52.2)	96 (21.6)
30-34	190 (27.5)	—	190 (42.9)
35-39	157 (22.7)	—	157 (35.5)
Education[b], n (%)			
High school	56 (8.1)	18 (7.2)	38 (8.5)
College level	138 (19.9)	40 (16.1)	98 (22.2)
University level	450 (65)	184 (73.9)	266 (60)
Master/doctoral	48 (6.9)	7 (2.8)	41 (9.2)
Occupation, n (%)			
Student	107 (15.5)	100 (40.2)	7 (1.6)
Employee	403 (58.2)	101 (40.6)	302 (68.2)
Homemaker	69 (10)	2 (0.8)	67 (15.1)
Self-employee	19 (2.7)	2 (0.8)	17 (3.8)
Others	38 (5.5)	14 (5.6)	24 (5.4)
No occupation	56 (3.8)	30 (12)	26 (5.9)
Marriage, n (%)			
Unmarried	475 (68.6)	237 (95.2)	238 (53.7)
Married	217 (31.4)	12 (4.8)	205 (46.3)

[a]Not applicable.

[b]Education level includes both students and graduates.

Menstruation Experience and Cycle Management

The results of the survey on menstrual experiences and behaviors are presented in Table 2. Approximately half of the millennial and gen Z individuals reported a menarche age between 12 and 13 years. The most frequently reported average period length was 27-29 days (n=257, 36.9%), followed by 30-32 days (n=160, 23%), and 505 out of 692 (75.3%) had regular or relatively regular menstrual cycles. The proportion of those who reported irregular cycles (including relatively irregular cycles) was higher among gen Z than individuals than among millennials. A total of 68 respondents reported having severe PMS, and half of the respondents indicated that they experience strong or severe PMS. Regarding dysmenorrhea, more than half of both millennial and gen Z individuals experienced severe or strong period cramps. Regarding gynecologist visits over the past 3 years, excluding visits during pregnancy or childbirth, 445 (63.8%) respondents had never visited the clinic. Participants' regularity in recording their period duration and date was evaluated by the period cycle management level. In total, 614 (88.7%) respondents routinely recorded their menstrual cycles. A total of 431 (62.3%) respondents reported using PTAs to record and track their menstrual cycles, with no significant difference between millennials and gen Z in PTA use.

XSL•FO
RenderX

Table 2. Survey participants menstrual experience and behavior.

Category	Overall	Generation Z	Millennials
Total, n (%)	692 (100)	249 (100)	443 (100)
Menarche, n (%)			
Underage 12	114 (16.4)	48 (19.3)	65 (14.7)
12-13	343 (49.2)	112 (45)	231 (52.1)
14-16	185 (26.5)	68 (27.3)	117 (26.4)
Over 16	50 (7.2)	21 (8.4)	29 (6.5)
Period cycle, n (%)			
Do not know	19 (2.7)	8 (3.2)	11 (2.5)
Less than 20 days	9 (1.3)	4 (1.6)	5 (1.1)
21-23 days	42 (6)	19 (7.6)	23 (5.2)
24-26 days	77 (11)	24 (9.6)	53 (12)
27-29 days	257 (36.9)	79 (31.7)	178 (40.2)
30-32 days	160 (23)	59 (23.7)	101 (22.8)
33-40 days	93 (13.3)	43 (17.3)	50 (11.3)
More than 41 days	35 (5)	13 (5.2)	22 (5)
Cycle Regularity, n (%)			
Regular	202 (29)	66 (26.5)	135 (30.5)
Relatively regular	303 (43.5)	90 (36.1)	211 (47.6)
Relatively irregular	119 (17.1)	61 (24.5)	58 (13.1)
Irregular/amenorrhea	73 (10.5)	32 (12.9)	39 (8.8)
PMS[a] severity, n (%)			
None	10 (1.4)	4 (1.6)	6 (1.4)
Mild	63 (9)	28 (11.2)	35 (7.9)
Moderate	266 (38.2)	101 (40.6)	165 (37.2)
Strong	285 (40.9)	96 (38.6)	189 (42.7)
Severe	68 (9.8)	20 (8)	48 (10.8)
Dysmenorrhea severity, n (%)			
None	9 (1.3)	4 (1.6)	5 (1.1)
Mild	82 (11.8)	28 (11.2)	54 (12.2)
Moderate	236 (33.9)	84 (33.7)	152 (34.3)
Strong	289 (41.5)	111 (44.6)	178 (40.2)
Severe	76 (10.9)	22 (8.8)	54 (12.2)
Visit OBGY[b] in 3 years[c], n (%)			
None	445 (63.8)	162 (65.1)	283 (63.9)
1-2 times	171 (24.5)	62 (24.9)	109 (24.6)
3-5 times	51 (7.3)	16 (6.4)	35 (7.9)
6-9 times	14 (2)	4 (1.6)	10 (2.3)
More than 10 times	11 (1.6)	5 (2.0)	6 (1.4)
Cycle management level, n (%)			
No record	25 (3.6)	12 (4.8)	13 (2.9)
Record sometimes	53 (7.6)	24 (9.6)	29 (6.5)

XSL•FO
RenderX

Category	Overall	Generation Z	Millennials
Record regularly	235 (33.7)	86 (34.5)	149 (33.6)
Record and predict next cycle	151 (21.7)	44 (17.7)	107 (24.2)
Record, predict cycle, and reference plans	150 (21.5)	47 (18.9)	103 (23.3)
Record, predict cycle, and change plans	78 (11.2)	36 (14.5)	42 (9.5)
Cycle management method, n (%)			
No record	25 (3.6)	12 (4.8)	13 (2.9)
Calendar/diary (Paper)	75 (10.8)	23 (9.2)	52 (11.7)
Calendar/diary (Digital)	161 (23.3)	56 (22.5)	105 (23.7)
PTA[d]	431 (62.3)	158 (63.5)	273 (61.6)

[a]PMS: premenstrual syndrome.

[b]OBGY: obstetrics and gynecology.

[c]Number of visits to obstetricians and gynecologists related to menstruation, such as menstrual irregularities. Visits owing to pregnancy or childbirth were excluded.

[d]PTA: Period tracker apps.

Factors Affecting PTA Use

PTA user and nonuser characteristics are shown in Table 3. Chi-square analysis was used to analyze the differences between the 2 groups according to several predictors. App users and nonusers showed statistically significant differences in marital status (P=.006), with unmarried individuals being more likely to use PTA. Childbirth experience significantly impacts PTA app usage (P<.001), with those who have not experienced childbirth being more likely to use the app. Logistic regression was used to determine factors that affect PTA use; marriage, childbirth experience, number of PMS symptoms, and number of dysmenorrhea symptoms were found to be statistically significant (Table 4). As an important social factor, childbirth experience was found to affect PTA use. In women with childbirth experience, the likelihood of using a PTA decreased by 52.5% compared to women without childbirth experience. To estimate the levels of PMS and dysmenorrhea, we used 3 indicators: perceived severity, effect on daily life, and number of symptoms. Among these indicators, only the number of dysmenorrhea symptoms showed statistical significance. A higher number of dysmenorrhea symptoms is significantly associated with increased PTA usage (P=.04). Each additional symptom increases the likelihood of using PTA by approximately 13.6%. The number of dysmenorrhea symptoms and cycle management level were positively associated with app use. Individuals with higher cycle management levels are significantly more likely to use PTA (P<.001). They are more than twice as likely to use PTA compared to those with lower cycle management levels.

XSL•FO
RenderX

Table 3. PTA[a] users and nonuser's characteristics.

Predictor	Nonuser	App user	*P* value[b]
Total, n (%)	261 (37.7)	431 (62.3)	__[c]
Social factor			
Generation, n (%)			.63
Millennial	170 (38.4)	273 (61.6)	
Generation Z	91 (37)	158 (63.5)	
Marriage, n (%)			.006
Married	98 (45.2)	119 (54.8)	
Unmarried	163 (34.3)	312 (65.7)	
Childbirth experience, n (%)			<.001
Yes	80 (52.6)	72 (47.0)	
No	181 (33.5)	359 (66.5)	
Education level[d], n (%)			.54
High school graduate	24 (42.9)	32 (57.1)	
College level	53 (38.4)	85 (61.6)	
University level	170 (37.8)	280 (62.2)	
Graduate school level	14 (29.2)	34 (70.8)	
Menstrual experience			
Period regularity, n (%)			.64
Regular	68 (33.8)	133 (66.2)	
Relatively regular	119 (39.5)	182 (60.5)	
Relatively irregular	44 (37.0)	75 (63.0)	
Irregular	19 (44.2)	24 (55.8)	
Amenorrhea	11 (39.3)	17 (60.7)	
PMS severity, n (%)			.92
None	4 (40.0)	6 (60.0)	
Mild	22 (34.9)	41 (65.1)	
Moderate	99 (37.2)	167 (62.8)	
Strong	107 (37.5)	178 (62.5)	
Severe	29 (42.6)	39 (57.4)	
Dysmenorrhea severity, n (%)			.67
None	4 (44.4)	5 (55.6)	
Mild	25 (30.5)	57 (69.5)	
Moderate	93 (39.4)	143 (60.6)	
Strong	111 (38.4)	178 (61.6)	
Severe	28 (36.8)	48 (63.2)	

[a]PTA: Period tracking app.

[b]Derived from Chi-square test.

[c]Not applicable.

[d]Education level includes both students and graduates.

XSL•FO
RenderX

Table 4. Results of statistical analyses using logistic regression to determine likelihood of period tracking app (PTA) usage.

Variables	Crude OR[a]	*P* value[a]	95% CI[a]	Adjusted OR[b]	*P* value[b]	95% CI[b]
Social Factor						
Childbirth Experience	0.454	<.001	0.315-0.654	0.475	.04	0.237-0.950
Education level						
Baseline (High school graduate)	__[c]	—	—	—	—	—
College level	1.203	.57	0.640-2.260	1.021	.96	0.479-2.178
University level	1.235	.46	0.704-2.168	0.964	.92	0.492-1.891
Graduate school level	1.821	.15	0.805-4.123	2.177	.12	0.825-5.742
Menstrual experience						
Period irregularity	0.923	.07	0.796-1.071	1.096	.35	0.906-1.325
Effect of PMS[d] on daily life	1.040	.11	0.853-1.268	0.839	.36	0.577-1.219
Effect of dysmenorrhea on daily life	1.110	.10	0.921-1.338	0.935	.71	0.653-1.337
Number of PMS symptoms	1.225	.06	1.110-1.351	1.100	.18	0.957-1.264
Number of dysmenorrhea symptoms	1.209	.05	1.117-1.308	1.136	.04	1.008-1.280
Cycle management level	2.209	.18	1.889-2.582	2.279	<.001	1.919-2.707

[a]Crude odds ratio value from univariate model.

[b]Multivariate model. Adjusted odds ratio using the following predictors: age, marriage, menarche age, area group, visit obstetrics and gynecology in 3 years and occupation.

[c]Not applicable.

[d]PMS: premenstrual syndrome.

Factors Affecting Cycle Management Level

Two models were used to analyze the factors affecting cycle management level (Table 5). As shown in Table 2, cycle management level was categorized into 6 levels, from “no record” (level 0) to “record, predict cycle, and change plan” (level 5). A larger value indicates a more active attitude in cycle management level. Model 1 was established to investigate social factors such as childbirth experience and education level, while Model 2 was established to investigate menstrual experience. Childbirth experience was negatively associated with cycle management level. In women with childbirth experience, the odds of having a certain cycle management level decreased by approximately 9.8%. Different education levels were compared with the baseline data (high school level). College-level and University-level education were associated with increased odds of having a certain cycle management level, with odds ratios of 1.119 and 1.122, respectively. However, graduate school level did not show a significant effect on cycle management level.

XSL•FO
RenderX

Table 5. Results of statistical analyses using logistic regression to determine likelihood of period tracking app (PTA) usage.

Variables	Crude OR[a]	95% CI[a]	*P* value[a]	Adjusted OR[b]	95% CI[b]	*P* value[b]
Model 1 social factor						
Childbirth Experience	0.942	0.882-1.002	.06	0.902	0.809-0.995	.04
Education level						
Baseline (High school graduate)	__[c]	—	—	—	—	—
College level	1.122	1.020-1.225	.02	1.119	1.015-1.224	.03
University level	1.129	1.038-1.221	.006	1.122	1.028-1.216	.01
Graduate school level	1.094	0.967-1.221	.15	1.071	0.940-1.201	.29
Model 2 menstrual experience						
Period irregularity	0.928	0.905-0.951	<.001	0.929	0.903-0.954	<.001
Number of PMS[d] symptoms	1.007	0.989-1.025	.45	1.006	0.987-1.024	.54
Number of Dysmenorrhea symptoms	1.024	1.010-1.038	.001	1.024	1.010-1.039	.001

[a]Crude odds ratio value from univariate model.

[b]Multivariate model. Adjusted odds ratio using the following predictors: age, marriage, menarche age, area group, visit obstetrics and gynecology in 3 years and occupation.

[c]Not applicable.

[d]PMS: premenstrual syndrome.

Among the menstrual experience variables, period irregularity and number of dysmenorrhea symptoms had statistically significant effects on the cycle management level (with 95% confidence). When period irregularity increased by one level, the odds of reaching a certain cycle management level decreased by approximately 0.929. In contrast, the number of dysmenorrhea symptoms and cycle management levels were positively associated with an odds ratio of 1.024. (Ordinal logistic regression results are shown in Table S5 in Multimedia Appendix 5).

Results of the FGIs

FGIs were conducted among eight participants and their responses were recorded and coded. The profiles of interview participants can be found in Table S6 in Multimedia Appendix 6. The participants' comments and opinions revealed four major themes and fourteen subthemes as follows: (I) menstruation management: (1) self-calendar, (2) pain management, and (3) app use; (II) barriers to following the cycle management: (4) period irregularity, and (5) childbirth and care; (III) motivators to use apps: (6) dysmenorrhea, (7) pregnancy plan, (8) reminders, (9) reference to doctor meeting, (10) set/adjust schedule (11) self-health check; (IV) barriers to continued use of PTAs: (12) lack of app literacy, (13) lack of app accuracy, and (14) time and update issue (Table 6).

XSL•FO
RenderX

Table 6. Results of focus group interviews in thematic analysis.

Theme	Quote
Menstruation management	
Self-calendar	• *"I don't record it separately since it's a normal cycle, but sometimes it's difficult when I go out without sanitary pads. I sometimes note it in my Naver calendar."* [NU2, Unmarried] • *"I used to record the first date of period in my diary."* [NU1, 34, Married] • *"Ever since I was in high school, my mother has told me to check my calendar myself. So, I've been checking it since then."* [NU4, 39, Married] • *"I know when I need to take medicine so I can prepare in advance, so I guess I type it into my diary."* [NU3, 33, Unmarried]
Pain management	• *"I carry two hot packs and put one on my stomach and one on my back…and of course I also taking pain killers."* [NU3, 33, Unmarried] • *"I heard that Omega 3 is good, so I am taking it."* [NU3, 33, Unmarried] • *"I usually take about three 400mg pills of Naproxen-type menstrual pain reliever."* [NU3, 33, Unmarried] • *"I took painkillers for migraine and also get a shot for relieve migraine."* [NU4, 39, Married] • *"I usually take Supplement known for women health such as Vitamin B, evening primrose oil and take pills only when menstruation cramp is severe."* [AU3, 26, Unmarried] • *"When I have menstrual cramps, I just take one or two of ibuprofen pill and it gets better."* [AU1, 25, Married]
App use	• *"I'm using a Bom calendar to manage my cycle."* [AU1, 25, Married] • *"I am using My calendar (app name)."* [AU2, 34, Married] • *"I am using Samsung Health."* [AU3, 26, Unmarried] • *"I use two applications, one called p-tracker, an app provided by iOS, and one that checks cycles on the Apple Health app."* [AU4, 36, Unmarried]
Barriers to follow the cycle management	
Irregular menstrual cycle (Period irregularity)	• *"I also have severe menstrual irregularity, and my periods may be delayed by two or three months. Therefore, I was unable to follow up my cycle management."* [NU1, 34, Married] • *"My friend has an irregular menstrual cycle, and she always carries sanitary pads and medicines with her 365 days a year. That is why she doesn't bother with cycle management."* [NU3, 33, Unmarried] • *"The reason for recording the cycle in this way is that the purpose is to know the due date, but in the case of people who has irregular menstrual cycle, even if they know the due date, there may be cases where they do not menstruate at that time, so the need for cycle management is probably less than that of people with regular period."* [NU2, 26, Unmarried]
Childbirth and Care	• *"After giving birth, my interest in cycle management is a bit loose. Since I was focused on raising a child, I became less concerned about menstruation than before marriage."* [NU4, 39, Married] • *"My menstrual regularities were so bad that I thought about whether I could get pregnant, so I paid a lot of attention on cycle management before giving birth. Now that I have one child, I really don't manage my menstrual cycle."* [NU1, 34, Married] • *"I spend most of my days caring for my children. Thus, I don't have sufficient time to spend taking care of my cycle management and general health."* [NU4, 39, Married]
Motivators to adopt period track apps	
Dysmenorrhea	• *"I visited a gynecologist because my menstrual pain was quite severe. The doctor just informed me that, if my cycle is normal, I should start taking painkillers the day before my period. So, I started using an app to estimate my period and prepare for drugs in advance."* [AU2, 34, Married] • *"I used to have terrible menstrual cramps. Therefore, I was unable to fulfill my planned activities or attend class. So, I looked at the menstrual cycle app to predict my next cycle as well as make a proper schedule."* [AU2, 34, Married]
Pregnancy plan	• *"I mentioned it earlier, it will be useful to consult it now in order to confirm the fertility window by ovulation period for all pregnancy planning."* [AU2, 34, Married] • *"I used to use a menstrual cycle app rarely. But now that I'm married and preparing for pregnancy, So, I use it more often."* [AU2, 34, Married]
Reminders	• *"The app is very useful because it reminds me at the time of period"* [AU4, 36, Unmarried]
Reference to doctor meeting	• *"I set my annual health checkup schedule with gynecologist based on expected cycle in the app"* [AU4, 36, Unmarried] • *"It is helpful when I talk with gynecologist"* [AU1, 25, Married]

XSL•FO
RenderX

Theme	Quote
Set/adjust schedule	• *"As for the advantage of using the app…I think it may be that it is easier to adjust the schedule. I can avoid Important schedules during the period."* [AU4, 36, Unmarried]
Self-health check	• *"In that app, I can now record details such as my weight and height, as well as my health status. So, I can see how my weight has changed over time, and I think it's great to be able to see in depth of the physical changes too."* [AU1, 25, Married] • *"I can notice my body condition is not good when the cycle became irregular."* [AU3, 26, Unmarried] • *"I could notice my emotion is unstable because of hormone in premenstrual period"* [AU4, 36, Unmarried]
Barriers to continued use of period track apps	
Lack of app literacy	• *"Some app features are complex. And there is an app need to touch every single day menstruation is in progress. That is inconvenient."* [AU4, 36, Unmarried]
Lack of app accuracy	• *"I tried to use the app but quit to use because of app's inaccuracy to predict next cycle"* [NU3, 33, Unmarried]
Time and update issue	• *"I often forget to record my cycle. So, I delete the app"* [NU2, 26, Unmarried] • *"I could not follow up the app use. Because I don't have much time and concentration"* [NU1, 34, Married] • *"I have no complaints about how to use it or the UI, but sometimes, when updates are made or something like that, the menstruation category disappears, which happens once or twice a year. I keep inquiring about things like that and I'm trying to resolve them somehow by updating."* [AU3, 26, Unmarried]

PTA User Experience

In total, 431 participants used PTAs as a tool to record their menstrual cycle. More than two-thirds (n=290, 67.3%) of the PTA users have been app users for more than 3 years, and 116 (26.9%) participants have been app users for more than 7 years. The most frequently used application was the Pink diary, which is an official PTA accredited by the Korean Association of Obstetricians and Gynecologists, and 103 (23.9%) respondents were using this app [28]. In addition, a few respondents (n=65, 15%) reported using a period tracker function embedded in their mobile phone health applications, such as Apple Health and Samsung Health (Table S7 in Multimedia Appendix 7).

Figure 2 shows the reasons for using PTAs. Most of the app users (n=368, 85%) used the app to predict their next cycle. More than half of the app users used a PTA to keep record (n=235, 54.5%) and to set or adjust their schedule (n=234, 54.3%). The most frequently used feature of PTAs was cycle prediction; 348 (80%) app users used cycle prediction features in their app and period records, while more than half of the respondents used the next period alarm and fertile window prediction features.

XSL•FO
RenderX

Figure 2. Reasons for using period tracker apps (A) and frequently used features in period tracker apps (B). OBGY: obstetrics and gynecology.

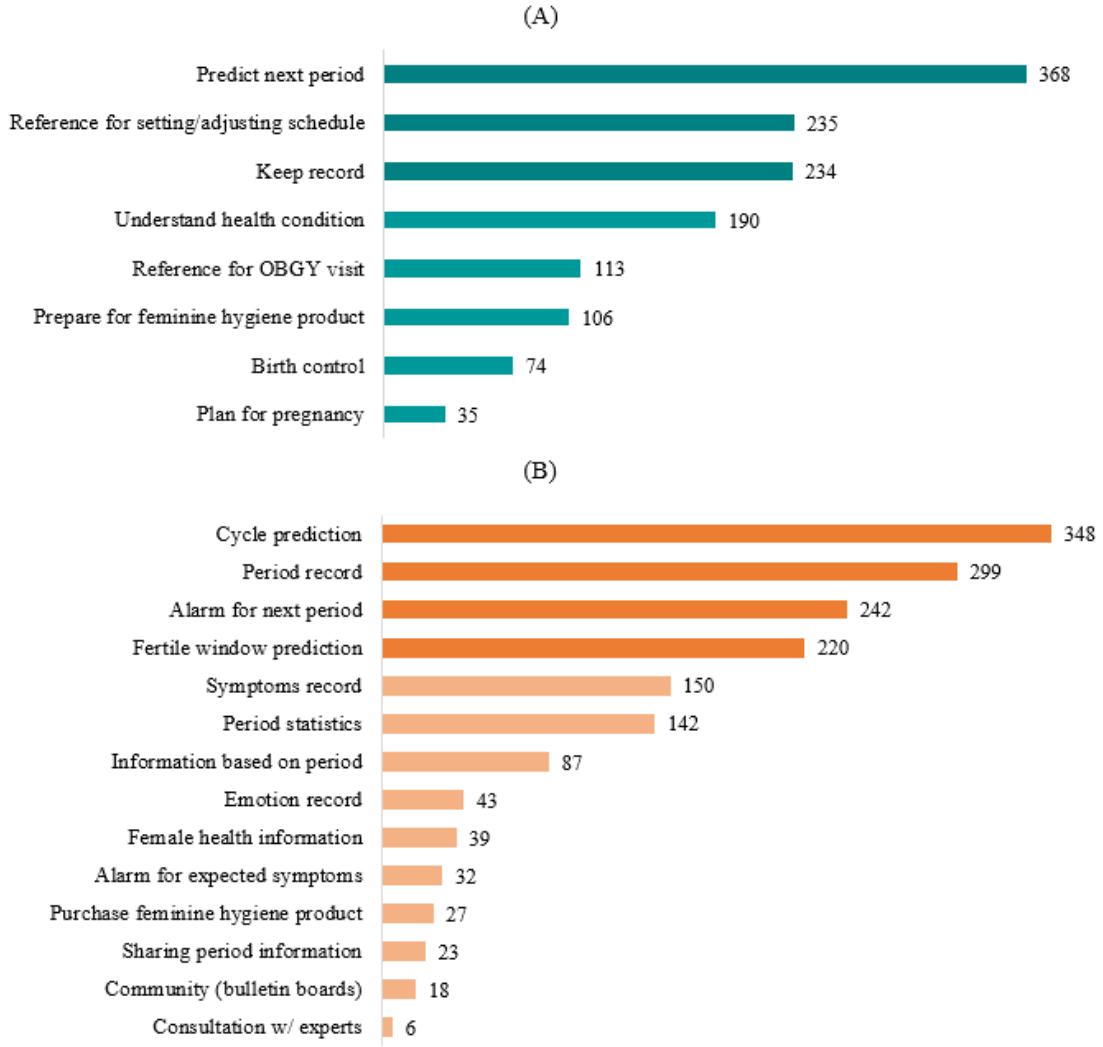

We asked about the accuracy of the cycle prediction and the utility level of the app; 199 (46.2%) app users replied that their app predicted the next cycle with a 2-day error range more than 10 times a year. Some of the respondents who answered that their app was inaccurate explained that this is because of their irregular menstrual cycle. Regarding PTA utility, 387 app users (90%) considered PTAs useful. PTAs played an important role in the healthy menstrual management of the app users in this study (Figure 3).

Figure 3. Users' response about predictive accuracy (A) and usability (B) of the period tracker app.

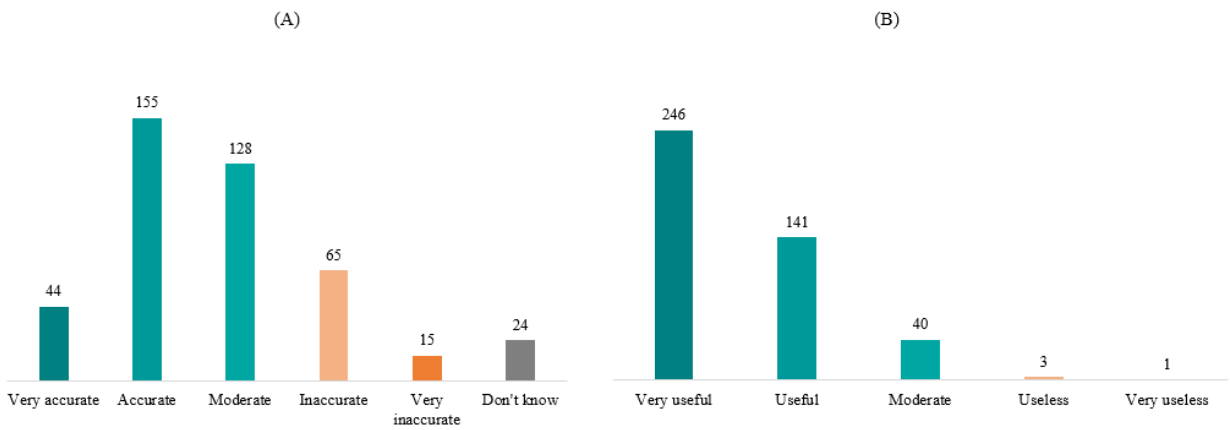

XSL•FO
RenderX

> *Using apps is always basic thing when doing something. Basically, I think I should be aware of my health. Women should keep records.* [AU2]
>
> *PTA account for 80point out of 100(in my menstrual management). Menstrual information is very critical and only information to determine my body condition and uterine health so that period tracking is very important.* [AU3]

In total, 91 (28%) app users complained about seeing too many advertisements on the app. Advertisement is related to the price policy of the apps. Most PTAs provide their apps free of charge to use the main feature and adopt an in-app charge policy to use additional features or eliminate advertisements (Figure S1 in Multimedia Appendix 8). Similarly, most health applications often display pertinent information and commercial ads that enable users to monitor and manage their own health [15,19].

In this study, 414 app users (96.1%) use their apps free of charge. Under these circumstances, it is difficult for app companies to invest in and improve their apps. However, app users are willing to pay when the app has improved accuracy in cycle prediction and features.

> *If it's a subscription concept, I'm willing to pay 1,000 won per month. so, if it's between 10,000 and 12,000 won per year, I'm willing to pay if I get satisfactory usability from the app.* [AU4]
>
> *If I'm really satisfied with the app, I think I'll pay about 20,000-30,000 won a year to use it.* [AU3]
>
> *It costs less than 10,000 won to download, and if it is really accurate and helpful, I think I will get it.* [AU2]
>
> *If it's a good app, I'd like to spend around 10,000 won per year, and if it's a one-time cost, I'll probably spend up to around 16,000 won.* [AU1]

Discussion

Principal Findings

This study aimed to investigate (1) menstrual cycle management and app usage status, (2) factors affecting PTA usage, (3) factors affecting period tracker cycle management, and (4) motivators and barriers to using PTAs, in millennial and gen Z women.

We found that PTAs are widely used among young women—millennials and gen Z—in South Korea and play an important role in menstrual cycle management. Approximately 90% of the women recorded their menstrual cycles regularly, and over 60% of the respondents used PTA.

The social factors influencing PTA usage included childbirth experience, while menstrual experience factors included the number of PMS symptoms, the number of dysmenorrhea symptoms, and the cycle management level. After childbirth, women were less likely to use PTAs to manage their cycles because of childcare and no concerns about infertility. In addition, women who have given birth tend to neglect their own health management due to childcare responsibilities, so if an app is released that allows them to record postpartum health management and health management according to their child's growth, and receive information, it is expected to further enhance user convenience. To better serve postpartum users, the app should include features tailored to their needs, such as tracking postpartum symptoms, fertility awareness, and contraceptive management. By addressing the specific needs of this demographic, the app can enhance its relevance and usefulness.

Regarding education level, college- and university-level education were positively associated with cycle management level, consistent with the finding that women with higher education levels are likely to manage their cycle more thoroughly compared with those with high school level education. Previous studies have reported that PMS and dysmenorrhea has a negative impact on young women's education and quality of life [29-38]. It was also consistent with previous research showing that people with dysmenorrhea have higher rates of mental health issues, including anxiety and depression [39], a lower quality of life [4,40-42], and a higher lifetime risk of developing comorbid pain conditions [43]. Another study indicated that adolescent girls' education is affected by dysmenorrhea because of the lack of proper hygiene during menstruation, leading to a higher proportion of dropouts [39]. Therefore, cycle management is an important aspect of women's education. The more menstrual symptoms a woman experiences, the more diligent she tends to be at managing her cycle. For student users, PTA should include educational tips to help them understand the implications of their symptoms and manage them effectively. This could cover topics such as PMS, dysmenorrhea, and effective cycle management strategies, tailored to accommodate academic schedules.

Our study findings reported that the period irregularity was negatively associated with cycle management level. Women with irregular cycle were less likely to manage their menstrual cycle because of their unpredictable cycle, causing them to prepare for their period every day and to give up on predicting their period. Consistent with previous research on PTA, women with irregular menstrual cycles often find the app to be unhelpful due to its inaccuracies [9]. The existing research offers minimal information on which monitoring capabilities of cycle management that are most crucial for menstruating app users, even though tracking cycle and regularity differed throughout apps [19,21,22]. Thus, it is important to include features that allow users to customize and adapt the app's predictions based on their unique cycle patterns. This could include flexible tracking options for users with irregular cycles, predictive adjustments based on past patterns, and alerts for irregularities. Enhanced algorithms that accommodate a range of cycle irregularities could also improve user experience.

The PTA is used by most app users to predict the next cycle, keep cycle records, and set or change schedules. Regarding the accuracy of PTAs in predicting the next cycle, less than half of app users reported that it was highly accurate. Regardless of the level of accuracy, nearly 90% of app users said that PTAs were useful for menstrual cycle management.

Strength of the Study

This study has several strengths. First, it is the first study to investigate PTA usage in South Korea. In this study, we

XSL•FO
RenderX

elucidated current menstrual management behaviors among millennials and gen Z women. Second, this is the first study to identify social and experiential factors that influence PTA usage and menstrual cycle management. Third, a mixed methods approach was employed, combining quantitative data analysis with qualitative insights gathered from FGIs with the survey participants. This approach allowed for a deeper understanding of the results and provided participants with the opportunity to discuss the data analysis outcomes and share their experiences in detail.

Limitations

This study has certain limitations. First, there are no universally agreed-upon criteria for evaluating the accuracy of PTAs in predicting the next menstrual cycle. We used arbitrary criteria in the survey to simplify participants' choices. Future studies should focus on developing standardized evaluation criteria for PTAs. Second, we conducted FGIs with only 2 groups (app user and nonuser) covering eight women which may not capture entire gamut of issues related to cycle management and PTAs. Third, owing to privacy concerns, we were unable to contact interviewees for follow-up questions after the FGIs. In future research, providing an opportunity for follow-up questions is essential to validate findings and gather additional insights. Fourth, there are potential biases introduced by recruiting participants from a panel, noting that they may not be representative of the general population and may have a particular interest in menstrual health.

Conclusions

In conclusion, PTAs are becoming the new normal among millennials and gen Z for managing their menstrual cycles. The use of a PTA empowers women by helping them gain a better understanding of their bodies, ultimately enhancing their social, academic, and health-related lives. Improving the accuracy and literacy of the app is an ongoing task for period-tracking apps, and it is important to provide added value tailored to users' needs. Therefore, the findings of this study should be considered when designing or upgrading PTAs to facilitate the adoption of digital technology among young women, thereby promoting their overall well-being and reproductive health.

Acknowledgments

We thank the participants for sharing their experiences. This study was supported by the Healthcare Management Research Fund of the Yonsei University Graduate School of Public Health. The funding body had no role in the study design, data collection and analysis, or data interpretation.

Data Availability

The datasets generated and analyzed in the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

MJ led the project's conceptualization, developed the survey questionnaire, conducted focus interviews, performed data analysis, created visualizations, and drafted the original manuscript. SG contributed to the conceptualization, secured funding, validated results, and reviewed and edited the manuscript. VR contributed to the content analysis of the focus group interviews and also reviewed and edited the manuscript. KY handled data curation and conducted the literature review. SY focused on statistical analysis. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey information sheet.
[DOCX File , 17 KB - jmir_v26i1e53146_app1.docx]

Multimedia Appendix 2
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 20 KB - jmir_v26i1e53146_app2.docx]

Multimedia Appendix 3
Focus group interview guide.
[DOCX File , 20 KB - jmir_v26i1e53146_app3.docx]

Multimedia Appendix 4
Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups.
[DOCX File , 20 KB - jmir_v26i1e53146_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Results of statistical analyses using ordinal logistic regression to determine factors to affect cycle management.
[DOCX File , 19 KB - jmir_v26i1e53146_app5.docx]

Multimedia Appendix 6
Focus group interview participants profile.
[DOCX File , 17 KB - jmir_v26i1e53146_app6.docx]

Multimedia Appendix 7
Information of period tracker application users.
[DOCX File , 18 KB - jmir_v26i1e53146_app7.docx]

Multimedia Appendix 8
App users' complaints when using the period tracker app, non–app user's reasons not to use the period tracker app, and improvements needed to make nonusers willing to use the period tracker app.
[DOCX File , 111 KB - jmir_v26i1e53146_app8.docx]

References

1. Sveinsdóttir H. Menstruation, objectification and health-related quality of life: a questionnaire study. J Clin Nurs 2018;27(3-4):e503-e513. [doi: 10.1111/jocn.14049] [Medline: 28833784]
2. Iacovides S, Avidon I, Bentley A, Baker FC. Reduced quality of life when experiencing menstrual pain in women with primary dysmenorrhea. Acta Obstet Gynecol Scand 2014;93(2):213-217 [FREE Full text] [doi: 10.1111/aogs.12287] [Medline: 24266425]
3. Tanaka E, Momoeda M, Osuga Y, Rossi B, Nomoto K, Hayakawa M, et al. Burden of menstrual symptoms in Japanese women: results from a survey-based study. J Med Econ 2013;16(11):1255-1266 [FREE Full text] [doi: 10.3111/13696998.2013.830974] [Medline: 24015668]
4. Ünsal A, Tözün M, Aslan G, Ayranci U, Alkan G. Evaluation of dysmenorrhea among women and its impact on quality of life in a region of western Turkey. Pak J Med Sci 2010;1:26.
5. Chen CX, Draucker CB, Carpenter JS. What women say about their dysmenorrhea: a qualitative thematic analysis. BMC Womens Health 2018;18(1):47 [FREE Full text] [doi: 10.1186/s12905-018-0538-8] [Medline: 29499683]
6. Schoep ME, Nieboer TE, van der Zanden M, Braat DDM, Nap AW. The impact of menstrual symptoms on everyday life: a survey among 42,879 women. Am J Obstet Gynecol 2019;220(6):569.e1-569.e7. [doi: 10.1016/j.ajog.2019.02.048] [Medline: 30885768]
7. Schantz JS, Fernandez CSP, Anne Marie ZJ. Menstrual cycle tracking applications and the potential for epidemiological research: a comprehensive review of the literature. Curr Epidemiol Rep 2021;8(1):9-19 [FREE Full text] [doi: 10.1007/s40471-020-00260-3] [Medline: 34055569]
8. Levy J, Romo-Avilés N. "A good little tool to get to know yourself a bit better": a qualitative study on users' experiences of app-supported menstrual tracking in Europe. BMC Public Health 2019;19(1):1213 [FREE Full text] [doi: 10.1186/s12889-019-7549-8] [Medline: 31481043]
9. Gambier-Ross K, McLernon DJ, Morgan HM. A mixed methods exploratory study of women's relationships with and uses of fertility tracking apps. Digit Health 2018;4:2055207618785077 [FREE Full text] [doi: 10.1177/2055207618785077] [Medline: 30225095]
10. Song M, Kanaoka H. Effectiveness of mobile application for menstrual management of working women in Japan: randomized controlled trial and medical economic evaluation. J Med Econ 2018;21(11):1131-1138 [FREE Full text] [doi: 10.1080/13696998.2018.1515082] [Medline: 30130990]
11. Grieger JA, Norman RJ. Menstrual cycle length and patterns in a global cohort of women using a mobile phone app: retrospective cohort study. J Med Internet Res 2020;22(6):e17109 [FREE Full text] [doi: 10.2196/17109] [Medline: 32442161]
12. Lupton D, Maslen S. How women use digital technologies for health: qualitative interview and focus group study. J Med Internet Res 2019;21(1):e11481 [FREE Full text] [doi: 10.2196/11481] [Medline: 30681963]
13. Earle S, Marston HR, Hadley R, Banks D. Use of menstruation and fertility app trackers: a scoping review of the evidence. BMJ Sex Reprod Health 2021;47(2):90-101 [FREE Full text] [doi: 10.1136/bmjsrh-2019-200488] [Medline: 32253280]
14. Worsfold L, Marriott L, Johnson S, Harper JC. Period tracker applications: what menstrual cycle information are they giving women? Womens Health (Lond) 2021;17:17455065211049905 [FREE Full text] [doi: 10.1177/17455065211049905] [Medline: 34629005]
15. Moglia ML, Nguyen HV, Chyjek K, Chen KT, Castaño PM. Evaluation of smartphone menstrual cycle tracking applications using an adapted applications scoring system. Obstet Gynecol 2016;127(6):1153-1160. [doi: 10.1097/AOG.0000000000001444] [Medline: 27159760]

XSL•FO
RenderX

16. Ko S, Lee J, An D, Woo H. Menstrual tracking mobile app review by consumers and health care providers: quality evaluations study. JMIR Mhealth Uhealth 2023;11:e40921. [doi: 10.2196/40921]
17. Epstein D, Lee N, Kang J, Agapie E, Schroeder J, Pina LR, et al. Examining menstrual tracking to inform the design of personal informatics tools. Proc SIGCHI Conf Hum Factor Comput Syst 2017;2017:6876-6888 [FREE Full text] [doi: 10.1145/3025453.3025635] [Medline: 28516176]
18. Jacobson AE, Vesely SK, Haamid F, Christian-Rancy M, O'Brien SH. Mobile application vs paper pictorial blood assessment chart to track menses in young women: a randomized cross-over design. J Pediatr Adolesc Gynecol 2018;31(2):84-88. [doi: 10.1016/j.jpag.2017.09.009] [Medline: 29030160]
19. Lee J, Kim J. Can menstrual health apps selected based on users' needs change health-related factors? A double-blind randomized controlled trial. J Am Med Inform Assoc 2019;26(7):655-666 [FREE Full text] [doi: 10.1093/jamia/ocz019] [Medline: 30946478]
20. Dantas LO, Carvalho C, Pena C, Breda CC, Driusso P, Ferreira CHJ, et al. Mobile health technologies for the monitoring of menstrual cycle: a systematic review of online stores in Brazil. J Obstet Gynaecol Res 2022;48(1):5-14. [doi: 10.1111/jog.15082] [Medline: 34706123]
21. Karasneh RA, Al-Azzam SI, Alzoubi KH, Muflih SM, Hawamdeh SS. Smartphone applications for period tracking: rating and behavioral change among women users. Obstet Gynecol Int 2020;2020:2192387 [FREE Full text] [doi: 10.1155/2020/2192387] [Medline: 32952563]
22. Eschler J, Menking A, Fox S, Backonja U. Defining menstrual literacy with the aim of evaluating mobile menstrual tracking applications. Comput Inform Nurs 2019;37(12):638-646. [doi: 10.1097/CIN.0000000000000559] [Medline: 31524688]
23. Starič KD, Trajkovik V, Belani H, Vitagliano A, Bukovec P. Smart phone applications for self-monitoring of the menstrual cycle: a review and content analysis. CEOG 2019;46(5):731-735. [doi: 10.12891/ceog4830.2019]
24. Ford EA, Roman SD, McLaughlin EA, Beckett EL, Sutherland JM. The association between reproductive health smartphone applications and fertility knowledge of Australian women. BMC Womens Health 2020;20(1):45 [FREE Full text] [doi: 10.1186/s12905-020-00912-y] [Medline: 32131809]
25. Creswell JW. A Concise Introduction to Mixed Methods Research. Washington, DC: SAGE publications; 2014.
26. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
27. Creswell JW, Klassen AC, Plano CV, Smith K. Best practices for mixed methods research in the health sciences. Bethesda (Maryland): National Institutes of Health 2013;2011:541-545.
28. KAOG. Korean association of obstetricians and gynecologist webpage. KAOG. URL: https://www.kaog.org [accessed 2023-10-09]
29. Ortiz MI, Rangel-Flores E, Carrillo-Alarcón LC, Veras-Godoy HA. Prevalence and impact of primary dysmenorrhea among Mexican high school students. Int J Gynaecol Obstet 2009;107(3):240-243. [doi: 10.1016/j.ijgo.2009.07.031] [Medline: 19716130]
30. Pitangui ACR, Gomes MRDA, Lima AS, Schwingel PA, Albuquerque APDS, de Araújo RC. Menstruation disturbances: prevalence, characteristics, and effects on the activities of daily living among adolescent girls from Brazil. Journal of Pediatric and Adolescent Gynecology 2013;26(3):148-152. [doi: 10.1016/j.jpag.2012.12.001]
31. Hillen TI, Grbavac SL, Johnston PJ, Straton JA, Keogh JM. Primary dysmenorrhea in young Western Australian women: prevalence, impact, and knowledge of treatment. J Adolesc Health 1999;25(1):40-45. [doi: 10.1016/s1054-139x(98)00147-5] [Medline: 10418884]
32. Roberts SC, Hodgkiss C, DiBenedetto A, Lee E. Managing dysmenorrhea in young women. Nurse Pract 2012;37(7):47-52. [doi: 10.1097/01.npr.0000415245.88638.bf] [Medline: 22924171]
33. Maity S, Wray J, Coffin T, Nath R, Nauhria S, Sah R, et al. Academic and social impact of menstrual disturbances in female medical students: a systematic review and meta-analysis. Front Med (Lausanne) 2022;9:821908 [FREE Full text] [doi: 10.3389/fmed.2022.821908] [Medline: 35242785]
34. Chéileachair FN, McGuire BE, Durand H. Coping with dysmenorrhea: a qualitative analysis of period pain management among students who menstruate. BMC Womens Health 2022;22(1):407 [FREE Full text] [doi: 10.1186/s12905-022-01988-4] [Medline: 36199106]
35. Aktaş D. Prevalence and factors affecting dysmenorrhea in female university students: effect on general comfort level. Pain Manag Nurs 2015;16(4):534-543. [doi: 10.1016/j.pmn.2014.10.004] [Medline: 26256218]
36. Shah V, Nabwera H, Sonko B, Bajo F, Faal F, Saidykhan M, et al. Effects of menstrual health and hygiene on school absenteeism and drop-out among adolescent girls in rural Gambia. Int J Environ Res Public Health 2022;19(6):3337 [FREE Full text] [doi: 10.3390/ijerph19063337] [Medline: 35329020]
37. Durand H, Monahan K, McGuire BE. Prevalence and impact of dysmenorrhea among university students in Ireland. Pain Med 2021;22(12):2835-2845 [FREE Full text] [doi: 10.1093/pm/pnab122] [Medline: 33822197]
38. Knox B, Azurah A, Grover S. Quality of life and menstruation in adolescents. Curr Opin Obstet Gynecol 2015;27(5):309-314. [doi: 10.1097/GCO.0000000000000199] [Medline: 26208045]
39. Bajalan Z, Moafi F, MoradiBaglooei M, Alimoradi Z. Mental health and primary dysmenorrhea: a systematic review. J Psychosom Obstet Gynaecol 2019;40(3):185-194. [doi: 10.1080/0167482X.2018.1470619] [Medline: 29745745]

XSL•FO
RenderX

40. Branecka-Woźniak D, Cymbaluk-Płoska A, Kurzawa R. The impact of premenstrual syndrome on women's quality of life - a myth or a fact? Eur Rev Med Pharmacol Sci 2022;26(2):598-609 [FREE Full text] [doi: 10.26355/eurrev_202201_27887] [Medline: 35113436]
41. Goker A, Artunc-Ulkumen B, Aktenk F, Ikiz N. Premenstrual syndrome in Turkish medical students and their quality of life. J Obstet Gynaecol 2015;35(3):275-278. [doi: 10.3109/01443615.2014.948820] [Medline: 25140580]
42. Trépanier LCM, Lamoureux É, Bjornson SE, Mackie C, Alberts NM, Gagnon MM. Smartphone apps for menstrual pain and symptom management: a scoping review. Internet Interv 2023;31:100605 [FREE Full text] [doi: 10.1016/j.invent.2023.100605] [Medline: 36761398]
43. Gagnon MM, Moussaoui D, Gordon JL, Alberts NM, Grover SR. Dysmenorrhea across the lifespan: a biopsychosocial perspective to understanding the dysmenorrhea trajectory and association with comorbid pain experiences. Pain 2022;163(11):2069-2075. [doi: 10.1097/j.pain.0000000000002649] [Medline: 35420567]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COREQ: consolidated criteria for reporting qualitative research
FGI: focus group interview
Gen Z: generation Z
PMS: premenstrual syndrome
PTA: Period Tracker App

Edited by N Cahill; submitted 27.09.23; peer-reviewed by X Yan, MR Pradhan; comments to author 06.04.24; revised version received 05.05.24; accepted 15.08.24; published 10.10.24.

Please cite as:
Hong M, Rajaguru V, Kim K, Jang SY, Lee SG
Menstrual Cycle Management and Period Tracker App Use in Millennial and Generation Z Individuals: Mixed Methods Study
J Med Internet Res 2024;26:e53146
URL: https://www.jmir.org/2024/1/e53146
doi:10.2196/53146
PMID:

©Minji Hong, Vasuki Rajaguru, KyungYi Kim, Suk-Yong Jang, Sang Gyu Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Associations of Wearable Activity Tracker Use With Physical Activity and Health Outcomes in Patients With Cancer: Findings from a Population-Based Survey Study

Weijiao Zhou[1*], PhD, RN; Shaomei Shang[1*], MSc, RN; Youmin Cho[2], PhD, RN, AGPCNP-BC

[1]School of Nursing, Peking University, Beijing, China

[2]College of Nursing, Chungnam National University, Daejeon, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Youmin Cho, PhD, RN, AGPCNP-BC
College of Nursing
Chungnam National University
266 Munhwa-ro, Jung-gu
Daejeon, 35015
Republic of Korea
Phone: 82 42 580 8336
Fax: 82 42 580 8309
Email: youmcho@cnu.ac.kr

Abstract

Background: Physical inactivity is a global issue for cancer survivors. Wearable activity trackers are promising to address physical inactivity by providing real-time feedback on physical activity and offering opportunities for self-monitoring and goal setting. Meta-analysis has reported the effects of interventions that incorporate wearable activity trackers on improved physical inactivity and related health outcomes (eg, BMI, anxiety and depression, and self-rated health status). However, wearable activity trackers were often used as an adjunct to physical activity interventions, and the effectiveness of wearable activity trackers alone is unknown.

Objective: This study aims to determine the association of wearable activity trackers with physical activity and health outcomes in patients with cancer.

Methods: Data from 957 cancer survivors from the Health Information National Trends Survey–Surveillance, Epidemiology, and End Results (HINTS-SEER) were analyzed. The outcome variables examined were time spent in moderate to vigorous physical activity, weekly frequency of strength training, BMI, anxiety and depression levels, and self-assessed health status. The primary independent variable was whether cancer survivors had used wearable activity trackers within the past 12 months. Design-based linear regression for continuous outcome variables and ordinal logistic regression for ordinal outcome variables were conducted to determine the associations after controlling for sociodemographic, cancer-related, and health-related factors. All data analyses accounted for the complex survey design and sample weights.

Results: Only 29% of cancer survivors reported wearable activity tracker use. Bivariate analyses showed that younger age ($P<.001$), higher education ($P=.04$), higher income ($P<.001$), and an employed status ($P<.001$) were significantly associated with wearable activity tracker use. Wearable activity tracker use was significantly associated with higher time spent in moderate to vigorous physical activity (adjusted =37.94, 95% CI 8.38-67.5; $P=.01$), more frequent strength training per week (adjusted odds ratio [OR] 1.50, 95% CI 1.09-2.06; $P=.01$), and better self-rated health status (adjusted OR 1.58, 95% CI 1.09-2.29; $P=.01$), but not with BMI or anxiety and depression.

Conclusions: This study suggests that the uptake of wearable activity trackers is low and highlights the digital divide among patients with cancer. This study has confirmed the associations of wearable activity tracker use with physical activity and self-rated health, supporting using wearable activity trackers as a promising tool to facilitate physical activity promotion.

(J Med Internet Res 2024;26:e51291) doi:10.2196/51291

XSL•FO
RenderX

KEYWORDS

physical activity; exercise; wearable tracker; wearable device; cancer; oncology; cancer survivorship; wearable; wearables; tracker; trackers; health outcome; health outcomes; HINTS; survey; surveys; national; bivariate; epidemiology; epidemiological; association; associations

Introduction

Physical activity improves health outcomes of patients with cancer, including cardiovascular fitness, muscle strength, cancer-related fatigue, health-related quality of life, and depression [1-3]. Physical activity guidelines for cancer survivors recommend at least 150 minutes of moderate physical activity per week [4]. However, only a minority of cancer survivors meet the physical activity recommendations [5,6]. Physical inactivity is a global issue for cancer survivors, and much work remains to be done to improve physical activity [7].

Wearable activity trackers offer an appealing, low-cost tool to address physical inactivity and, therefore, improve related health outcomes (eg, physiological outcomes, psychological outcomes, and self-rated health status). These devices provide real-time feedback on physical activity and offer opportunities for self-monitoring and goal setting. A review of systematic reviews [8] has reported that interventions that incorporate wearable activity trackers improve moderate to vigorous physical activity with an average effect size of 0.3 (6 minutes per day); an increase of 5-10 minutes per day of moderate to vigorous physical activity is considered clinically significant [9]. A meta-analysis of cancer survivors has found that wearable activity trackers improve moderate to vigorous physical activity with an effect size of 0.6 [10]. In addition, using wearable activity trackers has been shown to be associated with improved physiological outcomes (eg, BMI) [8], and it has the potential to improve psychological outcomes (eg, anxiety and depression) [8], and achieve positive self-rated health [11]. However, wearable activity trackers were often used as an adjunct to physical activity intervention or weight loss programs, which limits the ability to rule out the influence of other physical activity intervention components and determine the effectiveness of wearable activity trackers alone.

A few observational studies have examined the associations of wearable activity trackers with physical activity among general adults [12] and older adults [13], but the generalization to patients with cancer is limited. To address this gap, this study aimed to determine the association of wearable activity trackers with physical activity (moderate to vigorous physical activity and strength training) and health outcomes (BMI, anxiety and depression, and self-rated health status) in patients with cancer, using a population-based survey dataset obtained from 3 large cancer registries.

Methods

Data Sources and Study Population

We analyzed the Health Information National Trends Survey–Surveillance, Epidemiology, and End Results (HINTS-SEER) dataset. The Health Information National Trends Survey (HINTS) is a nationally representative survey that examined the knowledge, attitudes, and usage of cancer- and health-related information by American adults [14]. The initial HINTS encompassed both adults with and those without cancer, whereas the HINTS-SEER focused on a larger sample of adults with cancer from 3 SEER cancer registries in Iowa, New Mexico, and the Greater Bay Area of California, excluding individuals with only non–melanoma skin cancer diagnoses [15]. The data collection for HINTS-SEER took place by mail between January and August 2021. Among 1234 cancer survivors in the final HINTS-SEER sample, 277 respondents were excluded (21 had missing information about wearable activity use, 98 had missing information about outcome variables, and 158 had missing information about covariates), resulting in a study sample of 957 respondents in our final analyses. There was no significant difference in the outcome variables between individuals with complete and incomplete data ($P>.05$). The National Cancer Institute granted permission for the data used in this study.

Measures

Independent Variable—Wearable Activity Tracker Use

Respondents were asked whether they have used electronic wearable activity trackers, such as a Fitbit, Apple Watch, or Garmin Vivofit, to monitor or track their health or activity in the past 12 months (yes=1, no=0).

Outcome Variables

Physical activity variables included the following two items: (1) time spent in moderate to vigorous physical activity, that respondents were asked how many minutes in a week they do any physical activity or exercise of at least moderate intensity such as brisk walking, bicycling, or swimming, and (2) weekly frequency of strength training, that the respondents were asked how many days they do physical activity or exercise specifically designed to strengthen their muscles such as lifting weights or circuit training. Health-related outcome variables included the following three items: (1) BMI, (2) anxiety and depression evaluated by the Patient Health Questionnaire-4 (ranged 0-12, a higher score means a higher level of anxiety and depression), and (3) self-rated health status (0=poor to 4=excellent).

Covariates

Sociodemographic characteristics included age (18-49, 50-64, 65-75, and 76 years or older), sex (male or female), race (White and non-White), education level (less than high school, high school, some college, or college graduate or higher), income, and employment status (employed or unemployed). Cancer-related factors included cancer types (breast, prostate, colorectal, skin, or other), stage (localized, regional, or distant), and time since diagnosis (<1 year, 2-5 years, 6-10 years, and 11 or more years). Other health-related factors included BMI (categorized as underweight, normal, overweight, or obese) and smoking status (current, former, or never). Of note is that we controlled for BMI in all other analyses, except for the linear

XSL•FO
RenderX

regression, to examine BMI as the outcome variable. All data were self-reported.

Statistical Analysis

Descriptive statistics (mean, SE, and %) were calculated to describe wearable activity tracker use rate, physical activity, and health-related outcome variables, sociodemographic characteristics, and cancer- and health-related factors. Bivariate analyses were used to compare the sociodemographic and health-related characteristics of different groups (wearable activity tracker use or not); bivariate differences were assessed with Rao-Scott chi-square tests for categorical variables and design-based *F* tests for continuous variables. Design-based linear regression for continuous outcome variables and ordinal logistic regression for ordinal outcome variables were conducted to determine the associations after controlling for sociodemographic characteristics as well as cancer-related and general health–related clinical factors. For continuous outcome variables, the beta () coefficient was presented, while the odds ratio (OR) was provided for ordinal outcome variables. The findings were presented as weighted point estimates along with 95% CI. A significance level of .05 was used. The complex survey design and sample weights were considered in all analyses. Taylor series linearization methods were used for variance estimation. Stata/SE (version 17.0; StataCorp) was used to perform all the analyses.

Ethical Considerations

The HINTS-SEER was approved as an amendment to HINTS 5 by the Westat institutional review board (IRB; amendment ID #3212), with participating SEER registries obtaining independent IRB approvals from 3 cancer registries, namely Iowa, New Mexico, and the Greater Bay Area [16]. The HINTS 5 was approved by the Westat IRB (project #6048.14) and received a “Not Human Subjects Research” determination from the National Institutes of Health Office of Human Subjects Research (Exempt #13204) [16]. Informed consent was obtained from all the participants [15]. Due to the use of a deidentified dataset, this study was exempted from undergoing an institutional ethical review [17]. This study has been conducted using the HINTS-SEER dataset with permission from the HINTS.

Results

Characteristics of Study Population

We included 957 participants, representing 320,164 patients with cancer. The majority were 65 years old or older (69%) and identified as White (78%). Over half of them were women (55%). Most had at least a college degree (62%), an annual income of US $50,000 or more (72%), and were not employed (70%). Breast cancer was the most prevalent (24%), followed by prostate (22%), skin (11%), and colorectal cancers (7%), with the majority being at a localized stage (70%). Over half of the patients had received a cancer diagnosis 11 or more years ago (58%). The majority of patients with cancer reported their health as good or better (86%), were overweight or obese (61%), and had never smoked (61%). The average anxiety and depression score was within the normal range (mean=1.54). Patients with cancer engaged in moderate to vigorous physical activity for an average of 171 minutes, and approximately 60% did not participate in any strength training. Only 29% of participants reported wearable activity tracker use. Bivariate analyses showed wearable activity tracker use was significantly associated with younger age ($P<.001$), higher education level ($P=.04$), higher income ($P<.001$), and employed status ($P<.001$); more details in Table 1.

XSL•FO
RenderX

Table 1. Characteristics of study population (design adjusted mean or proportion).

	Total	Wearable activity device use		*P* value[a]
		No	Yes	
Sample size, N	957	__[b]	—	—
National estimates[c], %	—	70.89	29.11	—
Age (years), %				<.001
18-49	6.12	3.99	11.31	
50-64	24.89	22.43	30.89	
65-74	31.47	30.22	34.52	
75+	37.52	43.36	23.28	
Sex, %				.18
Male	45.08	46.64	41.27	
Female	54.92	53.36	58.73	
Race, %				.93
White	77.53	77.63	77.27	
Non-White	22.47	22.37	22.73	
Education level, %				.04
Less than high school	3.09	3.47	2.15	
High school	10	11.49	6.38	
Some college	25.24	27.09	20.76	
College graduate or higher	61.67	57.95	70.71	
Income (US $), %				<.001
<20,000	6.45	8.22	2.15	
20,000-35,000	10.65	12.91	5.16	
35,000-50,000	10.68	11.84	7.83	
50,000-75,000	15.19	15.61	14.18	
≥75,000	57.03	51.43	70.68	
Employment, %				<.001
Employed	30.18	24.76	43.37	
Unemployed	69.82	75.24	56.63	
Cancer site, %				.005
Breast	24.46	25.10	22.89	
Prostate	21.82	23.19	18.49	
Colorectal	7.45	8.78	4.22	
Skin	11.12	8.77	16.85	
Other	35.14	34.16	37.54	
Stage, %				.81
Localized	69.82	69.82	71	
Regional	23.46	23.46	21.39	
Distant	6.72	6.72	7.62	
Time since diagnosis (years), %				.26
<1	0.92	1.02	0.95	
2-5	16.65	14.74	16.65	

XSL•FO
RenderX

	Total	Wearable activity device use		*P* value[a]
		No	Yes	
6-10	24.63	24.98	24.63	
11+	57.78	59.36	57.78	
Smoking status, %				.03
Current	3.3	4.34	0.78	
Former	36.19	37.06	34.06	
Never	60.51	58.60	65.16	
BMI, %				.86
Underweight	1.65	1.43	2.19	
Normal	37.09	36.70	38.04	
Overweight	34.53	34.77	33.96	
Obese	26.73	27.10	25.81	
Self-rated general health, %				.002
Excellent	9.11	7.58	12.85	
Very good	35.85	32.99	42.84	
Good	40.86	42.54	36.78	
Fair	12.35	14.92	6.09	
Poor	1.82	1.98	1.45	
Anxiety and depression, mean (SE)	1.54 (0.08)	1.59 (0.10)	1.43 (0.13)	.35
Weekly minutes spent in moderate to vigorous physical activity, mean (SE)	171.08 (6.54)	157.34 (7.89)	204.54 (12.28)	.002
Weekly days spent in strength training, %				<.001
None	59.52	64.4	47.72	
1 day	7.59	6.3	10.74	
2 days	10.91	9.05	15.4	
3 days	9.63	7.68	14.36	
4 days	3.56	3	4.94	
5 days	3.68	4.12	2.62	
6 days	1.42	1.91	0.22	
7 days	3.68	3.55	4.00	

[a]*P*-value compares older adults in different subgroups.

[b]Not available.

[c]National estimates based on complex survey design.

Association Between Physical Activities and Wearable Activity Tracker Use

Table 2 illustrates the associations between 2 types of physical activities and wearable activity tracker use. The unadjusted regression analyses revealed a statistically significant association between wearable activity tracker use and higher weekly minutes spent in moderate to vigorous physical activity (=47.21, 95% CI 17.96-76.44, *P*=.002), as well as more weekly days spent in strength training (OR 1.67, 95% CI 1.24-2.24, *P*=.001). After accounting for covariates, these relationships remained statistically significant (Moderate to vigorous physical activity: =37.94, 95% CI 8.37-67.50, *P*=.01; Strength training: OR 1.49, 95% CI 1.08-2.06, *P*=.01). Among covariates, being female (= –61.30, 95% CI –98.14 to –24.45, *P*=.001), employed (=–36.33, 95% CI –67.87 to –4.80, *P*=.02), and obese (= –69.65, 95% CI –94.38 to –44.93, *P*<.001) were associated with decreased time in moderate to vigorous physical activity, while having an income of US $75,000 or higher (=72.29, 95% CI 19.28-125.31, *P*=.008) was associated with increased time in moderate to vigorous physical activity. Being White (OR 0.65, 95% CI 0.43-0.98, *P*=.04) and obese (OR 0.48, 95% CI 0.33-0.70, *P*<.001) were negatively associated with the frequency of strength training.

XSL•FO
RenderX

Table 2. Associations of wearable activity trackers used with physical activity (weighted estimates of adjusted coefficients and adjusted odds ratios [ORs]).

	Time spent in moderate to vigorous physical activity[a]			Frequency of strength training[b]		
	Adjusted OR	95% CI	*P* value	Adjusted OR	95% CI	*P* value
Wearable activity trackers use (reference: no use)	37.94	8.37-67.50	.01	1.49	1.08-2.06	.01
Age (years)						
18-49	Reference	__[c]	—	Reference	—	—
50-64	47.13	–11.62 to 105.90	.12	1.31	0.63-2.71	.45
65-74	57.70	–2.44 to 117.85	.06	1.42	0.68-2.98	.34
75+	45.65	–17.53 to 108.83	.16	0.94	0.42-2.09	.89
Sex						
Male	Reference	—	—	Reference	—	—
Female	–61.30	–98.14 to –24.45	.001	0.71	0.48-1.05	.08
Race						
Non-White	Reference	—	—	Reference	—	—
White	10.78	–18.37 to 39.93	.46	0.65	0.43-0.98	.04
Educational level						
Less than high school	Reference	—	—	Reference	—	—
High school	42.45	–28.49 to 113.40	.24	1.45	0.55-3.79	.44
Some college	10.97	–52.30 to 74.26	.73	1.12	0.48-2.58	.78
College graduate or higher	44.40	–15.24 to 104.06	.14	1.46	0.65-3.27	.34
Income (US $)						
<20,000	Reference	—	—	Reference	—	—
20,000-35,000	–44.00	–94.55 to 6.53	.08	1.06	0.46-2.47	.87
35,000-50,000	19.75	–35.24 to 74.75	.48	1.91	0.84-4.35	.12
50,000-75,000	31.83	–24.07 to 87.74	.26	1.16	0.50-2.66	.71
≥75,000	72.29	19.28 to 125.31	.01	1.46	0.71-2.98	.29
Employment						
Unemployed	Reference	—	—	Reference	—	—
Employed	–36.33	–67.87 to –4.80	.02	1.43	0.99-2.08	.05
Cancer site						
Breast	Reference	—	—	Reference	—	—
Prostate	–25.58	–77.78 to 26.61	.33	1.34	0.71-2.51	.35
Colorectal	–5.46	–60.57 to 49.65	.84	1.19	0.59-2.36	.61
Skin	–0.18	–38.05 to 37.69	.99	1.38	0.80-2.35	.23
Other	–26.73	–56.93 to 3.46	.08	1.04	0.64-1.68	.86
Cancer stage						
Localized	Reference	—	—	Reference	—	—
Regional	—	–15.93 to 44.00	.35	1.06	0.69-1.63	.77
Distant	—	–71.60 to 29.13	.41	1.03	0.52-2.03	.92
Time since diagnosis (years)						
1	Reference	—	—	Reference	—	—
2-5	84.30	–0.20 to 168.80	.05	0.99	0.33-2.93	.99

XSL•FO
RenderX

	Time spent in moderate to vigorous physical activity[a]			Frequency of strength training[b]		
	Adjusted OR	95% CI	*P* value	Adjusted OR	95% CI	*P* value
6-10	64.84	–13.98 to 143.67	.11	1.50	0.56-4.03	.41
11+	68.78	–11.51 to 149.09	.09	1.38	0.49-3.86	.52
BMI						
Underweight	—	–131.75 to 145.42	.92	1.30	0.60-2.85	.49
Normal	Reference	—	—	Reference	—	—
Overweight	—	–36.25 to 30.59	.86	0.69	0.48-1.00	.05
Obese	—	–94.38 to –44.93	<.001	0.48	0.33-0.70	<.001
Smoking status						
Current	Reference	—	—	Reference	—	—
Former	–21.83	–95.41 to 51.74	.55	0.73	0.31-1.68	.46
Never	7.53	–67.96 to 83.03	.84	0.84	0.35-1.96	.68

[a]Design-based linear regression, the residuals were close to a normal distribution.

[b]Design-based ordinal logistic regression.

[c]Not applicable.

Association Between Health Outcomes and Wearable Activity Tracker Use

The unadjusted regression analyses revealed a statistically significant association between wearable activity tracker use and better self-rated health status (OR 1.91, 95% CI 1.38-2.63, *P*<.001), and this association remained statistically significant after controlling for covariates (OR 1.58, 95% CI 1.09-2.28, *P*=.01; more details in Table 3). However, BMI and anxiety and depression were not significantly associated with wearable activity tracker use (more details in Multimedia Appendix 1).

XSL•FO
RenderX

Table 3. Associations of wearable activity trackers used with self-rated health status (weighted estimates of adjusted coefficients and adjusted odds ratios [ORs]).

	Self-rated general health[a]			Anxiety and depression[b]		
	Adjusted OR	95% CI	*P* value	Adjusted OR	95% CI	*P* value
Wearable activity device use (reference: no use)	1.58	1.09-2.28	.01	–0.16	–0.48 to 0.17	.35
Age (years)						
18-49	Reference	__[c]	—	Reference	—	—
50-64	0.97	0.46-2.03	.94	–0.34	–1.09 to 0.41	.37
65-74	0.95	0.45-1.99	.90	–0.48	–0.31 to 0.35	.26
75+	1.01	0.47-2.13	.98	–0.56	–1.48 to 0.36	.23
Sex						
Male	Reference	—	—	Reference	—	—
Female	0.89	0.57-1.41	.63	0.56	0.16-0.97	.01
Race						
Non-White	Reference	—	—	Reference	—	—
White	1.26	0.83-1.92	.26	–0.08	–0.57 to 0.42	.76
Educational level						
Less than high school	Reference	—	—	Reference	—	—
High school	1.41	0.66-3.04	.36	0.65	–0.31 to 1.61	.18
Some college	1.60	0.75-3.38	.21	0.40	–0.47 to 1.26	.37
College graduate or higher	2.54	1.20-5.36	.01	0.01	–0.82 to 0.85	.98
Income (US $)						
<20,000	Reference	—	—	Reference	—	—
20,000-35,000	0.96	0.46-2.03	.93	0.92	0.23-1.61	.01
35,000-50,000	2.58	1.27-5.23	.01	0.35	–0.50 to 1.20	.41
50,000-75,000	2.44	1.18-5.04	.01	0.04	–0.56 to 0.65	.88
≥75,000	2.23	1.14-4.33	.01	0.08	–0.47 to 0.62	.78
Employment						
Unemployed	Reference	—	—	Reference	—	—
Employed	1.39	0.96-2.01	.07	–0.09	–0.54 to 0.36	.70
Cancer site						
Breast	Reference	—	—	Reference	—	—
Prostate	1.59	0.87-2.90	.12	0.11	–0.59 to 0.81	.75
Colorectal	1.06	0.57-1.95	.84	–0.65	–1.23 to –0.07	.03
Skin	1.83	1.10-3.05	.02	0.51	–0.15 to 1.18	.13
Other	0.88	0.57-1.37	.58	–0.08	–0.56 to 0.40	.76
Cancer stage						
Localized	Reference	—	—	Reference	—	—
Regional	1.14	0.76-1.71	.50	–0.10	–0.48 to 0.28	.61
Distant	0.79	0.39-1.58	.50	0.14	–0.66 to 0.95	.73
Time since diagnosis (years)						
1	Reference	—	—	Reference	—	—
2-5	2.29	0.64-8.14	.19	–0.93	–3.05 to 1.20	.39

XSL•FO
RenderX

	Self-rated general health[a]			Anxiety and depression[b]		
	Adjusted OR	95% CI	*P* value	Adjusted OR	95% CI	*P* value
6-10	2.12	0.66-6.82	.20	–0.67	–2.74 to 1.41	.53
11+	2.07	0.61-6.98	.23	–1.05	–3.09 to 0.99	.31
BMI						
Underweight	0.31	0.12-0.80	.01	1.03	–0.13 to 2.20	.08
Normal	Reference	—	—	Reference	—	—
Overweight	0.58	0.42-0.80	.001	0.24	–0.15 to 0.62	.22
Obese	0.31	0.22-0.44	<.001	0.02	–0.44 to 0.47	.95
Smoking status						
Current	Reference	—	—	Reference	—	—
Former	1.69	0.66-4.31	.26	–0.64	–1.66 to 0.39	.22
Never	1.69	0.70-4.04	.23	–0.72	–1.65 to 0.21	.13

[a]Design-based ordinal logistic regression.

[b]Design-based linear regression, the residuals were close to a normal distribution.

[c]Not applicable.

Discussion

Principal Findings

This study used a national dataset to examine the association between using wearable activity trackers and physical activity and health outcomes among patients with cancer. The key findings were that (1) overall wearable activity tracker use was low (29%), and patients with no wearable activity tracker use tended to be older and of low socioeconomic status; (2) using wearable activity trackers was associated with increased weekly duration of moderate to vigorous physical activity, weekly frequency of strength training and improved self-rated health. Associations were not observed between using wearable activity trackers and BMI or anxiety and depression. Based on the nature of this cross-sectional study, we could not draw a causality of the benefits of using wearable activity trackers. However, the findings still advance our understanding of the use of wearable activity trackers and support using wearable activity trackers as a promising behavior modification tool to facilitate physical activity promotion.

This study showed the association between wearable activity tracker use and physical activity level (aerobic and strength) among patients with cancer. This is consistent with observational studies of general adults [12] and older adults [13] and meta-analyses of clinical trials [8,10,18]. These trackers provide real-time feedback on physical activity levels and allow users to self-monitor and set goals [10]; patients may perceive the devices as a reminder of physical activity, and they are motivated to do more physical activity when feeling they are observed [19]. A review of systematic review has reported an effect size of 0.3 for wearable activity trackers on moderate to vigorous physical activity in US adults, equating to a 6-minute-per-day increase in moderate to vigorous physical activity [8]. Interestingly, our study has found a similar result (an increase of 5.4 minutes per day) among patients with cancer. An increase of 5-10 minutes of moderate to vigorous physical activity per day is considered meaningful [9], suggesting the magnitude of moderate to vigorous physical activity increase is clinically significant. The evidence supports the use of wearable activity trackers to facilitate physical activity promotion.

Our study found an association between wearable activity trackers and self-rated health, which is consistent with another study using the HINTS dataset [11]. The association of wearable activity trackers and self-rated health is mediated through perceived health competence and physical activity; an approach to foster the perceived health competence includes setting challenges and meaningful feedback by wearable activity trackers [11]. We did not detect the association between wearable activity trackers and other physiological and psychosocial outcomes. One possible explanation is that these outcomes are probably downstream of physical activity behavior change and might require a long duration of wearable activity trackers use to manifest.

We found the prevalence of wearable activity tracker use among patients with cancer was low (29%), which is similar to the estimates of adults in the United States [12], although higher than the older adults [13]. Our study also found that patients with no wearable activity tracker use tended to be older, unemployed, and with lower income, highlighting the digital divide and health inequity. As the study findings support the positive association between wearable activity tracker use and improved physical activity and health status, this study's findings can inform researchers on how to mitigate the digital divide and promote digital health literacy [20] to facilitate wearable activity tracker use in patients with cancer. Wide reach to patients of older age and low socioeconomic status is key to guaranteeing that physical activity promotion using wearable activity trackers generates large effects on a broad population without worsening health inequities.

XSL•FO
RenderX

Limitations

While this study used a population-based dataset acquired from 3 prominent cancer registries in the United States, the geographical aspects involved could potentially affect the generalizability of the study findings. Also, the cross-sectional analysis prevents us from drawing solid conclusions about the direction of the relationships between wearable activity trackers and physical activity and health outcomes. Future longitudinal research addressing this research gap would be beneficial. In addition, the self-reported measures of physical activity are not as accurate as objectively measured physical activity, and the details about the wearable activity tracker use (eg, wear time, type of wearable activity trackers, etc) were not available in the HINTS-SEER dataset, which may moderate the relationship between wearable activity tracker use and physical activity. We only included several health outcome indicators and other important health outcomes (eg, quality of life, fatigue) were not available for analysis.

Conclusion

This research investigated the associations between the use of wearable activity trackers, physical activity, and health outcomes in patients with cancer using a population-based dataset. The results of the study indicate that the adoption of wearable activity trackers among patients with cancer remains relatively low. However, the findings also provide evidence that using wearable activity trackers is associated with increased physical activity and improved overall health, thus suggesting their potential as effective tools for promoting physical activity. Nonetheless, the study revealed a lower usage rate of wearable activity trackers among older individuals and those with a lower socioeconomic status, underscoring the presence of a digital divide within the population of patients with cancer. Further research is necessary to explore strategies for increasing the use of wearable activity trackers in cancer survivors who are older and have lower socioeconomic status.

Acknowledgments

YC was supported by UT Health Houston-Cancer Prevention and Research Institute of Texas Innovation for Cancer Prevention Research Postdoctoral Fellowship (grant #RP210042). The funder played no role in the study design, data collection, analysis, and interpretation of data or the writing of this manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Cancer Prevention and Research Institute of Texas. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2023-00271837).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Weighted estimates of adjusted coefficients in linear regression models of BMI and anxiety/depression.
[DOCX File , 24 KB - jmir_v26i1e51291_app1.docx]

References

1. Fuller JT, Hartland MC, Maloney LT, Davison K. Therapeutic effects of aerobic and resistance exercises for cancer survivors: a systematic review of meta-analyses of clinical trials. Br J Sports Med 2018;52(20):1311-1311. [doi: 10.1136/bjsports-2017-098285] [Medline: 29549149]
2. Friedenreich CM, Stone CR, Cheung WY, Hayes SC. Physical activity and mortality in cancer survivors: a systematic review and meta-analysis. JNCI Cancer Spectr 2020;4(1):pkz080 [FREE Full text] [doi: 10.1093/jncics/pkz080] [Medline: 32337494]
3. McTiernan A, Friedenreich CM, Katzmarzyk PT, Powell KE, Macko R, Buchner D, 2018 PHYSICAL ACTIVITY GUIDELINES ADVISORY COMMITTEE*. Physical activity in cancer prevention and survival: a systematic review. Med Sci Sports Exerc 2019;51(6):1252-1261 [FREE Full text] [doi: 10.1249/MSS.0000000000001937] [Medline: 31095082]
4. Rock CL, Doyle C, Demark-Wahnefried W, Meyerhardt J, Courneya KS, Schwartz AL, et al. Nutrition and physical activity guidelines for cancer survivors. CA Cancer J Clin 2012;62(4):243-274 [FREE Full text] [doi: 10.3322/caac.21142] [Medline: 22539238]
5. Blanchard CM, Courneya KS, Stein K, American CSS. Cancer survivors' adherence to lifestyle behavior recommendations and associations with health-related quality of life: results from the American cancer society's SCS-II. J Clin Oncol 2008;26(13):2198-2204 [FREE Full text] [doi: 10.1200/JCO.2007.14.6217] [Medline: 18445845]
6. Thraen-Borowski KM, Gennuso KP, Cadmus-Bertram L. Accelerometer-derived physical activity and sedentary time by cancer type in the United States. PLoS One 2017;12(8):e0182554 [FREE Full text] [doi: 10.1371/journal.pone.0182554] [Medline: 28806753]
7. Schmitz KH. Exercise Oncology: Prescribing Physical Activity Before and After a Cancer Diagnosis. New York City: Springer Nature; 2020.

XSL•FO
RenderX

8. Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022;4(8):e615-e626 [FREE Full text] [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]
9. Kraus WE, Powell KE, Haskell WL, Janz KF, Campbell WW, Jakicic JM, 2018 PHYSICAL ACTIVITY GUIDELINES ADVISORY COMMITTEE*. Physical activity, all-cause and cardiovascular mortality, and cardiovascular disease. Med Sci Sports Exerc 2019;51(6):1270-1281 [FREE Full text] [doi: 10.1249/MSS.0000000000001939] [Medline: 31095084]
10. Singh B, Zopf EM, Howden EJ. Effect and feasibility of wearable physical activity trackers and pedometers for increasing physical activity and improving health outcomes in cancer survivors: a systematic review and meta-analysis. J Sport Health Sci 2022;11(2):184-193 [FREE Full text] [doi: 10.1016/j.jshs.2021.07.008] [Medline: 34314878]
11. De La Torre S, Spruijt-Metz D, Farias AJ. Survivors' health competence mediates the association between wearable activity tracker use and self-rated health: HINTS analysis. J Cancer Surviv 2022;16(6):1268-1278. [doi: 10.1007/s11764-021-01112-9] [Medline: 35001258]
12. Xie Z, Jo A, Hong YR. Electronic wearable device and physical activity among US adults: an analysis of 2019 HINTS data. Int J Med Inform 2020;144:104297. [doi: 10.1016/j.ijmedinf.2020.104297] [Medline: 33059242]
13. Jiang Y, Zeng K, Yang R. Wearable device use in older adults associated with physical activity guideline recommendations: empirical research quantitative. J Clin Nurs 2023;32(17-18):6374-6383. [doi: 10.1111/jocn.16631] [Medline: 36740763]
14. Overview of the HINTS 5 cycle 4 (2020) survey and data analysis recommendations. National Cancer Institute. 2022. URL: https://hints.cancer.gov/data/download-data.aspx [accessed 2023-07-26]
15. Health information national trends survey 5 (HINTS 5) HINTS-SEER methodology report. Webstat. URL: https://hints.cancer.gov/data/methodology-reports.aspx [accessed 2023-07-26]
16. Institutional review board (IRB) approvals for the health information national trends survey (HINTS). Survey NCiHINT. URL: https://hints.cancer.gov/about-hints/institutional-review-board.aspx [accessed 2024-09-04]
17. Notice on issuing the measures for the ethical review of life sciences and medical research involving human subjects. China TSCTPsRo. URL: https://www.gov.cn/zhengce/zhengceku/2023-02/28/content_5743658.htm [accessed 2024-09-04]
18. Laranjo L, Ding D, Heleno B, Kocaballi B, Quiroz JC, Tong HL, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
19. Zhou W, Reddy RM, Lavoie Smith EM, Veliz PT, Chen W, Webster K, et al. Objectively measured physical activity in patients with lung cancer after surgey: feasibility and acceptability of a longitudinal approach. J Geriatr Oncol 2023;14(3):101467. [doi: 10.1016/j.jgo.2023.101467] [Medline: 36907109]
20. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]

Abbreviations

HINTS: Health Information National Trends Survey
HINTS-SEER: Health Information National Trends Survey–Surveillance, Epidemiology, and End Results
IRB: institutional review board
OR: odds ratio

Edited by A Schwartz; submitted 27.07.23; peer-reviewed by Z Xie, A Jo; comments to author 19.09.23; revised version received 08.10.23; accepted 09.09.24; published 22.10.24.

Please cite as:
Zhou W, Shang S, Cho Y
Associations of Wearable Activity Tracker Use With Physical Activity and Health Outcomes in Patients With Cancer: Findings from a Population-Based Survey Study
J Med Internet Res 2024;26:e51291
URL: https://www.jmir.org/2024/1/e51291
doi:10.2196/51291
PMID:39436693

©Weijiao Zhou, Shaomei Shang, Youmin Cho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Phenotypes of Mobile Keyboard Backspace Rates and Their Associations With Symptoms of Mood Disorder: Algorithm Development and Validation

Qimin Liu[1,2], PhD; Emma Ning[3]; Mindy K Ross[4]; Andrea Cladek[2]; Sarah Kabir[2]; Amruta Barve[2]; Ellyn Kennelly[5]; Faraz Hussain[2]; Jennifer Duffecy[2]; Scott A Langenecker[6]; Theresa M Nguyen[4]; Theja Tulabandhula[7]; John Zulueta[2]; Alexander P Demos[3]; Alex Leow[2,4]; Olusola Ajilore[2]

[1]Department of Psychological and Brain Sciences, Boston University, Boston, MA, United States
[2]Department of Psychiatry, University of Illinois Chicago, Chicago, IL, United States
[3]Department of Psychology, University of Illinois Chicago, Chicago, IL, United States
[4]Department of Biomedical Engineering and Computer Science, University of Illinois Chicago, Chicago, IL, United States
[5]Department of Psychology, Wayne State University, Detroit, MI, United States
[6]Department of Psychiatry and Behavioral Health, Ohio State University, College of Medicine, Columbus, OH, United States
[7]Department of Information and Decision Sciences, University of Illinois Chicago, Chicago, IL, United States

Corresponding Author:
Qimin Liu, PhD
Department of Psychological and Brain Sciences
Boston University
900 Commonwealth Ave
Boston, MA, 02215
United States
Phone: 1 6173532000
Email: qiminliu@bu.edu

Abstract

Background: Passive sensing through smartphone keyboard data can be used to identify and monitor symptoms of mood disorders with low participant burden. Behavioral phenotyping based on mobile keystroke data can aid in clinical decision-making and provide insights into the individual symptoms of mood disorders.

Objective: This study aims to derive digital phenotypes based on smartphone keyboard backspace use among 128 community adults across 2948 observations using a Bayesian mixture model.

Methods: Eligible study participants completed a virtual screening visit where all eligible participants were instructed to download the custom-built BiAffect smartphone keyboard (University of Illinois). The BiAffect keyboard unobtrusively captures keystroke dynamics. All eligible and consenting participants were instructed to use this keyboard exclusively for up to 4 weeks of the study in real life, and participants' compliance was checked at the 2 follow-up visits at week 2 and week 4. As part of the research protocol, every study participant underwent evaluations by a study psychiatrist during each visit.

Results: We found that derived phenotypes were associated with not only the diagnoses and severity of depression and mania but also specific individual symptoms. Using a linear mixed-effects model with random intercepts accounting for the nested data structure from daily data, the backspace rates on the continuous scale did not differ between participants in the healthy control and in the mood disorders groups (P=.11). The 3-class model had mean backspace rates of 0.112, 0.180, and 0.268, respectively, with a SD of 0.048. In total, 3 classes, respectively, were estimated to comprise 37.5% (n=47), 54.4% (n=72), and 8.1% (n=9) of the sample. We grouped individuals into Low, Medium, and High backspace rate groups. Individuals with unipolar mood disorder were predominantly in the Medium group (n=54), with some in the Low group (n=27) and a few in the High group (n=6). The Medium group, compared with the Low group, had significantly higher ratings of depression (b=2.32, P=.008). The High group was not associated with ratings of depression with (P=.88) or without (P=.27) adjustment for medication and diagnoses. The High group, compared with the Low group, was associated with both nonzero ratings (b=1.91, P=.02) and higher ratings of mania (b=1.46, P<.001). The High group, compared with the Low group, showed significantly higher odds of elevated mood (P=.03), motor activity (P=.04), and irritability (P<.05).

XSL•FO
RenderX

Conclusions: This study demonstrates the promise of mobile typing kinematics in mood disorder research and practice. Monitoring a single mobile typing kinematic feature, that is, backspace rates, through passive sensing imposes a low burden on the participants. Based on real-life keystroke data, our derived digital phenotypes from this single feature can be useful for researchers and practitioners to distinguish between individuals with and those without mood disorder symptoms.

(J Med Internet Res 2024;26:e51269) doi:10.2196/51269

KEYWORDS

keyboard typing; passive sensing; digital phenotyping; mood disorder; mixture model; phenotypes; mobile keyboard; smartphone; keyboard data; monitoring; clinical decision-making; depression; mania, mobile phone

Introduction

Passive sensing has increasingly become an important tool for identifying and monitoring mood disorder symptoms, given its low participant burden and potential to correlate with subjective ratings [1]. Particularly, behavioral phenotyping based on passive sensing data can help efficient clinical decision-making [2]. Much of passive sensing phenotyping to date has aimed to correlate with mood disorder diagnoses or severity, overlooking that mood disorders may not be unitary [3-5]. This study aims to derive a digital phenotype [6] (ie, collection of an individual's data representing their interactions with their environments as obtained through digital devices) based on backspace rates gathered through mobile keyboard passive sensors. This approach leverages the nuanced insights that can be gleaned from smartphone usage patterns to enhance our understanding of health, behavior, and well-being in real-time and in naturalistic settings. We show that derived phenotypes differentially correlate with not only the diagnoses and severity of depression and mania but also the individual symptoms.

Keyboard typing data have shown promise in differentiating and monitoring mood disorder symptoms [7-11]. Using unbiased and objective information for monitoring mood disorder symptoms is important: many of biological and neural markers still lack dependable diagnostic validity or reliability [12-15]; using self-reports to assess symptoms can create a burden for the patient and may cause bias from recall and the negative mood state [16,17]. Since 85% of Americans own a smartphone, using metadata from mobile keyboards can create unobtrusive and equitable access to mood disorder symptom monitoring [18]. Mobile keystroke kinematics can potentially reflect mental health, including psychomotor, cognitive, and social functions [11]. Previous research has shown that more severe depression is related to more variable typing speed, shorter session duration, and lower accuracy [10]. Overall, the number of typing sessions was positively correlated with depression, potentially reflecting loneliness and social withdrawal [11]. Typing speed, session length, as well as autocorrect and backspace, have been linked to brain aging [19].

Past studies have largely focused on data-driven approaches to identify keystroke features relevant to syndrome-level mood disorder diagnoses. However, cognitive theories of depression have long posited the role of negative cognitive styles, such as rumination, in the development and maintenance of depressive symptoms [20-22]. Specifically, rumination has been linked to mixed presentations of anxiety among depressed individuals [23-25]. Relatedly, rumination has also been found to be important among individuals with bipolar disorders, where additional rumination has been reported to associate with positive affect [26-28]. One potential keystroke feature that may reflect rumination is the backspace rate (ie, the percentage of backspace use in a total number of keypresses). The backspace rate can be particularly insightful because the backspace use may capture not just the frequency of typing errors and subsequent corrections but also the underlying cognitive pattern—potentially indicative of a ruminative typing style. This dual functionality elevates the backspace rate from a simple error metric to a potential marker of cognitive processes. As such, focusing on the correlation between backspace rates and mood disorder symptoms [11,29] may offer a novel pathway for digital phenotyping that is theory-driven.

Backspace rates, as a proxy of typing corrections, may reflect aspects of cognitive impairments by capturing the frequency and patterns of errors and corrections during typing. Specifically, executive dysfunction (which encompasses cognitive processes such as planning, problem-solving, and error correction) is among the most common features of unipolar and bipolar mood disorders [30,31]. Executive dysfunctions in mood disorders are linked to symptom severity, functional outcomes, and treatment responses [30,32-36]. Executive functions (which encompass cognitive processes such as planning, problem-solving, and error correction) have been demonstrated to be differentially impaired in unipolar and bipolar mood disorders [37]. The variability in backspace rates could be indicative of the degree of executive dysfunction. While patients with unipolar and bipolar disorder both exhibit deficits in executive functions, bipolar mood disorders appear to be associated with a wider range of cognitive impairment and higher executive dysfunctions than unipolar mood disorder [38]. As such, it is plausible that mobile keyboard backspace rates could serve as a nuanced marker for distinguishing between these conditions. This specificity in measurement could potentially offer insights into the underlying cognitive disruptions unique to each disorder, thereby providing a more targeted approach to identifying and understanding the cognitive dimensions of mood disorders.

Deriving digital phenotypes can facilitate efficient clinical decision-making. Whereas passive sensing data are continuous (eg, backspace rates can be any value between 0 and 1), clinical decisions are often discrete (eg, treat or not to treat with medication or psychotherapy). Digital phenotypes are ecologically valid, data-driven, and require low patient burden [1,39]. For clinicians, the ecological nature of digital data can help them better understand the underlying mechanisms and

XSL•FO
RenderX

manifestations of a patient's mood disorder symptoms in daily life. For researchers, digital phenotyping can also facilitate the development of new interventions and technologies that can be tailored to individual patients as informed by unobtrusively monitored digital behaviors. For administrators, clinically informative phenotypes may also guide equitable service allocation in limited-resourced mental health settings. Overall, digital phenotyping has the potential to transform the field of mood disorder research and clinical practice by providing a more accurate, efficient, and patient-centered approach to diagnosis and treatment.

In the context of digital phenotyping research, many studies have sought to derive phenotypes that correspond to psychiatric diagnoses. However, mood disorder diagnoses reflect consensus definitions of syndromes instead of unitary disease entities [5,40], creating challenges for the use and efficacy of such efforts. Another common approach is to correlate digital phenotypes with aggregated severity scores based on symptom counts or sum scores. This helps quantify how digital phenotypes may correspond to specific levels of severity but still overlooks the heterogeneity in the expression of mood disorders [3,4]. As mood disorder symptoms can occur in different combinations, this study aims to go beyond diagnoses and severity aggregates by additionally investigating how individual symptoms are related to derived digital phenotypes. By doing so, the study aims to provide a more nuanced understanding of derived digital phenotypes in relation to mood disorders.

This study uses backspace rates to derive digital phenotypes that correlate with mood disorder symptoms. We apply a Bayesian mixture modeling approach to identify digital phenotypes for each individual. We then use mood disorder diagnoses as a proxy to aid in model selection to ensure that the derived phenotypes are externally valid. We quantify how individuals in different derived phenotypes may differ in the severity of depression and mania symptoms. Furthermore, we identify how derived phenotypes correlate with individual symptoms of depression and mania.

Methods

Participants

The study participants consisted of healthy controls (n=27) and individuals with mood disorders (n=101) between the ages of 25 and 50 years. Participants with active suicidal ideation, severe cognitive impairment, active alcohol or substance use disorders, major medical or neurological illness interfering with protocol adherence, contraindications to magnetic resonance imaging, who were pregnant, or who did not own a smartphone were ineligible for the study. Participants were recruited using community outreach efforts such as listserv emails, social media posts, and clinical trial registries, and participant eligibility was determined after the completion of an initial screening session. Individuals in the mood disorder group were required to meet the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) criteria for major depressive disorder (n=83), bipolar disorder type I/II (n=14), persistent depressive disorder (n=3), or cyclothymia (n=1). Our broad diagnostic inclusion criteria aimed to ensure sufficient representation across different diagnostic categories, given the imprecision of current symptom clusters, as well as to account for the widespread occurrence of cognitive dysfunction across diagnostic categories..

Table 1 presents the descriptive statistics of the sample characteristics of our study participants.

XSL•FO
RenderX

Table 1. Descriptive statistics for the study participants.

	Healthy control (N=27)	Mood disorder (N=101)	*P* value
Age			.35
Mean (SD)	31.6 (7.93)	33.1 (6.58)	—
Median (Min, Max)	28.3 (24.8, 50.6)	31.1 (25.0, 50.2)	—
Gender, n (%)			.67
Female	13 (48.1)	56 (55.4)	—
Male	14 (51.9)	44 (43.6)	—
Other	0 (0)	1 (1)	—
Medications, n (%)			.001
No	21 (77.8)	41 (40.6)	—
Yes	6 (22.2)	60 (59.4)	—
Hamilton Depression Rating Scale			<.001
Mean (SD)	0.556 (1.09)	7.66 (5.54)	—
Median (Min, Max)	0 (0, 4.00)	7.00 (0, 20.0)	—
A quick inventory of depressive symptomatology			<.001
Mean (SD)	0.469 (0.73)	6.08 (4.15)	—
Median (Min, Max)	0 (0, 2.00)	5.67 (0, 18.0)	—
Young Mania Rating Scale			<.001
Mean (SD)	0.074 (0.39)	1.10 (2.09)	—
Median (Min, Max)	0 (0, 2.00)	0 (0, 14.0)	—

Procedures

Eligible study participants completed a remote screening visit where all eligible participants were instructed to download the custom-built BiAffect smartphone keyboard (University of Illinois) that replaced the standard iOS or Android keyboard on their smartphones. The custom BiAffect keyboard appears and functions much like the native keyboard with no notable delay or discrepancies for users. The BiAffect keyboard unobtrusively captures keystroke dynamics (ie, typing metadata such as keystroke entry date and time), but never the typing content. The data collection is unobtrusive as user interactions remain natural, uninterrupted, and no different from other real-life settings. Participants completed an additional e-consent that described the goal of capturing keystroke dynamics and the extent of metadata the custom keyboard collected before enabling the keyboard.

All eligible and consenting participants were instructed to use this keyboard exclusively for up to 4 weeks of the study in real life, and participants' adherence was checked at the 2 follow-up visits at week 2 and week 4. A keyboard session begins when the user presses the first key and terminates when the keyboard is no longer rendered or after 6 seconds of keyboard inactivity. Keyboard sessions are associated with unique identifiers with corresponding time stamps recorded in coordinated universal time (as well as the specific time zone the user is in). The keyboard records the timing of keypress events and their general category (eg, alphanumeric, backspace, autocorrection, punctuation, auto-suggestion, and special characters).

As part of the research protocol, every study participant underwent evaluations by a study psychiatrist during each visit. During the initial visit, participants were categorized into either the healthy control (HC) or mood disorder (MD) groups. In addition, their depression levels were assessed using measures including the Hamilton Depression Rating Scale (HAM-D), while their mania levels were assessed using the Young Mania Rating Scale (YMRS). Follow-up visits repeated neuropsychiatric assessments for participant engagement and monitoring, as well as allowing for within-subject replications of study findings on other study outcomes.

Measures

Depression

HAM-D [41] was used to measure symptoms of depression. HAM-D is a gold standard clinician-administered measure with 17 items, each rated on a 3- or 5-point Likert scale. HAM-D assesses depressed mood, feelings of guilt, suicide, initial or middle or terminal insomnia, work and activities, psychomotor retardation or agitation, anxiety psychic, anxiety somatic, gastrointestinal somatic symptoms, general somatic symptoms, genital symptoms, hypochondriasis, loss of weight, and insight.

Mania

The YMRS was used to measure symptoms of mania [42]. YMRS is a clinician-administered measure that consists of 11 items, each rated on a 5-point Likert scale. YMRS assesses elevated mood, increased motor activity or energy, sexual interest, sleep, irritability, speech, rate and amount, language,

XSL•FO
RenderX

thought disorder, content, disruptive or aggressive behavior, appearance, and insight.

Backspace Usage

Backspace usage is defined as the total number of backspaces used per day divided by the total number of keypresses per day. We used the longitudinal data on daily backspace use for individual participants.

Usability

We measured the usability through 4 questions on an ordinal scale from 1 (strongly disagree) to 7 (strongly agree): "I learn to use it quickly," "I easily remember how to use it," "It is easy to learn to use it," and "I quickly became skillful with it."

Data Analysis

We implemented a bounded Gaussian mixture model using Bayesian estimation through JAGS [43]. We used Bayesian estimation due to its advantages in handling limited sample size, enhancing the robustness and reliability of our model estimations. Since daily backspace rates for each person are bounded within (0,1), we included the bounded support correspondingly. To infer the model parameters, we used noninformative priors to avoid biasing the results and conjugate priors for computational efficiency. To ensure the adequacy of the posterior distributions, we used a burn-in process consisting of 5000 iterations. The burn-in process is a conventional Bayesian estimation step that is critical for stabilizing the Markov Chain Monte Carlo (MCMC) simulations. This was then followed by a sampling phase with 2 chains, each comprised of 10,000 iterations. We estimated the model with up to 4 classes, considering model parsimony and interpretability within our limited sample size. The final model was selected based on both how the model internally fitted the backspace rates data and how the model was externally relevant to mood disorder diagnoses. The internal fit was determined per deviance information criterion (DIC), a statistical measure used to compare the fit of different models, considering both the goodness of fit and the model complexity, where lower values indicate a better fit [44], as well as considerations of class proportions (classes with less than 5 individuals were considered too small). External fit was determined per chi-square tests to mood disorder diagnoses (unipolar mood disorder, bipolar disorder vs. healthy control).

To ascertain the use of identified digital phenotypes, we conducted regression to determine the differences between class membership and baseline ratings of depression. Since baseline ratings of mania showed zero inflation and after-log-transformation asymmetry, a zero-inflated gamma regression was used to examine between phenotype differences. We additionally examined whether derived phenotypes show differences in baseline depression and mania after adjusting for mood disorder diagnoses and medication (taking vs not taking medication). Furthermore, we examined how the identified phenotype relates to individual symptoms of depression and mania at the initial visit using ordinal regression with individual symptoms as the dependent variable and the identified phenotype as the independent variable.

Ethical Considerations

Institutional review board approval was obtained at the University of Illinois Chicago (protocol number: 2019-1333). We obtained informed consent from all participants before the screening protocols. Participants were informed of their right to withdraw at any time, and confidentiality was rigorously maintained with data anonymization and secure storage protocols. We also have provisions for handling sensitive information and incidental findings responsibly, ensuring that participants are supported throughout the study and in the aftermath of any discoveries made during research activities. To minimize potential risk from study procedures, (1) participants were carefully screened to exclude those who were physically, neurologically, or psychiatrically at risk or had a prior or current history of a major medical, neurological, or psychiatric disorder; (2) studies were conducted in areas located in close proximity to a hospital, where emergency assistance can be obtained; and (3) a complete medical history, review of medications and physical symptoms or signs were performed before entry into this study. Furthermore, the evaluation of mood disorders included suicide risk screening by a board-eligible or board-certified psychiatrist. Participants with active suicidal ideation were excluded from the study. If a participant exhibited active suicidal ideation, they were connected to a doctoral-level study staff for emergent psychiatric evaluation. If needed, patients with elevated suicide risk were offered inpatient hospitalization or an immediate clinical evaluation in either the emergency room or an outpatient clinic. Participants were compensated through cheques or gift cards 4-8 weeks after the end of their participation, with values commensurate to their involvement.

Results

Figure 1 plots the histogram of daily backspace rates across participants across days. The average daily backspace rate is 0.161 (SD 0.062). For participants in the mood disorder group, the average daily backspace rate is 0.165 (SD 0.066). For participants in the healthy control group, the average backspace rate is 0.145 (SD 0.051). Using a linear mixed-effects model with random intercepts accounting for the nested data structure from daily data, the backspace rates on the continuous scale did not differ between participants in the HC and in the MD groups (P=.11).

In terms of usability, the average across all questions was 5.28, indicating general ease of use. The average ratings for individual questions were as follows: (1) 5.47 for "I learn to use it quickly," (2) 5.57 for "I easily remember how to use it," (3) 5.24 for "It is easy to learn to use it," and (4) 4.84 for "I quickly became skillful with it." These detailed responses reflect a high level of user acceptance and positive interaction with the app.

XSL•FO
RenderX

Figure 1. Histogram of daily backspace use rates across participants across days.

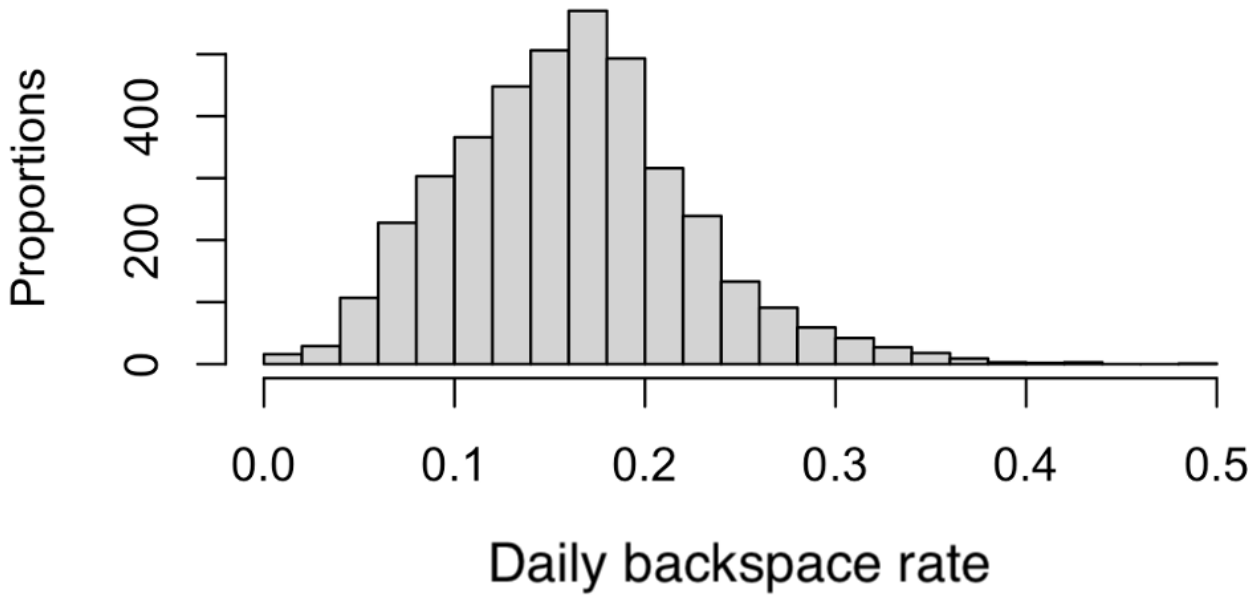

Phenotype Identification

Models with 2 to 4 classes, respectively, had DICs of –9061.370, –9669.93, and –9772.263. The 4-class model demonstrated the best internal fit per DIC but had 1 class with only 2 individuals. While considering the relation between estimated classes and mood disorder diagnoses, we found that both the 3-class ($\chi^2(4)=10.95$, $P=.03$) and the 4-class ($\chi^2(6)=13.48$, $P=.04$) models showed significant associations to mood disorder diagnoses. Taken together, the 3-class model was selected.

The 3-class model had mean backspace rates respectively at 0.112, 0.180, and 0.268 with an SD of 0.048. In addition, 3 classes respectively were estimated to comprise 37.5% (47/128), 54.4% (72/128), and 8.1% (9/128) of the sample. We term the groups respectively Low, Medium, and High backspace rates groups. Individuals from healthy control groups were either in the Low (n=13) or the Medium (n=14) backspace rates group. Individuals with bipolar disorders were distributed across all groups: 7 in the Low group, 4 in the Medium group, and 3 in the High group. Individuals with unipolar mood disorder were predominantly in the Medium group (n=54), with some in the Low group (n=27) and a few in the High group (n=6).

Between-Phenotype Differences in Baseline Severity of Depression and Mania

The Medium group, compared to the Low group, had significantly higher ratings of depression (b=2.32, $P=.008$). This is significant after accounting for medication and diagnoses (b=2.18, $P=.02$). The High group was not associated with ratings of depression with ($P=.88$) or without ($P=.27$) adjusting for medication and diagnoses.

The High group, compared to the Low group, was associated with both nonzero ratings (b=1.91, $P=.02$) and higher ratings of mania (b=1.46, $P<.001$). The Medium group, compared with the Low group, was only associated with higher ratings of mania (b=.79, $P=.01$). After controlling for medication and diagnoses, derived phenotypes were not significantly associated with nonzero ratings, but both High (b=.887, $P=.01$) and Medium (b=1.45, $P=.001$) groups, compared with the Low group remained significantly associated with higher ratings of mania.

Between-Phenotype Differences in Specific Symptoms of Depression

Figure 2 displays the average ratings by identified phenotypes in each item of depression. Models for guilt, suicide, hypochondriasis, loss of weight, and insight did not converge. Closer examination revealed that all individuals in the High group endorsed 0 for guilt, suicide, and loss of weight. Analyses excluding individuals in the High group did not reveal significant differences between the Low and Medium groups. For hypochondriasis, all individuals in the Low group endorsed 0. Analyses excluding individuals in the Low group did not reveal significant differences between the Medium and the High group. Only the Medium group endorsed nonzero values for insight. Results for all other items are presented in Table 2.

The Medium group, compared with the Low group, showed significantly higher odds of higher psychic anxiety ($P=.001$), somatic anxiety ($P=.002$), general somatic symptoms ($P=.03$), and genital symptoms ($P=.01$). The High group, compared with the Low group, showed significantly higher odds of initial ($P=.03$), middle ($P=.009$), terminal insomnia ($P=.003$), agitation ($P=.02$), and psychic anxiety ($P=.02$).

XSL•FO
RenderX

Figure 2. Average ratings of depression symptoms by backspace rates phenotype.

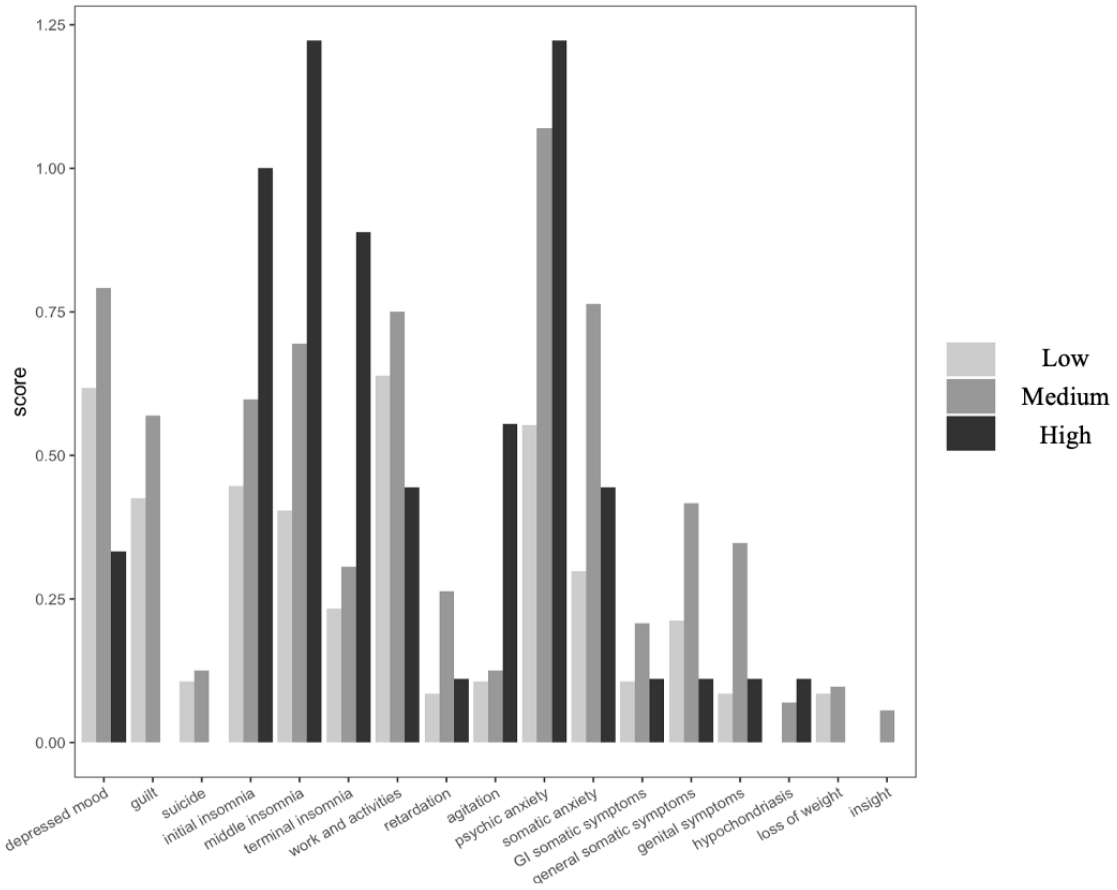

XSL•FO
RenderX

Table 2. Relations between specific depression symptoms and backspace rates phenotypes.

Variable and phenotype	OR (95% CI)	*z* score	*P* value
Depressed mood			
Medium	1.58 (0.761-3.262)	1.22	.22
High	0.49 (0.093-2.608)	–0.83	.41
Initial insomnia			
Medium	1.69 (0.776-3.703)	1.32	.19
High[a]	4.20 (1.181-14.939)	2.22	.03
Middle insomnia			
Medium	2.12 (0.977-4.584)	1.90	.06
High[a]	6.82 (1.606-28.97)	2.60	.009
Terminal insomnia			
Medium	1.26 (0.505-3.127)	0.49	.62
High[a]	7.61 (2.004-28.875)	2.98	.003
Work and activities			
Medium	1.51 (0.701-3.23)	1.05	.29
High	0.65 (0.121-3.499)	–0.50	.61
Retardation			
Medium	3.14 (0.981-10.073)	1.93	.05
High	1.34 (0.132-13.534)	0.25	.81
Agitation			
Medium	0.93 (0.279-3.135)	–0.11	.91
High[a]	6.87 (1.412-33.462)	2.39	.02
Psychic anxiety			
Medium[a]	3.30 (1.608-6.787)	3.25	.001
High[a]	4.62 (1.256-17.019)	2.30	.02
Somatic anxiety			
Medium[a]	3.64 (1.627-8.165)	3.14	.002
High	2.08 (0.517-8.360)	1.03	.30
Gastrointestinal somatic symptoms			
Medium	2.05 (0.685-6.124)	1.28	.20
High	1.05 (0.108-10.21)	0.04	.97
General somatic symptoms			
Medium[a]	2.78 (1.138-6.817)	2.24	.03
High	0.58 (0.064-5.336)	–0.48	.63
Genital symptoms			
Medium[a]	5.24 (1.456-18.851)	2.54	.01
High	1.76 (0.163-19.055)	0.47	.64

[a]Significant relations.

XSL•FO
RenderX

Between-Phenotype Differences in Specific Symptoms of Mania

Figure 3 displays the average ratings by identified phenotypes in each item of mania. Only models for elevated mood, motor activity, sleep, and irritability converged. Closer examination revealed that all individuals endorsed 0 for disruptive or aggressive behaviors, appearance, and insight. For thought content, only individuals from the Medium group endorsed a nonzero rating. For language or thought, only individuals from the High group had a nonzero rating. For speech and sexual interest, only individuals from Medium and High groups endorsed nonzero ratings. Analyses excluding the Low group revealed that sexual interest and speech did not differ between the Medium and High groups. Results for all other items are presented in Table 3.

The High group, compared with the Low group, showed significantly higher odds of elevated mood (*P*=.03), motor activity (*P*=.04), and irritability (*P*=.05).

Figure 3. Average ratings of mania symptoms by backspace rates phenotypes.

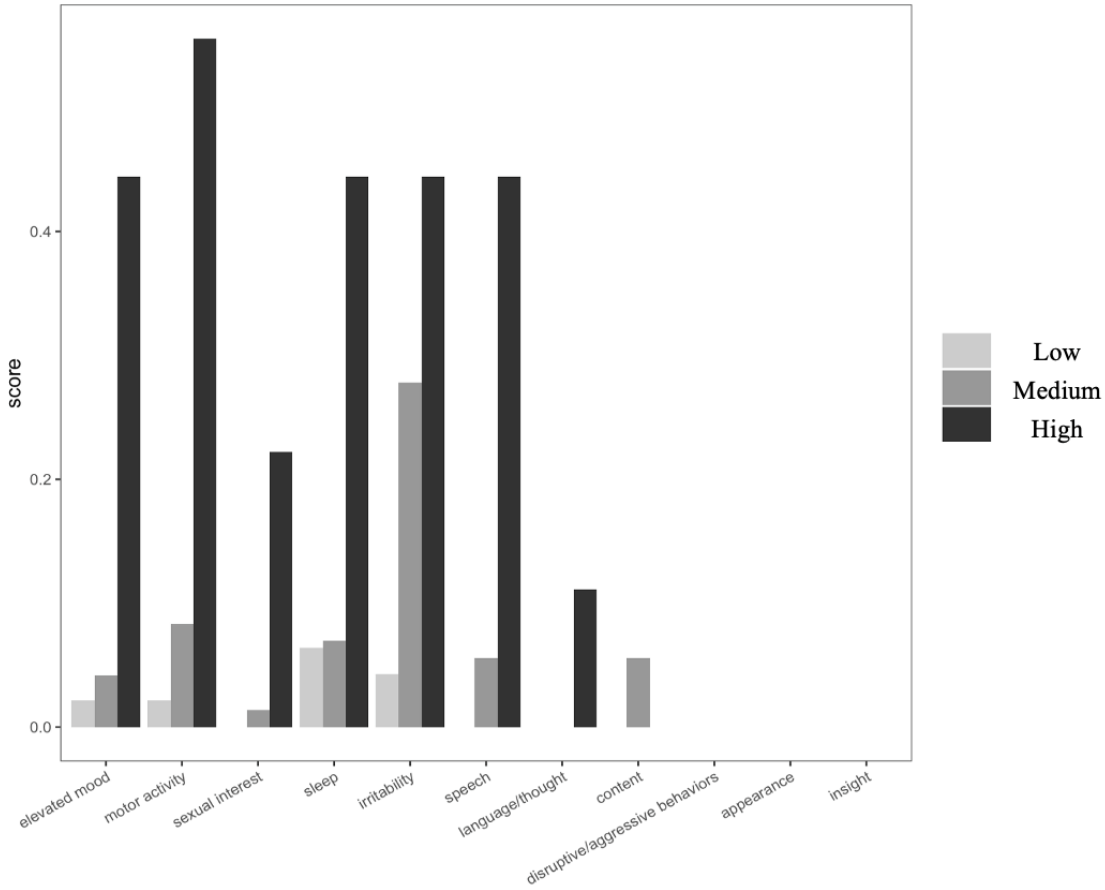

XSL•FO
RenderX

Table 3. Relations between individual mania symptoms and backspace rates phenotypes.

Variable and phenotype	OR (95% CI)	*z* score	*P* value
Elevated mood			
Medium	1.99 (0.2-19.679)	0.59	.56
High[a]	16.12 (1.265-205.427)	2.14	.03
Motor activity			
Medium	2.71 (0.294-25.012)	0.88	.38
High[a]	15.22 (1.206-192.227)	2.1	.04
Sleep			
Medium	1.09 (0.249-4.807)	0.12	.91
High	4.74 (0.66-34.027)	1.55	.12
Irritability			
Medium	6.61 (0.809-54.004)	1.76	.08
High[a]	12.83 (1.028-160.073)	1.98	.05

[a]Significant relations.

Discussion

This study demonstrates the promises of mobile typing kinematics in mood disorder research and practice. Monitoring single mobile typing kinematic feature—backspace rates—through passive sensing imposes a low burden on participants. Based on real-life keystroke data, our derived digital phenotypes from this single feature can be useful for researchers and practitioners to distinguish between individuals with and without mood disorder symptoms. We found that derived phenotypes were associated with not only the diagnoses and severity of depression and mania but also specific individual symptoms. These findings suggest that digital phenotyping based on mobile typing kinematics can provide a nuanced understanding of mood disorders beyond diagnoses and severity aggregates and can serve as a useful clinical tool.

We identified 3 digital phenotypes based on backspace rates: Low (11.2%), Medium (18%), and High (26.8%), with an SD of 4.8%. Practitioners may use these parameter estimates to intimate the likelihood of observed individual backspace rates from Low, Medium, and High classes through probability density functions. For example, if an individual presents with a backspace rate of 14%, then the unnormalized probability density for observing such rate given Low, Medium, and High class would be respectively 7.01, 5.87, and 0.24 (0.53, 0.45, 0.02 after normalization). To maximize the likelihood of observed data, the phenotype for the individual is likely to be low. Compared with the traditional approach of categorization based on cutoff thresholds that just assign an individual to 1 class, our model-driven likelihood approach allows a more nuanced categorization where the likelihood for each class can be derived for an individual. This additional information can be helpful, especially when an individual may share a similar likelihood to more than 1 class.

Our derived phenotypes showed differential associations with depression and mania. Mood disorder diagnoses were used as an external criterion to determine the number of phenotypes. Further analyses showed that the Medium type was associated with depression severity, and the High type was associated with mania severity. Furthermore, our data-driven approach showed predictive use incremental to diagnoses and medication use in differentiating the severity of depression and mania. As mobile phones have become increasingly accessible, identifying and benchmarking disordered digital typing behaviors can be cost-effective. Backspace rates, as we show in this study, can reflect mental health from pervasive behaviors of mobile phone use. This shows the potential of empirically derived digital markers for mood disorder research and practices.

Furthermore, as the severity and diagnoses encompass mood disorder symptoms at a gross level, we examined the use of our digital phenotypes for specific symptoms. We found the Medium phenotype to correlate with psychiatric anxiety, somatic anxiety, general somatic symptoms, and genital symptoms. Existing research has documented the relation between rumination and these symptoms [23,24]. Individuals in the Medium backspace rate phenotype may have a ruminative cognitive style that can also contribute to the stress generation behind their depression [45]. Psychosocial interventions, including cognitive behavioral therapy and mindfulness-based interventions, have demonstrated large effects in reducing rumination [46,47]. Future research can further uncover the shared mechanism between backspace typing and these psychiatric symptoms, potentially including alterations in potential or sustained threat processing, frustrative nonreward, attention, and response selection.

In our study, the high phenotype was correlated with insomnia, agitation, psychic anxiety, elevated mood, motor activity, and irritability. These symptoms have been found to be associated with deficits in inhibitory control, as well as a heightened risk for suicidal and self-injurious thoughts and behaviors [48-50]. This suggests the need for increased monitoring for potential suicide and for adjunctive sleep interventions such as CBT-i. Future research may uncover additional and distinct mechanisms

XSL•FO
RenderX

for this phenotype with respect to inhibition or suppression, goal selection, updating, representation, and maintenance.

Our study's findings suggest that mobile keyboard backspace rates offer a promising, low-burden method for continuous mood disorder symptom tracking. Our usability data indicate general ease to acclimate to the use of the BiAffect app. This echoes past studies where researchers were able to engage participants to use BiAffect on a daily basis for individuals with suicidality and affective instability for 54 days on average [51], for individuals with multiple sclerosis for 37 days on average [52], and for individuals with bipolar disorders for 8 weeks on average [11]. In our study, users found that they learned to use the app quickly, easily remembered how to use it, found it easy to learn to use, and quickly became skillful with it. These high ratings suggest that the app's design and functionality are well-received by users, indicating a generally positive user experience. However, the slightly lower score for becoming skillful with the app (4.84/7) highlights potential room for improvement in the onboarding process to help users achieve proficiency more quickly. For sustained use in real-world applications, it is critical that future iterations of the digital phenotype incorporate adaptive learning algorithms to accommodate changes in individual typing patterns over time, ensuring the app remains sensitive and specific to the user's symptomatology. In addition, real-life deployment must carefully consider privacy and data security concerns. Several practical strategies may be helpful in effectively integrating digital phenotypes into clinical workflows. First, training programs for health care providers to proficiently and appropriately interpret and apply digital biomarkers to clinical care are important. Second, developing standardized protocols for the inclusion of digital phenotypes in diagnostic and treatment processes will help streamline their use. Third, fostering a collaborative approach where digital phenotypes complement traditional diagnostic tools can enhance clinical decision-making. Finally, ongoing feedback loops among clinicians, patients, and researchers are essential to continuously refine and optimize the tool for clinical use. In integration with clinical workflows, health care providers should carefully access and interpret the data within the context of broader treatment plans.

Limitations of this study suggest directions for future research. A limitation of the study is the small sample size. Future research should explore larger samples and more diverse populations to assess the validity of the findings. Nevertheless, the study provides valuable insights into the potential of digital phenotyping in studying and treating mood disorders. As this study used real-life data, keystroke use was not monitored in a controlled laboratory setting. While the real-life data represents a strength in the generalizability of this study, our design precludes the possibility of monitoring the specific contexts of backspace use. This points to the future direction for studying time- and situation-specific backspace use in further elucidation of the real-time mechanism behind our demonstrated relevance of the derived phenotypes.

Acknowledgments

This study was funded by the National Institute of Mental Health (5R01MH120168). The funder played no role in the study design, data collection, analysis, and interpretation of data or the writing of this manuscript.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request. The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Authors' Contributions

QL contributed to conceptualization, methodology, software, formal analysis, data curation, writing-original draft, writing-review, editing, and visualization. EN contributed to data curation, writing-review, editing, and visualization. MKR contributed to the investigation, data curation, writing-review, and editing. AC performed the investigation and data curation. SK contributed to the investigation, data curation, writing-review, and editing. AB, EK, FH, JD, SLL, TN, and JZ contributed to the investigation and data curation. TT contributed to the investigation, data curation, writing-review, and editing. APD contributed to data curation, writing-review, and editing. AL and OA contributed to funding acquisition and supervision.

Conflicts of Interest

Dr Ajilore is a cofounder of KeyWise AI. He also serves as a consultant for Otsuka Pharmaceuticals and on the advisory boards of Blueprint, Embodied Labs, and Sage Therapeutics. Dr Leow is a cofounder of KeyWise AI and has served as a consultant for Otsuka US and Buoy Health. Dr Langenecker is the co-owner of Secondary Triad LLC, a graphic artist company. The other authors have no conflicts of interest to disclose.

References

1. De Angel V, Lewis S, White K, Oetzmann C, Leightley D, Oprea E, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022;5(1):3 [FREE Full text] [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]

XSL•FO
RenderX

2. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016;41(7):1691-1696 [FREE Full text] [doi: 10.1038/npp.2016.7] [Medline: 26818126]
3. Fried EI, Nesse RM, Zivin K, Guille C, Sen S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. Psychol Med 2014;44(10):2067-2076 [FREE Full text] [doi: 10.1017/S0033291713002900] [Medline: 24289852]
4. Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med 2015;13:72 [FREE Full text] [doi: 10.1186/s12916-015-0325-4] [Medline: 25879936]
5. Liu Q, Joiner RJ, Trichtinger LA, Tran T, Cole DA. Dissecting the depressed mood criterion in adult depression: the heterogeneity of mood disturbances in major depressive episodes. J Affect Disord 2023;323:392-399. [doi: 10.1016/j.jad.2022.11.047] [Medline: 36455714]
6. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023;25:e44502 [FREE Full text] [doi: 10.2196/44502] [Medline: 37792430]
7. Cao B, Zheng L, Zhang C, Yu PS, Piscitello A, Zulueta J, et al. DeepMood: modeling mobile phone typing dynamics for mood detection. : ACM; 2017 Presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017 August 13-17; Halifax, NS, Canada p. 747-755. [doi: 10.1145/3097983.3098086]
8. Moore R, Ajilore O, Leow A. Building a fitness tracker for the brain: a journey from lab to consumers. JCB 2022;27(1):27. [doi: 10.5912/jcb1017]
9. Ning E, Cladek AT, Ross MK, Kabir S, Barve A, Kennelly E, et al. Smartphone-derived virtual keyboard dynamics coupled with accelerometer data as a window into understanding brain health: smartphone keyboard and accelerometer as window into brain health. Proc SIGCHI Conf Hum Factor Comput Syst 2023;2023:326 [FREE Full text] [doi: 10.1145/3544548.3580906] [Medline: 38115842]
10. Vesel C, Rashidisabet H, Zulueta J, Stange J, Duffecy J, Hussain F, et al. Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: a BiAffect iOS study. J Am Med Inform Assoc 2020;27(7):1007-1018 [FREE Full text] [doi: 10.1093/jamia/ocaa057] [Medline: 32467973]
11. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker SA, et al. Predicting mood disturbance severity with mobile phone keystroke metadata: a BiAffect digital phenotyping study. J Med Internet Res 2018;20(7):e241. [doi: 10.2196/jmir.9775] [Medline: 30030209]
12. Blair RJR, Mathur A, Haines N, Bajaj S. Future directions for cognitive neuroscience in psychiatry: recommendations for biomarker design based on recent test re-test reliability work. Current Opinion in Behavioral Sciences 2022;44:101102. [doi: 10.1016/j.cobeha.2022.101102]
13. Moriarity DP, Ellman LM, Coe CL, Olino TM, Alloy LB. A physiometric investigation of inflammatory composites: comparison of "a priori" aggregates, empirically-identified factors, and individual proteins. Brain Behav Immun Health 2021;18:100391 [FREE Full text] [doi: 10.1016/j.bbih.2021.100391] [Medline: 34877552]
14. Moriarity DP, Alloy LB. Back to basics: the importance of measurement properties in biological psychiatry. Neurosci Biobehav Rev 2021;123:72-82 [FREE Full text] [doi: 10.1016/j.neubiorev.2021.01.008] [Medline: 33497789]
15. Shackman AJ, Fox AS. Getting serious about variation: lessons for clinical neuroscience (A Commentary on 'The Myth of Optimality in Clinical Neuroscience'). Trends Cogn Sci 2018;22(5):368-369 [FREE Full text] [doi: 10.1016/j.tics.2018.02.009] [Medline: 29576465]
16. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]
17. Brusco NK, Watts JJ. Empirical evidence of recall bias for primary health care visits. BMC Health Serv Res 2015;15:381. [doi: 10.1186/s12913-015-1039-1] [Medline: 26373712]
18. Langford AT, Solid CA, Scott E, Lad M, Maayan E, Williams SK, et al. Mobile phone ownership, health apps, and tablet use in US adults with a self-reported history of hypertension: cross-sectional study. JMIR Mhealth Uhealth 2019;7(1):e12228 [FREE Full text] [doi: 10.2196/12228] [Medline: 31344667]
19. Zulueta J, Demos AP, Vesel C, Ross M, Piscitello A, Hussain F, et al. The effects of bipolar disorder risk on a mobile phone keystroke dynamics based biomarker of brain age. Front Psychiatry 2021;12:739022 [FREE Full text] [doi: 10.3389/fpsyt.2021.739022] [Medline: 35002792]
20. Alloy LB, Just N, Panzarella C. Attributional style, daily life events, and hopelessness depression: subtype validation by prospective variability and specificity of symptoms. Cogn Ther Res 1997;21(3):321-344. [doi: 10.1023/A:1021878516875]
21. Ingram RE, Trenary L, Odom M, Berry L, Nelson T. Cognitive, affective and social mechanisms in depression risk: cognition, hostility, and coping style. Cognition & Emotion 2007;21(1):78-94. [doi: 10.1080/02699930600950778]
22. Beck AT, Rush AJ, Shaw BF, Emery G. Cognitive Therapy of Depression. New York: Guilford Press; 1979.
23. Nolen-Hoeksema S. The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. J Abnorm Psychol 2000;109(3):504-511. [doi: 10.1037//0021-843x.109.3.504]
24. Olatunji BO, Naragon-Gainey K, Wolitzky-Taylor KB. Specificity of rumination in anxiety and depression: a multimodal meta‐analysis. Clinical Psychology: Science and Practice 2013;20(3):225-257. [doi: 10.1037/h0101719]

XSL•FO
RenderX

25. McEvoy PM, Watson H, Watkins ER, Nathan P. The relationship between worry, rumination, and comorbidity: evidence for repetitive negative thinking as a transdiagnostic construct. J Affect Disord 2013;151(1):313-320. [doi: 10.1016/j.jad.2013.06.014] [Medline: 23866301]
26. Gruber J, Eidelman P, Johnson SL, Smith B, Harvey AG. Hooked on a feeling: rumination about positive and negative emotion in inter-episode bipolar disorder. J Abnorm Psychol 2011;120(4):956-961 [FREE Full text] [doi: 10.1037/a0023667] [Medline: 21553935]
27. Feldman GC, Joormann J, Johnson SL. Responses to positive affect: a self-report measure of rumination and dampening. Cognit Ther Res 2008;32(4):507-525 [FREE Full text] [doi: 10.1007/s10608-006-9083-0] [Medline: 20360998]
28. Gilbert KE. The neglected role of positive emotion in adolescent psychopathology. Clin Psychol Rev 2012;32(6):467-481. [doi: 10.1016/j.cpr.2012.05.005] [Medline: 22710138]
29. Wisco BE, Nolen-Hoeksema S. Ruminative response style. Risk Factors in Depression 2008:221-236. [doi: 10.1016/b978-0-08-045078-0.00010-1]
30. Kozicky J, Ha TH, Torres IJ, Bond DJ, Honer WG, Lam RW, et al. Relationship between frontostriatal morphology and executive function deficits in bipolar I disorder following a first manic episode: data from the systematic treatment optimization program for early mania (STOP-EM). Bipolar Disord 2013;15(6):657-668. [doi: 10.1111/bdi.12103] [Medline: 23919287]
31. Trivedi MH, Greer TL. Cognitive dysfunction in unipolar depression: implications for treatment. J Affect Disord 2014;152-154:19-27. [doi: 10.1016/j.jad.2013.09.012] [Medline: 24215896]
32. Trivedi MH, Morris DW, Wisniewski SR, Lesser I, Nierenberg AA, Daly E, et al. Increase in work productivity of depressed individuals with improvement in depressive symptom severity. Am J Psychiatry 2013;170(6):633-641. [doi: 10.1176/appi.ajp.2012.12020250] [Medline: 23558394]
33. Dunkin JJ, Leuchter AF, Cook IA, Kasl-Godley JE, Abrams M, Rosenberg-Thompson S. Executive dysfunction predicts nonresponse to fluoxetine in major depression. J Affect Disord 2000;60(1):13-23. [doi: 10.1016/s0165-0327(99)00157-3] [Medline: 10940443]
34. Robinson LJ, Ferrier IN. Evolution of cognitive impairment in bipolar disorder: a systematic review of cross-sectional evidence. Bipolar Disord 2006;8(2):103-116. [doi: 10.1111/j.1399-5618.2006.00277.x] [Medline: 16542180]
35. Selva G, Salazar J, Balanzá-Martínez V, Martínez-Arán A, Rubio C, Daban C, et al. Bipolar I patients with and without a history of psychotic symptoms: do they differ in their cognitive functioning? J Psychiatr Res 2007;41(3-4):265-272. [doi: 10.1016/j.jpsychires.2006.03.007] [Medline: 16762369]
36. Godard J, Baruch P, Grondin S, Lafleur MF. Psychosocial and neurocognitive functioning in unipolar and bipolar depression: a 12-month prospective study. Psychiatry Res 2012;196(1):145-153. [doi: 10.1016/j.psychres.2011.09.013] [Medline: 22370154]
37. Yang T, Zhao G, Mao R, Huang J, Xu X, Su Y, et al. The association of duration and severity of disease with executive function: differences between drug-naïve patients with bipolar and unipolar depression. Journal of Affective Disorders 2018;238:412-417. [doi: 10.1016/j.jad.2018.05.051] [Medline: 29909305]
38. Cotrena C, Branco LD, Shansis FM, Fonseca RP. Executive function impairments in depression and bipolar disorder: association with functional impairment and quality of life. J Affect Disord 2016;190:744-753. [doi: 10.1016/j.jad.2015.11.007] [Medline: 26606718]
39. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2:88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
40. Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J Affect Disord 2015;172:96-102 [FREE Full text] [doi: 10.1016/j.jad.2014.10.010] [Medline: 25451401]
41. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23(1):56-62 [FREE Full text] [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
42. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978;133:429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
43. Plummer M. JAGS: Just another Gibbs sampler. 2012. URL: https://ui.adsabs.harvard.edu/abs/2012ascl.soft09002P [accessed 2023-07-06]
44. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J Royal Statistical Soc B 2002;64(4):583-639. [doi: 10.1111/1467-9868.00353]
45. Alloy LB, Abramson LY, Metalsky GI, Hartlage S. The hopelessness theory of depression: attributional aspects. Br J Clin Psychol 1988;27(1):5-21. [doi: 10.1111/j.2044-8260.1988.tb00749.x] [Medline: 3281732]
46. Hawley LL, Schwartz D, Bieling PJ, Irving J, Corcoran K, Farb NAS, et al. Mindfulness practice, rumination and clinical outcome in mindfulness-based treatment. Cogn Ther Res 2013;38(1):1-9. [doi: 10.1007/s10608-013-9586-4]
47. Watkins E. Psychological treatment of depressive rumination. Curr Opin Psychol 2015;4:32-36. [doi: 10.1016/j.copsyc.2015.01.020]
48. Liu Q, Nestor BA, Cole DA. Differential associations of phasic and tonic irritability to suicidality among U.S. adults. J Affect Disord 2021;292:391-397. [doi: 10.1016/j.jad.2021.05.066] [Medline: 34139413]

XSL•FO
RenderX

49. Orri M, Perret LC, Turecki G, Geoffroy MC. Association between irritability and suicide-related outcomes across the life-course. Systematic review of both community and clinical studies. J Affect Disord 2018;239:220-233. [doi: 10.1016/j.jad.2018.07.010] [Medline: 30025311]
50. Walther S, Bernard J, Mittal VA, Shankman S. The utility of an RDoC motor domain to understand psychomotor symptoms in depression. Psychol Med 2019;49(2):212-216 [FREE Full text] [doi: 10.1017/S0033291718003033] [Medline: 30322416]
51. Knol L, Nagpal A, Leaning IE, Idda E, Hussain F, Ning E, et al. Smartphone keyboard dynamics predict affect in suicidal ideation. NPJ Digit Med 2024;7(1):54 [FREE Full text] [doi: 10.1038/s41746-024-01048-1] [Medline: 38429434]
52. Chen MH, Leow A, Ross MK, DeLuca J, Chiaravalloti N, Costa SL, et al. Associations between smartphone keystroke dynamics and cognition in MS. Digit Health 2022;8:20552076221143234 [FREE Full text] [doi: 10.1177/20552076221143234] [Medline: 36506490]

Abbreviations

DIC: Deviance Information Criterion
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
HAM-D: Hamilton Depression Rating Scale
HC: healthy control
MCMC: Markov Chain Monte Carlo
MD: mood disorder
YMRS: Young Mania Rating Scale

Edited by A Mavragani; submitted 26.07.23; peer-reviewed by SW Bae, W Qu; comments to author 05.12.23; revised version received 11.04.24; accepted 24.07.24; published 29.10.24.

Please cite as:
Liu Q, Ning E, Ross MK, Cladek A, Kabir S, Barve A, Kennelly E, Hussain F, Duffecy J, Langenecker SA, Nguyen TM, Tulabandhula T, Zulueta J, Demos AP, Leow A, Ajilore O
Digital Phenotypes of Mobile Keyboard Backspace Rates and Their Associations With Symptoms of Mood Disorder: Algorithm Development and Validation
J Med Internet Res 2024;26:e51269
URL: https://www.jmir.org/2024/1/e51269
doi:10.2196/51269
PMID:

©Qimin Liu, Emma Ning, Mindy K Ross, Andrea Cladek, Sarah Kabir, Amruta Barve, Ellyn Kennelly, Faraz Hussain, Jennifer Duffecy, Scott A Langenecker, Theresa M Nguyen, Theja Tulabandhula, John Zulueta, Alexander P Demos, Alex Leow, Olusola Ajilore. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Discriminant Power of Smartphone-Derived Keystroke Dynamics for Mild Cognitive Impairment Compared to a Neuropsychological Screening Test: Cross-Sectional Study

Jin-Hyuck Park[1], MPH, PhD

Department of Occupational Therapy, College of Medical Science, Soonchunhyang University, Asan, Republic of Korea

Corresponding Author:
Jin-Hyuck Park, MPH, PhD
Department of Occupational Therapy
College of Medical Science
Soonchunhyang University
Room 1401, Medical Science
22, Soonchunhayng-ro, Shinchang-myeon
Asan, 31538
Republic of Korea
Phone: 82 41 530 4773
Fax: 82 41 530 3085
Email: roophy@naver.com

Abstract

Background: Conventional neuropsychological screening tools for mild cognitive impairment (MCI) face challenges in terms of accuracy and practicality. Digital health solutions, such as unobtrusively capturing smartphone interaction data, offer a promising alternative. However, the potential of digital biomarkers as a surrogate for MCI screening remains unclear, with few comparisons between smartphone interactions and existing screening tools.

Objective: This study aimed to investigate the effectiveness of smartphone-derived keystroke dynamics, captured via the Neurokeys keyboard app, in distinguishing patients with MCI from healthy controls (HCs). This study also compared the discriminant performance of these digital biomarkers against the Korean version of the Montreal Cognitive Assessment (MoCA-K), which is widely used for MCI detection in clinical settings.

Methods: A total of 64 HCs and 47 patients with MCI were recruited. Over a 1-month period, participants generated 3530 typing sessions, with 2740 (77.6%) analyzed for this study. Keystroke metrics, including hold time and flight time, were extracted. Receiver operating characteristics analysis was used to assess the sensitivity and specificity of keystroke dynamics in discriminating between HCs and patients with MCI. This study also explored the correlation between keystroke dynamics and MoCA-K scores.

Results: Patients with MCI had significantly higher keystroke latency than HCs ($P<.001$). In particular, latency between key presses resulted in the highest sensitivity (97.9%) and specificity (96.9%). In addition, keystroke dynamics were significantly correlated with the MoCA-K (hold time: $r=-.468$; $P<.001$; flight time: $r=-.497$; $P<.001$), further supporting the validity of these digital biomarkers.

Conclusions: These findings highlight the potential of smartphone-derived keystroke dynamics as an effective and ecologically valid tool for screening MCI. With higher sensitivity and specificity than the MoCA-K, particularly in measuring flight time, keystroke dynamics can serve as a noninvasive, scalable, and continuous method for early cognitive impairment detection. This novel approach could revolutionize MCI screening, offering a practical alternative to traditional tools in everyday settings.

Trial Registration: Thai Clinical Trials Registry TCTR20220415002; https://www.thaiclinicaltrials.org/show/TCTR20220415002

(J Med Internet Res 2024;26:e59247) doi:10.2196/59247

KEYWORDS

digital biomarker; motor function; digital device; neuropsychological screening; screening tools; cognitive assessment; mild cognitive impairment; keystroke dynamics

XSL•FO
RenderX

Introduction

Mild cognitive impairment (MCI), a preclinical stage of Alzheimer disease (AD), causes cognitive declines, especially in memory and executive functions beyond what is expected from normal aging [1]. The importance of early intervention for patients with MCI has highlighted the need for continuous testing of patients with MCI to ensure they have not transitioned from MCI to AD [2]. Current test methods mainly depend on neuropsychological screening tools implemented in a clinical setting [3-5]. Unfortunately, previous studies have reported the low sensitivity of these tools, particularly the Montreal Cognitive Assessment (MoCA), which limits their effectiveness in discriminating early MCI from healthy aging [4-6]. In addition, various neuropsychological screening tools, such as the Rapid Cognitive Screen, Six-item Screener, Mini-Cog, and Clock Drawing Test, were investigated to determine their usefulness in screening for MCI. However, similar to the MoCA, these tools were found to have low sensitivity and specificity in detecting MCI, leading to limitations in their application. Although tools such as the Short Test of Mental Status and Memory Alteration Test have been reported to show high sensitivity and specificity, only one study has examined these tools [7], which leaves their effectiveness uncertain. Therefore, it has been concluded that existing neuropsychological screening tools that evaluate cognitive function alone are ineffective for screening MCI [7-9].

Apart from cognitive declines, there are several indications that people with MCI are associated, to a certain degree, with motor dysfunction in upper-extremity functions [10]. A previous study indicated that individuals with MCI have poor performance in dual-task gait tests, which involve a sensor-based, upper-extremity motor task during cognitive testing. This could be a marker to distinguish MCI from normal aging, with quite high sensitivity (over 80%) and specificity (over 70%) [11]. Several studies suggest that MCI markers, especially digital biomarkers using data derived from mobile and wearable devices, could be a promising new research field, along with the sensitivity and usability of sensors [11,12]. From this perspective, variations in typing on and touching a smartphone screen have been investigated for early MCI screening, showing considerable differences compared to healthy aging [13,14].

Smartphone-induced behaviors are significantly correlated with cognitive performance, such as the temporal characteristics of typing speed, removal rate, and interval time between pushing and releasing buttons. These keystroke dynamics have been identified as a potential marker for cognitive decline, particularly in attention and working memory [12,15,16]. Keystroke dynamics could be collected seamlessly as users interact with their smartphones during their daily routines, capturing their natural behavior and thereby offering high ecological validity. Therefore, keystroke dynamics have been used for the early screening of AD and Parkinson disease [14,17]. In a previous study, features of keystroke dynamics extracted from text-typing activities were investigated to monitor signs of early cognitive decline in MCI, showing their clinical efficacy [12,15,17]. Especially, the discriminant power of these features could be enhanced when combined with conventional features of cognitive decline such as memory and linguistic abilities [17]. These findings suggest the potential for multimodal features, indicating that this approach is needed to discriminate MCI.

However, most previous studies on sensor-based markers to discriminate MCI examined their efficacy by combining sensor-derived data that could be difficult to use and assess, such as those from accelerometers, voice recorders, and motion capture sensors. This approach requires special devices that are not typically used and must be installed on the body or in the home, creating an artificial environment that deviates from the subjects' natural state and reduces ecological validity, resulting in findings that are not applicable to everyday living [7,18,19]. Also, even if some prior studies revealed the potential of smartphone-derived markers for MCI, it is difficult to ensure their clinical applicability as no direct comparisons have been made with widely used screening tools [13-15,17].

Therefore, this study aimed to examine the efficacy of digital biomarkers obtained through routine interactions with a smartphone keyboard to distinguish MCI from healthy aging in a nonclinical setting, taking into consideration the pragmatic condition of everyday living. In addition, this study aimed to investigate whether classification performance could be improved through the combination of keystroke dynamics and conventional screening tools.

Methods

Participants

The participants were recruited from senior centers and daycare centers in Seoul and Asan, South Korea. The study included 2 groups: 64 healthy controls (HCs) and 47 patients with MCI. In accordance with a previous study [1], the inclusion criteria for MCI were as follows: (1) a subjective memory complaint; (2) memory impairment relative to age- and education-matched HCs, confirmed by performance on a neuropsychological battery (below 1.5 SD); (3) intact global cognitive function, confirmed by a score on the Korean version of the Mini-Mental Status Examination (MMSE-K); (4) independent activities of daily living; and (5) experience using a smartphone for at least 3 months. The exclusion criteria were as follows: (1) diagnosed with dementia by a clinician, (2) neurological or psychiatric disorders such as stroke and depression, and (3) visual or auditory impairments. These criteria are based on the original Petersen criteria, which restricts MCI to memory problems (amnestic MCI) only.

Ethical Consideration

This study was approved by the Institutional Review Board of Soonchunhyang University (202306-SB-070-04) and registered at the Thai Clinical Trials Registry (TCTR20220415002). This study adhered to all relevant guidelines and regulations concerning the ethical conduct of human subject research. All participants provided written informed consent prior to the initiation of any study activities. The consent process ensured that all participants were fully informed about the nature of the study, its objectives, and their rights as participants, including the voluntary nature of their participation. To protect participants' privacy and confidentiality, all study data were

XSL•FO
RenderX

anonymized and deidentified before analysis. No identifying information was collected or stored that could be traced back to individual participants. The study followed strict data protection protocols to ensure the confidentiality of the information provided by the participants. All participants were compensated with ₩30,000 (~US $21.74) for their study participation. No images that could potentially identify individual participants were included in the paper or supplementary materials.

Procedure

All participants performed the Korean version of the MoCA (MoCA-K) and the computerized Corsi block-tapping test (CBT) before their keystroke data were collected. Both the MoCA-K and the computerized CBT were used to assess baseline cognitive performance. Subsequently, the relationship between these assessments and keystroke dynamics was examined. The MoCA-K was selected for its role as a screening tool for MCI and its ability to assess global cognitive function, while the CBT was chosen for its efficacy in assessing working memory decline, which is common in patients with MCI.

Participants then installed the "Neurokeys" mobile app (Neurocast), developed for Android and iOS, which is available for free and allows users to measure health conditions through typing on a smartphone. The Neurokeys keyboard replaced the default keyboard, and participants were allowed to become accustomed to it for a week. During regular typing activities, keyboard interactions (keystroke dynamics) such as key presses and releases were recorded and time-stamped in the background. The keystroke dynamics were stored in a JSON format and indexed in a database, accessible only to the app. The app periodically delivered information on uniquely coded keystroke dynamics to a remote cloud server (Microsoft Azure), when the participant's device was connected to Wi-Fi and charging.

Data collection lasted a month, and all participants were instructed not to have anyone other than themselves type on their phones. Although it was impossible to confirm whether participants typed all data definitively, the author reconfirmed that only the participants themselves used their phones for 1 month before data analysis. This study did not impose a minimum daily usage requirement on participants, as the aim of this study was to capture keystroke dynamics that most accurately reflect the natural status of their daily lives. Within a month, 3530 typing sessions were collected, 2250 (63.7%) from 64 HCs and 1280 (36.3%) from 35 patients with MCI. Of the 3530 typing session, 2740 (77.6%) with more than 40 key presses per session were finally used to ensure sufficient data for meaningful comparison, in accordance with a previous study [14]. The flow diagram of the overall procedure of this study is presented in Figure 1.

Two variables of the keystroke dynamics stored in the JSON files were analyzed in this study: hold time (HT) and flight time (FT). These 2 variables were chosen because a previous study showed that they are significant in discriminating MCI from healthy aging among other keystroke dynamics [13]. HT refers to the time interval between pressing and releasing a key, while FT indicates the time interval between releasing a key and pressing the next key (Figure 2). To preprocess data, HT >700 microseconds and FT >3 seconds were excluded according to the guidelines in a previous study [14].

Figure 1. The flow diagram of the procedure of this study. CBT: Corsi block-tapping test; MCI: mild cognitive impairment; MoCA-K: Korean version of Montreal Cognitive Assessment.

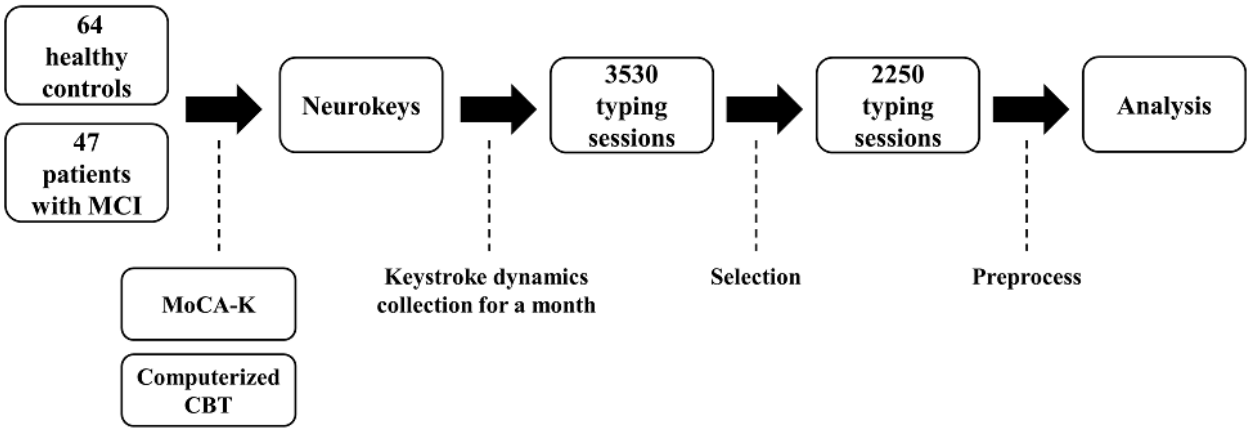

XSL•FO
RenderX

Figure 2. Schematic representation of keystroke dynamics derived from time-stamped keyboard interactions.

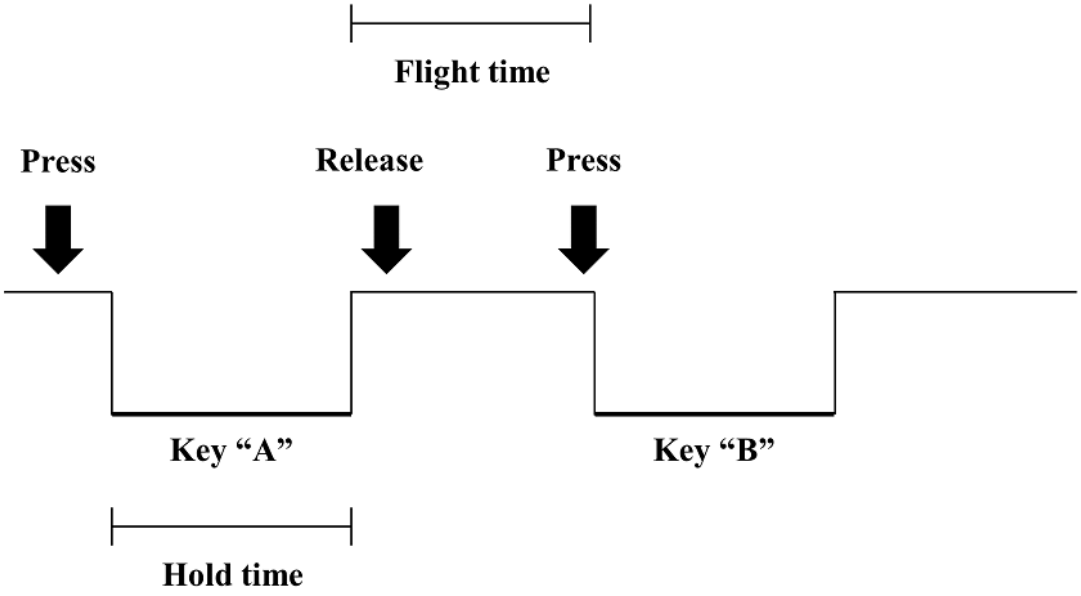

Outcome Measures

The original MoCA was developed to discriminate MCI from normal aging, and its domains consist of visuospatial/executive function, attention, memory, language, abstraction, and orientation [20]. Its score ranges from 0 to 30 points, where higher scores indicate better global cognitive function. The MoCA-K was adapted based on the original MoCA [21] (Multimedia Appendix 1). The cutoff score of the MoCA-K for MCI was 23, and 1 point could be added for participants with fewer than 6 years of education. In a previous study with a cutoff score of 23, its sensitivity and specificity were 94.2% and 40.5%, respectively [21].

The computerized CBT was used to assess attention and working memory. Since patients with MCI commonly exhibit a decline in memory and executive functions [1], a working memory test that evaluates both memory and executive function components was chosen [22]. In the task, 9 white squares were randomly located on a tablet monitor. Some of the squares changed their color from white to red in a sequential manner while participants were asked to memorize the locations and sequences of the changed squares. Afterward, participants were instructed to point out the changed squares in the order in which their colors had changed by touching the screen [23,24]. In this study, the number of squares to be changed was 5, and 15 trials were implemented. A previous study indicated that the average square span of HCs is 5 [23].

The MMSE is a widely used tool for screening cognitive impairment. The MMSE-K was the result of the MMSE being translated into Korean and standardized [25] (Multimedia Appendix 2). It comprises 12 items assessing 7 cognitive domains: orientation, memory registration, attention and calculation, recall, language, understanding and judgment, and visual construction. Its score ranges from 0 to 30, with scores below 24 indicating cognitive impairment [25].

All outcome measures were implemented by an occupational therapist with 5 years of clinical experience.

Statistical Analysis

SPSS for Windows (version 22.0; IBM Corp) was used to analyze the data. The demographic characteristics of the participants were analyzed using descriptive statistics. Independent 2-tailed *t* test and chi-square test were used to compare both groups. A receiver operating characteristic (ROC) curve analysis was used to confirm sensitivity and specificity, and a cutoff score for patients with MCI was determined according to the highest Youden Index (*sensitivity* + *specificity* – *1*), which could be a criterion for choosing an optimal cutoff score. To conduct ROC curve analysis of the combination of keystroke dynamics as a single variable, predictive probability through logistic regression was used in accordance with a previous study [26]. A Spearman correlation was performed to examine the relationship between keystroke dynamics and global cognitive function and working memory. Statistical significance was set at $P<.05$.

Results

General and Clinical Characteristics in Both Groups

There were no statistically significant differences in demographic characteristics between both groups (all $P>.05$) except for the score of the MoCA-K ($P<.001$). This finding indicated that HCs were age- and education-matched to patients with MCI except for cognitive function (Table 1).

XSL•FO
RenderX

Table 1. General and clinical characteristics of participants (N=111).

Characteristics	HCs[a] (n=64)	Patients with MCI[b] (n=47)	*t* test or chi-square test (*df*)	*P* value
Age (y), mean (SD)	75.52 (7.00)	74.45 (6.98)	.795 (109)[c]	.43
Sex, n (%)			.053 (1)[d]	.82
Male	30 (47)	21 (45)		
Female	34 (53)	26 (55)		
Education period (y), mean (SD)	6.48 (4.41)	6.34 (4.81)	.163 (109)[c]	.87
Typing experience (y), mean (SD)	11.41 (1.25)	11.87 (1.81)	1.600 (109)[c]	.11
Grooved pegboard test (s), mean (SD)				
Preferred hand	82.98 (5.81)	82.28 (6.02)	0.624 (109)[c]	.53
Nonpreferred hand	92.25 (5.48)	91.15 (4.90)	1.097 (109)[c]	.28
MMSE-K[e] (point), mean (SD)	26.86 (1.39)	26.55 (1.15)	1.229 (109)[c]	.22
MoCA-K[f] (point), mean (SD)	25.83 (2.11)	22.77 (2.26)	7.316 (109)[c]	<.001
CBT[g] (accuracy), mean (SD)	.773 (0.06)	.687 (0.07)	6.817 (109)[c]	<.001
HT[h] (ms), mean (SD)	108.77 (25.25)	184.85 (25.24)	15.689 (109)[c]	<.001
FT[i] (ms), mean (SD)	622.925 (135.14)	1351.51 (242.75)	20.151 (109)[c]	<.001

[a]HC: health control.
[b]MCI: mild cognitive impairment.
[c]2-tailed *t* test.
[d]Chi-square test.
[e]MMSE-K: Korean version of the Mini-Mental State Examination.
[f]MoCA-K: Korean version of the Montreal Cognitive Assessment.
[g]CBT: Corsi block-tapping test.
[h]HT: hold time.
[i]FT: flight time.

Sensitivity, Specificity, and Discriminant Power

For discriminating patients with MCI from the matched HC, both HT and FT showed a higher Youden Index than the MoCA-K (HT: .847; FT: .947; and MoCA-K: .469), suggesting that keystroke dynamics can better discriminate MCI compared to the conventional MCI screening tool (Table 2). Specifically, FT yielded a maximum sensitivity (97.9%) and specificity (94.7%). Interestingly, when combined with the MoCA-K, HT had a higher Youden Index, while FT did not (Table 2 and Figure 3).

Table 2. Sensitivity and specificity of MCI[a] detection (N=111).

Variable	Sensitivity	Specificity	Youden index	Cutoff	AUC[b] (95% CI)	*P* value
MoCA-K[c] (point)	.938	.532	.469	22.50	.831 (.756-.903)	<.001
HT[d] (ms)	.894	.953	.847	159.50	.976 (.954-1.000)	<.001
FT[e] (ms)	.979	.969	.947	884.00	.997 (.990-.1000)	<.001
MoCA-K + HT (probability)	.957	.953	.911	.427	.988 (.973-.1000)	<.001
MoCA-K + FT (probability)	.979	.969	.947	.255	.997 (.990-.1000)	<.001

[a]MCI: mild cognitive impairment.
[b]AUC: area under the curve.
[c]MoCA-K: Korean version of the Montreal Cognitive Assessment.
[d]HT: hold time.
[e]FT: flight time.

XSL•FO
RenderX

Figure 3. ROC curves of 5 predictors: (A) MoCA-K, (B) HT, (C) FT, (D), MoCA-K + HT, and (E) MoCA-K + FT. Greater AUC values indicate higher power in discriminating patients with MCI from healthy controls. AUC: area under the curve; FT: flight time; HC: healthy control; HT: hold time; MCI: mild cognitive impairment; MoCA-K: Korean version of Montreal Cognitive Assessment; ROC: receiver operating characteristic.

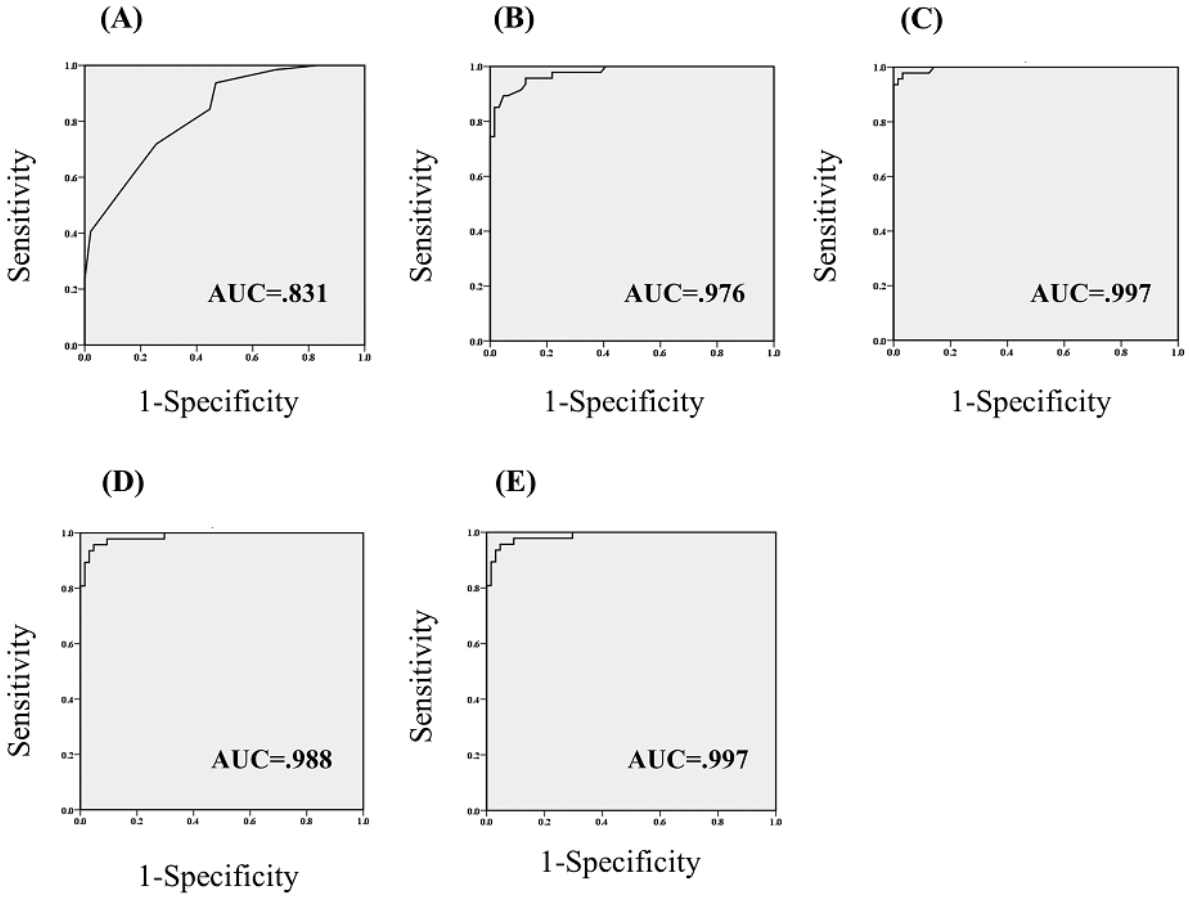

Correlation in Keystroke Dynamics and Cognitive Function

Keystroke dynamics were found to be significantly correlated with the MoCA-K (HT: r=–.468; P<.001; FT: r=–.491; P<.001) and the CBT (HT: r=–.487; P<.001; FT: r=–.492; P<.001; Table 3). These findings suggested that keystroke dynamics are associated with global cognitive function and working memory.

XSL•FO
RenderX

Table 3. Correlation of cognitive function with keystroke dynamics.

	MoCA-K[a]	CBT[b]	HT[c]	FT[d]
MoCA-K				
r	1	.302	–.468	–.491
P value	__[e]	.001	<.001	<.001
CBT				
r	.302	1	–.487	–.492
P value	.001	—	<.001	<.001
HT				
r	–.468	–.487	1	.733
P value	<.001	<.001	—	<.001
FT				
r	–.491	–.492	.733	1
P value	<.001	<.001	<.001	—

[a]MoCA-K: Korean version of the Montreal Cognitive Assessment.

[b]CBT: Corsi block-tapping test.

[c]HT: hold time.

[d]FT: flight time.

[e]Not applicable.

Discussion

Principal Findings

The aim of this study was to evaluate keystroke dynamics as potential digital biomarkers for MCI. For this purpose, keystroke dynamics were investigated in 64 HCs and 47 patients with MCI, alongside the conventional screening tool. By measuring keystroke dynamics through the natural spontaneous use of smartphones, this study aimed to identify distinct patterns that differentiate HCs and patients with MCI.

The findings of this study revealed that keystroke dynamics, particularly HT and FT, were effective in distinguishing between HCs and patients with MCI. Notably, the sensitivity and specificity of keystroke dynamics surpassed those of the conventional screening tool. Furthermore, FT showed the highest Youden Index for all outcome measures, and its Youden Index did not increase when combined with the MoCA-K, suggesting that FT can effectively distinguish MCI on its own; this finding highlights its potential as an alternative screening method for MCI.

Comparison to Prior Work

MCI-induced dysfunction includes motor as well as cognitive dysfunction; this finding is supported by several studies reporting that the degree of motor decline might help in dissociating MCI from healthy aging [27]. Specifically, dominant hand bradykinesia could be an efficient marker to distinguish MCI [14]. In addition, a previous study indicated that Parkinson disease signs characterized by slow movement are correlated with patients with MCI. The findings of this study, showing longer HT and FT in patients with MCI, are consistent with these studies, emphasizing the link between motor function decline and MCI [28]. Notably, motor impairment in patients with MCI considerably affects fine motor skills more than gross motor skills, which leads to a decline in dexterity skills in the upper limbs, including typing [29,30].

Declines in motor function could be attributed to working memory deficits [31,32]. Working memory is required to maintain motor chunk length in high-level motor sequence performance. However, patients with MCI, even patients with amnestic MCI, are typically characterized by deficits in working memory due to reduced brain volume in the prefrontal cortex [33]. This aligns with this study's findings, showing lower performance in the CBT among patients with MCI compared to HCs. Interestingly, no statistically considerable difference in hand dexterity between both groups was confirmed by the grooved pegboard test in this study, suggesting that longer HT and FT in individuals with MCI are more likely to be attributed to a deficit in working memory rather than a purely motor-related issue. Therefore, smartphone typing in individuals with MCI could be negatively affected by working memory deficits.

Prior studies reported the potential of keystroke dynamics derived from touchscreen typing to capture MCI, which supports the findings of this study [13-15]. However, these studies did not compare keystroke dynamics to traditional screening tools, which limits the practical clinical applicability of keystroke dynamics [13-16]. In contrast, this study demonstrated that both HT and FT showed a greater discriminant power for MCI than the MoCA-K, suggesting that keystroke dynamics could be more beneficial to discriminate MCI from normal aging than conventional screening tools. The main factor underlying this comparative finding is based on the metrics between the keystroke dynamics and the MoCA-K. Delayed HT and FT

XSL•FO
RenderX

could represent a working memory deficit, one of the hallmarks of MCI [33]. On the other hand, the MoCA-K includes orientation and language domains, which are not obviously impaired in MCI as it represents global cognitive function [5,6]. This difference in the metric could explain the superiority of the keystroke dynamics for MCI screening.

Notably, FT showed the highest Youden Index for all outcome measures, and its Youden Index did not increase when combined with the MoCA-K. This result suggests that FT is more useful in distinguishing MCI than HT. Furthermore, delayed FT can be used alone as a good marker of MCI, which indicates that a keystroke dynamics monitor could be a cost-effective way of screening MCI, considering that it does not require the additional implementation of neuropsychological tests to increase screening accuracy. Similarly, in a previous study, an ensemble model combining keystroke dynamics and linguistic features did not show superiority in discriminating MCI over keystroke dynamics alone [14], supporting the findings of this study. On the other hand, recent studies have expanded keystroke dynamics analysis by incorporating time feature analysis with graphical transformations analyzed using convolutional neural networks. Additionally, techniques such as natural language processing (eg, noun-to-verbs ratio) have been applied [15]. There have also been efforts to combine keystroke dynamics with brain imaging data analyzed through machine learning to enhance discriminant accuracy [34]. This reflects a growing interest in multimodal biomarkers of naturalistic behavioral data and machine learning methods, integrating various data sources to improve the robustness of biometric systems [35].

Strengths

One of the main strengths of this study is the naturalistic collection of data in an unobtrusive way, reflecting the natural status of the participants while showing a significant correlation with the clinical screening tool. Digital health is an emerging field that enhances disease detection and management through the objective capturing of behavioral characteristics in natural situations [17]. From this digital health perspective, keystroke dynamics derived from the natural interaction with smartphones could be key factors in digital health for individuals with MCI. In addition, keystroke dynamics could be noninvasively and longitudinally captured during routine typing in real daily life without any additional effort, ensuring long-term adherence. Long adherence plays a key role in a continuous process of monitoring, diagnosis, and treatment, which is of high importance toward personalized management [14]. Considering that individualized tests that ignore the longitudinal aspect of data collection could miss people with cognitive impairment in real-world settings [36], digital biomarkers from daily smartphone use could ensure that MCI is not missed, allowing for early intervention, compared to conventional screening tools that are periodically implemented in clinics.

Limitations and Future Directions

Despite the promising findings and strengths of this study, there are some limitations to be considered. First, these findings could not be generalized as MCI was limited to its amnestic type. Nevertheless, considering that individuals with amnestic MCI show less heterogeneity in cognitive function than those with nonamnestic MCI and multidomain MCI [5], the findings of this study have clinical implications. Second, although working memory decline in participants with MCI was investigated via the CBT, the prefrontal cortex, which underlies working memory, was not objectively observed. Thus, it is difficult to affirm whether changes in keystroke dynamics in people with MCI are due to working memory deficits caused by neurodegeneration in the prefrontal cortex. A neuroimaging study will provide more objective evidence of changes in the keystroke dynamics of patients with MCI, allowing for subgroup analysis based on neurodegeneration severity. Third, although this study confirmed through subjective reports that participants did not allow others to use their smartphones, this was not objectively verified. Finally, this study demanded technology familiarization with a smartphone to some extent.

In the future, therefore, to develop an ecologically valid MCI screening variable for a larger population, it is necessary to investigate activities that are easier to use and more widely available. Additionally, it is necessary to determine if different MCI subtypes exhibit distinct features of keystroke dynamics and incorporate these findings into brain imaging data to establish multimodal biomarkers.

Conclusion

Keystroke dynamics reflecting both motor and cognitive deficits were identified as a more clinically useful digital biomarker for MCI, compared to a neuropsychological screening tool. The findings of this study present a new perspective on detecting MCI via keystroke dynamics from smartphone use. These promising results suggest that measures from smartphone typing could serve as an ecologically valid digital biomarker in a nonclinical setting, acting as a surrogate for laboratory-based neuropsychological screening tools.

Acknowledgments

The author would like to thank all participants of the study. This work was supported by the Soonchunhyang University Research Fund and Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2021R1I1A3041487).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

JHP confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Korean version of the Montreal Cognitive Assessment (MoCA-K) form.
[PDF File (Adobe PDF File), 291 KB - jmir_v26i1e59247_app1.pdf]

Multimedia Appendix 2
Korean version of the Mini-Mental Status Examination (MMSE-K) form.
[PDF File (Adobe PDF File), 193 KB - jmir_v26i1e59247_app2.pdf]

References

1. Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med 2004 Sep 20;256(3):183-194 [FREE Full text] [doi: 10.1111/j.1365-2796.2004.01388.x] [Medline: 15324362]
2. Chang CY, Silverman DH. Accuracy of early diagnosis and its impact on the management and course of Alzheimer's disease. Expert Rev Mol Diagn 2004 Jan 09;4(1):63-69. [doi: 10.1586/14737159.4.1.63] [Medline: 14711350]
3. Waldemar G, Dubois B, Emre M, Georges J, McKeith IG, Rossor M, EFNS. Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia: EFNS guideline. Eur J Neurol 2007 Jan 09;14(1):e1-26. [doi: 10.1111/j.1468-1331.2006.01605.x] [Medline: 17222085]
4. Park JH, Park JH. A systematic review of computerized cognitive function tests for the screening of mild cognitive impairment. Korea J Occup Ther 2016;24(2):19-31. [doi: 10.14519/jksot.2016.24.2.02]
5. Park JH. Machine-learning algorithms based on screening tests for mild cognitive impairment. Am J Alzheimers Dis Other Demen 2020 Jun 30;35:1533317520927163 [FREE Full text] [doi: 10.1177/1533317520927163] [Medline: 32602347]
6. Park JH, Jung M, Kim J, Park HY, Kim JR, Park JH. Validity of a novel computerized screening test system for mild cognitive impairment. Int Psychogeriatr 2018 Oct;30(10):1455-1463. [doi: 10.1017/S1041610218000923] [Medline: 29923471]
7. Tran J, Nimojan T, Saripella A, Tang-Wai DF, Butris N, Kapoor P, et al. Rapid cognitive assessment tools for screening of mild cognitive impairment in the preoperative setting: a systematic review and meta-analysis. J Clin Anesth 2022 Jun;78:110682. [doi: 10.1016/j.jclinane.2022.110682] [Medline: 35193049]
8. Ghose S, Das S, Poria S, Das T. Short test of mental status in the detection of mild cognitive impairment in India. Indian J Psychiatry 2019;61(2):184-191 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_145_18] [Medline: 30992614]
9. Rami L, Molinuevo JL, Sanchez-Valle R, Bosch B, Villar A. Screening for amnestic mild cognitive impairment and early Alzheimer's disease with M@T (Memory Alteration Test) in the primary care population. Int J Geriatr Psychiatry 2007 Apr 22;22(4):294-304. [doi: 10.1002/gps.1672] [Medline: 16998781]
10. Rycroft SS, Quach LT, Ward RE, Pedersen MM, Grande L, Bean JF. The relationship between cognitive impairment and upper extremity function in older primary care patients. J Gerontol A Biol Sci Med Sci 2019 Mar 14;74(4):568-574 [FREE Full text] [doi: 10.1093/gerona/gly246] [Medline: 30358815]
11. Toosizadeh N, Ehsani H, Wendel C, Zamrini E, Connor KO, Mohler J. Screening older adults for amnestic mild cognitive impairment and early-stage Alzheimer's disease using upper-extremity dual-tasking. Sci Rep 2019 Jul 29;9(1):10911 [FREE Full text] [doi: 10.1038/s41598-019-46925-y] [Medline: 31358792]
12. Rabinowitz I, Lavner Y. Association between finger tapping, attention, memory, and cognitive diagnosis in elderly patients. Percept Mot Skills 2014 Aug;119(1):259-278. [doi: 10.2466/10.22.PMS.119c12z3] [Medline: 25153754]
13. Stringer G, Couth S, Brown L, Montaldi D, Gledson A, Mellor J, et al. Can you detect early dementia from an email? A proof of principle study of daily computer use to detect cognitive and functional decline. Int J Geriatr Psychiatry 2018 Jul 09;33(7):867-874 [FREE Full text] [doi: 10.1002/gps.4863] [Medline: 29424087]
14. Ntracha A, Iakovakis D, Hadjidimitriou S, Charisis VS, Tsolaki M, Hadjileontiadis LJ. Detection of mild cognitive impairment through natural language and touchscreen typing processing. Front Digit Health 2020 Oct 8;2:567158 [FREE Full text] [doi: 10.3389/fdgth.2020.567158] [Medline: 34713039]
15. Vizer LM, Sears A. Classifying text-based computer interactions for health monitoring. IEEE Pervasive Comput 2015 Oct;14(4):64-71 [FREE Full text] [doi: 10.1109/mprv.2015.85] [Medline: 34267607]
16. Dagum P. Digital biomarkers of cognitive function. NPJ Digit Med 2018 Mar 28;1(1):10 [FREE Full text] [doi: 10.1038/s41746-018-0018-4] [Medline: 31304295]

XSL•FO
RenderX

17. Iakovakis D, Hadjidimitriou S, Charisis V, Bostantjopoulou S, Katsarou Z, Klingelhoefer L, et al. Motor impairment estimates via touchscreen typing dynamics toward parkinson's disease detection from data harvested in-the-wild. Front ICT 2018 Nov 8;5:28. [doi: 10.3389/fict.2018.00028]
18. König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimers Dement (Amst) 2015 Mar;1(1):112-124 [FREE Full text] [doi: 10.1016/j.dadm.2014.11.012] [Medline: 27239498]
19. Fraser KC, Lundholm Fors K, Eckerström M, Öhman F, Kokkinakis D. Predicting MCI status from multimodal language data using cascaded classifiers. Front Aging Neurosci 2019 Aug 2;11:205 [FREE Full text] [doi: 10.3389/fnagi.2019.00205] [Medline: 31427959]
20. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
21. Lee JY, Lee DW, Cho SJ, Na DL, Jeon HJ, Kim SK, et al. Brief screening for mild cognitive impairment in elderly outpatient clinic: validation of the Korean version of the Montreal Cognitive Assessment. J Geriatr Psychiatry Neurol 2008 Jun 01;21(2):104-110. [doi: 10.1177/0891988708316855] [Medline: 18474719]
22. Baddeley A. Working memory and language: an overview. J Commun Disord 2003 May;36(3):189-208. [doi: 10.1016/s0021-9924(03)00019-4] [Medline: 12742667]
23. Monaco M, Costa A, Caltagirone C, Carlesimo GA. Forward and backward span for verbal and visuo-spatial data: standardization and normative data from an Italian adult population. Neurol Sci 2013 May 12;34(5):749-754. [doi: 10.1007/s10072-012-1130-x] [Medline: 22689311]
24. Kessels RPC, van Zandvoort MJE, Postma A, Kappelle LJ, de Haan EHF. The Corsi Block-Tapping Task: standardization and normative data. Appl Neuropsychol 2000;7(4):252-258. [doi: 10.1207/S15324826AN0704_8] [Medline: 11296689]
25. Kwon YC. Korean version of Mini-Mental State Examination (MMSE-K) part I: development of the test for the elderly. J Korean Neuropsychiatr Assoc 1989;28(1):135 [FREE Full text]
26. Doseeva V, Colpitts T, Gao G, Woodcock J, Knezevic V. Performance of a multiplexed dual analyte immunoassay for the early detection of non-small cell lung cancer. J Transl Med 2015 Feb 12;13(1):55 [FREE Full text] [doi: 10.1186/s12967-015-0419-y] [Medline: 25880432]
27. Aggarwal NT, Wilson RS, Beck TL, Bienias JL, Bennett DA. Motor dysfunction in mild cognitive impairment and the risk of incident Alzheimer disease. Arch Neurol 2006 Dec 01;63(12):1763-1769. [doi: 10.1001/archneur.63.12.1763] [Medline: 17172617]
28. Louis ED, Schupf N, Manly J, Marder K, Tang MX, Mayeux R. Association between mild parkinsonian signs and mild cognitive impairment in a community. Neurology 2005 Apr 12;64(7):1157-1161. [doi: 10.1212/01.WNL.0000156157.97411.5E] [Medline: 15824340]
29. Wu Q, Chan JS, Yan JH. Mild cognitive impairment affects motor control and skill learning. Rev Neurosci 2016 Feb;27(2):197-217. [doi: 10.1515/revneuro-2015-0020] [Medline: 26426886]
30. Yan JH, Rountree S, Massman P, Doody RS, Li H. Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. J Psychiatr Res 2008 Oct;42(14):1203-1212. [doi: 10.1016/j.jpsychires.2008.01.006] [Medline: 18280503]
31. Bangert AS, Balota DA. Keep up the pace: declines in simple repetitive timing differentiate healthy aging from the earliest stages of Alzheimer's disease. J Int Neuropsychol Soc 2012 Nov;18(6):1052-1063 [FREE Full text] [doi: 10.1017/S1355617712000860] [Medline: 22929329]
32. Anguera JA, Reuter-Lorenz PA, Willingham DT, Seidler RD. Failure to engage spatial working memory contributes to age-related declines in visuomotor learning. J Cogn Neurosci 2011 Jan;23(1):11-25. [doi: 10.1162/jocn.2010.21451] [Medline: 20146609]
33. Saunders NLJ, Summers MJ. Attention and working memory deficits in mild cognitive impairment. J Clin Exp Neuropsychol 2010 Apr 25;32(4):350-357. [doi: 10.1080/13803390903042379] [Medline: 19787522]
34. Rahman A, Chowdhury M, Khandakar A, Kiranyaz S, Zaman K, Reaz MBI, et al. Multimodal EEG and keystroke dynamics based biometric system using machine learning algorithms. IEEE Access 2021;9:94625-94643 [FREE Full text] [doi: 10.1109/access.2021.3092840]
35. Alfalahi H, Khandoker AH, Chowdhury N, Iakovakis D, Dias SB, Chaudhuri KR, et al. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. Sci Rep 2022 May 11;12(1):7690 [FREE Full text] [doi: 10.1038/s41598-022-11865-7] [Medline: 35546606]
36. de Haan-Rietdijk S, Voelkle MC, Keijsers L, Hamaker EL. Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. Front Psychol 2017 Oct 20;8:1849 [FREE Full text] [doi: 10.3389/fpsyg.2017.01849] [Medline: 29104554]

Abbreviations

AD: Alzheimer disease
CBT: Corsi block-tapping test

XSL•FO
RenderX

FT: flight time
HC: healthy control
HT: hold time
MCI: mild cognitive impairment
MMSE: Mini-Mental Status Examination
MMSE-K: Korean version of the Mini-Mental Status Examination
MoCA: Montreal Cognitive Assessment
MoCA-K: Korean version of the Montreal Cognitive Assessment
ROC: receiver operating characteristic

Edited by A Coristine; submitted 06.04.24; peer-reviewed by A Gallucci, C Mychajliw; comments to author 01.06.24; revised version received 25.06.24; accepted 11.10.24; published 30.10.24.

Please cite as:
Park JH
Discriminant Power of Smartphone-Derived Keystroke Dynamics for Mild Cognitive Impairment Compared to a Neuropsychological Screening Test: Cross-Sectional Study
J Med Internet Res 2024;26:e59247
URL: https://www.jmir.org/2024/1/e59247
doi:10.2196/59247
PMID:

©Jin-Hyuck Park. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Electronic Health Interventions and Cervical Cancer Screening: Systematic Review and Meta-Analysis

Xiaoxia Liu[1], MSN; Lianzhen Ning[1], MSN; Wenqi Fan[1], MSN; Chanyi Jia[2], MSN; Lina Ge[1], MSN

[1]Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, Nanhu Campus, Shenyang, China

[2]Department of Nursing, China Medical University, Shenyang, China

Corresponding Author:
Lina Ge, MSN
Department of Obstetrics and Gynecology
Shengjing Hospital of China Medical University, Nanhu Campus
No36 Sanhao Street
Heping District
Shenyang, 110004
China
Phone: 86 18940251669
Email: geln@sj-hospital.org

Abstract

Background: Cervical cancer is a significant cause of mortality in women. Although screening has reduced cervical cancer mortality, screening rates remain suboptimal. Electronic health interventions emerge as promising strategies to effectively tackle this issue.

Objective: This systematic review and meta-analysis aimed to determine the effectiveness of electronic health interventions in cervical cancer screening.

Methods: On December 29, 2023, we performed an extensive search for randomized controlled trials evaluating electronic health interventions to promote cervical cancer screening in adults. The search covered multiple databases, including MEDLINE, the Cochrane Central Registry of Controlled Trials, Embase, PsycINFO, PubMed, Scopus, Web of Science, and the Cumulative Index to Nursing and Allied Health Literature. These studies examined the effectiveness of electronic health interventions on cervical cancer screening. Studies published between 2013 and 2022 were included. Two independent reviewers evaluated the titles, abstracts, and full-text publications, also assessing the risk of bias using the Cochrane Risk of Bias 2 tool. Subgroup analysis was conducted based on subjects, intervention type, and economic level. The Mantel-Haenszel method was used within a random-effects model to pool the relative risk of participation in cervical cancer screening.

Results: A screening of 713 records identified 14 articles (15 studies) with 23,102 participants, which were included in the final analysis. The intervention strategies used in these studies included short messaging services (4/14), multimode interventions (4/14), phone calls (2/14), web videos (3/14), and internet-based booking (1/14). The results indicated that electronic health interventions were more effective than control interventions for improving cervical cancer screening rates (relative risk [RR] 1.464, 95% CI 1.285-1.667; $P<.001$; $I^2=84\%$), cervical cancer screening (intention-to-treat) (RR 1.382, 95% CI 1.214-1.574; $P<.001$; $I^2=82\%$), and cervical cancer screening (per-protocol; RR 1.565, 95% CI 1.381-1.772; $P<.001$; $I^2=74\%$). Subgroup analysis revealed that phone calls (RR 1.82, 95% CI 1.40-2.38), multimode (RR 1.62, 95% CI 1.26-2.08), SMS (RR 1.41, 95% CI 1.14-1.73), and video- and internet-based booking (RR 1.25, 95% CI 1.03-1.51) interventions were superior to usual care. In addition, electronic health interventions did not show a statistically significant improvement in cervical cancer screening rates among women with HPV (RR 1.17, 95% CI 0.95-1.45). Electronic health interventions had a greater impact on improving cervical cancer screening rates among women in low- and middle-income areas (RR 1.51, 95% CI 1.27-1.79). There were no indications of small study effects or publication bias.

Conclusions: Electronic health interventions are recommended in cervical cancer screening programs due to their potential to increase participation rates. However, significant heterogeneity remained in this meta-analysis. Researchers should conduct large-scale studies focusing on the cost-effectiveness of these interventions.

Trial Registration: CRD42024502884; https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=502884

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58066) doi:10.2196/58066

KEYWORDS

cervical cancer; electronic health record; cancer screening; HPV; women's health; electronic health interventions

Introduction

Cervical cancer is a prevalent gynecologic malignancy within the female reproductive system. Based on the most recent data from the World Health Organization's (WHO) International Agency for Research on Cancer in 2020, an estimated 6 million women worldwide are projected to be diagnosed with cervical cancer, accounting for 3.1% of all new cancer cases globally. In addition, about 3 million women are projected to die from the disease, representing 7.7% of all female deaths from cancer worldwide [1]. Cervical cancer develops gradually, often taking years or even decades to progress into a precancerous lesion, with clinical signs typically not apparent during the early stages of onset. Symptoms such as vaginal bleeding and discharge generally manifest in the intermediate to late stages, by which point the cancer may have invaded surrounding tissues and organs, surpassing the optimal window for treatment [2]. The WHO has initiated a worldwide effort to eliminate cervical cancer by promoting vaccination, screening, and treatment, given that the disease is preventable with current knowledge and technologies [3]. Therefore, cervical cancer screening is essential for the early detection of human papillomavirus (HPV) infection and cervical cytopathy, facilitating early intervention that can significantly reduce both the incidence and mortality rates of cervical cancer.

Over 1.5 billion women globally have never participated in cervical cancer screening [4]. Cancer screening begins by identifying and assessing eligible individuals, and it has been central to health care reform legislation and quality improvement initiatives. To optimize the benefits of screening, both population-based approaches and timely follow-up of abnormal test results are essential [5]. The primary methods for cervical cancer screening include Papanicolaou smears, HPV DNA testing, cytological examinations, and colposcopy. The WHO advocates HPV DNA testing as the preferred initial screening approach, regardless of whether a triage strategy is used. Women who test positive in the HPV DNA test are recommended to undergo further evaluation through genotyping, colposcopy, visual inspection with acetic acid, or cytological examination [6]. While many high-income countries have implemented screening programs, disparities persist among specific population subgroups [7]. Consequently, numerous studies have implemented interventions to enhance screening coverage among women.

Previously, 4 meta-analyses investigated interventions to boost the implementation and uptake of cervical cancer screening, with 2 reviews confirming the effectiveness of health education in improving screening rates. A review by Musa et al [8] demonstrated that cervical cancer education and screening reminders significantly increase screening rates, especially among women with lower educational levels. Brevik et al [9] showed that culturally tailored health education can improve cervical cancer screening rates among minority women. Research by Costa et al [10] confirmed that sending HPV self-testing kits directly to women is more effective at increasing cervical cancer screening rates than simply inviting women to screenings. Alam et al [7] conducted a systematic review of interventions aimed at improving cervical cancer screening among immigrant and refugee women, discovering that health education, screening guidance, and written information interventions effectively enhance screening rates. However, the scope and convenience of these measures are limited, and new strategies are needed to address this issue. The expanding use of electronic health technologies in medical informatics, public health, business, and emerging domains such as health care has broadened the applications of health services and information, which are provided or enhanced through the internet and associated technologies [11,12].

Electronic health allows for the delivery of health information in an accessible manner. During the COVID-19 pandemic, the use of electronic health platforms, such as websites and video conferencing, proliferated [13]. Numerous in-person appointments have been replaced by telehealth, which provides a scalable and adaptable means of offering support, monitoring patient-reported health outcomes, and facilitating continuity of care between hospital visits [14]. Significantly, patients find electronic health initiatives useful and acceptable, and through co-design, these initiatives can support patient-centered care [15]. Many high-income countries use electronic health interventions and media campaigns as educational strategies to promote cervical cancer screening among a broader population of women [16,17]. This approach might help achieve the WHO's suggested elimination targets, which call for 90% of women with precancer and cancer to be treated or managed, 70% of women to be tested by the age of 30, and 90% of girls to receive the full HPV vaccination [3].

Examples of electronic health intervention technologies include web-based tactics, email, mobile or smartphone apps, text messaging, digital games, wearable or monitoring devices, and telemedicine or telehealth [18]. A systematic review by Romli et al [19] on the effectiveness of electronic health interventions for enhancing cervical cancer knowledge found that a mixed approach, including electronic health movies, video education, and didactic sessions, increased cervical cancer screening rates. This indicates that electronic health interventions for health education hold significant potential to improve cervical cancer screening rates. However, empirical research assessing the efficacy of electronic health interventions is needed to provide supporting evidence. A systematic evaluation of mobile technology for cervical cancer screening in low- and middle-income countries, published by Zhang et al [20], revealed that telephone reminders or SMSs significantly increased cervical cancer screening rates compared with traditional communication methods, such as postal mail. The study systematically examined the effects of mobile health intervention strategies on cervical cancer screening rates,

XSL•FO
RenderX

specifically in relation to smartphones. No specific studies have reported the effectiveness of more comprehensive electronic health interventions in improving cervical cancer screening rates. Previous studies have been limited to specific groups or a single approach to electronic health intervention. However, the WHO advocates for a cervical cancer elimination program targeting the entire population [3]. This study encompasses global demographics and incorporates comprehensive electronic health interventions, including, but not limited to, text messages, video screens, apps, voice interactions, websites, and podcasts. Electronic health interventions constitute long-term interventions, which may lead to loss to follow-up, thereby impacting the assessment of their effectiveness. Previous meta-analyses did not further explore intention-to-treat (ITT) and per-protocol (PP) outcomes; therefore, this study extends the evaluation to include ITT and PP analyses of the effectiveness of electronic health interventions in cervical cancer screening.

We conducted a systematic review and meta-analysis to compare the effectiveness of comprehensive electronic health interventions with nonelectronic health interventions in terms of cervical cancer screening rates. In addition, we compared whether the impact of electronic health interventions on cervical cancer screening rates varied based on subjects, intervention type, and economic level. Comprehensive electronic health interventions are defined as all patient-oriented electronic health interventions, including, but not limited to, text messages, video screens, apps, voice interactions, websites, and podcasts. The results provide research evidence on the impact of electronic health interventions on cervical cancer screening rates, which may inform the design of strategies to improve global cervical cancer screening.

Methods

Study Design and Registration

This systematic review was registered with the PROSPERO (International Prospective Register of Systematic Reviews; CRD42024502884) and was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (checklist provided in Multimedia Appendix 1). In the course of the study, to clearly determine the effectiveness of electronic health interventions directly applied to patients and to ensure the accuracy and referability of the combined effect sizes after the meta-analysis, health care workers were excluded from the inclusion criteria, and only patient-related electronic health interventions were included.

Inclusion and Exclusion Criteria

We included peer-reviewed articles published in English that met the eligibility criteria based on the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) strategy (Textbox 1).

Textbox 1. Study inclusion and exclusion criteria defined by the Population, Intervention, Comparison, Outcome, and Study Design.

Inclusion criteria

- Population: women aged 18 years or older receiving cervical cancer screening.
- Intervention: examples of patient-directed electronic health treatments include email, video conferences, videos, activity trackers, websites, podcasts, chat rooms, mobile applications, and text messages or SMS.
- Comparison: any comparator was acceptable, including a nonintervention group or an alternative group using nonsocial media and nonelectronic health interventions.
- Outcome: primary outcome indicator such as, predominantly reported cervical cancer screening rates in the included randomized controlled trials; secondary outcome indicators, including results of the intention-to-treat analysis and the per-protocol analysis of cervical cancer screening rates. The receipt of cervical cancer screening was defined as the completion of at least 1 cervical cancer screening test, including the Papanicolaou test, human papillomavirus DNA test, cytological examinations or colposcopy, from the start of the study to the end of follow-up. Both self-reported screening status (ie, patient reports) and provider-verified screening status (ie, medical records) were acceptable methods for assessing cervical cancer coverage.
- Study design: randomized controlled trials.

Exclusion criteria

- Population: studies focusing on the impact of electronic health interventions on health care workers and community workers to improve cervical cancer screening rates.
- Intervention: studies discussing how electronic health can be used to improve the accuracy of cervical cancer detection.
- Comparison: control group interventions that included electronic health components.
- Outcome: studies focused solely on knowledge, attitudes, or intentions regarding cervical cancer.
- Study design: editorials, letters, reviews, commentary pieces, and other nonresearch articles.

Only English-language publications were considered. When full-text papers were not available, the study authors were contacted to obtain the articles. If the eligibility requirements were met and data on cervical cancer screening were available, the abstract was included in the event that the authors did not reply. In addition, we excluded studies that focused on using electronic health to improve cancer screening rates across multiple types when cervical cancer screening data could not be isolated.

XSL•FO
RenderX

Information Sources and Searches

A total of 8 electronic databases (from inception to present) were searched on December 29, 2023: MEDLINE, Cochrane Central Registry of Controlled Trials, Embase, PsycINFO, PubMed, Scopus, Web of Science, and the Cumulative Index to Nursing and Allied Health Literature. Authors LXX and NLZ conducted the search by combining terms indicative of cervical cancer screening (eg, cervical cancer, screen*, human papillomavirus DNA tests, and Papanicolaou test) with terms indicative of electronic health (eg, mobile phone, SMS, and videos). In addition, we checked the reference lists of the retrieved articles and previous meta-analyses and reviews for additional studies. The search queries used in this review are detailed in the Multimedia Appendix 2. Studies published between 2013 and 2022 were included.

Study Selection

We exported the search results into a citation management system. After duplicates were eliminated, LXX (all articles), NLZ, and FWQ (each evaluated half) independently reviewed the titles and abstracts against the inclusion criteria. For full-text review, abstracts containing ambiguous information were included. LXX (18 articles), NLZ (17 articles), JCY (16 articles), FWQ (17 articles), and GLN (21 articles) reviewed full-text articles. LXX reviewed each article to ensure it was appropriately included or excluded.

Data Collection Process

To extract the data, a prespecified electronic data extraction table was created using the Cochrane Collaboration's Risk of Bias (RoB) tool and the PRISMA guidelines. The extracted data included the study details (author, year, country of origin, study design, application model, and sample size); participant details (medical history and demographics); duration of the intervention and follow-up; details of the intervention and control groups; primary outcomes; and Cochrane rule of reference measures. Authors LXX (8 articles) and NLZ (6 articles) extracted data from full text publications. LXX verified the accuracy of all the data.

Risk of Bias

The RoB in the included studies was evaluated using the Cochrane Collaboration's RoB tool [21]. The domains assessed were selection bias (sequence generation and allocation concealment), reporting bias (selective outcome reporting), attrition bias (incomplete outcome data), and performance or detection bias (blinding of participants, staff, and outcome assessors). For the criteria of low, unclear, and high RoB, the Cochrane Handbook for Systematic Reviews of Interventions was consulted for both within-trial and across-trial assessments. Independent reviews of RoB were conducted by authors LXX (8 publications) and NLZ (6 papers). LXX reviewed each RoB evaluation to ensure accuracy. Publication bias was assessed by visually inspecting a funnel plot for the main outcome and conducting the Egger test [22]. To estimate the pooled effect while accounting for missing studies, the Duval and Tweedie trim-and-fill analysis was used [23].

Statistical Methods

A descriptive presentation of the study, participant, and intervention features, as well as the RoB assessments, is provided. The relevant outcomes are presented in terms of absolute and relative values, and a random-effects model was used to aggregate relative risk (RR) by applying the Mantel-Haenszel technique. The study used data from the longest follow-up period if outcomes were evaluated at different time points. In cases where studies included multiple intervention arms, the meta-analysis included only the most complex intervention, defined as having the most components. The I^2 statistic was used to compute statistical heterogeneity, with a cutoff value of ≥75% considered significant. The Cochran Q test (P<.1) was used to conduct a formal test of homogeneity, which is considered statistically significant [24]. Subgroup analysis was performed according to subjects (women with and without cervical cancer screening experience, women with a long absence from screening, and women with HPV), intervention type (phone call, SMS, multimode, web video, and internet-based booking), and economic level (high-income and middle- and low-income). This study used 2 types of sensitivity analyses: the leave-one-out method and the removal of highly biased studies. For statistical significance, a 2-tailed P<.05 was used. Meta-analyses were conducted using Stata (version 16.0, StataCorp LLC).

Results

Study Inclusion

After removing duplicates, the screening identified 14 articles (comprising 15 studies), [25-38], of which 11 [25-27,29-32,34-37] contained ITT results, and 12 [25-33,35,36,38] contained PP results (Figure 1), representing 23,102 unique patients (Table S1 in Multimedia Appendix 3). A total of 7 studies [27,28,31-33,35,36] had high RoB.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process.

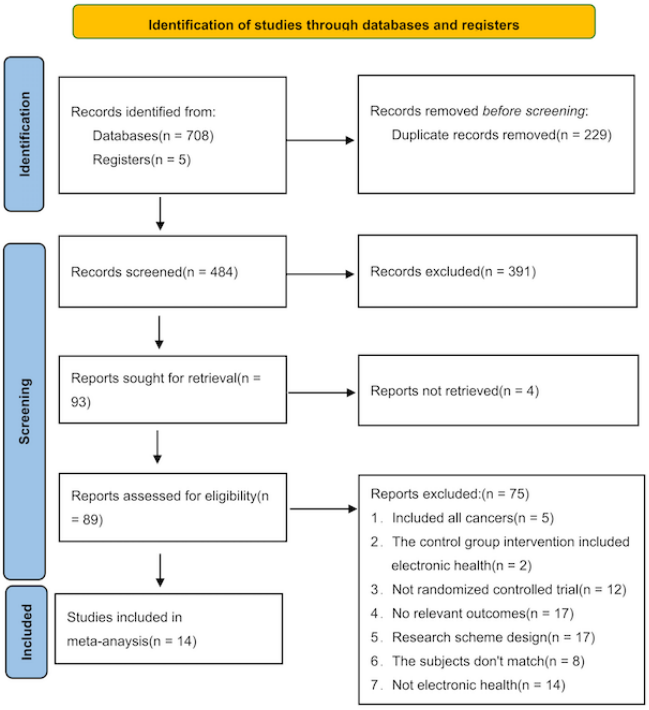

Study Characteristics

The included studies were conducted in America, Europe, Asia, and Africa, and they were published between 2013 and 2022. Single-intervention studies made up the majority of the studies (10/14, 71%) [25-27,29,30,32-35,38] while multimode intervention studies accounted for 29% (4/14) [28,31,36,37]. The interventions were conducted by university-based research teams, governmental or commercial screening initiatives, or health care facilities or departments. The intervention strategies used in these studies included SMS in 4 studies [27,30,32,35], multimode interventions in 4 studies [28,31,36,37], phone call interventions in 2 studies [25,26], web videos in 3 studies [29,33,38], and internet-based booking in 1 study [34]. Outcomes were measured at several time points, including 2, 3, 4, 5, 6, 7, 12, and 14 months. The interventions in most studies (8/14) were based on theoretical models, including the Transtheoretical Model [26-30,32,36,38], theory of planned behavior [27], Health Belief Model [28-30], MINDSPACE (Messenger, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitments, and Ego) framework [32], and social cognitive theory [29,36,38] (Table S1 in Multimedia Appendix 3).

There was a wide variation among the study participants. For example, in some studies, participants were targeted based on geographical region or by their profession as teachers [26,33], emergency patients [27], or young women [32,34]. Some of the participants were women from the general population (regardless of their participation in cervical cancer screening), some were women who had not participated in screening for a long time or had never participated in screening [26,27,37,38], and some were women with HPV [28,35]. Supplementary information regarding the interventions in the experimental and control groups is shown in Table S1 in Multimedia Appendix 3.

Quality Assessment

The RoB evaluations for the research included in the analysis are presented in Figures 2 [25-38] and 3. In summary, 50% (7/14) [27,28,31-33,35,36] of the randomized controlled trials (RCTs) included in the study were categorized as high risk, and the remaining RCTs (7/14) [26,27,29,30,33,34,37] were considered as having some concerns. Only 2 studies (2/14, 14%) [25,38] were classified as low risk. Studies were categorized as high risk due to issues in multiple areas, such as participant and staff blinding, outcome assessment blinding, and assignment effect blinding. Upon visual examination of the funnel plot (Figure 4), no indications of publication bias were observed. The effect sizes exhibited a relatively symmetrical distribution. Most effect sizes fell within the funnel plot, and those outside it were symmetrically distributed. Egger's regression intercept test yielded a significant result (Intercept=2.5, 95% CI 0.32-4.45; P=.03), indicating that the observed results might be influenced by publication bias. The trim and fill approach developed by Duval and Tweedie was used to evaluate the presence of publication bias. After 2 iterations of the analysis, no additional trials were deemed necessary, and the combined effect size remained consistent. After accounting for potential publication bias, the effect of the intervention remained statistically significant (P<.001). Although the observed results might have been influenced by publication bias, there was no indication of a small-study effect (Egger test: t=2.5; P=.03; 95% CI 0.32-4.45).

XSL•FO
RenderX

Figure 2. Risk of bias summary for the included randomized controlled trials (N=14) [25-38].

Figure 3. Analysis of overall risk of bias in the included randomized controlled trials (n=14).

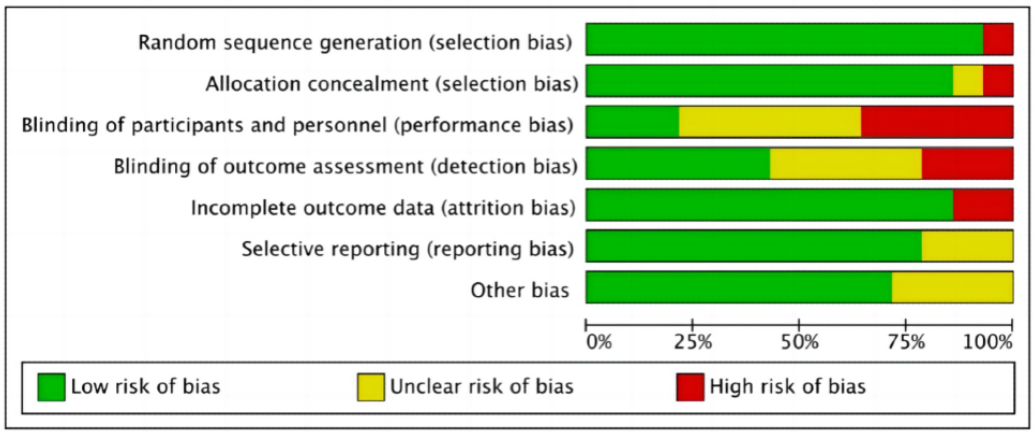

XSL•FO
RenderX

Figure 4. Funnel plot comparing electronic health interventions to nonelectronic health interventions in randomized controlled trials reporting on cervical cancer screening. RR: relative risk.

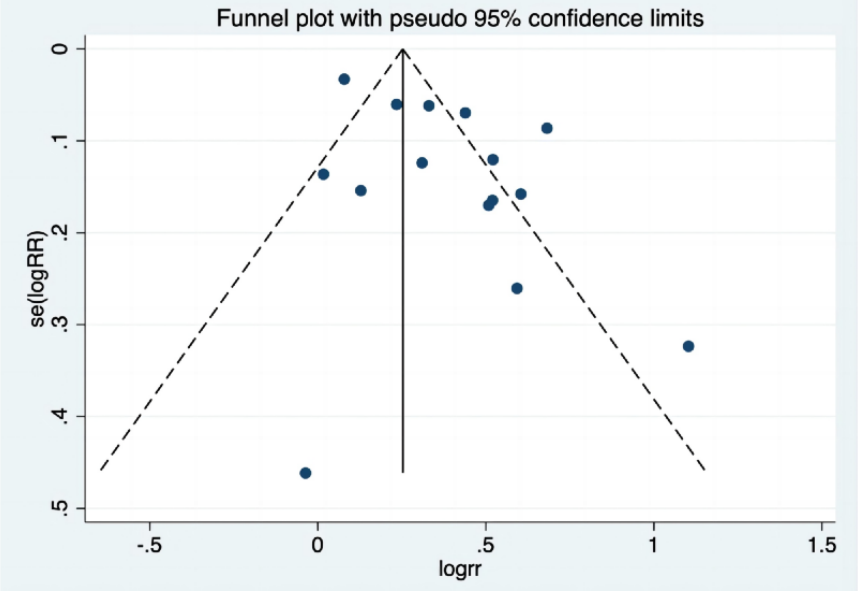

Cervical Cancer Screening

The cervical cancer screening rate was the primary outcome in 12 out of 14 articles (86%) [26-32,34-38]. In total, 10 studies [25,26,28-32,36-38] found that the intervention significantly increased cervical cancer screening rates. The intervention had an absolute impact of 34% (3125/9191) in the intervention arms. Between the intervention and comparison arms, there was an absolute risk difference of 11.9% (95% CI, 7.4-16.4), with 27% (2714/9940) screened in the comparison arm. Among the included RCTs, the overall pooled RR for participation in cancer screening was 1.464 (95% CI 1.285-1.667; Figure 5 [25-38]), meaning that receiving an electronic health intervention raised the odds of screening by 46%. Nonetheless, significant heterogeneity (I^2=84%) was noted.

Similar effect estimates were obtained after stratification according to the intervention method. The order of magnitude of the effects of the various intervention modalities was as follows: phone call (RR 1.82, 95% CI 1.40-2.38; I^2=0%), multimode (RR 1.62, 95% CI 1.26-2.08; I^2=85%), SMS (RR 1.41, 95% CI 1.14-1.73; I^2=71%), and web videos and internet-based booking (RR 1.25, 95% CI 1.03-1.51; I^2=67%; Figure S3 in Multimedia Appendix 3). Subgroup analysis based on the intervention group revealed no significant difference in cervical cancer screening rates between women with HPV in the electronic health interventions group and the nonelectronic health interventions group (RR 1.17, 95% CI 0.95-1.45; I^2=57%). Women with and those without cervical cancer screening experience (RR 1.56, 95% CI 1.32-1.90; I^2=90%) and women with a long absence from screening (RR 1.41, 95% CI 1.08-1.85; I^2=37%) showed a difference between the electronic health interventions group and the control group (Figure S4 in Multimedia Appendix 3). Subgroup analysis according to economic status showed a significant difference for the middle- and low-income group between the electronic health intervention group and the control group (RR 1.51, 95% CI 1.27-1.79; I^2=68%). A significant difference was observed between the intervention and control groups in high-income countries. The RR was 1.41, with a 95% CI of 1.15-1.73, and the heterogeneity (I^2) was 91%. Refer to Figure S5 in Multimedia Appendix 3 for further details. The sensitivity analysis demonstrated that the combined RR and I^2 value were consistent (RR 1.46, 95% CI 1.29-1.66; refer to Figure S6 in Multimedia Appendix 3). After removing studies with a high risk of bias, a meta-analysis (n=7) compared the intervention and control groups at the end of the intervention. The analysis showed an RR of 1.56 (95% CI 1.21-2.02) for cervical cancer screening (refer to Figure S7 in Multimedia Appendix 3).

XSL•FO
RenderX

Figure 5. Forest plot comparing electronic health interventions with nonelectronic health interventions in randomized controlled trials for cervical cancer screening participation (n=14). RR: relative risk [25-38].

Study	Intervention Event	Intervention Total	Comparison Event	Comparison Total	Weight(%)	RR(95%CI) M-H random
Fauziah Abd ullah (2013)	36	201	20	202	3.91	1.809(1.086,3.014)
David Adler (2019)	7	28	7	27	1.71	0.964(0.390,2.382)
Silvina Arrossi (2022)	314	445	163	292	9.07	1.264(1.123,1.423)
Theresa L. Byrd (2013)	64	155	38	153	5.97	1.662(1.191,2.321)
Erica Erwin (2019)	35	272	12	281	2.95	3.013(1.598,5.681)
João Firmino-Machad(2019)	179	273	121	365	8.41	1.978(1.670,2.342)
Sarah Huf (2020)	209	511	366	1388	8.85	1.551(1.353,1.778)
Sarah Huf (2020)	230	485	432	1268	9.04	1.392(1.233,1.571)
Aslı Karakuş Selç(2019)	121	383	79	341	7.31	1.364(1.069,1.739)
H Kitchener (2018)	1518	5267	1190	4467	9.57	1.082(1.014,1.154)
Ditte S Linde (2020)	84	350	80	339	6.94	1.017(0.778,1.329)
Zaahirah Mohammad (2022)	110	201	65	200	7.42	1.684(1.330,2.132)
Beti Thompson (2017)	58	150	50	147	6.42	1.137(0.840,1.538)
S. M. Peitzmeier (2018)	74	220	44	220	6.12	1.682(1.218,2.323)
Rima Marhayu Abdul R(2013)	86	250	47	250	6.31	1.830(1.343,2.493)
Total	3125	9191	2714	9940	100	1.464(1.285,1.667)

Heterogeneity: $Tau^2 = 0.045$; $chi^2 = 89.41$ df = 14 ($p < 0.00001$); I^2=84.3%
Test of overall effect : z= 5.73 P < 0.00001

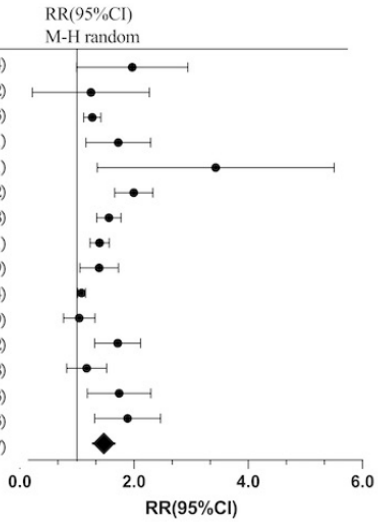

Intention-To-Treat of Cervical Cancer Screening

A total of 11 studies [25-27,29-32,34-37] reported the results of ITT analysis of cervical cancer screening rates. The results indicated that electronic health can improve the cervical cancer screening rate of participants more effectively than that of the control group (RR 1.382, 95% CI 1.214-1.574; *P*<.001; I^2=82.0%; Figure 6 [25-27,29-32,34-37]). The percentage of people screened in the comparison groups was 26.9% (2636/9795), while the absolute effect of screening in the intervention groups was 31.7% (3347/10558). The sensitivity analysis of the leave-one-out method, demonstrated that excluding any single study from the random-effects model did not result in a loss of significance (Figure S8 in Multimedia Appendix 3). The results of sensitivity analysis after excluding studies with high RoB show that the overall pooled RR were robust (RR 1.67, 95% CI 1.22-2.28; Figure S9 in Multimedia Appendix 3).

Figure 6. Forest plot comparing electronic health interventions with nonelectronic health interventions in randomized controlled trials for intention-to-treat analysis of cervical cancer screening participation (n=11). RR: relative risk [25-27,29,30-32,34-37].

Study	Intervention Event_I	Intervention Total_I	Comparison Event_C	Comparison Total_C	Wight(%)	RR(95%CI) M-H random
Fauziah Abd ullah (2013)	36	201	20	202	4.39	1.809(1.086,3.014)
David Adler (2019)	7	48	7	47	1.59	0.979(0.372,2.576)
Theresa L. Byrd (2013)	64	155	38	153	7.2	1.662(1.191,2.321)
Erica Erwin (2019)	35	272	12	281	3.21	3.013(1.598,5.681)
João Firmino-Machad(2019)	310	605	209	615	11.94	1.508(1.318,1.725)
Sarah Huf (2020)	466	1482	384	1453	12.4	1.190(1.062,1.334)
Sarah Huf (2020)	557	1507	540	1568	12.77	1.073(0.976,1.180)
H Kitchener (2018)	1518	5267	1190	4467	13.26	1.082(1.014,1.154)
Ditte S Linde (2020)	84	350	80	339	8.67	1.017(0.778,1.329)
Zaahirah Mohammad (2022)	110	201	65	200	9.43	1.684(1.330,2.132)
S. M. Peitzmeier (2018)	74	220	44	220	7.42	1.682(1.218,2.323)
Rima Marhayu Abdul R(2013)	86	250	47	250	7.71	1.830(1.343,2.493)
Total	3347	10558	2636	9795	100	1.382(1.214,1.574)

Heterogeneity: $Tau^2 = 0.032$; $chi^2 = 61.02$; df = 11($P < 0.00001$); $I^2 = 82.0\%$
Test of overall effect : z = 4.88 P < 0.00001

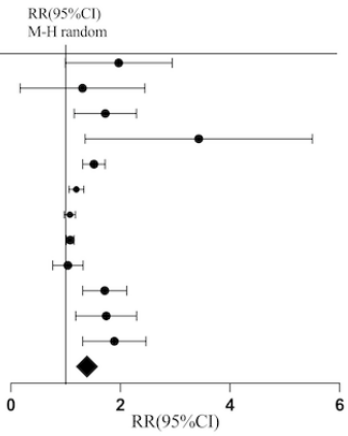

PP of Cervical Cancer Screening

The findings of the PP analysis of cervical cancer screening rates were published in 12 publications [25-33,35,36,38]. According to the findings, the total pooled effect estimate for PP of cervical cancer screening was 1.565 (95% CI 1.381-1.772, I^2=74%; Figure 7 [25-33,35,36,38]). The intervention and comparison groups had an absolute risk difference of 15% (95% CI 10.2-19.8). Comparatively, the intervention group's cancer screening rate was 44% (1549/3540), while the control group's rate was 29% (1474/5132). The sensitivity analysis of the leave-one-out method, demonstrated that excluding any single study from the random-effects model did not result in a loss of significance (Figure S10 in Multimedia Appendix 3). After removing studies with a high RoB (n=5), a sensitivity analysis revealed a pooled RR value of 1.93 (95% CI 1.64-2.27; Figure S11 in Multimedia Appendix 3).

XSL•FO
RenderX

Figure 7. Forest plot comparing electronic health interventions with nonelectronic health interventions in randomized controlled trials for per-protocol analysis of cervical cancer screening participation (n=12). RR: relative risk [25-33,35,36,38].

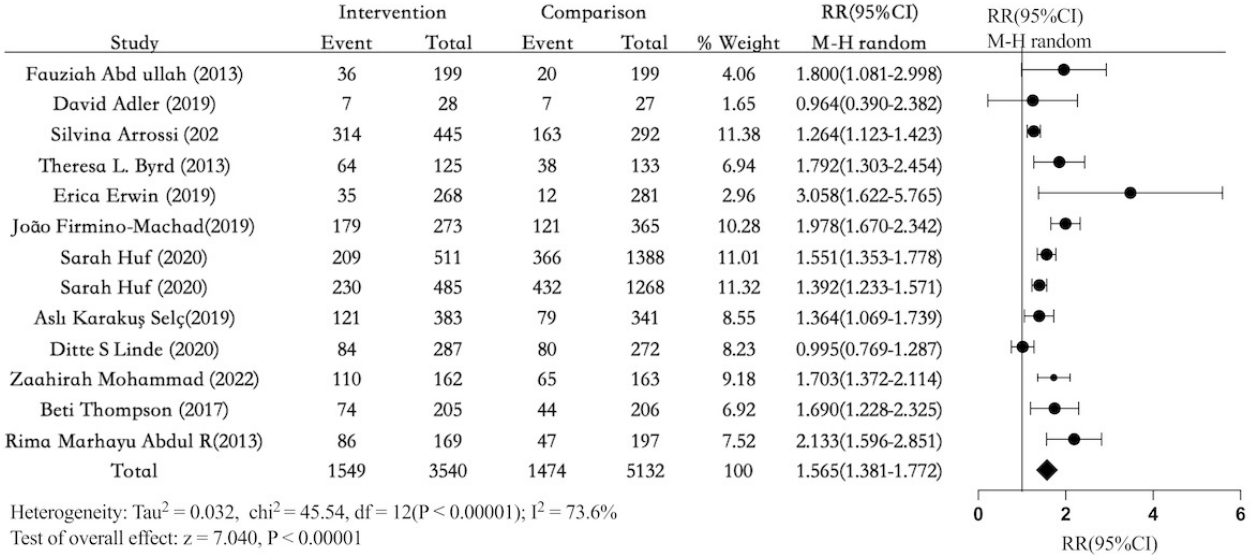

Discussion

Principal Findings

This meta-analysis evaluated the effectiveness of electronic health interventions in increasing cervical cancer screening rates. We found that electronic health interventions can significantly improve cervical cancer screening rates among women. The ITT and PP analyses also showed similar results.

Our systematic review included 14 papers (15 studies) that examined the impact of electronic health interventions on enhancing participation in cervical cancer screening. The overall effect size for cervical cancer screening rates significantly favored the electronic health interventions group (RR 1.464, 95% CI 1.285-1.667). In addition, the subgroup analysis revealed no discernible impact of electronic health interventions on increasing cervical cancer screening rates among women who tested positive for HPV. For individuals who qualified for a PP analysis, the effect size on cervical cancer screening rates was the largest (RR 1.565, 95% CI 1.381-1.772). The sensitivity analyses demonstrated the stability of the result estimates, indicating that electronic health interventions are effective in enhancing cervical cancer screening rates.

Three studies similar to the present study were published before 2021. Uy et al [39] and Ruco et al [12] evaluated the effectiveness of electronic health-related interventions for improving cancer screening rates. These studies included different types of cancer. Uy et al [39] focused on SMS interventions, and the results showed that SMS can improve the screening rates for cervical cancer and breast cancer. The main intervention studied by Ruco et al [12] was social media, and the results showed that mobile health technology, especially social media, can effectively improve cancer screening rates. However, the study by Ruco et al [12] included controlled clinical trials, whereas the present analysis included only RCTs, which may provide a more accurate pooled effect size. Tamuzi et al [40] looked at mobile health interventions for cervical cancer screening and found 17 papers in total, which were then categorized by type of intervention into a meta-analysis. In this study, the intervention is electronic health, which includes SMS, phone calls, multimode interventions, web videos, and internet-based booking. In their study, the intervention was evaluated in the context of mobile health, which included phone calls, letters, and text alerts. According to Tamuzi et al [40], the only intervention with statistically significant estimates of aggregate impacts was phone alerts. However, a meta-analysis of these treatments could not be carried out since only 1 article in their review discussed the effectiveness of SMS alerts. According to the current analysis, 4 out of the 14 publications included discussed how text messaging increased cervical cancer screening rates. The results of this study also suggest that text messaging is effective for improving cervical cancer screening rates (RR 1.41, 95% CI 1.14-1.73). In addition, his study was conducted in 2017, and most of the articles included in this study were published after 2017 (10/14). Our review offers a thorough and updated overview of this subject. In addition, this study analyzes the effect size of electronic health interventions on cervical cancer screening rates using PP and ITT analyses to clarify their effectiveness and extensibility.

The popularity of electronic health is increasing in parallel with technological advances [41]. In a survey of US adults, over 80% of individuals aged 18-49 years old and 73% of those aged 50-64 years old use social media sites [42]. Electronic health interventions include various methods, such as SMS, phone calls, and web videos [43]. The results of this study show that, regardless of the type of intervention, electronic health can improve cervical cancer screening rates. The results of this study show that the aggregate effect size was statistically significant for both ITT and PP analyses, which strongly supports the effectiveness of electronic health in improving cervical cancer screening rates.

When the studies were grouped according to the intervention method, the results showed that each of the intervention modalities (including phone calls, SMS, multimode, web videos,

XSL•FO
RenderX

and internet-based booking) could increase the cervical cancer screening rate, but the combined effect size was highest when phone calls were used as the means of intervention (RR 1.82, 95% CI 1.40-2.38; I^2=0%) [27,33]. Phone calls are a cost-effective intervention [44], because direct contact with the recipient allows timely responses to questions, which further validates the importance of direct communication [45]. However, phone calls intervention is associated with higher manpower costs. The development of voice interaction can save manpower costs while retaining some of the advantages of telephone intervention [46]. Further studies are necessary to understand the impact of voice interaction on cervical cancer screening rates.

When the studies were grouped according to economic level, the subgroup analysis results showed that the aggregate effect size of electronic health on cervical cancer screening rates was higher in low- and middle-income areas than in high-income areas, suggesting that electronic health is more effective in low- and middle-income populations. One possible reason for this result is that resources are scarce in low- and middle-income areas. Electronic health interventions can help the residents of low- and middle-income areas improve their awareness and attention to cervical cancer through internet-based health education, health knowledge dissemination, internet-based reminder services, and other forms, which may increase the motivation for screening [20,47]. In high-income areas, electronic health applications are more common, and electronic health incentives are therefore weaker for the high-income population than for the low-income population [48].

Subgroup analysis based on distinct intervention populations revealed no statistically significant differences in cervical cancer screening rates among women with HPV between the electronic health intervention group and the conventional intervention group. Most of the studies included in the analysis were based on the theoretical health belief model [28,30], and the main strategy was to provide patients with health education related to cervical cancer screening in the form of electronic health [26-28,35-37]. Overall, the screening rate in women with HPV (45%) was higher than the average rate (30.52%), possibly because women with HPV are more aware of their risk of cervical cancer as carriers of high-risk HPV [49,50]. Factors such as physician recommendations, secondary tests, and education also play a role in promoting their participation in cervical cancer screening, resulting in a relatively high screening rate. Because the main limitation to follow-up screening in women with HPV is not cervical cancer awareness, electronic health is not effective in improving screening rates in this population [51]. However, advances in science and technology are increasing the diversity of electronic health methods available, and the intervention mechanism is not limited to reminder education [52]. Further studies are needed to explore the effect of electronic health on cervical cancer screening rates in women with HPV.

The Egger's test results of this study suggest potential publication bias; however, after using the trim and fill method, no additional studies were deemed necessary, and the combined effect size and *P* values remained unchanged, thereby affirming the robustness of these findings. The RoB results indicated that 50% of the studies exhibited high bias risk, prompting a sensitivity analysis excluding these high-risk studies, which enhanced the precision and reliability of the outcomes. After excluding studies with high bias risk, the pooled results were as follows: cervical cancer screening: 1.56 (95% CI 1.21-2.02), cervical cancer screening (ITT): 1.67 (95% CI 1.22-2.28), cervical cancer screening (PP): 1.93 (95% CI 1.64-2.27). Consequently, although studies with high bias risk may reduce some intervention effects, the overall effect size exhibited minimal variation, thereby confirming the robustness of the study results.

Limitations

Despite the consistency of the pooled effect estimate in the subgroup and sensitivity analyses, significant heterogeneity remained in this meta-analysis. The variation in demographics, treatments, or outcome measures among the studies may be the cause of this. For instance, all adults aged 65 years or younger or highly specialized populations like ER patients or individuals with HPV were among the populations randomized in the studies included in this review. Furthermore, we only included papers published in English, which could have an impact on the meta-analysis's findings. Despite our best efforts, we were unable to locate some of the records that the search identified. Finally, because it was unable to identify the outcomes unique to cervical cancer, RCTs that enrolled patients with a variety of malignancies and reported their combined findings were excluded. Furthermore, non-RCT designs (such as adaptive trials) were disregarded.

Conclusion

In conclusion, the results of this study suggest that electronic health interventions may significantly impact increasing cervical cancer screening rates. However, the study exhibited a high degree of heterogeneity. Phone call–based electronic health interventions demonstrated the most substantial improvements in increasing cervical cancer screening rates, with low heterogeneity. Multimodal interventions constitute a growing segment of electronic health treatments. The RoB review highlighted variations in blinding processes, and only a few studies disclosed plans to upscale, assess the cost of the intervention, or maintain the outcomes. Future cost-effectiveness analyses should be conducted to further elucidate the scalability of electronic health interventions.

Acknowledgments

This study was conducted with the support of the Liaoning Provincial Department of Education's 2021 Top-Level Scientific Research Project (LJKR0279). We thank International Science Editing for editing this manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XL contributed to writing the original draft, software, methodology, formal analysis, data curation, and conceptualization. LN handled software, methodology, formal analysis, and data curation. WF managed software, methodology, and data curation. CJ performed methodology and data curation. LG contributed to writing–review & editing, visualization, validation, supervision, resources, project administration, methodology, funding acquisition, and formal analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist for the full text and abstract sections.
[DOCX File , 33 KB - jmir_v26i1e58066_app1.docx]

Multimedia Appendix 2
Search strategies of MEDLINE, Cochrane Central Registry of Controlled Trials, Embase, PsycINFO, PubMed, Scopus, Web of Science, and the Cumulative Index to Nursing and Allied Health Literature.
[DOCX File , 726 KB - jmir_v26i1e58066_app2.docx]

Multimedia Appendix 3
Supplementary material for this study, including general information on included studies, funnel plot results, subgroup analysis results, sensitivity analysis results.
[DOCX File , 2624 KB - jmir_v26i1e58066_app3.docx]

References

1. Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health 2020;8(2):e191-e203 [FREE Full text] [doi: 10.1016/S2214-109X(19)30482-6] [Medline: 31812369]
2. Singh D, Vignat J, Lorenzoni V, Eslahi M, Ginsburg O, Lauby-Secretan B, et al. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO global cervical cancer elimination initiative. Lancet Glob Health 2023;11(2):e197-e206 [FREE Full text] [doi: 10.1016/S2214-109X(22)00501-0] [Medline: 36528031]
3. Das M. WHO launches strategy to accelerate elimination of cervical cancer. Lancet Oncol 2021;22(1):20-21. [doi: 10.1016/S1470-2045(20)30729-4] [Medline: 33248466]
4. Bedell SL, Goldstein LS, Goldstein AR, Goldstein AT. Cervical cancer screening: past, present, and future. Sex Med Rev 2020;8(1):28-37 [FREE Full text] [doi: 10.1016/j.sxmr.2019.09.005] [Medline: 31791846]
5. Smith JL, Wilson KM, Orians CE, Byrd TL. AMIGAS: building a cervical cancer screening intervention for public health practice. J Womens Health (Larchmt) 2013;22(9):718-723 [FREE Full text] [doi: 10.1089/jwh.2013.4467] [Medline: 23930983]
6. Global HIV, Hepatitis and STIs Programmes (HHS), Guidelines Review Committee, Sexual and Reproductive Health and Research. WHO guideline for screening and treatment of cervical pre-cancer lesions for cervical cancer prevention.: World Health Organization; 2001. URL: https://www.who.int/publications/i/item/9789240030824 [accessed 2024-10-03]
7. Alam Z, Cairns JM, Scott M, Dean JA, Janda M. Interventions to increase cervical screening uptake among immigrant women: a systematic review and meta-analysis. PLoS One 2023;18(6):e0281976 [FREE Full text] [doi: 10.1371/journal.pone.0281976] [Medline: 37267330]
8. Musa J, Achenbach CJ, O'Dwyer LC, Evans CT, McHugh M, Hou L, et al. Effect of cervical cancer education and provider recommendation for screening on screening rates: a systematic review and meta-analysis. PLoS One 2017;12(9):e0183924 [FREE Full text] [doi: 10.1371/journal.pone.0183924] [Medline: 28873092]
9. Brevik TB, Laake P, Bjørkly S. Effect of culturally tailored education on attendance at mammography and the papanicolaou test. Health Serv Res 2020;55(3):457-468 [FREE Full text] [doi: 10.1111/1475-6773.13271] [Medline: 31994187]
10. Costa S, Verberckmoes B, Castle PE, Arbyn M. Offering HPV self-sampling kits: an updated meta-analysis of the effectiveness of strategies to increase participation in cervical cancer screening. Br J Cancer 2023;128(5):805-813 [FREE Full text] [doi: 10.1038/s41416-022-02094-w] [Medline: 36517552]
11. Vasankari V, Halonen J, Vasankari T, Anttila V, Airaksinen J, Sievänen H, et al. Physical activity and sedentary behaviour in secondary prevention of coronary artery disease: a review. Am J Prev Cardiol 2021;5:100146 [FREE Full text] [doi: 10.1016/j.ajpc.2021.100146] [Medline: 34327489]

XSL•FO
RenderX

12. Ruco A, Dossa F, Tinmouth J, Llovet D, Jacobson J, Kishibe T, et al. Social media and mHealth technology for cancer screening: systematic review and meta-analysis. J Med Internet Res 2021;23(7):e26759 [FREE Full text] [doi: 10.2196/26759] [Medline: 34328423]
13. Rittberg R, Mann A, Desautels D, Earle CC, Navaratnam S, Pitz M. Canadian cancer centre response to COVID-19 pandemic: a national and provincial response. Curr Oncol 2020;28(1):233-251 [FREE Full text] [doi: 10.3390/curroncol28010026] [Medline: 33704191]
14. Jiang Y, West BT, Barton DL, Harris MR. Acceptance and use of eHealth/mHealth applications for self-management among cancer survivors. Stud Health Technol Inform 2017;245:131-135 [FREE Full text] [Medline: 29295067]
15. Singleton AC, Raeside R, Hyun KK, Partridge SR, Di Tanna GL, Hafiz N, et al. Electronic health interventions for patients with breast cancer: systematic review and meta-analyses. J Clin Oncol 2022;40(20):2257-2270. [doi: 10.1200/JCO.21.01171] [Medline: 35500200]
16. Thompson B, Barrington WE, Briant KJ, Kupay E, Carosso E, Gonzalez NE, et al. Educating Latinas about cervical cancer and HPV: a pilot randomized study. Cancer Causes Control 2019;30(4):375-384 [FREE Full text] [doi: 10.1007/s10552-019-01150-w] [Medline: 30830494]
17. Lyson HC, Le GM, Zhang J, Rivadeneira N, Lyles C, Radcliffe K, et al. Social media as a tool to promote health awareness: results from an online cervical cancer prevention study. J Cancer Educ 2019;34(4):819-822 [FREE Full text] [doi: 10.1007/s13187-018-1379-8] [Medline: 29948924]
18. Vimarlund V, Koch S, Nøhr C. Advances in E-Health. Life (Basel) 2021;11(6):468 [FREE Full text] [doi: 10.3390/life11060468] [Medline: 34073659]
19. Romli R, Abd Rahman R, Chew KT, Mohd Hashim S, Mohamad EMW, Mohammed Nawi A. Empirical investigation of e-health intervention in cervical cancer screening: a systematic literature review. PLoS One 2022;17(8):e0273375 [FREE Full text] [doi: 10.1371/journal.pone.0273375] [Medline: 35984812]
20. Zhang D, Advani S, Waller J, Cupertino A, Hurtado-de-Mendoza A, Chicaiza A, et al. Mobile technologies and cervical cancer screening in low- and middle-income countries: a systematic review. JCO Glob Oncol 2020;6:617-627 [FREE Full text] [doi: 10.1200/JGO.19.00201] [Medline: 32302236]
21. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
22. Mavridis D, Salanti G. How to assess publication bias: funnel plot, trim-and-fill method and selection models. Evid Based Ment Health 2014;17(1):30. [doi: 10.1136/eb-2013-101699] [Medline: 24477535]
23. Chandeying N, Thongseiratch T. Systematic review and meta-analysis comparing educational and reminder digital interventions for promoting HPV vaccination uptake. NPJ Digit Med 2023;6(1):162 [FREE Full text] [doi: 10.1038/s41746-023-00912-w] [Medline: 37644090]
24. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
25. Abdul Rashid RM, Mohamed M, Hamid ZA, Dahlui M. Is the phone call the most effective method for recall in cervical cancer screening?--results from a randomised control trial. Asian Pac J Cancer Prev 2013;14(10):5901-5904 [FREE Full text] [doi: 10.7314/apjcp.2013.14.10.5901] [Medline: 24289597]
26. Abdullah F, Su TT. Applying the transtheoretical model to evaluate the effect of a call-recall program in enhancing pap smear practice: a cluster randomized trial. Prev Med 2013;57 Suppl:S83-S86. [doi: 10.1016/j.ypmed.2013.02.001] [Medline: 23415623]
27. Adler D, Abar B, Wood N, Bonham A. An intervention to increase uptake of cervical cancer screening among emergency department patients: results of a randomized pilot study. J Emerg Med 2019;57(6):836-843 [FREE Full text] [doi: 10.1016/j.jemermed.2019.07.021] [Medline: 31594738]
28. Arrossi S, Paolino M, Antelo VS, Thouyaret L, Kohler RE, Cuberli M, ATICA Study team. Effectiveness of an mHealth intervention to increase adherence to triage of HPV DNA positive women who have performed self-collection (the ATICA study): a hybrid type I cluster randomised effectiveness-implementation trial. Lancet Reg Health Am 2022;9:100199 [FREE Full text] [doi: 10.1016/j.lana.2022.100199] [Medline: 35655914]
29. Byrd TL, Wilson KM, Smith JL, Coronado G, Vernon SW, Fernandez-Esquer ME, et al. AMIGAS: a multicity, multicomponent cervical cancer prevention trial among Mexican American women. Cancer 2013;119(7):1365-1372 [FREE Full text] [doi: 10.1002/cncr.27926] [Medline: 23280399]
30. Erwin E, Aronson KJ, Day A, Ginsburg O, Macheku G, Feksi A, et al. SMS behaviour change communication and eVoucher interventions to increase uptake of cervical cancer screening in the Kilimanjaro and Arusha regions of Tanzania: a randomised, double-blind, controlled trial of effectiveness. BMJ Innov 2019;5(1):28-34 [FREE Full text] [doi: 10.1136/bmjinnov-2018-000276] [Medline: 31645991]
31. Firmino-Machado J, Varela S, Mendes R, Moreira A, Lunet N, SCAN-Cervical Cancer collaborators. A 3-step intervention to improve adherence to cervical cancer screening: the SCAN randomized controlled trial. Prev Med 2019;123:250-261. [doi: 10.1016/j.ypmed.2019.03.025] [Medline: 30936001]

XSL•FO
RenderX

32. Huf S, Kerrison RS, King D, Chadborn T, Richmond A, Cunningham D, et al. Behavioral economics informed message content in text message reminders to improve cervical screening participation: two pragmatic randomized controlled trials. Prev Med 2020;139:106170. [doi: 10.1016/j.ypmed.2020.106170] [Medline: 32610059]
33. Karakuş Selçuk A, Yanikkerem E. The effect of web-based education on pap smear behaviours of teachers. Eur J Cancer Care (Engl) 2020;29(3):e13202. [doi: 10.1111/ecc.13202] [Medline: 31820836]
34. Kitchener H, Gittins M, Cruickshank M, Moseley C, Fletcher S, Albrow R, et al. A cluster randomized trial of strategies to increase uptake amongst young women invited for their first cervical screen: the STRATEGIC trial. J Med Screen 2018;25(2):88-98 [FREE Full text] [doi: 10.1177/0969141317696518] [Medline: 28530513]
35. Linde DS, Andersen MS, Mwaiselage J, Manongi R, Kjaer SK, Rasch V. Effectiveness of one-way text messaging on attendance to follow-up cervical cancer screening among human papillomavirus-positive Tanzanian women (Connected2Care): parallel-group randomized controlled trial. J Med Internet Res 2020;22(4):e15863 [FREE Full text] [doi: 10.2196/15863] [Medline: 32238335]
36. Mohammad Z, Ahmad N, Baharom A. The effects of theory-based educational intervention and WhatsApp follow-up on papanicolaou smear uptake among postnatal women in Malaysia: randomized controlled trial. JMIR Mhealth Uhealth 2022;10(6):e32089 [FREE Full text] [doi: 10.2196/32089] [Medline: 35759319]
37. Peitzmeier SM, Khullar K, Potter J. Effectiveness of four outreach modalities to patients overdue for cervical cancer screening in the primary care setting: a randomized trial. Cancer Causes Control 2016;27(9):1081-1091. [doi: 10.1007/s10552-016-0786-6] [Medline: 27447961]
38. Thompson B, Carosso EA, Jhingan E, Wang L, Holte SE, Byrd TL, et al. Results of a randomized controlled trial to increase cervical cancer screening among rural Latinas. Cancer 2017;123(4):666-674 [FREE Full text] [doi: 10.1002/cncr.30399] [Medline: 27787893]
39. Uy C, Lopez J, Trinh-Shevrin C, Kwon SC, Sherman SE, Liang PS. Text messaging interventions on cancer screening rates: a systematic review. J Med Internet Res 2017;19(8):e296 [FREE Full text] [doi: 10.2196/jmir.7893] [Medline: 28838885]
40. Tamuzi JL, Muyaya LM, Tshimwanga JL, Zeng L. Effectiveness of mhealth to increase cervical cancer screening: systematic review of interventions. Int J Pul & Res Sci 2017;2(3). [doi: 10.19080/ijoprs.2017.02.555586]
41. Gyawali B, Bowman M, Sharpe I, Jalink M, Srivastava S, Wijeratne DT. A systematic review of eHealth technologies for breast cancer supportive care. Cancer Treat Rev 2023;114:102519. [doi: 10.1016/j.ctrv.2023.102519] [Medline: 36736125]
42. Schliemann D, Tan MM, Hoe WMK, Mohan D, Taib NA, Donnelly M, et al. mHealth interventions to improve cancer screening and early detection: scoping review of reviews. J Med Internet Res 2022;24(8):e36316 [FREE Full text] [doi: 10.2196/36316] [Medline: 35969450]
43. Heimer M, Schmitz S, Teschler M, Schäfer H, Douma ER, Habibovic M, et al. eHealth for maintenance cardiovascular rehabilitation: a systematic review and meta-analysis. Eur J Prev Cardiol 2023;30(15):1634-1651. [doi: 10.1093/eurjpc/zwad145] [Medline: 37154363]
44. Dutton GR, Nackers LM, Dubyak PJ, Rushing NC, Huynh TT, Tan F, et al. A randomized trial comparing weight loss treatment delivered in large versus small groups. Int J Behav Nutr Phys Act 2014;11:123 [FREE Full text] [doi: 10.1186/s12966-014-0123-y] [Medline: 25249056]
45. Shimada SL, Zocchi MS, Hogan TP, Kertesz SG, Rotondi AJ, Butler JM, et al. Impact of patient-clinical team secure messaging on communication patterns and patient experience: randomized encouragement design trial. J Med Internet Res 2020;22(11):e22307 [FREE Full text] [doi: 10.2196/22307] [Medline: 33206052]
46. Kim J, Shin E, Han K, Park S, Youn JH, Jin G, et al. Efficacy of smart speaker-based metamemory training in older adults: case-control cohort study. J Med Internet Res 2021;23(2):e20177 [FREE Full text] [doi: 10.2196/20177] [Medline: 33591276]
47. Al-Dhahir I, Reijnders T, Faber JS, van den Berg-Emons RJ, Janssen VR, Kraaijenhagen RA, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022;24(8):e34229 [FREE Full text] [doi: 10.2196/34229] [Medline: 36001380]
48. Marcus JZ, Cason P, Downs LS, Einstein MH, Flowers L. The ASCCP cervical cancer screening task force endorsement and opinion on the American cancer society updated cervical cancer screening guidelines. J Low Genit Tract Dis 2021;25(3):187-191. [doi: 10.1097/LGT.0000000000000614] [Medline: 34138787]
49. Maver PJ, Poljak M. Primary HPV-based cervical cancer screening in Europe: implementation status, challenges, and future plans. Clin Microbiol Infect 2020;26(5):579-583 [FREE Full text] [doi: 10.1016/j.cmi.2019.09.006] [Medline: 31539637]
50. Perkins RB, Wentzensen N, Guido RS, Schiffman M. Cervical cancer screening: a review. JAMA 2023;330(6):547-558. [doi: 10.1001/jama.2023.13174] [Medline: 37552298]
51. Isaacson S, Adewumi K, Smith JS, Novak C, Oketch S, Huchko MJ. A qualitative exploration of barriers to treatment among HPV-positive women in a cervical cancer screening study in western Kenya. Oncologist 2023;28(1):e9-e18 [FREE Full text] [doi: 10.1093/oncolo/oyac208] [Medline: 36239434]
52. Smith B, Magnani JW. New technologies, new disparities: the intersection of electronic health and digital health literacy. Int J Cardiol 2019;292:280-282 [FREE Full text] [doi: 10.1016/j.ijcard.2019.05.066] [Medline: 31171391]

XSL•FO
RenderX

Abbreviations

HPV: human papillomavirus
ITT: intention-to-treat
MINDSCAPE: Messenger, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitments, and Ego framework
PICOS: Population, Intervention, Comparison, Outcome, and Study design
PP: per-protocol
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
RoB: risk of bias
RR: relative risk
WHO: World Health Organization

Edited by A Mavragani; submitted 05.03.24; peer-reviewed by D Linde, Z Alam; comments to author 10.05.24; revised version received 30.05.24; accepted 10.09.24; published 31.10.24.

Please cite as:
Liu X, Ning L, Fan W, Jia C, Ge L
Electronic Health Interventions and Cervical Cancer Screening: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e58066
URL: https://www.jmir.org/2024/1/e58066
doi:10.2196/58066
PMID:

©Xiaoxia Liu, Lianzhen Ning, Wenqi Fan, Chanyi Jia, Lina Ge. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Experiences With mHealth Use Among Patient-Caregiver Dyads With Chronic Heart Failure: Qualitative Study

Xiaorong Jin[1], MSN; Yimei Zhang[1], BSN; Min Zhou[1], BSN; Xiong Zhang[1], MSN; Qian Mei[2], BSN; Yangjuan Bai[3], MSN; Wei Wei[4], MSN; Fang Ma[1], PhD

[1]Department of Nursing, The First Affiliated Hospital of Kunming Medical University, Kunming City, China
[2]Coronary Heart Disease Center, Fuwai Yunnan Cardiovascular Hospital, Kunming City, China
[3]Cardiology Department, The First Affiliated Hospital of Kunming Medical University, Kunming City, China
[4]Digestive Surgery Department, The First Affiliated Hospital of Kunming Medical University, Kunming City, China

Corresponding Author:
Fang Ma, PhD
Department of Nursing
The First Affiliated Hospital of Kunming Medical University
295 Xichang Road
Wuhua District
Kunming City, 650032
China
Phone: 86 65324888
Email: rebeccamalei@126.com

Abstract

Background: Chronic heart failure has become a serious threat to the health of the global population, and self-management is key to treating chronic heart failure. The emergence of mobile health (mHealth) provides new ideas for the self-management of chronic heart failure in which the informal caregiver plays an important role. Current research has mainly studied the experiences with using mHealth among patients with chronic heart failure from the perspective of individual patients, and there is a lack of research from the dichotomous perspective.

Objective: The aim of this study was to explore the experiences with mHealth use among patients with chronic heart failure and their informal caregivers from a dichotomous perspective.

Methods: This descriptive phenomenological study from a post-positivist perspective used a dyadic interview method, and face-to-face semistructured interviews were conducted with patients with chronic heart failure and their informal caregivers. Data were collected and managed using NVivo 12 software, and data analysis used thematic analysis to identify and interpret participants' experiences and perspectives. The thematic analysis included familiarizing ourselves with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report.

Results: A total of 14 dyads of patients with chronic heart failure and their informal caregivers (13 men and 15 women) participated in this study, including 3 couples and 11 parent-child pairs. We constructed 4 key themes and their subthemes related to the experiences with mHealth use: (1) opposing experiences with mHealth as human interaction or trauma (great experience with mHealth use; trauma), (2) supplement instead of replacement (it is useful but better as a reference; offline is unavoidable sometimes), (3) both agreement and disagreement over who should be the adopter of mHealth (achieving consensus regarding who should adopt mHealth; conflict occurs when considering patients as the adopter of mHealth), (4) for better mHealth (applying mHealth with caution; suggestions for improved mHealth).

Conclusions: This study reported that the experiences with mHealth use among patients with chronic heart failure and their informal caregivers were mixed, and it highlighted the human touch of mHealth and the importance of network security. These results featured mHealth as a complement to offline hospitals rather than a replacement. In the context of modern or changing Chinese culture, we encourage patients to use mHealth by themselves and their informal caregivers to provide help when necessary. In addition, we need to use mHealth carefully, and future mHealth designs should focus more on ease of use and be oriented more toward older adults.

(J Med Internet Res 2024;26:e57115) doi:10.2196/57115

XSL•FO
RenderX

KEYWORDS

chronic heart failure; informal caregiver; mHealth; experience; dyad

Introduction

Heart failure is defined as signs or symptoms caused by structural or functional cardiac abnormalities with objective evidence of elevated levels of natriuretic peptides and cardiogenic, pulmonary, or systemic congestion [1]. Chronic heart failure (CHF) is a persistent state of heart failure that can be stable, worsening, or decompensated [2]. It is a serious manifestation of the advanced stage of various cardiovascular diseases, is characterized by high morbidity and mortality, and has become more important as a public problem that threatens the lives and health of the world's population [3]. According to the global heart failure survey, the average incidence rate of CHF is 460/100,000 person-years: 5-year and 10-year survival rates of patients with CHF are only 57% and 35% [4,5], respectively, and about 60% of patients with CHF die within 5 years of diagnosis [6,7]. In addition, a recent multicenter, cross-sectional study suggested that almost one-half of patients with CHF were ≥65 years old [8].

Health Canada defines self-management as "decisions and actions taken by someone who is facing a health problem or issue in order to cope with it and improve his or her health," and self-management of patients with CHF is defined as the ability to assume responsibility for managing one or more aspects of CHF (eg, symptom monitoring, weight monitoring, medication dosage adjustment, or medication decision-making) [9]. Studies have shown that self-management is significant for patients with CHF; it can effectively improve cardiac function and prognosis and reduce mortality for patients with CHF [9,10]. With the promotion and popularization of smartphones and wearable devices, mobile health (mHealth) has emerged as a new option to support self-management of CHF [11,12]. mHealth is a general term for the use of mobile phones and other wireless technologies in medical practice [13], with common functions including booking of appointments, health information inquiries, and vital sign monitoring. mHealth acts as an extension of the clinical walls and enables a patient-to-provider link regardless of proximity to care, which helps meet the long-term health management needs of patients with CHF, and is widely used in real-time health monitoring, active intervention, and other aspects of care for patients with CHF. Patients with CHF can rapidly contact a provider using mHealth, which can save valuable time in the treatment of such a critical and chronic disease [14,15]. A systematic review showed that mHealth can enhance self-care and medication adherence, increase quality of life, reduce depression and anxiety, improve patient satisfaction, increase cardiac function, and reduce hospitalization rates and hospitalization costs of patients with CHF [15]. These outcomes indicate that mHealth is effective in improving the self-management of patients with CHF, thereby diminishing the burden of care [16,17].

Despite the many benefits of mHealth, there is inconsistent evidence regarding outcomes related to mHealth for managing CHF [14], and there are some problems with mHealth in CHF situations that challenge the practice of mHealth. For example, mHealth, such as remote telemedical management, did not reduce all-cause mortality in ambulatory patients with CHF [18]. In addition, among patients recently hospitalized for CHF, telemonitoring did not improve outcomes [19]. Additionally, some patients have expressed difficulty with downloading electronic health-related equipment and procedures and with practicing mHealth due to the complexity; this increased their anxiety about its use and led to patient dissatisfaction [20-22]. Furthermore, mHealth may not be suitable for all patients with CHF, such as older adult patients with low electronic health literacy. In addition, the limited accessibility of electronic devices, including the fees associated with some services, can affect the practice of mHealth [15]. Hence, patients might not accept mHealth due to these problems, leading to low mHealth use among patients with CHF [23,24]. A systematic review reported that more than 20% of patients with CHF had never used mHealth and 60% of patients with CHF no longer used mHealth after a single experience [25]. In order to increase mHealth adoption among patients with CHF, it is necessary to understand their perceptions of and experiences with mHealth use. Previous studies have stated that patients with CHF are satisfied with the use of mHealth for self-management [26,27]. However, studies have also proposed that patients with CHF experience difficulties using mHealth for self-management and are dissatisfied with it [20-22]. These inconsistent results suggest that this area needs further exploration. Frequent medical appointments are necessary for patients with CHF [28], and when they are unable to use mHealth independently, informal caregivers (ICs) [29] (individuals who provide unpaid care and assistance to friends or family members because of a health condition) play an important role [28,30]. According to the study by El-Dassouki et al [31], IC guidance not only enables patients with CHF to gradually overcome the digital divide but also greatly improves the experience of mHealth use for patients with CHF. Therefore, studies on mHealth experiences only from the perspectives of individual patients with CHF may lead to certain limitations [32,33]. Besides, according to the extension of the Technology Acceptance Model proposed by Davis et al [34], patients might be more likely to use services recommended by people who influence their health decisions, such as ICs [35]. Therefore, the purpose of this study was to explore the experience with mHealth use among CHF patients and their ICs from a dichotomous perspective, which can enrich and improve the data and provide a theoretical basis for future intervention, thereby improving mHealth use.

Methods

Study Design and Setting

A descriptive phenomenological study was performed from a post-positivist point of view to explore the dyadic experiences with mHealth use among patients with CHF and their ICs. Descriptive phenomenology advances human understanding by revealing the nature and organized structure of phenomena without imposing preconceived notions that may influence the understanding of the experiences being examined [36]. This

XSL•FO
RenderX

approach was chosen due to its emphasis on exploring the essence of the phenomenon as it is lived by the participants, aligning with the study's focus on understanding the subjective lived experiences of mHealth use among patients with CHF and their ICs [37]. The Standards for Reporting Qualitative Research (SRQR) checklist informed the development, analysis, and reporting of this study [38].

We conducted dyadic interviews, which were defined as semistructured interviews with 2 family members (ie, the patient with CHF and his or her IC) conducted by 1 researcher [39]. Unlike other interview methods, dyadic interviews can obtain the perspectives of patients and ICs at the same time, which helps the researchers to obtain more comprehensive data [40]. Second, dyadic interviews can demonstrate the interaction between the patient and IC, helping the researcher to explore complex perspectives, whether complementary or contradictory [41-43]. In addition, dyadic interviews allow participants to spark ideas that may not have been recognized nor remembered, thus enabling both participants to respond to and build on each other's contributions in their interactions during the interview process [39-41,44].

This study was conducted in the Department of Cardiology at a comprehensive tertiary hospital in Yunnan Province from October 2023 to December 2023. This hospital is located in the provincial capital of Yunnan province and provides services to a wide geographical area of the province and some patients from neighboring provinces. A variety of patients with different backgrounds come to this hospital for its reputation; this patient population might lead to diverse information and rich data about the topic to be studied. A purposive sampling method was used to recruit participants of different ages, genders, occupations, employment status, and types of mHealth resources the patients and ICs were using.

Participant Recruitment and Selection Process

This study used a purposive sampling method to select patients with CHF who were hospitalized in the Department of Cardiology in a tertiary hospital in Yunnan Province and their ICs from October 2023 to December 2023. The inclusion criteria were (1) 18 years or older for both patients with CHF and ICs, (2) patients were diagnosed with CHF and had at least one IC, (3) at least one participant in a dyad had used mHealth for the management of CHF, and (4) voluntary participation. The exclusion criteria were patients and caregivers (1) with other serious diseases (eg, malignant tumor, severe organ failure), (2) with hearing or speech impairments, or (3) who were also participating in other research projects.

Data Collection Process

Face-to-face interviews were conducted between the researchers and CHF patient-IC dyad in hospital wards from October 2023 to December 2023. All interviews were conducted by 2 researchers (XJ and YZ), with XJ acting as the primary interviewer and YZ taking field notes. They were current master's candidate students at a medical school with 1 year to 2 years of experience in a cardiology internship and had received training in qualitative interviewing. The participants and researchers did not know each other, and the interview was conducted only once. With informed consent obtained from the participants, the interviews were audio-recorded. Data collection was based on information saturation, and the absence of new themes emerging determined the number of interviews, resulting in 14 dyadic interviews. The interview outlines mainly included the following: (1) perceptions and experiences of patients with CHF and their ICs concerning mHealth use and (2) expectations of patients with CHF and their ICs about the development of mHealth. The full interview outline is detailed in Multimedia Appendix 1. Interviews lasted between 21 minutes and 42 minutes. Within 24 hours after the interviews, the researcher (XJ) transcribed the audio recordings of the interviews verbatim into textual material. To ensure the accuracy of the content, we asked another researcher (YZ) to re-listen to the recordings and check and verify the transcribed text.

Data Analysis

Based on Braun and Clarke [45], thematic analysis was carried out to identify and interpret participants' experiences and perspectives; the thematic analysis included familiarizing ourselves with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report [46]. We chose thematic analysis for its flexibility and because it allowed us to explore the depth of our data and better understand the commonalities of participants' discussions and responses. Furthermore, we hoped to generate unanticipated insights using thematic analysis because it allows for social and psychological interpretations of data. Thematic analysis is a family of methods, not a singular method, that includes coding reliability, codebook, reflexive thematic analysis, and thematic coding [47]. We used reflexive thematic analysis in our study, which embraces researcher subjectivity as a resource for research and emphasizes researcher reflexivity. The collated data were imported into the NVivo 12 software [48] for management and analysis. In the initial stages of the analysis, the data were read multiple times to gain insight into its content. Next, the researchers (XJ and FM) each performed independent preliminary coding of the data. After the coding was completed, the researchers compared and discussed the data in order to reach consensus on the final coding. When disagreements arose during the process, YZ intervened to resolve the issue through discussion. After the coding was finalized, we started looking for candidate themes. This process was done by recombining and grouping related codes under the same theme until data saturation was achieved. We defined and named each theme in detail to better understand what each represented. In the final stage of the analysis, the research team met to discuss and refine the existing themes in depth to ensure the accuracy and comprehensiveness of the findings. In the data collection and analysis process, we integrated several strategies to maintain reflexivity. We kept a journal of field notes and a record of the interviewer's reasoning, judgment, emotional reaction to the interview, and how this influenced the interview. These notes were primarily made after the interviews, and the interviewer recalled as much detail about the interviews as possible. In the data analysis process, we examined previous reflexive journals (field notes and records) from the data collection to provide more context and made additional notes on the transcript to aid understanding [49].

XSL•FO
RenderX

Importantly, all participants were identified by serial numbers (eg P1, P2, where "P" stands for patient, and IC1, IC2, where "IC" stands for informal caregiver). All personally identifiable information was removed from the transcripts. Interview records were stored in a password-secured computer file, and only the identified researcher could access these records. To achieve and ensure the methodological rigor of the study, the 4 criteria of credibility, dependability, transferability, and confirmability were used. To ensure credibility, long-term engagement of the researcher with the data, reflexivity, and peer debriefing were performed. Dependability was ensured by providing a detailed description of the research methods and having peers participate in the analysis process. Transferability was achieved by providing in-depth and rich descriptions of the demographics and geographic boundaries of the study. Confirmability was established using the researchers' field notes regarding personal feelings, biases, and insights immediately following the individual interviews [50]. Peer debriefing was performed to ensure credibility. Moreover, the disagreements in the research were discussed and resolved through frequent meetings.

Ethics Approval

The study was approved by the Ethics Committee of the First Affiliated Hospital of Kunming Medical University (2022-L-304). For all the participants, the study was fully explained orally and in written form, and their written informed consent was obtained before the data collection. Participation in the study was entirely voluntary, and no compensation was provided for their involvement. Participant data were deidentified and stored on a local secure server. Further, the participants had the right to withdraw from the study at any time, without giving any reasons. In addition, no identification of individual participants or users in any images occurred in the manuscript or supplementary material.

Results

Participants

A total of 14 patients with CHF and their ICs were interviewed in this study, including 3 couples and 11 parent-child dyads. The patients consisted of 9 men and 5 women, with one-half (7/14, 50%) of the patients older than 65 years old, which was similar to the demographics in other research [8]. The ICs were mostly female (10/14, 71%), with an age range of 22 years to 55 years. In terms of the highest level of education, the majority of the patients (9/14, 64%) had primary or junior high school diplomas, while 79% (11/14) of the ICs had obtained a diploma or higher degree. As for the type of mHealth resources the patients and ICs were using, patients used telemedicine platforms more frequently (7/14, 50%), while caregivers were more likely to use mHealth applications (13/14, 93%). In addition, we also collected employment status and economic situations, and the demographic information is detailed in Table 1.

Our qualitative analyses produced 4 salient themes: (1) opposing experiences with mHealth as human interaction or trauma, (2) supplement instead of replacement, (3) both agreement and disagreement over who should be the adopter of mHealth, (4) for better mHealth.

XSL•FO
RenderX

Table 1. Demographic characteristics of participants (N=28).

Variables	Patients (n=14), n (%)	Caregivers (n=14), n (%)
Age (years)		
18-45	0	8 (57)
46-65	7 (50)	6 (43)
66-75	5 (36)	0
≥76	2 (14)	0
Gender		
Male	9 (64)	4 (29)
Female	5 (36)	10 (71)
Relationship with the patient		
Parents	—[a]	0
Children	—	11 (79)
Couple	—	3 (21)
Highest level of education		
Primary or junior high school diploma	9 (64)	2 (14)
High school diploma	1 (7)	1 (7)
College trade or technical diploma	2 (14)	4 (29)
Bachelor's degree or higher	2 (14)	7 (50)
Employment status		
Retired	4 (29)	1 (7)
Unemployed	0	2 (14)
Working full-time or part-time	4 (29)	10 (71)
Other	6 (43)	1 (7)
Monthly income per capita (¥)[b]		
<2000	1 (7)	1 (7)
2000-4999	4 (29)	4 (29)
≥5000	9 (64)	9 (64)
Type of mHealth resources the patients and ICs[c] were using		
mHealth application	5 (36)	13 (93)
Wearable device	5 (36)	7 (50)
Telemedicine platform	7 (50)	11 (79)

[a]Not applicable.

[b]A currency exchange rate of ¥1=US $0.14 is applicable.

[c]ICs: informal caregivers.

Theme 1: Opposing Experiences With mHealth as Human Interaction or Trauma

Participants' experiences with mHealth were polarized, with some saying that using mHealth was a good experience, but others thought that it was an unpleasant experience for them. In this theme, there were 2 subthemes: (1) great experience with mHealth use and (2) trauma.

Subtheme 1: Great Experience With mHealth Use

Some patients and ICs reflected that mHealth was convenient because they could get expert suggestions in an easy way. In addition, they considered that online doctors were patient and they answered their questions one by one, no matter how many questions they had, which led to a sense of comfort and humanity compared with their experiences with offline doctors.

> *Compared with offline appointments, doctors' attitudes were much better during the online visit,*

XSL•FO
RenderX

and the responses were quite fast. It feels good. [IC 6]

The telemedicine platform sent messages to me periodically, which included health education information and encouragement words, which was a good experience. [IC 4]

We are able to book appointments online with leading specialists, find better doctors, and get specialized advice; that is more convenient. [IC 5]

Subtheme 2: Trauma

However, some participants recalled an experience of being scammed while using mHealth, which was a negative experience for them. These participants stated that, although a long time had passed, they were scared whenever they thought of it, and it was traumatic for them to some extent.

When I surfed online someday, I found a link to a well-known medical expert, which was the one that I wanted to visit but couldn't because of his busy appointments. Then I clicked the link, filled out the information, and finished the payment with the hope that I could see the doctor in this easy way. I took the payment voucher to the hospital only to know that this was a fake, I was cheated tens of Yuan! That is a terrible experience. [IC 2]

Theme 2: Supplement Instead of Replacement

Most participants stated that, despite the many advantages of mHealth such as convenience and speed that were the result of the digital age, it was constrained by its form and the authenticity and accuracy of the information. They thought that mHealth could not completely replace traditional health-seeking behaviors and it was a supplement to offline hospitals. In this theme, there were 2 subthemes: (1) it is useful but better as a reference and (2) offline is unavoidable sometimes.

Subtheme 1: It is Useful but Better as a Reference

In terms of content, patients and ICs suggested that mHealth could provide some information about their illnesses and medications, but they were troubled by the inconsistent advice that emerged from their inquiries. Hence, mHealth was better used as reference in their point of view.

It's useful to know what's wrong with your body online, but sometimes I don't know how to identify the right information, and it might be better to be used as a reference. [P 7]

Subtheme 2: Offline is Unavoidable Sometimes

All participants reflected that mHealth had some limitations and restrictions. Sometimes, they felt the information that mHealth could provide was vague, which did not really solve their problem. Furthermore, the online delivery of tests and treatments was impossible.

During an online consultation, the doctor gave you a general idea; in the end, you have to go to the hospital for examination and treatment. [IC 4]

Theme 3: Both Agreement and Disagreement Over Who Should be the Adopter of mHealth

In the dyad data, we found that patients' and ICs' views on who should adopt mHealth demonstrated both consensus and conflict. In this theme, there were 2 subthemes: (1) achieving consensus regarding who should adopt mHealth and (2) conflict occurs when considering patients as the adopter of mHealth.

Subtheme 1: Achieving Consensus About Who Should Adopt mHealth

In this subtheme, patients and ICs agreed on who should use mHealth. For the dyads reaching a consensus, a minority of dyads agreed that patients should use mHealth by themselves. In this situation, patients stated that they used mHealth by themselves, and their ICs agreed with them, considering that patients should have the ability to use mHealth for health care and disease management. However, most dyads who reached consensus agreed that ICs should use mHealth for the patients. The patients reflected that they were not capable of using mHealth independently due to age, illness, and other factors and ICs should do it all, which was also supported by their ICs, who held the belief that patients were incapable of using mHealth and they should use mHealth for their patients.

I can make an appointment to see the doctor and buy medicines on the mobile phone...I can solve the problems by myself, so I don't trouble my children. I suppose that can reduce the burden on my children. [P 5]

I support my dad with using mHealth for registration and health information; he has no problem with it. I am not always able to be with my father, and I hope that he has the ability to use mHealth. [IC 5]

I'm old. I can't see well, I can't read much, I can't type, and it's too difficult to use (mHealth). [*P* 14]

My father is getting older...he has heart failure, and I will help him as much as I can. It may be impractical for him to use it (mHealth) on his own; after all, he is too old. [IC 14]

Subtheme 2: Conflict Occurs When Considering Patients as the Adopter of mHealth

Within this subtheme, there was conflict between patients and ICs regarding users of mHealth. Some patients thought of themselves as the adopter of mHealth because they believed in their ability to do it, whereas their ICs didn't support patients' use of mHealth, as they held the belief that patients were vulnerable and could be easily cheated or it was hard for patients to express themselves online. In addition, a few patients reflected their unwillingness to use mHealth because of their vulnerability, yet their ICs insisted that the patients should do their jobs.

I think these things (mHealth) are simple, no big deal, I can master them myself, but my daughter won't let me use them myself...It's my disease. I want to manage it myself. [P 9]

I will not recommend my father to use mHealth; there are some risks. For example, the elderly are likely to consult Baidu (an encyclopedia website in China),

XSL•FO
RenderX

which is one-sided and often scares you. I don't want my father to search Baidu; the information from Baidu is not completely correct. To make things worse, the information from Baidu adds a heavy psychological burden to my father. [IC 9]

I can't; I already suffer from heart failure, my memory is failing, I'm little educated, and I don't think about these things (mHealth), I just think about living a few more years, and that's just fine. [P 2]

We all have our things now, that is, it is unrealistic to accompany the elderly for a long period of time. He (my father) should manage his condition, including using mHealth. [IC 2]

Theme 4: For Better mHealth

Participants reflected that they had applied some skills in using mHealth, which could be more convenient. In addition, they also made some suggestions for the future development of mHealth. In this theme, 2 subthemes were generated: (1) applying mHealth with caution and (2) suggestions for improved mHealth.

Subtheme 1: Applying mHealth With Caution

Patients and ICs mentioned that they could ensure the accuracy and authenticity of the information by using official channels, multiple queries, and double-checking strategies when using mHealth.

I'll read much more information about my question on the internet, and finding the official website is necessary. Anyway, I would not trust the information if it had a contradiction. If possible, I will confirm with some experts. [IC 1]

Subtheme 2: Suggestions for Improved mHealth

Patients and ICs offered several suggestions in terms of improving the mHealth experience, including stricter regulations, more official channels, simpler steps, and fee deductions for online consultation services.

For these online environments, published content should be more strictly regulated. [*P* 10]

It is best to have a way to allow me to identify the true and false, such as official channels. [P 7]

The (mHealth) easier the better, with no barriers is the best. I wish my father could learn to operate it for its simplicity. [IC 2]

The expert fee (Online Clinic Service) is too expensive. I asked her four or five questions for five or six hundred dollars, and this price is too high. If it can be cheaper, it would be better. [IC 10]

Discussion

Principal Findings

The results of current reports on mHealth experiences are mixed [13,51], which is consistent with our findings. Our findings are supported by those by Agnihothri et al [13] and Hilty et al [52]: mHealth can provide convenient care to patients and improve the efficiency of medical visits. However, some studies have shown that online medical visits lack the essential element of a doctor-patient relationship, which constitutes the human touch [53,54]. Our findings suggest that mHealth can provide users with a more comfortable health care experience and has characteristics of humanity as well, which is a novel finding. This might be explained by the theory of bureaucratic caring [55,56], which suggests that, under pressure to perform work with maximal efficiency in a minimal amount of time, doctors might pay more attention to enhancing the care of technological dimensions such as disease diagnosis and treatment, especially in the midst of numerous patients and high patient demands [57]. In contrast, online doctors have more time and less pressure due to the virtual environment of the work, which highlights the advantage of mHealth. In addition to the aforementioned advantages, we identified some concerns with mHealth. Previous studies have suggested that mHealth suffers from a lack of standardized diagnostic criteria and unregulated and invalidated processes or technologies, which significantly increase the risk of telecommunication fraud [13]. In this study, participants described their experiences of being scammed, and these ultimately led to an unpleasant experience, which highlights the importance of network security.

Previous studies have noted that, during the process of searching for health information, users usually retrieve a large amount of information of varying quality [58], which makes it difficult for users to discern the authenticity of the information, and this is why the results of this study support the idea that mHealth information is just for reference. Not only that, the limitations of the mHealth format lead users to obtain incomplete data through online counseling [59], and it is challenging to perform physical examinations and treatments online [60-62], making it difficult for doctors to make the correct diagnosis, which ultimately leads to incorrect referrals and poor outcomes [54]. The aforementioned reasons have led to patients trusting face-to-face medical consultations more, especially patients with CHF who often experience a life-threatening deterioration of their condition. Patients with CHF need medication or surgical treatments that mHealth cannot provide, making it necessary for them to go to the hospital to solve their problems.

In our study, we attempted to highlight the added value of using the dyad as the unit of analysis while collecting data from individual participants in qualitative research. While contrasting and overlapping 2 individual versions, we captured a third dyadic relationship (created by the researchers) without losing or corrupting the individual versions, and the dyadic perspective added dimension to our understanding [39]. First, using overlapping, we constructed a subtheme that patients with CHF and their ICs reached an agreement as to who should be the adopter of mHealth. The consensus was reached in 2 contexts: One was that patients and their ICs considered that patients should use mHealth, and the other was that patients and their ICs agreed that it was the ICs' responsibility to use mHealth for patients. The use of mHealth by patients highlights guaranteed patient empowerment, which is viewed as a key factor for improving health outcomes and bringing about better adherence with treatment regimens [63]. In addition, when patients are empowered, they are more likely to take an active

XSL•FO
RenderX

role in medical encounters, have a better understanding of different treatment options, and participate more in shared decision-making [64]. Furthermore, in Chinese culture, Confucius dictated that the senior partner owes the strong duties of care and benevolence to the junior partner. To avoid troubling their children in the parent-child pairs, parents might choose to use mHealth by themselves. In other cultural backgrounds, older patients with heart failure had high confidence for using telemonitoring devices, demonstrating that older adult patients with heart failure are ready to embrace mHealth management programs [65]. However, our study also suggested that it was the caregivers' responsibility to use mHealth, which is common in the Chinese cultural context. The Chinese are heavily influenced by Confucianism, in which Five Cardinal Relationships exist, requiring that, in the parent-child dyad, the junior partner (child) owes a strong duty of service and reverence to the senior partner (parent). As in this study, ICs in the parent-child pairs, instead of the patients, used mHealth to help them manage their illnesses and, in this way, showed their caring for family members. Similarly, in the parent-child dyad, patients also considered the responsibility of using mHealth should be put on their caregivers to show respect for their parents and loyalty to the family. In contrast, western countries have more robust social security and welfare systems that prioritize social pensions; hence, the caregivers' responsibilities are less pronounced in western societies [66].

Second, by studying the contrasts between the 2 individual versions, conflicts occurred. Studies have pointed out that a high prevalence of dyadic incongruence occurs in heart failure dyads, with disagreements on illness management that were not resolved, leading to poorer adherence to medical recommendations, worsening CHF symptoms, and damage to the dyadic health [67,68]. In our study, the incongruence was demonstrated as either the patients considered themselves as the adopter of mHealth but their ICs disagreed with the patients' independent use or the patients were unwilling to use mHealth but their ICs wanted the patients to use it. The former is associated with ICs' concerns about the potential risks of mHealth use by the patient. Previous studies have noted negative stereotypes of older adults as unmotivated, sickly, stubborn, and gullible [69,70]. Hence, caregivers, especially children in the parent-child dyads, refused to allow patients to use mHealth because they believed that patients would be deceived or make mistakes in the process. However, patients wanted to be the first person to be responsible for their health, which led to the conflict. The latter may be related to the Chinese culture becoming more individualistic [71]. Previous research has shown that, in individualistic cultures, family structures tend to be freer and looser than in collectivistic cultures, which leads to people living separately from other family members as well as smaller families [71,72]. In this study, there were instances where ICs and patients did not live together, and they expected patients to be able to use mHealth on their own when experiencing illness-related problems. This might be explained by the fact that the children in the parent-child dyads might hold the individualism aspect and prefer equity; they are much less concerned with relationships and take a flexible view toward social obligations, which might be due to the fact that filial piety and family obligations have to compete for resources with the demands of a modern society and the needs for personal achievement of family members [73]. However, older adult patients may refuse to use mHealth due to sensory impairment, cognitive decline, and memory loss [31], which leads to conflict as to who should use mHealth. When managing a chronic disease, people are increasingly seen as coproducers of their health and need to be empowered to take control of the determinants of their health [63]. According to systematic reviews, self-management support interventions are associated with improved outcomes among people with comorbid diabetes and chronic kidney disease, and self-management improves type 2 diabetes treatment by helping people stay healthy and adapt to their illness, highlighting the key role of self-management in chronic disease management [74,75]. Therefore, patients should be encouraged to use mHealth by themselves, and their ICs should help them if needed. Furthermore, with the juxtaposition of tradition and modernization and globalization, one might expect to see gradual shifts from interdependence to independence in Chinese culture, particularly in the younger generations [73]. Hence, self-reliance is advocated, especially in the management of chronic disease, which emphasizes that patients should use mHealth, with or without the support of their caregivers.

To achieve better effectiveness of mHealth, participants summarized several tips, including repeated queries, multiple comparisons, and the use of official channels. This is consistent with the study by Jia et al [76], which showed that more than 60% of health information consumers searched at least 3 different websites to locate and track health information and that authoritative and official channels were more trusted by consumers. We suggest that future mHealth designs should focus more on ease of use and user-friendliness for older adults, which is also supported by the findings of Zhao et al [58]. Consistent with the findings of Dang et al [77], we would like to see lower fees for mHealth services in the future. In addition, participants would like to have stricter regulations and auditing to help them access more accurate health care resources.

Limitations

This qualitative study has obvious strengths but also limitations. It compensates for the lack of existing data on the experiences with mHealth use by conducting dyadic interviews with patients with CHF and their ICs. However, only patients and 1 of their ICs were interviewed, and other family members' perspectives were ignored. This study was conducted in the context of CHF, so the results may not be easily generalizable to other illnesses. Future studies could be conducted more deeply on a family basis and in different diseases. In addition, this study involved only a small group of people with CHF and their caregivers in 1 hospital of Yunnan Province, China, which reduces the generalizability of this study, and caution should be taken when reading the results.

Conclusions

Our study reported that experiences with using mHealth varied between patients with CHF and their ICs. The fact that mHealth involves the human touch is emphasized in our study. Although we acknowledge the convenience and efficiency of mHealth, we also suggest that users should be aware of network security

XSL•FO
RenderX

while using it. In addition, mHealth should complement rather than replace offline hospitals, which are more practical for solving problems. Meanwhile, in the context of the modern or changing Chinese culture, we highlight that patients should use mHealth and ICs should only provide help and guidance when necessary. In addition, our study proposes that mHealth use should be cautious, and the design of mHealth should be more tightly regulated, more affordable, simpler, and more friendly to meet the needs of older adults in the future.

Acknowledgments

The authors would like to express their heartfelt gratitude to all the interviewees who participated in this study and the members of the research team.

This study was supported by Yunnan Provincial Department of Education Science Research Fund Project, China (number 2023Y0749).

Data Availability

The full interview transcripts and audio files are not available to the public due to data protection reasons. However, transcripts and findings are available on request from the corresponding author.

Authors' Contributions

XJ contributed to the study design, data collection and analysis, and writing the manuscript. YZ contributed to the data collection and data analysis. MZ and XZ contributed to the data analysis. QM, YB, and WW contributed to the review of the manuscript. FM contributed to the study design, data analysis, manuscript editing, and manuscript review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 14 KB - jmir_v26i1e57115_app1.docx]

References

1. Heidenreich P, Bozkurt B, Aguilar D, Allen L, Byun J, Colvin M, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2022 May 03;145(18):E. [doi: 10.1161/cir.0000000000001063]
2. Greene SJ, Bauersachs J, Brugts JJ, Ezekowitz JA, Lam CS, Lund LH, et al. Worsening heart failure: nomenclature, epidemiology, and future directions: JACC review topic of the week. J Am Coll Cardiol 2023 Jan 31;81(4):413-424 [FREE Full text] [doi: 10.1016/j.jacc.2022.11.023] [Medline: 36697141]
3. Arrigo M, Jessup M, Mullens W, Reza N, Shah AM, Sliwa K, et al. Acute heart failure. Nat Rev Dis Primers 2020 Mar 05;6(1):16 [FREE Full text] [doi: 10.1038/s41572-020-0151-7] [Medline: 32139695]
4. Emmons-Bell S, Johnson C, Roth G. Prevalence, incidence and survival of heart failure: a systematic review. Heart 2022 Aug 11;108(17):1351-1360 [FREE Full text] [doi: 10.1136/heartjnl-2021-320131] [Medline: 35042750]
5. Kitsiou S, Vatani H, Paré G, Gerber BS, Buchholz SW, Kansal MM, et al. Effectiveness of mobile health technology interventions for patients with heart failure: systematic review and meta-analysis. Can J Cardiol 2021 Aug;37(8):1248-1259. [doi: 10.1016/j.cjca.2021.02.015] [Medline: 33667616]
6. Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, American Heart Association Advocacy Coordinating Committee, Council on Arteriosclerosis, Thrombosis and Vascular Biology, Council on Cardiovascular Radiology and Intervention, Council on Clinical Cardiology, Council on Epidemiology and Prevention, Stroke Council. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail 2013 May;6(3):606-619 [FREE Full text] [doi: 10.1161/HHF.0b013e318291329a] [Medline: 23616602]
7. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics—2016 update. Circulation 2016 Jan 26;133(4):e38-e360. [doi: 10.1161/cir.0000000000000350]
8. Wang Y, Yang Q, Liu Y, Zheng G, Fan F, Tian H, et al. Psychosomatic mechanisms of heart failure symptoms on quality of life in patients with chronic heart failure: A multi-centre cross-sectional study. J Clin Nurs 2024 May 03;33(5):1839-1848. [doi: 10.1111/jocn.16955] [Medline: 38044710]
9. Jovicic A, Holroyd-Leduc JM, Straus SE. Effects of self-management intervention on health outcomes of patients with heart failure: a systematic review of randomized controlled trials. BMC Cardiovasc Disord 2006 Nov 02;6:43 [FREE Full text] [doi: 10.1186/1471-2261-6-43] [Medline: 17081306]

XSL•FO
RenderX

10. Apriningsih A. Health promotion in health care – vital theories and research. Health Communication 2022 Sep 30;38(4):852-852. [doi: 10.1007/978-3-030-63135-2] [Medline: 36315659]
11. Kitsiou S, Paré G, Jaana M. Effects of home telemonitoring interventions on patients with chronic heart failure: an overview of systematic reviews. J Med Internet Res 2015 Mar 12;17(3):e63 [FREE Full text] [doi: 10.2196/jmir.4174] [Medline: 25768664]
12. Leigh JW, Gerber BS, Gans CP, Kansal MM, Kitsiou S. Smartphone ownership and interest in mobile health technologies for self-care among patients with chronic heart failure: cross-sectional survey study. JMIR Cardio 2022 Jan 14;6(1):e31982 [FREE Full text] [doi: 10.2196/31982] [Medline: 35029533]
13. Agnihothri S, Cui L, Delasay M, Rajan B. The value of mHealth for managing chronic conditions. Health Care Manag Sci 2020 Jun;23(2):185-202. [doi: 10.1007/s10729-018-9458-2] [Medline: 30382448]
14. Kruse CS, Soma M, Pulluri D, Nemali NT, Brooks M. The effectiveness of telemedicine in the management of chronic heart disease - a systematic review. JRSM Open 2017 Mar;8(3):2054270416681747 [FREE Full text] [doi: 10.1177/2054270416681747] [Medline: 28321319]
15. Zhang N, Li Q, Chen S, Wu Y, Xin B, Wan Q, et al. Effectiveness of nurse-led electronic health interventions on illness management in patients with chronic heart failure: A systematic review and meta-analysis. Int J Nurs Stud 2024 Feb;150:104630. [doi: 10.1016/j.ijnurstu.2023.104630] [Medline: 38029453]
16. Piette JD, Striplin D, Marinec N, Chen J, Aikens JE. A randomized trial of mobile health support for heart failure patients and their informal caregivers: impacts on caregiver-reported outcomes. Med Care 2015 Aug;53(8):692-699 [FREE Full text] [doi: 10.1097/MLR.0000000000000378] [Medline: 26125415]
17. Piette JD, Rosland AM, Silveira M, Kabeto M, Langa KM. The case for involving adult children outside of the household in the self-management support of older adults with chronic illnesses. Chronic Illn 2010 Mar 22;6(1):34-45 [FREE Full text] [doi: 10.1177/1742395309347804] [Medline: 20308349]
18. Cheitlin M. Impact of remote telemedical management on mortality and hospitalizations in ambulatory patients with chronic heart failure: the Telemedical Interventional Monitoring in Heart Failure Study. Yearbook of Cardiology 2012 Jan;2012:312-315. [doi: 10.1016/j.ycar.2012.01.028]
19. Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, et al. Telemonitoring in patients with heart failure. N Engl J Med 2010 Dec 09;363(24):2301-2309 [FREE Full text] [doi: 10.1056/NEJMoa1010029] [Medline: 21080835]
20. Madujibeya I, Lennie TA, Pelzel J, Moser DK. Patients' experiences using a mobile health app for self-care of heart failure in a real-world setting: qualitative analysis. JMIR Form Res 2023 Aug 15;7:e39525 [FREE Full text] [doi: 10.2196/39525] [Medline: 37581912]
21. Mercer K, Giangregorio L, Schneider E, Chilana P, Li M, Grindrod K. Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation. JMIR Mhealth Uhealth 2016 Jan 27;4(1):e7 [FREE Full text] [doi: 10.2196/mhealth.4225] [Medline: 26818775]
22. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han H. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018 Aug;36(8):376-382 [FREE Full text] [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
23. Wali S, Keshavjee K, Nguyen L, Mbuagbaw L, Demers C. Using an electronic app to promote home-based self-care in older patients with heart failure: qualitative study on patient and informal caregiver challenges. JMIR Cardio 2020 Nov 09;4(1):e15885 [FREE Full text] [doi: 10.2196/15885] [Medline: 33164901]
24. Cajita MI, Hodgson NA, Budhathoki C, Han H. Intention to use mHealth in older adults with heart failure. J Cardiovasc Nurs 2017;32(6):E1-E7 [FREE Full text] [doi: 10.1097/JCN.0000000000000401] [Medline: 28248747]
25. Cajita MI, Gleason KT, Han H. A systematic review of mHealth-based heart failure interventions. J Cardiovasc Nurs 2016;31(3):E10-E22 [FREE Full text] [doi: 10.1097/JCN.0000000000000305] [Medline: 26544175]
26. Madujibeya I, Chung M, Moser D, Miller J, Humbert J, Chih M, et al. Patients experiences of using a publicly avaliable mobile health application for self-care of heart failure in a real-world setting. European Journal of Cardiovascular Nursing 2022;21(Supplement_1):zvac060.111 [FREE Full text] [doi: 10.1093/eurjcn/zvac060.111]
27. Sivakumar B, Lemonde M, Stein M, Mak S, Al-Hesayen A, Arcand J. Patient perspectives on the use of mobile apps to support heart failure management: A qualitative descriptive study. PLoS One 2023;18(5):e0285659 [FREE Full text] [doi: 10.1371/journal.pone.0285659] [Medline: 37167283]
28. McHorney CA, Mansukhani SG, Anatchkova M, Taylor N, Wirtz HS, Abbasi S, et al. The impact of heart failure on patients and caregivers: A qualitative study. PLoS One 2021 Mar 11;16(3):e0248240 [FREE Full text] [doi: 10.1371/journal.pone.0248240] [Medline: 33705486]
29. Definitions. Family Caregiver Alliance. URL: https://www.caregiver.org/resource/definitions-0/ [accessed 2024-10-27]
30. Wingham J, Frost J, Britten N, Jolly K, Greaves C, Abraham C, REACH-HF research investigators. Needs of caregivers in heart failure management: A qualitative study. Chronic Illn 2015 Dec 19;11(4):304-319 [FREE Full text] [doi: 10.1177/1742395315574765] [Medline: 25795144]
31. El-Dassouki N, Pfisterer K, Benmessaoud C, Young K, Ge K, Lohani R, et al. The value of technology to support dyadic caregiving for individuals living with heart failure: qualitative descriptive study. J Med Internet Res 2022 Sep 07;24(9):e40108 [FREE Full text] [doi: 10.2196/40108] [Medline: 36069782]

XSL•FO
RenderX

32. Bezerra Giordan L, Ronto R, Chau J, Chow C, Laranjo L. Use of mobile apps in heart failure self-management: qualitative study exploring the patient and primary care clinician perspective. JMIR Cardio 2022 Apr 20;6(1):e33992 [FREE Full text] [doi: 10.2196/33992] [Medline: 35442205]
33. Schmaderer M, Miller JN, Mollard E. Experiences of using a self-management mobile app among individuals with heart failure: qualitative study. JMIR Nurs 2021 Aug 09;4(3):e28139 [FREE Full text] [doi: 10.2196/28139] [Medline: 34406966]
34. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Management Science 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
35. Sun Y, Wang N, Guo X, Peng J. Understanding the acceptance of mobile health services: A comparison and integration of alternative models. Journal of Electronic Commerce Research 2013;14(2):183-200 [FREE Full text]
36. Bourgeault I, Dingwall R, De Vries R. The SAGE Handbook of Qualitative Methods in Health Research. London, England: SAGE Publications Ltd; 2010.
37. Morrow R, Rodriguez A, King N. Colaizzi's descriptive phenomenological method. The Psychologist 2015;28(8):643-644 [FREE Full text]
38. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
39. Eisikovits Z, Koren C. Approaches to and outcomes of dyadic interview analysis. Qual Health Res 2010 Dec 27;20(12):1642-1655. [doi: 10.1177/1049732310376520] [Medline: 20663940]
40. Morgan DL, Ataie J, Carder P, Hoffman K. Introducing dyadic interviews as a method for collecting qualitative data. Qual Health Res 2013 Sep 07;23(9):1276-1284. [doi: 10.1177/1049732313501889] [Medline: 23925406]
41. Kvalsvik F, Øgaard T. Dyadic interviews versus in-depth individual interviews in exploring food choices of Norwegian older adults: a comparison of two qualitative methods. Foods 2021 May 26;10(6):1199 [FREE Full text] [doi: 10.3390/foods10061199] [Medline: 34073332]
42. Arksey H. Collecting data through joint interviews. Social Research Update 1996;Winter(15):1 [FREE Full text] [doi: 10.1017/cbo9780511627156.003]
43. Morgan DL. Essentials of Dyadic Interviewing. London, England: Routledge; 2016.
44. Head A, Ellis-Caird H, Rhodes L, Mengoni S. 'It was really good, she sort of took some words what happened, like what I would say': Adapting dyadic interview techniques to capture the stories of marginalised voices in research. QMiP Bulletin 2021 Mar 01;1(31):35-44. [doi: 10.53841/bpsqmip.2021.1.31.35]
45. Braun V, Clarke V. Thematic Analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological. Washington DC: American Psychological Association; 2012:57-71.
46. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
47. Braun V, Clarke V. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
48. Elliott J. The craft of using NVivo12 to analyze open-ended questions: an approach to mixed methods analysis. TQR 2022:1. [doi: 10.46743/2160-3715/2022.5460]
49. Trainor LR, Bundon A. Developing the craft: reflexive accounts of doing reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2020 Nov 02;13(5):705-726. [doi: 10.1080/2159676X.2020.1840423] [Medline: 22489612]
50. Thomas E, Magilvy J. Qualitative rigor or research validity in qualitative research. J Spec Pediatr Nurs 2011 Apr;16(2):151-155. [doi: 10.1111/j.1744-6155.2011.00283.x] [Medline: 21439005]
51. Buck H, Kitko L, Hupcey J. Dyadic heart failure care types: qualitative evidence for a novel typology. J Cardiovasc Nurs 2013;28(6):E37-E46 [FREE Full text] [doi: 10.1097/JCN.0b013e31827fcc4c] [Medline: 23388704]
52. Hilty DM, Crawford A, Teshima J, Nasatir-Hilty SE, Luo J, Chisler LSM, et al. Mobile health and cultural competencies as a foundation for telehealth care: scoping review. J Technol Behav Sci 2021 Feb 15;6(2):197-230. [doi: 10.1007/s41347-020-00180-5]
53. Triantafillou V, Layfield E, Prasad A, Deng J, Shanti RM, Newman JG, et al. Patient perceptions of head and neck ambulatory telemedicine visits: a qualitative study. Otolaryngol Head Neck Surg 2021 May 14;164(5):923-931. [doi: 10.1177/0194599820943523] [Medline: 32662749]
54. Ftouni R, AlJardali B, Hamdanieh M, Ftouni L, Salem N. Challenges of telemedicine during the COVID-19 pandemic: a systematic review. BMC Med Inform Decis Mak 2022 Aug 03;22(1):207 [FREE Full text] [doi: 10.1186/s12911-022-01952-0] [Medline: 35922817]
55. Ray MA. The theory of bureaucratic caring for nursing practice in the organizational culture. Nurs Adm Q 1989;13(2):31-42. [doi: 10.1097/00006216-198901320-00007] [Medline: 2922119]
56. Manworren RC. Nurses' management of children's acute postoperative pain: A theory of bureaucratic caring deductive study. J Pediatr Nurs 2022 May;64:42-55. [doi: 10.1016/j.pedn.2022.01.021] [Medline: 35149258]

XSL•FO
RenderX

57. Zhou M, Zhou H, Zhang X, Jin X, Su X, Bai Y, et al. A qualitative study on patients' and health care professionals' perspectives regarding care delivered during CIED operation. BMC Health Serv Res 2024 Jan 15;24(1):73 [FREE Full text] [doi: 10.1186/s12913-024-10546-7] [Medline: 38225638]
58. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790 [FREE Full text] [doi: 10.2196/34790] [Medline: 35171099]
59. Howgego G, Sharma I, Kalu P. The rules for online clinical engagement in the COVID era. J Plast Reconstr Aesthet Surg 2020 Dec;73(12):2127-2135 [FREE Full text] [doi: 10.1016/j.bjps.2020.08.045] [Medline: 33051176]
60. Gao Y, Liu R, Zhou Q, Wang X, Huang L, Shi Q, COVID-19 Evidence and Recommendations Working Group. Application of telemedicine during the coronavirus disease epidemics: a rapid review and meta-analysis. Ann Transl Med 2020 May;8(10):626 [FREE Full text] [doi: 10.21037/atm-20-3315] [Medline: 32566563]
61. Wamsley C, Kramer A, Kenkel J, Amirlak B. Trends and challenges of telehealth in an academic institution: the unforeseen benefits of the COVID-19 global pandemic. Aesthet Surg J 2021 Jan 01;41(1):109-118 [FREE Full text] [doi: 10.1093/asj/sjaa212] [Medline: 32697289]
62. Murphy EP, Fenelon C, Murphy RP, O'Sullivan MD, Pomeroy E, Sheehan E, et al. Are virtual fracture clinics during the COVID-19 pandemic a potential alternative for delivering fracture care? A Systematic Review. Clin Orthop Relat Res 2020 Jun 26;478(11):2610-2621. [doi: 10.1097/corr.0000000000001388]
63. Kärner Köhler A, Tingström P, Jaarsma T, Nilsson S. Patient empowerment and general self-efficacy in patients with coronary heart disease: a cross-sectional study. BMC Fam Pract 2018 May 30;19(1):76 [FREE Full text] [doi: 10.1186/s12875-018-0749-y] [Medline: 29843619]
64. Jiang S, Street RL. The effects of patient-centered communication, social capital, and internet use on patient empowerment: a cross-sectional study in China. Glob Health Promot 2019 Dec 09;26(4):33-43. [doi: 10.1177/1757975917749197] [Medline: 29521571]
65. Nahm E, Blum K, Scharf B, Friedmann E, Thomas S, Jones D, et al. Exploration of patients' readiness for an eHealth management program for chronic heart failure: A preliminary study. The Journal of Cardiovascular Nursing 2008;23(6):463-471. [doi: 10.1097/01.JCN.0000317459.41015.d6]
66. Yu M, Wang C, Ning L, Jin J, Liu X, Chen W, et al. The relationship between filial values and preparedness in adult -child caregivers of patients with chronic heart failure: A cross-sectional study. Geriatr Nurs 2024 Sep;59:669-676. [doi: 10.1016/j.gerinurse.2024.08.008] [Medline: 39213988]
67. Yoong S, Tan R, Jiang Y. Dyadic relationships between informal caregivers and older adults with chronic heart failure: A systematic review and meta-synthesis. Eur J Cardiovasc Nurs 2024 Jun 14:e093. [doi: 10.1093/eurjcn/zvae093] [Medline: 38875163]
68. Cameron J, Thompson DR, Szer D, Greig J, Ski CF. Dyadic incongruence in chronic heart failure: Implications for patient and carer psychological health and self-care. J Clin Nurs 2017 Dec 20;26(23-24):4804-4812. [doi: 10.1111/jocn.13836] [Medline: 28370704]
69. Dionigi RA. Stereotypes of aging: their effects on the health of older adults. Journal of Geriatrics 2015:1. [doi: 10.1155/2015/954027]
70. Swift HJ, Barber SJ, Lamont RA, Weiss D, Chasteen AL. Editorial: Age-based stereotype threat effects on performance outcomes. Front Psychol 2021 Nov 2;12:773615 [FREE Full text] [doi: 10.3389/fpsyg.2021.773615] [Medline: 34795622]
71. Ogihara Y. Chinese culture became more individualistic: Evidence from family structure, 1953-2017. F1000Res 2023;12:10 [FREE Full text] [doi: 10.12688/f1000research.128448.3] [Medline: 37476505]
72. Zhang L, Wei X, Ma X, Ren Z. Physical health and quality of life among older people in the context of Chinese culture. Int J Environ Res Public Health 2021 Jun 24;18(13):6798 [FREE Full text] [doi: 10.3390/ijerph18136798] [Medline: 34202846]
73. Sun CTL. Themes in Chinese psychology. Singapore: Cengage Learning Asia Pte Ltd; 2013.
74. Zimbudzi E, Lo C, Misso ML, Ranasinha S, Kerr PG, Teede HJ, et al. Effectiveness of self-management support interventions for people with comorbid diabetes and chronic kidney disease: a systematic review and meta-analysis. Syst Rev 2018 Jun 13;7(1):84 [FREE Full text] [doi: 10.1186/s13643-018-0748-z] [Medline: 29898785]
75. Polsook R, Aungsuroch Y, Thontham A. The effect of self-management intervention among type 2 diabetes: A systematic review and meta-analysis. Worldviews Evid Based Nurs 2024 Feb 02;21(1):59-67. [doi: 10.1111/wvn.12688] [Medline: 37916757]
76. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
77. Dang S, Karanam C, Gómez-Orozco C, Gómez-Marín O. Mobile phone intervention for heart failure in a minority urban county hospital population: usability and patient perspectives. Telemed J E Health 2017 Jul;23(7):544-554. [doi: 10.1089/tmj.2016.0224] [Medline: 28051761]

Abbreviations

CHF: chronic heart failure

XSL•FO
RenderX

IC: informal caregiver
mHealth: mobile health
SRQR: Standards for Reporting Qualitative Research

Edited by A Schwartz; submitted 06.02.24; peer-reviewed by A Hassan, S Alsahli, CH Chan, A Hewison; comments to author 02.05.24; revised version received 27.06.24; accepted 18.10.24; published 31.10.24.

Please cite as:
Jin X, Zhang Y, Zhou M, Zhang X, Mei Q, Bai Y, Wei W, Ma F
Experiences With mHealth Use Among Patient-Caregiver Dyads With Chronic Heart Failure: Qualitative Study
J Med Internet Res 2024;26:e57115
URL: https://www.jmir.org/2024/1/e57115
doi:10.2196/57115
PMID:

©Xiaorong Jin, Yimei Zhang, Min Zhou, Xiong Zhang, Qian Mei, Yangjuan Bai, Wei Wei, Fang Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of a Wearable Activity Tracker With Step-by-Step Goal-Setting on Older Adults' Physical Activity and Sarcopenia Indicators: Clustered Trial

Mu-Hsing Ho[1], PhD; Chi-Yuan Peng[2], MSc; Yung Liao[3], PhD; Hsin-Yen Yen[2], PhD

[1]School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, China (Hong Kong)
[2]School of Gerontology and Long-term Care, College of Nursing, Taipei Medical University, Taipei City, Taiwan
[3]Graduate Institute of Sport, Leisure, and Hospitality Management, National Taiwan Normal University, Taipei, Taiwan

Corresponding Author:
Hsin-Yen Yen, PhD
School of Gerontology and Long-term Care
College of Nursing
Taipei Medical University
250 Wuxing St
Taipei City, 11031
Taiwan
Phone: 886 2 2736 1661 ext 6326
Email: yenken520@gmail.com

Abstract

Background: Smart wearable technology has potential benefits for promoting physical activity and preventing sarcopenia.

Objective: The purpose of this study was to explore the efficacy of a wearable activity tracker with 2-stage goal-setting for daily steps on older adults' physical activity and sarcopenia indicators.

Methods: The study used a clustered trial design and was conducted in March to June 2022. Participants were community-dwelling adults older than 60 years who were recruited from 4 community centers in Taipei City. The intervention was designed with 2-stage goals set to 5000 steps/day in the first 4 weeks and 7500 steps/day in the final 4 weeks while wearing a commercial wearable activity tracker. Data were collected by self-reported questionnaires, a body composition analyzer, a handle grip tester, and 5 sit-to-stand tests.

Results: All 27 participants in the experimental group and 31 participants in the control group completed the 8-week intervention. Total and light-intensity physical activities, skeletal muscle index, and muscle strength increased, while sedentary time, BMI, and the waist circumference of participants decreased in the experimental group, with significant group-by-time interactions compared to the control group.

Conclusions: A wearable activity tracker with gradual goal-setting is an efficient approach to improve older adults' physical activity and sarcopenia indicators. Smart wearable products with behavioral change techniques are recommended to prevent sarcopenia in older adult populations.

(J Med Internet Res 2024;26:e60183) doi:10.2196/60183

KEYWORDS

behavioral change technique; chronic disease prevention; health promotion; mHealth; sedentary behavior; smartwatch

Introduction

Sarcopenia is an urgent health challenge in later life. Older adults with sarcopenia have increased risks of falls, fractures, and disabilities; a poor quality of life; functional declines; and increased mortality [1]. Muscle mass, muscle strength, and physical performance are 3 indicators of sarcopenia [2]. Muscle mass of adults may decrease 8% per decade after the age of 40 years, and muscle mass rapidly decreases by 10% to 15% per decade after the age of 70 years [3]. Muscle strength decreases by 10% to 15% per decade after the age of 40 years, and by 25% to 40% per decade after the age of 70 years [4]. Physical functions decline more rapidly, at 1.5% per year between the ages of 50 and 60 years and 3% per year thereafter [5]. The

XSL•FO
RenderX

prevalence rate of sarcopenia was estimated to be 10% in both older men and older women [6].

Engaging in physical activity (PA) is one preventive strategy to slow the development of sarcopenia in older adult populations [7,8]. Sufficient PA can improve older adults' gait speed, balance, physical functions, coordination, and activities of daily living [9]. Sufficient PA is generally defined as engaging in PA to meet recommended guidelines for maintaining health and preventing disease, such as those proposed by the World Health Organization, American College of Sports Medicine, and other health care institutions [10], or achieving a minimum of 600 metabolic equivalents of task (MET)–minutes per week [11]. PA promotion and structured exercise training programs were found to improve muscle mass, grip and leg muscle strength, and physical performance in older adult populations both with and without sarcopenia [12,13].

Walking is a key way to stay active and expend energy in daily life. It is an easy, feasible, and convenient option for older adults of various ages and health conditions [14]. Older adults are recommended to walk at least 7000 steps/day, which is equivalent to at least 30 minutes/day (or 150 minutes/week) of PA [15]. A previous prospective cohort study found that middle-aged adults who walked more than 7000 steps/day had a 50% to 70% lower risk of death [16]. For older adults with an inactive and sedentary lifestyle, at least 4600 to 5500 steps per day are recommended to reduce the risk of developing chronic diseases. An increase of 1000 steps/day revealed additional health benefits [17]. A previous investigation of an older population aged over 60 years indicated that increasing walking may decrease the odds of sarcopenia. The link between sarcopenia and PA was particularly evident in those who walked [18].

Smart wearable technology is an innovative approach to modify PA in daily life to prevent sarcopenia. Commercial wearable activity trackers (WATs) have features that help users stay motivated and on track with their fitness goals. These devices continually monitor PA levels, provide real-time feedback, set achievable goals, offer reminders, and even encourage social interactions and competition [19-21]. For instance, a WAT allows users to effortlessly track their daily steps, monitor heart rate, and measure energy expenditure as part of their routine. It can also send gentle reminders to stand up and break up sedentary habits, while providing an easy way to review progress toward their PA goals [22,23]. The combination of WATs and smartphone apps can enhance the effects of PA interventions [24]. However, the validity of WATs as PA-measuring tools, using sensors like accelerometers and GPS to track movement, steps, and heart rate, is still a problem. The accuracy can vary by device and activity type and depends on consistent use. Their accuracy may be less reliable for PA like cycling or strength training compared to walking [25].

Older adults are increasingly embracing WATs due to their potential health benefits and personalized features, such as comfort, sleek designs, and user-friendly interfaces [26,27]. A key factor in the success of WATs is self-efficacy—the belief in one's own ability to succeed. This is essential for using WATs to enhance and maintain PA levels [28]. For older adults, low self-efficacy can diminish motivation, making it harder to maintain healthy habits [29]. WATs support users by helping them gradually achieve short-, medium-, and long-term goals through consistent encouragement and reminders, rather than pushing them toward unrealistic targets [30]. By providing behavioral change techniques, WATs effectively increase older adults' self-efficacy, awareness, and proactive engagement in PA and health self-management [23].

Previous experimental studies suggested that wearable-based interventions, including WATs, pedometers, and other fitness trackers, have potential benefits in improving PA and modifying healthy behaviors [22,24,31,32]. Also, PA had positive impacts on reducing the risks and incidence of sarcopenia [8,33] and improving muscle mass, muscle strength, and physical performance in older community-dwelling adults [7,34].

In this study, we developed an innovative wearable-based intervention with behavioral change technologies. The intervention was designed as 2-stage goal-setting for daily steps via a commercial WAT. A nutritional health educational program that focused on healthy eating was also conducted in both the control and experimental groups. The study purpose was to explore the efficacy and feasibility of an intervention among community-dwelling older adults and investigate PA and sarcopenia indicators. Outcomes included multi-intensity PA, sedentary time, waist circumference (WC), BMI, muscle mass, muscle strength, and physical performance.

Methods

Study Design

This study was a clustered trial conducted in March to June 2022. Participants were recruited from 4 community centers in Taipei City. Of these 4 centers, 2 were assigned to either an experimental group or a control group by a draw. Participants were not aware of the other groups because the community centers were in different locations.

Participants

Participants were community-dwelling older adults. All community centers funded as part of the local long-term care policy serve as primary care facilities. These centers are designated to assist healthy older adults who live independently, have no disabilities, can move about without physical limitations, and have chronic conditions that do not currently require urgent medical attention.

The inclusion criteria were (1) being an adult older than 60 years, (2) having a smartphone that can download the app connected to the WAT, (3) being able to learn how to use the WAT, and (4) having used no device to monitor daily steps in the previous 6 months. The exclusion criteria were (1) having any disease or cancer that was currently under medical care (ie, surgery, chemotherapy, or hemodialysis); (2) having any diagnosed disability (ie, dementia or hearing or visual impairment), (3) being unable to walk on one's own; and (4) having any electronic device inside one's body. The minimum sample size was precalculated by medium effect sizes via G*power (Universität Düsseldorf) with an added 20% attrition rate, resulting in a sample size of 41. In total, 70 participants

XSL•FO
RenderX

were recruited from 4 clusters; 58 participants were assessed as meeting the criteria. Finally, 27 participants were allocated to the experimental group, and another 31 participants were allocated to the control group.

The Intervention

The intervention used a WAT connected to a smartphone app, which was set to progressively increase daily step goals over an 8-week period. Participants in the experimental group were required to wear a commercial WAT, the Asus VivoWatch BP (HC-A04; Asus), on their wrist. Participants were carefully taught how to use the WAT and the smartphone app, and they were introduced to all features before the intervention. Features included tracking of daily steps, walking distance, heart rate, calories burned, exercise duration, and notifications of long sedentary time. By connecting to the smartphone app, the user could access integrated information about tracking results and gain personalized health advice. Participants were encouraged to use the WAT during their waking time and only take it off to charge during sleep and when bathing every day. Participants were allowed to freely use the smartphone app without specific restrictions. The social media account of the research team was provided in case participants had any difficulty using the WAT. They could contact the research team at any time to resolve problems.

The 2-stage goals were 5000 steps/day in the first month (weeks 1-4) and 7500 steps/day in the second month (weeks 5-8). The research team set the step goals on their smartphones. When a participant reached the target number of daily steps during the wearing time, the WAT would vibrate and display a “trophy” image to inform them. Achievements were recorded on the smartphone app, which allowed the older adults to check them at any time. Reminders and achievements can encourage older adults to keep pursuing their goals each day through self-efficacy. In addition, once a week, participants received group-based nutritional health education about healthy eating. During the weekly group session, a research team member checked use of the WAT, and participants could ask for help with any difficulties using the WAT to resolve the problem. Participants who mentioned having lower wearing time were reminded to use the WAT more frequently.

Participants in the control group received the same weekly nutritional health education during the 8 weeks of the study.

Outcome Measures

Data were collected twice, at the pretest and postintervention stages, by a trained master’s degree student. Participant characteristics were collected before the intervention, including age, sex, marital status, educational level, and presence of chronic diseases (yes/no).

PA and Sedentary Time

Subjective data were collected by structured questionnaires via one-on-one interviews. The International Physical Activity Questionnaire–Short Form (IPAQ-SF) was used to assess participants’ PA and sedentary time in the past week through 7 items. The IPAQ-SF has good validity and internal consistency [11]. Daily sedentary hours and MET-min/week of light-intensity PA (LPA), moderate-to-vigorous-intensity PA (MVPA), and total PA were calculated as outcomes.

Sarcopenia Indicators, BMI, and WC

Sarcopenia indicators were measured according to the Asian Working Group for Sarcopenia [2]. BMI and skeletal muscle index (SMI) were measured with a body composition analyzer (Inbody 270; Inbody) using bioelectrical impedance analytical technology. The participant stepped barefoot onto the foot electrodes while gripping the hand electrodes. During this time, the participant stood still and kept their arms straight without touching for 1 minute. A higher score (kg/m^2) for SMI indicates greater muscle mass. Muscle strength was assessed by a handle grip instrument (Tokyo Telecommunication Machine Corp). Participants were asked to stand, allowing the wrist and arm of the dominant hand to hang straight down. The individual squeezed the device with maximum effort for more than 3 seconds. This measurement was repeated 3 times. The highest value recorded from multiple trials was used as an indicator of muscle strength. Physical performance was measured by 5 sit-to-stand tests. The participant was asked to rise from a seated position in a chair to a full standing position and then sit back down with no assistance as quickly as possible 5 times. A timer measured the time required to complete the task. The total time taken to complete 5 repetitions was recorded. A longer time (in seconds) indicated worse physical performance. In addition, to measure WC, a flexible, nonstretchable tape was used. The tape was positioned horizontally around the abdomen, just above the belly button, between the bottom of the rib cage and the top of the hip bones. The measurement was taken after a normal exhalation.

Statistical Analysis

Descriptive analyses included the frequency and percentage of participants’ characteristics and the mean and SD of baseline outcome scores. Chi-square tests and independent 2-tailed *t* tests were performed to compare participants’ characteristics and baseline outcome scores between the control and experimental groups. An analysis of covariance (ANCOVA) was performed to compare participants’ posttest scores between the 2 groups by controlling for their age and pretest scores. Paired 2-tailed *t* tests were performed to analyze differences in subjective and objective outcomes between the pretest and postintervention stages in each group with the mean difference (MD) and 95% CI. Cohen *d* and η^2 were calculated for effect sizes. Generalized estimating equations (GEEs) were used to analyze group-by-time interactions on outcomes, which were adjusted for participants’ age and baseline scores in the pretest. After testing, all outcome values exhibited a normal distribution. SPSS (version 18.0; IBM Corp) was used for all statistical analyses.

Ethical Considerations

Participants were required to sign an informed consent form agreeing to participate before entering the study. All participants’ data were handled with strict confidentiality and privacy throughout the storage and analysis processes. Data were anonymized and securely stored in compliance with applicable data protection regulations, ensuring that personal

XSL•FO
RenderX

information was not accessible to unauthorized individuals. All participants were fully informed prior to the intervention that no financial compensation or material benefits would be provided upon its completion. Ethical approval for this study was obtained from the Taipei Medical University Joint Institutional Review Board (N202202031).

Results

The study flow is illustrated in Figure 1. Participant characteristics and outcome scores at baseline are described in Tables 1 and 2. Sex, marital status, educational level, and the presence of chronic disease had no significant differences between the experimental and control groups. However, participants in the control group (mean age 74.26, SD 6.30 years) were significantly older than those in the experimental group (mean age 70.26, SD 4.72 years; *P*=.01). Participants' sedentary time at baseline significantly differed between the 2 groups (*P*=.048). No other outcome score at baseline significantly differed between the 2 groups (*P*<.05). None of the participants dropped out of the study. All 27 participants in the experimental group and 31 participants in the control group completed the intervention and were included in subsequent analyses.

Figure 1. Consolidated Standards of Reporting Trials (CONSORT) flow diagram.

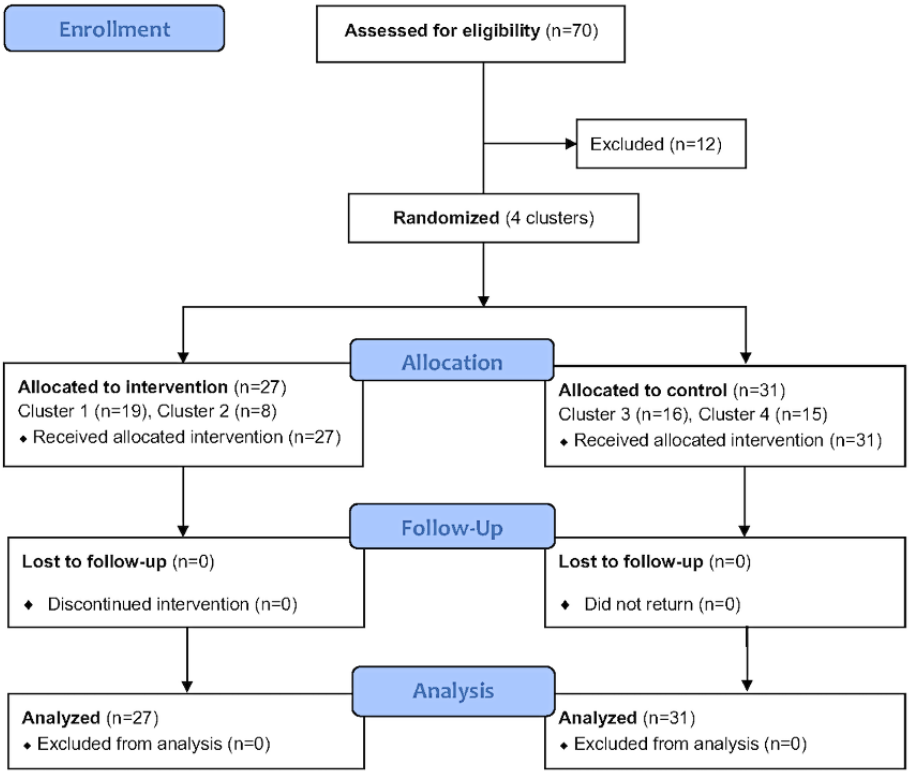

Table 1. Participant characteristics.

Characteristics	Experimental group (n=27)	Control group (n=31)	*P* value
Sex, n (%)			.39
Male	3 (11.11)	6 (19.35)	
Female	24 (88.89)	25 (80.65)	
Marital status, n (%)			.36
Married	17 (62.96)	23 (74.19)	
Single, divorced, or widowed	10 (37.04)	8 (25.81)	
Education, n (%)			.13
None to elementary school	7 (25.93)	14 (45.16)	
High school or higher	20 (74.07)	17 (54.84)	
Chronic diseases, n (%)			.88
No	11 (40.74)	12 (38.71)	
Yes	16 (59.26)	19 (61.29)	
Age (years), mean (SD)	70.26 (4.72)	74.26 (6.30)	.01

XSL•FO
RenderX

Table 2. Comparison of participant outcomes at the pre-and posttest stages between the groups.

Outcome	Experimental group	Control group	*P* value	Cohen *d*	η^2
Pretest scores, mean (SD)					
Waist circumference (cm)	85.35 (8.34)	84.47 (9.62)	.71	0.10	—[a]
BMI (kg/m^2)	24.39 (2.71)	23.86 (3.18)	.50	0.18	—
Physical activity					
Total (MET[b]-min/week)	1609.33 (1443.41)	1653.71 (1605.98)	.91	0.03	—
Moderate to vigorous (MET-min/week)	700.00 (841.77)	636.13 (888.41)	.78	0.07	—
Light intensity (MET-min/week)	909.33 (1034.58)	1016.94 (934.20)	.68	0.11	—
Sedentary time (h/day)	45.59 (8.18)	45.45 (4.66)	.048	0.02	—
Sarcopenia indicators					
Skeletal muscle index (kg/m^2)	6.20 (0.64)	5.99 (1.17)	.40	0.22	—
Muscle strength (kg)	22.69 (6.10)	20.26 (6.95)	.17	0.37	—
Physical performance (s)	7.97 (2.34)	9.18 (2.97)	.10	0.45	—
Posttest scores, mean (SE)					
Waist circumference (cm)	84.74 (0.77)	85.88 (0.71)	.30	—	0.02
BMI (kg/m^2)	24.06 (0.06)	24.49 (0.06)	<.001	—	0.31
Physical activity					
Total (MET-min/week)	2056.65 (244.26)	1225.37 (227.00)	.02	—	0.10
Moderate to vigorous (MET-min/week)	258.19 (104.67)	394.48 (97.27)	.36	—	0.02
Light intensity (MET-min/week)	1777.87 (201.38)	848.83 (187.14)	.002	—	0.17
Sedentary time (h/day)	5.53 (0.41)	6.69 (0.37)	.06	—	0.07
Sarcopenia indicators					
Skeletal muscle index (kg/m^2)	6.13 (0.04)	6.24 (0.04)	.06	—	0.07
Muscle strength (kg)	22.31 (0.44)	20.96 (0.41)	.03	—	0.08
Physical performance (s)	7.00 (0.28)	8.07 (0.26)	.01	—	0.12

[a]Not applicable.

[b]MET: metabolic equivalent of task.

Table 2 also shows comparisons of participants' outcomes at the pre- and posttests between the 2 groups. After adjusting for age and pretest scores, participants in the experimental group had a significantly lower BMI, higher total PA and LPA, and better muscle strength and physical performance than those in the control group. WC, MVPA, sedentary time, and SMI did not show a significant difference between the 2 groups at the posttest.

Paired *t* test and GEE results are presented in Figure 2 and Table 3. After the 8-week intervention, the total PA (MD 510.96 MET-min/week), LPA (MD 919.11 MET-min/week), SMI (MD 0.09 kg/m^2), and muscle strength (MD 0.84 kg) of participants in the experimental group had increased, while those of participants in the control group had decreased, with significant group-by-time interactions (P=.03, P=.002, P<.001, and P<.001, respectively). Sedentary time (MD –2.02 h/day), BMI (MD –0.21 kg/m^2), and WC (MD –0.48 cm) decreased for participants in the experimental group, while they increased in the control group, with significant group-by-time interactions (P=.001, P=.04, and P<.001, respectively). The MVPA of participants in both groups decreased without a significant group-by-time interaction (P=.41). Physical performance of participants in the experimental group (MD –1.50 s) improved compared to the control group (MD –0.64 s), with a significant group-by-time interaction (P=.04).

Within the experimental group, neither participants' BMI nor total PA showed a significant change after the intervention. Within the control group, participants' WC increased, and MVPA and total PA decreased after the 8 weeks. No other outcomes showed a significant change after the 8 weeks, including BMI, LPA, sedentary time, and all sarcopenia indicators.

XSL•FO
RenderX

Figure 2. Results of outcome changes within each group. MET: metabolic equivalent of task; PA: physical activity. *P<.05, **P<.01, ***P<.001.

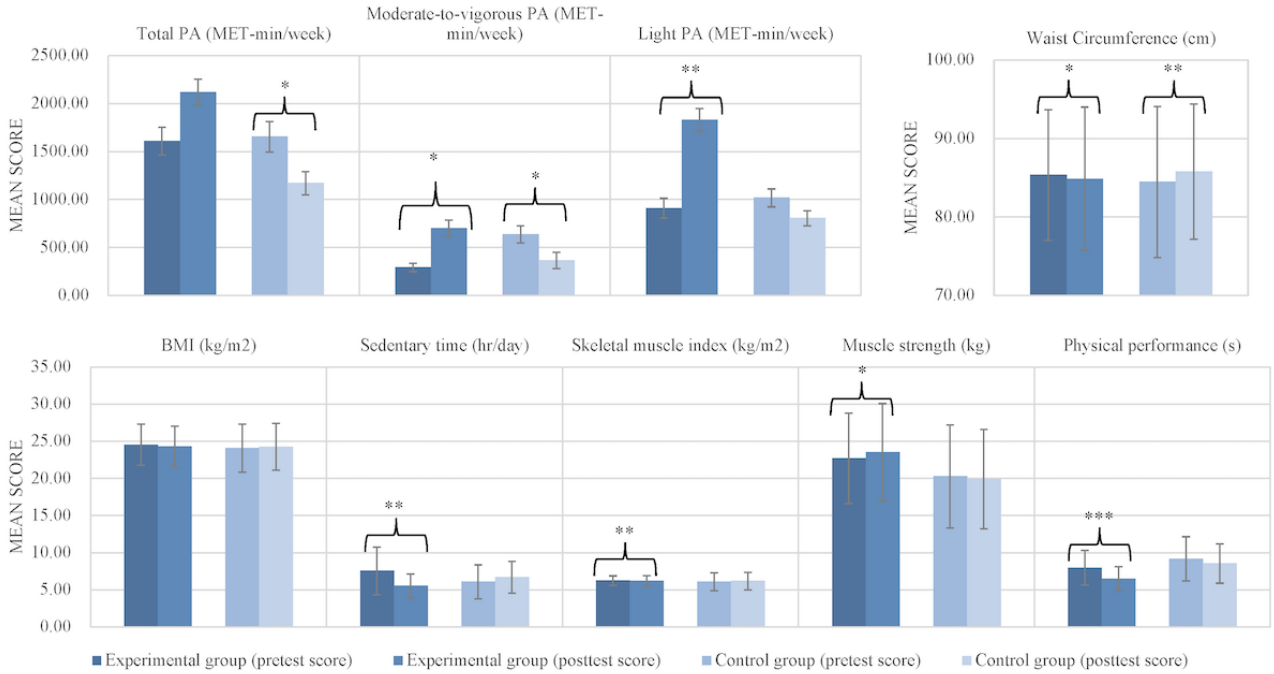

Table 3. Differences within and between the experimental and control groups. Generalized estimating equations were adjusted for participant age and pretest scores; Cohen d was used for effect size.

Outcome	Experimental group			Control group			Group × time interaction (P value)
	MD[a] (95% CI)	P value	Cohen d	MD (95% CI)	P value	Cohen d	
Waist circumference (cm)	–0.21 (–0.38 to 0.04)	.02	0.14	0.17 (0.08 to 0.27)	.001	0.30	<.001
BMI (kg/m^2)	–0.48 (–1.87 to 0.90)	.48	0.50	1.30 (–0.30 to 2.89)	.11	0.73	.04
Total PA[b] (MET[c]-min/week)	510.96 (–297.50 to 1319.43)	.21	0.25	–483.77 (–967.81 to 0.26)	.049	0.37	.03
MVPA[d] (MET-min/week)	–408.15 (–741.53 to 74.77)	.02	0.48	–270.97 (–432.29 to 109.64)	.002	0.62	.41
LPA[e] (MET–min/week)	919.11 (244.05 to 1594.17)	.01	0.54	–212.16 (–613.22 to 188.90)	.29	0.19	.002
Sedentary time (h/day)	–2.02 (–3.51 to 0.53)	.01	0.54	0.63 (–0.29 to 1.55)	.17	0.25	.001
Skeletal muscle index (kg/m^2)	0.09 (0.02–0.16)	.01	0.10	–0.02 (–0.10 to 0.06)	.61	0.54	<.001
Muscle strength (kg)	0.84 (0.03 to 1.64)	.04	0.41	–0.36 (–1.20 to 0.48)	.39	0.16	<.001
Physical performance (s)	–1.50 (–2.14 to 0.86)	<.001	0.92	–0.64 (–1.29 to 0.00)	.05	0.37	.04

[a]MD: mean difference between pre- and posttest score.
[b]PA: physical activity.
[c]MET: metabolic equivalent of task.
[d]MVPA: moderate-to-vigorous-intensity physical activity.
[e]LPA: light-intensity physical activity.

Discussion

Principal Findings

In this study, we designed an innovative intervention based on a WAT with 2-stage goal-setting for daily steps and explored its efficacy in improving older adults' PA and sarcopenia indicators. The total PA, LPA, sedentary time, BMI, WC, SMI, muscle strength, and physical performance of older adults who engaged in the 8-week intervention improved compared to those in the control group. Study results suggested that the WAT-based intervention with gradual goal-setting was

XSL•FO
RenderX

beneficial for community-dwelling older adults' PA and sarcopenia prevention. However, due to the cluster randomization, older adults in the experimental group were younger than those in the control group. The findings should be interpreted with caution.

A WAT with step-by-step goal setting improved older adults' LPA and sedentary time but not MVPA after the 8-week intervention. The intervention was designed with daily steps as a goal, and with LPA as a targeted outcome for improvement, instead of MVPA. This is possibly why LPA was the only indicator that improved among multi-intensity types of PA. Previous studies also suggested that WAT-based interventions can improve participants' PA, especially LPA. A previous randomized trial conducted a 6-week intervention with a pedometer with weekly goals. Their results also found that LPA improved but MVPA did not change after the intervention [33]. Another study conducted interventions with a WAT to record daily steps and PA with additional in-person health behavioral change techniques, including goal-setting, self-monitoring, building PA plans, and using social interactions, and the self-efficacy, daily steps, and the amount of PA of older adults who participated in the 3-month intervention improved; furthermore, participants' PA level was maintained after the intervention [35].

WAT-based interventions provide behavioral change techniques to support and motivate older adults' health self-management, PA engagement, and self-efficacy [19]. Results of a 3-arm randomized trial consisting of a control group, wearable group, and wearable group with goal-setting also found that the PA of older adults in the wearable group, which had a daily goal of 10,000 steps, had improved after 12 weeks [36]. Older adults are able to use WATs to understand physiological data, behavioral patterns, and health habits based on their personal records to promote awareness of their health conditions [23].

This study reinforced findings that WATs with 2-stage goal-setting for daily steps can improve older adults' BMI and all sarcopenia indicators. Previous studies found similar results; PA reduced older adults' sarcopenia risks. A previous randomized trial found that the physical performance and muscle strength of sedentary older adults engaged in a 2.6-year PA intervention with weekly goals improved compared to those who only received a health educational program [37]. Another randomized trial with a large sample size found that PA, physical performance, and muscle strength of older adults who received a 24-month walking intervention improved compared to those who only received discussion of a health-related topic [38]. Another study conducted a 6-month WAT-based intervention for sedentary community-dwelling older adults; PA, fear of falling, physical performance, and muscle mass of participants in the experimental group improved after the intervention [39].

This study found increased LPA and decreased MVPA after the intervention, but sarcopenia indicators improved nevertheless. In contrast, a previous study [40] suggested that there was a significant impact from reallocating sedentary time to at least LPA on reducing sarcopenia risks in older adults. Their findings indicated that engaging in LPA and MVPA was associated with a significantly lower sarcopenia risk. Reallocating more time to MVPA usually offers more protection for mitigating age-related muscle deterioration and promoting healthier aging [40].

Furthermore, both groups received identical weekly nutritional health education on healthy eating, which could have influenced their eating behaviors during the intervention and sarcopenia indicators. While sarcopenia indicators improved in the experimental group after the intervention, no such changes were observed in the control group. This suggests that the WAT-based intervention offered PA-promoting benefits. PA is an important protective factor for older adults that can be reinforced by goal-setting. Behavioral change techniques can be integrated into WAT-based interventions to prevent sarcopenia [7].

Limitations

There are several limitations to this study. First, the major limitation is that it was a clustered trial with a small sample size that randomly assigned the community centers to either the control or intervention group, instead of using individual randomization. Variations in the characteristics of the community centers may have increased the risk of bias, as reflected in significant differences between them, particularly in participants' ages (the experimental group mean age was 70.26 years, while the control group mean age was 74.26 years) and baseline scores. Although we used age and pretest scores as covariates for adjusting the ANCOVA and GEEs, the results should be interpreted with caution in future practice. Second, this study only used structured questionnaires to collect subjective data on PA and sedentary time. The accuracy might have been lower than if we had used objective PA measures, such as accelerometers and pedometers. In future studies, objective measures of PA and calories should be used during, before, and after the intervention, especially using WATs, as in this intervention. Third, the research team could not control users' behavior and use patterns of the WATs. It is difficult to know the actual wearing time, how long they did not wear the device due to abnormal connections or other technical problems, or how many times they reached the goal of daily steps. The adherence rate might potentially be a major bias. Finally, the intervention was only conducted for 8 weeks before the community centers closed because of the COVID-19 pandemic. Improvements in PA outcomes and sarcopenia indicators could not absolutely be attributed to the WATs and 2-stage goal-setting.

Implications

This study has implications for clinical application. Attention should be paid to sarcopenia risks in older adult populations. Health practitioners can advocate for older adults to adopt cost-efficient WATs to engage in walking. Due to advancements in technology, wearable devices feature increased accuracy without being invasive. The physiological data generated by WATs are expected to be an important reference for physicians and patients to make further medical decisions and health actions. The smartphone app may provide immediate health advice according to the data. WAT-based interventions are recommended as practical applications in older adults across different health care settings, such as long-term care facilities, daycare centers, and nursing homes.

XSL•FO
RenderX

This study also has implications for future research. Future studies can conduct WAT-based interventions to examine the health benefits of physiological indicators, such as blood pressure, blood glucose, and heart rate, across different older adult populations with chronic diseases. The goal of future WAT-based interventions could be to increase older adults' MVPA, given its major impact on the development of sarcopenia. The duration of the intervention could be prolonged. Innovative and interesting wearable-based interventions can be designed according to the preferences of targeted populations. For increased adherence to WATs, the interventions could be integrated with different behavioral change techniques, including social support, rewards, reviews, competitions, and goal-setting for sedentary or exercise time. In addition, we did not have a platform to access personal data during the wearing period (eg, PA, heart rate, and calories). These data can be recorded to test the reliability and validity of the current WAT compared to other sensors, accelerometers, commercial WATs, and questionnaires.

Conclusions

Smart wearable technologies are an innovative and efficient strategy for health behavioral modifications in older adult populations. Engaging in MVPA usually offers more protective benefits, while LPA by itself has a limited effect against developing sarcopenia. This study found that an 8-week intervention with a WAT and 2-stage goal-setting for daily steps improved older adults' PA and sarcopenia indicators, including total PA, LPA, sedentary time, BMI, WC, muscle mass, muscle strength, and physical performance. This suggests that a WAT-based intervention with nutritional health education is feasible and efficient in older populations. However, the findings should be interpreted with caution when considering practical applications due to the age difference between the 2 groups. Commercial WATs with behavioral change techniques are recommended to promote older adults' PA and prevent sarcopenia.

Acknowledgments

This work was financially supported by the National Taiwan Normal University within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

Data Availability

The datasets generated during and analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

MHH contributed to writing (review). CYP contributed to data curation and investigation. YL contributed to writing (review). HYY contributed to conceptualization, project administration, methodology, software, formal analysis, and writing.

Conflicts of Interest

None declared.

References

1. Yuan S, Larsson SC. Epidemiology of sarcopenia: prevalence, risk factors, and consequences. Metabolism 2023 Jul;144:155533 [FREE Full text] [doi: 10.1016/j.metabol.2023.155533] [Medline: 36907247]
2. Chen L, Woo J, Assantachai P, Auyeung T, Chou M, Iijima K, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020 Mar;21(3):300-307.e2. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
3. Mitchell WK, Williams J, Atherton P, Larvin M, Lund J, Narici M. Sarcopenia, dynapenia, and the impact of advancing age on human skeletal muscle size and strength; a quantitative review. Front Physiol 2012;3:260 [FREE Full text] [doi: 10.3389/fphys.2012.00260] [Medline: 22934016]
4. Goodpaster BH, Park SW, Harris TB, Kritchevsky SB, Nevitt M, Schwartz AV, et al. The loss of skeletal muscle strength, mass, and quality in older adults: the health, aging and body composition study. J Gerontol A Biol Sci Med Sci 2006 Oct;61(10):1059-1064. [doi: 10.1093/gerona/61.10.1059] [Medline: 17077199]
5. von Haehling S, Morley J, Anker S. An overview of sarcopenia: facts and numbers on prevalence and clinical impact. J Cachexia Sarcopenia Muscle 2010 Dec;1(2):129-133 [FREE Full text] [doi: 10.1007/s13539-010-0014-2] [Medline: 21475695]
6. Shafiee G, Keshtkar A, Soltani A, Ahadi Z, Larijani B, Heshmat R. Prevalence of sarcopenia in the world: a systematic review and meta-analysis of general population studies. J Diabetes Metab Disord 2017;16:21 [FREE Full text] [doi: 10.1186/s40200-017-0302-x] [Medline: 28523252]
7. Steffl M, Bohannon RW, Sontakova L, Tufano JJ, Shiells K, Holmerova I. Relationship between sarcopenia and physical activity in older people: a systematic review and meta-analysis. Clin Interv Aging 2017;12:835-845 [FREE Full text] [doi: 10.2147/CIA.S132940] [Medline: 28553092]

XSL•FO
RenderX

8. Oliveira JS, Pinheiro MB, Fairhall N, Walsh S, Chesterfield Franks T, Kwok W, et al. Evidence on physical activity and the prevention of frailty and sarcopenia among older people: a systematic review to inform the World Health Organization Physical Activity Guidelines. J Phys Act Health 2020 Aug 11;17(12):1247-1258 [FREE Full text] [doi: 10.1123/jpah.2020-0323] [Medline: 32781432]
9. Cunningham C, O' Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta-analyses. Scand J Med Sci Sports 2020 May;30(5):816-827. [doi: 10.1111/sms.13616] [Medline: 32020713]
10. Physical activity. World Health Organization. URL: https://www.who.int/dietphysicalactivity/pa/en/ [accessed 2024-10-21]
11. Guthold R, Ono T, Strong KL, Chatterji S, Morabia A. Worldwide variability in physical inactivity a 51-country survey. Am J Prev Med 2008 Jun;34(6):486-494. [doi: 10.1016/j.amepre.2008.02.013] [Medline: 18471584]
12. Wu P, Huang K, Chen K, Chou C, Tu Y. Exercise, nutrition, and combined exercise and nutrition in older adults with sarcopenia: a systematic review and network meta-analysis. Maturitas 2021 Mar;145:38-48. [doi: 10.1016/j.maturitas.2020.12.009] [Medline: 33541561]
13. Tieland M, Verdijk LB, de Groot LCPGM, van Loon LJC. Handgrip strength does not represent an appropriate measure to evaluate changes in muscle strength during an exercise intervention program in frail older people. Int J Sport Nutr Exerc Metab 2015 Feb;25(1):27-36. [doi: 10.1123/ijsnem.2013-0123] [Medline: 24903908]
14. Latorre-Román PÁ, Carmona-Torres JM, Cobo-Cuenca AI, Laredo-Aguilera JA. Physical activity, ability to walk, weight status, and multimorbidity levels in older Spanish people: the National Health Survey (2009-2017). Int J Environ Res Public Health 2020 Jun 17;17(12):4333 [FREE Full text] [doi: 10.3390/ijerph17124333] [Medline: 32560442]
15. Tudor-Locke C, Craig CL, Aoyagi Y, Bell RC, Croteau KA, De Bourdeaudhuij I, et al. How many steps/day are enough? For older adults and special populations. Int J Behav Nutr Phys Act 2011 Jul 28;8:80 [FREE Full text] [doi: 10.1186/1479-5868-8-80] [Medline: 21798044]
16. Paluch AE, Gabriel KP, Fulton JE, Lewis CE, Schreiner PJ, Sternfeld B, et al. Steps per day and all-cause mortality in middle-aged adults in the Coronary Artery Risk Development in Young Adults study. JAMA Netw Open 2021 Sep 01;4(9):e2124516 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.24516] [Medline: 34477847]
17. Hall KS, Hyde ET, Bassett DR, Carlson SA, Carnethon MR, Ekelund U, et al. Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. Int J Behav Nutr Phys Act 2020 Jun 20;17(1):78 [FREE Full text] [doi: 10.1186/s12966-020-00978-9] [Medline: 32563261]
18. Kim S, Kim T, Hwang H. The relationship of physical activity (PA) and walking with sarcopenia in Korean males aged 60 years and older using the Fourth Korean National Health and Nutrition Examination Survey (KNHANES IV-2, 3), 2008-2009. Arch Gerontol Geriatr 2013;56(3):472-477. [doi: 10.1016/j.archger.2012.12.009] [Medline: 23298535]
19. Yen H, Chiu H. The effectiveness of wearable technologies as physical activity interventions in weight control: A systematic review and meta-analysis of randomized controlled trials. Obes Rev 2019 Oct;20(10):1485-1493. [doi: 10.1111/obr.12909] [Medline: 31342646]
20. Gualtieri L, Rosenbluth S, Phillips J. Can a free wearable activity tracker change behavior? The impact of trackers on adults in a physician-led wellness group. JMIR Res Protoc 2016 Nov 30;5(4):e237 [FREE Full text] [doi: 10.2196/resprot.6534] [Medline: 27903490]
21. Naslund JA, Aschbrenner KA, Bartels SJ. Wearable devices and smartphones for activity tracking among people with serious mental illness. Ment Health Phys Act 2016 Mar;10:10-17 [FREE Full text] [doi: 10.1016/j.mhpa.2016.02.001] [Medline: 27134654]
22. Yen H, Liao Y, Huang H. Smart wearable device users' behavior is essential for physical activity improvement. Int J Behav Med 2022 Jun;29(3):278-285. [doi: 10.1007/s12529-021-10013-1] [Medline: 34363130]
23. Baig MM, Afifi S, GholamHosseini H, Mirza F. A systematic review of wearable sensors and IoT-based monitoring applications for older adults - a focus on ageing population and independent living. J Med Syst 2019 Jun 15;43(8):233. [doi: 10.1007/s10916-019-1365-7] [Medline: 31203472]
24. Gal R, May A, van Overmeeren EJ, Simons M, Monninkhof E. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. Sports Med Open 2018 Sep 03;4(1):42 [FREE Full text] [doi: 10.1186/s40798-018-0157-9] [Medline: 30178072]
25. Case MA, Burwick HA, Volpp KG, Patel MS. Accuracy of smartphone applications and wearable devices for tracking physical activity data. JAMA 2015 Feb 10;313(6):625-626. [doi: 10.1001/jama.2014.17841] [Medline: 25668268]
26. Puri A, Kim B, Nguyen O, Stolee P, Tung J, Lee J. User acceptance of wrist-worn activity trackers among community-dwelling older adults: mixed method study. JMIR Mhealth Uhealth 2017 Nov 15;5(11):e173 [FREE Full text] [doi: 10.2196/mhealth.8211] [Medline: 29141837]
27. Kwan R, Salihu D, Lee P, Tse M, Cheung D, Roopsawang I, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020;17:7 [FREE Full text] [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
28. Blom V, Drake E, Kallings LV, Ekblom MM, Nooijen CFJ. The effects on self-efficacy, motivation and perceived barriers of an intervention targeting physical activity and sedentary behaviours in office workers: a cluster randomized control trial. BMC Public Health 2021 Jun 02;21(1):1048 [FREE Full text] [doi: 10.1186/s12889-021-11083-2] [Medline: 34078342]

XSL•FO
RenderX

29. McMullan I, Bunting B, Blackburn N, Wilson J, Deidda M, Caserotti P, SITLESS group. The mediating role of self-regulation and self-efficacy on physical activity change in community-dwelling older adults (≥65 years): an experimental cross-lagged analysis using data from SITLESS. J Aging Phys Act 2021 Jun 16;29(6):931-940. [doi: 10.1123/japa.2020-0322] [Medline: 34135127]
30. Patel MS, Asch DA, Volpp KG. Wearable devices as facilitators, not drivers, of health behavior change. JAMA 2015 Feb 03;313(5):459-460. [doi: 10.1001/jama.2014.14781] [Medline: 25569175]
31. Chaudhry UAR, Wahlich C, Fortescue R, Cook DG, Knightly R, Harris T. The effects of step-count monitoring interventions on physical activity: systematic review and meta-analysis of community-based randomised controlled trials in adults. Int J Behav Nutr Phys Act 2020 Oct 09;17(1):129 [FREE Full text] [doi: 10.1186/s12966-020-01020-8] [Medline: 33036635]
32. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020 Dec 09;22(12):e21687 [FREE Full text] [doi: 10.2196/21687] [Medline: 33295292]
33. McLellan AG, Slaght J, Craig CM, Mayo A, Sénéchal M, Bouchard DR. Can older adults improve the identification of moderate intensity using walking cadence? Aging Clin Exp Res 2018 Jan;30(1):89-92. [doi: 10.1007/s40520-017-0746-3] [Medline: 28378230]
34. Nakahara S, Takasaki M, Abe S, Kakitani C, Nishioka S, Wakabayashi H, et al. Aggressive nutrition therapy in malnutrition and sarcopenia. Nutrition 2021 Apr;84:111109. [doi: 10.1016/j.nut.2020.111109] [Medline: 33453622]
35. Harris T, Kerry SM, Victor CR, Ekelund U, Woodcock A, Iliffe S, et al. A primary care nurse-delivered walking intervention in older adults: PACE (pedometer accelerometer consultation evaluation)-Lift cluster randomised controlled trial. PLoS Med 2015 Feb;12(2):e1001783 [FREE Full text] [doi: 10.1371/journal.pmed.1001783] [Medline: 25689364]
36. Rowley TW, Lenz EK, Swartz AM, Miller NE, Maeda H, Strath SJ. Efficacy of an individually tailored, internet-mediated physical activity intervention in older adults: a randomized controlled trial. J Appl Gerontol 2019 Jul;38(7):1011-1022. [doi: 10.1177/0733464817735396] [Medline: 29165018]
37. Santanasto AJ, Glynn NW, Lovato LC, Blair SN, Fielding RA, Gill TM, LIFE Study Group. Effect of physical activity versus health education on physical function, grip strength and mobility. J Am Geriatr Soc 2017 Jul;65(7):1427-1433 [FREE Full text] [doi: 10.1111/jgs.14804] [Medline: 28221668]
38. Wanigatunga AA, Tudor-Locke C, Axtell RS, Glynn NW, King AC, McDermott MM, et al. Effects of a long-term physical activity program on activity patterns in older adults. Med Sci Sports Exerc 2017 Nov;49(11):2167-2175 [FREE Full text] [doi: 10.1249/MSS.0000000000001340] [Medline: 29045323]
39. Yamada M, Nishiguchi S, Fukutani N, Aoyama T, Arai H. Mail-based intervention for sarcopenia prevention increased anabolic hormone and skeletal muscle mass in community-dwelling Japanese older adults: the INE (Intervention by Nutrition and Exercise) Study. J Am Med Dir Assoc 2015 Aug 01;16(8):654-660. [doi: 10.1016/j.jamda.2015.02.017] [Medline: 25858281]
40. Veen J, Montiel-Rojas D, Kadi F, Nilsson A. Effects of Reallocating time spent in different physical activity intensities on sarcopenia risk in older adults: an isotemporal substitution analysis. Biology (Basel) 2022 Jan 10;11(1):111 [FREE Full text] [doi: 10.3390/biology11010111] [Medline: 35053109]

Abbreviations

ANCOVA: analysis of covariance
GEE: generalized estimating equation
IPAQ-SF: International Physical Activity Questionnaire–Short Form
LPA: light-intensity physical activity
MD: mean difference
MET: metabolic equivalent of task
MVPA: moderate-to-vigorous-intensity physical activity
PA: physical activity
SMI: skeletal muscle index
WAT: wearable activity tracker
WC: waist circumference

XSL•FO
RenderX

Edited by YH Lin; submitted 03.05.24; peer-reviewed by MC Uchida, J Veen; comments to author 08.08.24; revised version received 10.09.24; accepted 12.10.24; published 01.11.24.

Please cite as:
Ho MH, Peng CY, Liao Y, Yen HY
Efficacy of a Wearable Activity Tracker With Step-by-Step Goal-Setting on Older Adults' Physical Activity and Sarcopenia Indicators: Clustered Trial
J Med Internet Res 2024;26:e60183
URL: https://www.jmir.org/2024/1/e60183
doi:10.2196/60183
PMID:

©Mu-Hsing Ho, Chi-Yuan Peng, Yung Liao, Hsin-Yen Yen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of a Mobile Health Approach to a Long-Lasting Insecticidal Net Uptake Intervention for Malaria Prevention Among Pregnant Women in Tanzania: Process Evaluation of the Hati Salama (HASA) Randomized Controlled Trial Study

Trinity Vey[1], BScH, MSc; Eleonora Kinnicutt[2], BA, MPH; Nicola West[1,3], BScN, MSc; Jessica Sleeth[1], BA, BPHE, MPH; Kenneth Bernard Nchimbi[4], BSc, PGDip; Karen Yeates[1,3], BSc, MD, MPH

[1]Department of Medicine, Queen's University, Kingston, ON, Canada

[2]Management Sciences for Health, Arlington, VA, United States

[3]Pamoja Tunaweza Women's Centre, Moshi, United Republic of Tanzania

[4]Tanzania Education Network/Mtandao wa Elimu Tanzania, Dar es Salaam, United Republic of Tanzania

Corresponding Author:
Trinity Vey, BScH, MSc
Department of Medicine
Queen's University
99 University Avenue
Kingston, ON, K7L 3P5
Canada
Phone: 1 6135332000
Email: 16tv7@queensu.ca

Abstract

Background: Malaria infection is associated with many adverse outcomes for pregnant women and neonates, yet pregnant women in East and Southern Africa remain frequently exposed to malaria. Long-lasting insecticidal nets (LLINs) can help prevent malarial infections and the associated adverse events. The Hati Salama (HASA) study was a cluster-randomized controlled trial implemented in 100 antenatal health facilities in urban and rural settings of Tanzania that provided pregnant women in both intervention and control groups with e-vouchers to redeem for LLINs for malaria prevention. The intervention group received behavior change communication mobile messages across a 14-day period while the e-voucher was active, and no significant difference between the rates of e-voucher redemption was found across the two groups.

Objective: This study was a process evaluation of the HASA randomized controlled trial to determine barriers and facilitators to e-voucher reception and LLIN acquisition for pregnant women enrolled in the trial, as well as challenges and lessons learned by nurses who worked at the antenatal health facilities supporting the trial.

Methods: Following the e-voucher's expiration at 14 days, voluntary phone follow-up surveys were conducted for nurses who supported the trial, as well as participants in both intervention and control groups of the trial who did not redeem their e-vouchers. Survey questions asked nurses about workflow, training sessions, network connectivity, proxy phone use, and more. Surveys asked participants about reasons for not redeeming e-vouchers. Both surveys provided lists of preset answers to questions, as well as the option to provide open-ended responses. Nurses and trial participants were contacted between January and June 2016 on up to three occasions.

Results: While nurses who supported the HASA trial seemed to recognize the value of the program in their communities, some barriers identified by nurses included network connectivity, workload increase, inadequate training and on-the-ground support, and difficulty following the workflow. Several barriers identified by trial participants included personal obligations preventing them from redeeming the e-voucher on time, network connectivity issues, losing the e-voucher number, no stock of LLINs at retailers when attended, inadequate explanation of where or how to redeem the e-voucher, or not receiving an SMS text message with the e-voucher number promptly or at all.

Conclusions: Large-scale e-voucher platforms for health-related commodity interventions, such as LLIN distribution in sub-Saharan Africa, are feasible, but challenges, including network connectivity, must be addressed. Nurses identified issues to be considered in a future scale-up, such that the number of nurses trained should be increased and the e-voucher issuance workflow

XSL•FO
RenderX

should be simplified. To address some of the key barriers impacting e-voucher redemption for trial participants, the network of retailers could be expanded and the e-voucher expiration period should be extended.

Trial Registration: ClinicalTrials.gov NCT02561624; https://clinicaltrials.gov/ct2/show/NCT02561624

(J Med Internet Res 2024;26:e51527) doi:10.2196/51527

KEYWORDS

mHealth; short message service; behavior change communication; pregnancy; long-lasting insecticidal nets; malaria; protozoan infections; parasitic diseases; vector borne diseases; insecticide; intervention; malaria prevention

Introduction

Malaria remains a significant cause of morbidity and mortality in sub-Saharan Africa [1]. Malaria infection in pregnant women is associated with a variety of adverse outcomes including miscarriage, neonate death, low birth weight, and maternal anemia and death [2]; yet in 2020, 22% of pregnant women in East and Southern Africa were exposed to malaria [1]. The use of insecticide-treated nets (ITNs) by pregnant women in malaria-endemic regions of Africa has shown decreased outcomes of placental malaria, low birth weight, and fetal loss [3]. As such, ITNs, and specifically, long-lasting insecticidal nets (LLINs), are a core malaria prevention tool, recommended by the World Health Organization [4].

In attempts to increase LLIN use in Tanzania, the Tanzania National Voucher Scheme (TNVS) provided pregnant women and women with young children with vouchers at antenatal clinics that could be redeemed for LLINs at a discounted price [5]. At the program's inception in 2004, vouchers were distributed in paper format, until 2011 when electronic "e-vouchers" were implemented. Despite the high risk of maternal and neonate morbidity and mortality from malaria, findings suggest that only 47% of women who attended an antenatal clinic subsequently received a net through the TNVS [5].

Seeking to address this gap, the Hati Salama (HASA) study aimed to increase LLIN uptake in pregnant women [6]. The HASA study was a cluster-randomized controlled trial (RCT) in which pregnant women older than 13 years of age in Tanzania were issued e-vouchers through nurses at antenatal clinics that could subsequently be redeemed for LLINs at a subsidized price at a local retailer, where LLINs were stored and distributed. The original cost of an LLIN was approximately US $8, but at a retailer registered with the intervention, participants with an e-voucher only would pay approximately US $0.25 [6]. Throughout the 14-day period in which the e-voucher was active, the intervention group of women received behavior change communication SMS text messages emphasizing the importance of LLIN use during pregnancy and reminding them to redeem their e-vouchers. While there was no significant increase in LLIN uptake for the pregnant women in the intervention group compared to the control group that did not receive SMS text messages, with a total of 3746 e-vouchers redeemed, the HASA study demonstrated that LLIN e-voucher distribution via nurses in antenatal clinics is feasible on a large scale.

Findings from several studies using mobile health (mHealth) as a tool for malaria prevention and treatment have found that SMS text messages were well accepted by recipients [7-9]. However, there is a paucity of literature investigating the barriers to mobile messaging or e-voucher schemes from the perspective of both study recipients and administrators for malaria prevention.

This paper presents the mixed methods evaluation of the HASA study that was designed to assess perceived barriers to e-voucher redemption and LLIN acquisition through follow-up surveys administered to nurses and study participants of the HASA study. In doing so, we highlight challenges, unintended findings, and lessons learned to ensure effective scale-up and implementation of the e-voucher platform for future health-related commodity interventions such as LLIN distribution in sub-Saharan Africa.

Methods

The HASA RCT

The HASA study was a cluster-randomized controlled double-blind 2-arm study implemented between July and December 2015, in 100 antenatal health facilities in urban and rural settings across 21 regions of Tanzania (ClinicalTrials.gov ID NCT02561624) [6]. There were 50 clusters randomly allocated to the intervention and 50 clusters to the control group. Participants were pregnant women older than 13 years of age attending one of the selected antenatal health facilities, who were either literate or living with someone who could assist them with reading SMS text messages. This research study presents the mixed methods evaluation of the implementation of the HASA cluster RCT. This mixed methods evaluation took place via nurse and participant surveys, described in more detail below, between January and June 2016.

Staff Training

Prior to the implementation of the HASA RCT, centralized meetings were organized with all relevant medical stakeholders from enrolled antenatal clinics including Regional Medical Officers, District Medical Officers, and Clinic Officers-in-Charge. The objectives of these meetings, led by one representative from Queen's University and one from Mennonite Economic Development Associates (MEDA) Tanzania, were to present the HASA trial, answer any questions or concerns, and define the stakeholder's involvement during the implementation and follow-up period of the trial. The meetings also provided an opportunity to obtain informed consent from all medical and clinic officers agreeing to have

XSL•FO
RenderX

their health facility participate in the study and adhere to all ethical procedures of the trial.

Medical officers from each enrolled health facility selected two nurses from the facility's antenatal department for training and participation in the HASA trial. Three nurse training sessions and three retailer training sessions were held between June 1 and June 8, 2015, in the Tanzanian cities of Mwanza, Dodoma, and Dar es Salaam. Sessions were attended by more than 180 nurses from the 100 health facilities selected for the study, along with 28 retailers. For facility sites for which only one nurse could attend the training, this nurse subsequently trained the second nurse involved at their site. The objectives of the nurse training were to introduce the HASA trial, explain and review the role and participation of nurses in the trial, provide training regarding obtaining informed consent and standard operating procedures, and provide an overview of the SMS text messaging workflow. The objectives of the retailer training were to provide a thorough participatory overview of the retailer contract and an overview of the SMS text messaging workflow.

The registered retailers received small incentives for participating in the study; specifically, they received a percentage of the profits for LLINs sold through the program, they received increased foot traffic from women enrolled in the program, and at the end of the program, the retailers were able to sell the unredeemed LLINs for their profit. The retailers involved in the e-voucher program were all involved in the program managed by MEDA which predated the launch of the e-voucher program in the HASA study [6]. The retailers were small to medium-sized local convenience shops known in Tanzania as "Dukas." To be included in the e-voucher program, they were within a 500-m radius of the catchment area of the health facility that was participating in the distribution of the e-vouchers.

Four zonal managers were selected from a pool of highly qualified former TNVS staff recommended by MEDA Tanzania to oversee the fieldwork. Each zonal manager was allotted a cluster of regions (zones) for which they were responsible for registering retailers and nurses prior to study implementation, as well as monitoring and managing the e-voucher workflow (both with regards to e-voucher issuance, as well as LLIN redemption at the retailer) to provide sufficient support, technical assistance, and retraining throughout the course of the study. Zonal managers were also responsible for data collection via iForm and providing weekly reports from facilities regarding women enrolled, vouchers issued, and supply of nets to retailers.

Nurse Follow-Up Survey

Following the e-voucher expiration period of 14 days, a structured follow-up survey was conducted for nurses who worked at the antenatal health facilities enrolled in the trial. The survey was conducted directly over the phone by the HASA field team members. Nurses were contacted on up to 3 occasions with the option to participate in the survey. Nurses were asked a series of yes or no questions in addition to 2 open-ended questions (4b and 11), as listed in Table S1 in Multimedia Appendix 1. Responses were deidentified and uploaded into an iForm. A total of 180 antenatal nurses were involved in the HASA study and contacted to be asked to participate in this phone survey, 67 of which did participate. Nurses from across the 21 regions of Tanzania involved in the study were contacted.

Participant Follow-Up Survey

Following the e-voucher expiration period, all participants (in both intervention and control groups) who did not redeem their e-vouchers for an LLIN at a registered retailer were contacted by a HASA field team member by phone on up to 3 occasions to complete a voluntary structured follow-up survey over the phone. The survey asked participants to indicate the reason they did not redeem the e-voucher prior to its expiration. The survey included a list of preset options for participants to choose from, along with the option to provide an open-ended response for why the voucher was not redeemed, as listed in Table S2 in Multimedia Appendix 1. Responses were deidentified and uploaded into an iForm. Of the 1702 participants of the HASA trial from across 21 regions of Tanzania who did not redeem their e-vouchers and were thus contacted to be asked to participate in this phone survey, 410 participants did participate.

Ethical Considerations

This study was conducted as a follow-up process evaluation under the HASA study, which received ethical approval from the National Institute of Medical Research Tanzania and Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (6013000). Ethical approval to assess the feasibility, acceptability, and adoptability of the HASA intervention among nurses and women who participated in the intervention via surveys was obtained under the approval by the National Institute of Medical Research and Health Sciences (original certificate number NIMR/HQ/R.8a/Vol. IX/1750) and Affiliated Teaching Hospitals Research Ethics Board for the initial HASA RCT study. Phone surveys administered to both nurses and intervention participants began with a statement about their purpose and informed consent. Verbal consent was obtained, and confidentiality and anonymity were explained to participants.

Results

Of the 180 antenatal clinic nurses involved in the study, 67 nurses provided answers to the phone survey following the e-voucher redemption period, which are listed in Table 1. The majority responded affirmatively that they felt HASA was helpful in improving the overall community understanding of malaria (n=65, 97%), and that women were eager and easily able to redeem for a net (n=65, 97%). All nurses surveyed (n=67, 100%) felt that their community needs programs like HASA. Only 67% (n=45) of nurses felt that the workflow was easy to follow, and only 37% (n=25) of nurses felt that the training and support provided were adequate. The majority (n=62, 93%) of nurses identified that they did experience challenges regarding network and connectivity, of which 92% (n=56) of nurses felt these issues impacted the number of vouchers issued or redeemed. While 91% (n=61) of nurses felt that using a proxy phone was helpful, just over half (n=35, 52%) of nurses experienced challenges finding a reliable proxy phone. A total of 73% (n=49) of nurses felt that involvement in HASA heavily increased their workload, and most (n=64, 96%) nurses agreed

XSL•FO
RenderX

that having more nurses in their facility trained as issuers would have been helpful.

Responses to the open-ended question to nurses "If (you did not find that the workflow [to issue vouchers] was easy to understand and follow), what would you change?" (4b from Table S1 in Multimedia Appendix 1) are listed in Table 2. These answers were all centered around shortening and simplifying either the workflow or the SMS text messages.

Responses to the open-ended question "In a potential scale-up, what are your recommendations for improving the program?" (11 from Table S1 in Multimedia Appendix 1), which were suggested by nurse respondents in the survey are listed in Table 3. Nurses were able to provide up to three suggestions each. Many responses suggested that network connectivity should be improved (n=38, 28.5%), more nurses should be trained to issue vouchers (n=27, 20%), and workflow should be shortened (n=10, 7.5%).

In total, of the 1702 participants who participated in the trial and did not redeem their e-vouchers, 410 participants completed the phone survey following the e-voucher redemption period, indicating the reason they did not redeem the e-voucher [6]. The answers to this survey are listed in Table 4. Common reasons that participants did not redeem their voucher included that they lost their voucher number (n=64, 16%), they went to the shop, but the shopkeeper did not use the e-voucher system (n=54, 13%), or they went to redeem the e-voucher, but the store had no nets available (n=23, 6%). Over half (n=217, 53%) of participants who responded to the survey indicated that there was an "other reason" they did not redeem the e-voucher.

The open-ended responses detailing other reasons that participants did not redeem their e-voucher for an LLIN are listed in Table 5. Participants were able to provide multiple reasons. There is some overlap between the responses detailed in Table 5 and those in the initial participant survey (Table 4), however, most of the open-ended responses were distinct from the initial survey. The most common of these was that participants could not redeem their e-voucher on time due to a variety of other obligations or health concerns (n=90, 35%), participants did not receive an SMS text message with the e-voucher at all or on time (n=38, 14.8%), and participants did not receive an adequate explanation of the HASA program and e-voucher redemption (n=26, 10.1%).

XSL•FO
RenderX

Table 1. Responses to the nurse follow-up survey.

Questions and responses	Count	Percentage
1. Did you find HASA[a] to be helpful in improving the overall perception and understanding of malaria and the importance of prevention? Both with regards to the bed net component as well as messages?		
Yes	65	97
No	2	3
2. Do you find that this is a program that your community needs?		
Yes	67	100
No	0	0
3. Do you find that your patients were eager to and easily able to redeem a net?		
Yes	65	97
No	2	3
4. Do you find that the workflow (to issue vouchers) was easy to understand and follow?		
Yes	45	67
No	22	33
5. Did you find the training sessions and on-the-ground support adequate?		
Yes	25	37
No	42	63
6a. Did you encounter challenges with regard to network or connectivity when issuing vouchers?		
Yes	62	93
No	5	7
6b. If so, did this affect the number of vouchers you issued or redemption by beneficiaries?		
Yes	56	92
No	5	8
7. Considering your catchment population, was it challenging to register women with phones?		
Yes	40	60
No	27	40
8a. Did you find that using a proxy phone was helpful?		
Yes	61	91
No	6	9
8b. Did you encounter challenges with finding a reliable proxy phone?		
Yes	35	52
No	32	48
9. Did you find that your involvement in HASA increased your workload heavily?		
Yes	49	73
No	18	27
10. Do you think it would have been helpful to have had more nurses trained as issuers in your facility?		
Yes	64	96
No	3	4

[a]HASA: Hati Salama.

XSL•FO
RenderX

Table 2. Responses from nurse survey to open-ended question 4b "If you did not find that the workflow was easy to understand and follow, what would you change?"

Response	Count (n=14)	Percentage
Workflow should be shortened and simplified	10	72
Only one SMS text message should be sent	3	21
SMS text messages should be shortened	1	7

Table 3. Responses from nurse survey to open-ended question 11 "In a potential scale-up, what are your recommendations for improving the program?"

Responses	Count (n=134)	Percentage
Network connectivity should be improved	38	28.5
More nurses or clinic staff should be trained to issue vouchers	27	20
Workflow should be shortened	10	7.5
LLINs[a] should be available within, or closer to, the health facility	8	6
An alternative way to issue vouchers to participants without phones should be considered	7	5
More training sessions for voucher issuers is needed	6	4.5
Staff bonus or motivation for voucher issuers should be considered and given on time	6	4.5
The health care facility or program should provide a phone that can be used for voucher issuance	6	4.5
The number of forms that voucher issuers need to fill out should be reduced	5	4
Nurses should receive a message with the voucher number to confirm that clients have received the voucher	4	3
The retailer shops should have a greater stock of LLINs	4	3
LLINs should also be issued to children younger than 5 years	3	2
Women without phones should continue to use a proxy phone	2	1.5
There should be more than one retailer associated with each health facility	2	1.5
LLINs and vouchers should be issued at regular mobile clinics	1	0.75
One SMS text message during issuance is enough	1	0.75
There should be posters that explain the program	1	0.75
LLINs of different sizes and colors should be available	1	0.75
LLIN mesh size should be smaller	1	0.75
Vouchers should be available more promptly	1	0.75

[a]LLIN: long-lasting insecticidal net.

Table 4. Participant follow-up survey responses.

	Responses	Count (n=410)	Percentage
1	I lost my voucher number	64	16
2	My voucher number was not valid when I took it to the shop	9	2
3	I went to redeem the voucher and the store had no nets available	23	6
4	I do not want a net	10	2
5	I have all the bed nets I need	3	1
6	I cannot afford to pay the additional payment for the net	12	3
7	I do not live close enough to the shops that are in the Hati Punguzo program	18	4
8	I went to the shop, but the shopkeeper failed to use the e-voucher system	54	13
9	Other reason	217	53

XSL•FO
RenderX

Table 5. Answers to participant follow-up survey provided for "other reason."

Responses	Count (n=257)	Percentage
Participant was not able to redeem e-voucher prior to expiration due to illness, giving birth, travel, or other obligations	90	35
Participant never received an SMS text message with the e-voucher number, or the SMS text message did not arrive in time to redeem	38	14.8
Participant did not receive an adequate explanation about how or where to redeem the e-voucher	26	10.1
Issue related to proxy phone use (such as the owner of the phone was traveling)	18	7
Retailer shop was closed when the participant tried to redeem	16	6.2
The phone was lost, stolen, or damaged	12	4.7
No particular reason indicated	9	3.5
Forgot to redeem or negligence	8	3.1
Network connectivity issues at issuance or redemption of the e-voucher	7	2.7
Retailer shops are too far from where the participant lives	6	2.3
No instruction was given at the retail shop or the retailer was not informed about HASA[a]	5	1.9
The SMS text message was deleted from the phone	5	1.9
The husband did not understand the HASA program	3	1.2
Confusion about the expiry date or unaware it expired	3	1.2
LLIN[b] stock issue	2	0.8
Participant already had sufficient LLINs	2	0.8
Did not like the retailer's attitude	2	0.8
The SMS text message was sent to the wrong number	2	0.8
Did not want white-colored LLIN	1	0.4
LLIN described as too stiff	1	0.4
Participant did not bring antenatal clinic card to the retailer staff and they would not sell them the LLIN	1	0.4

[a]HASA: Hati Salama.

[b]LLIN: long-lasting insecticidal net.

Discussion

Principal Results

This process evaluation study sought to understand the challenges experienced during the HASA trial by both participants who received e-vouchers to redeem for LLINs and the nurses at antenatal clinics who issued these vouchers.

The responses to the nurse follow-up survey indicated that nurses recognized the value of the HASA program, with 97% (n=65) of nurses responding that it was helpful in improving the overall perception and understanding of malaria and the importance of prevention. Similarly, 100% (n=67) of nurses surveyed felt that HASA is a program that their community needs. This positive reception of the program supports the finding from the HASA primary outcome paper that nurses likely played a role in motivating participants and encouraging LLIN uptake in both control and intervention groups [6]. This also suggests that such interventions surrounding LLIN use in pregnant women will be well-received and supported by community nurses.

The nurse survey did identify several key process barriers, with 33% (n=22) of nurses responding that the workflow was not easy to follow, and 63% (n=42) of nurses responding that the training sessions and on-the-ground support were not adequate. These findings were emphasized during the open-ended question responses, with suggestions focused on shortening workflow (n=10, 7.5%), reducing the number of forms for nurses to fill out (n=5, 4%), and providing more training sessions (n=6, 4.5%). Together, this suggests that the workflow and training for staff should be revised and simplified and that perhaps zonal managers should be empowered to provide greater on-demand support and retraining. It is also possible that since 33% (n=22) of nurses found the workflow not easy to follow, this may have impacted their explanations of the program to participants and subsequent redemption; 10% (n=26) of participants did not redeem their e-voucher due to feeling that they did not receive an adequate explanation of the program. As such, simplifying workflow in the future may not only improve nurse satisfaction with the program but may increase reception.

A total of 73% (n=49) of nurses surveyed felt that HASA increased their workload heavily, with 96% (n=64) of nurses agreeing that having more nurses trained as issuers would have

XSL•FO
RenderX

been helpful. During the implementation stage of the HASA trial, the ongoing absence of nurses from their postings due to conflicting obligations (work seminars and commitments, family obligations, illness, etc) was noted by the HASA field staff. Such understaffing of voucher issuers likely contributed to many nurses feeling that their workload was increased. Together, this demonstrates the need to train more staff to issue vouchers, including junior nurses with lower levels of additional responsibilities, who will often be present in the facility. Providing a staff bonus or additional incentive, as recommended by 4.5% (n=6) of respondents, may also help decrease the feeling of being overworked.

Comparison With Prior Work

An additional recommendation by 6% (n=8) of nurses was that LLINs should be sold within the antenatal clinic itself. This is consistent with suggestions from clinic staff in a similar Tanzania ITN study, who felt that the purchase of discounted nets should only take place at the clinic [10]. While this may improve redemption, an overall strength of the HASA study was that the retailers were generally in close proximity to the antenatal clinics, allowing some women to redeem vouchers immediately following their visit. The geographical accessibility of the retailers did not seem to present an especially large challenge to redemption for participants. Some nurses also suggested that only one SMS text message, or shortened messages, should be sent to participants. This is reflective of the literature on mHealth interventions in Ghana and Kenya showing that a short simple message can be as effective as sending longer or multiple messages and therefore could be considered for the scale-up of such a mHealth initiative [11,12].

A major challenge identified by both nurses and participants was issues surrounding network and connectivity. A total of 93% (n=62) of nurses encountered network challenges, the majority of which impacted voucher issuance or redemption. It is possible that network issues may be amplified for women of lower socioeconomic status, who may be more likely to live in an area with unreliable cellular connectivity. Women of lower socioeconomic status have already been found to have lower ITN voucher uptake compared to those of higher socioeconomic status [13,14]. This reflects the most common suggestion in the literature from participants of an ITN voucher scheme in Tanzania that the monetary value of vouchers should be increased [10]. In rural areas, which are typically already subject to unreliable connectivity, research suggests that women are less likely to be the sole owners of phones, which may cause additional barriers through the use of proxy phones [15]. In a future scale-up, network reliability may present less of an issue, as cellular coverage continues to increase in rural areas of Tanzania and sub-Saharan Africa. A technical backup system could also be implemented in the future to avoid issues with e-voucher issuance or redemption due to network connectivity.

One barrier to nonredemption identified by 6% (n=23) of participants was that the retailer had no nets available when they went to redeem the voucher. This supports a similar recommendation from the nurse survey that retailer shops should have a greater stock of LLINs. For the HASA study, the stock of LLINs was delivered to local retailers at multiple phases during the program with numbers of LLINs provided based on historical data and the number of participants each site was expected to enroll. Despite retailers signing a contract that the LLINs were intended to be exclusively used for the HASA program, some retailers may have sold stock to other customers. Particularly since part of the incentive for retailers to participate in the intervention was that they were able to sell unredeemed LLINs following the trial for profit, it is possible that retailers intentionally or unintentionally sold LLINs outside the trial before it ended, thus impacting stock. Concerns around fraudulent e-voucher creation by retailers' staff to sell LLINs on the commercial market were in part addressed by an e-voucher redesign; security for e-vouchers was increased by including a unique identifier number for all pregnant women, random barcode activation at the retailer, and reports to verify the number of e-vouchers issued by clinics. A similar voucher scheme for ITNs in Ghana also found that inadequate stock of ITNs was the greatest issue faced by retailers [16]. An additional barrier encountered at retailers was that 13% (n=54) of participants attended a shop to redeem their e-vouchers, but the shopkeeper did not recognize the e-voucher system. This could potentially be overcome in a scale-up by expanding the network of retailers using the e-voucher system.

While a Ghana voucher scheme found that the major reason (36%) for women not accepting a voucher at an antenatal clinic was that they already had a net, this was a minor reason for nonredemption in the HASA study [16]. Only 1% (n=3) of women stated that they already had all the nets they needed, suggesting that despite the need for LLINs, there were other barriers preventing the redemption of e-vouchers. Another reason identified preventing redemption for 10% (n=26) of participants was that they did not receive a sufficient explanation about how to redeem the voucher. This is consistent with suggestions from participants of another ITN voucher scheme in Tanzania that more information about the voucher scheme, malaria, and nets is required [10]. Another similar study suggests that a barrier to receiving vouchers at antenatal clinics was that clinic coordinators were not familiar enough with the voucher system [17]. More thorough training and regular retraining of nurses to ensure they are providing adequate explanations of the e-voucher scheme to women may alleviate this barrier. A total of 16% (n=64) of women indicated that they lost their voucher number, which may have been due to proxy phone use if another user of the phone deleted the SMS text message with the voucher code. It is also possible that some participants could not find the SMS text message on their phone at the time they went to redeem it, or that the message may have been deleted in error.

The main reason provided that participants did not redeem their e-voucher (n=90, 35%) reflected that the voucher expired before the participant had a chance to redeem it; this barrier could be resolved by extending the expiration period beyond 14 days. In 2011, prior to the switch from paper vouchers to e-vouchers, the TNVS vouchers had a validity of 60 days [5]. Extending the e-voucher validity period or giving participants the option to reregister if the deadline is missed may improve redemption.

XSL•FO
RenderX

Limitations

A potential limitation of this study was that surveys of nurses and participants were completed over the phone. While this decreased costs and travel time associated with in-person interviews, potentially facilitating responses from rural participants, this method likely limited the responses from those without access to their phones. Despite the frequency of yes and no questions and preset answers, nurses and participants were given the opportunity to provide their open-ended recommendations, enriching the data collected. Due to the survey format, there was an element of interpretation required for the open-ended responses given by both nurses and participants. Despite providing up to 3 opportunities for both nurses and intervention participants to participate in these surveys, response rates to the surveys were 67 of 180 nurses and 410 of 1702 study participants, which suggests that there are likely challenges and facilitators of the HASA study experienced by nurses and participants that are not captured in this paper.

Conclusions

Despite these identified challenges, the HASA study was a large robust trial with an overall positive perception by involved nurses, suggesting that large e-voucher platforms for health-related commodity interventions, such as LLIN distribution in sub-Saharan Africa, are feasible [6]. In a scale-up of HASA or similar programs, many of the issues identified by nurses could be overcome by facilitating the nurse workflow and training a greater number of nurses. To address some of the major barriers faced by participants, we suggest extending the validity of the e-voucher and expanding the network of retailers selling LLINs for the program. Improving network connectivity would facilitate the HASA program for both nurses and participants.

Acknowledgments

The authors acknowledge Michael Sarco, Thom Dixon, Moses Mokua, and the Mennonite Economic Development Associates for their support with this project. This project was supported by Grand Challenges Canada. Grand Challenges Canada is funded by the Government of Canada and is dedicated to supporting Bold Ideas with Big Impact.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

KY developed this research question, designed the analyses, and supervised the conduct of the HASA study. TV had primary responsibility for the writing of the report and completed data analysis. EK, JS, NW, and KN were involved in the coordination of the HASA study at the project office or in Tanzania and provided comments during the writing and editing of the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File , 16 KB - jmir_v26i1e51527_app1.docx]

References

1. World malaria report 2021. WHO. Geneva URL: https://www.who.int/publications-detail-redirect/9789240040496 [accessed 2022-05-27]
2. Schantz-Dunn J, Nour NM. Malaria and pregnancy: a global health perspective. Rev Obstet Gynecol 2009;2(3):186-192. [Medline: 19826576]
3. Gamble CL, Ekwaru J, Kuile FO. Insecticide‐treated nets for preventing malaria in pregnancy. Cochrane Database of Systematic Reviews. 2006. URL: http://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD003755.pub2/full [accessed 2022-05-27]
4. Long-lasting insecticidal nets (LLINs). WHO. URL: https://www.who.int/teams/immunization-vaccines-and-biologicals/essential-programme-on-immunization/integration/linking-with-other-health-interventions/long-lasting-insecticidal-nets-(llins) [accessed 2022-12-08]
5. Kramer K, Mandike R, Nathan R, Mohamed A, Lynch M, Brown N, et al. Effectiveness and equity of the Tanzania national voucher scheme for mosquito nets over 10 years of implementation. Malar J 2017;16(1):255 [FREE Full text] [doi: 10.1186/s12936-017-1902-0] [Medline: 28619076]

XSL•FO
RenderX

6. Vey T, Kinnicutt E, Day AG, West N, Sleeth J, Nchimbi KB, et al. A cluster-randomized controlled trial of targeted behaviour change communication using a mobile health platform to increase uptake of long-lasting insecticidal nets among pregnant women in Tanzania: the Hati Salama "Secure Voucher" (HASA) study. J Med Internet Res 2024 (forthcoming).
7. Ollivier L, Romand O, Marimoutou C, Michel R, Pognant C, Todesco A, et al. Use of short message service (SMS) to improve malaria chemoprophylaxis compliance after returning from a malaria endemic area. Malar J 2009;8:236 [FREE Full text] [doi: 10.1186/1475-2875-8-236] [Medline: 19852811]
8. Svege S, Kaunda B, Robberstad B, Nkosi-Gondwe T, Phiri KS, Lange S. Post-discharge malaria chemoprevention (PMC) in Malawi: caregivers' acceptance and preferences with regard to delivery methods. BMC Health Serv Res 2018;18(1):544 [FREE Full text] [doi: 10.1186/s12913-018-3327-z] [Medline: 29996833]
9. Kurumop SF, Bullen C, Whittaker R, Betuela I, Hetzel MW, Pulford J. Improving health worker adherence to malaria treatment guidelines in Papua New Guinea: feasibility and acceptability of a text message reminder service. PLoS One 2013;8(10):e76578 [FREE Full text] [doi: 10.1371/journal.pone.0076578] [Medline: 24116122]
10. Tami A, Mbati J, Nathan R, Mponda H, Lengeler C, Schellenberg JRMA. Use and misuse of a discount voucher scheme as a subsidy for insecticide-treated nets for malaria control in southern Tanzania. Health Policy Plan 2006;21(1):1-9. [doi: 10.1093/heapol/czj005] [Medline: 16301306]
11. Raifman JRG, Lanthorn HE, Rokicki S, Fink G. The impact of text message reminders on adherence to antimalarial treatment in northern Ghana: a randomized trial. PLoS One 2014;9(10):e109032 [FREE Full text] [doi: 10.1371/journal.pone.0109032] [Medline: 25350546]
12. Pop-Eleches C, Thirumurthy H, Habyarimana JP, Zivin JG, Goldstein MP, de Walque D, et al. Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders. AIDS 2011;25(6):825-834 [FREE Full text] [doi: 10.1097/QAD.0b013e32834380c1] [Medline: 21252632]
13. Mushi AK, Schellenberg JRMA, Mponda H, Lengeler C. Targeted subsidy for malaria control with treated nets using a discount voucher system in Tanzania. Health Policy Plan 2003;18(2):163-171. [doi: 10.1093/heapol/czg021] [Medline: 12740321]
14. Marchant T, Schellenberg D, Nathan R, Armstrong-Schellenberg J, Mponda H, Jones C, et al. Assessment of a national voucher scheme to deliver insecticide-treated mosquito nets to pregnant women. CMAJ 2010;182(2):152-156 [FREE Full text] [doi: 10.1503/cmaj.090268] [Medline: 20064944]
15. Vasudevan L, Ostermann J, Moses SM, Ngadaya E, Mfinanga SG. Patterns of mobile phone ownership and use among pregnant women in Southern Tanzania: cross-sectional survey. JMIR Mhealth Uhealth 2020;8(4):e17122 [FREE Full text] [doi: 10.2196/17122] [Medline: 32267240]
16. Kweku M, Webster J, Taylor I, Burns S, Dedzo M. Public-private delivery of insecticide-treated nets: a voucher scheme in Volta region, Ghana. Malar J 2007;6:14 [FREE Full text] [doi: 10.1186/1475-2875-6-14] [Medline: 17274810]
17. Njau RJA, de Savigny D, Gilson L, Mwageni E, Mosha FW. Implementation of an insecticide-treated net subsidy scheme under a public-private partnership for malaria control in Tanzania—challenges in implementation. Malar J 2009;8:201 [FREE Full text] [doi: 10.1186/1475-2875-8-201] [Medline: 19698109]

Abbreviations

HASA: Hati Salama
ITN: insecticide-treated net
LLIN: long-lasting insecticidal net
MEDA: Mennonite Economic Development Associates
mHealth: mobile health
RCT: randomized controlled trial
TNVS: Tanzania National Voucher Scheme

Edited by A Mavragani; submitted 02.08.23; peer-reviewed by R Zhuang, L Ahmed; comments to author 27.08.23; revised version received 31.10.23; accepted 11.04.24; published 05.11.24.

Please cite as:
Vey T, Kinnicutt E, West N, Sleeth J, Nchimbi KB, Yeates K
Implementation of a Mobile Health Approach to a Long-Lasting Insecticidal Net Uptake Intervention for Malaria Prevention Among Pregnant Women in Tanzania: Process Evaluation of the Hati Salama (HASA) Randomized Controlled Trial Study
J Med Internet Res 2024;26:e51527
URL: https://www.jmir.org/2024/1/e51527
doi:10.2196/51527
PMID:

XSL•FO
RenderX

©Trinity Vey, Eleonora Kinnicutt, Nicola West, Jessica Sleeth, Kenneth Bernard Nchimbi, Karen Yeates. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effectiveness of Remote Exercise Rehabilitation Based on the "SCeiP" Model in Homebound Patients With Coronary Heart Disease: Randomized Controlled Trial

Dandan Xu[1*], BD; Dongmei Xu[1*], BD; Lan Wei[1], BD; Zhipeng Bao[1], MD; Shengen Liao[1], PhD; Xinyue Zhang[1*], MD

Department of Cardiology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[*]these authors contributed equally

Corresponding Author:
Xinyue Zhang, MD
Department of Cardiology
The First Affiliated Hospital of Nanjing Medical University
300 Guangzhou Road
Gulou District
Nanjing, 210029
China
Phone: 86 15005162289
Fax: 86 025 68303110
Email: 15005162289@163.com

Abstract

Background: While exercise rehabilitation is recognized as safe and effective, medium- to long-term compliance among patients with coronary heart disease (CHD) remains low. Therefore, promoting long-term adherence to exercise rehabilitation for these patients warrants significant attention.

Objective: This study aims to investigate the impact of remote exercise rehabilitation on time investment and related cognitive levels in homebound patients with CHD. This study utilizes the SCeiP (Self-Evaluation/Condition of Exercise-Effect Perception-Internal Drive-Persistence Behavior) model, alongside WeChat and exercise bracelets.

Methods: A total of 147 patients who underwent percutaneous coronary intervention in the cardiovascular department of a grade III hospital in Jiangsu Province from June 2022 to March 2023 were selected as study participants through convenience sampling. The patients were randomly divided into an experimental group and a control group. The experimental group received an exercise rehabilitation promotion strategy based on the "SCeiP" model through WeChat and exercise bracelets, while the control group followed rehabilitation training according to a standard exercise rehabilitation guide. The days and duration of exercise, levels of cardiac rehabilitation cognition, exercise planning, and exercise input were analyzed before the intervention and at 1 month and 3 months after the intervention.

Results: A total of 81 men (55.1%) and 66 women (44.9%) were recruited for the study. The completion rate of exercise days was significantly higher in the experimental group compared with the control group at both 1 month (t145=5.429, P<.001) and 3 months (t145=9.113, P<.001) after the intervention. Similarly, the completion rate of exercise duration was significantly greater in the experimental group (t145=3.471, P=.001) than in the control group (t145=5.574, P<.001). The levels of autonomy, exercise planning, and exercise input in the experimental group were significantly higher than those in the control group at both 1 month and 3 months after the intervention (P<.001). Additionally, the experimental group exhibited a significant reduction in both process anxiety and outcome anxiety scores (P<.001). Repeated measures ANOVA revealed significant differences in the trends of cognitive function related to cardiac rehabilitation between the 2 patient groups over time: autonomy, F1,145(time×group)=9.055 (P<.001); process anxiety, F1,145(time×group)=30.790 (P<.001); and outcome anxiety, F1,145(time×group)=28.186 (P<.001). As expected, the scores for exercise planning (t145=2.490, P=.01 and t145=3.379, P<.001, respectively) and exercise input (t145=2.255, P=.03 and t145=3.817, P<.001, respectively) consistently demonstrated superiority in the experimental group compared with the control group at both 1 and 3 months after the intervention. Interestingly, we observed that the levels of exercise planning and exercise input in both groups initially increased and then slightly decreased over time, although both remained higher than the preintervention levels (P<.001).

XSL•FO
RenderX

Conclusions: The remote health intervention based on the "SCeiP" model effectively enhances exercise compliance, exercise planning, exercise input, and cognitive levels during cardiac rehabilitation in patients with CHD.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300069463; https://www.chictr.org.cn/showproj.html?proj=192461

(J Med Internet Res 2024;26:e56552) doi:10.2196/56552

KEYWORDS

coronary heart disease; exercise rehabilitation; promotion strategy; home rehabilitation

Introduction

Background

Exercise rehabilitation is a core component of cardiac rehabilitation and plays a vital role in the secondary prevention of coronary heart disease (CHD). It has been shown to significantly reduce adverse cardiovascular events and improve disease prognosis [1]. The center-based rehabilitation model is particularly recognized for its safety and effectiveness in achieving these outcomes [2]. However, the implementation of exercise rehabilitation is often hindered by traffic and medical conditions. Consequently, long-term exercise rehabilitation predominantly occurs in home settings, which presents challenges in ensuring compliance and effectiveness [3]. Home-based rehabilitation typically relies on remote coaching, involving indirect exercise performance, while supervision is conducted in a convenient and acceptable manner [4]. Existing systematic reviews indicate that there are no statistically significant differences in readmission rates or all-cause mortality when comparing the safety profiles of traditional center-based cardiac rehabilitation with those of home-based cardiac rehabilitation [5,6]. However, the participation rate in exercise rehabilitation remains low. Of the approximately 50% of countries worldwide that offer cardiac rehabilitation, only 30% of patients with heart disease participate in these programs [7]. Furthermore, the participation rate for patients with CHD in the general population ranges from 8% to 34% [8], while dropout rates vary from 12% to 56% [9]. Overall, the rate of participation in exercise rehabilitation remains low [10,11]. When extrinsic structural barriers, such as work obligations or scheduling conflicts, are removed [12], it becomes evident that the low participation rate and poor adherence to exercise rehabilitation are likely influenced by a lack of motivation and awareness regarding long-term rehabilitation. This complacency diminishes awareness of the importance of rehabilitation and negatively impacts exercise behavior. The in-depth analysis of adherence to exercise rehabilitation identified 2 main types of influencing factors [13]: self-related factors, which include age, sex, work status, economic conditions, and health knowledge [14]; and external health system factors, which encompass the medical environment, health care personnel, rehabilitation models, and social support [15]. Specifically, patients' lack of awareness regarding the benefits of exercise rehabilitation, coupled with a fear of pain, has contributed to decreased adherence to exercise [16,17]. In 1998, Johnson and Heller [18] proposed that perceptions of risk factors and benefit factors are significant predictors of behavior. The benefit-risk analysis model [19] further suggests that behavioral intention is closely linked to an individual's perception of risk and benefit. The level of risk-benefit perception is influenced by various interfering factors, offering new insights into how internal and external factors impact behavioral intention. Classical behavioral theories indicate that generating behavioral motivation is a crucial prerequisite for individual behavior [20,21]. According to social exchange theory, exercise motivation arises from perceptions based on benefits and risks. The formation of accurate perceptions is linked to the appropriate assessment of both internal and external conditions [22].

Previous approaches to behavioral promotion have primarily relied on classical research and analytical methods in psychology and sociology [23]. This reliance somewhat restricts the scope of research on behavioral promotion to existing frameworks, hindering the implementation of promotional measures and the quantification of their mechanisms of action. Building on classical theories in psychology and sociology, Wang [24] proposed the "SCeiP" (Self-Evaluation/Condition of Exercise-Effect Perception-Internal Drive-Persistence Behavior) model. This model emphasizes the self-evaluation of healthy behavior and exercise conditions. It integrates the concepts of effect perception, internal drive, and persistence behavior, forming a transmission mechanism that encompasses perception input, decision-making, motivation, and action output. This mechanism yields positive outcomes for exercise persistence (Multimedia Appendix 1; also see [24]). Based on the "SCeiP" model, this study developed a promotion strategy for exercise rehabilitation behavior and examined its effects on exercise planning, commitment, and adherence among individuals with CHD.

Objectives

To implement and evaluate the impact of remote exercise rehabilitation promotion strategies based on the "SCeiP" model, we conducted a randomized controlled trial comparing these strategies with current exercise rehabilitation management methods for patients with CHD in a home setting. The primary indicators we observed included compliance with exercise rehabilitation, encompassing both subjective cognitive levels and objective exercise input. To assess the effectiveness and persistence of the intervention, we measured the relevant indicators before the intervention, as well as at 1 month and 3 months after the intervention, conducting both intergroup and intragroup statistical analyses. We hypothesized was that implementing the exercise rehabilitation promotion strategy based on the "SCeiP" model would lead to modest improvements in exercise input and compliance among homebound patients.

XSL•FO
RenderX

Methods

Study Design and Setting

This was a single-center, randomized controlled trial, conducted in accordance with the CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines [25] at all stages (Multimedia Appendix 2; also see [26]). The sample size was calculated to compare 2 sample rates [27] using the following formula: N1=N2=$[2(U_\alpha+U_\beta)^2P(1-P)]/(P_1-P_2)^2$, where the test criterion was α=.05 and the probability of type II error was β=.2. Consequently, U_α=1.96 and U_β=0.84; P_0 and P_1 represent the preintervention and postintervention exercise rehabilitation participation rates, respectively, with $P = (P_0+P_1)/2$. Based on preexperimental findings from a small sample, we derived P_0 as 46% and P_1 as 72%, resulting in 63 study participants assigned to both the intervention and control groups. Considering a 20% attrition rate, a final sample size of 76 cases per group was established. A random number table was generated using SPSS 25.0 (IBM Corp.), and the enrolled patients were assigned to either the control group or the intervention group by individuals not involved in the study, according to the random number table.

Participants

The convenience sampling method was used to select patients who underwent coronary interventional procedures between June 2022 and March 2023 at a tertiary hospital in Jiangsu Province. The inclusion criteria for patients were as follows: (1) diagnosed with CHD; (2) aged 18-70 years; (3) no mental disabilities and capable of effective verbal or written communication; (4) classified according to risk stratification criteria for cardiac rehabilitation exercise, specifically patients with low to medium risk; (5) a Barthel index score of ≥60; and (6) voluntary participation in the investigation with signed informed consent. The exclusion criteria for individuals were as follows: (1) presence of combined diseases that are unsuitable for exercise, such as heart valve disease, severe arrhythmia, or osteoarthritis; (2) presence of serious heart, liver, kidney, or other complications, or other severe consumptive diseases; (3) inability to be followed up after hospital discharge; (4) voluntary withdrawal from the study during its course; and (5) inability to participate in the study as planned. The trial was prospectively registered with the Chinese Clinical Trial Registry (ChiCTR2300069463).

Experimental Group: Remote Exercise Rehabilitation Based on the "SCeiP" Model

According to the "SCeiP" model proposed by Wang [24], 2 intervention pathways were explored. These pathways begin with "self-evaluation of healthy behavior" and "condition of exercise," connecting the perceived effects of exercise with internal motivation to enhance adherence to exercise behavior. This results in the "SeiP" path during the exercise cognition stage and the "CeiP" path during the exercise promotion implementation stage. The first draft of the program was developed by integrating the characteristics of exercise rehabilitation for patients with CHD. The final draft was then refined through group discussions and validation by a panel of 5 senior nursing specialists, 3 physicians, and 3 cardiologists.

Cognitive Phase: The "SeiP" Pathway

This stage is primarily conducted in hospitals. Doctors and rehabilitation instructors utilize the exercise self-assessment form to develop exercise rehabilitation programs tailored to each patient's condition. Additionally, nurses gather information through a lifestyle self-assessment questionnaire to evaluate the patient's diet, sleep patterns, self-care abilities, and other relevant factors. Nurses use the patient's health status to enhance their health records based on test indicators. Additionally, a personal input self-assessment form was used to gather information on the amount of time each patient dedicates to exercise daily. This information is collected during a question-and-answer session with the nurses. The assessments mentioned above are completed 3-5 days after the patient's admission to the hospital, allowing for the establishment of initial records. Based on the patient's self-assessment, a cognitive intervention program was developed and delivered through group health education and patient communication sessions. These sessions were conducted 3-5 times, each lasting 30-40 minutes, and were supplemented by face-to-face guidance and question-and-answer sessions with the rehabilitation instructor and cardiologist (Multimedia Appendix 3).

Implementation Phase: "CeiP" Pathway

Exercise Conditioning Phase (3-5 Weeks)

First, we assessed the conditions under which patients engaged in out-of-hospital exercise rehabilitation, including the use of external exercise equipment and devices, the exercise environment, and the support system. Individualized exercise programs were designed based on patients' exercise preferences at various locations. Patients selected exercise equipment based on their personal preferences, including jump ropes, yoga mats, dumbbells, and other tools. The research team supplied patients with exercise bracelets to monitor vital signs, exercise duration, intensity, and energy expenditure, with data automatically uploaded. This approach enhanced the out-of-hospital management of patients with CHD and fostered a supportive exercise environment. The survey primarily focused on establishing WeChat (Tencent Holdings Limited) groups for patients with CHD. Participants were encouraged to complete and upload their exercise logs daily, organize offline exercise sessions, and perform a 6-minute walk test during outpatient follow-ups for any group exercise involving more than 2 participants. Online and offline interactive methods were integrated to enhance physical examinations. Additionally, the support system was strengthened through improved communication among medical staff, peers, and family members. Regular sharing of benefits aimed to increase patients' motivation. The WeChat group, exercise logbook, and exercise bracelet provided feedback on patients' exercise behaviors, allowing doctors and rehabilitators to adjust individualized rehabilitation exercise prescriptions, including the form, frequency, and duration of exercise.

XSL•FO
RenderX

Maintenance Phase of Exercise Behavior (5-16 weeks)

Based on the rehabilitation exercise program provided by the rehabilitator, a detailed exercise schedule was established. Patients were required to document the type of daily exercise they performed, along with the actual start and end times, in relation to the prescribed schedule. This allowed them to compare their results against the schedule to determine compliance with the exercise regulations. Each patient then recorded their exercises in a logbook, which was uploaded to the WeChat group daily. A return visit register was established, along with regular telephone follow-ups or home visits, with at least one follow-up conducted. The follow-up visits typically occurred at 6, 10, and 16 weeks, each lasting 30-60 minutes. Patients could report any special circumstances at any time through clinic visits or the WeChat group. Each follow-up visit included an assessment of the patient's basic condition, exercise rehabilitation program, peer co-training, family support, and availability of exercise equipment. After correcting incorrect rehabilitation methods and living habits, timely feedback from the doctor and rehabilitation division was provided regarding unsuitable exercise programs, facilitating easy adjustments. Follow-up visits were recorded in the patient's rehabilitation exercise book and retained as stubs. If patients had questions about the exercise program, they could request a case consultation appointment through the WeChat group. The rehabilitator and doctor held online question-and-answer sessions every 2 weeks, lasting 20-30 minutes each, to identify current issues and offer one-on-one guidance.

Control Group

The control group received a prescribed exercise rehabilitation program and was monitored during follow-up. Cardiologists and rehabilitation specialists collaborated to create individualized exercise rehabilitation programs for each patient. Nurses were responsible for implementing the exercise program log schedule. Patients and their families were instructed on how to record the number of exercise sessions and consultations regarding exercise rehabilitation through various channels. (2) During hospitalization, patients received education about their disease and related exercise rehabilitation. This education aimed to enhance their understanding of CHD and the importance of regular exercise. After discharge, patients were further supported with a guidance program focused on the secondary prevention of CHD and exercise rehabilitation. (3) After discharge, regular telephone follow-ups were conducted, and patients attended follow-up visits every 5 weeks for 30-60 minutes. The rehabilitation clinic and cardiovascular outpatient clinic took special circumstances into account, providing feedback and adjustments tailored to each patient's situation.

Measures

Adherence to Exercise Rehabilitation

When patients received their exercise rehabilitation prescriptions at 1 month and 3 months after the intervention, data on exercise duration and the number of exercise days per week were collected. A daily exercise duration of up to 80% of the theoretical exercise time was considered an indication of completing one day's exercise prescription. This study analyzed the percentage of participants who completed the exercise program, as well as the total duration and number of days of exercise. The formula for the completion rate of exercise duration was (actual daily exercise time/theoretical daily exercise time) × 100%, and the formula for the completion rate of exercise days was (actual exercise day/30 × 100%). The theoretical daily exercise time was established by the rehabilitation physician and cardiologist, referencing guidelines for cardiac rehabilitation [28]. Additionally, taking into account each patient's physical condition, the research group formulated an individualized daily exercise rehabilitation prescription that included the theoretical daily exercise duration. The difference in theoretical exercise duration per day between the 2 patient groups was not statistically significant (t_{145}=1.40, P=.16).

Cardiac Rehabilitation Scale

This scale was developed by Micklewright et al [29] and adapted for the Chinese context by Wang et al [30]. The assessment was conducted to evaluate patients' cognitive levels regarding personalized cardiac rehabilitation. The scale comprised 3 dimensions: process anxiety, outcome anxiety, and autonomy, totaling 18 items. A 5-point Likert scale was utilized, with the sum of the scores for the corresponding items in each dimension representing the score for that dimension. The total Cronbach α for the scale was 0.816, indicating good internal consistency, while the content validity index was 0.96. The scale was validated based on the following criteria.

Exercise Planning Scale

The Chinese version of the Runner Exercise Planning Scale was translated by Shen [31] and consists of 9 items divided into 2 dimensions. Each item is rated on a 5-point Likert scale, ranging from "1=not at all sure" to "5=completely sure." The Cronbach α for each dimension was 0.8832 and 0.8830, respectively, indicating strong internal consistency. A higher score reflects a stronger exercise planning ability among participants.

Exercise Input Scale

The scale comprised 4 dimensions: vigor persistence, focus satisfaction, value perception, and participation autonomy, totaling 20 items. A 5-point Likert scale was used, with scores ranging from 1 to 5, indicating "1=not at all conforming" to "5=fully conforming," respectively. The total Cronbach α for the scale was 0.906 [32], indicating excellent internal consistency.

Ethical Considerations

The study was designed in accordance with relevant ethical standards for human experimentation and the Helsinki Declaration. The Medical Ethics Committee of Jiangsu Province Hospital reviewed and approved the study protocol (2021-SRFA-329). All enrolled patients provided informed consent and participated voluntarily. Patient data were used exclusively for this study. The personal information of patients was kept confidential throughout the entire study. After the conclusion of the research, the data were securely sealed without requiring additional consent for secondary analysis. Each participating patient received a complimentary outpatient visit

XSL•FO
RenderX

to a specialist in the cardiology department. Additionally, the auxiliary exercise equipment used during the intervention was not removed but left available for continued use by the patients.

Process and Quality Control Method

The outcomes were measured at 3 time points: baseline (T0: week 0), 1 month after the start of the intervention (T1: week 4), and 3 months after the start of the intervention (T2: week 12). We utilized general information questionnaires to collect data on age, sex, marital status, education level, monthly income, ability to perform activities of daily living, New York Heart Association cardiac function classification, and the number of diseased vessels. Additionally, we assessed the effect of the intervention using the cardiac rehabilitation scale, along with measurements of exercise planning, exercise input, and exercise rehabilitation adherence. Two trained nurses (holding intermediate titles) provided uniform instructions to the participants for completing the scales before the study commenced. Patient data were collected using paper questionnaires and the Questionnaire Star electronic platform at 1 month and 3 months following the intervention. Consistent with previous studies, the dropout rate increased gradually during months 3 and 4 of the exercise rehabilitation process. To prevent omissions, the completion of the questionnaire was verified on-site. At the conclusion of data collection, 2 additional researchers were tasked with entering, checking, and correcting the data without knowledge of the group assignments.

Statistical Analyses

Data entry and statistical analysis were conducted using SPSS 25.0 software. The quantitative data in this study were normally distributed and presented as mean and SD. To determine whether significant differences existed between groups, the chi-square test was used to examine categorical or nominal variables (eg, gender, age stratification, marital status, education level, residence, income level, cardiac functional grading, and the number of diseased vessels). Additionally, independent samples *t* tests (unpaired, 2-tailed) were performed on continuous variables between the 2 groups, including baseline theoretical average daily exercise duration, completion rates, and score evaluation indicators. A paired samples *t* test was then applied to examine the completion rates of exercise days and duration among the same patients at 1 and 3 months after the intervention. Repeated measures ANOVA was utilized to compare the effects of multiple interactions between group and intervention modality on the levels of cardiac rehabilitation cognition, exercise planning, and exercise input. Differences were considered statistically significant at $P<.05$.

Results

Participant Demographics

Out of the initial sample, 5 participants (3.4%) dropped out before the final survey due to medical reasons, including gastrointestinal surgery and neoplastic diseases, as well as loss to follow-up resulting from personal life events. Consequently, a total of 147 participants were included in this study, comprising 81 men (55.1%) and 66 women (44.9%). The participants were divided into an experimental group (n=73, 49.7%) and a control group (n=74, 50.3%; Figure 1). Statistically, there were no significant differences between the 2 groups regarding sex ($P=.56$), age ($P=.73$), marital status ($P=.70$), education level ($P=.88$), place of residence ($P=.81$), personal monthly income ($P=.75$), New York Heart Association cardiac function classification ($P=.50$), number of diseased vessels ($P=.69$), or theoretical exercise duration (Table 1). Additionally, no cardiovascular-related adverse events were reported during the 3-month study period.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow of participant progression. SCeiP: Self-Evaluation/Condition of Exercise-Effect Perception-Internal Drive-Persistence Behavior.

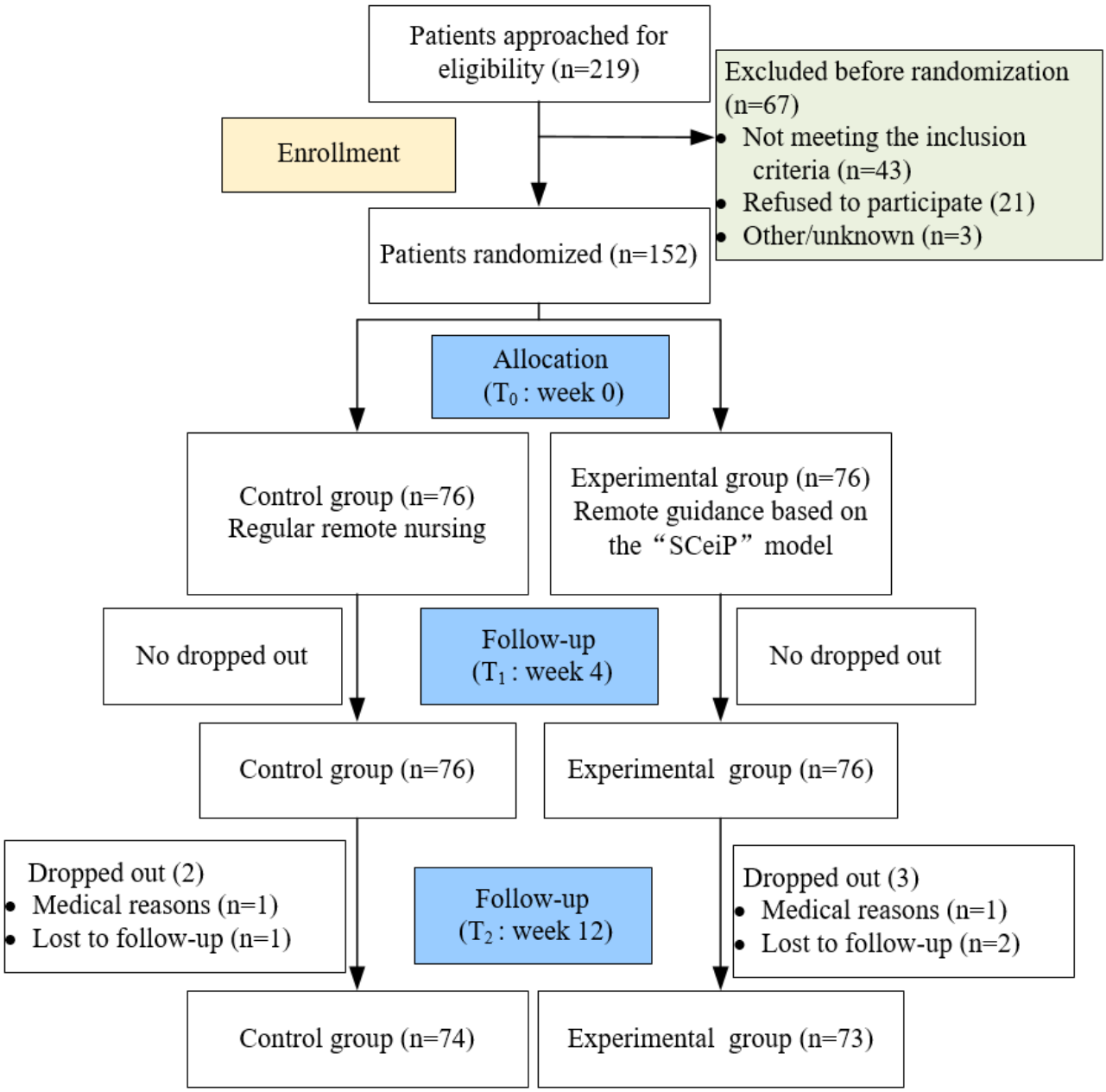

XSL•FO
RenderX

Table 1. Baseline characteristics of participants.

Variables	Experimental group (n=73)	Control group (n=74)	χ^2/*t* test (*df*)	*P* value
Gender, n (%)			0.347[a] (145)	.56
Female	31 (42.5)	35 (47.3)		
Male	42 (57.5)	39 (52.7)		
Age (years), n (%)			0.622[a] (145)	.73
<30	10 (13.7)	12 (16.2)		
30-60	48 (65.8)	44 (59.5)		
>60	15 (20.5)	18 (24.3)		
Marital status, n (%)			0.147[a] (145)	.70
Married	68 (93.2)	71 (95.9)		
Single	5 (6.8)	3 (4.1)		
Education level, n (%)			0.657[a] (145)	.88
Primary and below	17 (23.3)	21 (28.4)		
Junior high school	25 (34.2)	23 (31.1)		
Senior high school	14 (19.2)	15 (20.3)		
College and above	17 (23.3)	15 (20.3)		
Residence, n (%)			0.055[a] (145)	.81
City	44 (60.3)	46 (62.2)		
Town	29 (39.7)	28 (37.8)		
Monthly personal income (RMB[b]), n (%)			0.579[a] (145)	.75
<5000	26 (35.6)	22 (29.7)		
5000-10,000	37 (50.7)	41 (55.4)		
>10,000	10 (13.7)	11 (14.9)		
NYHA[c], n (%)			1.385[a] (145)	.50
Level 1	46 (63.0)	51 (68.9)		
Level 2	24 (32.9)	22 (29.7)		
Level 3	3 (4.1)	1 (1.4)		
Number of diseased vessels, n (%)			0.736[a] (145)	.69
1	52 (71.2)	48 (64.9)		
2	19 (26.0)	23 (31.1)		
3	2 (2.7)	3 (4.1)		
Theoretical average daily exercise duration (minutes), mean (SD)	45.12 (5.15)	43.76 (6.60)	1.401[d] (145)	.16

[a]Chi-square test.
[b]1RMB=US $0.14.
[c]NYHA: New York Heart Association.
[d]Independent samples *t* test.

Comparison of Adherence to Exercise Rehabilitation

The completion rates of exercise days declined gradually in both the intervention (t_{72}=2.973, *P*=.004) and control groups (t_{73}=9.457, *P*<.001). Similar results were observed for the completion rates of exercise duration (t_{72}=8.252, *P*<.001 for the intervention group and t_{73}=31.054, *P*<.001 for the control group). The results of the intervention study revealed that the strategies based on the "SCeiP" model exhibited significantly higher exercise completion rates at 1 and 3 months after the intervention (t_{145}=5.429, *P*<.001) compared with the control group (t_{145}=9.113, *P*<.001). Similarly, the completion rate of exercise duration (t_{145}=3.471, *P*=.001) was significantly greater than that of the control group (t_{145}=5.574, *P*<.001; Table 2).

XSL•FO
RenderX

Table 2. Comparison of adherence to exercise rehabilitation between the 2 groups.

Comparison	Completion rate of exercise days, mean (SD)			Completion rate of exercise duration, mean (SD)		
	T_1	T_2	*t* test (*df*)	T_1	T_2	*t* test (*df*)
Intervention group	67.76 (9.56)	64.38 (9.47)	2.973[a] (72)	84.01 (5.77)	82.17 (5.51)	8.252[b] (72)
Control group	58.78 (10.46)	50.09 (9.54)	9.457[b] (73)	80.58 (6.20)	76.77 (6.22)	31.054[b] (73)
t test (*df*)	5.429 (145)	9.113 (145)	N/A[c]	3.471 (145)	5.574 (145)	N/A
P	<.001	<.001	N/A	0.001	<.001	N/A

[a]*P*=.004.
[b]*P*<.001.
[c]N/A: not applicable.

Comparison of Cognitive Level of Cardiac Rehabilitation

There was no statistically significant difference in the scores for the 3 dimensions of cardiac rehabilitation—autonomy, process anxiety, or outcome anxiety—between the 2 groups of patients before the intervention (Tables 3-5). However, the level of autonomy in the intervention group showed a significant increase compared with the control group at 1 month (t_{145}=4.227, *P*<.001) and 3 months (t_{145}=6.277, *P*<.001) after the intervention, with a gradual improvement observed over time in both groups. The scores for process anxiety at 1 month (t_{145}=–2.950, *P*=.004) and 3 months (t_{145}=–6.367, *P*<.001) and for outcome anxiety at 1 month (t_{145}=–8.999, *P*<.001) and 3 months (t_{145}=–8.424, *P*<.001) exhibited a significant decrease in the SCeiP group. The trends in changes in cognitive function during cardiac rehabilitation differed significantly between the 2 groups over time: autonomy ($F_{1,145[time\times group]}$=9.055, *P*<.001), process anxiety ($F_{1,145[time\times group]}$=30.790, *P*<.001), outcome anxiety ($F_{1,145[time\times group]}$=28.186, *P*<.001; Tables 3-5). This indicates that the implementation of interventions resulted in a gradual improvement in patients' cognitive levels related to cardiac rehabilitation over time.

Table 3. Comparison of the autonomy dimension of Cardiac Rehabilitation Scale scores between the 2 groups.

Comparison	T_0, mean (SD)	T_1, mean (SD)	T_2, mean (SD)	F_{time} test (*df*)	F_{group} test (*df*)	$F_{time\times group}$ test (*df*)
Experimental group	12.10 (3.05)	19.01 (2.60)	20.03 (1.97)	461.821[a] (1, 145)	17.878[a] (1, 145)	9.055[a] (1, 145)
Control group	11.74 (3.33)	16.93 (3.32)	17.77 (2.37)	N/A[b]	N/A	N/A
t test (*df*)	0.67 (145)	4.227 (145)	6.277 (145)	N/A	N/A	N/A
P value	.50	<.001	<.001	N/A	N/A	N/A

[a]*P*<.001.
[b]N/A: not applicable.

Table 4. Comparison of the process anxiety dimension of Cardiac Rehabilitation Scale scores between the 2 groups.

Comparison	T_0, mean (SD)	T_1, mean (SD)	T_2, mean (SD)	F_{time} test (*df*)	F_{group} test (*df*)	$F_{time\times group}$ test (*df*)
Experimental group	20.60 (3.59)	15.51 (3.59)	12.78 (2.69)	1272.100[a] (1, 145)	13.408[a] (1, 145)	30.790[a] (1, 145)
Control group	21.03 (3.80)	17.26 (3.60)	16.01 (3.41)	N/A[b]	N/A	N/A
t test (*df*)	–0.696 (145)	–2.95 (145)	–6.367 (145)	N/A	N/A	N/A
P value	.49	.004	<.001	N/A	N/A	N/A

[a]*P*<.001.
[b]N/A: not applicable.

XSL•FO
RenderX

Table 5. Comparison of the outcome anxiety dimension of Cardiac Rehabilitation Scale scores between the 2 groups.

Comparison	T_0, mean (SD)	T_1, mean (SD)	T_2, mean (SD)	F_{time} test (*df*)	F_{group} test (*df*)	$F_{time\times group}$ test (*df*)
Experimental group	12.41 (2.36)	8.00 (1.87)	6.79 (2.44)	235.586[a] (1, 145)	72.392[a] (1, 145)	28.186[a] (1, 145)
Control group	13.04 (2.56)	11.05 (2.23)	9.95 (2.08)	N/A[b]	N/A	N/A
t test (*df*)	–1.549 (145)	–8.999 (145)	–8.424 (145)	N/A	N/A	N/A
P value	.12	<.001	<.001	N/A	N/A	N/A

[a]*P*<.001.
[b]N/A: not applicable.

Comparison of Exercise Planning

The difference in preintervention exercise planning scores between the 2 groups was not statistically significant (*P*=.17). However, the exercise planning scores in the experimental group were significantly higher than those in the control group at both 1 month (t_{145}=2.490, *P*=.01) and 3 months (t_{145}=3.379, *P*<.001) after the intervention. Interestingly, during the intervention, exercise planning scores for both groups initially increased and then decreased; however, both remained significantly higher than preintervention levels ($F_{1,145[time]}$=505.053, *P*<.001). Additionally, the trend of change in exercise planning differed significantly over time between the 2 groups ($F_{1,145[time\times group]}$=3.527, *P*=.03; Table 6).

Table 6. Comparison of exercise planning scores between the 2 groups.

Comparison	T_0, mean (SD)	T_1, mean (SD)	T_2, mean (SD)	F_{time} test (*df*)	F_{group} test (*df*)	$F_{time\times group}$ test (*df*)
Experimental group	31.82 (3.77)	36.93 (3.83)	34.68 (3.83)	505.053[a] (1, 145)	7.503[b] (1, 145)	3.527[c] (1, 145)
Control group	30.97 (3.63)	35.36 (3.80)	32.76 (3.05)	N/A[d]	N/A	N/A
t test (*df*)	1.392 (145)	2.49 (145)	3.379 (145)	N/A	N/A	N/A
P	.17	.01	.001	N/A	N/A	N/A

[a]*P*<.001.
[b]*P*=.007.
[c]*P*=.03.
[d]N/A: not applicable.

Comparison of Exercise Input Levels

The difference in preintervention exercise input scores between the 2 groups was not statistically significant (*P*=.68). However, at 1 and 3 months after the intervention, the exercise input scores in the experimental group (t_{145}=2.255, *P*=.03) were significantly higher than those in the control group (t_{145}=3.817, *P*<.001). Similarly, we observed that the scores for exercise input initially increased and then decreased, yet both remained significantly higher than the preintervention levels ($F_{1,145[time]}$=302.616, *P*<.001). Additionally, the pattern of change in exercise input showed significant temporal variation as the duration of the intervention period increased ($F_{1,145[time\times group]}$=10.886, *P*<.001; Table 7).

Table 7. Comparison of exercise input scores between the 2 groups.

Comparison	T_0, mean (SD)	T_1, mean (SD)	T_2, mean (SD)	F_{time} test (*df*)	F_{group} test (*df*)	$F_{time\times group}$ test (*df*)
Experimental group	66.07 (9.56)	76.37 (9.20)	74.48 (8.28)	302.616[a] (1, 145)	4.971[b] (1, 145)	10.886[a] (1, 145)
Control group	65.45 (8.94)	72.97 (9.06)	69.47 (7.61)	N/A[c]	N/A	N/A
t test (*df*)	0.408 (145)	2.255 (145)	3.817 (145)	N/A	N/A	N/A
P	.68	.03	<.001	N/A	N/A	N/A

[a]*P*<.001.
[b]*P*=.03
[c]N/A: not applicable.

XSL•FO
RenderX

Discussion

Overview

We conducted a randomized controlled trial to evaluate the efficacy of remote exercise rehabilitation based on the "SCeiP" model in patients with CHD. Our study was powered to detect small differences between the exercise rehabilitation strategies while ensuring no adverse cardiovascular events occurred. The findings indicated significant improvements over time in patients' cognitive levels regarding cardiac rehabilitation, the rate of implementation of exercise rehabilitation, and adherence compared with the traditional exercise rehabilitation model. Moreover, these indicators sustained marked improvement after the intervention.

Principal Findings

Based on the "SCeiP" model, we implemented an exercise rehabilitation promotion strategy utilizing information technology, including WeChat exercise groups, exercise bracelets, and online consultation applets. This approach enhanced the accuracy and convenience of acquiring knowledge related to cardiac rehabilitation, particularly for patients engaging in home-based exercise rehabilitation. A cross-sectional analysis [33] showed that patients with CHD who highly accepted personalized, user-friendly eHealth platforms with remote monitoring were more willing to maintain phase III cardiac rehabilitation. Additionally, Ramachandran et al [5] reported that the implementation of exercise rehabilitation strategies facilitated by informational tools can effectively enhance patients' quality of life, improve unhealthy lifestyle habits, and reduce risk factors for adverse cardiovascular events. The reasons for this improvement may include the real-time sharing of exercise prescriptions through WeChat groups, the ability to use keyword searches to obtain methods for monitoring vital signs such as blood pressure and heart rate, the application of techniques for evaluating exercise effects, and the utilization of smart exercise equipment that automatically records exercise results [34]. Moreover, with the assistance of smart exercise equipment, electrocardiograms, metabolic parameters, training intensity, and exercise prescriptions are automatically transmitted to medical personnel, enabling timely adjustments and improvements to exercise prescriptions [35]. Additionally, this smart equipment, combined with online diagnosis and treatment, can help individuals identify abnormal electrocardiograms and arrhythmias during exercise, providing early warnings and interventions related to exercise risks [36]. The implementation of eHealth intervention strategies, including wearable devices, monitoring tools, and web-based portals, has the potential to enhance the efficacy of exercise rehabilitation training [37]. The results indicated that the difference in cardiac rehabilitation–related cognitive status between the 2 groups began to widen at 1 month after the intervention and became more pronounced at 3 months. This suggests that the exercise rehabilitation facilitation strategy used in the experimental group had a lasting impact on improving cardiac rehabilitation cognition in patients with CHD and that its effectiveness progressively outperformed that of the control group. Guo et al [38] also demonstrated that remote rehabilitation can enhance the long-term effectiveness of cardiac rehabilitation for patients. In this study, the combination of WeChat and a smart bracelet facilitated its application and sustained usage.

The exercise rehabilitation strategies emphasized the creation of individualized exercise rehabilitation prescriptions and outcome tracking, offering targeted guidance for both in-hospital phase I and home phase II cardiac rehabilitation for patients with CHD. This approach led to improved levels of exercise planning and input. Nabutovsky et al [39] also found that extended monitoring of prescribed exercise implementation positively influenced patients' adherence to rehabilitation. This improvement may be attributed to the increased degree of patient participation. The process of formulating precise exercise prescriptions considers the patient's exercise preferences and adjusts the program based on their actual exercise conditions, thereby enhancing the patient's ability to engage in collaborative decision-making [40]. Additionally, the external support system was strengthened, as social support from family and health care providers positively influences patients' knowledge, beliefs, and behaviors, fostering a greater willingness to engage in exercise [41]. Thus, home-based cardiac rehabilitation can enhance patients' exercise planning and commitment when supported by information technology and personalized exercise program development [42]. This study demonstrated that both groups exhibited a greater willingness to exercise compared with the preintervention period. Moreover, the differences between the 2 groups in terms of exercise planning and input varied significantly over time. The results showed that patients' exercise planning initially increased, followed by a subsequent decrease; however, both levels remained significantly higher than the preintervention baseline. This indicates that exercise rehabilitation promotion strategies positively impacted compliance, and that interventions based on remote strategies could further optimize patient outcomes. This finding aligns with the study by Lahtio et al [43], which demonstrated that integrating remote technology into cardiac rehabilitation programs significantly enhances positive health outcomes for individuals diagnosed with cardiovascular diseases. Interestingly, the differences in the completion rates of exercise days and duration, as well as in exercise planning and exercise input, were more pronounced 1 month after the intervention. However, over time, these changes appeared to diminish. This may be attributed to the fatigue that often occurs after extended periods of exercise rehabilitation. Our previous qualitative interviews revealed that patients experienced a decline in motor performance during prolonged exercise rehabilitation. Relevant research findings on this topic have been documented and are currently being submitted.

Strengths and Future Directions

To our knowledge, studies on adherence in exercise rehabilitation have primarily focused on subjective measurement indicators. The evaluation indicators in this study not only include subjective indicators such as cognitive level and exercise planning, but also enhance the objectivity of evaluation indicators by documenting the frequency and duration of each exercise session. Furthermore, it is noteworthy that the introduced "SCeiP" model, which advocates for the preservation of motor behavior, aligns with the purpose of this study. The "SCeiP" model summarizes the internal motivations for exercise

XSL•FO
RenderX

behavior as “wanting to do” or “being able to do” (ie, motivation and self-efficacy). These motivations are influenced by the individual’s health, self-evaluation, and perceptions of the conditions and effects of exercise. The model considers “self-evaluation of healthy behavior” as an internal perspective and “condition of exercise” as an external perspective. Ultimately, it encourages patients to make informed exercise decisions and maintain consistent adherence to exercise rehabilitation behaviors. By mediating the perception of exercise effects, the model translates self-evaluation into internal motivation for exercise, optimizes factors associated with exercise behavior, and enhances the synergistic effect of these factors on exercise adherence. This process ultimately increases patients’ inclination to engage in physical activity. The implementation of the SCeiP model in this study enhances compliance with exercise rehabilitation and augments the cardiac benefits for patients with CHD. Previous literature has demonstrated that remote cardiac rehabilitation is advantageous for reducing readmission rates [44] and medication costs [45]. The incorporation of the SCeiP model not only improves participation and sustainability in exercise rehabilitation but also holds significant economic value, warranting further discussion. Moreover, WeChat and exercise bracelets were effectively utilized to promote and supervise the implementation of the SCeiP model. Existing studies have also reported the positive role of smartphone-based programs [46], wearable devices [47], and other information technologies in remote rehabilitation, suggesting that further development and application of artificial intelligence technology may unlock additional possibilities for remote cardiac rehabilitation.

Limitations

There were several limitations in this study. First, it is imperative to consider expanding the subject enrollment to encompass a larger sample size, with particular emphasis on enhancing exercise rehabilitation compliance among patients with CHD who also have obesity, diabetes, or other chronic diseases. Second, this study exclusively recruited participants from a single university in Nanjing, China, and its generalizability to a larger population remains undetermined. To address this limitation, future research could consider conducting multicenter randomized controlled trials. Third, the intervention period of 3 months could be extended to further observe long-term patient participation in exercise rehabilitation. Lastly, the research measures involve information tools such as WeChat groups and exercise bracelets, which may not fully represent the latest advancements in information technology. Therefore, it is advisable to incorporate new technological approaches, such as wearables, into the formulation of the scheme.

Conclusions

Based on the “SCeiP” model, this study developed a dual-path practice strategy to enhance exercise rehabilitation in patients with CHD. This strategy integrated both online and offline health education, along with the evaluation of subjective records and objective data that complemented each other. Patients’ adherence to exercise rehabilitation, exercise planning, exercise input, and cognitive levels in remote exercise rehabilitation improved significantly compared with the traditional model, showing better outcomes than before the intervention. Individual interviews with the targeted population revealed that the long-term nature of the exercise rehabilitation process could lead to a decrease in exercise commitment and adherence due to patients’ slackness at a later stage, a phenomenon also reported in the team’s follow-up study. Therefore, timely identification of the emergence of exercise slack needs to be explored in greater depth to develop targeted corrective strategies.

Acknowledgments

The research was partially funded by the National Natural Science Foundation of China (grant 72204120) and the Young Scholars Fostering Funding of the First Affiliated Hospital of Nanjing Medical University (grant PY2021047). We sincerely appreciate all the beneficial comments given by our instructor and members of the research group, with special thanks to Shiyin Zhai and Hongmei Sun for their assistance in data entry and verification. We extend our gratitude to the participants for their valuable contribution of the data.

Data Availability

The research data are uploaded to a publicly accessible personal database [48] with stringent usage restrictions. The study researchers can be contacted for utilization purposes if necessary.

Authors' Contributions

XZ, DdX, and LW contributed to the study’s conception. XZ and SL were responsible for designing the process and content of the SCeiP intervention. ZB assisted in the recruitment and screening of participants. XZ and DdX collaborated on the trial design and manuscript drafting. LW, DdX, and SL conducted the trial. ZB and DmX commented on the design of the trial and supervised data collection. XZ and DmX performed data analyses. All authors revised the content critically and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

The "SCeiP" model adopted from Dr. Wang, 2021 [24]. SCeiP: Self-Evaluation/Condition of Exercise-Effect Perception-Internal Drive-Persistence Behavior.
[PNG File , 76 KB - jmir_v26i1e56552_app1.png]

Multimedia Appendix 2
CONSORT-EHEALTH (V 1.6.1) checklist. CONSORT: Consolidated Standards of Reporting Trials.
[PDF File (Adobe PDF File), 3455 KB - jmir_v26i1e56552_app2.pdf]

Multimedia Appendix 3
Specific intervention programs for the cognitive phase.
[DOCX File , 15 KB - jmir_v26i1e56552_app3.docx]

References

1. Anderson LJ, Taylor RS. Cardiac rehabilitation for people with heart disease: an overview of Cochrane systematic reviews. Int J Cardiol 2014 Dec 15;177(2):348-361. [doi: 10.1016/j.ijcard.2014.10.011] [Medline: 25456575]
2. Dibben G, Faulkner J, Oldridge N, Rees K, Thompson D, Zwisler A, et al. Exercise-based cardiac rehabilitation for coronary heart disease: a meta-analysis. Eur Heart J 2023 Feb 07;44(6):452-469 [FREE Full text] [doi: 10.1093/eurheartj/ehac747] [Medline: 36746187]
3. Bakhshayeh S, Sarbaz M, Kimiafar K, Vakilian F, Eslami S. Barriers to participation in center-based cardiac rehabilitation programs and patients' attitude toward home-based cardiac rehabilitation programs. Physiother Theory Pract 2021 Jan;37(1):158-168. [doi: 10.1080/09593985.2019.1620388] [Medline: 31155986]
4. Thomas RJ, Beatty AL, Beckie TM, Brewer LC, Brown TM, Forman DE, et al. Home-based cardiac rehabilitation: a scientific statement from the American Association of Cardiovascular and Pulmonary Rehabilitation, the American Heart Association, and the American College of Cardiology. Circulation 2019 Jul 02;140(1):e69-e89 [FREE Full text] [doi: 10.1161/CIR.0000000000000663] [Medline: 31082266]
5. Ramachandran H, Jiang Y, Tam W, Yeo T, Wang W. Effectiveness of home-based cardiac telerehabilitation as an alternative to phase 2 cardiac rehabilitation of coronary heart disease: a systematic review and meta-analysis. Eur J Prev Cardiol 2022 May 25;29(7):1017-1043 [FREE Full text] [doi: 10.1093/eurjpc/zwab106] [Medline: 34254118]
6. Antoniou V, Kapreli E, Davos CH, Batalik L, Pepera G. Safety and long-term outcomes of remote cardiac rehabilitation in coronary heart disease patients: a systematic review. Digit Health 2024;10:20552076241237661 [FREE Full text] [doi: 10.1177/20552076241237661] [Medline: 38533308]
7. Grace SL, Turk-Adawi K, Santiago de Araújo Pio C, Alter DA. Ensuring cardiac rehabilitation access for the majority of those in need: a call to action for Canada. Can J Cardiol 2016 Oct;32(10 Suppl 2):S358-S364. [doi: 10.1016/j.cjca.2016.07.001] [Medline: 27692116]
8. Abreu A, Mendes M, Dores H, Silveira C, Fontes P, Teixeira M, et al. Mandatory criteria for cardiac rehabilitation programs: 2018 guidelines from the Portuguese Society of Cardiology. Rev Port Cardiol (Engl Ed) 2018 May;37(5):363-373 [FREE Full text] [doi: 10.1016/j.repc.2018.02.006] [Medline: 29724635]
9. Ruano-Ravina A, Pena-Gil C, Abu-Assi E, Raposeiras S, van 't Hof A, Meindersma E, et al. Participation and adherence to cardiac rehabilitation programs. A systematic review. Int J Cardiol 2016 Nov 15;223:436-443. [doi: 10.1016/j.ijcard.2016.08.120] [Medline: 27557484]
10. Sommer CG, Jørgensen LB, Blume B, Møller T, Skou ST, Harrison A, et al. Dropout during a 12-week transitional exercise-based cardiac rehabilitation programme: a mixed-methods prospective cohort study. Eur J Cardiovasc Nurs 2022 Aug 29;21(6):578-586 [FREE Full text] [doi: 10.1093/eurjcn/zvab119] [Medline: 35043160]
11. British Heart Foundation. The National Audit of Cardiac Rehabilitation: Quality and Outcomes Report 2018. British Heart Foundation. 2018 Nov 29. URL: https://www.bhf.org.uk/-/media/files/information-and-support/publications/hcps/nacr-report-2018_final.pdf?rev=9f3fe85b82b5411d9345e33b9608e334 [accessed 2024-10-21]
12. Cabrera-Aguilera I, Ivern C, Badosa N, Marco E, Salas-Medina L, Mojón D, et al. Impact of and reasons for not performing exercise training after an acute coronary syndrome in the setting of an interdisciplinary cardiac rehabilitation program: results from a risk-op-acute coronary syndrome ambispective registry. Front Physiol 2021;12:768199 [FREE Full text] [doi: 10.3389/fphys.2021.768199] [Medline: 34899392]
13. Santiago de Araújo Pio C, Chaves G, Davies P, Taylor R, Grace S. Interventions to promote patient utilisation of cardiac rehabilitation. Cochrane Database Syst Rev 2019 Feb 01;2(2):CD007131 [FREE Full text] [doi: 10.1002/14651858.CD007131.pub4] [Medline: 30706942]
14. Terada T, Vidal-Almela S, Tulloch H, Pipe A, Reed J. Cardiac rehabilitation following percutaneous coronary intervention is associated with superior psychological health and quality of life in males but not in females. J Cardiopulm Rehabil Prev 2021 Sep 01;41(5):345-350. [doi: 10.1097/HCR.0000000000000597] [Medline: 33797456]
15. Taylor JL, Holland DJ, Keating SE, Leveritt MD, Gomersall SR, Rowlands AV, et al. Short-term and long-term feasibility, safety, and efficacy of high-intensity interval training in cardiac rehabilitation: the FITR Heart Study randomized clinical

XSL•FO
RenderX

trial. JAMA Cardiol 2020 Dec 01;5(12):1382-1389 [FREE Full text] [doi: 10.1001/jamacardio.2020.3511] [Medline: 32876655]
16. Sumner J, Grace SL, Doherty P. Predictors of cardiac rehabilitation utilization in England: results from the national audit. J Am Heart Assoc 2016 Oct 21;5(10):e003903 [FREE Full text] [doi: 10.1161/JAHA.116.003903] [Medline: 27792657]
17. Yin Y, He Q, Zhang R, Cheng H, Zhang Y, Zhang J. Predictors of adherence of enhanced external counterpulsation in patients with coronary heart disease after discharge: a mixed-methods study. Front Cardiovasc Med 2022;9:1005958 [FREE Full text] [doi: 10.3389/fcvm.2022.1005958] [Medline: 36505377]
18. Johnson NA, Heller RF. Prediction of patient nonadherence with home-based exercise for cardiac rehabilitation: the role of perceived barriers and perceived benefits. Prev Med 1998;27(1):56-64. [doi: 10.1006/pmed.1997.0235] [Medline: 9465354]
19. Ueland Ø, Gunnlaugsdottir H, Holm F, Kalogeras N, Leino O, Luteijn J, et al. State of the art in benefit-risk analysis: consumer perception. Food Chem Toxicol 2012 Jan;50(1):67-76. [doi: 10.1016/j.fct.2011.06.006] [Medline: 21683114]
20. Fogg BJ. A behavior model for persuasive design. In: Proceedings of the 4th International Conference on Persuasive Technology. New York City, NY: Association for Computing Machinery; 2009 Apr 26 Presented at: 4th International Conference on Persuasive Technology; April 26-29, 2009; Claremont, CA, USA p. 1-7. [doi: 10.1145/1541948.1541999]
21. Galloway RD. Health promotion: causes, beliefs and measurements. Clin Med Res 2003 Jul;1(3):249-258 [FREE Full text] [doi: 10.3121/cmr.1.3.249] [Medline: 15931316]
22. McAlexander JH, Kim SK, Roberts SD. Loyalty: the influences of satisfaction and brand community integration. Journal of Marketing Theory and Practice 2015 Dec 02;11(4):1-11. [doi: 10.1080/10696679.2003.11658504]
23. Holdgaard A, Eckhardt-Hansen C, Lassen C, Kjesbu I, Dall C, Michaelsen K, et al. Cognitive-behavioural therapy reduces psychological distress in younger patients with cardiac disease: a randomized trial. Eur Heart J 2023 Mar 14;44(11):986-996. [doi: 10.1093/eurheartj/ehac792] [Medline: 36649937]
24. Wang L. Research on the influencing factors and promoting strategies of college students' exercise persistence behavior. Master's dissertation, Shandong University. CNKI 2021:1 [FREE Full text] [doi: 10.27272/d.cnki.gshdu.2021.000327]
25. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
26. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
27. Pan YS, Jin AM, Wang MX. Methods and common pitfalls of sample size estimation in clinical studies. Chinese Journal of Stroke 2022 Jan 20;17(01):31-35. [doi: 10.3969/j.issn.1673-5765.2022.01.004]
28. Chinese Medical Association, Chinese Medical Journals Publishing House, Chinese Society of General Practice, Prevention Committee of Chinese Society of Cardiology, Cardiac Rehabilitation Committee of Chinese Society of Cardiology, Editorial Board of Chinese Journal of General Practitioners of Chinese Medical Association. Guideline for primary care of cardiac rehabilitation of coronary artery disease (2020). Chinese Journal of General Practitioners 2024 Oct 8;20(2):150-165. [doi: 10.3760/cma.j.cn114798-20201124-01187]
29. Micklewright D, Northeast L, Parker P, Jermy M, Hardcastle J, Davison R, et al. The Cardiac Rehabilitation Inventory: a new method of tailoring patient support. J Cardiovasc Nurs 2016;31(2):175-185. [doi: 10.1097/JCN.0000000000000220] [Medline: 25829135]
30. Wang JH, Zhang ZX, Yang QF, Mei YX, Wang P. Translation and reliability and validity of the Chinese version of the Cardiac Rehabilitation Inventory. Chinese Journal of Nursing 2019;54(04):632-636. [doi: 10.3761/j.issn.0254-1769.2019.04.033]
31. Shen MY. Intervention strategies of Chinese adults' exercise behavior: the integration of the TPB with the HAPA. Master's dissertation. Beijing Sport University. Dissertation. 2011. URL: https://www.dissertationtopic.net/doc/1508862 [accessed 2024-10-24]
32. Dong BL. Physical exercise involvement of college students in China: measurement, antecedent and aftereffect. Journal of Tianjin University of Sport 2017 Mar 25;32(02):176-184. [doi: 10.13297/j.cnki.issnl005-0000.2017.02.013]
33. Schmitz B, Wirtz S, Sestayo-Fernández M, Schäfer H, Douma ER, Alonso Vazquez M, et al. Living lab data of patient needs and expectations for eHealth-based cardiac rehabilitation in Germany and Spain from the TIMELY study: cross-sectional analysis. J Med Internet Res 2024 Feb 22;26:e53991 [FREE Full text] [doi: 10.2196/53991] [Medline: 38386376]
34. Su JJ, Yu DS. Effects of a nurse-led eHealth cardiac rehabilitation programme on health outcomes of patients with coronary heart disease: a randomised controlled trial. Int J Nurs Stud 2021 Oct;122:104040. [doi: 10.1016/j.ijnurstu.2021.104040] [Medline: 34333211]
35. Akiash N, Mohammadi M, Mombeini H, Nikpajouh A. Myocardial strain analysis as a non-invasive screening test in the diagnosis of stable coronary artery disease. Egypt Heart J 2021 May 25;73(1):49 [FREE Full text] [doi: 10.1186/s43044-021-00173-6] [Medline: 34032936]

XSL•FO
RenderX

36. Chu JK, Zhao Q, Men L, Zhou XR, Li XM, Li QJ, et al. Application of the new diagnosis and treatment model of internet+medical union based on wearable ECG equipment in cardiovascular disease. Chinese Journal of Cardiovascular Research 2023 Feb 20;21(02):160-165. [doi: 10.3969/j.issn.1672-5301.2023.02.012]
37. Yu T, Xu H, Sui X, Zhang X, Pang Y, Yu T, et al. Effectiveness of eHealth interventions on moderate-to-vigorous intensity physical activity among patients in cardiac rehabilitation: systematic review and meta-analysis. J Med Internet Res 2023 Mar 29;25:e42845. [doi: 10.2196/42845] [Medline: 36989017]
38. Guo WT, Liu JP, Zhang XX, Zhang BB, Wu YZ, Wang WJ. Effectiveness and compliance of remote cardiac rehabilitation：an overview of systematic reviews. Chinese Journal of Nursing 2023 Feb 20;58(04):426-433. [doi: 10.3761/j.issn.0254-1769.2023.04.006]
39. Nabutovsky I, Ashri S, Nachshon A, Tesler R, Shapiro Y, Wright E, et al. Feasibility, safety, and effectiveness of a mobile application in cardiac rehabilitation. Isr Med Assoc J 2020 Jun;22(6):357-363 [FREE Full text] [Medline: 32558441]
40. Cao Q, Xu L, Wen S, Li F. Investigating the influence of the shared decision-making perception on the patient adherence of the home- and exercise-based cardiac rehabilitation after percutaneous coronary intervention. Patient Prefer Adherence 2021;15:411-422 [FREE Full text] [doi: 10.2147/PPA.S292178] [Medline: 33654386]
41. Sun L, Wu T, Zhang M, Huang S, Zeng Z, Wu Y. Investigation on family support system and willingness of patients to participate in cardiac rehabilitation after percutaneous coronary intervention. Evid Based Complement Alternat Med 2022;2022:3633188 [FREE Full text] [doi: 10.1155/2022/3633188] [Medline: 35711495]
42. Antoniou V, Davos CH, Kapreli E, Batalik L, Panagiotakos DB, Pepera G. Effectiveness of home-based cardiac rehabilitation, using wearable sensors, as a multicomponent, cutting-edge intervention: a systematic review and meta-analysis. J Clin Med 2022 Jun 29;11(13):3772 [FREE Full text] [doi: 10.3390/jcm11133772] [Medline: 35807055]
43. Lahtio H, Heinonen A, Paajanen T, Sjögren T. The added value of remote technology in cardiac rehabilitation on physical function, anthropometrics, and quality of life: cluster randomized controlled trial. J Med Internet Res 2023 Apr 12;25:e42455 [FREE Full text] [doi: 10.2196/42455] [Medline: 37043264]
44. Indraratna P, Biswas U, McVeigh J, Mamo A, Magdy J, Vickers D, et al. A smartphone-based model of care to support patients with cardiac disease transitioning from hospital to the community (TeleClinical Care): pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Feb 28;10(2):e32554 [FREE Full text] [doi: 10.2196/32554] [Medline: 35225819]
45. Oehler AC, Holmstrand EC, Zhou L, Harzand A, Vathsangam H, Kendall K, et al. Cost analysis of remote cardiac rehabilitation compared with facility-based cardiac rehabilitation for coronary artery disease. Am J Cardiol 2024 Jan 01;210:266-272. [doi: 10.1016/j.amjcard.2023.08.061] [Medline: 37973439]
46. Dwiputra B, Santoso A, Purwowiyoto BS, Radi B, Ambari AM, Desandri DR, et al. Smartphone-based cardiac rehabilitation program improves functional capacity in coronary heart disease patients: a systematic review and meta-analysis. Glob Heart 2023;18(1):42 [FREE Full text] [doi: 10.5334/gh.1253] [Medline: 37577291]
47. Mitropoulos A, Anifanti M, Koukouvou G, Ntovoli A, Alexandris K, Kouidi E. Exploring the effects of real-time online cardiac telerehabilitation using wearable devices compared to gym-based cardiac exercise in people with a recent myocardial infarction: a randomised controlled trial. Front Cardiovasc Med 2024;11:1410616 [FREE Full text] [doi: 10.3389/fcvm.2024.1410616] [Medline: 38903965]
48. GitHub. URL: https://github.com/Mengyuan-Li-Lab/Zhang-s-research-database/issues [accessed 2024-10-16]

Abbreviations

CHD: coronary heart disease
CONSORT: Consolidated Standards of Reporting Trials
SCeiP: Self-Evaluation/Condition of Exercise-Effect Perception-Internal Drive-Persistence Behavior

Edited by T de Azevedo Cardoso; submitted 03.02.24; peer-reviewed by L Batalik, H Liu; comments to author 03.06.24; revised version received 19.07.24; accepted 24.09.24; published 05.11.24.

Please cite as:
Xu D, Xu D, Wei L, Bao Z, Liao S, Zhang X
The Effectiveness of Remote Exercise Rehabilitation Based on the "SCeiP" Model in Homebound Patients With Coronary Heart Disease: Randomized Controlled Trial
J Med Internet Res 2024;26:e56552
URL: https://www.jmir.org/2024/1/e56552
doi:10.2196/56552
PMID:

XSL•FO
RenderX

©Dandan Xu, Dongmei Xu, Lan Wei, Zhipeng Bao, Shengen Liao, Xinyue Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Service Quality and Patient Satisfaction of Internet Hospitals in China: Cross-Sectional Evaluation With the Service Quality Questionnaire

Tao Han[1*], MA; Qinpeng Wei[1*], BSc; Ruike Wang[1], BSc; Yijin Cai[1], BSc; Hongyi Zhu[1], BSc; Jiani Chen[1], BSc; Zhiruo Zhang[1*], PhD; Sisi Li[1*], PhD

School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Sisi Li, PhD
School of Public Health
Shanghai Jiao Tong University School of Medicine
227 Chongqing South Road
Shanghai, 200025
China
Phone: 86 02163846590 ext 776145
Email: lisi8318@gmail.com

Abstract

Background: Internet hospitals, which refer to service platforms that integrate consultation, prescription, payment, and drug delivery based on hospital entities, have been developing at a rapid pace in China since 2014. However, assessments regarding their service quality and patient satisfaction have not been well developed. There is an urgent need to comprehensively evaluate and improve the service quality of internet hospitals.

Objective: This study aims to investigate the current status of patients' use of internet hospitals, as well as familiarity and willingness to use internet hospitals, to evaluate patients' expected and perceived service qualities of internet hospitals using the Chinese version of the Service Quality Questionnaire (SERVQUAL-C) with a national representative sample, and to explore the association between service quality of internet hospitals and patients' overall satisfaction toward associated medical platforms.

Methods: This cross-sectional survey was conducted through face-to-face or digital interviews from June to September 2022. A total of 1481 outpatient participants (635 men and 846 women; mean age 33.22, SD 13.22). Participants reported their use of internet hospitals, and then rated their expectations and perceptions of service quality toward internet hospitals via the SERVQUAL-C, along with their demographic information.

Results: Among the surveyed participants, 51.2% (n=758) of participants had used internet hospital service or services. Use varied across age, education level, and annual income. Although the majority of them (n=826, 55.8%) did not know internet hospital services well, 68.1% (n=1009) of participants expressed the willingness to adopt this service. Service quality evaluation revealed that the perceived service quality did not match with the expectation, especially the responsiveness dimension. Important-performance analysis results further alerted that reliable diagnosis, prompt response, clear feedback pathway, and active feedback handling were typically the services awaiting substantial improvement. More importantly, multiple linear regressions revealed that familiarity and willingness to use internet hospital services were significant predictors of satisfaction, above and over tangibles, reliability, and empathy service perspectives, and demographic characteristics such as gender, age, education level, and annual income.

Conclusions: In the future, internet hospitals should focus more on how to narrow the gaps between the expected and perceived service quality. Promotion of internet hospitals should also be facilitated to increase patients' familiarity with and willingness to use these services.

(J Med Internet Res 2024;26:e55140) doi:10.2196/55140

XSL•FO
RenderX

KEYWORDS

service quality; SERVQUAL; Service Quality Questionnaire; internet hospital; e-hospital; digital medical care; health care professionals; Chinese digital health care

Introduction

Internet Hospital Services in China

Internet hospital (or e-hospital) services have been developing at a rapid pace in China since 2014, which refer to hospital-based service platforms that integrate consultation, prescription, payment, and drug delivery, mostly featured in digital follow-up and routine consultations [1,2]. To date, China has witnessed the establishment of over 1700 internet hospitals with an extensive array of services [3]. The rapid growth of internet hospitals in China is expected to enhance the availability and accessibility of medical resources nationwide while controlling the costs associated with distant medical services [4]. Additionally, the strike of the COVID-19 pandemic has further expedited the advancement and acceptance of remote medical care, with internet hospitals playing a pivotal role [5,6]. However, despite the rapid development of internet hospitals, assessments regarding their service quality and patient satisfaction have not been well developed [7]. This study attempted to fill in this research gap.

Internet hospital services heavily rely on various digital platforms such as smartphone apps or miniprograms integrated into popular social media platforms (eg, WeChat or Alipay). The service quality provided by internet hospital platforms directly impacts the medical experience of patients as it encompasses (1) appointment, registration, guided triage, and referral during the previsit stage; (2) remote payment, and instant access to case and test reports during the visit stage; and (3) prescription management, drug delivery, health monitoring, and remote health education during the postvisit stage. Granted these benefits, it was undeniable that the rapid development and engagement of internet hospitals had raised several practical issues such as limited service accessibility to older people and risks of personal information leakage. Therefore, it was critical to comprehensively assess, and therefore, improve the service quality of internet hospitals [8].

The Service Quality Questionnaire

The service quality model was initially developed by Parasuraman et al [9] to investigate potential causes of service quality issues and assist managers in improving service quality. The model was further refined and a measurement tool was developed accordingly, namely, the Service Quality Questionnaire (SERVQUAL) [10,11], which measures customers' expectations and perceptions of the service quality provided by an entity, reflecting their attitudes toward the service provider [10], highlighting the gaps between what customers believe should be provided (expectations) and what they believe they have received (perceptions). The SERVQUAL questionnaire consists of 5 dimensions (22 items) [10]: tangibles, which refers to the physical facilities, equipment, and appearance of personnel; reliability, which assesses the ability to deliver services accurately and dependably; responsiveness, which examines the willingness to assist customers and provide prompt service; assurance, which evaluates employees' knowledge, courtesy, and ability to convey trust and confidence; and empathy, which assesses the level of care and individualized attention provided to customers. The SERVQUAL questionnaire has been applied to evaluate service quality in health-related contexts [12-14], demonstrating its suitability for assessing medical service quality. Therefore, it was adopted to evaluate the service quality of Chinese internet hospitals and to identify the possible gaps between expected and perceived service qualities.

This Study

To briefly summarize, the purpose of this study was threefold, (1) to investigate the current status of Chinese patients' use of internet hospitals, as well as their familiarity and willingness to use internet hospitals; (2) to evaluate Chinese patients' expected and perceived service qualities of internet hospitals using SERVQUAL with a nationally representative sample; and (3) to explore the association between service quality of internet hospitals and overall satisfaction of associated medical platforms.

Methods

Participants

A total of 1586 participants were recruited via convenient sampling from 31 provinces in China. Inclusion criteria were (1) had used medical services in China within 1 year and (2) consented to participate in this study.

Procedure

A cross-sectional e-survey was administered on the SoJump platform (a popular Chinese digital survey platform) from June to September 2022. Recruitment for the survey was carried out through WeChat, a widely used social media platform in China. Potential participants were invited to complete the survey via WeChat moments. Investigators received digital data collection training sessions in advance for survey distribution, data collection monitoring, and quality control. Research team members conducted quality control on submitted responses based on (1) duplicate IP addresses, (2) shorter-than-expected response time window, and (3) incomplete responses. A total of 105 responses that failed at these criteria were excluded from the following analyses. Therefore, the final sample consisted of 1481 participants. Additionally, the 758 people who self-reported having used internet hospital services were further included in the analysis.

Ethical Considerations

This study was approved by the Shanghai Jiao Tong University School of Medicine Public Health and Nursing Research Ethics Committee (reference: SJUPN-202203). All eligible participants who consented to participate were informed about the study's purpose and procedure and then were asked to respond to the e-survey without interference. Data collection was conducted

XSL•FO
RenderX

anonymously to ensure confidentiality. For participants deemed minor according to their reported age, further consent was requested from their legal guardians. Participants did not receive any material or monetary rewards for participation.

Materials

Participants responded to a battery of surveys including SERVQUAL-C, satisfaction with internet hospital services, and their demographic information.

Internet Hospital Use and Experience

Patients first reported whether they had used the internet hospital (0=no or 1=yes), together with the frequency of use: "How often do you use the internet hospital" (1=never used before to 5=more than five times per year). Reasons for and against using internet hospitals were followed via multiple-choice questions (ie, "What are your reasons for using internet hospitals?" and "What are your reasons against using internet hospitals?"). Then, participants reported their overall familiarity, willingness to use, and satisfaction with internet hospitals by answering the questions "How much do you know about internet hospitals?" (1=barely to 5=very well), "How willing are you to use internet hospitals?" (1=very unwilling to 5=very willing), and "All things considered, how satisfied were you toward Chinese internet hospitals in general?" (1=not at all satisfied to 10=extremely satisfied).

Service Quality of Internet Hospitals

The 22-item SERVQUAL questionnaire consists of five dimensions of service quality (4 to 5 items each; 1=very poor to 5=very good), namely tangibles (TG; eg, "The aesthetic degree of internet hospital platform page design"), reliability (RA; eg, "Reliability of treatment results in internet hospitals"), responsiveness (RS; eg, "The timeliness of the internet hospital doctor's response"), assurance (AS; eg, "The extent to which the internet hospital platform meets the needs of patients"), and empathy (EP; eg, "The extent to which internet hospitals provide personalized care for patients"), each question measures both the perception and the expectation perspectives, respectively. Therefore, SERVQUAL yields three scores, specifically, the perception (P) and the expectation (E) scores, which are computed by averaging the 22 items from each perspective, respectively, and the gap score, which is further computed by subtracting the latter from the former (ie, P–E).

Sociodemographic Information

At the end of the questionnaire, participants reported their sociodemographic information, including gender (0=man and 1=woman), age (1=19 years or younger, 2=20-29 years, 3=30-39 years, 4=40-49 years, and 5=50 years or older), location (1=eastern region, 2=central region, 3=western region, and 4=northeast region), marital status (1=unmarried and 2=married), education level (1=below high school's degree, 2=high school's degree, 3=bachelor's degree, and 4=higher than bachelor's degree), as well as personal annual income (1=¥50,000 or below, 2=¥50,001-100,000, 3=¥100,001-150,000, 4=¥150,001-200,000, and 5=¥200,001 or above). The exchange rate of US $ to RMB ¥ mentioned in this paper is 1:7.0354.

Statistical Analysis

All personal identifiers have been removed prior to data analysis to ensure anonymization. All data were analyzed with SPSS (version 25.0; IBM Corp) and R (version 4.2.0; R Core Team). The questionnaire was calculated to reflect internet hospital patients' evaluations. Additionally, important-performance analysis (IPA) [15] was used to compare patients' expectations and perceptions of the services provided by internet hospitals and to reveal critical management perspectives. Finally, multiple linear regressions were also conducted to examine the association between service quality, demographics, familiarity, and willingness in patient satisfaction. All continuous variables have been centralized.

Results

Table 1 presents the sociodemographic characteristics of all the participants (n=1481; 635 men and 846 women). Participants were mainly women (n=846, 57.1%), aged between 20 and 29 (n=429, 29%), and married (n=820, 55.4%). The majority of the sample received a bachelor's degree education (n=969, 65.4%) and received less than 50,000 Chinese Yuan per year.

It was found that 51.2% (n=758) of patients had used internet hospitals. Results suggested that there were significant differences in whether patients had used the internet hospital at different levels of age ($P<.001$), education level ($P=.006$), and annual income ($P<.001$), while differences in gender or marital status were not statistically significant.

Specifically, patients in their 20s (n=268, 62.5%) and 30s (n=190, 72%) had a higher percentage of using internet hospitals. Moreover, the use of internet hospitals varied among patients with different levels of education and annual income. A total of 59.6% (n=65) of patients with higher than bachelor's degrees had tried internet hospitals, compared with 40% (n=145) of those with lower than high school degrees. Only 39.9% (n=265) of the patients with an annual income of below ¥50,000 had participated in internet hospitals.

A total of 51.2% (n=758) of the patients had used internet hospitals, among which 42.9% (n=325) of the patients used internet hospitals 1-3 times per year. It was also found that there was diversity in the patients' familiarity and willingness among those who have ever used internet hospital service before. Patients who knew internet hospitals well or excellently only accounted for 26.9% (n=398) and 17.4% (n=258), respectively. A total of 20.3% (n=301) of the patients had poor (n=230, 15.5%) or terrible (n=71, 4.8%) familiarity with internet hospitals. In terms of willingness, 68.1% (n=1009) of the patients gave positive answers, showing a willingness to use internet hospitals. A total of 29% (n=429) of the patients had neutral willingness, and only 2.9% (n=43) of the patients were unwilling to participate in internet hospitals.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics and use of internet hospital service among Chinese patients from June to September 2022 (n=1481).

	Use of internet hospital service			*P* value
	Overall, n (%)	Yes, n (%)	No, n (%)	
Gender				.56
Man	635 (42.9)	331 (43.7)	304 (42)	
Woman	846 (57.1)	427 (56.3)	419 (58)	
Age (in years)				<.001
≤19	274 (18.5)	97 (12.8)	177 (24.5)	
20-29	429 (29)	268 (35.4)	161 (22.3)	
30-39	264 (17.8)	190 (25.1)	74 (10.2)	
40-49	332 (22.4)	130 (17.2)	202 (27.9)	
≥50	182 (12.3)	73 (9.6)	109 (15.1)	
Location				<.001
Eastern Region	906 (61)	476 (62.8)	427 (59.1)	
Central Region	272 (18.4)	128 (16.9)	144 (19.9)	
Western Region	264 (17.8)	118 (15.6)	146 (20.2)	
Northeast Region	42 (2.8)	36 (4.7)	6 (0.8)	
Marital status				.31
Unmarried	661 (44.6)	328 (43.3)	333 (46.1)	
Married	820 (55.4)	430 (56.7)	390 (53.9)	
Education level				.006
<High school degree	145 (9.8)	58 (7.7)	87 (12)	
High school degree	258 (17.4)	143 (18.9)	115 (15.9)	
Bachelor's degree	969 (65.4)	492 (64.9)	477 (66)	
>Bachelor's degree	109 (7.4)	65 (8.6)	44 (6.1)	
Annual income (US $1=RMB 7.0354)				<.001
≤ ¥50,000	664 (44.8)	265 (35)	399 (55.2)	
¥50,001-100,000	353 (23.8)	207 (27.3)	146 (20.2)	
¥100,001-150,000	258 (17.4)	158 (20.8)	100 (13.8)	
¥150,001-200,000	118 (8)	69 (9.1)	49 (6.8)	
≥ ¥200,001	88 (5.9)	59 (7.8)	29 (4)	

Table 2 presents patients' reasons for (n=758) and against (n=723) using internet hospitals. Major reasons for and against using internet hospitals were time-saving (n=500, 66%) and inaccuracy in diagnosis or treatment (n=427, 59.1%), respectively. Meanwhile, cost-saving (n=45, 5.9%) and hard-to-find suitable platforms (n=235, 32.5%) were the least concerned.

Table 3 shows the expectation, perception, and gap scores of the SERVQUAL questionnaire. In the evaluation of Chinese internet hospital service, reliability (mean E 4.12, SD 0.67) received the highest score while empathy (mean E 4.06, SD 0.40) received the lowest score for the expectation perspective, and when it comes to perception perspective, assurance (mean P 3.87, SD 0.74) received the highest score while empathy (mean P 3.79, SD 0.80) still received the lowest score. Across all dimensions, the perception did not meet the expectation of service quality of internet hospitals (namely, all gap scores were negative), and it was noteworthy that the largest gap was observed in the responsiveness dimension.

IPA results are presented in Figure 1. Specifically, the x-axis represents the perception while the y-axis represents the expectation scores, therefore, items in Quadrant I were service features where the performance matched the importance (eg, Q7, protected privacy), while Quadrant II were their counterparts where service qualities should be boosted (eg, Q13, active feedback handling). More importantly, most of the responsiveness features fell into Quadrant II, indicating the perceived responsiveness of internet hospitals in the eyes of patients is less than expected and awaits substantial improvement. Quadrant III consisted of items with low expectations and low perceptions (eg, Q19, personalized care), predominantly the empathy items, where lower expectations

XSL•FO
RenderX

led to these items being given a lower priority. Quadrant IV consisted of items with perceptions exceeding expectations, (eg, Q22, reasonable charge), reflecting the strengths and attractiveness of internet hospital service.

Table 4 presents the results of three multiple linear regression models predicting patients' overall satisfaction with Chinese internet hospitals. Model 1 included five dimensions of internet hospitals' service quality. In Model 2, two variables regarding familiarity and willingness to use internet hospitals were further added. Finally, Model 3 further included gender, age, location, marital status, education level, and annual income using the "enter" method to examine the associations between service quality perspectives, familiarity, and willingness to use internet hospitals above and beyond these control variables.

Model 1 (adjusted R^2=0.526; $F_{5,752}$=168.8; P<.001) revealed that among the included SERVQUAL-C dimensions, only tangibles (P<.001), reliability (P=.06), and empathy (P<.001) were significant and positively associated with patients' overall satisfaction while others were not (Bs<.250; Ps>.066). Among the significant predictors, empathy contributed the most to the explained variance in patient satisfaction.

Model 2 (ΔR^2=0.033; $\Delta F_{7,750}$=29.7; P<.001) revealed that familiarity and willingness were significantly and positively associated with patients' satisfaction, above and beyond the SERVQUAL dimensions, and the inclusion of which further resulted in nonsignificance of the reliability dimension (B=.260; P=.08). Willingness appeared a stronger predictor than familiarity in predicting patient's satisfaction while controlling the SERVQUAL-C dimensions.

Table 2. Reasons for and against using internet hospitals among Chinese patients from June to September 2022[a].

	Count, n (%)
Reasons for us ing internet hospital s (n=758)	
No queuing could save time	500 (66)
No need to commute	364 (48)
Unable or inconvenient to seek offline medical treatment	306 (40.4)
Simplified medical processes in internet hospitals, easy to operate	319 (42.1)
Only mild conditions so no need to go for offline consultation	159 (21)
Access to quality medical resources in more developed areas	126 (16.6)
Internet hospitals are cost-saving	45 (5.9)
Reasons against using internet hospitals (n=723)	
Inaccuracy in diagnosis or treatment, hard to trust	427 (59.1)
Delay in medical treatment due to untimely digital response	385 (53.3)
Internet hospital operation is complicated and inconvenient	290 (40.1)
Worried about the leakage of private, medical information	275 (38)
Hard to find suitable internet hospital platforms	235 (32.5)

[a]Responses are presented in descending order in frequency.

XSL•FO
RenderX

Table 3. Expectations, perceptions, and gaps in service quality of internet hospitals according to SERVQUAL[a] among Chinese patients from June to September 2022 (n=758).

Item	Mean (SD)		Gaps
	E[b]	P[c]	
TG[d]			
Q1: smooth operation	4.14 (0.76)	3.84 (0.87)	–0.30
Q2: rich functions	4.05 (0.82)	3.82 (0.87)	–0.23
Q3: clear layout	4.10 (0.78)	3.84 (0.89)	–0.26
Q4: aesthetical layout	4.01 (0.81)	3.85 (0.86)	–0.16
Dimension mean	4.08 (0.68)	3.84 (0.76)	–0.24
RA[e]			
Q5: comprehensive departments	4.10 (0.80)	3.83 (0.88)	–0.27
Q6: professional health care	4.11 (0.78)	3.89 (0.85)	–0.22
Q7: protected privacy	4.15 (0.81)	3.85 (0.88)	–0.30
Q8: reliable diagnosis	4.12 (0.78)	3.81 (0.87)	–0.31
Q9: dependable service	4.11 (0.75)	3.83 (0.90)	–0.28
Dimension mean	4.12 (0.67)	3.84 (0.73)	–0.28
RS[f]			
Q10: accurate guidance	4.10 (0.77)	3.84 (0.84)	–0.26
Q11: Prompt response	4.10 (0.80)	3.78 (0.97)	–0.32
Q12: Clear feedback pathway	4.10 (0.79)	3.78 (0.94)	–0.32
Q13: Active feedback handling	4.11 (0.80)	3.78 (0.95)	–0.33
Dimension mean	4.10 (0.69)	3.80 (0.81)	–0.30
AS[g]			
Q14: trustworthiness of the platform	4.09 (0.77)	3.82 (0.88)	–0.27
Q15: Satisfaction of the demand	4.06 (0.78)	3.79 (0.89)	–0.27
Q16: Attitude of the doctor	4.12 (0.77)	3.96 (0.82)	–0.16
Q17: Rationality of the process	4.13 (0.77)	3.88 (0.85)	–0.25
Q18: Standardization of the operation	4.12 (0.77)	3.90 (0.86)	–0.22
Dimension mean	4.10 (0.67)	3.87 (0.74)	–0.23
EP[h]			
Q19: personalized care	4.06 (0.82)	3.77 (0.94)	–0.29
Q20: understood need	4.07 (0.80)	3.76 (0.89)	–0.31
Q21: offered medical advice	4.07 (0.78)	3.80 (0.92)	–0.27
Q22: reasonable charge	4.05 (0.80)	3.84 (0.90)	–0.21
Dimension mean	4.06 (0.70)	3.79 (0.80)	–0.27
Overall	4.09 (0.63)	3.83 (0.71)	–0.26

[a]SERVQUAL: Service Quality Questionnaire.
[b]E: Expected service quality estimated by SERVQUAL-C.
[c]P: Perceived service quality estimated by SERVQUAL-C.
[d]TG: Tangibles.
[e]RA: Reliability.
[f]RS: Responsiveness.
[g]AS: Assurance.

XSL•FO
RenderX

[h]EP: Empathy.

Figure 1. Importance-performance analysis of internet hospital service quality among Chinese patients from June to September 2022 (n=758). Grid lines represent mean scores of expectation and perception, respectively. Solid dots represent tangibles (namely, Q1 to Q4), triangles represent reliability (namely, Q5 to Q9), diamonds represent responsiveness (namely, Q10 to Q13), squares represent assurance (namely, Q14 to Q18), and circles represent empathy (namely, Q19 to Q22).

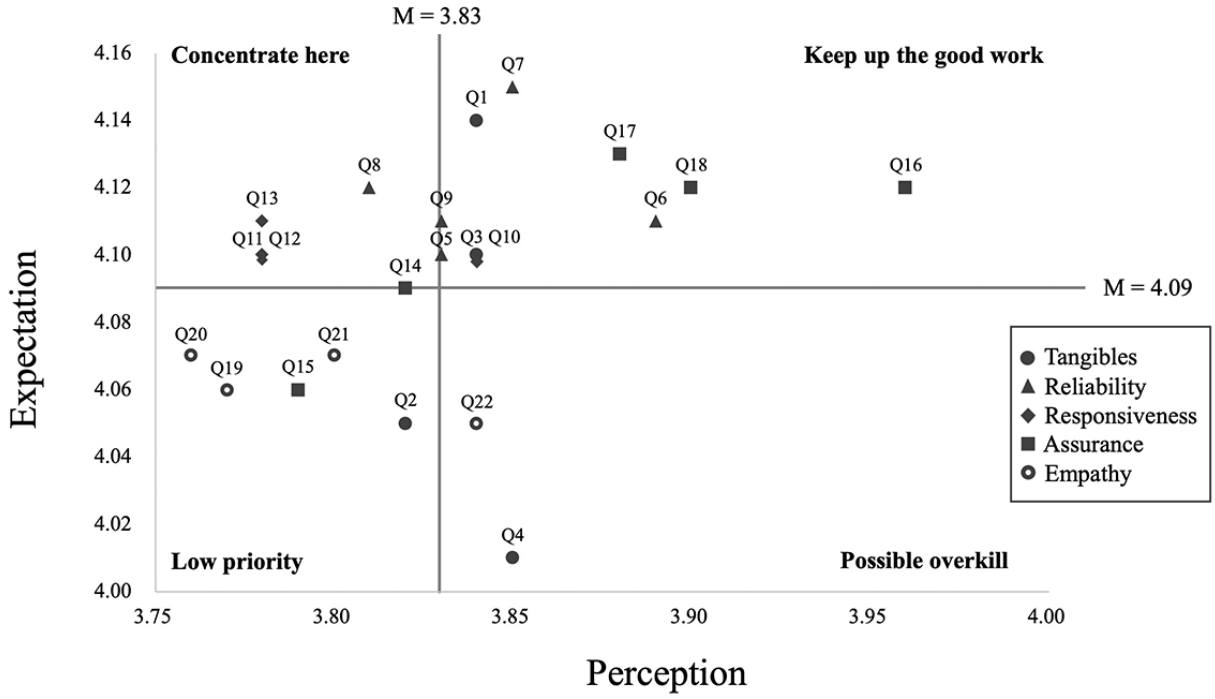

XSL•FO
RenderX

Table 4. Predictors of patient satisfaction with internet hospital services: service quality, familiarity, willingness, and demographics among Chinese patients from June to September 2022 (n=758).

	B	SE	*P* value	95% CI	η^2
Model 1[a]					
Tangibles	.445	0.124	<.001	0.201 to 0.689	0.016
Reliability	.373	0.153	.02	0.073 to 0.673	0.008
Responsiveness	.250	0.136	.07	–0.016 to 0.516	0.004
Assurance	.234	0.160	.15	–0.081 to 0.548	0.003
Empathy	.689	0.137	<.001	0.420 to 0.958	0.032
Model 2[b]					
Tangibles	.359	0.120	.003	0.123 to 0.595	0.012
Reliability	.260	0.148	.08	–0.031 to 0.551	0.004
Responsiveness	.210	0.132	.11	–0.048 to 0.468	0.003
Assurance	.217	0.155	.16	–0.087 to 0.520	0.003
Empathy	.636	0.132	<.001	0.376 to 0.896	0.030
Familiarity	.167	0.057	.004	0.055 to 0.280	0.011
Willingness	.397	0.074	<.001	0.252 to 0.541	0.037
Model 3[c]					
Tangibles	.347	0.119	.005	0.104 to 0.570	0.011
Reliability	.280	0.147	.06	–0.008 to 0.569	0.005
Responsiveness	.217	0.130	.09	–0.038 to 0.472	0.004
Assurance	.251	0.153	.10	–0.049 to 0.551	0.004
Empathy	.575	0.130	<.001	0.319 to 0.831	0.026
Familiarity	.131	0.059	.03	0.015 to 0.247	0.007
Willingness	.384	0.073	<.001	0.241 to 0.527	0.037
Gender	–.233	0.094	.01	–0.417 to –0.049]	0.008
Age (years; reference: ≤ 19)					0.036
20-29	.724	0.161	<.001	0.407 to 1.040	
30-39	.879	0.183	<.001	0.520 to 1.238	
40-49	.520	0.194	.007	0.140 to 0.901	
≥ 50	.631	0.215	.003	0.210 to 1.053	
Location (reference: Northeast Region)					0.005
Eastern Region	–.204	0.217	.35	–0.630 to 0.222	
Central Region	–.379	0.237	.11	–0.843 to 0.086	
Western Region	–.341	0.242	.16	–0.816 to 0.134	
Marital status	–.119	0.094	.20	–0.303 to 0.065	0.002
Education level (reference: < high school)					0.012
High school's degree	–.189	0.195	.33	–0.572 to 0.193	
Bachelor's degree	–.442	0.179	.01	–0.794 to –0.090	
> Bachelor's degree	–.432	0.235	.07	–0.893 to 0.029	
Annual income (reference: ≤ ¥50,000)					0.010
¥50,001-100,000	–.319	0.131	.02	–0.575 to 0.063	
¥100,001-150,000	–.108	0.146	.46	–0.395 to 0.179	

XSL•FO
RenderX

	B	SE	*P* value	95% CI	η^2
¥150,001-200,000	.010	0.186	.96	–0.354 to0.375	
≥ ¥200,001	–.193	0.196	.33	–0.579 to 0.193	

[a]Model 1: $F_{5,752}$=168.8; P<.001, R^2=0.526.

[b]Model 2: $\Delta F_{7,750}$=29.7; P<.001; ΔR^2=0.033.

[c]Model 3: $\Delta F_{23,734}$=6.48; P<.001; ΔR^2=0.055.

Model 3 (R^2=0.055; $\Delta F_{23,734}$=6.48; P<.001) revealed that gender, age, education level, and annual income were all significant predictors of patients' overall satisfaction with Chinese internet hospital service while marital status was not. Specifically, patients with a bachelor's degree were significantly less satisfied with internet hospitals compared to those with a high school (or below) degree (ie, the reference group; B=–.442; P=.01; 95% CI –0.794 to –0.090); and patients with an annual income between ¥50,001 and ¥100,000 were significantly less satisfied with internet hospitals compared to those with an annual income below ¥50,000 (US $ 7019.91; ie, the reference group; B=–.319; P=.02; 95% CI –0.575 to –0.063). Most importantly, after controlling for the patient's demographic characteristics, the effects of the service perspectives (tangibles and empathy), as well as familiarity and willingness remained significant. Overall, the 5 SERVQUAL-C dimensional scores contributed significantly to the explanation of variances in patients' satisfaction, while the inclusion of familiarity and willingness, as well as the sociodemographic variables, only slightly improved the explanation.

Discussion

Principal Findings

This study evaluated the service quality of internet hospitals in China via the Chinese version of SERVQUAL from the perspective of patients, and associated patients' satisfaction. To summarize, the current service quality of Chinese internet hospitals as perceived by patients did not yet meet their expectations, especially in terms of service responsiveness. Factors associated with patient satisfaction suggested that patient familiarity with and willingness to use internet hospitals cast on their overall satisfaction.

In our research, we found that more than half of the sampled participants had used internet hospitals. However, there is a big gap in the penetration rate of internet hospitals between the young and the older people older than 60 years, which may be related to the older people are not used to the internet, and the psychological resistance caused by the fear of fraud. Our results also demonstrated that education level and annual income played an important role in determining whether patients had used internet hospitals. This finding emphasized the necessity of lowering the threshold and fee of using internet hospitals, which can be achieved by creating more user-friendly web pages, providing more smooth operation, or introducing more favorable policies [16], etc.

Furthermore, there was diversity in the patients' familiarity and willingness to use internet hospitals. The majority of participants found it hard to evaluate their familiarity with internet hospitals, yet many more expressed interest and willingness to use this service. This could be attributable to the lack of publicity and promotion of internet hospital services so patients were not yet aware of the possible convenience and applicability of Chinese internet hospitals. At the same time, the patients' stronger willingness also pointed out that internet hospitals' functions could serve as an important supplement to the current health care system and associated health care service to further improve both the health care service and patients' experience. It was noteworthy that the COVID-19 pandemic had somehow fastened the development of the internet hospital service scope and quality, especially in China [17]. Therefore, it is expected that patients' willingness toward internet hospitals might further enhance to the extent that they become more and more familiar with its services.

Our results also indicated that patients' expectations as regards the quality of the provided services are not met. All of the five quality dimensions have a negative gap between patients' expectations and perceptions. These results are similar to other studies using the SERVQUAL model in China [18] and other countries such as Greece [19-21]. The greatest gap lies in responsiveness (–0.308) and reliability (–0.276). However, the difference between the two dimensions of expectations and perceptions does not necessarily indicate a low quality of service, but rather that the patient's requirements are not met, thus leading to patient dissatisfaction. IPA results showed that prompt response, clear feedback pathway, and active feedback handling demonstrate high expectations but low perception levels, which covered three of the four items of responsiveness. Gaps revealed in IPA corroborated patients' reported reasons for and against using internet hospitals. Specifically, timely responses, as well as diagnosis and treatment accuracy, were the most concerning reasons. Improvement in these perspectives would significantly enhance patients' willingness toward internet hospitals. Another item that warranted research attention was reliable diagnosis, it was undeniable that the convenience of digital interaction inevitably introduced higher risks of misdiagnosis. Therefore, developers should pay more attention to the importance of user-friendly interfaces and feedback from internet hospital users to promote efficient and effective responses. Furthermore, more policies and regulations ought to be adopted to regulate information security and digital health care processes to minimize the perceived information risk and medical risks.

Results revealed that patient-perceived tangibility, reliability, and empathy positively predicted satisfaction, with increased empathy being particularly effective in improving satisfaction. This may be due to the fact that patients originally have low expectations of empathy, and once a high-quality empathic

XSL•FO
RenderX

service is perceived, the sense of surprise that comes with it can substantially increase satisfaction. At the same time, responsiveness, which had a low patient perception score, was unexpectedly not a significant predictor of satisfaction, which may be the result of a large gap between perceived and expected aspects of responsiveness, and the fact that if responsiveness is not currently sufficient for normal service delivery, patients will also see improved responsiveness as a basic requirement that should have been achieved and will not be able to establish a link with satisfaction.

In addition, patients' familiarity and willingness to use the internet hospital were similarly significantly and positively associated with satisfaction, and the association of these dimensions on satisfaction was not affected by the introduction of demographic variables. Therefore, strengthening the publicity and promotion of internet hospitals among the patient population so that more people can learn to use internet hospitals more conveniently and quickly, as well as increasing the willingness of patients to use internet hospitals when they have a need for medical care, can better improve satisfaction and establish a more amicable and stable digital consultation platform.

Conclusions

Using the SERVQUAL questionnaire, this study shed light on gaps between the expected and the perceived service quality of Chinese internet hospitals from patients' perspectives. In the future, internet hospitals should focus more on how to improve reliable diagnosis, prompt response, and active and clear feedback pathways. In this way, internet hospitals could improve patients' satisfaction and provide superior telemedicine services, and consequently, patient's willingness to use internet hospitals in need.

Acknowledgments

We thank all the participants and staff of this study. No generative artificial intelligence was used in the writing of this manuscript.

Data Availability

The datasets analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TH designed the study, conducted all statistical analyses, and wrote the manuscript. QW organized project implementation and wrote the manuscript. TH and QW contributed equally and share the first authorship. RW assisted with the results section. YC assisted with the discussion section. HZ provided data for tables. JC provided data for the figure. ZZ designed and supervised the study. SL provided methodology and data verification and supervised the completion of manuscript revisions. ZZ (zhangzhiruo@sjtu.edu.cn) and SL contributed equally and share correspondence. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Xie X, Zhou W, Lin L, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]
2. Ge F, Qian H, Lei J, Ni Y, Li Q, Wang S, et al. Experiences and challenges of emerging online health services combating COVID-19 in China: retrospective, cross-sectional study of internet hospitals. JMIR Med Inform 2022;10(6):e37042 [FREE Full text] [doi: 10.2196/37042] [Medline: 35500013]
3. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021;23(9):e18307. [doi: 10.2196/18307] [Medline: 34342267]
4. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021;23(7):e22330. [doi: 10.2196/22330] [Medline: 34283025]
5. Chen J, Jie L, Su X, Zhang X, Guo D, Wang J, et al. COVID-19 pandemic and the operation of a tertiary hospital in Tianjin—the impact of internet healthcare on hospital business indicators. Front Public Health 2022;10:892773. [doi: 10.3389/fpubh.2022.892773] [Medline: 35669752]
6. Ding L, She Q, Chen F, Chen Z, Jiang M, Huang H, et al. The internet hospital plus drug delivery platform for health management during the COVID-19 pandemic: observational study. J Med Internet Res 2020;22(8):e19678. [doi: 10.2196/19678] [Medline: 32716892]
7. Qiu Y, Liu Y, Ren W, Qiu Y, Ren J. Internet-based and mobile-based general practice: cross-sectional survey. J Med Internet Res 2018;20(9):e266. [doi: 10.2196/jmir.8378] [Medline: 30257819]

XSL•FO
RenderX

8. Roncero C, Remon-Gallo D, Casado-Espada N, Aguilar L, Gamonal-Limcaoco S, Gallego MT, et al. Healthcare professionals' perception and satisfaction with mental health tele-medicine during the COVID-19 outbreak: a real-world experience in telepsychiatry. Front Psychiatry 2022;13:981346. [doi: 10.3389/fpsyt.2022.981346] [Medline: 36405902]
9. Parasuraman AP, Zeithaml VA, Berry LL. A conceptual model of service quality and its implications for future research. J Mark 1985;49(4):41-50. [doi: 10.1177/002224298504900403]
10. Parasuraman A, Zeithaml VA, Berry LL. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. J Retail 1988;64(1):12-40 [FREE Full text]
11. Parasuraman A, Berry LL, Zeithaml VA. Refinement and reassessment of the SERVQUAL scale. J Retail 1991;67(4):420-450 [FREE Full text]
12. Teshnizi SH, Aghamolaei T, Kahnouji K, Teshnizi SMH, Ghani J. Assessing quality of health services with the SERVQUAL model in Iran. A systematic review and meta-analysis. Int J Qual Health Care 2018;30(2):82-89. [doi: 10.1093/intqhc/mzx200] [Medline: 29408970]
13. Jonkisz A, Karniej P, Krasowska D. The servqual method as an assessment tool of the quality of medical services in selected Asian countries. Int J Environ Res Public Health 2022;19(13):7831. [doi: 10.3390/ijerph19137831] [Medline: 35805492]
14. Nadi A, Shojaee J, Abedi G, Siamian H, Abedini E, Rostami A. Patients' expectations and perceptions of service quality in the selected hospitals. Med Arch 2016;70(2):135-139 [FREE Full text] [doi: 10.5455/medarh.2016.70.135-139] [Medline: 27147790]
15. Martilla JA, James JC. Importance-performance analysis. J Mark 1977;41(1):77-79. [doi: 10.2307/1250495]
16. Li P, Luo Y, Yu X, Wen J, Mason E, Li W, et al. Patients' perceptions of barriers and facilitators to the adoption of E-hospitals: cross-sectional study in Western China. J Med Internet Res 2020;22(6):e17221. [doi: 10.2196/17221] [Medline: 32525483]
17. Xu X, Cai Y, Wu S, Guo J, Yang L, Lan J, et al. Assessment of internet hospitals in China during the COVID-19 pandemic: national cross-sectional data analysis study. J Med Internet Res 2021;23(1):e21825. [doi: 10.2196/21825] [Medline: 33417586]
18. Fan LH, Gao L, Liu X, Zhao SH, Mu HT, Li Z, et al. Patients' perceptions of service quality in China: an investigation using the SERVQUAL model. PLoS One 2017;12(12):e0190123 [FREE Full text] [doi: 10.1371/journal.pone.0190123] [Medline: 29272312]
19. Goula A, Stamouli MA, Alexandridou M, Vorreakou L, Galanakis A, Theodorou G, et al. Public hospital quality assessment. Evidence from Greek health setting using SERVQUAL model. Int J Environ Res Public Health 2021;18(7):3418. [doi: 10.3390/ijerph18073418] [Medline: 33806126]
20. Gadeka D, Esena R. Quality of care of medical imaging services at a teaching hospital in Ghana: clients' perspective. J Med Imaging Radiat Sci 2020;51(1):154-164. [doi: 10.1016/j.jmir.2019.12.005] [Medline: 32081678]
21. Zun AB, Ibrahim MI, Hamid AA. Level of satisfaction on service quality dimensions based on SERVQUAL model among patients attending 1 Malaysia clinic in Kota Bharu, Malaysia. Oman Med J 2018;33(5):416-422. [doi: 10.5001/omj.2018.76] [Medline: 30210721]

Abbreviations

IPA: important-performance analysis
SERVQUAL: Service Quality Questionnaire

Edited by A Mavragani; submitted 05.12.23; peer-reviewed by M Mota Morales, X Ma, LJ Jang; comments to author 23.05.24; revised version received 12.06.24; accepted 26.09.24; published 08.11.24.

Please cite as:
Han T, Wei Q, Wang R, Cai Y, Zhu H, Chen J, Zhang Z, Li S
Service Quality and Patient Satisfaction of Internet Hospitals in China: Cross-Sectional Evaluation With the Service Quality Questionnaire
J Med Internet Res 2024;26:e55140
URL: https://www.jmir.org/2024/1/e55140
doi:10.2196/55140
PMID:

©Tao Han, Qinpeng Wei, Ruike Wang, Yijin Cai, Hongyi Zhu, Jiani Chen, Zhiruo Zhang, Sisi Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Design Guidelines for Improving Mobile Sensing Data Collection: Prospective Mixed Methods Study

Christopher Slade[1,2], MS; Roberto M Benzo[3,4], MS, PhD; Peter Washington[2], PhD

[1]Computer Science Department, Brigham Young University—Hawaii, Laie, HI, United States

[2]Information and Computer Sciences Department, University of Hawaii at Manoa, Honolulu, HI, United States

[3]Division of Cancer Prevention & Control, Department of Internal Medicine, Wexner Medical Center, Ohio State University, Columbus, OH, United States

[4]Arthur G James Cancer Hospital, The Ohio State University Comprehensive Cancer Center, Columbus, OH, United States

Corresponding Author:
Christopher Slade, MS
Computer Science Department
Brigham Young University—Hawaii
55-220 Kulanui Street #1919
Laie, HI, 96762
United States
Phone: 1 8086753471
Email: christopher.slade@byuh.edu

Abstract

Background: Machine learning models often use passively recorded sensor data streams as inputs to train machine learning models that predict outcomes captured through ecological momentary assessments (EMA). Despite the growth of mobile data collection, challenges in obtaining proper authorization to send notifications, receive background events, and perform background tasks persist.

Objective: We investigated challenges faced by mobile sensing apps in real-world settings in order to develop design guidelines. For active data, we compared 2 prompting strategies: setup prompting, where the app requests authorization during its initial run, and contextual prompting, where authorization is requested when an event or notification occurs. Additionally, we evaluated 2 passive data collection paradigms: collection during scheduled background tasks and persistent reminders that trigger passive data collection. We investigated the following research questions (RQs): (RQ1) how do setup prompting and contextual prompting affect scheduled notification delivery and the response rate of notification-initiated EMA? (RQ2) Which authorization paradigm, setup or contextual prompting, is more successful in leading users to grant authorization to receive background events? and (RQ3) Which polling-based method, persistent reminders or scheduled background tasks, completes more background sessions?

Methods: We developed mobile sensing apps for iOS and Android devices and tested them through a 30-day user study asking college students (n=145) about their stress levels. Participants responded to a daily EMA question to test active data collection. The sensing apps collected background location events, polled for passive data with persistent reminders, and scheduled background tasks to test passive data collection.

Results: For RQ1, setup and contextual prompting yielded no significant difference (ANOVA $F_{1,144}$=0.0227; P=.88) in EMA compliance, with an average of 23.4 (SD 7.36) out of 30 assessments completed. However, qualitative analysis revealed that contextual prompting on iOS devices resulted in inconsistent notification deliveries. For RQ2, contextual prompting for background events was 55.5% (χ^2_1=4.4; P=.04) more effective in gaining authorization. For RQ3, users demonstrated resistance to installing the persistent reminder, but when installed, the persistent reminder performed 226.5% more background sessions than traditional background tasks.

Conclusions: We developed design guidelines for improving mobile sensing on consumer mobile devices based on our qualitative and quantitative results. Our qualitative results demonstrated that contextual prompts on iOS devices resulted in inconsistent notification deliveries, unlike setup prompting on Android devices. We therefore recommend using setup prompting for EMA when possible. We found that contextual prompting is more efficient for authorizing background events. We therefore recommend using contextual prompting for passive sensing. Finally, we conclude that developing a persistent reminder and requiring participants

XSL•FO
RenderX

to install it provides an additional way to poll for sensor and user data and could improve data collection to support adaptive interventions powered by machine learning.

(J Med Internet Res 2024;26:e55694) doi:10.2196/55694

KEYWORDS

mobile health sensing; mHealth; active data collection; passive data collection; ecological momentary assessment; mobile data; mobile phone; machine learning; real-world setting; mixed method; college; student; user data; data consistency

Introduction

Mobile and ubiquitous devices are valuable tools for gathering patient-generated health data in real-world settings [1-11]. Prior to the advent of the smartphone, mobile devices were predominantly used to collect data actively, usually through ecological momentary assessments (EMA). EMA involves "repeated sampling of participants' current behaviors and experiences in real-time in the participants' natural environment" [12-14]. EMA can include completing journals, diaries, and survey questions [15], providing audio or video samples [16,17], or participating in digital or physical tests [18]. As mobile devices evolved to include more sensors and access to health data, mobile apps started to use passive data collection methods [19-22], where mobile devices collect data without involving the end user. Today, machine learning models often use passively recorded sensor data streams as inputs to machine learning models that predict health outcomes or events captured through EMA [23-29] (Figure 1).

Figure 1. Machine learning models often use passively recorded sensor streams as inputs to predict outcomes captured through EMA. We explore the feasibility of collecting both passive and active data on consumer mobile devices to answer the following research questions (RQs): (RQ1) how do contextual prompting and setup prompting affect scheduled notification delivery and the response rate of notification-initiated EMA? (RQ2) Which authorization paradigm, setup or contextual prompting, is more successful in leading users to grant authorization to receive background events? and (RQ3) Which polling-based method, persistent reminders or scheduled background tasks, completes more background sessions? EMA: ecological momentary assessment.

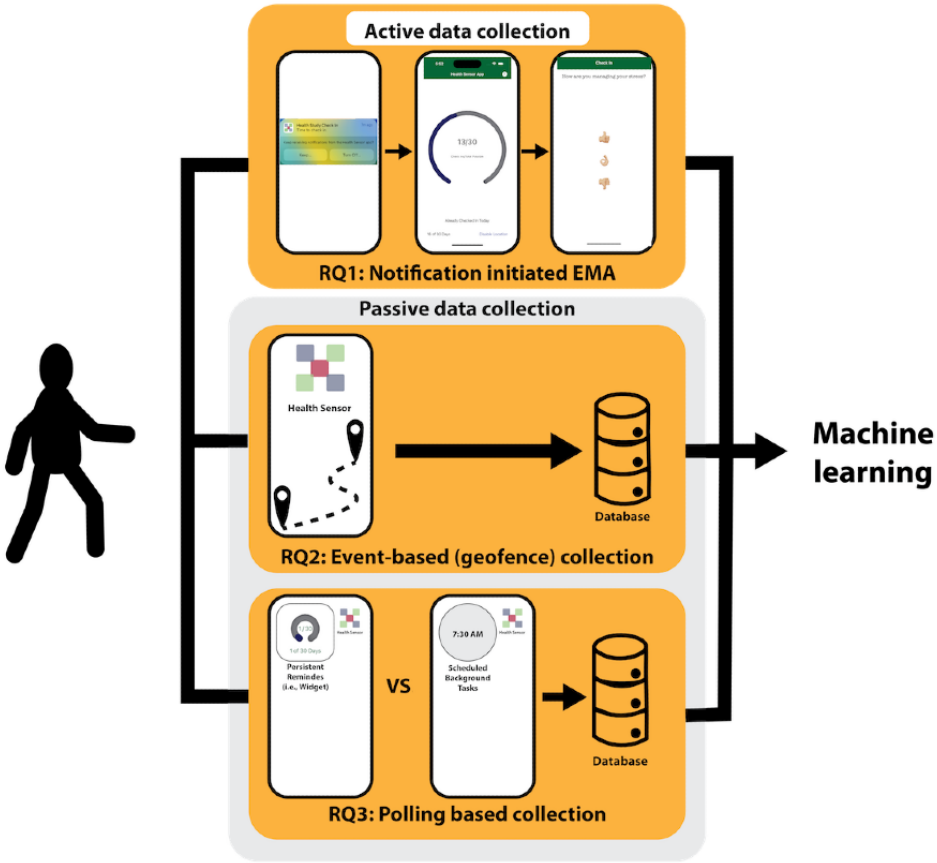

Despite the growth of mobile data collection in health research, several challenges for mobile sensing on consumer mobile devices persist. These challenges are due to implementation decisions made by the developers of major mobile devices (ie, iOS or iPhone and Android) to protect users' privacy, preserve battery life, and minimize distractions. Passive data collection requires access to private user data, whereas active data collection needs to interrupt users to initiate an EMA. In this work, we studied how various user interface (UI) decisions for obtaining authorization at the app level affect the success of both active and passive mobile sensing. Specifically, we tested the following authorization scenarios: (1) users explicitly

XSL•FO
RenderX

granting authorization to receive notifications to initiate EMA, (2) users explicitly granting authorization to access background events, and (3) the system implicitly granting background runtime to collect data through polling.

The specific authorization procedure for the first 2 scenarios varies depending on the device. An Android device obtains authorization through setup prompting, where the user is prompted during the initial app launch. On the other hand, iOS devices use contextual prompting, where the user is prompted when the first event or notification occurs.

Background runtime for passive sensing is often obtained implicitly. Android and iOS systems determine when to run background tasks based on user actions and battery status. Another way to obtain background runtime is through a persistent reminder. A persistent reminder is a permanent UI feature, like a home screen widget or persistent notification, that receives background runtime to update its UI.

To explore these active and passive sensing implementation scenarios, we developed mobile sensing apps for Android and iOS devices that logged both passive and active data. We tested our apps with a user study to answer the following research questions (RQs; Figure 1): (RQ1) how do contextual prompting and setup prompting affect scheduled notification delivery and the response rate of notification-initiated EMA? We hypothesize that contextual prompting will lead to better EMA compliance because participants will not have to respond to a setup prompt and will have more context when authorizing notifications; (RQ2) which authorization paradigm, setup or contextual prompting, is more successful in leading users to grant authorization to receive background events? We hypothesize that the contextual prompts will improve background event authorization because the added context will help participants understand and feel safe approving the permission prompt; and (RQ3) which polling-based method, persistent reminders or scheduled background tasks, completes more background sessions? We hypothesize that persistent reminders will poll for data more often than background tasks because the mobile operating system (OS) is willing to expend resources to keep the UI up-to-date.

Methods

Overview

This section is organized as follows. First, we introduce the authorization procedures that can affect the success of mobile sensing studies. Next, we describe how we used these procedures in our mobile sensing app. Finally, we describe a user study we performed to answer our RQs. A summary of the authorization methods used for each RQ is found in Table 1.

Table 1. Summary of authorization methods used in each research question (RQ).

Condition	Authorization method
RQ1[a]	
Android	Setup prompting for notifications
iOS	Contextual prompting for notifications
RQ2[b]	
Android	Setup prompting for background events
iOS	Contextual prompting for background events
RQ3[c]	
Background tasks	User actions, device status implies consent
Persistent reminders	Installation implies consent

[a]RQ1 compares authorization methods for notification-initiated ecological momentary assessment.

[b]RQ2 compares authorization methods for event-driven data collection.

[c]RQ3 compares polling-based collection methods.

Authorization Procedures

For both Android and iOS, mobile apps must gain explicit authorization before sending notifications or receiving events. Explicit authorization is when a user grants authorization through a permission prompt. At the time of writing, Android devices use setup prompting for explicit authorization, where the user is prompted to grant authorization during setup. The setup prompt can be displayed during the initial launch or when the feature requiring authorization is enabled. iOS devices, on the other hand, use contextual prompting for explicit authorization, where the user is prompted when the first event or notification occurs.

Contextual prompting provides the user with more context to enable a more informed decision. For example, with setup prompting, the user is asked for background location access after consenting, not knowing when or where the location will be accessed. With contextual prompting, the user is prompted when the location is first accessed, helping them understand when their location is accessed. Providing this additional context could help the users feel more comfortable with granting permissions, knowing that the location will be accessed only at a specific location. With setup prompting, on the other hand, we hypothesize that the user might be overwhelmed during setup and deny the request for location or notifications.

XSL•FO
RenderX

While explicit authorization procedures are required for notifications, implicit authorization is needed to grant the background runtime for sensing apps to automatically collect passive sensor and use data. Implicit authorization is when the mobile OS implies authorization based on user actions and device status. Gaining background runtime for scheduled tasks depends on user actions such as app use, swiping an app out of the recent app switcher, and enabling power-saving mode. Device status indicators, such as the battery level, charging state, and network connectivity, are additional factors determining when background runtime is granted. iOS devices always use implicit authorization to determine when to grant runtime. Android devices vary by model and version. However, modern devices include the "adaptive battery" [30] feature, which implicitly grants background runtime.

Installing a permanent UI feature, or persistent reminder, is another way to gain background runtime through implicit authorization. A persistent reminder is part of an app that is always displayed on the mobile device's home screen, lock screen, or notification center. A persistent reminder gains background runtime to keep its UI updated. In this case, the installation of the persistent reminder implicitly grants authorization to run in the background. The persistent reminder can also remind the participant to complete EMA tasks.

Mobile Sensing Apps

We developed a mobile sensing app for iOS and Android devices designed to collect active data through EMA and passive data through polling-based and event-based data collection. The native languages of each OS, Swift for iOS and Kotlin for Android, were used for development. Besides native UI differences and authorization procedures outlined above, iOS and Android apps were designed to appear and function identically. The apps featured a home screen widget displaying the study's progress and serving as a persistent reminder. The main screen, EMA screen, and home screen widget are shown in Figure 2.

Figure 2. Mobile sensing app. Left: the main screen displays the study progress. Middle: the EMA screen, which asks participants about their stress levels. Participants were asked to complete one assessment per day. Right: the home screen widget is used as a persistent reminder. EMA: ecological momentary assessment.

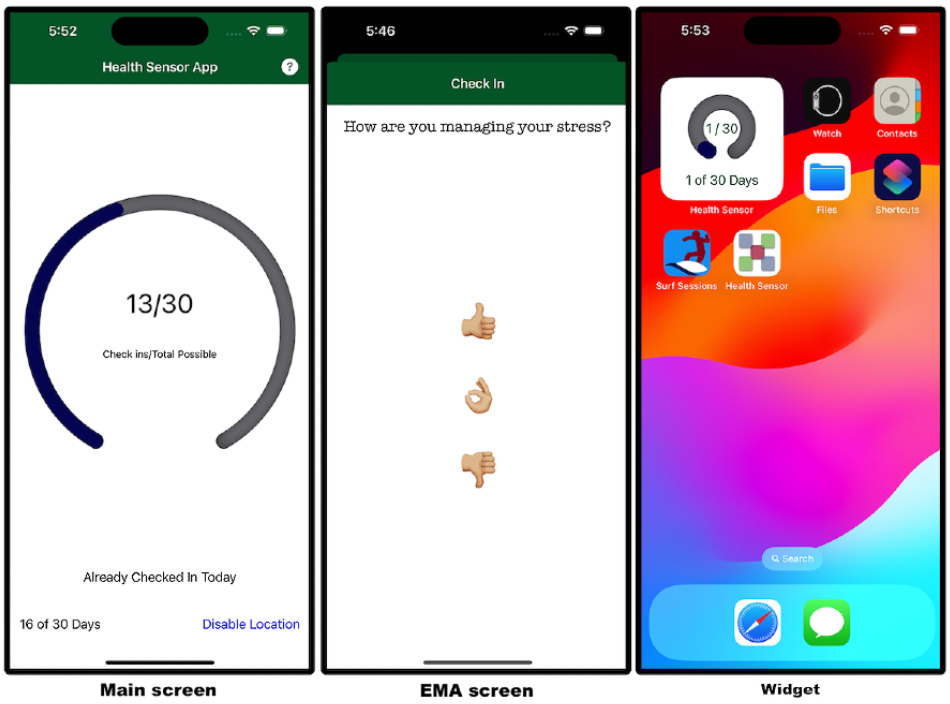

Development Process

We first developed the iOS app and then developed the Android app to match the look and functionality of the iOS app. We tested the apps simultaneously to ensure that Android and iOS devices reported the same data and functioned similarly. We ensured they reported the same number of location events, sent notifications at the same time, and otherwise behaved similarly.

The authorization methods we used in this study reflect each mobile OS' preferred method for gaining authorization. At the time of writing, Android devices do not support contextual prompts for notifications or background events. iOS does not directly support setup prompting for background events and discourages setup prompting for notifications. We chose the OS' preferred authorization method because users should be familiar with the procedure. Figure 3 highlights the authorization differences.

XSL•FO
RenderX

Figure 3. Overview of the difference between authorization procedures on Android and iOS devices. On iOS devices, both notification and background location authorization prompts are received in context. The notification authorization prompt is presented with the first EMA reminder notification, and the background location authorization is presented when the first location event occurs. For Android devices, those prompts are present during the initial launch of the app. EMA: ecological momentary assessment.

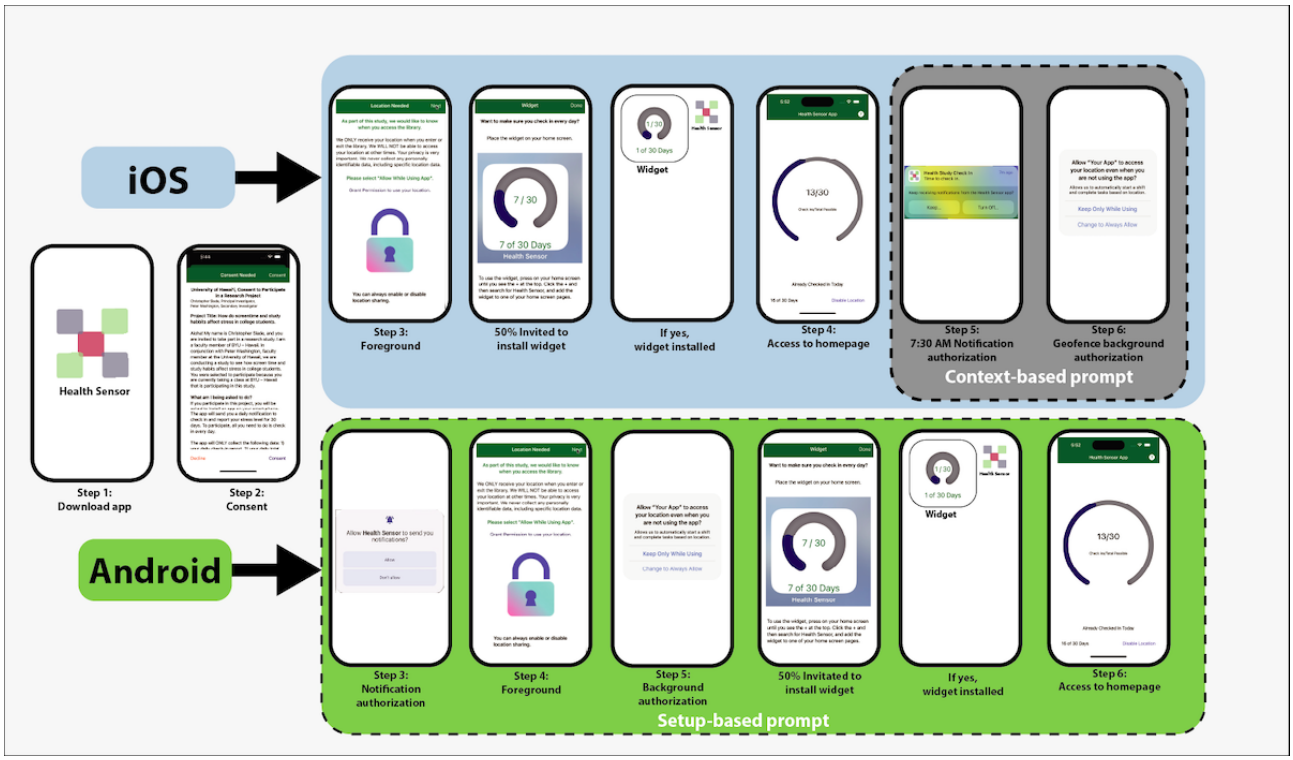

We originally planned to upload all data to secure storage through the background task. However, during testing, we noticed an inconsistency in uploading the data using only the background task. Thus, we also synced data whenever the app was launched in the foreground. To ensure all data were collected, the collection was confirmed before presenting a "Study Finished" screen along with instructions to take a screenshot.

RQ1: Notification-Initiated EMA

RQ1 compares setup and contextual prompting for notification authorization to measure their effect on EMA compliance. We compared EMA compliance using the iOS notification system, which uses contextual prompting, against Android's notification system, which uses setup prompting. To answer RQ1, the apps used notification-initiated EMA to collect active data. We performed a simple single-question EMA once per day, for which both Android and iOS apps reminded users to complete their EMA through a notification. The Android setup prompt and iOS contextual prompt are displayed in Figure 4.

Figure 4. Notification permission prompts. Left: contextual prompt on an iOS device. Right: setup prompt on an Android device. With setup prompting, users are asked to approve notifications during the initial run of the app. Using contextual prompting, users are asked to approve notifications when the first notification arrives.

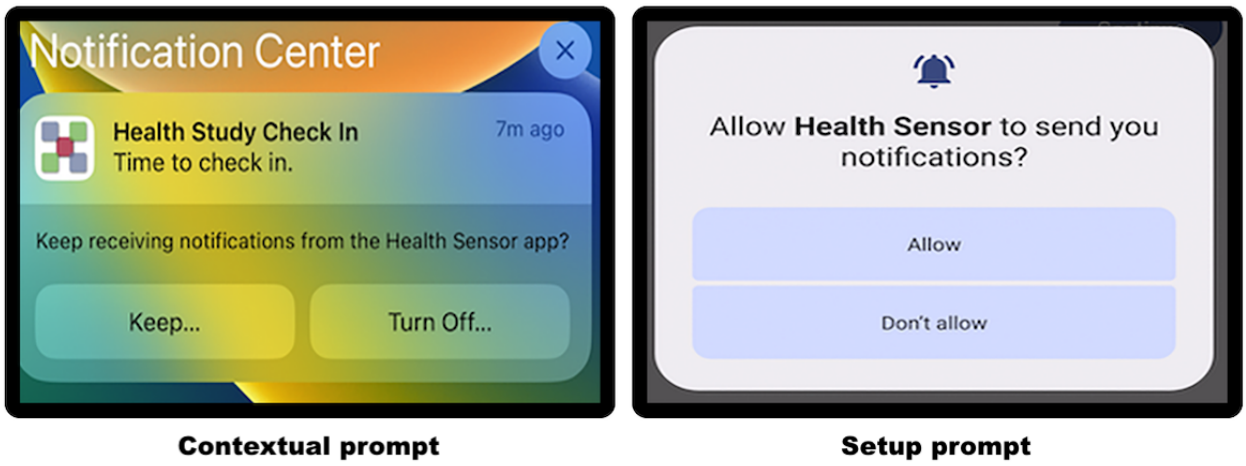

XSL•FO
RenderX

Notification systems on Android and iOS have other features that could affect compliance. iOS 15 introduced "Focus" times [31], which implements notification deferral, and notification summaries [32]. Android devices also implement notification deferral through a focus mode [33], pausing notifications from selected apps when activated. These features might reduce EMA compliance [34,35].

RQ2: Event-Based Passive Data Collection

RQ2 explores how setup and contextual prompting affect the success of event-based passive data collection. In the event-based collection, an event, such as a location change, phone call, message, or health alert, initiates the data collection. Sensing apps need explicit authorization to receive background events. Our apps used location changes as the event of interest and monitored a geofence, or circular region, where participants frequently entered and exited, triggering a location event. iOS devices used contextual prompting to authorize receipt of background location events, and Android devices used setup prompting.

RQ3: Polling-Based Passive Data Collection

RQ3 explores the ability of traditional background tasks and persistent reminders to collect data passively. Both Android and iOS apps implemented and scheduled a daily background task. The background task logged the time and uploaded all logs to cloud storage. We also uploaded the data upon app launch to ensure that all data were collected. For persistent reminders, we implemented a home screen widget, shown in Figure 2, and requested a daily refresh, which was also logged.

We explored users' willingness to install the persistent reminder. Each participant had an equal chance of being assigned to either (1) a control group that did not receive prompts or notifications to install the widget or (2) an experimental group that did receive prompts and notifications. One group of participants was provided verbal instructions for installing the widget.

Alternative Study Designs

We used the mobile OS' default or preferred authorization procedures to test their effectiveness in the wild, similar to other studies that test the feasibility of cross-platform mobile sensing [36-38]. The strength of this approach is that it does not limit the study to users of one particular device (iOS or Android), increasing the diversity of the participants. It also works to understand mobile sensing in the wild, revealing insights into potential pitfalls while developing cross-platform mobile sensing frameworks. However, this limits our ability to isolate study variables because other differences between the mobile OS implementations or device hardware could impact our results.

Studies focusing on a single mobile OS sometimes sacrifice user diversity in favor of isolating variables. These studies are favored when testing new features [39,40] or focusing on specific variables [41]. Android tends to be used in these studies due to its more open programming interface.

User Study

We tested our sensing apps through a user study that aimed to identify the impact of screen time and study time on college students' stress levels. Participants installed our apps on their mobile devices and participated in our study for 30 days. To test event-based collection, a geofence that contained the school library and classrooms was used to calculate study time. Background tasks purported to collect screen time statistics to test polling-based collection.

Active data was collected through a simple EMA that asked participants how they managed their stress levels, as shown in Figure 2. A notification was sent each morning at 7:30 AM to remind users to complete their EMA. Students could complete their daily EMA until the end of the day (midnight). Participants could respond with a thumbs up, thumbs down, or neutral.

Ethical Considerations

The University of Hawaii Institutional Review Board approved the study under protocol #2022-0722, and the Brigham Young University—Hawaii Institutional Review Board approved it under protocol #22-72. All participants consented to participate in the study through the sensing app and received extra credit in their courses for participating. The following measures were implemented to ensure user privacy and participant safety: (1) after installation, the mobile app required the participants to electronically consent before collecting any data; (2) students who did not want to participate were given an alternative extra credit assignment, representing the same time commitment as completing the study; (3) the app anonymized all data collected before being uploaded to cloud storage, so the instructors could not identify the participants' data, including their study habits; (4) follow-up paper surveys were collected by a volunteer, not the class instructor; and (5) at the end of the study, students were instructed by the app to submit a screenshot to their courses' learning management system, and all students were provided with the same amount of extra credit regardless of EMA compliance or the amount of data collected.

Recruitment

We recruited 145 college students in Computer Science, Information Technology, Business, and Science classes at Brigham Young University—Hawaii. The students were offered extra credit incentives to install our app on their mobile devices and actively engage in the study for 30 days.

To ensure students that their professors could not identify their study habits, specific demographic information was not collected, and the app anonymized all data before uploading logs to cloud storage. However, the general demographics of the recruitment base included college students aged 18-26 years, with around 60% being female and 40% being male. Students represented a variety of races and cultures, with a distribution of about 40% White or non-Hispanic, 15% Native Hawaiian and Other Pacific Islanders, 25% Asian, 15% two or more races, and 5% other.

Exit Survey

Upon completion of the study, a follow-up survey was provided to participants, which 48 participants completed. The survey was provided on paper during the course's final exam. To ensure privacy, a volunteer, not the instructor, distributed and collected the survey from participants. To preserve privacy, we did not correlate the survey with the data collected by the mobile app.

XSL•FO
RenderX

The survey allowed us to gather qualitative insights regarding the differing success rates across the authorization paradigms. The participants were asked (1) what reminded them to complete their assessment, (2) their general thoughts on granting background location access, (3) whether they knew about and installed the home screen widget and their thoughts on installing a widget, and (4) suggestions to improve the app.

Results

Data Cleaning and Analysis

We initially logged 178 (126 iOS and 52 Android) users who installed the app and consented to participate in the study. We removed users from the study if the logs showed they participated for fewer than 5 days, which meant that they deleted the app fewer than 5 days after starting. We did not detect any users who deleted the app after participating for 5 days. This yielded 145 participants, representing 105 iOS and 40 Android users. Each participant had an equal chance of being assigned to the widget or control group by the app. A total of 52 individuals (34 iOS and 18 Android) were randomly assigned to the widget group, where the app prompted them to install the widget. In total, 82 participants were randomly assigned to the control group, which did not receive the prompts, and 11 participants (8 iOS and 3 Android) received verbal instructions to install the widget on their home screens.

To conduct our qualitative analysis, we coded responses to the survey based on general patterns that emerged in at least five responses. The questions and the number of themed responses are listed in Table 2.

Table 2. Summary of exit survey responses for each question, separated by device. At the end of the study, participants were administered a paper survey to complete during their final exam, of which 48 participants (33 iOS and 15 Android) completed. Responses were coded based on general themes that emerged in at least 5 responses. Themes with fewer than 5 responses are not listed.

	iOS (n=33), n (%)	Android (n=15), n (%)
Question 1: What reminded you to check-in?		
Notifications	12 (36)	9 (60)
Set alarm	9 (27)	0 (0)
Widget	5 (15)	0 (0)
Saw app	5 (15)	3 (20)
Question 2: Did you allow the app to use your location? Why or why not?		
Yes	25 (76)	15 (100)
No	2 (6)[a]	0 (0)
It would help the study	7 (21)	6 (40)
It was required	4 (12)	3 (20)
Safe	4 (12)	1 (7)
Question 3: Did you install the widget? Why or why not?		
Yes	15 (45)	2 (13)
No	15 (45)	13 (87)
Did not know about it	7 (21)	4 (27)
Helped me remember to check-in	6 (18)	1 (7)
Did not want to change the home screen or do not use widgets	4 (12)	2 (13)
Question 4: Do you have any comments or suggestions on improving the app?		
Better notifications	15 (45)	3 (20)
Select notification time	4 (12)	2 (13)

[a]1 worried about tracking.

RQ1: Notification-Initiated EMA

To answer RQ1, we compared notification authorization procedures: contextual prompting implemented in iOS versus setup prompting implemented in Android by measuring the number of completed EMA. On average, participants across all conditions completed 23.4 (SD 7.38) assessments out of 30. iPhone users completed 23.46 (SD 7.02) assessments, and Android users completed 23.25 (SD 7.9) assessments on average. We observed no statistically significant difference in the number of completed assessments when examining device-specific differences as shown by an ANOVA test ($F_{1,144}$=0.0227; P=.88) or when comparing the widget group to the control group ($F_{1,144}$=1.33; P=.27).

Although our quantitative results demonstrated no difference between devices, our qualitative results indicated that contextual prompting, notification deferral, and notification summaries on iOS devices affected notification delivery. Nine Android users reported that notifications were the primary method of reminding

XSL•FO
RenderX

them to complete their EMA. Despite 9 iOS users stating that the notifications helped them remember to complete the task, 15 iOS users surveyed said the notifications did not appear consistently. Nine participants even set their alarms or reminders. For example, P32 stated, “Notifications weren’t working, so I had an event in my calendar to remind me.”

RQ2: Event-Driven Passive Data Collection Results

In RQ2, we compared the effectiveness of contextual prompting (iOS) and setup prompting (Android) in gaining authorization to access background location events. A total of 28% (11/40) of Android and 49% (51/105) of iOS users enabled background location permission. Contextual and setup prompting differed significantly (χ^2_1=4.4; *P*=04). Thus, contextual prompting was 55.5% more effective than setup prompting in gaining authorization for background location events.

Our poststudy survey asked participants about their willingness to share their location with the sensing app. A total of 40 surveyed participants said they authorized background location access versus 2 participants who reported not sharing their location. Only one participant, P10 (iOS), expressed apprehension about sharing their location, stating, “I was afraid it could track me.” However, 13 participants surveyed expressed a willingness to share their location to contribute to the study’s objectives and thought they granted the needed authorization to access background location events. P46 (Android) exemplified this sentiment: “Yes, if it helps the study then I don’t really mind if it has my location.” User’s willingness to share location data for nonresearch reasons will most likely vary, so we emphasize that these results only apply to the context of research studies.

RQ3: Background Tasks and Persistent Reminders

RQ3 compared traditional background tasks to persistent reminders for polling-based data collection. First, we present our results on how willing users were to install the persistent reminder, and then we present results comparing persistent reminder refreshes to background tasks completed. Finally, we compare background tasks completed by device type and present our qualitative results.

We studied how willing participants were to install the persistent reminder by randomly assigning users to an experimental group where they were notified by the app to install the widget and a control group that did not receive such prompts. In addition, a subgroup of 11 users were given verbal instructions to install the widget. Of 52 participants in the widget group, only 18 (35%) participants installed the home screen widget, with 16 being iOS users and 2 being Android users. One iOS participant in the control group installed the widget. Of the 11 participants who received verbal instructions to install the widget, 5 (46%) participants complied accordingly, of whom 4 were iOS users and 1 was an Android user.

Limiting our data to only users who installed the persistent reminder, on average, the persistent reminder was refreshed 61.2 (SD 16.7) times throughout the study, and the devices performed an average of 7.2 (SD 12.5) background tasks. The two Android devices performed 23 and 42 widget refreshes, with 8 and 6 background tasks, respectively. When users installed the widget, persistent reminders refreshed 266.5% more than background tasks were executed.

Focusing on the difference between Android and iOS, we also observed a significant difference in background tasks completed when comparing devices. iOS devices completed 7.73 (SD 16.32) background sessions during the study. Android devices completed 25.7 (SD 13.96) background sessions. Using data from all participants, an ANOVA test revealed a significant difference in the total number of background sessions between iOS devices and Android devices ($F_{1,144}$=37.8; *P*<.001), showing that Android devices were more permissive in granting background runtime. Most iOS devices completed fewer than 10 background sessions, and many did not execute a single background task. In addition, almost half of the Android devices did not complete a daily background task. This demonstrates our difficulty in consistently gaining background runtime, especially on iOS devices.

Our qualitative analysis showed mixed results regarding the installation of the home screen widget. Several participants indicated that the widget helped them remember to complete their daily assessment, as expressed by seven participants. For example, P37 (iOS) stated, “The widget helped me to remember to check in every day and what the study was about.” However, 6 users opted not to use the widget due to concerns about modifying their home screen or preferring not to use widgets. P22 (iOS) mentioned, “My home screen was already full and organized,” while P35 (Android) stated, “I just don’t want to change my home screen.” iOS users may be more willing to add the widget to their home screen because iOS allows for widget stacking, a feature not found in most Android devices. This functionality enables the rotation of different widgets within a stack, eliminating the need to reorganize the home screen to install the widget.

Effects of Stress Level

Because students experiencing a high stress level might be more motivated to authorize data collection and complete assessments, we divided the participants into 2 groups: a low-stress group (n=117) whose EMA responses averaged better than neutral and a high-stress group (n=28) whose EMA responses averaged lower than neutral. We then compared the two groups’ authorization and data collection rates. There was no statistical difference between the groups in the number of EMA responses ($F_{1,144}$=0.928; *P*=.34), the number of background tasks executed ($F_{1,144}$=1.54; *P*=.22), the number of participants that enabled location (χ^2_1=0.4; *P*=.52), or whether they installed the widget (χ^2_1=0.5; *P*=.57). The stress level of the participants did not significantly impact their willingness to grant authorization, provide EMA responses, or install the widget.

Discussion

Based on our results, we present design principles to enhance the success of mobile passive-sensing research studies.

XSL•FO
RenderX

Notification-Initiated EMA

Setup prompts should generally be used for notifications. Although our quantitative results showed no difference between the iOS and Android notification systems, our qualitative analysis revealed that many iOS users did not consistently receive notifications to remind them to complete their EMA. Nine iOS users reported setting their alarms or calendar reminders to remember to complete their EMA. We conjecture that if this study had required participants to complete multiple tasks per day or to follow a strict time frame for EMA tasks, then EMA compliance on iOS devices would have suffered.

We hypothesized that contextual prompts would work more effectively. However, they lose their context when notification deferrals and summaries are enabled. Only users who searched through their notification summaries would have found the contextual prompt and authorized future notifications. With setup prompting, by contrast, users authorize the notifications immediately after consenting.

Our qualitative results show that notification deferral and notification summaries can cause delayed and missed responses to EMA notifications on iOS devices. New methods should be developed to help participants remember to complete their assessments in a timely manner. iOS allows developers to mark notifications as time-sensitive, which increases their visibility to the user and should be used for time-sensitive EMA. This study did not use time-sensitive notifications because our assessment did not require a time-sensitive response. Future work could analyze the effectiveness of time-sensitive notifications versus basic notifications. However, even time-sensitive notifications could be delayed or ignored based on the user's settings. We recommend that mobile sensing researchers review notification settings during onboarding. Persistent reminders offer an additional way to remind users to complete tasks. Seven participants (6 iOS and 1 Android) reported that the persistent reminder helped them to remember to complete their assessment and could be an area to explore further.

Besides notification deferral, notification summaries, and contextual prompting, other notification system advances can potentially disrupt notification-initiated EMA. Because notifications disrupt users, sometimes causing stress [42-44], several modifications to notification systems could be introduced to consumer mobile devices. Lin et al [45] demonstrated how notification summaries could be improved by letting users determine the order of the notifications. Pejovic et al [46] explored user contexts to understand user interruptability, leading to the development of intelligent notification systems [34,47]. Kandappu et al [48] also explored intelligently interrupting users. Mobile sensing apps must adjust to these advanced notification systems or find a different way to initiate EMA. For example, Zhang et al [49] explored using a persistent reminder on the lock screen to initiate EMA.

Event-Driven Passive Data Collection

For event-driven data collection, use carefully designed contextual prompts. Our qualitative analysis revealed that most participants were willing to authorize background location access for research studies and were under the impression that they enabled background location access. However, our quantitative results revealed that many participants did not authorize background location access. Many failed to grant authorization for background events even though they intended to authorize it. As we hypothesized, contextual prompting on iOS was 55.5% more effective in gaining authorization for background location events. This coincides with previous work that shows additional context helps users make better privacy decisions [39,50,51].

Even with contextual prompts, less than half of the participants authorized access to background location events. Challenges arise when users misinterpret authorization prompts, limiting the success of event-based collection [39,41,51-53]. To improve contextual prompt accuracy, Wijesekera et al [40] developed a machine learning–based, contextual-aware permission model that improves the permission accuracy rate above context prompts, which should be considered if it becomes available on consumer mobile devices. However, the current model suggests providing "generic but well-formed data" when an app is denied access. In the context of health sensing apps, this can skew results and create adverse health interventions. We recommend that such systems be designed to communicate to the app about denials and that apps be constructed to handle denials to ensure data fidelity. Further research must be performed to balance the collection abilities of mobile sensing apps and protect the privacy and security of their users [54].

Polling-Based Passive Data Collection

For polling, implement a persistent reminder that can be used both as a means of collecting data and as a reminder to complete EMA tasks. Collecting data through background tasks presented significant challenges due to the use of an implied consent model for granting background runtime that is inherent to mobile OS, as described in previous work [36,55]. Although Android devices are generally more permissive, too many factors are under consideration to consistently guarantee background runtime. However, as we hypothesized, using a persistent reminder as a secondary means to poll for data yielded more successful data collection. Installing a persistent reminder on the user's home screen signifies to the OS the intent to allocate the necessary resources to maintain the reminder's regular updates, which can also be used for passive data collection.

Many users did not comply with installing the widget when prompted by the app, especially among Android users. Verbal prompts to install the widget did help to improve compliance on Android and iOS. Our qualitative analysis indicated some resistance to installing the widget because users did not want to change their home screen layout, and some were unfamiliar with how to install widgets or did not use widgets.

Several methods are available to overcome the resistance to using widgets. First, participants who more directly benefit from the data collection might be more willing to install a widget. For instance, participants affected by a disease might be more willing to install a widget, especially if the app offers just-in-time, adaptive interventions. In addition, study incentives can also be increased for participants who install the widget. To assist users unfamiliar with widgets, the app can provide a

XSL•FO
RenderX

video tutorial on installing the widget, or instructions can be provided during study onboarding.

Limitations

There are other differences between iOS and Android devices and users of those devices than what we tested for, which could be confounding factors. In addition, we studied passive sensing for research studies, and the results do not necessarily apply to passive sensing in other contexts such as crowdsourcing or commercial purposes.

This study involved college students, and while they represented diverse academic disciplines, races, and countries of origin, they were all aged 19-26 years and tended to be more familiar with mobile devices. Further work should be conducted to observe how these results would translate to a larger, more heterogeneous population. Individuals less familiar with mobile devices, such as geriatric patients, may exhibit more difficulty granting and declining authorization due to different privacy preferences or familiarity with smartphone technology.

For persistent reminders, only two Android users installed the widget, with one device refreshing more than once per day and the other refreshing 23 of the 30 days. These preliminary results are promising. However, further work is required to ensure the results are consistent with a larger sample size.

Future Work

Additional modifications to the mobile apps could be implemented and tested in follow-up studies. Although Android does not allow contextual prompts, contextual prompts could be mimicked by randomly sending a notification asking for background location access or notification access. Such a study would eliminate potential confounding factors that could have influenced the results.

Our sensing apps could be used in additional health studies incorporating other populations with different motivations to comply with study procedures. Comparing our results with studies involving more diverse health issues would be an interesting avenue for future work.

Comparison With Prior Work

Numerous studies have contributed to a comprehensive understanding of EMA compliance across research fields [56]. In a review of EMA studies, Wrzus and Neubauer [57] found that compliance cannot be predicted by the number of assessments or length of the study but that financial incentives did improve compliance rates. Murray et al [58] explored the role of participants' emotional states in affecting compliance. In the related field of crowdsourcing [59], gamification can improve response rates [60]. Other efforts to improve EMA compliance include the work by Schenider et al [61] on just-in-time, adaptive EMA to reduce the burden on participants. Zhang et al [49] explored unlock journaling, where users unlock their devices by completing an EMA. This work built upon this by examining how notification permissions affected EMA compliance.

Mobile sensing depends upon gaining proper authorization to collect data and interrupting participants to initiate an EMA. Research into permissions on mobile devices has shown that users often misinterpret permission prompts [52,53]. Wijesekera et al [39,40] showed that contextual prompts help users correctly interpret permission prompts. Alsoubai et al [41] profiled users to help understand differing privacy strategies, which helps improve intelligent permission systems [40]. Our mobile sensing apps explored contextual prompts and permissions and their role in passive data collection.

Prior work has demonstrated that collecting consistent data across iOS and Android devices remains challenging [62]. Most mobile sensing studies use Android devices due to their more open programming interface, but some work has been conducted to improve mobile sensing on iOS devices. The AWARE-iOS research team [55] explored background data collection methods on iOS and developed guidelines for sustainable data collection on iOS. AWARE-iOS has successfully collected passive data on iOS devices [63]. RADAR-base [64,65] is an open-source mobile health platform for collecting and analyzing large-scale data from various devices including passive and active mobile apps. We expand upon this inspirational prior work by adding design guidelines for consistent active and passive data collection across Android and iOS devices. Consistent data collection is needed to support just-in-time adaptive interventions on consumer mobile devices, necessitating further research [66].

Conclusions

We developed and tested mobile sensing apps for iOS and Android to answer our RQs: (RQ1) How do contextual prompting and setup prompting affect scheduled notification delivery and the response rate of notification-initiated EMA? (RQ2) Which authorization paradigm, setup or contextual prompting, is more successful in leading users to grant authorization to receive background events? and (RQ3) Which polling-based method, persistent reminders or scheduled background tasks, completes more background sessions? Although contextual prompts for notification authorization on iOS devices did not impact EMA compliance rates compared to setup prompts on Android devices, many iOS users reported not receiving notifications. For background event authorization, contextual prompts on iOS devices were 55.5% more effective in gaining authorization than setup prompts on Android devices. Finally, persistent reminders resulted in a completion of background sessions 266.5% more often than when using traditional background tasks. However, we observed some user resistance to installing persistent reminders. Although mobile sensing on consumer mobile devices continues to exhibit challenges, our results suggest that persistent reminders and proper authorization procedures can improve user compliance.

XSL•FO
RenderX

Acknowledgments

The project described was supported by award number 2406251 from the National Science Foundation (NSF) under the Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science program. This research was partially funded by the National Institute of General Medical Sciences grant U54GM138062 and the Medical Research Award fund of the Hawaii Community Foundation grant MedRes_2023_00002689.

Data Availability

The datasets generated during or analyzed during this study are not publicly available due to institutional review board restrictions but are available from the first author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
iCHECK Checklist for Digital Health Implementations.
[PDF File (Adobe PDF File), 172 KB - jmir_v26i1e55694_app1.pdf]

References

1. Jongs N, Jagesar R, van Haren NEM, Penninx BWJH, Reus L, Visser PJ, et al. A framework for assessing neuropsychiatric phenotypes by using smartphone-based location data. Transl Psychiatry 2020;10(1):211 [FREE Full text] [doi: 10.1038/s41398-020-00893-4] [Medline: 32612118]
2. Swain VD, Gao L, Wood WA, Matli SC, Abowd GD, De Choudhury M. Algorithmic power or punishment: information worker perspectives on passive sensing enabled AI phenotyping of performance and wellbeing. 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; April 19, 2023; Hamburg, Germany p. 1-17. [doi: 10.1145/3544548.358137]
3. Rooksby J, Morrison A, Murray-Rust D. Student perspectives on digital phenotyping: the acceptability of using smartphone data to assess mental health. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 02, 2019; Glasgow, Scotland Uk p. 1-14. [doi: 10.1145/3290605.330065]
4. Kline A, Voss C, Washington P, Haber N, Schwartz H, Tariq Q, et al. Superpower glass. GetMobile: Mobile Comput Commun 2019;23(2):35-38. [doi: 10.1145/3372300.3372308]
5. Washington P, Voss C, Kline A, Haber N, Daniels J, Fazel A, et al. SuperpowerGlass: a wearable aid for the at-home therapy of children with autism. Proc ACM Interact Mobile Wearable Ubiquitous Technol 2017;1(3):1-22. [doi: 10.1145/3130977]
6. Washington P, Kline A, Mutlu O, Leblanc E, Hou C, Stockham N, et al. Activity recognition with moving cameras and few training examples: applications for detection of autism-related headbanging. 2021 Presented at: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems; May 8, 2021; New York, NY p. 1-7. [doi: 10.1145/3411763.345170]
7. Washington P, Voss C, Haber N, Tanaka S, Daniels J, Feinstein C, et al. A wearable social interaction aid for children with autism. 2016 Presented at: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems; May 7, 2016; San Jose, CA p. 2348-2354. [doi: 10.1145/2851581.289228]
8. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. JAMA Pediatr 2019;173(5):446-454 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.0285] [Medline: 30907929]
9. Suruliraj B, Bessenyei K, Bagnell A, McGrath P, Wozney L, Orji R, et al. Mobile sensing apps and self-management of mental health during the COVID-19 pandemic: web-based survey. JMIR Form Res 2021;5(4):e24180 [FREE Full text] [doi: 10.2196/24180] [Medline: 33872181]
10. Sun Y, Kargarandehkordi A, Slade C, Jaiswal A, Busch G, Guerrero A, et al. Personalized deep learning for substance use in Hawaii: protocol for a passive sensing and ecological momentary assessment study. JMIR Res Protocol 2024;13:e46493 [FREE Full text] [doi: 10.2196/46493] [Medline: 38324375]
11. Kargarandehkordi A, Slade C, Washington P. Personalized AI-driven real-time models to predict stress-induced blood pressure spikes using wearable devices: proposal for a prospective cohort study. JMIR Res Protocol 2024;13:e55615 [FREE Full text] [doi: 10.2196/55615] [Medline: 38526539]
12. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
13. van Berkel N, Ferreira D, Kostakos V. The experience sampling method on mobile devices. ACM Comput Surv 2017;50(6):1-40. [doi: 10.1145/3123988]

XSL•FO
RenderX

14. Doherty K, Balaskas A, Doherty G. The design of ecological momentary assessment technologies. Interact Comput 2020;32(1):278. [doi: 10.1093/iwcomp/iwaa019]
15. Doherty K, Marcano-Belisario J, Cohn M, Mastellos N, Morrison C, Car J, et al. Engagement with mental health screening on mobile devices: results from an antenatal feasibility study. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 2, 2019; Glasgow, Scotland p. 1-15. [doi: 10.1145/3290605.330041]
16. Huang YN, Zhao S, Rivera M, Hong JI, Kraut R. Predicting well-being using short ecological momentary audio recordings. 2021 Presented at: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems; May 8, 2021; Yokohama, Japan p. 1-7. [doi: 10.1145/3411763.3451629]
17. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall D. A gamified mobile system for crowdsourcing video for autism research. 2018 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); July 26, 2018; New York, NY p. 350-352. [doi: 10.1109/ICHI.2018.00052]
18. Omberg L, Chaibub Neto E, Perumal TM, Pratap A, Tediarjo A, Adams J, et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat Biotechnol 2022;40(4):480-487. [doi: 10.1038/s41587-021-00974-9] [Medline: 34373643]
19. Lane ND, Miluzzo E, Lu H, Peebles D, Choudhury T, Campbell AT. A survey of mobile phone sensing. IEEE Commun Mag 2010;48(9):140-150. [doi: 10.1109/mcom.2010.5560598]
20. Meegahapola L, Gatica-Perez D. Smartphone sensing for the well-being of young adults: a review. IEEE Access 2021;9:3374-3399. [doi: 10.1109/access.2020.3045935]
21. Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. J Med Internet Res 2016;18(3):e72 [FREE Full text] [doi: 10.2196/jmir.5505] [Medline: 27025287]
22. Wang W, Mirjafari S, Harari G, Ben-Zeev D, Brian R, Choudhury T, et al. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. 2020 Presented at: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; April 23, 2020; New York, NY p. 1-15. [doi: 10.1145/3313831.33768]
23. Adler DA, Wang F, Mohr DC, Choudhury T. Machine learning for passive mental health symptom prediction: generalization across different longitudinal mobile sensing studies. PLoS One 2022;17(4):e0266516 [FREE Full text] [doi: 10.1371/journal.pone.0266516] [Medline: 35476787]
24. Assi K, Meegahapola L, Droz W, Kun P, De GA, Bidoglia M, et al. Complex daily activities, country-level diversity,smartphone sensing: a study in Denmark, Italy, Mongolia, Paraguay, and UK. 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; April 19, 2023; New York, NY p. 1-23. [doi: 10.1145/3544548.358119]
25. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and Fitbit data. JMIR Mhealth Uhealth 2019;7(7):e13209 [FREE Full text] [doi: 10.2196/13209] [Medline: 31342903]
26. LiKamWa R, Liu Y, Lane ND, Zhong L. Moodscope: building a mood sensor from smartphone usage patterns. 2013 Presented at: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services; June 25, 2013; New York, NY p. 389-402. [doi: 10.1145/2462456.246444]
27. Nepal S, Wang W, Vojdanovski V, Huckins J, daSilva A, Meyer M, et al. COVID Student Study: a year in the life of college students during the COVID-19 pandemic through the lens of mobile phone sensing. 2022 Apr Presented at: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 28, 2022; New York, NY p. 1-19. [doi: 10.1145/3491102.3502043]
28. Xu X, Chikersal P, Dutcher JM, Sefidgar YS, Seo W, Tumminia MJ, et al. Leveraging collaborative-filtering for personalized behavior modeling. Proc ACM Interact Mobile Wearable Ubiquitous Technol 2021;5(1):1-27. [doi: 10.1145/3448107]
29. Lind MN, Byrne ML, Wicks G, Smidt AM, Allen NB. The effortless assessment of risk states (EARS) tool: an interpersonal approach to mobile sensing. JMIR Ment Health 2018;5(3):e10334 [FREE Full text] [doi: 10.2196/10334] [Medline: 30154072]
30. Birney A. Android adaptive battery: everything you need to know. Android Authority. 2024 Feb 26. URL: https://www.androidauthority.com/android-adaptive-battery-explained-3223097/# [accessed 2024-11-12]
31. Use focus on your iPhone or iPad. Apple support. URL: https://support.apple.com/en-us/HT212608 [accessed 2023-11-13]
32. Auda J, Weber D, Voit A, Schneegass S. Understanding user preferences towards rule-based notification deferral. 2018 Presented at: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems; April 20, 2018; New York, NY p. 1-6. [doi: 10.1145/3170427.3188688]
33. Get into "Focus mode" at work with help from Android. Google Blog. URL: https://tinyurl.com/2rx37d59 [accessed 2023-12-13]
34. Li T, Haines J, De Eguino MFR, Hong J, Nichols J. Alert now or never: understanding and predicting notification preferences of smartphone users. ACM Trans Comput-Hum Interact 2023;29(5):1-33. [doi: 10.1145/3478868]
35. Mehrotra A, Pejovic V, Vermeulen J, Hendley R, Musolesi M. My phone and me: understanding people's receptivity to mobile notifications. 2016 Presented at: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 07, 2016; New York, NY p. 1021-1032. [doi: 10.1145/2858036.2858566]

XSL•FO
RenderX

36. Boonstra TW, Nicholas J, Wong QJ, Shaw F, Townsend S, Christensen H. Using mobile phone sensor technology for mental health research: integrated analysis to identify hidden challenges and potential solutions. J Med Internet Res 2018;20(7):e10131 [FREE Full text] [doi: 10.2196/10131] [Medline: 30061092]
37. Boonstra TW, Werner-Seidler A, O'Dea B, Larsen ME, Christensen H. Smartphone app to investigate the relationship between social connectivity and mental health. Annu Int Conf IEEE Eng Med Biol Soc 2017;2017:287-290. [doi: 10.1109/EMBC.2017.8036818] [Medline: 29059866]
38. Boonstra TW, Larsen ME, Townsend S, Christensen H. Validation of a smartphone app to map social networks of proximity. PLoS One 2017;12(12):e0189877 [FREE Full text] [doi: 10.1371/journal.pone.0189877] [Medline: 29261782]
39. Wijesekera P, Baokar A, Hosseini A, Egelman S, Wagner D, Beznosov K. Android permissions remystified: a field study on contextual integrity. 2015 Presented at: Proceedings of the 24th USENIX Conference on Security Symposium; August 12, 2015; Washington, DC URL: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/wijesekera
40. Wijesekera P, Reardon J, Reyes I, Tsai L, Chen JW, Good N, et al. Contextualizing privacy decisions for better prediction (and protection). 2018 Presented at: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; April 21, 2018; New York, NY p. 1-13. [doi: 10.1145/3173574.3173842]
41. Alsoubai A, Anaraky RG, Li Y, Page X, Knijnenburg B, Wisniewski PJ. Permission vs. app limiters: profiling smartphone users to understand differing strategies for mobile privacy management. 2022 Presented at: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 29, 2022; New York, NY p. 1-18. [doi: 10.1145/3491102.3517652]
42. Yoon SH, Lee SS, Lee JM, Lee K. Understanding notification stress of smartphone messenger app. 2014 Presented at: CHI '14 Extended Abstracts on Human Factors in Computing Systems; April 26, 2014; New York, NY p. 1735-1740. [doi: 10.1145/2559206.2581167]
43. Kang S, Park CY, Kim A, Cha N, Lee U. Understanding emotion changes in mobile experience sampling. 2022 Presented at: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 29, 2022; New Orleans, LA p. 1-14. [doi: 10.1145/3491102.3501944]
44. Chan L, Swain VD, Kelley C, de Barbaro K, Abowd GD, Wilcox L. Students' experiences with ecological momentary assessment tools to report on emotional well-being. Proc ACM Interact Mobile Wearable Ubiquitous Technol 2018;2(1):1-20. [doi: 10.1145/3191735]
45. Lin TC, Su YS, Yang EH, Chen YH, Lee HP, Chang YJ. “Put it on the Top, I’ll Read it Later”: investigating users’ desired display order for smartphone notifications. 2021 Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 7, 2021; Yokohama, Japan p. 1-13. [doi: 10.1145/3411764.3445384]
46. Pejovic V, Musolesi M, Mehrotra A. Investigating the role of task engagement in mobile interruptibility. 2015 Presented at: MobileHCI '15: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct; August 24, 2015; Copenhagen, Denmark p. 1100-1105. [doi: 10.1145/2786567.2794336]
47. Mehrotra A, Musolesi M. Intelligent notification systems: a survey of the state of the art and research challenges. ArXiv. Preprint posted online on November 28, 2017 2018. [doi: 10.48550/arXiv.1711.10171]
48. Kandappu T, Mehrotra A, Misra A, Musolesi M, Cheng S, Meegahapola L. PokeME: applying context-driven notifications to increase worker engagement in mobile crowd-sourcing. 2020 Presented at: CHIIR '20: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval; March 14, 2020; Vancouver BC p. 3-12. [doi: 10.1145/3343413.3377965]
49. Zhang X, Pina LR, Fogarty J. Examining unlock journaling with diaries and reminders for In situ self-report in health and wellness. Proc SIGCHI Conf Hum Factor Comput Syst 2016;2016:5658-5664 [FREE Full text] [doi: 10.1145/2858036.2858360] [Medline: 28804795]
50. Barth A, Datta A, Mitchell J, Nissenbaum H. Privacy and contextual integrity: framework and applications. 2006 Presented at: IEEE Symposium on Security and Privacy (S&P'06); May 21-24, 2006; Berkeley/Oakland, CA p. 1-15. [doi: 10.1109/SP.2006.32]
51. Bonné B, Peddinti S, Bilogrevic I, Taft N. Exploring decision making with Android's runtime permission dialogs using in-context surveys. 2017 Presented at: Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security; July 12-14, 2017; Santa Clara, CA p. 195-210 URL: https://tinyurl.com/2s37s4ss
52. Kelley P, Consolvo S, Cranor L, Jung J, Sadeh N, Wetherall D. A conundrum of permissions: installing applications on an Android smartphone. 2012 Presented at: A Conundrum of Permissions: Installing Applications on an Android Smartphone; March 2, 2012; Berlin, Heidelberg p. 68-79. [doi: 10.1007/978-3-642-34638-5_6]
53. Tan J, Nguyen K, Theodorides M, Negrón-Arroyo H, Thompson C, Egelman S, et al. The effect of developer-specified explanations for permission requests on smartphone user behavior. 2014 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 26, 2014; Toronto, ON p. 91-100. [doi: 10.1145/2556288.2557400]
54. Dehling T, Gao F, Schneider S, Sunyaev A. Exploring the far side of mobile health: information security and privacy of mobile health apps on iOS and android. JMIR Mhealth Uhealth 2015;3(1):e8 [FREE Full text] [doi: 10.2196/mhealth.3672] [Medline: 25599627]
55. Nishiyama Y, Ferreira D, Eigen Y, Sasaki W, Okoshi T, Nakazawa J, et al. IOS crowd–sensing won’t hurt a bit!: AWARE framework and sustainable study guideline for iOS platform. 2020 Presented at: Distributed, Ambient and Pervasive

XSL•FO
RenderX

Interactions: 8th International Conference, DAPI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings; July 19, 2020; Berlin, Heidelberg p. 223-243 URL: https://link.springer.com/chapter/10.1007/978-3-030-50344-4_17
56. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023;19:107-131 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]
57. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
58. Murray AL, Brown R, Zhu X, Speyer LG, Yang Y, Xiao Z, et al. Prompt-level predictors of compliance in an ecological momentary assessment study of young adults' mental health. J Affect Disord 2023;322:125-131 [FREE Full text] [doi: 10.1016/j.jad.2022.11.014] [Medline: 36372127]
59. Jaimes LG, Vergara-Laurens IJ, Raij A. A survey of incentive techniques for mobile crowd sensing. IEEE Internet Things J 2015;2(5):370-380. [doi: 10.1109/jiot.2015.2409151]
60. Morschheuser B, Hamari J, Koivisto J. Gamification in crowdsourcing: a review. 2016 Presented at: 49th Hawaii International Conference on System Sciences (HICSS); January 5-8, 2016; Koloa, HI p. 4375-4384. [doi: 10.1109/HICSS.2016.543]
61. Schneider S, Junghaenel DU, Smyth JM, Fred Wen CK, Stone AA. Just-in-time adaptive ecological momentary assessment (JITA-EMA). Behav Res Methods 2024;56(2):765-783 [FREE Full text] [doi: 10.3758/s13428-023-02083-8] [Medline: 36840916]
62. Blunck H, Kjærgaard M, Bouvin N, Lukowicz P, Franke T, Wüstenberg M, et al. On heterogeneity in mobile sensing applications aiming at representative data collection. 2013 Presented at: UbiComp '13 Adjunct: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication; September 8, 2013; Zurich, Switzerland p. 1087-1098. [doi: 10.1145/2494091.2499576]
63. Nishiyama Y, Ferreira D, Sasaki W, Okoshi T, Nakazawa J, Dey A, et al. Using iOS for inconspicuous data collection: a real-world assessment. 2020 Presented at: UbiComp/ISWC '20 Adjunct: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers; September 12, 2020; Virtual Event, Mexico p. 261-266. [doi: 10.1145/3410530.3414369]
64. Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, Hyve, et al. RADAR-Base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019;7(8):e11734 [FREE Full text] [doi: 10.2196/11734] [Medline: 31373275]
65. Sun S, Folarin AA, Ranjan Y, Rashid Z, Conde P, Stewart C, et al. Using smartphones and wearable devices to monitor behavioral changes during COVID-19. J Med Internet Res 2020;22(9):e19992 [FREE Full text] [doi: 10.2196/19992] [Medline: 32877352]
66. Teepe GW, Da Fonseca A, Kleim B, Jacobson NC, Sanabria AS, Tudor Car L, et al. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. J Med Internet Res 2021;23(9):e29412 [FREE Full text] [doi: 10.2196/29412] [Medline: 34309569]

Abbreviations

EMA: ecological momentary assessment
OS: operating system
RQ: research question
UI: user interface

Edited by A Mavragani; submitted 20.12.23; peer-reviewed by TAR Sure, S Mitra, E Larson; comments to author 06.04.24; revised version received 29.04.24; accepted 07.09.24; published 18.11.24.

Please cite as:
Slade C, Benzo RM, Washington P
Design Guidelines for Improving Mobile Sensing Data Collection: Prospective Mixed Methods Study
J Med Internet Res 2024;26:e55694
URL: https://www.jmir.org/2024/1/e55694
doi:10.2196/55694
PMID:

©Christopher Slade, Roberto M Benzo, Peter Washington. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Smartphone App for Improving Self-Awareness of Adherence to Edoxaban Treatment in Patients With Atrial Fibrillation (ADHERE-App Trial): Randomized Controlled Trial

Minjae Yoon[1], MD; Ji Hyun Lee[1], MD, PhD; In-Cheol Kim[2], MD, PhD; Ju-Hee Lee[3], MD, PhD; Mi-Na Kim[4], MD, PhD; Hack-Lyoung Kim[5], MD, PhD; Sunki Lee[6], MD, PhD; In Jai Kim[7], MD, PhD; Seonghoon Choi[8], MD, PhD; Sung-Ji Park[9], MD, PhD; Taeho Hur[1,10], PhD; Musarrat Hussain[11], PhD; Sungyoung Lee[10], PhD; Dong-Ju Choi[1], MD, PhD

[1]Division of Cardiology, Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Republic of Korea

[2]Division of Cardiology, Department of Internal Medicine, Keimyung University Dongsan Hospital, Keimyung University School of Medicine, Daegu, Republic of Korea

[3]Division of Cardiology, Department of Internal Medicine, Chungbuk National University Hospital, Chungbuk National University College of Medicine, Cheongju, Republic of Korea

[4]Department of Cardiology, Korea University Anam Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[5]Division of Cardiology, Seoul National University Boramae Medical Center, Seoul, Republic of Korea

[6]Department of Cardiology, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[7]Division of Cardiology, Department of Internal Medicine, Bundang CHA Medical Center, CHA University, Seongnam, Republic of Korea

[8]Division of Cardiology, Department of Internal Medicine, Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Republic of Korea

[9]Division of Cardiology, Department of Internal Medicine, Heart Vascular Stroke Institute, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[10]Department of Computer Science and Engineering, Kyung Hee University, Yongin, Republic of Korea

[11]Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway

Corresponding Author:
Dong-Ju Choi, MD, PhD
Division of Cardiology, Department of Internal Medicine
Seoul National University Bundang Hospital
Seoul National University College of Medicine
82, Gumi-ro 173 Beon-gil,
Bundang-gu
Seongnam, 13620
Republic of Korea
Phone: 82 31 787 7007
Fax: 82 31 787 7041
Email: djchoi.snu@gmail.com

Abstract

Background: Adherence to oral anticoagulant therapy is essential to prevent ischemic stroke in patients with atrial fibrillation (AF).

Objective: This study aimed to evaluate whether smartphone app–based interventions improve medication adherence in patients with AF.

Methods: This open-label, multicenter randomized controlled trial (ADHERE-App [Self-Awareness of Drug Adherence to Edoxaban Using an Automatic App Feedback System] study) enrolled patients with AF treated with edoxaban for stroke prevention. They were randomly assigned to app-conditioned feedback (intervention; n=248) and conventional treatment (control; n=250) groups. The intervention group received daily alerts via a smartphone app to take edoxaban and measure blood pressure and heart rate at specific times. The control group received only standard, guideline-recommended care. The primary end point was edoxaban

XSL•FO
RenderX

adherence, measured by pill count at 3 or 6 months. Medication adherence and the proportion of adequate medication adherence, which was defined as ≥95% of continuous medication adherence, were evaluated.

Results: Medication adherence at 3 or 6 months was not significantly different between the intervention and control groups (median 98%, IQR 95%-100% vs median 98%, IQR 91%-100% at 3 months, *P*=.06; median 98%, IQR 94.5%-100% vs median 97.5%, IQR 92.8%-100% at 6 months, *P*=.15). However, the proportion of adequate medication adherence (≥95%) was significantly higher in the intervention group at both time points (76.8% vs 64.7% at 3 months, *P*=.01; 73.9% vs 61% at 6 months, *P*=.007). Among patients aged >65 years, the intervention group showed a higher medication adherence value and a higher proportion of adequate medication adherence (≥95%) at 6 months.

Conclusions: There was no difference in edoxaban adherence between the groups. However, the proportion of adequate medication adherence was higher in the intervention group, and the benefit of the smartphone app–based intervention on medication adherence was more pronounced among older patients than among younger patients. Given the low adherence to oral anticoagulants, especially among older adults, using a smartphone app may potentially improve medication adherence.

Trial Registration: International Clinical Trials Registry Platform KCT0004754; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=28496&search_page=L

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-048777

(J Med Internet Res 2024;26:e65010) doi:10.2196/65010

KEYWORDS

mobile apps; digital health; atrial fibrillation; anticoagulants; medication adherence; mobile phone

Introduction

Stroke prevention with oral anticoagulants is essential in the management of patients with atrial fibrillation (AF) [1,2]. Previous large randomized controlled trials for AF have shown improved safety and comparable efficacy of nonvitamin K antagonist oral anticoagulants (NOACs) compared with warfarin [3]. Thus, current guidelines prefer NOACs over warfarin for ischemic stroke prevention in patients with AF [1,2]. However, with the relatively short half-life of NOACs, when doses are missed, there is a higher risk of a prothrombotic state [4]. Therefore, strict adherence to NOAC therapy is essential to maintain a consistent anticoagulant effect on stroke prevention.

Telemonitoring, feedback, and education result in higher NOAC adherence [5,6]. However, monitoring and feedback by health care professionals require high costs and human resources and, thus, are challenging to incorporate in real clinical practice. With advancements in mobile technology and artificial intelligence, smartphone apps may be a promising strategy to improve medication adherence with ease of use [7]. Although there is some debate about their effectiveness, smartphone apps are currently recognized as an effective method of improving medication adherence [8-14]. Smartphone apps can send push alerts to remind patients to take their medication, which can help improve drug adherence. In addition, active involvement in measuring patients' health status, including blood pressure (BP) using smartphone apps can improve awareness of their underlying conditions. This, in turn, influences self-care behavior and may increase medication adherence [15].

There have been previous studies on mobile apps for improving adherence to NOACs [16-18]. However, these studies have a small number of participants and limited app functions. Given the low adherence to NOACs in patients with AF in the real world [19-21], we developed a smartphone app that provides push alerts to take NOACs and measure BP and heart rate (HR) at specific times of the day. This study aimed to evaluate whether a smartphone app–based intervention would improve medication adherence compared with usual care in a large population of patients with AF requiring NOACs.

Methods

Study Design

The ADHERE-App (Self-Awareness of Drug Adherence to Edoxaban Using an Automatic App Feedback System) study is a prospective, open-label, multicenter randomized controlled trial aimed at evaluating the effect of a smartphone app on improving drug adherence to NOAC in patients with AF. Nine tertiary university hospitals in South Korea participated in this study, which was registered in the International Clinical Trials Registry Platform (KCT0004754).

Ethical Considerations

This clinical trial was approved by the Institutional Review Boards of Seoul National University Bundang Hospital (B-1904-532-301), Keimyung University Dongsan Hospital, Chungbuk National University Hospital, Korea University Anam Hospital, Seoul National University Boramae Medical Center, Bundang CHA Medical Center, Dongtan Sacred Heart Hospital, Kangnam Sacred Heart Hospital, and Samsung Medical Center. This study was conducted per the Declaration of Helsinki, and all study data were deidentified. All patients provided written informed consent on enrollment.

Study Population

We enrolled patients with AF aged ≥19 years requiring oral anticoagulation therapy (CHA_2DS_2-VASc score ≥1 point for men and ≥2 points for women) for stroke prevention [1]. Patients already taking edoxaban or patients who had switched to edoxaban from other NOACs were enrolled. Edoxaban (Lixiana, Daiichi Sankyo) was administered to patients at an on-label dosage of 60 mg once daily. The dose was reduced to 30 mg once daily in patients who met any of the following criteria:

XSL•FO
RenderX

moderate renal impairment (creatinine clearance, 30-50 mL/min), body weight ≤60 kg, or concomitant use of potent P-glycoprotein inhibitors. As the use of smartphones was essential in this study, the participants were required to be familiar with the use of Android smartphones and able to understand and follow Korean instructions for using the app. Patients with severe renal insufficiency (creatinine clearance <15 mL/min), moderate or severe mitral valve stenosis, or a history of mitral valve replacement or repair were excluded. The detailed inclusion and exclusion criteria are listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

1. Patients with atrial fibrillation aged 19 years or older requiring oral anticoagulation therapy (CHA_2DS_2-VASc score ≥1 point for men and ≥2 points for women)
2. Patients who are already taking edoxaban or are initiating newly prescribed edoxaban
3. Patients who can use Android smartphones and are fluent in Korean
4. Patients who voluntarily consent to participate in this clinical trial

Exclusion criteria

1. Creatinine clearance less than 15 mL/min
2. Patients with dual antiplatelet therapy
3. Moderate or severe mitral stenosis
4. Previous history of mitral valve replacement or mitral valve repair
5. Previous history of alcohol or drug abuse
6. Patients with a mean systolic blood pressure ≥200 mm Hg or diastolic blood pressure ≥110 mm Hg during the screening visit
7. Patients with potential pregnancy or those who are breastfeeding
8. Patients who are judged as both legally and psychologically inadequate to participate in the clinical study by the investigator
9. Patients who have participated in clinical studies with other investigational drug products within 4 weeks before screening
10. Patients unwilling to participate in the clinical study

The CHA_2DS_2-VASc score for ischemic stroke risk prediction was calculated by the summation of all assigned points: one point each for congestive heart failure (C), hypertension (H), being aged between 65 and 74 years (A), diabetes mellitus (D), vascular disease (V), and female sex (Sc) and 2 points each for a history of stroke, transient ischemic attack, or thromboembolism (S_2) or being aged ≥75 years (A_2) [1,22]. Congestive heart failure was defined as clinical heart failure regardless of ejection fraction, objective evidence of moderate to severe left ventricular dysfunction, or hypertrophic cardiomyopathy. Vascular disease was defined as angiographically significant coronary artery disease, previous myocardial infarction, peripheral vascular disease, or aortic plaques.

Patient Recruitment and Randomization

Any patient who presented to the clinic for AF management was considered a potential participant. The research nurses conducted an initial screening of the patients' eligibility criteria and access to a smartphone. Of these patients, after a final comprehensive screening and interview, written informed consent was obtained from eligible participants. After enrollment, the participants were randomized 1:1 to either the intervention group (app-conditioned feedback group) or the control group (conventional treatment group) using a web-based central randomization service [23]. Randomization was performed by research nurses using permuted block randomization derived from SAS (version 9.2, SAS Institute). Owing to the nature of the intervention, this study's participants and investigators interacting with the patients were not blinded to the group allocation. All the participants were informed that information on edoxaban adherence would be collected for analysis.

Intervention (Smartphone App–Conditioned Feedback) and Follow-Up

We developed a smartphone app that provides push alerts to take edoxaban and measure BP and HR at specific times of the day. After randomization, this study's participants in the intervention group installed this study's app provided by the research nurses on their Android smartphones. As it was not an open app accessible to the public, the control group could not access it. The operating system and display of the smartphone app are described in a previously published design paper [24] and in Multimedia Appendix 1.

In brief, the research nurses initially set an alarm for taking edoxaban at a certain time of the day in this study's app, and the participants were able to change the time of the alarm. When the alarm was heard at the preset time, the patient was prompted to take edoxaban, and the app also checked whether the patient took their medication. In addition to the alarm for taking medication, the app sent an alarm to measure BP and HR using an automated BP monitor (UA-651BLE, A&D Medical). The

XSL•FO
RenderX

measured BP and HR data were automatically transmitted to a smartphone via Bluetooth. To assess awareness of disease status, the app asked patients whether their measured BP results were optimal, and the patients were required to answer the question. The patients could review their historical vital sign data entered by date using the smartphone app. Physicians could obtain the patient's BP and HR readings by checking the patient's smartphone app at the follow-up visit or they could monitor the patient's vital signs in the standalone web dashboard system.

The control group received only standard, guideline-recommended care [1,2], including education about the disease and the importance of taking medication. The control group received the same automated BP monitor as the intervention group but did not have access to the smartphone app. Patients in both the intervention and control groups were followed up at 3 and 6 months after randomization. During this study, patients needed to continue taking edoxaban unless there were special circumstances that required an interruption or change. Discontinuation of edoxaban could be considered if creatinine clearance drops below 15 mL/min or in case of major bleeding or other serious events during this study.

Study Outcomes

The primary end point of this study was adherence to edoxaban at 3 and 6 months. Medication adherence was evaluated with the "pill count" measurements. The patients brought the remaining tablets to each scheduled visit, and trained research nurses counted the number of returned drugs and calculated medication adherence as follows:

Medication adherence at 3 and 6 months was defined as the value from baseline to 3 months and from baseline to 6 months, respectively. We also defined another primary end point, adequate medication adherence, as ≥95%, taking into account previous studies [25-27], and compared the proportion of adequate medication adherence (≥95%) between the intervention and control groups. Medication adherence was analyzed as a continuous variable, while adequate medication adherence was treated as a binary variable. Accordingly, we examined both the actual adherence values and the proportion of patients with adequate adherence.

The secondary end points were clinical composite end points, including stroke, systemic embolic events, major bleeding requiring hospitalization or transfusion, and death during the 6-month follow-up period. The exploratory clinical outcome was the frequency of app use during the 6 months, as assessed by the input of vital signs into the app.

Sample Size and Statistical Analyses

To the best of our knowledge, no large studies have evaluated the use of a smartphone app to improve NOAC adherence using pill count measurements. Thus, a precise sample size calculation was not possible. Considering previous studies, drug adherence was assumed to be 90% in the control group and 95% in the intervention group with a SD of 15%, considering previous studies [16,28]. With a 2-tailed α of .05, a power of 0.95, and a dropout rate of 6%, 500 patients (250 patients in each group) were required.

Categorical variables are reported as frequencies (percentages) and were compared using the Pearson chi-square test or Fisher exact test. Continuous variables were tested for normality using the Shapiro-Wilk normality test. Continuous variables with normal distribution are expressed as means (SDs) and were compared using the 2-tailed Student *t* test. Nonnormally distributed continuous variables are presented as medians with IQRs and were compared using the Mann-Whitney *U* test.

Analyses were performed according to the intention-to-treat analysis principle and included data from all patients who had undergone valid randomization. For patients who discontinued the trial early, we used the data available before discontinuation of the trial to calculate medication adherence at 3 or 6 months. If medication adherence could not be verified by pill count measurements or any other reliable method, the data were considered missing and excluded from the primary analysis. In the subgroup analyses, we evaluated the differential effects of the intervention on primary outcomes concerning sex and age. To assess the feasibility and effectiveness of the smartphone app function, an additional analysis was performed on the intervention group to determine whether adherence to edoxaban differed according to the frequency of app use.

All tests were 2-tailed, and a *P* value <.05 was considered statistically significant. Statistical analyses were performed using R (version 4.2.3, R Foundation).

Results

Patient Enrollment and Clinical Characteristics

From November 2020 to February 2023, a total of 501 patients with AF were screened for eligibility, of whom 498 were randomly assigned to the intervention (n=248) or control (n=250) group (Figure 1 and the CONSORT [Consolidated Standards of Reporting Trials] checklist in Multimedia Appendix 2). The demographic and clinical characteristics of the 2 groups were well balanced at baseline (Table 1). The mean age was 65.7 years, and 68.3% (340/498) were male. The median CHA_2DS_2-VASc score was not significantly different between the two groups (median 3, IQR 2-3 vs median 2, IQR 2-3; *P*=.66). The mean systolic BP, diastolic BP, and HR of the total population were 126.7 mm Hg, 78.5 mm Hg, and 75.6 beats per minute, respectively.

The diagram indicates the number of patients screened, enrolled, and randomized in the trial and this study's flow.

End-of-study visits occurred on September 12, 2023, which was the final date of the follow-up data. Among the 498 study participants, 401 (80.5%) completed the 6-month follow-up, and 97 patients discontinued follow-up in the trial before study completion (62 in the intervention group and 35 in the control group).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

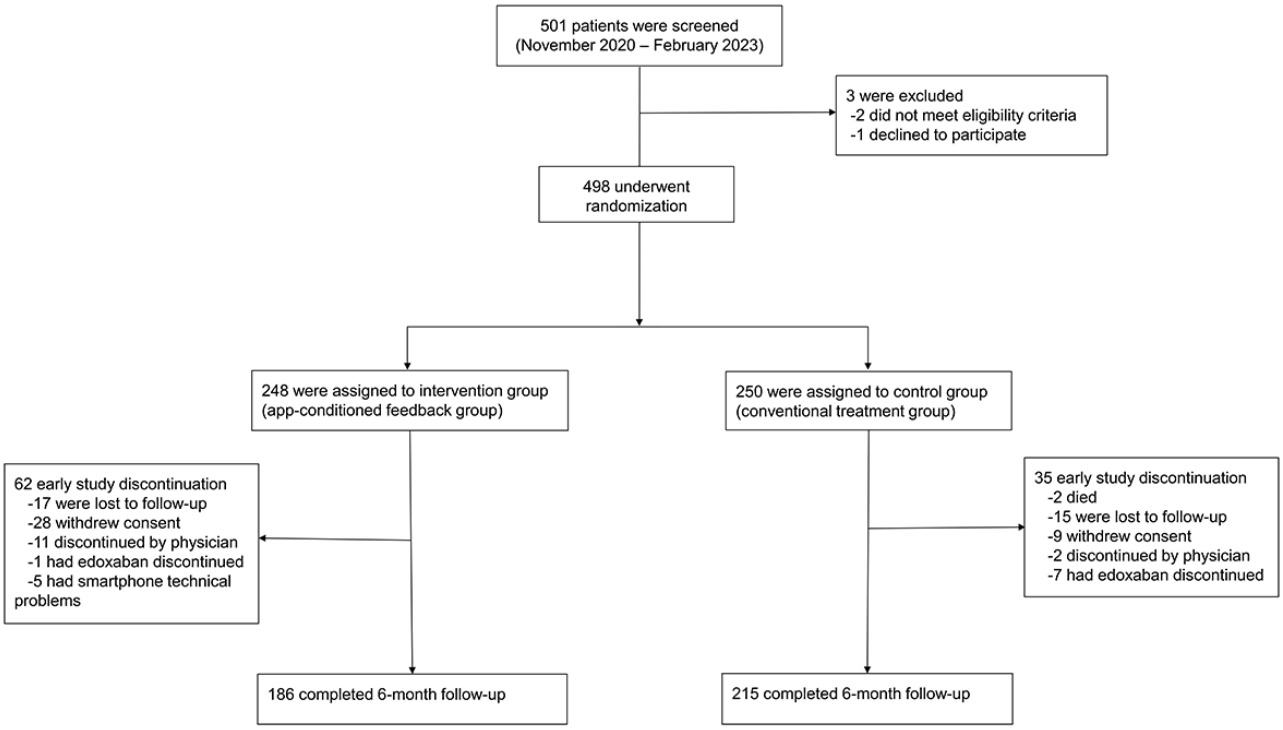

Table 1. Baseline characteristics according to treatment groups.

Variables	Intervention (n=248)	Control (n=250)	*P* value
Age (years), mean (SD)	65.7 (10.1)	65.6 (10.5)	.89
Male, n (%)	171 (69)	169 (67.6)	.82
Height (cm), mean (SD)	165.2 (9.3)	166.1 (8.8)	.28
Weight (kg), mean (SD)	69.3 (12.8)	69.8 (12.1)	.67
Alcohol, n (%)	84 (33.9)	81 (32.4)	.80
Smoking, n (%)	28 (11.3)	35 (14)	.44
Heart failure, n (%)	115 (46.4)	119 (47.6)	.85
Hypertension, n (%)	145 (58.5)	138 (55.2)	.52
Diabetes mellitus, n (%)	60 (24.2)	69 (27.7)	.44
Previous stroke, transient ischemic attack, or thromboembolism, n (%)	13 (5.2)	16 (6.4)	.72
Vascular disease, n (%)	48 (19.4)	42 (16.8)	.75
CHA_2DS_2-VASc score, median (IQR)	3 (2-3)	2 (2-3)	.66
Systolic blood pressure (mm Hg), mean (SD)	126.3 (18.2)	127.1 (17.4)	.62
Diastolic blood pressure (mm Hg), mean (SD)	78.2 (12.1)	78.8 (11.9)	.61
Heart rate (beat per minute), mean (SD)	75 (13.8)	76.1 (15.4)	.41

Medication Adherence According to Treatment Groups

Medication adherence according to treatment group and primary outcome is presented in Table 2. Medication adherence to edoxaban at 3 or 6 months was assessed in 405 patients (190 in the intervention group and 215 in the control group) and 422 patients (199 in the intervention group and 223 in the control group), respectively. Medication adherence at 3 or 6 months was not significantly different between the intervention and control groups (median 98%, IQR 95%-100% vs median 98%, IQR 91%-100% at 3 months, *P*=.06; median 98%, IQR 94.5%-100% vs median 97.5%, IQR 92.8%-100% at 3 months, *P*=.15; Figure 2). However, the proportion of adequate medication adherence (≥95%) was significantly higher in the intervention group than in the control group at both 3 and 6 months (146/190, 76.8% vs 139/215, 64.7% at 3 months, *P*=.01; 147/199, 73.9% vs 136/223, 61% at 3 months, *P*=.007).

XSL•FO
RenderX

Table 2. Medication adherence according to treatment groups.

Variables	Intervention (n=248)	Control (n=250)	Odds ratio (95% CI)[a]	*P* value[b]
Medication adherence at 3 months (%), median (IQR)[c]	98 (95-100)	98 (91-100)	__[d]	.06
The proportion of adequate medication adherence (≥95%) at 3 months, n/N (%)[e]	146/190 (76.8)	139/215 (64.7)	1.81 (1.17-2.81)	.01
Medication adherence at 6 months (%), median (IQR)	98 (94.5-100)	97.5 (92.8-100)	—	.15
The proportion of adequate medication adherence (≥95%) at 6 months, n/N (%)	147/199 (73.9)	136/223 (61)	1.8 (1.2-2.74)	.007

[a]Odds ratio for categorical variables.

[b]*P* value was calculated using the chi-square test for categorical variables and the Mann-Whitney *U* test for continuous variables.

[c]Medication adherence at 3 months was assessed in 405 patients (190 in the intervention group and 215 in the control group).

[d]Not available.

[e]Medication adherence at 6 months was assessed in 422 patients (199 in the intervention group and 223 in the control group).

Figure 2. Medication adherence according to treatment groups. (A) Medication adherence at 6 months and (B) proportion of patients achieving adequate medication adherence (≥95%) at 6 months.

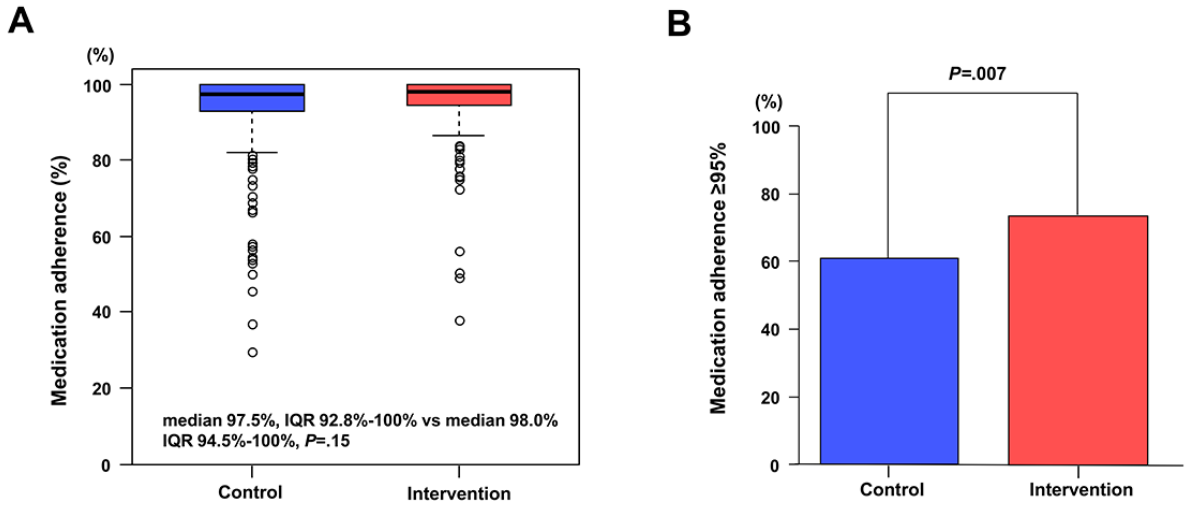

Subgroup Analysis of the Primary End Point

Table 3 shows subgroup analyses regarding the difference in medication adherence at 6 months according to the intervention and control groups. When stratified by sex, medication adherence at 6 months was not significantly different between the two groups for either male or female (median 98%, IQR 94.5%-100 vs median 97.5%, IQR 93%-100%, *P*=.42 for male; median 98.5%, IQR 94.5%-100% vs median 96.8%, IQR 92.5%-100%, *P*=.18 for female). However, among patients aged >65 years, the intervention group showed higher medication adherence at 6 months than the control group (median 98%, IQR 96%-100% vs median 97%, IQR 93%-100%; *P*=.03). Moreover, the proportion of adequate drug adherence (≥95%) was significantly higher in the intervention group than in the control group (82/101, 81.2% vs 66/112, 58.9%; *P*=.001). In patients aged ≤65 years, there was no significant difference between the 2 groups per medication adherence (*P*=.92) and the proportion of adequately adherent patients (*P*=.73).

XSL•FO
RenderX

Table 3. Subgroup analyses of the difference between the intervention and control groups regarding medication adherence at 6 months[a].

Variables	Intervention	Control	*P* value
Male (n=288)			
Medication adherence at 6 months (%), median (IQR)	98 (94.5-100)	97.5 (93-100)	.42
The proportion of adequate medication adherence (≥95%) at 6 months, n/N (%)	102/139 (73.4)	95/149 (63.8)	.10
Female (n=134)			
Medication adherence at 6 months (%), median (IQR)	98.5 (94.5-100)	96.8 (92.5-100)	.18
The proportion of adequate medication adherence (≥95%) at 6 months, n/N (%)	45/60 (75)	41/74 (55.4)	.03
Age >65 y (n=213)			
Medication adherence at 6 months (%), median (IQR)	98 (96-100)	97 (93-100)	.03
The proportion of adequate medication adherence (≥95%) at 6 months, n/N (%)	82/101 (81.2)	66/112 (58.9)	.001
Age ≤65 y (n=209)			
Medication adherence at 6 months (%), median (IQR)	97.8 (92.5-100)	97.5 (92-100)	.92
The proportion of adequate medication adherence (≥95%) at 6 months, n/N (%)	65/98 (66.3)	70/111 (63.1)	.73

[a]Medication adherence at 6 months was assessed in 422 patients (199 in the intervention group and 223 in the control group).

Secondary End Points and Application Frequency

During the 6-month period, one participant in the control group experienced an ischemic stroke, and one participant each in the intervention and control groups experienced major bleeding. Two patients in the control group died during follow-up. The composite clinical end points were not significantly different between the two groups. No adverse effects were associated with the use of the smartphone app during the trial.

Of the 186 patients in the intervention group who completed 6 months of follow-up, the median frequency of app use during this period, as assessed by the input of vital signs into the app, was 65.6% (IQR 16.7%-91.7%), and 107 (57.5%) patients had a frequency of app use of ≥50%. When patients were categorized into tertiles based on the frequency of app use, they were further divided into groups with values of <30% (n=60), between 30% and 80% (n=58), and >80% (n=68). There was no significant difference in medication adherence to edoxaban at 6 months between the 3 groups (median 98.5%, IQR 95.5%-100% vs median 97.5%, IQR 93%-100% vs median 98%, IQR 95.5%-100%; *P*=.11). The proportion of adequate medication adherence (≥95%) was not significantly different among the 3 groups (45/60, 75% vs 42/58, 72.4% vs 5568, 80.9%; *P*=.51), although the highest tertile of the frequency of app use showed a higher trend toward adequate medication adherence.

Discussion

Principal Findings

We developed a mobile app to provide NOAC alerts and measure vital signs. The main findings of this study are as follows: (1) contrary to our expectation, there were no significant differences in medication adherence to edoxaban between the app-based intervention and conventional treatment groups; (2) however, this smartphone app-based intervention resulted in a higher proportion of adequate medication adherence (≥95%) than the conventional treatment; and (3) moreover, in patients aged >65 years, the intervention group showed a higher medication adherence value and a higher proportion of adequate medication adherence at 6 months than the control group.

Comparison to Prior Work

Poor drug adherence to NOACs can increase the risk of ischemic stroke, leading to increased overall health care costs [19,29]. Therefore, consistent drug adherence to NOAC therapy is important for reducing the risk of ischemic stroke. Interventions such as mobile phone text messaging, telemonitoring, feedback, and education result in higher adherence to NOACs [5,6,30]. However, these interventions by health care providers require time and financial and human resources and are therefore challenging to implement in a real-world clinical setting. With the advent of mobile technology and artificial intelligence, mobile-based reminders or feedback may be a promising strategy to improve medication adherence in chronic diseases, including AF, using cost-effective and accessible methods [8]. We believe that our approach of using a smartphone app–based intervention would be an efficient way to improve NOAC adherence in AF with minimal cost and ease of use.

There have been some previous studies on mobile apps for improving drug adherence to NOACs [16-18]. However, the studies by Labovitz et al [16] and Turakhia et al [17] were limited by a small number of patients (28 and 139 patients, respectively). The study by Turakhia et al [17] did not show a difference in medication adherence following smartphone intervention; however, their smartphone app functions were different from those of ours, as our mobile app functions mainly focused on medication reminders and vital sign measurements. Recently, another study demonstrated that intake reminders

XSL•FO
RenderX

increased adherence to NOACs in patients with stroke attributable to AF [31]. We believe that these heterogeneous and diverse functions of mobile interventions and different frequencies of app use by participants can affect the benefits and outcomes of mobile apps, particularly for improving medication adherence.

Adherence to NOACs in clinical trials tends to be higher than that in the real world, which may be because patients feel monitored and measured [5,6,19-21]. In this trial, there was no statistically significant difference in the continuous value of adherence to NOACs between the intervention and control groups. This finding may be because both groups had higher levels of adherence than initially expected. However, the proportion of adequate medication adherence (≥95%), which may be the more important aspect than medication adherence value, was significantly higher in the intervention group than in the control group, demonstrating the benefit of the smartphone app–based intervention in improving medication adherence.

One of the interesting findings of this study was that smartphone-based intervention increased medication adherence to NOACs, particularly in older patients (>65 years). In general, older patients have lower medication adherence than younger patients for several reasons, particularly cognitive problems, which can cause them to forget to take their medication [32]. Specifically, in our study, the proportion of adequate medication adherence in the control group among patients aged >65 years was 58.9%, which is relatively low compared with those of the other subgroups. However, the benefit of smartphone app-based interventions was more pronounced in older patients than in younger patients, with 81.2% demonstrating adequate medication adherence. Therefore, the potential to improve medication adherence in older patients using smartphone apps is of great clinical importance.

The aim of this study differed from that of the ongoing RIVOX-AF (Reinforcement of Adherence via Self-Monitoring App Orchestrating Biosignals and Medication of RivaroXaban in Patients With Atrial Fibrillation and Co-Morbidities) study [33] at our institution. The RIVOX-AF trial aimed to evaluate whether a smartphone app-based intervention, including an alert for drug intake, visual confirmation of medication administration, and a list of medication intake histories, would increase drug adherence. In the ADHERE-App study, patients received a reminder to take their medication and measure their HR and BP according to a prespecified schedule. The RIVOX-AF and ADHERE-App studies have similar app functions for medication intake reminders; however, the ADHERE-App study was designed for active participation in vital sign measurement with the idea that using a smartphone app–based feedback system could help patients become more aware of their underlying conditions, which in turn could influence their self-care habits and improve medication adherence. The RIVOX-AF study mainly focused on confirming medication intake using a camera check.

Limitations and Strengths

This study has several limitations. First, actual engagement with the apps and actual participant responses to reminders or feedback from the app could not be analyzed. Additionally, the frequency of app use was evaluated based on the input of vital signs, which may not accurately reflect the actual use of the app. These limitations are common in studies using mobile apps. Second, in NOAC studies, it is important to determine whether more than one dose was skipped or an “unprotected day” [34], which was not assessed in this study. Third, an electronic Medication Event Monitoring System using a special cap that fits on a medication bottle to record the exact date and time of bottle opening is one of the best methods for measuring medication adherence and has been used in several other trials [5,6,25]; however, it was not used in this trial. However, pill count measurement by a clinical research nurse is also a relatively accurate assessment of medication adherence and may be more accurate than the drug score [35,36]. Fourth, data on satisfaction with the app or quality of life were not examined. Fifth, there was a significant difference between the intervention and control groups in the number of patients who dropped out or withdrew consent before the end of our study, suggesting that the use of smartphone apps may be associated with a high dropout rate. Sixth, additional methods to adjust for type I error inflation were not considered in the subgroup analyses. Therefore, results from subgroup analyses are exploratory and should be interpreted with caution, even for the subgroup (ie, age >65 years) that differs between the intervention and control groups. In addition, the proportion of adequate medication adherence was added as a coprimary end point after this study was first planned.

Despite these limitations, this study is a nationwide, multicenter randomized control trial with a relatively large number of participants compared with previous studies [16,17]. The major strength of this study is the use of a mobile app with various functions, and we believe that the features of our app can be effective solutions for improving medication adherence in patients with AF.

Conclusions

There is no difference in adherence to edoxaban between the app-conditioned feedback and conventional treatment groups. However, the proportion of patients with adequate adherence is higher in the intervention group than in the control group, and the benefit of the smartphone app–based intervention on medication adherence is more pronounced among older patients than among younger patients. Given the low adherence to NOACs, especially among older adults, a smartphone app has the potential to improve medication adherence.

Acknowledgments

This work was supported by a grant from Daiichi Sankyo Korea Co, Ltd (06-2019-233). This work was supported by the Institute for Information & communication Technology Planning & evaluation (IITP) grant funded by the Korean government, Ministry of Science and ICT (information and communications technology; MSIT; IITP-2022-0-00078, Explainable Logical Reasoning

XSL•FO
RenderX

for Medical Knowledge Generation). This research was supported by a grant from the Korea Health Technology R&D Project (Research and Development) through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant RS-2021-KH114109).

Data Availability

The data generated in this study are available from the corresponding author on reasonable request.

Authors' Contributions

MY, Ji Hyun Lee, ICK, and DJC contributed to the conception and design of this paper. MY, Ji Hyun Lee, ICK, Ju-Hee Lee, MNK, HLK, SL, IJK, SC, SJP, TH, MH, SL, and DJC contributed to data collection and analysis. MY drafted this original paper. All authors have revised this paper critically for important intellectual content. All authors approved the version submitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Operating system and sample display of the smartphone app.
[DOCX File , 347 KB - jmir_v26i1e65010_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1085 KB - jmir_v26i1e65010_app2.pdf]

References

1. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, et al. 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): the task force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. Eur Heart J 2021;42(5):373-498. [doi: 10.1093/eurheartj/ehaa612] [Medline: 32860505]
2. Joglar JA, Chung MK, Armbruster AL, Benjamin EJ, Chyou JY, Cronin EM, et al. 2023 ACC/AHA/ACCP/HRS guideline for the diagnosis and management of atrial fibrillation: a report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. Circulation 2024;149(1):e1-e156 [FREE Full text] [doi: 10.1161/CIR.0000000000001193] [Medline: 38033089]
3. Ruff CT, Giugliano RP, Braunwald E, Hoffman EB, Deenadayalu N, Ezekowitz MD, et al. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. Lancet 2014;383(9921):955-962. [doi: 10.1016/S0140-6736(13)62343-0] [Medline: 24315724]
4. How CH. Novel oral anticoagulants for atrial fibrillation. Singapore Med J 2015;56(12):657-659 [FREE Full text] [doi: 10.11622/smedj.2015184] [Medline: 26702159]
5. Desteghe L, Vijgen J, Koopman P, Dilling-Boer D, Schurmans J, Dendale P, et al. Telemonitoring-based feedback improves adherence to non-vitamin K antagonist oral anticoagulants intake in patients with atrial fibrillation. Eur Heart J 2018;39(16):1394-1403. [doi: 10.1093/eurheartj/ehx762] [Medline: 29300888]
6. Knaepen L, Delesie M, Vijgen J, Dendale P, Ector J, Desteghe L, et al. Adherence to oral anticoagulation measured by electronic monitoring in a Belgian atrial fibrillation population. Clin Res Cardiol 2023;112(12):1812-1823 [FREE Full text] [doi: 10.1007/s00392-023-02261-w] [Medline: 37498363]
7. Armitage LC, Kassavou A, Sutton S. Do mobile device apps designed to support medication adherence demonstrate efficacy? A systematic review of randomised controlled trials, with meta-analysis. BMJ Open 2020;10(1):e032045 [FREE Full text] [doi: 10.1136/bmjopen-2019-032045] [Medline: 32005778]
8. Pérez-Jover V, Sala-González M, Guilabert M, Mira JJ. Mobile apps for increasing treatment adherence: systematic review. J Med Internet Res 2019;21(6):e12505 [FREE Full text] [doi: 10.2196/12505] [Medline: 31215517]
9. Morawski K, Ghazinouri R, Krumme A, Lauffenburger JC, Lu Z, Durfee E, et al. Association of a smartphone application with medication adherence and blood pressure control: The MedISAFE-BP randomized clinical trial. JAMA Intern Med 2018;178(6):802-809 [FREE Full text] [doi: 10.1001/jamainternmed.2018.0447] [Medline: 29710289]
10. Huang Z, Tan E, Lum E, Sloot P, Boehm BO, Car J. A smartphone app to improve medication adherence in patients with type 2 diabetes in Asia: feasibility randomized controlled trial. JMIR mHealth uHealth 2019;7(9):e14914 [FREE Full text] [doi: 10.2196/14914] [Medline: 31516127]

XSL•FO
RenderX

11. Sarzynski E, Decker B, Thul A, Weismantel D, Melaragni R, Cholakis E, et al. Beta testing a novel smartphone application to improve medication adherence. Telemed J E Health 2017;23(4):339-348. [doi: 10.1089/tmj.2016.0100] [Medline: 27564971]
12. Wu YP, Linder LA, Kanokvimankul P, Fowler B, Parsons BG, Macpherson CF, et al. Use of a smartphone application for prompting oral medication adherence among adolescents and young adults with cancer. Oncol Nurs Forum 2018;45(1):69-76 [FREE Full text] [doi: 10.1188/18.ONF.69-76] [Medline: 29251285]
13. Yu C, Liu C, Du J, Liu H, Zhang H, Zhao Y, et al. Smartphone-based application to improve medication adherence in patients after surgical coronary revascularization. Am Heart J 2020;228:17-26. [doi: 10.1016/j.ahj.2020.06.019] [Medline: 32745732]
14. Johnston N, Bodegard J, Jerström S, Åkesson J, Brorsson H, Alfredsson J, et al. Effects of interactive patient smartphone support app on drug adherence and lifestyle changes in myocardial infarction patients: a randomized study. Am Heart J 2016;178:85-94 [FREE Full text] [doi: 10.1016/j.ahj.2016.05.005] [Medline: 27502855]
15. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012;14(6):e152. [doi: 10.2196/jmir.2104] [Medline: 23151820]
16. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. Stroke 2017;48(5):1416-1419 [FREE Full text] [doi: 10.1161/STROKEAHA.116.016281] [Medline: 28386037]
17. Turakhia M, Sundaram V, Smith SN, Ding V, Michael Ho P, Kowey PR, et al. Efficacy of a centralized, blended electronic, and human intervention to improve direct oral anticoagulant adherence: smartphones to improve rivaroxaban ADHEREnce in atrial fibrillation (SmartADHERE) a randomized clinical trial. Am Heart J 2021;237:68-78 [FREE Full text] [doi: 10.1016/j.ahj.2021.02.023] [Medline: 33676886]
18. Senoo K, Miki T, Ohkura T, Iwakoshi H, Nishimura T, Shiraishi H, et al. A smartphone app to improve oral anticoagulation adherence in patients with atrial fibrillation: prospective observational study. JMIR mHealth uHealth 2022;10(1):e30807 [FREE Full text] [doi: 10.2196/30807] [Medline: 34894626]
19. An J, Bider Z, Luong TQ, Cheetham TC, Lang DT, Fischer H, et al. Long-term medication adherence trajectories to direct oral anticoagulants and clinical outcomes in patients with atrial fibrillation. J Am Heart Assoc 2021;10(21):e021601 [FREE Full text] [doi: 10.1161/JAHA.121.021601] [Medline: 34713708]
20. Manzoor BS, Lee TA, Sharp LK, Walton SM, Galanter WL, Nutescu EA. Real-world adherence and persistence with direct oral anticoagulants in adults with atrial fibrillation. Pharmacotherapy 2017;37(10):1221-1230. [doi: 10.1002/phar.1989] [Medline: 28730619]
21. Kim D, Yang PS, Jang E, Yu HT, Kim TH, Uhm JS, et al. The optimal drug adherence to maximize the efficacy and safety of non-vitamin K antagonist oral anticoagulant in real-world atrial fibrillation patients. Europace 2020;22(4):547-557. [doi: 10.1093/europace/euz273] [Medline: 31598651]
22. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
23. MatrixMDR eCRF. URL: http://matrixmdr.com/eCRF/login.do [accessed 2024-10-23]
24. Kim IC, Lee JH, Choi D, Park S, Lee J, Park SM, et al. Rationale design and efficacy of a smartphone application for improving self-awareness of adherence to edoxaban treatment: study protocol for a randomised controlled trial (adhere app). BMJ Open 2022;12(4):e048777 [FREE Full text] [doi: 10.1136/bmjopen-2021-048777] [Medline: 35477876]
25. Osterberg L, Blaschke T. Adherence to medication. N Engl J Med 2005;353(5):487-497. [doi: 10.1056/NEJMra050100] [Medline: 16079372]
26. Burnier M, Wuerzner G, Struijker-Boudier H, Urquhart J. Measuring, analyzing, and managing drug adherence in resistant hypertension. Hypertension 2013;62(2):218-225. [doi: 10.1161/HYPERTENSIONAHA.113.00687] [Medline: 23753412]
27. Burnier M. Is there a threshold for medication adherence? Lessons learnt from electronic monitoring of drug adherence. Front Pharmacol 2019;9:1540 [FREE Full text] [doi: 10.3389/fphar.2018.01540] [Medline: 30687099]
28. Hwang J, Han S, Bae HJ, Jun SW, Choi SW, Lee CH, et al. NOAC adherence of patients with atrial fibrillation in the real world: dosing frequency matters? Thromb Haemost 2020;120(2):306-313. [doi: 10.1055/s-0039-1697954] [Medline: 31659737]
29. Deshpande CG, Kogut S, Willey C. Real-world health care costs based on medication adherence and risk of stroke and bleeding in patients treated with novel anticoagulant therapy. J Manag Care Spec Pharm 2018;24(5):430-439 [FREE Full text] [doi: 10.18553/jmcp.2018.24.5.430] [Medline: 29694285]
30. Thakkar J, Kurup R, Laba T, Santo K, Thiagalingam A, Rodgers A, et al. Mobile telephone text messaging for medication adherence in chronic disease: a meta-analysis. JAMA Intern Med 2016;176(3):340-349. [doi: 10.1001/jamainternmed.2015.7667] [Medline: 26831740]
31. Dietrich F, Polymeris AA, Albert V, Engelter ST, Hersberger KE, Schaedelin S, et al. Intake reminders are effective in enhancing adherence to direct oral anticoagulants in stroke patients: a randomised cross-over trial (MAAESTRO study). J Neurol 2024;271(2):841-851 [FREE Full text] [doi: 10.1007/s00415-023-12035-z] [Medline: 37831125]

XSL•FO
RenderX

32. Gellad WF, Grenard JL, Marcum ZA. A systematic review of barriers to medication adherence in the elderly: looking beyond cost and regimen complexity. Am J Geriatr Pharmacother 2011;9(1):11-23 [FREE Full text] [doi: 10.1016/j.amjopharm.2011.02.004] [Medline: 21459305]
33. Yoon M, Park JJ, Hur T, Hua C, Shim CY, Yoo B, et al. The reinforcement of adherence via self-monitoring app orchestrating biosignals and medication of RivaroXaban in patients with atrial fibrillation and co-morbidities: a study protocol for a randomized controlled trial (RIVOX-AF). Front Cardiovasc Med 2023;10:1130216 [FREE Full text] [doi: 10.3389/fcvm.2023.1130216] [Medline: 37324622]
34. Vrijens B, Heidbuchel H. Non-vitamin K antagonist oral anticoagulants: considerations on once- vs. twice-daily regimens and their potential impact on medication adherence. EP Europace 2015;17(4):514-523. [doi: 10.1093/europace/euu311] [Medline: 25694538]
35. Farmer KC. Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. Clin Ther 1999;21(6):1074-1090. [doi: 10.1016/S0149-2918(99)80026-5] [Medline: 10440628]
36. Erdine S, Arslan E. Monitoring treatment adherence in hypertension. Curr Hypertens Rep 2013;15(4):269-272. [doi: 10.1007/s11906-013-0369-9] [Medline: 23828148]

Abbreviations

ADHERE-App: Self-Awareness of Drug Adherence to Edoxaban Using an Automatic App Feedback System
AF: atrial fibrillation
BP: blood pressure
CONSORT: Consolidated Standards of Reporting Trials
HR: heart rate
NOAC: nonvitamin K antagonist oral anticoagulant
RIVOX-AF: Reinforcement of Adherence via Self-Monitoring App Orchestrating Biosignals and Medication of RivaroXaban in Patients With Atrial Fibrillation and Co-Morbidities

Edited by G Eysenbach, A Coristine; submitted 02.08.24; peer-reviewed by Y Wang, M Lv; comments to author 26.08.24; revised version received 20.09.24; accepted 09.10.24; published 21.11.24.

Please cite as:
Yoon M, Lee JH, Kim IC, Lee JH, Kim MN, Kim HL, Lee S, Kim IJ, Choi S, Park SJ, Hur T, Hussain M, Lee S, Choi DJ
Smartphone App for Improving Self-Awareness of Adherence to Edoxaban Treatment in Patients With Atrial Fibrillation (ADHERE-App Trial): Randomized Controlled Trial
J Med Internet Res 2024;26:e65010
URL: https://www.jmir.org/2024/1/e65010
doi:10.2196/65010
PMID:39570659

©Minjae Yoon, Ji Hyun Lee, In-Cheol Kim, Ju-Hee Lee, Mi-Na Kim, Hack-Lyoung Kim, Sunki Lee, In Jai Kim, Seonghoon Choi, Sung-Ji Park, Taeho Hur, Musarrat Hussain, Sungyoung Lee, Dong-Ju Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Long-Term Efficacy of an AI-Based Health Coaching Mobile App in Slowing the Progression of Nondialysis-Dependent Chronic Kidney Disease: Retrospective Cohort Study

Jianwei Ma[1,2*], MD; Jiangyuan Wang[3*], MD; Jiapei Ying[1*], MD; Shasha Xie[3*], MD; Qin Su[1], MD; Tianmeng Zhou[3], MS; Fuman Han[3], MD; Jiayan Xu[4], MD; Siyi Zhu[5], MD; Chenyi Yuan[1], MD; Ziyuan Huang[1], MD; Jingfang Xu[1], MD; Xuyong Chen[3], MD; Xueyan Bian[1], MD

[1]Department of Nephrology, The First Affiliated Hospital of Ningbo University, Ningbo, China
[2]Tongji University School of Medicine, Shanghai, China
[3]Beijing Kidney Health Technology Co., Ltd, Beijing, China
[4]Health Science Center, Ningbo University, ningbo, China
[5]School of Public Health, Health Science Center, Ningbo University, Ningbo, China
[*]these authors contributed equally

Corresponding Author:
Xueyan Bian, MD
Department of Nephrology
The First Affiliated Hospital of Ningbo University
The 5th floor of Building 6
Number 59 Liuting Street, Haishu District
Ningbo, 315000
China
Phone: 86 87085300
Email: fyybianxueyan@nbu.edu.cn

Abstract

Background: Chronic kidney disease (CKD) is a significant public health concern. Therefore, practical strategies for slowing CKD progression and improving patient outcomes are imperative. There is limited evidence to substantiate the efficacy of mobile app–based nursing systems for decelerating CKD progression.

Objective: This study aimed to evaluate the long-term efficacy of the KidneyOnline intelligent care system in slowing the progression of nondialysis-dependent CKD.

Methods: In this retrospective study, the KidneyOnline app was used for patients with CKD in China who were registered between January 2017 and April 2023. Patients were divided into 2 groups: an intervention group using the app's nurse-led, patient-oriented management system and a conventional care group that did not use the app. Patients' uploaded health data were processed via deep learning optical character recognition, and the artificial intelligence (AI) system provided personalized health care plans and interventions. Conversely, the conventional care group received suggestions from nephrologists during regular visits without AI. Monitoring extended for an average duration of 2.1 (SD 1.4) years. The study's objective is to assess the app's effectiveness in preserving kidney function. The primary outcome was the estimated glomerular filtration rate slope over the follow-up period, and secondary outcomes included changes in albumin-to-creatinine ratio (ACR) and mean arterial pressure.

Results: A total of 12,297 eligible patients were enrolled for the analysis. Among them, 808 patients were successfully matched using 1:1 propensity score matching, resulting in 404 (50%) patients in the KidneyOnline care system group and another 404 (50%) patients in the conventional care group. The estimated glomerular filtration rate slope in the KidneyOnline care group was significantly lower than that in the conventional care group (odds ratio –1.3, 95% CI –2.4 to –0.1 mL/min/1.73 m^2 per year vs odds ratio –2.8, 95% CI –3.8 to –1.9 mL/min/1.73 m^2 per year; P=.009). Subgroup analysis revealed that the effect of the KidneyOnline care group was more significant in male patients, patients older than 45 years, and patients with worse baseline kidney function, higher blood pressure, and heavier proteinuria. After 3 and 6 months, the mean arterial pressure in the KidneyOnline care group decreased to 85.6 (SD 9.2) and 83.6 (SD 10.5) mm Hg, respectively, compared to 94.9 (SD 10.6) and 95.2 (SD 11.6) mm Hg in the conventional care group (P<.001). The ACR in the KidneyOnline care group showed a more significant reduction

XSL•FO
RenderX

after 3 and 6 months (736 vs 980 mg/g and 572 vs 840 mg/g; *P*=.07 and *P*=.03); however, there was no significant difference in ACR between the two groups at the end of the follow-up period (618 vs 639 mg/g; *P*=.90).

Conclusions: The utilization of KidneyOnline, an AI-based, nurse-led, patient-centered care system, may be beneficial in slowing the progression of nondialysis-dependent CKD.

(J Med Internet Res 2024;26:e54206) doi:10.2196/54206

KEYWORDS

artificial intelligence; chronic kidney disease; eHealth care; mobile app; self-management; kidney function; telemedicine; app; health coaching; CKD; mobile phone

Introduction

Chronic kidney disease (CKD) is rapidly becoming a widespread noncommunicable chronic disease worldwide, presenting a significant public health challenge. It affects approximately 11%-13% of the global population. It is associated with a high mortality rate, substantial health care costs, particularly in its advanced stages, and a potential need for renal replacement therapy [1-4]. It is projected that by the year 2040, CKD will become the fifth leading cause of years of life lost globally [5]. CKD survivors often grapple with various systemic complications that significantly diminish their quality of life [4,6]. Therefore, mitigating the progression of CKD and managing its complications and comorbidities are critical to improving patient outcomes and reducing the associated risks of end-stage renal disease (ESRD) and mortality [6].

Despite being guided by evidence-based practices, conventional treatment methodologies that rely heavily on multidisciplinary health care teams and patient self-management exhibit limitations in efficiently managing and monitoring CKD [4,7-9]. The complexity of CKD management, given its numerous risk factors, complications, and comorbidities, poses a challenge to clinicians. To address this intricate issue, artificial intelligence (AI) has been introduced as a potential revolution [10]. Recent trends demonstrate that incorporating health information technology into disease management can enhance care management and patient self-care in chronic diseases, such as CKD. Mobile health (mHealth) apps, exemplifying this digital advancement, enable patients to conveniently access health services and information, promoting improved patient engagement and self-management [1,7,11-14]. Although existing studies have demonstrated the advantages of mobile app–based nursing systems in enhancing patient education, self-management, dietary guidance, doctor-patient communication, and disease monitoring, there is relatively limited evidence to substantiate the efficacy of slowing CKD progression of CKD [11,15-19].

In our previous related study, we found that “KidneyOnline,” a patient care system based on a mobile app, has the potential to significantly lower the mean arterial pressure (MAP) and reduce composite kidney outcomes in patients with CKD [7]. However, it is essential to mention that this study had a relatively short follow-up period, and the long-term effects of delayed renal function progression have not been researched. Therefore, the objective of this study was to evaluate the long-term efficacy of KidneyOnline in slowing kidney function decline in patients with CKD compared with conventional care. This study aims to provide robust evidence supporting the use of mHealth apps in CKD management and contribute to the broader acceptance and adoption of eHealth initiatives in chronic disease management.

Methods

Population

This retrospective cohort analysis was conducted using KidneyOnline, a mobile app in China. KidneyOnline provides an intelligent patient care system for patients with CKD. Information regarding recruitment for the KidneyOnline program was disseminated through WeChat, allowing many people to download the app and engage actively. The patients signed informed consent forms within the app to confirm their participation. All users from January 2017 to April 2023 were screened based on the recruitment criteria.

Patients were enrolled if they (1) were older than 18 years; (2) met the diagnostic criteria for CKD (ie, an estimated glomerular filtration rate [eGFR] <60 mL/min/1.73 m^2 or an eGFR less than 90 mL/min/1.73 m^2, concurrent with albuminuria or hematuria lasting at least 3 months, or as defined by other clinically significant indicators); (3) uploaded at least two distinct data entries upon registration, separated by a minimum of 6 months; (4) were not undergoing dialysis, with a baseline eGFR more significant than 15 mL/min/1.73 m^2; and (5) were capable and willing to provide informed consent. The exclusion criteria were as follows: (1) inability of the patient to operate a smartphone, (2) intention to initiate dialysis or undergo kidney transplantation within the subsequent 3 months, (3) absence of essential baseline or follow-up data, and (4) suspicion of acute kidney injury. To evaluate the efficacy of KidneyOnline in preserving kidney function, we identified a control group of patients with CKD. These patients underwent renal biopsies at the First Affiliated Hospital of Ningbo University between January 2018 and January 2021 but received conventional care exclusively.

Intervention

The KidneyOnline intelligent patient care system is a nurse-led, patient-focused, collaborative management system for patients with CKD, complementing standard clinic visits. This system leveraged AI and a health coach team composed of experienced nurses trained by nephrologists, dieticians, and social workers. The system incorporated a smartphone app designed for patients, a web-based clinical dashboard app for health care providers, and a cloud server for efficient data management. KidneyOnline

XSL•FO
RenderX

had access to at least 5 integral service elements provided by the system. These included the interpretation of disease conditions and lifestyle interventions, regular checkups, real-time question-and-answer sessions, early warnings, and clinical reminders, as detailed in our previous study [7].

Patients in the conventional care group only received routine face-to-face consultations with doctors in hospitals. They were unable to use the services offered by the KidneyOnline Care System.

Data Collection

The foundational elements of this system involve structuring patient health data using deep-learning optical character recognition. Generally, patient health data originates from various sources including (1) patient self-reported signs and symptoms, (2) intelligent home devices such as sphygmomanometers, and (3) patient medical history, clinical notes, drug prescriptions, laboratory results, pathology reports, and imaging examinations. Within the KidneyOnline intelligent system, patients upload the data by simply taking photos, after which the intelligent system efficiently extracts the data using deep-learning optical character recognition. Combined with manual verification, an organized database was established, enabling integrated and quantitative analysis of patient health data.

The data of patients in the conventional care group were entered by doctors via the KidneyOnline app, with informed consent from the patients. All electronic data and photographs were uploaded instantly to a secure cloud-based server. All data were encrypted and deidentified in adherence to security and privacy regulations. Only the research staff can access the data stored on the cloud platform.

eGFR results were calculated using the Chronic Kidney Disease Epidemiology Collaboration equation [20]. Blood pressure data were collected via the app, using home-based measurements uploaded by the patients using either a mercury or an electronic sphygmomanometer. Proteinuria measured by the total protein test or dipstick method was converted to albumin-to-creatinine ratio (ACR) using approximation formulas [21].

Outcomes

The primary outcome of this study was the eGFR slope during the follow-up period. Secondary endpoints included changes in the ACR and changes in MAP.

Ethical Considerations

This study was approved by the Medical and Research Ethics Committee of The First Affiliated Hospital of Ningbo University (2023110A) and was conducted in accordance with the Declaration of Helsinki. Implied consent was obtained from all participants when they registered on the KidneyOnline app, as the privacy policy included a clause allowing anonymized data to be used for research purposes. All data were anonymized and deidentified before extraction and securely stored in compliance with data protection regulations. No compensation was provided to participants, as this study involved a secondary analysis of existing data. No identifiable images of participants were included in the study or supplementary materials, eliminating the need for additional image consent.

Statistical Analysis

Data distribution properties are expressed as mean (SD) for continuous variables with a normal distribution or median (IQR) for variables with a skewed distribution. For continuous data, comparisons between groups were performed using the Student 2-tailed *t* test. The chi-square test was used for categorical variables and expressed as percentages.

To account for potential confounding factors between the KidneyOnline intelligent patient care system group and the conventional care group, propensity score matching (PSM) was conducted in a 1:1 ratio without replacement using the optimal method. The matching criteria were based on baseline characteristics including age, sex, BMI, baseline eGFR, baseline MAP, baseline ACR, nephropathy type, and renin-angiotensin-aldosterone system blocker (RASB) or immunosuppressive agent (ISA) treatment.

The eGFR slopes of both groups were estimated using a mixed linear regression model with a random intercept and random slope. This model was adjusted for baseline characteristics including age, sex, baseline MAP, log-transformed ACR, and RASB or ISA treatment. Subgroup analyses were performed after stratification based on sex, median age, median baseline eGFR, median baseline ACR, and median baseline MAP. All the missing data were treated as missing data without imputation. Statistical analyses were performed using the R software (version 4.1.2; R Core Team), and 2-sided $P<.05$ was considered statistically significant.

Results

Baseline Characteristics

Between January 2017 and April 2023, a total of 68,135 potential participants were screened for eligibility. Among them, 12,297 (18%) met the inclusion criteria and were enrolled in this study. Of those enrolled, 11,893 (96.7%) participants were included in the KidneyOnline care group, and 404 (3.3%) participants were included in the conventional care group (Figure 1). The mean age of the enrolled participants was 39.0 (SD 11.0) years, and 6280 (51.1%) participants were female. The mean BMI was 23.1 (SD 5.3) kg/m². In terms of CKD etiology, 5061 (41.2%) patients were diagnosed with biopsy-proven immunoglobulin A (IgA) nephropathy or IgA vasculitis. The baseline parameters revealed a mean eGFR of 84.8 (SD 30.2) mL/min/1.73 m², a median ACR of 410 (IQR 143-1143) mg/g, and a mean MAP of 88.9 (SD 10.7) mm Hg. During the follow-up period, 7152 (58.2%) patients were treated with a RASB, and 3219 (26.2%) patients were treated with ISA (Table 1).

Compared to the conventional care group, patients in the KidneyOnline care group were younger (mean 38.8, SD 10.8 years vs mean 46.0, SD 14.4 years; $P<.001$) and had a slightly lower BMI (mean 23.1, SD 5.4 kg/m^2 vs mean 23.8, SD 3.6 kg/m^2; $P<.001$), lighter ACR (median 406, IQR 143-1136 mg/g vs median 551, IQR 162-1562 mg/g; $P<.001$), lower MAP

XSL•FO
RenderX

(mean 88.5, SD 10.4 mm Hg vs mean 99.2, SD 13.8 mm Hg; *P*<.001), and lower use rate of ISA (25.9% vs 35.1%; *P*<.001). There were no significant differences in sex, baseline eGFR, or RASB therapy between the 2 groups.

Figure 1. Flow diagram of patient enrollment for the analysis. eGFR: estimated glomerular filtration rate.

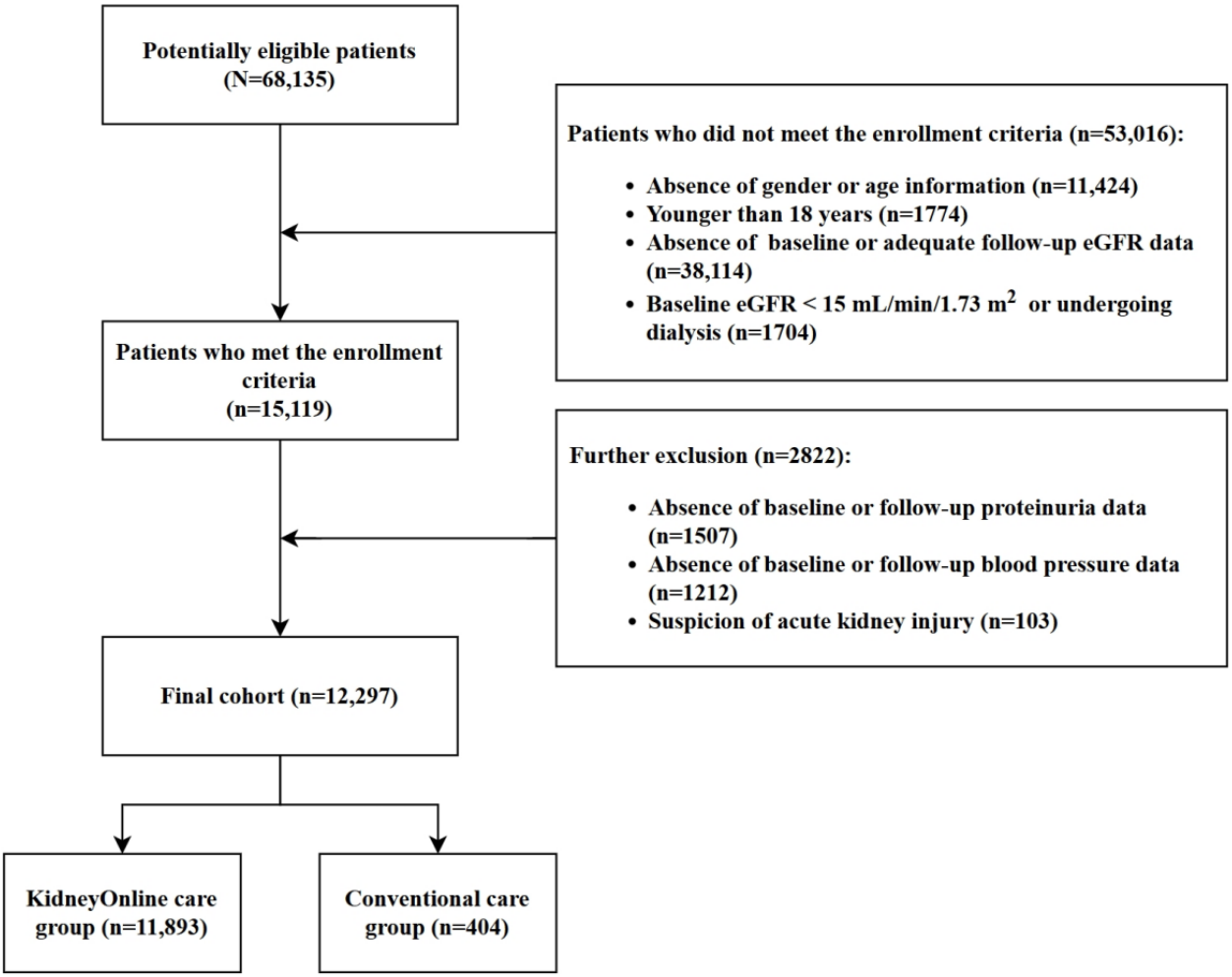

XSL•FO
RenderX

Table 1. Baseline characteristics of patients in the total cohort.

	Total (N=12,297)	Conventional care group (n=404)	KidneyOnline care group (n=11,893)	*P* value
Age (years)				<.001
Mean (SD)	39.0 (11.0)	46.0 (14.4)	38.8 (10.8)	
Median (IQR)	37.0 (31.0-45.0)	46.0 (34.0-57.0)	36.0 (31.0-45.0)	
Sex, n (%)				≥.99
Female	6280 (51)	206 (51)	6074 (51.1)	
Male	6017 (48.9)	198 (49)	5819 (48.9)	
BMI (kg/m^2)				<.001
Mean (SD)	23.1 (5.32)	23.8 (3.55)	23.1 (5.37)	
Median (IQR)	22.2 (20.1-24.7)	23.5 (21.5-25.8)	22.2 (20.0-24.7)	
Baseline eGFR[a] (mL/min/1.73 m^2)				.30
Mean (SD)	84.8 (30.2)	83.3 (29.8)	84.9 (30.2)	
Median (IQR)	88.5 (62.1-110.9)	89.5 (60.0-106.4)	88.5 (62.2-111.1)	
Baseline CKD[b] grade, n (%)				.60
G1	5970 (48.5)	201 (49.8)	5769 (48.5)	
G2	3487 (28.4)	102 (25.2)	3385 (28.5)	
G3a	1286 (10.5)	43 (10.6)	1243 (10.5)	
G3b	951 (7.7)	34 (8.4)	917 (7.7)	
G4	603 (4.9)	24 (5.9)	579 (4.9)	
baseline ACR[c] (mg/g)				<.001
Mean (SD)	929 (1540)	1270 (1780)	918 (1530)	
Median (IQR)	410 (143-1143)	551 (162-1562)	406 (143-1136)	
Baseline proteinuria category, n (%)				.12
Normal to mildly increased (ACR<30 mg/g)	1156 (9.4)	32 (7.9)	1124 (9.5)	
Moderately increased (ACR 30-300 mg/g)	3649 (29.7)	106 (26.2)	3543 (29.8)	
Severely increased (ACR>300 mg/g)	7492 (60.9)	266 (65.8)	7226 (60.8)	
Baseline MAP[d] (mm Hg)				<.001
Mean (SD)	88.9 (10.7)	99.2 (13.8)	88.5 (10.4)	
Median (IQR)	89.0 (81.7-95.0)	97.7 (90.0-107.1)	88.7 (81.7-94.7)	
Type of nephropathy, n (%)				<.001
IgA[e] nephropathy or IgA vasculitis	5061 (41.2)	158 (39.1)	4903 (41.2)	
Membranous nephropathy	1010 (8.2)	77 (19.1)	933 (7.8)	
Focal segmental glomerular sclerosis	342 (2.8)	16 (4)	326 (2.7)	
Diabetic nephropathy	108 (0.9)	35 (8.7)	73 (0.6)	
Hypertensive nephropathy	269 (2.2)	14 (3.5)	255 (2.1)	
Kidney transplant recipient	54 (0.4)	0 (0)	54 (0.5)	
Other types of nephritis	5453 (44.3)	104 (25.7)	5349 (45)	
Drug treatment, n (%)				

XSL•FO
RenderX

	Total (N=12,297)	Conventional care group (n=404)	KidneyOnline care group (n=11,893)	*P* value
RASB[f]	7152 (58.2)	230 (56.9)	6922 (58.2)	.65
ISA[g]	3219 (26.2)	142 (35.1)	3077 (25.9)	<.001

[a]eGFR: estimated glomerular filtration rate.

[b]CKD: chronic kidney disease.

[c]ACR: albumin-to-creatinine ratio.

[d]MAP: mean arterial pressure.

[e]IgA: immunoglobulin A.

[f]RASB: renin-angiotensin-aldosterone system blocker.

[g]ISA: immunosuppressive agents.

Estimation of eGFR Slopes

The average duration of follow-up was 2.1 (SD 1.4) years, and the average number of eGFR follow-ups was 5.4 (SD 3.8). After adjustment for age, sex, baseline MAP, ACR, and RASB or ISA treatment, the mixed linear model showed that the eGFR slope of patients in the KidneyOnline care group was significantly lower than that in the conventional care group (odds ratio [OR] –1.0, 95% CI –2.1 to 0.09 mL/min/1.73 m^2 per year vs OR –2.7, 95% CI –3.8 to –1.6 mL/min/1.73 m^2 per year; *P*=.002).

Baseline Characteristics After PSM

A total of 808 patients (404 per group) were successfully matched using 1:1 PSM. In total, the average age was 46.1 (SD 14.2) years, and 413 (51.1%) were female. The average BMI was 23.8 (SD 4.2) kg/m^2. The baseline eGFR, MAP, and ACR were 82.7 (SD 30.4) mL/min/1.73 m^2, 99.3 (SD 14.2) mm Hg, and 582 (IQR 158-1628) mg/g, respectively (Table 2). IgA nephropathy or IgA vasculitis was the most common cause of CKD, accounting for 38.5% (n=311) of cases. A total of 440 (54.5%) patients received RASB and 288 (35.6%) patients received ISA. After 1:1 matching, there were no significant differences in laboratory results or drug treatments between the two groups at baseline (Table 2).

XSL•FO
RenderX

Table 2. Baseline characteristics of patients after propensity score matching.

	Total (N=808)	Conventional care group (n=404)	KidneyOnline care group (n=404)	*P* value
Age (years)				.81
Mean (SD)	46.1 (14.2)	46.0 (14.4)	46.2 (14.1)	
Median (IQR)	45.0 (35.0-56.0)	46.0 (34.0-57.0)	45.0 (35.0-55.0)	
Sex, n (%)				≥.99
Female	413 (51.1)	206 (51)	207 (51.2)	
Male	395 (48.9)	198 (49)	197 (48.8)	
BMI (kg/m^2)				.72
Mean (SD)	23.8 (4.24)	23.8 (3.55)	23.7 (4.84)	
Median (IQR)	23.2 (21.2-25.7)	23.5 (21.5-25.8)	23.0 (20.8-25.6)	
Baseline eGFR[a] (mL/min/1.73 m^2)				.56
Mean (SD)	82.7 (30.4)	83.3 (29.8)	82.1 (31.0)	
Median (IQR)	88.1 (59.0-107.3)	89.5 (60.0-106.4)	87.2 (58.1-107.9)	
Baseline CKD[b] grade, n (%)				.91
G1	390 (48.3)	201 (49.8)	189 (46.8)	
G2	212 (26.2)	102 (25.2)	110 (27.2)	
G3a	85 (10.5)	43 (10.6)	42 (10.4)	
G3b	69 (8.5)	34 (8.4)	35 (8.7)	
G4	52 (6.4)	24 (5.9)	28 (6.9)	
baseline ACR[c] (mg/g)				.59
Mean (SD)	1240 (1690)	1270 (1780)	1200 (1600)	
Median (IQR)	582 (158-1628)	551 (162-1562)	616 (157-1637)	
Baseline proteinuria category, n (%)				.54
Normal to mildly increased (ACR<30 mg/g)	73 (9)	32 (7.9)	41 (10.1)	
Moderately increased (ACR 30-300 mg/g)	209 (25.9)	106 (26.2)	103 (25.5)	
Severely increased (ACR>300 mg/g)	526 (65.1)	266 (65.8)	260 (64.4)	
Baseline MAP[d] (mm Hg)				.87
Mean (SD)	99.3 (14.2)	99.2 (13.8)	99.4 (14.5)	
Median (IQR)	97.3 (90.0-106.7)	97.7 (90.0-107.1)	96.7 (90.3-106.6)	
Type of nephropathy, n (%)				.83
IgA[e] nephropathy or IgA vasculitis	311 (38.5)	158 (39.1)	153 (37.9)	
Membranous nephropathy	149 (18.4)	77 (19.1)	72 (17.8)	
Focal segmental glomerular sclerosis	30 (3.7)	16 (4)	14 (3.5)	
Diabetic nephropathy	66 (8.2)	35 (8.7)	31 (7.7)	
Hypertensive nephropathy	34 (4.2)	14 (3.5)	20 (5)	
Other types of nephritis	218 (27)	104 (25.7)	114 (28.2)	
Drug treatment, n (%)				

XSL•FO
RenderX

	Total (N=808)	Conventional care group (n=404)	KidneyOnline care group (n=404)	*P* value
RASB[f]	440 (54.5)	230 (56.9)	210 (52)	.18
ISA[g]	288 (35.6)	142 (35.1)	146 (36.1)	.83

[a]eGFR: estimated glomerular filtration rate.
[b]CKD: chronic kidney disease.
[c]ACR: albumin-to-creatinine ratio.
[d]MAP: mean arterial pressure.
[e]IgA: immunoglobulin A.
[f]RASB: renin-angiotensin-aldosterone system blocker.
[g]ISA: immunosuppressive agents.

Estimation of eGFR Slopes After PSM

After 1:1 matching, the average follow-up was 1.7 (SD 1.2) years, and the average number of eGFR follow-ups was 4.6 (SD 2.9). In consistence with the results before matching, the eGFR slope of patients in the KidneyOnline care group was significantly lower than that in the conventional care group (OR –1.3, 95% CI –2.4 to –0.1 mL/min/1.73 m^2 per year vs OR –2.8, 95% CI –3.8 to –1.9 mL/min/1.73 m^2 per year; *P*=.009). This suggests that utilization of the KidneyOnline intelligent care system was associated with a delay in the progression to ESRD.

Subgroup Analysis

To investigate the kidney-protective effect within the KidneyOnline care group across various characteristic populations, we conducted further subgroup analysis. Subgroups were categorized based on sex, median age (<45 vs ≥45 years), median baseline ACR (<719 vs ≥719 mg/g), median baseline MAP (<97 vs ≥ 97 mm Hg), and median baseline eGFR (<86.6 mL/min/1.73 m^2 vs ≥ 86.6 mL/min/1.73 m^2). The results indicated that, in all subgroups, the eGFR slope of patients in the KidneyOnline care group was lower than that of patients in the conventional care group. Statistical significance was observed in the subgroups of male patients, patients older than 45 years, patients with worse baseline kidney function (eGFR<86.6 mL/min/1.73 Â m^2), higher blood pressure (MAP≥97 mm Hg), and heavier proteinuria (ACR ≥719 mg/g; Table 3).

Table 3. Subgroup analysis of eGFR[a] slope.

Characteristic and subgroup	eGFR slope (mL/min/1.73 m^2 per year)		*P* value
	Conventional care group	KindeyOnline care group	
Age (years)			
<45	–1.4	–0.6	.42
≥45	–4.2	–1.7	.002
Sex			
Female	–2.7	–1.6	.18
Male	–3.0	–1.0	.02
Baseline eGFR (mL/min/1.73 m^2)			
<86.6	–3.0	–0.2	.02
≥ 86.6	–2.5	–1.5	.18
Baseline MAP[b] (mm Hg)			
<97	–2.3	–1.5	.27
≥97	–3.3	–1.1	.02
Baseline ACR[c] (mg/g)			
<719	–1.5	–1.0	.49
≥719	–4.2	–1.5	.007

[a]eGFR: estimated glomerular filtration rate.
[b]MAP: mean arterial pressure.
[c]ACR: albumin-to-creatinine ratio.

XSL•FO
RenderX

The Handling of Missing Data

All the missing data were treated as missing data without imputation. As illustrated in Figure 1, among 68,135 patients, 11,424 (16.8%) individuals had missing data on sex and age, while 38,114 (55.9%) individuals lacked baseline or follow-up serum creatinine data for a duration exceeding six months. Additionally, 2719 (4%) patients had missing data on proteinuria and blood pressure follow-up. These patients were excluded from the cohort based on the exclusion criteria. Among the 808 patients who met the inclusion criteria and were successfully matched through PSM, 175 (21.7%) patients had 2 follow-up serum creatinine measurements, while 633 (78.3%) patients had three or more follow-up serum creatinine measurements.

Regarding the follow-up periods, the number of patients with proteinuria data at 3, 6, 12, 18, and 24 months was 473 (58.5%), 419 (51.9%), 374 (46.3%), 332 (41.1%), and 234 (29%), respectively. The number of patients with blood pressure data at 3, 6, 12, 18, and 24 months were 278 (34.4%), 224 (27.7%), 188 (23.2%), 162 (20%), and 122 (15.1%), respectively.

Leffondre et al [22] compared three methods for estimating the slope of eGFR decline over time (linear regression on individual slopes, linear mixed models, and generalized estimating equations). The results indicated that linear mixed models (which are also used in this study) are robust to missing values and irregularly spaced data, where the measurement time points are not fixed. Even with one-third of the data missing, linear mixed models can still accurately estimate the slope of eGFR decline.

Changes in Blood Pressure, Proteinuria, and eGFR During Follow-Up

The presence of hypertension increases the risk of a decline in kidney function. Although subgroup analysis revealed that the kidney-protective effect of KidneyOnline care was more pronounced in patients with higher baseline blood pressure, we further analyzed the trend of blood pressure changes in both groups during follow-up. The MAP was similar in both groups (99.4, SD 14.5 mm Hg vs 99.2, SD 13.8 mm Hg; *P*=.87) at baseline. However, after 3 months, the MAP of the patients in the KidneyOnline care group decreased to 85.6 (SD 9.2) mm Hg, while that of the conventional care group was 94.9 (SD 10.6) mm Hg (*P*<.001). After 6 months, the MAP of the KidneyOnline care group further decreased to 83.6 (SD 10.5) mm Hg, while that of the conventional care group was 95.2 (SD 11.6) mm Hg (*P*<.001). During the 24-month follow-up period, the blood pressure level of patients in the KidneyOnline care group was significantly lower than that in the conventional care group and remained stable throughout the follow-up period (Figure 2A).

Albuminuria was indicated by a steep eGFR slope. At baseline, there was no significant difference in the average ACR between the KidneyOnline and conventional care groups (1270 vs 1200 mg/g; *P*=.59). During the follow-up period, the average ACRs in the KidneyOnline care and conventional care groups decreased to 736 and 980 mg/g, respectively, after three months (*P*=.07), and further decreased to 572 and 840 mg/g after 6 months (*P*=.03). However, at the end of 24 months, there was no significant difference in ACR between the 2 groups (618 vs 639 mg/g; *P*=.90; Figure 2B).

The changes in eGFR during the follow-up period are illustrated in Figure 2C. The decline in eGFR in the KidneyOnline care group is more gradual compared to the control group, which is consistent with the eGFR slopes observed in both groups.

Figure 2. Changes in (A) MAP, (B) ACR, and (C) eGFR during the follow-up period. ACR: albumin-to-creatinine ratio; eGFR: estimated glomerular filtration rate; MAP: mean arterial pressure.

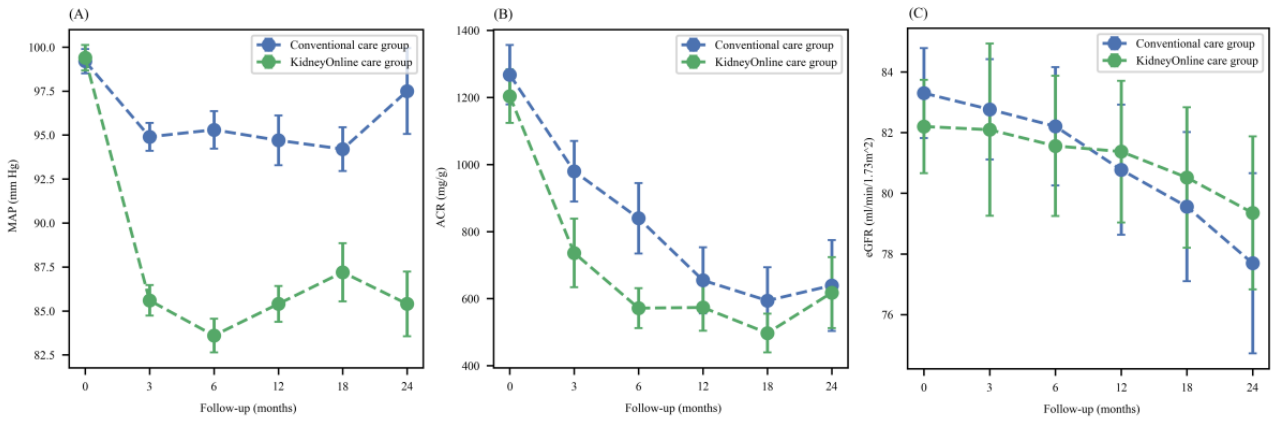

Discussion

Main Findings

This study describes a nurse-administered smartphone-based patient-centered system designed for disease management. Our results demonstrated that over an average period of up to 2.1 (SD 1.4) years, patients who received the integration of the KidneyOnline intelligent care system, an AI-based mobile app, exhibited a significantly slower decrease of eGFR, lower MAP throughout the follow-up period, and decreased ACR in the early period compared to those received conventional care. This provides a new strategy for long-term improvement of self-management and delayed renal function progression in patients with nondialysis-dependent CKD.

XSL•FO
RenderX

Analysis of Findings

The findings from this study indicated that using the KidneyOnline intelligent care system may be beneficial for delaying the progression to ESRD in patients with CKD. The eGFR slope is a valuable tool for monitoring kidney function over time, and our data reaffirm its significance. It revealed a substantial contrast between the KidneyOnline and conventional care groups, despite the balanced patient characteristics achieved through PSM. These findings are consistent. Previous research results have evaluated the value of similar telemedicine and intelligent care systems in CKD management [7,8,23]. Recently, in their 3-month prospective study, Li et al [23] found that patients with CKD at stages 1-4 who received dietary and exercise advice, along with motivation through a social media group via a wearable device and smartphone app, exhibited a slower decline in eGFR compared to those receiving conventional care alone (–0.56 vs –4.58 mL/min/1.73 m^2). Although the results were similar, this study featured a longer follow-up period (averaging 2.1 years). The effect of delaying the decline of eGFR may be attributed to the KidneyOnline care system's enhancement of the patient's capacity for self-management, promotion of self-monitoring, encouragement of adherence to lifestyle interventions, and improvement of long-term medication compliance. Remarkably, our subgroup analysis revealed critical information regarding potential beneficiaries of the KidneyOnline care system. We discovered that this system was particularly advantageous for male patients, individuals older than 45 years, and those with worse kidney function, higher blood pressure, and more significant proteinuria at baseline. These findings imply that a personalized approach to CKD care, taking into account patients' demographics, baseline kidney function, and clinical characteristics, is both feasible and beneficial.

Moreover, it is worth highlighting that those in the KidneyOnline care group displayed more promising trends in critical CKD progression markers. In the initial three months, patients using the KidneyOnline care system showed a significant reduction in blood pressure compared to the conventional care group. This improvement consistently persisted throughout the entire study duration. This reduction in blood pressure could have contributed to the slower decline in kidney function observed in the KidneyOnline care group [24-26]. In parallel, the KidneyOnline care group showed a more pronounced reduction in proteinuria, a potent predictor of CKD progression, highlighting the effectiveness of this intelligent care system [27-30]. Interestingly, by the end of the 24-month follow-up period, there was no significant difference in the ACR between the two groups, suggesting that the ability to sustain long-term effects on ACR through lifestyle and medication adjustments may be limited as CKD progresses, leading to a stabilization of benefits from such interventions. This finding indicates that more intensive or sustained intervention strategies may be required to influence long-term ACR outcomes. Furthermore, the convergence of ACR values also highlights the complexity of managing chronic diseases, underscoring the necessity for continuous innovation and timely adjustment of digital health strategies to meet the evolving needs of CKD patients. Nonetheless, it remains evident that the initial decrease in ACR may have potentially contributed to the kidney-protective effect of the KidneyOnline care system.

Advantages of KidneyOnline in CKD Management

Even though hundreds of apps for CKD management have been developed, research shows that only a tiny subset of CKD apps received high ratings from patients and nephrologists. This arises because most apps cannot simultaneously fulfill the demands of patient engagement safety, fostering excellent nursing team-patient interaction, providing treatment advice based on evidence-based medicine, and ensuring continuity of care [15,31]. The KidneyOnline intelligent care system is well-equipped to address these demands, offering disease condition interpretation and lifestyle interventions, regular checkups, real-time question-and-answer sessions, early warnings, and clinical reminders.

Many individuals with CKD frequently need to engage in complex disease management strategies. However, they often exhibit limited participation in self-management behaviors, potentially leading to unfavorable outcomes [32]. Therefore, the capacity of an app to enhance patient engagement in self-management is crucial. Singh et al [15] argue that patient engagement in an app should be evaluated from four dimensions: the provision of educational information, reminders or alerts to patients, tracking and summarization of health information, and guidance based on data input by patients. In order to fulfill these requirements, the KidneyOnline system provides lifestyle guidance, including dietary, exercise, and sleep recommendations, using AI to customize personalized meal plans in accordance with the patient's health status and food preferences. Moreover, it consistently monitors the patient's health status, blood pressure, dietary choices, and medication adherence through real-time data collection and an integrated intelligent character recognition system. Additionally, the system promptly identifies and alerts to potential risks via its built-in early warning algorithm. These features collectively enhance the patient's capacity for self-management and safety, thereby enhancing the patient's prognosis.

Some studies have shown that interactive treatment plans can enhance compliance, safety, and quality of life in patients [33,34]. It is a commonly held belief among both patients with CKD and CKD care providers that mobile apps hold the potential to enhance the self-management of CKD significantly. This enhancement is accomplished by facilitating communication between patients and providers and empowering patients to engage in self-care activities, including adhering to treatment plans. Mobile apps can serve as remote digital platforms for patients and care providers [17]. However, most existing apps designed for CKD management primarily focus on one-sided patient education, failing to incorporate clinical care and health behavior promotion through motivational feedback or interaction with health care providers [13,17,31]. The KidneyOnline care system is equipped to fulfill the functionality mentioned above, enabling health coaches to furnish real-time answers to patient queries promptly. We have also established a renal knowledge graph based on kidney disease: Improving Global Outcomes guidelines for CKD. This feature provides advice based on evidence-based medicine,

XSL•FO
RenderX

further augmenting patient care and treatment strategies. Moreover, the KidneyOnline system provides patients with a crucial element of long-term stable management: continuity of care. This is essential for patients' self-management and their collaboration with health care professionals to decelerate the progression of CKD. However, recent research shows that the continuity of patient-centered care for CKD, as provided by most mHealth apps, falls short of expectations. This deficiency adversely affects patients' ability to effectively self-manage their condition and hampers the support that health care professionals can offer [31]. A 5-year observational data from the real world have substantiated this notion.

During the COVID-19 pandemic, mHealth apps, such as KidneyOnline, demonstrated significant advantages by providing a secure and accessible platform for chronic disease management. The pandemic introduced unprecedented challenges to health care delivery, including movement restrictions and the risks associated with in-person visits [35,36]. These factors likely shifted patient preferences toward mHealth apps such as KidneyOnline. We observed a substantial increase in the utilization of KidneyOnline during periods of strict lockdown and social distancing. This trend illustrates the pandemic's role in accelerating the adoption of digital health tools and emphasizes their critical importance in sustaining chronic disease management when access to traditional health care services is constrained. These findings suggest that mHealth tools could be instrumental in maintaining continuity of care and effectively managing chronic conditions during future public health crises. The pandemic has underscored the need to integrate digital health strategies into standard health care practices, ensuring that these tools are readily available and optimized for patient care in routine and crises.

Strengths and Limitations

The major strength of this study is that KidneyOnline provides a comprehensive approach to CKD management by integrating real-time patient monitoring, AI-driven personalized interventions, and continuous clinical care. This distinguishes it from other CKD management apps, which often focus on singular aspects, like education or symptom tracking, without incorporating interactive features or personalized care plans. This study also offers long-term real-world evidence of KidneyOnline's efficacy in slowing CKD progression, whereas many studies lack extended follow-up. Additionally, the system's AI-driven scalability makes it a potential solution for broader health care apps.

However, this study also has some limitations. First, the retrospective nature of the cohort study restricted our ability to establish conclusive causality, which should be further validated through a randomized controlled trial. Second, potential bias may have been introduced by selecting the control group from a different setting. The control group, which was managed solely through conventional care at the First Affiliated Hospital of Ningbo University without any digital health intervention, was compared with patients using the KidneyOnline app. This setting difference may impact comparability, particularly regarding patient engagement, service access, and follow-up care. Although PSM reduced some disparities, it could not entirely eliminate biases. Future studies should consider using more homogeneous control groups or other methods to address these setting-related biases better. Third, we did not analyze other potential endpoints related to the utilization of the KidneyOnline care system, such as the incidence rates of cardiovascular comorbidities, complications, health economic benefits, and hospitalization frequencies. These additional endpoints would have provided a more comprehensive evaluation of our management model for patients with CKD. Fourth, the effectiveness of KidneyOnline may also be influenced by factors such as patients' levels of technological literacy, motivation to engage in self-management, and adherence to app use, all of which were not measured in this study. It is also important to note that this study population is predominantly comprised of young and middle-aged individuals from China, who generally demonstrate a high level of digital literacy. This demographic characteristic, combined with China's unique health care and technological environment, might influence the observed effectiveness of the AI-based app KidneyOnline. Moreover, the integration of AI into health care varies widely across different regions, indicating that the outcomes observed in this study might not be universally applicable. Therefore, to fully ascertain the efficacy and applicability of KidneyOnline, future research should include participants from a broader age spectrum, with varied levels of technological literacy, and from different health care systems to enhance the generalizability of the results.

Conclusions

In conclusion, our research further substantiates the therapeutic efficacy of the mobile app KidneyOnline in delaying the progression of CKD in patients without nondialysis. KidneyOnline is a smartphone-based, nurse-led, patient-oriented management system. Our findings emphasize the potential of AI and machine learning in health care interventions, particularly their capabilities for optimizing management strategies, tailoring personalized treatment plans, and improving long-term medication compliance of patients. Given the increasing global incidence of CKD, there is an urgent need for innovative solutions such as these [37-39]. We encourage further studies to validate these results and to explore the feasibility of implementing digital health interventions on a larger scale.

Acknowledgments

We are grateful to all the patients and controls for participating in this study. The authors were supported by the Nature Science Foundation of Ningbo (2023J160), the Zhejiang Province Medicine and Health Technology Plan Project (2022KY299, 2022KY309), and the Nursing Home Level Project of the First Affiliated Hospital of Ningbo University (H2020YJ011).

XSL•FO
RenderX

Data Availability

The datasets generated during or analyzed during this study are not publicly available due to business confidentiality reasons but are available from the corresponding author upon reasonable request.

Authors' Contributions

JM, JW, and SX wrote the manuscript and provided data for Tables 1 and 2; JY conducted the patient interviews; QS, JiayanXu, SZ, CY, ZH, and JingfangXu conducted the data curation; and TZ and FH conducted all statistical analyses. XB and XC conducted the research idea and study design. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

References

1. Krisanapan P, Tangpanithandee S, Thongprayoon C, Pattharanitima P, Cheungpasitporn W. Revolutionizing chronic kidney disease management with machine learning and artificial intelligence. J Clin Med 2023 Apr 21;12(8):3018 [FREE Full text] [doi: 10.3390/jcm12083018] [Medline: 37109354]
2. Jager KJ, Kovesdy C, Langham R, Rosenberg M, Jha V, Zoccali C. A single number for advocacy and communication-worldwide more than 850 million individuals have kidney diseases. Nephrol Dial Transplant 2019;34(11):1803-1805. [doi: 10.1093/ndt/gfz174] [Medline: 31566230]
3. Hu EA, Coresh J, Anderson CAM, Appel LJ, Grams ME, Crews DC, et al. Adherence to healthy dietary patterns and risk of CKD progression and all-cause mortality: findings from the CRIC (Chronic Renal Insufficiency Cohort) study. Am J Kidney Dis 2021;77(2):235-244 [FREE Full text] [doi: 10.1053/j.ajkd.2020.04.019] [Medline: 32768632]
4. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011) 2022;12(1):7-11 [FREE Full text] [doi: 10.1016/j.kisu.2021.11.003] [Medline: 35529086]
5. Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. Lancet 2018;392(10159):2052-2090 [FREE Full text] [doi: 10.1016/S0140-6736(18)31694-5] [Medline: 30340847]
6. Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: a review. JAMA 2019;322(13):1294-1304 [FREE Full text] [doi: 10.1001/jama.2019.14745] [Medline: 31573641]
7. Liu W, Yu X, Wang J, Zhou T, Yu T, Chen X, et al. Improving kidney outcomes in patients with nondiabetic chronic kidney disease through an artificial intelligence-based health coaching mobile app: retrospective cohort study. JMIR Mhealth Uhealth 2023;11:e45531 [FREE Full text] [doi: 10.2196/45531] [Medline: 37261895]
8. Yang F, Hou Y, Chang R. The impact of a social networking service-enhanced smart care model on stage 5 chronic kidney disease: quasi-experimental study. J Med Internet Res 2020;22(4):e15565 [FREE Full text] [doi: 10.2196/15565] [Medline: 32200348]
9. Ruan H. Access to health care in rural China: the influence of travel distance on healthcare decision-making. Innovation Aging 2017;1(suppl_1):569-570. [doi: 10.1093/geroni/igx004.2002]
10. Ravizza S, Huschto T, Adamov A, Böhm L, Büsser A, Flöther FF, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med 2019;25(1):57-59. [doi: 10.1038/s41591-018-0239-8] [Medline: 30617317]
11. Lightfoot CJ, Wilkinson TJ, Hadjiconstantinou M, Graham-Brown M, Barratt J, Brough C, et al. J Med Internet Res 2022 Nov 14;24(11):e39657 [FREE Full text] [doi: 10.2196/39657] [Medline: 36374538]
12. Shen H, van der Kleij R, van der Boog PJM, Wang W, Song X, Li Z, et al. Digital tools/eHealth to support CKD self-management: a qualitative study of perceptions, attitudes and needs of patients and health care professionals in China. Int J Med Inform 2022;165:104811 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104811] [Medline: 35753175]
13. Siddique AB, Krebs M, Alvarez S, Greenspan I, Patel A, Kinsolving J, et al. Mobile apps for the care management of chronic kidney and end-stage renal diseases: systematic search in app stores and evaluation. JMIR Mhealth Uhealth 2019;7(9):e12604 [FREE Full text] [doi: 10.2196/12604] [Medline: 31486408]
14. Ong SW, Jassal SV, Miller JA, Porter EC, Cafazzo JA, Seto E, et al. Integrating a smartphone-based self-management system into usual care of advanced CKD. Clin J Am Soc Nephrol 2016;11(6):1054-1062 [FREE Full text] [doi: 10.2215/CJN.10681015] [Medline: 27173169]
15. Singh K, Diamantidis CJ, Ramani S, Bhavsar NA, Mara P, Warner J, et al. Patients' and nephrologists' evaluation of patient-facing smartphone apps for CKD. Clin J Am Soc Nephrol 2019;14(4):523-529 [FREE Full text] [doi: 10.2215/CJN.10370818] [Medline: 30898873]
16. Chang AR, Bailey-Davis L, Hetherington V, Ziegler A, Yule C, Kwiecen S, et al. Remote dietary counseling using smartphone applications in patients with stages 1-3a chronic kidney disease: a mixed methods feasibility study. J Ren Nutr 2020;30(1):53-60 [FREE Full text] [doi: 10.1053/j.jrn.2019.03.080] [Medline: 31078403]

XSL•FO
RenderX

17. Markossian TW, Boyda J, Taylor J, Etingen B, Modave F, Price R, et al. A mobile app to support self-management of chronic kidney disease: development study. JMIR Hum Factors 2021;8(4):e29197 [FREE Full text] [doi: 10.2196/29197] [Medline: 34914614]
18. Vilasi A, Panuccio VA, Morante S, Villa A, Versace MC, Mezzatesta S, et al. Monitoring risk factors and improving adherence to therapy in patients with chronic kidney disease (Smit-CKD project): pilot observational study. JMIR Bioinform Biotechnol 2022;3(1):e36766 [FREE Full text] [doi: 10.2196/36766] [Medline: 38935948]
19. Stevenson JK, Campbell ZC, Webster AC, Chow CK, Tong A, Craig JC, et al. eHealth interventions for people with chronic kidney disease. Cochrane Database Syst Rev 2019;8(8):CD012379 [FREE Full text] [doi: 10.1002/14651858.CD012379.pub2] [Medline: 31425608]
20. Levin A, Bilous RW. Kidney disease: improving global outcomes (KDIGO) CKD work group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int Suppl 2013;3(1):1-150 [FREE Full text]
21. Sumida K, Nadkarni GN, Grams ME, Sang Y, Ballew SH, Coresh J, et al. Conversion of urine protein-creatinine ratio or urine dipstick protein to urine albumin-creatinine ratio for use in chronic kidney disease screening and prognosis: an individual participant-based meta-analysis. Ann Intern Med 2020;173(6):426-435 [FREE Full text] [doi: 10.7326/M20-0529] [Medline: 32658569]
22. Leffondre K, Boucquemont J, Tripepi G, Stel VS, Heinze G, Dunkler D. Analysis of risk factors associated with renal function trajectory over time: a comparison of different statistical approaches. Nephrol Dial Transplant 2015;30(8):1237-1243. [doi: 10.1093/ndt/gfu320] [Medline: 25326471]
23. Li W, Chiu F, Zeng J, Li Y, Huang S, Yeh H, et al. Mobile health app with social media to support self-management for patients with chronic kidney disease: prospective randomized controlled study. J Med Internet Res 2020;22(12):e19452 [FREE Full text] [doi: 10.2196/19452] [Medline: 33320101]
24. Borrelli S, Garofalo C, Mallamaci F, Tripepi G, Stanzione G, Provenzano M, et al. Short-term blood pressure variability in nondialysis chronic kidney disease patients: correlates and prognostic role on the progression of renal disease. J Hypertens 2018;36(12):2398-2405. [doi: 10.1097/HJH.0000000000001825] [Medline: 29995698]
25. Drawz PE, Alper AB, Anderson AH, Brecklin CS, Charleston J, Chen J, et al. Masked hypertension and elevated nighttime blood pressure in CKD: prevalence and association with target organ damage. Clin J Am Soc Nephrol 2016;11(4):642-652 [FREE Full text] [doi: 10.2215/CJN.08530815] [Medline: 26912547]
26. Agarwal R. Blood pressure components and the risk for end-stage renal disease and death in chronic kidney disease. Clin J Am Soc Nephrol 2009;4(4):830-837 [FREE Full text] [doi: 10.2215/CJN.06201208] [Medline: 19339424]
27. Carrero JJ, Grams ME, Sang Y, Ärnlöv J, Gasparini A, Matsushita K, et al. Albuminuria changes are associated with subsequent risk of end-stage renal disease and mortality. Kidney Int 2017;91(1):244-251 [FREE Full text] [doi: 10.1016/j.kint.2016.09.037] [Medline: 27927597]
28. Waheed S, Matsushita K, Astor BC, Hoogeveen RC, Ballantyne C, Coresh J. Combined association of creatinine, albuminuria, and cystatin C with all-cause mortality and cardiovascular and kidney outcomes. Clin J Am Soc Nephrol 2013;8(3):434-442 [FREE Full text] [doi: 10.2215/CJN.04960512] [Medline: 23258794]
29. Al-Wahsh H, Lam NN, Quinn RR, Ronksley PE, Sood MM, Hemmelgarn B, et al. Calculated versus measured albumin-creatinine ratio to predict kidney failure and death in people with chronic kidney disease. Kidney Int 2022;101(6):1260-1270. [doi: 10.1016/j.kint.2022.02.034] [Medline: 35398476]
30. Coresh J, Heerspink HJL, Sang Y, Matsushita K, Arnlov J, Astor BC, et al. Change in albuminuria and subsequent risk of end-stage kidney disease: an individual participant-level consortium meta-analysis of observational studies. Lancet Diabetes Endocrinol 2019;7(2):115-127 [FREE Full text] [doi: 10.1016/S2213-8587(18)30313-9] [Medline: 30635225]
31. Lee Y, Cui Y, Tu M, Chen Y, Chang P. Mobile health to maintain continuity of patient-centered care for chronic kidney disease: content analysis of apps. JMIR Mhealth Uhealth 2018;6(4):e10173 [FREE Full text] [doi: 10.2196/10173] [Medline: 29678805]
32. Schrauben SJ, Appel L, Rivera E, Lora CM, Lash JP, Chen J, et al. Mobile health (mHealth) technology: assessment of availability, acceptability, and use in CKD. Am J Kidney Dis 2021;77(6):941-950.e1 [FREE Full text] [doi: 10.1053/j.ajkd.2020.10.013] [Medline: 33309860]
33. Mertens A, Brandl C, Miron-Shatz T, Schlick C, Neumann T, Kribben A, et al. A mobile application improves therapy-adherence rates in elderly patients undergoing rehabilitation: a crossover design study comparing documentation via iPad with paper-based control. Medicine (Baltimore) 2016;95(36):e4446 [FREE Full text] [doi: 10.1097/MD.0000000000004446] [Medline: 27603339]
34. Hayashi A, Yamaguchi S, Waki K, Fujiu K, Hanafusa N, Nishi T, et al. Testing the feasibility and usability of a novel smartphone-based self-management support system for dialysis patients: a pilot study. JMIR Res Protoc 2017;6(4):e63 [FREE Full text] [doi: 10.2196/resprot.7105] [Medline: 28428168]
35. Hollander JE, Carr BG. Virtually perfect? Telemedicine for COVID-19. N Engl J Med 2020;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]

XSL•FO
RenderX

36. Schiffer L, Gertges R, Nöhre M, Schieffer E, Tegtbur U, Pape L, et al. Use and preferences regarding internet-based health care delivery in patients with chronic kidney disease. BMC Med Inform Decis Making 2021;21(1):34 [FREE Full text] [doi: 10.1186/s12911-020-01375-9] [Medline: 33522934]
37. Stauss M, Floyd L, Becker S, Ponnusamy A, Woywodt A. Opportunities in the cloud or pie in the sky? Current status and future perspectives of telemedicine in nephrology. Clin Kidney J 2021;14(2):492-506 [FREE Full text] [doi: 10.1093/ckj/sfaa103] [Medline: 33619442]
38. Tangri N, Ferguson TW. Artificial intelligence in the identification, management, and follow-up of CKD. Kidney360 2022;3(3):554-556 [FREE Full text] [doi: 10.34067/KID.0007572021] [Medline: 35582190]
39. Niel O, Bastard P. Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. Am J Kidney Dis 2019 Dec;74(6):803-810. [doi: 10.1053/j.ajkd.2019.05.020] [Medline: 31451330]

Abbreviations

ACR: albumin-to-creatinine ratio
AI: artificial intelligence
CKD: chronic kidney disease
eGFR: estimated glomerular filtration rate
ESRD: end-stage renal disease
IgA: immunoglobulin A
ISA: immunosuppressive agent
MAP: mean arterial pressure
mHealth: mobile health
OR: odds ratio
PSM: propensity score matching
RASB: renin-angiotensin-aldosterone system blocker

Edited by X Ma; submitted 01.11.23; peer-reviewed by N Koizumi, Y Li; comments to author 21.06.24; revised version received 16.08.24; accepted 14.10.24; published 25.11.24.

Please cite as:
Ma J, Wang J, Ying J, Xie S, Su Q, Zhou T, Han F, Xu J, Zhu S, Yuan C, Huang Z, Xu J, Chen X, Bian X
Long-Term Efficacy of an AI-Based Health Coaching Mobile App in Slowing the Progression of Nondialysis-Dependent Chronic Kidney Disease: Retrospective Cohort Study
J Med Internet Res 2024;26:e54206
URL: https://www.jmir.org/2024/1/e54206
doi:10.2196/54206
PMID:39402012

©Jianwei Ma, Jiangyuan Wang, Jiapei Ying, Shasha Xie, Qin Su, Tianmeng Zhou, Fuman Han, Jiayan Xu, Siyi Zhu, Chenyi Yuan, Ziyuan Huang, Jingfang Xu, Xuyong Chen, Xueyan Bian. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characterizing Performance on a Suite of English-Language NeuroUX Mobile Cognitive Tests in a US Adult Sample: Ecological Momentary Cognitive Testing Study

Emily W Paolillo[1], PhD; Jessica Bomyea[2,3], PhD; Colin A Depp[2,3], PhD; Ashley M Henneghan[4], RN, PhD; Anunay Raj[5], BTech; Raeanne C Moore[2], PhD

[1]Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, United States
[2]Department of Psychiatry, University of California San Diego School of Medicine, La Jolla, CA, United States
[3]VA San Diego Healthcare System, La Jolla, CA, United States
[4]School of Nursing, University of Texas at Austin, Austin, TX, United States
[5]NeuroUX, Inc., Dover, DE, United States

Corresponding Author:
Raeanne C Moore, PhD
Department of Psychiatry, University of California San Diego School of Medicine
9500 Gilman Dr.
La Jolla, CA, 92093
United States
Phone: 1 619 543 5378
Email: r6moore@health.ucsd.edu

Abstract

Background: Mobile cognitive testing is growing in popularity, with numerous advantages over traditional cognitive testing; however, the field lacks studies that deeply examine mobile cognitive test data from general adult samples.

Objective: This study characterized performance for a suite of 8 mobile cognitive tests from the NeuroUX platform in a sample of US adults across the adult lifespan.

Methods: Overall, 393 participants completed 8 NeuroUX cognitive tests and a brief ecological momentary assessment survey once per day on their smartphones for 10 consecutive days; each test was administered 5 times over the testing period. The tests tapped the domains of executive function, processing speed, reaction time, recognition memory, and working memory. Participants also completed a poststudy usability feedback survey. We examined alternate form test-retest reliability; practice effects; and associations between scores (averages and intraindividual variability) and demographics as well as test-taking context (ie, smartphone type, being at home vs not at home, and being alone vs not alone).

Results: Our final sample consisted of 393 English-speaking US residents (aged 20-79 y; female: n=198, 50.4%). Of the 367 participants who provided responses about their race and ethnicity, 258 (70.3%) were White. Of the 393 participants, 181 (46.1%) were iOS users, and 212 (53.9%) were Android users. Of 12 test scores derived from the 8 tests, 9 (75%) showed good to excellent test-retest reliability (intraclass correlation coefficients >0.76). Practice effects (ie, improvements in performance) were observed for 4 (33%) of the 12 scores. Older age was associated with worse performance on most of the test scores (9/12, 75%) and greater within-person variability for nearly all reaction time scores (3/4, 75%). Relationships with smartphone type showed better performance among iOS users and those with newer Android software versions compared to those with older software. Being at home (vs not at home) was associated with better performance on tests of processing speed. Being alone (vs not alone) was associated with better performance on tests of recognition and working memory. Poststudy feedback indicated that participants found NeuroUX easy to learn and use, an enjoyable experience, and an app that would be helpful in understanding their thinking skills. Only 4.2% (16/379) endorsed privacy concerns, and 77.3% (293/379) reported that they would be willing to share their results with their health care provider. Older age—but not other demographics—was associated with finding the tests more challenging.

Conclusions: In a sample of adults across a wide age range, this study characterized features that are particularly important for the interpretation of remote, repeated mobile cognitive testing performance, including test-retest reliability, practice effects,

XSL•FO
RenderX

smartphone type, and test-taking context. These data enhance the understanding and application of mobile cognitive testing, paving the way for improved clinical decision-making, personalized interventions, and advancements in cognitive research.

(J Med Internet Res 2024;26:e51978) doi:10.2196/51978

KEYWORDS

digital health; cognition; cognitive aging; neuropsychology; mobile health; psychometrics; mobile phone; Ecological Momentary Assessment; EMA; NeuroUX

Introduction

Background

Advances in smartphone technology have allowed researchers to couple brief cognitive testing with frequently repeated ecological momentary assessment (EMA), opening the door to studying cognitive performance as it occurs in the natural environment [1]. Mobile cognitive testing in the context of an EMA paradigm (also known as ecological momentary cognitive testing) has several advantages that enhance cognitive research, including the ability to (1) collect data as people go about their daily lives, (2) randomly sample cognitive assessments across various times and occasions to examine an individual's average performance and variability in performance, and (3) measure cognition alongside mood and behavior in real time and across the different contexts of daily life [2].

Traditionally, clinicians and researchers have relied on standardized paper-and-pencil tests or laboratory-based experiments to measure cognitive abilities. While these methods are gold standard assessments of cognitive functioning, they also have some limitations; for example, laboratory or clinical-based tests can be resource intensive because evaluations typically last >1 hour and require trained clinicians to administer and interpret the testing results [3]. They are also conducted in controlled (ie, artificial) testing environments, which may lack ecological validity [4]. Mobile cognitive tests are increasingly being used to assess cognitive performance as a method to enhance measurement of real-world cognitive function either in conjunction with traditional neuropsychological testing or as a stand-alone assessment [5]. This remote assessment method offers unique advantages over traditional neuropsychological testing in specific cases, including studying scientific questions related to cognitive performance in a naturalistic environment or acute fluctuations in cognitive performance [6]. It can also increase access to neuropsychological care by reducing geographic or mobility barriers for individuals [7]. Growing research has also shown that mobile cognitive testing is reliable and may be more sensitive to subtle or early cognitive deficits than single-administration neuropsychological tests [1,8-10].

Despite the growing popularity of mobile cognitive tests, there are challenges in the validation of these tests. For one, there is a notable lack of comprehensive data available from general, nonclinical populations for these assessment tools. These data provide a frame of reference for interpreting individual performance and enable comparisons across clinical populations when demographics are appropriately matched. They serve as a benchmark against which an individual's cognitive abilities may be evaluated, aiding in diagnostic decision-making, personalized interventions, and the identification of cognitive strengths and weaknesses. In a 2017 systematic review of the applications of mobile cognitive testing in clinical research, 12 articles were identified that used, and reported the psychometric properties of, self-administered mobile cognitive tests [1]. Since then, other groups have published additional work on the feasibility, reliability, and validity of different mobile cognitive testing platforms in small to midsize samples of various populations, including healthy adults [9,11]; healthy older adults [12]; and older adults at risk for Alzheimer disease (AD) [13], frontotemporal dementia [7], and Parkinson disease [14,15]. Second, as mobile cognitive tests produce within-person data, additional psychometric properties concerning within-person variability are possible that are not possible in single-administration cognitive tests; for example, Aschenbrenner et al [16] examined within-person variability in test performance across 28 repeated sessions of a mobile cognitive testing platform, administered over a 1-week period, and found increased variability on a processing speed task among persons at risk for AD compared to those without AD risk. In addition, with mobile cognitive tests, practice effects are measured over a far briefer period than in traditional tests, and a better understanding of these practice effect trends in healthy samples is needed. Third and last, because mobile cognitive tests are self-administered as participants go about their daily lives, context, such as the influence of being alone versus not alone and being at home versus not at home, needs to be considered when examining test performance.

Objectives

In our prior work in recent years, we have published on the psychometric properties of several of the NeuroUX mobile cognitive tests in clinical populations, including people with schizophrenia [17,18], those with bipolar disorder [17-19], and individuals with mild cognitive impairment [20]. The primary objective of this paper was to present data for a mobile cognitive testing protocol that includes a suite of 8 NeuroUX mobile cognitive tests, derived from a relatively large sample (n=393) of adults in the United States ranging in age from 20 to 79 years. Participants completed EMA surveys and mobile cognitive testing once daily for 10 days. The cognitive domains assessed included tests of processing speed, working memory, recognition memory, executive function, and reaction time. We present the distributions of scores for each test (averaged across study days), along with data on alternate form test-retest reliability, practice effects, and associations between scores (both averages and intraindividual variability) and demographics as well as test-taking context (ie, smartphone type, being at home vs not at home, being alone vs not alone, and the time of day). In addition, we report scores for each test by age bin and present usability feedback data from the participants.

XSL•FO
RenderX

Methods

Participants

A total of 394 participants from the United States were recruited via an opt-in, convenience sampling approach using Prolific, a web-based research and data collection platform (Prolific Academic Ltd). Prolific provides researchers access to a pool of >130,000 participants whose identities are verified using email, telephone number, and proof of identity (eg, driver's license), allowing quick and efficient remote data collection. Participants were recruited to Prolific through word of mouth—such as referrals from other Prolific participants—as well as through social media shares and flyers distributed on university campuses. Participants then sign up on Prolific through their web-based portal and enroll in various studies per individual preference. Each study must communicate the following to interested participants: the purpose of the research and the intended use of the collected data. Participants are fully anonymized, and Prolific provides an anonymized internal messaging service, which allows participants to message researchers (and vice versa). Researchers cannot access participants' identifiable information, and Prolific does not store any data provided within studies.

The inclusion criteria for this study included being aged between 20 and 79 years, having no diagnosis of cognitive impairment (per self-report), being of US nationality, being born in the United States, currently residing in the United States, and having English as the first language. No exclusionary criteria were applied. The target recruitment goal was 300 participants, with 50 participants per age decade (20-29 y, 30-39 y, etc), which was based on estimates from pervious similar short-term, repeated cognitive assessment studies. After careful review of the data, of the 394 participants recruited, 1 (0.3%) was excluded due to performing at the floor level of all NeuroUX tests, leaving 393 (99.7%) adults in the final sample.

Procedures

Participants completed 4 of the 8 NeuroUX cognitive tests and a brief EMA survey once per day on their personal smartphones for 10 consecutive days. Each daily session took 8 to 10 minutes to complete, and the link to complete the daily session was active from 7 AM to 8 PM, allowing participants flexibility to complete the tasks at their convenience over the course of the day. Each of the 8 NeuroUX cognitive tests described in this study was completed 5 times (Table 1). There are multiple versions of each test, and alternate forms of each test were used for each of the 5 administrations. To inform the context in which participants were completing the cognitive tests, daily EMA surveys asked participants to report their current location (being at home vs not at home) and whether they were alone or with others, as well as rate their mood. At the end of each session, participants completed EMA questions asking whether they were interrupted or distracted during the testing session. At the conclusion of the 10 days of testing, participants were sent a poststudy feedback survey, which asked questions about their experience with NeuroUX (answered with Likert-scale response options ranging from *strongly agree* to *strongly disagree*), whether they had privacy concerns related to their NeuroUX use, and whether they would be willing to share NeuroUX data with their health care provider.

Table 1. NeuroUX test administration protocol for this study in US adults.

Cognitive tests	Study day									
	1	2	3	4	5	6	7	8	9	10
Memory List	—[a]	2nd[b]	1st	—	1st	—	—	2nd	2nd	—
Memory Matrix	2nd	—	4th	—	—	3rd	1st	—	4th	—
Matching Pair	1st	—	—	2nd	—	2nd	—	4th	1st	—
Quick Tap 1	—	3rd	—	3rd	2nd	—	—	1st	—	2nd
Quick Tap 2	—	4th	—	4th	3rd	—	—	3rd	—	3rd
Odd One Out	3rd	—	2nd	—	4th	—	4th	—	—	1st
CopyKat	4th	—	3rd	—	—	1st	2nd	—	3rd	—
Hand Swype	—	1st	—	1st	—	4th	3rd	—	—	4th

[a]Not applicable.

[b]Values within the table represent the order in which the tests were administered on each study day. On the days when Memory List was administered, the recall portion was always administered at the end of the testing session.

Measures

Overview

NeuroUX is a proprietary platform designed to deliver mobile cognitive tests and EMA surveys. This platform has been implemented in several patient populations, including individuals with serious mental illness, older adults with mild cognitive impairment, and persons with cancer-related cognitive impairment [17,19-21]. Eight cognitive tests, as well as a poststudy feedback survey, described in the following subsections, were administered as part of this study.

Memory List

This test was designed to assess recognition memory (Figure 1A). Participants are asked to learn a list of 12 words presented simultaneously for 30 seconds, after which the list of words is removed, and participants complete another NeuroUX cognitive

XSL•FO
RenderX

test as a distractor task. Then, participants complete the recognition memory component of the Memory List test, which presents 24 words 1 by 1 (ie, the 12 target words mixed in with 12 distractor words). For each word, participants are asked to select "Yes" if the word was on the original list or select "No" if the word was not on the original list. The recognition trial takes approximately 1 minute to complete. Five different word lists were presented across sessions over the study period. The development of the word lists has been described previously [17]. The Memory List score is the total number of correct hits of target words and correct rejections of distractor words, with a total possible score of 24. Memory List scores of 0 were excluded due to suspected low effort or engagement.

Figure 1. Screenshots of each NeuroUX mobile cognitive test. (A) Memory List. (B) Memory Matrix. (C) Matching Pair. (D) Quick Tap 1. (E) Quick Tap 2. (F) Odd One Out. (G) CopyKat. (H) Hand Swype.

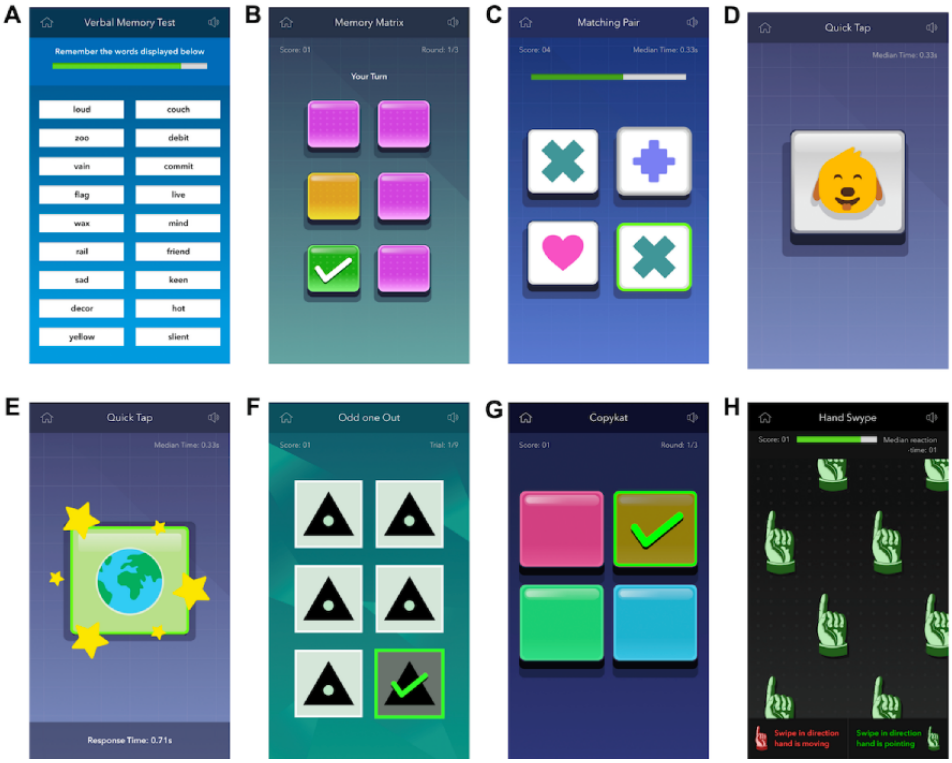

Memory Matrix

This test was designed to assess visual working memory (Figure 1B). Participants are shown a grid of tiles, with the number of tiles increasing gradually across correct trials. For each grid shown, some tiles are highlighted simultaneously for 1.5 seconds. When the highlighted color returns to the original tile color, participants are asked to select all tiles that were highlighted. The number of highlighted tiles increases with grid size across correct trials. Memory Matrix takes 1 to 2 minutes to complete. The total score represents the total number of tiles selected across all correct trials. The score does not change on trials when an incorrect response is given. Memory Matrix scores were excluded for instances in which participants were not able to move beyond >3 highlighted tiles, suggesting low effort or engagement.

Matching Pair

This test was designed to assess processing speed (Figure 1C). Participants are shown a matrix of tiles that each display various shapes of different colors. Participants are asked to select the 2 matching tiles as quickly as possible. The matrix gradually increases in size. Scores are calculated by adding total grid sizes of correct trials. The score does not change for incorrect responses. The task ends after 90 seconds.

Quick Tap 1

This test was designed to assess processing speed (Figure 1D). Participants are asked to tap the target image as quickly as possible when it appears on the screen. Incorrect responses are recorded when the target image is not tapped or if the display is tapped before the target image appears. Each session contains 12 trials, with the target appearing at randomly generated time intervals of between 1 and 5 seconds since the last target was displayed. The time to complete this task is approximately 1 minute. Total correct responses are recorded, with a maximum possible score of 12. Median response times (in ms) for correct trials are also recorded as the reaction time score for each session. Different images were shown as the target image across testing sessions over the study period.

Quick Tap 2

This task was designed to assess response inhibition and follows a classic Go–No-Go paradigm (Figure 1E). Participants are presented with either target images that they are asked to tap as quickly as possible or foil and trick images that they are asked to refrain from tapping. Similar to Quick Tap 1, each session contains 12 trials, with the target or foil images appearing at randomly generated time intervals of between 1 and 5 seconds since the last image was displayed. The time to complete this

XSL•FO
RenderX

task is 1 minute. Total correct responses are recorded, with a maximum possible score of 12, including either correctly tapping the target or not tapping the foil. Median response times (in ms) for correct target or Go trials were also recorded as the reaction time score for each session.

Odd One Out

This task was designed to assess both visual working memory and processing speed (Figure 1F). Participants are presented with 6 symbols, 5 of which are identical, and 1 is slightly different. Participants are asked to tap the symbol that is different as quickly as possible. Each session contains 9 trials and takes approximately 1 minute to complete. Total correct responses are recorded, with a maximum possible score of 9. Median response times (in ms) for correct responses are also recorded as the reaction time score for each session.

CopyKat

This task was designed to assess visual working memory (Figure 1G). The task operates similarly to the commercial game Simon such that in CopyKat, participants are shown a grid of colored tiles that light up 1 by 1 in a random order. Participants are asked to replicate the pattern by tapping the tiles in the same order that was shown. The grid size increases after each correct response. Responses are considered incorrect when the tiles are selected incorrectly or when there is no response within 20 seconds. The time to complete this task is approximately 2 to 3 minutes. The total score is the number of correct trials. CopyKat scores of <4 were excluded due to suspected low effort or engagement.

Hand Swype

This task was designed to assess cognitive flexibility (Figure 1H). Participants are shown several hands that are all pointing and moving across the screen in a particular direction (up, down, right, or left). The instructions change throughout the task and are always visible. Sometimes, participants are asked to swipe in the direction in which the hands are pointing regardless of the direction in which they are moving, and sometimes, they are asked to swipe in the direction in which the hands are moving. The proportion of errors made (ie, the number of errors/total trials) as well as the median response time (in ms) for correct trials are recorded for each session. The task ends after 60 seconds. Hand Swype median reaction times of <0.5 seconds were excluded due to suspected low effort or engagement.

Feedback Survey

At the completion of the 10 days of testing, participants were sent a feedback questionnaire that asked about feasibility, accessibility, and potential use.

Ethical Considerations

All participants provided informed consent for this study through the Prolific platform. Participants' data were fully anonymized during the data collection process, and researchers cannot access any identifiable participant information through Prolific. This study qualifies as exempt research under the 2018 Common Rule of the US Department of Health and Human Services and received exempt status approval from the University of California San Diego Institutional Review Board (811227). Data are securely stored on NeuroUX's Health Insurance Portability and Accountability Act (HIPAA)–compliant Amazon Web Services servers located in the United States.

For each daily session completed, participants were compensated US $1.50. Those who completed ≥8 of the 10 daily sessions, as well as the final feedback survey, received an additional bonus of US $2.50.

Statistical Analyses

Data from each test were aggregated across testing sessions within persons to assess both average performance and performance variability over the study period. Alternate form test-retest reliability was evaluated by calculating the intraclass correlation coefficient (ICC) for the 5 administrations of each test. ICC values for reliability were interpreted following published guidelines: poor (<0.50), moderate (0.50-0.75), good (0.76-0.90), and excellent (>0.90) [22]. To examine how the number of repeated administrations impacts the test-retest reliability of each test, we calculated cumulative ICCs, which represent the ICC values when including progressively more administrations. Practice effects were evaluated using linear mixed effects models. First, we examined linear changes in scores for each test over time (defined as study day). Person-specific random intercepts and effects of time were modeled. For scores with significant practice effects, mixed effects models with linear splines tested whether there was a changepoint at which improvements in performance leveled off [23]. Given the variable test-retest intervals across sessions, linear mixed effects models also examined test-retest interval as a predictor of score changes across sequential sessions for each test, covarying for session number and observed test scores from the prior session.

Next, associations between demographics and the average performance and within-person variability in performance for each NeuroUX test were evaluated using Pearson r correlation for continuous demographic variables (ie, age), Spearman correlation for ordinal demographic variables (ie, education), 2-tailed independent t tests for dichotomous demographic variables (ie, sex), and 1-way ANOVAs for multilevel categorical variables (ie, race and ethnicity). One-way ANOVAs were also used to examine differences in aggregate NeuroUX test performance by smartphone software type and version. Age-stratified analyses were subsequently conducted to understand whether relationships with sex, race and ethnicity, or smartphone type differed between younger (<50 y) and older (≥50 y) adults. To understand within-person fluctuations in NeuroUX test performance by contextual and environmental factors (reported for each session via EMA), separate linear mixed effects models were used for each test (outcome) and each time-varying contextual and environmental predictor: (1) being alone versus not alone, (2) being at home versus not at home, and (3) the time of day (before noon vs afternoon). Models examining the within-person relationship between being alone and test performance also covaried for the proportion of sessions during which participants reported being alone to appropriately parse apart between-person and within-person effects. Similarly, models examining the within-person

XSL•FO
RenderX

relationship between being at home and test performance covaried for the proportion of sessions during which participants reported being at home, and models examining the within-person relationship between the time of day and test performance covaried for the proportion of sessions during which they completed the tests before noon. Person-specific random intercepts were included in all models.

Finally, data from the poststudy feedback survey were interpreted using descriptive statistics. We also examined the relationship between demographics and the ratings of how challenging the tests were using correlation for continuous variables and ANOVA for categorical variables. All analyses were conducted in R (version 4.3.0; R Foundation for Statistical Computing).

Results

Participant Characteristics

Demographic and smartphone type data for the overall sample and by age (decade bins) are presented in Table 2. Participants were aged 44.60 (SD 16.10) years on average, and 50.4% (198/393) were female. Of the 367 participants who provided responses about their race and ethnicity, 258 (70.3%) were White. Of note, the representation of people of color was lower in older age bins (eg, 40/43, 93% White in the oldest age bin, 70-79 y). All participants had at least 12 years of education, with the greatest representation among those who had a bachelor's degree (138/373, 37%). Regarding smartphone type, a little more than half of the participants were Android users (212/393, 53.9%), and the remaining were iOS users (181/393, 46.1%).

Of the 40 unique test administrations that participants were asked to complete (8 tests, 5 times each), participants had a mean completion rate of 88% (SD 21%). Participants were most adherent to Matching Pair (mean completion rate 90%, SD 19%), followed by CopyKat (89%, SD 20%), Memory Matrix (88%, SD 20%), Odd One Out (88%, SD 20%), Quick Tap 1 (87%, SD 23%), Quick Tap 2 (87%, SD 23%), and Memory List (86%, SD 24%). This resulted in 20,662 tests taken by all participants over the study period.

XSL•FO
RenderX

Table 2. Participant characteristics.

	Overall sample (n=393)	Age bins (y)					
		20-29 (n=75)	30-39 (n=114)	40-49 (n=61)	50-59 (n=50)	60-69 (n=50)	70-79 (n=43)
Age (y), mean (SD)	44.60 (16.10)	24.77 (2.62)	34.54 (2.84)	44.20 (3.21)	53.70 (2.58)	65.10 (2.94)	72.28 (2.16)
Sex (male), n (%)	195 (49.6)	35 (46.7)	70 (61.4)	30 (49.2)	20 (40)	21 (42)	19 (44.2)
Race and ethnicity[a], n (%)							
Asian	21 (5.7)	11 (16.2)	7 (6.7)	3 (5.3)	0 (0)	0 (0)	0 (0)
Black	44 (12)	13 (19.1)	15 (14.4)	6 (10.5)	5 (10.6)	2 (4.2)	3 (7)
Hispanic or Latinx	27 (7.4)	10 (14.7)	12 (11.5)	2 (3.5)	2 (4.4)	1 (2.1)	0 (0)
White	258 (70.3)	30 (44.1)	64 (61.5)	41 (71.9)	40 (85.1)	43 (89.6)	40 (93)
Other	17 (4.6)	4 (5.9)	6 (5.8)	5 (8.8)	0 (0)	2 (4.2)	0 (0)
Education level[b], n (%)							
High school diploma or GED[c]	61 (16.4)	11 (16.3)	15 (14.1)	9 (14.5)	10 (20.4)	8 (16.3)	8 (18.5)
Some college, no degree	75 (20.1)	19 (27.9)	20 (18.9)	12 (20.7)	9 (18.4)	9 (18.4)	6 (14)
Associate degree	37 (9.9)	2 (2.9)	13 (12.3)	11 (19)	4 (8.2)	5 (10.2)	2 (4.7)
Bachelor's degree	138 (37)	31 (45.6)	43 (40.6)	20 (34.5)	18 (36.7)	15 (30.6)	11 (25.6)
Master's degree	49 (13.1)	3 (4.4)	12 (11.3)	5 (8.6)	7 (14.3)	7 (14.3)	15 (34.9)
Professional degree beyond bachelor's degree (eg, MD[d], PhD[e], DDS[f], DVM[g], JD[h], and EdD[i])	13 (3.5)	2 (2.9)	3 (2.8)	1 (1.7)	1 (2)	5 (10.2)	1 (2.3)
Smartphone type							
iOS	181 (46.1)	46 (61.3)	49 (43)	30 (49.2)	20 (40)	20 (40)	16 (37.2)
Android	212 (53.9)	29 (38.7)	65 (57)	31 (50.8)	30 (60)	30 (60)	27 (62.8)

[a]A total of 367 participants provided responses about their race and ethnicity.
[b]A total of 373 participants provided responses about their education level.
[c]GED: General Educational Development.
[d]MD: doctor of medicine.
[e]PhD: doctor of philosophy.
[f]DDS: doctor of dental surgery.
[g]DVM: doctor of veterinary medicine.
[h]JD: juris doctor.
[i]EdD: doctor of education.

Average and Within-Person Variability in NeuroUX Performance

Scores for each test were averaged within persons across all 5 administrations. Descriptive statistics of these average scores are reported (Table 3) to understand average performance and variability in performance between participants (histograms are presented in Figure 2). Given that participants took each test 5 times, it is also important to understand variability in tests scores within persons across administrations. SDs were calculated within persons across test administrations, and the sample average of these within-person SDs are reported in Table 4, along with the range of within-person SDs. Although the majority of within-person SDs fall below the between-person SD for each test score shown in Table 3, the maximum within-person SD values in Table 4 indicate that there were at least some participants whose performance varied widely over the study period.

XSL•FO
RenderX

Table 3. NeuroUX raw score distributions in US adults (n=393).

	Values, mean (SD; min-max)
Memory List score	19.8 (2.3; 12.8-24.0)
Memory Matrix score	44.3 (11.2; 21.0-89.2)
Matching Pair score	305.0 (74.5; 110.0-546.0)
Quick Tap 1 score	11.6 (0.6; 6.0-12.0)
Quick Tap 1 reaction time (ms)	404.0 (96.0; 253.8-966.5)
Quick Tap 2 score	10.9 (0.8; 7.8-12.0)
Quick Tap 2 reaction time (ms)	528.8 (98.2; 337.5-983.2)
Odd One Out score	8.4 (0.5; 5.2-9.0)
Odd One Out reaction time (ms)	1536.3 (490.5; 682.0-3784.0)
CopyKat, score	11.2 (3.5; 4.0-26.8)
Hand Swype reaction time (ms)	1808.6 (512.5; 662.4-4006.6)
Hand Swype proportion of errors	0.2 (0.1; 0.0-0.7)

Figure 2. Histograms of average scores for each NeuroUX test in US adults (n=393).

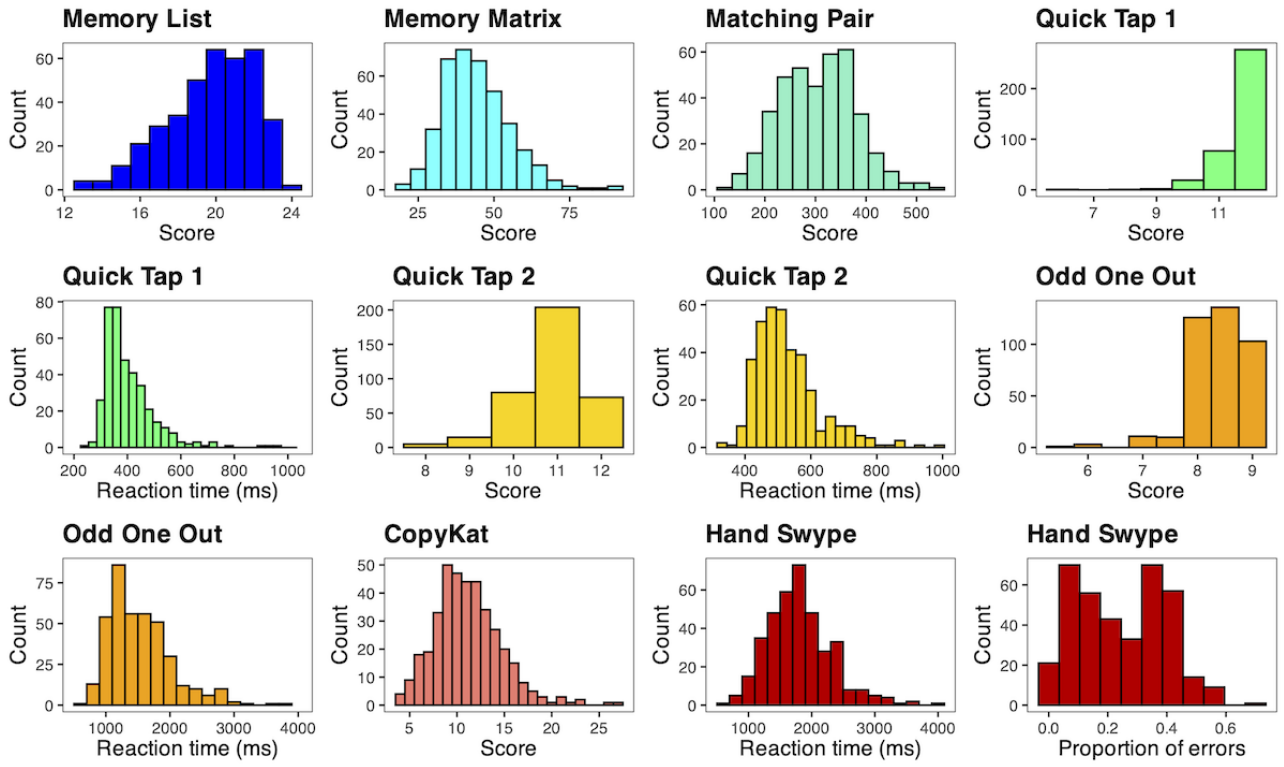

XSL•FO
RenderX

Table 4. Within-person variability and test-retest reliability across the administrations of each NeuroUX test in US adults (n=393).

	Within-person variability			Test-retest reliability, ICC[a]
	Average within-person SD	Minimum within-person SD	Maximum within-person SD	
Memory List score	2.1	0.0	8.5	0.762
Memory Matrix score	7.6	0.0	32.8	0.875
Matching Pair score	52.3	7.2	134.2	0.880
Quick Tap 1 score	0.5	0.0	4.2	0.608
Quick Tap 1 reaction time (ms)	45.8	3.8	653.7	0.912
Quick Tap 2 score	0.9	0.0	3.2	0.575
Quick Tap 2 reaction time (ms)	70.5	2.9	307.3	0.865
Odd One Out score	0.7	0.0	2.5	0.438
Odd One Out reaction time (ms)	382.3	18.0	1550.6	0.820
CopyKat score	3.0	0.0	9.2	0.807
Hand Swype reaction time (ms)	339.6	15.2	1917.3	0.864
Hand Swype proportion of errors	0.1	0.0	0.3	0.858

[a]ICC: intraclass correlation coefficient.

Alternate Form Test-Retest Reliability

ICCs for each test score are shown in Table 4. The majority of the test scores (9/12, 75%) showed good to excellent test-retest reliability (ICCs >0.760). However, the test scores for Odd One Out (ICC=0.438), Quick Tap 1 (ICC=0.608), and Quick Tap 2 (ICC=0.575) demonstrated poor to moderate test-retest reliability. Cumulative ICCs for each test score also showed that test-retest reliability generally increased with each repeated administration (Figure 3).

Figure 3. Cumulative intraclass correlation coefficients (ICCs) for each repeated NeuroUX test administration across 5 sessions in US adults (n=393).

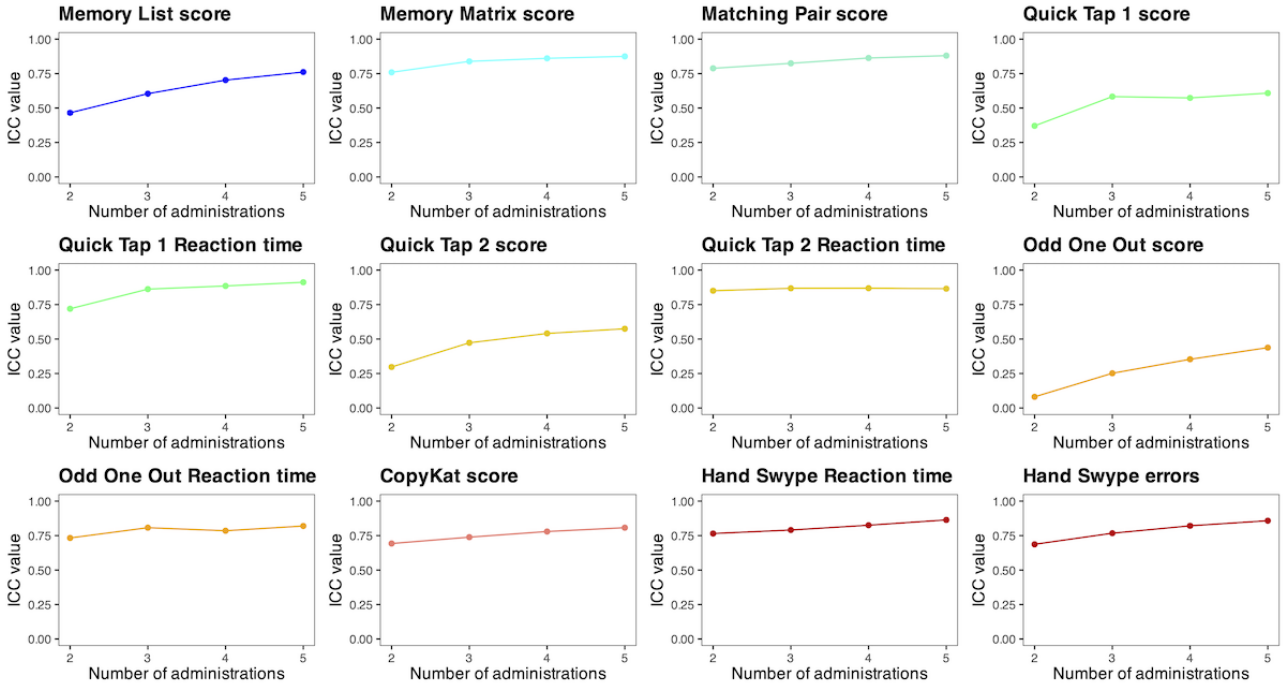

Practice Effects

Trends over time for each test score are shown in Figure 4. Statistically significant linear practice effects (ie, improvements in performance) were observed for Memory Matrix score (b=0.24, SE 0.07; P=.001), Matching Pair score (b=3.69, SE 0.47; P<.001), Quick Tap 1 reaction time (b=–1.38, SE 0.53; P=.009), and Hand Swype proportion of errors (b=–0.020, SE 0.001; P<.001). Subsequent mixed effects regression models with linear splines showed that practice effects in Memory Matrix score leveled off after study day 6, and Quick Tap 1 reaction time leveled off after study day 8 (Multimedia Appendix 1); however, there was no point at which Matching Pair performance or Hand Swype errors significantly leveled

XSL•FO
RenderX

off. Notably, several test scores (7/12, 58%) also showed that participants performed significantly worse over time on average, including Memory List score (*b*=–0.175, SE 0.020; *P*<.001), Quick Tap 2 score (*b*=–0.039, SE 0.010; *P*<.001), Quick Tap 2 reaction time (*b*=1.489, SE 0.680; *P*=.03), Odd One Out score (*b*=–0.072, SE 0.010; *P*<.001), Odd One Out reaction time (*b*=54.196, SE 3.180; *P*<.001), CopyKat score (*b*=–0.145, SE 0.030; *P*<.001), and Hand Swype reaction time (*b*=10.499, SE 3.710; *P*=.005). There was no significant change in Quick Tap 1 score over time (*b*=0.001, SE 0.006; *P*=.82). Additional analyses showed that test-retest interval was not a statistically significant predictor of change scores for any test (Multimedia Appendix 2).

Figure 4. Trends in performance over time (5 sessions across 10 days) for each NeuroUX test score in US adults (n=393). RT: reaction time.

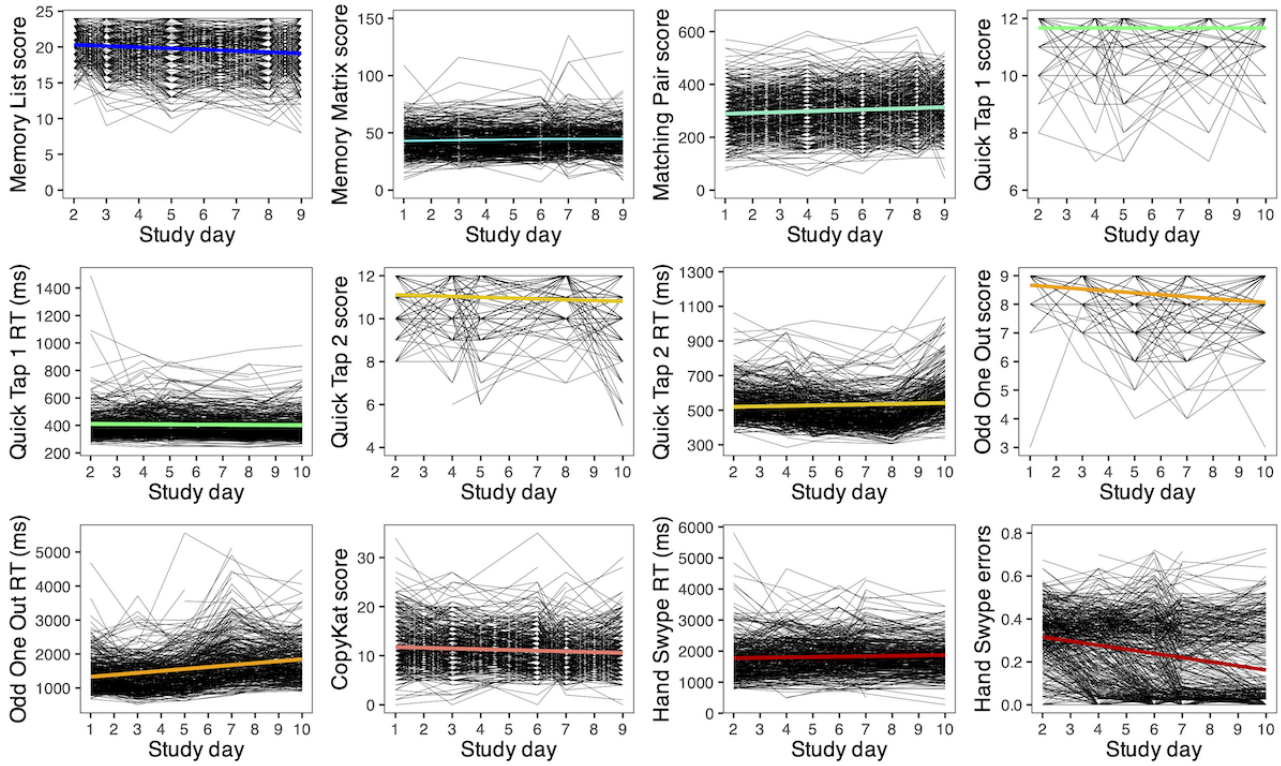

Associations With Demographics and Test-Taking Context

Older age was associated with worse average Memory Matrix score (*r*=–0.48; *P*<.001), Matching Pair score (*r*=–0.51; *P*<.001), Quick Tap 1 reaction time (*r*=0.44; *P*<.001), Quick Tap 2 reaction time (*r*=–0.48; *P*<.001), Odd One Out reaction time (*r*=0.61; *P*<.001), CopyKat score (*r*=–0.22; *P*<.001), Hand Swype reaction time (*r*=0.31; *P*<.001), and Hand Swype proportion of errors (*r*=0.12; *P*=.03). Notably, older age was also associated with better average Quick Tap 2 score (*r*=0.27; *P*<.001). Age was not significantly associated with the other average test scores (*r* values ranging from –0.019 to 0.078; *P* values >.13). We present means and SDs of average performance for each test in Table 5. Regarding within-person variability, older age was associated with greater within-person SDs for nearly all reaction time scores (3/4, 75%), including those for Quick Tap 1 (*r*=0.18; *P*<.001), Odd One Out (*r*=0.32; *P*<.001), and Hand Swype (*r*=0.14; *P*=.01). By contrast, older age was associated with lower within-person SDs (ie, less variability) on total score metrics for almost every test (6/8, 75%), including Memory List (*r*=–0.10; *P*=.04), Memory Matrix (*r*=–0.18; *P*=.001), Matching Pair (*r*=–0.29; *P*<.001), Quick Tap 1 (*r*=–0.12; *P*=.02), Quick Tap 2 (*r*=–0.19; *P*<.001), and CopyKat (*r*=–0.18; *P*<.001).

Education was significantly positively associated with average CopyKat score (ρ=0.14; *P*=.01). No other average test scores were strongly associated with education level (ρ values ranging from –0.09 to 0.09; *P* values >.08). Regarding within-person variability, higher education was associated with greater within-person SDs for Odd One Out reaction time (ρ=0.13; *P*=.02). Education was not associated with within-person variability on any other test (ρ values ranging from –0.09 to 0.07; *P* values >.08).

Sex differences in average performance were found across several tests. Women had better Memory List scores (mean 20.1, SD 2.1) than men (mean 19.4, SD 2.4; *P*=.002). Men had faster reaction times across almost all tasks compared to women, including Quick Tap 1 reaction time (mean difference 41 ms; *P*<.001), Quick Tap 2 reaction time (mean difference 36 ms; *P*<.001), and Hand Swype reaction time (mean difference 197 ms; *P*<.001). Within-person variability did not differ significantly by sex for any test score (*P* values >.06). Age-stratified analyses (<50 y and ≥50 y) showed that sex differences in average Memory List scores were driven by older adults (mean difference 1.33; *P*<.001), not younger adults (mean difference 0.44; *P*=.15), whereas sex differences in average reaction times were driven by younger adults (*P* values <.001), not older adults (*P* values >.05).

XSL•FO
RenderX

Table 6 presents differences in test performance by race and ethnicity. Differences in average scores by race and ethnicity were significantly different for Memory Matrix score, Matching Pair score, Quick Tap 1 reaction time, Quick Tap 2 score, Quick Tap 2 reaction time, Odd One Out score, Odd One Out reaction time, CopyKat score, and Hand Swype reaction time. Generally, Asian participants tended to perform best on these tasks compared to the other groups, with the exception of Odd One Out score for which the only significant pairwise difference showed that White participants scored higher than Black participants. Age-stratified analyses showed that these group differences tended to be driven by younger adults, with the exception of the significant group differences in average performance on Odd One Out reaction time, which was exhibited in both younger and older adults. Regarding within-person variability, average within-person SDs also differed by race and ethnicity for Memory List score, Memory Matrix score, Quick Tap 2 score, Odd One Out score, and Odd One Out reaction time. Group differences in within-person variability on these tasks were less consistent, with Asian participants sometimes having greater variability (ie, on Memory Matrix) and sometimes having less variability (ie, Odd One Out reaction time) than other groups. White participants also demonstrated less within-person variability than other groups on Memory List, Quick Tap 2, and Odd One Out. Again, group differences in within-person variability were all driven by younger adults; there were no significant differences in within-person variability by race and ethnicity among older adults (*P* values >.11).

Multimedia Appendix 3 presents differences in test performance by smartphone software type and version. Average scores across smartphone types were different for Memory Matrix score, Matching Pair score, Quick Tap 1 reaction time, Quick Tap 2 reaction time, Odd One Out reaction time, and CopyKat score. Participants using iOS or newer Android software versions (ie, Android 13) tended to perform better than participants with older Android software versions (eg, Android 10 or lower). These associations remained statistically significant after covarying for age, sex, education, and race and ethnicity (White vs people of color). In addition, age-stratified analyses showed consistent differences in average reaction times across smartphone type among both younger and older adults. Regarding within-person variability, average within-person SDs also differed by smartphone type for Quick Tap 1 reaction time and Odd One Out reaction time. Again, participants using newer software versions tended to be less variable than those using older versions. These associations also remained statistically significant after covarying for age, sex, education, and race and ethnicity (White vs people of color). However, in age-stratified analyses, none of the differences in within-person variability by smartphone type held in younger or older adults (*P* values >.05).

Separate linear mixed effects models for each test score showed that being at home during the NeuroUX testing session was associated with better participant performance for Matching Pair score (b=11.50, SE 5.18; P=.03), Quick Tap 1 score (b=0.14, SE 0.06; P=.02), and Quick Tap 1 reaction time (b=–12.46, SE 5.58; P=.03) compared to when participants were not at home while taking these tests. There were no significant within-person associations between being at home (vs not at home) and performance on any other tests (*P* values >.05). Similarly, linear mixed effects models showed that being alone during the NeuroUX testing session was associated with better performance for Memory List score (b=0.90, SE 0.20; P<.001), Memory Matrix score (b=1.92, SE 0.73; P=.009), and CopyKat score (b=0.60, SE 0.28; P=.03). There were no significant within-person associations between being alone (vs not alone) and performance on any other tests (*P* values >.05). Finally, linear mixed effects models showed that NeuroUX session start times before noon were associated with better performance for Odd One Out score (b=0.15, SE 0.05; P=.003) and CopyKat score (b=0.62, SE 0.22; P=.005). There were no significant within-person associations between the time of day (before noon vs afternoon) and performance on any other tests (*P* values >.05). Of note, there was a small yet significant association between older age and a higher proportion of tests taken at home (r=0.15; P=.002); however, age was not significantly associated with the proportion of tests taken alone (r=–0.06; P=.27) nor average NeuroUX session start time (ie, the hour of the day from 7 AM to 8 PM; r=–0.05; P=.36).

XSL•FO
RenderX

Table 5. Average NeuroUX raw scores by age.

NeuroUX test score	Age bins (y)					
	20-29 (n=75), mean (SD)	30-39 (n=114), mean (SD)	40-49 (n=61), mean (SD)	50-59 (n=50), mean (SD)	60-69 (n=50), mean (SD)	70-79 (n=43), mean (SD)
CopyKat score	11.89 (3.97)	12.00 (3.64)	10.72 (2.7)	11.72 (3.91)	10.12 (2.54)	9.26 (2.91)
Hand Swype reaction time	1641.68 (428.51)	1709.54 (410.71)	1748.54 (470.11)	1835.15 (517.44)	2085.11 (598.03)	2073.41 (611.28)
Hand Swype errors	0.24 (0.17)	0.21 (0.14)	0.28 (0.14)	0.24 (0.14)	0.26 (0.15)	0.29 (0.15)
Matching Pair score	343.58 (75.73)	336.25 (63.03)	302.27 (65.51)	284.69 (65.54)	254.30 (52.03)	237.31 (57.18)
Memory List score	19.50 (2.74)	20.06 (2.29)	19.82 (2.21)	19.69 (2.20)	19.89 (1.75)	19.41 (2.30)
Memory Matrix score	51.40 (11.40)	47.88 (10.61)	43.44 (10.57)	40.98 (9.38)	37.22 (6.85)	35.14 (6.62)
Odd One Out score	8.33 (0.56)	8.46 (0.42)	8.34 (0.58)	8.36 (0.60)	8.50 (0.35)	8.39 (0.59)
Odd One Out reaction time	1195.47 (283.89)	1344.09 (326.94)	1470.19 (346.64)	1692.96 (482.83)	1980.56 (481.33)	2038.25 (517.61)
Quick Tap 1 score	11.53 (0.53)	11.60 (0.54)	11.68 (0.39)	11.60 (0.62)	11.81 (0.28)	11.57 (1.09)
Quick Tap 1 reaction time	361.63 (59.24)	366.51 (63.28)	407.30 (80.13)	429.48 (104.47)	462.00 (119.48)	468.89 (117.34)
Quick Tap 2 score	10.46 (0.84)	10.90 (0.76)	11.01 (0.78)	11.03 (0.78)	11.19 (0.65)	11.14 (0.57)
Quick Tap 2 reaction time	484.33 (71.19)	486.62 (74.46)	524.27 (78.34)	554.99 (96.65)	615.66 (116.88)	585.35 (92.24)

XSL•FO
RenderX

Table 6. Aggregate performance on each NeuroUX test by race and ethnicity (n=367).

Age (y)	Asian (n=21), mean (SD)	Black (n=44), mean (SD)	Hispanic (n=27), mean (SD)	White (n=258), mean (SD)	Other (n=17), mean (SD)	*P* value	Pairwise comparisons[a] (White >Asian, Black, Hispanic, and other)
	30.67 (7.86)	38.23 (14.49)	33.74 (10.37)	49.14 (16.04)	37.88 (12.03)	<.001	
Average performance							
Memory List score	19.93 (2.39)	19.24 (2.97)	19.85 (2.35)	19.76 (2.16)	21.10 (1.73)	.09	N/A[b]
Memory Matrix score	57.24 (18.94)	41.92 (10.27)	45.86 (9.15)	42.99 (10.25)	44.68 (8.88)	<.001	Asian >Black, Hispanic, White, and other
Matching Pair score	349.94 (72.44)	287.76 (75.91)	322.40 (66.40)	299.41 (74.90)	335.41 (69.34)	.004	Asian >Black and White
Quick Tap 1 score	11.65 (0.52)	11.50 (0.54)	11.68 (0.25)	11.64 (0.64)	11.58 (0.32)	.63	N/A
Quick Tap 1 reaction time	353.88 (57.70)	428.52 (105.50)	382.18 (56.23)	410.16 (102.33)	361.89 (58.64)	.008	Asian <Black
Quick Tap 2 score	10.64 (1.00)	10.70 (0.91)	10.79 (0.76)	11.01 (0.72)	10.72 (1.01)	.02	__[c]
Quick Tap 2 reaction time	470.20 (38.80)	557.05 (113.38)	512.43 (75.13)	536.41 (100.12)	472.95 (82.57)	.001	Asian <Black and White
Odd One Out score	8.39 (0.44)	8.14 (0.70)	8.42 (0.30)	8.42 (0.50)	8.46 (0.38)	.02	White >Black
Odd One Out reaction time	1177.34 (228.32)	1755.51 (555.67)	1410.75 (327.70)	1585.60 (493.17)	1378.22 (475.68)	<.001	Asian <Black and White; Hispanic and other <Black
CopyKat score	12.29 (4.11)	9.65 (3.55)	11.20 (2.74)	11.21 (3.60)	11.64 (2.09)	.04	Asian >Black
Hand Swype reaction time	1564.90 (404.69)	1697.85 (504.13)	1731.54 (327.44)	1858.84 (525.08)	1818.65 (626.17)	.045	—
Hand Swype errors	0.22 (0.16)	0.29 (0.15)	0.23 (0.13)	0.24 (0.15)	0.24 (0.15)	.32	N/A
Within-person variability in performance (SD)							
Memory List score	1.97 (0.84)	2.55 (1.31)	2.66 (1.70)	2.04 (0.97)	1.61 (0.76)	.001	White and other <Black and Hispanic
Memory Matrix score	11.88 (7.55)	7.77 (3.07)	6.66 (3.82)	7.12 (3.78)	8.40 (2.56)	<.001	Asian >Black, Hispanic, White, and other
Matching Pair score	63.80 (25.29)	59.93 (26.02)	46.52 (22.36)	50.62 (23.24)	51.21 (17.71)	.01	—
Quick Tap 1 score	0.56 (0.59)	0.62 (0.57)	0.43 (0.29)	0.47 (0.57)	0.47 (0.25)	.46	N/A
Quick Tap 1 reaction time	38.51 (48.16)	45.15 (28.92)	36.62 (21.48)	48.29 (51.60)	34.61 (18.31)	.52	N/A
Quick Tap 2 score	1.02 (0.53)	1.12 (0.70)	1.05 (0.50)	0.83 (0.50)	0.75 (0.40)	.003	White <Black
Quick Tap 2 reaction time	67.09 (34.79)	73.83 (28.03)	72.38 (56.96)	71.11 (39.73)	58.60 (29.97)	.73	N/A
Odd One Out score	0.71 (0.42)	0.93 (0.49)	0.77 (0.30)	0.65 (0.41)	0.67 (0.44)	.001	White <Black
Odd One Out reaction time	252.29 (138.84)	465.28 (277.51)	342.55 (222.01)	389.48 (241.04)	358.96 (167.07)	.01	Asian <Black
CopyKat score	3.09 (1.36)	2.98 (1.38)	2.91 (1.80)	2.92 (1.49)	3.00 (0.83)	.99	N/A
Hand Swype reaction time	252.86 (210.82)	309.55 (223.32)	343.95 (139.10)	356.47 (277.72)	271.04 (144.03)	.26	N/A
Hand Swype errors	0.11 (0.05)	0.11 (0.06)	0.12 (0.05)	0.11 (0.06)	0.12 (0.05)	.85	N/A

[a]Pairwise comparisons were evaluated using the Tukey honestly significant difference test for scores with omnibus group differences.

[b]N/A: not applicable.

XSL•FO
RenderX

[C]No significant pairwise differences after Tukey honestly significant difference test adjustment for multiple comparisons.

Poststudy Feedback

Of the 393 participants, 379 (96.4%) completed the feedback survey. Feedback about NeuroUX use and experience was generally positive (Table 7). A majority of the participants agreed or strongly agreed that (1) NeuroUX is easy to use (379/379, 100%), (2) it is easy to learn to use (376/379, 99.2%), (3) they were satisfied with NeuroUX (365/379, 96.3%), (4) they enjoyed the experience (367/379, 96.8%), and (5) an app like this would be helpful to understand their thinking skills (308/379, 81.3%). Only 4.2% (16/379) of the participants endorsed privacy concerns related to NeuroUX use, and these 16 participants were evenly distributed within the age bins (ie, 2-4 participants within every age bin reported privacy concerns). In addition, 77.3% (293/379) reported being willing to share the results with their health care provider. Participants also rated how challenging they found the NeuroUX tests on a scale ranging from 1 (*not at all*) to 10 (*extremely*); the average rating was 5.7 (SD 2.1). Pearson *r* correlation showed that challenge ratings were positively associated with age (r=0.28; P<.001) such that older adults in this study found the tests more challenging. However, challenge ratings were not associated with sex (mean difference 0.07; P=.76), education (ρ=–0.04; P=.47), or race and ethnicity (all Tukey honestly significant difference–corrected pairwise comparisons; P values >.11). Covarying for age, challenge ratings were also associated with worse average performance on Memory List (b=–0.12, SE 0.06; P=.048) but were not significantly associated with performance on other tests.

Table 7. Results from the poststudy feedback survey assessing user experience with NeuroUX (n=379).

Item	Strongly disagree, n (%)	Disagree, n (%)	Neither agree nor disagree, n (%)	Agree, n (%)	Strongly agree, n (%)
Is it easy to use?	0 (0)	0 (0)	1 (0.3)	103 (27.2)	276 (72.8)
Is it easy to learn to use?	0 (0)	1 (0.3)	2 (0.5)	99 (26.1)	278 (73.4)
I am satisfied with it.	0 (0)	0 (0)	14 (3.7)	130 (34.3)	235 (62)
It was a burden to take the tests.	212 (55.9)	145 (38.3)	16 (4.2)	4 (1.1)	2 (0.5)
I enjoyed the experience.	0 (0)	1 (0.3)	12 (3.2)	147 (38.8)	219 (57.8)
An app like this would be helpful to understand my thinking skills.	2 (0.5)	9 (2.4)	60 (16)	153 (40.4)	155 (40.9)

Discussion

Principal Findings

This study fills a critical gap in existing literature by providing data derived from brief, repeated, and self-administered mobile cognitive tests across the adult lifespan. We presented NeuroUX data from a general sample of US adults, including average performance and within-person variability in performance for 8 tests, each with 5 in-the-wild testing sessions over 10 days. We characterized features that are particularly important for the interpretation of performance in the repeated testing format, including test-retest reliability showing good-to-excellent estimates for most of the tests (9/12, 75%) and practice effects showing variable trends in performance on each test over time. Importantly, we also characterized features that are important to consider in remote, smartphone-based testing contexts, including relationships between cognitive test performance and (1) smartphone software type showing that newer software related to better and less variable performance, as well as (2) test-taking context showing that participants performed better on some tests when they were at home, were alone, or took the tests before noon. With these data, we aim to enhance the understanding and application of mobile cognitive testing, paving the way for improved clinical decision-making, personalized interventions, and advancements in cognitive research.

These data contribute to our overarching understanding of the range of performance on this mobile testing platform among US adults and demonstrate that even a nonclinical sample of participants have a wide range of inter- and intraindividual variabilities in performance. The range of intraindividual variability, particularly given the varying contextual factors in a real-world testing format, means that it can be quite *normal* for a person's scores during any 1 testing session to be outside of a typically expected range (eg, –1 SD to +1 SD from the mean). Furthermore, it is also necessary to consider the influence of the test administration protocol (ie, the number of times each test is administered and the time between each administration) on test performance. The vast literature on repeat neuropsychological testing suggests that practice effects are highly affected by the length of the interval between testing sessions as well as the number of exposures to testing [24-26]. Thus, it is possible that the distribution of scores presented here may be different from that in other studies using different mobile test administration schedules. While our data provide a general sense of how adults across the lifespan tend to perform on these NeuroUX tests, they may be somewhat specific to our study conditions, and researchers are encouraged to use their own control sample to make accurate comparisons against specific clinical groups. Additional studies are needed to explore how different factors related to the test administration protocol and timing may affect aggregate score estimates.

The results showed good to excellent alternate form test-retest reliability for most of the tests (9/12, 75%); however, Odd One Out score, Quick Tap 1 score, and Quick Tap 2 score reliability estimates were suboptimal. This may be related to the limited

XSL•FO
RenderX

range of possible scores on these tests, as well as the tendency for participants to perform at the ceiling level. The lower estimates of alternate form test-retest reliability for these test scores negatively impact the reliability and validity of their aggregated scores. This is consistent with previous NeuroUX studies showing that these scores were not strongly related to gold standard in-person neuropsychological tests and did not discriminate clinical groups well (ie, healthy control vs bipolar disorder [19]). Notably, the reaction times for each of these tests still showed excellent alternate form test-retest reliability and showed strong associations with age and sex in expected directions [27], making it worthwhile for NeuroUX users to continue administering these tests to participants if additional reaction time measures are indicated. Furthermore, the observed increases in alternate form test-retest reliability per each additional test administration supports the repeated testing paradigm for mobile cognitive testing.

Practice effects are also very important to consider when using a repeated testing format. Although several test scores (4/12, 33%) showed expected practice effects (ie, Memory Matrix score, Matching Pair score, Quick Tap 1 reaction time, and Hand Swype proportion of errors), Quick Tap 1 score had no significant change over time, and the remaining test scores (7/12, 58%) unexpectedly showed worse performance over time. Similar decreases in performance over time have been shown previously across some NeuroUX tests [19]; however, the reason for this trend is unclear. It is possible that these decreases in performance over time may be related to unmeasured factors such as effort, fatigue, and boredom. Although it is possible that our convenience sampling approach may increase the likelihood of study disengagement over time, the consistency of these findings across other university-recruited clinical samples [19] argues against this as a primary cause. Indeed, attrition is common in longitudinal cognitive studies [28,29]. Other possibilities include factors related to the alternate forms of each test, including varying perceptions of difficulty, or possible proactive interference effects specifically on Memory List. Regardless of the underlying cause, our characterization of these trends in performance over time in a general sample of adults is a step toward improving the interpretation of NeuroUX performance in clinical populations.

The observed associations between each test and age support their validity; for example, most of the average test scores (8/12, 67%) were negatively associated with age, consistent with well-known effects of normal aging on cognitive performance [30], and support the tests' ability to be sensitive to age-related cognitive change. Intraindividual variability in performance across administrations was also sensitive to age such that reaction time variability was positively associated with age, and total score variability was negatively associated with age. Greater intraindividual variability in reaction times with age has been consistently supported throughout the literature [31] and may even be a marker of steeper incident cognitive decline [31,32]. Although the negative association between total score variability and age was somewhat unexpected, most of the existing literature on intraindividual variability has focused on reaction time tasks or other timed measures of processing speed. While more research may be needed to understand this paradoxical relationship between the variability in speed and accuracy metrics across the lifespan, it is possible that differences in environmental contextual factors during our in-the-wild test administrations (eg, location, noise, and external distractors), differences in study engagement or conscientiousness, or selection bias by age [33] may be playing a role in these associations.

Regarding test-taking context, we found that performance was generally better when participants were at home, were alone, or took the tests before noon, although this was only statistically significant for some of the tests. Depending on specific research questions for the future use of mobile cognitive testing platforms such as NeuroUX, researchers may consider requesting that participants take tests at home and alone in the morning hours. Alternatively, researchers may want to consider incorporating even more questions into the paired EMA survey about participants' environment while taking the tests, with the potential to covary for these contextual factors in analyses.

Participants' feedback about their study experience was generally positive. As the use of mobile cognitive testing platforms continues to grow, understanding participant burden, willingness, difficulty, and effort to complete cognitive tests in a "burst" repeated testing schedule is crucial for interpreting their validity to accurately assess cognitive functioning. Given our use of "gamified" cognitive tests developed in conjunction with a software design team to optimize the aesthetics, presentation, and user-friendliness of NeuroUX, our platform may be particularly well suited to create a more pleasant cognitive testing experience than traditional paper-and-pencil neuropsychological testing.

Limitations

A thorough review of study limitations is necessary to inform future research in this field. First, the convenience sampling approach to participant recruitment limits the generalizability of our sample because possible selection bias may have influenced the range of test scores captured here [34]. In addition, regarding our sample of participants, we had limited power to test differences in performance across racial and ethnic groups, particularly in the older age ranges. Furthermore, because the older age bins had very little representation from people of color, average performance from the older adults in this sample cannot be considered representative and would not reflect a culturally appropriate comparison against future older adult NeuroUX participants from minoritized populations [35]. Future directions include gathering data in more racially and ethnically diverse general adult samples.

Next, the length of our assessment burst allowed for each test to be administered only 5 times. While this may be sufficient for many future studies depending on any given research question, 5 administrations may not be sufficient to achieve the most optimal reliability and subsequent validity of the aggregated scores. In addition, this study only examined the 12-item Memory List, whereas NeuroUX also has word lists of other lengths available. In fact, the 12-item and 18-item word lists have shown similarly strong correlations with in-person, gold standard memory testing among adults with serious mental illness [17], and future studies may want to implement the

XSL•FO
RenderX

18-word list to improve sensitivity to more subtle memory difficulties. Some of the tests demonstrated ceiling effects, which limits the distribution of test scores and power to detect relationships with clinically meaningful variables (eg, cognitive impairment). On the basis of these data, those tests are being modified within the NeuroUX platform to improve their psychometric properties. Moreover, given the completely remote nature of this study and the use of a third-party recruitment source, we were limited in our ability to characterize much else about our participants beyond demographics. Future studies may benefit from collecting additional data about socioeconomic status, medical history, and perceived current functioning. Finally, as with any study using cognitive tests, we do not know whether participants were providing optimal effort on every test and every administration. Although true effort is difficult to measure, neuropsychologists in clinical practice often use performance validity tests as a proxy of effortful performance [36,37]. Additional research is needed to develop and assess potential performance validity metrics for use with mobile cognitive testing platforms.

Conclusions

In summary, this study represents an important step forward in the field of mobile cognitive testing by characterizing performance and the psychometric properties of several NeuroUX tests in a general sample of US adults across the lifespan. Given the ability for mobile cognitive testing to be delivered completely remotely, thereby increasing access to neuropsychological services for individuals who would not otherwise have access to an in-person clinic, there is great potential for a platform such as NeuroUX to be implemented in large-scale national or international research studies. The data generated through this study may be of particular use to studies that will implement the same protocol (ie, administer 4 tests once per day for 10 days) to use as a benchmark comparison. Mobile cognitive testing may also be useful in short prospective observational studies or to provide outcome measures in clinical trials in bursts (including before treatment, during treatment, and after treatment) or administered continuously throughout a trial. In conclusion, this study's comprehensive characterization of data on a suite of mobile cognitive tests in a general, nonclinical US adult sample highlights its potential for implementation in large-scale research studies and enabling remote access to neuropsychological services for broader populations.

Acknowledgments

This study was supported by NeuroUX Inc. The views expressed in this paper are the authors' own and do not necessarily represent the official position of the institution.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

AMH was a consultant for Prodeo, Inc at the time of writing. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its conflict-of-interest policies. AR is a cofounder of NeuroUX Inc. RCM is a cofounder of KeyWise AI, Inc, and NeuroUX Inc. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Results of mixed effects regression models with linear splines showing practice effects leveling off for 2 NeuroUX tests: Memory Matrix score and Quick Tap 1 reaction time.
[DOCX File , 14 KB - jmir_v26i1e51978_app1.docx]

Multimedia Appendix 2
Associations between test-retest interval and change in each NeuroUX score across sequential sessions.
[DOCX File , 15 KB - jmir_v26i1e51978_app2.docx]

Multimedia Appendix 3
Aggregate performance on each NeuroUX test by smartphone software type and version.
[DOCX File , 21 KB - jmir_v26i1e51978_app3.docx]

References

1. Moore RC, Swendsen J, Depp CA. Applications for self-administered mobile cognitive assessments in clinical research: a systematic review. Int J Methods Psychiatr Res 2017 Dec 31;26(4):e1562 [FREE Full text] [doi: 10.1002/mpr.1562] [Medline: 28370881]

XSL•FO
RenderX

2. Koo BM, Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. Innov Aging 2019 Jan;3(1):igy038 [FREE Full text] [doi: 10.1093/geroni/igy038] [Medline: 30619948]
3. Reitan RM, Wolfson D. The Halstead-Reitan neuropsychological test battery and aging. The title of the journal, book, proceedings, or website. Value updated based off doi data (Was: Clin Gerontol) 2008 Oct 25;5(1-2):39-61. [doi: 10.1300/j018v05n01_03]
4. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr 01;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
5. Staffaroni AM, Tsoy E, Taylor J, Boxer AL, Possin KL. Digital cognitive assessments for dementia: digital assessments may enhance the efficiency of evaluations in neurology and other clinics. Pract Neurol (Fort Wash Pa) 2020;2020:24-45 [FREE Full text] [Medline: 33927583]
6. Bomyea JA, Parrish EM, Paolillo EW, Filip TF, Eyler LT, Depp CA, et al. Relationships between daily mood states and real-time cognitive performance in individuals with bipolar disorder and healthy comparators: a remote ambulatory assessment study. J Clin Exp Neuropsychol 2021 Oct 08;43(8):813-824 [FREE Full text] [doi: 10.1080/13803395.2021.1975656] [Medline: 34493155]
7. Staffaroni AM, Clark AL, Taylor JC, Heuer HW, Sanderson-Cimino M, Wise AB, ALLFTD Consortium. Reliability and validity of smartphone cognitive testing for frontotemporal lobar degeneration. JAMA Netw Open 2024 Apr 01;7(4):e244266 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.4266] [Medline: 38558141]
8. Allard M, Husky M, Catheline G, Pelletier A, Dilharreguy B, Amieva H, et al. Mobile technologies in the early detection of cognitive decline. PLoS One 2014 Dec 23;9(12):e112197 [FREE Full text] [doi: 10.1371/journal.pone.0112197] [Medline: 25536290]
9. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment 2018 Jan 15;25(1):14-30 [FREE Full text] [doi: 10.1177/1073191116643164] [Medline: 27084835]
10. Papp KV, Jutten RJ, Soberanes D, Weizenbaum E, Hsieh S, Molinare C, et al. Early detection of amyloid-related changes in memory among cognitively unimpaired older adults with daily digital testing. Ann Neurol 2024 Mar 19;95(3):507-517. [doi: 10.1002/ana.26833] [Medline: 37991080]
11. Singh S, Strong R, Xu I, Fonseca LM, Hawks Z, Grinspoon E, et al. Ecological momentary assessment of cognition in clinical and community samples: reliability and validity study. J Med Internet Res 2023 Jun 02;25:e45028 [FREE Full text] [doi: 10.2196/45028] [Medline: 37266996]
12. Thompson LI, Harrington KD, Roque N, Strenger J, Correia S, Jones RN, et al. A highly feasible, reliable, and fully remote protocol for mobile app-based cognitive assessment in cognitively healthy older adults. Alzheimers Dement (Amst) 2022 Apr 05;14(1):e12283 [FREE Full text] [doi: 10.1002/dad2.12283] [Medline: 35415201]
13. Nicosia J, Aschenbrenner AJ, Balota DA, Sliwinski MJ, Tahan M, Adams S, et al. Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. J Int Neuropsychol Soc 2022 Sep 05;29(5):459-471. [doi: 10.1017/s135561772200042x]
14. Weizenbaum EL, Fulford D, Torous J, Pinsky E, Kolachalama VB, Cronin-Golomb A. Smartphone-based neuropsychological assessment in Parkinson's disease: feasibility, validity, and contextually driven variability in cognition. J Int Neuropsychol Soc 2022 Apr 17;28(4):401-413 [FREE Full text] [doi: 10.1017/S1355617721000503] [Medline: 33998438]
15. Lipsmeier F, Taylor KI, Postuma RB, Volkova-Volkmar E, Kilchenmann T, Mollenhauer B, et al. Reliability and validity of the Roche PD mobile application for remote monitoring of early Parkinson's disease. Sci Rep 2022 Jul 15;12(1):12081 [FREE Full text] [doi: 10.1038/s41598-022-15874-4] [Medline: 35840753]
16. Aschenbrenner AJ, Hassenstab J, Morris JC, Cruchaga C, Jackson JJ. Relationships between hourly cognitive variability and risk of Alzheimer's disease revealed with mixed-effects location scale models. Neuropsychology 2024 Jan;38(1):69-80. [doi: 10.1037/neu0000905] [Medline: 37079810]
17. Parrish EM, Kamarsu S, Harvey PD, Pinkham A, Depp CA, Moore RC. Remote ecological momentary testing of learning and memory in adults with serious mental illness. Schizophr Bull 2021 Apr 29;47(3):740-750 [FREE Full text] [doi: 10.1093/schbul/sbaa172] [Medline: 33219382]
18. Depp CA, Kamarsu S, Filip TF, Parrish EM, Harvey PD, Granholm EL, et al. Ecological momentary facial emotion recognition in psychotic disorders. Psychol Med 2022 Oct 12;52(13):2531-2539 [FREE Full text] [doi: 10.1017/S0033291720004419] [Medline: 33431072]
19. Moore RC, Parrish EM, Van Patten R, Paolillo E, Filip TF, Bomyea J, et al. Initial psychometric properties of 7 NeuroUX remote ecological momentary cognitive tests among people with bipolar disorder: validation study. J Med Internet Res 2022 Jul 29;24(7):e36665 [FREE Full text] [doi: 10.2196/36665] [Medline: 35904876]
20. Moore RC, Ackerman RA, Russell MT, Campbell LM, Depp CA, Harvey PD, et al. Feasibility and validity of ecological momentary cognitive testing among older adults with mild cognitive impairment. Front Digit Health 2022 Aug 5;4:946685 [FREE Full text] [doi: 10.3389/fdgth.2022.946685] [Medline: 35990012]
21. Henneghan AM, Van Dyk KM, Ackerman RA, Paolillo EW, Moore RC. Assessing cancer-related cognitive function in the context of everyday life using ecological mobile cognitive testing: a protocol for a prospective quantitative study. Digit Health 2023 Aug 13;9:20552076231194944 [FREE Full text] [doi: 10.1177/20552076231194944] [Medline: 37588154]

XSL•FO
RenderX

22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
23. Holm SP, Wolfer AM, Pointeau GH, Lipsmeier F, Lindemann M. Practice effects in performance outcome measures in patients living with neurologic disorders - a systematic review. Heliyon 2022 Aug;8(8):e10259 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10259] [Medline: 36082322]
24. Calamia M, Markon K, Tranel D. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol 2012 May;26(4):543-570. [doi: 10.1080/13854046.2012.680913] [Medline: 22540222]
25. Jutten RJ, Rentz DM, Fu JF, Mayblyum DV, Amariglio RE, Buckley RF, et al. Monthly at-home computerized cognitive testing to detect diminished practice effects in preclinical Alzheimer's disease. Front Aging Neurosci 2021 Jan 13;13:800126 [FREE Full text] [doi: 10.3389/fnagi.2021.800126] [Medline: 35095476]
26. Sanderson-Cimino M, Chen R, Tu XM, Elman JA, Jak AJ, Kremen WS. Misinterpreting cognitive change over multiple timepoints: when practice effects meet age-related decline. Neuropsychology 2023 Jul;37(5):568-581 [FREE Full text] [doi: 10.1037/neu0000903] [Medline: 37079809]
27. Der G, Deary IJ. Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. Psychol Aging 2006 Mar;21(1):62-73. [doi: 10.1037/0882-7974.21.1.62] [Medline: 16594792]
28. Salthouse TA. Selectivity of attrition in longitudinal studies of cognitive functioning. J Gerontol B Psychol Sci Soc Sci 2014 Jul 02;69(4):567-574 [FREE Full text] [doi: 10.1093/geronb/gbt046] [Medline: 23733858]
29. Young AF, Powers JR, Bell SL. Attrition in longitudinal studies: who do you lose? Aust N Z J Public Health 2006 Aug;30(4):353-361 [FREE Full text] [doi: 10.1111/j.1467-842x.2006.tb00849.x] [Medline: 16956166]
30. Salthouse TA. Trajectories of normal cognitive aging. Psychol Aging 2019 Feb;34(1):17-24 [FREE Full text] [doi: 10.1037/pag0000288] [Medline: 30211596]
31. Haynes BI, Bauermeister S, Bunce D. A systematic review of longitudinal associations between reaction time intraindividual variability and age-related cognitive decline or impairment, dementia, and mortality. J Int Neuropsychol Soc 2017 May 02;23(5):431-445 [FREE Full text] [doi: 10.1017/S1355617717000236] [Medline: 28462758]
32. Bielak AA, Hultsch DF, Strauss E, Macdonald SW, Hunter MA. Intraindividual variability in reaction time predicts cognitive outcomes 5 years later. Neuropsychology 2010 Nov;24(6):731-741. [doi: 10.1037/a0019802] [Medline: 20853957]
33. Golomb BA, Chan VT, Evans MA, Koperski S, White HL, Criqui MH. The older the better: are elderly study participants more non-representative? A cross-sectional analysis of clinical trial and observational study samples. BMJ Open 2012 Dec 14;2(6):e000833 [FREE Full text] [doi: 10.1136/bmjopen-2012-000833] [Medline: 23242479]
34. Olson R, Parkinson M, McKenzie M. Selection bias introduced by neuropsychological assessments. Can J Neurol Sci 2010 Mar 02;37(2):264-268. [doi: 10.1017/s0317167100010039] [Medline: 20437940]
35. Manly JJ, Jacobs DM, Sano M, Bell K, Merchant CA, Small SA, et al. Cognitive test performance among nondemented elderly African Americans and whites. Neurology 1998 May;50(5):1238-1245. [doi: 10.1212/wnl.50.5.1238] [Medline: 9595969]
36. Lippa SM. Performance validity testing in neuropsychology: a clinical guide, critical review, and update on a rapidly evolving literature. Clin Neuropsychol 2018 Apr 28;32(3):391-421. [doi: 10.1080/13854046.2017.1406146] [Medline: 29182444]
37. Greher MR, Wodushek TR. Performance validity testing in neuropsychology: scientific basis and clinical application-a brief review. J Psychiatr Pract 2017 Mar;23(2):134-140. [doi: 10.1097/PRA.0000000000000218] [Medline: 28291039]

Abbreviations

AD: Alzheimer disease
EMA: ecological momentary assessment
HIPAA: Health Insurance Portability and Accountability Act
ICC: intraclass correlation coefficient

Edited by A Mavragani; submitted 18.08.23; peer-reviewed by E Weizenbaum, P Harvey, J Hakun; comments to author 16.04.24; revised version received 17.05.24; accepted 18.10.24; published 25.11.24.

Please cite as:
Paolillo EW, Bomyea J, Depp CA, Henneghan AM, Raj A, Moore RC
Characterizing Performance on a Suite of English-Language NeuroUX Mobile Cognitive Tests in a US Adult Sample: Ecological Momentary Cognitive Testing Study
J Med Internet Res 2024;26:e51978
URL: https://www.jmir.org/2024/1/e51978
doi:10.2196/51978
PMID:

XSL•FO
RenderX

©Emily W Paolillo, Jessica Bomyea, Colin A Depp, Ashley M Henneghan, Anunay Raj, Raeanne C Moore. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Recruitment for Voluntary Video and Mobile HIV Testing on Social Media Platforms During the COVID-19 Pandemic: Cross-Sectional Study

Piao-Yi Chiou[1,2,3,4], PhD; Wei-Wen Tsao[1,3], BA; Chia-Lin Li[3,5], PhD; Jheng-Min Yu[6], MD; Wen-Han Su[1], BA; Zhi-Hua Liu[2], BSN; Cheng-Ru He[3,7], MSN; Yu-Chun Chang[1], MSN; Yi-Hsuan Tsai[8], MD

[1]School of Nursing, National Taiwan University College of Medicine, Taipei, Taiwan
[2]Department of Nursing, National Taiwan University Hospital, Taipei, Taiwan
[3]Taiwan AIDS Nurse Association, Taipei, Taiwan
[4]Taiwan Lourdes Association, Taipei, Taiwan
[5]Center for Neuropsychiatric Research, National Health Research Institutes, Taipei, Taiwan
[6]Institute of Health Policy and Management, College of Public Health, National Taiwan University, Taipei, Taiwan
[7]Department of Cardiovascular surgery, National Cheng Kung University Hospital, Tainan, Taiwan
[8]Department of traditional Chinese medicine, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan

Corresponding Author:
Piao-Yi Chiou, PhD
School of Nursing, National Taiwan University College of Medicine
No. 2-1, Xuzhou Rd., Zhongzheng Dist.,
Taipei, 10051
Taiwan
Phone: 886 0935883722 ext 288427
Fax: 886 2 23219913
Email: purechiou2@gmail.com

Abstract

Background: The COVID-19 pandemic prompted social distancing policies and caused misinformation that hindered in-person HIV screening for high-risk groups. Social media platforms provide additional options for voluntary counseling and testing (VCT) for HIV, overcoming these limitations. However, there is a lack of data on HIV testing recruitment through social media platforms and its outcomes during the pandemic.

Objective: This study aimed to measure the rate of face-to-face mobile and video VCT conducted after recruitment through social media platforms and friend referrals during the pandemic and compare the geographic distribution, risk feature targeting, testing outcome, and cost between the 2 models.

Methods: Data were collected from March 3 to December 31, 2021, during the COVID-19 outbreak in Taiwan. Participants engaging in unprotected sex were recruited. After one-on-one message discussions through the platforms, the well-trained research assistants provided mobile or video VCT based on the participants' availability. Primary outcomes were completion rate, testing results, and CD4 count. Secondary outcomes included demographic and HIV risk-taking and protective features from a questionnaire. Selection bias was controlled by adjusting for the testing site (Taipei vs non-Taipei) using univariable multinomial logistic regression.

Results: This study gathered 5142 responses on the social media platforms, recruiting 1187 participants. Video VCT had a completion rate of 31.8% (207/651), higher than mobile VCT's 21.8% (980/4491). Both rates were higher than those before the COVID-19 pandemic. Recruitment through friend referrals, instant messaging apps (eg, Line [LY Corporation]), and geosocial dating apps (eg, Hornet [Queer Networks Inc], Grindr [Grindr LLC], and Gsland [Tien-Hao Tsai]) resulted in higher acceptance and completion rates than social networks (eg, Facebook [Meta], X [formerly Twitter], and Instagram [Meta]). Mobile VCT had higher recruitment among urban residents and screening density, while video VCT reached a broader geographic area. The mobile group was more likely to have had more than 10 sexual partners (odds ratio [OR] 1.92, 95% CI 1.05-3.50; P=.03), history of sex work (OR 4.19, 95% CI 1.68-10.43; P=.002), and sexually transmitted diseases (OR 2.23, 95% CI 1.18-4.23; P=.01) within the past 3 months. The video group was more likely to meet sexual partners through social media. The HIV-positive rate in the mobile

XSL•FO
RenderX

group was 0.7% (7/973) with an average CD4 count of 460/μL, while in the video group, it was 1% (2/205) with an average CD4 count of 347/μL, indicating a later diagnosis. Both positivity rates were higher than those before the COVID-19 pandemic, with no significant difference between the groups. The video group cost US $54.68 per participant, slightly higher than the US $50.36 for the mobile group.

Conclusions: Recruiting through social media platforms that facilitate one-on-one message discussions can effectively target high-risk groups for mobile and video VCT. This approach should be integrated into the current screening model to enhance HIV case finding.

(J Med Internet Res 2024;26:e54420) doi:10.2196/54420

KEYWORDS

COVID-19; HIV testing; mobile health; risk-taking behavior; social media; video; mobile phone

Introduction

Although the number of confirmed cases and deaths associated with COVID-19 has gradually decreased, emerging infectious diseases continue to be a global public health concern [1-3]. In response to the COVID-19 outbreak, countries worldwide adopted prevention and control measures recommended by the World Health Organization to reduce virus transmission [4]. These measures included social and physical distancing mandates and implementing visitor control at hospitals and health care facilities [5], which directly or indirectly hindered the accessibility of face-to-face voluntary counseling and testing (VCT) services for HIV [6].

Studies have found that high-risk sexual behaviors and recreational drug use persisted or increased during the COVID-19 pandemic [7,8]. Despite lockdown policies, approximately 76% of the HIV target population, particularly men who have sex with men (MSM), continued sexual activities and 19% used recreational drugs owing to stress and sexual desires [9-12]. However, capacity for HIV testing was estimated to have dropped from 26% to 85% in Africa, America, Asia, and Europe; this reflected the situation pertaining to key populations who lacked access to VCT services during the pandemic [8,13-16]. Face-to-face mobile VCT distributes recruitment messages through social media platforms and offers flexibility in terms of testing locations and schedules for specific groups [17]. However, despite its high accessibility and convenience, service volume drastically decreased during the COVID-19 pandemic in Taiwan [18]. In contrast, home-based screening allowed participants to request a self-testing kit through postal services, with additional real-time counseling accessible internet, eliminating the need for face-to-face contact [19-23]. Therefore, implementing VCT for HIV through the video model provides an opportunity to adhere to social distancing guidelines and prioritize the safety of frontline screeners [14,24-28].

Social media platforms, including mobile instant messaging apps, geosocial network (GSN) apps using GPS, and social networks, serve as interactive digital channels. Considering the high ownership rate of smartphones in Taiwan [29], social media platforms, which have an expanding user base and lack face-to-face limitations, have emerged as effective channels for recruiting high-risk and hard-to-reach groups and maintaining uninterrupted VCT services [30-32]. The statistics on adult social media use in Taiwan indicate that the Line (LY Corporation) messaging app has the highest market share among instant messaging apps, accounting for a weighted value of 77.6% (2153/1670) [33]. The most popular social networks are Facebook (Meta), Instagram (Meta), and X (formerly Twitter), which cover 78.1% (16.95 million/21.71 million), 52.3% (11.35 million/21.71 million), and 21.4% (4.64 million/21.71 million) of the local internet user base, respectively [34,35]. Hornet (Queer Networks Inc), Grindr (Grindr LLC), and Gsland (Tien-Hao Tsai) are the most widely used GSN apps among lesbian, gay, bisexual, transgender, and queer (LGBTQ) users in Taiwan [36,37]. These platforms are popular for connecting with potential sexual partners among high-risk groups for HIV [38-40]. However, the use of social media platforms can vary due to discrepancies in internet infrastructure across different geographic areas. In rural areas of Taiwan, the personal internet access rate for individuals aged 12 years and older stands at 75.8%, which is 10.4% lower than the national average of 86.2% [41]. In addition, the COVID-19 infodemic, which refers to the spread of false and misleading information through social media platforms that causes confusion, risky behavior, and distrust toward health officials, might interfere with the recruitment of high-risk groups for HIV testing [42,43]. Existing data on participant recruitment and the outcomes of recruiting for HIV testing through social media platforms during the COVID-19 pandemic are scarce.

Therefore, this study aimed to evaluate the rate of face-to-face mobile and video VCT for HIV conducted after recruitment through social media platforms during the COVID-19 pandemic. It also examined the geographic distribution, targeted risk behavior features, testing outcomes, and CD4 counts to compare the health status of HIV-positive participants, as well as the cost between the groups tested using these 2 models.

Methods

Study Design

Data were collected from March 3 to December 31, 2021, during the COVID-19 outbreak in Taiwan. This study used a cross-sectional research design. Participants were selected using convenience and snowball sampling and recruited through popular social media platforms frequented by high-risk groups for HIV, based on data from the previous year (2020), as indicated by program CW108056. Participants were recruited through the Line messaging app, GSN dating apps, a fan page on Facebook, Instagram, and X. In total, 2 testing models,

XSL•FO
RenderX

mobile and video VCT, were conducted by proficient research assistants according to the participants' availability.

Participants

The inclusion criteria were being aged 20 years or older, having the literacy skills to understand the research consent form and related questionnaires, and self-reporting engagement in unprotected sexual intercourse before a 3-month window period. The exclusion criteria were reporting no involvement in any unprotected sexual intercourse and having a previous HIV-positive identification.

Procedures and Data Collection

Recruitment

For recruitment through Line, the QR code of the official account was disseminated across various platforms to encourage web-based users to engage in one-on-one discussions with our research assistants.

Recruitment through GSN dating apps was conducted through Hornet, Grindr, and Gsland. User profiles were refreshed every 2 hours during the daytime on weekdays to display our official account with the heading and profile describing HIV testing. Nearby web-based users interested in testing could initiate one-on-one discussions with our research assistants by tapping and sending a message.

A conspicuous headline about HIV testing and posters of the 2 VCT models were presented on our official Facebook fan page to encourage viewers to visit and initiate one-on-one message discussions with our research assistants.

In addition, recruitment was carried out through the Facebook, Instagram, and X accounts of 6 celebrities, each with an average follower count exceeding 10,000. Information regarding the 2 VCT models was posted and discussed at least twice every month during the research period on their pages. The audience members who were interested in HIV testing could link to our Line account.

Moreover, friend referrals were used. Screening information were sourced from individuals who had either seen our recruitment posters or undergone our screening process and then shared it with their friends. Those friends who had viewed our recruitment information actively participated in one-on-one discussions with the research assistants through our official account on any social media platform.

Grouping Condition

The 2 VCT model options were discussed through messages on each platform. Mobile VCT was provided to participants who lived in the Taipei area and were available for face-to-face meetings. Video VCT was provided to participants who lived in the Taipei area but were not available for face-to-face meetings owing to the social and physical distancing mandates of COVID-19 prevention, and for those who lived in remote areas distant from Taipei.

Mobile Voluntary HIV Counseling and Testing Model

Well-trained research assistants delivered the mobile VCT at the time and place designated for the participants. Protective equipment against COVID-19 infection, such as masks, gloves, and goggles, were used by the research assistants when conducting the mobile VCT. The location choices were based on convenience and had compartmentalized seating, such as a convenience store, fast food restaurant, coffee shop, or park. After checking the participants' identity through the messaging record on their mobile phones, participants were asked to complete the questionnaires by following a URL. Pretest counseling was provided and a blood test was performed by a skin prick using a rapid HIV self-testing kit. The results were interpreted immediately upon completion of the test, followed by posttest counseling and resource referrals. In addition, we accompanied the participants who tested positive to AIDS-designated hospitals for confirmatory diagnosis and CD4 testing.

Video Voluntary HIV Counseling and Testing Model

A rapid HIV self-testing kit was promptly and anonymously dispatched to participants' designated addresses using a password-protected system. Subsequently, a research assistant made a video call through the Line app to the participant at the appointed time and place, such as at home, at a coffee shop, or in a private location. Participants' identities were confirmed by verifying the messaging record and box number of the testing kits presented by the screen. They were provided a URL for research consent, where participants can click to agree to join the study and then complete the questionnaire. To maintain privacy, participants could choose only to display their hands-on testing process on the camera screen. Next, participants received pretest counseling before using a rapid HIV self-testing kit to perform the test. They obtained a blood drop by a skin prick and followed instructions provided by the research assistant (Figure 1). The results were displayed on the video screen and confirmed by the research assistant. Posttest counseling and resource referral sections mirrored those in the mobile VCT model.

XSL•FO
RenderX

Figure 1. Demonstration of video voluntary HIV counseling and testing in Taiwan during the COVID-19 pandemic.

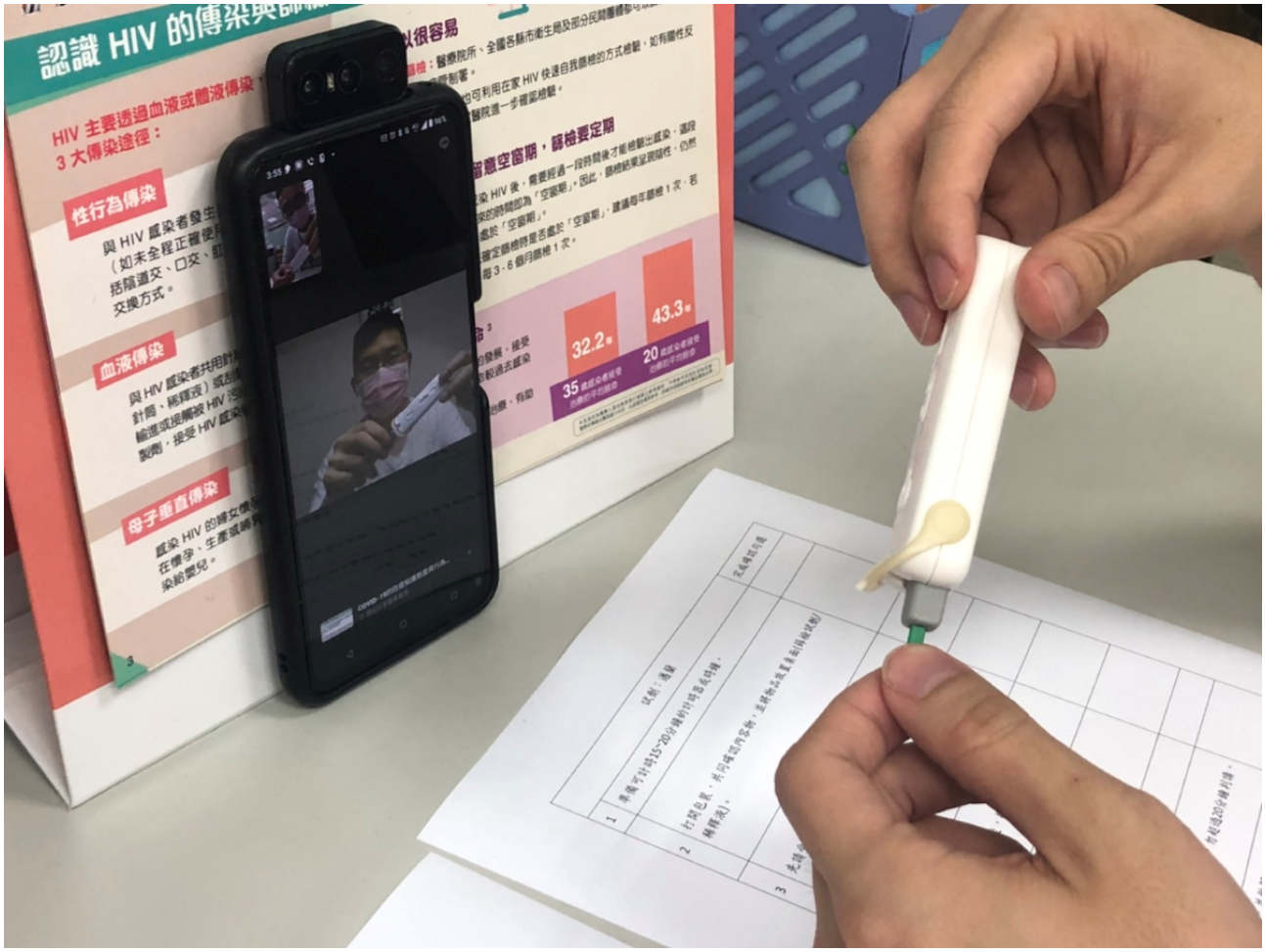

Instrument

Demographic Characteristics

The participants' demographic information, including age, sex, sexual orientation, marital status, education level, and presence of stable income, were collected using a questionnaire. Furthermore, the questionnaire inquired about the source of recruitment information leading to the present test and testing location.

HIV-Related Risk-Taking and Protective Behavior

Based on previous studies in Taiwan [37,44], this study inquired about high-risk and preventive behavior for HIV, such as preferred sexual position, meeting sexual partners through social media, having multiple sexual partners, number of sexual partners in the past 3 months, frequency of condom use during recent sexual intercourse, history of regular HIV testing, testing frequency in the preceding year, engagement in sex work, history of sexually transmitted diseases (STDs), recreational drug use in the past 3 months, and previous use of pre-exposure prophylaxis (PrEP) and postexposure prophylaxis (PEP).

Rapid HIV Self-Testing Kits

This study used 2 rapid HIV self-testing kits, namely SURE CHECK HIV 1/2 Assay (Chembio) and Mylan HIV Self-Test (Mylan). These kits were approved by the Taiwan Food and Drug Administration (procurement identification 1109042028A and 1109042028B). These tests yield results in approximately 15 minutes and have high sensitivity (95.2% to 97%) and specificity (99.6% to 100%).

Cost

This study recorded the consumables and associated administrative expenses used for each VCT session, including test kits, transportation expenses, personnel salaries, internet charges, and the COVID-19 protective equipment for mobile VCT; and test kits, postage, personnel salaries, internet charges, and video devices for video VCT. The total cost was calculated and divided by the average number of participants to determine the average cost per VCT session.

Statistical Analyses

The recruitment and completion rates for both VCT models were calculated as percentages. Geographic coverage was assessed by evaluating the area and participant density, represented as the number of participants per square kilometer. To control for the potential influence of a selection bias, the testing site (Taipei vs non-Taipei area) was adjusted using univariable multinomial logistic regression. The primary outcome comprised HIV testing results, while the secondary outcome included demographic data and data on behavior associated with HIV risk, which were compared between the mobile and video VCT groups.

XSL•FO
RenderX

Ethical Considerations

The study was approved by the research ethics committee of National Taiwan University (202012HM042). All participants provided consent after receiving a comprehensive explanation of the research details through a web page, including its purpose, methods, participant requirements, risks and benefits, privacy measures, voluntary participation, contact information, data policies, and compensation. Data collection and processing were conducted anonymously. The participants' privacy was also protected through the use of masks during both VCT models. Upon completion of all procedures, each participant received an electronic gift certificate valued at US $3.27.

Results

Recruitment and Completion

Table 1 shows the results of recruitment and completion of 2 VCTs through different recruitment sources. This study generated 5142 responses on social media platforms and friend referrals, and 1187 participants were recruited, with an overall completion rate of 23.08%. Among the participants, 980 (21.8%) out of 4491 and 207 (31.8%) out of 651 completed the mobile and video VCT, respectively.

Table 1. Recruitment and completion of the 2 voluntary HIV counseling and testing models for the HIV risk population in Taiwan during the COVID-19 pandemic.

	GSN[a] dating apps	Facebook	X (Twitter)	Instagram	Line app	Friend referral	Total
Mobile VCT[b]							
Responses, n	1126[c]	1098[d]	785[d]	672[d]	468[e]	342[f]	4491
One-on-one message discussion, n/N (%)	587/1126 (52.13)	378/1098 (34.43)	240/785 (30.6)	243/672 (36.2)	391/468 (83.6)	287/342 (83.9)	2126/4491 (47.34)
Acceptance of VCT, n/N (%)	334/587 (56.9)	122/378 (32.3)	76/240 (31.7)	80/243 (32.9)	199/391 (50.9)	169/287 (58.9)	980/2126 (46.1)
Completion of VCT, n/N (%)	334/1126 (29.66)	122/1098 (11.11)	76/785 (9.7)	80/672 (11.9)	199/468 (42.5)	169/342 (49.4)	980/4491 (21.82)
Video VCT							
Responses, n	177[c]	124[d]	102[d]	103[d]	89[e]	56[f]	651
One-on-one message discussion, n/N (%)	93/177 (52.5)	56/124 (45.2)	44/102 (43.1)	44/103 (42.7)	71/89 (79)	48/56 (85)	356/651 (54.7)
Acceptance of VCT, n/N (%)	64/93 (68)	25/56 (44)	18/44 (40)	18/44 (40)	46/71 (64)	36/48 (75)	207/356 (58.2)
Completion of VCT, n/N (%)	64/177 (36.2)	25/124 (20.2)	18/102 (17.6)	18/103 (17.5)	46/89 (51)	36/56 (64)	207/651 (31.8)

[a]GSN: geosocial network.

[b]VCT: voluntary counseling and testing.

[c]Number of users who tapped on our official account on Hornet, Grindr, and Gsland.

[d]Number of viewers on the recruitment information page of internet celebrities' pages on Facebook, X (Twitter), and Instagram, as well as our Facebook fan page.

[e]Number of internet users who actively searched for and browsed the internet to find our official Line account and QR code and requested to be added as friends through the Line app.

[f]Referred by friends who have either read our recruitment information or used our VCT service before.

Responses for mobile VCT comprised 1126 on GSN dating apps, 1098 on Facebook, 785 on X, 672 on Instagram, and 468 on Line, as well as 342 responses through friend referrals. The completion rate was 21.82% (980/4491). Among the recruitment sources, friend referral had the highest completion rate of mobile VCT (49.4%, 169/342), followed by Line (42.5%, 199/468) and GSN dating apps (29.66%, 334/1126). The main reasons for not completing mobile VCT after one-on-one discussion were not meeting the inclusion criteria (15.01%, 172/1146), no response (36.47%, 418/1146), concern about COVID-19 infection when going outside (16.58%, 190/1146), preferring to wait beyond the window period of HIV infection in anticipation of obtaining more accurate results (12.65%, 145/1146), feeling unprepared for the test result (8.29%, 95/1146), and missing the appointment without an explanation (3.32%, 38/1146).

The overall completion rate of video VCT was 31.8% (207/651). Friend referral exhibited the highest completion rate (64%, 36/56), followed by Line (52%, 46/89) and GSN dating apps (36.2%, 64/177). The main reasons for not completing video VCT after one-on-one discussion were not meeting the inclusion criteria (13.4%, 20/149), no response (33.6%, 50/149), not trusting self-testing and wishing to have blood collected by a well-trained person (14.8%, 22/149), fear of drawing blood

XSL•FO
RenderX

(12.8%, 19/149), feeling unprepared for the test result (8.1%, 12/149), concern that receiving the parcel may expose their privacy (7.4%, 11/149), and missing the appointment without an explanation (2.7%, 4/149).

Geographic Distribution

Among the participants in the mobile and video VCT groups, 100% (980/980) and 75.4% (156/207), respectively, lived in Taipei, and 24.6% (51/207) of the participants in the video VCT group lived in remote areas outside Taipei.

The total area covered by the mobile VCT was 390.2 km^2, with a participant density of 2.51 (980/390.2 km^2). This covered 1.08% of Taiwan's total land area (36,197.0 km^2). The video VCT covered 18 cities and counties, with 24.6% (51/207) of the participants residing in 16 cities and counties far from Taipei, where the mobile VCT did not reach. The coverage area of the video VCT was 26,391.04 km^2, with a participant density of 0.008 (207/26391.04 km^2). This covered 72.91% of Taiwan's total land area.

The participant density for mobile VCT was 313.8 (2.51/0.008 km^2) times higher compared to video VCT. However, the geographic area covered by the video VCT was 67.6 (26391.04/390.2 km^2) times larger than that of the mobile VCT.

Demographic Characteristics

The mean age of all participants was 30.59 (SD 6.7) years. The majority were men (1084/1187, 91.3%), identified as homosexual (855/1187, 72%), and reported being unmarried (1158/1187, 97.6%). In addition, most participants held a college or university-level education (837/1187, 70.5%) and had a stable income (1023/1187, 86.2%). The analysis of demographic characteristics indicated homogeneity between the 2 groups (Table 2).

Table 2. The comparison of demographic data between the 2 VCT[a] models in Taiwan during the COVID-19 pandemic (adjusted for the testing site in the Taipei area and non-Taipei area).

Variable	Total (N=1187), n (%)	Mobile VCT model (n=980), n (%)	Video VCT model (n=207), n (%)	OR[b] (95% CI)	*P* value
Age group (year)					
<30 (reference)	558 (47)	462 (47.1)	96 (46.4)	__[c]	—
≥30	629 (53)	518 (52.9)	111 (53.6)	0.97 (0.72-1.31)	.84
Sex					
Female (reference)	103 (8.7)	87 (8.9)	16 (7.7)	—	—
Male	1084 (91.3)	893 (91.1)	191 (92.3)	0.86 (0.49-1.50)	.59
Sexual orientation					
Others[d] (reference)	26 (2.2)	24 (2.4)	2 (1)	—	—
Homosexual	855 (72)	700 (71.4)	155 (74.9)	0.38 (0.09-1.61)	.19
Heterosexual	126 (10.6)	106 (10.8)	20 (9.7)	0.44 (0.10-2.02)	.29
Bisexual	180 (15.2)	150 (15.3)	30 (14.5)	0.42 (0.09-1.86)	.25
Marital status					
Yes (reference)	29 (2.4)	23 (2.3)	6 (2.9)	—	—
No	1158 (97.6)	957 (97.7)	201 (97.1)	1.24 (0.50-3.09)	.64
Education					
High school or less (reference)	105 (8.8)	89 (9.1)	16 (7.7)	—	—
University or college	837 (70.5)	687 (70.1)	150 (72.5)	0.82 (0.47-1.44)	.50
Above university	245 (20.6)	204 (20.8)	41 (19.8)	0.89 (0.45-1.68)	.73
Stable income					
No (reference)	164 (13.8)	137 (14)	27 (13)	—	—
Yes	1023 (86.2)	843 (86)	180 (87)	0.92 (0.59-1.44)	.72

[a]VCT: voluntary counseling and testing.

[b]OR: odds ratio.

[c]Not applicable.

[d]Included pansexuality and uncertainty.

XSL•FO
RenderX

Risk-Taking and Protective Behavior and Test Results

The majority of participants reported having a preference for both top and bottom sexual positions (455/1187, 38.3%), meeting sexual partners through social media (1003/1187, 84.5%), having 2-3 sexual partners (461/1187, 54.9%), inconsistently using condoms in the past 3 months (952/1187, 80.2%), and not having an HIV test regularly within the previous year (678/1187, 57.1%). Some participants reported a history of sex work, STDs, and recreational drug use in the past 3 months. Most participants had not used PrEP or PEP (950/1187, 80%; Table 3).

XSL•FO
RenderX

Table 3. The comparison of risk-taking features between the 2 voluntary HIV counseling and testing models in Taiwan during the COVID-19 pandemic (adjusted for the testing site in the Taipei area and non-Taipei area).

Variable	Total (N=1187), n (%)	Mobile VCT[a] model (n=980), n (%)	Video VCT model (n=207), n (%)	OR[b] (95% CI)	*P* value
Preferred sexual position					
Side (reference)	27 (2.3)	26 (2.7)	1 (0.5)	__[c]	—
Top	386 (32.5)	321 (32.8)	65 (31.4)	0.19 (0.03-1.42)	.11
Bottom	319 (26.9)	266 (27.1)	53 (25.6)	0.91 (0.02-1.45)	.11
Both top and bottom	455 (38.3)	367 (37.4)	88 (42.5)	0.16 (0.02-1.20)	.07
Meeting sex partners through social media					
No (reference)	184 (15.5)	166 (16.9)	18 (8.7)	—	—
Yes	1003 (84.5)	814 (83.1)	189 (91.3)	0.47 (0.28-0.78)	.004
Multiple sex partners[d]					
No (reference)	348 (29.3)	279 (28.5)	69 (33.3)	—	—
Yes	839 (70.7)	701 (71.5)	138 (66.7)	1.26 (0.91-1.73)	.16
Number of multiple sexual partners[d] (n=839)					
2-3 (reference)	461 (54.9)	379 (54.1)	82 (59.4)	—	—
4-9	240 (28.6)	198 (28.2)	42 (30.4)	1.02 (0.68-1.54)	.93
>10	138 (16.4)	124 (17.7)	14 (10.1)	1.92 (1.05-3.50)	.03
Frequency of condom usage[d]					
Always (reference)	235 (19.8)	185 (18.9)	50 (24.2)	—	—
Not always	952 (80.2)	795 (81.1)	157 (75.8)	1.01 (0.74-1.36)	.97
Regularly HIV testing[e]					
No (reference)	678 (57.1)	560 (57.1)	118 (57)	—	—
Yes	509 (42.9)	420 (42.9)	89 (43)	0.99 (0.74-1.35)	.97
Testing frequency[e] (n=509)					
Per half month to once a year (reference)	230 (45.2)	185 (44)	45 (50.6)	—	—
Per 3 months	279 (54.8)	235 (56)	44 (49.4)	1.29 (0.82-2.05)	.26
Experience of sex work					
No (reference)	1090 (91.8)	888 (90.6)	202 (97.6)	—	—
Yes	97 (8.2)	92 (9.4)	5 (2.4)	4.19 (1.68-10.43)	.002
Experience of STDs[d,f]					
No (reference)	1067 (89.9)	871 (88.9)	196 (94.7)	—	—
Yes	120 (10.1)	109 (11.1)	11 (5.3)	2.23 (1.18-4.23)	.01
Experiences of recreational drug use[d]					
No (reference)	1046 (88.1)	856 (87.3)	190 (91.8)	—	—
Yes	141 (11.9)	124 (12.7)	17 (8.2)	1.62 (0.95-2.75)	.08
Adopting PrEP[g] or PEP[h]					
No (reference)	950 (80)	791 (80.7)	159 (76.8)	—	—
Yes	237 (20)	189 (19.3)	48 (23.2)	0.79 (0.55-1.13)	.20
HIV testing result					
Negative (reference)	1178 (99.2)	973 (99.3)	205 (99)	—	—

XSL•FO
RenderX

Variable	Total (N=1187), n (%)	Mobile VCT[a] model (n=980), n (%)	Video VCT model (n=207), n (%)	OR[b] (95% CI)	*P* value
positive	9 (0.8)	7 (0.7)	2 (1)	0.74 (0.15-3.58)	.71

[a]VCT: voluntary counseling and testing.

[b]OR: odds ratio.

[c]Not applicable.

[d]Within the past 3 months.

[e]Within the previous year.

[f]STDs: sexual transmitted diseases.

[g]PrEP: pre-exposure prophylaxis.

[h]PEP: postexposure prophylaxis.

Participants in the video VCT group were more likely to meet sexual partners through social media than those in the mobile VCT group. Participants in the mobile VCT group were more likely to have had more than 10 sexual partners, a history of sex work, and STDs in the past 3 months than those in the video VCT group. No statistically significant difference was observed in HIV-positive rates between the 2 groups. The average CD4 count of HIV-positive participants was 347/µL in the video VCT model (n=2), which was lower than that in the mobile VCT model (460/µL; n=7).

Testing Costs

The average cost of the video VCT was approximately US $54.68 per participant, which was slightly higher than that of the mobile VCT (US $50.36).

Discussion

Principal Findings

This study recruited participants for 2 VCT models through social media platforms and friend referrals during the COVID-19 pandemic and evaluated the test results. The results demonstrated that recruiting for VCT through friend referral and social media platforms was feasible for reaching people with various HIV-related, risk-taking behaviors and HIV case findings while complying with COVID-19 prevention regulations. Furthermore, friend referral through Line had the best success rate for both VCT models. Our research findings indicate that this approach is highly valuable beyond the context of a pandemic and should be integrated into the current screening model to enhance HIV case findings.

The sample size was asymmetric for the 2 VCT models, with 4.7 times more participants opting for mobile than video VCT. Furthermore, most of the participants lived in Taipei. This may be because Taipei and New Taipei have the highest prevalence of HIV in Taiwan [45], creating a higher demand for screening of the HIV risk population in these cities than in other cities and remote areas.

The completion rate of mobile VCT in this study was higher than that reported in 2018 (20.71%), before the COVID-19 pandemic [17]. This increase may be attributed to the requirement to show identification to enter medical facilities for HIV testing during the COVID-19 pandemic, which discouraged individuals who preferred anonymous screening and led them to turn to mobile VCT. Mobile VCT could be performed anonymously outdoors, reducing the risk of COVID-19 transmission associated with entering crowded areas. Furthermore, eligible participants were informed during one-on-one message discussions that the screening personnel tested negative for COVID-19 and did not have any relevant symptoms; protective gear would be used while conducting HIV testing; and they could meet in an open and not crowded location, such as a park. In addition, the discussion dispelled common rumors regarding COVID-19, such as the possibility of COVID-19 transmission through mobile phones and that anti-HIV medication can prevent COVID-19 infection [46]. The mobile VCT was free and anonymous, provided timely and appropriate service, and disseminated accurate COVID-19 prevention information before the test; therefore, it promoted acceptance.

The video VCT model had an overall complete rate of 31.8%, which was higher than that of the mobile VCT. Previous studies before the COVID-19 pandemic recruited participants through a public website and Facebook advertisements and mailed at-home, self-testing kits for HIV and STDs to the participants; these studies reported return rates of 27.2% and 43.7%, respectively [23,44]. The present results for video VCT were consistent with these findings. People may have become aware of the limited HIV testing resources owing to social and physical distancing mandates and visitor control in hospital and health care facilities. In addition, some individuals resided out of range of mobile VCT. Therefore, participants who met the criteria for video VCT had high motivation to accept it. The advantages of video VCT were not only the ability to avoid personal information exposure but also the ability to receive individualized and visible guidance and immediate emotional support in a stigma-free environment from a well-trained research assistant. This result echoes the advantages of video intervention, including the lack of geographical restrictions for delivery, convenience, saving travel time, and reducing the risk of COVID-19 exposure and transmission [20,24,27,47-49]. However, the disadvantage is the decreased results of video interactions if the network speed is unstable or screen quality is poor. Furthermore, this study encountered 5 testing errors, including 3 unsuccessful blood collections by the participants and 2 incorrect results. The participants may feel distressed while waiting for new testing packages.

XSL•FO
RenderX

The completion rates for the 2 VCT models following recruitment through social media platforms, including X, Facebook, and Instagram, were lower than those of friend referral, Line, and GSN apps. Previous studies have found a considerable amount of misleading information spread through social networks during the COVID-19 pandemic, leading to unhealthy behaviors and psychological stress [50,51]. This could potentially divert users' attention away from HIV screening information. Unlike social media platforms, Line and GSN app interfaces do not contain excessive shared photos and posts on the COVID-19 pandemic and daily life information to divert users' attention. This may make it easier to enter one-on-one message discussion and complete HIV screening. Furthermore, referrals from friends, particularly those who had used VCT services and could attest to their privacy and credibility, through Line, had the highest completion rate among the channels used in this study [52]. Compared with urban areas, people in rural areas have limited HIV testing resources and experience more difficulty accessing testing owing to transportation to the screening location being inconvenient and distrust toward health care providers [53,54]. In this study, video VCT covered a larger geographic area than mobile VCT and expanded to remote cities and counties. Sharing of recruitment messages on social media platforms by internet celebrities and friend referral play a critical role in disseminating trustable information to areas with limited screening resources and offering additional testing options [17,55,56].

Participants recruited through the 2 models exhibited differences in risk-taking behaviors. The video VCT model required 4-7 days to complete, including distributing the screening kit and conducting testing. Conversely, the mobile VCT provided real-time service within 1-2 days of initial consultation and flexible test time and location. Individuals with a high frequency of risk-taking behaviors and immediate demand for HIV testing [57,58] were more likely to use the mobile VCT model. A previous study revealed a significantly positive correlation between meeting sexual partners through social media and seeking online sexual health–related information [59]. For internet users, the operation of video VCT is simple and familiar [60]. This could explain our results that participants who met sexual partners through social media may have more opportunities to receive recruitment information and use video VCT.

The newly identified HIV-positive rate was 1% in the video VCT group and 0.7% in the mobile VCT group, surpassing the 0.5% rate recorded for mobile VCT in 2020 (program CW108056), before the COVID-19 pandemic. These findings align with previous research findings indicating an elevated HIV-positive rate in MSM during the COVID-19 pandemic [7,8]. This may be because participants in both VCT groups were mostly MSM and their risk-taking behaviors, which included mainly having multiple sexual partners, not always using condoms, and recreationally using drugs, persisted and increased during the COVID-19 pandemic [12,13]. The average CD4 count of participants who are HIV positive in the video VCT group was lower than that in the mobile VCT group and reached the standard of a late diagnosis (<349/μL) [61]. A previous study demonstrated that the late diagnosis rate of HIV increased during the COVID-19 pandemic [62]. Further research using more extensive datasets is required to determine whether video VCT improves late diagnosis of HIV, particularly in remote areas.

The average costs of the 2 models in this study were significantly higher than the cost of self-screening using a kit purchased through a vending machine (US $33.81), without pre- and posttest counseling [63]. The higher costs in this study were mainly due to personnel costs. Considering the importance of one-on-one consultations, all positive cases identified through initial screening in this study were successfully referred to hospital case managers for a confirmed diagnosis and further treatment. Future studies should investigate the effectiveness of personal counseling on participants, focus on enhancing HIV prevention knowledge and behaviors, emotional regulation, reduced COVID-19 infections, and evaluating service aspects such as usability, performance, and satisfaction.

Limitations

This study had several limitations. First, the use of a cross-sectional design precluded identification of causal relationships between the independent and dependent variables. Second, as this study used a questionnaire survey to collect data, participants may have provided socially expected responses, thus affecting the validity of the research outcomes. Third, owing to the specific recruitment conditions, the study results can only be generalized to participants who completed the whole process. Fourth, the applicability of the study findings is limited to countries and regions characterized by high mobile phone usage and internet connectivity. Fifth, the recruitment may not have effectively reached HIV risk population in remote areas. Future studies could examine undiscovered social media platforms used by HIV risk population in remote areas to enhance the accessibility of video screening information and recruitment. Finally, the high cost of this study warrants further investigation of the cost-effectiveness and feasibility of this design.

Conclusions

This study demonstrated the feasibility of recruiting participants for HIV screening through social media platforms, particularly by one-on-one message discussions through instant messaging facilitated by friend referrals. Furthermore, the findings revealed the efficacy of continuing VCT for HIV during restrictions associated with the COVID-19 pandemic. Our findings suggest that mobile and video VCT can provide accessible screening options tailored to the testing needs of at-risk groups, thereby promoting their testing motivation and maintaining HIV detection during public health emergencies, such as the COVID-19 pandemic. The findings of this study provide valuable insights for enhancing HIV screening strategies amid emerging infectious disease pandemics.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Taiwan Centers for Disease Control for providing research funding. The authors sincerely thank all study participants for sharing our recruitment information to promote the 2 VCT models. This research was supported by Taiwan Centers for Disease Control (grant LA109042). The manuscript was not assisted by any generative artificial intelligence in its writing.

Data Availability

The analyzed data are unavailable owing to the privacy policy of the research ethics committee of this study.

Conflicts of Interest

None declared.

References

1. Albhaisi S, Wenzel RP. The value of medical registries and observational studies early in pandemics: the Coronavirus disease 2019 (COVID-19) experience. Clin Infect Dis 2022;74(6):1112-1116 [FREE Full text] [doi: 10.1093/cid/ciab634] [Medline: 34270715]
2. Khairnar P, Soni M, Handa M, Riadi Y, Kesharwani P, Shukla R. Recent highlights on Omicron as a new SARS-COVID-19 variant: evolution, genetic mutation, and future perspectives. J Drug Target 2022;30(6):603-613 [FREE Full text] [doi: 10.1080/1061186X.2022.2056187] [Medline: 35311601]
3. Weekly epidemiological update on COVID-19 - 16 March 2023. World Health Organization. 2023 Mar 16. URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---16-march-2023 [accessed 2023-03-19]
4. Infection prevention and control in the context of coronavirus disease (COVID-19) : a living guideline. World Health Organization. 2023 Aug 10. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-IPC-guideline-2023.2 [accessed 2023-08-20]
5. CECC raises epidemic alert level for Taipei City and New Taipei City to Level 3 and strengthens national restrictions and measures, effective from May 15 to May 28, in response to increasing level of community transmission. Ministry of Health and Welfare. 2021. URL: https://covid19.mohw.gov.tw/en/cp-4868-61352-206.html [accessed 2021-05-15]
6. Phillips TR, Fairley CK, Donovan B, Ong JJ, McNulty A, Marshall L, et al. Sexual health service adaptations to the coronavirus disease 2019 (COVID-19) pandemic in Australia: a nationwide online survey. Aust N Z J Public Health 2021;45(6):622-627 [FREE Full text] [doi: 10.1111/1753-6405.13158] [Medline: 34473388]
7. Ejima K, Koizumi Y, Yamamoto N, Rosenberg M, Ludema C, Bento AI, et al. HIV testing by public health centers and municipalities and new HIV cases during the COVID-19 pandemic in Japan. J Acquir Immune Defic Syndr 2021;87(2):e182-e187 [FREE Full text] [doi: 10.1097/QAI.0000000000002660] [Medline: 33625066]
8. Rick F, Odoke W, van den Hombergh J, Benzaken AS, Avelino-Silva VI. Impact of coronavirus disease (COVID-19) on HIV testing and care provision across four continents. HIV Med 2022;23(2):169-177 [FREE Full text] [doi: 10.1111/hiv.13180] [Medline: 34632685]
9. Camargo E, de Oliveira BIA, Siffoni IF, de Sousa AR, Teixeira JRB, Mendes IAC, et al. Low psychological well-being in men who have sex with men (MSM) during the shelter-in-place orders to prevent the COVID-19 spread: results from a nationwide study. Sex Res Social Policy 2022;19(1):391-400 [FREE Full text] [doi: 10.1007/s13178-021-00550-5] [Medline: 33643482]
10. Hyndman I, Nugent D, Whitlock GG, McOwan A, Girometti N. COVID-19 restrictions and changing sexual behaviours in HIV-negative MSM at high risk of HIV infection in London, UK. Sex Transm Infect 2021;97(7):521-524. [doi: 10.1136/sextrans-2020-054768] [Medline: 33462118]
11. Leung TYM, Chan AYL, Chan EW, Chan VKY, Chui CSL, Cowling BJ, et al. Short- and potential long-term adverse health outcomes of COVID-19: a rapid review. Emerg Microbes Infect 2020;9(1):2190-2199 [FREE Full text] [doi: 10.1080/22221751.2020.1825914] [Medline: 32940572]
12. Stephenson R, Chavanduka TMD, Rosso MT, Sullivan SP, Pitter RA, Hunter AS, et al. Sex in the time of COVID-19: results of an online survey of gay, bisexual and other men who have sex with men's experience of sex and HIV prevention during the US COVID-19 epidemic. AIDS Behav 2021;25(1):40-48 [FREE Full text] [doi: 10.1007/s10461-020-03024-8] [Medline: 32876905]
13. Howarth AR, Saunders J, Reid D, Kelly I, Wayal S, Weatherburn P, et al. 'Stay at home …': exploring the impact of the COVID-19 public health response on sexual behaviour and health service use among men who have sex with men: findings from a large online survey in the UK. Sex Transm Infect 2022;98(5):346-352 [FREE Full text] [doi: 10.1136/sextrans-2021-055039] [Medline: 34544888]
14. Odinga MM, Kuria S, Muindi O, Mwakazi P, Njraini M, Melon M, et al. HIV testing amid COVID-19: community efforts to reach men who have sex with men in three Kenyan counties. Gates Open Res 2020;4:117. [doi: 10.12688/gatesopenres.13152.1]

XSL•FO
RenderX

15. Pampati S, Emrick K, Siegler AJ, Jones J. Changes in sexual behavior, PrEP adherence, and access to sexual health services because of the COVID-19 pandemic among a cohort of PrEP-using MSM in the south. J Acquir Immune Defic Syndr 2021;87(1):639-643 [FREE Full text] [doi: 10.1097/QAI.0000000000002640] [Medline: 33512848]
16. Rao A, Rucinski K, Jarrett BA, Ackerman B, Wallach S, Marcus J, et al. Perceived interruptions to HIV prevention and treatment services associated with COVID-19 for gay, bisexual, and other men who have sex with men in 20 countries. J Acquir Immune Defic Syndr 2021;87(1):644-651 [FREE Full text] [doi: 10.1097/QAI.0000000000002620] [Medline: 33443963]
17. Chiou P, Ko NY, Chien CY. Mobile HIV testing through social networking platforms: comparative study. J Med Internet Res 2021;23(3):e25031 [FREE Full text] [doi: 10.2196/25031] [Medline: 33769298]
18. Post-pandemic era - let love continue uninterrupted. Taipei City United Hospital. URL: https://tpech.gov.taipei/mp109151/News_Content.aspx?n=535FED1409A8AEF2&s=F79D32AC2580B996 [accessed 2022-05-01]
19. Chan PS, Chidgey A, Lau J, Ip M, Lau JT, Wang Z. Effectiveness of a novel HIV self-testing service with online real-time counseling support (HIVST-Online) in increasing HIV testing rate and repeated HIV testing among men who have sex with men in Hong Kong: results of a pilot implementation project. Int J Environ Res Public Health 2021;18(2):729. [doi: 10.3390/ijerph18020729]
20. Stephenson R, Sullivan SP, Mitchell JW, Johnson BA, Sullvian PS. Efficacy of a telehealth delivered couples' HIV counseling and testing (CHTC) intervention to improve formation and adherence to safer sexual agreements among male couples in the US: results from a randomized control trial. AIDS Behav 2022;26(8):2813-2824 [FREE Full text] [doi: 10.1007/s10461-022-03619-3] [Medline: 35194698]
21. Sullivan SP, Sullivan PS, Stephenson R. Acceptability and feasibility of a telehealth intervention for STI testing among male couples. AIDS Behav 2021;25(12):4029-4043. [doi: 10.1007/s10461-021-03173-4] [Medline: 33507455]
22. Armstrong W, Agwu AL, Barrette EP, Ignacio RB, Chang JJ, Colasanti JA, et al. Innovations in human immunodeficiency virus (HIV) care delivery during the Coronavirus disease 2019 (COVID-19) pandemic: policies to strengthen the ending the epidemic initiative-a policy paper of the Infectious Diseases Society of America and the HIV Medicine Association. Clin Infect Dis 2021;72(1):9-14 [FREE Full text] [doi: 10.1093/cid/ciaa1532] [Medline: 33035296]
23. Nadarzynski T, Burton J, Henderson K, Zimmerman D, Hill O, Graham C. Targeted advertisement of chlamydia screening on social media: a mixed-methods analysis. Digit Health 2019;5:2055207619827193 [FREE Full text] [doi: 10.1177/2055207619827193] [Medline: 30746155]
24. Hill BJ, Anderson B, Lock L. COVID-19 pandemic, pre-exposure prophylaxis (PrEP) care, and HIV/STI testing among patients receiving care in three HIV epidemic priority states. AIDS Behav 2021;25(5):1361-1365 [FREE Full text] [doi: 10.1007/s10461-021-03195-y] [Medline: 33599879]
25. Warmbrunn I, Green-Jones M, Outlaw AY. Utilizing alternative testing technology for human immunodeficiency virus (HIV) in the COVID era. Curr Emerg Hosp Med Rep 2021;9(2):38-44 [FREE Full text] [doi: 10.1007/s40138-021-00228-7] [Medline: 33972891]
26. Budak JZ, Scott JD, Dhanireddy S, Wood BR. The impact of COVID-19 on HIV care provided via telemedicine-past, present, and future. Curr HIV/AIDS Rep 2021;18(2):98-104 [FREE Full text] [doi: 10.1007/s11904-021-00543-4] [Medline: 33616811]
27. Hoagland B, Torres TS, Bezerra DR, Benedetti M, Pimenta C, Veloso VG, et al. High acceptability of PrEP teleconsultation and HIV self-testing among PrEP users during the COVID-19 pandemic in Brazil. Braz J Infect Dis 2021;25(1):101037 [FREE Full text] [doi: 10.1016/j.bjid.2020.11.002] [Medline: 33285137]
28. Middleton M, Somerset S, Evans C, Blake H. Test@Work texts: mobile phone messaging to increase awareness of HIV and HIV Testing in UK construction employees during the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(21):101037 [FREE Full text] [doi: 10.3390/ijerph17217819] [Medline: 33114546]
29. Market survey analysis: observation and comparison of internet behavior trends in Taiwan from 2020 to 2021. Intercity Biweekly. URL: https://tw.stock.yahoo.com/news/ix-%E5%B8%82%E8%AA%BF%E8%A7%A3%E6%9E%90:2020%E8%87%B32021%E5%B9%B4%E5%8F%B0%E7%81%A3%E7%B6%B2%E8%B7%AF%E8%A1%8C%E7%82%BA%E8%B6%A8%E5%8B%A2%E8%A7%80%E5%AF%9F%E8%88%87%E6%AF%94%E8%BC%83-092939821.html [accessed 2022-05-06]
30. Iribarren SJ, Ghazzawi A, Sheinfil AZ, Frasca T, Brown W, Lopez-Rios J, et al. Mixed-method evaluation of social media-based tools and traditional strategies to recruit high-risk and hard-to-reach populations into an HIV prevention intervention study. AIDS Behav 2018;22(1):347-357 [FREE Full text] [doi: 10.1007/s10461-017-1956-6] [Medline: 29124420]
31. Nguyen LH, Tran BX, Rocha LEC, Nguyen HLT, Yang C, Latkin CA, et al. A systematic review of eHealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019;23(9):2253-2272 [FREE Full text] [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
32. Nebeker C, Dunseath SE, Linares-Orozco R. A retrospective analysis of NIH-funded digital health research using social media platforms. Digit Health 2020;6:2055207619901085 [FREE Full text] [doi: 10.1177/2055207619901085] [Medline: 32030195]
33. 2023 Taiwan internet report. Taiwan Network Information Center. 2023. URL: https://report.twnic.tw/2023/assets/download/TWNIC_TaiwanInternetReport_2023_CH_all.pdf [accessed 2024-04-01]

XSL•FO
RenderX

34. The current situation of social media in Taiwan: social platform development trends and user distribution data in 2023. OOSGA. 2023. URL: https://zh.oosga.com/social-media/twn/ [accessed 2024-04-01]
35. Digital 2024: Taiwan. DATAREPORTAL. URL: https://datareportal.com/reports/digital-2024-taiwan [accessed 2024-04-01]
36. Wittig C, Hornet. In: Das S, Zutshi A, Janardhanan J, editors. The Great Lockdown: Lessons Learned During the Pandemic from Organizations Around the World. Hoboken, NJ: Wiley; 2021.
37. Gibson L. World's largest dating app for gays sets up shop in Taiwan. Business Next Media Corp. 2018. URL: https://meet-global.bnext.com.tw/articles/view/43307? [accessed 2024-04-01]
38. Cao B, Gupta S, Wang J, Hightow-Weidman LB, Muessig KE, Tang W, et al. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017;19(11):e394 [FREE Full text] [doi: 10.2196/jmir.7997] [Medline: 29175811]
39. Macapagal K, Moskowitz DA, Li DH, Carrión A, Bettin E, Fisher CB, et al. Hookup app use, sexual behavior, and sexual health among adolescent men who have sex with men in the United States. J Adolesc Health 2018;62(6):708-715 [FREE Full text] [doi: 10.1016/j.jadohealth.2018.01.001] [Medline: 29784114]
40. Gibson LP, Kramer EB, Bryan AD. Geosocial networking app use associated with sexual risk behavior and pre-exposure prophylaxis use among gay, bisexual, and other men who have sex with men: cross-sectional web-based survey. JMIR Form Res 2022;6(6):e35548. [doi: 10.2196/35548] [Medline: 35699992]
41. Council ND. 108-Year digital application survey report in remote villages and towns. Taiwan Ministry of Digital Affairs. 2019. URL: https://www-api.moda.gov.tw/File/Get/moda/zh-tw/PEkbv7pfK1MNhu1 [accessed 2024-04-05]
42. Gisondi MA, Barber R, Faust JS, Raja A, Strehlow MC, Westafer LM, et al. A deadly infodemic: social media and the power of COVID-19 misinformation. J Med Internet Res 2022;24(2):e35552 [FREE Full text] [doi: 10.2196/35552] [Medline: 35007204]
43. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. Sci Rep 2020;10(1):16598 [FREE Full text] [doi: 10.1038/s41598-020-73510-5] [Medline: 33024152]
44. Rahib D, Delagreverie H, Gabassi A, Le Thi T, Vassel E, Vodosin P, et al. Online self-sampling kits to screen multipartner MSM for HIV and other STIs: participant characteristics and factors associated with kit use in the first 3 months of the MemoDepistages programme, France, 2018. Sex Transm Infect 2021;97(2):134-140. [doi: 10.1136/sextrans-2020-054790] [Medline: 33397802]
45. HIV/AIDS statistics. Taiwan Centers for Disease Control. 2023. URL: https://www.cdc.gov.tw/Category/Page/rCV9N1rGUz9wNr8lggsh2Q [accessed 2023-03-29]
46. Islam M, Sarkar T, Khan SH, Mostofa Kamal AH, Hasan SMM, Kabir A, et al. COVID-19-related infodemic and its impact on public health: a global social media analysis. Am J Trop Med Hyg 2020;103(4):1621-1629 [FREE Full text] [doi: 10.4269/ajtmh.20-0812] [Medline: 32783794]
47. Stephenson R, Todd K, Kahle E, Sullivan SP, Miller-Perusse M, Sharma A, et al. Project moxie: results of a feasibility study of a telehealth intervention to increase HIV testing among binary and nonbinary transgender youth. AIDS Behav 2020;24(5):1517-1530 [FREE Full text] [doi: 10.1007/s10461-019-02741-z] [Medline: 31760536]
48. Wang Z, Lau JTF, Ip M, Ho SPY, Mo PKH, Latkin C, et al. A randomized controlled trial evaluating efficacy of promoting a home-based HIV self-testing with online counseling on increasing HIV testing among men who have sex with men. AIDS Behav 2018;22(1):190-201. [doi: 10.1007/s10461-017-1887-2] [Medline: 28831616]
49. Miners A, Nadarzynski T, Witzel C, Phillips AN, Cambiano V, Rodger AJ, et al. Preferences for HIV testing services among men who have sex with men in the UK: a discrete choice experiment. PLoS Med 2019;16(4):e1002779 [FREE Full text] [doi: 10.1371/journal.pmed.1002779] [Medline: 30973868]
50. Alsmadi I, Rice NM, O'Brien MJ. Fake or not? Automated detection of COVID-19 misinformation and disinformation in social networks and digital media. Comput Math Organ Theory 2022:1-19 [FREE Full text] [doi: 10.1007/s10588-022-09369-w] [Medline: 36466587]
51. Xu Q, McMann T, Godinez H, Nali MC, Li J, Cai M, et al. Impact of COVID-19 on HIV prevention access: a multi-platform social media infodemiology study. AIDS Behav 2023;27(6):1886-1896 [FREE Full text] [doi: 10.1007/s10461-022-03922-z] [Medline: 36471205]
52. Chiou PY, Hung CC, Chen CY. Sexual partner referral for HIV testing through social networking platforms: cross-sectional study. JMIR Public Health Surveill 2022;8(4):e32156 [FREE Full text] [doi: 10.2196/32156] [Medline: 35380540]
53. Sarno EL, Bettin E, Jozsa K, Newcomb ME. Sexual health of rural and urban young male couples in the United States: differences in HIV testing, pre-exposure prophylaxis use, and condom use. AIDS Behav 2021;25(1):191-202 [FREE Full text] [doi: 10.1007/s10461-020-02961-8] [Medline: 32607917]
54. McKenney J, Sullivan PS, Bowles KE, Oraka E, Sanchez TH, DiNenno E. HIV risk behaviors and utilization of prevention services, urban and rural men who have sex with men in the United States: results from a national online survey. AIDS Behav 2018;22(7):2127-2136. [doi: 10.1007/s10461-017-1912-5] [Medline: 28986669]
55. Campbell CK, Lippman SA, Moss N, Lightfoot M. Strategies to increase HIV testing among MSM: a synthesis of the literature. AIDS Behav 2018;22(8):2387-2412. [doi: 10.1007/s10461-018-2083-8] [Medline: 29550941]

XSL•FO
RenderX

56. Okoboi S, Twimukye A, Lazarus O, Castelnuovo B, Agaba C, Immaculate M, et al. Acceptability, perceived reliability and challenges associated with distributing HIV self-test kits to young MSM in Uganda: a qualitative study. J Int AIDS Soc 2019;22(3):e25269 [FREE Full text] [doi: 10.1002/jia2.25269] [Medline: 30932364]
57. Verhaegh-Haasnoot A, Dukers-Muijrers NHTM, Hoebe CJPA. High burden of STI and HIV in male sex workers working as internet escorts for men in an observational study: a hidden key population compared with female sex workers and other men who have sex with men. BMC Infect Dis 2015;15:291 [FREE Full text] [doi: 10.1186/s12879-015-1045-2] [Medline: 26220287]
58. Tilchin C, Schumacher CM, Psoter KJ, Humes E, Muvva R, Chaulk P, et al. Human immunodeficiency virus diagnosis after a syphilis, gonorrhea, or repeat diagnosis among males including non-men who have sex with men: what is the incidence? Sex Transm Dis 2019;46(4):271-277 [FREE Full text] [doi: 10.1097/OLQ.0000000000000964] [Medline: 30870326]
59. Shrestha R, Lim SH, Altice FL, Copenhaver M, Wickersham JA, Saifi R, et al. Use of smartphone to seek sexual health information online among Malaysian men who have sex with men (MSM): implications for mhealth intervention to increase HIV testing and reduce HIV risks. J Community Health 2020;45(1):10-19 [FREE Full text] [doi: 10.1007/s10900-019-00713-x] [Medline: 31375976]
60. Turner D, Lockhart E, Marhefka SL. Willingness of MSM living with HIV to take part in video-groups: application of the technology readiness and acceptance model. AIDS Behav 2019;23(11):3165-3174. [doi: 10.1007/s10461-019-02534-4] [Medline: 31123856]
61. Croxford S, Stengaard AR, Brännström J, Combs L, Dedes N, Girardi E, et al. Late diagnosis of HIV: an updated consensus definition. HIV Med 2022;23(11):1202-1208 [FREE Full text] [doi: 10.1111/hiv.13425] [Medline: 36347523]
62. van Bremen K, Monin M, Schlabe S, Bischoff J, Rieke GJ, Schwarze-Zander C, et al. Impact of COVID-19 on HIV late diagnosis in a specialized German centre. HIV Med 2022;23(11):1209-1213 [FREE Full text] [doi: 10.1111/hiv.13426] [Medline: 36263724]
63. Stafylis C, Natoli LJ, Murkey JA, Gordon KK, Young SD, McGrath MR, et al. Vending machines in commercial sex venues to increase HIV self-testing among men who have sex with men. Mhealth 2018;4:51 [FREE Full text] [doi: 10.21037/mhealth.2018.10.03] [Medline: 30505849]

Abbreviations

GSN: geosocial network
LGBTQ: lesbian, gay, bisexual, transgender, and queer
MSM: men who have sex with men
OR: odds ratio
PEP: postexposure prophylaxis
PrEP: pre-exposure prophylaxis
STD: sexually transmitted disease
VCT: voluntary counseling and testing

Edited by A Mavragani; submitted 11.11.23; peer-reviewed by X Han, A Outlaw; comments to author 25.03.24; revised version received 26.05.24; accepted 23.08.24; published 28.11.24.

Please cite as:
Chiou PY, Tsao WW, Li CL, Yu JM, Su WH, Liu ZH, He CR, Chang YC, Tsai YH
Recruitment for Voluntary Video and Mobile HIV Testing on Social Media Platforms During the COVID-19 Pandemic: Cross-Sectional Study
J Med Internet Res 2024;26:e54420
URL: https://www.jmir.org/2024/1/e54420
doi:10.2196/54420
PMID:39607762

©Piao-Yi Chiou, Wei-Wen Tsao, Chia-Lin Li, Jheng-Min Yu, Wen-Han Su, Zhi-Hua Liu, Cheng-Ru He, Yu-Chun Chang, Yi-Hsuan Tsai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Challenges and Opportunities in Digital Screening for Hypertension and Diabetes Among Community Groups of Older Adults in Vietnam: Mixed Methods Study

Trang Thi Thu Nong[1], BPH; Giang Hoang Nguyen[2], MPH; Alexander Lepe[3], MPH, PhD; Thuy Bich Tran[1], MDM; Lan Thi Phuong Nguyen[4], MD, PhD; Jaap A R Koot[3], MD, MBA

[1]HelpAge International in Vietnam, Hanoi, Vietnam

[2]Health Strategy and Policy Institute, Vietnam Ministry of Health, Hanoi, Vietnam

[3]Department of Health Sciences, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[4]Faculty of Public Health, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam

Corresponding Author:
Trang Thi Thu Nong, BPH
HelpAge International in Vietnam
A201, No 2, Alley 59, Lang Ha street
Thanh Cong ward, Ba Dinh district
Hanoi, 100000
Vietnam
Phone: 84 0947375689
Email: trangnt@helpagevn.org

Abstract

Background: The project of scaling up noncommunicable disease (NCD) interventions in Southeast Asia aimed to strengthen the prevention and control of hypertension and diabetes, focusing on primary health care and community levels. In Vietnam, health volunteers who were members of the Intergenerational Self-Help Clubs (ISHCs) implemented community-based NCD screening and health promotion activities in communities. The ISHC health volunteers used an app based on District Health Information Software, version 2 (DHIS2) tracker (Society for Health Information Systems Programmes, India) to record details of participants during screening and other health activities.

Objective: This study aimed to assess the strengths, barriers, and limitations of the NCD screening app used by the ISHC health volunteers on tablets and to provide recommendations for further scaling up.

Methods: A mixed methods observational study with a convergent parallel design was performed. For the quantitative data analysis, 2 rounds of screening data collected from all 59 ISHCs were analyzed on completeness and quality. For the qualitative analysis, 2 rounds of evaluation of the screening app were completed. Focus group discussions with ISHC health volunteers and club management boards and in-depth interviews with members of the Association of the Elderly and Commune Health Station staff were performed.

Results: In the quantitative analysis, data completeness of all 6704 screenings (n=3485 individuals) was very high. For anthropomorphic measurements, such as blood pressure, body weight, and abdominal circumference, less than 1% errors were found. The data on NCD risk factors were not adequately recorded in 1908 (29.5%) of the screenings. From the qualitative analysis, the NCD screening app was appreciated by ISHC health volunteers and supervisors, as an easier and more efficient way to report to higher levels, secure data, and strengthen relationships with relevant stakeholders, using tablets to connect to the internet and internet-based platforms to access information for self-learning and sharing to promote a healthy lifestyle as the strengths. The barriers and limitations reported by the respondents were a non–age-friendly app, incomplete translation of parts of the app into Vietnamese, some issues with the tablet's display, lack of sharing of responsibilities among the health volunteers, and suboptimal involvement of the health sector; limited digital literacy among ISHC health volunteers. Recommendations are continuous capacity building, improving app issues, improving tablet issues, and involving relevant stakeholders or younger members in technology adoption to support older people.

Conclusions: The implementation of the NCD screening app by ISHC volunteers can be an effective way to improve community-led NCD screening and increase the uptake of NCD prevention and management services at the primary health care

XSL•FO
RenderX

level. However, our study has shown that some barriers need to be addressed to maximize the efficient use of the app by ISHC health volunteers to record, report, and manage the screening data.

(J Med Internet Res 2024;26:e54127) doi:10.2196/54127

KEYWORDS

NCD screening; DHIS2 tracker; District Health Information Software, version 2 tracker; digital application; ISHC health volunteers; non-communicable diseases; prevention; Vietnam; mobile phone

Introduction

Noncommunicable diseases (NCDs) have become a leading burden of disease, and accounted for 74% and 77% of all deaths globally and in Vietnam, respectively [1,2]. Older people are increasingly susceptible to NCDs and multimorbidity as they age [3] and Vietnam is among the most rapidly aging countries in the world [1-3]. The government is tackling an increased prevalence of NCDs [4,5], and one priority branch is to deal with hypertension and diabetes, which are recognized as having high prevalence, low percentage of awareness of their condition, and undertreatment [5,6]. In 2019, there were 11.8 million individuals with hypertension in Vietnam, while only 26.8% were under treatment [6]. It is estimated that 1 in every 20 Vietnamese adults has diabetes and more than 50% of patients with diabetes are unaware of their condition [7]. However, many obstacles still exist such as a lack of adequate resources and suboptimal policy implementation, hospitals are primarily responsible for providing NCD services, and provision of these services at primary health care (PHC) facilities is limited [4]. Therefore, the government has prioritized focusing on reaching more individuals for screening, diagnosis, and treatment of NCDs by strengthening the provision of services at the community level [4].

In this context, the research project "Scaling up non-communicable disease interventions in Southeast Asia" (SUNI-SEA) was implemented in Indonesia, Myanmar, and Vietnam. SUNI-SEA aimed to strengthen the quality and reach of NCD prevention and management services and to increase the linkage between the community and the PHC facilities [8].

In Vietnam, a component of community intervention in the SUNI-SEA project was implemented by 59 Intergenerational Self-Help Clubs (ISHCs), in which 295 ISHC health volunteers with a mean age of 65.3 years were working in Ninh Binh and Hai Phong provinces [8]. A detail of the ISHC model is presented in Multimedia Appendix 1 and widely documented [9-15]. In brief, ISHC is a community-based organization that has a dominant proportion of older members (about 70%). This model applies a comprehensive and inclusive approach that promotes healthy aging through a variety of interventions such as physical exercise, periodic health screenings, monthly monitoring of weight and blood pressure, health check-ups, and access to health insurance. During the project, ISHC health volunteers received initial training, health kits, and communication materials to conduct health promotion activities and NCD screening for risk factors of diabetes and hypertension with 6 tables of the screening process including (1) registration; (2) measuring waist circumference, height, weight, and BMI; (3) evaluating the Finnish Diabetes Risk Score (FINDRISC); (4) measuring blood pressure; (5) entering screening data into the NCD screening application; and (6) consultation (Multimedia Appendix 2). To maintain the sustainability and continuous strengthening of the capacity of ISHC health volunteers, 2 rounds of training, monitoring, and technical support trips were conducted [8]. Supervisors sometimes helped with data recording during screening activities. Members of ISHCs who were screened and classified as high risk of having hypertension or diabetes were advised to go to the PHC facility nearby for further diagnosis and treatment. Individuals who were diagnosed with an NCD in the PHC facility were treated and referred back to ISHCs after initial medical treatment, where ISHC health volunteers and club management boards helped to monitor their treatment adherence and promoted a healthy lifestyle (Multimedia Appendix 3).

There is mounting evidence of the relevance of eHealth apps in community health [16], including in Vietnam [17]. To further support ISHC health volunteers in capturing and using screening data, the SUNI-SEA project introduced a digital mobile NCD screening app in November 2021. This app was based on the open-source District Health Information Software, version 2 (DHIS2) tracker [18]. ISHC health volunteers used tablets to capture screening data, which were subsequently uploaded to a cloud server managed by HelpAge International in Vietnam (HAIV) once an internet connection was available. Training on using the tablet and the NCD screening app was integrated into the broader NCD prevention and management training. In addition, a tutorial video and user guidance for the app were provided to the ISHC health volunteers [19,20]. Due to the COVID-19 pandemic restrictions, the project implementation duration was approximately 1 year.

In the context of the digital era, eHealth interventions are supported for enhancing NCD management effectiveness, such as perspectives of clinical, behavioral changes, and service implementation outcomes. The recent reviews of eHealth interventions for NCD management in PHC settings of low- and middle-income countries found that eHealth interventions targeted health care providers or people with NCD mostly [16,17,21]. The studies, however, have not yet considered eHealth interventions for users who are volunteers, organizations, and older people in the community level. To fulfill this gap, this study aims to assess the strengths, barriers, and limitations of using DHIS2 on the tablet by ISHC health volunteers. Furthermore, it seeks to provide recommendations for the future use of digital NCD screening apps in community-based settings.

XSL•FO
RenderX

Methods

Study Design

This study uses a mixed methods observational approach with a convergent parallel design, aiming to achieve a comprehensive and multidimensional understanding of the implementation of a mobile health app in community-based organizations. In accordance with this design, quantitative data from the NCD screening app, collected by ISHC health volunteers in 2 provinces, were extracted and analyzed. These quantitative data facilitated the evaluation of ISHC health volunteers' practices and the effectiveness of the app in supporting NCD prevention and management within ISHCs. Concurrently, qualitative research was conducted during and after the implementation of the screening app among ISHCs. These qualitative data provide deeper insights into user experiences and highlight the barriers and challenges encountered in integrating digital health apps into the ISHCs' routine screening activities.

Sample

In Vietnam, the SUNI-SEA project took place in the provinces of Ninh Binh and Hai Phong. Within these provinces, we implemented the intervention in 7 districts, of which 3 were urban and 4 were rural. In all 7 districts, the project was implemented in 44 purposely selected communes, which had 1 or more ISHCs, as well as a district health center and commune health station (CHS) provided both treatment and preventive services, which made up a sample of ISHCs (n=59) participating in the community-based intervention.

The ISHC health volunteers conducted 2 rounds of screenings in 2022, to which all members were invited. The first screening took place during the first half of 2022, and the second during the second half of 2022 with an interval of at least 6 months. In total 3463 individuals took part in the first screening and 3241 individuals took part in the second screening. Most of these individuals participated in both screenings (n=3222). Ultimately, data were collected from 3485 unique individuals among 59 ISHCs in these 2 provinces for the quantitative study.

In the qualitative study, we selected the ISHCs and the participants in screening purposively for inclusion in focus group discussions (FGDs) and in-depth interviews (IDIs), to achieve optimal diversity of users' feedback. In total, 29 participants participated in FGDs in the first evaluation in September 2022 and 44 participants participated in FDGs and IDIs during the second evaluation in April 2023. In total, we had 70 unique participants; 3 individuals (each from a different ISHC) participated in both evaluations. We applied the following criteria for the selection of ISHCs and for individuals: (1) participants from ISHCs in both provinces, (2) participants from ISHCs in rural and urban areas, (3) participants from ISHCs who received previous financial and technical support from HAIV and ISHCs that did not receive previous support from HAIV (before the SUNI-SEA project), (4) ISHC participants who have experience entering screening data into the NCD screening application, and (5) participants who have not entered data into the NCD screening app before the first screening event in 2022. Figure 1 illustrates the sample of the qualitative study and the timeline of data collection.

To better understand the management aspects of the use of the NCD screening application and the potential for adoption and scaling up of the digital technology at the community level, various stakeholders were interviewed in the second evaluation, including representatives of the Association of the Elderly at the national, provincial, district and commune levels, and the CHS's representatives. The representatives of commune and district stakeholders were based in the same areas as the selected ISHCs.

Figure 1. Timeline of data collection and study population. FGD: focus group discussion; IDI: in-depth interview.

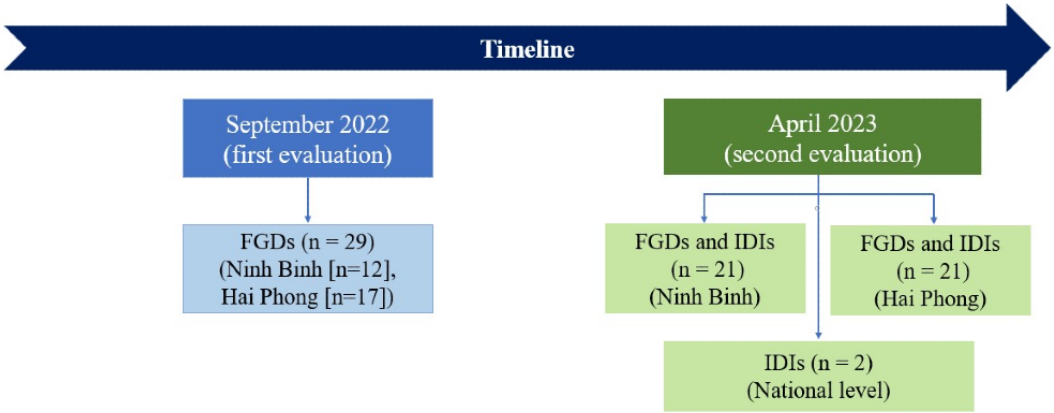

Data Collection

The DHIS2 tracker software was used as a digital mobile NCD screening app for health volunteers in ISHCs to record and report data, which is available on the Google Play Store. To manage data security, HelpAge provided each ISHC with a log-in account to use and report data. The app was able to record data offline but to synchronize data to the main server, the device needed to connect to the internet. To ensure data protection, only the assigned HelpAge staff could log in and check the data of all ISHCs but from the ISHC level, users can review and check the data from their ISHC only. As guided by the project manual, screening was performed in a standardized manner, with predefined questions and measurements. The NCD

XSL•FO
RenderX

screening app included questions about the demographics, that is, sex, age, marital status, place of living, and ISHC name. The survey included questions about previous diagnoses with an NCD, previous information about risk factors (eg, elevated glucose, overweight or obesity, and lifestyle issues), and the items from the FINDRISC questionnaire [22]. The ISHC health volunteers also measured the participants' weight, height, waist circumference, and systolic and diastolic blood pressure; BMI was automatically calculated by the NCD screening app. The screening data were collected by trained ISHC health volunteers using tablets. However, often paper forms were used first, and the data were later entered into the screening app. All the data collected from 59 ISHCs were securely retrieved and included for further data analysis in the current study.

The first evaluation was conducted in September 2022 and the second was conducted in April 2023. During the first evaluation, the FGD guidelines and questionnaires were developed in English, translated into Vietnamese, tested, and adapted to the local participants' backgrounds. The FGDs were conducted with selected ISHC health volunteers who met the inclusion criteria. Each FGD was conducted in a private place, sitting in a circle to ensure a friendly environment for the discussion with 1 facilitator and 1 note-taker who took minutes of the discussions. In the second evaluation, besides FGDs, IDIs were conducted. The IDIs were conducted individually and additional questions to obtain more in-depth information were asked. The interview guidelines were developed in English and translated into Vietnamese and adapted to the interviewees' backgrounds. During the second evaluation, there were interviewers, note-takers, and interpreters (to interpret the questions and answers from English to Vietnamese and vice versa). To ensure accurate note-taking and reduce the loss of information missing, 2 note-takers were assigned to capture all answers during the interviews. FGDs and IDIs were not recorded in audio in both evaluations.

Analysis

For the quantitative study, first, we provide a basic description of some of the demographic characteristics of our sample. To gain insight into the ISHC health volunteers' use of the NCD screening app, we first assessed the completeness of the data. Finally, we checked whether there were potential errors in collected data by performing crosstabs of the item asking if an individual had previously been diagnosed with an NCD with subquestions specifying which NCD had been diagnosed, and we checked if plausible values were input for the anthropomorphic measures. Statistical analyses were performed using R (version 4.2.0; R Core Team).

The information received from the FGDs and IDIs of the qualitative study was manually coded and analyzed by identifying major themes, subthemes, codes, organizing the data, and interpreting the themes and ideas in the context. The codes were categorized based on the subquestions and purposes of the study within the following groups' strengths, limitations or barriers, and recommendations for improving the use of app and further scaling up. The analysis was independently performed by 2 researchers (TTTN and GHN), who reconciled their findings afterward.

Ethical Considerations

Ethical approval for the SUNI-SEA project in Vietnam was initially granted on November 13, 2019, by the Medical and Health Research Ethics Committee (approval number 019-485/DD-YTCC). Due to project disruptions caused by the COVID-19 outbreak, this approval was renewed on April 25, 2023 (approval number 196/2023/YTCC-HD3). In addition, an amendment for the use of quantitative data from the DHIS2 database was approved (298/2024/YTCC-HD3).

For privacy and confidentiality, all data were anonymized before analyzing. No compensation was provided to participants for taking part in this project. All participants received a full explanation of the objectives of the research and the interview process and were informed that they were free to opt out of participating. Informed consent was requested from each participant before the measurements were taken or interviews were conducted with 100% acceptance.

Results

Quantitative Results

Demographic Characteristics of the Screened Individuals

In total, 3485 individuals were screened. Of these, the majority were female (2642, 75.8% individuals). On average, the individuals screened were 65.3 (SD 10.1) years old. Most of the sample were either married (2968, 85.2% individuals) or widowed (406, 11.6% individuals). The remaining 3.2% of individuals were either single (84 individuals), divorced (14 individuals), or separated (13 individuals).

Completeness of the Data

Overall, the ISHC health volunteers filled in the required fields in the NCD screening app, often after first entering the data in paper-based records. The demographic data were recorded for all participants. The anthropomorphic data were nearly complete in all 6704 entries; diastolic blood pressure was not recorded for 1 individual during the second screening. The ISHC health volunteers generally filled in all the questions regarding having been previously diagnosed with an NCD. During the first screening, this was only missing in 0.4% (15/3463) of individuals. Similarly, during the second screening, this was only missing in 1.1% (36/3241) of individuals. The data concerning risk factors were not adequately recorded in 29.5% (1980/6704) of the screenings; 1156 entries were missing from the first screening and 824 entries were missing from the second screening. The FINDRISC items were mostly filled in. FINDRISC scores could not be calculated in 5.4% (186/3463) and 4.6% (150/3241) of participants from the first and second screenings, respectively. This was due to at least 1 item missing.

The quantitative results have pointed out that there was still a certain percentage of missing information from data entry by ISHC health volunteers. The qualitative results recorded several barriers and limitations in the use of the NCD screening app that help to explain the reasons.

XSL•FO
RenderX

Data Entry Errors

When noting previous diagnoses of an NCD, in a few cases the ISHC health volunteers did not correctly record the diagnosis. This occurred 15 times during the first screening and 32 during the second screening. For the anthropometric data, we checked for plausible values by inspecting the distribution of these variables (Figures 2 and 3). Of the recorded values, 1.1% (71/6704) to 3% (204/6704) of values were outliers, for weight and height, respectively. However, at most 0.2% (11/6704) of the entries were defined as extreme outliers (a value that is 3 times the IQR away from the first or third quartile).

Figure 2. Distribution of the anthropometric measures from the first screening. DBP: diastolic blood pressure; SBP: systolic blood pressure.

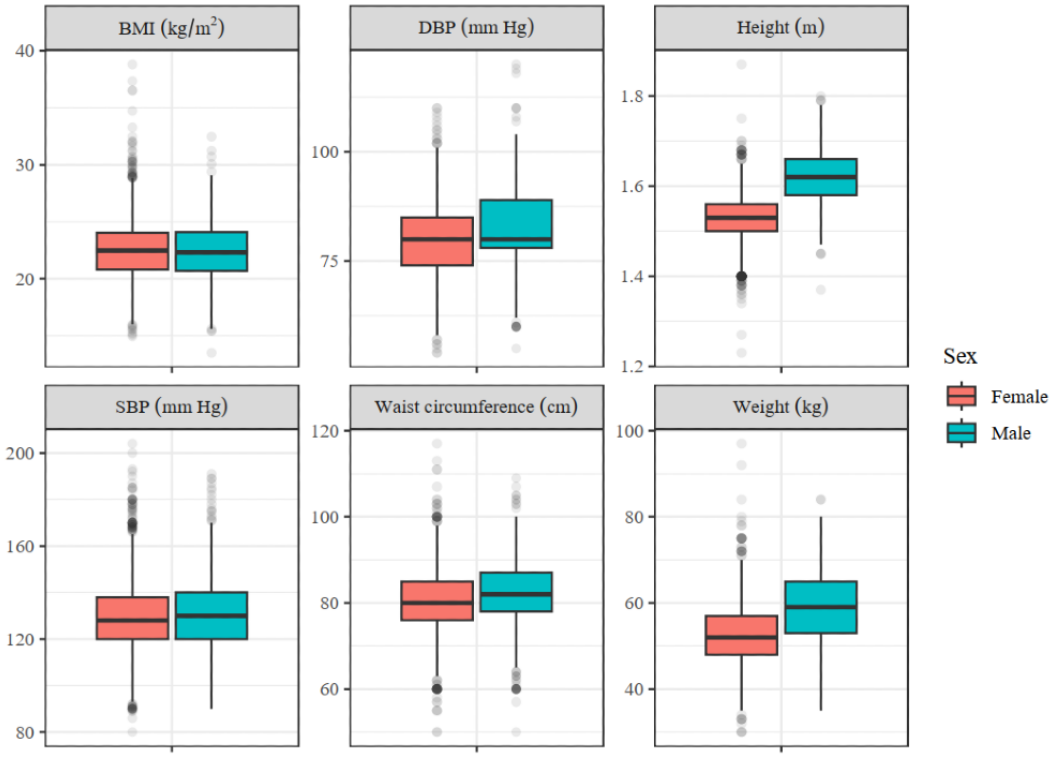

Figure 3. Distribution of the anthropometric measures from the second screening. DBP: diastolic blood pressure; SBP: systolic blood pressure.

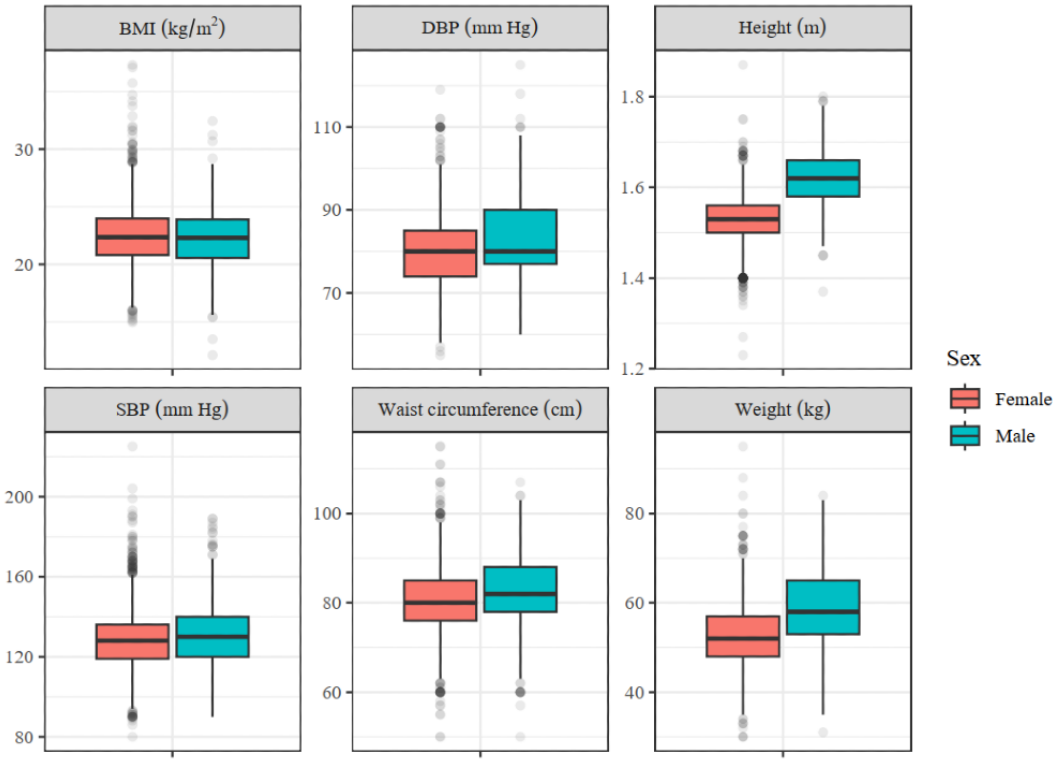

XSL•FO
RenderX

We checked the outliers individually and concluded that for height, weight abdominal circumference, and BMI, the outliers were generally plausible. Among older Asian females, few women with a low body weight are found. However, we suspect that in 5 cases errors were made when recording the height or weight, which resulted in a BMI below 15. The American Heart Association considers systolic blood pressure below 90 mm Hg and diastolic blood pressure below 60 mm Hg too low [23]. A total of 8 women had a systolic blood pressure below 90 mm Hg, and 11 women had a diastolic blood pressure below 60 mm Hg. Although it is clinically possible, these could also have been measuring errors.

Qualitative Results

For the first evaluation, a total of 6 in-person FGDs were conducted in September 2022 in 13 ISHCs (9 ISHCs in Hai Phong and 4 ISHCs in Ninh Binh) with the participation of 29 ISHC health volunteers and club management board members. For the second evaluation, in 16 ISHCs (8 ISHCs in Ninh Binh and 8 ISHCs in Hai Phong) FGDs and IDIs were performed with a total of 44 interviewees.

Characteristics of Participants

The percentage of females in the first qualitative evaluation was 76% (22/29) and in the second evaluation, it was 57% (25/44). The average age during the first evaluation was 63.0 years and 64.2 years during the second evaluation. The percentage of people with college or university diplomas in the first evaluation was 14% (4/29) and 36% (16/44) during the second evaluation (Table 1).

Table 1. Detailed characteristics of participants.

Characteristics	Participants from the first evaluation (N=29), n (%)	Participants from the second evaluation (N=44), n (%)
Sex		
Male	7 (24)	19 (43)
Female	22 (76)	25 (57)
Average age, mean (SD)	63.0 (9.7)	64.2 (10.2)
Geographic area		
Rural	19 (66)	22 (50)
Urban	10 (35)	22 (50)
Education status		
Secondary school	6 (21)	6 (14)
High school	19 (65)	22 (50)
College or university	4 (14)	16 (36)

Findings

In evaluating the usability of the NCD screening app for ISHC health volunteers and ISHCs, 3 main themes were identified through manual coding, including strengths in the use of the NCD screening app, barriers and limitations in the use of NCD screening app, and recommendations for improving the use of NCD screening app and further scaling up. This resulted in a wide range of sometimes contradicting opinions, both positive and negative. The answers per theme are listed in Table 2. The full table of answers with themes, subthemes, and codes is described in Multimedia Appendix 4. Interviewees reflected on the technical issues of the app, on the capacity of volunteers using it, and on organizational and strategic issues of electronic data collection. Besides giving positive and negative opinions they made concrete suggestions for improvement. The frequency of responses is provided combined with quotes in Tables 3-5. This helps to get a better insight into the importance of some remarks, and more context of comments provided during FGDs and IDIs.

XSL•FO
RenderX

Table 2. Categories of findings and answers from interviewees.

Themes	Subthemes
Strengths in the use of the NCD[a] screening app	• Advanced app functions • Building capacity for management boards, ISHC[b] health volunteers, and members • Improving the health of ISHC members • Increasing reputation for club management board members and ISHC health volunteers with their club members • Strengthening relationships with stakeholders • Helping the implementation of the national ISHC model • Storing photos of ISHC's activities on the tablet • Others
Barriers and limitations in the use of the NCD screening app	• App issues • Limited capacity of ISHC health volunteers/ older people in technology • Tablet issues • Irregular technical support after initial training due to COVID-19 restrictions • Others
Recommendations for improving the use of the NCD screening app and further scaling up	• Continuous capacity building • Improving app issues • Improving tablet issues • Involving relevant stakeholders • Involving younger members in tablet use or technology adoption to support older people • Others

[a]NCD: noncommunicable disease.

[b]ISHC: Intergenerational Self-Help Clubs.

Table 3. Quotes on strengths in the use of the noncommunicable disease screening app.

Strength	Quote	Participant description
Advanced app functions	"Using a tablet and the NCD[a] screening application to report data helps me not to have to report by paper forms, just enter data into the application and then press sync function then data is sent very quickly, saving money and time."	Female, ISHC[b] health volunteer, Hai Phong
Building capacity for management boards, ISHC health volunteers, and members	"In addition, my club also uses a tablet connecting to Zalo groups to receive information and read newspapers to raise awareness, and access information and technology."	Female, chairperson and ISHC health volunteer, Ninh Binh
Improving the health of ISHC members	"…. but the tablet is also used by the club management board to connect to the network with songs and performances for physical exercises for all club members. That's good for us."	Female, chairperson and ISHC health volunteer, Ninh Binh
Increasing reputation for club management board members and ISHC health volunteers with their club members	"… I feel more confident. Club members trust me, ISHC health volunteers, and the Club Management Board more because they saw us using the tablet phone and the NCD screening application to record their screen data. We can do the same as younger people. They (club members) feel happy and secure."	Female, chairperson, Hai Phong

[a]NCD: noncommunicable disease.

[b]ISHC: Intergenerational Self-Help Clubs.

XSL•FO
RenderX

Table 4. Quotes on barriers and limitations in the use of the noncommunicable disease screening app.

Barrier and limitation	Quote	Participant description
App issues	"Some parts of the software are still in English, so it is difficult to remember. After the training, when we go home and practice, we see the English parts then we would forget everything."	Female, ISHC[a] health volunteer, Hai Phong
Limited capacity of ISHC health volunteers or older people in technology	"I feel difficult and confused because the application is new and I'm old."	Female, ISHC health volunteer, Hai Phong
Limited capacity of ISHC health volunteers or older people in technology	"I'm afraid that when I press the wrong button then it will log out and difficult to log in back."	Female, chairperson and ISHC health volunteer, Ninh Binh
Tablet issues	"Letters and keyboard in the tablet phone are small and my eyesight is poor, so I often make a mess."	Female, ISHC health volunteer, Ninh Binh

[a]ISHC: Intergenerational Self-Help Clubs.

Table 5. Quotes on the recommendations.

Recommendation	Quote	Participant description
Continuous capacity building	"Training, refresh training, and technical support are necessary for us to use app fluently."	Female, chairperson and ISHC[a] health volunteer, Ninh Binh
Continuous capacity building	"Tutorial video is useful."	Female, ISHC health volunteer, Hai Phong
Improving app issues	"Converting language from English to Vietnamese in few parts so we can remember better."	Male, ISHC health volunteer, Hai Phong
Involving younger members in tablet use or technology adoption to support older people	"We're old, the application is new, we need support from younger members."	Male, chairperson, Hai Phong

[a]ISHC: Intergenerational Self-Help Clubs.

Strengths in the Use of the Noncommunicable Disease Screening App for Health Volunteers and Intergenerational Self-Help Clubs

The majority of the answers focused on advanced functions of the NCD screening app (eg, better data storage, faster and easier reporting to HAIV, more time efficiency, and saving money; 112 responses). The following are building capacity for ISHC health volunteers in using technology (26 responses), improving the health of ISHC members through connecting the internet for learning skills and doing exercise (26 responses), increasing the reputation of club management board members and ISHC health volunteers with their club members (13 responses), strengthening relationships with stakeholders (6 responses), and others (9 responses). These advantages could explain the great results of data entry by ISHC health volunteers.

Barriers and Limitations in the Use of the Noncommunicable Disease Screening App for Health Volunteers and Intergenerational Self-Help Clubs

Most of the answers focused on app issues (eg, a few parts of the app are still in English, typographical errors, bugs or errors, few parts of the display screen are covered by the keyboard, many steps to register screening data, and hard to add new members; 67 responses) and limited capacity of ISHC health volunteers or older people in technology (67 responses). The following includes tablet issues (small font size; 13 responses), irregular technical support after initial training due to COVID-19 restrictions (6 responses), and other issues (eg, Wi-Fi connection problems and changes in the personnel who is in charge of the tablet; 15 responses). These barriers and limitations could be the reasons of certain data that are missing and few errors that are recorded in quantitative results.

Recommendations for Improving the Use of the Noncommunicable Disease Screening App

The main recommendations recorded from the qualitative analysis are continuous capacity building (34 responses), improving app issues (27 responses), and improving tablet issues (24 responses). The following includes involving relevant stakeholders (7 responses), involving younger members in tablet use or technology adoption to support older people (7 responses), and others (13 responses; eg, financial support, more measurement devices for the screening, and supporting internet connection or internet packages).

Discussion

Principal Results

This study aimed to evaluate the strengths, limitations, opportunities, and barriers of the NCD screening app used by ISHC health volunteers, and to provide recommendations for future use of the digital NCD screening app in community settings. Quantitative analysis revealed that nearly one-third of the risk factor data were missing and there were a few errors in data entry. However, the NCD screening app demonstrated

XSL•FO
RenderX

significant advantages, including easier data storage, better space and time savings, more efficient submissions of reports to higher levels, and improved security compared with paper forms. The key barriers to using the app included the advanced age of some volunteers, lack of previous experience, and low digital literacy among ISHC health volunteers.

The quantitative data analysis showed that the ISHC volunteers generally filled in all required fields and there were few obvious errors in the data entry. The main source of missing data was the recording of risk factors. As indicated by FGDs with health volunteers, a possible explanation for this is that ISHC health volunteers may have only checked the boxes when a risk factor had been identified and skipped the question if no risk factors were identified. This could be attributed to a misunderstanding of how to fill in the questionnaire, which was occasionally noted during the training sessions by HAIV staff. Other reasons, identified as key findings of the qualitative evaluation of the use of the NCD screening app, include the limited digital literacy of older health volunteers, poor vision, the small keyboard, errors in the app, and irregular technical support during the COVID-19 pandemic. These factors likely contributed to data entry mistakes or unintentional data omissions by health volunteers.

It is important to note here that, initially, in most ISHCs, data were first entered manually in registers and later entered electronically in the tablet. The ISHC health volunteers did not want to slow down the screening process, as they were more familiar with paper-based systems. Supervisors in several instances assisted with entering the data. There is no record of the frequency of assistance by supervisors. Therefore, our finding that health volunteers were able to enter the screening data completely and correctly may be influenced by the support of supervisors during the screening.

The percentage of females who participated in the qualitative evaluations was much higher in the first evaluation (22/29, 76%) compared with the second evaluation (25/44, 57%) and the percentage of people with college or university diplomas in the second evaluation (16/44, 36%) was higher than in the first evaluation (4/29, 14%). This is partly to be expected as the participants of the first evaluation mostly consisted of ISHC health volunteers, club chairpersons (most of the club members were female [70%, as per ISHC's regulation]), and community members with lower levels of education. The second evaluation included more stakeholders from higher levels including representatives from the Association of the Elderly at all levels and representatives from CHSs were included. It is aligned with data from the quantitative survey with 75.8% (2642/3485) of screened people being female.

The qualitative analysis has shown that adoption of digital technology at the community level brings many benefits to health volunteers, ISHCs, and the local Association of the Elderly. ISHC health volunteers found that the NCD screening app helps them more easily store data, saves both space and time, facilitates sending reports to higher levels, is more secure compared with paper forms, and automatically calculates scores. The analysis of data using the dashboard provides faster in-depth information. The advantages of the NCD screening app make the work easier for ISHC health volunteers, which is aligned with previous studies [24,25]. The Vietnamese government has been moving toward increased digital transformation, particularly digital transformation in health care [23,26,27]; therefore, the use of digital technology at the community level is aligned with the country's strategy, helping health volunteers and ISHCs adopt IT knowledge which is vital so that they are not left behind [28]. Furthermore, the analysis of the data collected with the NCD screening app increases awareness of NCD symptoms leading to better self-care, more frequent visits to the health facilities and potentially contributing to improved overall health of ISHC members.

The most frequently mentioned barrier during the interviews was the health volunteers' old age with an average age of 63.8 years. Furthermore, many ISHC health volunteers and club management boards lacked experience with tablets, smartphones, the internet, and other IT knowledge [25,29]. As shared by the ISHC health volunteers during FGDs due to their perceived low digital literacy, they initially preferred to continue using paper-based records. This can be partially explained by the quantitative findings which indicated some errors or missing data. Although HAIV provided additional training, support, and e-learning tools, these efforts did not always seem sufficient.

A notable difference was observed between rural and urban ISHCs in their experience with the NCD screening app. Urban health volunteers and ISHCs more frequently mentioned the app's strengths likely due to their greater familiarity with smartphones and internet. In contrast, rural ISHCs reported more limitations. Learning IT was not a priority for many rural health volunteers who were primarily focused on agriculture, fishing, and family care. Some rural ISHC health volunteers mentioned that they had never used a smartphone. Digital literacy remains a significant challenge in the introduction of the community-based app, requiring more time and training than is typically allocated in projects [30]. Once the health volunteers become confident in using the NCD screening app, they can enter data on the spot and transition away from paper registers.

In addition, several issues related to the screening app were recorded, such as a non–age-friendly app, confusing display, incorrect translation of medical terminology, language barriers knowledge [25], small font size on the tablet, and so on. Furthermore, internet-based software requires an internet connection to upload the data to the server, which was a problem for some rural ISHC health volunteers with poor internet connections. These findings are similar to those of a study in Bangladesh [31].

The short implementation time due to the COVID-19 pandemic was recorded as one of the barriers where participants mentioned irregular technical support after the initial training for ISHC health volunteers and club management boards and a pause in project activities due to pandemic restrictions.

Health systems in low and middle-income countries are slow in adopting technology at the community level [24,25,32,33]. This may lead to limitations in sharing and using data at higher levels, especially the use of data for planning and policy-making [31]. These findings are aligned with our study because

XSL•FO
RenderX

interviewees from the national level mentioned that the NCD screening app we tested cannot be easily used in existing PHC systems.

An additional function of the NCD screening app that can be used for ISHCs in Vietnam is the monthly report. In ISHCs, where the functions of the app were better used, it helped club management boards save time and space to store reports, making reports easier and faster to be sent to higher levels of the Association of the Elderly. The ISHCs can use the dashboards in the app for improved monitoring and planning of activities. In the long run, expanding the NCD screening app with monthly reporting indicators can completely replace paper forms and reduce the workload for management boards and ISHCs [29].

Study Strengths and Limitations

A strength of this research is the use of convergent parallel design, combining quantitative data with detailed insights from qualitative data. With the use of semistructured interviews or FGDs, the experience of the users of the app was examined in depth.

There are a few limitations that should be discussed. With parallel implementation, some qualitative findings cannot explain or answer questions and issues from quantitative findings. In some instances, corrections of wrong data entries may have been made immediately after the screening sessions, when data were transferred from paper-based to digital systems. Furthermore, supervisors in several instances assisted in entering data. This may have led to a more positive picture of data quality.

In the first round of qualitative evaluation, there was only 1 note-taker and no data cross-check, which may have led to data loss. Due to the presence of HAIV staff, representatives of the Provincial Association of the Elderly, and the local Association of the Elderly, the interviewees may have had a bias in their responses in both evaluation rounds. In addition, in the second evaluation round data may have been lost during the translation process.

Implications

Applying the NCD screening app in NCD intervention through ISHCs in Ninh Binh and Hai Phong in Vietnam in the SUNI-SEA project has provided multiple lessons for the use of digital technology at the community level.

To strengthen the use of the tablet NCD screening app among health volunteers or ISHCs, simplifying the app, fixing errors, converting English text into Vietnamese, and redesigning it to be an age-friendly app that is easier for older people to use is necessary. Involving users, especially end users (local people or older people), in the design stage would help the design to be more user-friendly and aligned to the needs of the community and older people.

Providing continuous capacity-building opportunities, longer implementation time of the project, longer team support and commitment, allotting more time for practice sessions, adding separate general internet technology sessions for ISHC health volunteers and older people, provision of regular technical support, and recruiting younger members to join ISHCs and health volunteer groups are recommended for ISHC health volunteers and old-aged–groups. These recommendations from our study are consistent with the suggestions documented in other studies, which suggest that additional training and ongoing technical support should be continued to improve knowledge and efficiency with any update of the DHIS2 tracker [25,29,33].

To leverage the linkage between ISHC or the Association of the Elderly and the health sector, the use of screening data by health sectors and authorities, integrating the app into the existing health information system is recommended to provide an official referral system and use of the data for planning and decision-making. DHIS2 tracker is an open-source software with good interoperability, and linking this community-based information system to the PHC system should be the next step.

For the health sector, it would be beneficial if the referral mechanism is systematic and official from ISHC to PHCs and vice versa. With this mechanism, PHCs can diagnose NCDs earlier, provide proper treatment for patients, and better manage NCDs at the community level. Adopting technology at the community level is also consistent with the health system strategy and digital transformation of the Vietnamese government [26,27], while maximizing the use of local resources and support from health volunteers and ISHCs as a contribution to the NCD prevention and management programs in the context where there is a lack of resources within the health sector [34].

Adopting the NCD screening app in ISHCs is to help health volunteers and communities have an easier way and more effective to track and manage NCD screening data, especially with high-risk members. Using the NCD screening app can generate quality data that can be used for analysis, advocacy, and decision-making for the health sector in NCD programs across the country.

Conclusion

Our findings show that the NCD screening app brings a lot of benefits for health volunteers and ISHCs, where all ISHC health volunteers and club management boards are voluntarily donating their time and ability to contribute to NCD prevention and management at the community level, including managing screening data for their club members.

Adopting technology for NCD prevention and management at the community level is not an easy process. It requires the involvement of several stakeholders from the grassroots to higher levels in order to maintain sustainability and effective data use in NCD prevention and management, especially for the health sector in strategic planning and decision-making at higher levels.

In summary, the implementation of the NCD screening app at the community level in Vietnam can be an effective way to improve health screening for ISHCs and community members and NCD prevention and management for the health sector. However, our study has shown that there are still various barriers that limit the ISHC health volunteers who use the NCD screening app to record, report, and manage screening data in the app correctly.

XSL•FO
RenderX

Acknowledgments

The authors gratefully acknowledge the contributions for the draft manuscript by Mary Dunbar from HelpAge International (based in Chiang Mai, Thailand).

The manuscript would not be completed without qualitative evaluation, which is conducted by Evi Willemsen, Vera Poelman, Storm D. Vries, Susanne Reiling, Silke Wolfswinkel, Anna Boonstra, and Romy V. Dijk, medical students from the University of Groningen, the Netherlands.

We thank Nguyen Thao My, Hoang Phuong Nhung, and Dinh Hong Chung from HelpAge International in Vietnam, Hanoi, Vietnam for supporting the qualitative data collection and monitoring data in the NCD screening application recorded by ISHC health volunteers.

We thank Mary Muturi, University of Groningen, University Medical Center Groningen, Department of Health Sciences, Netherlands for her support with manuscript formatting.

We thank Vietnam Association of the Elderly, Ninh Binh Association of the Elderly, Hai Phong Association of the Elderly, ISHC health volunteers, club management board, local Association of the Elderly, and Commune Health Stations from Ninh Binh Province and Hai Phong City, Vietnam participated in the evaluations.

Scaling-Up NCD Interventions in South-East Asia (SUNI-SEA) was implemented by a research consortium in Indonesia, Myanmar, and Vietnam. This research into the use of the mobile data collection application was a part of the project and funded by HelpAge International.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Intergenerational Self-help Club Model.
[PNG File , 130 KB - jmir_v26i1e54127_app1.png]

Multimedia Appendix 2
Noncommunicable disease screening process.
[PNG File , 46 KB - jmir_v26i1e54127_app2.png]

Multimedia Appendix 3
The synergy approach of the SUNI-SEA (Scaling up non-communicable disease interventions in Southeast Asia) project.
[PNG File , 77 KB - jmir_v26i1e54127_app3.png]

Multimedia Appendix 4
Categories of findings and answers from interviewees (full Table 2).
[DOCX File , 21 KB - jmir_v26i1e54127_app4.docx]

References

1. World Health Organization. Noncommunicable diseases. Geneva: World Health Organization; 2023. URL: https://tinyurl.com/5n6s2unz [accessed 2024-08-01]
2. Noncommunicable Diseases (NCD) Country Profiles, 2018 - Vietnam. Word Health Organization. 2018. URL: https://www.who.int/publications/i/item/ [accessed 2024-10-23]
3. Liu J, Yu W, Zhou J, Yang Y, Chen S, Wu S. Relationship between the number of noncommunicable diseases and health-related quality of life in Chinese older adults: a cross-sectional survey. Int J Environ Res Public Health 2020;17(14):5150 [FREE Full text] [doi: 10.3390/ijerph17145150] [Medline: 32708844]
4. Decision 376 /Q?-TTg: National Strategy prevention of cancer, cardiovascular, diabetes, chronic obstructive pulmonary disease, asthma and other non-communicable diseases, the period 2015 - 2025. 2015. URL: https://tinyurl.com/bde2ahau [accessed 2024-10-23]
5. Nguyen TNT, Nguyen TTT, Tran BQ, Pham CT, Perry KE, Haregu T, et al. Putting non-communicable disease data to work in Vietnam: an investigation of community health surveillance capacity. BMC Public Health 2023;23(1):321 [FREE Full text] [doi: 10.1186/s12889-023-14986-4] [Medline: 36788519]

XSL•FO
RenderX

6. World Health Organization. Hypertension Viet Nam 2023 country profile. Hypertension Viet Nam 2023 country profile. Geneva: WHO Team - Surveillance, Monitoring and Reporting (SMR); 2023 Sep 19. URL: https://tinyurl.com/52k4345z [accessed 2024-10-05]
7. The growing burden of diabetes in Viet Nam. World Health Organization. 2016. URL: https://tinyurl.com/2s4bwcxn [accessed 2023-06-22]
8. Scaling-up NCD interventions in South-East Asia (SUNI-SEA). SUNI-SEA Project. URL: https://www.suni-sea.org/vi/ [accessed 2023-06-22]
9. The intergenerational self-help club (ISHC) development model. Asia Health and Wellbeing Initiative. 2020. URL: https://ahwin.org/helpage-vietnam-ishc/ [accessed 2024-10-23]
10. Pardoel ZE, Reijneveld SA, Lensink R, Widyaningsih V, Probandari A, Stein C, et al. Core health-components, contextual factors and program elements of community-based interventions in Southeast Asia - a realist synthesis regarding hypertension and diabetes. BMC Public Health 2021;21(1):1917 [FREE Full text] [doi: 10.1186/s12889-021-11244-3] [Medline: 34686171]
11. The Urban agenda platform. Hanoi's Intergenerational Self-Help Clubs. 2022. URL: https://tinyurl.com/yc85c45m [accessed 2024-10-23]
12. HelpAge International in Vietnam. Intergenerational Self-Help Club: Effective model adapting to population aging. 2022. URL: https://helpagevietnam.org/hoat-dong/cau-lac-bo-lien-the-he-tu-giup-nhau/ [accessed 2023-06-22]
13. The Vietnamese Government. Decision No. 1336/QD-TTg of the Prime Minister: Approving the project to replicate the model of Intergenerational Self-Help Club for the period up to 2025. 2020. URL: https://tinyurl.com/5hcm6apa [accessed 2024-10-23]
14. Vietnam Ministry of Health. The Prime Minister approved the Health Care Program for the Elderly until 2030. The Prime Minister approved the Health Care Program for the Elderly until 2030. Hanoi: Vietnam Ministry of Health; 2020 Oct 15. URL: https://tinyurl.com/5dhu9xpp [accessed 2024-09-23]
15. The Vietnamese Government. Decision No. 2156/QD-TTg of the Prime Minister: Approving the National Action Program on the elderly for the period 2021-2030. 2021. URL: https://tinyurl.com/c2dwuy37 [accessed 2024-10-23]
16. Early J, Gonzalez C, Gordon-Dseagu V, Robles-Calderon L. Use of mobile health (mHealth) technologies and interventions among community health workers globally: a scoping review. Health Promot Pract 2019;20(6):805-817. [doi: 10.1177/1524839919855391] [Medline: 31179777]
17. Lam JA, Dang LT, Phan NT, Trinh HT, Vu NC, Nguyen CK. Mobile health initiatives in Vietnam: scoping study. JMIR Mhealth Uhealth 2018;6(4):e106 [FREE Full text] [doi: 10.2196/mhealth.8639] [Medline: 29691214]
18. Individual Data Records with Tracker. DHIS2. 2022. URL: https://dhis2.org/tracker/ [accessed 2023-06-22]
19. HelpAge International in Vietnam. Tutorial video on the use of DHIS2 tracker. 2022. URL: https://www.youtube.com/watch?v=HeKdDiPzGRo [accessed 2023-06-22]
20. A set of educational materials on diabetes and hypertension in Vietnamese. SUNI-SEA Project. 2023. URL: https://www.suni-sea.org/en/resources/a-set-of-educational-materials-on-diabetes-and-hypertension-in-vietnamese/ [accessed 2023-06-22]
21. Xiong S, Lu H, Peoples N, Duman EK, Najarro A, Ni Z, et al. Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries. NPJ Digit Med 2023;6(1):12 [FREE Full text] [doi: 10.1038/s41746-023-00764-4] [Medline: 36725977]
22. Lim HM, Chia YC, Koay ZL. Performance of the Finnish Diabetes Risk Score (FINDRISC) and modified Asian FINDRISC (ModAsian FINDRISC) for screening of undiagnosed type 2 diabetes mellitus and dysglycaemia in primary care. Prim Care Diabetes 2020;14(5):494-500. [doi: 10.1016/j.pcd.2020.02.008] [Medline: 32156516]
23. The Vietnamese Government. Circular No. 54/2017/TT-BYT dated December 29, 2017 of the Ministry of Health on criteria for assessment of information technology application at health facilities. 2017. URL: https://tinyurl.com/ys35rwe6 [accessed 2024-10-23]
24. Dehnavieh R, Haghdoost A, Khosravi A, Hoseinabadi F, Rahimi H, Poursheikhali A, et al. The District Health Information System (DHIS2): a literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. Health Inf Manag 2019;48(2):62-75. [doi: 10.1177/1833358318777713] [Medline: 29898604]
25. Bakibinga P, Kamande E, Kisia L, Omuya M, Matanda DJ, Kyobutungi C. Challenges and prospects for implementation of community health volunteers' digital health solutions in Kenya: a qualitative study. BMC Health Serv Res 2020;20(1):888 [FREE Full text] [doi: 10.1186/s12913-020-05711-7] [Medline: 32957958]
26. The Vietnamese Government. Decision 749/QD-TTg on National digital transformation Program to 2025, Orientation to 2030. 2023. URL: https://tinyurl.com/4hv8z3dk [accessed 2024-10-23]
27. YCP Solidiance. A Look Forward: How Digitalization is Transforming Vietnam?s Healthcare System. Ho Chi Minh City. 2020. URL: https://tinyurl.com/yxf4mn9u [accessed 2024-10-23]
28. Right of older person ? Older people are part of the Millennium Development Goals agenda to ensure that nobody is left behind. The Ministry of Health of Vietnam. 2015. URL: https://tinyurl.com/mpasxsez [accessed 2023-06-22]
29. Youssef D, Yaghi A, Jouny A, Abou-Abbas L, Chammaa H, Ghosn N. Converting the existing disease surveillance from a paper-based to an electronic-based system using district health information system (DHIS-2) for real-time information: the Lebanese experience. BMC Health Serv Res 2022;22(1):395 [FREE Full text] [doi: 10.1186/s12913-022-07773-1] [Medline: 35337327]

XSL•FO
RenderX

30. Kemp E, Trigg J, Beatty L, Christensen C, Dhillon HM, Maeder A, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Austr 2021;32 Suppl 1:104-114. [doi: 10.1002/hpja.387] [Medline: 32681656]
31. Begum T, Khan SM, Adamou B, Ferdous J, Parvez MM, Islam MS, et al. Perceptions and experiences with district health information system software to collect and utilize health data in Bangladesh: a qualitative exploratory study. BMC Health Serv Res 2020;20(1):465 [FREE Full text] [doi: 10.1186/s12913-020-05322-2]
32. Leon N, Schneider H, Daviaud E. Applying a framework for assessing the health system challenges to scaling up mHealth in South Africa. BMC Med Inform Decis Mak 2012;12:123 [FREE Full text] [doi: 10.1186/1472-6947-12-123] [Medline: 23126370]
33. Karuri J, Waiganjo P, Daniel O, Manya A. DHIS2: the tool to improve health data demand and use in Kenya. J health informatics dev ctries 2014;8(1):1-23 [FREE Full text]
34. Duong DB, Minh HV, Ngo LH, Ellner AL. Readiness, availability and utilization of rural vietnamese health facilities for community based primary care of non-communicable diseases: a crosssectional survey of 3 provinces in Northern vietnam. Int J Health Policy Manag 2019;8(3):150-157 [FREE Full text] [doi: 10.15171/ijhpm.2018.104] [Medline: 30980631]

Abbreviations

CHS: commune health station
DHIS2: District Health Information Software, version 2
FGD: focus group discussion
FINDRISC: Finnish Diabetes Risk Score
HAIV: HelpAge International in Vietnam
IDI: in-depth interview
ISHC: Intergenerational Self-Help Club
NCD: noncommunicable disease
PHC: primary health care
SUNI-SEA: Scaling up non-communicable disease interventions in Southeast Asia

Edited by X Ma; submitted 03.11.23; peer-reviewed by N Baharom, D Podder; comments to author 03.07.24; revised version received 14.08.24; accepted 23.09.24; published 02.12.24.

Please cite as:
Nong TTT, Nguyen GH, Lepe A, Tran TB, Nguyen LTP, Koot JAR
Challenges and Opportunities in Digital Screening for Hypertension and Diabetes Among Community Groups of Older Adults in Vietnam: Mixed Methods Study
J Med Internet Res 2024;26:e54127
URL: https://www.jmir.org/2024/1/e54127
doi:10.2196/54127
PMID:

©Trang Thi Thu Nong, Giang Hoang Nguyen, Alexander Lepe, Thuy Bich Tran, Lan Thi Phuong Nguyen, Jaap A R Koot. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predicting and Monitoring Symptoms in Patients Diagnosed With Depression Using Smartphone Data: Observational Study

Arsi Ikäheimonen[1], MSc; Nguyen Luong[1], MSc; Ilya Baryshnikov[2,3], MD, PhD; Richard Darst[4], PhD; Roope Heikkilä[5], MA; Joel Holmen[6,7], MD, PhD; Annasofia Martikkala[2,3], MD; Kirsi Riihimäki[3,8], MD, PhD; Outi Saleva[3], MD; Erkki Isometsä[2,3], MD, PhD; Talayeh Aledavood[1], PhD

[1]Department of Computer Science, Aalto University, Espoo, Finland
[2]Department of Psychiatry, University of Helsinki, Helsinki, Finland
[3]Helsinki and Uusimaa Hospital District, Helsinki, Finland
[4]School of Science, Aalto University, Espoo, Finland
[5]City of Helsinki Mental Health Servcies, Helsinki, Finland
[6]University of Turku, Turku, Finland
[7]Turku University Central Hospital, Turku, Finland
[8]Finnish Institute for Health and Welfare, Helsinki, Finland

Corresponding Author:
Arsi Ikäheimonen, MSc
Department of Computer Science
Aalto University
Konemiehentie 2
Espoo, 02150
Finland
Phone: 358 449750110
Email: arsi.ikaheimonen@aalto.fi

Abstract

Background: Clinical diagnostic assessments and the outcome monitoring of patients with depression rely predominantly on interviews by professionals and the use of self-report questionnaires. The ubiquity of smartphones and other personal consumer devices has prompted research into the potential of data collected via these devices to serve as digital behavioral markers for indicating the presence and monitoring of the outcome of depression.

Objective: This paper explores the potential of using behavioral data collected with smartphones to detect and monitor depression symptoms in patients diagnosed with depression. Specifically, it investigates whether this data can accurately classify the presence of depression, as well as monitor the changes in depressive states over time.

Methods: In a prospective cohort study, we collected smartphone behavioral data for up to 1 year. The study consists of observations from 164 participants, including healthy controls (n=31) and patients diagnosed with various depressive disorders: major depressive disorder (MDD; n=85), MDD with comorbid borderline personality disorder (n=27), and major depressive episodes with bipolar disorder (n=21). Data were labeled based on depression severity using 9-item Patient Health Questionnaire (PHQ-9) scores. We performed statistical analysis and used supervised machine learning on the data to classify the severity of depression and observe changes in the depression state over time.

Results: Our correlation analysis revealed 32 behavioral markers associated with the changes in depressive state. Our analysis classified patients who are depressed with an accuracy of 82% (95% CI 80%-84%) and change in the presence of depression with an accuracy of 75% (95% CI 72%-76%). Notably, the most important smartphone features for classifying depression states were screen-off events, battery charge levels, communication patterns, app usage, and location data. Similarly, for predicting changes in depression state, the most important features were related to location, battery level, screen, and accelerometer data patterns.

Conclusions: The use of smartphone digital behavioral markers to supplement clinical evaluations may aid in detecting the presence and changes in severity of symptoms of depression, particularly if combined with intermittent use of self-report of symptoms.

(J Med Internet Res 2024;26:e56874) doi:10.2196/56874

XSL•FO
RenderX

KEYWORDS

data analysis; digital phenotyping; digital behavioral data; depression symptoms; depression monitoring; mHealth; mobile health; smartphone; mobile phone

Introduction

In recent years, digital tools and algorithms have become indispensable in health care, including mental health. Data-driven technologies have the potential to renew health care, providing new avenues for personalized care, remote monitoring, and improved service access. At the same time, mental health disorders, including depression, have remained a significant concern. Depressive disorders are estimated to be the second-leading cause of life-years lost to disability worldwide [1]. Alongside markedly impacting individuals' quality of life, depressive disorders impose a substantial economic burden, including costs to health care and societies overall due to disability, reduced employment, and impaired work productivity [2].

Psychiatric evaluations are based on clinical interviews, relying on patients' self-reflections and recollections, which are susceptible to memory biases and subjective inaccuracies [3]. Further, the absence of definitive physiological biomarkers for mental disorders complicates accurate diagnoses and treatment [4]. Given these challenges, a growing interest has been in data-driven clinical monitoring and decision-making, supplementing subjective evaluations with objective, longitudinal, physiological, and behavioral data collected via digital devices [5]. This approach, known as digital phenotyping, involves creating a digital representation of a patient's clinical phenotype using behavioral, social, and physiological markers. The premise of the data-driven approach lies in the inherent value of continuous monitoring, uncovering valuable insights unattainable through intermittent assessments [4].

Recent data-driven studies using devices like smartphones and activity trackers have effectively used digital behavioral data to monitor and detect participants' depressive moods [6-8]. These studies gather sensor data to identify behavioral patterns associated with depressive disorders, such as changes in physical activity, phone usage, and sleep routines. The primary goals include differentiating between patients with depression and healthy controls, classifying mood state transitions, and predicting future mood states. Alongside passively collected data, these studies often use established self-report questionnaires as the reference standard for subjects' severity of depressive symptoms.

However, some of the studies have used limited data collection, sample sizes of fewer than 50 participants [9-11], a sample of college students [12-15], and data collected over only a few weeks [9,12,16]. Due to these limitations, it may be challenging to generalize results to either a broader population or a free-living setting. Regarding methodologies, earlier research has used smartphone sensors and data categorized as smartphone usage [9,10,14,15], GPS location data-based features [9-15,17], physical activity data or step counts [11-17], communication patterns [12,14,17], Bluetooth data [13,14], sleep data [13,15], metrics for behavior regularity [15], and physiological measurements [17]. Furthermore, studies have used several metrics for depression as the ground truth, including the 9-item Patient Health Questionnaire (PHQ-9) [9-11,17,18], a compact version of the 4-item Patient Health Questionnaire (PHQ-4) [15,19], the Montgomery and Åsberg Depression Rating Scale [16,20], and the Beck Depression Inventory-II [13,14,21]. The analysis methods used in these studies vary, encompassing correlation analysis [9,10,12], machine learning [11,13,14,16,17], and deep learning [15,16].

This paper builds on previous research, exploring the potential of using behavioral data collected with smartphones to detect and monitor depression symptoms in outpatients diagnosed with depression. Our study aims to identify digital behavioral markers indicative of depressive states and assess the accuracy of this data in detecting depression. Key markers extracted from smartphone sensors, such as the accelerometer, app usage, battery status, communication log, screen activations, and GPS location, comprise metrics like screen-on activation count, total distance traveled, average battery level, phone call count, app usage duration, and maximum acceleration. We analyzed a comprehensive dataset, gathered through smartphones, from patients with depression who have a diagnosis of either major depressive disorder (MDD), major depressive episodes with bipolar disorder (MDE|BD), or MDD with comorbidborderline personality disorder (MDD|BPD) and healthy controls. The focus was on distinguishing patients self-reporting moderate or more severe depression symptoms and tracking changes in reported depression levels.

Methods

Dataset Description

We used the data from the Mobile Monitoring of Mood (MoMo-Mood) study, a 1-year multimodal digital phenotyping study of individuals undergoing treatment for mental disorders and healthy controls [22,23]. The MoMo-Mood study recruited 164 participants from 4 different groups: healthy controls (n=31) and patients with MDD (n=85), MDD|BPD (n=27), and MDE|BD (n=21). Voluntary patients were recruited in Finland from the mood disorder outpatient treatment facilities of the Helsinki University Hospital Mood Disorder Division, Turku University Central Hospital Department of Psychiatry, and City of Espoo Mental Health Services. The patients were diagnosed with structured interviews, namely the Mini-International Neuropsychiatric Interview [24] and the Structured Clinical Interview for DSM-IV Axis II Personality Disorders [25], as having ongoing major depressive episodes. Healthy controls were collected by contacting, via email, lists of students from the University of Helsinki and Aalto University, users of student health services from these institutions, and recruiting voluntary health care personnel from Helsinki University Hospital.

Each group had more female individuals than male individuals: (1) control group, 24 female individuals and 7 male individuals; (2) MDD group, 46 female individuals and 31 male individuals;

XSL•FO
RenderX

(3) MDE|BD group, 18 female individuals and 3 male individuals; and (4) MDD|BPD group, 23 female individuals and 1 male individual. On average, control group participants were older than patient group participants, with average ages as follows: (1) control group, 41.8 (SD 13.9) years, (2) MDD group, 39.0 (SD 14.2) years, (3) MDE|BD group, 37.1 (SD 10.3) years, and (4) MDD|BPD group, 28.3 (SD 6.0) years. A more detailed description is provided elsewhere [22,23].

Study participants were recruited on a rolling basis, allowing them to join and leave the research at various intervals. They were requested to stay involved in the study for 1 year. Data collection was carried out in 2 phases. In the initial 2 weeks, called the *active phase*, participants collected data continuously via personal devices (smartphones running Android operating system), bed sensors, and actigraphs, and they answered daily mood-related questions. The active phase was followed by the *passive phase*, lasting up to 1 year. During the passive phase, data collection via smartphones continued, and participants' depression was monitored by biweekly PHQ-9 surveys prompted via the smartphone. The PHQ-9 comprises 9 questions, each scored from 0 to 3, based on the frequency of depressive symptoms over the past 2 weeks. Thus, the total score ranges from 0 to 27, with high values representing more severe depression. The passive data originate from various smartphone sensors, including accelerometers, app usage, communication, battery level and screen status logs, and GPS location data. The data were collected through the Niima data collection platform [26]. This work exclusively focuses on the passive phase of the study, which uses smartphone data and PHQ-9 survey responses. This phase was selected due to its unobtrusive data collection methods and, thus, the minimal requirement for behavioral adjustment from the participants.

Data Preprocessing

We used Python and the *Niimpy* behavioral data analysis toolbox [27] for data preprocessing. We extracted 93 behavioral features from the raw data. Multimedia Appendix 1 provides a detailed description of data sources and extracted features. Furthermore, we segmented the data from the accelerometer, app usage, battery status, communication log, and smartphone screen activations into 6-hour bins (12:00 AM to 06:00 AM, 6:00 AM to 12:00 PM, 12:00 PM to 6:00 PM, and 6:00 PM to 12:00 AM). We extracted 308 additional features, resulting in a total of 401 features. The data from different sensors were resampled and averaged over 14-day periods. The data were merged with the PHQ-9 responses to align data from the preceding biweekly period with the questionnaire responses. Of the 164 participants in the active phase, 99 proceeded to the passive phase. For the analysis, we selected participants who had provided passive data for at least 14 days and had answered a PHQ-9 survey at the end of this period, yielding 83 participants. Each participant provided data for up to 1 year, yielding 26 biweekly data points. Due to missing observations and participant withdrawals from the study, 818 observations (37.9% of the possible 2158 observations) were available for further analysis. Figure 1 details the data collection and preparation for the analysis, while Figure 2 provides additional information on data aggregation and alignment.

Figure 1. The MoMo-Mood study data collection and preparation schema. MoMo-Mood: Mobile Monitoring of Mood; PHQ-9: 9-item Patient Health Questionnaire.

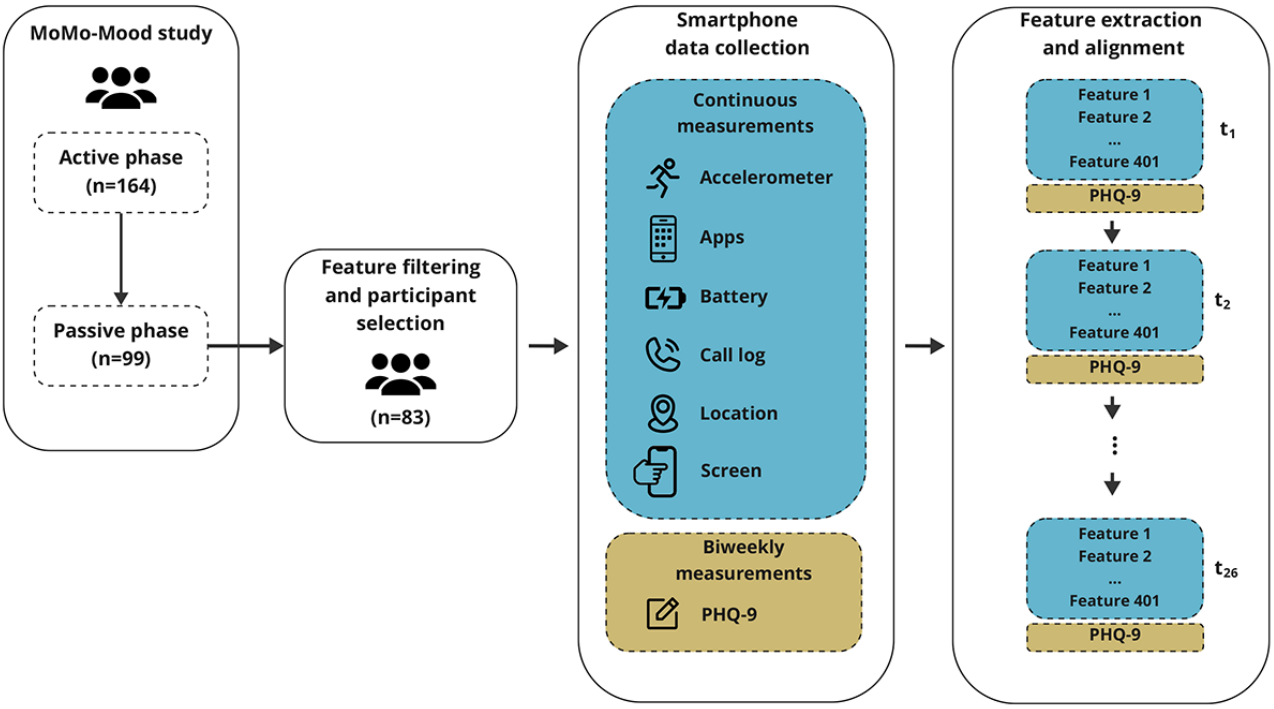

XSL•FO
RenderX

Figure 2. Schematics of data aggregation and alignment. PHQ-9: 9-item Patient Health Questionnaire.

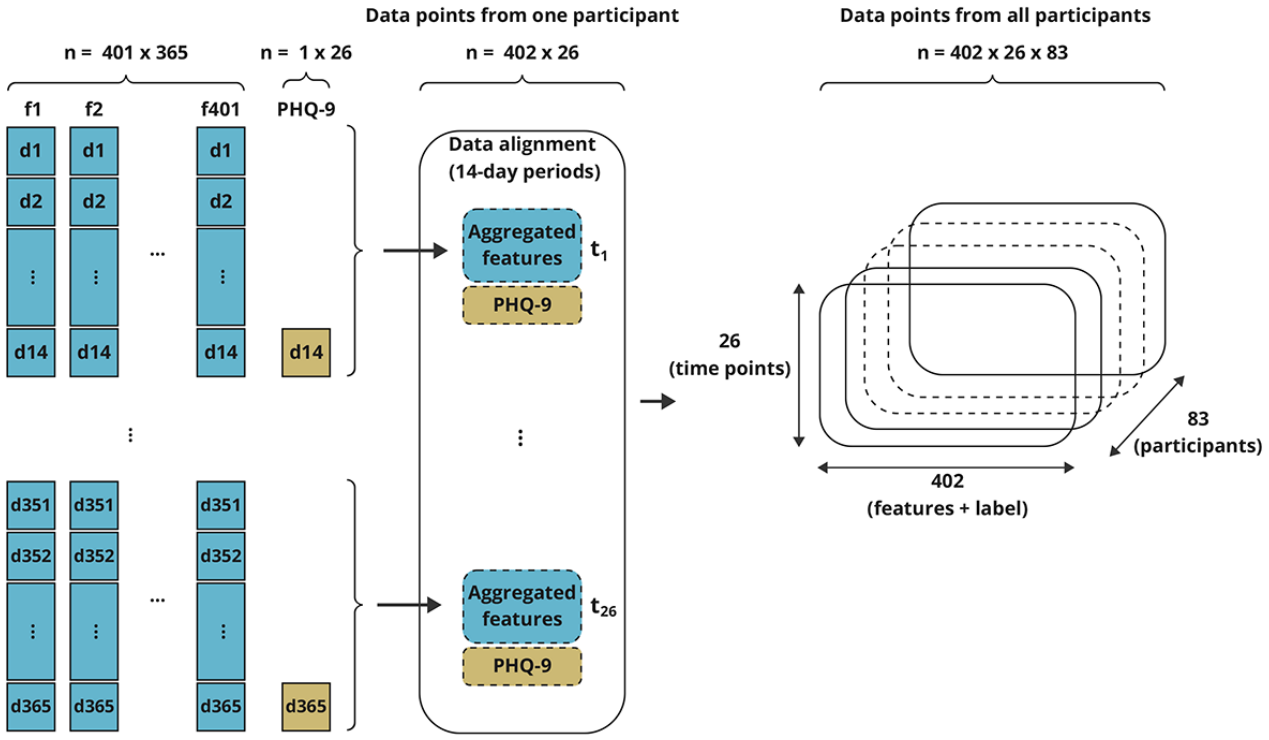

Statistical Analysis

Distributional Testing

To examine whether passively collected smartphone sensor data show differences between patient groups and control participants, we used distributional testing using the nonparametric 2-sample Kolmogorov-Smirnov test [28]. The test was chosen due to its capability to detect variations across the entire distribution, including the tails. For the test, we averaged the biweekly sampled data by participant, normalized the data, and omitted the missing values. For robustness against the risk of type I errors (false-positive) due to multiple comparisons, we implemented false discovery rate (FDR) correction using the Benjamini-Hochberg procedure [29] at a significance level of α=.05.

Correlation Analysis

We conducted a correlation analysis to assess the association between passive data features and PHQ-9 scores. We pooled passive data from all participants, omitted missing values, and applied the Spearman rank correlation coefficient to assess the strength of the relationship. Furthermore, we used FDR correction using the Benjamini-Hochberg procedure at a significance level of α=.05 to account for the multiple testing involved, controlling the expected proportion of false discoveries.

Predictive Modeling

To achieve the research goal, we deployed supervised machine learning models for predicting both the presence of depression and state transitions of depressive states. We used a cutoff PHQ-9 depression score of 10 for binary classification analyses. Scores of 10 or higher were considered *depressed*, and scores below 10 as *nondepressed*. We chose a cutoff value of 10 because it signifies clinical depression, typically warranting a treatment plan that may include counseling, follow-up sessions, and possibly pharmacotherapy for the individual. For defining the depression state transition, we used the same threshold of 10 and the previous depression state. Each transition is paired with a specific label, used as the target variable for the depression state transition modeling. The transition definitions are presented in Table 1.

Table 1. Overview of depression state transition definitions and corresponding labels.

Transition description	Label
Depressed→Depressed	Remains depressed
Depressed→Nondepressed	Improves
Nondepressed→Nondepressed	Remains nondepressed
Nondepressed→Depressed	Declines

We built a machine learning pipeline using Python (version 3.10.8) and the following libraries: *scikit-learn* (version 1.2.0) [30], *extreme gradient boosting* (*XGBoost*; version 1.7.3) [31], *OPTUNA* (version 3.1.0) [32], *imbalance-learn* (version 0.11.0) [33], and *Shapley additive explanations* (*SHAP*; version 0.41.0) [34]. Initially, we partitioned our dataset into a 75%:25%

XSL•FO
RenderX

train:test split, preventing data leakage by keeping each participant's data exclusively in either the training or test set. We conducted feature prefiltering by removing features with no or low variance, many missing observations, and a high correlation with other features. We compared filtering and wrapper-based methods and embedded feature selection methods with XGBoost classifiers for feature selection. We used data missingness, variance, and cross-correlation thresholding-based feature selection for the filtering-based method and the sequential forward selection method for the wrapper-based method. Standard preprocessing was applied to selected features, comprising imputation using median values, scaling transformations, and data normalization. To address the class imbalance and improve the robustness of our classification models against overfitting to the majority class, we used the synthetic minority oversampling technique (SMOTE) [35], a method for generating synthetic minority class samples to balance the dataset. We applied SMOTE at 2 stages of the model's development. First, the training data folds were balanced using SMOTE during the cross-validation process. We then applied SMOTE to the entire training dataset in preparation for the final model fitting. For pipeline details, refer to Multimedia Appendix 2.

In our study, we focused on the prediction task of identifying (1) the presence and (2) the state transitions of depression symptoms using passively sensed smartphone data and supervised machine learning models. Figure 3 outlines how data was used for prediction. Specifically, 3 models were examined, namely k-nearest neighbors, support vector classifier, and XGBoost, all of which are commonly used models in digital phenotyping studies [36,37]. To fine-tune feature filtering, transformation functions, classification model, and SMOTE hyperparameters, we used stratified grouped 5-fold cross-validation, using the OPTUNA framework [32]. The primary objective in the hyperparameter optimization process was to maximize the F_1-score, which balances precision and recall, thereby ensuring a more reliable evaluation of model performance. We used a pruning early stopping technique, which ceases training if there is no improvement in the F_1-score (our chosen validation metric). Finally, we used the test data and bootstrapping validation (using 10,000 bootstrap samples from training data) to evaluate the model performance, assessing the performances with accuracy, precision, recall, negative predictive value (NPV), and F_1-scores, as defined in Multimedia Appendix 3. F_1-score is a valuable metric because maximizing it ensures that both false positives (identifying a participant who is nondepressed as depressed) and false negatives (failing to identify a participant who is depressed) are minimized. High recall reflects low false-negative classification, so we emphasized its importance in model performance evaluation.

Figure 3. Schema for depression presence and transition prediction using passive behavioral data. An asterisk (*) depicts a model using the PHQ-9 measurement from the preceding biweekly period as a predictor. Time point t0 on the analysis timeline represents the active phase, and points from t1 to t26 represent the passive phase. PHQ-9: 9-item Patient Health Questionnaire.

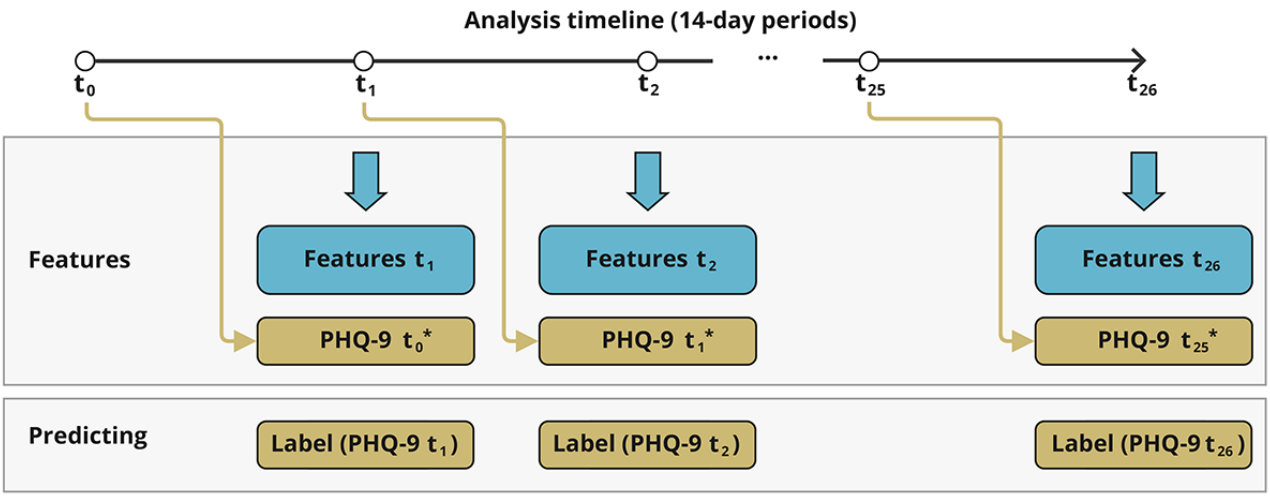

Measuring Feature Importance

For the final part, focusing on model interpretation, we assessed the importance of features (behavioral markers) for the best-performing XGBoost models to gain insight into the underlying classification mechanisms of the model. We evaluated the importance of each feature for depression presence and the state transition classifications using SHAP values [34]. SHAP values measure each feature's contribution to the model prediction, their relative importance compared with other features, and the significance of feature interactions.

Ethical Considerations

The Helsinki and Uusimaa Hospital District's Ethics Committee approved the research protocol for the MoMo-Mood study (approval number § 125/2018). A research permit was granted by Helsinki and Uusimaa Hospital District Psychiatry. This covered data streams, data collection platform security, and participants' consent information. All data in transit were encrypted, and participant privacy was a key design value. Local IT support and the research ethics committee approved the written data security statement. Study participants were presented with full study information and data collected prior to providing written consent. The participants were allowed to withdraw whenever they chose. As remuneration for their

XSL•FO
RenderX

participation, participants received 4 movie tickets at the end of the initial phase of the study.

Results

Descriptive Statistics

The raw data from the passive collection phase contained over 67 million data points, and 819 biweekly PHQ-9 surveys gathered data from 99 participants from 4 subgroups: 25 healthy controls, 46 patients with MDD, 16 patients with MDD|BPD, and 12 patients with MDE|BD. Participant selection, filtering, and preprocessing reduced the raw data to 327,200 data points (818 observations with 401 data features) and PHQ-9 scores to 818 observations. The resulting dataset had 83 participants, comprising 20 healthy controls, 41 patients with MDD, 12 patients with MDD|BPD, and 10 patients with MDE|BD.

PHQ-9 Scores

Most of the patients' PHQ-9 scores during the passive data collection phase remained within the range of 5-19, representing mild to moderate clinical depression, while most control scores remained within the range of 0-4, representing no depression. The group-wise mean scores over the passive phase were as follows: control group, 1.2 (SD 1.8); MDD group, 11.9 (SD 6.7); MDE|BD group, 13.7 (SD 6.5); and MDD|BPD group, 13.8 (SD 6.6). It is noteworthy that the patient group scores predominantly represent mild to moderately severe clinical depression. Figure 4 [38] presents these differences and the distribution of PHQ-9 scores across the various groups.

On average, PHQ-9 scores remain at similar levels within patient groups throughout the study, while all patient groups express a slightly decreasing trend at the beginning of the study. At the group level, MDE|BD and MDD|BPD groups exhibited more fluctuation in the scores toward the end of the study period as the number of participants within those groups decreased. Control group scores exhibited a slightly decreasing trend. Figure 5 shows the overall trends in PHQ-9 scores, averaged over each group throughout the study. It is worth noting that the number of participants decreased over time, leading to increased fluctuations in average scores.

Figure 4. PHQ-9 score distributions for control and patient groups.To assess differences in PHQ-9 scores across various groups, we used a generalized estimating equations approach. We chose the method due to its effectiveness in dealing with correlated response data and its ability to provide robust SEs. The analysis revealed statistically significant differences in PHQ-9 scores between the control group and each of the patient groups. The significance of these differences was high, with $P<.001$ for each comparison. MDD: major depressive disorder; MDD|BPD: major depressive disorder with comorbid borderline personality disorder; MDE|BD: major depressive episodes with bipolar disorder; PHQ-9: 9-item Patient Health Questionnaire.

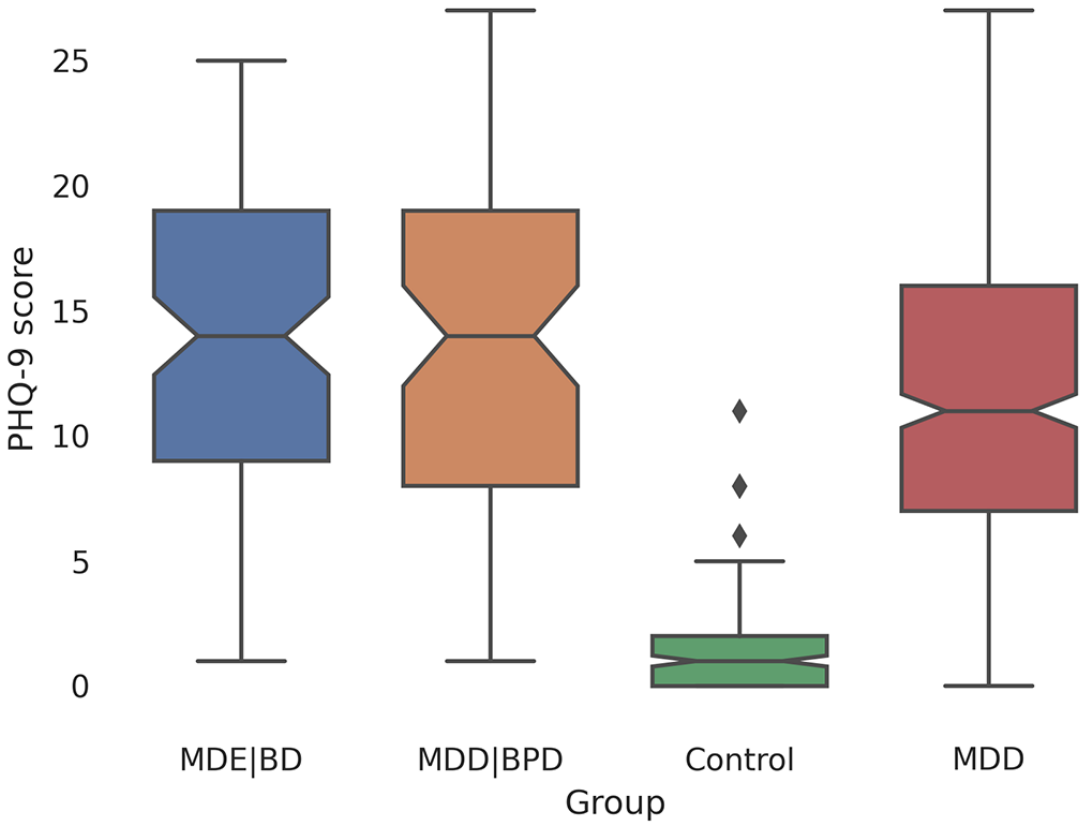

XSL•FO
RenderX

Figure 5. Averaged PHQ-9 score trends for controls and patient groups (standard deviations depicted by shaded regions). MDD: major depressive disorder; MDD|BPD: major depressive disorder with comorbid borderline personality disorder; MDE|BD: major depressive episodes with bipolar disorder; PHQ-9: 9-item Patient Health Questionnaire.

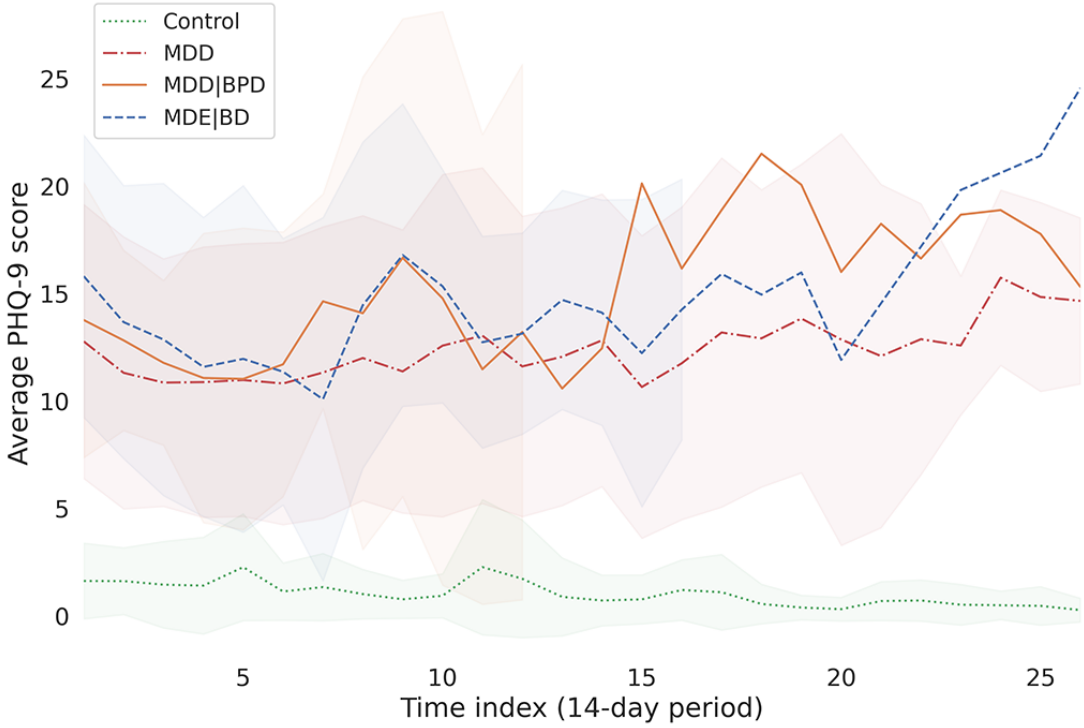

We compared the groups by depression severity by categorizing participants using a cutoff PHQ-9 score threshold of 10. Participants with a PHQ-9 score of 10 or higher were categorized as "Depressed," and those below 10 were categorized as "Nondepressed." Table 2 shows the prevalence of depression severity across different patient groups.

We categorized the 818 data points into 2 groups: 347 (42.4%) participants who are depressed and 471 (57.6%) participants who are nondepressed, resulting in mildly imbalanced classes considering the classification tasks. We assessed biweekly depression state transitions for each group, as described in Table 1. Table 3 summarizes these transitions. Notably, the number of transitions for "Declines" and "Improves" was significantly lower than those for "Remains Depressed" and "Remains nondepressed."

These results show that in the data, the state changes in depression are infrequent compared with the occurrences where the state remains the same. Here, we noticed that transition classes have a significant imbalance, as only 119 (14.5%) out of 818 state changes counted as transitions, and 699 (85.5%) were stationary. This pronounced imbalance could bias classification algorithms toward the majority class, necessitating corrective measures for reliable analysis in subsequent stages.

Table 2. Distribution of PHQ-9a scores by severity and group (n=818).

Group	Control	MDE|BD[b]	MDD|BPD[c]	MDD[d]	Total
Depression severity, n (%)					
Depressed	1 (0.1)	65 (7.9)	50 (6.1)	231 (28.2)	347 (42.4)
Nondepressed	204 (24.9)	36 (4.4)	24 (2.9)	207 (25.3)	471 (57.6)
Total	205 (25.1)	101 (12.3)	74 (9)	438 (53.5)	818 (100)

[a]PHQ-9: 9-Item Patient Health Questionnaire.

[b]MDE|BD: major depressive episodes with bipolar disorder.

[c]MDD|BPD: major depressive disorder with comorbid borderline personality disorder.

[d]MDD: major depressive disorder.

XSL•FO
RenderX

Table 3. Depression state transition counts for the control group and each patient group.

Group	Control	MDE|BD[a]	MDD|BPD[b]	MDD[c]	Total
Transition, n					
Declines	1	9	5	34	49
Improves	1	15	11	43	70
Remains depressed	0	57	42	222	320
Remains nondepressed	203	20	17	139	379
Total	205	101	74	438	818

[a]MDE|BD: major depressive episodes with bipolar disorder.

[b]MDD|BPD: major depressive disorder with comorbid borderline personality disorder.

[c]MDD: major depressive disorder.

Data Completeness

Participant compliance and, thus, data completeness decreased as the study's passive phase progressed. PHQ-9 survey answer compliance dropped below 70% after 6 weeks (3 biweekly periods) had passed, and after that, it continued to decline steadily. For further details, refer to Multimedia Appendix 4. Passive data collection compliance shows a pattern similar to answering the PHQ-9 survey. Most of the missing data occurred due to the participant dropping out of the study, while some participants had gaps in data collection. Notably, only a few participants remained in the study for the entire year. Also, the data collection for participants was incomplete due to missing features.

Statistical Analysis

Two-sample distributional testing using a 2-sample Kolmogorov-Smirnov test identified 20 significant features (5%), with P values ranging from .0045 to .0497. However, after applying the FDR correction for multiple comparisons at a significance level =.05, none of these features were statistically significant; thus, we found no evidence for patient group behavioral data differing from control data. For further details, see Table S1 in Multimedia Appendix 5.

Correlation analysis between the behavioral features and PHQ-9 scores using Spearman ranked correlation and FDR correction for multiple comparisons at significance level =.05 resulted in 32 (8%) out of 401 features exhibiting statistically significant correlations. The majority (18/32, 56%) of the correlations were very weak (absolute value from 0 to 0.19), and the rest (14/32, 44%) were weak (absolute value from 0.2 to 0.39). For more information, refer to Table S2 in Multimedia Appendix 5.

Depression Presence Classification

We used 2 distinct approaches for classifying the presence of depression. The initial approach treated all biweekly aggregated passive data features (aligned with corresponding biweekly PHQ-9 scores) as independent observations. Using the XGBoost classifier with filter-based feature selection, we achieved the highest accuracy of 66% (95% CI 56%-70%) and an F_1-score of 0.66 (95% CI 0.5-0.7). The performance comparison of various classifiers and feature selection methods is detailed in Table S1 in Multimedia Appendix 6, while Table S2 in Multimedia Appendix 6 provides a comprehensive summary of the model's performance.

For the second modeling approach, we included the measured PHQ-9 score from the previous biweekly period as a predictor in the model. Model performance improves notably after adding the predictor. XGBoost classifier with a filtering-based feature selection method achieved the best accuracy of 82% (95% CI 80%-84%) and a corresponding F_1-score of 0.82 (95% CI 0.80-0.85) across the test data of 208 samples.

This classifier outperformed the other classifiers (k-nearest neighbor and support vector classifier) by a small margin. The "Nondepressed" class (99 samples) achieved a precision of 0.80 and recall of 0.83, with an NPV of 0.84 and an F_1-score of 0.81, reflecting balanced performance. The "Depressed" class (109 samples) had a slightly higher precision of 0.84, a recall of 0.81, an NPV of 0.80, and an F_1-score of 0.82, indicating a similar level of predictive accuracy to the "Nondepressed" class. Both macro- and weighted averages across precision, recall, F_1-score, and NPV are 0.82, demonstrating consistent performance in detecting both the presence and absence of depression. Table S3 in Multimedia Appendix 6 summarizes the performance of selected classifiers and feature selection methods. Table 4 summarizes the XGBoost classifier's performance, and Figure S1 in Multimedia Appendix 6 presents the receiver operating characteristic curve for the classifier.

XSL•FO
RenderX

Table 4. XGBoost[a] model performance metrics for depression presence classification (using previous PHQ-9[b] score as a predictor).

Metric	Precision	Recall	NPV[c]	F_1-score	Support, n
Class					
Nondepressed	0.80	0.83	0.84	0.81	99
Depressed	0.84	0.81	0.80	0.82	109
Averages					
Macroaverage	0.82	0.82	0.82	0.82	208
Weighted average	0.82	0.82	0.82	0.82	208

[a]XGBoost: extreme gradient boosting.

[b]PHQ-9: 9-item Patient Health Questionnaire.

[c]NPV: negative predictive value.

Depression State Transition Classification

For depression state transition classification, we used the XGBoost classifier with feature filtering since it performed best in the depression presence classification. The model was able to classify relatively well the cases where a participant's state remains the same, while the accuracy is considerably lower for cases where the state changes. Applying SMOTE's synthetic oversampling technique to alleviate class imbalance significantly increased the recall of the minority classes (depression transitions for "Declines" and "Increases"). The model achieved an accuracy of 75% (95% CI 72%-76%) and a corresponding F_1-score of 0.67 (95% CI 0.63-0.69). Table 5 summarizes the model validation results for each transition type. For the transition of "Declines," the model shows high NPV (0.98) but lower precision (0.34). It indicates that while the model reliably identifies cases where the state will not decline, it is less accurate at correctly identifying the cases where it declines. The recall is 0.76, leading to an F_1-score of 0.47, signifying unbalanced classification performance. "Increases" shows a similar pattern with high NPV (0.96) and moderate recall (0.74) but lower precision (0.46), resulting in an F_1-score of 0.57, also indicating unbalanced classification performance. For the "Remains Depressed" and "Remains nondepressed" states, the model exhibits higher precision (0.93 and 0.95, respectively) and NPV (0.86 and 0.83, respectively), along with recall rates of 0.72 and 0.77, leading to a more balanced performance with F_1-scores of 0.81 and 0.85. The macroaverage F_1-score of 0.67, compared with the overall accuracy of 0.75, reflects the effect of class imbalance on the model's performance. Further, Figures S2 and S3 in Multimedia Appendix 6 display classification outcomes for the test data and a multiclass receiver operating characteristic curve for the XGBoost classifier.

The results show the model's ability to classify most cases correctly. With an overall accuracy of 75%, the model effectively balances precision across different cases. These findings demonstrate the model's potential for predicting depression state transitions, leveraging smartphone-sensed behavioral data and self-reported PHQ-9 scores.

Table 5. XGBoost[a] model performance metrics for depression state transition classification (using previous PHQ-9[b] score as a predictor).

Metric	Precision	Recall	NPV[c]	F_1-score	Support, n
Transition					
Declines	0.34	0.76	0.98	0.47	17
Increases	0.46	0.74	0.96	0.57	23
Remains depressed	0.93	0.72	0.86	0.81	74
Remains nondepressed	0.95	0.77	0.83	0.85	94
Averages					
Macroaverage	0.67	0.75	0.91	0.67	208
Weighted average	0.84	0.75	0.87	0.77	208

[a]XGBoost: extreme gradient boosting.

[b]PHQ-9: 9-Item Patient Health Questionnaire.

[c]NPV: negative predictive value.

Feature Importance Analysis Using SHAP Values

In our analysis of feature importance for classification of the presence of depression and depression state transition, we evaluated the relative significance of different features by examining the SHAP values in the best-performing XGBoost models. In summary, our findings highlight the previous PHQ-9 score as the most impactful feature when included in the model. For depression presence classification, additional significant features include smartphone screen status, app usage, battery

XSL•FO
RenderX

level, phone call, and location–related information. In addition to the previous PHQ-9 score for state transition classification, screen status, location–, battery level– and accelerometer–related features stand out as the most important. Conversely, app usage and communication–related features had a limited impact on the models.

The importance of the previous PHQ-9 score implies that the depression scores are autocorrelated, thus reflecting future depression levels. Smartphone screen status (e.g., screen on and off event counts) reveals users' interaction with the device, showing usage frequency and patterns. Similarly, battery level indicates phone usage, reflecting battery drains and charges. App usage features (especially apps labeled as leisure, sports, and social media) suggest behavioral patterns related to such activities as watching movies or listening to music, exercising, and communicating via social media. Finally, accelerometer-related features reveal physical activity and mobility patterns.

Figures S1-S3 in Multimedia Appendix 7 present the most important features of these classifications. Specifically, Figure S1 in Multimedia Appendix 7 illustrates the important features of depression presence classification without considering the previous biweekly PHQ-9 scores. Conversely, Figure S2 in Multimedia Appendix 7 shows the results for the model, including these scores as a predictor. Finally, Figure S3 in Multimedia Appendix 7 explores features pertinent to depression state transition classification.

Discussion

Principal Findings

Our analysis encompassed passively sensed digital behavioral data, which we compared against actively collected PHQ-9 survey data. Using the generalized estimating equation approach, we discovered a statistically significant difference in PHQ-9 score distributions between the control and patient groups. It is important to note that some patients likely experienced recovery after recruitment for the study, potentially lessening the severity of symptoms reflected in their PHQ-9 scores. Consequently, our data could underrepresent the depression severity spectrum, particularly among patients with more severe depression.

After adjusting for multiple comparisons, distributional testing on behavioral features revealed no significant differences between control and patient groups. This finding suggests that the differences in behavioral data at the group level are minimal. Therefore, our study implies that detecting these subtle differences might require larger sample sizes or alternative statistical methodologies that can leverage hierarchical structures and temporal correlations.

Correlation analysis identified 32 behavioral features with weak or very weak correlations with PHQ-9 scores, predominantly involving smartphone screen interaction (18 features) and accelerometer data (14 features). Despite most features showing no significant correlation with PHQ-9 scores, their potential value in classification tasks remains, especially considering possible nonlinear relationships or interactions with other features.

For the depression prediction tasks, we found that the XGBoost classifier with filtering-based feature selection performed the best in discriminating between participants who are depressed and nondepressed, achieving 66% accuracy. The accuracy increased to 82% when we added the PHQ-9 score from the previous biweekly period as a predictor. The difference implies the importance of the temporal structure of the data. Therefore, we propose to include temporal information in future analyses to improve the accuracy. Further, for clinical monitoring applications, information about participants' depression histories should be available, providing the temporal context necessary to enhance the model's predictive accuracy.

Furthermore, our results show that the XGBoost classifier, combined with filter-based feature selection and PHQ-9 measurement from the previous biweekly monitoring period as a predictor, can differentiate mood state transitions with a classification accuracy of 75%. While promising, this accuracy level suggests room for further improvement in the model's performance. Like the depression presence classification, we suggest using more comprehensive methods, personalized models, and temporal information. Additionally, we suspect that the data's limited sample size and sparsity of transition events hinder the classification performance. Therefore, model development should benefit from a larger sample.

Finally, feature importance analysis revealed insights into the key features of depression prediction models. The most significant predictor for detecting and classifying depression presence was previous biweekly PHQ-9 scores, complemented by features related to accelerometer, app usage, battery level, location, and screen events. The results emphasize the significance of daily behavioral patterns and time-of-day distinctions (morning, afternoon, evening, and nighttime) in accurately predicting depression. Interestingly, some features were identified by both the correlation and feature importance analyses for classifier models. While the methods and objectives of these analyses differ, the consistency in identifying the same key features across both approaches implies their potential relevance in depression prediction.

Comparison With Previous Studies

Our study aligns methodologically with previous research using validated depression assessments and analyzing passively collected smartphone behavioral features. Also, it focuses on statistical inference and machine learning techniques to classify depression among participants and distinguish participants based on behavioral data. Additionally, the identified important features are consistent with earlier research reporting features related to phone usage [9,10,14,15], physical activity [11-17], and location data [9-15,17]. By contrast, the importance of features related to communication [12,14,17] were slightly underrepresented in our analysis.

Our classification results are numerically comparable to previous studies using machine learning methods with smartphone data for depression detection. Using a cohort of college students, Chikersal et al [14] achieved an 85% accuracy and an F_1-score of 0.82 in the postsemester depression detection task. They also achieved an 85% accuracy and an F_1-score of 0.80 in detecting

XSL•FO
RenderX

a change in the depression state task. Similarly, Wang et al [15] used machine learning and deep learning models to detect depression using a subset of smartphones, also from a cohort of college students, achieving an F_1-score of 0.65 using a machine learning model and an F_1-score of 0.7 using deep learning.

However, our study differentiates itself by including a diverse cohort of real outpatients, clinically diagnosed with structured interviews, alongside control participants, thereby offering a broader perspective on depression. Additionally, the data are collected over an extended period in a naturalistic setting, enhancing the reliability of the findings. Unlike other studies that often focus on student populations, it demonstrates the feasibility of digital behavioral monitoring in real outpatients. Furthermore, it excludes certain data features like physiological measurements and social engagement metrics. Lastly, the study does not aim to predict future depressive states, setting it apart from other predictive modeling efforts in the field.

Limitations

While this research yields insightful outcomes, it is crucial to acknowledge certain limitations. First, dropouts and missing data increased substantially after the first 3 biweekly periods. Some participants provided data only for 1 biweekly period, limiting our model's ability to capture patient symptom fluctuations. Second, our analysis does not fully account for the hierarchical and autocorrelational structure of the data. We rely on simplified analysis, using aggregated features and pooled participants, resulting in the loss of available information. Finally, our study does not accommodate external factors that might impact the participants' behavior patterns and mood states. Given that the data collection partially took place during the COVID-19 era, factors such as social isolation could have played a role in changing the behavior patterns and emotional states of participants.

Recommendations for Future Work

This study lays the groundwork for multiple future research endeavors. A direct expansion of our work would be the implementation of personalized models designed to predict the depression state of individuals. These personalized models, which incorporate both group and participant variations and sample-level information, have demonstrated improved accuracy in depression classification tasks [39]. Furthermore, we recommend fully using the temporal structure of the data in classification tasks. Given the inherent variability in symptomatic periods among patients with depression, analyzing temporal patterns and trends from longitudinal data could offer a more accurate representation of their evolving mental states than single-point estimates. We also encourage the exploration of deep learning models in future studies, as these models tend to surpass conventional machine learning methods in predictive accuracy [15,16]. However, due to their complexity and less clear interpretability relative to more traditional methods, we suggest not starting with these models at the outset, instead gradually incorporating them into the analysis. Lastly, to address the challenges posed by the unbalanced dataset in our study, we suggest collecting additional data to enhance the robustness and generalizability of future research findings.

Conclusions

In summary, this study demonstrates the potential of using smartphone-sensed behavioral data for monitoring depression symptoms, thereby paving the way for personalized and more effective mental health care. The results contribute to an expanding body of evidence supporting the integration of data-driven methods into mental health services. These insights may complement and enhance clinical practices, supplementing conventional diagnostic and monitoring approaches.

Acknowledgments

We acknowledge the support and input from Jesper Ekelund regarding the study design and data collection. The computational resources provided by the Aalto Science-IT project are gratefully acknowledged. We thank Yunhao Yuan for providing support and feedback for the visualizations.

Data Availability

Due to the fact that the data are highly sensitive, the collected data cannot be shared with researchers outside of our consortium. Our research permit does not allow the free availability of these data.

Authors' Contributions

AI contributed to writing – original draft, writing – review & editing, visualization, methodology, formal analysis, validation, data curation, and conceptualization. NL contributed to writing – original draft, review & editing, methodology, and data curation. IB contributed to investigation, methodology, and writing – review & editing. RD contributed to investigation, data collection software, data curation, and writing – review & editing. RH contributed to investigation and methodology. JH contributed to investigation. AM contributed to investigation and writing – review & editing. KR contributed to investigation and methodology. OS contributed to investigation. EI contributed to writing – review & editing, investigation, project administration, investigation, methodology, supervision, funding acquisition, and conceptualization. TA contributed to writing – review & editing, supervision, project administration, investigation, methodology, supervision, funding acquisition, and conceptualization.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of data sources and extracted features for each sensor.
[DOCX File , 8 KB - jmir_v26i1e56874_app1.docx]

Multimedia Appendix 2
The machine learning pipeline for depression presence and state transition classification.
[DOCX File , 114 KB - jmir_v26i1e56874_app2.docx]

Multimedia Appendix 3
Definitions for key performance metrics accuracy, precision, recall, negative predictive value, and F1-score.
[DOCX File , 9 KB - jmir_v26i1e56874_app3.docx]

Multimedia Appendix 4
The percentage of participants providing PHQ-9 data over time by group. PHQ-9: 9-item Patient Health Questionnaire.
[DOCX File , 11 KB - jmir_v26i1e56874_app4.docx]

Multimedia Appendix 5
Kolmogorov-Smirnov test results for behavioral data distribution differences and Spearman rank correlation test results between behavioral data features and PHQ-9 scores. PHQ-9: 9-item Patient Health Questionnaire.
[DOCX File , 14 KB - jmir_v26i1e56874_app5.docx]

Multimedia Appendix 6
Performance comparison of depression presence classification models. Performance metrics for depression presence classification: nondepressed versus depressed. Comparative performance of depression presence classification models with biweekly PHQ-9 score as a predictor. ROC curves for depression presence classification and different depression transition classifications. Confusion matrix for depression state transition classification. PHQ-9: 9-item Patient Health Questionnaire; ROC: receiver operating characteristic.
[DOCX File , 371 KB - jmir_v26i1e56874_app6.docx]

Multimedia Appendix 7
Key features in depression presence XGBoost classification model based on SHAP values. Key features in depression presence classification XGBoost model with preceding biweekly PHQ-9 score as a predictor. Key features in depression state transition classification XGBoost model with previous PHQ-9 score. PHQ-9: 9-item Patient Health Questionnaire; SHAP: Shapley additive explanations; XGBoost: extreme gradient boosting.
[DOCX File , 175 KB - jmir_v26i1e56874_app7.docx]

References

1. World mental health report: transforming mental health for all. World Health Organization. 2022 Jun 16. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-01-26]
2. Health TLG. Mental health matters. Lancet Glob Health 2020;8(11):e1352 [FREE Full text] [doi: 10.1016/S2214-109X(20)30432-0] [Medline: 33069297]
3. Nelson B, McGorry PD, Wichers M, Wigman JTW, Hartmann JA. Moving from static to dynamic models of the onset of mental disorder: a review. JAMA Psychiatry 2017;74(5):528-534. [doi: 10.1001/jamapsychiatry.2017.0001] [Medline: 28355471]
4. Torous J, Kiang MV, Lorme J, Onnela J. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
5. Hsin H, Fromer M, Peterson B, Walter C, Fleck M, Campbell A, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. NPJ Digit Med 2018;1:37. [doi: 10.1038/s41746-018-0046-0] [Medline: 31304319]
6. Maatoug R, Oudin A, Adrien V, Saudreau B, Bonnot O, Millet B, et al. Digital phenotype of mood disorders: a conceptual and critical review. Front Psychiatry 2022 Jul 26;13:895860 [FREE Full text] [doi: 10.3389/fpsyt.2022.895860] [Medline: 35958638]

XSL•FO
RenderX

7. Leaning IE, Ikani N, Savage HS, Leow A, Beckmann C, Ruhé HG, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024;158:105541 [FREE Full text] [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
8. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]
9. Saeb S, Zhang M, Kwasny MM, Karr CJ, Kording K, Mohr DC. The relationship between clinical, momentary, and sensor-based assessment of depression. Int Conf Pervasive Comput Technol Healthc 2015;2015:e4273 [FREE Full text] [doi: 10.4108/icst.pervasivehealth.2015.259034] [Medline: 26640739]
10. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537 [FREE Full text] [doi: 10.7717/peerj.2537] [Medline: 28344895]
11. Masud MT, Mamun MA, Thapa K, Lee DH, Griffiths MD, Yang SH. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. J Biomed Inform 2020;103:103371 [FREE Full text] [doi: 10.1016/j.jbi.2019.103371] [Medline: 31935462]
12. Boukhechba M, Daros AR, Fua K, Chow PI, Teachman BA, Barnes LE. DemonicSalmon: monitoring mental health and social interactions of college students using smartphones. Smart Health 2018;9-10:192-203. [doi: 10.1016/j.smhl.2018.07.005]
13. Xu X, Chikersal P, Dutcher JM, Sefidgar YS, Seo W, Tumminia MJ, et al. Leveraging collaborative-filtering for personalized behavior modeling. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021;5(1):1-27. [doi: 10.1145/3448107]
14. Chikersal P, Doryab A, Tumminia M, Villalba DK, Dutcher JM, Liu X, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing. ACM Trans Comput Hum Interact 2021;28(1):1-41. [doi: 10.1145/3422821]
15. Wang W, Nepal S, Huckins JF, Hernandez L, Vojdanovski V, Mack D, et al. First-gen lens: assessing mental health of first-generation students across their first year at college using mobile sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(2):1-32 [FREE Full text] [doi: 10.1145/3543194] [Medline: 36561350]
16. Jakobsen P, Garcia-Ceja E, Riegler M, Stabell LA, Nordgreen T, Torresen J, et al. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. PLoS One 2020;15(8):e0231995 [FREE Full text] [doi: 10.1371/journal.pone.0231995] [Medline: 32833958]
17. Mullick T, Radovic A, Shaaban S, Doryab A. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. JMIR Form Res 2022;6(6):e35807 [FREE Full text] [doi: 10.2196/35807] [Medline: 35749157]
18. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
19. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ–4. Psychosomatics 2009;50(6):613-621. [doi: 10.1016/s0033-3182(09)70864-3]
20. Müller MJ, Himmerich H, Kienzle B, Szegedi A. Differentiating moderate and severe depression using the montgomery-asberg depression rating scale (MADRS). J Affect Disord 2003;77(3):255-260. [doi: 10.1016/s0165-0327(02)00120-9] [Medline: 14612225]
21. Beck AT, Steer RA, Ball R, Ranieri W. Depression inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
22. Baryshnikov I, Aledavood T, Rosenström T, Heikkilä R, Darst R, Riihimäki K, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: a prospective ecological momentary assessment study on 80 psychiatric outpatients. J Affect Disord 2023;324:170-174 [FREE Full text] [doi: 10.1016/j.jad.2022.12.127] [Medline: 36586594]
23. Aledavood T, Luong N, Baryshnikov I, Darst R, Heikkilä R, Holmén J, et al. Mobile monitoring of mood (MoMo-Mood): a multimodal digital phenotyping study with major depressive patients and healthy controls. JMIR Preprints Preprint posted online on June 26, 2024. [doi: 10.2196/preprints.63622]
24. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59(Suppl 20):quiz 34-quiz 57. [Medline: 9881538]
25. First MB, Benjamin LS, Gibbon M, Spitzer RL, Williams JB. Structured Clinical Interview for DSM-IV Axis II Personality Disorders. Washington, DC: American Psychiatric Press; 1997:1919-1920.
26. Aledavood T, Hoyos AMT, Alakörkkö T, Kaski K, Saramäki J, Isometsä E, et al. Data collection for mental health studies through digital platforms: requirements and design of a prototype. JMIR Res Protoc 2017;6(6):e110 [FREE Full text] [doi: 10.2196/resprot.6919] [Medline: 28600276]
27. Ikäheimonen A, Triana AM, Luong N, Ziaei A, Rantaharju J, Darst R, et al. Niimpy: a toolbox for behavioral data analysis. SoftwareX 2023;23:101472. [doi: 10.1016/j.softx.2023.101472]
28. Berger VW, Zhou Y. Kolmogorov-Smirnov test: overview. In: Wiley StatsRef: Statistics Reference Online. New York, NY: John Wiley & Sons, Ltd; 2014.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

XSL•FO
RenderX

30. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
31. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD -16. Association for Computing Machinery; August 13, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
32. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a Next-generation hyperparameter optimization framework. 2019 Presented at: 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; July 25, 2019; Anchorage, AK p. 2623-2631. [doi: 10.1145/3292500.3330701]
33. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017;18(17):1-5 [FREE Full text]
34. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
36. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. Harv Rev Psychiatry 2020;28(5):296-304. [doi: 10.1097/hrp.0000000000000268]
37. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020;22(5):e16875 [FREE Full text] [doi: 10.2196/16875] [Medline: 32348284]
38. Hardin JW, Hilbe JM. Generalized Estimating Equations: GEE. 2nd ed. New York, NY: Chapman & Hall/CRC; 2013.
39. Palmius N, Saunders KEA, Carr O, Geddes JR, Goodwin GM, de Vos M. Group-personalized regression models for predicting mental health scores from objective mobile phone data streams: observational study. J Med Internet Res 2018;20(10):e10194 [FREE Full text] [doi: 10.2196/10194] [Medline: 30348626]

Abbreviations

FDR: false discovery rate
MDD: major depressive disorder
MDD|BPD: major depressive disorder with comorbid borderline personality disorder
MDE|BD: major depressive episodes with bipolar disorder
MoMo-Mood: Mobile Monitoring of Mood
NPV: negative predictive value
PHQ-4: 4-item Patient Health Questionnaire
PHQ-9: 9-item Patient Health Questionnaire
SMOTE: synthetic minority over-sampling technique
SHAP: Shapley additive explanations
XGBoost: extreme gradient boosting

Edited by T de Azevedo Cardoso; submitted 29.01.24; peer-reviewed by R Zhang, AS John, A AL-Asadi, J Kallio; comments to author 21.06.24; revised version received 07.07.24; accepted 24.09.24; published 03.12.24.

Please cite as:
Ikäheimonen A, Luong N, Baryshnikov I, Darst R, Heikkilä R, Holmen J, Martikkala A, Riihimäki K, Saleva O, Isometsä E, Aledavood T
Predicting and Monitoring Symptoms in Patients Diagnosed With Depression Using Smartphone Data: Observational Study
J Med Internet Res 2024;26:e56874
URL: https://www.jmir.org/2024/1/e56874
doi:10.2196/56874
PMID:

©Arsi Ikäheimonen, Nguyen Luong, Ilya Baryshnikov, Richard Darst, Roope Heikkilä, Joel Holmen, Annasofia Martikkala, Kirsi Riihimäki, Outi Saleva, Erkki Isometsä, Talayeh Aledavood. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Dynamic Bidirectional Associations Between Global Positioning System Mobility and Ecological Momentary Assessment of Mood Symptoms in Mood Disorders: Prospective Cohort Study

Ting-Yi Lee[1], MEd; Ching-Hsuan Chen[1,2], MD, MSc; I-Ming Chen[3], MD, PhD; Hsi-Chung Chen[3,4], MD, PhD; Chih-Min Liu[3], MD, PhD; Shu-I Wu[5,6], MD, PhD; Chuhsing Kate Hsiao[1,7], PhD; Po-Hsiu Kuo[1,3,8], PhD

[1]Department of Public Health and Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

[2]Department of Obstetrics and Gynecology, Taipei City Hospital Heping Fuyou Branch, Taipei, Taiwan

[3]Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan

[4]Department of Psychiatry, Center of Sleep Disorders, National Taiwan University Hospital, Taipei, Taiwan

[5]Department of Psychiatry, Mackay Memorial Hospital, Taipei, Taiwan

[6]Department of Medicine, MacKay Medical College, New Taipei City, Taiwan

[7]Institute of Health Data Analytics and Statistics, College of Public Health, National Taiwan University, Taipei, Taiwan

[8]Psychiatric Research Center, Wan Fang Hospital, Taipei, Taiwan

Corresponding Author:
Po-Hsiu Kuo, PhD
Department of Public Health and Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University
Room 521
No 17, Xuzhou Road
Taipei, 10055
Taiwan
Phone: 886 2 33668015
Email: phkuo@ntu.edu.tw

Abstract

Background: Although significant research has explored the digital phenotype in mood disorders, the time-lagged and bidirectional relationship between mood and global positioning system (GPS) mobility remains relatively unexplored. Leveraging the widespread use of smartphones, we examined correlations between mood and behavioral changes, which could inform future scalable interventions and personalized mental health monitoring.

Objective: This study aims to investigate the bidirectional time lag relationships between passive GPS data and active ecological momentary assessment (EMA) data collected via smartphone app technology.

Methods: Between March 2020 and May 2022, we recruited 45 participants (mean age 42.3 years, SD 12.1 years) who were followed up for 6 months: 35 individuals diagnosed with mood disorders referred by psychiatrists and 10 healthy control participants. This resulted in a total of 5248 person-days of data. Over 6 months, we collected 2 types of smartphone data: passive data on movement patterns with nearly 100,000 GPS data points per individual and active data through EMA capturing daily mood levels, including fatigue, irritability, depressed, and manic mood. Our study is limited to Android users due to operating system constraints.

Results: Our findings revealed a significant negative correlation between normalized entropy (r=–0.353; P=.04) and weekly depressed mood as well as between location variance (r=–0.364; P=.03) and depressed mood. In participants with mood disorders, we observed bidirectional time-lagged associations. Specifically, changes in homestay were positively associated with fatigue (β=0.256; P=.03), depressed mood (β=0.235; P=.01), and irritability (β=0.149; P=.03). A decrease in location variance was significantly associated with higher depressed mood the following day (β=–0.015; P=.009). Conversely, an increase in depressed mood was significantly associated with reduced location variance the next day (β=–0.869; P<.001). These findings suggest a dynamic interplay between mood symptoms and mobility patterns.

XSL•FO
RenderX

Conclusions: This study demonstrates the potential of utilizing active EMA data to assess mood levels and passive GPS data to analyze mobility behaviors, with implications for managing disease progression in patients. Monitoring location variance and homestay can provide valuable insights into this process. The daily use of smartphones has proven to be a convenient method for monitoring patients' conditions. Interventions should prioritize promoting physical movement while discouraging prolonged periods of staying at home.

(J Med Internet Res 2024;26:e55635) doi:10.2196/55635

KEYWORDS

ecological momentary assessment; digital phenotyping; GPS mobility; bipolar disorder; major depressive disorder; GPS; global positioning system; mood disorders; assessment; depression; anxiety; digital phenotype; smartphone app; technology; behavioral changes; patient; monitoring

Introduction

Patients with major depressive disorder (MDD) and bipolar disorder (BP) frequently experience symptoms during abnormal mood swings. These mood symptoms not only affect an individual's overall functioning and quality of life but also influence their behavior and activity patterns. Traditional methods of measuring mood were limited to subjective self-reporting, which lacked the frequency, convenience, and long-term detection required for necessary assessment. With the advent of digital phenotyping [1], collecting both active and passive data is now possible in a more timely and efficient manner. Digital phenotyping represents an unobtrusive and nonintrusive surveillance approach, harnessing the growing availability of health-related data to enhance our understanding of disease-related status [2]. Currently, mental disorder assessment heavily relies on subjective measures such as paper-based questionnaires used in clinical diagnosis. Additionally, doctors frequently have only a few minutes to assess mood disorders, making a thorough evaluation of symptoms and behaviors challenging within a limited time frame.

A smartphone app has been developed to gather digital phenotypic data from users, enabling the use of participants' medical data as a foundation for developing systems for long-term tracking and personalized health care [3]. The usage and ownership of smartphones far surpass that of wearable devices, while wearable devices have gained some popularity [4]. Smartphones not only allow for passive data collection but also allow active data collection through techniques such as ecological momentary assessment (EMA) [5]. EMA involves real-time, active self-reporting with smartphones or actigraphy, which captures participants' current behavior and experiences in their natural environment [6]. Individuals can fill out surveys multiple times per day through EMA, enabling a better understanding of symptom variability over time [6]. EMA on smartphones aims to minimize recall bias, maximize generalizability, and study the effects of high-frequency surveys on behavior in real-world settings [5].

Recent studies have introduced novel approaches to measure behavior and activity by using global positioning system (GPS) geographic location data as reliable indicators of distance travelled or daily movement, which can be associated with mood fluctuations [7]. Symptoms associated with MDD include energy level changes and feelings of fatigue [8], while symptoms of BP often involve irritability. Several studies have revealed a correlation between reduced daily movement activity and increased depressive symptoms at a 2-week follow-up [9,10]. Additionally, studies have examined the predictive value of changes in GPS features related to locations and transitions in determining variations in depression symptom severity [11]. Conversely, emotions have been found to affect subsequent mobility behavior. Baseline measurements of depressive symptom severity have been shown to affect GPS-based indices over a 2-week follow-up period, including circadian movement, location variance, and normalized entropy, which measures the distribution of a participant's time across different location clusters [12,13]. Additionally, increased homestay is associated with greater depressive symptom severity [14]. These studies often investigate the relationships between mood and behavioral patterns from a time lag perspective.

The infrequent assessment of mood, which is typically conducted every 2 or 3 weeks [4,15-17], limits our ability to capture the dynamic changes in mood experienced by individuals. Tracking patients' behavior and mood in a timelier manner is essential to effectively monitor these fluctuations. Assessing emotions at baseline and again after 2 weeks is insufficient to capture individual changes over time. Therefore, this study aims to monitor daily mood variations, focusing on day-to-day changes rather than long-term shifts from baseline to 2 weeks later. This is achieved by implementing a daily self-reporting approach over 6 months, capturing participants' daily experiences. The aims of this study are to (1) perform the GPS feature selection by integrating EMA data across all participants and (2) explore the temporal bidirectional relationship between GPS features and EMA data over 6 months in individuals with mood disorders, with a particular emphasis on time lag effects.

Methods

Recruitment

We conducted a prospective study to continuously collect passive smartphone GPS data and daily EMA data on mood-related symptoms over a 6-month period. Participants were enrolled from psychiatric clinics in 3 major hospitals in Taipei between March 2020 and June 2022. Notably, the peak of the COVID-19 outbreak in Taiwan began in May 2022. Despite the global panic induced by the COVID-19 pandemic, Taiwan's prompt response effectively mitigated local transmission rates [18]. Consequently, the impact of actual

XSL•FO
RenderX

infections and related preventive measures on GPS data collection was significant only after May 2022 and not during the initial recruitment phase. This study targets outpatients aged 20-65 years who met the Diagnostic and Statistical Manual of Mental Disorders criteria for MDD or BP. All participants were diagnosed by psychiatrists and consented to formal interviews conducted by our team interviewers. Patients who are night-shift workers, with substance-induced mood disorders, intellectual developmental disorder, or schizophrenia were excluded. Healthy controls were recruited from the community near the recruitment hospitals in Taipei, specifically targeting non–night shift workers and individuals with no history of psychiatric disorders. Initially, 95 participants were eligible for recruitment. However, due to factors such as inability to establish contact or the experience of panic attacks, only 45 participants were successfully enrolled who completed the questionnaires. Consequently, our study focused on these 45 participants, tracking their daily moods over a 6-month period. This extended monitoring allowed us to observe variations in the mood states of the participants over time. In analyzing the correlation between mood levels and mobility features derived from GPS data, we included data from healthy controls to identify the key variables for subsequent time lag analysis. The time lag analysis itself was confined to data from participants diagnosed with MDD and BP.

Ethics Approval

All participants provided written informed consent, and this study received approval by the institutional review boards of National Taiwan University Hospital (Approval number: 201811087RINA).

Measurement of Symptom Severity

We investigated whether the patients were in the acute stage at baseline by using the Young Mania Rating Scale (YMRS) and Hamilton Depression Rating Scale (HAMD) scores for measurement to understand the disease severity. In this study, depressive symptoms and their severity were evaluated using HAMD [19]. The HAMD scores are interpreted as follows: 0-7 indicates a normal range, 8-16 suggests mild depression, 17-23 corresponds to moderate depression, and a score over 24 signifies severe depression. For assessing manic symptoms, YMRS [20] was employed. In YMRS, scores of 0-11 indicate remission, 12-15 suggest mild symptoms, 16-20 are indicative of moderate symptoms, and 21-60 represent severe symptoms. The cutoff value of <16 on YMRS denotes a nonacute stage of manic symptoms. The Beck Depression Inventory-II [21] measures the severity of symptoms, assessing the current depressive severity over the past 2 weeks. The scale classifies depression severity as follows: 0-13 indicates minimal, 14-19 indicates mild, 20-28 indicates moderate, and 29-63 indicates severe depression. All patients in our study were not in the acute stage of their conditions.

GPS Features and EMA Mood Collection

We collected GPS and EMA data by using the Beiwe platform developed by the Harvard research team Onnela Lab (Multimedia Appendix 1). It extracts biomedical and clinical insights from smartphone data by using research platforms, mobile apps, databases, and data analysis tools [2]. We collected data from participants who started to use the Beiwe app on March 16, 2020. The smartphone app is applied to Android and iOS systems as part of the Beiwe research platform [22]. Therefore, we could collect raw data from individual-level participants at every moment in actual real-world experiences everywhere. However, Beiwe was presented in English version; so, we worked with the Beiwe research team to develop a traditional Chinese version suitable for our study. Two types are currently available: Android and iOS versions. The Android version was easier to launch and has begun to enroll participants. However, problems remain with the data collection of the iOS system in the Chinese version; thus, our study can only extract data from the Android version. Therefore, out of the 95 eligible participants, 35 participants were excluded who were using the iOS version (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of the 6-month study period. BP: bipolar disorder; HC: healthy control; MDD: major depressive disorder.

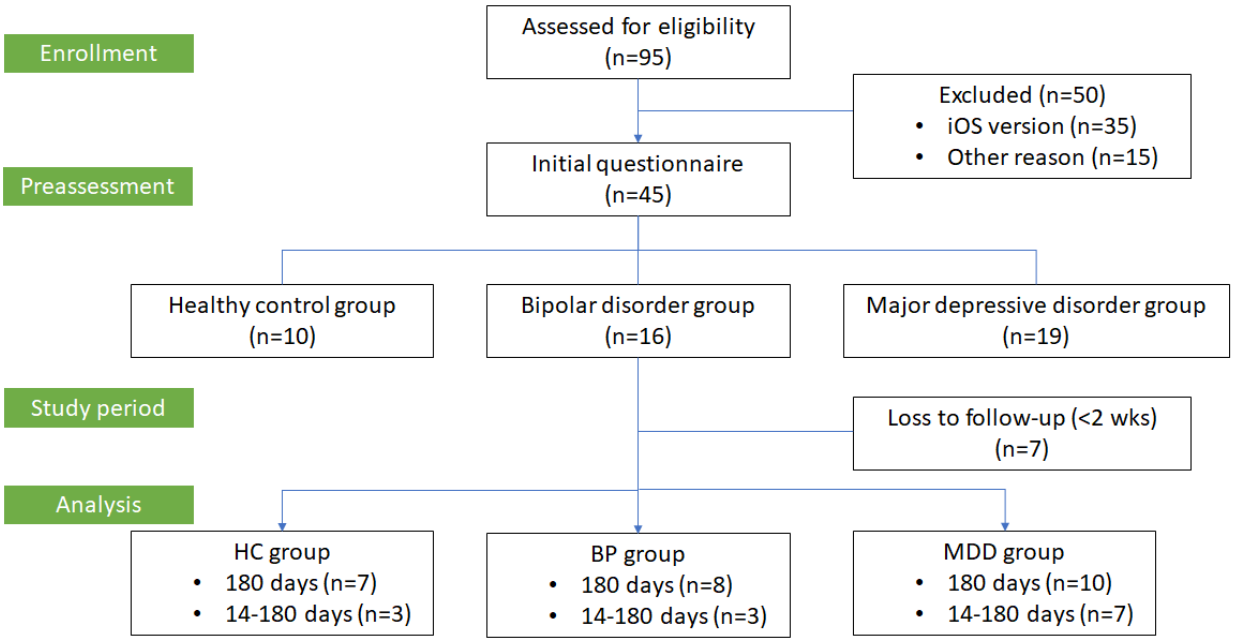

The GPS data includes a timestamp, Universal Time Coordinated, longitude, latitude, altitude, and accuracy. The time recorded by the Beiwe platform is Universal Time Coordinated. Taiwan time is 8 hours ahead of Universal Time Coordinated; thus, it was converted to Taiwan time. We collected GPS data generated by smartphone sensors 1000 times a day for 1 participant and realized the path of movement [22]. The GPS could not be sampled continuously because this would drain the smartphone's battery within a few hours; so, the app operated the GPS sensor at a prespecified frequency and regular duration intervals [2]. The Beiwe platform uploads an Excel file every hour, which is set as a rest interval of every 10 minutes to record 1 minute of data. Ideally, 1 data would record per second, but sometimes, the number of uploads would vary depending on the smartphone type [16]. The GPS data collected on the phone were encrypted at the moment of collection, and the recorded GPS data were not in the original latitude and longitude format. Additionally, the GPS data could not be accessed on the phone itself, and encryption measures protected the transmission process. Only when the data reached the researchers could it be decrypted to obtain the original GPS data. The mobility patterns following the features proposed by Saeb et al [12,13] included location variance, speed mean, speed variance, number of clusters, normalized entropy, homestay, transition time, and total distance. Location variance quantifies the variability in participants' GPS locations. A higher value of normalized entropy indicates a more uniform distribution of time spent across all locations. Homestay is defined as the proportion of time a participant spends at home. For detailed definitions and formulas of these and other GPS variables, refer to Multimedia Appendix 2.

In this study, active EMA data collection included monitoring participants' mood, sleep, and exercise states. Participants completed questionnaires every day for 180 consecutive days. The interviewers instructed participants to complete the questionnaire at the times prompted by the app. However, if they missed a prompt, they were allowed to complete it later at a more convenient time. The app prompted participants to complete the mood questionnaire (ie, EMA) twice daily at 12:45 PM and 5 PM, with response rates exceeding 95% for both times. Although participants were encouraged to respond at both times, the responses were generally consistent between the 2 times. Among those who reported twice per day, the intraclass correlations calculated for manic mood=0.888, irritability=0.999, depressed mood=0.920, and fatigue=0.747, indicating low within-day variations between the 2 prompted times. Thus, we selected the 12:45 PM responses for subsequent analysis due to a slightly higher completion rate. The mood questions addressed fatigue, irritability, manic mood, and depressed mood. EMA mood items were adopted from questionnaires that have already been published in papers related to treatment response [23]. They were asked 4 questions related to their current mood: (1) How fatigued do you feel right now? (2) How depressed do you feel right now? (3) How elated do

XSL•FO
RenderX

you feel right now? (4) How irritable/angry do you feel right now? Response options included 1 indicating “not at all, I’m in a good mood now,” 2 indicating “a little bit,” 3 indicating “moderate,” 4 indicating “serious,” and 5 indicating “serious and lasted all day.” Participants were given a compensation of approximately US $3.50 each time they completed a month of EMA questionnaires.

GPS Data Preprocessing and Imputation

Generally, the location coordinate error of GPS is less than 50 m. The raw data are processed following the protocol provided by the Onnela Lab research [22]. The process is as follows:

1. Take data points with an accuracy of <51, indicating high accuracy.
2. Collapse the data points into 10-second intervals. Additionally, calculate the value of longitude and latitude at every 10-second interval, mainly to solve the problem of sudden speed increase caused by GPS drift (Figure 2).
3. Check the distance and time interval between the 2 data points. Moving is a distance that exceeds √10 meters for a 10-second interval. The first stage mainly aimed to confirm (1) moving, (2) pause, (3) cannot define, and (4) missing.
4. The distance between any 2 locations within 300 seconds of <60 m was considered to be a pause at the same location. Represented by the distance length calculated by Mercator projection.
5. Do GPS tracking imputation. Data imputation is prone to achieve real-world distance because the GPS data of smartphones usually contain various degrees of missing data [24].

Figure 2. Average speed of the raw data and integrated 10-second data.

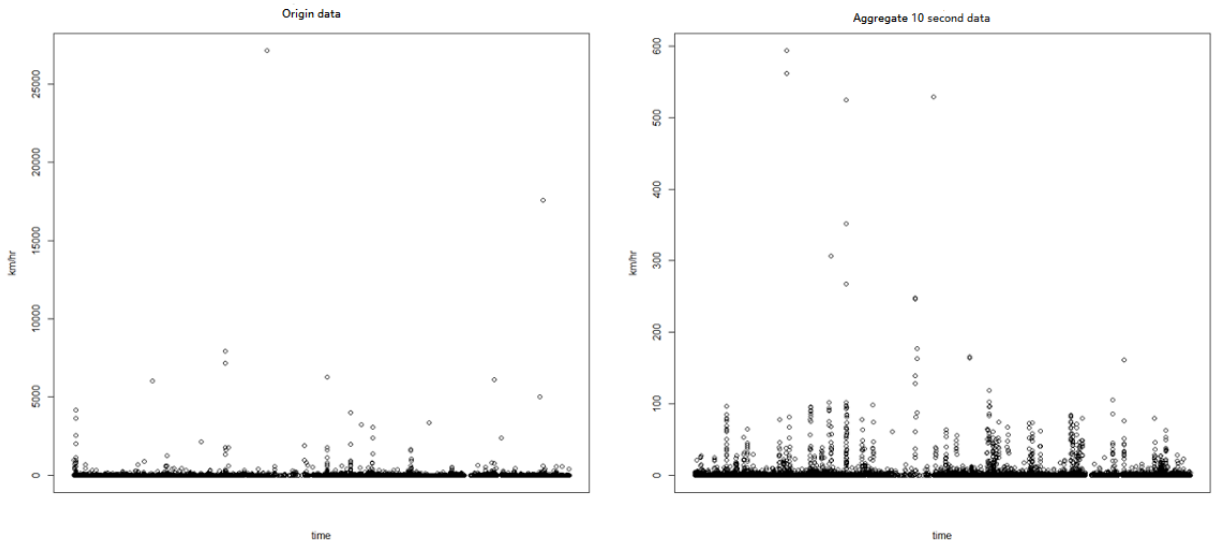

The following analysis is conducted based on a 10-time simulation average (Multimedia Appendix 3).

Statistical Analysis

We employed 1-way analysis of variance to analyze differences in the demographics, clinical characteristics, and data collection factors among the heathy controls, BP, and MDD groups. To explore the concurrent relationship between GPS features and EMA mood scores, polyserial correlation was computed. Initially, we analyzed all the samples to identify GPS features that showed a significant correlation with mood variations. The sample size was estimated based on the correlations between GPS features and mood-related symptoms reported in previous literature [12], which ranged between 0.49 and 0.63. Using a correlation coefficient of r=0.5, an α level of .05, and desired power levels of .80 and .90, the required sample sizes were 29 and 38, respectively. Thus, the current sample size is sufficient for feature selection in subsequent analyses. Additionally, we applied the generalized estimating equation models to capture the dynamic associations between mobility features and EMA mood states in patients with mood disorders, accounting for time lag associations. This decision was driven by our specific interest in exploring the fluctuations in mood, as captured by the EMA, within the patient group. The working correlation matrix in generalized estimating equation models can also select different types of correlation structures. Typically, the first-order autoregressive model is suitable for studies where the intervals between repeated measurements are of the same length, such as when the time intervals between different time points are identical. For example, in our study, the interval between day 1 and day 2 is one day apart. We considered a 2-sided P value of <.05 as statistically significant across all analyses. For the generalized estimating equation models, we adopted the first-order autoregressive working correlation structure to effectively accommodate the within-person correlations over the 6-month duration.

Results

Demographic and Clinical Characteristics

During the follow-up period of 6 months, 7 of the 45 participants were lost to follow-up and had <2 weeks of the GPS and EMA data due to various reasons such as the dislike of being located or not being used to filling in questionnaires on a smartphone daily. Therefore, this study analyzed data from 38 participants over a 6-month period, encompassing 5137 person-days. Table 1 demonstrates the sociodemographic characteristics of the

XSL•FO
RenderX

participants, including age, gender, social status, as well as the severity of depressive and bipolar symptoms according to affective status. The average age of the participants was 41.34 (SD 12.09) years, with approximately 63% (24/38) of them being female. Additionally, a higher percentage of individuals with BP were unemployed (6/11, 55%) compared to the MDD group and the healthy controls. The adherence rate to the Beiwe app, which calculated the proportion of days on which participants completed EMA mood assessments, was 93.5% (4803/5137) for the entire sample, and the clinical status collection rate was 100%.

Table 1. Sociodemographic characteristics of the participants.

	Healthy controls (n=10)	Bipolar disorder (n=11)	Major depressive disorder (n=17)	*P* value
Person-days (n)	1342	1537	2258	N/A[a]
Age (years), mean (SD)	34.40 (11.85)	44.09 (9.97)	43.65 (12.85)	.11
Male gender, n (%)	3 (30)	4 (36)	7 (41)	.86
BDI-II[b] scores, mean (SD)	4.60 (5.89)	20.73 (12.10)	20.41 (17.29)	.01
HAMD[c] scores, mean (SD)	—[d]	6.81 (4.79)	7.29 (4.95)	.80
YMRS[e] scores, mean (SD)	—	5.09 (4.89)	0.53 (1.33)	.001
Unemployment, n (%)	1 (10)	6 (55)	3 (18)	.04
Collection days, mean (SD)	134.20 (73.95)	156.18 (46.72)	135.12 (65.29)	.64
EMA[f] missing data percentage, mean (SD)	3.48 (3.51)	5.88 (6.29)	8.03 (13.10)	.88

[a]N/A: not applicable.
[b]BDI-II: Beck Depression Inventory-II.
[c]HAMD: Hamilton Depression Rating Scale.
[d]Not available.
[e]YMRS: Young Mania Rating Scale.
[f]EMA: Ecological Momentary Assessment.

Raw Data Quality and Preprocessing of GPS Data

We examined both the overall and individual accuracy distributions of the study participants to validate the selection of an accuracy threshold of 50 for the GPS data (Multimedia Appendix 4). Approximately 95% of the overall data points had an accuracy value of <47.59, and approximately 99% had an accuracy value of <76.03. Therefore, setting the threshold at 50 would encompass >90% of the data, with the majority of the accuracy values falling within the range of 10-20 m. This aligns with the recommendations of the research team at the Onnela Lab [24]. The simulated data exhibited a closer resemblance to the moving distance. We present the actual map distribution before and after imputation for 1 participant in Figure 3 to provide visual evidence.

Figure 3. Global positioning system tracking covered on maps. (A) Original track (B) Imputation track. The green circles show the number of clusters. The red circle indicates the location of home. GPS: global positioning system.

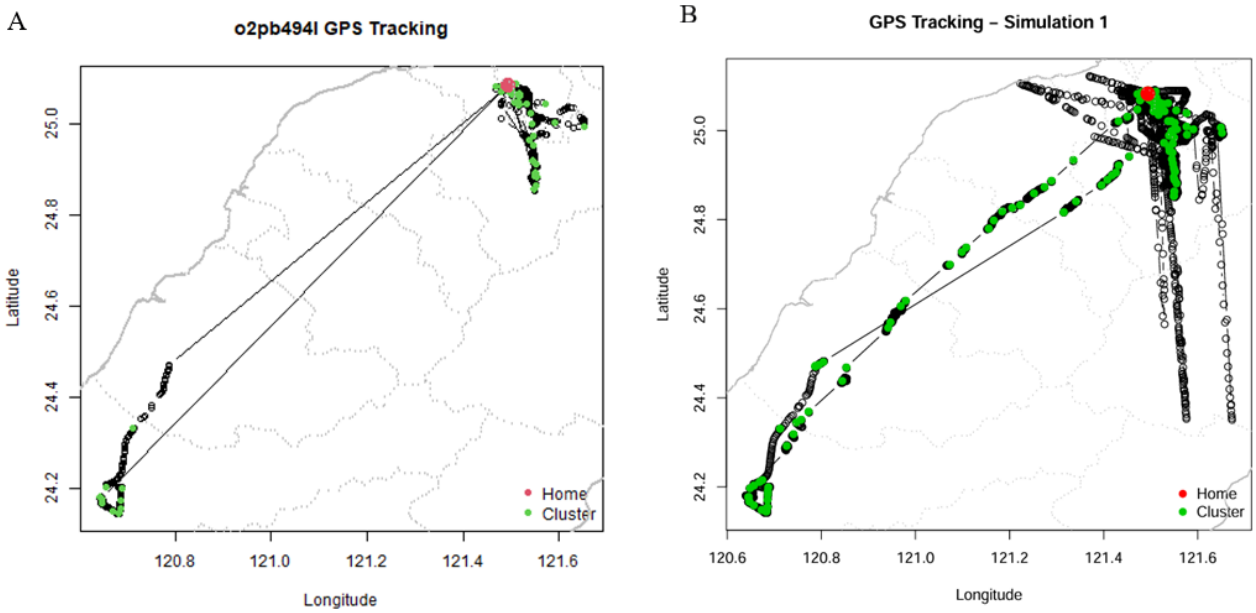

XSL•FO
RenderX

Two GPS Features Significantly Associated With Depressed Mood in EMA Data

To investigate the potential influences of moods on individual mobility, we conducted a correlation analysis involving GPS features and EMA data for all participants over a 1-week period. The correlation coefficients between 8 GPS features (location variance, speed mean, speed variance, total distance, transition time, number of clusters, homestay, and normalized entropy) and 4 EMA variables (fatigue, depressed mood, manic mood, and irritability) are presented in Table 2, revealing predominantly negative correlations. Notably, the most substantial absolute correlation was observed between depressed mood and location variance (r=–0.364; P=.03), closely followed by depressed mood and normalized entropy (r=–0.353; P=.04). Furthermore, it is noteworthy that depressed mood exhibited a positive correlation with homestay (r=0.239). Building on these preliminary findings, we then delved into the daily time lag relationship between mood and the aforementioned 3 GPS features by using the 6-month dataset.

Table 2. Correlation analysis of the global positioning system features and ecological momentary assessment mood.

	Fatigue		Depressed mood		Manic mood		Irritability	
	Estimate (r)	P value	Estimate (r)	P value	Estimate (r)	P value	Estimate (r)	P value
Location variance	–0.080	.64	–0.364	.03	–0.098	.57	–0.055	.75
Speed mean	–0.215	.21	–0.186	.28	–0.049	.78	–0.118	.49
Speed variance	–0.194	.26	–0.061	.72	–0.036	.84	–0.173	.31
Total distance	–0.102	.55	–0.305	.07	–0.159	.36	–0.189	.27
Transition time	–0.060	.73	–0.308	.07	–0.177	.30	–0.224	.19
Number of clusters	0.042	.81	–0.139	.42	–0.034	.84	0.090	.60
Homestay	0.142	.41	0.239	.16	0.059	.73	0.058	.74
Normalized entropy	–0.280	.09	–0.353	.04	–0.206	.23	–0.138	.42

Time Lagging Association of Changes in GPS Features With EMA Mood

We further examined the bidirectional relationship and lagging effect of GPS features on EMA mood by analyzing consecutive 6 months of EMA data and GPS coordinate collection. We evaluated the impact of changes in GPS features on subsequent EMA data as well as the influence of changes in EMA mood on subsequent GPS features. The patient group (Table 3) demonstrated significant associations between changes in GPS features and EMA mood. An increase in location variance from day 1 to day 2 was associated with lower depressed mood levels (β=–0.015; P=.009) on day 2. Furthermore, an increase in homestay from day 1 to day 2 was associated with higher fatigue (β=0.256; P=.03), depressed mood (β=0.235; P=.01), and irritability (β=0.149; P=.03) on day 2. However, changes in normalized entropy did not significantly predict any EMA mood within the patient group similar to the overall samples (see Multimedia Appendix 5).

XSL•FO
RenderX

Table 3. Time lag association of the change in the global positioning system features with mood for the patient group (bipolar disorder and major depressive disorder).

	$Fatigue_{day\ 2}$		$Depressed\ mood_{day\ 2}$		$Manic\ mood_{day\ 2}$		$Irritability_{day\ 2}$	
	β	*P* value	β	*P* value	β	*P* value	β	*P* value
Outcome 1								
$\Delta LV^{a}_{day\ 1\text{-}day\ 2}$	–0.007	.06	–0.015	.009	0.005	.35	–0.004	.31
Age	–0.022	.001	–0.017	.04	0.000	.98	–0.003	.62
Sex	–0.062	.72	–0.440	.07	–0.367	.04	–0.418	.004
$LV_{day\ 1}$	–0.014	.06	–0.029	.009	0.002	.76	–0.007	.32
Outcome 2								
$\Delta NEN^{b}_{day\ 1\text{-}day\ 2}$	0.206	.31	–0.279	.19	0.065	.58	–0.094	.48
Age	–0.023	.001	–0.018	.04	<0.001	.99	–0.003	.60
Sex	–0.084	.65	–0.454	.07	–0.372	.04	–0.425	.005
$NEN_{day\ 1}$	0.164	.64	–0.515	.22	0.243	.32	–0.021	.93
Outcome 3								
$\Delta HS^{c}_{day\ 1\text{-}day\ 2}$	0.256	.03	0.235	.01	–0.042	.74	0.149	.03
Age	–0.025	<.001	–0.020	.02	<0.001	.99	–0.005	.44
Sex	–0.017	.92	–0.419	.08	–0.361	.04	–0.393	.005
$HS_{day\ 1}$	0.460	.002	0.396	.01	0.020	.89	0.249	.07

[a]LV: location variance.

[b]NEN: normalized entropy.

[c]HS: homestay.

Time Lagging Association of Changes in EMA Mood With GPS Features

Our analysis of the associations between changes in EMA mood and subsequent GPS features revealed the following results (Table 4). BP and MDD revealed less location variance if fatigue (β=–0.492; *P*=.03) and depressed mood (β=–0.869; *P*<.001) increased from day 1 to day 2. However, the homestay would be greater if fatigue (β=0.047; *P*=.02) and depressed mood (β=0.042; *P*<.001) increased from day 1 to day 2. In all samples (Multimedia Appendix 6), location variance (β=–0.880; *P*=.001) would be less and homestay (β=0.038; *P*=.01) would be greater on day 2 if depressive status increased from day 1 to day 2. Additionally, homestay (β=0.034; *P*=.03) would be higher on day 2 if irritability increased from day 1 to day 2.

XSL•FO
RenderX

Table 4. Time lag association of the changes in the ecological momentary assessment mood with global positioning system features for the patient group (bipolar disorder and major depressive disorder).

	Location variance$_{day\ 2}$		Normalized entropy$_{day\ 2}$		Homestay$_{day\ 2}$	
	β	*P* value	β	*P* value	β	*P* value
Outcome 1						
ΔFatigue$_{day\ 1\text{-}day\ 2}$	–0.492	.03	0.011	.33	0.047	.02
Age	0.015	.65	0.000	.79	0.008	<.001
Sex	1.074	.21	0.038	.33	–0.132	.02
Fatigue$_{day\ 1}$	–1.040	.03	0.003	.89	0.097	<.001
Outcome 2						
ΔDepressed mood$_{day\ 1\text{-}day\ 2}$	–0.869	<.001	–0.014	.21	0.042	<.001
Age	0.011	.68	0.000	.93	0.007	.002
Sex	0.378	.64	0.025	.48	–0.108	.07
Depressed$_{day1}$	–1.553	<.001	–0.026	.21	0.063	.003
Outcome 3						
ΔManic mood$_{day\ 1\text{-}day\ 2}$	0.117	.73	0.008	.38	0.001	.96
Age	0.039	.26	0.000	.83	0.006	.02
Sex	1.100	.30	0.045	.28	–0.134	.04
Manic mood$_{day\ 1}$	–0.092	.84	0.021	.22	0.013	.76
Outcome 4						
ΔIrritability$_{day\ 1\text{-}day\ 2}$	–0.498	.15	–0.007	.58	0.040	.01
Age	0.036	.26	0.000	.83	0.006	.008
Sex	0.759	.47	0.036	.39	–0.109	.10
Irritability$_{day1}$	–0.872	.13	–0.004	.88	0.070	.03

In our study, we summarized the bidirectional correlations between changes in GPS features and EMA mood assessments in mood disorders, as illustrated in Figure 4. Our key findings are as follows:

1. Significant negative correlations were observed between changes in location variance and depressed mood levels. This suggests that greater variability in location is associated with lower depressed mood scores.
2. We also identified significant positive correlations between increase in homestay duration and elevated levels of fatigue, depressed mood, and irritability. This indicates that spending more time at home is linked to higher levels of these mood symptoms.
3. Additionally, a unidirectional relationship was discovered between increase in fatigue and subsequent changes in location variance. This implies that higher fatigue levels may lead to reduced variability in location.

XSL•FO
RenderX

Figure 4. Summary of the time lag effect between the global positioning system features and the ecological momentary assessment moods for bipolar disorder and major depressive disorder. EMA: ecological momentary assessment; GPS: global positioning system.

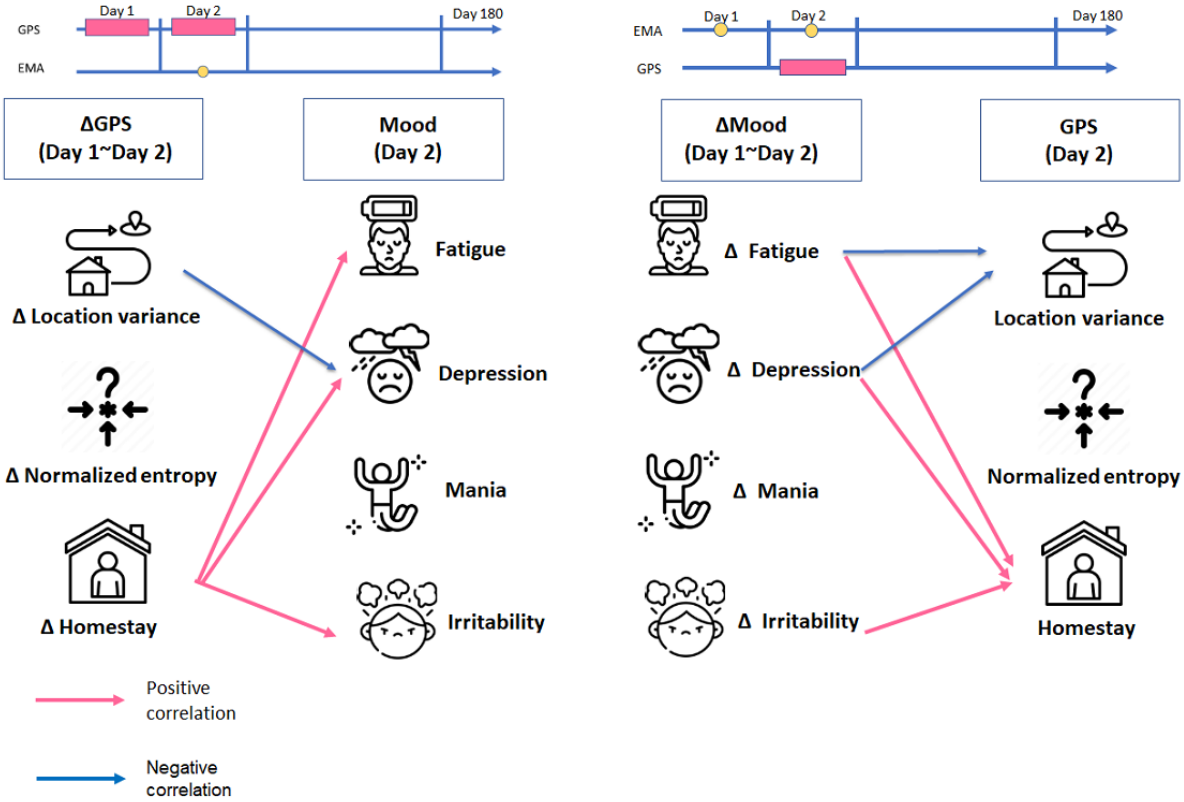

Discussion

This study investigates the association between GPS features and EMA mood in individuals with MDD and BP, focusing on correlation analysis and bidirectional time lag effects. Our study enables each participant to generate nearly 100,000 GPS data points by leveraging the abundant data collected from smartphones. GPS-derived mobility patterns are highly individual-specific and can accurately identify individuals based on factors such as their time spent at home, providing a comprehensive representation of their overall statistical mobility patterns [25]. Our study reveals that location variance and normalized entropy, among all GPS features, were moderately correlated with mood symptoms across the entire sample when analyzed on a weekly basis, similar to findings from previous research [12,13,17,26]. Although earlier studies have reported moderate to high correlations (–0.58 to –0.43), our correlations ranging from –0.35 to –0.36 are slightly lower. Additionally, in line with findings from another study [27], mobility patterns correctly classified participants with at least mild depression and those not classified as depressed in 82% of the cases. The most predictive mobility features in that study were related to location changes and entropy, with lower entropy associated with greater inequality in time spent across different locations and higher levels of depression [27]. Furthermore, a study focusing on Android systems revealed correlations among entropy, normalized entropy, location variance, and mood scores [28]. Collectively, our findings indicate that individuals with depressed mood tend to exhibit reduced mobility and a tendency to concentrate in specific locations.

Our study reveals intriguing findings regarding the time lag effect, demonstrating the dynamic bidirectional association between location variance and fatigue and homestay and depressed mood (Figure 4). The bidirectional relationship coincides with previous studies [17] that revealed significant correlations between location variance, homestay, and subsequent depressed mood. Additionally, changes in depressed mood status significantly affected the subsequent periodicity of mobility measured by circadian rhythm. BP and MDD are distinct diagnoses within the mood disorder spectrum, but they share a common feature—the presence of depression. Our study results are consistent with those of a previous study [14] that demonstrated a positive relationship between homestay and the severity of depressive symptoms in individuals with MDD and other affective disorders, including unipolar disorder and BP [29]. Thus, our results support the notion that increased homestay is linked to higher severity of depressed mood. This association suggests that individuals with MDD and BP who spend more time at home may experience decreased energy levels. Feelings of fatigue and loss of energy are relatively easy to observe in the early stages of depression diagnosis [30]. Fatigue, a common early symptom in depression diagnosis, is prevalent in MDD, with over 90% of the patients reporting it as a prodromal or residual symptom [31]. Therefore, prolonged homestay, a key indicator for mood monitoring in our study, is correlated with increased fatigue, depressive symptoms, and irritability in individuals diagnosed with MDD or BP.

Conversely, our findings did not establish a relationship between manic mood states and GPS features. We observed no predictive or time-lagged effects for mania symptoms in BP. This could be attributed to the nature of our study cohort, which, over the

XSL•FO
RenderX

6-month follow-up, predominantly consisted of nonsevere cases due to dropout of more unstable cases. As a result, the low mania scores in the EMA mood assessments were likely due to the lower prevalence of mania symptoms, rendering it challenging to detect any predictors. Additionally, the lack of insight into manic symptoms in BP is a factor that our study could not conclusively address. Although mania is a defining characteristic of BP, depressive episodes and symptoms are more prevalent over the course of the disorder [32]. Thus, even in patients with mania, the response variation to manic states was minimal, leading to more pronounced depressive outcomes in our findings.

Proxies for social and physical behavior, derived from smartphone sensor data such as GPS and phone usage, enable context-sensitive and personalized interventions for individuals with depressive symptoms. Long-term use of the app significantly reduced self-reported symptom severity among participants [33]. In line with this, a study by Rethorst et al [34] explored the efficacy of behavioral activation interventions, which specifically focus on increasing physical activity. Such interventions have shown promise in improving treatment outcomes for depression and enhancing physical health. This underscores the potential benefits of incorporating physical activity into treatment plans for individuals with depression. Additionally, research by Merikangas et al [35] has shown that individuals with bipolar illness exhibit increased cross-domain reactivity, affecting various aspects such as motor activity, sleep, mood, and energy levels. The concept of mobility patterns, akin to activity levels, aligns with the idea that in patients with BP, this heightened cross-domain reactivity suggests a lack of centralized coordination in the network governing activity and energy levels. This is indicative of a broader dysregulation rather than being confined to a single domain. Our study contributes novel insights to this area of research, further elucidating the relationship between activity or mobility levels and mood states. By examining these dynamics in the context of BP, our data provide additional evidence supporting the theory of widespread network dysregulation in these patients, particularly in how their mobility patterns correlate with mood variations.

Previous research has highlighted the predictive power of functional brain connectivity in determining individual mobility patterns. Notably, Xia et al [25] observed that increased connectivity between the somatomotor and limbic networks could predict mobility over medium time scales, ranging from weeks to months. Future research may explore the integration of studies on mobility and brain connectivity, as this could provide valuable insights into the interplay between physical movements and behavioral approaches. Such investigations may deepen our understanding of how brain function influences daily behavior and mobility.

This study has certain limitations that warrant consideration. First, the relatively small sample size poses questions regarding the representativeness of our findings. With a correlation coefficient of 0.364 this study, we achieved a moderate power of .70. A larger scale study is needed to validate these findings. Additionally, it is important to note that despite the limited number of participants, the study capitalizes on the extensive collection of GPS data. These data were gathered daily over a 6-month period, resulting in a substantial volume of person-days EMA and GPS data. Such a dataset offered a more comprehensive view of participants' behavior over time. Second, data on the Apple iOS version are lacking, which may introduce selection bias. In Taiwan, Apple smartphones hold a market share of 52.55%. Third, the reasons for loss to follow-up varied, with some data missing due to clinical management. When a patient experienced an episode requiring hospitalization during the follow-up period, they were unable to use a mobile phone while hospitalized in Taiwan. Consequently, movement variations during the hospitalization period and EMA completion were not captured, causing a significant amount of missing values due to inpatient health care. In our study, 2 participants lost their data for this reason. Fourth, although participants could complete the questionnaire at convenient times, potentially introducing recall bias, this flexibility likely had minimal impact on our findings. High completion rates and strong intraclass correlations indicate consistent responses across times, suggesting stable mood reporting. Thus, the potential bias from variable response times appears negligible in affecting result reliability. Lastly, the baseline participants demonstrated considerable interest and willingness to answer the app questionnaire daily for 6 months, but a dropout rate of approximately 16% (7/45) was observed at the first 2 weeks during the follow-up.

In conclusion, this study demonstrates a significant correlation between the reported mood states of patients and their mobility patterns. Utilizing a passive method such as carrying a smartphone daily proved to be a convenient and effective way to monitor patients' conditions. The concept of digital phenotyping employed in this study enabled nonintrusive and noninterfering measurement of mobile behaviors by collecting GPS data from smartphones without causing additional disruptions. Our findings showed a time lag relationship between real-time GPS monitoring combined with EMA. Mobility patterns (location variance and homestay) were significant indicators for detecting poor mood (irritability and depressed mood) and low energy levels (fatigue) in individuals with mood disorders. Therefore, interventions should prioritize promoting physical movement while discouraging prolonged periods of staying at home.

Acknowledgments

Part of this study was supported by the Ministry of Science and Technology (MOST 108-2314-B-002-136-MY3, 110-2314-B-002-067-MY3), National Taiwan University Career Development Project (109L7860, 110L7809), and Population Health and Welfare Research Center from Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan (grant: NTU-112L900402, 113L900402) to Kuo PH. The English editing

XSL•FO
RenderX

of this paper was sponsored by the National Taiwan University with the support of Higher Education Sprout Project from the Ministry of Education, Taiwan. In addition, we used ChatGPT to enhance the English grammar, word choice, and fluency but not for writing the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the institutional review board restrictions on direct data sharing. However, they can be made available from the corresponding author upon reasonable request.

Authors' Contributions

Lee TY wrote the first draft of manuscript and conducted analysis and interpretation of the global positioning system and ecological momentary assessment data. Chen CH was responsible for executing the statistical programming code. Chen IM, Chen HC, Liu CM, and Wu SI assisted in clinical data collection. CKH oversaw the statistical analyses and results interpretation. Kuo PH conceptualized the study, supervised the project, and interpreted the results. Both CKH and Kuo PH critically reviewed the manuscript. All authors approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Implementation procedures.
[DOCX File , 160 KB - jmir_v26i1e55635_app1.docx]

Multimedia Appendix 2
Definition and formula of global positioning system features.
[DOCX File , 16 KB - jmir_v26i1e55635_app2.docx]

Multimedia Appendix 3
Example of the global positioning system location simulation 10 times.
[DOCX File , 1139 KB - jmir_v26i1e55635_app3.docx]

Multimedia Appendix 4
The distribution of the accuracy value in all the participants (N=38).
[DOCX File , 26 KB - jmir_v26i1e55635_app4.docx]

Multimedia Appendix 5
Time lag association of the changes in the global positioning system features with the ecological momentary assessment mood for all samples (N=38).
[DOCX File , 19 KB - jmir_v26i1e55635_app5.docx]

Multimedia Appendix 6
Time lag association of the changes in the ecological momentary assessment mood with the global positioning system features for all the samples (N=38).
[DOCX File , 19 KB - jmir_v26i1e55635_app6.docx]

References

1. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol 2015 May;33(5):462-463. [doi: 10.1038/nbt.3223] [Medline: 25965751]
2. Torous J, Kiang MV, Lorme J, Onnela J. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 05;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
3. Wu C, Hsu J, Liou C, Su H, Lin EC, Chen P. Automatic bipolar disorder assessment using machine learning with smartphone-based digital phenotyping. IEEE Access 2023;11:121845-121858. [doi: 10.1109/access.2023.3328342]
4. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela J. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology 2018 Jul;43(8):1660-1666 [FREE Full text] [doi: 10.1038/s41386-018-0030-z] [Medline: 29511333]

XSL•FO
RenderX

5. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
6. Torous J, Onnela J, Keshavan M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Transl Psychiatry 2017 Mar 07;7(3):e1053 [FREE Full text] [doi: 10.1038/tp.2017.25] [Medline: 28267146]
7. Beiwinkel T, Kindermann S, Maier A, Kerl C, Moock J, Barbian G, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016 Jan 06;3(1):e2 [FREE Full text] [doi: 10.2196/mental.4560] [Medline: 26740354]
8. Chand SP, Arif H, Kutlenios RM. Depression (Nursing). StatPearls [Internet] 2023:1-10. [Medline: 33760492]
9. Faurholt-Jepsen M, Frost M, Vinberg M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as objective measures of bipolar disorder symptoms. Psychiatry Res 2014 Jun 30;217(1-2):124-127 [FREE Full text] [doi: 10.1016/j.psychres.2014.03.009] [Medline: 24679993]
10. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015:1293-1304. [doi: 10.1145/2750858.2805845]
11. Meyerhoff J, Liu T, Kording KP, Ungar LH, Kaiser SM, Karr CJ, et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. J Med Internet Res 2021 Sep 03;23(9):e22844 [FREE Full text] [doi: 10.2196/22844] [Medline: 34477562]
12. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
13. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537 [FREE Full text] [doi: 10.7717/peerj.2537] [Medline: 28344895]
14. Laiou P, Kaliukhovich DA, Folarin AA, Ranjan Y, Rashid Z, Conde P, RADAR-CNS. The association between home stay and symptom severity in major depressive disorder: preliminary findings from a multicenter observational study using geolocation data from smartphones. JMIR Mhealth Uhealth 2022 Jan 28;10(1):e28095 [FREE Full text] [doi: 10.2196/28095] [Medline: 35089148]
15. Staples P, Torous J, Barnett I, Carlson K, Sandoval L, Keshavan M, et al. A comparison of passive and active estimates of sleep in a cohort with schizophrenia. NPJ Schizophr 2017 Oct 16;3(1):37 [FREE Full text] [doi: 10.1038/s41537-017-0038-0] [Medline: 29038553]
16. Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela JP. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. NPJ Digit Med 2018;1(1):15 [FREE Full text] [doi: 10.1038/s41746-018-0022-8] [Medline: 31304300]
17. Zhang Y, Folarin AA, Sun S, Cummins N, Vairavan S, Bendayan R, RADAR-CNS consortium. Longitudinal relationships between depressive symptom severity and phone-measured mobility: dynamic structural equation modeling study. JMIR Ment Health 2022 Mar 11;9(3):e34898 [FREE Full text] [doi: 10.2196/34898] [Medline: 35275087]
18. Cheng H, Liu D. Early prompt response to COVID-19 in Taiwan: Comprehensive surveillance, decisive border control, and information technology support. J Formos Med Assoc 2024 Jan;123 Suppl 1:S2-S7 [FREE Full text] [doi: 10.1016/j.jfma.2022.11.002] [Medline: 36424239]
19. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb;23(1):56-62 [FREE Full text] [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
20. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978 Nov;133:429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
21. Kapci EG, Uslu R, Turkcapar H, Karaoglan A. Beck Depression Inventory II: evaluation of the psychometric properties and cut-off points in a Turkish adult population. Depress Anxiety 2008;25(10):E104-E110. [doi: 10.1002/da.20371] [Medline: 17876817]
22. Onnela J, Dixon C, Griffin K, Jaenicke T, Minowada L, Esterkin S, et al. Beiwe: A data collection platform for high-throughput digital phenotyping. JOSS 2021 Dec;6(68):3417. [doi: 10.21105/joss.03417]
23. Scott J, Hidalgo-Mazzei D, Strawbridge R, Young A, Resche-Rigon M, Etain B, et al. Prospective cohort study of early biosignatures of response to lithium in bipolar-I-disorders: overview of the H2020-funded R-LiNK initiative. Int J Bipolar Disord 2019 Sep 25;7(1):20 [FREE Full text] [doi: 10.1186/s40345-019-0156-x] [Medline: 31552554]
24. Barnett I, Onnela JP. Inferring mobility measures from GPS traces with missing data. Biostatistics 2020 Apr 01;21(2):e98-e112 [FREE Full text] [doi: 10.1093/biostatistics/kxy059] [Medline: 30371736]
25. Xia CH, Barnett I, Tapera TM, Adebimpe A, Baker JT, Bassett DS, et al. Mobile footprinting: linking individual distinctiveness in mobility patterns to mood, sleep, and brain functional connectivity. Neuropsychopharmacology 2022 Aug;47(9):1662-1671 [FREE Full text] [doi: 10.1038/s41386-022-01351-z] [Medline: 35660803]
26. Palmius N, Tsanas A, Saunders KEA, Bilderbeck AC, Geddes JR, Goodwin GM, et al. Detecting bipolar depression from geographic location data. IEEE Trans Biomed Eng 2017 Aug;64(8):1761-1771 [FREE Full text] [doi: 10.1109/TBME.2016.2611862] [Medline: 28113247]

XSL•FO
RenderX

27. Müller SR, Chen XL, Peters H, Chaintreau A, Matz SC. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. Sci Rep 2021 Jul 07;11(1):14007 [FREE Full text] [doi: 10.1038/s41598-021-93087-x] [Medline: 34234186]
28. Farhan A. Behavior vs introspection: refining prediction of clinical depression via smartphone sensing data. 2016 Presented at: IEEE Wireless Health (WH); October 25-27; Bethesda, MD. [doi: 10.1109/wh.2016.7764553]
29. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR Mhealth Uhealth 2018 Aug 13;6(8):e165 [FREE Full text] [doi: 10.2196/mhealth.9691] [Medline: 30104184]
30. Kennedy SH. Core symptoms of major depressive disorder: relevance to diagnosis and treatment. Dialogues in Clinical Neuroscience 2022 Apr 01;10(3):271-277. [doi: 10.31887/dcns.2008.10.3/shkennedy]
31. Surova G, Ulke C, Schmidt FM, Hensch T, Sander C, Hegerl U. Fatigue and brain arousal in patients with major depressive disorder. Eur Arch Psychiatry Clin Neurosci 2021 Apr;271(3):527-536 [FREE Full text] [doi: 10.1007/s00406-020-01216-w] [Medline: 33275166]
32. McIntyre RS, Berk M, Brietzke E, Goldstein BI, López-Jaramillo C, Kessing LV, et al. Bipolar disorders. Lancet 2020 Dec 05;396(10265):1841-1856. [doi: 10.1016/S0140-6736(20)31544-0] [Medline: 33278937]
33. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR Mhealth Uhealth 2016 Sep 21;4(3):e111 [FREE Full text] [doi: 10.2196/mhealth.5960] [Medline: 27655245]
34. Rethorst CD, Trombello JM, Chen P, Carmody TJ, Lazalde A, Trivedi MH. Adaption of tele-behavioral activation to increase physical activity in depression: Protocol for iterative development and evaluation. Contemp Clin Trials Commun 2023 Jun;33:101103 [FREE Full text] [doi: 10.1016/j.conctc.2023.101103] [Medline: 37128575]
35. Merikangas KR, Swendsen J, Hickie IB, Cui L, Shou H, Merikangas AK, et al. Real-time mobile monitoring of the dynamic associations among motor activity, energy, mood, and sleep in adults with bipolar disorder. JAMA Psychiatry 2019 Feb 01;76(2):190-198 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3546] [Medline: 30540352]

Abbreviations

BP: bipolar disorder
EMA: ecological momentary assessment
GPS: global positioning system
HAMD: Hamilton Depression Rating Scale
MDD: major depressive disorder
YMRS: Young Mania Rating Scale

Edited by T de Azevedo Cardoso; submitted 19.12.23; peer-reviewed by F Lamers, R de Filippis, C Simons; comments to author 06.08.24; revised version received 15.09.24; accepted 13.11.24; published 06.12.24.

Please cite as:
Lee TY, Chen CH, Chen IM, Chen HC, Liu CM, Wu SI, Hsiao CK, Kuo PH
Dynamic Bidirectional Associations Between Global Positioning System Mobility and Ecological Momentary Assessment of Mood Symptoms in Mood Disorders: Prospective Cohort Study
J Med Internet Res 2024;26:e55635
URL: https://www.jmir.org/2024/1/e55635
doi:10.2196/55635
PMID:

©Ting-Yi Lee, Ching-Hsuan Chen, I-Ming Chen, Hsi-Chung Chen, Chih-Min Liu, Shu-I Wu, Chuhsing Kate Hsiao, Po-Hsiu Kuo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Analysis of Physical Activity Using Wearable Health Technology in US Adults Enrolled in the All of Us Research Program: Multiyear Observational Study

Rujul Singh[1], BS; Macy K Tetrick[1], BS; James L Fisher[2], PhD; Peter Washington[3], PhD; Jane Yu[4], BS; Electra D Paskett[1,5], PhD; Frank J Penedo[6,7], PhD; Steven K Clinton[8], MD, PhD; Roberto M Benzo[1], PhD

[1]Division of Cancer Prevention and Control, Department of Internal Medicine, College of Medicine, The Ohio State University, Columbus, OH, United States

[2]Arthur G. James Cancer Hospital and Richard J. Solove Research Institute, Columbus, OH, United States

[3]Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI, United States

[4]College of Medicine, The Ohio State University, Columbus, OH, United States

[5]James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, United States

[6]Departments of Psychology and Medicine, University of Miami, Miami, FL, United States

[7]Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Miami, United States

[8]Division of Medical Oncology, Department of Internal Medicine, The Ohio State University, Columbus, OH, United States

Corresponding Author:
Roberto M Benzo, PhD
Division of Cancer Prevention and Control
Department of Internal Medicine
College of Medicine, The Ohio State University
Suite 200
3650 Olentangy River Road
Columbus, OH, 43124
United States
Phone: 1 614 293 3675
Email: roberto.benzo@osumc.edu

Abstract

Background: To date, no studies have examined adherence to the 2018 Physical Activity Guidelines for Americans (PAGA) in real-world longitudinal settings using objectively measured activity monitoring data. This study addresses this gap by using commercial activity monitoring (Fitbit) data from the All of Us dataset.

Objective: The primary objectives were to describe the prevalence of adherence to the 2018 PAGA and identify associated sociodemographic determinants. Additionally, we compared 3 distinct methods of processing physical activity (PA) data to estimate adherence to the 2008 PAGA.

Methods: We used the National Institutes of Health's All of Us dataset, which contains minute-level Fitbit data for 13,947 US adults over a 7-year time span (2015-2022), to estimate adherence to PAGA. A published step-based method was used to estimate metabolic equivalents and assess adherence to the 2018 PAGA (ie, ≥150 minutes of moderate- to vigorous-intensity PA per week). We compared the step-based method, the heart rate–based method, and the proprietary Fitbit-developed algorithm to estimate adherence to the 2008 PAGA.

Results: The average overall adherence to the 2018 PAGA was 21.6% (3006/13,947; SE 0.4%). Factors associated with lower adherence in multivariate logistic regression analysis included female sex (relative to male sex; adjusted odds ratio [AOR] 0.66, 95% CI 0.60-0.72; P<.001); BMI of 25.0-29.9 kg/m^2 (AOR 0.53, 95% CI 0.46-0.60; P<.001), 30-34.9 kg/m^2 (AOR 0.30, 95% CI 0.25-0.36; P<.001), or ≥35 kg/m^2 (AOR 0.13, 95% CI 0.10-0.16; P<.001; relative to a BMI of 18.5-24.9 kg/m^2); being aged 30-39 years (AOR 0.66, 95% CI 0.56-0.77; P<.001), 40-49 years (AOR 0.79, 95% CI 0.68-0.93; P=.005), or ≥70 years (AOR 0.74, 95% CI 0.62-0.87; P<.001; relative to being 18-29 years); and non-Hispanic Black race or ethnicity (AOR 0.63, 95% CI 0.50-0.79; P<.001; relative to non-Hispanic White race or ethnicity). The Fitbit algorithm estimated that a larger percentage of

XSL•FO
RenderX

the sample (10,307/13,947, 73.9%; 95% CI 71.2-76.6) adhered to the 2008 PAGA compared to the heart rate method estimate (4740/13,947, 34%; 95% CI 32.8-35.2) and the step-based method (1401/13,947, 10%; 95% CI 9.4-10.6).

Conclusions: Our results show significant sociodemographic differences in PAGA adherence and notably different estimates of adherence depending on the algorithm used. These findings warrant the need to account for these disparities when implementing PA interventions and the need to establish an accurate and reliable method of using commercial accelerometers to examine PA, particularly as health care systems begin integrating wearable device data into patient health records.

(J Med Internet Res 2024;26:e65095) doi:10.2196/65095

KEYWORDS

Physical Activity Guidelines for Americans; accelerometry; All of Us Research Program; wearable activity monitors; health equity; multiyear activity tracking; activity intensity estimation; US adult population; sociodemographic determinants of physical activity; physical activity; wearables; United States; older adults; observational studies; longitudinal setting; sociodemographic determinants; physical activity data; Fitbit data; step-based method; adherence

Introduction

The current Physical Activity Guidelines for Americans (PAGA) recommends 150 minutes of moderate-intensity aerobic physical activity (MPA), 75 minutes of vigorous-intensity aerobic physical activity (VPA), or a combination of moderate- and vigorous-intensity aerobic physical activity (MVPA) per week [1,2]. Adherence to these standards is linked to a 33% decline in all-cause mortality and provides health benefits including but not limited to decreased risk of hypertension, type 2 diabetes, cardiovascular disease, cancer, and dementia [2,3]. In addition, studies conducted on the All of Us (AoU) dataset have demonstrated additional associations between daily step count and decreased risk of obesity, gastroesophageal reflux disease, and sleep apnea [4]. The AoU program is a national initiative led by the National Institutes of Health (NIH) aimed at gathering health data (eg, survey, electronic medical record, and Fitbit wearable) from 1 million or more people in the United States to accelerate research and improve health outcomes through precision medicine.

Given the established health benefits of physical activity (PA), numerous studies have been conducted to evaluate PA patterns in US adults, including the frequency, intensity, and duration of PA [5-12]. However, these studies present several limitations. A significant number relied on self-report questionnaires [5,6,11,13,14]; evidence suggests that questionnaires are less robust in measuring PA and are susceptible to social desirability bias, recall bias, and seasonal response variation [15-21]. One approach to overcome this limitation is to use activity monitors, as activity monitors provide objective measures of movement and activity that are free from random and systemic errors that arise from self-report [22]. Most studies using such technology have collected less than 1 week's worth of data [7,9,10,12]. Evidence indicates that 7-day monitoring periods typically result in intraclass correlations ≥80% while also capturing data from both weekdays and weekend days [22]. However, evidence also suggests that determining long-term habitual behavior patterns may require longer observation windows given significant seasonal and longitudinal variation in PA [19,22]. Compounding these limitations, activity-monitor estimates of PA intensity levels (ie, metabolic equivalents) have been determined in different ways (eg, using step count–, heart rate [HR]–, and proprietary device–generated estimates), which can lead to differing results [23,24]. As many health care systems are moving toward integrating these measures—many of which are "black box" outputs from commercial activity monitors like the Fitbit and Apple Watch—into patient health records [25,26], it is important to understand how these measures compare to one another in real-world observational settings.

Therefore, our observational study seeks to use the AoU dataset to address many of the aforementioned gaps and consists of 3 aims:

- Aim 1: Describe patterns of PA and adherence to the 2018 PAGA by sociodemographic variables (ie, sex, race or ethnicity, BMI, and age group)
- Aim 2: Evaluate the association of specific sociodemographic variables such as sex, race or ethnicity, BMI, and age group on the likelihood of meeting the 2018 PAGA
- Aim 3: Compare the proportion of adults meeting the 2008 PAGA as determined by 3 methods (step intensity, HR, and Fitbit proprietary algorithm)

Methods

Study Population

The AoU study is an NIH initiative to accelerate research and improve health outcomes through precision medicine by collecting a variety of data from various sources (eg, Fitbit, EHR, and survey) from adults in the United States. Under the AoU Fitbit Bring Your Own Device program, participants can link data (including data collected prior to the enrollment date) from their personal Fitbit device. A more detailed description of the AoU dataset can be found on the NIH AoU website [27]. The AoU program updates its dataset annually, and this study used the seventh version of the controlled tier dataset, released in April 2023. The study included 13,947 participants who consented to share their Fitbit data and possessed intraday HR and step count data for at least 1 valid week (requirements described below) of PA data from 2015 to 2022. All data periods prior to each participant's 18th birthday were excluded from the sample.

Ethical Considerations

Institutional review board approval was obtained from the AoU research program under protocol identification 2021-02-TN-001.

XSL•FO
RenderX

Guidelines established by the NIH Office for Human Research Protections were followed, by which standardized protection of study participants' rights and interests was ensured. A comprehensive informed consent procedure was implemented during study enrollment, whereby the right to discontinue involvement at any point was guaranteed to participants. Data security and participant confidentiality were safeguarded through multiple measures: secure data storage systems were implemented, access to identifying information was restricted, and confidentiality protocols were mandated through contractual agreements. Access to the Researcher Workbench was limited to authorized personnel by whom mandatory training had been completed and whose institutions were covered by active data use agreements. Compensation of US $25 was provided via a cash payment, gift card, or electronic voucher when biological specimens (blood, saliva, or urine) were collected at designated partner facilities. To protect participant privacy in accordance with AoU Research Program's Data and Statistics Dissemination Policy, for groups with fewer than 20 participants, results were not presented to prevent potential identification.

Measures

Minutes spent in sedentary, light-, moderate-, and vigorous-intensity activity were estimated using three distinct methods: (1) step intensity, (2) HR, and (3) a Fitbit proprietary algorithm. These outputs were then used to determine adherence to PAGA, where adherence was defined as meeting or exceeding a weekly average of 150 minutes of MVPA, where minutes of VPA count as 2 of MPA (ie, MPA+2×VPA≥150 minutes of MVPA). For aims 1 and 2, the step intensity–based method was used to estimate PA levels. The HR algorithm or equation (further described in the *PA Data Processing* section) was not used for various reasons that could confound our findings. First, the HR algorithm uses HR max as an input, and the formula used to calculate HR max has been shown to underestimate HR max in older ages [28]. Second, the accuracy of photoplethysmography sensors to measure HR is diminished in individuals with higher levels of melanin in the skin, which could confound the findings [29]. Third, equations that use HR max–based calculations (which is what was used in this analysis) to determine PA are poorly validated [30]. The Fitbit algorithm was not chosen as the primary method to estimate PA for 2 reasons. First, the algorithm only counts activity bouts of 10 minutes or more, which is aligned with the 2008 PAGA and not the current 2018 PAGA [31]. Second, the algorithm is not open source, which, from a scientific standpoint, limits transparency, reproducibility, peer review, collaboration, and improvement. In aim 3, we used all 3 algorithms to compare adherence to the 2008 PAGA, which only counts MVPA occurring in bouts of at least 10 minutes, where interruptions of up to 2 minutes were allowed in each bout (aligned with previous research) [9].

Sociodemographic Classification

Participants were classified by sex (ie, male, female, and other or not specified), race or ethnicity (ie, Hispanic, non-Hispanic Asian or Pacific Islander, non-Hispanic White, non-Hispanic Black, and other or not specified), age group (18-29, 30-39, 40-49, 50-59, 60-69, and ≥70 years), and BMI (ie, normal: 18.5-24.9 kg/m^2, overweight: 25.0-29.9 kg/m^2, obese: 30-34.9 kg/m^2, severely obese: ≥35 kg/m^2, and other or not specified). Given that a 7-year collection period of data was used (2015-2022), some participants aged into the subsequent age group. In these instances, only data from the age grouping that had the most valid weeks were used in the analysis. BMI was determined by averaging all measurements taken between the first and the last valid week of data collection for each participant, and BMIs over 150 kg/m^2 and under 12 kg/m^2 were considered outliers and removed from the dataset (n=101 participants had 1 or more invalid BMI measurements removed using this criterion) [32]. Individuals with no recorded BMI measurements were grouped with the "other" category. Participants with BMIs under 18.5 kg/m^2 were also grouped with the "other" category since there were fewer than 20 individuals with such measurements, and AoU restricts displaying data in such cases.

PA Data Processing

Given the lack of sleep logs and the absence of accelerometry wear-log data typically used for classifying sleep or nonwear, the approach recommended by Claudel et al [33] was used, in which data from 11 PM to 5 AM were classified as sleep or nonwear and removed from the dataset. After the removal of sleep periods, valid wear days were defined as any 24-hour period (starting at midnight) with 10 or more hours of HR recordings and 100 or more steps, following conventionally defined guidelines [4,9,34,35]. All participants with at least 1 week of valid wear were included, where a valid week was defined as 3 valid days of wear within a 7-day interval (starting on January 1st of each year) [35].

Average weekly minutes spent sedentary in light-intensity aerobic physical activity (LPA), MPA, and VPA were calculated using three methods: (1) step intensity algorithm: a threshold was set at 60 steps per minute for LPA, at 100 steps per minute for MPA, and at 130 steps per minute for VPA [36]; (2) HR algorithm: a threshold was set at 57% of maximum HR for LPA, at 64% of maximum HR for MPA, and at 77% of maximum HR for VPA [37], and maximum HR was computed by the American College of Sports Medicine's recommended formula of 208 – 0.7×age [38]; and (3) Fitbit algorithm: the device outputs of "sedentary," "lightly active," "fairly active," and "very active" minutes were used as analogs for each of the 4 activity intensities defined in the 2018 PAGA [39]. Figure 1 shows the data processing workflow in detail.

The processing steps were used to determine valid weeks of data and to determine adherence to the 2008 and 2018 PAGA.

XSL•FO
RenderX

Figure 1. Data processing workflow. AoU: All of Us; HR: heart rate; LPA: light-intensity aerobic physical activity; MPA: moderate-intensity aerobic physical activity; MVPA: moderate- and vigorous-intensity aerobic physical activity; PAGA: Physical Activity Guidelines for Americans; VPA: vigorous-intensity aerobic physical activity.

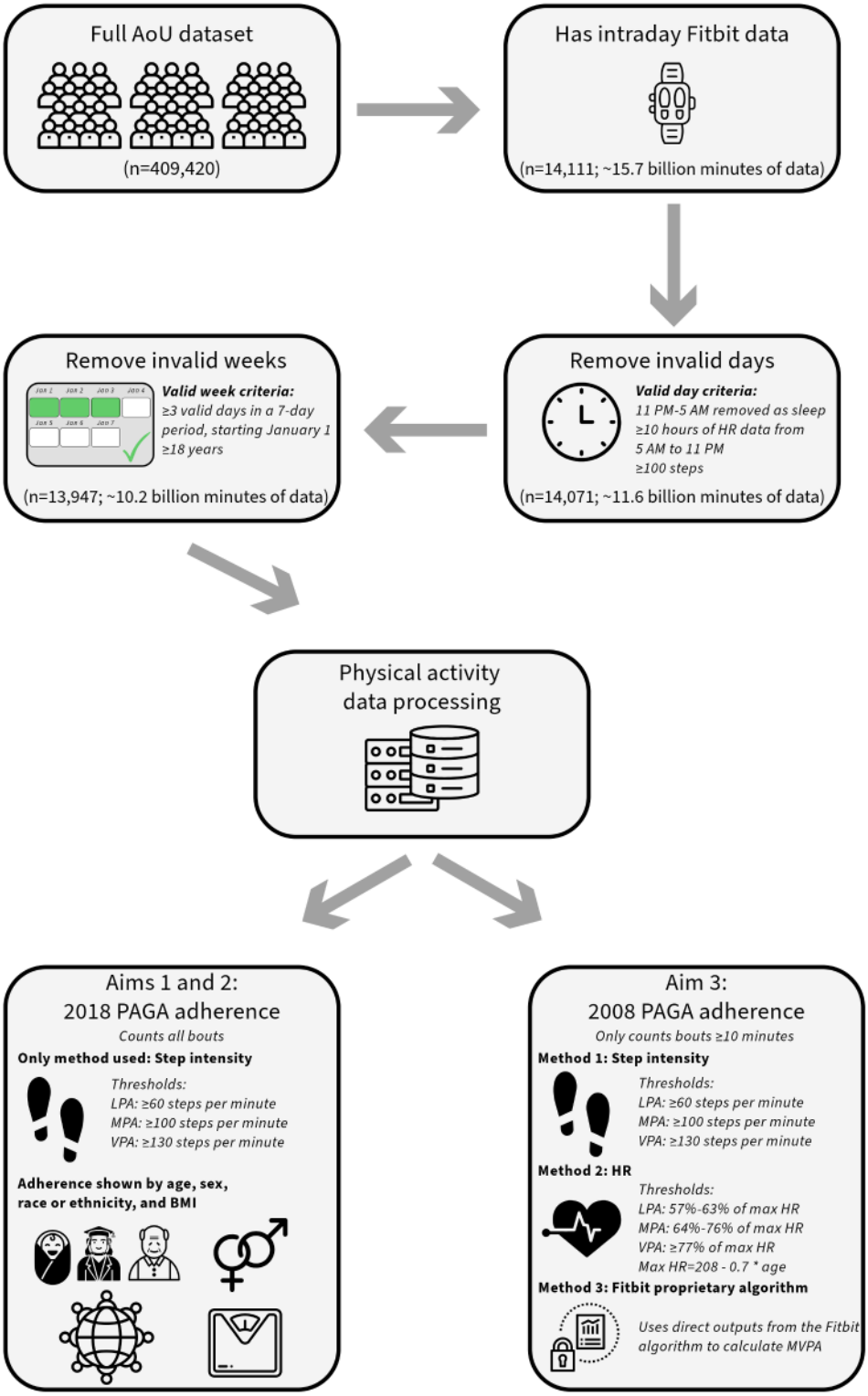

Statistical Analysis

A CONSORT (Consolidated Standards of Reporting Trials) diagram was prepared to show the number of individuals excluded from the dataset and can be found in Multimedia Appendix 1. We present descriptive summary statistics for our study sample using mean and SDs for continuous characteristics as well as frequencies for categorical variables.

Our research aim 1 was to describe PA levels and adherence to the 2018 PAGA across sociodemographic groups. To describe PA levels in our sample, we illustrate daily step counts, average weekly MVPA minutes, and IQRs by sex and age as well as sex and BMI. To describe the adherence of our sample to the guidelines, we present the proportion of participants meeting the 2018 PAGA (and 95% Wald CIs) [40] by sex, race or

XSL•FO
RenderX

ethnicity, BMI, and age group. Significant differences in proportions were assessed using pairwise chi-square tests with a Bonferroni-adjusted α level of 0.0012 (43 separate pairwise comparisons were considered) [41].

Our research aim 2 was to evaluate the association of specific sociodemographic variables with the likelihood of meeting the 2018 PAGA. Univariate and multivariate logistic regression were used to identify demographic factors associated with meeting the 2018 PAGA. Multivariate analysis controlled for age group, sex, BMI, and race or ethnicity, as these factors have been shown to have strong associations with PA (selected a priori) [42]. Linearity of logits was tested via examination of component residual plots, and multicollinearity was tested via the computation of variance inflation factors for each coefficient (with all variance inflation factor values confirmed to be under 5) [43]. Odds ratios (ORs) and 95% Wald CIs were reported for each coefficient. Significance was assessed using 2-sided Wald tests for significance on each coefficient, with an α level of 0.05 and 2 to 5 degrees of freedom for the univariate model (depending on the variable being tested) and 16 degrees of freedom for the multivariate model [40].

Our research aim 3 was to compare 3 different methods (step intensity, HR, and Fitbit proprietary) for estimating the proportion of individuals meeting the 2008 PAGA. To assess differences in proportions meeting guidelines as determined by each method, 2-sided pairwise McNemar tests were performed with a Bonferroni-adjusted α level of .0008, given the paired nature of the data (63 separate pairwise comparisons were considered) [44]. All analyses were performed using R (version 4.4.0; R Foundation for Statistical Computing) and RStudio (version 2024.4.0.735; Posit PBC) within the AoU Researcher Workbench.

Results

Participant Sociodemographic Characteristics

A total of 14,111 participants had intraday Fitbit HR, activity summary, and step count data available within the dataset (Table 1). After processing and removing for nonwear, the dataset included 13,947 individuals with over 1.6 million total weeks of data (averaging ~115 weeks of data collection for each individual participant). Our sample was predominantly female (9533/13,947, 68.5%) and identified predominantly as non-Hispanic White (11,109/13,947, 79.7%). The average age was 50.8 (SD 15.8) years, and the average BMI was 29.4 (SD 6.7) kg/m^2; nearly half (6590/13,947, 47.3%) of participants did not have valid BMI measurements available.

XSL•FO
RenderX

Table 1. Descriptive summary of the study sample[a].

Variables	Weeks of included data (n=1,607,789), n	Participants (n=13,947), n (%)
Sex		
Male	487,069	4007 (28.7)
Female	1,074,521	9553 (68.5)
Other or not specified	46,199	387 (2.8)
Race or ethnicity		
Hispanic	32,695	310 (2.2)
Non-Hispanic Asian or Pacific Islander	47,683	430 (3.1)
Non-Hispanic Black	71,482	684 (4.9)
Non-Hispanic White	1,311,659	11,109 (79.7)
Two or more races	31,603	293 (2.1)
Other or not specified	112,667	1121 (8)
BMI (kg/m^2)[b,c]		
18.5-24.9	276,329	1994 (14.3)
25-29.9	342,679	2487 (17.8)
30-34.9	203,851	1501 (10.8)
≥35	165,549	1375 (9.9)
Other or not specified	619,381	6590 (47.3)
Age group (years)[d]		
18-29	141,806	1511 (10.8)
30-39	274,727	2528 (18.1)
40-49	255,711	2259 (16.2)
50-59	322,264	2760 (19.8)
60-69	387,915	3102 (22.2)
≥70	225,366	1787 (12.8)

[a]Data include all weeks of valid data from all participants, ranging from 2015 to 2022.

[b]BMI was calculated as the average of all measured values measured between the first and last valid weeks for each participant. If no measured values were available, BMI was categorized as "other or not specified." Individuals with a BMI under 18.5 kg/m^2 are categorized as "other or not specified" due to the small sample size.

[c]Mean BMI 50.8, SD 15.8 kg/m^2.

[d]Mean age 29.4, SD 6.7 years.

Aim 1: Descriptive Analysis of PA

In total, 21.6% (3006/13,947) of individuals in the study sample met the 2018 PAGA, as determined by the step-intensity algorithm (Table 2). The average weekly minutes spent in MPA was 80.8 (SE 0.7) minutes, in VPA was 10.3 (SE 0.3) minutes, in MVPA (MPA+2×VPA) was 101.4 (SE 1.0) minutes, and the average daily step count was 7758 (SE 28) steps.

XSL•FO
RenderX

Table 2. Analysis of sociodemographic factors associated with meeting 2018 Physical Activity Guidelines for Americans (PAGA)[a].

	Total participants, n	Participants with MVPA[b] from step counts			Univariate regression		Multivariate regression[c]	
		n (%)	95% CI (%)	SE (%)	Odds ratio (95% CI)	*P* value	Multivariate odds ratio (95% CI)	*P* value
Total	13,947	3006 (21.6)	20.8-22.4	0.4	__[d]	—	—	—
Sex								
Male	4007	1069 (26.7)	24.7-28.7	1.0	—	—	—	—
Female	9553	1864 (19.5)	18.5-20.5	0.5	*0.67 (0.61-0.73)* [e]	<.001	*0.66 (0.60-0.72)*	<.001
Other or not specified	387	73 (18.9)	14.0-23.8	2.5	*0.64 (0.49-0.83)*	<.001	0.75 (0.55-1.01)	.07
Race or ethnicity								
Hispanic	310	60 (19.4)	13.9-24.9	2.8	0.85 (0.63-1.12)	.25	0.88 (0.65-1.17)	.39
Non-Hispanic Asian or Pacific Islander	430	135 (31.4)	24.9-37.9	3.3	*1.61 (1.31-1.98)*	<.001	*1.35 (1.08-1.67)*	.007
Non-Hispanic Black	684	87 (12.7)	9.8-15.6	1.5	*0.51 (0.41-0.64)*	<.001	*0.63 (0.50-0.79)*	<.001
Non-Hispanic White	11,109	2454 (22.1)	21.1-23.1	0.5	—	—	—	—
Two or more races	293	61 (20.8)	14.9-26.7	3.0	0.93 (0.69-1.23)	.60	0.96 (0.71-1.28)	.79
Other or not specified	1121	209 (18.6)	15.9-21.3	1.4	*0.81 (0.69-0.94)*	.008	0.84 (0.70-1.01)	.07
BMI (kg/m^2)[f]								
18.5-24.9	1994	727 (36.5)	33.2-39.8	1.7	—	—	—	—
25-29.9	2487	598 (24.0)	21.8-26.2	1.1	*0.55 (0.48-0.63)*	<.001	*0.53 (0.46-0.60)*	<.001
30-34.9	1501	224 (14.9)	12.7-17.1	1.1	*0.31 (0.26-0.36)*	<.001	*0.30 (0.25-0.36)*	<.001
≥35	1375	88 (6.4)	5.0-7.8	0.7	*0.12 (0.09-0.15)*	<.001	*0.13 (0.10-0.16)*	<.001
Other or not specified	6590	1369 (20.8)	19.6-22.0	0.6	*0.46 (0.41-0.51)*	<.001	*0.45 (0.41-0.51)*	<.001
Age group (years)								
18-29	1511	389 (25.7)	22.8-28.6	1.5	—	—	—	—
30-39	2528	446 (17.6)	15.8-19.4	0.9	*0.62 (0.53-0.72)*	<.001	*0.66 (0.56-0.77)*	<.001
40-49	2259	426 (18.9)	16.9-20.9	1.0	*0.67 (0.57-0.78)*	<.001	*0.79 (0.68-0.93)*	.005
50-59	2760	583 (21.1)	19.1-23.1	1.0	*0.77 (0.67-0.90)*	.001	0.88 (0.76-1.03)	.10
60-69	3102	773 (24.9)	22.9-26.9	1.0	0.96 (0.83-1.10)	.54	1.00 (0.86-1.16)	.97
≥70	1787	389 (21.8)	19.4-24.2	1.2	*0.80 (0.68-0.94)*	.007	*0.74 (0.62-0.87)*	<.001

[a]The table shows the proportion of individuals meeting the 2018 Physical Activity Guidelines for Americans (PAGA) aerobic activity guidelines and results from logistic univariate and multivariate regression analyses used to show the impact of various demographic factors on the probability of meeting the 2018 PAGA.

[b]MVPA: moderate- and vigorous-intensity aerobic physical activity.

[c]Multivariate regression adjusts for sex, race or ethnicity, BMI, and age.

[d]Not applicable (reference groups).

[e]Italicized odds ratios (95% CI) reflect *P* values less than .05.

[f]BMI was calculated as the average of all measured values measured between the first and last valid weeks for each participant. If no measured values were available, BMI was categorized as "other or not specified." Individuals with a BMI under 18.5 kg/m^2 are categorized as "other or not specified" due to the small sample size.

We found that male individuals were more likely than female individuals to meet PAGA (1069/4007, 26.7% vs 1864/9553, 19.5%, respectively; *P*<.0001). Among racial or ethnic groups, non-Hispanic Asian or Pacific Islander individuals had higher adherence compared to Hispanic (135/430, 31.4% vs 60/310, 19%, respectively; *P*=.0003), non-Hispanic White (2454/11,109, 22.1%; *P*<.0001), and non-Hispanic Black (87/684, 13%; *P*<.0001) individuals. Non-Hispanic White individuals also had higher adherence than non-Hispanic Black (*P*<.0001) individuals. Regarding BMI, there were significant differences in adherence across normal weight (727/1994, 36.5%), overweight (598/2487, 24%), obese (224/1501, 14.9%), and severely obese groups (88/1375, 6.4%; *P*<.0001 for each pairwise comparison), with adherence decreasing as BMI

XSL•FO
RenderX

increased. In terms of age, the 18-29 years age group showed the highest adherence (389/1511, 25.7%), significantly higher than the 30-39 years age group (446/2528, 17.6%; *P*<.0001), 40-49 years age group (426/2259, 18.9%; *P*<.0001), and 50-59 years age group (583/2760, 21.1%; *P*=.0007). The 30-39 years age group adhered less than both the 60-69 years age group (773/3102, 24.9%; *P*<.0001) and ≥70 years age group (389/1787, 21.8%; *P*=.0008). The 40-49 years age group and 50-59 years age group both had lower adherence in comparison to the 60-69 years age group (*P*<.0001 and *P*=.0007, respectively). A more detailed breakdown of the average time spent in MVPA by sex and age group and by sex and BMI is presented in Figure 2. Similarly, the average daily step count by age group and sex and by sex and BMI is also presented in Figure 2. Multimedia Appendix 2 presents minutes per week (means and SEs) of LPA, MPA, VPA, and MVPA among US adults by sex, age group, ethnicity, and BMI.

Figure 2. Weekly minutes of MVPA (computed via step count algorithm) and daily step counts across sex, age, and BMI. Total weekly minutes of MVPA are defined as minutes of moderate activity + 2×minutes of vigorous activity, shown segmented by (A) age and sex and (B) BMI and sex. Daily step counts are shown segmented by (C) age and sex and (D) BMI and sex. All 4 panels show 25th percentile, median, and 75th percentile values in box plot format. The lower whisker extends to the smallest value, which is greater than or equal to the 25th percentile value – 1.5×IQR. The upper whisker extends to the largest value, which is less than or equal to the 25th percentile value + 1.5×IQR. BMI was calculated as the average of all measured values between the first and last valid weeks for each participant. If no measured values were available, BMI was categorized as "other or not specified." Individuals with a BMI under 18.5 kg/m2 are categorized as "other or not specified" due to the small sample size. MVPA: moderate- and vigorous-intensity aerobic physical activity.

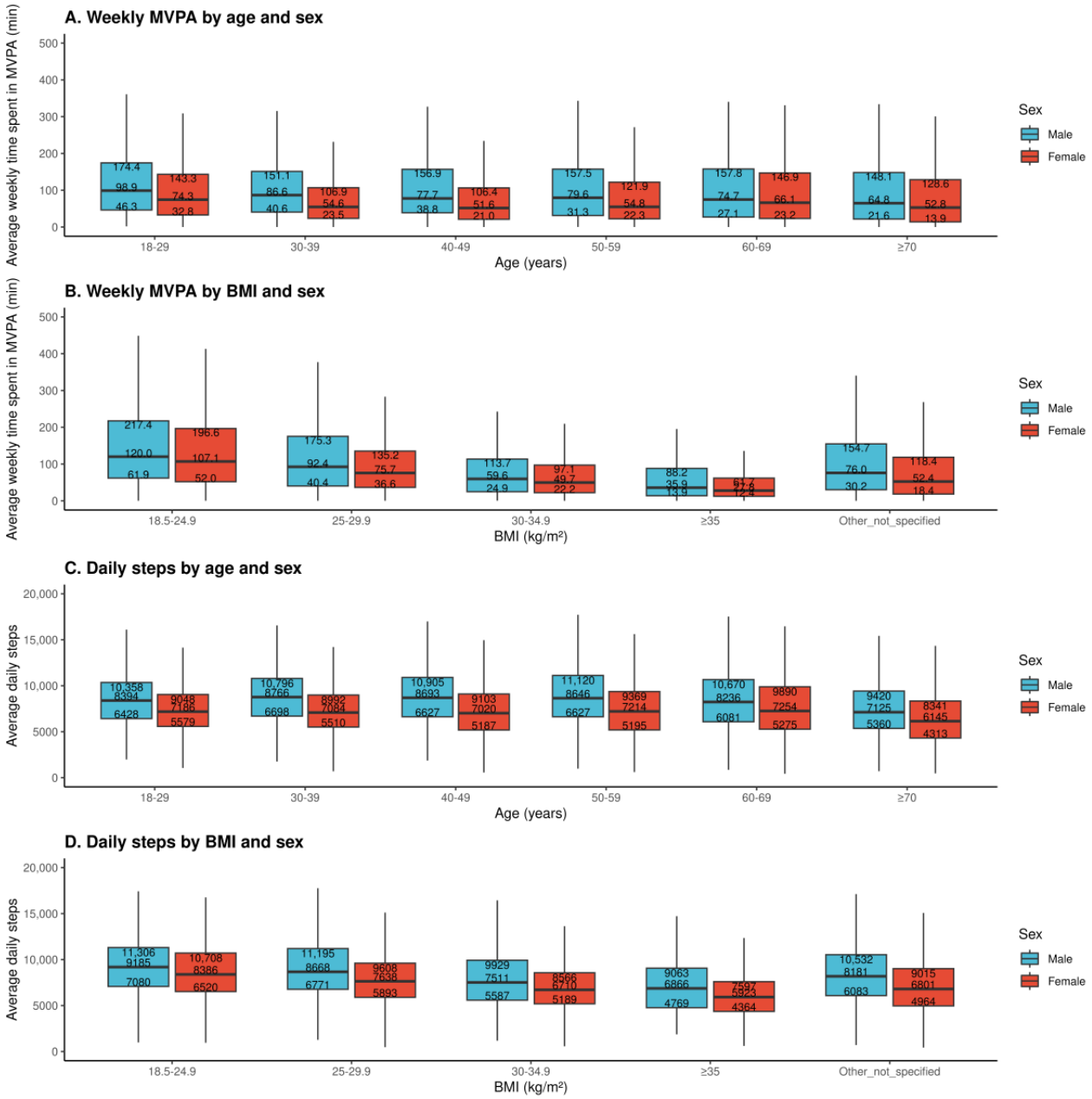

XSL•FO
RenderX

Aim 2: Sociodemographic Influences on Adherence to the 2018 PAGA

Univariate and multivariate logistic regression analyses identified sex, race or ethnicity, BMI, and age as significantly associated with meeting the 2018 PAGA at the 0.05 α level, as indicated by 2-sided tests (Table 2). Univariate analysis indicated that female individuals were significantly less likely than male individuals to meet the 2018 PAGA, with a univariate OR of 0.67 (95% CI 0.61-0.73). Non-Hispanic Black individuals also had a lower likelihood of adherence when compared to non-Hispanic White individuals, with a univariate OR of 0.51 (95% CI 0.41-0.64). Conversely, non-Hispanic Asian individuals were more likely to meet the guidelines than non-Hispanic White individuals in univariate analysis (OR 1.61, 95% CI 1.31-1.98). Higher BMI was inversely associated with decreased PA, and the univariate analysis found significant differences in adherence across all age groups except the 60-69 years age group compared to the 18-29 years reference group.

The multivariate analysis largely agreed with the univariate analysis. Both female and non-Hispanic Black individuals were less likely to meet the 2018 PAGA guidelines than their reference groups, with multivariate ORs of 0.66 (95% CI 0.60-0.72) and 0.63 (95% CI 0.50-0.79), respectively. Multivariate analysis also indicated that non-Hispanic Asian individuals were more likely to meet the guidelines than non-Hispanic White individuals, with an OR of 1.35 (95% CI 1.08-1.67). As with the univariate analysis, higher BMI was inversely associated with decreased PA. However, in multivariate analysis, only the 30-39, 40-49, and ≥70 years age groups showed significant differences compared to the reference group (18-29 years age group).

Aim 3: Evaluating 3 Methods for Assessing Compliance With the 2008 PAGA

All 3 methods (step intensity, HR, and Fitbit proprietary) showed significant differences in the total percentage of individuals meeting the 2008 PAGA, with *P* values less than .0001 for each comparison. The Fitbit algorithm indicated the highest adherence in our sample (10,307/13,947, 73.9%; 95% CI 71.2-76.6), followed by the HR-based algorithm (4740/13,947, 34%; 95% CI 32.8-35.2) and then step intensity–based algorithm (1401/13,947, 10%; 95% CI 9.4-10.6; Table 3). All pairwise comparisons across the 3 algorithms by each sociodemographic variable also showed significant differences ($P<.0001$ for each comparison), with 1 notable exception: for the 18-29 years age group, the HR method and the step-intensity algorithm were not statistically significantly different ($P=.39$).

XSL•FO
RenderX

Table 3. Percentage of US adults who meet 2008 physical activity guidelines based on Fitbit proprietary algorithm, heart rate (HR), and step counts from accelerometer data[a].

	Total participants, n	Participants with MVPA[b] from algorithm			Participants with MVPA from HR values			Participants with MVPA from step counts		
		n (%)	95% CI[c] (%)	SE (%)	n (%)	95% CI[c] (%)	SE (%)	n (%)	95% CI[c] (%)	SE (%)
Total	13,947	10,307 (73.9)	71.2-76.6	1.4	4740 (34.0)	32.8-35.2	0.6	1401 (10.0)	9.4-10.6	0.3
Sex										
Male	4007	3600 (89.8)	80.6-99.0	4.7	1476 (36.8)	24.4-39.2	1.2	544 (13.6)	12.4-14.8	0.6
Female	9553	6425 (67.3)	64.4-70.2	1.5	3115 (32.6)	31.2-34.0	0.7	822 (8.6)	8.0-9.2	0.3
Other or not specified	387	282 (72.9)	56.6-89.2	8.3	149 (38.5)	30.7-46.3	4.0	35 (9.0)	5.9-12.1	1.6
Race or ethnicity										
Hispanic	310	223 (71.9)	54.1-89.7	9.1	66 (21.3)	15.4-27.2	3.0	23 (7.4)	4.3-10.5	1.6
Non-Hispanic Asian or Pacific Islander	430	360 (83.7)	62.3-100.0	10.9	124 (28.8)	22.7-34.9	3.1	56 (13.0)	9.3-16.7	1.9
Non-Hispanic Black	684	428 (62.6)	53.0-77.2	4.9	232 (33.9)	28.6-39.2	2.7	41 (6.0)	4.0-8.0	1.0
Non-Hispanic White	11,109	8302 (74.7)	71.6-77.8	1.6	3913 (35.2)	33.8-36.6	0.7	1185 (10.7)	10.1-11.3	0.3
Two or more races	293	200 (68.3)	51.4-85.2	8.6	61 (20.8)	14.9-26.7	3.0	<20 (6.8)	N/A[d]	1.3
Other or not specified	1121	794 (70.8)	61.6-80.0	4.7	344 (30.7)	26.8-34.6	2.0	83 (7.4)	5.8-9.0	0.8
BMI (kg/m^2)[e]										
18.5-24.9	1994	1647 (82.6)	73.0-92.2	4.9	879 (44.1)	40.2-48.0	2.0	364 (18.3)	16.1-20.5	1.1
25-29.9	2487	2034 (81.8)	73.6-90.0	4.2	986 (39.6)	36.5-42.7	1.6	275 (11.1)	9.7-12.5	0.7
30-34.9	1501	1109 (73.9)	65.5-82.3	4.3	463 (30.8)	27.5-34.1	1.7	96 (6.4)	5.0-7.8	0.7
≥35	1375	856 (62.3)	55.4-69.2	3.5	324 (23.6)	20.7-26.5	1.5	25 (1.8)	1.0-2.6	0.4
Other or not specified	6590	4661 (70.7)	67.0-74.4	1.9	2088 (31.7)	30.1-33.3	0.8	641 (9.7)	8.9-10.5	0.4
Age group (years)										
18-29	1511	1106 (73.2)	64.8-81.6	4.3	119 (7.9)	6.3-9.5	0.8	108 (7.1)	5.7-8.5	0.7
30-39	2528	1798 (71.1)	65.0-77.2	3.1	323 (12.8)	11.2-14.4	0.8	161 (6.4)	5.4-7.4	0.5
40-49	2259	1570 (69.5)	63.2-75.8	3.2	547 (24.2)	21.8-26.6	1.2	188 (8.3)	7.1-9.5	0.6
50-59	2760	2022 (73.3)	67.0-79.6	3.2	1068 (38.7)	35.8-41.6	1.5	285 (10.3)	9.1-11.5	0.6
60-69	3102	2431 (78.4)	71.7-85.1	3.4	1639 (52.8)	49.1-56.5	1.9	428 (13.8)	12.4-15.2	0.7
≥70	1787	1380 (77.2)	68.6-85.8	4.4	1044 (58.4)	52.9-63.9	2.8	231 (12.9)	11.1-14.7	0.9

[a]This table shows the proportion of individuals meeting the 2008 Physical Activity Guidelines for Americans (PAGA) aerobic activity guidelines of 150 minutes a week of moderate to vigorous physical activity. Only activity bouts ≥10 minutes in duration included in determining physical activity.

[b]MVPA: moderate- and vigorous-intensity aerobic physical activity.

[c]95% Wald CIs shown for the proportion of each demographic group meeting the 2008 PAGA aerobic activity guidelines. CIs capped at a maximum value of 100%.

XSL•FO
RenderX

[d]All of Us policy prevents the display of any participant count less than 20.

[e]BMI was calculated as the average of all measured values measured between the first and last valid weeks for each participant. If no measured values were available, BMI was categorized as "other or not specified." Individuals with a BMI under 18.5 kg/m^2 are categorized as "other or not specified" due to the small sample size.

Discussion

Principal Findings

We found that approximately one-fifth (3006/13,947, 21.6%) of our sample met the 2018 PAGA using the step count algorithm, and the average minutes of MVPA per week (101.4 minutes) was nearly 50 minutes lower than recommended. After controlling for confounders, we found that several demographic factors were independently associated with meeting PA guidelines. Women were less likely to meet PA guidelines than men, and higher BMI was shown to be inversely associated with adhering to the 2018 PAGA. Race or ethnicity and age also showed strong associations, as non-Hispanic Black individuals were less likely than non-Hispanic White individuals to meet guidelines, whereas non-Hispanic Asian individuals were more likely to meet guidelines. Interestingly, age showed a more nuanced relationship with meeting guidelines, with PA declining until middle age, then increasing briefly, before declining again in old age (≥70 years) when compared to the 18-29 years reference group.

Comparison With Prior Work

Our report showed substantially lower proportions of participants meeting the 2018 PAGA compared to a report completed by the National Center for Health Statistics (NCHS; which showed that ~47% of US adults met the latest aerobic PAGA in 2020; N=31,568) [45] but higher proportions compared to previous studies completed by Tucker et al [10] (N=3082) as well as Troiano et al [9] (N=6329), which evaluated adherence to the 2008 PAGA and averaged around 5%. The lower proportion in comparison to NCHS reports was likely explained by response bias in surveys, as surveys have been shown to result in higher proportions of reported PA compared to objective measures [15,16]. In addition, studies have also shown that individuals are more likely to overreport PA for social acceptance reasons [46]. Finally, pedometer-measured PA may not capture activities that exclude or limit stepping (eg, yoga, tai chi, bicycling, and swimming), thus limiting the ability to accurately estimate PA levels [46]. Conversely, we observed higher proportions of participants meeting the PAGA in comparison to earlier studies using activity monitor data. We hypothesize that this may be caused by several factors. First, many previous studies measured adherence to a previous iteration of the 2008 PAGA, where only bouts of 10-minute MVPA are included [9,10]. Aligned with this observation, our estimates of adherence fall to 10% (95% CI 9.4-10.6) when following the 2008 PAGA 10-minute bout requirement. Second, selection bias is also likely to have occurred in this study; since individuals were self-selecting to participate in the AoU study and had chosen to purchase activity monitors of their own volition, they may have been more likely to be physically active compared to randomly sampled US adults (eg, National Health and Nutrition Examination Survey). Third, the convenience sampling strategy of our study likely resulted in a sample distribution that was not demographically reflective of the general US population. Finally, while prevalence estimates remain low, longitudinal studies have shown an increase in aerobic PA over time, as general PA levels have increased over the past 30 years [5]. Given that it has been over a decade since accelerometers have been used in the National Health and Nutrition Examination Survey [47], there is a unique opportunity to recreate this approach in a nationally representative sample, which could allow us to gain further insights into this large Fitbit database.

When evaluating sociodemographic associations, our findings align with previous studies regarding the relationship between sex, BMI, and PA. This alignment was observed both in self-reported data and data derived from accelerometry [9,10,12,45,48]. We found that women were significantly less physically active than men. Although further research is needed to fully understand this phenomenon, existing studies suggest that various factors contribute to lower PA participation among women. These factors include lower levels of motivation to participate in PA, diminished levels of peer and early parental social support for PA, and societal expectations that encourage women to prioritize other responsibilities, such as domestic duties and childcare [49]. We also found that higher BMI was inversely associated with PA. This finding also aligns with both self-reported data and data derived from accelerometry [5,7,12,48]. This phenomenon has been well studied; prior studies have established that higher levels of PA cause a reduction in BMI and that higher BMI also serves as a barrier to PA participation [50,51].

For age, our study showed slightly different associations compared to prior studies. Prior cross-sectional studies have shown a negative correlation between increased age and PA at all age ranges [8-12,48]. However, our study showed a decline in PA when compared to a reference group aged 18-29 years until age 50 years, at which point, PA increased back to proportions seen in young adulthood, then declined again after age 70 years. We hypothesize several possible mechanisms of action for this phenomenon. Some targeted studies focusing on older adults have shown an increase in leisure time PA following retirement. Given the 60-69 years age group falls into retirement age, this may play a factor in a temporary resurgence of PA [52]. Studies have also shown a lower proportion of PA among adults with children versus child-free adults. As children leave their households, individuals may have more time and energy to exercise [53].

Regarding race, we found that non-Hispanic Black individuals were less likely to meet the 2018 PAGA guidelines compared to non-Hispanic White individuals. This finding is consistent with most prior studies [10,54,55], which demonstrated similar racial disparities in PA levels using both objective measures (eg, accelerometry) [10] and subjective self-report methods [54,55]. Tucker et al [10] evaluated adherence to the 2008 PAGA using accelerometry data (N=4773) and found that 8%

XSL•FO
RenderX

(1.40% SE) of non-Hispanic Black individuals adhered to the 2008 PAGA while 10.1% (1.16% SE) of non-Hispanic White individuals adhered to the 2008 PAGA (statistical tests for significance were not conducted). Wilson-Frederick et al [54] found in a national survey–based study (N=67,790) that Black individuals had higher adjusted odds of physical inactivity (with physical inactivity defined as 0 minutes of weekly MVPA) compared to White individuals after controlling for confounders (adjusted odds ratio [AOR] 1.40, 95% CI 1.30-1.51). Similarly, a Centers for Disease Control and Prevention report using self-report data from the Behavioral Risk Factor Survey System administered from 2017 through 2020 (N=~1.6 million, exact sample size not disclosed) also showed that non-Hispanic Black individuals had a 30% prevalence of physical inactivity outside of work compared to 23% for non-Hispanic White individuals (statistical tests for significance were not conducted) [55]. In contrast, Matthews et al [11], in a 2019 study (N=2640), found no significant differences in daily PA levels between non-Hispanic Black and non-Hispanic White individuals in a study where PA was self-reported. The divergence between our findings and those of Matthews et al [11] may be attributed to methodological differences, including their smaller sample size, their use of 24-hour survey recalls administered on a single day (which have been shown to result in biased estimates of MVPA at both group and individual level in adults) [19,56,57], and the Activities Collected Over Time Over 24-hours survey instrument used in their study, which has not been validated across all race and ethnic groups investigated.

Our findings showed that non-Hispanic Asian individuals were more likely to adhere to the 2018 PAGA when compared to non-Hispanic White individuals. Although few objective accelerometry-based studies evaluating PA prevalence in Asian American individuals have been conducted, our findings align with a Centers for Disease Control and Prevention report using self-report data from the Behavioral Risk Factor Survey System administered from 2017 through 2020 (N=~1.6 million, exact sample size not disclosed). This report found that non-Hispanic Asian individuals had a 20.1% prevalence of physical inactivity outside of work compared to 23% for non-Hispanic White individuals when measured using surveys (statistical tests for significance were not conducted). In contrast to our findings, Stella et al [58] conducted a survey study in 2 major US cities (Los Angeles: N=17,462 and New York City: N=8036) and found that non-Hispanic White individuals were more likely than non-Hispanic Asian individuals to adhere to the 2018 PAGA guidelines after adjusting for confounders (AOR 1.35, 95% CI 1.09-1.68 and AOR 1.45, 95% CI 1.13-1.86, respectively). The divergence between our findings and those of Stella et al [58] may be attributed to methodological differences, including their use of subjective PA measures, which are generally less accurate than objective methods [15-18,20,21], and their study's exclusive focus on large metropolitan areas. We theorize that family structures and dynamics in Asian households may contribute to the higher PA rates observed among non-Hispanic Asian individuals in our study, although additional research is needed to confirm this relationship. Prior studies have suggested that single-parent households lead to reductions in PA participation for children [59], that childhood PA behavior patterns are strongly predictive of PA patterns in adulthood [60], and that middle-aged adults with strong family support networks are more likely to engage in PA [61]. Asian American individuals have among the lowest single-parentage rates and the highest proportion of individuals living in intergenerational family households, which may play a role in influencing lifelong PA behavior patterns [62].

When comparing the Fitbit proprietary algorithm, the HR-based algorithm, and the step intensity–based algorithm, we found significant discrepancies in measuring adherence to the 2008 PAGA. The Fitbit proprietary algorithm indicated the highest adherence proportion at 73.9% (10,307/13,947), followed by the HR algorithm at 34% (4740/13,947) and the step intensity algorithm at 10% (1401/13,947). Notably, the Fitbit proprietary algorithm showed a nearly 7-fold higher proportion of individuals meeting the 2008 PAGA compared to the step-intensity algorithm.

Our findings regarding the Fitbit algorithm align with prior validation studies, which have demonstrated consistent overestimation of MVPA compared to validated accelerometry-based methods in free-living environments [39,63-65]. The degree of overestimation has been estimated to be as much as an additional hour of MVPA per day, which could lead to notable changes in the estimated prevalence of adherence to PAGA [39]. This overestimation is concerning, as it could also lead to a significant portion of the US population being incorrectly classified as meeting PA guidelines. Consequently, these individuals might not be targeted for necessary PA interventions, potentially increasing their risk for adverse health outcomes. Furthermore, evidence suggests that the degree of overestimation increases as the volume of MVPA increases, suggesting that the accuracy of MVPA measurement is further reduced in active individuals [39]. Rosenberger et al [63] conducted a study to compare Fitbit against ActiGraph measures in a 24-hour wear protocol and found that Fitbit's estimates of daily MVPA had a mean absolute percentage error greater than 60% compared to the Actigraph [63]. Similarly, Feehan et al [65] found in a systematic review comparing Fitbit devices against validated Actical and ActiGraph activity monitors that the median difference in time spent in light to vigorous PA ranged from 52% to 390% across 8 separate studies. This overestimation has been found to occur in both wrist-worn and hip-worn Fitbit devices [64]. Given the increasing adoption of Fitbit devices in both the general population and clinical settings, it is crucial to accurately assess their measurement precision and identify the factors contributing to any inaccuracies.

Similarly, the HR-based algorithm also showed meaningfully higher estimates of PA than the step intensity–based algorithm. We hypothesize that this finding may primarily be due to two reasons. First, the commonly used formula for calculating maximum HR tends to systematically underestimate the maximum HR in older individuals, leading to a systemic upward bias in PA measurements for this demographic [28]. Second, the HR algorithm ignores the impact of sex on determining maximum HR, which may result in systemic overestimation of PA in women since women have higher HRs than men [66]. Overestimation of PA levels is a problem because it can lead to the underidentification of individuals who may benefit from

XSL•FO
RenderX

increased activity, thus increasing their risk of a plethora of adverse health events. This is particularly important, as Fitbit and other commercial activity monitors are being integrated into the electronic medical record to support physician monitoring of patient health and inform medical decisions [25,67].

Strengths and Limitations

The principal strength of this study is the extensive size of the sample. We analyzed an unprecedented 1.6 million weeks of data collected over 7 years, involving over 13,000 participants. This extensive dataset enabled us to examine sociodemographic associations with adherence to PA guidelines and to assess differences in algorithms for measuring PA in an ecologically valid manner. This ecological validity was further strengthened by our ability to use objective, accelerometry-derived measures rather than self-reported data to analyze PA. Finally, the paired nature of the data—comprising MVPA as measured by Fitbit, step count, and HR—enabled a statistically robust analysis across a very large dataset to compare the outcomes from each algorithm. However, this study also presents several limitations. First, the "Bring Your Own Device" approach of the AoU program likely resulted in a sample that is not nationally representative of US adults. While our analysis controlled for age, race, BMI, and sex, we did not address potential bias arising from participants possibly being more active due to owning activity monitors. Although the AoU program is piloting initiatives to distribute Fitbit devices to increase sample diversity, this information is not yet publicly available. Second, the sample was not reweighted to reflect demographic characteristics for the general American population. It is important to note that although our multivariate analysis controlled for sex, race or ethnicity, BMI, and age, the raw adherence proportions shown do not represent the target proportions, as our population was older, White, and more female compared to the general US population. Third, participants used various Fitbit models with differing firmware, which may have introduced measurement discrepancies due to differences in device and proprietary analyses. Fourth, the analysis did not control for confounding by income or educational status, which have been shown to be associated with PA levels [42]. Fifth, although the step-intensity method was chosen as the most accurate method for estimating PA and determining the 2018 PAGA adherence when using Fitbit devices, the accuracy of the device at measuring step counts degrades during activities involving slow ambulation and in individuals with mobility limitations [65]. Finally, we could not accurately remove sleep time from the wear log since we did not have sleep diaries, minute-level accelerometry count data, or validated algorithms that could be used to determine sleep times using Fitbit data. Future research should consider using sleep logs or raw accelerometry data to assess PA levels more accurately.

Conclusions

Overall, our study resulted in three primary conclusions: (1) adherence proportions to the 2018 PAGA are lower than indicated by NCHS reports and differ by sociodemographic classification; (2) key sociodemographic characteristics, including age, BMI, race, and sex, are strongly correlated with adherence proportions to PA guidelines as outlined by PAGA; and (3) various algorithms based on proprietary Fitbit data, HR, and step count produce significantly different outcomes in measuring PA. These findings have important implications. Given the significant differences in meeting guidelines across age, sex, race, and BMI, PA interventions should account for these disparities during implementation. Given that our sample consists of individuals who already own fitness trackers, a promising avenue for PA interventions is a mobile health (mHealth) approach. By harnessing digital technologies, mHealth has the potential to enhance the accessibility, engagement, and personalization of PA interventions. Mobile platforms, such as apps and websites, can address key social determinants of health by overcoming geographical, logistical, and financial barriers. These platforms offer continuous support and near real-time monitoring of PA and factors that could impact one's activity levels (eg, stress levels, mood, and symptom burden), providing interventionists with more accurate and near real-time PA tracking to tailor prescriptions [25,68,69]. mHealth solutions can customize the content, include behavior change techniques into their functionality, and provide exercise recommendations to meet individual needs, potentially boosting engagement and effectiveness [25,69-71]. This approach could be used to intervene in the sociodemographic groups that were less likely to adhere to the 2018 PAGA guidelines (ie, women, higher BMI individuals, non-Hispanic Black individuals, and those aged 30-49 or ≥70 years). Focusing these applications on those groups that would most benefit from increases in PA may lead to significant public health benefits.

Additionally, the algorithm used for estimating PA can have a significant effect on the accuracy of estimates. Thus, it is crucial for interventionists to recognize and adjust for potential systemic biases caused by different algorithms. Finally, health care systems should also be cautious about relying on "black box" algorithms, as the wide variability in results may interfere with decision-making accuracy. We are optimistic that the insights and results detailed in this paper will drive others to use more targeted and effective strategies to boost adherence to PA guidelines and use evidence-based methods for processing PA data.

Acknowledgments

The authors gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. The authors also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request. This study used data from the All of Us Research Program's Controlled Tier Dataset (version 7), available to authorized users on the Researcher Workbench.

Authors' Contributions

RS conceived the study idea, conducted data and statistical analysis, interpreted data, and wrote and revised the manuscript. MKT conceived the study idea and wrote and revised the manuscript. JLF conducted data and statistical analysis, interpreted data, and wrote and revised the manuscript. EDP, FJP, JY, PW, and SKC participated in scientific discussions and wrote and revised the manuscript. RMB conceived the study idea, conducted data and statistical analysis, interpreted data, wrote and revised the manuscript, and supervised the project. RMB serves as the corresponding author. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

EDP is an MPI of grants to the institution not related to this work from Genentech, Pfizer, Astra Zeneca, and Guardent Health and is on advisory boards for GSK and Merck. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) diagram.
[PDF File (Adobe PDF File), 92 KB - jmir_v26i1e65095_app1.pdf]

Multimedia Appendix 2
Minutes per week (means and SEs) of sedentary, light-, moderate-, and vigorous-intensity physical activity among US adults by sex, age group, ethnicity, and BMI according to various algorithms.
[PDF File (Adobe PDF File), 227 KB - jmir_v26i1e65095_app2.pdf]

References

1. Physical Activity Guidelines Advisory Committee scientific report 2008. US Department of Health and Human Services. 2008. URL: https://odphp.health.gov/sites/default/files/2019-09/paguide.pdf [accessed 2024-05-01]
2. 2018 Physical Activity Guidelines Advisory Committee scientific report. US Department of Health and Human Services. 2018. URL: https://odphp.health.gov/sites/default/files/2019-09/Physical_Activity_Guidelines_2nd_edition.pdf [accessed 2024-05-01]
3. Marquez DX, Aguiñaga S, Vásquez PM, Conroy DE, Erickson KI, Hillman C, et al. A systematic review of physical activity and quality of life and well-being. Transl Behav Med 2020 Oct 12;10(5):1098-1109 [FREE Full text] [doi: 10.1093/tbm/ibz198] [Medline: 33044541]
4. Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, et al. Association of step counts over time with the risk of chronic disease in the All of Us Research Program. Nat Med 2022 Nov;28(11):2301-2308 [FREE Full text] [doi: 10.1038/s41591-022-02012-w] [Medline: 36216933]
5. Hyde E, Whitfield G, Omura J, Fulton JE, Carlson SA. Trends in meeting the Physical Activity Guidelines: muscle-strengthening alone and combined with aerobic activity, United States, 1998-2018. J Phys Act Health 2021 Aug 01;18(S1):S37-S44 [FREE Full text] [doi: 10.1123/jpah.2021-0077] [Medline: 34465652]
6. Dong L, Block G, Mandel S. Activities contributing to total energy expenditure in the United States: results from the NHAPS study. Int J Behav Nutr Phys Act 2004 Feb 12;1(1):4 [FREE Full text] [doi: 10.1186/1479-5868-1-4] [Medline: 15169563]
7. Berkemeyer K, Wijndaele K, White T, Cooper AJM, Luben R, Westgate K, et al. The descriptive epidemiology of accelerometer-measured physical activity in older adults. Int J Behav Nutr Phys Act 2016;13:2 [FREE Full text] [doi: 10.1186/s12966-015-0316-z] [Medline: 26739758]
8. Matthews CE, Chen KY, Freedson PS, Buchowski MS, Beech BM, Pate RR, et al. Amount of time spent in sedentary behaviors in the United States, 2003-2004. Am J Epidemiol 2008;167(7):875-881 [FREE Full text] [doi: 10.1093/aje/kwm390] [Medline: 18303006]
9. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
10. Tucker JM, Welk GJ, Beyler NK. Physical activity in U.S.: adults compliance with the Physical Activity Guidelines for Americans. Am J Prev Med 2011;40(4):454-461. [doi: 10.1016/j.amepre.2010.12.016] [Medline: 21406280]
11. Matthews CE, Patel S, Saint-Maurice PF, Loftfield E, Keadle SK, Chen KY, et al. Physical activity levels (PAL) in US adults—2019. Med Sci Sports Exerc 2023;55(5):884-891 [FREE Full text] [doi: 10.1249/MSS.0000000000003102] [Medline: 36729673]

XSL•FO
RenderX

12. Bassett D, Wyatt H, Thompson H, Peters JC, Hill JO. Pedometer-measured physical activity and health behaviors in U.S. adults. Med Sci Sports Exerc 2010;42(10):1819-1825. [doi: 10.1249/MSS.0b013e3181dc2e54] [Medline: 20305579]
13. Omura JD, Whitfield GP, Chen TJ, Hyde ET, Ussery EN, Watson KB, et al. Surveillance of physical activity and sedentary behavior among youth and adults in the United States: history and opportunities. J Phys Act Health 2021;18(S1):S6-S24 [FREE Full text] [doi: 10.1123/jpah.2021-0179] [Medline: 34465651]
14. Matthews CE, Saint-Maurice P, Fulton JE, Patel S, Loftfield E, Sampson JN, et al. Changes in physical activity and sedentary time in United States adults in response to COVID-19. PLoS One 2022;17(9):e0273919 [FREE Full text] [doi: 10.1371/journal.pone.0273919] [Medline: 36084056]
15. Janz KF. Physical activity in epidemiology: moving from questionnaire to objective measurement. Br J Sports Med 2006;40(3):191-192 [FREE Full text] [doi: 10.1136/bjsm.2005.023036] [Medline: 16505072]
16. Sallis JF, Saelens BES. Assessment of physical activity by self-report: status, limitations, and future directions. Res Q Exerc Sport 2000;71(Suppl 2):1-14. [doi: 10.1080/02701367.2000.11082780] [Medline: 25680007]
17. Vanhees L, Lefevre J, Philippaerts R, Martens M, Huygens W, Troosters T, et al. How to assess physical activity? How to assess physical fitness? Eur J Cardiovasc Prev Rehabil 2005;12(2):102-114. [doi: 10.1097/01.hjr.0000161551.73095.9c] [Medline: 15785295]
18. Durante R, Ainsworth BE. The recall of physical activity: using a cognitive model of the question-answering process. Med Sci Sports Exerc 1996;28(10):1282-1291. [doi: 10.1097/00005768-199610000-00012] [Medline: 8897386]
19. Uitenbroek DG. Seasonal variation in leisure time physical activity. Med Sci Sports Exerc 1993;25(6):755-760. [doi: 10.1249/00005768-199306000-00017]
20. Klesges RC, Eck LH, Mellon MW, Fulliton W, Somes GW, Hanson CL. The accuracy of self-reports of physical activity. Med Sci Sports Exerc 1990;22(5):690-697. [doi: 10.1249/00005768-199010000-00022] [Medline: 2233209]
21. Baranowski T, Dworkin RJ, Cieslik CJ, Hooks P, Clearman DR, Ray L, et al. Reliability and validity of self report of aerobic activity: family health project. Res Q Exerc Sport 1984;55(4):309-317. [doi: 10.1080/02701367.1984.10608408]
22. Matthews CE, Hagströmer M, Pober DM, Bowles HR. Best practices for using physical activity monitors in population-based research. Med Sci Sports Exerc 2012;44(1 Suppl 1):S68-S76 [FREE Full text] [doi: 10.1249/MSS.0b013e3182399e5b] [Medline: 22157777]
23. Sylvia LG, Bernstein EE, Hubbard JL, Keating L, Anderson EJ. Practical guide to measuring physical activity. J Acad Nutr Diet 2014 Feb;114(2):199-208 [FREE Full text] [doi: 10.1016/j.jand.2013.09.018] [Medline: 24290836]
24. Hartman SJ, Chen R, Tam RM, Narayan HK, Natarajan L, Liu L. Fitbit use and activity levels from intervention to 2 years after: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2022;10(6):e37086 [FREE Full text] [doi: 10.2196/37086] [Medline: 35771607]
25. Dinh-Le C, Chuang R, Chokshi S, Mann D. Wearable health technology and electronic health record integration: scoping review and future directions. JMIR Mhealth Uhealth 2019;7(9):e12861 [FREE Full text] [doi: 10.2196/12861] [Medline: 31512582]
26. Interoperability: bring together patient data across systems for better care coordination and improved health outcomes. Epic. URL: https://www.epic.com/software/interoperability/ [accessed 2024-10-25]
27. Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, et al. The "All of Us" Research Program. N Engl J Med 2019;381(7):668-676 [FREE Full text] [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
28. Shookster D, Lindsey B, Cortes N, Martin JR. Accuracy of commonly used age-predicted maximal heart rate equations. Int J Exerc Sci 2020;13(7):1242-1250 [FREE Full text] [Medline: 33042384]
29. Shcherbina A, Mattsson CM, Waggott D, Salisbury H, Christle JW, Hastie T, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med 2017;7(2):3 [FREE Full text] [doi: 10.3390/jpm7020003] [Medline: 28538708]
30. Jamnick NA, Pettitt RW, Granata C, Pyne DB, Bishop DJ. An examination and critique of current methods to determine exercise intensity. Sports Med 2020;50(10):1729-1756. [doi: 10.1007/s40279-020-01322-8] [Medline: 32729096]
31. Fitbit Help Center. Google. URL: https://support.google.com/fitbit [accessed 2024-10-25]
32. Karnes JH, Arora A, Feng J, Steiner HE, Sulieman L, Boerwinkle E, et al. Racial, ethnic, and gender differences in obesity and body fat distribution: an All of Us Research Program demonstration project. PLoS One 2021;16(8):e0255583 [FREE Full text] [doi: 10.1371/journal.pone.0255583] [Medline: 34358277]
33. Claudel SE, Tamura K, Troendle J, Andrews MR, Ceasar JN, Mitchell VM, et al. Comparing methods to identify wear-time intervals for physical activity with the Fitbit Charge 2. J Aging Phys Act 2021;29(3):529-535 [FREE Full text] [doi: 10.1123/japa.2020-0059] [Medline: 33326935]
34. Wing D, Godino JG, Baker FC, Yang R, Chevance G, Thompson WK, et al. Recommendations for identifying valid wear for consumer-level wrist-worn activity trackers and acceptability of extended device deployment in children. Sensors (Basel) 2022;22(23):9189 [FREE Full text] [doi: 10.3390/s22239189] [Medline: 36501894]
35. Chan A, Chan D, Lee H, Ng CC, Yeo AHL. Reporting adherence, validity and physical activity measures of wearable activity trackers in medical research: a systematic review. Int J Med Inform 2022;160:104696. [doi: 10.1016/j.ijmedinf.2022.104696] [Medline: 35121356]

XSL•FO
RenderX

36. Tudor-Locke C, Han H, Aguiar EJ, Barreira TV, Schuna JM, Kang M, et al. How fast is fast enough? Walking cadence (steps/min) as a practical estimate of intensity in adults: a narrative review. Br J Sports Med 2018;52(12):776-788 [FREE Full text] [doi: 10.1136/bjsports-2017-097628] [Medline: 29858465]
37. Deborah R, Jonathan KE, Gary L, Meir M. ACSM's Guidelines for Exercise Testing and Prescription. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins; 2021.
38. Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. J Am Coll Cardiol 2001;37(1):153-156 [FREE Full text] [doi: 10.1016/s0735-1097(00)01054-8] [Medline: 11153730]
39. Redenius N, Kim Y, Byun W. Concurrent validity of the fitbit for assessing sedentary behavior and moderate-to-vigorous physical activity. BMC Med Res Methodol 2019;19(1):29 [FREE Full text] [doi: 10.1186/s12874-019-0668-1] [Medline: 30732582]
40. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc 1943;54(3):426-428. [doi: 10.2307/1990256]
41. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dubl Phil Mag J Sci 2009;50(302):157-175. [doi: 10.1080/14786440009463897]
42. Trost SG, Owen N, Bauman AE, Sallis JF, Brown W. Correlates of adults' participation in physical activity: review and update. Med Sci Sports Exerc 2002;34(12):1996-2001. [doi: 10.1097/00005768-200212000-00020] [Medline: 12471307]
43. Kim JH. Multicollinearity and misleading statistical results. Korean J Anesthesiol 2019;72(6):558-569 [FREE Full text] [doi: 10.4097/kja.19087] [Medline: 31304696]
44. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153-157. [doi: 10.1007/BF02295996] [Medline: 20254758]
45. Elgaddal N, Kramarow EA, Reuben C. Physical activity among adults aged 18 and over: United States, 2020. NCHS Data Brief 2022;443:1-8. [Medline: 36043905]
46. Welk G. Physical Activity Assessments for Health-Related Research. Champaign, IL: Human Kinetics; 2002.
47. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/nhanes [accessed 2024-09-24]
48. Zenko Z, Willis EA, White DA. Proportion of adults meeting the 2018 Physical Activity Guidelines for Americans according to accelerometers. Front Public Health 2019;7:135. [doi: 10.3389/fpubh.2019.00135] [Medline: 31231627]
49. Edwards ES, Sackett SC. Psychosocial variables related to why women are less active than men and related health implications. Clin Med Insights Womens Health 2016;9(Suppl 1):47-56. [doi: 10.4137/cmwh.s34668]
50. Hamer O, Larkin D, Relph N, Dey P. Fear-related barriers to physical activity among adults with overweight and obesity: a narrative synthesis scoping review. Obes Rev 2021;22(11):e13307. [doi: 10.1111/obr.13307] [Medline: 34170596]
51. Carrasquilla GD, García-Ureña M, Fall T, Sørensen TIA, Kilpeläinen TO. Mendelian randomization suggests a bidirectional, causal relationship between physical inactivity and adiposity. Elife 2022;11:70386 [FREE Full text] [doi: 10.7554/eLife.70386] [Medline: 35254260]
52. Barnett I, van Sluijs EMF, Ogilvie D. Physical activity and transitioning to retirement: a systematic review. Am J Prev Med 2012;43(3):329-336 [FREE Full text] [doi: 10.1016/j.amepre.2012.05.026] [Medline: 22898127]
53. Hull EE, Rofey D, Robertson R, Nagle EF, Otto AD, Aaron DJ. Influence of marriage and parenthood on physical activity: a 2-year prospective analysis. J Phys Act Health 2010;7(5):577-583 [FREE Full text] [doi: 10.1123/jpah.7.5.577] [Medline: 20864752]
54. Wilson-Frederick SM, Thorpe RJ, Bell CN, Bleich SN, Ford JG, LaVeist TA. Examination of race disparities in physical inactivity among adults of similar social context. Ethn Dis 2014;24(3):363-369 [FREE Full text] [Medline: 25065080]
55. Adult physical inactivity outside of work. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/physical-activity/php/data/inactivity-maps.html [accessed 2024-02-01]
56. Garriga A, Sempere-Rubio N, Molina-Prados MJ, Faubel R. Impact of seasonality on physical activity: a systematic review. Int J Environ Res Public Health 2021;19(1):2 [FREE Full text] [doi: 10.3390/ijerph19010002] [Medline: 35010262]
57. Welk GJ, Kim Y, Stanfill B, Osthus DA, Calabro MA, Nusser SM, et al. Validity of 24-h physical activity recall: physical activity measurement survey. Med Sci Sports Exerc 2014;46(10):2014 [FREE Full text] [doi: 10.1249/MSS.0000000000000314] [Medline: 24561818]
58. Stella SY, Roberts C, Lightstone AS, Shih M, Trinh-Shevrin C. Disparities in meeting physical activity guidelines for Asian-Americans in two metropolitan areas in the United States. Ann Epidemiol 2015;25(9):656-660 [FREE Full text] [doi: 10.1016/j.annepidem.2015.05.002] [Medline: 26065343]
59. Duriancik DM, Goff CR. Children of single-parent households are at a higher risk of obesity: a systematic review. J Child Health Care 2019;23(3):358-369. [doi: 10.1177/1367493519852463] [Medline: 31129999]
60. Telama R, Yang X, Leskinen E, Kankaanpää A, Hirvensalo M, Tammelin T, et al. Tracking of physical activity from early childhood through youth into adulthood. Med Sci Sports Exerc 2014;46(5):955-962. [doi: 10.1249/MSS.0000000000000181] [Medline: 24121247]

XSL•FO
RenderX

61. Yuan SC, Weng SC, Chou MC, Tang YJ, Lee SH, Chen DY, et al. How family support affects physical activity (PA) among middle-aged and elderly people before and after they suffer from chronic diseases. Arch Gerontol Geriatr 2011;53(3):274-247. [doi: 10.1016/j.archger.2010.11.029] [Medline: 21194758]
62. The rise of Asian Americans. Pew Research Center. 2012 Jun 19. URL: https://www.pewresearch.org/social-trends/2012/06/19/the-rise-of-asian-americans/ [accessed 2024-05-01]
63. Rosenberger ME, Buman MP, Haskell WL, McConnell MV, Carstensen LL. Twenty-four hours of sleep, sedentary behavior, and physical activity with nine wearable devices. Med Sci Sports Exerc 2016;48(3):457-465. [doi: 10.1249/MSS.0000000000000778] [Medline: 26484953]
64. Reid RE, Insogna JA, Carver TE, Comptour AM, Bewski NA, Sciortino C, et al. Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. J Sci Med Sport 2017;20(6):578-582. [doi: 10.1016/j.jsams.2016.10.015] [Medline: 27887786]
65. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]
66. Prabhavathi K, Tamarai Selvi K, Poornima KN, Sarvanan A. Role of biological sex in normal cardiac function and in its disease outcome—a review. JCDR 2014:BE01-BE04. [doi: 10.7860/jcdr/2014/9635.4771]
67. Shannahan Md A, Shah Md A, Wright Ph D, Clements Md DS. Physician monitoring of FitBit use for patient health. Glob Adv Health Med 2021;10:21649561211018999 [FREE Full text] [doi: 10.1177/21649561211018999] [Medline: 34104577]
68. Penedo FJ, Fox RS, Oswald LB, Moreno PI, Boland CL, Estabrook R, et al. Technology-based psychosocial intervention to improve quality of life and reduce symptom burden in men with advanced prostate cancer: results from a randomized controlled trial. Int J Behav Med 2020;27(5):490-505 [FREE Full text] [doi: 10.1007/s12529-019-09839-7] [Medline: 31898309]
69. Børøsund E, Ehlers SL, Varsi C, Clark MM, Andrykowski MA, Cvancarova M, et al. Results from a randomized controlled trial testing StressProffen; an application-based stress-management intervention for cancer survivors. Cancer Med 2020;9(11):3775-3785 [FREE Full text] [doi: 10.1002/cam4.3000] [Medline: 32243717]
70. Buneviciene I, Mekary RA, Smith TR, Onnela J, Bunevicius A. Can mHealth interventions improve quality of life of cancer patients? A systematic review and meta-analysis. Crit Rev Oncol Hematol 2021;157:103123. [doi: 10.1016/j.critrevonc.2020.103123] [Medline: 33190065]
71. Khoo S, Mohbin N, Ansari P, Al-Kitani M, Müller AM. mHealth interventions to address physical activity and sedentary behavior in cancer survivors: a systematic review. Int J Environ Res Public Health 2021;18(11):5798 [FREE Full text] [doi: 10.3390/ijerph18115798] [Medline: 34071342]

Abbreviations

AOR: adjusted odds ratio
AoU: All of Us
CONSORT: Consolidated Standards of Reporting Trials
HR: heart rate
LPA: light-intensity aerobic physical activity
mHealth: mobile health
MPA: moderate-intensity aerobic physical activity
MVPA: moderate- and vigorous-intensity aerobic physical activity
NCHS: National Center for Health Statistics
NIH: National Institutes of Health
OR: odds ratio
PA: physical activity
PAGA: Physical Activity Guidelines for Americans
VPA: vigorous-intensity aerobic physical activity

XSL•FO
RenderX

Edited by YH Lin; submitted 05.08.24; peer-reviewed by Z Ehtesham, E Cohn; comments to author 26.09.24; revised version received 08.10.24; accepted 09.10.24; published 10.12.24.

Please cite as:
Singh R, Tetrick MK, Fisher JL, Washington P, Yu J, Paskett ED, Penedo FJ, Clinton SK, Benzo RM
Analysis of Physical Activity Using Wearable Health Technology in US Adults Enrolled in the All of Us Research Program: Multiyear Observational Study
J Med Internet Res 2024;26:e65095
URL: https://www.jmir.org/2024/1/e65095
doi:10.2196/65095
PMID:

©Rujul Singh, Macy K Tetrick, James L Fisher, Peter Washington, Jane Yu, Electra D Paskett, Frank J Penedo, Steven K Clinton, Roberto M Benzo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Experience and Extended Technology Acceptance Model in Commercial Health Care App Usage Among Patients With Cancer: Mixed Methods Study

Ye-Eun Park[1*], BA; Yae Won Tak[1*], MSc; Inhye Kim[2], BS; Hui Jeong Lee[3], MSN; Jung Bok Lee[4], PhD; Jong Won Lee[5], MD, PhD; Yura Lee[1], MD, PhD

[1]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Graduate Program of Industrial Pharmaceutical Science, Yonsei University, Seoul, Republic of Korea
[3]MediRama, Seoul, Republic of Korea
[4]Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[5]Division of Breast Surgery, Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Yura Lee, MD, PhD
Department of Information Medicine
Asan Medical Center
University of Ulsan College of Medicine
88 Olympic-ro 43-gil
Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 1498
Fax: 82 2 3010 2531
Email: haepary@amc.seoul.kr

Abstract

Background: The shift in medical care toward prediction and prevention has led to the emergence of digital health care as a valuable tool for managing health issues. Aiding long-term follow-up care for cancer survivors and contributing to improved survival rates. However, potential barriers to mobile health usage, including age-related disparities and challenges in user retention for commercial health apps, highlight the need to assess the impact of patients' abilities and health status on the adoption of these interventions.

Objective: This study aims to investigate the app adherence and user experience of commercial health care apps among cancer survivors using an extended technology acceptance model (TAM).

Methods: The study enrolled 264 cancer survivors. We collected survey results from May to August 2022 and app usage records from the app companies. The survey questions were created based on the TAM.

Results: We categorized 264 participants into 3 clusters based on their app usage behavior: short use (n=77), medium use (n=101), and long use (n=86). The mean usage days were 9 (SD 11) days, 58 (SD 20) days, and 84 (SD 176) days, respectively. Analysis revealed significant differences in perceived usefulness (P=.01), interface satisfaction (P<.01), equity (P<.01), and utility (P=.01) among the clusters. Structural equation modeling indicated that perceived ease-of-use significantly influenced perceived usefulness (β=0.387, P<.01), and both perceived usefulness and attitude significantly affected behavioral intention and actual usage.

Conclusions: This study showed the importance of positive user experience and clinician recommendations in facilitating the effective usage of digital health care tools among cancer survivors and contributing to the evolving landscape of medical care.

(J Med Internet Res 2024;26:e55176) doi:10.2196/55176

XSL•FO
RenderX

KEYWORDS

mHealth; user experience; cancer; technology acceptance model; structural equation modeling; health care app; mixed-method study; medical care; digital health care; cancer survivors; disparities; health status; behavioral intervention; clinician

Introduction

As the paradigm of medical care has shifted toward prediction, prevention, personalization, and participation, digital health care has emerged as a promising tool for managing health-related issues [1-3]. Digital health care has the potential to become an effective aid for long-term follow-up care among cancer survivors, contributing to increased survival rates [4-8]. While the evidence for the effectiveness of digital health care has been accumulated through well-designed clinical studies, there are potential barriers. Compared to 78% of cancer survivors who are aged 60 years or older, smartphone ownership and internet usage are lower among older people [9-12]. Also, users' digital literacy, socioeconomic status, and user interface (UI) of apps are hurdles to successful digital health intervention [13-17].

Though long-term use of health care apps has been linked to better outcomes, many commercial health apps face challenges in user retention [18,19]. To understand how we could make patients to use health care applications effectively, it is important to identify the factors influencing the user's adoption of digital interventions. The technology acceptance model (TAM) provides a theoretical framework that identifies and measures various factors influencing users' adoption and usage of digital technology; it is a useful tool for understanding technology acceptance [20,21]. Furthermore, accessibility, convenience, and self-efficacy are key facilitators for adopting digital health applications; thus, it is essential to examine the impact of the ability of patients with cancer to use mobile technologies and their health status on the usage of mobile health interventions [5,22,23].

There have been many studies on the process of developing apps for specific health problems, considering patients' needs and user-centered design [24]. However, from the perspective of older patients with cancer who use conventional commercial health management apps, this consideration is far from a real-world situation. In addition, when recommending or promoting commercial health apps to improve the quality of life of patients with cancer, there is a lack of evidence on what factors can make patients use them better and find them more helpful [8,25].

Therefore, the objective of this study is to investigate the association between user characteristics, user experience, and the level of compliance with commercial smartphone health apps among patients with cancer. We analyzed the relationship between application usage logs and survey data to devise an extended TAM model for effective elucidation of the relationships among the variables that contribute to the actual usage of the apps.

Methods

Study Design

This study aimed to provide insights into the factors that affect digital health care app acceptance and sustained use, proposing an enhanced TAM. The survey included various categories such as socioeconomic demographics (D), usual usage of smartphones, health management (HM) habits, user experience (app usability [Us]), interface satisfaction (IS), equity (Eq), utility (Ut), enjoyment (Ej), active willingness to use the app (AW), attitude toward usefulness (AU), information management (IM), preference for human interaction (HI) via app, willingness to continuous use (CU), preference of app feature (PF), and external motivation given by researchers (ie, attending physician of the study subject). The study enrolled 264 patients, part of a randomized controlled trial (RCT) involving 960 cancer survivors (breast, colorectal, or lung cancer), excluding a control group and accounting for a 20% dropout rate, from a pool of patients who had used the app for at least 6 months, aimed at assessing the impact of mobile Health apps on recovery. The description of the features of the 4 apps used in the prospective RCT is provided in Multimedia Appendix 1. And we employed structural equation modeling (SEM) to test hypotheses regarding the influence of motivation, perceived usefulness, and perceived ease of use on users' attitudes and behavioral intentions toward the app.

Questionnaire Design

The development of the survey was based on the TAM and prior research exploring perceived ease-of-use and perceived usefulness, commonly used in studies related to mobile health care, smartphone adoption, and continuous usage intention. The survey categories and items were selected based on previous studies investigating the factors influencing the acceptance of digital health care products and services [26-36]. The set of variables and reference articles relating to each TAM variable, category, and operational definition are shown in Table S1 in Multimedia Appendix 1. The final survey categories were as follows: D, use of smartphone, HM, Us, IS, Eq, Ut, Ej, AW, AU, IM, HI, CU, PF, and motivation (M). The matching of each category with TAM variables is shown in Figure S1 in Multimedia Appendix 1.

The response formats for each survey item were measured according to the nature of the questionnaire. These formats included the 5-point Likert scale (1=strongly disagree to 5=strongly agree), yes or no responses, single-choice selections, multiple-choice selections, and open-ended responses. The survey was created as a web-based individual survey link and sent to respondents on their devices (smartphones or PCs). For a complete description of the survey, see Multimedia Appendix 1.

XSL•FO
RenderX

Enrollment Process

Among the 960 participants involved in a previous protocol study, a control group of 240 individuals who did not use the app was excluded [37-39]. During the recruitment period, patients who presented for medical visits after a minimum of 6 months since the initiation of app usage were approached to obtain research participation consent. The sample size for the research participants was calculated based on the survey sample size formula [40]. With a confidence level of 95% and a margin of error of 5%, considering the population proportions for breast, colorectal, and lung cancer (33%), the appropriate sample size was determined to be 252 out of the total population of 720. Additionally, hypothesis testing was conducted using Cronbach α with a minimum acceptable value of .95, an expected value of .97, a significance level of .05 (2-tailed), a power of 80%, an expected dropout rate of 5%, and a total of 90 survey items. The sample size calculator determined that a sample size of at least 63 respondents was required to achieve valid reliability.

The survey was conducted from May 17, 2022, to August 26, 2022. A dropout rate of 20% was considered during patient recruitment, and a total of 309 patients were available for enrollment during the recruitment period. Among them, 17 (5.5%) patients who refused to participate in the survey were excluded. After receiving the survey link, the participants were requested to complete the survey within the designated research period. Participants who failed to complete the survey within 1 week received a reminder, and those who did not finish it within an additional week were deemed to have dropped out of the study. Out of 292 (94.5%) participants who agreed to participate in the survey, 28 (9.1%) did not complete it during the study period and were excluded. The final number of participants who completed the survey was 264 (85.4%; Figure 1).

Figure 1. Flow diagram summarizing the enrolment process.

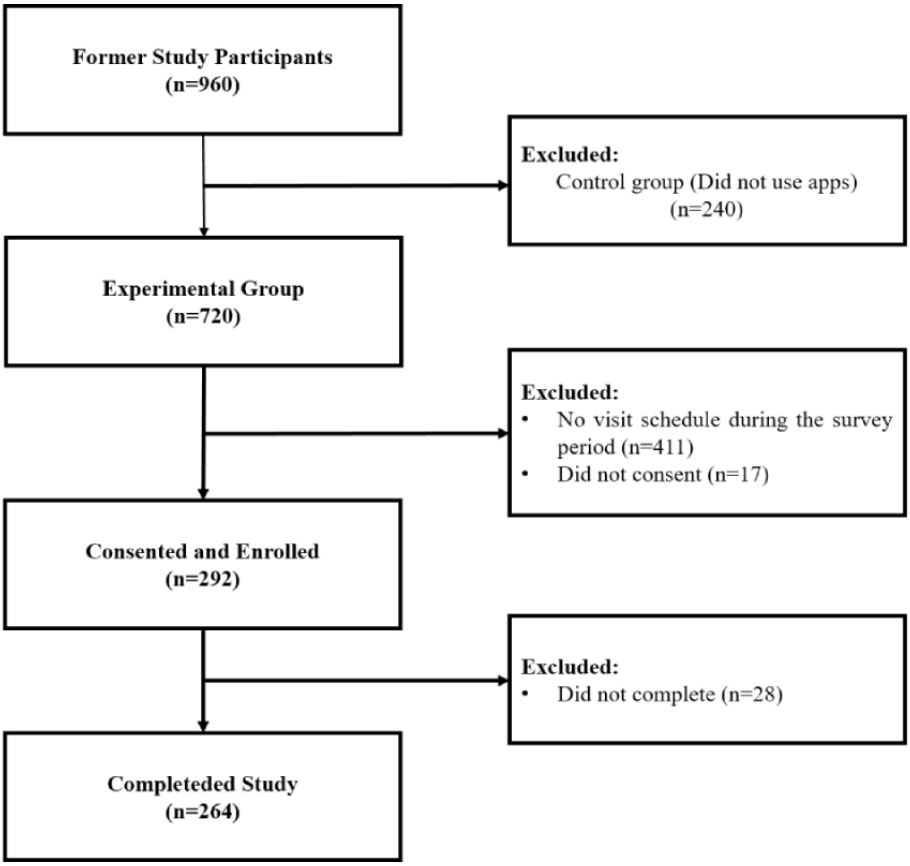

Data Collection

The app used in the study is commercially available and can be downloaded from the Korean App Store. In the case of paid apps, the research funds were used to cover the costs of using the app [37-39]. The app usage data were collected from the app development company, with the consent of the research participants. Information collected included app access timestamps and frequency of access.

Data Analysis

Questionnaire Result Analysis

We used chi-square cross-analysis to calculate the *P* values for survey items with binary responses. For survey items with continuous variable responses, such as Likert scale responses, we used the Wilcoxon rank sum test to calculate the p-values. In both cases, we compared 2 groups at a time rather than analyzing all 3 groups (short use, medium use, and long use) simultaneously. The data analysis was performed using R (version 4.3.0). For all survey items, excluding the responses related to continuous variables and open-ended responses, the responses were coded using numerical values without assigning

XSL•FO
RenderX

any inherent meaning to the magnitude of the numbers. These coded responses were then used for calculations as nominal variables.

For binary survey responses (yes or no), we coded the 264 participants' answers as follows: positive experiences as 1 and negative experiences as 0. We then calculated the ratio of positive responses within each category based on the total number of questions in that category. Participants were grouped as high or low within each category based on their scores compared to the category average. If an individual's score in a specific category exceeded the category average, they were categorized as high; otherwise, they were classified as low. Participants were labeled as positive or high if more than half of their category groups were high; otherwise, they were classified as negative or low.

Actual App Usage Data Analysis

The participants' actual app usage data were obtained using the dates and times when the app was opened for each usage during a period of 180 days. By counting the number of app access days for each patient, 3 components were calculated for K-means clustering: mean number of use, total term usage, and number of app access days [41]. The mean number of use was calculated as an average of app usage days based on the first usage date of the app. Total term usage, measured in days, was the difference between the dates of the first and last app access. The number of app access days refers to the count of days a patient accessed the app within the 180 days starting from their initial access date.

We classified patients into 3 clusters: short use, medium use, and long use [42]. The short-use group comprised patients who only used the app for 12 weeks during the intervention period. The medium-use group comprised patients who used the app for more than 12 weeks but did not use it persistently beyond that period. The long-use group consisted of patients who used the app continuously. The data analysis was performed using Python (version 3.8.5; Python Software Foundation).

Structural Equation Modeling

Descriptive statistics and SEM using latent variable analysis were performed using the Lavaan software (version 0.6-15; Ghent University) in R (version 4.3.0; R Foundation for Statistical Computing). As this study's sample was from 264 patients, the sample size was adequate for model analysis [43]. A *P* value less than .05 was considered statistically significant. In our proposed model, we introduced a new node called "Motivation" to the existing TAM [44] and hypothesized that it could influence (perceived) usefulness, perceived ease-of-use, and behavioral intention. Moreover, perceived usefulness was changed to (perceived) usefulness, as the apps were commercially available and had already demonstrated their Ut. The "Motivation" node represented how the participants felt while participating in the clinical study using the app. The schematic representation of the proposed theoretical model is presented in Figure 2. Moreover, the comparative fit index, Tucker-Lewis index, root mean square error of approximation, and standardized root mean square residual were used to identify the fit of the theoretical model.

Figure 2. Enhanced TAM (technology acceptance model; tailored for a digital health care app).

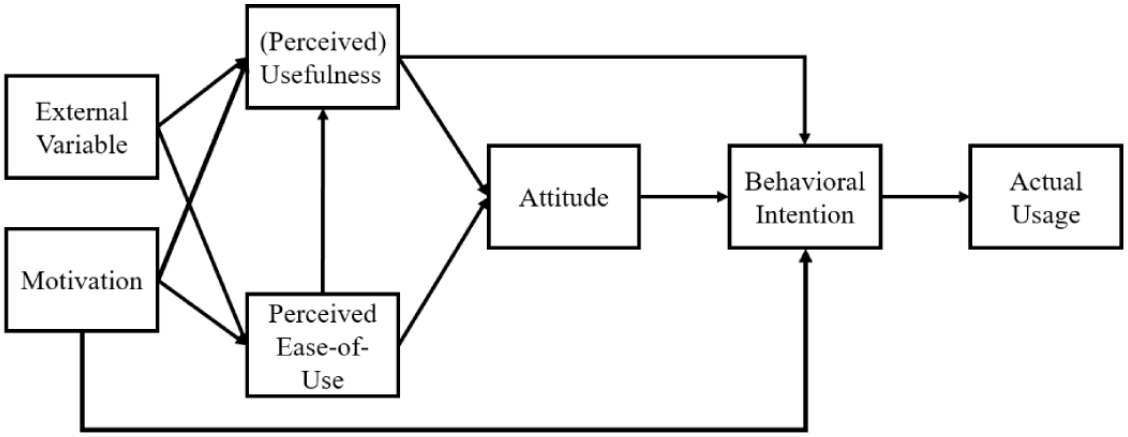

Therefore, through the SEM, we attempted to answer various hypotheses as follows: motivation will positively affect perceived usefulness and perceived ease of use. Moreover, external variables will positively affect perceived usefulness and perceived ease of use. The perceived ease-of-use of application users will positively affect their (perceived) usefulness, their attitude toward the application, and their behavioral intention to continue using the application. The (perceived) usefulness of application users will positively affect their attitude toward the application and their behavioral intention to continue using the application. The attitude of application users toward the application will positively affect their behavioral intention to continue using the application. Finally, the behavioral intention of application users will positively affect their actual usage of the application.

Ethical Considerations

This study was approved by the institutional review board of Asan Medical Center, Korea (IRB 2021-1631) and adhered to relevant ethical guidelines. Informed consent was obtained from all participants before their involvement in the study, with assurances of anonymity and confidentiality provided. Participants were briefed on the study's objectives and how the collected data would be used. Additionally, stringent measures were in place to protect the privacy and confidentiality of study data, including secure storage within the hospital premises. Each

XSL•FO
RenderX

participant received a US $7 gift voucher as compensation for their time.

Results

Categorizing Patients According to Their App Usage Behavior Using K-Means Clustering

We classified the participants into 3 clusters based on their app usage behavior. The clusters were labeled S, M, or L (short use, medium use, and long use, respectively) as shown in Figure 3. There were 77 short-use patients: the average of their mean number of use was 9 (SD 11) days, the mean of total term of use was 21 (SD 21) days, and the mean number of app access days was 7 (SD 14) days. The means for the 101 medium-use patients were 58 (SD 20) days for mean number of use, 144 (SD 36) days for total term of use, and 55 days for number of app access days. Finally, the 86 long-use patients had the following means: 84 days for mean number of use, 176 days for total term of use, and 154 days for number of app access days.

Figure 3. Results of K-means clustering on sample (n=264). Results of K-means clustering for (a) 3 components: number of app-access days (use_days), mean number of use (mean_usage), and total term of use (max_dd); (b) max_dd–mean_usage perspective; (c) max_dd–use_days perspective; and (d) mean_usage–use_days perspective.

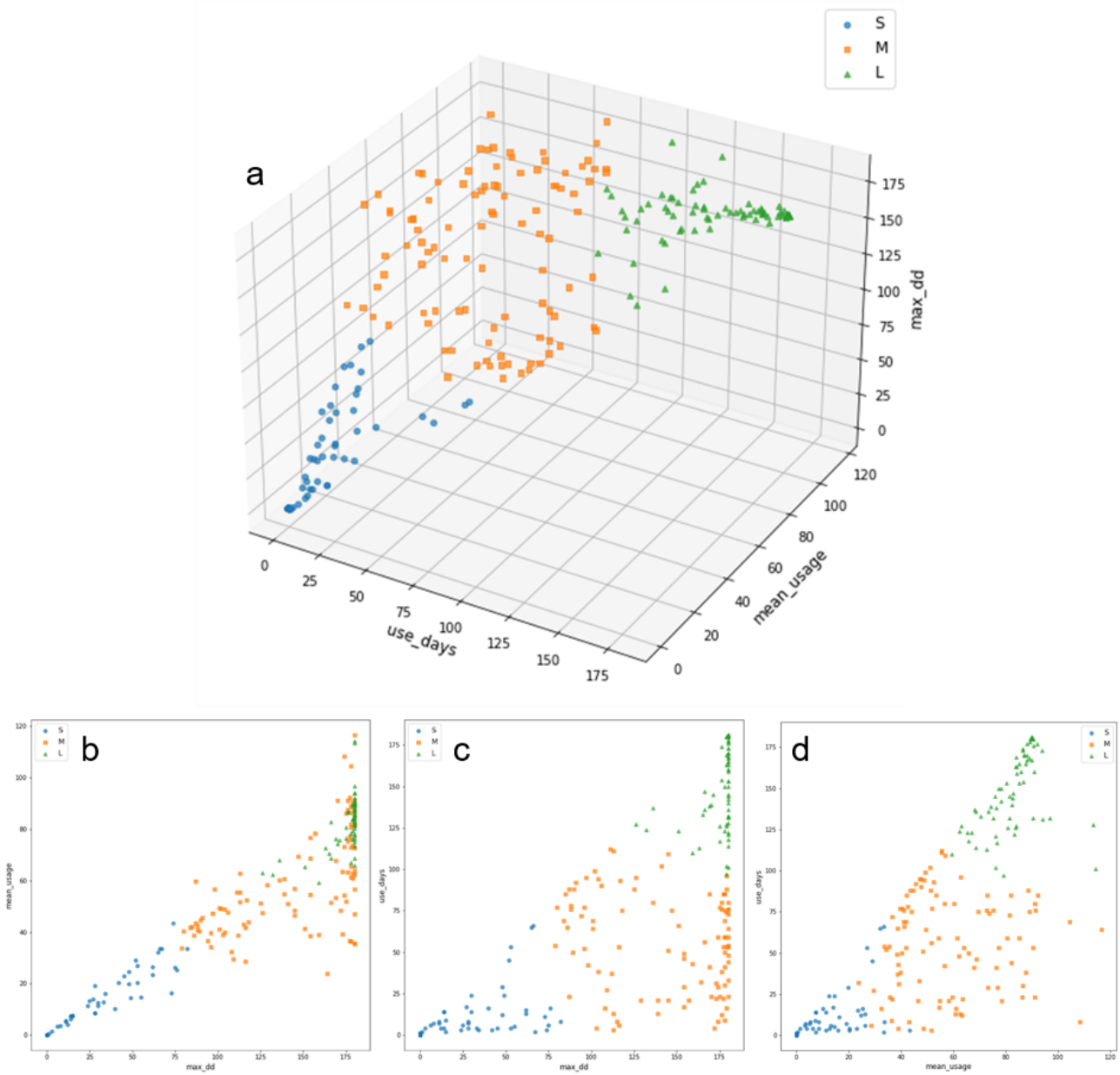

Gender distribution significantly varied across groups (*P*=.02), with a higher proportion of males in the short group (40.3%) compared to the long group (19.8%). The median age ranged from 54 to 57 years across groups (*P*=.08). Participants from urban areas were more prevalent in the long group (60.5%), while those from rural areas were less represented (*P*=.06). Education levels differed significantly across the group (*P*=.04), where a higher proportion of individuals in the short group had a more middle school graduate or lower compared to the medium and long groups. Even though high school graduates were fairly evenly distributed across the groups, university graduates or higher were most prevalent in the medium group.

XSL•FO
RenderX

Marital status did not differ significantly across groups *P*=.27). App usage varied notably (*P*<.01), with higher usage of Noom in the long group (52.3%), while Walkon and Second Doctor were more common in the medium group (46.5% and 28.7%, respectively). Breast cancer was the most common cancer type, especially in the long group (51.2%), with significant differences in cancer type distribution across groups (*P*=.03). Clinical stage distribution showed no significant differences (*P*=.39), though stage 1 was most prevalent in all groups, particularly in the long group (66.3%). The details of the demographics for each group are available at Table 1.

Table 1. Demographics of the participants in respect to their term use.

Variables	Total (N=264)	Short (n=77)	Medium (n=101)	Long (n=86)	*P* value
Gender, n (%)					*.02*[a]
Male	78 (29.5)	31 (40.3)	30 (29.7)	17 (19.8)	
Female	186 (70.5)	46 (59.7)	71 (70.3)	69 (80.2)	
Age (years), median (range)	54.0 (21-79)	57.0 (30-79)	54.0 (21-74)	54.5 (21-77)	.08
Residence, n (%)					.06
Urban	132 (50.0)	35 (45.5)	45 (44.6)	52 (60.5)	
Rural	132 (50.0)	42 (54.5)	56 (55.4)	34 (39.5)	
Education, n (%)					*.04*
Middle school graduate or lower	21 (7.9)	11 (14.3)	5 (5.0)	5 (5.9)	
High school graduate	87 (33.0)	27 (35.1)	28 (27.7)	32 (37.2)	
University graduate or higher	156 (59.1)	39 (50.6)	68 (67.3)	49 (56.9)	
Marital status, n (%)					.27
Married (having a spouse)	221 (83.7)	3 (3.9)	4 (4.0)	10 (11.6)	
Never married	17 (6.4)	64 (83.1)	90 (89.1)	67 (77.9)	
Other	26 (9.8)	10 (13.0)	7 (6.9)	9 (10.5)	
Monthly household income (US $), n (%)					.74
Less than 1150-2300	15 (5.7)	5 (6.5)	5 (5.0)	5 (5.8)	
2300-4600	106 (40.2)	31 (40.3)	44 (43.6)	31 (36.0)	
4600 and above	132 (50.0)	36 (46.8)	49 (48.5)	47 (54.7)	
Does not wish to answer	11 (4.2)	5 (6.5)	3 (3.0)	3 (3.5)	
Used app, n (%)					*<.01*
Walkon	97 (36.7)	17 (22.1)	47 (46.5)	33 (38.4)	
Noom	81 (30.7)	12 (15.6)	24 (23.8)	45 (52.3)	
Second Doctor	59 (22.3)	22 (28.6)	29 (28.7)	8 (9.3)	
Efilcare	27 (10.2)	26 (33.8)	1 (1.0)	0 (0)	
Cancer type, n (%)					*.03*
Breast cancer	114 (43.2)	22 (28.6)	48 (47.5)	44 (51.2)	
Colon cancer	64 (24.2)	25 (32.5)	26 (25.7)	13 (15.1)	
Lung cancer	86 (32.6)	30 (39.0)	27 (26.7)	29 (33.7)	
Clinical stage (p-stage), n (%)					.39
Stage 1	154 (58.3)	43 (55.8)	54 (53.5)	57 (66.3)	
Stage 2	62 (23.5)	21 (27.3)	24 (23.8)	17 (19.8)	
Stage 3	31 (11.7)	10 (13.0)	13 (12.9)	8 (9.3)	
Stage 4	17 (6.4)	3 (3.9)	10 (9.9)	4 (4.7)	

[a]The italicized values represent statistically significant (*P*<.05) comparisons.

XSL•FO
RenderX

Relationship Between App Usage and TAM Factors

Next, we conducted a comparative analysis using the survey categories corresponding with the TAM structure variables. We calculated the average positive response scores for each category and classified individuals into high or low groups based on whether their scores were above or below the mean score, respectively. Furthermore, we compared the proportions of the short-use, medium-use, and long-use cluster groups between the High and Low groups (Table 2). Moreover, the overall survey response groups are available in Table S2 in Multimedia Appendix 1.

Table 2. Distribution of clustering groups by survey category according to the technology acceptance model (TAM) structure.

Technology acceptance model (TAM) structure (elements, questionnaire, and response result groups)	Total (N=264)	Usage cluster category			*P* value
		Short (n=77)	Medium (n=101)	Long (n=86)	
Perceived usefulness, n (%)					
Utility					*.01*[a]
High	196 (74.2)	48 (62.3)	78 (77.2)	70 (81.4)	
Low	68 (25.8)	29 (37.7)	23 (22.8)	16 (18.6)	
Enjoyment					.54
High	120 (45.5)	23 (29.9)	45 (44.6)	52 (60.5)	
Low	144 (54.5)	54 (70.1)	56 (55.4)	34 (39.5)	
Perceived ease-of-use, n (%)					
Interface satisfaction					*<.01*
High	197 (74.6)	45 (58.4)	78 (77.2)	73 (84.9)	
Low	67 (25.4)	32 (41.6)	23 (22.8)	13 (15.1)	
Equity					*<.01*
High	193 (73.1)	44 (57.1)	77 (76.2)	72 (83.7)	
Low	71 (26.9)	33 (42.9)	24 (23.8)	14 (16.3)	
Usability					*.01*
High	183 (69.3)	40 (51.9)	75 (74.3)	68 (79.1)	
Low	81 (30.7)	37 (48.1)	26 (25.7)	18 (20.9)	
Attitude, n (%)					
Active willingness to use the app					.92
High	115 (43.6)	17 (22.1)	46 (45.5)	52 (60.5)	
Low	149 (56.4)	60 (77.9)	55 (54.5)	34 (39.5)	
Attitude toward usefulness					.10
High	189 (71.6)	45 (58.4)	75 (74.3)	69 (80.2)	
Low	75 (28.4)	32 (41.6)	26 (25.7)	17 (19.8)	
Information management					.32
High	168 (63.6)	41 (53.2)	65 (64.4)	62 (72.1)	
Low	96 (36.4)	36 (46.8)	36 (35.6)	24 (27.9)	
Human interaction					*.03*
High	135 (51.1)	34 (44.2)	47 (46.5)	54 (62.8)	
Low	129 (48.9)	43 (55.8)	54 (53.5)	32 (37.2)	

[a]The italicized values represent statistically significant ($P<.05$) comparisons.

The short-use clustering group exhibited higher proportions of low respondents than the other groups in all categories. Furthermore, among the categories corresponding to perceived usefulness and perceived ease-of-use, except for Ej, significant differences were observed among the cluster groups in all categories (Us: $P=.01$; IS: $P<.01$; E: $P<.01$; Ut: $P=.01$). The proportion of low respondents among the overall survey participants was 55% in the Ej category. The *P* values for the

XSL•FO
RenderX

individual group comparisons between the high or low groups and short-use or medium-use- or long-use cluster groups have been provided in Table S3 in Multimedia Appendix 1.

Structural Equation Modeling

For this extended TAM, the fit of data was favorable; comparative fit index: 0.905, Tucker-Lewis index: 0.890, root mean square error of approximation: 0.052, and standardized root mean square residual: 0.071. The final structural model with estimated standardized coefficients is presented in Figure 4, and the estimation results of the hypotheses are shown in Table S4 in Multimedia Appendix 1.

Figure 4. Results from structural equation modeling show the relevance among TAM (technology acceptance model) variables. Standardized estimates of the relationships between variables are indicated by solid lines. Significant relationships are denoted by *(*P*<.05), **(*P*<.01), and ***(*P*<.001), while broken lines represent non-significant relationships. The solid lines represent observed significant relationships, confirming the hypothesized connections.

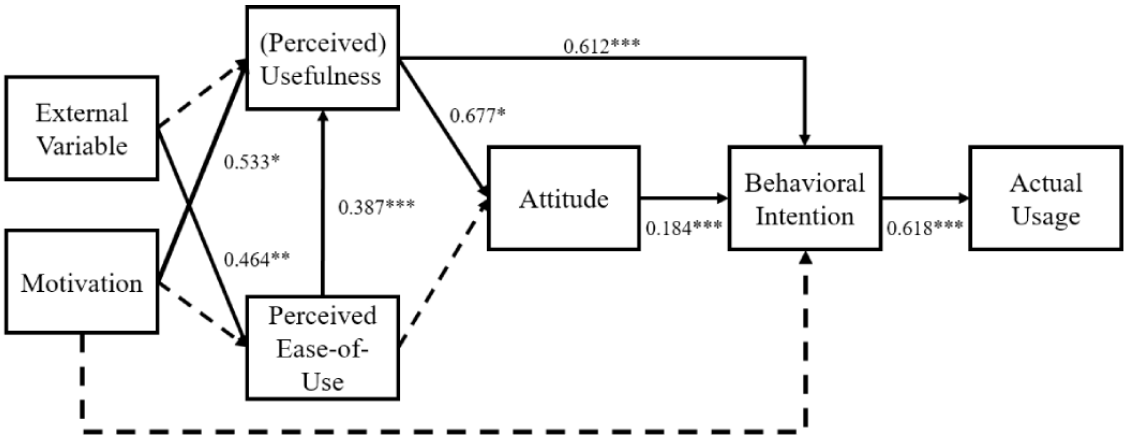

The path from external variables to (perceived) usefulness was not statistically significant (β=–0.061, *P*=.272). However, there was a significant and positive path from external variables to perceived ease-of-use (β=0.464, *P*=.004). While there was a significant path from motivation to (perceived) usefulness (β=0.533, *P*=.003), the paths from motivation to perceived ease-of-use (β=0.293, *P*=.401) and behavioral intention (β=–0.197, *P*=.201) were not statistically significant. Perceived ease-of-use significantly influenced (perceived) usefulness (β=0.387, *P*<.01). Perceived usefulness, in turn, had a significant positive effect on attitude (β=.677, *P*=.02). Furthermore, (perceived) usefulness had a significant positive effect on behavioral intention (β=0.612, *P*<.01). Attitude significantly influenced behavioral intention (β=0.184, *P*<.01). Finally, behavioral intention had a significant positive effect on actual usage (β=0.618, *P*<.01).

Discussion

Principal Findings

We analyzed questionnaire responses and app usage log data of patients with cancer who used a commercial smartphone health care app as an intervention in a prospective RCT, stratified users by whether they used the app faithfully or not, and identified factors affecting app usage. When we grouped patients using app usage logs, we considered not only total usage duration but also amount and access frequency, and the group (S) who used the app the least actively showed a maximum usage duration close to 12 weeks, the original RCT intervention period. Because the researchers did not discourage or encourage use after the 12-week intervention period, the above results suggest that the study setting had a meaningful effect on the continuation of app usage among the study subjects. Though we classified patients into 3 groups, “short use, medium use, and long use,” this is meaningful because the groups do not simply refer to the length of the maximum usage period but also to whether they used the app actively and loyally (Table S2 in Multimedia Appendix 1).

The analysis of the study population revealed notable differences among participants grouped by the duration of health intervention app usage. Gender distribution was significantly different across groups, where females may be more inclined to engage in prolonged use of health-related apps, potentially reflecting greater health-seeking behavior or adherence to digital interventions among women [45]. Residence patterns showed that participants from big cities were more likely to be in the long group, possibly due to better access to digital resources or a higher level of digital literacy in urban areas. App usage patterns varied considerably, with Noom being most popular in the long-duration group, while Walkon and Second Doctor were more common among those with medium-duration usage. This variation in app preference may be related to the specific features, usability, or perceived effectiveness of the apps, influencing users' commitment over time. Cancer type distribution also differed significantly, with patients with breast cancer more likely to be in the long-duration group [46].

In the category with TAM variables, significant differences among the cluster groups were observed in categories related to perceived usefulness and perceived ease of use, except for Ej. SEM results revealed that external variables significantly influenced perceived ease of use but not perceived usefulness. Motivation significantly influenced perceived usefulness but did not significantly affect perceived ease of use or behavioral intention. Perceived ease of use significantly influenced perceived usefulness, which in turn significantly affected attitude and behavioral intention. Attitude significantly influenced behavioral intention, which then had a significant positive effect on actual usage.

XSL•FO
RenderX

Interpretation Within the Context of Wider Literature

Despite continuing skepticism about commercial smartphone applications' effectiveness for HM, there is no doubt that smartphone apps have the potential to drive improved health through behavioral change by always being beside users [47]. Through this mixed methods study, we determined that the factors influencing the consistent use of apps are not solely dependent on user characteristics, such as age, gender, and typical smartphone usage experience [48]. A physician's recommendation, validated by an assessment tool and model, can exert a notable influence as external motivation. Also, daily smartphone use, although unmodifiable, aids in detecting non-compliance within high-risk groups using digital health tools [49]. Perceptions of ease-of-use and usefulness significantly influence app usage, highlighting the crucial role of UI design and app content in digital HM. The study emphasizes the importance of doctors understanding digital health apps for making effective recommendations. Proficiency in smartphone use correlates with high perceived ease-of-use, consistent with structural equation results. Further research is needed to determine if familiarity with smartphone interaction increases positive responses to notifications or messages.

Participant demographics such as age, gender, residence, education, and income did not significantly correlate with app usage experience. However, being female, below the median age, and having an income exceeding US $4600 were key characteristics of participants inclined to use apps, aligning with studies on app retention [50]. While smartphone ownership has reduced digital inequality, our findings suggest that factors associated with complex UI vulnerability still influence app usage.

Distinct differences were observed among the 3 groups regarding usage duration across all 3 criteria. Analyzing TAM factors in these groups revealed notable differences in IS and Eq. Us, Ut, AU, and HI also varied among the groups. Notably, IS and Eq exhibited the most significant differences, with higher percentages linked to increased app usage. This indicates that low Us stemmed from the challenging user experience while using the app. Previous studies have also indicated that a user-friendly interface offering simplicity and automation encourages the use of health care apps, particularly among older patients [51-54]. Our findings align with these studies, as patients who reported enjoying the app’s interactions and user-friendly interface were more prevalent in the group with higher app usage.

Overall, the significant connections positively affected the other variables. Except for the connections from external variables to (perceived) usefulness, motivation to perceived ease-of-use, and motivation to behavioral intention, the remaining connections from the improved TAM exhibited the following significant and positive paths: external variables to perceived ease-of-use, motivation to (perceived) usefulness, perceived ease-of-use to (perceived) usefulness, (perceived) usefulness to attitude, attitude to behavioral intention, and behavioral intention to actual usage. The presented findings align with existing research on mobile services and portable electronic devices. Kuo and Yen [55] investigated 3rd generation mobile value-added services and found a positive and significant relationship between perceived usefulness, behavioral intention, and attitude. Similarly, Chen and Chen [56] demonstrated a significant link between perceived ease-of-use and attitude among travelers using global positioning system devices.

We can better understand how different aspects correlate with the app's actual usage by closely examining the extended TAM at a more granular level. In terms of demographic data, the extended TAM revealed that female patients exhibited higher levels of app engagement compared to males. Furthermore, younger patients displayed more frequent and proficient app usage. This trend aligns with previous research findings, which have consistently highlighted the greater comfort and familiarity that female and younger individuals have with mobile apps, leading to prolonged usage [57,58]. This might be attributed to the fact that women tend to use more health apps related to fertility and pregnancy, indicating their openness to other health care apps as well [58]. Additionally, patients with higher education and income levels exhibited elevated app usage rates. Furthermore, individuals who reported more extensive smartphone usage in their current or previous professional roles were more inclined to use the app extensively. This association can be attributed to the increased comfort and familiarity that smartphone-familiar individuals have with mobile technology. Such users are more predisposed to engage with the app, as they already integrate smartphones more prominently into their daily routines than those without a history of smartphone use in their professional roles.

Digital interventions for clinical research are primarily designed for research purposes, with acknowledged limitations when using commercialized smartphone apps [59]. While some digital therapeutics and HM apps demonstrate effects proven through clinical studies, there is a lack of consistent indicators for app use compliance, even in research apps. Results may differ between clinical studies and real-world scenarios, as participants in clinical studies might not actively choose their application. Recognizing patients more as consumers, clinician recommendations play a crucial role in app retention, boosting confidence and adherence [60]. This underscores the potential of clinician recommendations to enhance adherence to digital health solutions.

Compared to other chronic diseases, the relationship between the patient with cancer and oncologist (surgical oncologist) is especially important in the dynamic treatment process and decision-making [61,62]. Previous research has attempted to define the patient with cancer–physician relationship in various ways [63]. Palmer Kelly et al [64] explained the patient with cancer–physician relationship with attachment model, and an appropriate attachment relationship between patients with cancer and physicians is associated with a better quality of life. Our findings support the idea that the patient-physician relationship can influence app acceptance and serve as a potential factor for increasing compliance with digital interventions. Physical activity is encouraged to cope with anxiety and depression in patients with cancer, and outdoor physical activity in patients with cancer can be promoted by organized activities [65,66]. Therefore, we can expect that if the use of digital interventions

XSL•FO
RenderX

is encouraged by his or her oncologist (surgical oncologist), the effect on the quality of life of patients could be further improved.

Strengths and Limitations

To our knowledge, this is the first study that integrated user experience evaluation in commercial health care apps for cancer survivors, examining their adherence through usage log data analysis. Moreover, this is the first study to use SEM to extend the TAM to be well suited in a mobile health context. The integration of these 2 approaches enables a comprehensive understanding of the factors influencing the CU intention of digital health care applications among patients with cancer, including the effects of health status and demographics. Our study provides insights into the key determinants of mobile health care app usage among patients with cancer, emphasizing the role of user experience, clinician recommendations, and the potential to bridge digital health care inequalities.

The traditional TAM faces limitations in capturing recent technological environment changes and fails to reflect the diverse information acquisition paths for consumers [67,68]. The classical TAM is criticized for not accounting for users' voluntary choice of technology, particularly in the context of HM apps where users may not opt for the technology voluntarily [69]. In addition, since this study targeted only those who could use a smartphone, it was difficult to determine how noncompliance with digital intervention is related to digital inequality [70]. All the participants in this study were from Korea, which could limit its generalizability. However, despite the homogeneity of race and culture in Korea compared to countries such as the United States, the characteristics of the patients would not pose a significant generalizability problem. The age of breast cancer onset varies by ethnicity, with Asian women having a younger age of breast cancer onset. In addition, breast cancer has a different peak onset age than colon and lung cancer, which may bias the results of this study [71,72]. However, this difference is not substantial enough to affect the study's overall applicability.

Conclusions

This study highlights the critical factors influencing cancer survivors' adherence to commercial health care apps, emphasizing the pivotal role of external motivation, particularly physician recommendations. Key determinants include perceived ease-of-use and usefulness, highlighting the significance of effective UI design. SEM reveals positive paths from motivation to perceived usefulness and from perceived usefulness to attitude and behavioral intention, impacting actual app usage. These findings stress the importance of positive user experience and clinician recommendations in facilitating the effective usage of digital health care tools among cancer survivors and contributing to the evolving landscape of medical care.

Acknowledgments

This research was supported by a grant (RS-2020-KH088470) from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) and funded by the Ministry of Health & Welfare, Republic of Korea.

Data Availability

The anonymously processed dataset used in this study can be shared by the corresponding author upon reasonable request. Any interested party should submit an appropriate proposal, analysis plan, and protocols for review and approval by the authors. Subsequently, the materials will be submitted for approval to the institutional review board of Asan Medical Center (Seoul, Republic of Korea). Once approved, the data will be made available.

Authors' Contributions

Formal analysis, investigation, software, visualization, and writing of the original draft were performed by YP and YT. Data curation and methodology were handled by IK and HL. Supervision and validation were conducted by JBL. Supervision was also provided by JWL. Conceptualization, methodology, project administration, and writing of the original draft were conducted by YL. All authors had full access to the data and had final responsibility for the decision to submit for publication. YP, YT, and YL directly accessed and verified the data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables and figures.
[DOCX File , 552 KB - jmir_v26i1e55176_app1.docx]

References

1. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]

XSL•FO
RenderX

2. Smart Healthcare Medical Device Technology/Standard Strategy Report. Seoul: Ministry of Food and Drug Safety (KR); 2018.
3. Slim K, Selvy M, Veziant J. Conceptual innovation: 4P medicine and 4P surgery. J Visc Surg 2021;158(3S):S12-S17. [doi: 10.1016/j.jviscsurg.2021.01.003] [Medline: 33714709]
4. Clauser SB, Wagner EH, Aiello Bowles EJ, Tuzzio L, Greene SM. Improving modern cancer care through information technology. Am J Prev Med 2011;40(5 Suppl 2):S198-S207 [FREE Full text] [doi: 10.1016/j.amepre.2011.01.014] [Medline: 21521595]
5. Hopstaken JS, Verweij L, van Laarhoven CJHM, Blijlevens NMA, Stommel MWJ, Hermens RPMG. Effect of digital care platforms on quality of care for oncological patients and barriers and facilitators for their implementation: systematic review. J Med Internet Res 2021;23(9):e28869 [FREE Full text] [doi: 10.2196/28869] [Medline: 34559057]
6. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, et al. Cancer treatment and survivorship statistics, 2022. CA Cancer J Clin 2022;72(5):409-436 [FREE Full text] [doi: 10.3322/caac.21731] [Medline: 35736631]
7. Penedo FJ, Oswald LB, Kronenfeld JP, Garcia SF, Cella D, Yanez B. The increasing value of eHealth in the delivery of patient-centred cancer care. Lancet Oncol 2020;21(5):e240-e251 [FREE Full text] [doi: 10.1016/S1470-2045(20)30021-8] [Medline: 32359500]
8. Roberts AL, Fisher A, Smith L, Heinrich M, Potts HWW. Digital health behaviour change interventions targeting physical activity and diet in cancer survivors: a systematic review and meta-analysis. J Cancer Surviv 2017;11(6):704-719 [FREE Full text] [doi: 10.1007/s11764-017-0632-1] [Medline: 28779220]
9. Age and Cancer Risk. US: National Cancer Institute; 2021. URL: https://www.cancer.gov/about-cancer/causes-prevention/risk/age.[accessed [accessed 2022-11-09]
10. Tonorezos E, Devasia T, Mariotto AB, Mollica MA, Gallicchio L, Green P, et al. Prevalence of cancer survivors in the United States. J Natl Cancer Inst 2024;116(11):1784-1790. [doi: 10.1093/jnci/djae135] [Medline: 39002121]
11. Laricchia F. Share of adults in the United States who owned a smartphone from 2015 to 2023, by age group. Statista. 2024. URL: https://www.statista.com/statistics/489255/percentage-of-us-smartphone-owners-by-age-group/ [accessed 2024-03-22]
12. Petrosyan A. Distribution of internet users worldwide as of February 2024, by age group. Statista. 2024. URL: https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/ [accessed 2024-06-27]
13. Boels AM, Rutten G, Zuithoff N, de Wit A, Vos R. Effectiveness of diabetes self-management education via a smartphone application in insulin treated type 2 diabetes patients - design of a randomised controlled trial ('TRIGGER study'). BMC Endocr Disord 2018;18(1):74 [FREE Full text] [doi: 10.1186/s12902-018-0304-9] [Medline: 30348142]
14. Madrigal L, Escoffery C. Electronic health behaviors among US adults with chronic disease: cross-sectional survey. J Med Internet Res 2019;21(3):e11240 [FREE Full text] [doi: 10.2196/11240] [Medline: 30835242]
15. Onyeaka HK, Zambrano J, Longley RM, Celano CM, Naslund JA, Amonoo HL. Use of digital health tools for health promotion in cancer survivors. Psychooncology 2021;30(8):1302-1310 [FREE Full text] [doi: 10.1002/pon.5677] [Medline: 33742737]
16. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
17. Global Diffusion of eHealth: Making Universal Health Coverage Achievable: Report of the third Global Survey on eHealth. Geneva: World Health Organization; 2016.
18. Clement I, Lorenz A, Ulm B, Plidschun A, Huber S. Implementing systematically collected user feedback to increase user retention in a mobile app for self-management of low back pain: retrospective cohort study. JMIR Mhealth Uhealth 2018;6(6):e10422. [doi: 10.2196/10422] [Medline: 29875088]
19. Coughlin JW, Martin LM, Zhao D, Goheer A, Woolf TB, Holzhauer K, et al. Electronic health record-based recruitment and retention and mobile health app usage: multisite cohort study. J Med Internet Res 2022;24(6):e34191 [FREE Full text] [doi: 10.2196/34191] [Medline: 35687400]
20. Davis FD. User Acceptance of Information Systems: The Technology Acceptance Model (TAM). 1987. URL: https://deepblue.lib.umich.edu/handle/2027.42/35547 [accessed 2024-11-13]
21. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319. [doi: 10.2307/249008]
22. Fernandes LG, Devan H, Fioratti I, Kamper SJ, Williams CM, Saragiotto BT. At my own pace, space, and place: a systematic review of qualitative studies of enablers and barriers to telehealth interventions for people with chronic pain. Pain 2022 Feb 01;163(2):e165-e181. [doi: 10.1097/j.pain.0000000000002364] [Medline: 34433776]
23. Shabir H, D'Costa M, Mohiaddin Z, Moti Z, Rashid H, Sadowska D, et al. The barriers and facilitators to the use of lifestyle apps: a systematic review of qualitative studies. Eur J Investig Health Psychol Educ 2022;12(2):144-165 [FREE Full text] [doi: 10.3390/ejihpe12020012] [Medline: 35200235]
24. Eliasen A, Abildtoft MK, Krogh NS, Rechnitzer C, Brok JS, Mathiasen R, et al. Smartphone app to self-monitor nausea during pediatric chemotherapy treatment: user-centered design process. JMIR Mhealth Uhealth 2020;8(7):e18564 [FREE Full text] [doi: 10.2196/18564] [Medline: 32706744]

XSL•FO
RenderX

25. Srivastava R. Role of smartphone devices in precision oncology. J Cancer Res Clin Oncol 2023;149(1):393-400. [doi: 10.1007/s00432-022-04413-3] [Medline: 36253632]
26. Cimperman M, Makovec Brenčič M, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an extended UTAUT model. Int J Med Inform 2016;90:22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
27. Jang HJ, Noh GY. The effect of health consciousness and playfulness on intention to use tangible fitness game: extended TAM. J Korea Contents Assoc 2017;17(1):1-11. [doi: 10.5392/jkca.2017.17.01.001]
28. Park D, Choi J, Kim D. The influence of health apps efficacy, satisfaction and continued use intention on wearable device adoption: a convergence perspective. J Digit Converg 2015;13(7):137-145. [doi: 10.14400/jdc.2015.13.7.137]
29. Ki YS, Ahn SM, Cho MG, Choi B. An analysis on affecting factors of healthcare applications continuous usage intention and their relationships. J Soc e-Bus Stud 2019;24:49-89. [doi: 10.7838/jsebs.2019.24.1.049]
30. Kim SS, Ryu S. Structural relationships among factors to adoption of telehealth service. Asia Pac J Inf Syst 2011;21(3):71-96.
31. Kim YW, Han S, Kim KS. Determinants of intention to use digital healthcare service of middle and older users. Inf Soc Media 2018;19(3):1-23. [doi: 10.52558/ism.2018.12.19.3.1]
32. Lee M, Lee H, Kim Y, Kim J, Cho M, Jang J, et al. Mobile app-based health promotion programs: a systematic review of the literature. Int J Environ Res Public Health 2018;15(12):2838 [FREE Full text] [doi: 10.3390/ijerph15122838] [Medline: 30551555]
33. Gao Y, Li H, Luo Y. An empirical study of wearable technology acceptance in healthcare. Ind Manage Data Syst 2015;115(9):1704-1723. [doi: 10.1108/IMDS-03-2015-0087]
34. Baek M, Choi H, Lee H. Age-specific acceptance intention over wearable smart healthcare device. Korean Acad Assoc Bus Admin 2015;28(12):3171-3189. [doi: 10.18032/kaaba.2015.28.12.3171]
35. Lee OH, Ham SW. A study on influence factors of mobile healthcare service using structural equation modeling. J Korea Acad Ind Cooperation Soc 2017;18(3):418-427. [doi: 10.5762/KAIS.2017.18.3.418 ISSN 1975-4701]
36. Wilson EV, Lankton NK. Modeling patients' acceptance of provider-delivered e-health. J Am Med Inform Assoc 2004;11(4):241-248 [FREE Full text] [doi: 10.1197/jamia.M1475] [Medline: 15064290]
37. Baek SY, Lee SB, Lee Y, Chung S, Choi CM, Lee HJ, et al. Effects of mobile healthcare applications on the lifestyle of patients with breast cancer: a protocol for a randomized clinical trial. J Breast Cancer 2022;25(5):425-435 [FREE Full text] [doi: 10.4048/jbc.2022.25.e42] [Medline: 36314766]
38. Kim YI, Park IJ, Kim CW, Yoon YS, Lim SB, Yu CS, et al. Lifestyle interventions after colorectal cancer surgery using a mobile digital device: a study protocol for a randomized controlled trial. Medicine (Baltimore) 2022;101(41):e31264 [FREE Full text] [doi: 10.1097/MD.0000000000031264] [Medline: 36254015]
39. Lee JH, Jeong JH, Ji W, Lee HJ, Lee Y, Jo MW, et al. Comparative effectiveness of smartphone healthcare applications for improving quality of life in lung cancer patients: study protocol. BMC Pulm Med 2022;22(1):175 [FREE Full text] [doi: 10.1186/s12890-022-01970-8] [Medline: 35501757]
40. SurveyMonkey. Calculating the Number of Respondents You Need. SurveyMonkey. 2023. URL: https://tinyurl.com/2ms32rm8 [accessed 2023-05-22]
41. Kim J, Lim S, Min YH, Shin Y, Lee B, Sohn G, et al. Depression screening using daily mental-health ratings from a smartphone application for breast cancer patients. J Med Internet Res 2016;18(8):e216 [FREE Full text] [doi: 10.2196/jmir.5598] [Medline: 27492880]
42. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol 2006;59(Pt 1):1-34. [doi: 10.1348/000711005X48266] [Medline: 16709277]
43. Little TD. Longitudinal Structural Equation Modeling. New York: The Guilford Press; 2013.
44. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
45. Fidan E, Çelik S. Factors affecting medical healthcare-seeking behaviours of female patients according to their stage of being diagnosed with breast cancer. Eur J Cancer Care (Engl) 2021;30(5):e13436. [doi: 10.1111/ecc.13436] [Medline: 33694269]
46. Kong LX, Zhao YH, Feng ZL, Liu TT. Personalized and continuous care intervention affects rehabilitation, living quality, and negative emotions of patients with breast cancer. World J Psychiatry 2024;14(6):876-883 [FREE Full text] [doi: 10.5498/wjp.v14.i6.876] [Medline: 38984338]
47. Apolinário-Hagen J, Menzel M, Hennemann S, Salewski C. Acceptance of mobile health apps for disease management among people with multiple sclerosis: web-based survey study. JMIR Form Res 2018;2(2):e11977 [FREE Full text] [doi: 10.2196/11977] [Medline: 30684408]
48. Nunes A, Limpo T, Castro SL. Acceptance of mobile health applications: examining key determinants and moderators. Front Psychol 2019;10:2791 [FREE Full text] [doi: 10.3389/fpsyg.2019.02791] [Medline: 31920836]
49. Rosenlund M, Kinnunen UM, Saranto K. The use of digital health services among patients and citizens living at home: scoping review. J Med Internet Res 2023;25:e44711 [FREE Full text] [doi: 10.2196/44711] [Medline: 36972122]
50. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the national cancer institute's health information national trends survey data. J Med Internet Res 2020;22(5):e16299 [FREE Full text] [doi: 10.2196/16299] [Medline: 32406865]

XSL•FO
RenderX

51. van den Berg N, Schumann M, Kraft K, Hoffmann W. Telemedicine and telecare for older patients--a systematic review. Maturitas 2012;73(2):94-114. [doi: 10.1016/j.maturitas.2012.06.010] [Medline: 22809497]
52. Arnhold M, Quade M, Kirch W. Mobile applications for diabetics: a systematic review and expert-based usability evaluation considering the special requirements of diabetes patients age 50 years or older. J Med Internet Res 2014;16(4):e104 [FREE Full text] [doi: 10.2196/jmir.2968] [Medline: 24718852]
53. Liew MS, Zhang J, See J, Ong YL. Usability challenges for health and wellness mobile apps: mixed-methods study among mHealth experts and consumers. JMIR Mhealth Uhealth 2019;7(1):e12160 [FREE Full text] [doi: 10.2196/12160] [Medline: 30698528]
54. Foster M, Xiong W, Quintiliani L, Hartmann CW, Gaehde S. Preferences of older adult veterans with heart failure for engaging with mobile health technology to support self-care: qualitative interview study among patients with heart failure and content analysis. JMIR Form Res 2022;6(12):e41317 [FREE Full text] [doi: 10.2196/41317] [Medline: 36538348]
55. Kuo YF, Yen SN. Towards an understanding of the behavioral intention to use 3G mobile value-added services. Comput Hum Behav 2009;25(1):103-110. [doi: 10.1016/j.chb.2008.07.007]
56. Chen CF, Chen PC. Applying the TAM to travelers' usage intentions of GPS devices. Expert Syst App 2011;38(5):6217-6221. [doi: 10.1016/j.eswa.2010.11.047]
57. Chowdhury SZ, Stevens S, Wu C, Woodward C, Andrews T, Ashall-Payne L, et al. An age-old problem or an old-age problem? A UK survey of attitudes, historical use and recommendations by healthcare professionals to use healthcare apps. BMC Geriatr 2023;23(1):110 [FREE Full text] [doi: 10.1186/s12877-023-03772-x] [Medline: 36823564]
58. Paradis S, Roussel J, Bosson JL, Kern JB. Use of smartphone health apps among patients aged 18 to 69 years in primary care: population-based cross-sectional survey. JMIR Form Res 2022;6(6):e34882 [FREE Full text] [doi: 10.2196/34882] [Medline: 35708744]
59. Sezgin E. Can we use commercial mobile apps instead of research mobile apps in healthcare research? Front Public Health 2021;9:685439 [FREE Full text] [doi: 10.3389/fpubh.2021.685439] [Medline: 34368058]
60. Kang SH, Baek H, Cho J, Kim S, Hwang H, Lee W, et al. Management of cardiovascular disease using an mHealth tool: a randomized clinical trial. NPJ Digit Med 2021;4(1):165 [FREE Full text] [doi: 10.1038/s41746-021-00535-z] [Medline: 34862449]
61. Leighl NB, Butow PN, Tattersall MHN. Treatment decision aids in advanced cancer: when the goal is not cure and the answer is not clear. J Clin Oncol 2004;22(9):1759-1762. [doi: 10.1200/JCO.2004.02.166] [Medline: 15118002]
62. Kim Y, Winner M, Page A, Tisnado DM, Martinez KA, Buettner S, et al. Patient perceptions regarding the likelihood of cure after surgical resection of lung and colorectal cancer. Cancer 2015;121(20):3564-3573 [FREE Full text] [doi: 10.1002/cncr.29530] [Medline: 26094729]
63. Palmer Kelly E, Agne JL, Hyer M, Meara A, Olsen G, Pawlik TM. A systematic review of the methods utilised to measure the relationship between cancer patients and oncologists: implications for future research and practice. Eur J Cancer Care (Engl) 2019;28(3):e12981. [doi: 10.1111/ecc.12981] [Medline: 30561074]
64. Palmer Kelly E, Tsilimigras DI, Hyer JM, Pawlik TM. Understanding the use of attachment theory applied to the patient-provider relationship in cancer care: recommendations for future research and clinical practice. Surg Oncol 2019;31:101-110. [doi: 10.1016/j.suronc.2019.10.007] [Medline: 31622916]
65. Andersen BL, Lacchetti C, Ashing K, Berek JS, Berman BS, Bolte S, et al. Management of anxiety and depression in adult survivors of cancer: ASCO guideline update. J Clin Oncol 2023;41(18):3426-3453. [doi: 10.1200/JCO.23.00293] [Medline: 37075262]
66. Lesser IA, Nienhuis CP, Belanger L. Active by nature: exploring cancer survivors' exercise barriers, facilitators, preferences, and psychosocial benefits of engaging in outdoor physical activity. Support Care Cancer 2021;29(7):4095-4103 [FREE Full text] [doi: 10.1007/s00520-020-05938-4] [Medline: 33404811]
67. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020;22(7):e17256 [FREE Full text] [doi: 10.2196/17256] [Medline: 32628122]
68. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018;9(3):604-634 [FREE Full text] [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
69. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
70. Cao L, Chongsuvivatwong V, McNeil EB. The sociodemographic digital divide in mobile health app use among clients at outpatient departments in inner Mongolia, China: cross-sectional survey study. JMIR Hum Factors 2022;9(2):e36962 [FREE Full text] [doi: 10.2196/36962] [Medline: 35587367]
71. Demicheli R, Retsky MW, Hrushesky WJM, Baum M, Gukas ID, Jatoi I. Racial disparities in breast cancer outcome: insights into host-tumor interactions. Cancer 2007;110(9):1880-1888 [FREE Full text] [doi: 10.1002/cncr.22998] [Medline: 17876835]
72. Stapleton SM, Oseni TO, Bababekov YJ, Hung YC, Chang DC. Race/ethnicity and age distribution of breast cancer diagnosis in the United States. JAMA Surg 2018;153(6):594-595 [FREE Full text] [doi: 10.1001/jamasurg.2018.0035] [Medline: 29516087]

XSL•FO
RenderX

Abbreviations

AU: attitude toward usefulness
AW: active willingness to use the app
CU: continuous use
D: demographics
Ej: enjoyment
Eq: equity
HI: human interaction
HM: health management
IM: information management
IS: interface satisfaction
M: motivation
PF: preference of app feature
RCT: randomized controlled trial
SEM: structural equation modeling
TAM: technology acceptance model
UI: user interface
Us: usability
Ut: utility

Edited by T de Azevedo Cardoso; submitted 04.12.23; peer-reviewed by T Park, TAR Sure, CW Zeng, V Girish, Z Yu; comments to author 21.06.24; revised version received 30.08.24; accepted 28.10.24; published 18.12.24.

Please cite as:
Park YE, Tak YW, Kim I, Lee HJ, Lee JB, Lee JW, Lee Y
User Experience and Extended Technology Acceptance Model in Commercial Health Care App Usage Among Patients With Cancer: Mixed Methods Study
J Med Internet Res 2024;26:e55176
URL: https://www.jmir.org/2024/1/e55176
doi:10.2196/55176
PMID:

©Ye-Eun Park, Yae Won Tak, Inhye Kim, Hui Jeong Lee, Jung Bok Lee, Jong Won Lee, Yura Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Methodological Challenges in Randomized Controlled Trials of mHealth Interventions: Cross-Sectional Survey Study and Consensus-Based Recommendations

Jesus Lopez-Alcalde[1,2,3], MPH, MD; L Susan Wieland[1,4], MPH, PhD; Yuqian Yan[1], PhD; Jürgen Barth[1], PhD; Mohammad Reza Khami[5,6], DDS, PhD; Siddharudha Shivalli[7], MD; Cynthia Lokker[8], MSc, PhD; Harleen Kaur Rai[9], PhD; Paul Macharia[10,11], MSc, PhD; Sergi Yun[12,13,14,15], MD; Elvira Lang[16], MD, PhD; Agnes Bwanika Naggirinya[17], MBChB, MSc, Mmed; Concepción Campos-Asensio[18], BPharm, MLiS; Leila Ahmadian[19], PhD; Claudia M Witt[20,21], MBA, MD

[1]Institute for Complementary and Integrative Medicine, University Hospital Zurich, Zurich, Switzerland

[2]Faculty of Medicine, Universidad Francisco de Vitoria, Madrid, Spain

[3]Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Unidad de Bioestadística Clínica, Hospital Universitario Ramón y Cajal, Center for Biomedical Research in Epidemiology and Public Health Network (CIBERESP), Madrid, Spain

[4]School of Medicine and Health Sciences, George Washington University, Washington, DC, United States

[5]Research Center for Caries Prevention, Dentistry Research Institute, Tehran University of Medical Sciences, Tehran, Iran

[6]Community Oral Health Department, School of Dentistry, Tehran University of Medical Sciences, Tehran, Iran

[7]Department of Medical Statistics, Faculty of Epidemiology and Public Health, London School of Hygiene & Tropical Medicine, London, United Kingdom

[8]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

[9]Digital Health and Wellness Research Group, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom

[10]Department of Research and Programmes, Kenyatta National Hospital, Nairobi, Kenya

[11]University of Nairobi, Faculty of Health Sciences, Nairobi, Kenya

[12]Bio-Heart Cardiovascular Diseases Research Group, Bellvitge Biomedical Research Institute, L'Hospitalet de Llobregat, Barcelona, Spain

[13]Community Heart Failure Program, Cardiology Department, Bellvitge University Hospital, L'Hospitalet de Llobregat, Barcelona, Spain

[14]Internal Medicine Department, Bellvitge University Hospital, Barcelona, Spain

[15]Center for Biomedical Research in Cardiovascular Diseases (CIBERCV), Instituto Salud Carlos III, Madrid, Spain

[16]Hypnalgesics, Comfort Talk, Brookline, MA, United States

[17]Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda

[18]Biblioteca Médica, Hospital Universitario de Móstoles, Móstoles, Spain

[19]Fakher Mechatronic Research Center, Kerman University of Medical Sciences, Kerman, Iran

[20]Institute for Complementary and Integrative Medicine, University of Zurich, Zurich, Switzerland

[21]Institute for Social Medicine, Epidemiology and Health Economics, Charité – Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Jesus Lopez-Alcalde, MPH, MD
Institute for Complementary and Integrative Medicine
University Hospital Zurich
Sonneggstrasse 6
Zurich, CH-8091
Switzerland
Phone: 41 43 253 21 31
Email: jesus.lopez@usz.ch

Abstract

Background: Mobile health (mHealth) refers to using mobile communication devices such as smartphones to support health, health care, and public health. mHealth interventions have their own nature and characteristics that distinguish them from traditional health care interventions, including drug interventions. Thus, randomized controlled trials (RCTs) of mHealth interventions

XSL•FO
RenderX

present specific methodological challenges. Identifying and overcoming those challenges is essential to determine whether mHealth interventions improve health outcomes.

Objective: We aimed to identify specific methodological challenges in RCTs testing mHealth interventions' effects and develop consensus-based recommendations to address selected challenges.

Methods: A 2-phase participatory research project was conducted. First, we sent a web-based survey to authors of mHealth RCTs. Survey respondents rated on a 5-point scale how challenging they found 21 methodological aspects in mHealth RCTs compared to non-mHealth RCTs. Nonsystematic searches until June 2022 informed the selection of the methodological challenges listed in the survey. Second, a subset of survey respondents participated in an online workshop to discuss recommendations to address selected methodological aspects identified in the survey. Finally, consensus-based recommendations were developed based on the workshop discussion and email interaction.

Results: We contacted 1535 authors of mHealth intervention RCTs, of whom 80 (5.21%) completed the survey. Most respondents (74/80, 92%) identified at least one methodological aspect as more or much more challenging in mHealth RCTs. The aspects most frequently reported as more or much more challenging were those related to mHealth intervention integrity, that is, the degree to which the study intervention was implemented as intended, in particular managing low adherence to the mHealth intervention (43/77, 56%), defining adherence (39/79, 49%), measuring adherence (33/78, 42%), and determining which mHealth intervention components are used or received by the participant (31/75, 41%). Other challenges were also frequent, such as analyzing passive data (eg, data collected from smartphone sensors; 24/58, 41%) and verifying the participants' identity during recruitment (28/68, 41%). In total, 11 survey respondents participated in the subsequent workshop (n=8, 73% had been involved in at least 2 mHealth RCTs). We developed 17 consensus-based recommendations related to the following four categories: (1) how to measure adherence to the mHealth intervention (7 recommendations), (2) defining adequate adherence (2 recommendations), (3) dealing with low adherence rates (3 recommendations), and (4) addressing mHealth intervention components (5 recommendations).

Conclusions: RCTs of mHealth interventions have specific methodological challenges compared to those of non-mHealth interventions, particularly those related to intervention integrity. Following our recommendations for addressing these challenges can lead to more reliable assessments of the effects of mHealth interventions on health outcomes.

(J Med Internet Res 2024;26:e53187) doi:10.2196/53187

KEYWORDS

digital health; eHealth; mobile health; mHealth; randomized controlled trial; survey; recommendations; intervention integrity; adherence; consensus; mobile phone

Introduction

Mobile Health: Concept and Potential Impact

Mobile health (mHealth) refers to using mobile communication devices, such as smartphones, tablets, or smartwatches, to support health, health care, and public health [1-3]. An mHealth app is software incorporated into a mobile device to improve health, health care services, and research [4]. Due to the pressure on health systems, there is an urgent need to invest in health and health systems via effective, feasible, and sustainable interventions [5,6]. mHealth has the potential to improve health and health care through different means. One example is apps that monitor users and collect information, such as exercise, heart rate, or medication adherence [7]. Other apps support health care providers in diagnostic procedures or clinical decision-making for treatments [8]. Finally, mHealth apps can also deliver health interventions, which is this study's focus. Examples are apps to increase medication adherence or provide psychotherapeutic interventions (eg, cognitive behavioral therapies) to manage mental and physical health problems [3,9-14].

Effectiveness of mHealth Interventions: Promising Results, but High-Quality Randomized Controlled Trials Are Still Needed

The mHealth market size is rapidly expanding. There are >350,000 health apps, and the global market will probably grow by 40% between 2022 and 2030 [15,16]. Developers often advertise mHealth apps as improving health and well-being, but sound research should support their claims on the effects of mHealth interventions on clinical outcomes. However, the evidence that mHealth interventions can improve health care processes, change patient behavior, and improve health outcomes is still uncertain [14,15,17]. Randomized controlled trials (RCTs) are considered the most rigorous method for evaluating the efficacy and effectiveness of health interventions because they minimize bias [18]. Therefore, RCTs are crucial for generating mHealth evidence and drawing firm conclusions about mHealth interventions' effects [19-21].

Specific Methodological Challenges of mHealth Intervention RCTs: A Potential Barrier for Rigorous mHealth Research

mHealth interventions have peculiarities that distinguish them from other health care interventions, such as drug interventions. For example, mHealth interventions are usually behavioral interventions that are more complex to define and standardize than pharmacological interventions [22]. In addition, evaluating

XSL•FO
RenderX

and reporting mHealth intervention integrity (the degree to which the study intervention was implemented as intended) in RCTs can be challenging [23]. A recent systematic review by Tønning et al [24] outlines key methodological challenges in conducting RCTs for smartphone-based treatments in psychiatry. The challenges highlighted include the rapid pace of technological advancements, which may make interventions obsolete by the time articles are published. In addition, privacy and security issues often do not receive the attention they deserve. Moreover, mHealth RCTs lack consistency in how adherence to interventions is measured and how primary outcome measures are selected, making it difficult to compare results across different RCTs. Challenges related to missing data and the methods used for statistical analysis further complicate the assessment of mHealth interventions' efficacy. To address these issues, Tønning et al [24] recommend adopting more flexible trial designs, giving priority to ethical considerations concerning privacy and security, using technology for improved adherence tracking, and following detailed protocols that align with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines [25], among other recommendations. Previous findings indicate that evaluating mHealth intervention effects via RCTs has specific methodological challenges. Our research is the first to solicit direct feedback from mHealth researchers on methodological challenges unique to mHealth RCTs.

A survey and a consensus exercise were chosen for this project to address the diverse methodological challenges in mHealth RCTs. Consensus methods allow for the integration of expert perspectives to produce recommendations. This approach is essential given the lack of consistent solutions to these challenges. The goal of the consensus exercise was to develop recommendations for researchers working in mHealth. The intended audience is global, with participation from experts across regions, making the recommendations relevant for both high- and low-resource settings.

Identifying and overcoming the methodological challenges specific to RCTs of mHealth interventions (from now on referred to as mHealth RCTs) can clarify whether mHealth improves health outcomes. This study had two aims: (1) to identify specific methodological challenges in RCTs evaluating the effects of mHealth interventions and (2) to develop consensus-based recommendations to address selected methodological challenges in mHealth RCTs.

Methods

Overview

We undertook a 2-phase participatory research study with mixed methods (quantitative and qualitative; Figure 1). This manuscript follows the Checklist for Reporting Results of Internet E-Surveys and the Accurate Consensus Reporting Document guidelines [26,27]. The study protocol was not registered.

The steering group, which included 4 researchers (CMW, LSW, JB, and JL-A), defined the project aims, collected potential methodological challenges of mHealth RCTs, identified the potential survey respondents, developed the survey, and selected the workshop participants. JB, an experienced researcher, facilitated the consensus meeting. The steering group also analyzed the survey results, summarized participants' comments during the workshop, and integrated their feedback to generate this manuscript. No members of the public, patients, or carers were invited to participate in this study.

XSL•FO
RenderX

Figure 1. Study phases. EQUATOR: Enhancing the Quality and Transparency of Health Research; mHealth: mobile health; RCT: randomized controlled trial; WHO: World Health Organization.

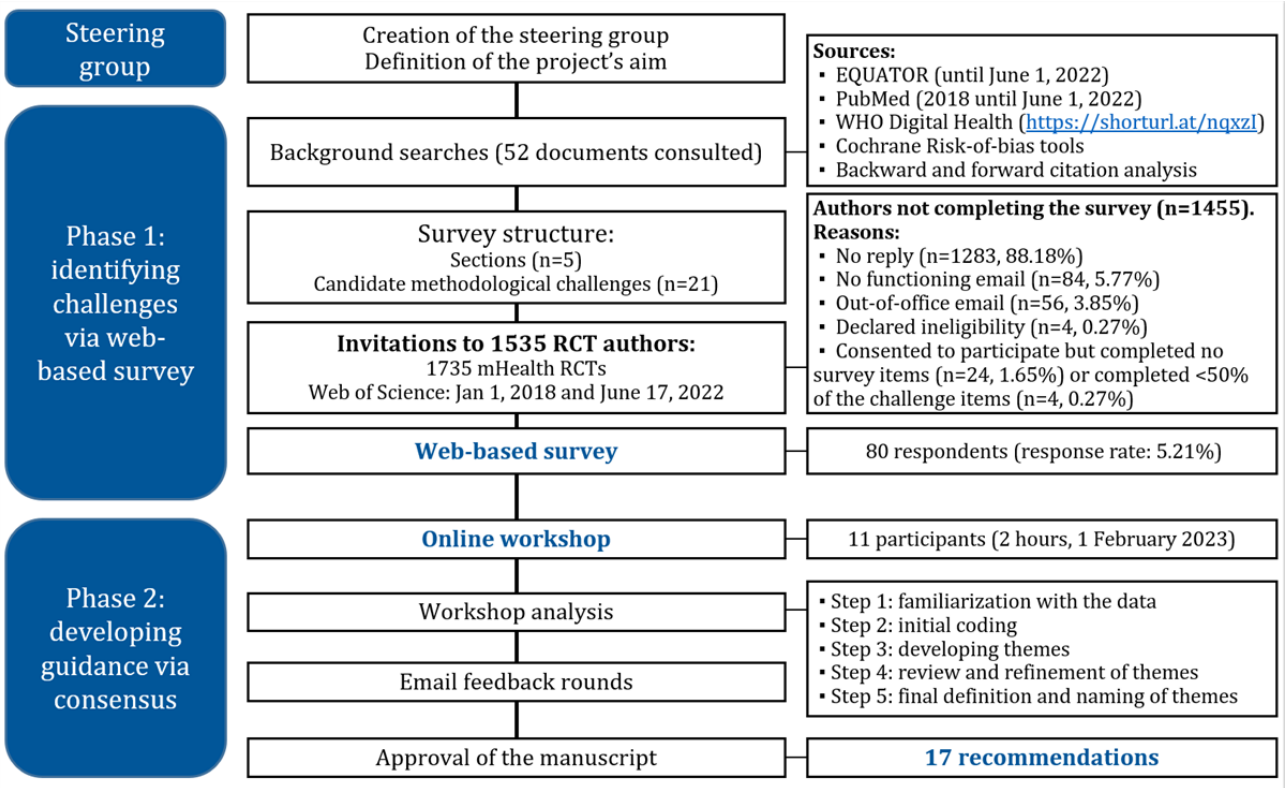

Phase 1: Identifying Methodological Challenges Specific to mHealth RCTs

The study design was a cross-sectional web-based survey.

Survey Items

We created an anonymous open voluntary web-based survey in English using SoSci Survey (SoSci Survey GmbH). The survey aimed at identifying methodological challenges specific to mHealth compared to non-mHealth RCTs. We performed nonsystematic searches in PubMed and EQUATOR (Enhancing the Quality and Transparency of Health Research) Network until June 1, 2022, looking for methodological challenges potentially relevant to mHealth RCTs (Figure 1; references of the documents consulted can be downloaded from Multimedia Appendix 1 [2,3,12,13,15,28-65]). On the basis of consensus, the steering group chose methodological aspects potentially specific to mHealth RCTs (initial list available upon request). The usability and technical functionality of the electronic questionnaire were tested by the steering group and 2 external researchers before fielding the questionnaire. The survey was available from September 23, 2022, to November 22, 2022 (Multimedia Appendix 2).

The first survey section listed 21 candidate methodological challenges in 5 sections (recruitment, randomization, intervention integrity, data quality, and data analysis). The questionnaire consisted of a total of 29 questions distributed over 6 pages with a maximum of 21 questions per page. The survey items were not randomized or alternated. Adaptive questioning (certain items were conditionally displayed based on responses to other items) was implemented to reduce the number of items. Participants described how challenging they found each methodological aspect in mHealth RCTs compared to non-mHealth RCTs. The response options were ranked on a 5-point scale: "much less challenging," "less challenging," "similar challenges," "more challenging," or "much more challenging." An "I don't know" option was also available. Participants could comment or propose additional challenges. Another survey section characterized the researchers' academic background and experience with mHealth and fully remote RCTs (recruitment, delivery of the intervention, and evaluation were conducted online using mobile devices or the internet). After each section of the questionnaire was filled in, participants had the opportunity to check for completeness. If any items were left unanswered, a message would appear to highlight the missing responses. Completing the missing items was not mandatory to minimize survey fatigue. Respondents were able to review and change their answers by clicking on the back button. Finally, the survey invited the respondents to the online workshop (participation was optional, and responses were stored separately).

Survey Sample

The sample frame included a convenience sample of 1535 authors from 1735 mHealth RCTs indexed in Web of Science (January 1, 2018, to June 17, 2022) who were invited via email to complete the survey. One invitation email (September 23, 2022) and a reminder (October 14, 2022) outlined the survey's purpose, duration, anonymity, funding source, the option to pause and resume the survey, and who the investigators were. Participants were encouraged to contact the team with questions. The email specified that the survey targeted trials evaluating mHealth interventions, excluding areas such as health

XSL•FO
RenderX

monitoring or diagnosis. No incentives were offered to participate. Multimedia Appendix 2 provides the full invitation email. Participants could complete the survey multiple times as the survey platform did not restrict submissions to unique visitors. Cookies were not used to assign a unique user identifier to each client computer. In addition, IP addresses were not used to detect potential duplicate entries from the same user, and no other methods were used to analyze the log file for identifying multiple submissions.

A librarian (CC-A) designed the search strategy for identifying mHealth RCTs and obtaining the authors' contact details (Multimedia Appendix 3). All survey respondents were required to confirm their participation in at least one mHealth RCT before completing the survey.

Survey Analysis

We used descriptive statistics (percentages, means and SDs, and medians and IQRs) to summarize dichotomous and quantitative data in narrative and tabular formats. We used R (R Foundation for Statistical Computing) to conduct statistical analyses [66]. We reported the proportion of participants who perceived each methodological aspect as more or much more challenging for mHealth compared to non-mHealth RCTs. Individuals who consented to participate but rated <50% of the listed challenges were excluded. Participants who responded "I don't know" or did not respond to a particular item were excluded from the analysis for that item. There was no time limit for completing the questionnaire. The results are presented for the entire survey sample and stratified according to the researchers' experience (experienced researchers were involved in more than one mHealth RCT). There were no statistical corrections to adjust for the nonrepresentative sample. Thematic analysis was used to group participants' comments into overarching categories [67] according to a 5-step process. JL-A conducted steps 1 to 3 (familiarization with the data, initial coding, and developing themes according to the topics of the survey). CMW reviewed and refined the themes in step 4, and then JL-A and CMW defined and named the themes in step 5.

Phase 2: Developing Consensus-Based Recommendations to Address Methodological Challenges in mHealth RCTs

The consensus method involved a consensus meeting, wherein participants engaged in discussions to reach an agreement. Recommendations from the workshop were endorsed through consensus without structured voting [26] (Multimedia Appendix 4).

Online Workshop

A 2-hour online workshop was held with mHealth researchers via a Zoom session (Zoom Video Communications) and was recorded with the participants' consent. We invited the 52 survey respondents who indicated their interest in participating in a workshop to develop consensus-based recommendations addressing methodological challenges specific to mHealth RCTs. The workshop included 4 sessions: (1) introduction of the participants, (2) project aims and survey results, (3) development of recommendations to overcome methodological challenges in mHealth RCTs, and (4) closing remarks. The workshop focused on 2 topics: mHealth intervention integrity and dealing methodologically with continuously updated apps (Figure 2). The first topic was frequently rated as more or much more challenging by the experienced mHealth researchers in the survey. The steering group chose to add the second topic for 2 reasons. First, the continuous requirement to update the apps is also related to mHealth intervention integrity. Second, the topic aligned with a parallel workshop addressing methodological challenges in mHealth systematic reviews [68].

Figure 2. Workshop topics. mHealth: mobile health.

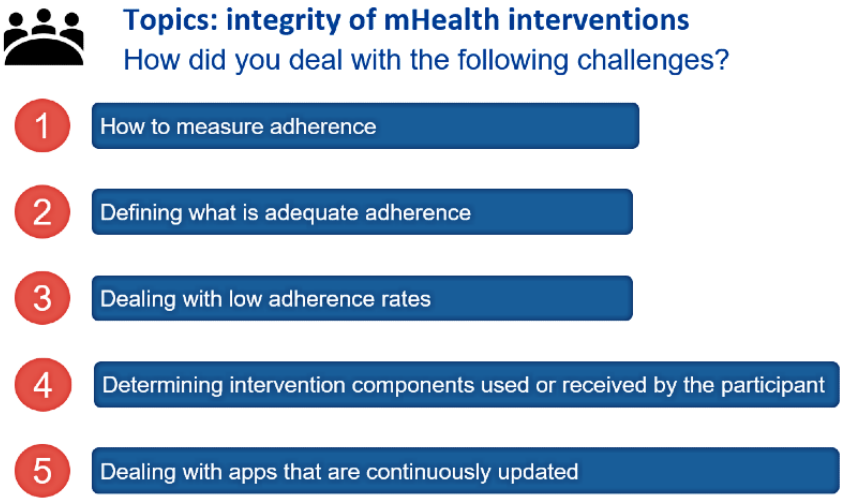

Workshop Analysis and Manuscript Feedback Rounds

JL-A and CMW performed a thematic analysis to summarize the workshop discussions [67] (refer to the steps in Figure 1). JL-A conducted steps 1 to 3 (familiarization with the data, initial coding from the transcription, and developing themes). CMW reviewed and refined the themes in step 4, and then JL-A and CMW defined and named the themes in step 5. Finally, all authors collaborated on step 6, organizing the final themes into a cohesive report, summarizing key insights, recommendations, and consensus points from the workshop discussions. The survey results and the workshop recommendations were integrated into the first manuscript draft, which was emailed to the workshop

XSL•FO
RenderX

participants. JL-A incorporated their feedback into the final version of the manuscript.

Ethical Considerations

We submitted the study synopsis to the Kantonale Ethikkommission (Kanton Zürich), Switzerland. After review, they stated that the study did not fall under the regulation of the Human Research Act of Switzerland (Business Administration System for Ethics Committees Req-2022-00839). The survey and meeting participants gave their consent to participate via the SoSci Survey tool. We conducted an anonymous web-based survey using the SoSci Survey platform hosted by University Hospital Zurich. Personal details such as names, email addresses, or IP addresses were not collected. In the final section of the survey, respondents were invited to participate in an optional online workshop. The responses to this invitation were stored in a separate SoSci Survey project. Participants interested in attending the workshop provided their name, email, and place of residence, among other details. To ensure privacy and confidentiality, all data were securely collected and stored at University Hospital Zurich. In addition, participants' data were managed in strict accordance with institutional data protection policies, ensuring that their privacy was rigorously protected throughout the research process. Survey participants did not receive compensation. While workshop participants were not offered financial compensation, they were given the opportunity to be coauthors of the manuscript if they met the authorship criteria.

Results

Survey Results

A total of 80 authors of mHealth RCTs completed the survey. The completion rate was 40.2% (106/264, corresponding to the number of people who agreed to participate divided by the number of visitors). During testing, it consistently took <10 minutes to complete the survey. Furthermore, a similar survey conducted by our team, which had a comparable number of items but focused on challenges in systematic reviews of mHealth interventions [68], took an average of 5 (SD 2) minutes to complete, with a range of 2 to 11 minutes.

Figure 1 details the reasons for not completing the survey. The respondents' academic backgrounds were diverse, and the most frequent were medicine (24/80, 30%), psychology (20/80, 25%), and epidemiology (12/80, 15%). Most participants (52/80, 65%) were experienced mHealth researchers who had conducted ≥2 mHealth RCTs. The survey respondents had been involved in at least 185 mHealth RCTs. Most trials (56/185, 30.3%) were conducted in North America and Europe. Few RCTs were conducted in Africa (7/185, 3.8%) and South or Central America (4/185, 2.2%; Table 1).

XSL•FO
RenderX

Table 1. Characteristics of the survey respondents (N=80).

Characteristic	Participants
Academic background[a], n (%)	
Medicine, nursing, or public health	32 (40)
Psychology or social sciences	25 (31.3)
Epidemiology	12 (15)
Physiotherapy or sports science	7 (8.8)
Biology or neuroscience	6 (7.5)
Nutrition	4 (5)
Engineering or computer science	5 (6.3)
Other backgrounds[b]	21 (26.3)
Clinical research experience in addition to mHealth[c] RCTs[d], n (%)	
Evaluation of nonpharmacological interventions	45 (56.3)
Evaluation of pharmacological interventions	0 (0)
Evaluation of nonpharmacological and pharmacological interventions	12 (15)
Other areas (eg, basic research, surveys, and prognosis research)	23 (28.8)
Experience in mHealth RCTs, n (%)	
Experienced researchers (≥2 mHealth RCTs)	52 (65)
Less experienced researchers (1 mHealth RCT)	28 (35)
Researchers involved in at least one fully remote mHealth RCT[e]	49 (61.3)
Number of mHealth RCTs per respondent	
Mean (SD)	2.3 (1.4)
Median (IQR)	2.3 (1-3)
≥5, n (%)	12 (15)
4, n (%)	4 (5)
3, n (%)	9 (11.3)
2, n (%)	27 (33.8)
1, n (%)	28 (35)
Location of the survey participants' mHealth RCTs[f] (n=185), n (%)	
United States	16 (8.6)
The Netherlands	9 (4.9)
Australia	8 (4.3)
United Kingdom and Canada (each)	7 (3.8)
Spain	6 (3.2)
China, India, and Uganda (each)	3 (1.6)
Finland, France, Iran, Kenya, Malaysia, New Zealand, and Sweden (each)	2 (1.1)
Argentina, Brazil, Germany, Guatemala, Hong Kong, Italy, Malawi, Norway, Peru, Portugal, Singapore, South Africa, South Korea, Switzerland, Taiwan, and Vietnam (each)	1 (0.5)
Participants starting an mHealth RCT under the new Medical Device Regulation in Europe (in place since May 2021), n (%)	9 (11.3)

[a]24% (19/80) of the participants had more than one background.

[b]Other backgrounds (one participant each): behavior science, biostatistics, dentistry, design science, digital health, economics, lifestyle medicine, mathematics, medical informatics, molecular biology and biochemistry, and statistics.

[c]mHealth: mobile health.

XSL•FO
RenderX

[d]RCT: randomized controlled trial.

[e]RCT in which the recruitment, intervention delivery, and evaluation were conducted online using mobile devices or the internet.

[f]3 RCTs took place in more than one country.

Most respondents (74/80, 92%) identified at least one methodological aspect as more or much more challenging in mHealth RCTs compared to non-mHealth RCTs. Survey respondents rated a median of 6 (IQR 4-9) out of 21 aspects as more or much more challenging (Multimedia Appendix 5). The methodological aspects most frequently perceived as more or much more challenging in mHealth RCTs were the following: managing low adherence to the mHealth intervention (43/77, 56%), defining adherence (39/79, 49%), measuring adherence (33/78, 42%), analyzing passive data (eg, data collected from smartphone sensors; 24/58, 41%), determining which active components of the mHealth intervention are used or received by the participant (31/75, 41%), and verifying the participants' identity during recruitment (participants can register more than once using different identities; 28/68, 41%). In contrast, the aspects described by the lowest number of respondents as challenging were analyzing patient-reported outcomes (6/75, 8%) and analyzing clustered data (7/62, 11%; Table 2). Multimedia Appendix 6 lists additional challenges reported by the respondents.

XSL•FO
RenderX

Table 2. Researchers perceiving each methodological aspect as more or much more challenging in mobile health (mHealth) randomized controlled trials (RCTs) compared to non-mHealth RCTs stratified according to experience in mHealth (N=80).

	Total sample, n/N (%)	Experienced researchers (n=52), n/N (%)[a]	Less experienced researchers (n=28), n/N (%)[a]	Absolute difference—experienced researchers – less experienced researchers (%)[b]
Challenges in recruitment				
Achieving the required sample size	22/78 (28)	10/51 (20)	*12/27 (44)* [c]	*–24*
Enrolling a representative sample of the desired trial population	27/77 (35)	15/50 (30)	*12/27 (44)*	–14
Verifying the participants' identity during recruitment (participants can register more than once using different identities)	*28/68 (41)*	*21/45 (47)*	7/23 (30)	17
Implementing the informed consent procedure	10/78 (13)	6/51 (12)	4/27 (15)	–3
Challenges related to randomization				
Concealing the randomization sequence during allocation	17/76 (22)	10/49 (20)	7/27 (26)	–6
Including the randomization process on the app	16/55 (29)	12/37 (32)	4/18 (22)	10
Challenges related to intervention integrity				
Defining adherence to the mHealth intervention	*39/79 (49)*	*25/52 (48)*	*14/27 (52)*	–4
Measuring adherence to the mHealth intervention	*33/78 (42)*	*22/52 (42)*	*11/26 (42)*	0
Determining the active mHealth intervention components used or received by the participants	*31/75 (41)*	20/51 (39)	*11/24 (46)*	–7
Managing low mHealth intervention adherence rates	*43/77 (56)*	*30/52 (58)*	*13/25 (52)*	6
Challenges related to data quality				
A high proportion of participants lost to follow-up	30/78 (38)	20/52 (38)	10/26 (38)	0
Differential follow-up rates between intervention and comparator groups	17/70 (24)	10/47 (21)	7/23 (30)	–9
Suitable outcome measurement instruments for mHealth trials	20/78 (26)	17/52 (33)	3/26 (12)	*21*
Large amounts of missing data	31/79 (39)	18/52 (35)	*13/27 (48)*	–13
Verifying the validity of the data	22/76 (29)	14/51 (27)	8/25 (32)	–5
Challenges in data analysis				
Analyzing clustered data	7/62 (11)	4/42 (10)	3/20 (15)	–5
Analyzing large amounts of data	15/72 (21)	10/50 (20)	5/22 (23)	–3
Analyzing patient-reported outcomes (ie, self-reports)	6/75 (8)	4/52 (8)	2/23 (9)	–1
Analyzing passive data (eg, data collected from smartphone sensors)	*24/58 (41)*	*18/41 (44)*	6/17 (35)	9
Analyzing data using repeated measurements across time	12/75 (16)	4/51 (8)	8/24 (33)	*–25*
Dealing with large amounts of missing data	21/72 (29)	12/50 (24)	*9/22 (41)*	–17

[a]Experienced mHealth researcher: involved in at least 2 mHealth RCTs. Less experienced mHealth researcher: authored 1 mHealth RCT.

[b]Absolute difference: experienced mHealth researchers (percentage) – less experienced mHealth researchers (percentage).

XSL•FO
RenderX

[c]Italics represent (1) challenges defined by at least 40% of the participants as more or much more challenging or (2) challenges with an absolute difference between experienced and less experienced mHealth researchers of >20 percentage points. Participants who responded "I don't know" or did not respond to a particular item were excluded from the analysis for that item.

Experienced and less experienced researchers' perceptions did not differ markedly. However, a higher number of less experienced researchers considered the following aspects as more or much more challenging: achieving the required sample size (12/27, 44% vs 10/51, 20%) and analyzing data using repeated measurements across time (8/24, 33% vs 4/51, 8%). On the other hand, more experienced researchers considered the availability of suitable outcome measurement instruments for mHealth trials as more or much more challenging (17/52, 33% vs 3/26, 12%).

Online Workshop

In total, 11 mHealth researchers attended the online workshop (2 hours; February 1, 2023; Table 3). Figure 2 outlines the topics that guided the discussion during the workshop.

Table 3. Characteristics of the workshop participants (N=11).

Characteristic	Participants, n (%)
Female participants	7 (64)
Geographical area of residence	
Europe	4 (36)
Switzerland	2 (18)
Spain	1 (9)
United Kingdom	1 (9)
North America	2 (18)
Canada	1 (9)
United States	1 (9)
Africa	2 (18)
Kenya	1 (9)
Uganda	1 (9)
Asia	3 (27)
Iran	2 (18)
India	1 (9)
Number of mHealth[a] RCTs[b] per participant	
8	1 (9)
6	1 (9)
3	4 (36)
2	2 (18)
1	3 (27)

[a]mHealth: mobile health.

[b]RCT: randomized controlled trial.

Recommendations for Addressing Methodological Challenges Specific to mHealth RCTs

The following recommendations were agreed upon during the workshop and refined via email afterward following a qualitative process.

How to Measure Adherence to the mHealth Interventions?

Adherence to the intervention is the degree to which the study participants followed the intervention as intended, that is, whether the participants performed the planned intervention and avoided proscribed procedures. Measuring and reporting adherence in RCTs is essential because a low adherence can affect the validity and generalizability of the trial results [69].

Recommendation 1: Researchers Should Use a Variety of Methods to Assess Adherence in mHealth Trials

Several methods can measure adherence in mHealth trials, but no gold standard exists, and each method has limitations [70]. Thus, adherence in mHealth trials should be estimated from a triangulation of sources, such as the following:

- Self-report can be used for measuring adherence in mHealth trials, but this method can provide inaccurate data [71].
- Information from interviews, focus groups, or surveys serves to assess the app's acceptability, check whether the

XSL•FO
RenderX

participants adhered to the planned intervention, or test whether the participants acquired the minimum knowledge required for the mHealth intervention to have an effect.

- Direct observation is an accurate way to measure adherence, but it is also more expensive and time-consuming. This method is typically only used in small-scale studies and may not apply to mHealth RCTs.
- Aggregated and anonymous use data collected by the app (eg, tracking and dashboard mapping) are another option. Examples are the percentage of participants who opened the app or the average time spent by the participants on the app.
- Biomarkers are measurable indicators of a biological process or condition [72]. Some biomarkers can measure adherence to mHealth interventions. For example, a study used a wireless glucometer to measure adherence to diabetes self-management among patients [73]. However, biomarkers are not always available or accurate.

Recommendation 2: Future Research Should Inform Methodological Guidance to Link Adherence Information Collected Through Different Methods

Methodological guidance is needed to link the adherence information collected through the interaction between researchers and participants (eg, via interviews) with aggregated and anonymous adherence data routinely collected by the app (not available at the individual level due to privacy issues). Researchers should try to explain discordant adherence findings depending on the source. For example, self-declaration by the user of the time spent on the app may under- or overestimate the time registered by the app.

Recommendation 3: Adherence Measurement in mHealth RCTs Should Consider That Interventions Initially Delivered via the App May Not Require App Engagement at a Later Stage

Some interventions (eg, relaxation techniques or exercise) may be initially delivered via the app used in the evaluation. However, patients may learn to apply the technique alone without the app. Thus, the app is no longer used (low adherence), but the main component of the intervention (eg, relaxation techniques or exercise) is applied. Methodological guidance to measure adherence in these situations is needed.

Recommendation 4: The Protocol and Consent Form of mHealth RCTs Should Detail How Adherence Data Will Be Collected, Managed, and Analyzed

Trial protocols should plan data-sharing activities transparently. If adherence data will be collected via the app, the purpose of the data collection and the analysis strategy should be detailed at the protocol development stage and on the trial consent form. Use data may contain sensitive personal information. Sharing and analyzing such data without proper consent and protection may violate the participants' privacy rights and expose them to potential risks.

Recommendation 5: The Trial Team Should Clarify With the App Developer Whether Data Privacy and Security Protection Issues Allow for Sharing Participants' Use Data Between the App Developer and the Research Team

Data privacy and security protection issues may vary depending on the country, region, or institution where the trial is conducted. Different regulations and standards may apply to collecting, storing, transferring, and analyzing participants' use data. Therefore, the trial team and the app developer should comply with the relevant laws and policies. Moreover, they should agree on the methods, formats, and frequency of data sharing, as well as the quality control and verification procedures.

Recommendation 6: The Measurement of mHealth Intervention Adherence Should Consider the Internet Access Requirements and Interoperability of the mHealth App

Access to the internet and interoperability should be carefully considered when measuring adherence to mHealth interventions. Researchers should differentiate between low adherence due to individual factors (eg, low app use by the participants because they did not like the app) or system factors (low app use due to connectivity difficulties).

Access to the internet is critical for mHealth intervention integrity as mHealth apps often require an internet connection to work. If users do not have good internet access, the functionality of the mHealth intervention can be compromised. In this line, poor internet access can reduce mHealth intervention integrity by worsening the app's functioning and the participants' adherence to the intervention.

Interoperability of the app with other software is important as this factor is related to the app's technical functioning, user satisfaction, and adherence. For example, the app may need to work with other apps, such as calendars, reminders, and sensors, to provide optimal user experience and support.

Recommendation 7: Adherence Assessment in mHealth Trials Should Not Apply the Same Approaches Used in Drug Efficacy Trials

Different approaches are needed to measure adherence to mHealth interventions, and guidance is needed to tailor the measurement of adherence in mHealth trials. In this line, measures from behavior change research might be helpful in mHealth RCTs. There are critical differences concerning the assessment of adherence in RCTs of mHealth and drug interventions.

First, adherence to mHealth interventions has more dimensions than adherence to drugs. It involves the frequency and duration of app use, the quality and intensity of engagement, compliance with behavioral recommendations, and achievement of personal goals, among other aspects [74,75]. This makes it more challenging to measure adherence to mHealth interventions.

Second, adherence to mHealth interventions can be affected by the design and functionality of the app, the feedback and support provided, and the integration with other health services. Moreover, mHealth intervention adherence can vary over time and across contexts depending on the intervention duration, intensity, and frequency [76-79].

Third, the dose or course of an mHealth intervention can be adjusted according to the specific characteristics and needs of

XSL•FO
RenderX

the participants. Consequently, adherence to mHealth interventions may affect the trial outcomes differently than drug adherence. For example, some participants may benefit from a short mHealth intervention, whereas others may need a longer course to achieve the desired effects [79,80].

Fourth, drug efficacy trials are typically conducted in controlled contexts, such as hospitals or specific clinical settings. This allows researchers to tightly control the environment and ensure that the intervention is delivered as intended. On the other hand, mHealth interventions are usually evaluated in real-world contexts (effectiveness), where many factors can affect adherence, such as access to the internet and technical problems. Thus, mHealth RCTs should not apply the same methods used in drug efficacy trials to assess adherence.

Defining Adequate Adherence

Recommendation 8: Predefine the mHealth Intervention Components Needed to Generate the Expected Intervention Effect and Measure Adherence Based on Those Components

mHealth interventions are multifaceted, making adherence measurement challenging. Specifying the critical mHealth intervention components (*must* items) and desirable components (*should* items) needed to generate the expected effect will help define what is adequate adherence.

There is no definitive answer to which mHealth intervention components are needed to generate the expected effects. Some factors that may contribute to the efficacy or effectiveness of mHealth interventions are the following: (1) the intervention intensity (frequency and delivery of the intervention over time); (2) its patient-centered approach (ie, tailoring the intervention content and interaction to the specific needs, preferences, and characteristics of the patients); and (3) their multifactorial approach, that is, the combination of the mHealth intervention with existing health systems and services [78,81].

If researchers are involved in designing the app or the mHealth intervention (rather than only its evaluation), they might consider conducting a scoping review to inform the logic model that outlines the components of the mHealth intervention. The selection of components should be underpinned by either existing evidence or an explanation for their choice. A logic model should also define whether the delivery sequence of the intervention components is a critical factor that can affect the outcome. If this is the case, the component sequence should also be considered to judge mHealth intervention integrity.

There is a need for a better description by app developers and researchers of each component of an mHealth intervention. Developing a common taxonomy for mHealth interventions and their components, including the behavioral techniques used, would help describe mHealth interventions in a standardized manner.

Recommendation 9: Future Research Should Guide How to Define Adequate Adherence in mHealth Interventions

There is no definitive answer to how to define adequate adherence to an mHealth intervention as it depends on the mHealth intervention itself, the outcome evaluated, the research aim (efficacy or effectiveness), and the measures of adherence used. The definition of adequate adherence should be based on the minimum intervention intensity or dose needed to achieve a clinically relevant effect. This information should be underpinned by existing evidence or a theoretical basis and reflect adherence to all the critical intervention components to achieve the desired effect.

The current understanding of factors that can either enhance or hinder adherence to mHealth apps remains limited, largely due to the fact that most of the underlying studies are pilot studies with short durations [70]. Future research on mHealth app adherence should clearly outline the app's intended use; report objective data on actual use relative to the intended use; and, ideally, inform the long-term use and retention of the mHealth intervention [70].

Dealing With Low Adherence Rates to the mHealth Intervention

Overview

Adherence in mHealth intervention studies is usually low. A systematic review including 99 studies of health apps for preventing or managing noncommunicable diseases indicated that the mean adherence across all interventions was 56% (range 3%-96%) [70]. Medication adherence in drug trials seems to be higher, although this aspect is frequently not reported. For example, a systematic review on medication adherence in RCTs of patients undergoing dialysis concluded that the mean medication adherence was 81% and 84% in the intervention and control arms, respectively [82].

Addressing low adherence rates in trials is essential because low adherence can reduce statistical power and introduce bias, particularly by making a potentially efficacious intervention appear less effective. Consequently, low adherence in mHealth RCTs can lead to misleading recommendations for patients, providers, and decision makers [83].

Recommendation 10: Consider Using Strategies to Increase Adherence in RCTs of mHealth Interventions

Researchers may consider using different strategies to increase adherence in mHealth RCTs (with the caveats addressed in recommendation 11). Possible strategies are the following [84-87]. The first is to provide real-time or regular feedback to the participants based on their performance, such as congratulating them for achieving their goals, encouraging them to improve their adherence, or suggesting tips or strategies to overcome barriers. The second is to send reminders to the participants via SMS text messages, voice calls, or push notifications to help them remember to use the mHealth app. The third is to apply behavior change techniques to enhance the participants' motivation and self-efficacy to adhere to the mHealth intervention. The fourth is to tailor the mHealth intervention to the specific needs and preferences of the participants, such as allowing them to choose the frequency, timing, content, and format of the messages or feedback or personalizing the intervention based on their demographic, clinical, or behavioral characteristics. The fifth is to integrate the mHealth intervention with the existing health systems and services, such as ensuring the coordination of the mHealth app

XSL•FO
RenderX

with health care providers or involving health care providers in delivering or supporting the intervention.

Recommendation 11: Define the Role of the Strategy to Increase Adherence to the mHealth Intervention

The trial protocol should specify the function of the strategy to enhance adherence to the mHealth intervention. Whether this strategy is an integral part of the mHealth intervention or a measure implemented merely in the RCT to boost adherence needs to be determined.

Researchers should be aware that incorporating strategies to improve adherence in RCTs can bias the study results, especially in studies with active comparators that did not use these strategies. Therefore, if one group receives additional strategies to improve adherence but the other group does not, this could introduce a performance bias [88]. Furthermore, if the strategy is not going to be a component of the mHealth intervention in real-world settings, the adherence observed in the RCT may not accurately represent adherence in actual settings. Researchers should carefully consider the ethics of using incentives to boost adherence in mHealth RCTs. While incentives are not inherently unethical, ethical concerns primarily emerge from their large-scale implementation in real-world scenarios. This is especially true if the incentives are expensive or challenging to sustain, raising issues regarding their long-term viability [4,5].

Recommendation 12: Involve Experts in Clinical Trial Design and Analysis to Handle High Amounts of Missing Participant Data and Low Adherence to the mHealth Intervention

Missing participant data, that is, missing outcome data for participants in a trial, is a common problem in RCTs. Missing participant data can occur when participants drop out of the study, fail to complete specific assessments, or are lost to follow-up [89]. Missing data problems are likely exacerbated in mHealth RCTs and affect their validity [90]. For example, early dropouts (recruited participants who did not start using the intervention) are probably a common and essential issue in mHealth RCTs [91]. Moreover, as stated previously, low adherence rates are frequent in mHealth RCTs [70].

mHealth RCTs should involve statisticians and experts in trial design and analysis to implement appropriate strategies to handle and analyze high amounts of missing participant data and low adherence. There is no agreed upon guidance, but the implemented approach should follow these general principles: (1) anticipate high amounts of missing participant data in the sample size calculation, (2) clearly report missing participant data (numbers and reasons per study arm), (3) conduct the primary analysis under the most plausible assumption as to why the data are missing, and (4) conduct sensitivity analyses considering alternative plausible assumptions to assess the robustness of the conclusions [89,92]. The importance of tracking adherence metrics and understanding the factors that lead to participant dropout should be quantified, analyzed, discussed, and reported. This also involves examining the attributes of the subset of participants who remain in the trial and use the mHealth intervention [91].

Addressing mHealth Intervention Components

Overview

The turnover in mHealth app development is fast. The rate at which mHealth apps are created, updated, or discontinued by developers is high. Consequently, obtaining relevant data to support future research and decision-making can be complex because critical information about the app (or the version evaluated in a trial) may no longer be available. However, the critical set of information that should be shared is still to be agreed upon.

Recommendation 13: Report the App's Technical Specifications and Cost and the Regulations Fulfilled During Its Development and Evaluation

It is essential to report the technical specifications of the app evaluated in the trial, such as version, launch date, compatibility with different devices and operating systems, storage, and security features. The cost of the app is also important, including any subscription or in-app purchase fees. All this information helps assess the accessibility of the app, particularly for populations with limited financial resources.

mHealth app development is subject to regulations and standards from different authorities and organizations, such as compliance with data collection and storage practices and data protection regulations. Knowing these regulations is vital for future researchers and app users as they impose specific requirements and limitations that may affect the apps' use.

Recommendation 14: App Developers and Researchers Should Report Users' Feedback Regarding the mHealth Intervention

App developers and researchers should report the findings of user-centered methods used to design or evaluate the mHealth intervention, for example, focus groups on the app's usability. Critical information could include the users' views on app functionality, usability, privacy, security, and quality [93]. In addition, the users' literacy with the mHealth intervention during the trial should be reported.

User feedback can provide valuable insights such as the following to understand the role of each mHealth intervention component and explain adherence: (1) the processes, mechanisms, and outcomes of mHealth interventions, which can help improve their design, implementation, and evaluation; (2) the strengths and weaknesses of the mHealth intervention, such as its usability, functionality, acceptability, and effectiveness, from the perspective of the end users; and (3) barriers and facilitators that influence the adoption, engagement, and retention of mHealth interventions. User feedback can contribute to the knowledge translation of mHealth interventions by sharing the experiences, opinions, and preferences of the target population with other researchers and stakeholders [93-97].

Recommendation 15: Consider Incorporating Scaling Lessons Learned in the Technology Field Into the mHealth Intervention Development Process and Evaluation

Traditional health intervention evaluation often relies on lengthy, static RCTs. However, this method fails to capture the dynamic

XSL•FO
RenderX

nature of mHealth app development. In contrast, technology companies such as Amazon use agile online randomized experiments on a large scale, running hundreds of simultaneous controlled experiments on millions of users. This approach is used to launch new products, assess their value, enhance customer experience, and implement code modifications. A/B testing is one such method. It is a quick experimentation process that efficiently evaluates multiple design options using live environments and real users. Essentially, A/B testing compares 2 versions of a variable to determine which one performs better [98-100].

Developers and researchers should consider using A/B testing methods to evaluate the impact of different features of the mHealth intervention. A/B testing can help identify app attributes that promote adherence and positive outcomes, minimize unintended consequences, or predict poor user experience. This approach can be applied to both major and minor design choices, such as the placement of icons on the app. Therefore, mHealth app developers and researchers should consider the involvement of experts who can analyze large amounts of data per user, swiftly implement improved versions, assess the relative impact of the modifications, and then modify the product for the next round [100]. However, A/B testing methods can be expensive and require that the individuals conducting the trial are also able to change targeted aspects of the software.

Recommendation 16: The Impact of App Updates Should Be Carefully Considered in RCTs of mHealth Interventions

App updates often include changes to the app's functionality and user interface and bug fixes. If some users have updated versions whereas others do not, it could introduce variability in data collection and user experience, potentially impacting the overall findings. For these reasons, mHealth RCTs must handle app updates carefully to ensure valid and reliable results. This information will help evaluators, users, and prescribers judge whether the mHealth intervention has evolved into a different intervention.

It is important to define the key aspects (a core set of items) that should be described regarding app updates. Some of these items for mHealth RCTs may be the following: (1) the version that was evaluated and whether newer app versions were implemented during the trial; (2) implemented app updates—key mHealth intervention components that were modified along the trial, reasons, and periodicity; and (3) app changes that were considered major changes (distinguishing between technical updates and content updates).

The RCT protocol should plan how to manage the app updates, and the ethics submission should also describe relevant aspects, such as the app changes that will be communicated to the ethics committee or the ethical implications of the use of apps that are not updated. It is also important that researchers and app developers collaborate and clarify the potential implications of updating the app for the validity and reliability of the study results.

Recommendation 17: The RCT Protocol Should Carefully Choose and Define the Control Intervention

Control conditions in mHealth trials are crucial for establishing the efficacy of the intervention. Each type of comparator has its strengths and limitations, and its choice should be carefully based on the research question, ethical considerations, and the nature of the mHealth intervention [101,102]. The following are some types of control conditions that can be used in mHealth trials.

- Sham control—a control group with a placebo or inactive treatment. This type of control can be used in mHealth trials in which the control group might receive a sham app with no active ingredients.
- Active control—a control group that is given an existing, proven intervention. This type of control is often used when it would be unethical to give a placebo, such as when an effective treatment already exists.
- A/B testing—the control group is another version of a variable to determine which one performs better (recommendation 15).
- Waitlist control—a control group that receives the intervention after the active group has already received it. This type of control is often used in psychological and behavioral interventions.
- Attention control—a control group that receives an intervention that is intended to control for the amount of attention received by the active group.
- Treatment as usual—a control group that continues to receive the standard care or treatment as usual while the experimental group receives the new intervention.

Discussion

Principal Findings

We contacted 1535 authors of mHealth intervention RCTs, of whom 80 (5.21%) completed the survey. Our research identified specific methodological challenges in RCTs of mHealth interventions. The aspects most frequently reported as challenging were those related to mHealth intervention integrity, in particular managing low adherence to the mHealth intervention, defining adherence, measuring adherence, and determining which mHealth intervention components are used or received by the participants. Other challenges were also frequent, such as analyzing passive data and verifying the participants' identity during recruitment to the RCT. The workshop concentrated on the integrity of mHealth interventions. Following the workshop and subsequent email discussions, we established 17 consensus-based recommendations to address selected methodological challenges related to mHealth intervention integrity.

Strengths and Limitations

Our study has several strengths. First, we likely identified the main methodological hurdles specific to mHealth RCTs. The process began with nonsystematic literature searches that informed the potential challenges listed in the survey. Interestingly, survey respondents identified only 1 methodological challenge not included in our survey: finding

XSL•FO
RenderX

a suitable control condition, such as creating a placebo app. Second, the survey and workshop participants had diverse experiences, having been involved in 1 to >5 mHealth RCTs each. This range of expertise allowed our study to encapsulate the challenges encountered by mHealth researchers with varying experience levels. Furthermore, the workshop participants were involved in mHealth RCTs conducted in high- and low-resource settings, such as the United States, Argentina, Uganda, and Iran. This diversity probably helped capture the challenges that researchers face in different contexts. Third, a key strength of our project was adherence to reporting guidelines for online surveys and consensus documents. This alignment with best practices enhanced the credibility and reproducibility of our work (Multimedia Appendices 7 and 8).

This study has several limitations. First, the low response rate may have affected the representativeness of the results as participants may have had a specific interest or experience in mHealth RCTs. Contributing factors include inoperative emails, out-of-office replies, the specificity of the topic, and survey fatigue. Online surveys generally suffer from lower response rates compared to other methods [103,104]. In our survey, despite sending reminders and attempting to reach a diverse pool of researchers, this limitation persisted. Future studies could use additional strategies such as incentives or targeted recruitment to improve participation rates.

Second, only 15% (12/80) of the respondents had experience with both mHealth and non-mHealth interventions, limiting the survey's ability to capture the challenges unique to mHealth RCTs. Including more researchers with experience in both areas could provide a more comprehensive understanding of these challenges in future studies.

Third, the study did not account for differences between various types of mHealth interventions. These interventions vary significantly, ranging from simple SMS text messaging to complex, multicomponent apps, each presenting distinct challenges. This variability may have influenced the findings. Future research should categorize mHealth interventions to provide more targeted recommendations for addressing specific challenges.

Fourth, the complexity of measuring adherence in mHealth RCTs was not fully explored. Adherence is difficult to measure due to factors such as user engagement, app use, and internet access. While adherence was identified as a key challenge, more research is needed to develop standardized methods for its measurement. Future studies should also examine how adherence differs across various types of mHealth interventions and how it impacts study outcomes.

Comparison With Prior Work

To our knowledge, our survey is the first to ask mHealth researchers about methodological challenges specific to mHealth RCTs. Our findings align with those of previous research on the challenges in mHealth intervention research. A review of RCTs evaluating the effects of mHealth interventions in chronic disease management identified the following challenges as the most important ones: designing high-quality studies, developing robust interventions in combination with health professional input, and identifying tools and methods to improve patient adherence [105]. Another systematic review pinpointed methodological challenges in RCTs of smartphone-based interventions in psychiatry. These included the need for flexible trial designs that can quickly adapt to the fast-evolving nature of mHealth interventions, privacy and security concerns, the lack of standardized adherence metrics, and the difficulties in choosing primary outcomes and handling missing data [24].

Our research advances mHealth methodology by addressing specific challenges not comprehensively covered in other studies [24,25,105]. First, we examined intervention integrity, highlighting the intricacies of defining and measuring adherence and pointing to the need to define the essential mHealth intervention components. Second, we spotlighted the challenge of analyzing passive smartphone sensor data, a significant digital health issue overlooked in previous research focused on traditional data analysis challenges. Third, our work underscores the unmet need for effective digital identity verification in studies relying on online recruitment, a crucial aspect often overlooked in previous research in favor of recruitment strategies and sample size considerations.

Our study sets itself apart from previous research by calling for a clear definition of mHealth intervention adherence, emphasizing the need to compare between the actual and intended use of the intervention against predefined thresholds. Previous research has broadly addressed mHealth adherence challenges, noting that poor reporting standards in mHealth RCTs hinder adherence assessment [24,70,106]. In addition, our study distinguishes itself from previous research by advocating for a transparent framework that requires researchers to predefine the intended use of the mHealth intervention and choose clear thresholds for adherence evaluation. However, we recognize our limitation in not offering specific guidelines for measuring adherence, for example, by proposing adherence thresholds. Nevertheless, advocating for a transparent and operative framework to measure mHealth intervention adherence is a first important step.

Most of the methodological challenges that this study highlighted are associated with mHealth intervention integrity. There is no gold standard for measuring adherence in mHealth trials [70,107]. This construct is challenging to define and measure in mHealth trials for several reasons. First, mHealth trials have different factors influencing adherence compared to drug interventions, such as more diverse participant characteristics, preferences, and context, such as internet access [19]. The key difference likely stems from drug interventions occurring in controlled settings versus mHealth interventions occurring in real-world settings, which are less controlled and more variable [8,9]. Second, adherence can be measured at different trial phases, such as at intervention initiation (when the participants begin the intervention), during the implementation period, or at discontinuation (the permanent cessation of the randomized intervention) [108]. Third, the data collection sources and methods to measure adherence are diverse, such as self-reporting, interviews, or electronic tracking via mobile devices [109]. In this line, adherence definitions and thresholds can vary, such as the number of uses of the app, the time spent using the app, or the proportion of participants who

XSL•FO
RenderX

achieved a certain level of adherence. Fourth, mHealth trials have different factors influencing adherence compared to drug interventions, such as more diverse participant characteristics, preferences, and context (eg, access to the internet network connectivity) [19].

Consequently, mHealth intervention integrity is seldom discussed in RCTs, likely due to the absence of universally accepted definitions and measurements [69]. Furthermore, existing reporting guidelines do not advise on how to report the integrity of mHealth interventions [25,110-114], and it remains uncertain how mHealth intervention integrity information should be managed by researchers and decision makers. Nevertheless, mHealth RCTs should strive to assess the integrity of interventions for each study arm and incorporate these data into their analyses, possibly using standardized data extraction forms. In this regard, a quantitative framework for evaluating the integrity of mHealth interventions could promote uniform assessment in RCTs. However, this is a complex task as there is no consensus on the model variables and their respective weights [69].

Conclusions

RCTs of mHealth interventions have specific methodological challenges compared to those of non-mHealth interventions, particularly those related to intervention integrity. Following our recommendations for addressing these challenges can lead to more reliable assessments of the effects of mHealth interventions on health outcomes.

Acknowledgments

The authors thank the survey respondents for their contribution to this research. During the preparation of this work, the authors used Microsoft Copilot and Grammarly to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. This work was supported by the National Institutes of Health (2R24AT001293, subaward 020468D). The funding source played no role in the design of this review, its execution, analysis and interpretation of the data, or the decision to submit the results for publication.

Authors' Contributions

CMW and LSW obtained funding. CMW, LSW, and JL-A defined the project's aims. CC-A and JL-A designed the searches and identified mobile health researchers. JL-A performed the narrative review to collect potential methodological challenges of mobile health randomized controlled trials. CMW, LSW, JB, and JL-A designed the survey. JL-A identified the potential survey respondents. YY analyzed the survey results. JL-A cross-checked the data analyses. CMW selected the workshop participants and moderated the workshop. AB, CL, CMW, EL, HR, JB, LA, MRK, PM, SS, and SY participated in the workshop as researchers. JL-A summarized the workshop discussions and integrated the survey results and the workshop recommendations. JL-A wrote the first version of this manuscript and incorporated the coauthors' feedback into the final version. All authors approved the final manuscript. CMW and LSW are the guarantors of this work.

Conflicts of Interest

CMW has active research grants for digital health projects awarded to the University of Zurich and Charité – Universitätsmedizin Berlin. These grants are funded by the German Health Care Innovation Fund, Krebsforschung Schweiz, Digitalisierungsinitiative der Zürcher Hochschulen, and Newsense Lab GmbH. Board positions related to digital health for mind and body (nonpaid) are as follows: codirector of the Digital Society Initiative of the University of Zurich. JB received honoraria for teaching activities about digital health interventions. EL owns a patent of an app and is the owner of a company (Hypnalgesics, LLC) as a requirement for having obtained grant support for its mobile health randomized controlled trial by the National Institutes of Health Small Business Innovation Research program through the National Center of Complementary and Integrative Health. All other authors do not have any conflict related to the manuscript's content.

Multimedia Appendix 1
References of documents on research methods consulted for methodological challenges.
[TXT File , 121 KB - jmir_v26i1e53187_app1.txt]

Multimedia Appendix 2
Invitation email and web-based survey.
[PDF File (Adobe PDF File), 632 KB - jmir_v26i1e53187_app2.pdf]

Multimedia Appendix 3
Identification of authors in Web of Science.
[DOCX File , 26 KB - jmir_v26i1e53187_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Workshop slides.
[PDF File (Adobe PDF File), 739 KB - jmir_v26i1e53187_app4.pdf]

Multimedia Appendix 5
Survey dataset.
[XLSX File (Microsoft Excel File), 557 KB - jmir_v26i1e53187_app5.xlsx]

Multimedia Appendix 6
Additional challenges.
[DOCX File , 41 KB - jmir_v26i1e53187_app6.docx]

Multimedia Appendix 7
Checklist for Reporting Results of Internet E-Surveys.
[DOCX File , 22 KB - jmir_v26i1e53187_app7.docx]

Multimedia Appendix 8
Checklist for Accurate Consensus Reporting Document guidelines.
[DOCX File , 36 KB - jmir_v26i1e53187_app8.docx]

References

1. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2024-04-29]
2. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://iris.who.int/handle/10665/44607 [accessed 2024-04-29]
3. Shaping Europe's digital future: mHealth. European Commission. 2020. URL: https://tinyurl.com/4jm9jhfk [accessed 2024-12-29]
4. Morse SS, Murugiah MK, Soh YC, Wong TW, Ming LC. Mobile health applications for pediatric care: review and comparison. Ther Innov Regul Sci 2018 May;52(3):383-391. [doi: 10.1177/2168479017725557] [Medline: 29714538]
5. Schofield P, Chambers S. Effective, clinically feasible and sustainable: key design features of psycho-educational and supportive care interventions to promote individualised self-management in cancer care. Acta Oncol 2015 May;54(5):805-812. [doi: 10.3109/0284186X.2015.1010016] [Medline: 25813474]
6. World health statistics 2023: monitoring health for the SDGs, sustainable development goals. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240074323 [accessed 2024-04-29]
7. Enricho Nkhoma D, Jenya Soko C, Joseph Banda K, Greenfield D, Li YJ, Iqbal U. Impact of DSMES app interventions on medication adherence in type 2 diabetes mellitus: systematic review and meta-analysis. BMJ Health Care Inform 2021 Apr;28(1):e100291 [FREE Full text] [doi: 10.1136/bmjhci-2020-100291] [Medline: 33853862]
8. Funnell EL, Spadaro B, Martin-Key N, Metcalfe T, Bahn S. mHealth solutions for mental health screening and diagnosis: a review of app user perspectives using sentiment and thematic analysis. Front Psychiatry 2022 Apr 27;13:857304 [FREE Full text] [doi: 10.3389/fpsyt.2022.857304] [Medline: 35573342]
9. Barth J, Wang J, Lopez-Alcalde J, Kramm C, Pach D, Álvarez-Díaz N, et al. Smartphone-RCCT: an online repository of randomized controlled clinical trials of smartphone applications for chronic conditions. Trials 2022 Oct 27;23(1):909 [FREE Full text] [doi: 10.1186/s13063-022-06849-x] [Medline: 36303168]
10. Aemissegger V, Lopez-Alcalde J, Witt CM, Barth J. Comparability of patients in trials of eHealth and face-to-face psychotherapeutic interventions for depression: meta-synthesis. J Med Internet Res 2022 Sep 14;24(9):e36978 [FREE Full text] [doi: 10.2196/36978] [Medline: 36103217]
11. Nolte E, McKee M. Caring for People with Chronic Conditions: A Health System Perspective: A Health System Perspective. New York, NY: Open University Press; 2008.
12. EU mHealth hub project - horizon 2020. World Health Organization. URL: https://tinyurl.com/44sha2p4 [accessed 2024-12-23]
13. Green paper on mobile Health ("mHealth"). European Commission. 2014. URL: https://tinyurl.com/t5zfbykh [accessed 2024-12-29]
14. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020;3:4 [FREE Full text] [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
15. Digital Health Trends 2021: innovation, evidence, regulation, and adoption. IQVIA. URL: https://tinyurl.com/f98ev8hk [accessed 2024-04-29]
16. mHealth apps market. Fortune Business Insights. URL: https://tinyurl.com/36yyp6y5 [accessed 2024-09-19]

XSL•FO
RenderX

17. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
18. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, International Stroke Trial Collaborative Group, European Carotid Surgery Trial Collaborative Group. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii-x, 1 [FREE Full text] [doi: 10.3310/hta7270] [Medline: 14499048]
19. Russell L, Pascoe MC, Seymour JF, Aranda S, Butow P, Gough K, et al. The trials and tribulations of conducting an m-health pilot randomized controlled trial to improve oral cancer therapy adherence: recommendations for future multisite, non-drug clinical trials. BMC Res Notes 2019 Apr 15;12(1):226 [FREE Full text] [doi: 10.1186/s13104-019-4264-6] [Medline: 30987685]
20. Kitsiou S, Paré G, Jaana M, Gerber B. Effectiveness of mHealth interventions for patients with diabetes: an overview of systematic reviews. PLoS One 2017;12(3):e0173160 [FREE Full text] [doi: 10.1371/journal.pone.0173160] [Medline: 28249025]
21. Lyons N, Cooper C, Lloyd-Evans B. A systematic review and meta-analysis of group peer support interventions for people experiencing mental health conditions. BMC Psychiatry 2021 Jun 23;21(1):315 [FREE Full text] [doi: 10.1186/s12888-021-03321-z] [Medline: 34162340]
22. Glasgow RE, Lichtenstein E, Marcus AC. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health 2003 Aug;93(8):1261-1267. [doi: 10.2105/ajph.93.8.1261] [Medline: 12893608]
23. Ginsburg LR, Hoben M, Easterbrook A, Anderson RA, Estabrooks CA, Norton PG. Fidelity is not easy! challenges and guidelines for assessing fidelity in complex interventions. Trials 2021 May 29;22(1):372 [FREE Full text] [doi: 10.1186/s13063-021-05322-5] [Medline: 34051830]
24. Tønning ML, Kessing LV, Bardram JE, Faurholt-Jepsen M. Methodological challenges in randomized controlled trials on smartphone-based treatment in psychiatry: systematic review. J Med Internet Res 2019 Oct 27;21(10):e15362 [FREE Full text] [doi: 10.2196/15362] [Medline: 31663859]
25. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
26. Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus reporting document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024 Jan;21(1):e1004326 [FREE Full text] [doi: 10.1371/journal.pmed.1004326] [Medline: 38261576]
27. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
28. Ahmed B, Dannhauser T, Philip N. A systematic review of reviews to identify key research opportunities within the field of eHealth implementation. J Telemed Telecare 2018 Apr 27;25(5):276-285 [FREE Full text] [doi: 10.1177/1357633x18768601]
29. Aji M, Gordon C, Stratton E, Calvo RA, Bartlett D, Grunstein R, et al. Framework for the design engineering and clinical implementation and evaluation of mHealth apps for sleep disturbance: systematic review. J Med Internet Res 2021 Feb 17;23(2):e24607 [FREE Full text] [doi: 10.2196/24607] [Medline: 33595441]
30. Alfawzan N, Christen M, Spitale G, Biller-Andorno N. Privacy, data sharing, and data security policies of women's mHealth apps: scoping review and content analysis. JMIR Mhealth Uhealth 2022 May 06;10(5):e33735 [FREE Full text] [doi: 10.2196/33735] [Medline: 35522465]
31. Arthurs N, Tully L, O'Malley G, Browne S. Usability and engagement testing of mHealth apps in paediatric obesity: a narrative review of current literature. Int J Environ Res Public Health 2022 Jan 27;19(3):22 [FREE Full text] [doi: 10.3390/ijerph19031453] [Medline: 35162470]
32. Asthana S, Jones R, Sheaff R. Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors. BMC Health Serv Res 2019 Dec 21;19(1):984 [FREE Full text] [doi: 10.1186/s12913-019-4790-x] [Medline: 31864370]
33. Bender JL, Yue RY, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. J Med Internet Res 2013 Dec 23;15(12):e287 [FREE Full text] [doi: 10.2196/jmir.2661] [Medline: 24366061]
34. THE digital health ecosystem: an in-depth examination of the players and tech trends reshaping the future of healthcare. Business Insider India. 2018. URL: https://tinyurl.com/2ppbzeyn [accessed 2024-04-29]
35. Cao W, Milks MW, Liu X, Gregory ME, Addison D, Zhang P, et al. mHealth interventions for self-management of hypertension: framework and systematic review on engagement, interactivity, and tailoring. JMIR Mhealth Uhealth 2022 Mar 02;10(3):e29415 [FREE Full text] [doi: 10.2196/29415] [Medline: 35234655]
36. Chib A, Lin SH. Theoretical advancements in mHealth: a systematic review of mobile apps. J Health Commun 2018;23(10-11):909-955 [FREE Full text] [doi: 10.1080/10810730.2018.1544676] [Medline: 30449261]

XSL•FO
RenderX

37. Davies AM, Mueller J. Developing Medical Apps and mHealth Interventions: A Guide for Researchers, Physicians and Informaticians. Cham, Switzerland: Springer; 2020.
38. Summary report on the public consultation on the green paper on mobile health. European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/summary-report-public-consultation-green-paper-mobile-health [accessed 2024-04-29]
39. Flemyng E, Moore TH, Boutron I, Higgins JP, Hróbjartsson A, Nejstgaard CH, et al. Using risk of Bias 2 to assess results from randomised controlled trials: guidance from Cochrane. BMJ Evid Based Med 2023 Aug 24;28(4):260-266. [doi: 10.1136/bmjebm-2022-112102] [Medline: 36693715]
40. Gold SM, Enck P, Hasselmann H, Friede T, Hegerl U, Mohr D, et al. Control conditions for randomised trials of behavioural interventions in psychiatry: a decision framework. Lancet Psychiatry 2017 Sep;4(9):725-732 [FREE Full text] [doi: 10.1016/S2215-0366(17)30153-0] [Medline: 28396067]
41. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14 [FREE Full text] [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
42. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
43. Hróbjartsson A, Boutron I, Turner L, Altman DG, Moher D, Cochrane Bias Methods Group. Assessing risk of bias in randomised clinical trials included in Cochrane reviews: the why is easy, the how is a challenge. Cochrane Database Syst Rev 2013 Apr 30;2013(4):ED000058 [FREE Full text] [doi: 10.1002/14651858.ED000058] [Medline: 23728703]
44. Igelström E, Campbell M, Craig P, Katikireddi SV. Cochrane's risk of bias tool for non-randomized studies (ROBINS-I) is frequently misapplied: a methodological systematic review. J Clin Epidemiol 2021 Dec;140:22-32 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.08.022] [Medline: 34437948]
45. Jokinen A, Stolt M, Suhonen R. Ethical issues related to eHealth: an integrative review. Nurs Ethics 2021 Mar;28(2):253-271 [FREE Full text] [doi: 10.1177/0969733020945765] [Medline: 32930041]
46. Jørgensen L, Paludan-Müller AS, Laursen DR, Savović J, Boutron I, Sterne JA, et al. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. Syst Rev 2016 May 10;5(1):80 [FREE Full text] [doi: 10.1186/s13643-016-0259-8] [Medline: 27160280]
47. Kulkarni P, Kirkham R, McNaney R. Opportunities for smartphone sensing in e-health research: a narrative review. Sensors (Basel) 2022 May 20;22(10):A [FREE Full text] [doi: 10.3390/s22103893] [Medline: 35632301]
48. Marcano Belisario JS, Huckvale K, Greenfield G, Car J, Gunn L. Smartphone and tablet self management apps for asthma. Cochrane Database Syst Rev 2013 Nov 27;2013(11):CD010013 [FREE Full text] [doi: 10.1002/14651858.CD010013.pub2] [Medline: 24282112]
49. Masterson Creber RM, Maurer MS, Reading M, Hiraldo G, Hickey KT, Iribarren S. Review and analysis of existing mobile phone apps to support heart failure symptom monitoring and self-care management using the mobile application rating scale (MARS). JMIR Mhealth Uhealth 2016 Jun 14;4(2):e74 [FREE Full text] [doi: 10.2196/mhealth.5882] [Medline: 27302310]
50. Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, et al. Evaluation of the risk of bias in non-randomized studies of interventions (ROBINS-I) and the 'target experiment' concept in studies of exposures: rationale and preliminary instrument development. Environ Int 2018 Nov;120:382-387 [FREE Full text] [doi: 10.1016/j.envint.2018.08.018] [Medline: 30125855]
51. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
52. Murray E, Khadjesari Z, White IR, Kalaitzaki E, Godfrey C, McCambridge J, et al. Methodological challenges in online trials. J Med Internet Res 2009 Apr 03;11(2):e9 [FREE Full text] [doi: 10.2196/jmir.1052] [Medline: 19403465]
53. Nahm ES, Bausell B, Resnick B, Covington B, Brennan PF, Mathews R, et al. Online research in older adults: lessons learned from conducting an online randomized controlled trial. Appl Nurs Res 2011 Nov;24(4):269-275 [FREE Full text] [doi: 10.1016/j.apnr.2009.09.004] [Medline: 20974077]
54. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence (NICE). URL: https://tinyurl.com/yc7kh2uh [accessed 2024-04-29]
55. Noura M, Rahmani A, Jahanfar S, Ellis UM. Mobile phone text messaging for the prevention of sexually transmitted infections. Cochrane Database Syst Rev 2019(10):CD013454 [FREE Full text] [doi: 10.1002/14651858.cd013454]
56. Perepletchikova F. On the topic of treatment integrity. Clin Psychol (New York) 2011 Jun;18(2):148-153 [FREE Full text] [doi: 10.1111/j.1468-2850.2011.01246.x] [Medline: 21769167]
57. Perepletchikova F, Hilt LM, Chereji E, Kazdin AE. Barriers to implementing treatment integrity procedures: survey of treatment outcome researchers. J Consult Clin Psychol 2009 Apr;77(2):212-218 [FREE Full text] [doi: 10.1037/a0015232] [Medline: 19309181]

XSL•FO
RenderX

58. Perepletchikova F, Treat TA, Kazdin AE. Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. J Consult Clin Psychol 2007 Dec;75(6):829-841. [doi: 10.1037/0022-006X.75.6.829] [Medline: 18085901]
59. Savović J, Weeks L, Sterne JA, Turner L, Altman DG, Moher D, et al. Evaluation of the Cochrane Collaboration's tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation. Syst Rev 2014 Apr 15;3:37 [FREE Full text] [doi: 10.1186/2046-4053-3-37] [Medline: 24731537]
60. Schünemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, GRADE Working Group. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol 2019 Jul;111:105-114 [FREE Full text] [doi: 10.1016/j.jclinepi.2018.01.012] [Medline: 29432858]
61. Varshney U, Singh N, Bourgeois AG, Dube SR. Review, Assess, Classify, and Evaluate (RACE): a framework for studying m-health apps and its application for opioid apps. J Am Med Inform Assoc 2022 Jan 29;29(3):520-535 [FREE Full text] [doi: 10.1093/jamia/ocab277] [Medline: 34939117]
62. Vis C, Bührmann L, Riper H, Ossebaard HC. Health technology assessment frameworks for eHealth: a systematic review. Int J Technol Assess Health Care 2020 Apr 16;36(3):204-216 [FREE Full text] [doi: 10.1017/s026646232000015x]
63. Wang Y, Min J, Khuri J, Xue H, Xie B, A Kaminsky L, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. JMIR Mhealth Uhealth 2020 Apr 28;8(4):e15400 [FREE Full text] [doi: 10.2196/15400] [Medline: 32343253]
64. Wolf JA, Moreau JF, Akilov O, Patton T, English 3rd JC, Ho J, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. JAMA Dermatol 2013 Apr;149(4):422-426 [FREE Full text] [doi: 10.1001/jamadermatol.2013.2382] [Medline: 23325302]
65. Yeaton WH, Sechrest L. Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. J Consult Clin Psychol 1981 Apr;49(2):156-167 [FREE Full text] [doi: 10.1037//0022-006x.49.2.156] [Medline: 7217482]
66. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/; [accessed 2024-04-29]
67. Kuckartz U. Qualitative text analysis: a systematic approach. In: Kaiser G, Presmeg N, editors. Compendium for Early Career Researchers in Mathematics Education. Cham, Switzerland: Springer; 2019:181-197.
68. Lopez-Alcalde J, Susan Wieland L, Barth J, Grainger R, Baxter N, Heron N, et al. Methodological challenges in systematic reviews of mHealth interventions: survey and consensus-based recommendations. Int J Med Inform 2024 Apr;184:105345. [doi: 10.1016/j.ijmedinf.2024.105345] [Medline: 38309237]
69. Lopez-Alcalde J, Yakoub N, Wolf M, Munder T, von Elm E, Flückiger C, et al. The RIPI-f (reporting integrity of psychological interventions delivered face-to-face) checklist was developed to guide reporting of treatment integrity in face-to-face psychological interventions. J Clin Epidemiol 2022 Nov;151:65-74 [FREE Full text] [doi: 10.1016/j.jclinepi.2022.07.013] [Medline: 35926822]
70. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
71. Thompson D, Mackay T, Matthews M, Edwards J, Peters NS, Connolly SB. Direct adherence measurement using an ingestible sensor compared with self-reporting in high-risk cardiovascular disease patients who knew they were being measured: a prospective intervention. JMIR Mhealth Uhealth 2017 Jun 12;5(6):e76 [FREE Full text] [doi: 10.2196/mhealth.6998] [Medline: 28606895]
72. Biomarker. Multiple Sclerosis Australia. URL: https://www.msaustralia.org.au/glossary/biomarker/ [accessed 2024-04-29]
73. Kleinman NJ, Shah A, Shah S, Phatak S, Viswanathan V. Improved medication adherence and frequency of blood glucose self-testing using an m-Health platform versus usual care in a multisite randomized clinical trial among people with type 2 diabetes in India. Telemed J E Health 2017 Sep;23(9):733-740. [doi: 10.1089/tmj.2016.0265] [Medline: 28328396]
74. Haldane V, Koh JJ, Srivastava A, Teo KW, Tan YG, Cheng RX, et al. User preferences and persona design for an mHealth intervention to support adherence to cardiovascular disease medication in Singapore: a multi-method study. JMIR Mhealth Uhealth 2019 May 28;7(5):e10465 [FREE Full text] [doi: 10.2196/10465] [Medline: 31140445]
75. Payne HE, Lister C, West JH, Bernhardt JM. Behavioral functionality of mobile apps in health interventions: a systematic review of the literature. JMIR Mhealth Uhealth 2015 Feb 26;3(1):e20. [doi: 10.2196/mhealth.3335] [Medline: 25803705]
76. Straw C, Sanchez-Antelo V, Kohler R, Paolino M, Viswanath K, Arrossi S. Implementation and scaling-up of an effective mHealth intervention to increase adherence to triage of HPV-positive women (ATICA study): perceptions of health decision-makers and health-care providers. BMC Health Serv Res 2023 Jan 18;23(1):47 [FREE Full text] [doi: 10.1186/s12913-023-09022-5] [Medline: 36653775]
77. Mao Y, Lin W, Wen J, Chen G. Impact and efficacy of mobile health intervention in the management of diabetes and hypertension: a systematic review and meta-analysis. BMJ Open Diabetes Res Care 2020 Sep;8(1):e001225 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001225] [Medline: 32988849]

XSL•FO
RenderX

78. Xiong S, Berkhouse H, Schooler M, Pu W, Sun A, Gong E, et al. Effectiveness of mHealth interventions in improving medication adherence among people with hypertension: a systematic review. Curr Hypertens Rep 2018 Aug 07;20(10):86. [doi: 10.1007/s11906-018-0886-7] [Medline: 30088110]
79. Li Y, Guo Y, Hong YA, Zeng Y, Monroe-Wise A, Zeng C, et al. Dose-response effects of patient engagement on health outcomes in an mHealth intervention: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2022 Jan 04;10(1):e25586 [FREE Full text] [doi: 10.2196/25586] [Medline: 34982724]
80. Karing C, Beelmann A. Evaluating the implementation and effectiveness of a low-dose mindfulness-based intervention in a student sample: a randomized controlled trial. Mindfulness 2021 Apr 01;12(6):1438-1450. [doi: 10.1007/S12671-021-01613-9]
81. Gandapur Y, Kianoush S, Kelli HM, Misra S, Urrea B, Blaha MJ, et al. The role of mHealth for improving medication adherence in patients with cardiovascular disease: a systematic review. Eur Heart J Qual Care Clin Outcomes 2016 Oct 01;2(4):237-244 [FREE Full text] [doi: 10.1093/ehjqcco/qcw018] [Medline: 29474713]
82. Murali KM, Mullan J, Chen JH, Roodenrys S, Lonergan M. Medication adherence in randomized controlled trials evaluating cardiovascular or mortality outcomes in dialysis patients: a systematic review. BMC Nephrol 2017 Jan 31;18(1):42 [FREE Full text] [doi: 10.1186/s12882-017-0449-1] [Medline: 28143438]
83. Mo Y, Lim C, Watson JA, White NJ, Cooper BS. Non-adherence in non-inferiority trials: pitfalls and recommendations. BMJ 2020 Jul 01;370:m2215 [FREE Full text] [doi: 10.1136/bmj.m2215] [Medline: 32611541]
84. Lee JA, Choi M, Lee SA, Jiang N. Effective behavioral intervention strategies using mobile health applications for chronic disease management: a systematic review. BMC Med Inform Decis Mak 2018 Feb 20;18(1):12 [FREE Full text] [doi: 10.1186/s12911-018-0591-0] [Medline: 29458358]
85. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
86. Arshed M, Mahmud AB, Minhat HS, Ying LP, Umer MF. Effectiveness of mHealth interventions in medication adherence among patients with cardiovascular diseases: a systematic review. Diseases 2023 Mar 01;11(1):25 [FREE Full text] [doi: 10.3390/diseases11010041] [Medline: 36975590]
87. Bond Z, Scanlon T, Judah G. Systematic review of RCTs assessing the effectiveness of mHealth interventions to improve statin medication adherence: using the behaviour-change technique taxonomy to identify the techniques that improve adherence. Healthcare (Basel) 2021 Sep 28;9(10):1282 [FREE Full text] [doi: 10.3390/healthcare9101282] [Medline: 34682962]
88. Higgins JT, Chandler J, Cumpston M, Li T, Page M, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5. Hoboken, NJ: Wiley-Blackwell; 2023.
89. Akl EA, Shawwa K, Kahale LA, Agoritsas T, Brignardello-Petersen R, Busse JW, et al. Reporting missing participant data in randomised trials: systematic survey of the methodological literature and a proposed guide. BMJ Open 2015 Dec 30;5(12):e008431 [FREE Full text] [doi: 10.1136/bmjopen-2015-008431] [Medline: 26719310]
90. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
91. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
92. Cro S, Morris TP, Kahan BC, Cornelius VR, Carpenter JR. A four-step strategy for handling missing outcome data in randomised trials affected by a pandemic. BMC Med Res Methodol 2020 Aug 12;20(1):208 [FREE Full text] [doi: 10.1186/s12874-020-01089-6] [Medline: 32787782]
93. Haggag O, Grundy J, Abdelrazek M, Haggag S. A large scale analysis of mHealth app user reviews. Empir Softw Eng 2022;27(7):196 [FREE Full text] [doi: 10.1007/s10664-022-10222-6] [Medline: 36246486]
94. Seiterö A, Thomas K, Löf M, Müssener U. Exploring the black box of an mHealth intervention (LIFE4YOUth): a qualitative process and outcome evaluation of end-user engagement. Int J Environ Res Public Health 2022 Oct 28;19(21):14022 [FREE Full text] [doi: 10.3390/ijerph192114022] [Medline: 36360903]
95. Barsky J, Hunter R, McAllister C, Yeates K, Campbell N, Liu P, et al. Analysis of the implementation, user perspectives, and feedback from a mobile health intervention for individuals living with hypertension (DREAM-GLOBAL): mixed methods study. JMIR Mhealth Uhealth 2019 Dec 09;7(12):e12639 [FREE Full text] [doi: 10.2196/12639] [Medline: 31815678]
96. Schroé H, Crombez G, De Bourdeaudhuij I, Van Dyck D. Investigating when, which, and why users stop using a digital health intervention to promote an active lifestyle: secondary analysis with a focus on health action process approach-based psychological determinants. JMIR Mhealth Uhealth 2022 Jan 31;10(1):e30583 [FREE Full text] [doi: 10.2196/30583] [Medline: 35099400]
97. Quanbeck A. Using stakeholder values to promote implementation of an evidence-based mobile health intervention for addiction treatment in primary care settings. JMIR Mhealth Uhealth 2019 Jun 07;7(6):e13301 [FREE Full text] [doi: 10.2196/13301] [Medline: 31237841]

XSL•FO
RenderX

98. Kohavi R, Tang D, Xu Y, Hemkens LG, Ioannidis JP. Online randomized controlled experiments at scale: lessons and extensions to medicine. Trials 2020 Feb 07;21(1):150 [FREE Full text] [doi: 10.1186/s13063-020-4084-y] [Medline: 32033614]
99. Ron K, Stefan T. The surprising power of online experiments. Harvard Business Review. 2017. URL: https://hbr.org/2017/09/the-surprising-power-of-online-experiments [accessed 2024-09-20]
100. Austrian J, Mendoza F, Szerencsy A, Fenelon L, Horwitz LI, Jones S, et al. Applying A/B testing to clinical decision support: rapid randomized controlled trials. J Med Internet Res 2021 Apr 09;23(4):e16651 [FREE Full text] [doi: 10.2196/16651] [Medline: 33835035]
101. Priebe JA, Toelle TR. Is there a right control condition in mHealth trials? A critical view on pain medicine. NPJ Digit Med 2019;2:107 [FREE Full text] [doi: 10.1038/s41746-019-0184-z] [Medline: 31701021]
102. Goldberg SS, Sun S, Carlbring P, Torous J. Selecting and describing control conditions in mobile health randomized controlled trials: a proposed typology. NPJ Digit Med 2023 Sep 30;6(1):181 [FREE Full text] [doi: 10.1038/s41746-023-00923-7] [Medline: 37775522]
103. Shiyab W, Ferguson C, Rolls K, Halcomb E. Solutions to address low response rates in online surveys. Eur J Cardiovasc Nurs 2023 May 25;22(4):441-444. [doi: 10.1093/eurjcn/zvad030] [Medline: 36827086]
104. Daikeler J, Bo'njak M, Lozar MK. Web versus other survey modes: an updated and extended meta-analysis comparing response rates. J Surv Stat Methodol 2019;8(3):513-539. [doi: 10.1093/jssam/smz008]
105. Triantafyllidis A, Kondylakis H, Votis K, Tzovaras D, Maglaveras N, Rahimi K. Features, outcomes, and challenges in mobile health interventions for patients living with chronic diseases: a review of systematic reviews. Int J Med Inform 2019 Dec;132:103984. [doi: 10.1016/j.ijmedinf.2019.103984] [Medline: 31605884]
106. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017 Dec 06;19(12):e402 [FREE Full text] [doi: 10.2196/jmir.8578] [Medline: 29212630]
107. Vitolins MZ, Rand CS, Rapp SR, Ribisl PM, Sevick MA. Measuring adherence to behavioral and medical interventions. Control Clin Trials 2000 Oct;21(5):S188-S194. [doi: 10.1016/S0197-2456(00)00077-5]
108. Dodd S, White IR, Williamson P. Nonadherence to treatment protocol in published randomised controlled trials: a review. Trials 2012 Jun 18;13(1):84 [FREE Full text] [doi: 10.1186/1745-6215-13-84] [Medline: 22709676]
109. Stephani V, Opoku D, Quentin W. A systematic review of randomized controlled trials of mHealth interventions against non-communicable diseases in developing countries. BMC Public Health 2016 Jul 15;16:572 [FREE Full text] [doi: 10.1186/s12889-016-3226-3] [Medline: 27417513]
110. Rhon DI, Fritz JM, Kerns RD, McGeary DD, Coleman BC, Farrokhi S, et al. TIDieR-telehealth: precision in reporting of telehealth interventions used in clinical trials - unique considerations for the Template for the Intervention Description and Replication (TIDieR) checklist. BMC Med Res Methodol 2022 Jun 02;22(1):161 [FREE Full text] [doi: 10.1186/s12874-022-01640-7] [Medline: 35655144]
111. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR Mhealth Uhealth 2021 Nov 19;9(11):e22890 [FREE Full text] [doi: 10.2196/22890] [Medline: 34806995]
112. Baker TB, Gustafson DH, Shaw B, Hawkins R, Pingree S, Roberts L, et al. Relevance of CONSORT reporting criteria for research on eHealth interventions. Patient Educ Couns 2010 Dec;81 Suppl:S77-S86 [FREE Full text] [doi: 10.1016/j.pec.2010.07.040] [Medline: 20843621]
113. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
114. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EQUATOR: Enhancing the Quality and Transparency of Health Research
mHealth: mobile health
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by N Cahill; submitted 29.09.23; peer-reviewed by Y Guo, D Liu; comments to author 26.01.24; revised version received 18.04.24; accepted 21.10.24; published 19.12.24.

Please cite as:
Lopez-Alcalde J, Wieland LS, Yan Y, Barth J, Khami MR, Shivalli S, Lokker C, Rai HK, Macharia P, Yun S, Lang E, Bwanika Naggirinya A, Campos-Asensio C, Ahmadian L, Witt CM
Methodological Challenges in Randomized Controlled Trials of mHealth Interventions: Cross-Sectional Survey Study and Consensus-Based Recommendations
J Med Internet Res 2024;26:e53187
URL: https://www.jmir.org/2024/1/e53187
doi:10.2196/53187
PMID:

©Jesus Lopez-Alcalde, L Susan Wieland, Yuqian Yan, Jürgen Barth, Mohammad Reza Khami, Siddharudha Shivalli, Cynthia Lokker, Harleen Kaur Rai, Paul Macharia, Sergi Yun, Elvira Lang, Agnes Bwanika Naggirinya, Concepción Campos-Asensio, Leila Ahmadian, Claudia M Witt. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

mHealth Engagement for Antiretroviral Medication Adherence Among People With HIV and Substance Use Disorders: Observational Study

Ranran Z Mi[1], PhD; Ellie Fan Yang[2], PhD; Alexander Tahk[3], PhD; Adati Tarfa[4], PharmD, PhD; Lynne M Cotter[5], MA; Linqi Lu[5], MA; Sijia Yang[5], PhD; David H Gustafson Sr[6], PhD; Ryan Westergaard[7], MPH, MD, PhD; Dhavan Shah[5], PhD

[1]Department of Communication, Media and Journalism, Kean University, Union, NJ, United States
[2]School of Communication, Illinois State University, Normal, IL, United States
[3]Department of Political Science, University of Wisconsin-Madison, Madison, WI, United States
[4]Yale School of Medicine, New Haven, CT, United States
[5]School of Journalism and Mass Communication, University of Wisconsin-Madison, Madison, WI, United States
[6]Industrial Engineering and Preventive Medicine, University of Wisconsin-Madison, Madison, WI, United States
[7]Department of Medicine, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, United States

Corresponding Author:
Ellie Fan Yang, PhD
School of Communication
Illinois State University
4480 School Of Communication
FEL Fell Hall 434
Normal, IL, 61790
United States
Phone: 1 (309) 438 3671
Email: fyang8@ilstu.edu

Abstract

Background: Despite the increasing popularity of mobile health (mHealth) technologies, little is known about which types of mHealth system engagement might affect the maintenance of antiretroviral therapy among people with HIV and substance use disorders.

Objective: This study aimed to use longitudinal and detailed system logs and weekly survey data to test a mediation model, where mHealth engagement indicators were treated as predictors, substance use and confidence in HIV management were treated as joint mediators, and antiretroviral therapy adherence was treated as the outcome. We further distinguished the initiation and intensity of system engagement by mode (expression vs reception) and by communication levels (intraindividual vs dyadic vs network).

Methods: Tailored for people with HIV living with substance use disorders, the mHealth app was distributed among 208 participants aged >18 years from 2 US health clinics. Supervised by medical professionals, participants received weekly surveys through the app to report their health status and medication adherence data. System use was passively collected through the app, operationalized as transformed click-level data, aggregated weekly, and connected to survey responses with a 7-day lagged window. Using the weekly check-in record provided by participants as the unit of analysis (N=681), linear regression and structure equation models with cluster-robust SEs were used for analyses, controlling within-person autocorrelation and group-level error correlations. Racial groups were examined as moderators in the structure equation models.

Results: We found that (1) intensity, not initiation, of system use; (2) dyadic message expression and reception; and (3) network expression positively predicted medication adherence through joint mediators (substance use and confidence in HIV management). However, intraindividual reception (ie, rereading saved entries for personal motivation) negatively predicts medication adherence through joint mediators. We also found Black participants have distinct usage patterns, suggesting the need to tailor mHealth interventions for this subgroup.

XSL•FO
RenderX

Conclusions: These findings highlight the importance of considering the intensity of system engagement, rather than initiation alone, when designing mHealth interventions for people with HIV and tailoring these systems to Black communities.

(J Med Internet Res 2024;26:e57774) doi:10.2196/57774

KEYWORDS

information and communication technologies; ICTs; mHealth; medication adherence; HIV care; antiretroviral therapy; substance use; social support; patient management; health disparities; information technology; communication technology; mobile health; app; clinic; United States; participants; mobile phone

Introduction

Background

Mobile health (mHealth) technology has emerged as a promising avenue in health care for addressing the complex challenges faced by people with HIV [1,2]. This population often grapples with their stigmatized condition while struggling with substance use disorders (SUDs) as an additional burden [3]. Substance use may hamper adherence to antiretroviral therapy (ART), which suppresses HIV viral load, thus reducing HIV transmission and preventing the health consequences of HIV infection [4]. As current human rights frameworks advocate that global HIV prevention needs to consider supporting individual treatment [5,6], this study focuses on understanding the potential benefits of mHealth technologies for people with HIV living with SUDs for maintaining ART medication adherence.

Most mHealth apps leverage the technological affordances of mobile phones to improve the health of users. The “anytime-anyplace” mobile access enables wide reach and improves how we coordinate and communicate [7-10]. In addition, mobile log data help identify evidence-based strategies for health intervention and message delivery [11-15]. Researchers have conducted studies that differentiate message expression from message reception across various communication levels and found that composing one-on-one messages is associated with greater perceived bonding, and that drafting public forum posts is linked to decreased risky drinking behaviors [16].

This study aims to understand how mHealth systems can influence adherence to ART, defined as consistent and proper compliance with the prescribed medication regimens for HIV viral load suppression. Numerous mobile technologies have been developed for interventions to enhance medication adherence among people with HIV [17]. For example, some focused on the effectiveness of electronic adherence monitoring (EAM) devices that remind patients to take pills [18]. Others adopted text message services to deliver motivational and skill-building content [19], launched social media campaigns that promote adherence and testing [20], and designed mobile games to make medication adherence enjoyable and goal-oriented [21]. However, there is a lack of understanding of how these interventions are beneficial for medication adherence. For instance, it is unclear how different communicative behaviors supported by the system would shape medication adherence and how within-person and group-level attributes should be considered when analyzing the system logs data.

To bridge this gap, we investigated both the initiation and intensity of mHealth engagement at the network (one-to-many), dyadic (one-to-one), and intraindividual (self-to-self) communication levels to predict ART medication adherence among people with HIV. In addition, we tested 2 sets of indirect pathways—different forms of substance use as risk factors and confidence in HIV management as a protective factor—that potentially mediated the relationships between mHealth engagement and ART medication adherence on a weekly basis. To examine these relationships, we used linear regression and structural equation models (SEMs) with cluster-robust SEs (CRSEs) to control for within-person autocorrelation and group-level error correlations. Finally, this paper explores whether the mediation model differs between Black and White populations and provides insights into designing and implementing mHealth interventions that promote medication adherence for different racial communities.

Medication Adherence and mHealth

Medication adherence has become a key factor in both individual health and public health efforts to control the spread of the virus. Traditional interventions have often fallen short in engaging and retaining people with HIV in care due to various barriers, such as stigma, limited access to health care facilities, and social isolation [22,23]. Interventions based on mHealth, including smartphone apps and text messaging systems, have shown promise in bridging these gaps. Studies have also shown that mobile apps are feasible and acceptable for use among people with HIV [24].

Although generally seen as providing social support, education, and monitoring, mHealth platforms can provide a wide range of distinct communication options: journal writing, one-to-one messaging, and group discussion posting [25,26]. Scholars have also theorized that distinct health outcomes are associated with message reception (ie, viewing existing content) and expression of messages (ie, sharing one’s own thoughts and feelings) at the intraindividual, dyadic, and network communication levels [16].

Message Reception and Expression at Each Communication Level

Message reception and expression likely function differently in the process of health behavior change and, more importantly, may trigger specific relational and psychological processes that occur when communicating, depending on the intended audience [16].

At the intraindividual level, expression occurs whenever individuals actively generate new, personal content (eg, creating

XSL•FO
RenderX

a journal entry), whereas reception refers to reflecting on self-generated content (eg, reading one's own previous writing). Pennebaker's [27] "expressive writing" research highlights the potential benefits of intraindividual expression, specifically through the "release effect" and cognitive processing. On the other hand, self-regulation [28,29] and self-perception [30] frameworks provided alternative accounts, emphasizing the potential benefits of intraindividual reception, which involves observing one's behavior and making necessary recalibration.

At the dyadic level, reception and expression occur when individuals receive and send, respectively, one-on-one messages. According to social penetration theory, dyadic expression is important because self-disclosure is the main driver for building and maintaining intimate relationships [31], and self-disclosure is most pronounced in dyads [32]. Dyadic reception also matters, according to optimal matching theory, as the help received and solicited in one-on-one messaging is more responsive and customized to specific individual needs [33].

At the network level, expression means crafting a message that can potentially be seen by many readers, whereas reception means viewing content available within such public forums. Network expression holds importance because, according to the concept of public commitment [34], expressing oneself to a group can compel individuals to adhere to their announced plans in order to maintain their reputation. Immersion in shared experiences and the sway of group norms provide potential explanations for why network reception might play a role [35]. By comparing the effects of system use at these distinct communication levels, as opposed to treating communication as a single entity, we can gain a more accurate understanding of the interactions within digital health interventions.

Patterns of human-app interactions can be understood in terms of the initiation and intensity of system use. Initiation indicates activation of system engagement on a given day, with regular initiation indicating consistency of use, while intensity reveals the depth of engagement on that day and the degree of penetration into system use. Studying human-technology interactions with attention to this distinction provides valuable insight into health intervention programs. For example, the level of engagement (ie, intensity), such as time spent per log-in, was more associated with intervention effects in comparison with the number of modules completed (ie, initiation) [36,37]. Numerically, mobile logs induce a significant volume of zero entries when users do not engage with the system, which must also be considered in the analysis. Distinguishing between these two metrics helps probe into subtle features of mHealth engagement: (1) determining whether an individual used a feature on a given day–which is binary, and (2) measuring the intensity of the engagement once initiated–which is continuous. This operationalization also echoes what digital media researchers refer to as the "session of use," meaning the extent of engagement users spend on a specific URL once they activate the interface [38]. Over time, initiation can be seen as a habitual repetition of engagement as the user consistently activates the feature. On the other hand, intensity represents the depth of use during the activated session. Although the initiation of system use is a prerequisite for an mHealth intervention, the intensity or depth of engagement, such as "lasting longer than 10 minutes", maybe just as critical in delivering the intervention [36,37].

We hypothesize that the initiation and intensity of message exchange, including reception and expression, at the 3 communication levels will predict ART medication adherence. As mHealth is designed to improve health, we hypothesize that engaging with the communication features will predict enhanced medication adherence.

- H1: The initiation of message reception at the (1) intraindividual, (2) dyadic, and (3) network level positively predicts ART medication adherence.
- H2: The intensity of message reception at the (1) intraindividual, (2) dyadic, and (3) network level positively predicts ART medication adherence.
- H3: The initiation of message expression at the (1) intraindividual, (2) dyadic, and (3) network level positively predicts ART medication adherence.
- H4: The intensity of message expression at the (1) intraindividual, (2) dyadic, and (3) network level positively predicts ART medication adherence.

SUDs as Risk Factors

Although mHealth interventions have shown significant potential in improving medication adherence, it is crucial to address potential risk factors, notably SUDs, which are linked to poor medication adherence. The misuse of a range of substances, including opioids, alcohol, and stimulants, poses substantial threats and could undermine the effectiveness of mHealth interventions [39,40].

SUDs can potentially mediate the relationship between mHealth engagement and medication adherence because mHealth engagement translates into a reduction of substance use, which eventually supports medication adherence. Past mHealth studies have examined whether mHealth interventions can potentially lead to a reduction in craving and substance use [41]. Such interventions can mitigate the challenges associated with SUDs through craving management, coping assistance, and tailored feedback and reminders. Reduction in substance use eventually improves medication adherence and health outcomes; a large body of evidence consistently shows that users of hard drug (eg, cocaine, heroin) and those with hazardous alcohol use report higher ART nonadherence as well as a greater chance of AIDS progression or death [4,42,43].

Thus, we hypothesize that the relationship between mHealth engagement and medication adherence will be mediated by patients' substance use, specifically their use of (1) heroin or opioids, (2) alcohol, and (3) stimulants. We also hypothesize that mHealth engagement predicts less substance use, which will, in turn, be associated with higher medication adherence.

- H5: The relationship between the engagement with mHealth and ART medication adherence will be mediated by substance use, including (1) heroin or opioids, (2) alcohol, and (3) stimulants.

Confidence in HIV Management as a Protective Factor

Confidence in HIV management serves as a protective factor that may encourage people with HIV to adhere to ART treatment

XSL•FO
RenderX

and potentially enhance the effectiveness of mHealth interventions. Confidence in HIV management denotes a patient's belief in the efficacy of their health care regimen, the trust they vest in their health care providers, and their self-care capability.

Confidence in one's HIV management can also potentially mediate the relationship between mHealth engagement and medication adherence. Research has shown patients who regularly interact with mHealth apps may develop a sense of self-efficacy and a stronger belief in the efficacy of their treatment. Likewise, confidence in HIV management may significantly motivate ART medication adherence. Research shows that individuals who have confidence in the care they are receiving are more likely to follow their prescribed medication regimens diligently [44,45]. Patients confident in their HIV management are more likely to overcome barriers and prioritize medication adherence, improving health outcomes.

As mHealth interactions enhance confidence in HIV management, patients may become more motivated and empowered to adhere to their medication regimens, ultimately improving their overall health and well-being. Thus, we hypothesize that mHealth engagement will positively predict confidence in HIV management, which will, in turn, be associated with higher medication adherence.

- H6: The relationship between engagement with the mHealth app and ART medication adherence will be mediated by confidence in HIV management.

Racial Disparities in Health

Racial disparities in health care have long been a concerning and pervasive issue with profound implications for health outcomes across various populations. Racial and ethnic minority groups such as Black and Hispanic populations often experience lower medication adherence rates, with an average of 7.5 percentage points lower than those of the White population [46]. The prevalence of comorbid conditions that intersect with HIV care, such as SUDs, is also highly prevalent in racial and ethnic minority groups [47] while facing unique challenges such as access to addiction treatment [48]. In addition, studies have indicated racial minority individuals had greater medical mistrust as compared with their White counterparts, which potentially hinders treatment discussions with their providers [49].

On the other hand, research has unveiled insights into the differences in mHealth use and preference among various racial groups. According to a recent national survey [50], Black adults exhibit a stronger inclination toward using mHealth for health decision-making, sharing health information, and engaging in discussions with their providers compared with non-Hispanic White adults. This is consistent with past reports indicating the Black US population has historically higher smartphone ownership (83%) than non-Hispanic White populations (74.2%) [51], suggesting the value of mHealth interventions for racial minorities.

Given the potential racial differences in mHealth use, substance use, confidence in HIV management, and health behaviors, we raise 3 research questions (RQs) below that are critical for developing targeted interventions that bridge racial disparities in medication adherence.

- RQ1: For Black participants, which types of mHealth engagement predict ART medication adherence, and through which mediators?
- RQ2: For White participants, which types of mHealth engagement predict ART medication adherence, and through which mediators?
- RQ3: How does the relationship between mHealth engagement and ART medication adherence differ between Black and White participants?

Methods

Clinical Setting and Recruitment

Recruitment took place at 2 US health clinics. Patients qualified for the study if they were adults aged >18 years, receiving medical care for HIV infection, and had a history of SUDs. This included those who either (1) actively used alcohol or other drugs within the past year or (2) were currently enrolled in an SUD treatment program. Recruitment methods included (1) posting flyers at two health clinics across three counties where the study took place, (2) project managers contacting social workers and case managers at these clinics to refer patients, and (3) asking physicians to review medical records of upcoming appointments and suggest ART-CHESS to eligible patients. A total of 208 participants were enrolled.

Data Collection

ART-CHESS was specifically designed for people with HIV living with SUDs and provided vulnerable patients with information resources, communication functions, and health support tools. It was distinguished from other apps by offering interactions in an unstigmatized environment that was also free of advertising and scammers [52].

The installation of the app on their smartphones was mandatory, through which they received weekly surveys capturing health status and medication adherence. Supervised by staff with experience in HIV care delivery settings, the app facilitated communication during critical moments to provide necessary support to participants.

Furthermore, 2 types of data were integrated by the user's unique ID. The first was system engagement data that were passively collected through the app usage logs. Note that while this paper only included system-use data from communication functionalities, the app provides additional noncommunication features. The second came from the baseline and weekly surveys, which asked about demographic information, substance use, confidence in HIV management, and ART medication adherence.

Measurement

mHealth Engagement

The "clicks," representing user interaction with app functions, served as a measure of participants' system engagement. Specifically, the "clicks" can capture every interface change when a user keystroked the mobile screen. It is a type of

XSL•FO
RenderX

behavioral measurement to track users' online navigation patterns and has been widely adopted to study user engagement [53,54]. In this study, the initiation and intensity metrics for 6 communication functions (Table 1) were assessed for the period before weekly surveys. For the initiation metric, "0" was coded when there was no click record for a user per day, while "1" was coded for any day with a click in the servers' record. As for the intensity metric, the number of clicks was aggregated among users who had engaged with particular system features (ie, initiation="1"). Notably, the intensity metric cannot be created when the initiation metric is coded as "0." After the 2 metrics were generated, they were transformed into their squared root format and lagged by 7 days before being incorporated into the modeling process. Table 1 shows more details regarding the operationalization of system engagement.

Table 1. Mobile health (mHealth) engagement across 3 communication levels on the studied intervention app for people with HIV and substance use disorders. Screenshots from the app interface are shown in Figures 3-6.

Engagement variables: communication level and message mode	Operationalization	App interface
Overall	Count of clicks per user per day on specific app page(s)	Home Page
Network		Discussion Forum
Message reception	App page indicates the user is viewing a message/comment posted on the discussion forum	
Message expression	App page indicates the user is creating, saving, or editing a message/comment posted on the discussion forum	
Dyadic		Private Messaging
Message reception	App page indicates the user is viewing a private one-on-one message thread	
Message expression	App page indicates the user is creating, saving, or editing a private one-on-one message	
Intraindividual		My Motivation
Message reception	App page indicates the user is viewing a self-written journal entry saved in the past	
Message expression	App page indicates the user is composing a new journal entry	

Figure 3. Home page of app interface.

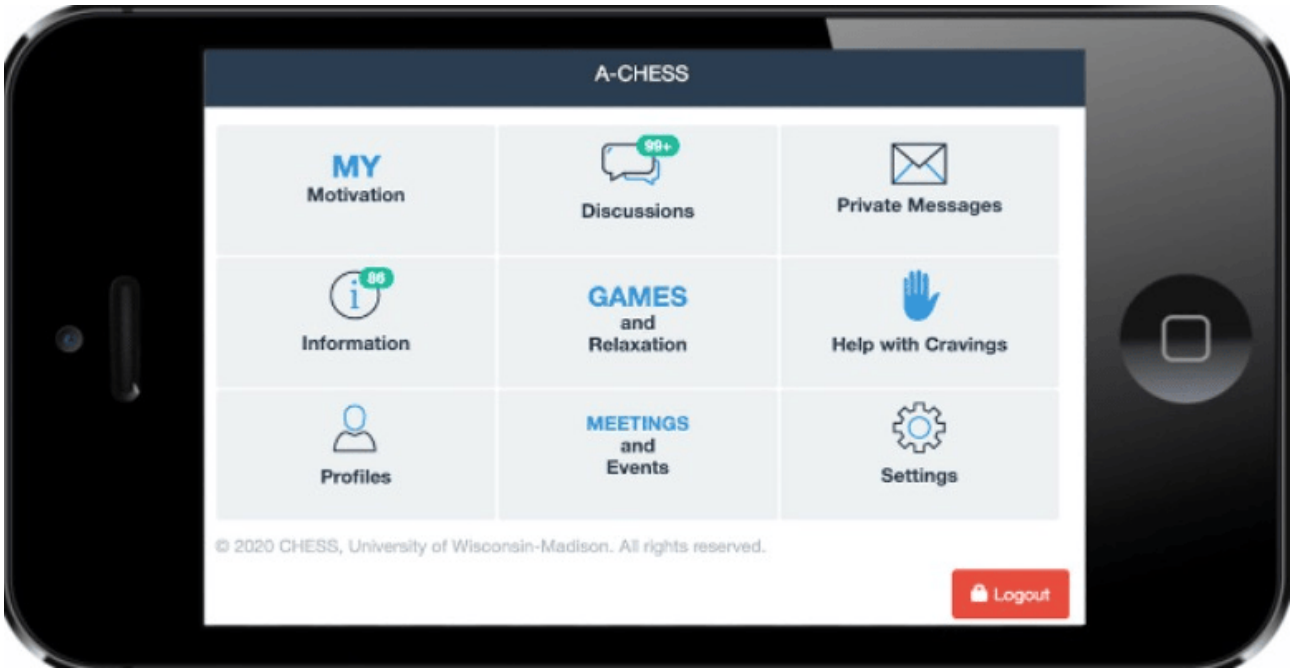

XSL•FO
RenderX

Figure 4. Discussion forum of app interface.

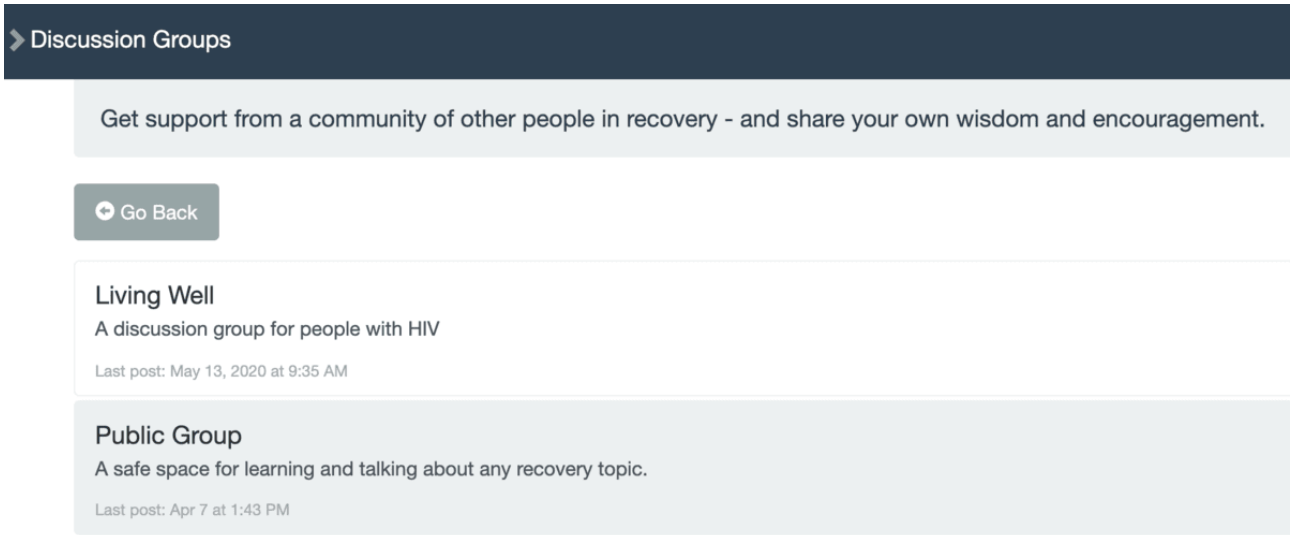

Figure 5. Private messaging on app interface.

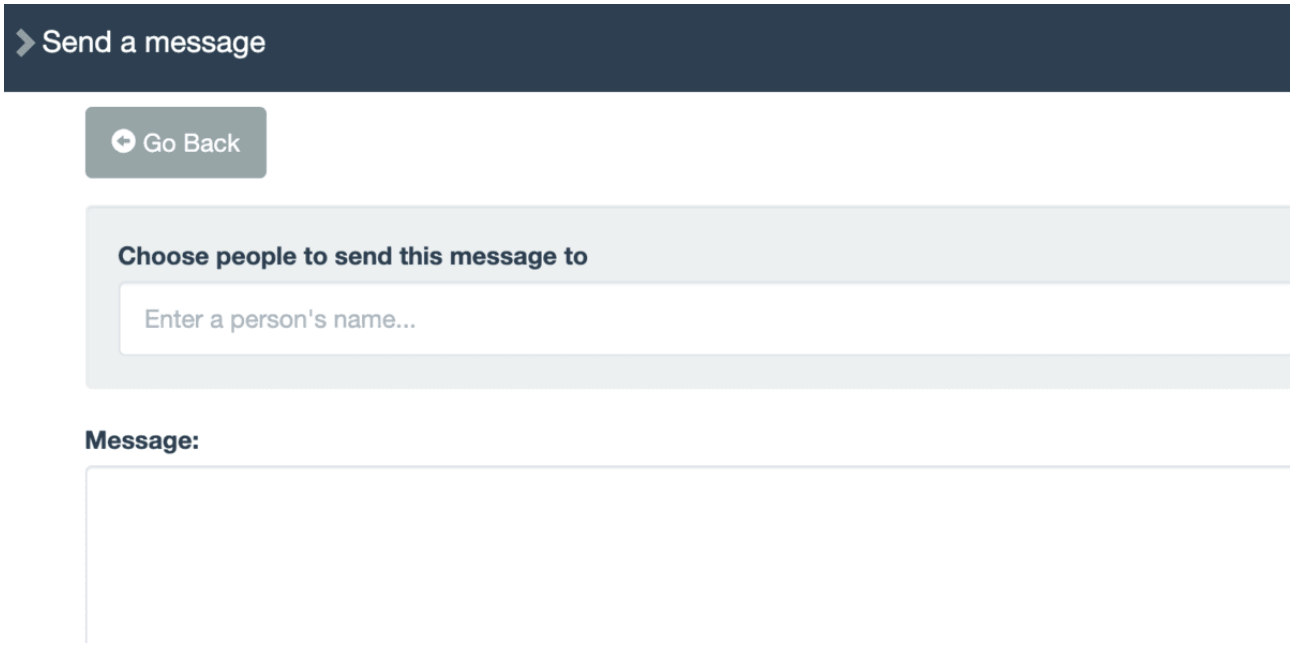

Figure 6. "My Motivation" section of app interface.

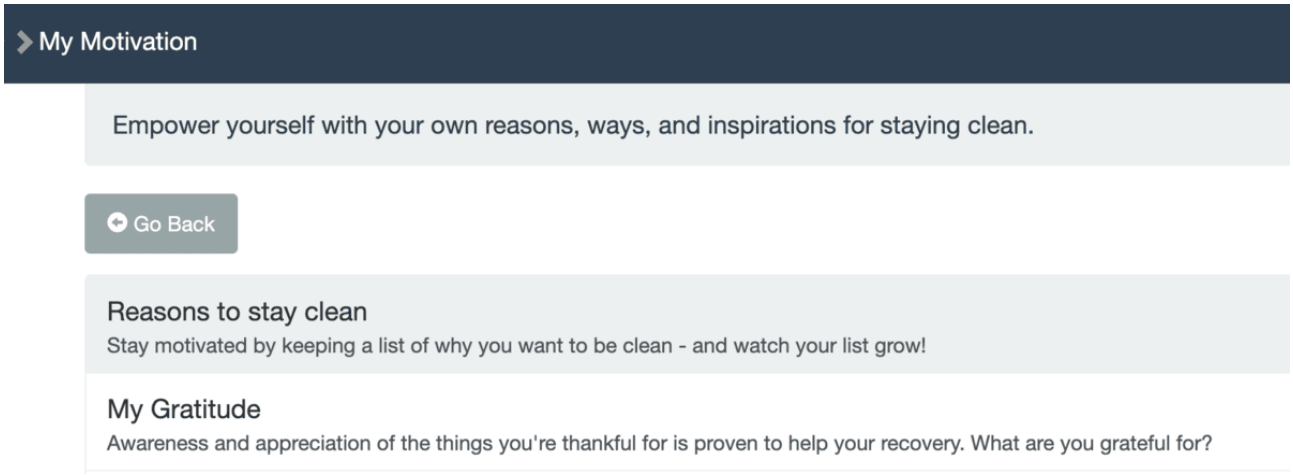

XSL•FO
RenderX

Medication Adherence

The app administered weekly surveys prompting participants to report the frequency of missed doses of HIV medications within the past week. Responses, ranging from 0-7, indicating the number of missed doses, were then reverse-coded with higher values indicating stronger adherence to the medication regimen.

Substance Use

The app administered weekly surveys presenting participants with 4 options (ie, heroin, opioid, alcohol, and stimulants) and asked, "Which of the following, if any, have you used in the last 7 days? (Check all that apply)." To capture whether users engaged with any substance use, checked options were encoded as binary variables. Heroin and opioids were combined as one category. Hence, 3 substance use variables were created to measure the recent use of heroin or opioids, alcohol, and stimulants.

Confidence in HIV Management

The app administered weekly surveys prompting participants to report their (1) confidence to keep the next appointment with the HIV care provider, (2) confidence to take all HIV medications during the next week, and (3) confidence in the belief that the HIV viral load "is undetectable right now." Responses on a 7-point Likert scale to the 3 items were averaged (Cronbach α=0.80).

Demographics

Gender, race, age, and years of education were acquired through a baseline survey.

Analytical Strategies

The unit of analysis was each response to the weekly survey provided by the 208 participants. The click-level data were aggregated into 7-day windows lagged 1 week before the survey input. We then transformed the aggregated clicks into the number of days with nonzero clicks (ie, initiation) and the sum of the square root of daily clicks (ie, intensity). As a result, when testing the hypotheses and RQs, the sample size was raised to 681 since the unit of analysis was not merely user based. Instead, the unit of analysis for the modeling process was the weekly check-in record per user (N=681).

The aggregated, transformed, and lagged system engagement variables were treated as predictors (see the *Measurement* section above for more details). System engagement predictors are deemed significant in the sense that they predict medication adherence through the joint mediators (opioid, alcohol, stimulant, and confidence in HIV). To control the temporal autocorrelation, we fit linear regression with CRSEs for H1-H4 (total) and structural equation modeling with CRSEs for H5-H6, which are typically used for controlling the error correlations at the group level. In our context, SEMs with CRSEs adjusted for the within-person temporal autocorrelation between observations for each app user. To answer RQ1-RQ3, a moderator indicating racial groups (Black or White participants) was added to the SEMs. For robustness check, we also specified an SEM in which system use is treated as a single latent component measured by the 12 system use variables. The results exhibited higher model fit and confirmed the association between app use and medication adherence.

Ethical Considerations

The study was approved by the Institutional Review Board at the University of Wisconsin-Madison (2016-1190) and funded by National Institutes of Health (grant DP2DA042424). Institutional Review Board–approved recruitment flyers were posted in the UW Health Infectious Disease Clinic, Johns Hopkins Moore Clinic, and the ALIVE study clinic during Phase 1 (2017-2018) and at the UW Health Infectious Disease Clinic during Phase 2 (2018-2020). Verbal consent using a brief script was obtained before administering the screening questionnaire. Eligible individuals were informed about the study details and asked to review the informed consent document, which described the risks and benefits of their participation. During the screening, participants who provided informed consent were given the option to install the study app on their mobile devices.

Participants created a username and password and provided the research team with the necessary information to build a new user account on the server. To protect participants' privacy, app use data were stored in a secure, encrypted database that was accessible only to the researchers. The information shared by the participants was anonymized before analysis. No monetary compensation was provided to participants. More details are available in the published protocol paper [52].

Results

A total of 173 participants used at least 1 communication feature (network, dyadic, or intraindividual) during the study period and provided weekly check-in data. The resulting cohort (N=173) had a mean age of 46 (SD 11.2) years and was 77.46% (n=134) male and 64.16% (n=111) Black. Overall, 42.2% (73/173) had at least some college-level education, and 60.12% (104/173) reported illegal or street drugs (including marijuana) or prescription medication abuse in the 30 days before joining this study. On average, participants used this mHealth app for a median of 13 (SD 23; range 1-141) days over the 6-month period. The zero-order correlations for all the numerical variables are presented in Table 2.

XSL•FO
RenderX

Table 2. Zero-order correlations for the numerical variables included in the structural equation model that examines how the use of mobile health (mHealth) predicts medication adherence (N=173).

Expression types	1	2	3	4	5	6	7	8	9
1. Medication adherence									
r	1	0.24	0.02	0.01	0.02	–0.04	0.01	–0.005	0.09
P value	—[a]	<.001	.69	.76	.58	.27	.73	.90	.02
2. Confidence in HIV management									
r	0.24	1	0.07	0.04	0.01	0.01	0.06	0.04	0.2
P value	<.001	—	.06	.28	.75	.73	.13	.25	<.001
3. Intensity of network reception									
r	0.02	0.07	1	0.7	0.43	0.3	0.37	0.2	–0.07
P value	.69	.06	—	<.001	<.001	<.001	<.001	<.001	<.001
4. Intensity of network expression									
r	0.01	0.04	0.70	1	0.36	0.20	0.25	0.11	–0.09
P value	.76	.28	<.001	—	<.001	<.001	<.001	<.001	<.001
5. Intensity of intraindividual reception									
r	0.02	0.01	0.43	0.36	1	0.62	0.41	0.51	0.04
P value	.58	.75	<.001	<.001	—	<.001	<.001	<.001	.08
6. Intensity of intraindividual expression									
r	–0.04	0.01	0.30	0.20	0.62	1	0.34	0.41	0.06
P value	.27	.73	<.001	<.001	<.001	—	<.001	<.001	.006
7. Intensity of dyadic reception									
r	0.01	0.06	0.37	0.25	0.41	0.34	1	0.54	–0.03
P value	.73	.13	<.001	<.001	<.001	<.001	—	<.001	.21
8. Intensity of dyadic expression									
r	–0.005	0.04	0.20	0.11	0.51	0.41	0.54	1	0.07
P value	.90	.25	<.001	<.001	<.001	<.001	<.001	—	<.001
9. Age									
r	0.09	0.20	–0.07	–0.09	0.04	0.06	–0.03	0.07	1
P value	.02	<.001	<.001	<.001	.08	.006	.21	<.001	—

[a]Not applicable.

Total Effects of System Use on Medication Adherence (H1-H4)

The total effects of the initiation and intensity of system use, either by message mode or by communication level (12 variables in total), were not significant in predicting medication adherence, except for the initiation of intraindividual reception (standardized β coefficient = 0.56, 95% CI 0.30-0.82; *P*<.001). Only part of H1 was supported. Nonetheless, the linear regression model suggested that the intensity metrics (eg, network reception, network expression, and dyadic reception) showed a more consistently positive association with medication adherence than initiation metrics. The divergent directions between predictors and the outcome led to the subsequent mediation tests to unpack these seemingly countervailing effects. For this reason, we use SEMs to decompose the mediating effects of the proposed theoretical mechanism [55,56]. Coefficient estimates for all system use variables in H1-H4 are reported in Multimedia Appendix 1.

Indirect Effects Through Substance Use and Confidence in HIV Management (H5-H6)

The mediation model based on the SEM showed an adequate model fit ($^2_6$=19.70, *P*=.003; comparative fit index=0.824; and standardized root mean squared residual=0.018). Substance use and confidence in HIV management jointly mediated the relationship between 4 types of system use and medication adherence. Through these joint mediators, 3 system use predictors indirectly predicted improved medication adherence: the intensity of dyadic reception (standardized β coefficient=0.036, 95% CI 0.01-0.06; *P*=.01), the intensity of dyadic expression (standardized β coefficient=0.067, 95% CI 0.01-0.13; *P*=.03), and the intensity of network expression (standardized β coefficient=0.039, 95% CI 0.01-0.07; *P*=.02).

XSL•FO
RenderX

Notably, the intensity of intraindividual reception predicted worsened medication adherence through the joint mediators (standardized β coefficient=–0.043, 95% CI –0.086 to –0.001; *P*=.05).

Figure 1 demonstrates the estimated indirect pathways between significant system engagement predictors and medication adherence through the 4 joint mediators by decomposing the indirect effects into the constituting a-paths (linking engagement metrics to mediators) and b-paths (linking mediators to medication adherence outcomes). For clarity, only statistically significant predictors are presented in the diagram. The 4 listed predictors are significant in that they predict medication adherence via the joint mediators. This presentation rule also applies to Figure 2. All the mediators were allowed to covary with each other. Full estimates of the coefficients of both the joint and individual mediators are reported in Multimedia Appendix 2.

Figure 1. Mediation model showing mobile health (mHealth) engagement predicting medication adherence by substance use and confidence in HIV management (all participants).

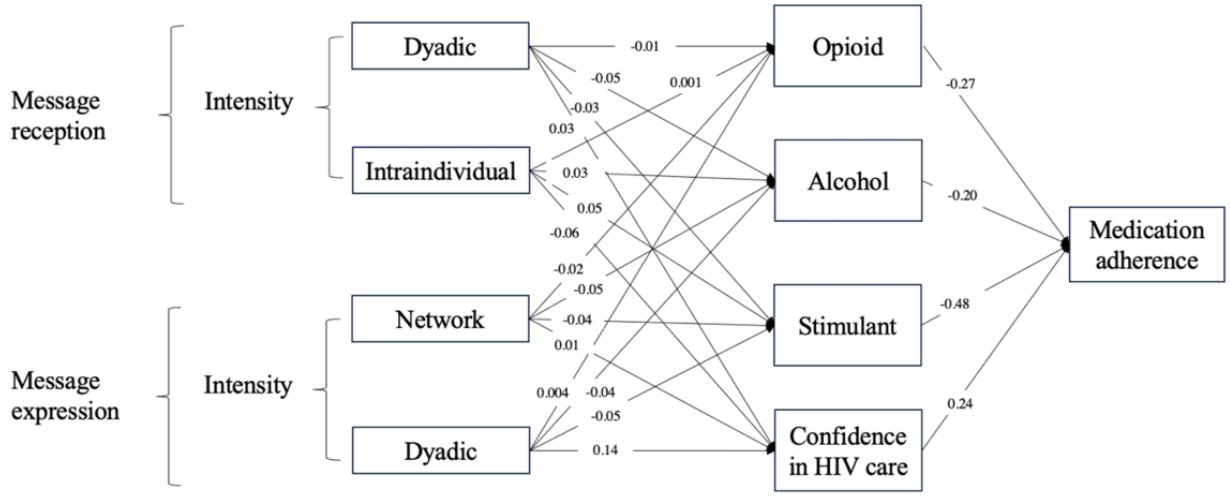

Figure 2. Mediation model showing mobile health (mHealth) engagement predicting medication adherence by substance use and confidence in HIV management (Black participants only).

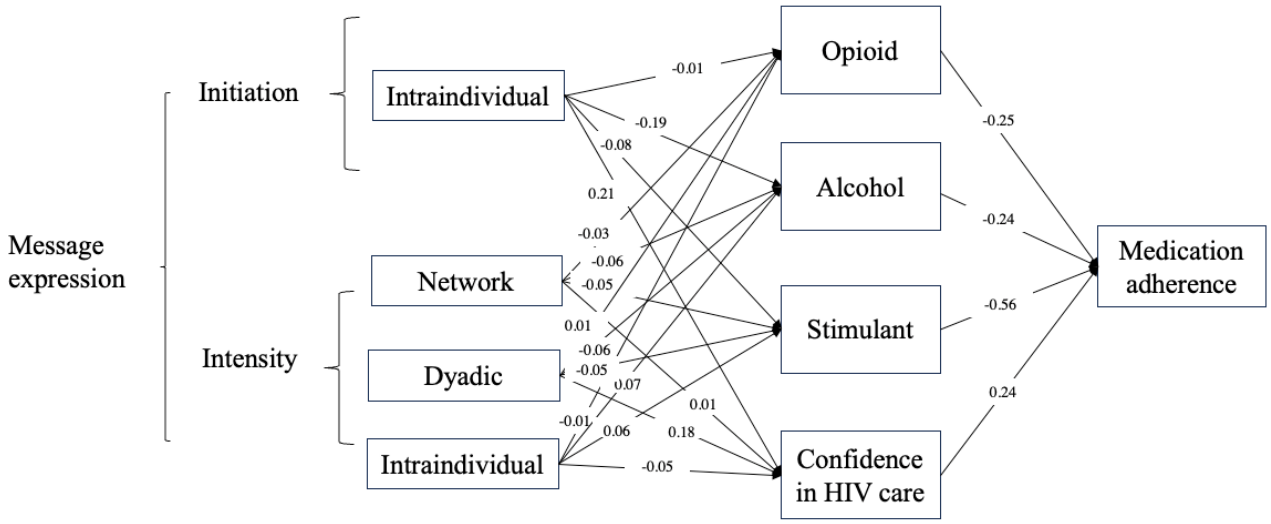

Racial Differences (RQ1-RQ3)

Race was entered as a binary moderator (ie, whether the users identified themselves as Black) to the SEM. The moderated mediation SEM showed an adequate model fit ($^2_6$=17.22, *P*=.009; comparative fit index=0.929; and standardized root mean squared residual=0.011). Subgroup-specific estimates were then extracted from this moderated mediation model.

For Black users (RQ1), results suggest that substance use and confidence in HIV management also jointly mediated the relationship between system engagement and medication adherence (more details are shown in Figure 2 for the diagram and Multimedia Appendix 3 for full estimates). Specifically,

XSL•FO
RenderX

through these joint mediators, 3 system use predictors indirectly predicted improved medication adherence: the initiation of intraindividual expression (standardized β coefficient=0.14, 95% CI 0.02=0.27; *P*=.02), the intensity of network expression (standardized β coefficient=0.05, 95% CI 0.01-0.09; *P*=.007), and the intensity of dyadic expression (standardized β coefficient=0.08, 95% CI 0.02-0.15; *P*=.01). Notably, the intensity of intraindividual expression predicted worsened medication adherence through the joint mediators (standardized β coefficient=–0.06, 95% CI –0.11 to –0.01; *P*=.01).

For White participants (RQ2), we lack the evidence that any of the systems use predictors significantly predict medication adherence through joint mediators (more details in Multimedia Appendix 4).

Statistically contrasting Black and White participants, the relationship between system engagement and medication adherence yielded distinct patterns (RQ3). Table 3 reports the difference coefficients for Black versus White participants. Specifically, both the initiation (coefficient=0.22, 95% CI 0.04-0.40; *P*=.02) and intensity (coefficient=–0.07, 95% CI –0.13 to –0.01; *P*=.03) of intraindividual expression differed in their relationship with medication adherence for Black and White participants.

Table 3. Difference coefficients estimated for Black and White participants in their mediated relationship between system engagement and medication adherence.

Predictors	Difference coefficient (95% CI), White vs Black	*P* value
Network reception initiation	0.06 (–0.12 to 0.25)	.51
Network reception intensity	–0.02 (–0.05 to 0.0001)	.051
Network expression initiation	0.01 (–0.32 to 0.35)	.93
Network expression intensity	0.03 (–0.07 to 0.13)	.57
Dyadic reception initiation	0.16 (–0.07 to 0.40)	.17
Dyadic reception intensity	–0.03 (–0.08 to 0.02)	.25
Dyadic expression initiation	–0.05 (–0.36 to 0.26)	.75
Dyadic expression intensity	0.08 (–0.06 to 0.22)	.25
Intraindividual reception initiation	0.41 (–0.01 to 0.83)	.06
Intraindividual reception intensity	–0.05 (–0.20 to 0.11)	.56
Intraindividual expression initiation	0.22 (0.04-0.40)	.02
Intraindividual expression intensity	–0.07 (–0.13 to –0.01)	.03

Discussion

Principal Findings

Emerging mHealth systems hold promise to enhance medication adherence among people with HIV and SUDs. Despite its potential, previous research has yet to fully explain how mobile technologies relate to medication adherence for this underserved population. Our study examines various aspects of mHealth engagement and their potential impact on medication adherence among people with HIV, overall and by racial groups.

Our findings reveal that most communicative engagement with mHealth does not directly predict medication adherence in people with HIV. Instead, certain metrics of mHealth engagement indirectly predicted medication adherence through substance use and confidence in HIV management as joint mediators. Specifically, at the dyadic level, we observed that greater intensity of message reception and expression (ie, sending and receiving one-on-one messages) were linked to reduced alcohol and stimulant use, as well as increased confidence in HIV management, all of which, in turn, were associated with greater medication adherence, all measured on a weekly basis. That said, the intensity of dyadic expression was associated with greater opioid use, which predicted worsened medication adherence. At the network level, greater intensity of message expression (ie, posting on the discussion forum) was related to decreased opioid, alcohol, and stimulant use, as well as increased confidence in management, all of which predicted improved medication adherence. In contrast, at the intraindividual level, a higher intensity of message reception (ie, reviewing past saved journal entries) was associated with elevated opioid, alcohol, and stimulant use, as well as decreased confidence in management, all of which signaled reduced medication adherence.

Comparison With Previous Work

Our study contributes to the growing body of mHealth research by disentangling system use along several significant dimensions, including message expression versus reception, communication levels, and initiation versus intensity, as noted in previous literature [14,16]. The findings regarding the overall participants highlight the critical role of message intensity, rather than initiation, in shaping mHealth engagement and its impact on medication adherence. This underscores the importance of considering not only habitual use but also the depth of engagement for people with HIV and SUDs to improve ART medication adherence.

Furthermore, our study underscores the need to account for all 3 communication levels in health interventions, especially when juxtaposing each level with message expression versus reception

XSL•FO
RenderX

in both design and analysis. For example, our results highlight that network expression (ie, posting on the discussion forum) was associated with improved health outcomes, while network reception (ie, browsing the discussion forum) may not yield the same benefit. This finding aligns with previous studies, which have indicated that individuals who actively engage in expressing themselves within online communities, often referred to as "expressors," tend to perceive more benefits [57,58]. In addition, providing emotional support in such interactions can also be beneficial for the message senders [59,60], enhancing their skills in developing coping strategies [61]. Another critical implication of our research is that patients who focus on intraindividual communication may signal potential challenges in SUDs and, eventually, in their medication adherence. This insight serves as a valuable indicator for health care providers to identify individuals who may require additional support and intervention to ensure the success of their medication regimens.

This study also contributes to our understanding of racial differences in their engagement with and outcomes of mHealth technologies. We observed notable distinctions between Black and White participants in how mHealth system engagement may (or may not) be associated with their medication adherence despite the fact that they were using the same mHealth app. Among Black participants, our results highlight the importance of message expression (Figure 2) as message expression at different communication levels, constituting a total of 12 distinct pathways, emerged as a set of critical predictors. These expression predictors, rather than any reception predictors, connect with both risk and protective factors, which, in turn, play a pivotal role in predicting medication adherence. In addition, the 2 racial groups also differ in the role played by intraindividual expression on medication adherence (Table 3). These disparities underscore the need for a nuanced, race-sensitive approach to medical interventions and health care support for people with HIV. Tailoring strategies based on these findings may help address racial disparities in health outcomes.

The third contribution of this study is methodological, where we introduced 2 engagement metrics based on longitudinal and timestamped system log data: daily initiation and intensity of system use within each active session. We tested the predictive power of these metrics using linear regression and SEMs with CRSEs to control for within-person autocorrelation and group-level error correlations. This approach can guide health practitioners in tailoring interventions for those showing less intensive use, revealing which app features supported ART medication adherence. Another avenue to explore is tracking the evolution of these 2 engagement types over time. By examining the fluctuations in mHealth engagement, we can extract crucial insights about the dynamic characteristics of these interventions. Attention to these dynamics may help further tailor the design of mHealth interventions to serve the evolving needs of people with HIV living with SUD. These insights might also inform the roles of health care professionals in promoting medication adherence among this population.

Practical Implications

As the adoption of mHealth interventions for improving medication adherence among people with HIV and SUDs increases, the findings from this study may provide insights into how to enhance their impact. Clinicians should encourage patients to participate in one-on-one messaging (dyadic communication) and actively post on discussion forums (network expression). In addition, monitoring system use patterns such as frequent reviewing of past journal entries and sending one-on-one messages may serve to identify patients in need of additional support. The findings from this study also indicate it is important to tailor mHealth interventions for specific subgroups. For instance, Black participants who showed distinct usage patterns may be more influenced by their expression across different communication levels.

Finally, our findings provide useful guidance for app developers. App developers should prioritize features that promote expressive behaviors in one-on-one messaging and discussion forums, such as reminders and notifications. Furthermore, developers should implement a monitoring and alert system to detect frequent reviewing of past journal entries, enabling timely connection of patients with health care providers for necessary support.

Limitations

We acknowledge this study has several limitations. First, this research did not consider the content of messages at the network, dyadic, or intraindividual level, ignoring the nature of the communication that may play a buffering role. Future research should examine message content to provide a deeper understanding of why, for instance, message reception at the intraindividual level negatively predicted medication adherence through the tested mediators. Second, some participants did not consistently complete the weekly check-in surveys, leading to missing data.

However, given that our unit of analysis is the user weekly survey, our analysis does not require that the surveys be completed at regular intervals. We model each user survey and allow for arbitrary correlation between surveys by the same user through robust SEs clustered on the user. Due to this approach, when users did not take a survey on a given week and failed to provide the dependent variable and all mediators, this missed survey is not considered missing data in the analysis in a practical sense. Future research may investigate the possibilities of data imputation techniques for simulation purposes [62]. It is also important to note that our sample is relatively small and does not represent the diverse population of people with HIV living with SUDs. Those who were willing to participate in our study were likely to have greater motivations to maintain medication adherence, potentially biasing results. Future research should aim to include a more diverse and larger sample to enhance generalizability.

Conclusions

To summarize, this study articulates the different aspects of mHealth engagement and emphasizes the importance of the intensity of engagement over initiation for promoting medication adherence in people with HIV and SUDs. The findings highlight the need for nuanced patient engagement strategies and targeted interventions due to the unique usage patterns and associations

XSL•FO
RenderX

with ART medication adherence of Black and White participants.

Data Availability

The datasets generated during this study are not publicly available due to the sensitive nature of the peer-to-peer communication data in recovery support forums and the fact that the participants did not consent to the broad sharing of the research data. However, numerical data underlying the regression analyses are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Effect estimates of system use on medication adherence (H1-H4).
[DOCX File , 16 KB - jmir_v26i1e57774_app1.docx]

Multimedia Appendix 2
Joint and separate indirect effects of all mediators connecting system engagement predictors and medication adherence (all participants).
[DOCX File , 21 KB - jmir_v26i1e57774_app2.docx]

Multimedia Appendix 3
Joint and separate indirect effects of all mediators connecting system engagement predictors and medication adherence (Black participants).
[DOCX File , 20 KB - jmir_v26i1e57774_app3.docx]

Multimedia Appendix 4
Joint and separate indirect effects of all mediators connecting system engagement predictors and medication adherence (White participants).
[DOCX File , 20 KB - jmir_v26i1e57774_app4.docx]

References

1. Schnall R, Bakken S, Rojas M, Travers J, Carballo-Dieguez A. mHealth technology as a persuasive tool for treatment, care and management of persons living with HIV. AIDS Behav 2015;19 Suppl 2(0 2):81-89 [FREE Full text] [doi: 10.1007/s10461-014-0984-8] [Medline: 25572830]
2. Cooper V, Clatworthy J, Whetham J, Consortium E. mHealth interventions to support self-management in HIV: a systematic review. Open AIDS J 2017;11:119-132 [FREE Full text] [doi: 10.2174/1874613601711010119] [Medline: 29290888]
3. Elkbuli A, Polcz V, Dowd B, McKenney M, Prado G. HIV prevention intervention for substance users: a review of the literature. Subst Abuse Treat Prev Policy 2019;14(1):1 [FREE Full text] [doi: 10.1186/s13011-018-0189-7] [Medline: 30606266]
4. Cohn SE, Jiang H, McCutchan JA, Koletar SL, Murphy RL, Robertson KR, et al. Association of ongoing drug and alcohol use with non-adherence to antiretroviral therapy and higher risk of AIDS and death: results from ACTG 362. AIDS Care 2011;23(6):775-785 [FREE Full text] [doi: 10.1080/09540121.2010.525617] [Medline: 21293986]
5. Meier BM, Brugh KN, Halima Y. Conceptualizing a human right to prevention in global HIV/AIDS policy. Public Health Ethics 2012;5(3):263-282 [FREE Full text] [doi: 10.1093/phe/phs034] [Medline: 23226723]
6. Huang YM, Shiyanbola OO, Smith PD. Association of health literacy and medication self-efficacy with medication adherence and diabetes control. Patient Prefer Adherence 2018;12:793-802 [FREE Full text] [doi: 10.2147/PPA.S153312] [Medline: 29785094]
7. Ling R. Taken for Grantedness: The Embedding of Mobile Communication Into Society. Cambridge, MA: MIT Press; 2012.
8. Campbell SW. Cutting the cord: social and scholarly revolutions as CMC goes mobile. Journal of Computer-Mediated Communication 2020;25(1):101-110 [FREE Full text] [doi: 10.1093/jcmc/zmz021]
9. Vanden Abeele M, de Wolf R, Ling R. Mobile media and social space: how anytime, anyplace connectivity structures everyday life. MaC 2018;6(2):5-14 [FREE Full text] [doi: 10.17645/mac.v6i2.1399]
10. Ling R, Oppegaard B. THIS IS NOT A DRILL: mobile telephony, information verification, and expressive communication during hawaii's false missile alert. Social Media & Society 2021;7(1):834-856 [FREE Full text] [doi: 10.1177/2056305121999661]

XSL•FO
RenderX

11. Boase J, Ling R. Measuring mobile phone use: self-report versus log data. J Comput-Mediat Comm 2013;18(4):508-519 [FREE Full text] [doi: 10.1111/jcc4.12021]
12. Boase J. Augmenting survey and experimental designs with digital trace data. Communication Methods and Measures 2016;10(2-3):165-166 [FREE Full text] [doi: 10.1080/19312458.2016.1150975]
13. Ross MQ, Bayer J, Rhee L, Potti I, Chang YJ. Tracking the temporal flows of mobile communication in daily life. New Media & Society 2023;25(4):732-755 [FREE Full text] [doi: 10.1177/14614448231158646]
14. Yang F, Shah DV, Tahk A, Vjorn O, Dietz S, Pe-Romashko K, et al. mHealth and social mediation: mobile support among stigmatized people living with HIV and substance use disorder. New Media & Society 2023;25(4):702-731 [FREE Full text] [doi: 10.1177/14614448231158653]
15. Wolfers LN, Utz S, Wendt R, Honecker J. Conditionally helpful? the influence of person-, situation-, and device-specific factors on maternal smartphone use for stress coping and on coping effectiveness. Cyberpsychology 2023;17(3) [FREE Full text] [doi: 10.5817/cp2023-3-1]
16. Mi RZ, Kornfield R, Shah DV, Maus A, Gustafson DH. Intraindividual, dyadic, and network communication in a digital health intervention: distinguishing message exposure from message production. Health Commun 2022;37(4):397-408 [FREE Full text] [doi: 10.1080/10410236.2020.1846273] [Medline: 33238732]
17. Whiteley LB, Olsen EM, Haubrick KK, Odoom E, Tarantino N, Brown LK. A review of interventions to enhance HIV medication adherence. Curr HIV/AIDS Rep 2021;18(5):443-457. [doi: 10.1007/s11904-021-00568-9] [Medline: 34152554]
18. Demonceau J, Ruppar T, Kristanto P, Hughes DA, Fargher E, Kardas P, et al. Identification and assessment of adherence-enhancing interventions in studies assessing medication adherence through electronically compiled drug dosing histories: a systematic literature review and meta-analysis. Drugs 2013;73(6):545-562 [FREE Full text] [doi: 10.1007/s40265-013-0041-3] [Medline: 23588595]
19. Mayer JE, Fontelo P. Meta-analysis on the effect of text message reminders for HIV-related compliance. AIDS Care 2017;29(4):409-417 [FREE Full text] [doi: 10.1080/09540121.2016.1214674] [Medline: 27477580]
20. Medich M, Swendeman DT, Comulada WS, Kao UH, Myers JJ, Brooks RA, et al. Promising approaches for engaging youth and young adults living with HIV in HIV primary care using social media and mobile technology interventions: protocol for the SPNS social media initiative. JMIR Res Protoc 2019;8(1):e10681 [FREE Full text] [doi: 10.2196/10681] [Medline: 30702434]
21. Hightow-Weidman LB, Muessig KE, Bauermeister JA, LeGrand S, Fiellin LE. The future of digital games for HIV prevention and care. Curr Opin HIV AIDS 2017;12(5):501-507 [FREE Full text] [doi: 10.1097/COH.0000000000000399] [Medline: 28692490]
22. Levi-Minzi MA, Surratt HL. HIV stigma among substance abusing people living with HIV/AIDS: implications for HIV treatment. AIDS Patient Care STDS 2014;28(8):442-451 [FREE Full text] [doi: 10.1089/apc.2014.0076] [Medline: 24983302]
23. Marziali ME, McLinden T, Card KG, Closson K, Wang L, Trigg J, et al. Social isolation and mortality among people living with HIV in British Columbia, Canada. AIDS Behav 2021;25(2):377-388 [FREE Full text] [doi: 10.1007/s10461-020-03000-2] [Medline: 32797358]
24. Westergaard RP, Genz A, Panico K, Surkan PJ, Keruly J, Hutton HE, et al. Acceptability of a mobile health intervention to enhance HIV care coordination for patients with substance use disorders. Addict Sci Clin Pract 2017;12(1):11 [FREE Full text] [doi: 10.1186/s13722-017-0076-y] [Medline: 28441962]
25. Chuang KY, Yang CC. Informational support exchanges using different computer - mediated communication formats in a social media alcoholism community. Asso for Info Science & Tech 2013;65(1):37-52 [FREE Full text] [doi: 10.1002/asi.22960]
26. Han JY, Shah DV, Kim E, Namkoong K, Lee SY, Moon TJ, et al. Empathic exchanges in online cancer support groups: distinguishing message expression and reception effects. Health Commun 2011;26(2):185-197 [FREE Full text] [doi: 10.1080/10410236.2010.544283] [Medline: 21318917]
27. Pennebaker JW. Writing about emotional experiences as a therapeutic process. Psychol Sci 1997;8(3):162-166. [doi: 10.1111/j.1467-9280.1997.tb00403.x]
28. Lepore SJ, Smyth JM. The Writing Cure: How Expressive Writing Promotes Health and Emotional Well-Being. Washington, DC: American Psychological Association; 2002.
29. King LA. The health benefits of writing about life goals. Personality and social psychology bulletin 2001;27(7):798-807. [doi: 10.1177/0146167201277003]
30. Bem DJ. Self-perception theory. Adv Exp Soc Psychol 1972;6:1-62. [doi: 10.1016/s0065-2601(08)60024-6]
31. Altman I, Taylor DA. Social Penetration: The Development of Interpersonal Relationships. New York, NY: Holt, Rinehart & Winston of Canada Ltd; 1973.
32. Bazarova NN, Choi YH. Self-Disclosure in social media: extending the functional approach to disclosure motivations and characteristics on social network sites. J Commun 2014;64(4):635-657 [FREE Full text] [doi: 10.1111/jcom.12106]
33. Green-Hamann S, Sherblom JC. The influences of optimal matching and social capital on communicating support. J Health Commun 2014;19(10):1130-1144. [doi: 10.1080/10810730.2013.864734] [Medline: 24580637]
34. Nyer PU, Dellande S. Public commitment as a motivator for weight loss. Psychology and Marketing 2009;27(1):1-12. [doi: 10.1002/mar.20316]

XSL•FO
RenderX

35. Wright KB, Bell SB, Wright KB, Bell SB. Health-related support groups on the internet: linking empirical findings to social support and computer-mediated communication theory. J Health Psychol 2003;8(1):39-54. [doi: 10.1177/1359105303008001429] [Medline: 22113899]
36. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
37. Fiore MC. A clinical practice guideline for treating tobacco use and dependence: a US public health service report. JAMA: Journal of the American Medical Association 200;283(24):3244-3254 [FREE Full text] [doi: 10.1001/jama.283.24.3244]
38. Peng TQ, Zhu JJH. Mobile phone use as sequential processes: from discrete behaviors to sessions of behaviors and trajectories of sessions. Journal of Computer-Mediated Communication 2020;25(2):129-146 [FREE Full text] [doi: 10.1093/jcmc/zmz029]
39. Shrestha R, Copenhaver MM. Viral suppression among HIV-infected methadone-maintained patients: the role of ongoing injection drug use and adherence to antiretroviral therapy (ART). Addict Behav 2018;85:88-93 [FREE Full text] [doi: 10.1016/j.addbeh.2018.05.031] [Medline: 29879611]
40. Socias ME, Milloy MJ. Substance use and adherence to antiretroviral therapy: what is known and what is unknown. Curr Infect Dis Rep 2018;20(9):1-6 [FREE Full text] [doi: 10.1007/s11908-018-0636-7]
41. Carreiro S, Newcomb M, Leach R, Ostrowski S, Boudreaux ED, Amante D. Current reporting of usability and impact of mHealth interventions for substance use disorder: a systematic review. Drug Alcohol Depend 2020;215:108201 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108201] [Medline: 32777691]
42. Gonzalez A, Barinas J, O'Cleirigh C. Substance use: impact on adherence and HIV medical treatment. Curr HIV/AIDS Rep 2011;8(4):223-234. [doi: 10.1007/s11904-011-0093-5] [Medline: 21858414]
43. Stringer KL, Marotta P, Baker E, Turan B, Kempf MC, Drentea P, et al. Substance use stigma and antiretroviral therapy adherence among a drug-using population living with HIV. AIDS Patient Care STDS 2019;33(6):282-293 [FREE Full text] [doi: 10.1089/apc.2018.0311] [Medline: 31166784]
44. Gifford AL, Bormann JE, Shively MJ, Wright BC, Richman DD, Bozzette SA. Predictors of self-reported adherence and plasma HIV concentrations in patients on multidrug antiretroviral regimens. J Acquir Immune Defic Syndr 2000;23(5):386-395. [doi: 10.1097/00126334-200004150-00005] [Medline: 10866231]
45. Duong M, Piroth L, Grappin M, Forte F, Peytavin G, Buisson M, et al. Evaluation of the patient medication adherence questionnaire as a tool for self-reported adherence assessment in HIV-infected patients on antiretroviral regimens. HIV Clin Trials 2001;2(2):128-135. [doi: 10.1310/M3JR-G390-LXCM-F62G] [Medline: 11590521]
46. Xie Z, St Clair P, Goldman DP, Joyce G. Racial and ethnic disparities in medication adherence among privately insured patients in the United States. PLoS One 2019;14(2):e0212117 [FREE Full text] [doi: 10.1371/journal.pone.0212117] [Medline: 30763400]
47. Mennis J, Stahler GJ. Racial and ethnic disparities in outpatient substance use disorder treatment episode completion for different substances. J Subst Abuse Treat 2016;63:25-33. [doi: 10.1016/j.jsat.2015.12.007] [Medline: 26818489]
48. Cummings JR, Wen H, Druss BG. Racial/ethnic differences in treatment for substance use disorders among U.S. adolescents. J Am Acad Child Adolesc Psychiatry 2011;50(12):1265-1274 [FREE Full text] [doi: 10.1016/j.jaac.2011.09.006] [Medline: 22115147]
49. Tekeste M, Hull S, Dovidio JF, Safon CB, Blackstock O, Taggart T, et al. Differences in medical mistrust between black and white women: implications for patient-provider communication about PrEP. AIDS Behav 2019;23(7):1737-1748 [FREE Full text] [doi: 10.1007/s10461-018-2283-2] [Medline: 30264207]
50. Chen Y, Kruahong S, Elias S, Turkson-Ocran RA, Commodore-Mensah Y, Koirala B, et al. Racial disparities in shared decision-making and the use of mHealth technology among adults with hypertension in the 2017-2020 health information national trends survey: cross-sectional study in the United States. J Med Internet Res 2023;25:e47566 [FREE Full text] [doi: 10.2196/47566] [Medline: 37703088]
51. Smartphone owners are as diverse as their devices. Nielsen. 2015. URL: https://www.nielsen.com/insights/2015/smartphone-owners-are-as-diverse-as-their-devices/#:~:text=Asian%2FPacific%20Islanders%20(86.6%25),%25%20of%20non%2DHispanic%20whites [accessed 2024-10-16]
52. Hochstatter KR, Akhtar WZ, Dietz S, Pe-Romashko K, Gustafson DH, Shah DV, et al. Potential influences of the COVID-19 pandemic on drug use and HIV care among people living with HIV and substance use disorders: experience from a pilot mHealth intervention. AIDS Behav 2021 Feb;25(2):354-359 [FREE Full text] [doi: 10.1007/s10461-020-02976-1] [Medline: 32705370]
53. Morrison C, Doherty G. Analyzing engagement in a web-based intervention platform through visualizing log-data. J Med Internet Res 2014;16(11):e252 [FREE Full text] [doi: 10.2196/jmir.3575] [Medline: 25406097]
54. Sundar SS, Xu Q, Dou X. The role of technology in online persuasion: a MAIN Model perspective. In: Advertising Theory. London, United Kingdom: Routledge; 2019:70-88.
55. Holbert RL, Stephenson MT. Structural equation modeling in the communication sciences, 1995-2000. Human Comm Res 2002;28(4):531-551. [doi: 10.1111/j.1468-2958.2002.tb00822.x]
56. Kline RB. Principles and Practice of Structural Equation Modeling. New York City, NY: Guilford Publications; 2023.

XSL•FO
RenderX

57. Namkoong K, McLaughlin B, Yoo W, Hull SJ, Shah DV, Kim SC, et al. The effects of expression: how providing emotional support online improves cancer patients' coping strategies. J Natl Cancer Inst Monogr 2013;2013(47):169-174 [FREE Full text] [doi: 10.1093/jncimonographs/lgt033] [Medline: 24395987]
58. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online japanese breast cancer communities: differences between lurkers and posters. J Med Internet Res 2011;13(4):e122 [FREE Full text] [doi: 10.2196/jmir.1696]
59. Moon TJ, Chih MY, Shah DV, Yoo W, Gustafson DH. Breast cancer survivors' contribution to psychosocial adjustment of newly diagnosed breast cancer patients in a computer-mediated social support group. Journal Mass Commun Q 2017;94(2):486-514 [FREE Full text] [doi: 10.1177/1077699016687724] [Medline: 34295137]
60. Yoo W, Namkoong K, Choi M, Shah DV, Tsang S, Hong Y, et al. Giving and receiving emotional support online: communication competence as a moderator of psychosocial benefits for women with breast cancer. Comput Human Behav 2014;30:13-22 [FREE Full text] [doi: 10.1016/j.chb.2013.07.024] [Medline: 24058261]
61. Kim E, Han JY, Moon TJ, Shaw B, Shah DV, McTavish FM, et al. The process and effect of supportive message expression and reception in online breast cancer support groups. Psycho-Oncology 2011;21(5):531-540 [FREE Full text] [doi: 10.1002/pon.1942]
62. Dai X, Acosta N, Lu X, Hubert CRJ, Lee J, Frankowski K, et al. A Bayesian framework for modeling COVID-19 case numbers through longitudinal monitoring of SARS-CoV-2 RNA in wastewater. Stat Med 2024;43(6):1153-1169 [FREE Full text] [doi: 10.1002/sim.10009] [Medline: 38221776]

Abbreviations

ART: antiretroviral therapy
CRSE: cluster-robust SE
EAM: electronic adherence monitoring
mHealth: mobile health
RQ: research question
SEM: structure equation model
SUD: substance use disorders

Edited by A Mavragani; submitted 26.02.24; peer-reviewed by T Davy-Mendez, Q Wu; comments to author 14.06.24; revised version received 23.09.24; accepted 24.09.24; published 20.12.24.

Please cite as:
Mi RZ, Yang EF, Tahk A, Tarfa A, Cotter LM, Lu L, Yang S, Gustafson Sr DH, Westergaard R, Shah D
mHealth Engagement for Antiretroviral Medication Adherence Among People With HIV and Substance Use Disorders: Observational Study
J Med Internet Res 2024;26:e57774
URL: https://www.jmir.org/2024/1/e57774
doi:10.2196/57774
PMID:

©Ranran Z Mi, Ellie Fan Yang, Alexander Tahk, Adati Tarfa, Lynne M Cotter, Linqi Lu, Sijia Yang, David H Gustafson Sr, Ryan Westergaard, Dhavan Shah. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Issues Related to the Use of Visual Social Networks and Perceived Usefulness of Social Media Literacy During the Recovery Phase: Qualitative Research Among Girls With Eating Disorders

Elena Faccio[1], Prof Dr; Margherita Reggiani[2], Dr; Michele Rocelli[1], PhD, Dr; Sabrina Cipolletta[3], PhD, Prof Dr

[1]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Padua, Italy
[2]Psychologist-Operator at Vivere Verde Onlus, Roma, Italy
[3]Department of General Psychology, Padova, Italy

Corresponding Author:
Elena Faccio, Prof Dr
Department of Philosophy, Sociology, Education and Applied Psychology
University of Padova
Via Venezia 14
35131
Padua, 35131
Italy
Phone: 39 3479607182
Email: elena.faccio@unipd.it

Abstract

Background: The patient-centered approach is essential for quality health care and patient safety. Understanding the service user's perspective on the factors maintaining the health problem is crucial for successful treatment, especially for patients who do not recognize their condition as clinically relevant or concerning. Despite the association between intensive use of visual social media and body dissatisfaction and eating disorders, little is known about the meanings users assign to posting or searching for edited photos and the strategies they use to protect themselves from digital risks.

Objective: This study aims to examine how young women recovering from eating disorders in Northern Italy perceive the health risks and potential benefits associated with visual social networks (ie, Instagram and Snapchat). The literature has found these platforms to be detrimental to online body comparisons. It also explores the perceived usefulness, willingness, and personal interest in coconstructing social media literacy programs with girls recovering from eating disorders.

Methods: A total of 30 semistructured interviews were conducted with adolescent girls aged 14-17 years at the end of their treatment for eating disorders. The following areas of research were addressed: (1) the meanings associated with the use of Instagram and Snapchat; (2) the investment in the photographic dimension and feedback; (3) the impact of visual social networks on body experiences; (4) the potential and risks perceived in their use; (5) the importance of supporting girls undergoing treatment for eating disorders in using social networks; and (6) the usefulness and willingness to co-design social network literacy programs. Content analysis was applied.

Results: A total of 7 main contents emerged: active or passive role in using social networks, the impact of online interactions on body image, investment in the photographic dimension, effects on self-representation, perceived risks, self-protective strategies, and potential benefits. The findings highlight a strong awareness of the processes that trigger body comparisons in the virtual context, creating insecurity and worsening the relationship with oneself. The self-protective behaviors identified are the development of critical thinking, the avoidance of sensitive content, increased control over social networking site use, and a certain skepticism toward developing antagonistic ideologies. All these topics were considered fundamental.

Conclusions: The findings provide important insights for health professionals working with youth in preparing media literacy programs. These programs aim to reduce potential risks and amplify the positive effects of online resources. They underscore the importance of addressing this issue during hospitalization to develop skills and critical thinking aimed at changing small habits that perpetuate the problem in everyday life. The inherent limitations in current service practices, which may not adequately address individual needs or impact posttreatment life, must also be considered.

(J Med Internet Res 2024;26:e53334) doi:10.2196/53334

XSL•FO
RenderX

KEYWORDS

visual social networks; body image; eating disorders; risks; potentials; social networks; social network; social media; literacy; food intake; appetite disorders; appetite disorder; eating disorder; patient safety; patient-centered approach; recovery; body comparison; users; semistructured interviews; semistructured interview; girls; adolescent; adolescents; content analysis; online

Introduction

Background

Although intensive use of visual social media and online body comparisons have been strongly linked to body dissatisfaction and eating disorders (EDs) [1,2], the literature has only partially explored the potential impact of competence development in the mindful use of social media during hospitalization for EDs [3].

Recent research on the topic has primarily been divided into 2 main strands: one focusing on the specific ways social media is used by individuals with EDs [4,5], and the other exploring the online environment as a potential "place" for therapy [6,7]. There are no examples of collaborative projects involving mental health professionals and service users during recovery from EDs that aim to jointly examine visual social media in new ways and co-construct new awareness about its use.

The concept that the process of care is a collaborative journey involving various stakeholders—researchers, health care providers, and service users—is a principle prominently featured in official health care documents [8] but less frequently implemented in practice [9]. This principle suggests that patients are "real experts" and that their embodied knowledge can form the foundation for coproducing the care pathway. This approach presupposes that care program design occurs within the context of people's lived experiences, including their discourses around the body, how they perceive it, and the environments in which they socialize about ideals of desirable bodies (eg, social media). It emphasizes the importance of engaging with users in their own worlds and languages, rather than being confined to the perspectives of health care professionals [9].

During hospitalization for ED treatment, girls are usually either not allowed to use mobile phones or only permitted to use them during restricted time slots. If the use of social networking sites (SNSs) is not discussed and explicitly addressed, returning home may lead patients to revert to their previous SNS usage and interpretations, potentially undermining the long-term effectiveness of the treatment. Social media literacy programs aimed at improving body image–related outcomes should be implemented not only in the context of ED prevention but also in rehabilitation [3]. According to Fitzsimmons-Craft et al [10], only 20% of therapists have asked patients with ED about the perceived impact of social media on their body image experiences. Understanding the meanings and ways in which girls use SNSs is essential for designing educational and rehabilitation pathways that help build self- and body perception based on criteria different from those that contributed to the development of the ED [11,12].

Our research represents the first step in drafting a collaborative project involving researchers, health professionals, and service users. The goal is to identify the most salient and relevant issues related to social media use from the perspective of girls undergoing treatment.

SN Use and Body Image Concern: What Psychological Processes Come Into Play?

Social media is a significant source of sociocultural pressure regarding appearance and influences adolescents' relationships with others as well as their perceptions of their own bodies. Active engagement in activities related to the photographic dimension on social networks (SNs) has been shown to contribute significantly to the development and perpetuation of concerns related to body image and food intake [13,14]. Such activities include publishing photographs, viewing others' photographs, commenting or liking posts, and receiving feedback. Compared with the passive use of SN platforms, these active behaviors are more insidious in fostering dynamics of comparison and competition between bodies, thereby exposing individuals to the scrutiny of social judgment [15-17].

A clear and direct connection between posting edited photos and risk factors for EDs has been consistently confirmed by various studies [18]. Indeed, social media usage is a plausible risk factor for the development of EDs, a finding supported by research from Asia, indicating that this association is not limited to traditionally Western cultures [19].

Intensive use of SNSs has been linked to the internalization of the thinness ideal, self-objectification, and body dissatisfaction [20-22]. The literature defines the internalization of the thinness ideal as the degree to which an individual cognitively adopts socially defined standards of attractiveness and engages in behaviors aimed at approximating these ideals [20]. This process of striving for an ideal physique through the intensive use of social media can lead to body dissatisfaction and subsequently to disordered eating behaviors. Individuals may pursue an idealized body image that is often unattainable, which can persist even during recovery from EDs and increase the risk of relapse. Moreover, the habit of viewing bodies as objects on social media can socialize girls into self-objectification, as they internalize an external spectator's perspective of their own bodies. This encourages habitual monitoring of one's physical appearance, a phenomenon known as self-surveillance. It can lead to feelings of anxiety, reduced awareness of one's own internal states, body shame, and a fear of internalized judgment from others. Another factor that could influence the relationship between exposure to SNSs and the development of EDs is the significance placed on feedback received through social media. Research has shown that the extent of one's SN, as measured by the number of "friends" [23], may predict a stronger inclination toward pursuing thinness. Facebook (Meta Platforms, Inc.) use has been identified as a prospective predictor of increased symptoms related to EDs [13]. Moreover, research indicates that high "appearance exposure" specifically through Facebook, rather than overall Facebook use, is positively correlated with increased body image issues among adolescent girls [14].

XSL•FO
RenderX

Visual SNs and the Risk of Developing EDs

Regular sharing of self-images on social media platforms such as Facebook, Instagram (Meta Platforms, Inc.), or Snapchat (Snap Inc.), along with actively manipulating these images before sharing, seems to be linked to heightened perceptions of body shape and weight, increased body dissatisfaction, and tendencies toward dietary restriction [24]. Facebook users tend to be older compared with Instagram users, and this age difference may contribute to younger Instagram users being at higher risk of developing EDs and expressing greater concerns about the impacts of sharing their images on SNs [4]. Analyzing fitness content on Instagram, particularly posts tagged with hashtags such as #fitspiration, #fitspo, and #thinspiration, reveals that the majority of images depict women with 2 prominent characteristics: thin bodies that are also visibly toned and muscular. This portrayal reinforces the notion that only a thin body can be considered fit [25]. The pairing of thinness with muscularity appears in as many as two-thirds of the photos posted by influencers [20], and these images have a more significant impact on body image concerns compared with photos posted by celebrities or models [25].

Another SNS that emphasizes the visual dimension and has experienced rapid growth is Snapchat. Currently, there is limited research dedicated to investigating the effects of this platform. One of its distinctive features is that users can apply filters to alter their appearance before sharing photos. Additionally, Snapchat is known for its privacy settings and the ephemerality of content, which is typically automatically deleted shortly after being viewed [26].

Comparative research indicates that the correlation between photo manipulation using specific editing tools (filters) and body image concerns appears to be stronger on Snapchat than on Instagram [27]. Those who frequently use Snapchat's lenses to enhance their photos appear to experience higher levels of dissatisfaction with their appearance, leading them to seek out environments that are highly appearance-oriented and focused on enhancing their appearance. Posts on Instagram tend to be more curated, with teenagers often spending considerable time and effort choosing the "right" content before taking a photo, rather than editing it after it has been taken [28]. This investment of time and energy before capturing the photo may strongly correlate with concerns about body image compared with postphoto retouching [29,30]. Moreover, while Snapchat is often used to connect with close contacts, users may be less focused on their physical appearance when taking pictures. By contrast, the larger audience on Instagram may have a more significant influence on body image concerns.

Methods

Research Aims

In light of these considerations, one might inquire about how adolescent girls undergoing treatment for EDs utilize SNs. What significance do they attribute to their SN usage? Do they differentiate between manipulated and unedited images? What strategies do they use to shield themselves from the impact of the content they post or view? Are they cognizant of which modes of SN use could potentially influence clinically significant outcomes?

Our focus is on exploring the usage of purely photographic platforms such as Instagram and Snapchat, which are particularly insidious in terms of fostering comparisons between bodies on the network [27-30]. We are interested in investigating the different emotions and experiences participants have when sharing their own photos in a private and intimate network among friends, such as Snapchat [27], versus sharing them in front of a larger audience, as is the case with Instagram [30].

Another area of interest for our investigation involves understanding the meanings associated with the permanence or short duration of content created and received on social media platforms. Specifically, while content published or shared on Instagram is permanent, on Snapchat, users determine the lifespan of their content, after which it is automatically deleted [27].

Additionally, we are interested in understanding whether girls use strategies to mitigate the potential influence of exposure to certain images, whether they can differentiate between edited and unedited photos, and whether they protect themselves from uncritically accepting messages from SNs. The literature [29] has demonstrated that visual social media can also be utilized during the recovery process from an ED to promote health messages and well-being. Paradoxically, the use of SNs that involve transforming one's own photos has also been identified as an important tool, albeit in a positive sense [26]. Therefore, it is not merely the photo itself but how it is used that determines whether it serves as an "exit" or "entry" point in managing eating behaviors.

The ability to alter one's images has both positive and negative implications. It allows individuals to transform their appearance inexpensively and democratically, fostering a sense of personal empowerment. However, the ease of digitally altering one's appearance also leads to an increase in the number of manipulated images that people view and are confronted with. This phenomenon may subsequently heighten dissatisfaction with one's physical appearance and intensify the pressure to alter it.

In conclusion, we also sought participants' opinions on the relevance of the proposed topics and what other subjects they would consider important in a collaboratively designed program focused on the responsible use of SNs, intended for potential users in the recovery process.

Sampling and Recruitment

After sharing the objectives and content of the research with health care staff, we informed users of the local health unit for the treatment of EDs within the National Health Service in Eastern Veneto (specifically, the Casa delle Farfalle Residential Protected Therapeutic Community for Children).

The health care staff presented the research to the girls, obtained their consent, requested authorization from their parents, and provided information about the study. They also organized the schedule for appointments. All the girls who were invited agreed to participate; none declined the invitation. The participants and

XSL•FO
RenderX

their families were informed of their right to withdraw from the study at any time. They signed a written informed consent form regarding their participation in the research and its potential publication.

Thirty adolescent girls, with a mean age of 15 (range 14-17) years, expressed interest in participating. The inclusion criteria were as follows: a diagnosis of an ED, being in the early stages of treatment with at least 15 days of hospitalization, having an active Instagram and Snapchat profile for at least six months before admission, and willingness to participate in interviews.

We extended the invitation to participants who were at the beginning of their treatment journey because we preferred individuals who had not yet been significantly influenced by the discussions occurring between users and health care providers during their hospitalization period.

The interviews were conducted by MRe, one of the authors of this study. After confirming their availability, she held several preliminary meetings to establish rapport and further explain the study's objectives and implications. The interviews were conducted individually in a room on the ward where each participant met with the research assistant.

Ethical Considerations

The research protocol was approved by the Ethical Committee of the School of Padova, University of Padova (approval number 2018/2745-10/7). Before commencing the interviews, participants and their parents were informed that participation in the research was voluntary and without any compensation, and their consent was obtained. The original informed consent that participants signed included permission to conduct secondary analysis without additional consent. The final data set is anonymized, ensuring that no identifiable private information linked to participants is included.

Data Collection and Analysis

A semistructured interview consisting of 9 open-ended questions was designed specifically for this research (Table 1). It covered the following areas: (1) the meanings attributed to the use of Instagram and Snapchat; (2) the level of engagement with the photographic aspect and the interactivity generated around it, particularly in terms of feedback; (3) the perceived impact of Instagram and Snapchat usage on body experiences; (4) the potential benefits and risks associated with the use of SNs; (5) the importance of providing support and guidance to girls undergoing treatment for EDs in their use of SNs; and (6) the willingness to participate in and contribute to the co-design of SN literacy programs.

Table 1. Structure of the semistructured interview and the open questions composing it.

Areas of research addressed by the semistructured interview	Open questions that compose the interview
• The meanings associated with the use of Instagram and Snapchat	• How do you consider virtual reality compared with the real world? • What is your main use of Instagram and Snapchat?
• The investment in the photographic dimension and in the interactivity generated around it in terms of feedback	• Have you ever received comments on the photos or stories you posted? How did you experience them? • To what end do you publish photos or stories? Under what circumstances do you usually edit them before publishing them? In what way? What aspects do you focus on?
• The perceived impact that the use of Instagram and Snapchat can have on the body experiences	• When you look at your account (Instagram or Snapchat), do you feel represented by the resulting image of you? • Would you like to recount a positive and a negative experience you have had with regard to Instagram, Snapchat, or both? • Have you ever compared your published photos with others? What aspects did you focus the comparison on? • Do you think that the information you received from the pages and the activities related to the photographic dimension had an influence on your and others' ideas of your body?
• The potential and the risk glimpsed in the use of social networking sites	• Which use of the platform do you associate with an improvement in your relationship with your body and which with a worsening? How could you protect yourself with regard to use and content that could be critical in your relationship with your body? • What potential and what risks do you see in the use of visual social networks?
• Importance of accompanying girls undergoing treatment for eating disorders in social network use • Willingness to participate in and co-design social network literacy programs	• What goals should these projects on the conscious use of social media have? • Would you be willing, once your treatment is over, to collaborate with the operators and share your experience as a testimonial during courses for other girls who will be hospitalized later on?

The researchers formulated the questions based on the research objectives. Initially, there were twice as many questions, but they were later streamlined to focus on the most comprehensible ones. The first interviews helped refine these questions further, addressing any areas that appeared unclear or prone to misunderstanding.

XSL•FO
RenderX

Finally, the girls were asked to provide feedback on the relevance of the suggested themes for developing programs on critical thinking regarding SN use during hospitalization. They were also encouraged to add any additional topics they considered important. The interviews took place between June and July 2021. Each interview lasted approximately 45 minutes and was audio-recorded and transcribed verbatim.

Content analysis was conducted following the guidelines outlined by Creswell [31]. The responses from participants were read and coded, creating meaning units that were categorized accordingly. This process resulted in the development of a codebook, which was then systematically applied to analyze all responses. A triangulation of analyses involving 3 researchers (MRe, EF, and SC) was conducted. This process included sharing one researcher's initial coding to critically discuss the initial coding patterns and subsequently reach a consensus on the overall interpretation of the data. This approach allowed for interpretations to be challenged, refined, and ultimately agreed upon through collaborative discussion among the researchers.

As the analysis moved from macrocategories (general) to microcategories (specific), there was a progressive refinement and detailed coding of the responses. Initially, responses were categorized at a broader and more anonymous level, providing a general commentary. As the analysis proceeded, categories became more specific, focusing on individual experiences and emotional impacts on personal lives. For instance, the macrocategory "virtual world configuration (VW)" is subdivided into 2 categories: "VW as an extension of reality" and "VW as a world distinct from real life." These categories represent opposite ends of a continuum that delineates how participants perceive the relationship between the virtual world and the real world. The category "virtual world as an extension of reality" is further detailed in the subcategory (self-reference) "the virtual world significantly influences self-perception." Participants who hold this view consider the virtual world relevant, viewing it as a source of inspiration for their real-life decisions (microcategory, fourth level of analysis). Conversely, individuals who perceive the virtual world as "separate and distinct from the real world" (category—second level of analysis) specify that the virtual world "has no impact on their self-concept" (subcategory—third level of analysis) and they "regard it as transient" (microcategory, fourth level of analysis). Another example of the categorization system and codebook created can be seen in describing the use of SNSs: we classified as "active use" everything related to actively seeking contacts and interactions, engaging in discussions, sharing personal information, and editing photos and stories. By contrast, we classified as "passive use" those actions oriented toward reading and viewing content made available by others, without active personal involvement. This position entails observing rather than actively participating in the virtual scene.

The second step was quantitative: each category was assigned a code of 1 or 0 based on its presence or absence. We focused on the number of respondents who mentioned a particular category, rather than the frequency of mentions for that category. This approach allowed us to provide an overview of the girls' attitudes toward the investigated issue. It was particularly important for us to gain a general understanding of the impact and prevalence of specific types of experiences, as well as to explore the personal meanings the girls attributed to the topic in a more qualitative manner.

After completing the analysis, all authors of the paper convened to review the results and make any required adjustments. The draft summary of the main results paper was also shared and discussed during the collaborative analysis and writing phases.

Results

Virtual Word's Configuration

An overview of the results and the codebook developed during the analysis can be found in Multimedia Appendix 1.

The thematic coding of the text collected in reference to the first answer revealed 2 main categories: the first, mentioned 28 times, portrays the virtual world as an extension of everyday reality, where online activities have a tangible impact on how respondents perceive themselves. The second category, mentioned only twice, contrasts with the first by viewing the virtual world as an ephemeral and transient realm where an idealized self-image is projected. In this perspective, the impact of online activities is perceived as less significant compared with activities in "real" life, which are seen as separate and more concrete.

SNs' Active or Passive Ways of Use

Regarding the mode of SN use, 2 main polarities emerged: one characterized by more active engagement and the other by more passive interaction. The proactive category included initiating relationships with others (5 mentions), sharing personal narratives (11 mentions), posting personal photos or stories (3 mentions), and searching for information (23 mentions). By contrast, the passive end of the spectrum included consulting content posted by others (21 mentions). Regarding proactive use, Instagram was noted for fostering positive emotional experiences through contact and exchange with other users (3 mentions), whereas Snapchat was valued for its utility in sharing daily life and disclosing personal information (2 mentions). The most frequently mentioned activity, active information seeking through SNs, was primarily described as a source of stimulation and inspiration, driven by curiosity (13 mentions), to reinforce and update interests (2 mentions), and also as focused information seeking on body-related issues (16 mentions).

> *As for the eating disorder, I used it a lot to watch food videos, videos of people cooking or even people doing particular sports, gym, like, I don't know, all those fitness posts. That was the main use.* [Participant 2]

Finally, the use of photo editing functions was noted. Passive use of Instagram and Snapchat refers to the reception of content generated by other users for entertainment purposes.

Online Interactions' Impact

The impact attributed to SNS use is expressed in terms of "social confirmation," as they are viewed as significant for providing a channel of self-recognition (20 mentions) that includes both approval and disapproval (12 mentions). This results in emotions of personal gratification (7 mentions), support (3 mentions),

XSL•FO
RenderX

and encouragement (5 mentions), but also anxiety related to the fear of judgment (5 mentions) and feelings of disconfirmation when few likes are received (10 mentions).

> *Back then, I thought that if more people commented on my content, that meant more people liked me. The more likes I received, the more others liked me [...].* [Participant 6]

The impact of online interactions varied depending on the specific use of Instagram or Snapchat (20 mentions): Instagram is perceived as having a more tangible and "real" impact compared with Snapchat (18 mentions). Feedback received on Instagram is deemed more significant, to the extent that some respondents consider it a metric for determining their personal worth (8 mentions). There was also a notable sense of distrust toward feedback gathered through SNSs (10 mentions). Consequently, receiving likes and comments was perceived as unimportant due to skepticism about the authenticity of posts. Some respondents rarely received likes and comments, feeling unpopular (6 mentions) and disregarded both on social media and in everyday life.

Regarding Snapchat, the automatic deletion of content and the narrow composition of the network seem to foster a playful, humorous atmosphere (5 mentions) and encourage creativity. This moderates the impact of feedback received (11 mentions).

Investment in the Photographic Dimension

The use of photographs on social media platforms shifts from total involvement (24 mentions) to complete disengagement (6 mentions). Photographs serve not only for self-promotion (20 mentions), fulfilling the desire to be seen (11 mentions), and presenting the best version of oneself (7 mentions), but also for gaining a deeper understanding of oneself (9 mentions) and celebrating specific events (4 mentions). Active involvement in the photographic dimension was specifically associated with Instagram, attributed to its high degree of self-image exposure (16 mentions). SNSs are portrayed as platforms where users consistently share only the best aspects of themselves to appear perfect (7 mentions). The use of Instagram, in particular, fosters second thoughts and inhibitions regarding photo publication (10 mentions). The high degree of exposure on Instagram promotes a tendency to hyper-control the quality of photographs, induces concerns about the perfection of one's images, and encourages individuals to scrutinize and edit their photos extensively.

> *When I was going to use Instagram, I never published photos [...] because many more people can see it there, even two hundred, three hundred people can see it [...]* [Participant 3]

The "disinvestment" contents (6 mentions) were primarily attributed to feelings of embarrassment (3 mentions) and discomfort (4 mentions). By contrast, Snapchat's emphasis on ephemerality (3 mentions) and frivolity (2 mentions) in its content contributes to a platform dominated by carefree attitudes toward self-image. This reflects a disengagement from the photographic dimension.

Self-Representation

The collected texts appear polarized: they describe a correlation between the body presented online and the body perceived in the real world (17 mentions), but also refute this congruence (10 mentions) in favor of emphasizing the dominance of the real. Users primarily discussed self-representation on Instagram due to Snapchat's ephemeral nature. Sharing photographs of themselves is viewed as a means to express their ways of being and interacting with others and the world (5 mentions), as well as a method to foster deeper personal understanding (3 mentions). There was considerable emphasis on portraying a "true" image of themselves (8 mentions). In some cases, it seems feasible to accept photos from the past, particularly those depicting a thinner body, alongside acceptance of the current body image (2 mentions). However, in other cases, confronting images of oneself from a thinner period can be agonizing (3 mentions), often evoking a sense of regret. Ten responses conveyed a sense of discrepancy between the real-life body and the online representation, attributed to efforts to present oneself in a more favorable light, sometimes accentuating flaws or concealing them (2 mentions). This can lead to a perception of projecting a false self-image (8 mentions) compared with one's subjective self-image. The existence of photos that diverge so significantly from each other also contributes to a sense of fragmentation in one's self-image (1 mention).

SNs' Risks

All the collected texts addressed the risks associated with SNSs' use, especially activities involving the photographic dimension (30 mentions). Content on Instagram, along with related interactive activities, is perceived as significantly influential in shaping ideas about the body and self-representation (21 mentions). This influence is evident in the promotion of a singular body prototype characterized by thinness, perfection, and muscularity (17 mentions). Viewing photos of others is seen as the initial step toward attaining this ideal, although it is perceived subjectively as deceptive (8 mentions). Unrealistic content shared on Instagram aims to bolster this ideal through maladaptive behaviors (11 mentions), sometimes progressing toward images that align with the "Pro-Ana" direction (5 mentions).

Many responses (24 mentions) emphasized the risk of deteriorating one's relationship with their body due to a heightened focus on physical appearance and an increased attention to the body dimension. Participants acknowledged the role played by Instagram in exacerbating concerns about weight and body shape (11 mentions), particularly through pages where individuals provide tips related to EDs (10 mentions). The heightened concerns are perceived to induce feelings of being "wrong" (2 mentions) and body dissatisfaction (8 mentions). Comparison with images posted online is a dominant theme (19 mentions), where physical attributes, particularly thinness, are frequently cited as comparison criteria, often accompanied by feelings of envy (4 mentions) and inferiority (9 mentions).

A sort of "vicious circle" of negative influence through imitation has been described:

XSL•FO
RenderX

> *A famous girl puts up a picture then maybe she sees that the picture has many comments from girls who say, "You're so thin, you're so beautiful", and all this, and then she continues to put up pictures of that type, and the girls obsess more and more.* [Participant 4]

In relation to the interactive dimension, risks are linked to social judgment and the pressure toward conformity (2 mentions). Finally, another risk identified is the underestimation and casual approach with which users engage with SNs, not fully recognizing the impact and influence that published content can have on people's lives (5 mentions).

Regarding Snapchat, the texts indicated a more self-deprecating attitude toward sharing content and less anxiety about posting images. This makes Snapchat perceived as a platform where users feel free from concerns about their physical appearance and comparison with others.

Self-Protective Strategies While Using SNSs

Regarding self-protective strategies, the primary defense that empowers the use of SNs without negative impacts on one's body image is subjective critical capacity (17 mentions). This includes the ability to recognize and distinguish true content from edited ones (4 mentions) and an awareness that each person has about their own unique body (11 mentions). Useful strategies to promote this critical capacity could be implemented by the Instagram platform itself (4 mentions). One form of defense involves actively avoiding sensitive content (7 mentions), self-exposure (2 mentions), or excessive SN use (1 mention). However, some participants noted the challenge of implementing this strategy due to their strong curiosity to view certain content. The Instagram platform itself should develop protective measures regarding the generation of certain content, which could include content control (2 mentions) or content removal (2 mentions).

From the text analysis, skepticism (8 mentions) and a sense of helplessness (4 mentions) emerged regarding the lack of control over online content, as well as a feeling of inescapability (5 mentions) due to the unavoidable accessibility of certain content on Instagram's home page, even when not intentionally sought. Additionally, other defensive strategies should involve parental controls (1 mention) and the sharing of experiences and content (3 mentions).

SNs' Potentials

Several potentials were identified regarding the use of SNs, especially concerning the photographic dimension. These potentials are linked to the possibility of improving one's relationship with their body by emphasizing spontaneity and naturalness in content creation (4 mentions), which is also perceived as a protective measure against the potential for negative comments (2 mentions). The act of searching for content is perceived as supportive during difficult times (3 mentions), providing an opportunity to follow individuals who have faced similar challenges and who, having overcome them, serve as sources of support (2 mentions) and encouragement (1 mention). Platforms such as Instagram also facilitate personal expression (9 mentions) and allow individuals to take an external perspective on their experiences (3 mentions). Exploring profiles of influencers and artists who promote diverse ideas about the body offers an opportunity to discover perspectives that might not be encountered otherwise. These individuals can inspire the exploration of values that challenge prevailing trends.

Another potential lies in actively creating content that emphasizes authenticity (6 mentions) and carefreeness (5 mentions) rather than solely focusing on physical appearance or presenting oneself in an idealized manner. By promoting authentic content, individuals can contribute to limiting the proliferation of unrealistic body ideals and reducing the circulation of idealized and unrealistic representations of the body (3 mentions). However, this aspiration is tempered by skepticism regarding the actual impact of sharing "realistic" content and finding nondeceptive body ideals. Finally, some texts emphasize the potential offered by the opportunity to connect with other people (6 mentions).

With regard to the importance of guiding girls undergoing treatment for EDs to reflect on the topic, there was unanimous agreement (30 mentions), confirming all proposed topics as relevant. If the project had been implemented, the girls would have enthusiastically participated in the group, viewing it as a safe space to find the courage and express themselves (28 mentions) and being inspired by individual choices (20 mentions). To advocate for content that goes beyond the pursuit of bodily perfection, one needs to observe behaviors in others (15 mentions). Expressing oneself with humor and levity is considered the best antidote (15 mentions), but it necessitates self-confidence and a willingness to reject the sanctity often placed on one's image (15 mentions).

The interest and willingness were confirmed for both participating in person to listen to other girls undergoing rehabilitation (30 mentions) and discussing these issues on an equal footing with peers in similar situations. Additionally, after completing the therapy course, there is interest in potentially serving as a testimonial to set a positive example of the conscious use of social media based on personal life experiences (28 mentions). This opportunity could enable several outcomes: providing support to others (15 mentions), creating a space for personal reflection (7 mentions), hearing about the challenges others have faced in overcoming EDs (8 mentions), and connecting with peers who share similar experiences (7 mentions).

Discussion

Principal Findings and Comparison With Prior Studies

This research investigated how young women recovering from EDs perceive both the health risks and potential benefits of visual SNs, specifically Instagram and Snapchat, which have been identified in the literature as platforms where harmful body comparisons can occur. Additionally, the study explored the participants' willingness and interest in collaboratively developing social media literacy programs aimed at supporting individuals recovering from EDs.

The study's findings underscore a keen awareness among participants of the mechanisms that trigger body comparisons

XSL•FO
RenderX

in the online realm, which contribute to feelings of insecurity and detrimentally impact self-relationships. However, the girls do not consistently harness this awareness to their benefit. The identified self-protective behaviors include the development of critical thinking, avoidance of sensitive content, increased control over SNSs, and a degree of skepticism toward the propagation of conflicting ideologies. All these topics were deemed fundamental by the participants.

For the young people participating in the study, 3 primary modes of using Instagram and Snapchat emerged. The first mode was characterized by the relational dimension, involving affective experiences of contact and exchange, which reduced feelings of loneliness and enhanced perceptions of social support. The second mode of use focused on self-presentation and self-disclosure, particularly evident in relation to Instagram. The third mode of use involved active information seeking, specifically the search for stimuli, inspiration, curiosity, and updates related to one's passions, interests, appearance, and eating behaviors.

The topics emerging from the data confirm findings in the literature regarding the problematic use of SNs and the development of EDs among preclinical samples. Following appearance-focused accounts on Instagram and engaging in photo-based activities, such as posting selfies or liking and commenting on photos, were associated with poorer body image outcomes [14,15,28,31,32]. Passive use characterized by role passivity, which involves the passive reception of content generated by other users for entertainment or spectatorship purposes, has been linked to a decrease in subjective well-being through social comparison [2]. Among our participants, even the passive mode of social media use—simply opening food- and body-related videos—was acknowledged by the girls as reinforcing problematic behaviors from a symptomological point of view, such as dieting and exercise obsession.

One of the most alarming findings from our research was the significant disparity between perspectives regarding the online environment either as a continuation of reality or as distinct from it (28 vs 2). Participants who viewed the virtual world as an extension of reality tended to perceive the feedback received as highly impactful on their self-perception and sense of self-worth. The significance of online feedback was closely tied to the validation it provided, through either approval or disapproval. However, some participants expressed skepticism about the authenticity of SNs, viewing the images and content as misleading. As a result, they perceived the impact of online feedback as less meaningful compared with feedback received in real life [33,34].

This finding could be pivotal for a project focused on training, discussing, and reflecting on the subject: Highlighting the distinction between individuals who trust the authenticity of the online environment and those who approach it with skepticism is an initial step toward fostering distance and increasing awareness of one's online presence and behavior.

Self-promotion emerged as a significant motivator for engaging in photographic activities on SNs. Users aimed to portray their best selves and sought positive recognition from others, aligning with existing findings in the literature [35,36]. The study also identified other motivations such as celebrating moments or events and seeking visibility among others [2], which are enhanced by Instagram's exposure and amplification features. Conversely, users' disengagement was primarily driven by feelings of embarrassment and discomfort associated with self-exposure, as well as the ephemeral and light-hearted nature of content shared on Snapchat [37].

All participants reported using criteria to select photographs for publication, particularly on Instagram. Those who noted a gap between their online self-presentation and their real selves emphasized an awareness that the virtual world perpetuates a sense of falsity by showcasing only the most desirable image of oneself. Regarding risks associated with photo-related activities, Instagram was seen as a highly visual platform centered around photos, with a strong focus on aesthetic content and a prevalence of edited photos that promote unhealthy, unrealistic, and deceptive body ideals [38]. These body ideals are perceived as highly influential in promoting and striving for perfection and the thin ideal [29]. Despite often being perceived as unattainable, media portrayal of this ideal leads women to view it as normative and central to attractiveness, thereby internalizing socially constructed appearance ideals [39,40]. This internalization and subsequent comparison lead to decreased satisfaction with one's own body, efforts to manage one's appearance to reach this ideal, and increased concerns about body image [41]. Social comparisons upward tend to evoke envy and feelings of inferiority, in terms of not only physical appearance but also the personality and lifestyle of individuals within one's SN. Moreover, participants emphasized the risks associated with the interactivity facilitated by SNs. This phenomenon is referred to in the literature as "body talk," which involves interpersonal interactions focused on bodies and physical appearance [42]. Body talk reinforces the value and significance of appearance and contributes to the construction of appearance ideals [43]. Previous research has demonstrated that body talk is positively correlated with body surveillance, body shame, and perceived pressure to conform to thinness ideals [42].

In terms of the potential of SNSs, participants emphasized the significance of being able to create, choose, and share content freely. They viewed promoting active sharing of naturalistic and unedited content, characterized by authenticity and spontaneity, as a potential means to enhance their relationship with their own bodies and for personal self-expression [44]. Improving the relationship with their own bodies was also seen as potentially helpful in finding support from people who have had a similar experience. Finally, users highlighted the potential of SNs to stabilize and enhance peer relationships with previously inaccessible groups, as well as the possibility of discovering things that might be difficult to encounter in everyday offline life. Our participants also highlighted the challenge of self-discipline and avoiding potentially harmful online content. Therefore, in the context of social media literacy programs, it could be beneficial to create spaces where girls can cultivate the courage to block the use of social media in such situations, as they suggested.

Regarding defense strategies against the influence of content published on SNSs, users primarily proposed empowering

XSL•FO
RenderX

subjective critical capacity. They suggested developing and enhancing social media awareness and adopting a critical approach to viewing and scrutinizing images and posts, including considerations about the realism of the images and the intentions of the posters. They also suggested enhancing the control that platforms themselves could exercise over the type and quality of images and content posted online, even advocating for automatic deletion of content that could pose risks to users' body image. However, some users expressed skepticism and uncertainty about the feasibility and legitimacy of such controls. They believed that the best strategy was to avoid content perceived as subjectively risky for their body image and that could potentially expose their image in an undesirable way.

It should be noted, however, that recent literature suggests enhancing critical thinking about the media may not be the sole or primary mechanism of change in effective media literacy interventions. There is uncertainty about whether literacy programs aimed at reducing the risk of EDs also effectively address critical thinking about the media. Future research will need to clarify this ambiguity [45].

What emerged from our research as a whole is that girls often have an awareness of risks but struggle to use this awareness to their advantage. They can recognize edited photos but still perceive them as authentic. Only a few responses indicated that "distrust" can serve as a strategy to mitigate perceived influence. These findings suggest that enhancing critical thinking about the media may not be the sole or primary mechanism of change for effective media literacy interventions.

However, we did not explore the influence of context, particularly the role of parents and the school environment, in moderating social media use and coping with body dissatisfaction. Among the limited qualitative research available on the meanings associated with social media use, Burnette et al's [46] study sheds light on this topic. The authors found that among adolescent girls in general (not necessarily those in treatment for EDs), parental control devices for social media and school environments may be particularly effective in helping adolescents enhance their strategies for filtering out the most harmful messages and developing a broader conception of beauty. These factors act as protective measures against the risks associated with social media use.

Practical Implications

Given the intricate connection between SN usage and its effects on mental health [47], it is crucial to investigate this phenomenon by valuing the personal perspectives of individuals who directly experience discomfort and pain related to SN use. Health professionals should integrate discussions about SN use into their treatment of body dissatisfaction and disordered eating [2]. Much of the content emerging from the research explores new dimensions of SNS use beyond existing literature. These include identifying 3 primary modes of use, understanding the challenge of distinguishing between real and online environments and the resulting behaviors, addressing emotional detachment strategies, and considering the implications of interacting with apps that feature permanent visual content and large followings. Additionally, the study discusses platforms with temporary photo availability and intimate networks, strategies to enhance positive social media use through peer support, and methods to mitigate risks shared between users and platform providers.

Considering the ineffectiveness of prohibiting adolescents from using SNs, it could be beneficial to focus on teaching them how to navigate these environments safely instead. Rather than restricting access [48], leveraging existing SN platforms for targeted interventions is possible. This approach can utilize anonymous access to health information available on the internet, which is appealing to young people seeking such resources [49]. This strategy aims to empower youth with the knowledge and skills needed to use SNs responsibly and effectively manage their online experiences.

We believe that every problem has its solution. In other words, starting from where young people are—specifically, their desire to engage and immerse themselves in the methods and meanings of SN use—we can empower credible and interested role models (such as girls with EDs) to influence others toward more mindful use of self-protective strategies on SNs. The most impactful approach to wield this influence is to engage these individuals as influencers and collaborate with them to develop literacy programs within treatment frameworks. This research introduces fresh insights and underscores the necessity of addressing sensitive issues to design and implement more effective and tailored educational programs tailored to individual health care recipients [47].

Limitations

Several study limitations should be noted, including the small sample size and the fact that all participants had undergone treatment at a specialized center for EDs, indicating they already had awareness of the issue. Future research could broaden the survey to include a wider range of participants, including those in the preclinical phase of EDs. This approach would enable professionals to enhance their understanding of potential histories, distress forms, and health promotion strategies, thereby guiding the management of early signs of distress. Another limitation is self-selection, as participants were invited by health care staff and chose to participate voluntarily rather than being randomly selected. This may result in a biased and unrepresentative sample of the population, as those who choose to participate may possess specific characteristics that distinguish them from the broader population. However, it is important to note that qualitative research such as this does not aim to generalize findings to the entire population. Instead, its focus is on exploring the lived experiences and meanings ascribed to those experiences by the participants themselves.

Another limitation arises from the context of ED treatment services where participants were restricted from using phones, which meant their responses relied on memory rather than current experiences. A third limitation that could be addressed in future research involves assessing the long-term retention and application of knowledge about the risks and potentials of social media use over time.

XSL•FO
RenderX

Conclusions

While previous literature extensively covers studies on the risks of using SNs and prevention programs, our study contributes numerous ideas on addressing this topic with young people undergoing therapy for EDs by involving them as experts by experience. It allows the identification of the most significant themes that may arise, enabling health professionals to incorporate them into discussions and share the self-protection strategies identified. There is a call to leverage the potential of SNSs, such as user-generated content and interactivity, to promote beauty ideals divergent from the prevailing model, alongside content that emphasizes humor rather than appearance-centric narratives. However, the crux of the research lies not only in what to propose but also in how to propose it. Enrolling girls as coparticipants in the programs represents, in our view, the real breakthrough.

Being able to share their experiences and offer their stories as a potential source of inspiration and influence among peers allows individuals to transform from passive recipients of life, influenced by others, into conscious leaders who guide and inspire others. This involves coconstructing care pathways, creating space, and negotiating ideas with those willing to take on the new role, such as that of a patient who, after completing their treatment journey, makes themselves available to assist others at the beginning of their own journey. Co-design indeed necessitates flexibility and a strong willingness from all participants. It is not about predefined content but rather entails collaborative decision-making where meetings and individuals together determine the objectives to pursue.

Acknowledgments

We express our gratitude to all the mental health workers who contributed to the research, actively participating in every stage of it, and to all the girls with whom the research project was collaboratively constructed. They gave their time to reflect and share their experiences sincerely and showed a genuine and generous interest in this work. We are also very grateful to our young colleagues Ludovica Aquili (PhD student in Social Sciences) and Lia Bitetti (psychologist, psychotherapist in training, graduated from the University of Padua) for their support and expertise in editing this work.

Data Availability

The data sets generated or analyzed during this study are not publicly available due to privacy restrictions. Indeed, they contain information that could compromise the privacy of research participants. They are available from the corresponding author upon reasonable request.

Authors' Contributions

EF contributed substantially to the conception and design of the work. MRe created the structure and collected the data. Together, both drafted the work and revised it critically for important intellectual content. All authors participated in the analysis and interpretation of the data.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Overview of the results and the codebook of analysis. The 4 categories are indicated from the general (macrocategories and categories) to the particular (subcategories and microcategories). The number in brackets refers to the number of respondents who evoked the relevant category upon the 30 involved. The categories were not always subdivided further; in some cases, in fact, no details were added during the interview. Interviews with young women in recovery from eating disorders (EDs) were held in North Italy in 2021.

[DOCX File , 23 KB - jmir_v26i1e53334_app1.docx]

References

1. Fardouly J, Rapee RM. The impact of no-makeup selfies on young women's body image. Body Image 2019 Mar;28:128-134. [doi: 10.1016/j.bodyim.2019.01.006] [Medline: 30665030]
2. Saunders J, Eaton A. Snaps, selfies, and shares: how three popular social media platforms contribute to the sociocultural model of disordered eating among young women. Cyberpsychol Behav Soc Netw 2018 Jun;21(6):343-354. [doi: 10.1089/cyber.2017.0713] [Medline: 29883209]
3. Gordon CS, Jarman HK, Rodgers RF, McLean SA, Slater A, Fuller-Tyszkiewicz M, et al. Outcomes of a cluster randomized controlled trial of the SoMe social media literacy program for improving body image-related outcomes in adolescent boys and girls. Nutrients 2021 Oct 27;13(11):3825 [FREE Full text] [doi: 10.3390/nu13113825] [Medline: 34836084]

XSL•FO
RenderX

4. Sanzari CM, Gorrell S, Anderson LM, Reilly EE, Niemiec MA, Orloff NC, et al. The impact of social media use on body image and disordered eating behaviors: content matters more than duration of exposure. Eat Behav 2023 Apr;49:101722. [doi: 10.1016/j.eatbeh.2023.101722] [Medline: 37060807]
5. Sidani JE, Shensa A, Hoffman B, Hanmer J, Primack BA. The Association between social media use and eating concerns among US young adults. J Acad Nutr Diet 2016 Sep;116(9):1465-1472 [FREE Full text] [doi: 10.1016/j.jand.2016.03.021] [Medline: 27161027]
6. Au ES, Cosh SM. Social media and eating disorder recovery: an exploration of Instagram recovery community users and their reasons for engagement. Eat Behav 2022 Aug;46:101651. [doi: 10.1016/j.eatbeh.2022.101651] [Medline: 35760017]
7. Kasson E, Vázquez MM, Doroshenko C, Fitzsimmons-Craft EE, Wilfley DE, Taylor CB, et al. Exploring social media recruitment strategies and preliminary acceptability of an mHealth tool for teens with eating disorders. Int J Environ Res Public Health 2021 Jul 28;18(15):7979 [FREE Full text] [doi: 10.3390/ijerph18157979] [Medline: 34360270]
8. National Collaborating Centre for Mental Health. Working well togethervidence and tools to enable co-production in mental health commissioning. National Collaborating Centre for Mental Health. 2019. URL: https://tinyurl.com/mtpdcrx3 [accessed 2024-06-25]
9. Ezaydi N, Sheldon E, Kenny A, Buck ET, Weich S. Service user involvement in mental health service commissioning, development and delivery: a systematic review of service level outcomes. Health Expect 2023 Aug 08;26(4):1453-1466 [FREE Full text] [doi: 10.1111/hex.13788] [Medline: 37292036]
10. Fitzsimmons-Craft EE, Krauss MJ, Costello SJ, Floyd GM, Wilfley DE, Cavazos-Rehg PA. Adolescents and young adults engaged with pro-eating disorder social media: eating disorder and comorbid psychopathology, health care utilization, treatment barriers, and opinions on harnessing technology for treatment. Eat Weight Disord 2020 Dec;25(6):1681-1692 [FREE Full text] [doi: 10.1007/s40519-019-00808-3] [Medline: 31679144]
11. Rocelli M, Aquili L, Giovanazzi P, Puecher A, Goglio MM, Faccio E. 'All about my ideal mental health service': users, family members and experts by experience discussing a co-designed service. Health Expect 2024 Apr;27(2):e13999 [FREE Full text] [doi: 10.1111/hex.13999] [Medline: 38439208]
12. Rocelli M, Sdrubolini F, Romania V, Faccio E. ‘Doctor, I’m not here to quit drugs!’ Hidden goals that undermine shared decision-making during access to services by young people who use substances. Addiction Research & Theory 2023 May 13;32(1):10-19. [doi: 10.1080/16066359.2023.2210840]
13. Mabe AG, Forney KJ, Keel PK. Do you "like" my photo? Facebook use maintains eating disorder risk. Int J Eat Disord 2014 Jul;47(5):516-523. [doi: 10.1002/eat.22254] [Medline: 25035882]
14. Meier EP, Gray J. Facebook photo activity associated with body image disturbance in adolescent girls. Cyberpsychol Behav Soc Netw 2014 Apr;17(4):199-206. [doi: 10.1089/cyber.2013.0305] [Medline: 24237288]
15. Cohen R, Newton-John T, Slater A. The relationship between Facebook and Instagram appearance-focused activities and body image concerns in young women. Body Image 2017 Dec;23:183-187. [doi: 10.1016/j.bodyim.2017.10.002] [Medline: 29055773]
16. McLean S, Paxton S, Wertheim E, Masters J. Selfies and social media: relationships between self-image editing and photo-investment and body dissatisfaction and dietary restraint. J Eat Disord 2015 Nov 23;3(S1):729-735. [doi: 10.1186/2050-2974-3-S1-O21]
17. Faccio E, Belloni E, Cipolletta S, Iudici A, Castiglioni M, Mannarini S. The power of weight and the weight of power in adolescence: a comparison between young and adult women. Journal of Family Studies 2016 Jun 20;25(1):46-60. [doi: 10.1080/13229400.2016.1187660]
18. Wick MR, Keel PK. Posting edited photos of the self: increasing eating disorder risk or harmless behavior? Int J Eat Disord 2020 Jun;53(6):864-872. [doi: 10.1002/eat.23263] [Medline: 32372445]
19. Dane A, Bhatia K. The social media diet: a scoping review to investigate the association between social media, body image and eating disorders amongst young people. PLOS Glob Public Health 2023;3(3):e0001091 [FREE Full text] [doi: 10.1371/journal.pgph.0001091] [Medline: 36962983]
20. Boepple L, Thompson JK. A content analytic comparison of fitspiration and thinspiration websites. Int J Eat Disord 2016 Jan;49(1):98-101. [doi: 10.1002/eat.22403] [Medline: 25778714]
21. de Vries DA, Peter J. Women on display: the effect of portraying the self online on women’s self-objectification. Computers in Human Behavior 2013 Jul;29(4):1483-1489. [doi: 10.1016/j.chb.2013.01.015]
22. Tiggemann M, Miller J. The internet and adolescent girls’ weight satisfaction and drive for thinness. Sex Roles 2010 Apr 28;63(1-2):79-90. [doi: 10.1007/s11199-010-9789-z]
23. Tiggemann M, Slater A. Facebook and body image concern in adolescent girls: a prospective study. Int J Eat Disord 2017 Jan 18;50(1):80-83. [doi: 10.1002/eat.22640] [Medline: 27753130]
24. McLean S, Paxton S, Wertheim E, Masters J. Photoshopping the selfie: self photo editing and photo investment are associated with body dissatisfaction in adolescent girls. Int J Eat Disord 2015 Dec;48(8):1132-1140. [doi: 10.1002/eat.22449] [Medline: 26311205]
25. Ahrens J, Brennan F, Eaglesham S, Buelo A, Laird Y, Manner J, et al. A longitudinal and comparative content analysis of Instagram fitness posts. Int J Environ Res Public Health 2022 Jun 03;19(11):6845 [FREE Full text] [doi: 10.3390/ijerph19116845] [Medline: 35682428]

XSL•FO
RenderX

26. Burnell K, George M, Kurup A, Underwood M, Ackerman. Associations between self-reports and device-reports of social networking site use: an application of the truth and bias model. Commun Methods Meas 2021;15(2):156-163 [FREE Full text] [doi: 10.1080/19312458.2021.1918654] [Medline: 34306291]
27. Maqsood A, Sangra S. The effect of Snapchat filters on self-image: a study of Netflix's 'The Social Dilemma'. Global Media Journal 2021;3(3):1-9 [FREE Full text]
28. Chua T, Chang L. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. Computers in Human Behavior 2016 Feb;55(Part A):190-197 [FREE Full text] [doi: 10.1016/j.chb.2015.09.011]
29. Cohen R, Fardouly J, Newton-John T, Slater A. #BoPo on Instagram: an experimental investigation of the effects of viewing body positive content on young women's mood and body image. New Media & Society 2019 Feb 06;21(7):1546-1564 [FREE Full text] [doi: 10.1177/1461444819826530]
30. McLean SA, Paxton SJ, Wertheim EH. The role of media literacy in body dissatisfaction and disordered eating: a systematic review. Body Image 2016 Dec;19:9-23. [doi: 10.1016/j.bodyim.2016.08.002] [Medline: 27572000]
31. Creswell J. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. New York, NY: SAGE Publications; 2013.
32. Cohen E. Should you 'Friend' your doctor on Facebook? CNN Health. 2009. URL: http://www.cnn.com/2009/HEALTH/09/03/friending.your.doctor/index.html [accessed 2024-06-25]
33. Sabik N, Falat J, Magagnos J. When self-worth depends on social media feedback: associations with psychological well-being. Sex Roles 2019 Jul 4;82(7-8):411-421 [FREE Full text] [doi: 10.1007/s11199-019-01062-8]
34. Hong S, Jahng M, Lee N, Wise K. Do you filter who you are?: Excessive self-presentation, social cues, and user evaluations of Instagram selfies. Computers in Human Behavior 2020 Mar;104:106159 [FREE Full text] [doi: 10.1016/j.chb.2019.106159]
35. Huang Y, Su S. Motives for Instagram use and topics of interest among young adults. Future Internet 2018 Aug 09;10(8):77 [FREE Full text] [doi: 10.3390/fi10080077]
36. Tiggemann M, Anderberg I, Brown Z. Uploading your best self: selfie editing and body dissatisfaction. Body Image 2020 Jun;33:175-182. [doi: 10.1016/j.bodyim.2020.03.002] [Medline: 32224447]
37. Xu B, Chang P, Welker C, Bazarova N, Cosley D. Automatic archiving versus default deletion: what Snapchat tells us about ephemerality in design. CSCW Conf Comput Support Coop Work 2016;2016:1662-1675 [FREE Full text] [doi: 10.1145/2818048.2819948] [Medline: 30294721]
38. Couture Bue AC, Harrison K. Visual and cognitive processing of thin-ideal Instagram images containing idealized or disclaimer comments. Body Image 2020 Jun;33:152-163. [doi: 10.1016/j.bodyim.2020.02.014] [Medline: 32193169]
39. Anixiadis F, Wertheim E, Rodgers R, Caruana B. Effects of thin-ideal instagram images: the roles of appearance comparisons, internalization of the thin ideal and critical media processing. Body Image 2019 Dec;31:181-190. [doi: 10.1016/j.bodyim.2019.10.005] [Medline: 31715553]
40. Casale S, Gemelli G, Calosi C, Giangrasso B, Fioravanti G. Multiple exposure to appearance-focused real accounts on Instagram: effects on body image among both genders. Curr Psychol 2019 Mar 14;40(6):2877-2886 [FREE Full text] [doi: 10.1007/s12144-019-00229-6]
41. Wenninger H, Cheung C, Krasnova H. College-aged users behavioral strategies to reduce envy on social networking sites: a cross-cultural investigation. Computers in Human Behavior 2019 Aug;97:10-23 [FREE Full text] [doi: 10.1016/j.chb.2019.02.025]
42. Wang Y, Yang J, Wang J, Yin L, Lei L. Body talk on social networking sites and body dissatisfaction among young women: a moderated mediation model of peer appearance pressure and self-compassion. Curr Psychol 2020 Mar 16;41(3):1584-1594 [FREE Full text] [doi: 10.1007/s12144-020-00704-5]
43. Jones D, Crawford J. The peer appearance culture during adolescence: gender and body mass variations. J Youth Adolescence 2006 Mar 2;35(2):243-255 [FREE Full text] [doi: 10.1007/s10964-005-9006-5]
44. Collin P, Rahilly K, Third A. The benefits of social networking services: a literature review. Western Sydney University. Melbourne, VIC, Australia: Cooperative Research Centre for Young People, Technology and Wellbeing; 2011. URL: https://www.westernsydney.edu.au/__data/assets/pdf_file/0003/476337/The-Benefits-of-Social-Networking-Services.pdf [accessed 2024-06-25]
45. Bennett BL, Martin-Wagar CA, Boswell RG, Forrest LN, Perelman H, Latner JD. Skepticism of and critical thinking about media messages: conflicting relationships with body dissatisfaction. Eat Behav 2023 Dec;51:101820. [doi: 10.1016/j.eatbeh.2023.101820] [Medline: 37769415]
46. Burnette C, Kwitowski MA, Mazzeo SE. "I don't need people to tell me I'm pretty on social media:" a qualitative study of social media and body image in early adolescent girls. Body Image 2017 Dec;23:114-125. [doi: 10.1016/j.bodyim.2017.09.001] [Medline: 28965052]
47. Singleton A, Abeles P, Smith I. Online social networking and psychological experiences: the perceptions of young people with mental health difficulties. Computers in Human Behavior 2016 Aug;61:394-403 [FREE Full text] [doi: 10.1016/j.chb.2016.03.011]
48. Mitchell K, Ybarra M. Social networking sites: finding a balance between their risks and benefits. Arch Pediatr Adolesc Med 2009 Jan;163(1):87-89. [doi: 10.1001/archpediatrics.2008.534] [Medline: 19124709]

XSL•FO
RenderX

49. Valentine L, McEnery C, D'Alfonso S, Phillips J, Bailey E, Alvarez-Jimenez M. Harnessing the potential of social media to develop the next generation of digital health treatments in youth mental health. Curr Treat Options Psych 2019 Sep 11;6(4):325-336 [FREE Full text] [doi: 10.1007/s40501-019-00184-w]

Abbreviations

ED: eating disorder
SN: social network
SNS: social networking site

Edited by A Mavragani; submitted 04.10.23; peer-reviewed by L Aquili, Y Wang, R Vitelli, A Nerini, A Mentasti, R von Brachel; comments to author 02.02.24; revised version received 29.03.24; accepted 20.05.24; published 02.07.24.

Please cite as:
Faccio E, Reggiani M, Rocelli M, Cipolletta S
Issues Related to the Use of Visual Social Networks and Perceived Usefulness of Social Media Literacy During the Recovery Phase: Qualitative Research Among Girls With Eating Disorders
J Med Internet Res 2024;26:e53334
URL: https://www.jmir.org/2024/1/e53334
doi:10.2196/53334
PMID:

©Elena Faccio, Margherita Reggiani, Michele Rocelli, Sabrina Cipolletta. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Gamified Crowdsourcing as a Novel Approach to Lung Ultrasound Data Set Labeling: Prospective Analysis

Nicole M Duggan[1*], MD; Mike Jin[1,2*], PhD; Maria Alejandra Duran Mendicuti[3], MD; Stephen Hallisey[1], MD; Denie Bernier[1], RDMS; Lauren A Selame[1], MD; Ameneh Asgari-Targhi[3], PhD; Chanel E Fischetti[1], MD; Ruben Lucassen[4], MSc; Anthony E Samir[5], MD, MPH; Erik Duhaime[2], PhD; Tina Kapur[3*], PhD; Andrew J Goldsmith[6*], MD, MBA

[1]Department of Emergency Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[2]Centaur Labs, Boston, MA, United States
[3]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[4]Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands
[5]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States
[6]Department of Emergency Medicine, Lahey Hospital, University of Massachusetts Medical School, Burlington, MA, United States
[*]these authors contributed equally

Corresponding Author:
Nicole M Duggan, MD
Department of Emergency Medicine
Brigham and Women's Hospital
Harvard Medical School
75 Francis Street
NH-2
Boston, MA, 02115
United States
Phone: 1 617 732 5640
Email: nmduggan@bwh.harvard.edu

Abstract

Background: Machine learning (ML) models can yield faster and more accurate medical diagnoses; however, developing ML models is limited by a lack of high-quality labeled training data. Crowdsourced labeling is a potential solution but can be constrained by concerns about label quality.

Objective: This study aims to examine whether a gamified crowdsourcing platform with continuous performance assessment, user feedback, and performance-based incentives could produce expert-quality labels on medical imaging data.

Methods: In this diagnostic comparison study, 2384 lung ultrasound clips were retrospectively collected from 203 emergency department patients. A total of 6 lung ultrasound experts classified 393 of these clips as having no B-lines, one or more discrete B-lines, or confluent B-lines to create 2 sets of reference standard data sets (195 training clips and 198 test clips). Sets were respectively used to (1) train users on a gamified crowdsourcing platform and (2) compare the concordance of the resulting crowd labels to the concordance of individual experts to reference standards. Crowd opinions were sourced from DiagnosUs (Centaur Labs) iOS app users over 8 days, filtered based on past performance, aggregated using majority rule, and analyzed for label concordance compared with a hold-out test set of expert-labeled clips. The primary outcome was comparing the labeling concordance of collated crowd opinions to trained experts in classifying B-lines on lung ultrasound clips.

Results: Our clinical data set included patients with a mean age of 60.0 (SD 19.0) years; 105 (51.7%) patients were female and 114 (56.1%) patients were White. Over the 195 training clips, the expert-consensus label distribution was 114 (58%) no B-lines, 56 (29%) discrete B-lines, and 25 (13%) confluent B-lines. Over the 198 test clips, expert-consensus label distribution was 138 (70%) no B-lines, 36 (18%) discrete B-lines, and 24 (12%) confluent B-lines. In total, 99,238 opinions were collected from 426 unique users. On a test set of 198 clips, the mean labeling concordance of individual experts relative to the reference standard was 85.0% (SE 2.0), compared with 87.9% crowdsourced label concordance (P=.15). When individual experts' opinions were compared with reference standard labels created by majority vote excluding their own opinion, crowd concordance was higher than the mean concordance of individual experts to reference standards (87.4% vs 80.8%, SE 1.6 for expert concordance; P<.001). Clips with discrete B-lines had the most disagreement from both the crowd consensus and individual experts with the expert

XSL•FO
RenderX

consensus. Using randomly sampled subsets of crowd opinions, 7 quality-filtered opinions were sufficient to achieve near the maximum crowd concordance.

Conclusions: Crowdsourced labels for B-line classification on lung ultrasound clips via a gamified approach achieved expert-level accuracy. This suggests a strategic role for gamified crowdsourcing in efficiently generating labeled image data sets for training ML systems.

(J Med Internet Res 2024;26:e51397) doi:10.2196/51397

KEYWORDS

crowdsource; crowdsourced; crowdsourcing; machine learning; artificial intelligence; point-of-care ultrasound; POCUS; lung ultrasound; B-lines; gamification; gamify; gamified; label; labels; labeling; classification; lung; pulmonary; respiratory; ultrasound; imaging; medical image; diagnostic; diagnose; diagnosis; data science

Introduction

Background

Machine learning (ML) models can improve medical diagnostic concordance and streamline health care processes [1]. This is particularly true when applied to medical image analysis. ML models require large-scale labeled data sets for model training [2]. Widespread ML tool development is limited by the need to acquire high-quality labels given the associated expert time and cost [3-7].

Combining opinions from multiple individuals on a given task can produce more accurate interpretations than from a single individual [8]. Crowdsourcing, the process of collecting large numbers of unskilled opinions, can improve efficiency, lower costs, and offer high quality in repetitive task completion [8,9]. Crowdsourced approaches to data set labeling are growing in popularity, and beneficial effects of crowdsourcing have been demonstrated in health care–related tasks including biomedical imaging analysis [10-14].

Using crowdsourcing for biomedical image labeling is challenged by the complexity of the tasks and the need to ensure label quality control. The user interface design for collecting crowd opinions and the metrics used for assessing opinion quality is key to successful results. Gamification, the persuasive system design that uses game-like tasks to engage participants competitively for rewards, can both encourage crowd participation and improve performance accuracy by selectively rewarding top users [15-18]. Combining crowdsourcing with gamification can become a performance measurement tool for identifying top crowd labelers [18-21].

Point-of-care ultrasound (POCUS) is a dynamic medical imaging technique used at patients' bedside to make accurate, real-time diagnoses [22-24]. Though POCUS has significant value in health care settings, advanced training is required to accurately apply this tool to clinical care [25]. As such, ML models that automate POCUS image interpretation hold exceptional potential clinical value. In lung POCUS, B-lines are hyperechoic linear artifacts that extend from the pleural line and appear dynamically with the respiratory cycle [23]. B-lines are known markers of pulmonary congestion and their presence, quantity, and thickness (discrete vs confluent B-lines) correlate with pathological severity of conditions such as congestive heart failure exacerbations, pneumonia, or inflammatory lung disease [23,26,27]. Early ML models for lung POCUS have shown promise in identifying and quantifying the presence of B-lines [28-33]. However, existing models have limited accuracy and generalizability due in part to the lack of large-scale, high-quality, labeled image databases for model training. Gamified crowdsourcing has not previously been applied to annotate lung POCUS clips or ultrasound imaging data overall.

Objectives

We examined whether a gamified crowdsourcing approach with inbuilt quality control measures can classify lung POCUS clips for presence and type of B-lines at comparable concordance to trained ultrasound experts. Our secondary objectives were to measure (1) the number of crowd opinions required to achieve maximal concordance, and (2) learning curves for expert and crowd individuals over time.

Methods

Study Design and Setting

This was a prospective analysis performed using retrospectively collected lung POCUS clips. All lung POCUS examinations performed in an academic tertiary care emergency department (ED) between March 1, 2020, and February 28, 2022, were retrospectively queried via the eHealth record.

Data Set Curation

In total, 2391 POCUS clips were downloaded in Digital Imaging and Communications in Medicine format from the hospital Picture Archiving and Communication System, and 6-second clips were acquired at frame rates between 15 and 46 Hz. Digital Imaging and Communications in Medicine files were converted to MP4 format using an open-source medical image viewer and subsequently deidentified using a software package [34,35].

The 2391 clips were randomly divided by patient into 2 sets: data set A (102 patients, 1271 clips) and data set B (101 patients, 1120 clips). A total of 200 random clips from data set A were selected as a crowd training set, and 200 random clips from data set B were selected as a test set to evaluate crowd label quality. A total of 5 training set clips and 2 test set clips were excluded for being flagged by at least 1 expert as not containing lung (Figure 1A).

XSL•FO
RenderX

Figure 1. Data set processing flow diagram. (A) Complete data set partitioning. All 2391-lung point-of-care ultrasound (POCUS) clips from the complete data set were divided evenly by the patient into data set A and data set B. A total of 200 clips from each sub-data set were randomly selected to create a training set and test set, respectively. A total of 5 clips from the training set and 2 clips from the test set were excluded for having no lung visible in the clip prior to deploying these sets to the crowd. (B) Lung POCUS clip labeling workflow. Of the complete data set, 400 clips (200 from data set A and 200 from data set B) were separated to become the training set, used for giving performance feedback to the crowd, and the test set, used to evaluate crowd performance and concordance. The remaining clips from the complete data set were not labeled by experts but instead were available to achieve a crowd-consensus label. Crowd users were given additional performance feedback on clips that had previously achieved a crowd-consensus label.

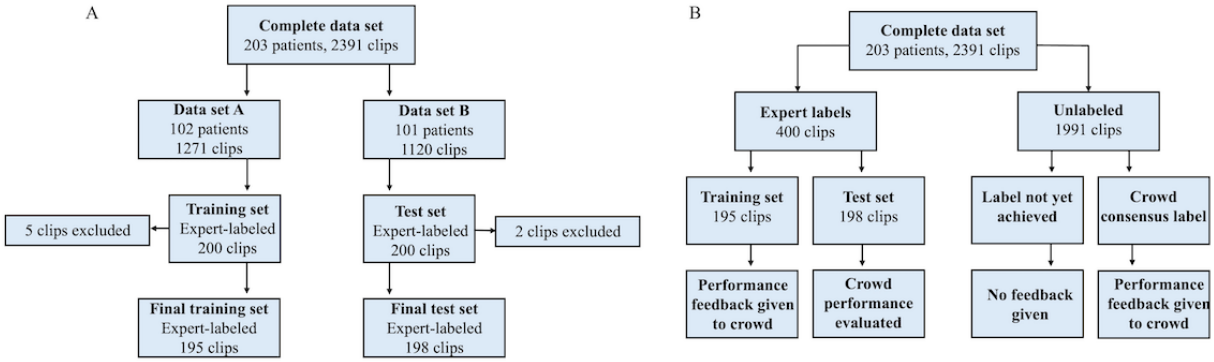

Task Definition

Expert and crowd users were asked to classify B-lines on lung POCUS clips into one of three classes: (1) no B-lines, (2) one or more discrete B-lines, or (3) confluent B-lines (Figure 2). Clips were classified based on the highest B-line severity present throughout the entire clip, with no B-lines < one or more discrete B-lines < confluent B-lines. Discrete B-lines were defined as hyperechoic lines originating from the pleural line, demonstrated sliding with the pleura, and extending to the bottom of the sonographic field [36,37]. Confluent B-lines were defined as hyperechoic sections originating from the pleural line, demonstrated sliding with the pleura, and had thickness along the pleura beyond that of discrete B-lines [36,37].

A total of 6 experts with advanced training in lung POCUS (4 ultrasound fellowship-trained emergency medicine physicians including 1 dual board-certified in intensive care; 1 emergency radiologist, and 1 registered diagnostic medical sonographer) were recruited. All 6 experts provided independent classification opinions for all training and test set clips via DiagnosUs, a free iOS app where users compete in medical data labeling contests to win cash prizes based on their labeling concordance. All clips could be continuously played or viewed frame-by-frame as many times as needed. Prior to participation, experts underwent a 30-minute training session familiarizing themselves with the app and ensuring agreement on the study-specific classifications.

Figure 2. Representative frames from lung point-of-care ultrasound (POCUS) images demonstrating the 3 potential classification label categories. (1) Two lung POCUS images demonstrating no B-lines; (2) Two lung POCUS images demonstrating discrete B-lines; (3) Two lung POCUS images demonstrating confluent B-lines.

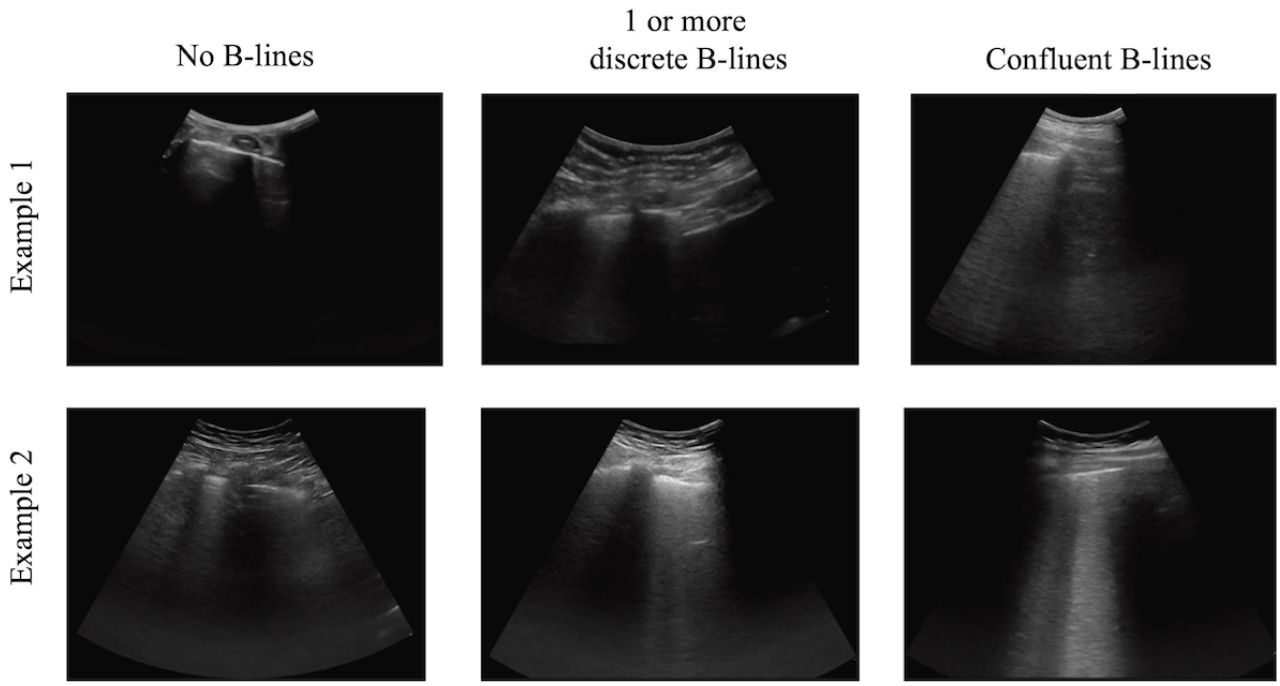

XSL•FO
RenderX

Reference Standard Calculations

Reference standard labels on training and test set clips were assigned using expert consensus—the majority rule of the 6 experts' opinions, with ties broken randomly. For the single reference standard based on all 6 experts' opinions, the average expert concordance with the reference standard was calculated as the average of the 6 individual expert concordances with the reference standard.

A collection of 6 "leave-one-out" reference standard label sets was created based on 5 experts each, 1 for each expert whose input was excluded from the majority rule consensus. This was done to measure each expert against a reference standard which their own opinion did not influence. For concordances with respect to leave-one-out reference standards, an individual expert's concordance was the proportion of their opinions that matched their respective set of leave-one-out reference standard labels (the set of labels excluding their own opinion), and the crowd concordance was the proportion of the 6 leave-one-out reference standard labels matched by the crowd consensus label. The overall crowd concordance was the average of the per-clip crowd concordances.

Gamified Crowdsourcing

Crowd opinions were collected using DiagnosUs via gamified contests. Crowd users included anyone in the general public with access to an iOS device who downloaded the iOS app. There were no criteria regarding the level of expertise to participate. Users voluntarily participated in labeling contests based on interest and the potential to earn rewards.

Crowd users were trained in two ways as follows: (1) optional tutorial cases with accompanying layperson explanations which they could access at any time and (2) after submitting an opinion, immediate feedback was given via revelation of the current label (ie, the reference standard label on training set clips, or the current crowd consensus label if one exists). Crowd consensus labels were computed and assigned for all clips using the majority rule of the top crowd labelers' opinions, and were assigned to any of the 1991 initially unlabeled clips once sufficient mutual agreement was reached among the top crowd labelers for that clip (Figure 1B). Top labelers were identified via continuous performance monitoring based on their trailing concordance on clips with labels. Clips were shown to users in random order so that they could not predict which clips would impact their contest score, and any test clip could be shown to a user multiple times. Additional details regarding the mechanics of crowd opinion collection and aggregation are provided in Multimedia Appendix 1.

Gamified Crowdsourcing Contest Mechanics

Users participating in classification contests were shown lung POCUS clips in random order. After submitting their opinion on a clip, they were either shown the correct label as feedback (on feedback clips) or a message thanking the user for their opinion (on nonfeedback clips). Feedback clips consisted of the 195 training set clips with reference standard labels, as well as any of the 1991 unlabeled clips that had acquired a crowd consensus label (Figure 1B). Feedback clips shown to the user were sampled to be equally likely to have a correct label in each of the 3 B-line classifications. Nonfeedback clips included the 198 test clips with reference standard labels, as well as any of the 1991 unlabeled clips that had yet to reach crowd consensus. Due to randomized clip ordering, users did not know which clips were feedback clips, and therefore, which clips would be used to grade their labeling performance for prizes. The live contest was run until all unlabeled clips reached crowd consensus.

Crowd Consensus Labeling and Quality Assessment

Crowd consensus (ie, the B-line classification with the most crowd votes, with ties broken randomly) was achieved for an initially unlabeled clip when either (1) a minimum 3-vote difference between the highest and second highest-voted classification choice was achieved, or (2) 15 opinions were collected for the clip. If a user contributed multiple opinions to the same clip, only their most recent opinion was considered. Once crowd consensus was reached on an unlabeled clip, the clip acquired a crowd consensus label and became a feedback clip.

Crowd label quality was assessed by comparing how well crowd consensus labels and individual expert opinions respectively matched the reference standard labels on the test set. Both of these were measured by concordance, calculated simply as the percent of matching labels across the test set clips. In addition, we also calculated concordances with "leave-one-out" reference standards.

A crowd user's opinion quality score (Qscore) was defined as the user's trailing average concordance (with respect to reference standards for training clips and crowd consensus labels for initially unlabeled clips) on the last 50 feedback clips they provided an opinion on. Each user's Qscore was dynamically tracked over time and was updated with each opinion they gave. If a user's Qscore was less than 80% at the time they gave an opinion on an individual clip, that opinion was not considered for determining the crowd consensus for that clip and instead was discarded. Users were assigned a Qscore of 0 until they provided at least 25 opinions on feedback clips. This ensured only the highest quality crowd opinions contributed to crowd consensus labels.

Analysis of Secondary Outcomes

To assess the minimum number of crowd opinions needed to achieve expert-level concordance, we estimated the concordance that the crowd labels would have had if we had collected fewer opinions on each clip by repeatedly sampling fewer opinions from all test clips and computing concordance with the reference standard based on the resulting crowd consensus labels. Just as crowd consensus labels were determined as described above ("Crowd Consensus Labeling and Quality Assessment"), only opinions submitted by users whose Qscore was at least 80% at the time of opinion submission were considered when sampling. For every opinion count, we used 1000 Monte Carlo samples per clip to estimate concordance.

To calculate learning curves, the concordance of individual crowd users or experts with the reference standard was computed each time the individual gave an opinion on a test set clip, based on the proportion of their last 25 opinions on test set clips that

XSL•FO
RenderX

matched the reference standard. The concordance at each opinion submission was averaged across individuals (all crowd users, skilled crowd users, or all experts) to produce the learning curves. Skilled crowd users were users who submitted at least 1 opinion on at least 1 test clip when they had a trailing average concordance (with respect to reference standards for training clips and crowd consensus labels for initially unlabeled clips) of 80% or better. For this analysis, all opinions from crowd users were considered, including multiple opinions per user per clip, if applicable.

Statistical Analysis

Analysis was performed using Python (version 3.10; Python Software Foundation) [38]. The difference between crowd concordance and average individual expert concordance was tested for significance using a paired-samples *t* test. For concordance based on the 6-expert consensus, the crowd concordance for each case was 1 if the crowd label was correct and 0 otherwise, and the average expert concordance for that case was the proportion of the 6 experts that were correct. For leave-one-out concordance based on 5-expert consensus, the crowd concordance for each case was the proportion of the 6 leave-one-out consensus labels that matched the crowd label, and the average expert concordance for that case was the proportion of the 6 experts that matched their respective leave-one-out consensus label. All mean calculations are reported as mean (SE).

Ethical Considerations

This study protocol was approved by the Mass General Brigham institutional review board (protocol number 2021P001446). Primary lung ultrasound data set creation was considered exempt from obtaining patient consent by the institutional review board given that the data set was collected retrospectively from the medical center's clinical image archive and all data were deidentified prior to use in this study. DiagnosUs users provided consent for their data to be used for research purposes prior to engaging in any labeling contests. No compensation was provided for participating in this study.

Results

Data Set Characteristics

The patients who contributed to the lung POCUS database had a mean age of 60.0 (SD 19.0) years, and 105 (51.7%) patients were female, 43 (21.2%) patients were Hispanic, 42 (20.7%) patients were Black, and 114 (56.1%) patients were White. From the ED, 122 (64%) of patients were admitted to the hospital floor, 57 (28.6%) were discharged home, 13 (6.4%) were admitted to the intensive care unit, 2 (1%) went directly to the operating room, 9 (4.4%) had an alternate disposition, and 0 (0%) expired in the ED (Table 1).

Table 1. Patient characteristics from the lung point-of-care ultrasound data set used for image annotation. Clips were queried retrospectively via the eHealth record from clinical lung ultrasound examinations performed for patients presenting to the emergency department between March 1, 2020, and February 28, 2022.

Characteristic	Subjects, (N=203), n (%)
Sex	
Male	98 (48.3)
Female	105 (51.7)
Ethnicity	
Hispanic or Latino	43 (21.2)
Non-Hispanic or Latino	160 (78.8)
Race	
American Indian or Alaskan Native	0 (0)
Asian	9 (4.4)
Black or African American	42 (20.7)
Caucasian	114 (56.2)
Native Hawaiian or Pacific Islander	0 (0)
Other	38 (18.7)
Emergency department disposition	
Discharge home	57 (28.6)
Floor admission	122 (64.0)
Intensive care unit admission	13 (6.4)
Operating room	2 (1.0)
Other	9 (4.4)

XSL•FO
RenderX

Opinion Collection

Experts spent an average of 1.7 (range 0.9-2.5) hours submitting opinions for all 393 training and test clips (3.9 opinions per minute on average). Over the 195 training clips, the reference standard label distribution (based on the experts' majority opinion) was 58% (n=114) no B-lines, 29% (n=56) discrete B-lines, and 13% (n=25) confluent B-lines. Over the 198 test clips, the reference standard label distribution was 70% (n=138) no B-lines, 18% (n=36) discrete B-lines, and 12% (n=24) confluent B-lines.

In total, 99,238 crowdsourced opinions were collected from 426 unique users across all 2384 clips. The number of users contributing an opinion to each test set clip ranged from 28 to 48. Of these, 34,363 opinions from 114 unique users were eligible to be considered for crowd consensus labels based on our quality thresholds. Of the 114 users who contributed to crowd consensus labels, 53% (n=60) had prior medical experience, compared to 45% (n=190) of users overall. A total of 56,874 (57.3%) opinions were from users who reported prior medical experience, as were 22,231 (64.7%) opinions eligible for consideration for crowd consensus labels. The live contest was launched over 138 hours which reflected a mean acquisition rate of 12.0 opinions per minute. The total cash prize payout throughout the entire competition was US $1100. The maximum prize earned by an individual user was US $25.

Label Concordance With Reference Standard

The 6 experts' concordances on the 198 test clips relative to the reference standard were 77.2%, 81.3%, 84.8%, 87.3%, 88.4%, and 90.9%, with a mean of 85.0 (SD 2.0). Comparatively, the crowd concordance on these clips was 87.9% relative to the reference standard (*P*=.15; Figure 3A). When individual expert concordances were computed against "leave-one-out" reference standards, expert concordances on the same 198 test clips were 75.8%, 77.8%, 79.8%, 81.8%, 83.3%, and 86.4% with a mean of 80.8 (SD 1.6), compared with a crowd concordance of 87.4% (*P*<.001; Figure 3B).

For clips designated by the reference standard as having no B-lines, experts had an average concordance with the reference standard of 91.5% (SD 2.3), compared with a crowd concordance of 99.3% (*P*<.001). For cases with discrete B-lines, experts had an average concordance of 63.9% (SD 13.2) compared with the crowd concordance of 50% (*P*=.09). For cases with confluent B-lines, expert average concordance (mean 79.2%, SD 6.5) and crowd concordance were both 79.2% (*P*>.99). The balanced multiclass concordance (ie, average per-class concordance) was 76.1% for the crowd and 78.2% (SD 4.0) for individual experts relative to the reference standard, and 76.9% for the crowd and 71.2% (SD 3.2) for individual experts relative to leave-one-out reference standard.

Using randomly sampled subsets of collected opinions, 7 quality-filtered opinions were sufficient to achieve near the maximum crowd concordance (Figure 4).

Figure 3. Expert and crowd label concordance with reference standards on test set clips stratified by B-line classification. (A) Expert and crowd label concordance for each B-line classification type compared with the reference standard comprised of consensus of 6 experts. Overall, on all clips combined, there was no difference in the crowd concordance of all clips compared with individual experts' concordance (*P*=.15). For clips designated by the reference standard as having no B-lines, crowd concordance exceeded average expert concordance compared with a reference standard (*P*<.001). (B) Expert and crowd label concordance for each B-line classification type compared with the reference standard comprised of consensus of 5 experts with the opinion of the individual expert being tested left out (ie, leave-one-out reference standard). When compared with the leave-one-out reference standard, crowd concordance exceeded expert concordance both for all B-line classifications combined, as well as for clips with no B-lines. **P*<.001.

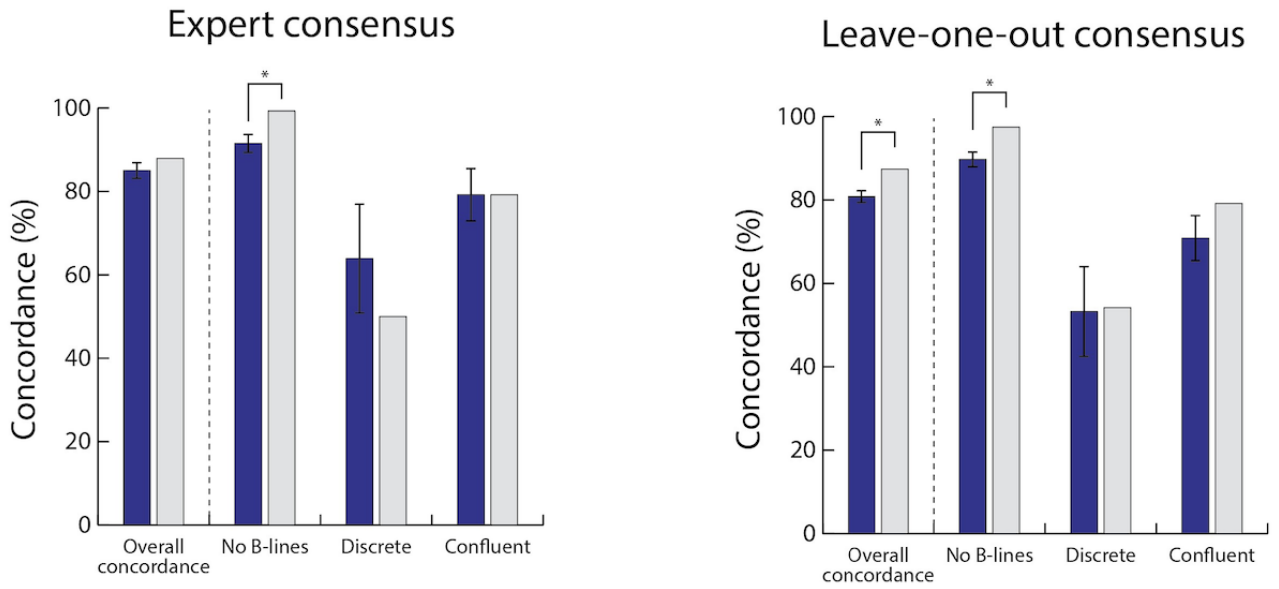

XSL•FO
RenderX

Figure 4. Determining the number of crowd opinions needed to maximize crowd-consensus concordance using randomly sampled subsets of collected opinions. The solid blue line indicates estimated crowd-consensus concordance as dependent on the number of crowd opinions collected. The vertical dotted line indicates that 7 crowd opinions are sufficient to achieve the desired crowd concordance, and the data suggest that further opinion collection shows diminishing marginal increases in crowd consensus concordance. Observed crowd concordance (horizontal dotted blue line) indicates the overall crowd-consensus concordance from using all crowd opinions on each clip. Average expert concordance (horizontal dotted red line) indicated the combined expert concordance. Error bars indicate the SE of the mean.

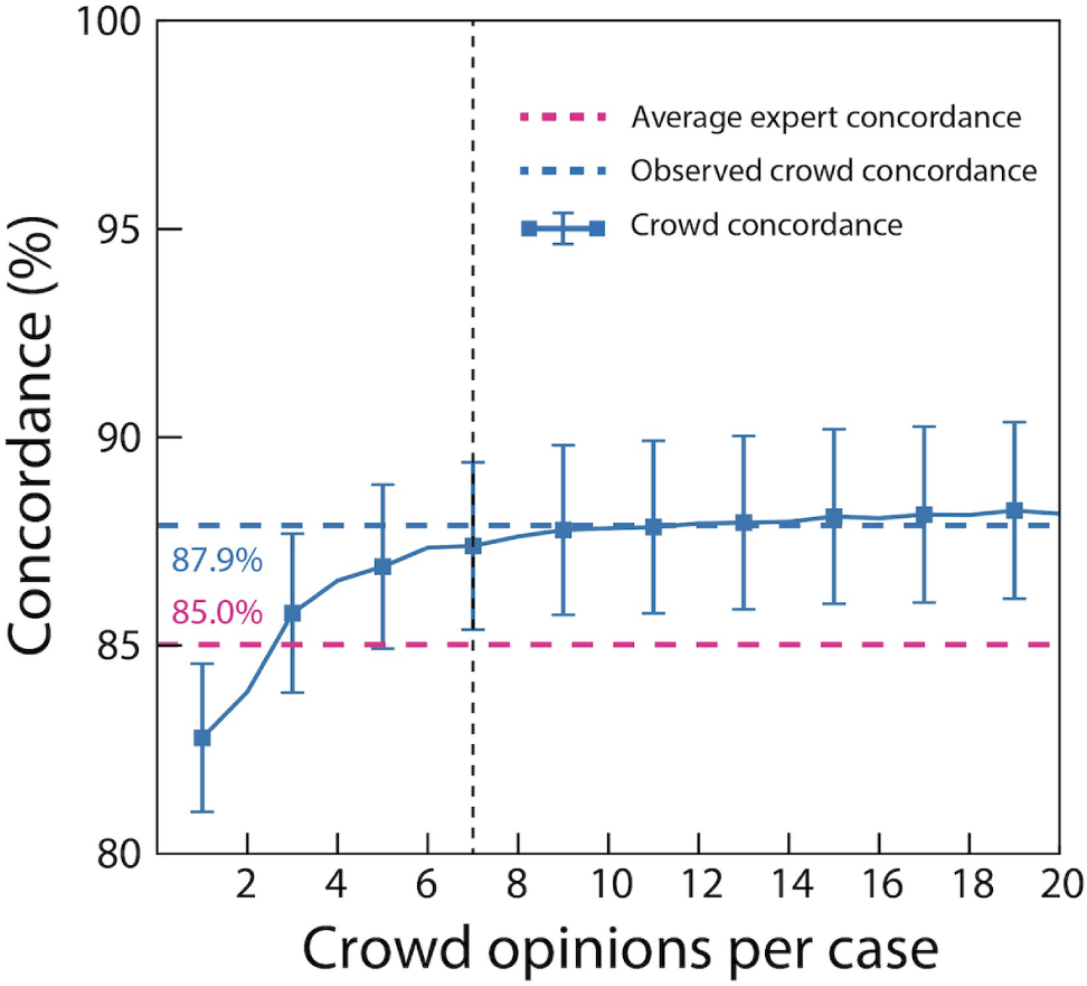

Expert and Crowd Labeling Disagreements

Expert opinions were unanimous on 50.3% (n=98) of training cases and 52.5% (n=104) of test cases. There was an expert supermajority (ie, two-thirds) on 89.2% (n=174) of training cases and 87.9% (n=174) of test cases. On average, 41.2 (SD 4.1) crowd opinions contributed toward the crowd consensus label for each of the test cases. Crowd opinions were unanimous in 35.4% of cases, and a crowd supermajority existed for 85.9% (n=170) of test cases.

The receiver operating characteristic curves based on using different thresholds of crowd opinion proportions to predict the expert-consensus label produced high area under the curve values for all B-line classifications: the area under the curve was 0.98 for the “No B-lines” classification, 0.95 for “discrete B-lines,” and 0.98 for “confluent B-lines” (Figure 5).

Clips with discrete B-lines had the most disagreement from both the crowd consensus and individual experts with the expert consensus, making up 75% (n=18) of cases where the crowd consensus differed from the expert consensus reference standard (Figure 6).

On a clip-by-clip basis, the level of internal crowd agreement and internal expert agreement on each test clip was significantly correlated (Pearson r=0.70, P<.001). When considering only test clips where the crowd opinions were at least 80% in agreement, the crowd consensus label concordance with the reference standard was 96%. The expert agreement was significantly higher on test clips where the crowd label matched the reference standard (89%) versus on test clips where it did not (60%; P<.001, Mann-Whitney U test).

XSL•FO
RenderX

Figure 5. Receiver operating characteristic (ROC) curves for crowd consensus labels relative to the expert-consensus label by B-line classification using different thresholds of crowd opinion proportions to predict the expert-consensus label. ROC curves for all B-line classifications produced a high AUC. Sensitivity and specificity for individual experts are plotted as green points, and the average sensitivity and specificity of all 6 experts are shown as a black cross. (A) ROC curve for clips with no B-lines, (B) ROC curve for clips with discrete B-lines, and (C) ROC curve for clips with confluent B-lines. AUC: area under the curve.

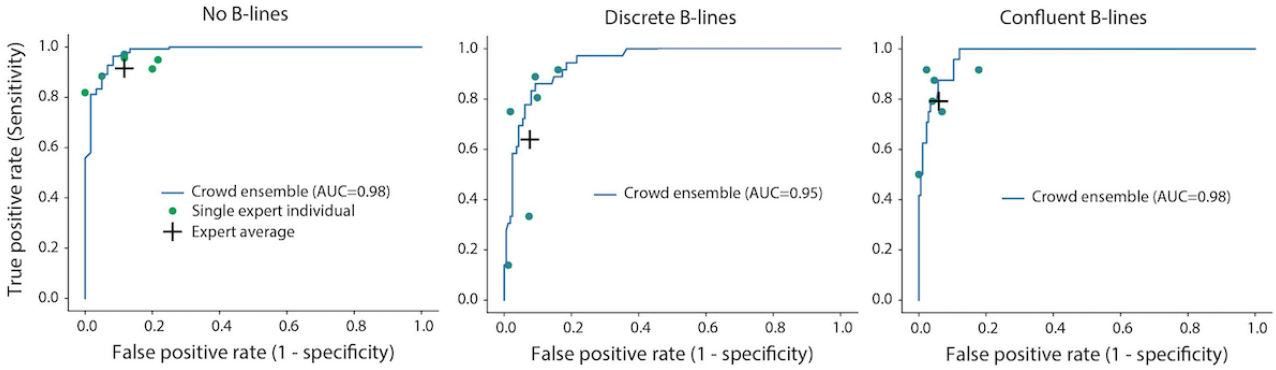

Figure 6. Confusion matrices showing disagreement by category between the expert consensus (A), the crowd consensus, and (B) the average of individual experts. Clips with discrete B-lines had the most disagreement from both the crowd consensus and individual experts with the expert consensus, making up 75% of cases where the crowd consensus differed from the expert consensus reference standard.

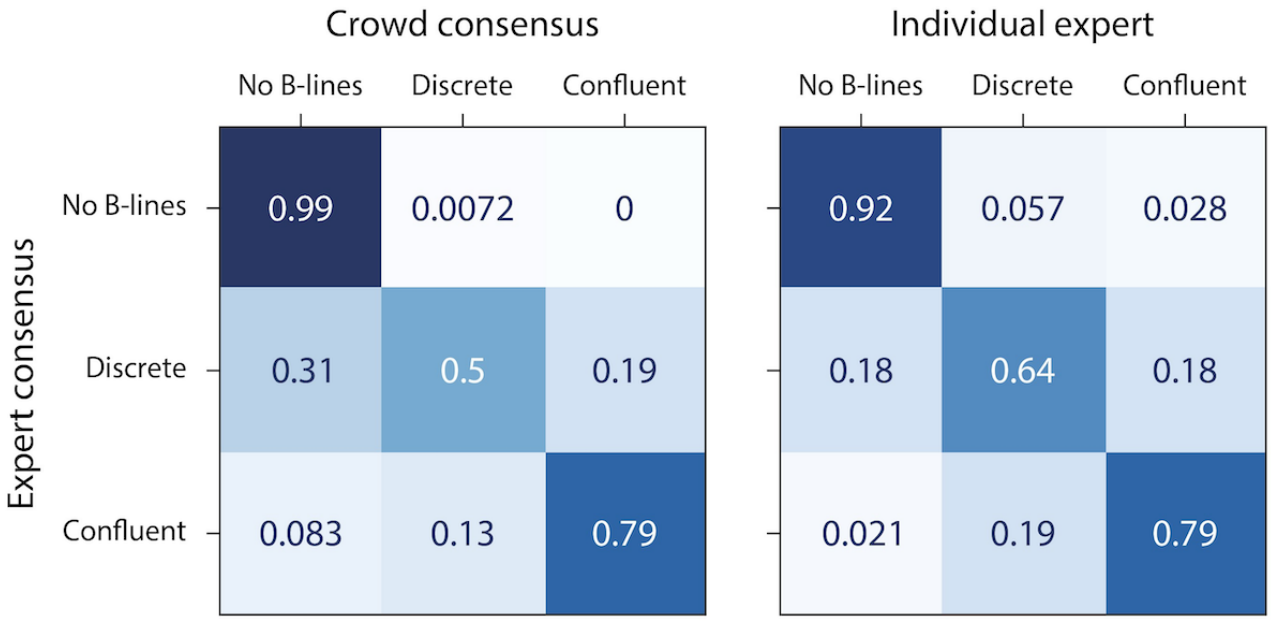

Expert and Crowd Learning

Individual crowd users showed improved concordance with the reference standard over time as each user gave opinions on more clips and thus received more feedback. Individual crowd users reached a final average concordance of 80%-81% after seeing around 75 test set clips (Figure 7). Experts maintained a relatively constant concordance level with the reference standard regardless of the number of test cases seen. The subset of skilled crowd users (n=114) showed an intermediate effect, having a slightly higher concordance than the overall crowd at all times and a less dramatic concordance increase, but still reaching a maximum individual concordance of around 81%.

XSL•FO
RenderX

Figure 7. Learning curves for experts and the crowd separated into all individuals and individuals who were designated as skilled based on a user opinion quality score of >80% as measured by trailing average concordance relative to the reference standard, plotted versus the number of test set cases seen. A window of 25 cases was used to compute trailing average concordance. Individual crowd users showed improved concordance with the reference standard over time as each user gave opinions on more clips and thus received more feedback, reaching a final average concordance of 80%-81% after seeing around 75 test set clips. Error bars represent the SE of the mean.

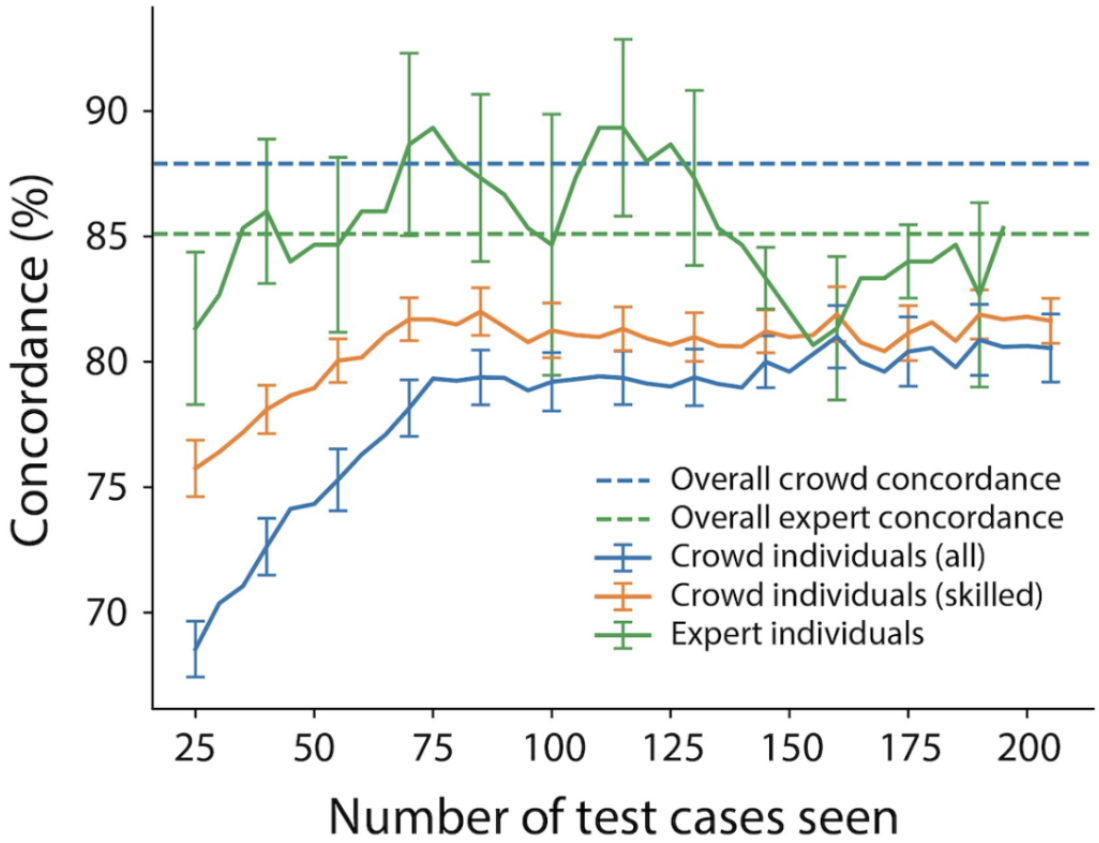

Discussion

Principal Findings

Our data suggest that gamified crowdsourcing can offer an efficient means of obtaining expert-quality labels for B-line classification on lung POCUS clips. While our crowdsourcing contests are open to and attract users of all skill levels from the general population, our approach of filtering highest performing opinions from the crowd allows us to collate a group of semiskilled labelers whose collective intelligence performs on par with, or in some cases outperforms, individual credentialed experts. This method offers a more efficient means of creating labels on medical imaging data than traditional expert-derived approaches. A filtering Qscore of 80% was selected in this work as this gives a reasonable balance between concordance and label acquisition speed based on preliminary data. Future studies could focus on identifying the lowest possible filtering Qscore at which crowd performance continues to outpace that of experts for a given labeling task. This may improve the efficiency of medical data labeling further and could help identify a skill threshold by which crowd input can be filtered while still achieving expert-level concordance. Individual expert accuracies for classifying B-lines ranged considerably. This finding could have 2 possible explanations. First, given that medical imaging data interpretation is often complex, it may be that a proportion of clips have inherent ambiguity that even experts disagree on. This is consistent with previously published work that shows expert interrater agreement for identifying B-lines while substantial, is often imperfect [39-42]. Further, crowd inaccuracy and internal disagreement correlated with expert internal disagreement. Thus, it is possible that clips demonstrating crowd "inaccuracy" may actually be clinically equivocal rather than reflecting poor skill by the crowd. Inherent ambiguity is likely a theme that extends across medical imaging data beyond lung POCUS and may represent a challenge for imaging database labeling overall.

Variability in expert concordance could also be attributable to variable expert baseline skill. Consistent with existing literature that uses medical experts to label POCUS images, our experts all had either fellowship-level training or advanced certification in interpreting lung ultrasound, as well as years of clinical experience [28-33]. Ground truth labeling for training POCUS-based ML models is commonly derived from a small handful of experts (typically 1-5 individuals). Thus, this work combining opinions from 6 experts (or 5 experts in the case of "leave-one-out consensus") to form our reference standard is consistent with accepted practices. Currently, there is no established method for defining ground truth in B-line identification beyond expert opinion. Given the recent widespread adoption of lung POCUS globally and the recognized utility of B-lines as a clinical disease marker, our work highlights the critical need for clarifying how ground truth interpretation of lung POCUS and POCUS overall is defined.

XSL•FO
RenderX

Individual crowd users improved by their opinions converging toward the reference standard over the course of contest participation whereas expert concordance with the reference standard remained high throughout. This is expected as experts should not gain significant benefits beyond their baseline skill level from additional cases and thus are not expected to demonstrate a learning curve. This suggests that nonexpert users can dynamically adapt to task-specific instructions or features based on real-time feedback. Even with tutorial- and feedback-based learning, individual crowd users generally did not improve to the concordance level of the average individual expert, but the combination of crowd users was able to exceed individual expert concordance. This illustrates the "wisdom of the crowd" effect of gaining concordance from incorporating multiple opinions from users who are individually less skilled than experts to reach a concordance level equal to that of experts [43].

While commercial crowdsourcing platforms that allow users to perform discrete, repetitive tasks, such as data labeling, exist, our approach increased labeling concordance and streamlined worker skill assessment via gamification. Gamified contests are structured to ensure crowd users put forth maximal effort, and offer users the opportunity for skill enhancement and learning via immediate feedback. Contest mechanics dynamically estimate user skill levels even as they fluctuate with learning from a subset of expert-annotated clips. The combination of continual user learning and selectively filtering for the most skilled users' opinions enhances labeling concordance. While this approach does initially require a small number of expert-quality labels for crowd training, this input is a significant improvement in expert effort over what is commonly required for data set labeling today. The novelty in this work is not through crowdsourcing alone, but rather in incorporating a gamified approach that encourages optimal user performance through a system of rewards. Using inbuilt quality control metrics in our platform, we are then able to filter the highest-performing labelers from the general crowd to curate a set of labels that have a level of concordance that exceeds that of the general participating crowd. Incorporating both gamification and quality filtering techniques into a crowdsourced task is what allows us to efficiently generate expert-quality labels.

The ultimate goal of our work is to identify strategies for scalable and accurate medical data labeling. When crowd opinions differed from each other, the split in crowd votes was a useful predictor of the expert consensus label. Taken together with possible inherent ambiguity in some of the clips, this highlights a possible triage approach to data set labeling using gamified crowdsourcing. Crowd opinion for clips with a high degree of crowd agreement would be accepted as truth, and expert review would only be necessary for cases where crowd agreement drops below a certain threshold. This approach could significantly decrease the proportion of clips requiring expert review and optimize both the time and cost associated with current expert-based data set labeling approaches.

Limitations

Our data set had an oversampling of clips with no lung pathology with more than 50% of clips in both the training and test data sets containing no B-lines. Since the crowd demonstrated higher concordance with reference standards than individual experts on clips with no B-lines but performed worse than experts on classifying discrete B-lines, it is possible that the crowd may be less efficient at identifying subtle diagnostic findings, but we are not adequately seeing this trend due to data set bias. While the similar or higher balanced multiclass concordance of the crowd compared with individual experts supports the conclusion that crowd performance was on par with experts, future work will apply more balanced data sets.

While this labeling iOS app is available to any user, users with medical backgrounds (medical students, etc) may be more likely to engage in this activity. We do have a general understanding of the proportion of our users with medical backgrounds, however, the precise demographic breakdown of prior medical experiences was poorly defined given that reports of demographic data from users were voluntary in the current platform. It is possible that our crowd is not representative of the general population and may be more consistent with a population of semiskilled labelers. The generalizability of our findings across variable crowd populations with clearly defined experience levels will need to be explored further. A breakdown of crowd performance by background medical expertise level will be helpful in future studies.

Our findings do show promise in streamlining the labeling of lung POCUS data; however, this work may not be generalizable to other more complex medical data labeling tasks. The next steps are to apply this approach to similar questions in lung POCUS data such as segmenting B-lines or evaluating for alternative findings beyond B-lines in lung POCUS. This will help us understand more about the generalizability of our labeling approach.

Conclusions

We demonstrate that gamified crowdsourcing with strategic inbuilt quality control measures can produce B-line classification labels that match expert consensus better than individual experts themselves. These methods illustrate general strategies for improving the reliability of crowdsourced opinions on the task of classifying B-lines on lung POCUS clips. Using innovative and scalable approaches to generate high-quality labeled image databases could contribute to streamlining ML model development which could help either standardize or automate lung POCUS interpretation in the future. Future work can validate whether labels produced by crowdsourcing that achieve expert quality based on standard concordance metrics when used as training data, can produce models with comparable performance to those trained using expert-sourced training data.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the Massachusetts Life Sciences Fund awarded to TK, AJG, NMD, and ED as well as an NIH grant (R21EB034075) awarded to TK, AJG, and NMD for supporting this work.

Data Availability

The data sets analyzed during this study are available in the Harvard Dataverse repository [43]. Annotation data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

NMD, MJ, ED, TK, and AJG conceived the study, designed the trial, obtained research funding, managed data acquisition, analysis, and interpretation, and drafted the manuscript. MADM, SH, DB, and LAS contributed to data acquisition, analysis, and interpretation. MJ, RL, and TK were responsible for statistical analysis. NMD and MJ created all the figures and tables. CEF, AES, and RL provided administrative, technical, or material support for the work. All authors contributed to the critical revision of the manuscript. TK and AJG take responsibility for the manuscript as a whole.

Conflicts of Interest

MJ is a current salaried employee and holds equity in Centaur Labs. CEF is a former salaried employee and currently holds equity in Centaur Labs. She has served as a compensated consultant for Philips and Level Ex. AES has served as a compensated consultant for Astra Zeneca, Bracco Diagnostics, Bristol Myers Squibb, General Electric, Gerson Lehman Group, Guidepoint Global Advisors, Supersonic Imagine, Novartis, Pfizer, Philips, Parexel Informatics, and WorldCare Clinical. He holds stock options in Rhino Healthtech Inc and Ochre Bio Inc. He holds advisory board or committee memberships for General Electric, the Foundation for the National Institutes of Health, and the Sano Center for Computational Personalized Medicine, and is a member of the Board of Governors of the American Institute for Ultrasound in Medicine. His institution has received research grant support and equipment for projects that he has led from the Analogic Corporation, Canon, Echosens, General Electric, Hitachi, Philips, Siemens, Supersonic Imagine/Hologic, Toshiba Medical Systems, the US Department of Defense, Fujifilm Healthcare, the Foundation for the National Institutes of Health, Partners Healthcare, Toshiba Medical Systems, and Siemens Medical Systems. He holds equity in Avira Inc, Autonomus Medical Technologies, Inc, Evidence Based Psychology LLC, Klea LLC, Katharos Laboratories LLC, Quantix Bio LLC, and Sonoluminous LLC. He receives royalties from Elsevier Inc and Katharos Laboratories LLC. ED is the chief executive officer of and holds equity in Centaur Labs. AJG has served as a compensated consultant for Philips and Exo.

Multimedia Appendix 1
Additional Methods.
[DOCX File , 15 KB - jmir_v26i1e51397_app1.docx]

References

1. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020 Aug;26(8):1229-1234 [FREE Full text] [doi: 10.1038/s41591-020-0942-0] [Medline: 32572267]
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410 [FREE Full text] [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
3. Irvin JA, Pareek A, Long J, Rajpurkar P, Eng DKM, Khandwala N, et al. CheXED: comparison of a deep learning model to a clinical decision support system for pneumonia in the emergency department. J Thorac Imaging 2022 May 01;37(3):162-167 [FREE Full text] [doi: 10.1097/RTI.0000000000000622] [Medline: 34561377]
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
5. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng 2019 Mar;3(3):173-182. [doi: 10.1038/s41551-018-0324-9] [Medline: 30948806]
6. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018 Oct 24;2(1):35 [FREE Full text] [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
7. Malone TW. The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Style, and Your Life. Boston, MA: Harvard Business School Press; 2004.
8. Surowiecki J. The Wisdom of Crowds. New York, NY: Knopf Doubleday Publishing Group; 2005.

XSL•FO
RenderX

9. Grote A, Schaadt NS, Forestier G, Wemmert C, Feuerhake F. Crowdsourcing of histological image labeling and object delineation by medical students. IEEE Trans Med Imaging 2019;38(5):1284-1294. [doi: 10.1109/TMI.2018.2883237] [Medline: 30489264]
10. Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. JAMA 2015 Jan 20;313(3):303-304 [FREE Full text] [doi: 10.1001/jama.2014.15770] [Medline: 25603003]
11. Heim E, Roß T, Seitel A, März K, Stieltjes B, Eisenmann M, et al. Large-scale medical image annotation with crowd-powered algorithms. J Med Imaging (Bellingham) 2018 Jul;5(3):034002 [FREE Full text] [doi: 10.1117/1.JMI.5.3.034002] [Medline: 30840724]
12. Meakin JR, Ames RM, Jeynes JCG, Welsman J, Gundry M, Knapp K, et al. The feasibility of using citizens to segment anatomy from medical images: accuracy and motivation. PLoS One 2019;14(10):e0222523 [FREE Full text] [doi: 10.1371/journal.pone.0222523] [Medline: 31600225]
13. Monu J, Triplette M, Wood DE, Wolff EM, Lavallee DC, Flum DR, et al. Evaluating knowledge, attitudes, and beliefs about lung cancer screening using crowdsourcing. Chest 2020 Jul;158(1):386-392 [FREE Full text] [doi: 10.1016/j.chest.2019.12.048] [Medline: 32035910]
14. Parry-Cruwys D, MacDonald J. Using gamification to promote accurate data entry of practicum experience hours in graduate students. Behav Anal Pract 2021 Mar;14(1):1-10 [FREE Full text] [doi: 10.1007/s40617-020-00421-2] [Medline: 33732573]
15. Foncubierta Rodríguez A, Müller H. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. New York, NY, United States: Association for Computing Machinery; 2012 Presented at: Proceedings of the ACM multimedia 2012 workshop on crowdsourcing for multimedia; October 29, 2012; Nara, Japan p. 9-14. [doi: 10.1145/2390803.2390808]
16. Von Ahn L. Duolingo: learn a language for free while helping to translate the web. New York, NY, US: Association for Computing Machinery; 2013 Presented at: Proceedings of the 2013 International Conference on Intelligent User Interfaces; March 19-22, 2013; Santa Monica, California, USA p. 1-2. [doi: 10.1145/2449396.2449398]
17. Kattan MW, O'Rourke C, Yu C, Chagin K. The wisdom of crowds of doctors: their average predictions outperform their individual ones. Med Decis Making 2016 May;36(4):536-540. [doi: 10.1177/0272989X15581615] [Medline: 25878196]
18. Kentley J, Weber J, Liopyris K, Braun RP, Marghoob AA, Quigley EA, et al. Agreement between experts and an untrained crowd for identifying dermoscopic features using a gamified app: reader feasibility study. JMIR Med Inform 2023;11:e38412 [FREE Full text] [doi: 10.2196/38412] [Medline: 36652282]
19. Kurvers RHJM, Herzog SM, Hertwig R, Krause J, Carney PA, Bogart A, et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci U S A 2016 Aug 02;113(31):8777-8782 [FREE Full text] [doi: 10.1073/pnas.1601827113] [Medline: 27432950]
20. Wolf M, Krause J, Carney PA, Bogart A, Kurvers RHJM. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PLoS One 2015;10(8):e0134269 [FREE Full text] [doi: 10.1371/journal.pone.0134269] [Medline: 26267331]
21. Moore CL, Copel JA. Point-of-care ultrasonography. N Engl J Med 2011 Feb 24;364(8):749-757. [doi: 10.1056/NEJMra0909487] [Medline: 21345104]
22. Lichtenstein DA. BLUE-protocol and FALLS-protocol: two applications of lung ultrasound in the critically ill. Chest 2015;147(6):1659-1670. [doi: 10.1378/chest.14-1313] [Medline: 26033127]
23. ACEP Board of Directors. Ultrasound guidelines: emergency, point-of-care and clinical ultrasound guidelines in medicine. Ann Emerg Med 2017;69(5):e27-e54 [FREE Full text] [doi: 10.1016/j.annemergmed.2016.08.457] [Medline: 28442101]
24. Blehar DJ, Barton B, Gaspari RJ. Learning curves in emergency ultrasound education. Acad Emerg Med 2015 May;22(5):574-582 [FREE Full text] [doi: 10.1111/acem.12653] [Medline: 25903780]
25. Demi L, Wolfram F, Klersy C, De Silvestri A, Ferretti VV, Muller M, et al. New international guidelines and consensus on the use of lung ultrasound. J Ultrasound Med 2023 Feb;42(2):309-344 [FREE Full text] [doi: 10.1002/jum.16088] [Medline: 35993596]
26. Pang PS, Russell FM, Ehrman R, Ferre R, Gargani L, Levy PD, et al. Lung ultrasound-guided emergency department management of acute heart failure (BLUSHED-AHF): a randomized controlled pilot trial. JACC Heart Fail 2021 Sep;9(9):638-648 [FREE Full text] [doi: 10.1016/j.jchf.2021.05.008] [Medline: 34246609]
27. Arntfield R, Wu D, Tschirhart J, VanBerlo B, Ford A, Ho J, et al. Automation of lung ultrasound interpretation via deep learning for the classification of normal versus abnormal lung parenchyma: a multicenter study. Diagnostics (Basel) 2021 Nov 04;11(11):2049 [FREE Full text] [doi: 10.3390/diagnostics11112049] [Medline: 34829396]
28. Baloescu C, Toporek G, Kim S, McNamara K, Liu R, Shaw MM, et al. Automated lung ultrasound B-line assessment using a deep learning algorithm. IEEE Trans Ultrason Ferroelectr Freq Control 2020 Nov;67(11):2312-2320. [doi: 10.1109/TUFFC.2020.3002249] [Medline: 32746183]
29. Moore CL, Wang J, Battisti AJ, Chen A, Fincke J, Wang A, et al. Interobserver agreement and correlation of an automated algorithm for B-line identification and quantification with expert sonologist review in a handheld ultrasound device. J Ultrasound Med 2022 Oct;41(10):2487-2495. [doi: 10.1002/jum.15935] [Medline: 34964489]

XSL•FO
RenderX

30. Pare JR, Gjesteby LA, Telfer BA, Tonelli MM, Leo MM, Billatos E, et al. Transfer learning for automated COVID-19 B-line classification in lung ultrasound. Annu Int Conf IEEE Eng Med Biol Soc 2022 Jul;2022:1675-1681. [doi: 10.1109/EMBC48229.2022.9871894] [Medline: 36086232]
31. van Sloun RJG, Demi L. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. IEEE J Biomed Health Inform 2020 Apr;24(4):957-964. [doi: 10.1109/JBHI.2019.2936151] [Medline: 31425126]
32. Wang J, Yang X, Zhou B, Sohn JJ, Zhou J, Jacob JT, et al. Review of machine learning in lung ultrasound in COVID-19 pandemic. J Imaging 2022 Mar 05;8(3):65 [FREE Full text] [doi: 10.3390/jimaging8030065] [Medline: 35324620]
33. Horos Project. 2018. URL: https://horosproject.org/ [accessed 2024-05-08]
34. Smith B. ClipDeidentifier. Core Ultrasound. 2022. URL: https://coreultrasound.com/clipdeidentifier/ [accessed 2024-05-08]
35. Lichtenstein DA, Mezière GA. Relevance of lung ultrasound in the diagnosis of acute respiratory failure: the BLUE protocol. Chest 2008 Jul;134(1):117-125 [FREE Full text] [doi: 10.1378/chest.07-2800] [Medline: 18403664]
36. Lichtenstein DA. Lung ultrasound for the cardiologist-a basic application: the B-profile of the bedside lung ultrasound in emergencies protocol for diagnosing haemodynamic pulmonary oedema. Arch Cardiovasc Dis 2020;113(8-9):489-491 [FREE Full text] [doi: 10.1016/j.acvd.2020.05.005] [Medline: 32739293]
37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
38. Anderson KL, Fields JM, Panebianco NL, Jenq KY, Marin J, Dean AJ. Inter-rater reliability of quantifying pleural B-lines using multiple counting methods. J Ultrasound Med 2013 Jan;32(1):115-120. [doi: 10.7863/jum.2013.32.1.115] [Medline: 23269716]
39. Gravel CA, Monuteaux MC, Levy JA, Miller AF, Vieira RL, Bachur RG. Interrater reliability of pediatric point-of-care lung ultrasound findings. Am J Emerg Med 2020 Jan;38(1):1-6. [doi: 10.1016/j.ajem.2019.01.047] [Medline: 30795947]
40. Herraiz JL, Freijo C, Camacho J, Muñoz M, González R, Alonso-Roca R, et al. Inter-rater variability in the evaluation of lung ultrasound in videos acquired from COVID-19 patients. Appl Sci 2023;13(3):1321 [FREE Full text] [doi: 10.3390/app13031321]
41. Šustić A, Mirošević M, Szuldrzynski K, Marčun R, Haznadar M, Podbegar M, et al. Inter-observer reliability for different point-of-care lung ultrasound findings in mechanically ventilated critically ill COVID-19 patients. J Clin Monit Comput 2022 Feb;36(1):279-281 [FREE Full text] [doi: 10.1007/s10877-021-00726-9] [Medline: 34050854]
42. Nowak S, Ruger S. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. New York, NY, United States: Association for Computing Machinery; 2010 Presented at: Proceedings of the international conference on Multimedia information retrieval; March 29-31, 2010; Philadelphia, Pennsylvania, USA p. 557-566. [doi: 10.1145/1743384.1743478]
43. Lucassen R, Duggan N, Goldsmith A, Kapur T. Replication data for: deep learning for detection and localization of B-lines in lung ultrasound. Harvard Dataverse. URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GLCZRB [accessed 2024-06-21]

Abbreviations

ED: emergency department
ML: machine learning
POCUS: point-of-care ultrasound
Qscore: quality score

Edited by A Mavragani; submitted 04.08.23; peer-reviewed by C Ordun, SA Javed; comments to author 13.09.23; revised version received 04.10.23; accepted 10.04.24; published 04.07.24.

Please cite as:
Duggan NM, Jin M, Duran Mendicuti MA, Hallisey S, Bernier D, Selame LA, Asgari-Targhi A, Fischetti CE, Lucassen R, Samir AE, Duhaime E, Kapur T, Goldsmith AJ
Gamified Crowdsourcing as a Novel Approach to Lung Ultrasound Data Set Labeling: Prospective Analysis
J Med Internet Res 2024;26:e51397
URL: https://www.jmir.org/2024/1/e51397
doi:10.2196/51397
PMID:

©Nicole M Duggan, Mike Jin, Maria Alejandra Duran Mendicuti, Stephen Hallisey, Denie Bernier, Lauren A Selame, Ameneh Asgari-Targhi, Chanel E Fischetti, Ruben Lucassen, Anthony E Samir, Erik Duhaime, Tina Kapur, Andrew J Goldsmith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.07.2024. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Twitter (X) to Mobilize Knowledge for First Contact Physiotherapists: Qualitative Study

Laura Campbell[1], BA (Hons), MPhil; Jonathan Quicke[2,3], BSc (Hons), MSc, PhD; Kay Stevenson[1,4], Grad Dip Phys, MPhil; Zoe Paskins[3,4], MbCHB, MMedED, PhD; Krysia Dziedzic[1], PhD; Laura Swaithes[1], BSc (Hons), MA, PhD

[1]Impact Accelerator Unit, School of Medicine, Keele University, Staffordshire, United Kingdom

[2]STARS Education and Research Alliance, Surgical Treatment and Rehabilitation Services (STARS), The University of Queensland, Queensland, Australia

[3]School of Medicine, Keele University, Staffordshire, United Kingdom

[4]Midlands Partnership University NHS Foundation Trust, Stoke on Trent, United Kingdom

Corresponding Author:
Laura Campbell, BA (Hons), MPhil
Impact Accelerator Unit
School of Medicine
Keele University
David Weatherall Building
Staffordshire, ST5 5BG
United Kingdom
Phone: 44 01782734727
Email: l.campbell@keele.ac.uk

Abstract

Background: Twitter (now X) is a digital social network commonly used by health care professionals. Little is known about whether it helps health care professionals to share, mobilize, and cocreate knowledge or reduce the time between research knowledge being created and used in clinical practice (the evidence-to-practice gap). Musculoskeletal first contact physiotherapists (FCPs) are primary care specialists who diagnose and treat people with musculoskeletal conditions without needing to see their general practitioner (family physician) first. They often work as a sole FCP in practice; hence, they are an ideal health care professional group with whom to explore knowledge mobilization using Twitter.

Objective: We aimed to explore how Twitter is and can be used to mobilize knowledge, including research findings, to inform FCPs' clinical practice.

Methods: Semistructured interviews of FCPs with experience of working in English primary care were conducted. FCPs were purposively sampled based on employment arrangements and Twitter use. Recruitment was accomplished via known FCP networks and Twitter, supplemented by snowball sampling. Interviews were conducted digitally and used a topic guide exploring FCP's perceptions and experiences of accessing knowledge, via Twitter, for clinical practice. Data were analyzed thematically and informed by the knowledge mobilization mindlines model. Public contributors were involved throughout.

Results: In total, 19 FCPs consented to the interview (Twitter users, n=14 and female, n=9). Three themes were identified: (1) How Twitter meets the needs of FCPs, (2) Twitter and a journey of knowledge to support clinical practice, and (3) factors impeding knowledge sharing on Twitter. FCPs described needs relating to isolated working practices, time demands, and role uncertainty. Twitter provided rapid access to succinct knowledge, the opportunity to network, and peer reassurance regarding clinical cases, evidence, and policy. FCPs took a journey of knowledge exchange on Twitter, including scrolling for knowledge, filtering for credibility and adapting knowledge for in-service training and clinical practice. Participants engaged best with images and infographics. FCPs described misinformation, bias, echo chambers, unprofessionalism, hostility, privacy concerns and blurred personal boundaries as factors impeding knowledge sharing on Twitter. Consequently, many did not feel confident enough to actively participate on Twitter.

Conclusions: This study explores how Twitter is and can be used to mobilize knowledge to inform FCP clinical practice. Twitter can meet the knowledge needs of FCPs through rapid access to succinct knowledge, networking opportunities, and professional reassurance. The journey of knowledge exchange from Twitter to clinical practice can be explained by considering the mindlines model, which describes how FCPs exchange knowledge in digital and offline contexts. Findings demonstrate that Twitter can be

XSL•FO
RenderX

a useful adjunct to FCP practice, although several factors impede knowledge sharing on the platform. We recommend social media training and enhanced governance guidance from professional bodies to support the use of Twitter for knowledge mobilization.

(J Med Internet Res 2024;26:e55680) doi:10.2196/55680

KEYWORDS

Twitter; X; social media; first contact physiotherapy; musculoskeletal; knowledge mobilisation; primary care; mindlines; qualitative; physiotherapy; implementation

Introduction

Digital social networks are an evolving way for health care professionals and researchers to quickly find, share, and use knowledge [1-3]. Although Facebook, YouTube, WhatsApp, Instagram, and TikTok are the world's most used social media platforms [4], the open, public arena of Twitter (now known as X) offers health care professionals access to many diverse sources of knowledge. Twitter is a popular, free-to-use forum for communication amongst the public and between health care professionals and patients [5] and can offer health care professionals additional insight and understanding into, for example, patient narratives, which are often hidden behind private support groups on platforms such as Facebook, and rarely mentioned on professional platforms such as LinkedIn. The use of Twitter involves users posting short messages that can rapidly be commented on, liked, or reposted by other users worldwide, providing health care professionals with access to a high volume of succinct knowledge in various formats, for example, images, text, videos, and links, unlike other platforms that may focus solely on images or videos. Yet, little is known about how health care professionals find, adapt, use, and share knowledge from Twitter to inform evidence-based clinical practice.

Twitter has been postulated as a solution to reduce the evidence-to-practice gap—the delay between the production of health care research knowledge and its uptake in clinical practice [6]. Knowledge mobilization (KM) seeks to accelerate and facilitate the dynamic nature of creating, sharing, and adapting research and other forms of knowledge across professional, public, and organizational domains, to the point where it can be most useful for stakeholders [7,8]. KM is a social process that acknowledges real world contextual complexities.

One approach to mobilizing knowledge among health care professionals is through enhancing mindlines [8]. Mindlines are "internalised, collectively reinforced and often tacit guidelines in the head" that underpin rapid clinical decision-making in highly pressurized, complex environments [9]. They are informed by a diverse combination of explicit and tacit individual and collective knowledge, experience and storytelling, clinical training, reading, and understanding of local contexts, among other sources, such as social media. Individual and collective mindlines are continuously constructed, challenged, and reinforced through informal conversation with peers [10] and are actively tried, tested, and contextualized in the real world [9]. The mindlines model can be used to study how people use social processes to interact with each other to find, adapt, use, and share knowledge on both an individual level and collectively, and it can be applied to the world of social media.

A recent UK policy push to increase the provision and remit of multispecialty teams and reduce pressures in primary care has seen the introduction of specialist musculoskeletal (MSK) first contact physiotherapists (FCPs) [11]. FCPs work in general practice to diagnose and treat MSK symptoms without patients needing to see their general practitioner (or family physician) [12]. In contrast to traditional physiotherapy departments, most FCPs typically work as the sole FCP in a general practice away from peers [13,14], and many split their time working across multiple practices, services, or additional roles. Within the primary care context of the rapid evolution and implementation of FCP, variation in practice and training [15], and feelings of uncertainty about the FCP role [14], little is known about where FCPs obtain knowledge for clinical practice outside of their teams.

The aim of this study was therefore to explore how Twitter is, and can be, used to mobilize knowledge to inform FCP clinical practice. The mindlines model is used as a lens through which to explore and understand the social processes associated with how FCPs use different types of knowledge from Twitter to inform their mindlines. Findings from this study could support FCPs in using Twitter to find, adapt, and share knowledge for clinical practice and have implications for the way that social media use is governed by professional bodies, such as the National Health Service (NHS) Integrated Care Systems (ICSs) and the Care Quality Commission.

Methods

Overview

A qualitative interview study was conducted. Purposive sampling was used to recruit English NHS MSK FCPs based in general practice with a range of employers (NHS hospital and community trusts, NHS service providers, Primary Care Networks) who used Twitter for professional purposes. To gain a breadth of perspectives, non–Twitter users (NTs) were also recruited. Participants were recruited via professional networks (including Versus Arthritis charity clinical networks, Keele University Allied Health Professionals Critically Appraised Topic Group [16], the Chartered Society of Physiotherapy FCP mailing list, Keele University Impact Accelerator Unit [17], health care professionals training networks, national and local FCP networks, and via Twitter itself). This was supplemented with snowball sampling to identify potential FCP participants via extended professional networks unknown to the research team. Participants were included in the study if they were

XSL•FO
RenderX

currently employed as an MSK FCP in NHS primary care in England. Recruitment continued until inductive thematic saturation was achieved [18].

The COVID-19 pandemic necessitated the use of Teams (Microsoft Corp) to conduct semistructured interviews virtually. Two pilot interviews were carried out with FCPs in the study Stakeholder Advisory Group (SAG). With informed consent, interviews were conducted by LC (Knowledge Broker, female, with training in KM research and practice and qualitative research). Topic guides were iteratively modified as interviews took place and new findings emerged. Final topic guides are included in Multimedia Appendix 1. A digital voice recorder was used to capture the interviews, which were transcribed verbatim and pseudonymized by LC. Participants were assigned codes according to the order in which they were interviewed (P01, P02, etc.) and whether they were a Twitter user (T) or NT. Video content was not recorded. The reporting of this study is in line with the COREQ (consolidated criteria for reporting qualitative research) checklist [19] (Multimedia Appendix 2).

Data Analysis

Data analysis was iterative and largely inductive, informed by the principles of reflexive thematic analysis [20,21]. Memos, decision logs, debriefing, regular meetings with the research team, a reflexive diary, and a researcher positionality statement maintained critical reflexivity throughout the study, allowing the findings to reflect the research question, aims, and objectives and not the bias of the researcher (Multimedia Appendix 3). Furthermore, analysis decisions were regularly discussed with the multidisciplinary SAG and KM experts.

Coding was conducted by LC, with double coding of a subset of transcripts by LS and JQ. Full details of the analysis steps taken are described in Multimedia Appendix 4.

Stakeholder Involvement

A multidisciplinary SAG consisting of patients and the public, academics, FCPs, physiotherapists, a marketing professional, and KM practitioners was convened at the start of the study to ensure relevance to the research topic, to develop, test and refine the interview topic guides and to inform interpretation of the data. SAG meetings were held via a videoconferencing platform (Teams).

Patient and public involvement and engagement was embedded throughout the study, and reporting is aligned to GRIPP (Guidance for Reporting Involvement of Patients and the Public) 2–Short Form [22]. Three public contributors with varying experience of MSK conditions, Twitter, and KM [23,24] were members of the SAG. They coproduced the topic guides, and informed interpretation of data, and codeveloped the plain language summary. Public contributors were reimbursed for their time in line with the National Institute for Health and Care Research's (NIHR's) Public Involvement Standards [25]. Further details are included in Multimedia Appendix 5.

Theoretical Underpinning

The mindlines model [9,10,26] was chosen as an additional lens through which to interpret the data from a KM perspective. This provided rich, contextualized insights into the social processes behind FCPs' use of knowledge from Twitter in clinical practice.

Ethical Considerations

Ethical approval was obtained from Keele University's Faculty of Medicine and Health Sciences Research Ethics Committee 28.10.21 (REC reference MH-210199) with no conditions. See Multimedia Appendix 6. Participants were fully informed through the provision of a participant information sheet before consent was taken and given the option to opt out at any time up to 2 weeks following the interview date. Interview transcripts were pseudonymized so that participants could not be identified. Participants took part on a voluntary basis.

Results

A total of 25 FCPs expressed an interest in the study. One potential participant did not meet the eligibility criteria (not based in England) and 5 did not respond after receiving the study information. In total, 19 MSK FCPs (Ts, n=14; NTs, n=5), from 6 different geographical regions of England consented to be interviewed via a videoconferencing platform (Teams) between November 2021 and February 2022. Interviews lasted between 22 and 92 minutes. Brief participant characteristics are included in Table 1.

XSL•FO
RenderX

Table 1. Participant characteristics.

Characteristic	Participants, n
Recruitment	
Keele University networks	3
Twitter	5
Snowball sampling—extended FCP[a] networks	10
Tagged[b] by colleague on Twitter	1
Gender	
Male	10
Female	9
Employment	
NHS[c]—Foundation Trust	11
NHS—Integrated Care Trust	2
NHS—Community musculoskeletal service	2
NHS—Clinical Commissioning Group	1
NHS and private practice	1
Social enterprise organization providing NHS community services	2
Additional roles	
NHS leadership	4
Split posts	15
Full-time FCP	4 (2 split across practices)
Experience	
Qualified—physiotherapy	Between 5 and 24 years
FCP role	Between 3 months and 3 years
Twitter user	
Yes	14
No	5

[a]FCP: first contact physiotherapist.

[b]Tagged on Twitter—when a Twitter user identifies someone else to bring information in a post to their attention.

[c]NHS: National Health Service.

Thematic maps were used to support data analysis by visualizing the relationships between codes, themes, and different levels of themes (Figures 1 and 2). Figure 1 outlines an initial set of themes and subthemes and the relationships between them. These themes and subthemes were then further refined and interpreted through iterative discussion between the study team and the SAG, as initial themes overlapped and were not clear. Figure 2 shows the final themes and subthemes. The boundaries between these were clarified, and the core concepts for each theme and subtheme were defined by the study team in the context of the overall narrative of the data. Subthemes were used to provide more interpretive depth. Figure 3 provides an overview of the coding structure.

XSL•FO
RenderX

Figure 1. Initial thematic map with themes, subthemes, and relationships. FCP: first contact physiotherapist.

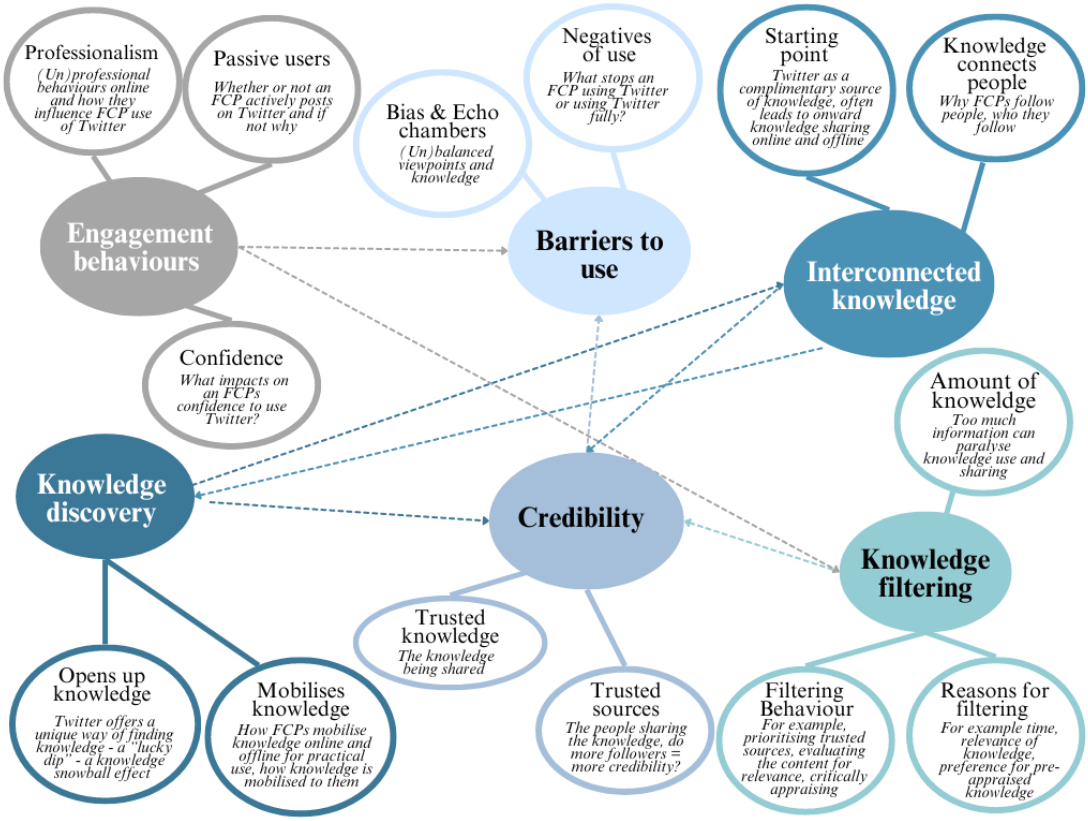

Figure 2. Final thematic map with themes, subthemes, and relationships. FCP: first contact physiotherapist.

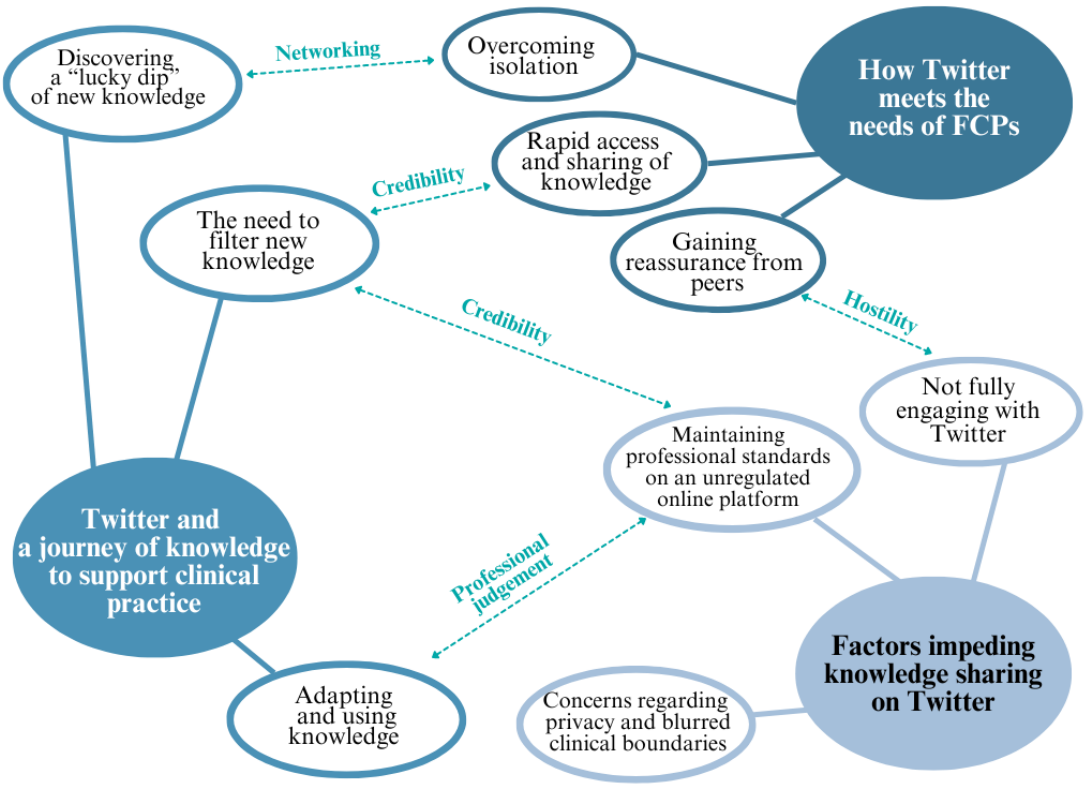

XSL•FO
RenderX

Figure 3. Overview of data structure. FCP: first contact physiotherapist.

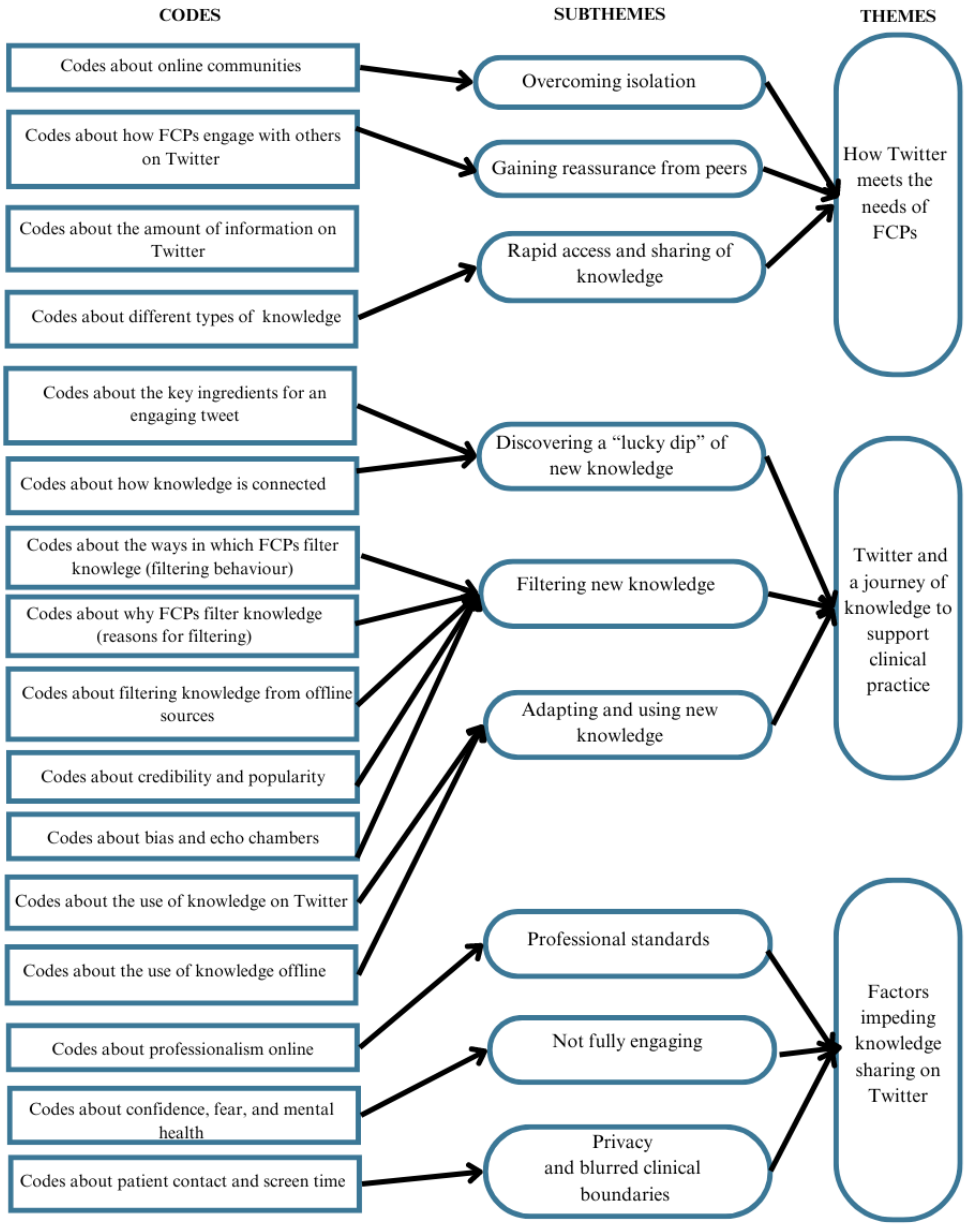

Three key themes were identified: (1) How Twitter meets the needs of FCPs; (2) Twitter and a journey of knowledge to support clinical practice; and (3) factors impeding knowledge sharing on Twitter.

A description of each theme and subtheme and supporting quotes are presented below.

Theme 1: How Twitter Meets the Needs of FCPs

Overcoming Isolation

Several participants described feeling isolated from other FCPs working as the sole therapists in primary care practice. Additional to this, fragmented working patterns between roles and sites, a lack of supervision and mentorship, remote working, and constantly changing policies and guidance contributed to feelings of professional and personal loneliness:

> *It's hard, you don't get the same interaction [...] You can feel a bit isolated.* [P07 T]

Some participants used Twitter more often after moving from being part of a unified secondary care team to an isolated primary care practice:

> *I didn't use Twitter as much in rheumatology as I do now and I think it's because the more remote you are, I mean, I work on my own in a room.* [P06 T]

Twitter was indicated as a potential solution to overcome professional isolation, offering social networking with others to "promote conversation" (P09 T). Additionally, it was thought to provide an opportunity to learn, share best practices and access this study for evidence-based care often hidden behind journal paywalls:

> *'What's the latest evidence on this' or 'What's been released recently?' or 'What discussions have been*

XSL•FO
RenderX

> *going on about this?' I find it [Twitter] really useful, really helpful for that.* [P05 T]

Rapid Access and Sharing of Knowledge

Rapidly changing knowledge could be accessed quickly and easily via Twitter within the time-pressured context of FCP roles to "keep up to date with current thinking" (P08 T) around research and clinical guidelines for best patient care, with short character limits of tweets conveying key, succinct knowledge:

> *I think with Twitter […] they give the kind of pertinent points of a research study or something that's easier to remember, more easily digestible and the kind of key take home messages.* [P13 T]

In contrast, research databases were described as "old fashioned" (P01 T).

Twitter offers many different types of knowledge and sources, allowing participants to be able to "keep on board with lots of things all in one place" (P11 T) and to share knowledge with patients and peers. Yet, this was also described as overwhelming and time-consuming to find, engage with, and share relevant knowledge.

Gaining Reassurance From Peers

Participants used Twitter to get "the feel of what other people are thinking and saying about different things" (P02 T), which gave them reassurance in their professional practice:

> *So there can be that reassurance that you look at a case study [on Twitter] and based on what everyone else is saying you think, oh yeah, well I'd have done that.* [P11 T]

Conversely, one participant highlighted that Twitter promoted feelings of professional insecurity when faced with not keeping up with all new knowledge shared within digital FCP social networks; however, the majority felt it was valuable for reassurance around wider contextual policy concerns, continuing professional development (CPD), and the development of the novel FCP role:

> *Is it only me that's stressed about it or is everybody thinking the same thing? […] it's that kind of support, even though you don't get factual information, it's people talking about things and they're going through the same things like we're going through.* [P14 T]

Seeing tweets from opinion leaders who were "on the same wavelength" (P06 T) and the visibility of senior staff in leadership roles using Twitter to offer guidance and inspiration was thought to be important, yet participants did not see themselves as having sufficient "status" to offer reassurance themselves or to share their own knowledge and experiences:

> *I think I would probably need to tag some big names so that more people saw it and get other people to retweet it and things like that, because I think me on my own probably wouldn't reach very far.* [P02 T]

Twitter's "always on" culture accentuated a perceived expectation that participants must use the social network "to be seen to be in the loop" (P08 T). Guilt around not having time to promote knowledge themselves and anxiety regarding missing out on knowledge were also acknowledged by NTs:

> *They may be getting loads of information and knowledge through Twitter that actually I could be getting and missing out on, and if that were the case, then I would want to know.* [P15 NT]

Twitter and a Journey of Knowledge to Support Clinical Practice

Discovering a "Lucky Dip" of New Knowledge

Participants did not purposefully look for specific knowledge on Twitter; instead, they browsed through their feeds to discover new knowledge from a range of multidisciplinary sources. Knowledge of value included new digital social networks, research evidence, clinical scenarios, imaging, guidelines, training opportunities, expert opinions, and service developments:

> *An example would've been that I don't think I'd have known about the chronic fatigue NICE guidelines coming out if I hadn't have seen somebody share that on Twitter.* [P16 T]

Conversely, participants also described anxiety around a lack of control over what knowledge they would see and expressed concerns about losing time when scrolling through Twitter:

> *I think you could get sucked down the rabbit hole of just endlessly scrolling and how much use would that actually be?* [P02 T]

Fellow FCPs, physiotherapists, governing bodies, opinion leaders, and researchers were followed because they were also followed by people in their digital social networks, conceptualized by one participant as "a bit of a chain really" (P09 T). Tweets containing pre-packaged, trustworthy content, graphics, videos, and simple concise messages that were of use for clinical practice were agreed to be most engaging for participants:

> *The ones that I really like are ones that have a little picture or graphic attached to it […] things that you can pass onto patients as well so, patient friendly information, there's a few things that you can just print off and put in the clinic room and bits like that and that's really helpful.* [P13 T]

The Need to Filter New Knowledge

There was an awareness of potential clinical misinformation and a subsequent need to filter both content (tweets) and sources (people posting) for credibility; yet, this was not done routinely or systematically. Although web-based debate was seen as informative, comments on tweets were mostly ignored unless backed up by research evidence; however, one participant talked about these acting as a mini "critical appraisal (P01 T)" system. There was concern that physiotherapists with more followers would be automatically seen as more credible, with the heightened power of physiotherapy influencers seen as potentially unethical.

For some, echo chambers (when the same ideas and opinions are repeated, reinforcing beliefs and encouraging bias) were

XSL•FO
RenderX

believed to be a risk due to potentially biased knowledge being shared within the small digital physiotherapy community, Twitter's algorithms, and FCPs working in isolation:

> *There is a big risk as clinicians as we develop to one school of thought because we follow the people we agree with only. And then we end up causing, not harm but possibly missing out on a lot of good information.* [P12 T]

In contrast, others felt the amount of knowledge available on Twitter plus debate and discussion in this "open forum" (P07 T) for digital communication reduced the risk of echo chambers and bias.

Sources that were considered credible were professional, respected national bodies such as the Chartered Society of Physiotherapy, academia, or the NHS:

> *The national bodies are pretty robust because a lot of that has already been filtered, so I'm aware that's already been reviewed before it's been put out so that's pretty trustworthy.* [P01 T]

Adapting and Using Knowledge

Participants did not use explicit knowledge (such as clinical guidelines or research findings) found on Twitter directly; instead, they combined this with their own experience and more tacit knowledge, discussing the importance of "the other bits beyond MSK that you need in this role" (P09 T), namely the tacit understandings of local processes and professional culture. Explicit knowledge from Twitter was included in CPD, training sessions, clinical case discussions, and shared through emails and WhatsApp groups. However, all participants who used Twitter adapted explicit knowledge to local contexts or summarized it in other formats:

> *When the MSK standards came out the other week, the first place I saw them was on Twitter. So I read them, summarised them, put a little PowerPoint presentation together for the whole service and said, look, I don't anticipate everyone's going to spend time reading 72 pages of this document but these are the key points.* [P18 T]

NTs interacted with knowledge on other social media platforms, such as LinkedIn, in similar ways:

> *Someone shared something on LinkedIn […] that summarised everything we talked about for three hours. So I immediately saved that, printed it off and I use it on a daily basis.* [P03 NT]

Many participants discussed seeing clinical case studies posted and discussed on Twitter, describing that these widened "clinical reasoning in terms of differential diagnosis" (P16 T) and enhanced their conceptual understanding of multiple clinical conditions.

Factors Impeding Knowledge Sharing on Twitter

Maintaining Professional Standards on an Unregulated Web-Based Platform

All participants who used Twitter described witnessing what they perceived to be unprofessional behavior within the digital physiotherapy community. They spoke about observing "heated arguments and swearing" (P02 T), a "toxic environment" (P08 T, P02 T, and P18 T), and "inflammatory comments" (P16 T), which made them feel sad, embarrassed, and concerned about the detrimental "issue with professionalism on Twitter" (P12 T). Even NTs were familiar with aggression on the platform:

> *((Names physiotherapist)) is bashing them saying, research doesn't suggest this works or that. But he doesn't really share any of his case studies, he's just bashing this guy because he does something which is a bit different.* [P17 NT]

Participants described having a professional responsibility to maintain standards on Twitter, particularly when posting clinical content to a public audience, and to not give clinical advice to members of the public. Although acknowledging that "free speech is important" (P02 T), participants suggested ways in which professional behavior needed to be "regulated" and "policed" (P01 T), and several suggested that the profession would benefit from social media training at the undergraduate, postgraduate, and CPD levels, with particular support for students and newly qualified practitioners.

Not Fully Engaging With Twitter

Despite being experienced and knowledgeable health care professionals, witnessing hostilities within the web-based physiotherapy community resulted in participants who used Twitter feeling anxious to share knowledge on the social network themselves:

> *You're doing something, or managing a service a certain way, or behaving in a certain way with your patients, and somebody disagrees, it can be quite a volatile place.* [P01 T]

This led to some participants preferring to use the more private direct message function over public tweeting, and the majority defining themselves as lurkers:

> *I don't post anything mainly because I'm not going to say the wrong thing and get loads of abuse, I'll just quietly lurk and look at what other people say.* [P02 T]

Reasons for lurking were multifaceted and included fear of making mistakes, a lack of time to accurately and actively share knowledge and the use of personal rather than organizational accounts. Several participants had positive attitudes toward the benefits of active engagement on Twitter; yet, they avoided giving knowledge in favor of taking it instead. Despite these passive, lurking behaviors, participants actively used knowledge from Twitter offline to inform their clinical practice, contributing to contextual, tacit understanding and "facilitating good conversations with teams" (P13 T):

XSL•FO
RenderX

> *I think it's good to know what the conversations being had are and perhaps what the kind of arguments both sides are [on Twitter] but I think ultimately it's more helpful to have the discussion with the colleagues and people that you're working with and also, the population that you're working with to see what's going to work best for your team, what's going to work best for the population that you serve.* [P13 T]

Concerns Regarding Privacy and Blurred Clinical Boundaries

Participants described how Twitter blurs the boundaries between professional and personal lives and expressed concern about 24/7 patient contact on digital platforms potentially breaching these boundaries:

> *That's my time. I don't want you impinging, I will see you in my clinic when it's your appointment, but I don't want you having access, to be in my thoughts and what I'm doing when I'm not at work. Because that's not the deal.* [P02 T]

Further concerns around negative comments and complaints from patients on Twitter were discussed, as well as a fear for privacy and personal security. One participant discussed how a patient had found and used a family image from her social media for his screensaver and how this had directly affected the way she uses Twitter:

> *I'm not really active at putting stuff on [Twitter], from a security point of view. But I do say this to staff, cos they put all their family information, their kids, their full name, you can see their house. It is really easy for people to be found then.* [P07 T]

Only one participant described a positive experience of engaging with patients on Twitter, although maintaining professional boundaries in this situation was also acknowledged:

> *I had a lady who wanted to run a half marathon and I discharged her six months, she'd got a training programme, she just contacted me on Twitter to say she'd done it and that's great, [...] equally we just have to be aware of those boundaries a little bit as well.* [P18 T]

Discussion

Principal Results and Comparison With Prior Work

To our knowledge, this semistructured interview study is the first to explore how FCPs use Twitter as a source of knowledge to inform their clinical practice and contributes to the growing literature on the use of social media for digital KM. This study shows that Twitter provides FCPs with rapid access to succinct knowledge, networking opportunities, and peer reassurance regarding clinical cases, evidence, and policy. It demonstrates the need for FCPs to understand how to find appropriate knowledge on Twitter, filter it for credibility, and adapt it for in-service training and clinical practice. This study highlights many factors impeding FCPs from sharing knowledge on Twitter and their consequent lack of confidence to actively participate on Twitter.

Participants reported that the functionality of Twitter supports their time pressured and fast paced roles by offering rapid access to brief summaries of diverse knowledge from diverse sources. This is particularly useful since FCPs work in a context where their isolated role is weighted toward rapid clinical assessments and access to current knowledge, training, and CPD can be difficult [15]. This reflects commentary noting Twitter as a valid way of keeping up to date with research that is inaccessible behind paywalls and without formally searching research databases [2,27]. FCPs reported being most likely to engage with tweets containing prepackaged knowledge and salient points of research evidence. This aligns with the emergence of knowledge translation tools such as actionable nuggets [28] and clinical knowledge summaries [29] and recent commentary discussing how Tweetorials (a collection of tweets that aim at educating users who engage with them) and tweet threads (a series of connected posts from one person) are useful to keep up to date with research findings [30]. Visual posts were reported by participants as being most engaging, aligning with a 2019 systematic review that found that health care professionals believe infographics reduce the time burden of reading full texts [31].

Social connection with peers, researchers, and opinion leaders through following, retweeting, and liking posts was seen as important in the FCP context, as most work in isolation away from FCP peers and many split their time working across multiple practices [13-15]. Participants described feeling reassured when reading tweets relating to clinical questions, evidence, and constantly changing policy and guidance for the FCP role; findings are useful in light of recent interviews with FCPs exploring common feelings of uncertainty regarding the FCP context and role [13,14]. Many expressed a fear of missing out on something on Twitter, perhaps echoing this uncertainty. Accessing wider perspectives and opinions helped to inform their own clinical decisions, highlighting the role that Twitter plays in cross-disciplinary knowledge sharing [32]. Furthermore, these social connections facilitated CPD and learning for FCPs, findings consistent with a systematic review highlighting Twitter as a vehicle for education and training amongst frontline clinical peers for professional development [1].

A key barrier to FCPs sharing knowledge on Twitter was the perceived hostile environment, which impacted their confidence to exchange knowledge and opinions on the platform (irrespective of their clinical experience). FCPs did not feel safe in this open public forum, believing they would encounter intimidation and toxicity, consistent with literature highlighting unprofessional behavior on Twitter as a concern for health care professionals using it [1,2,33,34]. This barrier to use was consistent with why some participants did not use Twitter. Instead, FCPs were happy to take knowledge from Twitter and adapt and use it in different contexts and communities offline, feeling more comfortable sharing knowledge in more familiar team meetings or with colleagues in person. These findings complement the analysis of 8711 web-based communities, which found that about 90% of digital community members are "lurkers" [35], and case studies, which found lurkers to be more active in sharing information offline [36]. FCP NTs welcomed the knowledge brought to team discussions by FCP Ts. FCP

XSL•FO
RenderX

roles often span role and organizational boundaries, offering a unique opportunity to share knowledge, skills, and ideas from Twitter across other networks [37], such as members of the wider primary care practice, primary care networks, ICSs, and FCP training networks. Although there is guidance for the broader physiotherapy profession to support the use of social media [38], there is no known existing guidance that is specific to FCPs, whose contexts, demands, and working environments differ from those of other physiotherapists.

Although participants reported the rapid access to diverse knowledge on Twitter as useful for clinical practice, at the same time they described the volume of information as occasionally overwhelming and requiring filtering for relevance. This echoes commentary in the field of medical education that accessing information on Twitter can resemble "drinking from the firehose" [2], which can conversely take up more valuable time. Additionally, participants were concerned about the risk of bias, echo chambers, privacy, blurred personal boundaries, and misinformation, often cited as pitfalls of social media for health care professionals [2,33,34,39,40]. FCPs in this study did not routinely or systematically appraise knowledge on Twitter for credibility and automatically trusted knowledge posted by authoritative national bodies and academia, findings that were similar to a cross-sectional survey of 203 physiotherapists and students in New Zealand that explored their use of electronic information for CPD [40].

How Key Findings Relate to the Mindlines Model

Applying the mindlines model to our analysis provided deeper insight into how knowledge from Twitter was shared, discussed, adapted, and used on both individual and collective levels. Knowledge from Twitter was frequently taken into offline discussions and training, allowing FCPs the opportunity to collectively combine, discuss, challenge, and make sense of knowledge to adapt it to fit local contexts, becoming what has been described in the literature as "Knowledge-in-Practice-in-Context" [9,41]. The reported hostile environment on Twitter reinforces why FCPs preferred not to contribute to web-based discussions and debate but to take knowledge offline to inform CPD and clinical practice. The knowledge gleaned from Twitter therefore did enhance FCP mindlines, aligning with KM literature, which outlines trusted "safe spaces" as a necessary prerequisite for mindlines to develop [8]. Conversely, misinformation on Twitter could pose a risk of inaccurate information becoming internalized into individual FCP mindlines. Subsequently, through ongoing knowledge sharing, "mindlines can spread collective folly" [9] and move into the collective FCP thought. This risk further explains why FCPs instinctively prefer to sense-check knowledge found on Twitter with trusted colleagues in face-to-face, offline contexts.

FCPs described taking multiple forms of knowledge they found on the platform and combining them with existing knowledge to make decisions. In this respect, Twitter resembles the insertion of knowledge into mindlines observed by Gabbay and le May in face-to-face contexts [25]. This study has shown how FCP informal debate and experience sharing on Twitter supported both FCPs explicit clinical knowledge and a more nuanced tacit contextual understanding of FCP policy, norms, and role expectations. This echoes research that determined that a blend of explicit and tacit knowledge is important for mindline development, both on digital social networks and offline [26,42]. This contemporary study has shown how knowledge sources informing FCP mindlines are moving beyond in person conversations and expanding to include digital social networks.

Implications for Research and Practice

As multispecialty teams in primary care continue to evolve [11], so too do the ways in which health care professionals such as FCPs access knowledge for clinical practice [1,34]. This study demonstrated how the use of Twitter in health care can support FCPs to be more informed by offering access to many different types of new knowledge and connections to peers. However, specific guidance for its use should be considered for implementation by local NHS MSK ICSs, professional bodies, and the Care Quality Commission. Further work is needed to cocreate this guidance with stakeholders based on the findings from this study. This should include the development and implementation of further social media training for FCPs to avoid inappropriate professional conduct and empower FCPs to use social media in a responsible and effective way for learning. Training could offer ways to overcome the potential barriers to the use of Twitter for KM highlighted by this study. Specifically, it should include how to access relevant information, how to identify misinformation and evaluate tweets, the importance of maintaining professional standards and behaviors, and ways of maintaining privacy if desired. In addition, guidance should also include increased visibility of senior leadership and governance for physiotherapists regarding digital professionalism, which would also contribute to psychological safety on digital social networks.

For people seeking to communicate and mobilize health knowledge, Twitter could play a role in developing 2-way relationships to share knowledge across professional, public, and organizational boundaries and between research and practice. Researchers and knowledge mobilisers must consider the drivers and challenges of the FCP community when mobilizing knowledge using Twitter and take into account the preferred ways in which knowledge is exchanged. Given that increased numbers of health care professionals, such as FCPs, are using social media to access knowledge, implementation, and KM strategies should include the use of social media. Based on findings from this study, Tweet-ups (when users can join a chat around a specific topic using a hashtag at a certain time) or Tweetorials are potential practical examples of how this could be done within the FCP community. Using visual tweets and infographics on Twitter is recommended to effectively engage FCPs and communicate knowledge. This study has shown how KM is complex and messy [43] and spans both digital and offline spaces. Knowledge mobilisers should therefore supplement the use of Twitter as a knowledge source with face-to-face means such as discussions or communities of practice—an established strategy for mobilizing knowledge [44].

Further research exploring the use of social media amongst other professional groups would develop a fuller picture of the

XSL•FO
RenderX

role that Twitter plays in health care practice, and questions still remain about what makes health care professionals frequent tweeters and how those who are less confident can be supported to use this digital social network effectively. Future work to address the ways in which social media may be used to mobilize knowledge between health care professionals, patients, and the public is also needed.

Limitations

Strengths of this work include the robust, theory-informed approach to reflexive thematic analysis to develop final themes and actionable outcomes [45]. The use of purposive and snowball sampling enabled the recruitment of a broad range of FCPs from varying employment and geographical backgrounds. Partnership working and coproduction in the SAG was a particular strength of the study, enabling the identification of areas for discussion in the topic guides not previously considered, for example, the ways in which patients may engage with FCPs on Twitter.

A potential limitation is that data collection took place before Twitter changed ownership and rebranded as X, changing the context in which this study was based. Recruitment via physiotherapy networks and Twitter risked potential response bias, however, this was mitigated by both purposive and snowball sampling methods to recruit a broad range of perspectives (including NTs). Although this study focused on one health care professional group, potentially impeding the transferability of results to other groups, the findings illustrate key issues likely to be comparable to those of others working in isolation, such as general practice nurses [46], general practitioners, community pharmacists, or practice managers. Furthermore, there is a limit to how well social media posts can comprehensively represent all knowledge and information. A number of authors researching in this area have identified temporal and geographic biases in tweets [47,48], which may influence the knowledge accessed by this study's participants through Twitter. While some findings may be applicable to other countries, this study was limited to the context of English NHS primary care. The study purposefully interviewed 5 NTs, which was enough to achieve theoretical saturation within the aims of this study. Semistructured interviews risk response bias and interviewer bias [49]; however, participants spoke openly about both positive and negative accounts of using Twitter, and researcher reflexivity was used throughout the study. Data regarding self-reported ethnicity was not collected—it is recognized that this would have been useful to describe the diversity of the sample. Finally, the study was conducted during the COVID-19 pandemic, which necessitated the use of Teams rather than face-to-face interviews. This has been suggested to stifle rapport building [50], but was convenient for busy clinicians and has been shown to produce similar richness and quality of data as traditional face-to-face methods [51,52].

Conclusions

To the best of our knowledge, this is the first study to explore how Twitter is and can be used to mobilize knowledge to inform FCP clinical practice through the exploration of FCP's perceptions and experiences of using the platform. It illustrates how Twitter can meet the knowledge needs of FCPs by providing rapid access to succinct knowledge, networking opportunities, and professional reassurance. The novel use of the KM model mindlines as a theoretical lens provided a deeper understanding of the journey of knowledge exchange from Twitter to clinical practice by describing how FCPs access, adapt, and share diverse knowledge with peers in digital and offline contexts. Although several factors impede knowledge sharing, we recommend social media training and enhanced governance guidance from professional bodies to support its potential to have a pivotal role in KM.

Acknowledgments

The authors would like to acknowledge the contributions of the Stakeholder Advisory Group and Public Contributors, the Knowledge Mobilisation Forum in the West Midlands, and John Gabbay and Andrée le May, whose insights contributed to the authors' theoretical understanding of mindlines and the practical application of the model to this work. KD and LC are partially funded by the National Institute for Health and Care Research (NIHR) Applied Health and Care Research Collaboration West Midlands (NIHR 200165), and support for this study was given through an NIHR Senior Investigator award (ID NIHR 200259) to KD.

Authors' Contributions

LC contributed to the study design, data collection, data analysis, and interpretation of findings and wrote the paper with input from all authors. JQ, LS, and KS contributed to the study conception, study design, data analysis, the interpretation of the findings, and the drafting of the paper. ZP contributed to interpretation of findings and drafting of the paper. KD contributed to the funding and the drafting of the paper.

Conflicts of Interest

ZP received funding from UCB Pharma.

Multimedia Appendix 1
Interview topic guides.
[DOCX File , 15 KB - jmir_v26i1e55680_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
COREQ (consolidated criteria for reporting qualitative research) checklist.
[DOCX File , 16 KB - jmir_v26i1e55680_app2.docx]

Multimedia Appendix 3
Positionality statement.
[DOCX File , 14 KB - jmir_v26i1e55680_app3.docx]

Multimedia Appendix 4
Analysis steps.
[DOCX File , 15 KB - jmir_v26i1e55680_app4.docx]

Multimedia Appendix 5
Patient and public involvement.
[DOCX File , 14 KB - jmir_v26i1e55680_app5.docx]

Multimedia Appendix 6
Ethics decision.
[PDF File (Adobe PDF File), 169 KB - jmir_v26i1e55680_app6.pdf]

References

1. Chan WS, Leung AY. Use of social network sites for communication among health professionals: systematic review. J Med Internet Res 2018;20(3):e117 [FREE Full text] [doi: 10.2196/jmir.8382] [Medline: 29592845]
2. Choo EK, Ranney ML, Chan TM, Trueger NS, Walsh AE, Tegtmeyer K, et al. Twitter as a tool for communication and knowledge exchange in academic medicine: a guide for skeptics and novices. Med Teach 2015;37(5):411-416. [doi: 10.3109/0142159X.2014.993371] [Medline: 25523012]
3. Tunnecliff J, Ilic D, Morgan P, Keating J, Gaida JE, Clearihan L, et al. The acceptability among health researchers and clinicians of social media to translate research evidence to clinical practice: mixed-methods survey and interview study. J Med Internet Res 2015;17(5):e119 [FREE Full text] [doi: 10.2196/jmir.4347] [Medline: 25995192]
4. Kemp S. Digital 2022: July global statshot. 2022. URL: https://datareportal.com/reports/digital-2022-july-global-statshot [accessed 2024-04-05]
5. Markham MJ, Gentile D, Graham DL. Social media for networking, professional development, and patient engagement. Am Soc Clin Oncol Educ Book 2017;37:782-787 [FREE Full text] [doi: 10.1200/EDBK_180077] [Medline: 28561727]
6. Currie G, Kiefer T, Spyridonidis D. From what we know to what we do: enhancing absorptive capacity in translational health research. BMJ Leader 2019;4(1):18-20. [doi: 10.1136/leader-2019-000166]
7. Ward V. Why, whose, what and how? A framework for knowledge mobilisers. Evid Policy 2017;13(3):477-497. [doi: 10.1332/174426416X14634763278725]
8. Powell A, Davies H, Nutley S. Facing the Challenges of Research‐Informed Knowledge Mobilization: ‘Practising What We Preach’?. New Jersey: Public Administration; 2017:36-52. [doi: 10.1111/padm.12365]
9. Gabbay J, le MA. Practice-Based Evidence for Healthcare: Clinical Mindlines. 1st Edition. Oxford: Routledge; 2011.
10. Gabbay J, le May A. Mindlines: making sense of evidence in practice. Br J Gen Pract 2016;66(649):402-403 [FREE Full text] [doi: 10.3399/bjgp16X686221] [Medline: 27481961]
11. Five year forward view. NHS England. 2014. URL: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf [accessed 2023-12-05]
12. Chartered Society of Physiotherapy. First contact physiotherapy posts in general practice. A guide for implementation in England. 2018. URL: https://www.csp.org.uk/system/files/001404_fcp_guidance_england_2018.pdf [accessed 2023-12-05]
13. Greenhalgh S, Selfe J, Yeowell G. A qualitative study to explore the experiences of first contact physiotherapy practitioners in the NHS and their experiences of their first contact role. Musculoskelet Sci Pract 2020;50:102267. [doi: 10.1016/j.msksp.2020.102267] [Medline: 33032036]
14. Ingram S, Stenner R, May S. The experiences of uncertainty amongst musculoskeletal physiotherapists in first contact practitioner roles within primary care. Musculoskelet Care 2023;21(3):644-654. [doi: 10.1002/msc.1735] [Medline: 36683250]
15. Halls S, Thomas R, Stott H, Cupples ME, Kersten P, Cramp F, et al. Provision of first contact physiotherapy in primary care across the UK: a survey of the service. Physiotherapy 2020;108:2-9. [doi: 10.1016/j.physio.2020.04.005] [Medline: 32693238]
16. Keele University. Evidence into practice groups. URL: https://www.keele.ac.uk/iau/evidenceintopracticegroups/ [accessed 2023-12-18]
17. Keele University. Impact Accelerator Unit. URL: https://www.keele.ac.uk/iau/ [accessed 2023-12-20]

XSL•FO
RenderX

18. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
19. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
20. Moltu C, Stefansen J, Svisdahl M, Veseth M. Thematic Analysis: A Practical Guide to Understanding and Doing. 1st Edition. United States: Sage Publications Ltd; 2012:1608-1616.
21. Braun V, Clarke V. Doing reflexive TA. 2022. URL: https://www.thematicanalysis.net/doing-reflexive-ta/ [accessed 2023-12-05]
22. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. Br Med J 2017;358:j3453 [FREE Full text] [doi: 10.1136/bmj.j3453] [Medline: 28768629]
23. Campbell L, Evans N, Skrybant M, Link G, Dziedzic K. Patient and public involvement in the implementation of research. Ann Rheum Dis 2023;82(1):187-188. [doi: 10.1136/annrheumdis-2023-eular.2658]
24. Swaithes L, Campbell L, Anthierens S, Skrybant M, Schiphof D, French H, et al. Series: public engagement with research. Part 4: maximising the benefits of involving the public in research implementation. Eur J Gen Pract 2023;29(1):2243037 [FREE Full text] [doi: 10.1080/13814788.2023.2243037] [Medline: 37609798]
25. NIHR. UK standards for public involvement. 2019. URL: https://sites.google.com/nihr.ac.uk/pi-standards/home [accessed 2023-12-18]
26. Gabbay J, le May A. Evidence based guidelines or collectively constructed "mindlines?" Ethnographic study of knowledge management in primary care. Br Med J 2004;329(7473):1013 [FREE Full text] [doi: 10.1136/bmj.329.7473.1013] [Medline: 15514347]
27. Hebron C. Physiotherapy past, present and future. Embracing the digital age: physiotherapy and social media. URL: https://www.ingentaconnect.com/content/ppa/pr/2018/00002018/00000045/art00002# [accessed 2023-12-15]
28. McColl MA, Aiken A, Smith K, McColl A, Green M, Godwin M, et al. Actionable nuggets: knowledge translation tool for the needs of patients with spinal cord injury. Can Fam Physician 2015;61(5):e240-e248 [FREE Full text] [Medline: 26167564]
29. National Institute for Health and Care Excellence. Clinical knowledge summaries. URL: https://cks.nice.org.uk/ [accessed 2023-12-15]
30. Coler-Reilly A, Graef ER, Kim AHJ, Liew JW, Putman MS, Sattui SE, et al. Social media for research discourse, dissemination, and collaboration in rheumatology. Rheumatol Immunol Res 2022;3(4):169-179 [FREE Full text] [doi: 10.2478/rir-2022-0031] [Medline: 36879836]
31. Zadro J, O'Keeffe M, Maher C. Do physical therapists follow evidence-based guidelines when managing musculoskeletal conditions? Systematic review. BMJ Open 2019;9(10):e032329 [FREE Full text] [doi: 10.1136/bmjopen-2019-032329] [Medline: 31591090]
32. Mohammadi E, Thelwall M, Kwasny M, Holmes K. Academic information on Twitter: a user survey. PLoS One 2018;13(5):e0197265 [FREE Full text] [doi: 10.1371/journal.pone.0197265] [Medline: 29771947]
33. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018;7(6):121 [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
34. Little JS, Romee R. Tweeting from the bench: Twitter and the physician-scientist benefits and challenges. Curr Hematol Malig Rep 2020 Dec;15(6):419-423 [FREE Full text] [doi: 10.1007/s11899-020-00601-5] [Medline: 33179209]
35. Muller M. Lurking as personal trait or situational disposition. 2012 Presented at: Lurking and contributing in enterprise and social media. Proceedings of the ACM Conference on Computer Supported Cooperative Work; 2012 Feb 11; USA p. 253-256. [doi: doi.org/10.1145/2145204.2145245]
36. Cranefield J, Yoong P, Huff S. Rethinking lurking: invisible leading and following in a knowledge transfer ecosystem. J Assoc Inf Syst 2015;16(4):213-247. [doi: 10.17705/1jais.00394]
37. Swaithes L, Walsh N, Quicke JG. Are physiotherapists too bound to be boundary spanning? Musculoskelet Care 2021;19(4):550-554. [doi: 10.1002/msc.1544] [Medline: 33651458]
38. Social media guide for CSP members. 2019. URL: https://www.csp.org.uk/system/files/publication_files/csp-social-media-guidance-may-2019-v02.pdf [accessed 2023-12-18]
39. Brugnoli E, Cinelli M, Quattrociocchi W, Scala A. Recursive patterns in online echo chambers. Sci Rep 2019;9(1):20118 [FREE Full text] [doi: 10.1038/s41598-019-56191-7] [Medline: 31882591]
40. Clode NJ, Darlow B, Rouse J, Perry M. What electronic information resources do physiotherapists prefer to use to support their CPD? Physiother Res Int 2021;26(1):e1881. [doi: 10.1002/pri.1881] [Medline: 32964592]
41. Gabbay J, le May A, Pope C, Brangan E, Cameron A, Klein JH, et al. Uncovering the processes of knowledge transformation: the example of local evidence-informed policy-making in United Kingdom healthcare. Health Res Policy Syst 2020;18(1):110 [FREE Full text] [doi: 10.1186/s12961-020-00587-9] [Medline: 32988405]

XSL•FO
RenderX

42. Panahi S, Watson J, Partridge H. Conceptualising social media support for tacit knowledge sharing: physicians' perspectives and experiences. In: JKM. 2016 Apr 04 Presented at: 23rd Australasian Conference on Information Systems; 2012 Dec 05; London p. 344-363. [doi: 10.1108/jkm-06-2015-0229]
43. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24. [doi: 10.1002/chp.47] [Medline: 16557505]
44. Swaithes L, Paskins Z, Quicke JG, Stevenson K, Fell K, Dziedzic K. Optimising the process of knowledge mobilisation in communities of practice: recommendations from a (multi-method) qualitative study. Implement Sci Commun 2023;4(1):11 [FREE Full text] [doi: 10.1186/s43058-022-00384-1] [Medline: 36703232]
45. Campbell K, Orr E, Durepos P, Nguyen L, Li L, Whitmore C, et al. Reflexive thematic analysis for applied qualitative health research. Qual Rep 2021;26(6):2011-2028. [doi: 10.46743/2160-3715/2021.5010]
46. Roberts-Lewis SF, Baxter HA, Mein G, Quirke-McFarlane S, Leggat FJ, Garner HM, et al. The use of social media for dissemination of research evidence to health and social care practitioners: protocol for a systematic review. JMIR Res Protoc 2023;12:e45684 [FREE Full text] [doi: 10.2196/45684] [Medline: 37171840]
47. Padilla JJ, Kavak H, Lynch CJ, Gore RJ, Diallo SY. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on twitter. PLoS One 2018;13(6):e0198857 [FREE Full text] [doi: 10.1371/journal.pone.0198857] [Medline: 29902270]
48. Gore RJ, Diallo S, Padilla J. You are what you tweet: connecting the geographic variation in America's obesity rate to twitter content. PLoS One 2015;10(9):e0133505 [FREE Full text] [doi: 10.1371/journal.pone.0133505] [Medline: 26332588]
49. Bowling A. Research Methods in Health: Investigating Health and Health Services. USA: Open University Press; 2009.
50. Sturges J, Hanrahan K. Comparing telephone and face-to-face qualitative interviewing: a research note. Qual Res 2016;4(1):107-118 [FREE Full text] [doi: 10.1177/1468794104041110]
51. Rubinger L, Gazendam A, Ekhtiari S, Nucci N, Payne A, Johal H, et al. Maximizing virtual meetings and conferences: a review of best practices. Int Orthop 2020;44(8):1461-1466 [FREE Full text] [doi: 10.1007/s00264-020-04615-9] [Medline: 32445031]
52. Vindrola-Padros C, Chisnall G, Cooper S, Dowrick A, Djellouli N, Symmons SM, et al. Carrying out rapid qualitative research during a pandemic: emerging lessons from COVID-19. Qual Health Res 2020;30(14):2192-2204 [FREE Full text] [doi: 10.1177/1049732320951526] [Medline: 32865149]

Abbreviations

COREQ: consolidated criteria for reporting qualitative research
CPD: continuing professional development
FCP: first contact physiotherapist
GRIPP: Guidance for Reporting Involvement of Patients and the Public
ICS: Integrated Care System
KM: knowledge mobilization
MSK: musculoskeletal
NHS: National Health Service
NIHR: National Institute for Health and Care Research
NT: non–Twitter user
SAG: Stakeholder Advisory Group
T: Twitter user

Edited by S Ma; submitted 02.01.24; peer-reviewed by R Gore, X Cheng, S Roberts-Lewis; comments to author 14.03.24; revised version received 26.04.24; accepted 06.05.24; published 08.07.24.

Please cite as:
Campbell L, Quicke J, Stevenson K, Paskins Z, Dziedzic K, Swaithes L
Using Twitter (X) to Mobilize Knowledge for First Contact Physiotherapists: Qualitative Study
J Med Internet Res 2024;26:e55680
URL: https://www.jmir.org/2024/1/e55680
doi:10.2196/55680
PMID:38742615

©Laura Campbell, Jonathan Quicke, Kay Stevenson, Zoe Paskins, Krysia Dziedzic, Laura Swaithes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Twitter Discussions on #digitaldementia: Content and Sentiment Analysis

Hyeongchan Cho[1], MA; Kyu-Min Kim[2], PhD; Jee-Young Kim[3], MD, PhD; Bo-Young Youn[4], MPH, MBA, PhD

[1]Department of Business Administration, Graduate School, Kyung Hee University, Seoul, Republic of Korea

[2]Department of Health Administration, Gyeonggi University of Science and Technology, Gyeonggi-do, Republic of Korea

[3]Medical R&D Center, Bodyfriend Co Ltd, Seoul, Republic of Korea

[4]Department of Bio-Healthcare, Hwasung Medi-Science University, Gyeonggi-do, Republic of Korea

Corresponding Author:
Bo-Young Youn, MPH, MBA, PhD
Department of Bio-Healthcare
Hwasung Medi-Science University
A-311, 400-5, Namyangjungang-ro
Namyang-eup, Hwaseong-si
Gyeonggi-do, 18274
Republic of Korea
Phone: 82 31 369 9116
Email: jamesyoun@hsmu.ac.kr

Abstract

Background: Digital dementia is a term that describes a possible decline in cognitive abilities, especially memory, attributed to the excessive use of digital technology such as smartphones, computers, and tablets. This concept has gained popularity in public discourse and media lately. With the increasing use of social media platforms such as Twitter (subsequently rebranded as X), discussions about digital dementia have become more widespread, which offer a rich source of information to understand public perceptions, concerns, and sentiments regarding this phenomenon.

Objective: The aim of this research was to delve into a comprehensive content and sentiment analysis of Twitter discussions regarding digital dementia using the hashtag #digitaldementia.

Methods: Retrospectively, publicly available English-language tweets with hashtag combinations related to the topic of digital dementia were extracted from Twitter. The tweets were collected over a period of 15 years, from January 1, 2008, to December 31, 2022. Content analysis was used to identify major themes within the tweets, and sentiment analysis was conducted to understand the positive and negative emotions associated with these themes in order to gain a better understanding of the issues surrounding digital dementia. A one-way ANOVA was performed to gather detailed statistical insights regarding the selected tweets from influencers within each theme.

Results: This study was conducted on 26,290 tweets over 15 years by 5123 Twitter users, mostly female users in the United States. The influencers had followers ranging from 20,000 to 1,195,000 and an average of 214,878 subscribers. The study identified four themes regarding digital dementia after analyzing tweet content: (1) cognitive decline, (2) digital dependency, (3) technology overload, and (4) coping strategies. Categorized according to Glaser and Strauss's classifications, most tweets (14,492/26,290, 55.12%) fell under the categories of wretched (purely negative) or bad (mostly negative). However, only a small proportion of tweets (3122/26,290, 11.86%) were classified as great (purely positive) or swell sentiment (mostly positive). The ANOVA results showed significant differences in mean sentiment scores among the themes ($F_{3,3581}$=29.03; P<.001). The mean sentiment score was –0.1072 (SD 0.4276).

Conclusions: Various negative tweets have raised concerns about the link between excessive use of digital devices and cognitive decline, often known as digital dementia. Of particular concern is the rapid increase in digital device use. However, some positive tweets have suggested coping strategies. Engaging in digital detox activities, such as increasing physical exercise and participating in yoga and meditation, could potentially help prevent cognitive decline.

(J Med Internet Res 2024;26:e59546) doi:10.2196/59546

XSL•FO
RenderX

KEYWORDS

digital dementia; dementia; public health; Twitter; social media; mobile phone

Introduction

Background

Excessive dependence on electronic devices such as the internet or smartphones can cause digital dementia, that is, cognitive impairment such as decreased attention or memory and, more seriously, promote the early onset of dementia [1]. According to a meta-analysis of 74 studies encompassing 2.8 million adults aged 30 to 64 years, 119 per 100,000 people develop young-onset dementia, amounting to 3.9 million cases worldwide; as a result, Alzheimer disease was found to be the most prevalent, followed by vascular and frontotemporal dementias [2]. Furthermore, the number of people with dementia is also estimated to nearly triple to >152 million by 2050 [3].

The term *digital dementia* was first introduced by the German neuroscientist Manfred Spitzer in 2012, suggesting that excessive use of digital technology can lead to cognitive decline and deterioration of short-term memory [4]. The widespread adoption of various digital screen-based media, including watching television; using the internet; texting; and using social media platforms such as Instagram, Facebook, and Twitter (subsequently rebranded as X), can impair attention and memory recall, which is a growing concern globally [5,6].

Excessive use of the internet can have long-term negative effects [7]. It can impact the brain's ability to process visual information and decrease concentration [8]. Furthermore, it has been linked to various mental health disorders and structural changes in the brain [9]. In addition, it can significantly impair cognitive function, attention control, decision-making processes, and working memory [10].

The COVID-19 pandemic has led to a rapid increase in the frequency of digital technology use. Unfortunately, this increased dependence on digital devices has resulted in physical inactivity, emotional instability, sleep disorders, and memory impairment [11-14]. Studies have shown that the mere presence of a smartphone can have negative effects on cognitive function, resulting in a reduced capacity of attention and impaired task performance [15-17]. Even when the device is not being consciously used, it can still impact cognitive ability [18].

The negative impact of the increased use of digital technology on mental health has been a topic of interest for researchers [19,20]. It can increase symptoms of attention-deficit/hyperactivity disorder, interfere with emotional and social intelligence, lead to addictive behavior, and contribute to social isolation; additionally, it can disrupt cognitive and brain development, affect sleep patterns, and have other adverse effects [8]. Furthermore, factors such as dependence on technology, information overload, and excessive screen time can harm brain health and cognitive capabilities, increasing the risk of Alzheimer disease [21]. Certain groups are at a high risk of experiencing digital dementia, particularly children and adolescents, whose brains are not yet fully matured and who rely heavily on electronic devices [22]. Overuse of such devices at a young age can lead to symptoms of cognitive impairment, including short-term memory loss, and developmental delay [23]. It is important to be aware of these risks and take measures to prevent overuse of electronic devices in young individuals [24,25]. Extensive research in the field of biopsychosocial studies has provided evidence that using digital devices excessively during crucial periods of brain development, for >2-3 hours a day, can have detrimental effects on individuals. These effects include impairments in learning and memory, attention and emotional disorders, substance use disorder, and negative changes in neurological disorders [26].

Given the rapidly increased frequency of the use of social media, it is vital to consider the potential of social media, also known as digital data, in public health surveillance and prevention. There have been promising developments of hybrid approaches that combine traditional surveillance data with digital data, which often capture the accurate spreading of information in a timelier manner [27]. Social media is known for the rapid identification of an outbreak of infectious diseases [28]. A study reported the prediction and detection of dengue fever in China using Weibo messages [29]; another study analyzed Twitter data and predicted the peak of the incidence of COVID-19 in Canada using various symptom keywords such as *cough*, *runny nose*, and *anosmia* [30]. Using social media was also recognized to detect mental illnesses; according to Guntuku et al [31], depression and other mental illnesses were detected based on screening surveys, public sharing of a diagnosis, and patterns in languages and web-based activities via Twitter, Facebook, and web forums.

Objective

With all that mentioned, this study aimed to gain a better understanding of digital dementia through a comprehensive analysis of Twitter data. The researchers used the hashtag #digitaldementia to conduct content and sentiment analysis, as Twitter is a popular social media platform with over 230 million monthly active users and 500 million daily tweets on average worldwide [32]. Twitter is recognized as a valuable source for monitoring public opinion on various health-related issues, including mental health [33-35].

Methods

Study Design and Data Collection

An exploratory content and sentiment analysis was performed. This study focused on a particular group of hashtags that had been used in the past to discuss the issue of digital dementia, including #digitaldementia, #digitalamnesia, #cyberdementia, and #digitalalzheimer. The study collected all tweets that contained at least 1 of these hashtags from January 1, 2008, to December 31, 2022, using Twitter's application programming interface and Python packages (*Tweepy*) [36]. The raw data collected in this study include the author's username, tweet content, time stamp, and any mentions or replies. The individual users were identified, and the user's profile information was

XSL•FO
RenderX

extracted through the appropriate application programming interface based on the collected data. All Twitter data collected and presented in this study were obtained in compliance with Twitter's terms and conditions, which permit the use of publicly available content for syndication, broadcasting, distribution, retweeting, promotion, or publication, with the exception of personal information such as home addresses or identity documents [37]. It is important to note that no compensation was paid to the individual tweeters whose tweets were used in this study, per the agreement upon using Twitter.

Inclusion and Exclusion Criteria

To create a data set of tweets related to digital dementia, publicly available tweets were searched in English or translated to English using Twitter's automated translation feature. Using various combinations of hashtags with Twitter's search function, we retrieved the most frequently hashtags, namely, #digitaldementia, #digitalamnesia, #cyberdementia, and #digitalalzheimer. These hashtags were coined by individuals who saw the excessive use of technology as a threat to human memory. To ensure the inclusion of tweets that specifically mentioned digital dementia without using the selected hashtags, the data set was expanded by searching for the hashtags as individual words. For instance, a search was performed for cases of "digital dementia" in addition to using the #digitaldementia hashtag.

At first, 33,498 tweets were collected using 4 specific search terms and imported into Microsoft Excel. After cleaning the data, 1964 (5.86%) tweets, written in the alphabet but not in English, were removed from the data set. Then, 884 (2.8%) non-retweeted duplicate posts were filtered out. In addition, tweets originating from social bots (ie, automated Twitter accounts) were detected and excluded to ensure an accurate representation of public discussions on digital dementia and related topics. Botometer was used to distinguish between genuine users and social bots; this tool assesses Twitter account characteristics and assigns a score to indicate the likelihood of an account being a bot, with a threshold set at ≥4 on a scale of 1 to 5 for English accounts [38,39]. All the accounts were screened after collection (not in real time). Out of the initial data set of 14,301 accounts, 1770 (12.38%) were identified as bots and removed. The analytical data set comprised 28,043 posts from 12,531 unique, verified, nonbot accounts after the comprehensive tweet cleansing process.

The analysis excluded all tweets containing emoticons. This is because tweets often contain typos, ad hoc abbreviations, phonetic substitutions, ungrammatical structures, and emoticons, which can pose challenges for text-processing tools and introduce bias problems in sentiment analysis [40]. After removing tweets with emoticons, the final set for analysis comprised 26,290 tweets from 11,134 unique nonbot accounts. The inclusion and exclusion process of total tweets according to the workflow is detailed in Figure 1.

Figure 1. Flowchart of the inclusion and exclusion process.

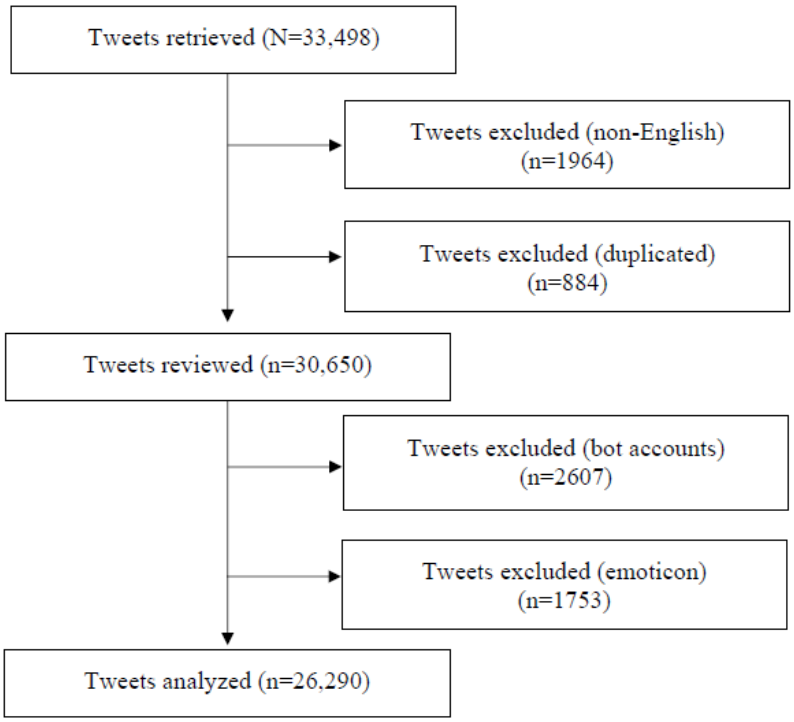

Data Analysis of Tweets

Two different data analysis methods were used to achieve the research objectives: summative content analysis and sentiment analysis. The main goal of content analysis is to measure and explain unfamiliar phenomena [41]. This study used content analysis to identify and describe the textual elements of tweets related to digital dementia. In addition, the study used sentiment analysis, a technique used to analyze written language and determine the emotions conveyed in the text. This method is frequently used in health care and social media research to interpret textual information about the patient experience, with a focus on understanding the individual's perspective [42].

XSL•FO
RenderX

Social media influencers are individuals who are considered opinion leaders and have the ability to influence the attitudes and behaviors of the audience [43]. Recent research has recognized the importance of studying the behavior of influencers on a particular issue to gain a better understanding of the societal context surrounding it [44]. On the basis of previous evidence highlighting the significant role of influencers in health care research [45], a content analysis was conducted on tweets authored by influencers among all the collected tweets. With that said, the aim of this study was to provide a more comprehensive insight into the current state of society's context regarding the issue of digital dementia.

All collected tweets from the influencers underwent qualitative content analysis, a systematic approach for making inferences from text to summarize communication content [46]. This study considered 4109 tweets written by users with a minimum of 20,000 followers. On the basis of previous studies, these users were identified as social media influencers [47-49]. A codebook was developed for this study to use the inductive coding process that is widely used in Twitter content analysis [50]. The analysis followed several steps. Initially, to ensure reliability, raters reviewed a subset of 200 tweets from the influencer tweets (n=4109) to apply preliminary classifications to each category. Any differences in categorization and discrepancies among the evaluators were discussed until consensus was achieved, leading to adjustments in the classification criteria based on the initial ratings. Next, pairs of researchers, working independently, rated another second training set of 500 tweets from influencer tweets using the refined codebook. Cohen κ statistic was used to measure the interoperator reliability of the 500 tweets during the process. The Cohen κ metric spans from –1 to +1. If Cohen κ is ≤0, it signifies no agreement; if Cohen κ is between 0.6 and 0.8, it indicates strong agreement; and if Cohen κ is exactly 1, it represents perfect agreement [51]. The pooled Cohen κ statistic for 500 tweets independently coded by 2 researchers was found to be 0.84, representing excellent reliability.

Two researchers independently reviewed each tweet and manually coded the 3409 remaining tweets from influencers. In cases of classification discrepancies (<12%), the entire research team reviewed the tweet's content and reached a final decision by a consensus of at least two-thirds of the members. Subsequently, 182 tweets with unclassifiable content were excluded. In total, 3927 influencer tweets were evaluated for content analysis.

The Valence Aware Dictionary and Sentiment Reasoner (VADER), a tool included in the *Natural Language Toolkit* library in Python, was used to analyze the overall sentiment of the collected tweets. VADER is a pretrained rule-based sentiment analysis system that evaluates sentiment by examining the words in a text. The toolkit operates based on a lexicon, where independent human evaluators have assigned ratings indicating each word's positive or negative connotations. VADER calculates a composite sentiment score for each tweet by identifying and assessing words in the lexicon. It uses an algorithm that considers punctuation, capitalization, emoticons, and word modifiers to generate a sentiment score ranging from –1 to +1; a score of+1 signifies the highest level of positive sentiment, while –1 indicates the highest level of negative sentiment [52].

Before calculating the sentiment scores, all tweets containing punctuation and emojis were removed during preprocessing. Before the sentiment analysis, the Python software version 3.11.3 (Python Software Foundation) was used for natural language processing (NLP). Emojis were stripped from all collected tweets before conducting NLP tasks to analyze plain texts only. Subsequently, *Natural Language Toolkit*, a Python package for NLP tasks, segmented each tweet into tokenized words, removing stop words with minimal analytic value. Following the NLP tasks, sentiment analysis for the collected tweets was performed.

After tokenizing the words, sentiment analysis was conducted on all 26,290 tweets in this study. The sentiments linked with each theme were evaluated using 6 codes for sentiment analysis (Table 1). Each tweet was given a sentiment score ranging from –1 to +1 and then placed into one of the 6 categories given by Glaser and Strauss [53]. The framework is commonly used in Twitter-based research focusing on sentiment analysis, as suggested by Glaser and Strauss [53,54].

XSL•FO
RenderX

Table 1. Six codes for sentiment analysis.

Code	Definition
No sentiment	Tweets in this category lack emotional language or distinctive punctuation and often focus solely on mentioning the topic without conveying clear positive or negative sentiments. For instance, a tweet simply asking about the symptoms of digital dementia, such as “What are the early signs of digital dementia?” would be classified under this category.
Wretched	Tweets categorized as “wretched” express overwhelmingly negative sentiments related to digital dementia, occasionally containing slightly positive terms. The tweets may convey frustration, dissatisfaction, or negative experiences associated with digital technology. An example tweet might be, “Digital dementia is ruining my life. Can’t remember anything without my phone!” Sentiment scores for this category range from –1 to –0.6.
Bad	“Bad” tweets predominantly feature negative phrases and words, often expressing disappointment or concerns about digital dementia. While there might be occasional positive statements, the negative feelings overshadow the statements. An example could be, “Digital dementia is terrifying. Losing cognitive function to screens is a nightmare.” Sentiment scores for this category range from >–0.6 to –0.2.
So-so	Tweets in the “so-so” category convey a mixed or neutral sentiment regarding digital dementia. Positive and negative statements may balance each other out, or the tweet may remain neutral overall. Even if there are more negative phrases, the positive ones use stronger language than the negative ones. An example tweet might be, “Learning more about digital dementia. It's concerning, but there are ways to mitigate risks.” Sentiment scores for this category range from >–0.2 to <0.2.
Swell	“Swell” tweets predominantly contain positive terms related to digital dementia, with occasional mild negative phrases. However, the positive expressions outweigh the negative ones. For example, “Research shows simple lifestyle changes can combat digital dementia. Encouraging news!” Sentiment scores for this category range from 0.2 to <0.6.
Great	Tweets categorized as “great” are entirely positive in tone and wording, expressing strong affirmative feelings about digital dementia without any complaints. Tweets in this category may contain minimal negative language but are predominantly filled with enthusiastic praise or optimism. An example tweet could be, “Exciting breakthroughs in digital dementia research! There's hope for prevention and treatment.” Sentiment scores for this category range from 0.6 to 1.

After the sentiment analysis for the total collected tweets, to determine if there were significant differences in sentiment scores among the different groups defined during the process of content analysis, a one-way ANOVA was conducted. The one-way ANOVA is an appropriate statistical method in this context for several reasons. First, the methodology allows for the comparison of the means of ≥3 independent groups, making it suitable for analyzing the sentiment scores across multiple themes and subthemes. Second, the observed differences in sentiment scores can be analyzed to determine statistical significance or the possibility of occurring by chance, which is crucial for validating the robustness of our sentiment analysis and ensuring that our findings are not merely artifacts of random variability [55]. Furthermore, ANOVA is particularly advantageous in handling the variation within and between groups, providing a comprehensive understanding of the data’s underlying structure. By quantifying the variances, ANOVA offers a clear picture of how sentiment varies across different categories, which is essential for our research’s objective of understanding the nuanced emotional responses associated with each theme. Given the multiple themes and subthemes in this study, it is imperative to pinpoint which groups exhibit significant differences in sentiment scores, enabling the researchers to draw more precise and actionable insights from the data.

Ethical Considerations

No ethics review was sought because the study only explored the publicly available data on social media and did not conduct any experiments on human participants. However, any personal study data, user IDs, followers, retweets, links, and emoticons were deidentified during analysis.

Results

Overview

This study analyzed 26,290 tweets spanning a 15-year period from 2008 to 2022. The study involved 11,134 Twitter users, with female users representing the highest proportion (n=5600, 50.3%). The primary countries of the users were the United States, followed by the United Kingdom and Northern Ireland (Table 2).

XSL•FO
RenderX

Table 2. Demographics of total Twitter users.

Characteristics	Users (n=11,134), n (%)	Tweets (n=26,290), n (%)
Sex		
Female	5600 (50.3)	14,012 (53.3)
Male	4365 (39.2)	8465 (32.2)
Unknown	1169 (10.5)	3812 (14.5)
Country		
The United States	5088 (45.7)	13,854 (52.7)
The United Kingdom and Northern Ireland	2973 (26.7)	4521 (17.2)
Canada	913 (8.2)	2655 (10.1)
Australia	857 (7.7)	1708 (6.5)
India	256 (2.3)	1498 (5.7)
France	212 (1.9)	289 (1.1)
Germany	178 (1.6)	236 (0.9)
New Zealand	167 (1.5)	197 (0.7)
Singapore	111 (1)	148 (0.6)
Hong Kong	89 (0.8)	155 (0.6)
Philippines	45 (0.4)	51 (0.2)
Unknown	245 (2.2)	978 (3.72)

Content Analysis

During the analysis of tweets by social media influencers, 4 main themes and 11 subthemes were identified using an inductive thematic analysis approach grounded in robust qualitative research principles [56]. The aim was to capture nuanced and meaningful insights from social media discourse. A codebook was developed based on preliminary classifications and refined through interrater reliability assessments to ensure consistent and reliable content categorization. To identify potential domains, the research team conducted an extensive literature review. The codebook was iteratively refined as the coding progressed to include emerging subthemes, ensuring that it captured the full spectrum of issues discussed in the tweets. This iterative refinement process aligns with best practices in qualitative research and ensures that the coding framework remains flexible and responsive to the data's nuances [57].

After analyzing tweet content, four themes emerged regarding digital dementia: (1) cognitive decline, (2) digital dependency, (3) technology overload, and (4) coping strategies. A comprehensive breakdown of primary themes and subthemes can be found in Table 3.

The content analysis identified 1526 (38.9%) of the total 3927 tweets, related to the theme of cognitive decline. These tweets were further categorized into 3 subthemes: the experience of cognitive decline due to the use of digital devices; scientific information about a decrease in brain activity; and problems with poor memory, concentration, and creativity.

XSL•FO
RenderX

Table 3. Content analysis of main themes and subthemes for influencers' tweets (n=3927).

Main themes and subthemes	Tweets, n (%)	Example of phrases or #hashtags
Cognitive decline (n=1526, 38.86%)		• "#Digitaldementia is real. Are we trading our brainpower for convenience?" • "My memory is fading, focus is fleeting, and creativity is dwindling #DigitalDementia #MemoryLoss"
Experience of cognitive decline due to the use of digital devices	859 (56.29)	
Scientific information about a decrease in brain activity	448 (29.36)	
Problems with poor memory, concentration, and creativity	219 (14.35)	
Digital dependency (n=1218, 31.02%)		• "Why is my productivity so tied to technology? #digitalalzheimer" • "Taking more time on my phone. decrease in my attention span and productivity. #DigitalAmnesia"
Issues that arise as a result of the increase in digital device use in reality	549 (45.07)	
The dependency of using digital devices	411 (33.74)	
Anxiety or frustration without digital devices	258 (21.18)	
Technology overload (n=779, 19.84%)		• "Too much digital input, too little mental output. #brainstrain #digitaldementia"
Issues about digital information overload and cognitive strain	496 (63.67)	
Stress caused by excessive information exposure on digital media	283 (36.33)	
Coping strategies (n=404, 10.29%)		• "Stop living in a digital dreamland and start experiencing the real world before your brain turns to mush #DigitalDementia"
Recommendations for alternative activities	234 (57.92)	
Strategies for reducing digital device use	135 (33.42)	
Public support	35 (8.66)	

A large number of tweets, 859 (56.3%), shared personal experiences of cognitive decline linked to excessive use of digital devices. These observations are consistent with research linking prolonged screen time to negative cognitive outcomes [58]. The tweet users frequently expressed concerns about the impact of digital device use on cognitive abilities:

> *My brain feels like it's in a fog after too much screen time. #DigitalDementia #BrainFog*

Another prevalent subtheme was identified through tweets that disseminated scientific information concerning cognitive decline. These tweets often discussed research findings and professional insights about how prolonged use of digital devices can lead to reduced brain activity. For example, recent studies suggest that excessive screen time can impact the brain structures responsible for critical cognitive functions [59]:

> *Studies show that prolonged digital exposure can lead to a decline in cognitive function. #Neuroscience #DigitalDementia*

The remaining 219 (14.4%) tweets discussed issues related to poor memory, concentration, and creativity. Users frequently mentioned struggling with memory retention, managing study time, and maintaining creative thinking and problem-solving skills. The subtheme was identified by grouping tweets that described specific cognitive difficulties and perceived links to digital device use:

> *I can't remember anything these days, and my focus is shot. Too much tech is ruining my brain. #MemoryLoss #DigitalDementia*

Of the total 3927 tweets, the content analysis found 1218 (31%) tweets related to the theme of digital dependency. Digital dependency was categorized into 3 subthemes: increased digital device use, reliance on digital devices, and anxiety or frustration experienced in the absence of digital devices.

The analysis revealed a significant subtheme related to the various issues stemming from the growing use of digital devices. Numerous users expressed concerns about productivity, which was closely tied to technology use, and described feeling helpless or dependent, hindering real-world efficiency. This tendency was evident in tweets where users reflected on the inability to perform tasks without digital assistance, expressing similar sentiments; a typical tweet read as follows:

XSL•FO
RenderX

> *I'm so tied to my phone for everything that it's becoming a problem. How do we function without them? #TechDependency*

One prominent subtheme revolved around the pervasive dependency on digital devices. Tweets within this category often described a compulsive need to use smartphones or computers, even when such use was detrimental to productivity or mental well-being. For example, users frequently noted spending more time on phones, leading to a noticeable decline in attention spans and overall productivity. In addition, the emotional tone of the tweets highlighted a common narrative of struggle and acknowledgment of the negative impacts of digital dependency. A representative tweet stated as follows:

> *I've noticed a drop in my productivity because I can't put my phone down. It's a serious issue. #ScreenTime #DigitalAmnesia*

The remaining 258 (21.2%) tweets addressed the anxiety or frustration experienced in the absence of digital devices. The final subtheme identified was the anxiety or frustration users experienced without digital devices. Tweets in the third subtheme often reflected a deep-rooted fear of missing out or a sense of unease when disconnected from the digital world. For example, some users mentioned needing access to platforms such as Twitter, underscoring the dependency on social media for daily interactions and information:

> *Being without my phone for a few hours made me so anxious. It's scary how dependent I've become. #TechAddiction #Anxiety #digitaldementia*

Of the total 3927 tweets, 779 (19.8%) tweets related to the theme of technology overload were identified. The theme of technology overload was further categorized into 2 subthemes: issues related to digital information overload and cognitive strain and stress caused by excessive exposure to information through digital media.

The most prominent subtheme related to technology overload comprised 496 (63.7%) tweets. The first subtheme, focusing on issues regarding digital information overload and cognitive strain, reflects the overwhelming amount of digital information people encounter daily. Previous research has shown that excessive information can lead to mental overload, reducing individuals' ability to process information effectively [60]. Numerous tweets expressed frustration with the sheer volume of data and difficulty discerning relevant information. For example, users frequently mentioned feeling overwhelmed by constant updates, notifications, and the pressure to stay informed:

> *My brain feels overloaded with all the digital info I consume daily. It's exhausting. #DigitalOverload #BrainStrain*

The second subtheme, which comprised 283 (36.3%) tweets, focused on the stress experienced due to excessive exposure to information on digital media. This subtheme captured the emotional and psychological toll of continuous exposure to digital content. It is well-documented in the literature that prolonged digital engagement can lead to significant stress and anxiety [61]. The tweets under this subtheme often mentioned feelings of anxiety, stress, and burnout associated with digital media use:

> *Scrolling through endless feeds is stressing me out. Too much information all the time. #InfoOverload #Stress*

Furthermore, of the total 3927 tweets, the analysis identified 404 (10.3%) tweets related to coping strategies. The coping strategies section was divided into 3 subthemes: recommendations for alternative activities, strategies for reducing digital device use, and public support.

The largest subtheme within coping strategies, comprising 234 (57.9%) tweets, focused on recommending alternative activities to reduce dependence on digital devices. Influencers often suggested physical exercise, hobbies, creative pursuits, and social interactions as substitutes for digital engagement:

> *Spent the day hiking and felt so much better without my phone. Nature is the best therapy. #DigitalDetox #RealLife*

The second subtheme, with 135 (33.5%) tweets, offered specific strategies for reducing digital device use. The tweets in this category provided practical advice for limiting screen time, such as setting time limits, using tracking apps, or designating device-free zones:

> *Try setting a daily limit on your screen time to help reduce digital dependency. It works! #ScreenTime #TechBalance*

The remaining 35 (8.7%) tweets discussed public support for managing digital dependency and cognitive decline. This category of tweets emphasized the importance of community and social support in coping with challenges. Influencers highlighted the role of peer support, community resources, and collective activities in enhancing resilience:

> *Join a local support group to share tips and get help managing digital addiction. You're not alone. #SupportGroup #DigitalHealth*

Sentiment Analysis

The sentiment analysis of the tweets gathered provided more insights into the emotional tone of the collected data in this research. Table 4 offers comprehensive information about the corresponding sentiments, including the words, phrases, or hashtags used in the collected tweets for this study.

On the basis of the classification given by Glaser and Strauss [53], most tweets (14,492/26,290, 55.12%) fell under the categories of wretched (purely negative) or bad (mostly negative). By contrast, a small proportion of tweets (n=3122, 11.86%) were classified as great (purely positive) or swell sentiment (mostly positive). Approximately 4.2% (n=1107) tweets were classified as so-so, meaning the text had a balanced mix of negativity and positivity. In addition, 9% (n=2366) of the tweets had no sentiments.

The descriptive statistics for a quantitative analysis was also conducted. The sentiment score variable, representing the sentiment score for each tweet, had 23,924 observations with a mean of around –0.197 (SD 0.360), indicating the spread of

XSL•FO
RenderX

sentiment scores around the mean. The sentiment code variable, following the classification by Glaser and Strauss [53], also had 23,924 observations. The mean sentiment code was approximately 2.509 (SD 0.853). The estimated mean sentiment score (–0.197, SE 0.002). The 95% CI for the mean sentiment score was approximately –0.202 to –0.193. These results indicate a high level of certainty (95% CI) that the true population mean sentiment score falls within this range. Since the CI does not include 0, it suggests that the mean sentiment score substantially differs from neutral sentiment. The results suggest an overall negative sentiment, as shown by the mean sentiment score being significantly below 0 and the CI not including 0, which represents neutral sentiment. The detailed statistics provide a solid basis for understanding the sentiment distribution and its implications within the data set.

Sentiment analysis provided valuable insights into the emotions associated with the main themes. Table 5 presents the main themes, sentiments, and phrases or hashtags that reflect the tweeters' experiences.

Sentiment analysis was performed on a total of 3927 tweets from influencers, revealing interesting patterns across 4 major themes. The theme "cognitive decline" accounted for 38.9% (n=1526) of the tweets and included discussions about brain health and cognitive function. Positive sentiments were found in a small fraction, with 2.7% (n=41) mentioning brain training and cognitive enhancement and 11.1% (n=169) discussing brain games, mindfulness, and yoga. A substantial portion (n=498, 32.6%) showed a moderate sentiment with terms such as brain supplements and sleep quality, while negative sentiments dominated, with 34.2% (n=522) mentioning issues such as sleep deprivation, poor diet, and depression. The most severe concerns, such as dementia and Alzheimer disease, were reflected in 10.4% (n=159) of the tweets, and 9% (n=137) of the mentions did not reflect any sentiment, focusing on brain function and neurological changes.

The analysis of tweets revealed that 31.02% (1218/3927) of the tweets reflected digital dependency, indicating a reliance on digital devices. Positive tweets were minimal, with only 1.8% (n=22) discussing detox and mindfulness, while 21.6% (n=263) showed a more favorable sentiment with mentions of support groups and regulation. Tweets with a "so-so" sentiment accounted for 25.9% (n=315) and included discussions on consumption and prevention. Negative sentiments were prominent, with 39.9% (n=486) mentioning addiction, withdrawal, and low productivity. In addition, 4.9% (n=60) of the tweets highlighted severe impacts such as isolation and harassment.

Technology overload accounted for 19.84% (779/3927) of the tweets and focused on the excessive use of technology. Within this category, 10.1% (n=79) of the tweets had a positive sentiment, discussing mindful use and moderation. In addition, 16% (n=125) of the tweets expressed favorable sentiments with terms similar to advancements and innovation. By contrast, 20.5% (n=160) of the tweets had a moderately negative tone, addressing distractions and social media. Furthermore, 29.7% (n=231) of the tweets conveyed negative sentiments such as digital fatigue and stress. Finally, 12.6% (n=98) of the tweets in this category highlighted severe negative impacts, including addiction and dehumanization.

Finally, coping strategies were mentioned in 10.29% (404/3927) of the tweets, discussing methods to manage stress and challenges. Positive tweets (n=28, 6.9%) referred to support and counseling, while 36.1% (n=146) expressed a more favorable sentiment toward meditation, yoga, and creative activities. Tweets with a so-so sentiment (n=65, 16.1%) included relaxation techniques and time management, whereas 21.8% (n=88) reflected harmful coping mechanisms such as substance use disorder and self-harm. The most serious concerns, such as suicide and hopelessness, were mentioned in 7.4% (n=30) of the tweets, and 11.6% (n=47) did not reflect any particular sentiment, discussing masks, sanitizers, and information changes.

Table 4. Sentiment analysis of total tweets (n=26,290; original theme: digital dementia).

Sentiment	Tweets, n (%)	Examples of words or phrase
Great	741 (2.82)	"wins award," "life story," "joy in patients," "better outcomes," and "meditation"
Swell	2381 (9.06)	"magic martin," "dementia diagnosis," "excellent outlook," "healthy tech," and "bright future"
So-so	6310 (24)	"digital dilemma," "robotic technology," "clock," "day," "rely on digital," "too much technology," "affected by," "irreversible," "digital research," "research community," "young people," "consultation," "assistive," "neuroscience," "short-term memory," "cognitive abilities," and "webinar"
Bad	13,385 (50.91)	"digital dementia: a modern plague," "destroying memory," "stress anxiety," "mental ill," "suffer," "cause death," "kill brain," "stop multitask," "note overload," "AI destroy," "young blind," "destroying ability," "neurobiologist warns," "destroying brain," reliant on devices," and "loss"
Wretched	1107 (4.21)	"making us stupid," "digital dementia s*cks," "dementia," "disorder," "diseases," Parkinson's," "Alzheimer," "worst," "enslaved by," "bad," "hate," "disrupting mental health," "children stupid," and "tech overuse"
No sentiment	2366 (9)	"diagnosis changed," "info," "usual," "identification," "contemplate," and "sum up"

XSL•FO
RenderX

Table 5. Sentiment analysis of the influencer tweets (n=3927).

Main themes and subthemes	Tweets, n (%)	Examples of words or #hashtags
Cognitive decline (n=1526, 38.86%)		
Great	41 (2.69)	“brain training,” “cognitive enhancement,” and “improvement”
Swell	169 (11.07)	“brain games,” “mindfulness,” “yoga,” and “healthy”
So-so	498 (32.63)	“brain supplements,” “sleep quality,” “engagement,” and “therapy”
Bad	522 (34.21)	“sleep deprivation,” “smoking,” “sedentary,” “poor diet,” “depression,” “chronic,” “injury,” and “neurological disorder”
Wretched	159 (10.42)	“dementia,” “Alzheimer’s,” “Parkinson’s,” “Huntington’s disease,” “injury,” and “impairment”
No sentiment	137 (8.98)	“brain function,” “brain structure,” “neurological changes,” and “age”
Digital dependency (n=1218, 31.02%)		
Great	22 (1.81)	“detox,” “mindfulness,” and “balanced”
Swell	263 (21.59)	“digital,” “support groups,” “regulation,” “control,” “coverage,” and “ongoing research”
So-so	315 (25.86)	“consumption,” “habits,” “overuse,” “prevention,” “dependence treatment,” “literacy,” and “media impact”
Bad	486 (39.9)	“addiction,” “compulsive,” “withdrawal,” “loss,” “low productivity,” “divide,” and “inequality”
Wretched	60 (4.93)	“isolation,” “loneliness,” “dehumanization,” “harassment,” “abuse,” “fraud,” “crime,” and “monopoly”
No sentiment	72 (5.91)	“screen time,” “digital media,” “virtual reality,” “statistics,” “facts,” and “knowledge”
Technology overload (n=779, 19.84%)		
Great	79 (10.14)	“mindful use,” “balance,” and “moderation”
Swell	125 (16.05)	“Advancements,” “innovation,” “new devices,” and “smart homes”
So-so	160 (20.54)	“Technology,” “screen,” “distractions,” “attention span,” “social media,” “online learning,” and “virtual meetings”
Bad	231 (29.65)	“digital fatigue,” “burnout,” “stress,” “anxiety,” “depression,” “loneliness,” “trolls,” “fake news,” and “scams”
Wretched	98 (12.58)	“addiction,” “dependency,” “isolate,” “disrupt,” “dehumanization,” “harmful,” and “digital divide”
No sentiment	86 (11.04)	“electronic device,” “visualization,” “network,” “found,” and “daily”
Coping strategies (n=404, 10.29%)		
Great	28 (6.93)	“support,” “counseling,” “services,” “forums,” “help,” and “therapy”
Swell	146 (36.14)	“meditation,” “yoga,” “breath,” “creative,” “hobbies,” and “exercise”
So-so	65 (16.09)	“thinking,” “relaxation,” “techniques,” “time,” “map,” “behavioral,” and “peer”
Bad	88 (21.78)	“substance,” “abuse,” “binge eating,” “self-harm,” “avoidance,” and “procrastination”
Wretched	30 (7.43)	“suicide,” “destructive,” “hopelessness,” “helplessness,” and “desperation”
No sentiment	47 (11.63)	“masks and sanitizers,” “info changing,” “prove hypothesis,” and “CDC”[a]

[a]CDC: Centers for Disease Control and Prevention.

To gather more detailed statistics about the selected tweets within each theme, a one-way ANOVA for each theme was conducted. The total number of observations for collected tweets decreased during the content analysis because we excluded the observations for “no sentiment” tweets, which were treated as missing values. Before conducting the ANOVA, Shapiro-Wilk tests were performed to assess the normality of sentiment scores within each group. The results revealed that the data for each theme did not significantly deviate from normality.

Table S1 in Multimedia Appendix 1 analyzed the sentiment scores for different themes, revealing varying perceptions among the themes. Overall, the total mean sentiment score across all themes was –0.1072 (SD 0.4276) from 3585 samples.

As shown in Table S2 in Multimedia Appendix 1, the ANOVA results indicate significant differences in mean sentiment scores among the themes. The data’s total variation was divided into 2 components: variation between groups and within groups.

The between-group variation was measured for this study; the between-group sum of squares (SS) was 15.5574, with 3 *df*, and the mean square (MS) for between-group variation was calculated as SS divided by *df*, resulting in 5.1858. The *F*-statistic, the ratio of the between-group MS to the

XSL•FO
RenderX

within-group MS, was 29.03, and the associated *P* value (probability>*F*) of <.001 indicates statistically significant differences in mean sentiment scores among the themes. This significant between-groups variation suggests that the average sentiment scores are different across the themes, meaning that each theme was perceived differently by the public, reflecting distinct emotional responses to all themes.

The study measured the variation within groups, finding a within-group SS of 639.7236 with 3581 *df*. The within-group MS was calculated as 0.1786 (ie, SS divided by *df*). The total SS of 655.281 represents the overall variability of the data, including both between-group and within-group variations. The within-group variation indicates the extent of individual differences in sentiment scores within each theme, highlighting the wide variability of individual responses within each theme; it is important for understanding the spread and consistency of sentiments within each theme.

A Bartlett test for equal variances was performed to assess the homogeneity of variances among the different themes (χ^2_3=82.1265; *P*<.001). This indicates that the variances among the themes are not equal, suggesting heterogeneity of variances. The variability in responses differs across themes, possibly due to the varied perceptions of the population.

The Scheffé test for multiple comparisons revealed significant differences between specific themes (Table S3 in Multimedia Appendix 1). "Digital dependency" had a significantly different mean sentiment score from "cognitive decline" by 0.065066 (*P*=.002). "Technology overload" differed substantially from "cognitive decline" by 0.097944 (*P*<.001) but not from "digital dependency" (0.032878; *P*=.455). "Coping strategies" had significantly higher mean sentiment scores compared with "cognitive decline" by 0.223505 (*P*<.001), "digital dependency" by 0.158439 (*P*<.001), and "technology overload" by 0.125561 (*P*<.001). These results indicate that while "cognitive decline," "digital dependency," and "technology overload" are associated with negative sentiments, "coping strategies" evoke a more positive response, with significant differences among these themes in public perception.

Discussion

Principal Findings

This study's purpose was to identify topics related to digital dementia by examining the content and sentiment of tweets. It aimed to provide insights that could help manage the excessive use of digital devices, which has been linked to early-onset dementia [62]. The analysis of tweets showed that digital dementia is generally perceived negatively. The study found that negative tweets were prevalent, especially those related to cognitive decline, with an overwhelmingly negative sentiment rate of almost 70%. Furthermore, it is worth noting that digital dementia is a prevalent term used among Twitter users, and 38.86% (1526/3927) of tweets gathered were associated with cognitive decline in this study, highlighting the recognition of digital dementia as an issue in contemporary society.

According to this study, 55.12% (14492/26290) of all the tweets expressed unfavorable emotions. In addition, negative feelings, which included cognitive decline, digital dependency, technology overload, and coping strategies, predominated across all main themes based on the content analysis. These findings highlight the widespread negative perception of digital dementia. Matthews et al [63] have highlighted the growing concern of digital dementia, where excessive screen time among millennials and Generation Z is leading to increasing symptoms of mild cognitive impairment and Alzheimer disease and related dementias (ADRD); the study also indicated that this phenomenon could result in significant societal and economic burdens on the already overburdened health care services, emphasizing the need for coping strategies from the public and private sectors to mitigate these medical threats [54].

On the basis of the findings from the tweets, users identified cognitive decline as the most prominent issue (1526/3972, 38.86%). In addition, users reported experiencing sleep disorders, depression, and neurological disorders. Excessive use of digital devices can lead to digital dementia, which negatively impacts cognitive function [64]. Individuals who excessively use digital devices may experience cognitive decline and significant negative changes in personality, mood, and social interaction and even exhibit psychiatric symptoms. Moreover, excessive digital device use may lead to poorer social functioning, disrupted self-care routines, higher anxiety scores, and increased sleep disturbances for those affected by ADRD [65,66]. A literature review on the effects of screen time on children and adolescents suggests that media type, duration, content, and after-dark use of mobile devices are highly associated with neurological adverse consequences [67].

Digital dependency and technology overload are also recognized as dilemmas. The analysis of the tweets from this study showed that 31.02% (1218/3927) of the users were concerned about being too dependent on digital devices, while 19.84% (779/3927) were worried about being overloaded with technology. Importantly, the users mentioned that addiction problems, low productivity, and depression could arise from using digital devices excessively. A study conducted in South Korea reported positive correlations between smartphone addiction levels and depression, and women were found to be more susceptible to smartphone addiction than men [68]. Another study indicated that excessive use of digital technology and work pressure lead to burnout in journalists. The study highlights the need for appropriate resources and rest to mitigate the effects of digital technology use [69]. These findings highlight the growing apprehension regarding the impact of digital device use on people's daily lives. It is vital to thoroughly examine the advantages and disadvantages of using such devices and increase awareness of the associated risks in the near future.

Among the 4 main themes, coping strategies were emphasized. The results showed that engaging in alternative activities, reducing digital device use, and seeking public support could prevent digital dementia. Regarding sentiment analysis of the coping strategies, 36.1% (146/404) of the users expressed a positive sentiment, proposing participation in yoga, meditation, and exercise.

Regular physical activity can help prevent ADRD by increasing cognitive reserve [70]. Engaging in high-intensity training for

XSL•FO
RenderX

at least 20 minutes on ≥3 days per week or participating in moderate-intensity training on ≥5 days per week is considered physical activity and positively impacts brain neurodegenerative diseases [71]. In addition, incorporating yoga and meditation interventions into the prevention process of dementia can be beneficial in preventing the onset of the condition. Modifying lifestyle factors such as dietary changes and physical activity through yoga and meditation intervention could be a cost-effective approach to managing dementia [72,73]. Wilmer et al [23] also indicated that the temporary cessation of smartphone use has positive effects, potentially alleviating digital fatigue. Recent research findings have shown that reducing the time spent using digital devices can significantly help prevent cognitive impairment and lessen the onset of digital dementia [74,75]. Several studies suggest reducing screen time can improve concentration, learning, and memory [76,77]. It can also enhance psychological well-being and reduce experiences of anxiety and depressive moods [78,79]. Furthermore, decreasing the use of digital devices can improve sleep quality and overall mental health [80]. Notably, there has been a global movement toward addressing digital dependency. World Digital Detox Day is observed to cope with the prolonged exposure to digital technologies while enhancing awareness of the negative effects of using digital devices [81]. Digital-free tourism has also become an emerging solution for family and social engagements [82]. A digital detox program in 2 preparatory governmental schools in Egypt indicated positive results; the high rate of screen addiction among students dropped to 14.3% in the posttest measures compared with 20% in the pretest measures [83].

Strengths and Limitations

This study explored the reactions and perceptions of digital dementia among Twitter users, taking a comprehensive approach to vividly portray users' attitudes toward digital dementia. Using Twitter as the data source facilitated a compelling exploration of public sentiment. The methodology used enabled the creation of a sizable data set, which is difficult to achieve through traditional survey methods. Importantly, the authors believe this is the first study seeking public perception of digital dementia via Twitter.

There are several limitations to note. First, it is important to note that the study's results may not accurately represent the entire population's views on digital dementia due to the high concentration of Twitter users [84]. Thus, it is crucial to understand that the opinions expressed in the study reflect only a subset of individuals rather than the entire population's views.

Although the study extracted a large amount of data, only Twitter was used for this study. Other social media platforms could provide more information regarding digital dementia. Twitter limits users to 140 characters per tweet, restricting users from fully expressing opinions or sentiments. This constraint raises concerns about the comprehensiveness of user contributions. For instance, there may be more 2-way communication on social networking sites such as Facebook, which may differ from that of Twitter [85]. As the proposed hashtags are also available on other social media platforms, such as Facebook and Instagram, future research could explore whether engagement of users differs from that of Twitter.

Previous research has shown a higher proportion of Twitter users in urban areas compared with rural areas, which could introduce biases [86]. Furthermore, Twitter is overrepresented by a younger population aged 25 to 34 years compared with the general population [87].

In addition, the analysis was restricted to English-language tweets; insights from other languages could provide valuable perspectives. Finally, tweets containing emojis were excluded from this study's NLP task, having a sufficient number of tweets to carry out sentiment and topic analysis. However, incorporating emojis could have provided valuable insights into users' emotions, resulting in a more nuanced sentiment analysis. That said, future studies should consider analyzing emojis based on a shorter period.

Conclusions

The study emphasizes the various concerns associated with digital dementia via Twitter. By exploring the content and sentiment of the recent tweets regarding digital dementia, the findings suggest that the perception of digital device use is increasingly negative, with rising concerns about its effects on mental health. Moreover, individuals are worrisome of digital dependency and technology overload in everyday lives. To address these concerns effectively, it is necessary to implement a range of coping strategies, such as yoga, exercise, and meditation, to prevent early-onset dementia. The results of this study could serve as the foundation for those performing research regarding digital dementia and for the betterment of the dementia community.

Acknowledgments

The authors received no financial support for the research, authorship, and publication of this manuscript.

Data Availability

The raw data for content analysis can be found in Multimedia Appendix 2, and the principal codes for data sets generated and analyzed during this study are available from GitHub [88].

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Sentiment score by theme, ANOVA, and comparison of sentiment score by theme (Scheffé).
[DOCX File , 16 KB - jmir_v26i1e59546_app1.docx]

Multimedia Appendix 2
Raw data used for content analysis.
[XLSX File (Microsoft Excel File), 342 KB - jmir_v26i1e59546_app2.xlsx]

References

1. Moledina S, Khoja A. Letter to the editor: digital dementia-is smart technology making us dumb? Ochsner J 2018;18(1):12 [FREE Full text] [Medline: 29559861]
2. Hendriks S, Peetoom K, Bakker C, van der Flier WM, Papma JM, Koopmans R, et al. Global prevalence of young-onset dementia: a systematic review and meta-analysis. JAMA Neurol 2021 Sep 01;78(9):1080-1090 [FREE Full text] [doi: 10.1001/jamaneurol.2021.2161] [Medline: 34279544]
3. Global dementia cases forecasted to triple by 2050. Alzheimer's Association®. URL: https://aaic.alz.org/releases_2021/global-prevalence.asp [accessed 2024-01-10]
4. Manwell LA, Tadros M, Ciccarelli TM, Eikelboom R. Digital dementia in the internet generation: excessive screen time during brain development will increase the risk of Alzheimer's disease and related dementias in adulthood. J Integr Neurosci 2022 Jan 28;21(1):28 [FREE Full text] [doi: 10.31083/j.jin2101028] [Medline: 35164464]
5. Bozzola E, Spina G, Agostiniani R, Barni S, Russo R, Scarpato E, et al. The use of social media in children and adolescents: scoping review on the potential risks. Int J Environ Res Public Health 2022 Aug 12;19(16):9960 [FREE Full text] [doi: 10.3390/ijerph19169960] [Medline: 36011593]
6. Cardoso-Leite P, Buchard A, Tissieres I, Mussack D, Bavelier D. Media use, attention, mental health and academic performance among 8 to 12 year old children. PLoS One 2021 Nov 17;16(11):e0259163 [FREE Full text] [doi: 10.1371/journal.pone.0259163] [Medline: 34788306]
7. Dienlin T, Johannes N. The impact of digital technology use on adolescent well-being . Dialogues Clin Neurosci 2020 Jun;22(2):135-142 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/tdienlin] [Medline: 32699513]
8. Small GW, Lee J, Kaufman A, Jalil J, Siddarth P, Gaddipati H, et al. Brain health consequences of digital technology use . Dialogues Clin Neurosci 2020 Jun;22(2):179-187 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
9. Montag C, Zhao Z, Sindermann C, Xu L, Fu M, Li J, et al. Internet communication disorder and the structure of the human brain: initial insights on WeChat addiction. Sci Rep 2018 Mar 01;8(1):2155 [FREE Full text] [doi: 10.1038/s41598-018-19904-y] [Medline: 29391461]
10. Ioannidis K, Hook R, Goudriaan AE, Vlies S, Fineberg NA, Grant JE, et al. Cognitive deficits in problematic internet use: meta-analysis of 40 studies. Br J Psychiatry 2019 Nov 20;215(5):639-646 [FREE Full text] [doi: 10.1192/bjp.2019.3] [Medline: 30784392]
11. Ward AF, Duke K, Gneezy A, Bos MW. Brain drain: the mere presence of one’s own smartphone reduces available cognitive capacity. J Assoc Consum Res 2017 Apr;2(2):140-154. [doi: 10.1086/691462]
12. Griffiths LJ, Dowda M, Dezateux C, Pate R. Associations between sport and screen-entertainment with mental health problems in 5-year-old children. Int J Behav Nutr Phys Act 2010 Apr 21;7(1):30 [FREE Full text] [doi: 10.1186/1479-5868-7-30] [Medline: 20409310]
13. Manganello JA, Taylor CA. Television exposure as a risk factor for aggressive behavior among 3-year-old children. Arch Pediatr Adolesc Med 2009 Nov 02;163(11):1037-1045. [doi: 10.1001/archpediatrics.2009.193] [Medline: 19884595]
14. Salicetia F. Internet addiction disorder (IAD). Procedia Soc Behav Sci 2015 Jun;191:1372-1376. [doi: 10.1016/j.sbspro.2015.04.292]
15. Tanil CT, Yong MH. Mobile phones: the effect of its presence on learning and memory. PLoS One 2020 Aug 13;15(8):e0219233 [FREE Full text] [doi: 10.1371/journal.pone.0219233] [Medline: 32790667]
16. Skowronek J, Seifert A, Lindberg S. The mere presence of a smartphone reduces basal attentional performance. Sci Rep 2023 Jun 08;13(1):9363 [FREE Full text] [doi: 10.1038/s41598-023-36256-4] [Medline: 37291199]
17. Canale N, Vieno A, Doro M, Rosa Mineo E, Marino C, Billieux J. Emotion-related impulsivity moderates the cognitive interference effect of smartphone availability on working memory. Sci Rep 2019 Dec 06;9(1):18519 [FREE Full text] [doi: 10.1038/s41598-019-54911-7] [Medline: 31811205]
18. Adovor E, Czaika M, Docquier F, Moullan Y. Medical brain drain: how many, where and why? J Health Econ 2021 Mar;76:102409. [doi: 10.1016/j.jhealeco.2020.102409] [Medline: 33465558]
19. Ostic D, Qalati SA, Barbosa B, Shah SM, Galvan Vela E, Herzallah AM, et al. Effects of social media use on psychological well-being: a mediated model. Front Psychol 2021 Jun 21;12:678766 [FREE Full text] [doi: 10.3389/fpsyg.2021.678766] [Medline: 34234717]

XSL•FO
RenderX

20. Reeves B, Robinson T, Ram N. Time for the human screenome project. Nature 2020 Jan 15;577(7790):314-317. [doi: 10.1038/d41586-020-00032-5] [Medline: 31942062]
21. Engineer M, Kot S, Dixon E. Investigating the readability and linguistic, psychological, and emotional characteristics of digital dementia information written in the English language: multitrait-multimethod text analysis. JMIR Form Res 2023 Oct 25;7:e48143 [FREE Full text] [doi: 10.2196/48143] [Medline: 37878351]
22. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the internet may be changing our cognition. World Psychiatry 2019 Jun 06;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]
23. Wilmer HH, Sherman LE, Chein JM. Smartphones and cognition: a review of research exploring the links between mobile technology habits and cognitive functioning. Front Psychol 2017;8:605 [FREE Full text] [doi: 10.3389/fpsyg.2017.00605] [Medline: 28487665]
24. Sparrow B, Liu J, Wegner DM. Google effects on memory: cognitive consequences of having information at our fingertips. Science 2011 Aug 05;333(6043):776-778. [doi: 10.1126/science.1207745] [Medline: 21764755]
25. Madigan S, Browne D, Racine N, Mori C, Tough S. Association between screen time and children's performance on a developmental screening test. JAMA Pediatr 2019 Mar 01;173(3):244-250 [FREE Full text] [doi: 10.1001/jamapediatrics.2018.5056] [Medline: 30688984]
26. Muppalla S, Vuppalapati S, Reddy Pulliahgaru A, Sreenivasulu H. Effects of excessive screen time on child development: an updated review and strategies for management. Cureus 2023 Jun;15(6):e40608 [FREE Full text] [doi: 10.7759/cureus.40608] [Medline: 37476119]
27. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41:101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
28. Yang YT, Horneffer M, DiLisio N. Mining social media and web searches for disease detection. J Public Health Res 2013 Apr 28;2(1):17-21 [FREE Full text] [doi: 10.4081/jphr.2013.e4] [Medline: 25170475]
29. Ye X, Li S, Yang X, Qin C. Use of social media for the detection and analysis of infectious diseases in China. Int J Geo-Inf 2016 Aug 30;5(9):156. [doi: 10.3390/ijgi5090156]
30. Yang Y, Tsao SF, Basri MA, Chen HH, Butt ZA. Digital disease surveillance for emerging infectious diseases: an early warning system using the internet and social media data for COVID-19 forecasting in Canada. Stud Health Technol Inform 2023 May 18;302:861-865. [doi: 10.3233/SHTI230290] [Medline: 37203518]
31. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 2017 Dec;18:43-49 [FREE Full text] [doi: 10.1016/j.copsyc.2017.02.009]
32. Aslam S. Twitter by the numbers: stats, demographics & fun facts. Omnicore. 2024. URL: https://www.omnicoreagency.com/twitter-statistics [accessed 2024-01-30]
33. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep 2017 Oct 11;7(1):13006 [FREE Full text] [doi: 10.1038/s41598-017-12961-9] [Medline: 29021528]
34. Berry N, Lobban F, Belousov M, Emsley R, Nenadic G, Bucci S. #WhyWeTweetMH: understanding why people use Twitter to discuss mental health problems. J Med Internet Res 2017 Apr 05;19(4):e107 [FREE Full text] [doi: 10.2196/jmir.6173] [Medline: 28381392]
35. Alvarez-Mon MA, Asunsolo Del Barco A, Lahera G, Quintero J, Ferre F, Pereira-Sanchez V, et al. Increasing interest of mass communication media and the general public in the distribution of tweets about mental disorders: observational study. J Med Internet Res 2018 May 28;20(5):e205 [FREE Full text] [doi: 10.2196/jmir.9582] [Medline: 29807880]
36. Tweepy: Twitter for Python. GitHub. 2023. URL: https://github.com/tweepy/tweepy [accessed 2023-09-17]
37. Terms of service. Twitter. URL: https://twitter.com/en/tos [accessed 2023-09-30]
38. Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. Commun ACM 2016 Jun 24;59(7):96-104. [doi: 10.1145/2818717]
39. Observatory on social media. Botometer by OSoMe. URL: https://botometer.osome.iu.edu/ [accessed 2023-10-16]
40. Ramagopalan S, Wasiak R, Cox AP. Using Twitter to investigate opinions about multiple sclerosis treatments: a descriptive, exploratory study. F1000Res 2014;3:216 [FREE Full text] [doi: 10.12688/f1000research.5263.1] [Medline: 25520780]
41. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep 11;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
42. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 01;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
43. Evans NJ, Phua J, Lim J, Jun H. Disclosing Instagram influencer advertising: the effects of disclosure language on advertising recognition, attitudes, and behavioral intent. J Interact Advert 2017 Sep 14;17(2):138-149. [doi: 10.1080/15252019.2017.1366885]
44. Buvár Á, Szilágyi SF, Balogh E, Zsila Á. COVID-19 messages in sponsored social media posts: the positive impact of influencer-brand fit and prior parasocial interaction. PLoS One 2022;17(10):e0276143 [FREE Full text] [doi: 10.1371/journal.pone.0276143] [Medline: 36240138]

XSL•FO
RenderX

45. Kostygina G, Tran H, Binns S, Szczypka G, Emery S, Vallone D, et al. Boosting health campaign reach and engagement through use of social media influencers and memes. Soc Media Soc 2020 May 06;6(2):205630512091247. [doi: 10.1177/2056305120912475]
46. How Americans use social media. Pew Research Center. URL: https://www.pewresearch.org/internet/2023/05/31/how-americans-use-social-media [accessed 2024-05-31]
47. Malik A, Heyman-Schrum C, Johri A. Use of Twitter across educational settings: a review of the literature. Int J Educ Technol High Educ 2019 Sep 25;16(1):36 [FREE Full text] [doi: 10.1186/s41239-019-0166-x]
48. Mabkhot H, Isa NM, Mabkhot A. The influence of the credibility of social media influencers SMIs on the consumers' purchase intentions: evidence from Saudi Arabia. Sustainability 2022 Sep 28;14(19):12323 [FREE Full text] [doi: 10.3390/su141912323]
49. Mardon R, Cocker H, Daunt K. How social media influencers impact consumer collectives: an embeddedness perspective. J Consum Res 2023;50(3):617-644. [doi: 10.1093/jcr/ucad003]
50. Keddem S, Agha A, Morawej S, Buck A, Cronholm P, Sonalkar S, et al. Characterizing Twitter content about HIV pre-exposure prophylaxis (PrEP) for women: qualitative content analysis. J Med Internet Res 2023 May 11;25:e43596 [FREE Full text] [doi: 10.2196/43596] [Medline: 37166954]
51. Nabi RL, Prestin A, So J. Facebook friends with (health) benefits? exploring social network site use and perceptions of social support, stress, and well-being. Cyberpsychol Behav Soc Netw 2013 Oct;16(10):721-727. [doi: 10.1089/cyber.2012.0521] [Medline: 23790356]
52. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May 16;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
53. Glaser B, Strauss A. Discovery of Grounded Theory: Strategies for Qualitative Research. Oxfordshire, UK: Routledge; 1999.
54. Jansen BJ, Zhang M, Sobel K, Chowdury A. Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci 2009 Jul 06;60(11):2169-2188. [doi: 10.1002/asi.21149]
55. Kirkpatrick W, Abboudi J, Kim N, Medina J, Maltenfort M, Seigerman D, et al. An assessment of online reviews of hand surgeons. Arch Bone Jt Surg 2017 May;5(3):139-144 [FREE Full text] [Medline: 28656160]
56. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
57. Saldaña J. The Coding Manual for Qualitative Researchers. 3rd edition. Thousand Oaks, CA: Sage Publications; 2017.
58. Twenge JM, Campbell WK. Associations between screen time and lower psychological well-being among children and adolescents: evidence from a population-based study. Prev Med Rep 2018 Dec;12:271-283 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.003] [Medline: 30406005]
59. Karim F, Oyewande AA, Abdalla LF, Chaudhry Ehsanullah R, Khan S. Social media use and its connection to mental health: a systematic review. Cureus 2020 Jun 15;12(6):e8627 [FREE Full text] [doi: 10.7759/cureus.8627] [Medline: 32685296]
60. Eppler MJ, Mengis J. The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. Inf Soc 2004 Nov;20(5):325-344. [doi: 10.1080/01972240490507974]
61. Kross E, Verduyn P, Sheppes G, Costello CK, Jonides J, Ybarra O. Social media and well-being: pitfalls, progress, and next steps. Trends Cogn Sci 2021 Jan;25(1):55-66 [FREE Full text] [doi: 10.1016/j.tics.2020.10.005] [Medline: 33187873]
62. Viguria I, Alvarez-Mon MA, Llavero-Valero M, Asunsolo Del Barco A, Ortuño F, Alvarez-Mon M. Eating disorder awareness campaigns: thematic and quantitative analysis using Twitter. J Med Internet Res 2020 Jul 14;22(7):e17626 [FREE Full text] [doi: 10.2196/17626] [Medline: 32673225]
63. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. Alzheimers Dement 2019 Jan;15(1):17-24 [FREE Full text] [doi: 10.1016/j.jalz.2018.06.3063] [Medline: 30243772]
64. What Is "digital dementia", and is it a harmful impact of technology? Discover. URL: https://www.discovermagazine.com/health/technology-overuse-and-the-fear-of-digital-dementia-what-you-need-to-know [accessed 2023-12-22]
65. Bacsu J, Cammer A, Ahmadi S, Azizi M, Grewal KS, Green S, et al. Examining the Twitter discourse on dementia during Alzheimer's awareness month in Canada: infodemiology study. JMIR Form Res 2022 Oct 26;6(10):e40049 [FREE Full text] [doi: 10.2196/40049] [Medline: 36287605]
66. Roberts K, Dowell A, Nie JB. Attempting rigour and replicability in thematic analysis of qualitative research data: a case study of codebook development. BMC Med Res Methodol 2019 Mar 28;19(1):66 [FREE Full text] [doi: 10.1186/s12874-019-0707-y] [Medline: 30922220]
67. Lissak G. Adverse physiological and psychological effects of screen time on children and adolescents: literature review and case study. Environ Res 2018 Jul;164:149-157. [doi: 10.1016/j.envres.2018.01.015] [Medline: 29499467]
68. Lee KE, Kim SH, Ha TY, Yoo YM, Han JJ, Jung JH, et al. Dependency on smartphone use and its association with anxiety in Korea. Public Health Rep 2016;131(3):411-419 [FREE Full text] [doi: 10.1177/003335491613100307] [Medline: 27252561]

XSL•FO
RenderX

69. Sanchez-Segura MI, Dugarte-Peña GL, Medina-Dominguez F, Amescua Seco A, Menchen Viso R. Digital transformation in organizational health and safety to mitigate burnout syndrome. Front Public Health 2023 Mar 21;11:1080620 [FREE Full text] [doi: 10.3389/fpubh.2023.1080620] [Medline: 37026125]
70. Cheng ST. Cognitive reserve and the prevention of dementia: the role of physical and cognitive activities. Curr Psychiatry Rep 2016 Sep 1;18(9):85 [FREE Full text] [doi: 10.1007/s11920-016-0721-2] [Medline: 27481112]
71. Ruscheweyh R, Willemer C, Krüger K, Duning T, Warnecke T, Sommer J, et al. Physical activity and memory functions: an interventional study. Neurobiol Aging 2011 Jul;32(7):1304-1319. [doi: 10.1016/j.neurobiolaging.2009.08.001] [Medline: 19716631]
72. Bali P, Kaur N, Tiwari A, Bammidi S, Podder V, Devi C, et al. Effectiveness of yoga as the public health intervention module in the management of diabetes and diabetes associated dementia in South East Asia: a narrative review. Neuroepidemiology 2020 Feb 19;54(4):287-303 [FREE Full text] [doi: 10.1159/000505816] [Medline: 32074622]
73. Russell-Williams J, Jaroudi W, Perich T, Hoscheidt S, El Haj M, Moustafa AA. Mindfulness and meditation: treating cognitive impairment and reducing stress in dementia. Rev Neurosci 2018 Sep 25;29(7):791-804. [doi: 10.1515/revneuro-2017-0066] [Medline: 29466242]
74. Nakshine VS, Thute P, Khatib MN, Sarkar B. Increased screen time as a cause of declining physical, psychological health, and sleep patterns: a literary review. Cureus 2022 Oct;14(10):e30051 [FREE Full text] [doi: 10.7759/cureus.30051] [Medline: 36381869]
75. Wu H, Gu Y, Du W, Meng G, Wu H, Zhang S, et al. Different types of screen time, physical activity, and incident dementia, Parkinson's disease, depression and multimorbidity status. Int J Behav Nutr Phys Act 2023 Nov 03;20(1):130 [FREE Full text] [doi: 10.1186/s12966-023-01531-0] [Medline: 37924067]
76. Christakis DA, Zimmerman FJ, DiGiuseppe DL, McCarty CA. Early television exposure and subsequent attentional problems in children. Pediatrics 2004 Apr 01;113(4):708-713. [doi: 10.1542/peds.113.4.708] [Medline: 15060216]
77. Christakis DA. The effects of infant media usage: what do we know and what should we learn? Acta Paediatr 2009 Jan;98(1):8-16. [doi: 10.1111/j.1651-2227.2008.01027.x] [Medline: 18793294]
78. Shakya HB, Christakis NA. Association of Facebook use with compromised well-being: a longitudinal study. Am J Epidemiol 2017 Mar 01;185(3):203-211. [doi: 10.1093/aje/kww189] [Medline: 28093386]
79. Twenge JM. The age of anxiety? birth cohort change in anxiety and neuroticism, 1952-1993. J Pers Soc Psychol 2000 Dec;79(6):1007-1021. [doi: 10.1037//0022-3514.79.6.1007] [Medline: 11138751]
80. Perrault AA, Bayer L, Peuvrier M, Afyouni A, Ghisletta P, Brockmann C, et al. Reducing the use of screen electronic devices in the evening is associated with improved sleep and daytime vigilance in adolescents. Sleep 2019 Sep 06;42(9):31260534. [doi: 10.1093/sleep/zsz125] [Medline: 31260534]
81. Chaudhari R, Sharma N. World digital detox day: a global movement to address digital dependency. Cambridge Open Engage 2024. [doi: 10.33774/coe-2024-h045s]
82. Hassan TH, Salem AE, Saleh MI. Digital-free tourism holiday as a new approach for tourism well-being: tourists' attributional approach. Int J Environ Res Public Health 2022 May 14;19(10):5974 [FREE Full text] [doi: 10.3390/ijerph19105974] [Medline: 35627511]
83. Mohamed SM, Abdallah LS, Ali FN. Effect of digital detox program on electronic screen syndrome among preparatory school students. Nurs Open 2023 Apr 14;10(4):2222-2228 [FREE Full text] [doi: 10.1002/nop2.1472] [Medline: 36373487]
84. Panda A, Sharma A, Dundar A, Packard A, Aase L, Kotsenas A, et al. Twitter use by academic nuclear medicine programs: pilot content analysis study. JMIR Form Res 2021 Nov 08;5(11):e24448 [FREE Full text] [doi: 10.2196/24448] [Medline: 34747708]
85. Thackeray R, Neiger BL, Burton SH, Thackeray CR. Analysis of the purpose of state health departments' tweets: information sharing, engagement, and action. J Med Internet Res 2013 Nov 11;15(11):e255 [FREE Full text] [doi: 10.2196/jmir.3002] [Medline: 24217361]
86. Hswen Y, Hawkins JB, Sewalk K, Tuli G, Williams DR, Viswanath K, et al. Racial and ethnic disparities in patient experiences in the United States: 4-year content analysis of Twitter. J Med Internet Res 2020 Aug 21;22(8):e17048 [FREE Full text] [doi: 10.2196/17048] [Medline: 32821062]
87. Distribution of X (formerly Twitter) users worldwide as of April 2024, by age group. Statista. URL: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/ [accessed 2024-01-10]
88. Brochan8 / ddm_sentiment. GitHub. URL: https://github.com/Brochan8/ddm_sentiment [accessed 2024-04-29]

Abbreviations

ADRD: Alzheimer disease and related dementias
MS: mean square
NLP: natural language processing
SS: sum of squares
VADER: Valence Aware Dictionary and Sentiment Reasoner

XSL•FO
RenderX

Edited by G Eysenbach, A Mavragani; submitted 15.04.24; peer-reviewed by C Ruiz-Nunez, L Southwick; comments to author 10.05.24; revised version received 30.05.24; accepted 24.06.24; published 16.07.24.

Please cite as:
Cho H, Kim KM, Kim JY, Youn BY
Twitter Discussions on #digitaldementia: Content and Sentiment Analysis
J Med Internet Res 2024;26:e59546
URL: https://www.jmir.org/2024/1/e59546
doi:10.2196/59546
PMID:39012679

©Hyeongchan Cho, Kyu-Min Kim, Jee-Young Kim, Bo-Young Youn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identifying Medicine Shortages With the Twitter Social Network: Retrospective Observational Study

Doerine J Postma[1,2], MSci, PharmD; Magali L A Heijkoop[1], MSci, PharmD; Peter A G M De Smet[3], Prof Dr; Kim Notenboom[4], PhD; Hubert G M Leufkens[1], Prof Dr; Aukje K Mantel-Teeuwisse[1], Prof Dr

[1]Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, Netherlands

[2]Royal Dutch Pharmacists Association, The Hague, Netherlands

[3]Departments of IQ Healthcare and Clinical Pharmacy, Radboud Institute for Health Sciences, Radboud University Medical Centre, Nijmegen, Netherlands

[4]Dutch Medicines Evaluation Board, Utrecht, Netherlands

Corresponding Author:
Aukje K Mantel-Teeuwisse, Prof Dr
Division of Pharmacoepidemiology and Clinical Pharmacology
Utrecht Institute for Pharmaceutical Sciences (UIPS)
Utrecht University
Universiteitsweg 99
Utrecht, 3584 CG
Netherlands
Phone: 31 302537324
Email: a.k.mantel@uu.nl

Abstract

Background: Early identification is critical for mitigating the impact of medicine shortages on patients. The internet, specifically social media, is an emerging source of health data.

Objective: This study aimed to explore whether a routine analysis of data from the Twitter social network can detect signals of a medicine shortage and serve as an early warning system and, if so, for which medicines or patient groups.

Methods: Medicine shortages between January 31 and December 1, 2019, were collected from the Dutch pharmacists' society's national catalog Royal Dutch Pharmacists Association (KNMP) Farmanco. Posts on these shortages were collected by searching for the name, the active pharmaceutical ingredient, or the first word of the brand name of the medicines in shortage. Posts were then selected based on relevant keywords that potentially indicated a shortage and the percentage of shortages with at least 1 post was calculated. The first posts per shortage were analyzed for their timing (median number of days, including the IQR) versus the national catalog, also stratified by disease and medicine characteristics. The content of the first post per shortage was analyzed descriptively for its reporting stakeholder and the nature of the post.

Results: Of the 341 medicine shortages, 102 (29.9%) were mentioned on Twitter. Of these 102 shortages, 18 (5.3% of the total) were mentioned prior to or simultaneous to publication by KNMP Farmanco. Only 4 (1.2%) of these were mentioned on Twitter more than 14 days before. On average, posts were published with a median delay of 37 (IQR 7-81) days to publication by KNMP Farmanco. Shortages mentioned on Twitter affected a greater number of patients and lasted longer than those that were not mentioned. We could not conclusively relate either the presence or absence on Twitter to a disease area or route of administration of the medicine in shortage. The first posts on the 102 shortages were mainly published by patients (n=51, 50.0%) and health care professionals (n=46, 45.1%). We identified 8 categories of nature of content. Sharing personal experience (n=44, 43.1%) was the most common category.

Conclusions: The Twitter social network is not a suitable early warning system for medicine shortages. Twitter primarily echoes already-known information rather than spreads new information. However, Twitter or potentially any other social media platform provides the opportunity for future qualitative research in the increasingly important field of medicine shortages that investigates how a larger population of patients is affected by shortages.

(J Med Internet Res 2024;26:e51317) doi:10.2196/51317

XSL•FO
RenderX

KEYWORDS

medicine shortages; signal detection; social media; Twitter social network; drug shortage; Twitter

Introduction

The continuous supply of quality-assured, safe, effective, and affordable medicines is one of the building blocks of a well-functioning health system [1]. European legislation stipulates that marketing authorization holders (MAHs) should ensure appropriate and continuous supplies so that the needs of patients in the member state in question are met [2]. Despite this requirement, access to medicines and vaccines is a concern, because of the persistent problems of shortages and stockouts [3]. With the COVID-19 pandemic, this reality was even more pronounced as the problem of shortages of essential medicines worsened [4,5]. It brought the long-standing vulnerability of the medical product supply chain into sharp focus [6].

For mitigating the impact of medicine shortages on patients, early identification is considered critical. It allows for early triaging, assessment, and coordination of medicine shortages to prevent them or minimize their impact [7,8]. The World Health Organization urged member states to implement effective notification systems that allow measures to avoid medicines and vaccine shortages [1].

MAHs in Europe are required to notify competent national authorities of expected supply interruptions 2 months in advance and of unexpected interruptions as soon as possible [2]. Competent authorities are recommended to communicate any nationwide shortage, depending on available resources and the communication needs within their territories [9]. In the Netherlands, competent authorities publish details, actions taken, and advice on the shortages of critical products for health care professionals. Products are considered critical if they meet 2 criteria. First, they are part of the treatment or prevention of a life-threatening or irreversibly progressive disease, and second, there is insufficient availability of equivalent alternatives [10,11]. The publication of all notifications submitted by MAHs to the Dutch authorities would not be meaningful because many of these notifications concern the risk of a supply disruption, but the supply is often unaffected. Most supply disruptions also concern a shortage of a product for which generic substitution is available and this substitution is considered to have a low impact on patients [12].

Health care professionals, especially pharmacists, are confronted daily with supply disruptions and seek information on and solutions for critical and noncritical shortages. In addition to the information authorities publish on shortages, many (European) pharmacists' societies publish national catalogs of identified shortages when a pharmacy cannot supply a requested product to patients. However, these catalogs are often published too late to mitigate the shortages' impact [13,14].

Patients may be directly impacted by shortages; they may, for example, receive an alternative therapy or face additional costs [15]. The research on the impact of medicine shortages on patients has been limited to qualitative research on a small number of patients [16]. They could potentially also play a role in the early identification of shortages, but this has not been studied yet. In pharmacovigilance, patients—or the general public—have often flagged medicine-related issues at an early stage, specifically adverse drug reactions [17-19]. How the general public learns about diseases and medicinal products has dramatically changed due to the internet [20,21]. Social media, in particular, are an emerging source of health data [22]. Social media platforms, such as Instagram (Meta Platforms), Facebook (Meta Platforms), and the Twitter social network (X Corp; recently transformed to X), allow users not only to inform themselves but also to reach a general audience of other internet or platform users. Despite the biases that arise when using social media in research [23], we believe that social media platforms should be studied as a possible source of health information.

In health research, the most commonly studied social media platforms are Facebook and Twitter. Studies have shown that Facebook has been primarily used for recruitment, whereas Twitter has been used for content analysis [24], early warning [25], and event detection [26]. Recent examples of Twitter as an early warning system include warnings for new COVID-19 outbreaks [26] and safety warnings on medicinal products [27,28]. Twitter provides a unique big data source for public health researchers because of the concise messages, the real-time nature of its content, and the ease of searching and accessing publicly available information on it [25]. Twitter users post messages ("posts"—formerly known as tweets) with a maximum of 280 characters at the time of research, visible to anyone following the sender. With 2.8 million users in the Netherlands in 2019, Twitter was one of the largest social media platforms [29].

Since patients or the general public can potentially identify shortages and provide early insight into them, this study aimed to explore whether the routine analysis of Twitter data can be used as an additional early warning system for medicine shortages in the Netherlands and, if so, for which medicines or patient groups.

Methods

Data Source

We collected data on temporary and permanent medicine shortages from Royal Dutch Pharmacists Association (KNMP) Farmanco between January 1 and December 31, 2019, the year before the COVID-19 pandemic. In the Netherlands, KNMP Farmanco is the most complete data set of medicine shortages. The Dutch pharmacists' society, KNMP, receives signals on possible shortages from Dutch pharmacy practice. The MAH validates these shortages and may also report shortages itself. KNMP Farmanco publishes data on its website within 24 hours of reports by pharmacists or others [30]. Information on the shortage (such as the product in shortage and the expected resolution date) is publicly available, but information on possible solutions for patients is restricted to health care professionals (mainly pharmacists).

XSL•FO
RenderX

On this website, a shortage is defined as the national unavailability for at least 2 weeks of an active pharmaceutical ingredient with at least 1 medicinal product for human use with marketing authorization. A period of fewer than 2 weeks of unavailability is likely to be mitigated by stock still present in the supply chain, as patients' remaining supply, or as other pharmacies' and wholesalers' stock. Therefore, the KNMP chooses 2 weeks as the cutoff point [31]. The output file that we retrieved for shortages during the study period contained data on the medicine's name, the active pharmaceutical ingredient, the dates the shortage was published and resolved, and possible alternative products for treating patients.

Data on the number and median age of patients affected by the shortages were obtained from the Dutch Foundation for Pharmaceutical Statistics (SFK) by analyzing the number of patients 12 months before the shortage. The SFK only has data on medicines dispensed in community and outpatient pharmacies. No data were available on products dispensed in hospital pharmacies.

We used posts from Twitter due to the advantage of the real-time nature of its content and the ease of searching and accessing publicly available information, despite not being the largest platform in the Netherlands (ranking after for example Facebook [29]). Posts on these shortages were retrieved from the Utrecht Data School. These posts were publicly available and published between January 1 and December 31, 2019. The output file contained data on the post ID, post text, user ID, date and time created, and the name of the medicine mentioned. When a post noted more than one medicine, we included this post separately for each medicine.

Data Selection

A total of 389 medicine shortages occurred in the Netherlands in 2019. We decided only to match a post to a shortage if the post was published no more than 30 days before the shortage was published and no more than 30 days after the shortage ended, according to KNMP Farmanco. We thus included shortages between January 31 and December 1, 2019, resulting in 341 shortages for this study.

For these 341 shortages, Dutch posts were collected by searching for the names of the medicines in shortage: the Anatomical Therapeutic Chemical (ATC) name on the fifth level, which is the active pharmaceutical ingredients [32] or the first word of the brand name. For instance, a dermal product with betamethasone may be best known by its brand name Diprosole, and tablets with clopidogrel may be best known by the brand name Plavix. We excluded reposts. We did not account for misspellings and common wording for groups of medicines, since we were aiming to assess the use of Twitter data in routine analysis.

Posts were then selected based on relevant keywords that potentially indicated a shortage (Multimedia Appendix 1). These keywords were developed in 2 rounds. We initially had a set of 6 keywords (almost out of, available, discontinued, scarce, shortage, and supply) that we as researchers expected to be used. Since this set did not identify the major shortages we had expected, we took a different approach. We analyzed all (>10,000) posts on 20 pharmaceutical substances. Based on this screening, we determined the final set of 30 keywords (adding keywords such as delivery date, deficiency, lack of, missing, present, and stock; Multimedia Appendix 1) that we used for this paper.

Using this final set of keywords, we selected all posts for the 341 shortages in our study. A researcher (MLAH) assessed these posts to determine if the post was about a mentioned medicine shortage. The assessment was based on individual posts and if they were ambiguous, a series of connected posts (thread) and replies to the post were studied. Inconclusive posts were assessed by a second researcher (DJP), and consensus was reached following discussion.

Data Analysis

Mentioned on the Twitter Social Network or Not

We characterized the products in shortage by their ATC class on the first level [32], according to the organ or the system on which they act. We compared the characteristics of the medicine shortages and the patients affected by these shortages, for shortages mentioned and not mentioned on Twitter. We used the Fisher test (exact if possible or Monte Carlo if necessary) for categorical data, and for continuous quantitative data, we used the Levene median test. P values <.05 were considered significant.

Timing of Posts

We analyzed the posts for their timing with the publication of shortages by KNMP Farmanco. If more than 1 post was published during the shortage, we selected the first post chronologically. If the difference in timing of the publications was 2 days or fewer, we considered the publications simultaneously since administrative processing (eg, reporting during weekends) can cause this difference. To determine differences in timing between stakeholder categories, we analyzed only stakeholder categories with more than 10 posts.

Content Analysis

We grouped the senders according to stakeholder category based on the sender's name, the profession stated in his or her profile, and the content of the post message (Textbox 1). The content of the posts was also analyzed and categorized based on its nature. While the first posts served as the main analysis (Textbox 2), all posts were analyzed and categorized. A researcher (MLAH) categorized these posts. A second researcher (DJP) independently categorized inconclusive posts, and if necessary, consensus was reached following the discussion.

XSL•FO
RenderX

Textbox 1. Stakeholder categories.

- Patient: post contains phrase such as “my medicine...”
- Health care professional: post contains phrase such as “my patients...” or profile states profession
- Media: profile or ID states name of media
- Others: profile or ID states name of pharmaceutical industry, government, or health insurers
- Unknown: insufficient information to determine the stakeholder

Textbox 2. Nature of posts.

Sharing personal experience

- Senders share how medicine shortages affect them (personally or professionally) or their loved ones.
- Post contains a phrase such as “My medicine is not available,” or “This medicine is unavailable, and I have to assure patients they can safely use an alternative.”

Sharing emotions

- Senders express their emotional responses to a medicine shortage.
- These posts use words expressing an emotion (eg, angry or sad) or figures of speech, such as sarcasm.
- Post contains a phrase such as “What a nightmare!”

Call to action

- Senders call on influential people or institutions to act.
- These posts use a direct mention (@) or a hashtag (#) combined with the name of the person or institution.
- Post contains a phrase such as “@Min of Health.”

Providing information

- Senders post objective information to a wider audience.
- Post contains phrase such as “There will be a temporary shortage of medicine X.”

Seeking medicine

- Senders try to find the medicine in shortage.
- Post contains a phrase such as “Does anyone happen to have some medicine left?”

Request for advice

- Senders ask for general or specific advice.
- Post contains a phrase such as “What should I do now?”

Explaining or advising someone

- Senders explain or advise an individual.
- Post contains a phrase such as “@Username: I want to refer you to this link for more information.”

Amplifying message

- Senders aim to reach a significant number of people.
- Post contains a phrase such as “This medicine is almost entirely unavailable in the Netherlands.”

All data used in this study were collected according to Twitter’s terms of use and were publicly available at the time of collection and analysis. Research on data from social media, such as Twitter, makes tracing direct quotations back to the user easy. Users expect that results from Twitter are anonymized [33]. Thus, reporting direct quotations is questionable. Modifying post text conflicts with Twitter display requirements prevents altering posts [34]. Thus, we did not quote individual posts and only paraphrased examples of posts in Textbox 1.

Statistical analysis was performed using SPSS (version 28; IBM Corp), and other data analyses were performed using Microsoft Office 365 Excel.

XSL•FO
RenderX

Ethical Considerations

This analysis did not receive approval from an institutional research board. Analysis of large bodies of text written in social media such as the Twitter social network is generally not considered "human subjects research" [35], and therefore, ethical approval was not sought.

Results

Selection of the Posts

The search for Dutch posts using the names of the 341 medicines in shortage retrieved 47,193 posts, 40,104 of which were unique. After selecting posts about medicine shortages by searching for relevant keywords, 2644 posts remained. After assessing the posts, we excluded 1909 posts because they mentioned keywords in a different context or medicine as an alternative to overcome a shortage of another medicine. The remaining posts (n=735) were analyzed.

Mentioned on the Twitter Social Network or Not

Of the 341 medicines in shortage, 102 (29.9%) were mentioned on Twitter during the shortage within the specified timeframe (±30 days). Among these 102 medicines, the median number of posts per medicine was 2 (IQR 1-5; range 1-50).

In comparing the number of medicines mentioned in posts to the number of medicines in shortage, the alimentary tract and metabolism (ATC class A) and the nervous system (ATC class N) were mentioned 1.5 and 1.4 times more often on Twitter, respectively, than expected, based on the observed number of shortages. Anti-infectives for systemic use (ATC class J) were mentioned 2.5 times less than expected (Figure 1; underlying data can be found in Multimedia Appendix 2). Differences were borderline significant (Fisher test Monte Carlo with a 99% CI: 21,390; *P*=.048, 99% CI .042-.054).

Figure 1. The number of medicines in shortage mentioned in posts compared to the number of medicines in shortage (per ATC1 class). At 1.0, the proportion of posts for this ATC class is equal to the proportion of published shortages. ATC: Anatomical Therapeutic Chemical; ATC1: Anatomical Therapeutic Chemical class on the first level.

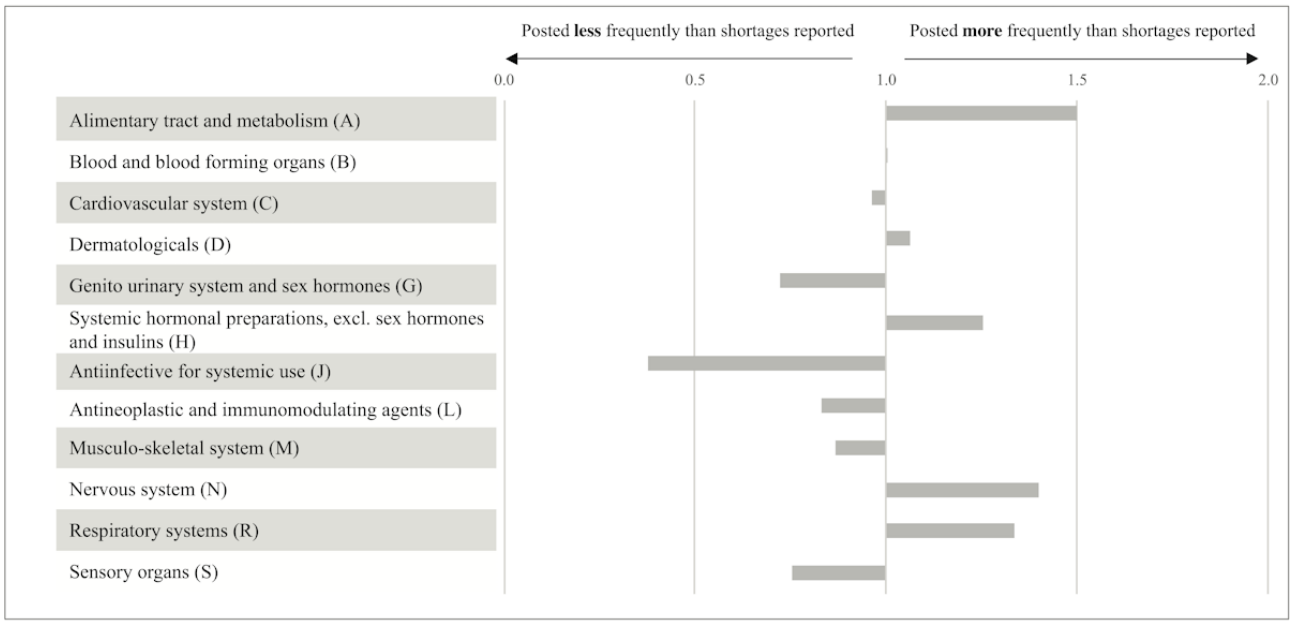

Shortages of oral preparations were mentioned more often on Twitter (n=69, 67.6% vs n=140, 58.6%), and parenteral preparations were mentioned less often on Twitter (n=11, 10.8% vs n=56, 23.4%), but the differences between routes of administration were—also borderline—not statistically significant (Fisher exact test=11,779; *P*=.051; Table 1).

Shortages mentioned on Twitter more often involved therapeutic substitution as a possible alternative product, according to KNMP Farmanco (n=18, 22.8% versus n=21, 13.6% among shortages not found on Twitter). This difference was not statistically significant (Fisher exact test=3441; *P*=.16; Table 1). Shortages mentioned on Twitter affected a higher number of patients than those that were not mentioned (median 31,760, IQR 8735-86,583 versus 4636, IQR 630-22,134; median test: *P*<.001). The median age of patients using medicines in shortage reported on Twitter was 60 (IQR 50-68) years; for medicines not mentioned on Twitter, it was 61 (IQR 48-69) years. The age difference was insignificant (median test: *P*=.68). Finally, shortages on Twitter had a significantly longer duration (149 days) than shortages not mentioned on Twitter (104 days; median test: *P*<.001).

XSL•FO
RenderX

Table 1. Characteristics of medicines in shortage and patients affected (mentioned on Twitter vs not mentioned on Twitter).

	Number of medicine shortages	
	Mentioned on Twitter (n=102)	Not mentioned on Twitter (n=239)
Product characteristics, n (%)		
Medicine group (ATC1[a])		
Alimentary tract and metabolism (A)	13 (12.7)	16 (6.7)
Blood and blood forming organs (B)	3 (2.9)	7 (2.9)
Cardiovascular system (C)	15 (14.7)	37 (15.5)
Dermatologicals (D)	7 (6.9)	15 (6.3)
Genito urinary system and sex hormones (G)	5 (4.9)	18 (7.5)
Systemic hormonal preparations, excl. sex hormones and insulins (H)	3 (2.9)	5 (2.1)
Anti-infective for systemic use (J)	4 (3.9)	31 (13.0)
Antineoplastic and immunomodulating agents (L)	5 (4.9)	15 (6.3)
Musculo-skeletal system (M)	6 (5.9)	17 (7.1)
Nervous system (N)	28 (27.5)	39 (16.3)
Antiparasitic products, insecticides and repellents (P)	0 (0.0)	3 (1.3)
Respiratory system (R)	8 (7.8)	12 (5.0)
Sensory organs (S)	5 (4.9)	17 (7.1)
Various (V)	0 (0.0)	7 (2.9)
Route of administration		
Oral	69 (67.6)	140 (58.6)
Parenteral	11 (10.8)	56 (23.4)
Nasal or inhalation	5 (4.9)	9 (3.8)
Cutaneous	8 (7.8)	14 (5.9)
Rectal	0 (0.0)	1 (0.4)
Ocular	4 (3.9)	15 (6.3)
Other	5 (4.9)	4 (1.7)
Patient impact, n (%)		
Alternative product[b]		
Generic substitution	59 (74.7)	130 (84.4)
Therapeutic substitution	18 (22.8)	21 (13.6)
Import	2 (2.5)	3 (1.9)
Number of patients affected, median (IQR)	31,760 (8735-86,583)	4636 (630-22,134)
Age of patients affected (years), median (IQR)	60 (50-68)	61 (48-69)
Shortage duration[c] (days), median (IQR)	149 (82-248)	104 (58-190)

[a]ATC1: Anatomical Therapeutic Chemical class on the first level.

[b]n=275 because not all advised alternatives could be traced.

[c]For permanent shortages, the end date was set at December 31, 2019, to calculate the duration.

Timing of Posts

For 18 of the 102 shortages mentioned on Twitter, posts were published before or simultaneous to the publication by KNMP Farmanco and 4 were posted more than 14 days earlier (Table 2). Of the 18 shortages, 8 (44%) concerned medicines for the nervous system (N). These shortages early mentioned on Twitter, were mainly posted by patients (n=11, 61%).

XSL•FO
RenderX

Table 2. Shortages mentioned on Twitter before publication by KNMP[a] Farmanco.

Active substance	Medicine group (ATC1[b])	Stakeholder	Days between first publication by Twitter versus KNMP Farmanco, n
Cabergoline	Genitourinary system and sex hormones (G)	Health care professionals	–27
Salmeterol and fluticasone	Respiratory systems (R)	Health care professionals	–26
Acamprosate	Nervous system (N)	Patients	–24
Olanzapine	Nervous system (N)	Health care professionals	–15
Naproxen	Musculo-skeletal system (M)	Patients	–13
Mirtazapine	Nervous system (N)	Patients	–13
Dexamfetamine	Nervous system (N)	Health care professionals	–12
Aciclovir	Anti-infective for systemic use (J)	Unknown	–11
Disulfiram	Nervous system (N)	Health care professionals	–11
Colestyramine	Cardiovascular system (C)	Patients	–7
Paracetamol	Nervous system (N)	Patients	–6
Mycophenolic acid	Antineoplastic and immunomodulating agents (L)	Patients	–5
Citalopram	Nervous system (N)	Patients	–5
Acetazolamide	Sensory organs (S)	Patients	–3
Tamoxifen	Antineoplastic and immunomodulating agents (L)	Patients	–2
Vancomycin	Anti-infective for systemic use (J)	Health care professionals	–1
Pyridostigmine	Nervous system (N)	Patients	0
Famotidine	Alimentary tract and metabolism (A)	Patients	1

[a]KNMP: Royal Dutch Pharmacists Association.

[b]ATC1: Anatomical Therapeutic Chemical class on the first level.

For the 102 shortages with matched posts, the median difference between publication by KNMP Farmanco and the first published post was 37 (IQR 7-81) days, meaning that KNMP Farmanco identified the shortages earlier than Twitter (Figure 2). There was no difference in the median delay of posts published by patients (39, IQR 7-69 days) or health care professionals (36, IQR 8-81 days; median test: P=.840; Figure 3). Posts about medicines for the nervous system (N) were published with a median delay of 30 days. Medicines for the alimentary tract and metabolism (A) and the cardiovascular system (C) were posted later (median of 48 and 67 days, respectively; Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Date of publication of a medicine shortage (n=102) by KNMP Farmanco versus the first time mentioned on Twitter, per stakeholder. KNMP: Royal Dutch Pharmacists Association.

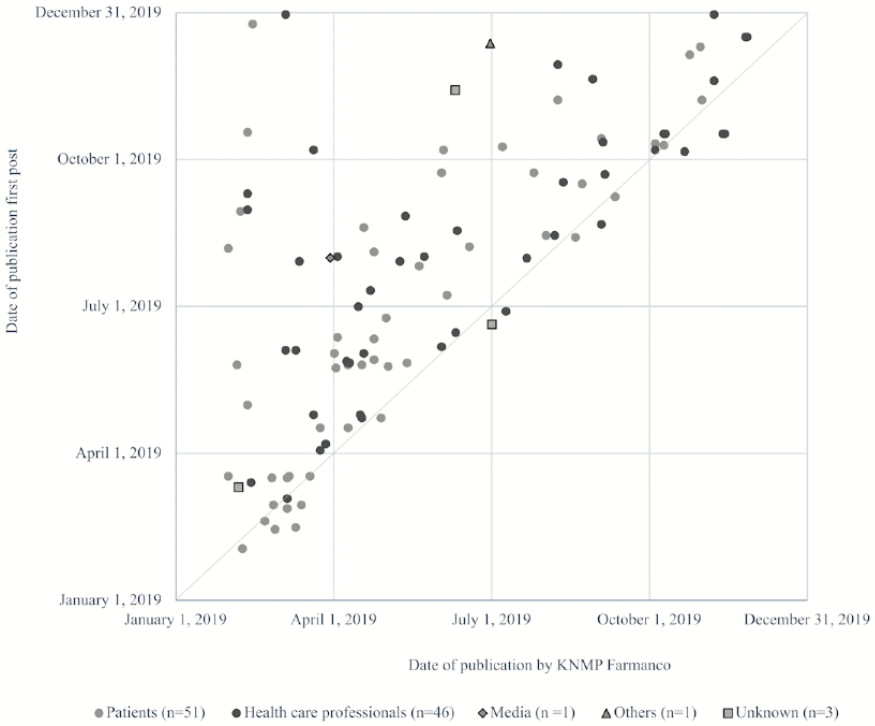

Figure 3. Days between the publication of a medicine shortage by KNMP Farmanco versus the first time mentioned on Twitter (patients and health care professionals). KNMP: Royal Dutch Pharmacists Association.

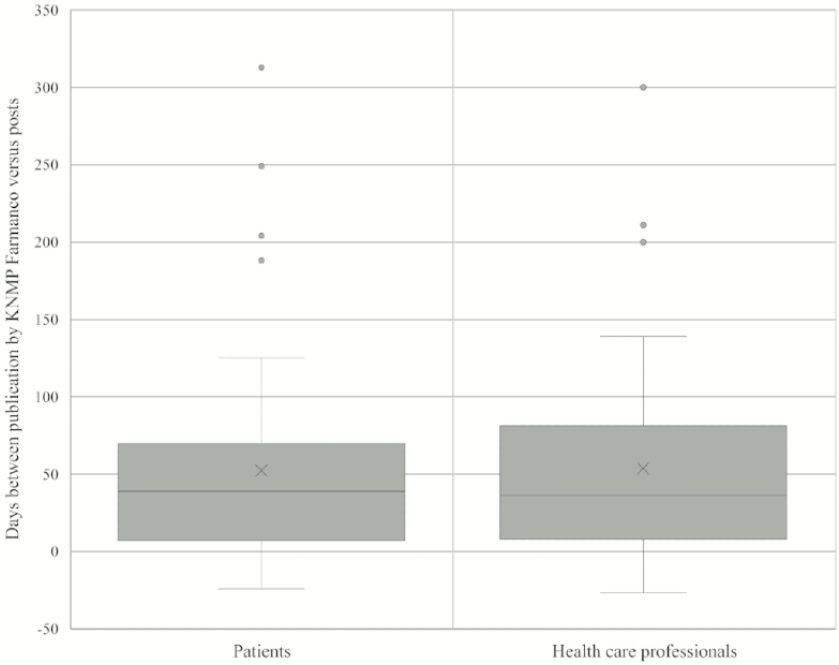

Content Analysis

The first posts on the 102 shortages were published by patients (n=51, 50%), health care professionals (n=46, 45.1%), the media (n=1, 1%), others (n=1, 1%), and unknown users (n=3, 2.9%). We identified 8 categories of posts (Table 3). Sharing personal experience (n=44, 43.1%) was the main category. Emotions, such as anger, frustration, and despair, were regularly shared (n=21, 20.6%), or calls to action were made (n=13, 12.7%). A similar pattern is visible in all 735 posts on medicine shortages (Multimedia Appendix 4).

XSL•FO
RenderX

Table 3. Nature of first posts on medicine shortages per stakeholder (n=102).

Nature	Patients, n	Health care professionals, n	Media, n	Others, n	Unknown, n
Sharing personal experience	18	26	N/A	N/A	N/A
Sharing emotions	12	9	N/A	N/A	N/A
Call to action	6	5	N/A	N/A	2
Providing information	4	3	N/A	1	N/A
Seeking for medicine	3	N/A[a]	N/A	N/A	N/A
Request for advice	5	N/A	N/A	N/A	N/A
Explaining or advising someone	3	3	N/A	N/A	1
Amplifying message	N/A	N/A	1	N/A	N/A

[a]N/A: not applicable.

Discussion

Principal Findings

Our study showed that a routine analysis of Twitter data is not well suited as an early warning system for medicine shortages. Of the 341 medicine shortages, only 18 (5.3%) were mentioned on Twitter prior to or simultaneous to publication by KNMP Farmanco. Only 4 (1.2%) shortages were mentioned on Twitter more than 14 days before publication by KNMP Farmanco. Although we observed differences, we could not conclusively relate the presence or absence on Twitter to a disease area or route of administration of the medicine in shortage and, thus, to certain patient groups.

Two factors may explain why only a limited number of medicine shortages are mentioned on Twitter simultaneously or prior to publication by KNMP Farmanco. First, Twitter does not represent the general population. In 2019, 2.8 million people (16.2% of the population of 17.3 million) used Twitter in the Netherlands, and the usage varied among age groups [36]. Patients affected by the medicine shortages (regardless of being mentioned on Twitter) were a median age of 60 years, an age group that uses Twitter less than other age groups [36]. Second, most terms searched on Twitter, such as the names of medicines, are specific. Spelling errors and references to more general terms, for instance, heart pills and birth control, are not detected, resulting in a limited number of posts. A recent study on early warnings for signals of adverse events and the pharmacovigilance of medicines also detected only a limited number of signals [28], probably for the same reason. Nevertheless, using Twitter as an early warning system for COVID-19 outbreaks was found to be an effective approach [26,37]. This search strategy contained more general terms.

Thus, we expected shortages with a high impact on patients to be mentioned on Twitter more often. We analyzed 2 elements that are core drivers of patient impact—alternative products and the number of patients affected [38]. The availability of an alternative product with the same substance and route of administration (generic substitution) was expected to have less patient impact than that of an alternative product with a different substance (therapeutic substitution). For example, a shortage of aciclovir eye ointment had a high patient impact since no eye drop or ointment was available to treat an eye infection caused by the herpes simplex virus, making therapeutic substitution necessary. Nevertheless, this shortage was not mentioned on Twitter. Although shortages solved by therapeutic substitution were mentioned more often on Twitter than shortages solved by generic substitution, this difference was not statistically significant. In contrast, the number of patients affected, another core driver, showed a significant difference. If more people were affected, it was likelier that the medicine shortage would be mentioned on Twitter.

Even though shared personal experiences comprised the main content of the analyzed posts, over 1 in 5 posts expressed emotions. As expected, patients showed anger, frustration, and despair, and health care professionals did as well. This finding demonstrated that medicine shortages are not solely an administrative problem; they also affect patients and health care professionals personally.

Detecting medicine shortages as early as possible creates time to find solutions to mitigate the shortages' impact on patients. However, identifying medicine shortages and subsequent data sharing may also lead to undesirable reactions such as panic buying and hoarding. Harnessing the dividends of information sharing requires standardized procedures, close collaboration and exchange of information between stakeholders, as well as accurate timing [8].

The major strength of this study is the direct comparison between a public catalog of medicine shortages and posts on medicine shortages on social media platforms such as Twitter. We could establish all the characteristics of the medicine shortages published by KNMP Farmanco and all those of Dutch posts on these shortages. To our knowledge, this study is the first to investigate the suitability of social media as an early warning system for medicine shortages. Detecting medicine shortages as early as possible is expected to create time to find solutions to mitigate the shortages' impact on patients. Thus, social media platforms should be researched as possible sources of information. Moreover, many countries outside of Europe do not yet publish national catalogs on medicine shortages and may be investigating information sources for detecting medicine shortages. Unfortunately, though, Twitter does not serve as an early warning system in the Netherlands.

XSL•FO
RenderX

During analyses, we read the emotional responses and perspectives of patients and health care professionals that have not been described in research. The impact of medicine shortages on patients has been limited to qualitative research on a small number of patients. For research on a larger population of patients and health care professionals that is affected by shortages, Twitter may provide the opportunity for future qualitative research [39]. At the time of research, Twitter was a social media platform with free access for users and nonusers. It is unclear if this access-policy will remain. A change in policy, terms of use, availability of viable alternatives, or public opinion has proven to result in a change for users [40]. Therefore, for future research other social media platforms should also be considered. Like signal detection in current pharmacovigilance systems, Twitter cannot be used in isolation, but it can be used in combination with traditional systems for early signal detection because it can provide a holistic profile [27].

Limitations

Our study has several limitations related to search strategies. First, we used a selection of shortages that KNMP Farmanco published, but we did not use potential shortages that people identified on Twitter that KNMP Farmanco did not publish. Second, we used the active pharmaceutical ingredient's name and the brand name's first word to identify relevant posts. As previously mentioned, our search strategy did not allow us to capture posts about medicines that contained a spelling error in the substance or brand name or posts that only mentioned a more general description of the medicine (for example, "heart pills" or "birth control"). Missing posts with these possible errors or using general terms may mean that specific shortages may have been reported earlier on Twitter than we currently found. However, since we were aiming to assess the role of posts on Twitter for routine analysis, our approach resembles this potential use.

Comparison With Prior Work

We expected lively discussions on Twitter on medicine shortages because there have been increasingly more medicine shortages in the Netherlands in recent years [31,41]. This phenomenon [42] and specific shortages have been discussed in the national news. For instance, a national shortage of contraceptive pills in late 2018 resulted in many news articles [43] and national television reports [44]. This media attention was related to that on Twitter, and we hypothesized that Twitter might serve as a new but still underused data source for identifying medicine shortages. However, the results did not confirm this hypothesis. Only 735 (1.6%) of the 47,193 posts on the medicines in shortage discussed the shortage, while the other 46,458 (98.4%) were concerned with other aspects of the medicines. Two-thirds of the shortages were not even discussed on Twitter.

Conclusions

Twitter cannot be considered an early warning system for medicine shortages, since only 102 of the 341 medicines in shortage were mentioned on Twitter. Of these 102, only 18 were mentioned on Twitter prior to or simultaneous to publication by KNMP Farmanco and only 4 shortages were mentioned on Twitter more than 14 days before publication by KNMP Farmanco. Posts on Twitter tend to echo known information rather than spread new information. However, Twitter has the potential added value of providing a broader view of the personal experiences of patients and health care professionals with medicine shortages. Twitter, or potentially any other social media platform, provides the opportunity for future qualitative research in the increasingly important field of medicine shortages that investigates how a larger population of patients is affected by shortages.

Acknowledgments

The authors wish to thank the Royal Dutch Pharmacists Association for supplying their Farmanco data and Joris Veerbeek (Utrecht Data School) for retrieving the posts and establishing keywords. This work was supported by the Dutch Medicines Evaluation Board and the Royal Dutch Pharmacists Association. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Data Availability

The data set on medicine shortages not publicly available, since data are related to specific products and manufacturers, which is confidential information. The data set on posts is not publicly available due to privacy or ethical restrictions but are available from the corresponding author upon reasonable request. Also, Twitter offers the service Twitter's API (application programming interfaces) [39]. Data on the number of users are not publicly available since it is third-party data. These data belong to the Dutch Foundation for Pharmaceutical Statistics (SFK). The SFK welcomes applications from researchers for access to data. Application forms for data access and further information are available [45].

Disclaimer

The views expressed in this paper are the personal views of the authors and must not be understood or quoted as being made on behalf of or reflecting the position of the Dutch Medicines Evaluation Board or the Royal Dutch Pharmacists Association.

XSL•FO
RenderX

Authors' Contributions

DJP, MLAH, PAGMDS, KN, HGML, and AKMT contributed to the conceptualization and methodology. MLAH, DJP, and AKMT performed the data analysis. DJP, MLAH, and AKMT contributed to writing—original draft preparation. PAGMDS, KN, and HGML contributed to writing—review and editing.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Key words potentially indicating a medicine shortage.
[PDF File (Adobe PDF File), 62 KB - jmir_v26i1e51317_app1.pdf]

Multimedia Appendix 2
Number of medicines in shortage mentioned in posts compared to number of medicines in shortage (per ATC1 class). ATC: Anatomical Therapeutic Chemical.
[PDF File (Adobe PDF File), 67 KB - jmir_v26i1e51317_app2.pdf]

Multimedia Appendix 3
Number of days between publication by KNMP Farmanco versus social network Twitter (per ATC1 class). ATC: Anatomical Therapeutic Chemical; KNMP: Royal Dutch Pharmacists Association.
[PDF File (Adobe PDF File), 56 KB - jmir_v26i1e51317_app3.pdf]

Multimedia Appendix 4
Content analysis of posts on medicine shortages in 2019 (n=735).
[PDF File (Adobe PDF File), 110 KB - jmir_v26i1e51317_app4.pdf]

References

1. Addressing the global shortage of medicines and vaccines. WHA resolution; sixty-ninth world health assembly, 2016. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/wha-69.25 [accessed 2024-04-16]
2. Directive 2001/83/EC of the European parliament and of the council of 6 November 2001 on the community code relating to medicinal products for human use - article 23a and article 81. The European Parliament and the Council of the European Union. 2011. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32001L0083 [accessed 2024-04-16]
3. Addressing the global shortage of, and access to, medicines and vaccines. World Health Organization. 2018. URL: http://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_12-en.pdf [accessed 2024-04-16]
4. Choo EK, Rajkumar SV. Medication shortages during the COVID-19 crisis: what we must do. Mayo Clin Proc 2020;95(6):1112-1115 [FREE Full text] [doi: 10.1016/j.mayocp.2020.04.001] [Medline: 32312491]
5. Cameron EE, Bushell MA. Analysis of drug shortages across two countries during pre-pandemic and pandemic times. Res Social Adm Pharm 2021;17(9):1570-1573. [doi: 10.1016/j.sapharm.2020.12.001] [Medline: 33323334]
6. Miller FA, Young SB, Dobrow M, Shojania KG. Vulnerability of the medical product supply chain: the wake-up call of COVID-19. BMJ Qual Saf 2020;30(4):331-335. [doi: 10.1136/bmjqs-2020-012133] [Medline: 33139342]
7. Guidance on detection and notification of shortages of medicinal products for marketing authorisation holders (MAHs) in the union (EEA). European Medicines Agency, Heads of Medicines Agencies. 2019. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guidance-detection-notification-shortages-medicinal-products-marketing-authorisation-holders-mahs_en.pdf [accessed 2024-04-16]
8. Vann Yaroson E, Quinn G, Breen L. Medicines shortages reporting systems (MSRS): an exploratory review of access and sustainability. Res Social Adm Pharm 2024;20(6):72-83 [FREE Full text] [doi: 10.1016/j.sapharm.2024.02.010] [Medline: 38458895]
9. Good practice guidance for communication to the public on medicines availability issues. European Medicines Agency, Heads of Medicines Agencies. 2019. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guidance-communication-public-medicines-availability-issues_en.pdf [accessed 2024-04-16]
10. European MA. Criteria for classification of critical medicinal products. European Medicines Agency. 2016. URL: https://www.ema.europa.eu/documents/other/criteria-classification-critical-medicinal-products_en.pdf [accessed 2024-04-16]
11. Good practice guidance for patient and healthcare professional organisations on the prevention of shortages of medicines for human use. European Medicines Agency, Heads of Medicines Agencies. URL: https://www.ema.europa.eu/en/documents/

XSL•FO
RenderX

other/good-practice-guidance-patient-healthcare-professional-organisations-prevention-shortages-medicines_en.pdf [accessed 2024-04-16]
12. Meldpunt geneesmiddelentekorten en -defecten 2022 [Report on medicine shortages and defects notification centre 2022]. Dutch Medicines Evaluation Board. 2023. URL: https://www.cbg-meb.nl/documenten/rapporten/2023/03/07/rapportage-meldpunt-geneesmiddelentekorten-en--defecten-2022 [accessed 2024-04-16]
13. Vogler S, Fischer S. How to address medicines shortages: findings from a cross-sectional study of 24 countries. Health Policy 2020;124(12):1287-1296 [FREE Full text] [doi: 10.1016/j.healthpol.2020.09.001] [Medline: 33032846]
14. Miljković N, Batista A, Polidori P, Kohl S, Horák P. Results of EAHP's 2019 medicines shortages survey. Eur J Hosp Pharm 2020;27(4):202-208 [FREE Full text] [doi: 10.1136/ejhpharm-2020-002341] [Medline: 32471816]
15. Postma DJ, Notenboom K, de Smet PAGM, Leufkens HGM, Mantel-Teeuwisse AK. Medicine shortages: impact behind numbers. J Pharm Policy Pract 2023;16(1):44 [FREE Full text] [doi: 10.1186/s40545-023-00548-x] [Medline: 36918981]
16. Atif M, Sehar A, Malik I, Mushtaq I, Ahmad N, Babar ZU. What impact does medicines shortages have on patients? a qualitative study exploring patients' experience and views of healthcare professionals. BMC Health Serv Res 2021;21(1):827 [FREE Full text] [doi: 10.1186/s12913-021-06812-7] [Medline: 34404420]
17. Egberts TC, Smulders M, de Koning FH, Meyboom RH, Leufkens HG. Can adverse drug reactions be detected earlier? a comparison of reports by patients and professionals. BMJ 1996;313(7056):530-531 [FREE Full text] [doi: 10.1136/bmj.313.7056.530] [Medline: 8789980]
18. Bian J, Topaloglu U, Yu F. Towards large-scale Twitter mining for drug-related adverse events. SHB12 (2012) 2012;2012:25-32 [FREE Full text] [doi: 10.1145/2389707.2389713] [Medline: 28967001]
19. Zhou Z, Hultgren KE. Complementing the US food and drug administration adverse event reporting system with adverse drug reaction reporting from social media: comparative analysis. JMIR Public Health Surveill 2020;6(3):e19266 [FREE Full text] [doi: 10.2196/19266] [Medline: 32996889]
20. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
21. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Saf 2014;37(10):777-790 [FREE Full text] [doi: 10.1007/s40264-014-0218-z] [Medline: 25151493]
22. Edo-Osagie O, de la Iglesia B, Lake I, Edeghere O. A scoping review of the use of Twitter for public health research. Comput Biol Med 2020;122:103770 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103770] [Medline: 32502758]
23. Zhao Y, He X, Feng Z, Bost S, Prosperi M, Wu Y, et al. Biases in using social media data for public health surveillance: a scoping review. Int J Med Inform 2022;164:104804. [doi: 10.1016/j.ijmedinf.2022.104804] [Medline: 35644051]
24. Dol J, Tutelman PR, Chambers CT, Barwick M, Drake EK, Parker JA, et al. Health researchers' use of social media: scoping review. J Med Internet Res 2019;21(11):e13687 [FREE Full text] [doi: 10.2196/13687] [Medline: 31719028]
25. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
26. Lopreite M, Panzarasa P, Puliga M, Riccaboni M. Early warnings of COVID-19 outbreaks across Europe from social media. Sci Rep 2021;11(1):2147 [FREE Full text] [doi: 10.1038/s41598-021-81333-1] [Medline: 33495534]
27. Pathak R, Catalan-Matamoros D. Can Twitter posts serve as early indicators for potential safety signals? a retrospective analysis. Int J Risk Saf Med 2023;34(1):41-61. [doi: 10.3233/JRS-210024] [Medline: 35491804]
28. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, van Le H, et al. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf 2017;40(4):317-331 [FREE Full text] [doi: 10.1007/s40264-016-0491-0] [Medline: 28044249]
29. Social media in Nederland 2020: TikTok is domein van kinderen, uittocht jongeren op Facebook. Social media in the Netherlands 2020: TikTok is domain of childeren, youth departs Facebook. Marketingfacts. 2020. URL: https://www.marketingfacts.nl/berichten/social-media-in-nederland-2020 [accessed 2024-04-16]
30. The website for Dutch medicine shortages. KNMP Farmanco. URL: https://www.farmanco.knmp.nl [accessed 2024-04-16]
31. About KNMP Farmanco. KNMP Farmanco. URL: https://farmanco.knmp.nl/about-knmp-farmanco [accessed 2024-04-16]
32. World Health Organization. Structure and principles. Norwegian Institute of Public Health. URL: https://www.whocc.no/atc/structure_and_principles/ [accessed 2024-04-16]
33. Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation. Sociology 2017;51(6):1149-1168 [FREE Full text] [doi: 10.1177/0038038517708140] [Medline: 29276313]
34. Display requirements: posts. X Corp. URL: https://developer.twitter.com/en/developer-terms/display-requirements [accessed 2024-04-16]
35. Do I need ethics approval for social media research? JMIR Publications Knowledge Base and Help Center. URL: https://support.jmir.org/hc/en-us/articles/115001620728-Do-I-need-ethics-approval-for-social-media-research [accessed 2024-04-16]

XSL•FO
RenderX

36. Social media in Nederland in 2020: Facebook stijgt licht, Instagram stijgt sterk. Social media in the Netherlands in 2020: Facebook small increase, Instagram big increase. The Big Story. 2020. URL: https://thebigstory.nl/social-media-in-nederland-in-2020-facebook-stijgt-licht-instagram-stijgt-sterk/ [accessed 2024-04-16]
37. Zhang Y, Chen K, Weng Y, Chen Z, Zhang J, Hubbard R. An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. Expert Syst Appl 2022;198:116882 [FREE Full text] [doi: 10.1016/j.eswa.2022.116882] [Medline: 35308584]
38. Postma DJ, de Smet PAGM, Notenboom K, Leufkens HGM, Mantel-Teeuwisse AK. Impact of medicine shortages on patients - a framework and application in the Netherlands. BMC Health Serv Res 2022;22(1):1366 [FREE Full text] [doi: 10.1186/s12913-022-08765-x] [Medline: 36397073]
39. About X's APIs. X Corp. URL: https://help.twitter.com/en/rules-and-policies/x-api [accessed 2024-04-16]
40. How Elon Musk is changing the Twitter experience. The New York Times. 2023. URL: https://www.nytimes.com/2023/04/07/technology/elon-musk-twitter-changes.html [accessed 2024-04-16]
41. Aantal meldingen mogelijke leveringsproblemen medicijnen verder gestegen. Number of reports of potential supply problems with medicines further increased. Dutch Medicines Evaluation Board. 2019. URL: https://www.meldpuntgeneesmiddelentekortendefecten.nl/actueel/nieuws/2019/03/13/aantal-meldingen-leveringsproblemen-medicijnen-verder-gestegen [accessed 2024-04-16]
42. van der Geest M. Medicijntekorten blijven maar toenemen. Hoe komt dat? Medicine shortages continue to increase. Why is that? deVolkskrant. 2019. URL: https://www.volkskrant.nl/wetenschap/medicijntekorten-blijven-maar-toenemen-hoe-komt-dat~b0f97a57/ [accessed 2024-04-16]
43. Nog steeds tekort aan meest gebruikte anticonceptiepil. Still short of most commonly used contraceptive pill. Nederlandse Omroep Stichting (Dutch Broadcasting Foundation). 2018. URL: https://nos.nl/artikel/2260098-nog-steeds-tekort-aan-meest-gebruikte-anticonceptiepil [accessed 2024-04-16]
44. Tekort aan anticonceptiepillen. Shortage of birth control pills. Avrotros. 2018. URL: https://radar.avrotros.nl/nieuws/item/tekort-aan-anticonceptiepillen/ [accessed 2024-04-16]
45. Duthch Foundation for pharmaceutical statistics. Stichting Farmaceutische Kengetallen (SFK). URL: https://www.sfk.nl/english/foundation-for-pharmaceutical-statistics [accessed 2024-04-16]

Abbreviations

ATC: Anatomical Therapeutic Chemical
KNMP: Royal Dutch Pharmacists Association
MAH: marketing authorization holder
SFK: Dutch Foundation for Pharmaceutical Statistics

Edited by S Ma, T Leung; submitted 22.08.23; peer-reviewed by D Carvalho, AR Ashraf, R Vida; comments to author 02.04.24; revised version received 10.05.24; accepted 21.06.24; published 06.08.24.

Please cite as:
Postma DJ, Heijkoop MLA, De Smet PAGM, Notenboom K, Leufkens HGM, Mantel-Teeuwisse AK
Identifying Medicine Shortages With the Twitter Social Network: Retrospective Observational Study
J Med Internet Res 2024;26:e51317
URL: https://www.jmir.org/2024/1/e51317
doi:10.2196/51317
PMID:

©Doerine J Postma, Magali L A Heijkoop, Peter A G M De Smet, Kim Notenboom, Hubert G M Leufkens, Aukje K Mantel-Teeuwisse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Peer Support for Chronic Pain in Online Health Communities: Quantitative Study on the Dynamics of Social Interactions in a Chronic Pain Forum

Aaron Necaise[1], MSc; Mary Jean Amon[2], PhD

[1]School of Modeling, Simulation, and Training, University of Central Florida, Orlando, FL, United States

[2]Department of Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN, United States

Corresponding Author:
Aaron Necaise, MSc
School of Modeling, Simulation, and Training
University of Central Florida
Partnership II
3100 Technology Parkway
Orlando, FL, 32816
United States
Phone: 1 321 300 4582
Email: aaron.necaise@ucf.edu

Abstract

Background: Peer support for chronic pain is increasingly taking place on social media via social networking communities. Several theories on the development and maintenance of chronic pain highlight how rumination, catastrophizing, and negative social interactions can contribute to poor health outcomes. However, little is known regarding the role web-based health discussions play in the development of negative versus positive health attitudes relevant to chronic pain.

Objective: This study aims to investigate how participation in online peer-to-peer support communities influenced pain expressions by examining how the sentiment of user language evolved in response to peer interactions.

Methods: We collected the comment histories of 199 randomly sampled Reddit (Reddit, Inc) users who were active in a popular peer-to-peer chronic pain support community over 10 years. A total of 2 separate natural language processing methods were compared to calculate the sentiment of user comments on the forum (N=73,876). We then modeled the trajectories of users' language sentiment using mixed-effects growth curve modeling and measured the degree to which users affectively synchronized with their peers using bivariate wavelet analysis.

Results: In comparison to a shuffled baseline, we found evidence that users entrained their language sentiment to match the language of community members they interacted with (t_{198}=4.02; P<.001; Cohen d=0.40). This synchrony was most apparent in low-frequency sentiment changes unfolding over hundreds of interactions as opposed to reactionary changes occurring from comment to comment ($F_{2,198}$=17.70; P<.001). We also observed a significant trend in sentiment across all users (β=–.02; P=.003), with users increasingly using more negative language as they continued to interact with the community. Notably, there was a significant interaction between affective synchrony and community tenure (β=.02; P=.02), such that greater affective synchrony was associated with negative sentiment trajectories among short-term users and positive sentiment trajectories among long-term users.

Conclusions: Our results are consistent with the social communication model of pain, which describes how social interactions can influence the expression of pain symptoms. The difference in long-term versus short-term affective synchrony observed between community members suggests a process of emotional coregulation and social learning. Participating in health discussions on Reddit appears to be associated with both negative *and* positive changes in sentiment depending on *how* individual users interacted with their peers. Thus, in addition to characterizing the sentiment dynamics existing within online chronic pain communities, our work provides insight into the potential benefits and drawbacks of relying on support communities organized on social media platforms.

(J Med Internet Res 2024;26:e45858) doi:10.2196/45858

XSL•FO
RenderX

KEYWORDS

social media; chronic pain; peer support; sentiment analysis; wavelet analysis; nonlinear dynamics; growth curve modeling; online health communities; affective synchrony

Introduction

Background

The National Institutes of Health describes chronic pain as pain persisting for >3 to 6 months [1,2], which is substantially longer than a "typical" acute response to injury [3]. Chronic pain can be attributed to a wide variety of underlying medical conditions [3], and it is one of the most common health complaints in the United States [4,5]. A recent report from the National Center for Health Statistics found that the prevalence of chronic pain among adults in the United States was approximately 20%, while an additional 7.4% of survey respondents reported pain severe enough to impact daily functioning [4,5]. Moreover, prevalence rates are similarly high in high-income [6-9] and low- and middle-income countries [9,10] throughout the world, contributing to a global public health problem [11,12].

Due to the complexities of pain processing, successful management often necessitates a multifaceted approach personalized to meet the needs of the individual [13]. Chronic pain is considered to be a complex process, whereby psychosocial factors such as emotions, expectations, and social relationships interact with an individual's physical health to contribute to the maintenance of symptoms [3,14,15]. Thus, it is encouraged to engage in health self-management strategies that contribute to mental and physical well-being [16-18]. One particularly impactful self-management strategy is participation in peer support, which occurs when individuals facing similar health challenges exchange advice, emotional validation, and educational resources [19-22]. Those experiencing chronic pain frequently report emotional distress and feelings of social isolation, and peer support can help alleviate these stressors [23].

Peer support, and pain self-management [23] more broadly, is increasingly taking place on social media platforms such as Reddit (Reddit, Inc), Facebook (Meta Platforms, Inc), and Instagram (Meta Platforms, Inc) [24]. Social media users with chronic health conditions frequently self-organize into large peer-to-peer communities that function as social and informational hubs [25]. For example, there are many peer-driven support communities on Reddit in which tens of thousands of members share private health information and respond to medical inquiries. Support interactions on social media platforms are unique in that they consist of naturally occurring discussions driven by users themselves, contrasting formally organized support groups [26] and structured web-based interventions [27]. Although users tend to have positive experiences with online support [28], interactions on the internet are not universally positive. As demonstrated by the COVID-19 pandemic, social media platforms have the potential to spread misinformation and hostility toward medical experts [29,30]. Thus, although writing about pain [31] and having supportive relationships [32] can benefit mental health, it is unclear how naturally occurring web-based health interactions influence chronic pain [24,33,34].

It is reasonable to assume that the impact of online peer support not only depends on the amount of support an individual receives but also on the affective qualities of their social interactions [14,35]. Negative affective thoughts and pain catastrophizing are theorized to contribute to the maintenance of chronic pain (eg, [3,14,36-38]), and social interactions can drive users to adopt or reject those beliefs [14]. Furthermore, the manner in which people communicate about their experiences plays an important role in shaping their emotions and expectations for the future [31,39]. It follows that health communities perpetuating overly pessimistic attitudes or hostility toward health care providers can be harmful to well-being, just as those providing emotional validation can be beneficial [22]. Thus, examining the affective qualities of web-based interactions using sentiment analysis may serve as an entry point to understanding how health is influenced by participation in online health communities (OHCs). Moreover, the analysis of online health discussions can provide broader insight into how social media users experiencing chronic pain communicate about their pain.

Related Work

Psychosocial Determinants of Chronic Pain

The biopsychosocial model is the prevailing approach to understanding chronic pain [14,15,40]. From this perspective, chronic pain results from the dynamic interactions among biological, psychological, and social processes [40,41]. Edwards et al [14] provided an overview of the most widely researched psychosocial factors associated with chronic pain, centering around negative affect and pain catastrophizing (ie, negative rumination regarding pain symptoms). Negative affective thoughts are associated with an increased risk of developing chronic pain [42-44] and are believed to contribute to a variety of negative health outcomes [3,14]. Several longitudinal studies have reported that psychological distress, pain catastrophizing, and self-efficacy were moderators between pain and disability [37,38,45], such that individuals with greater negative affect were more likely to report impairment at follow-up. Overall, there is substantial evidence that expectations and emotions toward pain can impact health outcomes.

Although the factors outlined earlier are primarily psychological, they are also inextricably linked to social interactions. Maintaining strong support systems provides benefits to both physical and mental well-being [46-49], such as being associated with reduced psychological distress [50-52]. Communication with physicians, friends, or family can drive specific health beliefs that are relevant to pain outcomes. For example, individuals who experience higher spousal autonomy support report increased need satisfaction and well-being [53], while interventions teaching couples how to communicate supportively reduce pain catastrophizing [54,55]. With respect to the specific

XSL•FO
RenderX

mechanisms, social interactions can perpetuate realistic treatment expectations, provide supportive (vs adversarial or solicitous [14]) feedback in response to pain expressions, and encourage health self-management [24]. Thus, social interactions are potentially facilitative or harmful depending on their qualities.

Building on the biopsychosocial approach to understanding pain, several theories have emerged that present a more explicit account of the social determinants of pain, for example, the study by Craig [56], the study by Hadjistavropoulos et al [57], and the study by Sullivan [58]. The social communication model of pain (SCMP) is an example of conceptualizing the specific biopsychosocial interactions underlying chronic pain [56,59]. The SCMP distinguishes between the effects of intrapersonal (eg, genetic predispositions and family history) and interpersonal (eg, social interactions and medical treatments) processes, emphasizing the contributions of "others" on health. According to this model, a pain-relevant social interaction occurs when a person in pain expresses their symptoms verbally or nonverbally, and an observer decodes and acts on that expression. The action taken by the observer can range from helpful to exacerbating, and it directly influences how the person in pain interprets their pain in the future. Notably, this model emphasizes that an observer's decoding of a pain expression is biased by their own background, and the subsequent actions they take depend on their relationship to the person in pain (ie, they might be a caregiver, friend, or stranger). This framing helps contextualize online support interactions where communication tends to be anonymous, expressions of pain are text based, and observers have little obligation to respond ethically compared to "real-life" associates [60-62]. The description of social influence provided by the SCMP can also be extended to the context of social media. That is to say that a social media user's interpretation of their pain may be iteratively updated by the feedback they receive on the web leading to changes in how they perceive and express their pain in the future. Thus, examining online health discussions is an avenue for investigating the social determinants of pain described by the SCMP.

Chronic Pain Support Interactions on Social Media

Social media is a term used to describe a collection of digital technologies that allow users to maintain a web-based presence, communicate and network with others, and share user-generated content [63,64]. The potential utility of social media platforms for health care purposes was recognized long before the rise of modern social networking sites, with initial research often focusing on peer-to-peer support occurring in chat rooms and messaging boards throughout the internet [65-67]. A review published in 2004 by Eysenbach et al [68] reported a lack of evidence to suggest that these early-forming OHCs had positive impacts on well-being, but the authors acknowledged that they were largely unmoderated and in the infancy of their development. With expanded access to the internet and the introduction of popular social networking sites (ie, Facebook and YouTube [Google LLC]), the growth of OHCs accelerated [69]. There is now a huge selection of communities available to social media users depending on their individual needs and preferences, varying with the amount of professional input provided [70], type of communication, and degree of anonymity [26].

Qualitative research examining pain discourse on social media platforms frequently highlights interactions containing positive emotional messages and information sharing [28,71-73]. In an analysis of 44 blogs collected across several websites, commenters overwhelmingly replied to pain blogs with messages of consolation and encouragement [73]. Furthermore, there was a "virtual online support sequence" underlying these interactions, in which commenters replied to blog posts with personal anecdotes and used their common experiences as an opportunity for emotional validation [73]. Similarly, a clinical study using semistructured interviews to gather perceptions about the use of social media platforms for pain self-management found that participants appreciated the ability to connect with others, share their personal experiences, and learn directly from their peers [28]. The perceived benefits of online peer support are not specific to chronic pain, as similar themes of emotional support, connectivity, and experiential knowledge sharing have been identified in online communities organized for a variety of chronic health conditions [19,25,74-76].

Quantitative studies investigating the effects of online support on chronic pain have been comparatively sparse and, in many cases, focus on the impact of structured intervention programs (eg, [77-79]) as opposed to the interactions occurring naturally among social media users. There is some evidence that the positive psychosocial benefits provided by traditional support groups can also be provided by casual web-based interactions, particularly when it comes to information sharing [80-82]. For example, individuals who were assigned to follow a Twitter (subsequently rebranded as X; X Corp) profile posting information about self-management strategies reported small improvements in pain, emotional distress, and quality of life after 6 months of virtual interactions [81]. When considering chronic illnesses more broadly, online peer support has been found to reduce feelings of isolation in adolescents [83], and the mere act of self-disclosing about life stressors (ie, not just chronic health) can reduce emotional distress [84]. Although peer support on social media platforms can play a beneficial role, these effects likely depend on the specific social dynamics in the community, such as the presence of peer role models to help guide conversations [85] or the type of messages that are circulated. Web-based health resources will be even more popular among future generations [24], highlighting the need for continued research on web-based health self-management. Specifically, social media research can lead to an improved understanding of how individuals experiencing chronic pain develop their health-related beliefs.

Objectives

This study investigated peer support interactions in OHCs by focusing on the sentiment of individual support interactions. Applying sentiment analysis to the comment histories of users in a popular chronic pain forum on Reddit, we measured the degree to which users synchronized the sentiment of their comments to match their peers, and we modeled the trajectories of their sentiment over the course of their community

XSL•FO
RenderX

participation. This work addresses 3 questions regarding the use of online peer support for chronic pain: Do users who engage in online peer support synchronize their pain expressions to match the language of other community members? How does a user's sentiment progress over time in response to online support interactions? And finally, how do specific interaction dynamics, such as affective synchrony with other users, relate to changes in sentiment?

Methods

Study Design

We analyzed comments posted to a popular chronic pain support community on Reddit. Reddit is a platform where users can create personalized forums dedicated to discussing specific topics [86] and within each user-created forum (ie, "subreddit"), users define their own rules regarding content and membership. The subreddit analyzed in this study describes itself as a forum for users to discuss their conditions with their peers and share advice; however, we have opted to withhold the exact name of the community out of concern for user privacy. The community guidelines provided by the moderators discourage direct medical advice and suggest consultation with professionals before participation. However, it is unclear how strictly these guidelines are enforced. In terms of content, discussion threads contain a mixture of advice-seeking, informational resources, and personal anecdotes. A recent paper using latent Dirichlet allocation to analyze chronic pain subreddits on Reddit similar to the one used in this study found that users most frequently mentioned phrases related to lower back pain in their posts in addition to words such as "doctors," "help," and "work" [87], providing evidence that discussions within these communities are highly focused on the topic of chronic pain.

We collected the post histories of 200 randomly selected Reddit users active on the subreddit and calculated the sentiment of their comments using a dictionary-based approach [88]. Sentiment analysis is a common natural language processing method used to classify and describe the emotional expressiveness of text by analyzing the valence, intensity, and structural features of language [89,90]. Next, we used bivariate wavelet analysis to estimate the degree of sentiment synchrony between users during interactions, which is a technique popularized for its applications on complex systems [91]. Unlike a standard linear approach (eg, cross-correlation), wavelet analysis is robust against nonstationarity and describes multiple types of synchronies, including matched intensity, comovement, and leader-follower dynamics [91,92]. Moreover, wavelet analysis can effectively separate changes in sentiment that are occurring reactively (eg, comment to comment), from those occurring globally over hundreds of interactions. Finally, we examined whether discourse on the subreddit was associated with negative or positive changes in the health attitudes of users by using growth curve analysis to model trajectories in comment sentiment.

Ethical Considerations

This study was reviewed by the institutional review board at the University of Central Florida (00001430). Observational public data from Reddit were collected, and there were no direct interactions with human participants. Reddit is a social media platform where users typically provide pseudonyms in place of their real names, which are not reported, and we did not collect personally identifiable information. We randomly sampled users who fit our eligibility criteria, meaning a particular profile's inclusion is not discernable. In addition, we have opted to withhold the name of the subreddit that we collected data from to further obfuscate the identity of its members. Given the large number of comments we collected, it was possible that some of the text contained identifiable information indirectly in the form of self-disclosures. To protect user anonymity, each user in our sample received a randomized identifier in place of their Reddit username, and the text content of their comments was aggregated using the procedures detailed in the sections that follow such that it would not be possible to link the stored data to any individual Reddit user or real-life individual.

Data Sampling

We collected publicly available Reddit data using the Pushshift database [93] and Reddit's application programming interface [94]. On Reddit, the term *submission* refers to user-created discussion threads posted within each community, while *comments* refer to the text replies within each discussion thread. To be included in our final sample, users had to be moderately active members of the support subreddit, with at least 100 total comments and submissions after excluding self-deleted posts and those deleted through moderation. This minimum activity threshold was applied to ensure there was an adequate number of data points for wavelet analysis at lower frequencies [92]. In addition, we excluded users serving as moderators and Reddit profiles posting automatic replies (eg, apparent bot accounts). To identify Reddit profiles meeting this criterion, we downloaded every publicly available comment and submission made to the subreddit in the last 10 years. We then calculated each user's total number of posts in the community and randomly selected 200 users with more than the minimum required activity. The sample size was determined using an a priori power analysis for a growth curve model with a small effect size, f^2=0.10, and α=.05, estimated based on similar previous research examining sentiment in online discussion forums [95-97]. Finally, we collected data from the public interactions these users had on Reddit. We define an interaction as the pairing between a user's personal comment and the post they were replying to when making that comment. For example, an interaction could involve a user posting a top-level comment in response to a discussion thread. An interaction could also involve posting a reply to someone else's comment inside of a discussion thread. Thus, for each user, we collected the following: (1) the text of all their comments, (2) the corresponding text of posts they interacted with, (3) comment scores (upvotes minus downvotes), and (4) time of posting.

Data Processing

Each user's comment history was formatted as a time series sorted in chronological order, and each point in those time series represented 1 interaction with the community. Activity occurring outside of the subreddit of interest was excluded from the analysis. In addition, the Reddit application programming interface returns an error message instead of the original text

XSL•FO
RenderX

when comments are removed by moderators or users themselves, and these error messages were excluded. A small portion of comments contained single-word phrases, such as a web address or emoji. Sentiment calculations for these items would have been unreliable, and we opted to remove comments containing <3 words. A total of 73,876 of comment interactions were collected from the pain support subreddit. Of those comments, 5670 (7.68%) were excluded based on low word count or missing data. Shortly before analysis, we also discovered 1 user acting as a moderator, and this individual's data were removed due to their unique role in the community. Thus, our final sample size included 68,206 (92.32%) of the 73,876 peer-to-peer interactions from 199 users and across 29,360 unique discussion threads.

After cleaning the data, we calculated sentiment scores using a lexicon-based approach [88], such that there were separate scores for both sides of each social interaction. To improve the reliability of these estimates, we compared 2 popular techniques in the psychology and social media literature. The Linguistic Inquiry and Word Count (LIWC) [98] dictionary is a natural language processing tool used to extract psychological language attributes and has been applied in a variety of digital health and social media contexts (eg, [99-102]). The LIWC dictionary calculates the percentage of text pertaining to specific topics (eg, the percentage of health-related words) and provides summary variables measuring higher-level cognitive processes. For our purposes, we used the LIWC dictionary to calculate sentiment, number of words, and percentage of health-related words in each comment. Sentiment scores from the LIWC dictionary ranged from –100 to 100, with higher scores indicative of more positive language, lower scores indicative of more negative language, and scores around 0 indicative of neutral language [98,103]. For comparison, we also calculated sentiment using the Valence Aware Dictionary for Sentiment Reasoning (VADER) [104]. Sentiment scores from VADER have been validated against manual human reviewers and were designed for the analysis of brief web-based internet interactions [104]. Like the LIWC dictionary, compound sentiment scores from VADER ranged from –1 to +1, with higher scores indicative of more positive language, lower scores indicative of more negative language, and scores around 0 indicative of neutral language [105]. We found both methods were highly consistent (α=.93) and strongly correlated (r=0.70; P<.001); consequently, we created a compound score by standardizing and averaging the results of both. Generating compound scores through aggregation has been shown to improve reliability [106], and, in our case, this approach minimized the influence of text passages where the 2 methods diverged. As a result, each user had one-time series representing their personal comment sentiment and one-time series for the sentiment of comments they interacted with when making their personal comments.

Bivariate Wavelet Analysis

Bivariate wavelet analysis was used to quantify the sentiment synchrony between users and their peers. We used the cross-wavelet transform (XWT) and wavelet coherence (WTC) to gain information about matching sentiment intensity and phase-locked (ie, correlated) changes in sentiment. Compared to cross-correlation, wavelet analysis describes changes in sentiment at multiple frequencies (eg, short- vs long-term changes in sentiment). This allowed us to isolate fluctuations in sentiment occurring comment to comment (ie, high-frequency changes) from those occurring over dozens or hundreds of interactions (ie, lower-frequency changes). In other words, this analysis provided insight into short-term reactionary changes in sentiment versus long-term global changes. We briefly discuss XWT and WTC subsequently; however, readers should refer to the work of Torrence and Compo [107] for a comprehensive guide to wavelet analysis or Issartel et al [92] for applications on behavioral synchrony. A discussion of the method and theoretical implications in the context of social media is provided by Necaise et al [95].

Wavelet calculations were completed using the *biwavelet* package in R (R Foundation for Statistical Computing) [108] with a Morlet wavelet function (ω_0=6). The Morlet wavelet was selected because it provides good resolution compared to other functions [107]. Furthermore, cross-wavelet calculations can be sensitive to large spikes in the data, so the data were transformed using a cumulative distribution function, so the values represented percentiles. This transformation has been suggested in previous literature [91]. The XWT was applied using the method outlined by Grinsted et al [91] and identified points in time (x-axis of Figure 1) and frequency ("period" on the y-axis of Figure 1) where users had fluctuations in sentiment that were similar in intensity as their conversation partners. Significance is determined in comparison to AR(1) background processes [91,92], and the significant points (circled in black in panel A of Figure 1) are referred to as regions of *high common power* [91]. The XWT also provided information about the *relative phase (RP) relationship* between users and their conversation partners within regions of high common power (depicted by the orientation of arrows in Figure 1). The RP angle was extracted from the XWT and described how user sentiment changed relative to their peers [91], providing information about leader-follower dynamics.

In addition to measuring common power with the XWT, we also calculated user *coherence* with their peers using WTC. The formula for WTC is similar to Pearson correlation except localized in frequency and time [91]. Therefore, it is helpful to think about coherence as an indicator of correlation or "comovement" between 2 signals. We used a standard smoothing factor of 0.6 for Morlet wavelets in the WTC calculation and tested for significant coherence against simulated data via Monte Carlo methods with 2000 random initializations [91]. Regions of significant coherence are circled in black in panel B of Figure 1.

XSL•FO
RenderX

Figure 1. Example cross-wavelet spectrum plots visualizing (A) high common power (cross-wavelet transform [XWT]) and (B) coherence (wavelet coherence [WTC]) generated from 1 user's sentiment compared to the sentiment of posts they replied to. The XWT identified significant regions across 2 signals with similarly intense sentiment fluctuations. The WTC plot identified regions across the 2 signals having high coherence. Finally, the relative phase (RP) angle is visualized as arrows within significant regions of the cross-wavelet plots. Arrows pointing right represent in-phase fluctuations, while arrows to the left represent antiphase fluctuations. The common power and coherence were calculated as the percentage of their respective plots having significant high common power or coherence (regions circled in black).

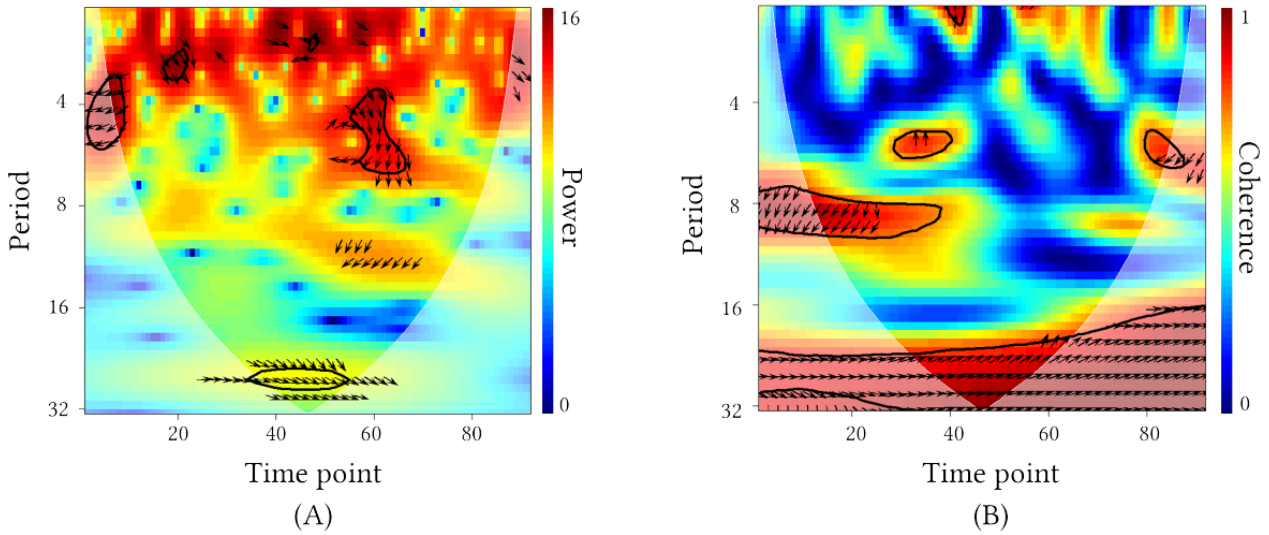

Statistical Approach

Measuring Sentiment Synchrony

Several metrics can be extracted from cross-wavelet power plots relevant to describing synchrony [92] (Table 1). We were interested in global estimates of synchrony and therefore calculated the percentage of each user's cross-wavelet plot exhibiting significant common power and coherence (Figure 1). Likewise, we calculated the mean circular angle [109] of the RP relationships in regions with significantly high common power. These calculations excluded points outside of the cone of influence where estimates can be unreliable, as depicted by the lightened regions in Figure 1. Each of these metrics was used to describe synchrony from a distinct perspective. Social media users with higher common power were more synchronous with their peers in terms of matched intensity, and higher coherence was an indicator of correlation in time-frequency space. RP angle describes *how* the sentiment of a user changed directionally with respect to the community. An RP angle of 0° would indicate the sentiments of a user and their peers were *in-phase* (ie, by fluctuating up and down at the same time), while an RP angle of 180° would indicate an *antiphase* relationship (ie, by alternating opposite of one another).

Table 1. Description of synchrony metrics derived from wavelet analysis, the label provided to them in the text, and a description of which aspect of synchrony each outcome measures.

Synchrony measures	Calculation	What it measures
Common power	Percentage of significant points in XWT[a] plot	Degree of matched sentiment intensity in time-frequency domain
Coherence	Percentage of significant points in WTC[b] plot	Degree of correlated behavior in time-frequency domain (eg, phase-locked behavior or comovement)
RP[c] angle	Mean circular angle of RP in regions of significant common power	Leader-follower dynamics (eg, fluctuating in perfect synchrony versus in an alternating pattern)

[a]XWT: cross-wavelet transform.

[b]WTC: wavelet coherence.

[c]RP: relative phase.

Testing for Significant Synchrony Against Shuffled Baseline

To determine whether our estimates of common power and coherence constituted a significant amount of synchrony beyond what could be explained by random variance, we repeated the wavelet estimations described above on a shuffled version of each user's data. We then used paired 2-tailed *t* tests to compare common power and coherence from the original data to the shuffled baseline. It should be noted that our calculations for common power and coherence were based on the percentage of significant points across all frequencies in the power spectrum, leading to deflated percentages that may be difficult to interpret. Thus, we provide effect size estimates to better depict the magnitude of differences between shuffled and real data [110].

To provide evidence of the construct validity of our wavelet-based outcome measures, we also compared common power and coherence to a more traditional synchrony estimate.

XSL•FO
RenderX

Pearson correlation coefficients have been used in previous literature as global estimates of synchrony (eg, [111]), so we calculated a Pearson coefficient for each user by correlating the sentiment of their personal comments to the sentiment of their peers. We then conducted a correlation analysis between Pearson coefficients and the bivariate wavelet metrics in both the real and shuffled data.

Comparing Degree of Synchrony Across Different Frequencies of Change

As described previously, although there is some overlap between cross-correlation and bivariate wavelet analysis in how they describe synchrony, wavelet analysis has the added benefit of decomposing signals into individual frequency components. This allowed us to analyze synchrony at specific frequency bands (ie, slow vs fasting moving sentiment changes). In addition to extracting common power and coherence across each user's entire cross-wavelet spectrum plot, we also calculated common power and coherence within 3 distinct frequency bands: low (fluctuations in sentiment unfolding between 20 and 32 interactions), medium (fluctuations unfolding between 10-20 interactions), and high (fluctuations unfolding in <10 interactions). We then used repeated measures ANOVA to compare the amount of synchrony present at each of these 3 frequency bands to investigate whether the degree of sentiment synchrony differed depending on the timescale.

Mixed-Effects Growth Curve for Modeling Sentiment Trajectories

Finally, we examined how users' comment sentiment changed as a function of activity on the subreddit. To examine the trajectories of comment sentiment, we used the *LME4* package in R to fit a mixed-effects growth curve model. A moving window transform was applied to improve the interpretability of model coefficients. We divided each user's data into a series of equally proportioned windows and calculated average comment sentiment within each *interaction window* such that each user had 100 data points representing 1% of their total activity. In other words, the value at interaction window 1 reflected the average sentiment over the first 1% of a user's comments, while the value at interaction window 100 reflected the average sentiment over the last 1% of a user's comments. This approach has been used in similar literature [112].

For the growth curve analysis, we fit a mixed-effects model with comment sentiment entered as the dependent variable, interaction window as a fixed-effect, and user ID as a random intercept. The common power estimate, average word count, average number of days active on the subreddit, and percentage of health-related words were entered as covariates. We included word count and the percentage of health-related words to control for the potentially confounding effects of abnormally long comments or medical terms on sentiment scores (ie, certain medical terms could be misclassified as negative). Furthermore, common power was included as an estimate of overall sentiment synchrony, and this allowed us to examine individual changes in sentiment independent of sentiment change related to entrainment with peers [91]. Finally, we included the number of days active in the subreddit to compare users with long versus short tenures in the community.

Results

Descriptive Statistics

Table 2 contains descriptive statistics related to the Reddit activity of users in our sample, the sentiment of their comments, feedback received from their peers, and wavelet synchrony metrics. Frequency distributions for comment sentiment, common power, and coherence can be found in Figure 2. Although we only analyzed interactions within the support community, it is worth noting users made a substantial number of posts to other subreddits. On average, activity inside of the support group constituted 22.03% (SD 24.01%) of users' total Reddit activity.

Table 2. Descriptive statistics.

Variable	Value, mean (SD)
Comments made per user	371.24 (518.18)
Comments made per day	2.87 (2.09)
Number of days active in the community	141.22 (142.56)
Sentiment of personal comments	–0.01 (0.24)
Sentiment of peer comments	–0.28 (0.15)
Number of replies received from peers	200.57 (272.63)
Comment score (upvotes–downvotes)	3.17 (1.27)
Common power (%XWT[a])	0.06 (0.02)
Coherence (%WTC[b])	0.08 (0.04)

[a]XWT: cross-wavelet transform.

[b]WTC: wavelet coherence.

XSL•FO
RenderX

Figure 2. Frequency distributions for primary outcome variables, including (A) user comment sentiment, (B) common power, and (C) coherence.

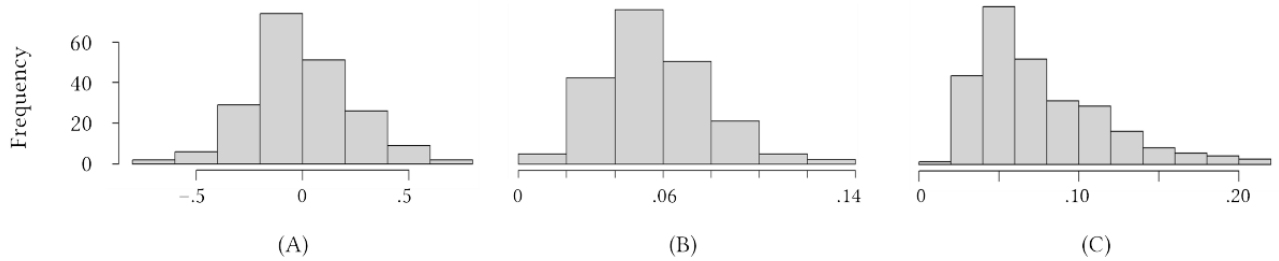

Users Exhibited Significant Synchrony Compared to Shuffled Baseline

We found users had significantly more common power during their real support interactions (mean 0.06, SD 0.02) compared to a shuffled baseline (mean 0.04, SD 0.01; t_{198}=4.02; P<.001; Cohen d=0.40). Likewise, there was significantly more sentiment coherence during real interactions (mean 0.08, SD 0.04) as compared to shuffled data (mean 0.05, SD 0.03; t_{198}=6.57; P<.001; Cohen d=0.64). These results suggest users synchronized in terms of matching each other's intensity (common power), and changes in user sentiment corresponded to changes in peer sentiment (coherence).

Comparing these results to a traditional Pearson correlation approach (Table 3), we found significant positive correlations between Pearson correlation coefficients, real-data coherence (r_{197}=0.46; P<.001), and real-data common power (r_{197}=0.18; P<.001). However, there were no relationships between Pearson coefficients and wavelet calculations in the shuffled data (P=.77 and .50). Taken together, we found evidence that users synchronized their sentiment during support interactions beyond what could be explained by random variance in their data. The results of our correlation analysis also supported the validity of our synchrony analysis.

Table 3. Correlation analysis relating traditional synchrony measure (Pearson correlation) to bivariate wavelet estimates (original and shuffled data)[a].

Variable	Pearson coefficients	Coherence	Coherence shuffled	Common power	Common power shuffled
Pearson coefficients					
r	1	0.46	–0.02	0.18	0.05
P value	__[b]	<.001	.77	.01	.50
Coherence					
r	0.46	1	0.07	0.41	–0.05
P value	<.001	—	.35	<.001	.48
Coherence shuffled					
r	–0.02	0.07	1	–0.04	0.14
P value	.77	.35	—	.60	.05
Common power					
r	0.18	0.41	–0.04	1	–0.01
P value	.01	<.001	.60	—	.87
Common power shuffled					
r	0.05	–0.05	0.14	–0.01	1
P value	.50	.48	.05	.87	—

[a]Pearson correlation, listwise deletion.

[b]Not applicable.

User Sentiment Fluctuated In-Phase With Peer Sentiment

Across all users, we found an average circular RP angle of 2.81°, suggesting users were *in-phase* with the sentiment of their peers during interactions (Figure 3). Thus, the sentiment intensity of users fluctuated in *the same direction* without a clear leader or follower, and periods of positive posting by a user were mirrored by periods of positive posting by their peers.

XSL•FO
RenderX

Figure 3. Frequency distribution for average relative phase (RP) angle by user.

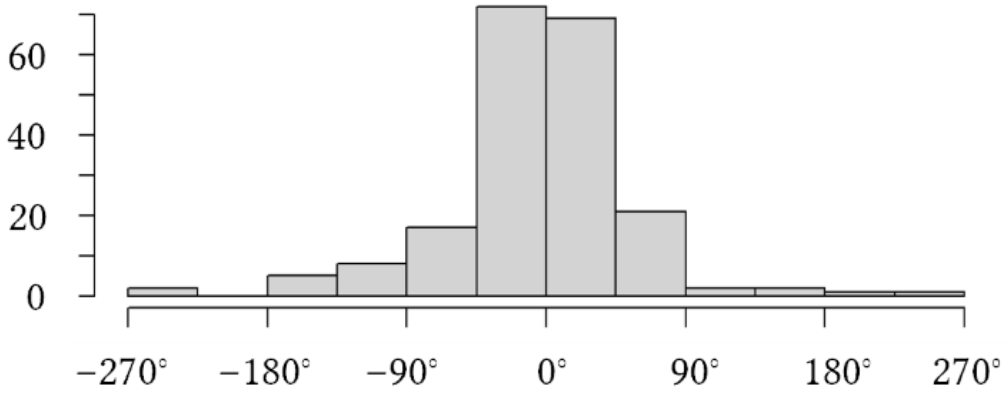

Users Had Greater Synchrony in Low- Versus High-Frequency Sentiment Changes

We found that the average common power varied significantly depending on frequency ($F_{2,198}$=17.70; P<.001). The common power in low-frequency bands (mean 0.08, SD 0.10) was significantly higher than medium- (mean 0.05, SD 0.05; t_{199}=3.66; P<.001) or high-frequency bands (mean 0.05, SD 0.01; t_{199}=5.05; P<.001). The same was true for coherence ($F_{2,198}$=3.79; P=.01), as there was significantly more coherence in low (mean 0.10, SD 0.15) compared to high frequencies (mean 0.07, SD 0.03; t_{199}=2.36; P=.02). There were no differences in coherence between medium and high frequencies (P=.13). These findings suggest that synchrony was most evident in long-term sentiment fluctuations occurring over hundreds of interactions as opposed to the more reactionary changes in sentiment occurring from comment to comment. There were no differences in RP angle (P=.11), indicating the sentiment of users and their peers were in-phase regardless of scale.

Sentiment Trajectories Depended on Activity and Degree of Synchrony

Across all users, sentiment decreased significantly when interacting with the subreddit (β=–.02; P=.003), and this negative trend was observed while controlling for differences in average word count (β=–.10; P<.001) and average health-related language (β=–.20; P<.001; Table 4). In addition, there was a significant 3-way interaction between interaction number, synchrony, and the number of days active (β=.02; P=.02), indicating the trajectory of comment sentiment depended on the degree of emotional synchrony and the duration of subreddit participation. Users who were active on the subreddit for a greater number of days exhibited a steeper decline in sentiment compared to those who were active for shorter durations. Thus, sentiment not only decreased as users interacted with their peers but also decreased longitudinally with the total number of days spent seeking online support. However, this compounding effect depended on affective synchrony. Greater affective synchrony was associated with negative changes in sentiment for those who spent less time on the subreddit, but it functioned as a protective factor for those who were active for longer (Figure 4).

Table 4. Mixed-effects growth curve model examining the trajectory of comment sentiment and interactions with synchrony and days active[a].

Predictors	Sentiment		
	β	Standard CI	*P* value
Intercept	–.00	–0.04 to 0.03	.80
Interaction window	–.02	–0.03 to –0.01	.003
Common power	–.01	–0.04 to 0.03	.71
Number of days active	0	–0.04 to 0.05	.74
Percent of health-related words	–.20	–0.24 to –0.16	<.001
Average word count	–.10	–0.14 to –0.06	<.001
Interaction window: common power	0	–0.01 to 0.02	.15
Interaction window: number of days	–.02	–0.03 to –0.00	.04
Common power: number of Days	.03	–0.01 to 0.07	.92
Interaction window: common power: number of days	.02	0.00 to –0.03	.02

[a]σ^2=0.88, τ_{00} (random intercept variance)=0.07; intraclass correlation coefficient=0.07; N_{author}=199; observations=19,892; marginal R^2=0.05; conditional R^2=0.12.

XSL•FO
RenderX

Figure 4. A 3-way interaction effect of common power, number of days active on subreddit, and sentiment scores. XWT: cross-wavelet transform.

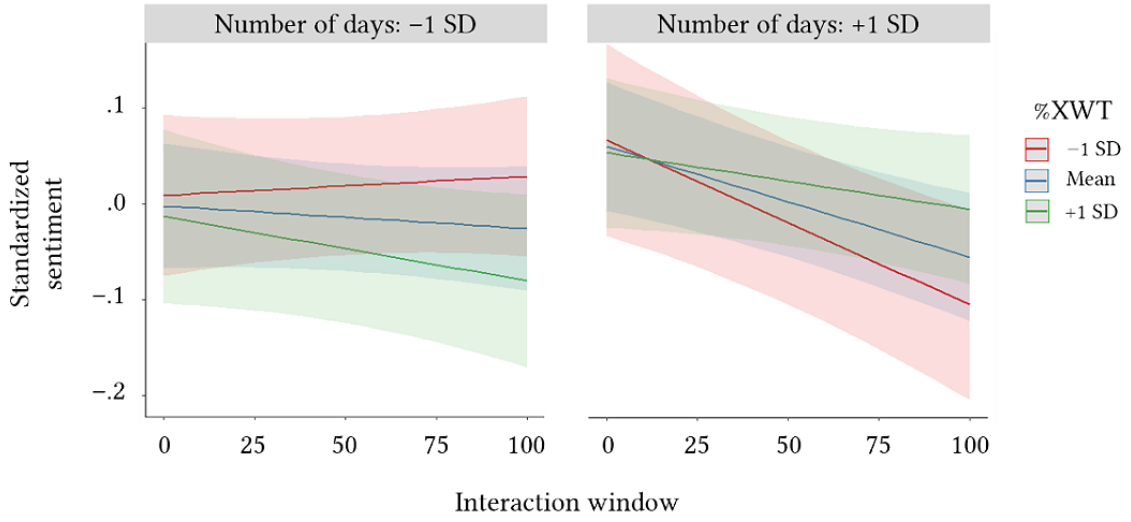

Discussion

Principal Findings

Sentiment Synchrony

We found evidence that social media users synchronized the emotion of their language during conversations within a pain support community on Reddit. Users not only matched the sentiment intensity of their peers but also exhibited significant phased-locked behavior, showing positively correlated (ie, in-phase) changes in sentiment at several different frequency scales. Synchrony during dyadic interactions is a well-documented phenomenon, as conversation partners tend to mimic each other's emotions [113], language style [114,115], and nonverbal patterns of behaviors [111,116]. Wood et al [113] suggest that affective synchrony assists with information processing, emotional regulation, and social bonding. Under this framework, synchrony by members of the support community may reflect a more general process underlying supportive interactions, whereby participants bond over shared experiences using similarly intense emotional expressions. Synchrony is also associated with positive benefits for conversation partners, including increased feelings of social connectedness [117] and positive changes in affect after interaction [117,118]. For example, a study on online support seeking reported that greater linguistic synchrony during text-based discussions predicted satisfaction with support and improvements in mood [114]. We suspect affective synchrony in the chronic pain subreddit was an indicator of supportive (vs adversarial [14]) interactions and empathetic language [113]. Peer support aims to connect individuals who have similar medical conditions with the hope that their common experiences promote feelings of acceptance and understanding [119]. By that basic conceptualization, affective and experiential synchrony can be seen as essential elements of peer support. Although we do not collect data about health status, future research should investigate whether sentiment synchrony is predictive of positive versus negative health outcomes.

Our investigation of synchrony by timescale (ie, low, medium, and high frequencies), using the XWT, revealed that users were most synchronous with their long-term sentiment dynamics. Each data point included a single interaction between a user and one of their peers, such that low-frequency synchrony (unfolding over hundreds of interactions) represented large-scale coordination across hundreds of community members. In other words, our findings reflect a tendency of users to affectively synchronize at the level of the community over long periods as opposed to at the level of individual interactions. This tendency can be interpreted as a type of social learning, showing how repeated community interactions influence emotional expressions over long periods. Social learning occurs when people observe and imitate the behaviors of others [120], and, in the context of affective synchrony, it involves multiple people coregulating their emotions to align with a larger group [113]. It is reasonable that online community members would steadily adjust their emotional expressions as they become more familiar with group dynamics.

A concern among medical professionals in recommending web-based resources is that social media users will be negatively influenced by misinformation or the attitudes of peers [121,122]. In addition, a study on user perceptions found that users often experienced negative web-based interactions that damaged their opinions of online support [123]. By contrast, our results are a promising indicator that users were partaking in supportive and emotionally synchronous interactions. However, synchrony has the potential to be harmful if it proliferates negative health attitudes. Related to this concern, we observed in our descriptive statistics that the sentiment of users' personal comments was more positive on average than the content they interacted with, suggesting users often interacted with content containing more negative language than their own. Affective synchrony among users highlights the importance of moderation in peer support communities for promoting positive as opposed to "toxic" (eg, [124,125]) interactions. For example, it is common for many of the pain support communities on Reddit to prohibit users from providing explicit medical advice and to encourage users to talk to their providers about their participation in the community. However, it is unclear how strictly these rules were enforced in the subreddit we analyzed or in OHCs more broadly.

XSL•FO
RenderX

Trajectories in User Sentiment

The biopsychosocial model of pain emphasizes how supportive relationships can promote mental well-being [14], but it is unclear how naturally occurring discussions in OHCs influence attitudes toward health. Our mixed effect growth curve model revealed a significant increase in negative sentiment over time across all users, controlling for differences in health-related language (eg, "pain" and "health") and word count. To our knowledge, this is the first study to examine sentiment trajectories in online *chronic pain* communities. Previous research has focused primarily on mental health support, and findings from those studies have been highly variable [96,97,126]. For example, Davcheva et al [96] reported that changes in sentiment on mental health forums depended on underlying conditions, such that seeking support for anxiety and depression was associated with positive changes in sentiment, and obsessive-compulsive disorder was associated with negative changes in sentiment. This study contributes to this literature by demonstrating how aspects of social media use can further enhance online support outcomes. Specifically, affective asynchrony and community tenure predicted negative versus positive sentiment trajectories. Moreover, our findings align with the SCMP, which describes pain expressions as evolving over time in response to pain-relevant social interactions.

Given the purported emotional benefits of peer support [14,32], findings of a negative trend in sentiment across users were somewhat unexpected. Chronic pain is referred to as a "vicious cycle" [127], with physical and psychological symptoms progressing many years beyond their onset [8,41]. Several studies have indicated long-term declines in affect co-occurring with chronic illness [128-130], and psychological distress and pain are theorized to be reciprocal [14,38,40,131]. Thus, our results may have reflected the continued frustrations experienced by users during their pain self-management journey as opposed to representing any specific adverse impact of web-based interactions. The fear-avoidance (FA) model describes chronic pain as a negative affective feedback loop between catastrophizing and pain, and this "downward spiral" [132] of affect propels symptoms forward until the loop can be interrupted [133,134]. The negative sentiment trajectories observed in this study are consistent with the downward spiral of affect described by the FA model, and, aligning with this idea, users who spent the most time on the subreddit exhibited the most pronounced increases in negative sentiment. It is also important to highlight that a portion of users progressed positively in our sample despite the overall negative trend, and these individual differences were related to having lower community synchrony and shorter community tenures. By identifying the factors that contribute to positive sentiment trajectories, it may be possible to develop personalized interventions that target users based on the dynamics of their interactions. According to the FA model of chronic pain, users may benefit from information about pain catastrophizing, mental health, and social support [134]. This could be provided in the form of a "stickied" thread that remains at the top of the forum or as an automated message sent to members.

Finally, affective synchrony emerged as a significant factor in predicting positive versus negative sentiment outcomes. For long-tenured community members, synchrony appeared to function as a protective factor against increased negativity. This is consistent with research conceptualizing synchrony as a process promoting social connectedness and emotional support [113,114]. In other words, we would expect individuals involved in more supportive interactions to be more emotionally resilient [135,136]. It is less clear as to why synchrony was dually associated with negative sentiment trajectories among *short*-tenured community members. One possibility is that these members differed in the type of support they were seeking. There was a significant correlation between the number of days active on the subreddit and word count, r_{197}=0.18; P<.001, indicating that short-tenured members were involved in less verbose discussions. It is possible these users were seeking a type of support in which affective synchrony was irrelevant, such as by seeking occasional medical information as opposed to the emotional validation achieved by sharing personal narratives (eg, [76]).

Strengths and Limitations

Our findings make several contributions to the literature. First, our study is unique in quantitatively examining the dynamics of online chronic pain support communities, contrasting the current literature that is primarily qualitative and focused on pain narratives [28,73,80,87,123,137,138]. Second, the qualities of pain are typically investigated through measures of central tendency (ie, by comparing average pain catastrophizing across groups). However, the biopsychosocial model describes chronic pain as a dynamic (changing over time) and complex (evolving from the interactions of multiple factors) process [14,139,140]. Consistent with this framing of chronic pain, we used wavelet analysis to analyze sentiment synchrony, which is an analytic technique better suited to describe complex system interactions and nonstationary behavior [92,107]. Next, previous studies of online pain support groups have often relied on small sample sizes and do not analyze data longitudinally. In comparison, we analyzed the entire comment histories of Reddit users consisting of over a million comments and approximately 70,000 interactions within 1 OHC. Thus, we present a comprehensive view of users' web-based interactions inside the chronic pain subreddit over nearly a decade. Finally, as opposed to examining the impact of a structured intervention (eg, [27,85]), we examined naturally occurring conversations that are more representative of everyday social interactions. Unlike a structured intervention or professionally managed support group, our analyses considered undirected web-based conversations where participants likely encountered a wide variety of positive and negative actors. Participants were free to express their emotions with little inhibition related to research participation, especially due to the relatively anonymous nature of Reddit [33]. This type of longitudinal and naturalistic look into an individual's pain expressions would be exceedingly difficult to collect in traditional settings [33].

Despite the strengths of our quantitative approach, there are several limitations. Notably, our investigations lacked data about user health outcomes. We investigated text sentiment as a correlate to chronic pain attitudes and emotions, and our results

XSL•FO
RenderX

provided novel insights regarding how sentiment progressed in 1 OHC. However, without data on health outcomes, we cannot know if these trends were indicative of declines in physical or psychological well-being. Users could have viewed the forum as an opportunity for emotional catharsis, venting about their pain increasingly over time. There are some circumstances where negative emotional expression can lead to positive outcomes [141-143], particularly if those expressions are positively reframed by the audience [144]. In contrast, emotional catharsis can be counterproductive and lead to agitation [145,146]. Thus, it is necessary to exercise caution when drawing conclusions about the merits of support seeking on social media platforms based on the findings of this study.

Another potential limitation relates to our selection of Reddit as a data source. Reddit blends features from popular social networking sites (eg, the ability to personalize a user profile, add friends, or form groups) with the formatting of a web-based discussion board. As a result, the interactions on Reddit tend to surround topical discussions as opposed to resharing or reposting user-generated content. This format facilitates verbose and focused conversations, and it allows us to collect rich longitudinal data about pain-related social interactions. Reddit also affords users increased anonymity compared to social networking sites such as Facebook, as a user's Reddit profile is not tied to their identity, friends, or family. There is evidence to suggest that social media users seeking chronic pain support prefer the ability to remain anonymous [28], thus, we suspected users would more candidly share their feelings on Reddit compared to a public-facing social networking site. However, given the unique features of Reddit, it is not clear how our findings generalize to more popular social media platforms that promote different styles of communication. It is also not clear if our findings concerning chronic pain support can be applied to OHCs for other types of chronic health conditions.

Conclusions

Although social media provides easy access to massive peer support networks, it also has the potential to spread negative attitudes and beliefs about pain. We found evidence that social media users synchronized the emotional intensity of their language during virtual conversations about chronic pain, which can be viewed as an indicator of social bonding [113] and highlights the influence of OHC interactions on pain expressions. Furthermore, affective synchrony functioned as a protective factor against continued negative language use among those with the most negative expressions of pain in the online community. Despite these promising findings, we noted a steady increase in negative sentiment averaged across users as they continued their community participation. This negative trend is consistent with the "downward spiral" of affect described by the FA model of pain [132], suggesting that the impact of OHCs on pain expressions varies to a large degree dependent on specific user behaviors. Our results highlight the importance of considering specific user and community dynamics when assessing the impact of OHCs. Furthermore, our synchrony indices show there is a high degree of peer influence on pain attitudes and expressions, consistent with the SCMP.

Data Availability

The data generated during and analyzed during this study are not publicly available due to the developer data terms provided by Reddit, which prevents us from distributing or resharing data collected using the Reddit application programming interface.

Authors' Contributions

AN was responsible for project design, carrying out the investigation, data analysis, and writing the manuscript. MJA served a supervisory role, contributing to conceptualization and methodology, in addition to writing, reviewing, and editing the manuscript.

Conflicts of Interest

None declared.

References

1. Chronic pain: what you need to know. National Institutes of Health National Center for Complementary and Integrative Health. URL: https://www.nccih.nih.gov/health/chronic-pain-what-you-need-to-know [accessed 2022-06-07]
2. Pain. National Institutes of Health National Institute of Neurological Disorders and Stroke. URL: https://www.ninds.nih.gov/health-information/disorders/pain [accessed 2022-06-07]
3. Cohen SP, Vase L, Hooten WM. Chronic pain: an update on burden, best practices, and new advances. Lancet 2021 May 29;397(10289):2082-2097. [doi: 10.1016/S0140-6736(21)00393-7] [Medline: 34062143]
4. Zelaya CE, Dahlhamer JM, Lucas JW, Connor EM. Chronic pain and high-impact chronic pain among U.S. adults, 2019. NCHS Data Brief 2020 Nov(390):1-8. [Medline: 33151145]
5. Yong RJ, Mullins PM, Bhattacharyya N. Prevalence of chronic pain among adults in the United States. Pain 2022 Feb 01;163(2):e328-e332. [doi: 10.1097/j.pain.0000000000002291] [Medline: 33990113]
6. Almalki MT, BinBaz SS, Alamri SS, Alghamdi HH, El-Kabbani AO, Al Mulhem AA, et al. Prevalence of chronic pain and high-impact chronic pain in Saudi Arabia. Saudi Med J 2019 Dec;40(12):1256-1266 [FREE Full text] [doi: 10.15537/smj.2019.12.24690] [Medline: 31828278]
7. Rustøen T, Wahl AK, Hanestad BR, Lerdal A, Paul S, Miaskowski C. Prevalence and characteristics of chronic pain in the general Norwegian population. Eur J Pain 2004 Dec;8(6):555-565. [doi: 10.1016/j.ejpain.2004.02.002] [Medline: 15531224]

XSL•FO
RenderX

8. Tunks ER, Crook J, Weir R. Epidemiology of chronic pain with psychological comorbidity: prevalence, risk, course, and prognosis. Can J Psychiatry 2008 Apr;53(4):224-234. [doi: 10.1177/070674370805300403] [Medline: 18478825]
9. Sá KN, Moreira L, Baptista AF, Yeng LT, Teixeira MJ, Galhardoni R, et al. Prevalence of chronic pain in developing countries: systematic review and meta-analysis. Pain Rep 2019 Dec 06;4(6):e779 [FREE Full text] [doi: 10.1097/PR9.0000000000000779] [Medline: 31984290]
10. Elzahaf RA, Tashani OA, Unsworth BA, Johnson MI. The prevalence of chronic pain with an analysis of countries with a Human Development Index less than 0.9: a systematic review without meta-analysis. Curr Med Res Opin 2012 Jul;28(7):1221-1229. [doi: 10.1185/03007995.2012.703132] [Medline: 22697274]
11. Carr DB. "Pain is a public health problem" --what does that mean and why should we care? Pain Med 2016 Apr;17(4):626-627. [doi: 10.1093/pm/pnw045] [Medline: 27052886]
12. Goldberg DS, McGee SJ. Pain as a global public health priority. BMC Public Health 2011 Oct 06;11:770 [FREE Full text] [doi: 10.1186/1471-2458-11-770] [Medline: 21978149]
13. Sarzi-Puttini P, Vellucci R, Zuccaro SM, Cherubino P, Labianca R, Fornasari D. The appropriate treatment of chronic pain. Clin Drug Investig 2012 Feb;32 Suppl 1:21-33. [doi: 10.2165/11630050-000000000-00000] [Medline: 23389873]
14. Edwards RR, Dworkin RH, Sullivan MD, Turk DC, Wasan AD. The role of psychosocial processes in the development and maintenance of chronic pain. J Pain 2016 Sep;17(9 Suppl):T70-T92 [FREE Full text] [doi: 10.1016/j.jpain.2016.01.001] [Medline: 27586832]
15. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 08;196(4286):129-136 [FREE Full text] [doi: 10.1126/science.847460] [Medline: 847460]
16. Devan H, Hale L, Hempel D, Saipe B, Perry MA. What works and does not work in a self-management intervention for people with chronic pain? Qualitative systematic review and meta-synthesis. Phys Ther 2018 May 01;98(5):381-397. [doi: 10.1093/ptj/pzy029] [Medline: 29669089]
17. Shery M, MacNeil C. Peer support: what makes it unique? Int J Psychosocial Rehabil 2004;10(2) [FREE Full text]
18. Nicholas MK, Asghari A, Blyth FM, Wood BM, Murray R, McCabe R, et al. Long-term outcomes from training in self-management of chronic pain in an elderly population: a randomized controlled trial. Pain 2017 Jan;158(1):86-95. [doi: 10.1097/j.pain.0000000000000729] [Medline: 27682207]
19. Bartone PT, Bartone JV, Violanti JM, Gileno ZM. Peer support services for bereaved survivors: a systematic review. Omega (Westport) 2019 Nov;80(1):137-166. [doi: 10.1177/0030222817728204] [Medline: 28871835]
20. Dennis CL. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321-332. [doi: 10.1016/S0020-7489(02)00092-5]
21. Farr M, Brant H, Patel R, Linton MJ, Ambler N, Vyas S, et al. Experiences of patient-led chronic pain peer support groups after pain management programs: a qualitative study. Pain Med 2021 Dec 11;22(12):2884-2895 [FREE Full text] [doi: 10.1093/pm/pnab189] [Medline: 34180996]
22. Pester BD, Tankha H, Caño A, Tong S, Grekin E, Bruinsma J, et al. Facing pain together: a randomized controlled trial of the effects of Facebook support groups on adults with chronic pain. J Pain 2022 Dec;23(12):2121-2134 [FREE Full text] [doi: 10.1016/j.jpain.2022.07.013] [Medline: 36096353]
23. Stenberg N, Gillison F, Rodham K. How do peer support interventions for the self-management of chronic pain, support basic psychological needs? A systematic review and framework synthesis using self-determination theory. Patient Educ Couns 2022 Nov;105(11):3225-3234 [FREE Full text] [doi: 10.1016/j.pec.2022.07.017] [Medline: 35985906]
24. Tolley JA, Michel MA, Williams AE, Renschler JS. Peer support in the treatment of chronic pain in adolescents: a review of the literature and available resources. Children (Basel) 2020 Sep 07;7(9):129 [FREE Full text] [doi: 10.3390/children7090129] [Medline: 32906581]
25. Naslund JA, Grande SW, Aschbrenner KA, Elwyn G. Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube. PLoS One 2014 Oct 15;9(10):e110171 [FREE Full text] [doi: 10.1371/journal.pone.0110171] [Medline: 25333470]
26. Daynes-Kearney R, Gallagher S. Online support groups for family caregivers: scoping review. J Med Internet Res 2023 Dec 13;25:e46858 [FREE Full text] [doi: 10.2196/46858] [Medline: 38090796]
27. Hossain SN, Jaglal SB, Shepherd J, Perrier L, Tomasone JR, Sweet SN, et al. Web-based peer support interventions for adults living with chronic conditions: scoping review. JMIR Rehabil Assist Technol 2021 May 25;8(2):e14321 [FREE Full text] [doi: 10.2196/14321] [Medline: 34032572]
28. Merolli M, Gray K, Martin-Sanchez F. Patient participation in chronic pain management through social media: a clinical study. Stud Health Technol Inform 2016;225:577-581. [Medline: 27332267]
29. Brennen JS, Simon F, Howard PN, Nielsen RK. Types, sources, and claims of COVID-19 misinformation. Reuters Institute for the Study of Journalism. 2020 Apr 7. URL: https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation [accessed 2024-08-13]
30. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
31. Ziemer KS, Fuhrmann A, Hoffman MA. Effectiveness of a positive writing intervention for chronic pain: a randomized trial. Myopain 2017 Apr 11;23(3-4):143-154. [doi: 10.1080/24708593.2017.1307893]

XSL•FO
RenderX

32. Guillory J, Chang P, Henderson CRJ, Shengelia R, Lama S, Warmington M, et al. Piloting a text message-based social support intervention for patients with chronic pain: establishing feasibility and preliminary efficacy. Clin J Pain 2015 Jun;31(6):548-556 [FREE Full text] [doi: 10.1097/AJP.0000000000000193] [Medline: 25565587]
33. Caes L, Jones A, Jordan A. Engaging use of social media as a research tool to capture the daily life experiences of young people with chronic pain. Evid Based Nurs 2018 Oct;21(4):91-93. [doi: 10.1136/eb-2018-102980] [Medline: 30158175]
34. Cooper K, Kirkpatrick P, Wilcock S. The effectiveness of peer support interventions for community-dwelling adults with chronic non-cancer pain: a systematic review. JBI Database System Rev Implement Rep 2014 May;12(5):319-348. [doi: 10.11124/jbisrir-2014-1590]
35. Palant A, Himmel W. Are there also negative effects of social support? A qualitative study of patients with inflammatory bowel disease. BMJ Open 2019 Jan 21;9(1):e022642 [FREE Full text] [doi: 10.1136/bmjopen-2018-022642] [Medline: 30670504]
36. Gómez Penedo JM, Rubel JA, Blättler L, Schmidt SJ, Stewart J, Egloff N, et al. The complex interplay of pain, depression, and anxiety symptoms in patients with chronic pain: a network approach. Clin J Pain 2020 Apr;36(4):249-259. [doi: 10.1097/AJP.0000000000000797] [Medline: 31899722]
37. Lee H, Hübscher M, Moseley GL, Kamper SJ, Traeger AC, Mansell G, et al. How does pain lead to disability? A systematic review and meta-analysis of mediation studies in people with back and neck pain. Pain 2015 Jun;156(6):988-997. [doi: 10.1097/j.pain.0000000000000146] [Medline: 25760473]
38. Lerman SF, Rudich Z, Brill S, Shalev H, Shahar G. Longitudinal associations between depression, anxiety, pain, and pain-related disability in chronic pain patients. Psychosom Med 2015 Apr;77(3):333-341. [doi: 10.1097/PSY.0000000000000158] [Medline: 25849129]
39. Knuppenburg RM, Fredericks CM. Linguistic affect: positive and negative emotion words are contagious, predict likability, and moderate positive and negative affect. Inq J 2021;13(03):1.
40. Blyth FM, Macfarlane GJ, Nicholas MK. The contribution of psychosocial factors to the development of chronic pain: the key to better outcomes for patients? Pain 2007 May;129(1-2):8-11. [doi: 10.1016/j.pain.2007.03.009] [Medline: 17398007]
41. Gatchel RJ, McGeary DD, McGeary CA, Lippe B. Interdisciplinary chronic pain management: past, present, and future. Am Psychol 2014;69(2):119-130. [doi: 10.1037/a0035514] [Medline: 24547798]
42. Diatchenko L, Fillingim RB, Smith SB, Maixner W. The phenotypic and genetic signatures of common musculoskeletal pain conditions. Nat Rev Rheumatol 2013 Jun;9(6):340-350 [FREE Full text] [doi: 10.1038/nrrheum.2013.43] [Medline: 23545734]
43. Şentürk İ, Şentürk E, Üstün I, Gökçedağ A, Yıldırım NP, İçen NK. High-impact chronic pain: evaluation of risk factors and predictors. Korean J Pain 2023 Jan 01;36(1):84-97 [FREE Full text] [doi: 10.3344/kjp.22357] [Medline: 36581599]
44. Wager J, Brown D, Kupitz A, Rosenthal N, Zernikow B. Prevalence and associated psychosocial and health factors of chronic pain in adolescents: differences by sex and age. Eur J Pain 2020 Apr;24(4):761-772. [doi: 10.1002/ejp.1526] [Medline: 31889351]
45. Hall AM, Kamper SJ, Maher CG, Latimer J, Ferreira ML, Nicholas MK. Symptoms of depression and stress mediate the effect of pain on disability. Pain 2011 May;152(5):1044-1051. [doi: 10.1016/j.pain.2011.01.014] [Medline: 21306826]
46. Ganster DC, Victor B. The impact of social support on mental and physical health. Br J Med Psychol 1988 Mar;61 (Pt 1):17-36. [doi: 10.1111/j.2044-8341.1988.tb02763.x] [Medline: 3282536]
47. Hale CJ, Hannum JW, Espelage DL. Social support and physical health: the importance of belonging. J Am Coll Health 2010 Aug 07;53(6):276-284. [doi: 10.3200/jach.53.6.276-284]
48. Kelly ME, Duff H, Kelly S, McHugh Power JE, Brennan S, Lawlor BA, et al. The impact of social activities, social networks, social support and social relationships on the cognitive functioning of healthy older adults: a systematic review. Syst Rev 2017 Dec 19;6(1):259 [FREE Full text] [doi: 10.1186/s13643-017-0632-2] [Medline: 29258596]
49. Vandervoort D. Quality of social support in mental and physical health. Curr Psychol 1999 Jun;18:205-221. [doi: 10.1007/s12144-999-1029-8]
50. Holahan CJ, Moos RH. Social support and psychological distress: a longitudinal analysis. J Abnorm Psychol 1981 Aug;90(4):365-370. [doi: 10.1037//0021-843x.90.4.365] [Medline: 7264067]
51. Lepore SJ. Social conflict, social support, and psychological distress: evidence of cross-domain buffering effects. J Pers Soc Psychol 1992 Nov;63(5):857-867. [doi: 10.1037//0022-3514.63.5.857] [Medline: 1447698]
52. Menec VH, Newall NE, Mackenzie CS, Shooshtari S, Nowicki S. Examining social isolation and loneliness in combination in relation to social support and psychological distress using Canadian Longitudinal Study of Aging (CLSA) data. PLoS One 2020 Mar 23;15(3):e0230673 [FREE Full text] [doi: 10.1371/journal.pone.0230673] [Medline: 32203553]
53. Uysal A, Ascigil E, Turunc G. Spousal autonomy support, need satisfaction, and well-being in individuals with chronic pain: a longitudinal study. J Behav Med 2017 Apr;40(2):281-292. [doi: 10.1007/s10865-016-9783-1] [Medline: 27541643]
54. Abbasi M, Dehghani M, Keefe FJ, Jafari H, Behtash H, Shams J. Spouse-assisted training in pain coping skills and the outcome of multidisciplinary pain management for chronic low back pain treatment: a 1-year randomized controlled trial. Eur J Pain 2012 Aug;16(7):1033-1043. [doi: 10.1002/j.1532-2149.2011.00097.x] [Medline: 22337646]

XSL•FO
RenderX

55. Keefe FJ, Caldwell DS, Baucom D, Salley A, Robinson E, Timmons K, et al. Spouse-assisted coping skills training in the management of osteoarthritic knee pain. Arthritis Care Res 1996 Aug;9(4):279-291. [doi: 10.1002/1529-0131(199608)9:4<279::aid-anr1790090413>3.0.co;2-6] [Medline: 8997917]
56. Craig KD. The social communication model of pain. Can Psychol 2009;50(1):22-32. [doi: 10.1037/a0014772]
57. Hadjistavropoulos T, Craig KD, Duck S, Cano A, Goubert L, Jackson PL, et al. A biopsychosocial formulation of pain communication. Psychol Bull 2011 Nov;137(6):910-939. [doi: 10.1037/a0023876] [Medline: 21639605]
58. Sullivan MJ. The communal coping model of pain catastrophising: clinical and research implications. Can Psychol 2012;53(1):32-41. [doi: 10.1037/a0026726]
59. Craig KD. Social communication model of pain. Pain 2015 Jul;156(7):1198-1199. [doi: 10.1097/j.pain.0000000000000185] [Medline: 26086113]
60. Lapidot-Lefler N, Barak A. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. Comput Hum Behav 2012 Mar;28(2):434-443. [doi: 10.1016/j.chb.2011.10.014]
61. Malik S, Coulson NS. The therapeutic potential of the internet: exploring self-help processes in an internet forum for young people with inflammatory bowel disease. Gastroenterol Nurs 2011;34(6):439-448. [doi: 10.1097/SGA.0b013e318237a9ba] [Medline: 22129797]
62. Omernick E, Sood SO. The impact of anonymity in online communities. In: Proceedings of the International Conference on Social Computing. 2013 Presented at: SocialCom 2013; September 8-14, 2013; Alexandria, VA. [doi: 10.1109/socialcom.2013.80]
63. Carr CT, Hayes RA. Social media: defining, developing, and divining. Atl J Commun 2015 Feb 06;23(1):46-65. [doi: 10.1080/15456870.2015.972282]
64. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 2010;53(1):59-68. [doi: 10.1016/j.bushor.2009.09.003]
65. Landro L. Alone together. Cancer patients and survivors find treatment--and support--online. It can make all the difference. Oncologist 1999;4(1):59-63 [FREE Full text] [Medline: 10337372]
66. Eysenbach G, Sa ER, Diepgen TL. Shopping around the internet today and tomorrow: towards the millennium of cyber medicine. BMJ 1999;319:1294 [FREE Full text] [doi: 10.1136/bmj.319.7220.1294]
67. Frank SR. Digital health care--the convergence of health care and the internet. J Ambul Care Manage 2000 Apr;23(2):8-17. [doi: 10.1097/00004479-200004000-00003] [Medline: 10848396]
68. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166 [FREE Full text] [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
69. Bennett GG, Glasgow RE. The delivery of public health interventions via the internet: actualizing their potential. Annu Rev Public Health 2009;30:273-292. [doi: 10.1146/annurev.publhealth.031308.100235] [Medline: 19296777]
70. Huh J, Kwon BC, Kim SH, Lee S, Choo J, Kim J, et al. Personas in online health communities. J Biomed Inform 2016 Oct;63:212-225 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.019] [Medline: 27568913]
71. Ashtari S, Taylor J, Lai G. Systematic review of social networking support groups for genetic-disorders pain management. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020 Presented at: HICSS 2020; January 7-10, 2020; Maui, HI. [doi: 10.24251/hicss.2020.423]
72. Dhar VK, Kim Y, Graff JT, Jung AD, Garrett J, Dick LE, et al. Benefit of social media on patient engagement and satisfaction: results of a 9-month, qualitative pilot study using Facebook. Surgery 2018 Mar;163(3):565-570. [doi: 10.1016/j.surg.2017.09.056] [Medline: 29195735]
73. Tsai S, Crawford E, Strong J. Seeking virtual social support through blogging: a content analysis of published blog posts written by people with chronic pain. Digit Health 2018 May 16;4:2055207618772669 [FREE Full text] [doi: 10.1177/2055207618772669] [Medline: 29942635]
74. Davidson L, Chinman M, Kloos B, Weingarten R, Stayner D, Tebes JK. Peer support among individuals with severe mental illness: a review of the evidence. Clin Psychol Sci Pract 1999;6(2):165-187. [doi: 10.1093/clipsy.6.2.165]
75. Hall SL, Ryan DJ, Beatty J, Grubbs L. Recommendations for peer-to-peer support for NICU parents. J Perinatol 2015 Dec;35 Suppl 1(Suppl 1):S9-13 [FREE Full text] [doi: 10.1038/jp.2015.143] [Medline: 26597805]
76. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness: a qualitative systematic review. Qual Health Res 2017 Jan;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
77. Merolli M, Gray K, Martin-Sanchez F. Health outcomes and related effects of using social media in chronic disease management: a literature review and analysis of affordances. J Biomed Inform 2013 Dec;46(6):957-969 [FREE Full text] [doi: 10.1016/j.jbi.2013.04.010] [Medline: 23702104]
78. Ruehlman L, Karoly P, Enders C. A randomized controlled evaluation of an online chronic pain self management program. Pain 2012 Feb;153(2):319-330 [FREE Full text] [doi: 10.1016/j.pain.2011.10.025] [Medline: 22133450]
79. Trudeau KJ, Pujol LA, DasMahapatra P, Wall R, Black RA, Zacharoff K. A randomized controlled trial of an online self-management program for adults with arthritis pain. J Behav Med 2015 Jun;38(3):483-496 [FREE Full text] [doi: 10.1007/s10865-015-9622-9] [Medline: 25686538]

XSL•FO
RenderX

80. Meldrum S, Savarimuthu BT, Licorish S, Tahir A, Bosu M, Jayakaran P. Is knee pain information on YouTube videos perceived to be helpful? An analysis of user comments and implications for dissemination on social media. Digit Health 2017 Mar 29;3:2055207617698908 [FREE Full text] [doi: 10.1177/2055207617698908] [Medline: 29942583]
81. Rod K. Finding ways to lift barriers to care for chronic pain patients: outcomes of using internet-based self-management activities to reduce pain and improve quality of life. Pain Res Manag 2016;2016:8714785 [FREE Full text] [doi: 10.1155/2016/8714785] [Medline: 27445632]
82. Willis E. Patients' self-efficacy within online health communities: facilitating chronic disease self-management behaviors through peer education. Health Commun 2016;31(3):299-307. [doi: 10.1080/10410236.2014.950019] [Medline: 26325224]
83. Berkanish P, Pan S, Viola A, Rademaker Q, Devine KA. Technology-based peer support interventions for adolescents with chronic illness: a systematic review. J Clin Psychol Med Settings 2022 Dec;29(4):911-942 [FREE Full text] [doi: 10.1007/s10880-022-09853-0] [Medline: 35147830]
84. Zhang R. The stress-buffering effect of self-disclosure on Facebook: an examination of stressful life events, social support, and mental health among college students. Comput Hum Behav 2017 Oct;75:527-537. [doi: 10.1016/j.chb.2017.05.043]
85. Young SD, Koussa M, Lee SJ, Perez H, Gill N, Gelberg L, et al. Feasibility of a social media/online community support group intervention among chronic pain patients on opioid therapy. J Addict Dis 2018;37(1-2):96-101 [FREE Full text] [doi: 10.1080/10550887.2018.1557992] [Medline: 30614403]
86. What is Reddit? Reddit. URL: https://support.reddithelp.com/hc/en-us/articles/204511479-What-is-Reddit [accessed 2022-06-20]
87. Goudman L, de Smedt A, Moens M. Social media and chronic pain: what do patients discuss? J Pers Med 2022 May 14;12(5):797 [FREE Full text] [doi: 10.3390/jpm12050797] [Medline: 35629218]
88. Hardeniya T, Borikar DA. Dictionary based approach to sentiment analysis - a review. Int J Adv Eng Manag Sci 2016;2(5):317 [FREE Full text]
89. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl Based Syst 2021 Aug 17;226:107134. [doi: 10.1016/j.knosys.2021.107134]
90. Wankhade M, Rao AC, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 2022 Feb 07;55(7):5731-5780. [doi: 10.1007/s10462-022-10144-1]
91. Grinsted A, Moore JC, Jevrejeva S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. Nonlin Processes Geophys 2004 Nov 18;11(5/6):561-566. [doi: 10.5194/npg-11-561-2004]
92. Issartel J, Bardainne T, Gaillot P, Marin L. The relevance of the cross-wavelet transform in the analysis of human interaction - a tutorial. Front Psychol 2015 Jan 09;5:1566 [FREE Full text] [doi: 10.3389/fpsyg.2014.01566] [Medline: 25620949]
93. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The Pushshift Reddit dataset. Proc Int AAAI Conf Web Soc Media 2020 May 26;14(1):830-839. [doi: 10.1609/icwsm.v14i1.7347]
94. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with Reddit data. Soc Sci Comput Rev 2019 Dec 18;39(5):943-960. [doi: 10.1177/0894439319893305]
95. Necaise A, Han J, Vrzáková H, Amon MJ. Understanding collective human behavior in social media networks via the dynamical hypothesis: applications to radicalization and conspiratorial beliefs. Top Cogn Sci 2023 Oct 18 (forthcoming). [doi: 10.1111/tops.12702] [Medline: 37850669]
96. Davcheva E, Adam M, Benlian A. User dynamics in mental health forums – a sentiment analysis perspective. In: Proceedings of 14th International Conference on Wirtschaftsinformatik. 2019 Presented at: Wi 2019; February 24-27, 2019; Siegen, Germany URL: https://aisel.aisnet.org/wi2019/track08/papers/4
97. Zhang S, Bantum E, Owen J, Elhadad N. Does sustained participation in an online health community affect sentiment? AMIA Annu Symp Proc 2014 Nov 14;2014:1970-1979 [FREE Full text] [Medline: 25954470]
98. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahwah, NJ: Erlbaum Publishers; 2001.
99. de Choudhury M. Anorexia on Tumblr: a characterization study. In: Proceedings of the 5th International Conference on Digital Health 2015. 2015 Presented at: DH '15; May 18-20, 2015; Florence, Italy. [doi: 10.1145/2750511.2750515]
100. del Pilar Salas-Zárate M, López-López E, Valencia-García R, Aussenac-Gilles N, Almela Á, Alor-Hernández G. A study on LIWC categories for opinion mining in Spanish reviews. J Inf Sci 2014 Aug 26;40(6):749-760. [doi: 10.1177/0165551514547842]
101. Sharma C, Whittle S, Haghighi PD, Burstein F, Keen H. Sentiment analysis of social media posts on pharmacotherapy: a scoping review. Pharmacol Res Perspect 2020 Oct 19;8(5):e00640 [FREE Full text] [doi: 10.1002/prp2.640] [Medline: 32813329]
102. Wang Y, Weber I, Mitra P. Quantified self meets social media: sharing of weight updates on Twitter. In: Proceedings of the 6th International Conference on Digital Health Conference. 2016 Presented at: DH '16; April 11-13, 2016; Montréal, QC. [doi: 10.1145/2896338.2896363]
103. Monzani D, Vergani L, Pizzoli SF, Marton G, Pravettoni G. Emotional tone, analytical thinking, and somatosensory processes of a sample of Italian tweets during the first phases of the COVID-19 pandemic: observational study. J Med Internet Res 2021 Oct 27;23(10):e29820 [FREE Full text] [doi: 10.2196/29820] [Medline: 34516386]

XSL•FO
RenderX

104. Hutto CJ, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. 2014 Presented at: ICWSM-14; June 1-4, 2014; Ann Arbor, MI. [doi: 10.1609/icwsm.v8i1.14550]
105. Elbagir S, Yang J. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In: Proceedings of the International MultiConference of Engineers and Computer Scientists 2019. 2019 Presented at: IMECS 2019; March 13-15, 2019; Hong Kong, China. [doi: 10.1142/9789811215094_0005]
106. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
107. Torrence C, Compo GP. A practical guide to wavelet analysis. Bull Am Meteorol Soc 1998 Jan;79(1):61-78. [doi: 10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2]
108. Gouhier TC, Grinsted A, Simko V. R package biwavelet: conduct univariate and bivariate wavelet analyses. GitHub. 2019. URL: https://github.com/tgouhier/biwavelet [accessed 2024-08-13]
109. Zar JH. Biostatistical Analysis. Hoboken, NJ: Prentice Hall; 1999.
110. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
111. Reddan MC, Young H, Falkner J, López-Solà M, Wager TD. Touch and social support influence interpersonal synchrony and pain. Soc Cogn Affect Neurosci 2020 Nov 10;15(10):1064-1075 [FREE Full text] [doi: 10.1093/scan/nsaa048] [Medline: 32301998]
112. Necaise A, Williams A, Vrzakova H, Amon MJ. Regularity versus novelty of users' multimodal comment patterns and dynamics as markers of social media radicalization. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media. 2021 Presented at: HT '21; August 30-September 2, 2021; Virtual Event. [doi: 10.1145/3465336.3475095]
113. Wood A, Lipson J, Zhao O, Niedenthal P. Forms and functions of affective synchrony. In: Robinson MD, Thomas LE, editors. Handbook of Embodied Psychology. Cham, Switzerland: Springer; 2021.
114. Doré BP, Morris RR. Linguistic synchrony predicts the immediate and lasting impact of text-based emotional support. Psychol Sci 2018 Oct 18;29(10):1716-1723. [doi: 10.1177/0956797618779971] [Medline: 30020863]
115. Ireland ME, Slatcher RB, Eastwick PW, Scissors LE, Finkel EJ, Pennebaker JW. Language style matching predicts relationship initiation and stability. Psychol Sci 2011 Jan 13;22(1):39-44. [doi: 10.1177/0956797610392928] [Medline: 21149854]
116. Ramseyer FT, Tschacher W. Synchrony in dyadic psychotherapy sessions. In: Rossler OE, Vrobel S, Marks-Tarlow T, editors. Simultaneity: Temporal Structures And Observer Perspectives. Singapore, Singapore: World Scientific Publishing Company; 2008:329-347.
117. Mogan R, Fischer R, Bulbulia JA. To be in synchrony or not? A meta-analysis of synchrony's effects on behavior, perception, cognition and affect. J Exp Soc Psychol 2017 Sep;72:13-20. [doi: 10.1016/j.jesp.2017.03.009]
118. Tschacher W, Rees GM, Ramseyer F. Nonverbal synchrony and affect in dyadic interactions. Front Psychol 2014 Nov 24;5:1323 [FREE Full text] [doi: 10.3389/fpsyg.2014.01323] [Medline: 25505435]
119. Matthias MS, Kukla M, McGuire AB, Bair MJ. How do patients with chronic pain benefit from a peer-supported pain self-management intervention? A qualitative investigation. Pain Med 2016 Dec 08;17(12):2247-2255 [FREE Full text] [doi: 10.1093/pm/pnw138] [Medline: 28025359]
120. Bandura A. Social Learning Theory. Englewood Cliffs, NJ: Prentice-Hall; 1977.
121. Areli E, Godfrey HK, Perry MA, Hempel D, Saipe B, Grainger R, et al. 'I think there is nothing . . . that is really comprehensive': healthcare professionals' views on recommending online resources for pain self-management. Br J Pain 2021 Nov;15(4):429-440 [FREE Full text] [doi: 10.1177/2049463720978264] [Medline: 34840791]
122. Devan H, Godfrey HK, Perry MA, Hempel D, Saipe B, Hale L, et al. Current practices of health care providers in recommending online resources for chronic pain self-management. J Pain Res 2019 Aug 12;12:2457-2472 [FREE Full text] [doi: 10.2147/JPR.S206539] [Medline: 31496788]
123. Sannon S, Murnane E, Bazarova N, Gay G. "I was really, really nervous posting it": communicating about invisible chronic illnesses across social media platforms. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK. [doi: 10.1145/3290605.3300583]
124. Pascual-Ferrá P, Alperstein N, Barnett DJ, Rimal RN. Toxicity and verbal aggression on social media: polarized discourse on wearing face masks during the COVID-19 pandemic. Big Data Soc 2021 Jun 10;8(1). [doi: 10.1177/20539517211023533]
125. Zaheri S, Leath J, Stroud D. Toxic comment classification. SMU Data Sci Rev 2020;3(1):13.
126. Stevens HR, Acic I, Rhea S. Natural language processing insight into LGBTQ+ youth mental health during the COVID-19 pandemic: longitudinal content analysis of anxiety-provoking topics and trends in emotion in LGBTeens Microcommunity Subreddit. JMIR Public Health Surveill 2021 Aug 17;7(8):e29029 [FREE Full text] [doi: 10.2196/29029] [Medline: 34402803]
127. Gatchel RJ, Neblett R, Kishino N, Ray CT. Fear-avoidance beliefs and chronic pain. J Orthop Sports Phys Ther 2016 Feb;46(2):38-43. [doi: 10.2519/jospt.2016.0601] [Medline: 26828236]

XSL•FO
RenderX

128. Carstens JK, Shaw WS, Boersma K, Reme SE, Pransky G, Linton SJ. When the wind goes out of the sail - declining recovery expectations in the first weeks of back pain. Eur J Pain 2014 Feb 03;18(2):269-278. [doi: 10.1002/j.1532-2149.2013.00357.x] [Medline: 23824730]
129. Ferro MA, Boyle MH. Longitudinal invariance of measurement and structure of global self-concept: a population-based study examining trajectories among adolescents with and without chronic illness. J Pediatr Psychol 2013 May 05;38(4):425-437. [doi: 10.1093/jpepsy/jss112] [Medline: 23132934]
130. Glette M, Stiles TC, Borchgrevink PC, Landmark T. The natural course of chronic pain in a general population: stability and change in an eight-wave longitudinal study over four years (the HUNT pain study). J Pain 2020 May;21(5-6):689-699 [FREE Full text] [doi: 10.1016/j.jpain.2019.10.008] [Medline: 31698134]
131. Karels CH, Bierma-Zeinstra SM, Burdorf A, Verhagen AP, Nauta AP, Koes BW. Social and psychological factors influenced the course of arm, neck and shoulder complaints. J Clin Epidemiol 2007 Aug;60(8):839-848. [doi: 10.1016/j.jclinepi.2006.11.012] [Medline: 17606181]
132. Vlaeyen JW, Linton SJ. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. Pain 2000 Apr;85(3):317-332 [FREE Full text] [doi: 10.1016/S0304-3959(99)00242-0] [Medline: 10781906]
133. Crombez G, Eccleston C, van Damme S, Vlaeyen JW, Karoly P. Fear-avoidance model of chronic pain: the next generation. Clin J Pain 2012 Jul;28(6):475-483. [doi: 10.1097/AJP.0b013e3182385392] [Medline: 22673479]
134. Zale EL, Ditre JW. Pain-related fear, disability, and the fear-avoidance model of chronic pain. Curr Opin Psychol 2015 Oct 01;5:24-30 [FREE Full text] [doi: 10.1016/j.copsyc.2015.03.014] [Medline: 25844393]
135. Sippel LM, Pietrzak RH, Charney DS, Mayes LC, Southwick SM. How does social support enhance resilience in the trauma-exposed individual? Ecol Soc 2015;20(4):10. [doi: 10.5751/ES-07832-200410]
136. Wilks SE. Resilience amid academic stress: the moderating impact of social support among social work students. Adv Soc Work 2008 Dec 12;9(2):106-125. [doi: 10.18060/51]
137. Gonzalez-Polledo E, Tarr J. The thing about pain: the remaking of illness narratives in chronic pain expressions on social media. New Media Soc 2016 Jul 09;18(8):1455-1472. [doi: 10.1177/1461444814560126]
138. Sendra A, Farré J. Communicating the experience of chronic pain through social media: patients' narrative practices on Instagram. J Commun Healthc 2020 Apr 21;13(1):46-54. [doi: 10.1080/17538068.2020.1752982]
139. Bevers K, Watts L, Kishino ND, Gatchel RJ. The biopsychosocial model of the assessment, prevention, and treatment of chronic pain. US Neurol 2016;12(2):98-104. [doi: 10.17925/USN.2016.12.02.98]
140. Vlaeyen JW, Haslbeck JM, Sjouwerman R, Peters ML. Towards a dynamic account of chronic pain. Pain 2022 Sep 01;163(9):e1038-e1039 [FREE Full text] [doi: 10.1097/j.pain.0000000000002706] [Medline: 35984384]
141. Broderick JE, Junghaenel DU, Schwartz JE. Written emotional expression produces health benefits in fibromyalgia patients. Psychosom Med 2005;67(2):326-334. [doi: 10.1097/01.psy.0000156933.04566.bd] [Medline: 15784801]
142. Gillis ME, Lumley MA, Mosley-Williams A, Leisen JC, Roehrs T. The health effects of at-home written emotional disclosure in fibromyalgia: a randomized trial. Ann Behav Med 2006 Oct;32(2):135-146 [FREE Full text] [doi: 10.1207/s15324796abm3202_11] [Medline: 16972811]
143. Junghaenel DU, Schwartz JE, Broderick JE. Differential efficacy of written emotional disclosure for subgroups of fibromyalgia patients. Brit J Health Psychol 2010 Dec 24;13(1):57-60. [doi: 10.1348/135910707x251162]
144. Nils F, Rimé B. Beyond the myth of venting: social sharing modes determine the benefits of emotional disclosure. Eur J Soc Psychol 2012 Apr 17;42(6):672-681. [doi: 10.1002/ejsp.1880]
145. Zhan J, Xu H, Ren J, Luo J. Is catharsis beneficial or harmful? The psychological intervention effect and potential harm of catharsis. Adv Psychol Sci 2020;28(1):22-32. [doi: 10.3724/SP.J.1042.2020.00022]
146. Prieto-Callejero B, Gómez-Salgado J, Alvarado-Gómez F, Dias A, García-Iglesias JJ, Ruiz-Frutos C. [Systematic review of the reduction of negative emotional effects in emergency and disaster response workers through catharsis techniques]. Arch Prev Riesgos Labor 2020;23(1):52-67. [doi: 10.12961/aprl.2019.22.01.05] [Medline: 31991077]

Abbreviations

FA: fear-avoidance
LIWC: Linguistic Inquiry and Word Count
OHC: online health community
RP: relative phase
SCMP: social communication model of pain
VADER: Valence Aware Dictionary for Sentiment Reasoning
WTC: wavelet coherence
XWT: cross-wavelet transform

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 24.01.23; peer-reviewed by M Merolli, S Stones; comments to author 10.03.24; revised version received 20.05.24; accepted 24.06.24; published 05.09.24.

Please cite as:
Necaise A, Amon MJ
Peer Support for Chronic Pain in Online Health Communities: Quantitative Study on the Dynamics of Social Interactions in a Chronic Pain Forum
J Med Internet Res 2024;26:e45858
URL: https://www.jmir.org/2024/1/e45858
doi:10.2196/45858
PMID:39235845

©Aaron Necaise, Mary Jean Amon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality Measurement of Consumer Health Questions: Content and Language Perspectives

Ashwag Alasmari[1,2], PhD; Lina Zhou[3], PhD

[1]Computer Science Department, King Khalid University, Abha, Saudi Arabia

[2]Center for Artificial Intelligence, King Khalid University, Abha, Saudi Arabia

[3]Department of Business Information Systems and Operations Management, The University of North Carolina at Charlotte, Charlotte, NC, United States

Corresponding Author:
Ashwag Alasmari, PhD
Computer Science Department
King Khalid University
King Khalid Road, Guraiger
Abha, 61421
Saudi Arabia
Phone: 966 2418000
Email: aasmry@kku.edu.sa

Abstract

Background: Health information consumers increasingly rely on question-and-answer (Q&A) communities to address their health concerns. However, the quality of questions posted significantly impacts the likelihood and relevance of received answers.

Objective: This study aims to improve our understanding of the quality of health questions within web-based Q&A communities.

Methods: We develop a novel framework for defining and measuring question quality within web-based health communities, incorporating content- and language-based variables. This framework leverages k-means clustering and establishes automated metrics to assess overall question quality. To validate our framework, we analyze questions related to kidney disease from expert-curated and community-based Q&A platforms. Expert evaluations confirm the validity of our quality construct, while regression analysis helps identify key variables.

Results: High-quality questions were more likely to include demographic and medical information than lower-quality questions ($P<.001$). In contrast, asking questions at the various stages of disease development was less likely to reflect high-quality questions ($P<.001$). Low-quality questions were generally shorter with lengthier sentences than high-quality questions ($P<.01$).

Conclusions: Our findings empower consumers to formulate more effective health information questions, ultimately leading to better engagement and more valuable insights within web-based Q&A communities. Furthermore, our findings provide valuable insights for platform developers and moderators seeking to enhance the quality of user interactions and foster a more trustworthy and informative environment for health information exchange.

(J Med Internet Res 2024;26:e48257) doi:10.2196/48257

KEYWORDS

question quality; quality measurement; health questions; ; information needs; information behavior; information sharing; consumer; health information; health information consumers; quality

Introduction

Background and Motivation

Health information consumers (HICs) are increasingly taking an active role in their health care, turning to a variety of sources for information. Web-based question-and-answer (Q&A) communities are one of these valuable resources, enabling HICs to post and receive answers from fellow community members [1]. Web-based Q&A communities serve as a direct and alternative means of obtaining information compared to search engines, [2].

Articulating health information needs in a concise and understandable question is essential because the relevancy, quality, and nature of the obtained answers are substantially linked to the nature and quality of the question representing the information needs [3,4]. Questions serve as the starting point

XSL•FO
RenderX

in the Q&A setting and the primary driver of what might happen next—who responds, and how quality and relevant obtained answers are. However, HICs may experience significant uncertainty in conceiving their information needs and subsequent difficulty in articulating them [5].

Writing high-quality questions can bring many potential benefits. Well-formed questions attract more high-quality answers than poorly formed questions, as subject-matter experts are more likely to assist users who already put in some effort [6]. The quality of the question can significantly impact the probability of receiving helpful answers [7], which eventually drives the popularity of a Q&A community. Previous studies [8-12] show that the features of the questions and the responsiveness to these questions are correlated. More specifically, there are language determinants of the quantity and quality of received answers in web-based communities and social networks [11]. For instance, stating the information needs in a question format, explicitly scoping the audience, and using only 1 sentence leads to high quality and timely answer [11]. Other studies indicate that a high-quality question should provide more details and examples [12]. Thus, HICs need to formulate their information needs in high-quality health questions to receive timely, relevant, and comprehensive answers.

There are 3 streams of research related to question quality. The first one is focused on the characteristics of answers with the assumption that high-quality questions generate answers, and the most straightforward measure of a good-quality question is whether it has received an acceptable answer [13]. In contrast, a poor-quality question likely fails to receive any answers. Some studies have further examined the quality of answers in Q&A communities [14]. This stream of research relies on the outcomes of Q&A interaction, which may not be available. Even if they are available, a lag is expected between the times when an HIC posts a question and when an answer is provided. The second stream of research examines the characteristics of the askers, such as reputation or expertise [15], on the basis that community recognition reflects the ability to construct appropriate and valuable questions. However, the asker's information is generally not publicly available due to the platform's privacy policy or the HIC's privacy concerns. The third stream is on analyzing question characteristics, such as the topics, number of views, and content and language variables that correlate with high-quality answers [16]. Examining question quality based on the question's intrinsic features holds promise for overcoming the limitations of the first 2 research streams. However, limited research has examined question quality and its measurement, particularly within the context of health-related questions.

This study seeks to enhance our understanding of the intrinsic variables influencing health question quality posed in web-based Q&A communities from the content and language perspectives. Specifically, we address the following primary research question: What are the content and language variables of high-quality questions in web-based health Q&A communities? Answering the question can not only assist HIC in soliciting answers from peer community members and actively engaging themselves in community discussions but also promote the success of web-based Q&A communities.

Our Proposed Constructs for Health Question Quality

In this section, we first introduce our proposed construct of the health question quality. We then develop quantitative measures for the construct using a clustering approach and finally validate the measurements through human assessment. Figure 1 illustrates the architecture of health question quality measurement, including data collection, feature extraction, clustering, and question quality validation.

Figure 1. Architecture of health question quality assessment.

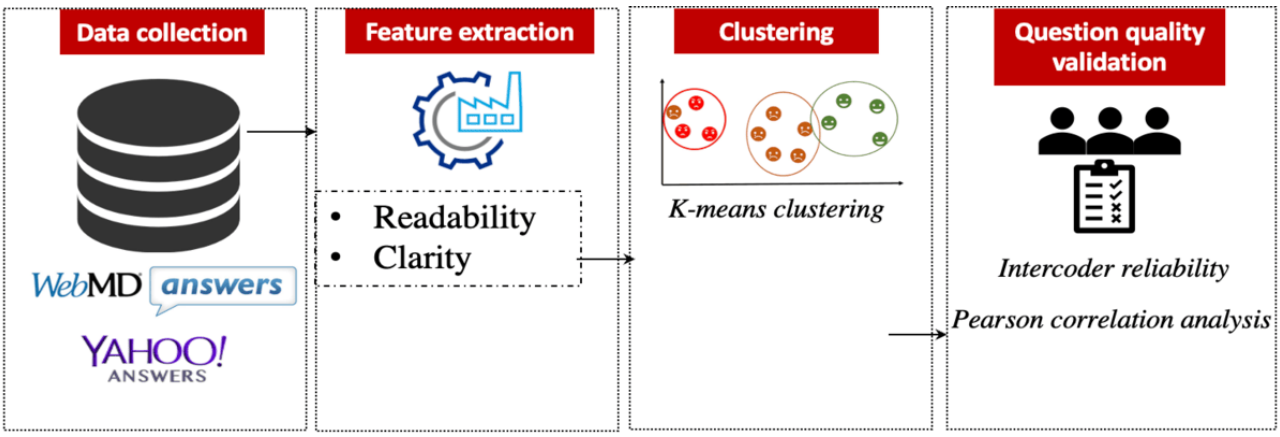

Construct Design

We introduce a conceptualization of question quality that centers on 2 key indicators: readability and clarity. We chose these 2 indicators because they represent the variety of measures that impact how people perceive questions. Quality questions should be readable in order to be accessible to a large and diverse audience and be clear about what is expected to be answered. To the best of our knowledge, these indicators have not yet been explored in the context of health-related questions, let alone their collective impacts. These proposed indicators will enable us to uncover fundamental aspects of question quality that would otherwise be missed through solely examining answers. Therefore, our proposed construct will facilitate the timely assessment of question quality before receiving any answer.

XSL•FO
RenderX

Readability relates to how well a question is written, how understandable it is, and the extent to which it is free of unnecessary complexity. The concept of readability is often studied in natural language processing and used as an indicator of article content quality in the digital library [17]. For web-based Q&A communities, we assume that most community members communicate at an average reading level [18]. Accordingly, questions with a high reading level tend to attract fewer potential answers because fewer community members would comprehend the question [18]. Clarity assesses how easily understood a question is and how readily it can be answered. This directly correlates to a higher probability of receiving high-quality answers.

Construct Measurement

Overview

We propose measures for each of the 2 indicators of health question quality construct.

Readability

A variety of automated measures, such as the Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG (Simple Measure of Gobbledygook) Index, Coleman-Liau Index, and Automated Readability Index, have been developed to assess text readability [19]. Among them, the Flesch-Kincaid Reading Ease metric (equation 1) is the most widely used [20] and has been adopted in question-answering in various significant contexts [8]. Similarly, we have incorporated it into our study.

The Flesch-Kincaid readability score uses the average length of sentences and the average number of syllables per word to calculate reading ease. The output is on a scale, which typically ranges from 1 to 100, with lower scores indicating a more difficult text to read. As laypersons typically compose health questions without a medical background, these questions are nonscientific and share similarities with informal discourses encountered in a high-school education setting. Hence, the Flesch-Kincaid Reading Ease method is applicable in this context.

Clarity

Clearly expressed questions are more readily understood by other HICs, facilitating relevant and timely answers. For a question to be considered explicit, it should include a minimum of 1 interrogative word [8]. Interrogative words (eg, who, what, where, when, why, and how) can indicate the clarity of health questions. Following the work of Kitzie et al [18], we developed a scale of clarity based on the number of interrogative words normalized by the question length. Specifically, a higher ratio of interrogative words to the total number of words suggests a clearer formulation of the question. We illustrate the measure of clarity with 2 sample questions as the following:

- Q1: “What's the first thing doctors do when a child is diagnosed with chronic kidney disease? And what happens next. Like when you first go in what do they do? What do they tell you? what happens?”
- Q2: “I have been early diagnosed with stage 3 kidney failure. Let me know thoughts on treatments? The condition stems from many years of high blood pressure. I am 55 and hope this doesn't lead to dialysis.”

Question Q1 contains multiple occurrences of interrogative words, such as what and when, accounting for 19.44% of words in the question. They clearly indicate the user’s informational needs. In contrast, question Q2 provides more background information about the HIC’s condition but does not explicitly state his or her question with any interrogative word, which results in a ratio of zero for clarity. Thus, the information needs expressed in Q2 are not as clear as those in Q1. We used linguistic inquiry and word count (LIWC) [21] to extract interrogative words.

Overall Question Measurement

One straightforward method to derive an overall quality measure is by averaging the measures of individual indicators, namely readability and clarity. However, assigning equal weight to all indicators may overlook the nuanced differences among individual constructs within questions. In this study, we propose a clustering analysis approach complemented with a human assessment to validate the measurement. This study used the k-means clustering technique [22] to group health questions into a small number of distinct clusters based on question similarities. The parameter k was empirically determined based on the number of clusters’ sets and the proportion of betweenness. We selected the elbow method to determine the number of clusters, which runs the k-means clustering algorithm on the data set for a range of values and then calculate the sum of squared error for each value [22]. Accordingly, we set k to 3. To validate the clustering results, we used the method of human validation.

Human Validation

Three human judges were recruited to manually assess question quality. Each judge had extensive experience searching for health information on the web and actively participating in web-based Q&A communities. All were English speakers with direct or indirect experience in disease management. We opted not to include human judges with medical backgrounds because web-based Q&A communities tend to attract people without medical expertise, as observed in prior studies [23-25].

We performed stratified sampling by randomly selecting 10% of the questions from each quality cluster. The order of the questions was randomized for each of the human judges. The judges were asked to rate each question independently at 1 of the 3 quality levels: high, average, and low.

The final quality rating of the health questions was determined based on a joint discussion of the 3 judges. Intercoder reliability was assessed with Cohen κ statistics, following the convention of κ>0.70 indicating “substantial” agreement [26]. In addition, we also performed a Pearson correlation analysis between the ratings of the human judges and the results of k-means clustering using the following coding scheme: high=1, average=2, and low=3. The results show that r>0.5, which indicates a very high correlation.

XSL•FO
RenderX

Methods

Overview

To investigate the content and language characteristics influencing the quality of health-related questions in web-based Q&A communities, we begin by describing the data set. Subsequently, we introduce content- and language-based variables and highlight the analysis methods used.

Data Collection

For this study, we collected health questions from Yahoo! Answers, a community-based Q&A site, and WebMD, an expert-based Q&A site. WebMD is one of the most influential web-based health sites, and users were able to post questions for certified health experts to answer in the Q&A section, covering more than 900 health topics. We collected all posts made from August 2008 until the closure of WebMD Answers in 2018. Yahoo! Answers features health as one of the top-level categories. Since its release in December 2005, Yahoo! Answers has become a popular internet reference site worldwide and the most frequented Q&A community in the United States. As of June 2019, the site ranked ninth in global internet traffic and engagement over the past 3 months and seventh in the United States.

Following the prior work [27-30], we treated postings in a question section of a Q&A site as a question regardless of whether they ended with a question mark or not. To investigate question quality in the context of web-based health communities, we focused on kidney disease. This condition, characterized by its prevalence and complex management, provides a valuable case study. Chronic kidney disease (CKD) is a noncommunicable health disease, often accompanied by multimorbidity (frequent co-occurrences with other diseases) [31]. According to the National Kidney Foundation, CKD affects 14% of adults in the United States [32]. In these platforms, patients navigate and manage complex health information, making this study particularly relevant. We screened for queries based on preselected key terms that directly refer to kidney conditions or early signs of CKD [33]. Using an application programming interface, we sampled 400 random questions from Yahoo! Answers and another 400 questions from WebMD relating to the above kidney-related vital terms. In other words, the questions included in our data set must be related to human kidney diseases. We eliminated repeated questions using a combination of user identification and question similarity techniques. In addition, we manually filtered irrelevant questions, including advertisements, those unrelated to kidney-related topics, those not involving human patients, or those associated with student research. For each of the relevant questions, we extracted the title, description, date of question, topic, and number of answers.

Content-Based Variables

We contextualized the health questions by drawing on relevant findings from a previous study on health information seeking [27]. The 2 main areas of interest in our analyses included health stages of disease development and information shared (ie, demographic and medical information). Further, 2 coders, with medical backgrounds, manually coded the stage of disease development and the type of information shared. To ensure coding consistency, a well-defined coding guideline was followed. Additionally, any discrepancies between the coders' initial assessments were resolved through decisiveness by a third coder to minimize subjective bias and ensure the accuracy of the coded data. Variables were coded as binary variables (1 indicating presence and 0 indicating absence) of specific information.

- Health stages of disease development: We followed 2 systems in representing the stages of disease development: stages of health questioning [27] and chronic disease stages [34]
  - Stages of health questioning: Managing a disease or condition is an ongoing process, and HIC at different stages often has different levels of information needs [29]. Additionally, HICs may display other information-seeking behaviors based on the nature and extent of their needs. To identify at which stage of their disease HIC asked questions (stage of health questioning), we used a model proposed by Zhang [27] consisting of eight stages: (1) being healthy—at this stage, the questions are related to disease prevention and health promotion, (2) self-diagnosed as being ill, (3) before having a medical test or checkup, (4) after being diagnosed or self-diagnosed as ill, (5) before treatment (such as surgery or medications), (6) during treatment (including medications or exercise); (7) after treatment, and (8) when the disease becomes chronic or reaches the terminal stage. To address data sparsity, we merged stages (2-4) as stage 2 and stages (5-7) as stage 3.
  - Chronic disease stages: We also drew on the stages of chronic illness to understand the questions of HICs who are chronically ill. To this end, we used the Corbin Chronic Illness Trajectory Framework because it includes all stages of chronic disease [34] and is used by clinicians in nursing care and chronic illness management [35]. This trajectory framework is built around the idea that chronic conditions vary and change over time, which consists of nine stages [34]: (1) pretrajectory: before the disease onset; (2) trajectory onset: the appearance of symptoms and diagnosis; (3) stable: condition and symptoms are under control, everyday life is unaffected, illness management is home-centered, and hospitalization not required; (4) unstable: condition and symptoms are not under control, everyday life is disrupted, but care remains centered in the home; (5) acute: symptoms or complications require hospitalization or other measures, everyday life activities are cut back or severely curtailed; (6) crisis: a life-threatening situation that requires emergency care, everyday life is placed on hold; (7) comeback: a return to everyday life activities, possibly with changed ability for everyday life activities; (8) downward: decline associated with increased disability and trouble controlling symptoms, requires adaptation in everyday life activities; and (9) dying: death of the patient. To address data sparsity, we grouped the nine stages into

XSL•FO
RenderX

2: stable and nonstable stages, with the latter covering stages 4 through 9 and the former covering the rest.

- Type of information shared: When HICs communicate their information needs by asking questions, they often include some demographic and medical information related to diagnosis, treatment, and prevention that represent their understanding of their disease [27]. Demographic information includes age, gender, ethnicity, weight, location, and profession. Medical information may pertain to diagnosis, treatment, or prevention. Diagnostic medical information includes symptoms, medical tests, personal and family medical history, etc. Treatment and prevention medical information include treatment options, duration of treatment, lifestyle, prescribed medications, length of hospitalization, etc.

Language-Based Variables

We leveraged LIWC [21] to extract language-based variables due to the tool's ability to measure language in multiple dimensions. The value of each dimension is calculated based on the percentage of words related to this particular dimension [21]. For this research, we selected the following variables that can be used to characterize the language and writing style of a health question:

- The total number of words and the number of words per phrase.
- Total number of pronouns: these include personal pronouns and impersonal pronouns.
- Social processes: this includes references to family or friends, as well as biological sex (male and female).
- Time orientation: this includes references to the past, present, and future, which reflect a general time orientation.
- Biological process: this includes references to the body and health.
- Affect: this includes 5-word variables: overall affect, positive and negative emotion, anxiety, anger, and sadness.

Analysis Methods

A multinomial logistic regression model was used to analyze the content and language variables determining the health question quality. The dependent variable was the overall question quality, and the independent variables were content- and language-based.

Ethical Considerations

The study's use of health-related questions from a public web-based platform raises potential privacy concerns. To address this, we implemented several measures to protect user privacy. While the data was sourced from a publicly accessible platform, we strictly adhered to the platform's terms of use (eg, Yahoo's agreement) for data collection. Due to copyright restrictions, we were unable to directly share the original questions but instead provided unique identifiers. These questions remain publicly accessible with appropriate platform permissions. To further safeguard privacy, all personally identifiable information was carefully removed from the data set. Recognizing the limitations of a small data set focused on a single disease, we implemented measures to mitigate potential biases in data collection and analysis, such as random sampling and rigorous coding guidelines.

Results

Descriptive Statistics

Table 1 reports the descriptive statistics and the statistical test for the content-based variables. The table shows that the proportion of medical information included in WebMD questions is significantly higher ($P<.01$) than that in Yahoo! Answers questions. More specifically, medical treatment and medical diagnosis information in WebMD are significantly higher ($P<.05$) than that in Yahoo! Answers. In contrast, demographic information is higher in Yahoo! Answers ($P<.01$) than in WebMD. The comparison results of the stages of general health questioning show that the proportions of all stages except for stage 1 are significantly higher in WebMD ($P<.01$) as compared to Yahoo! Answers. In contrast, the proportions of questions regarding the 3 chronic disease stages are higher in Yahoo! Answers ($P<.01$) than in WebMD.

Table 2 reports the descriptive statistics and the statistical test for language-based variables. It shows that the average word count per question, the total number of pronouns, personal pronouns, and affect in WebMD is significantly lower ($P<.01$) than in Yahoo! Answers. In contrast, the social process is lower in questions from Yahoo! Answers ($P<.01$) than those from WebMD. However, no statistically significant differences were detected in word per sentence and time orientation between the Yahoo! Answers and WebMD platforms ($P=.96$).

Based on the results of k-means clustering, the coders manually analyzed the indicators of the question quality within each of the 3 clusters and provided a quality rating for each cluster separately, ranging from low to high quality. The statistical details of each cluster are provided in Table 3. We discussed the method for determining the number of clusters in the second section (our proposed constructs for health question quality).

- Cluster 1 (high quality) comprised 101 questions with readability ranging between 39.01% and 80.5% and clarity of 33.33% or above.
- Cluster 2 (average quality) had 169 questions with readability ranging between 0.7% and 49.3% and clarity of 12.5% or lower.
- Cluster 3 (low quality) contained 354 questions with readability ranging between 1.5% and 14.7% and clarity being 9.09% or lower. While the clarity range of the current cluster is somehow similar to that of cluster 2, the latter is rated lower than the former on a range of readability.

XSL•FO
RenderX

Table 1. Descriptive statistics for content-based variables.

	Yahoo! Answers, n (%)	WebMD, n (%)	*P* value
Types of information shared			
Medical information	147 (46.52)	182 (59.09)	<.001***
Demographic information	100 (31.65)	58 (18.83)	<.001[a]
Medical diagnosis information	111 (35.13)	154 (50)	<.001[a]
Medical treatment information	78 (24.68)	96 (31.17)	.035[b]
Stages of health questioning			
1	12 (3.8)	11 (3.57)	.88
2	55 (17.41)	87 (28.25)	<.001[c]
3	4 (1.27)	36 (11.69)	<.001[c]
4	191 (60.44)	64 (20.78)	<.001[c]
5	54 (17.09)	110 (35.71)	<.001[c]
Chronic disease stage			
1	119 (37.66)	40 (12.99)	<.001[c]
2	50 (15.82)	18 (5.84)	<.001[c]
3	147 (46.52)	250 (81.17)	<.001[c]

[a]*P*<.001.
[b]*P*<.05.
[c]P<.01.

Table 2. Descriptive statistics for language-based variables.

	WebMD, mean (SD)	Yahoo! Answers, mean (SD)	*P* value
Word count	42.1 (36.9)	60.2 (59.4)	<.001[a]
Word or sentence	13.1 (7.9)	13.1 (7.1)	.96
Total pronouns	12.1 (7.1)	14.3 (7.2)	<.001[a]
Personal pronouns	6.6 (6.3)	7.9 (5.6)	.003[a]
Affect	1.0 (2.5)	1.5 (2.2)	.01[b]
Social process	3.3 (5.1)	6.6 (6.0)	<.001[a]
Time orientation	3.0 (3.8)	2.6 (3.1)	.08

[a]*P*<.001.
[b]*P*<.05.

XSL•FO
RenderX

Table 3. Descriptive statistics of quality indicators in the question clusters of different quality levels.

Cluster and indicators	Mean (SD)	Minimum	Maximum
1 (high quality)			
Readability	9.88 (8.95)	39.01	80.5
Clarity	2.28 (2.35)	9.09	33.33
2 (average quality)			
Readability	9.6 (6.9)	0.7	49.3
Clarity	1.85 (2.72)	0.0	12.5
3 (low quality)			
Readability	6.67 (2.62)	1.5	14.7
Clarity	15.35 (5.4)	0	9.09

Regression Analysis Results

We chose high-quality questions as the reference cluster because it is the expectation of all health questions. Accordingly, the regression coefficients indicate which independent variables significantly discriminate high-quality questions from low- and average-quality questions, respectively.

Tables 4 and 5 provide the coefficient estimates for all outcome comparisons, along with the SE for each independent variable. We found statistically significant differences between the reference quality cluster and the other clusters in some of the stages of health questioning, most types of shared information, and the language variables of the health questions. However, there is no evidence that expressions of anxiety differ among the different quality clusters of the health questions.

Further, 3 of the stages of questioning (before or after diagnosis, when chronically ill, and when in an unstable condition) were strong predictors of high quality among all comparisons with the reference cluster (high-quality questions; $P<.001$; Table 4). This reached statistical significance only when comparing high-quality to low-quality questions for unstable conditions ($b=0.87$, SD 0.34, $P<.005$). Thus, questions relating to diagnosis, or chronic or unstable conditions are less likely to be of high quality. Questions relating to preventative reasons, treatment, and stable stages of chronic disease did not significantly affect the prediction of question quality ($P>0.1$).

Demographic information and diagnostic medical information are significant predictors across the 2 comparisons ($P<.001$). High-quality questions are more likely to include more demographic information than average quality ($b=-0.84$, SD 0.21, $P<.001$) and low quality ($b=-3.39$, SD 1.02, $P<.001$). In addition, high-quality questions are more likely to include more diagnostic medical information than low-quality ($b=-2.96$, SD 0.73, $P<.001$), but least compared to average quality ($b=1.08$, SD 0.19, $P<.001$). Finally, high-quality questions are more likely to include more treatment and prevention information than low-quality questions ($b=-3.20$, SD 1.02, $P<.001$).

Regarding language variables of health questions (Table 5), we found that the number of words ($P<.001$) and the number of words per sentence ($P<.001$) tend to differ significantly between the 2 average and high-quality questions. Longer questions are more likely to be classified as high-quality ($b=-0.01$, SD 0.003, $P<.001$) than low-quality questions. However, high-quality questions are less likely to be associated with a higher word count per sentence than average-quality questions ($b=0.04$, SD 0.01, $P<.001$). That means average-quality questions tend to be overall shorter, but with more words per sentence than high-quality questions.

Regarding the effect and measures of emotions, a statistically significant difference was found between low or average and high-quality questions. High-quality questions were more likely to be effective compared to low-quality ($b=-2.12$, SD 0.03, $P<.001$), and yet less likely to include positive ($b=2.05$, SD 0.04, $P<.001$) or negative ($b=2.08$, SD 0.03, $P<.001$) emotions than low-quality questions (Table 5). The expression of anger was also associated with high-quality questions ($b=-0.07$, SD 0, $P<.001$). In addition, the expression of sadness was associated with high-quality questions ($b=-0.25$, SD 0.06, $P<.001$). However, there is no evidence that expressions of anxiety in health questions differ among the different quality clusters.

Questions that focused on the past were more likely to be high quality compared to low quality ($b=-0.30$, SD 0.14, $P<.001$). In comparison to low-quality questions, high-quality questions were more likely to focus on the future ($b=-0.16$, SD 0.06, $P<.001$). Questions focusing on the present were equally likely to be of any quality.

In addition, we found that among the social processes, only mentions of friends ($b=-1.74$, SD 0, $P<.001$) or male references ($b=-5.85$, SD 0.00, $P<.001$) tend to be strong predictors of high-quality questions compared to low-quality ones. No other differences were detected between any of the other social processes (including family and female references) and other pairs of comparisons ($P>.05$).

XSL•FO
RenderX

Table 4. Multinomial logistic regression analysis of the effect of content variables on question quality.[a]

Question and explanatory variable	Average versus high, coefficient (SE)	Low versus high, coefficient (SE)
Stages of health questioning		
Preventative reasons	0.82 (0.51)	–1.24 (0.83)
Before or after diagnosis	2.83[b] (0.78)	0.18 (1.33)
Before, after, or during treatment	–0.64 (0.5)	–0.8 (0.69)
Chronic disease stages		
When chronic	0.54 (0.53)	2.08[b] (0.68)
Stable	–0.41 (0.4)	0.26 (0.52)
Unstable	1.16[b] (0.24)	0.87[c] (0.34)
Demographic information	–0.84[b] (0.21)	–3.39[b] (1.02)
Medical information		
Diagnostic information	1.08[b] (0.19)	–2.96[b] (0.73)
Treatment and prevention	0.34[d] (0.2)	–3.2[b] (1.02)

[a]High quality (cluster 1) serving as a reference cluster.
[b]$P<.001$.
[c]$P<.05$.
[d]$P<.10$.

XSL•FO
RenderX

Table 5. Multinomial logistic regression analysis of the effect of language variables on question quality.[a]

Explanatory variable	Average vs high, coefficient (SE)	Very low vs high, coefficient (SE)
Word count	–0.01[b] (0.003)	–0.14[b] (0.03)
Word or sentence	0.04[b] (0.01)s	0.22[c] (0.08)
Total pronouns	–4.30[b] (0.01)	7.54[b] (0.01)
Personal pronouns	4.254[b] (0.01)	–7.53[b] (0.02)
Impersonal pronouns	4.160[b] (0.02)	–7.36[b] (0.02)
Social processes	–0.11[b] (0.03)	–0.07[c] (0.03)
Family	0.02 (0.09)	0.00 (0.28)
Friend	–0.29 (0.22)	–1.74[b] (0)
Female references	0.05 (0.06)	–0.04 (0.25)
Male references	0.11[c] (0.053)	–5.85[b] (0)
Past focus	0.11[b] (0.03)	–0.30[c] (0.14)
Present focus	0.01 (0.02)	0.02 (0.03)
Future focus	–0.16[c] (0.06)	–0.06 (0.09)
Body	0.04[d] (0.02)	0.024 (0.03)
Health	–0.11[b] (0.02)	–0.06[b] (0.02)
Affect	0.37 (0.26)	–2.12[b] (0.03)
Positive emotion	–0.37 (0.26)	2.05[b] (0.04)
Negative emotion	–0.33 (0.26)	2.08[b] (0.03)
Anxiety	–0.07 (0.00)	0.12 (0.11)
Anger	0.09 (0.15)	–0.71[b] (0.00)
Sadness	–0.25[b] (0.06)	–0.12[d] (0.07)

[a]High quality (cluster 1) serving as a reference cluster.
[b]$P<.001$.
[c]$P<.05$.
[d]$P<.10$.

Discussion

Principal Results

Formulating high-quality questions can potentially bring many benefits, directly impacting the relevancy, quality, and nature of the information acquired. Advancing HICs' understanding of question content will contribute significantly to identifying quality questions and facilitate HICs in formulating their questions to solicit better answers. This study highlights several content and language variables of high-quality questions. For example, asking questions at various stages of disease development is more likely to be associated with lower-quality questions. On the other hand, high-quality questions are more likely to include demographic and medical information than lower-level quality questions. These results are consistent with a previous study that reported that the chances of a question being answered are higher when background and personal information about the health situation are provided [8]. Health experts will provide relevant and high-quality information when the HIC includes more contextual information about them, such as age, gender, and other background information about the health condition in question.

While high-quality questions conveyed more information using shorter sentences, low-quality questions were shorter overall but contained lengthier sentences. This suggests that clear and concise language expressing specific information needs is crucial for perceived question quality. Complex questions, often with long sentences, may require excessive effort from potential respondents, hindering their willingness to answer [18]. This aligns with our observation that people with more chronic conditions (potentially having more experience formulating clear questions) tend to ask higher-quality questions with shorter sentences.

The expression of positive or negative emotion was more likely to be found in low than high-quality questions, except for anger and sadness which was more likely to be seen in high-quality

XSL•FO
RenderX

questions. One possible explanation is that HICs were seeking emotional support in low-quality questions and hence disclosed their emotions more readily [36] than those who posted high-quality questions and were seeking informational help.

Our analysis revealed significant discrepancies in the descriptive statistics between the WebMD Answers and Yahoo! Answers data sets. These differences likely stem from several factors. The platforms themselves cater to distinct audiences. WebMD Answers focused on health information, attracting users with specific medical questions and potentially greater health knowledge. Conversely, Yahoo! Answers served a broader audience with diverse interests, resulting in a wider range of question complexity and potentially lower health expertise. Despite these disparities, we believe that merging the data sets offers valuable insights. The combined data set encompasses a broader range of health-related questions and user experiences. This allows us to explore a more comprehensive picture of health information-seeking behavior, potentially revealing patterns not evident within a single platform.

Research Contribution

One of the main contributions of this study lies in quantifying health question quality. Although other quality measures already exist, such as receiving satisfactory answers and subjective judgment of quality [9], improving the interaction between users in health-related Q&A communities is of paramount importance. Any further insight into improving the presentation and expression of HIC needs is welcomed.

Unlike previous studies that focus on either the answers or the askers' profiles to measure the question quality in web-based Q&A communities, this study defined question quality as objective measures of the questions themselves based on their readability and clarity. More specifically, this research explores the idea that questions may share similar quality features. Rather than scoring each question on predefined measures, it used k-means clustering to facilitate the analysis and divided question quality into 3 clusters. Based on the means of each question quality measure, we identified high, average, and low-quality clusters. The results of the clustering process and the subsequent validation with human assessment support the idea that grouping questions is a viable analysis method because of shared characteristics.

Further, 1 final contribution of this study is that the proposed framework for measuring question quality is independent of the application context. The 2 measures, readability and clarity can be applied or extended to other types of web-based Q&A communities beyond health care.

Research Implications

This new question of quality measurement has significant research implications. First, compared to alternative quality measures (eg, receiving satisfactory answers or based on the asker's profiles [8,13]), the textual features of questions can provide a timely assessment of quality based on the characteristics of the questions themselves without relying on answers. This is particularly important for researchers aiming to build models that automatically predict or evaluate the quality of large volumes of health questions in web-based Q&A communities. Second, this study establishes a foundation for developing a set of objective quality criteria that can be used to create classifiers for identifying high-quality questions in the future. Such work will require the identification of those content and language variables most relevant to the specific health subject.

This study also offers practical suggestions for improving the effectiveness of web-based Q&A communities. The findings of this study may guide designers and developers of Q&A systems to design community support systems that encourage user contributions and control quality. Based on the measurement methods of question quality introduced in this work, Q&A communities could develop automated systems for prescreening the quality of health questions before HICs submit them to the community and facilitate the asker in formulating high-quality questions. Such approaches could include using an iterative feedback system, query expansion, or syntax checking correction. Quality questions are likely to lead to quality answers, critical to user engagement in those communities.

Iterative feedback mechanisms, such as query reformulation in interactive information retrieval [37], have demonstrated their effectiveness in helping users refine information needs based on initial relevance judgments and subsequent system feedback. Within the Q&A context, archived answers to similar questions could be presented to the asker, and the asker could pick the questions that are most relevant to their question to check out their answers. The asker may not only learn from others how to write questions but also find answers to their questions directly. In addition, by understanding the textual features of high-quality questions, consumers can tailor their questions to elicit optimal responses. Our research provides consumers with concrete steps to improve their questions. Highlighting details such as demographics, medical history, and symptoms clearly and concisely can lead to more targeted and relevant answers. This empowers consumers to make informed health care decisions based on reliable information.

One long-term implication of this study lies in helping health care professionals or artificial intelligence systems improve their performance by adding a layer of quality assurance before the question is processed. Already, artificial intelligence is finding its uses in health care such as diagnosis, personalized care, and treatment [38], and it relies on quality data to perform well. In addition, health care providers will also benefit from receiving high-quality questions that contain all relevant information at the outset when providing web-based health advice or telemedicine due to their limited availability. This is particularly important because health professionals rely more on the patient's information than the examination findings when managing a patient [39].

Limitations

One of the main limitations is the small number of health questions used in our study. The increasing diversity of analyzed health questions will strengthen the generalization of our conclusion. Future studies may include more health questions and use an automated approach to analyze the content variables of health questions. In addition to not fully exploring platform

XSL•FO
RenderX

effects, we examined health questions from only 2 distinct types of web-based Q&A communities. Health questions are posted on diverse web-based platforms with varying features and functionalities. This limited sample size and lack of platform-specific analysis restrict the generalizability of our findings to other types of Q&A communities. Future research could expand the scope by including a broader range of Q&A platforms and investigating how platform-specific factors such as guidelines, moderation, and user demographics interact with question characteristics to influence response dynamics across diverse web-based communities. This comprehensive approach would provide a deeper understanding of how the web-based environment shapes user behavior and information exchange in different Q&A contexts. Furthermore, this study establishes a strong foundation for investigating question quality assessment in the context of kidney disease. Future research can then expand on this work by conducting empirical validations in other chronic diseases.

In addition, the validation of this new quality measurement was based on human judges who have experience in finding web-based information but not necessarily in answering other HIC questions. Ideally, further validation with other quality measures that have already been evaluated will add another layer of validity to this study. For example, answers can reveal additional insights about the quality of the questions, and future exploration of textual quality indicators should also consider the quality of the answers to the questions. Our new quality measures could potentially be used to analyze the characteristics of high-profile HIC queries in web-based forums. While our current focus is on measures such as readability and clarity, future research can explore prompt engineering, showing how users craft effective questions for health care information systems. The rise of platforms such as ChatGPT highlights this need, as our findings provide a foundation for equipping users with the skills to formulate high-quality prompts regardless of the platform. Longitudinal studies can then assess the real-world impact on information access and communication success in AI-powered health care.

Conclusions

The increasing popularity and usage of web-based Q&A communities for seeking health information calls for an investigation into factors influencing content quality and ways to improve the quality of questions. By identifying the content and language variables of health questions that affect the quality, this research contributes to a deeper understanding of how users can formulate effective questions that receive accurate and relevant responses. This research would offer meaningful suggestions for platforms' managers and users as well. This knowledge can empower consumers to become more active participants in their web-based health information–seeking journeys. By formulating concise, specific, and focused questions, people can maximize the effectiveness of web-based Q&A platforms and increase the likelihood of receiving high-quality answers from health care professionals and other informed users. Furthermore, our findings provide valuable insights for platform managers seeking to enhance the quality of user interactions within their communities. By promoting best practices for question formulation through educational resources and user guidance, web-based Q&A platforms can foster a more efficient and trustworthy environment for the exchange of health information. This research paves the way for future investigations into the dynamic interplay between user behavior, platform characteristics, and response quality within web-based health communities. Through continued exploration, we can work toward optimizing the web-based environment for effective health information access and use.

Acknowledgments

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small group research (RGP1/337/45). The authors also thank the University of North Carolina at Charlotte for their support.

Conflicts of Interest

None declared.

References

1. Shah C, Oh S, Oh JS. Research agenda for social Q&A. Library & Information Science Research 2009 Dec;31(4):205-209. [doi: 10.1016/j.lisr.2009.07.006]
2. Liu Q, Agichtein E, Dror G, Maarek Y, Szpektor I. When web search fails, searchers become askers. 2012 Presented at: Proc 35th Int ACM SIGIR Conf Res Dev Inf Retr; August 12-16, 2012; Portland Oregon USA p. 801-810.
3. Barry CL. User-defined relevance criteria: an exploratory study. J Am Soc Inf Sci 1994;45(3):149-159. [doi: 10.1002/(sici)1097-4571(199404)45:3<149::aid-asi5>3.0.co;2-j]
4. Taylor RS. Question-negotiation and information seeking in libraries. CRL 1968;29(3):178-194. [doi: 10.5860/crl_29_03_178]
5. Belkin NJ. Anomalous states of knowledge as a basis for information retrieval. Can J Inf Libr Sci 1980;5:133-143 [FREE Full text]
6. Ameri F, Keeling K, Salehnejad R. You get what you pay for on health care question and answer platforms: nonparticipant observational study. J Med Internet Res 2020;22(1):e13534 [FREE Full text] [doi: 10.2196/13534] [Medline: 31939741]
7. Baltadzhieva A, Chrupa\la G. Question quality in community question answering forums: a survey. ACM SIGKDD Explor Newsl 2015;17(1):8-13. [doi: 10.1145/2830544.2830547]

XSL•FO
RenderX

8. Shah C, Radford ML, Connaway LS, Choi E, Kitzie V. "How much change do you get from 40$?"- Analyzing and addressing failed questions on social Q&A. Proc of Assoc for Info 2013 Jan 24;49(1):1-10. [doi: 10.1002/meet.14504901119]
9. Treude C, Barzilay O, Storey MA. How do programmers ask and answer questions on the web? 2011 Presented at: 33rd Int Conf Softw Eng; May 21-28, 2011; Waikiki, Honolulu HI USA p. 804-807. [doi: 10.1145/1985793.1985907]
10. Li B, Jin T, Lyu M, King I, Mak B. Analyzing and predicting question quality in community question answering services. USA: Association for Computing Machinery; 2012 Presented at: Proc 21st Int Conf World Wide Web; April 16-20, 2012; New York, NY p. 775-782. [doi: 10.1145/2187980.2188200]
11. Teevan J, Morris M, Panovich K. Factors affecting response quantity, quality, and speed for questions asked via social network status messages. 2021 Presented at: Proc Fifth Int AAAI Conf Weblogs Soc Media; July 17-21, 2011; Barcelona, Catalonia, Spain p. 630-633 URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2795/3219 [doi: 10.1609/icwsm.v5i1.14159]
12. Nasehi SM, Sillito J, Maurer F, Burns C. What makes a good code example?: A study of programming Q&A in StackOverflow. 2012 Presented at: IEEE Int Conf Softw Maintenance; September 23-28, 2012; Trento, Italy p. 25-34. [doi: 10.1109/icsm.2012.6405249]
13. Chatterjee P, Damevski K, Kraft NA, Pollock L. Automatically identifying the quality of developer chats for post hoc use. ACM Trans Softw Eng Methodol 2021;30(4):1-28. [doi: 10.1145/3450503]
14. Fu H, Oh S. Quality assessment of answers with user-identified criteria and data-driven features in social Q&A. Information Processing & Management 2019 Jan;56(1):14-28. [doi: 10.1016/j.ipm.2018.08.007]
15. Shi J, Du W, Xu W. Identifying Impact Factors of Question Quality in Online Health Q&A Communities: an Empirical Analysis on MedHelp. PACIS 2018 Proceedings 2018:173 [FREE Full text]
16. Zhao Y, Wu L, Zhang J, Le T. How question characteristics impact answer outcomes on social question-and-answer websites. J Glob Inf Manag 2021;29(6):1-21. [doi: 10.4018/jgim.20211101.oa20]
17. Rassbach L, Pincock T, Mingus B. Exploring the feasibility of automatically rating online article quality. Int Wikimedia Conf. 2007. URL: https://pdfs.semanticscholar.org/789e/3b36a645bec2a99bc0c4d9760be3f22c40a9.pdf [accessed 2019-12-17]
18. Kitzie V, Choi E, Shah C. Analyzing question quality through intersubjectivity: world views and objective assessments of questions on social question-answering. Proc Am Soc Info Sci Tech 2013;50(1):1-10. [doi: 10.1002/meet.14505001052]
19. Zhou S, Jeong H, Green PA. How consistent are the best-known readability equations in estimating the readability of design standards? IEEE Trans Prof Commun 2017;60(1):97-111. [doi: 10.1109/tpc.2016.2635720]
20. Kincaid J, Fishburne R, Rogers R, Chissom B. Derivation of new readability formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) for navy enlisted personnel. Institute for Simulation and Training. 1975. URL: https://stars.library.ucf.edu/istlibrary/56 [accessed 2019-12-27]
21. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of LIWC2007. LIWC Inc, Austin, Texas 78703 USA conjunction with LIWC2007 Softw Progr LIWC Inc, Austin, Texas 78703 USA in Conjunction with the LIWC2007 Software Program. 2007. URL: https://www.researchgate.net/publication/228650445_The_Development_and_Psychometric_Properties_of_LIWC2007 [accessed 2024-08-13]
22. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc 5th Berkeley Symp Math Stat Probab 1967;1(14):281-297.
23. Tavakoli L, Zamani H, Scholer F, Croft WB, Sanderson M. Analyzing clarification in asynchronous information-seeking conversations. J Assoc Inf Sci Technol 2022;73(3):449-471 [FREE Full text] [doi: 10.1002/asi.24562]
24. Liu Y, Bian J, Agichtein E. Predicting information seeker satisfaction in community question answering. New York, New York, USA: ACM Press; 2008 Presented at: Proc 31st Annu Int ACM SIGIR Conf Res Dev Inf Retr - SIGIR-08; July 20-24, 2008; Singapore p. 483. [doi: 10.1145/1390334.1390417]
25. Rath M, Shah C, Floegel D. Identifying the reasons contributing to question deletion in educational Q&A. Proc. Assoc. Info. Sci. Tech 2017 Oct 24;54(1):327-336 [FREE Full text] [doi: 10.1002/pra2.2017.14505401036]
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [doi: 10.2307/2529310]
27. Zhang Y. Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. 2010 Presented at: Proc 1st ACM Int Heal; November 11-12, 2010; Arlington Virginia USA p. 210-219. [doi: 10.1145/1882992.1883023]
28. Roberts K, Demner-Fushman D. Interactive use of online health resources: a comparison of consumer and professional questions. J Am Med Inf Assoc 2016;23(4):802-811. [doi: 10.1093/jamia/ocw024] [Medline: 27147494]
29. Alasmari A, Zhou L. Share to seek: the effects of disease complexity on health information—seeking behavior. J Med Internet Res 2021;23(3):e21642. [doi: 10.2196/21642] [Medline: 33759803]
30. Alasmari A, Zhou L. How multimorbid health information consumers interact in an online community Q&A platform. Int J Med Inform 2019 Nov;131:103958. [doi: 10.1016/j.ijmedinf.2019.103958] [Medline: 31521012]
31. Coyne DW. CKD Medscape C Expert Column Series: Issue 3 - Management of Chronic Kidney Disease Comorbidities. Medscape. 2011. URL: https://www.medscape.org/viewarticle/736181
32. Centers for Disease Control and Prevention. 2021. URL: https://nccd.cdc.gov/CKD/Documents/Chronic-Kidney-Disease-in-the-US-2021-h.pdf

XSL•FO
RenderX

33. Kidney problem symptoms, causes, and types. American Kidney Fund. 2019. URL: https://www.kidneyfund.org/all-about-kidneys/chronic-kidney-disease-ckd [accessed 2024-08-13]
34. Corbin JM. The Corbin and Strauss Chronic Illness Trajectory model: an update. Sch Inq Nurs Pract 1998;12(1):33-41. [Medline: 9805470]
35. Hinkle JL, Cheever KH. Brunner & Suddarth's Textbook of Medical-Surgical Nursing. Gurgaon, Haryana: Wolters kluwer India Pvt Ltd; 2018.
36. Wang YC, Kraut RE, Levine JM. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. J Med Internet Res 2015;17(4):e99. [doi: 10.2196/jmir.3558]
37. Hersh W. Information retrieval. In: Shortliffe EH, Cimino JJ, editors. Biomed Informatics Comput Appl Heal Care Biomed. Cham: Springer International Publishing; 2021:755-794.
38. Horgan D, Romao M, Morré SA, Kalra D. Artificial intelligence: power for civilisation—and for better healthcare. Public Health Genomics 2019;22(5-6):145-161 [FREE Full text] [doi: 10.1159/000504785] [Medline: 31838476]
39. Kuziemsky C, Maeder AJ, John O, Gogia SB, Basu A, Meher S, et al. Role of artificial intelligence within the telehealth domain. Yearb Med Inform 2019;28(1):35-40 [FREE Full text] [doi: 10.1055/s-0039-1677897] [Medline: 31022750]

Abbreviations

CKD: chronic kidney disease
HIC: health information consumer
LIWC: linguistic inquiry and word count
Q&A: question-and-answer
SMOG: Simple Measure of Gobbledygook

Edited by G Tsafnat; submitted 17.04.23; peer-reviewed by C Turvey, J Shan, R Kravitz; comments to author 10.11.23; revised version received 09.04.24; accepted 10.07.24; published 12.09.24.

Please cite as:
Alasmari A, Zhou L
Quality Measurement of Consumer Health Questions: Content and Language Perspectives
J Med Internet Res 2024;26:e48257
URL: https://www.jmir.org/2024/1/e48257
doi:10.2196/48257
PMID:

©Ashwag Alasmari, Lina Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Alternative Health and Conventional Medicine Discourse About Cancer on TikTok: Computer Vision Analysis of TikTok Videos

Roxana Mika Muenster[1], MSc; Kai Gangi[2]; Drew Margolin[1], PhD

[1]Department of Communication, Cornell University, Ithaca, NY, United States

[2]Department of Computer Science, Cornell University, Ithaca, NY, United States

Corresponding Author:
Roxana Mika Muenster, MSc
Department of Communication
Cornell University
450 Mann Library
237 Mann Drive
Ithaca, NY, 14853
United States
Phone: 1 (607) 255 2601
Email: rm858@cornell.edu

Abstract

Background: Health misinformation is abundant online and becoming an increasingly pressing concern for both oncology practitioners and patients with cancer. On social media platforms, including the popular audiovisual app TikTok, the flourishing alternative health industry is further contributing to the spread of misleading and often harmful information, endangering patients' health and outcomes and sowing distrust of the medical community. The prevalence of false and potentially dangerous treatments on a platform that is used as a quasi–search engine by young people poses a serious risk to the health of patients with cancer.

Objective: This study seeks to examine how cancer discourse on TikTok differs between alternative health and conventional medicine videos. It aims to look beyond mere facts and falsehoods that TikTok users may utter to understand the visual language and format used in the support of both misleading and truthful narratives, as well as other messages.

Methods: Using computer vision analysis and subsequent qualitative close reading of 831 TikTok videos, this study examined how alternative health and conventional medicine videos on cancer differ with regard to the visual language used. Videos were examined for the length of time and prominence in which faces are displayed, as well as for the background setting, location, and dominant color scheme.

Results: The results show that the alt-health and conventional health samples made different use of the audiovisual affordances of TikTok. First, videos from the alternative health sample were more likely to contain a single face that was prominently featured (making up at least 7.5% of the image) for a substantial period of time (35% of the shots), with these testimonial-style videos making up 28.5% (93/326) of the sample compared to 18.6% (94/505) of the conventional medicine sample. Alternative health videos predominantly featured cool tones ($P<.001$) and were significantly more likely to be filmed outdoors ($P<.001$), whereas conventional medicine videos were more likely to be shot indoors and feature warm tones such as red, orange, or yellow.

Conclusions: The findings of this study contribute to an increased understanding of misinformation as not merely a matter of individual falsehoods but also a phenomenon whose effects might be transported through emotive as well as rational means. They also point to influencer practices and style being an important contributing factor in the declining health of the information environment around cancer and its treatment. The results suggest that public health efforts must extend beyond correcting false statements by injecting factual information into the online cancer discourse and look toward incorporating both visual and rational strategies.

(J Med Internet Res 2024;26:e60283) doi:10.2196/60283

KEYWORDS

misinformation; social media; TikTok; alternative health; cancer; computer vision

XSL•FO
RenderX

Introduction

Overview

Misinformation has become a significant topic of concern not only in politics but increasingly also in nonpolitical spheres, especially in the realms of health and medicine [1-4]; for instance, members of the much-publicized bot and fake news farms in Macedonia, known for interfering in American elections, have also admitted to involvement in alternative health sites [5]. The COVID-19 pandemic not only resulted in a medical and public health crisis but also ushered in a public information crisis often referred to as an infodemic [6-10]. However, COVID-19 is not the only health issue about which falsehoods are prevalent on social media. Online conversations about cancer are similarly marred by misinformation [11]. Common claims suggest that standard treatments are neither safe nor effective and that supposedly corrupt pharmaceutical companies should not be trusted [12-14]. Some posit that the success of conspiracy theorists and those peddling falsehoods is a byproduct of the conservative nature of pronouncement: scientists' tendency to speak on a subject only after they have sufficient certainty of the answer leaves many patients with unanswered questions. Those who are not governed by the norms and responsibility of academia are all too eager to try and fill this vacuum [14].

As with COVID-19, uncertainty and concomitant fear are prominent among patients with cancer [15-17]. With the urgency of a diagnosis often triggering information-seeking behavior, patients and carers search for help and support online, where they encounter not only unproven "alternative" treatments but at times also confounding and conflicting information from reliable sources [16-19]. The consequences of this confusion and misinformation can be dire because even a small delay in cancer treatment can have significant consequences for patient mortality [20-22]. Patients with cancer thus find themselves in an especially vulnerable position with regard to the misinformation present in online alternative health spaces [23].

Medical misinformation about cancer does not merely exist in response to patient demand; it has also been spurred on by the growing and flourishing industry of alternative health and wellness culture [24-26]. Content creators and social media platforms play a key role in facilitating the supply and encouraging the spread of health misinformation [4,27-29]. The rise of TikTok—the fastest-growing social media platform and one where falsehoods are plentiful—as a quasi-search engine for young adults has raised concerns among experts who worry that the platform will soon emerge as a new hotbed of misinformation [30-32].

In light of these developments, this study aims to examine cancer-related discourse and misinformation on TikTok. More specifically, we contrast the standard medicine and alternative health communities and demonstrate that not only do they differ in their narratives of illness and portrayal of the treatment of cancer, but they also use distinct visual-narrative languages when presenting these narratives. Using computer vision, we demonstrate that the 2 sides of the cancer community—one adhering to conventional medicine and another to alternative treatments—use differing visual rhetoric in their portrayal of healing and illness.

Background

The Growth of Health Misinformation and the Alt-Health Industry

Research has uncovered 3 trends with respect to cancer-related misinformation on social media.

The first is that cancer misinformation is prevalent on the internet (as it is offline). In a study of popular social media platforms, Johnson et al [33] found that not only did a third of the articles about cancer contain misinformation, but many also included misinformation that could cause economic and physical harm to patients, such as encouraging a delay in treatment or touting supposed treatment with potentially toxic effects. This pattern of misinformation is consistent with other observations of cancer discourse on social media, where misinforming content often receives higher engagement than factual information [19,33-36].

A second, related concern is that this prevalence of misinformation distorts public understanding of cancer and its treatment. Results from the National Cancer Survey indicate that 40% of Americans believe that alternative medicine alone can treat cancer, and nearly half give credence to at least 1 medical conspiracy, such as the notion that the Food and Drug Administration is suppressing a cure for cancer, a view held by 37% of the respondents [37,38]. Even if individual falsehoods are not given credence, their profusion in online cancer discourse creates an aura of doubt and confusion, which is associated with reduced compliance with screening protocols [2,18,38]. The National Cancer Institute defines complementary and alternative medicine as "any medical and health care systems, practices, or products that are not thought of as standard medical care" [39]. Alternative health assumes a "holistic view of health [that] takes into account the 'whole,'" contrasting it with the "mainstream" health system's approach to illness [40]. In this way, the alternative medicine industry claims to meet patients' needs in a way that conventional medicine cannot, providing an escape from "objectification, devaluation, and disempowerment" [41]. In actuality, many alternative health treatments for cancer have not been proven beneficial and safe, and some have even been shown to cause harm, including death, to patients [42,43].

Third, social media platforms have been instrumental in promoting the growing alt-health industry. Since its inception as a counterculture movement in California in the 1960s and 1970s, the so-called wellness culture has expanded into an intricate, interwoven system of goods and services, with the industry now estimated at US $1.5 trillion worldwide [29,44]. Between 1990 and 1997, researchers found a 47% increase in visits with natural health practitioners, while today between 20% and 80% of patients with cancer use natural methods to supplement their conventional treatment [45-47]. Alt-health, here, is used to differentiate between the substantial number of patients who may be using an alternative health modality in addition to their conventional treatment, and those individuals who form part of a movement of online promotion of alternative

XSL•FO
RenderX

health bound in misinformation, conspiracy, and ideology [29]. The industry's reach now extends to the period before diagnosis: the National Cancer Survey found that individuals are increasingly taking supplements—produced by an unregulated and sometimes unsafe industry—in an attempt to prevent cancer [37,48]. There is also a distinct financial component to alt-health; for instance, for the unsafe and ineffective "Gerson therapy," patients are charged multiple thousands of dollars per week [43,49]. In 2018, an article in *The BMJ* posed the question, "Is cancer fundraising fuelling quackery?" It noted that, in the United Kingdom alone, US $10 million had been raised on crowdfunding platforms for supposed alternative cancer treatments since 2012 [50]. Social media platforms also serve to amplify the alt-health industry in other ways: alternative cancer clinics are using these platforms to advertise their wares, increasing the visibility of their unproven approaches to treating cancer [51].

A further connection between the alt-health industry and misinformation is its skepticism toward the motives and methods of conventional science and medicine. Health misinformation and alt-health discourse are commonly shrouded in theories suggesting that the corrupt pharmaceutical industry is motivated not by patient care but by profit—or, in more extreme cases, by a supposed desire to inflict harm on patients [52,53]. Scholarship highlights the disparagement of "big pharma," a conspiracy theory used to refer to "an abstract entity comprised of corporations, regulators, NGOs, politicians, and often physicians" perceived to be financially benefiting from patients who are sick and dying [54]. In this view, conventional treatments are seen as toxic agents harming patients instead of treating them [23]. Alt-health proponents thus privilege supposed "natural" treatments over synthetic, "chemical" medicines. Following the logic of this so-called natural fallacy, what is natural is by default good, while anything chemical must be bad [23,55]. What this means, practically, is that alt-health can be influential by both promoting specific, sometimes dangerous, practices and undermining public confidence in established, medically supported treatments and practices. In 2018, an editorial in *The Lancet Oncology* voiced concerns over patients' reliance upon such unproven therapies, arguing that increased distrust of the medical system leaves room for the growth in popularity of alternative, supposed treatments [56].

Visual Communication via Video Platforms

Digital platforms have been instrumental in the spread of falsehoods and antiscience attitudes. This is in part due to their infrastructure, which privileges engagement and virality, as well as their ability to connect disparate users, allowing them to share bias-confirming evidence and emotional support [57,58]. The rising popularity of video platforms adds another reason for concern because many of their affordances are well suited to promote alternative health information. Specifically, audiovisual social media platforms give nonexpert content creators the ability to create and disseminate content that is effective at persuasion through processes that do not follow logical reasoning.

Scholars argue that 3 elements are important with regard to an audiovisual message's persuasive effect: what is said, how it is said, and how it looks [59]. Audiences frequently draw inferences about the credibility of a message through peripheral processing based on nonverbal cues given off by the speaker as well as the features of their presentation, such as vocal tone and how they are dressed [60]. As video is a richer medium, it affords the presentation of more of these cues. Social media platforms such as TikTok provide the creators of alternative health videos with an even wider range of editing tools, independent of evidence that they can (or cannot) provide for their claims.

Research suggests that visually presented information can influence persuasion even without engaging with judgments of credibility. Work on narrative persuasion shows that the features of a message can change its persuasiveness and the way audiences engage with it [60]. Narrative persuasion is not logical in the sense that it reasons with evidence and propositions [61,62]. Instead, narrative can be persuasive through the immersion of the audience in a simulated world in which causal relations are learned and then enacted to predict behaviors. Hamby et al [63] found that narratives have a larger effect on behavioral intent than nonnarrative formats, with some scholars attributing this effect to the ability of narratives to transport experience, motives, and metaphors as opposed to privileging rationality [61,64]. The effect is through the processes of embodied, imaginative simulation and interactions with identification, mirroring, and transportation [62,65,66]. These are derived from a variety of visual factors such as eye contact, facial expressions, emotive voices, or shared social or ethnic backgrounds [60,66-69]. When individuals identify with the characters in a story, they are both more likely to feel transported (increasing the story's impact) and more likely to believe that the causal structure of the world portrayed in the story applies to them.

The role of these embodied simulation processes should be particularly important for persuasion regarding the effect of health-oriented interventions (such as disease treatments). This is because these interventions claim to make changes to an individual's *body* and its healthfulness. Implicit in these claims is the idea that the treatment will make a person *feel* differently. By the logic of narrative, visual persuasion, the portrayal of treatments that make the audience feel a change toward healthfulness will seem to be effective; that is, if an individual *feels* healthy or perceives the person in the video as healthy after watching the video, they should be inclined to believe that the treatment has a healthful effect. Consider 2 ways to claim that an alternative to chemotherapy is effective. One way is to make arguments, perhaps based on statistical or even anecdotal evidence, that suggest that the alternative works. This argument may be augmented by peripheral cues, such as an authoritative voice, that lend credibility to the claims. Another way is for a patient to tell the story of their journey using this alternative method and to demonstrate or portray this journey visually. They might show still photos or video footage documenting the sadness they felt when they received the diagnosis and the happiness and vitality they experienced after their treatment—similar to the narratives seen in prescription drug advertisements [70]. The video demonstrates the effect of the treatment's impact on the way the body feels and looks, rather

XSL•FO
RenderX

than arguing for the effectiveness of the treatment based on how the body works. Audiences learn that people who use this therapy go from feeling one way (sad and sick) to feeling another way (happy and healthful), thereby concluding that the therapy is effective.

Consistent with this rationale, narrative testimonials are common in many media formats, including social media videos, and have been found to be more effective than health communication strategies based on the communication of statistical evidence [71-74]. For instance, Morton [74] posits that the use of narrative testimonials may aid influencers in the creation of authenticity and, by extension, trustworthiness. Studies of alt-health influencers show that personal testimonials, anecdotes, and subjective experience take precedence over expert knowledge and opinion [28,29]. Wellman [28] describes the influence of these emotionally engaging personal stories to be akin to that of "a friend who knows what they're talking about." This trust is easily abused; for example, an Australian wellness influencer profited greatly by falsely claiming that she had used diets and lifestyle products to cure herself of brain cancer [28,75]. Alternative cancer treatment videos may thus differ from standard cancer treatment videos in ways that emotively and visually highlight the supposed efficacy of treatments, not by making explicit false claims but by portraying an understanding of the world that is favorable to the treatments they promote.

Unsurprisingly, social media platforms facilitate the engaging message format of videos. Writing about Instagram Stories, Bainotti et al [76] highlight the "proliferation of small stories," where short-form video invites what they term "(micro)storytelling" or "micro-narrations." It can be argued that TikTok's infrastructure privileges narratives and personal and engaging storytelling over fact-based information sharing. One way it does so is through its timeline, the ForYouPage: in a world in which many short videos are consumed in a continuous stream, there can be an effect where the whole is greater than the sum of its individual parts. Examining social media narrative genres, researchers see patterns of "small stories," and fragments of storytelling [77,78]. Online storytelling is subject to cotellership, where interactions between users and the continuous scroll of videos placing individual videos in context with one another lead to them being seen not as fragments but as a whole, bolstering the overall credibility of their narrative [77-79]. Small stories and micronarratives are thus individual visual artifacts that function as fragments of a greater narrative, which, even if appearing or being seen separately, trigger prior instances of the same or a similar narrative.

In this view, influencers within the alt-health space thus create "tiles" of micronarratives in the form of videos, testimonials, and personal storytelling, which then form part of a wider alt-health narrative, parsed together by individual viewers [79]. Examining social media content as individual fragments, especially when it is provided as part of the endless flow of TikTok's ForYouPage, neglects the impact the fragments have as recognizable examples of a genre. Examining them as representatives of a greater narrative instead highlights the fact that each video enters into an interdiscursive narrative with other examples of the same genre: users are unlikely to encounter merely 1 video pertaining to cancer; instead, they will encounter numerous microcontent fragments at different times throughout the day and actively work to relate them to previously seen content [79].

Social media platforms do not just enable the sharing of visual content; they are also instrumental in their creation and editing. The rich multimodality and editorial affordances of audiovisual platforms, especially TikTok, mean that there is an abundance of choices and features for creators to include, further opening opportunities for new forms of visual storytelling [80]. The theory of affordances, coined by Gibson [81], refers to what an environment—ecological or otherwise—offers to those within it in terms of possible actions or uses. This theory has since been applied to the study of social media, among other fields, where it is used to understand how social media features shape the way a user engages with, and behaves in, their environments [82-85]. TikTok shares similarities with the internet and other social media platforms, such as offering users the ability to express themselves or be creative [86]. TikTok also boasts of other, more unique affordances relevant to visual storytelling, such as its replacement of a home or landing page with an algorithmically curated short-video feed for the user to watch without exerting great mental effort on curation, supporting the aforementioned micronarrative effect [87]. TikTok also differs from other social media platforms more focused on building networks in that it instead centers, and allows for, escapist entertainment, similar to television [86].

Another affordance offered by TikTok is an extensive array of options for creating and editing content, such as filters and stickers, as well as the ability to respond to other videos or comments with a video called stitching or dueting [87]. Drawing on the concept of mise-en-scène and visual genres in the analysis of these social media videos can highlight the constructedness of TikTok videos: creators are not merely haphazardly turning on their camera; they are also making choices about angles, backgrounds, props, setting, music, and text subtitles, all of which carry implicit meaning [59]. Prior research has examined the prevalence of certain features, such as color, in social media posts of the same genre [88]. While the exact impact of color tone is still unclear, studies have found cool tones to be associated with lower engagement [89] and fact-checking videos to be associated with warmer colors [88]. Outside of TikTok, the persuasive effect of faces and the long history of the use and efficacy of testimonials in advertising have been established in prior research, but their use in social media has been discussed to a lesser extent [29,72]. Previous scholarship has pointed to the effect of the presence of faces on engagement rates, and humans' tendency to focus on the faces in an image has long been established [88,90-93]. It is not just the mere presence of faces but also the way in which they appear that has become of interest: research indicates that expressions of joy and surprise are positively correlated with greater attention and viewer retention [94].

TikTok is also increasingly relevant due to its widespread use, especially among young people, who often use it as a search engine, and because it ushered in a new type of shorter-form video content that has since been adopted by other platforms due to its success [32,95-97]. TikTok's success and popularity,

XSL•FO
RenderX

especially among young people, have altered the ways in which users experience social media [87]. Long known primarily as an app through which teenagers shared dance videos, the use of TikTok has since expanded to other groups—including political campaigns, which now use TikTok to reach young voters and difficult-to-reach age groups [98-100]. In doing so, the campaigns have also adapted to the norms of communication on TikTok, often sharing (or performing) personal moments and life content, and entertaining material, rather than focusing on communicating policies [101,102]. Others have found that political communication on TikTok is more participatory and collaborative than on other platforms [103]. These features of communication on TikTok have to do with its primary user base, the affordances it offers to them, and the norms that develop around their use [103].

For our analysis, these findings on visual analysis and narrative suggest that we should aim to analyze the visual language of the alt-health and conventional health discourse on TikTok as a whole, with individual videos as parts of a greater narrative, rather than separate fragments. Thus, in examining the public communication environment on TikTok, a representative of the increasingly audiovisual social media landscape, we focus our analysis on the way in which these narratives are represented through visual choices, as well as on whether specific features such as background, location, and color are associated with either side. In this way, a valuable overview is created of the cancer-related public communication environment on TikTok. On the basis of this review of extant literature, we propose the following research questions (RQs):

- RQ1: How do the visual narratives about cancer shared by the conventional medical community on TikTok differ from those shared by the alternative health community?
- RQ2: In what ways are these differences communicated through specific visual narrative features?

Methods

Inference Strategy

Overview

To answer our RQs, our analysis proceeded through 3 stages: observation, testing, and explanation. In stage 1 (observation), we used a pilot sample of videos to manually identify theoretically interesting patterns that computer vision could plausibly detect. We used this stage to generate hypotheses for the computer vision analysis. In stage 2 (testing), we tested these hypotheses on a larger sample of videos. In stage 3 (explanation), we conducted a qualitative analysis of the video contents in light of the hypotheses supported by the computer vision analysis.

Computer Vision

Images transport significant meaning, but their analysis is often cumbersome or limited due to the labor-intensive nature of manual coding. Computer vision analysis, an automated method of analyzing visual data, offers an ability to examine visual data at scale [104]. Common uses include image classification, where each image is assigned a label out of a set of categories (eg, correctly identifying an image of a bike); detecting the presence of an object within an image; or recognizing human faces and bodies [104]. Previous research by Lu and Shen [88] used computer vision analysis to examine the properties of fact-checking videos on Douyin, the Chinese version of TikTok, to help guide decision-making on content creation for fact checkers. In political science, Joo and Steinert-Threlkeld [104] introduced new methods to use computational methods not only for the analysis of written and spoken political data but also for the analysis of meaningful and influential imagery used in politics. As some features cannot be measured by the human eye alone and because manual measurement, where possible, often lacks scalability, computer vision analysis is a promising tool for examining and understanding the vast amounts of content and data uploaded to social media platforms every day.

Defining Cancer Communication

We identified public videos by searching TikTok for those containing the term *cancer*. However, an initial search using this term alone resulted in a large proportion of videos about either astronomy or cancer in pets. Thus, to operationalize references to cancer as a disease in humans, the search phrase was modified to include a second keyword to identify relevant videos, that is, those concerning a discussion of cancer treatment in humans, while excluding irrelevant results. These search terms, which were identified from prior iterations of the search, yielded search results explicitly discussing the disease and its treatment in humans, while minimizing false positives (Textbox 1).

Textbox 1. The search phrases used for the data collection process on TikTok.

Search phrases

- Conventional medicine: cancer AND treatment OR cancer AND chemo
- Alt-health: cancer AND healing OR cancer AND naturalhealing OR cancer AND holistic OR cancer AND holistichealth OR cancer AND cure

Two pilot samples, consisting of 100 videos each from the conventional medicine and alt-health sides, were drawn in March 2023 by the second author. Search terms were tested for precision and proved satisfactory. Upon identification of the appropriate search strategy and the variables of interest, another set of samples, consisting of 2412 videos in total, was drawn in August 2023 for the testing phase. The samples were then processed to ensure relevance. First, duplicate videos were excluded. Second, videos whose captions did not include a combination of the term *cancer* and any of the other search terms were discarded from analysis. Next, videos were sorted into alt-health and conventional medicine categories based on the inclusion of the respective search terms in the video captions. Videos that used terms from both conventional medicine and alt-health categories (eg, *cancer*, *treatment*, and *cure*) were sorted into both categories because average users would feasibly

XSL•FO
RenderX

encounter them with either of these search terms. Ultimately, of the original dataset of 2412 videos, we were left with 326 (13.52%) videos in the alt-health category and 505 (20.94%) videos in the conventional medicine category for analysis.

Initial Manual Inspection of Videos

The lead author first examined the sampled videos in an exploratory process to identify meaningful features for our computer vision analysis. The pilot sample (200 videos) was screened to identify visual patterns that were (1) detectable by computer vision; and (2) theoretically meaningful as it pertains to the arguments on visual storytelling outlined previously. For example, some videos from the alt-health sample showed people chopping vegetables as part of their routine when preparing a home remedy for cancer—while this is theoretically interesting and did seem to appear frequently in only a certain type of video, training the machine to reliably identify this particular behavior would require identifying hundreds of training examples for different vegetables in different forms. Similarly, the computer might be able to detect a feature such as the wearing of face masks, which satisfies the first point, but this feature's theoretical meaning (the second point) could be obscured by external factors. For example, patients seeking chemotherapy may be more likely to appear masked in videos simply because they are filming themselves in places where mask wearing was mandated, such as hospitals, rather than making a conscious decision to include face masks in their videos.

This exploratory process pointed to 4 ways of conducting analysis on the larger sample using computer vision: the presence and prominence of faces ("testimonials"), scene and nature, color choice, and emotion displays. The rationale for these is described in the following paragraphs. On the basis of the pilot, we developed hypotheses for what we would observe and test for in the larger sample.

The first tendency we observed was for videos to be in the form of a testimonial. We observed many videos that were (1) posted by an ordinary person (meaning neither a physician nor a celebrity) who is (2) personally affected and recounting their experience and (3) that show the person's face prominently and for a prolonged period. This is consistent with the definition of a testimonial video [71,74]. We also observed that, consistent with the aforementioned theoretical rationale, alt-health videos featured a greater number of testimonials than conventional medicine videos. However, not all of these features can be detected by the computer. Therefore, we hypothesized that the computer vision implication of this style—the prolonged and prominent display of one person's face, which we refer to as *testimonial presentation*—will be significantly associated with alt-health videos (H1).

Second, we observed tendencies with regard to the locations in which the videos were filmed and the presence of nature overall. Arguments for alternative medicine often rely on the so-called natural fallacy, which privileges what is perceived as "natural" over the "chemical" [23]. Consistent with this, we noted that many videos in the alt-health pilot sample featured natural elements such as trees or gardens. Conversely, we observed many features and props in the conventional medicine pilot sample that indicated a clinical or industrial environment. Although, as noted previously, it was not feasible to identify each object, we observed that simply identifying the setting of the video as "outdoors" (natural) versus "indoors" (not natural) provided an approximation of the natural versus not natural effect because most of the videos that showed "nature" did so in an outdoor setting, while those that did not used an indoor setting. Thus, we hypothesized that alt-health videos featured a greater number of outdoor settings than conventional medicine videos (H2).

The third tendency observed concerns the dominant color tone of the video. We noticed in our alt-health pilot sample a dominance of green and blue hues accompanying the outdoor and natural settings (greenery and the sky, respectively) and in our conventional health sample an often hospital-related presence of beige and gray. We expect this pattern to hold true in the larger sample (H3).

The fourth tendency relates to the emotions displayed by the people in the videos. Cancer and its treatment represent times of significant upheaval for patients, and in light of this and prior findings on human reaction and attention to emotions [94], we were interested in whether the range of emotions displayed differed between the conventional medicine and alt-health samples. We observed that the alt-health sample showed significantly more happy emotions, whereas the conventional medicine sample featured more expressions of sadness.

Computer Vision Analysis

Computer vision tools were used to assess certain features at scale, as well as other features that the human eye alone cannot measure. The *face_recognition* Python package [105], which boasts 99.38% accuracy, was used to determine the number of faces present throughout the video, as well as their proportional size of the screen (H1). For a given video, the face percentage refers to the percentage of frames in a video in which ≥1 faces are detected. Face size is calculated by taking the average face size calculated for each frame that contains a face. We manually checked 50 (6%) of the 831 videos to determine the precision and recall of the analysis and found it satisfactory with an F_1-score of 0.857.

Next, to assess the presence of nature and scene, we used the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory's computer vision, Places365 convolutional neural networks for scene classification [106]. This was used to analyze whether a video was recorded indoors or outdoors (H2). The convolutional neural network was trained on a Places2 database [107] of approximately 1.8 million images. It classifies frames as *indoor* or *outdoor* based on certain features, such as the presence of indoor lighting and man-made materials (eg, plastic) for indoor scenes, and the presence of a horizon, natural lighting, and natural fibers for outdoor scenes. While the program can provide additional scene specifications, such as categorizing the location further to, say, offices, courtyards, beer gardens, or dressing rooms, these capabilities were not used for this analysis due to a lack of theoretical reasoning or hypotheses for any of these locations. Precision and recall testing for scene classification proved satisfactory with an F_1-score of 0.978. In conjunction with this, ColorKit

XSL•FO
RenderX

[108] was used to determine color tone (H3). Color tone is measured directly by the computer rather than being an approximation of a human-observed ground truth; therefore, we did not evaluate precision and recall for this analysis.

Finally, facial expression recognition using computer vision analyses proved unsuccessful. Human inspection of the sample showed that while surprise was identified as the most common emotion by the machine, a large number of these assessments were inaccurate. We hypothesize that this might be due to the norms of speaking and engaging on social media platforms, which prioritize engaging narration, sensationalism, and emotiveness—all of which could contribute to facial expressions typically associated with surprise, a product of "social media speak," so to speak. As a result, we dispensed with this measure for testing our hypotheses.

Qualitative Investigation

Next, the lead author conducted an audiovisual close reading of a sample of the analyzed videos to identify themes and styles that highlight and explain the computational findings. This reading thus served to "identify, analyze, and report" the patterns in which the examined features and computational results manifest in the videos [109]. The qualitative investigation paid close attention to the themes or recurring styles in each set, how they differed between the 2 samples, and, crucially, how the computationally examined features appeared within them.

Ethical Considerations

This study did not require ethics approval from an institutional review board because the data that were collected and analyzed are publicly available, posted to public TikTok accounts. Individual accounts are neither named nor described in detail. Any screenshots reproduced in this study from the publicly posted videos have been anonymized, including the users' appearance and usernames. A reverse image search was conducted with the anonymized images to ensure that the original accounts would not be discoverable as a result of this study.

Results

Computer Vision Analysis

Testimonials were operationalized using a combination of face presence (ie, is a face present in the video in this frame?) and face size (ie, how large is the face?). Three-fold categorization yielded the results presented in Table 1.

H1 examines in which of the 2 samples face size and face presence are more prominent. Operationalizing face size and face presence as indicators of testimonial presentation, these videos were found to be more frequent in the alt-health sample than in the conventional medicine sample (H1) under the *loose* categorization ($P<.001$) but not under the *medium* and *strict* categorizations (medium: $P=.19$; strict: $P=.13$), although the direction is consistent with H1. This means that videos in the alt-health sample showed faces closer to the camera and for a longer period of time in the testimonial-style presentation that is commonly used in influencer content in video format and known to be persuasive in its narrative form in health communication [28,71,74].

H2 hypothesizes that outdoor settings will be more prominent in alt-health videos than in conventional medicine videos. The relationship between alt-health videos and outdoor scenery was found to be significant ($P<.001$). Videos uploaded by patients with cancer seeking conventional treatment were thus more often recorded indoors, whereas patients with cancer seeking and promoting supposed alternative treatments chose to record and upload videos shot outdoors or with greenery and nature present.

Finally, in examining color tone in our sample, we found that cool-toned videos were associated more with alt-health videos than with conventional medicine videos (overall: $P<.001$; first dominant color: $P<.001$, second dominant color: $P<.001$, and third dominant color: $P<.001$). Conventional medicine videos were therefore more likely to have dominant warm undertones such as red, orange, or yellow, whereas alt-health videos were more likely to be dominated by cool undertones such as blues and greens.

Table 1. Results for the facial recognition computer vision analyses of 2 samples from TikTok, one representing an alternative health approach to treating cancer and the other a conventional medicine approach.

	Conventional medicine (n=505), n (%)		Alt-health (n=326), n (%)	
	Testimonial presentation	Nontestimonial presentation	Testimonial presentation	Nontestimonial presentation
Loose (35% face presence, 7.5% face size)	94 (18.6)	411 (81.4)	93 (28.5)	233 (71.5)
Medium (40% face presence, 10% face size)	77 (15.2)	428 (84.8)	61 (18.7)	265 (81.3)
Strict (45% face presence, 12.5% face size)	47 (9.3)	458 (90.7)	41 (12.6)	285 (87.4)

Qualitative Investigation

An audiovisual close reading of a sample of the analyzed videos demonstrated that the computational findings correspond to distinct styles and themes between the 2 categories.

A distinct style of testimonial on the conventional medicine side shows the act of head shaving. Users frequently posted recordings of the shaving of their heads in a bathroom. The caption or in-video texts explain that this decision is in response to, or to preempt, the effects of chemotherapy, which often leads to drastic hair loss. Sometimes, the patient is seen alone, shaving

XSL•FO
RenderX

their own head; more often, other people—family or friends—are seen helping the patient in the act of shaving. The mood is somber, as is the music choice, and sad expressions and tears are frequent throughout. The videos usually conclude with the user revealing their shaved head (Figure 1).

Within the alt-health sample, "visuals of wellness" represents a common theme. Popular videos show the users, often women, in ways that do not make it apparent that they have been diagnosed with cancer: users are seen outdoors, with vibrant greenery, bright sunshine, and blue skies featuring prominently (Figure 2). The users are tanned, have long hair, and may be seen engaging in physical activity or preparing food. In the caption, the whole foods seen in the video—juicy fruits, greens, or spices—are claimed to represent much more than mere nutrition; they are purported to be the cure for the disease these users have, or had, been diagnosed with.

Visually, these videos exemplify the results of the computer vision analysis: the pattern of visual elements used within them mirrors itself across individual videos to form a genre. The visual choices of how the users in the 2 samples choose to portray their cancer journey, paint a drastically different picture: one highlights illness, while another obscures the signs of illness to highlight wellness; one shows the wide-reaching side effects of cancer treatment, while another shows a person who looks healthy after their cure; and one shows the patient with cancer visibly ill in windowless rooms or in hospitals, while another displays the patient apparently healthy and in nature.

These differences in visual genres, of course, are mirrored in the content of the sampled videos: conventional medicine videos usually highlight the pain, distress, and sadness of a cancer diagnosis and treatment cycle, often providing medical information about the exact type of cancer and its treatment. Alt-health videos, by contrast, contain misinformation casting doubt on the efficacy of chemotherapy or radiation (instead portraying them as poison), insinuating the existence of a cure, or presenting dangerous and ineffective ingredients or practices as safe and efficacious. Both video genres exemplify patient testimonials. In the conventional medicine sample, these patient testimonials differ from expert "lessons," where physicians explain the science and mechanics of the disease. These videos feature medical personnel who convey their credentials not only through words and text (such as an account handle with "Dr. med.") but also visually; by wearing scrubs, a laboratory coat, or a perfunctory stethoscope. In the alt-health sample, the "visuals of wellness" videos combine both patient and expert testimonials: users recount their experience as patients and build their expertise on the basis of this experience, proclaiming their cure, with their wellness projected as living proof of the treatment's efficacy and thus their expertise.

Figure 1. An example of the head-shaving genre in a TikTok video from the conventional health sample. The results from the computer vision analysis showed that videos, like this one, that feature or advocate for conventional medicine approaches to treating cancer are more likely to be filmed indoors than alternative health videos.

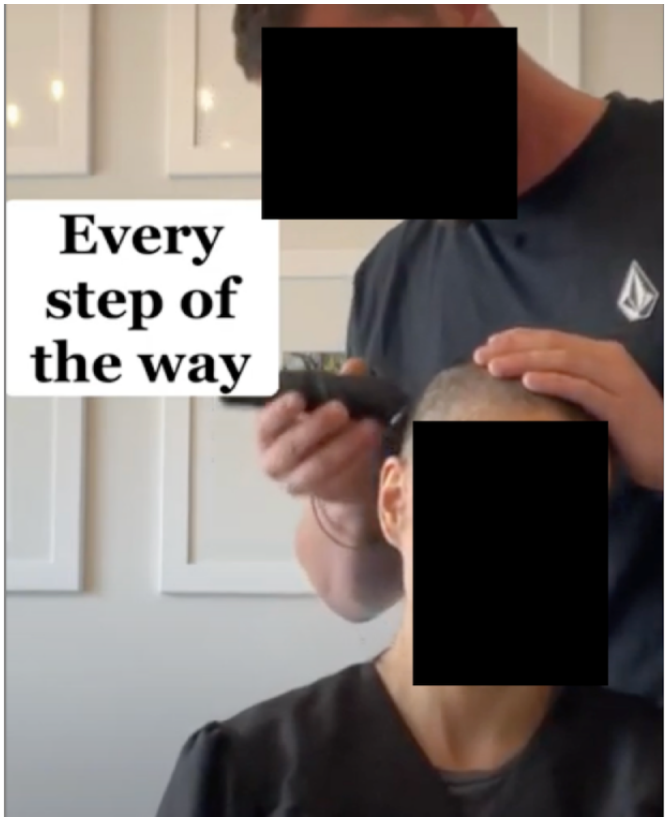

XSL•FO
RenderX

Figure 2. An example of the outdoors testimonial genre in a TikTok video from the alternative health sample. This screenshot exemplifies results from the computer vision analysis that showed that videos advocating for unproven alternative medicine approaches to treating cancer are more likely to be recorded outdoors and feature faces significantly (both in presentation duration and size).

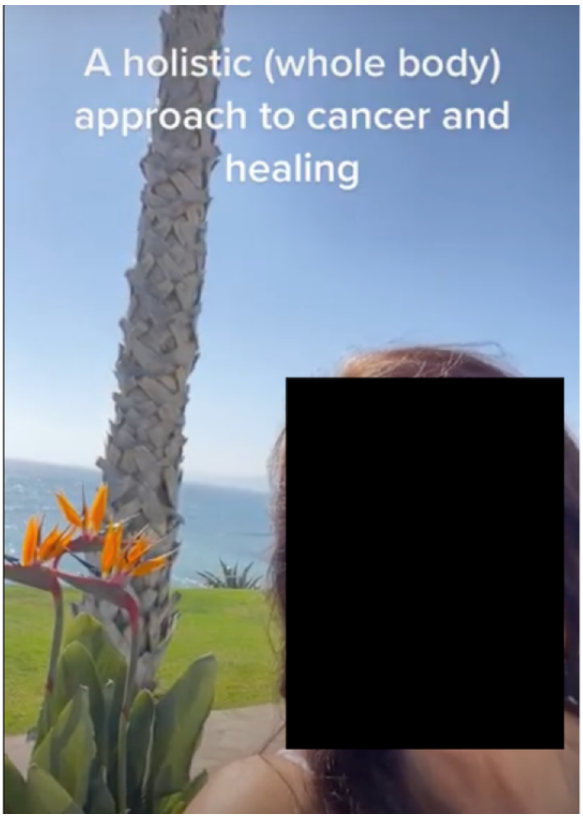

Discussion

Computer Vision Analysis Findings

This study set out to examine how the visual language of videos about cancer on TikTok differs between those who support alternative health treatments and those who adhere to conventional medicine. The results show that videos advocating for unproven alternative health cures for cancer will more frequently do so in a testimonial-style video than those advocating for conventional medicine approaches, meaning that the alternative health videos will show users' faces more prominently and for significant portions of the video. Similarly, support for natural health remedies is often underscored by natural scenery and backgrounds. When comparing the dominant color tones between the alt-health sample and the conventional medicine sample, the former tend toward cool tones and the latter toward warm tones.

In examining conversations about health, researchers have clearly established the prevalence of falsehoods and misleading claims and facts [51]. What our findings show is that there may be an additional layer to misinforming videos beyond the veracity of the claims made: there are substantial differences in the ways in which videos promoting alternative health "cures" for cancer attempt to persuade their audience visually. The alt-health videos we examined use visual social media to create a more transportive effect, one that fosters a personal connection with the audience, heightens the perceived authenticity of the user, and highlights the purported healthful life and path their chosen treatment affords. By contrast, the conventional medicine sample features videos that highlight the emotional distress of a diagnosis and the arduous process that treatment involves: videos displayed testimonial presentation to a lesser degree, showed a greater degree of sadness, and filmed themselves in hospital settings. It is possible that the patients with cancer in the conventional medicine sample did not feel the need to use persuasive video elements such as testimonial presentation because these videos often lacked the commercial impetus of alt-health videos or because they instead highlighted those aspects of their life that helped foster a connection to other patients with cancer by making their illness visible [110].

The 2 examples of common themes further exemplify the findings from the computational analysis: video recordings of patients shaving their head are usually shot from farther away than the personable close-ups in the alt-health sample, leading to the differences in face size and face presence between the 2 samples in what we label a testimonial presentation. Moreover, while the act of shaving might require a power outlet, thus making it more likely to take place indoors, other types of videos in the conventional health sample also did not feature outdoor backgrounds in the way that alt-health videos did. In addition, users could have chosen to share another type of video but instead adhered to a genre or style of video that is common within their community and puts them in communication with other members of this community.

Our findings indicate that there is little visual overlap between the alt-health group of patients with cancer and the conventional medicine group of patients with cancer on TikTok, mirroring the findings of Milani et al [111] that show that pro- and

XSL•FO
RenderX

antivaccine users on Twitter (subsequently rebranded X) formed 2 separate, barely interacting networks. Some of the features, such as the role of testimonials, reflect a common genre of social media content that can be used by public health practitioners to build familiarity and credibility with potential viewers. Our findings also point to the necessity of considering the evolving nature of social media in health communication research and practice: viewing digital misinformation through the lens of the impact of a single message could leave us blind to the reality of social media consumption, which relies on small stories spread across fragments of videos and posts spreading the same visual narrative and adhering to features of the same genre [76,77].

The visual languages used by the users in the 2 samples paint 2 starkly different pictures of the journey and treatment of a patient with cancer. Users promoting alt-health on TikTok videos and images espouse purported natural cancer cures, claiming that these alternatives can cure cancer in days or weeks and without the uncomfortable side effects of conventional (actual) treatments. This portrayal presents them as a more attractive alternative to the arduous chemotherapy or radiation treatments depicted in the conventional medicine sample videos. These claims are then reinforced through visual narratives highlighting nature and natural objects, building an aspirational image often at play in the creator culture and constructing narratives of wellness: healthy, happy people enjoying life in nature. By contrast, the conventional medicine side builds and displays narratives of illness: users are seen indoors, often in hospitals or in bathrooms recording the process of shaving their hair due to the effects of chemotherapy. Nature is rarely present, and both the physical appearance and emotional expressions of the patients with cancer in these videos is in line with what one would expect of those undergoing cancer treatment: subdued or sad emotions, teary expressions, and pale skin are prevalent throughout this sample. In effect, this acts as a counterpoint to the visions of health and vitality portrayed in the alt-health sample and presents a picture to viewers and potential patients of an arduous and painful treatment and cancer journey.

Of the between 20% and 80% of patients with cancer who claim that they use alternative treatments at least supplementary to their conventional treatment, many will venture onto the web to search for resources [37,46-48]. However, patients and carers on TikTok will not just be exposed to relatively harmless advice to consider nonmedical wellness behavior such as adopting a healthy diet and avoiding harmful substances alongside chemotherapy or radiation treatment but will also encounter content advising them to replace these medically proven treatments with supposed “natural” alternatives. Beyond this, users searching for natural treatment through the search terms employed in this study will invariably come across videos that combine misinformation about the effectiveness of these supposed treatments with conspiracy ideology. Times of crisis, turmoil, and confusion can make individuals vulnerable to the influence of conspiracy ideologies—for patients navigating the throes of a cancer diagnosis, encountering content claiming that governments and pharmaceutical companies are concealing effective cancer treatments (conspiracies that are often laced with anti-Semitic ideology) could prove especially dangerous. This becomes especially concerning as we consider which narratives users are exposed to over time as opposed to in just 1 post: users do not usually make up their mind on the basis of rational arguments or even just 1 post; instead, they examine content fragments as part of a greater narrative and ideological position [112].

This study did not aim to be a comprehensive study of all visual features of TikTok videos on the topic of cancer. We selected features that were deemed theoretically relevant to the topic at hand, 3 of which (testimonials, scene and nature, and color choice) were successfully analyzable using computer vision. Further research aiming to use similar methodology on other topics will likely have to adapt their selection of visual features accordingly (although we expect testimonials to be a relevant phenomenon of much audiovisual social media content overall). Just as the examination of visual features is not exhaustive, there remain other elements of social media videos that were not considered in this study that could be of note, such as captions, soundtracks, and emojis. Future research should examine other elements of the platform’s rich audiovisual environment [113]. Finally, videos were collected via search phrases, and the TikTok app’s algorithms are opaque in their functioning. However, this method mirrors the way users would typically search for and encounter content if they were to seek information about cancer on TikTok.

TikTok’s affordances privilege communication that is emotive and personable, and these meanings are substantially transported through visual features [87]. Political campaigns have made moves to adapt to this style of communication [101,102]. Our findings demonstrate that health communication cannot rely on the truthful and rational depiction of medical facts alone and must also consider adapting to this new audiovisual platform and the norms that govern it.

Conclusions

On topics ranging from vaccines and nutrition to wellness, the internet has become a central source of both factual and false information for patients [114]. Our examination of the public communication environment about cancer on TikTok, a central social media platform, found that different types of information are spread in different ways: compared to the conventional medicine sample videos, the alternative health sample videos rely more heavily on testimonial presentation and feature more natural elements and settings. We also add to a growing body of research on the role of color in images and videos: we found that cool-toned videos are more frequently found in the alt-health sample than in the conventional medicine sample. Our findings suggest that the alternative health community leverages the advantages of the persuasive power of audiovisual social media platforms such as TikTok. Through their visual choices and language, they portray wellness where the conventional medicine sample portrays illness, thus positing a choice to viewers between 2 unequal (supposed) solutions to their diagnosis.

Our findings also provide insights for public health scholars and practitioners: while it is important to provide patients with accurate, reliable information, including on social media platforms such as TikTok, this may not be the most successful

XSL•FO
RenderX

approach or align with what users are seeking on the internet. In part, this is due to the logic of social media, which privileges all but dry, factual information. Beyond this, it is because patients scouring social media may not be rationally looking for information to inform their treatment decisions but instead searching for ways to cope with the disease in a manner that gives them hope. Alt-health influencers use the visual language of the testimonial, among other things, to build authenticity and rapport with their users and, crucially, create hope for an easier, less arduous road to recovery. Medical professionals and public health scholars might consider using similar visual language to meet patients' emotional needs online, rather than focusing solely on their information needs.

Acknowledgments

This work was supported by the National Cancer Institute (grant R37CA259156). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute.

Data Availability

The datasets generated and analyzed during this study are not publicly available to preserve the anonymity of the uploaders of the videos included in the dataset but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Bovet A, Makse HA. Influence of fake news in Twitter during the 2016 US presidential election. Nat Commun 2019;10(7):1-14 [FREE Full text] [doi: 10.1038/s41467-018-07761-2] [Medline: 30602729]
2. Grimes DR. Medical disinformation and the unviable nature of COVID-19 conspiracy theories. PLoS One 2021;16(3):e0245900 [FREE Full text] [doi: 10.1371/journal.pone.0245900] [Medline: 33711025]
3. Kucharski A. Post-truth: study epidemiology of fake news. Nature 2016;540(7634):525. [doi: 10.1038/540525a] [Medline: 30905947]
4. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
5. Kitta A. Alternative health websites and fake news: taking a stab at definition, genre, and belief. J Am Folk 2018;131(522):405-412. [doi: 10.5406/jamerfolk.131.522.0405]
6. Cuan-Baltazar JY, Muñoz-Perez MJ, Robledo-Vega C, Pérez-Zepeda MF, Soto-Vega E. Misinformation of COVID-19 on the internet: infodemiology study. JMIR Public Health Surveill 2020;6(2):e18444 [FREE Full text] [doi: 10.2196/18444] [Medline: 32250960]
7. Immunizing the public against misinformation. World Health Organization. 2020 Aug 25. URL: https://www.who.int/news-room/feature-stories/detail/immunizing-the-public-against-misinformation [accessed 2024-11-14]
8. Nguyen A, Catalan-Matamoros D. Digital mis/disinformation and public engagement with health and science controversies: fresh perspectives from Covid-19. Media Commun 2020;8(2):323-328. [doi: 10.17645/mac.v8i2.3352]
9. Lee JJ, Kang KA, Wang MP, Zhao SZ, Wong JY, O'Connor S, et al. Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. J Med Internet Res 2020;22(11):e22205. [doi: 10.2196/22205] [Medline: 33048825]
10. UN tackles 'infodemic' of misinformation and cybercrime in COVID-19 crisis. United Nations. 2020 Mar 31. URL: https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-covid-19 [accessed 2024-11-14]
11. Chen L, Wang X, Peng TQ. Nature and diffusion of gynecologic cancer-related misinformation on social media: analysis of tweets. J Med Internet Res 2018;20(10):e11515 [FREE Full text] [doi: 10.2196/11515] [Medline: 30327289]
12. Hernandez D, McMillan R. Facebook, YouTube overrun with bogus cancer-treatment claims. The Wall Street Journal. 2019 Jul 02. URL: https://www.wsj.com/articles/facebook-youtube-overrun-with-bogus-cancer-treatment-claims-11562072401 [accessed 2024-11-14]
13. Grimes DR. How to survive the fake news about cancer. The Guardian. 2019 Jul 14. URL: https://www.theguardian.com/science/2019/jul/14/cancer-fake-news-clinics-suppressing-disease-cure [accessed 2024-11-14]
14. Wardle C. Tackling misinformation: a three-pronged approach. National Institutes of Health. 2023 Sep 06. URL: https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/perspectives/tackling-misinformation-three-pronged-approach [accessed 2024-11-14]
15. Dunwoody S. Science journalism and pandemic uncertainty. Media Commun 2020;8(2):471-474. [doi: 10.17645/mac.v8i2.3224]

XSL•FO
RenderX

16. Lee SY, Hawkins RP. Worry as an uncertainty-associated emotion: exploring the role of worry in health information seeking. Health Commun 2016;31(8):926-933. [doi: 10.1080/10410236.2015.1018701] [Medline: 26752071]
17. Liu J, King AJ, Margolin D, Niederdeppe J. Information seeking and scanning about colorectal cancer screening among Black and White Americans, ages 45-74: comparing information sources and screening behaviors. J Health Commun 2020;25(5):402-411. [doi: 10.1080/10810730.2020.1776424] [Medline: 32529955]
18. Gaysynsky A, Senft Everson N, Heley K, Chou WY. Perceptions of health misinformation on social media: cross-sectional survey study. JMIR Infodemiology 2024;4:e51127 [FREE Full text] [doi: 10.2196/51127] [Medline: 38687591]
19. Warner EL, Waters AR, Cloyes KG, Ellington L, Kirchhoff AC. Young adult cancer caregivers' exposure to cancer misinformation on social media. Cancer 2021;127(8):1318-1324 [FREE Full text] [doi: 10.1002/cncr.33380] [Medline: 33368215]
20. Chotipanich A, Sooksrisawat C, Jittiworapan B. Association between complementary and alternative medicine use and prolonged time to conventional treatment among Thai cancer patients in a tertiary-care hospital. PeerJ 2019;7:e7159 [FREE Full text] [doi: 10.7717/peerj.7159] [Medline: 31231600]
21. Davis GE, Bryson CL, Yueh B, McDonell MB, Micek MA, Fihn SD. Treatment delay associated with alternative medicine use among veterans with head and neck cancer. Head Neck 2006;28(10):926-931. [doi: 10.1002/hed.20420] [Medline: 16755583]
22. Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ 2020;371:m4087. [doi: 10.1136/bmj.m4087] [Medline: 33148535]
23. Grimes DR. The struggle against cancer misinformation. Cancer Discov 2022;12(1):26-30. [doi: 10.1158/2159-8290.CD-21-1468] [Medline: 34930788]
24. Abraham J, White KM. Tracking the changing landscape of corporate wellness companies. Health Aff 2017;36(2):222-228. [doi: 10.1377/hlthaff.2016.1138] [Medline: 28167709]
25. Kickbusch I, Payne L. Twenty-first century health promotion: the public health revolution meets the wellness revolution. Health Promot Int 2003;18(4):275-278. [doi: 10.1093/heapro/dag418] [Medline: 14695358]
26. Lagrosen S, Lagrosen Y, Lind L. Health leadership in the wellness industry. In: Proceedings of the 3rd International Conference on Management, Leadership And Governance. 2015 Presented at: ICMLG 2015; February 12-14, 2015; Auckland, New Zealand.
27. Lofft Z. When social media met nutrition. Health Sci Inq 2020;11(1):56-61. [doi: 10.29173/hsi319]
28. Wellman ML. “A friend who knows what they’re talking about”: extending source credibility theory to analyze the wellness influencer industry on Instagram. New Media Soc 2023;26(12):7020-7036. [doi: 10.1177/14614448231162064]
29. Baker SA. Alt. health influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the COVID-19 pandemic. Eur J Cult Stud 2022;25(1):3-24. [doi: 10.1177/13675494211062623]
30. Bhargava P, MacDonald K, Newton C, Lin H, Pennycook G. How effective are TikTok misinformation debunking videos? Harv Kennedy Sch Misinformation Rev 2023;4(2):1-17. [doi: 10.37016/mr-2020-114]
31. Bursztynsky J. TikTok says 1 billion people use the app each month. CNBC. 2021 Sep 27. URL: https://www.cnbc.com/2021/09/27/tiktok-reaches-1-billion-monthly-users.html [accessed 2024-11-14]
32. Huang K. For Gen Z, TikTok is the new search engine. The New York Times. 2022 Sep 17. URL: https://www.nytimes.com/2022/09/16/technology/gen-z-tiktok-search-engine.html [accessed 2024-11-14]
33. Johnson SB, Parsons M, Dorff T, Moran MS, Ward JH, Cohen SA, et al. Cancer misinformation and harmful information on Facebook and other social media: a brief report. J Natl Cancer Inst 2022;114(7):1036-1039 [FREE Full text] [doi: 10.1093/jnci/djab141] [Medline: 34291289]
34. Bal R, Sinha S, Dutta S, Joshi R, Ghosh S, Dutt R. Analysing the extent of misinformation in cancer related tweets. In: Proceedings of the International AAAI Conference on Web and Social Media. 2020 Presented at: ICWSM 2020; June 8-11, 2019; Virtual Event. [doi: 10.1609/icwsm.v14i1.7359]
35. Gage-Bouchard EA, LaValley S, Warunek M, Beaupin LK, Mollica M. Is cancer information exchanged on social media scientifically accurate? J Cancer Educ 2018;33(6):1328-1332. [doi: 10.1007/s13187-017-1254-z] [Medline: 28721645]
36. Forster K. Revealed: how dangerous fake health news conquered Facebook. The Independent. 2017 Jan 07. URL: https://www.independent.co.uk/tech/fake-news-health-facebook-cruel-damaging-social-media-mike-adams-natural-health-ranger-conspiracy-a7498201.html [accessed 2024-11-14]
37. ASCO 2018 cancer opinions survey. American Society of Clinical Oncology. 2018 Oct. URL: https://old-prod.asco.org/sites/new-www.asco.org/files/content-files/research-and-progress/documents/2018-NCOS-Results.pdf [accessed 2024-11-14]
38. Oliver JE, Wood T. Medical conspiracy theories and health behaviors in the United States. JAMA Intern Med 2014;174(5):817-818. [doi: 10.1001/jamainternmed.2014.190] [Medline: 24638266]
39. Complementary and Alternative Medicine for Patients. National Cancer Institute. URL: https://www.cancer.gov/about-cancer/treatment/cam/patient [accessed 2024-11-14]
40. McKee J. Holistic health and the critique of western medicine. Soc Sci Med 1988;26(8):775-784. [doi: 10.1016/0277-9536(88)90171-2] [Medline: 3287633]

XSL•FO
RenderX

41. Sointu E. The search for wellbeing in alternative and complementary health practices. Sociol Health Illn 2006;28(3):330-349 [FREE Full text] [doi: 10.1111/j.1467-9566.2006.00495.x] [Medline: 16573719]
42. Complementary and alternative medicine. National Institutes of Health National Cancer Institute. URL: https://www.cancer.gov/about-cancer/treatment/cam [accessed 2024-11-14]
43. Gerson therapy (PDQ®)–patient version. National Institutes of Health National Cancer Institute. URL: https://www.cancer.gov/about-cancer/treatment/cam/patient/gerson-pdq [accessed 2024-11-14]
44. Callaghan S, Lösch M, Pione A, Teichner W. Feeling good: the future of the $1.5 trillion wellness market. McKinsey & Company. 2021 Apr 8. URL: https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/feeling-good-the-future-of-the-1-5-trillion-wellness-market [accessed 2024-11-14]
45. Eisenberg DM, Davis RB, Ettner SL, Appel S, Wilkey S, Van Rompay M, et al. Trends in alternative medicine use in the United States, 1990-1997: results of a follow-up national survey. JAMA 1998;280(18):1569-1575. [doi: 10.1001/jama.280.18.1569] [Medline: 9820257]
46. Alsharif F. Discovering the use of complementary and alternative medicine in oncology patients: a systematic literature review. Evid Based Complement Alternat Med 2021:6619243 [FREE Full text] [doi: 10.1155/2021/6619243] [Medline: 33519943]
47. Johnson SB, Park HS, Gross CP, Yu JB. Complementary medicine, refusal of conventional cancer therapy, and survival among patients with curable cancers. JAMA Oncol 2018;4(10):1375-1381 [FREE Full text] [doi: 10.1001/jamaoncol.2018.2487] [Medline: 30027204]
48. Dietary supplements: underregulated, unknown and maybe unsafe. American Medical Association. 2022 May 11. URL: https://www.ama-assn.org/delivering-care/public-health/dietary-supplements-underregulated-unknown-and-maybe-unsafe [accessed 2024-11-14]
49. Maritess C, Small S, Waltz-Hill M. Alternative nutrition therapies in cancer patients. Semin Oncol Nurs 2005;21(3):173-176. [doi: 10.1016/j.soncn.2005.04.005] [Medline: 16092804]
50. Newman M. Is cancer fundraising fuelling quackery? BMJ 2018;362:k3829. [doi: 10.1136/bmj.k3829]
51. Zenone M, Snyder J, Bélisle-Pipon JC, Caulfield T, van Schalkwyk M, Maani N. Advertising alternative cancer treatments and approaches on meta social media platforms: content analysis. JMIR Infodemiology 2023;3:e43548 [FREE Full text] [doi: 10.2196/43548] [Medline: 37256649]
52. Carrion-Alvarez D, Tijerina-Salina PX. Fake news in COVID-19: a perspective. Health Promot Perspect 2020;10(4):290-291 [FREE Full text] [doi: 10.34172/hpp.2020.44] [Medline: 33312921]
53. Peng W, Lim S, Meng J. Persuasive strategies in online health misinformation: a systematic review. Inf Commun Soc 2022;26(11):2131-2148. [doi: 10.1080/1369118x.2022.2085615]
54. Blaskiewicz R. The Big Pharma conspiracy theory. Med Writing 2013;22(4):259-261. [doi: 10.1179/2047480613z.000000000142]
55. Derkatch C. The self-generating language of wellness and natural health. Rhetoric Health Med 2018;1(1-2):132-160. [doi: 10.5744/rhm.2018.1009]
56. The Lancet Oncology. Oncology, "fake" news, and legal liability. Lancet Oncol 2018;19(9):1135. [doi: 10.1016/S1470-2045(18)30610-7] [Medline: 30191835]
57. Nguyen A, Vu HT. Testing popular news discourse on the “echo chamber” effect: does political polarisation occur among those relying on social media as their primary politics news source? First Monday 2019;24(6):1. [doi: 10.5210/fm.v24i6.9632]
58. Warren KE, Wen LS. Measles, social media and surveillance in Baltimore City. J Public Health 2017;39(3):e73-e78. [doi: 10.1093/pubmed/fdw076] [Medline: 27521926]
59. Lee MT, Theokary C. The superstar social media influencer: exploiting linguistic style and emotional contagion over content? J Bus Res 2021;132:860-871. [doi: 10.1016/j.jbusres.2020.11.014]
60. Moyer-Gusé E. Toward a theory of entertainment persuasion: explaining the persuasive effects of entertainment-education messages. Commun Theory 2008;18(3):407-425. [doi: 10.1111/j.1468-2885.2008.00328.x]
61. Fisher WR. Narration as a human communication paradigm: the case of public moral argument. Commun Monogr 2009;51(1):1-22. [doi: 10.1080/03637758409390180]
62. Margolin DB. The theory of informative fictions: a character-based approach to false news and other misinformation. Commun Theory 2020;31(1) [FREE Full text] [doi: 10.1093/ct/qtaa002]
63. Hamby A, Daniloski K, Brinberg D. How consumer reviews persuade through narratives. J Bus Res 2015;68(6):1242-1250. [doi: 10.1016/j.jbusres.2014.11.004]
64. Fisher WR. The narrative paradigm: an elaboration. Commun Monogr 2009;52(4):347-367. [doi: 10.1080/03637758509376117]
65. Mar RA, Oatley K, Djikic M, Mullin J. Emotion and narrative fiction: interactive influences before, during, and after reading. Cogn Emot 2011;25(5):818-833. [doi: 10.1080/02699931.2010.515151] [Medline: 21824023]
66. Breil C, Böckler A. Look away to listen: the interplay of emotional context and eye contact in video conversations. Visual Cognition 2021;29(5):277-287. [doi: 10.1080/13506285.2021.1908470]
67. Burgoon JK, Birk T, Pfau M. Nonverbal behaviors, persuasion, and credibility. Human Comm Res 1990;17(1):140-169. [doi: 10.1111/j.1468-2958.1990.tb00229.x]

XSL•FO
RenderX

68. Lim H, Childs M. Visual storytelling on Instagram: branded photo narrative and the role of telepresence. J Res Interact Market 2020;14(1):33-50. [doi: 10.1108/jrim-09-2018-0115]
69. Pera R, Viglia G. Exploring how video digital storytelling builds relationship experiences. Psychol Mark 2016;33(12):1142-1150. [doi: 10.1002/mar.20951]
70. Grenard JL, Uy V, Pagán JA, Frosch DL. Seniors' perceptions of prescription drug advertisements: a pilot study of the potential impact on informed decision making. Patient Educ Couns 2011;85(1):79-84 [FREE Full text] [doi: 10.1016/j.pec.2010.10.008] [Medline: 21044826]
71. Appiah O. Rich media, poor media: the impact of audio/video vs. text/picture testimonial ads on browsers' evaluations of commercial web sites and online products. J Curr Issues Res Advert 2006;28(1):73-86. [doi: 10.1080/10641734.2006.10505192]
72. de Wit JB, Das E, Vet R. What works best: objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. Health Psychol 2008;27(1):110-115. [doi: 10.1037/0278-6133.27.1.110] [Medline: 18230021]
73. Martin BA, Wentzel D, Tomczak T. Effects of susceptibility to normative influence and type of testimonial on attitudes toward print advertising. J Advert 2008;37(1):29-43. [doi: 10.2753/joa0091-3367370103]
74. Morton H. The new visual testimonial: narrative, authenticity, and subjectivity in emerging commercial photographic practice. Multidiscip Stud Media Commun 2017;5(2):11-20. [doi: 10.17645/mac.v5i2.809]
75. Donelly B, Toscano N. The Woman Who Fooled The World: Belle Gibson's Cancer Con, and the Darkness at the Heart of the Wellness Industry. Melbourne, Australia: Scribe Publications; 2017.
76. Bainotti L, Caliandro A, Gandini A. From archive cultures to ephemeral content, and back: studying Instagram stories with digital methods. New Media Soc 2020;23(12):3656-3676. [doi: 10.1177/1461444820960071]
77. Page R. The narrative dimensions of social media storytelling: options for linearity and tellership. In: De Fina A, Georgakopoulou A, editors. The Handbook of Narrative Analysis. Hoboken, NJ: John Wiley & Sons; 2015.
78. Klein EJ. Fake news about Covid 19: communication strategies on WhatsApp in Brazil. In: Berube DM, editor. Pandemic Communication and Resilience. Cham, Switzerland: Springer; 2021.
79. Venditti S, Piredda F, Mattana W. Micronarratives as the form of contemporary communication. Design J 2017;20(sup1):S273-S282. [doi: 10.1080/14606925.2017.1352804]
80. Eisenlauer VJ. Multimodality and social actions in 'personal publishing' text: from the German 'poetry album' to web 2.0 'social network sites. In: Smith B, O'Halloran K, editors. Multimodal Studies: Exploring Issues and Domains. New York, NY: Routledge; 2011.
81. Gibson JJ. The Ecological Approach to Visual Perception: Classic Edition. Milton Park, UK: Taylor & Francis; 1979.
82. Struik LL, Bottorff JL, Baskerville NB, Oliffe JL. The Crush the Crave quit smoking app and young adult smokers: qualitative case study of affordances. JMIR Mhealth Uhealth 2018;6(6):e134 [FREE Full text] [doi: 10.2196/mhealth.9489] [Medline: 29884602]
83. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019;21(3):e11014 [FREE Full text] [doi: 10.2196/11014] [Medline: 30912754]
84. Chong I, Proctor RW. On the evolution of a radical concept: affordances according to Gibson and their subsequent use and development. Perspect Psychol Sci 2020;15(1):117-132. [doi: 10.1177/1745691619868207] [Medline: 31711365]
85. boyd D. Social network sites as networked publics: affordances, dynamics, and implications. In: A Networked Self. London, UK: Routledge; 2010.
86. Schellewald A. Understanding the popularity and affordances of TikTok through user experiences. Media Culture Soc 2023;45(8):1568-1582. [doi: 10.1177/01634437221144562]
87. Zhao H, Wagner C. How TikTok leads users to flow experience: investigating the effects of technology affordances with user experience level and video length as moderators. Internet Res 2022;33(2):820-849. [doi: 10.1108/intr-08-2021-0595]
88. Lu Y, (Cindy) Shen C. Unpacking multimodal fact-checking: features and engagement of fact-checking videos on Chinese TikTok (Douyin). Soc Media Society 2023;9(1). [doi: 10.1177/20563051221150406]
89. Peng Y, Jemmott JBIII. Feast for the eyes: effects of food perceptions and computer vision features on food photo popularity. Int J Commun 2018;12 [FREE Full text]
90. Kizilcec RF, Papadopoulos K, Sritanyaratana L. Showing face in video instruction: effects on information retention, visual attention, and affect. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON URL: https://dl.acm.org/doi/10.1145/2556288.2557207 [doi: 10.1145/2556288.2557207]
91. Li Y, Xie Y. Is a picture worth a thousand words? An empirical study of image content and social media engagement. J Mark Res 2019;57(1):1-19. [doi: 10.1177/0022243719881113]
92. Goren CC, Sarty M, Wu PY. Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics 1975;56(4):544-549. [Medline: 1165958]
93. Johnson CI, Mayer RE. An eye movement analysis of the spatial contiguity effect in multimedia learning. J Exp Psychol Appl 2012;18(2):178-191. [doi: 10.1037/a0026923] [Medline: 22309059]

XSL•FO
RenderX

94. Teixeira T, Wedel M, Pieters R. Emotion-induced engagement in internet video advertisements. J Mark Res 2012;49(2):144-159. [doi: 10.1509/jmr.10.0207]
95. Ostrovsky AM, Chen JR. TikTok and its role in COVID-19 information propagation. J Adolesc Health 2020;67(5):730 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.07.039] [Medline: 32873499]
96. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021;23(9):e30409 [FREE Full text] [doi: 10.2196/30409] [Medline: 34468327]
97. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021;23(12):e28318 [FREE Full text] [doi: 10.2196/28318] [Medline: 34931996]
98. García CA, Duncan P, Goodier M, Hunter-Green Z. #ukpolitics: how the 2024 general election has played out on TikTok. The Guardian. 2024 Jul 4. URL: https://www.theguardian.com/politics/article/2024/jul/04/ukpolitics-how-the-2024-general-election-has-played-out-on-tiktok [accessed 2024-11-14]
99. Ward J. Donald Trump joins TikTok and rapidly wins three million followers. Reuters. 2024 Jun 4. URL: https://www.reuters.com/world/us/donald-trump-joins-tiktok-video-platform-he-once-sought-ban-2024-06-02/ [accessed 2024-11-14]
100. Gilbert D. TikTok pushed young german voters toward far-right party. Wired. 2024 Jul 16. URL: https://www.wired.com/story/tiktok-german-voters-afd/ [accessed 2024-11-14]
101. Cervi L, Tejedor S, Blesa FG. TikTok and political communication: the latest frontier of politainment? A case study. Media Commun 2023;11(2):203-217. [doi: 10.17645/mac.v11i2.6390]
102. Bösch M. Broken promises: TikTok and the German election. The Mozilla Foundation. 2021. URL: https://foundation.mozilla.org/en/campaigns/tiktok-german-election-2021/ [accessed 2024-11-14]
103. Serrano JC, Papakyriakopoulos O, Hegelich S. Dancing to the partisan beat: a first analysis of political communication on TikTok. In: Proceedings of the 12th ACM Conference on Web Science. 2020 Presented at: WebSci '20; July 6-10, 2020; Southampton, UK. [doi: 10.1145/3394231.3397916]
104. Joo J, Steinert-Threlkeld ZC. Image as data: automated visual content analysis for political science. arXiv Preprint posted online October 3, 2018 [FREE Full text] [doi: 10.48550/arXiv.1810.01544]
105. Geitgey A. ageitgey / face_recognition. GitHub. 2021. URL: https://github.com/ageitgey/face_recognition [accessed 2024-11-14]
106. Zhou B. CSAILVision / places365. GitHub. 2017. URL: https://github.com/CSAILVision/places365/tree/master [accessed 2024-11-14]
107. Zhou B, Lapedriza A, Khosla A, Torralba A, Oliva A. Places. Massachusetts Institute of Technology. 2017. URL: http://places2.csail.mit.edu/ [accessed 2024-11-14]
108. Emorine B. Boris-Em / ColorKit. GitHub. 2020. URL: https://github.com/Boris-Em/ColorKit [accessed 2024-11-14]
109. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
110. Wellman ML, Holton AE, Kaphingst KA. Previvorship posting: why breast cancer previvors share their stories on social media. Health Commun 2023;38(11):2441-2449. [doi: 10.1080/10410236.2022.2074780] [Medline: 35582752]
111. Milani E, Weitkamp E, Webb P. The visual vaccine debate on Twitter: a social network analysis. Media Commun 2020;8(2). [doi: 10.17645/mac.v8i2.2847]
112. Wardle C. Misunderstanding misinformation. Issues Sci Technol 2023;39(3):38-40 [FREE Full text] [doi: 10.58875/zaud1691]
113. Fornara F, Lomicka LL. Using visual social media in language learning to investigate the role of social presence. CALICO J 2019;36(3):184-203. [doi: 10.1558/cj.37205]
114. Skafle I, Nordahl-Hansen A, Quintana DS, Wynn R, Gabarron E. Misinformation about COVID-19 vaccines on social media: rapid review. J Med Internet Res 2022;24(8):e37367 [FREE Full text] [doi: 10.2196/37367] [Medline: 35816685]

Abbreviations

RQ: research question

Edited by A Mavragani; submitted 06.05.24; peer-reviewed by L Cervi, K Wiley, D Boatman; comments to author 22.06.24; revised version received 21.08.24; accepted 22.10.24; published 09.12.24.

Please cite as:
Muenster RM, Gangi K, Margolin D
Alternative Health and Conventional Medicine Discourse About Cancer on TikTok: Computer Vision Analysis of TikTok Videos
J Med Internet Res 2024;26:e60283
URL: https://www.jmir.org/2024/1/e60283
doi:10.2196/60283
PMID:

XSL•FO
RenderX

©Roxana Mika Muenster, Kai Gangi, Drew Margolin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Spillover Effects of Paid Functions on Physicians' Unpaid Knowledge Activities: Quasi-Experimental Approach

Xuan Liu[1*], PhD; Xiaotong Chi[1*], MSc; Ming Chen[1], MSc; Wen Sun[1]; Jia Li[1], PhD

East China University of Science and Technology, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Xuan Liu, PhD
East China University of Science and Technology
Meilong Rd, No 130
Shanghai, 200237
China
Phone: 86 021 64253177
Email: xuanliu@ecust.edu.cn

Abstract

Background: To promote sustained contributions by physicians to online health care communities, these platforms have introduced a content payment model that offers economic incentives for physicians' online knowledge activities. However, the impact of these paid features on unpaid knowledge activities remains unexplored.

Objective: This study investigated how the introduction of economic incentives in online medical communities affects physicians' unpaid knowledge activities in the community.

Methods: The data for this study were obtained from the Haodf Online platform in China, which has implemented paid scenarios for its science popularization function, providing economic benefits to physicians. The dataset, which comprises panel data, includes 7453 physicians who participated in both paid and unpaid knowledge contributions on the website. This study examined the impact of paid knowledge activities on physicians' free knowledge contributions, focusing on dimensions including knowledge quantity, quality, and diversity. To address the timing discrepancies in physicians' participation in paid activities, we used a quasi-experimental design that combined the approach of propensity score matching and multi-period difference in differences.

Results: In the balance test results of the propensity score matching, the absolute values of the SDs of all matching variables were mostly <5% after matching, ensuring the accuracy of the results obtained from the difference in differences method. This study found that participation in paid knowledge activities had a positive spillover effect on physicians' free knowledge contributions, which manifested in the increase in post quantity (473.1%; $P<.001$), article length (108%; $P=.009$), function word frequency (0.6%; $P=.001$), causal word frequency (0.2%; $P<.001$), and content information entropy (6.6%; $P=.006$). The paid function led to a decrease in the consistency between titles and content (–115.5%; $P<.001$).

Conclusions: The findings of this study contribute to the existing literature on the impact of economic incentives in the medical context. For the platform, providing economic incentives to physicians can have positive significance in promoting the development of the platform's knowledge ecosystem and can effectively encourage physicians to contribute to both paid and free knowledge activities. This study provides a valuable reference for the platform to introduce a paid knowledge model, which is beneficial to the sustainable development of the platform.

(J Med Internet Res 2024;26:e58688) doi:10.2196/58688

KEYWORDS

health knowledge contribution; economic incentives; diversity; propensity score matching; multi-period difference in differences

Introduction

Background

The prosperity of online medical communities depends on the sustained participation of physicians, especially their knowledge contribution to the community. To encourage physicians' contributions, online medical communities have implemented a content-charging model to provide economic rewards for physicians' online knowledge activities. The success of platforms with similar revenue models, such as Quora and Zhihu

XSL•FO
RenderX

Live, demonstrates that economic incentives enhance users' performance in paid knowledge activities [1].

Providing incentives for physicians' knowledge contributions in online medical communities is crucial for both physicians and patients. For physicians, offering channels to monetize their knowledge enhances their earnings from online contributions, increasing their willingness to share and efficiency in sharing knowledge. For patients, the availability of paid knowledge supplements their access to information, providing higher-quality and more specialized medical knowledge. Figure 1 illustrates the paid and free science popularization pages on the Haodf platform in China.

Figure 1. Example of physicians' (A) paid and (B) unpaid knowledge contributions.

Incentive mechanisms are one of the most effective measures to encourage active participation in knowledge contribution within online communities. They hold significant value for the thriving development of online health communities. Effective incentive mechanisms can significantly enhance user engagement. Common online community incentives include economic incentives (such as monetary rewards and point systems) and social incentives (such as online reputation). This study focused on economic incentives, specifically manifested as patients paying fees to access paid articles published by physicians. The fees generated from this process directly translate into economic incentives for physicians. Thus, the core focus of this study was on the impact of economic incentives in the context of online health communities. Currently, paid knowledge and free knowledge coexist in online medical communities, and previous studies have found that economic incentives directly affect users' knowledge contributions [2,3] and sustained participation [4]. However, it is also crucial to understand how economic incentives influence unpaid contributions to knowledge activities, which may affect overall community performance [5]. On the one hand, free content, which has no payment threshold and accumulates significantly, can reach a wider range of users and maintain user engagement on the platform. On the other hand, medical knowledge products are of public interest, and free medical knowledge provides a barrier-free channel for users to obtain medical information, effectively protecting the basic interests of patients. Therefore, if economic incentives result in physicians contributing less to free content, it would be detrimental to the long-term development of the community. Conversely, if economic incentives encourage physicians to further contribute to free content, it would significantly enhance physician-patient interactions within the community. Previous studies have confirmed that, in generalized knowledge communities, economic incentives have a spillover effect on nonincentive activities, boosting users' overall participation in knowledge communities [5].

Although many scholars have conducted research in this field, several key issues are still unresolved. First, the findings from generalized knowledge communities may not apply to medical contexts. In online medical communities, physicians face significant pressures from offline diagnoses and clinical duties, limiting their time and energy for online knowledge services. Thus, the conclusions drawn from generalized knowledge communities may not be applicable. Second, there is no consensus on the spillover effects of economic incentives. Existing studies have not yet reached a unified conclusion on whether economic incentives have spillover effects on free service performance and the specific direction of the impact. The answer varies across different research contexts. In general, the impacts of participating in paid knowledge activities on contributions to free knowledge activities can be summarized as being both positive and negative. Positive impacts include (1) reputation establishment (engaging in free knowledge activities without thresholds can enhance reputation in the community and attract patients) [6], (2) community contribution (reciprocity motivates physicians to contribute to free knowledge activities due to the benefits received from paid activities) [7], and (3) exposure opportunity (the creation threshold of free knowledge is relatively low, and by sharing free knowledge, physicians can quickly obtain more exposure opportunities) [8].

XSL•FO
RenderX

Negative impacts include (1) the substitution effect (participation in paid knowledge activities leads to physicians' inability to allocate sufficient time and energy for free knowledge activities) [9] and (2) the crowding-out effect (charging for knowledge activities weakens physicians' intrinsic motivation to contribute freely, thus reducing the prosocial quality of physicians' participation in free knowledge contribution) [10,11].

How does physicians' participation in paid knowledge activities affect their contribution to free knowledge activities? For platforms, it is crucial to develop knowledge-paying models without harming the interests of the vast number of free knowledge-sharing groups and balance physicians' free and paid knowledge contributions. This study examined the experience of a Chinese leading medical platform, Haodf Online. To clarify the differences in physicians' free content contribution before and after economic incentives, this study used a quasi-experimental design that combines the propensity score matching (PSM) technique with the multi-period difference in differences (DID) analysis, and investigated on the spillover effect of paid knowledge services on physicians' free knowledge services. The propensity score matching (PSM) method was used to construct comparable groups of paid and nonpaid physicians to solve the sample selection bias problem. The multi-period difference in differences (DID) method was used to compare the changes in physicians' free content contribution before and after economic incentives. Its advantage lies in eliminating individual and time trends and accounting for policy stimuli across multiple time points. This research will contribute to the application of economic incentives in online medical communities.

Literature Review

Physicians' Online Knowledge Contribution

Research on online medical communities has emphasized the importance of knowledge services and knowledge activities. Scholars have extensively studied patient participation in these communities, finding that a primary motivation is to obtain informational support. Knowledge exchange and knowledge communication on the platform can improve the health status of patients. Knowledge exchange on these platforms provides answers with more quantity and more clinical significance than search engine results [12].

Nowadays, scholars have focused on the crucial role of physicians as professional service providers in promoting the prosperity of online medical communities. Current research primarily examines the factors and mechanisms influencing physicians' knowledge contribution and the impact of knowledge exchange on service performance. Scholars have primarily studied the influencing factors and internal mechanisms behind physicians' knowledge contribution while noting a trend from free sharing to paid knowledge exchange in online medical communities. Studies have found that both intrinsic and extrinsic motivations drive users to share knowledge voluntarily. Specifically, providing economic incentives for free question-and-answer services enhances users' willingness to share their expertise [13]. Yang et al [14] discovered that both personal and social motivations positively influence physicians' contributions to online health communities, with physicians' professional status moderating this relationship. Some scholars have attempted to explore the connection between general free knowledge sharing and specific paid medical services by examining patients' transition from seeking free information to engaging in paid consultations. For instance, Meng et al [15] confirmed the significant role played by reputation-establishing mechanisms in facilitating this transition. As the paid knowledge model in the medical field becomes commercialized, some scholars have started conducting more comprehensive research on payment models for medical expertise. For example, Qi et al [1] studied the external impact of physicians' paid question-and-answer services, finding improvements in both the quantity and quality of their unpaid knowledge contributions. Regarding the influence of knowledge exchange between physicians and patients, establishing knowledge interaction significantly affects patients' perceptions of medical services. Khurana et al [16] found that physicians' active participation in responding to queries on online medical platforms can mitigate information asymmetry between physicians and patients, thereby increasing physician recommendations.

Spillover Effect of Incentives

Incentive mechanisms play a pivotal role in online communities, significantly influencing user engagement and contribution levels [17]. These mechanisms can be broadly categorized into intrinsic and extrinsic incentives. Intrinsic incentives arise from the internal satisfaction that users derive from participation, such as self-fulfillment [18], personal interests, knowledge sharing, and social interaction. Conversely, extrinsic incentives include material and social rewards. Material rewards motivate users through economic incentives, gift cards, and coupons. Social rewards encompass reputation [18], honors, ranking systems, and badges [19], enhancing users' status and recognition within the community. Incentive mechanisms in online communities should be tailored to the community's nature and user characteristics to promote sustainable and healthy development.

Recently, the impact of economic incentives on content contribution has garnered significant attention from scholars. Most previous question-and-answer platforms were free, focusing solely on free knowledge sharing. Therefore, compared with earlier free knowledge exchange platforms, the leading role of economic incentives in knowledge sharing remains to be further explored. Typically, economic incentives are used as a direct means to stimulate user content contribution. In previous studies, the primary forms of economic incentives have included direct economic rewards, coupons, product discounts, and online currency rewards.

In empirical studies, conclusions on the effects of economic incentives are not uniform. Some scholars believe that, in the short term, economic incentives on platforms have a positive effect, potentially increasing the number of comments [3,6]. Sun and Zhu [20] found that revenue from shared advertising increased the number and quality of blog posts, prompting bloggers to shift content to more popular topics. Regarding quality, economic incentives are believed to have no impact on

XSL•FO
RenderX

performance [21], such as the quality of comments [3,6] or the quality of stock recommendations [22].

However, some studies have found that economic incentives are not effective and may even have a negative effect. Chen et al [23] found that price and tips had no significant effect on the length and quality of answers in question-and-answer communities. In addition, some scholars have found that economic incentives have a negative effect, with the introduction of economic incentives leading to a decrease in the number of online comments and the total contribution to the community [24].

Beyond examining how economic incentives directly affect paid activities, scholars have also focused on the interaction between incentive-based activities and other unpaid activities. There is no consensus on the spillover effect of economic incentives on free knowledge contributions. In the short term, economic incentives can enhance the quantity and quality of users' free knowledge contributions [1,6] and increase users' overall participation in the knowledge community but have no significant effect on users' knowledge-seeking behavior [5]. In the long term, this effect is difficult to sustain [6]. Some studies have also found that economic incentives negatively affect users' subsequent prosocial contributions, reducing their prosocial performance and efforts [25]. Regarding the impact of free knowledge sharing on paid knowledge activities, some scholars have found that free knowledge contributions can enhance user trust and promote user knowledge purchasing [15]. However, some scholars believe that, due to the limited time and energy of service providers, the spillover effect of free activities will turn negative after a certain point [26].

Research Gap

Despite extensive research in this field, several critical issues remain unresolved.

Generalized knowledge community findings may not apply to medical scenarios in which physicians' consultation time is a significant cost [27]. Given the substantial offline consultation and clinical workload, physicians have limited time and energy to engage in online knowledge services, suggesting that the findings from generalized knowledge communities may not hold true in medical contexts. Second, there are 2 distinct mechanisms through which economic incentives impact free knowledge activities, and their effects may vary depending on the research context. Third, there is no consensus on the spillover effect of economic incentives, especially regarding the quality and diversity of contributions.

This study examined Haodf, a leading online medical platform in China, to explore how economic incentives influence physicians' free knowledge activities using the PSM and multi-period DID methods.

Methods

Theoretical Foundation and Hypothesis Development

Motivation theory is a fundamental framework for analyzing user behavior patterns in online communities [28-30], categorizing user motivations as intrinsic and extrinsic [31]. Intrinsic motivation stems from the inherent satisfaction or enjoyment derived from engaging in specific behaviors, serving as a core driving force that prompts users to participate spontaneously. In contrast, extrinsic motivation is associated with external rewards or incentives that prompt users to undertake particular actions due to external stimuli. Existing research identifies 2 possible pivotal effects. First, overemphasizing or primarily relying on extrinsic incentives can trigger a "crowding-out effect," in which extrinsic incentives undermine or even replace intrinsic motivation. Second, extrinsic incentives can trigger a "spillover effect" if they aim to enhance individuals' perceived competence without compromising their sense of autonomy. That is to say, when extrinsic incentives are regarded as acknowledgments of individual effort and contribution, they are more likely to bolster intrinsic motivation rather than diminish it [29,31].

This study focused on online health knowledge contribution, distinguishing between free and paid health knowledge contributions. In this case, physicians who participate in free knowledge contributions are primarily driven by intrinsic motivation, which arises from their passion for knowledge sharing, professional achievement, and social responsibility. Conversely, physicians' involvement in paid knowledge contributions is largely influenced by extrinsic motivation, namely, economic incentives. Notably, given the typically strong prosocial orientation of physicians, their behavior in sharing knowledge often stems from a robust intrinsic motivation. Thus, we believe that the rewards from paid contributions not only elevate their sense of achievement and self-confidence but also further enhance their intrinsic motivation, prompting them to engage more actively in unpaid knowledge contributions. Moreover, receiving economic incentives may prompt physicians to devote additional time and effort to free knowledge activities to avoid being perceived as solely motivated by economic incentives and establish a positive reputation [23]. Consequently, we hypothesize that, on online medical platforms, physicians' engagement in paid knowledge activities will produce a spillover effect on their free knowledge activities rather than a crowding-out effect. This effect is manifested through a significant increase in the quantity of free knowledge contributions, notable improvements in their quality, and an expansion of content diversity. To evaluate this, this study developed measurement indexes for content contributions from 3 dimensions: quantity, quality, and diversity. Quantity directly measures the frequency of physicians' participation in knowledge services, serving as the most fundamental evaluation metric. It reflects the physicians' enthusiasm and level of engagement in knowledge contribution. Quality pertains to the professionalism and accuracy of the physicians' knowledge service content, with high-quality services meeting readers' medical needs. Diversity assesses the richness of the physicians' knowledge services, evaluating whether they provide a wide range of content in various formats to address different issues, emphasizing variety and comprehensiveness.

The quantity of physicians' free article contributions directly reflects the intensity of physicians' online knowledge contributions, which is crucial for patients to evaluate physicians' online performance. On the basis of this, this study

XSL•FO
RenderX

proposed that participation in paid knowledge activities has a positive spillover effect on the quantity of free content contributions made by physicians (hypothesis 1).

As important as quantity, the quality of free articles is also vital for the sustainable development of online medical platforms. By writing high-quality articles and showcasing their profound medical knowledge and rich clinical experience, physicians can establish a professional image and win the trust and respect of patients. High-quality content can attract more user attention and participation, accumulate a good reputation and reputation for the platform, and lay a solid foundation for future sustainable development.

Whether physicians will improve the quality of free content when they receive economic incentives has not yet been conclusive. Some studies have found that small incentives have a negative impact on contribution quality [32], such as reducing the length of comments [33] and conveying one-sided emotions [33,34], which is due to the weakening of intrinsic motivation. However, from the perspective of reputation building, improving the quality of free content helps physicians establish a reputation for quality content services in the community, thus reducing patients' uncertainty before purchasing content, which will have a positive impact on improving physicians' paid services. On the basis of this, this study proposed that participation in paid knowledge activities has a positive spillover effect on the quality of free content contributions by physicians (hypothesis 2).

Diversity is the distribution of differences in certain attributes among members [35]. Diversity is important for community performance, such as user diversity promoting community learning [36] and improving content collaboration efficiency and outcomes [35]. Question-and-answer communities that encourage participants to answer more new questions have a good performance in answering questions, creating long-term value and improving difficulty resolution.

Diversity in online medical communities can be divided into media diversity and content diversity. Media diversity refers to the diversity of media adopted by physicians to spread medical knowledge. High media diversity indicates that physicians have adopted more ways to participate in knowledge contribution. Richer media are more likely to attract consumers' attention and affect their evaluation of content [37]. The existence of economic incentives may encourage physicians to adopt more diverse means to present content to patients.

Content diversity refers to the differences in vocabulary involved in the creation of medical knowledge by physicians. These differences in vocabulary can affect the content richness of the physicians' articles. When the content diversity of an article is high, the article may have richer and more comprehensive explanations [38]. Research has found that economic incentives increase the content diversity of online community discussions [22]. On the basis of this, this study proposed that participation in paid knowledge activities has a positive spillover effect on the diversity of free content contributions by physicians (hypothesis 3).

The research model is shown in Figure 2.

Figure 2. Research model of the spillover effects of the paid knowledge function. H: hypothesis.

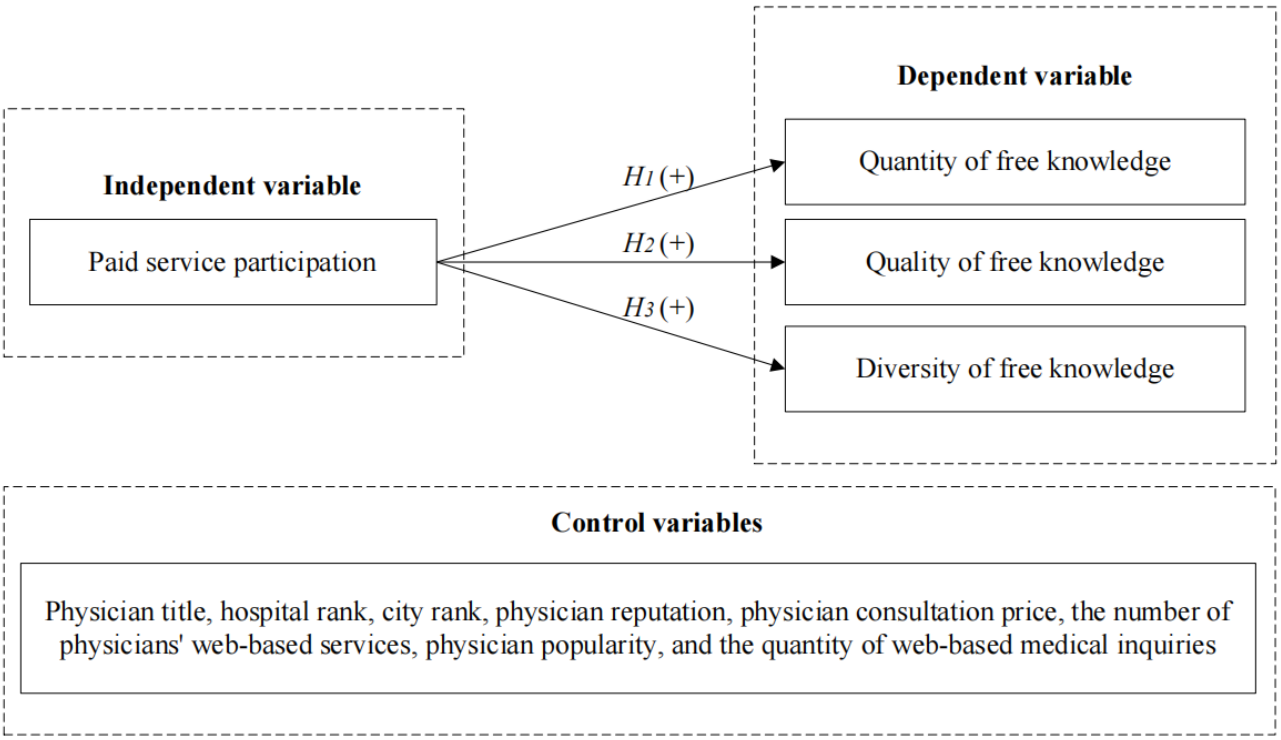

Sample Description

This study used the health knowledge column on the Haodf Online platform [39] as the data source. Haodf is one of the largest online medical websites in China, where many physicians provide diagnosis and treatment consultation services.

XSL•FO
RenderX

The paid medical knowledge service of the Haodf Online platform was introduced in 2016. The observation period for this study spanned 2016 to 2021. Due to the limited participation of physicians in knowledge activities in the short term, short intervals cannot effectively capture physicians' knowledge activities and trends. Therefore, this study examined the impact of economic incentives on physicians' knowledge activities over a relatively long period. Our time frame is divided into 6-month intervals [6], with the first interval covering data from January 1, 2016, to May 31, 2016, designated as spell 1. Each phase was set to 6 months to allow for a more accurate capture and analysis of the trends in physicians' participation in online knowledge activities. Finally, the entire observation window is divided into 12 spells.

First, we comprehensively and systematically collected all the articles shown in the "Popular Science and Knowledge" section of the Haodf Online platform, totaling 63,151 articles. Subsequently, the authors of these articles, namely, the physicians, were the focus of our study. These physicians come from various departments covering the major medical specialties on the platform. After data cleaning, which involved removing physicians who had left the community and whose basic information could no longer be accessed, a total of 14,459 physicians were retained. Of these 14,459 physicians, 2070 (14.32%) participated in paid knowledge activities from 2016 to 2021, which constitutes the focal sample of this study. Using PSM with a 1:3 nearest neighbor matching technique, we matched physicians who participated in paid contributions with those who did not on a period-by-period basis, resulting in a comparison sample of 7453 physicians. Within this group of 7453 physicians, 2070 (27.77%) attempted to publish paid articles (although not necessarily in every period), whereas 5383 (72.23%) only published free articles. After conducting PSM on a period-by-period basis over the 6-year observation window, a total of 15,820 physician samples were obtained, including 6617 (41.83%) physician samples who participated in paid knowledge activities and 9203 (58.17%) physician samples who participated in free knowledge activities. The sample in this study is representative of the overall physician population in the Haodf Online medical community.

Variable Definition

Dependent Variable

This study assessed the effectiveness of physicians' contributions to free knowledge from 3 dimensions: quantity, quality, and diversity.

Quantity Dimension

The quantity dimension measures physicians' contributions by the number of free articles published in the current period (*content_num*). Research indicates that economic incentives increase the number of content contributions in the community and cause widespread interest [22]. The community mainly takes the form of free articles without thresholds as knowledge contributions, and these free articles may be one of the important sources of physicians' exposure in the community. When physicians participate in paid services, they may increase the number of free knowledge contributions to improve their exposure in the community and provide patients with a channel to evaluate the quality of the content.

Quality Dimension

The quality of free articles pertains to the professionalism and accuracy of the knowledge content produced by physicians, which can be evaluated from dimensions such as completeness, consistency, social consensus, accuracy, and usefulness [40]. Considering that the assessment of the accuracy of medical content requires the participation of medical professionals and the measurement of the usefulness of the information needs to take into account the needs and perceptions of patients, we omitted the 2 dimensions of accuracy and usefulness due to the insufficient support of secondhand data. This study selected the following 3 dimensions: completeness, consistency, and social consensus.

Completeness (*content_len*) is measured through the average of the number of characters in all free articles published by physicians in the current period, reflecting the detail and thoroughness of the information. The length of the information is directly related to the detail of the information. For example, research has found that the length of answers can best predict the quality of answers [41]. This study believes that, in the medical field, the length of answers is an important indicator, and the total number of words can be used to measure the completeness of answers.

Consistency (*simi_fuzz*) is measured through the similarity between the title and the full text [42], serving as an important indicator of information quality [43]. Whether the content is relevant to the topic directly affects the accurate transmission of medical content. This study used the *FuzzyWuzzy* package in Python (Python Software Foundation) to calculate the consistency score of the 2 groups of content. This toolkit is used for fuzzy string matching, and the difference between 2 sequences is calculated according to the edit distance algorithm.

Social consensus evaluates the common understanding among readers. This study considered whether the content contribution of physicians can promote communication and understanding from the perspective of suitability. Suitability is measured through the pragmatic features of free knowledge, which refer to language characteristics related to context, the speaker's intention, the listener's understanding, and sociocultural background in language use, reflecting the medical professionalism, logic, and readability of the article. From the perspective of the logic of content expression, due to economic incentives, physicians should enhance their logic to improve the experience of patients with reading free content. From the perspective of professional expression of content, due to economic incentives, physicians may enhance their professionalism to provide users with more professional medical content.

Pragmatic features were calculated using the TextMind software (Computational Cyber-Psychology Lab, Institute of Psychology, Chinese Academy of Sciences) [44,45], which categorizes words based on Linguistic Inquiry and Word Count. This method compares text words with dictionaries, calculating the proportion of different word categories in the text [46]. The relevant

XSL•FO
RenderX

categories for this study included biological words, health words, functional words, and causal words. The proportion of biological (*bio*) and health (*health*) words reflects the medical professionalism of the content, improving the efficiency of physician-patient communication [45] and improving patient reading experiences [46]. The proportion of functional (*funct)* and causal (*cause*) words captures the language style and logical coherence of the content [47]. Textbox 1 shows the Linguistic Inquiry and Word Count word categories involved in this study and corresponding examples.

Textbox 1. Linguistic Inquiry and Word Count word categories in the Haodf online medical community.

Category and examples

- *bio: dizzy, hug,* and *sweat*
- *health: insomnia, doctor,* and *exercise*
- *funct: maybe, many,* and *those*
- *cause: cause, make,* and *become*

Diversity Dimension

In online health communities, the diversity of knowledge services underscores the richness and comprehensiveness of knowledge. The diversity of physicians' participation in knowledge services is primarily reflected in the presentation formats and the content they include. Physicians can use various media, such as images and text, to vividly present medical knowledge and cover a wide range of topics to ensure content richness. Therefore, the diversity dimension of physicians' free knowledge contributions is measured through media diversity and content diversity. The number of media types and the information entropy of full text content were selected as the evaluation indexes for this dimension.

Media diversity (*tab_type*) refers to the variety of media types involved in the current free articles of physicians. The articles can take various forms, including popular science graphics, audio, and video. This study calculated the number of media types used in each physician's article as a measure of media diversity.

Content diversity (*shanno_ent*) is measured through the information entropy of the current free articles of physicians. Information entropy represents the degree of uncertainty of the value, describing the amount of information content [48]. Higher entropy values indicate a higher degree of information complexity and greater information content, whereas lower entropy values indicate lower complexity and less information content. Information entropy has been applied in text sentiment classification [49], social influence analysis [50], comment usefulness analysis, personalized recommendation [51], and other fields. In this study, we preprocessed the free article content by cleaning and segmenting the text, counting the word frequency of each term, and then applying equations 1 and 2 to calculate the information entropy of each article. Finally, we calculated the average value of the information entropy of all free articles published by a physician in the current period as the content information entropy of the physician's free knowledge contributions in the current month.

In the aforementioned formulas, t represents the tth period, $i$represents the ith physician, m represents the mth word, and N_{it} represents the frequency of words in the free article published by physician i in the tth period.

Independent Variable

The main explanatory variable in this study was the participation of physicians in paid knowledge services, represented as $paid_service_{it}$. It is a dummy variable to depict the participation of physicians in paid knowledge services. If the physician participated in paid knowledge services in the current period, the variable was coded as 1; otherwise, it was coded as 0.

Matching Variables

Overview

In PSM, matching variables are used to estimate the propensity score for each individual. These propensity scores measure the probability of an individual receiving the treatment or not (treatment or control). Matching variables typically include covariates related to the outcome variable, which help control for confounding factors, making the treatment and control groups as similar as possible on these variables. We synthesized previous research related to online health communities, identifying 3 primary components for matching variables: physician professional characteristics pertaining to their competency [52], online effort as an indicator of their experience and engagement [53], and online physician reviews reflecting patient evaluations [54].

Physician Professional Characteristics

Physician professional characteristics refer to their knowledge, skills, experience, and professional qualities in the medical field. Due to the variations in evaluating physicians' titles across regions and hospitals, the professional capability dimension comprehensively considers physicians' titles, the grade of the hospitals they work at [53], and the grade of the cities they work in [55].

Physician title (*title_add*) adds the clinical and academic titles of physicians. Physicians with higher titles, who are more likely to participate in online services, are better equipped to provide quality services [56]. Physician title is an important status capital, positively impacting physicians' online participation performance [57,58].

XSL•FO
RenderX

Hospital rank (*hospital_rank*) ranks the hospitals where physicians work on a scale from 1 to 6, from low to high (first level=1, first class=2, second level=3, second class=4, third level=5, and third class=6). Hospital rank symbolizes the quality of the services provided by physicians and is another important offline capital for them [56].

City rank (*province_order*) is divided into 5 grades according to the development status of the province where the hospital is located and is recorded as high-quality area (5), medium-high–quality area (4), medium-quality area (3), medium-low–quality area (2), and element-lacking area (1). Cities with a higher economic development quality indicate that they have good resources, enabling physicians to gain better social advantages [58].

Online Effort

Online effort comprehensively measures physicians' medical service performance on the platform, including the number of services offered, online consultation volumes. The website uses these metrics to reflect physicians' activity and engagement levels [53]. This dimension also includes physicians' consultation price, reflecting their value and experience.

The number of physicians' online services (*service_num*) counts online services such as online diagnosis and treatment, physician team, private physician, and remote diagnosis. Considering the limited time and energy of physicians, they need to determine how to allocate their time to various tasks [59]. A wide service scope may lead to a decrease in the time that physicians allocate to free content creation, thus having a negative impact on the quantity, quality, and diversity of physicians' free content.

The quantity of online medical inquiries (*inquiry_ln*) is measured using the logarithm of the quantity of online medical inquiries, reflecting the experience and participation of physicians in online services. The quantity of online medical inquiries is also a form of online knowledge contribution by physicians, which may have 2 directions of influence on their free knowledge contribution. On the one hand, online medical inquiries will occupy physicians' time for online services, which could reduce the time available for them to contribute free content. On the other hand, physicians with more online inquiries are more active online and more likely to provide free knowledge for the community.

Physicians' consultation price (*wenzhen_prince_ln*) is directly displayed on their home page and measures their assessment of their self-worth [56]. Physicians' medical services are rewarded. For rational people's consideration, physicians may pursue economic interests and put more effort into rewarding diagnosis and treatment services, which will have a crowding-out effect on free content contribution [26].

Online Physician Reviews

Online physician reviews refer to patient opinions [54], including physician popularity and reputation [60].

Physician popularity (*hot*) comes from the scores given by patients to physicians, ranging from 0 to 5. Physician popularity reflects the degree of recognition of physicians by patients and the personal brand influence of physicians and may have a positive impact on physicians' online knowledge contribution. The reasons are as follows. First, the recommendation popularity to some extent conveys the degree of satisfaction of patients with physicians' medical technology and communication level. Physicians with high popularity are more capable of providing a large amount of high-quality and rich free knowledge. Second, to maintain high scores from patients, physicians with high recommendation popularity are more motivated to give back to the community.

Physician reputation (*vote_each_month_ln*) is the number of patient votes received by physicians. It is the feedback and evaluation from patients of a physician's service, indicating patient satisfaction. Physician reputation serves as an online status capital for physicians, motivating them to strengthen their unpaid content contributions to receive higher evaluations [56].

Table 1 shows the list of variables.

XSL•FO
RenderX

Table 1. Variable explanation.

Dimension and variable	Meaning	Explanation
Dependent variables		
Quantity		
content_num	Post number	Number of free articles posted by physicians
Quality		
Completeness—content_len	Character count of free articles	The total number of characters in a free article posted by a physician
Consistency—simi_fuzz	Free article theme compliance	The similarity between the title and the content of free articles published by physicians (fuzzy set); the similarity score between the title and the content is calculated using the Python (Python Software Foundation) *FuzzyWuzzy* package
Social consensus—funct, social consensus—health, social consensus—bio, and social consensus—cause	Pragmatic feature	Use TextMind (Computational Cyber-Psychology Lab, Institute of Psychology, Chinese Academy of Sciences) to calculate the pragmatic features of free articles published by physicians; *funct* is the frequency of functional words in free articles published by physicians, *health* is the frequency of health words in free articles published by physicians, *bio* is the frequency of biological words in free articles published by physicians, and *cause* is the frequency of causal words in free articles published by physicians
Diversity		
tab_type	Number of media types	Number of free articles with audio, video, and other multimedia content published by physicians
shanno_ent	Content information entropy	Equations 1 and 2 provide specific calculations
Independent variable		
$paid_service_{it}$	The participation of physicians in paid knowledge services	A dummy variable—1 if participating and 0 otherwise
Matching variables		
Physician professional characteristics		
title_add	Physician title	The clinical and academic titles of physicians
hospital_rank	The rank of the hospital where the physician works	The grade of the hospitals where the physicians work, from low to high, is recorded as first level=1, first class=2, second level=3, second class=4, third level=5, and third class=6
province_order	The rank of the city where the physician works	Divided into 5 grades according to the development status of the province where the hospital is located and recorded as high-quality area=5, medium-high–quality area=4, medium-quality area=3, medium-low–quality area=2, and element-lacking area=1
Online effort		
service_num	The number of physicians' online services	The number of online services, such as online diagnosis and treatment, physician team, private physician, and remote diagnosis
inquiry_ln	The quantity of online medical inquiries	The logarithm of the quantity of online medical inquiries
wenzhen_price_ln	Physicians' consultation price	The price of a single consultation provided by the physician
Online physician reviews		
hot	Physician popularity	The scores given by patients to physicians, ranging from 0 to 5, with higher values indicating greater popularity
vote_each_month_ln	Physician reputation	The number of votes the physician has received from patients

XSL•FO
RenderX

Model Construction

This study used a quasi-experimental design that combines PSM with DID to study the impact of physicians' participation in paid knowledge activities on their performance in free knowledge activities.

PSM Was Conducted to Reduce Selection Bias

PSM is a statistical method used to address data bias and confounding variables in observational studies, commonly used for causal inference. The presence of confounding variables can lead to an inaccurate estimation of treatment effects as certain characteristics may simultaneously influence both the likelihood of an individual receiving treatment and the study outcomes. This results in differences between the treatment and control groups that are not solely attributable to the treatment itself. Selection bias is also a significant concern. If the treatment and control groups differ substantially in baseline characteristics, these differences may arise from unobserved or difficult-to-quantify factors, leading to erroneous conclusions about the treatment effect. Directly comparing the 2 groups may, consequently, overestimate or underestimate the true impact of the treatment. The principle of PSM involves calculating the probability (propensity score) of each individual receiving a certain treatment (such as participating in a project or receiving a treatment) and then matching individuals in the treatment group with those in the control group who have similar propensity scores to reduce bias caused by confounding variables in the observed data, thus more accurately estimating the treatment effect. Common matching methods include one-to-one matching, one-to-many matching, one-to-many nearest neighbor matching, and match with replacement. One-to-one matching pairs each treatment group individual with a single closest control group individual, whereas one-to-many matching pairs each treatment group individual with multiple closest control group individuals. One-to-many nearest neighbor matching selects multiple closest control group individuals for each treatment group individual and chooses one for matching, improving matching accuracy. Match with replacement allows a control group individual to be matched with multiple treatment group individuals, enhancing flexibility. To improve the stability and accuracy of matched samples, this study used one-to-three nearest neighbor matching with replacement.

As physicians' choice to participate in knowledge-based payment services is likely nonrandom and associated with factors such as offline capital and online service performance, sample selection bias may arise. To eliminate the influence of other characteristic differences between physicians participating in paid knowledge services and those not participating, this study initially used PSM to match samples and estimate the probability of physicians choosing paid knowledge services based on logit regression analysis [61].

Given a set of observable variables Z, the conditional probability that a physician i will enter the treatment group can be expressed as follows:

In this formula, $treat_i$ represents whether the physician has entered the treatment group. When $treat$=1, it indicates that the physician has been treated; otherwise, it is 0. The matching variables Z are selected from the following: physician title, hospital rank, city rank, physician reputation, physicians' consultation price, number of physicians' online services, physician popularity, and quantity of online medical inquiries.

After conducting PSM, the balance test is used to examine the effectiveness of the matching process. It checks whether the treatment and control groups have achieved a good balance in terms of covariates. The balance test is typically performed through statistical methods such as the 2-tailed t test and chi-square test, which compare the means, proportions, and other statistics of the covariates between the treatment and control groups before and after matching. If the differences in covariates between the 2 groups are not significant after matching, it is considered that a good balance has been achieved. Therefore, the balance test is one of the key indicators for evaluating the effectiveness of PSM results.

The Multi-Period DID Method Was Used to Estimate the Effect of Physicians' Participation in Paid Knowledge Activities

The multi-period DID method is an econometric analysis technique developed from the traditional DID approach. It is used to evaluate the average causal effect of a policy or event on both the treatment and control groups across multiple time points. The principle of this method lies in comparing the changes within the treatment group before and after policy implementation (across multiple periods) with the changes within the control group during the same time frame. This allows for the estimation of the policy or event's effect. This method effectively accounts for unobservable individual heterogeneity between the treatment and control groups as well as other time-varying trends, thus providing more accurate causal inferences.

As physicians joining paid knowledge services in this study may have occurred at different time points, to solve the endogeneity problem, we used the multi-period DID method for research, which takes into account the differences in the times when physicians entered the paid knowledge service. The DID method absorbs the unobserved differences between the treatment and control groups and eliminates the missing time trend related to physicians' participation in paid knowledge services, which is widely used in policy evaluation. For the evaluation of multi-period policies, this study adopted the multi-period DID method [62], which can reduce the sample loss, include the data of physician participation in paid knowledge services in multiple periods, and make the results more general. In addition, physicians choose to participate in paid knowledge activities at different points in time, which poses multiple shocks to physicians' online knowledge contribution. When the community launches paid knowledge activities, physicians may not immediately participate but may choose to participate after a period of observation. Therefore, the design model is as follows:

XSL•FO
RenderX

i represents the physician, t represents time, μ_i is the individual fixed effect, τ_t is the fixed effect of time, and $control_{it}$ represents other control variables. According to the basic idea of multi-time DID, the variable coefficient β_1 reflects the change in physicians' participation in free knowledge activities before and after the launch of paid knowledge activities, which is the main parameter to be estimated in this study.

Ethical Considerations

This study did not require ethics approval as the authors only used publicly available content on the Haodf website and these data did not involve patient information.

Results

Results of PSM

In this study, we used the one-to-three nearest neighbor matching method with replacement for PSM [63]. In addition, in view of the time distribution of the articles published by physicians in each half-year span from 2016 to 2021, the year-by-year matching method (each spell as a *year*, representing half a year in our case) was adopted to find the matching control group for the treatment group in each spell. Taking the first period (ie, the first half of 2016) as an example, the predicted probability value of each physician's engagement in free knowledge activities was first calculated in combination with the observable matching variables, and then 3 control group physicians who were not involved in paid knowledge activities were found for each physician participating in paid knowledge activities (the treatment group). After matching, physicians who were not successfully matched were excluded, resulting in 336 physicians in the treatment group for the first period. This method was consistently applied to subsequent periods, as shown in Table 2. After 12 periods, we obtained a total of 15,820 samples.

Table 2. Data situation of each period after propensity score matching.

Time dummy variable (spell_dummy)	Physicians, n	Participation in payment, n	Not participating in payment, n
1	576	240	336
2	612	263	349
3	978	409	569
4	1809	758	1051
5	2102	919	1183
6	1840	801	1039
7	2145	878	1267
8	1812	757	1055
9	1793	742	1051
10	893	346	547
11	734	290	444
12	526	214	312

Following PSM, we conducted a balance test using the 2-tailed t test method, and the results are presented in Table 3. The test results in Table 3 show that the absolute values of the bias of all matching variables were basically <5% after matching, indicating that the matching variables and matching methods selected in this study were reasonable. In addition, the t statistics after matching were not significant, suggesting no significant differences between the treatment and control groups after matching. Therefore, the samples obtained after matching ensure the randomness of sample processing and the reliability of the estimated results in this study. After PSM, the kernel density plots of the treatment group and the control group are closer to each other (as shown in Figure 3), indicating that the matching process successfully reduced the differences in covariates between the two groups, achieving better balance.

XSL•FO
RenderX

Figure 3. Kernel density plots before and after propensity score matching in phase 1.

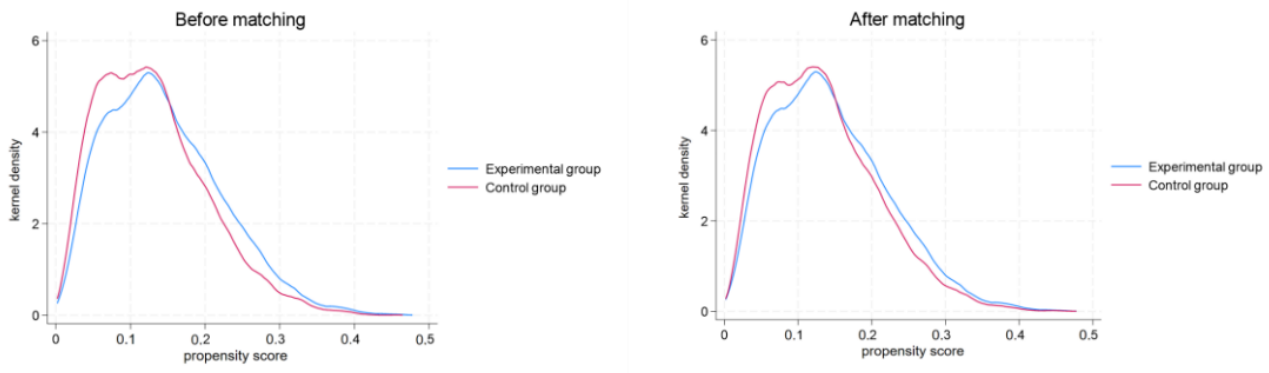

XSL•FO
RenderX

Table 3. Balance test results of propensity score matching for the first period.

Variable	Values, mean		Bias (%)	*t* test (*df*)	*P* value	V(T)[a]/V(C)[b]	Bias reduction (%)
	Treatment	Control					
hospital_rank[c]							100
Unmatched	5.713	5.687	2.4	0.26 (4075)	.80	0.83	
Matched	5.712	5.712	0	–0.00 (574)	>.99	0.94	
title_add[d]							25.9
Unmatched	5.092	5.078	0.6	0.07 (4075)	.94	0.99	
Matched	5.074	5.064	0.5	0.04 (574)	.97	1.00	
hot[e]							74.6
Unmatched	3.596	3.568	4.9	0.53 (4075)	.60	1.09	
Matched	3.599	3.606	–1.2	–0.11 (574)	.91	0.97	
vote_each_month_ln[f]							–68.8
Unmatched	2.668	2.623	3.3	0.35 (4075)	.73	0.81	
Matched	2.677	2.602	5.5	0.49 (574)	.62	0.76	
inquiry_ln[g]							97.2
Unmatched	7.836	7.720	8.5	0.94 (4075)	.35	1.13	
Matched	7.869	7.866	0.2	0.02 (574)	.98	1.19	
service_num[h]							67.9
Unmatched	1.933	1.837	8.4	0.90 (4075)	.37	0.93	
Matched	1.945	1.914	2.7	0.24 (574)	.81	0.87	
wenzhen_price_ln[i]							–3.6
Unmatched	3.419	3.375	3.4	0.37 (4075)	.71	1.31	
Matched	3.440	3.395	3.5	0.32 (574)	.75	1.33	
province_order[j]							–76.4
Unmatched	3.348	3.306	2.9	0.31 (4075)	.75	0.97	
Matched	3.350	3.276	5.1	0.46 (574)	.65	0.97	

[a]V(T): variance in treatment group.

[c]hospital_rank: the rank of the hospital where the physician works.

[d]title_add: physician title.

[e]hot: physician popularity.

[f]vote_each_month_ln: physician reputation.

[g]inquiry_ln: the quantity of online medical inquiries.

[h]service_num: the number of physicians' online services.

[i]wenzhen_price_ln: physicians' consultation price.

[j]province_order: the rank of the city where the physician works.

bV(C): variance in control group.

After PSM, we conducted descriptive statistical analyses of the data (Table 4). The descriptive statistics revealed that the participation rate of physicians in paid knowledge services was 30.3% (4793/15,820) across all physician-period combinations. The median of 0 (IQR 0.000-1.000) indicates that the overall sample had a low participation rate in paid knowledge services. Furthermore, the data distribution of physicians within the sample shows significant variation in the dimensions of quantity, quality, and diversity of knowledge services, indicating potential areas for further improvement.

XSL•FO
RenderX

Table 4. Summary statistics.

Variable name	Observations, n	Values, mean (SD)	Values, median (IQR; range)
Independent variable			
paid_service[a]	15,820	0.303 (0.460)	0.000 (0.000-1.000; 0.000-1.000)
Dependent variable			
content_num[b]	15,820	8.841 (33.272)	3.000 (1.000-7.000; 1.000-2029.000)
content_len[c]	15,820	1201.667 (1383.669)	925.535 (500.000-1493.250; 0.000-43,188.000)
simi_fuzz[d]	15,820	3.332 (5.895)	2.000 (1.000-3.438; 0.000-100.000)
funct[e]	15,820	0.250 (0.099)	0.273 (0.221-0.312; 0.000-0.625)
health[f]	15,820	0.054 (0.037)	0.058 (0.024-0.078; 0.000-0.750)
bio[g]	15,820	0.078 (0.051)	0.089 (0.041-0.113; 0.000-1.000)
cause[h]	15,820	0.019 (0.015)	0.019 (0.002-0.029; 0.000-0.222)
tab_type[i]	15,820	1.046 (0.236)	1.000 (1.000-1.000; 1.000-4.000)
shanno_ent[j]	15,532	7.288 (1.549)	7.673 (7.303-7.958; 0.000-9.397)
Matching variables			
title_add[k]	15,820	4.504 (2.144)	4.000 (3.000-6.000; 0.000-8.000)
hospital_rank[l]	15,820	5.570 (1.334)	6.000 (6.000-6.000; 0.000-6.000)
province_order[m]	15,820	3.351 (1.490)	4.000 (2.000-5.000; 0.000-5.000)
vote_each_month_ln[n]	15,820	1.802 (1.424)	1.792 (0.693-2.944; 0.000-6.842)
wenzhen_price_ln[o]	15,820	3.249 (1.218)	3.367 (2.398-4.111; 0.000-7.314)
service_num[p]	15,820	1.788 (1.136)	2.000 (1.000-3.000; 0.000-4.000)
hot[q]	15,820	3.507 (0.545)	3.400 (3.100-3.800; 0.000-5.000)
inquiry_ln[r]	15,820	7.235 (1.523)	7.412 (6.371-8.286; 0.000-11.752)

[a]paid_service: the participation of physicians in paid knowledge services.
[b]content_num: post number.
[c]content_len: character count of free articles.
[d]simi_fuzz: free article theme compliance.
[e]funct: the frequency of functional words in free articles published by physicians.
[f]health: the frequency of health words in free articles published by physicians.
[g]bio: the frequency of biological words in free articles published by physicians.
[h]cause: the frequency of causal words in free articles published by physicians.
[i]tab_type: number of media types.
[j]shanno_ent: content information entropy.
[k]title_add: physician title.
[l]hospital_rank: the rank of the hospital where the physician works.
[m]province_order: the rank of the city where the physician works.
[n]vote_each_month_ln: physician reputation.
[o]wenzhen_price_ln: physicians' consultation price.
[p]service_num: the number of physicians' online services.
[q]hot: physician popularity.
[r]inquiry_ln: the quantity of online medical inquiries.

XSL•FO
RenderX

Results and Analysis of the Multi-Period DID Estimation

On the basis of the PSM estimation, this study identified a new control group with similar characteristics to those of the treatment group. The treatment and control groups encompassed a total of 7453 physicians and 15,820 samples. For these matched samples, we constructed dependent variables from the dimensions of quantity, quality, and diversity. The multi-period DID method was used to examine the impact of physicians' engagement in paid knowledge activities on their participation in free knowledge activities. The detailed estimation results are shown in Tables 5 and 6. Table 5 illustrates the pure treatment effects on the dimensions of quantity, quality, and diversity of various dependent variables derived from a multi-period DID analysis after eliminating confounding factors. To validate the robustness of these findings, Table 6 incorporates additional regression analyses with control variables. The results of Table 6 are largely consistent with those of Table 5, thereby strengthening the reliability and validity of the treatment effects.

Table 5. Estimation results of difference in differences (no controls).

	Quantity—content_num[a] (model 1)	Quality						Diversity	
		content_len[b] (model 2)	simi_fuzz[c] (model 3)	funct[d] (model 4)	health[e] (model 5)	bio[f] (model 6)	cause[g] (model 7)	tab_type[h] (model 8)	shanno_ent[i] (model 9)
paid_service[j], estimated value (*P* value)	5.071 (<.001)	106.100 (.01)	–1.121 (<.001)	0.006 (.001)	0.000 (.76)	0.001 (.55)	0.002 (<.001)	0.004 (.02)	0.065 (.007)
Observations, N	15,820	15,820	15,820	15,820	15,820	15,820	15,820	15,820	15,532
R^2	0.005	0.006	0.010	0.054	0.038	0.049	0.015	0.033	0.075
Adjusted R^2	0.004	0.005	0.010	0.053	0.037	0.048	0.014	0.032	0.074

[a]content_num: post number.

[b]content_len: character count of free articles.

[c]simi_fuzz: free article theme compliance.

[d]funct: the frequency of functional words in free articles published by physicians.

[e]health: the frequency of health words in free articles published by physicians.

[f]bio: the frequency of biological words in free articles published by physicians.

[g]cause: the frequency of causal words in free articles published by physicians.

[h]tab_type: number of media types.

[i]shanno_ent: content information entropy.

[j]paid_service: the participation of physicians in paid knowledge services.

XSL•FO
RenderX

Table 6. Estimation results of difference in differences (with controls).

	Quantity—content_num[a] (model 10)	Quality						Diversity	
		content_len[b] (model 11)	simi_fuzz[c] (model 12)	funct[d] (model 13)	health[e] (model 14)	bio[f] (model 15)	cause[g] (model 16)	tab_type[h] (model 17)	shanno_ent[i] (model 18)
paid_service[j], estimated value (*P* value)	4.731 (<.001)	108.000 (.009)	–1.155 (<.001)	0.006 (.001)	0.001 (.64)	0.001 (.45)	0.002 (<.001)	0.003 (.09)	0.066 (.006)
hospital_rank[k], estimated value (*P* value)	—[l]	—	—	—	—	—	—	—	—
title_add[m], estimated value (*P* value)	–1.321 (.60)	102.900 (.72)	–3.763 (.15)	–0.067 (.11)	–0.007 (.64)	–0.037 (.001)	–0.003 (.60)	–0.023 (.13)	–1.006 (.37)
hot[n], estimated value (*P* value)	76.450 (.55)	1659.900 (.92)	120.800 (.23)	4.517 (.04)	0.355 (.60)	1.886 (.003)	0.225 (.50)	1.998 (.25)	—
vote_each_month_ln[o], estimated value (*P* value)	2.322 (<.001)	–13.140 (.39)	0.237 (.001)	–0.001 (.26)	–0.001 (.004)	–0.001 (.02)	0.000 (.78)	0.007[o] (.04)	–0.005 (.82)
inquiry_ln[p], estimated value (*P* value)	–256.800 (.72)	113,611.900 (.24)	82.680 (.74)	6.442 (.64)	5.590 (.09)	3.299 (.43)	0.816 (.72)	–15.460 (.26)	104.400 (.36)
service_num[q], estimated value (*P* value)	—	—	—	—	—	—	—	—	—
wenzhen_price_ln[r], estimated value (*P* value)	7.399 (.02)	2078.500 (<.001)	–0.756 (.37)	0.029 (.34)	–0.220 (<.001)	–0.214[t] (<.001)	0.094[t] (<.001)	0.055 (.13)	–0.164 (.56)
province_order[s], estimated value (*P* value)	—	—	—	—	—	—	—	—	—
Observations, N	15,820	15,820	15,820	15,820	15,820	15,820	15,820	15,820	15,532
R^2	0.008	0.006	0.012	0.056	0.040	0.051	0.016	0.034	0.075
Adjusted R^2	0.007	0.005	0.011	0.055	0.039	0.050	0.014	0.033	0.074

[a]content_num: post number.
[b]content_len: character count of free articles.
[c]simi_fuzz: free article theme compliance.
[d]funct: the frequency of functional words in free articles published by physicians.
[e]health: the frequency of health words in free articles published by physicians.
[f]bio: the frequency of biological words in free articles published by physicians.
[g]cause: the frequency of causal words in free articles published by physicians.
[h]tab_type: number of media types.
[i]shanno_ent: content information entropy.
[j]paid_service: the participation of physicians in paid knowledge services.
[k]hospital_rank: the rank of the hospital where the physician works.
[l]Not applicable.
[m]title_add: physician title.
[n]hot: physician popularity.
[o]vote_each_month_ln: physician reputation.
[p]inquiry_ln: the quantity of online medical inquiries.
[q]service_num: the number of physicians' online services.
[r]wenzhen_price_ln: physicians' consultation price.
[s]province_order: the rank of the city where the physician works.

In the quantity dimension, model 10 shows that physicians' participation in paid knowledge activities led to a significant 473.1% increase in the number of free articles (*P*<.001). This substantial increase aligns with the findings of previous research

XSL•FO
RenderX

[1,6]. The consumer's knowledge payment behavior helps to promote the contribution of free knowledge.

According to Table 6, in the dimension of quality, the participation in paid knowledge activities significantly promoted physicians' performance in terms of content length, frequency of functional words, and frequency of causal words. The result of model 11 demonstrates that physicians' participation in paid knowledge activities had a positive spillover effect on the length of free articles (108%; P=.009), which is inconsistent with the conclusion that economic incentives do not affect the quality of free services in previous study [6,22]. This may be due to physicians' desire to enhance the online reputation of their free content and establish a good online reputation, thereby attracting more traffic to their paid services. The result of model 12 indicates that physicians' participation in paid knowledge activities had a negative effect on the consistency between titles and content in free articles, with a reduction of 115.5% (P<.001). This is inconsistent with our initial hypothesis. Further investigation on examples of lower similarity scores is shown in Table S1 in Multimedia Appendix 1. It indicates that physicians often use eye-catching expressions and summaries in titles to engage readers, which may lead to less alignment between titles and content. We further conducted supplementary analyses (Table S2 in Multimedia Appendix 1). First, a t test comparing the consistency between titles and content in free versus paid knowledge after physicians engaged in paid knowledge activities was conducted. The results revealed that, within 12 months of initiating paid activities, the consistency between titles and content in free knowledge services remained significantly higher than in paid knowledge services. This may be because paid knowledge products require more attention-grabbing and creative titles to enhance click-through and conversion rates, thereby reducing consistency between titles and content. In the short term, physicians might also adjust the titles of their free knowledge services, resulting in a decrease in title-content consistency for free services after adopting the paid service, although consistency remained higher than for paid services. Second, over the 12 months following the initiation of the paid service, consistency between titles and content in both free and paid knowledge services gradually improved. This trend may be attributed to physicians accumulating more experience in providing knowledge services over time, enabling them to better balance creativity and content, thereby progressively enhancing the consistency between titles and content. The result of model 13 shows that physicians' participation in paid knowledge activities had a positive effect on the frequency of functional words in free articles, which increased by 0.6% (P=.001). Physicians adopt different language styles in the expression of free content, which may be a differentiation strategy to distinguish paid content from free content. Model 14 indicates that there was no significant effect on the frequency of health-related words in free articles. Similarly, the result of model 15 shows no significant effect on the frequency of biological words. The changes in health word and biological word frequency before and after receiving the economic incentive were not significant, which shows that physicians did not pursue the output of professional content, possibly to maintain the universal positioning of free content. The result of model 16 reveals that physicians' participation in paid knowledge activities positively impacted the frequency of causal words in free articles, which increased by 0.2% (P<.001). Physicians increased the frequency of causal words, which may be to narrate medical content with greater logic.

In the dimension of diversity, participation in paid knowledge activities had a significant impact on the content information entropy of free knowledge services. The result of model 8 indicates that physicians' participation in paid knowledge activities had a positive impact on the media diversity of free articles, but the significance level was relatively low (P=.02). However, model 17 shows that physicians' participation in paid knowledge activities had no significant impact on the media diversity of free articles. This discrepancy may arise from path dependency in physicians' choice of media as they tended to continue using previously used media, leading to minimal variation in media use before and after engaging in paid knowledge activities. The result of model 18 indicates that physicians' participation in paid knowledge activities had a positive effect on the content information entropy of free articles, which increased by 6.6% (P=.006). This finding aligns with the conclusion by Wang et al [6]. Economic incentives encourage physicians to produce more diverse and rich content to the community, contributing to its vitality and development.

Overall, the results demonstrate that the physicians' participation in paid knowledge activities significantly enhanced their engagement in free knowledge activities. In terms of quantity, participation in paid knowledge activities substantially increased the number of free articles published by physicians. In terms of quality, paid knowledge participation had a significant positive impact on the length, function word frequency, and causal word frequency of free articles while negatively affecting the consistency between title and content. It had no significant impact on the frequency of health-related and biological words. In terms of diversity, paid knowledge participation significantly improved the content information entropy of free articles but did not notably affect media diversity.

Discussion

Principal Findings

The research question addressed in this study was the impact of economic incentives on physicians' free knowledge contribution in online medical communities. To elucidate this issue, this study used a quasi-experimental method combining PSM with multi-period DID to verify the spillover effect of knowledge payment activities on free knowledge contributions across dimensions of quantity, quality, and diversity. The sample selection bias problem was controlled through PSM. The findings revealed that economic incentives exerted spillover effects on physicians' free knowledge contributions, affecting all 3 dimensions. Specifically, economic incentives had a positive spillover effect on the quantity, length, and frequency of functional words; frequency of causal words; and content information entropy of free articles while negatively impacting the consistency between titles and content of free articles.

XSL•FO
RenderX

Theoretical Implications

Theoretically, this study enriches understanding of economic incentives within the realm of online health care and explores physicians' online knowledge contributions from a multidimensional perspective. The contributions of this study are as follows. First, this study comprehensively explored the impact of economic incentives on physicians' free online medical knowledge contributions, enriching the literature on digital health care and knowledge contribution. Previous studies have mostly focused on nonspecific fields or broad knowledge-sharing platforms [5], with few investigating the motivational mechanisms for knowledge contribution in the specific context of health care. By introducing motivation theory, this study validated the spillover effect of paid knowledge contributions on free knowledge contributions by physicians, providing a theoretical basis and decision support for the sustainable and orderly development of paid knowledge services in online medical communities. Second, this study constructed a comprehensive, 3D knowledge evaluation framework. Previous research has primarily focused on the quantity and quality of knowledge [31,34]. This study further introduced the important dimension of *diversity*. In terms of measurement methods, earlier research has largely relied on structured data. In contrast, this study used text-mining techniques to extract information related to knowledge quality, such as the consistency between titles and content, function words, health-related vocabulary, physiological process terms, causal words, and entropy values related to knowledge diversity. Third, this study used a year-by-year PSM and a multi-period DID method to scientifically evaluate the impact of economic incentive models on the platform. By constructing a counterfactual framework, the PSM method was used to introduce physicians who did not participate in paid knowledge contributions as a control group, scientifically assessing the impact of participation in paid knowledge contributions on physicians' free knowledge activities. Simultaneously, the year-by-year DID reduced the selection bias of physicians who published paid content at different time points and, in combination with text-mining methods, explored the impact of economic incentives on physicians' online knowledge participation on the platform.

Practical Implications

This study also has certain practical implications. First, for the platform, providing economic incentives to physicians can promote the overall development of the platform's knowledge content ecosystem. The introduction of economic incentives can be a feasible business model for knowledge platforms to stimulate users' contributions in paid and related unpaid activities. Currently, some free medical knowledge–sharing platforms, such as Dingxiangyuan and Kuaishou Health, have introduced paid knowledge through paid courses or short videos. This study holds significant reference value and strategic guidance for these platforms. In addition, this study also provides guidance for nonpaid medical knowledge platforms such as Chunyu Doctor and Baidu Health, offering valuable insights and strategic direction for introducing brand-new paid knowledge models. For paid knowledge platforms in other countries that operate on a membership basis or support one-time payments, such as medical-related communities on Patreon and medical-related modules on Substack, the research framework and related conclusions of this study can also provide valuable insights and references. Second, regarding physicians, they need to maintain their attention and investment in free knowledge activities and effectively manage them as a traffic entrance for paid users. Third, for patients, economic incentives have spillover effects on their free experience, helping them enjoy more, better, and more diverse knowledge articles for free.

Limitations

There are still some limitations to this study, which can be studied in depth in the future. First, the observation period was relatively short. This study only chose half a year as an observation period for the release of paid articles, primarily focusing on the medium- to long-term economic incentive effects. In the future, the observation period can be extended to study the long-term changes in spillover effects and further explore the impact of economic incentives on physicians' long-term professional development and knowledge-sharing behaviors. Second, in addition to economic incentives, the influence of noneconomic incentives (such as reputation) on physicians' online knowledge service behaviors is a significant topic for future research. In the future, further in-depth exploration can be conducted on the noneconomic incentive measures of the platform and their impact on physicians' online knowledge service behaviors. Third, in terms of the presentation format of knowledge content, only the textual form of products was studied. In the future, we can use speech transcription–related tools to extract information from videos, speech, and other forms of products and fully consider all types of articles. Fourth, this study was based on an online medical platform in China. Given that cultural backgrounds would profoundly influence individuals' values, motivational systems, and decision-making processes, the conclusions drawn from Chinese platforms may offer only limited applicability, and there are some limitations in applying these findings to other countries and regions. Future research should fully consider the influence of cultural backgrounds and conduct more in-depth studies in this regard.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (grants 71971082, 71471064, 72271093, and 91646205) and was also funded by the Science and Technology Innovation Action Plan of the Science and Technology Commission of Shanghai Municipality (grants 22692110200 and 19692106700) and the Humanities and Social Sciences Youth Foundation of the Ministry of Education of China (grant 18YJC630068). The authors would like to thank the reviewers for their comments.

XSL•FO
RenderX

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XL was responsible for the conceptualization of the research and methodology and contributed to the writing and revision of the paper. XC was responsible for revision, editing, and data curation. MC conducted data collection and formal analysis. WS contributed to draft preparation. JL contributed to the review and editing of the paper and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Further analysis of the consistency between titles and content.
[DOCX File , 20 KB - jmir_v26i1e58688_app1.docx]

References

1. Qi TT, Wang TM, Yan JR. The spillover effects of different monetary incentive levels on health experts' free knowledge contribution behavior. Internet Res 2021 Jul 16;31(6):2143-2166. [doi: 10.1108/intr-08-2020-0445]
2. Garnefeld I, Iseke A, Krebs A. Explicit incentives in online communities: boon or bane? Int J Electron Commer 2014 Dec 08;17(1):11-38. [doi: 10.2753/jec1086-4415170101]
3. Burtch G, Hong YL, Bapna R, Griskevicius V. Stimulating online reviews by combining financial incentives and social norms. Manag Sci 2018 May;64(5):2065-2082. [doi: 10.1287/mnsc.2016.2715]
4. Gallus J. Fostering public good contributions with symbolic awards: a large-scale natural field experiment at Wikipedia. Manag Sci 2017 Dec;63(12):3999-4015. [doi: 10.1287/mnsc.2016.2540]
5. Kuang LN, Huang N, Hong YL, Yan ZJ. Spillover effects of financial incentives on non-incentivized user engagement: evidence from an online knowledge exchange platform. J Manag Inf Syst 2019 Mar 31;36(1):289-320. [doi: 10.1080/07421222.2018.1550564]
6. Wang J, Li G, Hui KL. Monetary incentives and knowledge spillover: evidence from a natural experiment. Manag Sci 2022 May;68(5):3549-3572. [doi: 10.1287/mnsc.2021.4048]
7. Falk A, Fischbacher U. A theory of reciprocity. Games Econ Behav 2006 Feb;54(2):293-315. [doi: 10.1016/j.geb.2005.03.001]
8. Tang Q, Gu B, Whinston AB. Content contribution for revenue sharing and reputation in social media: a dynamic structural model. J Manag Inf Syst 2014 Dec 08;29(2):41-76. [doi: 10.2753/MIS0742-1222290203]
9. Brüggen A, Moers F. The role of financial incentives and social incentives in multi‐task settings. J Manag Account Res 2007;19(1):25-50 [FREE Full text] [doi: 10.2308/jmar.2007.19.1.25]
10. Frey B, Oberholzer-Gee F. The cost of price incentives: an empirical analysis of motivation crowding- out. Am Econ Rev 1997;87(4):746-755 [FREE Full text]
11. Zhao L, Detlor B, Connelly CE. Sharing knowledge in social Q and A sites: the unintended consequences of extrinsic motivation. J Manag Inf Syst 2016 Jun 17;33(1):70-100. [doi: 10.1080/07421222.2016.1172459]
12. Kanthawala S, Vermeesch A, Given B, Huh J. Answers to health questions: internet search results versus online health community responses. J Med Internet Res 2016 Apr 28;18(4):e95 [FREE Full text] [doi: 10.2196/jmir.5369] [Medline: 27125622]
13. Zhou JJ, Zuo MY, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019 Oct;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
14. Yang HL, Du HS, He W, Qiao H. Understanding the motivators affecting doctors' contributions in online healthcare communities: professional status as a moderator. Behav Inf Technol 2019 Oct 23;40(2):146-160. [doi: 10.1080/0144929X.2019.1679887]
15. Meng FB, Zhang XF, Liu LB, Ren CC. Converting readers to patients? From free to paid knowledge-sharing in online health communities. Inf Process Manage 2021 May;58(3):102490. [doi: 10.1016/j.ipm.2021.102490]
16. Khurana S, Qiu LF, Kumar S. It shows: an empirical analysis of doctors' responses in a Q and A forum of an online healthcare portal. Inf Syst Res 2019 Sep;30(3):872-891. [doi: 10.1287/isre.2019.0836]
17. Liu Z, Li Y, Min Q, Chang M. User incentive mechanism in blockchain-based online community: an empirical study of steemit. Inf Manag 2022 Nov;59(7):103596. [doi: 10.1016/j.im.2022.103596]
18. Lou J, Fang Y, Lim KH, Peng JZ. Contributing high quantity and quality knowledge to online Q and A communities. J Am Soc Inf Sci Tec 2012 Dec 13;64(2):356-371. [doi: 10.1002/asi.22750]
19. Cavusoglu H, Li Z, Kim SH. How do virtual badges incentivize voluntary contributions to online communities? Inf Manag 2021 Jul;58(5):103483. [doi: 10.1016/j.im.2021.103483]

XSL•FO
RenderX

20. Sun M, Zhu F. Ad revenue and content commercialization: evidence from blogs. Manag Sci 2013 Oct;59(10):2314-2331. [doi: 10.1287/mnsc.1120.1704]
21. Jenkins Jr GD, Mitra A, Gupta N, Shaw JD. Are financial incentives related to performance? A meta-analytic review of empirical research. J Appl Psychol 1998;83(5):777-787. [doi: 10.1037//0021-9010.83.5.777]
22. Chen HL, Hu YJ, Huang S. Monetary incentive and stock opinions on social media. J Manag Inf Syst 2019 Jun 14;36(2):391-417. [doi: 10.1080/07421222.2019.1598686]
23. Chen Y, Ho TH, Kim YM. Knowledge market design: a field experiment at Google answers. J Public Econ Theory 2010 Jul 19;12(4):641-664. [doi: 10.1111/j.1467-9779.2010.01468.x]
24. Sun YC, Dong XJ, McIntyre S. Motivation of user-generated content: social connectedness moderates the effects of monetary rewards. Mark Sci 2017 May;36(3):329-337. [doi: 10.1287/mksc.2016.1022]
25. Qiao DD, Lee SY, Whinston AB, Wei Q. Financial incentives dampen altruism in online prosocial contributions: a study of online reviews. Inf Syst Res 2020 Dec;31(4):1361-1375. [doi: 10.1287/isre.2020.0949]
26. Zhang M, Guo XT, Wu TS. Impact of free contributions on private benefits in online healthcare communities. Int J Electron Commer 2019 Sep 19;23(4):492-523. [doi: 10.1080/10864415.2019.1655208]
27. Chen JH, Lan YC, Chang YW, Chang PY. Exploring doctors' willingness to provide online counseling services: the roles of motivations and costs. Int J Environ Res Public Health 2019 Dec 22;17(1):110 [FREE Full text] [doi: 10.3390/ijerph17010110] [Medline: 31877861]
28. Hung SY, Durcikova A, Lai HM, Lin WM. The influence of intrinsic and extrinsic motivation on individuals' knowledge sharing behavior. Int J Hum Comput Stud 2011 Jun;69(6):415-427. [doi: 10.1016/j.ijhcs.2011.02.004]
29. Osterloh M, Frey BS. Motivation, knowledge transfer, and organizational forms. Organ Sci 2000 Oct;11(5):538-550. [doi: 10.1287/orsc.11.5.538.15204]
30. Roberts JA, Hann IH, Slaughter SA. Understanding the motivations, participation, and performance of open source software developers: a longitudinal study of the Apache projects. Manag Sci 2006 Jul;52(7):984-999. [doi: 10.1287/mnsc.1060.0554]
31. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
32. Gneezy U, Rustichini A. Pay enough or don't pay at all*. Q J Econ 2000 Aug;115(3):791-810. [doi: 10.1162/003355300554917]
33. Khern-am-nuai W, Kannan K, Ghasemkhani H. Extrinsic versus intrinsic rewards for contributing reviews in an online platform. Inf Syst Res 2018 Dec;29(4):871-892. [doi: 10.1287/isre.2017.0750]
34. Cabral L, Li LF. A dollar for your thoughts: feedback-conditional rebates on eBay. Manag Sci 2015 Sep;61(9):2052-2063. [doi: 10.1287/mnsc.2014.2074]
35. Harrison DA, Klein KJ. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. Acad Manage Rev 2007 Oct;32(4):1199-1228. [doi: 10.5465/amr.2007.26586096]
36. de Almeida SO, Dholakia UM, Hernandez JM, Mazzon JA. The mixed effects of participant diversity and expressive freedom in online peer-to-peer problem solving communities. J Interact Mark 2014 Aug 01;28(3):196-209. [doi: 10.1016/j.intmar.2014.02.003]
37. Liu X, Wang X, Li J, Chen M. The effect of media richness on the stability of physician-patient relationships on e-consultation platforms. J Glob Inf Manag 2022;30(1):1-26. [doi: 10.4018/JGIM.315301]
38. Lu XC, Jiang JL, Head M, Yang JY. The impact of linguistic complexity on leadership in online Q and A communities: comparing knowledge shaping and knowledge adding. Inf Manag 2022 Sep;59(6):103675. [doi: 10.1016/j.im.2022.103675]
39. Good doctor online. Haodf Online platform. URL: https://www.haodf.com/ [accessed 2024-04-29]
40. Price R, Shanks G. A semiotic information quality framework: development and comparative analysis. J Inf Technol 2005 Jun 01;20(2):88-102. [doi: 10.1057/palgrave.jit.2000038]
41. Agichtein E, Castillo C, Donato D, Gionis A, Mishne G. Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008 Presented at: WSDM '08; February 11-12, 2008; Palo Alto, CA p. 183-194 URL: https://dl.acm.org/doi/10.1145/1341531.1341557 [doi: 10.1145/1341531.1341557]
42. Oh S, Worrall A, Yi YJ. Quality evaluation of health answers in Yahoo! answers: a comparison between experts and users. Proc Assoc Inf Sci Technol 2012 Jan 11;48(1):1-3 [FREE Full text] [doi: 10.1002/meet.2011.14504801269]
43. Zhang YL, Li XM, Fan WG. User adoption of physician's replies in an online health community: an empirical study. J Assoc Inf Sci Technol 2019 Nov 14;71(10):1179-1191. [doi: 10.1002/asi.24319]
44. Yu LX, Jiang WY, Ren ZH, Xu S, Zhang L, Hu XE. Detecting changes in attitudes toward depression on Chinese social media: a text analysis. J Affect Disord 2021 Feb 01;280(Pt A):354-363. [doi: 10.1016/j.jad.2020.11.040] [Medline: 33221722]
45. Jiang S, Liu X, Chi XT. Effect of writing style on social support in online health communities: a theoretical linguistic analysis framework. Inf Manag 2022 Sep;59(6):103683. [doi: 10.1016/j.im.2022.103683]
46. Zhang JL, Zhang J, Wang KL, Yan W. Should doctors use or avoid medical terms? The influence of medical terms on service quality of E-health. Electron Commer Res 2021 Oct 26;23(3):1775-1805. [doi: 10.1007/S10660-021-09516-6]
47. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 2009 Dec 08;29(1):24-54. [doi: 10.1177/0261927X09351676]

XSL•FO
RenderX

48. Shannon CE. The mathematical theory of communication. Bell Syst Tech J 1948;27(3):379-423 [FREE Full text] [Medline: 9230594]
49. Rao YH, Xie HR, Li J, Jin FM, Wang FL, Li Q. Social emotion classification of short text via topic-level maximum entropy model. Inf Manag 2016 Dec;53(8):978-986. [doi: 10.1016/j.im.2016.04.005]
50. Peng SC, Yang AM, Cao LH, Yu S, Xie DQ. Social influence modeling using information theory in mobile social networks. Inf Sci 2017 Feb;379:146-159. [doi: 10.1016/j.ins.2016.08.023]
51. Ni J, Zhang YL, Hu ZL, Song WJ, Hou L, Guo Q, et al. Ceiling effect of online user interests for the movies. Phys A Stat Mech Appl 2014 May;402:134-140. [doi: 10.1016/j.physa.2014.01.046]
52. Cao B, Huang W, Chao N, Yang G, Luo N. Patient activeness during online medical consultation in China: multilevel analysis. J Med Internet Res 2022 May 27;24(5):e35557 [FREE Full text] [doi: 10.2196/35557] [Medline: 35622403]
53. Deng Z, Hong Z, Zhang W, Evans R, Chen Y. The effect of online effort and reputation of physicians on patients' choice: 3-wave data analysis of China's good doctor website. J Med Internet Res 2019 Mar 08;21(3):e10170 [FREE Full text] [doi: 10.2196/10170] [Medline: 30848726]
54. Shah AM, Yan X, Tariq S, Ali M. What patients like or dislike in physicians: analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach. Inf Process Manage 2021 May;58(3):102516. [doi: 10.1016/j.ipm.2021.102516]
55. Wang JN, Chiu YL, Yu H, Hsu YT. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: longitudinal study. J Med Internet Res 2017 Dec 21;19(12):e427 [FREE Full text] [doi: 10.2196/jmir.9082] [Medline: 29269344]
56. Li YM, Yan XB, Song XL. Provision of paid web-based medical consultation in China: cross-sectional analysis of data from a medical consultation website. J Med Internet Res 2019 Jun 03;21(6):e12126 [FREE Full text] [doi: 10.2196/12126] [Medline: 31162129]
57. Wang JJ, Liu HY, Ye J. The impact of individual and team professional capital on physicians' performance in online health-care communities: a cross-level investigation. Internet Res 2022 Mar 15;33(1):152-177. [doi: 10.1108/intr-08-2021-0544]
58. Guo SS, Guo XT, Fang YL, Vogel D. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. J Manag Inf Syst 2017 Aug 17;34(2):487-519. [doi: 10.1080/07421222.2017.1334480]
59. Holmstrom B, Milgrom P. Multitask principal–agent analyses: incentive contracts, asset ownership, and job design. J Law Econ Organ 1991;7:24-52 [FREE Full text] [doi: 10.1093/jleo/7.special_issue.24]
60. Chiu YL, Wang JN, Yu H, Hsu YT. Consultation pricing of the online health care service in China: hierarchical linear regression approach. J Med Internet Res 2021 Jul 14;23(7):e29170 [FREE Full text] [doi: 10.2196/29170] [Medline: 34259643]
61. Tian X, Xu J. Do place-based policies promote local innovation and entrepreneurship? Rev Finance 2021;26(3):595-635. [doi: 10.1093/rof/rfab029]
62. Beck T, Levine R, Levkov A. Big bad banks? The winners and losers from bank deregulation in the United States. J Finance 2010 Sep 21;65(5):1637-1667. [doi: 10.1111/j.1540-6261.2010.01589.x]
63. Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. Stata J 2002 Dec 01;2(4):358-377. [doi: 10.1177/1536867X0200200403]

Abbreviations

DID: difference in differences
PSM: propensity score matching

Edited by A Schwartz; submitted 22.03.24; peer-reviewed by W He, J Mistry, Z Wang; comments to author 11.06.24; revised version received 01.08.24; accepted 04.11.24; published 10.12.24.

Please cite as:
Liu X, Chi X, Chen M, Sun W, Li J
Spillover Effects of Paid Functions on Physicians' Unpaid Knowledge Activities: Quasi-Experimental Approach
J Med Internet Res 2024;26:e58688
URL: https://www.jmir.org/2024/1/e58688
doi:10.2196/58688
PMID:39656521

©Xuan Liu, Xiaotong Chi, Ming Chen, Wen Sun, Jia Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of a Social Media Strategy for Public Health Promotion in Black, American Indian or Alaska Native, and Hispanic or Latino Communities During the COVID-19 Pandemic: Cross-Sectional Study

Maria Mora Pinzon[1], MD, MS; Ornella Hills[2], PhD; George Levy[1], AA; Taryn T James[1,3], PhD; Ashley Benitez[4], BS; Sacheen Lawrence[1], BS; Tiffany Ellis[1], BA; Venus Washington[5], BA; Lizbeth Solorzano[6], BS; Patricia Tellez-Giron[7], MD; Fernando Cano Ospina[7]; Melissa F Metoxen[8], MS; Carey E Gleason[1,3,9], PhD

[1]Division of Geriatrics and Gerontology, Department of Medicine, School of Medicine and Public Health, University of Wisconsin - Madison, Madison, WI, United States

[2]Department of Ethnic Studies, College of Arts and Sciences, Lawrence University, Appleton, WI, United States

[3]Wisconsin Alzheimer's Disease Research Center, Department of Medicine, School of Medicine and Public Health, University of Wisconsin – Madison, Madison, WI, United States

[4]School of Medicine and Public Health, University of Wisconsin – Madison, Madison, WI, United States

[5]Venus Inspires, Madison, WI, United States

[6]UnityPoint Health - Meriter, Madison, WI, United States

[7]Latino Health Council of Dane County, Madison, WI, United States

[8]Native American Center for Health Professions, School of Medicine and Public Health, University of Wisconsin - Madison, Madison, WI, United States

[9]Geriatric Research, Education and Clinical Center, William S. Middleton Memorial Veterans Hospital, Department of Veterans Affairs, Madison, WI, United States

Corresponding Author:
Maria Mora Pinzon, MD, MS
Division of Geriatrics and Gerontology, Department of Medicine
School of Medicine and Public Health
University of Wisconsin - Madison
610 Walnut st
Madison, WI, 53726
United States
Phone: 1 6088902524
Email: mmora2@wisc.edu

Abstract

Background: Individuals identifying as Black, American Indian or Alaska Native, or Hispanic or Latino lack access to culturally appropriate accurate information and are the target of disinformation campaigns, which create doubt in science and health care providers and might play a role in sustaining health disparities related to the COVID-19 pandemic.

Objective: This study aims to create and disseminate culturally and medically appropriate social media messages for Black, Latino, and American Indian or Alaska Native communities in Wisconsin and evaluate their reach and effectiveness in addressing the information needs of these communities.

Methods: Our team identified relevant COVID-19 topics based on feedback from their respective community, developed lay format materials, and translated materials into culturally appropriate social media messages that community advocates delivered across their respective communities. Social media metrics (reach, engagement, and impressions) were collected using Sprout Social and Facebook Analytics. We hosted 9 focus groups with community members to learn about their social media use. These data were analyzed using an inductive approach, using NVivo software (release 1.7) to code content.

Results: Between August 2021 and January 2023, we created 980 unique social media posts that reached 88,790 individuals and gathered >6700 engagements. Average reach per post was similar across the 3 communities, despite differences in the number of posts and followers on each page: 119.46 (Latino individuals), 111.74 (Black individuals), and 113.11 (Oneida Nation members).

XSL•FO
RenderX

The type of posts that had higher engagement rate per reached person (ERR) varied across communities and platforms, with the highest being live videos for the Latino community on Facebook (ERR 9.4%), videos for the Black community on Facebook (ERR 19.53%), and social media messages for the Oneida Nation community (ERR 59.01%).

Conclusions: Our project presents a unique and effective model for health messages and highlights the need for tailoring social media messages and approaches for minoritized audiences (eg, age, gender, race, and ethnicity). Further research studies are needed to explore how specific types of information affect the dissemination of information and the implications for health communications.

(J Med Internet Res 2024;26:e58581) doi:10.2196/58581

KEYWORDS

health communications; social media; Hispanic; Latino; Black; American Indian; Alaska Native; minority health; health disparities; COVID-19

Introduction

Background

Individuals from Black, American Indian or Alaska Native, and Hispanic or Latino communities were disproportionately impacted by the SARS-CoV-2 virus that caused the COVID-19 pandemic. Multiple studies reported that Black, Hispanic, and American Indian or Alaska Native communities had higher hospitalization rates and death rates across the country [1,2]. In Kaiser Family data samples from >40 states, it was found that the rate of infection for Hispanic patients was >3 times the rate for White patients, and from March to July 18, 2020, hospitalization rates due to the COVID-19 pandemic for Black, American Indian or Alaska Native, and Hispanic patients were approximately 5 times higher than those for White patients [3,4]. In Wisconsin, compared with non-Hispanic White Wisconsin residents, Hispanic or Latino residents had 1.7 times higher case rates, Black residents had 2.1 times higher hospitalization rates, and American Indian residents had 1.4 times greater death rates [5].

Greater prevalence of comorbid conditions, such as diabetes, obesity, and cardiovascular disease, among Black, American Indian or Alaska Native, and Hispanic or Latino patients elevated their risk of hospitalization once infected [1]. Poor access to health care, low physician trust delaying help seeking and systemic racism and mistreatment in medical care likely worsened these hospitalization rates and death rates [1]. Furthermore, individuals who identify as Black, American Indian or Alaska Native, or Hispanic or Latino are more likely to live in socioeconomically disadvantaged neighborhoods [6,7], tend to rely more on public transportation, are more likely to live in multigenerational households, and are more likely to hold jobs that cannot be performed remotely [1,6], leaving them at a higher risk of exposure to SARS-CoV-2 virus. Notably, McCormack et al [8], in their study of >3 million persons, found that Black individuals were overrepresented in multiple industries with essential workers, including transportation, public administration, and health care. In many cities, these neighborhoods with communities of color have been segregated, which increased the concentration and spread of the virus in these neighborhoods and likely isolated its impact to these racially segregated communities [6]. The disparities in vaccination rates only exacerbated the issue. By May 25, 2021, a total of 44% of White residents of Wisconsin had received at least 1 dose of the COVID-19 vaccine. However, only 23% of Black residents, 29% of American Indian residents, and 31% of Hispanic or Latino residents had received a dose [9]. The reasons for lower vaccination rates were varied, including systemic racism [10], disparities in accessing care, lack of information about vaccination centers, misinformation and disinformation campaigns [11], and confusion about benefits and eligibility criteria [12].

During the COVID-19 pandemic, the lack of in-person interactions inhibited the use of traditional outreach approaches to address the misinformation, resulting in an increased use of social media [13]. As more individuals spend more time on online social networks, information and misinformation spread farther and faster than ever before, requiring the development of new approaches for using social media for public health promotion [14,15]. These approaches also needed to account for the mistrust in governmental and research organizations that communities of color might experience due to the structural racism, historical inequities, language of preference, and health literacy of the communities.

In light of the misinformation campaigns that were targeting communities of color [16], the community advisory board of the Wisconsin Alzheimer's Disease Research Center made the recommendation that the team pivot and focus on providing accurate information to communities of color on the COVID-19 pandemic. After exploration with additional community partners, there was a high level of interest in a social media solution that sought to address the needs of Black, Latino, and American Indian or Alaska Native communities in the state of Wisconsin by creating culturally appropriate content that promotes medically accurate information about the COVID-19 pandemic.

Multiple psychological theories, such as cultivation theory [17] and attribution theory, can be used to explore how media messages, including social media messages, affect individuals' perceptions and behaviors. Cultivation theory suggests that long-term exposure to media content shapes an individual's beliefs, attitudes, and perceptions about the world. It argues that heavy viewers of television, for example, are more likely to adopt the values and perspectives presented in media programming, leading to a *cultivation* of their worldview [17]. In the context of social media, for example, if social media messages were to consistently portray smoking as a common and acceptable behavior, individuals who are heavy users of

XSL•FO
RenderX

social media may develop a skewed perception of smoking norms, potentially impacting their attitudes and behavior regarding smoking cessation or prevention [18]. Therefore, by strategically incorporating positive health messages in social media content, public health campaigns can leverage the cultivation theory to influence people's beliefs and behaviors in favor of healthier choices.

Increasingly, more Americans use social media platforms, according to a Pew Research Center Study in 2021, with vast demographic differences per platform [19]. A total of 81% of all respondents used YouTube (Google LLC) while 95% of adults aged 18 to 29 years used the platform. The study also reported racial and ethnic differences. Higher rates of WhatsApp (Meta Platforms, Inc) use were reported among Hispanic respondents, with 49% of them using WhatsApp compared with only 16% of White respondents. A total of 49% of Black respondents used Instagram (Meta Platforms, Inc) and 74% used Facebook compared with only 35% of White residents using Instagram and 67% of them using Facebook. Given that misinformation is reported to be more easily spread via social media [20] and Black, American Indian or Alaska Native, and Hispanic or Latino communities report heavier use of social media, public health professionals could harness cultivation theory effects to increase positive social media messaging that combat misinformation in social networks frequented by Black, American Indian or Alaska Native, and Hispanic or Latino individuals.

Attribution theory [21] focuses on how individuals attribute causes to the behavior of themselves and others. It suggests that people tend to make inferences about the underlying causes of behavior based on 3 main dimensions: internal factors (eg, personal characteristics), external factors (eg, situational factors), and stability (eg, whether the behavior is consistent or temporary). In public health messaging, attribution theory can play a role in understanding how people interpret health-related behaviors and social media messages shared on social media. Individuals may attribute the behavior of others, such as engaging in health activities or practicing preventive measures, to internal factors (eg, personal motivation or responsibility) or external factors (eg, societal norms of situational influences) [22]. Public health campaigns can leverage the attribution theory by highlighting positive health behaviors and framing them in a way that encourages attributions to internal factors (eg, personal agency) to motivate individuals to adopt healthier behaviors themselves.

Together, cultivation and attribution theories reflect how health professionals can develop more effective social media strategies to engage and influence their audience's health behaviors and shape their perceptions related to health issues.

Objectives

This implementation report aimed to describe the process used to develop social media content and deploy multimedia strategies to communicate with 3 different populations: Black, American Indian or Alaska Native, and Hispanic or Latino communities on digital platforms, as guided by these psychological frameworks. Furthermore, we present data from a comprehensive evaluation of our processes' effectiveness. Specifically, we provide quantitative data obtained from social media metrics and qualitative data obtained from focus groups with community members from the 3 communities. This information provides evidence that addresses some of the gaps in knowledge about using social media in health promotion [13] and provides a blueprint for other organizations seeking to partner with communities of color to address the community's informational needs.

Methods

The Communications Campaign

Team Formation

The core team of this project consisted of a media specialist with experience in digital campaigns, an expert on mass communications for communities of color, and community members to serve the role of community influencers (later self-renamed as community advocates). The community advocates did not need to have experience in using social media, and the main requirement was to have connections within their self-identified community and a desire to share information with their community. For the Latino community, we also partnered with members of the Latino Health Council of Dane County (a nonprofit organization that convenes organizations providing services to Latino communities), who created and disseminated additional content. As the project evolved, team members were added to the core team to provide expertise and assistance in translating scientific findings into layperson language, including a team scientist who belongs to one of the communities of interest with experience in presenting scientific information to lay audiences and health care professionals (eg, emergency medicine physicians, medical students, and other health professionals).

Platforms and Pages

To reach the Black community in Wisconsin, in August 2020, we created a Facebook (Meta Platforms, Inc) page titled "COVID-19 Information for the Black Community" [23], and in January 2022, we created an Instagram (Meta Platforms, Inc) professional account titled @blackwellness. For reaching members of the Oneida Nation, a federally recognized American Indian tribe in Wisconsin, we created a Facebook page titled "COVID-19 Information for the Oneida Nation" [24], which was usually shared or reshared by the Oneida Nation's official Facebook page (that has >14,000 followers). For the Latino community, all the posts and content were disseminated through the social media channels of the Latino Health Council of Dane County (Twitter [Twitter, Inc; subsequently rebranded as X, X Corp], Instagram, Facebook [25], and YouTube [Google LLC]); this organization was selected based on long-standing partnerships and the intention to synergize with existing education initiatives around the COVID-19 pandemic that they were implementing. The selection of the platforms was informed by the community advocates and based on the use of the platform by each community in the United States from 2020 to 2021, which, according to Pew Research Center, were Facebook and YouTube, with an increased use of Instagram by those aged <30 years [19].

XSL•FO
RenderX

Selection of Topics

Monthly meetings were scheduled with the principal investigators (MM, MMP, and CEG) and the team to define the topics, which were based on upcoming national and cultural holidays (eg, Indigenous Peoples' Day and Dia de los Muertos), relevant events (eg, the start of the school year and vacations), and new developments related to the COVID-19 pandemic (eg, new recommendations from the Centers for Disease Control and Prevention). The advocates' assessment of their communities' most pressing concerns was the primary driver for the topics selected.

Development and Dissemination of Content

The core team met weekly to prepare posts, evaluate performance, and develop new communication strategies. Advocates were asked to post content weekly on their platform of choice. Advocates were never required to post on topics to propagandize or promote a specific agenda. Developing content was iterative but started with creating a scientifically informed response to the topic. To support their posts, the group, including physicians and other professionals, met weekly to clarify information and ensure that the content was medically accurate, culturally relevant, and evidence based. This often included a weekly presentation by our team scientist (TTJ) on the week's topics to address gaps in knowledge of the scientific materials prepared in layperson format. The team scientist stayed abreast of local and national news on the COVID-19 pandemic and pertinent health topics and searched social media spaces targeted to communities of color to understand myths and disinformation surrounding the COVID-19 pandemic to better serve the communities. All team members had to feel comfortable with the content being disseminated as it was to ensure authenticity.

Finally, the community advocates (TE, AB, LS, VW, and SL) drafted their posts using this rapport. However, they were directly supported by the communications team to create graphics and tailor the social media message to align with the individual community. Community advocates also received training on communication tools and resources to improve their content (eg, video editing and creating infographics). Figure 1 shows a graphic representation of the process used to develop content and the people involved in each step of the process.

Some of the social media messages posted included how vaccines work, updates on approvals of vaccines, correct use of masks, mental health checkups, and interviews with experts (Multimedia Appendix 1). Advocates posted content on the platforms directly or through a social management platform. The hashtags used and the day and time of the posting were agreed upon with the communications team based on the performance of previous posts. We did not use paid advertising or boosts to disseminate the social media messages.

Figure 1. Graphical representation of the content creation process, including team roles and responsibilities.

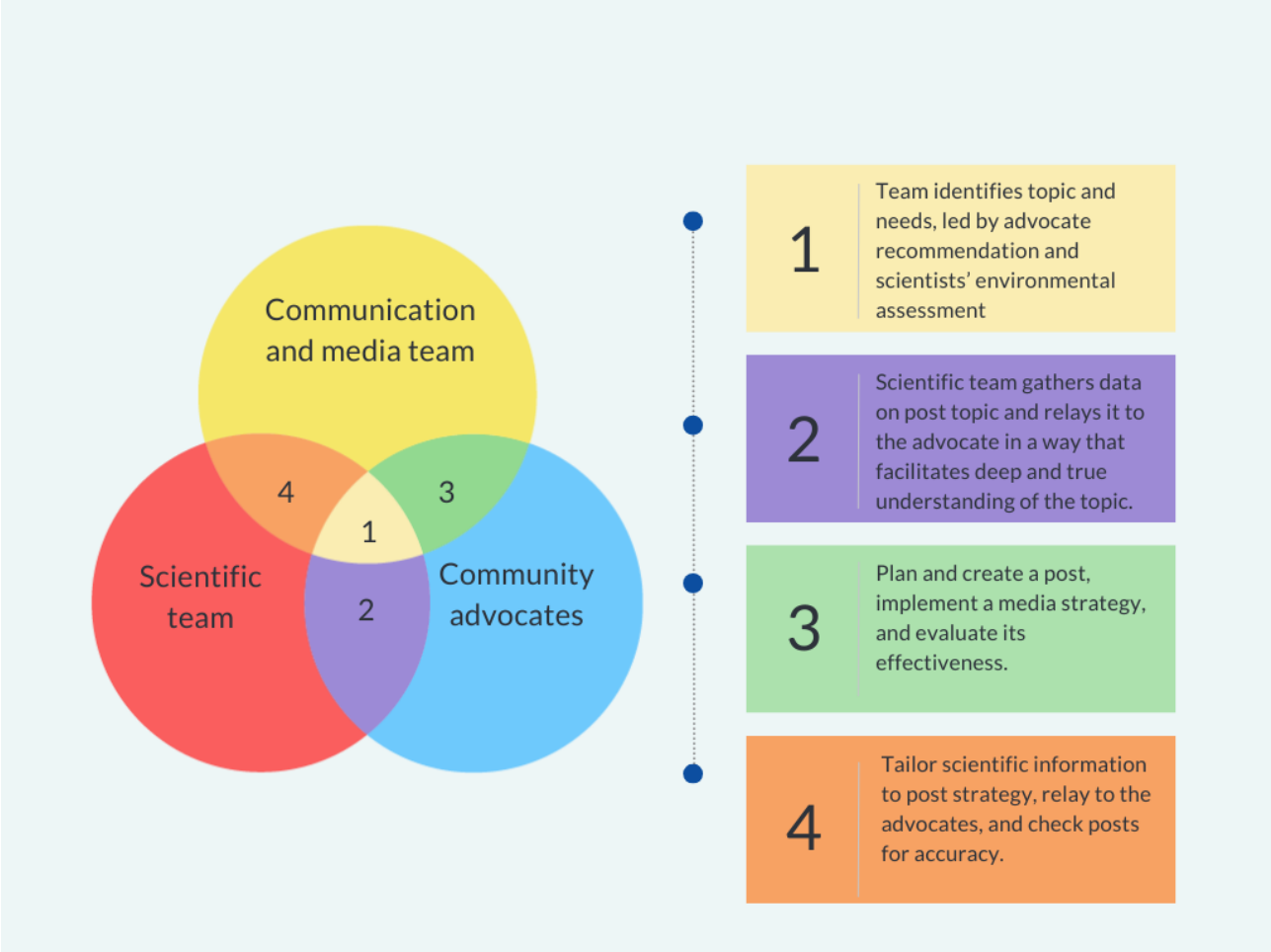

XSL•FO
RenderX

Data Collection and Analysis

Quantitative Data

We evaluated the success of our digital media efforts by monitoring reach and engagements. We used the analytic tools provided by the following platforms: Facebook and Sprout Social (Sprout Social, Inc). In March 2023, once the study was completed, we downloaded analytic reports for each community page's study period. Because the Latino pages included content created by additional volunteers and other sources, we selected a random sample of posts (1 per week during the study period) to collect the analytics of interest in the study.

The metrics of interest in this study are defined as follows:

1. *Reach*: the total number of unique individuals who see the content from the account
2. *Engagements*: unique individuals engaged with the content, either by liking the post, following the page or account, sharing the content, or clicking one of the links in the post.

Because metrics related to reach and engagement are dependent on the number of followers per page, we analyzed 2 rates to be able to compare the posts' performance across different communities and platforms: engagement rate per reached person (ERR, number of engagements divided by the number of persons reached) and average reach per post (ARPP, number of persons reached divided by the number of posts). We obtained the 95% CI for these rates using the MedCalc (MedCalc Software Ltd) software [26]. We conducted comparative analysis of the ERR according to the post type (link, live, photo or image, text, and video); and language (bilingual, English, and Spanish). Statistical significance was assessed using *P* values, which were calculated using chi-square tests or Fisher exact tests, where values <.05 were considered statistically relevant. Our results are presented separately for each community and platform, and no direct comparisons were made across communities.

Qualitative Data

Data were collected through 9 focus groups (3 per community). The interview guide was developed by the community advocates, communication specialist, and author MMP (coprincipal investigator) to ensure that the questions addressed the main concerns of the team. The interview guide was revised by CG and MM (coprincipal investigators) and further edited by experts in qualitative approach who facilitated the focus groups.

Participants in the focus groups were recruited through flyers and short videos the advocates posted on our social media platforms. In addition, we asked partner community organizations to share our posts to increase the reach and share via other platforms used by the community (eg, group chats).

Interested individuals were provided with a link, which directed them to a REDCap (Research Electronic Data Capture; Vanderbilt University) electronic data form hosted at the University of Wisconsin–Madison Department of Medicine [27]. The form collected information to assess eligibility for participation (eg, age, race or ethnicity, and use of social media platforms) and contact information.

Eligibility criteria included self-identifying as a Black or African American, American Indian or Alaska Native, or Hispanic or Latino individual; being aged >18 years; and having self-reported awareness of the community pages created for this study. For one of the focus groups of each community, we only included participants who reported following the social media page of interest.

The focus groups were audio recorded, and verbatim transcripts were made from the audio recordings, ensuring that these were deidentified for analysis. Data were analyzed using an inductive approach to code the transcripts, and the themes in this report were generated from participants' answers rather than defined in advance. Coding was done with NVivo (release 1.7; Lumivero).

Ethical Considerations

This study was not considered human subject research by the University of Wisconsin–Madison Institutional Review Board and was classified as quality improvement or program evaluation not requiring an institutional review board approval in accordance with federal regulations, as defined under 45 Code of Federal Regulations Part 46.102(d). This report follows the recommendations of the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) [28] guidelines, and the checklist is available in Multimedia Appendix 2.

The research team contacted individuals to obtain informed consent and provide focus group details. During the introduction of the section, participants were reminded about the volunteer nature of participation and provided additional opportunities to opt out. Participants received US $60 to US $75 as a token of appreciation for participating in the 90-minute focus group; the specific amount changed across the communities based on the timing of the recruitment.

Results

Use of Social Media

Overview of Focus Groups

A total of 9 focus groups were hosted across the 3 communities; 7 (78%) were via videoconferencing (Zoom; Zoom Video Communications), and 2 (22%) were conducted in person. Among a total of 59 individuals who participated, 45 (76%) were women and 19 (32%) self-identified as members of the Oneida Nation, 18 (30%) as Black or African American individuals, and 21 (36%) as Hispanic or Latino individuals. Classification according to the age group of participants is as follows: 17% (10/59) were aged 18 to 24 years, 20% (12/59) were aged 25 to 34 years, 31% (18/59) were aged 35 to 44 years, 9% (5/59) were aged 45 to 54 years, and 21% (12/59) were aged >55 years.

The following paragraphs reflect participants' responses to different topics. Where dialogue is needed to convey context or meaning, focus group participant and moderator responses are indicated by "P" and "M," respectively. For quotations that stand alone (ie, without dialogue or a speaker tag such as "P"), it can be assumed that the speaker is a focus group participant.

XSL•FO
RenderX

The racial or ethnic identity of the focus group participant is indicated at the end of the quotation with a parenthetical code that includes the participant's race or ethnicity and focus group number (Black or African American, Oneida Nation, or Hispanic or Latino individual).

Reasons to Use of Social Media

Participants liked using social media to keep up with people they knew, as stated as follows:

> *I like seeing other people's pictures, and what they're up to, how they've been.* [Oneida Nation individual, 1]
>
> *I keep in touch, I have different kinds of groups I have. I have my high school folks that I keep in touch with. I have my folks that I went to [college name] with. I have my church family. I have friends that I went to school with.* [Black or African American individual, 1]
>
> *I use Facebook, but more for its Marketplace or groups that I follow. It's also the only way I can see a certain part of my family.* [Hispanic or Latino individual, 1]

Participants liked and shared various posts, including memes, posts that make them laugh, posts about sports, and community events. Some participants talked about using social media to obtain information about the COVID-19 pandemic:

> *I don't really watch the news, so I use it a lot for keeping up with information about COVID, different health things with COVID, keeping up with the numbers of COVID, the guidelines, the things that I should and shouldn't be doing as far as COVID.* [Black or African American individual, 1]
>
> *I kind of just look at the page for myself to keep myself up to date.* [Black or African American individual, 3]
>
> *I usually share a lot of information about benefit to the community. So, I serve as a source of publicity, so to speak, or have served my family.* [Hispanic or Latino individual, 1]

Kinds of Posts That Participants Disliked

One thing many participants did not like about Facebook in particular was the number of unsolicited advertisements on their wall; they usually ended up blocking them even when they may be of interest. They felt that these advertisements were intrusive:

> *What I don't like is that messages appear that no one is asking for... I really have to be blocking many messages that are not what I have looked for.* [Hispanic or Latino individual, 1]

Similarly, individuals were asked what type of information or strategy would make posted content more persuasive, and they expressed that a lack of authenticity from the author would decrease the likelihood of changing their existing beliefs.

> *...that post actually sounds like that lady was paid to post that. So that wouldn't change my mind to... I mean, it just sounds too— I don't know, it just sounds too payish, like she got paid to write it, like it was supposed to sound perfect.* [Black or African American individual, 2]

Overall Use and Engagement Metrics

Between August 2021 and January 2023, we created 980 unique social media posts that reached 88,790 individuals and gathered >6700 engagements. Figure 2 describes the overall number of posts per community and platform and their associated social media metrics. Posts directed to members of the Black community showed an ARPP comparable with the other Facebook communities in this project (111.74 for Black or African American vs 119.46 for Hispanic or Latino vs 113.11 Oneida Nation) despite the lowest overall post count. Within the Latino community, Instagram posts had a higher ERR (9.2% vs 3.5% vs 3.2%; $P<.001$) despite the lower number of posts and reach. Multimedia Appendix 3 shows the demographics of the followers of each of the community platforms.

Some participants reported that they might not engage with a post because they want to avoid notifications or well-intended inquiries from others who can see their interaction.

> *I think, for me, if I were feeling anxiety or depressed or something, I probably would not. But if I was in a good mood, I'd probably like it or send a heart. Because... if you feel something that's down, you're going to get people saying, are you okay, or calling you or texting you, and I don't know that I'm always ready for that. People read into your reaction sometimes. I think about what others think. I know that. So sometimes I may not react at all because I don't want feedback.* [Black or African American individual, 1]

XSL•FO
RenderX

Figure 2. Overall performance of posts according to community (July 2021 to January 2023).

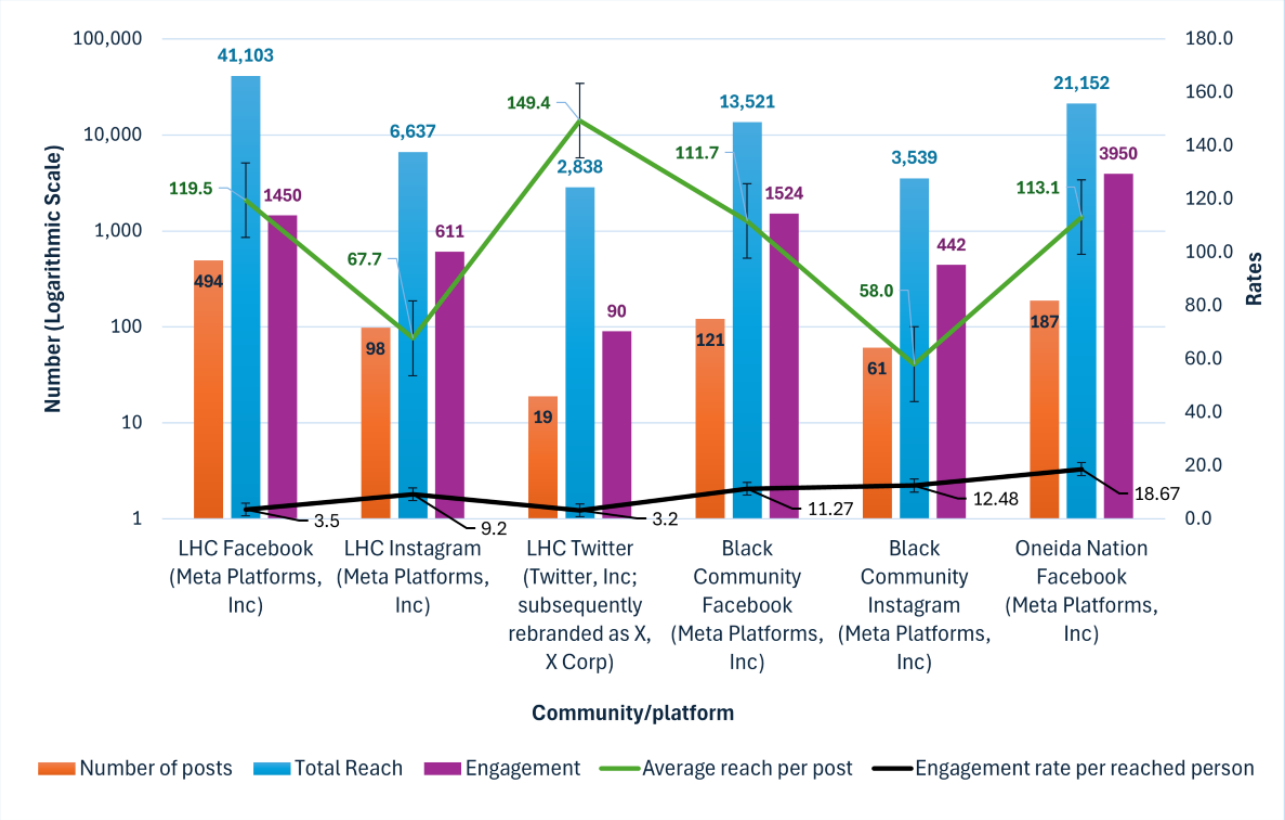

Reach and Engagement of Posts for Latino Communities

Overview

Table 1 presents the social media metrics according to the post type per platform. On Facebook, live videos had the highest ARPP and ERR (P<.001), while posts that comprised only social media messages had the lowest engagement. Meanwhile, on Instagram, videos had the highest ARPP (94.07) and photos or images had the highest ERR (10.17%; P=.002).

Focus group participants mentioned that combining graphics or images with text is more effective at catching the attention of Hispanic or Latino individuals rather than just catchy phrases with large fonts:

> *I am quite a very visual person. If I see this photo, it catches my attention and I say, if they went [to get vaccinated], why don't I go? So, our Latino community is quite visual, if you put a photo, it will reach you much more than put some big letters... That is to say, everything that we put in pure letters does not reach as much as a photo.* [Hispanic or Latino individual, 1]

XSL•FO
RenderX

Table 1. Metrics for Latino community platforms according to the type of posts (August 2021 to January 2023).

Community	Posts, n (%)	Total reach, n	Total engagement, n	ARPP[a] (95% CI)	ERR[b] (%; 95% CI)	*P* value
Facebook (n=135)						
Link	29 (21.5)	2306	193	79.52 (76.31-82.83)	8.37 (7.23-9.64)	<.001
Live	13 (9.6)	3871	364	*297.77 (288.46-307.3)* [c]	*9.40 (8.46-10.42)*	—[d]
Photo or image	76 (56.3)	11,734	454	154.39 (151.61-157.21)	3.86 (3.52-4.24)	—
Text only	2 (1.5)	300	9	150 (133.5-167.97)	3.00 (1.37-5.69)	—
Video	15 (11.1)	1471	53	98.07 (93.12-103.21)	3.60 (2.69-4.71)	—
Instagram (n=98)						
Photo or image	70 (71.4)	4003	407	57.18 (55.43-58.99)	*10.17 (9.2-11.2)*	.002
Video	28 (28.6)	2634	204	*94.07 (90.51-97.73)*	7.75 (6.72-8.88)	—

[a]ARPP: average reach per post.

[b]ERR: engagement rate per reached person.

[c]Italicized values indicate the highest category in the group.

[d]Not applicable.

Language

Table 2 shows that on Facebook, posts that were only in English had the largest ARPP, while those in Spanish had the highest ERR (6.99%; *P*<.001). On Instagram, there was no statistical difference in ARPP and ERR according to language (*P*=.10).

In contrast, focus group participants found that posts in both languages were more impactful because they were more accessible to a larger audience. Some participants felt that it is a misconception that to be a Latino, you have to speak Spanish; many individuals in the younger generations of Latinos have lost fluency of the Spanish Language and are uncomfortable with being excluded because of it. In contrast, many felt it harder to translate English-only posts to friends and family, as information is lost in translation. Many people found sharing and discussing content with family more appealing when there was a picture and content in English and Spanish:

> *Personally, I like it [bilingual posts] because I find it accessible. That is because perhaps there are terms that I do not handle in one language, and in another [language], that same term has more context. And that makes me able to use both languages to better understand the content.* [Hispanic or Latino individual, 1]

Table 2. Metrics for Latino community platforms according to the language of posts (August 2021 to January 2023).

Platform	Posts, n (%)	Total reach, n	Total engagement, n	ARPP[a] (95% CI)	ERR[b] (%; 95% CI)	*P* value
Facebook (n=135)						
Bilingual	6 (4.6)	519	31	86.5 (79.22-94.27)	5.97 (4.06-8.48)	<.001
English	28 (21.4)	5029	40	*179.61 (174.68-184.64)* [c]	0.79 (0.56-1.08)	—[d]
Spanish	97 (74)	11,818	827	121.84 (119.65-124.05)	*6.99 (6.53-7.49)*	—
Instagram (n=98)						
Bilingual	31 (31.6)	2544	249	82.07 (78.91-85.32)	9.79 (8.61-11.08)	.10
English	26 (26.5)	2203	177	*84.7 (81.23-88.35)*	8.03 (6.89-9.31)	—
Spanish	41 (41.8)	1890	185	46.09 (44.04-48.22)	*9.79 (8.43-11.31)*	—

[a]ARPP: average reach per post.

[b]ERR: engagement rate per reached person.

[c]Italicized values indicate the highest rate in the category.

[d]Not applicable.

Reach and Engagement of Posts for Black Communities

Table 3 presents the social media metrics according to the type of posts of each platform for the Black community. On Facebook, posts with photos had the largest ARPP (132.85), while posts containing only text had the lowest; videos had the largest ERR (19.53%; *P*<.001). For Instagram, videos had the largest ARPP, and carousel posts had the largest ERR (*P*<.001).

XSL•FO
RenderX

Table 3. Metrics for platforms for the Black community according to the type of post (August 2021 to January 2023).

Type of post	Posts, n (%)	Total reach, n	Total engagement, n	ARPP[a] (95% CI)	ERR[b] (%; 95% CI)	*P* value
Facebook (n=121)						
Link	19 (15.7)	657	83	34.58 (31.98-37.32)	12.63 (10.06-15.66)	<.001
Photo or image	74 (61.2)	9831	851	*132.85 (130.24-135.5)* [c]	8.65 (8.08-9.25)	—[d]
Text only	2 (1.7)	23	2	11.5 (7.29-17.26)	8.69 (1.05-31.4)	—
Video	26 (21.5)	3010	588	115.77 (111.67-119.98)	*19.53 (17.99-21.18)*	—
Instagram (n=61)						
Carousel	11 (18)	426	85	38.73 (35.14-42.59)	*19.95 (15.94-24.67)*	<.001
Photo or image	22 (36.1)	919	118	41.77 (39.12-44.56)	12.84 (10.63-15.38)	—
Video	28 (45.9)	2194	239	*78.36 (75.11-81.71)*	10.89 (9.56-12.37)	—

[a]ARPP: average reach per post.

[b]ERR: engagement rate per reached person.

[c]Italicized values indicate the highest rate in the category.

[d]Not applicable.

All participants reacted favorably to the example video played [29]. Participants cited the following reasons why this video caught their attention:

- It uses a popular song:

  It's a Beyonce song. And then I'm going to automatically read the information, and then I'm going to sing the song while I'm at it. [Black or African American individual, 1]

- It was funny:

  I got a good laugh out of it [Black or African American individual, 1]

- It feels comfortable and culturally familiar:

  It's easily hearable because like as for me and my house, like it feels familiar. It feels comfortable. [Black or African American individual, 1]

  I like to see things that appeal to me and my culture. [Black or African American individual, 1]

Reach and Engagement of Posts for the Oneida Nation Page

Table 4 shows the social media metrics for the Facebook page "COVID-19 Information for Oneida Nation Community" according to the post type. Video posts had an ARPP that was statistically significantly larger than other posts (179.27; *P*<.001), and the largest ERR was seen on posts that had only text.

All Oneida Nation members who participated in the focus groups were positive about posts that used images or videos. Several participants said they preferred videos that were ≤2 minutes in length, with 1 participant checking the video length first, skipping portions of long videos, and watching just the ending:

> *P7: ...like tops that I want to watch a video for is maybe like 1.5 or 2 minutes if it's something good....*
>
> *P8: I would say 2 minutes....*
>
> *M: Okay. Why 2 minutes?*
>
> *P8: It's short and to the point.*
>
> *P6: I'd say like a minute or 2 minutes, yeah, like everybody else was saying. Just get to the point, you know. Nobody really wants to sit there and have to scroll through or just watch it, you know. Really, if you really could just have it short, get to the point, it will be more interesting. More people will probably watch it....*
>
> *P5: I'd say 1-2 minutes as well.* [Oneida Nation individual, 1]

Participants stated that it is especially important that a video be short if it is a topic that people are not following or not already interested in:

> *If it's something like, you know, you actually like, I mean, I guess you're going to watch it no matter how long it's going to be. But if it's something that you really don't like, or you're not really interested in, I guess you're not going to have no—you know, you're not going to waste time and watch it.* [Oneida Nation individual, 1]

XSL•FO
RenderX

Table 4. Metrics for the Facebook (Meta Platforms, Inc) page for the Oneida Nation according to the type of post (August 2021 to January 2023).

Post characteristic	Posts (n=187), n (%)	Total reach, n	Total engagement, n	ARPP[a] (95% CI)	ERR[b] (%; 95% CI)	*P* value
Link	21 (11.2)	64	846	40.3 (37.6-43.09)	7.57 (5.82-9.66)	<.001
Photo or image	102 (54.5)	1509	9398	92.13 (90.28-94.02)	16.06 (15.3-16.89)	__[d]
Text only	10 (5.3)	724	1227	122.7 (115.93-129.76)	*59.01 (54.79-63.46)* [c]	—
Video	54 (28.9)	1653	9681	*179.27 (175.72-182.89)*	17.07 (16.26-17.92)	—

[a]ARPP: average reach per post.

[b]ERR: engagement rate per reached person.

[c]Italicized values indicate the highest rate in the category.

[d]Not applicable.

Perception of the Community Individuals to Our Social Media Campaigns

Everyone familiar with the COVID-19 Info for Madison's Black Community Facebook page praised it for having reliable and relatable information:

> *It's informational and it's easy to read and to understand what's there, which is part of what I think a lot of the issue is, is people don't necessarily like reading. And so the graphics and the wording is very clear, and anybody can understand it, no matter what their educational background is.* [Black or African American individual, 1]

Similarly, there was uniform favorability toward posts that offered resources and ideas to promote health. Participants cited the benefits to themselves, their families, and their communities:

> *I think they're helpful resources. They're just there, you know. You can take a look, take a glance. If you do need that type of information, then it's there for you to get it open to the public, you know, good information.* [Oneida Nation individual, 1]

> *It's kind of like you hear on the news, like when they have recalls for food and stuff, it's like very informational things that I would like to know. So, I mean, everybody kind of benefits from posts like that.* [Oneida Nation individual, 1]

Discussion

Principal Findings

Our study shows that social media communications for health promotion were favorably perceived by individuals from Black, American Indian or Alaska Native, and Hispanic and Latino communities when produced by reliable sources with whom they are already connected in real life. Similarly, there is evidence that social network sites are becoming essential to how people experience news, moving from a traditional news cycle that journalists primarily dominate to more complex information cycles that incorporate ordinary people within the process [30].

In our study results, individuals reflected that social media platforms are used for recreational purposes and staying up to date with family and friends. Prior research indicates that individuals from marginalized communities also use social media to find health information [31,32]. Given that racial and ethnic minority groups use these media platforms more than White individuals [13], our study developed processes to meet these populations on digital platforms in a way that promoted trusted and accurate dissemination of public health information. This model is distinct from prior studies using social media influencers as the community influencers were not merely tools for disseminating public health social media messaging created by officials and scientists [33-35].

As described in the *Methods* section, our community influencers determined what content was relevant to their communities and were empowered through skill building to be codrivers at every stage of the layperson translation process, from strategy to content creation to dissemination and analysis. This approach is similar to previous descriptions of tailored influencer marketing approaches that harness the social influencers' experiences and competencies [36]. An additional key distinguishing feature of our model is detailed knowledge sharing between the scientists and the community influencers, which led to the self-renaming of the influencers to advocates as they began to take ownership and gain a proper understanding of the scientific content being shared on their pages. According to the attribution theory, these interactions within our team affected how the information was presented and ensured that the most relevant information for the audience was presented first, which might, in turn, influence the audience's perception of the stability and controllability of health information, leading to a higher likelihood of message acceptance and behavior change.

Our results also highlight that there is not 1 unique strategy that works across various Black, American Indian, Alaska Native, and Hispanic or Latino communities. Health promotion social media messages must be tailored to the audience of interest and the platform used. Facebook audiences were mostly middle-aged individuals, while Instagram had more young adults and teenagers, which is consistent with the existing literature [37]. The Instagram account also reached a higher percentage of men than Facebook and even reached other communities of color in low- and middle-income countries (Colombia, India, Mexico, Nigeria, and Venezuela).

To increase the dissemination of social media messages, these must be easy to understand and attention grabbing to generate interest in audiences while being authentic without creating the

XSL•FO
RenderX

sensation of spreading propaganda. The most successful strategies we identified to accomplish these goals were live streaming and videos. Videos had the advantage of overcoming literacy limitations, and when they displayed people, these created a sense of connection and authenticity with the audience, which partially addressed one of the limitations of social media: the lack of reliable sources [13]. These results are similar to those obtained by other studies that reflect that the content needs to be attractive and easy to understand to be effective [38].

The focus group data indicated that people use social media to keep in touch with people they already know or are connected to, which suggests that for future endeavors, it is paramount to identify an advocate who is highly connected to their community, as the audiences would be more likely to pay attention and trust the information presented.

Unsurprisingly, given the historical abuse endured by marginalized communities, trust and authenticity emerged as critical issues in our focus groups. The choice to empower the advocates to drive and create the content is a novel and distinguishing feature of our model and was instrumental in facilitating trust and ensuring authenticity [36]. This approach led to content production that aligned with the personal brand of each advocate and facilitated connectedness to each community, which was identified in the focus groups as important. These findings that social media communications are favorably perceived when produced by reliable sources align with cultivation theory, as consistent social media messaging from trusted sources could cultivate a perception of credibility and reliability in health promotion.

Comparison With Previous Work

There are multiple studies describing strategies for health promotion that engage community representatives in the dissemination of information to improve trust and facilitate the dissemination of social media messages [39], and that crossposting across multiple platforms is necessary to combat misinformation [40]. However, our study is the first to show how different content types behave across platforms, emphasizing the need to design public health content according to the platform and audience.

Previous studies have used impressions [41,42], reach [41], and engagements [41-45] or specific elements (eg, shares and comments) to assess the dissemination of social media messages. These metrics in raw form are a reflection of the number of followers and might not be an accurate representation of the reach of the messages, and there is limited evidence of how these metrics relate to subsequent behaviors; however, previous studies suggest that these are related. A study by Nadarzynski et al [38] showed that Facebook advertisements could lead to increased traffic to websites promoting chlamydia testing and related sexual health information, resulting in a 41% increase in test requests, showcasing how increases in engagement in social media can be associated with new behaviors.

Selecting the right metrics to evaluate outreach efforts must be individualized to the campaign's goal. On the basis of our results, we suggest using a combination of reach and engagement metrics, maximizing the use of tools available in social media professional accounts, and evaluating these metrics regularly to facilitate tailoring to the audience [14,37]. Contrary to common recommendations by social media influencers and marketing experts, in our experience, more posts did not lead to more engagement and reach. This might reflect the relevance of the social media messages to the audiences.

In our experience, producing large amounts of content at a fast pace does not allow enough time to create intentional and relevant social media messages and might affect individuals' recognition of the messages. According to Vakratsas and Ambler [46], a minimum of 3 exposures is often necessary for a social media message to have an impact, with diminishing returns after 10 to 20 exposures. The frequency of these social media messages and the complexity of the message also affect the effectiveness in promoting a new behavior or increasing interest in a product; according to a white paper by Facebook from 2016, one to two social media messages per week might be enough to generate recognition [47], which support the frequency selected for our messages.

Limitations

Although our study explores many perspectives, 1 limitation was that our focus groups engaged a small number of individuals across the communities, which might affect the richness of our results and potential generalizability to other communities and geographic locations. Considering that our study was performed in Wisconsin, it is important to acknowledge that context factors affect the perceptions and potential interactions that individuals have with our social media posts. For example, in Wisconsin, vaccination against COVID-19 was free to all individuals, and there was no requirement for proof of residence or US citizenship to access these services. This resulted in less fear of deportation and lower mistrust in vaccination social media messages compared with what could be seen in other localities in the United States, where these services were only available to US residents.

Another limitation of our study is that our metrics do not account for reach and engagements after our posts were shared. In July 2021, Facebook analytics changed how it collected and presented metrics for pages; since then, only the metrics of the page are shown, not accounting for engagement on third-party pages, resulting in lower overall numbers. Similarly, our results do not reflect what happened outside the social media platforms, such as sharing information in group chats, WhatsApp, or SMS text messages, which are platforms frequently used by our communities [13]. To ensure uniformity of the metrics, all data for the study period were downloaded in March 2023 and used for all the analyses.

Conclusions

This project created social media content that was medically accurate and culturally appropriate related to the COVID-19 pandemic and its physical and mental health consequences. One key aspect of our work is that we relied on building trustworthy relationships with partners and communities within and outside social media; this was achieved through the work of our advocates and team members who served as a connection with the communities. We identified that stories in videos, photos,

XSL•FO
RenderX

or messages that contained only text created meaningful connections with individuals and provided trustworthy information with very encouraging results that we have been sharing with academicians and community organizations. However, our results showcase how analytics reports of the social media platforms (eg, share, likes, and views) remain inadequate metrics for measuring behavioral changes in the target audiences [48] and require the development of further population-health approaches to identify whether social media campaigns such as ours can translate on increased adoption and maintenance of health behavior and subsequently improved health outcomes. Future studies could explore whether social media engagements are associated with changes in beliefs and the acquisition of new health behaviors to promote health and well-being and how social media campaigns can build on community partnerships to address misinformation or disinformation at the population level in ways that are culturally appropriate and result in improved health outcomes.

Acknowledgments

The authors would like to sincerely thank the other contributors, without whom this work would not have been possible. These invaluable individuals include Shiva Bidar Sielaff and the team of dedicated volunteers at the Latino Health Council of Dane County; Andrea Ochoa, Mirna Arroyo-Miranda, PhD, and Luis Medina, PhD, for their invaluable assistance in conducting the focus groups within the Latino communities; Nasia Safdar, MD, PhD, Ryan Westergaard, MD, Jasmine Zapata, MD, MPH, James Conway, MD, Shenikqua Bouges, MD, and Gerald Ignace, MD for generously sharing their time and expertise with our communities; Kenneth Cross and the remarkable team at the University of Wisconsin Survey Center; and William Hatton Swain, Ilan Kolkowitz, Daniel Summers, Hannah Glenn, and Maria del Carmen Rosales for their outstanding collaboration throughout this project.

This project was funded through a COVID-19 Response Grant from the Wisconsin Partnership Program at the University of Wisconsin School of Medicine and Public Health. MMP was supported during this study period by the University of Wisconsin Department of Family Medicine and Community Health Primary Care Research Fellowship, funded by grant T32HP10010 from the Health Resources and Services Administration. CEG's work is supported by resources and facilities at the William S. Middleton Memorial Veterans Hospital, Madison WI and NIH/NIA grants: AA-FAIM, R01 AG054059; ICARE, R01 AG062307; AMICA, R01 AG074231.

During the preparation of this work, the author or authors used Grammarly to proofread and improve the readability of the *Introduction* and *Discussion* sections. After using these tools, the author or authors reviewed and edited the content as needed and take or takes full responsibility for the content of the publication. No other artificial intelligence generative tool was used to generate text or create images.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MMP, OH, GL, MFM, and CEG conceptualized the study. MMP, GL, and AB conducted the formal analysis of the study. MMP and CEG collected resources for the study. MMP, OH, GL, TTJ, LS, VW, TE, SL, and AB curated the study data. MMP was involved in writing the original draft. MMP, PTG, FCO, and CEG supervised the study. MMP, MFM, and CEG conducted the project administration. OH, GL, TTJ, LS, and AB contributed to the visualization of the study. OH, GL, TTJ, AB, LS, VW, TE, SL, FCO, and PTG conducted the investigation for the study. OH and GL were involved in the software analysis of the study. MFM and CEG were involved in the validation and funding acquisition of the study. All the authors were involved in the methodology of the study and in reviewing and editing the manuscript.

Conflicts of Interest

CEG serves on the following scientific and external advisory boards: Stanford University Alzheimer’s Disease Center’s External Advisory Board; Scientific Advisory Board for Huntington Medical Research Institute; University of New Mexico’s Alzheimer’s Disease Center’s External Advisory Board; Advisory Board for documentary film produced in partnership with PBS for national broadcast, entitled Matter of Mind: My Alzheimer’s Chair; Institutional Data and Safety Monitoring Board (IDSMB) Arterial stiffness, Cognition and Equol (ACE) trial, funded by NIA (R01AG074971; PI: Akira Sekikawa); University of Pittsburgh Board; Scientific Advisory Board for UH2AG083258 study entitled “Research Infrastructure for the Study of Alzheimer's Disease and Alzheimer's Disease-related Dementias in Older Asian Americans.

Multimedia Appendix 1
Sample content created by the community advocates.
[PDF File (Adobe PDF File), 571 KB - jmir_v26i1e58581_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) checklist.
[DOCX File , 30 KB - jmir_v26i1e58581_app2.docx]

Multimedia Appendix 3
Demographics of the followers of the community platforms.
[DOCX File , 174 KB - jmir_v26i1e58581_app3.docx]

References

1. Lopez 3rd L, Hart 3rd LH, Katz MH. Racial and ethnic health disparities related to COVID-19. JAMA 2021 Feb 23;325(8):719-720. [doi: 10.1001/jama.2020.26443] [Medline: 33480972]
2. Benitez J, Courtemanche C, Yelowitz A. Racial and ethnic disparities in COVID-19: evidence from six large cities. J Econ Race Policy 2020;3(4):243-261 [FREE Full text] [doi: 10.1007/s41996-020-00068-9] [Medline: 35300199]
3. Rubin-Miller L, Alban C, Artiga S, Sullivan S. COVID-19 racial disparities in testing, infection, hospitalization, and death: analysis of Epic patient data. Kaiser Family Foundation (KFF). 2020 Sep 16. URL: https://www.kff.org/report-section/covid-19-racial-disparities-in-testing-infection-hospitalization-and-death-analysis-of-epic-patient-data-issue-brief/ [accessed 2024-04-29]
4. Artiga SC, Corallo B, Pham O. Racial disparities in COVID-19: key findings from available data and analysis. Kaiser Family Foundation (KFF). 2020 Aug 17. URL: https://www.kff.org/racial-equity-and-health-policy/issue-brief/racial-disparities-covid-19-key-findings-available-data-analysis/ [accessed 2024-04-29]
5. COVID-19: racial and ethnic disparities. Wisconsin Department of Health Services. 2021. URL: https://www.dhs.wisconsin.gov/covid-19/disparities.htm [accessed 2024-04-29]
6. Wrigley-Field E, Garcia S, Leider JP, van Riper D. COVID-19 mortality at the neighborhood level: racial and ethnic inequalities deepened in minnesota in 2020. Health Aff (Millwood) 2021 Oct 01;40(10):1644-1653 [FREE Full text] [doi: 10.1377/hlthaff.2021.00365] [Medline: 34524913]
7. Tung EL, Peek ME, Rivas MA, Yang JP, Volerman A. Association of neighborhood disadvantage with racial disparities in COVID-19 positivity in Chicago. Health Aff (Millwood) 2021 Nov 01;40(11):1784-1791 [FREE Full text] [doi: 10.1377/hlthaff.2021.00695] [Medline: 34724418]
8. McCormack G, Avery C, Spitzer AK, Chandra A. Economic vulnerability of households with essential workers. JAMA 2020 Jul 28;324(4):388-390 [FREE Full text] [doi: 10.1001/jama.2020.11366] [Medline: 32556217]
9. COVID-19: vaccine data. Wisconsin Department of Health Services. URL: https://www.dhs.wisconsin.gov/covid-19/vaccine-data.htm [accessed 2024-04-29]
10. Khazanchi R, Evans CT, Marcelin JR. Racism, not race, drives inequity across the COVID-19 continuum. JAMA Netw Open 2020 Sep 01;3(9):e2019933 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.19933] [Medline: 32975568]
11. Sanchez GR, Bennett C. Why Spanish-language mis- and disinformation is a huge issue in 2022. Brookings. 2022. URL: https://www.brookings.edu/articles/why-spanish-language-mis-and-disinformation-is-a-huge-issue-in-2022/ [accessed 2024-04-29]
12. Hamel L, Lopes L, Sparks G, Stokes M, Brodie M. KFF COVID-19 vaccine monitor: April 2021. Kaiser Family Foundation (KFF). URL: https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-april-2021/ [accessed 2024-04-29]
13. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
14. Stellefson M, Paige SR, Chaney BH, Chaney JD. Evolving role of social media in health promotion: updated responsibilities for health education specialists. Int J Environ Res Public Health 2020 Feb 12;17(4):1153 [FREE Full text] [doi: 10.3390/ijerph17041153] [Medline: 32059561]
15. Mendoza-Herrera K, Valero-Morales I, Ocampo-Granados ME, Reyes-Morales H, Arce-Amaré F, Barquera S. An overview of social media use in the field of public health nutrition: benefits, scope, limitations, and a Latin American experience. Prev Chronic Dis 2020 Aug 06;17:E76 [FREE Full text] [doi: 10.5888/pcd17.200047] [Medline: 32762809]
16. COVID's devastation of Black community used as 'marketing' in new anti-vaccine film. NBC News. 2021. URL: https://www.nbcnews.com/news/nbcblk/covid-s-devastation-black-community-used-marketing-new-anti-vaxxer-n1260724 [accessed 2024-06-05]
17. Gerbner G, Gross L. Living with television: the violence profile. J Commun 1976;26(2):173-199. [doi: 10.1111/j.1460-2466.1976.tb01397.x] [Medline: 932235]
18. Yzer M, Rhodes K, Nagler RH, Joseph A. Effects of culturally tailored smoking prevention and cessation messages on urban American Indian youth. Prev Med Rep 2021 Dec;24:101540 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101540] [Medline: 34976614]

XSL•FO
RenderX

19. Social media use in 2021. Pew Research Center. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-04-29]
20. Su Y. It doesn't take a village to fall for misinformation: social media use, discussion heterogeneity preference, worry of the virus, faith in scientists, and COVID-19-related misinformation beliefs. Telemat Inform 2021 May;58:101547 [FREE Full text] [doi: 10.1016/j.tele.2020.101547] [Medline: 36570475]
21. Debanjan B, Chehak G, Rao TS S. The role of “attributions” in social psychology and their relevance in psychosocial health: a narrative review. Indian J Soc Psychiatry 2020;36(4):277-283. [doi: 10.4103/ijsp.ijsp_315_20]
22. Jackson M. Utilizing attribution theory to develop new insights into tourism experiences. J Hosp Tour Manag 2019 Mar;38:176-183. [doi: 10.1016/j.jhtm.2018.04.007]
23. blackwellnesswi Instagram page. 2021. URL: https://www.instagram.com/blackwellnesswi/ [accessed 2024-10-17]
24. COVID-19 Information for the Oneida Nation Facebook page. URL: https://www.facebook.com/covid19oneida [accessed 2024-10-17]
25. Consejo Latino Para la Salud Facebook page. 2014. URL: https://www.facebook.com/LatinoHealthCouncil [accessed 2024-10-17]
26. Confidence interval for a rate. MedCalc Software Ltd. URL: https://www.medcalc.org/calc/rate_ci.php [accessed 2024-10-17]
27. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
28. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
29. @blackwellnesswi: how are you keeping your young child(ren) safe? blackwellnesswi Instagram page. 2022 Mar 19. URL: https://www.instagram.com/reel/CbSYJRGlylM/?utm_source=ig_web_copy_link&igsh=MzRlODBiNWFlZA%3D%3D [accessed 2024-10-17]
30. Bergström A, Jervelycke Belfrage M. News in social media. Digit Journal 2018 Jan 12;6(5):583-598. [doi: 10.1080/21670811.2018.1423625]
31. Hills O, Shah D. Online health information seeking, medical care beliefs and timeliness of medical check-ups among African Americans. Patient Educ Couns 2020 Jun 18;103(12):2468-2476. [doi: 10.1016/j.pec.2020.06.006] [Medline: 32616321]
32. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625. [doi: 10.1177/0033354919874074] [Medline: 31513756]
33. Bozzola E, Staiano AM, Spina G, Zamperini N, Marino F, Roversi M, et al. Social media use to improve communication on children and adolescent's health: the role of the Italian Paediatric Society influencers. Ital J Pediatr 2021 Aug 11;47(1):171 [FREE Full text] [doi: 10.1186/s13052-021-01111-7] [Medline: 34380533]
34. Bonnevie E, Smith SM, Kummeth C, Goldbarg J, Smyser J. Social media influencers can be used to deliver positive information about the flu vaccine: findings from a multi-year study. Health Educ Res 2021 Jul 12;36(3):286-294 [FREE Full text] [doi: 10.1093/her/cyab018] [Medline: 34252187]
35. Al-Oraiqat AM, Ulichev OS, Meleshko YV, AlRawashdeh HS, Smirnov OO, Polishchuk LI. Modeling strategies for information influence dissemination in social networks. J Ambient Intell Human Comput 2021 Jun 28;13(5):2463-2477. [doi: 10.1007/s12652-021-03364-w]
36. Lutkenhaus RO, Jansz J, Bouman MP. Tailoring in the digital era: stimulating dialogues on health topics in collaboration with social media influencers. Digit Health 2019 Jan 28;5:2055207618821521 [FREE Full text] [doi: 10.1177/2055207618821521] [Medline: 30729023]
37. Korda H, Itani Z. Harnessing social media for health promotion and behavior change. Health Promot Pract 2013 Jan;14(1):15-23. [doi: 10.1177/1524839911405850] [Medline: 21558472]
38. Nadarzynski T, Burton J, Henderson K, Zimmerman D, Hill O, Graham C. Targeted advertisement of chlamydia screening on social media: a mixed-methods analysis. Digit Health 2019;5:2055207619827193 [FREE Full text] [doi: 10.1177/2055207619827193] [Medline: 30746155]
39. Lee A, Moore RC, Hancock JT. Designing misinformation interventions for all: perspectives from AAPI, Black, Latino, and Native American community leaders on misinformation educational efforts. Harvard Kennedy School. URL: https://misinforeview.hks.harvard.edu/wp-content/uploads/2023/02/lee_misinformation_interventions_for_-all_20230207.pdf [accessed 2024-04-29]
40. Cruickshank I, Ginossar T, Sulskis J, Zheleva E, Berger-Wolf T. Content and dynamics of websites shared over vaccine-related tweets in COVID-19 conversations: computational analysis. J Med Internet Res 2021 Dec 03;23(12):e29127 [FREE Full text] [doi: 10.2196/29127] [Medline: 34665760]
41. Ramos AK, Duysen E, Carvajal-Suarez M, Trinidad N. Virtual outreach: using social media to reach Spanish-speaking agricultural workers during the COVID-19 pandemic. J Agromedicine 2020 Sep 12;25(4):353-356. [doi: 10.1080/1059924x.2020.1814919]

XSL•FO
RenderX

42. Maafs-Rodríguez A, Otis B, Mattei J. Cultural adaptation and social media promotion of healthy eating guides for Spanish speakers. J Nutr Educ Behav 2022 Sep;54(9):863-871. [doi: 10.1016/j.jneb.2022.03.008] [Medline: 35750617]
43. Andrade EL, Evans WD, Barrett N, Edberg MC, Cleary SD. Strategies to increase Latino immigrant youth engagement in health promotion using social media: mixed-methods study. JMIR Public Health Surveill 2018 Dec 19;4(4):e71 [FREE Full text] [doi: 10.2196/publichealth.9332] [Medline: 30567689]
44. Edney S, Bogomolova S, Ryan J, Olds T, Sanders I, Maher C. Creating engaging health promotion campaigns on social media: observations and lessons from fitbit and garmin. J Med Internet Res 2018 Dec 10;20(12):e10911 [FREE Full text] [doi: 10.2196/10911] [Medline: 30530449]
45. Gabarron E, Larbi D, Årsand E, Wynn R. Engaging social media users with health education and physical activity promotion. Stud Health Technol Inform 2021 May 27;281:789-793. [doi: 10.3233/SHTI210283] [Medline: 34042686]
46. Vakratsas D, Ambler T. How advertising works: what do we really know? J Mark 1999 Jan 01;63(1):26-43. [doi: 10.1177/002224299906300103]
47. Effective frequency: reaching full campaign potential. Facebook Marketing Science. URL: https://www.facebook.com/business/news/insights/effective-frequency-reaching-full-campaign-potential [accessed 2024-04-29]
48. Adewuyi EO, Adefemi K. Behavior change communication using social media: a review. Int J Commun Health 2016;9:109-116 [FREE Full text]

Abbreviations

ARPP: average reach per post
ERR: engagement rate per reached person
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by A Wani, L Oliveira, K Singh, A Ajibade, D Ghosh; comments to author 23.05.24; revised version received 17.07.24; accepted 25.09.24; published 10.12.24.

Please cite as:
Mora Pinzon M, Hills O, Levy G, James TT, Benitez A, Lawrence S, Ellis T, Washington V, Solorzano L, Tellez-Giron P, Cano Ospina F, Metoxen MF, Gleason CE
Implementation of a Social Media Strategy for Public Health Promotion in Black, American Indian or Alaska Native, and Hispanic or Latino Communities During the COVID-19 Pandemic: Cross-Sectional Study
J Med Internet Res 2024;26:e58581
URL: https://www.jmir.org/2024/1/e58581
doi:10.2196/58581
PMID:

©Maria Mora Pinzon, Ornella Hills, George Levy, Taryn T James, Ashley Benitez, Sacheen Lawrence, Tiffany Ellis, Venus Washington, Lizbeth Solorzano, Patricia Tellez-Giron, Fernando Cano Ospina, Melissa F Metoxen, Carey E Gleason. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characteristics and Popularity of Videos of Abusive Head Trauma Prevention: Systematic Appraisal

Luc Goethals[1,2], PhD; Victoria Prokofieva Nelson[3], PhD; Fabien Fenouillet[3], PhD; Karine Chevreul[2,4], MD, PhD; Manon Bergerat[1,5], MD; Christine Lebreton[6], MSc; Yacine Refes[1], PhD; Flora Blangis[1,5], PhD; Martin Chalumeau[1,5*], MD, PhD; Enora Le Roux[2,7*], PhD

[1]Obstetrical, Perinatal and Pediatric Epidemiology Research team (EPOPé), Centre of Research in Epidemiology and Statistics (CRESS), Université Paris Cité, Inserm, Paris, France

[2]Epidémiologie clinique-évaluation économique appliqué aux populations vulnérables (ECEVE), Université Paris Cité, Inserm, Paris, France

[3]Learning, Physical Activity and Health, Interdisciplinary Laboratory in Neuroscience, Physiology and Psychology, Université Paris Nanterre, Nanterre, France

[4]Kastafiore, Paris, France

[5]Child protection unit, Department of General Pediatrics, Necker-Enfants malades hospital, Assistance Publique des Hôpitaux de Paris (AP-HP), Université Paris Cité, Paris, France

[6]Synakène, Ivry-sur-Seine, France

[7]Unit of Clinical Epidemiology (CIC-EC 1426), Hôpital Universitaire Robert-Debré, Assistance Publique des Hôpitaux de Paris (AP-HP), Inserm, Paris, France

[*]these authors contributed equally

Corresponding Author:
Luc Goethals, PhD
Obstetrical, Perinatal and Pediatric Epidemiology Research team (EPOPé), Centre of Research in Epidemiology and Statistics (CRESS)
Université Paris Cité
Inserm
123 Bd de Port-Royal
Paris, 75014
France
Phone: 33 0699222119
Email: luc.goethals@inserm.fr

Abstract

Background: Numerous strategies for preventing abusive head trauma (AHT) have been proposed, but controlled studies failed to demonstrate their effectiveness. Digital tools may improve the effectiveness of AHT prevention strategies by reaching a large proportion of the adult population.

Objective: This study aimed to describe the characteristics of videos of AHT prevention published on the internet, including their quality content, and to study their association with popularity.

Methods: From a systematic appraisal performed in June 2023, we identified videos addressing the primary prevention of AHT in children younger than 2 years that were published in English or French on the internet by public organizations or mainstream associations. We analyzed the characteristics of the videos; their quality with the Global Quality Scale (GQS); and their association with an index of popularity, the Video Power Index, using multivariable quasi-Poisson modeling.

Results: We included 53 (6.6%) of the 804 videos identified. Videos were mainly published by public organizations (43/53, 81%). The median time spent on the web was 6 (IQR 3-9) years, the median length was 202 (IQR 94-333) seconds, and the median GQS score was 4 (IQR 3-4). Infants were often depicted (42/53, 79%), including while crying (35/53, 66%) and being shaken (21/53, 40%). The characterization of shaking as an abuse and its legal consequences were cited in 47% (25/53) and 4% (2/53) of videos, respectively. The main prevention strategies in the videos were to raise awareness of the noxious outcome of shaking (49/53, 93%) and convince viewers of the effectiveness of coping strategies for infants' cries (45/53, 85%). The Video Power Index was positively correlated with the GQS (r=0.38; P=.007) and was independently associated with depicting an infant being shaken (P=.03; β=1.74, 95% CI 1.06-2.85) and the use of text or headers (P=.04; β=2.15, 95% CI 1.08-4.26).

XSL•FO
RenderX

Conclusions: AHT prevention videos had high quality but did not frequently deal with parental risk factors. The characteristics identified as being associated with the popularity of AHT prevention videos could help improve the impact of future prevention programs by enhancing their popularity.

(J Med Internet Res 2024;26:e60530) doi:10.2196/60530

KEYWORDS

abusive head trauma; child physical abuse; shaken baby syndrome; SBS; primary prevention; web-based videos; digital tools; head trauma; prevention; video; internet; infant; mortality; morbidity; parent; caregivers; communication

Introduction

Abusive head trauma (AHT), the most severe form of child physical abuse, is defined as "an injury to the skull or intracranial contents of an infant or young child due to inflicted blunt impact and/or violent shaking" [1]. AHT most often occurs in infants younger than 2 years, at a mean age of 2 to 4 months [2-5], and its incidence has been estimated at 14 to 34 per 100,000 infants per year in countries with developed economies (CDE) [6-9]. Perpetrators of AHT are most often adult men acting alone [10]. Persistent crying is reported as the primary trigger for most AHT perpetration [11]. AHT primary prevention is pivotal because its mortality rate ranges from 7% to 26% [12-14], and nonlethal forms are associated with severe long-term morbidity, notably neurodevelopmental impairment [15].

Numerous AHT prevention strategies have been proposed, targeting the public, parents, or caregivers of children through mass communication campaigns (brochures, television, and radio) or specific programs that mainly rely on educating about the management of infants' cries and the dangers of shaking a child [16,17]. These prevention programs also emphasize emotional regulation, particularly through the promotion of coping strategies when dealing with stress or frustration in front of infants' persistent crying [16]. The effectiveness of these strategies has been evaluated in several controlled trials, and although some individual studies have reported reductions in AHT incidence [18-20], meta-analyses and the US Preventive Services Task Force have concluded no significant summed preventive effects [16,21-24]. This discrepancy may be due to differences in the study design used to measure the impact of the intervention and the level of proof used to define effectiveness [24]. More generally, despite all these efforts, the incidence of AHT is not declining in CDE, particularly in the wake of the COVID-19 pandemic and its prevention and mitigation measures [25,26]. Consequently, there is a pressing need for innovative approaches that can reach a wider population and lead to effective behavioral change among the parents and caregivers of tens of millions of newborns each year [27].

Public organizations increasingly diffuse prevention videos on wide-audience digital platforms to support and educate new parents [28]. This approach can reach a large part of the population and become an effective vector for health promotion and prevention [29]. Thus, internet videos could participate in the effective prevention of AHT [30,31]. Young adults are major consumers of internet videos [32] and frequently use the internet to search for health information [33,34]. Consequently, knowing the characteristics of AHT prevention videos that are associated with a wide reach (ie, popularity) will help build more impactful public health interventions. Although measuring the direct impact of such videos on the incidence of AHT is complex and requires large-scale studies [35,36], understanding the factors that contribute to their popularity can serve as an intermediate step toward developing innovative prevention programs. The effectiveness of these programs could ultimately be evaluated by the reduction in AHT incidence. We aimed to describe the characteristics of videos of AHT prevention published on the internet, including their quality content, and to study their association with popularity.

Methods

Ethical Considerations

This study was determined to be exempt from review by the Research Project Ethics Evaluation Committee of the Robert Debré Hospital (IRB 00006477).

General Design

We performed a systematic appraisal of videos available on the internet whose main topic was the primary prevention of AHT in infants <2 years old. Because of the lack of methodological guidelines to conduct and report this type of systematic appraisal, we followed, after adaptation, the Center for Reviews and Dissemination guidance [37] and the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) statement (Multimedia Appendix 1) [38].

Search Strategy and Selection Criteria

The search was conducted on June 29, 2023. We included all videos addressing the primary prevention [39] of AHT in children <2 years old that were published in English or French (arbitrary) on the internet by public organizations (medical or scientific societies, governments or agencies, hospitals) or mainstream associations in the field of child physical abuse. We excluded commercial videos, videos not specific to children <2 years old, duplicates, and videos proposed by professionals or individuals acting outside organizations or mainstream associations.

Two investigators (a public health researcher [LG] and an educational sciences researcher [VN]) systematically searched potentially eligible videos on official websites of the national pediatric and public health societies and ministries of health of CDE [40] and on the YouTube platform, the most-used website in the world to upload and share videos [41]. On YouTube, one investigator (LG) sequentially used the terms "child physical abuse," "shaken baby syndrome," "abusive head trauma," and "crying baby" in English and French to identify videos. The

XSL•FO
RenderX

YouTube search engine was used in "incognito status," with no filters, customized search methods, or personal log-in. The results were sorted by relevance according to the default YouTube search, and the first 100 videos of each search were extracted as suggested in the literature [42-45].

We first screened video titles and descriptions to identify and exclude content that was not specific to children under 2 years of age. We then evaluated each selected video in its entirety to determine its relevance to our study's focus on AHT in infants, excluding commercial content.

Data Extraction and Collection Process

Because of the lack of guidelines to identify the variables to be extracted in the analysis of prevention videos on the internet, the authors established a list based on previous research syntheses of health videos [45-48], public health behavior theoretical models [49-52], and AHT literature [2,3,6,10,12,15,25,53]. This list was revised and enriched after the visualization of a sample of 5 eligible videos and then discussed among researchers until a consensus was reached.

The final data extraction form included 6 sections, 28 independent items, and 3 scales (Multimedia Appendices 2-4). The first section related to the general characteristics of the videos as follows: author (public organizations or associations); author's country (United States, France, or other); video duration; persistence (years since web publication); and the number of views, likes, and dislikes. The second section related to video creation features, including the type of narrative style used (interview, monologue, or conversational), the use of storytelling, the use of metaphor, the type of language used (colloquial and formal), and the use of audiovisual elements (sound, music, image, photograph, text or headers, subtitles, or motion design). The third section related to the characters depicted in the videos: the presence of an infant, the infant's age (younger than 6 months), whether the infant was crying or represented being shaken, the presence of the infant's parents and family, the professionals (legal, health, or early childhood ones) and association members involved, and the perceived origin of these characters classified as suggested by the French regulation authority for audiovisual and numeric contents [54]. The fourth section related to AHT information delivered by the videos, including the risk factors for AHT related to the infant (male sex and disability) [55] and the parents (low socioeconomic level [56], parental vulnerability [2], and poor parent-infant bonding [57]), symptoms and medical consequences of AHT (temporary or long term), characterization of shaking as an abuse, and its penal consequences. The fifth section was related to the perceived purpose of the video, which included the prevention strategies used (fear-appeal, persuasion of the severity of the depicted events, increasing the awareness of the noxious outcome of shaking, and persuasion of the effectiveness of coping strategies for infants' cries) [49] and the perceived general outcomes searched (informational, educational, giving solution, call for action, or call to prevent from action). The sixth section included the Patient Education Materials Assessment Tool (PEMAT) Audio Visual A/V score [58,59] and the Global Quality Scale (GQS; Multimedia Appendix 4) [60].

Two investigators (LG and VN) extracted the data for each video independently. In case of disagreement between the 2 investigators, consensus was sought with the help of a third investigator (a public health researcher [EL]).

Analyses

First, we calculated the popularity of the videos by using the Video Power Index (VPI; Multimedia Appendix 4), as suggested [61]. Second, the Cohen κ value was calculated to measure the interinvestigator agreement on each extracted variable [62]. Third, we described the characteristics of the analyzed videos. Fourth, we studied the associations between the VPI and the videos' characteristics and GQS and PEMAT scores using first, univariate tests and correlation analyses, and then a multivariable quasi-Poisson regression (due to the overdispersion of residuals in our regression model) including all variables statistically significant ($P<.05$) in univariate analyses. The GQS and PEMAT scores were not included in the multivariable model given the redundancy of their information (Multimedia Appendix 4) with some included variables. Hypothesis tests were 2-sided, with a priori level of significance of .05. R (version 4.3.0; R Foundation for Statistical Computing) was used for all statistical analysis [63].

Results

Results of the Search

Among the 804 videos identified by the search strategy, 53 (6.6%; Multimedia Appendix 5) met the inclusion criteria and were analyzed (Figure 1). All 53 videos were available on the YouTube platform; however, 2 (4%) had not been identified by our defined search on YouTube, but by searching the official websites of public organizations. The terms used to tag videos on the YouTube platform were "shaken baby syndrome" (42/53, 82%), "abusive head trauma" (11/53, 22%), and "crying baby" (2/53, 4%); none used the term "child physical abuse."

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) flowchart.

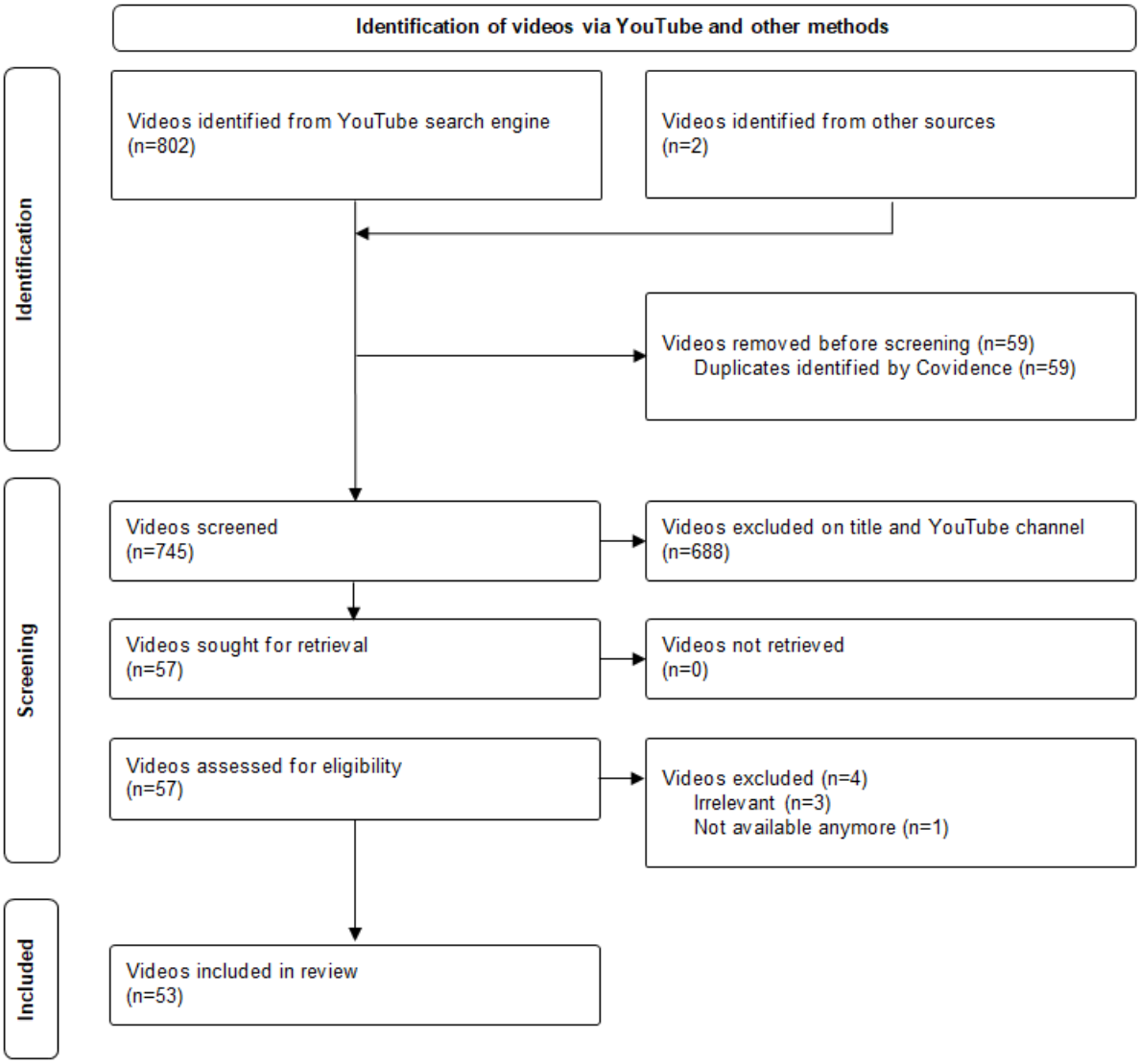

General Characteristics, Quality, and Popularity of Videos

Videos were published by public organizations (43/53, 81%) and associations (10/53, 19%) from the United States (34/53, 64%), France (11/53, 21%), or other countries (8/53, 15%; Table 1). They had a median length of 202 (IQR 94-333) seconds and had remained accessible via the web for a median of 6 years (IQR 3.3-9.3; Table 2). Videos accumulated a median of 7917 (IQR 1389-45,272) views, 21 (IQR 4-147) likes, and 2 (IQR 0-13) dislikes. The VPI was calculated for 48 (91%) of the 53 videos included due to missing data (number of likes and dislikes not available). The median VPI was 3 (IQR 1-14; mean 29, SD 113).

For the 28 video items rated, the median Cohen κ value was 0.65 (IQR 0.47-0.80; Multimedia Appendix 3). The videos included audiovisual elements such as text or headers (43/53, 81%), music (39/53, 74%), images (26/53, 49%), motion design (18/53, 34%), photographs (11/53, 21%), and subtitles (5/53, 9%; Table 1). Characters depicted were infants (42/53, 79%; various forms of presentation were used: real infants, photographs, motion design, and dolls), including while crying (35/53, 66%) and being shaken (21/53, 40%). The process of shaking was illustrated by motion design, the use of dolls, and metaphorical representations of an infant's brain. Other characters included parents and family (36/53, 68%), and professionals (40/53, 75%). In 57% (30/53) of videos, these characters were from diverse origins. The cited AHT risk factors referred mainly to the infant (44/53, 83%) and then the parents (8/53, 15%). The characterization of shaking as an abuse and its penal consequences were mentioned in 47% (25/53) and 4% (2/53) of the videos, respectively. The main prevention strategies used in the videos aimed at increasing awareness of the noxious outcomes of shaking (49/53, 93%) and the persuasion of the effectiveness of coping strategies for infants' cries (45/53, 85%).

The median understandability and actionability PEMAT scores were 83% (IQR 70%-89%) and 100% (IQR 67%-100%), respectively. The median GQS score was 4 (IQR 3-4; Table 2).

XSL•FO
RenderX

Table 1. Univariate associations between Video Power Index (VPI) and video characteristics.

Video characteristics	Videos (n=53), n (%)	VPI (n=48), median (IQR)	*P* value[a]
Video creation feature			
Author of the video			.41
Public organizations	43 (81)	3.0 (1.2-14.8)	
Association	10 (19)	3.9 (0.7-6.9)	
Author's country			.28
United States	34 (64)	6.6 (1.2-17.6)	
France	11 (21)	1.9 (1.6-3.2)	
Other	8 (15)	1.8 (0.9-5.7)	
Narrative style			
Interview			.96
Yes	20 (38)	4.8 (1.4-14.3)	
No	33 (62)	2.9 (1.1-13.6)	
Monologue			.02
Yes	45 (85)	3.9 (1.3-14.8)	
No	8 (15)	0.5 (0.2-1.6)	
Conversational			.41
Yes	14 (26)	7.3 (2.7-14.4)	
No	39 (74)	2.3 (1.1-12.4)	
Use of storytelling			.31
Yes	12 (23)	8.0 (2.2-14.5)	
No	41 (77)	2.7 (1.0-11.6)	
Use of metaphor			.12
Yes	4 (8)	35.1 (3.1-78.6)	
No	49 (92)	2.9 (1.1-13.1)	
Type of language used			
Colloquial			.22
Yes	41 (77)	3.9 (1.2-14.8)	
No	12 (19)	1.5 (0.7-7.0)	
Formal			.17
Yes	46 (87)	4.6 (1.2-14.8)	
No	7 (13)	1.9 (0.9-2.9)	
Audiovisual element			
Sound			.22
Yes	34 (64)	2.7 (0.9-10.5)	
No	19 (36)	4.6 (1.8-16.5)	
Music			.12
Yes	39 (74)	5.6 (1.3-14.4)	
No	14 (26)	1.5 (0.6-6.1)	
Image			.01
Yes	26 (49)	9.8 (2.8-23.7)	
No	27 (51)	1.7 (0.7-6.6)	

XSL•FO
RenderX

Video characteristics	Videos (n=53), n (%)	VPI (n=48), median (IQR)	*P* value[a]
Photograph			.02
Yes	11 (21)	14.5 (8.1-23.7)	
No	42 (79)	2.3 (0.8-8.0)	
Text or headers			.02
Yes	43 (81)	6.5 (1.5-15.6)	
No	10 (19)	1.1 (0.7-1.8)	
Subtitles			.94
Yes	5 (9)	3.6 (2.4-5.5)	
No	48 (91)	3.0 (1.1-14.8)	
Motion design			.003
Yes	18 (34)	14.8 (3.9-25.0)	
No	35 (66)	1.8 (0.8-6.6)	
Characters depicted in the videos			
Infant			.07
Yes	42 (79)	5.6 (1.3-14.8)	
No	11 (21)	1.5 (0.7-2.9)	
Infant's age younger than 6 months			.12
Yes	37 (70)	6.5 (1.3-14.8)	
No	16 (30)	1.7 (1.0-3.9)	
Representation of an infant being shaken			<.001
Yes	21 (40)	13.5 (5.4-25.2)	
No	32 (60)	1.8 (0.7-5.7)	
Representation of a crying infant			.28
Yes	35 (66)	6.5 (1.1-16.4)	
No	18 (34)	1.8 (1.3-5.1)	
Infant's parents and family			.47
Yes	36 (68)	6.5 (1.2-14.8)	
No	17 (32)	2.7 (0.9-5.6)	
Professional			.57
Yes	40 (75)	3.0 (1.2-14.8)	
No	13 (25)	3.1 (0.6-8.2)	
Perceived origin of the characters			.08
Only white	23 (43)	1.9 (0.8-6.6)	
Mixed origin	30 (57)	6.5 (1.3-22.9)	
AHT[b] information delivered in the videos			
Infant risk factors			.19
Yes	44 (83)	1.9 (1.0-13.5)	
No	9 (17)	6.6 (2.7-16.5)	
Parent risk factors			.30
Yes	8 (15)	8.0 (3.9-17.7)	
No	45 (85)	2.7 (1.0-12.7)	

XSL•FO
RenderX

Video characteristics	Videos (n=53), n (%)	VPI (n=48), median (IQR)	*P* value[a]
Symptoms of AHT			.049
Yes	18 (34)	10.3 (2.6-22.3)	
No	35 (66)	1.9 (0.9-8.1)	
Temporary medical consequences of AHT			.11
Yes	25 (47)	7.3 (1.5-16.0)	
No	28 (53)	2.3 (0.8-7.7)	
Long-term medical consequences of AHT			.17
Yes	42 (80)	3.9 (1.2-16.0)	
No	11 (20)	1.8 (0.8-5.8)	
Characterization of shaking as abuse			.09
Yes	25 (47)	8.0 (2.7-14.8)	
No	28 (53)	1.8 (0.9-7.5)	
Penal consequences spelled out			.86
Yes	2 (4)	4.6 (4.6-4.6)	
No	51 (96)	3.0 (1.1-14.5)	
Video perceived purposes			
Prevention strategy used			
Fear-appeal, persuasion of severity of the depicted events			.14
Yes	26 (49)	6.6 (1.2-18.8)	
No	27 (51)	1.9 (1.1-8.0)	
Increasing the awareness of the noxious outcome of shaking			.06
Yes	49 (93)	3.9 (1.2-14.8)	
No	4 (7)	0.9 (0.0-2.1)	
Persuasion of the effectiveness of coping strategies for infants' cries			.68
Yes	45 (85)	3.0 (0.9-14.8)	
No	8 (15)	3.7 (2.3-7.9)	
Perceived general outcome searched			
Educational			.91
Yes	49 (93)	3.1 (1.1-14.8)	
No	4 (7)	2.7 (2.4-4.1)	
Informational			.37
Yes	50 (94)	3.1 (1.2-14.2)	
No	3 (6)	0.2 (0.1-56.6)	
Giving solution			.56
Yes	43 (81)	3.1 (1.1-14.8)	
No	10 (19)	2.7 (1.2-6.1)	
Call for action			.12
Yes	43 (81)	4.9 (1.2-16.0)	
No	10 (19)	2.0 (0.9-4.1)	
Call to prevent from action			.38
Yes	48 (91)	3.0 (1.2-14.8)	

XSL•FO
RenderX

Video characteristics	Videos (n=53), n (%)	VPI (n=48), median (IQR)	*P* value[a]
No	5 (9)	3.1 (0.7-6.6)	

[a]Mann-Whitney and Kruskal-Wallis tests.
[b]AHT: abusive head trauma.

Table 2. General characteristics of the included videos and association with the Video Power Index (VPI).

Video features	Value (n=53), n (%)	Value (n=48), median (IQR)	*r* (n=48)	*P* value[a] (n=48)
Duration (seconds)	53 (100)	202 (94-333)	0.36	.01
Number of years accessible via the web	53 (100)	6 (3-9)	N/A[b]	N/A
Number of views	53 (100)	7917 (1389-45,272)	N/A	N/A
Number of likes	48 (91)	21 (4-147)	N/A	N/A
Number of dislikes	48 (91)	2 (0-13)	N/A	N/A
VPI	48 (91)	3 (1-14)	N/A	N/A
GQS[c] score	53 (100)	4 (3-4)	0.38	.007
PEMAT[d] understandability score (%)	53 (100)	83 (70-89)	0.27	.06
PEMAT actionability score (%)	53 (100)	100 (67-100)	0.27	.07

[a]Spearman correlation test.
[b]Not applicable.
[c]GQS: Global Quality Scale.
[d]PEMAT: Patient Education Materials Assessment Tool.

Determinants of Popularity

In the univariate analyses (Table 1), only monologue used as narrative style was significantly associated with increased VPI (median 3.9 vs 0.5 for no monologue; *P*=.02). The VPI was significantly increased when videos contained images (median 9.8 vs 1.7; *P*=.01), photographs (median 14.5 vs 2.3; *P*=.02), text or headers (median 6.5 vs 1.1; *P*=.02), and motion design (median 14.8 vs 1.8; *P*=.003) as audiovisual elements. The presence of an infant in the videos was not associated with the VPI. The VPI was significantly increased with the depiction of an infant being shaken (median 13.5 vs 1.8; *P*<.001). Character types represented in the video (infant's parents and family, professionals) and their perceived origin were not significantly associated with the VPI (all *P*≥.05). The VPI was significantly increased when symptoms of AHT were mentioned (median 10.3 vs 1.9; *P*=.049) but not when shaking was characterized as an abuse (*P*=.09) or when its penal consequences were stated (*P*=.86). None of the prevention strategies used in the videos and the perceived video's general outcomes were associated with increased VPI. The VPI was significantly positively correlated with the duration of the videos (*r*=0.36; *P*=.01) and the GQS score (*r*=0.38; *P*=.007) but not PEMAT scores (all *P*≥.05).

In the multivariable analysis (Table 3), increased VPI was independently significantly associated only with the use of text or headers in the videos (*P*=.04; β=2.15, 95% CI 1.08-4.26) and the representation of an infant being shaken (*P*=.03; β=1.74, 95% CI 1.06-2.85).

Table 3. Multivariable analysis of the association of variables with Video Power Index (n=48).

Video features	β	95% CI	*P* value
Duration (seconds)	1.00	1.00-1.00	.60
Narrative style: monologue	2.44	0.98-6.10	.06
Audiovisual element			
Image	0.85	0.50-1.45	.55
Photograph	1.26	0.79-2.02	.34
Text or headers	2.15	1.08-4.26	.04
Motion design	1.00	0.58-1.75	.99
Representation of an infant being shaken	1.74	1.06-2.85	.03
Symptoms of abusive head trauma	1.35	0.85-2.13	.21

XSL•FO
RenderX

Discussion

Principal Findings

This systematic appraisal highlights the relationship between the characteristics and content of health web-based prevention videos and their popularity. It leads to practical implications for developing more effective prevention campaigns. In our study, VPIs were significantly increased for videos that included images, photographs, text or headers, and motion design. Although numerous studies examined the popularity of preventive health videos or health topics, none analyzed the relation between visual elements within the videos and their popularity [46,47,64-66]. In another educational field, instructional videos, one study had the same conclusion as ours on the determinants of popularity [67]. In our study, the representation of an infant being shaken was associated with increased VPI, a result providing additional evidence to support considering the use of shocking content on this topic. Shocking content is often used in prevention campaigns because it is supposed to better capture the audience's attention [68-70], but there is no consensus on its effectiveness [68-72]. Finally, in our study, as in other studies in various fields [73-75], the quality of the videos was positively correlated with their popularity.

In our study, the overall content of 53 videos of AHT prevention had good quality as indicated by a median GQS score of 4. This result can be explained in part by the videos being mainly published by public organizations, whose videos generally have higher GQS score than those published by nonprofessionals [48,64,76]. The videos we analyzed mostly contained information about the symptoms and medical consequences of AHT as well as the dangerous consequences of shaking an infant. This observation is consistent with the main content of AHT prevention interventions that do not use the internet [16,21]. The videos were mostly educational and used a strategy of persuasion about the effectiveness of coping, which is consistent with the strategies used in most AHT prevention interventions that attempt to convince parents of the efficacy of coping strategies in managing the response to infant crying (eg, walk away, leave the infant alone, or call a support person) [16,21]. However, some information was not addressed, particularly regarding parental risk factors, a finding similar to the content of other AHT prevention interventions [16,21]. By focusing solely on child risk factors and neglecting parental and more generally perpetrator risk factors, the videos analyzed could induce a deleterious shift of the responsibility for shaking from the adults to the infant, a mechanism similar to victim blaming [77-79]. Including perpetrators' risk factors in prevention programs would help raise awareness of the personal responsibility of perpetrators and limit their ability to claim to be victims themselves, passive and will-less in their wrongdoing [78]. The characterization of shaking as an abuse and its penal consequences were often missing in the analyzed videos and were not found to be associated with popularity. The use of these arguments in prevention campaigns is debated. Emphasizing legal consequences might decrease audience participation in prevention campaigns, particularly because of stigma [80,81]. However, some studies suggest that education about the legality, consequences, and responsibilities associated with abusive behavior might enhance the target audience's understanding and commitment to preventing such behavior [82-84]. The optimal balance between legal deterrence and educational strategies should be adapted according to the objectives of the prevention campaigns.

The analyzed videos had a mean VPI of 29. There is no validated threshold to interpret a VPI value. The mean VPI in our study seems to be rather low compared with VPIs reported in the medical literature, which range from <1 (eg, 0.17 in health prevention videos) [46] to >1000 (eg, 1078 in chronic medical conditions videos) [47,64-66], and are lower when the publisher is a public organization [46,64,66]. The video platform that we used, YouTube, may itself use a recommendation algorithm that will not prioritize or suggest preventive AHT content because of potential violent content linked to the representation of a shaken infant. Furthermore, only 4% (2/53) of the analyzed videos resulted from the search term "crying baby," whereas crying is reported as the primary trigger in most perpetrations of AHT [11] and is probably the most frequent term used by parents rather than "shaken baby syndrome," the most frequent tag used in the analyzed videos. In the future, publishers should systematically tag their videos with the term "crying baby".

Limitations

The first limitation is that we included only videos published by public organizations in CDE and mainstream associations, on YouTube and on public organization websites, in English and French. As well, the search was restricted to the first 100 videos for each search term. The characteristics and popularity of videos posted on other popular platforms, such as Facebook, Instagram, and TikTok, and by other publishers may differ. Our findings, which are limited to videos using English and French languages, may not be generalized to all contexts. The direction and magnitude of the probable selection biases introduced by these criteria are unknown. Second, because of these strict selection criteria, the sample size resulted in limited statistical power to identify the determinants of VPI. Third, the median Cohen κ indicated only substantial agreement between investigators, thus highlighting the subjectivity of quoted criteria. Finally, although the GQS and VPI are increasingly used tools in the literature, they are not validated and there is no threshold for their interpretation.

Implications

As a first step to developing an effective AHT prevention program, this systematic appraisal of available videos suggests that developing high-quality content for AHT prevention videos can have a significant positive impact on their popularity. Future videos may also include specific features such as text or headers and the depiction of an infant being shaken to improve their popularity. However, popularity alone does not ensure the effectiveness of these videos in preventing AHT. Therefore, understanding popularity must be complemented by evaluating the effectiveness of the content and the needs and expectations of the target audience. Increasing popularity could lead to more impactful prevention programs.

Another implication is the need for the development and validation of methodological tools to scientifically conduct and

XSL•FO
RenderX

report health video synthesis research. Videos will probably become the main channel for public health programs in the coming years, and the standardization of methods is critical. Currently, there is an urgent need to validate or refine the PEMAT, GQS, and VPI as tools to evaluate the quality and popularity of videos. Furthermore, consensus is needed on core items that must be evaluated when analyzing videos. Methods to maximize the objectivity of the assessments should also be produced and made available to the research community.

By establishing an association between video quality and popularity, this research provides insights for the development of more effective digital prevention strategies that could be of use beyond AHT. The methodology developed to study the videos, in the absence of a recommended approach, could serve as a basis for further methodological developments that would improve research practices in this field.

Conclusion

Videos addressing the primary prevention of AHT in children <2 years old that were published on the internet in English or French by public organizations or mainstream associations were of high quality but did not frequently deal with parental risk factors. The inclusion of characteristics identified as associated with the popularity of AHT prevention videos could help improve the impact of future prevention programs by enhancing their popularity. Finally, there is an urgent need for the development and validation of methodological tools to standardize the conduct and report of studies on prevention videos.

Acknowledgments

The authors thank Pauline Scherdel, PhD, and Marianne Jacques, PhD student, of the Inserm for their technical advice on the statistical analysis of this study. They did not receive compensation for their contribution. This study was supported by a grant from the health sponsorship program of Mutuelles AXA. The funder had no role in the study design; the collection, analysis, and interpretation of data; the writing of the report; and the decision to submit the paper for publication.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request (Multimedia Appendix 6).

Authors' Contributions

LG contributed to conceptualization, methodology, validation, formal analysis, investigation, data curation, writing the original draft, reviewing and editing the manuscript, visualization, and project administration. VPN was involved in conceptualization, methodology, formal analysis, investigation, writing the original draft, and reviewing and editing the manuscript. FF contributed to conceptualization, methodology, formal analysis, writing the original draft, and reviewing and editing the manuscript. KC and MB reviewed and edited the manuscript. CL and YR contributed to conceptualization, methodology, writing the original draft, reviewing and editing the manuscript, and project administration. FB was responsible for conceptualization, methodology, software, formal analysis, writing the original draft, and reviewing and editing the manuscript. MC was involved in conceptualization, methodology, validation, formal analysis, resources, writing the original draft, reviewing and editing the manuscript, visualization, supervision, project administration, and securing funding. ELR contributed to conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing the original draft, reviewing and editing the manuscript, visualization, supervision, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) checklist.
[DOCX File , 21 KB - jmir_v26i1e60530_app1.docx]

Multimedia Appendix 2
Metrics of the videos analyzed.
[DOCX File , 33 KB - jmir_v26i1e60530_app2.docx]

Multimedia Appendix 3
Video characteristics and their κ coefficient for investigator agreement.
[DOCX File , 20 KB - jmir_v26i1e60530_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Description of the Global Quality Scale (GQS), Patient Education Materials Assessment Tool Audio Visual (PEMAT) A/V scale, and Video Power Index (VPI) .
[DOCX File , 18 KB - jmir_v26i1e60530_app4.docx]

Multimedia Appendix 5
Videos included in the appraisal.
[DOCX File , 23 KB - jmir_v26i1e60530_app5.docx]

Multimedia Appendix 6
Data sharing statement.
[DOCX File , 13 KB - jmir_v26i1e60530_app6.docx]

References

1. Parks SE, Annest JL, Hill HA, Karch DL. Pediatric abusive head trauma : recommended definitions for public health surveillance and research. Centers for Disease Control and Prevention. URL: https://stacks.cdc.gov/view/cdc/26243 [accessed 2023-05-09]
2. Rebbe R, Mienko JA, Martinson ML. Incidence and risk factors for abusive head trauma: a population-based study. Child Abuse Rev 2020;29(3):195-207 [FREE Full text] [doi: 10.1002/car.2630] [Medline: 33071539]
3. Lee C, Barr RG, Catherine N, Wicks A. Age-related incidence of publicly reported shaken baby syndrome cases: is crying a trigger for shaking? J Dev Behav Pediatr 2007;28(4):288-293. [doi: 10.1097/DBP.0b013e3180327b55] [Medline: 17700080]
4. Paget LM, Gilard‐Pioc S, Cottenet J, Beltzer N, Quantin C. Identification of infant victims of abusive head trauma hospitalised in France from 2015 to 2017. Child Abuse Review 2020;29(3):182-194. [doi: 10.1002/car.2639]
5. Fujiwara T, Barr RG, Brant RF, Rajabali F, Pike I. Using international classification of diseases, 10th edition, codes to estimate abusive head trauma in children. Am J Prev Med 2012;43(2):215-220. [doi: 10.1016/j.amepre.2012.04.028] [Medline: 22813688]
6. Keenan HT, Runyan DK, Marshall SW, Nocera MA, Merten DF, Sinal SH. A population-based study of inflicted traumatic brain injury in young children. JAMA 2003;290(5):621-626. [doi: 10.1001/jama.290.5.621] [Medline: 12902365]
7. Emrick BB, Smith E, Thompson L, Mullett C, Pino E, Snyder K, et al. Epidemiology of abusive head trauma in West Virginia children <24 months: 2000-2010. Child Abuse Negl 2019;93:215-221 [FREE Full text] [doi: 10.1016/j.chiabu.2019.05.011] [Medline: 31125851]
8. Parrish J, Baldwin-Johnson C, Volz M, Goldsmith Y. Abusive head trauma among children in Alaska: a population-based assessment. Int J Circumpolar Health 2013;72:eCollection 2013 [FREE Full text] [doi: 10.3402/ijch.v72i0.21216] [Medline: 23986886]
9. Fanconi M, Lips U. Shaken baby syndrome in Switzerland: results of a prospective follow-up study, 2002-2007. Eur J Pediatr 2010 Aug;169(8):1023-1028. [doi: 10.1007/s00431-010-1175-x] [Medline: 20213304]
10. Edwards GA, Maguire SA, Gaither JR, Leventhal JM. What do confessions reveal about abusive head trauma? a systematic review. Child Abuse Rev 2020;29(3):253-268 [FREE Full text] [doi: 10.1002/car.2627] [Medline: 37982093]
11. Barr RG, Trent RB, Cross J. Age-related incidence curve of hospitalized shaken baby syndrome cases: convergent evidence for crying as a trigger to shaking. Child Abuse Negl 2006;30(1):7-16. [doi: 10.1016/j.chiabu.2005.06.009] [Medline: 16406023]
12. Niederkrotenthaler T, Xu L, Parks SE, Sugerman DE. Descriptive factors of abusive head trauma in young children--United States, 2000-2009. Child Abuse Negl 2013;37(7):446-455. [doi: 10.1016/j.chiabu.2013.02.002] [Medline: 23535075]
13. Hobbs C, Childs AM, Wynne J, Livingston J, Seal A. Subdural haematoma and effusion in infancy: an epidemiological study. Arch Dis Child 2005;90(9):952-955 [FREE Full text] [doi: 10.1136/adc.2003.037739] [Medline: 16113132]
14. Ferguson NM, Sarnaik A, Miles D, Shafi N, Peters MJ, Truemper E, et al. Abusive head trauma and mortality-an analysis from an international comparative effectiveness study of children with severe traumatic brain injury. Crit Care Med 2017;45(8):1398-1407 [FREE Full text] [doi: 10.1097/CCM.0000000000002378] [Medline: 28430697]
15. Chevignard MP, Lind K. Long-term outcome of abusive head trauma. Pediatr Radiol 2014;44 Suppl 4:S548-S558. [doi: 10.1007/s00247-014-3169-8] [Medline: 25501726]
16. Scott LJ, Wilson R, Davies P, Lyttle MD, Mytton J, Dawson S, et al. Educational interventions to prevent paediatric abusive head trauma in babies younger than one year old: a systematic review and meta-analyses. Child Abuse Negl 2022;134:105935 [FREE Full text] [doi: 10.1016/j.chiabu.2022.105935] [Medline: 36308894]
17. Poole MK, Seal DW, Taylor CA. A systematic review of universal campaigns targeting child physical abuse prevention. Health Educ Res 2014;29(3):388-432 [FREE Full text] [doi: 10.1093/her/cyu012] [Medline: 24711483]
18. Altman RL, Canter J, Patrick PA, Daley N, Butt NK, Brand DA. Parent education by maternity nurses and prevention of abusive head trauma. Pediatrics 2011;128(5):e1164-e1172. [doi: 10.1542/peds.2010-3260] [Medline: 22025587]
19. Barr RG, Barr M, Rajabali F, Humphreys C, Pike I, Brant R, et al. Eight-year outcome of implementation of abusive head trauma prevention. Child Abuse Negl 2018;84:106-114. [doi: 10.1016/j.chiabu.2018.07.004] [Medline: 30077049]

XSL•FO
RenderX

20. Dias MS, Smith K, DeGuehery K, Mazur P, Li V, Shaffer ML. Preventing abusive head trauma among infants and young children: a hospital-based, parent education program. Pediatrics 2005;115(4):e470-e477. [doi: 10.1542/peds.2004-1896] [Medline: 15805350]
21. Chang HY, Chang YC, Chang YT, Chen YW, Wu PY, Feng JY. The effectiveness of parenting programs in preventing abusive head trauma: a systematic review and meta-analysis. Trauma Violence Abuse 2024;25(1):354-368. [doi: 10.1177/15248380231151690] [Medline: 36762510]
22. Viswanathan M, Rains C, Hart LC, Doran E, Sathe N, Hudson K, et al. Primary care interventions to prevent child maltreatment: evidence report and systematic review for the US preventive services task force. JAMA 2024;331(11):959-971. [doi: 10.1001/jama.2024.0276] [Medline: 38502070]
23. Barry MJ, Nicholson WK, Silverstein M, Chelmow D, Coker TR, Davis EM, et al. Primary care interventions to prevent child maltreatment: US preventive services task force recommendation statement. JAMA 2024;331(11):951-958. [doi: 10.1001/jama.2024.1869] [Medline: 38502069]
24. Schilling S, Wood JN, Christian CW. Struggling to stem the tide of child maltreatment. JAMA 2024;331(11):918-919 [FREE Full text] [doi: 10.1001/jama.2024.0899] [Medline: 38502084]
25. Lăzărescu AM, Benichi S, Blauwblomme T, Beccaria K, Bourgeois M, Roux CJ, et al. Abusive head trauma in infants during the COVID-19 pandemic in the Paris Metropolitan area. JAMA Netw Open 2022;5(8):e2226182 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.26182] [Medline: 36040743]
26. Cercone DJ, Berger RP, Manole MD, Soung JK, Coombs CM, Noorbakhsh KA. Increased severity of abusive head trauma during the first year of the COVID-19 pandemic. Child Abuse Negl 2023;135:105971 [FREE Full text] [doi: 10.1016/j.chiabu.2022.105971] [Medline: 36427395]
27. Total births by sex. United Nations, Department of Economic and Social Affairs, Population Division. URL: https://tinyurl.com/4kwaw3yb [accessed 2024-03-18]
28. Wiley M, Schultheis A, Francis B, Tiyyagura G, Leventhal JM, Rutherford HJV, et al. Parents' perceptions of infant crying: a possible path to preventing abusive head trauma. Acad Pediatr 2020;20(4):448-454. [doi: 10.1016/j.acap.2019.10.009] [Medline: 31629119]
29. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021;23(11):e26931 [FREE Full text] [doi: 10.2196/26931] [Medline: 34787575]
30. Fujiwara T, Isumi A, Sampei M, Yamada F, Miyazaki Y. Effectiveness of using an educational video simulating the anatomical mechanism of shaking and smothering in a home-visit program to prevent self-reported infant abuse: a population-based quasi-experimental study in Japan. Child Abuse Negl 2020;101:104359. [doi: 10.1016/j.chiabu.2020.104359] [Medline: 31945514]
31. Coster D. Changing parents' behaviour using a psycho‐educational film as a preventative measure to reduce the risk of non‐accidental head injury. Child Abuse Rev 2017;26(6):465-476. [doi: 10.1002/car.2469]
32. Global time spent with online video 2023. Statista. URL: https://www.statista.com/statistics/611707/online-video-time-spent/ [accessed 2023-09-28]
33. EU citizens: over half seek health information online. Eurostat. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220406-1 [accessed 2023-09-28]
34. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625. [doi: 10.1177/0033354919874074] [Medline: 31513756]
35. Dias MS, Rottmund CM, Cappos KM, Reed ME, Wang M, Stetter C, et al. Association of a postnatal parent education program for abusive head trauma with subsequent pediatric abusive head trauma hospitalization rates. JAMA Pediatr 2017;171(3):223-229. [doi: 10.1001/jamapediatrics.2016.4218] [Medline: 28135348]
36. Barr RG, Barr M, Fujiwara T, Conway J, Catherine N, Brant R. Do educational materials change knowledge and behaviour about crying and shaken baby syndrome? a randomized controlled trial. CMAJ 2009;180(7):727-733 [FREE Full text] [doi: 10.1503/cmaj.081419] [Medline: 19255065]
37. Centre for Reviews and Dissemination. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. Layerthorpe, England: University of York; 2009.
38. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Health promotion and disease prevention through population-based interventions, including action to address social determinants and health inequity. World Health Organization; Eastern Mediterranean Region. URL: http://www.emro.who.int/about-who/public-health-functions/health-promotion-disease-prevention.html [accessed 2023-07-27]
40. Country classifications. United Nations. 2022. URL: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/WESP2022_ANNEX.pdf [accessed 2023-05-09]
41. Top websites ranking-most visited websites. Similar Web. URL: https://www.similarweb.com/top-websites/ [accessed 2024-01-02]

XSL•FO
RenderX

42. Morra S, Collà Ruvolo C, Napolitano L, La Rocca R, Celentano G, Califano G, et al. YouTube as a source of information on bladder pain syndrome: a contemporary analysis. Neurourol Urodyn 2022;41(1):237-245. [doi: 10.1002/nau.24802] [Medline: 34559920]
43. Cilio S, Collà Ruvolo C, Turco C, Creta M, Capece M, La Rocca R, et al. Analysis of quality information provided by "Dr. YouTube™" on Phimosis. Int J Impot Res 2022;35(4):398-403. [doi: 10.1038/s41443-022-00557-5] [Medline: 35332276]
44. Capece M, Di Giovanni A, Cirigliano L, Napolitano L, La Rocca R, Creta M, et al. YouTube as a source of information on penile prosthesis. Andrologia 2022;54(1):e14246. [doi: 10.1111/and.14246] [Medline: 34519075]
45. Chen Z, Pan S, Zuo S. TikTok and YouTube as sources of information on anal fissure: a comparative analysis. Front Public Health 2022;10:1000338. [doi: 10.3389/fpubh.2022.1000338] [Medline: 36407987]
46. Kiss A, Soós S, Temesi Á, Unger-Plasek B, Lakner Z, Tompa O. Evaluation of the reliability and educational quality of YouTube™ videos on sport nutrition topics. J Int Soc Sports Nutr 2023;20(1):2278632. [doi: 10.1080/15502783.2023.2278632] [Medline: 37953602]
47. Kwak D, Park JW, Won Y, Kwon Y, Lee JI. Quality and reliability evaluation of online videos on carpal tunnel syndrome: a YouTube video-based study. BMJ Open 2022;12(4):e059239. [doi: 10.1136/bmjopen-2021-059239] [Medline: 35428646]
48. Li HOY, Pastukhova E, Brandts-Longtin O, Tan MG, Kirchhof MG. YouTube as a source of misinformation on COVID-19 vaccination: a systematic analysis. BMJ Glob Health 2022 Mar;7(3):e008334. [doi: 10.1136/bmjgh-2021-008334] [Medline: 35264318]
49. Rogers RW, Cacioppo JR, Petty RE. Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Social Psychology: A Source Book. New York City, NY: Guilford Press; 1983:153-176.
50. Maddux JE, Rogers RW. A protection motivation theory of fear appeals and attitude change. J Psychol 1975;91(1):93-114. [doi: 10.1080/00223980.1975.9915803] [Medline: 28136248]
51. Ajzen I, Fishbein M. Understanding Attitudes and Predicting Social Behavior. Englewood Cliffs, NJ: Prentice-Hall; 1980.
52. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. Educ Psychol 2010;38(1):43-52. [doi: 10.1207/S15326985EP3801_6]
53. Notrica DM, Kirsch L, Misra S, Kelly C, Greenberg J, Ortiz JB, et al. Evaluating abusive head trauma in children <5 years old: risk factors and the importance of the social history. J Pediatr Surg 2021;56(2):390-396. [doi: 10.1016/j.jpedsurg.2020.10.019] [Medline: 33220974]
54. Représentation de la diversité. Autorité de Régulation de la Communication Audiovisuelle et Numérique. URL: https://www.arcom.fr/nous-connaitre-nos-missions/garantir-le-pluralisme-et-la-cohesion-sociale/representation-de-la-diversite [accessed 2024-01-22]
55. Sullivan PM, Knutson JF. Maltreatment and disabilities: a population-based epidemiological study. Child Abuse Negl 2000;24(10):1257-1273. [doi: 10.1016/s0145-2134(00)00190-3] [Medline: 11075694]
56. Judd N, Hughes K, Bellis MA, Hardcastle K, Amos R. Is parental unemployment associated with increased risk of adverse childhood experiences? a systematic review and meta-analysis. J Public Health (Oxf) 2023;45(4):829-839. [doi: 10.1093/pubmed/fdad069] [Medline: 37253685]
57. Loiseau M, Cottenet J, François-Purssell I, Bechraoui-Quantin S, Jud A, Gilard-Pioc S, et al. Hospitalization for physical child abuse: associated medical factors and medical history since birth. Child Abuse Negl 2023;146:106482. [doi: 10.1016/j.chiabu.2023.106482] [Medline: 37776729]
58. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
59. Shoemaker S, Wolf M, Brach C. The Patient Education Materials Assessment Tool (PEMAT) and user's guide. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/health-literacy/patient-education/pemat.html [accessed 2023-04-05]
60. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the world wide web. Am J Gastroenterol 2007;102(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
61. Erdem MN, Karaca S. Evaluating the accuracy and quality of the information in kyphosis videos shared on YouTube. Spine (Phila Pa 1976) 2018;43(22):E1334-E1339. [doi: 10.1097/BRS.0000000000002691] [Medline: 29664816]
62. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159. [doi: 10.2307/2529310]
63. R Core Team (2023). R Found Stat Comput Vienna Austria. R: The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2024-04-24]
64. Kunze KN, Krivicich LM, Verma NN, Chahla J. Quality of online video resources concerning patient education for the meniscus: a YouTube-based quality-control study. Arthroscopy 2020;36(1):233-238. [doi: 10.1016/j.arthro.2019.07.033] [Medline: 31864582]
65. Etzel CM, Bokshan SL, Forster TA, Owens BD. A quality assessment of YouTube content on shoulder instability. Phys Sportsmed 2022;50(4):289-294. [doi: 10.1080/00913847.2021.1942286] [Medline: 34121601]

XSL•FO
RenderX

66. Zhang X, Yang Y, Shen YW, Zhang KR, Ma LT, Ding C, et al. Quality of online video resources concerning patient education for neck pain: a YouTube-based quality-control study. Front Public Health 2022;10:972348. [doi: 10.3389/fpubh.2022.972348] [Medline: 36211682]
67. Hove PT, Meij H. Like it or not. what characterizes YouTube's more popular instructional videos? Semantic Scholar. 2015. URL: https://www.semanticscholar.org/paper/Like-It-or-Not.-What-Characterizes-YouTube's-More-Hove-Meij/db1038640b6f95675c66e2be533785e8f006d721 [accessed 2024-02-02]
68. Witte K, Allen M. A meta-analysis of fear appeals: implications for effective public health campaigns. Health Educ Behav 2000;27(5):591-615. [doi: 10.1177/109019810002700506] [Medline: 11009129]
69. Ruiter RAC, Kessels LTE, Peters JGY, Kok G. Sixty years of fear appeal research: current state of the evidence. Int J Psychol 2014;49(2):63-70. [doi: 10.1002/ijop.12042] [Medline: 24811876]
70. Tannenbaum MB, Hepler J, Zimmerman RS, Saul L, Jacobs S, Wilson K, et al. Appealing to fear: a meta-analysis of fear appeal effectiveness and theories. Psychol Bull 2015;141(6):1178-1204. [doi: 10.1037/a0039729] [Medline: 26501228]
71. Dahl DW, Frankenberger KD, Manchanda RV. Does it pay to shock? reactions to shocking and nonshocking advertising content among university students. J Adv Res 2003;43(03):268-280. [doi: 10.1017/s0021849903030332]
72. Lull RB, Bushman BJ. Do sex and violence sell? a meta-analytic review of the effects of sexual and violent media and ad content on memory, attitudes, and buying intentions. Psychol Bull 2015;141(5):1022-1048. [doi: 10.1037/bul0000018] [Medline: 26191956]
73. Güvener O, Da' F, Sarikaya P. Popularity, quality and reliability of YouTube videos as an information resource on kinesio taping. Dicle T'p Derg Dicle Üniversitesi 2023;50(1):22-29. [doi: 10.5798/dicletip.1266703]
74. Aydin CK, Gunec HG. Quality of information on YouTube about artificial intelligence in dental radiology. J Dent Educ 2020;84(10):1166-1172. [doi: 10.1002/jdd.12362] [Medline: 32813894]
75. Nishizaki N, Hirano D, Oishi K, Shimizu T. YouTube videos in Japanese as a source of information on nocturnal enuresis: a content-quality and reliability analysis. Pediatr Int 2022;64(1):e15049. [doi: 10.1111/ped.15049] [Medline: 34747553]
76. Li HOY, Bailey A, Huynh D, Chan J. YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Glob Health 2020;5(5):e002604. [doi: 10.1136/bmjgh-2020-002604] [Medline: 32409327]
77. Suarez E, Gadalla TM. Stop blaming the victim: a meta-analysis on rape myths. J Interpers Violence 2010;25(11):2010-2035. [doi: 10.1177/0886260509354503] [Medline: 20065313]
78. Lamb S. The Trouble with Blame: Victims, Perpetrators, and Responsibility. Cambridge, MA: Harvard University Press; 1996.
79. Strömwall LA, Landström S, Alfredsson H. Perpetrator characteristics and blame attributions in a stranger rape situation. Eur J Psychol Appl Leg Context 2014;6(2):63-67. [doi: 10.1016/j.ejpal.2014.06.002]
80. Kemshall H, Moulden HM. Communicating about child sexual abuse with the public: learning the lessons from public awareness campaigns. J Sex Aggress 2016;23(2):124-138. [doi: 10.1080/13552600.2016.1222004]
81. Beier KM, Neutze J, Mundt IA, Ahlers CJ, Goecker D, Konrad A, et al. Encouraging self-identified pedophiles and hebephiles to seek professional help: first results of the Prevention Project Dunkelfeld (PPD). Child Abuse Negl 2009;33(8):545-549. [doi: 10.1016/j.chiabu.2009.04.002] [Medline: 19765825]
82. Schober DJ, Fawcett SB, Bernier J. The enough abuse campaign: building the movement to prevent child sexual abuse in Massachusetts. J Child Sex Abus 2012;21(4):456-469. [doi: 10.1080/10538712.2012.675423] [Medline: 22809049]
83. Bailey A, Allen L, Stevens E, Dervley R, Findlater D, Wefers S. Pathways and prevention for indecent images of children offending: a qualitative study. Sex Offending 2022;17:1-25. [doi: 10.5964/sotrap.6657]
84. Newman EF, Efthymiadou E, Quayle E, Squire T, Denis D, Wortley R, et al. The impact of a public health campaign to deter viewing of child sexual abuse images online: a case study of the UK stop it now! campaign. Sex Abuse 2024;36(6):635-661. [doi: 10.1177/10790632231205784] [Medline: 37837444]

Abbreviations

AHT: abusive head trauma
CDE: countries with developed economies
GQS: Global Quality Scale
PEMAT: Patient Education Materials Assessment Tool
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis
VPI: Video Power Index

XSL•FO
RenderX

Edited by F Xie; submitted 14.05.24; peer-reviewed by VJ Palusci, O Berthold, X Jiang; comments to author 17.07.24; revised version received 02.09.24; accepted 25.09.24; published 10.12.24.

Please cite as:
Goethals L, Prokofieva Nelson V, Fenouillet F, Chevreul K, Bergerat M, Lebreton C, Refes Y, Blangis F, Chalumeau M, Le Roux E
Characteristics and Popularity of Videos of Abusive Head Trauma Prevention: Systematic Appraisal
J Med Internet Res 2024;26:e60530
URL: https://www.jmir.org/2024/1/e60530
doi:10.2196/60530
PMID:

©Luc Goethals, Victoria Prokofieva Nelson, Fabien Fenouillet, Karine Chevreul, Manon Bergerat, Christine Lebreton, Yacine Refes, Flora Blangis, Martin Chalumeau, Enora Le Roux. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences

Julia Walsh[1*], PhD; Jonathan Cave[2*]; Frances Griffiths[1,3*]

[1]Warwick Medical School, University of Warwick, Coventry, United Kingdom
[2]Department of Economics, University of Warwick, Coventry, United Kingdom
[3]Centre for Health Policy, University of the Witwatersrand, Johannesburg, South Africa
[*]all authors contributed equally

Corresponding Author:
Julia Walsh, PhD
Warwick Medical School
University of Warwick
Gibbet Hill
Coventry, CV4 7AL
United Kingdom
Phone: 44 02476528009
Email: julia.walsh@warwick.ac.uk

Abstract

Background: Patient experience data from social media offer patient-centered perspectives on disease, treatments, and health service delivery. Current guidelines typically rely on systematic reviews, while qualitative health studies are often seen as anecdotal and nongeneralizable. This study explores combining personal health experiences from multiple sources to create generalizable evidence.

Objective: The study aims to (1) investigate how combining unsupervised natural language processing (NLP) and corpus linguistics can explore patient perspectives from a large unstructured dataset of modafinil experiences, (2) compare findings with Cochrane meta-analyses on modafinil's effectiveness, and (3) develop a methodology for analyzing such data.

Methods: Using 69,022 posts from 790 sources, we used a variety of NLP and corpus techniques to analyze the data, including data cleaning techniques to maximize post context, Python for NLP techniques, and Sketch Engine for linguistic analysis. We used multiple topic mining approaches, such as latent Dirichlet allocation, nonnegative matrix factorization, and word-embedding methods. Sentiment analysis used TextBlob and Valence Aware Dictionary and Sentiment Reasoner, while corpus methods including collocation, concordance, and n-gram generation. Previous work had mapped topic mining to themes, such as health conditions, reasons for taking modafinil, symptom impacts, dosage, side effects, effectiveness, and treatment comparisons.

Results: Key findings of the study included modafinil use across 166 health conditions, most frequently narcolepsy, multiple sclerosis, attention-deficit disorder, anxiety, sleep apnea, depression, bipolar disorder, chronic fatigue syndrome, fibromyalgia, and chronic disease. Word-embedding topic modeling mapped 70% of posts to predefined themes, while sentiment analysis revealed 65% positive responses, 6% neutral responses, and 28% negative responses. Notably, the perceived effectiveness of modafinil for various conditions strongly contrasts with the findings of existing randomized controlled trials and systematic reviews, which conclude insufficient or low-quality evidence of effectiveness.

Conclusions: This study demonstrated the value of combining NLP with linguistic techniques for analyzing large unstructured text datasets. Despite varying opinions, findings were methodologically consistent and challenged existing clinical evidence. This suggests that patient-generated data could potentially provide valuable insights into treatment outcomes, potentially improving clinical understanding and patient care.

(J Med Internet Res 2024;26:e54321) doi:10.2196/54321

XSL•FO
RenderX

KEYWORDS

unstructured text; natural language processing; NLP; topic modeling; sentiment analysis; corpus linguistics; social media data; patient experience; unsupervised; modafinil

Introduction

Background

Spontaneously generated online patient experience (SGOPE) data collected from social media platforms are a rich data source for natural language processing (NLP) tasks [1-4]. Providing patient-centered perspectives [5,6] on the posters' experiences of disease, treatments, and health service delivery rather than the researcher-driven focus of published literature [7], SGOPE data are increasingly recognized as having the potential to transform clinical care and research [6,8-14].

Current estimates suggest that 3.6 billion people worldwide are currently using social media, with numbers forecast to increase to 4.4 billion by 2025 [15]. Social media were originally seen as being mostly used by younger people, but a 2019 US study showed that 73% of individuals aged 50 to 64 years and 45% of those aged ≥65 years used at least 1 form of social media [16]. SGOPE is recognized as being able to include a wider range of demographic groups, including many who may previously have been seen as "hard to reach" [17-19].

Modafinil is an oral wakefulness-promoting drug originally developed in the 1990s that is licensed by the UK National Health Service purely for narcolepsy, although its Food and Drug Administration classification in the United States allows it to be prescribed "off-label" for a wide variety of conditions [20]. Modafinil targets symptoms of fatigue seen in many clinical presentations; however, current randomized controlled trial (RCT)–based evidence regarding its efficacy for treating other conditions is inconclusive [21]. Having acquired a reputation as a "study drug," modafinil has sparked a large volume of online discussion about posters' experiences of taking it for both therapeutic and enhancement purposes.

Patient narrative is already recognized as a tool that can help patients, clinicians, and researchers [22,23]. Containing a mix of both objective and subjective views, SGOPE data provide a unique perspective on the way that patients perceive, manage, and react to their conditions, as well as how such conditions impact their life, their treatments, or other aspects of their health [24].

Although evidence-based medicine has been defined as the integration of the best research evidence with real-world clinical expertise and patient values (Sackett et al [25]), in reality, the pyramid-shaped hierarchy of evidence quality ensures that it is the findings from RCTs and subsequent systematic reviews, rather than any other form of knowledge, that tend to dominate and be reflected in the clinical guidelines [26-28].

The need for a plurality of evidence-generating methods is already recognized [29-31]. SGOPE represents a type of data that fall under the umbrella terms of real-world data (RWD) and real-world evidence. RWD include health care data generated from sources other than conventional RCTs, while real-world evidence is defined as evidence derived from the aggregation and analysis of RWD [32] and is argued to have significant advantages that can be used to supplement or augment RCT findings, including the ability to identify "clinical gaps" [33], indicating the effectiveness of an intervention in the real world, on much larger populations, and much faster than can be achieved within the artificial and highly constrained confines of an RCT [34,35]. Combining data sources such as SGOPE with new methods of analyzing unstructured data will enable the development of new and different approaches to knowledge and evidence generation.

Our previous study compared a thematic qualitative analysis with an NLP-based analysis of a small number of posts related to the therapeutic use of modafinil [21]. Eight main themes were identified from the posts, including details of the reasons for taking modafinil, conditions or symptoms, dosage, side effects, effectiveness, and outcomes in terms of quality of life, as well as details of other interventions whether previously tried, used concurrently, or subsequently moved on to. In this paper, we scale up this approach, using a combination of NLP and linguistic techniques to analyze a much larger dataset of modafinil experiences from a wide variety of social media platforms. We also compare the findings from some of the NLP tools used for the analysis to help future analysis of this type of data for health research.

Methodology

NLP approaches can be divided into 2 main types: supervised, which requires large quantities of the data to be labeled with the features of interest; and unsupervised, which uses clustering techniques that allow the data to tell their own story. Despite the development of ever-larger language models, such as GPT-3, which can be extremely resource heavy [36,37], there is an argument that to try to move nearer to the ultimate goal of natural language understanding, which is required to understand the complexity of patient experiences, entails stepping back toward combining unsupervised, rules-based methods with those from corpus linguistics [38,39]. To replicate the inductive data–driven approach of qualitative studies, but on a much larger scale, this study uses unsupervised methods. These include varied methods of topic modeling, sentiment analysis, and linguistic analysis.

Whichever approach is selected, cleaning the data is one of the most important and time-consuming components of the study. The cleaning process is specific to each project—each dataset has its own characteristics, and each project requires specific features from the dataset to answer the research question—but it is important to try to maximize the quality of the processed dataset for each subtask; for instance, in topic modeling, the aim of preprocessing is to reduce noise and incoherence from the data [40], allowing the themes to emerge. Stemming and lemmatizing words to their root form enables this, whereas when assessing effectiveness, it is important to retain all relevant details to understand the nuanced context within the text. Taking

XSL•FO
RenderX

too blunt an approach can result in the loss of potentially useful data.

Particularly suitable for exploratory and descriptive analysis, topic modeling can be used as a method for determining what people are talking about in social media by looking for underlying structure within the text [41]. Combining an inductive approach with quantitative measurement, topic modeling is a useful method for obtaining an insight into the concepts that are contained within documents in a similar manner to grounded theory [42], although it is not yet widely used in clinical NLP [43].

Sentiment analysis is a well-known and widely used technique within NLP that analyzes text for positive, neutral, or negative sentiment or emotion, aiming to extract an understanding of the meaning, mood, context, and intent. It has already been shown to be capable of reasonable agreement with online comments, including those rated using a Likert scale [44].

Causation is central to health care, both in understanding the onset of diseases or symptoms and the effectiveness of interventions or management strategies used to treat them [45]. Showing causation in health care using non-RCT data has been viewed as problematic. At both structural and cultural levels, causation is generally seen as something that can only be shown in empirical settings such as RCTs, where all confounding factors are controlled for, and the Humean principle of "same cause, same effect" can be repeatedly shown [46,47].

Causal dispositionalism is an alternative approach to causation, which may be relevant to this type of data. This takes a more nuanced view of how the characteristics or dispositions of both the intervention and the individual combine in complex ways to affect the effectiveness [48]. It suggests that population-level health research should be only 1 part of the evidence-generation process, and that it is listening to the patient narrative that can be the key to understanding their individual health needs [47]. One of the strengths of narrative data, such as SGOPE data, is that they enable both author and reader to make sense of the interplay of actions and contexts in the text in a way that conveys perceived causality [22]. The mantra "correlation does not equal causation" is justifiably used, but that leaves the question of how it is possible to determine causation.

Causation can be defined as a reaction between 2 events: a cause event and its consequence. The cause must precede the consequence and is counterfactual in that the consequence would not have occurred without the cause. While this sounds quite logical and straightforward, causation theories are not necessarily definitive explanations of how events occur but rather represent how humans make sense of, and understand, the world [49]. Williamson [50] argues that causation can be shown by identifying or understanding the underlying mechanism between a correlated cause and effect.

NLP methods still struggle with identifying potential causality; therefore, we used linguistic analysis to aid in this process. The language used to describe cause and effect can be crucial to understanding the semantic meaning of a text but is not always easy to identify. One method involves using transition words that link a reason to a consequence or indicate a sequence of events (Textbox 1).

Textbox 1. Examples of text that indicate sequential events.

Transition words

- Firstly
- to begin with
- next
- then following this
- at this time
- now
- at this point
- previously
- before this
- after
- afterward
- subsequently
- finally
- at last
- simultaneously
- meanwhile

Traditionally, findings from health-based qualitative studies have been seen as anecdotal, unrepresentative, and not generalizable across populations [51]. This study examines how we can move toward combining personal evidence of a health effect from sufficient numbers of people to the point where it

XSL•FO
RenderX

could be generalized and added to existing population-level evidence [47].

The aim of this study was to assess what can be learnt from an NLP-based analysis of a large quantity of unstructured SGOPE data. This can be broken down into 5 subquestions:

- To assess whether topic mining can elicit the themes that are contained in the data
- To explore how sentiment analysis can be used to assess perceived effectiveness
- To compare various methods of theme and effectiveness identification
- To assess whether linguistic analysis can identify perceived causality from the text
- To establish whether these techniques can be used to develop a methodology for this type of analysis

Methods

Overview

The dataset contained 69,022 publicly available social media posts and threads that included the terms *modafinil*, *provigil*, *armodafinil*, or *nuvigil* as of July 2017. The dataset was supplied by Treato Ltd, which was a web-based social media data mining service that collected publicly available health-related posts (ie, posts viewable by anyone without requiring log-in) from >10,000 global blogs and online forums. The company agreed to supply English-language data relating to modafinil use, using its own proprietary algorithms based on the Unified Medical Language System to create a searchable dataset that can be analyzed in aggregate [52].

Analysis code was developed using Python (version 3.8.5; Python Software Foundation) [53] in JupyterLab (version 3.0.15; Project Jupyter) [54]. Bearing in mind the need to retain as much context to the data as possible, as described in the Methodology subsection, we took a staged approach to data cleaning, initially performing a minimal level of transformation and parsing of fields. The time stamp field, originally formatted as *2011-01-01 00:00:00 UTC*, was simplified to *PostYear* to represent the year the post was published. Line breaks, paragraph breaks, and other extra spaces were removed. The URL field was parsed to identify the main website or forum name. New fields were created for subsite names. Having extracted the site name, it became obvious that many of the URLs contained either the name of the condition that was of primary interest to the poster or the title of the thread or question that they were referring to. Using clustering techniques, we were able to group and extract this detail from the URL. Three new fields were created to represent the second-level domain name, the site's focus condition (if applicable), and the extracted thread titles. To maximize the options for analysis, the cleaned data were structured to include 3 additional fields: *TextOnly* (response only), *Title* (thread title), and *TextWithTitle* (thread title preceding each response). All references to dosage amount in *mg* were standardized to *xxxmg*. Exact duplicate posts and obvious spam posts were removed. After data deduplication and spam removal, all forms of author identification were removed. The restructured file was saved in CSV format for the next stages. The *TextOnly* and *Title* fields were exported as 2 separate corpora text files for linguistic analysis. Keeping them distinct avoided the possibility of the repetition of the title words skewing any frequency-based analysis. These steps enabled us to obtain a dataset that retained an optimal level of quality and flexibility and upon which further preprocessing could be performed specific to the individual task.

Topic Modeling to Identify Themes

Topic modeling was the main method for theme detection. On the basis of a previous study that evaluated 4 of the most widely used bag-of-words topic modeling methods [55], we selected latent Dirichlet allocation (LDA) and nonnegative matrix factorization (NMF) for comparison because they were seen to deliver the most meaningful extracted topics. Both LDA and NMF use the bag-of-words approach, which disregards any order within the corpus and uses word frequency to generate topics. Although the LDA method has been the most widely used method for patient experience feedback [56], a previous study found that NMF yields better results than LDA when used for short texts [57]. Other comparisons between the 2 methods found that LDA output was more semantically interpretable with more distinct categories [58], while NMF was faster and therefore less resource intensive [59]. However, another comparison found the opposite [60]. Yet another study suggested that NMF returned higher quality topics than LDA on smaller datasets [61]. As part of the project involves identifying a methodology for this type of data that can be developed for use on other datasets, we compared the findings of both methods using the *gensim* (version 3.8.3) [62] and *sklearn* (version 0.23.1) [63] libraries as they relate to SGOPE data. Another package—Top2Vec (version 1.0.24) [64]—using word-embedding methods was released during the study and was included for comparison. Word-embedding methods work by considering each word in the context of its neighbors, creating a numeric vector where words with similar meanings are grouped together, which has been seen as a significant advance in trying to establish the meaning or topics of posts [65].

Additional preprocessing for the LDA and NMF methods included removing stop words and punctuation and converting all text to lowercase. The stop word list was extended to include common name variations for modafinil. Bigrams and trigrams were generated; text tokens were lemmatized; and part-of-speech (POS) tags relating to nouns, adjectives, verbs, and adverbs were retained. Coherence and perplexity values were generated to help assess the performance of each model. The LDA outputs included generating the 10 most discriminative words for each topic; the weighting of each word within the allocated topic; and, for the *gensim* LDA model, a computer-based visualization (pyLDAvis [version 2.1.2]) that demonstrated the words for each topic and the degree of overlap between topics [66]. This visualization could also be used to show varying values of alpha and beta, the balance between words per topic and topics per document.

For the embedding-based method, no preprocessing of the text or prespecified number of topics was required because the

XSL•FO
RenderX

Top2Vec algorithm calculates the number of topics contained within the corpus.

Sentiment Analysis to Evaluate Effectiveness

Two widely used lexicon-based methods—TextBlob (version 0.15.3) [67] and Valence Aware Dictionary and Sentiment Reasoner (VADER; version 3.3.2) [68]—were compared and the strengths and limitations of both identified. The original cleaned *TextOnly* field was selected for the sentiment analysis because this contained only the responses to the posts. Word counts were calculated for each post. Capitalization, punctuation, and stop words were retained for this part of the analysis because each can contribute meaning or intensity to the analysis. TextBlob [67] calculates values for polarity and subjectivity for each post. The lexicon it uses derives from a separate library in the Natural Language Toolkit. It focuses on adjectives from customer product reviews that have been tagged by humans for polarity and subjectivity. Subjectivity analysis assesses how objective or subjective the text is, whereas polarity classification determines whether the text is positive or neutral. It uses the sentiment lexicon to assign scores for polarity and subjectivity for each word, which are then averaged out using a weighted average to provide an overall sentence sentiment score. Basic statistics were generated for both values, and the numerical polarity score was converted to categorical values of positive (>0), neutral (0), and negative (<0). Plots showing the distribution and the relationship between the polarity and subjectivity scores were generated.

The methods behind the design of the VADER library make it possibly a better choice for sentiment analysis of social media–type posts than TextBlob [69]. Rather than calculating the polarity and subjectivity of a post, it scores each post on 4 aspects: positive, negative, neutral, and compound. The positive, negative, and neutral scores represent the proportions of the post that fall in these categories. The compound score is calculated from the other 3 scores, normalized to a value between –1 and 1, and represents the overall sentiment of the post [68]. The lexicon VADER uses is based on general language rather than reviews [70] and contains approximately 7500 words.

Although the basic sentiment is calculated on the individual words, VADER looks at the whole text and can take negations into account [71]. This can help to give a balanced assessment when the post contains contradictory words out of context. This approach is intended to take into account some of the characteristics often seen in SGOPE data where features such as repeated punctuation or capital letters can be used to signify stronger sentiment [68].

The VADER lexicon is easily modified. After reviewing the positive and negative words it had identified from a sample of posts at each end of the sentiment spectrum, we modified the lexicon, removing the positive words *credit*, *free*, *accepted*, and *approval* because these words were frequently included in spam posts. We also added frequently mentioned effects to the negative lexicon, including *headache*, *jittery*, *rash*, *tired*, *harmful*, *disappointed*, *sleepy*, *nightmare*, and *intolerable*. In addition, we modified the positive lexicon to include *awake*, *focus*, *concentrate*, *normal*, *productive*, *helped*, *grateful*, *miracle* and *lifesaver*.

The results from each method were then compared against each other.

Linguistic Analysis

We extracted the narrative fields from each post to form a corpus, which was then imported into Sketch Engine [72], a corpus linguistics tool. Each token was assigned a POS tag from the English TreeTagger POS tagset with Sketch Engine modifications [72]. Using the English Web corpus 2020 as a reference corpus [73], we generated lists of the top 1000 keywords, key terms, and n-grams specific to the dataset to help identify both themes and examples of causal text. N-grams are sequences of words, numbers, or symbols that appear in a specific order within the text and are helpful in identifying commonly used phrases of up to *n* words within the corpus [74]. For each word or term in the lists, we recorded its frequency in the focus corpus, the number of posts it appeared in, and a calculated score based on its relative frequency in each corpus. We then classified the top 100 highest-scoring keywords and key terms into themes and summarized the results to see how this technique compared to the topic modeling. N-grams that indicated a possible cause and effect or temporal dimension were identified. Combining these selected n-grams with concordance techniques revealed specific relevant sentences that expressed the poster's understanding of these sequential events.

Ethical Considerations

Ethics approval for the study was granted by the University of Warwick (BSREC Ref 11/19-20) in October 2019. No personally identifiable information other than the online "user handle" was included in the data collection, and this was removed and replaced with a unique ID for each post as part of the cleaning and preparation process.

Results

Descriptive

The cleaned dataset contained 68,559 records from a 6-year period (2011-2016). A total of 790 unique top-level sites were identified, with the number of posts per site ranging from 25,355 to 1. Reddit was the largest overall source, with 36.98% (25,355/68,559) of the posts from 213 subreddits, each of which represents a separate community. Of the 213 subreddits, 5 (2.3%) contributed >1000 posts, with the largest being the *afinil* subreddit (n=12,870, 18.77% posts). Post lengths ranged from 1 to 1577 (mean 100.4, SD 100.86; IQR 34-132) words. The *TextOnly* field comprised 7.99 million tokens, 6.84 million words, 104,565 unique words, and 388,516 sentences. Parsing the site or forum URLs revealed 166 separate health conditions. Multimedia Appendix 1 shows analysis by the number of posts posted to the top 10 condition-specific sites. This does not assume that the specified condition was the primary or sole condition of the poster but rather reflects the poster's choice in selecting where to post their contribution.

XSL•FO
RenderX

Topic Modeling

Overview

First, using the *gensim* LDA library, initial parameters were set to 8 topics (as per the earlier themes identified [21]) and 50 iterations. The default output is the top 10 words per topic, together with the weighting of each word within the topic. Although the returned topic word lists could all be seen to relate to the poster's experience, they did not seem to be clearly distinguishable from each other. The visualization (Multimedia Appendix 2) indicates a substantial overlap of topics 1 to 4, which between them represented 72.7% (49,842/68,559) of the tokens.

Coherence model testing (Figure 1) using the NMF method (range 5-50) suggested that the optimal number of topics was 27; therefore, we ran the model again with varying numbers of iterations across the data.

Figure 1. Coherence testing model (range 5-50).

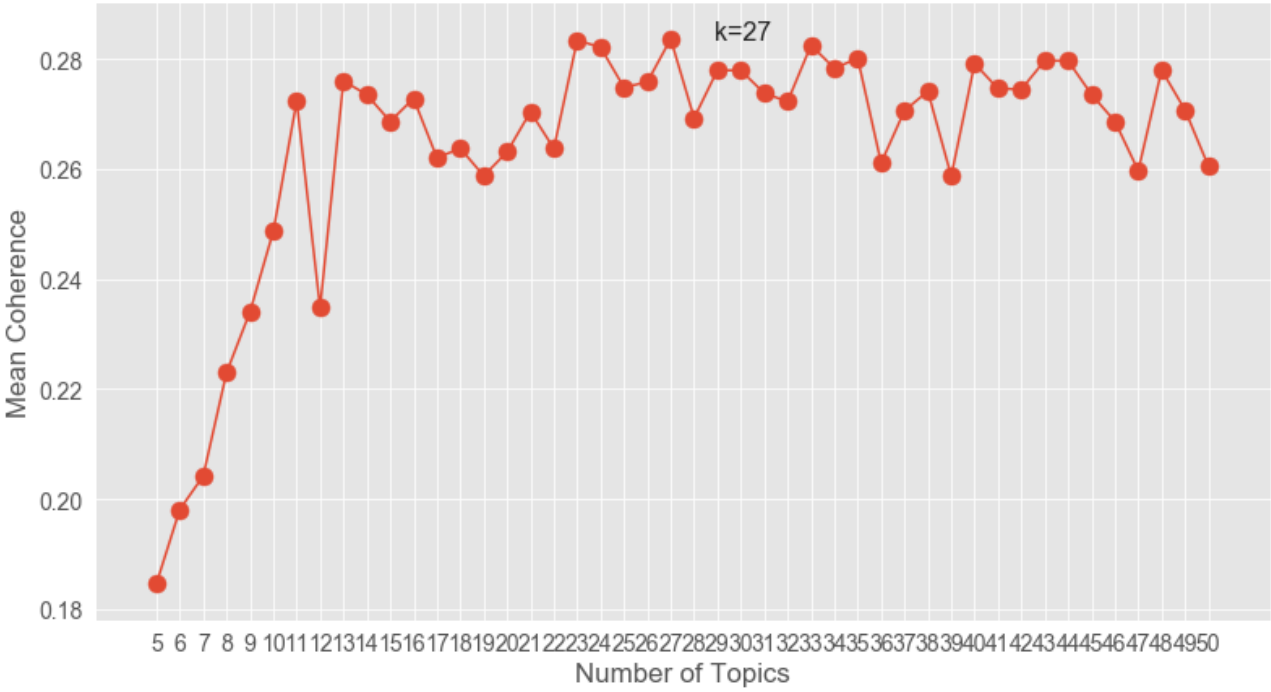

Gensim LDA

Running the LDA model with parameters of 27 topics and 200 passes (Multimedia Appendix 3) showed a clearer distribution of topics, but there was still a substantial degree of overlap of topics 1 to 6. Increasing the number of passes to 1000 did not seem to significantly improve the visual evaluation (Multimedia Appendix 4), although it took >5 times as long to run.

Although both visualizations show some distinct topic circles that are not overlapped by others, the categorization of the topics into themes was not possible because most of them could have multiple interpretations. The top 10 topic words for each of the 27 topic models and the attempted mapping are shown in Figures 2 and 3.

XSL•FO
RenderX

Figure 2. Latent Dirichlet allocation model: 27 topics.

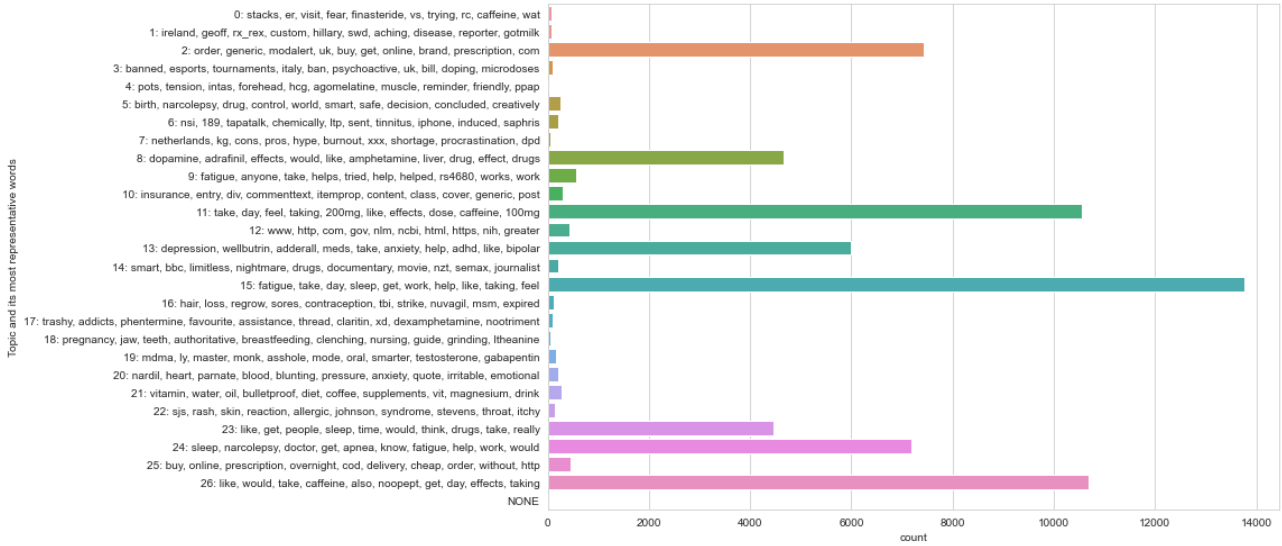

In terms of the processing load, the timings of the *gensim* LDA models were impacted far more by the number of iterations through the data than the number of topics selected, with the simplest configuration—8 topics and 50 iterations—taking 32 minutes, 27 topics and 200 iterations taking 2 hours 16 minutes, and 27 topics and 1000 iterations taking 11 hours and 6 minutes. Adjusting the memory handling parameters reduced the processing time significantly (13 min, 1 h 44 min, and 8 h 13 min, respectively) but gave the highest coherence score to a model with just 2 topics and 10 passes, which did not seem a plausible result.

Sklearn LDA and NMF Methods

Running the same 27-topic model with the *sklearn* library enabled a direct comparison of the LDA and NMF methods. Multimedia Appendix 5 presents a comparison of the top 10 words per topic and the number of posts each model classified as belonging to each topic, together with the percentage of the corpus per topic in descending order for each method. It also includes our evaluation of the theme that the topic words most closely indicated. As with the earlier *gensim* LDA models, trying to map each of the returned topic word lists to the identified themes was complicated by the degree of overlap in most of the lists. The bar graphs (Figures 2 and 3) show that the NMF method returned topics that were distributed slightly more evenly throughout the corpus, whereas the LDA version identified some topics that were much less represented. The *sklearn* LDA model allocated 94.45% (64,753/68,559) of the posts to just 8 (30%) of the 27 topics; the remaining 19 (70%) topics each represented <1% (3806/68,559) of the posts. In comparison, the largest NMF topic was assigned to 16.6% (11,381/68,559) of the posts, with the remaining 26 ranging from 5.4% (3702/68,559) to 2% (1371/68,559) of the posts. Future work could look at going back to the posts included in some of the smaller topics to assess their relevance to the research question.

Mapping the topics found by both models, even at a superficial level, to distinct themes was problematic. For the *sklearn* LDA model, only 26% (7/27) of the topics could be mapped to the general themes. The NMF model was slightly more interpretable with 52% (14/27) of the topics that could be seen as relating to themes.

XSL•FO
RenderX

Figure 3. Nonnegative matrix factorization model: 27 topics.

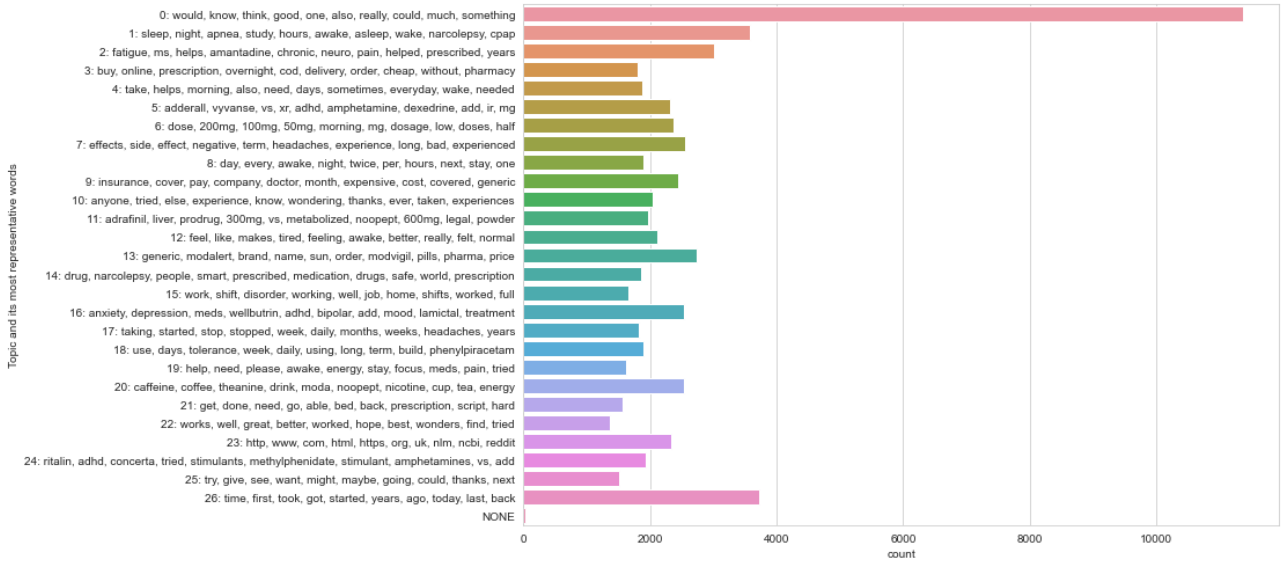

Top2Vec Library

The Top2Vec library demonstrated substantially faster performance compared to the LDA method. By default, it returns the number of detected topics, the top 50 words per topic, and the number of posts per topic. The optimal DeepLearn parameter took 2 hours 15 minutes to generate 367 topics from the dataset, while the Learn parameter took 19 minutes to generate 566 topics.

The results from the DeepLearn model were used for analysis. The percentage of posts per topic ranged from 2.94% (2017/68,559) in the largest group to 0.07% (45/68,559) in the smallest. Overall, 70% (257/367) of the posts could be mapped to either the P1 themes or the codes used during the thematic analysis. The P1 study refer to the previous part of the study where we compared a sample of 260 posts using a qualitative analysis with a basic NLP or corpus [21]. In total, 186 (50.7%) of the 367 topics representing 38,637 (56.36%) of the 68,559 posts could be mapped to the P1 themes. A further 71 (19.3%) of the 367 topics representing 15,557 (22.69%) of the 68,559 posts were mapped to the codes.

In total, 110 (30%) of the 367 topics representing 14,345 (20.92%) of the 68,559 posts were initially categorized as being uninterpretable without taking a deeper look at the specific posts. Of the 367 topics, 31 (8.4%; 3913/68,559, 5.7% posts) combined multiple themes and were classed as mixed; 50 (13.6%; 7019/68,559, 10.24% posts) were uninterpretable and were labeled unclear; and 29 (7.9%; 3413/68,559, 5% posts) contained words indicating that the topics related to possible spam posts.

Sentiment Analysis

The TextBlob library returns values for both polarity and subjectivity. Of the 68,559 posts, the initial results for polarity were as follows: 47,282 (69%) positive, 6229 (9.09%) neutral, and 15,048 (21.95%) negative. The polarity scores extended across the whole range from −1 to +1 (mean +0.1003). The subjectivity scores also covered the entire range from 0 to +1 (mean +0.4638).

Using the previously mentioned parameters of positive (>0), neutral (0), and negative (<0), the initial results returned from the standard VADER analysis were 64.03% (43,898/68,559) positive, 6.7% (4592/68,559) neutral, and 29.27% (20,070/68,559) negative. Modifying the lexicon yielded the following results: 65.01% (44,610/68,559) positive, 6.44% (4417/68,559) neutral, and 28.49% (19,533/68,559) negative. The compound score values ranged −0.9991 to +0.9997 (mean +0.2825). The distribution is shown in Table 1.

Table 1. Basic statistics for the extended VADER analysis (n=68,559).

	Compound	Positive	Neutral	Negative
Scores, mean (SD; min-max)	0.28250790 (0.61562543; –0.99910000 to 0.99970000)	0.11785168 (0.09204523; 0.00000000-1.00000000)	0.81442440 (0.10185110; 0.00000000-1.00000000)	0.06772396 (0.06403353; 0.00000000-0.67000000)
Percentile values				
25%	–0.1779	0.0590	0.7590	0.0120
50%	0.4515	0.1070	0.8200	0.0580
75%	0.8407	0.1600	0.8760	0.1010

XSL•FO
RenderX

Although the results from both Vader and TextBlob methods were similar, with both showing a majority of posts being assessed as positive, comparing the distribution shape of the sentiment values between the methods showed distinct differences. Both are skewed toward the right, indicating the positive mean value; however, TextBlob showed a normal type of distribution of polarity apart from those posts classified as neutral, whereas Vader showed a similar peak at 0 but seemed to assess more of the posts as being at the extremes of the available range (Figure 4).

Figure 4. TextBlob and Valence Aware Dictionary and Sentiment Reasoner (VADER) distributions.

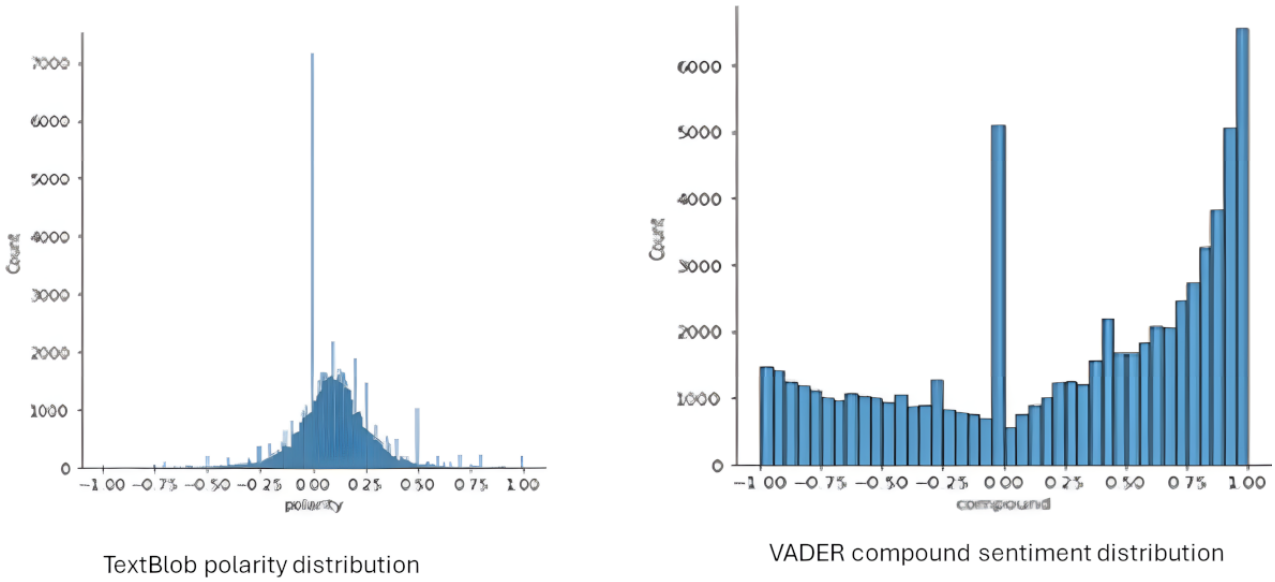

The average word count of the 10 highest-rated posts based on the VADER analysis was 704, and that of the lowest-rated posts was 1095. For TextBlob, the average word count of the 10 highest-rated posts was 39, and that of the lowest-rated posts was 23. VADER is reported as performing better on short texts [68]. The P3 dataset (total 68,559 posts) contained 1232 posts with a word count of >400 and 8496 posts longer than 200 words. However, running VADER again on the reduced datasets showed little difference in the percentages of posts rated in each category (Table 2; Figure 5).

Table 2. Valence Aware Dictionary and Sentiment Reasoner results from limiting post length.

	All posts, standard (n=68,559)	All posts, extended (n=68,559)	<400 words, extended (n=67,327)	<200 words, extended (n=60,063)
Compound scores, mean (SD; IQR)	0.2658 (0.613580; –0.2040 to +0.8250)	0.2819 (0.61587235; –0.1794 to +0.8404)	0.2816 (0.609968; –0.1779 to +0.8438)	0.2658 (0.587878; –0.1655 to +0.7984)
Positive, n (%)	43,898 (64.03)	44,586 (65.03)	43,781 (64.18)	38,546 (64.18)
Neutral, n (%)	4592 (6.70)	4416 (6.44)	4416 (6.56)	4414 (7.35)
Negative, n (%)	20,070 (29.27)	19,557 (28.53)	19,130 (28.41)	17,103 (28.48)

XSL•FO
RenderX

Figure 5. The impact of word count on sentiment. VADER: Valence Aware Dictionary and Sentiment Reasoner.

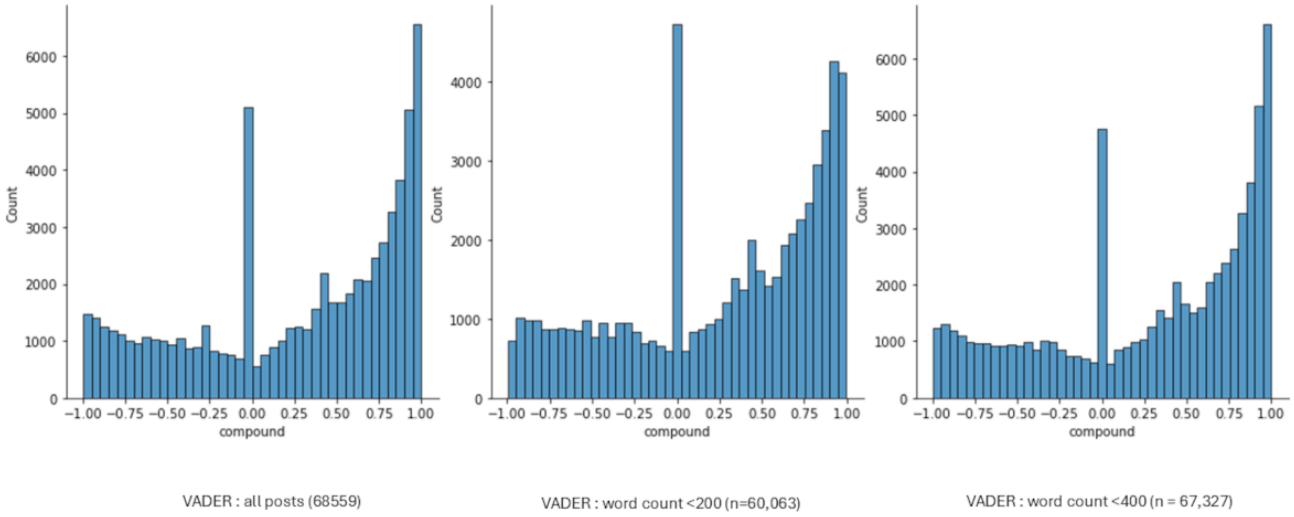

Corpus Linguistics

Using the corpus linguistic tool Sketch Engine, we generated 1000 key n-grams specific to the SGOPE corpus, identifying many phrases that could suggest a form of causality. Attempting to map these key n-grams to the individual themes was problematic. Unlike the key words and terms, only 16 (16%) of the top 100 n-grams specific to the corpus could be directly mapped to themes. A full analysis would require looking at the n-grams in the context of the post. However, the key n-grams are helpful in detecting expressions of causality. Unlike the individual words, all of which have a POS tag that can indicate tense, n-grams are combinations of words. It was possible to label many of them as relating to past, present, or future tense or as indicating possible belief. Examples are shown in Table 3.

XSL•FO
RenderX

Table 3. Key n-grams indicating possible belief.

Key n-gram	Frequency (n=68,559), n	Total number of documents including the phrase (n=68.559), n	Score (relative frequency compared to the reference corpus)	Theme	Tense
keep me awake	406	396	50.0	Effect	Present
works for me	408	398	49.2	__[a]	Present
i have found	458	440	48.8	—	Past
but it does	488	485	48.3	—	Present
i find that	403	388	46.1	—	Present
was able to	610	579	46.0	—	Past
that i can	474	460	45.4	Outcome	Present
i felt like	396	377	45.2	Effect	Past
i find it	407	400	44.4	—	Present
gave me a	395	389	44.3	—	Past
in my experience	377	365	43.8	—	N/A[b]
because i have	381	380	42.6	—	Present
because i was	377	363	41.2	—	Past
because of the	576	561	34.7	—	N/A
and i think	368	363	32.3	—	Present
in my opinion	301	293	29.7	—	N/A
and it seems	258	257	29.7	—	Present
i have noticed	242	235	29.1	—	N/A
but i feel	241	237	28.7	—	Present
it gives me	230	226	27.9	—	Present
to kick in	225	217	27.9	Effect	N/A
seems to work	229	226	27.6	—	Present
it seems to be	237	237	27.4	—	Present
has helped me	225	219	27.2	—	Present
because i do	236	233	27.1	—	Present
effect on me	216	212	26.9	Effect	N/A
me feel like	220	216	26.9	Effect	N/A
it gave me	218	213	26.7	Effect	Past
changed my life	216	209	26.5	Outcome	Past
but it seems	231	231	26.3	—	Present
gives me a	216	210	26.3	Effect	Present
think it is	255	247	26.3	—	Present
as soon as i	227	223	25.9	—	Present
i can say	229	218	25.6	—	Present
it does help	205	204	25.6	—	Present
for me is	212	208	25.5	—	Present
i still feel	206	200	25.4	Effect	Present
my experience with	204	201	25.0	—	N/A
and i know	228	225	24.8	—	Present
thought i was	211	204	24.7	—	Past

XSL•FO
RenderX

Key n-gram	Frequency (n=68,559), n	Total number of documents including the phrase (n=68.559), n	Score (relative frequency compared to the reference corpus)	Theme	Tense
thought it was	237	233	24.6	—	Past
and it helps	196	194	24.4	—	Present
know if i	208	204	24.3	—	N/A
i felt like i	198	188	24.1	Effect	Past
i found it	209	202	24.0	—	Past
i thought it	229	227	23.9	—	Past
seems to have	242	234	23.5	—	Present
it helps with	185	183	23.2	—	Present
it has helped	187	185	23.2	—	Past
it seems that	232	227	23.2	—	Present
i know this	200	197	23.2	—	Present
feel like it	190	186	22.9	—	Present
because of my	191	188	22.9	—	N/A
am able to	189	178	22.9	—	Present
great for me	182	182	22.8	—	N/A
i can sleep	181	177	22.8	Effect	Present
i started to	197	186	22.8	—	Past
and it worked	186	186	22.7	Effect	Past
have found that	198	195	22.7	—	Past
give you a	228	226	22.7	—	N/A
and i felt	188	184	22.6	Effect	Past
it wears off	176	172	22.2	Dosage	N/A
a huge difference	183	180	22.2	Effect	N/A
better for me	177	176	22.2	—	N/A
this is a	642	627	22.2	—	Present
i found out	187	181	21.7	—	Past

[a]Could not be mapped.

[b]N/A: not applicable.

The n-gram “have found that” was shown to be indicative of causal expression in the exploratory study [21]. Using it on the P3 dataset and filtering out any of the sentences that did not explicitly mention modafinil or one of its name variants in the concordance sentence returned the examples presented in Textbox 2.

XSL•FO
RenderX

Textbox 2. Concordance examples for the n-gram "have found that.

Tolerance

• "I have been on Nuvigil for about 2 years now, and I have found that I have to skip my medication at least one day per week in order to not lose its effectiveness." [Post ID 6289]

Side effects

• "I have found that I get visuals from modafinil anyways, for the first few hours of it's effects I have mild visuals and a solid body load." [Post ID 7711]

• "After taking modafinil 200mg next day i have found that i have a skin rash on the right hand and itchy skin on both hands." [Post ID 26,660]

Dosage

• "Forgetting and False Memories I am on Nuvigil, and I have found that I become a 'zombie' when they have my dosage too high." [Post ID 29,323]

Other Intervention, but effect or outcome

• "I have found that I have been able to reduce my Prozac dosage while taking Provigil." [Post ID 53,387]

Outcome

• "I also have found that I am much more confident since started on provigil (200mg/day)." [Post ID 59,900]

Comparison

• "I have tried Adderall and Provigil and have found that I prefer a sister drug to the Provigil called Nuvigil, but my insurance company won't pay for it so I'm stuck with the Provigil or Adderall." [Post ID 67, 037]

The word sketch tool can be used to demonstrate the context of how any word or phrase is used within the corpus. Many of the key n-grams for this corpus relate to an observation the poster has made or an effect they have noticed in relation to the subject of their post. The most frequent key n-gram in the corpus is "in the morning," which appears 3016 times in 2627 posts. Using the corpus query language to filter down to only those concordances that included modafinil in the same sentence returned 183 examples of dosage patterns, amounts, drug combinations, timing advice, and effects. As with the P1 study, posters reported how the standard dose can be excessive for some people [21]:

...my Dr prescribed starting dose of 200mg modafinil...once in the morning...with the instruction that if the200mg did not keep me awake that I should double the dose to 400mg once a day in the a.m...the 200mg was too much all at once...all it did was enhance the side effects to the point that I wasn't able to notice if the medicine was doing what it was supposed to.because I was too busy cradling my cracked feeling skull and drinkn insane amounts of water. [Post ID 3209]

Another frequent lemma related to effectiveness in the n-grams is "feel," which has been used by post writers in many ways. As a verb, it was used 22,767 times in the corpus. Splitting the occurrences into grammatical categories, as shown in Table 4, highlights the categories, some of the most frequent examples of each phrase from the corpus, and the number of occurrences for each category. A visual representation of the most frequent adjectives and objects associated with the verb "feel" is shown in Figure 6, while Figure 7 displays the most frequent collocates. The size of each circle represents the frequency of the collocate. Of note, "good" is the most prominent adjective collocate of "feel," supporting the hypothesis that modafinil is perceived as effective by many of the posters. The full list of collocates of "feel," together with their frequencies in the corpus, is available in Multimedia Appendix 6.

Feeling normal was identified as being an important outcome for some posters in the earlier study [21]. Textbox 3 presents examples of n-gram concordances for "makes me feel," filtered by "normal."

XSL•FO
RenderX

Table 4.

Grammatical categories	Examples	Frequency (n=68,559), n (%)
pronominal subjects of “feel”	I feel, you feel, made me feel, it feels	12,026 (17.54)
modifiers of “feel”	Don’t feel, I still feel, I just feel, really feel	6842 (9.98)
adjectives after “feel”	feel better, feel tired, feel worse, feel great, feel sleepy, feel normal	5342 (7.92)
objects of “feel”	feel the effects, feel a bit, felt nothing	4354 (6.61)
prepositional phrases associated with “feel”	feel like, feel in, feel on, feel though	2163 (3.15)
subjects of “feel”	I feel, my body feels, I don’t feel	2032 (2.96)
pronominal objects of “feel”	feel it, you feel you, feel myself	689 (1)
complements of “feel”	feel a lot better, felt it more, felt a bit weird	289 (0.42)
“wh-” words following “feel”	feel when, feel what, I feel that, feel how, feel normal which	179 (0.26)
“feel” and or	sleep and feel, yawning and feeling	150 (0.22)
“-ing” objects of “feel”	felt taking, felt amazing	81 (0.12)
particles after “feel”	feel up to it, feeling down,	74 (0.11)
infinitive objects of “feel”	it feels to be	37 (0.05)
particles after “feel” with object	feel hyped up, to feel out	19 (0.03)

Figure 6. Most frequent adjectives and objects of “feel.”.

Figure 7. Word sketch of the verb “feel.”.

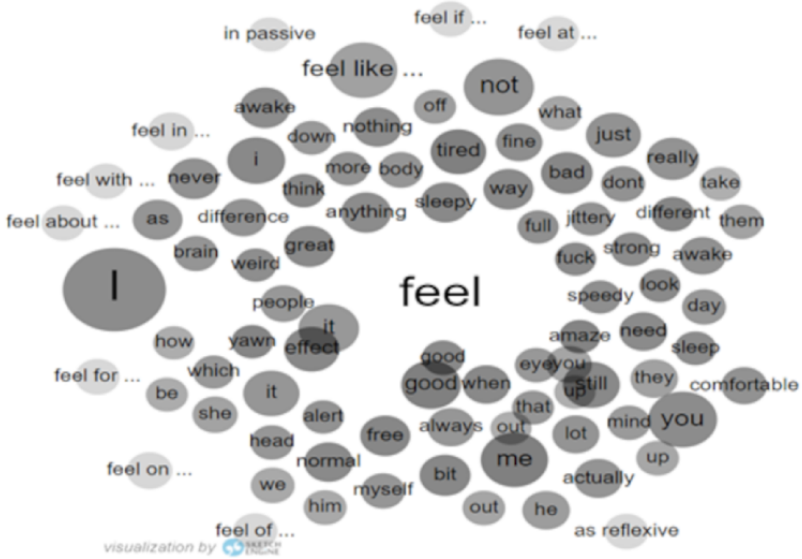

XSL•FO
RenderX

Textbox 3. Concordance of "makes me feel" with "normal."

Examples of n-gram concordances for "makes me feel," filtered by "normal"

- I have never noticed excessive energy or anything out of the ordinary; it just makes/make me/me feel/feel like a normal person would.
- Taking the whole thing almost makes/make me/me feel/feel normal for a while.
- Anything that makes me less sleepy makes/make me/me feel/feel more "normal" (i.e., less tired), and not high (course I am not shooting it in my arm or anything).
- While Modafinil *feels* like a some sort of drug-induced happiness, Zoloft actually makes/make me/me feel/feel naturally normal and happy.
- Cheers. :) I am on Modafinil which makes/make me/me feel/feel normal most of the time. @Nicole – I'm showing my age, but as a student it was ProPlus every time for me!
- It just makes/make me/me feel/feel closer to normal.
- At first I did feel speedy but now it just makes/make me/me feel/feel normal (ish)!!
- Doesn't jack me up or give me jitters - just makes/make me/me feel/feel as "normal people normal" as I can imagine.
- My epileptologist has just put me on nuvigil for sleepiness and it really helps, there is only a day here and there it doesn't but it's awesome now most of the time I have the energy that my family has (2 kids) doesn't make me hyper just honestly makes/make me/me feel/feel more normal.
- I take Nuvigil, and, unlike stimulants, it just makes/make me/me feel/feel normal without the waves of crippling exhaustion or a crash at the end of the day.
- Nuvigil makes/make me/me feel/feel like a normal person again and without it, my quality of life is severely decreased.
- I love nuvigil and it makes/make me/me feel/feel "normal" and have a "normal" life but somedays I feel like I could use another pill and if its *safe* to take it twice a day then that may help me ALOT!!
- I have read posts where people talk about feeling revved up from it but for me it just makes/make me/me feel/feel normal.
- The provigil makes/make me/me feel/feel normal.
- It just makes/make me/me feel/feel normal which is perfect...no jitters.
- It makes/make me/me feel/feel normal.
- It makes/make me/me feel/feel pretty normal like I used too.
- I usually take it around noon at work during the week and it makes/make me/me feel/feel normal, and I can get through the rest of the day.
- It makes/make me/me feel/feel normal.
- I am taking 200mg an hour before work and it makes/make me/me feel/feel normal.I try not to take it every day, but it definitely helps...makes/make me/me feel/feel normal almost.
- I've now been feeling like it makes/make me/me feel/feel more "normal" (normal energy & focus) for a few hours past my dose (8am and 2pm) and the other times are like a complete drop in energy, not even normal tired....just SO exhausted.
- (It wasn't my first choice.) The only thing that makes/make me/me feel/feel close to normal is use of stimulants such as Nuvigil, but those give me serious insomnia.
- I hate that a pill/pills makes/make me/me feel/feel normal.
- It doesn' make me feel buzzed or jittery, it just makes/make me/me feel/feel "normal."

Comparison With Existing Trial Evidence

The effectiveness of modafinil suggested by this study contrasts with the existing RCT and systematic review evidence that is used to determine treatment pathway options for clinicians [28]. Rather than searching for every review or RCT of modafinil, we used Cochrane reviews as a comparison. Cochrane reviews critically appraise individual trials, are recognized as providing high-quality assessment and evidence synthesis, and are also used to contribute to the development of clinical guidelines [75]. As of May 2021, a search of the Cochrane Library [76] showed that there were 16 published Cochrane reviews for various conditions that included the term modafinil in the title, abstract, or keywords. To compare the findings, we extracted the authors' evidence summaries, the quality assessments of the evidence, and suggestions for addressing the remaining uncertainties relevant to this project (Multimedia Appendix 7). All reviews were inconclusive, with either insufficient [77-85] or low-quality [86-92] evidence of effectiveness. One of the main findings of this study was that although modafinil is only currently licensed by the National Institute for Health and Care Excellence for a single condition within the United Kingdom, posters were finding it effective for a wide range of conditions, including central disorders of hypersomnolence, multiple sclerosis, attention-deficit disorder and attention-deficit/hyperactivity disorder, social anxiety, depression, sleep-related breathing disorders, general fatigue, myalgic encephalomyelitis and chronic fatigue syndrome, and fibromyalgia (Figure 1). Other conditions for which modafinil was used included cancer fatigue, traumatic brain injury, diabetes, epilepsy, autoimmune conditions, pain, irritable bowel

XSL•FO
RenderX

syndrome, hepatitis C, and poststroke fatigue. Multimorbidity was a regular feature.

Discussion

Principal Findings

Although a range of positive and negative experiences were reported, our analysis indicates that posters found modafinil effective for their symptoms with similar levels of effectiveness found across all methods. Similar themes were identified by both qualitative and computational analyses. Difficulties in obtaining a prescription or acquiring modafinil were common. All topic-modeling methods returned topics containing words that clearly related to and could be mapped to the themes and subcodes from the earlier qualitative study [21]. Linguistic analysis identified expressions of causal belief.

The overall methodology of the study was designed so that it can be applied to other health-related research questions that use unstructured data. The principles underlying the methods used in this study have shown that they can be used inductively on large volumes of unstructured text to extract the themes, sentiment, and expressions of perceived causality.

As an inductive and iterative method, topic modeling shows potential for scaling up qualitative analysis [43,61,93] when working with large volumes of data. The requirement of both the LDA and NMF methods for a defined number of topics to be determined before running the models is problematic. Previous comparisons of findings from both manual coders and NMF topic modeling found that neither group could agree on the ideal number of topics [61]. Using the Top2Vec method had the advantage that it did not require a predetermined number of topics or themes to be specified. The Top2Vec embedding-based method was more effective in eliciting topics that mapped to those previously identified through qualitative analysis [21]. A possible disadvantage of this model is that, depending on the dataset, it may return too many topics [94], but this can be mitigated in a later version of the model through the use of hierarchical topic reduction [64].

Previous studies have commented on how lexicon-based tools trained on general language do not perform as well on health-related text [3]. Although lexicon-based sentiment analysis can provide an accurate assessment of text that contains words that express a strong positive or negative sentiment, posts that do not contain many of these predefined words are harder to evaluate. One of the features of the informal nature of SGOPE data is that the writers assume that readers can readily infer the affective reaction they are describing. Descriptive phrases such as “I could go back to work” or “It gave me a headache” suggest the effect of the event but would be viewed as neutral statements by most sentiment analysis models. Developing lexicons that are more relevant to health outcomes would improve and refine the results.

The inclusion of linguistic analysis added a depth of understanding to the findings that would not have been possible with a pure NLP approach [38,39]. The reported rapid onset of the effect of modafinil, whether positive or negative, together with the temporal sequencing, allowed the identification of text indicating perceived causality.

Unsupervised methods align more with the inductive approach of qualitative studies and are shown to be effective for exploring SGOPE data. Although topic modeling has not yet been widely used within health research, previous studies have shown how it can be used to generate findings in a similar fashion to grounded theory [21,42]. Both topic modeling and the extraction of keywords, key terms, and key n-grams identify what is being spoken about but not how the word or phrase is used in context. Combining NLP with corpus linguistics draws on the strengths of both disciplines [38,39] and allows the researcher to identify the content that is most relevant to the research question [95].

This research could be extended in a variety of ways that could be used to improve health outcomes. Extending the case study approach, these could include extracting features such as dosage detail and treatment duration, examining more granular topics, further refining the lexicons used for sentiment analysis, and conducting tense analysis of POS tags of modafinil or other interventions. Combining NLP with linguistics on large quantities of unstructured data could be a valuable source both to identify “off-label” indications and obtain a deeper understanding of the outcomes that patients and their families prioritize and how they are managing their conditions. In terms of methodological development, these methods could also be applied to many different types of unstructured text sources, such as qualitative interview transcripts or the free-text sections of clinical notes.

Strengths and Limitations

The use of unsupervised methods allows for an inductive approach to analysis, and the comparison of findings from multiple methods with those from the exploratory dataset is a strength of this study. SGOPE data analysis relies on the poster’s self-description of their condition, which may include self-diagnosis rather than a clinician’s assessment. The reporting of symptoms and outcomes may not be as accurate or complete as it could be, although this limitation could apply to any form of self-reported data, whether collected in a trial, clinical encounter, or on the web. Self-reported data, particularly regarding hard-to-measure factors such as fatigue and cognition, are subjective but generally reflect the normative value of the patient. The natural, nonclinical language used in informal texts may contain valuable, unexplored, or overlooked information relevant to clinical or research purposes [96], but it can also contain spelling or grammatical errors and inappropriate slang or colloquialisms that pose challenges for NLP methods [97]. Keyword comparison with a reference corpus was found to be effective in identifying such terms and common misspellings.

SGOPE data have several known strengths and limitations [98-100] as a single data source. Using multiple data sites enhanced the representativeness and validity of the sample and reduced the potential for demographic bias and emotional contagion (18), while mitigating the impact of spam or nongenuine posts through the cleaning process. We do recognize the limitation of only including posts written in English. Although social media use is widespread, those who create posts represent a self-selected subset of users, with only 10%

XSL•FO
RenderX

estimated to be active posters, while 90% read other users' posts without contributing their own comments [101].

Conclusions

The study demonstrated the value of combining NLP and linguistic techniques for analyzing large quantities of unstructured text that can then be used as evidence of improved patient outcomes. In contrast to the current systematic review–based evidence, posters with a wide range of conditions found modafinil effective. The methods we used successfully identified the entities and topics contained in posts. The perceived experiences of causality and effectiveness were identified using 2 different methods. Our study indicates that this NLP- and linguistics-based approach can be used to look beyond the literal meaning of the words in posts, gaining an understanding of how posters assess the effectiveness of a health care intervention and the outcomes they value.

Data Availability

The code generated and analyzed during this study is available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Postfrequency across condition-specific sites derived from parsing of data source URLs.
[PNG File , 56 KB - jmir_v26i1e54321_app1.png]

Multimedia Appendix 2
Latent Dirichlet allocation: 8 topics and 50 iterations.
[PNG File , 57 KB - jmir_v26i1e54321_app2.png]

Multimedia Appendix 3
Latent Dirichlet allocation: 27 topics and 200 iterations.
[PNG File , 69 KB - jmir_v26i1e54321_app3.png]

Multimedia Appendix 4
Latent Dirichlet allocation: 27 topics and 1000 iterations.
[PNG File , 68 KB - jmir_v26i1e54321_app4.png]

Multimedia Appendix 5
Comparison of the top ten words of the sklearn latent Dirichlet allocation and nonnegative matrix factorization topics (n=27).
[DOCX File , 23 KB - jmir_v26i1e54321_app5.docx]

Multimedia Appendix 6
Collocates and frequency of "feel".
[DOCX File , 28 KB - jmir_v26i1e54321_app6.docx]

Multimedia Appendix 7
Cochrane library reviews including modafinil: May 2021.
[XLSX File (Microsoft Excel File), 79 KB - jmir_v26i1e54321_app7.xlsx]

References

1. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
2. Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O. A scoping review of the use of Twitter for public health research. Comput Biol Med 2020 Jul;122:103770 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103770] [Medline: 32502758]
3. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 2020 Jan 28;8(1):e16023 [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
4. Walsh J, Dwumfour C, Cave J, Griffiths F. Spontaneously generated online patient experience data - how and why is it being used in health research: an umbrella scoping review. BMC Med Res Methodol 2022 May 14;22(1):139 [FREE Full text] [doi: 10.1186/s12874-022-01610-z] [Medline: 35562661]

XSL•FO
RenderX

5. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy 2018 Aug;38(8):822-841 [FREE Full text] [doi: 10.1002/phar.2151] [Medline: 29884988]
6. Lafferty NT, Manca A. Perspectives on social media in and as research: a synthetic review. Int Rev Psychiatry 2015 Apr;27(2):85-96. [doi: 10.3109/09540261.2015.1009419] [Medline: 25742363]
7. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of textual health information from Reddit: analysis of chronic diseases with extracted entities and their relations. J Med Internet Res 2019 Jun 13;21(6):e12876 [FREE Full text] [doi: 10.2196/12876] [Medline: 31199327]
8. Suresh S. Patient-generated health data can provide value in clinical care, research settings. American Academy of Pediatrics News. 2020 Jul 01. URL: https://www.aappublications.org/news/2020/07/01/hit070120 [accessed 2021-05-05]
9. Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform 2018 Sep 28;19(5):863-877 [FREE Full text] [doi: 10.1093/bib/bbx010] [Medline: 28334070]
10. Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry 2020 Apr 22;10(1):116 [FREE Full text] [doi: 10.1038/s41398-020-0780-3] [Medline: 32532967]
11. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res 2016 Jun;25(2):86-100 [FREE Full text] [doi: 10.1002/mpr.1481] [Medline: 26184780]
12. Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. Yearb Med Inform 2016 Nov 10(1):224-233 [FREE Full text] [doi: 10.15265/IY-2016-017] [Medline: 27830255]
13. Döbrössy B, Girasek E, Susánszky A, Koncz Z, Győrffy Z, Bognár VK. "Clicks, likes, shares and comments" a systematic review of breast cancer screening discourse in social media. PLoS One 2020 Apr 15;15(4):e0231422 [FREE Full text] [doi: 10.1371/journal.pone.0231422] [Medline: 32294139]
14. Kim SJ, Marsch LA, Hancock JT, Das AK. Scaling up research on drug abuse and addiction through social media big data. J Med Internet Res 2017 Oct 31;19(10):e353 [FREE Full text] [doi: 10.2196/jmir.6426] [Medline: 29089287]
15. Number of social media users worldwide from 2017 to 2028. Statista. URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [accessed 2024-11-25]
16. Social media fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2021-08-09]
17. Bour C, Ahne A, Schmitz S, Perchoux C, Dessenne C, Fagherazzi G. The use of social media for health research purposes: scoping review. J Med Internet Res 2021 May 27;23(5):e25736 [FREE Full text] [doi: 10.2196/25736] [Medline: 34042593]
18. Cesare N, Grant C, Nsoesie EO. Understanding demographic bias and representation in social media health data. In: Proceedings of the Companion Publication of the 10th ACM Conference on Web Science. 2019 Presented at: WebSci '19; June 30-July 3, 2019; Boston, MA. [doi: 10.1145/3328413.3328415]
19. Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. Br J Clin Pharmacol 2015 Oct;80(4):878-888 [FREE Full text] [doi: 10.1111/bcp.12746] [Medline: 26271492]
20. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res 2011 Jan 21;13(1):e6 [FREE Full text] [doi: 10.2196/jmir.1643] [Medline: 21252034]
21. Walsh J, Cave J, Griffiths F. Spontaneously generated online patient experience of modafinil: a qualitative and NLP analysis. Front Digit Health 2021 Feb 17;3:598431 [FREE Full text] [doi: 10.3389/fdgth.2021.598431] [Medline: 34713085]
22. Greenhalgh T. Cultural Contexts of Health: The Use of Narrative Research in the Health Sector. Copenhagen, Denmark: WHO Regional Office for Europe; 2016.
23. Drewniak D, Glässel A, Hodel M, Biller-Andorno N. Risks and benefits of web-based patient narratives: systematic review. J Med Internet Res 2020 Mar 26;22(3):e15772 [FREE Full text] [doi: 10.2196/15772] [Medline: 32213468]
24. McKenna B, Myers MD, Newman M. Social media in qualitative research: challenges and recommendations. Inf Organ 2017 Jun;27(2):87-99 [FREE Full text] [doi: 10.1016/j.infoandorg.2017.03.001]
25. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996 Jan 13;312(7023):71-72 [FREE Full text] [doi: 10.1136/bmj.312.7023.71] [Medline: 8555924]
26. Kones R, Rumana U, Merino J. Exclusion of 'nonRCT evidence' in guidelines for chronic diseases - is it always appropriate? The Look AHEAD study. Curr Med Res Opin 2014 Oct;30(10):2009-2019. [doi: 10.1185/03007995.2014.925438] [Medline: 24841173]
27. Ogilvie D, Adams J, Bauman A, Gregg EW, Panter J, Siegel KR, et al. Using natural experimental studies to guide public health action: turning the evidence-based medicine paradigm on its head. J Epidemiol Community Health 2020 Feb;74(2):203-208 [FREE Full text] [doi: 10.1136/jech-2019-213085] [Medline: 31744848]
28. Schlegl E, Ducournau P, Ruof J. Different weights of the evidence-based medicine triad in regulatory, health technology assessment, and clinical decision making. Pharmaceut Med 2017;31(4):213-216 [FREE Full text] [doi: 10.1007/s40290-017-0197-3] [Medline: 28824273]

XSL•FO
RenderX

29. EBM+: integrating diverse evidence in evidence-based medicine. University of Kent. URL: https://ebmplus.org/ [accessed 2024-11-25]
30. Greenhalgh T. Will COVID-19 be evidence-based medicine's nemesis? PLoS Med 2020 Jun 30;17(6):e1003266 [FREE Full text] [doi: 10.1371/journal.pmed.1003266] [Medline: 32603323]
31. Anjum RL, Copeland S, Rocca E. Conclusion: causehealth recommendations for making causal evidence clinically relevant and informed. In: Rethinking Causality, Complexity and Evidence for the Unique Patient. Cham, Switzerland: Springer; Jun 03, 2020.
32. Real-world evidence. U.S. Food and Drug Administration. 2024 Sep 19. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2024-11-15]
33. Schilsky RL. Finding the evidence in real-world evidence: moving from data to information to knowledge. J Am Coll Surg 2017 Jan;224(1):1-7. [doi: 10.1016/j.jamcollsurg.2016.10.025] [Medline: 27989954]
34. Miani C, Robin E, Horvath V, Manville C, Cave J, Chataway J. Health and healthcare: assessing the real world data policy landscape in Europe. Rand Health Q 2014 Jun 01;4(2):15 [FREE Full text] [Medline: 28083344]
35. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ Digit Med 2020 May 11;3:67 [FREE Full text] [doi: 10.1038/s41746-020-0277-8] [Medline: 32411828]
36. Bender E, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual Event. [doi: 10.1145/3442188.3445922]
37. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv Preprint posted online on June 5, 2019 [FREE Full text] [doi: 10.48550/arXiv.1906.02243]
38. Lee J. Writing linguistic rules for natural language processing. Medium. 2019 Nov 28. URL: https://towardsdatascience.com/linguistic-rule-writing-for-nlp-ml-64d9af824ee8 [accessed 2020-04-28]
39. Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.463]
40. Abdellaoui R, Foulquié P, Texier N, Faviez C, Burgun A, Schück S. Detection of cases of noncompliance to drug treatment in patient forum posts: topic model approach. J Med Internet Res 2018 Mar 14;20(3):e85 [FREE Full text] [doi: 10.2196/jmir.9222] [Medline: 29540337]
41. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, et al. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun Methods Meas 2018 Feb 16;12(2-3):93-118 [FREE Full text] [doi: 10.1080/19312458.2018.1430754]
42. Baumer EP, Mimno D, Guha S, Quan E, Gay GK. Comparing grounded theory and topic modeling: extreme divergence or unlikely convergence? J Assoc Inf Sci Technol 2017 Apr 28;68(6):1397-1410 [FREE Full text] [doi: 10.1002/asi.23786]
43. Spasic I, Button K. Patient triage by topic modeling of referral letters: feasibility study. JMIR Med Inform 2020 Nov 06;8(11):e21252 [FREE Full text] [doi: 10.2196/21252] [Medline: 33155985]
44. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 01;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
45. Kerry R. Causal dispositionalism and evidence based healthcare. In: Anjum RL, Copeland S, Rocca E, editors. Rethinking Causality, Complexity and Evidence for the Unique Patient. Cham, Switzerland: Springer; 2020.
46. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med 2018 Aug;210:2-21 [FREE Full text] [doi: 10.1016/j.socscimed.2017.12.005] [Medline: 29331519]
47. Anjum RL. Dispositions and the unique patient. In: Anjum RL, Copeland S, Rocca E, editors. Rethinking Causality, Complexity and Evidence for the Unique Patient. Cham, Switzerland: Springer; Jun 03, 2020.
48. Edwards R. Living with complexity and big data. Uppsala Monitoring Centre. URL: https://view.publitas.com/uppsala-monitoring-centre/uppsala-reports-78/page/28 [accessed 2018-11-06]
49. Neeleman A, van de Koot H. The linguistic expression of causation. In: The Theta System: Argument Structure at the Interface. Oxford, UK: Oxford University Press; 2012.
50. Williamson J. Establishing causal claims in medicine. Int Studies Philos Sci 2019 Jun 27;32(1):33-61 [FREE Full text] [doi: 10.1080/02698595.2019.1630927]
51. Greenhalgh T, Snow R, Ryan S, Rees S, Salisbury H. Six 'biases' against patients and carers in evidence-based medicine. BMC Med 2015 Sep 01;13:200 [FREE Full text] [doi: 10.1186/s12916-015-0437-x] [Medline: 26324223]
52. Davies N. Social media: the voice of the patient. Reuters Events. 2015 Jul 27. URL: https://www.reutersevents.com/pharma/commercial/social-media-voice-patient#.VbX9zTjUk84.linkedin [accessed 2024-11-29]
53. van Rossum G. Python reference manual. Centrum Wiskunde & Informatica. 1995. URL: https://ir.cwi.nl/pub/5008 [accessed 2024-11-15]
54. jupyterlab. GitHub. URL: https://github.com/jupyterlab/jupyterlab [accessed 2024-05-03]

XSL•FO
RenderX

55. Albalawi R, Yeap TH, Benyoucef M. Using topic modeling methods for short-text data: a comparative analysis. Front Artif Intell 2020;3:42 [FREE Full text] [doi: 10.3389/frai.2020.00042] [Medline: 33733159]
56. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar 02;28(1):e100262 [FREE Full text] [doi: 10.1136/bmjhci-2020-100262] [Medline: 33653690]
57. Chen Y, Zhang H, Liu R, Ye Z, Lin J. Experimental explorations on short text topic mining between LDA and NMF based schemes. Knowl Based Syst 2019 Jan;163:1-13 [FREE Full text] [doi: 10.1016/j.knosys.2018.08.011]
58. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021 Feb 10;23(2):e25431 [FREE Full text] [doi: 10.2196/25431] [Medline: 33497352]
59. Suri P, Roy NR. Comparison between LDA and NMF for event-detection from large text stream data. In: Proceedings of the 3rd International Conference on Computational Intelligence & Communication Technology (CICT). 2017 Presented at: CICT 2017; February 9-10, 2017; Ghaziabad, India. [doi: 10.1109/ciact.2017.7977281]
60. Birks D, Coleman A, Jackson D. Unsupervised identification of crime problems from police free-text data. Crime Sci 2020 Oct 07;9(1):18 [FREE Full text] [doi: 10.1186/s40163-020-00127-4]
61. Bakharia A. On the equivalence of inductive content analysis and topic modeling. In: Proceedings of the First International Conference on Advances in Quantitative Ethnography. 2019 Presented at: ICQE 2019; October 20-22, 2019; Madison, WI URL: https://link.springer.com/chapter/10.1007/978-3-030-33232-7_25 [doi: 10.1007/978-3-030-33232-7_25]
62. Rehurek R. gensim: python framework for vector space modelling. Machine Learning Open Source Software. 2010 Sep 7. URL: https://mloss.org/revision/view/546/ [accessed 2020-06-23]
63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825-2830.
64. Angelov D. Top2Vec: distributed representations of topics. arXiv Preprint posted online on August 19, 2020 [FREE Full text]
65. A guide on word embeddings in NLP. Turing. URL: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp [accessed 2024-05-03]
66. Sievert C, Shirley KE. LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014 Presented at: ACL 2014; June 27, 2014; Baltimore, MD URL: https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf [doi: 10.3115/v1/w14-3110]
67. Loria S. TextBlob: simplified text processing. TextBlob. URL: https://textblob.readthedocs.io/en/dev/index.html [accessed 2021-11-15]
68. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May 16;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]
69. Bonta V, Kumaresh N, Janardhan N. A comprehensive study on lexicon based approaches for sentiment analysis. Asian J Comput Sci Technol 2019;8(S2):1-6 [FREE Full text] [doi: 10.51983/ajcst-2019.8.S2.2037]
70. Soma J. Comparing sentiment analysis tools. Data Science for Journalism. URL: https://investigate.ai/investigating-sentiment-analysis/comparing-sentiment-analysis-tools/ [accessed 2021-06-30]
71. Caren N. Word lists and sentiment analysis. Neal Caren. 2019 May 1. URL: https://nealcaren.org/lessons/wordlists/ [accessed 2021-07-01]
72. Sketch Engine. URL: https://www.sketchengine.eu/ [accessed 2024-11-25]
73. enTenTen: corpus of the English web. Sketch Engine. URL: https://www.sketchengine.eu/ententen-english-corpus/ [accessed 2018-02-27]
74. What is an N-Gram? MathWorks. URL: https://uk.mathworks.com/discovery/ngram.html [accessed 2024-05-03]
75. Alper BS, Fedorowicz Z, van Zuuren EJ. Limitations in conduct and reporting of cochrane reviews rarely inhibit the determination of the validity of evidence for clinical decision-making. J Evid Based Med 2015 Aug 21;8(3):154-160. [doi: 10.1111/jebm.12161] [Medline: 26107648]
76. Cochrane library homepage. Cochrane Library. URL: https://www.cochranelibrary.com/ [accessed 2024-11-26]
77. Ruthirakuhan MT, Herrmann N, Abraham EH, Chan S, Lanctôt KL. Pharmacological interventions for apathy in Alzheimer's disease. Cochrane Database Syst Rev 2018 May 04;5(5):CD012197 [FREE Full text] [doi: 10.1002/14651858.CD012197.pub2] [Medline: 29727467]
78. Castells X, Cunill R, Pérez-Mañá C, Vidal X, Capellà D. Psychostimulant drugs for cocaine dependence. Cochrane Database Syst Rev 2016 Sep 27;9(9):CD007380 [FREE Full text] [doi: 10.1002/14651858.CD007380.pub4] [Medline: 27670244]
79. Day J, Yust-Katz S, Cachia D, Wefel J, Tremont Lukats IW, Bulbeck H, et al. Interventions for the management of fatigue in adults with a primary brain tumour. Cochrane Database Syst Rev 2022 Sep 12;9(9):CD011376 [FREE Full text] [doi: 10.1002/14651858.CD011376.pub3] [Medline: 36094728]
80. Dougall D, Poole N, Agrawal N. Pharmacotherapy for chronic cognitive impairment in traumatic brain injury. Cochrane Database Syst Rev 2015 Dec 01;2015(12):CD009221 [FREE Full text] [doi: 10.1002/14651858.CD009221.pub2] [Medline: 26624881]

XSL•FO
RenderX

81. Elbers RG, Verhoef J, van Wegen EE, Berendse HW, Kwakkel G. Interventions for fatigue in Parkinson's disease. Cochrane Database Syst Rev 2015 Oct 08;2015(10):CD010925 [FREE Full text] [doi: 10.1002/14651858.CD010925.pub2] [Medline: 26447539]
82. Mücke M, Mochamat, Cuhls H, Peuckmann-Post V, Minton O, Stone P, et al. Pharmacological treatments for fatigue associated with palliative care. Cochrane Database Syst Rev 2015 May 30;2015(5):CD006788 [FREE Full text] [doi: 10.1002/14651858.CD006788.pub3] [Medline: 26026155]
83. Koopman FS, Beelen A, Gilhus NE, de Visser M, Nollet F. Treatment for postpolio syndrome. Cochrane Database Syst Rev 2015 May 18;2015(5):CD007818. [doi: 10.1002/14651858.CD007818.pub3] [Medline: 25984923]
84. Day J, Zienius K, Gehring K, Grosshans D, Taphoorn M, Grant R, et al. Interventions for preventing and ameliorating cognitive deficits in adults treated with cranial irradiation. Cochrane Database Syst Rev 2014 Dec 18;2014(12):CD011335 [FREE Full text] [doi: 10.1002/14651858.CD011335.pub2] [Medline: 25519950]
85. Pérez-Mañá C, Castells X, Torrens M, Capellà D, Farre M. Efficacy of psychostimulant drugs for amphetamine abuse or dependence. Cochrane Database Syst Rev 2013 Sep 02;2013(9):CD009695. [doi: 10.1002/14651858.CD009695.pub2] [Medline: 23996457]
86. Ortiz-Orendain J, Covarrubias-Castillo SA, Vazquez-Alvarez AO, Castiello-de Obeso S, Arias Quiñones GE, Seegers M, et al. Modafinil for people with schizophrenia or related disorders. Cochrane Database Syst Rev 2019 Dec 12;12(12):CD008661 [FREE Full text] [doi: 10.1002/14651858.CD008661.pub2] [Medline: 31828767]
87. Castells X, Blanco-Silvente L, Cunill R. Amphetamines for attention deficit hyperactivity disorder (ADHD) in adults. Cochrane Database Syst Rev 2018 Aug 09;8(8):CD007813 [FREE Full text] [doi: 10.1002/14651858.CD007813.pub3] [Medline: 30091808]
88. Gibbons C, Pagnini F, Friede T, Young CA. Treatment of fatigue in amyotrophic lateral sclerosis/motor neuron disease. Cochrane Database Syst Rev 2018 Jan 02;1(1):CD011005 [FREE Full text] [doi: 10.1002/14651858.CD011005.pub2] [Medline: 29293261]
89. Liira J, Verbeek JH, Costa G, Driscoll TR, Sallinen M, Isotalo LK, et al. Pharmacological interventions for sleepiness and sleep disturbances caused by shift work. Cochrane Database Syst Rev 2014 Aug 12;2014(8):CD009776 [FREE Full text] [doi: 10.1002/14651858.CD009776.pub2] [Medline: 25113164]
90. Ker K, Edwards PJ, Felix LM, Blackhall K, Roberts I. Caffeine for the prevention of injuries and errors in shift workers. Cochrane Database Syst Rev 2010 May 12;2010(5):CD008508 [FREE Full text] [doi: 10.1002/14651858.CD008508] [Medline: 20464765]
91. Candy M, Jones L, Williams R, Tookman A, King M. Psychostimulants for depression. Cochrane Database Syst Rev 2008 Apr 16(2):CD006722. [doi: 10.1002/14651858.CD006722.pub2] [Medline: 18425966]
92. Annane D, Moore DH, Barnes PR, Miller RG. Psychostimulants for hypersomnia (excessive daytime sleepiness) in myotonic dystrophy. Cochrane Database Syst Rev 2006 Jul 19;2006(3):CD003218 [FREE Full text] [doi: 10.1002/14651858.CD003218.pub2] [Medline: 16855999]
93. Gonzalez G, Vaculik K, Khalil C, Zektser Y, Arnold C, Almario CV, et al. Using large-scale social media analytics to understand patient perspectives about urinary tract infections: thematic analysis. J Med Internet Res 2022 Jan 25;24(1):e26781 [FREE Full text] [doi: 10.2196/26781] [Medline: 35076404]
94. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022;7:886498 [FREE Full text] [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
95. Isoaho K, Gritsenko D, Mäkelä E. Topic modeling and text analysis for qualitative policy research. Policy Stud J 2019 Jun 19;49(1):300-324 [FREE Full text] [doi: 10.1111/psj.12343]
96. Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S. Collecting and analyzing patient experiences of health care from social media. JMIR Res Protoc 2015 Jul 02;4(3):e78 [FREE Full text] [doi: 10.2196/resprot.3433] [Medline: 26137885]
97. Dirkson A, Verberne S, Kraaij W. Lexical normalization of user-generated medical text. In: Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. 2019 Presented at: SMM4H@ACL 2019; August 2, 2019; Florence, Italy URL: https://arxiv.org/pdf/1812.04265#page=5
98. Dalmer NK. Questioning reliability assessments of health information on social media. J Med Libr Assoc 2017 Jan 17;105(1):61-68 [FREE Full text] [doi: 10.5195/jmla.2017.108] [Medline: 28096748]
99. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
100. Staccini P, Fernandez-Luque L. Secondary use of recorded or self-expressed personal data: consumer health informatics and education in the era of social media and health apps. Yearb Med Inform 2017 Sep 11;26(01):172-177. [doi: 10.15265/iy-2017-037]
101. van Mierlo T. The 1% rule in four digital health social networks: an observational study. J Med Internet Res 2014 Feb 04;16(2):e33 [FREE Full text] [doi: 10.2196/jmir.2966] [Medline: 24496109]

XSL•FO
RenderX

Abbreviations

LDA: latent Dirichlet allocation
NLP: natural language processing
NMF: nonnegative matrix factorization
POS: part-of-speech
RCT: randomized controlled trial
RWD: real-world data
SGOPE: spontaneously generated online patient experience
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by A Mavragani; submitted 06.11.23; peer-reviewed by M Chatzimina, M Mahmic Kaknjo; comments to author 22.03.24; revised version received 19.06.24; accepted 27.09.24; published 11.12.24.

Please cite as:
Walsh J, Cave J, Griffiths F
Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences
J Med Internet Res 2024;26:e54321
URL: https://www.jmir.org/2024/1/e54321
doi:10.2196/54321
PMID:

©Julia Walsh, Jonathan Cave, Frances Griffiths. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Relationship Between Public Social Media Accounts, Adolescent Mental Health, and Parental Guidance in England: Large Cross-Sectional School Survey Study

Wakithi Siza Mabaso[1,2], MBChB, MSc; Sascha Hein[3,4], PhD; Gabriela Pavarini[5,6], PhD; The OxWell Study Team[7]; Mina Fazel[1], DM

[1]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[2]Department of Clinical Child and Adolescent Psychology, University of Giessen, Giessen, Germany
[3]Department of Education & Psychology, Freie Universität Berlin, Berlin, Germany
[4]Child Study Center, Yale University, New Haven, CT, United States
[5]Ethox Centre, Oxford Population Health, University of Oxford, Oxford, United Kingdom
[6]Department of Social Policy and Intervention, University of Oxford, Oxford, United Kingdom
[7]see Acknowledgments

Corresponding Author:
Wakithi Siza Mabaso, MBChB, MSc
Department of Psychiatry
University of Oxford
University Department of Psychiatry
Warneford Hospital
Oxford, OX3 7JX
United Kingdom
Phone: 44 01865 618200
Email: wakithi.mabaso@trinity.ox.ac.uk

Abstract

Background: Although associations between social media use and adolescent mental health have been described, more information is needed on the potential components characterizing this complex exposure, in particular, those related to maintaining a public social media account.

Objective: This study aims to investigate the association between having a public social media account and anxiety and depression in school-going adolescents.

Methods: Overall, 80 secondary schools and further education colleges in England were sampled using a cross-sectional web-based survey as part of the 2023 OxWell Student Survey. Social media exposure was categorized among the adolescents as having a public social media account versus not having a public social media account. The risk of clinical anxiety and depression was determined using the Revised Child Anxiety and Depression Scale-11. Adolescents self-reported the content and platforms accessed in the previous 24 hours. Associations between having a public social media account and symptoms of anxiety and depression were assessed using logistic regression controlling for age, sex, the experience of being bullied, parental guidance of online behavior (describing perceived parental approaches to adolescents' online activity), the proportion of close friendships engaged with online, poverty status, and placement in statutory care. Age, sex, and parental guidance of online behavior were assessed for primary association effect modification.

Results: Data collected from 16,655 adolescents (aged 11-18 y) were analyzed. Of these 16,655 adolescents, 6734 (40.43%) had a public social media account, while 9921 (59.57%) either had a private social media account or no social media account. Moreover, 32.6% (5429/16,655) of the adolescents screened positive for symptoms of anxiety and depression. Those with a public social media account had higher odds of anxiety and depression (odds ratio [OR] 1.41, 95% CI 1.32-1.50) than those without a public social media account in an unadjusted and fully adjusted model (OR 1.39, 95% CI 1.29-1.49). Adolescents reporting active parental guidance had lower odds of anxiety and depression (OR 0.85, 95% CI 0.75-0.93) than those reporting no parental guidance, and these parental approaches to online behaviors significantly modified the association between having a public social media account and symptoms of anxiety and depression (P=.004; χ^2_2=11.1).

XSL•FO
RenderX

Conclusions: Our OxWell study findings suggest a potential mental health risk for adolescents with a public social media account. We show evidence indicating some protection from anxiety and depression among adolescents who do not have a public social media account and those reporting some form of parental guidance of their online behavior. This was pronounced in adolescents reporting active parental guidance compared to stricter regulatory approaches or no guidance at all. The specific roles that social media account choices and parental guidance of online behavior may play in supporting the mental health of adolescents are highlighted for further investigation.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-052717

(J Med Internet Res 2024;26:e57154) doi:10.2196/57154

KEYWORDS

social media; adolescent health; privacy; parental guidance; mood disorders; adolescent; anxiety; depression; cross-sectional; mental health; public; account; school-going; school; England; survey; logistic regression; observational

Introduction

Background

Associations between social media use and adolescent mental health have been described; however, new characterizations of social media use are required to advance a nuanced understanding of this complex exposure, prompted by concerns over adolescents' susceptibilities on social media platforms (SMPs). Existing findings draw on more than a decade of broad research into time spent on social media and online engagement patterns, including both active and passive social media use [1-3]. The rapidly changing nature of social media related behaviors exacerbates the difficulty of inquiry in this field and necessitates focused social media research questions to try to extend and explain the findings from existing studies to explore this complexity in greater depth.

One such area broadly concerns online privacy behaviors on social media through a simple characterization: having a public social media account versus not having a public social media account (ie, having a private social media account). Social media privacy behavior refers to actions or settings that allow users to control whom they communicate and share information with directly [4]. Adolescents can express such behaviors or phenotypes in the following ways: having a public social media account, not having a public social media account (ie, having a private social media account), having both a public social media account and a private one, and having no social media account at all [5]. In practice, social media privacy settings extend further to a set of complex parameters, including who can contact a user, customizable information disclosure, privacy concerns, the transparency of users' online behavior to other users in their network, and the accessibility of their information to organizations and platforms through data-sharing agreements. Exercising their social media privacy by this behavioral definition facilitates young people's expectations of SMPs and confers subjective control over the information they share and access. Understanding the differences between having a public social media account and not having one offers insights into the potential risks on adolescent mental health outcomes conferred by social media exposure, where private social media accounts require a follow request and permissions to access (and interact with) the user's profile and content, while publicly available social media accounts do not [5].

Prior Findings

Overview

Broadly, cross-sectional studies on social media and mental health outcomes have predominantly examined the duration of time spent online, finding significant associations between spending ≥2 hours on SMPs a day and depressive symptoms in adolescent girls [6-8], with discrete windows of sensitivity emerging early in boys (aged 14 y and 19 y) and in girls (aged 13-15 y and 19 y) [9]. Longitudinal studies on social media use and adolescent mental health have shown conflicting results. The UK Household Longitudinal Study of 3228 adolescents looked at the association between the duration of time spent on SMPs at ages 12 to 13 years and predicted mental health problems 2 years later at ages 14 to 15 years using the Strengths and Difficulties Questionnaire, a brief screening measure of behavioral and emotional problems, and found no association between time spent on SMPs and mental health problems [10]. A Swedish study of 3502 students found that adolescents who spent more time on social media reported higher levels of mental health problems measured using the Strengths and Difficulties Questionnaire [11]. In contrast to these findings, a US study of 388 adolescents using live mobile device use measurements and the Positive and Negative Affect Schedule for Children showed that adolescents' daily duration of time spent on entertainment content was associated with less same-day worry and that sending more text or online messages per day was associated with lower depression scores [12].

The Stanford Social Media Lab conducted a systematic review and meta-analysis of 226 peer-reviewed papers involving 257,728 participants; the researchers demonstrated an overall null effect, indicating that social media use was not associated with well-being in bidirectional analysis [2]. However, analysis looking exclusively at negative indicators of well-being showed a small significant association between greater social media use and greater anxiety and depression symptoms. Analysis looking only at positive indicators of well-being also showed a small significant association, this time between greater social media use and greater social well-being symptoms. The presence of small significant associations in both negative and positive well-being indicators explained their aggregated null effect [2], and cross-lagged effects analysis showed a solitary, significant, unidirectional association between positive psychological well-being and lower social media use [2]. These findings hint

XSL•FO
RenderX

at underlying drivers of heterogeneity, prompting the investigation of adolescents' social media use and their potential effect modifiers, including parental approaches to adolescents' online behaviors measured by adolescents' self-reported perceptions.

Parental Guidance of Online Behavior

Parental approaches to the guidance of young people's online activities are suspected to predict positive mental health because parents who demonstrated awareness of their children's activities significantly halved their children's odds of reporting negative moods and feelings [13]. The type of web-related parental guidance reported by adolescents significantly influences their online privacy behaviors [14,15], and established social media privacy behaviors among young people are shown to be even more influential than their personality traits in predicting future privacy-related choices [15].

Social Media Privacy

Regarding social media privacy, Barsova et al [8], demonstrated that participants who attempted to exercise greater control over the social media material they were exposed to by using tools such as "unfollow" and managing notification settings to control the content in their feeds showed significantly lower levels of anxiety, depression, and stress. These actions resembled adolescents' greater concern for their personal privacy behaviors online [7,13,15]. These associations illuminate gaps for further work strengthening the characterization and description of associations between social media privacy and mental health outcomes. The findings by Gruzd and Hernández-García [5] demonstrated distinct patterns between online privacy concerns and the protective posting behaviors of a subset of individuals who reported having both a public social media account and a private one but showed no intraindividual differences in protective posting behaviors across a given individual's public and private accounts. Few studies have characterized interindividual differences between people with a public social media account and those without one. This highlights the importance of research investigating social media account types and their influence on mental health as a primary level of distinction between appreciably nuanced privacy behaviors and in keeping with a complex understanding of social media use behaviors [8].

Relevance

With significant heterogeneity in the literature, it is suggested that newer characterizations of social media use, including social media privacy, SMP use, and content consumption, are useful toward a greater understanding of the specific behaviors from which effects on mental health may be expected [8]. This is important because of the rapid integration of social media with normative development in adolescents, who, in some aspects of their development, need to be considered a vulnerable group [16]. This developmental group vulnerability calls for the consideration of an ethical framework of accountability as it relates to the SMPs' instrumentalization of content and users' data against arising conflicts with adolescents' social media use patterns and observed associations with their mental health outcomes per the goals of the Lancet Commission on Adolescent Health and Wellbeing [16].

Using data from the 2023 OxWell Student Survey, this study aims to explore the relationship between social media account privacy (ie, whether an adolescent has a public social media account versus whether they do not have a public social media account) as a behavioral proxy and their mental health outcomes across this behavior in a sample of English school-going adolescents. We ask whether there is a difference in the symptoms of anxiety and depression between adolescents who have a publicly available social media account and those who do not have a publicly available social media account (ie, have a private account). Next, we describe the distribution of SMPs and content accessed by adolescents, and finally, we discuss some of the ethical considerations of SMPs and their accountability surrounding adolescents as a vulnerable group.

Hypotheses

The following hypotheses are examined:

- Small but significant associations are described between positive psychological well-being and decreased social media use, a potential proxy for agency over one's social media exposure, which may extend to privacy behaviors [2]. Therefore, we hypothesize (H1) that the odds of anxiety and depression in adolescents with a publicly available social media account are higher than the odds of anxiety and depression in adolescents without a publicly available social media account.
- Adolescent privacy behaviors are shown to be influenced by an open discussion with adults, a component of parental guidance of online behavior [14], and prior work looking at age and sex differences in mental illness describes higher rates of anxiety and depression in young adolescent girls compared to boys [17]. Therefore, we hypothesize (H2) that the strength of the association between social media account type and anxiety and depression is modified by age, sex, and parental guidance of online behavior [14,17].

Methods

Study Characteristics

The OxWell Student Survey is a repeated cross-sectional survey of schools and further education colleges (FECs) in England, asking students various questions relating to their mental health and well-being, life experiences, and behaviors [18]. In 2023, it had 2 age-appropriate versions (divided into English school years 5 to 6 [primary] and 7 to 13 [secondary], collectively covering ages 9 to 18 years). This observational study used only secondary school and FEC data because the questions of interest were only asked to the older students. Participants came from 80 secondary schools and FECs, primarily in 4 regions in England: Oxfordshire, Berkshire, Liverpool, and Milton Keynes. The OxWell Student Survey 2023 (administered in February-March 2023) collected 42,215 valid consenting responses from 43,735 survey log-ins. Of the 42,215 valid consenting responses, 32,965 (78.09%) were from students in school years 7 to 13 (ages 11-18 y) in 80 secondary schools and FECs. The proportion of students who, upon logging in to the

XSL•FO
RenderX

survey, declined to consent to participate was 3.48% (1520/43,735). Information on school-recorded opt-outs effected by nonconsenting parents was not shared with the study team in keeping with the study's data protection protocols and ethics guidelines [18].

Participant Sampling

The final study sample is defined as 16,655 adolescents in secondary schools and FECs aged 11 to 18 years responding to social media use questions. Adolescents were excluded from the sample if they spent <10 minutes on the survey as a minimum engagement threshold, ensuring validity in the integrity of the responses (Figure 1).

Figure 1. Flow diagram of participant eligibility.

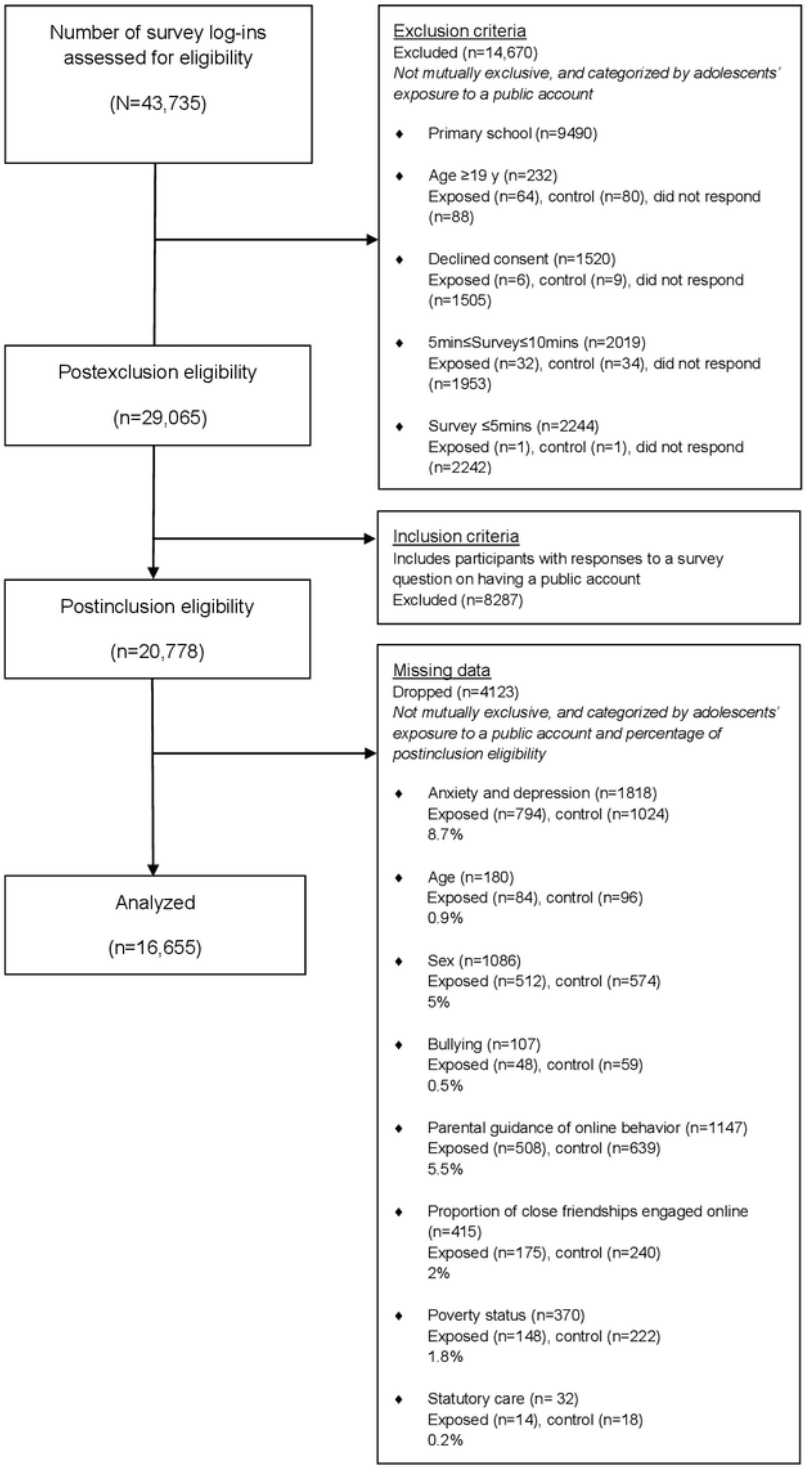

XSL•FO
RenderX

Measures

Primary Exposure: Public Social Media Account

The primary exposure describes social media account privacy behavior in response to the question, "Do you have a publicly available social media account?" It is binarized into "having a publicly available social media account" defined as any account that does not require a follow request, friend request, or permission to access the account's information, and "not having a publicly available social media account," which includes "having a private social media account" requiring a follow request, friend request, or permission to access the account's information [8].

Primary Outcome: Anxiety and Depression

The primary outcome in adolescents is "anxiety and depression," a state of mental illness defined as being above the threshold for risk of clinical anxiety and depression using the Revised Child Anxiety and Depression Scale-11 (RCADS-11). The RCADS-11 includes 4-point scale items on depression (5 items) and anxiety (6 items), such as "I feel worthless" and "All of a sudden, I feel really scared for no reason at all." A 2-item symptom impact score includes, "How much do these difficulties upset or distress you?" and improves screening sensitivity to 84% and specificity to 72%. The item scores align with the criteria of the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. They are summed to produce a combined, internalizing symptom score for anxiety and depression that is binarized and interpreted at cutoffs of ≥14 for boys and ≥18 for girls [19].

Descriptive Variables: Social Media Use

Descriptive information on the distribution of SMPs and social media content accessed by adolescents was collected by asking, "In the last 24 hours, which online platforms and content have you accessed?" Responses to these questions provide a snapshot of the variability in adolescent social media use; platforms included YouTube, WhatsApp, Snapchat, TikTok, Instagram, Pinterest, BeReal, Discord, Twitter, Facebook, Reddit, Telegram, Omegle, and Signal, as well as "other," and "none of these" options. Content response options included entertainment, humor, celebrities and influencers, family and friends, lifestyle, health and fitness, support and wellness, beauty and fashion, education, and forums, as well as "other" [8].

Covariates

Established Confounders

The International Alliance of Mental Health Research Funders recommends the use of RCADS-11 for the measurement of anxiety and depression in adolescents, with age and sex as demographic confounders [10,17,20]. In addition, the following a priori confounders are included: the experience of bullying [13], the perception of different parental guidance approaches to online behavior [14,15], the proportion of close friendships engaged with online [21-23], poverty status and exposure to placement in statutory care [13,24].

Bullying, especially online-mediated bullying (cyberbullying), is recognized as being particularly impactful on the mental health of adolescents, with girls being disproportionately affected [25]. Together with income inequality [25], as captured by poverty status in this study, these covariates contribute to significantly greater odds of negative moods and feelings in those who are bullied very often compared to those bullied less often [25] and poorer mental health symptoms overall [25,26]. Close friendships and the strength of these relationships are demonstrated to significantly predict adolescent mental health [21-23]. Greater proportional time spent engaging in these close friendships is demonstrated to impact young people's mental health positively, with lower odds of negative moods and low self-esteem [13]. Parental guidance of online behavior captures adolescents' perception of parental control or approaches to their online activities and their level of emotional closeness to their parents, which is shown to be a strong predictor of positive mental health [13]. Exposure to childhood maltreatment, expectedly, worsens mental health outcomes in young people and is represented in this analysis by adolescents who were exposed to being placed in statutory care [24].

Age and sex will be used in the analysis of a minimally adjusted model, and a priori confounders will be included in a fully adjusted model. Age, sex, and parental guidance of online behavior will be explored for effect modification of the relationship between social media privacy and anxiety and depression [14,15].

Confounder Categorization

The experience of bullying is categorized into 5 groups ranging from *never* to *every day*. Parental guidance of online behavior is categorized into 3 groups: *none*, *active guidance*, and *regulation*. *Active guidance* describes parents who "Encourage you to explore and learn things on the internet," "Talk to you about what you do online," "Explain why some websites are appropriate or inappropriate," or "Help you when something bothers you on the internet." *Regulation* describes parents who "Restrict your internet use," "Log in to access your online social media accounts (eg, Facebook, Twitter, Instagram)" or "Monitor your internet use." The proportion of close friendships primarily engaged with online is grouped into 3 categories in response to the question, "Out of all the individuals you would identify as close friends, how many do you primarily meet with online?": *none*, *a few or some*, or *most or all*. Poverty status is binarized from responses to the question, "At home, I go to bed hungry because there is not enough food in the house." Placement in statutory care is binarized from responses to the question, "Are you a child in care, a looked-after child, or fostered?"

Statistical Analysis

Statistical Modeling

Data distributions are assessed for normality, and univariable association tests are performed between the primary exposure and outcome variables using crude logistic regression likelihood ratio tests (LRTs). Chi-square tests and ANOVA are used to independently test the primary exposure and outcome variables against a priori confounders and assess confounder collinearity. A multivariable logistic regression model is built to assess the association between having a public social media account and anxiety and depression, with stepwise inclusion of the confounders. LRTs are used to test categorical variables for

XSL•FO
RenderX

departure from linearity, conduct sequential model adjustment for the strength of confounding, and assess for effect modification. Statistical significance is set at $P<.05$ using 95% CIs, and analyses are performed using Stata software (version 17.0; StataCorp LLC).

Sensitivity Analysis

A small proportion of adolescents (204/16,655, 1.22%) who report not using any survey-listed SMPs or any "other" SMPs not listed in the survey in response to a question that asks them to indicate which SMPs they had accessed in the past 24 hours are excluded from the full sample in a sensitivity analysis to determine whether the information they represent biases the observed associations between having a public social media account and anxiety and depression. This sensitivity analysis aims to achieve the closest comparison between adolescents with a public social media account and those without a public social media account who had a private social media account.

Power

Given that the estimated prevalence of probable mental disorders in UK adolescents ranges from 18% to 22% [27], to detect a 15% relative difference in the odds of depression and anxiety with a 2-tailed significance level of 95% and power of 80%, a total of 10,500 participants would be required for statistical analysis. Missing values independently constitute <10% of the selected covariates (Figure 1). Complete case analysis is performed to maintain statistical integrity accounted for by sufficient study power.

Ethical Considerations

Ethics Approval

The study was approved by the University of Oxford Research Ethics Committee (R62366/RE0014). The research team provided all schools that signed up with detailed ethics procedures for the study, information for parents or guardians and pupils, and school-specific log-in IDs and passwords for their pupils.

Informed Consent

Participation in the OxWell survey was voluntary. Students aged <16 years were asked to assent to participation after their parents were given an opportunity to opt their child out of the study; those aged ≥16 years could consent to participation. Participants did not receive any incentives to take part in the study. Consent procedures are fully described in the protocol [18].

Pupils were informed and frequently reminded during the survey that they could stop at any time, leave questions blank, and move on to the next question if they did not wish to answer. At the end of the survey, pupils were advised what they could do if they felt worried or upset, including seeking support from available school staff. Information on appropriate online resources supporting adolescent mental health was also provided. If a survey response raised concern about a participant's safety, the lack of identifiers in the survey meant that we could not positively identify pupils to provide direct referral to appropriate support.

Privacy and Confidentiality

Nonidentifiable data were collected without students' names, addresses, or date of birth information. No unique individual log-ins were used, and IP addresses were not retained. The survey provider used encryption when transferring the data to the research team, and participants were instructed and reminded to close their browser window when they were finished.

Results

Participant Characteristics

The characteristics of 16,655 study participants are summarized and grouped by social media account (Table 1). Female individuals comprised 55.48% (9241/16,655) of all adolescents. The participants' mean age was 13.8 (SD 1.84; range 11-18) years. Approximately two-thirds of the participants (10,860/16,655, 65.21%) fell into the young adolescent category (age 10-14 y), and the percentage of adolescents with a publicly available social media account was consistent across the age range with no statistically significant differences across the 2 age groupings ($P=.84$; Table 2). Of the 16,655 adolescents, 6734 (40.43%) reported having a public social media account. This group differed significantly from those without a public social media account across the remaining covariates, including age ($P<.001$; Table 3). Of the 16,655 adolescents, 5429 (32.6%) scored above the RCADS-11 threshold for risk of clinical anxiety and depression. A third of the adolescents (5511/16,655, 33.09%) reported experiencing *no parental guidance* over their online activities, while two-thirds (11,144/16,655, 66.91%) reported experiencing *active guidance* or *regulation*. One-fifth of the adolescents (3356/16,655, 20.15%) experienced all or most of their closest friendships online, while 35% (5836/16,655) reported experiencing none of their closest friendships online (Table 1).

XSL•FO
RenderX

Table 1. Participant characteristics by social media privacy.

Variables	Publicly available social media account		Total (n=16,655)	*P* value
	Yes (n=6734)	No (n=9921)		
Continuous variables				
Age (y), median (IQR)	14.0 (12.0-15.0)	14.0 (12.0-15.0)	__[a]	—
Age (y), mean (SD; range)	—	—	13.8 (1.84; 11.0-18.0)	—
Categorical variables, n (%)				
Anxiety and depression (RCADS-11) [b]				<.001
Yes	2499 (37.11)	2930 (29.53)	5429 (32.6)	
No	4235 (62.89)	6991 (70.47)	11 226 (67.4)	
Sex				<.001
Female	3293 (48.9)	5948 (59.95)	9241 (55.48)	
Male	3441 (51.1)	3973 (40.05)	7414 (44.52)	
Bullying				<.001
Negligible	5291 (78.57)	8240 (83.06)	13,531 (81.24)	
Occasionally	716 (10.63)	882 (8.89)	1598 (9.59)	
Weekly	231 (3.43)	281 (2.83)	512 (3.07)	
Frequently	318 (4.72)	366 (3.69)	684 (4.11)	
Daily	178 (2.64)	152 (1.53)	330 (1.98)	
Parental guidance of online behavior [c]				<.001
None	2508 (37.24)	3003 (30.27)	5511 (33.09)	
Active guidance	2338 (34.72)	3241 (32.67)	5579 (33.5)	
Regulation	1888 (28.04)	3677 (37.06)	5565 (33.41)	
Proportion of close friendships engaged with online [d]				<.001
None	1892 (28.1)	3944 (39.75)	5836 (35.04)	
A few or some	3244 (48.17)	4219 (42.53)	7463 (44.81)	
Most or all	1598 (23.73)	1758 (17.72)	3356 (20.15)	
Poverty status				<.001
Yes	360 (5.35)	248 (2.5)	608 (3.65)	
No	6374 (94.65)	9673 (97.5)	16,047 (96.35)	
Statutory care				<.001
Yes	470 (6.98)	462 (4.66)	932 (5.6)	
No	6264 (93.02)	9459 (95.34)	15,723 (94.4)	

[a]Not applicable.

[b]RCADS-11: Revised Child Anxiety and Depression Scale-11 (symptom impact score: 0-39; positive anxiety and depression screen: ≥18 for girls, ≥14 for boys).

[c]Adolescent-reported parental approaches to their online activity.

[d]Adolescent-reported proportion of existing close friendships engaged with online.

XSL•FO
RenderX

Table 2. Percentage of adolescents with a publicly available social media account by age group and age.

	Publicly available social media account		Total (n=16,655), n (%)	*P* value
	Yes (n=6734), n (%)[a]	No (n=9921), n (%)[a]		
Age group (y)				.84
Young adolescents (10-14)	4397 (40.49)[b]	6463 (59.51)	10,860 (65.21)	
Middle adolescents (15-19)	2337 (40.33)	3458 (59.67)	5795 (34.79)	
Age (y)				<.001
11	507 (35.14)	936 (64.86)	1443 (8.66)	
12	1257 (38.15)	2038 (61.85)	3295 (19.78)	
13	1428 (42.7)	1916 (57.3)	3344 (20.08)	
14	1205 (43.38)	1573 (56.62)	2778 (16.68)	
15	995 (42.91)	1324 (57.09)	2319 (13.92)	
16	748 (39.14)	1163 (60.86)	1911 (11.47)	
17	414 (37.98)	676 (62.02)	1090 (6.54)	
18	180 (37.89)	295 (62.11)	475 (2.85)	

[a]Row percentages derived from row totals.
[b]No age group differences in proportion of public accounts.

Table 3. Univariable associations between social media privacy and anxiety and depression and covariates.

	Chi-square (*df*)	*F* test (*df*)[a]	*P* value
Publicly available social media account (binary)			
Anxiety and depression	104.2 (1)	—[b]	<.001
Age	52.3 (7)	—	<.001
Sex	198.4 (1)	—	<.001
Bullying	—	15.73 (4)	<.001
Parental guidance of online behavior	—	81.69 (2)	<.001
Proportion of close friendships engaged with online	—	130 (2)	<.001
Poverty status	92.4 (1)	—	<.001
Statutory care	41.0 (1)	—	<.001
Anxiety and depression (binary)			
Age	158.9 (7)	—	<.001
Sex	433.2 (1)	—	<.001
Bullying	1040.4 (4)	—	<.001
Parental guidance of online behavior	14.2 (2)	—	<.001
Proportion of close friendships engaged with online	26.5 (2)	—	<.001
Poverty status	339.3 (1)	—	<.001
Statutory care	21.3 (1)	—	<.001

[a]ANOVA reporting.
[b]Not applicable.

Social Media Use Patterns

The 5 leading SMPs accessed by adolescents in the previous 24 hours were YouTube, WhatsApp, Snapchat, TikTok, and Instagram (Table 4). Three-quarters of the adolescents (12,619/16,655, 75.77%) accessed entertainment content compared to 41.03% (6833/16,655) who engaged with content related to family and friends. Only 12.89% (2147/16,655) of the adolescents accessed support and wellness content, half of whom (1059/2147, 49.32%) screened positive for anxiety and depression (Table 5).

XSL•FO
RenderX

Table 4. Number of adolescents accessing social media platforms in the previous 24 hours proportioned by adolescents reporting having a publicly available social media account.

Social media platform	Total (N=16,655), n (%)	Publicly available social media account (yes; n=6734), n (%)	Proportion with public account (%)
YouTube	13,914 (83.53)	5762 (85.57)	41.41
WhatsApp	13,613 (81.68)	5554 (82.48)	40.8
Snapchat	11,595 (69.58)	5298 (78.66)	45.69
TikTok	11,552 (69.35)	5623 (83.5)	48.68
Instagram	9274 (55.68)	4358 (64.72)	46.99
Pinterest	4971 (29.85)	2099 (31.17)	42.22
BeReal	4839 (29.1)	2293 (34.05)	47.39
Discord	3889 (23.4)	2062 (30.62)	53.02
Twitter	3270 (19.63)	1911 (28.38)	58.44
Facebook	2454 (14.73)	1407 (20.89)	57.33
Other[a]	2065 (12.39)	1038 (15.41)	50.27
Reddit	1861 (11.17)	1026 (15.24)	55.13
Omegle	705 (4.23)	491 (7.29)	69.65
Telegram	375 (2.25)	242 (3.59)	64.53
Signal	215 (1.29)	117 (1.74)	54.42
None of these[b]	204 (1.22)	61 (0.91)	29.9

[a]Platforms accessed are not mutually exclusive: 2.08% (43/2065) of adolescents selecting “other” also selected “none of these.”

[b]None of these, n=204 are dropped in a sensitivity analysis.

Table 5. Number of adolescents accessing social media content in the previous 24 hours proportioned by adolescents screening positive for anxiety and depression.

Social media content	Total (N=16,655), n (%)	Anxiety and depression (yes; n=5429), n (%)	Proportion with anxiety and depression (%)
Entertainment	12,619 (75.77)	4218 (77.69)	33.43
Humor	10,467 (62.85)	3630 (66.86)	34.68
Celebrities and influencers	7323 (43.97)	2694 (49.62)	36.79
Family and friends	6833 (41.03)	2357 (43.41)	34.49
Lifestyle	5875 (35.27)	2223 (40.95)	37.84
Health and fitness	5669 (34.04)	2069 (38.11)	36.50
Beauty and fashion	5639 (33.86)	2341 (43.12)	41.51
Education	4643 (27.88)	1637 (30.15)	35.26
Other	4073 (24.46)	1259 (23.19)	30.91
Support and wellness	2147 (12.89)	1059 (19.51)	49.32
Forums	794 (4.77)	353 (6.50)	44.46

Social Media Privacy: Associations

In an unadjusted model, adolescents with a publicly available social media account showed 41% higher odds of scoring above the threshold for the risk of clinical anxiety and depression compared to those with no public social media account (odds ratio [OR] 1.41, 95% CI 1.32-1.50; $P<.001$). In a model minimally adjusted for age and sex, this association was 13% higher (OR 1.54, 95% CI 1.44-1.65; $P<.001$). In a fully adjusted model, this association was attenuated to 39% higher odds of anxiety and depression in adolescents with a public social media account compared to those with no public social media account, controlled for age, sex, bullying, parental guidance of online behavior, proportion of close friendships engaged with online, poverty status, and statutory care (OR 1.39, 95% CI 1.29-1.49; $P<.001$; Figure 2). The LRT for heterogeneity showed that anxiety and depression outcomes differed significantly by social media account exposure ($P<.001$; Figure 2).

XSL•FO
RenderX

Figure 2. Forest plot showing the association between having a publicly available social media account and anxiety and depression using odds ratios (ORs) and 95% CIs comparing adolescents with a public social media account to those with no public social media account in 3 separately adjusted models. Likelihood ratio test for heterogeneity: *P*<.001. *Minimally adjusted model adjusted for age and sex. **Fully adjusted model adjusted for age, sex, bullying, parental guidance of online behavior, proportion of close friendships engaged with online, poverty status, and statutory care.

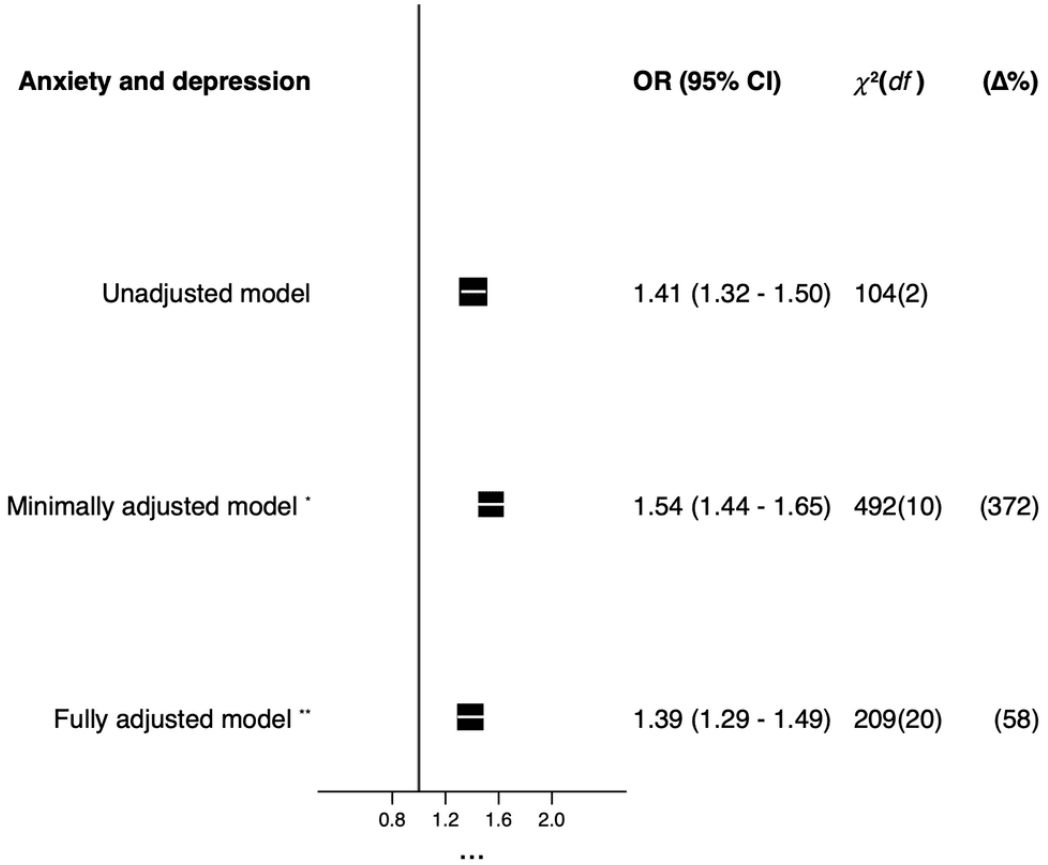

Social Media Privacy: Effect Modification

ORs for the association between anxiety and depression and social media privacy were not shown to be modified by age (*P*=.27) or sex (*P*=.34). However, ORs for the association between anxiety and depression and having a public social media account were found to be statistically significantly modified by parental guidance of online behavior (*P*=.007). Adolescents with a public social media account reporting either *active guidance* or *regulation* had 51% and 52% higher odds of anxiety and depression, respectively, compared to 18% higher odds of anxiety and depression in adolescents with a public social media account reporting *no parental guidance* (Table 6). Analysis in a fully adjusted model showed that, controlling for social media account exposure, adolescents reporting *active guidance* showed 15% lower odds of anxiety and depression compared to those reporting *no parental guidance* (*P*<.001). In adolescents reporting *regulation* compared to *no parental guidance*, the association was not statistically significant (*P*=.11; Figure 3).

XSL•FO
RenderX

Figure 3. Shape plot showing the association between parental guidance of online behavior and anxiety and depression, accounting for social media account privacy, age, sex, bullying, proportion of close friendships engaged with online, poverty status, and statutory care. OR: odds ratio.

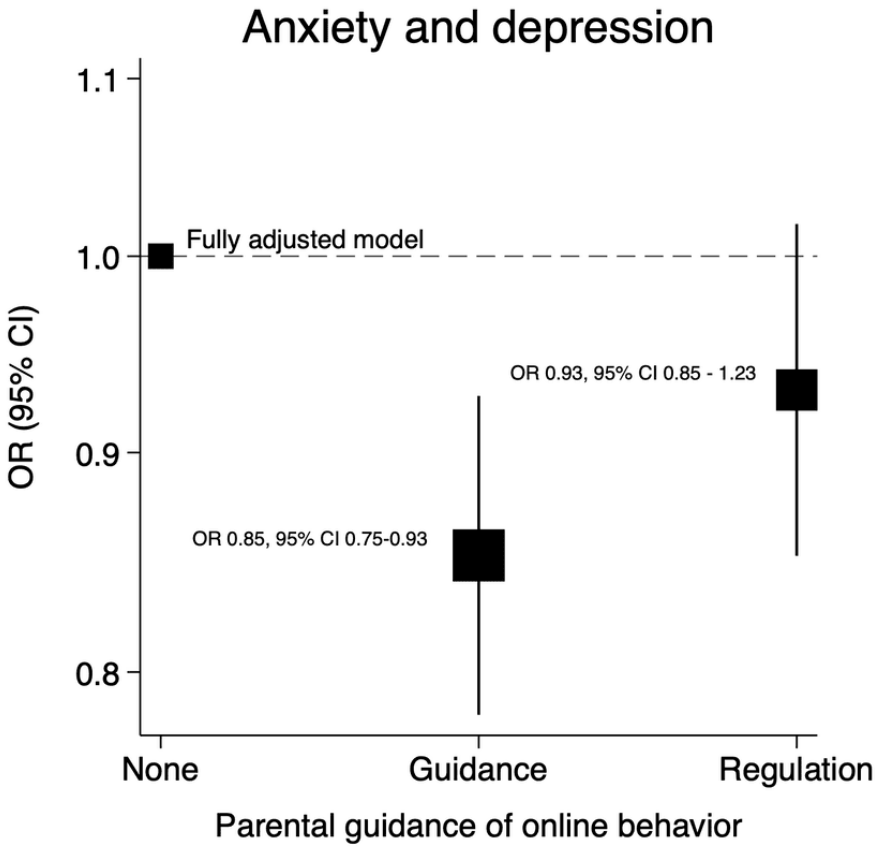

Table 6. Odds ratios (ORs) for the anxiety and depression outcome in adolescents with a publicly available social media account compared to those without a public social media account, showing effect modification by parental guidance of online behavior.

Parental guidance of online behavior	OR (95% CI)	*P* value	LRT[a] interaction chi-square (*df*)
No publicly available social media account	Reference	—	—
Publicly available social media account		.004	11.1 (2)
No parental guidance	1.18 (1.04-1.33)	.007	—
Parental guidance	1.51 (1.34-1.71)	<.001	—
Parental regulation	1.52 (1.34-1.73)	<.001	—

[a]LRT: likelihood ratio test.

Sensitivity Analysis

By removing a small proportion of adolescents (204/16,655, 1.22%) who were most likely to represent those without a public social media account who had no social media account at all from the pool of those without a public social media account who had a private social media account, the analysis was able to demonstrate that there were no significant differences in the odds of anxiety and depression in a fully adjusted sensitivity analysis model (OR 1.38, 95% CI 1.28-1.48; *P*<.001; Multimedia Appendix 1) compared to a fully adjusted model in the full sample (Figure 2).

The effect modification of parental guidance of online behavior on the association between having a public social media account and anxiety and depression also remained significant (*P*=.005) in a fully adjusted sensitivity analysis model (Multimedia Appendix 2), and predictive margins demonstrated that the greatest differences in anxiety and depression outcomes were observed between adolescents who were not exposed to having a public social media account and who reported no parental guidance of their online behavior compared to those who received *active guidance* or *regulation* (Figure 4). This was supported by comparable effect modification associations between the full sample and the sensitivity analysis sample (Figure 5).

XSL•FO
RenderX

Figure 4. Predictive margins plot from the sensitivity sample showing effect modification of social media account privacy by parental guidance of online behavior on the probability of anxiety and depression. LRT: likelihood ratio test.

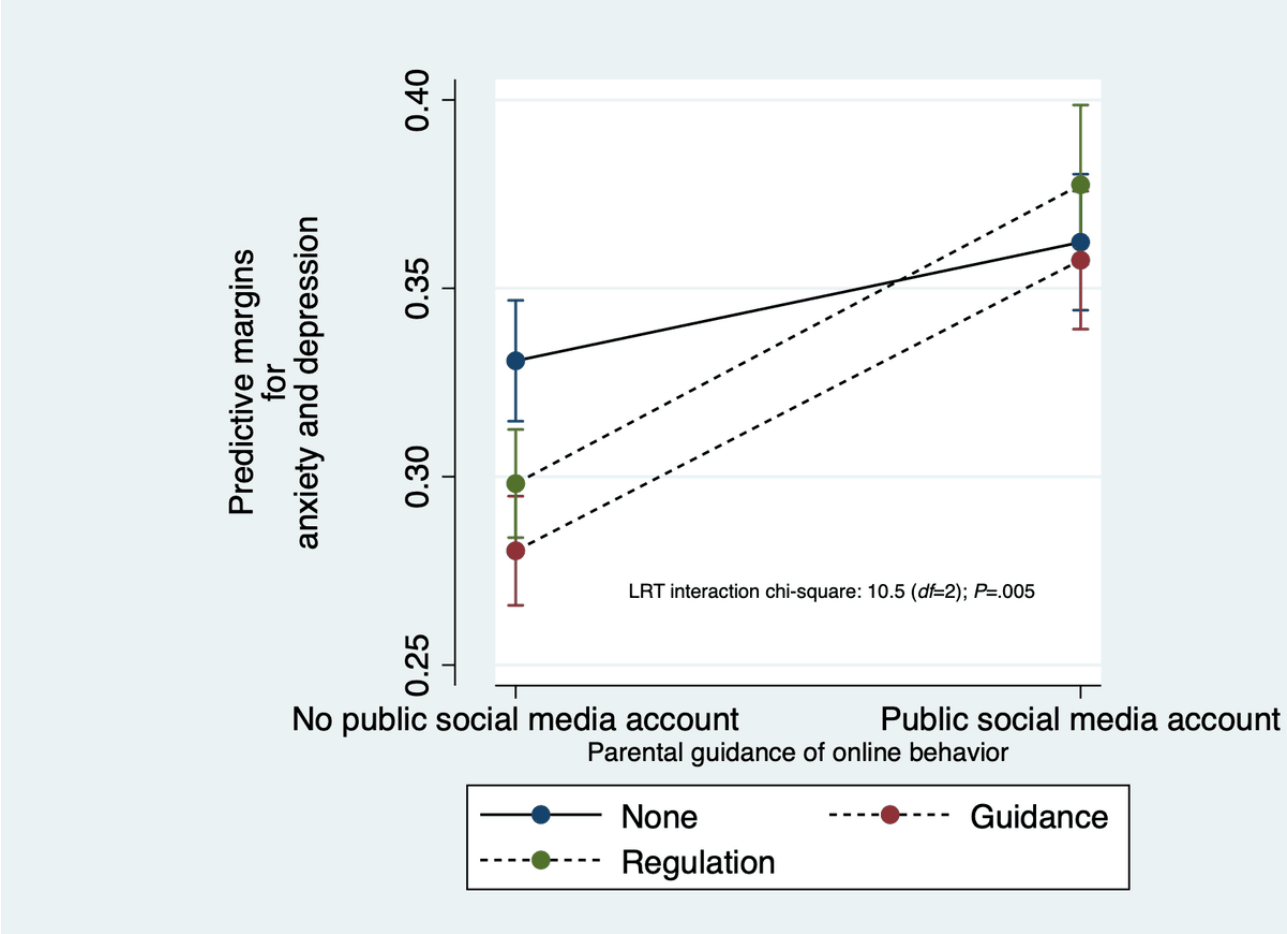

Discussion

Principal Findings

This study set out to describe the association of social media account privacy, determined by whether participants had a public social media account or no public social media account, and symptoms of clinical anxiety and depression in a population of 16,655 school-going adolescents. We found that adolescents with a public social media account had 1.39 times higher odds of symptoms of anxiety and depression compared to those with no public social media account. This association suggests that having a public social media account is likely to be an important factor in the complex behavioral exposure of social media use and should be incorporated in guidance pertaining to online behaviors. We also found that parental approaches to their adolescents' online activity are likely to be an important factor in their mental health.

One interpretation is that adolescents without a public social media account maintain some control over the amount and quality of the material they share publicly, as suggested by Gruzd and Hernández-García [5], and possibly exercise greater awareness of the nature of the material that they, in turn, are exposed to online. A potentially beneficial role of such self-regulation is echoed in findings from an experimental study demonstrating how a 4-week randomized, financially compensated Facebook deactivation intervention increased offline activities among participants [28]. This resulted in increased personal social activity with family and friends, as well as improved depression and anxiety, with an effect equivalent to a quarter of that achieved by psychological therapy at the most conservative estimate [28,29].

In a study investigating specific online disclosure behaviors, nearly half of the social media users reported difficulty managing their privacy settings [5]. The more an individual was concerned about "horizontal" privacy threats from peer group social contacts, the greater the amount and depth of their online disclosures and the weaker the accuracy and awareness of their online disclosures, specifically describing the intentional awareness of self-disclosing online. The more an individual was concerned about "vertical" privacy threats from hierarchical corporations, the greater the accuracy and awareness of their online disclosures and the weaker the amount and depth of their online disclosures; private social media account holders in this group were associated with having more positive content in their online disclosures [5], partially supported by evidence of lower odds of anxiety and depression in adolescents who did not have a public social media account from our findings.

Social Media Privacy Interaction With Parental Guidance of Online Behavior

The association between social media privacy and anxiety and depression was, of interest, only modified by parental guidance of online behavior, characterizing parents who discuss concepts

XSL•FO
RenderX

around the risks of online activity as providing *active gui*dance and parents who monitor their children's online activity as providing *regulation* [14,15]. Adolescents with either *active guidance* or *regulation* had approximately 50% higher odds of anxiety and depression compared to 18% higher odds in adolescents with a public social media account reporting *no parental guidance*. One interpretation of this finding suggests that adolescents experiencing anxiety and depression may have reported behavior from their caregivers that indicated caregiver recognition of their adolescent's mental distress, prompting the implementation of parental guidance and regulation approaches in addressing what may have been perceived by caregivers as the exacerbating effect of their children's social media use on anxiety and depression. This interpretation is supported by analysis showing 15% lower odds of anxiety and depression in adolescents reporting *active guidance* compared to those adolescents reporting *none* (*P*<.001), with no significant effect in those reporting *regulation* (*P*=.45).

Interestingly, upon closer examination of the interaction between social media privacy and parental guidance of online behavior, it was shown that regulatory parental guidance also decreased the odds of anxiety and depression compared to no parental guidance, but only in adolescents who did not have a public social media account (*P*=.002), and not in adolescents with a public social media account (*P*=.28; Figure 5). This reduction was also statistically significant (*P*=.003) in a sensitivity analysis after excluding adolescents who reported not accessing any SMP in the previous 24 hours (Figure 5). Strict regulation may demonstrate lesser benefit in adolescents with a public social media account despite probably intending to reduce the amount of social media use, while open discussion might cautiously be interpreted to be more effective in influencing adaptive online behaviors.

Figure 5. Shape plot of original sample and sensitivity analysis sample showing effect modification of the association between parental guidance of online behavior and anxiety and depression by social media account. OR: odds ratio.

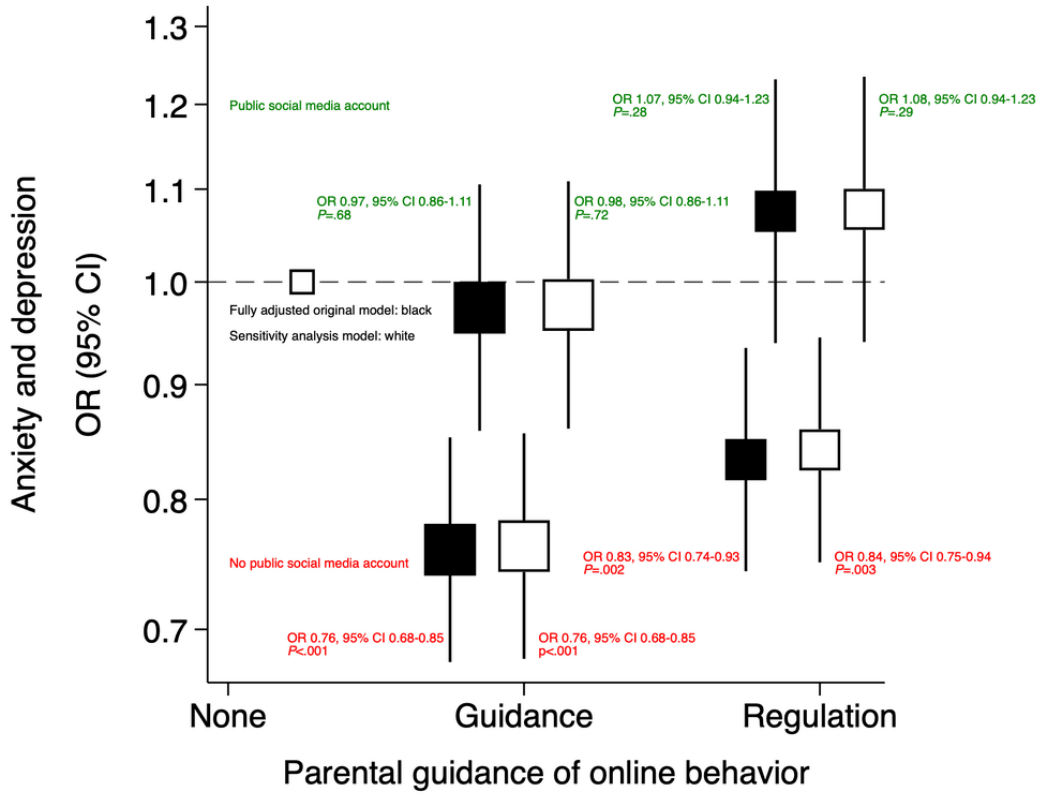

An equally plausible interpretation is that parents who are more regulatory in their approach may be doing so in response to symptomatic mental illness, such as observing severe depressive episodes or self-harming behaviors in their children or responding to concern over their children's disclosure of information online. The relationship between parental guidance of online behavior and anxiety and depression can only be clarified with longitudinal data and in their present form generate the basis of valuable hypotheses. One such valuable association hypothesis is that parents demonstrating *active guidance* may enjoy, on average, healthier relationships with their children, potentially conferring a decreased odds of anxiety and depression risk due to the quality of their interpersonal relationship captured within the variable [6].

Social Media Use Patterns

The primary descriptive analysis of social media use patterns found that YouTube, WhatsApp, and Snapchat were the most

XSL•FO
RenderX

accessed SMPs in the previous 24 hours. Adolescents' preferred SMP activities settled on video and direct messaging formats, while global social media use statistics in 2023 ranked Facebook, YouTube, and WhatsApp as the top SMPs [30]. This highlights shifting generational priorities, with only 14.7% (2454/16,655) accessing Facebook in our study [30]. Almost 49.32% (1059/2147) of the adolescents accessing support and wellness content (2147/16,655, 12.89% of the sample accessed this content) were at risk of anxiety and depression. Given curiosity about the role that online interventions might be able to play, our findings indicate that service uptake in support and wellness is among the least accessed categories, markedly lower than the 32.6% (5429/16,655) of adolescents at risk for anxiety and depression—consistent with previous OxWell data findings about limited online support sought after self-harm [31].

Of the 16,655 adolescents, 9921 (59.57%) reported having no public social media account. New developments in the social media space may complicate the interpretation of such observations. The Threads platform, released in July 2023 by Meta, delivers an unfiltered feed of information that users cannot control while allowing them not to make their accounts public. We cannot know if associations between having a public social media account and anxiety and depression are driven by what young people share versus what they see on their feeds—and future studies should try to disentangle this experimentally. The blurring of public and private social media account features is a rising trend among SMPs. It demonstrates how notions of private and public are becoming less binary, intended by SMPs pursuing behavioral reinforcement through app engagement that adolescents are more vulnerable to. This foregrounds the case for generating useful evidence in sustaining the momentum behind policy-backed social media accountability and mitigation frameworks [5].

Ethical Accountability

A helpful approach in conceptualizing frameworks of responsibility for SMPs and the potential implications of these technologies on adolescents as a vulnerable group draws elements from ethical scholarship. Artificial intelligence algorithms have captured the public imagination. Their processing power exceeds human capabilities, leaving ample room for the application of ethical oversight. To satisfy demands for accountability, SMPs should operate with the understanding that they know what they are doing as agents of responsibility. This condition is presupposed by the moral requirement to provide reasons for their actions to those to whom they are answerable, which, crucially, includes adolescent users who are demonstrably vulnerable to the effects of their algorithms. Recent artificial intelligence algorithmic administration seems to mystify the human agents behind their implementation, thus necessitating a transfer of responsibility from the technology itself to the humans in charge. Coeckelbergh [32] describes how giving the right of explanation rather than the right to information can be more beneficial because it places the onus on SMPs to package such explanations in response to adolescent concerns against growing policy pressures.

Stakeholder involvement has been integral to collecting data within the OxWell study as a model for building trust with adolescents, their caregivers, and school communities. This approach creates the necessary scaffolding for the potential cocreation of online privacy literacy among adolescents [33]. In 2023, the American Psychiatric Association published a health advisory on social media use in adolescence, recommending that in young adolescents, adult monitoring by way of ongoing review, discussion, and coaching around social media content is advised for most youths' social media use and that their autonomy may increase as they age and gain digital literacy [34]. However, this should be balanced with youths' appropriate needs for privacy. Recent policy developments have culminated in parallel bills introduced in the US Senate and the UK Parliament. The US Protecting Kids on Social Media Act would restrict children aged <13 years from creating social media accounts and require parental consent for teenagers aged <18 years to sign up for SMPs [35]. The UK Online Safety Act, which was passed in 2023, recognizes age groups judged to be at risk of harm by SMP features, requiring SMPs to institute platform functionalities allowing for control over the content encountered by children [36].

Limitations

This study has certain limitations. Findings from cross-sectional studies cannot determine the causal inference of association, which remains beyond the scope of this study design. The multilevel structure of data collected from different schools in this study may have benefited from multilevel analytical techniques suited to instances of data of this nature—beyond this study's appropriately used pooled, regression-based analytical techniques. In addition, a distinct exposure classification might have included a third category accounting for the small proportion of adolescents (204/16,655, 1.22%) who reported not accessing SMPs in the past 24 hours, with statistical analysis using multiordinal regression techniques; however, because of predicted challenges with the use of underpowered categorical variables, we excluded these participants from analysis in a sensitivity analysis that showed no significant differences against all hypothesized associations. Observed association risks between individuals with a public social media account versus those without a public social media account and anxiety and depression do not necessarily represent associative risk within the same individual over time, demonstrating the need for a cautious interpretation of the behavioral social media account phenotype's effect as being confined to the point in time at which it was recorded.

Spurious findings from subgroup analyses are not infrequent among well-powered studies, and the magnitude of ORs in our analysis reflects our primary outcome's high prevalence at 32.6% (5429/16,655) screening positive for anxiety and depression. These ORs tend to be higher than the absolute effect size. Not all plausible confounders are accounted for in the regression model due to limitations arising from the characterization of legacy variables in the original dataset of the OxWell Student Survey. However, our conservative approach toward including a priori confounders was balanced against the risk of residual confounding. Plausible confounders for inclusion in future work include parental educational attainment as a gold standard measure for socioeconomic status

XSL•FO
RenderX

and school deprivation indices as a measure of the availability of school resources in addition to parental factors.

The OxWell Student Survey could not determine an overall response rate due to the study team's student recruitment pathway through local education authorities and schools according to a prespecified data protection protocol. However, diverse representation in the study is considerably strengthened by the breadth of exposure to 80 geographically and economically varying schools from the south and north of England, spanning 8 local education authorities. Only a proportion of the eligible sample (16,655/29,065, 57.3%) was included in this study. A fair proportion of the students (8287/29,065, 28.51%) had to be excluded because they did not answer the question on their type of social media account, which appeared further along the survey for which students might not have been allocated sufficient time. The study sample had 90% power with the sample size of 16,655 students, drawn from a considerably representative pool, demonstrating plausible generalizability to an English population of *young* and *middle adolescents*.

Conclusions

The findings from this study confirm the dynamic nature of social media use, specifically highlighting that adolescents with a public social media account may face a higher behavioral risk for anxiety and depression than those without a public social media account. These findings highlight the future role of parental guidance of online activity in supporting the mental health of all adolescents, especially those with a public social media account, and demonstrate how adolescent perceptions of how their parents approach their online activity is associated with significant reductions in the odds of anxiety and depression among young people who do not have a public social media account. Official recommendations in keeping with these findings have been published by the American Psychiatric Association, and attempts to regulate SMPs targeting adolescents are actively underway [34-36]. Social media behaviors and associated patterns of use are complex, and determining which specific behaviors are likely to be associated with risk can help us better understand how best to inquire about and understand the impact of social media use on adolescent mental health.

Future Work

Our findings, along with those of prior researchers, are exploratory steps highlighting the complexity of the behavioral mechanisms at play within social media use. Future research using experimental and longitudinal designs to investigate causal differences between additional behavioral phenotypes, including adolescents with both public and private social media accounts and those without social media accounts, and the resulting impact on mental illness may benefit from the application of network and factor analytic methodology in addressing the complexity of this primary exposure. The value of elucidating parental approaches to the guidance of adolescents' behaviors on SMPs also emerges as a parallel feature that may prove increasingly important. Details on the motivations that parents and caregivers report behind their specific guidance of adolescent online behaviors become key. None of this should be attempted without the coproduction of research questions and the investigation of social media phenomena that are relevant to adolescents, both as a source of insight into adolescent needs and behaviors and as a source of clarity in interpreting findings emerging from such research.

In conclusion, including social media account privacy behaviors and further phenotypes in social media and digital mental health analysis associations promises to contribute to an increasingly detailed picture of this multidimensional, exposure-outcome relationship.

Acknowledgments

The authors would like to thank all students who took the time to participate in the OxWell Student Survey and the many staff at participating schools for the considerable time and planning they put into the survey—without them, this research would not have been possible. The authors thank Janette Fulwood from Brighter Futures Together for her invaluable contribution to the overall OxWell study strategy. The authors thank many colleagues in local authorities and clinical commissioning groups, including James Dove, Giovanni Ferri, Donna Husband, Lajla Johansson, Jonathan Jones, and Chris Price. The authors would also like to thank Pauline Foster and Elaine Purse at Foster and Brown Research Limited, which provided the infrastructure for the 2023 OxWell survey. The authors would also like to acknowledge Ms Shona Leary and Dr Priyanka Panchal for their tireless work on project administration and Dr Simon R White for data curation. The authors also thank the members of the OxWell study team (in alphabetical order): Dr Holly Bear, University of Oxford; Dr Rohan Borschmann, The University of Melbourne; Dr Galit Geulayov, University of Oxford; Dr Gaby Illingworth, University of Oxford; Ms Tanya Manchanda, University of Oxford; Dr Karen L Mansfield, University of Oxford; Ms Rasanat Fatima Nawaz, University of Cambridge; Dr Emma Soneson, University of Oxford; and Dr Simon R White, University of Cambridge.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this paper.

WSM is a Rhodes Scholar with funding support provided by the Rhodes Trust and Trinity College, University of Oxford. GP is funded by a Nuffield Department of Population Health Intermediate Fellowship from Oxford Population Health. This research was funded by the National Institute for Health and Care Research Applied Research Collaboration Oxford and Thames Valley at Oxford Health National Health Service (NHS) Foundation Trust awarded to MF and the Westminster Foundation with support

XSL•FO
RenderX

from Liverpool City Council; Milton Keynes City Council; NHS Frimley Integrated Care Board; NHS Buckinghamshire, Oxfordshire, and Berkshire West Integrated Care Board; and Oxfordshire City Council.

The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to ethical and information governance restrictions. To ensure that uses of the data fall under the remit of the intended purposes set out in the privacy information and to prevent duplication of analyses, data are available from the research team on reasonable request after a review process. The full list of questions as well as other details are available on a project-specific OxWell Open Science Framework website [37] along with the preregistration of the study protocol [38].

Authors' Contributions

WSM was responsible for conceptualization, methodology, formal analysis, visualization, and writing (original draft preparation and review). SH and GP were responsible for writing (review and editing). MF was responsible for writing (review and editing), funding acquisition, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analysis showing odds ratios for anxiety and depression outcomes in adolescents who have a publicly available social media account compared to adolescents without a public social media account.
[DOCX File , 19 KB - jmir_v26i1e57154_app1.docx]

Multimedia Appendix 2
Sensitivity analysis showing effect modification of having a public social media account by online parental guidance on anxiety and depression outcomes.
[DOCX File , 16 KB - jmir_v26i1e57154_app2.docx]

References

1. Verduyn P, Gugushvili N, Kross E. The impact of social network sites on mental health: distinguishing active from passive use. World Psychiatry 2021 Feb 12;20(1):133-134. [doi: 10.1002/wps.20820] [Medline: 33432746]
2. Hancock J, Liu SX, Luo M, Mieczkowski H. Psychological well-being and social media use: a meta-analysis of associations between social media use and depression, anxiety, loneliness, eudaimonic, hedonic and social well-being. SSRN Journal. Preprint posted online March 9, 2022 2022 [FREE Full text] [doi: 10.2139/ssrn.4053961]
3. Burke M, Marlow C, Lento M. Social network activity and social well-being. In: Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems. 2010 Presented at: CHI '10; April 10-15, 2010; Atlanta, GA p. 1909-1912 URL: https://dl.acm.org/doi/10.1145/1753326.1753613 [doi: 10.1145/1753326.1753613]
4. Valkenburg PM, Meier A, Beyens I. Social media use and its impact on adolescent mental health: an umbrella review of the evidence. Curr Opin Psychol 2022 Apr;44:58-68 [FREE Full text] [doi: 10.1016/j.copsyc.2021.08.017] [Medline: 34563980]
5. Gruzd A, Hernández-García Á. Privacy concerns and self-disclosure in private and public uses of social media. Cyberpsychol Behav Soc Netw 2018 Jul 01;21(7):418-428 [FREE Full text] [doi: 10.1089/cyber.2017.0709] [Medline: 29995525]
6. Abi-Jaoude E, Naylor KT, Pignatiello A. Smartphones, social media use and youth mental health. CMAJ 2020 Feb 10;192(6):E136-E141 [FREE Full text] [doi: 10.1503/cmaj.190434] [Medline: 32041697]
7. Twenge JM, Haidt J, Lozano J, Cummins KM. Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. Acta Psychol (Amst) 2022 Apr;224:103512 [FREE Full text] [doi: 10.1016/j.actpsy.2022.103512] [Medline: 35101738]
8. Barsova T, Cheong Z, Mak A, Liu J. Predicting psychological symptoms when Facebook's digital well-being features are used: cross-sectional survey study. JMIR Form Res 2022 Aug 29;6(8):e39387 [FREE Full text] [doi: 10.2196/39387] [Medline: 36036971]
9. Orben A, Przybylski AK, Blakemore SJ, Kievit RA. Windows of developmental sensitivity to social media. Nat Commun 2022 Mar 28;13(1):1649 [FREE Full text] [doi: 10.1038/s41467-022-29296-3] [Medline: 35347142]
10. Plackett R, Sheringham J, Dykxhoorn J. The longitudinal effect of social media use on adolescent mental health in the UK: findings from the UK Longitudinal Household Study. Lancet 2022 Nov;400:S71. [doi: 10.1016/s0140-6736(22)02281-4]

XSL•FO
RenderX

11. Beeres D, Andersson F, Vossen HG, Galanti MR. Social media and mental health among early adolescents in Sweden: a longitudinal study with 2-year follow-up (KUPOL study). J Adolesc Health 2021 May;68(5):953-960 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.07.042] [Medline: 32943289]
12. Jensen M, George MJ, Russell MR, Odgers CL. Young adolescents' digital technology use and mental health symptoms: little evidence of longitudinal or daily linkages. Clin Psychol Sci 2019 Nov 20;7(6):1416-1433 [FREE Full text] [doi: 10.1177/2167702619859336] [Medline: 31929951]
13. Hartas D. The social context of adolescent mental health and wellbeing: parents, friends and social media. Res Pap Educ 2019 Dec 18;36(5):542-560 [FREE Full text] [doi: 10.1080/02671522.2019.1697734]
14. Liu C, Ang RP, Lwin MO. Cognitive, personality, and social factors associated with adolescents' online personal information disclosure. J Adolesc 2013 Aug 14;36(4):629-638. [doi: 10.1016/j.adolescence.2013.03.016] [Medline: 23849657]
15. Ho SS, Lwin MO, Yee AZ, Lee EW. Understanding factors associated with Singaporean adolescents' intention to adopt privacy protection behavior using an extended theory of planned behavior. Cyberpsychol Behav Soc Netw 2017 Sep;20(9):572-579. [doi: 10.1089/cyber.2017.0061] [Medline: 28876967]
16. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health 2018 Mar;2(3):223-228. [doi: 10.1016/S2352-4642(18)30022-1] [Medline: 30169257]
17. Booker CL, Kelly YJ, Sacker A. Gender differences in the associations between age trends of social media interaction and well-being among 10-15 year olds in the UK. BMC Public Health 2018 Mar 20;18(1):321 [FREE Full text] [doi: 10.1186/s12889-018-5220-4] [Medline: 29554883]
18. Mansfield KL, Puntis S, Soneson E, Cipriani A, Geulayov G, Fazel M. Study protocol: the OxWell school survey investigating social, emotional and behavioural factors associated with mental health and well-being. BMJ Open 2021 Dec 08;11(12):e052717 [FREE Full text] [doi: 10.1136/bmjopen-2021-052717] [Medline: 34880020]
19. Radez J, Waite P, Chorpita B, Creswell C, Orchard F, Percy R, et al. Using the 11-item version of the RCADS to identify anxiety and depressive disorders in adolescents. Res Child Adolesc Psychopathol 2021 Sep 01;49(9):1241-1257 [FREE Full text] [doi: 10.1007/s10802-021-00817-w] [Medline: 33792821]
20. Driving the adoption of common measures. International Alliance of Mental Health Research Funders. URL: https://iamhrf.org/projects/driving-adoption-common-measures#faqs [accessed 2024-04-29]
21. Rajkumar K, Saint-Jacques G, Bojinov I, Brynjolfsson E, Aral S. A causal test of the strength of weak ties. Science 2022 Sep 16;377(6612):1304-1310. [doi: 10.1126/science.abl4476] [Medline: 36107999]
22. Haythornthwaite C. Social networks and internet connectivity effects. Inf Commun Soc 2005 Jun;8(2):125-147. [doi: 10.1080/13691180500146185]
23. Granovetter MS. The strength of weak ties. Am J Sociol 1973 May;78(6):1360-1380. [doi: 10.1086/225469]
24. Dubois-Comtois K, Bussières EL, Cyr C, St-Onge J, Baudry C, Milot T, et al. Are children and adolescents in foster care at greater risk of mental health problems than their counterparts? A meta-analysis. Child Youth Serv Rev 2021 Aug;127:106100. [doi: 10.1016/j.childyouth.2021.106100]
25. Napoletano A, Elgar FJ, Saul G, Dirks M, Craig W. The view from the bottom: relative deprivation and bullying victimization in Canadian adolescents. J Interpers Violence 2016 Dec;31(20):3443-3463. [doi: 10.1177/0886260515585528] [Medline: 25985976]
26. Twenge JM, Martin GN, Campbell WK. Decreases in psychological well-being among American adolescents after 2012 and links to screen time during the rise of smartphone technology. Emotion 2018 Sep;18(6):765-780. [doi: 10.1037/emo0000403] [Medline: 29355336]
27. Treloar W. Mental health of children and young people in England. National Health Service, England. 2022. URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2022-follow-up-to-the-2017-survey/part-1---mental-health [accessed 2024-04-29]
28. Allcott H, Braghieri L, Eichmeyer S, Gentzkow M. The welfare effects of social media. Am Econ Rev 2020 Mar 01;110(3):629-676. [doi: 10.1257/aer.20190658]
29. Cheng C, Wang HY, Sigerson L, Chau CI. Do the socially rich get richer? A nuanced perspective on social network site use and online social capital accrual. Psychol Bull 2019 Jul;145(7):734-764. [doi: 10.1037/bul0000198] [Medline: 31094537]
30. Kemp S. Digital 2023 April global Statshot report. DataReportal. URL: https://datareportal.com/reports/digital-2023-april-global-statshot [accessed 2024-04-29]
31. Geulayov G, Borschmann R, Mansfield K, Hawton K, Moran P, Fazel M. Utilization and acceptability of formal and informal support for adolescents following self-harm before and during the first COVID-19 lockdown: results from a large-scale English schools survey. Front Psychiatry 2022 Jul 19;13(S1):881248 [FREE Full text] [doi: 10.3389/fpsyt.2022.881248] [Medline: 35815012]
32. Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 2020 Aug 24;26(4):2051-2068 [FREE Full text] [doi: 10.1007/s11948-019-00146-8] [Medline: 31650511]
33. Al Zou'bi HW, Khatatbeh M, Alzoubi KH, Khabour OF, Al-Delaimy WK. Attitudes and knowledge of adolescents in Jordan regarding the ethics of social media data use for research purposes. J Empir Res Hum Res Ethics 2020 Jan 23;15(1-2):87-96 [FREE Full text] [doi: 10.1177/1556264620901390] [Medline: 31971059]

XSL•FO
RenderX

34. Canady VA. APA panel calls for safe, social media use by adolescents. Ment Health Wkly 2023 May 12;33(19):7-8. [doi: 10.1002/mhw.33641]
35. Canady VA. Bill would ban kids under 13 from joining social media platforms. Ment Health Wkly 2023 Apr 28;33(17):8. [doi: 10.1002/mhw.33624]
36. Online safety act 2023: parliamentary bills. UK Parliament. URL: https://bills.parliament.uk/bills/3137 [accessed 2024-04-29]
37. The OxWell student survey on mental health and wellbeing. Open Science Framework. URL: https://osf.io/sekhr/ [accessed 2024-04-29]
38. Adolescent social media use: privacy and mental health. Open Science Framework. URL: https://osf.io/knrdc [accessed 2024-04-29]

Abbreviations

FEC: further education college
LRT: likelihood ratio test
OR: odds ratio
RCADS-11: Revised Child Anxiety and Depression Scale-11
SMP: social media platform

Edited by T de Azevedo Cardoso; submitted 14.02.24; peer-reviewed by A Raffoul, O Itani; comments to author 03.06.24; revised version received 27.08.24; accepted 15.11.24; published 17.12.24.

Please cite as:
Mabaso WS, Hein S, Pavarini G, The OxWell Study Team, Fazel M
Exploring the Relationship Between Public Social Media Accounts, Adolescent Mental Health, and Parental Guidance in England: Large Cross-Sectional School Survey Study
J Med Internet Res 2024;26:e57154
URL: https://www.jmir.org/2024/1/e57154
doi:10.2196/57154
PMID:39688898

©Wakithi Siza Mabaso, Sascha Hein, Gabriela Pavarini, The OxWell Study Team, Mina Fazel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Influencers in Policy Fields on Social Media: Global Longitudinal Study of Dietary Sodium Reduction Posts, 2006-2022

Alana Montoya[1*], MS; Lingchao Mao[2*], MS; Adam Drewnowski[3], PhD; Joshua Chen[4], HS; Ella Shi[5], HS; Aileen Liang[6], HS; Bryan J Weiner[7], PhD; Yanfang Su[8], SCD

[1]Information School, University of Washington, Seattle, WA, United States
[2]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States
[3]Department of Epidemiology, University of Washington, Seattle, WA, United States
[4]Skyline High School, Sammamish, WA, United States
[5]Redmond High School, Redmond, WA, United States
[6]Interlake High School, Bellevue, WA, United States
[7]Department of Global Health, University of Washington, Seattle, WA, United States
[8]Evans School of Public Policy and Governance, University of Washington, Seattle, WA, United States
[*]these authors contributed equally

Corresponding Author:
Yanfang Su, SCD
Evans School of Public Policy and Governance
University of Washington
4105 George Washington Lane Northeast
Seattle, WA, 98105
United States
Phone: 1 2066165418
Email: yfsu@uw.edu

Abstract

Background: Excessive sodium intake is a major concern for global public health. Despite multiple dietary guidelines, population sodium intakes are above recommended levels. Lack of health literacy could be one contributing issue and contemporary health literacy is largely shaped by social media.

Objective: This study aims to quantify the posting behaviors and influence patterns on dietary sodium–related content by influencers in the policy field on X (formerly Twitter) across time.

Methods: We first identified X users with a scope of work related to dietary sodium and retrieved their posts (formerly Tweets) from 2006 to 2022. Users were categorized into the policy groups of outer-setting organization, inner-setting organization, or individual, based on their role in the conceptual policy field. Network analysis was used to analyze interactions among users and identify the top influencers in each policy group. A 4D influence framework was applied to measure the overall influence, activity, priority, originality, and popularity scores. These measures were used to reveal the user-level, group-level, and temporal patterns of sodium-related influence.

Results: We identified 78 users with content related to dietary sodium, with 1,099,605 posts in total and 14,732 dietary sodium posts. There was an increasing volume of sodium posts from 2010 to 2015; however, the trend has been decreasing since 2016, especially among outer-setting organizations. The top influencers from the three policy groups were the World Health Organization (WHO), the American Heart Association, and Tom Frieden. Simon Capewell and the WHO ranked the highest in activity; the World Action on Salt, Sugar, and Health and Action on Salt had the highest priority for dietary sodium content; General Mills and Tom Frieden had the highest originality; and WHO, Harvard University School of Medicine, and Tom Frieden received the highest popularity. Outer-setting organizations tend to interact with more users in the network compared to inner-setting organizations and individuals, while inner-setting organizations tend to receive more engagements from other users in the network than the other two groups. Monthly patterns showed a significant peak in the number of sodium posts in March compared with other months.

Conclusions: Despite the increased use of social media, recent trends of sodium intake education on social media are decreasing and the priority of sodium among other topics is low. To improve policy implementation effectiveness and meet recommended

XSL•FO
RenderX

dietary targets, there is an increasing need for health leaders to consistently and collectively advocate for sodium intake reduction on social media.

(J Med Internet Res 2024;26:e54506) doi:10.2196/54506

KEYWORDS

policy field; sodium intake; sodium consumption; cardiovascular disease; social media; health education; health promotion; dissemination; influence; Twitter; X; activity; priority; originality; popularity

Introduction

The excessive sodium content in most diets remains a significant challenge to global public health [1]. The World Health Organization (WHO) has estimated the mean global sodium intake to be 4310 mg per day, more than double the WHO recommendation of <2000 mg per day [1]. Excessive sodium intake is an established cause of raised blood pressure and elevated risk of cardiovascular disease (CVD) [1,2]. The Global Burden of Disease study identified a high-sodium diet as the leading dietary risk factor, with an estimated 3 million deaths attributable to it annually [3,4]. Unlike nonmodifiable risk factors, such as age and family history, sodium intake has been documented as a modifiable risk factor for CVD [5]. Interventions to reduce sodium intake are highly cost-effective in both developed and developing countries [6-9]. Specifically, a US study showed that a 10% average reduction in sodium intake, a reduction level similar to successful efforts in the United Kingdom, could prevent over 510,000 strokes and 480,000 myocardial infarctions in American adults [2].

Despite the release of several dietary guidelines, recent studies have observed no significant decline in sodium intake among American adults, and adherence to these guidelines is poor [10,11]. Many individuals are unaware of their salt intake or the effect of this intake on their long-term health, creating an increasing concern for and a need to raise awareness of sodium overconsumption [12]. Health literacy, defined as “the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions,” has been shown to relate to disparities in the use of health care services and consequently impact individual health outcomes [13-15]. A key venue for increasing health literacy is for health care professionals and health systems to provide access to health information [16]. Health systems can leverage channels for information dissemination to inform individuals about the negative health impacts of sodium overconsumption. As traditional information dissemination venues such as radio broadcasting, television, newspapers, and magazines adapt to the digital landscape, there is a need for increased research on digital means of information dissemination.

The use of social media for health education and promotion can play a key role in informing individuals on important health topics. Among several social media platforms, X (formerly Twitter) has emerged as a popular platform, with 70% of American adults using it as a source of news [17]. X has been widely used for health literacy promotion, infoveillance of health-related content, analyzing audience reach, and identifying stakeholder characteristics [17-19]. There is a need for an accessible tool that can quantify the impact of information dissemination on social media platforms and explain its driving factors. Mao et al [20] proposed a framework for measuring topic-specific influence on X, which quantifies the overall influence of each user by decomposing it into four dimensions: activity, popularity, originality, and priority. This framework was illustrated in a case study of dietary sodium–related posts [20]; however, the analysis was only conducted for 16 stakeholders, which may not have been representative of the health field landscape and provided limited insights for policy implementation.

This study aims to provide a more comprehensive understanding of the trends of dietary sodium–related information dissemination on X among stakeholders in the policy field from 2006 to 2022. We borrow Moulton and Sandfort’s [21] concept of policy fields in identifying sodium policy–related stakeholders on X. We define a policy field as a bounded network among organizations and individuals carrying out substantive policy making and information dissemination on a specific topic, and in this study, we focus on the topic of sodium reduction. We placed X users into one of three policy groups based on their role in the policy field: inner-setting organizations, such as public agencies that are directly involved in policy making; outer-setting organizations, such as research organizations that provide scientific evidence to facilitate policy making; and individuals, such as policy entrepreneurs, who may act across both the inner and outer settings. We analyzed the in-network interactions among users to identify the top influencers in different groups. After applying the 4D framework of Mao et al [20] to measure topic-specific influence, group-level and temporal trends of influence were investigated to understand the health leaders’ dissemination efforts on sodium intake reduction.

Methods

Overview

To understand the trends of dietary sodium–related information dissemination, we identified health domain users with a scope of work related to dietary sodium. Users were positioned into one of the 3 policy groups. All posts from these users from 2006 to 2022 were retrieved. We used social network analysis to identify the top influencers within each policy group and computed the 4D measures of influence for the top influencers. These results were used to reveal user-level, group-level, and temporal patterns of sodium-related influence.

Account Selection

We took a 2-step approach, starting with a knowledge-informed “macrostructure” to identify influencers and then refined the

XSL•FO
RenderX

selection using data-informed "micro behavior" analysis, as depicted in Multimedia Appendix 1. In this context, the macrostructure was outlined by 10 functional categories spanning public agencies, research and evaluation organizations, international organizations and initiatives, nongovernmental organizations, professional associations, policy entrepreneurs, philanthropies, scholars, food manufacturers, and advocacy organizations. We identified candidate influencers under each category by collecting a list of relevant organizations based on our domain knowledge, a list of experts from the New York City (NYC) Health Consensus Statement on Sodium [22], and top-ranked universities from US News regarding the top public health programs [23]. We then refined our list based on data, checking for whether these organizations and individuals had active accounts on X, posted any public health or sodium-related content, and what other accounts they interacted with in their sodium-related posts. This approach for account selection allowed us to comprehensively cover relevant stakeholders in dietary sodium reduction and ensure representatives from each category are included. In total, 78 X accounts were identified that encompassed domestic and international organizations across the 10 categories (see the full list in Multimedia Appendix 2).

Account Categorizations

We conceptualize all X accounts as actors in policy fields, interacting with each other to carry out a substantive policy or program. Based on each organization or individual's role in policy implementation, we categorized the corresponding X handle into one of three policy groups as either an inner-setting organization, an outer-setting organization, or an individual (Multimedia Appendix 3). We conceptualize the inner setting as a setting within the policy field in which policies are actively created. We define outer-setting organizations as organizations outside of the direct policy-making setting that still influence policy. We define individuals as actors with the potential to influence policy.

Data Collection

All posts from the 78 accounts from X's conception in March 2006 through May 31, 2022, were retrieved using the Academic Research licensed X (Twitter) Full-Archive Application Programming Interface V2. Post data contained the post ID, author username, post creation date and time, language of the post, post text, public metrics (count of reposts, replies, likes, and quotes), hashtags, whether the post was a repost, and mentions. Definitions of these terminologies can be found in X Glossary [24].

Data Preprocessing

To determine which posts were related to dietary sodium, filtering was performed based on keyword criteria to label dietary sodium–related posts. First, English expressions containing "salt" with their meanings rooted in cultural or historical contexts that do not refer to dietary sodium were removed from the posts. Examples of slang terms used include "salt into the wound" or "salt of the earth." Then, predefined sodium-related keywords and combinations with food-related keywords were used to flag dietary sodium content. Posts not containing keywords but using hashtags related to dietary sodium were also labeled as dietary sodium related. The full list of English expressions, keywords, and hashtags can be found in Multimedia Appendix 4. Note that only English posts were considered, as indicated in the language attribute of the post data.

Network Analysis

Next, we used a data-driven approach to rank the initial sample of accounts. This sample can be seen as a community of stakeholders with a scope of work related to dietary sodium. We identified the most important accounts based on interactions within the community. First, we constructed a directed graph in which nodes represent users and edges are weighted based on the number of dietary sodium–related engagements from one user to another. User interactions on X included (1) a post mentioning another user, (2) a post quoting another user's post, (3) a repost of another user's post, and (4) a reply to another user's post. We then used PageRank [25] to measure the in-network importance of each user based on sodium-related interactions. PageRank was originally proposed to assess the importance of web pages in the context of search engine results, with the underlying assumption that more important websites are likely to receive more and higher quality (ie, from other websites considered as important) links from other websites. These dynamics naturally translate to the X environment. In this context, PageRank scores can be interpreted as the likelihood that a person randomly browsing X will arrive at a user's profile. We conducted the social network analysis using the networkx package [26]. We then selected the same number of top accounts from each policy group based on their PageRank scores to create an equally distributed sample for group-level analysis.

Influence Analysis

The 4D framework of influence [20] was applied to measure the sodium-specific influence of each X account. Denote N as the number of posts, R as the number of reposts, P as the number of public engagements, and superscript ^{t} as dietary sodium specific. Influence is decomposed into four dimensions: (1) activity, which measures the level of participation in X, (2) priority of the topic within the scope of work of the user, (3) originality, which represents the proportion of posts that involve the creation of original content, and (4) popularity, which quantifies the average public attention received per post. A summary of the four measures and their equations is presented in Table 1. The overall measure of influence can be computed as:

$$\text{Influence}^t = \text{Activity} \times \text{Priority} \times \text{Originality}^t \times \text{Popularity}^t,$$

which can be shown to be equivalent to the total public engagements received from all topic-specific posts.

XSL•FO
RenderX

Table 1. 4D framework of influence.

Dimension	Description	Measure
Activity	Level of participation	N
Priority	Relative importance of a given topic within the scope of work	N^t/N
Originality	Proportion of posts that required creation of original content	$(N^t - R^t)/N^t$
Popularity	Average number of public engagements received per (content-creation) post	$(P^t_{like} + P^t_{reply} + P^t_{quote} + P^t_{repost}) / (N^t - R^t)$

Ethical Considerations

Since the data used in this study are publicly available on X with minimal risks to human participants, this study was deemed exempt from review by the University of Washington's institutional review board (#18401). The data used in this study were obtained with explicit consent from X for research purposes. As users of X data, we only accessed data that is publicly available.

Results

Overall Trends

Of the initial 78 users, 6 accounts did not have any dietary sodium–related posts (Departmentofh14, WHOSTP, SJPHonline, UCLeHealth, ChanZuckerberg, and Kcferdmd). The remaining 72 accounts totaled 1,099,605 posts and 14,732 dietary sodium–related posts. The number of dietary sodium–related posts from these accounts significantly increased from 2010 to 2015, increasing from 178 dietary sodium posts in 2010 to 1647 dietary sodium posts in 2015 (Figure 1). However, the number of dietary sodium posts began to decrease after 2016, with small increases in 2019 and 2021.

Figure 1. Total number of dietary sodium posts from all X accounts from March 2006 to May 2022.

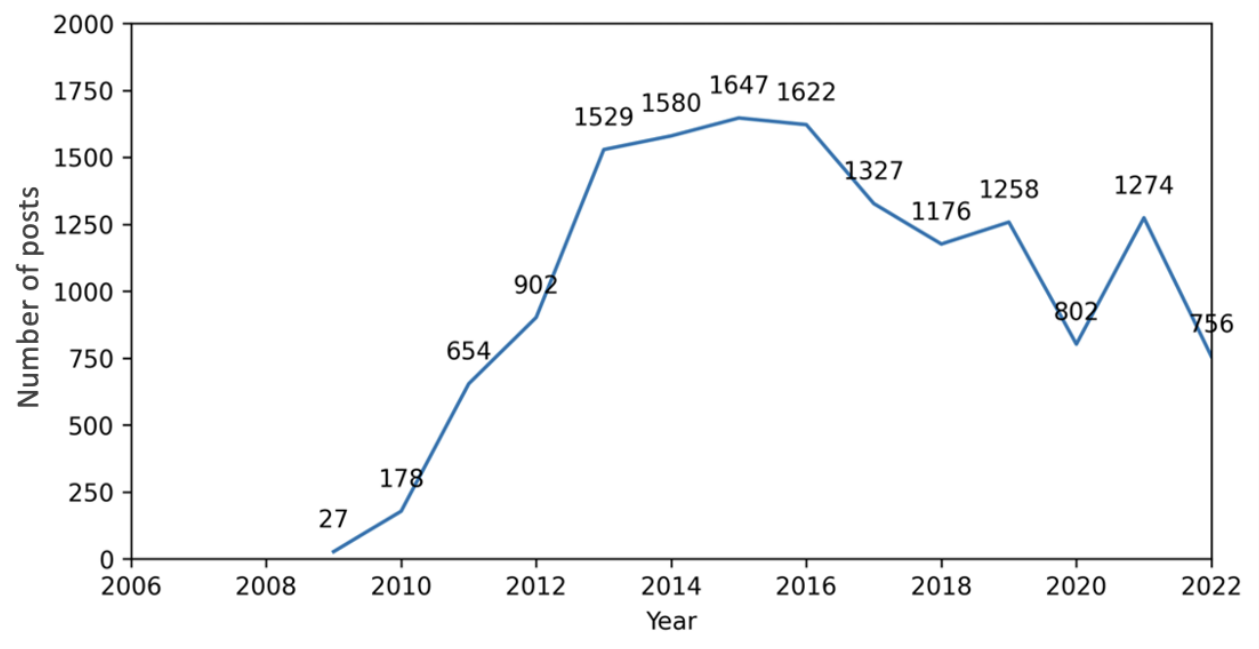

In-Network User Interactions

The top 11 users with the largest PageRank scores from each policy group were selected for a more detailed influence analysis. Figure 2 [27] shows the dietary sodium–related interactions among the top influencers in each policy group. X accounts from the WHO, the American Heart Association (AHA), and Tom Frieden had the greatest overall influence at each policy level. The WHO had significantly higher dietary sodium–related influence than other users, with an overall influence measure of 55,593, while the AHA and Tom Frieden had an overall influence of 26,395 and 12,672, respectively.

XSL•FO
RenderX

Figure 2. Dietary sodium–related engagements among top users from each policy group selected based on in-network importance. Each node represents an X user, with the color of the node indicating the policy group of the users and the size of the node representing the overall influence score of the user (ie, the larger the node, the more dietary sodium influence the user had). The arrows indicate the direction of the engagements, and the thickness of the arrows represents the number of engagements between users (ie, the thicker the line, the more engagements from one user to the other user). For an interactive figure, refer to [27]. AHA: American Heart Association; APHA: American Public Health Association; CDC: Centers for Disease Control and Prevention; CFSAN: Center for Food Safety and Applied Nutrition; CSPI: Center for Science in the Public Interest; FDA: Food and Drug Administration; HHS: Department of Health and Human Services; NYC-DOH: New York City Department of Health and Mental Hygiene; RTSL: Resolve to Save Lives; USDA: US Department of Agriculture; USDA-FNS: US Department of Agriculture Food and Nutrition Service; WASSH: World Action on Salt, Sugar, and Health; WHO: World Health Organization.

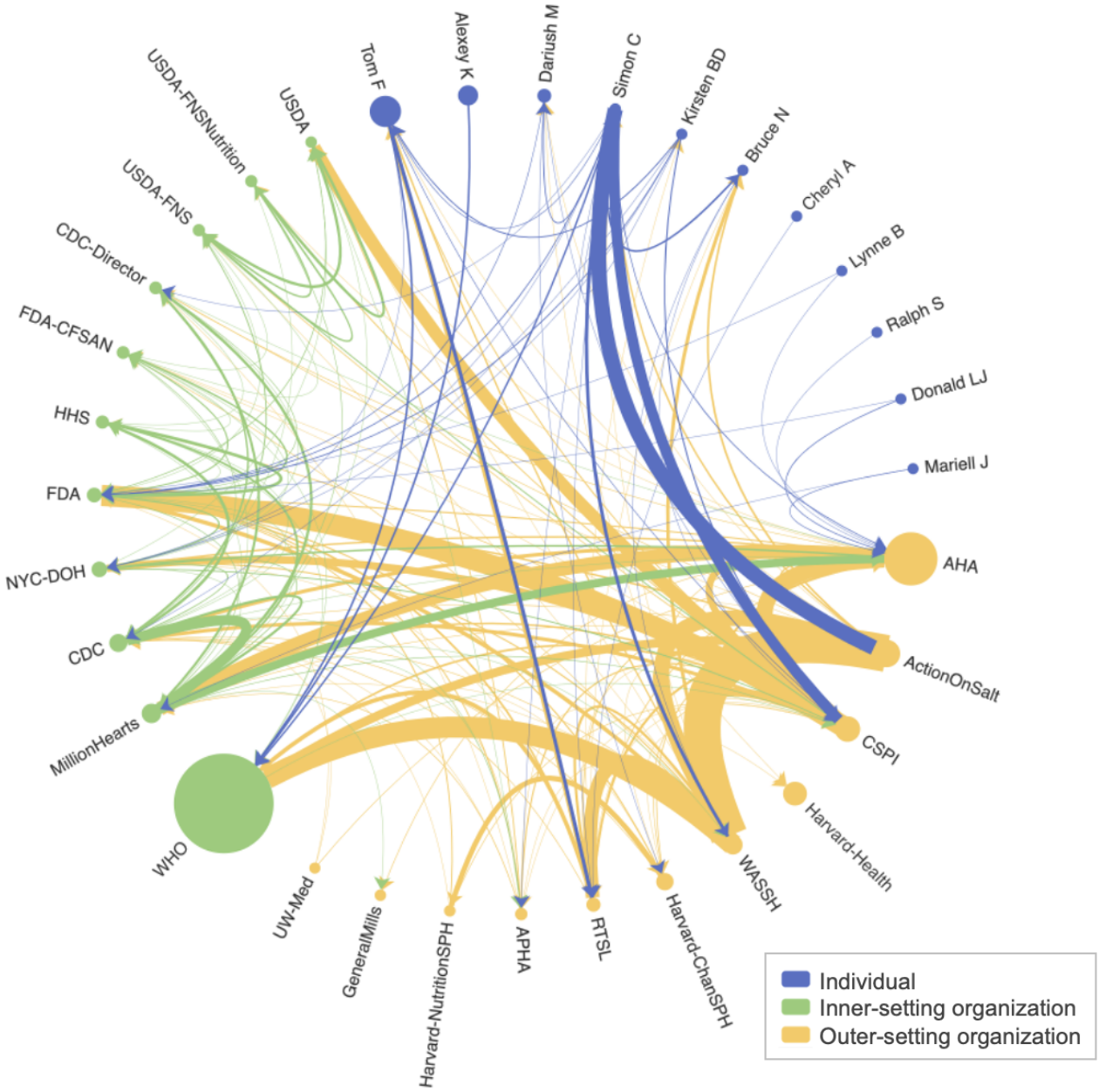

Frequent interactions were observed within the top influencers (Figure 2 [27]). The WHO received a high volume of engagements from users of all three policy groups but did not frequently engage with other users in the network. The strongest amount of dietary sodium–related engagement the WHO received within the network was from the World Action on Salt, Sugar, and Health (WASSH), a global group founded with the mission to reduce salt intake to the WHO-recommended levels. WASSH not only actively interacted with large organizations including WHO but also had reciprocal interactions with similar initiatives including Action on Salt, as well as with global public health initiatives such as Resolve to Save Lives (RTSL). The Center for Science in the Public Interest (CSPI) is another example of a highly active organization that engaged with many other national organizations, including AHA, the US Department of Agriculture (USDA), the US Food and Drug Administration (FDA), the NYC Department of Health and Mental Hygiene (NYC-DOH), and the Centers for Disease Control and Prevention (CDC). Frequent mutual interactions were also observed between partnered entities such as those between the AHA and Million Hearts, a national initiative co-led by the CDC and the Centers for Medicare and Medicaid Services with

XSL•FO
RenderX

the mission to prevent CVD; and those between RTSL and Tom Frieden.

Outer-setting organizations tend to engage with more users in the network showing broader in-network outreach efforts. Specifically, outer-setting organizations engaged with 8.4 (SD 6.4) other users within the network on average, followed by inner-setting organization users engaged with 5.3 (SD 3.0) other users on average, and individual users with 3.4 (SD 3.5) other users. The CSPI, WASSH, and Action on Salt had the broadest sodium-related in-network outreach, with 18, 17, and 15 other users, respectively. On the other hand, inner-setting organizations received engagements from more users in the network from 9.7 (SD 5.2) other users on average, compared to outer-setting organizations from 5.7 (SD 5.3) users on average, and individuals from 1.5 (SD 2.1) other users on average. FDA and CDC received engagements from 21 and 17 other users in the network. Simon Capewell was the most active individual user, engaging with 13 other users in the network, while Tom Frieden received the most engagement from other users. Multimedia Appendix 5 shows a summary of these results.

Annual Trend of Influence Across Policy Groups

The annual trend of the total dietary sodium–related influence of top users in each policy group is shown in Figure 3. Inner-setting and outer-setting organizations had greater influence than individual users overall. The sodium-related influence of inner-setting organizations fluctuated considerably across all years. Alternating spikes and sudden decreases in dietary sodium influence occurred roughly every other year, varying by more than 60% in some years. In contrast, outer-setting organizations experienced a steady increase in dietary sodium–related influence from the launch of X in 2006 through 2016, followed by a decline thereafter. A dip in dietary sodium influence was observed in 2020 for both inner-setting and outer-setting organizations although not for individual users. The influence of individual users has been increasing over the years, with a significant spike in 2021.

Figure 3. Annual trends of total dietary sodium–related influence of top users from the three policy groups. Years without 12 months of data were excluded (years 2006 and 2022) for a fair comparison.

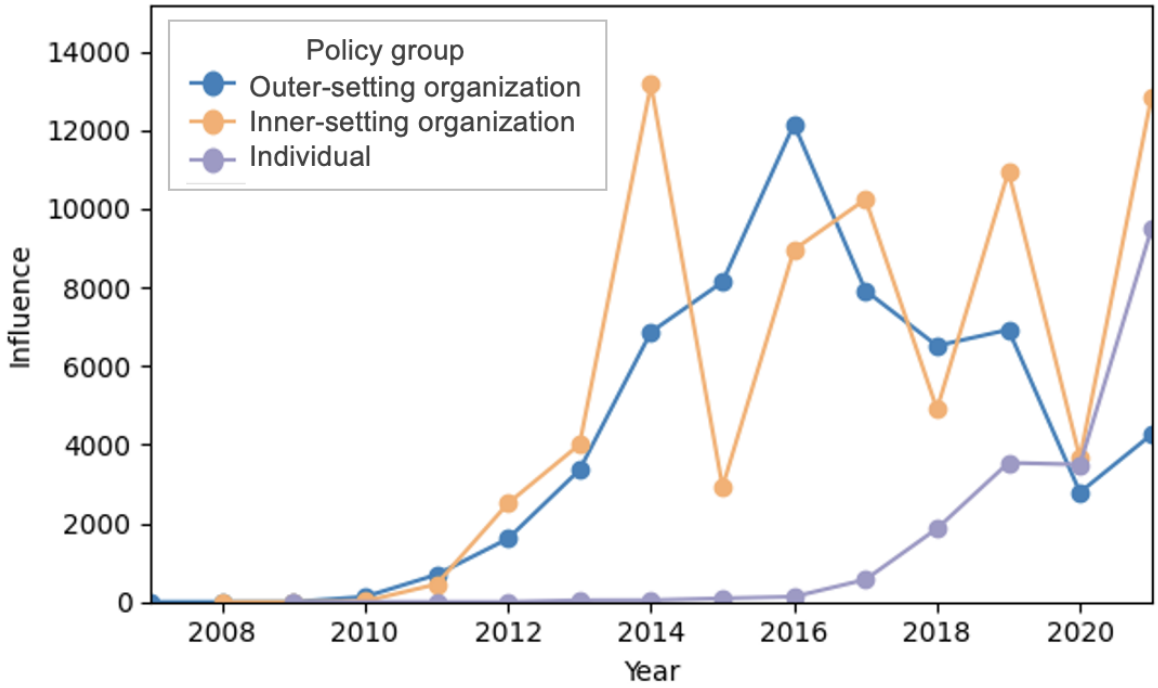

The WHO emerged as the leading influencer among inner-setting organizations, the AHA led among outer-setting organizations, and Tom Frieden stood out as the top influencer among individual users. The WHO had a steadily increasing trend of posting volume since joining X in 2008 and received a similar increase in average public attention per original post (Figure 4, left panel). A significant increase in activity and popularity was observed in 2020, which was coincident with the COVID-19 pandemic. Despite having one of the largest posting volumes in this sample, the WHO's priority for dietary sodium–related posts remained very low across all years, given WHO's large scope of work. The WHO did not post any dietary sodium–related posts until 2011 and consequently, priority was zero in the first 3 years prior and remained near zero across later years. In terms of originality, the WHO posted the highest percentage of original sodium-related content in 2013, coinciding with the Member States' adoption of the global action plan for the prevention and control of noncommunicable diseases [28]. From 2014 to 2019, the WHO steadily relied more on reposts, but originality began increasing again in 2020 as the target year for achieving a 30% relative reduction in population sodium intake approached

XSL•FO
RenderX

Figure 4. Top influencers in each policy group and their four dimensions of influence. Years without 12 months of data were excluded (years 2006 and 2022) for a fair comparison. The three panels in the figure represent the top influencers in each policy group: the WHO for inner-setting organizations, the AHA for outer-setting organizations, and Tom Frieden for individual users. Each plot displays a yearly comparison of the four influence dimensions. Originality and popularity ranged from 0 to 1. Activity ranged from 168 to 11,612. Popularity ranged from 0 to 528. For visualization purposes, activity and popularity were scaled to be between 0 and 1. AHA: American Heart Association; WHO: World Health Organization.

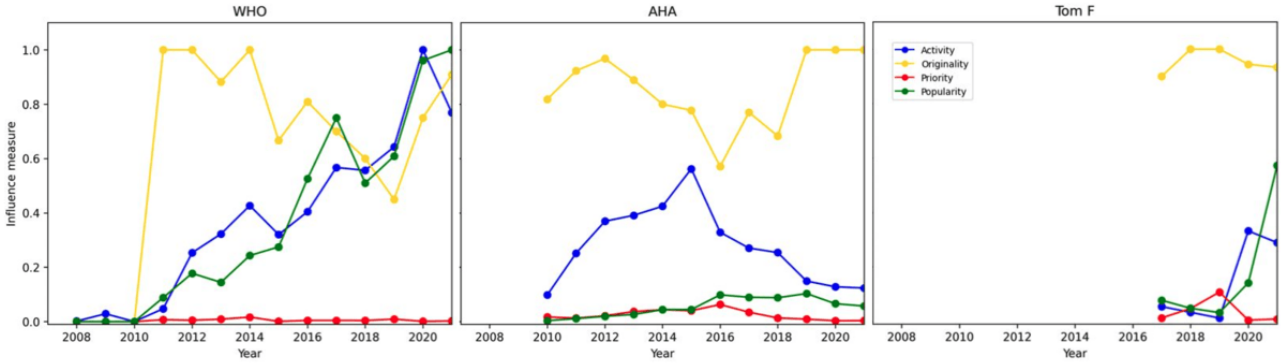

The AHA had increasing posting activity from the creation of their account in 2010 through 2015 but the posting volume started decreasing following that year (Figure 4, middle panel). Their priority of dietary sodium posts was comparatively higher than the WHO for many years, especially from 2013 through 2017, which is expected given that the AHA's scope of work in CVD is closely related to sodium intake. The originality of their dietary sodium posts was in a similar range as the WHO in earlier years, reaching 100% originality starting in 2019 indicating the AHA's efforts in content creation for sodium-related topics.

Despite being one of the newest to join the social media platform, Tom Frieden was by far the top influencer among the individual users (Figure 4, right panel). His posting activity level was typically lower than WHO and AHA. Tom Frieden had significantly higher priority placed on dietary sodium–related topics in 2018 and 2019 than the other top influencers and tended to prefer posting original content rather than reposts. We present a more detailed case study for Tom Frieden later in this section.

Leaders in Four Dimensions of Influence

A comparison of the four influence measures across the users in each policy group across all years is shown in Figure 5. Overall, outer-setting organizations tend to have higher activity levels than inner-setting organizations or individuals. Among inner-setting organizations, the WHO demonstrated the highest level of posting activity, nearly doubling that of the next most active organization. This further underscores the WHO's global leadership in sodium reduction efforts. Despite most individual users having relatively low activity levels as compared to organizations, Simon Capewell was an outlier who posted the largest number of posts in this sample, with nearly 80,000 posts since conception, more than the total number of posts by the WHO. It is worth noting that roughly 80% of Simon Capewell's posts were reposts, indicating that the main role of his account is a disseminator rather than a content creator.

Figure 5. Comparison of activity, priority, originality, and popularity across users in each policy group. Activity is displayed in the blue bars and measured based on the left y-axis. Priority and originality are displayed as the red triangles and yellow boxes, respectively, and measured based on the right y-axis. Popularity is displayed in parentheses next to the X handle names. Within each policy group, users are arranged by popularity. AHA: American Heart Association; APHA: American Public Health Association; CDC: Centers for Disease Control and Prevention; CFSAN: Center for Food Safety and Applied Nutrition; CSPI: Center for Science in the Public Interest; FDA: Food and Drug Administration; HHS: Department of Health and Human Services; NYC-DOH: New York City Department of Health and Mental Hygiene; RTSL: Resolve to Save Lives; USDA: US Department of Agriculture; USDA-FNS: US Department of Agriculture Food and Nutrition Service; WASSH: World Action on Salt, Sugar, and Health; WHO: World Health Organization.

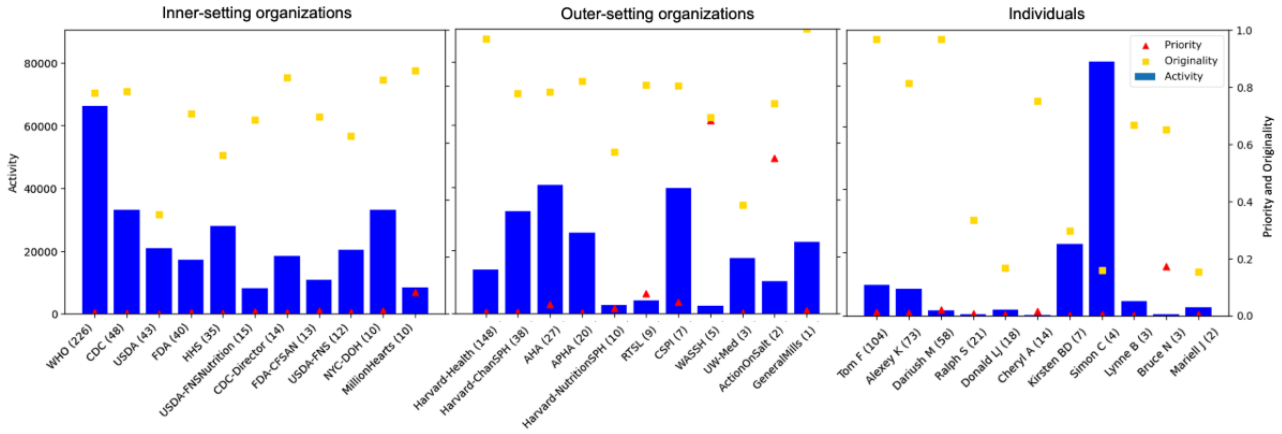

XSL•FO
RenderX

The presence of dietary sodium posts out of total posts (measured by priority) is very low among the overall efforts of these users regardless of their policy group. Among outer-setting organizations, the WASSH and Action on Salt had the highest priority for dietary sodium content, with over half of their contents related to dietary sodium topics. The individual user Bruce Neal, who is a UK-trained physician and currently executive director at the George Institute for Global Health Australia, received priority scores of approximately 20%, ranking highest among inner-setting organizations. The US-based initiative Million Hearts had a priority score of around 10%. Following them are outer-setting organizations RTSL, the AHA, and CSPI. The remaining users had less than 2% of their posts related to dietary sodium.

In terms of popularity, the WHO also ranked the highest across all policy groups. Notably, larger posting volume is not always associated with higher public attention per post. Tom Frieden and Harvard School of Medicine (Harvard-Health) are among the highest top three in the sample despite having low-mid level activity.

Inner-setting and outer-setting organizations had relatively similar levels of originality. General Mills, Harvard-Health, Tom Frieden, and Dariush Mozaffarian had the highest levels of originality with over 95% of their sodium-related posting efforts dedicated to original content. In contrast, the USDA and UW School of Medicine (UW-Med) had the lowest originality among the inner-setting and outer-setting organizations with over 60% of their sodium-related posts being reposts. There is more variability of originality within the individual category. About half of individual users focus more on content creation and the other half focus more on content dissemination.

Monthly Posting Patterns

The total number of dietary sodium–related posts in each month across all policy-type users from 2007 through 2021 is shown in Table 2. A noticeable peak of posting volume is observed in March. The other months tend to stay relatively consistent with the number of dietary sodium posts, averaging about 857 dietary sodium posts per month, over four times less than the volume in March.

Table 2. Monthly posting patterns of dietary sodium–related posts from 2007 to 2021. Years without 12 months of data were excluded (years 2006 and 2022) for a fair comparison.

Month	Number of dietary sodium–related posts
January	792
February	857
March	3432
April	930
May	801
June	971
July	989
August	795
September	1161
October	690
November	761
December	678

A Notable Example: Tom Frieden

To provide a more specific example of analysis, we consider Tom Frieden, a physician and notable public figure in public health who has played numerous roles in both the inner-setting and outer-setting organizations. As shown in Figure 2 [27], Tom Frieden was not only the top-ranked individual user in terms of overall influence of dietary sodium content but also had close relationships with three other accounts identified in our analyses as being influential (NYC-DOH, CDC-Director, RTSL). He serves as President and Chief Executive Officer at RTSL starting in 2017, and previously was Commissioner at the NYC-DOH from 2002 to 2009 and Director of the CDC from 2009 to 2017 [29]. Tom Frieden created his personal X account in 2017. The dietary sodium–related influence of the various X accounts that have been associated with Tom Frieden is displayed in Figure 6 with partitions provided to indicate when he worked for each organization.

Having held leadership roles in multiple health organizations, Tom Frieden's personal account had a higher overall influence than the accounts of the organizations he served at. Tom Frieden's personal account reached the greatest dietary sodium–related influence in 2021, over six times that of the next highest influence account, which was that of the NYC-DOH, in 2016. The NYC-DOH account did not have any dietary sodium posts from 2006 to 2009 when Tom Frieden was Commissioner. The CDC-Director account was created in 2011 when Tom Frieden was serving as Director at the CDC and had consistent growth in sodium-related influence from 2011 to 2014. However, the CDC-Director's activity on dietary sodium topics dropped to zero after his departure from the CDC in 2017 and in the years after. The RTSL account was established when Tom Frieden joined RTSL in 2017, and its sodium-related influence has steadily grown since its creation, aligning with the sodium-focused priorities of Tom Frieden's personal

XSL•FO
RenderX

account. RTSL had a relatively higher dietary sodium influence with about 319 average public engagements each year, which may be partly attributed to Tom Frieden's frequent interactions with RTSL to increase its visibility. Given the creation of both the CDC-Director and RTSL during his roles at the respective organizations, as well as the lack of dietary sodium influence with the CDC-Director after he left the CDC, Tom Frieden appears to have catalyzed the information dissemination of dietary sodium topics in these organizations.

Figure 6. An example of a policy entrepreneur: Tom Frieden. The line plot visualizes the influence evolution of Tom Frieden's personal account and the three accounts of the organizations he served at. Partitions provided to indicate when Tom Frieden worked for each group. CDC: Centers for Disease Control and Prevention; NYC-DOH: NYC Department of Health and Mental Hygiene; RTSL: Resolve to Save Lives.

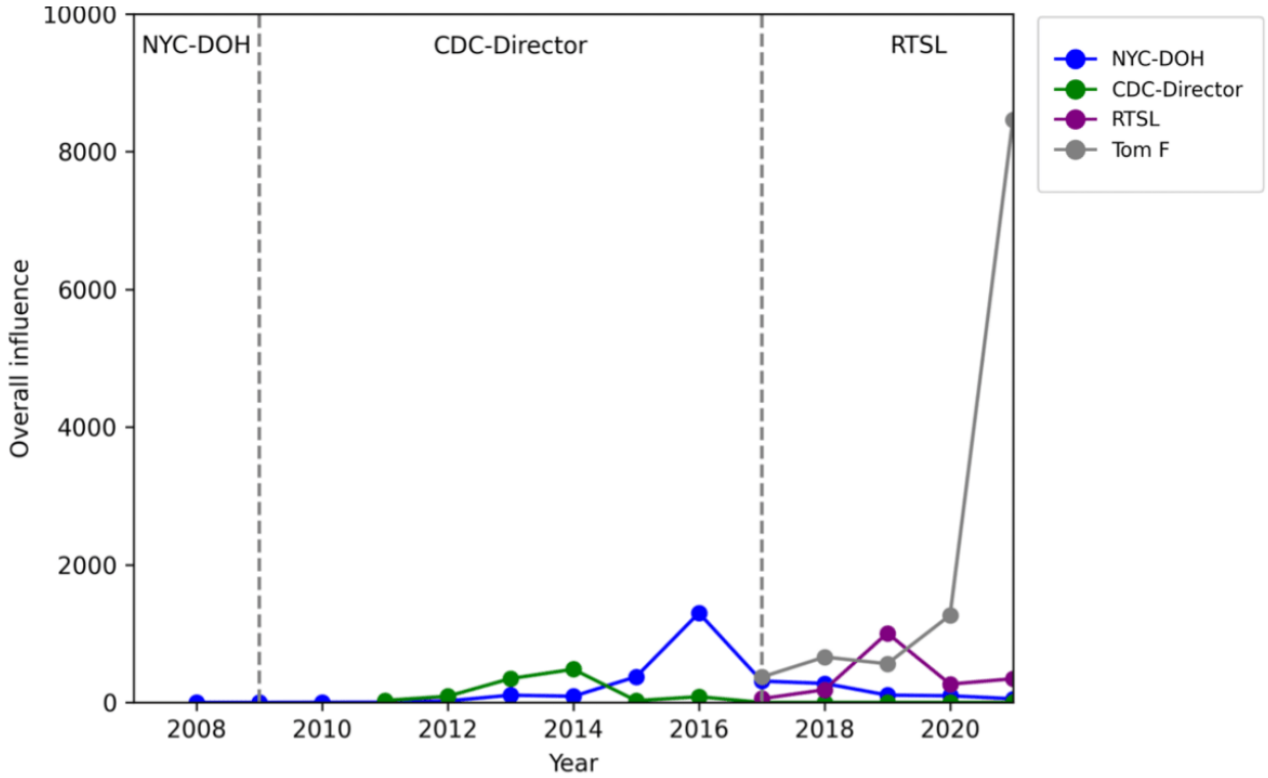

Discussion

Overall Trends

This study leveraged the 4D influence framework and the concept of policy fields to analyze health leaders' efforts in disseminating dietary sodium–related content on X in the past 16 years. Overall, the presence of sodium intake–related content is very low among the overall dissemination efforts of health leaders across all policy groups, except for initiatives specialized in sodium consumption such as WASSH and Action on Salt. Initially, dissemination efforts about dietary sodium content increased after the WHO's announcement of sodium reduction targets in 2013 [28]. Two years later, the USDA released Dietary Guidelines for Americans [30], which could explain the rise of dietary sodium intake discussion on social media and the spike in outer-setting and inner-setting organizations' sodium-related influence in 2016. However, the overall dissemination efforts on sodium intake have been decreasing since then. A local peak was observed in 2021, which may be attributed to the National Academy of Sciences, Engineering, and Medicine's establishment of a chronic disease–specific recommendation for dietary sodium in 2019 [31] followed by the USDA 2020 Dietary Guidelines for Americans [30] attracting some attention in sodium reduction. The sodium-related influence of outer-setting organizations, in particular, has largely decreased from 2020 to 2022, which is possible due to these organizations redirecting efforts to other health topics related to the COVID-19 pandemic. In contrast, individual users, particularly Tom Frieden, gained significantly more public attention during the pandemic and continued advocating for the prevention of CVD. Inner-setting organizations exhibit a sporadic pattern of influence on sodium-related topics with large fluctuations year over year. This could be related to the inconsistency of their efforts in addressing the topic of dietary sodium intake or related to external factors that would impact the engagement they received on their content.

Comparison Across Policy Groups

Comparing the three policy groups, inner-setting and outer-setting organizations had generally greater overall influence as opposed to individual users. This may be because organizations typically constitute larger groups with more access to resources for health advocacy and social media dissemination. Individual users likely do not have comparable resources but may instead focus on increasing the visibility of their posts on social media through strategies such as using hashtags and interacting with larger organizations or increasing their activity through less time-consuming reposting. It is worth noting that the top two individual influencers, Tom Frieden and Alexey Kulikov, both serve or served in leadership positions in inner-setting organizations.

XSL•FO
RenderX

In-Network User Interactions

A significant volume of interactions was observed among the top influencers. A potential reason behind these interactions is the collaboration of efforts to reduce dietary sodium consumption. For example, RTSL and WASSH were observed to have reciprocal interactions, possibly due to the funding provided by RTSL for a salt reduction toolkit led by WASSH [32]. Another reason the organizations interact is to hold each other accountable to drive positive change. We observed that CSPI actively engaged with the FDA. A recent press release noted that the CSPI has been recommending the FDA to decrease the sodium in foods since 1978 and continues to advocate for sodium reduction, including filing petitions for the FDA to better regulate sodium in processed food [33]. Outer-setting organizations tend to interact with more users in the network compared to inner-setting organizations and individuals, while inner-setting organizations tend to receive more engagements from other users in the network than the other two groups. Leveraging the professional authority can enhance the legitimacy of policy dissemination. For instance, WASSH and Action on Salt engaged the WHO as an effective strategy to promote international sodium reduction guidelines and scientific evidence.

Leaders in Four Dimensions of Influence

The top influencers from inner-setting organizations, outer-setting organizations, and individual policy groups were the WHO, AHA, and Tom Frieden, respectively. As a large global organization, the WHO had a steadily increasing presence and popularity on social media over the years. The WHO is the most widely engaged health organization among all health leaders, which reinforces its influential position as the goal-setting organization for sodium reduction. However, the priority of advocating for sodium intake reduction is very low (<1% of total posts) in the WHO's broad scope of work. Expectedly, the AHA had relatively higher priority for sodium content than the WHO given its goal of fighting heart disease through education. The AHA's priority for advocating for sodium intake reduction peaked in 2016, likely due to the release of Dietary Guidelines for Americans [30]. However, less than 2% of their annual posts were related to sodium since 2018. Having held leadership roles in both inner-setting and outer-setting organizations, Tom Frieden has been a prominent health policy entrepreneur and social media influencer. Since 2019, Tom Frieden's personal account has posted an annual volume of posts similar to large organizations such as the AHA. He peaked in priority for sodium content in 2019 since the creation of RTSL in 2017, periodically interacting with RTSL's account through his personal account, and continued advocating for combating high blood pressure and CVD but with a lower portion of posts discussing sodium intake in 2019. Tom Frieden is a notable example of a driver of health education through social media, not only through his personal account but also by promoting the accounts of organizations he served at. Tom Frieden is likely a common factor contributing to the rise in influence of the CDC-Director and RTSL accounts during his service at those organizations.

Originality, which indicates a preference for content creation, is an important but not sole driver of influence. While both inner-setting and outer-setting organizations have similar levels of originality, with about one-third of posts being reposts, there is more variability of behavior among individual users. Some individual users are original content creators (eg, Tom Frieden and Dariush Mozaffarian), who focus on sharing their expertise in health and sodium consumption, and others are content disseminators (eg, Simon Capewell, Mariell Jessup, and Donald M Lloyd-Jones), who actively disseminate the content created by other authorities. While content creators may attract public attention by providing their unique expertise, disseminators have the advantage of time efficiency which allows them to publish a large volume of posts with less effort. A notable example is Simon Capewell, who had the highest level of activity in the sample despite being an individual user.

Monthly Posting Patterns

Analyzing the monthly trends of posts across all years reveals a significant peak of dietary sodium–related posts in March. This is likely because National Nutrition Month (#NationalNutritionMonth), as well as World Salt Awareness Week (#SaltAwarenessWeek, #WorldSaltAwarenessWeek) typically occur in March, both of which can lead to an increase in sodium-related discussions on social media [34,35]. Such an increase in attention for sodium consumption was not observed for other sodium-related events such as Stroke Awareness Month (#StrokeAwarenessMonth) and Hypertension Awareness Month (#HighBloodPressureMonth) in May or the National Heart Month (#NationalHeartMonth) in February.

Factors that contributed to a higher sodium intake include knowledge, attitude, and habits, which can be modified by education [36,37]. Given the increasing use of social media by the general public, there is an increasing need for health leaders to advocate for sodium intake reduction and use social media as an important platform for health education. Theories of implementation effectiveness convey the importance of (1) building a strong implementation climate, which refers to the targeted user's shared perceptions about the impact of the policy, and (2) delivering the fit of that policy to the targeted user's values [38,39]. Additionally, fostering shared beliefs and collection action among interdependent stakeholders is critical for successful implementation [39,40]. Social network advertising has become a key component of strategic communication and audience engagement [41]. Thus, it is crucial for health leaders to collectively and consistently advocate for the recommended dietary guidelines of sodium intake and the long-term health risks of current sodium intake levels.

Comparison With Prior Work

To the best of our knowledge, the first work to analyze the public influence of sodium-related posts of health organizations and leaders is Mao et al [20]. We include a broader range of stakeholders selected from 10 different categories to provide results that are more representative of the public health landscape. By including a larger sample, our analysis found influencers who displayed relatively high priority (Million Hearts, Bruce Neal, and RTSL), popularity (Harvard-Health),

XSL•FO
RenderX

originality (Dariush Mozaffarian, Harvard-Health, and General Mills), and activity (Department of Health and Human Services, NYC-DOH, and American Public Health Association), which were not included in Mao et al's analysis. In terms of the accounts analyzed by both studies, our analysis found similar patterns. Note that the top influencers found by our analysis account for a fifth dimension of influence which is the in-network importance. As a result, the Food and Agriculture Organization of the United Nations was a top influencer in the study by Mao et al [20] but not considered a top influencer in our PageRank analysis. In terms of study objectives, Mao et al [20] focused on evaluating the dissemination behavior of stakeholders to reveal their strengths and weaknesses. In contrast, we aim to understand stakeholders' role in policy implementation and its relationship to their health information dissemination behaviors. This was accomplished by positioning stakeholders within one of the three policy groups, analyzing their interactions via network analysis, and comparing the patterns of outer-setting organizations, inner-setting organizations, and individuals at a policy group level. Our results reveal that these three types of users exhibited different temporal evolution of sodium-related influence possibly due to their distinct reactions to national and international policy events such as the WHO's sodium reduction target announcement, the USDA's release of dietary guidelines.

We applied Moulton and Sandfort's [21] concept of policy fields to categorize the users into three policy groups, though there exist other frameworks. Cruden et al [42] defined five categories of policy actors: (1) policy developers, who shape the vision or intention of a policy; (2) policy disseminators, who communicate the policy with relevant stakeholders; (3) policy implementers, who usually have decision-making authority in their organization and lead the adoption of the new policy into their practice settings; (4) policy influencers, who can be formal or informal actors that impact which evidence is used in which ways during policy implementation; and (5) policy enforcers, who monitor, communicate, and support the adoption and implementation quality of the policy. While this categorization is more detailed, it is difficult to apply to our analysis because many times an organization or individual could fall into multiple categories of policy actors.

Limitations

There are several limitations in this study. First, we used keyword-based criteria to determine whether a post was dietary sodium–related. Although we intended to curate a comprehensive list of salt-related expressions, hashtags, and food-related synonyms, this labeling method may still incorrectly classify some posts. Using more advanced natural language processing techniques is a possible improvement direction. Second, our account selection focused on organizations and individuals with the scope of work in public health and dietary sodium–related posts and may exclude unconventional users with potential impact on this topic. Third, our analysis only focused on English posts, while some international organizations such as the WHO published a small portion of posts in other languages. Fourth, the 4D framework used in this study focused on measuring influence in terms of quantifiable engagements with social media content. Content analysis of individual posts is outside the scope of this work. Fifth, our analysis focuses on X and excludes other social media platforms where influencers might be active. Future research is warranted to continuously track the dietary-related sodium influence on social media after 2022 and extend the analysis to other nutritional policies on sugar and trans-fat.

Conclusions

This study analyzed the dietary sodium–related dissemination efforts of national and international public health leaders on the social media platform X from 2006 to 2022. We used the 4D framework to measure the public influence of sodium-related posts [20] and applied Moulton and Sandfort's [21] policy field concept to relate dissemination behaviors with the users' role in policy implementation. Our analysis identified top influencers in sodium content dissemination (WHO, AHA, Tom Frieden, and others), analyzed in-network interactions among the health leaders, and contrasted the behavior of the inner-setting, outer-setting, and individual policy groups. Despite the increased use of social media over the last decade, the overall dissemination efforts in sodium intake education have been decreasing from 2016 to date, particularly for outer-setting organizations. The priority of sodium-related topics is very low among the health leaders across all policy groups except for specialized initiatives. Given the cost-effectiveness of sodium reduction on CVD prevention and the lack of awareness of the health risks of high sodium intake by the public, there is an increasing need for health leaders to consistently and collectively advocate for the reduction of sodium consumption on social media to achieve the recommended dietary targets.

Acknowledgments

We sincerely thank Emily Chu (Stanford University), Luyanda Z Ngongoma (University of Washington), Yuhang Ding (University of Washington), Jinghong Gu (University of Washington), Jingning Wang (International School), Meranda Wang (Interlake High School), Owen Fei (Interlake High School), Ariel V Lin (University of Washington), Hilma N Nakambale (University of Washington), Hung Yi Kuo (University of Washington), reviewers and editors for providing valuable feedback and review to improve this manuscript. This project is funded through grants from the National Institutes of Health (K01HL166688 and R25DK123008). The contents are the responsibility of the authors and do not reflect the views of the funders. The funders had no role in the study design, conduct of the study, analysis, or dissemination of findings.

XSL•FO
RenderX

Data Availability

The datasets generated during or analyzed during this study are not publicly available due to X (Twitter)'s data sharing policies but are available for retrieval from X's Full Archive Application Programming Interface on access approval from X [43].

Authors' Contributions

YS is the principal investigator for this project and led the study design, literature review, framework development, and result interpretation. AM led data analysis, result interpretation, and manuscript writing. LM led data collection, result interpretation, and manuscript refinement. AD and BJW contributed to the result interpretation. JC, ES, and AL contributed to data quality control and the manuscript writing. All authors critically revised the manuscript and approved the final version of the paper.

Conflicts of Interest

AD is the original developer of the Nutrient Rich Food Index, a nutrient profiling model; has held grants and contracts related to nutrient profiling, and has advised governments, foundations, and the private sector on issues related to nutrient profiling of foods.

Multimedia Appendix 1
A two-step approach for account selection leveraging knowledge-informed macro-structure and data-informed micro behaviors.
[PNG File , 22 KB - jmir_v26i1e54506_app1.png]

Multimedia Appendix 2
All accounts categorized by policy group.
[XLSX File (Microsoft Excel File), 22 KB - jmir_v26i1e54506_app2.xlsx]

Multimedia Appendix 3
Policy groups in information dissemination.
[PNG File , 341 KB - jmir_v26i1e54506_app3.png]

Multimedia Appendix 4
List of dietary sodium keywords, English expression terms excluded from tweets, dietary sodium-related hashtags, and food-related keywords.
[DOCX File , 26 KB - jmir_v26i1e54506_app4.docx]

Multimedia Appendix 5
Number of users to interact with within the network for each top influencer.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v26i1e54506_app5.xlsx]

References

1. WHO global report on sodium intake reduction.: World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240069985 [accessed 2023-03-29]
2. Smith-Spangler CM, Juusola JL, Enns EA, Owens DK, Garber AM. Population strategies to decrease sodium intake and the burden of cardiovascular disease: a cost-effectiveness analysis. Ann Intern Med 2010;152(8):481-487 [FREE Full text] [doi: 10.7326/0003-4819-152-8-201004200-00212] [Medline: 20194225]
3. GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2019;393(10184):1958-1972 [FREE Full text] [doi: 10.1016/S0140-6736(19)30041-8] [Medline: 30954305]
4. GBD 2021 Risk Factors Collaborators. Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2162-2203 [FREE Full text] [doi: 10.1016/S0140-6736(24)00933-4] [Medline: 38762324]
5. Zhu Y, Zhang J, Li Z, Liu Y, Fan X, Zhang Y, et al. Association of sodium intake and major cardiovascular outcomes: a dose-response meta-analysis of prospective cohort studies. BMC Cardiovasc Disord 2018;18(1):192 [FREE Full text] [doi: 10.1186/s12872-018-0927-9] [Medline: 30340541]
6. Bibbins-Domingo K, Chertow GM, Coxson PG, Moran A, Lightwood JM, Pletcher MJ, et al. Projected effect of dietary salt reductions on future cardiovascular disease. N Engl J Med 2010;362(7):590-599 [FREE Full text] [doi: 10.1056/NEJMoa0907355] [Medline: 20089957]
7. Joffres MR, Campbell NRC, Manns B, Tu K. Estimate of the benefits of a population-based reduction in dietary sodium additives on hypertension and its related health care costs in Canada. Can J Cardiol 2007;23(6):437-443 [FREE Full text] [doi: 10.1016/s0828-282x(07)70780-8] [Medline: 17487286]

XSL•FO
RenderX

8. Selmer RM, Kristiansen IS, Haglerod A, Graff-Iversen S, Larsen HK, Meyer HE, et al. Cost and health consequences of reducing the population intake of salt. J Epidemiol Community Health 2000;54(9):697-702 [FREE Full text] [doi: 10.1136/jech.54.9.697] [Medline: 10942450]
9. Wang G, Labarthe D. The cost-effectiveness of interventions designed to reduce sodium intake. J Hypertens 2011;29(9):1693-1699 [FREE Full text] [doi: 10.1097/HJH.0b013e328349ba18] [Medline: 21785366]
10. Brouillard AM, Kraja AT, Rich MW. Trends in dietary sodium intake in the United States and the impact of USDA guidelines: NHANES 1999-2016. Am J Med 2019;132(10):1199-1206.e5. [doi: 10.1016/j.amjmed.2019.04.040] [Medline: 31103645]
11. Clarke LS, Overwyk K, Bates M, Park S, Gillespie C, Cogswell ME. Temporal trends in dietary sodium intake among adults aged ≥19 years—United States, 2003-2016. MMWR Morb Mortal Wkly Rep 2021;70(42):1478-1482 [FREE Full text] [doi: 10.15585/mmwr.mm7042a4] [Medline: 34673747]
12. Michael V, You YX, Shahar S, Manaf ZA, Haron H, Shahrir SN, et al. Barriers, enablers, and perceptions on dietary salt reduction in the out-of-home sectors: a scoping review. Int J Environ Res Public Health 2021;18(15):8099 [FREE Full text] [doi: 10.3390/ijerph18158099] [Medline: 34360392]
13. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
14. Cocchi S, Mis CC, Mazzocut M, Barbieri I, Bassi MC, Cavuto S, et al. Content and face validity of the evaluation tool of health information for consumers (ETHIC): getting health information accessible to patients and citizens. Healthcare (Basel) 2023;11(8):1154 [FREE Full text] [doi: 10.3390/healthcare11081154] [Medline: 37107988]
15. Ratzan SC, Parker R, Selden CR, Marcia Z. National library of medicine current bibliographies in medicine: health literacy. 2000. URL: https://www.researchgate.net/publication/230877250_National_Library_of_Medicine_Current_Bibliographies_in_Medicine_Health_Literacy [accessed 2023-06-28]
16. Coleman C, Kurtz-Rossi S, Mckinney J, Pleasant A, Rootman I, Shohet L. Calgary charter on health literacy: Rationale and core principles for the development of health literacy curricula.: The Centre for Literacy; 2010. URL: https://www.centreforliteracy.qc.ca/sites/default/files/CFL_Calgary_Charter_2011.pdf [accessed 2024-05-01]
17. Singhal A, Mago V. Exploring how healthcare organizations use Twitter: a discourse analysis. Informatics 2023;10(3):65. [doi: 10.3390/informatics10030065]
18. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. JCM 2018;7(6):121 [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
19. Ek S, Curci M, Yang X, Lin B, Liu P, Xu H. Sentiment analysis of long-term social data during the COVID-19 pandemic. URL: https://scholarworks.utrgv.edu/cs_fac/71/ [accessed 2021-09-21]
20. Mao L, Chu E, Gu J, Hu T, Weiner B, Su Y. A 4D theoretical framework for measuring topic-specific influence on Twitter: development and usability study on dietary sodium tweets. J Med Internet Res 2023 Jun 13;25:e45897 [FREE Full text] [doi: 10.2196/45897] [Medline: 37310774]
21. Moulton S, Sandfort JR. The strategic action field framework for policy implementation research. Policy Stud J 2017;45(1):144-169. [doi: 10.1111/psj.12147]
22. Consensus statement on sodium. NYC Health Department. URL: https://www1.nyc.gov/assets/doh/downloads/pdf/cardio/consensus-statement.pdf [accessed 2024-11-02]
23. Top social sciences and public health schools in the world—US news education. U.S. News. 2023. URL: https://www.usnews.com/education/best-global-universities/social-sciences-public-health [accessed 2024-04-29]
24. X Glossay. Twitter. 2023. URL: https://help.twitter.com/en/resources/glossary [accessed 2024-11-02]
25. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Stanford InfoLab. 1999. URL: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768 [accessed 2023-07-21]
26. Hagberg A, Schult D, Swart P. Exploring network structure, dynamics, and function using NetworkX. 2008 Jan 01 Presented at: SciPy2008 Conference; August 19-24, 2008; Pasadena, CA p. 11-15 URL: https://www.osti.gov/servlets/purl/960616 [doi: 10.25080/TCWV9851]
27. Montoya A, Mao L, Drewnowski A, Chen J, Shi E, Liang A, et al. Interactions among top users on dietary sodium reduction on Twitter (X). GitHub. 2024. URL: https://lingchm.github.io/sodium_twitter_influence [accessed 2024-11-12]
28. Global action plan for the prevention and control of noncommunicable diseases 2013-2020. World Health Organization. 2013 Nov 13. URL: https://www.who.int/publications/i/item/9789241506236 [accessed 2024-12-10]
29. Tom Frieden. URL: https://www.linkedin.com/in/tom-frieden/ [accessed 2023-06-25]
30. US Department of Agriculture, US Department of Health and Human Services. Dietary guidelines for Americans, 2020-2025, 9th edition. DietaryGuidelines.gov. 2020 Dec. URL: https://www.dietaryguidelines.gov/resources/2020-2025-dietary-guidelines-online-materials [accessed 2024-05-02]
31. Korsen D. Sodium and potassium dietary reference intake values updated in new report; introduces new category for sodium based on chronic disease risk reduction. National Academies of Sciences, Engineering, and Medicine. 2019. URL: https://www.nationalacademies.org/news/2019/03/sodium-and-potassium-dietary-reference-intake-values-updated-in-new-report [accessed 2023-10-22]

XSL•FO
RenderX

32. Eastern mediterranean salt reduction. 2023. URL: https://www.worldactiononsalt.com/resources/salt-and-sugar-reduction/eastern-mediterranean-salt-reduction-toolkit/ [accessed 2024-11-02]
33. FDA urged to reduce Americans' sodium consumption by setting 10-year reduction targets in food. Center for Science in the Public Interest. 2023. URL: https://www.cspinet.org/press-release/fda-urged-reduce-americans-sodium-consumption-setting-10-year-reduction-targets-food [accessed 2023-07-26]
34. National Nutrition Month®. FDA. 2024 Mar 06. URL: https://tinyurl.com/4znfxw24 [accessed 2023-02-27]
35. Salt awareness week. URL: https://www.worldactiononsalt.com/awarenessweek/salt-awareness-week/ [accessed 2023-06-25]
36. Lallukka T, Laaksonen M, Rahkonen O, Roos E, Lahelma E. Multiple socio-economic circumstances and healthy food habits. Eur J Clin Nutr 2007;61(6):701-710. [doi: 10.1038/sj.ejcn.1602583] [Medline: 17180154]
37. Sarmugam R, Worsley A, Wang W. An examination of the mediating role of salt knowledge and beliefs on the relationship between socio-demographic factors and discretionary salt use: a cross-sectional study. Int J Behav Nutr Phys Act 2013;10(1):25 [FREE Full text] [doi: 10.1186/1479-5868-10-25] [Medline: 23418906]
38. Klein KJ, Sorra JS. The challenge of innovation implementation. Acad Manage Rev 1996;21(4):1055-1080 [FREE Full text] [doi: 10.5465/amr.1996.9704071863]
39. Weiner BJ, Lewis MA, Linnan LA. Using organization theory to understand the determinants of effective implementation of worksite health promotion programs. Health Educ Res 2009;24(2):292-305. [doi: 10.1093/her/cyn019] [Medline: 18469319]
40. Xin W, Wang K, Fu Z, Zhou L. Let community rules be reflected in online content moderation. ArXiv. Preprint posted online on August 21, 2024 2024 [FREE Full text]
41. Yang Q, Li P, Xu X, Ding Z, Zhou W, Nian Y. A comparative study on enhancing prediction in social network advertisement through data augmentation. 2024 Presented at: 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE); June 28-30, 2024; Zhuhai, China p. 214-218. [doi: 10.1109/mlise62164.2024.10674203]
42. Cruden G, Crable EL, Lengnick-Hall R, Purtle J. Who's "in the room where it happens"? A taxonomy and five-step methodology for identifying and characterizing policy actors. Implement Sci Commun 2023;4(1):113 [FREE Full text] [doi: 10.1186/s43058-023-00492-6] [Medline: 37723580]
43. X developer platform. About the Twitter API. 2023. URL: https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api [accessed 2023-10-22]

Abbreviations

AHA: American Heart Association
CDC: Centers for Disease Control and Prevention
CSPI: Center for Science in the Public Interest
CVD: cardiovascular disease
FDA: US Food and Drug Administration
NYC-DOH: New York City Department of Healthand Mental Hygiene
RTSL: Resolve to Save Lives
USDA: US Department of Agriculture
WASSH: World Action on Salt, Sugar, and Health
WHO: World Health Organization

Edited by A Mavragani; submitted 12.11.23; peer-reviewed by E de Quincey, E Vashishtha, X Cheng; comments to author 08.04.24; revised version received 29.04.24; accepted 16.10.24; published 30.12.24.

Please cite as:
Montoya A, Mao L, Drewnowski A, Chen J, Shi E, Liang A, Weiner BJ, Su Y
Influencers in Policy Fields on Social Media: Global Longitudinal Study of Dietary Sodium Reduction Posts, 2006-2022
J Med Internet Res 2024;26:e54506
URL: https://www.jmir.org/2024/1/e54506
doi:10.2196/54506
PMID:

©Alana Montoya, Lingchao Mao, Adam Drewnowski, Joshua Chen, Ella Shi, Aileen Liang, Bryan J Weiner, Yanfang Su. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Information and Communication Technology for Managing Social Isolation and Loneliness Among People Living With Parkinson Disease: Qualitative Study of Barriers and Facilitators

Gomathi Thangavel[1], MSc; Mevludin Memedi[1,2], PhD; Karin Hedström[1,3], PhD

[1]Centre for Empirical Research on Information Systems, Örebro University School of Business, Örebro, Sweden

[2]College of Business, Alfaisal University, Riyadh, Saudi Arabia

[3]Department of Communication, Quality Management and Information Systems, Mid Sweden University, Sundsvall, Sweden

Corresponding Author:
Gomathi Thangavel, MSc
Centre for Empirical Research on Information Systems
Örebro University School of Business
Fakultetsgatan 1
Örebro, 70281
Sweden
Phone: 46 19302471
Email: gomathi.thangavel@oru.se

Abstract

Background: Parkinson disease (PD) is a complex, noncurable, and progressive neurological disease affecting different areas of the human nervous system. PD is associated with both motor and nonmotor symptoms, which negatively affect patients' quality of life and may cause changes in socialization such as intentional social withdrawal. This may further lead to social isolation and loneliness. The use of information and communication technology (ICT) plays an important role in managing social isolation and loneliness. Currently, there is a lack of research focusing on designing and developing ICT solutions that specifically address social isolation and loneliness among people living with PD.

Objective: This study addresses this gap by investigating barriers and social needs in the context of social isolation, loneliness, and technology use among people living with PD. The insights gained can inform the development of effective ICT solutions, which can address social isolation and loneliness and improve the quality of life for people living with PD.

Methods: A qualitative study with 2 phases of data collection were conducted. During the first phase, 9 health care professionals and 16 people living with PD were interviewed to understand how PD affects social life and technology use. During the second phase, 2 focus groups were conducted with 4 people living with PD in each group to gather insights into their needs and identify ways to manage social isolation and loneliness. Thematic analysis was used to analyze both data sets and identify key themes.

Results: The results showed that the barriers experienced by people living with PD due to PD such as "fatigue," "psychological conditions," "social stigma," and "medication side effects" affect their social life. People living with PD also experience difficulties using a keyboard and mouse, remembering passwords, and navigating complex applications due to their PD-related physical and cognitive limitations. To manage their social isolation and loneliness, people living with PD suggested having a simple and easy-to-use solution, allowing them to participate in a digital community based on their interests, communicate with others, and receive recommendations for social events.

Conclusions: The new ICT solutions focusing on social isolation and loneliness among people living with PD should consider the barriers restricting user's social activities and technology use. Given the wide range of needs and barriers experienced by people living with PD, it is more suitable to adopt user-centered design approaches that emphasize the active participation of end users in the design process. Importantly, any ICT solution designed for people living with PD should not encourage internet addiction, which will further contribute to the person's withdrawal from society.

(J Med Internet Res 2024;26:e48175) doi:10.2196/48175

KEYWORDS

social isolation; loneliness; Parkinson disease; ICT; information and communication technology

XSL•FO
RenderX

Introduction

Background

Parkinson disease (PD) is a complex, noncurable, and progressive neurological disease affecting different areas of the human nervous system. PD negatively impacts the ability of people to perform their daily activities. PD is most notably characterized by visible motor symptoms such as rigidity, tremor, slowness of movements, and impaired balance and coordination while walking [1-3]. PD is also associated with nonmotor symptoms such as sleep disturbances, problems with swallowing, communication issues, anxiety, and depression [4]. These motor and nonmotor symptoms cause changes in socialization such as intentional social withdrawal, that is, avoiding social participation and interactions with people [5,6]. These behavioral shifts may be prompted by various factors, including decreased facial expressions and slurred speech and may lead to social isolation and loneliness [7]. Both social isolation and loneliness contribute to reduced quality of life and well-being [8,9]. Social isolation refers to the objective lack of social connections, while loneliness is the subjective state characterized by negative feelings about having a lower level of social contact than desired [10]. Hence, understanding the complex interplay between motor and nonmotor symptoms and their impact on social interactions is crucial and may have a profound effect on quality of life of people living with PD.

PD predominantly affects individuals aged older than 50 years, and mean age of onset is around 60 years [1,11]. Approximately 1% of people aged older than 60 years and more than 3% of those aged older than 80 years are affected by this disease [1,12]. Currently, clinical examinations are done at infrequent time intervals during hospital clinical visits. Such examinations provide only a snapshot of PD symptoms at that point in time and do not completely reflect a patient's daily living, including his or her social activities. Unfortunately, the medical community has a limited understanding of social isolation and loneliness and their impact into the health of their patients and lacks effective interventions to evaluate and integrate them into routine clinical practice [13].

The use of information and communication technology (ICT) plays an important role in managing social isolation and loneliness [14,15]. For example, videoconferencing platforms such as Zoom (Zoom Video Communication, Inc) and Skype (Skype Technologies [Microsoft]), allow individuals to connect with their friends, family, and colleagues in real time through video calls. Similarly, social media platforms such as Facebook (Meta Platforms, Inc) and X, previously named Twitter (X Corp), provide a space for people to connect with their loved ones remotely. It is, however, important to recognize that one-size-fits-all solutions may not be effective. Instead, solutions need to more tailored to the specific needs of different individuals and groups. The diverse range of specific needs, experiences, and preferences must also be met within these groups [16,17].

ICT solutions targeting people living with PD are mainly developed to monitor the disease, assess symptoms to provide optimal and individualized treatments [18-20], and empower the users to manage their disease [21]. A few more studies also exist on analyzing online discussion forums for people living with PD [22,23]. Further, few studies focused on improving the quality of life and well-being among people living with PD [24,25]. There needs to be more research focusing on designing and developing customized ICT solutions that specifically address social isolation and loneliness among people living with PD. Additionally, due to PD symptoms, some people have issues accessing online peer support or using technology [22]. So, it is important to understand the barriers faced by people living with PD regarding socialization and technology use while developing an ICT solution for managing social isolation and loneliness.

Literature Review

There are different barriers associated with PD symptoms that may cause social isolation and loneliness. A scoping review study analyzed the existing evidence regarding social withdrawal in PD and found that physical, cognitive, and psychiatric symptoms of PD and perceived stigma have reduced social activities among people living with PD [7]. The effects of PD on people's social interactions were explored through in-depth interviews conducted in Tehran by Soleimani et al [26], who found out that "progressive physical disability," "mood disturbances," "shrinking of social activities," and "secluding oneself" caused social disconnections. Similarly, another study investigated the impact of living with PD among 19 participants and found that participants compromised their social activities due to embarrassment of their changed persona and speech problems [27]. This embarrassment experienced by people living with PD were reviewed by Maffoni et al [28]. They reviewed 14 papers and synthesized social stigma into 4 constructs: "stigma arising from symptoms," "stigma linked to relational and communication problems," "social stigma arising from sharing perceptions," and "caregiver's stigma." The construct "stigma arising from symptoms" includes not only the embarrassment due to visible physical symptoms but also the progressive loss of functionality. Similarly, stigma linked to relational and communication problems includes oral communication issues (hoarseness and delayed thinking process) and body gesture issues (rigid and unexpressive face). Furthermore, Maffoni et al [28] classified the construct of "social stigma arising from sharing perceptions" into two distinct categories: (1) "perception exchange of others toward the patient" and (2) "perception exchange of patient toward the other." The former one deals with the people living with PD considering other's beliefs about their physical and mental status, whereas the latter one includes the perspective from people living with PD that they are a burden to their family due to a change or loss of social roles. Finally, the last construct includes the stigma experienced by family caregivers due to their ill family member's embarrassing and visible symptoms [28].

Concerning ICT solutions that support social connection for people living with PD, an online support group was introduced during COVID-19 period through a series of weekly lectures by expert neurologists using Zoom [24]. The main goal of the support group was to improve the quality of life and prevent the worsening of PD symptoms by keeping the patients

XSL•FO
RenderX

connected, educated, and empowered. Furthermore, Boulos et al [25] developed a web platform (LiveWell) to enable people living with PD to manage their health condition and promote well-being through social communities. In the platform, the main feature of social communities was designed as a discussion forum among people living with PD and facilitated interaction with carers and clinicians. A study in the United Kingdom analyzed different online discussion forums and communities for people living with PD and their caregivers [23]. They found that the discussion forums helped participants to share their knowledge and develop friendships. Further, a systematic review by Gerritzen et al [22] revealed that online discussion forums, either through the web or Facebook groups, keep people living with PD socially connected, create awareness of PD, and exchange social support among them. All the above studies, except Subramanian [24], did not focus on addressing social isolation and loneliness issues for people living with PD, despite supporting social connections. Due to this limitation, there was also a lack of knowledge about user needs for managing social isolation and loneliness.

Previous studies about technology solutions for social isolation and loneliness highlight the need for customized ICT solutions to suit the needs of individuals, specific groups, or degrees of social isolation and loneliness [16,17]. A review by Thangavel et al [17] analyzed customized ICT solutions targeting social isolation and loneliness among older adults. They found that there are different aspects contributing to reduced social isolation or loneliness among older adults using ICT solutions. Such aspects include social communication, social participation, a sense of belonging, companionship, and feelings of being seen. Increasing social communication through ICT helps older adults to enhance their social connections, while ICT that promotes social participation enables older adults to engage in social activities with others [17]. Further, increasing a sense of belonging, companionship, and feelings of being seen through ICT fulfils older adults' emotional relationships and reduces their loneliness [17]. Given that PD is more commonly diagnosed aged older than 50 years, the findings from the review on customized ICT solutions targeting social isolation and loneliness among older adults may have implications for developing similar solutions for people living with PD.

Objective

This study investigates the multifaceted barriers experienced by people living with PD in the context of social isolation, loneliness, and technology use. Furthermore, this study also investigates the social needs of people living with PD, which can facilitate the design and development of potential ICT solutions to manage social isolation and loneliness among people living with PD.

Therefore, the study aims to address the following research questions (RQs).

1. What are the barriers experienced by people living with PD in the context of social isolation and loneliness?
2. What are the barriers experienced by people living with PD in relation to their technology use?
3. What are the social needs of people living with PD to facilitate the design and development of ICT solutions for managing social isolation and loneliness?

The expected results of this study are twofold. First, a comprehensive understanding of the barriers and challenges faced by people living with PD in the context of social isolation, loneliness, and technology use will be obtained. This knowledge will inform the development of customized ICT solutions and strategies that effectively target these specific challenges. Second, the study will identify the social needs of people living with PD, enabling the formulation of an evidence-based framework for designing ICT solutions that cater to these needs. Through these outcomes, we aim to provide valuable insights that can lead to enhancing the well-being and social connections of people living with PD.

Methods

Study Scope and Framework

This qualitative study extensively examined the multifaceted barriers encountered by people living with PD in the context of social isolation, loneliness, and technology use. With people living with PD as the main focus, health care professionals' perspectives were also considered to gain a holistic understanding of the disease's impact. By encompassing these diverse participant categories, a thorough exploration of the disease's impact was achieved. The research, carried out in Sweden, aimed to shed light on the social experiences of people living with PD. Furthermore, the study explored the social needs of people living with PD, which could serve as a foundation for potential ICT solutions aimed at alleviating social isolation and loneliness. The study framework depicted in Table 1 encompasses the phases of research, the categories of participants involved as sources of data, study objectives, research methods for data generation, and an approach to analyze the generated data, providing a comprehensive overview of the research process. In total, 2 distinct phases of data collection were undertaken to achieve the study objectives. In the initial phase, our focus centered around understanding the diverse barriers arising from PD diagnosis and its implications on social isolation and loneliness through interviews with people living with PD and health care professionals. Additionally, we explored the impact of PD on the technology use of people living with PD. Subsequently, the second phase involved facilitating focus groups with people living with PD to identify their social needs, providing insights crucial for the development of strategies to alleviate social isolation and loneliness.

XSL•FO
RenderX

Table 1. Study framework.

Phases of research and participant category	Period	Objectives	Research method	Data analysis approach
Phase 1: health care professionals	November 2019	• To understand how PD[a] affects a person's social life and quality of life	Interview	Thematic analysis
Phase 1: people living with PD and relatives	April to June 2021	• To understand the barriers of being diagnosed with PD and how that relates to social isolation and loneliness. • To understand how PD affects the technology use of people living with PD	Interview	Thematic analysis
Phase 2: people living with PD	January 2022	• To understand the social needs of people living with PD to facilitate the design and develop ICT[b] solutions for managing social isolation and loneliness	Focus group discussions	Thematic analysis

[a]PD: Parkinson disease.

[b]ICT: information and communication technology.

Study Participants and Recruitment Procedure

The health care professionals who took part in the study included physicians, nurses, and rehabilitation team (physiotherapists, occupational therapists, and social workers). These professionals were actively associated with the neurology department at a university hospital in Sweden, where they provided dedicated care and support to people living with PD. To ensure a comprehensive and insightful participation of health care professionals, a neurologist who was part of a previous study played a pivotal role. The neurologist assisted us in sending invitations to coworkers within the neurology department. Furthermore, the neurologist shared contact information of interested health care professionals and facilitated the scheduling of interviews based on each expert's availability.

We extended invitations to people living with PD who met the specific inclusion criteria, which included being aged older than 50 years, having been diagnosed with PD, and demonstrating the ability to understand and express willingness to sign the informed consent form. People living with PD with severe cognitive problems, which could potentially impact their ability to provide reliable responses, were excluded from participation. Further, people living with PD were also invited to bring relatives or another close person if needed. The information from those relatives was used to complement the information given by the people living with PD. Invitations to participate in the study were facilitated through the national Parkinson association, which played a crucial role in identifying and reaching out to eligible individuals who fulfilled the study's inclusion and exclusion criteria. The association's support was instrumental in connecting us with people living with PD from 2 different regions in Sweden.

Data Collection

Interviews With Health Care Professionals

Semistructured interviews were conducted with open-ended questions related to PD focusing on the quality of life and social aspects. In addition, general questions related to treatments or therapies and ICT used in relation to PD were included (see the interview guide in Multimedia Appendix 1). The interviews with physicians and nurses were conducted individually, whereas the interview with the rehabilitation team were a group interview, due to their work nature as a team. In total, the interviews were conducted with 2 physicians, 2 nurses, 2 occupational therapists, 2 physiotherapists, and 1 social worker. The interviews took place at the university hospital and were recorded with their consent; on average, each interview lasted approximately 45 (SD 8.62) minutes.

Interviews With People Living With PD

Individual interviews with people living with PD were conducted during the period from April to June 2021. Due to the COVID-19 pandemic, physical meetings were not possible, so online interviews were conducted via Zoom. For the participants who did not have previous experience with Zoom, we also sent a link with information about how to join Zoom meetings. We also gave telephone numbers and meeting IDs to connect to the Zoom meetings. Further, 2 participants connected through this feature. Questions were asked about their background, challenges of living with PD, social isolation and loneliness, technology use, and issues. Questions related to social isolation and loneliness were adapted from Cornwell and Waite [29] since their study included both isolation and loneliness aspects (see the interview guide in Multimedia Appendix 1). As the interview took place during the COVID-19 pandemic, the discussion did not center around their experiences of socializing amid the pandemic. Instead, the people living with PD were requested to share their perspectives on how PD had impacted their socialization. Interviews were conducted in the language (English or Swedish) based on the preference and comfort of the participants. In total, 18 interviews were conducted, each taking 35 (SD 6.06) minutes on average. All the interviews were recorded with the participants' consents. However, it's important to note that 1 participant chose to withdraw from the study after their interview.

Focus Group With People Living With PD

As a follow-up, 2 focus group discussions were conducted with people living with PD in January 2022 to understand how they

XSL•FO
RenderX

would manage or would like to manage social isolation or loneliness with the use of ICT. Invitations were sent to all the people living with PD who participated in the initial interviews. Based on their acceptance and availability, the focus groups were arranged. In total, 4 individuals participated in each group using Zoom. Further, 1 group preferred the Swedish language, while the other chose English. Each group discussion was recorded with their consent and lasted about 1 hour and 30 minutes. Discussions were concentrated around 3 topics: coping mechanisms for social isolation and loneliness; social communication, social participation, and sense of belonging; and technology barriers and needs. These topics were selected based on the insights from the results obtained during initial individual interviews with people living with PD and findings from a previous literature review [17].

Data Analysis

The audio-recorded data were transcribed, and later, the Swedish interviews were translated into English because the main author needed help in understanding Swedish. The interview data was analyzed using an Excel (Microsoft) sheet following the thematic analysis process [30]. Initially, transcripts were manually coded line by line based on the content by the first author. For example, codes such as challenges, diagnosis, tiredness, anxiety, apathy, embarrassment, stigma, social connection, social network, events information, and awareness were formed. The codes were then sorted and collated based on the RQs, that is, barriers to socialization, barriers to technology use, and needs for managing social isolation and loneliness. Subsequently, the sorted codes under each RQ were further categorized into themes by inductive and deductive approaches as shown below by the first author.

The codes linked to the RQ1 and RQ2 were further categorized by organizing the codes together based on the relationship between codes. For example, apathy and anxiety were put together since both are related to the psychological conditions of a person. The part related to stigma is further categorized using the social stigma model from Maffoni et al [28].

For the codes related to RQ3, we classified using the aspects identified by Thangavel et al [17], which include social communication, social participation, and a sense of belonging. For example, events information is categorized under the social participation theme since information about events enhances social activities with others.

Finally, all the authors reviewed and refined the themes until a consensus was reached. For example, under RQ2, the code “learning difficulties” was originally mapped under the theme “ageing concern.” However, through our collaborative review, we collectively agreed that this code would be more appropriately categorized under “cognitive limitations.” Further, we also extracted the quotes that capture each theme’s essence.

Ethical Considerations

This study followed the Swedish ethical guidelines and received approval from the Swedish Ethical Review Authority (2020-05855). For all the participants, the study was fully explained orally and in written form, and their written or digital informed consent was obtained before the data collection. Participation in the study was entirely voluntary and no compensation was provided for their involvement. Further, the participants had the right to withdraw from the study at any time, without giving any reasons.

Results

Participants’ Characteristics

Information about the people living with PD and health care professionals who participated in this study is described in Tables 2 and 3, respectively. The average age of people living with PD in this study is 67 (SD 5.88) years. Out of 17 interviews, 16 were people living with PD and 1 was a relative. This 1 relative represented her recently deceased mother, diagnosed with PD for 17 years. Apart from that, 2 other relatives (spouse) joined along with people living with PD. At the time of the interview with people living with PD, the participants had been diagnosed with PD for an average of 8 (SD 4.52) years. Further, 4 people living with PD were living alone and others were with their spouses.

XSL•FO
RenderX

Table 2. Characteristics of people living with PD[a] (n=16).

Characteristics	Value, n (%)
Age group (y)	
56-60	3 (19)
61-65	5 (31)
66-70	3 (31)
>70	5 (19)
Gender	
Women	7 (44)
Men	9 (56)
Living situation	
Alone	4 (25)
With partner	12 (75)
Working status	
Retired	13 (81)
Working	3 (19)
Years since diagnosis	
≤5	6 (38)
6-10	4 (25)
11-15	6 (38)

[a]PD: Parkinson disease.

Table 3. Characteristics of the health care professionals (n=9).

Health care professionals	Value, n (%)
Physicians	2 (22)
Nurses	2 (22)
Rehabilitation team	5 (56)

Emerged Themes

The outcomes of the interviews and focus group discussions yielded distinct themes in response to each RQ. For RQ1, the identified themes encompassed “fatigue,” “psychological conditions,” “social stigma,” and “medication side effects.” Notably, the subthemes within the “social stigma” category were developed following the social stigma model proposed by Maffoni et al [28]. Furthermore, addressing RQ2, the exploration of barriers associated with the use of ICT revealed themes centered around “physical limitations” and “cognitive limitations.” Lastly, RQ3 centered around the user needs for managing social isolation and loneliness, were categorized into “increase social connections,” “increase social participation,” “increase a sense of belonging,” and “other suggestions.” Some of the example codes associated with each theme are described in Table 4.

XSL•FO
RenderX

Table 4. Themes, subthemes, and codes.

RQ[a], themes, and subthemes	Example codes
1. Barriers in the context of social isolation and loneliness	
Fatigue	Energy depletion, slowed movements, and extreme tiredness
Psychological conditions	Anxiety (fear about future and excessive worry) or apathy (lack of initiative)

Social stigma
• Arising from symptoms
• Due to relational and communicational problems
• Arising from sharing perceptions
  • Perception of others toward people with PD[b]
  • Perception of people with PD toward others
	Embarrassment, shame, and stigma

Medication side effects	Gambling, hallucinations, and impulse control disorder
2. Barriers in relation to use of ICT[c]	
Physical limitations	Difficulties using keyboard and mouse
Cognitive limitations	Not remembering passwords, complex to navigate, and learning difficulties
3. User needs for managing social isolation and loneliness	
Increase social connections	Connect similar interest people, video chat, and group establishment
Increase social participations	Events information, create event, and recommendations
Increase a sense of belonging	Awareness, community forum, information from experts, and tips and suggestions
Other suggestions	Simple, easy-to-use, and technology guidance

[a]RQ: research question.

[b]PD: Parkinson disease.

[c]ICT: information and communication technology.

Barriers in the Context of Social Isolation and Loneliness

Overview

Results from the interviews with people living with PD show that most people living with PD withdraw from socializing with others due to PD and its symptoms. In total, 13 people living with PD reported being socially isolated; 10 people living with PD also expressed feeling lonely at some point of time due to PD. Further, 3 out of 4 people living with PD who live alone were feeling lonely. Apart from this, the following factors which cause social withdrawal were identified.

Fatigue

One of the main causes identified by health care professionals and people living with PD is extreme tiredness due to PD. Therapists pointed out that this was attributed to the motor fluctuations experienced by the patients on that particular day.

> *If they are tired due to physical and cognitive fluctuations, it will obstruct their routine and plans...one day they feel really good, and the next day they decide to eat food at a restaurant but motor fluctuations on that day will make them tired and stop them from going out.* [Physiotherapist 1]

Most people living with PD also expressed feeling exhausted due to their deteriorating physical and mental abilities. Due to this experience, they avoided spontaneous meetings with others. Additionally, they expressed that they need to plan and organize everything based on their bodily needs.

> *I'm not able to decide myself when I want to meet people, Parkinson is deciding how I am able to go to a social meeting. The disease determines when and how and what I should be able to do. I am being imprisoned by the disease more and more.* [Person living with PD 11]

Psychological Conditions

People living with PD also experience an on-off state, that is, a switch between mobility and immobility occurs at the end of the dose of medication. During the off state, motor functions will deteriorate and concurrently, the mental well-being of people living with PD is significantly affected, leading to increased anxiety. The challenges presented by this fluctuation contribute to the development of psychological conditions, notably apathy (lack of initiative) and anxiety. For some people living with PD, these symptoms can result in a loss of interest in activities and a reluctance to engage in social interactions and outings.

XSL•FO
RenderX

> *...can also have mental on and mental off which is ability to think and no ability to think or being really anxious or not anxious. So, this can fluctuate with levels of medication. This can sufficiently repair their ability to engage in social activities.* [Physician 1]

> *...Yes, she isolated. She had friends, but she didn't call them. She wouldn't pick up the phone and I took friends to her house for coffee and then she enjoyed it and then it worked, but she would not take the initiative herself.* [Relative 1]

Due to the nature of this disease, they worry a lot about their balance while they walk. They are not going out thinking they cannot come back without other's help.

> *Some patients get physically isolated due to anxiety they get that they cannot go back home on their own.* [Physician 2]

Further, this anxious nature also prevents people living with PD from meeting others diagnosed with PD. They worry that when they see others who have worse symptoms, they will have those symptoms later.

> *I get more depressed from going to the other meetings where you see those that are significantly worse. And actually, maybe it's the same for them. They can imagine: "Why have I got this, and am not as healthy as the other?" So, it's kind of not good for anyone.* [Person living with PD 5]

> *She didn't want to see how sick they were, she was afraid because some of them...of course some were better than her, but some were terribly ill and that frightened her.* [Relative 1]

Social Stigma

Social Stigma Arising From Symptoms

Some of the most common visible symptoms of PD are tremors and drooling. A tremor (shaking) is an involuntary movement affecting part of their body, particularly the hands. Due to these visible motor symptoms, people living with PD expressed that they are embarrassed to show their symptoms in front of others.

> *I am very fond of going to the theatre, the opera, concerts, but sometimes I feel, no I can't sit among people when my arms start shaking, so I am afraid that everybody can see. And it is difficult to concentrate on the music or whatever.* [Person living with PD 14]

This symptom makes people living with PD stay at home and avoid going out. Health care professionals also confirmed that this complaint is the biggest one they receive from their patients when enquiring about their social activities.

> *There are like emotions such as shame that can stop people from engaging. If you shake and you feel shame of that in a social situation, you may stop being part of the family dinner. If you start drooling because you don't think of swallowing your saliva, you may avoid people, so people cannot see drooling and people think of you like the person you were in your better days. So, people do some social hiding there too.* [Physician 1]

Apart from the visible symptoms, some people living with PD felt uncomfortable because of their progressive loss of functionality and autonomy. Further, they felt bad about themselves and were not able to ask for help easily if they cannot do certain things like cutting meat or putting on shoes.

> *...I went by myself to a restaurant, and then I get the main course in and it's a piece of meat, and I can't cut the meat. And then I felt like this: "Yes, it was not so easy to ask for help." Yes, I've thought it is. I've always said this: "Yes, but if I don't know anything, then I'll ask someone for help." But it wasn't that easy.* [Person living with PD 6]

Social Stigma and Relational and Communication Problems

Further, during the PD off state, some people living with PD mentioned that their thinking speed had decreased, making it challenging to find the right words to communicate with others.

> *That they don't listen. They don't hear what I say and that they don't listen as much as before, on what I have to say. That's the negative. And for my wife it was much worse, because people didn't hear what she said and then when she said something, they either ignored it or they had to ask what she said, or the other was silent just to try to figure out what she had said, or someone could hear what she said. So, it was...that meant that she was more and more silent in groups, yeah, at the end there.* [Person living with PD 10]

The nurses and the therapist also mentioned that some patients have swallowing issues and difficulty eating due to PD, making them not go to restaurants with their family or friends.

> *Some patients have problems with talking, they have soft or hoarse voice, and very hard time to speak. And some patients have hard to swallow and that's the reason they won't go out to restaurants or family dinner or eat with other people.* [Nurse 2]

Social Stigma Arising From Sharing Perceptions

Overview

This category was also one of the major concerns raised by people living with PD. There were two different perceptions shared: (1) perception of others toward people living with PD, and (2) perceptions of people living with PD toward others (caregivers and family).

Perception of Others Toward People Living With PD

The people living with PD expressed that, opinions from their family or friends such as "it's a terrible disease" or "you don't fit in" make them upset and unwilling to socialize with their loved ones. Due to this perception, some people did not feel comfortable sharing information about PD with their friends.

> *...They don't invite...they didn't invite us, and don't invite me for some activities where they think we don't fit in because of the Parkinson. For example, some party with dance and so on, so I think naturally they*

XSL•FO
RenderX

> *choose which ones fit in to a party, and then they think we were less attractive.* [Person living with PD 10]

> *I think it's a little bit embarrassing to tell new friends about this. My old friends, they know all about me, it's not embarrassing at all, but if I meet a new...friend and then...they don't know what Parkinson is and they don't really understand.* [Person living with PD 11]

Further, in certain cases, their family and friends are not able to understand their illness and they avoid asking the people living with PD about their health condition.

> *...most of the time I put in tablets and stuff like that to be able to meet them, but they can't, or don't really understand my illness and this. And they don't want to talk about it, don't want to know. My son he says only briefly: 'How are you? Good, or bad', like. And they don't want to know any more, and it's a little frustrating sometimes.* [Person living with PD 4]

Perception of People Living With PD Toward Others

Some people living with PD feel they have lost their personal identity. They want to be treated normally, just like they were treated before their diagnosis. Their self-esteem was shattered by others in the family. Their sense of belonging was lost in some cases. Further, people living with PD feel they burden their family since they depend on others for everything.

> *...It affects their psyche too, that I'm not feeling well. I lie down on the floor, they pick the yoga mat, they pick the ball. I would like to...have a separate room, but it is not possible, to isolate me when I exercise, and that they do not see me suffer.* [Person living with PD 8]

> *No matter what you suffer from, it is well that the surroundings have difficulty relating to how to treat the person who has received something, Parkinson's or something else. But I want to keep my identity and not become a Parkinson patient...I want everybody to treat me as a person not as person with Parkinson.* [Person living with PD 17]

Medication Side Effects

Health care professionals highlighted that, at an advanced stage, due to more dosages of medications, there are side effects like hallucinations, behavioral problems, gambling addiction, and sex addiction.

> *...they have confusions or hallucinations due to the side effects of medicines. We get complaints from their partners. Because patients get suspicious if their partners go out and it affects their relationship.* [Nurse 1]

Due to these side effects, health care professionals further emphasized that patients may have problems with their life partners and their family members. Further, they expressed that some patients may be more prone to gambling or online shopping due to high dopamine effects. This impulsive control disorder and gambling makes the person withdraw from their social circle.

> *Due to increase in dopamine, they have side effects. They want to do some things repeatedly or totally focusing on something or become hyper – if they get some rewards. Some may lose money in gambling; some may do more shopping online.* [Physician 2]

Barriers in Relation to the Use of ICT

Overview

In addition to the difficulties faced by people living with PD in socializing, there are also obstacles related to their PD symptoms and aging that impede their ability to use ICT. These barriers are discussed below.

Physical Limitations

Due to Parkinson tremors and shaking symptoms, some people living with PD expressed difficulties in using a keyboard and mouse. With the shaking and tremor symptoms, keys were pressed wrongly, and sometimes the same key was pressed for a long time.

> *And then not only that, but you are fiddling with all the other letters as well, even if you don't want to at the same time, I think. That's a problem, I think too.* [Person living with PD 6]

Due to this issue, they needed help to complete the task. In the case of mouse handling, it was more difficult since the precision of the mouse pointer is important when they want to click something on the screen.

> *I don't shake so much now, but I...sometimes when I shake, when I'm in an off- or on-state, to run the mouse is like whisking your cream to morning coffee...To move the mouse exactly when it's a millimetre is hell sometimes.* [Person living with PD 16]

Furthermore, many people living with PD are unaware of accessibility options for keyboards such as "filter keys," and "slow keys" to ignore repeated keys and slow down. Further, most of them do not know about alternative options like voice control. Some people living with PD expressed that it was difficult for them to change or use those options, even if they were aware.

> *I think it's possible that you can set on the keyboard how hard you have to press to make something happen. But it's not something I can do, and it's something like that that would make it easier, so that... it should be touch controls, of course, but it shouldn't be that it's enough to breathe on them to trigger a response.* [Person living with PD 3]

> *You mention voice control now, I haven't used it at all, I don't know there's anything like that. So that it is a small obstacle.* [Person living with PD 15]

Cognitive Limitations

Some people living with PD expressed that they could not remember codes or passwords because of cognitive decline symptoms. This situation also makes them anxious in some situations. Due to that, some people living with PD also needed

XSL•FO
RenderX

help when the application was too complex to navigate. They explicitly stated that they were looking for help.

> *...passwords get very complicated. So, you need some support to make it simple and also to have someone to help out...I am getting a little bit nervous, and I fear that something will happen, that I would lose something, some information or...lock the data...the computer and so on.* [Person living with PD 15]

> *Whatever solution you have digitally...if it is complex...You need help sometimes. You need someone to ask.* [Person living with PD 17]

Some people living with PD also expressed that learning a new technology is difficult due to aging. They also conveyed that recent technology developments are not user centered and did not consider their requirements while designing new technologies.

> *...it's more difficult to learn about new technology today, if I compare for...when I was younger. But I think that's...maybe...I don't think it's not just about Parkinson as much as the age.* [Person living with PD 7]

> *...those who make these computers and technologies, they are not...the users. They should ask us who use these devices, ask us what we need and how and so on, but they are not interested in other than their own investors.* [Person living with PD 14]

User Needs for Managing Social Isolation and Loneliness

Overview

In order to manage their social isolation and loneliness, people living with PD offered several suggestions that could guide the development of relevant ICT solutions.

Increase Social Connections

The people living with PD wanted a simple and easy-to-use digital solution that could help them connect with other people with similar interests or hobbies and establish themselves as a group.

> *...so it's easy to pick up and see where you can find people that have the same interest as you have, for example golf, tennis, or cars or whatever. That could be...easy way to find new friends...If you can find other people that way, as a...maybe you can...easier to take new contact with new people and...learn from them as well...and share experience.* [Person living with PD 7]

In that way, they can communicate about their interests, share their understanding of what they know, and learn from others. In the long run, this group establishment can also help people living with PD who are lonely to rely on and interact with someone in the group whenever they want to have an emotional connection.

> *It's this that you have someone you can talk to when needed, huh, that when you feel lonely you would have someone to call. Oh, I do, so I don't really miss it, but maybe that's what's most important, that you have someone to talk to.* [Person living with PD 2]

Further, most of them recommended that digital solutions should facilitate connections not only among people living with PD but also with individuals who share similar interests.

> *...someone has to connect people in one way or another, because if you sit on your own and feel lonely and have Parkinson, you are not that likely to establish some kind of contact with other people. And I mean, that might be a good idea to connect to people with Parkinson, but it could also be a good idea to connect with people that don't have Parkinson.* [Person living with PD 3]

Since many of them become exhausted easily because of PD, they expressed that having a digital solution to communicate with groups through video chats would help them interact with others without losing much energy.

> *it's easier to meet someone by the computer, it doesn't require so much energy. You don't have to go somewhere and take the bus or walk. You can sit at home, and you can meet people, talk to them.* [Person living with PD 11]

Increase Social Participation

Apart from initiating group communication, some also suggested having recommendations about events (both digital and offline) and clubs and deals based on their locality and their interests or hobbies. The people living with PD also suggested a solution that helps them to create an event if they would like to give training or share their expertise with others. Currently some people living with PD were unaware of the existing classes such as ballet, boxing, and dance for PD.

> *Training...if you find training, if you can for example show us where...for example, some people don't know. Still some Parkinson's patients do not know that there is dance, ballet dance for Parkinson's. If you refer to offers, that one pays, free etc...It also sounds perfect. Very good.* [Person living with PD 8]

They suggested that information should include the time, date, location, and whether it is a recurring event. This information helps them to plan in advance. It makes them participate in social activities. Further, they stated that this kind of information would help people living with PD who have difficulties taking the initiatives to go out.

> *...this communication about different activities, what times...time it starts, where is it, how many times, how many weeks. So, if you want to be in touch with other people, then I think it's the best to know when and where.* [Person living with PD 14]

Increase a Sense of Belonging

In order to create a sense of belonging and feeling valued, many people living with PD suggested having a community forum to share information about their experience with PD. In that way they feel they can help others and have a sense of purpose. They are also looking for the reliable latest information about PD from experts.

XSL•FO
RenderX

> *When I was diagnosed three years ago, I told my children and then, it was not more than week or fortnight until they had googled and were very well informed and told me about it. So, information to people is the best way to feel belonging.* [Person living with PD 13]

> *..there are several people who call me who are Parkinson sick, and are worried, and have not received the care that they want...you have to help, give them tips and suggestions...create a forum where you can ask questions, and also answer others...it could be...it's probably quite rewarding for a lot of people, I think, who are alone and don't have anyone to talk to.* [Person living with PD 2]

Other Suggestions

The people living with PD also suggested having a simple, easy-to-use solutions where they do not need to remember any passwords or codes. Some people living with PD felt that proper technology help and guidance would help them better use computer, phone, and tablets.

> *But what I really want to stress is that it should be simple to use, and I think also you need to have support if something happens...I always say that crucial aspect is that you have to...to have user-friendly technology and both physically, using the fingers, and then also mentally, that you should have systems where you do not have difficult password, that makes you nervous and so on.* [Person living with PD 15]

Discussion

Principal Findings

The results of this study describe difficulties faced by people living with PD in the context of social isolation and loneliness along with their technology use. In addition, the results highlight the high variability among people living with PD for intended social withdrawal. For some people living with PD, it was the fatigue caused by PD that restricted their participation, whereas the psychological conditions and social stigma due to PD played a role for others. The social comparison theory [31] may explain the anxious nature observed in certain people living with PD, as they worry about encountering individuals with more advanced symptoms. Alternatively, seeking interactions with peers facing similar experiences may provide comfort and support, fostering a sense of belonging and reducing isolation. Understanding these influences can inform targeted interventions, such as support groups, to address the social and emotional needs of people living with PD effectively. When compared with social stigma constructs such as "stigma arising from symptoms," "stigma linked to relational and communication problems," "social stigma arising from sharing perceptions," and "caregiver's stigma" [28], we found that the people living with PD experienced stigma due to their symptoms, relational and communication problems, and also due to sharing perceptions with others. However, the relatives who participated in this study did not express any stigma ("caregivers' stigma") because of their family member's condition. This situation may be because very few relatives participated in this study and no question was asked about their life experiences separately. We also found that the medication side effects were related to increased impulse control disorder (eg, gambling), which could further lead to social withdrawal which was not highlighted in previous studies [7,26,27].

Not surprisingly, we found that the majority of the people living with PD in our study stated that they experienced difficulties using technology, especially, the keyboard and the mouse due to their motor symptoms. This aligns with previous research that has shown people living with PD encountering problems with typing and using a mouse [32,33]. Adopting alternate input devices or adapted (larger keys) keyboards and mouse can be instrumental in addressing these challenges and making ICT solutions more accessible and user-friendly for people living with PD. Further, we also found out that due to PD, people living with PD also have cognitive decline making it difficult to remember passwords and complete complex tasks. Furthermore, there is a lack of IT support to help people living with PD in case of any issues. Some people living with PD explicitly stated that they did not know where to find IT support. Besides PD symptoms, aging also contributed to difficulties in using a computer, such as learning new technologies.

As mentioned earlier, most of the ICT solutions such as online support groups or discussion forums are not tailored to the needs of people living with PD for managing social isolation or loneliness [22,23]. This point was also raised as a concern by people living with PD, who mentioned that including them during the design phase of new solutions is essential. Therefore, given the wide range of needs and barriers experienced by people living with PD, it is more suitable to adopt user-centered design approaches that emphasizes the active participation of end users in the design process [34,35].

Our results demonstrated that the solutions should consider the barriers faced by people living with PD which restrict their social activities. For example, some may experience "apathy" and withdraw from social activities, hence the ICT solution should address this and help them to socialize by incorporating features that actively encourage socialization and engagement. This can be done by providing features such as customized content, online community, online events, and reminders or prompts for people living with PD to participate in social activities. Further, to enhance the accessibility of ICT solutions for people living with PD, developers should consider incorporating features like voice recognition, enabling users to convert spoken words into text or commands. This feature could be particularly useful for people living with PD who have problems typing or using a mouse, allowing them to dictate text and navigate the solution using voice commands. However, it is important to acknowledge that speech-related challenges are prevalent among some people living with PD, and standard voice recognition technologies may not always be effective in recognizing weak or impaired speech. Therefore, specialized software and adaptations should be developed to cater to the unique speech needs of people living with PD, ensuring that the ICT solutions are truly inclusive and beneficial for all users. Further, our people living with PD emphasized the need for

XSL•FO
RenderX

simple easy-to-use ICT solutions, which has also been highlighted in many studies [20,36].

Concerning the needs highlighted by people living with PD, they emphasized that ICT solutions should not only be used by people living with PD but should include everyone. Currently, discussion forums or platforms designed for people living with PD mostly include people living with PD alone or may also include carers and health care professionals [23,25]. Furthermore, some people living with PD wanted information about public events near their residence to be customized based on their interests. In contrast, some people prefer communicating online with others based on their interests. For people who experience social stigma, connecting with others who have similar experiences with the disease can help the affected individuals to overcome their stigma [37]. Hence the PD people's suggestion of an online discussion forum not only helps to create a sense of belonging but may also help them to overcome their stigma by coming to know other people's experiences. These findings indicate that ICT solutions need to be adapted and tailored based on the barriers experienced by people living with PD and their ICT needs, in order to facilitate managing their social isolation and loneliness.

The study also revealed a significant finding related to impulse control disorder caused by medication side effects highlights the importance of designing ICT solutions that do not promote internet addiction. Excessive internet use can lead to social withdrawal and further isolate individuals from society [38,39]. Therefore, it is crucial for technology developers and designers to consider the potential impact of their solutions on users' behavior and well-being. Incorporating features that prioritize meaningful engagements instead of addictive features such as aimless scrolling, automatic video playback, and gamification elements (such as points and rewards) can promote positive social connections and reduce the risk of internet addiction.

Strengths and Limitations

The main strength of this study was the inclusion of both people living with PD and health care professionals. This approach has provided unique insights and perspectives that might not be accessible to researchers who do not have direct experience with the disease. Furthermore, this inclusion allowed us to better understand the barriers faced daily by people living with PD. For example, the people living with PD shared their personal experiences and explained how PD had affected their social life, while the health care professionals provided medical explanations for these challenges. Further, the health care professionals' involvement helped us to identify a new barrier "medication side effects." Overall, involving both people living with PD and health care professionals in this study gave us a complete picture of the challenges that people living with PD face and helped us to identify important issues that might have been missed if we had only spoken to 1 group. Another strength of this study was the choice of qualitative methods (interviews and focus groups). This approach allowed us to gather in-depth information about the experiences of people living with PD and health care professionals regarding their perspectives on the challenges of living with PD. This type of data is difficult to capture through quantitative methods alone, such as surveys or questionnaires. Further, the focus group provided a supportive environment for people living with PD to discuss their experiences and concerns with others who share similar experiences. This approach promoted a sense of community among participants.

The main limitation of this study was the use of an online platform for interviews and focus groups due to the COVID-19 pandemic. While this approach allowed for people living with PD from 2 different regions in Sweden to participate, it also created bias by excluding individuals with limited technological knowledge. Additionally, the recruitment through the Parkinson association limited the possibility of including more isolated and lonelier individuals. Another limitation of the study was the challenge of scheduling a focus group workshop with the same people living with PD who were interviewed initially. Different proposed timings made it difficult to find a suitable time for everyone. Some participants had health conditions that made it challenging for them to attend a 2-hour focus group interview. This situation resulted in a smaller sample size for the focus group than originally intended. Future research can focus on conducting a larger study with a more diverse sample of participants. It could consider recruiting individuals who do not have access to technology or who may be more isolated, to better understand their experiences and perspectives.

Conclusions

This study showed that the barriers experienced by people living with PD due to PD, such as "fatigue," "psychological conditions," "social stigma," and "medication side effects," affect their social life. Additionally, people living with PD experience difficulties using keyboards and mouse, remembering passwords, and navigating complex applications due to their physical and cognitive limitations. The people living with PD suggested having a simple and easy-to-use digital community based on their interests and looking for recommendations for events nearby to manage their social isolation and loneliness. Designers need to consider these barriers and facilitators while developing ICT solution among people living with PD. Importantly, any ICT solution designed for people living with PD should not encourage internet addiction, which will further contribute to the person's withdrawal from society.

Acknowledgments

The authors would like to thank all the participants who agreed to participate in this study for their valuable time and contributions. This project received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant (754285).

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are not publicly available as they include sensitive data but are available from the corresponding author on reasonable request, including the need of Swedish ethical vetting.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[PDF File (Adobe PDF File), 104 KB - jmir_v26i1e48175_app1.pdf]

References

1. Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, et al. Parkinson disease. Nat Rev Dis Primers 2017;3:17013. [doi: 10.1038/nrdp.2017.13] [Medline: 28332488]
2. Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015;30(12):1591-1601. [doi: 10.1002/mds.26424] [Medline: 26474316]
3. Balestrino R, Schapira AHV. Parkinson disease. Eur J Neurol 2020;27(1):27-42. [doi: 10.1111/ene.14108] [Medline: 31631455]
4. Chaudhuri KR, Healy DG, Schapira AHV. Non-motor symptoms of Parkinson's disease: diagnosis and management. Lancet Neurol 2006;5(3):235-245. [doi: 10.1016/S1474-4422(06)70373-8] [Medline: 16488379]
5. Schrag A, Jahanshahi M, Quinn N. How does Parkinson's disease affect quality of life? a comparison with quality of life in the general population. Mov Disord 2000;15(6):1112-1118. [doi: 10.1002/1531-8257(200011)15:6<1112::aid-mds1008>3.0.co;2-a] [Medline: 11104193]
6. Perepezko K, Hinkle JT, Shepard MD, Fischer N, Broen MPG, Leentjens AFG, et al. Social role functioning in Parkinson's disease: a mixed-methods systematic review. Int J Geriatr Psychiatry 2019;34(8):1128-1138 [FREE Full text] [doi: 10.1002/gps.5137] [Medline: 31069845]
7. Ahn S, Springer K, Gibson JS. Social withdrawal in Parkinson's disease: a scoping review. Geriatr Nurs 2022;48:258-268 [FREE Full text] [doi: 10.1016/j.gerinurse.2022.10.010] [Medline: 36332441]
8. Nicholson NR. A review of social isolation: an important but underassessed condition in older adults. J Prim Prev 2012;33(2-3):137-152. [doi: 10.1007/s10935-012-0271-2] [Medline: 22766606]
9. Luanaigh CO, Lawlor BA. Loneliness and the health of older people. Int J Geriatr Psychiatry 2008;23(12):1213-1221. [doi: 10.1002/gps.2054] [Medline: 18537197]
10. de Jong Gierveld J, Van Tilburg T, Dykstra P. Loneliness and social isolation. In: Vangelisti AL, Perlman D, editors. The Cambridge Handbook of Personal Relationships. Cambridge: Cambridge University Press; 2006:485-500.
11. Schrag A, Hovris A, Morley D, Quinn N, Jahanshahi M. Young- versus older-onset Parkinson's disease: impact of disease and psychosocial consequences. Mov Disord 2003;18(11):1250-1256. [doi: 10.1002/mds.10527] [Medline: 14639664]
12. Gazewood JD, Richards DR, Clebak K. Parkinson disease: an update. Am Fam Physician 2013;87(4):267-273 [FREE Full text] [Medline: 23418798]
13. Perissinotto C, Holt-Lunstad J, Periyakoil VS, Covinsky K. A practical approach to assessing and mitigating loneliness and isolation in older adults. J Am Geriatr Soc 2019;67(4):657-662. [doi: 10.1111/jgs.15746] [Medline: 30762228]
14. Chen YRR, Schulz PJ. The effect of information communication technology interventions on reducing social isolation in the elderly: a systematic review. J Med Internet Res 2016;18(1):e18 [FREE Full text] [doi: 10.2196/jmir.4596] [Medline: 26822073]
15. Khosravi P, Rezvani A, Wiewiora A. The impact of technology on older adults' social isolation. Comput Hum Behav 2016;63:594-603. [doi: 10.1016/j.chb.2016.05.092]
16. Fakoya OA, McCorry NK, Donnelly M. Loneliness and social isolation interventions for older adults: a scoping review of reviews. BMC Public Health 2020;20(1):129 [FREE Full text] [doi: 10.1186/s12889-020-8251-6] [Medline: 32054474]
17. Thangavel G, Memedi M, Hedström K. Customized information and communication technology for reducing social isolation and loneliness among older adults: scoping review. JMIR Ment Health 2022;9(3):e34221 [FREE Full text] [doi: 10.2196/34221] [Medline: 35254273]
18. Al Mamun KA, Alhussein M, Sailunaz K, Islam MS. Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications. Future Gener Comput Syst 2017;66:36-47. [doi: 10.1016/j.future.2015.11.010]
19. Kim DY, Hwang SH, Kim MG, Song JH, Lee SW, Kim IK. Development of Parkinson Patient Generated Data Collection Platform Using FHIR and IoT Devices. Stud Health Technol Inform 2017;245:141-145. [Medline: 29295069]

XSL•FO
RenderX

20. Kenny L, Moore K, O' Riordan C, Fox S, Barton J, Tedesco S, et al. The views and needs of people with Parkinson disease regarding wearable devices for disease monitoring: mixed methods exploration. JMIR Form Res 2022;6(1):e27418 [FREE Full text] [doi: 10.2196/27418] [Medline: 34989693]
21. Memedi M, Tshering G, Fogelberg M, Jusufi I, Kolkowska E, Klein G. An interface for IoT: feeding back health-related data to Parkinson's disease patients. J Sens Actuator Netw 2018;7(1):14 [FREE Full text] [doi: 10.3390/jsan7010014]
22. Gerritzen EV, Lee AR, McDermott O, Coulson N, Orrell M. Online peer support for people with Parkinson disease: narrative synthesis systematic review. JMIR Aging 2022;5(3):e35425 [FREE Full text] [doi: 10.2196/35425] [Medline: 35896025]
23. Attard A, Coulson NS. A thematic analysis of patient communication in Parkinson's disease online support group discussion forums. Comput Hum Behav 2012;28(2):500-506. [doi: 10.1016/j.chb.2011.10.022]
24. Subramanian I. Virtual Parkinson's disease support groups in the COVID-19 era: social connection in the time of social distancing. Mov Disord Clin Pract 2020;7(6):739-740 [FREE Full text] [doi: 10.1002/mdc3.12994] [Medline: 32775536]
25. Boulos MNK, Ifeachor E, Escudero J, Zhao P, Carroll C, Costa P, et al. LiveWell—promoting healthy living and wellbeing for Parkinson patients through social network and ICT training: lessons learnt and best practices. Int J Healthc Inf Syst Inform 2015;10(3):24-41. [doi: 10.4018/ijhisi.2015070102]
26. Soleimani MA, Negarandeh R, Bastani F, Greysen R. Disrupted social connectedness in people with Parkinson's disease. Br J Community Nurs 2014;19(3):136-141 [FREE Full text] [doi: 10.12968/bjcn.2014.19.3.136] [Medline: 24897835]
27. Hammarlund CS, Westergren A, Åström I, Edberg AK, Hagell P. The impact of living with Parkinson's disease: balancing within a web of needs and demands. Parkinsons Dis 2018;2018:4598651 [FREE Full text] [doi: 10.1155/2018/4598651] [Medline: 30151098]
28. Maffoni M, Giardini A, Pierobon A, Ferrazzoli D, Frazzitta G. Stigma experienced by Parkinson's disease patients: a descriptive review of qualitative studies. Parkinsons Dis 2017;2017:7203259 [FREE Full text] [doi: 10.1155/2017/7203259] [Medline: 28243481]
29. Cornwell EY, Waite LJ. Measuring social isolation among older adults using multiple indicators from the NSHAP study. J Gerontol B Psychol Sci Soc Sci 2009;64(Suppl 1):i38-i46 [FREE Full text] [doi: 10.1093/geronb/gbp037] [Medline: 19508982]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Gerber JP, Wheeler L, Suls J. A social comparison theory meta-analysis 60+ years on. Psychol Bull 2018;144(2):177-197. [doi: 10.1037/bul0000127] [Medline: 29144145]
32. Cunningham LM, Nugent CD, Finlay DD, Moore G, Craig D. A review of assistive technologies for people with Parkinson's disease. Technol Health Care 2009;17(3):269-279. [doi: 10.3233/THC-2009-0547] [Medline: 19641262]
33. Nunes F, Silva PA, Cevada J, Barros AC, Teixeira L. User interface design guidelines for smartphone applications for people with Parkinson's disease. Univ Access Inf Soc 2016;15(4):659-679. [doi: 10.1007/s10209-015-0440-1]
34. Abras C, Maloney-Krichmar D, Preece J. User-centered design. In: Bainbridge W, editor. Encyclopedia of Human-Computer Interaction. Thousand Oaks, CA: Sage Publications; 2004:445-456.
35. Kramer J, Noronha S, Vergo J. A user-centered design approach to personalization. Commun ACM 2000;43(8):44-48 [FREE Full text] [doi: 10.1145/345124.345139]
36. Ahmed M, Marín M, How D, Judica E, Tropea P, Bentlage E, et al. End users' and other stakeholders' needs and requirements in the development of a personalized integrated care platform (PROCare4Life) for older people with dementia or Parkinson disease: mixed methods study. JMIR Form Res 2022;6(11):e39199 [FREE Full text] [doi: 10.2196/39199] [Medline: 36449340]
37. Gray AJ. Stigma in psychiatry. J R Soc Med 2002;95(2):72-76 [FREE Full text] [doi: 10.1177/014107680209500205] [Medline: 11823548]
38. Shaw M, Black DW. Internet addiction: definition, assessment, epidemiology and clinical management. CNS Drugs 2008;22(5):353-365. [doi: 10.2165/00023210-200822050-00001] [Medline: 18399706]
39. Meshi D, Cotten SR, Bender AR. Problematic social media use and perceived social isolation in older adults: a cross-sectional study. Gerontology 2020;66(2):160-168. [doi: 10.1159/000502577] [Medline: 31522178]

Abbreviations

ICT: information and communication technology
PD: Parkinson disease
RQ: research question

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 14.04.23; peer-reviewed by H Pilabré, J Taylor; comments to author 14.06.23; revised version received 17.08.23; accepted 20.11.23; published 17.01.24.

Please cite as:
Thangavel G, Memedi M, Hedström K
Information and Communication Technology for Managing Social Isolation and Loneliness Among People Living With Parkinson Disease: Qualitative Study of Barriers and Facilitators
J Med Internet Res 2024;26:e48175
URL: https://www.jmir.org/2024/1/e48175
doi:10.2196/48175
PMID:38231548

©Gomathi Thangavel, Mevludin Memedi, Karin Hedström. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Addressing Cyberscams and Acquired Brain Injury ("I Desperately Need to Know What to Do"): Qualitative Exploration of Clinicians' and Service Providers' Perspectives

Kimberly Ann Chew[1,2], BA (Hons); Jennie Ponsford[1,2], BA (Hons), MA, PhD; Kate Rachel Gould[1,2], BBNSc (Hons), DPsych

[1]Monash-Epworth Rehabilitation Research Centre, Richmond, Victoria, Australia

[2]Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton, Victoria, Australia

Corresponding Author:
Kimberly Ann Chew, BA (Hons)
Monash-Epworth Rehabilitation Research Centre
185-187 Hoddle Street
Richmond, Victoria, 3121
Australia
Phone: 61 03 9426 8923
Email: kimberly.chew@monash.edu

Abstract

Background: People with acquired brain injury (ABI) may be more susceptible to scams owing to postinjury cognitive and psychosocial consequences. Cyberscams result in financial loss and debilitating psychological impacts such as shame and mistrust, interference with neurorehabilitation, and reduced independence. Despite these significant consequences, there are no psychological treatments to support cyberscam survivors. There is limited evidence regarding how the current workforce is addressing post-ABI cyberscams.

Objective: This study aims to understand the perspectives and needs of clinicians and service providers in addressing post-ABI cyberscams.

Methods: Overall, 20 multidisciplinary clinicians and service providers were recruited through purposive sampling across Australia. Semistructured interviews explored post-ABI scam experiences and vulnerabilities, treatments and their efficacy, and recommendations for future cybersafety recovery interventions. Reflexive thematic analysis was used.

Results: In total, 8 themes encompassing a biopsychosocial understanding of scam vulnerabilities and impacts were identified: "genuine lack of awareness: cognitive-executive difficulties"; "not coping with the loss of it all"; "needing trust and connection"; "strong reactions of trusted others"; "nothing structured to do"; "financial stress and independence"; "cyberability"; and "scammer persuasion." Each theme informed clinical recommendations including the need to provide psychological and cognitive support, enhance financial and cybersafety skills, promote meaningful social engagement, and foster collaboration between families and clinical support teams.

Conclusions: The multifaceted range of scam vulnerabilities and impacts highlighted the need for individualized, comprehensive, and targeted treatments using a biopsychosocial approach to enable cyberscam recovery among people with ABI. These findings will guide the development of a co-designed intervention.

(J Med Internet Res 2024;26:e51245) doi:10.2196/51245

KEYWORDS

cyberscam; cyberscams; fraud; cybercrime; cybersafety; brain injury; disability; neurorehabilitation; interventions; treatment; qualitative

XSL•FO
RenderX

Introduction

Background

Acquired brain injuries (ABIs), such as traumatic brain injury and stroke, are leading causes of disability, affecting 1 in 45 Australians [1]. People with ABI often experience significant cognitive-communication difficulties, mental health challenges, increased social isolation, and reduced participation in meaningful activities [2,3]. These factors often interact with each other, highlighting the need to understand and address the impacts of biological, psychological, and social (ie, biopsychosocial) processes [4]. To reduce isolation and form meaningful relationships, people with ABI may engage in web and social media use [5,6]. Although there are many benefits to online engagement, cybercrimes, such as scams, are rapidly increasing internationally. In 2022, more than Aus $3 billion (US $1.97 billion) was lost owing to scams in Australia, with 71% increase in scams reported by people with disability in 2021 [7]. Anyone can be scammed; however, post-ABI sequelae including lack of insight and impulsivity may place people with ABI at increased risk [8,9].

Cyberscams are crimes that leverage mass communication tools to defraud people for financial gain [10]. Common cyberscams include investment, dating and romance, false billing, phishing, remote access, threats to life and arrest, identity theft, job and employment, online shopping, and classified scams [7]. Reports have highlighted the digital vulnerabilities of community groups such as adults with disability and young people [11-13]. People with disability are also more susceptible to cybercrimes related to threats to life, arrest, ransomware, other malware, and mobile premium services [7]. Although no study has directly investigated the relationship between cognitive impairment in ABI and scam susceptibility, studies of older adults with mild cognitive impairment and vascular pathology lend some support to this contention [14-16]. There is currently a paucity of established evidence specifically regarding cyberscams or safety and ABI [17,18]. A recent survey of 101 ABI clinicians and service providers found that 53.5% of participants had at least 1 client with ABI affected by cyberscams, with romance scams encompassing the majority, highlighting the need for clinical and research attention toward this matter within neurorehabilitation [9]. Participants described that their current management approaches include psychologist referrals, building awareness, and increasing community engagement, yet they did not endorse any strategy as particularly effective.

Qualitative interviews exploring the lived experiences of people with ABI who have been scammed and close others (COs) revealed a complex scam journey, with multiple significant difficulties extricating from the scam or its impacts [8]. Factors contributing to scams included scam vulnerability, scammer manipulation, reduced cyberscam awareness, financial and psychosocial functioning, and unhelpful or restrictive cyberscam responses from families and services (eg, limiting internet and financial access). Distressing psychological consequences such as shame, fear, and mistrust were often experienced [8,9], similar to the general population [19-21]. These psychological consequences are often debilitating owing to a “double-hit” of both financial and emotional distress [20]. Post-ABI scams have also interfered with neurorehabilitation treatment and reduced independence [9]. Intervention recommendations from those with lived experience included increasing scam awareness; reducing shame through hearing scam stories; and normalizing scams through the provision of education, acknowledging how common scams are, and reassuring survivors not to blame themselves [8].

There is a critical need for evidence-based strategies to address the significant unmet financial and psychological impacts of cyberscams. Cyberscam prevention campaigns have been implemented for the general population [22,23]. However, despite the significant distress experienced by survivors, no known intervention has been identified to address the psychosocial impacts of cyberscams. Furthermore, people with ABI who are scammed may require tailored interventions, which account for their cognitive, emotional, and functional impairments. Our group recently co-designed CyberAbility [17], an ABI-tailored, prevention-focused, cybersafety training program. “Cyberability” refers to the ability to learn and adapt to current and emerging technologies (eg, safe and responsible online use) [24]. Social-ABI-lity, an ABI-specific web-based social media skills training program, has also been recently developed [18]. Nonetheless, there remains a need for ABI-tailored interventions that specifically support psychological recovery after the cyberscam. Interestingly, participants involved in the CyberAbility [17] co-design process experienced it as “an intervention addressing shame” [25]. This suggests potential for supporting scam recovery through sharing stories, peer bonding, and meaningful contribution.

Objective

Clinicians and service providers play a central role in supporting and providing treatment to people with ABI who have been cyberscammed. Currently, there is limited research investigating the current treatment landscape among people with ABI [9]. Moreover, the secondary impacts of addressing cyberscams on health professionals is unknown. Therefore, this study aimed to qualitatively understand the experiences and recommendations of clinicians and service providers in managing the impacts of cyberscams among people with ABI.

Methods

Design

The study had a qualitative design. This paper is presented based on the COREQ (Consolidated Criteria for Reporting Qualitative Research) 32-item checklist [26].

Ethical Considerations

Institutional ethics approval (Monash University Human Research Ethics Committee; 17984) and informed consent were obtained.

Procedure

Authors KRG and KAC conducted the interviews between January 2019 and October 2022. Both interviewers were women with at least an honors degree in psychology and were trained in qualitative interviewing and analysis. Interviews were

XSL•FO
RenderX

conducted face to face (2/25, 8%) or via videoconference (23/25, 92%). Interviews were audio recorded and transcribed externally or by KAC. Interviews ranged from 37 to 81 minutes in duration. Interviewers kept reflection journals for each interview. An interview summary was sent to participants to ensure that their views were accurately captured and to seek additional feedback. Comments were integrated into the findings before analysis.

Participants

Multidisciplinary clinicians and service providers were recruited via an existing research database, professional networks, and social media, with invitations to participate sent via email. Eligibility criteria were the following: clinicians or service providers with experience in providing assessment or intervention to adults with ABI, experience in cybercrime vulnerability or supporting an adult with ABI who had been affected by a cybercrime, and working within Australia. Overall, 20 participants were recruited. There were preexisting interviewer-participant relationships (professional colleagues) among 35% (7/20) of the participants. Participant characteristics are displayed in Table 1. Most participants (18/20, 90%) were female, with an average of 18 (SD 9) years of experience and a bachelor's degree or higher qualification. Most (13/20, 65%) were clinical neuropsychologists and occupational therapists working in community settings.

Table 1. Participant demographics (N=20).

Characteristics	Value
Age at the time of interview (y), mean (SD); range	43.89 (0.54; 25-63)
Sex, n (%)	
Female	18 (90)
Male	2 (10)
Highest level of education, n (%)	
Doctoral degree	6 (30)
Master's degree	9 (45)
Graduate diploma	1 (5)
Bachelor's degree	4 (20)
Experience at the time of interview, mean (SD; range)	18.43 (9.99 y; 8 mo-42 y)
Discipline, n (%)	
Clinical neuropsychology	9 (45)
Clinical psychology	2 (10)
Occupational therapy	4 (20)
Psychologist with no endorsement	2 (10)
Service provider	3 (15)
Work setting, n (%)	
Community or private practice	16 (80)
Hospital	1 (5)
Forensic department	1 (5)
Not-for-profit organization	2 (10)

Measures

An initial semistructured interview schedule was designed to explore client cyberscam incidents, prevention strategies, resources, confidence in personal cybersafety, and skills in addressing cyberscam impacts with clients. Following a reflexive design, the interview schedule was revised iteratively based on the generated themes, gaps, and areas of interest identified during the interviews. Additional questions were related to current and future interventions, scam and intervention engagement, and recommendations for intervention content and structure. Refer to Multimedia Appendix 1 for the interview schedules. Of the 20 participants, 5 (25%) participants completed the initial interview schedule only, 5 (25%) completed the initial interview and a second interview covering the additional questions, and 10 (50%) completed a single interview covering all questions (ie, a total of 25 interviews were conducted with the 20 participants). Participants' demographics were obtained using a brief questionnaire.

Data Analysis

Descriptive statistics for participant demographics were summarized using Microsoft Excel (2019). Interview transcripts were analyzed using the 6-phase thematic analysis by Braun and Clarke [27] to identify themes reflecting the perceptions about client experiences and intervention recommendations. A nonpositivist [28] and nonlinear format [29,30] was used to analyze themes. First, data familiarization was conducted, where

XSL•FO
RenderX

data were repeatedly reviewed. Second, initial codes were generated inductively. Codes were cataloged using NVivo (version 12.5.0; Lumivero) [31]. KRG coded the first 36% (9/25) of the interviews, and KAC coded the remaining 64% (16/25) of the interviews. Acknowledging that reflexive thematic analysis does not require cross-coding [32], 8% (2/25) of the transcripts were cross-coded by KAC as part of her training. Excellent intercoder reliability was found (mean Cohen κ=0.83; range 0.31-1) [31,33]. In total, 3 codes with Cohen κ<0.50 were reviewed by both coders [34], and consensus was achieved through discussion. Third, codes were classified into high-order themes following repeated review with all authors. Provisional themes were discussed and refined 7 times among all authors. Themes were visually mapped using Miro (Realtime Board, Inc), a collaborative whiteboard [35]. Fourth, the thematic map and themes were checked against the original transcripts and refined to ensure accuracy. Fifth, once consensus was achieved, themes were defined and labeled. Sixth, a written conceptualization of the findings was prepared, and representative quotes were selected.

Results

Overview

In total, 20 clinicians and service providers described 34 cyberscams experienced by their clients with ABI who had predominantly moderate to severe traumatic brain injury or stroke. Most cyberscams reported were romance or friendship scams (17/34, 50%), with other scam examples including buying and selling, prize, service provider, and investment scams. Scam impacts included near-misses to significant financial loss (eg, Aus $250,000 [US $168,801]; “all his savings”) and psychological distress. Scam survivors were also placed in personal danger and experienced sexual safety impacts. Striking romance scam examples included a woman who traveled to Africa alone, requiring rescue by the Australian Federal Police; a man who was sexually extorted; and a woman living in a 24-hour accommodation who was taken advantage of by registered sexual offenders she met on the web, resulting in childbirth.

Overall, 8 interlinked themes relating to cyberscam vulnerabilities and impacts on people with ABI were generated by the researchers through reflexive thematic analysis (Figure 1). Cyberscam vulnerabilities and impacts were organized into biological, psychological, and social aspects and visually displayed by a ship’s helm to represent the interrelationships of these themes with equal influence on the conceptualization of survivors’ scam experiences. ABI-related cognitive impairment underpinned vulnerability and impacts by reducing the individual’s ability to recognize and protect themselves from scams. Furthermore, ABI was associated with impaired coping capacity and reduced ability to navigate important relationships with support people concerned about the scam. Owing to ABI, support people commonly implemented restrictive practices that exacerbated rather than mitigated cyberscam risks and impacts. Theme-specific intervention recommendations, shown in the outer ring of the helm, are described separately after exploration of the thematic results.

XSL•FO
RenderX

Figure 1. Thematic map of clinicians' and service providers' perspectives.

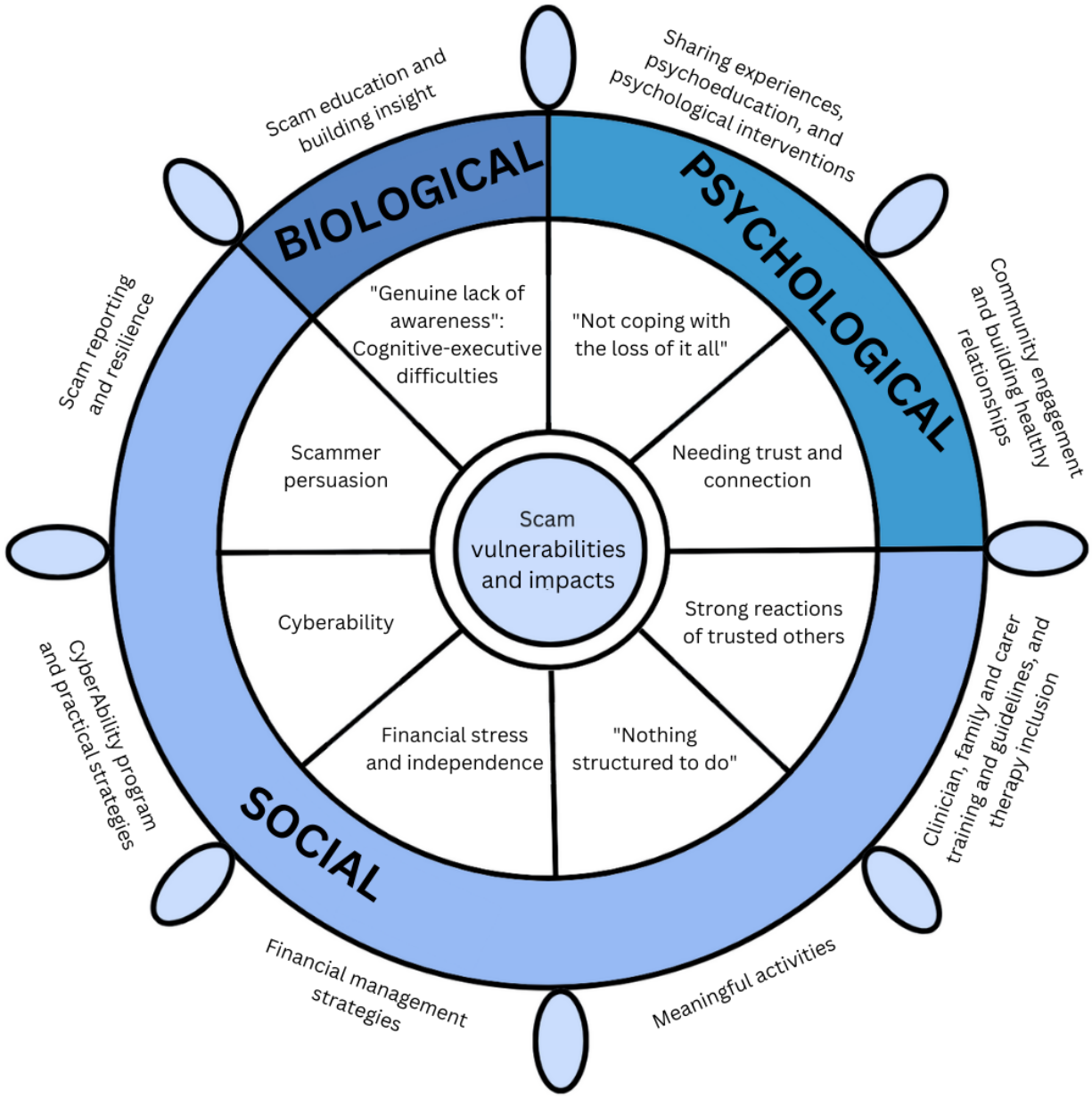

"Genuine Lack of Awareness": Cognitive-Executive Difficulties

Overview

ABI was a key biological factor underpinning vulnerability to cyberscams. Specifically, participants highlighted that cognitive-executive impairments associated with ABI increased cyberscam vulnerability owing to lack of insight, difficulties in generalizing between experiences, and impaired memory for vital information. Although lack of insight occasionally protected against the emotional impacts of cyberscams, people with ABI remained vulnerable to future cyberscams owing to difficulties in learning from mistakes.

Cognitive Vulnerability

Most participants (15/20, 75%) noted that anyone can be scammed; however, "if we're all vulnerable, then...someone with an ABI is even more vulnerable" (participant 13). Specifically, all participants (20/20, 100%) identified executive dysfunction as a significant cyberscam vulnerability, including lack of insight; cognitive rigidity; and impairments in problem-solving, reasoning, decision-making, understanding complex information, generalizing, and learning from mistakes. Executive deficits and reduced insight placed some people with ABI in dangerous situations owing to difficulties in identifying red flags and appreciating both real and potential risks—an occupational therapist described the physical danger her client faced when she traveled to Africa to meet the scammer she thought was her boyfriend:

> *She came so close to losing her life in many ways or being assaulted or she was in a hotel room on her own for a week with no food.... She took $50 Australian with her, she caught taxis on her own...but still, she doesn't seem to have that level of severity*

XSL•FO
RenderX

> *of the concern at all of that situation...she'll be, "Really? I didn't see that."... I think that's just genuine lack of awareness because of her brain injury.* [Participant 2]

Executive difficulties also underpinned behavioral dysregulation. More than half of the participants (11/20, 55%) shared that their clients with ABI engaged in impulsive and risk-taking behaviors. People with ABI were left vulnerable to scammer tactics (eg, time pressure) owing to "inability to regulate her behaviour in the moment when she sees an instant reward" (participant 18).

Brain injury–related reductions in complex communication skills such as comprehension of implied information, judgment, and self-monitoring increased risks during online discourse:

> *It's a lot easier to misinterpret what you say.* [Participant 5]

In combination, these cognitive-executive, behavioral, and communication difficulties underpinned by ABI resulted in increased vulnerability to being repeatedly scammed and reduced self-awareness of this susceptibility.

Barriers to Recovery

Participants described that concrete and inflexible thinking presented barriers to cyberscam recovery as it took time and additional support "because of his brain injury to put two and two together" (participant 15). Moreover, disbelief and denial, in conjunction with the grooming by scammers, made it difficult for clients to accept the scam and engage in intervention as they still believed that they were "getting special scam treatment" (participant 2). Cognitive difficulties of impaired memory and reduced insight interfered with intervention effectiveness:

> *You weren't gonna be able to do any kinds of therapy with him, any intervention of that nature. He just wouldn't have been able to...hold on to it, and...believe that he needed it.* [Participant 17]

For those who accepted the need for cybersafety practices, memory and executive difficulties reduced an individual's ability to learn from mistakes and use strategies when responding to high-risk cybersafety situations in real time, contributing to cyberscam vulnerability and re-engagement:

> *He remembers that it's something he should be doing so he does hang onto that information...when he's prompted...But in the heat of the moment, that doesn't necessarily come into his mind as the first thing he's gonna do.* [Participant 8]

Short-Term Protection

Interestingly, 20% (4/20) of the participants shared that lack of insight initially appeared to protect people with ABI from negative emotional impacts of the cyberscam. However, as a result of the absence of a negative reaction to the scam, they were less likely to learn from their experiences, contributing to repeated exposure to scams:

> *She recovered from [a little bit of shame] pretty quickly. But she also re-engaged with another scammer. So she wasn't down for long. It's almost like – yes, her heart was broken, but then she kind of went back and found someone else.* [Participant 9]

Nonetheless, preexisting psychological difficulties, in combination with limited social connection and cognitive impairment, increased scam susceptibility:

> *His isolation and depression...combined with his TBI [traumatic brain injury] just left him really vulnerable.* [Participant 17]

This was compounded by the trusting and generous nature of participants, who empathized with the convincing stories told by cyberscammers as part of their grooming techniques, which made it hard for survivors to accept the scam:

> *He still reflects on how sorry he felt for her [cyberscammer], how difficult her life was, how he felt he could offer her a stable and good life.* [Participant 16]

"Not Coping With the Loss of it All"

Overview

Psychological difficulties before the scam, combined with reduced cognition, were reported to increase cyberscam vulnerability. All participants (20/20, 100%) also identified significant psychological repercussions of the scam among people with ABI, particularly, shame, depression, anxiety, and denial. Clients who found it hard to accept the cyberscam often re-engaged with the same scammer:

> *That scam had gone on for many years...and he, even at the time that I exited working with him, was still not convinced that story wasn't true.* [Participant 16]

Debilitating psychological impacts also interfered with treatment and rehabilitation.

"Life Plummeting to New Depths"

Several significant psychological impacts were identified by participants where their clients were "psychologically, emotionally, financially abused, and it can lead to suicide" (participant 9). Depression, anger, embarrassment, mistrust, shame, and guilt were the most common psychological impacts reported:

> *She was angry. She's like, "I've been so stupid. I should never have done this. I can't believe I've been scammed."* [Participant 3]

For those embroiled in romance scams, losing the relationship with the scammer often resulted in depression:

> *To think you're going to get a relationship and you're not is hugely devastating for our client group.* [Participant 4]

Participants expressed that their clients felt a loss of hope and confidence "in her ability to make friendships and more intimate relationships" (participant 10). For some clients, getting scammed facilitated awareness of differences in cognitive and functional abilities after the injury. However, increased insight into postinjury function also exacerbated emotional distress:

> *He spent a lot of time crying.* [Participant 15]

XSL•FO
RenderX

The injustice of the scam was particularly difficult to cope with:

> *Frustration...he seemed to struggle with managing his emotional response to wanting to help, feeling like a good guy...like he's doing the right thing for someone and then yet having all this other bad stuff happen as a result, and his life plummeting to new depths.* [Participant 16]

Anger was directed toward family and friends:

> *He's got anger with his mum, which he never would have done before, and with a friend that maybe hasn't called him back in time.* [Participant 15]

People with ABI also felt "silly...like an easy prey...robbed him of his pride" (participant 11). Anger owing to being scammed also contributed to broad escalations in challenging behavior:

> *Like he might grab them on the jugular and do stuff to them. Strangle them.* [Participant 11]

Some people with ABI attempted to retaliate against the scammers by re-engaging with them. Despite this appearing to be a reasonable motive, people with ABI increased their risk of another cyberscam in the future as "[scammers] are still keeping you on the line" (participant 16).

Extreme Mistrust

For some people with ABI, anger following the recognition of the cyberscam led to scam disengagement and heightened awareness of cyberscam red flags. However, cognitive difficulties related to ABI resulted in overgeneralizing from the online scam experience into everyday life, becoming extremely mistrusting and "cutting off other people that were legitimately his friend...because he's worried about people ripping him off now" (participant 15). Difficult experiences in reaching out to formal and informal supports contributed to additional distress and relationship impact for scam survivors with ABI. Reporting the scam to services such as the police and the resultant unhelpful outcomes exacerbated their "lack of confidence in engaging with services" (participant 10). In contrast, clients who did not accept that they had been scammed were angry with the support people's attempts to increase their awareness, contributing to a significant "level of mistrust with all services that were trying to help" (participant 1). A psychologist reported that by revealing their client's scam, their client was "momentarily angry that I ruined her friendship" (participant 12). Relationship breakdowns with family members and friends were common as "he's got anger with his mum, which he never would have done before" (participant 15). These experiences highlighted the compounding effect of cognitive difficulties (ie, lack of insight and overgeneralization) and psychological factors (eg, anger and loneliness) on important social relationships in the person's life.

Intervention Roadblocks

Psychological responses to the scam acted as major barriers to intervention by interfering with problem-solving to seek support:

> *The guilt and shame is so much that he's not even thought, "Hang on, I should go to the police. I've been ripped off."* [Participant 13]

Shame, in particular, served as an obstacle to help seeking from family owing to fear of judgment:

> *That kind of embarrassment, and the fact that he didn't want the son involved in this process much when he probably would have been a good support for him.* [Participant 13]

Significant anger and challenging behaviors triggered by discussing the scam prevented clinicians from directly addressing the scam during their interventions:

> *He would ramp up his agitation so then you'd have to switch topics, otherwise, he would get quite nasty.* [Participant 15]

When able to be provided, support from trusted others through reassurance and counseling curbed some negative emotional impacts. In contrast, restrictive practices and increased supervision were considered to have undesired effects on independence and relationships through deteriorations in trust:

> *[His mother] still had a sense of protectionism...she was checking his bank statements...the biggest impact was the fact that she didn't trust him anymore.* [Participant 16]

Needing Trust and Connection

Overview

Social factors increased the vulnerability to cyberscams and interacted with the life-changing impact of ABI. Psychological vulnerability to scams was closely tied to a strong desire for trusting romantic relationships and social connection as people with ABI "wanted the white picket fence family future that he'd always dreamt of" (participant 16). However, presence of a trusting and supportive relationship did not protect against cyberscams for people with ABI if their caregivers were not cyber aware.

Strong Desire for a Trusted Relationship and Intimacy

Although participants reported that some of their clients with ABI were aware of romance scam red flags, their strong drive for a relationship outweighed the potential for financial or emotional loss:

> *[The client] agreed not to do it again. But he...desperately wanted a relationship and he saw [the scammer] as the only solution.* [Participant 7]

This was often driven by social isolation and a "deep loneliness a lot of clients have that they just wanna be loved [and] have kids" (participant 8). People with ABI were often isolated, had difficulty in integrating into their community, and lacked proximal social support, increasing their motivation to search on the internet for connection:

> *She had to separate from her husband after a brain injury...had a string of relationships that didn't work out...a lack of meaningful activities in the community and few social supports so a lot of her world was online.* [Participant 2]

These same drivers for connection were also reinforcing for family members and carers, who encouraged people with ABI

XSL•FO
RenderX

to engage in online relationships. Compounded by their own limited cybersafety knowledge, family members were described as facilitating scam engagement:

> *Mum probably saw these as potential legitimate relationships...she would get angry with him and say, "Why are you being so rude to these people who are trying to be friends with you?"* [Participant 15]

Clinician-Client Relationship

The presence or absence of a strong therapeutic relationship was an instrumental social factor in addressing cyberscam vulnerability and interventions for people with ABI. People with ABI who were isolated and had limited clinical and support services were more susceptible to the grooming and psychological influence of scammers. They also lacked support from trusted therapists in identifying and managing cyberscam impacts. Not having a strong, trusted clinician-client relationship was a barrier to treatment effectiveness as "if they don't trust you, they just don't tell you anything" (participant 9). In contrast, interventions were considered to be effective when clinicians had a good awareness of their clients' needs and an established therapeutic alliance:

> *I knew her and how far I could push her.* [Participant 12]

Nonetheless, it was acknowledged that building a trusting working relationship was more challenging for clinicians who engaged with clients in a once-off setting (eg, conducting neuropsychological assessments).

Strong Reactions of Trusted Others

Overview

The combination of reduced insight and difficulty in identifying red flags, coupled with a strong desire for social connection kept people with ABI hooked onto the scam. This often led to significant financial loss, prompting strong reactions from clinicians and caregivers. Although some were supportive and involved, others were unsupportive and angry. Participants felt frustrated and helpless and lacked confidence in their own capacity to support their clients' recoveries from cyberscams.

"Tumultuous Relationships"

For people with ABI who commonly experienced postinjury social changes, "tumultuous relationship[s]" (participant 4) with family and other caregivers interfered with cyber awareness and seeking support. These relationships made it difficult for clients to share what was happening in their lives:

> *He'd only just resumed contact after they've had a falling out...he wouldn't consent to me contacting his son for an informant history.* [Participant 13]

Arguments between people with ABI and caregivers were common after the cyberscam—"huge fights with his family" (participant 4) and "breakdown of friendship" (participant 10). This was attributed by participants to the lack of insight and inability to accept the cyberscam among the people with ABI, combined with caregiver's blame of the person with ABI for the financial loss, reinforcing feelings of shame. People with ABI therefore lost important social supports that could serve as cyberscam-protective factors. Existing challenging behaviors also deteriorated relationships as difficulties with self-regulation "means that the discussions escalate to a point where you can't reasonably talk about it [and] have helpful discussions" (participant 3). Some caregivers therefore limited their interactions to avoid conflict. In contrast, some clients had supportive and reassuring caregivers helping them through the process by supporting online participation access and trying "to maintain that normal home life rather than a knee-jerk reaction" (participant 10).

Clinician's Distress

Participants described the experience of their client being scammed as "heartbreaking" (participant 2), noting that it "lingered for a while but more so than any other devastating thing that happens to our clients" (participant 6). The severity and unexpected nature of the scam situation and clients' significant financial, resource, and independence losses contributed to these persistent feelings. Participants working on established rehabilitation goals, expressed anger as "it takes you away from the pathway and the goals you would have been working towards" (participant 3). Cyberscam re-engagement contributed to feelings of frustration as "we'd built this man's independence really significantly...I didn't wanna see that disappear" (participant 7). Nevertheless, some participants reported addressing cyberscam experiences as an opportunity for professional development and change in clinical practice:

> *I ask about online relationships now, which I never used to.* [Participant 9]

"Nothing Structured to Do"

Overview

Participants noted that unstructured and purposeless time meant that people with ABI were spending more time on the internet and were more socially isolated, highlighting increased cyberscam vulnerability. Loss of funds and restrictions in self-management of finances owing to the scam often resulted in reduced capacity to participate in meaningful activities and engage with the wide community.

Reduced Meaningful Engagement

Unstructured time exacerbated the vulnerability to scams through extensive online engagement, such as a neuropsychologist's client with ABI who had an "addictive personality and...was using his iPad for a lot of his waking hours" (participant 3). Although it served a purpose as a form of community engagement ("it was filling a gap in her social isolation" [participant 2]), more screen time increased clients' cyberscam exposure. Even people with ABI who had structured daily activities were vulnerable and likely to engage in cyberscams during evenings and nights:

> *It's very hard to fill those gaps, from a therapeutic, rehab, work perspective.* [Participant 7]

Being isolated during this time frame increased cyberscam vulnerability owing to ABI-related cognitive impairments (eg, reduced insight into scams and impaired memory for red flags), which was compounded by other factors such as unsupervised access to technology and desire for relationship.

XSL•FO
RenderX

Following the scam, participants described that people with ABI became less likely to engage in meaningful activities for several reasons. For some, this related to loss of self-confidence in navigating social events and meeting new people:

> *Because of the risks of doing the wrong thing – especially if it's a relationship scam – trusting people again.* [Participant 19]

For others, infatuation with the scammer reduced interest in meaningful participation such as "Men's Shed or...a woodwork group...because [the scam] was his focus" (participant 3).

Scams that resulted in significant financial loss also led to the loss of ability to afford leisure activities and led to clients having excessive unfilled time.

Financial Stress and Independence

Overview

Participants identified that clients experienced financial stress and changes to independence before and after the cyberscam. This was often paired with the strong reactions of trusted others who sometimes turned to restrictive practices as a way of financially protecting their loved ones. Being under financial administration before the cyberscam often left clients without sufficient financial self-management skills. Some clients had disposable income and financial resources to accommodate scammer requests; wealthy clients ended up with significant financial losses. Some clients who independently handled their financial affairs were deemed to have lost their financial decision-making capacities in the face of the scam.

Limited Financial Knowledge and Independence

Cognitive impairments after ABI often reduced financial literacy. More than half of participants (11/20, 55%) reported that their clients received financial oversight (eg, administration orders) or informal support. Financial administration alone, however, was insufficient to prevent cyberscam engagement:

> *Even though his funds are protected...he had enough access to that amount of money [$500] to transfer that.* [Participant 5]

Nonetheless, restricted financial access limited significant financial loss to scammers:

> *He couldn't give her the money because of this "stupid administration order" and [financial trustee].* [Participant 3]

Although financial loss was significantly reduced, financial administration orders were often described as paternalistic and left clients without appropriate financial management skills as "he wouldn't have developed those skills in managing his own money and setting up those practical things" (participant 13).

Some people with ABI arranged to have financial administration orders removed at the behest of the cyberscammer. Attempts by people with ABI to demonstrate their ability to manage their finances independently interfered with client-clinician relationships as clients hid key information:

> *He felt that he needed to prove that he wasn't spending his money in a poor way, that his decision-making wasn't bad, so he wasn't always keen to let us know what he was spending his money on.* [Participant 4]

Moreover, restricted financial access increased clients' desire for control in other aspects of their life. Romance scam vulnerability was increased as clients were "more driven towards this goal [finding someone] because it's the only thing he thinks he could independently do...that's not restricted" (participant 3).

Loss of Finances and Decision-Making Capacity

Cyberscam engagement resulted in the loss of finances and important resources. Several participants (17/20, 85%) reported that their clients had lost all of their savings, inheritance, and compensation payouts:

> *Sold all his shares...AUD$250,000 [USD $168,801] that went overseas.* [Participant 6]
>
> *Wiped out most of his savings.* [Participant 14]
>
> *A really large financial compensation payout...all gone.* [Participant 16]

Financial loss contributed to significant loss of independence and participation in meaningful activities, which often exacerbated psychological distress. A client "couldn't afford to do activities that he previously...enjoyed because his resources were completely depleted" (participant 16). Examples include the inability to pay rent, pay for car registration, and continue living their current lifestyle (eg, traveling overseas and going to concerts and football games). The reduced ability to engage in these important social and leisurely pursuits significantly affected the quality of life and increased social isolation.

Financial independence was at risk, with some clients requiring a review of their decision-making capacity after the scam:

> *He was so clearly hugely vulnerable and unaware of how vulnerable he was that I actually made the recommendation that he probably shouldn't even be managing his disability pension.* [Participant 17]

Loss of financial capacity often left clients feeling vulnerable about "losing his independence separate to the scam experience" (participant 16). For clients with limited insight into postinjury impairments, losing financial access and dignity to take risks resulted in emotional distress:

> *The depression...didn't come from being scammed. It came from quite the opposite, almost like he...feels like it's his right to have access and leave himself open to those things.* [Participant 17]

In contrast, some participants, albeit a minority, reported little to no financial loss for their clients. Restricted access to funds, carer scam knowledge, and supervision contributed to "near-misses":

> *His carer very quickly identified that he was having [scammer] requests.... He just didn't have the access to money to be able to send that through.* [Participant 10]

XSL•FO
RenderX

Cyberability

Overview

Participants described that online engagement was seen as a means to fill the gaps of reduced meaningful participation and subsequent social isolation for people with ABI; however, basic levels of cybersafety knowledge and awareness was required. Participants provided examples of their own cybersafety experiences such as awareness of social media privacy settings, blocking and deleting scam calls or emails, and being skeptical of unexpected requests for information. Participants reported that reduced levels of cyberability in their clients and themselves increased cyberscam susceptibility. In contrast, adequate cyberability might also create a false sense of security and overconfidence in online behavior.

Limited Technological Exposure and Knowledge

Limited cyberability presented as a lack of understanding of the internet and computers. Preinjury factors including socioeconomic disadvantage or incarceration led to limited knowledge and exposure to technology:

> *Coming out of correction orders, they haven't had any technology for five to ten years.* [Participant 1]

Lack of technological exposure also reduced awareness of personal privacy protection, which was perpetuated by ABI-related cognitive impairments:

> *She wasn't a high internet user prior to her brain injury, now all of sudden she found herself on social media sites...using an iPad...and really has no awareness of protecting her privacy, much more than the general person would experience.* [Participant 2]

Lack of direct observation by clinicians made it challenging to ascertain whether people with ABI participated in safe online behavior, even those residing in 24-hour supported accommodation. With limited cybersafety awareness, unsupervised access to technology was concerning to participants as "you wouldn't have a clue what they're up to" (participant 5). For people with ABI who were dependent on others, having trusted relationships protected them from cyberscam attempts as "he has somebody in the room with him...they just happened to notice what he was doing and intervened" (participant 10). However, this was with the caveat that caregivers had strong cyberability:

> *I've had stories of people whose support worker has clicked the [scam] link for them.* [Participant 16]

In contrast, sound technological knowledge and online independence was considered to be insufficient in protecting against cyberscams as it created a false sense of security:

> *I think perversely his technical prowess with IT stuff...and obviously scammers rather exploit that vulnerability as well.* [Participant 8]

This was compounded by the cognitive-executive difficulties experienced by people with ABI who "...don't have the ability to filter and work out what's safe and what's not safe" (participant 11). Individuals might therefore appear overconfident about their cybersafety skills, potentially masking their vulnerability, and this results in low motivation to engage in scam prevention education.

Overall, 25% (5/20) of the participants spontaneously reported that they had used the web-based CyberAbility [17] training program and found this to be helpful in improving the cybersafety skills of their clients with ABI, with a particular benefit of reducing scam stigma and improving scam prevention:

> *We did the CyberAbility, it's like, "Yes, it's happened to lots of people!" and it's never happened again.* [Participant 15]

The program helped to facilitate discussions with clients as it "allowed me to have more conversation with her about this than I ever have...because it's a more formalised way" (participant 18).

"Anti-Facebook"

Some participants noted a change after the cyberscam where clients stopped using social media, emails, and internet owing to generalized mistrust of technology. Despite this conferring protection against future cyberscams, this negatively affected their social communication, as some participants expressed that "it's very excessive because she's reluctant to go on social media...emails" (participant 10). A participant shared that although their client restricted his digital engagement, this was only short term as "he likes...the connection to the world through his computer and the convenience. But he will not go back on to any type of online banking" (participant 20). This may negatively affect function with reduced accessibility to services, transaction delays, and potential security concerns. Online re-engagement also did not return to prescam patterns for many clients.

Scammer Persuasion

Overview

Scammers leveraged the limited levels of cyberability among people with ABI and their desire for a relationship to keep them hooked onto the scam. Participants described that scammers used tactics such as love bombing, personalization, constant reinforcement, and isolation to keep the clients hooked onto the cyberscam:

> *The person online was very nice, made her feel very happy and attractive and confident.* [Participant 10]

Coupled with cognitive-executive difficulties, people with ABI therefore had difficulty in disengaging from ongoing contact with the scammer and resisting subsequent scam attempts.

"Sucked in 'Cause it Sounds So Convincing"

Participants identified that scammers used love bombing strategies to groom and manipulate clients by making them feel loved and wanted and promising a future together. Scammers used clients' trusting personalities to keep them hooked:

> *It was a person saying, "You're amazing, you're capable, you're worthy of being loved," and he [client] was saying, "that is attractive."* [Participant 16]

XSL•FO
RenderX

Scammers also manipulated clients' generosity and fabricated common experiences, such as trauma:

> *There is this Aussie spirit of helping a mate when your chips are down, coupled with...that sense of, "We've got this shared trauma, let's get together and we'll be stronger for it."* [Participant 16]

Scammer techniques were increasingly sophisticated and personalized ("he'd say he had similar experiences and it was all just too perfect" [participant 9]), therefore reinforcing the desire for connection. Scammer strategies evolved with clients' patterns of online engagement, which made it difficult even for clinicians to identify the red flags and recognize the scam:

> *As a therapist, I get sort of bit sucked in too 'cause it sounds so convincing.* [Participant 8]

Relentless Pressure

Participants described that constant reinforcement and availability was a grooming tactic that scammers used to create positive experiences for people with ABI. The relentless requests for money made it difficult for clients to disengage:

> *He would express a lot of love for her...then it started to be scattered through quite often, "What about the money, babe? What about the money?"* [Participant 3]

The grooming was so effective that some clients continued engaging with the scammer and "would confide less in them [family]...because most of his life was thinking about her [scammer]" (participant 3). Scammers isolated people with ABI from their social and support networks and reinforced the scam relationship, with participants noting reduced client contact "because they [scammers] would speak negatively of anyone she trusted and was close to...she became more and more separate from her relationships" (participant 9). People with ABI ultimately trusted the scammers instead "because he thought it was everybody else that was doing the wrong thing, not her [scammer]" (participant 16). Most people with ABI were only able to disengage from the scam when scammers stopped contact or when participants provided education or implemented strategies.

Treatment Recommendations

Overview

The significant cyberscam vulnerabilities and impacts experienced by individuals with ABI, identified through each of the 8 themes, warrant the need for targeted interventions. Our participants provided specific cyberscam recovery recommendations to address each of these factors. Given the multifaceted aspects of cyberscam vulnerability and disconnection between knowing and doing, which is common for people with ABI, a multicomponent biopsychosocial approach to interventions was considered to be necessary.

Participants recommended guidelines such as "a clinician toolkit" (participant 8), with clear "referral pathway or resources" (participant 13), especially for clinicians working in assessment roles. Professional development capacity building workshops were suggested:

> *More extensive training around this concept for the rehab community generally.* [Participant 7]

Clinicians who attended previous CyberAbility workshops had implemented these learnings into their practice, improving cyberscam awareness among people with ABI, which reduced cyberscam engagement and financial loss:

> *...That awareness is there...those are part of our conversations we have with [clients] in preventing it.* [Participant 11]

Participants highlighted the importance of working in multidisciplinary teams as it "sped up the process a bit" (participant 15).

Need to Build Awareness

Participants described that building awareness was key to understanding and increasing scam knowledge. Given the cognitive-executive vulnerabilities to cyberscams among people with ABI, scam education needs to be provided in a tailored way as "people are not going to accept interventions until there's some self-awareness or ownership" (participant 20). Examples included scam education (eg, red flags) and positive behavior support strategies to challenge the veracity of beliefs about the scammer (eg, "Your Theory is that this is your boyfriend, My Theory is that this is a scammer").

Normalization and Psychological Interventions

Most participants (15/20, 75%) recommended psychoeducation and cyberscam normalization to reduce the psychological impacts experienced after the cyberscam. People with ABI could share and listen to lived experiences of others who had been scammed as "maybe this is a safe place and I can say what's been on my mind that I maybe was wanting to tell someone but hadn't felt brave enough" (participant 16). Mood support drawn from cognitive behavioral therapy and acceptance commitment therapy approaches were also recommended.

Trusted Person in the Community

Reduced everyday social interactions for people with ABI also required targeted input to establish alternative pathways to check if something is a scam in real time. Participants recommended having therapists as a consistent ally and engaging a trusted person in the community (eg, lived experience experts and guardians) for this role. This could reduce clients' cyberscam engagement and potentially reduce social isolation—for example, regular therapy sessions that emphasized building rapport and trust and teaching people with ABI how to establish healthy relationships. Strong therapeutic rapport and trusting relationships within the broad support team also facilitated positive discussions about scam situations:

> *Even though he didn't like or agree necessarily with some of the conversations and their angle, he liked them as humans.* [Participant 16]

Having trusted relationships in the community helped to reduce social isolation by increasing community engagement such as "organising Bingo, social events, raffles, and the responsibility of looking after the money involved" (participant 20). Workshops for caregivers were also recommended to increase the awareness and understanding of cyberscams and ABI.

XSL•FO
RenderX

Involving caregivers whom clients "can do homework with" (participant 17) could also ensure intervention sustainability. Nonetheless, some participants were cautious about including caregivers to protect their client's privacy.

Cyberability Practical Strategies

Other recommendations to address social factors included practical strategies and skill building. Most participants (18/20, 90%) recommended using practical strategies to reduce cyberscam vulnerability and stop scammer contact—for example, reporting the cyberscam to authorities, supporting clients to screen and block numbers or emails, and teaching password and account management. Participants suggested increasing clients' cybersafety and technological skills and awareness when communicating and socializing with others by "not disclosing all that information until you've seen them in person" (participant 11). Some participants also suggested increasing cybersafety through the web-based CyberAbility [17] training program and working on the root cause of the client's cyberscam engagement (eg, strong desire for a relationship).

"Least Restrictive Practice"

Participants noted that financial skill building (eg, budget management) could reduce clients' cyberscam vulnerability through increasing ownership and understanding of their funds. Although some participants recommended imposing financial restrictions and administration orders, these were mainly considered as a last resort. Most participants (14/20, 70%) suggested operating "from a least restrictive practice" (participant 7) and "getting other measures in place to protect him other than just appointing [a financial trustee]" (participant 13) based on the clients' stage in the cyberscam journey "because it's also about the person's learning...and skill development" (participant 19). Although more restrictive practices such as limiting phone and internet access were also recommended, these were implemented for clients with low awareness or ability to learn from mistakes:

> *She's just unable to flexibly change or adapt...you can only concretely manage...restrict her environment and her access to money or phones. That's the only way she would be 100% safe, which is very challenging.* [Participant 2]

A strong and trusting clinician-client relationship was important to address clients' fear of losing independence if they opened up to clinicians about being scammed. Participants also recommended broad monitoring of clients' rehabilitation progress to facilitate gradual return to self-management of their finances, online activities, lifestyle, and relationships.

Meaningful Activities

Community engagement was highlighted as essential to reducing cyberscam engagement and social isolation. A way to do that was to engage clients in meaningful activities to help them integrate and re-engage with others in the community. For instance, volunteering, returning to work, developing hobbies, participation in projects they valued, and sharing lived experiences helped clients find meaning through contributing:

> *He loves being in the place of giving...He liked being the person to tell his story and help somebody else.* [Participant 11]

Clinician's Capacity

Clinician-related factors reflected a barrier to treatment when they did not feel confident in addressing the cyberscam with their clients:

> *It's very, very complex...it brings in every skill that you have as a clinician to work in this area with the client.* [Participant 2]

Limited knowledge, experience, and resources reduced confidence:

> *I felt awful. I had no concept, no idea. I felt silly and dumb...completely unprepared...unaware and unable to support as well.* [Participant 14]

Moreover, participants were unsure if addressing the scam was within their scope of practice and role:

> *What is my responsibility?* [Participant 6]

Some participants shared that their interventions had insufficient intensity and reach. Reduced capacity limited the level of support and expertise that participants could provide to their clients.

Discussion

Principal Findings

People with ABI may be more susceptible to cybercrimes owing to their postinjury consequences, specifically, cognitive and psychosocial difficulties [8,9]. This study aimed to qualitatively explore clinicians and service providers' experiences in managing cyberscams among people with ABI. Reflexive thematic analysis identified 8 themes that described cyberscam vulnerabilities and impacts among people with ABI, reflecting a broad array of biopsychosocial aspects.

Cyberscams are rife within the wide community and population with disability. These scams often target individuals, businesses, or organizations and can lead to financial losses, emotional distress, and other adverse outcomes [21]. Although numerous forms of cyberscams such as investment scams and identity theft exist, most scams reported to affect people with ABI were romance scams. As such, most of our discussion will focus on the vulnerabilities and impacts of romance scams. Romance scams are a low-volume scam but have significant financial and emotional impacts.

Clinicians identified a range of financial, emotional, cognitive, and psychosocial factors that increased the vulnerability of people with ABIs to cyberscams. The vulnerability to cyberscams bestowed by cognitive-executive impairment, social isolation, unstructured time, and trusting personalities was consistent with reports from people with ABI and their COs with cyberscam experience [8] and are commonly experienced after the injury [2,36]. Impaired insight and difficulty in generalizing from past experiences were described by clinicians as key contributing factors to being scammed. Although participants recommended building insight through scam

XSL•FO
RenderX

education, it is worthwhile to note that a discrepancy between "knowing and doing" may remain. Specifically, even when people with ABI possess the awareness of how to determine authenticity, limited insight and awareness make executing this challenging in practice, particularly when manipulated by scammer grooming, threats, and time-pressure tactics [37].

Survivors in the general community often report a "double-hit" of both financial and emotional impacts following romance scams [20]. In this study, participants reported significant financial losses experienced by their clients owing to various scams. In ABI, this financial impact is of additional concern owing to great difficulty in reaccumulating funds after the injury, commonly owing to a once-off compensation payout and high rates of unemployment [2]. This was compounded by the psychosocial impact on participation and loss of independence in the daily activities and social pursuits in which their clients engaged previously.

Similar to previous studies in the general community and among people with ABI [8,20,38,39], significant emotional distress and feelings of shame, anger, and loss of trust were experienced after the scam. Taken together, people with ABI may actually experience a "triple-hit" regarding the loss of finances, reduced independence, and emotional consequences after the scam. Whether this is unique to people with ABI or extends to others with disabilities and psychosocial vulnerabilities or the wide population requires verification.

The emerging evidence for a preponderance of romance scams among people with ABI in this and other related studies [8,9] warrants attention. Loneliness and a strong desire for a trusted relationship were described as factors that increase romance scam vulnerability among people with ABI. In contrast, literature evaluating these vulnerabilities in the general population is mixed [40-42]. Scammers typically leverage these vulnerabilities to groom and isolate people from their support networks [43]. Unfortunately, after the scam, people with ABI may experience even great social disconnection owing to conflict with family and reduced financial means for community participation. Therefore, people with ABI may be more at risk of continuing to engage in the current or additional romance scams that offer what is perceived as a meaningful connection. Potentially, interventions that target the development and maintenance of healthy relationships may help to reduce this underlying vulnerability.

Overall, it is essential to recognize the compounding interaction between the psychosocial consequences of ABI and experiencing cyberscams. Cognitive-executive impairments unique to ABI further complicate cyberscam vulnerability, creating a perfect storm that amplifies both vulnerability and the complexity of managing the distress and aftermath of the cyberscam. Crucially, restrictive interventions that initially appeared to be protective (eg, restricted financial control and checking bank statements) had unintended consequences owing to the interaction between the cyberscam and the cognitive aspects of ABI (eg, increased scam awareness resulted in globalized mistrust, affecting relationships). Consequently, a biopsychosocial, flexible, individualized, multidisciplinary, and multicomponent approach to supporting clients to recover from cyberscams was frequently adopted, similar to most other ABI-specific interventions [44-47]. Key recommended strategies included scam education and behavioral modifications (eg, practicing safe password account management and community engagement), consistent with those reported by people with ABI and their COs [8]. Most participants (14/20, 70%) emphasized the use of least restrictive practices to ensure clients' autonomy and trust. On the basis of our findings, clinical experience, and wide literature, we recommend a biopsychosocial framework that targets the psychological, social, and functional vulnerabilities and impacts of cyberscams, while accounting for the ABI-related cognitive impairments that complicate intervention and recovery. Importantly, digital upskilling of individuals with ABI needs to be prioritized, given the rapid technological advancements and scams that accompany it. Nonetheless, it is also essential for the technological sector to actively address the needs of people with ABI and other disabilities to ensure a safe and more inclusive space for online participation.

Within the context of these treatment experiences and the difficulties in obtaining good recovery outcomes despite implementing routine interventions, significant clinician distress was identified. To the best of our knowledge, our paper highlights original findings of the secondary impact on professionals who support scam survivors. Most clinicians (19/20, 95%) described lacking knowledge, clinical guidance, and capacity to provide support specifically tailored to cyberscams for people with ABI. In addition to feeling underequipped, they also described feeling distressed and frustrated by learning about and trying to support individuals with ABI who had been scammed. Together with the lack of resources, these findings underscore the need to provide clinicians with the desired targeted training and resources to guide practice and increase confidence and capacity to support people with ABI to avoid and recover from cyberscams. Directed toward prevention, the web-based CyberAbility [17] training program was frequently described as helpful in facilitating conversations and educating clients about cyberscams and had resulted in several examples of scam avoidance. The utility of this approach in prevention and treatment requires additional evidence.

Although we captured a broad range of perspectives regarding the treatment landscape across disciplines in Australia, given the global threat of cybercrime, future studies could explore international experiences in supporting people with ABI to recover from postcyberscam impacts. The generalizability of these findings beyond ABI to other disability and vulnerable cohorts also requires examination.

Conclusions

To the best of our knowledge, this is the first study to qualitatively explore the perspectives of clinicians and service providers in addressing cyberscam experiences among people with ABI. We identified a framework to assist in understanding the vulnerabilities and impacts of cyberscams, which aligns with a biopsychosocial model. The framework guides the design of cyberscam interventions for people with ABI, which are individualized, comprehensive, and targeted to biopsychosocial factors. These findings are now being used to co-design a

XSL•FO
RenderX

cyberscam psychological recovery intervention with people with ABI.

Acknowledgments

The authors would like to acknowledge the participants for their time. No funding was obtained for this study.

Authors' Contributions

All the authors contributed to the study conception and design. Data collection and analysis were led by KAC and KRG. JP supported the data analysis stage. KRG and JP oversaw the whole study and provided guidance as needed as KAC's doctoral supervisors. The first draft of the manuscript was written by KAC. Critical review, revisions, and approval were provided by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview schedule.
[DOCX File , 17 KB - jmir_v26i1e51245_app1.docx]

References

1. Disability in Australia: acquired brain injury. Australian Institute of Health and Welfare. 2007. URL: https://www.aihw.gov.au/getmedia/1f719b27-6b93-434a-b0e6-997b4ead061a/bulletin55.pdf [accessed 2023-03-29]
2. Ponsford JL, Downing MG, Olver J, Ponsford M, Acher R, Carty M, et al. Longitudinal follow-up of patients with traumatic brain injury: outcome at two, five, and ten years post-injury. J Neurotrauma 2014 Jan 01;31(1):64-77. [doi: 10.1089/neu.2013.2997] [Medline: 23889321]
3. Adams AG, Schweitzer D, Molenberghs P, Henry JD. A meta-analytic review of social cognitive function following stroke. Neurosci Biobehav Rev 2019 Jul;102:400-416. [doi: 10.1016/j.neubiorev.2019.03.011] [Medline: 30922978]
4. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 08;196(4286):129-136. [doi: 10.1126/science.847460] [Medline: 847460]
5. Brunner M, Hemsley B, Togher L, Dann S, Palmer S. Social media and people with traumatic brain injury: a metasynthesis of research informing a framework for rehabilitation clinical practice, policy. Am J Speech Lang Pathol 2021 Jan 27;30(1):19-33. [doi: 10.1044/2020_AJSLP-20-00211] [Medline: 33332986]
6. Brunner M, Palmer S, Togher L, Hemsley B. 'I kind of figured it out': the views and experiences of people with traumatic brain injury (TBI) in using social media-self-determination for participation and inclusion online. Int J Lang Commun Disord 2019 Mar 05;54(2):221-233 [FREE Full text] [doi: 10.1111/1460-6984.12405] [Medline: 29873159]
7. Targeting scams: report of the ACCC on scams activity 2022. Australian Competition and Consumer Commission. 2023. URL: https://www.accc.gov.au/system/files/Targeting%20scams%202022.pdf [accessed 2023-03-29]
8. Gould KR, Carminati JJ, Ponsford JL. "They just say how stupid I was for being conned". Cyberscams and acquired brain injury: a qualitative exploration of the lived experience of survivors and close others. Neuropsychol Rehabil 2023 Mar 27;33(2):325-345. [doi: 10.1080/09602011.2021.2016447] [Medline: 34957919]
9. Gould KR, Carolan M, Ponsford JL. Do we need to know about cyberscams in neurorehabilitation? A cross-sectional scoping survey of Australasian clinicians and service providers. Brain Impair 2023 Sep;24(2):229-244. [doi: 10.1017/BrImp.2022.13] [Medline: 38167197]
10. Whitty MT. Mass-marketing fraud: a growing concern. IEEE Secur Priv 2015 Jul;13(4):84-87 [FREE Full text] [doi: 10.1109/msp.2015.85]
11. Henry N, Flynn A, Powell A. Image-based sexual abuse: victims and perpetrators. Trends and Issues in Crime and Justice, Government of Australia. 2019. URL: https://www.aic.gov.au/sites/default/files/2020-05/imagebased_sexual_abuse_victims_and_perpetrators.pdf [accessed 2023-03-29]
12. Online hate speech. eSafety Commissioner. URL: https://www.esafety.gov.au/research/online-hate-speech [accessed 2023-03-29]
13. State of play: youth, kids and digital dangers. eSafety Commissioner. URL: https://www.esafety.gov.au/research/youth-digital-dangers [accessed 2023-03-29]
14. Han SD, Boyle PA, James BD, Yu L, Bennett DA. Mild cognitive impairment and susceptibility to scams in old age. J Alzheimers Dis 2016;49(3):845-851 [FREE Full text] [doi: 10.3233/JAD-150442] [Medline: 26519434]
15. Ueno D, Daiku Y, Eguchi Y, Iwata M, Amano S, Ayani N, et al. Mild cognitive decline is a risk factor for scam vulnerability in older adults. Front Psychiatry 2021 Dec 20;12:685451 [FREE Full text] [doi: 10.3389/fpsyt.2021.685451] [Medline: 34987422]

XSL•FO
RenderX

16. Kapasi A, Schneider JA, Yu L, Lamar M, Bennett DA, Boyle PA. Association of stroke and cerebrovascular pathologies with scam susceptibility in older adults. JAMA Neurol 2023 Jan 01;80(1):49-57 [FREE Full text] [doi: 10.1001/jamaneurol.2022.3711] [Medline: 36315115]
17. CyberAbility. URL: https://cyberability.org.au/ [accessed 2023-03-29]
18. Brunner M, Rietdijk R, Summers K, Southwell K, Avramovic P, Power E, et al. 'It gives you encouragement because you're not alone': a pilot study of a multi-component social media skills intervention for people with acquired brain injury. Int J Lang Commun Disord (Forthcoming) 2022 Nov 23. [doi: 10.1111/1460-6984.12806] [Medline: 36417179]
19. Cross C, Dragiewicz M, Richards K. Understanding romance fraud: insights from domestic violence research. Br J Criminol 2018;58(6):1303-1322 [FREE Full text] [doi: 10.1093/bjc/azy005]
20. Whitty MT, Buchanan T. The online dating romance scam: the psychological impact on victims – both financial and non-financial. Criminol Crim Justice 2015 Sep 03;16(2):176-194 [FREE Full text] [doi: 10.1177/1748895815603773]
21. Button M, Lewis C, Tapley J. Not a victimless crime: the impact of fraud on individual victims and their families. Secur J 2012 Apr 23;27(1):36-54 [FREE Full text] [doi: 10.1057/sj.2012.11]
22. Victim resources. ActionFraud. 2022. URL: https://www.actionfraud.police.uk/ [accessed 2023-03-29]
23. The little black book of scams. Australian Competition and Consumer Commission. URL: https://www.accc.gov.au/about-us/publications/the-little-black-book-of-scams [accessed 2023-03-29]
24. Gould KR, Brokenshire C. Scams and brain impairment: a clinician's treatment recommendations and a survivor's perspective. Brain Impair 2017;18(3):337-419 [FREE Full text]
25. Carminati JJ, Ponsford JL, Gould KR. "This group… I felt like I was medicating myself from this cyberscam illness that was living with me." A qualitative evaluation of co-designing cybersafety training resources with and for people with acquired brain injury. Disabil Rehabil 2023 Nov 07;45(22):3719-3729. [doi: 10.1080/09638288.2022.2139418] [Medline: 36342759]
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: Sage Publications; Mar 01, 2022.
28. Lincoln YS, Lynham SA, Guba EG. Paradigmatic controversies, contradictions, and emerging confluences, revisited. In: Denzin NK, Lincoln YS, editors. The SAGE Handbook of Qualitative Research. Thousand Oaks, CA: Sage Publications; 2011.
29. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
30. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020 Aug 12;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
31. NVivo. QSR International. 2019. URL: https://help-nv11.qsrinternational.com/desktop/welcome/welcome.htm [accessed 2023-03-29]
32. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922 [FREE Full text] [doi: 10.1177/1609406919899220]
33. Campbell JL, Quincy C, Osserman J, Pedersen OK. Coding in-depth semistructured interviews: problems of unitization and intercoder reliability and agreement. Sociol Methods Res 2013 Aug 21;42(3):294-320 [FREE Full text] [doi: 10.1177/0049124113500475]
34. Guest G, Bunce A, Johnson L. How many interviews are enough?: An experiment with data saturation and variability. Field Methods 2016 Jul 21;18(1):59-82 [FREE Full text] [doi: 10.1177/1525822X05279903]
35. Collaborative whiteboard. Miro. 2011. URL: https://miro.com/ [accessed 2023-03-29]
36. Ezekiel L, Collett J, Mayo NE, Pang L, Field L, Dawes H. Factors associated with participation in life situations for adults with stroke: a systematic review. Arch Phys Med Rehabil 2019 May;100(5):945-955. [doi: 10.1016/j.apmr.2018.06.017] [Medline: 29981316]
37. Cross C, Layt R. "I suspect that the pictures are stolen": romance fraud, identity crime, and responding to suspicions of inauthentic identities. Soc Sci Comput Rev 2021 Mar 19;40(4):955-973 [FREE Full text] [doi: 10.1177/0894439321999311]
38. Button M, Cross C. Cyber Frauds, Scams and Their Victims. London, UK: Taylor & Francis Group; 2017.
39. Notté R, Leukfeldt ER, Malsch M. Double, triple or quadruple hits? Exploring the impact of cybercrime on victims in the Netherlands. Int Rev Vict 2021 Jun 21;27(3):272-294 [FREE Full text] [doi: 10.1177/02697580211010692]
40. Cross C. 'They're very lonely': understanding the fraud victimisation of seniors. Int J Crime Justice Soc Democr 2016 Dec 01;5(4):60-75 [FREE Full text] [doi: 10.5204/ijcjsd.v5i4.268]
41. Coluccia A, Pozza A, Ferretti F, Carabellese F, Masti A, Gualtieri G. Online romance scams: relational dynamics and psychological characteristics of the victims and scammers. a scoping review. Clin Pract Epidemiol Ment Health 2020 Apr 20;16(1):24-35 [FREE Full text] [doi: 10.2174/1745017902016010024] [Medline: 32508967]
42. Whitty MT. Do you love me? Psychological characteristics of romance scam victims. Cyberpsychol Behav Soc Netw 2018 Feb;21(2):105-109 [FREE Full text] [doi: 10.1089/cyber.2016.0729] [Medline: 28657792]

XSL•FO
RenderX

43. Williams EJ, Beardmore A, Joinson AN. Individual differences in susceptibility to online influence: a theoretical review. Comput Human Behav 2017 Jul;72:412-421 [FREE Full text] [doi: 10.1016/j.chb.2017.03.002]
44. Ponsford J, Lee NK, Wong D, McKay A, Haines K, Alway Y, et al. Efficacy of motivational interviewing and cognitive behavioral therapy for anxiety and depression symptoms following traumatic brain injury. Psychol Med 2016 Apr;46(5):1079-1090. [doi: 10.1017/S0033291715002640] [Medline: 26708017]
45. Sander AM, Clark AN, Arciniegas DB, Tran K, Leon-Novelo L, Ngan E, et al. A randomized controlled trial of acceptance and commitment therapy for psychological distress among persons with traumatic brain injury. Neuropsychol Rehabil 2021 Aug 15;31(7):1105-1129. [doi: 10.1080/09602011.2020.1762670] [Medline: 32408846]
46. Brunner M, Rietdijk R, Togher L. Training resources targeting social media skills to inform rehabilitation for people who have an acquired brain injury: scoping review. J Med Internet Res 2022 Apr 28;24(4):e35595 [FREE Full text] [doi: 10.2196/35595] [Medline: 35482369]
47. Togher L, Douglas J, Turkstra LS, Welch-West P, Janzen S, Harnett A, et al. INCOG 2.0 guidelines for cognitive rehabilitation following traumatic brain injury, part IV: cognitive-communication and social cognition disorders. J Head Trauma Rehabil 2023;38(1):65-82 [FREE Full text] [doi: 10.1097/HTR.0000000000000835] [Medline: 36594860]

Abbreviations

ABI: acquired brain injury
CO: close other
COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by A Mavragani; submitted 04.08.23; peer-reviewed by C Jarko, M Button; comments to author 13.09.23; revised version received 15.10.23; accepted 25.12.23; published 29.01.24.

Please cite as:
Chew KA, Ponsford J, Gould KR
Addressing Cyberscams and Acquired Brain Injury ("I Desperately Need to Know What to Do"): Qualitative Exploration of Clinicians' and Service Providers' Perspectives
J Med Internet Res 2024;26:e51245
URL: https://www.jmir.org/2024/1/e51245
doi:10.2196/51245
PMID:38285489

©Kimberly Ann Chew, Jennie Ponsford, Kate Rachel Gould. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The EMPOWER Occupational e–Mental Health Intervention Implementation Checklist to Foster e–Mental Health Interventions in the Workplace: Development Study

Alberto Raggi[1], PsyD, PhD; Renaldo M Bernard[2], PhD; Claudia Toppo[1], MSc; Carla Sabariego[2,3,4], PhD; Luis Salvador Carulla[5,6], PhD; Sue Lukersmith[5], PhD; Leona Hakkaart-van Roijen[7], PhD; Dorota Merecz-Kot[8], PhD; Beatriz Olaya[9,10], PhD; Rodrigo Antunes Lima[9,10], PhD; Desirée Gutiérrez-Marín[9], PhD; Ellen Vorstenbosch[2,9,10], MSc; Chiara Curatoli[1], MSc; Martina Cacciatore[1], MSc

[1]Neurology, Public Health and Disability Unit, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milano, Italy
[2]Swiss Paraplegic Research, Nottwil, Switzerland
[3]Faculty of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland
[4]Center for Rehabilitation in Global Health Systems, University of Lucerne, Lucerne, Switzerland
[5]Health Research Institute, University of Canberra, Canberra, Australia
[6]Healthcare Information Systems (CTS553), University of Cadiz, Cadiz, Spain
[7]Erasmus School of Health Policy and Management, Erasmus University, Rotterdam, Netherlands
[8]Institute of Psychology, University of Lodz, Lodz, Poland
[9]Research, Innovation and Teaching Unit, Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain
[10]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain

Corresponding Author:
Alberto Raggi, PsyD, PhD
Neurology, Public Health and Disability Unit
Fondazione IRCCS Istituto Neurologico Carlo Besta
Via Celoria 11
Milano, 20133
Italy
Phone: 39 0223943105
Email: alberto.raggi@istituto-besta.it

Abstract

Background: Occupational e–mental health (OeMH) interventions significantly reduce the burden of mental health conditions. The successful implementation of OeMH interventions is influenced by many implementation strategies, barriers, and facilitators across contexts, which, however, are not systematically tracked. One of the reasons is that international consensus on documenting and reporting the implementation of OeMH interventions is lacking. There is a need for practical guidance on the key factors influencing the implementation of interventions that organizations should consider. Stakeholder consultations secure a valuable source of information about these key strategies, barriers, and facilitators that are relevant to successful implementation of OeMH interventions.

Objective: The objective of this study was to develop a brief checklist to guide the implementation of OeMH interventions.

Methods: Based on the results of a recently published systematic review, we drafted a comprehensive checklist with a wide set of strategies, barriers, and facilitators that were identified as relevant for the implementation of OeMH interventions. We then used a 2-stage stakeholder consultation process to refine the draft checklist to a brief and practical checklist comprising key implementation factors. In the first stage, stakeholders evaluated the relevance and feasibility of items on the draft checklist using a web-based survey. The list of items comprised 12 facilitators presented as statements addressing “elements that positively affect implementation” and 17 barriers presented as statements addressing “concerns toward implementation.” If a strategy was deemed relevant, respondents were asked to rate it using a 4-point Likert scale ranging from “very difficult to implement” to “very easy to implement.” In the second stage, stakeholders were interviewed to elaborate on the most relevant barriers and facilitators shortlisted from the first stage. The interview mostly focused on the relevance and priority of strategies and factors affecting OeMH intervention implementation. In the interview, the stakeholders’ responses to the open survey’s questions were further

XSL•FO
RenderX

explored. The final checklist included strategies ranked as relevant and feasible and the most relevant facilitators and barriers, which were endorsed during either the survey or the interviews.

Results: In total, 26 stakeholders completed the web-based survey (response rate=24.8%) and 4 stakeholders participated in individual interviews. The OeMH intervention implementation checklist comprised 28 items, including 9 (32.1%) strategies, 8 (28.6%) barriers, and 11 (39.3%) facilitators. There was widespread agreement between findings from the survey and interviews, the most outstanding exception being the idea of proposing OeMH interventions as benefits for employees.

Conclusions: Through our 2-stage stakeholder consultation, we developed a brief checklist that provides organizations with a guide for the implementation of OeMH interventions. Future research should empirically validate the effectiveness and usefulness of the checklist.

(J Med Internet Res 2024;26:e48504) doi:10.2196/48504

KEYWORDS

implementation; workplace; mental health; well-being; digital health; mobile health; mHealth; eHealth; e–mental health; stakeholder consultation; intervention; occupational; stakeholders; consultation; barrier; checklist

Introduction

Background

Mental health is a major concern in the workplace and is considered a worldwide public health priority. In Europe, stress, depression, and anxiety are the third-most common mental health problems caused or worsened by work [1]. About half of the European workers report stress and psychosocial risks as a common phenomenon in the workplace, causing approximately 50% of all working days lost [2]. Of the working-age population, 6%-8% of females and 4.0%-5.5% of males suffer from depression and 5%-6% of females and around 3% of males suffer from anxiety [3]. Depression and anxiety are common mental disorders that impact the working status of individuals in terms of unemployment, absenteeism (ie, loss of workdays due to a condition), and presenteeism (ie, reduced ability at work due to a condition). A review study showed that 21% of unemployed people have depression and that depression and unemployment are strongly interrelated [4]. In addition, research shows that past and current anxiety or depressive disorders are associated with higher absenteeism, with the highest effect in those with comorbid anxiety and depression [5]. This has economic consequences: mental ill health was estimated to cost the world economy about US $2.5 trillion a year in 2010, with a projection of US $6 trillion by 2030 [6]. Investment in mental health promotion is needed, and it is estimated that each dollar invested in scaled-up treatment for depression and anxiety will produce a US $4 return on investment in improved health and productivity [6].

A growing body of evidence identifies psychosocial and contextual risk factors related to mental health at the workplace, such as working conditions (eg, job insecurity, work-life balance, job pressure), organizational culture, and type of work [7-9]. For instance, anxiety and barriers at the workplace, including social, attitudinal, and health system–related environmental ones, are key determinants of work performance for workers with depression [10]. Moreover, other factors, such as highly conflicting and excessive work demands, time pressure, low autonomy levels, excessive authority, and lack of social support within the workplace, can increase stress and the risk of developing or worsening mental health conditions [11-13]. These workplace-based factors are critical as good mental health in the workplace is important for the well-being and success of employees, employers, and society.

Addressing mental health issues in the workplace ideally serves a dual purpose: to enhance the well-being of both employees and employers and to strengthen employers' ability to manage stressful situations. In this regard, existing evidence on occupational e–mental health (OeMH) interventions indicates their effectiveness in promoting mental health and alleviating the associated burdens [14-16]. In particular, a meta-analysis by Carolan et al [15] showed that workplace mental health interventions delivered via mobile technologies, computers, or the internet are effective in promoting psychological well-being and reducing anxiety and depression. Although evidence on their effectiveness has been exhaustively evaluated, research addressing implementation strategies, barriers, and facilitators is limited [17-19]. A contributing factor is the lack of guidance for documenting and reporting successful implementation of OeMH interventions, as well as related barriers and facilitators [17].

Although limited, implementation research on OeMH interventions has revealed a myriad of strategies, barriers, and facilitators that influence the successful implementation of these interventions [17]. A brief and user-friendly checklist comprising key implementation factors is potentially instrumental, if used prior to implementation, in ensuring the successful uptake of OeMH interventions. Such a checklist would be particularly useful for those who are interested in implementing OeMH interventions, including those who develop interventions, the employers, and the implementation team.

To develop such a checklist, the knowledge and perspective of stakeholders are crucial. Developers of OeMH interventions typically focus on the technical and content-related aspects of the interventions. The available research on the strategies for OeMH intervention implementation and particularly implementation in the workplace is limited [17]. Furthermore, we could not identify research, based on a real-wold experience, that accounted and systematically analyzed the knowledge and opinion of end users and of other organizational stakeholders on factors that should be considered for OeMH intervention implementation. Stakeholder consultations regarding the uptake of interventions are essential during the phases of development,

XSL•FO
RenderX

implementation, and evaluation of interventions in the workplace [20]. The experience-based knowledge and expertise gained from stakeholders can provide unique insights into the major factors that facilitate or hinder implementation in a workplace setting [21].

Objective

The objective of this study was to develop a brief checklist including the most relevant and feasible strategies, barriers, and facilitators for the implementation of OeMH interventions. We achieved this objective through 2 different stages: first, by reducing a comprehensive set of strategies, barriers, and facilitators, compiled in a recently published systematic review [17], through an online survey that involved a set of relevant stakeholders and, second, by refining the information through individual interviews with a restricted number of stakeholders who were available to provide additional feedback.

Methods

Overview

We carried out a 2-stage stakeholder consultation as part of the EMPOWER (European Platform to Promote Wellbeing and Health in the workplace) project; a more detailed description of the project can be found in the EMPOWER protocol paper [22].

As part of the EMPOWER project activity, we contacted a group of stakeholders through contacts from EMPOWER consortium collaborators, and additional possible contacts were provided by the already contacted stakeholders. A total of 120 stakeholders were included on our list: representatives from advocacy groups, labor organizations, government organizations, nongovernmental organizations (NGOs), the scientific community, and other interested parties, including the European Commission, universities, vocational training institutions, employment advisors, and members of the European Parliament, specifically those participating in the interparliamentarian group on disability and mental health. The 120 stakeholders from our list were predominantly female (n=71, 59.2%) and on average 50 years old. Presumably, they had on average 20 years of experience; 44 (36.7%) worked for NGOs or policy institutions and 30 (25%) were from related research areas. In most cases (n=68, 56.7%), they had a managerial or executive role within their organizations. Additionally, we contacted 25 senior authors of papers included in the scoping review by Bernard et al [17] to complement the practice-level perspective with a more technical perspective of those who had implemented an OeMH intervention and described strategies, barriers, and facilitators in a publication.

Stakeholder Consultation

First Stage: Web-Based Survey

The first consultation stage involved participants completing a web-based survey in English, implemented using Google Forms. The survey included a brief section requesting sociodemographic information, such as country of employment, highest career position, work sector, and organization type and size (based on the Organization for Economic Co-operation and Development [OECD] classification: microenterprises for <10 employees, small enterprises for 10-49 employees, medium-size enterprises for 50-249 employees, and large enterprises for 250 or more employees) [23]. Respondents were asked to rate the relevance and feasibility of items on a comprehensive checklist comprising strategies, barriers, and facilitators relevant to OeMH intervention implementation identified in the scoping review by Bernard et al [17]: specifically, 13 strategies for the implementation of OeMH interventions, 12 associated facilitators, and 17 barriers were included in the survey. A detailed description of the online survey is included in Multimedia Appendix 1.

Questions about the implementation strategies addressed the relevance and feasibility of each strategy operationalized in terms of ease or difficulty of practical implementation. For each strategy, the first response option was “not relevant,” which was meant to exclude the strategy from further consideration. If a strategy was deemed relevant, respondents were asked to rate it using a 4-point Likert scale ranging from “very difficult to implement” to “very easy to implement.” The facilitators and barriers were presented to participants as statements addressing “elements that positively affect implementation” and “concerns toward implementation,” respectively. Participants were asked to indicate the extent to which they agreed with the statements on a 5-point Likert scale ranging from “strongly disagree” to “strongly agree,” with a neutral middle position. The survey also included optional open-ended questions where participants could elaborate on responses and suggest additional implementation strategies, barriers, or facilitators not mentioned in the closed-ended questions. Responses were collected between January 31 and March 31, 2022, and reminder emails to complete the survey were sent every 2 weeks during this time.

Second Stage: Semistructured Individual Interviews

Qualitative semistructured individual interviews were conducted with stakeholders who agreed on participating in a follow-up interview after the web-based survey or in the interview only. An interview guide was developed based on the survey results:

- Stakeholder background, including the field of work and level of experience
- General comments on the survey’s results (showed to the interviewee during the interview)
- Analysis of the relevance and priority of strategies and factors affecting OeMH intervention implementation
- Analysis of responses to the survey’s open questions

The interviews were held online and in English, led by a senior researcher and assisted by 2 researchers, who took notes. A summary of the interview transcript was sent to interviewees for approval after each interview.

Ethical Considerations

Prior to the consultation, participants were asked to read and provide informed consent. The protocol for the consultation was approved by the Ethics Committee of the Fundació Sant Joan de Déu (ref: PS-19-20). Participation was on a voluntary basis: no compensation was offered.

XSL•FO
RenderX

The 120 stakeholders included on our list already provided consent to being contacted again for research purposes at the time they were included on our list. The 25 senior authors of the papers included in Bernard et al's [17] review were contacted using the same approach we used for the stakeholders on the list, the only difference being that they were directly informed that the survey was online.

The information produced from the web-based survey was anonymous (ie, no survey field with names and contact information was mandatory). Therefore, to be contacted also for the interview, participants willing to contribute had to explicitly disclose their identities.

Data Analysis

First Stage: Web-Based Survey

Microsoft Excel was used for calculations. We used frequency-based percentages to report how strategies were perceived by respondents. We ranked the strategies from the easiest to the most difficult to implement based on the response options (very easy, easy, difficult, and very difficult); strategies were defined as easy to implement if ≥50% of the respondents reported them as easy or very easy and difficult to implement if ≥50% of the respondents reported them as difficult or very difficult. Frequency-based percentages were also used to report the level of agreement among respondents regarding the identification of barriers and facilitators. We ranked the barriers and facilitators based on the degree of agreement, from the most agreed ones (ie, those with higher percentages of "strongly agree" and "agree") to the less agreed ones (ie, those with higher percentages of "strongly disagree" and "disagree"). Barriers and facilitators were considered relevant to our analysis if ≥50% of the respondents agreed with the particular item.

Second Stage: Semistructured Individual Interviews

A qualitative description was provided. We used the notes taken during interviews to (1) address whether strategies, barriers, and facilitators not endorsed from the survey were instead endorsed by the interviewees; (2) address whether open-ended survey questions might suggest a new construct, which should be included in the final checklist, or whether it was a different formulation of an item included in the comprehensive list of strategies, barriers, and facilitators; and (3) define additional strategies, barriers, and facilitators. Factors identified during the interviews, and not previously raised, were then retained for inclusion in the OeMH intervention implementation checklist.

EMPOWER OeMH Intervention Implementation Checklist Development

The final EMPOWER OeMH intervention implementation checklist was developed considering the 2-stage consultation process. Strategies were retained if ≥50% of the respondents judged them as "easy" or "very easy" to implement, whereas barriers and facilitators were retained if ≥50% of the respondents agreed with considering an item as either a barrier or a facilitator, respectively. In addition, we retained those items that did not reach the 50% threshold but that were endorsed during the individual interviews, as well as those added by interviewees and not included in the survey (and thus not identified by the literature review). Therefore, the checklist included strategies ranked as relevant and feasible and the most relevant facilitators and barriers. To improve the user-friendliness of the checklist, we rephrased all items so that users can check whether a condition is met or avoided (in the case of barriers).

Results

First Stage of the Stakeholder Consultation: Web-Based Survey

Descriptive Statistics

Of the 145 individuals to whom emails were sent, 105 (72.4%) participants were reached as 40 (27.6%) emails were inactive. Of those 105 participants reached, 26 (24.8%) completed the survey. They reported being from Poland (n=4, 15%), Finland (n=4, 15%), Spain (n=3, 12%), Switzerland (n=2, 8%), and Malta (n=2, 8%). The remaining 42% (n=11) were originally from other European countries (Slovenia, Austria, Belgium, Germany, Serbia, Lithuania, the United Kingdom, the Czech Republic, Croatia, Sweden, and Denmark). Half of the stakeholders worked in health and social work activities, 26.9% (n=7) in education, 15.4% (n=4) in the field of research, 3.8% (n=1) in the information and communication sector, and 3.8% (n=1) in a patient NGO. Most worked mainly in the public sector (n=22, 84.6%). Some (n=7, 26.9%) represented senior management, 26.9% (n=7) middle management, 23.1% (n=6) executive management (eg, chief executive officer [CEO]), 15.4% (n=4) operations (eg, staff), and 7.7% (n=2) a consultant position. At the time of the survey, 65.4% (n=17) worked in a large enterprise, 15.4% (n=4) in a medium-size one, and 19.2% (n=5) in a microenterprise or a small enterprise (ie up to 49 employees).

Implementation Strategies

Almost all 13 implementation strategies were rated as relevant (see Multimedia Appendix 2 for frequency tables). The strategy rated as havin the least relevance was "customizing recruitment activities to enhance reach" (n=3, 12%). Figure 1 shows the frequency-based ranking of the strategies.

XSL•FO
RenderX

Figure 1. Frequency-based ranking of strategies' easiness of implementation.

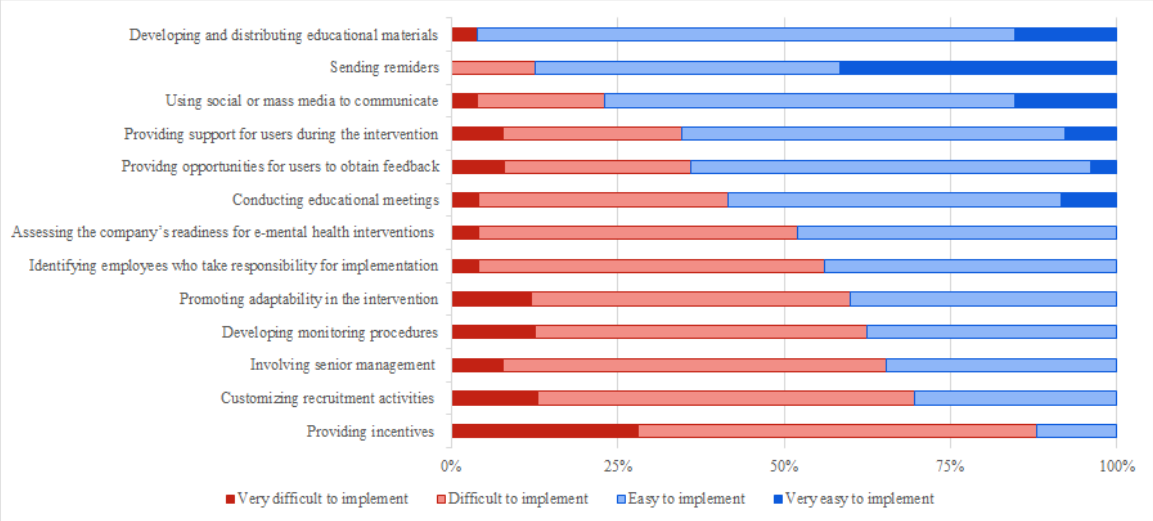

In addition, 6 (46.2%) implementation strategies were identified as "easy to implement" or "very easy to implement" by ≥50% of the respondents: "developing and distributing education material," "sending reminders for completing the intervention," "using social or mass media to increase reach," "providing support for users during the intervention," "providing opportunities for users to obtain feedback on progress," and "conducting educational meetings." Conversely, "involving senior management," "customizing recruitment," and "providing incentives" were perceived as the most difficult to implement.

Barriers and Facilitators

Of the 17 elements presented as potential barriers towards implementation, 8 (47.1%) were considered as such by participants. Specifically, the most negatively perceived barriers were "apps providing generic, irrelevant, or contradictory information," "long and effortful activities negatively impact usage of apps," and "apps are not tailored to an employee's situation and organization's culture." Conversely, 9 (52.9%) barriers were not endorsed by participants: in particular, the less negatively perceived barriers were "mental health symptoms will hinder app usage," "employees refusing mental health support will not use the app," and "users cannot progress at their own pace." Figure 2 shows the frequency-based ranking of the barriers.

All the 12 items presented as potential facilitators toward implementation were perceived as such, with the exception of "employers should plan contingencies for organizational restructuring that could hinder implementation." Among the endorsed facilitators, the 3 (27.3%) rated as most relevant were "employers should guarantee anonymity and confidentiality," "employers should allow employees enough time to use the app," and "the intervention should use reliable data storage systems." Figure 3 shows the frequency-based ranking of the facilitators.

Figure 2. Frequency-based ranking of barriers.

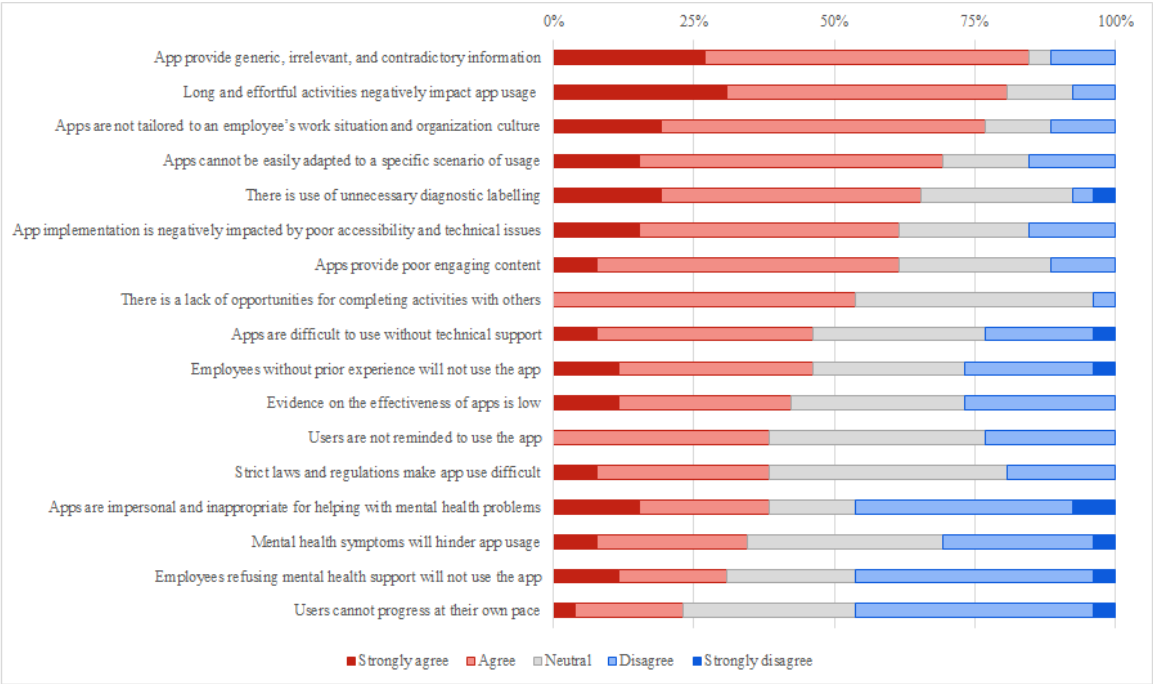

XSL•FO
RenderX

Figure 3. Frequency-based ranking of facilitators.

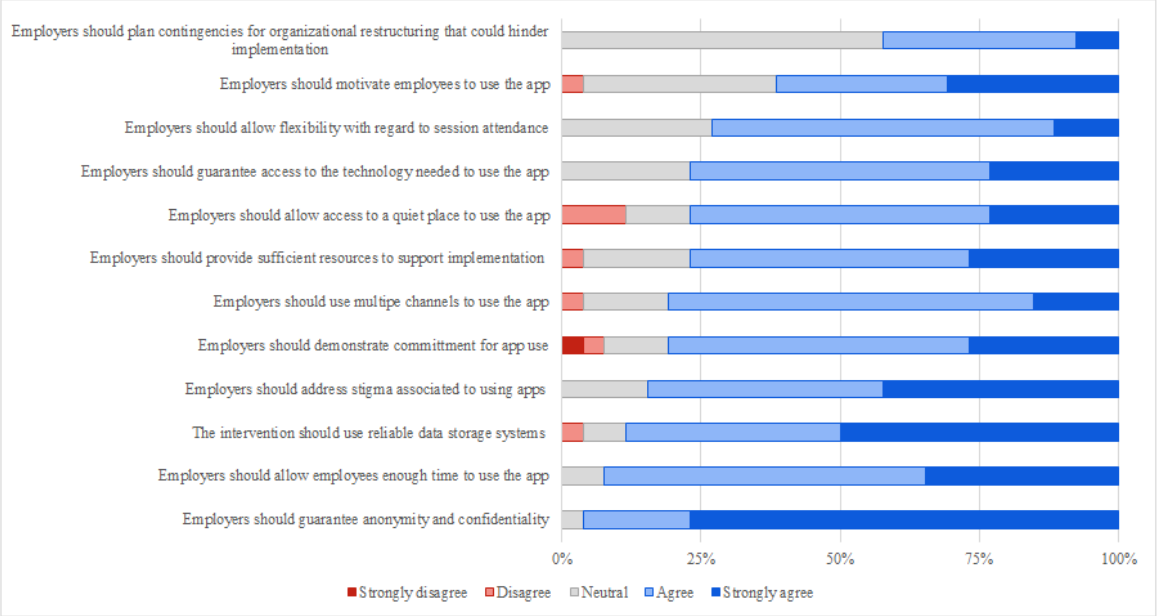

Overall, the responses to the web-based survey showed greater heterogeneity in the evaluations of barriers, whereas a general agreement toward facilitators was observed. Specifically, the key elements identified included the importance of the intervention's usability, the ease of access and use, the presence of engaging content, the exclusion of excessively effortful activities, and the presence of accurate information without diagnostic labeling.

Optional Open-Ended Questions

In addition to the predefined list of strategies, barriers, and facilitators, stakeholders provided some additional suggestions in optional open-ended questions included in the web-based survey, including 7 strategies, 3 barriers, and 3 facilitators (see Textbox 1).

Textbox 1. Additional strategies, facilitators, and barriers suggested in the first consultation stage (web-based survey), provided by the 26 participants who completed the web-survey.

Strategies

- Encouraging dialogue with trade unions and workers
- Providing access to a helpline
- Involving occupational health specialists
- Conducting market analysis to identify similar apps and address posed risks to the successful implementation of the newly proposed app
- Involving people with experience of mental disorders who are already using a similar tool
- Conducting usability tests before launch to verify the app's actual usability and, if required, improving it before its full implementation
- Testing the occupational e–mental health (OeMH) intervention with health managers (eg, from the human resources departments) of large companies

Barriers

- Noninvolvement of end users in cocreation
- Confidentiality issues (eg, will others, such as my boss or colleagues, know that I am using such an app?)
- Unclear data protection policies that poorly convey who, especially within the organization, can view usage information

Facilitators

- Having an open dialogue with managers showing willingness to share information about their past problems with mental health difficulties in the workplace (eg, stress or anxiety)
- Clearly identifying who is in the position of deciding whether to use the app (ie, the employer or single employees?)
- Providing mental health interventions in established mental health and community services

XSL•FO
RenderX

Second Stage of the Stakeholder Consultation: Semistructured Individual Interviews

A total of 4 stakeholders, 2 (50%) from Italy, 1 (25%) from Serbia, and 1 (25%) from Denmark, agreed to participate in an individual interview. These stakeholders came from different work sectors, namely human resources, disability management, and health education and public health, respectively. During the interviews, 5 topics were related to implementation strategies and 4 to barriers and facilitators.

Main Interview Topics on Implementation Strategies

Five main topics related to implementation strategies were addressed during the interviews, and 1 new strategy was suggested by an interviewee.

First, the assessment of an organization's readiness to implement OeMH interventions was considered a crucial preliminary step to be taken regardless of its difficulty.

Second, strategies exclusively focused on the simple "distribution of information" were deemed at a high risk of failure due to being too impersonal and because they cannot be used alone. Examples of these included "developing and distributing education material," "sending reminders for completing the intervention," and "using social or mass media to increase reach." These strategies could be perceived as ways employers provide quick mental health solutions that are inadequately personalized. One interviewee described such kind of information-focused strategies as "a fit for all approach that may not fit anyone."

Third, a similar consideration was made about strategies aimed at involving occupational health specialists and senior management (eg, sharing of experience). Such strategies cannot work if used alone. On the one hand, they might be effective for fostering the promotion of interventions; on the other hand, much of these strategies' success depends on the "personality" of the managers (ie, the degree to which they are available to "get involved" also on a personal level), rather than the overall culture of specific workplaces, regardless of whether such kind of personal exposure is realistic.

Fourth, involving employees (ie, the end users) in the implementation processes is a necessary strategy to avoid them perceiving interventions as being imposed by management. In fact, the level of adherence to an intervention might be seriously hindered if the intervention itself is perceived as mandatory.

Fifth, a new strategy was suggested by an interviewee, which involved offering interventions as an employee benefit provided by the company. This would have the advantage of employees seeing the intervention as an investment for their welfare. Further, it might be particularly valuable for those working in small and medium-size enterprises, where health and social interventions may be not available through the workplace, in contrast to those working in large enterprises.

Main Interview Topics on Barriers and Facilitators

Four main topics focused on barriers and facilitators: none of them constituted a new topic, and they were a further endorsement of topics already included in the survey.

First, interviewees stressed that using unnecessary diagnostic labels within interventions (eg, marketing it as "a way to combat depression") might hinder their acceptance.

Second, interviewees highlighted the risk associated with the stigma surrounding the use of OeMH interventions, a topic that in the web-based survey was included among facilitators (ie, "Employers should address the stigma associated with app use."). The role of employers in combating stigma associated with app use was, of course, emphasized.

Third, interviewees stressed the importance of employers as guarantors of anonymity and confidentiality of any information that is shared when using the intervention.

Fourth, it is important that employers demonstrate their commitment to adopting mental health interventions, while motivating employees to use such interventions and maintaining a dialogue with them.

EMPOWER OeMH Intervention Implementation Checklist

There was widespread agreement between findings from the survey and interviews, except regarding the level of management participation and the idea of proposing the OeMH intervention as a benefit for employees. By combining the results of Bernard et al's [17] scoping review and this stakeholder consultation, we developed an EMPOWER OeMH intervention implementation checklist (Textbox 2). It comprises 28 items: 9 (32.1%) address implementation strategies, 8 (28.6%) refer to implementation barriers, and 11 (39.3%) refer to facilitators for implementation.

XSL•FO
RenderX

Textbox 2. The EMPOWER (European Platform to Promote Wellbeing and Health in the workplace) occupational e–mental health (OeMH) intervention implementation checklist.

Implementation strategies (n=9)

- Promotion messages are sent via mass media channels to reach a large audience.
- Educational material has been developed and distributed (eg, leaflets or emails).
- Educational meetings are planned or have been conducted.
- Organization readiness has been assessed before intervention implementation.
- The intervention is proposed as a benefit for employees.
- End users have been involved throughout the implementation process.
- Support for using the intervention (eg, online support for technical problems with the app or platform) is available.
- Opportunities for users to obtain feedback on progress (eg, percentage of completion) are available.
- Reminders for completing the intervention are envisaged.

Barriers (n=8)

- Avoid generic, irrelevant, contradictory, or inaccurately translated content.
- Avoid long and effortful tasks.
- Avoid interventions that are not tailored to employees' work situation and the organization's culture.
- Avoid intervention components that exclusively target specific use cases and are difficult to adapt to other scenarios.
- Avoid unnecessary diagnostic labeling in the intervention (eg, people with depression or stress).
- Avoid poor accessibility, technical issues, and complicated user interfaces.
- Avoid content that is not engaging and presented in a single media format.
- Avoid having all tasks be single-user tasks (ie, no group activities).

Facilitators (n=11)

- Guarantee anonymity and confidentiality.
- Allow employees enough time to use the intervention.
- Use reliable data storage systems for the intervention.
- Address any stigma associated with using an OeMH intervention.
- Demonstrate the employers' commitment to and interest in employees' participation in the intervention .
- Use multiple communication channels (eg, leaflets, emails, or meetings) to promote the intervention.
- Provide sufficient resources (eg, money or personnel) to support the implementation of the intervention.
- Provide a quiet and private space for employees to use the intervention.
- Ensure access to the technology required (eg, devices or WiFi) to use the intervention.
- Allow flexibility regarding intervention attendance during work hours.
- Motivate employees to use the intervention.

Discussion

Principal Findings and Comparison With Prior Work

We developed an EMPOWER OeMH intervention implementation checklist, a 28-item checklist including 9 strategies, 8 barriers, and 11 facilitators, which guides the implementation of OeMH interventions. The checklist constitutes a step forward in documenting different implementation strategies for OeMH interventions and might fill the gap in the understanding of the reasons some OeMH interventions succeed and others fail.

Two major difficulties were identified in the implementation of OeMH interventions: first, involving the actors of the organizations, particularly those in management roles, in the promotion of the intervention and, second, adapting and customizing the intervention to the individual organization and end-user needs. The 3 strategies evaluated as most relevant and easy to implement (namely, "developing and distributing education material," "sending reminders," and "using social or mass media to increase reach") include simple activities in terms of the time needed to implement them and are expected to easily reach more end users, in line with a previous study [24]. Unsurprisingly, such automated and less time-consuming strategies were considered easier to implement. As shown in a

XSL•FO
RenderX

recent systematic review [25], automated strategies foster a positive user experience and offer a low-cost alternative to human support. On the contrary, our interviewees issued a warning about the use of automated strategies as they could be perceived as impersonal, which may result in a low level of commitment from employers. For these reasons, we support a multiplicity of strategies, not just automated ones.

A low level of commitment is a disadvantage: if employees think that employers are uncommitted, the likelihood of benefiting from the OeMH intervention will be much lower. Our results highlight the importance of organizational commitment in promoting an OeMH intervention, including the need to identify and dedicate resources to the implementation, such as giving employees flexibility, time, and space to participate in the intervention. These findings are in line with previous research showing that the lack of allocated time and adequate private space could hinder engagement with these interventions [25,26]. Organizational commitment is also relevant to addressing stigma, which is known to be a significant barrier to the implementation of mental health services in the community [27]. In contrast, antistigma interventions at the workplace are overwhelmingly effective [28]. Therefore, addressing stigma, including cultural factors that perpetuate stigma attached to mental illness [29], could be critical to foster the implementation of OeMH interventions.

Usability is an essential feature to enhance user engagement, as outlined in different studies on mental health interventions [25,30,31]. This is in line with the recommendations of the "Use of Digital Technology" guidelines [32-34], which advocate for enhancing content clarity through the use of explicit labels and sentences to describe content and instructions in the OeMH interventions.

The interview and web-based survey results were closely aligned on barriers and facilitators, whereas a different perspective on implementation strategies was offered by interviewees. A key point was the need to actively involve both management and end users in the process of implementation in order to avoid perceiving OeMH interventions as mandatory. The involvement of all actors, in particular end users, is a general principle in our checklist, recognized through one of the items. Of course, involvement has a voluntary basis, as no one can be forced to engage in any kind of intervention: from the side of companies, this means providing all possible facilitators that enable use and involvement in OeMH intervention uptake in the workplace. Such a general principle is thus recognized in the last facilitators included in our checklist, which focus on resources and space provision; access to technologies, if needed; time flexibility; and fostering of employees' motivation. The survey findings about the inclusion of senior management, judged as less feasible, and about preferring strategies that are time limited apparently contradict the interview results, which pointed out that an initial investment of time, including time from managers, might be preferred to foster the uptake of OeMH interventions. The latter is in line with previous work suggesting the importance of management participation in the early stages of implementation in order to make strategic decisions that meet the specific needs of organizations [26], thus positively influencing intervention success [35,36].

Assessing organizational readiness before implementing an intervention is a crucial step [37-40], and it was judged among the most relevant strategies by the interviewees. Such a readiness assessment should consider any stigma attached to mental illness within the organization and ensure the anonymity and data protection of end users. Organizations failing to account for these issues carry the risk that OeMH interventions will be negatively perceived by employees (eg, seen as an attempt to deceive employees or to force them to reveal their mental health status or as an untrustworthy system incapable of keeping employee health information private and secure).

Interviewees suggested presenting OeMH interventions as an economic benefit for employees. Highlighting both a personal and an economic value conveys the message that organizations invest in improving employees' health and well-being, exactly in the way in which other services, either health related or otherwise, are presented as part of their remuneration. Such a strategy was deemed to be potentially effective as it demonstrates a strong commitment from companies. Indeed, effectively communicating the intervention's proven benefits for employee end users is an important strategy for uptake [26].

A due consideration is that no single strategy exists to manage the complex process of OeMH intervention implementation. The strategies that have been identified because of the different stages of our activity are summative and highly complementary: the possibility to use a combined set of different approaches, composed of different strategies and facilitators, in the process of OeMH intervention implementation is therefore the preferred approach. The kind of OeMH intervention, the cultural context (including the organization's culture and values as well as the country's culture), and the circumstance in which the OeMH intervention is proposed (eg, response to a particular event or as a more general approach for promoting employees' mental health) may influence implementation.

Implications and Recommendations for Practice and Future Research

The EMPOWER OeMH intervention implementation checklist is a good starting point to determine the factors that enhance the success of OeMH interventions. Implementation research on OeMH interventions can also benefit from our checklist to collect information about the most usfeful strategies and about the most important barriers to and facilitators of implementation. More research is needed on the general validity, usefulness, and acceptability of different implementation strategies, as well as a broadly accepted definition of successful implementation (eg, number of users, usage over time, outcomes).

Limitations

The main limitation of this study is related to the small stakeholder sample size. A total of 26 stakeholders responded to the online survey, corresponding to a response rate of 24.8% and a limited number of them participated in the individual interviews. Such a low response rate is likely due to staff turnover of people who held institutional positions. Although a common problem for web-based surveys, studies with our observed ~25% response rate have demonstrated consistent and accurate results [41,42]. Although the invitation was sent to

XSL•FO
RenderX

several stakeholders from different geographical areas, it has to be remembered that potential respondents were limited to those referred by EMPOWER consortium members and to the senior authors of the papers included in Bernard et al. [17] review. Participants had heterogeneous backgrounds and position levels within the respective organizations, which was an advantage as it likely led to obtaining a wide and heterogeneous set of items. Results may have been different, in terms of the amount of additional implementation strategies, barriers, or facilitators suggested during this study, if there was a larger group of stakeholders. A second limitation is that the checklist has not been piloted and evaluated in an implementation project. Future research is thus needed to validate the checklist.

Conclusion

We developed an EMPOWER OeMH intervention implementation checklist comprising key strategies, barriers, and facilitators relevant to the implementation of OeMH interventions. The checklist could serve as a guide for organizations implementing OeMH interventions and those reporting on these implementations. Future studies that implement this checklist and assess its validity, usefulness, and acceptability in different contexts should be planned.

Acknowledgments

The research leading to these results received funding from the European Union's Union Horizon 2020 Research and Innovation Programme (grant 848180) and the National Health and Medical Research Council of Australia (grant APP1195937).

AR is supported by the Italian Ministry of Health (RRC). BO is supported by the Miguel Servet (CP20/00040) contract, funded by the Instituto de Salud Carlos III (Spain) and cofunded by the European Union (European Regional Development Fund/European Social Funds, "Investing in your Future"). RAL is supported by the Beatriu de Pinós program, Secretaria d'Universitats i Recerca, Departament d'Empresa I Coneixement, Generalitat de Catalunya (grant 2020 BP 00085).

The authors would like to thank the stakeholders who participated to the web-based survey and the individual interviews.

Data Availability

The deidentified quantitative data sets generated and analyzed during the study are available from the corresponding author upon reasonable request as long as the main objective of the data-sharing request is to replicate the analysis and findings, as reported in this paper (without investigator support), after approval of a proposal and with a signed data access agreement. The qualitative data sets generated and analyzed during the study are not publicly available due the risk of reidentification of participants.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Web survey form.
[DOCX File , 38 KB - jmir_v26i1e48504_app1.docx]

Multimedia Appendix 2
Frequency tables for implementation strategies, barriers, and facilitators.
[DOCX File , 33 KB - jmir_v26i1e48504_app2.docx]

References

1. Leclerc C, De Keulenaer F, Belli S. OSH Pulse – occupational safety and health in post-pandemic workplaces: Flash Eurobarometer report. European Agency for Safety and Health at Work. 2022. URL: https://tinyurl.com/4r838w9d [accessed 2024-02-16]
2. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on a comprehensive approach to mental health. COM(2023) 298 final. European Commmission. 2023. URL: https://tinyurl.com/22xd2j57 [accessed 2023-10-31]
3. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://apps.who.int/iris/handle/10665/254610 [accessed 2023-04-18]
4. Amiri S. Unemployment associated with major depression disorder and depressive symptoms: a systematic review and meta-analysis. Int J Occup Saf Ergon 2022 Dec;28(4):2080-2092. [doi: 10.1080/10803548.2021.1954793] [Medline: 34259616]
5. Hendriks SM, Spijker J, Licht CMM, Hardeveld F, de Graaf R, Batelaan NM, et al. Long-term work disability and absenteeism in anxiety and depressive disorders. J Affect Disord 2015 Jun 01;178:121-130. [doi: 10.1016/j.jad.2015.03.004] [Medline: 25805404]

XSL•FO
RenderX

6. The Lancet Global Health. Mental health matters. Lancet Glob Health 2020 Nov;8(11):e1352 [FREE Full text] [doi: 10.1016/S2214-109X(20)30432-0] [Medline: 33069297]
7. Young KP, Kolcz DL, O'Sullivan DM, Ferrand J, Fried J, Robinson K. Health Care Workers' Mental Health and Quality of Life During COVID-19: Results From a Mid-Pandemic, National Survey. Psychiatr Serv 2021 Feb 01;72(2):122-128. [doi: 10.1176/appi.ps.202000424] [Medline: 33267652]
8. Cahill J, Cullen P, Anwer S, Wilson S, Gaynor K. Pilot Work Related Stress (WRS), Effects on Wellbeing and Mental Health, and Coping Methods. The International Journal of Aerospace Psychology 2021 Jan 14;31(2):87-109. [doi: 10.1080/24721840.2020.1858714]
9. Sampson H, Ellis N. Stepping up: the need for Proactive Employer Investment in Safeguarding Seafarers' Mental Health and Wellbeing. Maritime Policy & Management 2020 Dec 30;48(8):1069-1081. [doi: 10.1080/03088839.2020.1867918]
10. Ivandic I, Kamenov K, Rojas D, Cerón G, Nowak D, Sabariego C. Determinants of work performance in workers with depression and anxiety: a cross-sectional study. Int J Environ Res Public Health 2017 Apr 26;14(5):466 [FREE Full text] [doi: 10.3390/ijerph14050466] [Medline: 28445433]
11. Day A, Kelloway EK, Hurrell JJ. Workplace Well-Being: How to Build Psychologically Healthy Workplaces. New York, NY: John Wiley & Sons; 2014.
12. Attridge M. A Global Perspective on Promoting Workplace Mental Health and the Role of Employee Assistance Programs. Am J Health Promot 2019 May;33(4):622-629. [doi: 10.1177/0890117119838101c] [Medline: 31006254]
13. LaMontagne AD, Martin A, Page KM, Reavley NJ, Noblet AJ, Milner AJ, et al. Workplace mental health: developing an integrated intervention approach. BMC Psychiatry 2014 May 9;14(1):131. [doi: 10.1186/1471-244x-14-131]
14. Phillips EA, Gordeev VS, Schreyögg J. Effectiveness of occupational e-mental health interventions: a systematic review and meta-analysis of randomized controlled trials. Scand J Work Environ Health 2019 Nov 01;45(6):560-576 [FREE Full text] [doi: 10.5271/sjweh.3839] [Medline: 31184758]
15. Carolan S, Harris PR, Cavanagh K. Improving Employee Well-Being and Effectiveness: Systematic Review and Meta-Analysis of Web-Based Psychological Interventions Delivered in the Workplace. J Med Internet Res 2017 Jul 26;19(7):e271 [FREE Full text] [doi: 10.2196/jmir.7583] [Medline: 28747293]
16. Tan L, Wang M, Modini M, Joyce S, Mykletun A, Christensen H, et al. Erratum to: preventing the development of depression at work: a systematic review and meta-analysis of universal interventions in the workplace. BMC Med 2014 Nov 13;12(1):74. [doi: 10.1186/s12916-014-0212-4]
17. Bernard RM, Toppo C, Raggi A, de Mul M, de Miquel C, Pugliese MT, et al. Strategies for Implementing Occupational eMental Health Interventions: Scoping Review. J Med Internet Res 2022 Jun 01;24(6):e34479 [FREE Full text] [doi: 10.2196/34479] [Medline: 35648457]
18. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 01;8:139 [FREE Full text] [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
19. Powell BJ, McMillen JC, Proctor EK, Carpenter CR, Griffey RT, Bunger AC, et al. A compilation of strategies for implementing clinical innovations in health and mental health. Med Care Res Rev 2012 Apr;69(2):123-157 [FREE Full text] [doi: 10.1177/1077558711430690] [Medline: 22203646]
20. Schmidt L, Falk T, Siegmund-Schultze M, Spangenberg JH. The Objectives of Stakeholder Involvement in Transdisciplinary Research. A Conceptual Framework for a Reflective and Reflexive Practise. Ecological Economics 2020 Oct;176:106751. [doi: 10.1016/j.ecolecon.2020.106751]
21. Salvador-Carulla L, Lukersmith S, Sullivan W. From the EBM pyramid to the Greek temple: a new conceptual approach to Guidelines as implementation tools in mental health. Epidemiol Psychiatr Sci 2017 Mar 14;26(2):105-114. [doi: 10.1017/s2045796016000767]
22. Olaya B, Van der Feltz-Cornelis CM, Hakkaart-van Roijen L, Merecz-Kot D, Sinokki M, Naumanen P, EMPOWER Consortium, et al. Study protocol of EMPOWER: A cluster randomized trial of a multimodal eHealth intervention for promoting mental health in the workplace following a stepped wedge trial design. Digit Health 2022;8:20552076221131145 [FREE Full text] [doi: 10.1177/20552076221131145] [Medline: 36276189]
23. Enterprises by business size. Organization for Economic Co-operation and Development. URL: https://data.oecd.org/entrepreneur/enterprises-by-business-size.htm [accessed 2023-04-18]
24. Kayrouz R, Dear BF, Karin E, Titov N. Facebook as an effective recruitment strategy for mental health research of hard to reach populations. Internet Interv 2016 May;4:1-10 [FREE Full text] [doi: 10.1016/j.invent.2016.01.001] [Medline: 30135786]
25. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and Facilitators of User Engagement With Digital Mental Health Interventions: Systematic Review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
26. Muuraiskangas S, Harjumaa M, Kaipainen K, Ermes M. Process and Effects Evaluation of a Digital Mental Health Intervention Targeted at Improving Occupational Well-Being: Lessons From an Intervention Study With Failed Adoption. JMIR Ment Health 2016 May 11;3(2):e13 [FREE Full text] [doi: 10.2196/mental.4465] [Medline: 27170553]

XSL•FO
RenderX

27. Patel V, Maj M, Flisher A, De Silva MJ, Koschorke M, Prince M, WPA ZonalMember Society Representatives. Reducing the treatment gap for mental disorders: a WPA survey. World Psychiatry 2010 Oct;9(3):169-176 [FREE Full text] [doi: 10.1002/j.2051-5545.2010.tb00305.x] [Medline: 20975864]
28. Hanisch SE, Twomey CD, Szeto ACH, Birner UW, Nowak D, Sabariego C. The effectiveness of interventions targeting the stigma of mental illness at the workplace: a systematic review. BMC Psychiatry 2016 Jan 06;16:1 [FREE Full text] [doi: 10.1186/s12888-015-0706-4] [Medline: 26739960]
29. Mascayano F, Toso-Salman J, Ho YCS, Dev S, Tapia T, Thornicroft G, et al. Including culture in programs to reduce stigma toward people with mental disorders in low- and middle-income countries. Transcult Psychiatry 2020 Feb;57(1):140-160. [doi: 10.1177/1363461519890964] [Medline: 31856688]
30. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital Mental Health Interventions for Depression, Anxiety, and Enhancement of Psychological Well-Being Among College Students: Systematic Review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
31. Lemon C, Huckvale K, Carswell K, Torous J. A Narrative Review of Methods for Applying User Experience in the Design and Assessment of Mental Health Smartphone Interventions. Int J Technol Assess Health Care 2020 Jan 24;36(1):64-70. [doi: 10.1017/s0266462319003507]
32. Web Content Accessibility Guidelines (WCAG) 2.0. W3C recommendation 11 December 2008. World Wide Web Consortium. 2008. URL: http://www.w3.org/TR/WCAG20/ [accessed 2023-04-18]
33. Treviranus J, McCathieNevile C, Jacobs I, Richards J. Authoring Tool Accessibility Guidelines 1.0. W3C recommendation 3 February 2000. World Wide Web Consortium. 2000. URL: https://www.w3.org/TR/ATAG10/atag10.pdf [accessed 2024-02-19]
34. Jacobs I, Gunderson J, Hansen E. User Agent Accessibility Guidelines 1. W3C recommendation 17 December 2002. World Wide Web Consortium. 2002. URL: http://www.w3.org/TR/WAI-USERAGENT/ [accessed 2023-04-18]
35. Gayed A, LaMontagne AD, Milner A, Deady M, Calvo RA, Christensen H, et al. A New Online Mental Health Training Program for Workplace Managers: Pre-Post Pilot Study Assessing Feasibility, Usability, and Possible Effectiveness. JMIR Ment Health 2018 Jul 03;5(3):e10517 [FREE Full text] [doi: 10.2196/10517] [Medline: 29970359]
36. Saksvik P, Nytrø K, Dahl-Jørgensen C, Mikkelsen A. A process evaluation of individual and organizational occupational stress and health interventions. Work Stress 2002 Mar;16(1):37-57. [doi: 10.1080/02678370110118744]
37. Khan S, Timmings C, Moore JE, Marquez C, Pyka K, Gheihman G, et al. The development of an online decision support tool for organizational readiness for change. Implement Sci 2014 May 10;9:56 [FREE Full text] [doi: 10.1186/1748-5908-9-56] [Medline: 24886072]
38. Weiner BJ, Amick H, Lee SD. Conceptualization and measurement of organizational readiness for change: a review of the literature in health services research and other fields. Med Care Res Rev 2008 Aug;65(4):379-436. [doi: 10.1177/1077558708317802] [Medline: 18511812]
39. Attieh R, Gagnon M, Estabrooks CA, Légaré F, Ouimet M, Roch G, et al. Organizational readiness for knowledge translation in chronic care: a review of theoretical components. Implement Sci 2013 Nov 28;8:138 [FREE Full text] [doi: 10.1186/1748-5908-8-138] [Medline: 24283365]
40. Holt DT, Helfrich CD, Hall CG, Weiner BJ. Are you ready? How health professionals can comprehensively conceptualize readiness for change. J Gen Intern Med 2010 Jan;25 Suppl 1(Suppl 1):50-55 [FREE Full text] [doi: 10.1007/s11606-009-1112-8] [Medline: 20077152]
41. Morton SM, Bandara DK, Robinson EM, Carr PEA. In the 21st Century, what is an acceptable response rate? Aust N Z J Public Health 2012 Apr;36(2):106-108 [FREE Full text] [doi: 10.1111/j.1753-6405.2012.00854.x] [Medline: 22487341]
42. Walsh EI, Chung Y, Cherbuin N, Salvador-Carulla L. Experts' perceptions on the use of visual analytics for complex mental healthcare planning: an exploratory study. BMC Med Res Methodol 2020 May 07;20(1):110 [FREE Full text] [doi: 10.1186/s12874-020-00986-0] [Medline: 32380946]

Abbreviations

EMPOWER: European Platform to Promote Wellbeing and Health in the workplace
OeMH: occupational e–mental health
NGO: nongovernmental organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.04.23; peer-reviewed by M Gasmi , E Phillips, C Lemon; comments to author 18.10.23; revised version received 29.11.23; accepted 21.12.23; published 15.03.24.

Please cite as:
Raggi A, Bernard RM, Toppo C, Sabariego C, Salvador Carulla L, Lukersmith S, Hakkaart-van Roijen L, Merecz-Kot D, Olaya B, Antunes Lima R, Gutiérrez-Marín D, Vorstenbosch E, Curatoli C, Cacciatore M
The EMPOWER Occupational e–Mental Health Intervention Implementation Checklist to Foster e–Mental Health Interventions in the Workplace: Development Study
J Med Internet Res 2024;26:e48504
URL: https://www.jmir.org/2024/1/e48504
doi:10.2196/48504
PMID:38488846

©Alberto Raggi, Renaldo M Bernard, Claudia Toppo, Carla Sabariego, Luis Salvador Carulla, Sue Lukersmith, Leona Hakkaart-van Roijen, Dorota Merecz-Kot, Beatriz Olaya, Rodrigo Antunes Lima, Desirée Gutiérrez-Marín, Ellen Vorstenbosch, Chiara Curatoli, Martina Cacciatore. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of COVID-19 and Exposure to Violent Media Content on Cyber Violence Victimization Among Adolescents in South Korea: National Population-Based Study

Eugene Lee[1], MA; Peter J Schulz[2], PhD; Hye Eun Lee[3], PhD

[1]Annenberg School for Communication and Journalism, University of Southern California, Los Angeles, CA, United States
[2]Department of Communication, Culture and Society, University of Lugano, Lugano, Switzerland
[3]Department of Communication & Media, Ewha Womans University, Seoul, Republic of Korea

Corresponding Author:
Hye Eun Lee, PhD
Department of Communication & Media
Ewha Womans University
#403 Ewha-POSCO, 52 Ewhayeodae-gil, Seodaemun-gu
Seoul, 03760
Republic of Korea
Phone: 82 3277 6937
Fax: 82 3277 4010
Email: hyeeunlee77@ewha.ac.kr

Abstract

Background: Because of the COVID-19 pandemic and consequent stay-at-home mandates, adolescents faced isolation and a decline in mental health. With increased online activity during this period, concerns arose regarding exposure to violent media content and cyber victimization among adolescents. Yet, the precise influence of pandemic-related measures on experiences of cyber violence remains unclear. Hence, it is pertinent to investigate whether the pandemic altered the dynamics of cyber violence victimization for individuals.

Objective: This study aims to investigate the effects of COVID-19 and exposure to violent media content on cyber violence victimization among adolescents in South Korea.

Methods: We used national survey data from 2019 (n=4779) and 2020 (n=4958) to investigate the potential impact of COVID-19 on the prevalence of cyber violence among young adolescents. The data encompassed responses from elementary fourth-grade students to senior high school students, probing their exposure to violent media content, average internet use, as well as experiences of victimization and perpetration.

Results: The analysis revealed a noteworthy decline in cyber victimization during 2020 compared with 2019 (B=–0.12, t=–3.45, P<.001). Furthermore, being a perpetrator significantly contributed to cyber victimization (B=0.57, t=48.36, P<.001). Additionally, younger adolescents (β=–.06, t=–6.09, P<.001), those spending more time online (β=.18, t=13.83, P<.001), and those exposed to violent media (β=.14, t=13.89, P<.001) were found to be more susceptible to victimization.

Conclusions: Despite the widespread belief that cyber violence among adolescents surged during COVID-19 due to increased online activity, the study findings counter this assumption. Surprisingly, COVID-19 did not exacerbate cyber victimization; rather, it decreased it. Given the strong correlation between cyber victimization and offline victimization, our attention should be directed toward implementing real-life interventions aimed at curbing violence originating from in-person violence at school.

(J Med Internet Res 2024;26:e45563) doi:10.2196/45563

KEYWORDS

cyber violence; adolescents; victimization; perpetration; COVID-19

XSL•FO
RenderX

Introduction

Background

In a desperate search for meaning in the virus, many academics compared the public health dangers of the COVID-19 pandemic to the "Black Death" of the second half of the 14th century [1,2] and the influenza affliction at the end of World War I [3]. The "War," or the Great War, as it was then called, eclipsed the number of deaths caused by the flu so that the flu, despite a tremendous number of fatalities, is not remembered as widely as the Black Death. Today's standard living conditions are incomparable to the dire environment of medieval times. Still, violence persists in different manifestations and can severely threaten public health, thereby changing people's sensitivity toward the pain of others. A case in point is cyber violence. Individuals inclined to harass their peers might find it easier to do so when the bully's threshold for action is lowered. Adolescents are especially exposed to the risks of cyber violence as a result of increased online communication and learning during COVID-19.

Previous studies have shown that exposure to media violence increases the risk of cyber aggression [4]. "Cyber violence" refers to using images or videos or both in cyberspace to verbally harm or harass others [5]. Similarly, "cyber aggression" refers to behaviors that are intentionally aimed at harming others by communicating through electronic devices such as computers and mobile phones [6]. "Cyberbullying" is a form of aggression that is carried out by an individual or by a group through electronic communication consistently, over time, and with the intention of targeting a particular individual exposed to violence [7]. It is crucial to treat adolescent exposure to cyber violence seriously, especially given the heightened duration of online activities and the widespread adoption of remote learning during and in the wake of the COVID-19 pandemic. As adolescents spend more time online, they are likely to consume online content depicting violent behavior, which risks exposing them to cyber violence [8]. In this study, we examine the impact of COVID-19 on cyber violence victimization by examining survey data conducted between 2019 and 2020 among South Korean adolescents. We also examine the impact that exposure to violent media content—such as sensational content, derogatory comments attacking other people, portrayals of illegal behavior, and misinformation—has on cyber victimization.

Marciano et al [9] conducted a meta-analysis and found that adolescents around the world spent more time on digital media during COVID-19 and were more likely to experience negative effects on their mental health. Although the study did not include samples from South Korea, it illustrated the impact of COVID-19 on mental health using samples from other Asian countries (eg, China, Japan, Indonesia, and Bangladesh) and from European and Northern American countries. As exposure to media violence increases the risk of cyber aggression [4], some may argue that COVID-19 could have accelerated cyber violence. Yet, a report from Canada [10] revealed that 17% of young people actually reported a decrease in cyber victimization during the pandemic in comparison to the previous period. Similarly, American youth reported an increase in negative mental health experiences while reporting a decrease in cyber victimization during the COVID-19 pandemic [11]. Hence, due to the absence of face-to-face interactions to instigate or sustain bullying in online environments, students might have experienced a reduction in cyber victimization during the pandemic [11].

Adolescents experiencing cyber victimization are widespread across many countries. For example, 69.9% of US adolescents reported some level of being victimized by cyber harassment, with 60.8% stating that they experienced cyber harassment repeatedly [12]. According to the 2021 Cyber Violence National Survey conducted in South Korea, 30% of adolescents reported experiencing some level of cyber violence [13]. When the prolonged negative effects of cyber violence and isolation during COVID-19 converge, it can pose a significant threat to the mental health of adolescents.

Previous longitudinal studies have illustrated the prolonged negative effects of cyber violence on individuals. For instance, individuals exposed to cyberbullying continuously experience higher levels of anxiety and depression, as well as more frequent suicidal thoughts [14]. Camerini et al [14] conducted a systematic review of countries in Europe, North America, Oceania, and Asia—including South Korea—and found a prolonged negative effect of cyber violence on individuals. Hinduja and Patchin [15] demonstrated that adolescents in the United States who experienced traditional bullying or cyberbullying had more suicidal thoughts and were more likely to attempt suicide than those who had not experienced such levels of aggression.

Because of stringent government measures in response to COVID-19 in South Korea, all public facilities—including places frequented by adolescents (eg, private academies, private chat rooms, libraries, and study cafés)—were closed, thereby denying adolescents a physical space to connect with their peers during the pandemic. Byeon [16] identified experiences such as depression, subjective stress recognition, loneliness, and changes in economic status due to COVID-19 as major factors influencing suicidal ideation among Korean adolescents. Being isolated during COVID-19 has made South Korean adolescents feel disconnected from each other and from the broader community; there has also been an increase in anxiety, depression, and other indicators of psychological stress, which has increased not only the frequency of suicidal ideation but also the number of suicide attempts [17].

As adolescence marks a period of life when individuals freely consume more media due to the increased accessibility of content, adolescents constitute a significant population for studying the effects of cyber violence [18]. Engaging in activities such as surfing the internet, watching videos, social networking, and interacting with others online and offline may expose adolescents to violent content and cyber violence; this is especially likely when parental control over media consumption is limited [18,19]. More surveillance challenges arise because smartphones, unlike television or personal computers, are personalized devices that one carries throughout the day, which renders monitoring nearly impossible [20]. Collectively, these conditions and behavioral trends may explain

XSL•FO
RenderX

why many adolescents have reported that they have experienced cyber violence or cyber aggression.

South Korean adolescents provide a particularly unique opportunity for researchers to study cyber violence among young adolescents. Unlike adolescents in many other parts of the world, Korean adolescents tend to possess personal mobile devices from a somewhat early age, with most starting to use smartphones during early adolescence [21]. Previous research has shown that 87.7% of elementary schoolers between fourth and sixth grades carry their mobile phones [22]. Smartphone ownership increases with age, with 92.6% of sixth graders reporting that they possess a personal smartphone. As media technology evolves, the age of owning one's first smartphone continually decreases. Thus, the case study of Korea provides a great opportunity to prepare for a future where adolescents have more freedom and exposure to various media contents in earlier stages of their lives. In this study, we use national surveys conducted in South Korea to examine the relationship between media consumption patterns, exposure to violent content, and the experience of being a perpetrator or exposed to cyber violence.

COVID-19 and Its Effect on Adolescents in South Korea/Asian Countries and Victimization of Cyber Violence

Previous literature centers on the mental health of Korean/Asian adolescents during COVID-19, as well as on the social impact of COVID-19 on adolescents. Mental health status and social impact, such as the level of exposure to the media, are essential factors that reflect the vulnerability of adolescents to cyber violence during COVID-19. Concerning mental health and COVID-19, studies have discussed suicidal tendencies [16,23-25], anxiety and depression [26], posttraumatic stress disorder (PTSD) [27], and the general mental health of adolescents [28,29].

Byeon [16] found that changes in economic status and deteriorated mental health during COVID-19 influenced the suicidal ideation of Korean adolescents. Similar to previous findings about isolation and suicide attempts [24], isolation during COVID-19 resulted in feelings of disconnection and increased anxiety, depression, and psychological stress among Korean adolescents, leading to increased suicidal ideation and suicide attempts [17]. Suicidal tendencies were particularly exacerbated in multicultural Korean adolescents when they were exposed to sexual intercourse, depressive moods, and unhappiness during COVID-19 [25]. Some studies have even shown that COVID-19 is associated with PTSD [27,28]. A short-term longitudinal study found that 15%-28% of the adolescents in the study showed PTSD symptoms due to COVID-19 [27].

Additionally, Field [26] reviewed several studies looking into the psychological effects of COVID-19 on adolescent mental health and found that overall, there is a significant prevalence of psychological problems—such as anxiety and depression—among adolescents in Asia and around the world. Many other psychological problems, such as stress, insomnia, social anxiety, PTSD, obsessive-compulsive disorder, and suicidality, were comorbid. These conditions were also found to be more prevalent among women than among men, and more prevalent among high school and university students compared with children. Lundström [29] reported similar findings through an analysis of 108 studies, which indicated that depression and anxiety, worrying, sleeping problems, posttraumatic stress, obsessive-compulsive disorder, deliberate self-harm, and suicidal ideation were all prevalent mental health outcomes in adolescents due to COVID-19.

The social impact category contains articles focusing on behaviors or results related to adolescents' social life and work life [29-31]. Some studies have illustrated that COVID-19 exacerbated adolescents' addiction to media, specifically internet addiction and game addiction. For instance, a Taiwanese study looking at 1244 junior high school students found that the prevalence rate of internet addiction during COVID-19 was as high as 24.4%. Similarly, internet game disorder became a major issue among adolescents, with a prevalence rate of 20% in South Korea during COVID-19 [30].

Some studies identified lockdowns as an important factor that led to internet or video game addiction [30]. In addition, higher levels of depression and lower levels of self-esteem, as well as subjective well-being, actual social support, and lower family function, increased the likelihood of internet addiction [31]. Similarly, previous studies have identified the relationship between internet gaming disorder and conditions of psychological well-being, such as depression, anxiety, and stress symptoms [30]. This shows that underserved adolescents with lower well-being can be more vulnerable to cyber violence during COVID-19 [32]. Most importantly, a previous study argued that increased social media use leads to cyberbullying because adolescents are more likely to be exposed to cyber violence when they spend more time online [29].

From this existing literature, we have found that COVID-19 has compromised the mental health of adolescents while increasing the time they spend consuming media, which puts them in a vulnerable situation that easily exposes them to cyber violence.

The Relationship Between Being a Cyber Violence Perpetrator and Victimization

Cyberbullying is one of the most significant categories of cyber violence. Previous research has shown that 76% of Korean adolescents have experienced cyberbullying, and as a result, severe secondary consequences to their well-being, such as depression and possibly suicide. Cyberbullying is noted as an important aspect of examining adolescents and media. Previous studies have linked internet use [33,34], previous offline bullying, and experience of exposure to violence/bullying in school [35,36] with cyberbullying behavior. To be specific, the more adolescents use the internet, the more likely they are to be associated with cyberbullying behavior. Additionally, Hinduja and Patchin [35] argued that people who get mistreated in the real world are likely to be mistreated in cyberspace. Similarly, Ybarra and Mitchell [36] indicated that about one-half of those who are engaged in cyberbullying activities or experienced cyberbullying also experience offline bullying [36]. The literature shows no clear line between perpetrators and

XSL•FO
RenderX

individuals experiencing cyber violence; being a perpetrator and victimization are correlated [37]. Further, we can say that there is no clear distinction between offline bullying and cyberbullying. In other words, offline bullying behavior can often spill over to cyberspace.

Adolescents who are exposed to cyber violence might be more susceptible to committing cyber violence toward others. As there is a correlation between perpetrators and individuals who experienced cyber violence, we include the experience of being a perpetrator in the model to explain the variance of victimization.

The quality of social interaction online seems to be an essential aspect related to well-being online. A study by Kim [38] found that for Korean adolescents, "time spent online has a deleterious effect on adolescent psychological well-being: for every one-unit increase in online social networking, the odds of suicidality rise by more than a third." By contrast, findings from the study by Bae [39] showed that increased use of smartphones for communication results in increased social capital, which, in turn, increases adolescents' well-being. These studies show that the types of content and kinds of people adolescents are exposed to online would have a significant impact on behaviors regarding cyber violence. If adolescents are exposed to violent content or are vulnerable interacting with others online, they are likely to get victimized. By contrast, if adolescents are having positive interactions with people online or are not exposed to violent content, it may not necessarily lead to victimization despite COVID-19 increasing their time spent online.

The Relationship Between Violent Media Exposure, Gender, and Cyber Violence Victimization

The association between exposure to violent media content and aggressive behavior has been studied in various contexts [37,40-42]. The common premise is that people are more likely to display aggressive behaviors when exposed to violent media content. In the context of cyber violence, people exposed to violent content are more likely to be involved in cyber violence.

What about the relationship between violent media exposure and victimization? Previous studies have shown evidence of a positive correlation between being a perpetrator and being exposed to cyber violence [37,43]. In other words, individuals who commit perpetration online are also likely to experience violence, and vice versa. Therefore, we can expect a positive relationship between exposure to violent media content and victimization. However, we can also expect that people exposed to violent media content are at more risk due to the type of online community, people, and space to which they are exposed.

Gender is another crucial predictor of victimization in cyber violence. Choi and Lee [44] have examined 2 types of cyber violence: cyber harassment and cyber impersonation, and found that females are more likely to experience higher levels of victimization in cyber violence. In addition, Backe et al [45] argued that cyber violence is associated with adverse psychological and social outcomes both online and offline, disproportionately affecting women, girls, and sexual and gender minorities. By contrast, some studies have indicated that cyber offending is more associated with boys than with girls [46].

Based on our literature review, we propose the following hypotheses:

- H1: COVID-19 will affect the level of cyber victimization of adolescents.
- H2: The relationship between exposure to violent media and victimization will vary across age, gender, COVID-19, and internet usage.
- H3: Exposure to violent media content will have a positive relationship with victimization among adolescents.
- H4: Age will have a negative relationship with victimization among adolescents.

Methods

Overview

This study focused on exposure to violent content and cyber violence victimization before and after COVID-19. The National Survey on Cyber Violence Among Adolescents was initiated in 2013 by the Korea Communications Commission (KCC), a presidential consensus-based administrative organization. The KCC is responsible for regulating broadcasting and communication services, protecting users, and addressing related issues to maintain the independence of broadcasting services in Korea. The KCC decided to conduct the national survey in response to the serious and increasing concerns about cyberbullying as a social problem. The goal is to identify the current issues and challenges related to cyberbullying in Korea. Data from the annual survey from 2013 to 2022 will be made available upon valid request. We only selected data from 2019 and 2020 because data from other years have different items to measure the variables or incongruent response formats. The KCC delegated the 2019 and 2020 national surveys of South Korean adolescents to the National Information Society Agency, a statutory agency established in 1987 under the Ministry of Science and Information and Communications Technology and the Ministry of the Interior and Security of Korea.

Data Collection

The data were collected among students ranging from elementary school fourth graders to seniors in high school. The academic system in South Korea has 6 years of elementary school, 3 years of middle school, and 3 years of high school. Students who start the fourth grade are usually aged 10, and third-year high school students are usually aged 18.

The data were collected using stratified random sampling. Whole classes were selected based on a regional and grade list from the School Statistics Database at the Korea Educational Development Institute. Teachers of these selected classes supervised students as they completed the survey. While students were allowed to skip individual questions, they were not allowed to skip the survey altogether. Because of this approach, the participation rate was 100%, but there are cases of missing data for certain questions.

Surveys were conducted across schools in both urban and rural South Korea, yielding 9737 responses in total (4779 participants in 2019 and 4958 participants in 2020). Participants completed the survey in schools from October 1, 2019, to November 23, 2019. However, because of the circumstances caused by

XSL•FO
RenderX

COVID-19, the 2020 survey was conducted online between October 6 and November 13.

Participants comprised 5157/9737 men (52.96%) and 4580/9737 women (47.04%), and each grade had more than 1000 students participating in the study (1119, 1128, 1068, 1085, 1037, 1083, 1047, 1083, and 1087 participants from each grade starting from elementary fourth grade to third-year high school students).

Ethics Approval

This study was granted ethical approval by the Institutional Research Board of Ewha Womans University (approval number ewha-202204-0022-02).

Measures

A questionnaire was constructed in Korean.

Exposure to Violent Media

Participants were asked to indicate instances in which they had been exposed to each of the 5 categories of cyber violence through various online content formats (eg, writing, cartoons, pictures, or videos) including pop-up windows. The 5 categories of cyber violence were (1) violent content, (2) sensational material involving parts of the male or female body or sexual interactions, (3) derogatory comments about well-known figures such as celebrities and politicians, (4) content related to illegal behaviors (eg, fraud or stealing), and (5) false information presented as if it were true. A 5-point response scale was used (0=none, 1=once or twice a year, 2=once or twice in a few months, 3=once or twice a month, and 4=once or twice a week). The sum of the 5 responses was used for the final analysis.

Victimization

Six items were used to measure whether or not participants experienced cyber violence. The instruction question was as follows: “In the most recent year, have you experienced cyber violence as listed below while you were using the internet/smartphone? If so, how often have you experienced them?” The 6 types of cyber violence outlined were “Someone made derogatory comments about me or hurt my feelings,” “Someone made a false rumor or spread an exaggerated story about me,” “Someone kept sending emails/messages or posting pictures or messages on my blog and social networking site knowing that I do not like,” “Someone sent me sensational pictures or videos without my consent when knowing I do not like,” “Someone posted my personal information (eg, name, address, the name of the school, photos) without my consent,” and “Someone stopped me from exiting the chatroom and made fun of me or made derogatory comments about me and did not let me participate in the conversation.” A 6-point response scale was used (0=none, 1=once or twice a year, 2=once or twice in a few months, 3=once or twice a month, 4=once or twice a week, and 5=almost every day). The sum of the 6-item responses was used to measure victimization.

Perpetration

Six items were used to measure whether or not participants had been perpetrators of cyber violence. The instruction question was as follows: “In the most recent year, have you done anything that is listed below using the internet/smartphone? If so, how often have you done them?” Six items were asked: “I made derogatory comments or hurt other people’s feelings,” “I made a false rumor or spread an exaggerated story about others,” “I kept sending emails/messages or posting pictures or messages on other people’s blogs and social networking site knowing that other people don’t like it,” “I sent sensational pictures or videos without consent when knowing other people don’t like it,” “I spread the personal information of others (eg, name, address, the name of the school, photos) without a consent,” and “I stopped someone from exiting the chatroom and made fun of them or made derogatory comments and did not let them participate in the conversation.” A 6-point response scale was used (0=none, 1=once or twice a year, 2=once or twice in a few months, 3=once or twice a month, 4=once or twice a week, and 5=almost every day). The sum of the 6-item responses was used to measure perpetration.

Average Internet Use

Average internet use per day was measured using a 6-point scale (0=less than 1 hour, 1=1-2 hour(s), 2=2-3 hours, 3=3-4 hours, 5=4-5 hours, and 6=5 or more hours).

Data Analysis Procedure

For data analysis, we conducted first-, second-, and third-order regression analyses to predict factors that are related to the victimization of cyber violence. In the first-order regression analysis, we included age, average internet usage, COVID-19, and gender as predictors of cyber violence victimization. For the second-order regression model, we added exposure to violent media content and the experience of being a perpetrator to the model. As it has been reported that victimization and perpetration are highly correlated [36], we included perpetration in our second model to account for its impact on victimization. For the third-order regression, we included interaction terms with COVID-19 to account for the impact of COVID-19 on cyber violence victimization. To reduce multicollinearity, all the continuous predictors were mean-centered [47].

Results

The Incidences of Exposure to Cyber Violence, Victimization, and Perpetration

In terms of exposure to the 5 categories of cyber violence, 1859/9737 (19.09%) participants reported exposure to at least one category. The most common category in each year was violent content, with 2356/4779 (49.30%) of participants in 2019 and 2573/4958 (51.90%) in 2020 reporting that they had seen violent content at least once or twice per year. Additionally, 329/4779 (6.88%) of participants in 2019 and 436/4958 (8.79%) of participants in 2020 indicated that they had been exposed to violent content once or twice per week. In total, 3192/4779 (66.79%) of 2019 participants and 3326/4958 (67.08%) of 2020 participants reported that they had been exposed to violent media content at least once or twice per year.

Regarding victimization, 903/4779 (18.90%) of students in 2019 and 956/4958 (19.28%) of students in 2020 reported that they had experienced cyber violence at least once or twice per year. Furthermore, 855/4779 (17.89%) of 2019 students and 461/4958

XSL•FO
RenderX

(9.30%) of 2020 students reported that they had been perpetrators of cyber violence at least once or twice per year.

Hierarchical Moderated Regression Models

The hypotheses were examined with a hierarchical moderated regression. Tables 1-3 present the regression results.

Table 1. First-order regression analysis on victimization (N=9737).[a]

Analysis	B	SE	β	*t*	Significance
Constant	6.89	0.04	N/A[b]	194.08	<.001
Age	–0.04	0.01	–.06	–6.09	<.001
Average internet use per day	0.16	0.01	.18	13.83	<.001
COVID-19 (2019=0; 2020=1)	–0.30	0.04	–.10	–0.77	<.001
Gender (male=0; female=1)	–0.12	0.03	–.04	–3.90	<.001

[a]$F_{4,9732}$=53.75, adjusted R^2=0.02, P<.001.

[b]N/A: not applicable.

Table 2. Second-order regression analysis on victimization (N=9737).[a]

Analysis	B	SE	β	*t*	Significance
Constant	6.65	0.31	N/A[b]	211.29	<.001
Age	–0.06	0.01	–.10	–9.65	<.001
Average internet use per day	0.07	0.01	.08	6.80	<.001
COVID-19 (2019=0; 2020=1)	–0.12	0.04	–.04	–3.45	<.001
Gender (male=0; female=1)	–0.01	0.03	–.00	–0.32	.75
Exposure to violent content	0.05	0.00	.14	13.89	<.001
Experience of being a perpetrator	0.57	0.01	.44	48.36	<.001

[a]$F_{2,9730(change)}$=1448.08, R^2_{change}=.22, P<.001.

[b]N/A: not applicable.

Table 3. Third-order regression analysis on victimization (N=9737).[a]

Analysis	B	SE	β	*t*	Significance
Constant	6.66	0.04	N/A[b]	163.36	<.001
Age	–0.04	0.01	–.07	–4.93	<.001
Average internet use per day	0.05	0.02	.06	3.68	<.001
COVID-19 (2019=0; 2020=1)	–0.07	0.05	–.02	–1.42	.16
Gender (male=0; female=1)	0.03	0.04	.01	0.87	.38
Exposure to violent content	0.04	0.01	.12	8.31	<.001
Experience of being a perpetrator	0.46	0.01	.36	32.19	<.001
COVID-19 × age	–0.03	0.01	–.04	–2.675	.01
COVID-19 × average internet use	0.03	0.02	.02	1.41	.16
COVID-19 × gender	–0.07	0.06	–.02	–1.32	.19
COVID-19 × exposure to violent content	0.01	0.01	.03	1.74	.08
COVID-19 × experience of being a perpetrator	0.23	0.03	.10	7.76	<.001
COVID-19 × exposure to violent content × experience of being a perpetrator	0.02	0.00	.06	5.54	<.001

[a]$F_{6,9724(change)}$=40.32, R^2_{change}=0.02, P<.001.

[b]N/A: not applicable.

XSL•FO
RenderX

H2 predicted that the relationship between exposure to violent media and victimization would vary with age, gender, COVID-19, and average internet use. Gender was coded as male=0 and female=1, with male as the reference. COVID-19 was coded as 2019=0 and 2020=1, with 2019 as the reference. To test interaction effects, the dependent variable was regressed onto product terms of independent variables. Simple slope analyses were conducted for significant interaction effects if needed.

The first-order predictors explained a significant amount of variance in victimization: $F_{4,9732}$=53.75, adjusted R^2=0.02, P<.001. Age (β=–.06, t=–6.09, P<.001), average internet use (β=.18, t=13.83, P<.001), COVID-19 (β=–.10, t=–7.70, P<.001), and gender (β=–.04, t=–3.90, P<.001) were significant. Those who were older experienced less victimization, and those who spent more time online experienced more victimization. These results were consistent with H2, which predicted that the relationship between exposure to violent media and victimization will vary across age, average internet use, COVID-19, and gender. The results are also consistent with H4, which suggested a negative relationship between age and victimization. Males (Y_{exp}=6.89) reported significantly more victimization than females (Y_{exp}=6.77). Victimization was significantly higher before COVID-19 (Y_{exp}=6.89) than after COVID-19 (Y_{exp}=6.59). H1 predicted that COVID-19 would affect victimization. The results showed that COVID-19 changed its pattern, supporting H1.

The second-order predictor explained a significant amount of variance in the dependent variable: $F_{2,9730(change)}$=1448.08, R^2_{change}=0.22, P<.001. Exposure to violent media (β=.14, t=13.89, P<.001) as a newly added predictor was significant. Those who are exposed to violent media experience more victimization. The experience of being a perpetrator was included as a control variable, which significantly predicted victimization (β=.44, t=48.36, P<.001). In the second-order regression analysis, the impact of gender did not hold as a significant variable explaining victimization (β=.00, t=–0.32, P=.75). However, age (β=–.10, t=–9.65, P<.001), average internet use (β=.08, t=6.8, P<.001), and COVID-19 (β=–.04, t=–3.45, P<.001) held consistent patterns as the first-order predictor model, significantly explaining victimization. Therefore, H1, H3, and H4 were supported in the second-order regression model.

The third-order predictors explained a significant amount of variance in victimization: $F_{6,9727(change)}$=40.32, R^2_{change}=0.02, P<.001. Results showed that COVID-19 did not moderate the relationship between exposure to violent media and victimization (unstandardized coefficient B=0.01, SE=0.00, t=1.74, P=.08). Results also showed that COVID-19 did not moderate the relationship between average internet use and victimization (unstandardized coefficient B=0.03, SE=0.02, t=1.41, P=.16). COVID-19 did not moderate the relationship between gender and victimization (unstandardized coefficient B=–0.07, SE=0.06, t=–1.32, P=.19). However, COVID-19 did moderate the relationship between age and victimization (unstandardized coefficient B=–0.03, SE=0.01, t=–2.67, P<.01). In 2019, age was negatively related to victimization (simple slope b=–0.04), but it was more negatively related to victimization in 2020 (simple slope b=–0.07). COVID-19 also moderated the relationship between perpetration and victimization (unstandardized coefficient B=0.23, SE=0.03, t=7.76, P<.001). In 2019, perpetration was positively related to victimization (simple slope b=0.46), but it was more positively related to victimization in 2020 (simple slope b=0.69).

Compared with the second-order regression model, the impact of COVID-19 on victimization lost its significance (β=–.02, t=–1.42, P=.16), while age (β=–.07, t=–4.93, P<.001), average internet use (β=.06, t=3.68, P<.001), exposure to violent content (β=.12, t=8.31, P<.001), and the experience of being a perpetrator (β=.36, t=32.19, P<.001) held their patterns regarding the impact on victimization. H1 was only supported in the first- and second-order regression models, whereas H3 and H4 were supported across all 3 models.

Table 4 is a summary table indicating the variables that were significant in the first-, second-, or third-order regression analyses. It illustrates the patterns of significance among different variables in the models.

Table 4. Summary table.

Significant variables	First order	Second order	Third order
Age	Significant	Significant	Significant
Average internet use	Significant	Significant	Significant
COVID-19	Significant	Significant	Not significant
Gender	Significant	Not significant	Not significant
Exposure to violent content	N/A[a]	Significant	Significant
Experience of being a perpetrator	N/A	Significant	Significant
COVID-19 × age	N/A	N/A	Significant
COVID-19 × experience of being a perpetrator	N/A	N/A	Significant

[a]N/A: not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we aimed to investigate the impact of COVID-19 and exposure to violent media content on cyber violence victimization among adolescents using national surveys conducted in South Korea. The findings are as follows: (1) COVID-19 did not exacerbate the likelihood of adolescents being victimized; instead, it was significantly associated with a decrease in the likelihood of victimization; (2) the impact of internet use, gender, and exposure to violent content on cyber violence victimization did not change depending on whether it was before or after COVID-19, but age and being a perpetrator did; (3) young adolescents were more victimized to cyber violence than older adolescents; (4) exposure to violent content led to cyber violence victimization; and (5) the experience of being a perpetrator was correlated with experiencing cyber violence victimization.

In accordance with prior research on cyber violence, it is conceivable that the heightened internet use [33,34] and exposure to violent content [37,40-42] during COVID-19 might have led many to anticipate a worsening of cyber victimization among adolescents. However, it turns out that COVID-19 did not worsen the likelihood of adolescents being victimized; rather, it was significantly related to the reduction in the likelihood of being victimized. This finding is consistent with the research by Garthe et al [11], who reported a decrease in cyber victimization during the COVID-19 pandemic when comparing data collected from US middle school students from 2019 to 2021. Our study confirms a similar pattern across an even broader age range among South Korean adolescents.

Furthermore, in our third model, we tested various interaction terms with the time factor to compare the impact of COVID-19 on victimization. The explained variance only increased by 2% in the third model, and most interaction terms were not significant. In other words, in terms of internet use, gender, and exposure to violent content, the results did not change depending on whether it was before or after COVID-19. The only significant changes were regarding age and the experience of being a perpetrator.

The findings have implications on how we should view the impact of COVID-19 on cyber violence. Many have speculated that COVID-19 would lead to heightened rates of cyber violence among adolescents, attributing this assumption to the natural increase in time spent online and heightened exposure to various types of content. This argument is strengthened by the fact that some studies have identified a significant increase in abusive content generated on social media platforms during the pandemic [48]. However, it turns out that COVID-19 did not intensify cyber victimization; rather it decreased during the pandemic, at least in South Korea. The study shows that we need to be aware of how we should interpret the impact of COVID-19 and that we need to carefully examine other factors that may have contributed to the decrease in cyber victimization.

One possible explanation is that cyber victimization may be closely related to offline victimization. Because of the stay-home measures, most schools had to implement remote learning during COVID-19, which may have helped break the vicious cycle of offline bullying and victimization among adolescents. Previous studies suggest that perpetration and victimization in cyberbullying are related to interpersonal peer aggression [14,36]. For instance, two-thirds of the adolescents who have reported being exposed to cyberbullying in the previous month mentioned that they were also bullied in person at school during the same period [35]. Similarly, over three-quarters of cyberbullying perpetrators admitted that they also bullied others in person within 1 month [35]. These findings indicate the possibility of offline violence continuing in cyberspace as a form of cyber violence. In that sense, COVID-19 may have prevented individuals from being exposed to other in-person violence among peers, and that may have led to less victimization online. This argument has been raised in a previous study addressing the role of face-to-face interactions and how this may fuel online bullying [11].

Reduced cyber victimization can also be explained by the quality of social interaction made online. Previous studies have indicated that increased smartphone use and online communication build social capital, leading to greater psychological well-being among adolescents [39]. This underscores the importance of considering the types of people adolescents interact with, the online communities they are exposed to, and the content they encounter, rather than solely focusing on the total time spent online, when evaluating cyber violence victimization. Marciano et al [9] found a positive correlation between ill health and social media use, and between ill health and media addiction, during COVID-19; however, they emphasized that not all forms of digital media use are detrimental to one's mental health. Specifically, engaging in one-on-one communication within online friendships in the midst of the pandemic alleviated feelings of loneliness and stress. This demonstrates that the use of media technology is relevant to the mental health of adolescents and to their exposure to cyber violence victimization.

As both the previous literature and this study suggest the potential for interpersonal violence to translate into cyber violence, future studies need to examine the intricate connections between offline and online violence, as well as how violence crosses over between these 2 spaces. In addition, future research should place a greater emphasis on the agency of adolescents, particularly by examining the types of media that adolescents choose to engage with and the communities to which they are most exposed. A comprehensive understanding of the context of the online environment, as well as the proclivities of individuals navigating specific online spaces, will provide valuable insights that can guide adolescents in minimizing cyber violence victimization.

While interventions directly suggested by our findings are beyond the scope of this study, future researchers should consider creative approaches to intervening in media use among adolescents. Rather than focusing on regulating internet usage, we may have to focus more on real-life interventions so that adolescents are not exposed to cyber violence that stems from in-person violence at school.

XSL•FO
RenderX

According to our data, younger adolescents are more likely to be victimized by cyber violence; therefore, it may be worthwhile to intervene earlier rather than later in adolescence to reduce cyber victimization. Early education on cyber violence can prevent both victimization and perpetration. If we manage to lower perpetration, it also has a significant impact on victimization because being a perpetrator and being exposed to cyberbullying are correlated [37]. Exposure to cyber victimization is also associated with exposure to violent content. Paying attention to what type of media content early adolescents are exposed to is crucial because it impacts not only what they see online but also what kind of online communities they build.

Limitations and Future Research

This study does not nuance the specific types of media content to which adolescents were exposed, nor does it address how offline and online social systems may contribute to the content that adolescents encounter. Maghsoudi et al [49] proposed that online social capital moderates the relationship between online risk exposure and psychological distress. Their study suggested that adolescents who heavily rely on online social capital may be more vulnerable to risk exposure than those who rely less on online social capital. Hence, it is critically important to establish offline support systems for adolescents who are exposed to violent content or cyber violence. Emphasizing online literacy education is also vital for empowering adolescents to cultivate positive online social capital and avoid communities that may pose a higher risk of cyber victimization. Exploring these areas could be valuable for future research. Another limitation of the study is the difference in survey distribution modalities between 2019 and 2020 due to the impact of the COVID-19 pandemic. In 2019, the survey of adolescents was conducted on-site in a classroom setting. However, the 2020 survey was conducted entirely online due to the restrictions of COVID-19. While some research has shown minimal significant discrepancies in results between offline and online surveys, our study was not able to account for the potential differences in results that could stem from the differences in these survey methods. Caution should also be exercised in interpreting the results, as the data are second-hand and based on a self-report survey. Although questions continue to be raised about the veracity of information from self-report methods [50], self-report data are neither inherently valid nor invalid; rather, they vary with the methodological sophistication of the data collection [51]. The validity of self-report information can be improved by minimal standardization of the procedures and high motivation and clear cognitive processes of the participants [51]. As the survey has been conducted annually since 2013, and focus groups of students and teachers have been studied regarding the questionnaire items and the procedures [52], the minimum validity of the data would be guaranteed.

In this study, we examined cyber violence data from a national survey conducted in South Korea, limited to the years 2019 and 2020. Our findings are specific to these 2 years, and we recognize that the trends may differ because COVID-19 would presumably have long-term effects. At the same time, adolescents have become more accustomed to remote interactions since the advent of COVID-19. For future research, we intend to conduct longitudinal data analyses and examine whether the trends in victimization have remained consistent.

Conclusions

This study sheds light on the complex relationship between COVID-19, exposure to violent media content, and cyber violence victimization among adolescents. The unexpected finding that COVID-19 was associated with a decrease rather than an increase in cyber victimization, or that it had no significance in cyber victimization, challenges common assumptions and underscores the need for nuanced approaches to understanding and addressing cyber violence. Furthermore, the study raises the possibility of the interconnectedness of offline and online violence, suggesting that interventions targeting cyber victimization should consider broader social contexts. Moving forward, policy makers, educators, and researchers must collaborate to develop comprehensive strategies that address both offline and online forms of violence, promote positive online interactions, and empower adolescents with the necessary skills to navigate digital spaces safely. By doing so, we can work toward cultivating a safer and more inclusive online environment for adolescents worldwide.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of the Korea and National Research Foundation of Korea (NRF-2020S1A3A2A02095619).

Data Availability

The data sets analyzed for this study will be available from the corresponding author upon reasonable request.

Authors' Contributions

EL, PJS, and HEL designed the research. HEL analyzed the data. EL drafted the manuscript. PJS and HEL critically reviewed the manuscript. HEL had primary responsibility for the final content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Van Schaik KD, DeWitte SN. COVID-19 and the black death: nutrition, frailty, inequity, and mortality. J Health Soc Sci 2020;5(4):471-484 [FREE Full text] [doi: 10.19204/2020/cvdn3]
2. Wagner A, Reifegerste D. In: Ribeiro N, Schwarzenegger C, editors. From Black Death to COVID-19: The Mediated Dissemination of Fear in Pandemic Times. Cham, Switzerland: Palgrave Macmillan; 2022.
3. Holton E, Naid C. The 1918 flu and COVID-19: a tale of two pandemics. The NIH Catalyst. 2022. URL: https://irp.nih.gov/catalyst/29/2/the-1918-flu-and-covid-19-a-tale-of-two-pandemics [accessed 2024-03-07]
4. den Hamer AH, Konijn EA. Adolescents' media exposure may increase their cyberbullying behavior: a longitudinal study. J Adolesc Health 2015 Feb;56(2):203-208. [doi: 10.1016/j.jadohealth.2014.09.016] [Medline: 25620303]
5. Prot S, Anderson CA, Barlett CP, Coyne SM, Saleem M. Content effects: Violence in the media. In: Roessler P, Hoffner CA, van Zoonen L, editors. The International Encyclopedia of Media Effects. Hoboken, NJ: Wiley-Blackwell; 2017:1-9.
6. Schoffstall CL, Cohen R. Cyber aggression: the relation between online offenders and offline social competence. Soc Dev 2011 Mar 25;20(3):587-604. [doi: 10.1111/j.1467-9507.2011.00609.x]
7. Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N. Cyberbullying: its nature and impact in secondary school pupils. J Child Psychol Psychiatry 2008 Apr;49(4):376-385. [doi: 10.1111/j.1469-7610.2007.01846.x] [Medline: 18363945]
8. Sasson H, Mesch G. Parental mediation, peer norms and risky online behavior among adolescents. Comput Hum Behav 2014 Apr;33:32-38. [doi: 10.1016/j.chb.2013.12.025]
9. Marciano L, Ostroumova M, Schulz PJ, Camerini A. Digital media use and adolescents' mental health during the COVID-19 pandemic: a systematic review and meta-analysis. Front Public Health 2021 Feb 1;9:793868 [FREE Full text] [doi: 10.3389/fpubh.2021.793868] [Medline: 35186872]
10. Worlds Apart: Canadian Companion to UNICEF Report Card 16. UNICEF Canada. 2020. URL: https://tinyurl.com/kk6cesw5 [accessed 2024-03-11]
11. Garthe RC, Kim S, Welsh M, Wegmann K, Klingenberg J. Cyber-victimization and mental health concerns among middle school students before and during the COVID-19 pandemic. J Youth Adolesc 2023 Apr;52(4):840-851 [FREE Full text] [doi: 10.1007/s10964-023-01737-2] [Medline: 36754916]
12. Pereira F, Spitzberg BH, Matos M. Cyber-harassment victimization in Portugal: prevalence, fear and help-seeking among adolescents. Computers in Human Behavior 2016;62:136-146. [doi: 10.1016/j.chb.2016.03.039]
13. 2018 Cyberviolence survey. Korea Communications Commission & National Information Society Agency. 2019. URL: https://tinyurl.com/ymf3z44s [accessed 2023-02-24]
14. Camerini A, Marciano L, Carrara A, Schulz PJ. Cyberbullying perpetration and victimization among children and adolescents: a systematic review of longitudinal studies. Telematics and Informatics 2020 Jun;49:101362. [doi: 10.1016/j.tele.2020.101362]
15. Hinduja S, Patchin JW. Bullying, cyberbullying, and suicide. Arch Suicide Res 2010 Jul 28;14(3):206-221. [doi: 10.1080/13811118.2010.494133] [Medline: 20658375]
16. Byeon H. Prediction of adolescent suicidal ideation after the COVID-19 pandemic: a nationwide survey of a representative sample of Korea. Front Pediatr 2022 Jul 25;10:951439 [FREE Full text] [doi: 10.3389/fped.2022.951439] [Medline: 35958177]
17. Lee I, Choi J, Kim KS, Suh J, Kim JH, Kim S. Suicide attempts presenting to the emergency department before and during the COVID-19 pandemic: a comparative study. Clin Exp Emerg Med 2022 Jun;9(2):120-127 [FREE Full text] [doi: 10.15441/ceem.21.088] [Medline: 35843612]
18. Bullo A, Schulz PJ. Do peer and parental norms influence media content-induced cyber aggression? Comput Hum Behav 2022 Apr;129:107136. [doi: 10.1016/j.chb.2021.107136]
19. Suter L, Waller G, Bernath J, Külling C, Willemse I, Süss D. JAMES: Youth, Activities, Media - Swiss Survey. ZHAW Zurich University of Applied Sciences. Zurich, Switzerland: Zürcher Hochschule für Angewandte Wissenschaften; 2018. URL: https://tinyurl.com/ywauvjz9 [accessed 2024-03-14]
20. Blackwell L, Gardiner E, Schoenebeck S. Managing expectations: technology tensions among parents and teens. New York, NY: Association for Computing Machinery; 2016 Presented at: The 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing; February 27, 2016; Minneapolis, MN p. 1390-1401. [doi: 10.1145/2818048.2819928]
21. Odgers CL, Jensen MR. Annual research review: adolescent mental health in the digital age: facts, fears, and future directions. J Child Psychol Psychiatry 2020 Mar;61(3):336-348 [FREE Full text] [doi: 10.1111/jcpp.13190] [Medline: 31951670]
22. Korea Press Foundation. 2020 Children Media Usage Report. Korea Press Foundation. 2021. URL: https://tinyurl.com/5x73e663 [accessed 2023-03-11]
23. Choi MJ, Kim KJ, Han S. Factors related to suicidal ideation in Korean high school students. Journal of the Korea Convergence Society 2021;12(11):487-499. [doi: 10.15207/JKCS.2021.12.11.487]
24. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. Psychol Rev 2010 Apr;117(2):575-600 [FREE Full text] [doi: 10.1037/a0018697] [Medline: 20438238]
25. Park JY, Lee I. Factors influencing suicidal tendencies during COVID-19 pandemic in Korean multicultural adolescents: a cross-sectional study. BMC psychology 2022;10(1):1-13 [FREE Full text] [doi: 10.1186/s40359-022-00867-9] [Medline: 35729606]
26. Field T. Anxiety and depression in youth during COVID-19: a narrative review. Journal of Psychiatry Research Reviews & Reports 2021;3(2):1-8. [doi: 10.47363/jpsrr/2021(3)126]

XSL•FO
RenderX

27. Shek DT, Zhao L, Dou D, Zhu X, Xiao C. The impact of positive youth development attributes on posttraumatic stress disorder symptoms among chinese adolescents under COVID-19. J Adolesc Health 2021 Apr;68(4):676-682 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.01.011] [Medline: 33583684]
28. Liang L, Ren H, Cao R, Hu Y, Qin Z, Li C, et al. The effect of COVID-19 on youth mental health. Psychiatr Q 2020 Sep;91(3):841-852 [FREE Full text] [doi: 10.1007/s11126-020-09744-3] [Medline: 32319041]
29. Lundström M. Young in pandemic times: a scoping review of COVID-19 social impacts on youth. Int J Adolescence Youth 2022 Aug 30;27(1):432-443. [doi: 10.1080/02673843.2022.2117637]
30. Chang CI, Fong Sit H, Chao T, Chen C, Shen J, Cao B, et al. Exploring subtypes and correlates of internet gaming disorder severity among adolescents during COVID-19 in China: a latent class analysis. Curr Psychol 2022 Apr 29:1-12 [FREE Full text] [doi: 10.1007/s12144-022-03133-8] [Medline: 35505828]
31. Lin M. Prevalence of internet addiction during the COVID-19 outbreak and its risk factors among junior high school students in Taiwan. Int J Environ Res Public Health 2020 Nov 18;17(22):8547 [FREE Full text] [doi: 10.3390/ijerph17228547] [Medline: 33218018]
32. Rideout V, Robb MB. Social Media, Social Life: Teens Reveal Their Experiences. San Francisco, CA: Common Sense Media; 2018.
33. Snell PA, Englander E. Cyberbullying victimization and behaviors among girls: applying research findings in the field. J Soc Sci 2010;6:510-514. [doi: 10.3844/jssp.2010.510.514]
34. You S, Lim SA. Longitudinal predictors of cyberbullying perpetration: evidence from Korean middle school students. Pers Individual Differences 2016 Jan;89:172-176. [doi: 10.1016/j.paid.2015.10.019]
35. Hinduja S, Patchin JW. Cyberbullying: neither an epidemic nor a rarity. Eur J Developmental Psychol 2012 Sep;9(5):539-543. [doi: 10.1080/17405629.2012.706448]
36. Ybarra ML, Mitchell KJ. Online aggressor/targets, aggressors, and targets: a comparison of associated youth characteristics. J Child Psychol Psychiatry 2004 Oct;45(7):1308-1316. [doi: 10.1111/j.1469-7610.2004.00328.x] [Medline: 15335350]
37. Bae SM. The relationship between exposure to risky online content, cyber victimization, perception of cyberbullying, and cyberbullying offending in Korean adolescents. Child Youth Services Rev 2021 Apr;123:105946. [doi: 10.1016/j.childyouth.2021.105946]
38. Kim HH. The impact of online social networking on adolescent psychological well-being (WB): a population-level analysis of Korean school-aged children. Int J Adolescence Youth 2016 Jun 16;22(3):364-376. [doi: 10.1080/02673843.2016.1197135]
39. Bae SM. The relationship between smartphone use for communication, social capital, and subjective well-being in Korean adolescents: Verification using multiple latent growth modeling. Child Youth Services Rev 2019 Jan;96:93-99. [doi: 10.1016/j.childyouth.2018.11.032]
40. Greitemeyer T, Mügge DO. Video games do affect social outcomes: a meta-analytic review of the effects of violent and prosocial video game play. Pers Soc Psychol Bull 2014 May;40(5):578-589. [doi: 10.1177/0146167213520459] [Medline: 24458215]
41. Huesmann LR, Moise-Titus J, Podolski CL, Eron LD. Longitudinal relations between children's exposure to TV violence and their aggressive and violent behavior in young adulthood: 1977-1992. Developmental Psychol 2003;39(2):201-221. [doi: 10.1037/0012-1649.39.2.201]
42. Saleem M, Anderson CA, Gentile DA. Effects of prosocial, neutral, and violent video games on children's helpful and hurtful behaviors. Aggress Behav 2012;38(4):281-287. [doi: 10.1002/ab.21428] [Medline: 25363697]
43. Wang J, Iannotti RJ, Nansel TR. School bullying among adolescents in the United States: physical, verbal, relational, and cyber. J Adolesc Health 2009 Oct;45(4):368-375 [FREE Full text] [doi: 10.1016/j.jadohealth.2009.03.021] [Medline: 19766941]
44. Choi KS, Lee JR. Theoretical analysis of cyber-interpersonal violence victimization and offending using cyber-routine activities theory. Comput Hum Behav 2017 Aug;73:394-402. [doi: 10.1016/j.chb.2017.03.061]
45. Backe EL, Lilleston P, McCleary-Sills J. Networked individuals, gendered violence: a literature review of cyberviolence. Violence Gender 2018 Sep;5(3):135-146. [doi: 10.1089/vio.2017.0056]
46. Yuliati R, Saptyasari A. Cyberbullying involvement: impacts of violence exposure in the media, family, society, and school. Jurnal Komunikasi Indonesia 2019;8(1):5-11. [doi: 10.7454/jki.v3i1.9857]
47. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed). New York, NY: Routledge; 2002.
48. Babvey P, Capela F, Cappa C, Lipizzi C, Petrowski N, Ramirez-Marquez J. Using social media data for assessing children's exposure to violence during the COVID-19 pandemic. Child Abuse Negl 2021 Jun;116(Pt 2):104747 [FREE Full text] [doi: 10.1016/j.chiabu.2020.104747] [Medline: 33358281]
49. Maghsoudi R, Shapka J, Wisniewski P. Examining how online risk exposure and online social capital influence adolescent psychological stress. Comput Hum Behav 2020 Dec;113:106488. [doi: 10.1016/j.chb.2020.106488]
50. Schwarz N. Self-reports: how the questions shape the answers. Am Psychol 1999;54(2):93-105 [FREE Full text] [doi: 10.1037//0003-066x.54.2.93]
51. Del Boca FK, Noll JA. Truth or consequences: the validity of self‐report data in health services research on addictions. Addiction 2002;95(11s3):347-360. [doi: 10.1046/j.1360-0443.95.11s3.5.x]

XSL•FO
RenderX

52. Hwang SW, Jo JS, Oh YH, Lee JH, Im DK, Kim S. 2013 Cyber violence field investigation. Korea Communications Commission/Korea Internet & Security Agency. 2013. URL: https://www.kcc.go.kr/ [accessed 2024-03-10]

Abbreviations

KCC: Korea Communications Commission
PTSD: posttraumatic stress disorder

Edited by A Mavragani; submitted 07.01.23; peer-reviewed by H Hochheiser, A AL-Asadi; comments to author 30.11.23; revised version received 09.01.24; accepted 23.02.24; published 22.03.24.

Please cite as:
Lee E, Schulz PJ, Lee HE
The Impact of COVID-19 and Exposure to Violent Media Content on Cyber Violence Victimization Among Adolescents in South Korea: National Population-Based Study
J Med Internet Res 2024;26:e45563
URL: https://www.jmir.org/2024/1/e45563
doi:10.2196/45563
PMID:38517467

©Eugene Lee, Peter J Schulz, Hye Eun Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Moderating Effect of Coping Strategies on the Association Between the Infodemic-Driven Overuse of Health Care Services and Cyberchondria and Anxiety: Partial Least Squares Structural Equation Modeling Study

Richard Huan Xu[1], PhD; Caiyun Chen[2], PhD

[1]Department of Rehabilitation Sciences, Faculty of Health and Social Sciences, Hong Kong Polytechnic University, Hung Hom, China (Hong Kong)
[2]Nanjing Academy of Administration, Nanjing, China

Corresponding Author:
Richard Huan Xu, PhD
Department of Rehabilitation Sciences
Faculty of Health and Social Sciences
Hong Kong Polytechnic University
11 Yuk Choi Rd
Hung Hom
China (Hong Kong)
Phone: 852 27664199
Email: richard.xu@polyu.edu.hk

Abstract

Background: The COVID-19 pandemic has led to a substantial increase in health information, which has, in turn, caused a significant rise in cyberchondria and anxiety among individuals who search for web-based medical information. To cope with this information overload and safeguard their mental well-being, individuals may adopt various strategies. However, the effectiveness of these strategies in mitigating the negative effects of information overload and promoting overall well-being remains uncertain.

Objective: This study aimed to investigate the moderating effect of coping strategies on the relationship between the infodemic-driven misuse of health care and depression and cyberchondria. The findings could add a new dimension to our understanding of the psychological impacts of the infodemic, especially in the context of a global health crisis, and the moderating effect of different coping strategies on the relationship between the overuse of health care and cyberchondria and anxiety.

Methods: The data used in this study were obtained from a cross-sectional web-based survey. A professional survey company was contracted to collect the data using its web-based panel. The survey was completed by Chinese individuals aged 18 years or older without cognitive problems. Model parameters of the relationships between infodemic-driven overuse of health care, cyberchondria, and anxiety were analyzed using bootstrapped partial least squares structural equation modeling. Additionally, the moderating effects of coping strategies on the aforementioned relationships were also examined.

Results: A total of 986 respondents completed the web-based survey. The mean scores of the Generalized Anxiety Disorder-7 and Cyberchondria Severity Scale-12 were 8.4 (SD 3.8) and 39.7 (SD 7.5), respectively. The mean score of problem-focused coping was higher than those of emotion- and avoidant-focused coping. There was a significantly positive relationship between a high level of infodemic and increased overuse of health care (bootstrapped mean 0.21, SD 0.03; 95% CI 0.1581-0.271). The overuse of health care resulted in more severe cyberchondria (bootstrapped mean 0.107, SD 0.032) and higher anxiety levels (bootstrapped mean 0.282, SD 0.032) in all the models. Emotion (bootstrapped mean 0.02, SD 0.008 and 0.037, SD 0.015)- and avoidant (bootstrapped mean 0.026, SD 0.009 and 0.049, SD 0.016)-focused coping strategies significantly moderated the relationship between the overuse of health care and cyberchondria and that between the overuse of health care and anxiety, respectively. Regarding the problem-based model, the moderating effect was significant for the relationship between the overuse of health care and anxiety (bootstrapped mean 0.007, SD 0.011; 95% CI 0.005-0.027).

Conclusions: This study provides empirical evidence about the impact of coping strategies on the relationship between infodemic-related overuse of health care services and cyberchondria and anxiety. Future research can build on the findings of this study to further explore these relationships and develop and test interventions aimed at mitigating the negative impact of the infodemic on mental health.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53417) doi:10.2196/53417

KEYWORDS

infodemic; health care; cyberchondria; anxiety; coping; structural equation modeling

Introduction

COVID-19–Related Mental Health Problems

In today's technologically advancing society, widespread and rapid digitization has led to a substantial increase in the use of social media and the internet. This, in turn, has facilitated the rapid dissemination of all types of information. Although this can be beneficial in filling information gaps quickly, it has its drawbacks. A prominent drawback is the amplification of harmful messages, which can have negative effects on individuals [1,2]. The World Health Organization (WHO) acknowledged the presence of an infodemic during the COVID-19 pandemic and subsequent responses. WHO defines an infodemic as an excessive amount of information, including both accurate and inaccurate content [3]. This abundance of information makes it difficult for individuals to distinguish reliable sources from unreliable sources and to find trustworthy guidance when they need it.

Excessive use of health care services can have adverse effects on individuals and the overall sustainability of health care systems. Although challenges associated with the overuse of health care services were evident before the COVID-19 pandemic [4,5], the urgent need for sustainable health care systems was exacerbated by the pandemic. Because large portions of the population were instructed to self-isolate at home and had limited access to health care professionals during the pandemic, the internet became the primary source of information for numerous individuals seeking answers to health-related questions. However, the abundance of web-based information, including both true and false content, can leave individuals feeling overwhelmed and struggling to make informed choices. This information overload can lead to depression because individuals bombarded with conflicting messages may feel unsure of what to believe [6-10].

Besides depression, cyberchondria has also emerged as a significant public health challenge since the onset of the COVID-19 pandemic. This refers to the repeated and excessive search for health-related information on the internet, leading to a significant increase in distress or anxiety [11]. Although the global emergency caused by the COVID-19 pandemic is over, telehealth remains a growing trend. An increasing number of studies have indicated that telehealth can improve health care access, outcomes, and affordability by offering a bridge to care and an opportunity to reinvent web-based care models [12]. However, increasing internet exposure increases the risk of cyberchondria, especially under conditions of uncertainty and increased risk, due to the large volume of information it contains. Thus, it is crucial to understand how to provide support and guidance to help people adopt appropriate strategies for using web-based resources safely in the context of an infodemic.

Current Research on the COVID-19–Related Infodemic

The harms of infodemic are well documented. An Italian study suggested developing early warning signals for an infodemic, which can provide important cues for implementing effective communication strategies to mitigate misinformation [13]. Other studies have shown that successful use of coping strategies can help individuals manage stressful events and reduce negative emotions during a pandemic. For example, Yang [14] found a positive correlation between emotion-focused coping and cyberbullying and depression during the COVID-19 pandemic. A large-scale UK study indicated that supportive coping was associated with a faster decrease in depression and anxiety symptoms [15]. Shigeto et al [16] emphasized the importance of training young adults to develop resilience, flexibility, and specific coping skills to offset the psychological effects of significant lifestyle changes resulting from pandemics or other health crises in the future. A recent study used machine learning technology to enhance the accuracy and efficiency of automated fact-checking and infodemic risk management at a strategic level [17]. However, the impact of coping strategies on the relationship among the infodemic, cyberchondria, and anxiety at an individual level during the COVID-19 pandemic is still unknown.

Importance of Coping Strategies

The ability of individuals to discern and adopt appropriate coping strategies can have a profound impact on their mental health, particularly in relation to conditions such as depression and anxiety. The ability to select and implement coping strategies is not uniform across all individuals, and these differences can significantly influence the trajectory of their mental health outcomes. For some, the ability to effectively choose and implement coping strategies can serve as a protective factor, mitigating the severity of the symptoms of depression or anxiety and promoting overall health and well-being. Conversely, for others, inability or difficulty in selecting and implementing effective coping strategies can exacerbate mental health conditions, leading to increased severity of depression and anxiety. This, in turn, can have detrimental effects on individuals' overall health and well-being. Therefore, understanding the factors that influence individuals' ability to select and implement effective coping strategies is of paramount importance in the field of mental health research and intervention [18].

Research has demonstrated the importance of appropriate coping mechanisms in managing mental health problems. Coping strategies, which are essential for dealing with stress or challenging situations, can be categorized into 3 primary types: emotion focused, problem focused, and avoidant focused [19]. Emotion-focused strategies are centered around managing and regulating emotions. They serve as a means to cope with stress or difficult situations. These strategies might involve seeking

XSL•FO
RenderX

emotional support from others, using relaxation techniques, or practicing mindfulness. In contrast, problem-focused strategies actively address the problem or stressor. These strategies might encompass problem-solving, devising a plan of action, or seeking information and resources to effectively tackle the situation. Avoidant-focused strategies involve evading or distancing oneself from the stressor or problem. These strategies might include denial, distraction, or engaging in activities to escape or avoid contemplating the issue [18]. The effectiveness of different coping strategies can vary depending on the situation. Individuals often use different or a combination of strategies, tailoring their approach to their circumstances.

Coping Strategies in the COVID-19–Related Infodemic

From a social perspective, this study underscores the importance of mental health in the context of public health emergencies such as the COVID-19 pandemic. It highlights the need for society to recognize and address the mental health burden that such emergencies can place on individuals, particularly in relation to the phenomenon of cyberchondria, which is the unfounded escalation of concerns about common symptoms based on reviews of web-based literature and resources.

Practically, this study provides valuable insights for policy makers and practitioners. It emphasizes the need for the development of effective coping strategies and programs to manage the negative impact of an overload of misinformation and disinformation on mental health. This is particularly relevant in the digital age, where individuals have access to a plethora of information, not all of which is accurate or reliable. Policy makers and practitioners can use the findings of this study to design interventions that not only provide accurate information but also equip individuals with the skills to distinguish reliable sources from unreliable sources and to cope with the anxiety that misinformation can cause. From a research standpoint, this study fills a gap in the literature by assessing the impact of the infodemic on cyberchondria and the moderating effect of coping strategies in this relationship. It opens up new avenues of research into the complex interplay among public health emergencies, infodemic, cyberchondria, and coping strategies. Future research could build on the findings of this study to further explore these relationships and develop and test interventions aimed at mitigating the negative impact of infodemic on mental health.

Objective of the Study

Currently, the association between the overuse of health care services and mental health problems in the context of an infodemic remains unclear, as is the moderating effect of different coping strategies on this association. Thus, this study investigated the moderating effect of coping strategies on the relationship between the infodemic-driven misuse of health care and depression and cyberchondria.

Hypotheses of the Study

The study used a hypothesis-driven format. Specifically, there are five hypotheses: (1) a positive relationship exists between infodemic and the misuse of health care, (2) a positive relationship exists between the misuse of health care and depressive disorders, (3) a positive relationship exists between the misuse of health care and cyberchondria, (4) coping strategies mitigate the negative effect of the misuse of health care on depression, and (5) coping strategies mitigate the negative effect of the misuse of health care on cyberchondria. Hypotheses 2-5 are separately evaluated for the three types of coping strategies: problem focused (H2.1), emotion focused (H2.2), and avoidant focused (H2.3).

Methods

Study Design and Sample Size

The data used in this study were obtained from a cross-sectional and web-based survey conducted between April and May 2023 in China.

There is no gold standard for sample estimation in partial least squares structural equation modeling (PLS-SEM). Following Hair et al [20], we set the significance level at 5% and the minimum path coefficients to between 0.05 and 0.1. Based on these criteria, a minimum sample size of 619 was determined.

Data Source and Collection

A professional surveying company, WenJuanXing, was invited to collect the data through its web-based panel. The panel of WenJuanXing consists of 2.6 million members, with an average of over 1 million questionnaire respondents daily. At the beginning of the project, a survey manager collaborated with the research team to screen and recruit participants using the company's internal social network platform. All of the eligible panel members received a survey invitation, and a voluntary response sampling method was used. The survey manager checked the data quality using WenJuanXing's artificial intelligence data quality control system to ensure that respondents met our inclusion criteria and provided valid responses, thus ensuring a high level of data accuracy and integrity. The inclusion criteria were (1) aged older than 18 years, (2) able to understand and read Chinese, and (3) agreed to provide informed consent. All eligible respondents were invited to participate in a web-based survey. The first section of the survey was the informed consent, which the participants were required to read and agree to before proceeding. All the participants who agreed to participate in the survey were asked to complete six questionnaires covering (1) demographics and socioeconomic status, (2) COVID-19 information–related questions, (3) a cyberchondria questionnaire, (4) an eHealth literacy questionnaire, (5) an anxiety questionnaire, and (6) a coping strategy questionnaire. The English translations of the questionnaires are presented in Multimedia Appendix 1. To ensure data quality, we collaborated with the survey company and implemented various indicators. We monitored completion time, excluding responses that took less than 6 minutes. We also tracked ID addresses, ensuring that each ID address could only complete the questionnaire once. To minimize random errors, we used an artificial intelligence formula developed by the survey company to identify and filter any response patterns that appeared to be generated in parallel.

Ethical Considerations

The study protocol and informed consent process were approved by the institutional review board of the Hong Kong Polytechnic

XSL•FO
RenderX

University (HSEARS20230502006). Informed consent was collected from all participants. The survey was conducted anonymously, and no personally identifiable information was collected. No compensation was provided by the research team.

Instruments

Cyberchondria Severity Scale-12

The Cyberchondria Severity Scale-12 (CSS-12), derived from the 33-item CSS, was used to measure the severity of cyberchondria. The CSS-12 exhibited equally good psychometric properties as the original version and has been validated in Chinese populations [21]. The CSS-12 items are scored on a Likert-type scale ranging from 1="never" to 5="always," giving total scores ranging from 12 to 60. A higher score indicates a higher severity of suspected cyberchondria. The psychometric properties of the Chinese version of the CSS-12 were reported by Peng et al [22].

Generalized Anxiety Disorder Assessment

The Generalized Anxiety Disorder Assessment-7 (GAD-7) was used to screen for generalized anxiety disorder and related anxiety disorders [23]. This scale consists of 7 items designed to assess the frequency of anxiety symptoms during the 2 weeks preceding the survey. The GAD-7 score is calculated by assigning scores of 0, 1, 2, and 3 to the response categories of "not at all," "several days," "more than half the days," and "nearly every day," respectively. The scores of the 7 questions are then summed, giving a total ranging from 0 to 21, with higher scores indicating a higher severity of anxiety disorders. Many studies have reported the psychometric properties of the GAD-7 in Chinese populations, such as that conducted by Sun et al [24].

Coping Orientation to Problems Experienced Inventory

The Coping Orientation to Problems Experienced Inventory (Brief-COPE) is a 28-item self-report questionnaire used to measure effective and ineffective strategies for coping with a stressful life event [25]. The Brief-COPE assesses how a person deals with stressors in their daily life. The questionnaire measures 3 coping strategy dimensions: problem focused, emotion focused, and avoidant focused [26]. Each item is rated on a 4-point scale. The scores for the 3 overarching coping styles are calculated as average scores. This is done by dividing the sum of the item scores by the number of items. These average scores indicate the extent to which the respondent engages in each coping style. A higher score indicates that the respondent does not have many coping skills. The Chinese version of the Brief-COPE and its psychometric properties in Chinese populations were reported by Wang et al [27].

Infodemic- and Misinformation-Driven Overuse of Health Care Services

The COVID-19–related infodemic and misinformation-driven medical misbehavior were assessed using 2 self-developed items. The first item was "Do you believe there is an excessive amount of information regarding the COVID virus and vaccine on a daily basis?" The second item was "Has misinformation or disinformation about COVID-19 led you to engage in the overuse of health care services (eg, frequently visiting the doctor/psychiatrist or buying unnecessary medicine)?" The respondents were required to indicate their response to these 2 questions by selecting 1 of 2 options presented dichotomously: yes or no.

Statistical Analysis

Descriptive statistics were used to describe the participants' background characteristics. Continuous variables (eg, age) were calculated as means and SDs. Categorical variables (eg, sex) were calculated as frequencies and proportions. The Pearson correlation coefficient (r) was used to examine the association between measures, where $r \geq 0.3$ and $r \geq 0.5$ indicated moderate and large effects, respectively [28,29].

In this study, we used PLS-SEM to estimate the research model parameters, as it works efficiently with small samples and complex models. Compared with covariance-based structural equation modeling, PLS-SEM has several advantages, such as the ability to handle non-normal data and small samples [30]. Unlike covariance-based structural equation modeling, which focuses on confirming theories, PLS-SEM is a causal-predictive approach that explains variance in the model's dependent variables [31]. To improve the model fit, we used the bootstrapping method with 10,000 replications to obtain the estimates of the mean coefficients and 95% CIs [32]. Composite reliability rho_a (>0.7), composite reliability rho_c (>0.7), and average variance extracted (>0.5) were used to examine the model performance.

PLS-SEM encompasses measurement models that define the relationship between constructs (instruments) and indicator variables and a structural model. The structural model used in this study is presented in Figure 1. We hypothesized that the infodemic significantly affects misinformation-driven medical misbehavior, resulting in cyberchondria and high anxiety levels. Furthermore, we speculated that coping strategies significantly modify this relationship. To test these hypotheses, we used 3 models that used the full sample to separately investigate the moderating effect of the 3 types of coping strategies (problem focused, emotion focused, and avoidant focused). We analyzed the data and estimated the PLS-SEM parameters using the "SEMinR" package in R (R Foundation for Statistical Computing). A P value of ≤.05 was considered statistically significant.

XSL•FO
RenderX

Figure 1. Conceptual framework of this study.

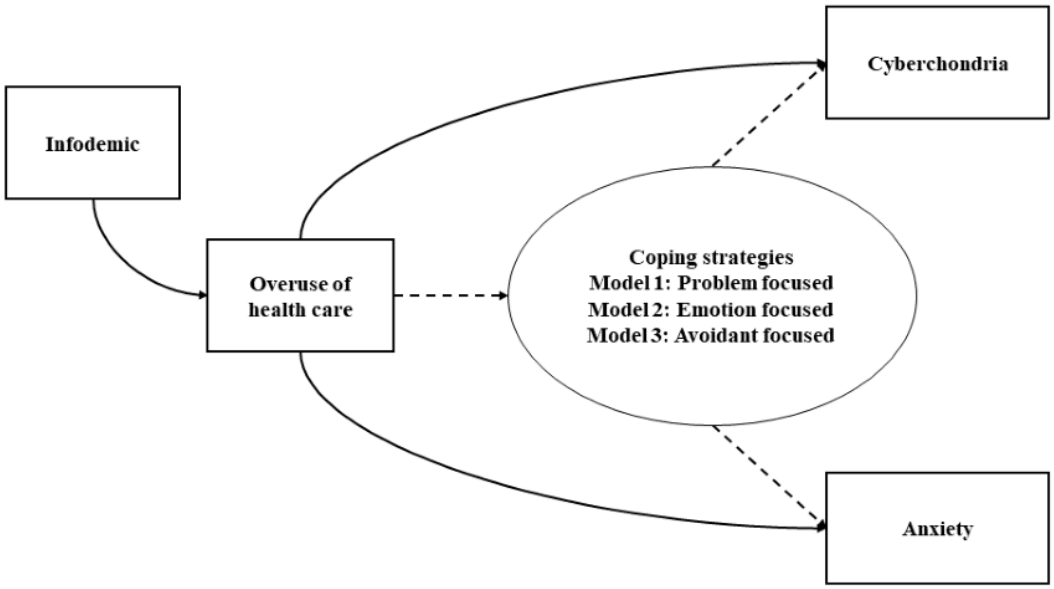

Results

Background Characteristics of Participants

A total of 986 respondents completed the web-based survey and provided valid responses, resulting in a response rate of 84%. Among the participants, 51.7% (n=510) were female, approximately 95% (n=933) had completed tertiary education or above, and 71.2% (n=702) resided in urban areas. The participants' background characteristics are listed in Table 1.

XSL•FO
RenderX

Table 1. Respondents' background characteristics (n=986).

	Respondents, n (%)
Sex	
Male	476 (48.3)
Female	510 (51.7)
Educational level	
Secondary or below	53 (5.4)
Tertiary or above	933 (94.6)
Household registry	
Urban	702 (71.2)
Rural	284 (28.8)
Employment	
Active	894 (90.7)
Nonactive	92 (9.3)
Marital status	
Single	232 (23.5)
Married	750 (76.1)
Divorce or widowed	4 (0.4)
Family annual income (CNY)[a]	
≤50,000	57 (5.8)
50,001-100,000	191 (19.4)
100,001-200,000	341 (34.6)
200,001-300,000	249 (25.3)
300,001-400,000	89 (9)
> 400,000	59 (6)
Diagnosed with chronic disease	
Yes	321 (32.6)
No	665 (67.4)

[a]A currency exchange rate of 7.23 CNY=US $1 applies.

Mean Scores and Frequency of Responses

The mean score of the GAD-7 was 8.4 (SD 3.8), while the mean score of the CSS-12 was 39.7 (SD 7.5). Problem-focused coping had a higher mean score than emotion- and avoidant-focused coping. Respondents with active employment reported statistically significantly higher mean scores on the GAD and avoidant-focused coping subscale compared to those with nonactive employment. A higher proportion of respondents with chronic diseases experienced an infodemic and exhibited the overuse of health care services relative to those without chronic diseases (Table 2). The correlations between all of the measures are presented in Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. Mean score of GAD-7[a], CSS-12[b], Brief-COPE[c], frequency of infodemic, and overuse of health care in different groups of background characteristics (n=986).

	Infodemic, yes, n (%)	Overuse of health care, mean (SD)	GAD-7, mean (SD)	CSS-12, mean (SD)	Brief-COPE, mean (SD)		
					Problem	Emotion	Avoidant
Overall	581 (58.9)	596 (60.4)	8.4 (3.8)	39.7 (7.5)	3.3 (0.4)	2.6 (0.3)	2.1 (0.5)
Sex							
Male	283 (59.5)	282 (59.2)	8.2 (3.7)	39.8 (7.6)	3.3 (0.5)	2.6 (0.4)	2.1 (0.5)
Female	298 (58.4)	314 (61.6)	8.6 (3.8)	39.6 (7.4)	3.3 (0.4)	2.6 (0.3)	2.1 (0.5)
Educational level							
Secondary or below	32 (59.6)	29 (55.8)	9 (4.5)	37.3 (7.8)[d]	3.2 (0.5)	2.6 (0.4)	2.2 (0.5)
Tertiary or above	549 (58.8)	567 (60.8)	8.3 (3.7)	39.9 (7.5)	3.3 (0.4)	2.6 (0.3)	2.1 (0.4)
Family registry							
Urban	428 (61)[d]	430 (61.3)	8.1 (3.7)[e]	40 (7.6)[f]	3.3 (0.5)	2.6 (0.3)	2.1 (0.5)
Rural	153 (53.9)	166 (58.5)	9 (4)	39 (7.1)	3.2 (0.4)	2.7 (0.3)	2.2 (0.5)
Employment							
Active	525 (58.7)	557 (62.3)[g]	9.8 (4.2)[g]	38.1 (7.3)[h]	3.2 (0.4)	2.6 (0.3)	2.2 (0.5)[i]
Nonactive	56 (60.9)	39 (42.4)	8.2 (3.7)	40 (7.5)	3.3 (0.5)	2.6 (0.3)	2.1 (0.4)
Chronic disease							
Yes	208 (64.8)[j]	239 (74.5)[g]	9.1 (4)[g]	41.2 (6.6)[g]	3.2 (0.4)	2.7 (0.3)[k]	2.2 (0.5)[g]
No	373 (56.1)	357 (53.7)	8 (3.7)	39 (7.8)	3.3 (0.5)	2.6 (0.3)	2.0 (0.4)

[a]GAD-7: Generalized Anxiety Disorder Assessment-7.
[b]CSS-12: Cyberchondria Severity Scale-12.
[c]COPE: Coping Orientation to Problems Experienced Inventory.
[d]P=.04.
[e]P=.002.
[f]P=.05.
[g]P<.001.
[h]P=.03.
[i]P=.02.
[j]P=.009.
[k]P=.01.

Measurement Models

Tables 3-5 present the performance of the measurement models for the 3 coping strategies. The values of rho_C and rho_A were above 0.7, indicating acceptable construct reliability. All 3 constructs had Cronbach α values exceeding the cutoff of 0.7, indicating adequate reliability. Table 2 presents the models' convergent validity. All the bootstrapped item loadings exceeded 0.3 and were significant at <.05 for the problem- and avoidant-focused models. However, for cyberchondria and the Brief-COPE, none of the average variance extracted values were above 0.5, indicating unsatisfactory model convergent validity.

XSL•FO
RenderX

Table 3. Performance of the measurement model: problem-focused model.

	Bootstrapped loadings (95% CI)	Cronbach α	rho_C	AVE[a]	rho_A
GAD-7[b]		0.842	0.878	0.505	0.865
GAD1	0.737 (0.692-0.774)				
GAD2	0.772 (0.738-0.804)				
GAD3	0.712 (0.661-0.755)				
GAD4	0.637 (0.58-0.688)				
GAD5	0.714 (0.665-0.756)				
GAD6	0.641 (0.584-0.689)				
GAD7	0.759 (0.721-0.794)				
Cyberchondria		0.833	0.865	0.350	0.834
CYB1	0.581 (0.514-0.638)				
CYB2	0.559 (0.427-0.65)				
CYB3	0.599 (0.538-0.653)				
CYB4	0.632 (0.54-0.693)				
CYB5	0.539 (0.362-0.678)				
CYB6	0.661 (0.609-0.703)				
CYB7	0.562 (0.432-0.651)				
CYB8	0.649 (0.56-0.709)				
CYB9	0.635 (0.537-0.704)				
CYB10	0.481 (0.344-0.581)				
CYB11	0.58 (0.458-0.66)				
CYB12	0.534 (0.374-0.655)				
Brief-COPE[c]		0.770	0.831	0.387	0.791
BF2	0.723 (0.668-0.773)				
BF7	0.729 (0.683-0.771)				
BF10	0.453 (0.34-0.546)				
BF12	0.67 (0.611-0.723)				
BF14	0.669 (0.61-0.72)				
BF17	0.567 (0.499-0.629)				
BF23	0.463 (0.345-0.554)				
BF25	0.614 (0.55-0.671)				
BF19	0.723 (0.668-0.773)				
Infodemic					
IF1	1.0 (1.0-1.0)	1.0	1.0	1.0	1.0
Overuse of HC[d]					
OHC1	1.0 (1.0-1.0)	1.0	1.0	1.0	1.0

[a]AVE: average variance extracted.

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]COPE: Coping Orientation to Problems Experienced Inventory.

[d]HC: health care.

XSL•FO
RenderX

Table 4. Performance of the measurement model: emotion-focused model.

	Bootstrapped loadings (95% CI)	Cronbach α	rho_C	AVE[a]	rho_A
GAD-7[b]		0.842	0.880	0.512	0.848
GAD1	0.755 (0.72 to 0.787)				
GAD2	0.752 (0.716 to 0.784)				
GAD3	0.726 (0.683 to 0.763)				
GAD4	0.648 (0.596 to 0.692)				
GAD5	0.719 (0.678 to 0.757)				
GAD6	0.677 (0.632 to 0.718)				
GAD7	0.723 (0.682 to 0.758)				
Cyberchondria		0.833	0.861	0.349	0.845
CYB1	0.565 (0.5 to 0.625)				
CYB2	0.64 (0.593 to 0.682)				
CYB3	0.555 (0.487 to 0.616)				
CYB4	0.695 (0.654 to 0.731)				
CYB5	0.36 (0.252 to 0.454)				
CYB6	0.65 (0.601 to 0.693)				
CYB7	0.66 (0.615 to 0.7)				
CYB8	0.686 (0.643 to 0.725)				
CYB9	0.704 (0.664 to 0.739)				
CYB10	0.574 (0.515 to 0.624)				
CYB11	0.465 (0.375 to 0.542)				
CYB12	0.385 (0.282 to 0.472)				
Brief-COPE[c]		0.764	0.718	0.381	0.732
BF5	0.182 (0.028 to 0.325)				
BF9	0.2 (0.058 to 0.329)				
BF13	0.594 (0.519 to 0.669)				
BF15	0.13 (–0.026 to 0.279)				
BF18	0.316 (0.205 to 0.415)				
BF20	–0.219 (–0.332 to–0.096)				
BF21	0.609 (0.541 to 0.666)				
BF22	0.552 (0.449 to 0.643)				
BF24	–0.28 (–0.394 to –0.142)				
BF26	0.666 (0.605 to 0.719)				
BF27	0.586 (0.508 to 0.651)				
BF28	–0.111 (–0.248 to 0.026)				
Infodemic					
IF1	1.0 (1.0 to 1.0)	1.0	1.0		0 . 1.0
Overuse of HC[d]					
OHC1	1.0 (1.0 to 1.0)	1.0	1.0		0 . 1.0

[a]AVE: average variance extracted.

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]COPE: Coping Orientation to Problems Experienced Inventory.

XSL•FO
RenderX

[d]HC: health care.

Table 5. Performance of the measurement model: avoidant-focused model.

	Bootstrapped loadings (95% CI)	Cronbach α	rho_C	AVE[a]	rho_A
GAD-7[b]		0.842	0.879	0.511	0.854
GAD1	0.750 (0.712-0.785)				
GAD2	0.763 (0.732-0.794)				
GAD3	0.714 (0.671-0.752)				
GAD4	0.643 (0.591-0.689)				
GAD5	0.716 (0.673-0.754)				
GAD6	0.666 (0.62-0.708)				
GAD7	0.740 (0.703-0.772)				
Cyberchondria		0.833	0.861	0.348	0.842
CYB1	0.566 (0.499-0.624)				
CYB2	0.652 (0.607-0.692)				
CYB3	0.552 (0.489-0.609)				
CYB4	0.684 (0.642-0.724)				
CYB5	0.351 (0.247-0.446)				
CYB6	0.644 (0.593-0.688)				
CYB7	0.661 (0.618-0.702)				
CYB8	0.686 (0.645-0.724)				
CYB9	0.692 (0.651-0.726)				
CYB10	0.597 (0.544-0.643)				
CYB11	0.460 (0.369-0.538)				
CYB12	0.384 (0.289-0.471)				
Brief-COPE[c]		0.710	0.796	0.346	0.749
B1	0.372 (0.179-0.364)				
BF3	0.558 (0.491-0.618)				
BF4	0.653 (0.595-0.699)				
BF6	0.647 (0.593-0.695)				
BF8	0.672 (0.621-0.715)				
BF11	0.690 (0.636-0.734)				
BF16	0.712 (0.669-0.751)				
BF19	0.313 (0.216-0.395)				
Infodemic					
IF1	1.0 (1.0-1.0)	1.0	1.0	1.0	1.0
Overuse of HC[d]					
OHC1	1.0 (1.0-1.0)	1.0	1.0	1.0	1.0

[a]AVE: average variance extracted.

[b]GAD-7: Generalized Anxiety Disorder.

[c]COPE: Coping Orientation to Problems Experienced Inventory.

[d]HC: health care.

XSL•FO
RenderX

Structural Models

The structural model analysis involved estimating path coefficients for the conceptual model. We performed PLS-SEM on the research model 3 times to estimate path coefficients for the models with different coping strategies. We found that H1 was supported. A significant and positive relationship was observed between a high level of infodemic exposure and increased overuse of health care services (coefficient=0.212, 95% CI 0.151-0.271). In addition, the overuse of health care services was correlated with more severe cyberchondria and higher anxiety levels in all the 3 models, supporting H2 and H3. The effect of the overuse of health care services on cyberchondria was larger than its effect on anxiety. All these relationships were statistically significant (Tables 3-5).

Moderating Effects

In our moderation analyses (Figure 2 and Tables 6 and 7), we found that emotion- and avoidant-focused coping strategies significantly moderated the relationship between the overuse of health care services and cyberchondria and that between the overuse of health care services and anxiety, respectively, supporting H5 and H6. For the problem-based model (H4), the moderating effect was not significant for the relationship between the overuse of health care services and cyberchondria (coefficient=0.002, 95% CI −0.011 to 0.006), indicating that H4.1 was not supported. Compared with the direct effects on the relationship between the overuse of health care services and cyberchondria or anxiety, a strong ability to cope with difficulties can effectively mitigate the negative effects of the infodemic-driven overuse of health care services on cyberchondria and anxiety.

XSL•FO
RenderX

Figure 2. Conceptual framework with partial least squares structural equation modeling estimates.

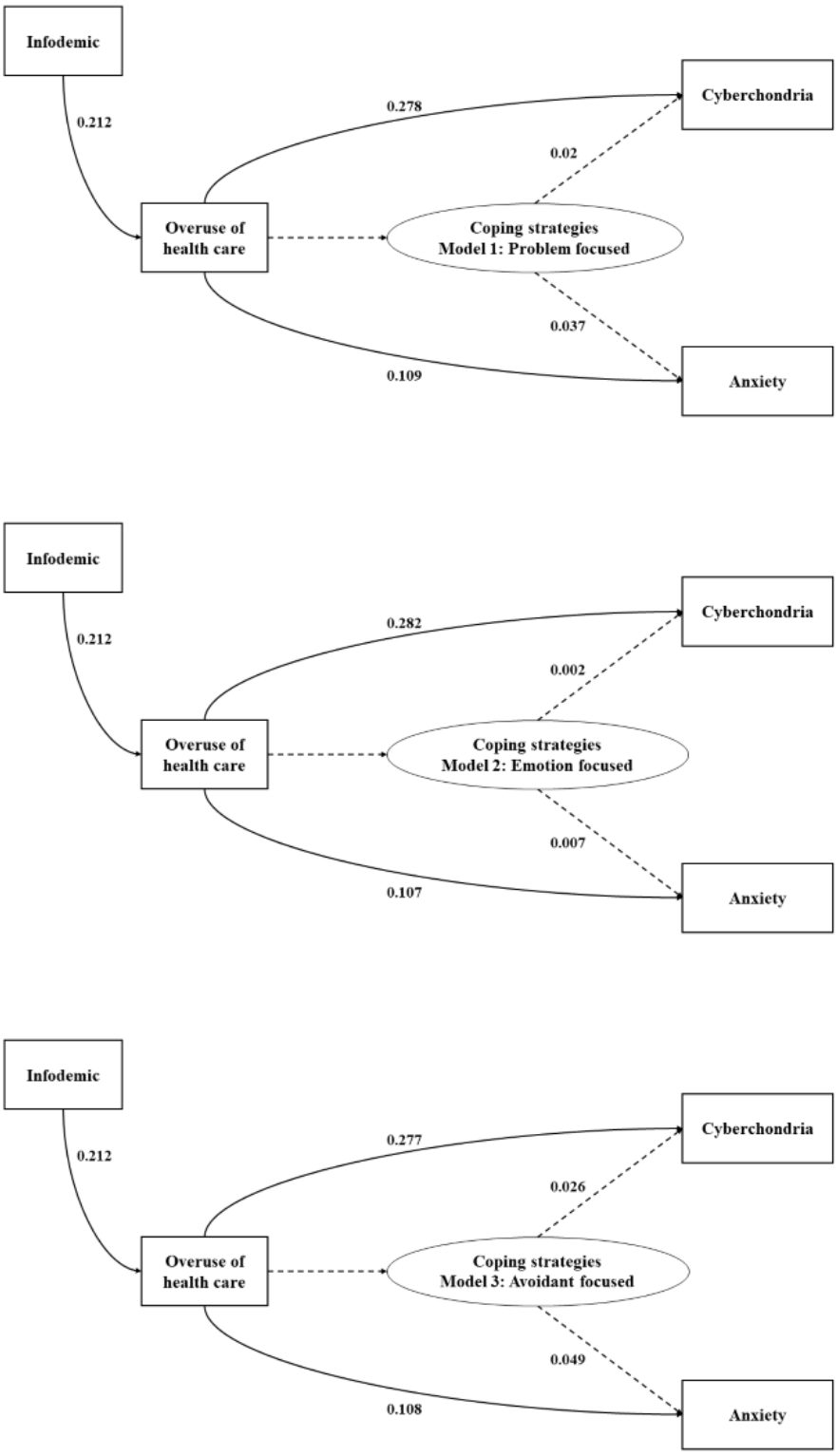

XSL•FO
RenderX

Table 6. Direct effects for the models.

Relationship	Bootstrapped mean (SD)	95% CI	Hypothesis	Responses
Infodemic→Overuse of HC[a]	0.212 (0.032)	0.151-0.271	H1	Yes
Problem-focused model				
Overuse of HC→Cyberchondria	0.282 (0.032)	0.215-0.338	H2.1	Yes
Overuse of HC→GAD[b]	0.107 (0.032)	0.044-0.171	H3.1	Yes
Emotion-focused model				
Overuse of HC→Cyberchondria	0.278 (0.031)	0.210-0.337	H2.2	Yes
Overuse of HC→GAD	0.109 (0.032)	0.046-0.172	H3.2	Yes
Avoidant-focused model				
Overuse of HC→Cyberchondria	0.277 (0.031)	0.208-0.336	H2.3	Yes
Overuse of HC→GAD	0.108 (0.032)	0.046-0.171	H3.3	Yes

[a]HC: health care.
[b]GAD: Generalized Anxiety Disorder Assessment.

Table 7. Moderating effects of the coping strategies.

Relationship	Bootstrapped mean (SD)	95% CI	Hypothesis	Responses
Problem-focused model				
Overuse of HC[a]→CS[b]→Cyberchondria	0.002 (0.004)	–0.011 to 0.006	H4.1	No
Overuse of HC→CS→GAD[c]	0.007 (0.011)	0.005 to 0.027	H5.1	Yes
Emotion-focused model				
Overuse of HC→CS→Cyberchondria	0.020 (0.008)	0.004 to 0.039	H4.2	Yes
Overuse of HC→CS→GAD	0.037 (0.015)	0.008 to 0.066	H5.2	Yes
Avoidant-focused model				
Overuse of HC→CS→Cyberchondria	0.026 (0.009)	0.010 to 0.044	H4.3	Yes
Overuse of HC→CS→GAD	0.049 (0.016)	0.019 to 0.081	H5.3	Yes

[a]HC: health care.
[b]CS: coping strategy.
[c]GAD: Generalized Anxiety Disorder Assessment.

Discussion

Principal Findings

We performed a series of PLS-SEM analyses to examine the relationships between the infodemic-driven overuse of health care services and cyberchondria and anxiety and determine the moderating effects of 3 types of coping strategies on these relationships. We observed that the individuals who were exposed to an overload of COVID-19–related information were more likely to seek and use extra and unnecessary health care services during the pandemic. Such behavior may lead to a considerable wastage of health resources that are particularly limited during a public health crisis. Although some studies have indicated that during the COVID-19 pandemic individuals with increasing mental health symptoms rarely used mental health services [33-35], we found that the overuse of health care services may contribute to higher levels of depression and cyberchondria during a pandemic. This finding has never been reported before. However, we did not differentiate between the types of health care services, either physical or mental, that the individuals overused during the pandemic. This limitation may affect the implications of our findings for policy making purposes.

Comparisons With Previous Studies

We observed that enhanced coping strategies can mitigate the adverse effects of overusing health care on depression and cyberchondria. Studies have confirmed the association between pandemics and depression, have identified several sources of depression [6,7,10,36,37], and have determined the relationship between depression and cyberchondria [38]. However, few studies have investigated the relationship between depression or cyberchondria and the infodemic-driven overuse of health care services. Our findings demonstrate that the adverse effects of the pandemic are diverse and require the investigation of individuals' health from multiple perspectives (ie, infodemic in health communication, the use of health care in health service research, and depression in psychiatry). These effects might not be immediately apparent, but they are all linked to each other

XSL•FO
RenderX

and collectively cause harm. Thus, policy makers should develop a comprehensive and cost-effective strategy to address the potential adverse effects of pandemics on people's health and well-being and better prepare for the next public health crisis.

This study offers new insights into the role of coping strategies in mediating the relationship between health care overuse and depression or cyberchondria during the COVID-19 pandemic. Overall, individuals with strong coping abilities were more likely to report lower levels of depression or cyberchondria than those with weak coping abilities. However, the moderating effects of different coping strategies varied slightly. We discovered that problem-focused coping strategies resulted in lower levels of depression and cyberchondria than avoidant-focused coping strategies. Additionally, emotion-focused coping strategies led to lower levels of depression than the other 2 types of coping strategies. These findings partially align with previous studies. For instance, Li [39] demonstrated that using both problem-focused and emotion-focused coping strategies was beneficial for psychological well-being. However, previous studies have reported mixed findings. For example, AlHadi et al [40] indicated that emotion-focused coping strategies were associated with increased depression, anxiety, and sleep disorders in the Saudi Arabian population. Few studies have examined the effect of avoidant-focused coping strategies. In this study, we found that respondents who reported living with chronic diseases exhibited a higher ability to use avoidant-focused coping. This finding is partially consistent with a previous study that found a positive relationship between avoidance-focused coping strategies and mental health in women with heart disease [41]. Individuals with medical conditions are more likely to adopt avoidant coping strategies. Firouzbakht et al [42] explained that avoidance is an effective strategy for handling short-term stress and is more likely to be adopted by certain patient groups.

We found that individuals who favor emotion-focused coping strategies to overcome difficulties are able to effectively mitigate the adverse effects of excessive health care use on depression and cyberchondria relative to those who opt for the other 2 coping strategies. This finding is not entirely surprising or unexpected. It is, in fact, quite reasonable when one considers that scholars and researchers in the field have previously indicated that people have a tendency to adopt emotion-focused strategies, especially when they find themselves in situations that are uncontrollable or unpredictable, such as the ongoing global pandemic [43]. Some studies have found that age can have a significant impact on an individual's coping strategy preferences. For instance, younger adults were more likely to use emotion-focused coping strategies during the acute phase of the SARS outbreak, whereas older adults used this particular strategy several months after the outbreak had initially occurred [44]. This suggests that the coping strategies adopted by individuals can vary greatly depending on their age and the stage of the crisis they are experiencing. However, in the context of this study, we did not observe any significant differences in the coping strategy preferences of the different age groups. This could be due to a variety of factors, but a possible explanation is that our model incorporated the COVID-19 infodemic. In this context, it is understandable that providing emotional support might be more important than providing real solutions. This is particularly true in the current digital age, where the internet offers unlimited information sources for people to explore, which can often lead to information overload and increased anxiety. Therefore, emotion-focused coping strategies could be more beneficial in helping individuals navigate the sea of information and manage their emotional responses effectively.

In this study, we used self-developed items to measure the infodemic and overuse of health care services. While this approach allowed us to collect data that were directly related to the research questions, it may have introduced some potential issues. First, self-developed items may have less validity and reliability than standardized questionnaires. This could affect the accuracy of measurements and the validity of findings. Second, using self-developed items may limit comparability with other studies that use standardized questionnaires. Standardized questionnaires allow for easy comparison across studies and populations. The lack of a common metric may make it challenging to compare the findings of this study to other studies or to aggregate them in future meta-analyses. Finally, self-developed items may be more susceptible to response bias. They may not have considered factors like social desirability bias or acquiescence bias as standardized questionnaires do. This could have skewed the responses and affected the accuracy of the findings. Despite these limitations, the study's findings provide valuable insights and pave the way for future research in this area.

Main Contributions of This Study

The importance of preparedness, prevention, and emergency response to infodemiology is highly encouraged by the WHO [45]. This study makes a significant contribution by exploring and empirically evaluating the relationship between the infodemic, the overuse of health care services, cyberchondria, and anxiety in the context of the COVID-19 pandemic. It provides empirical evidence supporting the assertion that a high level of infodemic can lead to the increased overuse of health care services, resulting in more severe cyberchondria and heightened anxiety levels. This finding adds a new dimension to our understanding of the psychological impacts of the infodemic, especially in the context of a global public health crisis. Additionally, this study highlights that adopting appropriate coping strategies can potentially reduce the severity of cyberchondria and anxiety, even among people facing high levels of the infodemic and the overuse of health care services.

Future Research

The study's findings emphasize the importance of coping strategies in reducing the negative effects of the infodemic and the excessive use of health care. Future research could focus on developing and testing interventions to improve coping skills, such as cognitive-behavioral, mindfulness-based, or psychoeducational approaches. Additionally, other factors like social support, personality traits, or health literacy may moderate the relationship between infodemic, health care overuse, cyberchondria, and anxiety. Future research could further explore these variables. This study's findings may not apply to all populations, so future research could investigate these relationships in different groups, including those with

XSL•FO
RenderX

pre-existing mental health conditions, health care professionals, or diverse cultural contexts. By pursuing these future directions, researchers could build on this study's findings, thereby enhancing our understanding of the psychological impact of infodemic and developing effective interventions.

Limitations

This study has several limitations that need to be addressed. A primary limitation is that the data were cross-sectional and self-reporting, which can introduce several biases. Social desirability bias may occur when respondents provide answers they believe are socially acceptable rather than truthful. Recall bias may also be present, as the respondents were asked to recall experiences from months or even a year ago. The data are also prone to response bias, as respondents may agree or disagree with statements regardless of their content. These biases may have affected the accuracy of the findings. In the future, we will try to collect data at multiple time points to reduce the biases and identify changes over time. Second, the data used in this analysis were obtained from a web-based survey, which excluded individuals who are not familiar with web-based surveys or do not have access to the internet. This could have resulted in selection bias. Additionally, due to the nature of the web-based survey, the demographic information of our sample was highly skewed. The majority of the respondents were young and highly educated and were frequent internet users who may have experienced more infodemic effects than older and less educated individuals. This may have affected the reliability of our findings. A quota sampling method could be used in future studies to improve the representativeness of the sample. Third, the study was conducted in China; thus, it is important to consider the unique context of China when interpreting the results. It is necessary to conduct further research in different cultural and regional contexts to determine the generalizability of the results. Finally, the evaluation of health care service overuse and strength of the infodemic relied on 2 self-developed items, which may have affected the measurement properties and limited the reliability of our findings. The development of standardized questionnaires to measure the infodemic and the overuse of health care services during a pandemic would be a valuable contribution to future research in this field.

Conclusions

This study is the first to demonstrate a significant correlation between the infodemic-driven overuse of health care services and high levels of depression and cyberchondria in the Chinese population during the COVID-19 pandemic. We find that 3 types of coping strategies can effectively mitigate the adverse effects of infodemic-driven health care overuse on depression and cyberchondria. Among them, emotion-focused coping strategies have stronger moderating effects than the other 2 types of coping strategies. These findings provide empirical evidence that can guide policy makers in developing strategies to reduce cyberchondria, provide accurate information about public health crises, and promote adaptive coping strategies to effectively manage future public health crises.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RHX contributed to developing the study concept and design, data analysis and interpretation, software, writing the original draft, and review and editing. CC contributed to data collection, software, and review and editing. Both authors approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
English-translated questionnaire.
[DOCX File , 29 KB - jmir_v26i1e53417_app1.docx]

Multimedia Appendix 2
Correlations between measures.
[DOCX File , 16 KB - jmir_v26i1e53417_app2.docx]

References

1. Joseph AM, Fernandez V, Kritzman S, Eaddy I, Cook OM, Lambros S, et al. COVID-19 misinformation on social media: a scoping review. Cureus 2022;14(4):e24601 [FREE Full text] [doi: 10.7759/cureus.24601] [Medline: 35664409]
2. Farsi D, Martinez-Menchaca HR, Ahmed M, Farsi N. Social media and health care (Part II): narrative review of social media use by patients. J Med Internet Res 2022;24(1):e30379 [FREE Full text] [doi: 10.2196/30379] [Medline: 34994706]
3. Health topic - Infodemic. World Health Organization. 2023. URL: https://www.who.int/health-topics/infodemic#tab=tab_1 [accessed 2023-05-14]

XSL•FO
RenderX

4. Gogol M, Siebenhofer A. [Choosing wisely—against overuse in healthcare systems—activities in Germany and Austria in geriatric medicine]. Wien Med Wochenschr 2016;166(5-6):155-160 [FREE Full text] [doi: 10.1007/s10354-015-0424-z] [Medline: 26811241]
5. Morgan DJ, Leppin AL, Smith CD, Korenstein D. A practical framework for understanding and reducing medical overuse: conceptualizing overuse through the patient-clinician interaction. J Hosp Med 2017;12(5):346-351 [FREE Full text] [doi: 10.12788/jhm.2738] [Medline: 28459906]
6. Alnazly E, Khraisat OM, Al-Bashaireh AM, Bryant CL. Anxiety, depression, stress, fear and social support during COVID-19 pandemic among Jordanian healthcare workers. PLoS One 2021;16(3):e0247679 [FREE Full text] [doi: 10.1371/journal.pone.0247679] [Medline: 33711026]
7. Choi EPH, Hui BPH, Wan EYF. Depression and anxiety in Hong Kong during COVID-19. Int J Environ Res Public Health 2020;17(10):3740 [FREE Full text] [doi: 10.3390/ijerph17103740] [Medline: 32466251]
8. Tomasoni D, Bai F, Castoldi R, Barbanotti D, Falcinella C, Mulè G, et al. Anxiety and depression symptoms after virological clearance of COVID-19: a cross-sectional study in Milan, Italy. J Med Virol 2021;93(2):1175-1179 [FREE Full text] [doi: 10.1002/jmv.26459] [Medline: 32841387]
9. Tran HTT, Nguyen MH, Pham TTM, Kim GB, Nguyen HT, Nguyen NM, et al. Predictors of eHealth literacy and its associations with preventive behaviors, fear of COVID-19, anxiety, and depression among undergraduate nursing students: a cross-sectional survey. Int J Environ Res Public Health 2022;19(7):3766 [FREE Full text] [doi: 10.3390/ijerph19073766] [Medline: 35409448]
10. Bareeqa SB, Ahmed SI, Samar SS, Yasin W, Zehra S, Monese GM, et al. Prevalence of depression, anxiety and stress in China during COVID-19 pandemic: a systematic review with meta-analysis. Int J Psychiatry Med 2021;56(4):210-227 [FREE Full text] [doi: 10.1177/0091217420978005] [Medline: 33243029]
11. Starcevic V, Berle D. Cyberchondria: towards a better understanding of excessive health-related internet use. Expert Rev Neurother 2013;13(2):205-213 [FREE Full text] [doi: 10.1586/ern.12.162] [Medline: 23368807]
12. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020;26(5):309-313 [FREE Full text] [doi: 10.1177/1357633X20916567] [Medline: 32196391]
13. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nat Hum Behav 2020;4(12):1285-1293 [FREE Full text] [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
14. Yang F. Coping strategies, cyberbullying behaviors, and depression among Chinese netizens during the COVID-19 pandemic: a web-based nationwide survey. J Affect Disord 2021;281:138-144 [FREE Full text] [doi: 10.1016/j.jad.2020.12.023] [Medline: 33321378]
15. Fluharty M, Bu F, Steptoe A, Fancourt D. Coping strategies and mental health trajectories during the first 21 weeks of COVID-19 lockdown in the United Kingdom. Soc Sci Med 2021;279:113958 [FREE Full text] [doi: 10.1016/j.socscimed.2021.113958] [Medline: 33965772]
16. Shigeto A, Laxman DJ, Landy JF, Scheier LM. Typologies of coping in young adults in the context of the COVID-19 pandemic. J Gen Psychol 2021;148(3):272-304 [FREE Full text] [doi: 10.1080/00221309.2021.1874864] [Medline: 33475048]
17. Hang CN, Yu PD, Chen S, Tan CW, Chen G. MEGA: Machine Learning-Enhanced Graph Analytics for infodemic risk management. IEEE J Biomed Health Inform 2023;27(12):6100-6111 [FREE Full text] [doi: 10.1109/JBHI.2023.3314632] [Medline: 37713230]
18. Lazarus RS. Coping theory and research: past, present, and future. Psychosom Med 1993;55(3):234-247 [FREE Full text] [doi: 10.1097/00006842-199305000-00002] [Medline: 8346332]
19. Wechsler B. Coping and coping strategies: a behavioural view. Appl Anim Behav Sci 1995;43(2):123-134 [FREE Full text] [doi: 10.1016/0168-1591(95)00557-9]
20. Hair JF, Hair J, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, CA: SAGE; 2022.
21. McElroy E, Kearney M, Touhey J, Evans J, Cooke Y, Shevlin M. The CSS-12: development and validation of a short-form version of the cyberchondria severity scale. Cyberpsychol Behav Soc Netw 2019;22(5):330-335 [FREE Full text] [doi: 10.1089/cyber.2018.0624] [Medline: 31013440]
22. Peng XQ, Chen Y, Zhang YC, Liu F, He HY, Luo T, et al. The status and influencing factors of cyberchondria during the COVID-19 epidemic. a cross-sectional study in Nanyang City of China. Front Psychol 2021;12:712703 [FREE Full text] [doi: 10.3389/fpsyg.2021.712703] [Medline: 34858254]
23. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
24. Sun J, Liang K, Chi X, Chen S. Psychometric properties of the Generalized Anxiety Disorder Scale-7 Item (GAD-7) in a large sample of Chinese Adolescents. Healthcare (Basel) 2021;9(12):1709 [FREE Full text] [doi: 10.3390/healthcare9121709] [Medline: 34946435]
25. Carver CS. You want to measure coping but your protocol's too long: consider the brief COPE. Int J Behav Med 1997;4(1):92-100 [FREE Full text] [doi: 10.1207/s15327558ijbm0401_6] [Medline: 16250744]

XSL•FO
RenderX

26. Poulus D, Coulter TJ, Trotter MG, Polman R. Stress and coping in Esports and the influence of mental toughness. Front Psychol 2020;11:628 [FREE Full text] [doi: 10.3389/fpsyg.2020.00628] [Medline: 32390900]
27. Wang AWT, Cheng CP, Chang CS, Chen DR, Chen ST, Shieh V, et al. Does the factor structure of the brief COPE fit different types of traumatic events? Eur J Psychol Assess 2018;34(3):162-173 [FREE Full text] [doi: 10.1027/1015-5759/a000321]
28. Brown TA. Confirmatory Factor Analysis for Applied Research, Second Edition. New York, NY: Guilford Publications; 2015.
29. Nunnally JC. Psychometric Theory, 3rd Edition. New York, NY: Tata McGraw-Hill Education; 1994.
30. Hair JF, Sarstedt M, Pieper TM, Ringle CM. The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. Long Range Plann 2012;45(5-6):320-340 [FREE Full text] [doi: 10.1016/j.lrp.2012.09.008]
31. Streukens S, Leroi-Werelds S. Bootstrapping and PLS-SEM: a step-by-step guide to get more out of your bootstrap results. Eur Manag J 2016;34(6):618-632 [FREE Full text] [doi: 10.1016/j.emj.2016.06.003]
32. Hair JF, Hult GTM, Ringle CM, Sarstedt M, Danks NP, Ray S. Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook. Cham, Switzerland: Springer International Publishing; 2021.
33. Liberman JN, Bhattacharjee S, Rui P, Ruetsch C, Rothman B, Kulkarni A, et al. Impact of the COVID-19 pandemic on healthcare resource utilization in individuals with major depressive disorder. Health Serv Res Manag Epidemiol 2022;9:23333928221111864 [FREE Full text] [doi: 10.1177/23333928221111864] [Medline: 35832488]
34. Saunders R, Buckman JEJ, Leibowitz J, Cape J, Pilling S. Trends in depression and anxiety symptom severity among mental health service attendees during the COVID-19 pandemic. J Affect Disord 2021;289:105-109 [FREE Full text] [doi: 10.1016/j.jad.2021.04.020] [Medline: 33962366]
35. Lee J, Jeong HJ, Kim S. Stress, anxiety, and depression among undergraduate students during the COVID-19 pandemic and their use of mental health services. Innov High Educ 2021;46(5):519-538 [FREE Full text] [doi: 10.1007/s10755-021-09552-y] [Medline: 33907351]
36. Luceño-Moreno L, Talavera-Velasco B, García-Albuerne Y, Martín-García J. Symptoms of posttraumatic stress, anxiety, depression, levels of resilience and burnout in Spanish health personnel during the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(15):5514 [FREE Full text] [doi: 10.3390/ijerph17155514] [Medline: 32751624]
37. Chen J, Liu X, Wang D, Jin Y, He M, Ma Y, et al. Risk factors for depression and anxiety in healthcare workers deployed during the COVID-19 outbreak in China. Soc Psychiatry Psychiatr Epidemiol 2021;56(1):47-55 [FREE Full text] [doi: 10.1007/s00127-020-01954-1] [Medline: 32914298]
38. Jungmann SM, Witthöft M. Health anxiety, cyberchondria, and coping in the current COVID-19 pandemic: which factors are related to coronavirus anxiety? J Anxiety Disord 2020;73:102239 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102239] [Medline: 32502806]
39. Li Q. Psychosocial and coping responses toward 2019 coronavirus diseases (COVID-19): a cross-sectional study within the Chinese general population. QJM 2020;113(10):731-738 [FREE Full text] [doi: 10.1093/qjmed/hcaa226] [Medline: 32678901]
40. AlHadi AN, Alarabi MA, AlMansoor KM. Mental health and its association with coping strategies and intolerance of uncertainty during the COVID-19 pandemic among the general population in Saudi Arabia: cross-sectional study. BMC Psychiatry 2021;21(1):382 [FREE Full text] [doi: 10.1186/s12888-021-03370-4] [Medline: 34320930]
41. Sidhu A, Arora AK. A study of mental health and coping among women with coronary heart disease and women without coronary heart disease. Int J Adv Res Dev 2017;2(6):1-3 [FREE Full text]
42. Firouzbakht M, Rahmani N, Nia HS, Omidvar S. Coping strategies and depression during the COVID-19 pandemic in pregnant women: a cross sectional study. BMC Psychiatry 2022;22(1):153 [FREE Full text] [doi: 10.1186/s12888-022-03792-8] [Medline: 35232424]
43. Chew QH, Wei KC, Vasoo S, Chua HC, Sim K. Narrative synthesis of psychological and coping responses towards emerging infectious disease outbreaks in the general population: practical considerations for the COVID-19 pandemic. Singapore Med J 2020;61(7):350-356 [FREE Full text] [doi: 10.11622/smedj.2020046] [Medline: 32241071]
44. Yeung DYL, Fung HH. Age differences in coping and emotional responses toward SARS: a longitudinal study of Hong Kong Chinese. Aging Ment Health 2007;11(5):579-587 [FREE Full text] [doi: 10.1080/13607860601086355] [Medline: 17882596]
45. Wilhelm E, Ballalai I, Belanger ME, Benjamin P, Bertrand-Ferrandis C, Bezbaruah S, et al. Measuring the burden of infodemics: summary of the methods and results of the Fifth WHO Infodemic Management Conference. JMIR Infodemiology 2023;3:e44207 [FREE Full text] [doi: 10.2196/44207] [Medline: 37012998]

Abbreviations

Brief-COPE: Coping Orientation to Problems Experienced Inventory
CSS-12: Cyberchondria Severity Scale-12
GAD-7: Generalized Anxiety Disorder Assessment-7

XSL•FO
RenderX

PLS-SEM: partial least squares structural equation modeling
WHO: World Health Organization

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 05.10.23; peer-reviewed by K Wang, J Chen, CN Hang, E Vashishtha, D Liu; comments to author 06.11.23; revised version received 14.11.23; accepted 22.03.24; published 09.04.24.

Please cite as:
Xu RH, Chen C
Moderating Effect of Coping Strategies on the Association Between the Infodemic-Driven Overuse of Health Care Services and Cyberchondria and Anxiety: Partial Least Squares Structural Equation Modeling Study
J Med Internet Res 2024;26:e53417
URL: https://www.jmir.org/2024/1/e53417
doi:10.2196/53417
PMID:38593427

©Richard Huan Xu, Caiyun Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Empowering School Staff to Support Pupil Mental Health Through a Brief, Interactive Web-Based Training Program: Mixed Methods Study

Emma Soneson[1,2], PhD; Emma Howarth[3], PhD; Alison Weir[4,5], MA, MSc; Peter B Jones[2*], PhD; Mina Fazel[1*], DM

[1]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[2]Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom
[3]School of Psychology, University of East London, London, United Kingdom
[4]Faculty of Education, University of Cambridge, Cambridge, United Kingdom
[5]Howard Community Academy, Anglian Learning multi-academy trust, Bury St Edmunds, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Emma Soneson, PhD
Department of Psychiatry
University of Oxford
Warneford Lane
Oxford, OX3 7JX
United Kingdom
Phone: 44 1865 613127
Email: emma.soneson@psych.ox.ac.uk

Abstract

Background: Schools in the United Kingdom and elsewhere are expected to protect and promote pupil mental health. However, many school staff members do not feel confident in identifying and responding to pupil mental health difficulties and report wanting additional training in this area.

Objective: We aimed to explore the feasibility of Kognito's *At-Risk for Elementary School Educators*, a brief, interactive web-based training program that uses a simulation-based approach to improve school staff's knowledge and skills in supporting pupil mental health.

Methods: We conducted a mixed methods, nonrandomized feasibility study of *At-Risk for Elementary School Educators* in 6 UK primary schools. Our outcomes were (1) school staff's self-efficacy and preparedness to identify and respond to pupil mental health difficulties, (2) school staff's identification of mental health difficulties and increased risk of mental health difficulties, (3) mental health support for identified pupils (including conversations about concerns, documentation of concerns, in-class and in-school support, and referral and access to specialist mental health services), and (4) the acceptability and practicality of the training. We assessed these outcomes using a series of questionnaires completed at baseline (T1), 1 week after the training (T2), and 3 months after the training (T3), as well as semistructured qualitative interviews. Following guidance for feasibility studies, we assessed quantitative outcomes across time points by comparing medians and IQRs and analyzed qualitative data using reflexive thematic analysis.

Results: A total of 108 teachers and teaching assistants (TAs) completed T1 questionnaires, 89 (82.4%) completed T2 questionnaires, and 70 (64.8%) completed T3 questionnaires; 54 (50%) completed all 3. Eight school staff members, including teachers, TAs, mental health leads, and senior leaders, participated in the interviews. School staff reported greater confidence and preparedness in identifying and responding to mental health difficulties after completing the training. The proportion of pupils whom they identified as having mental health difficulties or increased risk declined slightly over time ($median_{T1}$=10%; $median_{T2}$=10%; $median_{T3}$=7.4%), but findings suggested a slight increase in accuracy compared with a validated screening measure (the Strengths and Difficulties Questionnaire). In-school mental health support outcomes for identified pupils improved after the training, with increases in formal documentation and communication of concerns as well as provision of in-class and in-school support. Referrals and access to external mental health services remained constant. The qualitative findings indicated that school staff perceived the training as useful, practical, and acceptable.

XSL•FO
RenderX

Conclusions: The findings suggest that brief, interactive web-based training programs such as *At-Risk for Elementary School Educators* are a feasible means to improve the identification of and response to mental health difficulties in UK primary schools. Such training may help address the high prevalence of mental health difficulties in this age group by helping facilitate access to care and support.

(J Med Internet Res 2024;26:e46764) doi:10.2196/46764

KEYWORDS

mental health; children; schools; teachers; training; digital intervention; pupil mental health; mental health training; intervention; empowerment; student; pupil; support; school staff; web-based training

Introduction

Background

In recent years, there has been an increased emphasis on the role of schools in supporting children's mental health [1-3]. This enhanced focus has been driven in large part by an apparent increase in mental health difficulties (including behavioral, social, and emotional difficulties) present in school-aged populations [4-6]—a concern that became increasingly prominent in the context of the COVID-19 pandemic and the associated school closures and social distancing measures [7,8]. There is also a growing recognition of the many unique advantages of using the school setting to promote and protect pupil mental health [9]. First, most lifetime disorders begin during the schooling years [10], which suggests that schools may be an ideal setting for early identification and intervention. Second, schools have access to most children, meaning that they are an important component of any public health approach to address child mental health difficulties [11-14]. Third, schools benefit from prolonged engagement with pupils, which can facilitate the implementation of mental health promotion and prevention strategies as well as support and interventions for pupils with identified mental health needs [12]. Finally, mental health support in schools is often more accessible to families than other types of support [15].

However, while school staff are increasingly expected to support children's mental health [1], many do not feel prepared to do so [16-19] due in part to receiving limited training and supervision in this area [20]. Therefore, improving school staff's confidence and preparedness are important considerations for supporting them in taking an expanded role in pupil mental health [21]. Most schools offer some form of mental health training [22,23], but many staff members believe that they could benefit from additional training [18-20,24-26]. One area where staff training may be particularly beneficial is the identification of and first response to pupils who have mental health difficulties or who are believed to be at increased risk of developing them. However, although there is evidence suggesting that school staff, parents, and practitioners see such training as an acceptable, feasible, and potentially useful way to support pupil mental health [20,27-29], empirical evidence for the effectiveness of such training is limited and focuses primarily on intermediate outcomes (eg, staff knowledge and confidence) rather than downstream outcomes (eg, accurate identification, access to support, and mental health outcomes) [30,31]. Furthermore, there are several potential barriers to implementing training programs in schools, including time, cost, and resource requirements [28].

At-Risk for Elementary School Educators: A Brief, Interactive Web-Based Training

Training programs that address these barriers may be beneficial for supporting schools in identifying and responding to pupil mental health difficulties. Brief, interactive web-based training programs are a particularly promising avenue as they have the potential to be more affordable, flexible, and scalable than other training formats. One such training is *At-Risk for Elementary School Educators* (hereinafter, *At-Risk*), a virtual simulation-based program developed by the American company Kognito [32]. The program, which has been completed by >125,000 teachers in the United States, aims to improve pupil mental health by "[building] awareness, knowledge, and skills about mental health, and [preparing] users to lead real-life conversations with pupils, parents, and caregivers about their concerns and available support" [33].

The program addresses many common implementation barriers to school-based mental health training. For example, *At-Risk* only requires approximately 1 hour to complete, which is much shorter than many other available training programs [31,34]. This comparatively low time commitment may address the concern that training programs are overly time intensive and, thus, make the training more feasible for busy schools [28,34,35]. The web-based format of *At-Risk* may also address concerns about school-based mental health programs being resource intensive [28]. Nearly all school mental health training programs documented in the literature are face-to-face sessions led by external facilitators [34,36], with only a few examples of web-based training [37-39]. For schools with limited budgets, programs requiring external facilitators can prove unsustainable and have limited scalability. In terms of financial resources, the costs of *At-Risk* vary depending on the number of licenses purchased, but the maximum cost is approximately £22 (US $30) per user, a price point that is feasible for many UK schools. In the United States, there have been many examples of bulk purchases at the district or state level that have made the training even more affordable per teacher. In many areas, the training is even free at point of use due to state- or district-wide licensing agreements [40].

To date, 3 randomized studies have examined the effectiveness and acceptability of *At-Risk* among samples of American teachers [17,41] and teachers in training [42] across school years. Each study found high satisfaction ratings, with between 75% and 85% of participants rating the training as useful, well

XSL•FO
RenderX

constructed, relevant, and easy to use, and nearly all (88%-95%) reporting that they would recommend it to colleagues. The training also improved teachers' self-rated preparedness, self-efficacy, and likelihood of identifying and discussing concerns about pupils' mental health and referring them to appropriate support when needed. These improvements were reflected in the teachers' behaviors—compared with teachers in the control group, those who completed *At-Risk* self-reported significantly more helping behaviors (eg, identifying psychological distress, discussing concerns with pupils and parents, and consulting with parents about options for care and support) and gatekeeping behaviors (ie, connecting pupils with care and support) after the training and at 3 months after the training. The findings of these studies indicate that *At-Risk* may help improve teachers' ability to identify and respond to pupil mental health needs and lead to positive behavior change in terms of discussing concerns and facilitating access to care and support.

At-Risk in a UK Context: Considerations for Transportability

These 3 studies suggest that *At-Risk* may be a promising intervention for improving children's mental health; however, there is still much to be learned about the training's effectiveness, feasibility, and acceptability. Furthermore, to date, no evaluation of the training has been conducted outside the United States. There is increasing focus on the influence of context on the effectiveness of complex interventions [43-48], and while some interventions have shown success in terms of transportability [48], other interventions that have evidence of effectiveness in one context have demonstrated null or even negative effects in another [46]. Furthermore, information that could inform "transportability" is often not collected as part of evaluations [44], making it difficult to determine the likelihood of success in a new setting.

There are many contextual differences between the United States and the United Kingdom that could mean that school-based interventions developed in one country may not translate well to the other. Cross-country differences in education systems and (mental) health services are particularly relevant to this study. Differences in the education system include the length and content of initial teacher training, the number and roles of teaching assistants (TAs), and school funding structures. There are also key differences in the structure and availability of school-based mental health provision. In the United States, schools often have staff whose sole or at least main responsibility is mental health, such as school psychologists. While these roles are becoming more common in the United Kingdom with the implementation of the Green Paper recommendations [1], in most UK primary schools, mental health is included within the broader roles of the special educational needs coordinator (SENCo) and pastoral team. Finally, differences in the wider health care systems across the countries also mean that the process and outcomes of external referrals to specialist mental health services vary across settings, another fact that may influence the transportability of school-based interventions such as *At-Risk*.

This Study

Given these uncertainties regarding intervention transportability, additional evaluation of *At-Risk* is needed to understand whether it is a potentially useful and feasible tool to improve the identification of and response to mental health difficulties in UK primary schools. To explore the potential value of the training in this new context, we conducted a mixed methods feasibility study of *At-Risk* in 6 UK primary schools covering pupils aged 4 to 11 years. We aimed to examine the influence of *At-Risk* on staff confidence and preparedness, identification of pupils with mental health difficulties or increased risk of developing mental health difficulties, mental health support outcomes for identified children, and intervention acceptability and practicality.

Methods

Study Design

We used a mixed methods, nonrandomized, pretest-posttest study design to explore the feasibility of *At-Risk* in UK primary schools. While feasibility studies are acknowledged as a key stage of intervention design and evaluation [49,50], there is no universally agreed-upon definition of a feasibility study [50,51]. Therefore, we focused on 3 criteria from the guidance by Bowen et al [52]: acceptability, practicality, and limited effectiveness testing.

Intervention: *At-Risk for Elementary School Educators*

At-Risk is a web-based training that is delivered individually and requires only a log-in and internet connection. Using a simulation-based teaching model, the training aims to (1) improve mental health awareness and knowledge, (2) empower users to approach pupils about what they have noticed, (3) impart skills to have meaningful conversations with pupils and parents, and (4) train users to refer pupils to further support. The diagram in Figure 1 illustrates how the training might lead to improved mental health outcomes for pupils.

The simulation begins with an introduction by a virtual coach, who defines and explains how to recognize the warning signs of psychological distress and specific mental health difficulties and provides guidance and practical advice for discussing and acting upon concerns. Users then practice 2 virtual scenarios. The first scenario involves a fifth-grade (UK Year 6; ages of 10-11 years) teacher speaking with the parent of a pupil showing signs of behavioral difficulties. The second involves a third-grade (UK Year 4; ages of 8-9 years) teacher speaking with a pupil showing signs of emotional difficulties. During the conversations, users choose what to say via drop-down menus organized into categories (eg, "bring up concerns" or "ask a question") and phrases (eg, "Mia sometimes seems a little agitated in class"). Throughout the conversation, users receive feedback through a "comfort bar" (based on how the pupil or parent perceives the conversation), opportunities to "see" the thoughts of the pupil or parent, and suggestions from the virtual coach.

Importantly, there is no one "right" way to conduct the conversations, and several approaches can lead to a positive outcome. Throughout the conversation, users can "undo" actions

XSL•FO
RenderX

to backtrack after receiving an undesirable response or to explore what the response would have been had they chosen another option. At the end of each conversation, the pupil or parent provides feedback on the conversation. The training finishes with a short segment on connecting pupils with further support.

For this feasibility study, we used an unmodified version of the training (ie, the standard training designed for American schools, not tailored to the UK context) provided free of cost by Kognito. The potential need for adaptation and tailoring was an important consideration that we explicitly examined as part of our exploration of the acceptability and practicality of *At-Risk* in this new setting.

Figure 1. Theory of how the *At-Risk for Elementary School Educators* training may improve school staff's recognition of and response to pupil mental health difficulties.

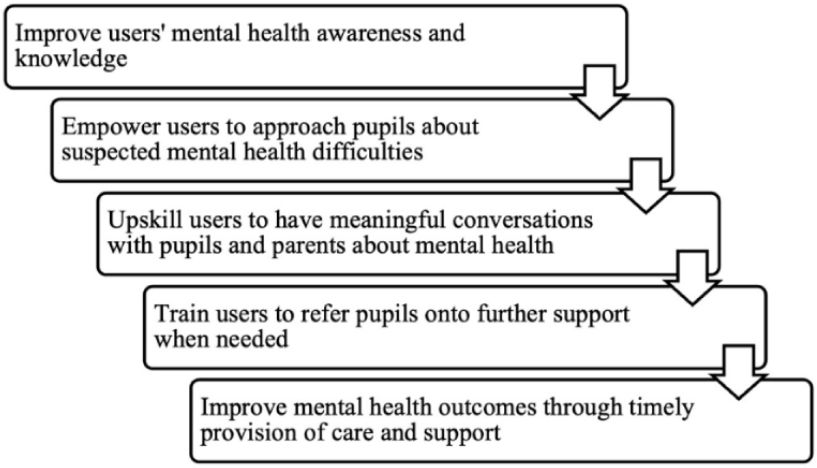

Recruitment

Schools

We originally sought to purposively sample 5 primary schools from Cambridgeshire or Norfolk that (1) had a higher-than-average proportion of pupils eligible for free school meals or (2) were located in an area in the top tertile of deprivation as measured using the Index of Multiple Deprivation [53], which we calculated with the publicly available Schools, pupils and their characteristics data [54]. We emailed headteachers, SENCos, and mental health leads from 131 candidate schools in September and October 2019 about participating in the study. To increase recruitment, we contacted additional schools in January 2020 for a study start date of March 2020. However, the study was suspended in March 2020 due to the in-person school closures associated with the onset of the COVID-19 pandemic. As some of the participating schools dropped out due to the pandemic, we reopened recruitment for a January 2021 study start date. In this round, we did not restrict participation by the 2 deprivation criteria described previously (ie, free school meal eligibility and Index of Multiple Deprivation), so any UK-based mainstream primary school was eligible to participate. The January 2021 start date was again delayed by the pandemic, but there was no subsequent recruitment.

Teachers and TAs

Schools were responsible for recruiting individual teachers and TAs to participate in the training. We encouraged schools to invite all teachers and TAs to participate, but schools made a variety of decisions in this regard. Three schools (schools D, E, and F) had all staff complete the training during inset (in-service training) days or other designated times, 2 schools (schools A and C) had staff volunteer to participate, and 1 school (school B) selected 2 to 3 staff members in each year group to participate.

Measures and Materials

School Characteristics

The characteristics of the participating schools, including school type, school sex (ie, whether they were single or mixed sex), urbanicity, head count, area-level deprivation, level of free school meal eligibility, ethnic composition, and proportion of pupils with special educational needs, were obtained from publicly available data from the Department for Education [54,55].

Teacher and TA Identification Form

The purpose of the Teacher and TA Identification Form (Multimedia Appendix 1) was to understand which pupils participants would identify as having mental health difficulties or an increased risk of developing mental health difficulties. As systematic reviews in this area have identified no suitable questionnaires [28,30], we developed a bespoke questionnaire, which was reviewed by a school staff advisory group to ensure accuracy and relevance. The questionnaire begins with instructions, including explanations and examples of what is meant by "mental health difficulties or risk for mental health difficulties." Full definitions are provided in Multimedia Appendix 1, but in brief, "mental health difficulties" are described as "behavioural and social-emotional problems" regardless of formal diagnosis, and "risk for mental health difficulties" is described as experiences that increase the chance of a child developing mental health difficulties in the future.

XSL•FO
RenderX

For all pupils in their class, participants first indicated whether they thought a pupil had mental health difficulties or increased risk. If yes, they answered 9 subsequent questions about mental health support outcomes. The first four outcomes were about communication of concerns, namely whether they had (1) formally documented their concerns with the school, (2) communicated concerns to the SENCo, pastoral care lead, or mental health lead, (3) communicated concerns to another member of the school staff, or (4) communicated concerns to the child or their parents. The next five outcomes pertained to the provision of mental health support, namely whether the pupil (5) received in-class support; (6) received in-school support or had an in-house support plan; (7) had documented social, emotional, and mental health (SEMH) status (a type of special educational need focused on mental health difficulties); (8) had been referred to external mental health services; or (9) had access to external mental health services.

Strengths and Difficulties Questionnaire

The teacher-report Strengths and Difficulties Questionnaire (SDQ) [56-59] served as the comparator for findings about teachers' and TAs' identification of pupils. The SDQ includes 25 positive and negative psychological attributes across 5 scales: emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behavior. The first 4 scales add up to a Total Difficulties Score (0-40, with higher scores representing greater difficulties). The SDQ has demonstrated acceptable psychometric properties in primary school samples [60]. It is important to note that the SDQ is not an exact comparator as it measures a narrower concept than the Teacher and TA Identification Form (which also includes *increased risk*). However, this comparison could potentially yield valuable information regarding feasibility.

Pre- and Posttraining Surveys

Kognito uses pre- and posttraining surveys to assess their training. These surveys (based on the validated Gatekeeper Behavior Scale [61]) explore teachers' self-efficacy in identifying and responding to mental health difficulties and whether their attitudes, self-efficacy, or practice have changed since completing the *At-Risk* training. The posttraining survey also includes questions on perceptions of the training's impact. We independently (ie, with no input from Kognito) reviewed the merits of these questionnaires and decided to use them in this study because (1) they covered relevant and useful concepts related to our aims and (2) using them increased comparability to the other 3 US-based studies of *At-Risk*. We slightly adapted the surveys to make them more relevant to the UK context (Multimedia Appendix 2).

Interview Schedules

For the pretraining interviews with SENCos and mental health leads, we developed a topic guide about current practice (Multimedia Appendix 3) with the specific purpose of creating Mental Health Resource Maps for each school (refer to the *Procedures* section). The main topics pertained to formal and informal procedures for when staff members suspect that a child might have mental health difficulties or increased risk, as well as the types of support available.

For the posttraining interviews with teachers, TAs, and strategic stakeholders (ie, those with key leadership roles, including senior leadership teams [SLTs], school governors, and SENCos and mental health leads), we developed 3 separate topic guides (Multimedia Appendix 3), which were informed by our research questions, systematic reviews [28,30], and the Consolidated Framework for Implementation Research [62,63]. For teachers and TAs who completed *At-Risk* and strategic stakeholders, interview topics included the acceptability of the training, the practicality of implementing it in schools, the utility of further refinement and testing, possible harms associated with the training (if any), and suggestions for adaptations. For teachers and TAs who did not complete the training, topics included reasons for not completing it, barriers to acceptability and practicality, and suggestions for adaptations.

Procedures

Interviews With SENCos and Mental Health Leads

We conducted a pretraining interview with each school's SENCo or mental health lead to develop a "Mental Health Resource Map" with information on referral processes and available support. These maps served an ethical purpose by ensuring that pupils identified as potentially having mental health difficulties would have the best possible chance of being linked to care and support.

Completing At-Risk

Schools' timelines for the study varied due to the pandemic and other commitments. School D completed the training in December 2020; school E completed the training in March 2021; schools B, C, and F completed the training in May 2021; and school A completed the training in June 2021. At baseline (T1), participants completed a Teacher and TA Identification Form and the pretraining survey. They then completed the *At-Risk* training. We encouraged schools to designate specific time for the training, which 3 schools (schools D, E, and F) did. One week after training (T2), participants were asked to complete a second Teacher and TA Identification Form and the posttraining survey. Three months after the training or at the end of the school year (whichever came first; T3), participants completed a third Teacher and TA Identification Form as well as SDQs for all pupils. All questionnaires were completed on the University of Cambridge Qualtrics platform (Qualtrics International Inc).

Feedback Provision

After T2, we provided all SENCos and mental health leads but not teachers or TAs with feedback regarding which children had been identified as having mental health difficulties or increased risk. After T3, we provided SDQ scores for each child as well as whole-class distributions (where available). This feedback was provided to ensure the ethical conduct of the study.

XSL•FO
RenderX

Interviews With Teachers, TAs, and Strategic Stakeholders

We aimed to recruit at least 3 teachers or TAs who completed the training per school, 3 to 5 teachers or TAs who had not completed the training across all schools, and up to 3 strategic stakeholders per school for posttraining semistructured interviews. Schools contacted staff members directly with an invitation to complete a virtual interview.

Analysis

Quantitative Outcomes

Analytical Samples

For the main analysis, participants were included if they (1) completed at least the pretraining (T1) questionnaires and the training itself and (2) had what we judged to be a typical number of children they regularly worked with. For the latter criterion, given that the average UK primary school class size is approximately 27 to 28 pupils [64], we excluded teachers and TAs who worked with <10 children (as we suspected this would not be a random selection of pupils and would therefore influence aggregate identification rates) and those who worked with >60 children (as we believed that it would be difficult for a teacher or TA to know >2 classes' worth of children well enough to make accurate judgments about their mental health).

Teacher and TA Self-Efficacy and Preparedness

To assess teachers' and TAs' preparedness, self-efficacy, and perceptions of training impact, we calculated the absolute and relative frequencies of responses to the pre- and posttraining surveys. Participants were eligible for inclusion in this analysis only if they had pretraining (T1) data.

Identification Outcomes

On the basis of the Teacher and TA Identification Forms, we calculated the number and percentage of pupils in each class whom teachers and TAs perceived as having mental health difficulties or increased risk at each time point. We summarized these across all participants using medians and IQRs.

We then calculated SDQ scores, which we compared with responses from the Teacher and TA Identification Form by calculating (1) the median and IQR for the percentage of children identified by participants who *did not* have elevated SDQ scores and (2) the median and IQR for the percentage of children with elevated SDQ scores who *were not identified by* participants. To be included in these analyses, participants had to have completed all 3 time points. For the first outcome, they had to have completed an SDQ for *all children they identified in the Teacher and TA Identification Form*. For the second outcome, they had to have completed SDQs for at least 80% of their class. Where it was possible to match pupil IDs between teachers and TAs, we pooled SDQ data such that, if one participant did not meet the inclusion criteria themselves, they could still be included if the SDQ data were available from another staff member working with the same children.

Mental Health Support Outcomes

Finally, for each time point, we calculated medians and IQRs for the proportion of identified children with each of the 9 mental health support outcomes (refer to the *Teacher and TA Identification Form* section for the outcomes).

Sensitivity Analyses

We also conducted 2 post hoc sensitivity analyses. The first sensitivity analysis excluded all participants from school D. When we prepared feedback for school D (the first school to complete the training), we learned that most participants at the school had misinterpreted the Teacher and TA Identification Form. We edited the form and instructions accordingly to address this issue, but therefore, school D participants completed a slightly different form than the other schools. The second sensitivity analysis was a complete case analysis intended to explore observed differences in outcomes according to whether participants had completed all 3 time points. For the analysis of outcomes pertaining to preparedness, self-efficacy, and perceptions of training impact, we included all participants who completed the surveys *at least* at T1 and T2.

Statistical Analysis

For all quantitative outcomes, we focused on preliminary, descriptive comparisons across the 3 time points and did not perform any formal hypothesis testing. This aligns with established recommendations for feasibility studies, which generally lack the statistical power necessary for a clear interpretation of hypothesis-testing results [65-68]. We conducted all quantitative analyses in R (version 4.0.3; R Foundation for Statistical Computing) [69] except for the comparison of Teacher and TA Identification Forms and SDQ scores, for which we used Microsoft Excel (Microsoft Corp). We created all plots using the *ggplot2* [70] and *likert* packages [71]. To score the SDQs, we used the freely available R code on the Youthinmind website [72].

Qualitative Outcomes

We considered 3 analysis approaches for the interview and qualitative questionnaire data: content analysis [73], framework analysis [74], and reflexive thematic analysis [75,76]. We initially decided to use content analysis for the survey comments and reflexive thematic analysis for the interviews; however, as we familiarized ourselves with the data, we realized that there was significant overlap between the survey comments and interviews and decided that analyzing them separately was not a useful distinction. As our main aim was to generate insights into the program and its future potential, we decided to use the 6-phase reflexive thematic analysis by Braun and Clarke [76] for all qualitative data due to its flexibility and ability to generate themes both inductively and deductively. ES developed the initial themes, and MF and EH helped clarify and enrich them. ES and MF worked together to name and refine the themes before the final write-up. We managed and coded all qualitative data in ATLAS.ti (version 9.1.3; ATLAS.ti Scientific Software Development GmbH) and additionally created manual thematic maps to better visualize and understand patterns between our data.

Ethical Considerations

This study was approved by the University of Cambridge Psychology Research Ethics Committee (PRE 2019.076). We obtained active informed consent from all teachers and TAs

XSL•FO
RenderX

who took part in the study. We used an opt-out model for parental consent whereby parents received (directly from the schools via their preferred communication routes) an information sheet detailing study aims, procedures, how data would be used, and the right to opt their child out of participation. Parents had 2 weeks to opt their child out of the study by returning a hard copy of the opt-out form or emailing or calling the school. Schools kept track of all opt-outs and instructed teachers and TAs not to include these children in their forms. All quantitative data were collected using anonymous pupil and staff identifiers generated by the participating schools, and all qualitative data were deidentified before analysis, with identifiable information stored on secure servers at the University of Cambridge. Teachers and TAs received £20 (approximately US $28) vouchers for completing the training and questionnaires for at least 2 of the 3 time points and an additional £10 (approximately US $14) for taking part in an interview. School staff members who created the anonymous identifiers received £10 (approximately US $14) vouchers to thank them for their time.

Results

Participants

Schools

A total of 6 schools participated in this study (Table S1 in Multimedia Appendix 4 [40]). Among these 6 schools, there were 4 (67%) from Cambridgeshire and 1 (17%) each from Greater London and Merseyside; 5 (83%) were located in urban areas and 1 (17%) was located in a rural area. All but 1 school (5/6, 83%) were situated in areas of above-average deprivation, and 50% (3/6) of the schools had a higher-than-average proportion of pupils eligible for free school meals. In total, 67% (4/6) of the schools had a high proportion of White pupils (>80%), and 33% (2/6) of the schools were more diverse, with approximately 20% of pupils from Black, Black British, Caribbean, or African backgrounds (school B) or Asian or Asian British backgrounds (school E).

Teachers and TAs

A total of 108 teachers and TAs completed the T1 questionnaires and the training itself, 89 (82.4%) completed the T2 questionnaires, and 70 (64.8%) completed the T3 questionnaires (Table 1), with 54 (50%) having completed all 3. After excluding those teachers and TAs who did not meet the inclusion criteria for the analyses, the final analytical samples were as follows:

1. Main analysis of identification and mental health support outcomes: n=97 at T1, n=75 at T2, and n=57 at T3.
2. Main analysis of preparedness, self-efficacy, and training impact outcomes: n=107 at T1 and n=83 at T2.
3. Main analysis comparing identification outcomes with SDQ scores: n=28 and n=25 (refer to the following section).
4. Complete case sensitivity analysis: n=51 at T1, T2, and T3.
5. Sensitivity analysis excluding all teachers and TAs from school D: n=70 at T1, n=54 at T2, and n=41 at T3.

Compared with the 2019-2020 national workforce statistics for teachers and TAs working in state-funded nursery and primary schools [77], our sample had a similar proportion of women (81/89, 91% in our sample vs 90.9% nationally) and a slightly higher proportion of White staff members (82/89, 92% in our sample vs 90.5% nationally).

A total of 7.4% (8/108) of school staff members from 67% (4/6) of the schools completed an interview (Table 2).

XSL•FO
RenderX

Table 1. Characteristics of participating teachers and teaching assistants (TAs) (N=89[a]).

Characteristics	Values
Gender, n (%)	
Woman	81 (91)
Man	6 (7)
Missing	2 (2)
Age (y), mean (SD)[b]	40.3 (10.9)
School, n (%)	
A	7 (8)
B	15 (17)
C	2 (2)
D	28 (31)
E	22 (25)
F	15 (17)
Position[c], n (%)	
Teacher	42 (47)
TA (including higher-level TAs)	40 (45)
SENCo[d], mental health champion, or mental health lead	3 (3)
School counselor or psychologist	1 (1)
Member of senior leadership team	6 (7)
Prefer not to say	1 (1)
Missing	1 (1)
Years of experience, mean (SD)[e]	12.2 (7.7)
Ethnicity, n (%)	
Asian or Asian British	2 (2)
Mixed or multiple ethnic groups	1 (1)
White	82 (92)
Other ethnic group	2 (2)
Prefer not to say	1 (1)
Missing	1 (1)

[a]N=89 because this information was collected only at T2.

[b]N_{NA}=4 (number with missing data for this question).

[c]Percentages add up to >100 because some participants had multiple roles.

[d]SENCo: special educational needs coordinator.

[e]N_{NA}=7 (number with missing data for this question).

XSL•FO
RenderX

Table 2. Characteristics of the interviewees (n=8).

Interviewee ID	School	Main role or roles
1	B	Assistant headteacher for mental health and well-being, PSHE[a] lead, and class teacher (Reception)
2	B	Assistant headteacher and class teacher (Year 4)
3	E	Class teacher (Year 6)
4	E	SENCo[b]
5	F	SENCo (note: oversaw training but did not complete it)
6	C	Mental health lead and TA[c] (across years)
7	E	Mental health champion and HLTA[d] (Year 5 and 6)
8	F	TA (Early Years Foundation Stage)

[a]PSHE: Personal, Social, Health and Economic.

[b]SENCo: special educational needs coordinator.

[c]TA: teaching assistant.

[d]HLTA: higher-level teaching assistant.

Findings

Teacher and TA Self-Efficacy and Preparedness

Pretest-posttest changes suggested that participating in the training was beneficial for the staff and that they had positive perceptions of the training. Findings regarding preparedness (Figure 2) suggest improvements across all domains of recognizing and acting upon concerns about pupils' mental health, particularly in terms of using key communication strategies and working with parents. Findings regarding self-efficacy (Figure 3) suggest that participants were more confident in their abilities to discuss their concerns about pupils' mental health after the training than before. Again, the largest changes were observed in discussing concerns with parents and applying key communication strategies. Finally, findings regarding teachers' and TAs' perceptions of the impact of applying the skills of the training (Figure 4) suggest that they were generally positive about the possible effects of the training on pupil outcomes (ie, attendance and academic success), teacher-pupil rapport, and the classroom environment. The results from the complete case analysis (Multimedia Appendix 5) were nearly identical to those of the main analysis (all differences were ≤3 percentage points in magnitude).

XSL•FO
RenderX

Figure 2. School staff preparedness to identify and respond to pupil mental health difficulties before and 1 week after completing the *At-Risk for Elementary School Educators* training (n=107 at T1; n=83 at T2). (A) Before training; (B) After training.

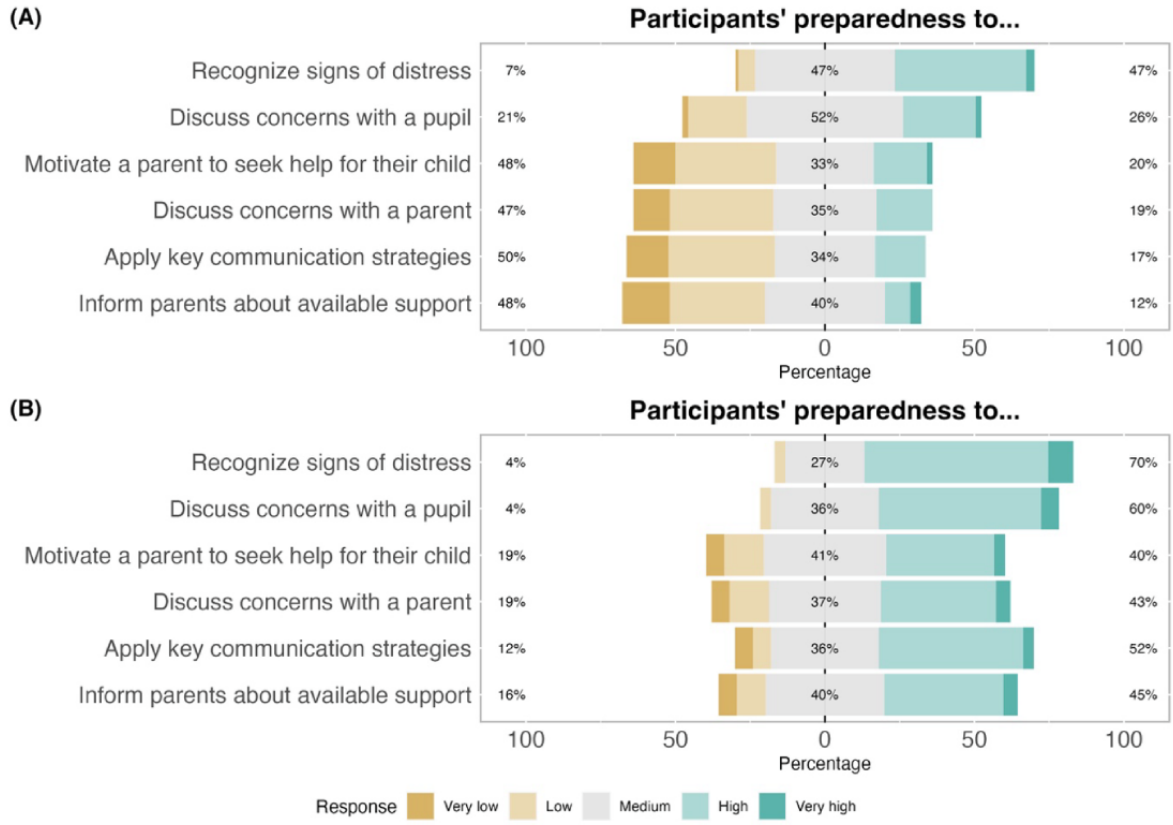

Figure 3. School staff self-efficacy to identify and respond to pupil mental health difficulties before and 1 week after completing the *At-Risk for Elementary School Educators* training (n=107 at T1; n=83 at T2). (A) Before training; (B) After training.

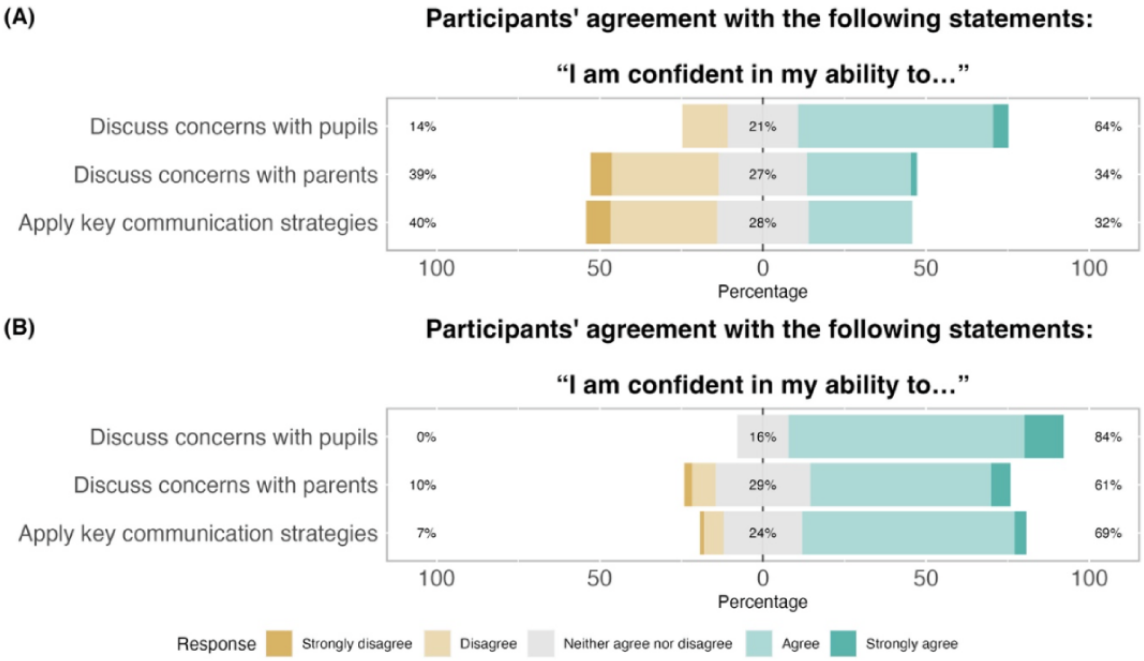

XSL•FO
RenderX

Figure 4. School staff perceptions of the impact of the *At-Risk for Elementary School Educators* training, assessed 1 week after completing it (n=83).

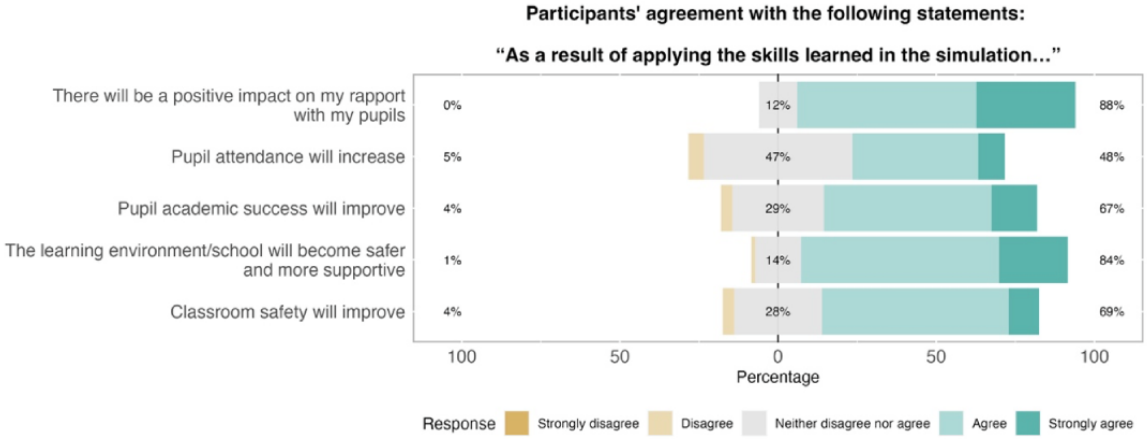

Identification Outcomes

In terms of how many pupils were identified as having mental health difficulties or increased risk, participants identified similar proportions of their pupils before and immediately after the training and then fewer over time. The median percentage of pupils whom participants believed had mental health difficulties or increased risk was 10% (IQR 6.7%-18.2%) at T1, 10% (IQR 4.5%-16.7%) at T2, and 7.4% (IQR 5.0%-16.7%) at T3. The directions of change were similar for both sensitivity analyses (whereby teachers and TAs identified *fewer* children over time), with slight differences. For the sensitivity analysis excluding school D (Multimedia Appendix 6), the percentages were slightly (approximately 2 percentage points) higher. For the complete case analysis, the decrease was also notable 1 week after the training, decreasing from 10% (IQR 6.7%-17.3%) at T1 to 8% (IQR 3.9%-16.7%) at T2 and 7.4% (IQR 5.7%-16.7%) at T3.

In terms of the accuracy of identification, it seems that teachers and TAs became slightly more accurate over time in comparison to pupils' SDQ scores (although it is important to acknowledge the limitations described in the *Methods* section regarding questionnaire comparability). The median percentage of children identified by participants who *did not have elevated SDQ scores* was 40% (IQR 0%-50%) at T1, 27.2% (IQR 0%-50%) at T2, and 25% (IQR 0%-50%) at T3. The median percentage of children with elevated SDQ scores *who were not identified by participants* was 68.8% (IQR 42.9%-87.5%) at T1, 66.7% (IQR 50%-88.2%) at T2, and 57.1% (IQR 33.3%-87.5%) at T3. In the sensitivity analysis excluding school D, the results were similar (typically within 5 percentage points); one small difference was that the median percentage of children identified by teachers and TAs who did not have elevated SDQ scores was 0% (IQR 0%-50%) at T2. The results of the complete case analysis were identical to those of the main analysis.

Mental Health Support Outcomes

Overall, the findings suggest that the training may be beneficial for facilitating conversations and access to school-based support (but not external support) for pupils with identified mental health difficulties or increased risk. Figure 5 presents the findings for the 9 mental health support outcomes among identified children across the 3 study time points. As with before the training, there was typically a wide variation in outcomes.

A comparison across time points suggests that participants formally documented their concerns and spoke with the SENCo, pastoral lead, or mental health lead for a greater proportion of identified pupils after the training than before. For example, at T1, teachers and TAs documented concerns for a median of 50% (IQR 0%-100%) of identified pupils; this increased to 56.3% (IQR 4.2%-100%) at T2 and 75.7% (IQR 0%-100%) at T3. The equivalent statistics for speaking with the SENCo, pastoral lead, or mental health lead were a median of 66.7% (IQR 16.7%-100%) at T1, 75% (IQR 50.0%-100%) at T2, and 95.5% (IQR 50.0%-100%) at T3. There was no change in speaking with another staff member, but this was because nearly all participants did so across all time points. Finally, the percentage of pupils whom teachers and TAs spoke with (or whose parents they spoke with) also increased after the training, with a median of 33.3% (IQR 0%-87.5%) at T1, 61.9% (IQR 0%-100%) at T2, and 50% (IQR 0%-100%) at T3.

A comparison across time points also suggests increases in school-based support for identified children after the training compared with before. The median percentage of pupils identified by teachers and TAs who received in-class support increased from 75% (IQR 35.4%-100%) at T1 to 100% at T2 and T3 (IQR 50%-100% and 66.7%-100%, respectively). There was a more modest increase in the receipt of in-school support or in-house support plans, with a median of 40% (IQR 0%-71.4%) of identified pupils receiving them at T1 compared with 50% at T2 and T3 (IQR 3.6%-100% and 8.3%-81.4%, respectively). There was very little change in documented SEMH status or referral or access to specialist mental health services. For each of these outcomes, the median percentage of identified pupils was 0% across time points.

The findings from the sensitivity analyses were similar to those of the main analysis in terms of direction, although improvements across time in the complete case analysis (Multimedia Appendix 5) tended to be more modest than for the main analysis.

XSL•FO
RenderX

Figure 5. Mental health support outcomes assessed before (T1), 1 week after (T2), and approximately 3 months after (T3) completing the *At-Risk for Elementary School Educators* training. Staff-related outcomes include (A) formally documenting concerns with the school; (B) communicating concerns to the special educational needs coordinator (SENCo), pastoral care lead, or mental health (MH) lead; (C) communicating concerns to another member of the school staff; and (D) communicating concerns to the child or their parents. Pupil outcomes include (E) receiving in-class support; (F) receiving in-school support or having an in-house support plan; (G) having documented social, emotional, or MH (SEMH) status; (H) having a referral to external MH services; and (I) having access to external MH services. These outcomes exclude participants who were not concerned about any child (n=2 at T1; n=5 at T2; n=5 at T3).

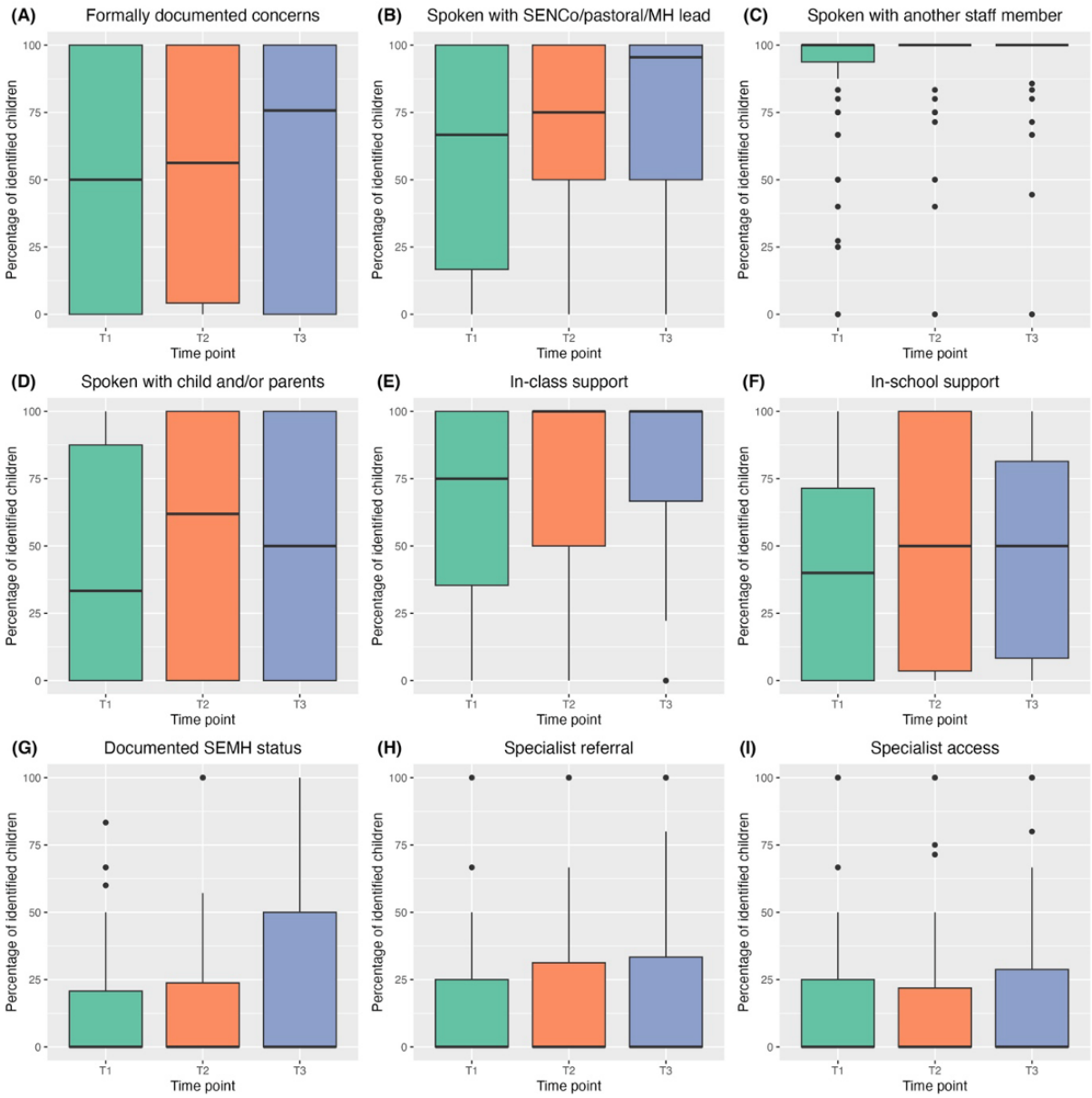

Acceptability and Practicality

Quantitative Findings

Quantitative data from the posttraining survey showed that participants were generally positive about the training. Of the 83 participants who completed the survey, 53 (64%) rated it as “good” and 13 (16%) rated it as “very good.” An additional 17% (14/83) rated it as “fair,” 2% (2/83) rated it as “poor”, and 1% (1/83) as “very poor.” A total of 84% (70/83) of the teachers and TAs said that the scenarios in the training were relevant to them. Finally, most participants (74/83, 89%) would recommend the training to other educators.

Qualitative Findings

Qualitative data also suggested that school staff generally found the training practical and acceptable. We generated three themes from our survey and interview data:

1. Individual fit: positive perceptions, self-efficacy, and change.
2. Institutional fit: alignment with school values and context.
3. Taking it forward: improvements and implementation.

Additional findings on possible harms are presented in Multimedia Appendix 7.

XSL•FO
RenderX

Individual Fit: Positive Perceptions, Self-Efficacy, and Change

In general, participants perceived the program to be a "good fit" with their personal philosophies and practice. Regarding the training itself, many appreciated the included scenarios, particularly in terms of their relevance to their practice. The format of the training—primarily that it was web-based and required active role-play—was also viewed as useful, engaging, and novel and might have contributed to its perceived usefulness. For example, one teacher commented:

> *The interactive elements of the training were brilliant and something which I have never encountered before!* [Survey respondent (SR) 56; school E]

One teacher and well-being lead described:

> *I think it definitely made you think. [...] you had to really think about what was being said and the response that you would give, reflecting back on sort of the knowledge that they'd given you beforehand, so I thought that was good.* [Interviewee 1]

Other participants suggested that opportunities to practice skills during the training improved the likelihood of using those skills in day-to-day practice.

Participants also believed that they had learned a lot from the training, especially in terms of *skills and strategies.* These included but were not limited to the skills within the *At-Risk* "EASING" strategy (check your Emotions, Ask for permission, be Specific, use *I* statements, keep it Neutral, and show Genuine curiosity). Importantly, there was evidence that participants had also *applied* new skills. Several participants described having new conversations with pupils or parents facilitated by the skills and strategies from the training. For example, one teacher described:

> *It was that permission thing [...] I wanted to ask [a child] about his home life [...] and kind of he just cried and didn't want to speak about it anymore, and then when I asked him if we were OK to talk about it, he said, "Actually no, because I think I'm going to cry again," so then we left it. And then he came to me the following week, and [...] said, "Can we talk about it now?" [...] so actually me asking that, it was the wrong time for him to talk about it, he wasn't ready, he would have just been emotional, and wouldn't have been able to get his words out, and actually the week after, him coming to me and saying, "Can we have a little chat," works perfectly [...] And now we're more aware of his situation.* [Interviewee 2]

This skill seems to have enabled this pupil to have this conversation with the teacher in a manner (in terms of time, place, and identified person) that suited him. Other participants provided similar examples, referencing how skills from the training had facilitated better outcomes.

However, it is important to note that the perceived usefulness of the training varied. Most notably, some participants indicated that their previous training or role made the training less impactful. Illustratively, when asked how the training had impacted their practice, one TA responded:

> *Having previously received similar training, due to my role, I do not have any recent cases where the training would have changed the way I carried out discussions.* [SR 60; school E]

Institutional Fit: Alignment With School Values and Context

Sustainable school-based programs should also align with the values of the school more broadly. Participants often referenced the importance of schools' prioritization of pupil mental health. For example, one teacher described:

> *[Mental health is] a conversation which is constantly ongoing and trying to constantly better our practices and make sure we're looking after them as best as we can and spotting things as best we can as well.* [Interviewee 3]

This description demonstrates how prioritizing mental health can promote the critical evaluation of related school practices as well as the additional provision of training opportunities. In many cases, support from the SLT led to *formal* recognition of pupil mental health within school policies or plans. One strategic stakeholder explained:

> *I think because our school have well-being and mental health as such a focus, SLT are very supportive of doing things like this and they're very accommodating. So when I said we had the training and people were going to have to take part in the training, it was very flexible, although they had other ones lined up, they were quite happy to move things around to make things work. And I think, the fact it is such a priority in our school definitely makes that easier.* [Interviewee 1]

In this school, mental health and well-being were one of three main school priorities. As indicated previously, direction setting from the SLT is key to ensuring momentum and impetus. However, as others noted, it is important that support from the SLT is genuine rather than being "just another tick box" (Interviewee 4) exercise.

Another facet of institutional fit pertained to the practical aspects of the training. Schools are time- and resource-limited settings, so mental health training needs to fit within this context. The format of *At-Risk*, especially its flexibility and relatively low time requirements, was viewed as beneficial, with comments such as "For the amount of time [...] I got a huge amount from it" (Interviewee 4). Others made direct comparisons with other training courses. For example, one higher-level TA had previously completed a 1-day, in-person training course with a similar purpose to that of *At-Risk.* While she preferred the in-person training, she listed the benefits of both types:

> *[In the in-person training] you can then query and question to your trainer, so you've got that interaction, so that obviously isn't there, is it, on the computer one. [...] if I was looking from a management point of view, I would say, budgetary, I'm sure it's cheaper [...] to use [At-Risk], not just*

XSL•FO
RenderX

> *cheaper as in [...] money, [...] but also cheaper in time [...] So probably if I was looking [...] with my management hat on, I would say the computer-based [training] would get the same message, or similar message, across for a wider audience for probably a cheaper cost.* [Interviewee 7]

In terms of efficiency, this participant highlighted the favorable input-to-output ratio of *At-Risk*, which could allow more staff members to participate in training. This quote also highlights that schools could use *At-Risk* flexibly. For example, schools might assign staff members to different training programs based on their roles and previous experience, with more intensive, in-person training for staff members with more significant mental health roles and *At-Risk* for those with fewer responsibilities or less experience.

Taking It Forward: Improvements and Implementation

Participants offered key insights into how to take the training forward in terms of both changes to the training itself and how best to implement it, primarily by tailoring it to the UK context. In terms of language, there was some reference to the American accent, but more so, participants highlighted the need to adapt some of the terminology and signposting resources to reflect UK support systems. They also made suggestions about additional training that could be useful with different topics (such as bullying) and age groups (particularly for younger children).

In addition to improvements and adaptations to the training itself, participants illustrated the importance of implementation. A common theme was that, to maximize impact, the training should include follow-up discussions or live workshopping. One teacher suggested:

> *I think some kind of "live" element to conclude the training—to have a "real" person to ask questions to as part of a group video chat could have been useful. Also, maybe to ask advice about particular scenarios that we may have found ourselves in in the past.* [SR 56; school E]

By facilitating greater engagement and critical thinking, a live element could enhance the impact of the training and potentially make *At-Risk* more acceptable to those who generally prefer face-to-face training. Participants indicated that someone internal, for example, the SENCo, would be best placed to lead a live element and would enable staff to practice role-playing based on situations and scenarios specific to each school.

There was also wide acknowledgment that any training had to lie within a strong support system. This began with having a clear referral pathway for identified concerns, which was viewed as important for facilitating access to support. In some cases, teachers and TAs were able to find new ways to support children after completing the training. However, in many cases, participants—and strategic stakeholders in particular—explained that support had not always been readily available. For example, one strategic stakeholder recounted what happened after the training:

> *A lot of them are people saying to me, "What are you going to do about it?" about different children. And I, because some of our support staff don't know the sort of route for getting extra support, or they're really shocked to find actually there's nothing out there for these children...it's about what we can do in school, and I think people have been really quite shocked about that. You know, they just presume I can make a phone call and these children will get face-to-face counselling.* [Interviewee 5]

This shows the importance of embedding the training within a wider support system, including collaboration with external agencies. However, many interviewees referenced the systemic issues that schools face in helping pupils access specialist support, particularly in terms of the high thresholds and long waiting lists that exist for many external services. While schools may be able to provide beneficial support for children, particularly for those with lower-level difficulties, this indicates an ongoing area of need for schools and their pupils.

Discussion

Summary of Findings

This study offers the first UK evidence for Kognito's *At-Risk for Elementary School Educators*, extending findings from 3 US-based trials and providing needed evidence regarding the potential utility, acceptability, and practicality of brief, interactive web-based mental health training for school staff. Overall, the findings showed that *At-Risk* is a feasible means of improving the identification of and response to pupil mental health difficulties in UK primary schools. Quantitative findings showed that staff preparedness and self-efficacy in identifying and responding to mental health difficulties increased after the training. Identification rates did not increase (and, in fact, decreased at the 3-month follow-up), but there was some suggestion that teachers' and TAs' identification became slightly more accurate in comparison with SDQ scores. Crucially, for those pupils identified as having mental health difficulties or increased risk, in-school mental health support outcomes (ie, documentation or discussion of concerns, conversations with pupils and parents, and in-class and in-school support) increased after the training, but more "downstream" outcomes (ie, documented SEMH status and referral and access to external mental health services) did not. Qualitative findings indicated that participants generally found the training acceptable and practical, with many explaining how they intended to use or had already used the skills they learned to improve their practice. Participants also suggested several useful improvements for the training and its implementation, including making it more relevant to the United Kingdom, adding more scenarios, and including a live element in the implementation of the training.

Findings regarding confidence and preparedness reflect those of the 3 US-based studies of *At-Risk* [17,41,42] and the wider literature surrounding teacher mental health training [31]. In general, mental health training seems to be effective in improving staff confidence. For example, 2 Australian-based studies [37,78] found that secondary school teachers who completed training felt more confident discussing their concerns and helping pupils with their mental health. Another UK-based study of a psychoeducational training program to improve

XSL•FO
RenderX

recognition of depression in secondary schools [79] found significant pretest-posttest improvements in teacher confidence in their knowledge of symptoms, ability to recognize symptoms, and knowledge about how to speak with pupils about their mental health. However, not all studies have shown an impact, with a prominent UK-based study of mental health first aid training finding no effect on educators' confidence in helping pupils with their mental health [80].

The general decrease in the proportion of pupils identified as having mental health difficulties or increased risk stands in contrast to previous studies of *At-Risk*, which found that school staff identified significantly more pupils of concern after completing the training [17,41]. Evidence of the effect of other training programs on identification is extremely limited [30,31,36], and differences in context, training content and delivery, baseline knowledge, and outcome measurement make it difficult to compare findings across studies. Two vignette-based studies showed little effect of either mental health first aid [78] or psychoeducational [81] training on identification (although each study also reported high recognition of difficulties before the training), whereas studies focused on real-world identification have shown mixed results [79,82]. However, changes in the proportion of identified pupils must be contextualized within the accuracy of identification. There are consequences of both over- and underidentification [83,84], most notably in terms of inefficient allocation of limited mental health support resources. While comparison with the SDQ suggested that there was some improvement in terms of the accuracy of identification following the training, underidentification remained a substantial challenge, with between one-half and two-thirds of pupils with elevated SDQ scores remaining unidentified by teachers and TAs. The underidentification of children with mental health difficulties in educational settings, particularly for children with internalizing as opposed to externalizing problems [85], has been well documented in the literature [30], and it is likely that a combination of identification models is required to address this challenge [27,29].

Promisingly, the training appeared to be useful in terms of connecting pupils with care and support, an outcome not frequently measured in other studies [30,31,34]. First, the findings suggested that participants had conversations about or documented concerns for a greater proportion of identified pupils following the training, which reflects findings from previous studies of *At-Risk* [17,41]. This is a rather unique outcome in the literature as other training evaluations have found no difference between training and control groups in terms of conversations with pupils and colleagues [78]. Importantly, this study went beyond conversations to include outcomes pertaining to in-school and external support. The increases in in-class and in-school support for identified pupils reflect findings of the UK-based study by Kidger et al [80] of mental health first aid training and the Australian-based pilot study by Parker et al [37] of a web-based training program, each of which found a positive effect of the training on helping behaviors. Although in-class and in-school support seemed to increase following the training, it is notable that referrals and access to specialist services did not. There are several plausible explanations for this finding. For example, it is likely that school staff were already aware of children with the most severe mental health difficulties and were confident and able to support newly identified pupils—who might have had lower-level mental health difficulties—within the school setting. However, if the training *did* lead to the identification of children who might benefit from specialist care, there are many barriers to accessing such support (eg, availability and long waiting lists) that might have influenced these outcomes, as reflected in both the qualitative interviews and the wider literature [23,86].

In addition, quantitative and qualitative findings suggested that the program was a good fit for individuals and schools, which aligns with previous research on the acceptability and perceived need for mental health training for school staff [18,20,27-29,87,88]. The training's format seemed to be a key contributor to its feasibility. With a few exceptions [37,39,89], the web-based simulation-based format of *At-Risk* is unique among training programs and is well aligned with teachers' preferences. For example, in their focus group study of UK secondary school teachers, Shelemy et al [20] found that participants wanted engaging, interactive, and concise training that included practical strategies and illustrative case studies, all of which are central to *At-Risk*. While the authors found that teachers disagreed over the usefulness of web-based training, it is possible that these concerns would have decreased during the COVID-19 pandemic as staff became more accepting of web-based opportunities to learn.

Qualitative findings also demonstrated the importance of school context and culture, which have been highlighted in previous research [27]. In particular, participants noted the importance of school culture in adopting mental health interventions into regular practice. In their systematic review, Moore et al [90] identified school culture, values, and policies as key facilitators of sustaining mental health interventions. A related area of focus was support from the SLT. This support is a well-recognized factor contributing to intervention success and sustainability for several reasons, including these leaders' practical role in communicating about interventions and allocating specific time and resources to them [43,90,91]. However, it is important to recognize that mental health training for school staff may be even more needed and impactful in schools where mental health is not as much of a priority.

Limitations

Our mixed methods approach, wide range of outcomes, and diverse sample of participating schools offer rich information regarding the feasibility of *At-Risk* in the United Kingdom. These strengths notwithstanding, there are also several limitations to consider when interpreting the findings. The nonrandomized design, while common for feasibility studies, prevents any conclusions regarding causality and also limits the exploration of other factors that may have influenced outcomes (eg, providing teachers and TAs with the Mental Health Resource Maps or SENCos and mental health leads with feedback on identified pupils). In terms of recruitment and retention, the study had 50% (54/108) attrition. Several factors may have influenced this, including the increased pressure on school staff due to the COVID-19 pandemic, the timing of the

XSL•FO
RenderX

study within the school year, and the requirement to communicate with participants only via the study link person. While we tried to explore the effect of attrition through a complete case sensitivity analysis, we lacked important information on the characteristics of those who dropped out as this information was collected only at T2. Furthermore, we were only able to recruit 8 staff members for the posttraining interviews, which was far below our recruitment target. Low participation rates could again be due to several factors, including the impact of the COVID-19 pandemic or competing priorities. Of note, we were not able to recruit anyone who did not complete the training, any headteachers, or any staff from 2 of the schools (schools A and D). This could mean that we lack viewpoints that may be important for understanding the feasibility and utility of the training.

There were also limitations associated with the study measures. While the Teacher and TA Identification Form was informed by the literature and reviewed by our primary school staff advisory group, its validity and reliability are unknown. In addition, the questionnaire only measured mental health support outcomes for those pupils *identified as having mental health difficulties or increased risk*. Therefore, we do not have information on those who were not identified. The measure is also based on teacher and TA reports and so may not have complete information about all types of support that pupils receive. Another important limitation pertains to the comparator used to assess the identification outcomes. To understand the potential utility of the training program, it is important to have a robust comparator. While we chose to use the teacher-report SDQ, it would also have been interesting to compare identification outcomes with parent-rated mental health difficulties, particularly in light of the low interrater agreement of common measures of child mental health difficulties [92]. An even stronger comparator would be to assess the teacher and TA identification outcomes against a clinical interview; however, this was not feasible in this study.

Finally, at the time of writing, *At-Risk* is currently not available for use as Kognito restructures its offerings. This demonstrates a trend that unfortunately is a common occurrence in the field of mental health, whereby many evidence-based digital tools are not available to potential end users [93]. Nonetheless, the learnings from this feasibility study offer rich information on what type of content and format may be useful for training programs in this area and, as such, can support further development and evaluation in the field.

Implications for Practice

Studies have consistently demonstrated that school staff would appreciate additional training on how best to support pupil mental health [18,20,87,88]. However, to be scalable, such programs must be realistic in terms of time, cost, and resource requirements [28,90,91]. Contextualized within the wider literature on school-based mental health interventions, the findings from this study suggest that mental health training is a feasible option for upskilling school staff to identify and respond to pupil mental health difficulties. They further highlight several specific factors that might positively contribute to feasibility and scalability, many of which are reflected in the broader literature on mental health training [20,28]. For example, teachers and TAs appreciated that the training actively engaged them in learning and applying new skills and that it used realistic examples to demonstrate the real-world applicability of the training, whereas school leaders identified the relatively low time and cost requirements and flexibility as key factors that could make the training feasible for their school context.

However, this is not to say that there are no implementation barriers associated with *At-Risk* or similar training programs. While the resources required to implement *At-Risk* are relatively low compared with other training programs, they must still be considered within the context of other school priorities. As demonstrated in the interviews and the wider literature [3,43,90,91], support from school leadership is essential for securing the time and budget required to implement a training such as *At-Risk*, and in schools where mental health is not a priority, there are likely to be many barriers to implementation. Even in schools with strong support from the leadership team, it may be difficult to find the requisite budget, time, and human resources to devote to the training. Finally, as is the case with any school-based mental health intervention, it is important that schools do not take sole responsibility for pupils' mental health. Active partnership between schools and mental health services is key to ensuring that schools feel empowered and supported in this role [21,90,94]. While the schools in this study worked hard to support pupils as best they could, interviewees expressed frustration about the difficulty of accessing external support for children who could benefit from it. This is not an uncommon theme in the wider literature surrounding school-based interventions [20,23,91] and is a key consideration for scaling up training programs.

Implications for Future Research

The promising findings of this study suggest that additional research is needed to explore the role of scalable mental health training in supporting schools to protect and promote children's mental health. On the basis of gaps in the literature, particular areas of interest include training for primary school staff (as most are focused on secondary school staff), web-based training (as opposed to traditional time- and resource-intensive in-person training), and training that takes a "whole school approach" by including all school staff members (rather than only teachers). This final area is especially interesting as findings from this study and others [27] have highlighted stakeholders' preference that training programs include all school staff members. While our study jointly analyzed findings for teachers and TAs, future research would do well to consider how the unique roles and perspectives of these professionals—as well as other staff members within the school setting—might influence outcomes. Furthermore, future research should be more inclusive about their choice of outcomes, as too often evaluations of school staff training programs have focused on intermediate outcomes such as knowledge or confidence [31] without considering more "downstream" outcomes such as access to support. Finally, as demonstrated in our study, there is great value in using mixed methods approaches and including information about wider issues of feasibility and implementation, and studies that take this broader lens can help identify programs that are scalable, sustainable, and effective.

XSL•FO
RenderX

Conclusions

School staff would welcome additional mental health training to enable them to respond to pupil mental health difficulties, but there are many barriers to implementing such training at scale. Therefore, training programs that have relatively low time and resource requirements have great potential to fulfill an unmet need in schools. This mixed methods feasibility study showed that *At-Risk for Elementary School Educators*—an example of a brief, interactive web-based training program—is a feasible means of empowering school staff to accurately identify and respond to pupil mental health difficulties and increased risk.

Acknowledgments

The authors would like to thank Professor Paul Ramchandani for his input in the design of this study, the Cambridge Mental Health in Schools Advisory Group for sharing their views and advice throughout the study, and the staff at the 6 schools that took part in this study for their time and effort. This study was funded by the UK Research and Innovation Emerging Minds network (grant ES/S004726/2), and the training was provided to the schools free of cost by Kognito. ES was funded by a Gates Cambridge Scholarship (grant OPP1144) for the duration of the study. MF is funded by the National Institute for Health and Care Research (NIHR) Oxford and Thames Valley Applied Research Collaboration at the Oxford Health National Health Service (NHS) Foundation Trust. PBJ is funded by the NIHR (grant 0616-20003). All research in the Cambridge Department of Psychiatry is supported by the NIHR Applied Research Collaboration East of England and the Cambridge Biomedical Research Centre at the Cambridgeshire and Peterborough NHS Foundation Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or the Bill & Melinda Gates Foundation.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to restrictions associated with our ethics approvals but are available from the corresponding author on reasonable request.

Conflicts of Interest

Kognito is a for-profit company. After reviewing their questionnaires on preparedness and self-efficacy and having found them rigorous and unbiased, we independently decided to include them as outcomes in our study. Kognito had no role in the study design, analysis, or publication. PBJ was a scientific advisory board member for MSD. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Teacher and Teaching Assistant Identification Form.
[PDF File (Adobe PDF File), 3204 KB - jmir_v26i1e46764_app1.pdf]

Multimedia Appendix 2
Kognito pre- and posttraining surveys.
[PDF File (Adobe PDF File), 88 KB - jmir_v26i1e46764_app2.pdf]

Multimedia Appendix 3
Interview topic guides.
[PDF File (Adobe PDF File), 93 KB - jmir_v26i1e46764_app3.pdf]

Multimedia Appendix 4
School characteristics.
[PDF File (Adobe PDF File), 59 KB - jmir_v26i1e46764_app4.pdf]

Multimedia Appendix 5
Results from complete case sensitivity analysis.
[PDF File (Adobe PDF File), 1746 KB - jmir_v26i1e46764_app5.pdf]

Multimedia Appendix 6
Results from the sensitivity analysis excluding school D.
[PDF File (Adobe PDF File), 716 KB - jmir_v26i1e46764_app6.pdf]

Multimedia Appendix 7

XSL•FO
RenderX

Quantitative and qualitative findings pertaining to the potential harms of At-Risk.
[PDF File (Adobe PDF File), 36 KB - jmir_v26i1e46764_app7.pdf]

References

1. Transforming children and young people's mental health provision: a green paper. Department of Health and Department for Education, United Kingdom Government. 2017 Dec. URL: https://assets.publishing.service.gov.uk/media/5a823518e5274a2e87dc1b56/Transforming_children_and_young_people_s_mental_health_provision.pdf [accessed 2024-03-25]
2. Future in mind: promoting, protecting and improving our children and young people's mental health and wellbeing. National Health Service England. URL: https://assets.publishing.service.gov.uk/media/5a80b26bed915d74e33fbe3c/Childrens_Mental_Health.pdf [accessed 2024-03-25]
3. Promoting children and young people's mental health and wellbeing. Office for Health Improvement and Disparities and Department for Education, United Kingdom Government. 2015 Mar 20. URL: https://www.gov.uk/government/publications/promoting-children-and-young-peoples-emotional-health-and-wellbeing [accessed 2024-03-25]
4. Newlove-Delgado T, Williams T, Robertson K, McManus S, Sadler K, Vizard T, et al. Mental health of children and young people in England 2021 - wave 2 follow up to the 2017 survey. National Health Service Digital. 2021. URL: https://openaccess.city.ac.uk/id/eprint/28718/ [accessed 2024-03-25]
5. Vizard T, Sadler K, Ford T, Newlove-Delgado T, McManus S, Marcheselli F, et al. Mental health of children and young people in England, 2020. National Health Service Digital. 2020 Oct 22. URL: https://www.infocoponline.es/pdf/mhcyp_2020_rep.pdf [accessed 2024-03-25]
6. Pitchforth J, Fahy K, Ford T, Wolpert M, Viner RM, Hargreaves DS. Mental health and well-being trends among children and young people in the UK, 1995-2014: analysis of repeated cross-sectional national health surveys. Psychol Med 2019 Jun;49(8):1275-1285 [FREE Full text] [doi: 10.1017/S0033291718001757] [Medline: 30201061]
7. Newlove-Delgado T, McManus S, Sadler K, Thandi S, Vizard T, Cartwright C, et al. Child mental health in England before and during the COVID-19 lockdown. Lancet Psychiatry 2021 May;8(5):353-354 [FREE Full text] [doi: 10.1016/S2215-0366(20)30570-8] [Medline: 33444548]
8. Creswell C, Shum A, Pearcey S, Skripkauskaite S, Patalay P, Waite P. Young people's mental health during the COVID-19 pandemic. Lancet Child Adolesc Health 2021 Aug;5(8):535-537 [FREE Full text] [doi: 10.1016/S2352-4642(21)00177-2] [Medline: 34174991]
9. Fazel M, Hoagwood K, Stephan S, Ford T. Mental health interventions in schools 1: mental health interventions in schools in high-income countries. Lancet Psychiatry 2014 Oct;1(5):377-387 [FREE Full text] [doi: 10.1016/S2215-0366(14)70312-8] [Medline: 26114092]
10. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
11. Greenberg MT, Domitrovich CE, Weissberg RP, Durlak JA. Social and emotional learning as a public health approach to education. Future Child 2017;27(1):13-32. [doi: 10.1353/foc.2017.0001]
12. Humphrey N, Wigelsworth M. Making the case for universal school-based mental health screening. Emot Behav Difficulties 2016;21(1):22-42. [doi: 10.1080/13632752.2015.1120051]
13. Kutcher S, Wei Y. School mental health: a necessary component of youth mental health policy and plans. World Psychiatry 2020 Jun;19(2):174-175 [FREE Full text] [doi: 10.1002/wps.20732] [Medline: 32394554]
14. Fazel M, Soneson E. Current evidence and opportunities in child and adolescent public mental health: a research review. J Child Psychol Psychiatry 2023 Dec;64(12):1699-1719. [doi: 10.1111/jcpp.13889] [Medline: 37771261]
15. Stephan SH, Weist M, Kataoka S, Adelsheim S, Mills C. Transformation of children's mental health services: the role of school mental health. Psychiatr Res 2007 Oct;58(10):1330-1338. [doi: 10.1176/appi.ps.58.10.1330]
16. Day L, Blades R, Spence C, Ronicle J. Mental health services and schools link pilots: evaluation brief. Department for Education, United Kingdom Government. 2017 Feb. URL: https://assets.publishing.service.gov.uk/media/5a817dfaed915d74e33fe812/Evaluation_of_the_MH_services_and_schools_link_pilots-RB.pdf [accessed 2024-03-25]
17. Long MW, Albright G, McMillan J, Shockley KM, Price OA. Enhancing educator engagement in school mental health care through digital simulation professional development. J Sch Health 2018 Sep;88(9):651-659. [doi: 10.1111/josh.12670] [Medline: 30133775]
18. Rothì DM, Leavey G, Best R. On the front-line: teachers as active observers of pupils' mental health. Teach Teach Educ 2008 Jul;24(5):1217-1231. [doi: 10.1016/j.tate.2007.09.011]
19. Mazzer KR, Rickwood DJ. Teachers' role breadth and perceived efficacy in supporting student mental health. Adv Sch Ment Health Promot 2015;8(1):29-41. [doi: 10.1080/1754730x.2014.978119]
20. Shelemy L, Harvey K, Waite P. Supporting students' mental health in schools: what do teachers want and need? Emot Behav Difficulties 2019;24(1):100-116. [doi: 10.1080/13632752.2019.1582742]
21. Fazel M, Soneson E, Sellars E, Butler G, Stein A. Partnerships at the interface of education and mental health services: the utilisation and acceptability of the provision of specialist liaison and teacher skills training. Int J Environ Res Public Health 2023 Feb 24;20(5):4066 [FREE Full text] [doi: 10.3390/ijerph20054066] [Medline: 36901077]

XSL•FO
RenderX

22. Marshall L, Wishart R, Dunatchik A, Smith N. Supporting mental health in schools and colleges: quantitative survey. Department for Education, United Kingdom Government. 2017 Aug. URL: https://assets.publishing.service.gov.uk/media/5a82e5bb40f0b6230269d46c/Supporting_Mental-Health_survey_report.pdf [accessed 2024-03-25]
23. Sharpe H, Ford T, Lereya ST, Owen C, Viner RM, Wolpert M. Survey of schools' work with child and adolescent mental health across England: a system in need of support. Child Adolesc Ment Health 2016 Sep;21(3):148-153. [doi: 10.1111/camh.12166] [Medline: 32680352]
24. Danby G, Hamilton P. Addressing the 'elephant in the room'. The role of the primary school practitioner in supporting children's mental well-being. Pastor Care Educ 2016;34(2):90-103. [doi: 10.1080/02643944.2016.1167110]
25. Reinke WM, Stormont M, Herman KC, Puri R, Goel N. Supporting children's mental health in schools: teacher perceptions of needs, roles, and barriers. Sch Psychol Q 2011;26(1):1-13. [doi: 10.1037/a0022714]
26. Kidger J, Gunnell D, Biddle L, Campbell R, Donovan J. Part and parcel of teaching? Secondary school staff's views on supporting student emotional health and well-being. Br Educ Res J 2013 Jan 02;36(6):919-935. [doi: 10.1080/01411920903249308]
27. Soneson E, Burn AM, Anderson JK, Humphrey A, Jones PB, Fazel M, et al. Determining stakeholder priorities and core components for school-based identification of mental health difficulties: a Delphi study. J Sch Psychol 2022 Apr;91:209-227 [FREE Full text] [doi: 10.1016/j.jsp.2022.01.008] [Medline: 35190077]
28. Soneson E, Howarth E, Ford T, Humphrey A, Jones PB, Thompson Coon J, et al. Feasibility of school-based identification of children and adolescents experiencing, or at-risk of developing, mental health difficulties: a systematic review. Prev Sci 2020 Jul;21(5):581-603 [FREE Full text] [doi: 10.1007/s11121-020-01095-6] [Medline: 32062764]
29. Childs-Fegredo J, Burn AM, Duschinsky R, Humphrey A, Ford T, Jones PB, et al. Acceptability and feasibility of early identification of mental health difficulties in primary schools: a qualitative exploration of UK school staff and parents' perceptions. Sch Ment Health 2020 Nov 18;13:143-159. [doi: 10.1007/s12310-020-09398-3]
30. Anderson JK, Ford T, Soneson E, Coon JT, Humphrey A, Rogers M, et al. A systematic review of effectiveness and cost-effectiveness of school-based identification of children and young people at risk of, or currently experiencing mental health difficulties. Psychol Med 2018 Sep 13;49(1):9-19. [doi: 10.1017/s0033291718002490]
31. Yamaguchi S, Foo JC, Nishida A, Ogawa S, Togo F, Sasaki T. Mental health literacy programs for school teachers: a systematic review and narrative synthesis. Early Interv Psychiatry 2020 Feb;14(1):14-25. [doi: 10.1111/eip.12793] [Medline: 30740884]
32. Kognito homepage. Kognito. URL: https://kognito.com/ [accessed 2024-03-26]
33. At-risk for elementary school educators. Kognito. URL: https://kognito.com/products/at-risk-for-elementary-schools [accessed 2024-03-25]
34. Anderson M, Werner-Seidler A, King C, Gayed A, Harvey SB, O'Dea B. Mental health training programs for secondary school teachers: a systematic review. Sch Ment Health 2018 Oct 3;11:489-508. [doi: 10.1007/s12310-018-9291-2]
35. Ekornes S. Teacher perspectives on their role and the challenges of inter-professional collaboration in mental health promotion. Sch Ment Health 2015 Mar 17;7:193-211. [doi: 10.1007/s12310-015-9147-y]
36. Ohrt JH, Deaton JD, Linich K, Guest JD, Wymer B, Sandonato B. Teacher training in K–12 student mental health: a systematic review. Psychol Sch 2020 Mar 03;57(5):833-846. [doi: 10.1002/pits.22356]
37. Parker BL, Anderson M, Batterham PJ, Gayed A, Subotic-Kerry M, Achilles MR, et al. Examining the preliminary effectiveness and acceptability of a web-based training program for Australian secondary school teachers: pilot study of the BEAM (building educators' skills in adolescent mental health) program. JMIR Ment Health 2021 Oct 22;8(10):e29989 [FREE Full text] [doi: 10.2196/29989] [Medline: 34677134]
38. McVey G, Gusella J, Tweed S, Ferrari M. A controlled evaluation of web-based training for teachers and public health practitioners on the prevention of eating disorders. Eat Disord 2009;17(1):1-26. [doi: 10.1080/10640260802570064] [Medline: 19105058]
39. Pereira CA, Wen CL, Miguel EC, Polanczyk GV. A randomised controlled trial of a web-based educational program in child mental health for schoolteachers. Eur Child Adolesc Psychiatry 2015 Aug;24(8):931-940. [doi: 10.1007/s00787-014-0642-8] [Medline: 25395381]
40. Kognito At-Risk in PK-12. Oklahoma Department of Mental Health and Substance Abuse Services. URL: https://oklahoma.gov/odmhsas/trainings/training-institute/at-risk-pk-12.html [accessed 2024-03-14]
41. Albright G, Fazel M, Khalid N, McMillan J, Hilty D, Shockley K, et al. High school educator training by simulation to address emotional and behavioral concerns in school settings: a randomized study. J Technol Behav Sci 2022;7(3):277-289 [FREE Full text] [doi: 10.1007/s41347-022-00243-9] [Medline: 35372668]
42. Greif Green J, Levine RS, Oblath R, Corriveau KH, Holt MK, Albright G. Pilot evaluation of preservice teacher training to improve preparedness and confidence to address student mental health. Evid Based Pract Child Adolesc Ment Health 2020;5(1):42-52. [doi: 10.1080/23794925.2020.1727793]
43. Domitrovich CE, Bradshaw CP, Poduska JM, Hoagwood K, Buckley JA, Olin S, et al. Maximizing the implementation quality of evidence-based preventive interventions in schools: a conceptual framework. Adv Sch Ment Health Promot 2008 Jul;1(3):6-28 [FREE Full text] [doi: 10.1080/1754730x.2008.9715730] [Medline: 27182282]

XSL•FO
RenderX

44. Bonell C, Prost A, Melendez-Torres GJ, Davey C, Hargreaves JR. Will it work here? A realist approach to local decisions about implementing interventions evaluated as effective elsewhere. J Epidemiol Community Health 2021 Jan;75(1):46-50. [doi: 10.1136/jech-2020-214287] [Medline: 32907917]
45. Domitrovich CE, Greenberg MT. The study of implementation: current findings from effective programs that prevent mental disorders in school-aged children. J Educ Psychol Consult 2000;11(2):193-221. [doi: 10.1207/S1532768XJEPC1102_04]
46. Evans RE, Craig P, Hoddinott P, Littlecott H, Moore L, Murphy S, et al. When and how do 'effective' interventions need to be adapted and/or re-evaluated in new contexts? The need for guidance. J Epidemiol Community Health 2019 Jun;73(6):481-482 [FREE Full text] [doi: 10.1136/jech-2018-210840] [Medline: 30787038]
47. Gardner F, Montgomery P, Knerr W. Transporting evidence-based parenting programs for child problem behavior (age 3-10) between countries: systematic review and meta-analysis. J Clin Child Adolesc Psychol 2016;45(6):749-762. [doi: 10.1080/15374416.2015.1015134] [Medline: 25785902]
48. Leijten P, Melendez-Torres GJ, Knerr W, Gardner F. Transported versus homegrown parenting interventions for reducing disruptive child behavior: a multilevel meta-regression study. J Am Acad Child Adolesc Psychiatry 2016 Jul;55(7):610-617. [doi: 10.1016/j.jaac.2016.05.003] [Medline: 27343888]
49. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new medical research Council guidance. BMJ 2008 Sep 29;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
50. Moore L, Hallingberg B, Wight D, Turley R, Segrott J, Craig P, et al. Exploratory studies to inform full-scale evaluations of complex public health interventions: the need for guidance. J Epidemiol Community Health 2018 Oct;72(10):865-866 [FREE Full text] [doi: 10.1136/jech-2017-210414] [Medline: 30030296]
51. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS One 2016;11(3):e0150205 [FREE Full text] [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]
52. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
53. McLennan D, Noble S, Noble M, Plunkett E, Wright G, Gutacker N. The English indices of deprivation 2019: technical report. Ministry of Housing, Communities and Local Government, United Kingdom Government. 2019 Sep. URL: https://dera.ioe.ac.uk/34259/1/IoD2019_Technical_Report.pdf [accessed 2024-03-25]
54. Schools, pupils and their characteristics: January 2019. Department for Education, United Kingdom Government. 2019 Jun. URL: https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2019 [accessed 2024-03-25]
55. Special educational needs in England: January 2019. Department for Education, United Kingdom Government. 2019 Jul 4. URL: https://assets.publishing.service.gov.uk/media/5d1c91cce5274a08df3d35bd/SEN_2019_Text.docx.pdf [accessed 2024-03-25]
56. Goodman R. The Strengths and Difficulties Questionnaire: a research note. J Child Psychol Psychiatry 1997 Jul 07;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
57. Goodman R. The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. J Child Psychol Psychiatry 1999 Jul;40(5):791-799. [doi: 10.1017/s0021963099004096]
58. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br J Psychiatry 2000 Dec 02;177(6):534-539. [doi: 10.1192/bjp.177.6.534] [Medline: 11102329]
59. Goodman R, Meltzer H, Bailey V. The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. Eur Child Adolesc Psychiatry 1998 Sep 12;7(3):125-130. [doi: 10.1007/s007870050057] [Medline: 9826298]
60. Stone LL, Otten R, Engels RC, Vermulst AA, Janssens JM. Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: a review. Clin Child Fam Psychol Rev 2010 Sep;13(3):254-274 [FREE Full text] [doi: 10.1007/s10567-010-0071-2] [Medline: 20589428]
61. Albright GL, Davidson J, Goldman R, Shockley KM, Timmons-Mitchell J. Development and validation of the Gatekeeper Behavior Scale. Crisis 2016 Jul;37(4):271-280. [doi: 10.1027/0227-5910/a000382] [Medline: 27245815]
62. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
63. Lyon AR, Bruns EJ. From evidence to impact: joining our best school mental health practices with our best implementation strategies. School Ment Health 2019 Mar 1;11(1):106-114 [FREE Full text] [doi: 10.1007/s12310-018-09306-w] [Medline: 31709018]
64. Schools, pupils and their characteristics: January 2020. Department for Education, United Kingdom Government. 2020 Jun 25. URL: https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2020 [accessed 2024-03-25]

XSL•FO
RenderX

65. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
66. Wilson DT, Walwyn RE, Brown J, Farrin AJ, Brown SR. Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. Stat Methods Med Res 2016 Jun;25(3):997-1009 [FREE Full text] [doi: 10.1177/0962280215589507] [Medline: 26071430]
67. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004 May;10(2):307-312. [doi: 10.1111/j..2002.384.doc.x] [Medline: 15189396]
68. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010 Jan 06;10:1 [FREE Full text] [doi: 10.1186/1471-2288-10-1] [Medline: 20053272]
69. R: a language and environment for statistical computing. The Comprehensive R Archive Network. 2015 Feb 10. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2024-03-25]
70. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer; 2009.
71. Bryer J, Speerschneider K. Analysis and visualization of Likert based items. GitHub. URL: http://github.com/jbryer/likert
72. YouthinMind homepage. YouthinMind. URL: https://youthinmind.com/ [accessed 2024-03-26]
73. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
74. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
75. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
76. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications; 2021.
77. School workforce in England. United Kingdom Government. 2023 Jun 8. URL: https://explore-education-statistics.service.gov.uk/find-statistics/school-workforce-in-england [accessed 2024-03-14]
78. Jorm AF, Kitchener BA, Sawyer MG, Scales H, Cvetkovski S. Mental health first aid training for high school teachers: a cluster randomized trial. BMC Psychiatry 2010 Jun 24;10:51 [FREE Full text] [doi: 10.1186/1471-244X-10-51] [Medline: 20576158]
79. Moor S, Maguire A, McQueen H, Wells EJ, Elton R, Wrate R, et al. Improving the recognition of depression in adolescence: can we teach the teachers? J Adolesc 2007 Feb;30(1):81-95. [doi: 10.1016/j.adolescence.2005.12.001] [Medline: 16500701]
80. Kidger J, Stone T, Tilling K, Brockman R, Campbell R, Ford T, et al. A pilot cluster randomised controlled trial of a support and training intervention to improve the mental health of secondary school teachers and students - the WISE (Wellbeing in Secondary Education) study. BMC Public Health 2016 Oct 06;16(1):1060 [FREE Full text] [doi: 10.1186/s12889-016-3737-y] [Medline: 27716226]
81. Vieira MA, Gadelha AA, Moriyama TS, Bressan RA, Bordin IA. Evaluating the effectiveness of a training program that builds teachers' capability to identify and appropriately refer middle and high school students with mental health problems in Brazil: an exploratory study. BMC Public Health 2014 Feb 28;14:210 [FREE Full text] [doi: 10.1186/1471-2458-14-210] [Medline: 24580750]
82. McLaughlin RJ, Holcomb JD, Jibaja-Rusth ML, Webb J. Teacher ratings of student risk for substance use as a function of specialized training. J Drug Educ 1993;23(1):83-95. [doi: 10.2190/eugt-g9ha-rm65-glm6]
83. Soneson E, Childs-Fegredo J, Anderson JK, Stochl J, Fazel M, Ford T, et al. Acceptability of screening for mental health difficulties in primary schools: a survey of UK parents. BMC Public Health 2018 Dec 22;18(1):1404 [FREE Full text] [doi: 10.1186/s12889-018-6279-7] [Medline: 30577830]
84. Levitt JM, Saka N, Hunter Romanelli L, Hoagwood K. Early identification of mental health problems in schools: the status of instrumentation. J Sch Psychol 2007 Apr;45(2):163-191. [doi: 10.1016/j.jsp.2006.11.005]
85. Loades ME, Mastroyannopoulou K. Teachers' recognition of children's mental health problems. Child Adolesc Ment Health 2010 Sep;15(3):150-156. [doi: 10.1111/j.1475-3588.2009.00551.x] [Medline: 32847228]
86. Anderson JK, Howarth E, Vainre M, Jones PB, Humphrey A. A scoping literature review of service-level barriers for access and engagement with mental health services for children and young people. Child Youth Serv Rev 2017 Jun;77:164-176. [doi: 10.1016/j.childyouth.2017.04.017]
87. Moon J, Williford A, Mendenhall A. Educators' perceptions of youth mental health: implications for training and the promotion of mental health services in schools. Child Youth Serv Rev 2017 Feb;73:384-391. [doi: 10.1016/j.childyouth.2017.01.006]
88. Graham A, Phelps R, Maddison C, Fitzgerald R. Supporting children’s mental health in schools: teacher views. Teach Teach 2011;17(4):479-496. [doi: 10.1080/13540602.2011.580525]
89. Atkins MS, Graczyk PA, Frazier SL, Abdul-Adil J. Toward a new model for promoting urban children's mental health: accessible, effective, and sustainable school-based mental health services. Sch Psychol Rev 2003;32(4):503-514. [doi: 10.1080/02796015.2003.12086214]

XSL•FO
RenderX

90. Moore A, Stapley E, Hayes D, Town R, Deighton J. Barriers and facilitators to sustaining school-based mental health and wellbeing interventions: a systematic review. Int J Environ Res Public Health 2022 Mar 17;19(6):3587 [FREE Full text] [doi: 10.3390/ijerph19063587] [Medline: 35329276]
91. Gee B, Wilson J, Clarke T, Farthing S, Carroll B, Jackson C, et al. Review: delivering mental health support within schools and colleges - a thematic synthesis of barriers and facilitators to implementation of indicated psychological interventions for adolescents. Child Adolesc Ment Health 2021 Feb;26(1):34-46. [doi: 10.1111/camh.12381] [Medline: 32543016]
92. Fält E, Wallby T, Sarkadi A, Salari R, Fabian H. Agreement between mothers', fathers', and teachers' ratings of behavioural and emotional problems in 3-5-year-old children. PLoS One 2018 Nov 01;13(11):e0206752 [FREE Full text] [doi: 10.1371/journal.pone.0206752] [Medline: 30383861]
93. Bear HA, Ayala Nunes L, DeJesus J, Liverpool S, Moltrecht B, Neelakantan L, et al. Determination of markers of successful implementation of mental health apps for young people: systematic review. J Med Internet Res 2022 Nov 09;24(11):e40347 [FREE Full text] [doi: 10.2196/40347] [Medline: 36350704]
94. Durbin K, Harmon S. We are on the same team: child psychiatry and the school system. JAACAP Connect 2020;7(1):32-36. [doi: 10.62414/001c.92411]

Abbreviations

SDQ: Strengths and Difficulties Questionnaire
SEMH: social, emotional, and mental health
SENCo: special educational needs coordinator
SLT: senior leadership team
TA: teaching assistant

Edited by T Leung; submitted 24.02.23; peer-reviewed by E Widnall, B Fernandes, J Burns, K Cohen; comments to author 27.08.23; accepted 01.03.24; published 23.04.24.

Please cite as:
Soneson E, Howarth E, Weir A, Jones PB, Fazel M
Empowering School Staff to Support Pupil Mental Health Through a Brief, Interactive Web-Based Training Program: Mixed Methods Study
J Med Internet Res 2024;26:e46764
URL: https://www.jmir.org/2024/1/e46764
doi:10.2196/46764
PMID:38652534

©Emma Soneson, Emma Howarth, Alison Weir, Peter B Jones, Mina Fazel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Video-Based Microinterventions on Attitudes Toward Mental Health and Help Seeking in Youth: Web-Based Randomized Controlled Trial

Diana Lemmer[1,2], MSc; Markus Moessner[1], PD, PhD; Nicolas Arnaud[3], PhD; Harald Baumeister[4], Prof Dr; Agnes Mutter[4], MSc; Sarah-Lena Klemm[5], BSc; Elisa König[6], Dipl-Psych; Paul Plener[6,7], Prof Dr Med; Christine Rummel-Kluge[5], Prof Dr Med; Rainer Thomasius[3], Prof Dr Med; Michael Kaess[8,9], Prof Dr Med; Stephanie Bauer[1,2,10], Prof Dr

[1]Center for Psychotherapy Research, Center for Psychosocial Medicine, University Hospital Heidelberg, Heidelberg, Germany
[2]Ruprecht-Karls University Heidelberg, Heidelberg, Germany
[3]German Centre for Addiction Research in Childhood and Adolescence, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany
[4]Department of Clinical Psychology and Psychotherapy, Ulm University, Ulm, Germany
[5]Department of Psychiatry and Psychotherapy, University of Leipzig Medical Center, Leipzig, Germany
[6]Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital Ulm, Ulm, Germany
[7]Department of Child and Adolescent Psychiatry, Medical University of Vienna, Vienna, Austria
[8]University Hospital of Child and Adolescent Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland
[9]Clinic of Child and Adolescent Psychiatry, Center for Psychosocial Medicine, University Hospital Heidelberg, Heidelberg, Germany
[10]German Center for Mental Health (DZPG), Partner site Mannheim/Heidelberg/Ulm, Heidelberg, Germany

Corresponding Author:
Stephanie Bauer, Prof Dr
Center for Psychotherapy Research
Center for Psychosocial Medicine
University Hospital Heidelberg
Bergheimer Str. 54
Heidelberg, 69115
Germany
Phone: 49 6221 56 7345
Fax: 49 6221 56 7350
Email: stephanie.bauer@med.uni-heidelberg.de

Abstract

Background: Mental health (MH) problems in youth are prevalent, burdening, and frequently persistent. Despite the existence of effective treatment, the uptake of professional help is low, particularly due to attitudinal barriers.

Objective: This study evaluated the effectiveness and acceptability of 2 video-based microinterventions aimed at reducing barriers to MH treatment and increasing the likelihood of seeking professional help in young people.

Methods: This study was entirely web based and open access. The interventions addressed 5 MH problems: generalized anxiety disorder, depression, bulimia, nonsuicidal self-injury, and problematic alcohol use. Intervention 1 aimed to destigmatize and improve MH literacy, whereas intervention 2 aimed to induce positive outcome expectancies regarding professional help seeking. Of the 2435 participants who commenced the study, a final sample of 1394 (57.25%) participants aged 14 to 29 years with complete data and sufficient durations of stay on the video pages were randomized in a fully automated manner to 1 of the 5 MH problems and 1 of 3 conditions (control, intervention 1, and intervention 2) in a permuted block design. After the presentation of a video vignette, no further videos were shown to the control group, whereas a second, short intervention video was presented to the intervention 1 and 2 groups. Intervention effects on self-reported potential professional help seeking (primary outcome), stigma, and attitudes toward help seeking were examined using analyses of covariance across and within the 5 MH problems. Furthermore, we assessed video acceptability.

XSL•FO
RenderX

Results: No significant group effects on potential professional help seeking were found in the total sample ($F_{2,1385}$=0.99; P=.37). However, the groups differed significantly with regard to stigma outcomes and the likelihood of seeking informal help ($F_{2,1385}$=3.75; P=.02). Furthermore, separate analyses indicated substantial differences in intervention effects among the 5 MH problems.

Conclusions: Interventions to promote help seeking for MH problems may require disorder-specific approaches. The study results can inform future research and public health campaigns addressing adolescents and young adults.

Trial Registration: German Clinical Trials Register DRKS00023110; https://drks.de/search/de/trial/DRKS00023110

(J Med Internet Res 2024;26:e54478) doi:10.2196/54478

KEYWORDS

help seeking; mental health; stigma; mental health literacy; psychoeducation; web-based experiment; web-based randomized controlled trial; microinterventions; video-based interventions

Introduction

Background

Mental health (MH) problems in youth are prevalent and pose severe health-related, social, and financial burdens on individuals [1-5] and societies [6,7]. Approximately half of all mental disorders first manifest before the age of 18 years [8], and MH problems in youth often persist and aggravate over the life span [9-14]. Therefore, the need for effective prevention and intervention programs targeting young people is an important public health goal. However, while effective MH services exist, most youth with MH problems do not seek professional help. Low uptake has been reported for various conventional [15-18] as well as digital MH services [19-22]. The burden of mental illness can only be alleviated at the population level if a substantial proportion of the population uses the available services [23,24]. Otherwise, the public health impact of MH services remains limited. Thus, increasing the reach of MH services (ie, fostering the uptake and use of professional help) is vital for the improvement of youth MH at the population level.

To facilitate service use, specific barriers to help seeking need to be addressed. Previous research has indicated that attitudinal factors pose larger impediments to help seeking than structural factors (eg, treatment costs and inconvenient scheduling) [25,26]. Specifically, self-reliance, a low perceived need for help [25-27], low treatment expectations [28], stigma [26,27,29,30], and poor MH literacy [27,31,32] have been identified as major contributors to the lack of professional help seeking.

Different approaches to facilitate help seeking and promote positive attitudes toward MH issues and help seeking in youth have been evaluated in previous research, including face-to-face and digital interventions. In a systematic mapping review, 84% (106/126) of the studies focused on school-based interventions, whereas only 10 (8%) articles covered internet-based approaches to improve MH literacy, MH-related attitudes, stigma, and help-seeking behavior in adolescents [33]. The internet-based interventions included both minimal, single-session interventions [34,35] and multisession approaches intended to be used over several weeks [36,37], with different outcome measures. A total of 4 studies focused on MH more broadly, whereas 6 studies investigated interventions for specific MH problems (depression: n=5; eating disorders: n=1). Keeping the limited number of studies in this area of research in mind, the results nevertheless point to the potential of internet-based interventions with respect to reduced stigma (2 studies), enhanced help-seeking intentions (2 studies), and improved help-seeking behaviors (1 study).

Clearly, there is a need for more research in this area, particularly with respect to digital brief and microinterventions (ie, highly focused in-the-moment interventions with a narrower scope and time frame than standard interventions [38]), which allow for a flexible, easily accessible, scalable, and efficient delivery of MH content. Initial research on such brief and microinterventions with psychoeducational and destigmatizing components has shown promising results. For instance, a brief acceptance-facilitating intervention that included a text-based personalized psychoeducation component had a small but significant effect on the intention to use MH services in German university students [39]. More recently, randomized controlled trials (RCTs) in young adults, university students, and adolescents with short video interventions demonstrated effects with regard to public stigma toward schizophrenia [40,41] and depression [42,43], as well as help-seeking intentions [42] and attitudes [43]. Furthermore, an Australian pilot study with international students found that a brief, web-based MH literacy intervention alleviated MH stigma. However, it had no significant effect on help-seeking intentions or MH literacy [44].

Another component of previous help seeking–facilitating strategies has been storytelling. A pilot study on a video-based intervention indicated that storytelling was well accepted and perceived as engaging [45]. In addition, an RCT evaluated internet-based storytelling programs with varying interactivity and stigma-related content. Significant reductions in MH stigma and microaggression toward individuals with MH problems were observed [46].

Concerning the theoretical foundation of interventions, few studies have investigated help seeking–promoting strategies that were explicitly based on the premises of health behavior models. Logsdon et al [47] evaluated an internet-based depression intervention for adolescent mothers, which was conceptualized according to the theory of planned behavior. The intervention led to significant improvements in help-seeking attitudes, intentions, and behavior. Another well-established and yet more recent health behavior model, which incorporates elements of previously developed approaches, is the Health Action Process Approach (HAPA) [48]. It encompasses a stage

XSL•FO
RenderX

theoretical perspective on health behavior and includes a motivational, intention-forming phase as well as a volitional phase, where planning and behavior maintenance occur. In both the HAPA model and the updated version of the theory of planned behavior, namely, the reasoned action approach, outcome expectancies (or instrumental attitudes) play a crucial role in the formation of intentions, and intentions significantly predict actual behavior [49,50]. The results of previous research on a trauma recovery internet intervention support the use of the HAPA model for the prediction of e-MH engagement. Specifically, outcome expectations significantly predicted the intention to use the intervention (β=.36) [51]. Skepticism about treatment effectiveness has further been identified as a predictor for not using MH services in another study with university students [28].

Building on the findings of previous research, this study investigates the short-term effectiveness of 2 brief animated video interventions to promote potential professional help seeking in a general sample of adolescents and young adults aged 14 to 29 years using a web-based RCT approach. Both interventions aimed to improve participants' willingness to seek professional help (ie, psychotherapists, psychiatrists, and counseling services) for 5 MH problems (generalized anxiety disorder [GAD], depression, bulimia, nonsuicidal self-injury [NSSI], and problematic alcohol use). The inclusion of various MH problems allowed for the investigation of potential differential effects. While one intervention followed a destigmatizing and psychoeducational approach, the other intervention aimed to induce positive outcome expectancies in accordance with the HAPA model through storytelling. The interventions were both compared to each other and to a nonintervention control group (CG) where participants were presented with a stand-alone video vignette without an additional intervention video. This approach was chosen due to both contextual (ie, vignette characters were described as experiencing difficulties in several life domains, and thus, additional control videos referring to the vignettes were unfeasible) and practical (ie, the creation of 10 additional videos was not necessary) considerations.

Objectives

This study had the following objectives:

1. To investigate the short-term effectiveness of the 2 interventions in the promotion of potential MH help seeking (professional and informal), whereby self-reported professional help seeking was defined as the primary outcome.
2. To investigate the interventions' effectiveness in the improvement of self-reported attitudes toward MH problems and MH service use (stigmatization and attitudes toward seeking MH services).
3. To evaluate the interventions' self-reported acceptability.

Within the framework of this study, the videos were evaluated as stand-alone interventions. They were not developed to replace existing interventions. However, in case of favorable outcomes, they have the potential to complement existing health care services. Results and procedures are reported in accordance with the Checklist for Reporting Results of Internet E-Surveys [52] and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) [53]. The study was preregistered at the German Clinical Trials Register on September 23, 2020 (DRKS00023110).

Methods

Study Design

This anonymous, fully automated, web-based, parallel-group exploratory RCT compared the effects of intervention 1 (psychoeducational intervention) and intervention 2 (positive consequences of help seeking) against those of the CG (no further videos after the case vignette) with regard to potential help seeking, attitudes toward help seeking, and stigma. The design comprised 15 conditions in total (5 MH problems × 3 interventional conditions). Randomization was stratified by gender and implemented using a permuted block design (block sizes: 15 and 30). Due to anonymous participation and automated randomization, researchers were unable to assign specific conditions to individuals. However, 2 of the authors were able to view the randomization list. The video material was aligned with the participants' gender to increase identification with the character (ie, participants who identified as woman, female, or nonbinary viewed videos with a female protagonist [Paula], and participants identifying as man or male viewed videos with a male protagonist [Paul]). The study components and conditions as well as the study procedure are shown in Figure 1.

XSL•FO
RenderX

Figure 1. Study design and sample sizes. Participants were directly randomized into 1 of 15 experimental conditions. To enhance clarity and understanding, the figure illustrates the randomization process in 2 steps: mental health issue and experimental condition. GAD: generalized anxiety disorder; INT1: intervention 1; INT2: intervention 2; nincl: included cases (number of subgroup cases with complete data and sufficient duration on video page or video pages that were included in the data analysis); nrand: randomized cases (number of randomized subgroup cases irrespective of data completeness and video page duration); NSSI: nonsuicidal self-injury.

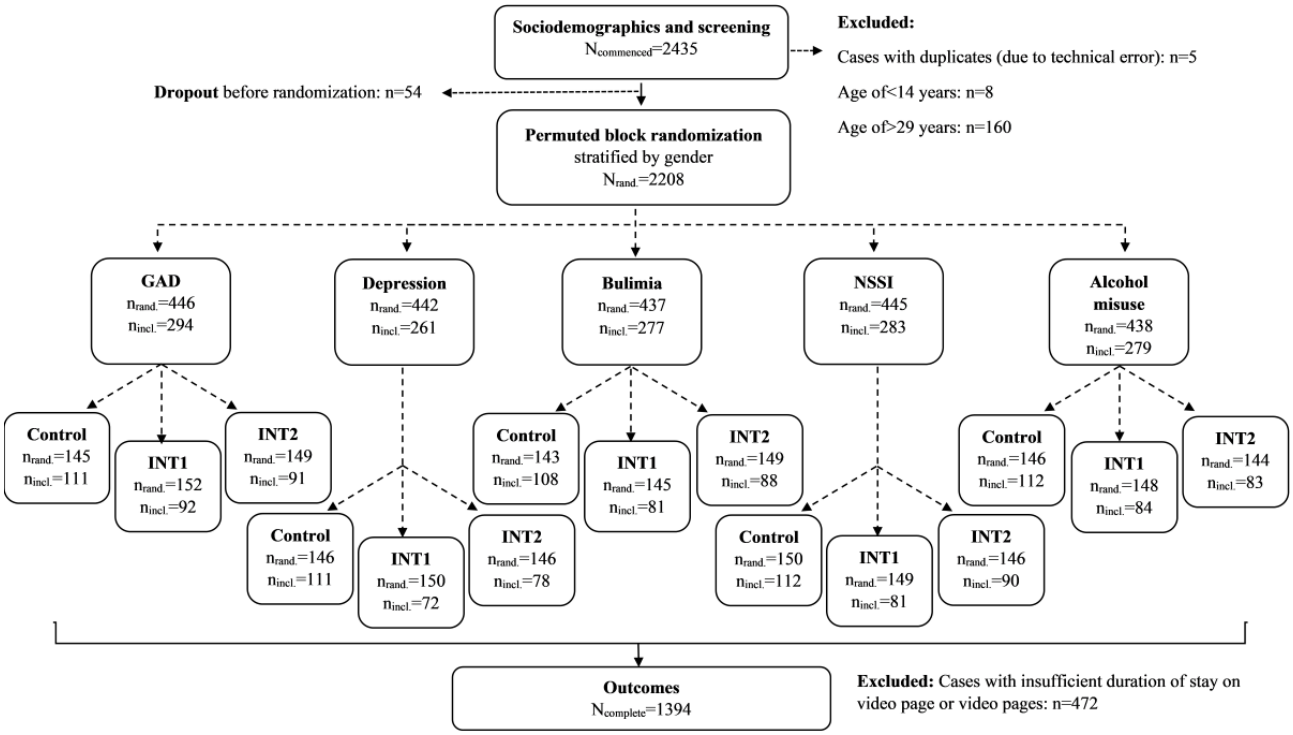

Recruitment and Sample

Recruitment started in October 2020 and ended in May 2022. Youth aged between 14 and 29 years with sufficient German language skills were eligible for participation. The age of 14 years is widely accepted as appropriate to provide informed consent for medical decisions and participation in studies [54,55]. The upper age limit of 29 years aligns with the definition of emerging adulthood, a separate life stage between adolescence and adulthood [56,57]. Participants were primarily recruited through the web on social media platforms and via mailing lists, web-based marketplaces, and forums for adolescents and young adults (eg, accounts and emails of youth clubs and student associations). As an incentive to complete the study, participants were offered to take part in an optional gift card lottery at the end of the study (100 gift cards of €20 [US $21.58]). We asked participants for a valid email address if they were interested in the lottery and stored email addresses separately from other study data and user IDs to ensure anonymous participation.

We recorded page change time stamps. Participants whose time stamp data indicated that the video or the videos they were assigned to had not been fully viewed (ie, duration of stay<length of the respective videos) were excluded from statistical analyses. Furthermore, only data from participants who completed all questionnaires were included in the final analysis (n=1394; completion rate: 1394/2435, 57.25%). We also excluded 5 cases with duplicate user IDs, which occurred due to a technical error and indicated repeated participation (Figure 1). HTTP cookies were used to assign individual user IDs to participants. For each session, new cookies were generated and used. Therefore, duplicate participation was possible after the completion of each study session and was not registered by the system. In the 5 aforementioned cases, duplicate IDs were mistakenly generated when participants tried to use the “back” button of their web browser and restarted their participation.

Procedure

Overview

This study was conducted in an open access, voluntary web-based setting. A website was established to provide study information and enable participation. The ASMO software (Center for Psychotherapy Research) [58] was used to implement the RCT. A randomization list with numbers representing the conditions was generated and embedded in our ASMO database [58] before recruitment. Data were collected at the Center for Psychotherapy Research, Heidelberg. The study’s technical functionality and usability were extensively tested before recruitment by the authors and their colleagues at their respective institutions. Before their participation, the youth received detailed information about the aims, scope, procedures, data processing, and data storage of the study on the website. Participants were informed that they would be randomly assigned to 1 of 5 MH problems and 1 of 3 video versions. They were not informed about the specific health issues or the conditions’ details before participation. As the aim of the conditions was to provide information about a specific MH problem, blinding of participants after assignment to the interventions was not possible. Only participants who provided informed consent through a web-based checkbox were eligible for participation. After study completion, participants were

XSL•FO
RenderX

debriefed in writing about the objectives on the study website. The debriefing form also included contact information for formal help services. Study duration amounted to approximately 30 minutes. Participants were first asked to complete sociodemographic and screening questionnaires; were then randomly assigned to 1 of the 15 experimental conditions; and, finally, were presented with the outcome questionnaires. The whole study (including informed consent and gift card lottery pages) comprised 26 pages with 1 to 12 items on each page. Each segment or measure was presented on 1 or 2 separate pages depending on its respective length. Some items were conditional for adaptive questioning (eg, when lifetime NSSI was denied, no further questions about NSSI were presented). Changes to the item responses could only be made while they had not been confirmed through a click on the "next" button, which brought participants to the next page. There was no "back" button.

Sociodemographics and Screening

All measures were self-reported. The sociodemographic form asked participants about their age, gender, migration background, education, whether they knew someone with MH problems, and participants' previous or current MH service use (actual help seeking). Thereafter, participants' subjective psychological distress was assessed using several screening instruments.

Anxiety symptoms were measured using the 7-item Generalized Anxiety Disorder Scale (GAD-7) [59]. Symptom frequency within the previous 2 weeks was indicated on a 4-point response scale. Total scores (potential range 0-21) were used for further analyses. Scores of ≥5 indicate a mild anxiety symptomatology, scores of ≥10 indicate a moderate anxiety symptomatology, and scores of ≥15 indicate a severe anxiety symptomatology [59].

The 9-item Patient Health Questionnaire (PHQ-9) [60] was used for depression symptomatology assessment. Frequencies of depression symptoms within the previous 2 weeks were measured on a 4-point scale. Total scores (potential range 0-27) were calculated for further analyses. Total scores of ≥5 were interpreted as mild, scores of ≥10 were interpreted as moderate, scores of ≥15 were interpreted as moderately severe, and scores of ≥20 were interpreted as severe depression symptomatology [60].

The Weight Concerns Scale (WCS) [61,62] assessed weight and body shape concerns. It consists of 5 items with varying response scale types (4- to 7-point scales). The response categories of each item represent scores between 0 and 100. The mean across all items was used for further analyses. Scores of ≥57 are indicative of a high risk of eating disorders [61].

Problematic alcohol use during the previous 12 months was measured using the Alcohol Use Disorders Identification Test for Consumption (AUDIT-C) [63,64]. It comprises 3 items with 5-point response scales. Sum scores range between 0 and 12. A score of 0 indicates abstinence, whereas scores between 1 and 3 are interpreted as moderate alcohol consumption. Scores of ≥4 indicate hazardous alcohol consumption [63,65].

A total of 4 items of the Self-Injurious Thoughts and Behaviors Interview [66] served to assess NSSI. The first item identified whether participants had ever harmed themselves without suicidal intention. If participants reported lifetime NSSI, the 3 subsequent questions were presented. These items measured the frequency of NSSI within the last year, the age at the first occurrence of NSSI, and the age at the last occurrence of NSSI. Item responses were analyzed separately and descriptively.

Experimental Conditions and Materials

Overview

The interventional strategies were applied using short animated videos. The videos were created with the Pro+ version of the web-based animation tool Powtoon (Powtoon Limited) [67]. Each research group involved in this study prepared materials for 1 of the 5 MH problems based on their respective field of expertise. The materials were structured in a similar fashion across MH problems. The main characters, Paul and Paula, were introduced as students aged 16 years in each condition. In total, 30 videos were created: 5 MH problems × 2 main character genders × 3 video types. Participants in the control condition only viewed a vignette, whereas participants in both intervention groups each viewed 1 additional video (either for intervention 1 or intervention 2). A subset of the videos was pretested between July 2020 and September 2020 with a convenience sample of 9 youths (mean age 18.56, SD 3.74 years; range 14-24 years; 3/9, 33% male), who confirmed comprehensibility and overall acceptability.

Vignettes

All participants viewed a case vignette. Each vignette depicted the respective main character, who was affected by 1 of 5 MH problems (GAD, depression, bulimia, NSSI, or problematic alcohol use). The vignettes introduced the characters to the viewers in a third-person perspective and described their challenges in their everyday lives due to their MH conditions (eg, difficult emotions and cognitions, physical symptoms, and social and school-related issues). The accurate diagnostic labels were not presented in the vignettes [68]. Vignette duration ranged from 2 minutes, 19 seconds to 2 minutes, 47 seconds (mean 2 min, 29 s; SD 11 s). The bulimia vignettes were developed first. They were inspired by the vignettes by Mond et al [69] and adapted in accordance with *International Classification of Diseases, 10th Revision* and *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* diagnostic criteria, as well as further literature on the symptomatology and psychological strain of bulimia [70]. The bulimia vignettes then served as a template for the vignettes of the other 4 MH problems.

Intervention 1

Intervention 1 aimed to improve MH literacy and decrease stigmatization through the presentation of psychoeducational information to encourage help seeking. These intervention videos first presented the correct diagnostic label, prevalence rates, and core symptoms of the condition shown in the vignette. Next, 5 destigmatizing and psychoeducational facts about the respective condition were presented (eg, "Bulimia is a serious illness and not a lifestyle"), which were inspired by the work by Bulik [71]. The videos then presented treatment options, information about potential challenges in professional help seeking, and encouraging statements about the benefits of

XSL•FO
RenderX

professional MH support. Intervention 1 video durations ranged from 4 to 5 minutes (mean 4 min, 27 s; SD 21 s). The information provided in these intervention videos was based on epidemiological, etiological, diagnostic, barrier-related, and interventional findings on the respective MH problems (eg, the studies by Bulik [71], Keski-Rahkonen and Mustelin [72], and Nagl et al [73] for bulimia).

Intervention 2

The second strategy (intervention 2) was based on the premises of the HAPA [48]. Intervention 2 was designed to induce positive outcome expectancies of professional help seeking through the continuation of Paul and Paula's stories. The videos showed the main characters 1 year after their initial situation as described in the vignettes. Intervention 2 videos first demonstrated the help-seeking process of the main characters in a retrospective fashion. Encouraged by their teachers, friends, or parents, the main characters sought and received professional support from a psychotherapist. The psychotherapist's gender matched the gender of the main character. The videos showed how the psychotherapist informed the main character about the correct diagnostic label of their condition and shortly portrayed the therapeutic process. The process included initial difficulties of the main character, such as feelings of insecurity about disclosing their experiences to their therapist, which were resolved over time, and the main characters became invested in their psychotherapy. Then, 5 positive consequences of psychotherapy were presented, such as decreased impairment and an improved quality of life. The videos ended with the notion that the main character was still facing occasional difficulties, but substantial improvements in overall well-being and satisfaction with their decision to seek help were emphasized. Intervention 2 video durations ranged from 4 minutes, 1 second to 4 minutes, 29 seconds (mean 4 min, 15 s; SD 14 s). These interventions were designed in accordance with previous literature on the therapeutic process in MH conditions, including treatment expectations, experiences, and consequences [74].

Outcome Measures

Primary Outcome Measure

Our primary outcome was the potential use of professional MH services (ie, the hypothetical likelihood of seeking formal sources of help if participants experienced Paul's or Paula's MH problem), measured using a 12-item version of the General Help Seeking Questionnaire (GHSQ) [75]. The GHSQ measures the willingness of seeking various formal and informal sources of help within the next 4 weeks for an indicated MH problem on a 7-point rating scale (1="extremely unlikely"; 7="extremely likely"). The maximum score among the 3 items, which measured potential help seeking with professional MH services (psychotherapists, psychiatrists, and counseling services), was used as our primary outcome. The GHSQ is the most frequently used instrument for help seeking [76].

Secondary Outcome Measures

GHSQ data on the potential use of informal sources (romantic partner, friend, parent, or other family member) and no intended help seeking (1 item) were used as secondary outcomes. For informal sources of support, the items' maximum score was used for the analyses.

Attitudes toward help seeking were measured using the Inventory of Attitudes Toward Seeking Mental Health Services (IASMHS) [77] on a 5-point rating scale. It comprises 24 items. Its 3 dimensions—"psychological openness," "help-seeking propensity," and "indifference to stigma"—are represented with 8 items. Subscale scores range from 0 to 32. Higher scores indicate more positive attitudes.

The Universal Stigma Scale (USS) [78] was used for stigma measurement. It consists of 11 statements in 2 subscales ("blame/personal responsibility": 5 items; "impairment/distrust": 6 items). The extent of agreement with these statements is indicated on a 5-point Likert scale. Means were calculated for each of the 2 subscales. Lower scores indicate higher stigmatization.

Transportation (ie, the immersiveness of the stories presented in the videos) was measured using an adapted version of the Transportation Scale–Short Form [79]. Adjustments were made to suit the medium of the narratives (ie, video material in contrast to written stories). Our adapted version contained 5 items on a 7-point Likert scale.

Video acceptability was measured using a translated and adapted 4-item version of the acceptability and likability scale used by Gaudiano et al [45]. In total, 3 items measured overall likability, comprehensibility, and interestingness of the videos on a 5-point rating scale.

Statistical Analysis

Sociodemographic, screening, and outcome data were first analyzed descriptively. Intervention effects on potential professional help seeking (primary outcome) and secondary outcomes in the total sample (ie, across all MH problems and across participants with and without actual help seeking as reported in the screening) were analyzed via analyses of covariance (ANCOVAs) at an α level of $P<.05$. In addition to the intervention group, the models included age as a covariate, the participants' actual help seeking (fixed effects), and the 5 MH problems (random effects) as control variables. The results of the main ANCOVA in the total sample are presented in the Results section.

Subgroup ANCOVAs were conducted for each of the 5 MH problems separately. In this case, the respective screening scores (GAD-7, PHQ-9, WCS, number of NSSI events during the last year, and AUDIT-C) were included as additional covariates. Subgroup analyses were further conducted for cases with and without actual help seeking in the total sample and within each of the 5 MH issue groups.

In case of significant ($P<.05$) and trend ANCOVA effects, pairwise group comparisons were conducted using 2-tailed t tests. All tests were 2-sided with an α level of 5%. Mean differences (MDs) adjusted for covariates are reported in the Results section.

An a priori power analysis was conducted using G*Power (Heinrich-Heine-Universität Düsseldorf) [80]. Under the assumption of a medium effect size (f=0.25), a minimum of 240

XSL•FO
RenderX

participants (80 per condition) were needed to test the expected effect within each of the 5 MH problems via ANCOVAs with a significance criterion of α=.05 and a power of 90%. Statistical analyses were performed using R (version 4.3; R Foundation for Statistical Computing) [81] and SPSS (version 28; IBM Corp) [82]. R was also used to generate the random allocation sequence. Authors involved in data analysis and interpretation were not blinded with respect to the assigned experimental conditions.

Deviations From the Protocol

In the beginning of recruitment, the upper age limit was raised from 25 years originally to 29 years due to the aforementioned findings of previous research.

Ethical Considerations

Ethics approval was obtained from Ethics Committee I of the Heidelberg Medical Faculty on July 27, 2020 (protocol S378/2020). The procedures were in accordance with the Helsinki Declaration of 1975, as revised in 2000. All participants received information about the study's aims, scope, procedures, data processing, and data storage on the study website in written form. Furthermore, all participants received contact information if they wished to clarify questions via telephone or email. Only participants who provided their informed consent through a web-based checkbox were eligible for participation. Participants were able to opt out of the study at any time by closing the study website, which they were informed of before their participation. Participants were offered to take part in an optional gift card lottery at the end of the study (100 gift cards of €20 [US $21.58] each). If they were interested in the lottery, they could enter their email address. Email addresses were stored separately from other study data and user IDs to ensure anonymity. All other data were collected and are reported anonymously. Thus, this study does not contain any individual data of identifiable participants.

Results

Sample Characteristics

Figure 1 shows the flow of participants. Of the 2208 participants who were randomized to 1 of the 15 conditions, 472 (21.38%) were excluded because their time spent on the video pages fell below the durations of the videos they were assigned to, indicating that they did not view the entire videos. Of the remaining 1736 participants, 342 (19.7%) were excluded due to incomplete data (ie, they did not complete all the relevant scales that the study entailed [beginning with informed consent up to and including the last acceptability item]). Our final sample consisted of 1394 youths aged 14 to 29 years (mean 20.97, SD 3.67 years). Sociodemographic and screening results are shown in Table 1.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics and screening results (n=1394).

Characteristics	Values
Age (years), mean (SD)	20.97 (3.67)
Gender, n (%)	
Girl and woman	1109 (79.6)
Boy and man	254 (18.2)
Nonbinary	31 (2.2)
Actual help seeking, n (%)	
None	770 (55.2)
Current	273 (19.6)
Past	351 (25.2)
Knew someone with MH[a] problems, n (%)	
Yes	1285 (92.2)
No	109 (7.8)
Education (highest secondary school diploma), n (%)	
Still in school	264 (18.9)
No diploma	0 (0)
Lower secondary school diploma (*Hauptschulabschluss*)	8 (0.6)
Intermediate secondary school diploma (*Realschulabschluss*)	89 (6.4)
Higher secondary school diploma (*Abitur*)	1016 (72.9)
Other	17 (1.2)
Migration background, n (%)	
None	1104 (79.2)
First generation or not born in Germany	64 (4.6)
Second generation	226 (16.2)
GAD-7[b] score, mean (SD)	8.38 (5.00)
Minimal or no anxiety (0-4)	370 (26.5)
Mild (5-9)	501 (35.9)
Moderate (10-14)	330 (23.7)
Severe (≥15)	193 (13.9)
PHQ-9[c] score, mean (SD)	9.56 (6.07)
Minimal or no depression (0-4)	333 (23.9)
Mild (5-9)	430 (30.9)
Moderate (10-14)	347 (24.9)
Moderately severe (15-19)	171 (12.3)
Severe (≥20)	113 (8.1)
WCS[d] score, mean (SD)	34.50 (24.65)
High risk (≥57)	273 (19.6)
Low risk (<57)	1121 (80.4)
SITBI-G[e] score, mean (SD)	
Lifetime NSSI[f]	479 (34.4)
12-month NSSI	265 (19.0)

XSL•FO
RenderX

Characteristics	Values
Number of NSSI events (previous 12 months; lifetime NSSI sample [n=479])	11.98 (45.36)
Number of NSSI events (previous 12 months; 12-month NSSI sample [n=265])	21.66 (59.29)
Age of first NSSI (n=479)	14.20 (3.01)
Age of last NSSI – age of first NSSI (n=475)[g]	3.92 (3.71)
AUDIT-C[h] score, mean (SD)	2.51 (2.08)
Abstinent (0)	319 (22.9)
Moderate (1-3)	645 (46.3)
Hazardous (≥4)	430 (30.9)

[a]MH: mental health.

[b]GAD-7: 7-item Generalized Anxiety Disorder Scale.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]WCS: Weight Concerns Scale.

[e]SITBI-G: German version of the Self-Injurious Thoughts and Behaviors Interview.

[f]NSSI: nonsuicidal self-injury.

[g]We excluded 4 cases in “Age of last NSSI – age of first NSSI” due to invalid values (age of first NSSI>age of last NSSI).

[h]AUDIT-C: Alcohol Use Disorders Identification Test for Consumption.

A total of 79.56% (1109/1394) of the sample identified as woman or girl, and 44.76% (624/1394) were help seekers (ie, they used professional MH services at the time of or before data collection). On average, the youth were moderately anxious (mean GAD-7 score 8.38, SD 5.00) and depressed (mean PHQ-9 score 9.56, SD 6.07). While 22.88% (319/1394) reported abstinence in the AUDIT-C, 30.85% (430/1394) indicated hazardous alcohol consumption. A total of 19.58% (273/1394) were at high risk of developing an eating disorder according to the WCS. One-third (479/1394, 34.36%) of the sample reported a lifetime history of NSSI according to the Self-Injurious Thoughts and Behaviors Interview, with a 12-month prevalence rate of 19.01% (265/1394).

Intervention Effects

The main results are presented in Table 2.

XSL•FO
RenderX

Table 2. Analysis of covariance results and pairwise comparisons for primary outcomes (total sample; n=1394)[a].

	Total, mean (SD)	CG[b] (n=554), mean (SD)	Intervention 1 (n=410), mean (SD)	Intervention 2 (n=430), mean (SD)	*F* test (*df*=2, 1385)	*P* value	Pairwise comparisons
Potential help seeking (GHSQ[c])[d]							
Professional maximum	4.74 (1.81)	4.65 (1.88)	4.82 (1.74)	4.76 (1.79)	0.99	.37	—[e]
Informal maximum	5.86 (1.38)	5.87 (1.37)	5.73 (1.43)	5.95 (1.35)	3.75	.02	Intervention 2>intervention 1
None	3.07 (2.01)	3.08 (2.02)	3.20 (1.99)	2.94 (2.01)	2.68	.07	—
Stigma (USS[f])[g]							
Blame	4.47 (0.64)	4.41 (0.68)	4.50 (0.65)	4.50 (0.58)	3.25	.04	Intervention 1; intervention 2>CG
Distrust	3.94 (0.79)	3.84 (0.79)	3.97 (0.79)	4.02 (0.78)	8.01	<.001	Intervention 1; intervention 2>CG
Help-seeking attitudes (IASMHS[h])[g]							
Psychological openness	21.20 (4.82)	21.05 (4.84)	21.65 (4.70)	20.97 (4.89)	1.67	.19	—
Help seeking propensity	20.95 (5.29)	20.78 (5.41)	21.04 (5.14)	21.09 (5.27)	0.48	.62	—
Indifference to stigma	23.37 (6.31)	23.83 (6.06)	23.02 (6.51)	23.13 (6.40)	3.18	.04	CG>intervention 1
Video acceptability and transportation[d]							
General likability	3.94 (0.81)	3.85 (0.78)	4.10 (0.79)	3.90 (0.84)	12.20	<.001	Intervention 1>CG; intervention 2
Comprehensibility	4.82 (0.45)	4.79 (0.50)	4.82 (0.45)	4.85 (0.40)	2.01	.13	—
Interestingness	3.86 (0.96)	3.84 (0.93)	3.98 (0.94)	3.76 (0.99)	6.39	.002	Intervention 1>CG; intervention 2
Transportation (TS-SF[i])	4.49 (1.24)	4.53 (1.21)	4.58 (1.25)	4.37 (1.27)	4.23	.02	CG; intervention 1>intervention 2

[a]Results controlled for help seeking (fixed factor), mental health issue (random factor), and age (covariate).

[b]CG: control group.

[c]GHSQ: General Help Seeking Questionnaire.

[d]Higher scores represent a greater level of agreement.

[e]Pairwise comparisons were conducted in case of significant or trend analysis of covariance effects. Empty cells indicate that pairwise comparisons were not conducted due to the analysis of covariance results.

[f]USS: Universal Stigma Scale.

[g]Higher scores represent more positive attitudes toward mental health issues and help seeking.

[h]IASMHS: Inventory of Attitudes Toward Seeking Mental Health Services.

[i]TS-SF: Transportation Scale–Short Form.

Figure 2 summarizes the results of the overall efficacy and the MH issue–specific subgroup analyses graphically. Specific results of the subgroup analyses can be found in Multimedia Appendices 1, 2, and 3.

XSL•FO
RenderX

Figure 2. Analysis of covariance (ANCOVA) results (overview). Black spots represent group effects on the total sample (with and without previous help seeking), gray spots represent group effects on one or both subsamples, dotted circles indicate nonsignificant trend group differences with $P<.06$ in the ANCOVAs, and striped areas indicate nonsignificant trend group differences with $P<.06$ in the post hoc comparisons. Asterisks indicate more favorable outcomes for the control group (CG) in comparison to both intervention groups. In the GHSQ, higher scores represent a greater level of agreement. In the other outcome measures, higher scores represent more positive attitudes toward MH issues and help seeking. b: both subsamples (with and without previous help seeking); GAD: generalized anxiety disorder; GHSQ: General Help Seeking Questionnaire; HSP: help seeking propensity; IASMHS: Inventory of Attitudes Toward Seeking Mental Health Services; INT1: intervention 1; INT2: intervention 2; ITS: indifference to stigma; MH: mental health; NSSI: nonsuicidal self-injury; PO: psychological openness; USS: Universal Stigma Scale; w/o: without previous help seeking; w: with previous help seeking.

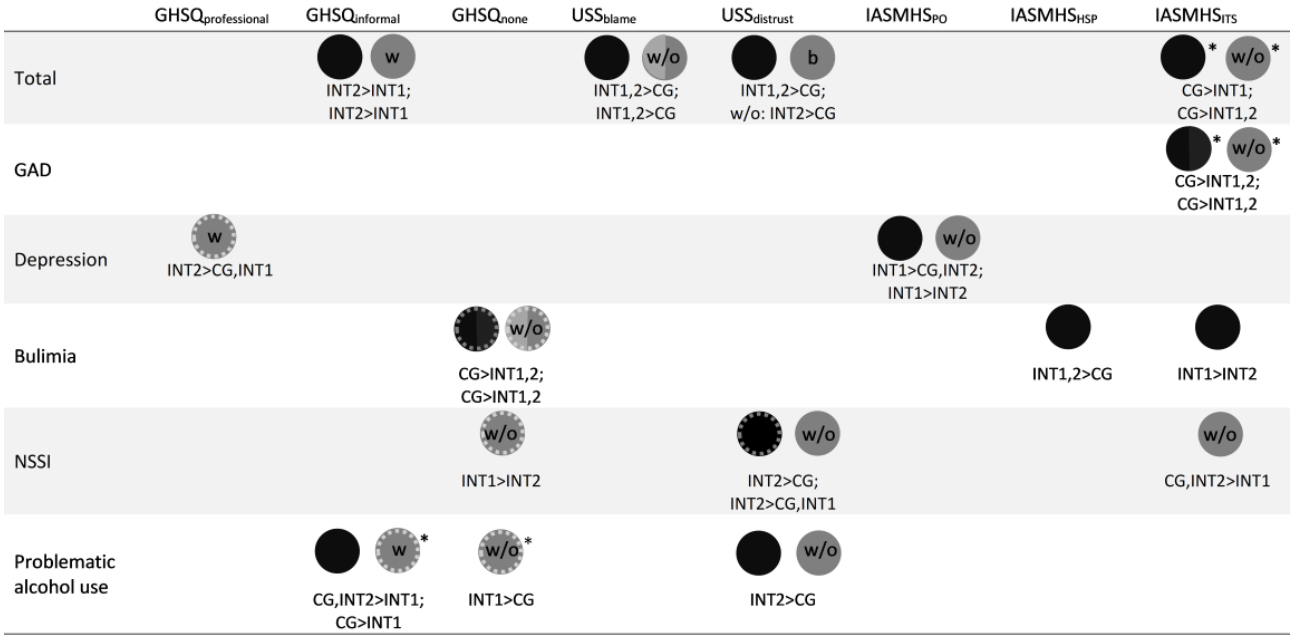

Primary Outcome: Potential Professional Help Seeking (GHSQ)

On the 7-point scale of the GHSQ, most participants (1046/1394, 75.04%) selected a score of ≥4 (CG: 409/554, 73.8%; intervention 1: 316/410, 77.1%; intervention 2: 321/430, 74.7%). In total, 19.23% (268/1394; CG: 105/554, 19%; intervention 1: 74/410, 18%; intervention 2: 89/430, 20.7%) of participants reported a score of 7 ("extremely likely"), whereas 6.74% (94/1394; CG: 47/554, 8.5%; intervention 1: 24/410, 5.9%; intervention 2: 23/430, 5.3%) responded with a score of 1 ("extremely unlikely"). Across all MH problems, no statistically significant group main effect was found on potential professional help seeking ($F_{2,1385}$=0.99; P=.37; Table 2).

Secondary Outcomes

Potential Informal Help Seeking (GHSQ)

For informal sources of support, most participants (1190/1394, 85.37%) selected a score of ≥5 on the 7-point scale (CG: 478/554, 86.3%; intervention 1: 338/410, 82.4%; intervention 2: 374/430, 87%). For 43.69% (609/1394) of the participants, informal help seeking was "extremely likely," with a selected score of 7 (CG: 244/554, 44%; intervention 1: 161/410, 39.3%; intervention 2: 204/430, 47.4%), whereas a minority of 1% (14/1394; CG: 4/554, 0.7%; intervention 1: 7/410, 1.7%; intervention 2: 3/430, 0.7%) responded with a score of 1 ("extremely unlikely"). In the total sample, significant group differences were found regarding informal help seeking ($F_{2,1385}$=3.75; P=.02), with intervention 2 showing a significantly higher mean score than intervention 1 (adjusted MD=0.25; P=.007; Table 2). In the subsample of help seekers across MH problems, the same pattern was observed ($F_{2,616}$=3.21; P=.04; adjusted MD=0.37; P=.01; Multimedia Appendix 1). A significant group effect was also found for the total sample in the problematic alcohol use conditions ($F_{2,273}$=3.51; P=.03; Multimedia Appendix 2). Both the CG (adjusted MD=0.42; P=.02) and intervention 2 (adjusted MD=0.41; P=.03) had greater mean scores than intervention 1.

No Potential Help Seeking (GHSQ)

With regard to no intention of seeking help with any of the potential sources listed in the GHSQ ("I would not seek help from anyone" item), almost half (674/1394, 48.35%) of participants selected a score of 1 or 2 (1="extremely unlikely"; CG: 264/554, 47.7%; intervention 1: 185/410, 45.1%; intervention 2: 225/430, 52.3%), whereas 15.42% (215/1394) responded with a score of 6 or 7 (7="extremely likely"; CG: 89/554, 16.1%; intervention 1: 68/410, 16.6%; intervention 2: 58/430, 13.5%). There were no statistically significant group differences in the total sample (P=.07; Table 2). However, there were trends for group differences in some of the MH issue subgroups (Multimedia Appendices 2 and 3 and Figure 2).

Public Stigma: Blame and Personal Responsibility (USS)

With regard to the USS blame and personal responsibility subscale, statistically significant group differences were found in the total sample ($F_{2,1385}$=3.25; P=.04; Table 2) and in non–help seekers across MH problems ($F_{2,762}$=3.21; P=.04; Multimedia Appendix 1). In the total sample, both intervention 1 and intervention 2 had significantly greater means compared to the CG (intervention 1>CG: adjusted MD=0.084 and P=.03;

XSL•FO
RenderX

intervention 2>CG: adjusted MD=0.085 and P=.03). In the subgroup of non–help seekers, there was a significant difference between intervention 2 and the CG (adjusted MD=0.13; P=.02). Further subgroup analyses revealed no additional differences between experimental conditions. It should be noted that blame and personal responsibility data distributions were heavily skewed to the left (total sample: skew=−1.58). As logarithmic, natural logarithm, square root, and reciprocal transformations did not normalize the distributions, we decided to perform ANCOVAs using the untransformed blame data. Therefore, results should be interpreted with caution.

Public Stigma: Impairment and Distrust (USS)

For the USS distrust subscale, ANCOVAs revealed statistically significant group differences in the total sample ($F_{2,1385}$=8.01; P<.001; Table 2) in both help seekers ($F_{2,616}$=4.39; P=.01) and non–help-seekers across MH problems ($F_{2,762}$=3.74; P=.02; Multimedia Appendix 1). Moreover, statistically significant group differences were found in the total problematic alcohol use subsample ($F_{2,273}$=4.49; P=.01; Multimedia Appendix 2) and its subgroup of non–help seekers ($F_{2,144}$=4.00; P=.02; Multimedia Appendix 3). In the NSSI subgroup of non–help seekers, a significant group main effect was observed ($F_{2,160}$=4.50; P=.01; Multimedia Appendix 3). Across MH problems, both in the total sample (intervention 1>CG: adjusted MD=0.13 and P=.005; intervention 2>CG: adjusted MD=0.17 and P<.001) and the subsample of help seekers (intervention 1>CG: adjusted MD=0.16 and P=.02; intervention 2>CG: adjusted MD=0.17 and P=.01), significantly larger means in both interventions as compared to the CG were observed. Among participants without previous help seeking across MH problems, post hoc comparisons only revealed a statistically significant difference between intervention 2 and the CG (adjusted MD=0.16; P=.007). In the NSSI subgroup of non–help seekers, intervention 2 differed significantly from both the CG (adjusted MD=0.36; P=.005) and intervention 1 (adjusted MD=0.31; P=.02). For problematic alcohol use, in both the total sample and the subsample of non–help seekers, significant post hoc differences between intervention 2 and the CG (MD for the total=0.32 and P=.003; MD for those without previous help seeking=0.42 and P=.006) were found.

Psychological Openness (IASMHS)

No statistically significant group main effect on the IASMHS psychological openness subscale was found in the total sample (Table 2). Significant effects were found in the total depression sample ($F_{2,255}$=4.59; P=.01; Multimedia Appendix 2) and its subgroup of non–help seekers ($F_{2,138}$=4.20; P=.02; Multimedia Appendix 3). In the total depression sample, intervention 1 showed a greater mean in comparison to the CG (adjusted MD=1.38; P=.046) and intervention 2 (adjusted MD=2.24; P=.003). In the subsample of non–help seekers in the depression conditions, intervention 1 was found to have a greater mean than intervention 2 (adjusted MD=2.75; P=.004), but no significant difference was found with the CG (adjusted MD=1.55; P=.09). No significant group main effects were observed in the other subsamples.

Help Seeking Propensity (IASMHS)

In the total sample, no significant group main effect was found for the IASMHS help seeking propensity subscale (Table 2). Subgroup analyses revealed significant differences in the total bulimia sample ($F_{2,271}$=3.27; P=.04), where both intervention 1 (adjusted MD=1.51; P=.03) and intervention 2 (adjusted MD=1.40; P=.04) showed larger means than the CG (Multimedia Appendix 2). No further group differences were found in the other subsamples.

Indifference to Stigma (IASMHS)

For the IASMHS indifference to stigma subscale, differential group main effects were found in the total sample ($F_{2,1385}$=3.18; P=.04; Table 2), in the subsample of non–help seekers ($F_{2,762}$=3.74; P=.02; Multimedia Appendix 1), in the total ($F_{2,288}$=3.22; P=.04; Multimedia Appendix 2) and non–help-seeking ($F_{2,176}$=4.48; P=.01; Multimedia Appendix 3) GAD samples, in the total bulimia sample ($F_{2,271}$=3.45; P=.03; Multimedia Appendix 2), and in the NSSI subsample of non–help seekers ($F_{2,160}$=3.23; P=.04; Multimedia Appendix 3). Across MH problems, the CG showed a larger mean than intervention 1 in the total sample (adjusted MD=0.97; P=.02), whereas a greater mean score in the CG compared to those of both intervention 1 (adjusted MD=1.20; P=.02) and intervention 2 (adjusted MD=1.15; P=.02) was found in the subsample without previous help seeking. A similar pattern emerged in the total GAD sample and its subsample of non–help seekers, where the CG's means were significantly larger in comparison to those of intervention 1 (adjusted MD=2.09; P=.02) in the total sample and of both intervention 1 (adjusted MD=2.57; P=.02) and intervention 2 (adjusted MD=2.91; P=.008) among non–help seekers. In the total bulimia sample, intervention 1 had a significantly higher mean than intervention 2 (adjusted MD=2.26; P=.009), whereas both the CG (adjusted MD=2.10; P=.04) and intervention 2 had greater means than intervention 1 (adjusted MD=2.59; P=.02) in the NSSI subsample of non–help seekers.

Video Acceptability and Transportation

In the total sample, most participants (1041/1394, 74.68%) rated the videos with a score of "4" (705/1394, 50.57%) or "5" (336/1394, 24.1%) on the overall likability item. Regarding comprehensibility, 83.93% (1170/1394) rated the videos as "very comprehensible" ("5" on the 5-point scale), whereas 14.13% (197/1394) assigned them a score of "4." With respect to the videos' interestingness, the responses were distributed across the 5-point scale as follows: 27.4% (382/1394) of participants gave a rating of "5," a total of 41.61% (580/1394) gave the videos a rating of "4," a total of 22.02% (307/1394) assigned them a score of "3," and 7.32% (102/1394) gave them a rating of "2." A minority of participants (23/1394, 1.65%) rated the videos with a score of "1" on the interestingness scale.

In the total sample (Table 2), the intervention 1 videos were rated as generally more likable ($F_{2,1385}$=12.20; P<.001; intervention 1>CG: adjusted MD=0.25 and P<.001; intervention 1>intervention 2: adjusted MD=0.20 and P<.001) and interesting ($F_{2,1385}$=6.39; P=.002; intervention 1>CG: adjusted MD=0.06

XSL•FO
RenderX

and P=.02; intervention 1>intervention 2: adjusted MD=0.07 and P<.001) in comparison to those of the CG and intervention 2. The groups did not differ significantly in video comprehensibility ($F_{2,1385}$=2.01; P=.13). Participants felt more "transported" into the videos' narratives in the CG and intervention 1 as compared to participants in intervention 2 ($F_{2,1385}$=4.23; P=.02; CG>intervention 2: adjusted MD=0.17 and P=.03; intervention 1>intervention 2: adjusted MD=0.23 and P=.006; Table 2). Most subgroup analyses revealed either similar patterns with regard to general likability and interestingness (eg, total help seekers, total GAD sample, and GAD non–help seekers) or no significant differences (eg, GAD help seekers, bulimia help seekers, all depression samples, and all alcohol use samples; Multimedia Appendices 1-3). In the cases of bulimia (total and non–help-seeking subsamples; Multimedia Appendices 2 and 3) and NSSI (Multimedia Appendix 2), different patterns emerged. In the total bulimia sample, the videos of both the CG and intervention 1 scored significantly higher on the interestingness scale than those of intervention 2 ($F_{2,271}$=4.49; P=.01; CG>intervention 2: adjusted MD=0.33 and P=.02; intervention 1>intervention 2: adjusted MD=0.44 and P=.005). In the total NSSI sample, the videos of both intervention 1 and intervention 2 were rated as significantly more likable than those of the CG ($F_{2,277}$=10.31; P<.001; intervention 1>CG: adjusted MD=0.51 and P<.001; intervention 2>CG: adjusted MD=0.29 and P=.008).

Discussion

Principal Findings

This study developed and tested the short-term effectiveness of 2 brief video-based strategies targeted at adolescents and young adults (aged 14 to 29 years) aiming to foster potential professional help seeking (main outcome) and related attitudes for 5 MH problems. In the total sample, we did not find effects of either intervention 1 (psychoeducation) or intervention 2 (positive outcome expectancies) on our primary outcome. However, significant group effects were found with respect to potential informal help seeking, stigma toward others, and indifference to stigma in the total sample. While both intervention groups showed more favorable attitudes than the CG with regard to public stigma, this did not translate to participants' own indifference to stigma. In this case, the CG showed significantly more positive attitudes in comparison to intervention 1. However, this finding was not apparent in the MH issue–specific subgroup analysis with the exception of GAD. Unintended adverse effects of MH interventions have been reported in previous research [83-85], which underlines the need for thorough evaluations of such interventions before their public dissemination. Accordingly, we would advise against the implementation of our GAD interventions at the current stage and would recommend the development and evaluation of other tailored strategies for this MH problem.

With regard to informal help, participants in intervention 2 showed a greater willingness to approach friends, family members, or romantic partners for help than participants in intervention 1. This might have been due to the interventions' design as intervention 2 explicitly depicted improvements in social relationships after the main characters in the videos had sought professional support. Overall, the videos were well accepted and rated as quite interesting, with some room for improvement and with the videos of intervention 1 outperforming those of the other 2 conditions. All videos were, on average, rated as very comprehensible, and no significant group differences were observed in this regard. Interestingly, participants felt more transported into the narratives in the CG and intervention 1 as compared to those in intervention 2. As intervention 2 followed a narrative approach, continuing Paul and Paula's vignette stories, this was surprising. However, as previously stated, intervention 1 was generally more liked and viewed as more interesting in comparison to intervention 2. The animated and fictional third-person approach of intervention 2 seemed to not have sparked as much interest in participants as the facts presented in intervention 1. While we aimed to increase identification with our main characters through the alignment of their genders with those of the participants, intervention 2 might have been insufficient with regard to the perceived "realness" of the story and the characters, which has been identified as crucial for the formation of narrative transportation and, in turn, attitudes and intentions [86]. The rather optimistic portrayal of the help-seeking process might have contributed to a lack of perceived authenticity in this sample as well. Furthermore, implicit MH statements in the videos' scenarios could have been more fruitful. For instance, the viewer could have watched directly how Paul and Paula discussed their issues with a psychotherapist rather than having a narrator describe the situation to them. These types of videos have been associated with improved health literacy and more beneficial attitudes toward cervical cancer [87], and their application to the field of MH would be interesting.

We further observed differential outcomes with respect to the assigned MH problems and participants' actual help-seeking status. While the CG outperformed either intervention 1 (total) or both interventions (subgroup of non–help seekers) for GAD with regard to stigma indifference and no further outcomes were found for GAD, different patterns emerged in the other MH issue groups. Results were mixed, where both intervention 1 and intervention 2 outperformed the other conditions in some of the outcomes but not in others (eg, in depression), or a clearer tendency toward the superiority of one of the interventions emerged (eg, in NSSI). In summary, our results point toward the usefulness of tailored interventions with regard to MH issue type and previous help-seeking experiences of potential target groups. Our finding that different strategies might work differently for each of the 5 MH problems included in our study is in accordance with those of previous research. For example, Ebneter and Latner [78] found varying stigmatizing attitudes among different MH problems. The participants in their study blamed a vignette character with an eating disorder more for their condition than a character with depression, whereas the latter was regarded as more impaired. Our finding that a destigmatizing and psychoeducational intervention such as intervention 1 might work better for bulimia fits their recommendation to target stigmatizing attitudes toward specific MH problems [78]. Similarly, alcohol dependency and self-endangering behaviors were perceived as particularly dangerous in a Swiss vignette study [88]. The humanizing

XSL•FO
RenderX

depiction of our NSSI and problematic alcohol use characters within a framework of close supportive relationships, which improved in quality through psychotherapy, might have been a relevant factor for reduced distrust scores in the intervention 2 condition as compared to the CG and in the case of non–help seekers in the NSSI condition as compared to intervention 1 as well. This approach might be advantageous to reduce public stigma regarding MH problems that are viewed as particularly dangerous. Decisions for one or another interventional strategy may also depend on the specific goal and targeted outcome. While more research is needed, our study provides preliminary evidence for the tailored strategies suggested in Table 3.

Table 3. Mental health (MH) problem–specific intervention recommendations based on the study results[a].

MH problem	Recommended intervention		Aims
	Psychoeducation (intervention 1)	Outcome expectancies (intervention 2)	
All MH problems	?	✓	Antistigma (others), motivate informal help seeking or peer support
GAD[b]	X	X	__[c]
Depression	?	?	Intervention 1: normalize MH-related topics; intervention 2: motivate previous help seekers to reseek professional support in case of relapse
Bulimia	✓	?	Antistigma (self), strengthen help seeking propensity
NSSI[d]	X	✓	Antistigma (others)
Problematic alcohol use	X	✓	Antistigma (others)

[a]Check marks (✓) represent recommended use of an intervention; crosses (X) represent advice against the use of an intervention; and question marks (?) represent inconclusive results and, therefore, no clear recommendation. Recommendations are solely based on the results of this study.

[b]GAD: generalized anxiety disorder.

[c]No recommended use.

[d]NSSI: nonsuicidal self-injury.

Limitations

One limitation of this study lies within the sole investigation of effects on hypothetical intentions and attitudes instead of actual help-seeking behavior. While intentions are substantially associated with behavior and provide valuable insights, they do not translate directly to behavior change [49]. Moreover, only short-term effects were investigated. Previous studies on short video-based interventions have demonstrated sustaining destigmatizing effects for 1 [40] and 5 months after their delivery [43]. While potential long-term effects of our interventions in particular remain unknown and might be investigated in the future, the current state of research points toward potentially impactful long-term effects of low-threshold microinterventions. Related to this, we did not investigate dose-response effects. Research on optimal doses (ie, durations, frequencies, and amounts of intervention components) needed for sustainable change through microinterventions and interventions in general is crucial for well-founded recommendations for or against specific interventions [89].

Just as in other microintervention studies among youth [39], group differences in our study were small. However, the potential high reach of easily accessible, low-threshold interventions such as the ones evaluated in this study is apparent in the final sample size of 1394. As it depends on both effectiveness and reach, this allows for a comparably high public health impact [23,24]. We also included trend effects in our overview (Figure 2), which should be interpreted with caution. However, these findings might be useful to inform the planning of subsequent research in this field.

Future research should focus on improvements in the effectiveness of microinterventions. One approach could be the investigation of interventional framing. In a Japanese study on depression, loss-framed messages (ie, emphasizing negative consequences of refraining from help seeking) had a greater impact on help-seeking intentions than gain-framed (positive consequences of help seeking) or neutral (eg, prevalence rates) messages as well as unformatted, plain-text messages in middle-aged adults [90]. Thus, it might be interesting to conduct future studies on the effects of video-based microinterventions with differently framed messaging as the videos in this study emphasized potential gains of help seeking rather than potential losses of help seeking restraint.

Furthermore, our sample showed, on average, a high level of education, very little public stigma, a pronounced willingness to seek help, and a high rate of actual professional help seeking (624/1394, 44.76%), which limits the generalizability of our findings. More than 90% (1285/1394, 92.18%) of our sample knew someone with MH problems. While we were careful not to recruit MH experts, such as university students of medicine and psychology, youth with a personal interest in MH-related topics seemed to have been more inclined to participate. Related to this, we aimed for a community youth sample rather than a clinical sample. The scenarios that our items referred to were hypothetical and did not necessarily reflect participants' own experiences due to the random assignment to 1 of the 5 MH

XSL•FO
RenderX

problems. A similar approach with targeted interventions according to youth's actual MH status and more individualized elements with regard to gender-related [91] and cultural [92,93] aspects could be promising in future research. Thus, upcoming studies should strive to align their research objectives more closely with the characteristics and needs of the selected target groups.

Finally, we did not include a comprehension check to assess participants' understanding of and engagement with the content presented in the videos. While we accounted for the time participants spent on the video pages and only included participants with sufficient durations of stay in the final analyses, they may not have fully comprehended or attended to the video material. The substantial number of excluded participants who completed the study without meeting the time threshold (472/2208, 21.38%) underlines this potential issue. Thus, future studies should address this limitation by including comprehension checks to improve the robustness of the findings.

Conclusions

The low uptake of professional MH services in youth reflects the need for appropriate strategies to facilitate professional help seeking. This study investigated the effectiveness of 2 short video-based strategies targeted at youth (aged 14 to 29 years) on potential professional help seeking and related attitudes for 5 MH problems. While we did not find intervention effects on potential professional help seeking (with the exception of previous help seekers in the depression conditions), differential intervention effects depending on each MH problem and participants' actual help-seeking status were found in our secondary outcomes, such as public stigma. Our study results can be used to inform the development of new antistigma interventions, which, based on our results, we would recommend tailoring to specific MH problems, target groups, and outcomes. While group differences were small, such low-threshold interventions can be easily disseminated and, therefore, hold potential for a high reach and, thus, a meaningful impact at the population level. More research is needed for more robust and generalizable recommendations.

Acknowledgments

This study was funded by the German Federal Ministry of Education and Research (funding identifier: 01GL1904). The Federal Ministry of Education and Research had no influence on the design of the study and was not involved in data collection, analysis, and interpretation or the writing of the manuscript. For the publication fee, the authors acknowledge financial support from Heidelberg University. The authors thank Sabrina Baldofski, Elisabeth Kohls, Felicitas Mayr, and Maria I. Austermann for their support in the creation of video materials for this study (depression and alcohol abuse videos). The authors thank Lutfi Arikan (University Hospital Heidelberg) for enabling the technical implementation of this study.

Data Availability

Individual participant data underlying the results presented in this publication (text, tables, figures, and appendices) and statistical code will be shared during the period of 3 months to 5 years following publication of the paper with researchers who provide a methodologically sound proposal to the corresponding author.

Authors' Contributions

SB, MM, and MK conceptualized the study. SB led the grant application. SB, MM, MK, and DL contributed to the study design and selection of screening and outcome measures. DL prepared the technical implementation of this study. DL, AM, EK, and SK prepared the video interventions under the supervision of MM, NA, HB, PP, CRK, RT, and SB, who also provided feedback and information with regard to their fields of expertise. MM generated the random allocation sequence. DL was responsible for study recruitment, which was supported by student assistants. DL analyzed the data, wrote the first draft of this manuscript, and created its tables and figures, with SB and MM providing further feedback, guidance, and supervision during each step. All authors provided feedback on the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

PP is an advisor for Boehringer Ingelheim and has received speaker's honoraria from InfectoPharm, GSK, Janssen, and Oral B.

Multimedia Appendix 1
Separate analysis of covariance results and pairwise comparisons for participants with and without previous help-seeking experience.
[DOC File , 95 KB - jmir_v26i1e54478_app1.doc]

Multimedia Appendix 2
Separate analysis of covariance results and pairwise comparisons for outcomes per mental health problem.
[DOC File , 180 KB - jmir_v26i1e54478_app2.doc]

XSL•FO
RenderX

Multimedia Appendix 3
Separate analysis of covariance results and pairwise comparisons for participants with and without previous help-seeking experience per mental health problem.
[DOC File , 366 KB - jmir_v26i1e54478_app3.doc]

Multimedia Appendix 4
CONSORT-eHEALTH Checklist V 1.6.1.
[PDF File (Adobe PDF File), 9119 KB - jmir_v26i1e54478_app4.pdf]

References

1. Gibb SJ, Fergusson DM, Horwood LJ. Burden of psychiatric disorder in young adulthood and life outcomes at age 30. Br J Psychiatry 2010 Aug;197(2):122-127. [doi: 10.1192/bjp.bp.109.076570] [Medline: 20679264]
2. Weitkamp K, Daniels JK, Romer G, Wiegand-Grefe S. Health-related quality of life of children and adolescents with mental disorders. Health Qual Life Outcomes 2013 Jul 31;11(1):129 [FREE Full text] [doi: 10.1186/1477-7525-11-129] [Medline: 23902824]
3. Sawyer MG, Whaites L, Rey JM, Hazell PL, Graetz BW, Baghurst P. Health-related quality of life of children and adolescents with mental disorders. J Am Acad Child Adolesc Psychiatry 2002 May;41(5):530-537. [doi: 10.1097/00004583-200205000-00010] [Medline: 12014785]
4. Romanowska S, MacQueen G, Goldstein BI, Wang J, Kennedy SH, Bray S, et al. Social and role functioning in youth at risk of serious mental illness. Early Interv Psychiatry 2020 Aug 10;14(4):463-469. [doi: 10.1111/eip.12872] [Medline: 31502422]
5. Gardner A, Filia K, Killackey E, Cotton S. The social inclusion of young people with serious mental illness: a narrative review of the literature and suggested future directions. Aust N Z J Psychiatry 2019 Jan;53(1):15-26. [doi: 10.1177/0004867418804065] [Medline: 30309242]
6. Erskine HE, Moffitt TE, Copeland WE, Costello EJ, Ferrari AJ, Patton G, et al. A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. Psychol Med 2015 May;45(7):1551-1563 [FREE Full text] [doi: 10.1017/S0033291714002888] [Medline: 25534496]
7. Suhrcke M, Pillas D, Selai C. Economic aspects of mental health in children and adolescents. Social Cohesion for Mental Wellbeing among adolescents. 2008. URL: https://pure.york.ac.uk/portal/en/publications/economic-aspects-of-mental-health-in-children-and-adolescents [accessed 2024-04-08]
8. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan 02;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
9. Kessler RC, Avenevoli S, Costello EJ, Georgiades K, Green JG, Gruber MJ, et al. Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. Arch Gen Psychiatry 2012 Apr 01;69(4):372-380 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2011.160] [Medline: 22147808]
10. Hofstra MB, Van Der Ende J, Verhulst FC. Adolescents' self-reported problems as predictors of psychopathology in adulthood: 10-year follow-up study. Br J Psychiatry 2001 Sep 02;179(3):203-209. [doi: 10.1192/bjp.179.3.203] [Medline: 11532796]
11. Pardini D, White HR, Stouthamer-Loeber M. Early adolescent psychopathology as a predictor of alcohol use disorders by young adulthood. Drug Alcohol Depend 2007 Apr;88 Suppl 1(Suppl 1):S38-S49 [FREE Full text] [doi: 10.1016/j.drugalcdep.2006.12.014] [Medline: 17257781]
12. Heijmens Visser JH, van der Ende J, Koot HM, Verhulst FC. Predictors of psychopathology in young adults referred to mental health services in childhood or adolescence. Br J Psychiatry 2000 Jul 02;177(1):59-65. [doi: 10.1192/bjp.177.1.59] [Medline: 10945090]
13. Reef J, Diamantopoulou S, van Meurs I, Verhulst F, van der Ende J. Child to adult continuities of psychopathology: a 24-year follow-up. Acta Psychiatr Scand 2009 Sep 03;120(3):230-238. [doi: 10.1111/j.1600-0447.2009.01422.x] [Medline: 19522882]
14. Pearson CM, Miller J, Ackard DM, Loth KA, Wall MM, Haynos AF, et al. Stability and change in patterns of eating disorder symptoms from adolescence to young adulthood. Int J Eat Disord 2017 Jul 15;50(7):748-757 [FREE Full text] [doi: 10.1002/eat.22692] [Medline: 28199037]
15. Hintzpeter B, Metzner F, Pawils S, Bichmann H, Kamtsiuris P, Ravens-Sieberer U, et al. Inanspruchnahme von ärztlichen und psychotherapeutischen Leistungen durch Kinder und Jugendliche mit psychischen Auffälligkeiten. Kindh Entwickl 2014 Oct;23(4):229-238. [doi: 10.1026/0942-5403/a000148]
16. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 2007 Sep 08;370(9590):841-850 [FREE Full text] [doi: 10.1016/S0140-6736(07)61414-7] [Medline: 17826169]

XSL•FO
RenderX

17. Cotter P, Kaess M, Corcoran P, Parzer P, Brunner R, Keeley H, et al. Help-seeking behaviour following school-based screening for current suicidality among European adolescents. Soc Psychiatry Psychiatr Epidemiol 2015 Jun 6;50(6):973-982. [doi: 10.1007/s00127-015-1016-3] [Medline: 25656270]
18. Vanheusden K, Mulder CL, van der Ende J, van Lenthe FJ, Mackenbach JP, Verhulst FC. Young adults face major barriers to seeking help from mental health services. Patient Educ Couns 2008 Oct;73(1):97-104. [doi: 10.1016/j.pec.2008.05.006] [Medline: 18584997]
19. Kaltenthaler E, Sutcliffe P, Parry G, Beverley C, Rees A, Ferriter M. The acceptability to patients of computerized cognitive behaviour therapy for depression: a systematic review. Psychol Med 2008 Nov;38(11):1521-1530. [doi: 10.1017/S0033291707002607] [Medline: 18205964]
20. Lawrence D, Johnson S, Hafekost J, Boterhoven de Haan K, Sawyer M, Ainley J, et al. The mental health of children and adolescents: report on the second Australian child and adolescent survey of mental health and wellbeing. Government of Australia. 2015. URL: https://www.health.gov.au/sites/default/files/documents/2020/11/the-mental-health-of-children-and-adolescents_0.pdf [accessed 2024-04-08]
21. Lillevoll KR, Vangberg HC, Griffiths KM, Waterloo K, Eisemann MR. Uptake and adherence of a self-directed internet-based mental health intervention with tailored e-mail reminders in senior high schools in Norway. BMC Psychiatry 2014 Jan 21;14(1):14 [FREE Full text] [doi: 10.1186/1471-244X-14-14] [Medline: 24443820]
22. Woodford J, Farrand P, Bessant M, Williams C. Recruitment into a guided internet based CBT (iCBT) intervention for depression: lesson learnt from the failure of a prevalence recruitment strategy. Contemp Clin Trials 2011 Sep;32(5):641-648. [doi: 10.1016/j.cct.2011.04.013] [Medline: 21570485]
23. Glasgow RE, Estabrooks PE. Pragmatic applications of RE-AIM for health care initiatives in community and clinical settings. Prev Chronic Dis 2018 Jan 04;15:E02 [FREE Full text] [doi: 10.5888/pcd15.170271] [Medline: 29300695]
24. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
25. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2013 Aug 09;44(6):1303-1317. [doi: 10.1017/s0033291713001943]
26. Mojtabai R, Olfson M, Sampson NA, Jin R, Druss B, Wang PS, et al. Barriers to mental health treatment: results from the National Comorbidity Survey replication. Psychol Med 2011 Aug;41(8):1751-1761 [FREE Full text] [doi: 10.1017/S0033291710002291] [Medline: 21134315]
27. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
28. Eisenberg D, Golberstein E, Gollust SE. Help-seeking and access to mental health care in a university student population. Med Care 2007 Jul;45(7):594-601. [doi: 10.1097/MLR.0b013e31803bb4c1] [Medline: 17571007]
29. Clement S, Schauman O, Graham T, Maggioni F, Evans-Lacko S, Bezborodovs N, et al. What is the impact of mental health-related stigma on help-seeking? a systematic review of quantitative and qualitative studies. Psychol Med 2015 Jan;45(1):11-27. [doi: 10.1017/S0033291714000129] [Medline: 24569086]
30. Schnyder N, Panczak R, Groth N, Schultze-Lutter F. Association between mental health-related stigma and active help-seeking: systematic review and meta-analysis. Br J Psychiatry 2017 Apr;210(4):261-268. [doi: 10.1192/bjp.bp.116.189464] [Medline: 28153928]
31. Furnham A, Swami V. Mental health literacy: a review of what it is and why it matters. Int Perspect Psychol 2018 Oct;7(4):240-257. [doi: 10.1037/ipp0000094]
32. Jorm AF. Mental health literacy: empowering the community to take action for better mental health. Am Psychol 2012 Apr;67(3):231-243. [doi: 10.1037/a0025957] [Medline: 22040221]
33. Patafio B, Miller P, Baldwin R, Taylor N, Hyder S. A systematic mapping review of interventions to improve adolescent mental health literacy, attitudes and behaviours. Early Interv Psychiatry 2021 Dec 03;15(6):1470-1501. [doi: 10.1111/eip.13109] [Medline: 33393231]
34. Howard KA, Griffiths KM, McKetin R, Ma J. Can a brief biologically-based psychoeducational intervention reduce stigma and increase help-seeking intentions for depression in young people? a randomised controlled trial. J Child Adolesc Ment Health 2018 May 15;30(1):27-39. [doi: 10.2989/17280583.2018.1467323] [Medline: 29764293]
35. Jaber R, Farroukh M, Ismail M, Najda J, Sobh H, Hammad A, et al. Measuring depression and stigma towards depression and mental health treatment among adolescents in an Arab-American community. Int J Cult Ment Health 2015 Sep 24;8(3):247-254 [FREE Full text] [doi: 10.1080/17542863.2014.953188] [Medline: 26257824]
36. Bruning Brown J, Winzelberg AJ, Abascal LB, Taylor CB. An evaluation of an internet-delivered eating disorder prevention program for adolescents and their parents. J Adolesc Health 2004 Oct;35(4):290-296. [doi: 10.1016/j.jadohealth.2003.10.010] [Medline: 15450542]
37. O'Kearney R, Kang K, Christensen H, Griffiths K. A controlled trial of a school-based internet program for reducing depressive symptoms in adolescent girls. Depress Anxiety 2009;26(1):65-72. [doi: 10.1002/da.20507] [Medline: 18828141]

XSL•FO
RenderX

38. Baumel A, Fleming T, Schueller SM. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J Med Internet Res 2020 Oct 29;22(10):e20631 [FREE Full text] [doi: 10.2196/20631] [Medline: 33118946]
39. Ebert DD, Franke M, Kählke F, Küchler AM, Bruffaerts R, Mortier P, WHO World Mental Health - International College Student collaborators. Increasing intentions to use mental health services among university students. Results of a pilot randomized controlled trial within the World Health Organization's World Mental Health International college student initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1754 [FREE Full text] [doi: 10.1002/mpr.1754] [Medline: 30456814]
40. Amsalem D, Markowitz JC, Jankowski SE, Yang LH, Valeri L, Lieff SA, et al. Sustained effect of a brief video in reducing public stigma toward individuals with psychosis: a randomized controlled trial of young adults. Am J Psychiatry 2021 Jul 26;178(7):635-642. [doi: 10.1176/appi.ajp.2020.20091293] [Medline: 33900809]
41. Amsalem D, Jankowski SE, Pagdon S, Valeri L, Smith S, Yang LH, et al. "It is hard to be a woman with schizophrenia": randomized controlled trial of a brief video intervention to reduce public stigma in young adults. J Clin Psychiatry 2022 Dec 19;84(1):22m14534. [doi: 10.4088/JCP.22m14534] [Medline: 36541795]
42. Amsalem D, Martin A. Reducing depression-related stigma and increasing treatment seeking among adolescents: randomized controlled trial of a brief video intervention. J Child Psychol Psychiatry 2022 Feb 06;63(2):210-217. [doi: 10.1111/jcpp.13427] [Medline: 33821507]
43. Conceição V, Rothes I, Gusmão R. The effects of a video-based randomized controlled trial intervention on depression stigma and help-seeking attitudes in university students. Psychiatry Res 2022 Feb;308:114356. [doi: 10.1016/j.psychres.2021.114356] [Medline: 34972028]
44. Clough BA, Nazareth SM, Casey LM. Making the grade: a pilot investigation of an e-intervention to increase mental health literacy and help-seeking intentions among international university students. Br J Guid Couns 2019 Oct 03;48(3):347-359. [doi: 10.1080/03069885.2019.1673312]
45. Gaudiano BA, Davis CH, Miller IW, Uebelacker L. Pilot randomized controlled trial of a video self-help intervention for depression based on acceptance and commitment therapy: feasibility and acceptability. Clin Psychol Psychother 2020 May 28;27(3):396-407 [FREE Full text] [doi: 10.1002/cpp.2436] [Medline: 32087610]
46. Fong TH, Mak WW. The effects of internet-based storytelling programs (amazing adventure against stigma) in reducing mental illness stigma with mediation by interactivity and stigma content: randomized controlled trial. J Med Internet Res 2022 Aug 12;24(8):e37973 [FREE Full text] [doi: 10.2196/37973] [Medline: 35969460]
47. Logsdon CM, Myers J, Rushton J, Gregg JL, Josephson AM, Davis DW, et al. Efficacy of an internet-based depression intervention to improve rates of treatment in adolescent mothers. Arch Womens Ment Health 2018 Jun;21(3):273-285 [FREE Full text] [doi: 10.1007/s00737-017-0804-z] [Medline: 29260321]
48. Schwarzer R, Luszczynska A. How to overcome health-compromising behaviors. Eur Psychol 2008 Jan;13(2):141-151. [doi: 10.1027/1016-9040.13.2.141]
49. Zhang CQ, Zhang R, Schwarzer R, Hagger MS. A meta-analysis of the health action process approach. Health Psychol 2019 Jul;38(7):623-637. [doi: 10.1037/hea0000728] [Medline: 30973747]
50. Fishbein M, Ajzen I. Predicting and Changing Behavior: The Reasoned Action Approach. New York, NY: Psychology Press; 2010.
51. Yeager CM, Shoji K, Luszczynska A, Benight CC. Engagement with a trauma recovery internet intervention explained with the Health Action Process Approach (HAPA): longitudinal study. JMIR Ment Health 2018 Apr 10;5(2):e29 [FREE Full text] [doi: 10.2196/mental.9449] [Medline: 29636323]
52. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet E-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
53. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
54. Santelli JS, Rosenfeld WD, DuRant RH, Dubler N, Morreale M, English A, et al. Guidelines for adolescent health research: a position paper of the society for adolescent medicine. J Adolesc Health 1995 Nov;17(5):270-276. [doi: 10.1016/1054-139x(95)00181-q] [Medline: 8924431]
55. Schachter D, Kleinman I, Harvey W. Informed consent and adolescents. Can J Psychiatry 2005 Aug 01;50(9):534-540. [doi: 10.1177/070674370505000906] [Medline: 16262108]
56. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
57. Mehta CM, Arnett JJ, Palmer CG, Nelson LJ. Established adulthood: a new conception of ages 30 to 45. Am Psychol 2020 May;75(4):431-444. [doi: 10.1037/amp0000600] [Medline: 32378940]
58. Wilhelm M, Feldhege J, Bauer S, Moessner M. Einsatz internetbasierter verlaufsmessung in der psychotherapieforschung. Psychotherapeut 2020 Oct 12;65(6):505-511. [doi: 10.1007/s00278-020-00461-7]
59. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]

XSL•FO
RenderX

60. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
61. Killen JD, Taylor CB, Hammer LD, Litt I, Wilson DM, Rich T, et al. An attempt to modify unhealthful eating attitudes and weight regulation practices of young adolescent girls. Int J Eat Disord 1993 May;13(4):369-384. [doi: 10.1002/1098-108x(199305)13:4<369::aid-eat2260130405>3.0.co;2-0] [Medline: 8490639]
62. Killen JD, Taylor CB, Hayward C, Wilson DM, Haydel KF, Hammer LD, et al. Pursuit of thinness and onset of eating disorder symptoms in a community sample of adolescent girls: a three-year prospective analysis. Int J Eat Disord 1994 Nov;16(3):227-238. [doi: 10.1002/1098-108x(199411)16:3<227::aid-eat2260160303>3.0.co;2-l] [Medline: 7833956]
63. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory care quality improvement project (ACQUIP). Alcohol use disorders identification test. Arch Intern Med 1998 Sep 14;158(16):1789-1795. [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
64. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction 1993 Jun;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
65. Kuitunen-Paul S, Roerecke M. Alcohol Use isorders Identification Test (AUDIT) and mortality risk: a systematic review and meta-analysis. J Epidemiol Community Health 2018 Sep;72(9):856-863. [doi: 10.1136/jech-2017-210078] [Medline: 29921648]
66. Fischer G, Ameis N, Parzer P, Plener PL, Groschwitz R, Vonderlin E, et al. The German version of the self-injurious thoughts and behaviors interview (SITBI-G): a tool to assess non-suicidal self-injury and suicidal behavior disorder. BMC Psychiatry 2014 Sep 18;14(1):265 [FREE Full text] [doi: 10.1186/s12888-014-0265-0] [Medline: 25230931]
67. Home page. Powtoon. URL: https://www.powtoon.com/ [accessed 2024-04-08]
68. Wright A, Jorm AF, Mackinnon AJ. Labels used by young people to describe mental disorders: which ones predict effective help-seeking choices? Soc Psychiatry Psychiatr Epidemiol 2012 Jun;47(6):917-926. [doi: 10.1007/s00127-011-0399-z] [Medline: 21626056]
69. Mond JM, Hay PJ, Paxton SJ, Rodgers B, Darby A, Nillson J, et al. Eating disorders "mental health literacy" in low risk, high risk and symptomatic women: implications for health promotion programs. Eat Disord 2010;18(4):267-285. [doi: 10.1080/10640266.2010.490115] [Medline: 20603729]
70. DeJong H, Hillcoat J, Perkins S, Grover M, Schmidt U. Illness perception in bulimia nervosa. J Health Psychol 2012 Apr;17(3):399-408. [doi: 10.1177/1359105311416874] [Medline: 21859799]
71. Bulik CM. Towards a science of eating disorders: replacing myths with realities: the fourth Birgit Olsson lecture. Nord J Psychiatry 2016 Sep 10;70(3):224-230. [doi: 10.3109/08039488.2015.1074284] [Medline: 26355341]
72. Keski-Rahkonen A, Mustelin L. Epidemiology of eating disorders in Europe: prevalence, incidence, comorbidity, course, consequences, and risk factors. Curr Opin Psychiatry 2016 Nov;29(6):340-345. [doi: 10.1097/YCO.0000000000000278] [Medline: 27662598]
73. Nagl M, Jacobi C, Paul M, Beesdo-Baum K, Höfler M, Lieb R, et al. Prevalence, incidence, and natural course of anorexia and bulimia nervosa among adolescents and young adults. Eur Child Adolesc Psychiatry 2016 Aug;25(8):903-918. [doi: 10.1007/s00787-015-0808-z] [Medline: 26754944]
74. Laberg S, Törnkvist Å, Andersson G. Experiences of patients in cognitive behavioural group therapy: a qualitative study of eating disorders. Cogn Behav Ther 2001 Jan;30(4):161-178. [doi: 10.1080/02845710127474]
75. Wilson CJ, Deane FP, Ciarrochi J, Rickwood D. Measuring help-seeking intentions: properties of the general help-seeking questionnaire. Can J Couns 2005;39(1):15-28 [FREE Full text]
76. Divin N, Harper P, Curran E, Corry D, Leavey G. Help-seeking measures and their use in adolescents: a systematic review. Adolescent Res Rev 2018 Jan 17;3(1):113-122. [doi: 10.1007/S40894-017-0078-8]
77. Mackenzie CS, Knox VJ, Gekoski WL, Macaulay HL. An adaptation and extension of the attitudes toward seeking professional psychological help scale. J Applied Social Pyschol 2006 Jul 31;34(11):2410-2433. [doi: 10.1111/j.1559-1816.2004.tb01984.x]
78. Ebneter DS, Latner JD. Stigmatizing attitudes differ across mental health disorders: a comparison of stigma across eating disorders, obesity, and major depressive disorder. J Nerv Ment Dis 2013 Apr;201(4):281-285. [doi: 10.1097/NMD.0b013e318288e23f] [Medline: 23538972]
79. Appel M, Gnambs T, Richter T, Green MC. The Transportation Scale–Short Form (TS–SF). Media Psychol 2015 Feb 03;18(2):243-266. [doi: 10.1080/15213269.2014.987400]
80. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/BF03193146]
81. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.r-project.org/ [accessed 2024-04-08]
82. IBM SPSS Statistics for Windows, version 28.0. IBM Corp. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-28010 [accessed 2024-04-08]

XSL•FO
RenderX

83. Lienemann BA, Siegel JT, Crano WD. Persuading people with depression to seek help: respect the boomerang. Health Commun 2013 Oct;28(7):718-728. [doi: 10.1080/10410236.2012.712091] [Medline: 23066999]
84. Lienemann BA, Siegel JT. State psychological reactance to depression public service announcements among people with varying levels of depressive symptomatology. Health Commun 2016 Jun 18;31(1):102-116. [doi: 10.1080/10410236.2014.940668] [Medline: 26086674]
85. Siegel JT, Lienemann BA, Rosenberg BD. Resistance, reactance, and misinterpretation: highlighting the challenge of persuading people with depression to seek help. Soc Pers Psychol Compass 2017 Jun 05;11(6):e12322. [doi: 10.1111/spc3.12322]
86. van Laer T, de Ruyter K, Visconti LM, Wetzels M. The extended transportation-imagery model: a meta-analysis of the antecedents and consequences of consumers' narrative transportation. J Consum Res 2014 Feb 01;40(5):797-817 [FREE Full text] [doi: 10.1086/673383]
87. Murphy ST, Frank LB, Chatterjee JS, Baezconde-Garbanati L. Narrative versus non-narrative: the role of identification, transportation and emotion in reducing health disparities. J Commun 2013 Feb 07;63(1):116-137 [FREE Full text] [doi: 10.1111/jcom.12007] [Medline: 24347679]
88. Sowislo JF, Gonet-Wirz F, Borgwardt S, Lang UE, Huber CG. Perceived dangerousness as related to psychiatric symptoms and psychiatric service use - a vignette based representative population survey. Sci Rep 2017 Apr 03;8(1):45716 [FREE Full text] [doi: 10.1038/srep45716] [Medline: 28367993]
89. McVay MA, Bennett GG, Steinberg D, Voils CI. Dose-response research in digital health interventions: concepts, considerations, and challenges. Health Psychol 2019 Dec;38(12):1168-1174 [FREE Full text] [doi: 10.1037/hea0000805] [Medline: 31580127]
90. Suka M, Yamauchi T, Yanagisawa H. Comparing responses to differently framed and formatted persuasive messages to encourage help-seeking for depression in Japanese adults: a cross-sectional study with 2-month follow-up. BMJ Open 2018 Nov 12;8(11):e020823 [FREE Full text] [doi: 10.1136/bmjopen-2017-020823] [Medline: 30420341]
91. Kaushik A, Kostaki E, Kyriakopoulos M. The stigma of mental illness in children and adolescents: a systematic review. Psychiatry Res 2016 Sep 30;243:469-494. [doi: 10.1016/j.psychres.2016.04.042] [Medline: 27517643]
92. Misra S, Jackson VW, Chong J, Choe K, Tay C, Wong J, et al. Systematic review of cultural aspects of stigma and mental illness among racial and ethnic minority groups in the United States: implications for interventions. Am J Community Psychol 2021 Dec;68(3-4):486-512. [doi: 10.1002/ajcp.12516] [Medline: 33811676]
93. Papadopoulos C, Foster J, Caldwell K. 'Individualism-collectivism' as an explanatory device for mental illness stigma. Community Ment Health J 2013 Jun 27;49(3):270-280. [doi: 10.1007/s10597-012-9534-x] [Medline: 22837106]

Abbreviations

ANCOVA: analysis of covariance
AUDIT-C: Alcohol Use Disorders Identification Test for Consumption
CG: control group
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
GAD: generalized anxiety disorder
GAD-7: 7-item Generalized Anxiety Disorder Scale
GHSQ: General Help Seeking Questionnaire
HAPA: Health Action Process Approach
IASMHS: Inventory of Attitudes Toward Seeking Mental Health Services
MD: mean difference
MH: mental health
NSSI: nonsuicidal self-injury
PHQ-9: 9-item Patient Health Questionnaire
RCT: randomized controlled trial
USS: Universal Stigma Scale
WCS: Weight Concerns Scale

XSL•FO
RenderX

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 13.11.23; peer-reviewed by A Morgan; comments to author 15.12.23; revised version received 22.12.23; accepted 08.03.24; published 24.04.24.

Please cite as:
Lemmer D, Moessner M, Arnaud N, Baumeister H, Mutter A, Klemm SL, König E, Plener P, Rummel-Kluge C, Thomasius R, Kaess M, Bauer S
The Impact of Video-Based Microinterventions on Attitudes Toward Mental Health and Help Seeking in Youth: Web-Based Randomized Controlled Trial
J Med Internet Res 2024;26:e54478
URL: https://www.jmir.org/2024/1/e54478
doi:10.2196/54478
PMID:38656779

©Diana Lemmer, Markus Moessner, Nicolas Arnaud, Harald Baumeister, Agnes Mutter, Sarah-Lena Klemm, Elisa König, Paul Plener, Christine Rummel-Kluge, Rainer Thomasius, Michael Kaess, Stephanie Bauer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Rapport Building in Written Crisis Services: Qualitative Content Analysis

Laura Schwab-Reese[1], BSc, MA, PhD; Caitlyn Short[1], BSc, MPH; Larel Jacobs[2], MEd, MSc; Michelle Fingerman[2], MSc

[1]Department of Public Health, Purdue University, West Lafayette, IN, United States

[2]Childhelp, Scottsdale, AZ, United States

Corresponding Author:
Laura Schwab-Reese, BSc, MA, PhD
Department of Public Health
Purdue University
MTHW 214F
West Lafayette, IN, 47907
United States
Phone: 1 765 496 6723
Email: lschwabr@purdue.edu

Abstract

Background: Building therapeutic relationships and social presence are challenging in digital services and maybe even more difficult in written services. Despite these difficulties, in-person care may not be feasible or accessible in all situations.

Objective: This study aims to categorize crisis counselors' efforts to build rapport in written conversations by using deidentified conversation transcripts from the text and chat arms of the National Child Abuse Hotline. Using these categories, we identify the common characteristics of successful conversations. We defined success as conversations where help-seekers reported the hotline was a good way to seek help and that they were a lot more hopeful, a lot more informed, a lot more prepared to address the situation, and experiencing less stress, as reported by help-seekers.

Methods: The sample consisted of transcripts from 314 purposely selected conversations from of the 1153 text and chat conversations during July 2020. Hotline users answered a preconversation survey (ie, demographics) and a postconversation survey (ie, their perceptions of the conversation). We used qualitative content analysis to process the conversations.

Results: Active listening skills, including asking questions, paraphrasing, reflecting feelings, and interpreting situations, were commonly used by counselors. Validation, unconditional positive regard, and evaluation-based language, such as praise and apologies, were also often used. Compared with less successful conversations, successful conversations tended to include fewer statements that attend to the emotional dynamics. There were qualitative differences in how the counselors applied these approaches. Generally, crisis counselors in positive conversations tended to be more specific and tailor their comments to the situation.

Conclusions: Building therapeutic relationships and social presence are essential to digital interventions involving mental health professionals. Prior research demonstrates that they can be challenging to develop in written conversations. Our work demonstrates characteristics associated with successful conversations that could be adopted in other written help-seeking interventions.

(J Med Internet Res 2024;26:e42049) doi:10.2196/42049

KEYWORDS

empathy; crisis hotline; child maltreatment; text hotline; chat hotline; telehealth; digital empathy; counseling; child abuse; family violence; crisis; hotline; chat; tele; emotional dynamics; therapeutic relationships; therapy; content analysis; text; inductive; deductive; emotion; affect; emotional dynamic; counseling psychology

Introduction

Background

Since the 1990s, mental health providers have explored how to support clients via internet-based communication [1]. Prior work suggests that young people may be particularly interested in these approaches, as digital communication feels more private and emotionally safe [2]. However, internet-based communication, particularly written communication, may have significant barriers for providers and clients, including the

XSL•FO
RenderX

inability to express emotion and challenges in communicating clearly [2]. Currently, there is limited evidence on how to overcome these communication issues in counseling settings [3]. Understanding how to do so may help telehealth providers build stronger therapeutic relationships, thus improving the help-seeking process. Further, this understanding may help agencies improve services and training for providers.

Technology-Based Mental Health Interventions

Technology-based (ie, telehealth) mental health services may not be as effective as in-person services. One recent meta-analysis suggests that videoconferencing-based therapeutic relationships may be inferior to those developed during in-person therapy [4]. However, there may be times when in-person care is not accessible or feasible. Nearly half of people in the United States live in a mental health shortage area, and there are areas with less than 2 psychiatrists per 100,000 residents [5,6]. As a result, increasing access to mental health care may depend on telehealth approaches. Within telehealth studies, interventions retaining elements of human contact are more effective than entirely computer-based interventions [7,8].

Two critical aspects of the helping relationship, therapeutic relationship and social presence, may be challenged when engaging digitally. A therapeutic relationship based on mutual trust, respect, empathy, and positive regard is essential in counseling [9]. Hundreds of studies have confirmed the importance of this collaborative relationship [10]. For most help-seekers, confidence in the provider, including perceptions of empathy and expertise, is key to developing a strong relationship [11-13].

Social presence [14], the sense of connecting and being with another, is another element that may be compromised during digital communication. Social presence may also be defined as the degree to which the other person feels “real” [15]. Although it is a natural element of face-to-face counseling, telehealth providers may have to be intentional in building a social presence. When conversing with unknown entities through written technology, it is common to question whether the other person is a human or a bot [16], in part because people are not reliably able to differentiate between the two [17]. Some prior work suggests that social presence is an important aspect of digital helping relationships because it assists in building therapeutic partnerships, professional bonds, and open communication [18,19].

Much of the literature on telehealth counseling focuses on verbal communication via videoconferencing or phone [20]. Few studies examine written mental health counseling services, and there is reason to believe that spoken and written communication are substantially different. In a recent review of the digital therapeutic relationship, Bantjes and Slabbert [20] suggest practical strategies for establishing rapport in digital spaces, such as maintaining eye contact, having high-speed internet to avoid lags, and attending to lighting and microphone placement. These strategies improve the audio and visual cues, which are not applicable to written communication.

Written Interventions

The literature on written counseling is limited [21,22]. In the 1990s, a small group of mental health providers offered therapy via email [23]. This early work identified several possible strengths and limitations. It was helpful for clients to write about their feelings, and the anonymity of email allowed them to share more openly. This asynchronous approach also increased many individuals’ sense of control, as they could choose when and where to engage with the therapist. Conversely, building a relationship and understanding nuances could be difficult without the usual social cues [23]. Two more recent literature reviews support many of the impressions formed by the early adopters, although most of the studies had very small sample sizes [22,24].

Written interventions may be challenging for the provider and patient, and both experience similar challenges. One randomized controlled trial of a chat-based cognitive behavioral therapy demonstrated reduced depression symptoms after 10 sessions [25]. In a parallel qualitative study, participants reported mixed perceptions of the experience [26]. Some reported feeling more able to share openly and process because of the anonymous platform. Others felt it was challenging to develop a relationship and express complex feelings and thoughts via writing [26]. Another study assessed differences between email-based cognitive behavioral therapy and unguided treatment. The email and unguided programs had better outcomes than the wait-list control group for some, but not all, outcomes [27,28]. Other studies, with and without in-person or telephone comparison groups, showed similarly mixed results [29,30]. In one unpublished dissertation, counselors who provided email services reported feeling substantial anxiety due to uncertainty, limited sensory information, and concerns about misunderstanding clients’ intentions [31]. The lack of visual, verbal, and social cues was particularly challenging [31]. They often focused more on the tasks and transactional aspects of helping to manage these uncertain dynamics [31]. Many also talked about needing much more time than usual to build the therapeutic relationship, although it did eventually happen for most [31].

Beyond mental health counseling, some recent work has examined written communication for brief counseling and advocacy [3,32,33]. Overall, privacy, autonomy, control, anonymity, and accessibility are seen as benefits of written services [34-36]. Building social presence and connection is an important aspect of the experience [34]. Often this professional connection builds over time, but because the help-seeker and crisis counselor or advocate do not have an ongoing relationship, it may be particularly difficult to communicate adequately and build a relationship [2,30,37,38]. Correctly understanding sarcasm, humor, and other nuanced language is particularly challenging in these brief interventions [3,39]. Like mental health counseling, the impact of written crisis and advocacy services is unclear in the current literature and may depend on geographical location, counselor training, and the help-seekers’ situations [33,40-46].

XSL•FO
RenderX

Study Purpose

Overall, establishing a human connection based on a strong therapeutic relationship and social presence will likely result in more effective, acceptable interventions. Providing crisis services is complex, and the confines of written communication create additional challenges. Rapport-building is particularly difficult, and mistakes may cause the help-seeker to feel worse [47]. However, there are not yet best practices for building rapport in these conversations, as existing approaches to rapport-building often depend on verbal and nonverbal cues [48]. As part of a larger study focused on building best practices for written hotlines, we worked with a child maltreatment-focused text or chat hotline. This analysis aims to categorize crisis counselors' efforts to build rapport and convey active listening in written conversations. Using these categories, we identified characteristics associated with successful conversations, as reported by help-seekers. This work provides an important foundation for how to build therapeutic relationships in written mental health and hotline services.

Methods

Data Source

The data for this study are from the PACTECH (Prevent Abuse of Children Text and Chat Hotline), the text and chat arm of the Childhelp National Child Abuse Hotline [49]. Since 1982, Childhelp has offered 24/7 phone-based hotline services focused on support and resources related to child maltreatment. In 2018, the hotline expanded to include text and chat capabilities. Crisis counselors are employees rather than volunteers. Most are master-level professionals with specialized training in hotline services and child maltreatment. After conducting a quantitative pilot evaluation for 2 years, hotline leadership partnered with the lead author to use qualitative and mixed method approaches to identify best practices for services. As part of the data sharing agreement, the lead author and her research team received access to deidentified transcripts and metadata from conversations that were purposefully selected to represent a wide range of experiences and perceived outcomes.

Ethical Considerations

The Purdue University Institutional Review Board approved the research protocol (IRB-2020-965). The service terms and conditions disclosed that data may be shared with researchers. As a secondary data analysis of deidentified data, additional consent from participants was not required by the Institutional Review Board. The contract teams from Purdue University and Childhelp negotiated the terms of the data sharing agreement, including data security and access. As a result of the data sharing agreement, the data may not be released publicly.

Sample

The sample consists of 314 purposely selected conversations out of the 1153 text and chat conversations during July 2020. In addition to maintaining the written transcript of the conversation for 60 days, Childhelp collects preconversation and postconversation surveys from the help-seekers. The preconversation surveys focus on help-seeker characteristics (ie, age, gender, state of residence, and referral source), while the postconversation survey focuses on their perceptions of the conversation (eg, do they feel more hopeful, less stressed, and more prepared). We used maximum variation sampling to capture diverse help-seekers and outcomes, although not necessarily in the proportions present in the overall data [50]. This approach is particularly useful when looking for diverse perspectives, as was the case for our study. We sampled based on the preconversation and postconversation surveys. In our sample, 297 (94.6%) help-seekers answered at least 1 presurvey question, and 263 (83.8%) answered at least 1 postconversation survey question. First, we selected conversations where help-seekers reported that they were satisfied, unsatisfied, or mixed. We also included some conversations without surveys to reduce survey response bias. Then, we reviewed the demographic characteristics of the selected conversations to ensure help-seekers of different ages, races or ethnicities, and genders were included in the sample. For example, most help-seekers are girls, so there were relatively few conversations with boys in our initial sample. We added additional conversations with boys to ensure the results were not only relevant to girls.

We analyzed and reported the findings from all 314 conversations. When reporting quotes, however, we were particularly interested in the 45 conversations where help-seekers reported in the postconversation survey that the hotline was a good way to seek help and that they were a lot more hopeful, a lot more informed, a lot more prepared to address the situation, and experiencing less stress. Except when specifically referencing less successful conversations, all example quotes come from these conversations, as they represent those most successful from the help-seekers' perspectives. All quotes are reported verbatim from the conversations, including any errors.

Analysis

We used qualitative content analysis to process the conversations. We used both inductive and deductive processes to develop the codes. The first draft of the coding frame was based on our work with child maltreatment–related conversations within the Crisis Text Line [51-53]. Then, we revised the framework based on the content of the conversations.

Our development process followed the adaptation of grounded theory described by Schreier [54]. The lead author and her graduate research assistant reviewed all the conversations. During a second review of the conversations, we took notes on commonalities within the conversations, emphasizing material not captured in the first draft of the codebook. As we refined the codebook, all team members met weekly to discuss emerging materials and define and develop codes. After completing the framework and definitions, we coded 30 conversations and met to compare the code applications. We discussed differences in coding and refined the framework with the entire team. Then, we coded 30 additional conversations and assessed the coder agreement. After the second round of pilot coding, we reached 95% agreement on the codes and moved to code the full data set. In sum, we had 127 codes in the codebook, which were applied 22,326 times. After coding all the conversations, we reviewed the materials within each code. This process followed

XSL•FO
RenderX

the segmentation process described by Schreier [54], where coded materials are decontextualized and reviewed to identify commonalities and themes. Through this process, we also assessed whether we met saturation, which occurs when all categories have been identified in the data set. Schrier's [54] definition of qualitative content analysis saturation is different from other forms of qualitative methods. In other forms of qualitative analysis, saturation refers to the point at which reviewing additional material does not provide new information. We informally assessed this type of saturation by examining whether all codes were used if we considered only half of the sample. We found that all codes were used when we reviewed 2 different randomly selected split samples, which suggests that few new insights would be gained if we added additional conversations to our sample. After conducting these checks, we categorized the conversations by the outcomes and focused on similarities and differences across the groups.

For this analysis, we focused on the codes related to rapport building and active listening conversations. There were two main types of approaches used by crisis counselors: (1) counseling approaches and (2) evaluation-based language. Active listening skills, otherwise known as attending skills, are how counselors build connections with clients, express empathy, and convey that they are listening [48]. These skills may be defined slightly differently; asking questions, paraphrasing, reflecting feelings, and interpreting or summarizing the situation are generally recognized skills. We added validation [55] and unconditional positive regard [56], which are also commonly incorporated into helping relationships. Evaluation-based language, such as praise and apologies, is commonly used by adults when talking with children [57]. These statements differ from other approaches because the counselor's evaluation of the situation is included.

We also examined how these approaches differed between the help-seekers most satisfied with the conversation (ie, answered all after-conversation survey questions as "Yes") and those who were least satisfied with the conversation. We intended to define the least satisfied as those who answered all the after-conversation survey questions as "No." However, only 4 people fit that criterion, so we included all help-seekers who answered most of the questions negatively.

Research Team

The research team included the lead author, a graduate research assistant, and 2 collaborators at Childhelp. The lead author is a family violence prevention researcher with a PhD in public health and an MA in counseling. She has experience conducting qualitative analyses of written hotline transcripts. The graduate research assistant was a master of public health student and had worked on the lead author's research team for 3 years. She had experience with qualitative child maltreatment research. The Childhelp collaborators have substantial experience in hotline counseling and leadership. One has an MS in counseling psychology. The second has an MS in family and human development and an MEd in guidance counseling.

Results

Help-Seeker Characteristics

Overall, our sample of help-seekers was generally similar to Childhelp's overall text and chat users (Table 1) [58,59]. Help-seekers tended to be female, young, and seeking help for themselves. Overall, they were generally at least a little more hopeful, informed, and prepared to deal with the situation after the conversation (Table 2).

XSL•FO
RenderX

Table 1. Characteristics of help-seekers (n=314).

Characteristics	Help-seekers, n (%)
Gender	
Female	220 (70.1)
Man	53 (16.9)
Gender expansive	22 (7)
No response	19 (6)
Age (years)	
10-11	9 (2.9)
12-13	45 (14.3)
14-15	80 (25.5)
16-17	61 (19.4)
18-24	35 (11.2)
≥25	52 (16.6)
No Response	32 (10.2)
Relationship to maltreated child	
Self[a]	172 (54.8)
Family	40 (12.7)
Friend	37 (11.8)
Other adult	17 (5.4)
Unknown to child	22 (7)
Other	26 (8.3)

[a]Includes children who were distressed but did not necessarily describe events consistent with maltreatment.

XSL•FO
RenderX

Table 2. Help-seekers' perceptions of conversations (n=314).

Perceptions	Help-seekers, n (%)
More positive or hopeful[a]	
A lot	106 (33.8)
A little	121 (38.5)
Not at all	36 (11.5)
No response	51 (16.2)
More informed[b]	
A lot	154 (49)
A little	84 (26.8)
Not at all	21 (6.7)
No response	55 (17.5)
More prepared[c]	
A lot	93 (29.6)
A little	112 (35.7)
Not at all	45 (14.3)
No response	64 (20.4)
Less stressed[d]	
Yes	120 (38.2)
Maybe	60 (19.1)
No	71 (22.6)
No response	63 (20.1)
Good approach[e]	
Yes	205 (65.3)
Maybe	39 (12.4)
No	18 (5.7)
No response	52 (16.6)

[a]Do you feel more positive or hopeful after this chat/text session?

[b]Did you get the information you needed from this chat or text session?

[c]Do you feel better prepared to deal with the situation after this chat or text session?

[d]Do you feel less stress after the chat or text session?

[e]Was using chat or text a good way for you to get help?

Active Listening Skills

Paraphrasing Information and Feelings

When paraphrasing (387 times across 170 conversations), the crisis counselor repeated what was said by the help-seeker in a way that honed the focus of the conversation. Often, it included the most important words shared by the help-seeker, along with a shortened, clarified version of the essential information or feelings. For example, when seeking to understand the situation, a crisis counselor said, "It does not sound like she is able to listen to your needs and wants at this time." At other times, the crisis counselor wanted to convey that they have been listening. Saying, "...you mentioned that they are screaming at him and from what you have said it sounds like they might be being really aggressive with him" demonstrated that they have been paying attention to the information shared.

Sometimes, the crisis counselors reflected the feelings shared by the help-seeker, saying things like, "That sounds like it can be frustrating from what you shared," "it sounds very overwhelming and scary," or "I can see how stressful this is." In these situations, the crisis counselor was often distilling the feelings to support the help-seeker in identifying what is most bothering them about the situation or what feeling is driving their response to the situation. Once the help-seeker recognized the most troubling aspect of the situation, they were often more able to brainstorm ways to address it with the crisis counselor.

XSL•FO
RenderX

Interpretation

Interpreting the situation was also common (236 times across 125 conversations). Often, help-seekers were confused or had ambivalent thoughts about the situation. In these cases, they usually struggled to identify the next steps or reduce their emotional activation. By interpreting the situation, the crisis counselors offered a coherent overview of the situation and a different perspective. In most active listening skills, crisis counselors stayed quite close to the information provided by the help-seeker (eg, paraphrasing or reflecting what was said). When interpreting the situation, crisis counselors often included their perspectives on the situation with the intent of supporting the help-seeker to see themes or new ideas. For example, one help-seeker shared that their caregivers regularly say hurtful things about their gender identity and sexual orientation, scream and yell, and tell the help-seeker that they are a disappointment. In response, the crisis counselor said, "Sounds like it would be very hard to be happy living with people who treat you like that." Although the help-seeker had not overtly shared about their unhappiness, this interpretation led to the help-seeker sharing about active suicidal ideation.

Open Questions

Open questions (208 across 124 conversations) served multiple purposes. At the beginning of the conversations, they invited the help-seeker to share about the experience. For example, "Could you tell me what's going on?" or "What's making you feel unsafe?" was used to begin the conversation in a nonthreatening way. As the conversation moved to explore the issues, open questions could elicit specific details (eg, "What's happened since then?" and "What does that mean?") or focus attention on feelings (eg, "How does it make you feel when your mom lashes out?" and "How are you feeling about all this happening?").

Other Common Counseling Approaches

Validation

Validation was the most used approach to active listening (647 times across 226 conversations), and it took many forms depending on the situation. Throughout the conversations, it was used to affirm the help-seeker, their feelings, and their thoughts. For example, one counselor said, "It can be hard living in a house where you don't feel supported and respected." In this situation, the help-seeker had a difficult relationship with a father, who regularly called the help-seeker "overdramatic or a crybaby." By validating the difficulty of feeling unsupported, the crisis counselor communicated that the help-seeker and their feelings were important.

In other instances, the crisis counselor validated the help-seeker's perspectives about what was or was not appropriate behavior within families. In one instance, a help-seeker shared concerns about an older sibling's treatment of an infant. The brother was rough with the infant and burned the infant with hot milk. In response, the crisis counselor said, "I can see why you would be concerned for the baby's safety." In doing so, the crisis counselor communicated that the help-seeker's feelings were valid but without confirming that the infant was being maltreated. The crisis counselor had not seen evidence of the situation, so they could not accurately validate whether the infant was being maltreated. Simple phrases, such as "I hear you. This is difficult," "That must be really hard for you," and "It's okay to feel stressed that is normal," also validated the help-seeker and their perspectives.

Unconditional Positive Regard

Unconditional positive regard (102 times across 66 conversations) occurred when crisis counselors provided basic acceptance and support of the help-seeker, regardless of their behavior or things that have been done to them. Unconditional positive regard primarily focused on the abuse experience. It was common for counselors to say things like, "No one deserves to be abused" or "No one deserves to be treated like that." These statements were often particularly well received by help-seekers, like this example:

> *You don't deserve to be emotionally abused. It's not o.k.* [Counselor]
>
> *Thank you for saying that. You are the first person ive ever talked about this personally with.* [Help-seeker]

Evaluation-Based Language

Overview

Evaluation-based language involved a judgment by the crisis counselor about whether an aspect of the help-seekers' experiences was good (eg, behavior worthy of praise) or bad (eg, an apology for something that happened to the help-seeker). Evaluation and judgment are generally not a part of helping relationships [48,60,61] but are quite common when adults speak with children [57,62]. Although these approaches are not generally part of counseling relationships, there is nothing inherently wrong with using them intentionally.

Praise

Praise (268 times across 145 conversations) occurred when the crisis counselor conveyed that they approved of the help-seekers or their behavior. Sometimes, praise focused on the behaviors occurring during the conversations, like "Thank you for sharing with me" and "I'm glad you reached out today." At other times, the praise centered on behaviors that they would do in the future, such as "Yes, I believe you're doing the right thing by calling)" and "I think that will be a good move for you."

Apologies

Apologies (372 times across 213 conversations) tended to focus on the help-seeker's situation or issues with the hotline. Apologies for the hotline were usually about a technical difficulty (eg, "sorry, our system is not working well"). Apologies about the help-seeker's situation could be very broad, such as "I'm so sorry to hear about all of this" and "I'm so sorry that you're having to go through this." Apologies could also be specific to the situation, like "I am sorry to hear Mom yelled at you yesterday too."

XSL•FO
RenderX

Differences Between Successful and Less Successful Conversations

There were some differences in active listening skills, other counseling skills, and evaluation-based language between successful and less successful conversations. Although the sample of successful and less successful conversations was too small for formal statistical analysis, some commonalities emerged. First, although conversations were approximately the same length, less successful conversations tended to have more statements that attended to rapport building. Second, there were also differences in how the counselors applied these approaches. Unlike the preceding sections, this section includes quotes from both successful and less successful conversations.

Overall, counselors in less successful conversations tended to be vague or to directly repeat what was said by the help-seeker. These differences were particularly apparent when counselors were paraphrasing, asking open questions, or apologizing. For example, paraphrasing in less successful conversations tended to be either very vague (eg, “It sounds like you are being hurt already”) or very specific (eg, “I am hearing you have some future plans to get a job and earn your own money...”). In the last example, the help-seeker used the same phrasing in their previous statement. Conversely, successful conversations tended to be specific without direct repetition (eg, “Sounds like they are something to help you cope”). Similarly, less successful conversations tended to include open questions that were either broad (eg, “What’s happening?”) or focused on clarifying how the crisis counselor could help (eg, “How are you hoping that I can help?”). Some successful conversations also included questions clarifying how the crisis counselor could help, but it was more common to ask more specific questions, like “What is it that you would like to vent about?” and “What are their thoughts on CPS involvement?” Finally, crisis counselors used generally vague apologies about the situation in less successful conversations. Saying things like “I am so sorry this happened to you” or “I’m sorry to hear that” was common. Although some successful conversations also included these types of apologies, it was more common to pair the apology with a specific reason, such as “I’m so sorry that you have been experiencing this for so long” or “I am sorry to hear Mama is sick.”

Discussion

Principal Results

Overall, our study suggests that it is possible to build therapeutic relationships via a text and chat hotline with individuals seeking child maltreatment–related information and support. Approximately 15% (n=45) of our sample reported that the hotline was a good way to seek help and that they were a lot more hopeful, a lot more informed, a lot more prepared to address the situation, and experiencing less stress. However, our sample was intentionally selected to represent a wide range of help-seeker perceptions, so this does not indicate that 15% of the hotline’s help-seekers felt this way. Based on the 2022 Childhelp data report, about 85% of help-seekers reported getting the information they needed, 80% of help-seekers reported feeling more hopeful after the conversation, and 75% reported feeling better prepared to deal with the situation [59]. These percentages suggest that the hotline provides a well-received service.

Generally, counselors built rapport through active listening skills, other counseling techniques, and evaluation-based language (ie, apologies, praise). Through active listening skills and other counseling techniques, counselors often expressed that they were listening, wanted to understand the help-seekers, and cared for them. They expressed their approval or disapproval of the help-seekers and aspects of their experiences through evaluation-based language. Although there is nothing inherently wrong with using apologies and praise, they tend to be avoided in many therapeutic approaches. Praise may undermine intrinsic motivation (ie, internal drive) and reduce engagement in the process [63-65]. Further, these types of evaluation-based language are rooted in control, as they are given based on something that another individual (ie, the crisis counselor) deems desirable [63]. As a result, the help-seeker might seek praise by giving answers that they believe the crisis counselor wants to receive instead of accurate answers, which may reduce the benefit of the conversation. However, praise and compliments may be a quick way to build encouraging feelings [66]. As it is challenging to build relationships via writing, praise may be one way to build a relationship quickly. Additional research into the impact of evaluation-based language is necessary to understand its role in written crisis counseling.

There were some differences between successful and less successful conversations. Surprisingly, less successful conversations tended to include more attending language than successful conversations. However, there were differences in the ways that crisis counselors apply these techniques. Overall, the crisis counselors in successful conversations tended to be more specific and tailor their responses to the help-seekers. Possibly, counselors who gave tailored responses built rapport more quickly; thus, fewer attending statements were required. If this is the case, they could move to problem-solving more quickly, which may also contribute to help-seekers’ perceptions that they were more prepared to address the situation and were more informed. These tailored responses may increase help-seekers’ perceptions that the crisis counselor is invested in the conversation. Several help-seekers explicitly asked if they were speaking with a bot in this sample. Having tailored responses may increase crisis counselors’ social presence and reduce help-seekers’ concerns about whether they are “real.” As organizations consider using large language models and chatbots in these types of services, careful attention should be given to help-seekers’ perceptions about the service and its appropriateness for the audience. As the National Eating Disorders Association learned when its wellness chatbot began providing diet information, large language models trained on outside data may not be a good fit for conversations with help-seekers [67].

Limitations

Our work has several limitations, including some inherent to secondary data analysis. First, we could not speak with the help-seekers or the counselors about the conversations. Although we were able to identify similarities across well-received conversations, it is possible that other aspects of the

XSL•FO
RenderX

conversations contributed to help-seekers' perceptions. Second, we do not know how these conversations shaped long-term outcomes. Moreover, it is difficult to follow up with help-seekers, as evidenced by the 6% response rate to a 2-week follow-up survey conducted by the National Domestic Violence Hotline [68]. Further, many of the help-seekers in this sample indicated that it is unsafe to speak aloud about their experiences, so qualitative data collection with this sample would likely have an even lower response rate. It would be more feasible to speak with counselors about their experiences, but their perspectives may be disconnected from those of the help-seekers. Despite this limitation, we incorporated the help-seekers' perspectives through the postconversation survey, which is more than is usually possible in secondary data analysis.

Our work may not generalize to conversations unrelated to child maltreatment. As a child maltreatment–specific hotline, all conversations included elements of child maltreatment. Conversations about other topics may require other approaches. However, our results are consistent with prior work on building rapport in other forms of counseling [48,55,56], so it is reasonable to expect these findings would translate to written conversations about other topics.

Comparison With Prior Work

To the best of our knowledge, there is no other work examining specific ways to build a therapeutic relationship in written mental health counseling or crisis counseling. However, the ways that crisis counselors attended to the dynamics of the conversations were generally like those found in in-person counseling [48,55,56].

Telehealth approaches to counseling may be particularly important for young people experiencing maltreatment. Other formal resources, such as law enforcement, schools, and child protection systems, often fail to respond adequately [53,69,70]. Further, internet-based approaches, particularly written approaches, are highly acceptable to young people experiencing maltreatment [69]. In our sample and past research, children shared that they could not call resources because an audible conversation would cause parents to know they were seeking help. In work conducted with Crisis Text Line, it was common for young people sharing child maltreatment to report that the abuse escalated when parents discovered their attempts to seek help. Written, anonymous communication that is available 24/7 may be a safer way for these young people to seek help. Thus, written communication may be particularly important for children in unsafe homes.

There is also limited evidence on how to respond when young people share maltreatment experiences. Regardless of the ability to impact or end the maltreatment, individuals who receive a child maltreatment disclosure need to receive an appropriate, supportive response [71-74]. Supportive responses encourage the young person experiencing maltreatment to reframe their experience, which substantially reduces the likelihood of poor outcomes otherwise associated with maltreatment [75]. Conversely, unsupportive experiences often have long-lasting consequences [74,76]. Receiving a hurtful or unsupportive response increases the likelihood that the young person will experience more significant physical and mental health issues [74,76,77]. Unfortunately, many young people receive unsupportive responses to their disclosures [53,70]. Often, they report that others, particularly adults, do not believe them and are unwilling to help [70,78]. These experiences reduce their willingness to seek help or share their experiences in the future [70]. Our work suggests that responding to these disclosures adequately in written conversations is possible.

Our work also contributes to a small body of literature on using text and chat hotlines to provide services to people experiencing violence more generally. Michigan State University added chat services to its existing sexual assault support and advocacy hotline. Their evaluation was consistent with many of the benefits and limitations of other forms of written counseling, including challenges with nuance, misunderstanding written language, and communicating empathy [3]. However, the format also gave help-seekers a greater sense of control [3]. Another study focused on agencies providing digital violence-related support and advocacy services [32]. This work also emphasized the importance of clear communication and building rapport, although help-seeker perceptions of these factors were not assessed [32].

Conclusions

Building therapeutic relationships and social presence are important components of digital interventions involving mental health professionals. Prior research suggests that they can be challenging to develop in written conversations. Our work demonstrates characteristics of conversations associated with greater satisfaction among help-seekers. These findings may be adopted by other organizations building mental health or support interventions that include written communication. However, additional research is needed to identify how to train providers to adopt these strategies while also tailoring their approach to the help-seeker. Further, our findings may inform future work with large language models, including how large language models could contribute to these interventions. However, future research is needed to understand how help-seekers would interface with these methods and to ensure that the models consistently convey appropriate, supportive information.

Acknowledgments

This project was supported by the Children's Bureau (CB) and Administration for Children and Families (ACF) of the US Department of Health and Human Services (HHS) as part of a financial assistance award in the amount of US $6,000,000 that was 100% funded by the CB and ACF of the HHS. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, the CB and ACF of the HHS or the US government. For more information, please visit administrative and national policy requirements.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Murphy LJ, Mitchell DL. When writing helps to heal: e-mail as therapy. Br J Guid Counc 1998;26(1):21-32. [doi: 10.1080/03069889808253835]
2. King R, Bambling M, Lloyd C, Gomurra R, Smith S, Reid W, et al. Online counselling: the motives and experiences of young people who choose the internet instead of face to face or telephone counselling. Couns Psychother Res 2006;6(3):169-174. [doi: 10.1080/14733140600848179]
3. Moylan CA, Carlson ML, Campbell R, Fedewa T. "It's hard to show empathy in a text": developing a web-based sexual assault hotline in a college setting. J Interpers Violence 2022;37(17-18):NP16037-NP16059. [doi: 10.1177/08862605211025036] [Medline: 34134570]
4. Norwood C, Moghaddam NG, Malins S, Sabin-Farrell R. Working alliance and outcome effectiveness in videoconferencing psychotherapy: a systematic review and noninferiority meta-analysis. Clin Psychol Psychother 2018;25(6):797-808. [doi: 10.1002/cpp.2315] [Medline: 30014606]
5. Cheng N, Mohiuddin S. Addressing the nationwide shortage of child and adolescent psychiatrists: determining factors that influence the decision for psychiatry residents to pursue child and adolescent psychiatry training. Acad Psychiatry 2022;46(1):18-24. [doi: 10.1007/s40596-021-01554-4] [Medline: 34767240]
6. Morales DA, Barksdale CL, Beckel-Mitchener AC. A call to action to address rural mental health disparities. J Clin Transl Sci 2020;4(5):463-467 [FREE Full text] [doi: 10.1017/cts.2020.42] [Medline: 33244437]
7. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
8. Spek V, Cuijpers P, Nyklícek I, Riper H, Keyzer J, Pop V. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. Psychol Med 2007;37(3):319-328. [doi: 10.1017/S0033291706008944] [Medline: 17112400]
9. Torous J, Hsin H. Empowering the digital therapeutic relationship: virtual clinics for digital health interventions. NPJ Digit Med 2018;1:16 [FREE Full text] [doi: 10.1038/s41746-018-0028-2] [Medline: 31304301]
10. Karver MS, Handelsman JB, Fields S, Bickman L. Meta-analysis of therapeutic relationship variables in youth and family therapy: the evidence for different relationship variables in the child and adolescent treatment outcome literature. Clin Psychol Rev 2006;26(1):50-65. [doi: 10.1016/j.cpr.2005.09.001] [Medline: 16271815]
11. Finsrud I, Nissen-Lie HA, Vrabel K, Høstmælingen A, Wampold BE, Ulvenes PG. It's the therapist and the treatment: the structure of common therapeutic relationship factors. Psychother Res 2022;32(2):139-150 [FREE Full text] [doi: 10.1080/10503307.2021.1916640] [Medline: 33938407]
12. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic) 2018;55(4):399-410 [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
13. Nienhuis JB, Owen J, Valentine JC, Black SW, Halford TC, Parazak SE, et al. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: a meta-analytic review. Psychother Res 2018;28(4):593-605. [doi: 10.1080/10503307.2016.1204023] [Medline: 27389666]
14. Short J, Williams E, Christie B. The Social Psychology of Telecommunications. Hoboken, NJ: Jon Wiley & Sons; 1976.
15. Gunawardena CN. Social presence theory and implications for interaction and collaborative learning in computer conferences. Int J Educ Telecommun 1995;1(2):147-166 [FREE Full text]
16. Batish R. Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger. Birmingham, UK: Packt Publishing Ltd; 2018.
17. Warwick K, Shah H. The importance of a human viewpoint on computer natural language capabilities: a turing test perspective. AI Soc 2016;31(2):207-221. [doi: 10.1007/s00146-015-0588-5]
18. Lopez A. An investigation of the use of internet based resources in support of the therapeutic alliance. Clin Soc Work J 2014;43(2):189-200. [doi: 10.1007/s10615-014-0509-y]
19. Holmes C, Foster V. A preliminary comparison study of online and face-to-face counseling: client perceptions of three factors. J Technol Hum Serv 2012;30(1):14-31. [doi: 10.1080/15228835.2012.662848]
20. Bantjes J, Slabbert P. The digital therapeutic relationship: retaining humanity in the digital age. In: Stein DJ, Fineberg NA, Chamberlain SR, editors. Mental Health in a Digital World. Amsterdam: Elsevier; 2022:223-237.
21. Berger T. The therapeutic alliance in internet interventions: a narrative review and suggestions for future research. Psychother Res 2017;27(5):511-524. [doi: 10.1080/10503307.2015.1119908] [Medline: 26732852]
22. Richards D, Viganó N. Online counseling: a narrative and critical review of the literature. J Clin Psychol 2013;69(9):994-1011. [doi: 10.1002/jclp.21974] [Medline: 23630010]
23. Chechele PJ, Stofle G. Individual therapy online via email and internet relay chat. In: Anthony K, editor. Technology in Counselling and Psychotherapy: A Practitioner's Guide. London: Palgrave Macmillan; 2003:39-58.

XSL•FO
RenderX

24. Stoll J, Müller JA, Trachsel M. Ethical issues in online psychotherapy: a narrative review. Front Psychiatry 2019;10:993 [FREE Full text] [doi: 10.3389/fpsyt.2019.00993] [Medline: 32116819]
25. Kessler D, Lewis G, Kaur S, Wiles N, King M, Weich S, et al. Therapist-delivered internet psychotherapy for depression in primary care: a randomised controlled trial. Lancet 2009;374(9690):628-634. [doi: 10.1016/S0140-6736(09)61257-5] [Medline: 19700005]
26. Beattie A, Shaw A, Kaur S, Kessler D. Primary-care patients' expectations and experiences of online cognitive behavioural therapy for depression: a qualitative study. Health Expect 2009;12(1):45-59 [FREE Full text] [doi: 10.1111/j.1369-7625.2008.00531.x] [Medline: 19250152]
27. Vernmark K, Lenndin J, Bjärehed J, Carlsson M, Karlsson J, Oberg J, et al. Internet administered guided self-help versus individualized e-mail therapy: a randomized trial of two versions of CBT for major depression. Behav Res Ther 2010;48(5):368-376. [doi: 10.1016/j.brat.2010.01.005] [Medline: 20152960]
28. Andersson G, Paxling B, Roch-Norlund P, Östman G, Norgren A, Almlöv J, et al. Internet-based psychodynamic versus cognitive behavioral guided self-help for generalized anxiety disorder: a randomized controlled trial. Psychother Psychosom 2012;81(6):344-355 [FREE Full text] [doi: 10.1159/000339371] [Medline: 22964540]
29. Dowling M, Rickwood D. Investigating individual online synchronous chat counselling processes and treatment outcomes for young people. Adv Ment Health 2015;12(3):216-224. [doi: 10.1080/18374905.2014.11081899]
30. King R, Bambling M, Reid W, Thomas I. Telephone and online counselling for young people: a naturalistic comparison of session outcome, session impact and therapeutic alliance. Couns Psychother Res 2006;6(3):175-181. [doi: 10.1080/14733140600874084]
31. Francis-Smith C. Email counselling and the therapeutic relationship: a grounded theory analysis of therapists' experiences [dissertation]. University of the West of England. 2014. URL: https://uwe-repository.worktribe.com/index.php/preview/806312/Thesis%20amended%20for%20repository.pdf [accessed 2024-04-17]
32. Wood L, Hairston D, Schrag RV, Clark E, Parra-Cardona R, Temple JR. Creating a digital trauma informed space: chat and text advocacy for survivors of violence. J Interpers Violence 2022;37(19-20):NP18960-NP18987. [doi: 10.1177/08862605211043573] [Medline: 34715764]
33. Gould MS, Chowdhury S, Lake AM, Galfalvy H, Kleinman M, Kuchuk M, et al. National suicide prevention lifeline crisis chat interventions: evaluation of chatters' perceptions of effectiveness. Suicide Life Threat Behav 2021;51(6):1126-1137 [FREE Full text] [doi: 10.1111/sltb.12795] [Medline: 34331471]
34. Gibson K, Cartwright C. Young people's experiences of mobile phone text counselling: balancing connection and control. Child Youth Serv Rev 2014;43:96-104. [doi: 10.1016/j.childyouth.2014.05.010]
35. Evans WP, Davidson L, Sicafuse L. Someone to listen: increasing youth help-seeking behavior through a text-based crisis line for youth. J Community Psychol 2013;41(4):471-487. [doi: 10.1002/jcop.21551]
36. Predmore Z, Ramchand R, Ayer L, Kotzias V, Engel C, Ebener P, et al. Expanding suicide crisis services to text and chat. Crisis 2017;38(4):255-260. [doi: 10.1027/0227-5910/a000460] [Medline: 28468558]
37. Chardon L, Bagraith KS, King RJ. Counseling activity in single-session online counseling with adolescents: an adherence study. Psychother Res 2011;21(5):583-592. [doi: 10.1080/10503307.2011.592550] [Medline: 21756193]
38. Bambling M, King R, Reid W, Wegner K. Online counselling: the experience of counsellors providing synchronous single-session counselling to young people. Couns Psychother Res 2008;8(2):110-116. [doi: 10.1080/14733140802055011]
39. Rodda SN, Lubman DI, Cheetham A, Dowling NA, Jackson AC. Single session web-based counselling: a thematic analysis of content from the perspective of the client. Br J Guid Counc 2015;43(1):117-130. [doi: 10.1080/03069885.2014.938609]
40. Fukkink RG, Hermanns JMA. Children's experiences with chat support and telephone support. J Child Psychol Psychiatry 2009;50(6):759-766. [doi: 10.1111/j.1469-7610.2008.02024.x] [Medline: 19207634]
41. Fukkink R, Hermanns J. Counseling children at a helpline: chatting or calling? Am J Community Psychol 2009;37(8):939-948. [doi: 10.1002/jcop.20340]
42. Sindahl TN, van Dolen W. Texting at a child helpline: how text volume, session length and duration, response latency, and waiting time are associated with counseling impact. Cyberpsychol Behav Soc Netw 2020;23(4):210-217 [FREE Full text] [doi: 10.1089/cyber.2019.0100] [Medline: 31990599]
43. van Dolen W, Weinberg CB. Child helplines: how social support and controllability influence service quality and well-being. J Serv Mark 2017;31(4/5):385-396. [doi: 10.1108/jsm-07-2016-0254]
44. van Dolen W, Weinberg CB. An empirical investigation of factors affecting perceived quality and well-being of children using an online child helpline. Int J Environ Res Public Health 2019;16(12):2193 [FREE Full text] [doi: 10.3390/ijerph16122193] [Medline: 31234285]
45. Navarro P, Bambling M, Sheffield J, Edirippulige S. Exploring young people's perceptions of the effectiveness of text-based online counseling: mixed methods pilot study. JMIR Ment Health 2019;6(7):e13152 [FREE Full text] [doi: 10.2196/13152] [Medline: 31271149]
46. Thompson LK, Sugg MM, Runkle JR. Adolescents in crisis: a geographic exploration of help-seeking behavior using data from crisis text line. Soc Sci Med 2018;215:69-79. [doi: 10.1016/j.socscimed.2018.08.025] [Medline: 30216891]
47. Fildes D, Williams K, Bradford S, Grootemaat P, Kobel C, Gordon R. Implementation of a pilot SMS-based crisis support service in Australia. Crisis 2022;43(1):46-52. [doi: 10.1027/0227-5910/a000758] [Medline: 33475017]

XSL•FO
RenderX

48. Ivey AE, Packard NG, Ivey MB. Basic Attending Skills. San Diego, CA: Cognella; 2018.
49. The Childhelp National Child Abuse Hotline. Childhelp. 2020. URL: https://www.childhelp.org/hotline/ [accessed 2024-04-17]
50. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
51. Cash SJ, Murfree L, Schwab-Reese L. "I'm here to listen and want you to know I am a mandated reporter": understanding how text message-based crisis counselors facilitate child maltreatment disclosures. Child Abuse Negl 2020;102:104414 [FREE Full text] [doi: 10.1016/j.chiabu.2020.104414] [Medline: 32070934]
52. Schwab-Reese L, Kanuri N, Cash S. Child maltreatment disclosure to a text messaging-based crisis service: content analysis. JMIR Mhealth Uhealth 2019;7(3):e11306 [FREE Full text] [doi: 10.2196/11306] [Medline: 30907745]
53. Schwab-Reese LM, Cash SJ, Lambert NJ, Lansford JE. "They aren't going to do jack shit": text-based crisis service users' perceptions of seeking child maltreatment-related support from formal systems. J Interpers Violence 2022;37(19-20):NP19066-NP19083 [FREE Full text] [doi: 10.1177/08862605211043577] [Medline: 34507498]
54. Schreier M. Qualitative Content Analysis in Practice. Thousand Oaks, CA: Sage; 2012.
55. Linehan MM. Validation and psychotherapy. In: Bohart AC, Greenberg LS, editors. Empathy Reconsidered: New Directions in Psychotherapy. Washington, DC: American Psychological Association; 1997:353-392.
56. Wilkins P. Unconditional positive regard reconsidered. Br J Guid Counc 2010;28(1):23-36. [doi: 10.1080/030698800109592]
57. Brummelman E, Crocker J, Bushman BJ. The praise paradox: when and why praise backfires in children with low self-esteem. Child Dev Perspect 2016;10(2):111-115. [doi: 10.1111/cdep.12171]
58. Wolfersteig W, Moreland D, Diaz M, Gotlieb E. Prevent Abuse of Children Text and Chat Hotline (PACTECH) project: semi-annual data report. Childhelp. Scottsdale, Arizona; 2022. URL: https://www.childhelphotline.org/wp-content/uploads/2022/05/PACTECH-Data-Report-April-2022.pdf [accessed 2024-04-17]
59. Hotline impact report. Childhelp. 2022. URL: https://www.childhelphotline.org/wp-content/uploads/2022/10/Hotline-Impact-Report-FY22.pdf [accessed 2024-04-17]
60. Nicholas A, Pirkis J, Reavley N. What responses do people at risk of suicide find most helpful and unhelpful from professionals and non-professionals? J Ment Health 2022;31(4):496-505. [doi: 10.1080/09638237.2020.1818701] [Medline: 32930018]
61. Rogers CR. A Way of Being. Boston, MA: Houghton Mifflin Harcourt; 1980.
62. Brummelman E, Nelemans SA, Thomaes S, de Castro BO. When parents' praise inflates, children's self-esteem deflates. Child Dev 2017;88(6):1799-1809. [doi: 10.1111/cdev.12936] [Medline: 28857141]
63. Kelsey J. The negative impact of rewards and ineffective praise on student motivation. ESSAI 2011;8(1):24 [FREE Full text]
64. Kakinuma K, Nishiguti F, Sonoda K, Tajiri H, Tanaka A. The negative effect of ability-focused praise on the "praiser's" intrinsic motivation: face-to-face interaction. Front Psychol 2020;11:562081 [FREE Full text] [doi: 10.3389/fpsyg.2020.562081] [Medline: 33329190]
65. Pellecchia M, Nuske HJ, Straiton D, Hassrick ME, Gulsrud A, Iadarola S, et al. Strategies to engage underrepresented parents in child intervention services: a review of effectiveness and co-occurring use. J Child Fam Stud 2018;27(10):3141-3154. [doi: 10.1007/s10826-018-1144-y]
66. Landrum RE, Gurung RA, Nolan SA, McCarthy MA, Dunn DS. Everyday Applications of Psychological Science: Hacks to Happiness and Health. Milton Park, UK: Routledge; 2022.
67. McCarthy L. A wellness chatbot is offline after its 'harmful' focus on weight loss. The New York Times. 2023. URL: https://www.nytimes.com/2023/06/08/us/ai-chatbot-tessa-eating-disorders-association.html [accessed 2024-04-17]
68. McDonnell K, Nagaraj N, Fuerst M. Short-term outcomes following contact with the national domestic violence hotline and loveisrespect. U.S. Department of Health & Human Services. 2020. URL: https://www.acf.hhs.gov/opre/report/short-term-outcomes-following-contact-national-domestic-violence-hotline-and [accessed 2024-04-17]
69. Al-Eissa MA. Utilization of child helpline (CHL) among adolescents in Saudi Arabia: results from a national survey. Child Fam Soc Work 2019;24(1):84-89. [doi: 10.1111/cfs.12584]
70. Tucker S. Listening and believing: an examination of young people's perceptions of why they are not believed by professionals when they report abuse and neglect. Child Soc 2011;25(6):458-469. [doi: 10.1111/j.1099-0860.2010.00291.x]
71. Collin-Vézina D, De La Sablonnière-Griffin M, Palmer AM, Milne L. A preliminary mapping of individual, relational, and social factors that impede disclosure of childhood sexual abuse. Child Abuse Negl 2015;43:123-134. [doi: 10.1016/j.chiabu.2015.03.010] [Medline: 25846196]
72. Goodman-Brown TB, Edelstein RS, Goodman GS, Jones DPH, Gordon DS. Why children tell: a model of children's disclosure of sexual abuse. Child Abuse Negl 2003;27(5):525-540. [doi: 10.1016/s0145-2134(03)00037-1] [Medline: 12718961]
73. Jensen TK, Gulbrandsen W, Mossige S, Reichelt S, Tjersland OA. Reporting possible sexual abuse: a qualitative study on children's perspectives and the context for disclosure. Child Abuse Negl 2005;29(12):1395-1413. [doi: 10.1016/j.chiabu.2005.07.004] [Medline: 16293304]

XSL•FO
RenderX

74. Palmer SE, Brown RA, Rae-Grant NI, Loughlin MJ. Responding to children's disclosure of familial abuse: what survivors tell us. Child Welfare 1999;78(2):259-282. [Medline: 10418117]
75. Briere J, Jordan CE. Violence against women: outcome complexity and implications for assessment and treatment. J Interpers Violence 2004;19(11):1252-1276. [doi: 10.1177/0886260504269682] [Medline: 15534329]
76. Arata CM. To tell or not to tell: current functioning of child sexual abuse survivors who disclosed their victimization. Child Maltreatment 1998;3(1):63-71. [doi: 10.1177/1077559598003001006]
77. Palo AD, Gilbert BO. The relationship between perceptions of response to disclosure of childhood sexual abuse and later outcomes. J Child Sex Abus 2015;24(5):445-463. [doi: 10.1080/10538712.2015.1042180] [Medline: 26301435]
78. Cossar J, Belderson P, Brandon M. Recognition, telling and getting help with abuse and neglect: young people's perspectives. Child Youth Serv Rev 2019;106:104469. [doi: 10.1016/j.childyouth.2019.104469]

Edited by T de Azevedo Cardoso; submitted 19.08.22; peer-reviewed by K Zhang, V Franzoni, B Li, Z Aghaei; comments to author 28.03.23; revised version received 21.08.23; accepted 26.03.24; published 15.05.24.

Please cite as:
Schwab-Reese L, Short C, Jacobs L, Fingerman M
Rapport Building in Written Crisis Services: Qualitative Content Analysis
J Med Internet Res 2024;26:e42049
URL: https://www.jmir.org/2024/1/e42049
doi:10.2196/42049
PMID:38748472

©Laura Schwab-Reese, Caitlyn Short, Larel Jacobs, Michelle Fingerman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Dynamics and Thematic Exploration in r/Depression During the COVID-19 Pandemic: Insights From Overlapping r/SuicideWatch Users

Jianfeng Zhu[1], PhD; Ruoming Jin[1], Prof Dr; Deric R Kenne[2], Prof Dr; NhatHai Phan[3], Prof Dr; Wei-Shinn Ku[4], Prof Dr

[1]Department of Computer Science, Kent State University, Kent, OH, United States

[2]Center for Public Policy and Health, College of Public Health, Kent State University, Kent, OH, United States

[3]Department of Data Science, New Jersey Institute of Technology, Newark, NJ, United States

[4]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, United States

Corresponding Author:
Jianfeng Zhu, PhD
Department of Computer Science
Kent State University
800 E. Summit St.
Kent, OH, 44242
United States
Phone: 1 3306729980
Email: jzhu10@kent.edu

Abstract

Background: In 2023, the United States experienced its highest- recorded number of suicides, exceeding 50,000 deaths. In the realm of psychiatric disorders, major depressive disorder stands out as the most common issue, affecting 15% to 17% of the population and carrying a notable suicide risk of approximately 15%. However, not everyone with depression has suicidal thoughts. While “suicidal depression” is not a clinical diagnosis, it may be observed in daily life, emphasizing the need for awareness.

Objective: This study aims to examine the dynamics, emotional tones, and topics discussed in posts within the r/Depression subreddit, with a specific focus on users who had also engaged in the r/SuicideWatch community. The objective was to use natural language processing techniques and models to better understand the complexities of depression among users with potential suicide ideation, with the goal of improving intervention and prevention strategies for suicide.

Methods: Archived posts were extracted from the r/Depression and r/SuicideWatch Reddit communities in English spanning from 2019 to 2022, resulting in a final data set of over 150,000 posts contributed by approximately 25,000 unique overlapping users. A broad and comprehensive mix of methods was conducted on these posts, including trend and survival analysis, to explore the dynamic of users in the 2 subreddits. The BERT family of models extracted features from data for sentiment and thematic analysis.

Results: On August 16, 2020, the post count in r/SuicideWatch surpassed that of r/Depression. The transition from r/Depression to r/SuicideWatch in 2020 was the shortest, lasting only 26 days. Sadness emerged as the most prevalent emotion among overlapping users in the r/Depression community. In addition, physical activity changes, negative self-view, and suicidal thoughts were identified as the most common depression symptoms, all showing strong positive correlations with the emotion tone of disappointment. Furthermore, the topic “struggles with depression and motivation in school and work” (12%) emerged as the most discussed topic aside from suicidal thoughts, categorizing users based on their inclination toward suicide ideation.

Conclusions: Our study underscores the effectiveness of using natural language processing techniques to explore language markers and patterns associated with mental health challenges in online communities like r/Depression and r/SuicideWatch. These insights offer novel perspectives distinct from previous research. In the future, there will be potential for further refinement and optimization of machine classifications using these techniques, which could lead to more effective intervention and prevention strategies.

(J Med Internet Res 2024;26:e53968) doi:10.2196/53968

XSL•FO
RenderX

KEYWORDS

reddit; natural language processing; NLP; suicidal ideation; SI; online communities; depression symptoms; COVID-19 pandemic; bidirectional encoder representations from transformers; BERT; r/SuicideWatch; r/Depression

Introduction

Background

In the United States, approximately 36% of Americans have self-reported experiences of serious loneliness during the COVID-19 pandemic, with an even more concerning 43% having acknowledged an intensified sense of isolation since its inception [1]. This intensified loneliness, compounded by the challenges of the pandemic, has worsened the struggles of individuals already dealing with preexisting mental health issues. Suicidal ideation (SI) is common among people with mental health problems, particularly depression [2], and the combination of pandemic-induced loneliness and existing mental health concerns further escalates the risk.

Identifying individuals at risk of suicide can be challenging due to the overlap between its warning signs and the symptoms of depression. Key indicators include significant changes in eating and sleep patterns; loss of interest in work, school, and community; social withdrawal; reckless behaviors; substance abuse; neglect of personal appearance; and an obsession with death. These challenging circumstances have created a fertile ground for a distressing increase in substance abuse; traumatic stressors; and, most alarmingly, suicidal thoughts [3]. However, detecting and measuring these thoughts is not easy, as thoughts of death, dying, or suicidal plans are often not expressed directly. The challenge is further amplified by the fact that such thoughts can change in intensity and frequency [4]. Clinicians are now requested to provide an in-depth investigation into the mind of a person who is suicidal, an assessment adjunctive to the psychiatric evaluation [5].

A closer examination reveals that working with individuals at risk of suicide reveals a complexity that defies the constraints of the outdated medical model. Warning signs often precede suicidal thoughts and actions, and individuals with suicidal tendencies frequently grapple with mental anguish as a common thread woven through various adverse life events [6]. In contrast, the widespread use of social media platforms offers a new way for people to express their emotions continuously. This digital space has the potential to better understand people's mental health. Importantly, using social media platforms allows researchers to study larger groups, making their findings stronger and more useful.

Our main goal is to gain insights into early signs of distress by analyzing social media content, focusing on transition patterns, emotion tones, and themes associated with suicidal depression. We achieve this through the application of natural language processing (NLP) and statistical methodology. By comprehending the multifaceted perspectives, coupled with appropriate interventions, depression can be treated and suicide can be prevented.

Prior Research

The emergence of online communities, largely facilitated by platforms like social media, has carved out a space for people to connect, share experiences, and support each other [3]. This surge in online activity has also empowered the social NLP research community, allowing for advancements in understanding mental health indicators from self-reported text [7].

Research in this area has focused on extracting features and patterns from social media data to identify potential risks, and various behavioral features such as the number of posts, the rate of retweets [8], the frequency of absolutist word [9], and the prevalence of emotional characteristics [10,11]. In addition, studies have explored the transition of users to subreddits like r/SuicideWatch [11], harnessing social media data in identifying mental health symptoms [12,13]. Meanwhile, another line of investigation focuses on detecting and predicting different mental health categories. Sarkar et al [12] explored the prediction of depression and anxiety on Reddit using a multitask learning approach that combines word embedding features, pretrained BERT models, and topic modeling features (ie, latent Dirichlet allocation and BERTopic). In a parallel, Kayalvizhi and Thenmozhi [13] established a gold standard data set for identifying levels of depression, such as not depressed, moderately depressed, and severely depressed from social media posts . In addition, Ghosal and Jain [14] proposed a novel framework to differentiate between depression and suicidal risk content using fastText embedding for contextual analysis, Term Frequency–Inverse Document Frequency vector for the relevance of terms, and machine learning classifier XGBoost for accurate classification. More Recently, Cai et al [15] leveraged depressive symptom score calculation to obtain the cosine similarity between posts and depressive symptom descriptions and then predict the presence of depression.

NLP has demonstrated strong performance in tasks such as text classification and sentiment mining. Tadesse et al [16] effectively used NLP techniques to identify specific terms indicating depression, emphasizing the potential of NLP in uncovering mental health indicators. In addition, Zhang et al [17] used NLP to characterize changes in mental health support groups on Reddit during the pandemic. The use of BERT-based models has shown promise in detecting mental health illnesses [17]. Kim et al [18] used BERT to extract textual features from mental health subreddits, analyzing linguistic characteristics. Demszky et al [19] used BERT in transfer learning experiments to assess emotional content on Reddit. Furthermore, Ramirez-Cifuentes et al [20] developed predictive models for substance abuse and mental health issues using enhanced representations with statistical and deep learning approaches on social media posts. Baird et al [21] and Alhaj et al [22] used BERTopic to analyze mental health texts on social media platform, enhancing classification accuracy.

XSL•FO
RenderX

However, existing studies [8-19,21-25] have several challenges, including limited data set sizes and time frames and a focus on specific mental health issues without considering the overlap between them. To address these challenges, our study focused on posts from overlapping users in r/Depression and r/SuicideWatch, using trend analysis and survival analysis, emotion feature identification, and machine learning classification to gain insights into posting behavior, emotional expressions, and prevalent topics associated with individuals who may be struggling with suicidal depression.

Our research aims to answer the following research questions (RQs): (RQ1) What are the patterns and trends in user engagement behaviors across both r/Depression and r/SuicideWatch subreddits over time? (RQ2) How do users transition between participating in r/Depression to engaging with r/SuicideWatch subreddits? (RQ3) What emotional expressions and signs of depression are prevalent among overlapping users in r/Depression? (RQ4) What topics are discussed by overlapping users in the r/Depression, categorized based on SI labels?

Methods

Data Collection

In our research, we used Reddit, a widely used online platform, to collect data from 2 mental health–related subreddits: r/Depression and r/SuicideWatch. To compile our data set, we accessed the archives of Reddit submission [24], covering the time span from 2019 to 2022. During this data collection period, we gathered around half a million posts. Specifically, we collected over 285,000 posts from >165,000 unique users in the r/SuicideWatch subreddit. For the r/Depression subreddit, our data set included >444,000 posts contributed by over 248,000 distinct users.

While we do not possess detailed demographic information about users within these specific subreddits, we do have insights into the broader Reddit user base. Reddit boasts an impressive 430 million monthly active users, with 74% of them being men and 25.8% being women. Most users fall within the age range of 18 to 29, and a substantial 63% of the users hold at least a bachelor's degree [25]. This reflects the quality of discussions and content found on Reddit, as users tend to be well-informed and educated on a variety of topics. This statistic is also important to consider when creating content for Reddit, as it suggests that the audience is likely to be more sophisticated than other social media platforms.

Data Cleaning

The proper execution of data preprocessing techniques is essential to ensure the accuracy of subsequent data analysis [26]. In this study, we used thematic and topic modeling strategies that generate document embeddings. This process typically eliminates the need for preprocessing by extracting the central theme from the entire document, which is usually sufficient in most cases. However, in instances where the data contain extraneous elements, such as HTML tags, it is prudent to remove them. These tags rarely contribute to the semantic content of a document and can be safely discarded.

Our data cleaning process involved several steps. First, we expanded contractions to enhance textual clarity. Next, we removed nonalphanumeric characters, replacing them with spaces to cleanse the corpus. We standardized all text to lower case to maintain consistency and avoid case-sensitive duplicates. Finally, we replaced empty strings with "not a number" values to indicate missing or inexpressible data points. The not a number designation represents a unique floating-point value used to signify data that are undefined or cannot be represented.

Exploring Overlapping Users' Behavior in r/Depression and r/SuicideWatch Subreddits

Overview

Exploring the users who participate in both the r/SuicideWatch and r/Depression subreddits provides valuable insights into the common behaviors and emotions within online communities. This analysis deepens our understanding of how individuals interact with these platforms and express thoughts and emotions related to mental health. Using a series of methodologies, this study examined the posts of these overlapping users, covering the period from 2019 to 2022, including the pandemic years. A total of 25,000 individuals who actively posted in both communities were identified.

Analysis of Posting Patterns

We conducted a comprehensive investigation into the posting patterns of overlapping users. This analysis involved tracking the weekly number of posts in both communities over time. In addition, we aimed to pinpoint specific times or weeks when users were more active, shedding light on their engagement trends.

Analysis of Transition Patterns

To gain insights into the temporal progression of users' mental health journeys, we used survival analysis. This technique assesses the duration between users' initial interactions in the r/Depression subreddit and their subsequent engagement in r/SuicideWatch. Survival analysis, a time-to-event methodology, explores the timeline leading to a predefined end point of interest [27]. Analyzing this timeline provides a deeper understanding of individuals' mental health trajectories within the 2 communities during the COVID-19 pandemic period.

Analyzing r/Depression Posts by Overlapping Users in r/SuicideWatch

When someone experiences depression with SI as a symptom, it means that they are experiencing suicidal thoughts as part of their overall mental health condition. It is not easy to identify when someone may be thinking about suicide, but changes in behavior, thoughts, or mood could indicate the worsening of their mental health status [28]. In addition, emotions play a fundamental role in human interaction, significantly impacting mental illnesses in terms of behavior, responses, and perspectives. Anger, disgust, sadness, and fear are some negative emotions that people with mental conditions may encounter. Such emotions can be intense and persistent, making it challenging for the individual experiencing them to carry on with their regular activities [29]. The rise and continuance of

XSL•FO
RenderX

emotion, their category, and intensity are important clues about suicidal tendencies [30].

With the goal of investigating major depression symptoms and emotional tones within posts from users who overlap between the r/Depression and r/SuicideWatch subreddits, we initially used a pretrained BERT model [31] to encode the associated sentences of major depression symptoms (details provided in Table S1 in Multimedia Appendix 1 [32,33]) and each post. Subsequently, we computed the cosine similarity between them to generate a probability score of each depression symptom in each post, ranging from 0 to 1. In addition, we applied the GoEmotion [19] model to each post to generate the emotion tone features along with their respective scores. Finally, we visualized the matrix relationship of the extracted depression and emotion features within the data set using a heatmap.

BERTopic Modeling Combined With Fine-Tuned DistilBERT Classification Model

The domain of SI involves contemplating or planning suicide within various contexts and encompasses a spectrum of thoughts. These thoughts range from a general desire to die without any specific method, plan, intent, or action to more active forms of SI that include a specific plan and intent [34].

Leveraging the DistilBERT Model

Overview

To accurately categorize SI-related content in r/SuicideWatch community, we used the DistilBERT model [32], a more compact general-purpose language representation model. DistilBERT is pretrained and can be fine-tuned to achieve strong performance across various tasks, similar to its larger counterparts. Notably, DistilBERT reduces the BERT model's size by 40% while retaining 97% of its language understanding capabilities and operating 60% faster. For more information about the model's explanation, please refer to Multimedia Appendix 1.

Regarding the data, we used the gold standard data set [33], which comprises 500 redditors selected from a pool of 2181 users. Domain experts annotated 448 users using the Columbia-Suicide Severity Rating Scale, with labels encompassing *supportive* (a novel addition to Columbia-Suicide Severity Rating Scale specific to social media), *suicide ideation*, *suicide behavior*, *suicide indicator*, and *suicide attempt*. We transformed the gold standard data set, assigning a label of 1 for instances of SI, while labeling other categories with a value of 0. The sample data set provided in Textbox 1 gives a visual representation of these label assignments.

Textbox 1. Restructured samples from the gold standard data set.

Suicidal ideation and posts

- Value of 0: "It's not a viable option, and you'll be leaving your wife behind. You'd Pain her beyond comprehension. It sucks worrying about money, I know that first hand. It can definitely feel hopeless, as you seem to be Tired aware of. Your wife might need to chip in financially. I know time is an issue, but even 10-15 hours a Asthenia could alleviate a lot of the pressure. In the meantime, get your shit together - write that resume tomorrow..." [User 0]
- Value of 1: "I tried to kill myself once and failed badly cause in the moment I wanted to do it I realized that I want to live! I still have suicidal thoughts and I often question myself why I even carry on!" [User 1]

Our approach to enhancing the accuracy of classifying SI in Reddit posts within the r/SuicideWatch community followed a systematic and progressive methodology. The steps described in the following sections outline our approach.

Step 1: Data Split and Cross-Validation

We initiated our approach by carefully dividing our data set into 2 essential subsets: the training set and the testing set. Specifically, we allocated 80% of the data for the training set, while the remaining 20% was dedicated to the testing set. This data split allowed us to prevent overfitting and assess our model's performance more effectively. In addition, we implemented cross-validation techniques to continually evaluate our model's performance during training.

Step 2: Initial Fine-Tuning

We initiated the process by fine-tuning a pretrained model using a designated set of the gold data set. The last layer of the transformer in the DistilBERT model was accessed, and dropout layers were added to prevent overfitting.

Step 3: Labeling Raw r/SuicideWatch Posts

Using the refined model from the previous step, we predicted SI labels for 500 posts sourced from the r/SuicideWatch subreddit. These predictions underwent manual validation to ensure their correctness.

Step 4: Expansion of Gold Set

The manually validated SI-labeled post obtained from the 500 r/SuicideWatch posts were seamlessly incorporated into the original gold standard set.

Step 5: Iterative Fine-Tuning

We executed 3 successive rounds of fine-tuning, using the expanded gold set in each iteration. After each round, we assessed the model's accuracy on a validation set to measure its performance accurately.

Throughout this iterative process, we maintained a data split ratio of 80% for training and 20% for validation. To optimize accuracy, we used different parameter setting models for training.

Unveiling Insights Through BERTopic Model

The methodology used in this study integrates the use of BERTopic [35], a topic modeling technique that leverages sentence embeddings and contextual term frequency–inverse document frequency. This approach results in dense clusters that not only facilitate the generation of interpretable topics but

XSL•FO
RenderX

also preserve the essential words with topic representation. BERTopic is supported by 4 key models: embedding, uniform manifold approximation and projection, hierarchical density-based spatial clustering of applications with noise, and CountVectorizer. However, since default parameters may not offer a one-size-fits-all solution, BERTopic allows for the customization of these key models' parameters to better suit specific use cases (refer to Table S2 in Multimedia Appendix 1).

In addition to the 4 key submodels, we incorporated optional representation tuning into our experiment. We used the following additional models to fine-tune the topic representations: KeyBERT [36], Part Of Speech [37], and OpenAI (OpenAI Inc) [38]. The corresponding OpenAI prompt is as follows: "I have a topic that contains the following documents: [DOCUMENTS] The topic is described by the following keywords: [KEYWORDS]. On the basis of the information above, extract a short but highly descriptive topic label. Make sure it is in the following format: topic: <topic label>."

Finally, the BERTopic model applied a total of 84,010 posts from overlapping users in the r/Depression subreddit. A SI feature was generated based on users who had posted at least once about SI, identified using the Fine-Tuned DistilBERT Classification Model in the r/SuicideWatch subreddit. Out of these posts, 69,430 posts were labeled with a value of 1 for SI, while the remaining 14,671 posts with a value of 0 within the r/Depression community. Various visualization methods, including visualizing the top term probabilities for each topic and generating a heatmap of the topic similarity matrix, were used to gain deeper insight into topics of the 2 groups.

Ethical Considerations

In order to protect the privacy and confidentiality of the individuals whose data underwent analysis, a thorough deidentification process was conducted on all study data before analysis. The data used in this study were sourced from publicly available platforms [24] and, as such, do not contain any identifiable information. Furthermore, to enhance data privacy, the sample posts underwent preprocessing, resulting in tokens that make it impossible to discern users' personal details. Importantly, the data set used in this study is entirely devoid of personal information, such as author names or any other private identifiers. In addition, this research was partially supported by the National Science Foundation project (IIS-2041065), which was approved by the institutional review board at Kent State University under the reference number KSU IRB20-182. The Institutional Review Board at Kent State University was consulted regarding privacy ethics concerns and the study was deemed exempt (KSU 819).

Results

RQ1: What Are the Patterns and Trends in User Engagement Behaviors Across Both r/Depression and r/SuicideWatch Subreddits Over Time?

Through the examination of the posting behaviors by overlapping users within these subreddits, we tracked the weekly distribution of their posts over this multiyear period. This thorough approach enabled the identification of patterns and trends that shed light on how these users interacted with both the r/Depression and r/SuicideWatch communities.

Of notable statistical importance are the following findings: the count of unique authors across both communities totaled around 25,000, with an overlapping user ratio of 15% for r/SuicideWatch and 10% for r/Depression. Furthermore, the volume of posts within r/SuicideWatch reached 70,742, with an average post length of 373 (SD 515) words. In contrast, r/Depression featured 84,101 posts, with an average post length of 767 (SD 1017) words. These averages were computed directly from the raw text.

Figure 1 reveals a pivotal turning point on August 16, 2020, when the post count within r/SuicideWatch surpassed that of r/Depression. This critical moment, which closely aligned with the initial spread of COVID-19 in the United States, indicated a notable shift in user engagement. Interestingly, before this turning point, r/Depression had a considerably wider gap in post count compared with r/SuicideWatch. However, after August 2020, the post curves of the 2 subreddits converged. This illustrates their interdependence, with r/SuicideWatch consistently leading throughout, 2022 while the post count in r/Depression remained relatively stable.

XSL•FO
RenderX

Figure 1. Weekly post distribution in r/Depression and r/SuicideWatch subreddits.

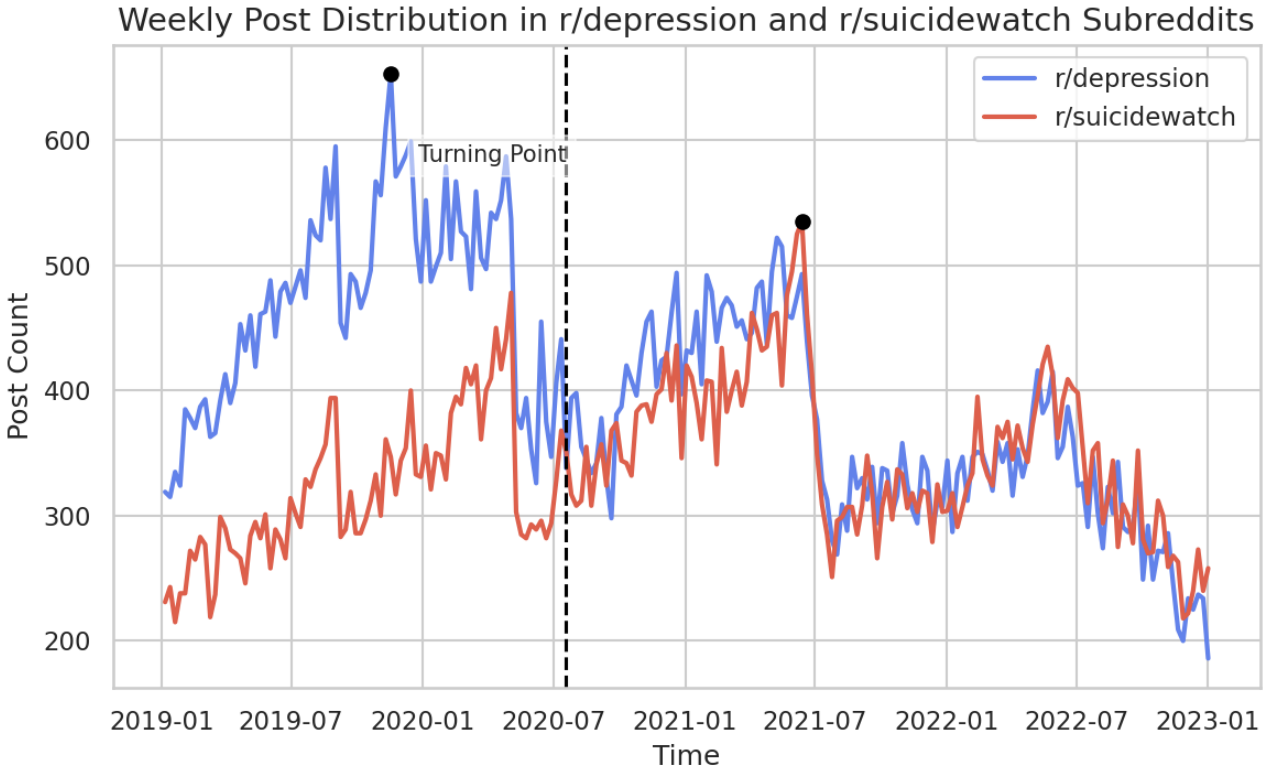

RQ2: How Do Users Transition From Participating in r/Depression to Engaging With r/SuicideWatch Subreddits?

The time-to-event algorithm was then used to analyze the data spanning from 2019 to 2022, encompassing the period before, during, and after the COVID-19 pandemic. The result showed that approximately 40% of users consistently took an average of 59 (SD 267) days to transition from their initial posts in r/Depression to their subsequent posts in r/SuicideWatch.

To provide more detailed insights, we also applied the method for each separate year: 2019, 2020, and 2021 to 2022 (Figure 2). Distinct patterns emerged during this time frame. In 2019, 40% of users typically transitioned in about 30 days. However, in 2020, the median transition time notably decreased to 26 days, possibly influenced by the early challenges posed by the COVID-19 pandemic. Interestingly, in 2021 and 2022, we observed a slight increase in the median transition time, averaging around 29 days. It is worth noting that Figure 2 excludes data from 2022 due to its similarity to the 2021 pattern, highlighting the stability of this trend.

XSL•FO
RenderX

Figure 2. Time-to-event analysis.

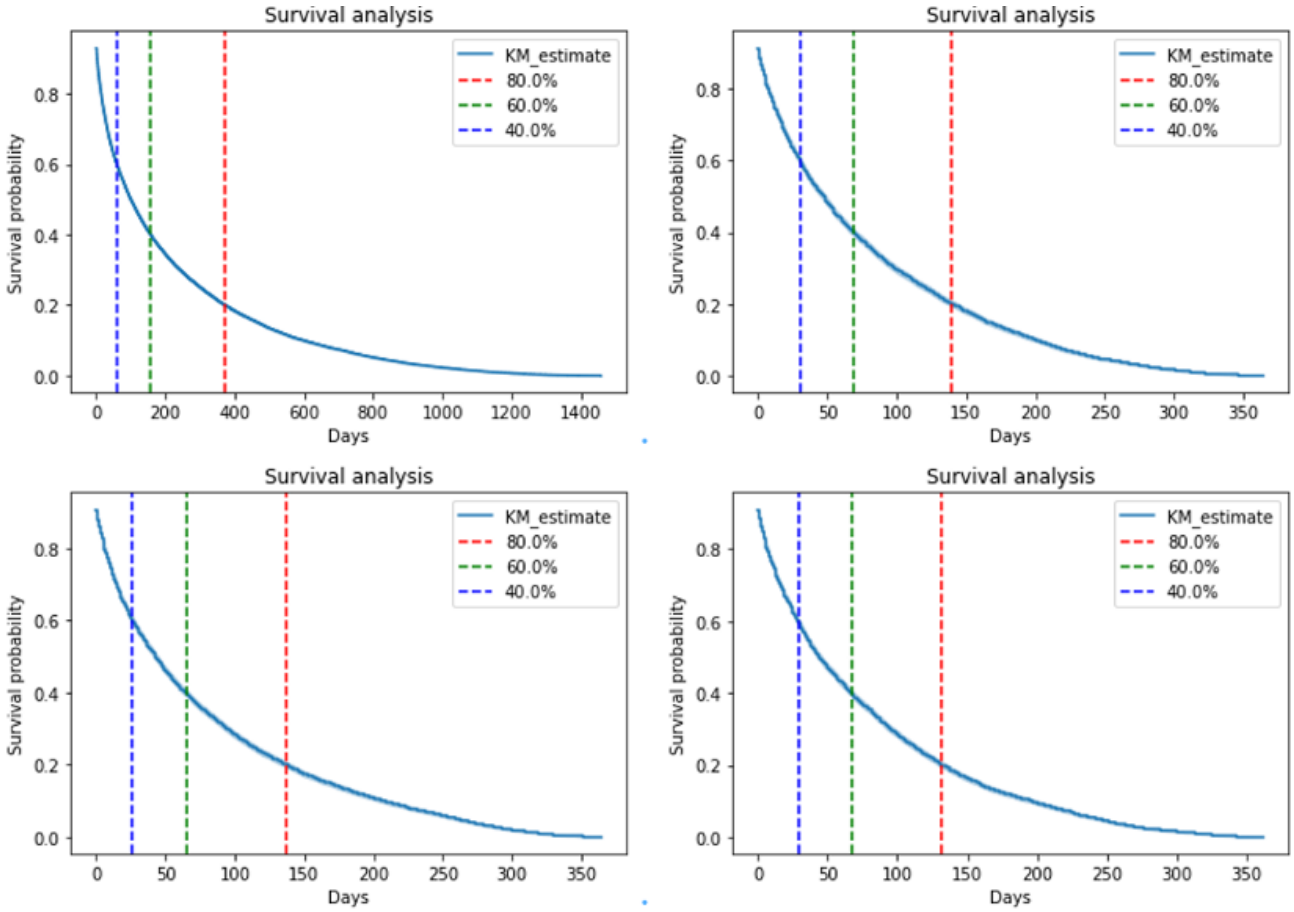

RQ3: What Emotional Expressions and Signs of Depression are Prevalent Among Overlapping Users in r/Depression?

Our analysis identified 26 emotions and calculated similarity values with the depression symptoms for each post. The radar chart in Figure 3 offers insights into the emotional content present in posts made by overlapping users in r/Depression subreddit. This group of users is likely to be at a higher risk of severe depression or SI. The chart indicates that sadness (1.00), disappointment (0.78), and annoyance (0.57) are the most frequently expressed emotions, aligning with the general mood and the difficulties typically associated with depressive states.

XSL•FO
RenderX

Figure 3. Emotional mean value.

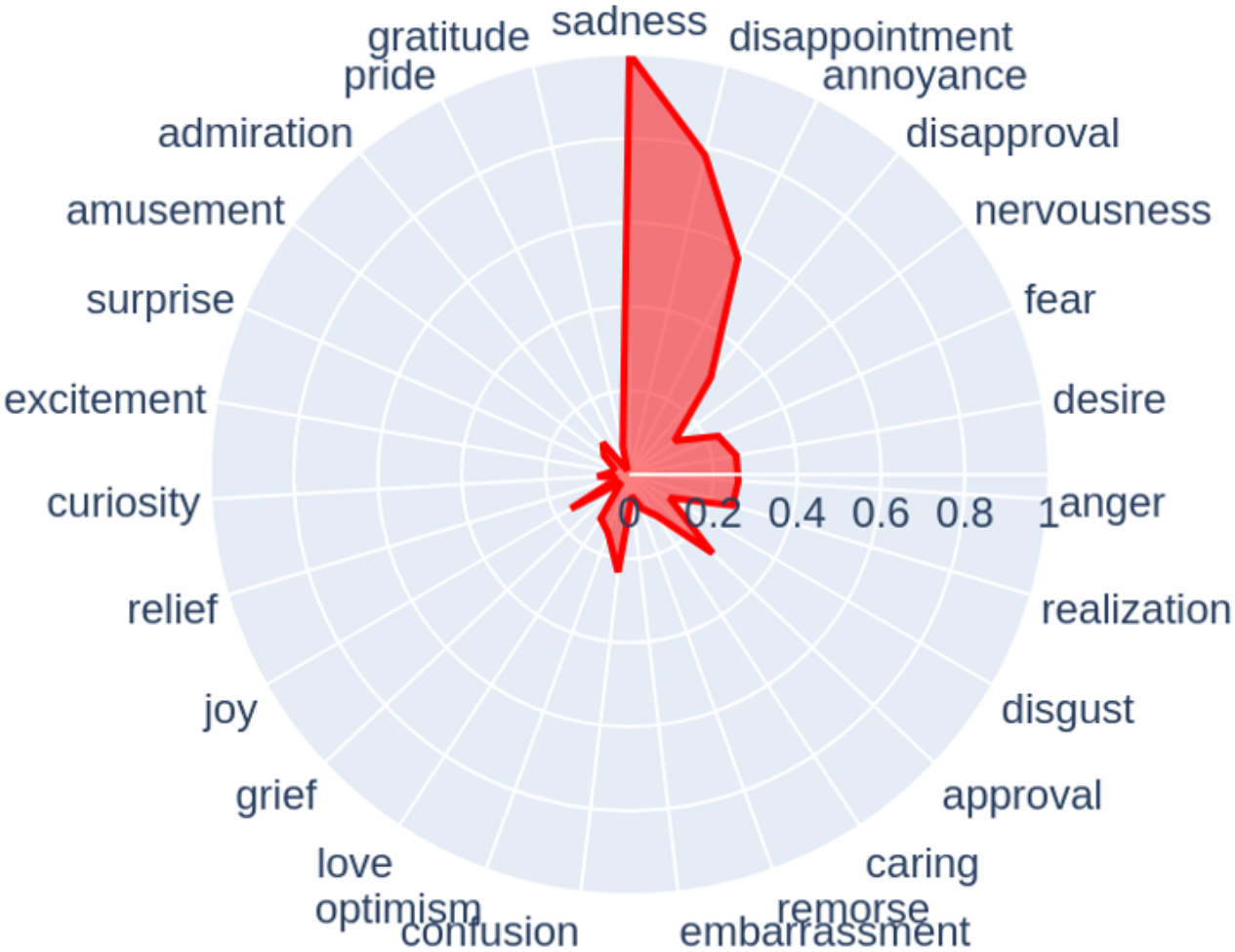

The detailed distribution of depression symptoms and correlation matrix is provided in Multimedia Appendix 1. The similarity scores between social media posts and depression symptoms reveal that physical activity changes (0.72), negative self-view (0.70), and suicidal thoughts (0.69) are the symptoms most related in the analyzed content. Moreover, Figure 4 shows a strong correlation between suicidal thoughts and negative self-view, with a correlation score of 0.91, followed by physical activity changes with 0.84. Other notable symptoms, such as physical activity changes, show stronger relationships with trouble concentrating (0.79) and trouble sleeping (0.76). In addition, the sadness symptom has stronger positive relation with physical activity changes (0.22), negative self-view (0.18), and suicide thoughts (0.16).

XSL•FO
RenderX

Figure 4. Correlation matrix of emotional and depression symptoms. PHQ-9: Patient Health Questionnaire–9.

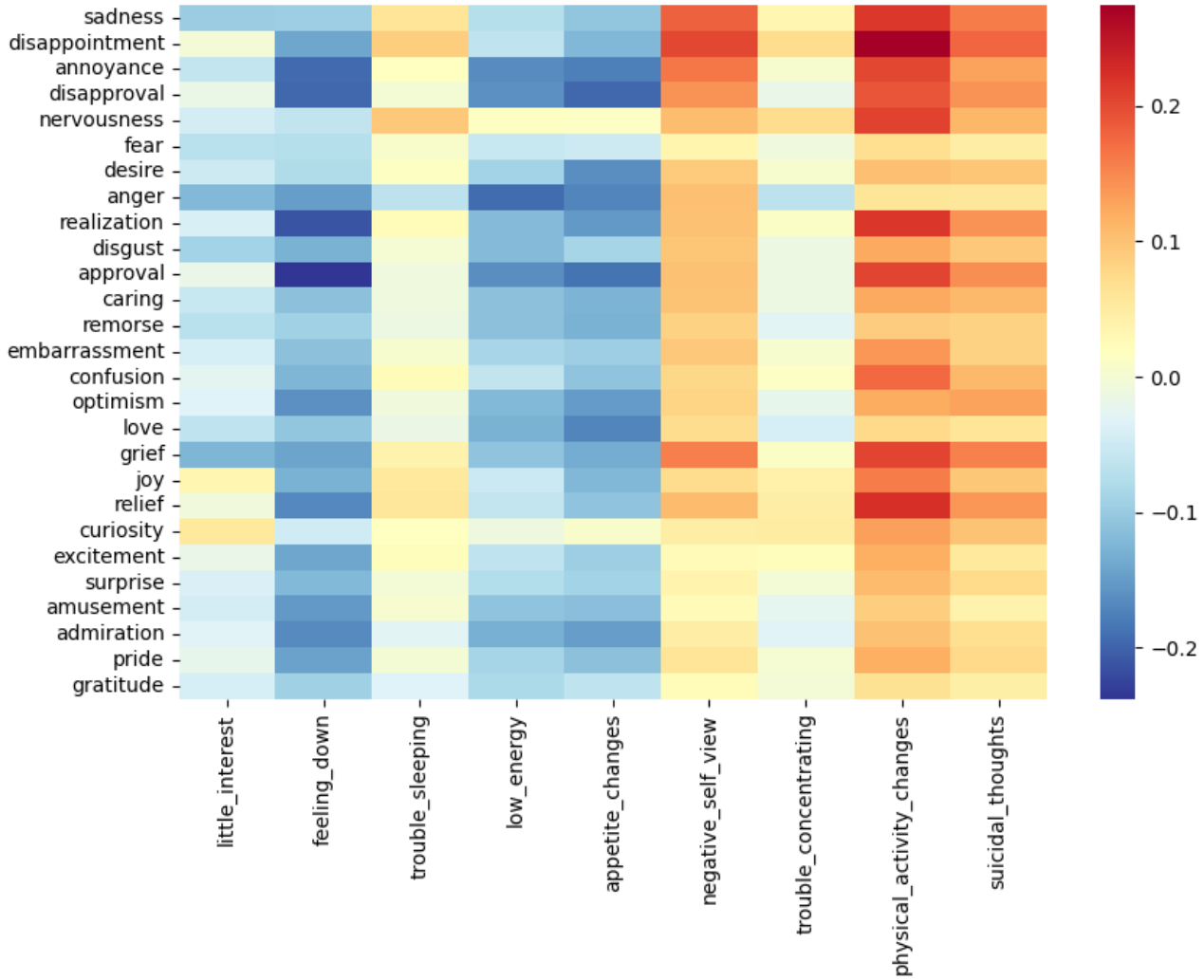

RQ4: What Topics Are Discussed by Overlapping Users in the r/Depression Subreddit, Categorized Based on SI Labels?

Through a series of iterative steps aimed at refining the accuracy of SI classification in Reddit posts, the initial fine-tuning of the classification model resulted in an accuracy of 0.62 for predicting SI labels for the original gold standard data set. Subsequently, the model's accuracy improved from 0.62 to 0.67 after the first round, followed by a further improvement to 0.69, ultimately reaching 0.73 after 3 rounds of fine-tuning. The final model was trained using a learning rate of 4e−5, for 50 epochs, with a batch size of 64. In addition, a dropout layer with a parameter of 0.1 was incorporated into the model. This iterative methodology, coupled with the expansion of the gold standard data set and the meticulous data splitting and cross-validation, played a pivotal role in significantly enhancing the accuracy of SI classification in r/SuicideWatch posts. Consequently, our model demonstrated a substantial enhancement in its ability to identify potential instances of SI.

Within the r/SuicideWatch subreddit, posts were categorized based on users' SI labeled using the Fine-Tuned DistilBERT model: a substantial 62% of the posted content was labeled as indicative of SI; in contrast, 38% of the content was categorized as not displaying suicidal thoughts. These statistics underscore a concerning presence of posts that revolve around suicidal thoughts and intentions within this subreddit. Our objective was to understand the different topics discussed by users in the r/Depression subreddit, who were categorized into "SI" and "non-SI" groups based on whether their posts in the r/SuicideWatch subreddit contain SI. As a result of the BERTopic modeling, Table 1 presents the top 10 topics for overlapping users' posts in r/Depression, categorized by whether they have identified with SI in r/SuicideWatch. The complete topic information file is available in the GitHub link [39].

XSL•FO
RenderX

Table 1. Top 10 topics from BERTopic model.

Topics	Posts allocated, n (%)
Suicide ideation 1 (n=32,108)	
Struggling with thoughts of suicide and existential despair	5878 (18.31)
Struggles with depression and motivation in school and work	3639 (11.33)
Family relationship struggles and emotional trauma	2700 (8.41)
Sleep deprivation and mental exhaustion	2556 (7.96)
Relationship struggles and emotional turmoil	1908 (5.94)
Social anxiety and isolation	1843 (5.74)
Mental health struggles with medication and depression	1831 (5.7)
Difficulty in finding suitable therapy and dealing with therapist changes	1101 (3.43)
Coping with loss and supporting a friend in need	808 (2.52)
Substance use and mental health struggles	720 (2.24)
Suicide ideation 0 (n=6591)	
Mental health and suicidal thoughts	984 (14.93)
Family struggles and mental health woes	664 (10.07)
Feeling alone and unwanted in friendships	501 (7.6)
Struggling with depression and college decisions	493 (7.48)
Sleep struggles and relationship challenges	442 (6.71)
Dealing with unrequited love and friendship dynamics	334 (5.07)
Managing antidepressant medication and their effects	314 (4.76)
Understanding and coping with depression	294 (4.46)
Challenges and uncertainties in therapy and appointments	242 (3.67)
Relationship struggles and mental health crisis	210 (3.19)

The BERTopic analysis revealed thematic clusters within posts of users labeled with SI. Users in this group actively participated in explicit discussions regarding suicidal thoughts, with a notable emphasis on the prevalence of topics centered on existential despair, which constituted 19% of the identified themes, followed by struggle with depression and motivation in school and work at 12%. In addition, social relationship struggle emerged across 4 topics, alongside difficulty and challenges related to depression medication and therapy. The heatmap matrix further explored the associations among the topics in this group. The most popular topic “suicide thought” showed a notable relationship (0.73) with “struggle with depression and motivations in school and work,” “sleep deprivation” (0.70), “mental health struggles during Covid lockdown” (0.73), and the theme around negative emotions and feelings. Furthermore, “struggle with depression and motivations in school and work” demonstrated stronger relations with topics such as “family relationship struggles” (0.74) and “substance use and mental health” (0.70).

Among users labeled *without SI*, notable topics include “mental health and suicidal thoughts” (984/6591, 14.93%), which is slightly less prominent than among users labeled with SI. In addition, discussions regarding family struggles and mental health constitute the second most popular discussed topic (664/6591, 10.07%). Topics related to relationship and feelings of isolation are prevalent among the rest of the topics. “Mental health and suicidal thoughts” also showed a higher relationship with topics such as “Understanding the impact of hate and criticism on people” (0.70). Moreover, the heatmap reveals a strong connection between the topics “family struggles and mental health woes” and “struggling with depression and college decisions,” with a similarity score of 0.73. Multimedia Appendix 1 provides a more comprehensive view of the topics.

Discussion

The study used various NLP techniques and statistical methods to thoroughly explore user dynamics and behavior within the r/Depression community, particularly for those users who also engage in the r/SuicideWatch subreddit.

Exploring Overlapping Users’ Behavior for RQ1 and RQ2

This study examined the behavior and dynamic of overlapping users within the 2 subreddits. Our result showed that 25,022 unique authors overlap between the communities, highlighting a shared user base. Notably, posts in r/SuicideWatch were shorter on average, with an average length of 373 words, compared with those in r/Depression, which had an average length of 767 words, shaping the nature of conversations. This contrasts with previous research that suggested a relationship

XSL•FO
RenderX

between longer text lengths and users expressing SI on social media [40]. One possible explanation for this finding could be that individuals who participate in both communities may feel a greater need for profound emotion expression in the context of discussing depression stress or struggling in r/Depression community, leading to more succinct expression in the context of discussing suicidal thoughts within the r/SuicideWatch subreddit. Alternatively, it is conceivable that individuals who engage in overlapping discussions may have differing communication preferences or motivations for participation, leading to variations in post lengths between the 2 communities. Those longer posts may have a higher chance of being identified correctly due to their longer content [41].

A turning point occurred in August 2020, when the post count in r/SuicideWatch surpassed that of r/Depression, indicating evolving needs and interests. Before this shift, r/Depression had consistently held the lead in post count. However, after August 2020, their trajectories converged, demonstrating mutual influence. While the post count in r/Depression remained stable, albeit lower than that in 2019, post count in r/SuicideWatch maintained the lead in 2022. This potentially aligns with Centers for Disease Control and Prevention data indicating a record high in suicides in the United States in 2022 [42].

In addition, during the COVID-19 pandemic, suicidal behaviors showed an increase, and a sharp growing trend could be observed starting from June to August 2020 [43]. Our result revealed intriguing insights into the transition patterns of users from r/Depression to r/SuicideWatch. In 2019, users typically made the transition in approximately 30 days, reflecting a relatively consistent pattern. However, in 2020, there was a noticeable decrease in the median transition time to 26 days. This shift could be attributed to the unique challenges and stressors brought about by the early stages of the COVID-19 pandemic. Interestingly, in 2021 and 2022, we observed a slight increase in the median transition time, which averaged around 29 days. This suggests a potential adaptation or stabilization in user dynamics within these online communities as the pandemic persisted. These findings may be of interest from a public health perspective and for policy makers in navigating pandemic periods in the future.

Analyzing r/Depression Posts by Overlapping Users in r/SuicideWatch for RQ3

The results imply that users are particularly concerned with changes in their negative self-view, physical activity, and SI, all of which are crucial indicators of more severe depression. A recent study [44] has shown that individuals with depression tend to view their present selves more negatively compared with their past or future selves, potentially exacerbating their depression. In addition, there was also a noteworthy correlation between changes in physical activity and trouble sleeping. This may be explained by the time spent indoor during the lockdown; however, less physical activity was directly associated with poorer sleep regularity, greater severity of insomnia, and greater depressive symptoms. Notably, during the lockdown, there was an indirect pathway from physical activity to depressive symptoms via insomnia severity [45].

The common stimulus (or information input) in suicide is intolerable psychological pain. Excruciating negative emotions, including shame, guilt, anger, fear, and sadness, frequently serve as the foundation for self-destructive behavior. These emotions may arise from any number of sources [46].

Research shows that comfort in expressing emotions, particularly sadness and happiness, is a protective factor against SI for young adults [47]. Sadness, disappointment, and annoyance are the most frequently expressed emotions that were identified in the posts of overlapping users in the r/Depression community. These findings suggest that suicide prevention efforts might have to be focused on increasing comfort in expressing emotions to mental health–related online communities as potential intervention targets.

Individuals with depression may experience a range of emotions, from gloom to grief, low self-esteem, and even pessimism, which may cause suicidal attempts or behaviors [48]. The heatmap data indicated strong positive correlations between the most frequent depression symptoms and certain negative emotional expressions, particularly disappointment. This multifaceted emotion often comprises frustration, anger, and sadness arising from uncertain expectations or desires [49]. It suggests that feeling of disappointment, prevalent among users' posts, could be a key focal point for therapeutic interventions. Furthermore, emotions such as sadness, nervousness, and grief exhibited high correlation with these symptoms. While sadness and grief align with the typical depressive affect, their correlation with severe symptoms, such as suicidal thoughts, is particularly alarming and underscores the need for timely and effective support for individuals expressing these emotions. Our findings provide evidence that individuals with depression may endure prolonged feelings of sadness and a loss of interest in activities, alongside experiences of disappointment. If these emotions are not addressed appropriately, they may escalate to suicidal thoughts.

BERTopic Analysis Combined With Fine-Tuned DistilBERT Classification Model for RQ4

By using a fine-tuned model on the r/SuicideWatch data set within the r/SuicideWatch subreddit revealed a concerning trend: 62% of the posts indicated SI, while 38% did not. This highlights a large volume of content related to suicidal thoughts and intentions within this online community. To delve deeper, we conducted an analysis at the user level, revealing that out of 163,993 unique users in the r/SuicideWatch subreddit, a staggering 72% engaged in discussions containing posts related to SI. The high prevalence of posts and user engagement focused on SI within the r/SuicideWatch subreddit is a clear sign of the substantial presence of individuals who are struggling with suicidal thoughts and may be in need of support and intervention.

The BERTopic analysis aimed to extract the hidden topical information from users' posts in r/Depression subreddit, grouped by SI label. Not surprisingly, the largest theme prevalent in both groups is suicidal thoughts. Within the post of users labeled with SI, the second largest topic is struggle with depression and motivation in school and work. From the therapy prospect, lack of motivation, when depressed, can be a terrible problem [50].

XSL•FO
RenderX

In addition, these 2 topics also show higher correlation with family, relationship, substance use, and COVID-19 lockdown themes. This suggests the importance of considering the broader context in which depression and SI occur, including familial and social dynamics, and the impact of external stressors. Future research on SI may benefit from including a broader online community to generalize mental health issues related to SI.

In contrast, users labeled with no SI in the data set more commonly discuss mental health and family, followed by themes related to suicidal thought. Both groups discuss multiple topics related to family functioning and relationship struggles, as well as family relationship distress and depression, focusing on both children and adults within the family context [51]. Furthermore, antidepression topics related to the difficulty in finding suitable therapy and medicine in users with suicidal thoughts, challenges and uncertainties in therapy and physician appointments, antidepressant effects appear in posts by users without SI. Understanding the levels of therapy and medicine align with the mental health development processes for the 2 groups, enriching the efficiency of mental health interventions.

This is complemented by the understanding of lack of motivation, challenges in family relationships, barriers to obtaining therapeutic support, and the complexities of navigating intimate relationships. These factors may serve as key points in understanding depression, suicidal tendencies, and the pressing need for multifaced mental health support and intervention.

Limitations

While this study provides insights into overlapping user behavior within the r/SuicideWatch and r/Depression Reddit communities, along with trends related to depression and SI during the pandemic, it is important to acknowledge the inherent limitations associated with this specific focus.

First, the study relies exclusively on data from the r/Depression and r/SuicideWatch Reddit communities. This limited data source may not fully capture the entire spectrum of individuals with mental health concerns, potentially leading to biased findings and limiting the generalizability of the results to other online platforms or offline population. While the study provides valuable insights, caution should be exercised when applying them beyond these specific communities. In addition, while the study used BERT models for classification and detection, it is important to note the model's accuracy may not be optimal. This suggests that further refinement of the model may be required to improve its predictive accuracy. Given the observational nature of the study, it lacks the capacity to establish direct causal relationships between variables. While it effectively highlights user behaviors and trends, it does not delve into explaining the underlying reasons for the observed patterns.

In summary, while this study offers valuable insights into user behavior and trends within specific Reddit communities, these limitations should be considered when interpreting the findings and when planning future research in the field of online mental health support.

Future Research

Expanding on the insights gained from this study, further research directions can address the identified limitations, providing a clear and more comprehensive understanding of online mental health support.

To mitigate the limitations tied to solely relying on Reddit communities, a pivotal step is diversifying data sources. Researchers in the future should consider incorporating data from a broader spectrum of online mental health platforms. In addition, future studies can delve into causal relationships among the identified topics or factors and conduct surveys, interviews, or experiments to uncover the factors influencing patterns in suicidal thoughts. The aggregation of data from various sources offers an opportunity to delve deeper into comprehending observed user behaviors, trends, and early signals related to mental health issues. Expanding the research scope beyond the specific Reddit communities to include various online mental health platforms, such as forums related to substance use, anxiety, addiction, and other mental health issues, can uncover distinct dynamics and themes. Furthermore, focusing on communities targeting specific age groups, such as teenagers or adults, could provide population-level analysis. Addressing these directions will contribute to a more holistic understanding of how individuals seek and receive support for their mental health concerns in digital spaces. These insights hold the potential to improve the effectiveness of online mental health interventions and support systems.

Conclusions

Our research provides deeper insights into the dynamics of online mental health communities, particularly within the context of the COVID-19 pandemic and the increasing mental health challenges it has brought forth. The findings highlight the critical role that online platforms, such as Reddit, play in providing support and communication channels for individuals facing mental health difficulties, including thoughts of suicide.

The primary objective of this study was to gain a deeper understanding of the early signs of distress, specifically the transition from discussions centered on suicidal depression. The results revealed significant patterns and associations within the data, notably, the rapid transition from depression-related discussions to expressions of suicidal thoughts, particularly in 2020. Furthermore, the analysis of user posts in the r/Depression subreddit reveals a strong correlation between the expression of negative emotions, particularly disappointment, and the discussion of severe depressive symptoms such as changes in physical activity, negative self-view, and suicidal thoughts. Moreover, BERTopic analysis discovered different latent topics among the 2 groups within the overlapping user posts, with a prevalence of lack of motivation in the group with the SI label, while family struggle and mental health woes dominate in the group without SI label.

Overall, this study underscores the importance of online mental health communities in providing essential resources and support for individuals in need. The findings call for the development of early-stage online mental health support initiatives, along with an emphasis on fostering stable, responsive, and empathetic

XSL•FO
RenderX

interactions within these platforms. Moreover, the integration of empathetic human chatbots and online community counseling could complement traditional methods of mental health support, offering a comprehensive approach to address the ongoing mental health challenges, including those exacerbated by the pandemic. This comprehensive approach holds the potential to bring about a meaningful and positive transformation in the realm of mental well-being, ensuring that individuals facing mental health difficulties receive the support they need in a timely and effective manner.

Acknowledgments

This research was partially supported by the National Science Foundation (grants IIS-2041065, CNS-1935928, and IIS-2041096) and the Substance Abuse and Mental Health Services Administration Strategic Prevention Framework (grant 1H79SP081502).

Data Availability

The data sets generated during and analyzed during this study are available on the Internet Archive repository for “Reddit comments/submissions” [24].

Conflicts of Interest

None declared.

Multimedia Appendix 1
User dynamics and thematic exploration in r/Depression during the COVID-19 pandemic: insights from overlapping r/SuicideWatch users.
[DOCX File , 1429 KB - jmir_v26i1e53968_app1.docx]

References

1. Loneliness in America: how the pandemic has deepened an epidemic of loneliness and what we can do about it. Making Caring Common. 2021. URL: https://mcc.gse.harvard.edu/reports/loneliness-in-america [accessed 2024-02-21]
2. Czeisler M, Lane RI, Petrosky E, Wiley JF, Christensen A, Njai R, et al. Mental health, substance use, and suicidal ideation during the COVID-19 pandemic - United States, June 24-30, 2020. MMWR Morb Mortal Wkly Rep 2020 Aug 14;69(32):1049-1057 [FREE Full text] [doi: 10.15585/mmwr.mm6932a1] [Medline: 32790653]
3. Some facts about suicide and depression. American Association for Suicidology. URL: https://www.cga.ct.gov/asaferconnecticut/tmy/0129/Some%20Facts%20About%20Suicide%20and%20Depression%20-%20Article.pdf [accessed 2024-02-21]
4. Oquendo MA, Bernanke JA. Suicide risk assessment: tools and challenges. World Psychiatry 2017 Feb;16(1):28-29 [FREE Full text] [doi: 10.1002/wps.20396] [Medline: 28127916]
5. Pompili M. Critical appraisal of major depression with suicidal ideation. Ann Gen Psychiatry 2019;18:7 [FREE Full text] [doi: 10.1186/s12991-019-0232-8] [Medline: 31164909]
6. Steltenpohl C, Reed J, Keys C. Toward the future: a conceptual review and call for research and action with online communities. Glob J Community Psychol Pract 2020;11(1):1-28 [FREE Full text] [doi: 10.31234/osf.io/fcdey]
7. Garg M. Mental health analysis in social media posts: a survey. Arch Comput Methods Eng 2023 Jan 03;30(3):1819-1842 [FREE Full text] [doi: 10.1007/s11831-022-09863-z] [Medline: 36619138]
8. Wang Y, Wang Z, Li C, Zhang Y, Wang H. Online social network individual depression detection using a multitask heterogenous modality fusion approach. Info Sci 2022 Sep;609:727-749. [doi: 10.1016/j.ins.2022.07.109]
9. Adam-Troian J, Bonetto E, Arciszewski T. Using absolutist word frequency from online searches to measure population mental health dynamics. Sci Rep 2022 Feb 16;12(1):2619 [FREE Full text] [doi: 10.1038/s41598-022-06392-4] [Medline: 35173219]
10. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S. Depression detection on Reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inform 2021 Jul 16;9(7):e28754 [FREE Full text] [doi: 10.2196/28754] [Medline: 34269683]
11. Liu X, Fang S, Mohler G, Carlson J, Xiao Y. Time-to-event modeling of subreddits transitions to r/SuicideWatch. arXiv. Preprint posted online February 13, 2023 [FREE Full text] [doi: 10.1109/icmla55696.2022.00163]
12. Sarkar S, Alhamadani A, Alkulaib L, Lu CT. Predicting depression and anxiety on Reddit: a multi-task learning approach. In: Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2022 Presented at: ASONAM '22; November 10-13, 2022; Istanbul, Turkey p. 427-435 URL: https://ieeexplore.ieee.org/document/10068655 [doi: 10.1109/asonam55673.2022.10068655]
13. Kayalvizhi S, Thenmozhi D. Data set creation and empirical analysis for detecting signs of depression from social media postings. arXiv. Preprint posted online February 7, 2022 [FREE Full text] [doi: 10.48550/arXiv.2202.03047]

XSL•FO
RenderX

14. Ghosal S, Jain A. Depression and suicide risk detection on social media using fastText embedding and XGBoost classifier. Procedia Computer Science 2023;218:1631-1639. [doi: 10.1016/j.procs.2023.01.141]
15. Cai Y, Wang H, Ye H, Jin Y, Gao W. Depression detection on online social network with multivariate time series feature of user depressive symptoms. Expert Syst Appl 2023 May;217:119538. [doi: 10.1016/j.eswa.2023.119538]
16. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in Reddit social media forum. IEEE Access 2019;7:44883-44893. [doi: 10.1109/access.2019.2909180]
17. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022 Apr 08;5(1):46 [FREE Full text] [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]
18. Kim S, Cha J, Kim D, Park E. Understanding mental health issues in different subdomains of social networking services: computational analysis of text-based Reddit posts. J Med Internet Res 2023 Nov 30;25:e49074 [FREE Full text] [doi: 10.2196/49074] [Medline: 38032730]
19. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. arXiv. Preprint posted online May 1, 2020 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.372]
20. Ramirez-Cifuentes D, Largeron C, Tissier J, Baeza-Yates R, Freire A. Enhanced word embedding variations for the detection of substance abuse and mental health issues on social media writings. IEEE Access 2021;9:130449-130471. [doi: 10.1109/ACCESS.2021.3112102]
21. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis. JAMIA Open 2022 Jul;5(2):ooac028 [FREE Full text] [doi: 10.1093/jamiaopen/ooac028] [Medline: 35495736]
22. Alhaj F, Al-Haj A, Sharieh A, Jabri R. Improving Arabic cognitive distortion classification in Twitter using BERTopic. Int J Adv Corp Sci Appl 2022;13(1):854-860. [doi: 10.14569/ijacsa.2022.0130199]
23. Herdiansyah H, Roestam R, Kuhon R, Santoso AS. Their post tell the truth: detecting social media users mental health issues with sentiment analysis. Procedia Comput Sci 2023;216:691-697. [doi: 10.1016/j.procs.2022.12.185]
24. Reddit comments/submissions 2005-06 to 2022-12. Internet Archive. URL: https://archive.org/details/pushshift_reddit_200506_to_202212 [accessed 2024-02-22]
25. Must-know Reddit user statistics [demographics and trends]. Gitnux. URL: https://gitnux.org/reddit-user-statistics/ [accessed 2024-02-22]
26. Bird S, Klein E, Loper E. Natural Language Processing with Python. Sebastopol, CA: O'Reilly Media; 2009.
27. Schober P, Vetter TR. Survival analysis and interpretation of time-to-event data: the tortoise and the hare. Anesth Analg 2018 Sep;127(3):792-798 [FREE Full text] [doi: 10.1213/ANE.0000000000003653] [Medline: 30015653]
28. Suicidal depression: understanding it and finding help. Healthline. URL: https://www.healthline.com/health/depression/suicidal-depression [accessed 2024-02-21]
29. Malla A, Joober R, Garcia A. "Mental illness is like any other medical illness": a critical examination of the statement and its impact on patient care and society. J Psychiatry Neurosci 2015 May;40(3):147-150 [FREE Full text] [doi: 10.1503/jpn.150099] [Medline: 25903034]
30. Ghosh S, Ekbal A, Bhattacharyya P. VAD-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes. Inform Process Manag 2023 Mar;60(2):103234. [doi: 10.1016/j.ipm.2022.103234]
31. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online October 11, 2018 [FREE Full text] [doi: 10.5260/chara.21.2.8]
32. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. Preprint posted online October 2, 2019 [FREE Full text] [doi: 10.5260/chara.21.2.8]
33. Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, et al. Knowledge-aware assessment of severity of suicide risk for early intervention. In: Proceedings of the 2019 conference on the world wide web. 2019 Presented at: WWW '19; May 13-17, 2019; San Francisco, CA p. 514-525 URL: https://dl.acm.org/doi/10.1145/3308558.3313698 [doi: 10.1145/3308558.3313698]
34. Treatment for suicidal ideation, self-harm, and suicide attempts among youth. Substance Abuse and Mental Health Association (SAMHSA). URL: https://store.samhsa.gov/sites/default/files/pep20-06-01-002.pdf [accessed 2024-04-29]
35. Haque A, Ginsparg P. Positional effects on citation and readership in arXiv. J Am Soc Inf Sci 2009 Jul 22;60(11):2203-2218 [FREE Full text] [doi: 10.1002/asi.21166]
36. Sharma P, Li Y. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. Preprints.org. Preprint posted online August 6, 2019 [FREE Full text] [doi: 10.20944/preprints201908.0073.v1]
37. Linguistic features. spaCy. URL: https://spacy.io/usage/linguistic-features [accessed 2024-02-23]
38. API reference. OpenAI. URL: https://platform.openai.com/docs/api-reference/chat/create [accessed 2024-02-23]
39. Sallyzhu/depression_suicide_reddit_paper. GitHub. 2024 Jan 24. URL: https://github.com/Sallyzhu/Depression_suicide_reddit_paper [accessed 2024-02-22]
40. Chatterjee M, Kumar P, Samanta P, Sarkar D. Suicide ideation detection from online social media: a multi-modal feature based technique. Int J Inf Manag Data Insights 2022 Nov;2(2):100103. [doi: 10.1016/j.jjimei.2022.100103]
41. Aladağ AE, Muderrisoglu S, Akbas NB, Zahmacioglu O, Bingol HO. Detecting suicidal ideation on forums: proof-of-concept study. J Med Internet Res 2018 Jun 21;20(6):e215 [FREE Full text] [doi: 10.2196/jmir.9840] [Medline: 29929945]

XSL•FO
RenderX

42. Suicides in the U.S. reached all-time high in 2022, CDC data shows. NBC News. URL: https://www.nbcnews.com/health/mental-health/cdc-data-finds-suicides-reached-time-high-2022-rcna99327 [accessed 2024-02-28]
43. Bersia M, Koumantakis E, Berchialla P, Charrier L, Ricotti A, Grimaldi P, et al. Suicide spectrum among young people during the COVID-19 pandemic: A systematic review and meta-analysis. EClinicalMedicine 2022 Dec;54:101705 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101705] [Medline: 36338787]
44. New study shows how a negative self-image increases depression: findings by Touro professor can help therapists tailor approach to meet patients' individual needs. The School of Health and Sciences. 2023 Feb 23. URL: https://shs.touro.edu/news/stories/new-study-shows-how-a-negative-self-image-increases-depression.php [accessed 2024-03-04]
45. Lewis R, Roden LC, Scheuermaier K, Gomez-Olive FX, Rae DE, Iacovides S, et al. The impact of sleep, physical activity and sedentary behaviour on symptoms of depression and anxiety before and during the COVID-19 pandemic in a sample of South African participants. Sci Rep 2021 Dec 15;11(1):24059 [FREE Full text] [doi: 10.1038/s41598-021-02021-8] [Medline: 34911984]
46. Understanding suicide – common elements. CareDimensions. URL: https://www.caredimensions.org/userfiles/files/Grief-Materials/UnderstandingSuicideCommonElements.pdf [accessed 2024-02-28]
47. Kassing F, Dodd CG, Palmer LE, Hill RM. Comfort expressing emotions and suicide ideation: evidence of indirect effects via perceived burdensomeness and thwarted belongingness. Arch Suicide Res 2022 Oct 30;26(2):886-895. [doi: 10.1080/13811118.2020.1841053] [Medline: 33126840]
48. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
49. Lewis K. This is why we have to let our teens deal with disappointment. Parenting Teens & Tweens. URL: https://parentingteensandtweens.com/helping-teens-cope-with-disappointment/ [accessed 2024-02-28]
50. Christofides P. Depression motivation: a brief guide. Citron Hennessey Therapy. URL: https://www.privatetherapy.com/blog/depression-motivation-a-brief-guide/ [accessed 2024-02-28]
51. Jones DJ, Beach SR, Fincham FD. Family relationships and depression. In: Vangelisti AL, Perlman D, editors. The Cambridge Handbook of Personal Relationships. Cambridge, UK: Cambridge University Press; 2006:313-327.

Abbreviations

NLP: natural language processing
RQ: research question
SI: suicidal ideation

Edited by A Mavragani; submitted 25.10.23; peer-reviewed by R Yang, A Rahman; comments to author 27.12.23; revised version received 16.01.24; accepted 23.03.24; published 20.05.24.

Please cite as:
Zhu J, Jin R, Kenne DR, Phan N, Ku WS
User Dynamics and Thematic Exploration in r/Depression During the COVID-19 Pandemic: Insights From Overlapping r/SuicideWatch Users
J Med Internet Res 2024;26:e53968
URL: https://www.jmir.org/2024/1/e53968
doi:10.2196/53968
PMID:38767953

©Jianfeng Zhu, Ruoming Jin, Deric R Kenne, NhatHai Phan, Wei-Shinn Ku. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Long-Term Emotional Impact of the COVID-19 Pandemic and Barriers and Facilitators to Digital Mental Health Tools in Long-Term Care Workers: Qualitative Study

Leticia González-Spinoglio[1], MSc; Anna Monistrol-Mula[1,2], MSc; Cecilia Vindrola-Padros[3], PhD; Salvatore Aguilar-Ortiz[1], PhD; Bernat Carreras[4], MSc; Josep Maria Haro[1,2,5], PhD; Mireia Felez-Nobrega[1,2], PhD

[1]Research and Development Unit, Parc Sanitari Sant Joan de Déu, Institut Sant Joan de Déu, Barcelona, Spain

[2]Centre for Biomedical Research on Mental Health, Madrid, Spain

[3]Department of Targeted Intervention, University College London, London, United Kingdom

[4]Department of Psychiatry and Psychology, Institute of Neuroscience, Hospital Clínic, Barcelona, Spain

[5]Departament de Medicina, Universitat de Barcelona, Barcelona, Spain

Corresponding Author:
Josep Maria Haro, PhD
Research and Development Unit, Parc Sanitari Sant Joan de Déu
Institut Sant Joan de Déu
C/Dr Antoni Pujada 42
Sant Boi de Llobregat
Barcelona, 08830
Spain
Phone: 34 93 640 63 50
Email: josepmaria.haro@sjd.es

Abstract

Background: The overall pandemic created enormous pressure on long-term care workers (LTCWs), making them particularly vulnerable to mental disorders. Despite this, most of the available evidence on professional well-being during COVID-19 has exclusively focused on frontline health care workers.

Objective: This study aimed to identify the long-term psychological needs of LTCWs derived from the COVID-19 pandemic and to explore barriers and facilitators related to digital mental health tools. This is part of a project that seeks to develop a digital mental health intervention to reduce psychological distress in this population group.

Methods: We performed a qualitative study with a rapid research approach. Participants were LTCWs of the autonomous community of Catalonia. We conducted 30 semistructured interviews between April and September 2022. We used a qualitative content analysis method with an inductive-deductive approach.

Results: The period of the pandemic with the highest mental health burden was the COVID-19 outbreak, with almost all workers having experienced some form of emotional distress. Emotional distress persisted over time in more than half of the participants, with fatigue and nervousness being the main emotions expressed at the time of the interview. High workload, the feeling that pandemic times are not over, and poor working conditions that have remained since then have been the most frequently expressed determinants of such emotions. Potential barriers and facilitators to engagement with digital tools were also identified in terms of previous experience and beliefs of the target population, possibilities for the integration of a digital tool into daily life, preferences regarding the level of guidance, the possibility of social connectedness through the tool, and privacy and confidentiality. The identified factors may become especially relevant in the context of the pandemic remission phase.

Conclusions: More than 2 years after the pandemic outbreak, emotional distress is still relevant. The persistent burden of psychological distress points to a need for institutions to take action to improve working conditions and promote employees' well-being. Considering factors that act as barriers and facilitators for the use of digital mental health tools, it is important to develop tailored tools that could offer valuable support to this population during and after a pandemic.

(J Med Internet Res 2024;26:e47546) doi:10.2196/47546

XSL•FO
RenderX

KEYWORDS

COVID-19 pandemic; digital technology; health care professionals; long-term care; mental health; well-being; digital mental health; digital mental health interventions; mobile phone

Introduction

Long-term care facilities, such as nursing homes and other assisted living facilities, have been hit particularly hard by the COVID-19 pandemic. By February 2021, approximately 41% of the global COVID-19-associated mortality occurred in long-term care residents [1]. In Spain, the COVID-19 outbreak entailed especially tragic consequences, mainly caused by the precariousness of these care systems [2]. According to data from the Spanish Ministry of Health, 27,359 long-term residents died between April 6 and June 20, 2020, which represents approximately 70% of the total COVID-19 deaths. Long-term care workers (LTCWs) have been constantly exposed to anguish and death, have witnessed the social isolation and loneliness of residents due to health restrictions, and have experienced a drastic increase in their working demands [3-5]. Consequently, the mental health and well-being of this population group have been compromised, with clinically relevant levels of stress, anxiety, and depression increasing to nearly 60% from March 2020 to June 2020 [4].

As we approach the third year of the pandemic, health and care workers continue to respond to uncertainties, potential new waves, and the long-term effects of COVID-19. Importantly, recent evidence highlights increased concerns for burnout in this population [6], which can ultimately jeopardize the quality of health care as well as patient safety [7]. While mounting evidence has assessed the mental health experiences, views, and needs of care professionals during the early stages of the pandemic, there is scarce knowledge about the long-term emotional impact. This information is crucial to develop tailored interventions in response to their emotional demands.

Digital mental health tools (DMHTs) are a promising strategy to mitigate psychological consequences in the context of a long-lasting pandemic since they can be delivered remotely, avoiding the risk of transmission of the infection [8]. Yet, low levels of user engagement have been recognized as important barriers that may compromise the efficacy and scalability of interventions [9]. Increasing (prepandemic) evidence is beginning to unravel key factors that influence user engagement and the effectiveness of engagement strategies. For instance, positive prior experiences with mental health services and technology, digital literacy, the credibility of content, guidance within the intervention, increased social connectedness, a greater number of engagement features, reminders, and tailored feedback have been identified as potential facilitators for promoting engagement or effectiveness of digital interventions [10-13]. Nonetheless, these are general features reported in studies that included a wide range of different populations (eg, refugees, general population, people who are homeless, and people with a diagnosis of mental disorders), and it is well recognized that understanding the settings and population's needs (eg, COVID-19 pandemic and long-term care facility characteristics) is a key factor for user engagement and intervention implementation success [14].

In this context, this study aimed to identify the long-term psychological needs of LTCWs as well as explore barriers and facilitators in relation to DMHTs. These findings will be used to inform and guide the development of a digital intervention to reduce psychological distress in this population group. The effectiveness of the psychological digital intervention will be tested through a randomized clinical trial (ClinicalTrials.gov NCT05526235).

Methods

Study Design, Sampling, and Participants

This qualitative study was conducted with a phenomenological orientation, in which we used a rapid research approach. In time-sensitive contexts, such as a pandemic, rapid appraisals allow us to collect and analyze data in a targeted and iterative way within limited time frames and, consequently, adapt and design expedient interventions for improving health care [15]. For reporting purposes, we adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [16].

Semistructured interviews were conducted with workers from long-term care facilities (including nursing homes and other long-term care facilities), who were on duty during the pandemic in the autonomous community of Catalonia. For the recruitment of participants, we contacted several long-term care facilities and presented the project to each representative. Once the facility agreed to participate, the managerial staff shared the study information sheet among their employees (via their usual official means of communication). When needed, the researcher conducted informative talks (in person or via teleconference) with LTCWs in order to disseminate the study and resolve doubts, if any. A total of 6 long-term care facilities from the metropolitan area of Barcelona participated in the study.

Ethical Considerations

Ethics approval was provided by the Fundació Sant Joan de Déu Ethics Committee, Barcelona, Spain (PIC 214-21). Participants were fully informed about the objectives and procedures of the study prior to signing the informed consent. The data set was pseudoanonymized, and personal data were stored separately and securely on institution servers.

Data Collection

Interviews were conducted between April and September 2022. They were individual, by videoconference, and lasted between 25 and 40 minutes. The interview guide (provided in Multimedia Appendix 1) was reviewed by 2 experts in quality-of-life research to ensure its adequacy, and it was then tested with the first 2 interviews. No adaptations had to be made. The interview guide related to the digital mental health section was based on the framework developed by Borghouts et al [11]. Participants were given the option of conducting the interview in Spanish or Catalan, according to their preferences. Two researchers (AM-M and LG-S) conducted the interviews. Both are female

XSL•FO
RenderX

PhD candidates. AM-M conducted the interviews in Catalan (8/30, 27% interviews), and LG-S conducted the interviews in Spanish (22/30, 73%). Both interviewers were previously trained in qualitative interview methods. Participants were informed of the interviewers' educational and professional background, and no working or personal relationship existed between participants and interviewers. All interviews were audio recorded, and the main points were documented in real-time notes. At the conclusion of each interview, the emerging findings were synthesized using a RREAL (Rapid Research Evaluation and Appraisal Lab) Sheet (provided in Multimedia Appendix 2). The RREAL Sheet is a flexible working document, normally presented as a table, designed for the collection, synthesis, reporting, and analysis of data [17]. It enhances familiarization with the data and facilitates analysis while data collection is still ongoing [15,17]. First, a table organized by categories was designed using the interview script as a guide, and we piloted or amended this RREAL Sheet during initial data collection. Then, after each interview, we registered the key data obtained using real-time notes. In instances where pertinent details were absent, we referred to the audio recordings for further clarification. Importantly, the categories represented in the table were modified whenever considered necessary during data collection. Finally, the RREAL Sheet was used to guide an in-depth analysis [17]. The sample size was determined by thematic saturation [18].

Data Analysis

We used a qualitative content analysis method with an inductive-deductive approach. We performed a 1-page table that summarized the core components registered on the RREAL Sheet to identify key themes [17]. Once key themes and specific issues (subthemes) were recognized, we selected quotes from the interviews that could exemplify these themes. Investigator triangulation ensured that the themes reflected the full range and depth of the data.

Results

Sample Characteristics

A total of 30 LTCWs participated in the study. The mean age was 44 (SD 11.4) years, most of our sample were women (n=26, 87%), and one-third (n=10, 33%) were from foreign nationalities. The vast majority of the participants were geriatric nursing assistants (n=17, 57%), followed by nurses (n=5, 17%). A more detailed description of the sociodemographic and occupational characteristics of the sample can be found in Table 1.

XSL•FO
RenderX

Table 1. Sample characteristics (N=30).

Variables	Values, n (%)
Age range (years)	
<35	7 (23)
35-49	10 (33)
≥50	13 (43)
Gender	
Woman	26 (87)
Man	4 (13)
Nationality	
Spanish	20 (67)
Foreign	10 (33)
Professional role	
Nurse	5 (17)
Psychologist	2 (7)
Geriatric nursing assistant	17 (57)
Physiotherapist	2 (7)
Others[a]	4 (13)
Years in service	
<5	7 (23)
5-15	13 (43)
>15	10 (33)
Workday	
Full time	21 (70)
Part time	9 (30)

[a]Director (n=1), occupational therapist (n=1), social worker (n=1), and cleaning staff (n=1).

Impact of COVID-19 on LTCWs' Mental Health

Almost all participants identified the COVID-19 outbreak as the worst period for their mental health and reported having experienced emotional distress during this stage. The main emotions expressed were fear, helplessness, abandonment, loneliness, and sadness. Less frequently expressed emotions also included frustration, anger, uncertainty, and exhaustion. Five themes emerged as the main perceived determinants of these emotions:

1. Unexpected and sudden nature of the crisis: Most participants expressed problems related to facing something new and completely unknown for which they were not prepared and its consequences. The reported main causes of distress were lack of information and knowledge, shortage of materials and human resources, problems with institutional organization, and the constant change of protocols and having to readjust to them.
2. High exposure to emotional anguish and death: Half of the participants referred to high exposure to death and emotional anguish. They mainly mentioned the high mortality among residents and the helplessness of witnessing their isolation due to suspended family visits and reduced interactions resulting from health restrictions. Several participants also referred to the emotional anguish caused by the deaths of colleagues, family members, or friends as well as seeing the anguish in their colleagues on a daily basis.
3. Fear of infection: Almost half of the participants mentioned the strong fear of infecting themselves, residents, and family members as well as the potential consequences of infection (eg, the severity of the disease and death).
4. Moral distress: Many workers reported feelings of hopelessness, expressing that nothing they did for the residents in terms of care was enough. This was accompanied, in many cases, by a sense of loss of control of the situation during their caregiving duties.
5. Lack of support: Many workers referred to a lack of recognition and support from authorities (both at the workplace and from the overall health authorities).

Textbox 1 provides an overview of these themes with example quotes for supporting data.

XSL•FO
RenderX

Textbox 1. Main determinants of emotional distress during the outbreak and example quotes.

Unexpected and sudden nature of the crisis

- "We were not prepared; it came from one moment to the next." [REPICAL (Reducing the psychosocial impact of the Covid-19 pandemic on workers of assisted living facilities)-007]
- "I also had to guide others in my charge, about things that I did not know either." [REPICAL-004]
- "The protocols that we received were not clear...We were all in a drifting boat." [REPICAL-008]
- "Overnight we were alone with co-workers. We were very lost; we did not have personal protective equipment and we did not know how to act. There were 4 of us and we managed as best as we could." [REPICAL-026]

High exposure to emotional anguish and death

- "The worst thing I experienced was seeing so many residents die in such a short time, people who were fine and the next day they were gone." [REPICAL-019]
- "We were not aware of the magnitude of this until residents began to die, they began to die one after the other, 4 or 5 per day (when before maybe one died every 2 months)." [REPICAL-026]
- "Seeing the loneliness in the residents affected me a lot, seeing that they were going to die alone, that they were not able to be with their family. It also affected me to see my co-workers, the most cheerful ones, defeated." [REPICAL-021]
- "I felt so helpless seeing the isolated residents. We were asked not to have contact with them, but they needed someone to hold their hand." [REPICAL-030]

Moral distress

- "We had to take measures that, while necessary, we knew were not good for the residents." [REPICAL-003]
- "No matter what we did, people kept dying." [REPICAL-004]
- "We saw that they were unwell and we could not help them...we wanted to address everything, but we could not." [REPICAL-028]

Fear of infection

- "I even moved and lived in the residence for a while, so as not to infect the residents." [REPICAL-019]
- "I had to isolate myself at home and I could not be with my child because I was afraid of infecting him. I also used to think: if something happens to me and my husband, who will take care of my child?" [REPICAL-023]

Lack of support

- "What I felt most was abandonment." [REPICAL-002]
- "Feeling of helplessness of not being able to do anything and not having help from anyone." [REPICAL-010]

In relation to their current emotional state, over half of the participants expressed some type of persistence of emotional distress over time. Among all participants, a minority reported maintaining intense distress, and a significant portion expressed only some improvement since the outbreak. The emotions most often mentioned were fatigue and nervousness, followed by anger and sadness. Three themes emerged as the main determinants of the persistent emotional distress:

1. High workload: Long shifts due to staff shortages and coping with postcrisis backlog.
2. Pandemic times are not over: Fatigue and residual discomfort from being under stress for so long as well as the feeling of not being able to completely move forward from COVID-19 due to the nature of their work.
3. Still poor working conditions: Although the pandemic clearly highlighted the precariousness of these health care systems, some participants claimed that no measures have been taken to address this situation and that, as a result, working conditions remain poor. Relatedly, LTCWs felt that there is an overall failure in learning from the lessons of the COVID-19 pandemic.

Textbox 2 provides an overview of these themes with example quotes for supporting data.

On the other hand, among those participants who expressed a noticeable improvement over time in their emotional discomfort, hope and relief were the most commonly mentioned emotions. Improvement was associated with an increased sense of control over the situation, confidence in vaccines, decreased measures of isolation, and the feeling of regaining prepandemic working dynamics.

XSL•FO
RenderX

Textbox 2. Main determinants of persistent emotional distress and example quotes.

High workload

- "I am overwhelmed. In 2020, they increased my working hours. I was told it would only be during the state of emergency, but I am still not doing my usual schedule. In addition, people have left the workforce, and they have not hired more. I do not have as much time as I would like for other things." [REPICAL-004]
- "Many projects were stopped because of the pandemic and now they have to be ready overnight. The workload is unreal." [REPICAL-001]
- "Now we are facing our patients' pathologies that were neglected during these two years. I am exhausted." [REPICAL-005]

Pandemic times are not over

- "I am tired and burned out from dragging on for so long in tension, needing it to be over and it will not end." [REPICAL-003]
- "Everyone has moved on, but I am still there: with the masks, the protective equipment, the heat." [REPICAL-031]
- "During the crisis, we were alert at night in case a co-worker called us for help, and we always called each other; now I keep waking up even if they do not call me." [REPICAL-026]

Still poor working conditions

- "With the pandemic, deficits that already existed in our field have come to light, and the thing is that nothing has changed, it's just going back to the old ways. Nothing has been learned." [REPICAL-002]

Digital Mental Health Tools

Previous Experience and Beliefs

Almost all participants reported having no previous experience with DMHTs and emphasized their lack of confidence in their own digital skills. Nonetheless, the vast majority reported believing that such tools could be helpful in improving their mental health. The minority who reported not having confidence in digital tools mainly referred to their preference for personal, face-to-face contact.

Integration Into Daily Life

When participants were asked about what characteristics a DMHT should have to make it easy for them to integrate it into their daily lives, the majority of participants referred to the importance of accessibility: "easy to use," "intuitive," "didactic," and "visual." Some respondents also highlighted the importance of flexibility in using the tool in a time-convenient manner as well as being able to access from a mobile phone.

Level of Guidance and Social Connectedness

Regarding the level of guidance or support when using DMHTs, the majority of our sample reported preferring a guided intervention via a coach rather than a self-guided tool, expressing that the figure of a coach generates more confidence in the intervention and a greater sense of "closeness" despite the lack of in-person contact. Others mentioned the possibility of combining both modalities based on preference or demand. As for the profile of the coach, the most frequently mentioned professional role was a psychologist. However, rather than a professional role, most participants placed greater emphasis on the coach's personal characteristics and skills, mainly highlighting empathy, active listening, receptivity, dynamism, and the ability to inspire confidence. As for the type of communication with the coach, the most preferred modality was videoconferencing, followed by telephone contact, and finally written messages. Some also mentioned the possibility of combining modalities, depending on the content and personal needs. Most of the participants preferred to accessing the content of the tool in stages or organized in modules rather than having it all available from the beginning. They indicated that this organization in modules would increase learnability by helping them navigate the tool in a more structured manner. Regarding content, some participants spontaneously emphasized the importance of having specific mental health content tailored to their needs rather than broad or general information. In addition, most noted that a reminder system, such as alarms, would be useful to aid engagement. Finally, regarding the possibility of being able to communicate with other users of the tool, a majority expressed that it would be a good idea, while a smaller group had doubts, and a few stated they would not use this option even if available.

Privacy and Confidentiality

Most participants reported that they would trust confidentiality protection mainly on the premise that when such an intervention is implemented within the context of a research study and with institutional backing, confidentiality should already be guaranteed. Among the few who expressed concerns, these were related to data privacy (sharing data with third parties), mentioning their concern about a possible link with social networks and the use of private data for advertising. On the other hand, almost all participants preferred nonanonymity with the figure of the coach to further "humanize" the intervention and increase confidence.

Discussion

Principal Findings

Our results showed that the worst stage of the pandemic in terms of mental health was the outbreak, with almost all workers experiencing some form of emotional distress. The main emotions that participants expressed during this stage, as well as their determinants, are consistent with those reported by other national and international studies conducted on frontline health care workers during the first wave of the pandemic [3-5,19-22].

XSL•FO
RenderX

Importantly, our results showed that perceived emotional distress, main emotions, and their determinants changed over time. We found that almost half of the sample reported an improvement in their emotional discomfort over time, expressing hope and relief. This improvement was associated with a greater feeling of control over the situation, confidence in vaccines, decreased measures of isolation, and the feeling of regaining prepandemic working dynamics. Nonetheless, more than 2 years after the outbreak, emotional distress persisted for over half of the sample. Fatigue and nervousness were the main expressed emotions, which were determined by a high workload, the feeling that pandemic times are not over, and sustained poor working conditions. This result is consistent with other epidemiological studies, which showed that poor mental health outcomes among health care workers tend to persist over time [23-25] and that the main symptoms of poor mental health tend to change with the different COVID-19 stages as well [26].

Fatigue, nervousness, and emotional exhaustion, which may occur in response to chronic work stressors, are common signs of burnout [27,28]. Fatigue, which is not only linked to lower job satisfaction and increased rates of absenteeism [29], can also impair concentration and slow reaction times, elevating the risk of workplace accidents or fatigue-related incidents [30]. The assessment of the determinants of such emotional distress provides valuable input into potential intervention strategies to respond to such calls. Multifaceted interventions with a holistic approach and implemented in a timely manner are needed to protect the mental health and well-being of these workers during and after the pandemic. Similar to recommendations and claims made for the health care workforce, interventions for care workers should also be targeted at the organizational or institutional level via systematic support and at the individual level by implementing psychological interventions for those workers with emotional distress [29,31,32]. It is not enough that the institutions have been reactive to the pandemic with concrete measures to deal with a sudden crisis. They need to implement long-lasting workplace changes to improve working conditions and promote the long-term well-being of their staff. At the individual level, offering tailored digital psychological interventions using cognitive behavioral therapy is a promising option for workers experiencing emotional distress [29].

Since digital mental health interventions can represent a promising individual-level strategy to improve the mental health of LTCWs, we identified several barriers and facilitators for their successful engagement in the context of the COVID-19 remission phase.

Not only did almost all participants deny having any previous experience in the use of DMHTs, but almost a half of them also spontaneously referred to their lack of skills in the use of digital tools. This is a potential barrier that may be linked to the mean age of the sample (44, SD 11.4 years), since previous studies found that younger people have higher adherence to this type of intervention than older people [33,34]. Despite this lack of experience, the vast majority of participants reported believing that using DMHTs could be helpful in improving their mental health. This is an important facilitator since people's expectations and preconceived beliefs about whether a tool will be effective positively influence experience and engagement [35-37]. In this regard, addressing the expectations and beliefs of the target population before developing a DMHT (or as an early step during the development process) could help identify and consequently manage early barriers [37].

Most participants highlighted the importance of accessibility when using a DMHT. Similar to previous studies [9,37], we found that achieving "user-friendly" tools is a key facilitator for engagement. In addition, participants also highlighted the importance of flexibility, which would enable them to use the tool according to their needs and changing working shifts.

Regarding the level of guidance or support, the majority of participants reported a preference for a guided intervention with a human coach rather than a self-guided tool, which is also in line with findings from previous studies [11,37]. Increased support may enhance and extend engagement in DMHTs. In the context of a pandemic, where social contact may be limited, this feature becomes even more relevant. Reminder systems are also important facilitators since they could prevent forgetfulness and encourage users to stay committed [37]. Furthermore, satisfaction with the type of content and the manner the materials are offered is critical to engagement. Interestingly, most of our participants express a preference for the tool's content to be delivered in stages or modules, aiming to enhance learnability and facilitate structured navigation. Additionally, some participants underscore the significance of receiving personalized content that fits their needs or interests. Therefore, addressing user needs is key to preventing dropouts and ensuring sustained engagement [9]. In turn, most of our sample expressed that having the possibility to communicate with other users of the tool would be beneficial. Prior to the pandemic, social connectedness through DMHTs had proven to be not only a facilitator for engagement [38] but also a tool with therapeutic value per se [39]. In the current context, where social interactions have changed, this possibility could be even more relevant.

Finally, most participants in our study reported a high level of trust in confidentiality. This places significant responsibility on those developing mental health interventions through DMHTs. Research has indicated that the transmission of data to third parties by mental health smartphone apps is prevalent, thereby denying users an informed choice regarding whether or not to accept such sharing [40]. This issue may be a consequence of most DMHTs existing outside of health care regulation and also falling outside of health care privacy legislation [9]. A clear and transparent written privacy policy should always be available to inform users about how their data are going to be used.

At this juncture, it is crucial to contemplate the transferability of our findings to other work environments and populations. Although our research focused on a specific group in a particular context, the emerging themes and shared emotional experiences might resonate in similar situations. The fluctuation in perceptions over time provides a dynamic perspective that can be insightful for understanding long-term impacts in the context of future global pandemics. On the other hand, the impact of variables such as age or digital skills, perceived as potential barriers to the adoption of DMHTs, could be extrapolated to other populations, such as informal carers. The preference for

XSL•FO
RenderX

guided interventions and the significance of accessibility and flexibility, among other variables, are also elements that might have broader applications in the design of DMHTs. It is imperative to acknowledge that each work environment and demographic group has its own unique characteristics, and the direct application of our results may necessitate adjustments. Nevertheless, by presenting these findings in a detailed and transparent manner, we aim to provide valuable insights that can guide future research and intervention strategies in diverse settings.

Limitations

This study acknowledges certain limitations that are crucial for contextualizing and evaluating the robustness of the findings. First, we did not account for participants' preexisting mental health conditions (prior to COVID-19) when we assessed mental health needs. This could bear significant implications, as individuals with preexisting mental disorders are particularly vulnerable to the mental health threat of the pandemic. Second, the uneven gender distribution in our sample, with only 4 men, poses a limitation in terms of generalizing findings to this group. While our sample was not centralized in a single location, it is worth noting that all participating centers are located in the metropolitan area of Barcelona, which may limit the generalizability of our findings to the broader Catalonia region. Future studies should use more heterogeneous samples. Finally, inherent in qualitative research, there may be a potential for selection bias since workers were to some extent self-selected.

Conclusions

We identified that the worst stage of the pandemic in terms of LTCWs' mental health was the COVID-19 outbreak, in which all workers reported having experienced some form of emotional distress. We also identified a persistence of emotional distress over time in more than a half of the participants, with fatigue and nervousness being the most frequently expressed emotions. This suggests that, although many workers significantly improved their psychological discomfort, mental health problems in this group are still relevant even more than 2 years after the pandemic outbreak. Future studies are needed to determine the factors that promote or hinder resilience among this underrepresented population group in order to shape implementation strategies to promote well-being. Finally, this study also identified new barriers and facilitators to engagement with DMHTs during the remission phase of the COVID-19 pandemic. These findings provide key information for the development of tailored digital mental health interventions among LTCWs.

Acknowledgments

This work was supported by the Instituto de Salud Carlos III (CD20/00036 to MF-N) and by la Fundació La Marato TV3 (202114 to JMH). The funding bodies had no role in study design, data collection and analysis, interpretation of data, or writing the manuscript.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LG-S conceptualized the study, conducted methodology, performed formal analysis, and drafted the original manuscript. AM-M conducted methodology and performed formal analysis. CV-P contributed to conceptualization and methodology. SA-O and BC contributed to conceptualization. JMH conceptualized the study, contributed to methodology, and acquired funding. MF-N conceptualized the study, conducted methodology, managed project administration, and participated in writing, reviewing, and editing. LG-S, AM-M, CV-P, SA-O, BC, and JMH participated in reviewing and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[PDF File (Adobe PDF File), 97 KB - jmir_v26i1e47546_app1.pdf]

Multimedia Appendix 2
Rapid Research Evaluation and Appraisal Lab Sheet.
[PDF File (Adobe PDF File), 64 KB - jmir_v26i1e47546_app2.pdf]

References

1. Comas-Herrera A, Zalakaín J, Lemmon E, Henderson D, Litwin C, Hsu AT, et al. Mortality associated with COVID-19 in care homes: international evidence. Article in LTCcovid.org, International Long-Term Care Policy Network, CPEC-LSE.

XSL•FO
RenderX

2021. URL: https://ltccovid.org/wp-content/uploads/2021/02/LTC_COVID_19_international_report_January-1-February-1-2.pdf [accessed 2024-05-02]
2. Rada AG. COVID-19: the precarious position of Spain's nursing homes. BMJ 2020;369:m1554 [FREE Full text] [doi: 10.1136/bmj.m1554] [Medline: 32312710]
3. Blanco-Donoso LM, Moreno-Jiménez J, Amutio A, Gallego-Alberto L, Moreno-Jiménez B, Garrosa E. Stressors, job resources, fear of contagion, and secondary traumatic stress among nursing home workers in face of the COVID-19: the case of Spain. J Appl Gerontol 2021;40(3):244-256 [FREE Full text] [doi: 10.1177/0733464820964153] [Medline: 33025850]
4. Hering C, Gangnus A, Budnick A, Kohl R, Steinhagen-Thiessen E, Kuhlmey A, et al. Psychosocial burden and associated factors among nurses in care homes during the COVID-19 pandemic: findings from a retrospective survey in Germany. BMC Nurs 2022;21(1):41 [FREE Full text] [doi: 10.1186/s12912-022-00807-3] [Medline: 35139842]
5. Palacios-Ceña D, Fernández-Peña R, Ortega-López A, Fernández-Feito A, Bautista-Villaécija O, Rodrigo-Pedrosa O, et al. Long-term care facilities and nursing homes during the first wave of the COVID-19 pandemic: a scoping review of the perspectives of professionals, families and residents. Int J Environ Res Public Health 2021;18(19):10099 [FREE Full text] [doi: 10.3390/ijerph181910099] [Medline: 34639401]
6. Shanafelt TD, West CP, Dyrbye LN, Trockel M, Tutty M, Wang H, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022;97(12):2248-2258 [FREE Full text] [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]
7. Kaushik D. COVID-19 and health care workers burnout: a call for global action. EClinicalMedicine 2021;35:100808 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100808] [Medline: 33778435]
8. Moreno C, Wykes T, Galderisi S, Nordentoft M, Crossley N, Jones N, et al. How mental health care should change as a consequence of the COVID-19 pandemic. Lancet Psychiatry 2020;7(9):813-824 [FREE Full text] [doi: 10.1016/S2215-0366(20)30307-2] [Medline: 32682460]
9. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
10. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016;18(1):e6 [FREE Full text] [doi: 10.2196/jmir.4790] [Medline: 26747176]
11. Borghouts J, Eikey E, Mark G, de Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
12. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020;22(12):e21687 [FREE Full text] [doi: 10.2196/21687] [Medline: 33295292]
13. Wu A, Scult MA, Barnes ED, Betancourt JA, Falk A, Gunning FM. Smartphone apps for depression and anxiety: a systematic review and meta-analysis of techniques to increase engagement. NPJ Digit Med 2021;4(1):20 [FREE Full text] [doi: 10.1038/s41746-021-00386-8] [Medline: 33574573]
14. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
15. Vindrola-Padros C, Chisnall G, Cooper S, Dowrick A, Djellouli N, Symmons SM, et al. Carrying out rapid qualitative research during a pandemic: emerging lessons from COVID-19. Qual Health Res 2020;30(14):2192-2204 [FREE Full text] [doi: 10.1177/1049732320951526] [Medline: 32865149]
16. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
17. Vindrola-Padros C, Chisnall G, Polanco N, Juan NVS. Iterative cycles in qualitative research: introducing the RREAL sheet as an innovative process. Soc Sci Humanit Open 2022 [FREE Full text] [doi: 10.31219/osf.io/9dp2w]
18. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
19. Fisher E, Cárdenas L, Kieffer E, Larson E. Reflections from the "forgotten front line": a qualitative study of factors affecting wellbeing among long-term care workers in New York City during the COVID-19 pandemic. Geriatr Nurs 2021;42(6):1408-1414 [FREE Full text] [doi: 10.1016/j.gerinurse.2021.09.002] [Medline: 34624698]
20. Mediavilla R, Monistrol-Mula A, McGreevy KR, Felez-Nobrega M, Delaire A, Nicaise P, et al. Mental health problems and needs of frontline healthcare workers during the COVID-19 pandemic in Spain: a qualitative analysis. Front Public Health 2022;10:956403 [FREE Full text] [doi: 10.3389/fpubh.2022.956403] [Medline: 35968478]
21. White EM, Wetle TF, Reddy A, Baier RR. Front-line nursing home staff experiences during the COVID-19 pandemic. J Am Med Dir Assoc 2021;22(1):199-203 [FREE Full text] [doi: 10.1016/j.jamda.2020.11.022] [Medline: 33321076]

XSL•FO
RenderX

22. Zhang X, Sheng Q, Wang X, Cai C. The experience of frontline nurses four months after COVID-19 rescue task in China: a qualitative study. Arch Psychiatr Nurs 2021;35(4):358-363 [FREE Full text] [doi: 10.1016/j.apnu.2021.05.007] [Medline: 34176576]
23. Alonso J, Vilagut G, Alayo I, Ferrer M, Amigo F, Aragón-Peña A, et al. Mental impact of COVID-19 among Spanish healthcare workers. A large longitudinal survey. Epidemiol Psychiatr Sci 2022;31:e28 [FREE Full text] [doi: 10.1017/S2045796022000130] [Medline: 35485802]
24. Mediavilla R, Fernández-Jiménez E, Martinez-Morata I, Jaramillo F, Andreo-Jover J, Morán-Sánchez I, et al. Sustained negative mental health outcomes among healthcare workers over the first year of the COVID-19 pandemic: a prospective cohort study. Int J Public Health 2022;67:1604553 [FREE Full text] [doi: 10.3389/ijph.2022.1604553] [Medline: 35814735]
25. Sasaki N, Asaoka H, Kuroda R, Tsuno K, Imamura K, Kawakami N. Sustained poor mental health among healthcare workers in COVID-19 pandemic: a longitudinal analysis of the four-wave panel survey over 8 months in Japan. J Occup Health 2021;63(1):e12227 [FREE Full text] [doi: 10.1002/1348-9585.12227] [Medline: 34021683]
26. Wang Y, Hu Z, Feng Y, Wilson A, Chen R. Changes in network centrality of psychopathology symptoms between the COVID-19 outbreak and after peak. Mol Psychiatry 2020;25(12):3140-3149 [FREE Full text] [doi: 10.1038/s41380-020-00881-6] [Medline: 32929212]
27. Denning M, Goh ET, Tan B, Kanneganti A, Almonte M, Scott A, et al. Determinants of burnout and other aspects of psychological well-being in healthcare workers during the COVID-19 pandemic: a multinational cross-sectional study. PLoS One 2021;16(4):e0238666 [FREE Full text] [doi: 10.1371/journal.pone.0238666] [Medline: 33861739]
28. Peng P, Liang M, Wang Q, Lu L, Wu Q, Chen Q. Night shifts, insomnia, anxiety, and depression among Chinese nurses during the COVID-19 pandemic remission period: a network approach. Front Public Health 2022;10:1040298 [FREE Full text] [doi: 10.3389/fpubh.2022.1040298] [Medline: 36544790]
29. De Kock JH, Latham HA, Cowden RG. The mental health of healthcare workers during the COVID-19 pandemic: a narrative review. Curr Opin Psychiatry 2022;35(5):311-316. [doi: 10.1097/YCO.0000000000000805] [Medline: 35855506]
30. Wong I, O'Connor MB. COVID-19 and workplace fatigue: lessons learned and mitigation strategies. Centers for Disease Control and Prevention. 2021. URL: https://blogs.cdc.gov/niosh-science-blog/2021/01/13/covid-19-fatigue/ [accessed 2024-05-02]
31. Muller AE, Hafstad EV, Himmels JPW, Smedslund G, Flottorp S, Stensland S, et al. The mental health impact of the COVID-19 pandemic on healthcare workers, and interventions to help them: a rapid systematic review. Psychiatry Res 2020;293:113441 [FREE Full text] [doi: 10.1016/j.psychres.2020.113441] [Medline: 32898840]
32. Murray E, Kaufman KR, Williams R. Let us do better: learning lessons for recovery of healthcare professionals during and after COVID-19. BJPsych Open 2021;7(5):e151 [FREE Full text] [doi: 10.1192/bjo.2021.981] [Medline: 34457351]
33. Abel EA, Shimada SL, Wang K, Ramsey C, Skanderson M, Erdos J, et al. Dual use of a patient portal and clinical video telehealth by veterans with mental health diagnoses: retrospective, cross-sectional analysis. J Med Internet Res 2018;20(11):e11350 [FREE Full text] [doi: 10.2196/11350] [Medline: 30404771]
34. Kannisto KA, Korhonen J, Adams CE, Koivunen MH, Vahlberg T, Välimäki MA. Factors associated with dropout during recruitment and follow-up periods of a mHealth-based randomized controlled trial for Mobile.Net to encourage treatment adherence for people with serious mental health problems. J Med Internet Res 2017;19(2):e46 [FREE Full text] [doi: 10.2196/jmir.6417] [Medline: 28223262]
35. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
36. Nitsch M, Dimopoulos CN, Flaschberger E, Saffran K, Kruger JF, Garlock L, et al. A guided online and mobile self-help program for individuals with eating disorders: an iterative engagement and usability study. J Med Internet Res 2016;18(1):e7 [FREE Full text] [doi: 10.2196/jmir.4972] [Medline: 26753539]
37. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
38. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
39. Townsend L, Zippay A, Caler K, Forenza B. Technology and opportunity: people with serious mental illness and social connection. J Soc Soc Work Res 2016;7(2):371-393 [FREE Full text] [doi: 10.1086/686882]
40. Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw Open 2019;2(4):e192542-e192542 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2542] [Medline: 31002321]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

XSL•FO
RenderX

DMHT: digital mental health tool
LTCW: long-term care worker
RREAL: Rapid Research Evaluation and Appraisal Lab
REPICAL: reducing the psychosocial impact of the Covid-19 pandemic on workers of assisted living facilities

Edited by T Leung; submitted 24.03.23; peer-reviewed by A AL-Asadi, P Petri-Romao, S Rennick-Egglestone; comments to author 19.01.24; revised version received 08.02.24; accepted 14.02.24; published 29.05.24.

Please cite as:
González-Spinoglio L, Monistrol-Mula A, Vindrola-Padros C, Aguilar-Ortiz S, Carreras B, Haro JM, Felez-Nobrega M
Long-Term Emotional Impact of the COVID-19 Pandemic and Barriers and Facilitators to Digital Mental Health Tools in Long-Term Care Workers: Qualitative Study
J Med Internet Res 2024;26:e47546
URL: https://www.jmir.org/2024/1/e47546
doi:10.2196/47546
PMID:38809605

©Leticia González-Spinoglio, Anna Monistrol-Mula, Cecilia Vindrola-Padros, Salvatore Aguilar-Ortiz, Bernat Carreras, Josep Maria Haro, Mireia Felez-Nobrega. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tracking and Profiling Repeated Users Over Time in Text-Based Counseling: Longitudinal Observational Study With Hierarchical Clustering

Yucan Xu[1], MA; Christian Shaunlyn Chan[2], PhD; Evangeline Chan[3], MA; Junyou Chen[4], MA; Florence Cheung[4], MA; Zhongzhi Xu[5], PhD; Joyce Liu[4], MA; Paul Siu Fai Yip[6], PhD

[1]School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China (Hong Kong)
[2]Department of Psychology, The University of Hong Kong, Hong Kong, China (Hong Kong)
[3]Department of Social and Behavioural Sciences, City University of Hong Kong, Hong Kong, China (Hong Kong)
[4]The Hong Kong Jockey Club Centre for Suicide Research and Prevention, The University of Hong Kong, Hong Kong, China (Hong Kong)
[5]School of Public Health, Sun Yat-sen University, Guang Zhou, China
[6]Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Paul Siu Fai Yip, PhD
Department of Social Work and Social Administration
The University of Hong Kong
Pokfulam
Hong Kong
China (Hong Kong)
Phone: 852 91401568
Email: sfpyip@hku.hk

Abstract

Background: Due to their accessibility and anonymity, web-based counseling services are expanding at an unprecedented rate. One of the most prominent challenges such services face is *repeated users*, who represent a small fraction of total users but consume significant resources by continually returning to the system and reiterating the same narrative and issues. A deeper understanding of repeated users and tailoring interventions may help improve service efficiency and effectiveness. Previous studies on repeated users were mainly on telephone counseling, and the classification of repeated users tended to be arbitrary and failed to capture the heterogeneity in this group of users.

Objective: In this study, we aimed to develop a systematic method to profile repeated users and to understand what drives their use of the service. By doing so, we aimed to provide insight and practical implications that can inform the provision of service catering to different types of users and improve service effectiveness.

Methods: We extracted session data from 29,400 users from a free 24/7 web-based counseling service from 2018 to 2021. To systematically investigate the heterogeneity of repeated users, hierarchical clustering was used to classify the users based on 3 indicators of service use behaviors, including the duration of their user journey, use frequency, and intensity. We then compared the psychological profile of the identified subgroups including their suicide risks and primary concerns to gain insights into the factors driving their patterns of service use.

Results: Three clusters of repeated users with clear psychological profiles were detected: episodic, intermittent, and persistent-intensive users. Generally, compared with one-time users, repeated users showed higher suicide risks and more complicated backgrounds, including more severe presenting issues such as suicide or self-harm, bullying, and addictive behaviors. Higher frequency and intensity of service use were also associated with elevated suicide risk levels and a higher proportion of users citing mental disorders as their primary concerns.

Conclusions: This study presents a systematic method of identifying and classifying repeated users in web-based counseling services. The proposed bottom-up clustering method identified 3 subgroups of repeated users with distinct service behaviors and psychological profiles. The findings can facilitate frontline personnel in delivering more efficient interventions and the proposed method can also be meaningful to a wider range of services in improving service provision, resource allocation, and service effectiveness.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50976) doi:10.2196/50976

KEYWORDS

web-based counseling; text-based counseling; repeated users; frequent users; hierarchical clustering; service effectiveness; risk profiling; psychological profiles; psycholinguistic analysis

Introduction

Background

In response to the growing demand for accessible mental health support, the number of text-based counseling services has grown exponentially over the past decade [1-3]. Whereas phone-based counseling or telephone counseling services provide mental health support through phone calls or video chats [4], text-based counseling services do so through SMS text messages or emails [5,6]. Similar to telephone counseling services, many text-based counseling services also follow a model of crisis intervention with the intent to provide one-off interventions [7]. The aim of such services is to provide timely mental health support to a wider population, including those who cannot afford consultation fees or are unwilling to speak over the phone.

Several unique features contribute to the popularity of text-based counseling. First, the anonymous nature of text-based communication can reduce the potential stigma that offline services or phone-based services carry [5]. Second, text-based counseling allows users to contact the service and send messages at any time and from anywhere at their own pace, which further reduces their barriers to seeking help [6]. Third, text-based counseling is appealing to people who prefer written communication or have social anxiety [8]. However, new challenges emerge when these services struggle to meet the diverse needs of different types of users, whose service use may vary greatly. While most users may use the service once or just a few times, some might do so frequently. A systematic review of frequent callers to helplines has addressed the concern of inefficient use of service caused by the disproportionate resources taken up by frequent users [9]. In text-based services, frequent users might hinder the effectiveness of the service platforms, as they could overburden the system and reduce other users' service accessibility [10]. Individuals who frequently reach out for support without any constraints would occupy many sessions and service hours, which can result in extended waiting times for other help seekers, thereby impeding their access to the service. These users may also develop a reliance on the counselor, creating a challenge for both parties involved [11]. Therefore, it is imperative for service providers to understand the varying levels of needs and dependence of their users.

Frequent Users

In the studies of hotlines and helplines, *frequent users* is a term used to describe those who make significant volumes of calls within a certain period [9]. The characteristics of this group of users have been extensively studied, especially in Australia and the United States. A few demographic characteristics of identified frequent users seem to be consistent across studies. For example, frequent users are more likely to be unmarried, have significant mental health conditions, and be unemployed or have financial difficulties [12-17]. These patterns suggest that, to some extent, users' frequent and intensive use of the services is associated with some demographic characteristics; that is, there are systematic individual differences in how likely a person would repeatedly use a service.

In terms of suicide risk, although many studies have shown that frequent users are more likely to present suicide ideation [12,17,18], a systematic review suggested otherwise; that is, there was no difference in the level of suicidality between frequent and nonfrequent users [14]. One possible reason for such contradicting results may be the differences in the nature and features of the services. However, the more apparent and plausible reason is that the categorization of frequent users is more often than not ad hoc, arbitrary, and thus inconsistent across studies. The lack of a clear and consistent definition for frequent users poses both theoretical and practical challenges to those seeking to understand and better serve this subgroup of service users.

Repeated Users

Another term used in the literature is *repeated users*, which refers to those who seek help from helplines or crisis lines more than once [9,18]. Other earlier studies have adopted more restrictive definitions, including inappropriate use [19] or generating multiple calls [20]. The contact pattern of repeated users could vary greatly. Some might return once while others might return daily or monthly. Similarly, some repeated users might frequently contact a service intermittently over a concentrated period, whereas others might do so consistently across a longer time span. Those who use a service occasionally with genuine needs and distress usually do not pose a significant burden to the service providers. Similarly, users who contact the service repeatedly during a certain period and stop when their problems have been resolved are also likely considered a desired outcome for both users and providers. However, some repeated users may become persistent and intensive users who treat these services as long-term support without recognizing the services' limits or boundaries [15,18]; this type of users can become a significant challenge to the service provider, as they only represent a small fraction of all users while occupying a disproportionate amount of resources [17] at the expense of other nonfrequent users' opportunity to access the service [18]. As one of the counseling goals is to promote the autonomy and self-efficacy of clients, frequent, on-demand support might become counterproductive in the long run [21].

Many studies have used the terms *repeated users* and *frequent users* interchangeably, rather than distinguishing them according to their service behaviors, characteristics, and help-seeking needs. A systematic review of frequent or repeated helpline callers found that one-third of the studies provided no information on how these groups were defined [9]. Another one-third of the reviewed studies mixed frequent callers with

XSL•FO
RenderX

repeated callers, and the remaining one-third used *greater than 8 calls a month* as their definition of frequent users. However, no clear explanation was given for this 8 times-per-month threshold. The mixing of frequent and repeated users as one homogenous group and the lack of a shared understanding of categorization and definition make it impossible to develop standardized strategies to identify and systematically and constructively assist different types of repeated users. Indeed, Spittal et al [17] point out specifically that the arbitrary threshold for the average number of contacts within a moving window cannot adequately reveal the heterogeneity of frequent and repeated users. More information should be considered, including the service journey from the beginning to termination of use, the duration of each session, and use patterns, to distinguish different frequent user subgroups and identify how services might tailor their strategies to meet different users' needs based on these patterns.

Taken together, there are several research gaps that necessitate a more nuanced classification and understanding of repeated users of hotline and text-line services: (1) most studies of frequent or repeated users focused on telephone-based services, while the profiling in text-based counseling remains relatively unknown; (2) the definition and operationalization of frequent or repeated users are not well-established, as previous studies tended not to differentiate repeated users from frequent users, and most studies classified frequent users based on an arbitrary number of contacts in a period, which neglects the heterogeneity of frequent users; and (3) other aspects of the profiles of frequent or repeated users, including their contact patterns and downstream service satisfaction, have thus far been largely unexplored. To fill these gaps, we first propose to differentiate repeated users from frequent users. We define *repeated users* as those who make >1 contact with the service. In contrast, the term *frequent users* refers to those with heavy use of or dependent behaviors toward the service. Inherently, *frequent users* is a subcategory of *repeated users*. We adopted the term *repeated users* as the master category to investigate service use because it is more inclusive than *frequent users*. Our objective was to develop a systematic way to identify and classify repeated users of text-based services. This would enable service workers to offer targeted interventions to different types of users and improve the service effectiveness. On the basis of previous literature suggesting that suicidal individuals are more likely to be repeated users [17,18,22,23], we hypothesized that compared with one-time users, repeated users require more contacts and longer session duration because they have higher suicide risk and more complicated presenting problems such as addictive behaviors. In addition, we anticipated that suicide risk and severe presenting issues are associated with a higher frequency and intensity of service use.

Methods

Participants and Data Collection

Data for this study were extracted from a free 24/7 text-based web-based counseling service called Open Up, which is a free service established to provide round-the-clock support to Hong Kong youth who are aged between 11 and 35 years and are experiencing psychological distress, common mental disorders, and suicidality [10,24]. The platform is staffed by professional social workers and trained volunteers. On average, registered social workers at Open Up have around 8 years of working experience, including 1 to 2 years in web-based counseling. Most volunteers are students with a background in psychology, counseling, or social work. The training information has been documented elsewhere [25]. Users can contact the platform 24/7 via different channels, including the service's web portal, WhatsApp, Facebook Messenger, WeChat (WeChat channel was terminated in June 2020 due to security concerns), and SMS text messaging. A user successfully connected to the service is assigned to an available staff. After a brief greeting, the service provider applies different approaches to assist the user, including identifying the main issue of concern, showing empathy and care, providing useful suggestions and helpful resources, and referring to offline services if needed. Toward the end of the session, the service provider may refer the user to a follow-up service (either on the web or offline), if deemed necessary.

Ethical Considerations

This study was conducted under the Education & Knowledge Division of the Open Up project. This study has involved human participants and was approved by the Human Research Ethics Committee at the University of Hong Kong (EA210185; 12/4/2021). Participants provided informed consent to the use of data for research and evaluation purposes before participating by accepting the Terms of Service. Except for the service provided and referrals, participants did not receive any financial incentives. We extracted session text and metadata based on the unique identifier of each user (ie, IP address or social media ID). The unique identifier was first transformed into an 8-digit pseudo-ID before being analyzed. Other private information disclosed during the session was masked to the research team. Finally, this study was conducted with health care professionals of 5 local nongovernmental organizations (The Boys' and Girls' Clubs Association of Hong Kong, Caritas Hong Kong, Hong Kong Children and Youth Services, The Hong Kong Federation of Youth Groups, and St. James' Settlement).

Procedure

In this study, we included session transcripts and the metadata from those whose first contact was between October 2018 and July 2021 (n=29,400). Users' first incoming sessions and sessions during the first year (ie, 365 days) were extracted to investigate their contact patterns. We chose to use a 1-year window for several reasons. First, it allowed us to conduct a standardized comparison between users who contacted the service in different periods. Second, the service that we evaluated began in October 2018, and we conducted this study in August 2022, resulting in a data set spanning around 3 years. Using an inclusion cutoff longer than 1 year would exclude more recent data after 2021, as these users would not have qualified to be enrolled. Conversely, using a shorter period may have resulted in insufficient information and use patterns to be investigated. Third, using a 1-year window allowed us to include a larger number of participants, accounting for approximately 92% of the total users of the service. The session metadata,

XSL•FO
RenderX

session transcripts, and postsession survey results of the users were extracted for analyses.

Figure 1 reports the data included for analysis. A total of 39,859 users used Open Up from October 2018 to June 2022, contributing a total of 112,166 valid sessions (defined by Open Up as those with at least 4 message exchanges between counselors and help seekers). The threshold of 4 message exchanges is based on a previous study that has shown that sessions with <4 message exchanges typically indicate that clients are logging onto the platform without any intention of using the service [25]. Among a portion of the users (n=29,400) whose first contact was before July 2021, which was the cutoff point for the inclusion window at the time of the data analysis, most users (n=26,910, 91.5%) met the inclusion criteria. Those users did not reconnect with the platform only after 1 year or drop out of the service during the chatbot triage process. Consequently, this yielded a sample of one-time (n=17,624, 65.5%) and repeated users (n=9286, 34.5%) over a 1-year period for comparison. The later 9286 repeated users were also included in the clustering analysis.

Figure 1. Data inclusion flow.

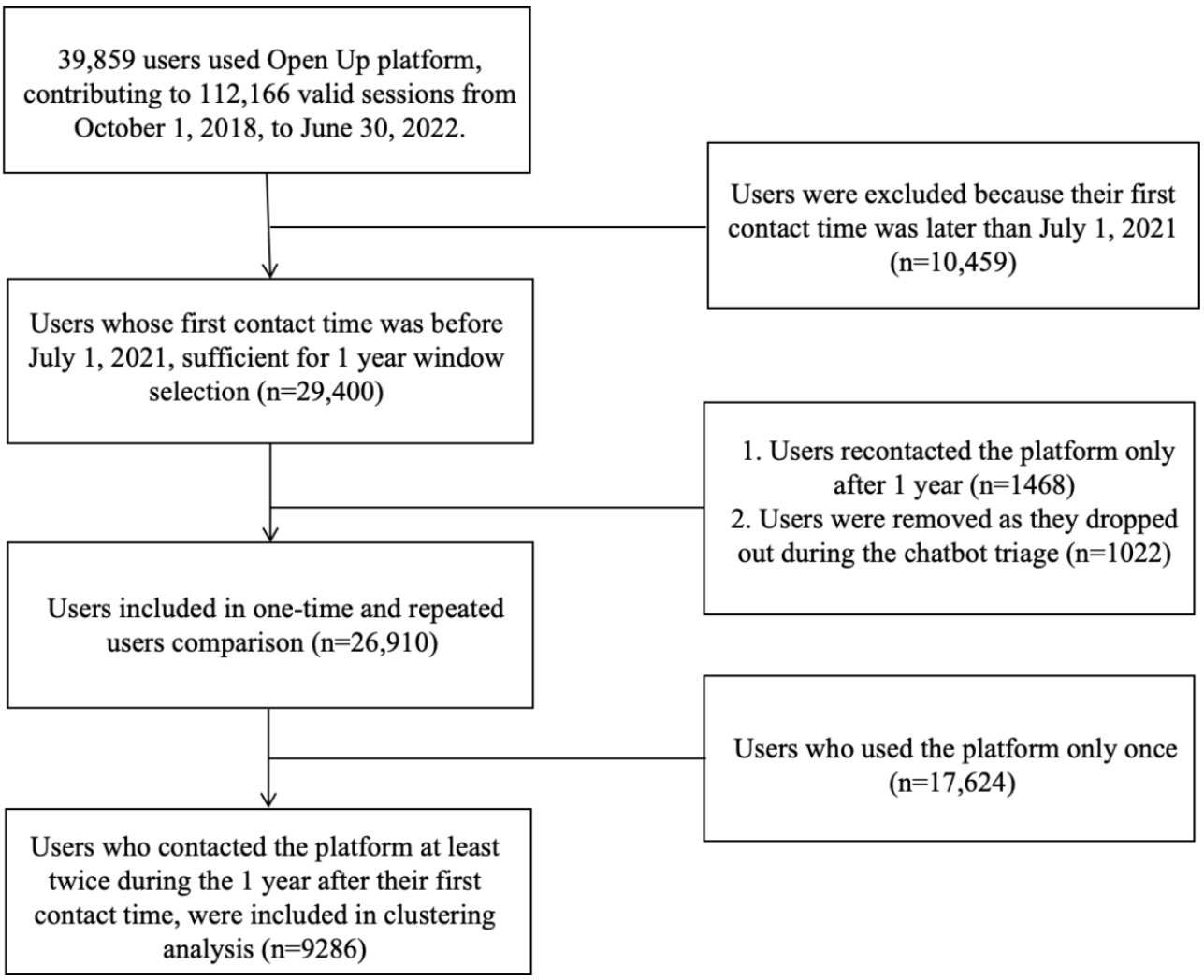

User Identification

Open Up is an anonymous and multichannel supported platform. Therefore, the identification of unique users was based on unique identifiers from their incoming platforms, including the IP address for web portal, phone number for SMS text messaging and WhatsApp, and social media account for Facebook and WeChat. No personal identifying information was extracted.

Operationalization of Repeated Users and One-Time Users

By our proposed definition, users who contacted the platform more than once within 365 days were classified as repeated users, whereas users who contacted Open Up once and did not return within the subsequent 365 days were defined as one-time users (17,624/26910, 65.5%).

Measures

We evaluated users' service use behaviors, user profile, and self-reported satisfaction from the postsession survey.

Service Use Behaviors

Session metadata collected automatically by the system were used to evaluate users' service use behaviors. Session metadata includes the date of their contacts, duration of the session, number of message exchanges between counselors and users, and the platforms used (eg, web portal).

XSL•FO
RenderX

User Profile

User profile includes demographic information, suicide risk, and presenting issues. Users' demographic information including age group and gender were extracted from the counselors' notes if the user disclosed them during the session. Users' suicide risk levels were assessed by using a 7×5 suicide risk assessment scale derived from previous studies [26,27]; to fit the local context, the scale was modified by 2 counseling experts, a professor of social work and a counseling psychologist, and endorsed by the frontline counselors of the service (Multimedia Appendix 1). The suicide risk assessment scale comprises 7 factors, including 5 risk factors and 2 protective factors. The 5 risk factors include suicide ideation, suicide plan, access to suicide means, suicide action or past attempts, and functioning (ie, physical and psychological functioning such as sleep disturbance and diet problems). The 2 protective factors include the level of perceived social support and use of health or psychosocial support services. Risk factors were assessed using a 5-point Likert scale with scores ranging from 0=no risk to 4=severe risk. Protective factors were rated on a similar but reversed scale. An automated 2-layer process, which was implemented to scan through the session content in real time, first picked up keywords and phrases that belonged to any of the 7 factors, then identified their score based on the keywords list of that factor, and concurrently summed the scores from the 7 factors to provide a total score. The accumulative risk score was then transformed to a risk level based on the rubric, including no risk, low risk, medium risk, high risk, and crisis. *No risk* indicates that the texter does not have any suicidal thoughts. *Low risk* suggests that the texter might occasionally think about suicide but does not have any specific plans. *Medium risk* indicates that the texter might have a suicide plan. *High risk* indicates that the texter not only has a suicide plan but also has access to the means to carry it out. *Crisis risk* signifies that the texter might engage in self-injurious and suicidal behaviors. Counselors were also tasked to check and, if needed, manually modify the risk level based on their clinical judgment. Users' presenting issues were detected by the keywords and phrases related to different topics. Open Up has 18 predefined topics (based on the early stage of human exploration and annotation of the data), including mental distress, family issues, and intimate relationships (Multimedia Appendix 2).

User Satisfaction

User's postsession satisfaction was assessed to evaluate the service effectiveness, including three questions from the user postsession survey: (1) "How do you feel after the session compared to before?" (2) "Do you find this service helpful?" (3) "Would you recommend our service to your friend?" For the first 2 questions, a 5-point Likert scale was used to specify the level of user's agreement on the statement, with scores ranging from 1=not helpful at all or much worse to 5=very helpful or much better. For the third question, users responded either "Yes" or "No." Similar questions were used to evaluate the service effectiveness in comparable services, such as Crisis Text Line [28] and National Suicide Prevention Lifeline [29].

Furthermore, we also included another indicator, premature departure, which refers to dropping out from a session without an agreed closure or an indication of the intention to leave. Premature departure was found to be associated with user dissatisfaction [22,30,31]. Premature departure was detected using the method developed by Xu et al [31].

Strategies to Classify Repeated Users

Instead of using an arbitrary number to define frequent users adopted by previous studies, which inevitably suffers from the loss of interpretability across different frequency levels, we aimed to bridge the research gap by incorporating more features that are indicative of a user's contact behaviors. The classification of repeated users was based on each user's 1-year use pattern since their first contact. The clustering analysis used three indicators of each user's service use (Figure 2): (1) the interval period between the first and last contact date, (2) the number of valid contacts that occurred during the first 365 days, and (3) the number of days with at least 1 contact. We investigated users' service use based on these 3 indicators because they not only reflect the overall use of the service but also consider the intensity. By considering the user journey in the platform (feature 1), intensity (feature 2), and frequency (feature 3), we can render a more comprehensive picture of users' help-seeking behavior and classify service users into meaningful groups with distinct profiles and needs.

XSL•FO
RenderX

Figure 2. Illustration of the 3 indicators of users' contacting behaviors.

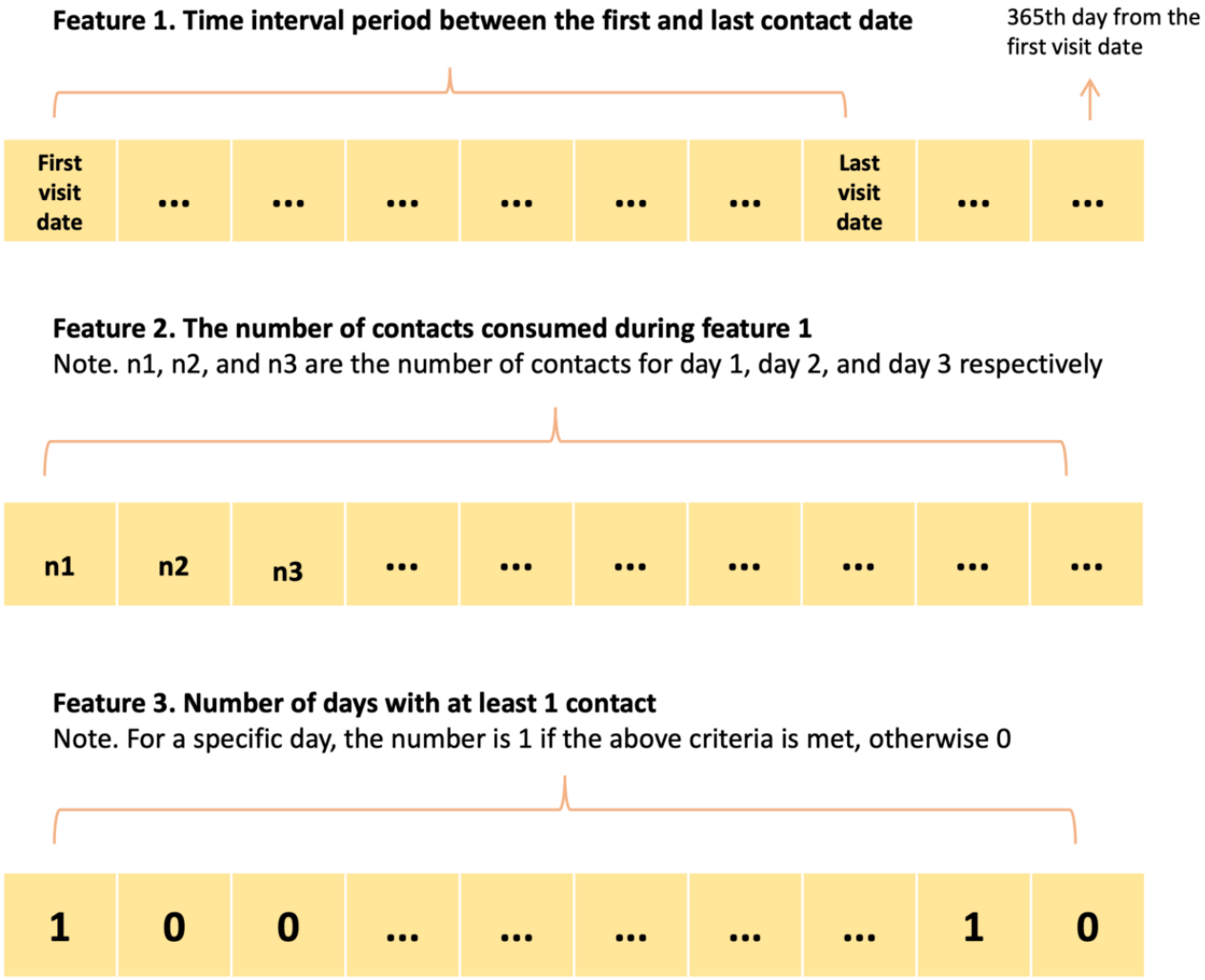

Statistical Analysis

As a starting point, descriptive statistical analyses were performed on one-time and repeated users' first sessions to investigate the reasons why repeated users return and one-time users do not. We decided to use features from their first session for 2 main reasons. First, according to the reviewed studies, users' contacting behaviors and service use patterns are at least partially associated with their background characteristics. Second, including subsequent sessions from repeated users may introduce confounding factors, such as newly emerged presenting issues, which could potentially influence their characteristics. Hence, focusing on the first session helps to eliminate these confounders and ensures a more rigorous analysis. Subgroup analyses were conducted using chi-square tests to identify the difference regarding user profiles and user satisfaction. Mann-Whitney-*U* tests were used to compare user contact patterns, including session duration, number of message exchanges between counselors and users, and users' satisfaction after the session. Cramér V (ϕ_c) was used to represent the effect size of the chi-square test, while common language effect size was applied to describe the effect size of the Mann-Whitney *U* tests.

Then, we used a bottom-up hierarchical clustering method to systematically classify repeated users into different groups and further investigated the differences between clusters of repeated users. Hierarchical clustering is a clustering algorithm that seeks to build nested clusters by merging or splitting them distinctly in a hierarchical fashion. The root of the structure represents the one inclusive cluster for all samples, and every sample of data in the structure is a cluster itself. The subclusters are merged together at intermediate levels to the root node by their linkage metrics [32]. Hierarchical clustering is widely used for classifying documents in a variety of contexts including language models on unstructured text data [33] and structured electronic health records [34,35], with its advantage of producing meaningful hierarchies and interactive visualizations. In this study, we used hierarchical agglomerative clustering, where each data point starts in its own cluster, and pairs of clusters are merged hierarchically based on the linkage between them [36]. Several linkage algorithms can be used in hierarchical cluster analysis to determine how the clusters are merged. In this study, we used the Ward linkage method, which minimizes the sum of squared differences within all groups. Python package sklearn (Python Software Foundation) [37] was used to perform the hierarchical clustering. A dendrogram was used to visualize and decide the optimal number of clusters, which was decided by the largest vertical gap without any horizontal

XSL•FO
RenderX

line passing through the area [38]. The clusters determined by this line are more distinct and separate, as there is more separation space between the clusters. The length of the vertical line represents the distance to be bridged to form the new cluster. Weekly heat maps were used to visualize the use patterns of users from different clusters.

Statistical analyses were performed between different clusters of repeated users regarding user profiles, contact patterns, and user satisfaction. Chi-square test was used to compare the user profiles and satisfaction between clusters, and the Kruskal-Wallis *H* test was used to compare the service use behavior and satisfaction between clusters. Post hoc pairwise comparisons were performed with false discovery rate correction for the chi-square test [39] and with Dunn test for Kruskal-Wallis *H* test [40]. Cramér V (ϕ_c) and epsilon-squared (ε^2) were computed to calculate the effect sizes [41].

Results

Hierarchical Clustering

Figure 3 visualizes the results of hierarchical clustering. The algorithm starts by considering each point as a separate cluster and joining points to clusters hierarchically. Each vertical line represents 1 cluster. The cutoff for selecting the optimal number of clusters was visualized as the horizontal line (in black) in Figure 3, which gives us 3 clusters (blue vertical lines).

In total, 3 clusters of repeated users were detected. Table 1 reports the basic information about them. They were classified as follows: (1) *episodic users*, accounting for 64.63% (6002/9286) of total repeated users and 36.27% (20,549/56,660) of total sessions among repeated users, only sporadically contacted the platform for the first 2 weeks (median 2 [IQR 1.0-3.0] days of contact); (2) *intermittent users*, accounting for 34.61% (3214/9286) of total repeated users and 44.75% (25,356/56,660) of total sessions, had a long interval of service contact (contact interval, mean 233.6, SD 83.6 days); however, they did not contact the platform too many times, and the number of days with contact was low (median 3 [IQR 2.0-5.0] days); and (3) *persistent-intensive users*, comprising 0.75% (70/9286) of total repeated users and 18.99% (10,758/56,660) of the total sessions, came to the platform nearly every week for the whole year after their first contact, and their median days with contacts were 83 (IQR 71.0-103.5). The intensity and frequency of contact behaviors were different between the 3 groups of users. Figures 4-6 visualize the contact pattern of 3 clusters.

Figure 3. Dendrograms of repeated users with hierarchical clustering. The x-axis represents pseudoidentifiers of users. The y-axis represents the distance by Ward linkage between clusters. Each point (ie, pseudoidentifier) starts with its own cluster, and each point joins other points into a new cluster hierarchically based on the Ward linkage distance until forming the ultimate 1 cluster. Each vertical line represents 1 formed cluster. The length of the vertical line represents the distance to be bridged to form the new cluster. The black solid line represents the cutoff of choosing the optimal number of clusters. The orange, green, and red colors represent different clusters in the final results.

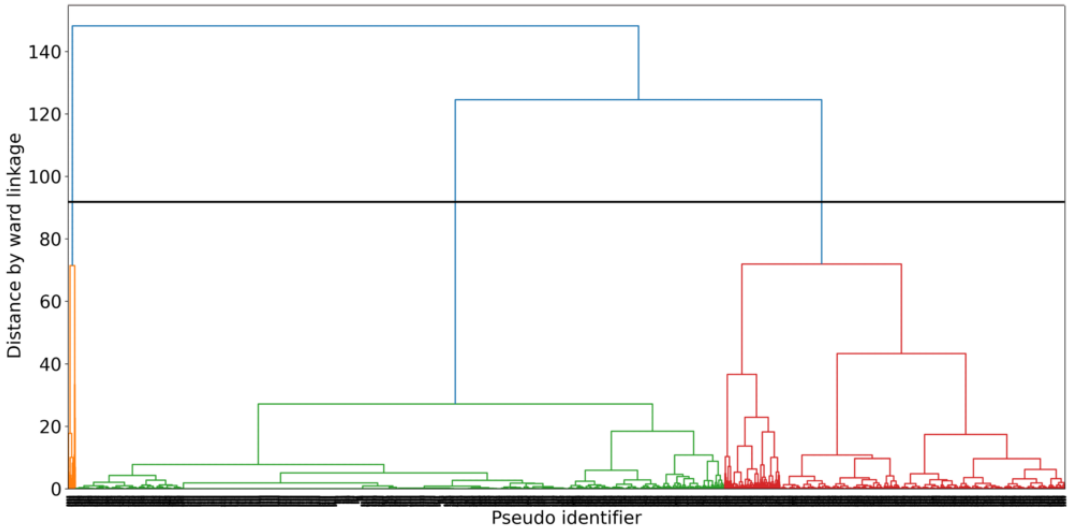

Table 1. Contact patterns of the 3 clusters of repeated users.

Cluster category	Total number of contacts, mean (SD)	Interval (days), mean (SD)	Days with contacts, mean (SD)	Total number of contacts, median (IQR)	Interval (days), median (IQR)	Days with contacts, median (IQR)	Users (n=9286), n (%)	Sessions (n=56,660), n (%)
Episodic users	3.4 (3.1)	18.5 (28.5)	2.4 (2.0)	2.0 (2.0-3.0)	3.0 (0-27.0)	2.0 (1.0-3.0)	6002 (64.63)	20,546 (36.27)
Intermittent users	7.9 (11.9)	233.6 (83.6)	6.2 (8.5)	3.0 (2.0-7.0)	234.0 (164.0-309.0)	3.0 (2.0-5.0)	3214 (34.61)	25,356 (44.75)
Persistent-intensive users	153.7 (114.8)	317.8 (69.1)	94.4 (35.5)	115.0 (97.3-146.5)	354.0 (301.3-362.0)	83.0 (71.0-103.5)	70 (0.75)	10,758 (18.99)

XSL•FO
RenderX

Figure 4. Intensity and contact patterns of episodic users during the 1 year after their first contact (n=6002). The x-axis represents the unique identifier of each user, and the y-axis represents the week index from one's first contact. The number represents the number of contacts in a particular week, and the shades of color represent the intensity of service use. The figure was sorted in ascending order based on the total number of contacts of each user, from left to right. Consequently, the user identifier on the x-axis indicates a progression in service use from left to right. Only the sampled user identifiers are displayed on the x-axis due to the limitation of the figure size. The dense concentration of numbers during the initial week is attributable to the first-week use of all users in this cluster. As time passed, individual variations started to emerge, and some users dropped out, resulting in a decrease in intensity in the following weeks.

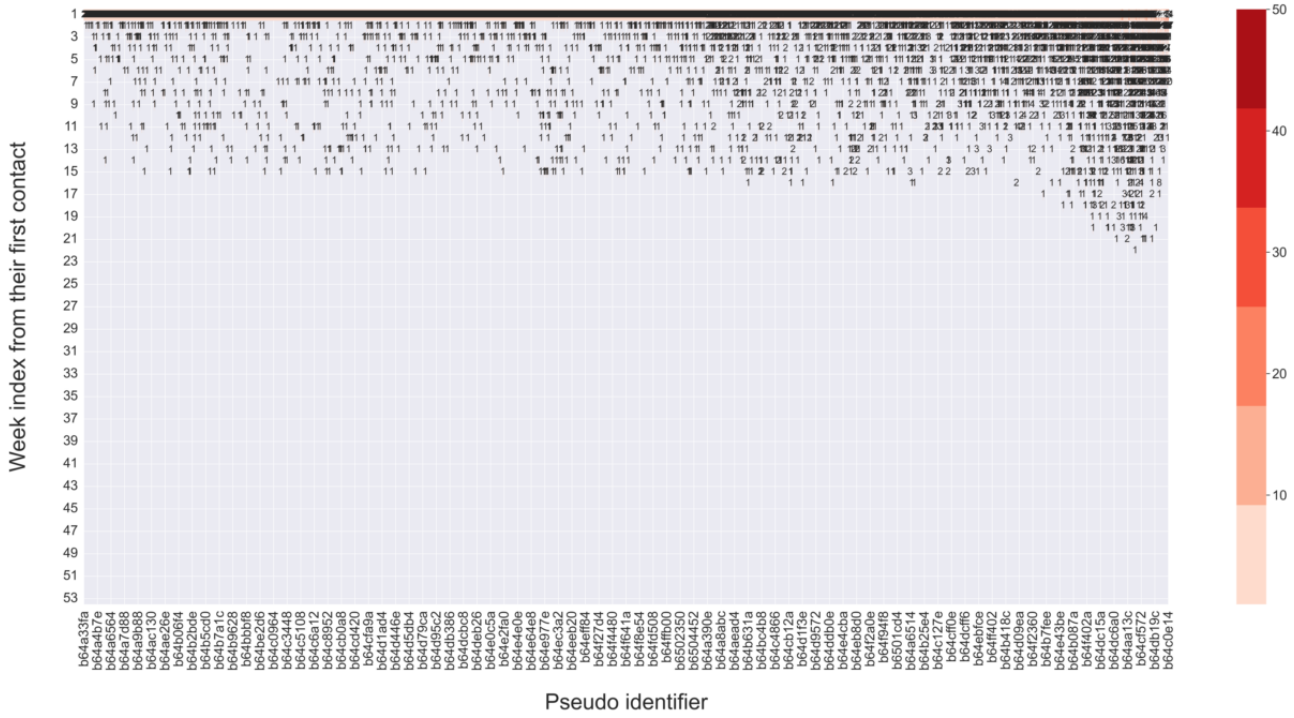

Figure 5. Intensity and contact patterns of intermittent users during the 1 year after their first contact (n=3214). The x-axis represents the unique identifier of each user, and the y-axis represents the week index from one's first contact. The number represents the number of contacts in a particular week, and the shades of color represent the intensity of service use. The figure was sorted in ascending order based on the total number of contacts of each user, from left to right. Consequently, the user identifier on the x-axis indicates a progression in service use from left to right. Only the sampled user identifiers are displayed on the x-axis due to the limitation of the figure size. The dense concentration of numbers during the initial week is attributable to the first-week use of all users in this cluster. As time passed, individual variations started to emerge, and some users dropped out, resulting in a decrease in intensity in the following weeks.

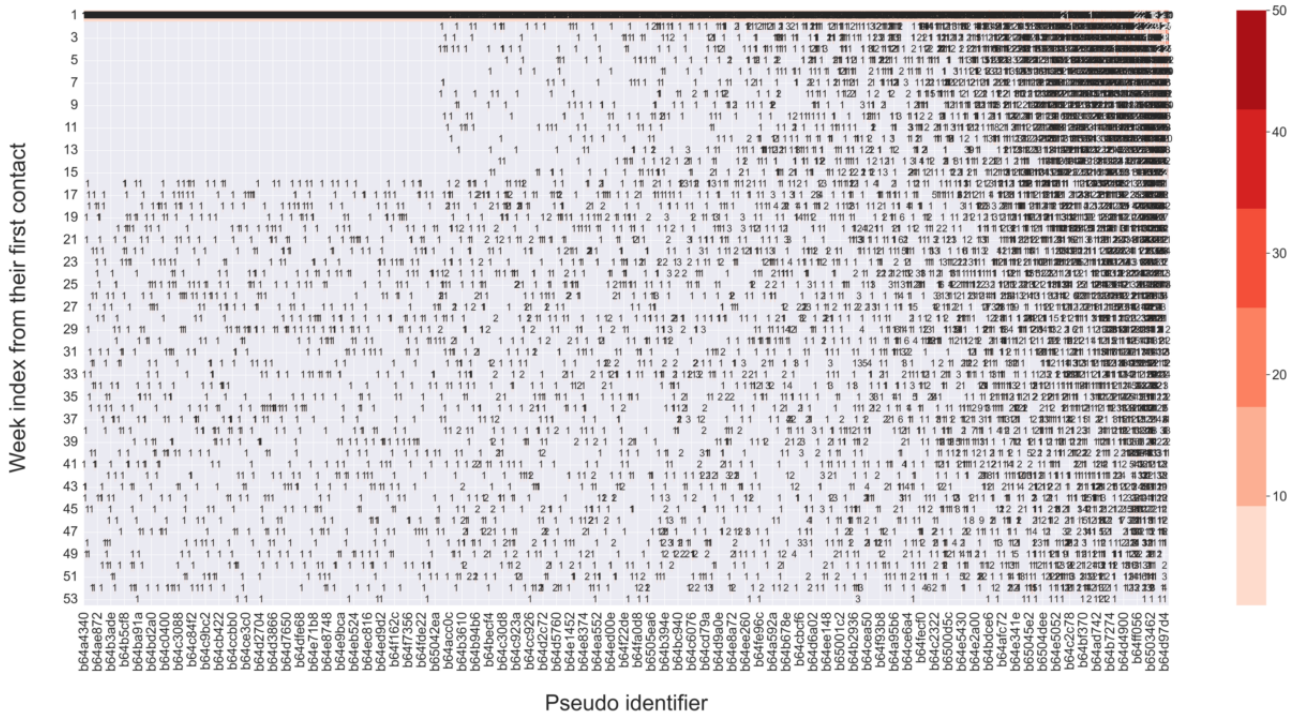

XSL•FO
RenderX

Figure 6. Intensity and contact patterns of persistent-intensive users during the 1 year after their first contact (n=70). The x-axis represents the unique identifier of each user, and the y-axis represents the week index from one's first contact. The number represents the number of contacts in a particular week, and the shades of color represent the intensity of service use. The figure was sorted in ascending order based on the total number of contacts of each user, from left to right. Consequently, the user identifier on the x-axis indicates a progression in service use from left to right. The dense concentration of numbers during the initial week is attributable to the first-week use of all users in this cluster. As time passed, individual variations started to emerge, and some users dropped out, resulting in a decrease in intensity in the following weeks.

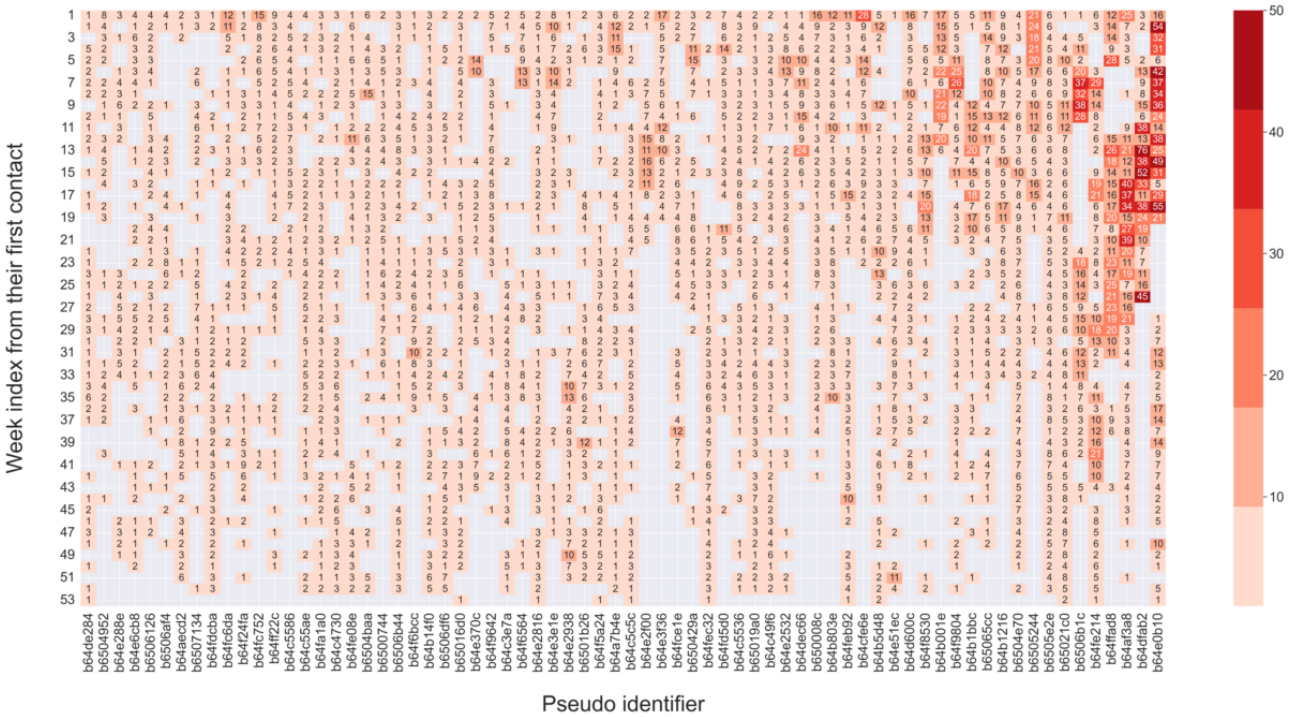

First Session Comparison of One-Time and Repeated Users

Detailed characteristics of users regarding their contact behaviors, profiles, and satisfactions are reported in Table 2. Among the 26,910 users who contacted the platform during the study period, 65.49% (n=17,624) were one-time users, and 34.51% (n=9286) were repeated users. The session use results revealed that the average session duration of the first session was longer among repeated users than among one-time users (cl=0.509; *P*=.02). The number of messages exchanged with counselors was also higher among repeated users than one-time users (cl=0.516; *P*<.001). 64.01% (17,224/26,910) of users revealed their age and gender. There were more female than male users in general and in both the one-time and repeated user groups. Nonstudent youth and middle school students (usually aged 12 to 18 years) and below were more represented than other categories in both one-time and repeated user groups. The web portal was the dominant channel for both one-time and repeated users. However, repeated users used more social media channels than one-time users (ϕc=0.046; *P*<.001). The risk profile comparison indicated that repeated users showed a higher percentage of high suicide risk and crisis suicide risk sessions than one-time users (ϕc=0.040; *P*<.001). The results regarding presenting issues revealed that repeated users had more suicide- or self-harm–related concerns, as well as those concerning interpersonal relationships, medical issues, bullying, and addictive behaviors (eg, substance abuse and gambling), than one-time users (ϕc=0.029, ϕc=0.023, ϕc=0.027, ϕc=0.029, and ϕc=0.031, respectively; all *P*<.001). By contrast, one-time users had more concerns related to family issues (ϕc=0.026; *P*<.001), intimate relationships (ϕc=0.017; *P*=.006), and trending topics (eg, social unrest and COVID-19; ϕc=0.018; *P*=.004).

As a comparison of service satisfaction, repeated users had a higher rate of premature departure than one-time users (ϕc=0.065; *P*<.001). The completion rate of the postsession survey was similar across the 2 groups (2965/17,624, 16.82%, among one-time users vs 1960/9286, 14.65%, among repeated users). Among those who filled out the survey, repeated users gave more positive feedback compared with one-time users in all 3 satisfaction questions, including feeling better after the session (cl=0.532; *P*<.001), finding the service more helpful (cl=0.540; *P*<.001), and more likely to recommend the service to others (ϕc=0.068; *P*<.001).

XSL•FO
RenderX

Table 2. Comparing the contact behaviors, profiles, and satisfaction between one-time and repeated users on their first contact (n=26,910).

Variables	One-time users (n=17,624, 65.49%)	Repeated users (n=9286, 34.51%)	Chi-square (*df*)/*U*	*P* value	ϕc/cl
Session duration (minutes)[a]			83,222,903	.02	0.509
Values, mean (SD)	55 (38)	57 (43)			
Values, median (IQR)	47 (27-74)	49 (26-78)			
Message exchanges			84,442,795	<.001	0.516
Values, mean (SD)	38 (29)	42 (34)			
Median (IQR)	32 (15-53)	33 (15-57)			
Sex[b], n (%)			98.675 (2)	<.001	0.061
Male	3543 (20.10)	1615 (17.39)			
Female	8102 (45.97)	3964 (42.69)			
Unknown	5979 (33.93)	3707 (39.92)			
Age group[b], n (%)			97.174 (4)	<.001	0.060
Middle school students and below	2618 (14.85)	1405 (15.13)			
University students	2099 (11.91)	914 (9.84)			
Nonstudent youth	4766 (27.04)	2266 (24.40)			
Middle aged	1001 (5.68)	426 (4.59)			
Unknown	7140 (40.51)	4275 (46.04)			
Platform, n (%)			56.762 (4)	<.001	0.046
Web portal	15166 (86.05)	7691 (82.82)			
WhatsApp	1314 (7.46)	886 (9.54)			
Facebook	431 (2.45)	295 (3.18)			
WeChat[c]	554 (3.14)	305 (3.28)			
Mobile phone	159 (0.90)	109 (1.17)			
Premature departure, n (%)	7656 (43.44)	4790 (51.58)	161.861 (1)	<.001	0.078
Risk level[d], n (%)			43.646 (4)	<.001	0.040
Crisis risk	351 (1.99)	275 (2.96)			
High risk	731 (4.15)	471 (5.07)			
Medium risk	6518 (36.98)	3354 (36.12)			
Low risk	5992 (34.00)	3002 (32.33)			
No risk	4032 (22.88)	2184 (23.52)			
Presenting issues[e], n (%)					
Mental disorders or emotional problems	9129 (51.80)	4708 (50.70)	2.911 (1)	.09	0.010
Family issue	5660 (32.12)	2744 (29.56)	18.543 (1)	<.001	0.026
Intimate relationships	5173 (29.35)	2577 (27.75)	7.536 (1)	.006	0.017
Suicide or self-harm	3972 (22.54)	2332 (25.11)	22.324 (1)	<.001	0.029
Interpersonal relationships	3789 (21.50)	2185 (23.53)	14.389 (1)	<.001	0.023
Study	3601 (20.43)	1926 (20.74)	0.333 (1)	.56	0.004
Career prospects or unemployment	2582 (14.65)	1358 (14.62)	0.002 (1)	.97	0.000
Medical issue	2235 (12.68)	1355 (14.59)	19.016 (1)	<.001	0.027
Work pressure	2234 (12.68)	1124 (12.10)	1.772 (1)	.18	0.008

XSL•FO
RenderX

Variables	One-time users (n=17,624, 65.49%)	Repeated users (n=9286, 34.51%)	Chi-square (*df*)/*U*	*P* value	ϕc/cl
Economic issue	1905 (10.81)	1022 (11.01)	0.221 (1)	.64	0.003
Trending topics (eg, social unrest and COVID-19)	1525 (8.65)	708 (7.62)	8.329 (1)	.004	0.018
Abuse or sexual assault	1175 (6.67)	676 (7.28)	3.465 (1)	.06	0.011
Workplace relationships	1221 (6.93)	654 (7.04)	0.106 (1)	.75	0.002
Public examinations	917 (5.20)	558 (6.01)	7.464 (1)	.006	0.017
Bullying	780 (4.43)	534 (5.75)	22.688 (1)	<.001	0.029
Addictive behaviors (eg, substance abuse and gambling)	610 (3.46)	438 (4.72)	25.273 (1)	<.001	0.031
Debt	397 (2.25)	199 (2.14)	0.289 (1)	.59	0.003
Sexual orientation or gender distress	364 (2.07)	172 (1.85)	1.309 (1)	.25	0.007
Postsession outcomes[f]					
Filling rate, n (%)	2965 (16.82)	1360 (14.65)	1.067 (1)	.30	0.006
How do you feel after the session vs before?[g] mean (SD)	3.91 (0.86)	4.01 (0.81)	2,106,417	<.001	0.532
Do you find this service helpful?[h] mean (SD)	3.44 (1.08)	3.60 (1.03)	2,081,828	<.001	0.540
Would you recommend our service to a friend in need?[i] n (%)	1798 (62.98)	880 (68.16)	19.715 (2)	<.001	0.068

[a]Counselor attend time to last message time.

[b]Extracted from counselor postsession survey.

[c]The use of WeChat was terminated on June 20, 2020.

[d]Suicide risks were scanned by the risk assessment scale.

[e]Presenting issues were detected by keywords and phrases.

[f]n=4325; 2965 versus 1360 for one-time users and repeated users, respectively.

[g]n=4286; 2939 versus 1347 for one-time users and repeated users, respectively (the score ranges from 1 to 5).

[h]n=4235; 2909 versus 1326 for one-time users and repeated users, respectively (the score ranges from 1 to 5).

[i]n=4146; 2855 versus 1291 for one-time users and repeated users, respectively (binary choice).

First Session Comparison Among 3 Clusters of Repeated Users

A further comparison between clusters of repeated users is tabulated in Table 3. As users' contact frequency and intensity increased, the average session duration (ε^2=0.008; *P*<.001) and messages exchanged also increased (ε^2=0.005; *P*<.001). The percentage of females in intermittent users was significantly higher than in the episodic users (ϕc=0.039; *P*<.001). Intermittent users also consisted of a different distribution of age groups compared with episodic users, where intermittent users had a higher percentage of middle school students and below, university students, and nonstudent youth (ϕc=0.038; *P*<.001). As for platform preference, users with more frequent and intensive contacts tended to choose social media platforms rather than the web portal (ϕc=0.099; *P*<.001). Intermittent users show a lower percentage of premature departure than episodic users (ϕc=0.073; *P*<.001). The profile comparisons show that users with more frequent and intensive contact behaviors have a higher percentage of high suicide risk and crisis suicide risk sessions than episodic users and intermittent users (ϕc=0.036; *P*<.001). Contact frequency and intensity were also found to be positively associated with the percentage of users mentioning mental disorders or emotional problems (ϕc=0.044; *P*<.001). Intermittent users also had a higher percentage of economic issues (ϕc=0.026; *P*=.048) and study issues (ϕc=0.028; *P*=.03) and a lower percentage of intimate relationships than episodic users (ϕc=0.036; *P*=.002). Comparisons in postsession survey questions show no statistical difference between groups of repeated users, potentially owing to the low response rate of the postsession survey.

XSL•FO
RenderX

Table 3. Contact behaviors, profiles, and satisfaction between clusters of repeated users[a] (n=9286).

Variables	Episodic users (n=6002, 64.63%)	Intermittent users (n=3214, 34.61%)	Persistent-intensive users (n=70, 0.75%)	Chi-square (*df*)/Kruskal-Wallis *H* statistics	*P* value	Post hoc or Dunn test	$\phi c/\varepsilon^2$
Session duration (minutes)[b]				71.220 (2)	<.001	1<2<3	0.008
Values, mean (SD)	55 (43)	60 (42)	81 (49)				
Values, median (IQR)	46 (24-76)	53 (30-81)	73 (48-111)				
Message exchanges				47.210 (2)	<.001	1<2<3	0.005
Values, mean (SD)	40 (34)	43 (34)	58 (43)				
Values, median (IQR)	31 (15-56)	37 (17-59)	48 (32-77)				
Sex[c], n (%)				28.228 (4)	<.001	1≠2	0.039
Male	1040 (17.33)	564 (17.55)	11 (16)				
Female	2459 (40.97)	1466 (45.61)	39 (56)				
Unknown	2503 (41.70)	1184 (36.84)	20 (29)				
Age[c], n (%)				27.199 (8)	<.001	1≠2	0.038
Middle school students and below	887 (14.61)	519 (16.15)	9 (13)				
University students	559 (9.31)	348 (10.83)	7 (10)				
Nonstudent youth	1414 (23.56)	834 (25.95)	18 (26)				
Middle aged	274 (4.57)	149 (4.64)	3 (4)				
Unknown	2878 (47.95)	1364 (42.44)	33 (47)				
Platforms, n (%)				181.530 (8)	<.001	1≠2≠3	0.099
Web portal	5127 (85.42)	2537 (78.94)	27 (39)				
WhatsApp	491 (8.18)	378 (11.76)	17 (24)				
Facebook	161 (2.68)	125 (3.89)	9 (13)				
WeChat[d]	156 (2.60)	136 (4.23)	13 (19)				
Mobile phone	67 (1.12)	38 (1.18)	4 (6)				
Premature departure, n (%)	3257 (54.27)	1504 (46.80)	29 (41)	49.678 (2)	<.001	1≠2	0.073
Risk level[e], n (%)				23.483 (8)	.003	1≠2≠3	0.036
Crisis risk	166 (2.77)	106 (3.30)	3 (4)				
High risk	291 (4.85)	171 (5.32)	9 (13)				
Medium risk	2115 (35.24)	1210 (37.65)	29 (41)				
Low risk	1975 (32.91)	1008 (31.36)	19 (27)				
No risk	1455 (24.24)	719 (22.37)	10 (14)				
Presenting issues[f], n (%)							
Mental disorders or emotional problems	2972 (49.52)	1687 (52.49)	49 (70)	17.909 (2)	<.001	1≠2≠3	0.044
Family issue	1775 (29.57)	946 (29.43)	23 (33)	0.390 (2)	.82	—[g]	0.006
Intimate relationships (lovers)	1737 (28.94)	824 (25.64)	16 (23)	12.229 (2)	.002	1≠2	0.036
Suicide or self-harm	1472 (24.53)	837 (26.04)	23 (33)	4.811 (2)	.09	—	0.023
Interpersonal relationships	1393 (23.21)	765 (23.80)	27 (39)	9.278 (2)	.01	1=2≠3	0.032
Study	1195 (19.91)	714 (22.22)	17 (24)	7.306 (2)	.03	1≠2	0.028

XSL•FO
RenderX

Variables	Episodic users (n=6002, 64.63%)	Intermittent users (n=3214, 34.61%)	Persistent-intensive users (n=70, 0.75%)	Chi-square (*df*)/Kruskal-Wallis *H* statistics	*P* value	Post hoc or Dunn test	$\phi c/\varepsilon^2$
Career prospects or unemployment	861 (14.35)	485 (15.09)	12 (17)	1.289 (2)	.53	—	0.012
Medical issue	850 (14.16)	487 (15.15)	18 (26)	8.649 (2)	.01	1=2≠3	0.031
Work pressure	694 (11.56)	422 (13.13)	8 (11)	4.863 (2)	.09	—	0.023
Economic issue	626 (10.43)	389 (12.10)	7 (10)	6.057 (2)	.048	1≠2	0.026
Trending topics (eg, social unrest and COVID-19)	445 (7.41)	258 (8.03)	5 (7)	1.141 (2)	.57	—	0.011
Abuse or sexual assault	432 (7.20)	238 (7.41)	6 (9)	0.308 (2)	.86	—	0.006
Workplace relationships	417 (6.95)	233 (7.25)	4 (6)	0.481 (2)	.79	—	0.007
Public examinations	357 (5.95)	196 (6.10)	5 (7)	0.244 (2)	.89	—	0.005
Bullying	328 (5.46)	202 (6.29)	4 (6)	2.598 (2)	.27	—	0.017
Addictive behaviors (eg, substance abuse and gambling)	264 (4.40)	167 (5.20)	7 (10)	7.342 (2)	.03	1=2=3	0.028
Debt	116 (1.93)	82 (2.55)	1 (1)	3.992 (2)	.14	—	0.021
Sexual orientation or gender distress	103 (1.72)	69 (2.15)	0 (0)	3.468 (2)	.18	—	0.019
Postsession outcomes[h]							
Filling rate, n (%)	13.7	16.5	10	13.701 (2)	.001	1≠2	0.038
How do you feel after the session vs before?[i], mean (SD)	3.98 (0.80)	4.06 (0.83)	4.29 (0.49)	6.322 (2)	.04	1=2=3	0.005
Do you find this service helpful?[j], mean (SD)	3.58 (1.03)	3.63 (1.04)	3.86 (1.22)	1.423 (2)	.49	—	0.001
Would you recommend our service to a friend in need?[k], n (%)	70.9	71.5	71.4	5.443 (2)	.25	—	0.046

[a]Repeated users were detected by unique identifiers, including IP address for web portal, account number for social media channels, and phone number for SMS text messaging.

[b]Counselor attend time to last message time.

[c]Extracted from counselor postsession survey.

[d]The use of WeChat was terminated on June 20, 2020.

[e]Suicide risks were scanned by the risk assessment scale.

[f]Topics were detected by keywords and phrases.

[g]Not applicable.

[h]n=1360; 824 versus 529 versus 7 for episodic, intermittent, and persistent-intensive users, respectively.

[i]n=1347; 819 versus 521 versus 7 for episodic, intermittent, and persistent-intensive users, respectively (the score ranges from 1 to 5).

[j]n=1326; 803 versus 516 versus 7 for episodic, intermittent, and persistent-intensive users, respectively (the score ranges from 1 to 5).

[k]n=1291; 782 versus 502 versus 7 for episodic, intermittent, and persistent-intensive users, respectively (binary choice).

Discussion

Implications for Web-Based Counseling Services

In this study, we analyzed the service use and postsession satisfaction of users of a text-based web-based counseling service to identify the potential characteristics of different clusters of users. By examining 3 key indicators of users' service use, including the interval between the first and the last contacts, the total number of contacts, and the number of days with at least 1 contact, we identified 3 groups of repeated users: episodic users, intermittent users, and persistent-intensive users. Generally, compared to one-time users, repeated users showed higher suicide risks and more complicated backgrounds, including more severe presenting issues such as suicide or self-harm, bullying, and addictive behaviors. Higher frequency and intensity of service usage were also associated with elevated suicide risk levels and a higher proportion of users citing mental

XSL•FO
RenderX

disorders as their primary concerns. The findings are potentially beneficial for developing efficient interventions tailored to the specific needs of different repeated user groups and improving service effectiveness.

Data from Open Up showed that 65.5% (17,624/26,910) of users were one-time users, and 34.5% (9286/26,910) had used the service more than once (ie, repeated users). The ratio is similar to other text-based services such as Crisis Text Line (63.8% and 36.2%, respectively) in the United States [42]. Contrary to telephone helplines, where the percentage of male individuals was higher for repeated users than for one-time users, we did not find a significant gender difference between one-time and repeated users [14,15]. Comparing data from their first sessions provides insights into why repeated users tend to recontact the service. Repeated users displayed higher suicide risk levels and disclosed more topics of concern related to suicide or self-harm, interpersonal relationships, medical issues, bullying, and addictive behaviors than one-time users in their first session, implying that they generally had more complicated needs that may require more attention and support, which likely and understandably drove them to use the service again. In addition, repeated users had a more positive first-session experience than one-time users. The fact that they were more satisfied with the service might also motivate them to return. The satisfaction rate of one-time users was on average lower than that of repeated users; some users of the former group did not find the service useful, and perhaps, as a result, these users did not seek help again from this service. This may imply that there is a degree of heterogeneity within the one-time user group. Future research could explore qualitatively and quantitatively the differences among different concerns, such as how counselors handle different presenting issues, user-counselor dynamics, user satisfaction, and rate of premature departure.

In total, 3 clusters of repeated users with unique characteristics were identified. The episodic users display the least service use. Their sessions were shorter in duration and lower in number of exchanges. Episodic users had the lowest suicide risk among the 3 subgroups of repeated users. Their primary concerns tended to be related to intimate relationships. While it might be the case that these users were able to adequately address or resolve those issues in the few sessions, the postsession results show that their satisfaction with the first contact was lower than the other 2 groups of repeated users, which indicates another possibility that they subsequently did not find the service to be helpful, that the sessions did not help them make progress, or that they did not match their expectations. Intermittent users used the service over long intervals but in relatively low intensity. Their risk level was higher than that of the episodic users but lower than that of the persistent-intensive users. The motivation for their use of Open Up could be periodical stressors, such as study, work, and economic stress; they would use the service as needed. Persistent-intensive users are likely the ones that impose the greatest burden on frontline service providers and the system. This group, while small in size (70/9286, 0.75%), accounted for 18.99% (10,758/56,660) of service capacity among repeated users and 14.48% (10,758/74,284) of the total service capacity. Similar to previous studies on telephone helplines, persistent-intensive users tended to present more mental health concerns [13,17]; higher suicide risk [17]; and, perhaps as a result, evidently had longer session time and more contacts to meet their needs. Persistent-intensive users had more concerns about medical issues and addictive behaviors (eg, substance abuse and gambling), which tend to be more chronic and difficult to address. This might lead them to return to the service frequently and intensively. In addition, persistent-intensive users were more likely to experience interpersonal relationship problems, which is found to be one of the main issues that drive suicide ideation [43]. Their higher suicide risk, complicated presenting issues, and exposure to long-term stressors could jointly drive their high frequency and intensity of service use. Preference for different service platforms also helps explain the service use of persistent-intensive users. They tended not to prefer using the anonymous web portal; rather, they generally accessed the service using social media platforms or SMS text messaging services that revealed their phone numbers. It is possible that they were in greater need and looking for a more long-term counseling service instead of a one-off session, and thus, they did not mind disclosing their social media accounts or phone numbers.

From the user postsession satisfaction surveys, there was an observable trend and pattern in the mean scores across the 3 subgroups of repeated users, which may explain why persistent-intensive users (ie, the most satisfied group) returned most frequently. However, due to the low response rate (ie, around 11%), the results were not statistically different. The demographical differences across the 3 categories suggest that there were individual differences that may help explain users' use of the service. However, the effect sizes of the characteristics were small, and thus, their implications should be considered with caution. Simultaneously, the lack of strict time limits and the idiosyncratic response patterns between different counselors may have intensified the dependent behavior of some users. Text-based network analysis or in-depth qualitative analysis is needed to examine whether there are structural or systematic differences between the ways in which the subgroups were served. Some standardization of service protocols could be considered to improve service effectiveness.

Service capacity and resource allocation are highly influenced by repeated users. The average session duration for persistent-intensive users was >1 hour (mean 81, SD 49 minutes), which was around 50% longer than that of one-time users (mean 55, SD 38 minutes). The capacity taken up by persistent-intensive users should be considered in the service capacity management plan to strategically improve service efficiency and limit users' unlimited use. Several suggestions could be considered for persistent-intensive users. Since they prefer social media channels more than the anonymous web portals, service providers could arrange scheduled sessions with the same counselors (eg, when the incoming traffic is less busy), minimize their random access to the system (ie, to change the reinforcement schedule from random to fixed), collectively set specific goals per session, and track progress. This will also reduce the need for users to repeat their presenting issues every time they access the service and thus improve service efficiency. We also found that persistent-intensive users displayed higher

XSL•FO
RenderX

risk and experienced more long-term stressors than intermittent and episodic users. Therefore, when serving persistent-intensive users, service providers could consider allocating them to more experienced counselors with skills that can cater to the specific issues of concerns. Regular case studies and workshops should also be conducted to equip service workers and update their knowledge and skills. In addition, it was observed that persistent-intensive users displayed a higher proportion of mental disorders or emotional problems, interpersonal problems, substance abuse, and medical problems. While this study did not identify the comorbidity profile of each individual, that is, whether the same user had experienced mental disorders, interpersonal problems, substance abuse, and medical problems or a different combination of these issues, we speculate that some of them might have some degree of personality dysfunctionality [44]. This suggests that intervention strategies for personality disorders, such as modules in dialectical behavior therapy, schema-focused therapy, or mentalization-based treatment, could potentially be helpful to those clients [45]. However, whether these strategies can be effectively delivered in a web- and text-based context such as the current one warrants further investigation [46]. Alternatively, when persistent-intensive users are willing to identify themselves, counselors can refer them to longer-term counseling services or offline services that might better suit their needs. They can also set more structured and standardized time limits. This could potentially be one of the optimal intervention solutions because by doing so, they could prioritize other users that can benefit more from their low-intensity service, as well as reduce persistent-intensive users' overdependence on the platform without denying or excluding them from service.

At the time of writing in November 2023, ChatGPT and other natural language processing models have attracted growing interest in the field of health care [47] and mental health [48]. There are at least several possible ways that artificial intelligence (AI) models can enhance the efficacy of service workers and support their service provision to help repeated users. In the presession stage, AI models can help triage users by detecting their mental health states and help-seeking issues based on sentiments and linguistic patterns. By doing so, AI models can help prioritize users and provide appropriate support. During the session, AI models can collect and analyze users' information in real time, assisting in the discovery of effective treatment and intervention strategies. This real-time analysis can greatly aid service workers in providing timely and personalized care. After the session, AI models can review the complete session history of repeat users, aiding service workers in developing a comprehensive understanding of these users' needs and progress. However, before applying AI models in the service, potential issues including data privacy and security, ethical consideration, integration, and supervision should be considered.

This study has several methodological contributions adding to the existing evidence. First, we used features including total number of contacts in a year, number of days with contacts, and time length between the first contact and the last contact to differentiate repeated users. We see this as an advancement from the prior practice of using an arbitrary number of contacts over a period to categorize frequent users. This approach also allows each service to define its own threshold based on actual use, as services can differ greatly in terms of their targeted clientele, nature, service protocol, and available resources. Second, previous studies used a top-down definition to identify frequent users, which varied across studies and thus made comparisons across studies difficult. To solve this issue, we proposed a bottom-up, data-driven method to classify users, supplementing the current definitions and providing a more objective perspective on understanding frequent and repeated users. Third, our method of classifying repeated users is transparent and straightforward. This can help improve the reproducibility and generalizability of our results and thus aid its application to other counseling or health services. Most importantly, the proposed method successfully identified 3 clusters with distinct service use patterns. This approach can benefit related services in clustering their service users, providing support for their intervention strategy for different types of users, and optimizing their resource allocation. Further in-depth investigations and interventions on different groups of repeated users are needed to identify the best practice for handling different types of frequent users.

Limitations

This study has several limitations. First, repeated users were identified based on their unique identifiers, which has limitations. IP addresses are not fixed even for the same location. Furthermore, a user may connect to the service from home, school, or work, resulting in multiple IP addresses for the same user. Moreover, the same IP address can be shared by multiple users, for example, an IP address from a school or a cafe. The same user can also create several social media accounts or have different phone numbers, making it impossible to distinguish the user without analyzing the session content. The current analytical practice respects and maintains the privacy and anonymity of the service users. To reduce the noise of data while maintaining user anonymity, future studies could explore the use of digital footprints to match repeated users' sessions for a fairer comparison.

Second, we did not investigate the long-term behavioral pathways for repeated users. Instead, we focused our analyses on the traits of their first contact and the factors that could predict their return. Future studies can use users' longitudinal sessions to investigate the changes in dependent behaviors.

Third, while the proposed methods provide a holistic view of repeated users that can facilitate frontline workers for better service provision, it is important to recognize that people who share similar psychological profiles and use patterns may not experience mental illness in the same way. The proposed method cannot provide a one-size-fits-all solution. Therefore, tailored assessment and treatment plans are necessary on a case-by-case basis. Future studies should not neglect the heterogeneity of users from the same cluster with varying experiences of mental illness.

Fourth, due to the low response rate of the postsession survey (4325/26,910, 16.07%), the analysis of between-group differences in satisfaction might be biased and may limit the generalizability of this study's findings. Low response rates to

XSL•FO
RenderX

postservice surveys are not uncommon. For example, Crisis Text Line reported that only 23% of all text conversations were followed by a completed postconversation survey [42], and 9% was reported in another web chat service developed for youth mental health [49]. Also, studies have found that those who opted to fill out a postservice survey tended to be also more satisfied with the service [50,51]. This adds biases to the results. Future studies can seek to improve the response rate of postsession surveys and, as an alternative approach, apply text mining to identify relevant factors to indicate user satisfaction.

Conclusions

Adequately serving repeated users is one of the most challenging aspects of web-based counseling and crisis intervention services. In this study, we used a bottom-up method based on features of service use to identify the heterogeneity of repeated users. Three groups of repeated users with distinct patterns of service use and risk profiles were identified. The findings can be meaningful to the frontline services in better allocating resources and building service model to improve efficiency and effectiveness. More in-depth analyses of one-time users and repeated users are needed, especially to develop best practices in handling most persistent-intensive users as well as to understand the different reasons why some users do not return.

Acknowledgments

This research was funded by the Hong Kong Jockey Club Charities Trust (AR170018) with the partnership of 5 nongovernmental organizations (The Boys' and Girls' Clubs Association of Hong Kong, Caritas Hong Kong, Hong Kong Children and Youth Services, The Hong Kong Federation of Youth Groups, and St James' Settlement). The strategic theme-based research was supported by the University of Hong Kong, a Collaborative Research Fund grant (C7151-20G), and the Humanities and Social Science Prestigious Fellowship (37000320). We would also like to express our sincere appreciation and gratitude to the reviewers and Professor Eric Caine for their invaluable contribution in enhancing the quality of this study.

Authors' Contributions

YX and CSC conceived the study. YX, PSFY, and CSC constructed the general outline. YX developed the model and performed the statistical analysis. YX and JC wrote the first draft of the manuscript, which was critically reviewed by ZX, CSC, and PSFY. FC and JL evaluated the chat histories. All authors contributed to the analysis. All authors interpreted the data, read the manuscript, and provided final approval for publication submission. CSC (shaunlyn@hku.hk) and PSFY (sfpyip@hku.hk) are co-corresponding authors for this study.

Conflicts of Interest

CSC received consultation fees from Open Up. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Risk assessment scale.
[DOCX File , 48 KB - jmir_v26i1e50976_app1.docx]

Multimedia Appendix 2
Presenting issues detected in Open Up.
[DOCX File , 10 KB - jmir_v26i1e50976_app2.docx]

References

1. Social return on investment forecast of the lifeline online crisis support chat service. netbalance. 2014. URL: https://www.lifeline.org.au/media/esypbhml/social-return-investment-forecast-2014.pdf [accessed 2024-04-29]
2. Fang L, Tarshis S, McInroy L, Mishna F. Undergraduate student experiences with text-based online counselling. Brit J Soc Work 2018;48(6):1774-1790. [doi: 10.1093/bjsw/bcx111]
3. Gatti FM, Brivio E, Calciano S. "Hello! I know you help people here, right?": A qualitative study of young people's acted motivations in text-based counseling. Child Youth Serv Rev 2016 Dec;71:27-35. [doi: 10.1016/j.childyouth.2016.10.029]
4. Reese RJ, Conoley CW, Brossart DF. The attractiveness of telephone counseling: an empirical investigation of client perceptions. J Couns Dev 2006;84(1):54-60. [doi: 10.1002/j.1556-6678.2006.tb00379.x] [Medline: 5637250]
5. Berrouiguet S, Baca-García E, Brandt S, Walter M, Courtet P. Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health. J Med Internet Res 2016 Jun 10;18(6):e135 [FREE Full text] [doi: 10.2196/jmir.5066] [Medline: 27287668]
6. Dwyer A, de Almeida Neto A, Estival D, Li W, Lam-Cassettari C, Antoniou M. Suitability of text-based communications for the delivery of psychological therapeutic services to rural and remote communities: scoping review. JMIR Ment Health 2021 Feb 24;8(2):e19478 [FREE Full text] [doi: 10.2196/19478] [Medline: 33625373]

XSL•FO
RenderX

7. Kalafat J, Gould MS, Munfakh JL, Kleinman M. An evaluation of crisis hotline outcomes. Part 1: nonsuicidal crisis callers. Suicide Life Threat Behav 2007 Jun 31;37(3):322-337. [doi: 10.1521/suli.2007.37.3.322] [Medline: 17579544]
8. Paul CL, Cox ME, Small HJ, Boyes AW, O'Brien L, Rose SK, et al. Techniques for improving communication of emotional content in text-only web-based therapeutic communications: systematic review. JMIR Ment Health 2017 Oct 24;4(4):e46 [FREE Full text] [doi: 10.2196/mental.6707] [Medline: 29066426]
9. Mishara BL, Côté LP, Dargis L. Systematic review of research and interventions with frequent callers to suicide prevention helplines and crisis centers. Crisis 2023 Mar;44(2):154-167 [FREE Full text] [doi: 10.1027/0227-5910/a000838] [Medline: 35086356]
10. Yip PS, Chan WL, Chan CS, He L, Xu Y, Chan E, et al. The opportunities and challenges of the first three years of open up, an online text-based counselling service for youth and young adults. Int J Environ Res Public Health 2021 Dec 14;18(24):13194 [FREE Full text] [doi: 10.3390/ijerph182413194] [Medline: 34948802]
11. Stoll J, Müller JA, Trachsel M. Ethical issues in online psychotherapy: a narrative review. Front Psychiatry 2019;10:993 [FREE Full text] [doi: 10.3389/fpsyt.2019.00993] [Medline: 32116819]
12. Bassilios B, Harris M, Middleton A, Gunn J, Pirkis J. Erratum to: characteristics of people who use telephone counseling: findings from secondary analysis of a population-based study. Adm Policy Ment Health 2015 Sep;42(5):633. [doi: 10.1007/s10488-014-0613-x] [Medline: 25403259]
13. Burgess N, Christensen H, Leach LS, Farrer L, Griffiths KM. Mental health profile of callers to a telephone counselling service. J Telemed Telecare 2008 Jan 01;14(1):42-47. [doi: 10.1258/jtt.2007.070610] [Medline: 18318929]
14. Middleton A, Gunn J, Bassilios B, Pirkis J. Systematic review of research into frequent callers to crisis helplines. J Telemed Telecare 2014 Mar;20(2):89-98 [FREE Full text] [doi: 10.1177/1357633X14524156] [Medline: 24518928]
15. Middleton A, Gunn J, Bassilios B, Pirkis J. The experiences of frequent users of crisis helplines: a qualitative interview study. Patient Educ Couns 2016 Nov;99(11):1901-1906. [doi: 10.1016/j.pec.2016.06.030] [Medline: 27401827]
16. Middleton A, Woodward A, Gunn J, Bassilios B, Pirkis J. How do frequent users of crisis helplines differ from other users regarding their reasons for calling? Results from a survey with callers to Lifeline, Australia's national crisis helpline service. Health Soc Care Community 2017 May;25(3):1041-1049. [doi: 10.1111/hsc.12404] [Medline: 27862572]
17. Spittal MJ, Fedyszyn I, Middleton A, Bassilios B, Gunn J, Woodward A, et al. Frequent callers to crisis helplines: who are they and why do they call? Aust N Z J Psychiatry 2015 Jan 27;49(1):54-64. [doi: 10.1177/0004867414541154] [Medline: 24972602]
18. Pirkis J, Middleton A, Bassilios B, Harris M, Spittal MJ, Fedszyn I, et al. Frequent callers to telephone helplines: new evidence and a new service model. Int J Ment Health Syst 2016;10:43 [FREE Full text] [doi: 10.1186/s13033-016-0076-4] [Medline: 27247615]
19. Hall B, Schlosar H. Repeat callers and the Samaritan telephone crisis line--a Canadian experience. Crisis 1995;16(2):66-89. [doi: 10.1027/0227-5910.16.2.66] [Medline: 7587293]
20. Houry D, Parramore C, Fayard G, Thorn J, Heron S, Kellermann A. Characteristics of household addresses that repeatedly contact 911 to report intimate partner violence. Acad Emerg Med 2004 Jun;11(6):662-667 [FREE Full text] [Medline: 15175205]
21. Georgoulas N. Counseling psychology: types, objectives and the role of counseling psychologist. Stud Humanit Soc Sci 2019:103. [doi: 10.32591/coas.e-conf.04.09103g]
22. Szlyk HS, Roth KB, García-Perdomo V. Engagement with crisis text line among subgroups of users who reported suicidality. Psychiatr Serv 2020 Apr 01;71(4):319-327 [FREE Full text] [doi: 10.1176/appi.ps.201900149] [Medline: 31795857]
23. Sindahl TN, Côte LP, Dargis L, Mishara BL, Bechmann Jensen T. Texting for help: processes and impact of text counseling with children and youth with suicide ideation. Suicide Life Threat Behav 2019 Oct 23;49(5):1412-1430. [doi: 10.1111/sltb.12531] [Medline: 30468267]
24. Yip P, Chan WL, Cheng Q, Chow S, Hsu SM, Law YW, et al. A 24-hour online youth emotional support: opportunities and challenges. Lancet Reg Health West Pac 2020 Nov;4:100047 [FREE Full text] [doi: 10.1016/j.lanwpc.2020.100047] [Medline: 34327390]
25. Xu Y, Chan CS, Tsang C, Cheung F, Chan E, Yip PS. Evaluating the effectiveness of concurrent sessions and counselors' attention allocation in online counseling. J Consult Clin Psychol 2023 Nov;91(11):640-651. [doi: 10.1037/ccp0000831] [Medline: 37616124]
26. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
27. Ribeiro M, Singh S, Guestrin C. “Why should i trust you?”: explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2016 Presented at: NAACL '16; June 12-17, 2016; San Diego, CA p. 97-101 URL: https://aclanthology.org/N16-3020.pdf [doi: 10.1145/2939672.2939778]
28. Gould MS, Pisani A, Gallo C, Ertefaie A, Harrington D, Kelberman C, et al. Crisis text-line interventions: evaluation of texters' perceptions of effectiveness. Suicide Life Threat Behav 2022 Jun;52(3):583-595 [FREE Full text] [doi: 10.1111/sltb.12873] [Medline: 35599358]

XSL•FO
RenderX

29. Gould MS, Chowdhury S, Lake AM, Galfalvy H, Kleinman M, Kuchuk M, et al. National Suicide Prevention Lifeline crisis chat interventions: evaluation of chatters' perceptions of effectiveness. Suicide Life Threat Behav 2021 Dec;51(6):1126-1137 [FREE Full text] [doi: 10.1111/sltb.12795] [Medline: 34331471]
30. Nesmith A. Reaching young people through texting-based crisis counseling. Adv Soc Work 2019 Jan 02;18(4):1147-1164. [doi: 10.18060/21590]
31. Xu YC, Chan CS, Tsang C, Cheung F, Chan E, Fung J, et al. Detecting premature departure in online text-based counseling using logic-based pattern matching. Internet Interv 2021 Dec;26:100486 [FREE Full text] [doi: 10.1016/j.invent.2021.100486] [Medline: 34877263]
32. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Department of Computer Science and Engineering, University of Minnesota. 2000. URL: https://hdl.handle.net/11299/215421 [accessed 2024-04-29]
33. Wong TY, Fang ZQ, Yu YT, Cheung C, Hui CL, Elvevåg B, et al. Discovering the structure and organization of a free Cantonese emotion-label word association graph to understand mental lexicons of emotions. Sci Rep 2022 Nov 15;12(1):19581 [FREE Full text] [doi: 10.1038/s41598-022-23995-z] [Medline: 36380119]
34. Berrouiguet S, Billot R, Larsen ME, Lopez-Castroman J, Jaussent I, Walter M, et al. An approach for data mining of electronic health record data for suicide risk management: database analysis for clinical decision support. JMIR Ment Health 2019 May 07;6(5):e9766 [FREE Full text] [doi: 10.2196/mental.9766] [Medline: 31066693]
35. Choudhury J, Ashraf FB. An analysis of different distance-linkage methods for clustering gene expression data and observing pleiotropy: empirical study. JMIR Bioinform Biotech 2022 Jun 17;3(1):e30890. [doi: 10.2196/30890]
36. Nielsen F. Introduction to HPC with MPI for Data Science. Cham, Switzerland: Springer; 2016.
37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text] [doi: 10.3389/fninf.2014.00014]
38. Yim O, Ramdeen KT. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. Quant Meth Psychol 2015 Feb 01;11(1):8-21. [doi: 10.20982/tqmp.11.1.p008]
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 2018 Dec 05;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
40. Dinno A. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. The Stata Journal 2015 Apr 01;15(1):292-300. [doi: 10.1177/1536867x1501500117]
41. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Wielkopolska Digital Library. 2014. URL: https://www.wbc.poznan.pl/publication/413565 [accessed 2024-04-29]
42. Pisani AR, Gould MS, Gallo C, Ertefaie A, Kelberman C, Harrington D, et al. Individuals who text crisis text line: key characteristics and opportunities for suicide prevention. Suicide Life Threat Behav 2022 Jun 26;52(3):567-582 [FREE Full text] [doi: 10.1111/sltb.12872] [Medline: 35615898]
43. Drum DJ, Brownson C, Burton Denmark A, Smith SE. New data on the nature of suicidal crises in college students: Shifting the paradigm. Prof Psychol Res Pr 2009 Jun;40(3):213-222. [doi: 10.1037/a0014465]
44. Stone MH. Long-term outcome in personality-disorders. Brit J Psychiat 1993;162:299-313. [doi: 10.1192/bjp.162.3.299] [Medline: 5637250]
45. Bateman AW, Gunderson J, Mulder R. Treatment of personality disorder. Lancet 2015 Feb 21;385(9969):735-743. [doi: 10.1016/S0140-6736(14)61394-5] [Medline: 25706219]
46. Lakeman R, King P, Hurley J, Tranter R, Leggett A, Campbell K, et al. Towards online delivery of dialectical behaviour therapy: a scoping review. Int J Ment Health Nurs 2022 Aug 19;31(4):843-856 [FREE Full text] [doi: 10.1111/inm.12976] [Medline: 35048482]
47. Asch DA. An interview with ChatGPT about health care. NNEJM Catal Innov Care Deliv 2023;4(2) [FREE Full text]
48. Hamdoun S, Monteleone R, Bookman T, Michael K. AI-based and digital mental health apps: balancing need and risk. IEEE Technol Soc Mag 2023 Mar;42(1):25-36. [doi: 10.1109/mts.2023.3241309]
49. Rickwood D, Wallace A, Kennedy V, O'Sullivan S, Telford N, Leicester S. Young people's satisfaction with the online mental health service eheadspace: development and implementation of a service satisfaction measure. JMIR Ment Health 2019 Apr 17;6(4):e12169 [FREE Full text] [doi: 10.2196/12169] [Medline: 30994470]
50. Groves RM, Peytcheva E. The impact of nonresponse rates on nonresponse bias - a meta-analysis. Public Opin Q 2008 May 02;72(2):167-189. [doi: 10.1093/poq/nfn011]
51. Perneger TV, Peytremann-Bridevaux I, Combescure C. Patient satisfaction and survey response in 717 hospital surveys in Switzerland: a cross-sectional study. BMC Health Serv Res 2020 Mar 02;20(1):158 [FREE Full text] [doi: 10.1186/s12913-020-5012-2] [Medline: 32122346]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 18.07.23; peer-reviewed by H Szlyk, A Kerber, D Nagel; comments to author 10.10.23; revised version received 14.11.23; accepted 12.04.24; published 30.05.24.

Please cite as:
Xu Y, Chan CS, Chan E, Chen J, Cheung F, Xu Z, Liu J, Yip PSF
Tracking and Profiling Repeated Users Over Time in Text-Based Counseling: Longitudinal Observational Study With Hierarchical Clustering
J Med Internet Res 2024;26:e50976
URL: https://www.jmir.org/2024/1/e50976
doi:10.2196/50976
PMID:38815258

©Yucan Xu, Christian Shaunlyn Chan, Evangeline Chan, Junyou Chen, Florence Cheung, Zhongzhi Xu, Joyce Liu, Paul Siu Fai Yip. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Transdiagnostic Video-Based Internet Intervention (Uni Virtual Clinic-Lite) to Improve the Mental Health of University Students: Randomized Controlled Trial

Louise M Farrer[1], PhD; Hayley M Jackson[1], BA; Amelia Gulliver[1], PhD; Alison L Calear[1], PhD; Liana Leach[2], PhD; Penelope Hasking[3], PhD; Natasha Katruss[1], Bsc; Philip J Batterham[1], PhD

[1]Centre for Mental Health Research, National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia

[2]Department of Health Economics Wellbeing and Society, National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia

[3]Curtin enAble Institute and School of Population Health, Faculty of Health Sciences, Curtin University, Curtin, Australia

Corresponding Author:
Louise M Farrer, PhD
Centre for Mental Health Research
National Centre for Epidemiology and Population Health
The Australian National University
63 Eggleston Rd
Canberra, 2601
Australia
Phone: 61 6125 8859
Email: louise.farrer@anu.edu.au

Abstract

Background: Numerous studies have demonstrated the effectiveness of digital interventions for improving the mental health of university students. However, low rates of engagement with these interventions are an ongoing challenge and can compromise effectiveness. Brief, transdiagnostic, web-based video interventions are capable of targeting key mental health and related issues affecting university students and may be more engaging and accessible for this population.

Objective: This study used a 2-arm randomized controlled trial to evaluate the effectiveness of Uni Virtual Clinic-Lite (UVC-Lite), a fully automated, transdiagnostic, web-based video intervention, relative to an attention-control condition. The primary outcomes were symptoms of depression and generalized anxiety disorder. The secondary outcomes included psychological distress, social anxiety symptoms, body appreciation, quality of life, well-being, functioning, general self-efficacy, academic self-efficacy, and help seeking. Program use (intervention uptake and engagement) and satisfaction were also assessed.

Methods: University students (n=487) with mild to moderate symptoms of distress were recruited from universities across Australia and randomly allocated to receive access to the UVC-Lite intervention or an attention-control condition targeting general health for a period of 6 weeks. UVC-Lite includes 12 modules, each comprising a brief animated video and an accompanying exercise. Of the 12 modules, 7 also included a brief symptom screening quiz. Outcomes were assessed at baseline, postintervention, and 3- and 6-months postintervention.

Results: The primary and secondary outcomes were analyzed on an intention-to-treat basis using mixed models repeated measures ANOVA. The intervention was not found to be effective relative to the control condition on any of the primary or secondary outcomes. While 67.9% (114/168) of participants accessed at least 1 module of the intervention, module completion was extremely low. Subgroup analyses among those who engaged with the program (completed at least 1 video) and those with higher baseline distress (Distress Questionnaire-5 score ≥15) did not reveal any differences between the conditions over time. However, uptake (accessing at least 1 video) and engagement (completing at least 1 video) were higher among those with higher baseline symptoms. Satisfaction with the intervention was high.

Conclusions: The UVC-Lite intervention was not effective relative to a control program, although it was associated with high satisfaction among students and was not associated with symptom deterioration. Given the challenges faced by universities in meeting demand for mental health services, flexible and accessible interventions such as UVC-Lite have the potential to assist students to manage symptoms of mental health problems. However, low uptake and engagement (particularly among students with lower levels of symptomatology) are significant challenges that require further attention. Future studies should examine the

XSL•FO
RenderX

effectiveness of the intervention in a more highly symptomatic sample, as well as implementation pathways to optimize effective engagement with the intervention.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12621000375853; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380146

(J Med Internet Res 2024;26:e53598) doi:10.2196/53598

KEYWORDS

university students; internet-based intervention; randomized controlled trial; mental health; transdiagnostic

Introduction

Background

There is growing concern about the mental health of university students worldwide. Most students commence university study during the transition from late adolescence to early adulthood, a time when the first onset of many common mental disorders is high [1,2]. In addition, this group faces numerous stressors such as transition to independent living, instability in relationships, financial stress, and academic pressure [3-6]; these stressors increase their risk of developing mental health problems [7]. Moreover, the COVID-19 pandemic and the associated mitigation measures severely disrupted the daily lives and educational experiences of university students around the world [8], with students reporting increased academic challenges due to a shift to remote learning, disruptions to paid work, concerns about their future academic and career prospects, and increased social isolation due to campus closures and physical distancing mandates [8-11].

Before the COVID-19 pandemic, research indicated that approximately one-third of university students met criteria for at least 1 mental disorder during a 12-month period, with generalized anxiety and mood disorders being the most prevalent diagnoses [12,13]. Furthermore, studies of the prevalence of psychological distress among university student populations have indicated that 19% to 48% of students experience high levels of psychological distress and that as many as 65% experience subsyndromal symptoms indicative of mild to moderate mental illness [14,15]. Research conducted since the onset of the pandemic has demonstrated an increase in the prevalence of mental health problems among university students [11,16,17], and approximately 30% to 40% of students are estimated to have experienced symptoms of depression or anxiety during the first 2 years of the pandemic [18-20].

Untreated mental disorders in young people place them at higher risk for developing comorbid mental disorders in later adulthood [21,22]. Moreover, mental health problems among university students can negatively impact academic performance [23] and have been associated with an increased risk of dropping out of university [24]. However, only approximately one-sixth of university students with a mental disorder receive minimally adequate treatment [12]. The most frequently reported barriers to using mental health services by students include a preference for self-management, concerns about stigma, high treatment costs, and scheduling difficulties [25]. Moreover, traditional models of delivering psychotherapy (ie, individual face-to-face sessions) are unlikely to be sufficient to meet the growing demand for services in universities and may not be feasible during times of crisis (such as during the COVID-19 pandemic). Therefore, there is an urgent need to reimagine existing models of mental health service delivery in tertiary education settings.

Internet-based and other digital mental health interventions are proposed as one solution to increase access to evidence-based treatment and prevention among the university student population. University students have indicated a willingness to use digital programs [26,27] and perceive numerous advantages in doing so, including increased treatment accessibility, anonymity, privacy, and confidentiality, as well as the potential to mitigate financial concerns by accessing support at a lower cost [28]. Moreover, several systematic reviews and meta-analyses have demonstrated that digital mental health interventions can be effective for university students [29-33]. In particular, these reviews have shown that interventions based on cognitive behavioral therapy (CBT) have been the most frequently used therapeutic approaches and demonstrate larger effects than other approaches [32]. However, the effectiveness of current interventions is also limited by low levels of adherence and high rates of dropouts [34].

Due to high rates of comorbidity between mental disorders in university student populations [35], transdiagnostic approaches that deliver therapeutic content targeting the common mechanisms that underlie mental disorders may have more clinical utility than disorder-specific approaches. Internet-based programs targeting the reduction of depression and anxiety symptoms using transdiagnostic approaches have generally been shown to be effective in improving one or more mental health outcomes among university students relative to control groups [36-38] although there are exceptions to this [7,39]. However, the potential of transdiagnostic internet-based interventions in university student populations has not yet been fully realized. Recent trials of transdiagnostic internet-based interventions with university students have predominantly focused on interventions delivered with guidance from a health professional or trained nonprofessional [7,40]. Although the provision of therapeutic guidance in an intervention can have small positive effects on the rates of intervention completion [38], this also requires increased resources, which substantially reduces the potential for scalability [41]. Moreover, sustained and active engagement with existing transdiagnostic interventions may be challenging for university students for different reasons. Those with few symptoms may not have much motivation for change or may not see a need to engage. Those who experience mental ill health or struggle to balance competing demands may find it challenging to engage with complex interventions that contain many pages of information; are predominantly text based; or

XSL•FO
RenderX

include multiple components (eg, written text, audio and visual information, and exercises and worksheets) [42]. In addition, most previous randomized controlled trials have tested transdiagnostic interventions using relatively small sample sizes (eg, ≤100 participants), which have not been sufficiently powered to detect small, yet clinically meaningful, effects in mental health outcomes that may be present in this population [7].

There is a need to explore innovative methods of delivering transdiagnostic programs that are low intensity and use potentially more engaging methods (eg, videos) to deliver therapeutic content directly to the user. A previous trial of a brief video-based transdiagnostic intervention in a sample of adults from the general population with elevated levels of psychological distress demonstrated significant reductions in depression, panic, and social anxiety symptoms relative to an attention-control condition [43]. Video-based interventions may be particularly accessible for university students who have limited time or motivation, as videos require less reading, scrolling, or shifting between pages of text-based materials or different intervention components. A recent study found that university students show a preference for internet-based mental health interventions that contain video content [44]. However, there are currently no video-based internet interventions that have been developed for and trialed with the university student population. In recognition of the potential utility of a video-based transdiagnostic internet intervention for university students, our research group developed the Uni Virtual Clinic-Lite (UVC-Lite), a brief intervention designed to help university students manage mental health problems and related issues.

Objective

In this 2-arm randomized controlled trial, we compared the unguided use of the UVC-Lite intervention for 6 weeks with an attention-control condition among students recruited from universities across Australia. The primary aim of the trial was to assess the feasibility and effectiveness of the intervention in reducing the symptoms of depression and generalized anxiety disorder at immediate postintervention and longer term at 3- and 6-month follow-ups. A range of secondary outcomes, as well as satisfaction with and use of the intervention, were also examined.

Methods

Participants (Recruitment and Eligibility Criteria)

Undergraduate and postgraduate students from all 42 universities in Australia were targeted using the recruitment strategies outlined below. Recruitment took place between August 2021 and May 2022, and all the follow-up data were collected by January 2023. Multiple strategies were used to recruit students into the trial. Students were provided with information about the trial via targeted Instagram and Facebook advertisements; via posts to university-affiliated groups on social media sites (Facebook, Instagram, Reddit, and Discord); and through university media channels such as newsletters. Recruitment materials targeted students interested in learning more about mental health and well-being. Student advocacy departments and associations, student housing coordinators, marketing teams, survey management departments, university counselors and well-being staff, university psychology clinics, and academic course conveners were contacted via email and social media pages to request assistance in distributing information about the trial to students via emails and flyers. Individuals involved in distributing recruitment materials were not asked to target any specific student groups.

To be eligible for the trial, students were required to (1) be enrolled at an Australian university; (2) currently reside in Australia; (3) be aged between 18 and 25 years; and (4) score between 8 and 17 on the Distress Questionnaire-5 (DQ5), which indicates moderate levels of psychological distress [45]. UVC-Lite was designed to be a brief internet-based self-help intervention; thus, students with a moderate level of psychological distress were targeted, as these programs are thought to be most suitable for this group [46]. However, because of the concerns noted in the ethical review about offering this program to people with higher levels of distress, students with high psychological distress (DQ5 score of 18-25) were deemed ineligible and were instead provided with specific instructions on how to access resources for help seeking. Students with a diagnosis of bipolar disorder, schizophrenia, posttraumatic stress disorder, or a personality disorder could participate if they fulfilled the eligibility criteria and were also currently receiving support or treatment for their disorder.

Procedure

Students interested in participating in the trial were provided with a link to a Qualtrics survey (Qualtrics International Inc) containing the participant information sheet, a question to obtain informed consent, and self-report questions to determine trial eligibility. Eligible participants were asked to provide an email address and complete the baseline survey. Participants who completed the baseline survey were randomized to either the UVC-Lite intervention condition or an attention-control condition for a period of 6 weeks. An independent researcher who was not involved with the trial generated a random sequence of integers between the values of 1 and 2 using a web-based application [47] and manually allocated participants to the trial conditions according to this sequence to ensure allocation concealment. Trial allocations were sent to a research assistant (NK), who enrolled participants in the trial.

During the 6-week intervention period, participants were sent 2 automated emails per week containing a link to an intervention or control module hosted in Qualtrics software. Participants could not access modules until they received these emails. Emails contained information about the topic of each module. Participants were not restricted in their use of usual services or support during the intervention or follow-up periods. All participants received an email containing a link to the postintervention survey 6 weeks after completing the baseline survey, and links to the follow-up surveys were sent 3 and 6 months after the postintervention survey was distributed. Participants were sent 2 email reminders to complete the postintervention and follow-up surveys. Participants who completed all the assessments were entered into a prize draw to win one of ten AUD $100 (USD $67) electronic gift cards.

XSL•FO
RenderX

In case of distress, participants could contact the research team and be followed up by the primary researcher (LMF), a registered psychologist, and provided with support and referral to relevant health, counseling, and crisis support services. No students in the trial contacted the research team in distress.

Ethics Approval

The trial was registered at the Australian New Zealand Clinical Trials Registry (ACTRN12621000375853), and ethical approval for the study was obtained from the Australian National University Human Research Ethics Committee (protocol #2020/412).

Intervention Condition: UVC-Lite

UVC-Lite is a brief internet-based intervention that comprises 12 modules targeting the common mechanisms that underlie mental health problems in university students (Multimedia Appendix 1). Each module includes a 3- to 6-minute video in which a series of animated characters introduce and share their lived experience of one or more mental health problems or issues affecting mental health in students. A therapist character presents relevant psychoeducation and therapeutic techniques, and a downloadable worksheet exercise designed to facilitate the practice of the therapeutic techniques presented follows each video. Additional features include self-monitoring quizzes for participants to track and receive feedback on their symptoms (included in modules 1, 3, 4, 5, 7, and 11) and links to additional help-seeking resources provided at their university or in the community if participants score above a certain symptom severity cutoff. Module transcripts and exercises were available for participants to download. The 12 modules covered the following topics and therapeutic strategies and were presented to participants in this order: (1) dealing with depression and low mood (behavioral activation: pleasant events scheduling); (2) tackling negative and anxious thoughts (cognitive reframing: thought diary), (3) dealing with anxiety (cognitive reframing, breathing, and grounding techniques); (4) managing study issues: procrastination and time management (practical strategies for time management); (5) perfectionism (challenging perfectionistic thoughts and behaviors); (6) coping with stress (mindfulness practice: focus on the present moment and mindful eating activity); (7) managing sleep issues (sleep hygiene); (8) social anxiety and shyness (behavioral experiments or exposure techniques); (9) relationships and loneliness (communication skills and social support); (10) social media use (practical strategies for reducing social media use); (11) body image (body functionality appreciation exercise); and (12) thoughts of suicide (safety planning exercise).

The content for the UVC-Lite program was drawn from a previously developed internet-based mental health intervention for university students, the Uni Virtual Clinic. Although this intervention was not effective in reducing the symptoms of depression, anxiety, or distress compared with a waitlist control group in a previous trial, it was shown to be effective in reducing the symptoms of social anxiety and improving academic self-efficacy [43]. The original Uni Virtual Clinic program, which was developed using extensive co-design processes involving students, university teaching and administrative staff, and service providers, was designed to be a comprehensive portal for student mental health [48]. We developed the UVC-Lite program based on the feedback from users of the original Uni Virtual Clinic program to ensure that it was focused primarily on the most critical issues affecting student mental health. The UVC-Lite incorporates key elements from the original program but is delivered in a more concise format involving brief videos containing student stories, tailored psychoeducation, and a range of psychotherapeutic strategies.

Attention-Control Condition: General Health Information

The attention-control condition received biweekly emails containing a link to the information pertaining to general health rather than mental health. The health information was delivered in a PDF document accessed through a weblink and was approximately matched to the UVC-Lite program on completion time. A similar program has previously been shown not to be associated with therapeutic reductions in depression [49]. A total of 12 general health topics were covered, including bone health, sun exposure, food hygiene, dietary supplements, kidney health, microbes, household burns, respiratory viruses, heart health, allergens, posture, and pancreas health. The general health information was derived from web-based public domain articles about general health topics published by the National Institutes of Health [50].

Measures

Overview

At baseline, the following demographic and study data were collected: age, sex, ethnicity, languages spoken, country of birth, living situation, state or territory of residence, hours in paid employment, financial stress, relationship status, study discipline and year of degree, study load, international or domestic student status, academic performance, and engagement with student life. Primary and secondary outcomes (described in subsequent sections) were assessed at the baseline, postintervention time point, and 3- and 6-month follow-ups. Satisfaction with the intervention program was assessed at the postintervention time point.

Primary Outcomes

Primary outcomes were symptoms of depression (measured by the Patient Heath Questionnaire-9 [PHQ-9]) [51] and generalized anxiety (measured by the Generalized Anxiety Disorder Scale-7 [GAD-7]) [52]. The PHQ-9 comprises 9 items rated on a 4-point scale, ranging from “not at all” to “nearly every day.” Item scores are summed to produce an overall severity score ranging from 0 to 27, with higher scores indicating greater symptom severity (0-4=no symptoms, 5-9=mild symptoms, 10-14=moderate symptoms, and 15-27=severe symptoms). The GAD-7 comprises 7 items rated on the same 4-point scale as the PHQ-9. Summed scores produce an overall severity score ranging from 0 to 21 (0-4=no symptoms, 5-9=mild symptoms, 10-14=moderate symptoms, and 15-21=severe symptoms). Higher scores indicate greater symptom severity. Both scales have been shown to have robust psychometric properties in general population samples and detect change over time [53,54]. In the current sample, internal consistency was acceptable at baseline (PHQ-9: α=.80; GAD-7: α=.84).

XSL•FO
RenderX

Secondary Outcomes

The secondary outcomes included social anxiety symptoms, psychological distress, body appreciation, quality of life, well-being, functioning, general self-efficacy, academic self-efficacy, and help-seeking behavior.

Social Anxiety

Symptoms of social anxiety were assessed using the Social Anxiety Disorder Screener [55], a brief 4-item scale with symptom severity scores ranging from 0 to 16. This scale has been developed and validated using Australian community–based samples, has demonstrated good convergent and divergent validity with diagnostic measures [55], and had very good internal consistency in this sample (α=.90).

Psychological Distress

Psychological distress was measured using the DQ5 [45], which consists of 5 items rated on a 5-point scale that assess the symptoms of common mental disorders experienced over the last 30 days (range 5-25; 5-7-mild distress, 8-17=moderate distress, and 18-25=high distress). The DQ5 has been shown to accurately identify individuals at risk for specific mental disorders relative to the Kessler-6 and Kessler-10 screeners [45]. Previous studies have demonstrated that DQ5 displays high internal consistency and external validity [45,56]. Internal consistency in the current sample was α=.58; however, this is an artifact of the restricted range on the DQ5 of the recruited sample [57,58].

Body Appreciation

Body appreciation was assessed using the 10-item Body Appreciation Scale-2 (BAS-2) [59]. The BAS-2 comprises 10 items rated on a 5-point scale. Item responses are averaged to produce a total score ranging from 1 to 5, with higher scores indicating greater body appreciation. The BAS-2 has demonstrated good internal consistency, test-retest reliability, and construct validity [59]. The scale had good internal consistency in this sample (α=.95).

Quality of Life

Quality of life was assessed using the European Health Interview Survey-8 (EUROHIS-8), an 8-item measure designed to assess the psychological, physical, social, and environmental aspects of quality of life (range 0-32) [60]. The EURO-HIS 8 has satisfactory discriminant validity and acceptable internal consistency [61], and internal consistency was adequate in the current sample (α=.79).

Well-Being

Well-being was assessed using the 5-item World Health Organization Well-Being Index, a 5-item scale with item responses ranging from 0 to 6, which produces a raw well-being score ranging from 0 to 25. Raw scores are multiplied by 4 to provide a final score ranging from 0 (worst-imaginable well-being) to 100 (best-imaginable well-being) [62]. The 5-item World Health Organization Well-Being Index has demonstrated sensitivity and specificity in screening for depression and is sensitive to change in clinical trials [63]. The internal consistency of the scale in the current sample was adequate (α=.82).

Functioning

Functioning was measured by the Recovering Quality of Life scale, a 20-item measure assessing the dimensions of mental health functioning including activity, hope, belonging and relationships, self-perception, well-being, autonomy, and physical health. The items are rated on a 5-point scale ranging from 0 to 4, with total scores ranging from 0 to 80. The scale has acceptable internal consistency, test-retest reliability, and convergence with related measures [64]. Internal consistency in the current sample was good (α=.88).

General Self-Efficacy

Self-efficacy was measured using the 10-item General Self-Efficacy Scale, which measures perceived coping with daily hassles and adaptation to stressful life events on a 4-point scale (range 0-30) [65]. The 10-item General Self-Efficacy Scale has been validated in 31 countries and languages and has acceptable internal consistency and good discriminant and concurrent validity [66,67]. Internal consistency in this study was good (α=.88).

Academic Self-Efficacy

Academic self-efficacy (ie, confidence in one's ability to successfully complete university-related tasks) was assessed using the Study and Social subscales of the College Self-Efficacy Inventory (CSEI; 15 items, 5-point Likert scale ranging from very unconfident to very confident, range 0-60) [68]. The CSEI Course subscale measures self-efficacy to complete tasks associated with university study (eg, "How confident do you feel to keep up to date with your university work?"). The CSEI Social subscale measures self-efficacy related to social interactions at university (eg, "How confident do you feel to talk to your lecturers/tutors?"). Both subscales have demonstrated adequate internal consistency (α=.79-.86) and have been found to be negatively correlated with social anxiety and academic worry or concern and positively correlated with grade point average [69,70]. Internal consistency in this study was good (α=.87).

Help Seeking

Help seeking was measured with the Actual Help Seeking Questionnaire [71]. Participants were asked whether they had sought help from a range of formal (ie, general practitioner, psychologist, counselor, and helpline) and informal (ie, friends, family, and partner) help sources in the past month. The Actual Help Seeking Questionnaire has been found to adequately differentiate help-seeking behavior for different problems and help sources [71].

Intervention Satisfaction and Adherence

Program use was assessed using author-developed JavaScript embedded within Qualtrics to track clicks and percentage of videos watched. Measures included uptake (number of modules and videos started) and engagement (viewing at least 50% of each video, number of exercises accessed, and number of quizzes completed). Satisfaction was assessed by asking participants how satisfied they were with the intervention

XSL•FO
RenderX

(5-point scale, *very unsatisfied* to *very satisfied*); whether they would recommend the intervention to other students (4-point scale, *definitely not* to *yes, definitely*); how much they learned from the intervention (4-point scale, from *almost nothing* to *a great deal*); and which modules they liked the best.

Blinding

Trial researchers were blinded to treatment allocation. Participants were blinded to whether they received the active or attention-control intervention. Specifically, participants were informed that they would be randomized to receive 1 of 2 programs: information and strategies for improving mental health (UVC-Lite) or information and strategies for improving general health (attention control). They were not provided with information about which program was expected to be more effective. Assessments were also blinded, as self-report questionnaires were delivered on the web.

Statistical Analyses

Statistical analyses were conducted by the primary author (LMF) and author PJB using SPSS (version 28; IBM Corp) for Windows [72]. Sample characteristics were described at baseline, and differences between the intervention and control groups were examined using 2-tailed *t* tests and chi-square statistics. Predictors of dropout (failure to complete trial assessments) were examined at postintervention and at 3- and 6-month follow-ups using logistic regression. Predictors of uptake and engagement with the UVC-Lite intervention were also examined using logistic regression. The primary and secondary outcome variables were analyzed on an intention-to-treat basis using mixed models repeated measures ANOVA, with measurement occasion as a within-groups factor and trial condition as a between-groups factor [73]. Mixed models incorporate all available data under the assumption that data are missing at random. Within-person variation was modeled using an unstructured covariance matrix. To examine whether intervention effectiveness was related to intervention adherence or baseline symptoms, analyses were re-estimated among participants who watched at least 1 video and those with higher levels of baseline distress (DQ5 score ≥15). Logistic regression was used to examine the relationship between trial condition and help-seeking behavior at postintervention and at 3- and 6-month follow-ups.

Power

A power calculation based on repeated measures ANOVA (conducted using G*Power) indicated that a target sample size of 418 would enable the detection of an effect of Cohen d=0.3 between conditions, with 90% power and a significance level of α=.01. The recruited sample size was 487, allowing the detection of an effect of Cohen d=0.3 at a significance level of .01 with slightly higher power (~95%). However, high attrition was observed in the trial, with a sample of 265 retained at postintervention, 105 at 3 months, and 172 at 6 months.

Results

Trial Flow and Survey Completion

The flow of participants through the trial is shown in Figure 1. A total of 2865 individuals clicked on the study invitation. Of them, 523 (18.25%) completed the baseline survey and were randomized into trial conditions. The most common reasons for not progressing to randomization were clicking on the link but not reading the study information and providing consent (n=1132, 39.5%) and scoring >17 on the DQ5 (n=640, 22.3%). After completing the baseline survey, 36 participants were identified as being outside the eligible age range for the trial and were excluded from analyses, leaving 487 eligible participants. A total of 265 participants (54.4%) completed the postintervention survey, 105 (21.6%) completed the 3-month follow-up survey, and 172 (35.3%) completed the 6-month follow-up survey. In total, 2 errors were discovered during the trial that resulted in a delay in participants receiving either their condition allocation after completing the baseline survey or their 3-month follow-up survey. As a result of administrative errors, 129 (26.5%) participants experienced a delay of >3 days between completing their baseline assessment and receiving confirmation of their allocated trial condition and instructions for accessing their materials. A technical error with the survey delivery platform resulted in 188 (38.6%) participants not receiving their 3-month follow-up survey on the date when it was due to be administered. Baseline depression and generalized anxiety symptoms, trial condition, and whether trial condition allocation or 3-month follow-up survey was administered on time were examined as predictors of survey completion at postintervention and at 3-and 6-month follow-ups. Baseline symptoms and trial condition were unrelated to survey completion at all time points. However, delay in the administration of the 3-month follow-up survey due to technical error was significantly associated with increased odds of survey completion at 3 months (B=2.61; P<.001).

XSL•FO
RenderX

Figure 1. Trial flow. DQ-5: Distress Questionnaire; MH: mental health; UVC-Lite: Uni Virtual Clinic-Lite.

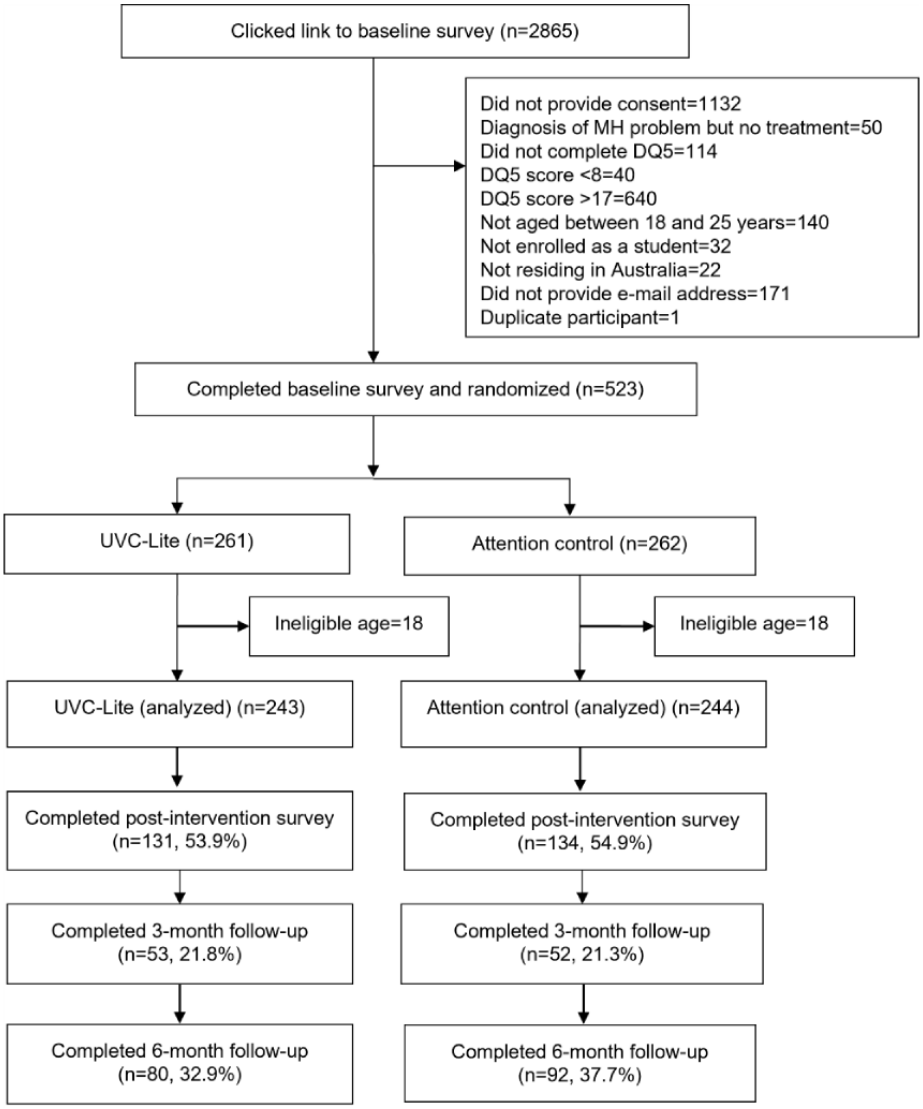

Sample Characteristics

Table 1 presents the baseline demographic and clinical characteristics of the sample. Students from 40 institutions in Australia participated in the study. The sample (n=487) was predominantly female (n=355, 72.9%), and participants were aged on average 21 (SD 2.1) years. Most participants were domestic students (411/487, 84.4%) studying full time (449/487, 92.2%) at the undergraduate level (416/485, 85.8%). Participants were studying degrees from a range of different disciplines. Most students were in some form of paid employment (328/487, 67.4%) and were living either with family (209/487, 42.9%) or on campus (152/487, 31.2%). There were no differences between participants randomized to the intervention or control conditions on any of the baseline demographic or symptom variables. As expected, due to screening, distress scores were moderate at baseline, whereas baseline depression and generalized anxiety symptoms were within the mild range.

XSL•FO
RenderX

Table 1. Demographic and study characteristics of the sample.

	Intervention (n=243)	Control (n=244)	Total (n=487)	Chi-square (*df*)	*t* test (*df*)	*P* value
Categorical variables, n (%)						
Sex				0.4 (2)	—[a]	.82
Female	180 (74.1)	175 (71.7)	355 (72.9)			
Male	52 (21.4)	58 (23.8)	110 (22.6)			
Other	11 (4.5)	11 (4.5)	22 (4.5)			
Ethnicity				3.2 (4)	—	.53
Aboriginal and/or Torres Strait Islander and/or Pacific Islander	6 (2.5)	2 (0.8)	8 (1.6)			
African/Middle Eastern	6 (2.5)	6 (2.5)	12 (2.5)			
Asian/Indian	65 (26.7)	75 (30.7)	140 (28.7)			
White/European	154 (63.4)	152 (62.3)	306 (62.8)			
Other	11 (4.5)	8 (3.3)	19 (3.9)			
Current living situation				4.6 (6)	—	.59
With parents/family	100 (41.2)	109 (44.7)	209 (42.9)			
On-campus housing	82 (33.7)	70 (28.7)	152 (31.2)			
Friends off-campus	24 (9.9)	21 (8.6)	45 (9.2)			
Others off-campus	19 (7.8)	18 (7.4)	37 (7.6)			
With partner/children	10 (4.1)	17 (7)	27 (5.5)			
Alone	7 (2.9)	9 (3.7)	16 (3.3)			
Other	1 (0.4)	0 (0)	1 (0.2)			
Hours per week in paid employment				6.05 (5)	—	.30
None	89 (36.6)	70 (28.7)	159 (32.6)			
1-9	53 (21.8)	58 (23.8)	111 (22.8)			
10-19	68 (28)	67 (27.5)	135 (27.7)			
20-29	24 (9.9)	33 (13.5)	57 (11.7)			
30-39	8 (3.3)	15 (6.1)	23 (4.7)			
≥40	1 (0.4)	1 (0.4)	2 (0.4)			
Relationship status				2.50 (3)	—	.48
Single	170 (70)	165 (67.6)	335 (68.8)			
In relationship (not living together)	51 (21)	57 (23.4)	108 (22.2)			
In relationship (living together)	20 (8.2)	22 (9)	42 (8.6)			
Other	2 (0.8)	0 (0)	2 (0.4)			
Discipline of degree studied (n=438)				5.46 (7)	—	.60
Health/medicine	83 (38.4)	71 (32)	154 (35.2)			
Arts/social sciences	37 (17.1)	42 (18.9)	79 (18)			
Science	39 (18.1)	45 (20.3)	84 (19.2)			
Engineering/computing	21 (9.7)	28 (12.6)	49 (11.2)			
Business/economics	11 (5.1)	14 (6.3)	25 (5.7)			
Law/criminology	14 (6.5)	15 (6.8)	29 (6.6)			
Education	11 (5.1)	6 (2.7)	17 (3.9)			
Tertiary preparation course	0 (0)	1 (0.5)	1 (0.2)			
Year of degree				4.30 (3)	—	.23

XSL•FO
RenderX

	Intervention (n=243)	Control (n=244)	Total (n=487)	Chi-square (*df*)	*t* test (*df*)	*P* value
First-year undergraduate	93 (38.3)	86 (35.2)	179 (36.8)			
Later-year undergraduate	107 (44)	112 (45.9)	219 (45)			
Honors	5 (2.1)	13 (5.3)	18 (3.7)			
Postgraduate	37 (15.2)	32 (13.1)	69 (14.2)			
Study load				0.44 (1)	—	.51
Full time	226 (93)	223 (91.4)	449 (92.2)			
Part time	17 (7)	21 (8.6)	38 (7.8)			
Student status				0.07 (1)	—	.79
Domestic	204 (84)	207 (84.8)	411 (84.4)			
International	39 (16)	37 (15.2)	76 (15.6)			
Average mark/grade achieved last semester (n=352)				0.86 (4)	—	.93
High distinction	41 (23.4)	44 (24.9)	85 (24.1)			
Distinction	79 (45.1)	76 (42.9)	155 (44)			
Credit	42 (24)	47 (26.6)	89 (25.3)			
Pass	12 (6.9)	9 (5.1)	21 (6)			
Fail	1 (0.6)	1 (0.6)	2 (0.6)			
Engagement with university life				1.40 (4)	—	.84
Not at all (only attend classes)	60 (24.7)	53 (21.7)	113 (23.2)			
Somewhat	80 (32.9)	80 (32.8)	160 (32.9)			
Moderate (participate in some university-based activities)	53 (21.8)	57 (23.4)	110 (22.6)			
High	29 (11.9)	27 (11.1)	56 (11.5)			
Extremely high (involved in student leadership activities within university)	21 (8.6)	27 (11.1)	48 (9.9)			
Continuous variables , mean (SD)						
Age (y)	20.67 (2.15)	20.57 (2.09)	20.62 (2.12)	—	0.55 (485)	.58
Depression (PHQ-9[b])	9.70 (5.11)	9.02 (4.34)	9.36 (4.75)	—	1.59 (485)	.11
Generalized anxiety (GAD-7[c])	7.84 (4.53)	7.28 (4.13)	7.56 (4.34)	—	1.41 (485)	.16
Social anxiety (SAD[d])	5.91 (3.69)	6.29 (3.89)	6.10 (3.79)	—	–1.12 (485)	.26
Psychological distress (DQ5[e])	14.10 (2.62)	13.92 (2.56)	14.01 (2.59)	—	0.77 (485)	.44
Body appreciation (BAS-2[f])	3.40 (0.90)	3.35 (0.87)	3.38 (0.88)	—	0.73 (485)	.47
Quality of life (EURO-HIS-8[g])	28.63 (5.16)	28.51 (4.87)	28.57 (5.01)	—	0.27 (485)	.79
Well-being (WHO-5[h])	47.00 (18.43)	47.57 (17.39)	47.28 (17.90)	—	–0.35 (484)	.72
Functioning (ReQoL[i])	45.24 (11.43)	45.64 (11.02)	45.44 (11.22)	—	–0.39 (484)	.70
Self-efficacy (GSE-10[j])	19.14 (5.21)	18.87 (4.64)	19.01 (4.93)	—	0.61 (484)	.54

XSL•FO
RenderX

	Intervention (n=243)	Control (n=244)	Total (n=487)	Chi-square (*df*)	*t* test (*df*)	*P* value
Academic self-efficacy (CSEI[k])	34.53 (10.86)	33.97 (10.17)	34.25 (10.51)	—	0.59 (484)	.55

[a]Not applicable.
[b]PHQ-9: Patient Health Questionnaire-9.
[c]GAD-7: Generalized Anxiety Disorder Scale-7.
[d]SAD: Social Anxiety Disorder Screener.
[e]DQ5: Distress Questionnaire-5.
[f]BAS-2: Body Appreciation Scale-2.
[g]EURO-HIS-8: European Health Interview Surveys-8.
[h]WHO-5: 5-item World Health Organization Well-Being Index.
[i]ReQoL: Recovering Quality of Life.
[j]GSE-10: General Self-Efficacy Scale-10.
[k]CSEI: College Self-Efficacy Inventory.

Intervention Adherence and Use

Uptake of and engagement with the UVC-Lite intervention was low (Table 2). On average, participants (n=243) started (eg, clicked on the link directing them to the intervention) 5.1 (SD 5.08) of the 12 intervention modules, with 168 (69.1%) participants starting at least 1 module, and 51 (21%) participants starting all 12 modules. Among those who clicked through to the intervention modules, participants started an average of 3.58 (SD 4.23) of the 12 videos, with 114 (67.9%) participants starting at least 1 of the 12 videos and only 14 (8.3%) participants starting all 12 videos. Only 11 (6.5%) participants watched all 12 videos. Encouragingly, among those who started 1 or more videos, 92.6% of participants watched over half of the video once they started (on average). Engagement with the module exercises was lower than the completion of the quizzes. Participants accessed an average of 1.89 (SD 3.02) of the 12 module exercises, with 86 (51.2%) participants accessing 1 or more exercises, and only 1 (0.6%) participant accessing all 12. In contrast, participants completed an average of 4.14 (SD 2.73) of 7 quizzes, with 151 (89.9%) participants completing at least 1 quiz and 57 (33.9%) participants completing all 7. Age, gender, student status (domestic vs international), baseline distress (DQ5 score: 8-14 vs 15-17), receiving trial allocation information on time, and year of study (first year vs later year) were examined as predictors of uptake (starting at least 1 module) and engagement (watching at least 1 video). Those with higher levels of baseline distress (DQ5 score≥15) were more likely to start at least 1 module (B=2.18; *P*=.01) and watch at least 1 video (B=2.14, *P*=.007) compared with those with lower distress. In addition, those who were in a later year of their degree were more likely to engage (watch at least 1 video) than those in their first year (B=2.36; *P*=.008). Although it was not significant, there was a trend toward higher intervention uptake among those who received their trial condition allocation on time (within 3 days of completing the baseline assessment; B=1.78, *P*=.08). More detailed descriptive statistics for uptake and engagement with each intervention module are provided in Multimedia Appendix 2. The number of participants accessing each module declined over the order of modules. The most frequently accessed module was module 1 (145/243, 59.7%), and the least frequently accessed module was module 12 (83/243, 34.2%).

XSL•FO
RenderX

Table 2. Uptake and engagement with the Uni Virtual Clinic-Lite intervention.

Uptake and engagement	Values
Participants who received the intervention (n=243), n (%)	
Started at least 1 module	168 (69.1)
Started all 12 modules	51 (21)
Participants who clicked through to the intervention modules (n=168), n (%)	
Started at least 1 video	114 (67.9)
Started all 12 videos	14 (8.3)
Watched ≥50% of at least 6 videos	42 (25)
Watched ≥50% of all 12 videos	11 (6.5)
Accessed at least 1 exercise	86 (51.2)
Accessed all 12 exercises	1 (0.6)
Accessed at least 1 quiz	151 (89.9)
Accessed all 7 quizzes	57 (33.9)
Modules started, mean (SD)	5.10 (5.08)
Videos started, mean (SD)	3.58 (4.23)
Exercises accessed, mean (SD)	1.89 (3.02)
Quizzes completed, mean (SD)	4.14 (2.73)

Primary Outcomes

Table 3 shows the estimated marginal means for the intervention and control groups on the primary outcomes at each measurement occasion. Significant decreases in depression ($F_{3,191.2}$=4.3; P=.006) and generalized anxiety ($F_{3,192.1}$=14.8; P≤.001) were observed over time in both trial groups, but no significant time-by-condition interactions were found, indicating that there was no significant difference in symptom reduction over time between conditions.

XSL•FO
RenderX

Table 3. Estimated marginal means for the intervention (Uni Virtual Clinic-Lite; UVC-Lite) and control groups at each measurement occasion and results of the mixed models analysis.

	Intervention (UVC-Lite) condition, mean (SE)				Control condition, mean (SE)				Occasion×condition	
	Baseline	Postintervention time point	3 months	6 months	Baseline	Postintervention time point	3 months	6 months	*F* test (*df*)	*P*
Depression[a]	9.7 (0.3)	8.9 (0.4)	9.1 (0.6)	9.1 (0.6)	9.0 (0.3)	8.1 (0.4)	7.9 (0.6)	8.9 (0.5)	0.51(3)	.68
Generalized anxiety[b]	7.8 (0.3)	6.7 (0.3)	7.6 (0.5)	8.1 (0.5)	7.3 (0.3)	5.9 (0.3)	6.6 (0.5)	7.4 (0.5)	0.22 (3)	.88
Social anxiety[c]	5.9 (0.2)	5.6 (0.3)	6.3 (0.4)	5.9 (0.4)	6.3 (0.2)	5.5 (0.3)	5.3 (0.4)	5.8 (0.3)	1.65 (3)	.18
Distress[d]	14.1 (0.2)	12.0 (0.3)	12.5 (0.5)	12.4 (0.5)	13.9 (0.2)	12.0 (0.3)	11.3 (0.5)	12.5 (0.5)	1.21 (3)	.31
Body appreciation[e]	3.4 (0.1)	3.4 (0.1)	3.6 (0.1)	3.4 (0.1)	3.3 (0.1)	3.3 (0.1)	3.4 (0.1)	3.4 (0.1)	0.38 (3)	.77
Quality of life[f]	28.6 (0.3)	28.5 (0.4)	28.1 (0.5)	28.8 (0.5)	28.5 (0.3)	28.8 (0.4)	29.3 (0.5)	28.9 (0.5)	1.23 (3)	.30
Well-being[g]	47.0 (1.1)	50.0 (1.5)	50.3 (2.2)	53.5 (1.9)	47.6 (1.1)	49.8 (1.5)	50.5 (2.2)	49.9 (1.8)	0.99 (3)	.40
Functioning[h]	45.2 (0.7)	46.4 (1.0)	46.3 (1.4)	46.9 (1.3)	45.6 (0.7)	47.3 (1.0)	47.5 (1.4)	45.4 (1.3)	1.01 (3)	.39
Self-efficacy[i]	19.1 (0.3)	19.1 (0.4)	19.2 (0.6)	19.2 (0.5)	18.9 (0.3)	19.1 (0.4)	19.5 (0.6)	19.7 (0.5)	0.47 (3)	.71
Academic self-efficacy[j]	34.5 (0.7)	36.1 (0.9)	36.1 (1.2)	37.3 (1.0)	34.0 (0.7)	35.2 (0.8)	37.0 (1.2)	35.9 (1.0)	0.73 (3)	.53

[a]Patient Health Questionnaire-9; range 0 to 27.
[b]Generalized Anxiety Disorder Scale-7; range 0 to 21.
[c]SAD: Social Anxiety Disorder Screener; range 0 to 16.
[d]Distress Questionnaire-5; range 5 to 25.
[e]Body Appreciation Scale-2; range 1 to 5.
[f]European Health Interview Surveys-8; range 0 to 32.
[g]5-item World Health Organization Well-Being Index; range 0 to 100.
[h]Recovering Quality of Life; range 0 to 80.
[i]General Self-Efficacy Scale-10; range 0 to 30.
[j]College Self-Efficacy Inventory; range 0 to 60.

Secondary Outcomes

No significant differences were found between conditions over time on any secondary outcome variables (Table 3). However, both conditions showed significant reductions in distress ($F_{3,187.6}$=25.0; P≤.001) and significant increases in body appreciation ($F_{3,173.5}$=2.8; P=.04), quality of life ($F_{3,188.1}$=4.3; P=.006), and academic self-efficacy ($F_{3,175.3}$=5.6; P=.001) over time. Trial condition was not associated with help seeking from different sources at postintervention, 3-month follow-up, or 6-month follow-up.

Analyses Among Engagers and Those With Higher Distress

No significant differences were found between conditions over time on any of the primary or secondary outcomes among those who engaged with the intervention (watched ≥1 videos). However, one time-by-condition interaction approached significance; nonsignificant reductions in distress were observed at the postintervention time point and 6-month follow-up among intervention group participants who watched ≥1 videos, compared with those in the control group ($F_{3,161.1}$=2.3=; P=.08). No significant differences were found between conditions over time among those with higher levels of distress (DQ5 ≥15; n=260, 53.4%) at baseline.

Satisfaction

Most participants who returned a postintervention survey reported being satisfied with the intervention; 81.3% (100/123) of the participants were either “very satisfied” or “somewhat satisfied” following the use of the UVC-Lite. Most participants (115/124, 92.7%) also reported that they would be willing to recommend the UVC-Lite to other students, and 75.6% (93/123) reported that they learned either “a lot” or “a fair bit” from the program. In terms of specific modules, 41.2% (47/114) of the participants reported that they liked the module on tackling negative and anxious thoughts the most, followed by the social anxiety module (45/114, 39.5%) and the study issues and

XSL•FO
RenderX

procrastination module (44/114, 38.6%); a low proportion of students reported liking the module on thoughts of suicide (13/114, 11.4%) and social media and internet use (24/114, 21.1%).

Discussion

Principal Findings

This trial examined the effectiveness of a brief, video-based intervention for improving the symptoms of depression and generalized anxiety disorder among university students with moderate psychological distress. Brief, self-guided interventions delivered at scale have the potential to confer significant public health benefits, particularly in university-based settings where demand for resource-intensive services outstrips supply. However, in this trial, the intervention was not effective relative to the control condition on either of the primary outcomes or any of the secondary outcomes. This is at odds with the findings of systematic reviews and meta-analyses suggesting that internet-based programs can be effective for a range of mental health outcomes in university students [31-33]. Despite a degree of heterogeneity in digital interventions for university students in terms of the level of guidance, program length and mode of delivery, therapeutic approach, and program components, most digital interventions for students comprise text-based modules that are based on CBT. Reviews suggest that the effects generally tend to be higher in programs that are guided, are based on CBT, and target an indicated sample [30,32,33]. The results from the very few studies to date that have examined digital transdiagnostic interventions for university students are mixed. A recent randomized controlled trial of a guided digital transdiagnostic intervention was not found to be effective [7], whereas an older study comparing a transdiagnostic intervention to a waitlist control group found positive outcomes for anxiety and depression [36]. The findings of this trial, together with the limited and mixed evidence to date, suggest that uncertainty remains regarding the utility of digital transdiagnostic interventions for students or what components are needed in an effective intervention. More high-powered trials are needed, particularly those examining brief interventions and the use of video and other alternatives to text-based intervention modules.

There are several possible explanations for the null findings in this trial. First, engagement with the intervention was low. Approximately two-thirds of participants (168/243, 69.1%) accessed at least 1 component of the intervention, which is comparable to the rates of uptake in other studies of digital mental health interventions for university students (79.1%; range 32.2%-100%) [74]. However, among those who accessed the intervention, very few participants either completed these components (11/168, 6.5%) or accessed all the program modules (14/168, 8.3%), which is also consistent with other studies of digital interventions with students [33] and a study of a comparable transdiagnostic video-based intervention trialed in the general community [75]. Possible reasons for the low levels of engagement in this trial include low perceived need, administration issues with the trial, or lack of interest among participants in engaging with the trial materials. The trial was originally designed to recruit students with at least moderate levels of psychological distress at baseline (DQ5 score>8). However, due to concerns expressed by the university ethics review committee about the appropriateness of a low-intensity digital intervention for students with high levels of distress, eligibility criteria for the trial were revised to exclude students who screened positive for high levels of distress (DQ5 score>17). As indicated in the trial flow diagram, a significant number of students expressed interest in the trial but were deemed ineligible on the basis of high psychological distress (n=640). Students who were screened as eligible on the basis of moderate psychological distress (DQ5 score=8-17) scored within the mild ranges of the PHQ-9 and GAD-7 at baseline, possibly indicating a low perceived or actual need for the intervention. Higher symptom severity has been shown to be positively associated with engagement with digital interventions [76], and in this trial, those with higher baseline distress were more likely to engage with the intervention, supporting the theory that low rates of adherence were potentially driven by low perceived need.

Second, due to an administrative error, a proportion of trial participants (n=129, 26.5%) did not receive their condition allocation and materials within a timely window after registering for the trial and completing the baseline assessment. For some participants, there was a delay of almost 1 month. Although these errors were not significantly associated with negative impacts on intervention uptake or survey completion in this study, they nonetheless jeopardized the fidelity of the trial and may have somewhat contributed to the low intervention adherence and missing data. There is a risk that participants may lose interest in research participation if there are long gaps in receiving contact from the research team, which has led researchers to use strategies to maintain ongoing contact with research participants (email or phone reminders, newsletters, and social media groups). Randomization and initial email contact were conducted manually in this trial and were thus vulnerable to administrative error. Automating these procedures can help to minimize delays and errors in administering trial materials. Technical issues have been previously shown to be a major barrier to engagement with digital mental health interventions [76].

Third, the trial materials were delivered twice weekly to participants via email, which might have been missed or ignored by students who receive a large volume of email or check their email infrequently. Despite attempts to make these emails more engaging using eye-catching subject lines and wording, other methods for linking students to intervention materials should be considered, such as SMS text message reminders or prompts. Finally, it is possible that the intervention was not engaging for participants or that participants lacked interest in engaging with the trial materials. Young people may be changing in the ways that they engage with technology over time, such that an internet-based intervention may be less novel and attractive now than it was in the past. Moreover, the version of the intervention that was trialed in this study was neither co-designed with the students nor personalized or customizable. Lack of perceived relevance and personalization have been previously reported as the barriers to engagement with digital mental health interventions [76]. In this study, using co-design

XSL•FO
RenderX

principles may have better aligned the intervention with the needs and expectations of the target users, leading to improved engagement.

The intervention was acceptable among the sample, as indicated by most participants indicating high satisfaction with the program and a willingness to recommend the program to others. Participants indicated that they liked the CBT-based modules the most (tackling negative and anxious thoughts and social anxiety), possibly because these modules contained tangible written exercises that users could complete or because these topics resonated the most with the participants in the sample. Engagement with the intervention was also higher among the later-year students than with the first-year students. This is noteworthy, given previous research suggesting that students may be at a heightened risk of developing mental health problems during their first year of university [77]. Possible delays in problem recognition or perceived need for treatment among students may account for the higher rates of engagement with the intervention among the later-year students in this trial. Students in their first year of university are often targeted with programs designed to help them adjust to university study and build social connections and study skills. Preventative mental health interventions could also be delivered at this time to assist students to build mental health literacy and skills in managing distress. However, findings from this trial suggest an ongoing need for accessible mental health interventions throughout a student's university candidature.

Strengths and Limitations

The study recruited a relatively larger sample than other trials of digital transdiagnostic interventions in university student populations [7,36]. The study also recruited a national sample from multiple universities, enhancing the generalizability of the findings. The study used a robust attention-control condition rather than a waitlist or no-treatment control. However, there are also several notable limitations, most of which are discussed in the previous section in relation to the study findings. In addition, there was low participation among male and gender-diverse students, limiting the generalizability of the findings to these groups. There was also high attrition in the trial, and although the data analytic methods used robustly accounted for missing data, the trial was likely underpowered due to not reaching the recruitment target and greater than expected dropouts from the assessments.

Future Directions

The findings highlight several implications for both future research on the UVC-Lite intervention and the implementation of this and other digital interventions in higher education settings. UVC-Lite was designed to be a highly accessible and flexible intervention that does not require significant financial or technological resourcing to implement. It was associated with high satisfaction among students and was not associated with symptom deterioration. Given the challenges faced by universities in meeting demand for mental health services, interventions such as UVC-Lite have significant potential to assist students to manage low-level distress and prevent the development of severe symptomatology when delivered universally. However, low uptake and engagement among students with low levels of symptomatology is a significant challenge that requires further attention. Future studies should examine the effectiveness of the intervention in a more appropriately symptomatic sample. Moreover, very few studies have examined the drivers of engagement with digital interventions among university students [78] or robustly tested strategies to facilitate engagement. The order of delivery of the modules appeared to affect the engagement in the current trial, with earlier modules being accessed more frequently than later modules. This may suggest that 12 modules are too many or that tailoring modules to participant symptoms, rather than expecting participants to engage with all 12 modules may improve engagement. Delivery of the intervention was fully automated, with no human support or contact. Further studies are needed to test different implementation pathways for digital interventions in university settings, particularly those involving greater support (eg, through counseling centers or other health services or via the classroom). Longitudinal studies of mental health in university settings may provide useful data about critical time points for targeting digital interventions to different student groups. There is capacity for interventions such as UVC-Lite to be expanded and adapted to address a wider range of topics relevant to students from different ethnic and cultural backgrounds, and this should also be a focus of future research and implementation efforts.

Conclusions

This study was the first to examine the effectiveness of a video-based transdiagnostic internet intervention for improving the mental health of university students. Although the intervention was designed to be highly engaging for a university student population, it was not effective in improving mental health outcomes, and sustained use of the intervention was low, possibly due to the low perceived need for the intervention among the study sample. Nonetheless, satisfaction with the intervention was high, and given the potential utility of brief transdiagnostic digital interventions in higher education settings, further research should examine the effectiveness of the UVC-Lite program among students with higher need for the program, as well as implementation pathways and strategies that optimize effective engagement with the program.

Acknowledgments

Funding for this project was provided by Australian Rotary Health. LMF is supported by a fellowship from the Australian Research Council (DE190101382). ALC is supported by a National Health and Medical Research Council (NHMRC) fellowship (1173146).

XSL•FO
RenderX

Conflicts of Interest

LMF, AG, ALC, LL, PH, NK, and PJB are codevelopers of the Uni Virtual Clinic-Lite intervention. LMF, AG, and NK were involved in the development of the original Uni Virtual Clinic intervention.

Multimedia Appendix 1
Screenshots of the Uni Virtual Clinic-Lite intervention.
[DOCX File , 1502 KB - jmir_v26i1e53598_app1.docx]

Multimedia Appendix 2
Descriptive statistics for uptake and engagement with each intervention module.
[DOCX File , 15 KB - jmir_v26i1e53598_app2.docx]

Multimedia Appendix 3
CONSORT-EHEALTH checklist.
[PDF File (Adobe PDF File), 3384 KB - jmir_v26i1e53598_app3.pdf]

References

1. de Girolamo G, Dagani J, Purcell R, Cocchi A, McGorry PD. Age of onset of mental disorders and use of mental health services: needs, opportunities and obstacles. Epidemiol Psychiatr Sci 2012 Mar;21(1):47-57. [doi: 10.1017/s2045796011000746] [Medline: 22670412]
2. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustün TB. Age of onset of mental disorders: a review of recent literature. Curr Opin Psychiatry 2007 Jul;20(4):359-364 [FREE Full text] [doi: 10.1097/YCO.0b013e32816ebc8c] [Medline: 17551351]
3. Cleary M, Walter G, Jackson D. "Not always smooth sailing": mental health issues associated with the transition from high school to college. Issues Ment Health Nurs 2011;32(4):250-254. [doi: 10.3109/01612840.2010.548906] [Medline: 21355760]
4. Deasy C, Coughlan B, Pironom J, Jourdan D, Mannix-McNamara P. Psychological distress and coping amongst higher education students: a mixed method enquiry. PLoS One 2014;9(12):e115193 [FREE Full text] [doi: 10.1371/journal.pone.0115193] [Medline: 25506825]
5. Eisenberg D, Hunt J, Speer N, Zivin K. Mental health service utilization among college students in the United States. J Nerv Ment Dis 2011 May;199(5):301-308. [doi: 10.1097/NMD.0b013e3182175123] [Medline: 21543948]
6. Farrer LM, Gulliver A, Bennett K, Fassnacht DB, Griffiths KM. Demographic and psychosocial predictors of major depression and generalised anxiety disorder in Australian university students. BMC Psychiatry 2016 Jul 15;16:241 [FREE Full text] [doi: 10.1186/s12888-016-0961-z] [Medline: 27422284]
7. Karyotaki E, Klein AM, Ciharova M, Bolinski F, Krijnen L, de Koning L, et al. Guided internet-based transdiagnostic individually tailored cognitive behavioral therapy for symptoms of depression and/or anxiety in college students: a randomized controlled trial. Behav Res Ther 2022 Mar;150:104028 [FREE Full text] [doi: 10.1016/j.brat.2021.104028] [Medline: 35066365]
8. Aristovnik A, Keržič D, Ravšelj D, Tomaževič N, Umek L. Impacts of the COVID-19 pandemic on life of higher education students: a global perspective. Sustainability 2020 Oct 13;12(20):8438. [doi: 10.3390/su12208438]
9. Appleby JA, King N, Saunders KE, Bast A, Rivera D, Byun J, et al. Impact of the COVID-19 pandemic on the experience and mental health of university students studying in Canada and the UK: a cross-sectional study. BMJ Open 2022 Jan 24;12(1):e050187 [FREE Full text] [doi: 10.1136/bmjopen-2021-050187] [Medline: 35074809]
10. Chakraborty P, Mittal P, Gupta MS, Yadav S, Arora A. Opinion of students on online education during the COVID-19 pandemic. Human Behav Emerg Technol 2021 Jul;3(3):357-365. [doi: 10.1002/hbe2.240]
11. Fruehwirth JC, Biswas S, Perreira KM. The Covid-19 pandemic and mental health of first-year college students: examining the effect of Covid-19 stressors using longitudinal data. PLoS One 2021 Mar 5;16(3):e0247999 [FREE Full text] [doi: 10.1371/journal.pone.0247999] [Medline: 33667243]
12. Auerbach RP, Alonso J, Axinn WG, Cuijpers P, Ebert DD, Green JG, et al. Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychol Med 2016 Oct;46(14):2955-2970 [FREE Full text] [doi: 10.1017/S0033291716001665] [Medline: 27484622]
13. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. WHO World Mental Health Surveys International College Student Project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
14. Leahy CM, Peterson RF, Wilson IG, Newbury JW, Tonkin AL, Turnbull D. Distress levels and self-reported treatment rates for medicine, law, psychology and mechanical engineering tertiary students: cross-sectional study. Aust N Z J Psychiatry 2010 Jul;44(7):608-615. [doi: 10.3109/00048671003649052] [Medline: 20560848]
15. Stallman HM. Psychological distress in university students: a comparison with general population data. Aust Psychol 2010 Nov 06;45(4):249-257. [doi: 10.1080/00050067.2010.482109]

XSL•FO
RenderX

16. Hamza CA, Ewing L, Heath NL, Goldstein AL. When social isolation is nothing new: a longitudinal study on psychological distress during COVID-19 among university students with and without preexisting mental health concerns. Can Psychol 2021;62(1):20-30. [doi: 10.1037/cap0000255]
17. Lanza ST, Whetzel CA, Linden-Carmichael AN, Newschaffer CJ. Change in college student health and well-being profiles as a function of the COVID-19 pandemic. PLoS One 2022 May 02;17(5):e0267724 [FREE Full text] [doi: 10.1371/journal.pone.0267724] [Medline: 35499988]
18. Deng J, Zhou F, Hou W, Silver Z, Wong CY, Chang O, et al. The prevalence of depressive symptoms, anxiety symptoms and sleep disturbance in higher education students during the COVID-19 pandemic: a systematic review and meta-analysis. Psychiatry Res 2021 Jul;301:113863 [FREE Full text] [doi: 10.1016/j.psychres.2021.113863] [Medline: 33984824]
19. Li Y, Wang A, Wu Y, Han N, Huang H. Impact of the COVID-19 pandemic on the mental health of college students: a systematic review and meta-analysis. Front Psychol 2021 Jul 14;12:669119 [FREE Full text] [doi: 10.3389/fpsyg.2021.669119] [Medline: 34335381]
20. Wang C, Wen W, Zhang H, Ni J, Jiang J, Cheng Y, et al. Anxiety, depression, and stress prevalence among college students during the COVID-19 pandemic: a systematic review and meta-analysis. J Am Coll Health 2023 Oct;71(7):2123-2130. [doi: 10.1080/07448481.2021.1960849] [Medline: 34469261]
21. McGorry PD, Purcell R, Goldstone S, Amminger GP. Age of onset and timing of treatment for mental and substance use disorders: implications for preventive intervention strategies and models of care. Curr Opin Psychiatry 2011 Jul;24(4):301-306. [doi: 10.1097/YCO.0b013e3283477a09] [Medline: 21532481]
22. Pedrelli P, Nyer M, Yeung A, Zulauf C, Wilens T. College students: mental health problems and treatment considerations. Acad Psychiatry 2015 Oct;39(5):503-511 [FREE Full text] [doi: 10.1007/s40596-014-0205-9] [Medline: 25142250]
23. Bruffaerts R, Mortier P, Kiekens G, Auerbach RP, Cuijpers P, Demyttenaere K, et al. Mental health problems in college freshmen: prevalence and academic functioning. J Affect Disord 2018 Jan 01;225:97-103 [FREE Full text] [doi: 10.1016/j.jad.2017.07.044] [Medline: 28802728]
24. Mojtabai R, Stuart EA, Hwang I, Eaton WW, Sampson N, Kessler RC. Long-term effects of mental disorders on educational attainment in the national comorbidity survey ten-year follow-up. Soc Psychiatry Psychiatr Epidemiol 2015 Oct;50(10):1577-1591 [FREE Full text] [doi: 10.1007/s00127-015-1083-5] [Medline: 26082040]
25. Ebert DD, Mortier P, Kaehlke F, Bruffaerts R, Baumeister H, Auerbach RP, et al. Barriers of mental health treatment utilization among first-year college students: first cross-national results from the WHO World Mental Health International College Student Initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1782 [FREE Full text] [doi: 10.1002/mpr.1782] [Medline: 31069905]
26. Ahuvia IL, Sung JY, Dobias ML, Nelson BD, Richmond LL, London B, et al. College student interest in teletherapy and self-guided mental health supports during the COVID-19 pandemic. J Am Coll Health 2024 Apr;72(3):940-946. [doi: 10.1080/07448481.2022.2062245] [Medline: 35427460]
27. Dunbar MS, Sontag-Padilla L, Kase CA, Seelam R, Stein BD. Unmet mental health treatment need and attitudes toward online mental health services among community college students. Psychiatr Serv 2018 May 01;69(5):597-600. [doi: 10.1176/appi.ps.201700402] [Medline: 29540117]
28. Montagni I, Tzourio C, Cousin T, Sagara JA, Bada-Alonzi J, Horgan A. Mental health-related digital use by university students: a systematic review. Telemed J E Health 2020 Feb;26(2):131-146. [doi: 10.1089/tmj.2018.0316] [Medline: 30888256]
29. Bolinski F, Boumparis N, Kleiboer A, Cuijpers P, Ebert DD, Riper H. The effect of e-mental health interventions on academic performance in university and college students: a meta-analysis of randomized controlled trials. Internet Interv 2020 Apr 23;20:100321 [FREE Full text] [doi: 10.1016/j.invent.2020.100321] [Medline: 32382515]
30. Conley CS, Durlak JA, Shapiro JB, Kirsch AC, Zahniser E. A meta-analysis of the impact of universal and indicated preventive technology-delivered interventions for higher education students. Prev Sci 2016 Aug;17(6):659-678. [doi: 10.1007/s11121-016-0662-3] [Medline: 27225631]
31. Farrer L, Gulliver A, Chan JK, Batterham PJ, Reynolds J, Calear A, et al. Technology-based interventions for mental health in tertiary students: systematic review. J Med Internet Res 2013 May 27;15(5):e101 [FREE Full text] [doi: 10.2196/jmir.2639] [Medline: 23711740]
32. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]
33. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
34. Becker TD, Torous JB. Recent developments in digital mental health interventions for college and university students. Curr Treat Options Psych 2019 Jun 14;6(3):210-220. [doi: 10.1007/s40501-019-00178-8]
35. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. Mental disorder comorbidity and suicidal thoughts and behaviors in the World Health Organization World Mental Health Surveys International College Student initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1752 [FREE Full text] [doi: 10.1002/mpr.1752] [Medline: 30450753]

XSL•FO
RenderX

36. Day V, McGrath PJ, Wojtowicz M. Internet-based guided self-help for university students with anxiety, depression and stress: a randomized controlled clinical trial. Behav Res Ther 2013 Jul;51(7):344-351. [doi: 10.1016/j.brat.2013.03.003] [Medline: 23639300]
37. Musiat P, Conrod P, Treasure J, Tylee A, Williams C, Schmidt U. Targeted prevention of common mental health disorders in university students: randomised controlled trial of a transdiagnostic trait-focused web-based intervention. PLoS One 2014 Apr 15;9(4):e93621 [FREE Full text] [doi: 10.1371/journal.pone.0093621] [Medline: 24736388]
38. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022 Jan;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
39. Melnyk BM, Amaya M, Szalacha LA, Hoying J, Taylor T, Bowersox K. Feasibility, acceptability, and preliminary effects of the COPE online cognitive-behavioral skill-building program on mental health outcomes and academic performance in freshmen college students: a randomized controlled pilot study. J Child Adolesc Psychiatr Nurs 2015 Aug;28(3):147-154. [doi: 10.1111/jcap.12119] [Medline: 26268362]
40. Rahmadiana M, Karyotaki E, Schulte M, Ebert DD, Passchier J, Cuijpers P, et al. Transdiagnostic internet intervention for Indonesian university students with depression and anxiety: evaluation of feasibility and acceptability. JMIR Ment Health 2021 Mar 05;8(3):e20036 [FREE Full text] [doi: 10.2196/20036] [Medline: 33666553]
41. Andersson G, Titov N. Advantages and limitations of internet-based interventions for common mental disorders. World Psychiatry 2014 Feb;13(1):4-11 [FREE Full text] [doi: 10.1002/wps.20083] [Medline: 24497236]
42. Farrer LM, Gulliver A, Katruss N, Fassnacht DB, Kyrios M, Batterham PJ. A novel multi-component online intervention to improve the mental health of university students: randomised controlled trial of the Uni Virtual Clinic. Internet Interv 2019 Aug 28;18:100276 [FREE Full text] [doi: 10.1016/j.invent.2019.100276] [Medline: 31890625]
43. Batterham PJ, Calear AL, Farrer L, Gulliver A, Kurz E. Efficacy of a transdiagnostic self-help internet intervention for reducing depression, anxiety, and suicidal ideation in adults: randomized controlled trial. J Med Internet Res 2021 Jan 22;23(1):e22698 [FREE Full text] [doi: 10.2196/22698] [Medline: 33480860]
44. Özer Ö, Köksal B, Altinok A. Understanding university students' attitudes and preferences for internet-based mental health interventions. Internet Interv 2024 Feb 06;35:100722 [FREE Full text] [doi: 10.1016/j.invent.2024.100722] [Medline: 38356613]
45. Batterham PJ, Sunderland M, Carragher N, Calear AL, Mackinnon AJ, Slade T. The Distress Questionnaire-5: population screener for psychological distress was more accurate than the K6/K10. J Clin Epidemiol 2016 Mar;71:35-42. [doi: 10.1016/j.jclinepi.2015.10.005] [Medline: 26464194]
46. Batterham PJ, Sunderland M, Calear AL, Davey CG, Christensen H, Teesson M, et al. Developing a roadmap for the translation of e-mental health services for depression. Aust N Z J Psychiatry 2015 Sep;49(9):776-784. [doi: 10.1177/0004867415582054] [Medline: 25907269]
47. What's this fuss about true randomness? RANDOM.ORG. URL: https://www.random.org/ [accessed 2024-07-08]
48. Farrer LM, Gulliver A, Katruss N, Bennett K, Bennett A, Ali K, et al. Development of the Uni Virtual Clinic: an online programme for improving the mental health of university students. Br J Guid Couns 2020 Feb 20;48(3):333-346. [doi: 10.1080/03069885.2020.1729341]
49. Glozier N, Christensen H, Naismith S, Cockayne N, Donkin L, Neal B, et al. Internet-delivered cognitive behavioural therapy for adults with mild to moderate depression and high cardiovascular disease risks: a randomised attention-controlled trial. PLoS One 2013;8(3):e59139 [FREE Full text] [doi: 10.1371/journal.pone.0059139] [Medline: 23555624]
50. Keeping bones strong and healthy. NIH News in Health. 2010 Feb. URL: https://newsinhealth.nih.gov/2010/02/keeping-bones-strong-healthy [accessed 2024-07-08]
51. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
52. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
53. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
54. Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
55. Sunderland M, Batterham PJ, Calear AL, Carragher N. The development and validation of static and adaptive screeners to measure the severity of panic disorder, social anxiety disorder, and obsessive compulsive disorder. Int J Methods Psychiatr Res 2017 Dec;26(4):e1561 [FREE Full text] [doi: 10.1002/mpr.1561] [Medline: 28370638]
56. Batterham PJ, Sunderland M, Slade T, Calear AL, Carragher N. Assessing distress in the community: psychometric properties and crosswalk comparison of eight measures of psychological distress. Psychol Med 2018 Jun;48(8):1316-1324. [doi: 10.1017/S0033291717002835] [Medline: 28967345]

XSL•FO
RenderX

57. Fife DA, Mendoza JL, Terry R. The assessment of reliability under range restriction: a comparison of α, ω, and test–retest reliability for dichotomous data. Educ Psychol Meas 2012 Apr 19;72(5):862-888. [doi: 10.1177/0013164411430225]
58. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. BMC Med Educ 2010 Jun 02;10:40 [FREE Full text] [doi: 10.1186/1472-6920-10-40] [Medline: 20525220]
59. Tylka TL, Wood-Barcalow NL. The Body Appreciation Scale-2: item refinement and psychometric evaluation. Body Image 2015 Jan;12:53-67. [doi: 10.1016/j.bodyim.2014.09.006] [Medline: 25462882]
60. Schmidt S, Mühlan H, Power M. The EUROHIS-QOL 8-item index: psychometric results of a cross-cultural field study. Eur J Public Health 2006 Aug;16(4):420-428. [doi: 10.1093/eurpub/cki155] [Medline: 16141303]
61. da Rocha NS, Power MJ, Bushnell DM, Fleck MP. The EUROHIS-QOL 8-item index: comparative psychometric properties to its parent WHOQOL-BREF. Value Health 2012 May;15(3):449-457 [FREE Full text] [doi: 10.1016/j.jval.2011.11.035] [Medline: 22583455]
62. Wellbeing measures in primary health care/the DepCare Project: report on a WHO meeting: Stockholm, Sweden, 12–13 February 1998. World Health Organization. 1998. URL: https://iris.who.int/handle/10665/349766 [accessed 2024-07-08]
63. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
64. Keetharuth AD, Brazier J, Connell J, Bjorner JB, Carlton J, Taylor Buck E, et al. Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. Br J Psychiatry 2018 Jan;212(1):42-49 [FREE Full text] [doi: 10.1192/bjp.2017.10] [Medline: 29433611]
65. Schwarzer R, Jerusalem M. Generalized self-efficacy scale. In: Weinman J, Wright S, Johnston M, editors. Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs. Windsor, UK: NFER-Nelson Publisher; 1995:35-37.
66. Luszczynska A, Scholz U, Schwarzer R. The general self-efficacy scale: multicultural validation studies. J Psychol 2005;139(5):439-457. [doi: 10.3200/jrlp.139.5.439-457]
67. Scholz U, Gutiérrez Doña B, Sud S, Schwarzer R. Is general self-efficacy a universal construct? Psychometric findings from 25 countries. Eur J Psychol Assess 2002 Sep;18(3):242-251. [doi: 10.1027//1015-5759.18.3.242]
68. Solberg VS, O'Brien K, Villareal P, Kennel R, Davis B. Self-efficacy and Hispanic college students: validation of the college self-efficacy instrument. Hisp J Behav Sci 1993 Feb;15(1):80-95. [doi: 10.1177/07399863930151004]
69. Barry CL, Finney SJ. A psychometric investigation of the college self-efficacy inventory. In: Proceedings of the Annual Meeting of the Northeastern Educational Research Association. 2007 Presented at: NERA 2007; October 17-19, 2007; Rocky Hill, CT.
70. Gore PAJ, Leuwerke WC, Turley SE. A psychometric study of the college self-efficacy inventory. J Coll Stud Retent Res Theory Pract 2005;7(3):227-244. [doi: 10.2190/5CQF-F3P4-2QAC-GNVJ]
71. Wilson CJ, Deane FP, Ciarrochi J, Rickwood D. Measuring help-seeking intentions: properties of the general help-seeking questionnaire. Can J Couns 2005;39(1):15-28.
72. IBM SPSS statistics for windows, version 28.0. IBM Corporation. 2021. URL: https://www.ibm.com/products/spss-statistics [accessed 2024-07-08]
73. Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York, NY: Springer; 2013.
74. D'Adamo L, Paraboschi L, Grammer AC, Fennig M, Graham AK, Yaeger LH, et al. Reach and uptake of digital mental health interventions based on cognitive-behavioral therapy for college students: a systematic review. J Behav Cogn Ther 2023 Jun;33(2):97-117 [FREE Full text] [doi: 10.1016/j.jbct.2023.05.002] [Medline: 37724304]
75. Batterham PJ, Calear AL, Farrer L, McCallum SM, Cheng VW. FitMindKit: randomised controlled trial of an automatically tailored online program for mood, anxiety, substance use and suicidality. Internet Interv 2017 Aug 23;12:91-99 [FREE Full text] [doi: 10.1016/j.invent.2017.08.002] [Medline: 30135773]
76. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
77. Karyotaki E, Cuijpers P, Albor Y, Alonso J, Auerbach RP, Bantjes J, et al. Sources of stress and their associations with mental disorders among college students: results of the World Health Organization World Mental Health Surveys international college student initiative. Front Psychol 2020 Jul 30;11:1759 [FREE Full text] [doi: 10.3389/fpsyg.2020.01759] [Medline: 32849042]
78. Melcher J, Camacho E, Lagan S, Torous J. College student engagement with mental health apps: analysis of barriers to sustained use. J Am Coll Health 2022;70(6):1819-1825. [doi: 10.1080/07448481.2020.1825225] [Medline: 33048626]

Abbreviations

BAS-2: Body Appreciation Scale-2
CBT: cognitive behavioral therapy
CSEI: College Self-Efficacy Inventory
DQ5: Distress Questionnaire

XSL•FO
RenderX

EUROHIS-8: European Health Interview Survey-8
GAD-7: Generalized Anxiety Disorder Scale-7
PHQ-9: Patient Health Questionnaire-9
UVC-Lite: Uni Virtual Clinic-Lite

Edited by T McCall; submitted 12.10.23; peer-reviewed by A Hildebrand, L D'Adamo; comments to author 10.02.24; revised version received 27.02.24; accepted 24.05.24; published 13.08.24.

Please cite as:
Farrer LM, Jackson HM, Gulliver A, Calear AL, Leach L, Hasking P, Katruss N, Batterham PJ
A Transdiagnostic Video-Based Internet Intervention (Uni Virtual Clinic-Lite) to Improve the Mental Health of University Students: Randomized Controlled Trial
J Med Internet Res 2024;26:e53598
URL: https://www.jmir.org/2024/1/e53598
doi:10.2196/53598
PMID:39137012

©Louise M Farrer, Hayley M Jackson, Amelia Gulliver, Alison L Calear, Liana Leach, Penelope Hasking, Natasha Katruss, Philip J Batterham. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of an Online Discussion Forum on Self-Guided Internet-Delivered Cognitive Behavioral Therapy for Public Safety Personnel: Randomized Trial

Hugh C McCall[1,2], MA; Heather D Hadjistavropoulos[1,2], PhD

[1]Department of Psychology, University of Regina, Regina, SK, Canada
[2]PSPNET, University of Regina, Regina, SK, Canada

Corresponding Author:
Heather D Hadjistavropoulos, PhD
Department of Psychology
University of Regina
3737 Wascana Pkwy
Regina, SK, S4S 0A2
Canada
Phone: 1 306 585 5133
Email: Heather.Hadjistavropoulos@uregina.ca

Abstract

Background: Internet-delivered cognitive behavioral therapy (ICBT) is an effective and accessible treatment for various mental health concerns. ICBT has shown promising treatment outcomes among public safety personnel (PSP), who experience high rates of mental health problems and face barriers to accessing other mental health services. Client engagement and clinical outcomes are better in ICBT with therapist guidance, but ICBT is easier to implement on a large scale when it is self-guided. Therefore, it is important to identify strategies to improve outcomes and engagement in self-guided ICBT and other self-guided digital mental health interventions. One such strategy is the use of online discussion forums to provide ICBT clients with opportunities for mutual social support. Self-guided interventions accompanied by online discussion forums have shown excellent treatment outcomes, but there is a need for research experimentally testing the impact of online discussion forums in ICBT.

Objective: We aimed to evaluate a transdiagnostic, self-guided ICBT intervention tailored specifically for PSP (which had not previously been assessed), assess the impact of adding a therapist-moderated online discussion forum on outcomes, and analyze participants' feedback to inform future research and implementation efforts.

Methods: In this randomized trial, we randomly assigned participating PSP (N=107) to access an 8-week transdiagnostic, self-guided ICBT course with or without a built-in online discussion forum. Enrollment and participation were entirely web-based. We assessed changes in depression, anxiety, and posttraumatic stress as well as several secondary outcome measures (eg, treatment engagement and satisfaction) using questionnaires at the pre-enrollment, 8-week postenrollment, and 20-week postenrollment time points. Mixed methods analyses included multilevel modeling and qualitative content analysis.

Results: Participants engaged minimally with the forum, creating 9 posts. There were no differences in treatment outcomes between participants who were randomly assigned to access the forum (56/107, 52.3%) and those who were not (51/107, 47.7%). Across conditions, participants who reported clinically significant symptoms during enrollment showed large and statistically significant reductions in symptoms ($P<.05$ and $d>0.97$ in all cases). Participants also showed good treatment engagement and satisfaction, with 43% (46/107) of participants fully completing the intervention during the course of the study and 96% (79/82) indicating that the intervention was worth their time.

Conclusions: Previous research has shown excellent clinical outcomes for self-guided ICBT accompanied by discussion forums and good engagement with those forums. Although clinical outcomes in our study were excellent across conditions, engagement with the forum was poor, in contrast to previous research. We discuss several possible interpretations of this finding (eg, related to the population under study or the design of the forum). Our findings highlight a need for more research evaluating the impact of online discussion forums and other strategies for improving outcomes and engagement in self-guided ICBT and other digital mental health interventions.

Trial Registration: ClinicalTrials.gov NCT05145582; https://clinicaltrials.gov/study/NCT05145582

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59699) doi:10.2196/59699

KEYWORDS

internet; cognitive behavioral therapy; forum; persuasive design; generalized anxiety disorder; major depressive disorder; posttraumatic stress disorder; public safety personnel

Introduction

Background

Internet-delivered cognitive behavioral therapy (ICBT) is a psychological treatment in which clients learn evidence-based cognitive behavioral therapy treatment strategies via web-based modules, often with therapist guidance via email or phone. Hundreds of randomized trials have demonstrated that ICBT is similarly effective to face-to-face psychotherapies for treating depression and anxiety [1]. There are 2 key advantages of ICBT and other digital mental health interventions (DMHIs): the ability of clients to access them privately and conveniently at practically any time and location [2-4] and the tendency for DMHIs to require less therapist time per client than traditional face-to-face psychotherapies [2,4].

ICBT is sometimes offered in a purely self-guided format (ie, without a therapist). Meta-analyses have shown that self-guided ICBT is at least slightly less effective than therapist-guided ICBT; this assertion is based on differences in effect sizes observed in separate meta-analyses of guided [1] and self-guided [5,6] ICBT, meta-analyses including subgroup analyses of both guided and self-guided ICBT [7-9], a meta-regression in which human contact predicted more favorable ICBT outcomes [10], a meta-analysis of randomized trials directly comparing guided and self-guided ICBT [11], and an individual-participant meta-analysis evaluating both guided and self-guided ICBT for depression [12]. In addition, client engagement with self-guided ICBT and other self-guided DMHIs tends to be low [13-15], particularly in real-world observational research, where completion rates were found in one systematic review to range from 0.5% to 28.6% [14]. However, self-guided ICBT and other self-guided DMHIs can be implemented on a large scale with minimal human or financial inputs required [7,16], making them cost-effective [17] and—many have argued—justifiable despite their tendency to be less effective than therapist-guided DMHIs or face-to-face psychotherapies [16,18].

There appears to be a growing consensus that DMHIs can be designed to be more engaging for clients [5,19-21], which may have particular implications for mitigating the problem of low engagement in self-guided DMHIs. The persuasive system design framework [22] describes 28 specific design principles for improving engagement divided into four categories: (1) primary task support principles, which facilitate completion of treatment tasks (eg, tailoring content for specific user groups and presenting complex tasks in a series of simple steps); (2) dialogue support principles, which facilitate dialogue between an intervention and its users (eg, automated praise, reminders, or virtual rewards); (3) system credibility support principles, which help ensure that users perceive interventions as credible (eg, endorsements from credible third parties and inclusion of experts in the design process); and (4) social support principles, which enable users to support each other in their use of an intervention (eg, opportunities for users to support and learn from each other). Systematic reviews and meta-analyses have shown that persuasive design characteristics can predict treatment engagement [20] and symptom change [5], but research assessing the impact of specific persuasive design principles is limited.

Online discussion forums facilitate social support principles of persuasive design. Previous research suggests that they may help support engagement and outcomes in self-guided DMHIs. For example, participants in forum-only control conditions across several studies have demonstrated promising outcomes [23-26], prompting the authors of one paper to conclude that they "could be regarded as an intervention" in and of themselves [23]. In total, 3 randomized trials have shown that self-guided ICBT [27,28] or self-guided bibliotherapy [3] accompanied by an online discussion forum exhibited equivalent outcomes to those of therapist-guided ICBT. Another randomized trial experimentally demonstrated that adding an online discussion forum to guided ICBT improved engagement [29]. Together, these studies suggest that forums could help improve engagement and outcomes in ICBT—potentially bridging the engagement and efficacy gap between guided and self-guided ICBT—but there are no previous randomized trials experimentally evaluating the impact of a forum on engagement and outcomes in self-guided ICBT.

In 2019, a clinical research unit called PSPNET was founded to develop, deliver, and conduct research on free ICBT interventions tailored specifically for Canadian first responders and other public safety personnel (PSP), who frequently experience potentially psychologically traumatic events [30], report high rates of mental health problems [31], and face unique barriers to accessing mental health care (eg, stigma within their workplaces) [32,33]. At the time this study was conducted, PSPNET offered 2 therapist-guided ICBT interventions to Canadian PSP—one transdiagnostic and the other posttraumatic stress disorder (PTSD) specific—both of which have shown promising outcomes with respect to symptom change, engagement, and treatment satisfaction [34-36]. However, at the time of this study, PSPNET was unable to offer guided ICBT services to PSP in all Canadian provinces and territories, highlighting an opportunity to develop a self-guided ICBT intervention that could be delivered with minimal resources required to PSP anywhere in Canada. No previous research has evaluated self-guided ICBT tailored specifically for PSP.

Objectives and Hypotheses

Broadly speaking, we designed this study to evaluate self-guided ICBT among Canadian PSP while addressing several questions concerning the role of online discussion forums in self-guided ICBT. Specifically, we sought to address the following four objectives:

XSL•FO
RenderX

1. To evaluate transdiagnostic, self-guided ICBT tailored for PSP with respect to treatment engagement, outcomes, and satisfaction.
2. To evaluate whether adding an online discussion forum to transdiagnostic, self-guided ICBT tailored for PSP would improve engagement and outcomes.
3. To evaluate whether participant engagement in the online discussion forum would moderate treatment outcomes.
4. To conduct a mixed methods analysis of participant feedback on the discussion forum.

We hypothesized that participants in both conditions would experience at least small to moderate reductions in symptoms of depression, anxiety, and posttraumatic stress, consistent with recent meta-analytic evidence [5,6]. Second, we hypothesized that participants randomly assigned to receive access to an online discussion forum would show greater engagement and more favorable treatment outcomes than those randomly assigned to receive ICBT without a discussion forum.

Methods

Study Design

We used a randomized trial design with 2 conditions: an ICBT plus peer support forum condition and an ICBT-only condition. Participants in both conditions were given free access to a self-guided ICBT program called the *Self-Guided PSP Wellbeing Course*. For participants in the ICBT plus peer support forum condition, but not for those in the ICBT-only condition, this ICBT course included a built-in online discussion forum. Participants were not blinded to their own condition as it is not possible to hide therapy content from those receiving therapy, but the experimental manipulation was described only in general terms (ie, without reference to forums) such that participants were blind to how the condition to which they were assigned differed from the condition to which they were not assigned. We adopted a simple randomization approach [37], which we implemented via a random number generator with a 1:1 ratio. We registered the methodological protocol for this research on ClinicalTrials.gov (ID NCT05145582) and made 2 deviations from it. First, we removed the Sheehan Disability Scale [38] from our planned outcome measures because we were unable to obtain permission to use it. Second, we ultimately carried out our primary quantitative analyses using multilevel modeling (MLM) instead of generalized estimating equations, as we had originally planned, because a paper was published during the course of this research that provided a compelling rationale and detailed recommendations for using MLM in treatment-control pretest-posttest-follow-up study designs [39]. We followed the CONSORT (Consolidated Standards of Reporting Trials) guidelines [40] in reporting the findings of this research. This research was conducted within the context of author HCM's doctoral dissertation, and we would refer interested readers to this dissertation (expected to be publicly available in or around October 2024) for further details about this research.

Setting

This study was conducted in Canada, where publicly funded mental health services have not met public demand, leading many Canadians to access private mental health care instead [41]. Canadians have access to DMHIs through various Canadian organizations [42]. There are also thousands of mental health–related phone apps and websites available in Canada and other countries [43], but many of these services are not empirically supported. All research activities pertaining to this study were carried out at the University of Regina in Saskatchewan, Canada.

Participants, Recruitment, and Enrollment

A power analysis indicated that a minimum of 110 participants would be required to achieve adequate power to detect moderate between-group differences (see Multimedia Appendix 1 [3,5,8,27,34,44-47] for details on our sample size planning). Participants were informed about this research via paid social media advertisements (ie, Twitter [subsequently rebranded X] and Facebook), emails forwarded to PSP by leaders of PSP organizations, presentations to PSP organizations by author HCM and PSPNET's clinicians, and word of mouth. To be eligible to take part, prospective participants were required to self-report (1) being aged ≥18 years; (2) residing in Canada; (3) working, volunteering, or having previously worked or volunteered as a PSP; (4) being able to access the internet via a computer; and (5) not experiencing significant ongoing concerns related to alcohol or drug use, psychotic symptoms, or manic symptoms.

Prospective participants accessed this study through a web page on PSPNET's website, which provided information about the study. Upon reviewing a consent form and consenting to participate, they accessed a series of eligibility screening questionnaires through Qualtrics (Qualtrics International Inc). We contacted prospective participants by email to inform them of their eligibility, and eligible participants were asked to confirm their intent to take part in the study, after which they were randomly assigned to 1 of the 2 conditions and provided with a temporary password to access the version of the *Self-GuidedPSP Wellbeing Course* (ie, with or without the peer support forum) to which they had been assigned. All randomization and enrollment procedures were carried out by author HCM and research assistant Julia Gregory (see the *Acknowledgments* section). Recruitment took place between December 6, 2021, and September 26, 2022.

Measures

Primary Outcome Measures

Patient Health Questionnaire–9

The Patient Health Questionnaire–9 (PHQ-9) is a psychometrically sound, 9-item questionnaire assessing depressive symptoms [48,49]. Possible total scores range from 0 to 27, and a score of ≥10 suggests that a respondent's symptoms are clinically significant [50].

Generalized Anxiety Disorder–7

The Generalized Anxiety Disorder–7 (GAD-7) is a 7-item questionnaire assessing generalized anxiety that has demonstrated strong psychometric properties [49,51]. Total scores can range from 0 to 21, with a score of ≥10 suggesting clinically significant symptoms [49,51].

XSL•FO
RenderX

PTSD Checklist for DSM-5

PTSD Checklist for DSM-5 (PCL-5) is a psychometrically sound questionnaire assessing PTSD symptoms [52]. Responses to its 20 items sum to a total score ranging from 0 to 80, and a score of ≥33 indicates that a respondent likely meets criteria for a PTSD diagnosis [53,54].

Secondary Outcome Measures

Brief Resilience Scale

The Brief Resilience Scale (BRS) is a 6-item questionnaire measure of resilience that has shown good psychometric properties [55,56]. Each item has 5 response options with associated numerical values ranging from 1 to 5, and a higher mean score across items indicates greater resilience.

Flourishing Scale

The Flourishing Scale (FS) is an 8-item questionnaire assessing flourishing across various domains of life (eg, relationships, meaning and purpose, and feeling of competence). It has demonstrated good psychometric properties [57,58]. Total scores can range from 8 to 56, with greater scores indicating a greater degree of flourishing.

Treatment Satisfaction Questionnaire

We administered a bespoke questionnaire designed to assess treatment satisfaction and solicit feedback on the *Self-Guided PSP Wellbeing Course* through a mix of yes or no, Likert-scale, and open-ended text response items. For participants in the ICBT plus peer support forum condition, the Treatment Satisfaction Questionnaire included several additional items pertaining to the forum. Open-ended questions about both the course and the forum were designed to solicit both positive and constructive feedback.

Adapted Session Rating Scale

We administered a modified version of the Session Rating Scale (SRS), a 4-item questionnaire originally designed to assess client perspectives on the quality of the therapeutic alliance in face-to-face therapy [59]. It has good psychometric qualities [60,61]. Items are rated on a 7-point Likert scale from 0 (*absolutely disagree*) to 6 (*absolutely agree*) and assess the therapeutic bond, goal agreement, task agreement, and overall alliance quality. Following an approach taken in another study [62], we adapted the SRS to measure patient-program alliance.

Program Use Questionnaire

We administered a brief bespoke questionnaire assessing engagement with the *Self-Guided PSP Wellbeing Course* and, if applicable, the peer support forum. Specifically, this questionnaire was designed to assess effort put into the course; the perceived helpfulness of the course; and, if applicable, use and perceived helpfulness of the peer support forum. Program use patterns were also assessed via automatic collection of program use data (eg, number of lessons and additional resources accessed).

Health Service Use Questionnaire

We also administered a bespoke questionnaire to assess use of health care services for mental health challenges during eligibility screening and at 8 and 20 weeks after enrollment. In the interest of brevity, and because this questionnaire is peripheral to the primary objectives of this study, we do not describe the outcomes of this questionnaire in this paper.

Pre-Enrollment Measures

During eligibility screening, we administered a bespoke participant information questionnaire assessing demographic, occupational, and clinical characteristics; an ICBT feedback questionnaire assessing pre-enrollment knowledge and attitudes toward ICBT; the Credibility/Expectancy Questionnaire (CEQ) [63]; the Alcohol Use Disorders Identification Test [64]; and the Drug Use Disorders Identification Test [65].

Administration of Measures

During eligibility screening, we administered the PHQ-9, GAD-7, PCL-5, BRS, FS, participant information questionnaire, ICBT feedback questionnaire, CEQ, Alcohol Use Disorders Identification Test, and Drug Use Disorders Identification Test. At 2, 4, 6, and 8 weeks after enrollment, we administered the Program Use Questionnaire. At 8 weeks after enrollment, we also administered the PHQ-9, GAD-7, PCL-5, BRS, FS, and Treatment Satisfaction Questionnaire. At 20 weeks after enrollment, we administered the PHQ-9, GAD-7, PCL-5, BRS, and FS. Our research team encouraged participants to complete the questionnaires via emails and phone calls but did not urge participants to use the *Self-Guided PSP Wellbeing Course*.

Intervention

The *Self-GuidedPSP Wellbeing Course* is an 8-week self-guided, transdiagnostic ICBT program that can be accessed through a web browser. It includes 5 core lessons, each consisting of a welcome video, a series of slides with instructive text and diagrams, an audio file covering the same clinical content as the lesson slides, illustrative case stories, frequently asked questions, downloadable homework activities called “DIY Guides,” and quotes from previous clients. These lessons included an introduction to the cognitive behavioral model and psychoeducation to help participants recognize and understand their symptoms (lesson 1); skills to help participants recognize and challenge unhelpful thoughts (lesson 2); skills for managing physiological underarousal and overarousal symptoms (lesson 3); skills for managing behavioral symptoms (lesson 4); and strategies for maintaining treatment gains, setting goals, and preventing future relapses (lesson 5). The course also included 14 additional resources covering a wide range of topics (eg, assertiveness, physical pain, and intimate relationships) and automated email reminders to encourage engagement.

The *Self-GuidedPSP Wellbeing Course* is effectively a self-guided version of a previously developed therapist-guided ICBT course called the *PSP Wellbeing Course* [34,36]; aside from the provision of therapist guidance in the latter but not the former, the courses are practically identical. PSPNET developed the *PSP Wellbeing Course* by tailoring an existing ICBT program called the *Wellbeing Course* to meet the needs of Canadian PSP based on feedback provided by Canadian PSP in a series of interviews, focus groups, and questionnaires [33,66]. The original *Wellbeing Course* was initially developed by the eCentreClinic at Macquarie University, Australia, and

XSL•FO
RenderX

has since shown excellent outcomes among the general population in Australia [67] and Canada [68].

Discussion Forum

The peer support forum was built into the *Self-Guided PSP Wellbeing Course*. It included 11 sections (eg, one for each of the 5 lessons, one for discussing families and relationships, and one for discussing workplace issues). It was monitored daily and moderated as required each business day by author HCM, who posed questions to spark discussion and responded to participants' posts.

Ethical Considerations

This study was approved by the University of Regina Research Ethics Board (file 2021-130). Before taking part, all participants were provided with an informed consent form, which described the following: the objectives of the research, the research team, what participation would entail (ie, the intervention and questionnaires), possible risks and benefits of participating, project funding, considerations regarding concurrent mental health treatments, right to withdraw, limits to confidentiality, risks to privacy, precautions to improve security of participant information (both PSPNET's precautions and precautions that participants could take), uses of participants' data (ie, eligibility determination and research), information on accessing research results, a statement indicating that participants would not be compensated for taking part, and an invitation to contact our team with any questions or concerns. All participant data were deidentified before analysis. Due to ethical concerns related to the exclusion of individuals reporting suicidal ideation from DMHI research [69], we tried to refer prospective participants reporting suicidal ideation to more intensive services and clarified that the *Self-Guided PSP Wellbeing Course* is not a crisis service, but we allowed them to participate if they met the eligibility criteria.

Analyses

Quantitative Data Analyses

We carried out all quantitative analyses using SPSS (version 28; IBM Corp). We did not statistically test for group differences in pre-enrollment variables as it is not meaningful to test the probability that group differences occurred by chance when it is already known—due to random assignment—that they did [70]. Instead, we inspected the magnitude of group differences and planned to conduct sensitivity analyses to assess the impact of marked differences should we observe any. We compared changes in scores on the PHQ-9, GAD-7, PCL-5, FS, and BRS across conditions using an MLM approach recommended by Sharpe and Cribbie [39]. We used an intention-to-treat approach [71] including all participants in the analyses, and we accounted for missing data using the restricted maximum likelihood estimation method, which previous research suggests is preferable to maximum likelihood estimation for MLM when random effects are included [72]. Each model was run using a random intercept and fixed effects of group, time (as a categorical variable), and the interaction between group and time. We used a variance components covariance structure [73]. We also produced a *G* matrix for each model consistent with the recommendations by Sharpe and Cribbie [39]. We used scatterplots and histograms to test the assumptions of linearity, homoscedasticity of residuals, and normality of residuals [74]. For each of the 5 outcome variables, we conducted one model for the entire sample and one model for the subset of participants with clinically significant scores at the pre-enrollment time point, which we defined using established cutoff scores for the PHQ-9 (≥10), GAD-7 (≥10), or PCL-5 (≥33) and scores in the lower 3 quartiles on the FS (<48) and BRS (<4.0). Therefore, we ran 10 models in total.

In each of the 10 MLM models, we conducted 5 contrasts. In total, 2 contrasts were designed to assess for interactions between group and time—that is, to identify any differences between groups with respect to changes in dependent variables over time—including one contrast for the period between the pre-enrollment time point and 8 weeks after beginning treatment and one for the period between the pre-enrollment time point and 20 weeks after beginning treatment. We collapsed the 2 groups for 3 additional contrasts to determine whether changes in questionnaire scores over time were statistically significant—one contrast for the period between the pre-enrollment time point and 8 weeks after beginning treatment, one for the period between the pre-enrollment time point and 20 weeks after beginning treatment, and one for the period between 8 weeks and 20 weeks after beginning treatment. These latter 3 contrasts were designed to assess whether participants in the *Self-Guided PSP Wellbeing Course* experienced significant changes in their mental health.

Finally, in each of the 10 MLM models, we investigated the effects of five covariates on changes in questionnaire scores over time: (1) the number of lessons that participants accessed, (2) the number of additional resources that participants accessed, (3) CEQ credibility scores, (4) CEQ expectancy scores, and (5) gender. These analyses are not central to the objectives of this study but may be of interest to some readers; accordingly, a rationale for the inclusion of these specific covariates, methods and results pertaining to our covariate analyses, and a brief discussion of the findings of those analyses are shown in Multimedia Appendix 2 [63,75-80].

In addition to the MLM models, we used 2-tailed independent-sample *t* tests and chi-square tests to assess for group differences in treatment satisfaction and program use. Upon observing possible group differences in rates of questionnaire completion, we conducted additional (non-prespecified) chi-square tests to evaluate their significance.

Qualitative Data Analyses

We conducted qualitative analyses using a content analysis approach to explore participant feedback on the peer support forum and the *Self-Guided PSP Wellbeing Course* in general [81]. After removing identifying information from the data, author HCM identified categories using a descriptive, inductive approach and grouped those categories into overarching themes. Given the relatively small amount of data, this was carried out using an Excel (Microsoft Corp) spreadsheet. The initial codebook was refined through meeting with author HDH and Dr Janine Beahm (see the *Acknowledgments* section).

XSL•FO
RenderX

It is a common practice for researchers using qualitative methods to engage in reflexivity, which is a practice of reflection on how the researchers' positionality might affect the process or outcomes of qualitative research [82]. Being neither PSP nor ICBT clients, the authors do not identify as members of the population under study, potentially granting the authors a degree of neutrality in describing participants' reported experiences but also potentially impeding their ability to fully understand those experiences [82]. In addition, the authors held certain attitudes and beliefs (eg, that ICBT can be helpful for many people and that forums may be able to enhance ICBT) that may have influenced the process and outcomes of this research. Nevertheless, we endeavored to minimize the risk of bias in this research by (1) including neutrally worded questions to solicit both positive and negative feedback; (2) conducting content analysis as descriptively as possible and avoiding even minor inferences and assumptions; (3) separating qualitative data from other data that could cause bias in coding (eg, demographic and clinical characteristics) before analysis; and (4) involving 3 researchers, as noted previously, in checking the accuracy of our coding.

Results

Participants

Of the 188 prospective participants who completed the web-based screening, 153 (81.4%) were enrolled in the study and randomized, and 107 (56.9%) were included in our analyses. A flowchart displaying enrollment, program use, and questionnaire completion is shown in Figure 1. Of note, Figure 1 shows that 36 participants in the ICBT-only condition completed symptom measures at 20 weeks after enrollment; one of these participants completed the PHQ-9 and GAD-7 but not the PCL-5. Participant characteristics are shown in Table 1. Chi-square tests evidenced that the difference between conditions with respect to the proportion of participants who completed posttreatment questionnaires was statistically significant at 8 weeks (n=107, χ^2_1=6.4, *P*=.01) but not at 20 weeks (n=107, χ^2_1=0.5, *P*=.47).

XSL•FO
RenderX

Figure 1. Flowchart displaying enrollment, program use, and questionnaire completion. ICBT: Internet-delivered cognitive behavioral therapy; PSP: public safety personnel.

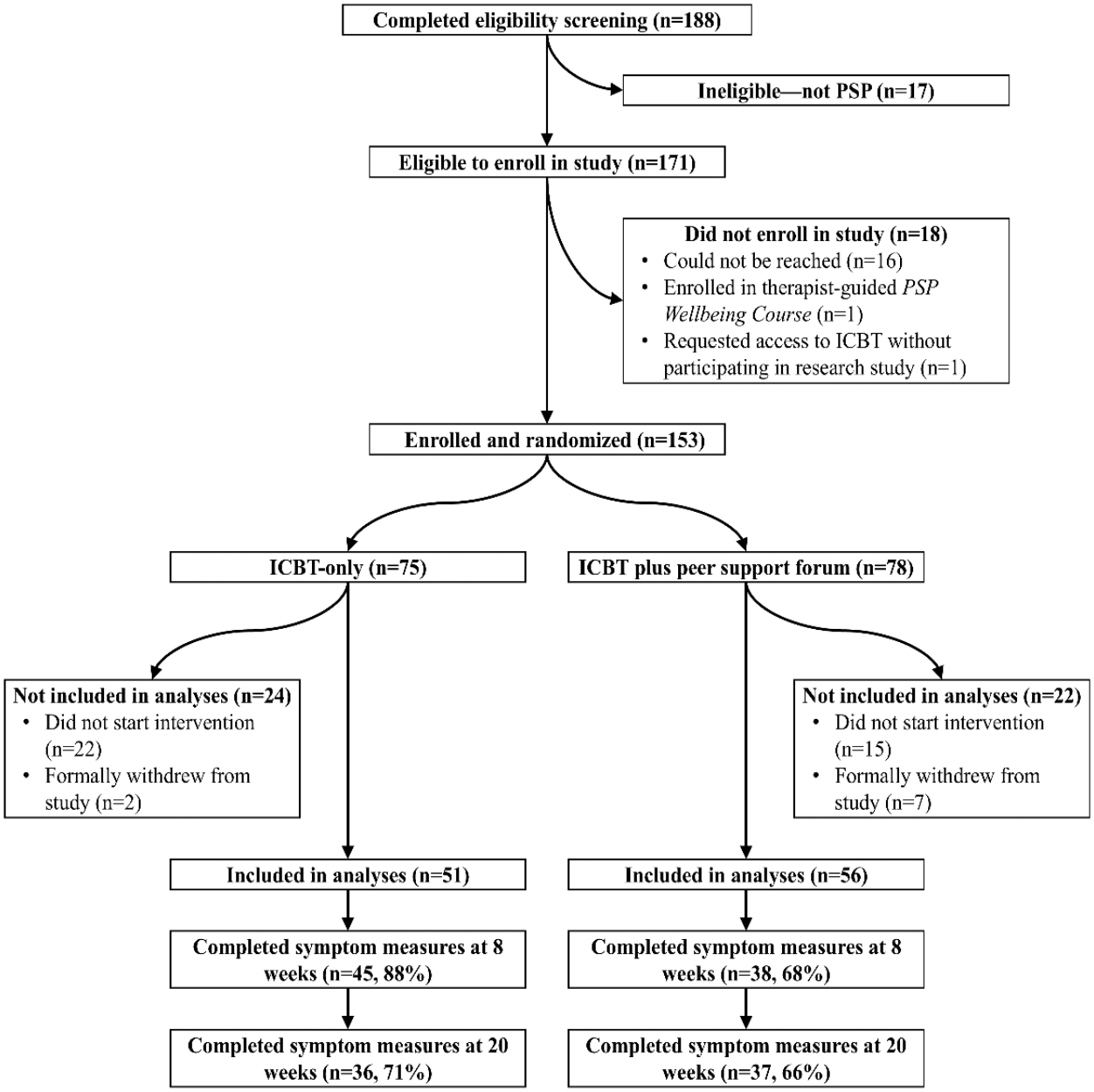

XSL•FO
RenderX

Table 1. Characteristics of participating public safety personnel (PSP; n=107).

Characteristics	All participants	ICBT[a]-only condition (n=51)	ICBT plus peer support forum condition (n=56)
Gender, n (%)			
Women	62 (57.9)	28 (54.9)	34 (60.7)
Men	45 (42.1)	23 (45.1)	22 (39.3)
Marital status, n (%)			
Married, common-law marriage, or living with a partner	78 (72.9)	38 (74.5)	40 (71.4)
Not married, in a common-law marriage, or living with a partner	29 (27.1)	13 (25.5)	16 (28.6)
Parental status, n (%)			
Has ≥1 children	70 (65.4)	36 (70.6)	34 (60.7)
Has no children	37 (34.6)	15 (29.4)	22 (39.3)
Province or territory, n (%)			
British Columbia	23 (21.5)	13 (25.5)	10 (17.9)
Ontario	21 (19.6)	10 (19.6)	11 (19.6)
Alberta	15 (14)	9 (17.6)	6 (10.7)
New Brunswick	12 (11.2)	4 (7.8)	8 (14.3)
Nova Scotia	11 (10.3)	4 (7.8)	7 (12.5)
Prince Edward Island	10 (9.3)	5 (9.8)	5 (8.9)
Saskatchewan	8 (7.5)	4 (7.8)	4 (7.1)
Manitoba	4 (3.7)	0 (0)	4 (7.1)
Newfoundland and Labrador	1 (0.9)	1 (2)	0 (0)
Northwest Territories	1 (0.9)	1 (2)	0 (0)
Quebec	1 (0.9)	0 (0)	1 (1.8)
Community size, n (%)			
<100,000 residents	72 (67.3)	32 (62.7)	40 (71.4)
≥100,000 residents	35 (32.7)	19 (37.3)	16 (28.6)
Level of education, n (%)			
No university degree	56 (52.3)	26 (51)	30 (53.6)
University degree	51 (47.7)	25 (49)	26 (46.4)
Years of experience as PSP[b], n (%)			
≥10	75 (70.1)	37 (72.5)	38 (67.9)
0-9	32 (29.9)	14 (27.5)	18 (32.1)
PSP occupational area, n (%)			
Police	37 (34.6)	19 (37.3)	18 (32.1)
Corrections	23 (21.5)	10 (19.6)	13 (23.2)
Paramedics or related emergency service	16 (15)	9 (17.6)	7 (12.5)
Fire	11 (10.3)	5 (9.8)	6 (10.7)
Communications (eg, 911 dispatch)	7 (6.5)	2 (3.9)	5 (8.9)
Other	13 (12.1)	6 (11.8)	7 (12.5)
Ethnicity, n (%)			
Indigenous (ie, First Nations, Inuit, or Metis)	7 (6.5)	6 (11.8)	1 (1.8)
White	96 (89.7)	42 (82.4)	54 (96.4)

XSL•FO
RenderX

Characteristics	All participants	ICBT[a]-only condition (n=51)	ICBT plus peer support forum condition (n=56)
Other ethnic minority group	3 (2.8)	2 (3.9)	1 (1.8)
Prefer not to answer	1 (0.9)	1 (2)	0 (0)
Age group (y), n (%)			
20-29	7 (6.5)	4 (7.8)	3 (5.4)
30-39	27 (25.2)	16 (31.4)	11 (19.6)
40-49	47 (43.9)	22 (43.1)	25 (44.6)
50-59	21 (19.6)	9 (17.6)	12 (21.4)
60-69	5 (4.7)	0 (0)	5 (8.9)
Age (y), mean (SD)	44.50 (9.28)	42.97 (8.94)	45.90 (9.45)

[a]ICBT: internet-delivered cognitive behavioral therapy.

[b]PSP: public safety personnel.

Changes in Questionnaire Scores

We observed a common pattern of results across all 10 MLM models: (1) all statistical assumptions were met; (2) we did not identify statistically significant effects of group or group-by-time interactions ($P \geq .17$ in all cases); (3) contrasts showed no effect of group on score change at 8 or 20 weeks for any measure; (4) there was a statistically significant and favorable effect of time (ie, scores on the PHQ-9, GAD-7, and PCL-5 decreased and scores on the FS and BRS increased over time); (5) contrasts showed statistically significant improvement in scores at 8 and 20 weeks for all measures; and (6) we identified residual variance, suggesting that models were likely missing predictor variables that could have helped account for estimates of dependent variables (which was expected given that covariates were tested separately via contrasts rather than being included in MLM models).

There were also some differences across MLM models. Certain contrasts for the PHQ-9 and PCL-5 showed further improvement in symptoms from 8 to 20 weeks. Further details of MLM results are reported in Table 2 (estimated means and percentage changes) and Table 3 (contrasts). Unaltered questionnaire scores observed among respondents at each time point are shown in Multimedia Appendix 3 [83-86].

XSL•FO
RenderX

Table 2. Summary of estimated means of scores on questionnaires and percentage changes over time.

Questionnaires and time points	Entire sample			Clinical subsamples		
	Both conditions	ICBT[a] only	ICBT plus peer support forum	Both conditions	ICBT only	ICBT plus peer support forum
PHQ-9[b,c]						
Pre-enrollment time point, estimated mean	9.51	9.69	9.34	14.21	14.35	14.07
8 weeks, estimated mean (% change from pre-enrollment time point)	7.26 (–23.7)	6.96 (–28.2)	7.53 (–19.4)	9.58 (–32.6)	9.86 (–31.3)	9.31 (–33.9)
20 weeks, estimated mean (% change from pre-enrollment time point)	6.01 (–36.7)	5.94 (–38.7)	6.08 (–34.9)	7.94 (–44.1)	7.96 (–44.5)	7.93 (–43.6)
GAD-7[d,e]						
Pre-enrollment time point, estimated mean	8.10	8.34	7.88	13.79	14.47	13.15
8 weeks, estimated mean (% change from pre-enrollment time point)	6.19 (–23.5)	5.92 (–29.0)	6.44 (–18.3)	8.71 (–36.9)	9.08 (–37.3)	8.35 (–36.5)
20 weeks, estimated mean (% change from pre-enrollment time point)	5.31 (–34.5)	5.47 (–34.4)	5.16 (–34.4)	8.25 (–40.2)	8.46 (–41.6)	8.05 (–38.8)
PCL-5[f,g]						
Pre-enrollment time point, estimated mean	26.86	24.73	28.80	46.64	47.47	46.08
8 weeks, estimated mean (% change from pre-enrollment time point)	18.71 (–30.3)	17.61 (–28.8)	19.71 (–31.6)	28.68 (–38.5)	33.84 (–28.7)	25.17 (–45.4)
20 weeks, estimated mean (% change from pre-enrollment time point)	16.49 (–38.6)	15.52 (–37.2)	17.37 (–39.6)	23.67 (–49.3)	24.78 (–47.8)	22.91 (–50.3)
FS[h,i]						
Pre-enrollment time point, estimated mean	40.83	40.92	40.75	37.63	37.16	38.02
8 weeks, estimated mean (% change from pre-enrollment time point)	42.22 (3.4)	41.86 (2.3)	42.55 (4.4)	39.78 (5.7)	38.49 (3.6)	40.86 (7.5)
20 weeks, estimated mean (% change from pre-enrollment time point)	43.35 (6.2)	43.28 (5.8)	43.41 (6.5)	41.11 (9.3)	40.34 (8.6)	41.76 (9.8)
BRS[j,k]						
Pre-enrollment time point, estimated mean	3.28	3.33	3.24	2.92	2.96	2.88
8 weeks, estimated mean (% change from pre-enrollment time point)	3.51 (6.8)	3.47 (4.0)	3.54 (9.3)	3.29 (12.9)	3.32 (12.3)	3.27 (13.5)
20 weeks, estimated mean (% change from pre-enrollment time point)	3.48 (5.9)	3.55 (6.5)	3.41 (5.3)	3.26 (11.8)	3.33 (12.5)	3.20 (11.0)

[a]ICBT: internet-delivered cognitive behavioral therapy.

[b]PHQ-9: Patient Health Questionnaire–9.

[c]Entire sample—both conditions: n=107, ICBT-only condition: n=51, and ICBT plus peer support forum: n=56; clinical subsamples—both conditions: n=53, ICBT-only condition: n=26, and ICBT plus peer support forum: n=27.

[d]GAD-7: Generalized Anxiety Disorder–7.

[e]Entire sample—both conditions: n=107, ICBT-only condition: n=51, and ICBT plus peer support forum: n=56; clinical subsamples—both conditions: n=39, ICBT-only condition: n=19, and ICBT plus peer support forum: n=20.

[f]PCL-5: PTSD Checklist for DSM-5.

XSL•FO
RenderX

[g]Entire sample—both conditions: n=107, ICBT-only condition: n=51, and ICBT plus peer support forum: n=56; clinical subsamples—both conditions: n=42, ICBT-only condition: n=17, and ICBT plus peer support forum: n=25.

[h]FS: Flourishing Scale.

[i]Entire sample—both conditions: n=107, ICBT-only condition: n=51, and ICBT plus peer support forum: n=56; clinical subsamples—both conditions: n=81, ICBT-only condition: n=37, and ICBT plus peer support forum: n=44.

[j]BRS: Brief Resilience Scale.

[k]Entire sample—both conditions: n=107, ICBT-only condition: n=51, and ICBT plus peer support forum: n=56; clinical subsamples—both conditions: n=78, ICBT-only condition: n=37, and ICBT plus peer support forum: n=41.

Table 3. Summary of contrasts assessing changes in questionnaire scores over time.

Variables	Entire sample			Clinical subsamples		
	t test (*df*)	*P* value	Cohen *d*	*t* test (*df*)	*P* value	Cohen *d*
PHQ-9[a,b]						
Pre-enrollment time point to 8 weeks after beginning treatment	–4.44 (158.78)	<.001	–0.70	–6.80 (76.52)	<.001	–1.55
Pre-enrollment time point to 20 weeks after beginning treatment	–6.55 (160.38)	<.001	–1.03	–8.39 (78.87)	<.001	–1.89
8-20 weeks after beginning treatment	–2.20 (156.57)	.03	–0.35	–2.11 (77.13)	.04	–0.48
GAD-7[c,d]						
Pre-enrollment time point to 8 weeks after beginning treatment	–4.18 (158.10)	<.001	–0.66	–6.67 (54.53)	<.001	–1.81
Pre-enrollment time point to 20 weeks after beginning treatment	–5.77 (159.88)	<.001	–0.91	–6.56 (56.78)	<.001	–1.74
8-20 weeks after beginning treatment	–1.71 (155.65)	.09	–0.27	–0.52 (54.08)	.61	–0.14
PCL-5[e,f]						
Pre-enrollment time point to 8 weeks after beginning treatment	–5.67 (151.91)	<.001	–0.92	–8.39 (57.98)	<.001	–2.20
Pre-enrollment time point to 20 weeks after beginning treatment	–6.88 (152.70)	<.001	–1.11	–10.58 (58.75)	<.001	–2.76
8-20 weeks after beginning treatment	–1.42 (149.50)	.16	–0.23	–2.51 (56.12)	.02	–0.67
FS[g,h]						
Pre-enrollment time point to 8 weeks after beginning treatment	2.10 (159.08)	.04	0.33	2.61 (118.39)	.01	0.48
Pre-enrollment time point to 20 weeks after beginning treatment	3.67 (159.97)	<.001	0.58	4.11 (119.64)	<.001	0.75
8-20 weeks after beginning treatment	1.60 (157.12)	.11	0.26	1.55 (117.03)	.12	0.29
BRS[i,j]						
Pre-enrollment time point to 8 weeks after beginning treatment	3.00 (160.80)	.003	0.47	4.94 (115.02)	<.001	0.92
Pre-enrollment time point to 20 weeks after beginning treatment	2.55 (162.06)	.01	0.40	4.23 (116.22)	<.001	0.78
8-20 weeks after beginning treatment	–0.29 (158.65)	.77	–0.05	–0.39 (112.77)	.70	–0.07

[a]PHQ-9: Patient Health Questionnaire–9.

[b]Entire sample: n=107; clinical subsamples: n=53.

[c]GAD-7: Generalized Anxiety Disorder–7.

[d]Entire sample: n=107; clinical subsamples: n=39.

[e]PCL-5: PTSD Checklist for DSM-5.

[f]Entire sample: n=107; clinical subsamples: n=42.

[g]FS: Flourishing Scale.

[h]Entire sample: n=107; clinical subsamples: n=81.

[i]BRS: Brief Resilience Scale.

[j]Entire sample: n=107; clinical subsamples: n=78.

Program Use

There was no statistically significant difference between groups with respect to the number of lessons participants accessed by 8 weeks (t_{105}=–0.28; *P*=.78; Cohen *d*=–0.05) or 20 weeks (t_{105}=0.82; *P*=.42; Cohen *d*=0.16). Collapsing across groups, a sizeable minority of participants accessed all 5 lessons of the *Self-Guided PSP Wellbeing Course* by 8 weeks (30/107, 28%) or 20 weeks (46/107, 43%). Nearly half (48/107, 44.9%) accessed 4 of 5 lessons by 8 weeks, whereas more than half (59/107, 55.1%) accessed 4 of 5 lessons by 20 weeks. Participants accessed an average of 3.33 (SD 5.00) additional

XSL•FO
RenderX

resources. Responses to the Program Use Questionnaire collapsed across groups and averaged across time points (ie, 2, 4, 6, and 8 weeks) showed that participants most commonly reported putting “some effort” into the course (39%), followed by “a little effort” (30.9%), “no effort” (17.5%), and “a lot of effort” (12.5%), with no participants reporting “a great deal of effort” at any time point.

Treatment Satisfaction

The Treatment Satisfaction Questionnaire and SRS were completed by 68% (38/56) of the participants in the ICBT plus peer support forum condition and 88% (45/51) of the participants in the ICBT-only condition. Two-tailed independent-sample *t* tests and chi-square tests showed no statistically significant differences between groups with respect to any treatment satisfaction variables ($P \geq .47$ in all cases). Accordingly, we present the results collapsed across groups in Table 4.

We qualitatively analyzed responses to open-ended questions from 61.7% (66/107) of the participants, which we organized into 3 main themes: positive feedback, negative or constructive feedback, and comments about personal circumstances or preferences that do not reflect the perceived helpfulness of the *Self-Guided PSP Wellbeing Course*. The results are shown in Table 5.

XSL•FO
RenderX

Table 4. Descriptive statistics on treatment satisfaction collapsed across conditions.

Variables	Values
Was it worth your time doing this course? (n=82), n (%)	
Yes	79 (96)
No	3 (4)
Would you feel confident recommending this treatment to a friend? (n=83), n (%)	
Yes	78 (94)
No	5 (6)
Overall, how satisfied were you with the treatment? (n=83)	
Very dissatisfied (0), n (%)	0 (0)
Dissatisfied (1), n (%)	0 (0)
Neutral (2), n (%)	31 (37)
Satisfied (3), n (%)	45 (54)
Very satisfied (4), n (%)	7 (8)
Values, mean (SD)	2.71 (0.62)
How has participating in this course affected your confidence that you can learn to manage your symptoms? (n=82)	
Greatly reduced (0), n (%)	3 (4)
Reduced (1), n (%)	3 (4)
No change (2), n (%)	21 (26)
Increased (3), n (%)	49 (60)
Greatly increased (4), n (%)	6 (7)
Values, mean (SD)	2.63 (0.82)
I would have preferred to communicate with a therapist by email while working through the PSP Wellbeing Course (n=82), n (%)	
Strongly disagree	2 (2)
Disagree	14 (17)
Neutral	32 (39)
Agree	27 (33)
Strongly agree	7 (9)
I would have preferred to communicate with a therapist by phone while working through the PSP Wellbeing Course (n=82), n (%)	
Strongly disagree	3 (4)
Disagree	19 (23)
Neutral	32 (39)
Agree	20 (24)
Strongly agree	8 (10)
Adapted SRS[a], mean (SD)	
Bond (n=83)	4.55 (1.13)
Goals (n=82)	4.67 (1.01)
Tasks (n=82)	4.37 (1.50)
Overall (n=82)	4.55 (1.29)

[a]SRS: Session Rating Scale.

XSL•FO
RenderX

Table 5. Results of content analysis of feedback on the Self-Guided PSP Wellbeing Course (n=66).

Theme, subtheme, and category	Example quote	Frequency, n (%)
Positive feedback		
Clinical content		
Positive feedback on stories or case examples	"I liked the stories cause it helped me relate and see other people are having these experiences." [Participant 1801]	18 (27)
Positive feedback on DIY[a] guides	"DIY Guides are very informative and easy to understand." [Participant 1256]	17 (26)
Positive feedback on additional resources	"I liked the resource library to be accessed for follow up and reminders." [Participant 1757]	13 (20)
Positive feedback on course content (eg, thorough, understandable, and relatable)	"It goes into explanations that in person therapy doesn't seem to have time for, or, that in person therapists don't think to cover." [Participant 1392]	8 (12)
Positive feedback on lessons	"Lessons were straight forward and easy to comprehend." [Participant 1583]	5 (8)
Positive feedback on tools and skills taught in the course	"Gave me a framework to understand what has been affecting me and how to work on it productively. I have taken my time, more than intended by the course I think, to practice the skills." [Participant 1175]	6 (9)
Course acted as a helpful reminder of previously learned skills and information	"Good refresher and reminder of important concepts." [Participant 1342]	4 (6)
Format and delivery		
Liked that the course was self-guided, self-paced, or accessible at any time and location	"[Liked] being able to work on the course on my own timeline, when I was in the right headspace. It didn't feel forced." [Participant 1091]	12 (18)
Liked the format or structure of the course or the presentation or delivery of information	"I liked how the course was structured." [Participant 1648]	10 (15)
Liked being able to download or print course materials or review them again in the future	"It is nice to have the resources to go back to in the future." [Participant 1648]	5 (8)
Liked the reminder emails	"[Liked] reminders to keep at it." [Participant 1225]	3 (5)
Other positive feedback		
General statement of liking the course	"[Liked] honestly, all of it." [Participant 1092]	2 (3)
No positive feedback provided	__[b]	2 (3)
Constructive or negative feedback		
Clinical content		
Disliked the stories, did not find them helpful, or provided feedback on them	"I didn't find the stories particularly helpful." [Participant 1154]	5 (8)
Course was too basic or recommendation for a second course with more tools	"I thought it would be longer and more in depth." [Participant 1742]	3 (5)
Some content seemed redundant or unnecessary	"I found the lessons and DIY guides a bit repetitive (they covered a lot of the same material)." [Participant 1173]	2 (3)
Other suggestions for improving clinical content	"I was approaching this as a preventative course as opposed to a treatment course so I found that the examples were not something I identified with. It would be wonderful if there was a separate course for individuals looking to build skills to help prevent a slide into negative mental health." [Participant 1258]	3 (5)
Format and delivery		
Difficulty or dislike concerning the current use of timelines and reminders to motivate completion	"I needed more time and felt somewhat anxious when the reminders were coming about a new section and I was behind." [Participant 1503]	7 (11)
Would prefer if the course included therapist support	"I wished I also had the therapist to help keep me on track and discuss some of my thoughts and feelings that came up while taking the course." [Participant 1801]	5 (8)

XSL•FO
RenderX

Theme, subtheme, and category	Example quote	Frequency, n (%)
Disliked amount of reading or suggested more video content	"[Disliked] a lot of reading. Hard to stay focused." [Participant 1546]	3 (5)
Course was not mobile friendly	"The slides were difficult to read in a phone. Sitting at a computer isn't always an option for privacy." [Participant 1181]	1 (2)
Other constructive or negative feedback		
No dislikes identified or constructive feedback provided	"There is nothing I didn't like." [Participant 1242]	29 (44)
Comments about personal circumstances or preferences		
Limited time, energy, or capacity to work on the course or unexpected life circumstances posing a barrier to progression in the course	"Nothing you can do but life threw me a curve the past couple weeks, very sick kitten so that was my immediate concern and this fell to the wayside." [Participant 1816]	6 (9)
Hard time with web-based courses in general, preference for in-person courses or would benefit more from in-person courses	"I would benefit more from in-person treatment, but am reluctant to participate." [Participant 1225]	3 (5)

[a]DIY: do-it-yourself.

[b]Not applicable.

Online Discussion Forum Use and Satisfaction

Only 9% (5/56) of the participants in the ICBT plus peer support forum condition posted in the forum, creating 9 posts in total. The moderator (author HCM) created an additional 16 posts in an effort to spark discussion. The Treatment Satisfaction Questionnaire was completed by 38 participants in the ICBT plus peer support forum condition, 14 (37%) of whom reported that they did not use the forum. Of the remaining 24 participants, 1 (4%) reported feeling "very satisfied" with the forum overall, 11 (46%) reported feeling "satisfied," 8 (33%) reported feeling "neutral," 4 (17%) reported feeling "dissatisfied," and none reported feeling "very dissatisfied." Among the 38 participants in the ICBT plus peer support forum who completed the Treatment Satisfaction Questionnaire were 15 (39%) who reported reading between "a few" posts and "all or nearly all" posts, including participants who reported that reading others' posts was "highly beneficial" (n=2, 13%), "beneficial" (n=3, 20%), "somewhat beneficial" (n=6, 40%), and "not beneficial at all" (n=4, 27%).

We received meaningful, analyzable feedback on the peer support forum from 52% (29/56) of the participants, identifying 17 categories of feedback, which we grouped into 3 general themes: positive feedback, constructive or negative feedback, and other personal reactions to the forum. The results of this content analysis are shown in Table 6.

XSL•FO
RenderX

Table 6. Results of content analysis of feedback on the peer support forum (n=29).

Theme, subtheme, and category	Example quote	Frequency, n (%)
Positive feedback		
Liked that the forum was supportive, open, or free of judgment	"Judgement-free, supportive space to use when/if/how helpful." [Participant 1217]	2 (7)
Liked reading others' comments or seeing a variety of perspectives	"[Liked] variety of viewpoints." [Participant 1217]	2 (7)
Liked not feeling alone	"It's nice to know you're not alone." [Participant 1816]	2 (7)
Liked that the forum was an option	"[Liked] that it was an option." [Participant 1978]	1 (3)
Did not like anything about the forum	"I did not [like the forum]. Possibly I didn't connect properly?" [Participant 1721]	2 (7)
Constructive or negative feedback		
Dislike, discomfort, or difficulty opening up to or being vulnerable with others	"[Did not post because] police culture does not encourage sharing or vulnerability with mental health. Peer forums are not a tool we are comfortable with. Especially with the association of privacy and information release in our jobs." [Participant 1181]	8 (28)
Disliked the low level of forum activity or did not post due to low activity level	"[Disliked that] it was not an active forum and often nothing had been posted." [Participant 1978]	7 (24)
Unaware of forum or comment that more prompts would result in greater forum use	"[Did not post because the forum] wasn't emphasized enough as an available tool or resource during the course. I also did not know it was available to me." [Participant 1095]	4 (14)
Would prefer a scheduled live chat to asynchronous posts	"Wished it was more of a real time chat." [Participant 1801]	2 (7)
General statement of dislike for or disinterest in the forum	"I did not like it." [Participant 1130]	2 (7)
Too much involvement from the moderator	"[Disliked that the forum was]...very monitored?" [Participant 1721]	1 (3)
Disliked nothing about the forum	"[Disliked] nothing." [Participant 1241]	1 (3)
Other personal reactions to the forum		
Did not post because of other demands or not enough time	"[Did not post because] work and life demands paused my participation in the program." [Participant 1584]	4 (14)
Participant did not feel that they had anything of value to add to forum	[Did not post because] "I felt I didn't have anything to add of value." [Participant 1816]	3 (10)
Comment on how it felt to post on the forum	"[It felt] very difficult, vulnerable to do, felt unburdened/heard after posting." [Participant 1217]	2 (7)
Misconception that participants cannot respond to each other	"[Disliked that the forum] seemed like a question and answer type without being able to respond to each other." [Participant 1721]	1 (3)
Did not feel a need to post because other aspects of the course were sufficient	"[Did not post because] I am still stuck on capturing my thoughts and found that the FAQ suffices." [Participant 1168]	1 (3)

Discussion

Principal Findings

ICBT is an effective mental health treatment [1], but clients demonstrate somewhat less favorable clinical outcomes [5-12] and engagement [13-15] when it is offered in a purely self-guided format. Persuasive design principles represent a possible means of improving engagement and outcomes in DMHIs [5,20], and preliminary evidence supports the use of social support principles of persuasive design implemented via online discussion forums [3,23-29], but there is a dearth of experimental research directly evaluating the impact of forums in ICBT or other DMHIs. Research has also shown that Canadian PSP benefit considerably from tailored, guided ICBT [34-36], but previous research has not evaluated self-guided ICBT among Canadian PSP. We conducted a randomized trial to assess the impact of adding an online discussion forum to self-guided ICBT and evaluate outcomes of tailored, self-guided ICBT among Canadian PSP.

Participants showed large improvements in symptoms of depression, anxiety, and posttraumatic stress, which supported and surpassed our hypothesis of at least small to moderate reductions in symptoms. Most meta-analyses of self-guided DMHIs have not reported pretest-posttest effect sizes, but our results for changes in depression over time compare favorably to the pretest-posttest effect size of d=0.78 reported in one meta-analysis of self-guided DMHIs for depression [87]. Changes in flourishing and resilience were more modest,

XSL•FO
RenderX

potentially because the *Self-Guided PSP Wellbeing Course* was designed to reduce symptoms of mental disorders but not explicitly designed to improve flourishing or resilience. Nevertheless, the finding of improvements in flourishing and resilience makes an important contribution to the research literature as research on the effects of ICBT on these constructs is scarce.

The symptom change and treatment satisfaction demonstrated in this study were roughly comparable to those observed in research on the guided version of the *PSP Wellbeing Course*, likely because the 2 courses included practically identical treatment materials. However, engagement with the guided version was markedly better, with only 5.7% of enrolled PSP failing to access or withdrawing from the intervention (compared to 46/153, 30% in this study), 76.1% of participants accessing at least 4 of the 5 lessons of the course within 8 weeks (compared to 48/107, 44.9% in this study), and 57.3% completing the course within 8 weeks (compared to 30/107, 28% in this study) [36]. These findings suggest that therapists play a pivotal role in both initiating and sustaining Canadian PSP's engagement in ICBT. This conclusion aligns with those of previous research showing that PSP very frequently cite therapist guidance as a liked aspect of ICBT [88] and with a broader research literature showing that engagement tends to be lower in self-guided than in guided DMHIs [13-15]. Nevertheless, engagement with the *Self-Guided PSP Wellbeing Course* in this study appears to compare favorably to that of other research on self-guided DMHIs. In a systematic review, Kelders et al [20] found, on average, a 54.2% rate of "intended use" (ie, some engagement but not necessarily completion) of internet interventions for mental health, including both guided and self-guided interventions. Another systematic review found real-world completion rates ranging from 0.5% to 28.6% for self-guided DMHIs [14]. Engagement in the present study may have been enhanced by the structure provided by the randomized trial design as this kind of study design was found to predict greater engagement in the review by Kelders et al [20]. Interestingly, only a minority of participants in the present study indicated that they would have preferred to receive therapist guidance via email or phone while taking the course.

The results failed to support our hypothesis that participants assigned to the ICBT plus peer support forum condition would demonstrate greater engagement and treatment outcomes. Given participants' limited engagement with the peer support forum, this was unsurprising. The proportion of participants who posted in the forum (5/56, 9%) was far lower than proportions of 53% [26] and 50.6% [28] reported in previous studies of forums in DMHIs. Similarly, the mean number of posts per participant (0.16) was far lower than the means of 13.1 [89], 4.5 [89], 2.2 [29], and 1.86 [28] reported in previous studies. We are aware of only one previous study in which a lower proportion (ie, 7%) of participants posted in a forum accompanying a DMHI [90]. Despite low engagement with the peer support forum, some participants reported feeling satisfied with it and indicated that reading posts was beneficial, with qualitative feedback suggesting that some participants felt that the forum was supportive and helped them feel that they were not alone.

There are several possible reasons for the low engagement with the forum in this study. First, qualitative feedback suggested that many participants felt uncomfortable opening up to others and showing vulnerability, with one participant explicitly attributing this to "police culture," suggesting that forums may be a poor fit for many PSP. Second, the fact that the *Self-Guided PSP Wellbeing Course* was transdiagnostic, with different participants experiencing different symptoms, may have led participants to feel that they did not have much in common with other forum users and, therefore, may also have detracted from their comfort in sharing their experiences. Third, forums may be particularly helpful as an adjunct to treatment for certain conditions; indeed, much of the past research supporting the use of forums in ICBT has been conducted within the context of ICBT for social anxiety [3,23,27,28]. Fourth, the PSP who self-selected into this study may have been particularly interested in independently accessing a self-guided treatment, whereas PSP who were interested in sharing their experiences with others may have opted for other mental health care options (including PSPNET's therapist-guided ICBT for PSP in provinces where it was available). Finally, there was likely room for improvement with respect to the structure and implementation of the forum and our efforts to encourage participants to use it.

There was only one statistically significant difference observed between conditions: a greater proportion of participants completed questionnaires at 8 weeks in the ICBT-only condition. It remains unclear why this occurred. It could be a spurious finding. It could also be that participants in the ICBT plus peer support forum condition inferred from the minimal forum engagement that engagement with the study *as a whole* was low and were less likely to complete questionnaires due to the phenomenon of social normative influence [91].

Strengths, Limitations, and Future Research

This study benefitted from a mixed methods approach, allowing for both a quantitative evaluation of treatment outcomes and a qualitative exploration of participants' experiences. Another strength of this study was its ecological validity as we evaluated the *Self-Guided PSP Wellbeing Course* and the peer support forum under the conditions in which they were designed to be implemented. This study also had important limitations. We did not include a control condition with which to compare outcomes of the course, and our inclusion of multiple outcome measures in separate analyses increased our familywise error rate. We expect that every discussion forum is unique and its social dynamics are unpredictable; therefore, a key limitation of this study is that it is, in a sense, a case study of a single forum with results that may not generalize well to other forums. This study was also sufficiently powered to detect only moderate differences between conditions, and due to an unexpectedly high rate of withdrawal from the study or failure to begin the intervention after we had ceased recruitment, we ultimately included 3 fewer participants in our analyses than originally planned. Finally, because we did not exclude participants with minimal or mild pre-enrollment symptoms from this research, floor effects are likely present in quantitative analyses conducted among our entire sample.

XSL•FO
RenderX

Future research can expand on this study and address the limitations noted previously in several ways. Although the peer support forum in this study had no demonstrable effect on treatment outcomes, previous research has shown excellent outcomes for online discussion forums [3,23-29], highlighting a need for further experimental research to evaluate the impact of forums on treatment outcomes in self-guided DMHIs. We are aware of a large factorial randomized trial assessing, among other things, the impact of an online discussion forum on treatment outcomes in ICBT, but the results of this trial are not yet available [92]. It would also be helpful for future research to identify common characteristics of forums that function well and those that do not, further explore DMHI users' perspectives on forums, and identify strategies for improving engagement with forums drawing on the persuasive system design framework and other work. With respect to outcomes of self-guided ICBT tailored for PSP, future research could compare treatment outcomes against a control condition, assess longer-term outcomes, and assess additional outcomes that we did not assess.

Conclusions

ICBT has shown excellent outcomes for treating a range of psychological concerns among PSP [34-36] and the general population [1]. Self-guided ICBT is more scalable but shows poorer engagement and outcomes than therapist-guided ICBT [5-15]. There is emerging evidence suggesting that persuasive design may help improve engagement and outcomes in ICBT [5,19-21], but further research is needed. We conducted a randomized trial, finding that transdiagnostic self-guided ICBT tailored specifically for PSP showed good outcomes, but PSP randomly assigned to receive access to a built-in online discussion forum showed limited engagement with it and no evidence of benefitting from it. Our findings support the continued implementation of self-guided ICBT. Our findings contrast with those of previous research on discussion forums in DMHIs, which have generally shown promising engagement and outcomes [3,23-29,89], highlighting a need for more research to clarify the circumstances under which forums may help improve engagement and outcomes in DMHIs. More broadly, as DMHIs become increasingly popular, there is a great need for more research identifying possible strategies to make them more engaging and effective, including—but not limited to—further research evaluating the impact of specific persuasive design principles.

Acknowledgments

The authors would like to thank Dr Janine Beahm for her feedback on their qualitative analyses and Dr Donald Sharpe for his recommendations regarding their quantitative analyses. The authors would like to thank all past and current members of the PSPNET team and all public safety personnel who participated in this research for making it possible. Particular thanks are due to Julia Gregory for her work as a research assistant in carrying out this study. The authors would like to thank Drs Nick Titov and Blake Dear for sharing treatment materials used to develop the *Self-Guided PSP Wellbeing Course*. This research was supported by funding from the Canadian Institutes of Health Research and Public Safety Canada; neither funder had any involvement in carrying out this research. Finally, acknowledgments are due to the Public Safety Canada Steering Committee, the Canadian Institute for Public Safety Research and Treatment, the Online Therapy Unit, and Information Services at the University of Regina for supporting this research.

Data Availability

The data sets generated and analyzed during this study are not publicly available due our commitment to our participants to securely store their data and refrain from sharing them with anyone outside of our research team.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample size planning.
[DOCX File , 39 KB - jmir_v26i1e59699_app1.docx]

Multimedia Appendix 2
Covariate analyses.
[DOCX File , 40 KB - jmir_v26i1e59699_app2.docx]

Multimedia Appendix 3
Observed descriptive statistics on questionnaire scores.
[DOCX File , 46 KB - jmir_v26i1e59699_app3.docx]

References

XSL•FO
RenderX

1. Andersson G, Carlbring P, Titov N, Lindefors N. Internet interventions for adults with anxiety and mood disorders: a narrative umbrella review of recent meta-analyses. Can J Psychiatry 2019 Jul 16;64(7):465-470 [FREE Full text] [doi: 10.1177/0706743719839381] [Medline: 31096757]
2. Christensen H, Hickie IB. E-mental health: a new era in delivery of mental health services. Med J Aust 2010 Jun 07;192(S11):S2-S3. [doi: 10.5694/j.1326-5377.2010.tb03684.x] [Medline: 20528702]
3. Furmark T, Carlbring P, Hedman E, Sonnenstein A, Clevberger P, Bohman B, et al. Guided and unguided self-help for social anxiety disorder: randomised controlled trial. Br J Psychiatry 2009 Nov;195(5):440-447. [doi: 10.1192/bjp.bp.108.060996] [Medline: 19880935]
4. Richards D, Enrique A, Palacios J, Duffy D. Internet-delivered cognitive behaviour therapy. In: Şenormancı Ö, Şenormancı G, editors. Cognitive Behavioral Therapy and Clinical Applications. London, UK: IntechOpen; 2018.
5. McCall HC, Hadjistavropoulos HD, Sundström CR. Exploring the role of persuasive design in unguided internet-delivered cognitive behavioral therapy for depression and anxiety among adults: systematic review, meta-analysis, and meta-regression. J Med Internet Res 2021 Apr 29;23(4):e26939 [FREE Full text] [doi: 10.2196/26939] [Medline: 33913811]
6. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
7. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205. [doi: 10.1080/16506070903318960] [Medline: 20183695]
8. Wright JH, Owen JJ, Richards D, Eells TD, Richardson T, Brown GK, et al. Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis. J Clin Psychiatry 2019 Mar 19;80(2):18r12188 [FREE Full text] [doi: 10.4088/JCP.18r12188] [Medline: 30900849]
9. Spek V, Cuijpers P, Nyklícek I, Riper H, Keyzer J, Pop V. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. Psychol Med 2007 Mar;37(3):319-328. [doi: 10.1017/S0033291706008944] [Medline: 17112400]
10. Krieger T, Bur OT, Weber L, Wolf M, Berger T, Watzke B, et al. Human contact in internet-based interventions for depression: a pre-registered replication and meta-analysis of randomized trials. Internet Interv 2023 Apr;32:100617 [FREE Full text] [doi: 10.1016/j.invent.2023.100617] [Medline: 37273939]
11. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
12. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
13. Andersson G, Carlbring P, Berger T, Almlöv J, Cuijpers P. What makes internet therapy work? Cogn Behav Ther 2009;38 Suppl 1:55-60. [doi: 10.1080/16506070902916400] [Medline: 19675956]
14. Fleming TM, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
15. Gan DZ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022 Jun 01;8:205520762210982. [doi: 10.1177/20552076221098268]
16. Andersson G, Titov N. Advantages and limitations of internet-based interventions for common mental disorders. World Psychiatry 2014 Feb;13(1):4-11 [FREE Full text] [doi: 10.1002/wps.20083] [Medline: 24497236]
17. Gerhards SA, de Graaf LE, Jacobs LE, Severens JL, Huibers MJ, Arntz A, et al. Economic evaluation of online computerised cognitive-behavioural therapy without support for depression in primary care: randomised trial. Br J Psychiatry 2010 Apr;196(4):310-318. [doi: 10.1192/bjp.bp.109.065748] [Medline: 20357309]
18. Lorenzo-Luaces L, Johns E, Keefe JR. The generalizability of randomized controlled trials of self-guided internet-based cognitive behavioral therapy for depressive symptoms: systematic review and meta-regression analysis. J Med Internet Res 2018 Nov 09;20(11):e10113 [FREE Full text] [doi: 10.2196/10113] [Medline: 30413400]
19. Fleming TM, de Beurs D, Khazaal Y, Gaggioli A, Riva G, Botella C, et al. Maximizing the impact of e-therapy and serious gaming: time for a paradigm shift. Front Psychiatry 2016;7:65 [FREE Full text] [doi: 10.3389/fpsyt.2016.00065] [Medline: 27148094]
20. Kelders SM, Kok RN, Ossebaard HC, van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
21. Ludden GD, van Rompay TJ, Kelders SM, van Gemert-Pijnen JE. How to increase reach and adherence of web-based interventions: a design research viewpoint. J Med Internet Res 2015 Jul 10;17(7):e172 [FREE Full text] [doi: 10.2196/jmir.4201] [Medline: 26163456]

XSL•FO
RenderX

22. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009;24:485-501. [doi: 10.17705/1CAIS.02428]
23. Andersson G, Carlbring P, Holmström A, Sparthan E, Furmark T, Nilsson-Ihrfelt E, et al. Internet-based self-help with therapist feedback and in vivo group exposure for social phobia: a randomized controlled trial. J Consult Clin Psychol 2006 Aug;74(4):677-686. [doi: 10.1037/0022-006X.74.4.677] [Medline: 16881775]
24. Boettcher J, Aström V, Påhlsson D, Schenström O, Andersson G, Carlbring P. Internet-based mindfulness treatment for anxiety disorders: a randomized controlled trial. Behav Ther 2014 Mar;45(2):241-253 [FREE Full text] [doi: 10.1016/j.beth.2013.11.003] [Medline: 24491199]
25. Johansson P, Westas M, Andersson G, Alehagen U, Broström A, Jaarsma T, et al. An internet-based cognitive behavioral therapy program adapted to patients with cardiovascular disease and depression: randomized controlled trial. JMIR Ment Health 2019 Oct 03;6(10):e14648 [FREE Full text] [doi: 10.2196/14648] [Medline: 31584000]
26. Hesser H, Gustafsson T, Lundén C, Henrikson O, Fattahi K, Johnsson E, et al. A randomized controlled trial of internet-delivered cognitive behavior therapy and acceptance and commitment therapy in the treatment of tinnitus. J Consult Clin Psychol 2012 Aug;80(4):649-661. [doi: 10.1037/a0027021] [Medline: 22250855]
27. Titov N, Andrews G, Schwencke G, Solley K, Johnston L, Robinson E. An RCT comparing effect of two types of support on severity of symptoms for people completing internet-based cognitive behaviour therapy for social phobia. Aust N Z J Psychiatry 2009 Jan 01;43(10):920-926. [doi: 10.1080/00048670903179228]
28. Berger T, Caspar F, Richardson R, Kneubühler B, Sutter D, Andersson G. Internet-based treatment of social phobia: a randomized controlled trial comparing unguided with two types of guided self-help. Behav Res Ther 2011 Mar;49(3):158-169. [doi: 10.1016/j.brat.2010.12.007] [Medline: 21255767]
29. Carolan S, Harris PR, Greenwood K, Cavanagh K. Increasing engagement with an occupational digital stress management program through the use of an online facilitated discussion group: results of a pilot randomised controlled trial. Internet Interv 2017 Dec;10:1-11 [FREE Full text] [doi: 10.1016/j.invent.2017.08.001] [Medline: 30135747]
30. Carleton RN, Afifi TO, Taillieu T, Turner S, Krakauer R, Anderson GS, et al. Exposures to potentially traumatic events among public safety personnel in Canada. Can J Behav Sci 2019 Jan;51(1):37-52. [doi: 10.1037/cbs0000115]
31. Carleton RN, Afifi TO, Turner S, Taillieu T, Duranceau S, LeBouthillier DM, et al. Mental disorder symptoms among public safety personnel in Canada. Can J Psychiatry 2018 Jan 28;63(1):54-64 [FREE Full text] [doi: 10.1177/0706743717723825] [Medline: 28845686]
32. Ricciardelli R, Carleton RN, Mooney T, Cramm H. "Playing the system": structural factors potentiating mental health stigma, challenging awareness, and creating barriers to care for Canadian public safety personnel. Health (London) 2020 May 16;24(3):259-278. [doi: 10.1177/1363459318800167] [Medline: 32283964]
33. McCall HC, Beahm JD, Fournier AK, Burnett JL, Carleton RN, Hadjistavropoulos HD. Stakeholder perspectives on internet-delivered cognitive behavioural therapy for public safety personnel: a qualitative analysis. Can J Behav Sci 2021 Jul;53(3):232-242. [doi: 10.1037/cbs0000242]
34. Hadjistavropoulos HD, McCall HC, Thiessen DL, Huang Z, Carleton RN, Dear BF, et al. Initial outcomes of transdiagnostic internet-delivered cognitive behavioral therapy tailored to public safety personnel: longitudinal observational study. J Med Internet Res 2021 May 05;23(5):e27610 [FREE Full text] [doi: 10.2196/27610] [Medline: 33949959]
35. McCall H, Dear BF, Landry C, Beahm JD, Gregory J, Titov N, et al. Internet-delivered cognitive behavioural therapy for symptoms of PTSD among public safety personnel: initial outcomes of an open cohort preference trial of transdiagnostic and disorder-specific therapy. Internet Interv 2023 Sep;33:100656 [FREE Full text] [doi: 10.1016/j.invent.2023.100656] [Medline: 37609530]
36. Hadjistavropoulos HD, McCall HC, Dear BF, Beahm JD, Carleton RN, Titov N. Outcomes of transdiagnostic internet-delivered cognitive behavioural therapy tailored to public safety personnel: a longitudinal observational study. J Anxiety Disord 2024 Jun;104:102861 [FREE Full text] [doi: 10.1016/j.janxdis.2024.102861] [Medline: 38640867]
37. Kang M, Ragan BG, Park JH. Issues in outcomes research: an overview of randomization techniques for clinical trials. J Athl Train 2008;43(2):215-221 [FREE Full text] [doi: 10.4085/1062-6050-43.2.215] [Medline: 18345348]
38. Sheehan DV. The Sheehan disability scales: the anxiety disease and how to overcome it. Mapi Research Trust. URL: https://eprovide.mapi-trust.org/instruments/sheehan-disability-scale [accessed 2024-04-29]
39. Sharpe D, Cribbie RA. Analysis of treatment-control pre-post-follow-up design data. Quant Method Psychol 2023 Feb 01;19(1):25-46. [doi: 10.20982/tqmp.19.1.p025]
40. Butcher NJ, Monsour A, Mew EJ, Chan AW, Moher D, Mayo-Wilson E, et al. Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension. JAMA 2022 Dec 13;328(22):2252-2264. [doi: 10.1001/jama.2022.21022] [Medline: 36511921]
41. Sunderland A, Findlay LC. Perceived need for mental health care in Canada: results from the 2012 Canadian Community Health Survey-Mental Health. Health Rep 2013 Sep;24(9):3-9 [FREE Full text] [Medline: 24258361]
42. Young C, Helis E, Williams D. Internet-delivered cognitive behavioural therapy for major depressive disorder and anxiety disorders: an environmental scan. Canadian Agency for Drugs and Technologies in Health. 2018. URL: https://www.cadth.ca/sites/default/files/pdf/ES0331_iCBT_for_Major_Depression_and_Anxiety.pdf [accessed 2024-04-29]

XSL•FO
RenderX

43. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019 Feb;18(1):20-28 [FREE Full text] [doi: 10.1002/wps.20610] [Medline: 30600624]
44. Donohue MC, Edland SD, Hu N. longpower: sample size calculations for longitudinal data. Cran R. URL: https://cran.r-project.org/web/packages/longpower/longpower.pdf [accessed 2024-04-29]
45. Dear BF, Staples LG, Terides MD, Fogliati VJ, Sheehan J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for social anxiety disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2016 Aug;42:30-44 [FREE Full text] [doi: 10.1016/j.janxdis.2016.05.004] [Medline: 27261562]
46. Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther 2011;40(4):251-266. [doi: 10.1080/16506073.2011.616531] [Medline: 22060248]
47. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
48. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
49. Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
50. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012 Feb 21;184(3):E191-E196 [FREE Full text] [doi: 10.1503/cmaj.110829] [Medline: 22184363]
51. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
52. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. J Trauma Stress 2015 Dec;28(6):489-498. [doi: 10.1002/jts.22059] [Medline: 26606250]
53. Bovin MJ, Marx BP, Weathers FW, Gallagher MW, Rodriguez P, Schnurr PP, et al. Psychometric properties of the PTSD checklist for diagnostic and statistical manual of mental disorders-fifth edition (PCL-5) in veterans. Psychol Assess 2016 Nov;28(11):1379-1391. [doi: 10.1037/pas0000254] [Medline: 26653052]
54. Wortmann JH, Jordan AH, Weathers FW, Resick PA, Dondanville KA, Hall-Clark B, et al. Psychometric analysis of the PTSD Checklist-5 (PCL-5) among treatment-seeking military service members. Psychol Assess 2016 Nov;28(11):1392-1403. [doi: 10.1037/pas0000260] [Medline: 26751087]
55. Leontjevas R, de Beek WO, Lataster J, Jacobs N. Resilience to affective disorders: a comparative validation of two resilience scales. J Affect Disord 2014 Oct;168:262-268. [doi: 10.1016/j.jad.2014.07.010] [Medline: 25069081]
56. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
57. Diener E, Wirtz D, Biswas-Diener R, Tov W, Kim-Prieto C, Choi DW, et al. New measures of well-being. In: Diener E, editor. Assessing Well-Being: The Collected Works of Ed Diener. Cham, Switzerland: Springer; 2009:247-266.
58. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]
59. Miller SD, Duncan BL, Johnson LD, Reynolds LR, Brown J. The session rating scale: preliminary psychometric properties of a “working” alliance measure. J Brif Ther 2000;3(1):3-12 [FREE Full text]
60. Campbell A, Hemsley S. Outcome rating scale and session rating scale in psychological practice: clinical utility of ultra-brief measures. Clin Psychol 2009 Mar 11;13(1):1-9. [doi: 10.1080/13284200802676391]
61. Duncan BL, Miller SD, Sparks JA, Claud DA, Reynolds LR, Brown J, et al. The session rating scale: preliminary psychometric properties of a “working” alliance measure. J Brif Ther 2003;3(1):3-12 [FREE Full text]
62. Cavanagh K, Herbeck Belnap B, Rothenberger SD, Abebe KZ, Rollman BL. My care manager, my computer therapy and me: the relationship triangle in computerized cognitive behavioural therapy. Internet Interv 2018 Mar;11:11-19 [FREE Full text] [doi: 10.1016/j.invent.2017.10.005] [Medline: 30135755]
63. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
64. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction 1993 Jun 24;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
65. Berman AH, Bergman H, Palmstierna T, Schlyter F. Evaluation of the drug use disorders identification test (DUDIT) in criminal justice and detoxification settings and in a Swedish population sample. Eur Addict Res 2005;11(1):22-31. [doi: 10.1159/000081413] [Medline: 15608468]

XSL•FO
RenderX

66. McCall HC, Sison AP, Burnett JL, Beahm JD, Hadjistavropoulos HD. Exploring perceptions of internet-delivered cognitive behaviour therapy among public safety personnel: informing dissemination efforts. Int J Environ Res Public Health 2020 Aug 19;17(17):6026 [FREE Full text] [doi: 10.3390/ijerph17176026] [Medline: 32824994]
67. Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. MindSpot clinic: an accessible, efficient, and effective online treatment service for anxiety and depression. Psychiatr Serv 2015 Oct;66(10):1043-1050. [doi: 10.1176/appi.ps.201400477] [Medline: 26130001]
68. Hadjistavropoulos HD, Peynenburg V, Thiessen DL, Nugent M, Karin E, Staples L, et al. Utilization, patient characteristics, and longitudinal improvements among patients from a provincially funded transdiagnostic internet-delivered cognitive behavioural therapy program: observational study of trends over 6 years. Can J Psychiatry 2022 Mar 12;67(3):192-206 [FREE Full text] [doi: 10.1177/07067437211006873] [Medline: 33840264]
69. McCall HC, Hadjistavropoulos HD, Loutzenhiser L. Reconsidering the ethics of exclusion criteria in research on digital mental health interventions. Ethics Behav 2019 Oct 31;31(3):171-180. [doi: 10.1080/10508422.2019.1684295]
70. Harvey LA. Statistical testing for baseline differences between randomised groups is not meaningful. Spinal Cord 2018 Oct 4;56(10):919. [doi: 10.1038/s41393-018-0203-y] [Medline: 30287888]
71. Gupta SK. Intention-to-treat concept: a review. Perspect Clin Res 2011 Jul;2(3):109-112 [FREE Full text] [doi: 10.4103/2229-3485.83221] [Medline: 21897887]
72. McNeish D. Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward-Roger correction. Multivariate Behav Res 2017 Jul 17;52(5):661-670. [doi: 10.1080/00273171.2017.1344538] [Medline: 28715244]
73. Field AP. Discovering Statistics Using IBM SPSS Statistics. 5th edition. Thousand Oaks, CA: Sage Publications; 2018.
74. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. 7th edition. Harlow, UK: Pearson; 2014.
75. Edmonds M, Hadjistavropoulos HD, Schneider LH, Dear BF, Titov N. Who benefits most from therapist-assisted internet-delivered cognitive behaviour therapy in clinical practice? predictors of symptom change and dropout. J Anxiety Disord 2018 Mar;54:24-32 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.003] [Medline: 29421369]
76. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016 Dec;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
77. Constantino MJ, Arnkoff DB, Glass CR, Ametrano RM, Smith JZ. Expectations. J Clin Psychol 2011 Feb;67(2):184-192. [doi: 10.1002/jclp.20754] [Medline: 21128304]
78. Thompson-Hollands J, Bentley KH, Gallagher MW, Boswell JF, Barlow DH. Credibility and outcome expectancy in the unified protocol: relationship to outcomes. J Exp Psychopathol 2014 Mar 30;5(1):72-82. [doi: 10.5127/jep.033712]
79. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? a systematic review of predictors and moderators of treatment outcome. Internet Interv 2023 Sep;33:100635 [FREE Full text] [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
80. Lambert MJ. Psychotherapy outcome research: implications for integrative and eclectical therapists. In: Norcross JC, Goldfried MR, editors. Handbook of Psychotherapy Integration. New York, NY: Basic Books; 1992:94-129.
81. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
82. Berger R. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. Qual Res 2013 Jan 03;15(2):219-234. [doi: 10.1177/1468794112468475]
83. The improving access to psychological therapies manual: appendices and helpful resources. National Collaborating Centre for Mental Health. 2019. URL: https://www.england.nhs.uk/wp-content/uploads/2018/06/iapt-manual-appendices-and-helpful-resources-v3.pdf [accessed 2020-12-15]
84. Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. The first 30 months of the mindspot clinic: evaluation of a national e-mental health service against project objectives. Aust N Z J Psychiatry 2017 Dec;51(12):1227-1239. [doi: 10.1177/0004867416671598] [Medline: 27733709]
85. Titov N, Dear BF, Nielssen O, Wootton B, Kayrouz R, Karin E, et al. User characteristics and outcomes from a national digital mental health service: an observational study of registrants of the Australian MindSpot Clinic. Lancet Digit Health 2020 Nov;2(11):e582-e593 [FREE Full text] [doi: 10.1016/S2589-7500(20)30224-7] [Medline: 33103097]
86. Byllesby BM, Dickstein BD, Chard KM. The probability of change versus dropout in veterans receiving cognitive processing therapy for posttraumatic stress disorder. Behav Res Ther 2019 Dec;123:103483. [doi: 10.1016/j.brat.2019.103483] [Medline: 31710956]
87. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
88. Beahm JD, McCall HC, Carleton RN, Titov N, Dear B, Hadjistavropoulos HD. Insights into internet-delivered cognitive behavioural therapy for public safety personnel: exploration of client experiences during and after treatment. Internet Interv 2021 Dec;26:100481 [FREE Full text] [doi: 10.1016/j.invent.2021.100481] [Medline: 34815951]
89. Ljótsson B, Falk L, Vesterlund AW, Hedman E, Lindfors P, Rück C, et al. Internet-delivered exposure and mindfulness based therapy for irritable bowel syndrome--a randomized controlled trial. Behav Res Ther 2010 Jun;48(6):531-539. [doi: 10.1016/j.brat.2010.03.003] [Medline: 20362976]

XSL•FO
RenderX

90. O'Mahen HA, Woodford J, McGinley J, Warren FC, Richards DA, Lynch TR, et al. Internet-based behavioral activation--treatment for postnatal depression (Netmums): a randomized controlled trial. J Affect Disord 2013 Sep 25;150(3):814-822. [doi: 10.1016/j.jad.2013.03.005] [Medline: 23602514]
91. Goldstein NJ, Cialdini RB, Griskevicius V. A room with a viewpoint: using social norms to motivate environmental conservation in hotels. J Consum Res 2008 Oct 01;35(3):472-482. [doi: 10.1086/586910]
92. Carlbring P, Lindqvist K, Mechler J, Philips B. Transdiagnostic self-help treatments for anxiety and/or depression: a full-factorial RCT investigating psychotherapeutic modalities, discussion boards and treatment length. In: Proceedings of the 11th Scientific Meeting on International Society for Research on Internet Interventions. 2022 Presented at: ISRII '22; September 18-21, 2022; Pittsburgh, PA p. 1-19 URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1700061&dswid=-46

Abbreviations

BRS: Brief Resilience Scale
CEQ: Credibility/Expectancy Questionnaire
CONSORT: Consolidated Standards of Reporting Trials
DMHI: digital mental health intervention
FS: Flourishing Scale
GAD-7: Generalized Anxiety Disorder–7
ICBT: internet-delivered cognitive behavioral therapy
MLM: multilevel modeling
PCL-5: PTSD Checklist for DSM-5
PHQ-9: Patient Health Questionnaire–9
PSP: public safety personnel
PTSD: posttraumatic stress disorder
SRS: Session Rating Scale

Edited by A Mavragani; submitted 19.04.24; peer-reviewed by L Yang; comments to author 10.06.24; revised version received 26.06.24; accepted 22.07.24; published 14.08.24.

Please cite as:
McCall HC, Hadjistavropoulos HD
Impact of an Online Discussion Forum on Self-Guided Internet-Delivered Cognitive Behavioral Therapy for Public Safety Personnel: Randomized Trial
J Med Internet Res 2024;26:e59699
URL: https://www.jmir.org/2024/1/e59699
doi:10.2196/59699
PMID:

©Hugh C McCall, Heather D Hadjistavropoulos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of a Smartphone-Based Stress Management Program for Depression in Hospital Nurses During COVID-19 in Vietnam and Thailand: 2-Arm Parallel-Group Randomized Controlled Trial

Kazuhiro Watanabe[1], PhD; Thuy Thi Thu Tran[2], PhD; Narisara Sripo[3], MNS; Asuka Sakuraya[4], PhD; Kotaro Imamura[4], PhD; Plernpit Boonyamalik[5], PhD; Natsu Sasaki[6], PhD; Thanate Tienthong[3], MNS; Hiroki Asaoka[6], PhD; Mako Iida[6], PhD; Quynh Thuy Nguyen[2], PhD; Nga Thi Nguyen[7], MPH; Son Thai Vu[2], MPH; Thuy Thi Ngo[8], PhD; Tham Thi Luyen[8], BN; Long Duc Nguyen[9], PhD; Nga Thi Viet Nguyen[9], MPP; Binh Thanh Nguyen[9], MN; Yutaka Matsuyama[10], PhD; Yukie Takemura[11], PhD; Daisuke Nishi[6], MD, PhD; Akizumi Tsutsumi[1], MD; Huong Thanh Nguyen[7], PhD; Orawan Kaewboonchoo[5], PhD; Norito Kawakami[4], MD, PhD

[1]Department of Public Health, Kitasato University School of Medicine, Sagamihara, Japan
[2]Faculty of Environmental and Occupational Health, Hanoi University of Public Health, Hanoi, Vietnam
[3]Princess Agrarajakumari College of Nursing, Chulabhorn Royal Academy, Bangkok, Thailand
[4]Department of Digital Mental Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[5]Faculty of Public Health, Mahidol University, Bangkok, Thailand
[6]Department of Mental Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[7]Faculty of Social Science and Behavior, Hanoi University of Public Health, Hanoi, Vietnam
[8]Pho Noi Hospital, Hung Yen Province, Vietnam
[9]Xanh Pon Hospital, Ha Noi, Vietnam
[10]Department of Biostatistics, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[11]Nursing Department, University of Tokyo Hospital, Tokyo, Japan

Corresponding Author:
Norito Kawakami, MD, PhD
Department of Digital Mental Health
Graduate School of Medicine
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 3 5800 9621
Email: nkawakami@m.u-tokyo.ac.jp

Abstract

Background: During the COVID-19 pandemic, health care professionals experienced high levels of depression. However, extant research has not highlighted effective internet-based psychological interventions to improve the mental health in this population during the pandemic. It remains unclear whether self-guided, internet-based cognitive behavioral therapy (iCBT) programs are effective in improving the mental health of health care workers during the COVID-19 pandemic.

Objective: The aim of this study was to evaluate the effectiveness of a smartphone-based iCBT stress management program for reducing the depression experienced by nurses in Vietnam and Thailand.

Methods: From March to April 2022, a 2-arm, parallel-group randomized controlled trial was implemented. One arm offered a 7-week self-guided iCBT program, and the other offered treatment as usual as a control arm. Full-time nurses were recruited from 6 hospitals: 2 hospitals in Vietnam and 4 hospitals in Thailand. The primary outcome of this program was the severity of depression measured by the Depression Anxiety Stress Scale-21 items. Follow-up surveys were conducted to measure the change in depression severity at 3 months (July-August 2022) and at 6 months (October-November 2022) after baseline. Mixed modeling for repeated measures was used to test the effects of the intervention compared with the control for the follow-up.

XSL•FO
RenderX

Results: A total of 1203 nurses were included in this study: 602 in the intervention group and 601 in the control group. The follow-up rate at 3 and 6 months ranged from 85.7% (515/601) to 87.5% (527/602). The completion rate for the program was 68.1% (410/602). The group difference in depression was significant at the 3-month follow-up (coefficient=–0.92, 95% CI –1.66 to –0.18; *P*=.02) and nonsignificant at the 6-month follow-up (coefficient=–0.33, 95% CI –1.11 to 0.45; *P*=.41). The estimated effect sizes were –0.15 and –0.06 at the 3- and 6-month follow-ups, respectively.

Conclusions: Our study shows that the smartphone-based iCBT program was effective in reducing depression at the 3-month follow-up among hospital nurses in Vietnam and Thailand during the COVID-19 pandemic. However, the effect size was small, and therefore, these results may not be clinically meaningful.

Trial Registration: UMIN Clinical Trials Registry UMIN000044145; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000050128

International Registered Report Identifier (IRRID): RR2-10.20944/preprints202303.0450.v1

(J Med Internet Res 2024;26:e50071) doi:10.2196/50071

KEYWORDS

digital mental health intervention; unguided program; universal prevention; health care workers; nurses; COVID-19; depression; mobile phone

Introduction

Background

The COVID-19 pandemic started in late December 2019 and is still impacting the physical and mental health of people worldwide [1]. Health care workers were the most affected during the COVID-19 pandemic, as they experienced increased workload while providing care for patients, along with anxiety about the high risk of infection from exposure to COVID-19 and dealing with stigmatization and discrimination from the public at large. During the COVID-19 pandemic, health care professionals experienced high levels of depression, anxiety, and posttraumatic stress symptoms [2]. In particular, nurses experienced poorer mental health than other health care professionals [2]. A universal approach is important for improving the mental health of nurses, including that of the healthy nurses, given that every nurse is at risk of their mental health being negatively affected by COVID-19. Improving their mental health will make the quality of their work life better and ensure that the workforce provides appropriate health care by preventing their mental health issues from taking a toll on them.

Very few studies have highlighted effective interventions that can improve the mental health of health care workers during infectious disease epidemics such as SARS, Ebola, and Middle East respiratory syndrome [3]. During the Ebola epidemic, a training program on Psychological First Aid helped prevent burnout among primary health care workers in the community [4]. During the SARS outbreak, a comprehensive organizational intervention, including infection measures, protection training, and psychological support teams, for patients and health care workers reduced depression and improved sleep among nurses in a hospital [5]. A few studies have reported positive responses to group-based resilience training [6] and an improvement in self-efficacy and interpersonal problems after conducting a web-based resilience training program among health care workers during the influenza pandemic [7]. In addition, other types of psychological interventions have been found to be effective in improving the mental health of frontline health care workers and other responders: eye movement desensitization and reprocessing, trauma risk management [8], supportive psychotherapy [9], peer support programs [10], and mindfulness- and mediation-based interventions [11].

As close face-to-face personal contact was limited to prevent the spread of COVID-19 in an organization or community, using information and communication technology to provide training on stress management skills, such as internet stress management programs [12], has been encouraged [13]. Internet-based programs have additional strengths in accessibility, low cost, and confidentiality. These programs can be completed at convenient times, which is considered suitable for health care workers who work irregular schedules. A randomized controlled trial (RCT) reported that an internet-based cognitive behavioral therapy (iCBT) program reduced perceived work-related stress among nurses in the United States [14]. The effectiveness of non–face-to-face CBT interventions for improving the mental health of health care workers has been reported to be similar to that of face-to-face interventions [15]. We found that a self-guided iCBT program reduced depression among nurses in Vietnam during a 3-month follow-up [16].

However, these studies were conducted before the COVID-19 pandemic. In the early stages of the pandemic, e-mental health solutions were summarized to assist health care workers [17]. CBT was also delivered to nurses for insomnia as a digital sleep intervention, and its effect on insomnia was confirmed [18]. An RCT conducted in Spain during the COVID-19 pandemic reported that an internet-based stress management program focusing on psychoeducation and mindfulness skills training yielded no significant improvement in depression, anxiety, or stress among health care workers [19]. A recent RCT reported that a therapist-guided, 6-week, internet- and CBT-based stress recovery intervention for nurses improved stress recovery, anxiety, and depression symptoms [20]. However, it remains unclear whether self-guided iCBT programs are effective in improving the mental health of health care workers during the COVID-19 pandemic, where health care workers experienced far higher stress than usual.

XSL•FO
RenderX

Objectives

Southeast Asia is one of the most populous geographical areas in the world, where rapid aging of the society is currently underway. The shortage of the health care workforce was serious in this region even before the COVID-19 pandemic [21]. As of the end of 2022, this region had the largest total mortality owing to COVID-19 followed by Europe and the United States [22]. Our study aims to evaluate the effectiveness of a 7-week smartphone iCBT stress management program for improving the mental health of nurses in this region during the COVID-19 pandemic by using a modified version of our previous iCBT program for nurses in Vietnam [16] adapted to suit the COVID-19 situation. To increase the generalizability of our findings, we selected 2 rapidly industrialized middle-income countries in the Southeast Asian region, namely, Vietnam and Thailand, to conduct the effectiveness study as part of the "Coping with COVID-19 helps Nurses be Active, Tough, and Smiling (COCONATS)" project. As the primary outcome, depression symptoms were measured. We hypothesized that depression scores in the intervention group would significantly improve when compared with those in the control group.

Methods

Trial Design

The trial design was an open, 2-arm, parallel-group RCT. Of the 2 arms, one offered the iCBT program for 7 weeks, and the other received occupational health and other services in their hospitals as a control group. After the completion of a web-based survey at baseline, nurses in each hospital were randomly assigned to the intervention or control groups at a 1:1 ratio. Randomization was conducted by stratifying the sample by countries, hospitals, and severity of depression at baseline. Follow-up surveys to measure the change in the outcome were conducted at 3 and 6 months after baseline. The web-based baseline and follow-up surveys were conducted independently in (not as a part of) the iCBT program. The study protocol was registered at the University Hospital Medical Information Network Clinical Trials Registry (UMIN000044145) and was published elsewhere [23]. This report of a randomized trial conforms to the guidelines in the CONSORT (Consolidated Standards of Reporting Trials) [24,25] (Multimedia Appendix 1).

Study Setting

From March to April 2022, this RCT was conducted in 6 hospitals across Vietnam and Thailand. In Vietnam, 1 hospital in Hanoi and another province hospital in the north of Vietnam were selected as the study fields. After obtaining consent from the hospitals, lists comprising 723 full-time registered clinical nurses in the hospitals were sent to researchers at the Hanoi University of Public Health. These researchers sent the research project advertisement (flyers) and documents that explained the aim and procedure of the study, along with the baseline questionnaire (February to March 2022) to the potential participants. When the nurses agreed to participate, they signed an informed consent form and completed the baseline survey (March to April 2022). After the randomization, the intervention was provided to the intervention group for 10 weeks. The follow-up surveys were conducted at 3 months (July 2022) and 6 months (October 2022) after baseline.

In Thailand, 4 provincial hospitals at the tertiary level and above, located in Bangkok, Samut, Sakhon, and Lop Buri, joined this study. The principal investigator of Mahidol University set up a web-based meeting with local collaborators of the target hospitals to explain the research project and sent the advertisement to the collaborators who introduced the research project to employed nurses (N=4837). Later, the main research team invited all the potential participants to a web-based meeting to reread the aims and steps of the project (February to March 2022). In the meeting, the potential participants accessed and pressed accept on the web-based informed consent form to electronically sign it and begin completing the baseline questionnaire (March to April 2022). Those who were unable to attend the meeting could request project documents or ask for more information from the principal investigator or local collaborators during the application period. After the randomization, the intervention was provided to the intervention group for 10 weeks. The follow-up surveys were conducted at 3 (July to August 2022) and 6 months (October to November 2022) after baseline.

Ethics Approval

The study protocol was approved by the research ethics review board of the Graduate School of Medicine/Faculty of Medicine, The University of Tokyo (32021082NI-1), Hanoi University of Public Health (353/2021/YTCC-HD3), and Mahidol University (MU MOU COA 2021-001). Informed consent was obtained from all the participants after full disclosure and explanation of the purpose and procedures of the study. The participants were informed that their participation was voluntary, that they could withdraw from the study at any time without stating reasons, and that nonparticipation and withdrawal would not result in any disadvantage to them.

Recruitment

This RCT recruited nurses working in hospitals in Vietnam and Thailand. Nurses who were included were employed full-time as registered nurses at one of the 6 hospitals in Vietnam or Thailand and were able to access the internet via a mobile device such as a smartphone. The study excluded those who were (1) on sick or maternity leave at baseline or planned to take maternity leave or to leave or change their job in the next 6 months, (2) assistant nurses or helpers, (3) nonregular or part-time employees, (4) on sick leave for 10 or more days for a physical or mental condition in the past 4 weeks, and (5) receiving treatment for a mental health problem from any health care professional.

Interventions

iCBT Program as the Intervention

The participants allocated to the intervention group were offered a 7-week smartphone-based self-guided iCBT program for nurses during the COVID-19 pandemic in Vietnam and Thailand. "The ABC Stress Management–COVID-19 version" was developed. This program, which consists of texts and static visuals of cartoon characters talking on the study topic and

XSL•FO
RenderX

guiding participants through the learning (Figure S1 in Multimedia Appendix 2), was originally developed in our previous study for nurses in Vietnam [16]. This program can be accessed using smartphone apps downloadable from the Apple App Store and Google Play Store as well as on personal computers and other devices connected to the internet. The original program [16] was developed in Vietnamese by researchers from Japan and Vietnam. Appropriate cultural adaptations were safeguarded through an extensive discussion among researchers and a group of senior nurses. A case story that would be used in the program to illustrate the stress process and stress management was developed, considering popular job stressors in the target group, such as job overload, role conflict, and work-family interference. The cartoon character images of a counselor and a nurse were carefully created to be appropriate for the setting. These adaptations were assessed to be appropriate for Thailand's culture and setting. The COVID-19 module was developed through a discussion among researchers from 3 countries; it was accommodated and adapted to the cultures and COVID-19 epidemic situations of Vietnam and Thailand. The program comprised 7 modules that were presented in a fixed order. New modules were made available each week, starting from modules 1 to 7. The modules were based on 2 already established CBT packages: cognitive therapy by Beck [26] and coping with depression by Lewinsohn et al [27]. The modules focused on the following themes: a transactional model of stress and coping (module 1), a self-case formulation based on the cognitive behavioral model (module 2), behavioral activation (module 3), cognitive restructuring (module 4), cognitive restructuring and relaxation (module 5), problem-solving (module 6), and stress and coping for better mental health during the COVID-19 pandemic (module 7). Modules 1-6 had been developed using the modules of the previous iCBT program that had successfully reduced depression, as seen at the 3-month follow-up among nurses in Vietnam [16]. Each module required 5-10 minutes to read, without any exercise or quiz being offered. In the last part of each module, 4-panel comics were provided to engage the participants in the programs. The seventh module was developed afresh to specifically teach awareness of stress and coping with stress during the COVID-19 pandemic, based on a literature review [28-30]. Through a discussion by researchers from Japan, Vietnam, and Thailand, a draft was developed, which consisted of 6 sections [28-30]: (1) job stressors and mental health symptoms of nurses during the COVID-19 pandemic, (2) keeping a healthy lifestyle, (3) communicating with others, (4) applying stress management techniques, (5) seeking professional help if needed, and (6) coping with discrimination and bullying to health care professionals due to COVID-19. Most importantly, the fourth section was designed to introduce how participants could use the CBT skills that they learned in the earlier modules during the COVID-19 pandemic. This draft was reviewed and approved by research teams, including nurses from Vietnam and Thailand, considering the local situations and experiences of nurses during the COVID-19 pandemic. In Vietnam, a smartphone app was prepared and used by the participants to access the program. In Thailand, the participants were asked to access the program using a URL leading to the web program server. Some of the participants in Vietnam who could not install the app used a web-based version. We allowed participants to access the intervention program for an additional 3 weeks to allow those who did not complete all the modules during the initial 7 weeks to study the remaining modules. To facilitate the use of the program, the researchers sent a notice for opening a new module and a reminder to complete an unfinished module to the participants every week during the first 7 weeks of the intervention period in both countries. In Vietnam, the messages were sent via group chat, SMS text messages, and hotline support. In addition, head nurses in the hospitals also reminded the participants through group chats and meetings. In Thailand, the messages were sent only from the researchers via emails, texts, and chat apps individually.

Control as Treatment-As-Usual

The participants allocated to the control group did not receive any intervention program for the first 6 months. Participants in both the intervention and control groups were able to use an internal regular occupational health service at the targeted hospitals. They also could use employee assistance program services if the hospitals offered them. Further, they had access to a mental health care service in their community. The participants in the control group were provided a chance to use the iCBT program that was offered to the intervention group after the 6-month follow-up survey.

Outcomes

Choice of Primary Measure

The primary outcome was the depressive symptoms that were measured by the Depression Anxiety Stress Scale-21 items (DASS21). The DASS21 was administered both in the intervention and control groups at baseline and at 3- and 6-month follow-ups. The DASS21 is a screening tool to assess depression, anxiety, and stress in the preceding 7 days [31]. It is free to use and is widely used for measuring depressive symptoms as the indicator in RCTs among nonclinical populations, including Asian countries. Vietnamese and Thai versions of DASS21 were developed and tested, and their internal consistency (Cronbach α) and validity were confirmed in previous studies [32,33]. In the sample of this study, Cronbach α coefficients of the scale were .83 in Vietnam and .86 in Thailand. The depression subscale comprises 7 items measuring dysphoria, hopelessness, and devaluation of life, among others. Each item is scored on a 4-point scale ranging from 0 (did not apply to me at all) to 3 (applied to me very much, or most of the time). The DASS21 depression score can range from 0 to 42, with higher scores indicating more severe depression; a cut-off score of 10 indicates mild depression [31]. The cut-off scores of moderate, severe, and extremely severe depression are 14, 21, and 28, respectively. Therefore, the 7-point change in DASS21 is considered a clinically significant change.

Process Evaluation

For the process evaluation, the completion rate of each module of the intervention program was collected from the digital records of the server system. A completion rate was calculated as the proportion of the participants who accessed, read, and worked on all the contents of the module until the close of the intervention program (10 weeks from the baseline survey).

XSL•FO
RenderX

Sample Size Calculation

The sample size that was necessary to detect the effects of the intervention on reducing depression was calculated. According to a previous meta-analysis, a web-based universal prevention psychological intervention on reducing depression and anxiety in the workplace yielded a summary effect size of 0.25 [34]. Our study of an unguided universal prevention iCBT program among nurses in Vietnam reported a small effect size on depression (*d*=0.18 at the 3-month follow-up) [16]. To detect a minimal effect size of 0.15 at an α of .05 and power of .80, the estimated sample size was 699 participants in each group. A total of 1398 nurses participated in the study, combining samples from both countries. The statistical power was calculated using the G*Power version 3.1.9.2 [35].

Randomization and Masking

The participants who met the inclusion criteria were randomly allocated to the intervention or control group. The randomization was stratified into 2×6×2 = 24 strata, based on country (Vietnam or Thailand), 6 hospitals, and subgroups based on the cutoff point of mild depression by the DASS21 scores at baseline (10 or greater and less than 10). The stratified permuted block randomization method was used, with a block size of 4. An independent biostatistician generated a password-protected stratified random table using SAS software (SAS Institute Inc). An independent research assistant at the Department of Mental Health, The University of Tokyo, conducted the assignment using the random table, which was blinded to the researchers. Because this trial offered a psychological intervention, it was impossible to blind the results of the allocation to the participants and researchers.

Statistical Analysis

For the main analysis, mixed modeling for repeated measures was used to test the effects of the intervention compared with the control group at 3- and 6-month follow-ups. The interaction effect of the group (1=intervention, 0=control) by time was considered the indicator of the intervention effect. Variables used in the stratified randomization were entered as covariates: countries, hospitals, and subgroups based on the cutoff point of mild depression by DASS21. Intention-to-treat analysis was adopted, which included all the participants in the analysis who completed the baseline survey. Effect sizes were calculated as the standardized mean differences of the score in the DASS21 based on the estimated mean differences at the 3- and 6-month follow-ups. The effectiveness of the intervention was tested in subgroups of nurses in Vietnam and Thailand. There were no missing values during the baseline survey because we asked participants to respond to all the questions by using a web-based survey system. Missing values for DASS21 scores due to attrition were imputed by applying the maximum likelihood estimation of the MIXED procedure of SPSS Statistics (IBM Corp).

We conducted a post hoc subgroup analysis to test the prevention effect of the intervention by examining the difference in the incidences of mild depression (ie, ≥10 on the DASS21 score) among participants with no depression (ie, <10 on the DASS21 score) at the baseline between the intervention and control groups. We applied a logistic regression analysis to estimate the odds ratio of mild depression for the intervention group compared with the control group at the 3- and 6-month follow-ups. Missing values for DASS21 scores due to attrition were imputed by multiple imputations using MULTIPLE IMPUTATION procedure of SPSS statistics. A total of 50 pseudocomplete data sets were generated by the Markov chain Monte Carlo methods using regression models. We referred to pooled estimates of the 50 results. All statistical analyses were conducted using SPSS Statistics (version 29.0; IBM Corp).

Change to the Protocol

The above post hoc subgroup analysis on the incidence of mild depression was conducted to clarify the pure prevention effect of the intervention, responding to a reviewer's request, which was not initially described in the registered protocol [23].

Results

Participant Recruitment

Recruitment and the baseline survey were conducted from March to April 2022 in Thailand and Vietnam. Among the 5560 recruited nurses, 1321 (23.8%) completed the baseline survey. After screening for eligibility, 118 nurses were considered ineligible and excluded. Most ineligible nurses were those who were/planned to go on sick or maternity leave or planned to change jobs (n=98). Following the stratified randomization of the included nurses (N=1203), 602 nurses were allocated to the intervention group and 601 nurses to the control group. The 3- and 6-month follow-up surveys were conducted from July to August 2022 and from October to November 2022, respectively. At the 3-month follow-up, 87.5% (527/602) of the nurses and 85.7% (515/601) of the nurses in the intervention and control groups, respectively, completed the follow-up survey. At the 6-month follow-up, 86% (518/602) of the nurses and 86.4% (519/601) of the nurses in the intervention and control groups, respectively, completed the survey (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of participant recruitment and allocation.

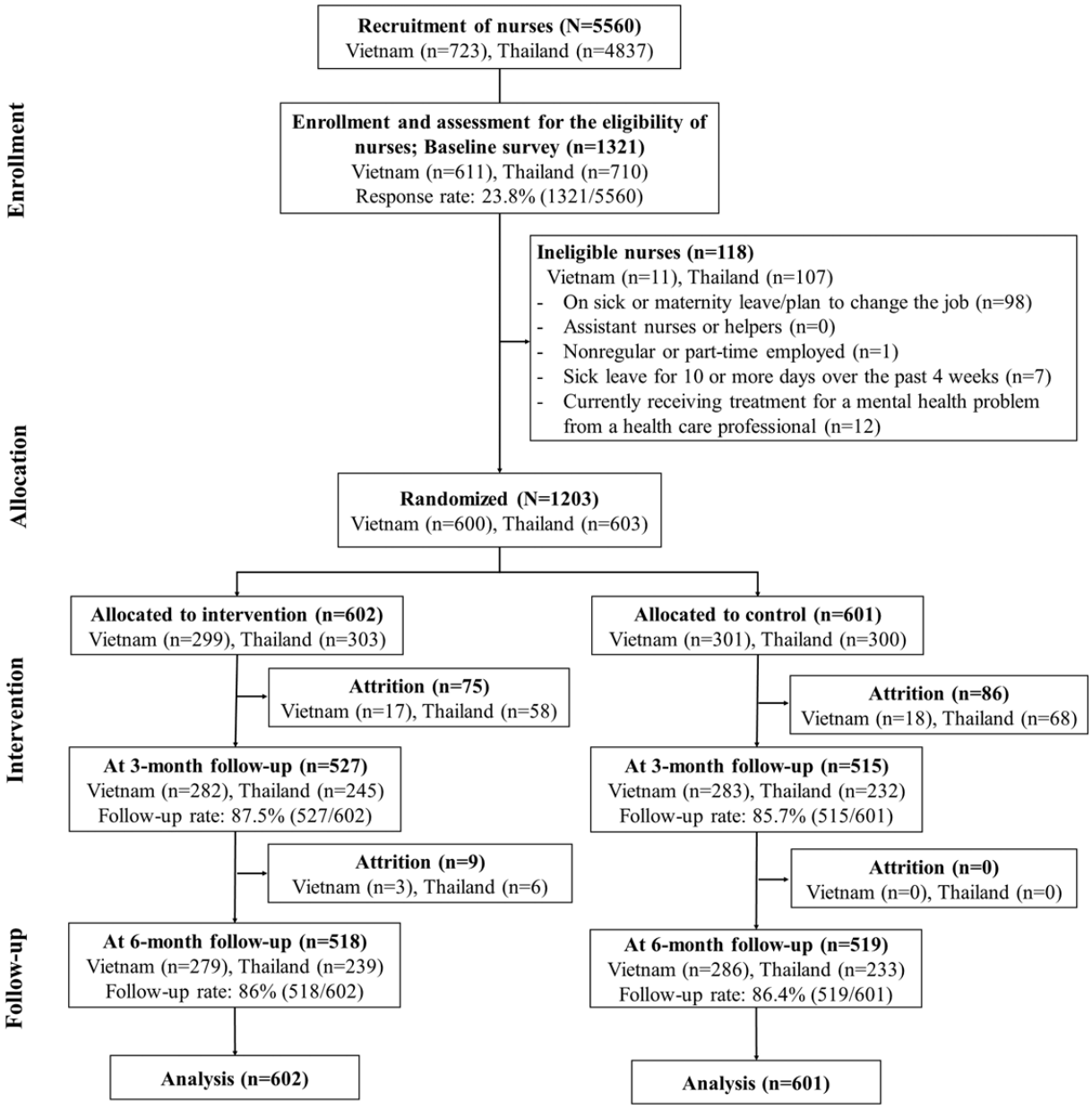

Baseline Characteristics

At baseline, 88.1% (1060/1203) of the participants were females, and 60.9% (733/1203) of them had a spouse (Table 1). Almost all participants (1162/1203) had obtained a college or university degree, and 83.4% (1004/1203) had no-term contracts with hospitals. About 22.6% (272/1203) of the participants indicated having mild or severe depression at baseline. Demographic characteristics were not very different between the 2 groups. Statistical tests (χ^2 tests and 2-sided *t* tests) indicated that the demographic differences were not significant (*P* for differences ranged from .17 to .87).

XSL•FO
RenderX

Table 1. Characteristics of the participants at baseline.

	Intervention group			Control group			*P* value (difference)
	Vietnam (n=299)	Thailand (n=303)	Total (n=602)	Vietnam (n=301)	Thailand (n=300)	Total (n=601)	
Hospital, n (%)							N/A[a]
Vietnam 1	100 (33.4)	N/A	100 (16.6)	100 (33.2)	N/A	100 (16.6)	
Vietnam 2	199 (66.6)	N/A	199 (33.1)	201 (66.8)	N/A	201 (33.4)	
Thailand 1	N/A	78 (25.7)	78 (13)	N/A	75 (25)	75 (12.5)	
Thailand 2	N/A	79 (26.1)	79 (13.1)	N/A	78 (26)	78 (13)	
Thailand 3	N/A	88 (29)	88 (14.6)	N/A	87 (29)	87 (14.5)	
Thailand 4	N/A	58 (19.1)	58 (9.6)	N/A	60 (20)	60 (10)	
Employment contract, n (%)							.87
Fixed-term contract	27 (9)	59 (19.5)	86 (14.3)	26 (8.6)	57 (19)	83 (13.8)	
No-term contract	267 (89.3)	238 (78.5)	505 (83.9)	260 (86.4)	239 (79.7)	499 (83)	
Others	2 (0.7)	6 (2)	8 (1.3)	2 (0.7)	4 (1.3)	6 (1)	
Missing	3 (1)	0 (0)	3 (0.5)	13 (4.3)	0 (0)	13 (2.2)	
Gender, n (%)							.27
Male	60 (20.1)	6 (2)	66 (11)	68 (22.6)	7 (2.3)	75 (12.5)	
Female	239 (79.9)	295 (97.4)	534 (88.7)	233 (77.4)	293 (97.7)	526 (87.5)	
Others	0 (0)	2 (0.7)	2 (0.3)	0 (0)	0 (0)	0 (0)	
Age (years), mean (SD)	36.70 (7.7)	37.19 (10.7)	36.95 (9.3)	36.15 (6.9)	36.69 (9.9)	36.42 (8.5)	.31
Missing, n (%)	0 (0)	0 (0)	0 (0)	1 (0.3)	0 (0)	1 (0.2)	
Marital status, n (%)							.83
Unmarried	40 (13.4)	173 (57.1)	213 (35.4)	45 (15)	173 (57.7)	218 (36.3)	
Have a spouse	248 (82.9)	122 (40.3)	370 (61.5)	251 (83.4)	112 (37.3)	363 (60.4)	
Separated/divorced/widowed	9 (3)	8 (2.6)	17 (2.8)	5 (1.7)	15 (5)	20 (3.3)	
Missing	2 (0.7)	0 (0)	2 (0.3)	0 (0)	0 (0)	0 (0)	
Highest educational qualification, n (%)							.31
Intermediate	25 (8.4)	0 (0)	25 (4.2)	14 (4.7)	0 (0)	14 (2.3)	
College	150 (50.2)	0 (0)	150 (24.9)	149 (49.5)	0 (0)	149 (24.8)	
University	117 (39.1)	252 (83.2)	369 (61.3)	124 (41.2)	259 (86.3)	383 (63.7)	
Postgraduate	7 (2.3)	51 (16.8)	58 (9.6)	12 (4)	41 (13.7)	53 (8.8)	
Missing	0 (0)	0 (0)	0 (0)	2 (0.7)	0 (0)	2 (0.3)	
Income in Vietnam (per month), n (%)[b]							.40
< 2.1 million	3 (1)	N/A	3 (0.5)	1 (0.3)	N/A	1 (0.2)	
2.1 to < 4.2 million	15 (5)	N/A	15 (2.5)	14 (4.7)	N/A	14 (2.3)	
4.2 to < 6.3 million	145 (48.5)	N/A	145 (24.1)	129 (42.9)	N/A	129 (21.5)	
6.3 to < 8.4 million	77 (25.8)	N/A	77 (12.8)	96 (31.9)	N/A	96 (16)	
≥ 8.4 million	59 (19.7)	N/A	59 (9.8)	58 (19.3)	N/A	58 (9.7)	
Income in Thailand (per month), n (%)[c]							.17
<THB 20,000	N/A	2 (0.7)	2 (0.3)	N/A	9 (3)	9 (1.5)	
THB 20,000 to <THB 40,000	N/A	159 (52.5)	159 (26.4)	N/A	159 (53)	159 (26.5)	

XSL•FO
RenderX

	Intervention group			Control group			*P* value (difference)
	Vietnam (n=299)	Thailand (n=303)	Total (n=602)	Vietnam (n=301)	Thailand (n=300)	Total (n=601)	
THB 40,000 to <THB 60,000	N/A	107 (35.3)	107 (17.8)	N/A	107 (35.7)	107 (17.8)	
THB 60,000 to <THB 80,000	N/A	30 (9.9)	30 (5)	N/A	20 (6.7)	20 (3.3)	
≥THB 80,000	N/A	5 (1.7)	5 (0.8)	N/A	5 (1.7)	5 (0.8)	
Missing	0 (0)	0 (0)	0 (0)	3 (1)	0 (0)	3 (0.5)	
DASS21[d] depression subscale score, n (%)							.59
No depression (<10 on the scale)	243 (81.3)	219 (72.3)	462 (76.7)	243 (80.7)	226 (75.3)	469 (78)	
Mild depression (≥10 on the scale)	56 (18.7)	84 (27.7)	140 (23.3)	58 (19.3)	74 (24.7)	132 (22)	

[a]N/A: not applicable.

[b]As of March 2022, US $1= 22,830.

[c]As of March 2022, US $1=THB 33.

[d]DASS21: Depression Anxiety Stress Scale-21 items.

Effects of the Intervention Program on Depression

In both groups, the average scores of depression at the 3- and 6-month follow-ups were lower than those at the baseline (Table 2). The scores in the intervention group decreased steeply at the 3-month follow-up and returned slightly to the baseline scores. The scores in the control group decreased monotonically. Cohen *d* scores calculated from the descriptive statistics were –0.13 (95% CI –0.26 to –0.02) at the 3-month follow-up and –0.05 (95% CI –0.17 to –0.02) at the 6-month follow-up. For the country subgroups, Cohen *d* scores at the 3-month follow-up were almost the same in Thailand (*d*=–0.15, 95% CI –0.33 to 0.03) and Vietnam (*d*=–0.13, 95% CI –0.31 to 0.03), while the scores at the 6-month follow-up in Thailand (*d*=–0.12, 95% CI –0.31 to 0.06) were greater than those in Vietnam (*d*=0.02, 95% CI –0.14 to 0.19). The number of missing values for depression due to attrition was not significantly different between the intervention and the control groups (*P*=.82). Compared with those who were survey completers, the participants who had fixed-term contracts (*P*<.001), were young (*P*=.02), were unmarried (*P*<.001), and had postgraduate qualifications (*P*<.001) tended to drop out from the follow-up surveys. The difference in the rate of mild or severe depression at baseline was not significant (*P*=.10).

Table 2. Mean (SD) of Depression Anxiety and Stress Scale-21 items depression subscale scores at baseline and at the 3- and 6-month follow-ups.

	Intervention group (n=602)				Control group (n=601)				Effect size (intervention vs control), Cohen *d* (95% CI)		
	V[a], mean (SD)	T[b], mean (SD)	Total, mean (SD)	Missing, n (%)	V, mean (SD)	T, mean (SD)	Total, mean (SD)	Missing, n (%)	V	T	Total
Baseline	5.05 (5.6)	6.27 (6.1)	5.66 (5.9)	N/A[c]	4.92 (5.3)	6.40 (6.5)	5.66 (6.0)	N/A	N/A	N/A	N/A
3-month follow-up	4.07 (4.8)	3.40 (5.3)	3.75 (5.0)	101 (16.8)	4.69 (5.4)	4.46 (6.4)	4.58 (5.9)	80 (13.3)	–0.13 (–0.31 to 0.03)	–0.15 (–0.33 to 0.03)	–0.13 (–0.26 to –0.02)
6-month follow-up	4.75 (6.2)	3.04 (4.8)	3.95 (5.6)	86 (14.3)	4.49 (6.2)	3.96 (6.1)	4.25 (6.1)	89 (14.8)	0.02 (–0.14 to 0.19)	–0.12 (–0.31 to 0.06)	–0.05 (–0.17 to 0.07)

[a]V: Vietnam.

[b]T: Thailand.

[c]N/A: not applicable.

Following mixed modeling (Table 3), the group difference in the score change was significant and nonsignificant at the 3-month (coefficient=–0.92, 95% CI –1.66 to –0.18; *P*=.02) and 6-month follow-ups (coefficient=–0.33, 95% CI –1.11 to 0.45; *P*=.41), respectively. The estimated effect sizes were –0.15 and –0.06 at the 3- and 6-month follow-ups, respectively. For the country subgroups, a significant group difference was observed at the 3-month follow-up in Thailand (coefficient=–1.24, 95% CI –2.26 to –0.23; *P*=.02) and about twice as greater than that in Vietnam (coefficient=–0.66, 95% CI –1.70 to 0.39; *P*=.22). The estimated effect size at the 3-month follow-up in Thailand was –0.21. The coefficients at the 6-month follow-up were nonsignificant in either country; the direction of the effect was maintained in Thailand

XSL•FO
RenderX

(coefficient=–0.83, 95% CI –1.78 to 0.12; *P*=.09) but reversed in Vietnam (coefficient=0.12, 95% CI –1.04 to 1.29; *P*=.84).

Table 3. Effects of the intervention program on depression.

	Coefficient[a] (95% CI)	SE	*t* test *(df)*	*P* value	Estimated effect size (95% CI)[b]
Total sample (N=1203)					
3-month follow-up	–0.92 (–1.66 to –0.18)	0.38	–2.44 (1331.34)	.02	–0.15 (–0.28 to –0.03)
6-month follow-up	–0.33 (–1.11 to 0.45)	0.40	–0.83 (1296.85)	.41	–0.06 (–0.19 to 0.08)
Pooled[c]	–0.06 (–0.19 to 0.08)	0.07	–0.85 (1196.36)	.39	–0.01 (–0.03 to 0.01)
Vietnam (n=600)					
3-month follow-up	–0.66 (–1.70 to 0.39)	0.53	–1.23 (709.20)	.22	–0.11 (–0.29 to 0.07)
6-month follow-up	0.12 (–1.04 to1.29)	0.60	0.21 (696.80)	.84	0.02 (–0.18 to 0.22)
Pooled	0.14 (–0.06 to 0.34)	0.10	0.14 (650.99)	.89	0.02 (–0.01 to 0.06)
Thailand (n=603)					
3-month follow-up	–1.24 (–2.26 to –0.23)	0.52	–2.40 (636.48)	.02	–0.21 (–0.38 to –0.04)
6-month follow-up	–0.83 (–1.78 to 0.12)	0.49	–1.70 (608.97)	.09	–0.14 (–0.30 to 0.02)
Pooled	–0.13 (–0.30 to 0.03)	0.08	–1.60 (541.01)	.11	–0.02 (–0.05 to 0.01)

[a]Coefficients were adjusted by the variables used in the stratified randomization: country (Vietnam and Thailand), hospital (6 hospitals), and severity of depression (≥10 or <10 on the Depression Anxiety Stress Scale-21 items depression subscale).

[b]Effect sizes were calculated by dividing the estimated effect by the pooled SD.

[c]Mixed models for repeated measures conditional growth model analyses were conducted.

Among the participants with no depression at the baseline (n=931), 85 (9.1%) reported mild depression at the 3-month follow-up. The proportions of new cases of mild depression were 7.8% (36/462) in the intervention group and 10.4% (49/469) in the control group at the 3-month follow-up. For the result of logistic regression, the pooled estimate of odds ratio for mild depression of the intervention group compared with the control group was 0.70 (95% CI 0.45-1.10; *P*=.12). At the 6-month follow-up, 88 (9.5%) reported mild depression. Proportions of mild depression were 8.9% (41/462) in the intervention group and 10% (47/469) in the control group. The pooled estimate of odds ratio was 0.94 (95% CI 0.61-1.45; *P*=.79).

Process Evaluation

The completion rates in each module ranged from 87.5% (527/602) to 74.3% (447/602) and decreased monotonically from modules 1 to 7 (Table 4). Approximately 68.1% (410/602) of the participants completed all the 7 modules. The completion rates of modules 1 to 5 in Vietnam were significantly higher than those in Thailand (Table 4). There were no significant differences in the rates for modules 6 and 7 as well as the proportions of completers of all the 7 modules between Vietnam and Thailand.

XSL•FO
RenderX

Table 4. Completion rates for each module and all 7 modules.

	Vietnam (n=299), n (%)	Thailand (n=303), n (%)	Total (n=602), n (%)	Difference in 2 countries, χ^2 (*df*)	*P* value
Module 1				9.1 (1)	.003
Complete	274 (91.6)	253 (83.5)	527 (87.5)		
Incomplete	25 (8.4)	50 (16.5)	75 (12.5)		
Module 2				11.9 (1)	<.001
Complete	272 (91)	246 (81.2)	518 (86)		
Incomplete	27 (9)	57 (18.8)	84 (14)		
Module 3				12.2 (1)	<.001
Complete	270 (90.3)	243 (80.2)	513 (85.2)		
Incomplete	29 (9.7)	60 (19.8)	89 (14.8)		
Module 4				6.6 (1)	.01
Complete	262 (87.6)	242 (79.9)	504 (83.7)		
Incomplete	37 (12.4)	61 (20.1)	98 (16.3)		
Module 5				5.7 (1)	.02
Complete	258 (86.3)	239 (78.9)	497 (82.6)		
Incomplete	41 (13.7)	64 (21.1)	105 (17.4)		
Module 6				0.3 (1)	.57
Complete	225 (75.3)	234 (77.2)	459 (76.2)		
Incomplete	74 (24.7)	69 (22.8)	143 (23.8)		
Module 7				3.5 (1)	.06
Complete	212 (70.9)	235 (77.6)	447 (74.3)		
Incomplete	87 (29.1)	68 (22.4)	155 (25.7)		
All 7 modules				1.8 (1)	.18
Complete	196 (65.6)	214 (70.6)	410 (68.1)		
Incomplete	103 (34.4)	89 (29.4)	192 (31.9)		

Discussion

Principal Results

The findings of our multi-country RCT show that the smartphone-based self-guided iCBT program, "The ABC Stress Management–COVID-19 version," significantly reduced depression at the 3-month follow-up among hospital nurses in Vietnam and Thailand. However, the estimated effect size was small (–0.15), and therefore, these results may not be clinically meaningful. Furthermore, the effect of the program leveled off and was nonsignificant at the 6-month follow-up. The incidence of mild depression was reduced by 30% among participants with no depression at the baseline in the intervention group at the 3-month follow-up, but this prevention effect was not statistically significant. The effect of the program in Thailand was slightly greater than that in Vietnam, while the pattern of change in the depression scores was similar. About two-third or more of the participants completed all the 7 modules during the intervention period.

This is the first study, to the best of our knowledge, that has demonstrated the effect of a smartphone-based iCBT universal program for prevention of depression among nurses during the COVID-19 pandemic. Our findings are consistent with those of a previous RCT on the effectiveness of an internet-based program for psychological distress among non–health care workers during the COVID-19 pandemic [36]. A self-guided smartphone-based stress management program may be effective for reducing depression among hospital nurses during the COVID-19 pandemic. A factor for this positive finding could be the high completion rate of the program: the process evaluation revealed that more than 80% of the participants in the intervention group completed the minimum intervention (module 1) and more than two-thirds completed all the modules. This was possibly due to the intensive follow-up procedures that we used during the intervention period. In a similar previous study [19], which failed to show the effect of an internet stress management program, there might not have been enough high intensity for the follow-up to achieve a significant finding on the effectiveness; however, the completion rate was not reported.

The effect of our program attenuated at the 6-month follow-up. A similar pattern was observed in our previous RCT using a similar self-guided smartphone-based stress management program among hospital nurses in Vietnam before the COVID-19 pandemic [16]. When the intervention is

XSL•FO
RenderX

nonintensive such as a self-guided program, a web-based psychosocial intervention may be effective for a short period but not for a longer period [37]. However, a group-based psychological first aid training program was reported to improve professional efficacy at the 6-month follow-up among health care workers during an Ebola outbreak in Africa [4]. The attenuation of the effect might be attributable to the low-intensity, self-guided nature of the program. Probably, the participants were not able to retain their learnings from the program in their memory for a longer duration. The other possibility is that the overwhelming stressful experiences that they may have had due to the COVID-19 pandemic could have limited them from using coping strategies learned from the program for a longer period [38]. One study showed that a group-based psychological intervention for the mental health of mothers after a nuclear disaster was effective for 1 month but not for 3 months [39]. Therefore, the severity of a disaster experience might diminish the long-term effect of a stress management program. This should be investigated further by, for instance, looking at the moderating effect of disaster-related (ie, COVID-19–related in this context) stressful experiences during a postintervention follow-up on the effect of a stress management program.

Although the effect of our program on depression was similar in Vietnam and Thailand, the effect in Thailand was slightly larger than that in Vietnam. This may be attributed to the differences in the sampling process: in Vietnam, all the nurses from the collaborating target hospitals were potential participants, while in Thailand, selected nurses from target hospitals were recruited for this study. The participants in Thailand may have included those who were more motivated to participate in the program, thereby enhancing the effect of the program. The other possibility could be the different timings and patterns of COVID-19 outbreaks and related responsiveness between Vietnam and Thailand. At the country level, the total number of confirmed cases in Vietnam was much more than that in Thailand in 2022, although the peaks of the outbreak were similar around March 2022 [21]. It is possible that COVID-19 hit the target hospitals in Vietnam more severely than that in Thailand, and participants in Vietnam may have been more overwhelmed by the situation and therefore less able to cope using what they learned in the program. The completion rates of all 7 modules by nurses in Thailand were greater than those by nurses in Vietnam; in particular, higher proportions of participants in Thailand studied module 7 (coping with COVID-19 stress), while the completion rates for modules 1 to 5 were higher in Vietnam. Module 7 (stress and coping for better mental health during the COVID-19 pandemic) completion rate was higher in Thailand. Module 7 might suit the participants' stressful situations during the COVID-19 pandemic period and lead to higher effectiveness among nurses in Thailand. This, in turn, may have led to slightly better findings among nurses in Thailand. A module connecting CBT skills with stress in the COVID-19 pandemic may assist participants in understanding how to cope with stress in the COVID-19 situation.

Limitations

This study had several limitations. The follow-up rates were relatively high at 3- and 6-month follow-ups in both groups. However, there was a possible attrition bias that may affect the findings. If the less motivated participants dropped out, the effects of the program would be overestimated. Because the iCBT program was a psychosocial intervention, it was impossible to blind the participants to the group. The intervention group could have been more motivated to report their depression as less than the exact level at the follow-ups. If the participants in the intervention group shared the information about the program content with the participants in the control group, the observed effect might have been underestimated. Generalization of the findings may also be limited. Our findings were obtained from nurses from a limited number of large provincial hospitals and may not apply to other countries, including lower or higher-income countries and countries in other regions and with different cultures. In addition, because the participants in Thailand were volunteers who were motivated to complete the program, the findings may not be generalizable to the entire population of nurses at the target hospitals.

Conclusions

Our study demonstrates that the smartphone-based iCBT program was effective in improving depression at the 3-month follow-up among hospital nurses in Vietnam and Thailand during the COVID-19 pandemic. This program was self-guided, not requiring support from professional therapists. This program did not need face-to-face contact to study the content. Thus, it could be disseminated to a large population of hospital nurses in low-resource settings during the outbreak crisis, with social distancing restrictions and reduced risk of infection. In addition, this program is suitable for health care workers who work irregular schedules because it can be completed at their convenient time. This program may be beneficial for the mental health and well-being of nurses who face stress due to the COVID-19 pandemic in these countries. It may also contribute to developing a psychologically resilient nursing workforce during such challenging times, which would help maintain health care services to patients. This program should be evaluated for its effectiveness in other middle-income countries in Southeast Asia and during other infectious disease outbreaks. Nonetheless, it is a promising approach for reducing the depression of nurses in a future infectious disease pandemic.

Acknowledgments

This study was supported by the grants-in-aid for scientific research (KAKENHI) fund for the Promotion of Joint International Research (Fostering Joint International Research B) from Japan Society for the Promotion of Science (JSPS) 2020-2022 (JSPS KAKENHI grant/award: 20KK0215). The funder had no role in this study's design; collection, analysis, or interpretation of data; writing of the report; or decision to submit this paper for publication. The authors declare that there was no use of generative artificial intelligence technology in the generation of texts, figures, or other informational content in this paper.

XSL•FO
RenderX

Authors' Contributions

HTN, OK, and NK contributed to conceptualization. KW, KI, PB, YM, TTTT, QTN, NTN, HTN, OK, and NK contributed to the methodology. KW, TTTT, N Sripo, and N Sasaki analyzed the data. KW, TTTT, N Sripo, AS, and NK contributed to writing the original draft. KW, TTTT, N Sripo, AS, KI, PB, N Sasaki, TT, HA, MI, QTN, NTN, STV, TTN, TTL, LDN, NTVN, BTN, YM, YT, DN, AT, HTN, OK, and NK contributed to editing the manuscript. All authors confirm that they had full access to all the data in the study and accept responsibility for submission of the manuscript for publication.

Conflicts of Interest

NK reports personal fees from Junpukai Foundation, Riken Institute, JAXA, Sekisui Chemical Co. Ltd, and SB@Work outside the submitted work. NK, KI, and AS are employed with the Department of Digital Mental Health, an endowment department that is supported by an unrestricted grant from 15 enterprises [40].

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1154 KB - jmir_v26i1e50071_app1.pdf]

Multimedia Appendix 2
Screenshot of The ABC Stress Management–COVID-19 version.
[PNG File , 342 KB - jmir_v26i1e50071_app2.png]

References

1. Cénat JM, Blais-Rochette C, Kokou-Kpolou CK, Noorishad P, Mukunzi JN, McIntee S, et al. Prevalence of symptoms of depression, anxiety, insomnia, posttraumatic stress disorder, and psychological distress among populations affected by the COVID-19 pandemic: A systematic review and meta-analysis. Psychiatry Res 2021 Jan;295:113599 [FREE Full text] [doi: 10.1016/j.psychres.2020.113599] [Medline: 33285346]
2. Lee BEC, Ling M, Boyd L, Olsson C, Sheen J. The prevalence of probable mental health disorders among hospital healthcare workers during COVID-19: A systematic review and meta-analysis. J Affect Disord 2023 Jun 01;330:329-345 [FREE Full text] [doi: 10.1016/j.jad.2023.03.012] [Medline: 36931567]
3. Serrano-Ripoll MJ, Meneses-Echavez JF, Ricci-Cabello I, Fraile-Navarro D, Fiol-deRoque MA, Pastor-Moreno G, et al. Impact of viral epidemic outbreaks on mental health of healthcare workers: a rapid systematic review and meta-analysis. J Affect Disord 2020 Dec 01;277:347-357 [FREE Full text] [doi: 10.1016/j.jad.2020.08.034] [Medline: 32861835]
4. Sijbrandij M, Horn R, Esliker R, O'May F, Reiffers R, Ruttenberg L, et al. The effect of psychological first aid training on knowledge and understanding about psychosocial support principles: a cluster-randomized controlled trial. Int J Environ Res Public Health 2020 Jan 11;17(2):484 [FREE Full text] [doi: 10.3390/ijerph17020484] [Medline: 31940865]
5. Chen R, Chou K, Huang Y, Wang T, Liu S, Ho L. Effects of a SARS prevention programme in Taiwan on nursing staff's anxiety, depression and sleep quality: a longitudinal survey. Int J Nurs Stud 2006 Feb;43(2):215-225 [FREE Full text] [doi: 10.1016/j.ijnurstu.2005.03.006] [Medline: 15927185]
6. Aiello A, Khayeri M, Raja S, Peladeau N, Romano D, Leszcz M, et al. Resilience training for hospital workers in anticipation of an influenza pandemic. J Contin Educ Health Prof 2011;31(1):15-20. [doi: 10.1002/chp.20096] [Medline: 21425355]
7. Maunder RG, Lancee WJ, Mae R, Vincent L, Peladeau N, Beduz MA, et al. Computer-assisted resilience training to prepare healthcare workers for pandemic influenza: a randomized trial of the optimal dose of training. BMC Health Serv Res 2010 Mar 22;10(1):72 [FREE Full text] [doi: 10.1186/1472-6963-10-72] [Medline: 20307302]
8. Hooper JJ, Saulsman L, Hall T, Waters F. Addressing the psychological impact of COVID-19 on healthcare workers: learning from a systematic review of early interventions for frontline responders. BMJ Open 2021 May 21;11(5):e044134 [FREE Full text] [doi: 10.1136/bmjopen-2020-044134] [Medline: 34020974]
9. Shatri H, Putranto R, Faisal E, Fitrianti F, Adhiatma K, Zulmiyusrini P, et al. Psychotherapy for healthcare provider during COVID-19 pandemic: an evidence based clinical review. Acta Med Indones 2021;53(2):225-232. [Medline: 34251353]
10. David E, DePierro JM, Marin DB, Sharma V, Charney DS, Katz CL. COVID-19 pandemic support programs for healthcare workers and implications for occupational mental health: a narrative review. Psychiatr Q 2022 Mar;93(1):227-247 [FREE Full text] [doi: 10.1007/s11126-021-09952-5] [Medline: 34606067]
11. Lekagul A, Piancharoen P, Chattong A, Suradom C, Tangcharoensathien V. Living through the psychological consequences of COVID-19 pandemic: a systematic review of effective mitigating interventions. BMJ Open 2022 Jul 26;12(7):e060804 [FREE Full text] [doi: 10.1136/bmjopen-2022-060804] [Medline: 35882462]
12. Wan Mohd Yunus WMA, Musiat P, Brown JSL. Systematic review of universal and targeted workplace interventions for depression. Occup Environ Med 2018 Jan;75(1):66-75. [doi: 10.1136/oemed-2017-104532] [Medline: 29074553]
13. Liu S, Yang L, Zhang C, Xiang YT, Liu Z, Hu S, et al. Online mental health services in China during the COVID-19 outbreak. Lancet Psychiatry 2020 Apr;7(4):e17-e18 [FREE Full text] [doi: 10.1016/S2215-0366(20)30077-8] [Medline: 32085841]

XSL•FO
RenderX

14. Hersch RK, Cook RF, Deitz DK, Kaplan S, Hughes D, Friesen MA, et al. Reducing nurses' stress: A randomized controlled trial of a web-based stress management program for nurses. Appl Nurs Res 2016 Nov;32:18-25 [FREE Full text] [doi: 10.1016/j.apnr.2016.04.003] [Medline: 27969025]
15. Yoo IG. The effects of the type of delivery of cognitive-behavioral therapy for healthcare workers: A systematic review. J Clin Psychol 2022 Feb;78(2):149-166. [doi: 10.1002/jclp.23215] [Medline: 34240422]
16. Imamura K, Tran TTT, Nguyen HT, Sasaki N, Kuribayashi K, Sakuraya A, et al. Effect of smartphone-based stress management programs on depression and anxiety of hospital nurses in Vietnam: a three-arm randomized controlled trial. Sci Rep 2021 May 31;11(1):11353 [FREE Full text] [doi: 10.1038/s41598-021-90320-5] [Medline: 34059737]
17. Drissi N, Ouhbi S, Marques G, de la Torre Díez I, Ghogho M, Janati Idrissi MA. A systematic literature review on e-mental health solutions to assist health care workers during COVID-19. Telemed J E Health 2021 Jun;27(6):594-602. [doi: 10.1089/tmj.2020.0287] [Medline: 32970532]
18. Zhang H, Yang Y, Hao X, Qin Y, Li K. Effects of digital sleep interventions on sleep and psychological health during the COVID-19 pandemic: A systematic review and meta-analysis. Sleep Med 2023 Oct;110:190-200. [doi: 10.1016/j.sleep.2023.07.036] [Medline: 37625326]
19. Fiol-DeRoque MA, Serrano-Ripoll MJ, Jiménez R, Zamanillo-Campos R, Yáñez-Juan AM, Bennasar-Veny M, et al. A mobile phone-based intervention to reduce mental health problems in health care workers during the COVID-19 pandemic (PsyCovidApp): randomized controlled trial. JMIR Mhealth Uhealth 2021 May 18;9(5):e27039 [FREE Full text] [doi: 10.2196/27039] [Medline: 33909587]
20. Dumarkaite A, Truskauskaite I, Andersson G, Jovarauskaite L, Jovaisiene I, Nomeikaite A, et al. The efficacy of the internet-based stress recovery intervention FOREST for nurses amid the COVID-19 pandemic: A randomized controlled trial. Int J Nurs Stud 2023 Feb;138:104408 [FREE Full text] [doi: 10.1016/j.ijnurstu.2022.104408] [Medline: 36527859]
21. Guilbert JJ. The World Health Report 2006: working together for health. Educ Health (Abingdon) 2006 Nov;19(3):385-387. [doi: 10.1080/13576280600937911] [Medline: 17178522]
22. Number of COVID-19 cases reported to WHO. WHO. URL: https://covid19.who.int [accessed 2023-06-06]
23. Kawakami N, Nguyen HT, Kaewboonchoo O, Imamura K, Tran TTT, Boonyamalik P, et al. Effect of a smartphone-based stress management program on depression and other health- and work-related outcomes among hospital nurses in the COVID-19 pandemic in Vietnam and Thailand: A protocol of a randomized controlled trial of the COCONATS Project. Preprints. Preprint posted online on March 27, 2023 [FREE Full text] [doi: 10.20944/preprints202303.0450.v1]
24. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
25. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
26. Beck A. Cognitive Therapy of Depression. New York: Guilford Press; 1979.
27. Lewinsohn PM, Antonuccio DO, Steinmetz J, et al. The Coping With Depression Course: A Psychoeducational Intervention for Unipolar Depression. Eugene, OR: Castalia Pub Co; 1984.
28. Basic psychosocial skills: a guide for COVID-19 responders. IASC. 2020. URL: https://interagencystandingcommittee.org/sites/default/files/migrated/2020-05/Basic%20Psychosocial%20Skills-%20A%20Guide%20for%20COVID-19%20Responders.pdf [accessed 2023-06-06]
29. Support for public health workers and health professionals. US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/mentalhealth/public-health-workers/index.html [accessed 2023-06-06]
30. COVID-19: Support for patients and carers. UK Royal College of Psychiatrists. URL: https://www.rcpsych.ac.uk/mental-health/treatments-and-wellbeing/covid-19-and-mental-health [accessed 2023-06-06]
31. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol 2005 Jun;44(Pt 2):227-239. [doi: 10.1348/014466505X29657] [Medline: 16004657]
32. Tran TD, Tran T, Fisher J. Validation of the depression anxiety stress scales (DASS) 21 as a screening instrument for depression and anxiety in a rural community-based cohort of northern Vietnamese women. BMC Psychiatry 2013 Jan 12;13(1):24. [doi: 10.1186/1471-244x-13-24]
33. Supasiri T, Lertmaharit S, Rattananupong T, Kitidumrongsuk P, Lohsoonthorn V. Mental health status and quality of life of the elderly in rural Saraburi. Chula Med J 2019;63(2):95-101. [doi: 10.58837/chula.cmj.63.2.5]
34. Carolan S, Harris PR, Cavanagh K. Improving employee well-being and effectiveness: systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. J Med Internet Res 2017 Jul 26;19(7):e271 [FREE Full text] [doi: 10.2196/jmir.7583] [Medline: 28747293]
35. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 2007 May;39(2):175-191. [doi: 10.3758/BF03193146]
36. Imamura K, Sasaki N, Sekiya Y, Watanabe K, Sakuraya A, Matsuyama Y, et al. The effect of the Imacoco care psychoeducation website on improving psychological distress among workers during the COVID-19 pandemic: randomized controlled trial. JMIR Form Res 2022 Mar 10;6(3):e33883 [FREE Full text] [doi: 10.2196/33883] [Medline: 35133972]

XSL•FO
RenderX

37. Imamura K, Kawakami N, Tsuno K, Tsuchiya M, Shimada K, Namba K. Effects of web-based stress and depression literacy intervention on improving symptoms and knowledge of depression among workers: A randomized controlled trial. J Affect Disord 2016 Oct;203:30-37. [doi: 10.1016/j.jad.2016.05.045] [Medline: 27280960]
38. Tran HTT, Nguyen YH, Vuong TD, Bui LV, Doan HT, Le HTT, et al. High prevalence of post-traumatic stress disorder and psychological distress among healthcare workers in COVID-19 field hospitals: a cross-sectional study from Vietnam. Psychol Res Behav Manag 2023 May;16:1663-1675. [doi: 10.2147/prbm.s407583]
39. Imamura K, Sekiya Y, Asai Y, Umeda M, Horikoshi N, Yasumura S, et al. The effect of a behavioral activation program on improving mental and physical health complaints associated with radiation stress among mothers in Fukushima: a randomized controlled trial. BMC Public Health 2016 Nov 08;16(1):1144 [FREE Full text] [doi: 10.1186/s12889-016-3819-x] [Medline: 27825326]
40. Participating companies. Department of Digital Mental Health, Graduate School of Medicine, The University of Tokyo. URL: https://dmh.m.u-tokyo.ac.jp/en [accessed 2024-07-09]

Abbreviations

COCONATS: Coping with COVID-19 helps Nurses be Active, Tough, and Smiling
CONSORT: Consolidated Standards of Reporting Trials
DASS21: Depression Anxiety Stress Scale-21 items
iCBT: internet-based cognitive behavioral therapy
RCT: randomized controlled trial

Edited by T de Azevedo Cardoso; submitted 19.06.23; peer-reviewed by A Pienkowska, R Pearson; comments to author 02.11.23; revised version received 25.12.23; accepted 05.06.24; published 30.08.24.

Please cite as:
Watanabe K, Tran TTT, Sripo N, Sakuraya A, Imamura K, Boonyamalik P, Sasaki N, Tienthong T, Asaoka H, Iida M, Nguyen QT, Nguyen NT, Vu ST, Ngo TT, Luyen TT, Nguyen LD, Nguyen NTV, Nguyen BT, Matsuyama Y, Takemura Y, Nishi D, Tsutsumi A, Nguyen HT, Kaewboonchoo O, Kawakami N
Effectiveness of a Smartphone-Based Stress Management Program for Depression in Hospital Nurses During COVID-19 in Vietnam and Thailand: 2-Arm Parallel-Group Randomized Controlled Trial
J Med Internet Res 2024;26:e50071
URL: https://www.jmir.org/2024/1/e50071
doi:10.2196/50071
PMID:

©Kazuhiro Watanabe, Thuy Thi Thu Tran, Narisara Sripo, Asuka Sakuraya, Kotaro Imamura, Plernpit Boonyamalik, Natsu Sasaki, Thanate Tienthong, Hiroki Asaoka, Mako Iida, Quynh Thuy Nguyen, Nga Thi Nguyen, Son Thai Vu, Thuy Thi Ngo, Tham Thi Luyen, Long Duc Nguyen, Nga Thi Viet Nguyen, Binh Thanh Nguyen, Yutaka Matsuyama, Yukie Takemura, Daisuke Nishi, Akizumi Tsutsumi, Huong Thanh Nguyen, Orawan Kaewboonchoo, Norito Kawakami. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Clinical Effects of Asynchronous Provider-Guided Practice Sessions During Blended Care Therapy for Anxiety and Depression: Pragmatic Retrospective Cohort Study

Hallie M Espel-Huynh[1], PhD; Lu Wang[1], PhD; Emily G Lattie[1], PhD; Robert E Wickham[2], PhD; Alethea Varra[1], PhD; Connie E Chen[1], MD; Anita Lungu[1], PhD; Jennifer L Lee[1], PhD

[1]Lyra Health, Burlingame, CA, United States

[2]Department of Psychological Sciences, Northern Arizona University, Flagstaff, AZ, United States

Corresponding Author:
Hallie M Espel-Huynh, PhD
Lyra Health
270 East Lane
Burlingame, CA, 94010
United States
Phone: 1 (877) 505 7147
Email: hespel-huynh@lyrahealth.com

Abstract

Background: Blended care therapy models are intended to increase the efficiency and effectiveness of evidence-based psychotherapy by combining synchronous and asynchronous components of care.

Objective: This retrospective cohort study evaluated the clinical effects of synchronous video therapy sessions and asynchronous guided practice session elements on anxiety and depression in a blended care therapy program, with a novel focus on asynchronous provider feedback messages.

Methods: Participants were adults (N=33,492) with clinical symptoms of anxiety (Generalized Anxiety Disorder 7-item scale [GAD-7] score of ≥8) and depression (Patient Health Questionnaire 9-item scale [PHQ-9] score of ≥10) at intake. Symptom trajectories were evaluated via individual growth curve models. Time-varying covariates evaluated effects of synchronous video session attendance and the presence or absence of each asynchronous guided practice session element occurring within 7 days and 8-14 days prior to each clinical outcome assessment. Guided practice session elements included client digital lesson completion, client digital exercise completion, and feedback messages sent by providers.

Results: Approximately 86.6% (29,012/33,492) of clients met criteria for clinical improvement by end of care (median 6, IQR 4-8 synchronous sessions). Synchronous video session attendance and client digital lesson completion in the past 7 days and in the past 8-14 days were each uniquely and significantly associated with lower GAD-7 scores (video session effects: $b_{session7}$=–0.82, $b_{session8\text{-}14}$=–0.58, *P* values<.001; digital lesson effects: $b_{lesson7}$=–0.18, $b_{lesson8\text{-}14}$=–0.26, *P* values <.001) and PHQ-9 scores (video session effects: $b_{session7}$=–0.89, $b_{session8\text{-}14}$=–0.67, *P* values <.001; digital lesson effects: $b_{lesson7}$=–0.12, $b_{lesson8\text{-}14}$=–0.30, *P* values <.001). Client digital exercise completion in the past 8-14 days was significantly associated with lower GAD-7 scores ($b_{exercise8\text{-}14}$=–0.10; *P*<.001) but exercise completion in the 7 days prior to clinical outcome assessment was not ($b_{exercise7}$=0.00; *P*=.89). Exercise completion in the past 7 days was significantly associated with lower PHQ-9 scores ($b_{exercise7}$=–0.16; *P*<.001) but exercise completion in the past 8-14 days was not ($b_{exercise8\text{-}14}$=–0.05; *P*=.09). Provider feedback messaging in the past 7 days and in the past 8-14 days was significantly associated with lower GAD-7 and PHQ-9 scores, respectively (GAD-7: $b_{feedback7}$=–0.12, *P*<.001; $b_{feedback8\text{-}14}$=–0.07, *P*=.004; PHQ-9: $b_{feedback7}$=–0.15, *P*<.001; $b_{feedback8\text{-}14}$=–0.08, *P*=.01).

Conclusions: Provider feedback between synchronous therapy sessions provided significant benefit for symptom reduction, beyond the effects of client digital engagement and synchronous video sessions. When guided practice sessions are well integrated into care, blended care therapy provides meaningful improvements upon the traditional, synchronous session–only therapy model. Provider guidance and feedback for clients between synchronous sessions support more efficient and effective mental health care overall.

(J Med Internet Res 2024;26:e60502) doi:10.2196/60502

XSL•FO
RenderX

KEYWORDS

anxiety; depression; blended care therapy; guided practice sessions; mental health care; digital mental health; psychotherapy outcomes

Introduction

Evidence-based psychological treatments for anxiety and depression (eg, cognitive behavioral therapy) are highly effective in reducing symptoms and improving functioning [1-3]. These evidence-based psychotherapies include "active elements" such as therapy concepts, emotion regulation skills, and behavior change skills that impart unique therapeutic benefits [4,5]. Clients benefit from these active elements through a multistage process of provider sharing, client understanding, and client application of skills and concepts [5]. In traditional therapy models, therapists *share* or teach the active elements of these treatments solely during live, synchronous, once-weekly sessions. To fully benefit, clients must take in and *understand* this information as it pertains to their experience and then *apply* it independently via "homework" assignments that are completed between therapy sessions [5]. More consistent homework completion is associated with greater symptom reductions in psychotherapy [6,7]. However, multiple inefficiencies to this model have been identified. In the traditional model, clients often struggle to complete homework independently, with less than 40% of homework typically completed as assigned [8,9]. In addition, therapists follow up on homework with clients only approximately 50% of the time, which reduces opportunities to reinforce and consolidate client learning [10,11]. Without sufficient out-of-session skills practice, clients tend to experience suboptimal outcomes, including increased likelihood of premature treatment dropout, slower symptom improvements (presumably due to slower skills acquisition), and ultimately worse symptoms at end of care [7,12-14]. The traditional therapy model is therefore limited in the extent to which it supports client understanding, retention, and application of the skills and concepts that are necessary for treatment effectiveness [11].

Blended care therapy models have emerged as a potential solution to these challenges. These treatments bridge the gap between synchronous therapy sessions with asynchronous therapeutic engagement and digital content [15]. Preliminary evidence for blended care therapy models suggests that they can produce comparable or significantly better clinical outcomes in fewer live therapy sessions, relative to face-to-face psychotherapy [16,17]. Blended care therapy is an umbrella term that encompasses many treatment approaches. These approaches vary widely in the extent of asynchronous provider engagement between synchronous sessions, as well as the degree to which digital content is fully integrated into care [15]. This paper specifically examines a model of blended care therapy called "Lyra Care Therapy" (LCT). LCT blends synchronous video therapy sessions with intensive asynchronous *guided practice sessions*. In guided practice sessions, clients learn new evidence-based therapy concepts (beyond what is learned in synchronous sessions), receive messaging-based support from their providers, and apply therapy skills in their daily lives. Prior research on the LCT program has shown that this type of care is highly effective, with 89% of clients achieving either reliable improvement or recovery on validated outcome measures, and 74% achieving both reliable improvement and recovery [18]. This is substantially higher than the treatment response rates for depression found in standard, face-to-face mental health care settings. For example, a recent meta-analysis estimated that depression treatment response rates in "usual care" mental health treatment are approximately 20% [19].

LCT guided practice sessions contain multiple components, including digital video lessons, digital exercises, client message exchanges with their providers, and direct provider feedback to clients on their exercises. During synchronous sessions, providers set clients up for success by collaborating to select relevant and clinically appropriate digital activities to be completed. Clients then complete digital video lessons and exercises asynchronously during the guided practice session. Digital video lessons reinforce therapeutic skills taught in synchronous sessions or teach new concepts. Digital exercises then support clients in applying their newly learned skills in daily life, which is essential to facilitate symptom reductions [20].

Providers play a key role during LCT guided practice sessions by viewing clients' progress on digital activities in the secure platform, sharing prompt feedback messages to reinforce clients' efforts, and providing support through direct client-provider messaging. At their core, guided practice sessions offer a more engaging and effective alternative to traditional "homework" assignments in psychotherapy by facilitating asynchronous provider support and providing a more enriching digital experience for clients. Information from clients' digital engagement also complements providers' impressions from synchronous sessions to inform case conceptualization and care planning.

Past research evaluating this specific blended care model has demonstrated its effectiveness for reducing symptoms of anxiety and depression, including across diverse racial and ethnic groups and in large samples of up to approximately 6000 participants [18,21]. A prior LCT component analysis study also provided preliminary support for the role of client engagement with guided practice sessions in improving symptoms [22]. However, this prior component analysis focused exclusively on client engagement and did not assess the role of provider engagement. In LCT, the provider also plays a key role in supporting guided practice, by curating what is assigned, providing positive reinforcement, answering client questions, and offering corrective feedback when needed. Prior results from research on homework effectiveness in evidence-based psychotherapies suggest that these provider-level factors may also contribute to therapy effectiveness [11,23,24]. In traditional therapy models, however, provider support is limited to in-session interactions. In contrast, LCT providers additionally give feedback and encouragement to their clients between synchronous sessions via written messages in the platform. Thus, clients can readily access them and apply the feedback to their daily lives.

XSL•FO
RenderX

Although there is theoretical support for the importance of provider engagement in a blended care model, there is no known research evaluating the link between provider-specific digital engagement and clinical therapy outcomes. Therefore, a primary aim of this study was to explore the clinical impact of digital provider feedback in LCT, alongside client-focused engagement during guided practice sessions. We hypothesized that engagement with each element of the LCT program (ie, session attendance, client digital lesson completion, client digital exercise completion, and provider feedback) in the 7 days and 8-14 days prior to an outcome assessment would be associated with lower symptom severity at that assessment.

Methods

Study Design and Setting

This observational study used a retrospective cohort design and is reported in line with guidelines for STROBE (Strengthening the Reporting of Observational Studies in Epidemiology; see Multimedia Appendix 1 for STROBE checklist). Participants were adults in the United States who were eligible for an employer-sponsored mental health benefit, Lyra Health, offered by Lyra Clinical Associates. Individuals seeking care completed a brief, questionnaire-based assessment via a secure web-based platform to evaluate baseline symptoms and recommend appropriate care. All care activities were conducted via a proprietary, HIPAA (Health Insurance Portability and Accountability Act)–compliant platform that is accessible via web browser and mobile device.

Ethical Considerations

This deidentified data analysis was determined to be exempt by the WCG institutional review board (WCG IRB Tracking ID#20220388). Participants provided informed consent to take part in care and have their deidentified data used for research purposes as a part of that consent for care. Data were collected as part of routine practice in the LCT program to support measurement-based clinical care and for quality assurance. Participants did not receive compensation for engagement with the LCT program or for completing assessments.

Participants

Individuals were eligible for this study if they began care on or after January 1, 2022, and completed or dropped out of care by November 5, 2023. This time frame was selected to allow for accumulation of a large sample size over nearly 2 years that was sufficiently representative and diverse to provide adequate statistical power to achieve study objectives. Participants were included if they had a baseline score in the clinical range on the Generalized Anxiety Disorder 7-item scale (GAD-7; total score of ≥8) [25] or Patient Health Questionnaire 9-item scale (PHQ-9; total score ≥10) [26]. Individuals were excluded from the study if they did not have a baseline assessment that was completed ≤2 weeks before their first session and before their second session; or they did not complete at least 1 follow-up assessment either during care or a maximum of 5 weeks after the date of the final session. See the participant flow diagram in Figure 1 for more detail. Consistent with prior research, only clinical outcome assessments and sessions occurring within 1 SD of the average therapy episode duration (16.85 weeks) were included in this study.

XSL•FO
RenderX

Figure 1. Participant inclusion and exclusion. GAD-7: Generalized Anxiety Disorder—7-item scale; PHQ-9: Patient Health Questionnaire-9 item scale.

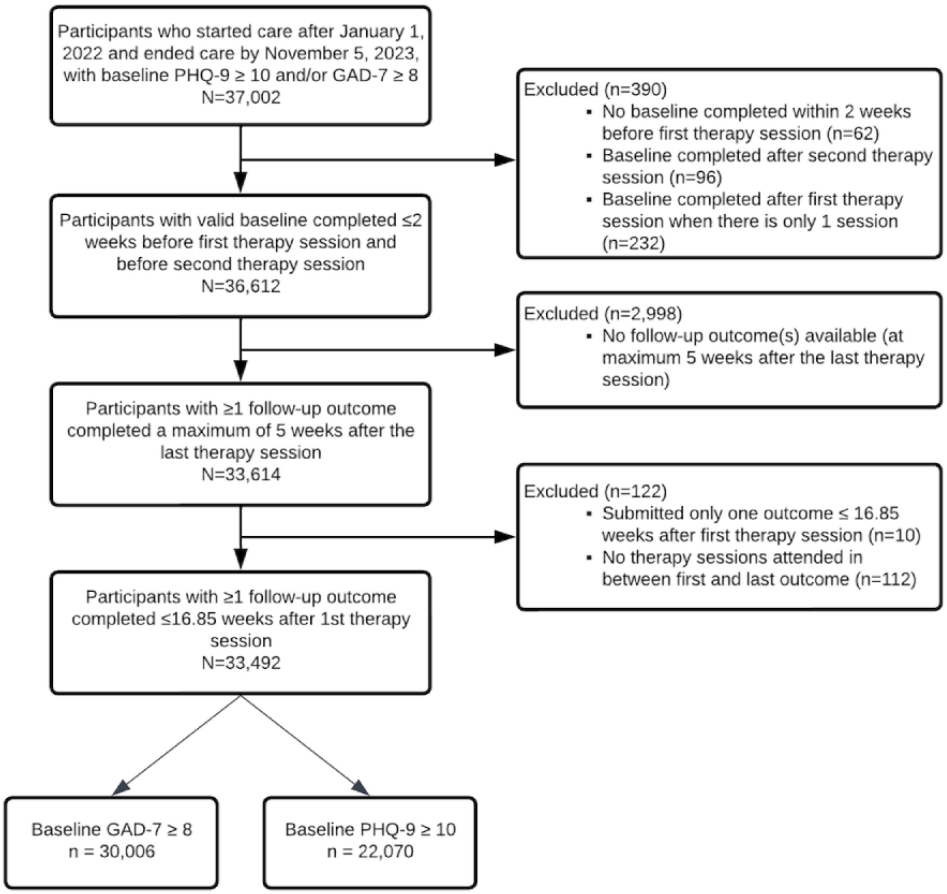

Clinical Program

Clinical Approach

A detailed description of LCT has been provided in prior research [18,21,22]. Briefly, the program consists of synchronous video therapy sessions combined with asynchronous guided practice sessions (ie, digital activities assigned by providers, therapist feedback, and direct client-provider messaging) via a secure, web-based digital care platform. The platform is accessible via web browser and mobile app. The program uses evidence-based clinical approaches with the highest-quality scientific evidence (eg, cognitive behavioral therapy, dialectical behavior therapy, and acceptance and commitment therapy) and is grounded in culturally responsive care principles to meet the social identity-related needs of a diverse client population [21,27-29].

Providers and Clinical Quality Supervision

Therapy is provided by licensed mental health professionals (eg, psychologists, marriage and family therapists, social workers, and professional counselors). Provider hiring is highly selective, emphasizing skillful use of evidence-based clinical practice and culturally responsive care. All providers receive >60 hours of intensive internal training upon hire, ongoing individual supervision, regular group consultation with a licensed clinical manager, and access to internal clinical consultation and training opportunities.

Guided Practice Sessions

Guided practice sessions refer to all interactions among the client, their provider, and the digital platform between synchronous sessions (see panels A and B in Figure 2 for illustrative examples of digital content). Two core elements of the guided practice sessions are digital video lessons and digital exercises. Lessons and exercises are based on evidence-based, transdiagnostic treatment approaches, including the Unified Protocol, dialectical behavior therapy, acceptance and commitment therapy, and other treatments rooted in cognitive behavioral principles [27-29]. Digital video lessons teach evidence-based therapeutic skills and concepts through a unique, narrative storytelling approach. Each lesson consists of 1-2 videos and concludes with a multiple-choice knowledge review to check clients' comprehension. Digital exercises provide opportunities for clients to apply concepts and practice skills learned in therapy or in digital lessons and reflect on how these concepts apply to their daily lives. Providers have access to a large library of digital lessons and exercises for their clients, which apply to a range of presenting concerns. Providers collaborate with clients to select relevant and clinically appropriate digital activities for each guided practice session. Providers can also tailor activity instructions for each client. This process allows for personalization of guided practice sessions based on a client's unique needs. Prior research has shown that use of these digital tools in LCT is associated with greater reductions in symptoms of anxiety and depression [22,30].

XSL•FO
RenderX

As clients complete their digital lessons and exercises, providers can view client progress and send asynchronous feedback messages. This provider feedback is intended to reinforce client learning and encourage consistent engagement for maximal clinical benefit [5,11]. The platform also allows for direct message exchanges between clients and providers. Through direct messages, clients can ask questions about their digital activities and ask their providers for help to troubleshoot challenges encountered when practicing skills. Providers can send check-in messages, answer questions, provide encouragement, and address any logistical or scheduling needs between sessions.

Finally, all clients receive validated clinical outcome assessments throughout care to monitor treatment progress. Both providers and clients are able to view clients' responses, assessment scores, and trends over time, which supports measurement-based care.

Figure 2. Illustrative examples of the client digital interface. (A) Client home page in mobile app (client name and provider's name and image are fictional). (B) Client view of the "Understanding emotions" digital lesson.

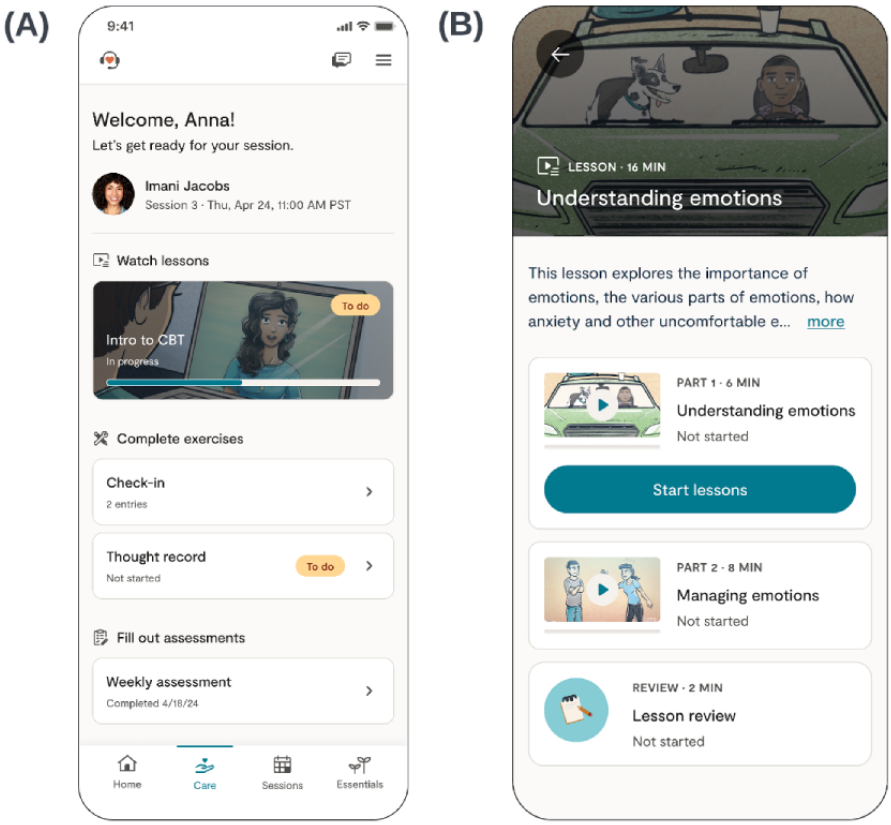

Measures

Demographic Characteristics, Session Attendance, and Treatment Duration

Clients may optionally self-report race, ethnicity, and gender identity during the intake process. Participant age was documented as age in years at the time of the first synchronous video session. Session attendance was automatically tracked on the platform. Treatment duration was computed as the time in weeks between the first session and the last session.

Digital Engagement During Guided Practice Sessions

Engagement with each element of guided practice sessions (client digital lesson completion, client digital exercise completion, provider feedback messages, and client-provider direct messages) was recorded automatically in the platform.

Clinical Outcome Measures

For quality assurance and clinical assessment to support measurement-based care, all clients in the program received the PHQ-9 [26] and GAD-7 [31] on a weekly basis and after the final session. For the purposes of this manuscript, client responses to these assessments are referred to as clinical outcome assessments. Reliable improvement was defined as a reduction ≥6 on the PHQ-9 and ≥4 on the GAD-7 [32,33]. Recovery was defined by a final score <10 for the PHQ-9 and <8 for the GAD-7 [25,26]. Reliable improvement and recovery were evaluated only for the measures on which the baseline score was in the clinical range for a given client. If a client scored in the clinical range on both the GAD-7 and the PHQ-9, they were included in analyses for both measures (including growth curve modeling). Growth curve models used all available clinical outcome assessments for each participant, including but not limited to the first and last measurements completed within the data collection period (see "Participants" section for more information).

Statistical Analysis

All statistical analyses were performed in Python (version 3.10.9; Python Software Foundation) and R 4.2.3 (R Core Team)

XSL•FO
RenderX

[34]. The statistical packages used for analyses are described in the remainder of the "Statistical Analysis" section where applicable.

Overall Treatment Outcome

Observed change in PHQ-9 and GAD-7 scores across the LCT episode was examined with 2-tailed dependent samples *t* tests, examining the difference between each individual's baseline and last available clinical outcome assessments (a=.05). The SciPy package (version 1.10.0) was used for analysis [35].

Symptom Trajectories and Time-Varying Effects of Treatment Engagement

Symptom change throughout the LCT episode was evaluated using linear mixed-effects models to fit individual growth curves for the PHQ-9 and GAD-7, respectively, using the lme4 package (version 1.1.35) and full-information restricted maximum likelihood estimation [36]. Data explorations were conducted to evaluate whether data missingness patterns for the clinical outcome assessments of interest were consistent with the missing at random assumption and are summarized in Multimedia Appendix 2. Results were consistent with those expected if the missing at random assumption was met.

For growth curve models, model 1 included linear and quadratic fixed effects for the continuous time variable (weeks). Participant-level random effects were estimated for the intercept, linear, and quadratic time components. In model 2, binary indicator variables (ie, absent=0, present=1) were used to model the time-varying effects of engagement with each treatment component (therapy sessions, client digital lesson completion, client digital exercise completion, and provider feedback messages) during the past 7 days. Direct messages were not included as a time-varying covariate (TVC) effect in growth curve models, because not all direct messages were expected to have therapeutic content and therefore their effects on clinical outcomes would be heterogeneous. In model 3, additional TVCs were added for treatment engagement over the prior 8-14 days. This modeling approach allowed for testing of whether engagement with each program element was uniquely associated with greater symptom reduction in the subsequent 1-2 weeks.

Supplemental analyses were conducted using count variables for the treatment engagement TVCs, in place of the binary covariates used in the primary analysis. This would allow for testing of whether engagement with treatment components was associated with clinical outcomes in a linear manner, such that each additional engagement with the same treatment element in a given 1-2 weeks was associated with a fixed incremental benefit for symptom reduction. This approach is consistent with prior evaluations of this program but is limited in that it treats every digital element as an equal unit (ie, assumes a linear dose-response relationship). Findings from these analyses are reported in Multimedia Appendix 3.

An additional sensitivity analysis was conducted for the binary covariate models with participant demographics (age, gender identity, and race and ethnicity) added as fixed effects. Significant effects emerged for these covariates; however, addition of these variables provided only marginal improvement in model fit and did not change the pattern of results for the primary outcomes of interest. Results are reported in Multimedia Appendix 3.

Results

Participant Characteristics

A total of 33,492 clients were included in the sample (see Figure 1 for participant flow). Participants self-identified as Asian or Pacific Islander (5401/33,492, 16.1%), Black or African American (3026/33,492, 9.0%), Hispanic or Latino (3709/33,492, 11.1%), White (17,347/33,492, 51.8%), multiple identities (2687/33,492, 8.0%), or another identity (683/33,492, 2.0%). Participants self-reported their gender identities as female (21,217/33,492, 63.4%), male (10,547/33,492, 31.5%), nonbinary (295/33,492, 0.9%), and other identities (262/33,492, 0.8%). Nearly 9 in 10 participants (30,006/33,492, 89.5%) screened positive for anxiety on the GAD-7 [25] (score ≥8 at baseline), and nearly two-thirds (22,070/33,492, 65.9%) screened positive for depression on the PHQ-9 [26] (baseline score ≥10). See Table 1 for additional details on participant characteristics.

Participants who were excluded due to insufficient clinical outcome assessment data were compared with included participants on baseline demographics, baseline clinical severity, and session count. As expected, given the large sample size, statistically significant differences were found, but effect sizes were very small for demographic and clinical variables (Cohen *d*=0.06-0.07, Cramer *V*=.02-.05). Notably, the median session count for the excluded group was 1 (IQR 1-2), which is much smaller than that for included participants (median 6, IQR 4-8). This indicates that most excluded clients did not continue therapy after attending the intake session (see Multimedia Appendix 4 for full results).

XSL•FO
RenderX

Table 1. Participant characteristics.

	Entire sample (N=33,492)	Anxiety sample (baseline GAD-7[a] ≥8) (N=30,006)	Depression sample (baseline PHQ-9[b] ≥10) (N=22,070)
Age (years), mean (SD)	33.67 (9.51)	33.60 (9.44)	33.59 (9.65)
Gender, n (%)			
Female	21,217 (63.35)	19,061 (63.52)	13,945 (63.19)
Male	10,547 (31.49)	9385 (31.28)	6929 (31.40)
Nonbinary	295 (0.88)	264 (0.88)	236 (1.07)
Other	262 (0.78)	229 (0.76)	191 (0.87)
Missing/unknown	1171 (3.50)	1067 (3.56)	769 (3.48)
Race and ethnicity, n (%)			
Asian or Pacific Islander	5401 (16.13)	4867 (16.22)	3417 (15.48)
Black or African American	3026 (9.03)	2704 (9.01)	2169 (9.83)
Hispanic or Latino	3709 (11.07)	3335 (11.11)	2638 (11.95)
Multiple	2687 (8.02)	2392 (7.97)	1837 (8.32)
Other	683 (2.04)	633 (2.11)	452 (2.05)
White	17,347 (51.79)	15,506 (51.68)	11,131 (50.43)
Prefer not to disclose/missing	639 (1.91)	569 (1.90)	426 (1.93)
Baseline GAD-7 severity, n (%)			
Minimal (0-4)	918 (2.74)	N/A[c]	918 (4.16)
Mild/below clinical cutoff (5-7)[d]	2568 (7.67)	N/A[c]	2568 (11.64)
Mild/above clinical cutoff (8-9)[d]	6296 (18.8)	6296 (20.98)	2531 (11.47)
Moderate (10-14)	13,298 (39.71)	13,298 (44.32)	7684 (34.82)
Severe (15-21)	10,412 (31.09)	10,412 (34.70)	8369 (37.92)
Baseline PHQ-9 severity, n (%)			
Minimal (0-4)	2748 (8.20)	2748 (9.16)	N/A
Mild (5-9)	8674 (25.90)	8674 (28.91)	N/A
Moderate (10-14)	11,977 (35.76)	9430 (31.43)	11,977 (54.27)
Moderately severe (15-19)	6996 (20.89)	6196 (20.65)	6996 (31.70)
Severe (20-27)	3097 (9.25)	2958 (9.86)	3097 (14.03)
Baseline GAD-7, mean (SD)	12.28 (4.26)	13.08 (3.70)	12.67 (4.71)
Baseline PHQ-9, mean (SD)	11.85 (5.35)	11.71 (5.54)	14.80 (3.91)
Final GAD-7, mean (SD)	6.04 (4.46)	6.25 (4.51)	6.44 (4.69)
Final PHQ-9, mean (SD)	5.82 (4.93)	5.83 (4.98)	6.89 (5.21)
Treatment duration in weeks, median (IQR)	8.00 (4.00-12.71)	8.00 (4.00-12.57)	7.86 (3.86-12.43)

[a]GAD-7: Generalized Anxiety Disorder 7-item scale.

[b]PHQ-9: Patient Health Questionnaire 9-item scale.

[c]N/A: not applicable; no participants scored in this range within the participant subsample.

[d]To be included in the anxiety sample, clients were required to have a GAD-7 score above the clinical cutoff of 8 or higher. The mild severity category was therefore split to distinguish between those above and below the clinical cutoff.

Treatment Characteristics and Overall Clinical Outcome

Participants attended a median of 6.0 live synchronous therapy sessions (IQR 4.0-8.0) over the course of a median of 8.0 (IQR 4.0-12.7) weeks. Participant engagement with digital tools during guided practice sessions was also high. Over the course of a therapy episode, participants viewed a median of 6.0 digital lessons (IQR 3.0-9.0), completed 6.0 exercises (IQR 2.0-12.0), and exchanged 16.0 direct messages (IQR 9.0-28.0) with their

XSL•FO
RenderX

providers during guided practice sessions. In addition, participants received a median of 2.0 (IQR 0-4.0) feedback messages from their providers on their completed exercises. Descriptive statistics for engagement with all treatment components are detailed in Table 2.

Differences in symptom severity from the first to last clinical outcome assessment were evaluated via dependent samples *t* tests. Results indicated that participants experienced statistically significant reductions in symptoms of anxiety (n=30,006, $t_{30,005}$=236.98; *P*<.001) and depression (n=22,070, $t_{22,069}$=213.56; *P*<.001), as measured by the GAD-7 and PHQ-9, respectively, with very large observed effect sizes (Cohen *d*=1.37 for GAD-7; Cohen *d*=1.44 for PHQ-9). Furthermore, among the full sample, 86.6% (29,012/33,492) of participants experienced clinical improvement (defined as experiencing reliable improvement or recovery on the GAD-7, PHQ-9, or both) from their first to last assessment (Table 3).

Table 2. Session attendance and engagement with guided practice session elements[a].

Treatment component	Full sample (N=33,492)	Anxiety sample (baseline GAD-7[b] ≥8) (N=30,006)	Depression sample (baseline PHQ-9[c] ≥10) (N=22,070)
Synchronous therapy sessions			
Sessions attended	6.0 (4.0-8.0)	6.0 (4.0-8.0)	6.0 (4.0-8.0)
Guided practice sessions: provider engagement			
Lessons assigned	8.0 (6.0-11.0)	8.0 (6.0-11.0)	8.0 (6.0-12.0)
Exercises assigned	7.0 (4.0-12.0)	7.0 (4.0-12.0)	8.0 (4.0-12.0)
Exercise feedback messages sent	2.0 (0.0-4.0)	1.0 (0.0-4.0)	2.0 (0.0-4.0)
Direct messages sent	10.0 (6.0-16.0)	10.0 (6.0-16.0)	10.0 (6.0-16.0)
Guided practice sessions: client engagement			
Lessons completed	6.0 (3.0-9.0)	6.0 (3.0-9.0)	6.0 (3.0-9.0)
Exercises completed	6.0 (2.0-12.0)	6.0 (2.0-12.0)	6.0 (2.0-12.0)
Direct messages sent	6.0 (3.0-12.0)	6.0 (3.0-12.0)	6.0 (3.0-12.0)
Clinical outcome assessments completed (GAD-7 or PHQ-9)[d]	12.0 (8.0-18.0)	6.0 (4.0-9.0)	6.0 (4.0-9.0)

[a]All descriptive statistics are reported as median (IQR) due to nonnormal distributions of variables within the sample. Engagement is totaled across a full episode of care.

[b]GAD-7: Generalized Anxiety Disorder 7-item scale.

[c]PHQ-9: Patient Health Questionnaire 9-item scale.

[d]For the full sample, each unique clinical outcome assessment completed was counted. GAD-7 and PHQ-9 measures were counted separately. For the anxiety and depression subsamples, only the relevant clinical outcome measures were included in assessment counts (ie, GAD-7 for anxiety subsample, PHQ-9 for depression subsample).

Table 3. Rates of reliable improvement and recovery[a].

Baseline symptoms	n	Reliable improvement, n (%)	Recovery, n (%)	Reliable improvement and recovery, n (%)	Reliable improvement or recovery, n (%)
Anxiety symptoms (GAD-7[b] ≥8)	30,006	22,807 (76.01)	20,748 (69.15)	18,955 (63.17)	24,600 (81.98)
Depression symptoms (PHQ-9[c] ≥10)	22,070	15,122 (68.52)	16,208 (73.44)	13,644 (61.82)	17,686 (80.14)
Anxiety and depression symptoms (GAD-7 ≥8 and PHQ-9 ≥10)	18,584	11,411 (61.40)	10,845 (58.36)	13,036 (70.15)	16,304 (87.73)
Anxiety or depression symptoms (full sample)	33,492	26,672 (79.64)	26,111 (77.96)	23,255 (69.43)	29,012 (86.62)

[a]Reliable improvement was defined by meeting at least one of the following criteria: (1) ≥4-point decrease on the final GAD-7 among those with baseline GAD-7 score of ≥8 and (2) ≥6-point decrease on the final PHQ-9 among those with baseline PHQ-9 score of ≥10. Recovery was defined by meeting at least one of the following criteria: (1) final GAD-7 score of <8 among those with baseline GAD-7 score of ≥8 and (2) final PHQ-9 score of <10 among those with baseline score of ≥10.

[b]GAD-7: Generalized Anxiety Disorder 7-item scale.

[c]PHQ-9: Patient Health Questionnaire 9-item scale.

XSL•FO
RenderX

Anxiety Symptom Trajectories and Time-Varying Effects of Engagement

Results from individual growth curve modeling for anxiety symptom trajectories and the time-varying effects of engagement with each program element are summarized in this section. Full modeling results are shown in Table 4. Individuals with initial GAD-7 scores in the clinical range were included in the analysis (n=30,006).

Coefficients from an initial model examined linear and quadratic fixed effects of time in weeks on GAD-7 scores (model 1). On average, participants exhibited an initial decline in anxiety symptoms of more than 1 unit per week (*b*=–1.23, 95% CI –1.24 to –1.22) that attenuated over time (*b*=0.06, 95% CI 0.06-0.06). Model 2 incorporated fixed effects for each TVC indicating the presence (or absence) of a given program element during the 7 days prior to the clinical outcome assessment. The occurrence of 1 or more synchronous video sessions was significantly associated with a –0.70 unit decrease in anxiety scores (*b*=–0.70, 95% CI –0.73 to –0.67). The coefficients for TVCs identifying client completion of 1 or more digital lessons (*b*=–0.20, 95% CI –0.24 to –0.17), client completion of 1 or more digital exercises (*b*=–0.13, 95% CI –0.17 to –0.09), and 1 or more provider feedback messages (*b*=–0.21, 95% CI –0.26 to –0.17) also indicated that these forms of client and provider engagement were uniquely and significantly associated with lower GAD-7 scores. A similar pattern emerged in model 3, which incorporated TVC effects during the 8-14 days prior to completion of the clinical outcome assessment. As expected, synchronous video session occurrence (*b*=–0.58, 95% CI –0.61 to –0.54), client digital lesson completion (*b*=–0.26, 95% CI –0.30 to –0.22), client digital exercise completion (*b*=–0.10, 95% CI –0.14 to –0.06), and the presence of provider feedback messaging (*b*=–0.07, 95% CI –0.12 to –0.02) were each significantly associated with lower GAD-7 scores. The pattern of findings for the 7-day coefficients was virtually unchanged from model 2, with the exception of the coefficient for exercise completion in the 7 days prior to a clinical outcome assessment. In model 2, the coefficient was larger, negative, and statistically significant, whereas in model 3, it was not statistically different from 0. Likelihood ratio tests across these models (all *P* values of <.01), as well as information criteria indices (Akaike information criterion [AIC] and Bayesian information criterion [BIC]), suggested that model 3 provided the best fit to the observed data. Panels A and B in Figure 3 depict the possible combinations of digital engagement and their expected impact on clinical symptoms in the 2 weeks leading up to the clinical outcome assessment, based on the coefficients from model 3.

Table 4. Anxiety symptom trajectories and time-varying effects of engagement with elements of blended care therapy[a].

	Model 1		Model 2		Model 3	
	b (95% CI)	*P* value	*b* (95% CI)	*P* value	*b* (95% CI)	*P* value
Intercept	11.50 (11.45 to 11.54)	<.001	11.92 (11.87 to 11.97)	<.001	12.10 (12.05 to 12.15)	<.001
Weeks	–1.23 (–1.24 to –1.22)	<.001	–1.17 (–1.18 to –1.15)	<.001	–1.05 (–1.06 to –1.03)	<.001
Weeks²	0.06 (0.06 to 0.06)	<.001	0.05 (0.05 to 0.05)	<.001	0.04 (0.04 to 0.05)	<.001
Therapy sessions last 7 days	N/A[b]	N/A	–0.70 (–0.73 to –0.67)	<.001	–0.82 (–0.85 to –0.78)	<.001
Digital lessons last 7 days	N/A	N/A	–0.20 (–0.24 to –0.17)	<.001	–0.18 (–0.22 to –0.15)	<.001
Digital exercises last 7 days	N/A	N/A	–0.13 (–0.17 to –0.09)	<.001	0.00 (–0.04 to 0.04)	.89
Provider feedback last 7 days	N/A	N/A	–0.21 (–0.26 to –0.17)	<.001	–0.12 (–0.16 to –0.08)	<.001
Therapy sessions 8-14 days	N/A	N/A	N/A	N/A	–0.58 (–0.61 to –0.54)	<.001
Digital lessons 8-14 days	N/A	N/A	N/A	N/A	–0.26 (–0.30 to –0.22)	<.001
Digital exercises 8-14 days	N/A	N/A	N/A	N/A	–0.10 (–0.14 to –0.06)	<.001
Provider feedback 8-14 days	N/A	N/A	N/A	N/A	–0.07 (–0.12 to –0.02)	.004
Deviance (–2LL[c])	1,059,750	N/A	1,056,354	N/A	1,054,123	N/A
AIC[d]	1,059,772	N/A	1,056,384	N/A	1,054,161	N/A
BIC[e]	1,059,884	N/A	1,056,536	N/A	1,054,355	N/A

[a]The outcome of interest was client scores on the Generalized Anxiety Disorder 7-item scale (GAD-7). The analyzed sample included 30,006 individuals with a baseline GAD-7 score ≥8. $df_{residual}$=196,023 for model 1, $df_{residual}$=196,019 for model 2, and $df_{residual}$=196,015 for model 3.

[b]Not applicable.

[c]–2LL: –2 × log likelihood.

[d]AIC: Akaike information criterion.

[e]BIC: Bayesian information criterion.

XSL•FO
RenderX

Figure 3. Estimated GAD-7 score reductions associated with different combinations of client and provider digital engagement in the (A) 0-7 days and (B) 8-14 days prior to a clinical outcome assessment. GAD-7: Generalized Anxiety Disorder-7 item scale.

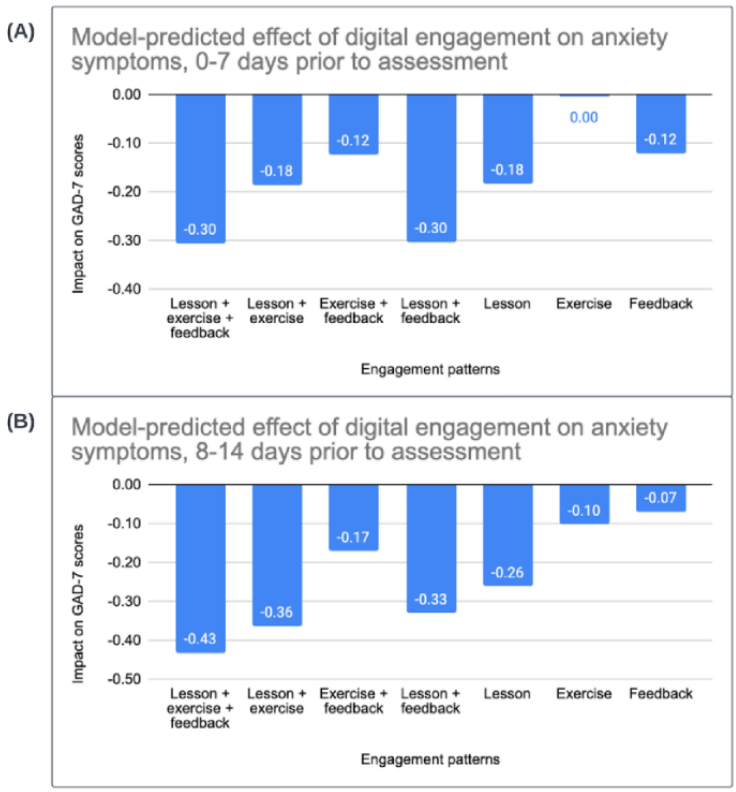

Depression Symptom Trajectories and Time-Varying Effects of Engagement

Results from individual growth curve modeling for depression symptom trajectories and time-varying effects of engagement with each program element are summarized in this section. Full modeling results are reported in Table 5. Participants with initial PHQ-9 scores in the clinical range were included in the analysis (n=22,070).

Coefficients from an initial model examined linear and quadratic fixed effects of time in weeks on PHQ-9 scores (model 1). The coefficients in model 1 were consistent with a steep initial decline in depression symptoms of more than 1 unit per week (*b*=–1.45, 95% CI –1.46 to –1.43) that attenuated over time (*b*=0.07, 95% CI 0.07-0.07). The TVC effects in model 2 indicated that the occurrence of 1 or more synchronous video sessions during the previous week was associated with a –0.76 unit decrease in PHQ-9 scores (*b*=–0.76, 95% CI –0.80 to –0.72). Similarly, the past-week coefficients for client completion of 1 or more digital lessons (*b*=–0.14, 95% CI –0.19 to –0.10), client completion of 1 or more digital exercises (*b*=–0.31, 95% CI –0.36 to –0.26), and the occurrence of 1 or more provider feedback messages (*b*=–0.25, 95% CI –0.30 to –0.19) indicated that these forms of client and provider engagement were uniquely and significantly associated with lower PHQ-9 scores. In model 3, the coefficients for the occurrence of 1 or more synchronous video sessions (*b*=–0.67, 95% CI –0.71 to –0.62) 8-14 days prior to the clinical outcome assessment, as well as client completion of 1 or more digital lessons (*b*=–0.30, 95% CI –0.34 to –0.25), and the presence of feedback messaging from providers (*b*=–0.08, 95% CI –0.14 to –0.02), suggest that these forms of engagement are uniquely and significantly associated with lower PHQ-9 scores. In contrast, the coefficient for client digital exercise completion in the past 8-14 days was not significant (*b*=–0.05, 95% CI –0.10 to 0.01). The 7-day coefficients for therapy sessions, client lesson completion, and provider feedback messages were very similar across models 2 and 3. The coefficient for client exercise completion was numerically smaller relative to model 2 but still statistically significant (*b*=–0.15, 95% CI –0.21 to –0.10). Likelihood ratio tests across these models (all *P* values of <.01), as well as information criteria indices (AIC and BIC), suggested that model 3 provided the best fit to the observed data. Panels A and B in Figure 4 depict the possible combinations of digital engagement and their expected impact on clinical symptoms in the 2 weeks leading up to the clinical outcome assessment, based on the coefficients from model 3.

XSL•FO
RenderX

Table 5. Depression symptom trajectories and time-varying effects of engagement with elements of blended care therapy[a].

	Model 1		Model 2		Model 3	
	b (95% CI)	*P* value	*b* (95% CI)	*P* value	*b* (95% CI)	*P* value
Intercept	12.90 (12.84 to 12.96)	<.001	13.39 (13.33 to 13.46)	<.001	13.59 (13.53 to 13.66)	<.001
Weeks	–1.45 (–1.46 to –1.43)	<.001	–1.36 (–1.38 to –1.34)	<.001	–1.23 (–1.25 to –1.21)	<.001
Weeks²	0.07 (0.07 to 0.07)	<.001	0.06 (0.06 to 0.06)	<.001	0.05 (0.05 to 0.05)	<.001
Therapy sessions last 7 days	N/A[b]	N/A	–0.76 (–0.80 to –0.72)	<.001	–0.89 (–0.93 to –0.85)	<.001
Digital lessons last 7 days	N/A	N/A	–0.14 (–0.19 to –0.10)	<.001	–0.12 (–0.16 to –0.08)	<.001
Digital exercises last 7 days	N/A	N/A	–0.31 (–0.36 to –0.26)	<.001	–0.16 (–0.21 to –0.11)	<.001
Provider feedback last 7 days	N/A	N/A	–0.25 (–0.30 to –0.19)	<.001	–0.15 (–0.21 to –0.10)	<.001
Therapy sessions 8-14 days	N/A	N/A	N/A	N/A	–0.67 (–0.71 to –0.62)	<.001
Digital lessons 8-14 days	N/A	N/A	N/A	N/A	–0.30 (–0.34 to –0.25)	<.001
Digital exercises 8-14 days	N/A	N/A	N/A	N/A	–0.05 (–0.10 to 0.01)	.09
Provider feedback 8-14 days	N/A	N/A	N/A	N/A	–0.08 (–0.14 to –0.02)	.01
Deviance (–2LL[c])	808,568.9	N/A	805,755.1	N/A	804,048.6	N/A
AIC[d]	808,590.9	N/A	805,785.1	N/A	804,086.6	N/A
BIC[e]	808,699.6	N/A	805,933.3	N/A	804,274.3	N/A

[a]The outcome of interest was client scores on the Patient Health Questionnaire 9-item scale (PHQ-9). The analyzed sample included 22,070 individuals with a baseline PHQ-9 score of ≥10. $df_{residual}$=144,505 for model 1, $df_{residual}$=144,501 for model 2, and $df_{residual}$=144,497 for model 3.

[b]Not applicable.

[c]–2LL: –2 × log likelihood.

[d]AIC: Akaike information criterion.

[e]BIC: Bayesian information criterion.

Figure 4. Estimated PHQ-9 score reductions associated with different combinations of client and provider digital engagement in the (A) 0-7 days and (B) 8-14 days prior to a clinical outcome assessment. PHQ-9: Patient Health Questionnaire—9 item scale.

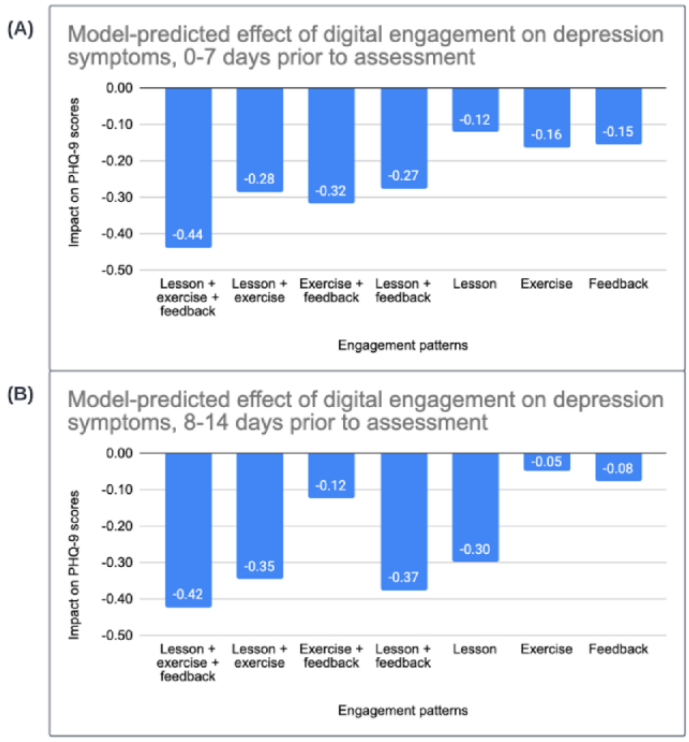

XSL•FO
RenderX

Discussion

Principal Findings

This study evaluated a unique blended care therapy program for anxiety and depression, consisting of synchronous video sessions and asynchronous guided practice sessions. Guided practice sessions are designed to provide more integrated and engaging opportunities for new client learning and skills practice outside of synchronous sessions. Providers encourage clients' practice through asynchronous feedback and messaging. This study's analyses evaluated the associations between guided practice session engagement and symptom reduction, with a novel focus on the time-varying effects of asynchronous provider feedback on client symptom outcomes.

Results supported the unique contributions of provider feedback messages to symptom reductions during care. Specifically, the presence of provider feedback messaging in the past 7 days and 8-14 days, respectively, was related to greater reductions in anxiety (b_{7days}=–0.12, $b_{8\text{-}14days}$=–0.07) and depression (b_{7days}=–0.15, $b_{8\text{-}14days}$=–0.08), even when accounting for effects of client-completed digital lessons, client-completed digital exercises, and synchronous video sessions. Results also provided further support for the clinical effects of client-level engagement with guided practice session elements (ie, viewing digital lessons and completing digital exercises), beyond the effects of synchronous video session attendance. More specifically, completing 1 or more digital lessons in the past 7 days and 8-14 days, respectively, was associated with significantly greater reductions in symptoms of anxiety (b_{7days}=–0.18, $b_{8\text{-}14days}$=–0.26) and depression (b_{7days}=–0.12, $b_{8\text{-}14days}$=–0.30). Completing an exercise in the past 7 days was associated with significantly lower depression symptoms (b_{7days}=–0.16), and completing an exercise in the past 8-14 days was associated with significantly lower anxiety symptoms ($b_{8\text{-}14days}$=–0.10). Finally, attending 1 or more synchronous sessions in the past 7 and 8-14 days, respectively, was also associated with significantly lower anxiety (b_{7days}=–0.82, $b_{8\text{-}14\ days}$=–0.58) and depression (b_{7days}=–0.89, $b_{8\text{-}14\ days}$=–0.67). If all types of guided practice session engagement were present over the course of 2 weeks, the model estimates suggest an expected total symptom reduction of 0.73 points on the GAD-7 and 0.86 points on the PHQ-9 during that time. Considering that the average baseline score on each measure was approximately 13 at baseline for the GAD-7 and 15 for the PHQ-9, the level of engagement observed in our sample would be expected to have a clinically meaningful impact over the course of care. Altogether, these findings indicate that each element of LCT provides distinct clinical benefit. In combination, these components can support more effective care overall.

It is interesting to note that digital exercise engagement was found to be slightly differentially associated with reductions in symptoms of anxiety versus depression. Specifically, engagement with digital exercises in the past 8-14 days was significantly associated with lower anxiety symptoms, but there was no significant relationship observed between exercise engagement in the past 7 days and anxiety outcomes. A converse pattern was observed for depression symptoms. Engagement with digital exercises in the past 7 days was significantly associated with lower depression symptoms, but no significant effect of engagement in the past 8-14 days was observed. Together, these results suggest that the impact of exercise engagement on depression symptoms may be more immediate, and the impact of exercises on anxiety symptoms may be slightly delayed. Clinically, this pattern is logical. Among other intervention strategies, evidence-based treatment for depression typically involves behavioral activation, which is intended to provide immediate positive reinforcement to clients and can support improved mood [4,37]. Thus, any exercises that may increase behavioral activation would be expected to have a more immediate effect on symptoms of depression. In contrast, evidence-based anxiety treatment typically involves reducing avoidance and gradually increasing exposure to previously feared or avoided situations [27]. When addressing anxiety symptoms, completion of exercises that involve increasing approach behaviors (and reducing avoidance) could initially correspond to consistent levels of distress before eventually leading to habituation and hence reduced symptoms [38]. Future research on the LCT program could explore how different types of digital exercises may differentially impact symptoms throughout care and how their effects may differ across presenting concerns.

This is the first known study to demonstrate that *both* provider and client digital engagement via asynchronous guided practice sessions contribute to symptom reduction, beyond the effects of synchronous video sessions. Moreover, each core element of the LCT model contributed uniquely and meaningfully to symptom reduction, including synchronous sessions and all guided practice session elements evaluated (digital lessons, digital exercises, and digital provider feedback messages). These effects also corresponded to strong overall clinical treatment outcomes. Approximately 87% (29,012/33,492) of clients achieved clinical improvement (ie, reliable improvement or recovery) over a median of 6 live synchronous sessions, compared with 22%-67% improvement rates over 8-15 sessions (or even longer) in traditional mental health care settings [19,39-41]. This particular model of blended care therapy therefore may support more efficient and effective care. In line with the LCT program's commitment to data-informed clinical practice and continuous quality improvement, these findings will also be used in the LCT program to inform clinical practice by providing further evidence to encourage client and provider use of all guided practice session elements as a means of supporting greater improvements in therapy.

One possible interpretation of these findings is that this blended care model extends opportunities for clients to engage with and benefit from the "active elements" of psychotherapy, outside of their synchronous video sessions [5]. In traditional therapy models, synchronous sessions typically last for only 1 hour every 1-2 weeks. These sessions are the only setting in which provider teaching of "active elements" and client understanding of this information can occur, and thus opportunities are limited. In this blended care model, synchronous sessions occur at a similar cadence, but providers also facilitate new learning and skills practice for clients during guided practice sessions. They

XSL•FO
RenderX

do this by first tailoring the selection of digital activities for each client and providing personalized instructions to help clients apply the skills to their unique situation. Providers then support client engagement between synchronous sessions through asynchronous feedback and messages. In this way, provider teaching and client understanding are extended well beyond the traditional bounds of the synchronous sessions.

Exploring the mechanisms for *why* provider feedback was associated with lower anxiety and depression symptoms is beyond the scope of this study. Some hypotheses can be made from prior research on the relationship between in-session provider homework-related behaviors and client symptom outcomes. Even when homework is assigned in traditional therapy settings, research has shown that providers do not consistently follow up on it in the next session [10,24]. Timely review and feedback from providers is critical because it provides clients with positive reinforcement and a sense of mastery for accomplishing the very hard work of behavior change [11]. In LCT, prompt provider feedback messages may give this reinforcement and signal to the client that they are supported when times are challenging. Thoughtful and personalized provider feedback messages may also remind the client of the link between the individual assignment completed and the client's broader goals for therapy, potentially enhancing motivation to continue practicing and improving. Finally, provider feedback may prompt a client to revisit the exercise or give the client a new perspective on their experience. In turn, this could facilitate consolidation of learning and ultimately support longer-lasting behavior change. In these ways, well-designed digital platforms and tools—particularly those that support provider-client interactions between therapy sessions—may offer effective solutions for key barriers to homework engagement in psychotherapy [42]. Future research evaluating the LCT program could explore whether quantitative evidence supports these potential mechanisms, as well as how the content and depth of messaging affects future client engagement and outcomes.

Results demonstrated a high level of engagement with key guided practice session elements. In turn, this provider and client digital engagement was linked to better anxiety and depression outcomes. This clear link between engagement and clinical outcomes is often overlooked in digital mental health research [43,44]. Demonstrating correspondence between engagement and clinical outcomes is imperative when evaluating blended care therapy models and digital mental health intervention components, especially given that digital engagement is typically lower than intended or desired for these interventions [44,45]. By bolstering the therapeutic impact that occurs between synchronous therapy sessions, providers can serve clients more effectively and efficiently (ie, in fewer synchronous sessions but with stronger clinical outcomes). In turn, more clients can be served and benefit from high-quality mental health care.

Strengths and Limitations

This study is strengthened by the inclusion of a very large participant sample of more than 30,000 individuals, a sample size that is uncommon in psychotherapy research. The design of the digital platform provided the opportunity for analysis of highly detailed data on client engagement and clinical symptom outcomes throughout care. To date, there are no other known studies with this level of detail on treatment engagement and clinical outcomes in such a large clinical sample. The sample was also diverse in terms of racial and ethnic identity. The study applied a rigorous statistical analysis approach, using objectively measured platform and session engagement data for both clients and providers. The results also represent clinical outcomes from real-world implementation of this program throughout the United States, and this enhances external validity and generalizability.

The study also has several limitations. Results from the growth curve modeling analyses provide information on time-varying associations among provider and client engagement variables and clinical outcomes; thus, they do not allow for causal conclusions. In addition, due to the high sensitivity of the data involved, this study did not evaluate the effects of client-provider direct messaging, aside from provider feedback messages that were focused specifically on clients' digital exercise completion. It is possible that other direct message exchanges during guided practice sessions could also contribute to clinical outcomes. However, the direct messaging category is heterogeneous, because it includes clinically focused messages as well as those that are administrative in nature (eg, focused on scheduling or logistics). Effects of client-provider direct messaging may be studied in the future if the content of these messages could be further classified without identifying participants, such that analyses could differentiate between message types.

Conclusions

In sum, this study yielded strong evidence that this blended care therapy program provided effective and efficient mental health care for clients at a large scale, and that provider feedback during guided practice sessions contributed to these clinical outcomes. This program leverages innovative digital tools to enhance therapeutic engagement for both providers and clients between synchronous sessions, providing substantial clinical benefit for clients.

Acknowledgments

The authors would like to express their appreciation for the Lyra providers who work every day to support clients in leading more fulfilling, productive lives. The authors would also like to acknowledge Katie Shakman, PhD, and Jennifer Chen, MS, for their assistance with code reviews during the analysis process.

XSL•FO
RenderX

The study was funded by Lyra Health. The funder of the study was involved in study design, data interpretation, and writing of the report, and informed during data collection and data analysis. The corresponding author had final responsibility for the decision to submit for publication.

The authors declare that there was no use of generative artificial intelligence technology in the generation of texts, figures, or other informational content in this paper.

Data Availability

The data sets generated and/or analyzed during this study are not publicly available due to the fact that the data are related to the delivery of health care and subject to HIPAA (Health Insurance Portability and Accountability Act of 1996).

As a business associate under HIPAA, Lyra is permitted to use patient information only as outlined in our Business Associate Agreements with our customers, which are covered entities. Many of our customer agreements do not permit us to disclose deidentified patient data; given this, we are unable to share the deidentified data due to contractual restrictions.

Authors' Contributions

HE led the study, including specifying study objectives, final selection of the analytic approach, interpreting findings, original manuscript draft writing, and review and editing of the manuscript. LW organized and analyzed all data required to meet the objectives of this study; generated all results, tables, and figures required for manuscript writing; contributed to model selection and results interpretation; and contributed to writing, review, and editing of the original manuscript. EL provided clinical and digital health expertise in interpreting the findings and contributed to original manuscript writing, review, and editing. RW provided statistical expertise and guidance to inform analysis planning, model selection, and results interpretation; contributed to original manuscript writing; and reviewed and edited the manuscript. JL, AL, and CC provided supervisory guidance and programmatic oversight throughout the research process, as well as manuscript review and editing. JL and AL also contributed to study design and methodology. CC and AV led clinical operations and quality assurance, including generation of client and participant data as part of routine measurement-based care and overseeing quality assurance practices in the therapy program. AV also contributed to the manuscript through review and editing. All authors read and approved the final manuscript.

Conflicts of Interest

HE, JL, and LW are employed by Lyra Health, receive income from Lyra Health, and have been granted equity in Lyra Health. CC, EL, AL, and AV are employed by Lyra Health and Lyra Clinical Associates, receive income from Lyra Health and Lyra Clinical Associates, and have been granted equity in Lyra Health. RW is a paid consultant for Lyra Health.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[DOCX File , 35 KB - jmir_v26i1e60502_app1.docx]

Multimedia Appendix 2
Missing data exploration.
[DOCX File , 3957 KB - jmir_v26i1e60502_app2.docx]

Multimedia Appendix 3
Extended growth curve modeling results.
[DOCX File , 36 KB - jmir_v26i1e60502_app3.docx]

Multimedia Appendix 4
Excluded participant characteristics.
[DOCX File , 26 KB - jmir_v26i1e60502_app4.docx]

References

1. Hofmann SG, Asnaani A, Vonk IJJ, Sawyer AT, Fang A. The efficacy of cognitive behavioral therapy: a review of meta-analyses. Cognit Ther Res 2012;36(5):427-440 [FREE Full text] [doi: 10.1007/s10608-012-9476-1] [Medline: 23459093]
2. Cuijpers P, Berking M, Andersson G, Quigley L, Kleiboer A, Dobson KS. A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. Can J Psychiatry 2013;58(7):376-385 [FREE Full text] [doi: 10.1177/070674371305800702] [Medline: 23870719]

XSL•FO
RenderX

3. Gloster A, Walder N, Levin M, Twohig M, Karekla M. The empirical status of acceptance and commitment therapy: A review of meta-analyses. Journal of Contextual Behavioral Science 2020 Oct;18:181-192 [FREE Full text] [doi: 10.1016/j.jcbs.2020.09.009]
4. Southward MW, Lane SP, Shroyer SE, Sauer-Zavala S. Do Unified Protocol modules exert general or unique effects on anxiety, depression, and transdiagnostic targets? J Mood Anxiety Disord 2023;3:100022 [FREE Full text] [doi: 10.1016/j.xjmad.2023.100022] [Medline: 37920572]
5. Cohen ZD, Barnes-Horowitz NM, Forbes CN, Craske MG. Measuring the active elements of cognitive-behavioral therapies. Behav Res Ther 2023;167:104364 [FREE Full text] [doi: 10.1016/j.brat.2023.104364] [Medline: 37429044]
6. Lebeau RT, Davies CD, Culver NC, Craske MG. Homework compliance counts in cognitive-behavioral therapy. Cogn Behav Ther 2013;42(3):171-179. [doi: 10.1080/16506073.2013.763286] [Medline: 23419077]
7. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. Behav Ther 2016;47(5):755-772. [doi: 10.1016/j.beth.2016.05.002] [Medline: 27816086]
8. Helbig S, Fehm L. Problems with homework in CBT: rare exception or rather frequent? Behav Cogn Psychother 2004;32(3):291-301. [doi: 10.1017/S1352465804001365]
9. Brown AJ, Bollini AM, Craighead LW, Astin MC, Norrholm SD, Bradley B. Self-monitoring of reexperiencing symptoms: a randomized trial. J Trauma Stress 2014;27(5):519-525. [doi: 10.1002/jts.21950] [Medline: 25322881]
10. Spinrad A, Taylor CB, Ruzek JI, Jefroykin S, Friedlander T, Feleke I, et al. Action recommendations review in community-based therapy and depression and anxiety outcomes: a machine learning approach. BMC Psychiatry 2024;24(1):133 [FREE Full text] [doi: 10.1186/s12888-024-05570-0] [Medline: 38365635]
11. Kazantzis N, Miller AR. A comprehensive model of homework in cognitive behavior therapy. Cogn Ther Res 2021;46(1):247-257. [doi: 10.1007/s10608-021-10247-z]
12. Callan JA, Kazantzis N, Park SY, Moore CG, Thase ME, Minhajuddin A, et al. A propensity score analysis of homework adherence-outcome relations in cognitive behavioral therapy for depression. Behav Ther 2019;50(2):285-299 [FREE Full text] [doi: 10.1016/j.beth.2018.05.010] [Medline: 30824246]
13. Conklin LR, Strunk DR. A session-to-session examination of homework engagement in cognitive therapy for depression: do patients experience immediate benefits? Behav Res Ther 2015;72:56-62 [FREE Full text] [doi: 10.1016/j.brat.2015.06.011] [Medline: 26183022]
14. Conklin LR, Curreri AJ, Farchione TJ, Barlow DH. Homework compliance and quality in cognitive behavioral therapies for anxiety disorders and obsessive-compulsive disorder. Behav Ther 2021;52(4):1008-1018. [doi: 10.1016/j.beth.2021.01.001] [Medline: 34134818]
15. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
16. Andersson G, Topooco N, Havik O, Nordgreen T. Internet-supported versus face-to-face cognitive behavior therapy for depression. Expert Rev Neurother 2016;16(1):55-60. [doi: 10.1586/14737175.2015.1125783] [Medline: 26610160]
17. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
18. Owusu JT, Wang P, Wickham RE, Varra AA, Chen C, Lungu A. Real-world evaluation of a large-scale blended care-cognitive behavioral therapy program for symptoms of anxiety and depression. Telemed J E Health 2022;28(10):1412-1420 [FREE Full text] [doi: 10.1089/tmj.2021.0590] [Medline: 35263185]
19. Cuijpers P, Miguel C, Harrer M, Ciharova M, Karyotaki E. The outcomes of mental health care for depression over time: a meta-regression analysis of response rates in usual care. J Affect Disord 2024;358:89-96 [FREE Full text] [doi: 10.1016/j.jad.2024.05.019] [Medline: 38710332]
20. Jackson HM, Calear AL, Batterham PJ, Ohan JL, Farmer GM, Farrer LM. Skill enactment and knowledge acquisition in digital cognitive behavioral therapy for depression and anxiety: systematic review of randomized controlled trials. J Med Internet Res 2023;25:e44673 [FREE Full text] [doi: 10.2196/44673] [Medline: 37256673]
21. Owusu JT, Wang P, Wickham RE, Cottonham DP, Varra AA, Chen C, et al. Blended care therapy for depression and anxiety: outcomes across diverse racial and ethnic groups. J Racial Ethn Health Disparities 2023;10(6):2731-2743 [FREE Full text] [doi: 10.1007/s40615-022-01450-z] [Medline: 36459363]
22. Lungu A, Wickham RE, Chen S, Jun JJ, Leykin Y, Chen CE. Component analysis of a synchronous and asynchronous blended care CBT intervention for symptoms of depression and anxiety: pragmatic retrospective study. Internet Interv 2022;28:100536 [FREE Full text] [doi: 10.1016/j.invent.2022.100536] [Medline: 35433277]
23. Haller E, Watzke B. The role of homework engagement, homework-related therapist behaviors, and their association with depressive symptoms in telephone-based cbt for depression. Cogn Ther Res 2020;45(2):224-235. [doi: 10.1007/s10608-020-10136-x]
24. Aguilera A, Ramos Z, Sistiva D, Wang Y, Alegria M. Homework completion via telephone and in-person cognitive behavioral therapy among Latinos. Cogn Ther Res 2018;42(3):340-347. [doi: 10.1007/s10608-017-9884-3]

XSL•FO
RenderX

25. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
26. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
27. Barlow DH, Farchione TJ, Sauer-Zavala S, Murray-Latin H, Bullis JR, Ellard K, et al. Unified Protocol for Transdiagnostic Treatment of Emotional Disorders: Therapist Guide. 2nd ed. New York, NY: Oxford University Press; 2017.
28. Hayes SC. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. Behav Ther 2004;35(4):639-665. [doi: 10.1016/S0005-7894(04)80013-3]
29. Linehan MM. DBT Skills Training Manual Second Edition. New York, NY: Guilford Press; 2014.
30. Wu MS, Wickham RE, Chen S, Chen C, Lungu A. Examining the impact of digital components across different phases of treatment in a blended care cognitive behavioral therapy intervention for depression and anxiety: pragmatic retrospective study. JMIR Form Res 2021;5(12):e33452 [FREE Full text] [doi: 10.2196/33452] [Medline: 34927591]
31. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
32. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037//0022-006x.59.1.12] [Medline: 2002127]
33. Gyani A, Shafran R, Layard R, Clark DM. Enhancing recovery rates: lessons from year one of IAPT. Behav Res Ther 2013;51(9):597-606 [FREE Full text] [doi: 10.1016/j.brat.2013.06.004] [Medline: 23872702]
34. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
35. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
36. Bates DM, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1):1-48.
37. Martell CR, Dimidjian S, Herman-Dunn R. Behavioral Activation for Depression: A Clinician's Guide. New York, NY: Guilford Press; 2021.
38. McGuire JF, Lewin AB, Storch EA. Enhancing exposure therapy for anxiety disorders, obsessive-compulsive disorder and post-traumatic stress disorder. Expert Rev Neurother 2014;14(8):893-910 [FREE Full text] [doi: 10.1586/14737175.2014.934677] [Medline: 24972729]
39. Karlin BE, Brown GK, Jager-Hyman S, Green KL, Wong M, Lee DS, et al. Dissemination and implementation of cognitive behavioral therapy for depression in the Kaiser Permanente health care system: evaluation of initial training and clinical outcomes. Behav Ther 2019;50(2):446-458. [doi: 10.1016/j.beth.2018.08.002] [Medline: 30824258]
40. Goldberg SB, Hoyt WT, Nissen-Lie HA, Nielsen SL, Wampold BE. Unpacking the therapist effect: impact of treatment length differs for high- and low-performing therapists. Psychother Res 2018;28(4):532-544. [doi: 10.1080/10503307.2016.1216625] [Medline: 27616739]
41. Population Health, Clinical Audit and Specialist Care Team. NHS Talking Therapies, for anxiety and depression, Annual reports, 2022-23. URL: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-talking-therapies-for-anxiety-and-depression-annual-reports/2022-23 [accessed 2024-09-18]
42. Dobson KS. A commentary on the science and practice of homework in cognitive behavioral therapy. Cogn Ther Res 2021;45(2):303-309. [doi: 10.1007/s10608-021-10217-5]
43. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
44. Lipschitz JM, Van Boxtel R, Torous J, Firth J, Lebovitz JG, Burdick KE, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022;24(10):e39204 [FREE Full text] [doi: 10.2196/39204] [Medline: 36240001]
45. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]

Abbreviations

AIC: Akaike information criterion
BIC: Bayesian information criterion
GAD-7: Generalized Anxiety Disorder 7-item scale
HIPAA: Health Insurance Portability and Accountability Act
LCT: Lyra Care Therapy
PHQ-9: Patient Health Questionnaire 9-item scale

XSL•FO
RenderX

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TVC: time-varying covariate

Edited by X Ma; submitted 15.05.24; peer-reviewed by K Mikkonen, A Parks; comments to author 13.06.24; revised version received 18.07.24; accepted 07.09.24; published 18.10.24.

Please cite as:
Espel-Huynh HM, Wang L, Lattie EG, Wickham RE, Varra A, Chen CE, Lungu A, Lee JL
Clinical Effects of Asynchronous Provider-Guided Practice Sessions During Blended Care Therapy for Anxiety and Depression: Pragmatic Retrospective Cohort Study
J Med Internet Res 2024;26:e60502
URL: https://www.jmir.org/2024/1/e60502
doi:10.2196/60502
PMID:39422996

©Hallie M Espel-Huynh, Lu Wang, Emily G Lattie, Robert E Wickham, Alethea Varra, Connie E Chen, Anita Lungu, Jennifer L Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Usage and Usability of a Mental Health Advice Telephone Service in Uganda: Mixed Methods Study

Johnblack K Kabukye[1,2], MBChB, MSc, PhD; Juliet Nakku[3], MBChB, MMed; Jackline Niwemuhwezi[3], BA, MSW; James Nsereko[3], BSc, MSc; Rosemary Namagembe[4], BSN; Iris Dorothee Emilie Groen[5], BA, MSci; Ritah Neumbe[3], BA; Denis Mubiru[3], Dip (Psy & PSW); Caroline Kisakye[3], BSc, MSc; Roseline Nanyonga[3], PCO; Marie Sjölinder[6], PhD; Susanne Nilsson[7], PhD; Caroline Wamala-Larsson[1], PhD

[1]Swedish Program for ICT in Developing Regions (SPIDER), Department of Computer and Systems Science, Stockholm University, Stockholm, Sweden
[2]Uganda Cancer Institute, Kampala, Uganda
[3]Butabika National Referral Mental Hospital, Kampala, Uganda
[4]Hutchinson Centre Research Institute of Uganda, Uganda Cancer Institute, Kampala, Uganda
[5]MoleMann Mental Health, Almere, Netherlands
[6]Research Institutes of Sweden (RISE), Stockholm, Sweden
[7]Unit for Integrated Product Development and Design, Department of Machine Design, KTH Royal Institute of Technology, Stockholm, Sweden

Corresponding Author:
Johnblack K Kabukye, MBChB, MSc, PhD
Swedish Program for ICT in Developing Regions (SPIDER)
Department of Computer and Systems Science
Stockholm University
Borgarfjordsgatan 12
Kista
Stockholm, 164 55
Sweden
Phone: 46 8162000
Email: kabukye@dsv.su.se

Abstract

Background: Harnessing mobile health (mHealth) solutions could improve the delivery of mental health services and mitigate their impact in Uganda and similar low-resource settings. However, successful adoption requires that mHealth solutions have good usability. We have previously implemented a telephone service to provide mental health information and advice in English and Luganda, utilizing an automated interactive voice response (IVR) system linked to live agents, including mental health care workers and peer support workers.

Objective: This study aims to assess the usage and usability of this mental health telephone service.

Methods: We obtained usage data from the system's call logs over 18 months to study call volumes and trends. We then surveyed callers to gather their characteristics and assess usability using the Telehealth Usability Questionnaire. Additionally, call recordings were evaluated for conversation quality by 3 independent health care professionals, using the Telephone Nursing Dialogue Process, and correlations between quality and usability aspects were investigated.

Results: Over 18 months, the system received 2863 meaningful calls (ie, calls that went past the welcome message) from 1125 unique telephone numbers. Of these, 1153 calls (40.27%) stopped at the prerecorded IVR information, while 1710 calls (59.73%) opted to speak to an agent. Among those who chose to speak with an agent, 1292 calls (75.56%) were answered, 393 calls (22.98%) went to voicemail and were returned in the following working days, and 25 calls (1.46%) were not answered. Usage was generally sustained over time, with spikes in call volume corresponding to marketing events. The survey (n=240) revealed that most callers were caregivers of patients with mental health issues (n=144, 60.0%) or members of the general public (n=46, 19.2%), while a few were patients with mental health issues (n=44, 18.3%). Additionally, the majority were male (n=143, 59.6%), spoke English (n=180, 75.0%), had postsecondary education (n=164, 68.3%), lived within 1 hour or less from Butabika Hospital (n=187, 77.9%), and were aged 25-44 years (n=160, 66.7%). The overall usability score for the system was 4.12 on a 5-point scale, significantly higher than the recommended target usability score of 4 (P=.006). The mean scores for usability components ranged from 3.66 for reliability to 4.41 for ease of use, with all components, except reliability, scoring higher than 4 or falling within its CI. Usability

XSL•FO
RenderX

scores were higher for Luganda speakers compared with English speakers, but there was no association with other participant characteristics such as sex, distance from the hospital, age, marital status, duration of symptoms, or treatment status. The quality of call conversations (n=50) was rated at 4.35 out of 5 and showed a significant correlation with usability (Pearson r=0.34, *P*=.02).

Conclusions: We found sustained usage of the mental health telephone service, along with a positive user experience and high satisfaction across various user characteristics. mHealth solutions like this should be embraced and replicated to enhance the delivery of health services in Uganda and similar low-resource settings.

(J Med Internet Res 2024;26:e65692) doi:10.2196/65692

KEYWORDS

mHealth; mental health; telephone service; usability; satisfaction; evaluation; mixed method; Uganda; Africa

Introduction

Mobile health (mHealth) solutions have been widely implemented in Africa to support various clinical goals and processes, including medication adherence, appointment attendance, patient communication, health education, clinical data collection and reporting, and clinical decision support, and guideline adherence [1-7]. Generally, positive results have been reported regarding acceptability, usability, and clinical outcomes; however, most implementations have focused on HIV/AIDS and maternal and child health, with limited application in other noncommunicable diseases such as cancer, hypertension, and diabetes [1-6]. Studies on the implementation of mHealth in mental health in Africa are limited. For instance, in a meta-analysis of mobile phone–based interventions for mental health conducted by Goldberg et al [8], none of the 145 identified randomized trials originated from Africa. Another review by Ding et al [9], which focused on mHealth for mental health among youth, found that only 3 out of 151 studies were from Africa (Nigeria). Similarly, in a review of mHealth in Sub-Saharan Africa by Aboye et al [1], only 2 out of 59 studies addressed mental health conditions: an app for dementia screening by lay health workers in Tanzania and a case study of telepsychiatry in Kenya following the outbreak of COVID-19. Other reviews by Ødegård et al [4] (31 clinical trials on 2-way SMS text message interventions) and Osei et al [5] (12 studies on mHealth for disease diagnosis and treatment in Africa) did not include any studies focused on mental health.

A key issue in implementing mHealth solutions is usability. The International Organization for Standardization [10] defines usability as the degree to which a product can be used by specific users to achieve particular goals with effectiveness, efficiency, and satisfaction in a given context. Usability directly affects users' intention to adopt and continue using technology [11], and thus, ensuring good usability is crucial for engagement with mental health mHealth solutions [12,13]. Moreover, individuals with mental illnesses, particularly in Africa, often face challenges such as not owning mobile devices or sharing them with family members or caregivers, which compromises privacy and confidentiality. Many live in rural areas with poor telecommunication coverage and limited access to electricity for charging devices, have low general and digital literacy, and experience significant stigma, among other difficulties [14-16]. These factors can hinder their ability to learn and use mHealth solutions or affect their satisfaction with them [12,13]. Mental health mHealth solutions, particularly smartphone apps, have also been criticized for being unnecessarily complex [13] and for imposing a high personal quantification burden [12], as they often require users to frequently input data about their behavior or mood for ecological momentary assessments.

Therefore, conducting research on the usability of mental health mHealth interventions [17] is crucial to inform improvements in their design, implementation strategies, and the evidence base for these technologies in Africa. Additionally, because usability is subjective, studying actual usage (eg, the number of returning users or the type of information sought) is essential, as it provides deeper insights into user satisfaction, perceived utility, and their unique challenges and needs.

Uganda, like many other low- and middle-income countries, faces a significant mental health burden and a substantial gap in mental health care. Approximately 1 in 3 Ugandans experience some form of mental illness, with depression (22.2%) and anxiety (20.2%) being the most prevalent [18,19]. Additionally, Uganda has one of the highest per-capita alcohol consumption rates, with alcohol and substance use disorders being widespread, particularly among men, while also affecting children and youth [20]. Several factors contribute to this critical situation, including a severe shortage of mental health care workers and underfunded and limited mental health facilities that are often distant, overcrowded, and prone to medication shortages. Additionally, a lack of awareness and negative sociocultural norms and beliefs hinder health-seeking behavior [21-23]. Previously, we implemented a mental health advice telephone service at the Butabika National Mental Referral Hospital in Kampala, Uganda [24] as a step toward addressing the challenges in Uganda's mental health care system. This paper evaluates the usage and usability of this tele–mental health service.

Methods

Study Design

A detailed description of the development of the mental health service is provided elsewhere [24]. In summary, we used a mixed methods, user-centered, and participatory design approach to assess the information needs and mental health care challenges in Uganda. Based on this assessment, we cocreated and implemented a telephone service that delivers mental health information through an automated interactive voice response (IVR) system, along with the option to connect users to a live agent for personalized advice. The service is available 24/7, accessible via any mobile phone without the need for internet,

XSL•FO
RenderX

and is toll-free (reverse-billed). The live agents comprise professional health care workers, including nurses, psychologists, and psychiatric clinical officers, as well as peer support workers—individuals recovering from mental illnesses who share their lived experiences.

This paper assesses the usage and usability of the service through a quantitative study that utilizes telephone system logs to evaluate usage patterns, a cross-sectional telephone survey of callers to assess usability, and rating of call conversation quality by healthcare workers. The study follows the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) guidelines [25].

Telephone System Usage

Data Collection

Data were obtained from the system call logs, which included the phone number, time stamp, duration, and selected IVR menu options. The data covered a period of 18 months, from August 2022, when the system deployment and testing were completed, to February 2024.

Analysis

Trends across calendar time, day of the week, and time of day were extracted from the call logs, along with the language used and the destination of the call (IVR, live agent, or voicemail). The data were summarized using descriptive statistics, including frequencies, percentages, medians, and ranges. Graphs were created to visualize these trends.

Usability Survey

Participants

Participants included individuals who had called the telephone system. Survey calls were made once or twice a week, and participants were recruited sequentially from the call logs of the preceding days. A survey log was maintained in a spreadsheet to track callers who had been successfully surveyed versus those for whom the survey could not be completed, such as instances where their phones were unreachable. If more than a month had passed since an individual last called the system without being successfully surveyed, they were excluded from the list of eligible participants, as it was deemed too long for the caller to accurately recall their call experience. The surveys were conducted in either English or Luganda, depending on the participant's preference. Data were entered in real-time into a Google Form (Google LLC/Alphabet Inc.), which was configured to shuffle the order of the questionnaire items to minimize any potential ordering effects [26].

The survey was administered by trained research assistants, including nurses, peer support workers, and psychiatric clinical officers, who were part of the project team and served as call agents. In the month leading up to data collection, the first author (JKK) trained the research assistants on the questionnaire content, the process for obtaining informed consent over the phone, and techniques to avoid "leading" respondents toward specific answers. Additionally, the first author (JKK) monitored quality control by listening to at least four randomly selected survey calls each month.

Sample Size

We estimated the sample size using the recommendation by Park and Jung [27], which suggested a target of 241 participants for a Likert scale encompassing 5 dimensions or usability attributes. This estimation assumes a correlation coefficient of 0.5 and a tolerable error margin of 5%, allowing for a maximum of 4 items per dimension if each dimension is analyzed separately.

Data Collection

We utilized the Telehealth Usability Questionnaire (TUQ), developed and validated by Parmanto et al [28]. The TUQ builds on established usability questionnaires and technology acceptance theories, aiming to evaluate a wide range of telehealth setups. Unlike traditional videoconferencing systems designed for single purposes and typically operated by clinicians or technical teams, the TUQ applies to both general-purpose computers and mobile phones. The TUQ can be used to evaluate usability from the patient's perspective. The original questionnaire comprises 21 items that cover 5 usability attributes: usefulness, ease of use and learnability, interface quality, interaction quality, and reliability, as well as 1 overall satisfaction dimension: satisfaction and future use. For this study, the language was adapted to fit the context by replacing "telehealth system" with "Butabika call center service." Additionally, the item "the system is simple and easy to understand" was modified to "the information is simple and easy to understand." The item "I believe I could become productive quickly using this system," deemed irrelevant to our survey, was also removed. Each item was scored on a 5-point Likert scale, with 1 representing the lowest score and 5 the highest. Additionally, the questionnaire included questions about demographic characteristics and mental health information, as well as a section for free-text comments.

Analysis

Usability survey data were analyzed by summarizing participants' characteristics (frequencies and percentages). Means and SDs were calculated for the TUQ items, subscales (usability attributes), and the overall scale. The scale's reliability was assessed using Cronbach α. Differences in usability scores among participants' characteristics were analyzed using the 2-tailed, unpaired *t* test (for language, sex, distance from the hospital, and duration of illness) and ANOVA (for caller category, level of education, and marital status).

The free-text comments were analyzed using a deductive approach [29]. The first (JKK) and fifth (RNamagembe) authors independently coded the comments into themes related to participants' general opinions about the system, usability challenges, and recommendations for improvement.

Call Conversation Quality Rating

Participants

The quality assessment was conducted by the third, fourth, and fifth co-authors (JaNi, JaNs, RNamagembe), all of whom have been involved in the development and operation of the call center system. JaNi is a social worker, and JaNs is a psychologist

XSL•FO
RenderX

at Butabika Hospital, while RNamagembe is a nurse at the Uganda Cancer Institute.

Data Collection

Fifty recordings of telephone conversations with the usability survey participants were randomly selected and independently assessed using the telephone nursing dialogue process [30]. The assessment focused on various aspects, including building *rapport*, actively *listening*, gathering and *analyzing* information, informing and *motivating* the caller, and *ending* with mutual agreement and safety netting. The agent who answered the call (peer support work vs professional health care provider) was also recorded, as well as the call issue, that is, if it was solvable on the phone (defined as "caller needed general information such as how to access mental health services, was providing information to us such as reporting a patient who "escaped," or just checking if the system works") or not (defined as "a thorough clinical assessment was needed, prescription of medication, urgent intervention for a patient with suicidal ideation, etc").

The recordings were accessed via a password-protected, time-limited cloud folder, and the assessments were documented in Google Forms.

Analysis

The ratings of the call conversations were summarized as means and SDs. The reliability of the quality rating scale was evaluated using Cronbach α. Additionally, the correlation between the ratings provided by health professionals and the callers' usability scores was computed to determine if a relationship exists. Chi-square tests were conducted to assess associations between the agent type and the call issue.

All data cleaning and analyses of usage, usability, and conversation quality were conducted using MS Excel (version 16; Microsoft Corp.) and SPSS (version 29; IBM Corp.).

Ethics Considerations

Ethical approval for the research study was obtained from the Makerere University School of Public Health Research Ethics Committee (approval number SPH-2021-153) and the Uganda National Council of Science and Technology (approval number HS1868ES). The research assistant obtained informed consent from the participants before administering the survey, clarifying that participation was voluntary and would not affect their medical care. As compensation for their time, participants received UGX 5000 (approximately US $1.3) in phone credit, as recommended by the research ethics committee.

Results

Usage Patterns

From August 2022 to February 2024, a total of 5120 calls were made to the system. Of these, 2257 calls (44.08%) did not interact with the system; these callers were primarily checking to confirm the system's operation or failed to understand the IVR instructions, resulting in no selection of IVR options. The remaining 2863 calls (55.92% of all calls, from 1125 unique telephone numbers) involved meaningful interactions, where callers chose options from the IVR menu. Notably, nearly half of these calls (1153 calls, or 40.27%, from 609 unique phone numbers) ended after accessing prerecorded IVR information. Out of the total calls (n=2863), 1710 (59.73%) opted to speak with a live agent. Among these, 1292 calls (75.56%, from 593 unique phone numbers) were answered, 25 calls (1.46%, from 25 unique numbers) were not answered, and 393 calls (22.98%, from 254 unique phone numbers) occurred during out-of-office hours and were directed to voicemail; agents returned these calls within the following working days. In total, the calls accumulated to 138.3 hours, with each unique number making a median of 2 calls (minimum=1, maximum=77), and the average duration of each call was approximately 3 minutes. The distribution of calls in English was roughly equal to that in Luganda, and there was no significant difference in call duration (to the agents) between the 2 languages (mean 2.99 vs 3.06 minutes, P=.31). Details per call type are presented in Table 1.

Table 1. Summary of the calls.

Call type	IVR[a]	Agent	Voicemail	Overall[b]
Number of calls, n	1153	1292	393	2838
Total duration of calls (hours)	49.5	65.0	23.7	138.3
Average duration of each call (minutes)	2.6	3.0	3.6	2.9
Calls in Luganda, n (%)	633 (54.90)	656 (50.77)	189 (48.09)	N/A[c]
Unique telephone numbers, n	609	593	254	1456
Median number of calls from a single number	1	2	1	2
Maximum number of calls from a single number, n	34	25	24	77

[a]IVR: interactive voice response.

[b]There is an overlap in the callers to IVR and live agents (187 phone numbers) and live agents and voicemail (86 phone numbers), that is, multiple calls from a single number choosing IVR, live agent, or voicemail on different calls. Therefore, the total in the table exceeds the figure (1125 unique telephone numbers) mentioned above. The numbers in the "Overall" column are calculated from the combined call data set of IVR, live agent, and voicemail rather than the summation of the other column figures. Further, the 25 unanswered calls are not included.

[c]N/A: not applicable.

XSL•FO
RenderX

Figure 1 illustrates the mental health information topics accessed through the IVR, along with their respective frequencies (direct calls and replays combined). The most frequently sought information pertained to the general understanding of mental illnesses, including their signs, symptoms, and causes. This accounted for 670 (44.61%) of the 1502 calls for which the IVR topic was recorded in the system. In addition, Luganda was the predominant language for IVR messages (911/1502, 60.65%).

Figure 1. Mental health messages listened to in the IVR. IVR: interactive voice response.

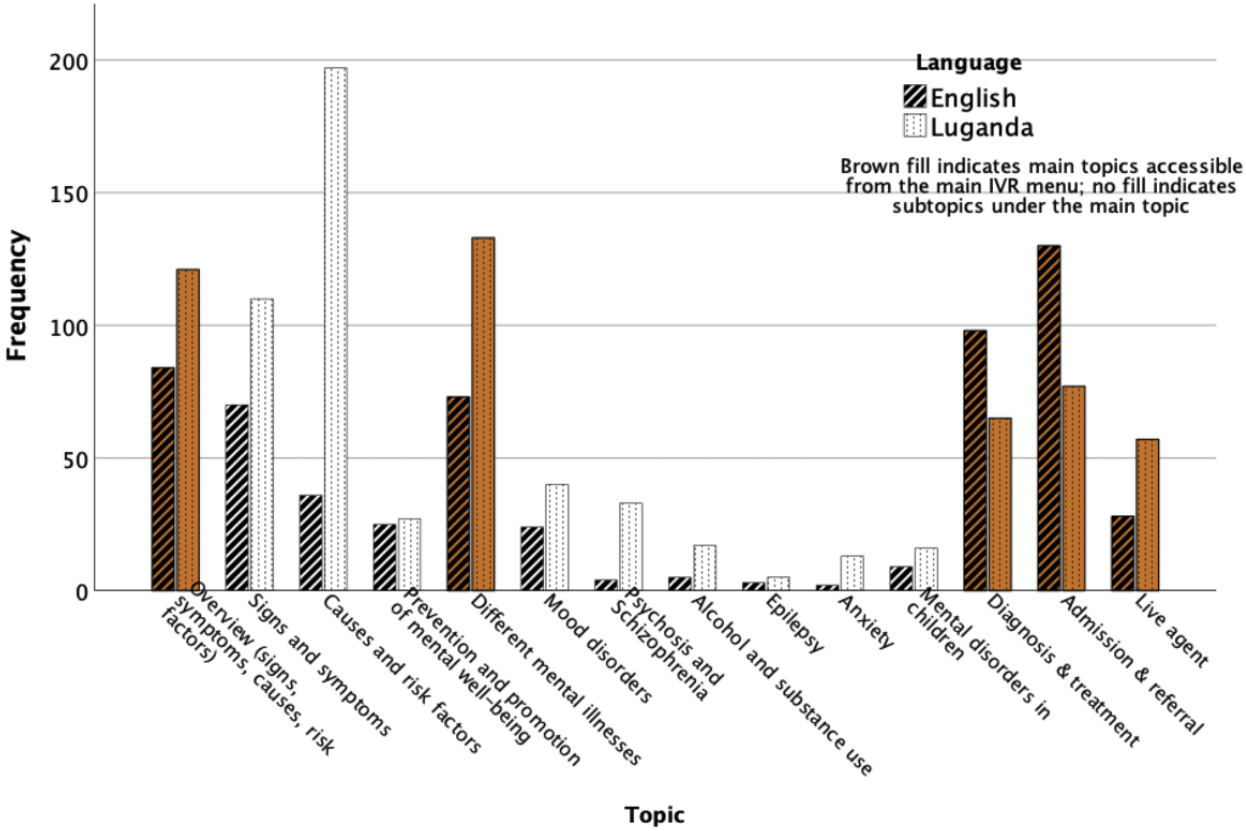

Multimedia Appendix 1 presents the call trends over the 18-month period, while Multimedia Appendix 2 displays the distribution of calls throughout the day, and Figure 2 illustrates call distribution across the days of the week. Overall, usage remains consistently sustained over time, with notable spikes in call volume corresponding to marketing and sensitization events, such as community outreach initiatives where the call center number was shared with patients and the general public. Relatively more calls were made earlier in the week and during daytime hours compared with weekends and nighttime. Call agents, including nurses, psychiatric clinical officers, psychologists, and peer support workers, were available only during office hours (8 AM to 6 PM, Monday to Friday). The system was configured not to route any calls to live agents outside of these hours, which explains the absence of calls to live agents and the prevalence of voicemails on Saturdays and Sundays.

XSL•FO
RenderX

Figure 2. Distribution of call types per day of the week. There were no live agents on Saturdays and Sundays. IVR: interactive voice response; VMS: voicemail system.

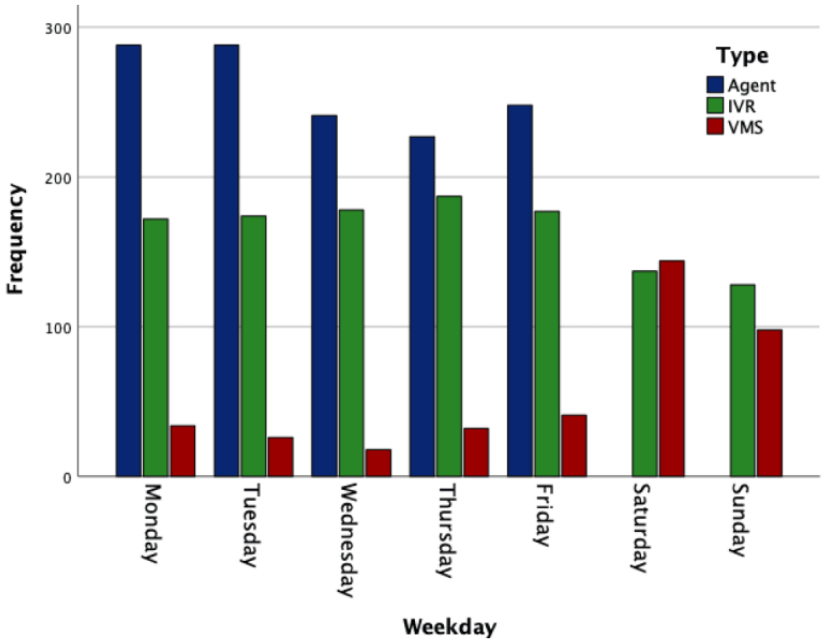

Usability

Characteristics of Usability Survey Participants

A total of 252 callers were surveyed, but 12 were excluded from the analysis due to incomplete usability question responses (eg, resulting from poor phone network connections or lack of time) or because they reported difficulty recalling their experience with the call center service. The remaining 240 surveys analyzed included responses from caregivers of patients with mental health issues (n=144, 60.0%), members of the general public (n=46, 19.2%), patients with mental health issues (n=44, 18.3%), health care providers (n=5, 2.1%), and peer support workers (n=1, 0.4%).

The majority of participants were male (n=143, 59.6%), spoke English (n=180, 75.0%), had postsecondary education (n=164, 68.3%), lived within 30 km (approximately 1 hour or less) from Butabika Hospital (n=187, 77.9%), and were aged 25-34 years (n=97, 40.4%) or 35-44 years (n=63, 26.3%). Among the patients (ie, those categorized as either a patient or caregiver; n=188), most had experienced their illness for over a year (n=147, 78.2%) and were already receiving treatment (n=134, 71.3%). The most common mental health issues reported were alcohol and substance-related disorders (n=48, 25.5%) and mood disorders (n=43, 22.9%). Additional participant characteristics are summarized in Table 2.

XSL•FO
RenderX

Table 2. Characteristics of usability survey participants (N=240).

Characteristic and categories	Values, n (%)
Category of participant (caller)	
Caregiver	144 (60.0)
Other[a]	46 (19.2)
Patients with mental health issues	44 (18.3)
Health care provider	5 (2.1)
Peer support worker	1 (0.4)
Language	
English	180 (75.0)
Luganda	60 (25.0)
Sex of participant (caller)	
Male	143 (59.6)
Female	97 (40.4)
Age (years) group of participant (caller)	
18-24	39 (16.3)
25-34	97 (40.4)
35-44	63 (26.3)
45 and above	41 (17.1)
Does the phone belong to you?	
Yes	237 (98.8)
No	3 (1.3)
Highest level of education completed	
University degree or higher	114 (47.5)
Tertiary but not degree	50 (20.8)
“A” level	19 (7.9)
“O” level	33 (13.8)
Primary school	22 (9.2)
None	2 (0.8)
Marital status	
Married/cohabiting	117 (48.8)
Never married	107 (44.6)
Separated/widowed	16 (6.7)
Distance to Butabika (30 km is approximately 1 hour travel time)	
≤30 km	187 (77.9)
>30 km	53 (22.1)
Sex of the patient [b] (N=188)	
Male	108 (57.4)
Female	80 (42.6)
Age group (years) of the patient [b] (N=188)	

XSL•FO
RenderX

Characteristic and categories	Values, n (%)
Under 18	33 (17.6)
18-24	33 (17.6)
25-34	68 (36.2)
35-44	33 (17.6)
≥45	21 (11.2)
Complaints [c] (N=188)	
Alcohol and substance-related disorders	48 (25.5)
Mood disorder (eg, depression, bipolar)	43 (22.9)
Psychotic disorders	26 (13.8)
Epilepsy	21 (11.2)
General enquiry or problem not clear	20 (10.6)
Physical symptoms/illness	11 (5.9)
Cognitive problems	9 (4.8)
Anxiety disorders	5 (2.7)
Developmental disorders	3 (1.6)
Sexual and gender identity disorders	2 (1.1)
When did the illness start? (N=188)	
A year or more ago	147 (78.2)
Less than a year ago	41 (21.8)
Is the patient on treatment? (N=188)	
Yes	134 (71.3)
Alternative medicine	3 (1.6)
No	51 (27.1)

[a]Other, for example, included community members reporting a person in the community who seems to have a mental illness and requesting guidance on how to deal with them, news agencies, and students looking for internships.

[b]Callers who said they were caregivers of a patient with mental health issues were asked for details about their patient. Together with callers who were patients themselves, these formed the total number of patients.

[c]Diagnoses reported by the participant; otherwise the research team derived the most likely diagnosis from the reported symptoms.

Usability Scores

The Cronbach α for the overall scale was 0.89, while for the subscales it ranged from 0.60 (Ease of Use) to 0.76 (Interaction Quality). Deleting any item resulted in a decrease in the α value.

The overall mean usability score for the system was 4.12 (SD 0.69) on a 5-point scale, which was significantly higher than the midpoint score of 3 (P<.001) and also significantly exceeded the recommended target usability score of 4 (P=.006) [31,32].

The mean scores for the individual TUQ items ranged from 3.50 (SD 1.28) for the item “The system is able to do everything I would want it to be able to do” to 4.61 (SD 0.80) for the item “I would use the mental health call center again.” All items had scores higher than the midpoint of 3 (all Ps<.001), and 14 of the 20 items exceeded the recommended score of 4 on a 5-point scale [31,32]. The 6 items that scored below 4 included all 3 items in the Reliability component, as well as 1 item from each of the Usefulness, Interface Quality, and Satisfaction components (Table 3).

XSL•FO
RenderX

Table 3. Usability components, questionnaire items, and their means, SDs, and Cronbach α for the subscales (usability attributes).[a]

Usability component and items	Item mean (▫)	SD	*P* value[b]	Mean difference (▫–4)	Component mean[c]	SD	(▫–4)	*P* value	Component α[c]
Usefulness					3.98	0.91	–0.02	.35	.61
U1: The mental health call center service improves my access to health care services	4.09	1.11	.11	0.09					
U2: The mental health call center service saves me time traveling to a hospital or specialist clinic	4.04	1.28	.32	0.04					
U3: The mental health call center service provides for my health care needs	*3.83*	*1.22*	*.02*	*–0.17*					
Ease of use					4.41	0.80	0.41	<.001	.60
E1: It was simple to use the call center system	4.43	0.89	<.001	0.44					
E2: It was easy to learn to use the call center system	4.42	0.94	<.001	0.42					
Interface quality					4.18	0.75	0.18	<.001	.66
UI1: The way I interact with the providers over the call center system is pleasant	4.54	0.91	<.001	0.54					
UI2: I like using the call center service	4.26	1.06	<.001	0.26					
UI3: The information is simple and easy to understand	4.47	0.86	<.001	0.47					
UI4: This system is able to do everything I would want it to be able to do	*3.50*	*1.28*	*<.001*	*–0.50*					
Interaction quality					4.30	0.81	0.30	<.001	.76
UX1: I could easily talk to the mental health provider using the call center service	4.19	1.18	.008	0.19					
UX2: I could hear the health provider clearly using the call center system	4.54	0.91	<.001	0.54					
UX3: I felt I was able to express myself effectively	4.43	0.96	<.001	0.43					
UX4: Using the call center service, I could interact with the health care provider or peer support worker as well as if we met in person	4.07	1.20	.19	0.07					
Reliability					3.66	1.03	–0.34	<.001	.60
R1: I think the care provided over the call center service is the same as in-person visits	*3.54*	*1.32*	*<.001*	*–0.46*					
R2: Whenever I made a mistake using the system, I could recover easily and quickly	*3.83*	*1.24*	*.03*	*–0.17*					
R3: The system gave error messages that clearly told me how to fix problems	*3.68*	*1.40*	*.001*	*–0.32*					
Satisfaction and future use					4.17	0.80	0.17	<.001	.66

XSL•FO
RenderX

Usability component and items	Item mean (▫)	SD	*P* value[b]	Mean difference (▫–4)	Component mean[c]	SD	(▫–4)	*P* value	Component α[c]
S1: I feel comfortable communicating with the mental health care provider using the call center service	4.24	1.15	<.001	0.24					
S2: The call center service is an acceptable way to receive health care services (compared for example to face-to-face)	*3.68*	*1.31*	*<.001*	*–0.32*					
S3: I would use mental health call center service again	4.61	0.80	<.001	0.61					
S4: Overall, I am satisfied with this mental health call center service	4.18	1.11	.007	0.18					
Overall					4.12	0.69	0.12	.006	.89

[a]Items in italics are less than 4, whereas the rest are higher than 4 or equivalent (within its CI).

[b]*P* value: 1-tailed *t* test (whether the mean is significantly less than 4).

[c]Component mean and α were calculated for items in the individual subscale.

The mean scores for the usability components (subscales) ranged from 3.66 (SD 1.03) for Reliability to 4.41 (SD 0.80) for Ease of Use. All scores were significantly higher than the midpoint of 3 (*P*<.001), and all components, except for Reliability, exceeded 4 or fell within their respective CIs.

The overall usability was slightly higher for Luganda speakers (4.24/5) compared with English speakers (4.07/5), but not statistically significant (*P*=.05). However, Luganda speakers demonstrated significantly higher scores in several usability components: interface quality (4.36/5 vs 4.12/5, *P*=.01), interaction quality (4.47/5 vs 4.25/5, *P*=.04), reliability (3.99/5 vs 3.55/5, *P*=.002), and satisfaction and future use (4.39/5 vs 4.10/5, *P*=.008).

There were no statistically significant differences in usability scores based on gender (*P*=.63), among participants living within 30 km (1 hour) from Butabika Hospital compared with those living farther away (*P*=.72), for patients who had experienced mental illness for a year or less versus those with a longer duration (*P*=.60), or for those receiving treatment versus those who were not (*P*=.65). Additionally, no differences in usability scores were observed across various participant categories (*P*=.67), age groups (*P*=.46), levels of education (*P*=.51), or marital status (*P*=.51).

Qualitative Comments

A total of 155 comments were coded into 6 themes (Table 4). The majority of the comments (n=60) expressed *appreciation* for the call center initiative, highlighting positive experiences with the service. One participant remarked that she was “very pleasantly surprised that [the hospital has] a toll-free line and it actually worked”. Conversely, 26 comments addressed *usability challenges*, including poor network connectivity, long telephone queues, callback delays, and difficulties navigating the IVR instructions for non–tech-savvy callers.

XSL•FO
RenderX

Table 4. Qualitative comments from the usability survey (N=155).

Theme	Description	Frequency, n (%)[a]	Illustrative quotes
Appreciation	Comments where participants expressed appreciation, approval, compliments, or other positive opinions (eg, because it is convenient, free, or improves access to mental health services).	60 (38.7)	• interactive and caring people • the customer care is good • happy [to get] the right information about the prices of the private services before coming to the hospital • very pleasantly surprised that [the hospital has] a toll-free line and it actually worked • Just thanking you for the service, because otherwise as a person in a rural area far away [I]was stuck with [my] grand daughter and had no other means of getting help. [I] was attend to well. • Having access to this information free of charge is wonderful using the toll free number. • I t's a good innovation, it's good to get the information you need before coming to the hospital
Usability challenges	Comments where participants reported difficulty using the system (eg, due to poor network; long telephone queues before connecting to an agent; difficulty in following the interactive voice response instructions; and mismatch in expectations, such as expecting an agent 24/7 instead of voicemail over the weekend or trying to book an appointment that was not possible).	26 (16.8)	• For a village person [someone who is not tech-savvy]the system is difficult • There should be clear network and quite environment. • T he length of time it takes to get to talk to a responder is too long, it should be the first option and not the last • G et more call responders so that a caller does not wait for long on line • T he call center responder[s] take long to return voicemail call.
Recommendations for improving the call center	Suggestions for improving the reach and impact of the call center service (eg, more marketing and sensitization of the public about it, adding a video call option, more training for the agents, adding more agents to ensure faster response and 24/7 availability, including specialists to give therapy on phone, adding more languages and agents who understand cultural and religious perspectives of calls, and improve continuity by integrating with other information systems).	48 (31.0)	• If the system could be operating 24/7 with a respondent and not just voice mail on weekend and after 8pm. • They should keep updating the call center responders to the current information in different departments in the hospital. • The call center should have all information or file data concerning the patients so that if one calls in to find out about their patient, there is an answer for them • Do more of the sensitization of the number to the public. • There should be professional personnel like doctors to answer some technical questions. • It would be good if there is a person who can give information depending on the cultural, religious and other affiliation • The stuff responding to the callers should have phone etiquette

XSL•FO
RenderX

Theme	Description	Frequency, n (%)[a]	Illustrative quotes
Recommendations for improving mental health care services in general	Comments or suggestions for improving mental health care services beyond the call center (eg, reducing medicine stock-outs, availing ambulances and pick-up services for patients who might be difficult to handle due to their mental illness, improving patient security and care while they are in Butabika Hospital, having more mental health care providers in rural areas, and other digital services that supplement the call center).	26 (16.8)	• We are not getting enough drugs when we come for review yet the drugs are very expensive and patients keep relapsing whenever they miss. • ...need to put a psychologist in different centers for easy access to services • There should be proper directions when someone is in the hospital • Security [should] be tightened so that patients do not escape easily • An emergency team should be put in place to pick-up people that have lost their sanity from the community • On the Butabika website prices of all services offered at the hospital should be shown to the public clearly for one to know what they need off hand • Avail download of prescriptions
Questions about mental illnesses	Participants asked questions related to mental illnesses	6 (3.9)	• Is depression curable? • What [is the] time span for someone to heal from mental illness?
Issues related to the survey	Comments related to the telephone survey or the items in the questionnaire	4 (2.6)	• Send the questionnaire via email • Instead of asking someone: “I like using the call center” use “I prefer using the call center [compared to...

[a]There were a total of 155 comments, but some comments were coded into more than 1 theme. Hence, the sum of the frequencies is 170.

Several comments included recommendations for improvement. The majority (48 comments) focused on *enhancing the call center service.* Suggestions included increasing marketing and sensitization efforts to improve outreach, adding more staff to reduce call waiting times and ensure 24/7 availability, providing training for agents, incorporating clinical specialists to offer therapy over the phone, and adding video calling options. Other recommendations focused on *improvements in mental health care services in general.* Suggestions included improving the availability of medications, enhancing patient care and security at Butabika Hospital, increasing access to mental health services in rural areas, and establishing community outreach and ambulance services to assist in transporting patients with unstable mental health conditions. A few comments were related to the usability survey (2 comments) or posed questions about mental illness (6 comments).

Assessment of Call Conversations by Health Workers

Call conversation ratings are presented in Table 5. The overall mean quality of the call conversations, as assessed by health workers, was 4.35 (SD 0.69) out of 5. Among the raters, rater 3 provided the lowest rating of 3.84 (SD 1.08), while rater 2 assigned the highest score of 4.66 (SD 0.75). The individual quality aspects varied, with scores ranging from 4.27 (SD 0.82) for ending to 4.49 (SD 0.57) for listening.

XSL•FO
RenderX

Table 5. Call conversation assessment by health workers.

Quality aspect and rater	Mean (SD)
All (average)	
Rater 1	4.55 (0.75)
Rater 2	4.66 (0.69)
Rater 3	3.84 (1.08)
All raters	4.35 (0.69)
Rapport	
Rater 1	4.69 (0.68)
Rater 2	4.64 (0.74)
Rater 3	3.55 (1.39)
All raters	4.29 (0.72)
Listening	
Rater 1	4.70 (0.65)
Rater 2	4.73 (0.67)
Rater 3	4.04 (1.16)
All raters	4.49 (0.57)
Analyzing	
Rater 1	4.40 (0.99)
Rater 2	4.64 (0.74)
Rater 3	3.98 (1.20)
All raters	4.34 (0.77)
Motivating	
Rater 1	4.45 (0.97)
Rater 2	4.66 (0.76)
Rater 3	3.94 (1.27)
All raters	4.35 (0.80)
Ending	
Rater 1	4.51 (0.90)
Rater 2	4.61 (0.83)
Rater 3	3.69 (1.29)
All raters	4.27 (0.82)

The quality aspects of the call conversations were highly correlated, with correlation coefficients ranging from 0.72 (between rapport and ending) to 0.92 (between analyzing and motivating). Additionally, the overall reliability of the quality rating scale was excellent, with a Cronbach α of 0.97.

The overall call quality showed a significant correlation with overall usability (Pearson r=0.34, P=.02). The correlations between individual quality aspects and usability components are presented in Multimedia Appendix 3, with the strongest correlation observed between ending and satisfaction (Pearson r=0.55, P<.001).

There was no association between the type of agent (peer support worker vs professional health care worker) or the nature of the call issue (whether solvable over the phone or not) and the usability scores provided by the callers.

Discussion

Principal Findings

This study evaluated the usage and usability of a mental health telephone advice service in Uganda. The findings indicate that all 3 access channels of the mental health telephone service (IVR, live calls, and voicemail) were utilized, and usage remained consistent throughout the evaluation period. The substantial proportion of callers (1153/2863 calls, 40.27%) who stopped at the IVR, despite having the option to speak with a live agent, is encouraging as it suggests a level of acceptability for the IVR system, potentially even a preference for it over

XSL•FO
RenderX

direct interaction with an agent. Utilizing IVR alleviates the burden on limited human resources and facilitates access to information outside of regular office hours. The call distribution data (see Multimedia Appendix 2 and Figure 2) illustrate that a significant portion of calls occurred on weekends and during the evening and night when staff were unavailable to respond, highlighting the IVR's utility in meeting callers' needs during these times. Additionally, all callers who opted to speak with an agent first navigated through the IVR, thereby accessing automated information before their conversation. In the formative research for this service [24], participants expressed a desire for anonymity to mitigate stigma, and they felt that the call center service provided this necessary confidentiality. Similarly, in another study focused on a cancer awareness service in Uganda [33], callers noted that a key advantage of IVR was the ability to listen to information multiple times, a feature that is often impractical or uncomfortable when interacting with a live agent (eg, when repeatedly requesting the same information). These perspectives may help clarify why a significant number of callers opted not to speak with an agent. Conversely, some survey participants indicated a preference for having the option to speak to an agent presented as the first choice. This suggests that the higher percentage of IVR users could be attributed to our system design, which intentionally prioritized IVR options over direct access to an agent, rather than reflecting a definitive user preference.

It is noteworthy that only a small proportion of the callers (44/240, 18.3%) were patients themselves, while the majority were caregivers (144/240, 60.0%) or other members of the general public (46/240, 19.2%). Butabika Hospital serves as a national referral center for individuals with severe mental health disorders, and prior studies in Uganda have indicated that people typically seek care only when mental health issues reach a severe level [34,35]. Such patients may struggle to arrange their own care or comprehend information provided through telephone services. Additionally, individuals with mental disorders often experience social isolation and may not have access to phones, further limiting their ability to seek assistance [24,36]. The significant role of caregivers (family members) must be taken into account when designing mental health interventions. For telephone services like ours, it is crucial to thoroughly explore and address the information needs of caregivers, as these may differ from those of the patients themselves. Understanding and accommodating these unique needs can enhance the effectiveness of mental health services and improve support for both patients and their caregivers [36].

The results indicated significantly high scores across various usability aspects and for the overall system, surpassing the recommended target usability scores [31,32]. This can be attributed to the user-centered approach applied during the design and implementation of the system. By prioritizing the unique circumstances of the target users in Uganda—such as their literacy levels, information needs, mental health care experiences, cultural norms, health-seeking behaviors, and familiarity with technology—we were able to effectively address their specific requirements and enhance the system's usability. Similar findings of high usability—particularly regarding satisfaction, intention for future use, ease of use, and perceived usefulness—have been reported in other studies on tele-mental health care in the United States [37], South Korea [38], the United Kingdom [39], and Italy [40], albeit within the context of COVID-19 lockdowns. This indicates that cultural factors, such as the "power distance" [41,42] between health care workers and patients—which tends to be higher in Africa [43-45]—did not significantly impact usability scores. The lack of fear among participants in rating the service may suggest a shift in attitudes toward the provider-patient relationship, which is often characterized by paternalism in African contexts. We could not find any studies in Uganda that have assessed aspects of the usability of digital health solutions from the client's perspective. As triangulated from the qualitative comments and call quality ratings provided by health care providers in our study, participants appreciated the convenience and the ability to access information through the telephone service at no cost. Additionally, it was evident that some participants had previously encountered negative experiences with other call center services or did not expect this hospital to establish one that "actually worked" and featured "pleasant and caring" agents. These factors contributed to the high usability scores observed in our study.

There was no difference in usage (number of calls) and efficiency (duration of calls) overall between English and Luganda. However, a slight majority of calls that stopped at the IVR were in Luganda. Additionally, the IVR options (ie, mental health information topics) sought differed for the 2 languages. Luganda-speaking callers primarily sought information to understand mental illnesses, including what mental illnesses are and what causes them. By contrast, English-speaking callers were more interested in the management of mental illnesses, specifically information on diagnosis, treatment, and the referral process. English proficiency serves as a proxy for formal literacy in Uganda, as English is taught in schools and used in health care and other official communications, except for clients with low literacy. The finding that callers with less literacy required more information for a basic understanding of mental illnesses is not surprising. Literate community members can access various mental health information sources, including online resources and print media, which may lead to fewer unmet mental health information needs among them. This is further supported by the finding that usability scores were slightly higher for Luganda speakers than for English speakers, suggesting that they appreciate the service more or consider it more useful, as it represents one of the few sources of mental health information available to them. As noted by Park et al [38], when users lack alternatives—whether due to COVID-19 lockdowns or, in Uganda's case, limited access to mental health care services in general—they may rate the telephone service more optimistically.

Similar to our findings, other studies have reported that the reliability component of usability is often ranked the lowest, with users expressing a desire for physical encounters with mental health care providers [37,39], despite their appreciation of the telephone service. Nonverbal communication plays a crucial role in mental health, as it contributes to comprehensive assessment and the development of a therapeutic alliance. Therefore, telephone services should supplement rather than

XSL•FO
RenderX

replace physical clinic visits. Videoconferencing options have also been proposed in our study and in others [37,39]; however, this could be constrained in Uganda due to limited access to the internet and smart devices in rural areas.

Connolly and colleagues [13] state that the evidence base for the usability of mental health apps is still uncertain and argue that simple telephone features can effectively deliver mental health services instead of relying on expensive smart apps with poor usability. Our study contributes to the literature on usability and demonstrates how basic phone features, such as IVR and voice calls, can provide significant value and a positive user experience. We also demonstrated efficient use of human resources, as we did not hire new staff but instead utilized over 130 hours from existing staff and volunteer peer support workers. With very few unanswered calls (25/1710 calls, 1.46%) and no reports from callers of congested telephone lines or long queues, we could potentially gain more hours from staff if we market the service more effectively to increase demand. Analyzing usage trends can help identify the busiest days of the week or hours of the day, thereby informing staff scheduling, further improving efficiency, and reducing the number of unanswered calls even when demand rises.

Limitations

There are a few limitations to this study. First, the usability survey was not conducted in real time; instead, it occurred a few days or even weeks after the participant called the mental health telephone service. This delay could have introduced recall bias, as some participants reported difficulty remembering their experience, leading to the termination of their surveys. However, it is also possible that some participants continued with the survey despite their poor recall. Future studies should aim to collect user feedback immediately after the call to enhance the accuracy of the responses. Moreover, the possibility of selection bias exists, where individuals who had a positive experience with the service may have been more likely to agree to participate in the survey, while those who encountered network connectivity challenges or had negative experiences may have been less reachable or less willing to engage. This potential bias could affect the overall assessment of usability and user satisfaction.

Another limitation is the cross-sectional nature of the survey. Users of the mental health telephone service called the system a median of 2 times, and their experiences may have varied across these encounters. Collecting feedback immediately after each call and tracking responses from each caller during different encounters could provide more insights into factors influencing usability, such as the specific agent who answered the call. Qualitative responses indicated that some participants felt the communication skills of certain agents were lacking; however, we could not identify the agents to subanalyze the usability scores associated with each individual.

Lastly, this study focused on a single system at a single site, which may limit the generalizability of the findings. Without a comparative system, interpreting individual usability scores could be challenging. Typically, usability evaluations are conducted using an A/B setup to compare different systems or versions of the same system. In our study, however, we utilized the recommended cutoffs derived from hundreds of usability studies as reference means for our *t* tests, rather than relying solely on the midpoint of the Likert scale [31,32].

Study Implications

First, the implications of our findings for policy and practice suggest that digital health tools are both acceptable and useful for addressing gaps in mental health care delivery in low- and middle-income countries. Therefore, these tools, particularly those utilizing simple and widely accessible mobile phones, should be embraced as a viable solution for improving mental health services. Second, using user-centered design and participatory methods is essential to ensure good usability in digital health services. Additionally, assessing both usage and usability is crucial to enhance the scientific evidence on mHealth. This includes understanding the information consumed by users, preferred communication channels, types of users, and any challenges they face. Such assessments will inform targeted improvements in digital health services, ultimately leading to better user experiences and outcomes.

Conclusion

This study found sustained usage of a mental health telephone service in Uganda, with positive user experiences, high satisfaction, and strong intentions for future use across diverse user groups. mHealth solutions like ours should be embraced and replicated to enhance the delivery of mental health services and address other disease areas in Uganda and similar low-resource settings. However, attention should be paid to user experience, and the design and implementation of the system (eg, content and access channels) should align with the preferences and characteristics of different users to ensure good usability and sustained use.

Acknowledgments

This project was funded by Digital Futures, a research collaboration between KTH Royal Institute of Technology, Stockholm University, and RISE Research Institutes of Sweden. The funders had no influence on the conduct of the research study.

Data Availability

The data sets generated during or analyzed in this study are not publicly available because they contain personal data (eg, telephone numbers), but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

JKK contributed to the conception and design of the study, data collection, analysis, and interpretation of results, and the preparation of the draft manuscript. JuNa oversaw the conception and design of the study, along with the analysis and interpretation of results. JaNi, JaNs, and RNamagembe performed data collection, analysis and interpretation of results. IDEG oversaw the conception and design, analysis and interpretation of results, and draft manuscript preparation. RNeumbe, DM, CK, and RNanyonga contributed to data collection. MS and SN were responsible for the analysis and interpretation of results, and draft manuscript preparation. CWL contributed to the conception and design, analysis and interpretation of results, and draft manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Call trends over time.
[PNG File , 21 KB - jmir_v26i1e65692_app1.png]

Multimedia Appendix 2
Calls trends across the time of the day.
[PNG File , 19 KB - jmir_v26i1e65692_app2.png]

Multimedia Appendix 3
Correlations between call conversation aspects and usability components.
[DOCX File , 17 KB - jmir_v26i1e65692_app3.docx]

References

1. Aboye G, Vande Walle M, Simegn G, Aerts J. Current evidence on the use of mHealth approaches in Sub-Saharan Africa: a scoping review. Health Policy and Technology 2023 Dec;12(4):100806-100821. [doi: 10.1016/j.hlpt.2023.100806]
2. Karamagi HC, Muneene D, Droti B, Jepchumba V, Okeibunor JC, Nabyonga J, et al. eHealth or e-Chaos: the use of digital health interventions for health systems strengthening in sub-Saharan Africa over the last 10 years: a scoping review. J Glob Health 2022 Dec 03;12:04090 [FREE Full text] [doi: 10.7189/jogh.12.04090] [Medline: 36462201]
3. Aboye GT, Vande Walle M, Simegn GL, Aerts JM. mHealth in sub-Saharan Africa and Europe: a systematic review comparing the use and availability of mHealth approaches in sub-Saharan Africa and Europe. Digit Health 2023;9:20552076231180972 [FREE Full text] [doi: 10.1177/20552076231180972] [Medline: 37377558]
4. Ødegård ES, Langbråten LS, Lundh A, Linde DS. Two-way text message interventions and healthcare outcomes in Africa: systematic review of randomized trials with meta-analyses on appointment attendance and medicine adherence. PLoS One 2022;17(4):e0266717 [FREE Full text] [doi: 10.1371/journal.pone.0266717] [Medline: 35421134]
5. Osei E, Kuupiel D, Vezi PN, Mashamba-Thompson TP. Mapping evidence of mobile health technologies for disease diagnosis and treatment support by health workers in sub-Saharan Africa: a scoping review. BMC Med Inform Decis Mak 2021 Jan 06;21(1):11 [FREE Full text] [doi: 10.1186/s12911-020-01381-x] [Medline: 33407438]
6. Manby L, Aicken C, Delgrange M, Bailey JV. Effectiveness of eHealth interventions for HIV prevention and management in sub-Saharan Africa: systematic review and meta-analyses. AIDS Behav 2022 Feb;26(2):457-469 [FREE Full text] [doi: 10.1007/s10461-021-03402-w] [Medline: 34427813]
7. Hampshire K, Mwase-Vuma T, Alemu K, Abane A, Munthali A, Awoke T, et al. Informal mhealth at scale in Africa: oOpportunities and challenges. World Dev 2021 Apr;140:105257 [FREE Full text] [doi: 10.1016/j.worlddev.2020.105257] [Medline: 33814676]
8. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002 [FREE Full text] [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
9. Ding X, Wuerth K, Sakakibara B, Schmidt J, Parde N, Holsti L, et al. Understanding mobile health and youth mental health: scoping review. JMIR Mhealth Uhealth 2023 Jun 16;11:e44951 [FREE Full text] [doi: 10.2196/44951] [Medline: 37220197]
10. International Organization for Standardization (ISO). ISO/TS 20282-2:2013(en): Usability of consumer products and products for public use - Part 2: Summative test method. ISO. URL: https://www.iso.org/obp/ui/#iso:std:iso:ts:20282:-2:ed-2:v1:en [accessed 2024-04-28]
11. Lah U, Lewis JR, Šumak B. Perceived usability and the modified technology acceptance model. International Journal of Human–Computer Interaction 2020 Feb 17;36(13):1216-1230. [doi: 10.1080/10447318.2020.1727262]

XSL•FO
RenderX

12. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
13. Connolly SL, Hogan TP, Shimada SL, Miller CJ. Leveraging implementation science to understand factors influencing sustained use of mental health apps: a narrative review. J Technol Behav Sci 2021;6(2):184-196 [FREE Full text] [doi: 10.1007/s41347-020-00165-4] [Medline: 32923580]
14. Madonsela S, Ware L, Scott M, Watermeyer J. The development and use of adolescent mobile mental health (m-mhealth) interventions in low- and middle-income countries: a scoping review. South African Journal of Psychology 2023 Jul 08;53(4):471-483. [doi: 10.1177/00812463231186260]
15. Nicholas A, Joshua O, Elizabeth O. Accessing mental health services in Africa: current state, efforts, challenges and recommendation. Ann Med Surg (Lond) 2022 Sep;81:104421 [FREE Full text] [doi: 10.1016/j.amsu.2022.104421] [Medline: 35996570]
16. Alloh FT, Regmi P, Onche I, Teijlingen EV, Trenoweth S. Mental health in low-and middle income countries (LMICs): going beyond the need for funding. Health Prospect 2018 Jun 19;17(1):12-17. [doi: 10.3126/hprospect.v17i1.20351]
17. Inal Y, Wake JD, Guribye F, Nordgreen T. Usability evaluations of mobile mental health technologies: systematic review. J Med Internet Res 2020 Jan 06;22(1):e15337 [FREE Full text] [doi: 10.2196/15337] [Medline: 31904579]
18. Opio JN, Munn Z, Aromataris E. Prevalence of mental disorders in Uganda: a systematic review and meta-analysis. Psychiatr Q 2022 Mar;93(1):199-226. [doi: 10.1007/s11126-021-09941-8] [Medline: 34427855]
19. Asiimwe R, Nuwagaba-K R, Dwanyen L, Kasujja R. Sociocultural considerations of mental health care and help-seeking in Uganda. SSM - Mental Health 2023 Dec;4:100232-100211 [FREE Full text] [doi: 10.1016/j.ssmmh.2023.100232]
20. Aber-Odonga H, Babirye JN, Engebretsen IMS, Nuwaha F. Prevalence of probable substance use disorders among children in Ugandan health facilities. BMC Public Health 2024 Jan 29;24(1):314 [FREE Full text] [doi: 10.1186/s12889-024-17732-6] [Medline: 38287328]
21. Kaggwa MM, Harms S, Mamun MA. Mental health care in Uganda. The Lancet Psychiatry 2022 Oct;9(10):766-767. [doi: 10.1016/S2215-0366(22)00305-4]
22. Molodynski A, Cusack C, Nixon J. Mental healthcare in Uganda: desperate challenges but real opportunities. BJPsych Int 2017 Nov;14(4):98-100 [FREE Full text] [doi: 10.1192/s2056474000002129] [Medline: 29093962]
23. Kagaari J. Mental health in Uganda. American Psychological Association. 2021 Feb 18. URL: https://www.apa.org/international/global-insights/uganda-mental-health [accessed 2022-05-18]
24. Kabukye JK, Namagembe R, Nakku J, Kiberu V, Sjölinder M, Nilsson S, et al. Implementing a hospital call center service for mental health in Uganda: user-centered design approach. JMIR Hum Factors 2024 Jun 06;11:e53976 [FREE Full text] [doi: 10.2196/53976] [Medline: 38843515]
25. Brender J, Talmon J, de Keizer N, Nykänen P, Rigby M, Ammenwerth E. STARE-HI - Statement on Reporting of Evaluation Studies in Health Informatics: explanation and elaboration. Appl Clin Inform 2013;4(3):331-358 [FREE Full text] [doi: 10.4338/ACI-2013-04-RA-0024] [Medline: 24155788]
26. Strack F. “Order effects” in survey research: activation and information functions of preceding questions. In: Schwarz N, Sudman S, editors. Context Effects in Social and Psychological Research. New York City, NY: Springer-Verlag; 1992:23-34.
27. Park J, Jung M. A note on determination of sample size for a Likert scale. Communications for Statistical Applications and Methods 2009 Jul 31;16(4):669-673 [FREE Full text] [doi: 10.5351/CKSS.2009.16.4.669]
28. Parmanto B, Lewis AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). Int J Telerehabil 2016;8(1):3-10 [FREE Full text] [doi: 10.5195/ijt.2016.6196] [Medline: 27563386]
29. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
30. Gustafsson S, Wahlberg A. The telephone nursing dialogue process: an integrative review. BMC Nurs 2023 Sep 28;22(1):345 [FREE Full text] [doi: 10.1186/s12912-023-01509-0] [Medline: 37770869]
31. Brooke J. SUS: a retrospective. J Usability Studies 2013;8(2):29-40 [FREE Full text]
32. Nielsen J, Levy J. Measuring usability. Commun ACM 1994 Apr;37(4):66-75. [doi: 10.1145/175276.175282]
33. Kabukye JK, Ilozumba O, Broerse JEW, de Keizer N, Cornet R. Implementation of an interactive voice response system for cancer awareness in Uganda: mixed methods study. JMIR Mhealth Uhealth 2021 Jan 26;9(1):e22061 [FREE Full text] [doi: 10.2196/22061] [Medline: 33496672]
34. Abbo C, Ekblad S, Waako P, Okello E, Musisi S. The prevalence and severity of mental illnesses handled by traditional healers in two districts in Uganda. Afr Health Sci 2009 Aug 01;9 Suppl 1(Suppl 1):S16-S22 [FREE Full text] [Medline: 20589155]
35. Mwesiga EK, Nakasujja N, Nakku J, Nanyonga A, Gumikiriza JL, Bangirana P, et al. One year prevalence of psychotic disorders among first treatment contact patients at the National Psychiatric Referral and Teaching Hospital in Uganda. PLoS One 2020;15(1):e0218843 [FREE Full text] [doi: 10.1371/journal.pone.0218843] [Medline: 31995567]
36. Verity F, Turiho A, Mutamba BB, Cappo D. Family care for persons with severe mental illness: experiences and perspectives of caregivers in Uganda. Int J Ment Health Syst 2021 May 20;15(1):48 [FREE Full text] [doi: 10.1186/s13033-021-00470-2] [Medline: 34016125]

XSL•FO
RenderX

37. Lee H, Palmer J, Mullick P, Kiely C. Patient experience with telehealth service in a mental health setting. Arch Psychiatr Nurs 2023 Apr;43:150-152 [FREE Full text] [doi: 10.1016/j.apnu.2023.03.001] [Medline: 37032009]
38. Park HY, Kwon YM, Jun HR, Jung SE, Kwon SY. Satisfaction survey of patients and medical staff for telephone-based telemedicine during hospital closing due to COVID-19 transmission. Telemed J E Health 2021 Jul;27(7):724-732 [FREE Full text] [doi: 10.1089/tmj.2020.0369] [Medline: 33216710]
39. Naik K, Allen L, Gnanavel M. Patient satisfaction with telephone consultations at a community mental health service: a descriptive survey. British Journal of Healthcare Management 2022 Nov 02;28(11):284-291 [FREE Full text] [doi: 10.12968/bjhc.2021.0115]
40. Maresca G, Anchesi S, Bonanno L, Bramanti A, Carnazza L, Cimino V, et al. Feasibility, usability, and customer satisfaction of the Tele-COVID19 project, Sicilian model. Medicina (Kaunas) 2022 Aug 16;58(8):1110. [doi: 10.3390/medicina58081110] [Medline: 36013577]
41. Lodge C. The impact of culture on usability: designing usable products for the international user. In: Aykin N, editor. Usability and Internationalization. Berlin, Germany: Springer-Verlag; 2007:365.
42. Ford G, Kotzé P. Designing usable interfaces with cultural dimensions. In: Costabile M, Paternò F, editors. Human-Computer Interaction - INTERACT 2005. Berlin, Germany: Springer; 2005:713-726.
43. Heller LD. Palliative care in sub-Saharan Africa: a narrative review. J Palliat Med 2024 Aug;27(8):1083-1089. [doi: 10.1089/jpm.2023.0501] [Medline: 38466990]
44. Camara BS, Belaid L, Manet H, Kolie D, Guillard E, Bigirimana T, et al. What do we know about patient-provider interactions in sub-Saharan Africa? a scoping review. Pan Afr Med J 2020;37:88 [FREE Full text] [doi: 10.11604/pamj.2020.37.88.24009] [Medline: 33244351]
45. Norman I. Blind trust in the care-giver: is paternalism essential to the health-seeking behavior of patients in sub-Saharan Africa? AASoci 2015;05(02):94-104. [doi: 10.4236/aasoci.2015.52008]

Abbreviations

IVR: interactive voice response
mHealth: mobile health
STARE-HI: Statement on Reporting of Evaluation Studies in Health Informatics
TUQ: Telehealth Usability Questionnaire

Edited by A Mavragani; submitted 22.08.24; peer-reviewed by K Elechi, D Singh; comments to author 02.09.24; revised version received 08.09.24; accepted 24.09.24; published 21.10.24.

Please cite as:
Kabukye JK, Nakku J, Niwemuhwezi J, Nsereko J, Namagembe R, Groen IDE, Neumbe R, Mubiru D, Kisakye C, Nanyonga R, Sjölinder M, Nilsson S, Wamala-Larsson C
Assessing the Usage and Usability of a Mental Health Advice Telephone Service in Uganda: Mixed Methods Study
J Med Internet Res 2024;26:e65692
URL: https://www.jmir.org/2024/1/e65692
doi:10.2196/65692
PMID:

©Johnblack K Kabukye, Juliet Nakku, Jackline Niwemuhwezi, James Nsereko, Rosemary Namagembe, Iris Dorothee Emilie Groen, Ritah Neumbe, Denis Mubiru, Caroline Kisakye, Roseline Nanyonga, Marie Sjölinder, Susanne Nilsson, Caroline Wamala-Larsson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Perspectives on Blended Internet-Based and Face-to-Face Cognitive Behavioral Therapy for Alcohol Use Disorder: Qualitative Study

Kristine Tarp[1,2,3], MA, PhD; Regina Christiansen[4,5], MA, PhD; Randi Bilberg[4,5,6], MSc, PhD; Simone Borkner[2], BSc; Caroline Dalsgaard[2], MSc; Marie Paldam Folker[2], MA; Anette Søgaard Nielsen[4,7,8], MA, PhD

[1]Research Unit for Digital Psychiatry, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[2]Centre for Digital Psychiatry, Mental Health Services, Region of Southern Denmark, Odense, Denmark
[3]The National Research Centre for the Working Environment, Copenhagen, Denmark
[4]Unit of Clinical Alcohol Research, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[5]Psychiatric University Hospital, University Function, Region of Southern Denmark, Odense, Denmark
[6]Department for finance and planning, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle, Denmark
[7]Brain Research, Inter-Disciplinary Guided Excellence, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[8]OPEN, Open Patient data Explorative Network, Odense University Hospital, Odense, Denmark

Corresponding Author:
Regina Christiansen, MA, PhD
Psychiatric University Hospital
University Function
Region of Southern Denmark
J.B. Winsløws Vej 18
Odense, 5000
Denmark
Phone: 45 6550 7221
Email: reginachristiansen@health.sdu.dk

Abstract

Background: Harmful alcohol consumption has been identified as a major contributor to disease, mortality, and social harm, accounting for 5.3% of worldwide deaths annually. In Denmark, an estimated 150,000 people suffer from alcohol use disorder (AUD), but a low proportion seek treatment due to person- and treatment-related barriers. Internet-based cognitive behavioral therapy (iCBT) has shown positive effects on the treatment gap, with patients reporting benefits such as increased knowledge and flexibility. However, there is a lack of research on blended cognitive behavioral therapy (bCBT), which combines face-to-face CBT (FtF CBT) and iCBT for AUD.

Objective: This study aims to investigate user experiences of bCBT. More specifically, it seeks to explore the advantages and disadvantages that users have experienced with bCBT for AUD, as well as their motivations for choosing this treatment format.

Methods: A total of 30 patients who had participated in the Blend-A (Blending Internet Treatment into Conventional Face-to-Face Treatment for AUD) study and received the intervention were contacted and offered the opportunity to participate in semistructured individual telephone interviews. Of these, 12 patients consented to participate. Furthermore, an additional participant was approached at a municipal clinic and agreed to engage in an individual FtF interview. Thus, the final sample consisted of 13 patients. The interviews explored their background, experiences with digital technology, motivations for choosing internet-based treatment, and experiences with the program during AUD treatment. The interviews were audio-recorded and transcribed in full length and analyzed using thematic analysis. All data were anonymized and securely stored.

Results: We found that users experienced several advantages of iCBT over a larger part of the treatment course, including increased anonymity and privacy. Most importantly, it offered flexibility, allowing patients to focus on their rehabilitation process at their own pace. Patients appreciated the availability of written text in the online program, finding it helpful for gaining knowledge and understanding of AUD and its impact on the individual with the condition. They emphasized how the assignments helped them fully engage in treatment by first acknowledging their problem with alcohol and then dedicating time to self-reflection

XSL•FO
RenderX

before FtF sessions, allowing for more in-depth discussions with the therapist. They also appreciated the reminders, which motivated them to complete their assignments.

Conclusions: Overall, patients perceived more benefits than disadvantages in using bCBT. Essentially, bCBT offers a form of assisted autonomy that cannot be fully achieved through iCBT or FtF CBT alone. It is only through their combination that patients can fully appreciate the benefits of the treatment, as they have time for self-reflection, with guidance from the therapist between FtF CBT sessions.

Trial Registration: ClinicalTrials.gov NCT04535258; https://clinicaltrials.gov/ct2/show/NCT04535258

(J Med Internet Res 2024;26:e47083) doi:10.2196/47083

KEYWORDS

internet-based; alcohol use disorder treatment; user perspective; qualitative; blended treatment; blended learning; cognitive behavioral therapy; alcoholism; alcohol use disorder; treatment; barriers; patient perspectives; rehabilitation

Introduction

Background

Unhealthy alcohol consumption represents a significant public health concern globally, contributing to increased risks of disease, mortality, and social harm [1]. It is estimated that 100 million people worldwide, or 1.3% of the total adult population, suffer from an alcohol use disorder (AUD) [2]. In Denmark, approximately 150,000 adults are estimated to be dependent drinkers, accounting for about 3% of the total adult population [3]. Despite the availability of free and easily accessible treatment for AUD in Denmark, the proportion of Danes seeking treatment remains low. This phenomenon is believed to stem from various factors, including a lack of awareness about available treatment options, insufficient perceived need for intervention, and concerns regarding the social stigma associated with seeking help for AUDs [4].

To address this treatment gap, internet-based modalities such as internet-based cognitive behavioral therapy (iCBT) have emerged as promising alternatives, effective in reducing alcohol consumption [1]. Patients who have participated in iCBT for mental health issues, including AUD, have reported that the treatment was helpful. For example, in a user experience study on internet-based treatment for problematic alcohol use, Ekström and Johansson [5] found that various components of iCBT, such as therapist support and discussion forums, were perceived as helpful and useful by patients, particularly when they could relate to the content, which in turn facilitated reflection on their alcohol consumption. In a qualitative study on user motivations and experiences with engagement in an internet intervention for alcohol reduction, Black et al [1] found that evidence-based behavior change techniques, such as social support, normative strategies, goal setting, and self-monitoring, were acceptable to patients. Similarly, in a qualitative study on experiences with iCBT for depression and anxiety, Lindegaard et al [6] found that patients experienced varying levels of success in applying the new information they had learned from treatment to their everyday lives.

Furthermore, patients reported that iCBT offered greater flexibility and accessibility compared with face-to-face (FtF) treatment, making it easier to adapt to changing routines for self-help tools and monitoring. For example, in the user experience study by Ekström and Johansson [5], availability was identified as an important factor for patients when choosing iCBT. In a review of women's expectations and experiences regarding eHealth treatment, Verhoeks et al [7] found that eHealth lowered the threshold for women seeking health care, as barriers to eHealth treatment, such as time constraints, were less significant compared with those for FtF treatment.

Moreover, iCBT may provide opportunities for discretion and anonymity. For instance, in a qualitative study on individuals seeking online help to reduce their drinking, Khadjesari et al [8] found that patients viewed the privacy of the internet as an important factor when searching for assistance, as online help allowed them to avoid the stigma and embarrassment often associated with seeking help in person. Similarly, Ekström and Johansson [5] identified anonymity as a crucial factor for patients when choosing internet-based support, noting that feelings of shame were among the reasons patients decided to reduce their drinking. Verhoeks et al [7] reported that barriers to eHealth treatment, such as feelings of shame, were less significant compared with those associated with FtF treatment. Results from Lindegaard et al [6] also indicated that iCBT could, to some extent, mitigate mental health stigma.

In addition, iCBT addressing mental health issues may be particularly useful for individuals who do not perceive their problems as serious enough for traditional treatment, as they can skip irrelevant modules in iCBT. For instance, the study by Black et al [1] found it possible to incorporate evidence-based behavior change techniques, such as social support, normative strategies, goal setting, and self-monitoring, into iCBT to engage patients who might otherwise delay or refuse help-seeking. Results from Lindegaard et al [6] also indicate that participation in iCBT can serve as a precursor to seeking other forms of treatment.

However, iCBT also presents challenges, particularly in ensuring patient adherence to the program. Verhoeks et al [7] found that reduced feelings of obligation and a lack of motivation were the greatest challenges women faced in completing eHealth treatment. The women expressed a desire for more support during treatment, preferably through blended care. To address this, a blended approach that combines FtF CBT with iCBT—known as blended cognitive behavioral therapy (bCBT)—has been proposed [9]. This blended approach aims to capitalize on the advantages of both formats, potentially enhancing treatment accessibility, reducing stigma, and

XSL•FO
RenderX

providing a more tailored treatment experience. While FtF CBT offers the benefits of personal interaction, immediate feedback, and tailored adjustments, iCBT allows for deeper reflection and exploration [10]. Combining these formats into bCBT may therefore present a meaningful approach.

Despite the theoretical benefits associated with bCBT, the literature currently lacks empirical investigations into its effectiveness in the treatment of AUD. To the authors' knowledge, this specific area of research remains unexplored. This study aims to address this gap by providing qualitative insights from the Blend-A Study (*Blending Internet Treatment into Conventional Face-to-Face Treatment for AUD*), which investigates the experiences of patients engaged in bCBT for AUD.

Aim

This study aims to uncover patients' perceptions and experiences related to the use of bCBT. Based on interviews with patients participating in the Blend-A Study, we present the findings in 2 categories: (1) patient experiences regarding bCBT as a treatment form and (2) patient experiences with the practical use of the platform.

Methods

Design

This study used a qualitative design based on semistructured interviews with patients participating in bCBT for AUD. We followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) [11] checklist for reporting our findings.

Setting

In Denmark, AUD treatment is provided through municipalities at no cost to patients. The treatment primarily consists of evidence-based methods such as CBT and motivational interviewing [12]. Conventional AUD treatment is often delivered through individual FtF sessions or group therapy, with the duration of treatment varying between 4 and 6 months, depending on the patient's goals and challenges. The staff at treatment centers typically includes nurses, psychologists, pedagogues, and social workers who specialize in AUD treatment and are referred to as therapists.

The implementation of the Blend-A Study commenced in June 2020 and continued until the end of December 2022. A total of 18 Danish municipal treatment institutions participated; however, 4 dropped out during the study due to resource challenges, which hindered their ability to dedicate time and focus on project meetings, patient recruitment, and the implementation of the blended format.

The Blend-A treatment program protocol originated from the Netherlands [13] and was translated and adjusted to better align with the Danish language and culture [14]. The Blend-A intervention is a form of bCBT, grounded in CBT and motivational interviewing. This blended format allows for the integration and substitution of FtF sessions within a treatment course with internet-based modules, which are available on a web-based treatment platform. The frequency of FtF sessions and the balance between FtF sessions and the use of internet-based modules were agreed upon by the patient and therapist. Typically, patients had a few FtF sessions before engaging with the internet-based modules. Patients often met FtF every 2-4 weeks, working on the modules on the platform in between. However, some patients utilized the platform almost exclusively, attending only 1 or 2 initial FtF sessions and possibly a later follow-up session, if any.

Based on experiences from the pilot phase, the internet-based platform was hosted by the Dutch company Minddistrict. It was accessible online via any web browser and offered a flexible format, allowing patients to access the platform according to their individual daily routines. Patients could also access the platform anonymously, if preferred. The platform included 21 online sessions, maintenance modules, and diaries, which the therapist could gradually add to the patient's individual account based on their treatment goals and challenges. The first 12 sessions were considered core sessions and were offered to all patients engaged in bCBT, while the subsequent sessions were optional and tailored to the patient's needs. Each session followed a fixed structure, starting with information about the background of the specific session, followed by multiple exercises and assignments. The online sessions included textboxes and videos. In most sessions, patients received online feedback from their therapist on the assignments they completed, which therapists aimed to provide once a week at a set time. After receiving feedback on their assignments, the next session would be made available to the patient by the therapist. Upon completing treatment, patients continued to have access to the online treatment platform, allowing them to re-read information and review previously completed exercises. An overview of the platform sessions can be seen in Table 1.

XSL•FO
RenderX

Table 1. Platform sessions.

Session	Content	Activity
Session 1	Welcome to the Blend-A[a] online treatment	N/A[b]
Session 2	Support from your social network	N/A
Session 3	Explanation of treatment course	N/A
Session 4	Preparation to change	Feedback moment
Session 5	Goals and techniques for self-control	Feedback moment
Session 6	List of alcohol use risk situations	N/A
Session 7	Alcohol use function analysis	Feedback moment
Session 8	Alcohol use emergency plan	N/A
Session 9	Tackling craving	Feedback moment
Session 10	Restructuring thoughts	Feedback moment
Session 11	Rejecting alcohol offers	Feedback moment
Session 12	Evaluation of the Blend-A main module	N/A
Session 13	Selection of themes	Feedback moment
Session 14	Theme: Social skills—small talk	Feedback moment
Session 15	Theme: Social skills—tackling criticism	Feedback moment
Session 16	Theme: Social skills—giving criticism	Feedback moment
Session 17	Theme: Tackling feeling sad and moody	Feedback moment
Session 18	Theme: Tackling stress	Feedback moment
Session 19	Theme: Solving problems effectively	Feedback moment
Session 20	Theme: Tackling relapse	Feedback moment
Session 21	Introduction to maintenance	N/A
Maintenance month 1	Alcohol status and Quality of life	N/A
Maintenance month 2	Alcohol status and Quality of life—your assessment	N/A
Maintenance month 3	Alcohol status, Quality of life—your assessment, and Support	N/A
Maintenance month 4	Alcohol status, Quality of life—your assessment, and Support—your assessment	N/A
Maintenance month 5	Alcohol status, Quality of life—your assessment, Support—your assessment, And Motivation	N/A
Maintenance month 6	Alcohol status, Quality of life—your assessment, Support—your assessment, Motivation—your assessment, and Evaluation of the maintenance phase	N/A
Diary	Alcohol use	N/A
Diary	Evaluation of turning alcohol offers down	N/A

[a]Blend-A: Blending Internet Treatment into Conventional Face-to-Face Treatment for AUD.

[b]N/A: not applicable.

Recruitment and Data Collection

To explore patients' experiences with bCBT, we decided to conduct a qualitative study involving the 30 most recently enrolled patients in the Blend-A study who utilized the option of combining FtF sessions with internet-based sessions. We invited them to participate in either telephone-based interviews or FtF interviews to share their experiences. Of these 30 patients, 13 agreed to participate, while the others opted out or could not be reached.

Data were collected through semistructured individual interviews using an interview guide (Multimedia Appendix 1). The guide was developed based on themes identified as important in earlier studies [5,8,15-17]. The questions were open-ended, asking patients about their background, experiences with digital technology in their daily lives, motivations for choosing the offered internet-based treatment, and their experiences with the program during their AUD treatment. The guide was not pilot-tested. The interviews were conducted by the first author (KT), a female anthropologist and PhD, who was a postdoc at the time of the study with approximately 10 years of training in qualitative research. No prior relationship was established between KT and the patients before the interviews. As a result, the patients had no prior knowledge of KT and were informed that she was a researcher not connected to the treatment institutions. The interviews lasted between 9

XSL•FO
RenderX

and 43 minutes, with a mean duration of 30 minutes. No one was present at the interviews besides the patient and the researcher; no repeat interviews were conducted, and no field notes were taken. The interviews were audio-recorded, transcribed in full in NVivo (Lumivero), and checked for accuracy. Transcripts were not presented to the patients for feedback. Data saturation was discussed within the research group. All data were anonymized and securely stored.

Data Analysis

The transcribed interviews with patients were qualitatively analyzed using the thematic analysis approach [18]. Thematic analysis is a method for identifying and describing patterns or themes within data, without being constrained by a preexisting theoretical framework. In this study, thematic analysis was used as a realist or essentialist method, focusing on reporting the experiences of the patients. Patterns and themes within the data were identified inductively and expressed in multiple ways. First, RC and KT read the transcripts multiple times independently and coded the interviews using headings that captured the essence of specific excerpts. Second, the headings and the content under each heading were discussed between the 2 coders. In cases of disagreement, either the heading was revised to better reflect the content or the content was moved to a different heading that more accurately defined it. Lastly, RC collated the categories with related content into overall themes using the qualitative software system NVivo. The themes were then refined by KT. Overall, the identified themes represented the participants' first-person perspectives on bCBT, contributing new knowledge on aspects important to treatment. The final analysis was written with relevant quotations included, but no patients provided feedback on the findings.

Ethics Considerations

The study was conducted according to current ethical standards. The protocol was approved by the Scientific Research Ethics Committee of the Region of Southern Denmark (project identification number S-20190166G). The Danish Data Protection Agency gave permission to collect and store data (record number 20/12692). After receiving oral and written information about the study, the patients signed a consent form.

Results

Sample Description and Thematic Analysis of Patient Experiences

A description of the sample of participants is presented in Table 2. The forthcoming analysis of the transcribed interviews is structured as follows: (1) patient experiences of how bCBT impacts treatment, comprising 4 themes; and (2) patient experiences with the practical use of the platform, also comprising 4 themes. An overview of the identified categories, themes, and subthemes is available for reference in Table 3 and will be described in detail in the following sections.

Table 2. Participant sample description (N=13).

Characteristics	Values[a]
Sociodemographic	
Age (years)	53 (31-75)
Sex (female)	4 (31)
Married/in a relationship (yes)	9 (69)
Short, intermediate, or long education (yes)	13 (100)
Employment (yes)	7 (54)
On leave/early retired/retired (yes)	6 (46)
Alcohol use	
Excessive alcohol use (yes)	13 (100)
Using alcohol every day (yes)	5 (38)
Using alcohol for self-medication, for example, anxiety, depression, pain, sleeping (yes)	4 (31)
Had sought treatment to reduce or stop alcohol use (yes)	13 (100)
Technology use	
Familiar technology user (yes)	13 (100)
Feeling safe using the internet for treatment purposes (yes)	13 (100)
Had completed the platform sessions at the time of the interview (yes)	7 (54)
Had used the blended format (yes, as opposed to primarily using the platform)	10 (77)

[a]Presented as mean (range) or N/n (%).

XSL•FO
RenderX

Table 3. Identified categories, themes, and subthemes (N=13).

Categories and themes	Frequencies for naming the themes, n (%)	Subthemes
Patient experiences of how bCBT[a] impacts treatment		
bCBT adds anonymity and privacy to the treatment course	6 (46)	• bCBT helps when I find it transgressive to go to the clinic in person • bCBT helps when I fear running into someone I know at the clinic
bCBT adds flexibility to the treatment course	8 (62)	• bCBT fits well when my transportation options are limited • bCBT fits well with my work life • bCBT fits well with my home life
bCBT enables time and space for reflection	4 (31)	• I can think about it at my own pace when I use the platform • I have thought about it, when I attend a session, leaving room for new angles • I will continue to think about it after the session, qualifying it
bCBT increases an assisted form of autonomy	N/A[b]	• I have to take responsibility for my own rehabilitation process • I want to go through the process on my own, with support from the therapist • Constructive feedback leads to further progress in my change process • The interaction between the 2 formats leaves room for deeper discussions
Patient experiences with the practical use of the platform		
User-friendliness	9 (69)	• The internet-based treatment program is user-friendly, intuitive, and easy to use • I appreciate how the program is tailored in accordance with the feedback given by the therapist
Written material and assignments	6 (46)	• The online material is well-written and usable for acquiring knowledge and facts about drinking • Reading the material helps me gain a greater reflection on my drinking habits • The material can be a good starting point for a conversation about drinking
Diary and registrations	8 (62)	• I used the diary for self-reflection, taking the time to think and evaluate my thoughts and feelings during the rehabilitation process • It helped me to gain insights into my use of alcohol, habits leading to cravings, and how to deal with cravings • This facilitated an internal dialog about whether the process was progressing as I hoped for
Motivating reminders	5 (38)	• Reminders are helpful and well-intended • There is a person behind the platform who expects my attendance at the program • This adds to my motivation and progress in the treatment course

[a]bCBT: blended cognitive behavioral therapy.

[b]N/A: not applicable.

Patient Experiences of How bCBT Impacts Treatment

Key Themes in bCBT Patient Experiences

This category includes 4 themes: how the possibility of bCBT added anonymity, privacy, and flexibility to the treatment course; how bCBT provided time and space for reflection for the patients; and how bCBT fostered a supported form of autonomy for the patients. Each of the 4 themes is presented in detail below.

bCBT Adds Anonymity and Privacy to the Treatment Course

Some patients felt that the relatively private or anonymous nature of bCBT influenced their decision to choose this particular treatment format.

One factor that influenced the decision to engage in bCBT was its ability to reduce the uncomfortable experience of entering a clinic:

> *To me, it is very transgressive to sit in a waiting room in the clinic. So, I think it is nice that some of it is*

XSL•FO
RenderX

> *private. That is primarily why I wanted online* [Patient, ID 1273]

Another factor that made bCBT attractive was that it minimized the risk of encountering someone the patient knew at the clinic:

> *During this time where I have been attending the clinic, I have been offered group therapy and group meetings. And I have plainly refused it, terrified of meeting someone I know* [Patient, ID 1268]

For them, entering a treatment center required a significant amount of courage; staying at home and avoiding frequent visits were perceived as less anxiety-provoking.

bCBT Adds Flexibility to the Treatment Course

The majority of patients indicated that the flexible nature of bCBT motivated them to actively seek out or engage in the treatment. They found that the ability to participate from home made it easier to incorporate the treatment into their everyday lives. One important factor influencing this was limited transportation options, as illustrated in the following example:

> *You do not have to involve others in the fact that you are leaving for treatment. And I will say that, especially when you live in a thinner populated area as we do. [...] There are not so many options. And you must drive for everything. Just that you don't need to drive for it. That makes a big difference* [Patient, ID 1114]

Consistent with the previous example, patients found bCBT easier to integrate into their work lives. They expressed that adding weekly FtF CBT to their already busy schedules would be difficult, if not impossible. The option to engage in bCBT alleviated these obstacles, as explained by one of the patients:

> *I was also at work and had to go in there and I did not want it to become too stressful for me. Then I think that it is fine and with the simultaneously sessions in person sometimes. I have been happy about this. [...] also, because I think it could stress me a little if I had to go in there all the time with work and the late hours [...] I think it was good because it took away the stress* [Patient, ID 1256]

Thus, the availability of bCBT offered patients a pressure-reducing model for receiving treatment, either partly or completely. They found bCBT convenient because it allowed them to engage with the treatment course at times that suited their home life. This was elaborated on by one of the patients:

> *And you can control it in your own tempo. That is also probably why I also wanted something online. [...] Every other week, I spend a lot of time with my partner. And it is suitable that the weeks where I am not then I can focus more on this.* [Patient, ID 1273]

The patients overall felt that bCBT provided a flexibility that aligned well with their daily lives. The ability to engage with the online format at times that suited each individual was viewed as a significant advantage to the treatment.

bCBT Enables Time and Space for Reflection

About one-third of the patients explained that they chose to receive bCBT because it was more likely to support them at their own pace in the rehabilitation process compared with FtF sessions alone. Working with the program at home allowed for a more thoughtful process, facilitating deeper realizations than those possible in a FtF CBT session with limited time. One patient elaborated on how using the platform at home provided uninterrupted time for reflection, leading to more comprehensive responses:

> *At home you are able go to and from. Understood in that way that you can answer more nuanced because you don't have to do it right here and right now. But you can spend some time reflecting on the questions, your thoughts, and reaction patterns* [Patient, ID 1184]

The patients were committed to the idea of completing assignments at home in a peaceful and quiet environment, allowing them to reflect on their answers as they worked through the tasks. This reflection time and level of nuance made the treatment course feel more thorough—and even more effective—for them.

Second, another factor that made patients find bCBT more effective than they had initially imagined FtF CBT alone would be was the reflection time, which enriched their dialog with the therapist during FtF sessions. One patient explained:

> *I think having the background information and assignments to do in the program and think about and write down what it meant to you that is something that can make you think before you show up for the session and then you can unfold your thoughts at the session better than if you first were to start doing the thinking when you are in the room* [Patient, ID 1350]

The patients felt that this made the FtF sessions more fruitful, as the reflection time between completing assignments and meeting with the therapist allowed space for new thoughts and perspectives. They experienced that this reflection time made the FtF sessions more intense and effective, leading them to emphasize that it had the potential to enhance the effectiveness of FtF CBT.

A third and final factor influencing the patients' perceptions of bCBT as being more thorough than solely FtF therapy was that the reflections continued even after the sessions. This was explained by one patient:

> *It had two functions. One was the immediate when I was working with the assignments where I was thinking quite some about the things in the assignment. And then, the other thing was, in the days after it kept being there in my head and I could continue working and think and use some of the things that I had been working with* [Patient, ID 1316]

The topics discussed on the platform and by the therapist during FtF CBT remained part of the patients' reflections throughout their rehabilitation process. In this way, bCBT facilitated

XSL•FO
RenderX

increased self-reflection during the treatment course—before, during, and after the FtF sessions.

bCBT Increases an Assisted Form of Autonomy

The patients felt that the combination of platform use and FtF sessions enhanced their treatment experience. One factor influencing this was the realization that they needed to take their alcohol-related problems seriously and assume responsibility for their own rehabilitation. This was explained by one patient:

> *First and foremost, it has been an important factor in making me more conscious about the fact that I had a huge problem with alcohol that I had to do something about. And for the first time in my life, I was deeply aware about that it was the most important thing I could do for myself. And it has contributed into making that clear for me somehow. Because the themes and the subjects that were raised were relevant and it was done in a proper manner that I could understand and relate to* [Patient, ID 1284]

The patients further explained that personal motivation is essential for recognizing the need for change. Another important aspect closely connected to their sense of autonomy was the opportunity to engage with the program in private. One patient elaborated:

> *I wanted it to myself. It was moments that I wanted to myself. And I did not want to explain why I went down here. Or I could but I did not need all the questions. I wanted it to myself together with the therapists. It was my little world* [Patient, ID 1345]

The patients distinguished between the rational decision to seek treatment and the deeper, personal realization of needing to make a significant change. For them, it was much more than simply concluding that entering treatment was the healthiest or most appropriate choice. Once they decided to pursue treatment, they emphasized the importance of taking control over their lives without external interference—an autonomy that bCBT fosters and strengthens.

The autonomy offered by bCBT was, however, a supported form, endorsed by the therapist. Patients described receiving constructive online feedback from their therapist as a key element in advancing their change process. They found the feedback to be an acknowledgment of their efforts, often including reflections on their completed assignments, along with additional suggestions for new, motivating challenges. One patient elaborated:

> *It had two functions. One was to point at things that I should be attentive to or work more with. And at the same time, it also gave me the feeling that someone else than me could see that there was a meaningful process in what I was doing and that was fine* [Patient, ID 1316]

A final note on how the combination of the 2 intervention formats assisted the rehabilitation process, as experienced by the patients, was that completing some sessions on the platform freed up time during FtF sessions to focus on deeper issues that required the therapist's in-person perspective. One patient elaborated:

> *After I started using the app there were some things that the therapist did not have to spend time explaining to me, which they otherwise would have done in the sessions. But when you only have one hour every other week then it is nice that you don't have to spend so much time explaining. Then I think it is nice that you can just dive into how it works for you, personally* [Patient, ID 1350]

The patients found that being able to read background information before the FtF CBT session with the therapist saved time for deeper conversations during the session. As a result, they felt that the combination of completing assignments at home and attending FtF CBT supported them more than they had anticipated from solely participating in FtF sessions.

Patient Experiences With the Practical Use of the Platform

Patient Engagement and Usability in Treatment: Key Themes

This category includes 4 themes that describe user-friendliness and patient engagement with written materials, assignments, diaries, registrations, and motivating reminders. Each of the 4 themes is presented in detail below.

User-Friendliness

Almost all patients found the internet-based treatment program to be user-friendly. Upon introduction to the program, they found it manageable, intuitive, and easy to use. They appreciated how the program was tailored based on the feedback provided by the therapist. One patient explained the usability as follows:

> *I answered the assignments and sent them to the therapist for feedback on the scheduled day for her to give that. That means I followed the rhythm in the program and that worked well.* [Patient, ID 1316]

The setup of the reading material and assignments were important, as was the digital component, the patient added:

> *It has worked very well for me because it is easy access, and I can do it regardless time and space. I would also be able to do that with a piece of paper but that would be heavier. It would certainly be a much heavier process to work with it on paper. It has also been easy for me sometimes to go back in the process and see what I was thinking at that time. What was it that I was talking to myself about at that time. And then I could go forward again. Very simple* [Patient, ID 1316]

Having the ability to navigate back and forth between completed assignments and new ones was found to be helpful in initiating their rehabilitation process.

Visual aids were also mentioned; 1 patient appreciated the animated films for providing an informal yet understandable overview of the problem area. Another found the videos too polished, while a different patient found the tools presented in the videos to be very helpful. When encountering technical

XSL•FO
RenderX

challenges, one suggested asking the therapist for assistance to improve navigation within the program. We also inquired whether the patients had missed anything on the platform or had suggestions for adjustments for future iterations. Their needs and wishes for the platform can be seen in Textbox 1.

Textbox 1. Needs and wishes for the further development of the platform.

- Tempering expectations may be helpful before starting an online treatment course.
- The examples in the assignments should be more diverse, addressing a wider range of patient groups.
- Include a videoconferencing option in the program, allowing patients to choose video sessions with the therapist instead of FtF sessions.
- It should be more intuitive to navigate the treatment program and easily go back and forth to find relevant sections to read, watch, or fill out again.
- Being able to reply directly to the therapist's feedback, rather than having to copy it into a separate message, would be helpful.
- Having the written material reviewed by a communication or rhetorical specialist who is not otherwise connected to the field would be beneficial.
- It would have been more relevant for some if the program better reflected that patients could also be in control, highlighting the potential for a process stemming from that.
- A module focused on pain relief.
- Free navigation and log-in within the program.
- More challenging assignments, examples, and questions would have enhanced the program's relevance.
- A timeline showing your progress in the program would provide a better overview.
- An individualized program where different answers lead patients to tailored treatment paths based on their unique problems and needs.
- Would have preferred the entire program, not just the diary and registrations, to be smartphone app-based.
- Would have liked the option to access additional information and assignments beyond the released session, as there could be a significant wait for feedback and the next session.
- A function that allows you to upload pictures of written materials.
- A clearer presentation of the self-selected themes.
- That the program involved families, provided explanations about alcohol use, and included monitoring of their observations while addressing the challenges they might encounter.
- The program could do more to celebrate victories and motivate users by incorporating visual positive reinforcements.

Written Material and Assignments

Overall, the patients found that the online material was well-written and effective for acquiring knowledge about how their drinking problem affected them. Reading the material helped them gain greater insights into their drinking habits, providing information they did not have beforehand. They also noted that the written material was occasionally patronizing in its choice of wording, emphasizing that this could potentially have a negative effect on patients. One final observation was that the reading material could serve as a good starting point for conversations with those close to the individual struggling with alcohol use.

The patients found that the assignments and tasks were relevant and aligned with their own experiences regarding problematic drinking. However, one patient felt that some tasks were somewhat superficial and requested a more challenging and pertinent program. Many patients noted that if the assignments lacked meaningful content related to their everyday lives and specific challenges, they found it difficult to answer the questions. They emphasized the importance of including a broad range of questions and examples in the assignments. One patient elaborated on this point:

> *It is difficult to identify with the questions and then the answers also become irrelevant. And I think that the learning that was intended with questions and answers it does not appear* [Patient, ID 1273]

The patients emphasized that it can be challenging to formulate questions that resonate with every individual. To overcome this obstacle, some patients suggested a practical solution: the program could offer options for patients to select assignments that are relevant to them while allowing them to skip those that are not. One patient elaborated:

> *Because it is super complex why you do what you do, which factors that is a starting point for a misuse. And I don't think that that complexity can be reflected in such a very linear progress. I'm sure that there are good considerations behind, but I think that if it needs to be relevant for both the well-functioning drinker and the one who are not in control at all, I think you need to differentiate it* [Patient, ID 1273]

The patient highlighted the complexity inherent in being an individual with a drinking problem, noting that this makes it challenging to organize assignments in a way that aligns with each person's unique experiences.

XSL•FO
RenderX

Diary and Registrations

The majority of patients reported using the diary and registration function on the platform. However, a few felt that the diary primarily concentrated on misuse and relapse, along with follow-up questions. For them, this focus was not relevant, as they aimed to lower their consumption; they would have preferred the program to also highlight strategies for reducing drinking.

The remaining patients found it rewarding to gain insights into their own alcohol use, the habits that led to cravings, and strategies for managing those cravings. They noted that using the program encouraged self-reflection, allowing them to take the time to think about and evaluate their thoughts and feelings throughout the process. The patients emphasized that this self-reflection facilitated an internal dialog about whether their rehabilitation journey was progressing as they hoped. As one patient expressed, regarding the use of the diary and registration features:

> *I had good help from my therapist to become aware of what happened inside of me. At one point, I just wrote number of units and clicked 'send'. And then she asked me why I did not fill out the rest with feelings, thoughts, and where I was etc. And then I said, it is because it is the same every day and it is just a routine now. And then she convinced me to try. She said, something still happens in your thoughts and in your feelings and in your body, try to really feel. And it was quite interesting because it was maybe also especially through that that I realized that this is about being afraid of choice. One big avoidance maneuver from my side against meeting my thoughts when I go to sleep. Corresponding, in the periods where I managed not to drink, then she asked me to also register what happened emotionally, mentally, and bodily. And it was quite good exercises, especially with her guidance, where I managed to dive into it because it is so easy for me to scroll over it* [Patient, ID 1268]

The patient expressed how the therapist guided and motivated them to engage more deeply with the program to fully understand the nuances of alcohol misuse. This written reflection enhanced the patient's awareness of their thoughts, making previously vague issues surrounding their alcohol use more tangible. As a result, this clarity contributed to progress in reducing their alcohol intake or achieving abstinence. Most patients noted that during the FtF CBT sessions, the therapist would begin the discussion where the written feedback had concluded, enriching the conversation with additional perspectives. This approach made the dialog more nuanced than the written feedback alone. One patient described how the concise nature of their written assignment answers and the subsequent feedback was designed to be expanded upon during the conversation.

> *I imagine that if the therapist had produced more comprehensive written feedback, one could be nervous that the question or conversation would close with the answer he then gives. And then it doesn't have the conversation-generating effect when you show up for the next session because you can be of the understanding that the subject is closed. And typically, it is not a subject that can or should be closed but ought to be unfolded during more future conversations* [Patient, ID 1350]

Most patients found it helpful to receive brief written feedback on their completed assignments, which they could then actively discuss during the FtF CBT sessions. However, a couple of patients expressed concern that the work they had done in the online treatment program did not always connect with the topics addressed by their therapist during the FtF sessions.

Motivating Reminders

About one-half of the patients felt motivated by receiving reminders from their therapists when they had not engaged with the program for several days. One patient elaborated on this point:

> *I also think that the therapist in the other end was good at following up on my platform use. If I was too slow, she wrote to me. Or not if I was too slow, but she made sure to follow up so the whole thing did not come to a standstill, I think. [...] And it worked very well. The motivation was rising. And when you get a reminder like that, you know there is another person in the other end, and then you also become motivated to do something about it, I think* [Patient, ID 1320]

The patients felt that knowing there was a person behind the platform expecting their engagement added to their motivation and progress in the treatment. Although the reminders might come across as somewhat supervisory, they were viewed as helpful and well-intended.

Discussion

Overview of Findings

In this study, we found that users experienced several advantages of using iCBT, including increased anonymity and privacy. Most importantly, it offered flexibility in the treatment process, allowing patients to focus on their rehabilitation at their own pace. The availability of written material in the online program was appreciated, as it provided valuable knowledge about AUD and its effects on individuals. Systematic access to written material that could be revisited helped patients better understand the mechanisms behind their problematic drinking—knowledge they felt they had not previously received. For some patients, the written content also served as a useful starting point for discussions with relatives, reducing the time therapists needed to spend on explaining background information and facts about AUD. We noted that patients found the platform particularly beneficial for gaining insights into their alcohol problems compared with solely FtF treatment. The assignments not only facilitated acknowledgment of their issues with alcohol but also encouraged self-reflection before FtF sessions, allowing for deeper exploration with the therapist. Reminders were well-received and helped motivate patients to complete assignments. Only a few disadvantages of combining internet-based iCBT with FtF treatment were mentioned.

XSL•FO
RenderX

Consistent with previous studies, patients in this study found the iCBT program user-friendly, describing it as intuitive and easy to navigate. However, a small number of patients reported difficulty answering some questions within the program, as they felt those questions were not relevant to their specific issues. They suggested that the assignments should include more examples to better address a wider range of patients and to make the program more personalized. This aligns with findings from various studies indicating that online programs often fail to meet all patients' individual needs, leading to frustration [7] and decreased motivation [7,17,19]. Reduced motivation can also stem from programs being overly time-consuming or demanding [7,20].

Our findings also support those of the qualitative study by Khadjesari et al [8], which examined people searching for online help to reduce their drinking, as well as the user experience study on internet-based treatment for problematic alcohol use by Ekström and Johansson [5]. Both studies concluded that some individuals felt embarrassed about their AUD and preferred to keep their struggles private. Ekström and Johansson [5] also reported that some patients feared that if local authorities became aware of their drinking problem, they might risk losing their children or jobs, which led them to avoid attending alcohol treatment centers. We found that a key aspect of choosing a blended treatment course was the added anonymity and privacy it provided for patients. Practical factors that influenced their decision to select the online format included the flexibility it offered, making it easier and more convenient to integrate into their daily lives, including family commitments, work, and transportation challenges.

Perhaps more importantly, we found that the time patients spent online on tasks they could complete independently seemed to free up time for FtF sessions with the therapist, allowing for deeper therapeutic exploration of more complex issues. This phenomenon is known as the "flipped clinic" [21,22], where digital resources are used for aspects of treatment that do not require the therapist's presence, thereby reserving in-person time for when it is most needed. This finding aligns with the qualitative study by Black et al [1] on user motivations and experiences with an internet intervention for alcohol reduction, as well as with Hadjistavropoulos et al [16], who explored patient perspectives on the strengths and challenges of therapist-assisted iCBT. Both studies found that patients appreciated the flexibility of online treatment, allowing them to complete the course at their own pace and use it as a substitute for FtF CBT when necessary. This conclusion was also reached in the review by Verhoeks et al [7], which discussed women's expectations and experiences regarding eHealth treatment. It suggests that eHealth may be seen as a complementary or integrated aspect of traditional health care rather than as a standalone solution. Therefore, iCBT could serve as an important tool in the recovery process to help prevent relapse. Patients who are already familiar with the treatment content may find it easier to engage with an online program when needed, or conversely, they might feel more comfortable interacting with a therapist after gaining foundational knowledge from online modules. Further studies on the effects of flipped clinics are warranted.

Most patients in this study found that using a digital diary enabled them to track changes in their alcohol consumption, recognize triggers for cravings, and practice strategies for managing those cravings, ultimately strengthening their sense of autonomy. Furthermore, as noted by Ekström and Johansson [5] in their user experience study, patients expressed appreciation for the therapists behind the screen, praising their professionalism and support. The therapists were perceived as honest and genuine in their approach, which enhanced the patient's experience by providing helpful feedback and supporting their change process. This finding is consistent with the results of this study.

Patients expressed that the desire for change regarding their alcohol problem should originate from within themselves. This introspection required time for self-reflection at their own pace, which they felt was facilitated by bCBT. This aligns with findings from the user experience study by Ekström and Johansson [5], where access to written text led to a high level of recognition among patients, who remarked, "that's why I react like I do or behave as I do." In a patient perspective study on iCBT for alcohol misuse, Hadjistavropoulos et al [17] similarly found that written material from iCBT enhanced self-awareness and provided insights into patients' problems with AUD, helping them identify reasons for making changes. In a related qualitative study on experiences of iCBT for depression and anxiety, Lindegaard et al [6] also reported that patients found the written texts informative. These texts were perceived as helpful in changing routines, limiting negative thoughts, and fostering a shift toward more positive thinking. In summary, we aim to highlight that self-reflection is essential for patients with AUD to strengthen their autonomy. This autonomy is further supported when FtF CBT is combined with iCBT in a blended treatment approach (bCBT).

Increased autonomy through self-reflection involves a deep and conscious examination of thoughts, as well as recognizing the origins and causes of one's feelings, emotions, and actions. Achieving this level of reflection requires a substantial amount of uninterrupted time, which is often challenging in FtF CBT sessions where time is limited. In this context, bCBT facilitates an effective interaction by providing a valuable combination of time, quietness, and contemplation for the patient. The patients in the Blend-A study emphasized that using the platform before FtF sessions allowed the therapist to start the session based on the patient's feedback from the assignments. This approach fostered a more engaging, in-depth, and meaningful conversation between the therapist and the patient. Therefore, the patients found that both the iCBT and the FtF CBT sessions with the therapist contributed significantly to their reflection process. This dual approach enhanced their sense of perceived autonomy, which in turn improved their ability to succeed in treatment and ultimately in their rehabilitation.

In general, autonomy is often understood as *substantive independence* and *self-determination* [23,24]. However, in this context, the concept of relational autonomy [25,26] appears more relevant. Relational autonomy emphasizes the influence of situations marked by oppression or deprivation—such as struggles with alcohol—and recognizes how external factors can constrain an individual's autonomy and ability to act. It also

XSL•FO
RenderX

considers how values, attitudes, and standards associated with these constraints can shape an individual's perception of what they believe they can achieve and how they can perform [27]. Based on studies of traumatized individuals, Bernstein [28] suggested that people may confront "underlying and intractable dimensions of vulnerability, dependence, and potential helplessness that are normally hidden from consciousness." We propose that the journey of transitioning from alcohol addiction similarly involves an experience of relative vulnerability, dependence, and helplessness during the change process. Before deciding to seek treatment, patients with AUD may have experienced a loss of self-care, increased cravings that lead to diminished control, and a lack of awareness of their surroundings—conditions that self-reflection can painfully illuminate. In the Blend-A study, we found that patients recognized how engaging with iCBT assignments heightened their awareness and sensitivity to their drinking behavior, enabling them to reflect on their issues in a broader and deeper context than what they could achieve through FtF sessions alone. This finding is supported by a qualitative study in which therapists noted that patients were more likely to attribute their treatment success to themselves in the context of online therapy compared with FtF treatment [29]. In our study, therapists facilitated the patients' autonomy by bridging the gap between the forward-looking goal of achieving abstinence or reducing alcohol intake and the self-realization process that took place in a private setting.

Moreover, when viewed through the lens of relational autonomy, the patients in this study found that the diary function helped them better understand themselves, recognize what was happening in their lives, and assess their progress throughout treatment, thereby strengthening their autonomy. Some patients reported that reminders from the therapists were motivating, especially if they had not engaged with the program for several days. This finding aligns with previous studies that concluded check-ins serve as motivating factors, helping patients hold themselves accountable [30]. Some patients described the feedback from their therapist as positive, acknowledging, and supportive, viewing it as a solid foundation for subsequent FtF CBT sessions. They appreciated that the therapists tailored their feedback to benefit each patient, aligning the feedback objectives with the outcomes of the patient's self-reflection. For example, the ability of a patient to drink less may not initially hold intrinsic value in isolation. However, if through self-reflection the patient recognizes the need to stop drinking for reasons such as health, family, or job security, then that realization aligns the treatment goals with a meaningful emotional experience. The therapist then helps the patient develop an action plan to achieve the goal of reducing or eliminating alcohol consumption. For this endeavor to succeed, both the patient and the therapist must actively engage with each other, identifying new opportunities and goals through collaboration.

In summary, successful cooperation may not always be achievable if the patient is not genuinely committed to the treatment, particularly in settings with limited time and insufficient opportunities for careful reflection. This highlights the importance of therapists helping patients cultivate a balanced sense of autonomy, while also addressing the risks of losing control over alcohol, experiencing relapse, or discontinuing treatment. Patients especially noted that they used the online program to read materials and reflect on assignments before their clinic visits. This approach may have allowed them to advance further in their treatment than they might have with FtF sessions alone.

Limitations and Strengths

This study has both limitations and strengths. One limitation is that the interview guide was not pilot-tested, which may have influenced how themes were addressed and how questions were posed, potentially impacting participants' responses. Additionally, the lack of stakeholder checks on the transcripts and codes means there was no external validation of the findings [31]. As we approached the 30 most recently enrolled patients in the Blend-A study, selection bias may have occurred during recruitment. This could have resulted in a less representative participant sample, potentially affecting the external validity of the results [32]. Moreover, it may introduce bias into the findings that all patients considered themselves skilled technology users and felt comfortable using the internet for treatment purposes. However, this aligns with the distribution of digital technology use in the general Danish population, where 97 out of 100 families had internet access, 96% owned a mobile phone, and 91% had a computer in 2020 [33].

We consider it a strength that the same researcher interviewed 13 patients for this study, as this number is, according to Guest et al [34], sufficient to achieve data saturation, allowing for an inductive analysis process where significant themes can emerge from the raw data. This number of participants also facilitates a stronger interviewer-interviewee alliance, which enhances the validity of the interviews [35]. Additionally, the credibility of the study findings is bolstered by the fact that 2 researchers conducted the coding of the raw data to ensure internal validation [31,36,37].

Perspectives and Implications to Practice

In this study, we found it essential to consider all the reasons for introducing new technology, ensuring that users understand the exact nature of the application. For users, technology cannot stand alone; ideally, it should not replace FtF treatment. Instead, it should be clear to users how technology can serve as an additional resource in treatment and potentially enhance the overall quality of care. If this is not taken into account, there is a risk of diminishing users' desire and enthusiasm for new treatment approaches that incorporate technology.

Conclusions

In conclusion, our findings indicate that patients engaging in bCBT perceive more benefits than drawbacks associated with this blended intervention. bCBT enhances anonymity, privacy, and flexibility, providing ample time and space for reflection. Essentially, it offers patients a form of assisted autonomy that cannot be achieved through iCBT or FtF CBT alone. The combination of these approaches allows patients to fully appreciate the benefits of treatment, as they have the opportunity for self-reflection, supported by their therapist between FtF sessions.

XSL•FO
RenderX

Acknowledgments

The study was financially supported by TrygFonden (grant 127727). The authors thank patients and staff at the municipal AUD treatment institutions participating in the Blend-A study. The authors also thank the project team associated with the Blend-A study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[PDF File (Adobe PDF File), 129 KB - jmir_v26i1e47083_app1.pdf]

References

1. Black N, Loomes M, Juraskova I, Johnston I. Engagement in a novel internet intervention for alcohol reduction: a qualitative study of user motivations and experiences. Cyberpsychol Behav Soc Netw 2020 Apr;23(4):225-233. [doi: 10.1089/cyber.2019.0289] [Medline: 32083488]
2. Ritchie H, Roser M. Alcohol consumption, who consumes the most alcohol? How has consumption changed over time? And what are the health impacts? Our World in Data. 2018. URL: https://ourworldindata.org/alcohol-consumption [accessed 2024-10-02]
3. Hansen ABG, Hvidtfeldt UA, Grønbæk M, Becker U, Nielsen AS, Tolstrup JS. The number of persons with alcohol problems in the Danish population. Scand J Public Health 2011 Mar;39(2):128-136. [doi: 10.1177/1403494810393556] [Medline: 21382854]
4. Wallhed Finn S, Bakshi A, Andréasson S. Alcohol consumption, dependence, and treatment barriers: perceptions among nontreatment seekers with alcohol dependence. Subst Use Misuse 2014 May;49(6):762-769. [doi: 10.3109/10826084.2014.891616] [Medline: 24601784]
5. Ekström V, Johansson M. Choosing internet-based treatment for problematic alcohol use-why, when and how? Users' experiences of treatment online. Addict Sci Clin Pract 2020 Jun 29;15(1):22 [FREE Full text] [doi: 10.1186/s13722-020-00196-5] [Medline: 32600456]
6. Lindegaard T, Kashoush F, Holm S, Halaj A, Berg M, Andersson G. Experiences of internet-based cognitive behavioural therapy for depression and anxiety among Arabic-speaking individuals in Sweden: a qualitative study. BMC Psychiatry 2021 Jun 03;21(1):288 [FREE Full text] [doi: 10.1186/s12888-021-03297-w] [Medline: 34082745]
7. Verhoeks C, Teunissen D, van der Stelt-Steenbergen A, Lagro-Janssen A. Women's expectations and experiences regarding e-health treatment: a systematic review. Health Informatics J 2019 Sep;25(3):771-787 [FREE Full text] [doi: 10.1177/1460458217720394] [Medline: 28764600]
8. Khadjesari Z, Stevenson F, Godfrey C, Murray E. Negotiating the 'grey area between normal social drinking and being a smelly tramp': a qualitative study of people searching for help online to reduce their drinking. Health Expect 2015 Dec;18(6):2011-2020 [FREE Full text] [doi: 10.1111/hex.12351] [Medline: 25676536]
9. Sundström C, Peynenburg V, Chadwick C, Thiessen D, Wilhems A, Nugent M, et al. Optimizing internet-delivered cognitive behaviour therapy for alcohol misuse-a randomized factorial trial examining effects of a pre-treatment assessment interview and guidance. Addict Sci Clin Pract 2022 Jul 23;17(1):37 [FREE Full text] [doi: 10.1186/s13722-022-00319-0] [Medline: 35871010]
10. Holmberg TT, Sainte-Marie M, Jensen EK, Linnet J, Runge E, Lichtenstein MB, et al. An analysis of patient motivation for seeking online treatment for binge eating disorder-a mixed methods study combining systematic text condensation with sentiment analysis. Front Psychiatry 2022;13:969115 [FREE Full text] [doi: 10.3389/fpsyt.2022.969115] [Medline: 36405908]
11. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
12. Becker U, Bygholm H, Broholm K, Nielsen A, Nielsen P. Kvalitet i Alkoholbehandling - et Rådgivningsmateriale. Copenhagen, Denmark: Sundhedsstyrelsen, Center for Forebyggelse; 2008.
13. Riper H, Hoogendoorn A, Cuijpers P, Karyotaki E, Boumparis N, Mira A, et al. Effectiveness and treatment moderators of internet interventions for adult problem drinking: an individual patient data meta-analysis of 19 randomised controlled trials. PLoS Med 2018 Dec;15(12):e1002714 [FREE Full text] [doi: 10.1371/journal.pmed.1002714] [Medline: 30562347]
14. Tarp K, Rasmussen J, Mejldal A, Folker MP, Nielsen AS. Blended treatment for alcohol use disorder (Blend-A): explorative mixed methods pilot and feasibility study. JMIR Form Res 2022 Apr 25;6(4):e17761 [FREE Full text] [doi: 10.2196/17761] [Medline: 35468082]

XSL•FO
RenderX

15. Runz-Jørgensen SM, Schiøtz ML, Christensen U. Perceived value of eHealth among people living with multimorbidity: a qualitative study. J Comorb 2017;7(1):96-111 [FREE Full text] [doi: 10.15256/joc.2017.7.98] [Medline: 29359124]
16. Hadjistavropoulos HD, Faller YN, Klatt A, Nugent MN, Dear BF, Titov N. Patient perspectives on strengths and challenges of therapist-assisted internet-delivered cognitive behaviour therapy: using the patient voice to improve care. Community Ment Health J 2018 Oct;54(7):944-950 [FREE Full text] [doi: 10.1007/s10597-018-0286-0] [Medline: 29804226]
17. Hadjistavropoulos HD, Chadwick C, Beck CD, Edmonds M, Sundström C, Edwards W, et al. Improving internet-delivered cognitive behaviour therapy for alcohol misuse: patient perspectives following program completion. Internet Interv 2021 Dec;26:100474 [FREE Full text] [doi: 10.1016/j.invent.2021.100474] [Medline: 34722165]
18. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
19. Postel MG, de Haan HA, ter Huurne ED, Becker ES, de Jong CA. Effectiveness of a web-based intervention for problem drinkers and reasons for dropout: randomized controlled trial. J Med Internet Res 2010 Dec 16;12(4):e68 [FREE Full text] [doi: 10.2196/jmir.1642] [Medline: 21163776]
20. Postel MG, de Haan HA, ter Huurne ED, van der Palen J, Becker ES, de Jong CA. Attrition in web-based treatment for problem drinkers. J Med Internet Res 2011 Dec 27;13(4):e117 [FREE Full text] [doi: 10.2196/jmir.1811] [Medline: 22201703]
21. Davenport TA, Cheng VWS, Iorfino F, Hamilton B, Castaldi E, Burton A, et al. Flip the clinic: a digital health approach to youth mental health service delivery during the COVID-19 pandemic and beyond. JMIR Ment Health 2020 Dec 15;7(12):e24578 [FREE Full text] [doi: 10.2196/24578] [Medline: 33206051]
22. Hadjistavropoulos H, Nugent M, Dirkse D, Pugh N. Implementation of internet-delivered cognitive behavior therapy within community mental health clinics: a process evaluation using the consolidated framework for implementation research. BMC Psychiatry 2017 Sep 12;17(1):331 [FREE Full text] [doi: 10.1186/s12888-017-1496-7] [Medline: 28899365]
23. Christman J. Autonomy and liberalism: a troubled marriage? In: Wall S, editor. The Cambridge Companion to Liberalism. Cambridge, MA: Cambridge University Press; Mar 05, 2015:141-162.
24. Colburn B. Autonomy and Liberalism. New York, NY: Routledge; Feb 11, 2010.
25. Heidenreich K, Bremer A, Materstvedt LJ, Tidefelt U, Svantesson M. Relational autonomy in the care of the vulnerable: health care professionals' reasoning in moral case deliberation (MCD). Med Health Care Philos 2018 Dec;21(4):467-477. [doi: 10.1007/s11019-017-9818-6] [Medline: 29243015]
26. Mackenzie C, Stoljar N. Relational autonomy. In: Feminist Perspectives on Autonomy, Agency, and the Social Self. Oxford, UK: Oxford University Press; 2000.
27. Mackenzie C. Three dimensions of autonomy: a relational analysis. In: Veltman A, Piper M, editors. Autonomy, Oppression, and Gender. Oxford, UK: Oxford University Press; Aug 21, 2014:15-41.
28. BERNSTEIN JM. Trust: on the real but almost always unnoticed, ever‐changing foundation of ethical life. Metaphilosophy 2011 Jul 04;42(4):395-416. [doi: 10.1111/j.1467-9973.2011.01709.x]
29. Mol M, van Genugten C, Dozeman E, van Schaik DJF, Draisma S, Riper H, et al. Why uptake of blended internet-based interventions for depression is challenging: a qualitative study on therapists' perspectives. J Clin Med 2019 Dec 30;9(1):91 [FREE Full text] [doi: 10.3390/jcm9010091] [Medline: 31905834]
30. Horsch C, Spruit S, Lancee J, van Eijk R, Beun RJ, Neerincx M, et al. Reminders make people adhere better to a self-help sleep intervention. Health Technol (Berl) 2017;7(2):173-188 [FREE Full text] [doi: 10.1007/s12553-016-0167-x] [Medline: 29201588]
31. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. American Journal of Evaluation 2006 Jun 01;27(2):237-246. [doi: 10.1177/1098214005283748]
32. Hegedus EJ, Moody J. Clinimetrics corner: the many faces of selection bias. Journal of Manual & Manipulative Therapy 2013 Jul 18;18(2):69-73. [doi: 10.1179/106698110x12640740712699]
33. Tassy A, Törnfelt C. It-anvendelse i befolkningen 2022. Danmarks Statistik. 2023 Mar 01. URL: https://www.dst.dk/da/Statistik/nyheder-analyser-publ/Publikationer/VisPub?cid=44692 [accessed 2024-10-03]
34. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2006 Feb 01;18(1):59-82. [doi: 10.1177/1525822x05279903]
35. Crouch M, McKenzie H. The logic of small samples in interview-based qualitative research. Social Science Information 2006 Dec 01;45(4):483-499. [doi: 10.1177/0539018406069584]
36. Malterud K. Qualitative research: standards, challenges, and guidelines. The Lancet 2001 Aug;358(9280):483-488. [doi: 10.1016/s0140-6736(01)05627-6]
37. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research. Academic Medicine 2014 Sep 01;89(9):1245-1251. [doi: 10.1097/acm.0000000000000388]

Abbreviations

AUD: alcohol use disorder
bCBT: blended cognitive behavioral therapy

XSL•FO
RenderX

Blend-A: Blending Internet Treatment into Conventional Face-to-Face Treatment for AUD
CBT: cognitive behavioral therapy
COREQ: Consolidated Criteria for Reporting Qualitative Research
FtF: face-to-face
iCBT: internet-based cognitive behavioral therapy

Edited by T McCall; submitted 07.03.23; peer-reviewed by J Freund, S Hughes; comments to author 12.02.24; revised version received 10.04.24; accepted 15.08.24; published 23.10.24.

Please cite as:
Tarp K, Christiansen R, Bilberg R, Borkner S, Dalsgaard C, Paldam Folker M, Søgaard Nielsen A
Patient Perspectives on Blended Internet-Based and Face-to-Face Cognitive Behavioral Therapy for Alcohol Use Disorder: Qualitative Study
J Med Internet Res 2024;26:e47083
URL: https://www.jmir.org/2024/1/e47083
doi:10.2196/47083
PMID:39441642

©Kristine Tarp, Regina Christiansen, Randi Bilberg, Simone Borkner, Caroline Dalsgaard, Marie Paldam Folker, Anette Søgaard Nielsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Different Types of Social Media Use on the Mental Health of UK Adults: Longitudinal Observational Study

Yue Yu[1], MSc; Jennifer Dykxhoorn[2], PhD; Ruth Plackett[3], PhD

[1]Department of Epidemiology and Public Health, University College London, London, United Kingdom

[2]Division of Psychiatry, University College London, London, United Kingdom

[3]Research Department of Primary Care and Population Health, University College London, London, United Kingdom

Corresponding Author:
Ruth Plackett, PhD
Research Department of Primary Care and Population Health
University College London
Rowland Hill Street
London, NW3 2PF
United Kingdom
Phone: 44 203108309
Email: ruth.plackett.15@ucl.ac.uk

Abstract

Background: Previous studies have explored the association between social media use and mental health among adolescents. However, few studies using nationally representative longitudinal data have explored this relationship for adults and how the effect might change depending on how people use social media.

Objective: This study investigated the longitudinal relationship between the frequency of viewing and posting on social media and mental health problems among UK adults.

Methods: This study included 15,836 adults (aged 16 years and older) who participated in Understanding Society, a UK longitudinal survey. Social media use was measured with questions about the frequency of viewing social media and posting on social media in Understanding Society Wave 11 (2019-2021). We explored viewing and posting separately, as well as a combined exposure: (1) high viewing, high posting; (2) high viewing, low posting; (3) low viewing, high posting; and (4) low viewing, low posting. Mental health problems were measured in Wave 12 (2020-2022) using the General Health Questionnaire (GHQ-12), a validated scale for identifying symptoms of common mental health problems, where higher scores indicated more mental health problems (0 to 36). Unadjusted and adjusted linear regression models were estimated for viewing social media and posting on social media, adjusting for the baseline GHQ score, gender, age, ethnicity, employment, and education. We found no evidence for effect modification by gender and age so overall associations were reported.

Results: In our adjusted models, we found no evidence of an association between the frequency of viewing social media and mental health problems in the following year. We found that adults who posted daily on social media had more mental health problems than those who never posted on social media, corresponding to a 0.35-point increase in GHQ score (β=0.35, 95% CI 0.01-0.68; *P*=.04). When we considered both social media behaviors, we found that those who frequently viewed and posted on social media scored 0.31 points higher on the GHQ score (β=0.31, 95% CI 0.04-0.58; *P*=.03) in the following year compared to those who rarely viewed or posted on social media.

Conclusions: We found that a high frequency of posting on social media was associated with increased mental health problems a year later. However, we did not find evidence of a similar association based on the frequency of viewing social media content. This provides evidence that some types of active social media use (ie, posting) have a stronger link to mental health outcomes than some types of passive social media use (viewing). These results highlighted that the relationship between social media use and mental health is complex, and more research is needed to understand the mechanisms underlying these patterns to inform targeted interventions and policies.

(J Med Internet Res 2024;26:e56950) doi:10.2196/56950

XSL•FO
RenderX

KEYWORDS

social media; mental health; depression; anxiety; mental disorders; cohort studies; United Kingdom; longitudinal observational study

Introduction

In the United Kingdom, around 1 in 4 people experience mental health problems each year [1], and improving mental health is a public health priority [2]. The growth of social media in recent decades has caused speculation that the increasing prevalence of mental health problems could be linked to social media use [3,4]. While social media use has become an integral part of everyday life for many people, social media users can interact with these internet-based tools in different ways. Individuals generally engage with social media passively or actively. Passive social media use includes browsing content and viewing posts, while active social media use includes creating public posts or sending private messages to friends, family, and others, and many engage in both passive and active social media use [5].

Due to the hypothesized link between social media use and mental health problems, several studies have explored this relationship, but the findings to date are mixed. Some studies have found a small association between social media use and increased mental health problems, such as anxiety, depressive symptoms, and psychological distress, particularly for female participants [6-8]. Other studies have shown that social media can benefit mental health, as it allows people to interact with others, can strengthen social connectedness [9], and reduce loneliness and social isolation [6,10]. Other studies have found no evidence of an association between social media use and mental health problems [11-13].

Most of the existing literature has focused on the overall time spent on social media, not specifically considering the types of social media use [6,14,15]. Some studies that have distinguished between active and passive social media use have suggested that active use can elicit social support and positive feedback, having a positive effect on mental health, whereas passive use might be more harmful to mental health than active use because it encourages upward social comparison and envy [14]. For example, some studies have concluded that active social media use, defined as posting or interacting with other's content, was related to less loneliness and less mental health problems, while passive social media was associated with increased loneliness and more mental health problems, especially for females [16-18]. However, a systematic review found that overall existing studies do not support the hypothesis that active social media use is more beneficial for mental health than passive use as these behaviors were often crudely measured with varying definitions and measurements of active and passive use [5].

Many of the aforementioned studies were conducted on adolescents and young adults [6,15,17], and the evidence for adult population remains limited [10,19,20]. Importantly, different age groups have been shown to use different social media platforms and may have different patterns of social media use. For example, data have indicated that a majority of younger adults use Instagram and Snapchat the most but older adults usually use Facebook and YouTube [21]. Evidence has also shown that younger age groups use social media more for self-presentation, but older adults use social media more for enjoyment or keeping up with family and friends [22], which may have different implications for mental health. For example, younger adults may experience more adverse mental health effects from social media use due to social comparison [23], while older adults may be less likely to engage in social comparison and instead benefit from enhanced social connectedness and reduced loneliness resulting from keeping in touch with family and friends [24].

Most previous studies in this area have explored the relationship between social media use and mental health using cross-sectional designs with nonrepresentative data [6,25]. This restricts what temporal conclusions can be drawn about the relationship and how generalizable the findings are to the general population. Some longitudinal studies have been conducted in the United Kingdom, but many suggested nonsignificant associations for adolescents and only focused on time spent on social media rather than the different types of activities on social media [11,26,27]. This study aimed to use nationally representative data to examine the longitudinal association between viewing social media content and posting on social media and mental health problems in UK adults. We expected to find an association between high levels of viewing and posting and mental health problems in the following year but expected there to be stronger evidence for viewing social media (passive use) and increased mental health problems than posting (active use). We also expected the effects to be different according to different age groups and gender. For example, the effect of social media use on mental health will be smaller among older adults than young adults but stronger in females than in males, based on previous literature. This study will provide evidence of the temporal relationship between social media use and mental health problems, as well as help to inform strategies to mitigate the risks of different types of social media use in adults.

Methods

Study Population

This study used data from Understanding Society (USoc), a longitudinal household survey, and a representative sample of the UK population [28]. All adults (aged 16 years or older) who completed the survey in Wave 11 (2019-2021; baseline) and Wave 12 (2020-2022; follow-up) were included in our sample. The flowchart in Multimedia Appendix 1 displays the sample selection process.

Exposures

Two social media exposures were measured at baseline: (1) frequency of viewing social media, and (2) frequency of posting on social media. Frequency of viewing social media was measured by the question, "How often do you use the internet for personal use by looking at content on social media or

XSL•FO
RenderX

websites and apps (eg, looking at text, images, and videos on Facebook, Twitter, Instagram)?" Frequency of posting was measured by the question, "How often do you use the internet for personal use by posting content on social media or websites and apps (eg, posting text, images, videos on Facebook, Twitter, Instagram)?"; including posting to private accounts. The frequency for both questions was categorized as: (1) every day, (2) several times a week, (3) several times a month, (4) once a month, (5) less than once a month, and (6) never. Participants who had indicated that they did not have access to or use the internet were recorded as "never" for both frequency of viewing and frequency of posting. Those who reported "refusal" and "don't know" were counted as "missing."

We created a combined variable to capture both types of social media use. Participants were categorized as having a high frequency of viewing and posting if they reported "Every day," and "Several times a week," and those who reported "Several times a month," "Once a month," "Less than once a month," and "Never" were classified as low frequency. Participants were assigned to one of the following groups: (1) high viewing, high posting; (2) high viewing, low posting; (3) low viewing, high posting; and (4) low viewing, low posting.

Outcome

Mental health problems were measured in Wave 12 using the General Health Questionnaire (GHQ-12). The GHQ-12 includes 12 questions that assess psychological distress and symptoms of anxiety or depression over the past weeks. We used the recommended Likert scoring method [29], where each question is scored from 0 to 3, resulting in a total score from 0 to 36. Higher GHQ-12 scores indicate higher levels of psychological distress and more mental health problems.

Confounders

We included gender (female and male), age group (16-29, 30-44, 45-59, and 60 and older), education (6 levels: higher degree, 1st degree or equivalent, other higher degree, A level, General Certificate of Secondary Education/O level, and other qualification), employment (employed and not in paid employment), and ethnicity (White and all other ethnic groups combined) as confounders in this analysis. All confounders were measured at baseline, and we also adjusted for mental health problems at baseline (baseline GHQ-12 score).

Effect Modification

We tested for effect modification by gender and age group. As we did not find evidence for effect modification, we reported overall estimates.

Missing Data

Any exposure, covariate, or outcome that was missing data, as well as any responses of "don't know," or "refusal," were imputed. All variables in the regression model were used in the imputation model. Age, gender, and ethnicity had no missing data, so were included in the imputation models as auxiliary variables. We created 20 imputed datasets using multiple imputations with chained equations. Our primary analysis was based on the imputed data, which was compared to a sensitivity analysis of complete cases (Multimedia Appendix 2).

Weighting

To account for the household clustered, stratified survey design, correct for nonresponse and loss-to-follow-up, we used the longitudinal weight from Wave 12 provided by USoc. This allowed our analyses to better represent the adult population in the United Kingdom and to improve the reliability and validity of our findings. We also excluded from the analysis any participants who had a zero weight due to sample design or fieldwork issues in USoc, as the weighted analysis compensated for these cases being missing. We used the svy command in Stata (version 17; StataCorp LLC) to account for the complex survey design and the clustering at individual and area levels.

Statistical Analysis

We conducted unadjusted and adjusted linear regression on the imputed analytic sample. Model 1 (unadjusted) investigated the associations between social media use and mental health, model 2 included the adjustment for baseline mental health problems, and model 3 (fully adjusted) additionally adjusted for gender, age, education, employment, and ethnicity. These models were estimated for all three exposures: (1) frequency of viewing social media, (2) frequency of posting on social media, and (3) social media viewing and posting frequency. All analysis was conducted using Stata.

Ethical Considerations

This study involved secondary data analysis of the USoc dataset, which had already received ethical approval, and this approval stated that secondary analyses did not require additional approval. The University of Essex Ethics Committee has approved all data collection on USoc main study and innovation panel waves (ethics approval ETH1920-0123 for Wave 12). Participants of USoc provided informed consent for the use of their data in future primary or secondary research contexts. All data used in this study were anonymized by USoc prior to being accessed, ensuring that no identifying information was available to the researchers. Participants were informed at the time of data collection that their data would be used anonymously and for research purposes only, with no analysis being conducted at the individual level. Participants were compensated for taking part in USoc.

Results

Descriptive Analysis

There were 15,836 participants included in our analytic sample, with an average age of 53 (SD 18.08) years (Table 1). There were more female (n=9105, 57.5%) participants than male participants, and most of the sample was of White ethnicity (n=13,978, 88.3%). Just over half (n=8298, 52.4%) of the sample were employed, while the rest were not in paid employment. More than half of the participants (n=8761, 55.3%) viewed social media every day at baseline and only 15.4% (n=2433) of participants posted on social media every day. Around a fifth of participants (n=3504, 22.1%) have never viewed social media, and a larger percentage of participants (n=5767, 36.4%) never posted on social media.

XSL•FO
RenderX

Table 1. Descriptive statistics of the sample in terms of the key variables (N=15,836).

Variables	Value, n (%)	Mean (SD)
Age (in years)		53.50 (18.1)
16-29	2062 (13)	
30-44	2621 (16.6)	
45-59	4556 (28.8)	
60 and older	6597 (41.7)	
Missing	0 (0)	
Sex		N/A[a]
Male	6731 (42.5)	
Female	9105 (57.5)	
Missing	0 (0)	
Education		N/A
Higher degree	1739 (11)	
1st degree or equivalent	2491 (15.7)	
Other higher degree	1546 (9.8)	
A level[b]	1236 (7.8)	
GCSE[c]/O level	2798 (17.7)	
Other qualification	1200 (7.6)	
Missing	4826 (30.5)	
Employment		N/A
Employed	8298 (52.4)	
Not in paid employment	7509 (47.4)	
Missing	29 (0.2)	
Ethnicity		N/A
White background	13,978 (88.3)	
All other ethnic groups combined	1858 (11.7)	
Missing	0 (0)	
Baseline GHQ-12 score (wave 11)		
Score	N/A	11.40 (5.45)
Missing	142 (0.9)	N/A
Follow-up GHQ-12 score (wave 12)		
Score	N/A	11.74 (5.67)
Missing	152 (1.0)	N/A
Frequency of viewing social media		N/A
Everyday	8761 (55.3)	
Several times a week	1990 (12.6)	
Several times a month	700 (4.4)	
Once a month	269 (1.7)	
Less than once a month	579 (3.7)	
Never	3504 (22.1)	
Missing	33 (0.2)	
Frequency of posting on social media		N/A

XSL•FO
RenderX

Variables	Value, n (%)	Mean (SD)
Everyday	2433 (15.4)	
Several times a week	2217 (14)	
Several times a month	2112 (13.3)	
Once a month	998 (6.3)	
Less than once a month	2277 (14.4)	
Never	5767 (36.4)	
Missing	32 (0.2)	
Frequency of viewing and posting on social media		N/A
Low viewing, low posting	4970 (31.4)	
Low viewing, high posting	75 (0.5)	
High viewing, low posting	6175 (39)	
High viewing, high posting	4569 (28.9)	
Missing	47 (0.3)	

[a]N/A: not applicable.

[b]A level: advanced level qualifications.

[c]GCSE: General Certificate of Secondary Education.

A large proportion of participants (n=6175, 39%) viewed social media frequently but posted less frequently, 31.4% (n=4970) of participants viewed and posted on social media both at low frequency, 28.9% (n=4569) of the participants viewed and posted both at high frequency, it is unusual (n=75, 0.5%) for participants to view social media less frequently but post frequently. The mean GHQ score at baseline was 11.40 (SD 5.45) and was 11.74 (SD 5.67) at follow-up.

Imputation

The proportion of missing data at baseline can be found in Table 1. There was no missing data for gender, age, and ethnicity, but a high proportion of missing data for education (n=4826, 30.5%), and less than 1% missing for employment, GHQ scores, and social media variables (Table 1).

Effect Modification

We tested for the presence of effect modification by gender and age but did not find evidence for effect modification for either variable once we adjusted for baseline mental health problems and other confounders (Multimedia Appendices 3 and 4). Therefore, both gender and age were included as confounders and were adjusted with other covariates in the fully adjusted model.

Linear Regression Analysis

Overview

Table 2 shows the results of the linear regression models estimating the association between social media exposures and later mental health problems.

XSL•FO
RenderX

Table 2. Linear regression model results for the association between types of social media use and mental health problems (N=15,836).

Variables	Value, n (%)	Model 1: unadjusted (crude association)			Model 2: adjusted for baseline GHQ[a]			Model 3: fully adjusted[b]		
		β	95% CI	*P* value	β	95% CI	*P* value	β	95% CI	*P* value
Frequency of viewing social media										
Never	3517 (22.2)	reference	reference	reference	reference	reference	reference	reference	reference	reference
Less than once a month	579 (3.66)	.25	–0.28 to 0.78	.35	.00089	–0.46 to 0.46	.99	–.0003	–0.46 to 0.46	.99
Once a month	271 (1.71)	.94	0.04 to 1.83	.04	–.12	–0.81 to 0.57	.73	–.16	–0.85 to 0.53	.64
Several times a month	701 (4.43)	.86	0.30 to 1.41	.003	.49	0.04 to 0.93	.03	.43	–0.07 to 0.87	.05
Several times a week	1997 (12.6)	.77	0.35 to 1.19	<.001	.36	–0.01 to 0.73	.06	.23	–0.13 to 0.58	.22
Every day	8771 (55.4)	1.34	1.05 to 1.64	<.001	.53	0.29 to 0.76	<.001	.12	–0.13 to 0.38	.34
Frequency of posting on social media										
Never	5781 (36.5)	reference	reference	reference	reference	reference	reference	reference	reference	reference
Less than once a month	2279 (14.4)	.69	0.33 to 1.05	<.001	.26	–0.04 to 0.56	.09	.06	–0.24 to 0.36	.69
Once a month	1001 (6.32)	.57	0.10 to 1.04	.02	.20	–0.22 to 0.61	.35	–.09	–0.50 to 0.32	.66
Several times a month	2118 (13.4)	.90	0.52 to 1.28	<.001	.39	0.08 to 0.71	.01	.10	–0.22 to 0.42	.54
Several times a week	2222 (14)	1.39	1.02 to 1.77	<.001	.71	0.41 to 1.01	<.001	.38	0.07 to 0.69	.02
Every day	2435 (15.4)	1.56	1.16 to 1.95	<.001	.73	0.41 to 1.06	<.001	.35	0.01 to 0.68	.04
Frequency of viewing and posting on social media										
Low viewing, low posting	4991 (31.5)	reference	reference	reference	reference	reference	reference	reference	reference	reference
Low viewing, high posting	77 (0.50)	.93	–0.58 to 2.44	.23	.95	–.16 to 2.06	.09	.76	–0.38 to 1.90	.19
High viewing, low posting	6188 (39.1)	.74	0.46 to 1.01	<.001	.26	0.03 to 0.49	.03	–.03	–0.26 to 0.20	.81
High viewing, high posting	4580 (28.9)	1.53	1.21 to 1.85	<.001	.71	0.46 to 0.97	<.001	.31	0.04 to 0.58	.03

[a]GHQ: General Health Questionnaire.

[b]Covariates: gender, age, employment, education, and ethnicity.

XSL•FO
RenderX

Viewing Social Media

In the unadjusted model, adults who viewed social media once a month or more had higher GHQ scores a year later than those who never viewed social media content. The largest increase was observed in those who viewed social media every day, where the mean GHQ scores were 1.34 points higher (95% CI 1.05-1.64; *P*<.001) than those who never viewed social media content. Once we adjusted for baseline mental health problems, those who viewed social media every day, or several times a month continued to have elevated GHQ scores compared to those who never viewed social media. However, once we adjusted for all confounders, this association was fully attenuated, and we found no evidence of increased mental health problems based on the frequency of viewing social media content.

Posting on Social Media

In the unadjusted model, participants who ever posted on social media had higher GHQ scores a year later than those who never posted, with the largest increase in those who posted every day (β=1.56, 95% CI 1.16-1.95; *P*<.001), with a similar increase in those who posted several times a week (β=1.39, 95% CI 1.02-1.77; *P*<.001). Following adjustment for the baseline GHQ, those who posted every day, several times a week, or several times a month continued to have higher GHQ scores than those who never posted. Finally, in the fully adjusted model, those who posted the most frequently (every day or several times a week) had elevated GHQ scores, but the magnitude of this increase had attenuated (every day: β=0.35, 95% CI 0.01-0.68; *P*=.04; several times a week: β=0.38, 95% CI 0.066-0.69; *P*=.02).

Viewing and Posting on Social Media

When we considered both social media exposures simultaneously, we found that those with a high frequency of both viewing and posting had an increased GHQ score across all models, although the magnitude of the association attenuated with each adjustment (unadjusted: β=1.53; 95% CI 1.21-1.85; *P*<.001; baseline GHQ adjusted: β=0.71, 95% CI 0.46-0.97; *P*<.001; fully adjusted: β=0.31, 95% CI: 0.04-0.58; *P*=.03). We found higher GHQ scores in the unadjusted and partially adjusted models for those who frequently viewed social media but posted infrequently, however, this association fully attenuated in the adjusted model (Table 2).

Complete Case Sensitivity Analysis

We repeated the linear regression models in the complete case sample (n=10,869), finding similar patterns (Multimedia Appendix 2). The complete case analysis also found no evidence of an association between the frequency of viewing social media and GHQ in the fully adjusted model, some evidence of an association between higher frequency of posting and higher GHQ scores, but the magnitude of effect attenuated with each adjustment. There was no evidence of an association between both viewing and posting frequency on social media and GHQ scores later.

Discussion

Principal Findings

This longitudinal study investigated the relationship between different types of social media use and later mental health problems in UK adults. We found no evidence for an association between the frequency of viewing social media content and later mental health problems. In contrast, we found that individuals who posted social media content several times a week or more had increased mental health problems at follow-up compared to those who never posted, even after accounting for baseline mental health and a range of covariates. This suggested that some types of active use of social media, specifically posting, might have a more negative effect on later mental health than passive use of social media. When looking at both behaviors, we found that individuals who frequently viewed and posted on social media had increased mental health problems when compared to individuals who rarely viewed or posted on social media.

Comparisons With Previous Studies

The results of this study contradicted some previous studies exploring the links between types of social media use and mental health. For example, a longitudinal study found an association between active social media use and reduced depressive symptoms for female adolescents [18]. Their study measured active use by measuring how often an adult posted messages or pictures or videos about themselves or others in the past 6 months. A cross-sectional study based on US adults aged 18-49 years old found evidence that a higher frequency of passive social media use related to increased depressive symptoms, while active use was associated with reduced depressive symptoms [20]. Their study defined active use as how often they liked or shared or posted and commented on others' content. In contrast, we did not find evidence of an association between a higher level of passive social media use and more mental health problems, but we concluded that a higher frequency of active social media use was related to more mental health problems a year later, and this did not vary by gender. The differences in how social media activities were measured could account for the inconsistency between these conclusions. Our study assessed active use of social media by asking how often somebody posted messages or pictures or videos on social media, which may not have been as comprehensive as some of the activities included in the other studies and may account for the different findings. To enable better comparison of findings, future research should use more specific and consistent measures of active and passive use. The discrepancies between these studies and our findings could also be due to the different age groups included in the analysis and our analysis was the only analysis to be conducted with adults on longitudinal nationally representative data.

Our findings were more consistent with a cross-sectional study that concluded that for adults, active social media use was associated with severe anxiety and stress but not severe depression, whereas, passive social media use showed no significant association with mental health problems [30]. Like our study, this study measured active use by how often a person

XSL•FO
RenderX

posted or commented on posts, but the analysis was not longitudinal or conducted on nationally representative data. Our conclusion that more frequent posting on social media was related to poorer mental health a year later, aligned with a recent longitudinal study conducted by Lowthian et al [16] that explored social media use of adolescents in the COVID-19 pandemic. They distinguished between 2 types of active social media use: posting and messaging with internet-only friends, their findings suggested that both active social media use and passive social media use were associated with poorer mental well-being. However, "avid" users who were defined as having the highest communication frequency with people they did not know on social media had the poorest mental health outcomes, compared to those who used social media to speak with friends and family and for social or political engagement. This suggests that some public active social media activities, like messaging strangers on social media or posting publicly, may be more harmful to some people than other active uses like private messaging with friends and family or engaging in social or political activities. We also found that adults who both frequently viewed and posted on social media have more mental health problems a year later than those who rarely used social media for either purpose. This conclusion was also consistent with a longitudinal study that suggested that adolescents who frequently shared content, messaged, and browsed social media were more likely to have poorer mental health a year later than those who minimally used social media [31]. Overall, these findings suggest that both time spent on social media and type of activity are important to consider in the relationship between social media use and mental well-being and should be measured in future studies. However, we also need to measure more specific active uses of social media both public and private to get a more nuanced understanding of what kinds of activities might be harmful or helpful for our mental health and for which groups of people.

Our findings were contradictory to our hypotheses that passive use would be more harmful to adult mental health than active use, but some research has suggested why some types of active social media use may lead to poorer mental health. Posting actively on social media may have negative effects on mental health because people may post unrealistic content about themselves or because they are replacing in-person relationships with web-based relationships [30]. Some types of active use, such as messaging friends, may have positive effects on mental health as it can improve social connectedness and reduce loneliness. Further, some types of passive social media use may not have a negative impact on mental health because people may be positively comparing themselves to others or receive admiration or inspiration from others [32,33]. Theories about social media use and the relationship with mental health are still developing and further qualitative research will help to develop our understanding of the underlying mechanisms in this relationship. In addition, the effects of social media on mental health are likely to be affected by the type of content posted or viewed on social media. For example, studies have found that viewing self-harm–related images or words on social media has a range of harmful effects on people's mental health, such as a higher level of emotional distress and more self-harming thoughts and behaviors [34]. Further research is needed to explore not only the range of activities people do on social media but also the types of content people view or post on social media and how they might affect mental health.

There was no evidence from this study showing gender or age differences in the association between social media use and mental health problems. This is despite previous literature suggesting that younger age groups and female groups might have worse mental health outcomes as a result of problematic social media use than male and older age groups [6,23,24]. Understanding the specific activities that can be helpful and harmful for mental health for different groups could help to target social media use management interventions at those who are most susceptible to the negative effects of social media use and harness some of the benefits of social media to improve mental health.

Strengths and Weakness

USoc is a large longitudinal study providing a nationally representative sample of UK adults, making the findings from this study generalizable to the UK population. A key strength of this study was that we were able to consider a more detailed picture of social media use by examining viewing and posting frequency on social media. Given that most existing literature has only used time spent on social media as the exposure, this study provided a more comprehensive understanding of social media behaviors. This longitudinal analysis can shed light on the potential causal relationship between social media use and mental health by accounting for the temporal order of exposures and outcomes. We adjusted for mental health at baseline, to account for the reverse causality of social media on mental health that has been suggested by previous research: poor mental health conditions may drive heavier social media use [35].

It is important to consider various limitations of this study. This study was only based on two waves of data because the two main exposures were only available in these two waves. This may not be sufficient to generate conclusions about long-term trends in mental health conditions. Future studies, with longer follow-ups, would explore the long-term impacts of social media use on mental health.

The complexity of the study design raised challenges in developing accurate weights for this population. While USoc was designed (in Wave 1) to be representative of the UK population, wave nonresponse and loss-to-follow-up affect the sample in subsequent waves. In line with the recommendations from USoc, we used the provided longitudinal weights. These weights were created for monotone longitudinal response and include design weights to account for differences in selection probabilities and wave nonresponse adjustments. Where a participant missed one or more previous waves, they will receive a zero-weight in this analysis, so were dropped from this analysis. While this reduces the sample size, there was sufficient statistical power to conduct this analysis and it ensured that the results were generalizable to the UK population.

Additionally, the outcome GHQ-12 was based on 12 self-report survey questions to measure mental health problems, which may lead to memory bias as participants may have difficulty accurately recalling their mental health symptoms over a specific

XSL•FO
RenderX

period before the survey, leading to inaccuracies in the scores. Hence, future studies may consider supplementing self-report measures with clinical evaluations by mental health professionals for a comprehensive understanding of participants' mental health problems. Furthermore, this study used social media use data from Wave 11 as the exposure variables, which was collected from 2019 to 2021, when the COVID-19 pandemic was prevalent. Studies suggest that the COVID-19 outbreak largely increased social media use in many countries [36,37], which could have affected the posting frequency on social media [38]. Evidence also showed that during the pandemic, many people experienced poorer mental health due to isolation policies, illness, job loss, and other factors [39,40]. Therefore, the conclusions of this study might be less applicable to the context before or after the pandemic. We ran a sensitivity analysis, to attempt to account for the influence of COVID-19, by controlling for the year in which people took part in the study at baseline and found this did not change the results significantly.

This study offered insights into the impact of social media use on mental health in adults. Future research could further explore the impact of increased mental health problems resulting from frequent active social media use by incorporating additional measures of social functioning such as social connectedness, family relations, and leisure activities [41]. By further assessing the social functioning of participants, we could enhance the understanding of the real-life impact of the increased mental health problems observed in this study. This could then inform more targeted interventions and support strategies to improve social outcomes for adults who experienced more mental health challenges due to frequent or problematic social media use.

Conclusions

We found that the ways in which adults use social media have a differential impact on mental health. While we found no evidence that the frequency of viewing social media content was linked to mental health outcomes, we found that frequently posting on social media was associated with increased mental health problems a year later compared to those who never posted on social media. This increase in mental health problems was also observed in those who both frequently viewed social media content and frequently posted on social media, compared to those who rarely used social media for either purpose. To help mitigate the negative impacts on adult mental health, it would be helpful to encourage people to be mindful of how they use social media, and the content they interact with, to develop healthy digital behaviors. Some types of active use, like public posting and interacting with people we do not know on social media, might have more harmful effects on mental health than actively using social media for more private uses such as messaging friends. Social media companies also have a responsibility to protect users from harmful content, which can also affect mental health, and better regulations from governments could help to ensure greater progress on social media safety.

Further research is needed to explore the range of activities people do on social media, how they might affect mental health, and which groups are most negatively affected. This could help us to understand how to harness the strengths of social media to improve mental well-being and to address the negative effects of social media use on mental health.

Acknowledgments

RP was funded by the National Institute for Health and Care Research (NIHR) Three Schools' Mental Health Programme (project reference MHF013) and supported by NIHR ARC North Thames. JD was funded by an NIHR Advanced Fellowship (NIHR302266). This work was supported by the National Institute for Health Research, University College London Hospital, Biomedical Research Centre (JD). The views expressed are those of the author or authors and not necessarily those of the NIHR or the Department of Health and Social Care. Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by the National Centre for Social Research (NatCen) and Verian (formerly Kantar Public). The research data are distributed by the UK Data Service. Fieldwork for the web survey was carried out by Ipsos MORI and for the telephone survey by Kantar.

Conflicts of Interest

RP has conducted consultancy work for Meta on a related topic of how to protect young people's mental health on social media.

Multimedia Appendix 1
Flowchart showing the sample selection process.
[PDF File (Adobe PDF File), 105 KB - jmir_v26i1e56950_app1.pdf]

Multimedia Appendix 2
Linear regression model results for the association between types of social media use and mental health problems for complete case analysis (N=10,869).
[DOCX File , 21 KB - jmir_v26i1e56950_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Results for effect modification for gender and age (adjusted data, imputed sample, N=15,836).
[DOCX File , 15 KB - jmir_v26i1e56950_app3.docx]

Multimedia Appendix 4
The association between posting frequency and mental health stratified by age (adjusted data, imputed sample, N=15,836).
[DOCX File , 18 KB - jmir_v26i1e56950_app4.docx]

References

1. Adult psychiatric morbidity in England—2007, results of a household survey. NHS Digital. URL: https://digital.nhs.uk/data-and-information/publications/statistical/adult-psychiatric-morbidity-survey/adult-psychiatric-morbidity-in-england-2007-results-of-a-household-survey [accessed 2023-12-06]
2. Improving mental health is a priority for public health, human rights and sustainable development. URL: https://www.who.int/publications/m/item/improving-mental-health-is-a-priority-for-public-health--human-rights-and-sustainable-development [accessed 2023-11-09]
3. The Online Safety Bill can be used against us. Socialist Worker. URL: https://socialistworker.co.uk/features/the-online-safety-bill-can-be-used-against-us/ [accessed 2023-08-15]
4. Online Safety Bill. URL: https://publications.parliament.uk/pa/cm5803/cmpublic/OnlineSafetyBill/memo/OSB76.htm [accessed 2023-08-15]
5. Valkenburg PM, van Driel II, Beyens I. The associations of active and passive social media use with well-being: a critical scoping review. New Media Soc 2021;24(2):530-549 [FREE Full text] [doi: 10.1177/14614448211065425]
6. Keles B, McCrae N, Grealish A. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. Int J Adolesc Youth 2019;25(1):79-93. [doi: 10.1080/02673843.2019.1590851]
7. Karim F, Oyewande AA, Abdalla LF, Ehsanullah RC, Khan S. Social media use and its connection to mental health: a systematic review. Cureus 2020;12(6):e8627 [FREE Full text] [doi: 10.7759/cureus.8627] [Medline: 32685296]
8. O'Reilly M, Dogra N, Whiteman N, Hughes J, Eruyar S, Reilly P. Is social media bad for mental health and wellbeing? Exploring the perspectives of adolescents. Clin Child Psychol Psychiatry 2018;23(4):601-613 [FREE Full text] [doi: 10.1177/1359104518775154] [Medline: 29781314]
9. Vaingankar JA, van Dam RM, Samari E, Chang S, Seow E, Chua YC, et al. Social media-driven routes to positive mental health among youth: qualitative enquiry and concept mapping study. JMIR Pediatr Parent 2022;5(1):e32758 [FREE Full text] [doi: 10.2196/32758] [Medline: 35254285]
10. Fu L, Xie Y. The effects of social media use on the health of older adults: an empirical analysis based on 2017 Chinese general social survey. Healthcare (Basel) 2021;9(9):1143 [FREE Full text] [doi: 10.3390/healthcare9091143] [Medline: 34574917]
11. Plackett R, Sheringham J, Dykxhoorn J. The longitudinal impact of social media use on UK adolescents' mental health: longitudinal observational study. J Med Internet Res 2023;25:e43213 [FREE Full text] [doi: 10.2196/43213] [Medline: 36961482]
12. Jensen M, George M, Russell M, Odgers C. Young adolescents' digital technology use and mental health symptoms: little evidence of longitudinal or daily linkages. Clin Psychol Sci 2019;7(6):1416-1433 [FREE Full text] [doi: 10.1177/2167702619859336] [Medline: 31929951]
13. Zhong B, Jiang Z, Xie W, Qin X. Association of social media use with mental health conditions of nonpatients during the COVID-19 outbreak: insights from a national survey study. J Med Internet Res 2020;22(12):e23696 [FREE Full text] [doi: 10.2196/23696] [Medline: 33302256]
14. Verduyn P, Ybarra O, Résibois M, Jonides J, Kross E. Do social network sites enhance or undermine subjective well-being? A critical review. Social Issues Policy Rev 2017;11(1):274-302 [FREE Full text] [doi: 10.1111/sipr.12033]
15. Thorisdottir IE, Sigurvinsdottir R, Asgeirsdottir BB, Allegrante JP, Sigfusdottir ID. Active and passive social media use and symptoms of anxiety and depressed mood among Icelandic adolescents. Cyberpsychol Behav Soc Networking 2019;22(8):535-542. [doi: 10.1089/cyber.2019.0079] [Medline: 31361508]
16. Lowthian E, Fee G, Wakeham C, Clegg Z, Crick T, Anthony R. Identifying protective and risk behavior patterns of online communication in young people. J Adolesc 2024;96(2):235-250. [doi: 10.1002/jad.12270] [Medline: 37929875]
17. Machado VC, Mcilroy D, Adamuz FMP, Murphy R, Palmer-Conn S. The associations of use of social network sites with perceived social support and loneliness. Curr Psychol 2022:1-14 [FREE Full text] [doi: 10.1007/s12144-021-02673-9] [Medline: 35103040]
18. Fredrick SS, Nickerson AB, Livingston JA. Adolescent social media use: pitfalls and promises in relation to cybervictimization, friend support, and depressive symptoms. J Youth Adolesc 2022;51(2):361-376 [FREE Full text] [doi: 10.1007/s10964-021-01561-6] [Medline: 35066709]

XSL•FO
RenderX

19. Perlis RH, Green J, Simonson M, Ognyanova K, Santillana M, Lin J, et al. Association between social media use and self-reported symptoms of depression in US adults. JAMA Netw Open 2021;4(11):e2136113 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.36113] [Medline: 34812844]
20. Escobar-Viera CG, Shensa A, Bowman ND, Sidani JE, Knight J, James AE, et al. Passive and active social media use and depressive symptoms among United States adults. Cyberpsychol Behav Soc Networking 2018;21(7):437-443. [doi: 10.1089/cyber.2017.0668] [Medline: 29995530]
21. Anderson M, Auxier B. Social media use in 2021. Pew Research Center: Internet, Science & Tech. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2023-11-11]
22. Sharifian N, Kraal AZ, Zaheed AB, Sol K, Morris EP, Zahodne LB. Measurement invariance of social media use in younger and older adults and links to socioemotional health. Innovation Aging 2021;5(2):igab009 [FREE Full text] [doi: 10.1093/geroni/igab009] [Medline: 33928196]
23. Lewin KM, Meshi D, Schuster AM, Cotten SR. Active and passive social media use are differentially related to depressive symptoms in older adults. Aging Ment Health 2023;27(1):176-183. [doi: 10.1080/13607863.2022.2068133] [Medline: 35470731]
24. Chen E, Wood D, Ysseldyk R. Online social networking and mental health among older adults: a scoping review. Can J Aging 2022;41(1):26-39. [doi: 10.1017/S0714980821000040] [Medline: 33736725]
25. Orben A, Przybylski AK, Blakemore S, Kievit RA. Windows of developmental sensitivity to social media. Nat Commun 2022;13(1):1649 [FREE Full text] [doi: 10.1038/s41467-022-29296-3] [Medline: 35347142]
26. Orben A, Przybylski AK. The association between adolescent well-being and digital technology use. Nat Hum Behav 2019;3(2):173-182. [doi: 10.1038/s41562-018-0506-1] [Medline: 30944443]
27. Thorisdottir IE, Sigurvinsdottir R, Kristjansson AL, Allegrante JP, Lilly CL, Sigfusdottir ID. Longitudinal association between social media use and psychological distress among adolescents. Prev Med 2020;141:106270 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106270] [Medline: 33031868]
28. University of Essex. Understanding Society: Waves 1-13, 2009-2022 and Harmonised BHPS: Waves 1-18, 1991-2009 18th Edition. United Kingdom: Institute for Social and Economic Research; 2022.
29. Goldberg DP, Gater R, Sartorius N, Ustun TB, Piccinelli M, Gureje O, et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychol Med 1997;27(1):191-197. [doi: 10.1017/s0033291796004242] [Medline: 9122299]
30. Scarpulla E, Stosic MD, Weaver AE, Ruben MA. Should I post? The relationships among social media use, emotion recognition, and mental health. Front Psychol 2023;14:1161300 [FREE Full text] [doi: 10.3389/fpsyg.2023.1161300] [Medline: 37287775]
31. Winstone L, Mars B, Haworth CMA, Heron J, Kidger J. Adolescent social media user types and their mental health and well-being: results from a longitudinal survey of 13-14-year-olds in the United Kingdom. JCPP Adv 2022;2(2):e12071 [FREE Full text] [doi: 10.1002/jcv2.12071] [Medline: 37431459]
32. Meier A, Gilbert A, Börner S, Possler D. Instagram inspiration: how upward comparison on social network sites can contribute to well-being. J Commun 2020;70(5):721-743. [doi: 10.1093/joc/jqaa025]
33. Meier A, Krause HV. Does passive social media use harm well-being? An adversarial review. J Media Psychol 2023;35(3):169-180. [doi: 10.1027/1864-1105/a000358]
34. Susi K, Glover-Ford F, Stewart A, Bevis RK, Hawton K. Research review: viewing self-harm images on the internet and social media platforms: systematic review of the impact and associated psychological mechanisms. J Child Psychol Psychiatry 2023;64(8):1115-1139. [doi: 10.1111/jcpp.13754] [Medline: 36940718]
35. Hartanto A, Quek FYX, Tng GYQ, Yong JC. Does social media use increase depressive symptoms? A reverse causation perspective. Front Psychiatry 2021;12:641934 [FREE Full text] [doi: 10.3389/fpsyt.2021.641934] [Medline: 33833700]
36. Lee Y, Jeon YJ, Kang S, Shin JI, Jung Y, Jung SJ. Social media use and mental health during the COVID-19 pandemic in young adults: a meta-analysis of 14 cross-sectional studies. BMC Public Health 2022;22(1):995 [FREE Full text] [doi: 10.1186/s12889-022-13409-0] [Medline: 35581597]
37. Finding social support through social media during COVID lockdowns. UConn Today. 2022. URL: https://today.uconn.edu/2022/06/finding-social-support-through-social-media-during-covid-lockdowns/ [accessed 2023-08-01]
38. Feehan B. The impact of coronavirus on social media engagement for brands. Rival IQ. 2020. URL: https://www.rivaliq.com/blog/coronavirus-on-social-media-engagement-for-brands/ [accessed 2023-08-01]
39. Panchal N, Saunders H, Rudowitz R, Cox C. The implications of COVID-19 for mental health and substance use. KFF. 2023. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/ [accessed 2023-08-01]
40. COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. URL: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide [accessed 2023-08-01]
41. Carlson JA, Sarkin AJ, Levack AE, Sklar M, Tally SR, Gilmer TP, et al. Evaluating a measure of social health derived from two mental health recovery measures: the California quality of life (CA-QOL) and mental health statistics improvement

XSL•FO
RenderX

program consumer survey (MHSIP). Community Ment Health J 2011;47(4):454-462 [FREE Full text] [doi: 10.1007/s10597-010-9347-8] [Medline: 20878235]

Abbreviations

GHQ: General Health Questionnaire
USoc: Understanding Society

Edited by T de Azevedo Cardoso; submitted 31.01.24; peer-reviewed by L Winstone, Z Haime, C Shen; comments to author 28.02.24; revised version received 25.04.24; accepted 16.09.24; published 30.10.24.

Please cite as:
Yu Y, Dykxhoorn J, Plackett R
The Impact of Different Types of Social Media Use on the Mental Health of UK Adults: Longitudinal Observational Study
J Med Internet Res 2024;26:e56950
URL: https://www.jmir.org/2024/1/e56950
doi:10.2196/56950
PMID:

©Yue Yu, Jennifer Dykxhoorn, Ruth Plackett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predictors of Engagement in Multiple Modalities of Digital Mental Health Treatments: Longitudinal Study

Molly Aideen Nowels[1], MA, MS, PhD; Meghan McDarby[2], PhD; Lilla Brody[3], BA; Evan Kleiman[4], PhD; Sara Sagui Henson[5], PhD; Cynthia Castro Sweet[5], PhD; Elissa Kozlov[1], PhD

[1]Division of Geriatrics and Palliative Medicine, Weill Cornell Medicine, New York, NY, United States

[2]Department of Psychiatry & Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York City, NY, United States

[3]Divison of Geriatrics & Palliative Medicine, Weill Cornell Medicine, New York City, NY, United States

[4]Department of Psychology, Rutgers University, New Brunswick, NJ, United States

[5]Modern Health, San Francisco, CA, United States

Corresponding Author:
Molly Aideen Nowels, MA, MS, PhD
Division of Geriatrics and Palliative Medicine
Weill Cornell Medicine
525 E 68th St
New York, NY, 10065
United States
Phone: 1 212 746 4888
Email: mon2007@med.cornell.edu

Abstract

Background: Technology-enhanced mental health platforms may serve as a pathway to accessible and scalable mental health care; specifically, those that leverage stepped care models have the potential to address many barriers to patient care, including low mental health literacy, mental health provider shortages, perceived acceptability of care, and equitable access to evidence-based treatment. Driving meaningful engagement in care through these platforms remains a challenge.

Objective: This study aimed to examine predictors of engagement in self-directed digital mental health services offered as part of an employer-based mental health benefit that uses a technology-enabled care platform.

Methods: Using a prospective, longitudinal design, we examined usage data from employees who had access to an employer-sponsored mental health care benefit. Participants had access to a digital library of mental health resources, which they could use at any time, including daily exercises, interactive programs, podcasts, and mindfulness exercises. Coaching and teletherapy were also available to. The outcome was engagement with the self-directed digital mental health resources, measured by the number of interactions. Poisson regression models included sociodemographic characteristics, patient activation, mental health literacy, well-being, PHQ-9 and GAD-7 scores at baseline, primary concern for engaging in treatment, and the use of coaching or teletherapy sessions.

Results: In total 950 individuals enrolled in the study, with 38% using any self-directed digital mental health resources. Approximately 44% of the sample did not use the app during the study period. Those using both self-directed digital and 1:1 modalities made up about one-quarter of the sample (235/950, 24.7%). Those using only coaching or therapy (170/950, 17.9%) and those using only self-directed digital mental health resources (126/950, 13.3%) make up the rest. At baseline, these groups statistically significantly differed on age, PHQ-9, GAD-7, MHLS, and primary concern. Receipt of coaching and teletherapy was associated with the number of self-directed digital mental health resources interactions in adjusted Poisson regression modeling. Use of any coach visit was associated with 82% (rate ratio [RR] 1.82, 95% CI 1.63-2.03) more self-directed digital mental health resource interactions while use of any teletherapy session was associated with 80% (RR 1.80, 95% CI 1.55-2.07) more digital mental health resources interactions (both $P<.001$). Each additional year of age was associated with increased digital mental health resources interactions (RR 1.04, 95% CI (1.03-1.05), and women had 23% more self-directed digital resources interactions than men (RR 1.23, 95% CI 1.09-1.39).

Conclusions: Our key finding was that the use of coaching or teletherapy was associated with increased self-directed digital mental health resource use. Higher self-directed digital resource engagement among those receiving coaching or therapy may be

XSL•FO
RenderX

a result of provider encouragement. On the other hand, when a participant engages with 1 modality in the platform, they may be more likely to begin engaging with others, becoming “super users” of all resources.

(J Med Internet Res 2024;26:e48696) doi:10.2196/48696

KEYWORDS

digital health; mental health; health care benefit; prediction; technology; digital mental health; employer-based; teletherapy; coaching; utilization; mobile phone

Introduction

Mental health and substance use disorders are leading causes of disease burden in the United States [1]. However, while nearly 25% of adults experience a mental disorder at any given time [2], less than a quarter (21.7%) of individuals with a mental disorder receives treatment from a licensed mental health clinician [3]. This discrepancy between the prevalence of mental health disorders and rates of mental health service use is largely attributed to barriers to health care access, such as not being able to schedule an appointment in a timely manner [4] and persistent shortages in the mental health workforce—especially in rural communities and for people with serious mental illness [5,6]. As the demand for psychological services outpaces the supply of mental health providers, innovative mental health care delivery models are increasingly needed.

Digital mental health (digital health) platforms represent a promising approach to deliver accessible, scalable, and timely mental health care. These platforms often provide care through multiple modalities with varying intensities (eg, traditional one-on-one telepsychotherapy or telecoaching, asynchronous self-guided treatments, and telegroup-based treatments). As digital health platforms can be equipped to provide a range of treatments to patients, they have the potential to dismantle key barriers to patient care, including mental health provider shortages and inequitable access to mental health treatment [7-9].

However, digital health platforms can be challenging to implement, largely because user adoption and continued engagement are relatively low [10-12]. While traditional psychotherapy also struggles with early dropout rates ranging between 8% and 20% [13,14], digital health historically has seen much higher rates of discontinuation [11]. As a result, while they hold the potential to promote increased, equitable access to mental health care, additional research is needed to determine how to facilitate increased engagement in mental health care from digital health platforms. Recent research investigating engagement in digital health applications after initial adoption suggests that several factors, including user attributes (eg severity of depressive symptoms), technology issues (eg concerns about anonymity and privacy), components of user experience (eg, interface design of a digital health application) [15] and motivation to engage in the health goal of interest, are associated with continued use [16]. Other research indicates that users’ beliefs about the severity of their health conditions may also be associated with digital health adoption and continued use [17] and that improvement in psychological outcomes of interest during early treatment may predict continued engagement [18]. Finally, it is important to understand the appropriate level of engagement in different services given a user’s needs and preferences. Not all users will need long-term unlimited engagement with a digital health platform, and engagement likely involves evaluating the need for continued use [19,20].

As the number of digital mental health platforms continues to grow and more modalities of digital mental health care (eg, podcasts, online courses, meditations, discussion boards, etc.) become available to consumers, it is critical to understand predictors of engagement in digital mental health care. Elucidating the range of factors that predict initial and continued use of digital health services for mental health will support these platforms in further customizing care approaches to maximize user engagement, and ultimately, optimize personalized mental health treatment. The purpose of this study was to identify predictors of engagement in multiple modalities of mental health treatments offered on an employer-sponsored digital mental health care benefit.

Methods

Design and Participants

This investigation was part of a prospective, longitudinal, observational study of individuals who received services through an employer-sponsored digital mental health benefits platform (Modern Health Inc). The study timeframe was from September 20, 2021, to May 31, 2022.

Eligible participants were 18 years or older, based in the United States, registered for the benefit, and had access to a smartphone, tablet, or computer.

Procedures

Eligible employees registered for a Modern Health account through a mobile app or website and completed several onboarding assessments, which asked about areas of focus, care modality preferences, and clinical symptoms. The platform’s proprietary algorithm recommended an initial care pathway to participants based on this information. Recommendations used a stratified stepped care model, meaning that patients with higher acuity and with certain areas of focus were more likely to be recommended to more intensive care options (eg, one-on-one teletherapy) (more details in Figure 1). Participants could use their recommended care or self-refer or be referred by a provider to any combination of care.

XSL•FO
RenderX

Figure 1. Stratified stepped care model incorporating areas of support, care preferences, and clinical assessments into personalized care recommendations. PHQ-9: 9-item Patient Health Questionnaire; GAD-7: 7-item Generalized Anxiety Disorder Questionnaire; WHO-5: 5-item World Health Organization Well-Being Index.

Members who matched with a provider or used at least 1 digital resource after onboarding were invited by email to complete a screening questionnaire within 2 weeks of onboarding, which contained questions about demographics. We used the demographic screening information for purposeful recruitment to ensure representation in racial and ethnic identities, gender identities, socioeconomic status, age groups, and mental health symptoms at baseline. After the screening questionnaire, eligible participants were directed to a consent form to provide informed consent. Consenting participants were sent the baseline survey asking questions about their mental, social, and physical health by email along with instructions and compensation. Participants were sent a follow-up survey at 12 weeks post baseline that contained the same mental, social, and physical health questions. All forms and surveys were hosted on Qualtrics' online survey platform.

Digital Mental Health Services

Individuals in this study were able to engage in all of the following digital mental health services, within the limits of their employer-covered sessions for 1:1 care modalities. Self-directed digital resources had no limits.

Digital Resources

Participants had access to a digital library of programs and resources they could use in unlimited quantities at any time. These included daily meditative exercises, interactive programs and podcasts, mindfulness exercises like meditations and breathing exercises, and self-paced structured educational lessons (similar to self-help workbooks). All digital materials were developed and designed by an in-house team of clinical psychologists. Digital health programs were designed to cover topics such as emotions, relationships, professional life, healthy lifestyles, and finances. Engagement across all digital resources was combined in analyses to represent total digital program engagement.

Telecoaching

Coaches were certified by an International Coaching Federation accredited program and underwent a careful screening process by Modern Health. They provided 30-minute telecoaching services to eligible employees through a videoconferencing platform (Zoom; Zoom Video Communications). Coaches and participants were also able to communicate by messaging within the app. All coaches had at least 150 hours of coaching experience in addition to completing additional training in evidence-based techniques and Modern Health's proprietary model of care. Rather than being confined to any specific protocol, coaches were encouraged to implement evidence-based techniques where appropriate. The number of coaching sessions attended by participants depended on the number of sessions covered by their employer, as well as on personal preferences and their level of need.

Teletherapy

Furthermore, 50 minutes one-on-one teletherapy sessions were provided through a videoconferencing platform (Zoom) by licensed and vetted therapists who had an advanced degree in clinical psychology or a related field (eg, PhD, PsyD, LCSW, LMFT, or LPC). Similar to telecoaching, therapists and participants were able to communicate through messaging in-app. All therapists were skilled in evidence-based practices, such as cognitive behavioral therapy, acceptance and commitment therapy, and dialectical behavior therapy and had completed additional training in Modern Health's proprietary model of care. Similar to telecoaching, the number of therapy

XSL•FO
RenderX

sessions attended by participants depended on the number of sessions covered by their employer, as well as on personal preferences and their level of need [21].

Outcome Measures

Engagement in digital mental health services: We created 4 outcome variables to capture engagement in platform services. First, we created a 4-level categorical variable to represent use of the digital health services: (1) only self-directed digital resource use, (2) only telecoach or teletherapist use, (3) use of both digital and telecoach or teletherapist modalities, and (4) no use of any modality. Second, we created a binary variable (0/1) indicating whether participants used at least 1 service within the platform (inclusive of self-directed digital resources, one-on-one telecoaching visits, and one-on-one teletherapy visits). Third, we created a binary variable (0/1) indicating whether participants used at least one of the self-directed digital resources collapsed across resource types: meditative exercises, programs, meditations, and podcasts. Fourth, we created a count of the number of self-directed digital resources each participant used. The use of both self-directed digital resources and 1:1 modalities was collected automatically by the digital mental health platform. Engagement data during the study period (between the baseline and follow-up surveys) was extracted from the platform for analysis.

Predictor Measures

Demographics

Participants self-identified their age, race and ethnicity (American Indian or Alaskan Native, Asian, Black or African American, Hawaiian or Pacific Islander, Hispanic, Latino, or Spanish origin, non-Hispanic white, prefer to self-describe, multiracial; collapsed to non-Hispanic White vs non-White), gender identity (man, woman, non-binary), and educational attainment (less than a bachelors' degree, bachelors' degree, and more than a bachelors' degree).

Depression

The 9-item Patient Health Questionnaire (PHQ-9) [22] was used to assess the presence of depression symptoms over the past 2 weeks. Participants responded on a 4-point scale (0= "Not at all," 3= "Nearly every day"), and total scores were categorized into 1 of 5 severity ratings: none, minimal, mild, moderate, or moderately severe. Higher scores indicate higher severity of depression symptomatology.

Anxiety Symptoms

The 7-item Generalized Anxiety Disorder Questionnaire (GAD-7) [23] was used to assess the presence of anxiety symptoms over the past 2 weeks. Participants responded on a 4-point scale (0="Not at all," 3="Nearly every day") and total scores were categorized into 1 of 4 severity ratings: minimal, mild, moderate, or severe. Higher scores indicate higher severity of anxiety symptomatology.

Well-Being

The 5-item World Health Organization Well-Being Index (WHO-5) [24] was used to assess well-being over the past 2 weeks. This construct was assessed during the onboarding process, approximately 2 weeks before the baseline assessment. Participants responded on a 6-point scale (0="At no time," 5="All of the time"). Higher scores indicate greater subjective well-being.

Patient Activation

The 13-item Patient Activation Measure for Mental Health (PAM-MH) was used [25]. Participants chose from 5 possible response options ranging from 1=Strongly Disagree to 4=Strongly Agree to 5=Not Applicable. Higher scores indicate higher patient activation.

Mental Health Literacy

Mental health literacy was measured using 13 items from the Mental Health Literacy Scale (MHLS) [26]. Response options ranged from 1=Strongly Disagree to 5=Strongly Agree.

Area of Focus

Participants were able to choose from over 40 topics of focus, or reasons for registering for the platform, during the onboarding process. These topics were wide-ranging and organized by the following well-being dimensions: "my emotions," "my professional life," "my physical well-being," "my relationships," and "my finances."

One-on-One Care Engagement

Variables indicating whether participants had used at least telecoaching (0/1) or teletherapy (0/1) sessions within the platform were also included as predictors in analyses predicting the use of self-directed digital resources.

Statistical Analyses

We conducted 4 sets of analyses examining the association between all the predictors and each engagement in digital mental health services outcome. First, we used descriptive analyses to investigate differences in predictor variables across the 4-level categorical variable of use, with ANOVA performed for continuous predictors and chi-square tests for categorical predictors.

Second, we examined predictors of any digital health service use in the app using logistic regression. The outcome variable was the binary variable representing if participants had used at least 1 service. In this regression model, we included age, race and ethnicity (non-White as reference group), gender identity (man as reference group), educational attainment (less than bachelor's degree completion as reference group), patient activation, mental health literacy, well-being, area of focus ("my emotions" as reference group), baseline depression (entered as a linear predictor), and baseline anxiety (entered as a linear predictor) as predictors.

Third, we examined predictors of any digital health self-directed digital resource use in the app using logistic regression. The outcome variable was the binary variable representing if participants had at least one self-directed digital resource in the 3-month follow-up period. In this regression, we included age, race and ethnicity, gender identity, educational attainment, patient activation, mental health literacy, well-being, primary concern, baseline depression (entered as a linear predictor), and

XSL•FO
RenderX

baseline anxiety (entered as a linear predictor) as predictors, with the same reference groups as in the previous analysis.

Fourth, using Poisson regression, we examined if the use of any telecoaching or teletherapy visits was associated with the number of self-directed digital resources used. We created binary variables for both telecoaching and teletherapy, reflecting any use of each modality in the follow-up period. We used both unadjusted and adjusted approaches to investigate this association. In the unadjusted model, we only included the telecoaching and teletherapy visit variables as predictors. In the adjusted model, we additionally included age, race and ethnicity, gender identity, educational attainment, patient activation, mental health literacy, well-being, and area of focus, with reference groups mirroring the previous analyses. We also performed the same adjusted Poisson regression analyses on a subgroup of our sample: those with elevated GAD-7 or PHQ-9 scores (scores on either scale above 9 indicating moderate symptoms [27,28]) at baseline.

Ethical Considerations

All participants provided informed consent, and this study protocol was reviewed and approved by the Western Clinical Group institutional review board (protocol #1316167). Compensation for this study was a US $25 digital gift card per completed survey. All data were deidentified.

Results

In total, 950 individuals were enrolled in this study. More than half of participants (n=531, 55.9%) engaged with at least 1 service within the platform (eg, self-directed digital resource, telecoach, or teletherapist session); 44.1% (n=419) did not engage in any service. Among engagers, 44.3% (n=235) used both self-directed digital resources and 1:1 care (teletherapy or telecoaching), 32% (n=170) used only 1:1 care, and 23.7% (n=126) used only self-directed digital resources. At baseline, these engagement groups statistically significantly differed in age, depressive symptoms, anxiety symptoms, health literacy, and area of focus (Table 1). Among participants who only used self-directed digital resources, the average number of digital engagements was 2.0 (median 1.0). Among those using both 1:1 care and self-directed digital resources, the average number of digital engagements was 3.4 (median 2.0).

In the logistic regression analysis predicting engagement in any service, only higher educational attainment was associated with increased odds of engaging in at least 1 service (Table 2). Relative to participants without a bachelor's degree, those with a bachelor's had 54.5% (P=.03) higher odds of engaging, and those with education greater than a bachelor's had 56.2% (P=.05) higher odds of engaging. No other variables in the model were statistically significant.

In our logistic regression analysis predicting engagement in any self-directed digital resources, certain demographic characteristics were associated with an increased likelihood of engagement. For every year of increasing age, subjects were 2.6% more likely to have used any self-directed resources (P=.003). Relative to men, women had 45.4% higher odds of using any self-directed resources (P=.01). Finally, educational attainment was also significantly associated with self-directed resource use, such that those with bachelor's degrees (odds ratio 1.56 95% CI 1.026-2.414; P=.04) and with education higher than a bachelor's degree (odds ratio 1.65 95% CI 1.045-2.625; P=.03) had higher odds of self-directed resource engagement than those with less than a bachelor's degree (Table 3).

In an unadjusted Poisson regression examining if engagement in at least 1 telecoaching or teletherapy was associated with self-directed digital resources use, engagement in any telecoach sessions was associated with a 71.6% increase in the number of self-directed digital resources used, while engagement in any teletherapy sessions was associated with a 41.4% increase in the number of digital resources used (Table 4). After adjusting for the additional predictor variables described above, engagement in any telecoaching sessions was associated with an 82% increase in the number of self-directed resources used, and engagement in any teletherapy sessions was associated with a 79.5% increase in the number of digital resources used. In the adjusted analysis, age was a significant predictor of engagement, with each additional year of age associated with a 3.9% increase in the number of self-directed digital resources used. Relative to men, those identifying as women used 23% more digital resources and non-binary individuals did not statistically significantly differ. The primary topic of focus was also a significant predictor of self-directed resource engagement, with finances (relative risk [RR] 1.46, P=.02), physical well-being (RR 2.28 P<.001), professional life (RR 1.71, P<.001), and relationships (RR 1.37, P<.001) all being associated with increased digital resource use as compared with those who chose emotions. In the subgroup analysis conducted in participants with an elevated GAD-7 or PHQ-9 score, engagement in any telecoaching was associated with a 101% increase in the number of self-directed digital resources used (P<.001), and engagement in any teletherapist visit was associated with a 77.8% increase in the number of digital resources used (P<.001) (Table 5).

XSL•FO
RenderX

Table 1. Demographics and baseline characteristics, overall and by modality of app use. *P* values are derived from chi-square tests for categorical variables and from ANOVA tests for continuous variables.

	Overall (N=950)	No use (N=419)	Therapy or coach only (N=170)	Self-directed digital resources only (N=126)	Both modalities (N=235)	*P* value
Age						<.001
Mean (SD)	33.9 (8.70)	33.9 (8.68)	31.5 (7.02)	36.8 (9.79)	34.0 (8.76)	
Median (range)	32 (19-65)	32 (19-63)	30 (22-57)	36.5 (22-60)	32 (21-65)	
Race and ethnicity (White vs non-White), n (%)						.74
Non-White	399 (42)	184 (43.9)	73 (42.9)	49 (38.9)	93 (39.6)	
White	548 (57.7)	232 (55.4)	97 (57.1)	77 (61.1)	142 (60.4)	
Missing	3 (0.3)	3 (0.7)	0 (0)	0 (0)	0 (0)	
Gender identity, n (%)						.66
Gender-nonbinary	46 (4.8)	23 (5.5)	8 (4.7)	4 (3.2)	11 (4.7)	
Man	343 (36.1)	164 (39.1)	59 (34.7)	48 (38.1)	72 (30.6)	
Woman	559 (58.8)	230 (54.9)	103 (60.6)	74 (58.7)	152 (64.7)	
Prefer not to say	2 (0.2)	2 (0.5)	0 (0)	0 (0)	0 (0)	
Educational attainment, n (%)						.40
<Bachelors	140 (14.7)	75 (17.9)	23 (13.5)	17 (13.5)	25 (10.6)	
Bachelors	545 (57.4)	230 (54.9)	105 (61.8)	71 (56.3)	139 (59.1)	
>Bachelors	265 (27.9)	114 (27.2)	42 (24.7)	38 (30.2)	71 (30.2)	
Baseline PHQ-9[a]						.03
Mean (SD)	8.58 (5.96)	9.13 (6.51)	8.51 (5.41)	7.11 (5.52)	8.54 (5.51)	
Median (range)	8 (0-27)	8 (0-27)	8 (0-23)	6 (0-23)	8 (0-25)	
Missing, n (%)	35 (3.7)	35 (8.4)	0 (0)	0 (0)	0 (0)	
Baseline GAD-7[b]						.01
Mean (SD)	7.73 (5.51)	8.03 (5.75)	7.86 (5.21)	6.10 (5.09)	8.01 (5.44)	
Median (range)	7 (0-21)	7 (0-21)	7 (0-21)	5 (0-20)	7 (0-21)	
Missing, n (%)	34 (3.6)	34 (8.1)	0 (0)	0 (0)	0 (0)	
Baseline PAM[c]						.85
Mean (SD)	58.3 (13.0)	57.9 (13.1)	59.3 (13.2)	58.3 (12.8)	58.2 (12.7)	
Median (range)	55.6 (17.9-100)	53.2 (17.9-100)	55.6 (34.2-100)	55.6 (30.4-100)	55.6 (31.7-100)	
Missing, n (%)	35 (3.7)	34 (8.1)	1 (0.6)	0 (0)	0 (0)	
Baseline MHLS[d]						.02
Mean (SD)	55.9 (6.43)	55.2 (7.13)	56.9 (6.01)	56.9 (6.16)	55.8 (5.46)	
Median (range)	57 (29-65)	57 (29-65)	58.5 (34-65)	58 (29-65)	57 (37-65)	
Missing, n (%)	26 (2.7)	26 (6.2)	0 (0)	0 (0)	0 (0)	
Baseline WHO-5[e]						.16
Mean (SD)	43.3 (18.3)	42.1 (17.7)	42.8 (19.4)	46.7 (18.7)	44.1 (18.2)	
Median (range)	44.0 (0-100)	40 (0-100)	40 (4-88)	48 (8-84)	44 (0-100)	
Any past mental health care, n (%)						.88
No	233 (24.5)	112 (26.7)	43 (25.3)	33 (26.2)	45 (19.1)	
Yes	244 (25.7)	111 (26.5)	42 (24.7)	34 (27)	57 (24.3)	
Missing	473 (49.8)	196 (46.8)	85 (50)	59 (46.8)	133 (56.6)	
Any current mental health care, n (%)						.66

XSL•FO
RenderX

	Overall (N=950)	No use (N=419)	Therapy or coach only (N=170)	Self-directed digital resources only (N=126)	Both modalities (N=235)	*P* value
No	464 (48.8)	217 (51.8)	82 (48.2)	62 (49.2)	103 (43.8)	
Yes	275 (28.9)	114 (27.2)	49 (28.8)	44 (34.9)	68 (28.9)	
Missing	211 (22.2)	88 (21)	39 (22.9)	20 (15.9)	64 (27.2)	
Primary concern, n (%)						.02
My emotions	466 (49.1)	198 (47.3)	102 (60)	44 (34.9)	122 (51.9)	
My finances	33 (3.5)	15 (3.6)	5 (2.9)	5 (4)	8 (3.4)	
My physical well-being	95 (10)	40 (9.5)	9 (5.3)	23 (18.3)	23 (9.8)	
My professional life	141 (14.8)	65 (15.5)	17 (10)	27 (21.4)	32 (13.6)	
My relationship	214 (22.5)	100 (23.9)	37 (21.8)	27 (21.4)	50 (21.3)	
Missing	1 (0.1)	1 (0.2)	0 (0)	0 (0)	0 (0)	

[a]PHQ-9: 9-item Patient Health Questionnaire.

[b]GAD-7: 7-item Generalized Anxiety Disorder Questionnaire.

[c]PAM: Patient Activation Measure.

[d]MHLS: Mental Health Literacy Scale.

[e]WHO-5: 5-item World Health Organization Well-Being Index.

XSL•FO
RenderX

Table 2. Adjusted logistic regression analysis predicting any app use (N=908).

Variables	Odds ratio (95% CI)	*P* value
Age	1.002 (0.986-1.019)	.78
Race and ethnicity		
White	1.074 (0.813-1.419)	.61
Non-White	REF[a]	—[b]
Gender identity		
Nonbinary	1.084 (0.559-2.125)	.81
Woman	1.310 (0.977-1.756)	.07
Man	REF	—
Educational attainment		
Greater than bachelors	1.562 (1.01-2.422)	.05
Bachelors	1.545 (1.037-2.305)	.03
Less than bachelors	REF	—
Patient activation		
PAM[c,d]	0.999 (0.987-1.01)	.82
Mental health literacy		
MHLS[c,e]	1.022 (0.999-1.045)	.07
Well-being		
WHO-5[c,f]	1.003 (0.994-1.013)	.51
Primary concern		
My emotions	REF	—
My finances	0.864 (0.409-1.871)	.70
My physical well-being	0.926 (0.577-1.495)	.75
My professional life	0.779 (0.52-1.171)	.23
My relationships	0.808 (0.568-1.151)	.24
Mental health at baseline		
PHQ-9[c,g]	0.975 (0.938-1.013)	.19
GAD-7[c,h]	1.005 (0.968-1.044)	.79

[a]REF: reference range.
[b]Not applicable.
[c]PAM, MHLS, WHO-5, PHQ-9, and GAD-7 scores were entered as linear predictors.
[d]PAM: Patient Activation Measure.
[e]MHLS: Mental Health Literacy Scale.
[f]WHO-5: 5-item World Health Organization Well-Being Index.
[g]PHQ-9: 9-item Patient Health Questionnaire.
[h]GAD-7: 7-item Generalized Anxiety Disorder Questionnaire.

XSL•FO
RenderX

Table 3. Adjusted logistic regression analysis predicting any self-directed digital resources use (N=908).

Variable	Odds ratio (95% CI)	*P* value
Age	1.026 (1.009-1.044)	.003
Race and ethnicity		
White	1.115 (0.841-1.48)	.45
Non-White	REF[a]	—[b]
Gender identity		
Non-binary	1.122 (0.55-2.22)	.75
Woman	1.454 (1.079-1.965)	.01
Man	REF	—
Educational attainment		
Greater than bachelors	1.648 (1.045-2.625)	.03
Bachelors	1.564 (1.026-2.414)	.04
Less than bachelors	REF	—
Patient activation		
PAM[c,d]	0.995 (0.983-1.007)	.40
Mental health literacy		
MHLS[c,e]	1.003 (0.98-1.027)	.82
Well-being		
WHO-5[c,f]	1.008 (0.998-1.017)	.11
Primary concern		
My emotions	REF	—
My finances	1.143 (0.524-2.43)	.73
My physical well-being	1.432 (0.894-2.292)	.13
My professional life	1.158 (0.769-1.737)	.48
My relationships	0.989 (0.689-1.414)	.95
Mental health at baseline		
PHQ-9[c,g]	0.992 (0.954-1.031)	.68
GAD-7[c,h]	1.004 (0.967-1.043)	.83

[a]REF: reference.
[b]Not applicable.
[c]PAM, MHLS, WHO-5, PHQ-9, and GAD-7 scores were entered as linear predictors.
[d]PAM: Patient Activation Measure.
[e]MHLS: Mental Health Literacy Scale.
[f]WHO-5: 5-item World Health Organization Well-Being Index.
[g]PHQ-9: 9-item Patient Health Questionnaire.
[h]GAD-7: 7-item Generalized Anxiety Disorder Questionnaire.

XSL•FO
RenderX

Table 4. Unadjusted and adjusted Poisson regression models predicted the number of self-directed digital resources used.

Variable	Unadjusted model (N=950)		Adjusted model (N=911)	
	RR[a] (95% CI)	*P* value	RR (95% CI)	*P* value
Coaching or therapy				
Any coach sessions	1.716 (1.542-1.908)	<.001	1.820 (1.629-2.033)	<.001
Any therapy sessions	1.414 (1.247-1.6)	<.001	1.795 (1.554-2.068)	<.001
Age	__[b]	—	1.039 (1.033-1.045)	<.001
Race and ethnicity				
White	—	—	1.120 (0.998-1.258)	.06
Non-white	—	—	REF[c]	—
Gender identity				
Nonbinary	—	—	1.794 (1.425-2.236)	.95
Woman	—	—	1.230 (1.091-1.388)	<.001
Man	—	—	REF	—
Educational attainment				
Less than bachelors	—	—	REF	—
Greater than bachelors	—	—	0.991 (0.83-1.186)	.92
Bachelors	—	—	1.147 (0.975-1.356)	.10
Patient activation				
PAM[d,e] score	—	—	0.991 (0.987-0.996)	<.001
Mental health literacy				
MHLS[d,f] score	—	—	1.000 (0.991-1.01)	.97
Well-being				
WHO-5[d,g] score	—	—	1.001 (0.998-1.004)	.52
Primary concern				
My emotions	—	—	REF	—
My finances	—	—	1.456 (1.058-1.954)	.02
My physical well-being	—	—	2.257 (1.925-2.64)	<.001
My professional life	—	—	1.713 (1.457-2.009)	<.001
My relationships	—	—	1.365 (1.174-1.584)	<.001

[a]RR: rate ratio.
[b]Not available.
[c]REF: reference.
[d]PAM, MHLS, and WHO-5 scores were entered as linear predictors.
[e]PAM: Patient Activation Measure.
[f]MHLS: Mental Health Literacy Scale.
[g]WHO-5: 5-item World Health Organization Well-Being Index.

XSL•FO
RenderX

Table 5. Subgroup Poisson regression analysis conditional on elevated GAD-7 or PHQ-9 score (≥ 10) at baseline, predicting number of self-directed resources used (n=400).

Variable	RR[a] (95% CI)	*P* value
Use of therapy or coach		
Any coach visits	2.099 (1.76-2.503)	<.001
Any therapist visits	1.778 (1.455-2.165)	<.001
Age	1.036 (1.025-1.047)	<.001
Race and ethnicity		
Non-White	REF[b]	—[c]
White	1.218 (1.014-1.467)	.04
Gender identity		
Man	REF	—
Woman	0.864 (0.71-1.053)	.14
Non-binary	1.266 (0.859-1.819)	.22
Educational attainment		
<Bachelors	REF	—
>Bachelors	1.005 (0.758-1.339)	.97
Bachelors	1.203 (0.943-1.553)	.15
Patient activation		
PAM[d,e]	0.985 (0.977-0.993)	<.001
Mental health literacy		
MHLS[d,f]	1.008 (0.993-1.023)	.31
Well-being		
WHO-5[d,g]	0.993 (0.987-0.999)	.03
Primary concern		
My emotions	REF	—
My finances	2.015 (1.202-3.191)	.005
My physical well-being	2.411 (1.865-3.095)	<.001
My professional life	2.318 (1.814-2.945)	<.001
My relationships	1.350 (1.04-1.735)	.02

[a]RR: rate ratio.
[b]REF: reference range.
[c]Not applicable.
[d]PAM, MHLS, and WHO-5 scores were entered as linear predictors.
[e]PAM: Patient Activation Measure.
[f]MHLS: Mental Health Literacy Scale.
[g]WHO-5: 5-item World Health Organization Well-Being Index.

Discussion

Principal Findings

Our study aimed to evaluate predictive factors of engagement in multiple digital health modalities offered by a digital mental health employer-based benefit. Over half of the participants (55.9%, 531/950) used at least 1 component of the digital mental health platform at least once. It is important to note that due to the nature of employer benefits programs, employers often choose to market their mental health services benefit at specific time points, such as open enrollment periods, which leads to an increase in sign-ups. However, not everyone needs or wants mental health care at the time they receive the email advertising the benefit. Thus, it is possible that people initiate with a mental health digital platform before they need or are ready to engage in care. Or, if they do not sign up at the time of the email, they may forget about the program at a time when they do need care. These findings underscore the need to continue adapting digital health and other digital resources in ways that engage broader

XSL•FO
RenderX

populations of users curious about mental health resources and ensure that people know how to access care if they need it. Our study did identify several factors that were associated with increased use of the platform, including the use of a coach or therapist, older age, identifying as a woman, baseline symptom severity, and presenting primary concerns of finances, physical well-being, professional life, and relationships.

A key finding of our study was that the use of coaching or therapy was associated with increased self-directed digital resource use. This relationship was even stronger for individuals with elevated GAD-7 and PHQ-9 scores. Higher self-directed use in this population may be the result of provider encouragement. Coaches and therapists seeing participants with elevated GAD-7 and PHQ-9 scores might be inclined to recommend additional use of digital health resources outside of one-on-one sessions, resulting in more self-directed use for those groups. Thus, additional education for platform mental health coaches and therapists about available digital support tools and their quality and effectiveness might facilitate greater user engagement in app-based resources. Training for the network of platform providers could also include information on the barriers and facilitators that sustain engagement in digital resources. Indeed, previous research has highlighted the impact clinicians can have on digital health interest, use, and engagement [29-31]. While clinicians may be recommending self-directed strategies and driving use, alternatively, it may be that when a participant engages in 1 modality of the platform, they are more likely to become "super users" across multiple modalities. Future research is needed to disentangle the relationship between live support and self-directed resource engagement.

Furthermore, certain topics of focus were more associated with self-directed resource use, including finances, physical well-being, professional life, and relationships. Participants whose primary chosen topic was "[their] emotions" were less likely to engage with self-directed digital resources, though almost half (48.1%, n=224) of this group used one-on-one modalities. Given that higher baseline depression was also associated with less self-directed digital engagement, perhaps features of depression, such as lack of motivation, energy, or concentration, could decrease the likelihood of using these resources, even when the need for them is higher. This result is consistent with findings from a systematic review that human factors, such as severe mental health symptoms, can be a barrier to engagement [15]. This may also be a feature of the personalized care recommendations, which were more likely to suggest users work with a provider one-on-one if they were experiencing higher levels of distress. Future research may be helpful in determining how to make digital health self-directed resources more appealing for individuals struggling with low mood who are not already engaging in therapy or coaching. Research is also needed to better determine the circumstances when self-guided care is a clinically effective stand-alone approach for people with depressive symptoms.

Several sociodemographic characteristics, including identifying as a woman, higher education, and older age, were also associated with greater self-guided resource use. The extant literature has demonstrated that women are more likely to receive mental health treatment in general [32,33] and to use digital mental health resources specifically [34]. Our finding extends previous literature by revealing that at the intersection of digital health and mental health treatment, women users are more likely to engage with the technology than men. Higher education was also related to increased use of self-directed resources, as well as overall increased use of the platform more generally. Some research indicates that education is a significant predictor of mental health service use [35-37]. Finally, older age was associated with more self-directed resource use, though given the restricted age range of this sample, this result may be more challenging to interpret. It is possible that people in their 30s and 40s are more time-constrained by career and caregiving responsibilities and may be more likely to engage in mental health services they can access when it is convenient to them.

This study had several limitations with regard to the sample and restrictions of data. First, participants were eligible for the study if they had matched with a provider or used at least 1 digital mental health resource. This inclusion criteria likely selected participants who were more predisposed to engaging with digital mental health resources, possibly biasing our sample to more resource use. In addition, the digital mental health resource use that made participants eligible for inclusion in the study is included in the total number of times a participant used these resources, likely inflating the count of times used. Next, our sample may not be generalizable to other samples, given that we used participants from companies that were contracted with 1 mental health platform. Thus, the sample was limited to working adults with benefits-eligible positions, which skews towards a higher level of socioeconomic status.

Another limitation of this study was our inability to determine the sequence of events for each participant, for example, whether they engaged in a coaching session and then self-directed resource use or vice versa. This information would be helpful in further breaking down predictors of digital health use. For example, do sessions with a coach then lead to subsequent self-directed resource use, or does an attempt at self-directed resources lead to a request for additional live support by coaching or therapy? Future research can disentangle the timing of the use of various modalities. Third, the platform design and user experience may have driven engagement to some extent, and future research is needed to explore intervention-level factors (eg, ease of use, perceived effectiveness) that may affect treatment engagement.

Next, we used a definition of engagement based on the number of times a participant used the self-directed digital mental health resources. In addition to the volume of care used, it is important to understand whether users are engaging in the right level of care based on their needs. Although we found alignment between self-guided resource use and user needs (eg, users with lower depression symptoms and less mental health-focused topics of concern engaged in resources on their own, while users with higher depression and anxiety were more likely to engage with resources in combination with one-on-one care), additional data could provide a more nuanced look at engagement. For example, future studies could examine whether users engaged in their recommended care pathway and if they saw improved mental health outcomes as a result of engaging.

XSL•FO
RenderX

Conclusion

Digital mental health platforms, including those using a stepped-care model, have the potential to deliver appropriate and evidence-based care to large populations of individuals in need. However, these platforms are affected by similar challenges related to engagement that plague many of the high-quality, evidence-based psychological treatment modalities currently available. This study revealed that users were more likely to use self-directed digital mental health resources if they were also engaged in at least 1 coaching or therapy session, which suggests that the key to successful engagement with digital therapeutics may lie in additional live support.

Acknowledgments

We would like to thank Camille Welcome Chamberlain for contributing to this project. We would also like to thank the platform members and providers for their contributions to this work, as well as the employers who made this mental health benefit available to their employees. The authors received research support from Modern Health to conduct this work. The funder had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Data Availability

The data set analyzed in this study is not publicly available as it is proprietary information from Modern Health, but are available from the corresponding author upon reasonable request.

Authors' Contributions

MN contributed to conceptualization, formal analysis, methodology, and writing-original draft. MM contributed to conceptualization, methodology, and writing–original draft. LB contributed to writing the original draft. EK contributed to supervision, writing-review and editing. SSH contributed to investigation, writing–review and editing. CCS contributed to investigation, writing–review and editing. EK contributed to conceptualization, methodology, supervision, and writing–original draft.

Conflicts of Interest

SSH and CSS were employed by Modern Health and received salary and stock options at the time this study was conducted.

References

1. Kamal R, Cox C, Rousseau D, Kaiser Family Foundation. Costs and outcomes of mental health and substance use disorders in the US. JAMA 2017 Aug 01;318(5):415. [doi: 10.1001/jama.2017.8558] [Medline: 28763535]
2. Reeves WC, Strine TW, Pratt LA, Thompson W, Ahluwalia I, Dhingra SS, Centers for Disease Control Prevention (CDC). Mental illness surveillance among adults in the United States. MMWR Suppl 2011 Sep 02;60(3):1-29 [FREE Full text] [Medline: 21881550]
3. Wang PS, Lane M, Olfson M, Pincus HA, Wells KB, Kessler RC. Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):629-640 [FREE Full text] [doi: 10.1001/archpsyc.62.6.629] [Medline: 15939840]
4. Coombs NC, Meriwether WE, Caringi J, Newcomer SR. Barriers to healthcare access among U.S. adults with mental health challenges: A population-based study. SSM Popul Health 2021 Sep;15:100847 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100847] [Medline: 34179332]
5. Olfson M. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. Health Aff (Millwood) 2016 Jun 01;35(6):983-990 [FREE Full text] [doi: 10.1377/hlthaff.2015.1619] [Medline: 27269013]
6. Thomas D, Macdowell M, Glasser M. Rural mental health workforce needs assessment - a national survey. Rural Remote Health 2012;12:2176 [FREE Full text] [Medline: 23088609]
7. Agarwal S, Jalan M, Wilcox HC, Sharma R, Hill R, Pantalone E, Thrul J, Rainey JC, Robinson KA. Evaluation of Mental Health Mobile Applications. Agency for Healthcare Research and Quality (US) technical brief, No. 41. 2022 May. URL: https://www.ncbi.nlm.nih.gov/books/NBK580948/ [accessed 2023-07-02]
8. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014 Apr;44(6):1303-1317 [FREE Full text] [doi: 10.1017/S0033291713001943] [Medline: 23931656]
9. Mojtabai R, Olfson M, Sampson NA, Jin R, Druss B, Wang PS, et al. Barriers to mental health treatment: results from the National Comorbidity Survey Replication. Psychol Med 2011 Aug;41(8):1751-1761 [FREE Full text] [doi: 10.1017/S0033291710002291] [Medline: 21134315]
10. Akter S, D'Ambra J, Ray P. Service quality of mHealth platforms: development and validation of a hierarchical model using PLS. Electron Markets 2010 Nov 18;20(3-4):209-227 [FREE Full text] [doi: 10.1007/s12525-010-0043-x]
11. Krebs P, Duncan DT. Health app use among us mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 04;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]

XSL•FO
RenderX

12. Kayyali R, Peletidi A, Ismail M, Hashim Z, Bandeira P, Bonnah J. Awareness and use of mHealth Apps: A study from England. Pharmacy (Basel) 2017 Jun 14;5(2):33 [FREE Full text] [doi: 10.3390/pharmacy5020033] [Medline: 28970445]
13. Kullgard N, Holmqvist R, Andersson G. Premature dropout from psychotherapy: prevalence, perceived reasons and consequences as rated by clinicians. Clin Psychol Eur 2022 Jun;4(2):e6695 [FREE Full text] [doi: 10.32872/cpe.6695] [Medline: 36397946]
14. Leichsenring F, Sarrar L, Steinert C. Drop-outs in psychotherapy: a change of perspective. World Psychiatry 2019 Feb;18(1):32-33 [FREE Full text] [doi: 10.1002/wps.20588] [Medline: 30600632]
15. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
16. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR Mhealth Uhealth 2019 Aug 29;7(8):e12983 [FREE Full text] [doi: 10.2196/12983] [Medline: 31469081]
17. Rai A, Chen L, Pye J, Baird A. Understanding determinants of consumer mobile health usage intentions, assimilation, and channel preferences. J Med Internet Res 2013 Aug 02;15(8):e149 [FREE Full text] [doi: 10.2196/jmir.2635] [Medline: 23912839]
18. Kim M, Kim C, Kim E, Choi M. Effectiveness of mobile health-based exercise interventions for patients with peripheral artery disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Feb 15;9(2):e24080 [FREE Full text] [doi: 10.2196/24080] [Medline: 33587042]
19. Doherty K, Doherty G. Engagement in HCI: Conception, theory and measurement. ACM Comput. Surv 2018 Nov 19;51(5):1-39. [doi: 10.1145/3234149]
20. Sidner CL, Kidd CD, Lee C, Lesh N. Where to look: a study of human-robot engagement. : ACM; 2004 Presented at: Proceedings of the 9th international conference on Intelligent user interfaces; January 13, 2004; Funchal, Madeira , Portugal p. 78-84. [doi: 10.1145/964442.964458]
21. Kozlov E, Bantum E, Pagano I, Walser R, Ramsey K, Taylor K, et al. The reach, use, and impact of a free mhealth mindfulness app in the general population: mobile data analysis. JMIR Ment Health 2020 Nov 27;7(11):e23377 [FREE Full text] [doi: 10.2196/23377] [Medline: 33245289]
22. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
23. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
24. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
25. Green CA, Perrin NA, Polen MR, Leo MC, Hibbard JH, Tusler M. Development of the Patient Activation Measure for mental health. Adm Policy Ment Health 2010 Jul;37(4):327-333 [FREE Full text] [doi: 10.1007/s10488-009-0239-6] [Medline: 19728074]
26. O'Connor M, Casey L. The Mental Health Literacy Scale (MHLS): A new scale-based measure of mental health literacy. Psychiatry Res 2015 Sep 30;229(1-2):511-516. [doi: 10.1016/j.psychres.2015.05.064] [Medline: 26228163]
27. Giebel C, Lord K, Cooper C, Shenton J, Cannon J, Pulford D, et al. A UK survey of COVID-19 related social support closures and their effects on older people, people with dementia, and carers. Int J Geriatr Psychiatry 2021 Mar;36(3):393-402 [FREE Full text] [doi: 10.1002/gps.5434] [Medline: 32946619]
28. Shiwaku H, Doi S, Miyajima M, Matsumoto Y, Fujino J, Hirai N, et al. Novel brief screening scale, Tokyo Metropolitan Distress Scale for Pandemic (TMDP), for assessing mental and social stress of medical personnel in COVID-19 pandemic. Psychiatry Clin Neurosci 2021 Jan;75(1):24-25 [FREE Full text] [doi: 10.1111/pcn.13168] [Medline: 33225480]
29. Kozlov E, McDarby M, Pagano I, Llaneza D, Owen J, Duberstein P. The feasibility, acceptability, and preliminary efficacy of an mHealth mindfulness therapy for caregivers of adults with cognitive impairment. Aging Ment Health 2022 Oct;26(10):1963-1970 [FREE Full text] [doi: 10.1080/13607863.2021.1963949] [Medline: 34428993]
30. Yang WE, Shah LM, Spaulding EM, Wang J, Xun H, Weng D, et al. The role of a clinician amid the rise of mobile health technology. J Am Med Inform Assoc 2019 Nov 01;26(11):1385-1388 [FREE Full text] [doi: 10.1093/jamia/ocz131] [Medline: 31373364]
31. Simblett S, Pennington M, Quaife M, Theochari E, Burke P, Brichetto G, et al. Key drivers and facilitators of the choice to use mhealth technology in people with neurological conditions: observational study. JMIR Form Res 2022 May 23;6(5):e29509 [FREE Full text] [doi: 10.2196/29509] [Medline: 35604761]
32. Cunningham R, Crowe M, Stanley J, Haitana T, Pitama S, Porter R, et al. Gender and mental health service use in bipolar disorder: national cohort study. BJPsych Open 2020 Nov 06;6(6):e138 [FREE Full text] [doi: 10.1192/bjo.2020.117] [Medline: 33153508]
33. Terlizzi E, Zablotsky B. Mental health treatment among adults: United States, 2019. Hyattsville, MD: National Center for Health Statistics; 2020. URL: https://www.cdc.gov/nchs/data/databriefs/db380-H.pdf#:~:text=In%202019%2C%2019.2%25%20of%20adults%20had%20received%20any,men%20to%20have%20received%20any%20mental%20health%20treatment [accessed 2024-10-11]

XSL•FO
RenderX

34. Smail-Crevier R, Powers G, Noel C, Wang J. Health-related internet usage and design feature preference for e-mental health programs among men and women. J Med Internet Res 2019 Mar 18;21(3):e11224 [FREE Full text] [doi: 10.2196/11224] [Medline: 30882361]
35. Johnson EM, Possemato K. Correlates and predictors of mental health care utilization for veterans with PTSD: A systematic review. Psychol Trauma 2019 Nov;11(8):851-860 [FREE Full text] [doi: 10.1037/tra0000461] [Medline: 30998061]
36. Smith LD, McGovern RJ, Peck PL. Factors contributing to the utilization of mental health services in a rural setting. Psychol Rep 2004 Oct;95(2):435-442. [doi: 10.2466/pr0.95.2.435-442] [Medline: 15587204]
37. Steele LS, Dewa CS, Lin E, Lee KLK. Education level, income level and mental health services use in Canada: associations and policy implications. Healthc Policy 2007 Aug;3(1):96-106 [FREE Full text] [Medline: 19305758]

Abbreviations

GAD-7: 7-item Generalized Anxiety Disorder Questionnaire
MHLS: Mental Health Literacy Scale
PAM-MH: Patient Activation Measure for Mental Health
PHQ-9: 9-item Patient Health Questionnaire
RR: rate ratio
WHO-5: 5-item World Health Organization Well-Being Index

Edited by T de Azevedo Cardoso; submitted 03.05.23; peer-reviewed by L Balcombe, D Adler; comments to author 14.11.23; revised version received 07.02.24; accepted 13.09.24; published 07.11.24.

Please cite as:
Nowels MA, McDarby M, Brody L, Kleiman E, Sagui Henson S, Castro Sweet C, Kozlov E
Predictors of Engagement in Multiple Modalities of Digital Mental Health Treatments: Longitudinal Study
J Med Internet Res 2024;26:e48696
URL: https://www.jmir.org/2024/1/e48696
doi:10.2196/48696
PMID:

©Molly Aideen Nowels, Meghan McDarby, Lilla Brody, Evan Kleiman, Sara Sagui Henson, Cynthia Castro Sweet, Elissa Kozlov. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of an Internet-Based Self-Help Acceptance and Commitment Therapy Program on Medical Students' Mental Well-Being: Follow-Up Randomized Controlled Trial

Difan Wang[1,2*], MEd; Bingyan Lin[3*], PhD; Shuangxi Zhang[4], PhD; Wei Xu[2], PhD; Xinying Liu[5], MA

[1]Department of Internal Medicine, Psychological Counseling and Service Center, Graduate School of Medical College of Chinese PLA General Hospital, Beijing, China

[2]Faculty of Psychology, Beijing Normal University, Beijing, China

[3]Department of Primary and Long-term Care, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[4]School of Education, Hebei University, Baoding, China

[5]Department of Otolaryngology Head and Neck Surgery, The Third Medical Center of Chinese PLA General Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Wei Xu, PhD
Faculty of Psychology
Beijing Normal University
19 Xinjiekouwai Street Haidian District
Beijing, 100875
China
Phone: 86 010 66939114
Email: livingxw@163.com

Abstract

Background: Psychological distress is a growing problem among medical students worldwide. This highlights the need for psychological interventions to focus on mental health and improve well-being in this population.

Objective: This study developed an internet-based, self-help, acceptance and commitment therapy program (iACT 2.0), aiming to examine its effectiveness in reducing depression, anxiety, stress, psychological inflexibility (PI), and obsessive-compulsive symptoms (OCSs) among medical students.

Methods: A total of 520 Chinese postgraduate medical students were randomly assigned to either an iACT 2.0 intervention group (n=260; six online lessons, once every 5 days) or a control condition (n=260; without intervention). Participants completed questionnaires including the 21-item Depression Anxiety Stress Scale, the revised Obsessive-Compulsive Inventory, and the Multidimensional Psychological Flexibility Inventory at the preintervention (T1), postintervention (T2), and 1-month follow-up time points (T3). No therapist support was provided during the 1-month iACT 2.0 intervention period. Data were collected via an online platform and analyzed using repeated-measures ANOVA.

Results: Participants in the intervention group demonstrated a significant decrease in depression, anxiety, stress, PI, and OCSs compared to the control group after the intervention (F=22.9-672.04, all P<.001). Specifically, the intervention group showed significant reductions in all measured outcomes from the preintervention to postintervention time point and at the 1-month follow-up (all P<.001). In contrast, no significant changes were observed in the control group over the same period (all P>.05). The groups did not differ significantly at baseline (all P>.05). Significant differences were noted at both the postintervention and follow-up time points (all P<.001).

Conclusions: This study demonstrated that the newly developed iACT 2.0 was effective in reducing depression, anxiety, stress, PI, and OCSs. Notably, the positive effects of the intervention persisted at the 1-month follow-up. This program can offer a useful addition to existing mental illness treatment and lead to improvements in clinical and psychotherapy planning while simultaneously reducing the burden on traditional counseling and services.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300070725; https://tinyurl.com/2h75wx8n

(J Med Internet Res 2024;26:e50664) doi:10.2196/50664

XSL•FO
RenderX

KEYWORDS

internet-delivered self-help acceptance and commitment therapy; depression; anxiety; stress; psychological inflexibility; obsessive-compulsive symptoms; medical students; iACT 2.0 program

Introduction

Background

Psychological distress (depression, anxiety, and stress) is highly prevalent among medical students, raising significant concerns [1-3]. Research indicates that medical students experience significantly higher levels of anxiety, depression, and stress compared to their age-matched peers in the general population [4]. For example, a meta-analysis revealed that approximately 1 in 3 medical students worldwide experiences anxiety, a rate significantly exceeding that of the general population [5]. Moreover, medical students exhibit a notable prevalence rate of 27.2% for depression, compared to 9.3% in the general populace [6]. The demanding journey through medical school frequently triggers substantial stress and anxiety [7-10]. Anxiety involves anticipating future threats, often accompanied by muscle tension and heightened vigilance, while depression is characterized by feelings of sadness or emptiness, impairing effective functioning [11]. Stress, defined as feeling unable to meet requirements or expectations, manifests in physical, psychological, or social complaints [12,13].

Previous studies highlight obsessive-compulsive disorder (OCD) as a significant mental health concern among Chinese medical students [14]. The prevalence of symptoms suggestive of OCD among medical students surpasses that of the general population and is correlated with depressive symptoms [15]. Additionally, the COVID-19 pandemic has exacerbated OCD prevalence among medical students [16,17]. Additionally, enforced prevention measures such as repeated disinfection and strict hand hygiene heighten their sense of compulsion. OCD is characterized by the presence of obsessions and compulsions. Obsessions refer to recurrent and persistent thoughts, urges, or images that are intrusive and undesired, whereas compulsions are repetitive behaviors or mental actions that an individual feels driven to perform in response to an obsession or according to rules that must be applied rigidly [11]. Therefore, based on existing literature, depression, anxiety, stress, and OCD are prominent emotional distresses among medical students, warranting further attention.

Psychological distress presents potential risks for both patients and medical students, raising concerns about mental health challenges within this group [2]. Poor mental health can lead to unprofessional behavior, including difficulties in decision-making and increased errors [18,19], reduced attention and concentration [19], and loss of empathy [20], ultimately affecting education, employment, and social relationships and causing poorer quality of life, lower self-esteem, and a sense of hopelessness [21]. Moreover, OCD can seriously interfere with an individual's social and occupational functioning, potentially affecting academic performance [22,23], social interactions, and cognitive abilities [24], thereby impeding normal life and work.

Despite the high prevalence and risk of psychological distress among medical students, many medical students with mental health problems do not receive treatment. Research indicates that only 22% of medical students who are depressed actually access mental health services [4]. Common barriers that prevent students from using these services include time constraints, limited accessibility, concerns about confidentiality, stigma, fear of unwanted interventions, and cost [4,25]. Consequently, it is crucial to use all available resources to develop sound innovations to tackle these challenges effectively.

Potential of Acceptance and Commitment Therapy in the Treatment of Mental Illness Among Medical Students

While medical students face challenges in managing mental health problems, and receive limited and inadequate behavioral health education and training [26,27], studies indicate that they have better recognition of mental illnesses compared to students in other disciplines [28,29]. Acceptance and commitment therapy (ACT) emphasizes the willingness to experience unwanted thoughts and emotions and contact with the present moment rather than avoid emotional distress [30], and it involves mindfulness and acceptance processes and behavior change processes [31]. The primary goal of ACT is to improve psychological flexibility (PF)—the ability to pursue value—and live a meaningful life in the presence of discomfort and other unwanted inner experiences [30,32].

A growing body of evidence indicates that ACT can alleviate depressive symptoms, anxiety, stress, and emotional distress in various populations [31,33,34]. For instance, prior meta-analyses have shown that ACT yields small to medium effects in reducing depressive symptoms, anxiety, and stress among family caregivers [34]. Positive mental health outcomes have also been observed among university students [35,36], parents [37], adolescents [38], and others. Moreover, ACT has demonstrated effectiveness in addressing psychological distress while achieving high satisfaction rates [39,40]. Notably, Pang et al [41] conducted a small-scale randomized controlled trial (n=22) involving clinical medical students in Borneo to assess the efficacy of a 1-day, therapist-delivered ACT intervention in alleviating performance anxiety. The intervention group exhibited significant improvements in anxiety immediately after the intervention and at the 1-month follow-up, indicating the potential for conducting a larger randomized controlled trial to verify the efficacy of ACT among medical students.

Furthermore, ACT is one of the newest treatment strategies that has been rapidly developed to improve the treatment of patients with OCD, by teaching individuals to shift their relationship with obsessive thinking rather than resist or reduce the number of obsessions and compulsions [42-44]. Studies have consistently found that ACT can improve the PF of patients with OCD, which in turn predicts the reduction of obsessive-compulsive symptoms (OCSs) [45-47]. Furthermore,

XSL•FO
RenderX

ACT has been shown to be as effective as drug therapy in improving the quality of life of patients with OCD [42].

Advantages of Internet-Based Interventions

Internet-based psychological interventions have gained popularity and proven to be beneficial [48-50]. Notably, ACT-based, internet-delivered treatments have been developed for health anxiety [51,52] and internet-based cognitive behavioral therapy has been used for OCD treatment [53]. Online interventions offer benefits such as breaking distance and time limits [54], providing more privacy and flexibility [36], and allowing access from anywhere at any time for rural medical students [7].

Although previous systematic reviews and meta-analyses have reported the effectiveness of internet-based ACT (iACT) programs in reducing depressive symptoms, anxiety, stress, and psychological distress across diverse populations [31], few studies explore the impact of iACT interventions on medical students, particularly those in China. Given their familiarity and regular use of digital technology and the internet for accessing health-related information [55], there is a compelling rationale for developing a highly feasible and engaging intervention program that can be perceived as a means of enhancing the mental health of Chinese medical students.

Objective and Our Study

Cultural adaptation involves the systematic modification of an evidence-based treatment to align with the client's language, culture, and context, ensuring compatibility with their cultural values, including factors such as interdependence and spirituality [56,57]. Research consistently shows that culturally adapted treatments are more effective, acceptable, and satisfying for patients [58-60]. Studies have also demonstrated that participants find culturally adapted ACT enjoyable and acceptable [58,61,62]. Therefore, considering cultural adaptation can be highly beneficial in our ACT interventions.

Furthermore, traditional ACT typically requires multiple sessions to systematically explore various facets of PF and psychological inflexibility (PI), which can impose significant economic and time constraints. Research has shown that these constraints often serve as barriers for medical students in accessing psychological services [4,25]. Given the proven effectiveness of single-session counseling [63], it is both logical and timely to develop an intervention that has the potential to evolve into a single-session therapy in the future.

In this study, we developed an internet-based, self-help ACT program (iACT 2.0) by incorporating cultural and spiritual elements into traditional ACT. Each session follows the 4 steps of "启-承-转-合" (means starting-proceeding-transforming-integrating in Chinese) to illustrate the process of PF and PI, with each step representing its corresponding therapeutic goal within a single session. Additionally, we used the Bagua diagram from traditional Chinese Daoism to illustrate these 4 steps, symbolizing the continuous change and thought patterns, thereby enhancing the relatability and comprehensibility of the intervention. Unlike traditional ACT, we introduced all 6 facets in the first lesson and used subsequent sessions to reinforce core ACT concepts with diverse examples. This approach encourages participants to initially focus on their most inflexible facet and potentially find solutions during the first session, facilitating problem-solving and promoting flexibility through internal exploration. The core content was delivered in 6 online lessons, spaced 5 days apart, to systematically reinforce PF. To maximize engagement, participants were asked to practice daily mindfulness during the intervention. Importantly, no direct therapist interaction was included, allowing us to isolate the effects of the online program itself. Self-guided internet-based therapies have shown positive results in treating depression [64]. In conclusion, our intervention was designed to enhance accessibility and acceptance for participants.

To the best of our knowledge, this study represents the pioneering application of iACT to a cohort of Chinese postgraduate medical students. The objective of our study was to evaluate the efficacy of iACT 2.0, an internet-based, self-help ACT program, in reducing depression, anxiety, stress, OCSs, and PI in medical students. We hypothesized that iACT 2.0 would yield significant improvements in primary outcomes when compared to the waitlist controls, with these improvements sustained for a 1-month follow-up.

Methods

Study Design

This study was designed as a single-blinded randomized controlled trial consisting of 2 parallel groups with a blocked 1:1 allocation. During recruitment, all participants were informed about the intervention but not their group assignment, ensuring their unawareness of group allocation throughout this study. Maintaining participant blinding was crucial to prevent potential bias in questionnaire responses, particularly for those in the control group. The randomization was performed by a research assistant who was not part of this study. The researcher remained unaware of group assignments until after finalizing randomization. The trial was conducted from April 2023 to June 2023. To ensure adherence to randomized controlled trial guidelines, the trial was registered in the Chinese Clinical Trials Registry under the registration ChiCTR2300070725.

Participants

Medical students from the Graduate School of Medical College of Chinese PLA Hospital were recruited in February 2023 via WeChat (Tencent) advertisements, stating that there is a newly developed psychotherapy program aimed at reducing psychological distress. It also stated that the intervention would be delivered via the internet and that participants would be expected to practice daily. Participants can participate in this study by searching the name of our WeChat Mini Program and then logging in to the platform to complete an online screening. Sign-up for this study was open from April 15 to 23, 2023. This study's inclusion criteria were (1) being a postgraduate student enrolled at the Graduate School of Medical College of Chinese PLA Hospital; (2) being aged 18 years and older; and (3) having 21-item Depression Anxiety Stress Scale (DASS-21) scores indicating at least mild depression, anxiety, and stress levels (depression subscale ≥10, anxiety subscale ≥8, and stress subscale ≥15) [65,66]. The exclusion criteria were (1) individuals who were currently receiving psychotherapy, (2)

XSL•FO
RenderX

organic mental disorder or schizophrenia, (3) visual or hearing impairment, (4) severe drug or alcohol use disorders, and (5) pregnant and lactating women.

A total of 563 students expressed interest in participating in this study. Before the preintervention measurement, all participants were provided with an online written consent form to review and sign. On review of eligibility criteria, 43 participants were excluded from this study, leaving a final sample of 520 participants. Participants were randomly assigned to either the iACT 2.0 intervention group (iACT group, n=260) or the waitlist control group (control group, n=260) at a 1:1 allocation ratio. The randomization process was conducted by an independent individual, unaffiliated with this study, who used computer-generated pseudorandom numbers to allocate participants, ensuring the concealment of allocation from participants.

In total, 17 participants in the iACT group and 14 participants in the control group did not complete the posttest, while 3 participants in the iACT group and 8 participants in the control condition did not complete the follow-up test, Therefore, the final number of participants in this study was 478, with a dropout rate of 8% (42/520). The flow diagram is presented in the *Results* section.

Procedures

Participants in this study were assessed at the preintervention (T1), postintervention (T2), and 1-month follow-up time points (T3). The pre- to postintervention comparison investigated whether the iACT 2.0 had been effective in comparison to the no-intervention control condition (control). The 1-month follow-up studied the maintenance of the intervention effects. At the preintervention measurement, participants were asked to fill out the relevant online questionnaires and sociodemographic information. When the preintervention measurement was finished, participants in the intervention group were given access to the online intervention lesson, including the link, username, and password. The postintervention measurement was conducted 30 days after the start of the intervention by filling in the questionnaires. Participants were asked to complete the same questionnaires again for the 1-month follow-up measurement. Control participants were given access to the intervention lesson after the follow-up test.

Intervention

The 1-month internet intervention (Textbox 1; see Table S1 in Multimedia Appendix 1 for detailed intervention protocol) was delivered through a patented online system (Figure 1A and B), without any personal support provided. The intervention was based on the 6 facets of ACT and consisted of 6 online lessons delivered in various media formats, including text, video, audio, and illustrations. Participants in the iACT group were instructed to practice daily mindfulness for 30 days (20-30 minutes per day) to increase their awareness. The lessons were released sequentially, with access to the next lesson granted on completion of the previous lesson. The control group did not receive any intervention.

The intervention process was designed to align with the principles of ancient Chinese Daoism and used the 4 steps of “启-承-转-合” (which means starting-proceeding-transforming-integrating in Chinese; Figure 1C). This iterative process introduced the 6 facets of PF and PI in the initial lesson, followed by increasingly sophisticated examples of these facets in subsequent lessons. Participants were instructed to proactively identify areas of PI and select the appropriate facet to apply to their daily lives. This process enabled participants to enhance their self-awareness and cultivate greater PF through the application of ACT principles.

XSL•FO
RenderX

Textbox 1. Overview of the internet-based, self-help, acceptance and commitment therapy program (iACT 2.0) treatment content showing time and content summary.

Lesson 1 (day 1)

- Start: Set objectives.
- Proceed: Review their previous coping styles.
- Transform:
  - Elucidate the 6 facets of psychological flexibility (PF) and psychological inflexibility (PI) of acceptance and commitment therapy (ACT).
  - Mindfulness practice: body scan in mindfulness.
- Integrate: Choose one facet of PF and use it as a coping strategy in the following 5 days.
- Homework: Practice mindfulness (body scan) daily until the next lesson.

Lesson 2 (day 6)

- Start: Review the coping style of the past 5 days.
- Proceed: Think about the long-term effects of previous coping manners.
- Transform:
  - Elucidate the 6 facets of PF and PI of ACT with different examples.
  - Mindfulness practice: body stretch.
- Integrate: Choose one facet of PF and use it as a coping strategy in the following 5 days.
- Homework: Practice mindfulness (body stretch) daily until the next lesson.

Lesson 3 (day 11)

- Start: Review the coping style of the past 5 days.
- Proceed: Assess the consequences of coping manners.
- Transform:
  - Elucidate the 6 facets of PF and PI of ACT with different examples.
  - Mindfulness practice: mindful breathing.
- Integrate: Choose one facet of PF and use it as a coping strategy in the following 5 days.
- Homework: Practice mindfulness (breathing) daily until the next lesson.

Lesson 4 (day 16)

- Start: Clarify which facet of PI is most likely to be their previous coping manners.
- Proceed: Think about the changes in their coping manners after having joined this study.
- Transform:
  - Elucidate the 6 facets of PF and PI of ACT with different examples.
  - Mindfulness practice: mindful walking.
- Integrate: Choose one facet of PF and use it as a coping strategy in the following 5 days.
- Homework: Practice mindfulness (walking) daily until the next lesson.

Lesson 5 (day 21)

- Start: Clarify which facet of PF is most likely to be their previous coping manners.
- Proceed: Think about the changes in their coping manners after having joined this study.
- Transform:
  - Elucidate the 6 facets of PF and PI of ACT with different examples.
  - Mindfulness practice: Metta meditation.
- Integrate: Choose one facet of PF and use it as a coping strategy in the following 5 days.

XSL•FO
RenderX

- Homework: Practice mindfulness (Metta meditation) daily until the next lesson.

Lesson 6 (day 26)

- Start: Determine medium and long-term objectives.
- Proceed: Think about how to achieve objectives.
- Transform:
  - Elucidate the 6 facets of PF and PI of ACT with different examples.
  - Mindfulness practice: breathing space: the door to actions.
- Integrate:
  - Use the 6 facets of PF in their life.
  - Actions are a must.
- Homework: Practice mindfulness (the door to actions) daily until the next lesson.

XSL•FO
RenderX

Figure 1. (A) The copyright registration certificate for our intervention system; (B) screenshot of the program; (C) English translation version (see Figure S1 in Multimedia Appendix 1 for the English version).

(A)

(B)

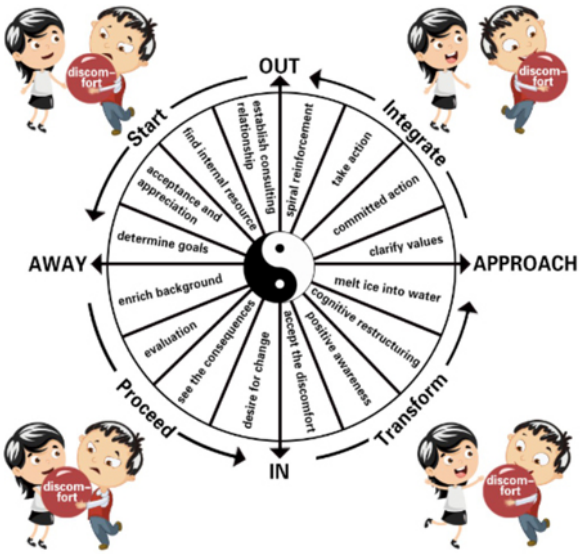

(C)

Measures

Outcome Measures

Depression, Anxiety, and Stress

The DASS-21 [67] consists of 3 self-report subscales assessing depression, anxiety, and stress. Participants indicate their experience over the past week using a 4-point Likert scale (0=not at all and 3=very much so). Each subscale score is calculated by summing the scores of its 7 items and multiplying by 2, resulting in a range of 0 to 42, where higher scores reflect greater severity of the respective emotional state. Cutoff scores for mild levels of distress are depression subscale ≥10, anxiety subscale ≥8, and stress subscale ≥15 [65,66]. The DASS-21 is widely used in both clinical practice and research (T1: Cronbach

XSL•FO
RenderX

α=0.71, McDonald ω=0.73; T2: Cronbach α=0.86, McDonald ω=0.88; T3: Cronbach α=0.87, McDonald ω=0.88).

About OCSs

The 18-item revised Obsessive-Compulsive Inventory [68] was used to assess the frequency and pain of individual OCSs (eg, "I always check the doors, windows, and drawers again and again"). It consists of 6 dimensions: washing, checking, obsessing, neutralizing, ordering, and hoarding. Participants were asked to rate the frequency of OCSs that had occurred in the past month (0 being "never" and 4 being "almost always") and how painful they had been (0 being "not painful at all" and 4 being "extremely painful"). The total score ranged from 0 to 72, with higher scores reflecting greater symptom severity (T1: Cronbach α=0.84, McDonald ω=0.74; T2: Cronbach α=0.91, McDonald ω=0.91; T3: Cronbach α=0.91, McDonald ω=0.91).

About PI

The Multidimensional Psychological Flexibility Inventory [69] contains 2 subscales: PF and PI. Each subscale contains six dimensions, corresponding to 6 facets of PF (eg, "I try to reconcile my negative thoughts and feelings instead of fighting them") and 6 facets of PI, respectively (eg, "When I have negative thoughts or emotions, it's hard to get out of them"), with a total of 60 items rated on a 1 to 6 scale. Scores for each subscale range from 30 to 180, where higher scores indicate greater PI or PF. The content validity, structure validity, and scale validity of the total scale and each subscale all meet the statistical requirements [69]. The PI subscale was adopted in this study (T1: Cronbach α=0.88, McDonald ω=0.88; T2: Cronbach α=0.92, McDonald ω=0.92; T3: Cronbach α=0.92, McDonald ω=0.92).

Statistical Analyses

Statistical analyses were conducted using SPSS (version 25; IBM Corp). The chi-squared test was used to analyze qualitative variables (eg, demographic data). For continuous variables, the independent-sample *t* test (2-tailed) and 1-way ANOVA were applied. Analyses of the intervention effects were executed with repeated-measures ANOVA, which included between-group factors, within-group factors, and group×time interaction. Whenever the sphericity assumption was not met, the Greenhouse-Geisser ε (<0.75) was used. Bonferroni correction was applied to adjust the *P* values for multiple comparisons. Specifically, after selecting the estimated marginal means in the analysis, we checked the main effect comparisons and selected Bonferroni for the CI adjustment. Effect sizes were calculated through partial Eta² (ηp^2), and the level of statistical significance for all comparisons was set at $P<.05$. An intent-to-treat analysis was applied, and the principle of last-observation-carrying-forward was used to address missing data. If no significant effect of the group×time interaction was found, no further statistical tests were conducted. Conversely, if a difference was observed, simple effect analyses were performed to assess potential interaction effects between time and group.

Ethical Considerations

This study was approved by the Institutional Review Board of the Medical Ethics Committee of Chinese PLA General Hospital (S2023-051-02), and the authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. Informed consent was obtained from all participants, who were also informed of their right to withdraw at any time without penalty. The data collected were anonymized or de-identified, with no identifiable personal information included. There was no compensation for participation, which was entirely voluntary.

Results

Demographics and Baseline Data

Participant characteristics between the 2 groups are presented in Table 1. There were no significant differences found between the iACT group and control group for participants' characteristics and each outcome variable at baseline (all $P>.05$). The CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram is presented in Figure 2.

However, based on the demographics analysis results, the depression, anxiety, stress, and DASS scores varied significantly among medical students with different grades ($F_{2,517}=26.28$, $P<.001$; $F_{2,517}=42.36$, $P<.001$; $F_{2,517}=167.79$, $P<.001$; $F_{2,517}=252.91$, $P<.001$). Ad hoc test results showed that final-year medical students' scores of depression, anxiety, stress, and DASS are significantly higher than first-year students ($P<.001$), indicating they have a lower level of mental health status.

XSL•FO
RenderX

Table 1. Demographics and baseline data.

	iACT[a] group (n=260)	Control group (n=260)	Chi-square or *t* test (*df*)	*P* value
Age (years)			0.6 (518)[b]	.54
Mean (SD)	27.07 (3.53)	26.88 (3.09)		
Range	23-45	22-39		
Sex, n (%)			0.1 (1)[c]	.71
Male	143 (55)	147 (56.7)		
Female	117 (45)	113 (43.3)		
Grade, n (%)			3.9 (2)[c]	.14
First year	83 (34.6)	75 (31.5)		
Second year	60 (25)	79 (33.2)		
Final year	97 (40.4)	84 (35.3)		
DASS-21[d] total score, mean (SD)	52.07 (10.93)	51.14 (9.76)	0.9 (518)[b]	.33
Depression score, mean (SD)	12.91 (3.16)	13.3 (2.35)	–1.5 (518)[b]	.12
Anxiety score, mean (SD)	15.38 (5.3)	14.55 (4.71)	1.8 (518)[b]	.07
Stress score, mean (SD)	23.78 (7.67)	23.29 (7.15)	0.7 (518)[b]	.48
PI[e] score, mean (SD)	102.75 (22.81)	101.3 (2)	0.7 (518)[b]	.46
OCS[f] score, mean (SD)	46.03 (8.31)	45.06 (8.86)	1.2 (518)[b]	.22

[a]iACT: internet-based acceptance and commitment therapy.
[b]*t* test
[c]Chi-square test
[d]DASS-21: 21-item Depression Anxiety Stress Scale.
[e]PI: psychological inflexibility.
[f]OCS: obsessive-compulsive symptom.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram. iACT: internet-based acceptance and commitment therapy.

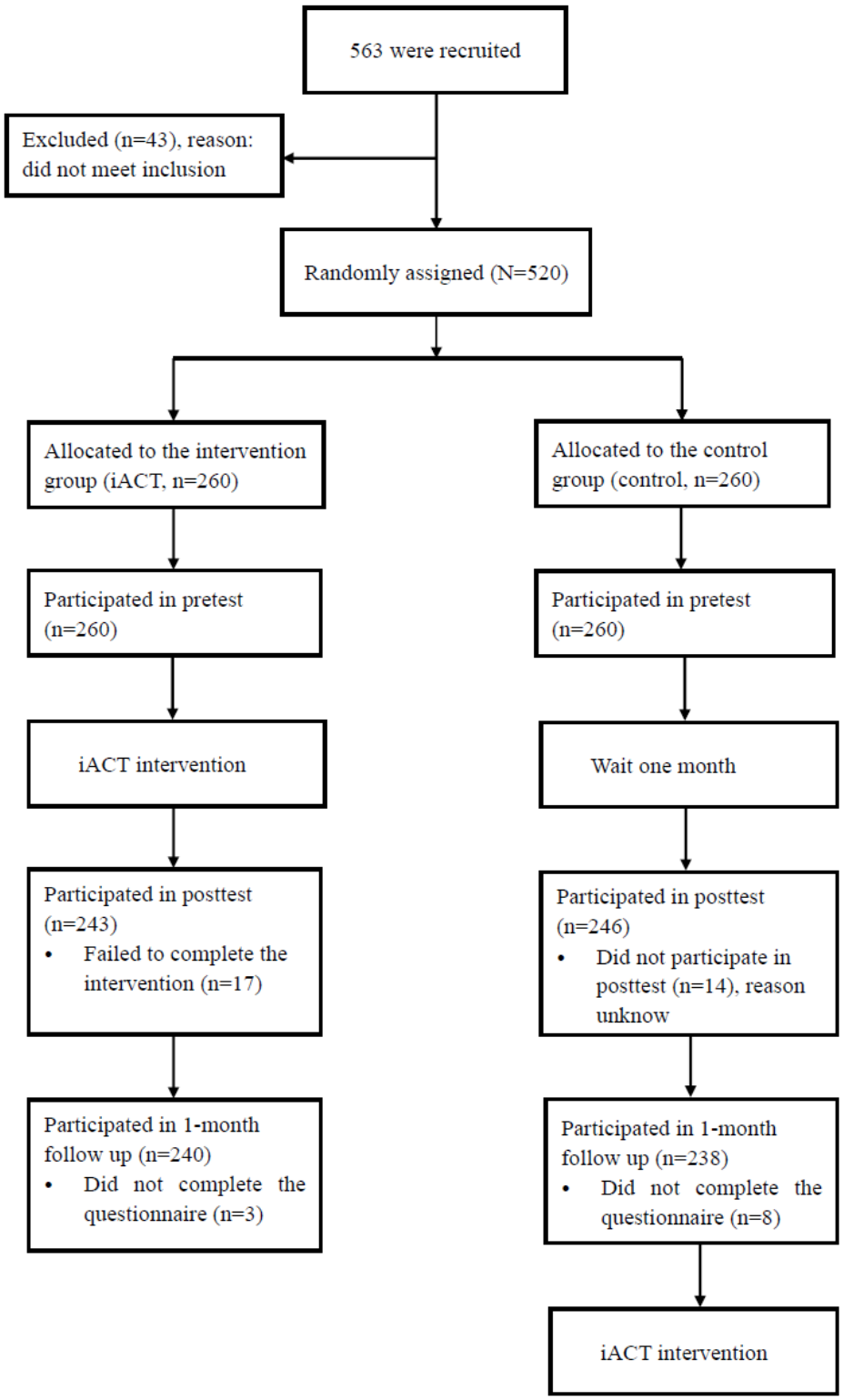

Effect of the iACT Intervention on Outcome Variables

Effect of the iACT Intervention on DASS-21 Score

As shown in Table S2 in Multimedia Appendix 1 and Figure 3A-D, the analysis revealed significant effects of time, group, and group×time interactions on the DASS-21 total score and its subscales. Participants in the intervention group demonstrated lower scores in total DASS-21, depression, anxiety, and stress compared to the control group after the intervention. Specifically, the simple effect analysis revealed that DASS-21 total score, depression score, anxiety score, and stress score were significantly lower in the intervention group at T2 and T3 (all $P<.001$), while there was no significant difference between the 3 measurements in the control group (all $P>.05$). The results of simple effect analysis between the 2 groups showed that there was no significant difference at T1 between the 2 groups, while there were significant differences at T2 and T3 (all $P<.001$). Overall, after the intervention, participants in the intervention

XSL•FO
RenderX

group progressed toward a better depression, anxiety, and stress status, suggesting the intervention was effective and the improvements were maintained during the 1-month follow-up period.

Figure 3. Changes of outcome variable scores over time: (A) DASS-21 total score, (B) depression score, (C) anxiety score, (D) stress score, (E) PI score, and (F) OCS score. iACT: internet-based acceptance and commitment therapy; DASS: Depression Anxiety Stress Scale; PI: psychological inflexibility; OCS: obsessive-compulsive symptom; TI: preintervention measurement; T2: postintervention measurement; T3: follow-up measurement.

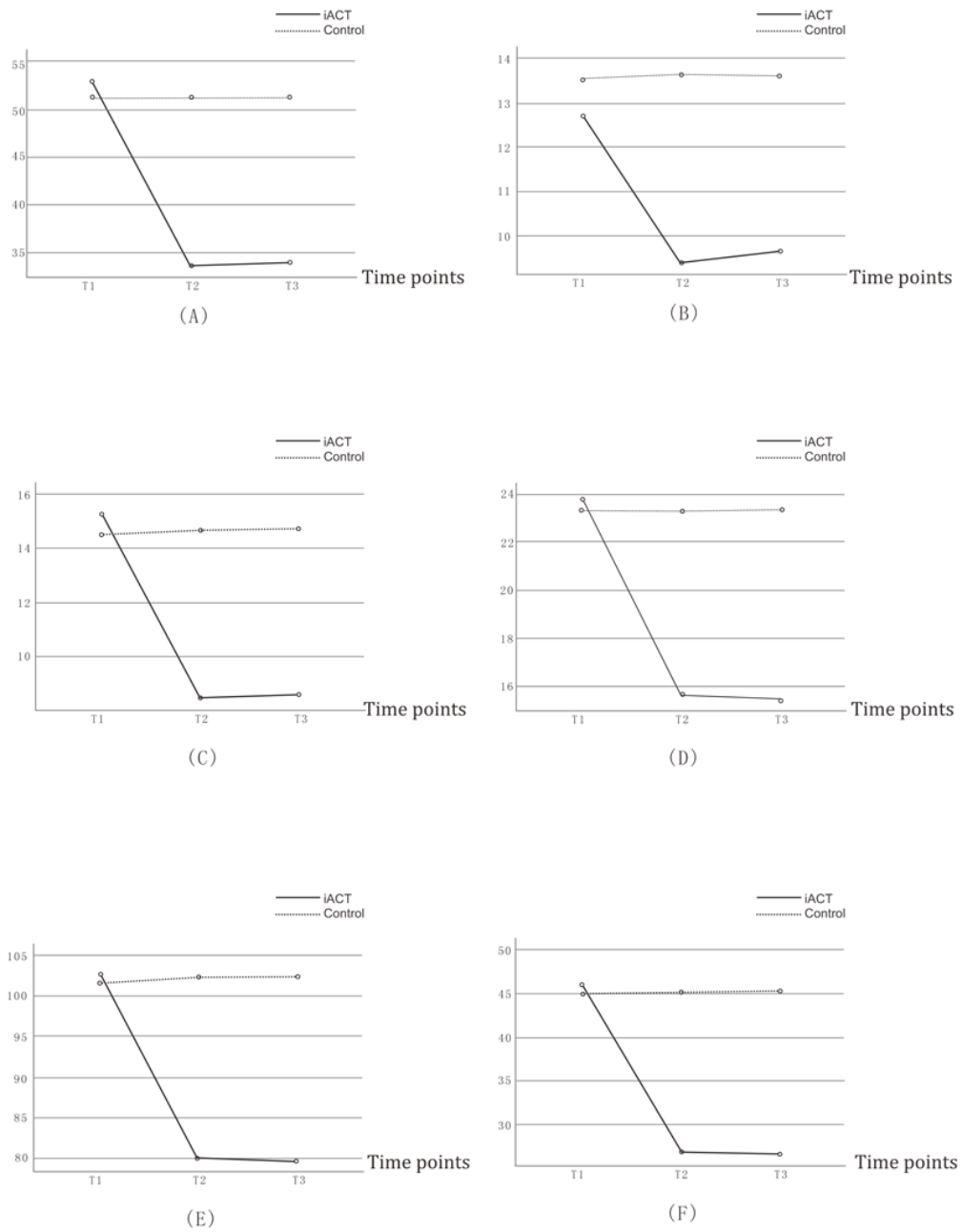

Effect of iACT Intervention on PI

The results presented in Table S2 in Multimedia Appendix 1 and Figure 3E indicated significant effects of time, group, and group×time interactions on PI scores. Participants in the intervention group demonstrated significantly lower levels of PI at both the postintervention (T2) and 1-month follow-up time points (T3) compared to the control group (all $P<.001$). In contrast, the control group exhibited no significant changes across the 3 measurement points (all $P>.05$). The analysis revealed that there were no significant differences between the 2 groups at baseline (T1), but significant differences emerged at T2 and T3 (all $P<.001$). Overall, following the intervention, participants in the intervention group showed improved PF,

XSL•FO
RenderX

suggesting that the iACT program effectively reduced PI scores, with these improvements being maintained during the follow-up period.

Effect of iACT Intervention on OCSs

Table S2 in Multimedia Appendix 1 and Figure 3F revealed significant effects of time, group, and group×time interactions on OCS scores. Participants in the intervention group showed a significant reduction in OCSs at both the postintervention (T2) and 1-month follow-up time points (T3) compared to the control group (all $P<.001$). In contrast, the control group did not demonstrate any significant changes across the 3 measurement points (all $P>.05$). The analysis indicated no significant differences between the 2 groups at baseline (T1), but significant differences were observed at T2 and T3 (all $P<.001$). Overall, after the intervention, participants in the intervention group progressed toward a better OCS status, suggesting the intervention was effective in reducing OCSs, and the improvements were maintained during the 1-month follow-up period.

Discussion

Principal Findings

This study developed and implemented a novel intervention program, iACT 2.0, rooted in ACT. We examined its suitability for online delivery and its effectiveness in improving mental health compared to a waitlist control group. Our findings indicate that the iACT 2.0 intervention significantly reduced symptoms of mental health distress, including depression, anxiety, stress, and OCSs. Moreover, it led to improvements in PF among Chinese medical students experiencing at least mild levels of depression, anxiety, and stress. These positive effects were sustained at the 1-month follow-up assessment.

A unique aspect of our study is the development of a new intervention program based on ACT, which offers several advantages. First, it integrates cultural and spiritual elements, using the Bagua diagram from traditional Chinese Daoism to illustrate dynamic thinking patterns, helping participants grasp core intervention principles within a familiar context. Second, each lesson follows the 4 steps of “启-承-转-合 (starting, proceeding, transforming, integrating)” to teach the 6 facets of PF and PI. The “启” step introduces participants to the session and sets goals. The “承” step encourages participants to review the past and connect with the present. The “转” step suggests the participants to undergo the transformation and seek change. The “合” step guides participants in integrating session content into actionable steps. This structured approach enhances understanding and acceptance of the intervention content. Third, our program provides participants with a solution from the first lesson, which attracts them to complete the 6 lessons, thereby improving adherence and retention (with a drop rate of 20/260, 7.7% in the iACT group). Fourth, each lesson is a further explanation of the 6 facets, step by step, with the whole process being a kind of recurrent reinforcement. Throughout the 6 lessons, the contents are reviewed, consolidated, and improved repeatedly, enabling participants to master ACT and gain insights. Fifth, each lesson is followed by homework, such as practicing targeted daily mindfulness and selected facet practice in the following 5 days before the next lesson, which could promote the intervention’s effectiveness. Lastly, our program gives participants more autonomy to begin with the most inflexible facet and then guides them to explore their internal resources to proceed with the whole process. Overall, our intervention aims to maximize the benefits for participants in a limited time as much as possible. It could be applied to one-time psychological assistance and consulting, offering a good opportunity to address common mental health conditions in medical students while simultaneously reducing the burden on traditional counseling and services.

In recent years, there has been increasing research focused on interventions to improve the mental health of medical students. However, existing studies are limited by small sample sizes and lack of representativeness. For instance, one study included only 91 medical students from An-Najah National University, Palestine, in a cognitive behavioral therapy program for mental health status [70]. Another study involved 120 undergraduate nursing students in China to examine the effectiveness of an online mindfulness intervention course on their mental health [71]. By contrast, our study included 478 postgraduate medical students in China, and the results demonstrated a significant reduction in mean values of depression, anxiety, stress, total DASS-21, OCS score, and PI score compared to a waitlist control group after the intervention. Importantly, no personal therapist contact was incorporated during our intervention, allowing us to isolate the influence of therapist instruction and demonstrate the effectiveness of a self-help model. Although previous research has not reached a consensus regarding the role of therapist support during online interventions, with some studies finding unsupported interventions did not lead to significant changes in PF, and others reporting that web-based, stand-alone treatments are comparably effective to therapist-assisted interventions [54,72], our findings contribute to the growing research on brief iACT interventions. Our findings also have cost implications, given that personnel costs often constitute the largest proportion of online interventions and face-to-face consulting costs [73]. If the developed self-help intervention is effective without the assistance of paid personnel, the participant cost in research and psychotherapy costs in real-world practice may be reduced.

Our study revealed that final-year postgraduate medical students reported significantly lower levels of positive mental health and higher levels of mental health distress compared to first-year students, which is consistent with the findings of a recent study by Wimberly et al [74]. These results provide further evidence for the association between grade level and mental health distress among medical students. They also suggest that university administrators and educators should pay more attention to the mental health of final-year students and consider providing additional psychological support resources to this particular cohort.

Limitations and Future Research

Overall, this study could make a significant contribution to promoting online psychotherapy for improving medical students’ mental well-being, with strengths including the self-help iACT intervention formulation, follow-up study, and relatively large

XSL•FO
RenderX

sample size. However, it also had several limitations. First, the participants in our study were postgraduate medical students, who are highly educated and familiar with modern technology, thereby limiting the representativeness of the sample. Future research should investigate the program's effectiveness among populations lacking digital literacy or within clinical demographics. Second, the use of self-report scales may potentially cause response bias, participants may also lose the patience to fill in carefully. To address this, future research could use adaptive algorithms to customize differentiated scales for different participants, allowing for real-time jumping to the most relevant items through the interaction with participants' answers. Third, this study integrated the entire 6 facets of ACT into the core component of the intervention and examined its efficacy, future research should explore and examine which specific facet is most important during the intervention. Finally, this study did not explore intervention-related placebo effects, which should be considered in future research.

Conclusion

In conclusion, this study demonstrates that the newly developed iACT 2.0 effectively alleviates depression, anxiety, stress, OCSs, and PI among medical students experiencing at least mild levels of psychological distress. Furthermore, the observed positive effects were sustained at the 1-month follow-up, suggesting the program's viability as a practical intervention for addressing prevalent mental health issues. This offers a promising alternative to conventional counseling services, potentially alleviating their burden.

Acknowledgments

We sincerely thank all the participants in this study for their support and cooperation. This work was supported by the nursing special project (2021HL087).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DW and BL contributed to the conceptualization, data analysis, original draft writing, and review. SZ contributed to the review and editing. WX contributed to the review, proofreading, and editing, and is the corresponding author. XL contributed to the review, editing, and funding acquisition, and is the co-corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 630 KB - jmir_v26i1e50664_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1187 KB - jmir_v26i1e50664_app2.pdf]

References

1. Nishimura Y, Ochi K, Tokumasu K, Obika M, Hagiya H, Kataoka H, et al. Impact of the COVID-19 pandemic on the psychological distress of medical students in Japan: cross-sectional survey study. J Med Internet Res 2021;23(2):e25232 [FREE Full text] [doi: 10.2196/25232] [Medline: 33556033]
2. Jin T, Sun Y, Wang H, Qiu F, Wang X. Prevalence of depression among Chinese medical students: a systematic review and meta-analysis. Psychol Health Med 2022;27(10):2212-2228. [doi: 10.1080/13548506.2021.1950785] [Medline: 34232090]
3. Zhang X, Yue H, Sun J, Liu M, Li C, Bao H. Regulatory emotional self-efficacy and psychological distress among medical students: multiple mediating roles of interpersonal adaptation and self-acceptance. BMC Med Educ 2022;22(1):283 [FREE Full text] [doi: 10.1186/s12909-022-03338-2] [Medline: 35421953]
4. Givens JL, Tjia J. Depressed medical students' use of mental health services and barriers to use. Acad Med 2002;77(9):918-921. [doi: 10.1097/00001888-200209000-00024] [Medline: 12228091]
5. Quek TTC, Tam WWS, Tran BX, Zhang M, Zhang Z, Ho CSH, et al. The global prevalence of anxiety among medical students: a meta-analysis. Int J Environ Res Public Health 2019;16(15):2735 [FREE Full text] [doi: 10.3390/ijerph16152735] [Medline: 31370266]

XSL•FO
RenderX

6. Rotenstein LS, Ramos MA, Torre M, Segal JB, Peluso MJ, Guille C, et al. Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis. JAMA 2016;316(21):2214-2236 [FREE Full text] [doi: 10.1001/jama.2016.17324] [Medline: 27923088]
7. Moore S, Barbour R, Ngo H, Sinclair C, Chambers R, Auret K, et al. Determining the feasibility and effectiveness of brief online mindfulness training for rural medical students: a pilot study. BMC Med Educ 2020;20(1):104 [FREE Full text] [doi: 10.1186/s12909-020-02015-6] [Medline: 32252750]
8. Dyrbye LN, Thomas MR, Shanafelt TD. Systematic review of depression, anxiety, and other indicators of psychological distress among U.S. and Canadian medical students. Acad Med 2006;81(4):354-373. [doi: 10.1097/00001888-200604000-00009] [Medline: 16565188]
9. Hope V, Henderson M. Medical student depression, anxiety and distress outside North America: a systematic review. Med Educ 2014;48(10):963-979. [doi: 10.1111/medu.12512] [Medline: 25200017]
10. Ludwig AB, Burton W, Weingarten J, Milan F, Myers DC, Kligler B. Depression and stress amongst undergraduate medical students. BMC Med Educ 2015;15:141 [FREE Full text] [doi: 10.1186/s12909-015-0425-z] [Medline: 26311538]
11. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). Arlington, VA: American Psychiatric Publishing; 2022.
12. Work-related stress. European Foundation for the Improvement of Living and Working Conditions. URL: https://www.eurofound.europa.eu/sites/default/files/ef_files/docs/ewco/tn1004059s/tn1004059s.pdf [accessed 2024-11-01]
13. Þórarinsdóttir H, Kessing LV, Faurholt-Jepsen M. Smartphone-based self-assessment of stress in healthy adult individuals: a systematic review. J Med Internet Res 2017;19(2):e41 [FREE Full text] [doi: 10.2196/jmir.6397] [Medline: 28193600]
14. Yang F, Meng H, Chen H, Xu XH, Liu Z, Luo A, et al. Influencing factors of mental health of medical students in China. J Huazhong Univ Sci Technolog Med Sci 2014;34(3):443-449. [doi: 10.1007/s11596-014-1298-9] [Medline: 24939314]
15. Torres AR, Cruz BL, Vicentini HC, Lima MCP, Ramos-Cerqueira ATA. Obsessive-compulsive symptoms in medical students: prevalence, severity, and correlates. Acad Psychiatry 2016;40(1):46-54. [doi: 10.1007/s40596-015-0357-2] [Medline: 26108391]
16. Mazhar K, Khaliq F, Arshad D. Association of COVID 19 pandemic with new onset obsessive-compulsive disorder (OCD) symptomology in the medical students – a cross sectional study. BJPsych open 2021;7(S1):S270-S270. [doi: 10.1192/bjo.2021.718]
17. Ji G, Wei W, Yue KC, Li H, Shi LJ, Ma JD, et al. Effects of the COVID-19 pandemic on obsessive-compulsive symptoms among university students: prospective cohort survey study. J Med Internet Res 2020;22(9):e21915 [FREE Full text] [doi: 10.2196/21915] [Medline: 32931444]
18. Fahrenkopf AM, Sectish TC, Barger LK, Sharek PJ, Lewin D, Chiang VW, et al. Rates of medication errors among depressed and burnt out residents: prospective cohort study. BMJ 2008;336(7642):488-491 [FREE Full text] [doi: 10.1136/bmj.39469.763218.BE] [Medline: 18258931]
19. Meyers JE, Grills CE, Zellinger MM, Miller RM. Emotional distress affects attention and concentration: the difference between mountains and valleys. Appl Neuropsychol Adult 2014;21(1):28-35. [doi: 10.1080/09084282.2012.721148] [Medline: 24826493]
20. Dyrbye LN, Thomas MR, Shanafelt TD. Medical student distress: causes, consequences, and proposed solutions. Mayo Clin Proc 2005;80(12):1613-1622. [doi: 10.4065/80.12.1613] [Medline: 16342655]
21. Picco L, Seow E, Chua BY, Mahendran R, Verma S, Chong SA, et al. Recognition of mental disorders: findings from a cross-sectional study among medical students in Singapore. BMJ Open 2017;7(12):e019038 [FREE Full text] [doi: 10.1136/bmjopen-2017-019038] [Medline: 29273669]
22. Mojsa-Kaja J, Golonka K, Gawłowska M. Preliminary analyses of psychometric characteristics of the polish version of the obsessive-compulsive inventory-revised (OCI-R) in a non-clinical sample. Int J Occup Med Environ Health 2016;29(6):1011-1021 [FREE Full text] [doi: 10.13075/ijomeh.1896.00792] [Medline: 27869250]
23. Sultan S, Fallata EO, Bashar MA, Olaqi EE, Alsharif GH, BinSaleh RA, et al. Prevalence, sociodemographic and academic correlates of obsessive-compulsive disorder in the students of college of applied medical sciences, Umm Al-Qura university. J Obsessive-Compulsive Relat Disord 2021;28:100604. [doi: 10.1016/j.jocrd.2020.100604]
24. Johansen T, Dittrich WH. Cognitive performance in a subclinical obsessive-compulsive sample 1: cognitive functions. Psychiatry J 2013;2013:565191 [FREE Full text] [doi: 10.1155/2013/565191] [Medline: 24236282]
25. Phillips MS, Steelesmith DL, Brock G, Benedict J, Muñoz J, Fontanella CA. Mental health service utilization among medical students with a perceived need for care. Acad Psychiatry 2022;46(2):223-227. [doi: 10.1007/s40596-021-01584-y] [Medline: 35006590]
26. Sahhar D, O'Connor D. How well do Australian medical schools prepare general practitioners to care for patients with mental disorders? Australas Psychiatry 2004;12(1):26-30. [doi: 10.1046/j.1039-8562.2003.02055.x] [Medline: 15715735]
27. Chur-Hansen A, Carr JE, Bundy C, Sanchez-Sosa JJ, Tapanya S, Wahass SH. An international perspective on behavioral science education in medical schools. J Clin Psychol Med Settings 2008;15(1):45-53. [doi: 10.1007/s10880-008-9092-0] [Medline: 19104953]
28. Amarasuriya SD, Jorm AF, Reavley NJ. Quantifying and predicting depression literacy of undergraduates: a cross sectional study in Sri Lanka. BMC Psychiatry 2015;15:269 [FREE Full text] [doi: 10.1186/s12888-015-0658-8] [Medline: 26518153]

XSL•FO
RenderX

29. Furnham A, Cook R, Martin N, Batey M. Mental health literacy among university students. J Public Ment Health 2011;10(4):198-210. [doi: 10.1108/17465721111188223]
30. Hayes SC, Luoma JB, Bond FW, Masuda A, Lillis J. Acceptance and commitment therapy: model, processes and outcomes. Behav Res Ther 2006;44(1):1-25. [doi: 10.1016/j.brat.2005.06.006] [Medline: 16300724]
31. Han A, Kim TH. Efficacy of internet-based acceptance and commitment therapy for depressive symptoms, anxiety, stress, psychological distress, and quality of life: systematic review and meta-analysis. J Med Internet Res 2022;24(12):e39727 [FREE Full text] [doi: 10.2196/39727] [Medline: 36485030]
32. McCracken LM, Morley S. The psychological flexibility model: a basis for integration and progress in psychological approaches to chronic pain management. J Pain 2014;15(3):221-234 [FREE Full text] [doi: 10.1016/j.jpain.2013.10.014] [Medline: 24581630]
33. Bai Z, Luo S, Zhang L, Wu S, Chi I. Acceptance and commitment therapy (ACT) to reduce depression: a systematic review and meta-analysis. J Affect Disord 2020;260:728-737. [doi: 10.1016/j.jad.2019.09.040] [Medline: 31563072]
34. Han A, Yuen HK, Jenkins J. Acceptance and commitment therapy for family caregivers: a systematic review and meta-analysis. J Health Psychol 2021;26(1):82-102. [doi: 10.1177/1359105320941217] [Medline: 32659142]
35. Levin ME, Haeger JA, Pierce BG, Twohig MP. Web-based acceptance and commitment therapy for mental health problems in college students: a randomized controlled trial. Behav Modif 2017;41(1):141-162. [doi: 10.1177/0145445516659645] [Medline: 27440189]
36. El Morr C, Maule C, Ashfaq I, Ritvo P, Ahmad F. A student-centered mental health virtual community needs and features: a focus group study. Stud Health Technol Inform 2017;234:104-108. [Medline: 28186024]
37. Chong YY, Mak YW, Leung SP, Lam SY, Loke AY. Acceptance and commitment therapy for parental management of childhood asthma: an RCT. Pediatrics 2019;143(2):e20181723. [doi: 10.1542/peds.2018-1723] [Medline: 30659063]
38. Puolakanaho A, Lappalainen R, Lappalainen P, Muotka JS, Hirvonen R, Eklund KM, et al. Reducing stress and enhancing academic buoyancy among adolescents using a brief web-based program based on acceptance and commitment therapy: a randomized controlled trial. J Youth Adolesc 2019;48(2):287-305 [FREE Full text] [doi: 10.1007/s10964-018-0973-8] [Medline: 30560515]
39. Räsänen P, Lappalainen P, Muotka J, Tolvanen A, Lappalainen R. An online guided ACT intervention for enhancing the psychological wellbeing of university students: a randomized controlled clinical trial. Behav Res Ther 2016;78:30-42. [doi: 10.1016/j.brat.2016.01.001] [Medline: 26848517]
40. Lappalainen P, Langrial S, Oinas-Kukkonen H, Tolvanen A, Lappalainen R. Web-based acceptance and commitment therapy for depressive symptoms with minimal support: a randomized controlled trial. Behav Modif 2015;39(6):805-834. [doi: 10.1177/0145445515598142] [Medline: 26253644]
41. Pang NTP, Koh E, Mohd Kassim MA, Shoesmith WD, Kamu A, Ho CM, et al. Enhancing final-year medical students' clinical examinations performance via a transdiagnostic ACT-based intervention. Sustainability 2022;14(13):7522. [doi: 10.3390/su14137522]
42. Ebrahimi A, Nasre Esfahan E, Akuchekian S, Izadi R, Shaneh E, Mahaki B. A randomized clinical trial: comparison of group acceptance and commitment therapy with drug on quality of life and depression in patients with obsessive-compulsive disorder. J Res Med Sci 2022;27:9 [FREE Full text] [doi: 10.4103/jrms.jrms_449_21] [Medline: 35342452]
43. Levin ME, Hildebrandt MJ, Lillis J, Hayes SC. The impact of treatment components suggested by the psychological flexibility model: a meta-analysis of laboratory-based component studies. Behav Ther 2012;43(4):741-756. [doi: 10.1016/j.beth.2012.05.003] [Medline: 23046777]
44. Rohani F, Rasouli-Azad M, Twohig MP, Ghoreishi FS, Lee EB, Akbari H. Preliminary test of group acceptance and commitment therapy on obsessive-compulsive disorder for patients on optimal dose of selective serotonin reuptake inhibitors. J Obsessive-Compulsive Relat Disord 2018;16:8-13. [doi: 10.1016/j.jocrd.2017.10.002]
45. Zou J, Wu S, Yuan X, Hu Z, Tang J, Hu M. Effects of acceptance and commitment therapy and repetitive transcranial magnetic stimulation on obsessive-compulsive disorder. Front Psychiatry 2022;12:720518 [FREE Full text] [doi: 10.3389/fpsyt.2021.720518] [Medline: 35095583]
46. Armstrong AB, Morrison KL, Twohig MP. A preliminary investigation of acceptance and commitment therapy for adolescent obsessive-compulsive disorder. J Cogn Psychother 2013;27(2):175-190. [doi: 10.1891/0889-8391.27.2.175] [Medline: 32759121]
47. Twohig MP, Vilardaga JCP, Levin ME, Hayes SC. Changes in psychological flexibility during acceptance and commitment therapy for obsessive compulsive disorder. J Contextual Behav Sci 2015;4(3):196-202. [doi: 10.1016/j.jcbs.2015.07.001]
48. Sherifali D, Ali MU, Ploeg J, Markle-Reid M, Valaitis R, Bartholomew A, et al. Impact of internet-based interventions on caregiver mental health: systematic review and meta-analysis. J Med Internet Res 2018;20(7):e10668 [FREE Full text] [doi: 10.2196/10668] [Medline: 29970358]
49. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014;16(5):e130 [FREE Full text] [doi: 10.2196/jmir.3142] [Medline: 24836465]

XSL•FO
RenderX

50. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]
51. Hoffmann D, Rask CU, Hedman-Lagerlöf E, Jensen JS, Frostholm L. Efficacy of internet-delivered acceptance and commitment therapy for severe health anxiety: results from a randomized, controlled trial. Psychol Med 2021;51(15):2685-2695 [FREE Full text] [doi: 10.1017/S0033291720001312] [Medline: 32404226]
52. Kelson J, Rollin A, Ridout B, Campbell A. Internet-delivered acceptance and commitment therapy for anxiety treatment: systematic review. J Med Internet Res 2019;21(1):e12530 [FREE Full text] [doi: 10.2196/12530] [Medline: 30694201]
53. Andersson E, Enander J, Andrén P, Hedman E, Ljótsson B, Hursti T, et al. Internet-based cognitive behaviour therapy for obsessive-compulsive disorder: a randomized controlled trial. Psychol Med 2012;42(10):2193-2203 [FREE Full text] [doi: 10.1017/S0033291712000244] [Medline: 22348650]
54. McCall HC, Richardson CG, Helgadottir FD, Chen FS. Evaluating a web-based social anxiety intervention among university students: randomized controlled trial. J Med Internet Res 2018;20(3):e91 [FREE Full text] [doi: 10.2196/jmir.8630] [Medline: 29563078]
55. Lattie EG, Kashima K, Duffecy JL. An open trial of internet-based cognitive behavioral therapy for first year medical students. Internet Interv 2019;18:100279 [FREE Full text] [doi: 10.1016/j.invent.2019.100279] [Medline: 31534911]
56. Bernal G, Jiménez-Chafey MI, Domenech Rodríguez MM. Cultural adaptation of treatments: a resource for considering culture in evidence-based practice. Prof Psychol: Res Pract 2009;40(4):361-368. [doi: 10.1037/a0016401]
57. Nagayama Hall GC, Hong JJ, Zane NWS, Meyer OL. Culturally-competent treatments for Asian Americans: the relevance of mindfulness and acceptance-based psychotherapies. Clin Psychol (New York) 2011;18(3):215-231 [FREE Full text] [doi: 10.1111/j.1468-2850.2011.01253.x] [Medline: 22065893]
58. Chowdhary N, Jotheeswaran AT, Nadkarni A, Hollon SD, King M, Jordans MJD, et al. The methods and outcomes of cultural adaptations of psychological treatments for depressive disorders: a systematic review. Psychol Med 2014;44(6):1131-1146 [FREE Full text] [doi: 10.1017/S0033291713001785] [Medline: 23866176]
59. Hall GCN, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
60. Soto A, Smith TB, Griner D, Domenech Rodríguez M, Bernal G. Cultural adaptations and therapist multicultural competence: two meta-analytic reviews. J Clin Psychol 2018;74(11):1907-1923. [doi: 10.1002/jclp.22679] [Medline: 30091201]
61. Perry A, Gardener C, Oliver JE, Taş Ç, Özenç C. Exploring the cultural flexibility of the ACT model as an effective therapeutic group intervention for Turkish speaking communities in East London. TCBT 2019;12:1-25. [doi: 10.1017/s1754470x18000041]
62. Perry A, Gardener C, Shieh J, H QT, Doan A, Bhui K. Investigating the acceptability of a culturally adapted acceptance and commitment therapy group for UK Vietnamese communities: A practice-based feasibility study. Transcult Psychiatry 2024 Aug;61(4):626-651. [doi: 10.1177/13634615241228071] [Medline: 38529626]
63. Jia XM, You LY, An Q. Specifics and operational procedures of the psychological assistance hotline during the COVID-19 pandemic in China. Intl J Ment Health Promot 2020;22(3):115-121. [doi: 10.32604/IJMHP.2020.011120] [Medline: 39055887]
64. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
65. Kyrou I, Christou A, Panagiotakos D, Stefanaki C, Skenderi K, Katsana K, et al. Effects of a hops (Humulus lupulus L.) dry extract supplement on self-reported depression, anxiety and stress levels in apparently healthy young adults: a randomized, placebo-controlled, double-blind, crossover pilot study. Hormones (Athens) 2017;16(2):171-180 [FREE Full text] [doi: 10.14310/horm.2002.1738] [Medline: 28742505]
66. Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the depression anxiety stress scales in clinical groups and a community sample. Psychological Assessment 1998;10(2):176-181. [doi: 10.1037/1040-3590.10.2.176]
67. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. Behav Res Ther 1995;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
68. Foa EB, Huppert JD, Leiberg S, Langner R, Kichic R, Hajcak G, et al. The obsessive-compulsive inventory: development and validation of a short version. Psychol Assess 2002;14(4):485-496. [Medline: 12501574]
69. Rolffs JL, Rogge RD, Wilson KG. Disentangling components of flexibility via the hexaflex model: development and validation of the multidimensional psychological flexibility inventory (MPFI). Assessment 2018;25(4):458-482. [doi: 10.1177/1073191116645905] [Medline: 27152011]
70. Hanani A, Badrasawi M, Zidan S, Hunjul M. Effect of cognitive behavioral therapy program on mental health status among medical student in Palestine during COVID pandemic. BMC Psychiatry 2022;22(1):310 [FREE Full text] [doi: 10.1186/s12888-022-03915-1] [Medline: 35501772]

XSL•FO
RenderX

71. Dai Z, Jing S, Wang H, Xiao W, Huang Y, Chen X, et al. Mindfulness-based online intervention on mental health among undergraduate nursing students during coronavirus disease 2019 pandemic in Beijing, China: a randomized controlled trial. Front Psychiatry 2022;13:949477 [FREE Full text] [doi: 10.3389/fpsyt.2022.949477] [Medline: 36465283]
72. Lappalainen P, Langrial S, Oinas-Kukkonen H, Muotka J, Lappalainen R. ACT for sleep - internet-delivered self-help ACT for sub-clinical and clinical insomnia: a randomized controlled trial. J Contextual Behav Sci 2019;12:119-127. [doi: 10.1016/j.jcbs.2019.04.001]
73. Ahmad F, El Morr C, Ritvo P, Othman N, Moineddin R, Team MVC. An eight-week, web-based mindfulness virtual community intervention for students' mental health: randomized controlled trial. JMIR Ment Health 2020;7(2):e15520 [FREE Full text] [doi: 10.2196/15520] [Medline: 32074061]
74. Wimberly CE, Rajapakse H, Park LP, Price A, Proeschold-Bell RJ, Østbye T. Mental well-being in Sri Lankan medical students: a cross-sectional study. Psychol Health Med 2022;27(6):1213-1226. [doi: 10.1080/13548506.2020.1858488] [Medline: 33356528]

Abbreviations

ACT: acceptance and commitment therapy
CONSORT: Consolidated Standards of Reporting Trials
DASS-21: 21-item Depression Anxiety Stress Scale
iACT: internet-based acceptance and commitment therapy
iACT 2.0: internet-based, self-help, acceptance and commitment therapy program
OCD: obsessive-compulsive disorder
OCS: obsessive-compulsive symptom
PF: psychological flexibility
PI: psychological inflexibility

Edited by A Coristine; submitted 08.07.23; peer-reviewed by Y Wang, M Yu; comments to author 28.02.24; revised version received 09.04.24; accepted 23.10.24; published 04.12.24.

Please cite as:
Wang D, Lin B, Zhang S, Xu W, Liu X
Effectiveness of an Internet-Based Self-Help Acceptance and Commitment Therapy Program on Medical Students' Mental Well-Being: Follow-Up Randomized Controlled Trial
J Med Internet Res 2024;26:e50664
URL: https://www.jmir.org/2024/1/e50664
doi:10.2196/50664
PMID:

©Difan Wang, Bingyan Lin, Shuangxi Zhang, Wei Xu, Xinying Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Based Suicide Risk Prediction Model for Suicidal Trajectory on Social Media Following Suicidal Mentions: Independent Algorithm Validation

Zachary Kaminsky[1,2,3,4], PhD; Robyn J McQuaid[1,5], PhD; Kim GC Hellemans[5], PhD; Zachary R Patterson[5], PhD; Mysa Saad[1,6], MSc, MD; Robert L Gabrys[5], PhD; Tetyana Kendzerska[6,7], MD, PhD; Alfonso Abizaid[5], PhD; Rebecca Robillard[1,8], PhD

[1]University of Ottawa Institute of Mental Health Research at The Royal, Ottawa, ON, Canada
[2]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, ON, Canada
[3]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[4]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[5]Department of Neuroscience, Carleton University, Ottawa, ON, Canada
[6]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[7]The Ottawa Hospital Research Institute, University of Ottawa, Ottawa, ON, Canada
[8]Department of Psychology, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:
Zachary Kaminsky, PhD
University of Ottawa Institute of Mental Health Research at The Royal
1145 Carling Avenue
Ottawa, ON, K1Z 7K4
Canada
Phone: 1 6137226521 ext 7003
Email: Zachary.Kaminsky@theroyal.ca

Abstract

Background: Previous efforts to apply machine learning–based natural language processing to longitudinally collected social media data have shown promise in predicting suicide risk.

Objective: Our primary objective was to externally validate our previous machine learning algorithm, the Suicide Artificial Intelligence Prediction Heuristic (SAIPH), against external survey data in 2 independent cohorts. A second objective was to evaluate the efficacy of SAIPH as an indicator of changing suicidal ideation (SI) over time. The tertiary objective was to use SAIPH to evaluate factors important for improving or worsening suicidal trajectory on social media following suicidal mention.

Methods: Twitter (subsequently rebranded as X) timeline data from a student survey cohort and COVID-19 survey cohort were scored using SAIPH and compared to SI questions on the Beck Depression Inventory and the Self-Report version of the Quick Inventory of Depressive Symptomatology in 159 and 307 individuals, respectively. SAIPH was used to evaluate changing SI trajectory following suicidal mentions in 2 cohorts collected using the Twitter application programming interface.

Results: An interaction of the mean SAIPH score derived from 12 days of Twitter data before survey completion and the average number of posts per day was associated with quantitative SI metrics in each cohort (student survey cohort interaction β=.038, SD 0.014; $F_{4,94}$=3.3, P=.01; and COVID-19 survey cohort interaction β=.0035, SD 0.0016; $F_{4,493}$=2.9, P=.03). The slope of average daily SAIPH scores was associated with the change in SI scores within longitudinally followed individuals when evaluating periods of 2 weeks or less (ρ=0.27, P=.04). Using SAIPH as an indicator of changing SI, we evaluated SI trajectory in 2 cohorts with suicidal mentions, which identified that those with responses within 72 hours exhibit a significant negative association of the SAIPH score with time in the 3 weeks following suicidal mention (ρ=–0.52, P=.02).

Conclusions: Taken together, our results not only validate the association of SAIPH with perceived stress, SI, and changing SI over time but also generate novel methods to evaluate the effects of social media interactions on changing suicidal trajectory.

(J Med Internet Res 2024;26:e49927) doi:10.2196/49927

XSL•FO
RenderX

KEYWORDS

suicide; prediction; social media; machine learning; suicide risk model; validation; prediction; natural language processing; suicide risk; Twitter; suicidal ideation; suicidal mention

Introduction

Suicide is a major public health problem that affects individuals and their families. Previous work has demonstrated that youth may disclose suicidal thoughts on social media [1], and in turn, those expressions of suicidality on social media are associated with suicide attempts [2] and death by suicide [3], suggesting the application of machine learning and natural language processing (NLP) tools to social media text may represent promising areas to generate methods for suicide risk prediction [4]. Several methods focusing on the identification of potentially suicidal tweets have been demonstrated in the past, as outlined by Roy et al [5]. Since that publication, Ophir et al [6] demonstrated that a multilayered machine learning model using deep neural nets incorporating known suicide risk factors and Facebook posts detects suicide risk with the area under the receiver operator characteristic curve (AUC) ranging between 0.70 and 0.75. Notably, the authors observed higher performance of their multilayered machine learning model that incorporated suicide theory than their single-task model that predicts risk from social media posts directly. More recently, novel methods have been developed to attempt to optimize risk detection performance. Sawhney et al [7] used adversarial methods for social media text–based suicide risk classification and demonstrated improvements over other deep learning methods such as convolutional neural networks. Naseem et al [8] applied graph-based hierarchical attention networks to a public Reddit dataset to improve performance of suicide risk prediction, and Li et al [9] used Label-Text Correlation and Deviation Punishment methods to penalize false predictions and generate improved performance. Xu et al [10] demonstrated the power of integrating transfer learning in a targeted fusion approach, potentially applicable to any data source including social media, to generate and demonstrate the efficacy of a predictor of youth's suicide risk in a hospital setting using a clinical database. While recent efforts have demonstrated the efficacy of using hospital electronic health record data in temporally informed, random forest models [11], importantly, few of the methods leveraging social media data prognosticate future risk of suicidal ideation (SI) and behaviors. The ability of these technologies to predict risk in the future, before periods of acute distress, may allow for the rapid delivery of noninvasive psychosocial interventions and may help to improve access by moving care upstream from primary care facilities. Burkhardt et al [12] applied bidirection transfer learning, integrating clinical data with social media data, to generate improved performance on models indicating suicidal messages and subsequently demonstrated the potential of an automated triage approach to improve response time by 15 minutes. In a recent pilot study, the implementation of online screening of at-risk individuals and delivery of information and counseling-based interventions was well tolerated by responsive individuals and yielded an increase in help-seeking behavior in those who previously reported being unmotivated to seek help [13]. This highlights the promise that online screening technologies may have for suicide screening and intervention; however, as with most of the studies above, this study relies on mentions of suicide-related keywords or in-the-moment expressions of suicide-related thought and, as such, will be limited to those already expressing suicidal sentiments online.

Previous work from our group generated an algorithm entitled the Suicide Artificial Intelligence Prediction Heuristic (SAIPH) [5], which is an NLP algorithm designed to predict future risk of suicidal thoughts and behaviors from social media text data. A key feature of SAIPH is that it efficaciously predicts the risk of SI using data collected before the expression of SI. Briefly, SAIPH converts text-based communications such as tweets into scores for psychological constructs that can be associated with suicide, including depression, anxiety, stress, loneliness, perceived burdensomeness, and hopelessness, among others, and subsequently generates a risk score through the integration of these metrics by a random forest model trained on those with SI mentions (those expressing their SI on Twitter [subsequently rebranded as X]) against controls. Our previous work demonstrated that SAIPH accurately predicted an independent cohort of Twitter users with SI mentions from those without such mentions with an AUC of 0.87 using data before SI mentions, and generated an AUC of 0.70 for identifying those with a history of suicide attempts or plans from those with SI. While promising, these early results were based solely on publicly available SI mentions on Twitter and require external validation in independent datasets of consenting individuals who have answered standardized questions about levels of SI. The primary objective of this study was to validate the efficacy of SAIPH for the prediction of suicidal phenotypes assessed through psychometrically validated scales and to assess the temporal variation of SAIPH scores.

A second important question is not only if our algorithm can prognosticate risk but also if it can act as an indicator of changing SI. Therefore, our secondary objective was to assess the temporal variation of SAIPH scores with changing suicidality in a longitudinal cohort.

Having a proxy indicator of changing SI based on social media will allow for a longitudinal assessment of factors influencing changing SI trajectory, such as the pandemic or being responded to after a suicidal mention. As a third objective, we sought to use SAIPH to evaluate factors important for improving or worsening suicidal trajectory on social media following suicidal mention.

Methods

Study Design

This study consisted of 4 cohorts, including 2 recruited through independent online survey efforts and 2 random samplings of individual Twitter users collated with the Twitter application programming interface (API) through *tweepy* (version 3.8.0). Online survey cohort data were derived from the Cannabis Use,

XSL•FO
RenderX

Social Media, and Mental Health During the COVID-19 Pandemic Study—a study of undergraduate students at Carleton University in Ottawa (student survey cohort)—and a COVID-19 survey (COVID-19 survey cohort) [14,15]. Both datasets contained basic age and sex demographics, as well as various psychological scales. From these, mood and SI scores were derived as outlined below. Only individuals with active Twitter accounts (those accounts with posts present during this study's period) who consented to the analysis of tweet data were included in this report. No overlap was present in participants across cohorts.

Student Survey Cohort

Carleton University students were recruited through the university's online research system. Participants completed this study during the COVID-19 pandemic, between January and December 2020. Throughout most of this study recruitment (March to December), classes were all being offered online, campuses were closed, and several social restrictions were in place. Participants completed an online survey hosted on Qualtrics. Once the questionnaires were completed, all participants received an online debriefing form and were compensated with course credit.

In total, 159 individuals consented to the use of their Twitter data, enabling the generation of SAIPH scores (mean age 19.32, SD 1.85 y; female to male ratio=67.1%). SI metrics were derived from question 9 on the Beck Depression Inventory (BDI), which was temporally nested in the past 7 days:

- 0=I don't have any thoughts of harming myself.
- 1=I have thoughts of harming myself, but I would not carry them out.
- 2=(A) I feel I would be better off dead, (B) I have definite plans about committing suicide, (C) I feel my family would be better off if I were dead.
- 3=I would kill myself if I could.

COVID-19 Survey Cohort

A longitudinal online survey assessing the financial and psychological correlates of the COVID-19 pandemic was distributed via websites, social media, and multiple organizations and hospitals across Canada (please see the full list in the *Acknowledgments* section) [14,15]. The sole inclusion criterion was to be aged 12 years or older. Data included in this report were collected between April 2020 and December 2020. All individuals included in the final sample were from Canada. The survey included the Perceived Stress Scale, the Generalized Anxiety Disorder-7 scale, and the Self-Report version of the Quick Inventory of Depressive Symptomatology (QIDS-SR16).

From the parent study containing 4298 consenting individuals, a total of 307 had Twitter data and consented to their use to enable the generation of SAIPH scores (mean age 49.04, SD 16.53 y; female to male ratio=64.2%). SI was derived using question 12 on the QIDS-SR16, which was also temporally nested in the past 7 days:

- 0=I do not think of suicide or death.
- 1=I feel that life is empty or wonder if it's worth living.
- 2=I think of suicide or death several times a week for several minutes.
- 3=I think of suicide or death several times a day in some detail, or I have made specific plans for suicide or have actually tried to take my life.

For SI prediction, we modeled SAIPH score data over a range of time frames before the survey completion days to predict SI scores greater than or equal to 1, 2, and 3 in both cohorts and to discriminate between individuals above and below these thresholds. As a subset of 193 individuals consented to longitudinal follow-up and completed surveys at multiple time points, we modeled SI prediction using data from each survey completion period in addition to the baseline survey completion in a manner commensurate with the suicidal mention "event" modeling in our original publication [5].

For proxy SAIPH score analysis (below), 240 individuals from the total cohort with 3 or more longitudinal survey entries and no Twitter accounts were selected.

Algorithmic Methods

Overview

Algorithmic methods for SAIPH, including hyperparameter optimization and performance evaluation, are detailed in our original publication by Roy et al [5], where a higher SAIPH score is indicative of a higher risk of SI. Novel algorithms used in this paper involve the generation of personalized, future SAIPH score–predicting models and the generation of regional proxies for individuals not on Twitter.

Personalized, Future SAIPH Score–Predicting Model

For a given individual, an ensemble of linear kernel support vector regression (SVR) models is generated using the *LinearSVR* function in *scikit-learn* (French national Institute for Research in Computer Science and Automation, version 0.21.3) using a stacked generalization approach, trained on mean daily SAIPH data per individual with the outcome being the next day's mean daily SAIPH value. Training data windows are created separately per individual based on the following. Twitter timeline data are downloaded and processed with SAIPH, after which the number of days for all available past SAIPH data is calculated and divided by 2 to generate a maximum range of possible windows. SVRs are then trained with a minimum window of 14 days and every 7 days prior until the maximum window value is reached or exceeded. Future SAIPH scores are generated through the averaging of all model predictions derived from the most recent data within the model input data range. Linear SVRs were selected over other machine learning methods due to their training speed and success in generating scores correlated with actual future SAIPH scores at the algorithm-building stage (data not shown). A validation strategy comparing predicted values to true metrics was selected over internal k-fold cross-validation strategies and hyperparameter optimization due to the algorithmic generation of multiple machine learning models per individual.

XSL•FO
RenderX

Personalized Models to Score Responses to Suicidal Mention Per Individual

We sought to gather and score responses to suicidal mentions on Twitter. Leveraging our existing corpus of suicidal mentions collected between 2016 and 2019 published previously [5], we designed a web interface program using *Selenium* in Python (Selenium Project, GitHub) to pull up the original Tweet ID for the suicidal mention from individuals in the independent testing (not training) set and scrape any direct responses occurring in the thread of that tweet. This method was necessary as the Twitter API allows only for the gathering of response data within the API usage window of 7 days. We limited responses to within 72 hours of the original tweet. From the corpus of 291 suicidal mentions and after eliminating any mentions that were retweets and those for which an account no longer existed, we generated a set of 125 individuals. Each suicidal mention was treated as a separate event, resulting in a total of 471 suicidal mentions, after which the mean SAIPH score per day for 21 days was calculated to act as an indicator of changing SI trajectory over time. We also collected a replication dataset in real time using the Twitter API over several periods during 2019, 2020, 2021, and 2022. Individuals with SI were found by querying for the term "I am suicide thinking, planning." Tweets were read for those believed to be credible SI by 2 independent raters, after which a custom Python (Python Software Foundation) script was designed to pull responses to that tweet using the Twitter API. Retweets were excluded, while in both cohorts, responses to the original tweet from the original individual who mentioned SI were excluded.

Next, we generated a method to score responses with a custom NLP algorithm designed to assess the personalized relevance of the response for the individual. Briefly, for each individual with a suicidal mention, this personalized model is generated by using SAIPH to score available historical tweets before the SI mention, from which the top and bottom first percentile of tweets were binarily coded using bag-of-words and used to train a personalized SVR model. We can then apply that person's specific model to the responses they received after their suicidal mention, where messages with low and high scores represent user model–specific positive (favorable) and negative (unfavorable) messages, respectively.

Ethical Considerations

For SI mention and response cohorts on Twitter and other randomly selected individuals from Twitter, no participant consent was obtained. According to Canada's *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*—Tri-Council Policy Statement 2 Article 2.5, research on data "in the public domain and the individuals to whom the information refers have no reasonable expectation of privacy (nonintrusive, does not involve direct interaction between the researcher and individuals through the internet)" is exempt from institutional review board or research ethics board (REB) review. This policy is consistent with US regulations where according to US Department of Health and Human Services Policy 45 CFR 46.104 under the heading "exempt research, Secondary research for which consent is not required," secondary research uses of identifiable private information or identifiable biospecimens are exempt if the identifiable private information or identifiable biospecimens are publicly available. Exemption for the need for REB review was confirmed through the Office of Research Ethics at the Royal Institute of Mental Health Research.

All participants in the COVID-19 survey cohort and student survey cohort gave informed consent. Study protocols for the COVID-19 survey cohort were approved by the Clinical Trials Ontario—Qualified Research Ethics Board via the Ottawa Health Science Network (protocol #2131), and those for the student survey cohort were approved under REB #111775.

Statistical Analysis

All statistical tests were performed in R (R Foundation for Statistical Computing). Using an Anderson-Darling test from the *nortest* package, all distributions of data that rejected the null hypothesis of normality were subsequently evaluated with nonparametric tests or robust regression. All statistical tests performed were 2-tailed, and $P<.05$ was considered significant. No correction for multiple comparisons was performed. AUC metrics were generated using the *pROC* package in R. Significance of the AUC was performed using a Monte Carlo card sorting task, randomly permuting case and control status 10,000 times to generate a null distribution of the data for each AUC comparison reported. Where data are referred to as being corrected, this involved subtracting linearly modeled residuals from the dependent variable.

Results

Independent Association of SAIPH Scores With Quantitative SI Metrics

We investigated SAIPH scores for association with continuous SI metrics from 2 independent mental health metric surveys, including one of Carleton University students in Ottawa, Ontario (student survey cohort; n=153), and one in a national sample of respondents (COVID-19 survey cohort; n=307). Using a sliding window approach and progressively integrating data from successive days before the survey completion day, we observed significant interactions of the mean SAIPH score during these periods and the number of social media posts on the quantitative SI score across both the student survey cohort and the COVID-19 survey cohort (Table S1 and Figure S1 in Multimedia Appendix 1). One day before survey completion, a significant association is observed in the student survey cohort with the quantitative BDI-9 SI score (interaction β=.026, SD 0.011; $F_{4,54}$=1.93, P=.02), while a nonsignificant trend in the same direction is observed on this day for the COVID-19 survey cohort with the quantitative QIDS-SR score (interaction β=.0017, SD 0.001; $F_{4,353}$=3.49, P=.09; Table S1 in Multimedia Appendix 1). Consistent overlapping of significant associations next occur at 12 days (student survey cohort interaction β=.038, SD 0.014; $F_{4,94}$=3.3, P=.01; and COVID-19 survey cohort interaction β=.0035, SD 0.0016; $F_{4,493}$=2.9, P=.03) and remain significant for up to 16 days (Figure 1A, and Table S1 and Figure S2 in Multimedia Appendix 1). In general, in the student survey cohort, all assessed windows exhibited either significant or nonsignificant trend level associations, while in the

XSL•FO
RenderX

COVID-19 survey cohort, similar relationships were observed except for the period of 2 to 11 days after the survey completion day (Figure 1A, and Table S1 and Figure S1A in Multimedia Appendix 1). A post hoc analysis of the combined dataset demonstrated the most robust associations with a minimum of 3 tweets per person per day at 1, 13, and 40 days before the survey completion day (Figure S3 in Multimedia Appendix 1).

Figure 1. SI associations with SAIPH scores. (A) A plot of the negative natural log of the *P* value (y-axis) of an interaction of the mean daily SAIPH score and number of posts per day on quantitative SI scores as a function of the time in days from the day of survey completion (x-axis). Vertical lines set at 12 and 16 days before the survey completion day demonstrate the earliest period of overlapping significant associations between the student and COVID-19 survey cohorts. Box plots depicting the average 14-day SAIPH score from individuals scoring 0, 1, 2, or 3 on the SI question from the (B) student survey cohort (BDI question 9) and (C) the COVID-19 survey cohort (QIDS-SR16 question 12). BDI: Beck Depression Inventory; QIDS-SR16: Self-Report version of the Quick Inventory of Depressive Symptomatology; SAIPH: Suicide Artificial Intelligence Prediction Heuristic; SI: suicidal ideation.

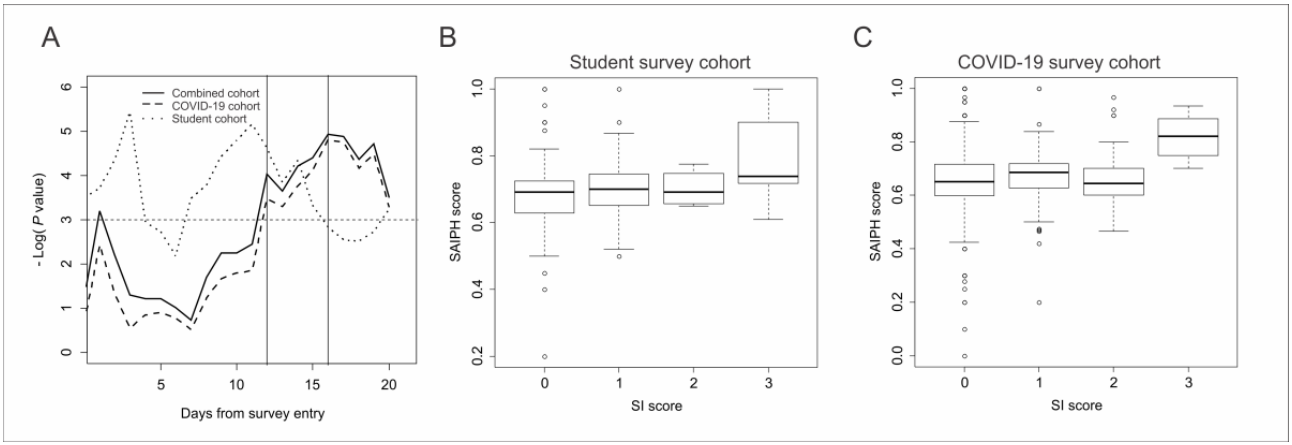

Assessment of AUC metrics for binary prediction was also performed on both raw and imputed SAIPH score data and demonstrated similar temporal patterns to those identified in the above interaction models (Sections S1-S3, Tables S2-S4, Figure S4, and Figure S5 in Multimedia Appendix 1; Figures 1B and1C). Cumulatively, the data demonstrate evidence for a moderate association of SAIPH scores with quantitative SI metrics in the initial days before survey completion but that these results become more robust on the integration of more data within the 12- to 16-day period. These results are consistent with our previous work that demonstrated that mean SAIPH score data from a multiple-day window performed better at predicting SI than that from smaller window periods [5]. Notably, SAIPH scores were associated with perceived stress and not depression or anxiety (Section S4 in Multimedia Appendix 1).

Temporal Assessment of SI Risk

We assessed variations in SAIPH scores over time relative to changes in SI levels in longitudinal follow-up data available in 109 individuals from the COVID-19 survey cohort. Of these, a total of 9 individuals exhibited levels of SI of 1 or greater at a minimum of 1 time point and had tweet information available on the day of survey completion. For this analysis, data points were derived from each successive pair of survey entries. In these participants, we observed a nonsignificant trend in the slope of mean SAIPH scores derived from a survey completion day for each pair of survey completion days relative to the slope of SI levels between 2 data entries (modeled as a function of time between data entry points; n=79 data points; ρ=0.21, *P*=.06).

To increase the available data beyond those who tweeted on the survey response days, we assessed the slope of mean daily SAIPH scores of all days between survey response points relative to the slope of the SI scores as a function of the time in days between survey entries in 109 individuals, of which 35 had variation in SI scores. We observed a significant interaction of the slope of SAIPH scores with the time in days between survey entries on the between survey entry SI slope using a linear regression (β=.00462, SD 0.00085; $F_{4,282}$=15.36, *P*<.001). This result remained significant when limiting the sample to 35 individuals demonstrating changes in SI levels (β=.0097, SD 0.0024; $F_{4,93}$=15.9, *P*<.001), which remained significant when applying robust regression (β=.013, SD 0.0013; *P*<.001). We performed a sliding window post hoc analysis, limiting the interrogated sample to a range of data points occurring below a range of thresholds from 10 days to 100 days. Significant positive associations of the slope of SI scores and SAIPH score slopes occurred before 15 days (eg, 14 days ρ=0.27, *P*=.04; Figure S6 in Multimedia Appendix 1). Notably, a sliding window analysis evaluating the slope of SAIPH scores between only the 2 survey entry days exhibited nonsignificant trends toward positive associations at longer periods (Figure S6 in Multimedia Appendix 1), suggesting the possibility that variation in SI during longer periods between surveys may be detected by SAIPH and affect the measured slope over longer periods. Similar results were observed when evaluating the slope of perceived stress with the slope of SAIPH scores between periods (Figure S7 in Multimedia Appendix 1). Together, these data suggest that changes in mean SAIPH scores over time reflect temporal changes in SI and can be used as an indicator of changing suicidal distress. For example, the application of SAIPH scores to social media profiles before and after the beginning of the COVID-19 pandemic in 2020 demonstrates significant postpandemic deviations in SAIPH scores between individuals endorsing moderate SI as compared to those who do not (Section S5 and Figure S8 in Multimedia Appendix 1).

XSL•FO
RenderX

Suicide Indicator Trajectory as a Function of Responses to Suicidal Mentions

In light of data suggesting that we could use the SAIPH scoring of Twitter profiles as an indicator of changing suicidal distress over time, we sought to investigate factors that may influence the SAIPH score trajectory after a suicidal mention. To that end, we used 471 suicidal mentions from the corpus of individuals derived from the independent test cohort of our originally published SAIPH paper [5]. Of these, 63 had responses to the suicidal mention within 72 hours, while 408 did not. The mean SAIPH score following each mention was used as an indicator of changing suicidal distress over time. We observed a significant negative association of mean daily SAIPH score in those 63 individuals with responses (ρ=–0.52, *P*=.01), while the group with no responses exhibited no change over time (ρ=–0.27, *P*=.22; Figure 2A). Furthermore, the last 7 days of SAIPH scores in the response group were significantly lower than that of the nonresponse group (Student 2-tailed *t* test: response group, mean 0.72, SD 0.072 and no-response group, mean 0.74, SD 0.085; *P*<.001). These data suggest that individuals who received responses to their suicidal mentions were more likely to improve over 3 weeks (Figure 2A).

Figure 2. SAIPH trajectory as a function of responses to suicidal mentions. (A) Line graph depicting the mean daily SAIPH score (y-axis) as a function of time in days from suicidal mention (x-axis) in a cohort of individuals with responses to their suicidal mentions (blue) and those with no responses (red). (B) Line graph depicting the mean daily SAIPH score (y-axis) as a function of time in days from suicidal mention (x-axis) in a cohort of individuals with responses scoring as "personally positive" (favorable; blue) and "personally negative" (unfavorable; red) in our originally published corpus of suicidal mentioners from 2016 to 2019. (C) Line graph depicting the mean daily SAIPH score (y-axis) as a function of time in days from suicidal mention (x-axis) in a cohort of individuals with responses scoring as "personally positive" (favorable; blue) and "personally negative" (unfavorable; red) in a replication corpus of those with suicidal mentions collected from 2019 to 2022. SAIPH: Suicide Artificial Intelligence Prediction Heuristic; SI: suicidal ideation.

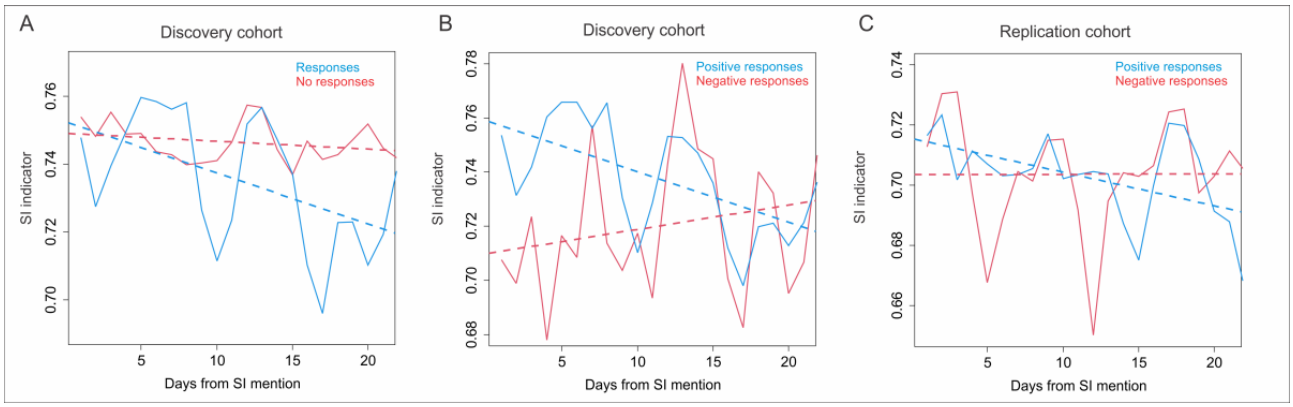

As some suicide intervention techniques such as ASIST (Applied Suicide Intervention Skills Training) focus on identifying concepts important to an individual as a key step in convincing them to focus on staying safe, we next reasoned that the quality of responses to suicidal mentions may further influence the trajectory of improvement. To that end, we designed an algorithm to score the importance of responses to suicidal mentions based on the social media profile of users with suicidal mentions. For each individual, all individual tweets were scored with SAIPH, and the 1st (highest scoring) and 99th percentile (lowest scoring) were used to train an SVR per individual. This person-specific model was then used to score those responses to their suicidal mentions and derive a personalized "positivity" score. Using the mean positivity score of 0.243 (SD 0.383) as a threshold, we split the group into 43 individuals with highly positive (favorable) versus 20 highly negative (unfavorable) responses. We observed a significant declining temporal direction in SAIPH scores in the 3 weeks following suicidal mention in those with positive (favorable; ρ=–0.60, *P*=.003) but not negative (unfavorable; ρ=0.22, *P*=.33) responses (Figure 2B). To validate this finding, we collected a replication cohort of those with suicidal mentions from 2019 to 2022 and generated personalized positivity scores for responses to suicidal mentions. Using the same threshold to segregate high and low responders, we observed a similar significant declining temporal direction of association of mean daily SAIPH scores with time over the following 3 weeks in the highly positive response group (N=27, ρ=–0.51, *P*=.02), while no association was observed in the highly negative response group (N=12, ρ=0.0038, *P*=.99; Figure 2C). Together, the data suggest that receiving responses to suicidal mentions on Twitter and specifically those personally scored as positive for an individual is associated with improving SAIPH scores over time following suicidal mentions. We next assessed for overrepresentation of word usage in responses above and below the personalized score threshold in the combined sample of responses from 2016 to 2022. As some individuals received multiple responses for their suicidal mentions, this resulted in a dataset of 99 negatively scored responses and 243 positively scored responses. Among these, no words were enriched in the positively (favorably) scored response corpus (Table S4 in Multimedia Appendix 1); however, in the negatively (unfavorably) scored corpus, several seemingly positive words such as "hope" were significantly overrepresented (Table S5 in Multimedia Appendix 1). Performing the same analysis for pairs of co-occurring words showed that several word pairs exhibited nonsignificant trends for enrichment in the negatively (unfavorably) scored corpus, including "everything will" and "be alright," while none were significant in the positive (favorable) corpus (Table S4 in Multimedia Appendix 1). We next divided the corpus of responses into those where the user with a suicidal mention had a positive slope of SAIPH scores (becoming worse; n=135 responses) following their suicidal mention, or a negative slope (becoming better; n=191

XSL•FO
RenderX

responses). Again, no enrichment for words in those with improving trajectories was observed (Table S5 and Figure S9 in Multimedia Appendix 1); however, several word pairs were identified in those without improving trajectory such as "please don't" and "do it," and the most significant single word was "don't" (Table S5 in Multimedia Appendix 1).

Discussion

Principle Findings

We performed a series of analyses toward the end goal of validating our SI prediction model, SAIPH [5], in 2 cohorts that provided online survey data based on psychometrically validated scales with suicide questions, and in random samples of Twitter users. Both cohorts represented different demographics of the population, with the student survey cohort deriving from an undergraduate university cohort and the COVID-19 survey cohort deriving from individuals with an expansive range of ages throughout the life span. We were limited in our ability to perform a prediction analysis like our original publication due to the small number of high SI scores in both cohorts, as well as the fact that many individuals with high SI scores did not have a large body of tweets in the initial days before survey completion. While our supplementary results identified elevated AUC values across both independent cohorts, these derive from small sample sizes and must be interpreted with caution. As such, we performed linear modeling against the quantitative SI scores to increase the statistical robustness of our analysis and identified significant interactions of SAIPH scores with the number of tweets per day in both cohorts. The interactions demonstrate that SAIPH scores derived from a higher number of tweets on average are more robust. While the associations from the cumulative data from 50 days before the survey completion were significant in both cohorts, we performed a sliding window analysis to identify periods of peak association at periods more temporally proximal to the survey completion day. We observed a general agreement across both cohorts by periods, with more robust associations at approximately 2 weeks before survey completion and, to a lesser degree, on the day before survey completion. The overlapping significance across both cohorts around 2 weeks is consistent with our previous work that showed a higher frequency of high-risk scoring SAIPH scores preceded death by suicide at approximately 20 days before death, while the elevated signal just before the survey completion day is consistent with our previous work that demonstrated the highest predictive accuracy of SI admissions on Twitter using data most temporally proximal to the SI mention. All associations derived from integrating over 42 days of SAIPH scores were significant across both cohorts, which is consistent with the observation that the association with SI is more robust when there is a higher amount of tweet data from which to derive the score. These findings are consistent with our previous work that demonstrated that mean SAIPH score data from a multiple-day window performed better at predicting SI than that from smaller window periods [5]. Cumulatively, both the temporal patterns and observed associations were largely consistent across both cohorts, adding confidence that SAIPH is detecting SI-associated signals.

The SAIPH algorithm is developed on established suicide theories, integrating scores of intermediate suicide-associated constructs such as loneliness, stress, depression, and perceived burdensomeness, among others. Recent work by Ophir et al [6] has demonstrated the success of comparative models using the integration of suicide theory over merely leveraging a bulk of social media posts, suggesting the integration of these constructs may be important for model performance.

It is our view that it is important to consider the application of suicide risk–detecting tools such as SAIPH and others in the context of the challenges of the suicidal state and in what area of suicide prevention these tools may generate the most benefit. One major challenge of suicide risk prediction is the so-called "low base-rate problem" [16], wherein even the best-performing predictors will generate large numbers of false positives. In light of this major problem, it is our view that risk prediction methods could best be applied in contexts where the interventions are tolerable to both false positives and the prevention systems in which they are deployed. For example, it may not be appropriate to deploy in a clinical setting where the response to a false positive will use vital resources in short supply. By contrast, with the presence of future temporal risk predicting tools, one possibility becomes providing a just-in-time adaptive intervention, such as those leveraging digital data sources to prompt the timing of intervention [17], in the community at a time before suicidal crisis that is tailored such that it is tolerable to false positives. One example for which we envision SAIPH might be developed and ultimately applied would be prompting peer support interventions from a list of confidants (such as friends) selected by the at-risk individual, who could be prompted digitally to reach out at an individually set threshold of risk representing a period before the crisis. Such an intervention, if delivered before a period of crisis, would leverage the protective nature of connectedness without the need for mental health intervention training, where the "intervention" might be an invitation to social interaction such as talking, shared activities such as sports, hiking, or going out socially, for example. This may provide protective benefits while also being tolerable to false positives, would be scalable, and could be evaluated for efficacy.

Our temporal analysis in individuals from the COVID-19 survey cohort demonstrated that variation of average daily SAIPH values within people was associated with variation in SI scores over time. This association was strongest when evaluating survey metrics repeated in shorter time frames. We posit that longer periods between surveys may have periods of higher or lower distress that might be marked by variations in SAIPH score that are not captured by survey time points. This supposition is supported by the observation that limiting measurements to those taken on the day of survey completion despite longer periods exhibits a trend toward positive association with the slope of SI changes over the period. Taken together, SAIPH scores may be capable of acting as an indicator of changing SI over time.

We evaluated the ability of SAIPH scores to mark changes in perceived stress, anxiety, and depression. Within participants, the strongest association we observed was with perceived stress. The ability of SAIPH to mark changing perceived stress may

XSL•FO
RenderX

be linked with its association as an indicator of changing SI within participants. Notably and consistent with this hypothesis, we observed significant deviations of high SI endorsing individuals with low SI or control individuals as a function of the declaration of a national emergency and the beginning of social distancing due to the start of the COVID-19 pandemic in the middle of March 2020 (Section S5 in Multimedia Appendix 1). Based on the stress diathesis theory, our interpretation is that vulnerable individuals may have experienced higher distress as a function of this shared stressor relative to more resilient controls. This analysis demonstrates the potential efficacy of using SAIPH as an indicator of SI risk and perceived stress to study temporal trends relative to major events, both on an individual and a population basis. This work is consistent with our previous work that demonstrated that averaging of regionally and randomly sampled SAIPH scores from certain areas was associated with suicide death rates in those regions [5].

In light of our data demonstrating that we could use SAIPH as an indicator of changing SI, this allowed us the opportunity to evaluate the effect of responses to suicidal mentions in 2 cohorts. The strongest result we observed was that individuals who had responses to their suicidal mentions demonstrated a significant improvement in their SAIPH scores over the following 3 weeks, suggesting that responding to those in crisis may contribute to their improvement. This result is intuitive and helps corroborate the ability of SAIPH to act as a proxy of SI trajectory. We next attempted to look for a more nuanced scoring method for the responses to qualitatively score them relative to the individual. The impetus for this theory came from the ASIST suicide intervention that teaches that identification of factors important to the individual is a key step in getting the individual with SI to agree to be safe for the moment. As such, our algorithm effectively measures content in their historical posts and builds a new model relative to their most extreme high and low content. If a response to this person mentions something that was positively reflected in their timeline, it will receive a score denoting it is a “positive” response. By building machine learning models specific to the individual, we demonstrated that we could score the response to their suicidal mention and that those with low scores, denoting a response scored as “positive” to the user with a suicidal mention, were more likely to improve. In our analysis of words and 2-word combinations in responses to those with suicidal mentions, no words appeared overrepresented in those individuals who improved over the following 3 weeks as measured by SAIPH. This may be because, as hypothesized, the responses important for someone to improve contained more contextual and personalized content relevant to the individual and, as such, were not overrepresented across the group. By contrast, words such as “don’t” and phrases such as “please don’t,” “do it,” “everything will,” or “be alright” were overrepresented in responses to those with suicidal mentions that did not improve over the 3-week period. This finding suggests that more standard responses may lack the personal context important for an individual’s improvement. These findings, if corroborated, could lead to the development of a tool allowing for quick scanning and identification of topics of importance to a crisis line user who consents to social media scanning. In combination, we envision a possible application of both the SAIPH method and personalized importance metric that could be used by crisis line responders where, through the future integration of large language models and these scores, would allow responders to interrogate the user’s risk and protective factors preceding and following a user’s historical periods of highest risk to gain insight into the person in need. As individuals with SI in crisis can experience cognitive constriction [18-20] and are not always in a mindset for collaborative problem-solving with the crisis responder, these tools could augment suicide intervention efforts, safety planning, and problem-solving and should be further developed. Beyond prompting an ideal response, the data suggest that any response is better than no response. These findings, while interesting, are preliminary and should be further replicated in larger datasets to validate them and further explore their implications.

Limitations

This study is not without limitations. The SI metrics, while collected as part of psychometrically validated surveys, represented answers to a single question and not, by contrast, the cumulative scores from comprehensive suicide risk scales. Despite this, these metrics are stronger than suicidal mentions on social media as a measure of suicide risk for external validation, as they represent data collections occurring independently from a participant’s social media use and are answers to a standardized set of questions. This fact partially addresses the legitimacy that suicidal mentions on social media are linked to actual suicidal intent as demonstrated in 2 studies from Japan [2,3], and by the fact that our model performs best for identifying BDI SI level 3, which states a desire to kill oneself if possible and thus suggests a level of intent. Despite this, the number of individuals endorsing high levels of SI in both survey cohorts was small and, in some cases, the amount of Twitter data available at the times closest to survey completion was lacking, necessitating the creation of our imputation algorithm. Future work should investigate the efficacy of data imputation using our future predicting SVR method to assess if this improves the robustness of SI prediction in individuals with less data. The primary strategy used involved looking for associations with quantitative SI scores using linear models. The use of the sliding window approach allowed us to investigate the times of most significant association; however, a limitation of this analysis is that we did not correct for multiple testing. Data gathered for this study occurred during the COVID-19 pandemic, which represented a rather anomalous year where in general, people’s usual routines, external influences, and environments were disrupted. In the context of the stress diathesis model of suicide risk, this may have caused a higher efficacy of our prediction methods and observed associations than would not have been observed in more normal times. Our analysis of SAIPH scores before and after this period indicated that the pandemic did seem to affect SAIPH scores in individuals at high suicide risk. Alternatively, it remains possible that the pandemic elevated SAIPH scores in individuals who are not suicidal and led to a weaker observed predictive accuracy than would have otherwise been observed in nonpandemic times. Despite this, the success of SAIPH in predicting SI during pandemic times suggests it is an efficacious tool for suicide risk prevention that is robust in the context of

XSL•FO
RenderX

external stressors. Future work should investigate if more personalized approaches, leveraging an individual's own data to prognosticate risk, may improve predictive performance.

Conclusions

In conclusion, the above study expands upon the previously published validation efforts of the previous model, SAIPH, on its ability to use Twitter profile data to associate with SI derived from external scales in 2 independent cohorts. Additionally, this study expands on our previous work, demonstrating an association of SAIPH scores with longitudinal changes in perceived stress, which can be used as an indicator of population changes in response to global stressors such as the COVID-19 pandemic. Moreover, our work produced a method to generate future SAIPH score predictions in a personalized manner. Finally, our study demonstrates that analysis of responses to suicidal mentions on social media and the resulting SI trajectory has the potential not only to generate novel algorithms for understanding risk but may inform novel and personalized interventions leveraging an individual's unique situation. Cumulatively, the work suggests that NLP techniques designed to infer mood from text-based social media content may be efficacious as indicators of individual suicide risk and periods of heightened stress. In the future, methods such as SAIPH may be integrated as tools to aid clinician decisions related to individual suicide risk. The psychiatry field does not have the extensive battery of objective diagnostic tests relative to other fields of medicine and will most likely benefit from the addition of more tools to aid in patient assessment. Longitudinal monitoring of suicide risk indicators on an individual scale will pave the way for the development of novel suicide interventions.

Acknowledgments

We would like to thank DIFD (Do It For Daron), the Mach-Gaensslen Foundation, and attendees of the Rideau View Inspirational Golf Classic for their generous support of this research. We wish to thank all the participants who gave their time to fill out these surveys. We thank the Royal Ottawa Mental Health Centre, the Centre for Addiction and Mental Health, the Clinical Investigation Unit at the Ottawa Hospital Research Institute, and the University of Ottawa Heart Institute for assistance with participant recruitment. We also thank the following organizations who helped in circulating the survey: Canadian Nurses Association, Canadian Physiotherapy Association, Canadian Association of Occupational Therapists, Ontario Public Health, Ontario Medical Association, Ontario Psychiatric Association, Ontario Society of Occupational Therapists, Canadian Counselling and Psychotherapy Association, Ordre des Psychologues du Québec, Mood Disorders Society of Canada, Canadian Arthritis Patient Alliance, Patients for Patient Safety Canada, the COVID-19 Resources Canada Platform, and Sleep On It! Canada.

Data Availability

In accordance with Twitter policies of data sharing, data used in the generation of the algorithm for this study will not be made publicly available. Due to the sensitive nature of study survey cohort data, the research ethics boards have not granted approval for these datasets to be for public dissemination. Due to the sensitive and potentially stigmatizing nature of this tool, code used for algorithm generation or implementation on individual Twitter profiles will not be made publicly available. All Python scripts were created in Python (version 3.6.8), and the machine learning algorithms used *scikit-learn* (version 0.21.3). Specific parameters used to generate, test, and process the current dataset are mentioned in detail in the *Methods* section. Access to Suicide Artificial Intelligence Prediction Heuristic (SAIPH) score generation for research purposes will be granted on reasonable request.

Authors' Contributions

RR, MS, KGCH, RJM, ZRP, TK, RLG, and AA were responsible for survey creation, data acquisition, and curation from the COVID-19 survey and the student survey. RR, RJM, and ZK were involved in the interpretation of the data. ZK was responsible for conception, data generation from Twitter, algorithm development, analysis, and paper writing.

Conflicts of Interest

ZK is a cofounder and acts as a consultant for Dionysus Digital Health Inc. He has received shares in exchange for a license to further develop and implement the Suicide Artificial Intelligence Prediction Heuristic (SAIPH) algorithm.

Multimedia Appendix 1
Supplementary results, figures, and tables.
[DOCX File , 3136 KB - jmir_v26i1e49927_app1.docx]

References

1. L. Belfort E, Mezzacappa E, Ginnis K. Similarities and differences among adolescents who communicate suicidality to others via electronic versus other means: a pilot study. Adolesc Psychiatry (Hilversum) 2012 Jul 01;2(3):258-262. [doi: 10.2174/2210676611202030258]
2. Sueki H. The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. J Affect Disord 2015;170:155-160. [doi: 10.1016/j.jad.2014.08.047] [Medline: 25240843]

XSL•FO
RenderX

3. Fahey RA, Boo J, Ueda M. Covariance in diurnal patterns of suicide-related expressions on Twitter and recorded suicide deaths. Soc Sci Med 2020;253:112960 [FREE Full text] [doi: 10.1016/j.socscimed.2020.112960] [Medline: 32251933]
4. Pourmand A, Roberson J, Caggiula A, Monsalve N, Rahimi M, Torres-Llenza V. Social media and suicide: a review of technology-based epidemiology and risk assessment. Telemed J E Health 2019;25(10):880-888. [doi: 10.1089/tmj.2018.0203] [Medline: 30362903]
5. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA. A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit Med 2020;3:78 [FREE Full text] [doi: 10.1038/s41746-020-0287-6] [Medline: 32509975]
6. Ophir Y, Tikochinski R, Asterhan CSC, Sisso I, Reichart R. Deep neural networks detect suicide risk from textual facebook posts. Sci Rep 2020;10(1):16685 [FREE Full text] [doi: 10.1038/s41598-020-73917-0] [Medline: 33028921]
7. Sawhney R, Joshi H, Gandhi S, Jin D, Shah RR. Robust suicide risk assessment on social media via deep adversarial learning. J Am Med Inform Assoc 2021;28(7):1497-1506 [FREE Full text] [doi: 10.1093/jamia/ocab031] [Medline: 33779728]
8. Naseem U, Kim J, Khushi M, Dunn A. Graph-based hierarchical attention network for suicide risk detection on social media. 2023 Presented at: WWW '23: The ACM Web Conference 2023; April 30 to May 4, 2023; Austin, TX p. 1995-1003. [doi: 10.1145/3543873.3587587]
9. Li Z, An Z, Cheng W, Zhou J, Hu B. Using label-text correlation and deviation punishment for fine-grained suicide risk detection in social media. 2022 Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2022); December 6-8, 2023; Las Vegas, NV. [doi: 10.1109/bibm55620.2022.9995476]
10. Xu W, Su C, Li Y, Rogers S, Wang F, Chen K, et al. Improving suicide risk prediction via targeted data fusion: proof of concept using medical claims data. J Am Med Inform Assoc 2022;29(3):500-511 [FREE Full text] [doi: 10.1093/jamia/ocab209] [Medline: 34850890]
11. Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, et al. Temporally informed random forests for suicide risk prediction. J Am Med Inform Assoc 2022;29(1):62-71 [FREE Full text] [doi: 10.1093/jamia/ocab225] [Medline: 34725687]
12. Burkhardt HA, Ding X, Kerbrat A, Comtois KA, Cohen T. From benchmark to bedside: transfer learning from social media to patient-provider text messages for suicide risk prediction. J Am Med Inform Assoc 2023;30(6):1068-1078 [FREE Full text] [doi: 10.1093/jamia/ocad062] [Medline: 37043748]
13. Liu X, Liu X, Sun J, Yu NX, Sun B, Li Q, et al. Proactive Suicide Prevention Online (PSPO): machine identification and crisis management for Chinese social media users with suicidal thoughts and behaviors. J Med Internet Res 2019;21(5):e11705 [FREE Full text] [doi: 10.2196/11705] [Medline: 31344675]
14. Robillard R, Daros AR, Phillips JL, Porteous M, Saad M, Pennestri M, et al. Emerging new psychiatric symptoms and the worsening of pre-existing mental disorders during the COVID-19 pandemic: a Canadian multisite study. Nouveaux symptômes psychiatriques émergents et détérioration des troubles mentaux préexistants durant la pandémie de la COVID-19: une étude Canadienne multisite. Can J Psychiatry 2021;66(9):815-826 [FREE Full text] [doi: 10.1177/0706743720986786] [Medline: 33464115]
15. Robillard R, Saad M, Edwards J, Solomonova E, Pennestri M, Daros A, et al. Social, financial and psychological stress during an emerging pandemic: observations from a population survey in the acute phase of COVID-19. BMJ Open 2020;10(12):e043805 [FREE Full text] [doi: 10.1136/bmjopen-2020-043805] [Medline: 33310814]
16. Large MM. The role of prediction in suicide prevention. Dialogues Clin Neurosci 2018;20(3):197-205. [doi: 10.31887/dcns.2018.20.3/mlarge]
17. Coppersmith DDL, Dempsey W, Kleiman EM, Bentley KH, Murphy SA, Nock MK. Just-in-time adaptive interventions for suicide prevention: promise, challenges, and future directions. Psychiatry (New York) 2022;85(4):317-333 [FREE Full text] [doi: 10.1080/00332747.2022.2092828] [Medline: 35848800]
18. Heinrich M, Berzlanovich A, Willinger U, Eisenwort B. Measurement of cognitive constriction in suicide notes [Article in German]. Neuropsychiatrie 2008;22(4):252-260. [Medline: 19080996]
19. Keilp JG, Gorlyn M, Russell M, Oquendo MA, Burke AK, Harkavy-Friedman J, et al. Neuropsychological function and suicidal behavior: attention control, memory and executive dysfunction in suicide attempt. Psychol Med 2013;43(3):539-551. [doi: 10.1017/s0033291712001419]
20. Fernández-Sevillano J, González-Pinto A, Rodríguez-Revuelta J, Alberich S, Gónzalez-Blanco L, Zorrilla I, et al. Suicidal behaviour and cognition: a systematic review with special focus on prefrontal deficits. J Affect Disord 2021 Jan 01;278:488-496 [FREE Full text] [doi: 10.1016/j.jad.2020.09.044] [Medline: 33017675]

Abbreviations

API: application programming interface
ASIST: Applied Suicide Intervention Skills Training
AUC: area under the receiver operator characteristic curve
BDI: Beck Depression Inventory

XSL•FO
RenderX

NLP: natural language processing
QIDS-SR16: Self-Report version of the Quick Inventory of Depressive Symptomatology
REB: research ethics board
SAIPH: Suicide Artificial Intelligence Prediction Heuristic
SI: suicidal ideation
SVR: support vector regression

Edited by A Mavragani; submitted 13.06.23; peer-reviewed by L Shang, T Tillmann, J Washburn; comments to author 20.01.24; revised version received 01.05.24; accepted 08.09.24; published 05.12.24.

Please cite as:
Kaminsky Z, McQuaid RJ, Hellemans KGC, Patterson ZR, Saad M, Gabrys RL, Kendzerska T, Abizaid A, Robillard R
Machine Learning–Based Suicide Risk Prediction Model for Suicidal Trajectory on Social Media Following Suicidal Mentions: Independent Algorithm Validation
J Med Internet Res 2024;26:e49927
URL: https://www.jmir.org/2024/1/e49927
doi:10.2196/49927
PMID:

©Zachary Kaminsky, Robyn J McQuaid, Kim GC Hellemans, Zachary R Patterson, Mysa Saad, Robert L Gabrys, Tetyana Kendzerska, Alfonso Abizaid, Rebecca Robillard. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Automated Conversational Agent Self-Help Program: Randomized Controlled Trial

Heather M Foran[1], PhD; Christian Kubb[1], Dr rer nat; Janina Mueller[1], Dr rer nat; Spencer Poff[2], BS; Megan Ung[2], BS; Margaret Li[2], BS; Eric Michael Smith[2], PhD; Akinniyi Akinyemi[2], BS; Melanie Kambadur[2], PhD; Franziska Waller[1], MSc; Mario Graf[1], Dr rer nat; Y-Lan Boureau[2], PhD

[1]Department of Health Psychology, University of Klagenfurt, Klagenfurt, Austria

[2]Meta AI Fundamental Research, Menlo Park, CA, United States

Corresponding Author:
Heather M Foran, PhD
Department of Health Psychology
University of Klagenfurt
Universitaetstrasse 65067
Klagenfurt, 9020
Austria
Phone: 43 46327001641
Email: Heather.Foran@aau.at

Abstract

Background: Health promotion and growth-based interventions can effectively improve individual well-being; however, significant gaps in access and utilization still exist.

Objective: This study aims to develop and test the effectiveness and implementation of a new, widely targeted conversational agent prevention program (Zenny) designed to enhance well-being.

Methods: A total of 1345 individuals in the United States were recruited online and randomly assigned to either (1) a self-help program intervention delivered via an automated conversational agent on WhatsApp or (2) an active control group that had access to evidence-based wellness resources available online. The primary outcomes were well-being (measured using the 5-item World Health Organization Well-being Scale), psychosocial flourishing (assessed with the Flourishing Scale), and positive psychological health (evaluated with the Mental Health Continuum-Short Form). Outcome measures were collected at baseline and again 1 month postassessment. All analyses were conducted using an intention-to-treat approach.

Results: Both groups showed significant improvements in well-being (self-help program intervention group effect size: Cohen d=0.26, $P<.001$; active control group effect size: d=0.24, $P<.001$), psychosocial flourishing (intervention: d=0.19, $P<.001$; active control: d=0.18, $P<.001$), and positive psychological health (intervention: d=0.17, $P=.001$; active control: d=0.24, $P<.001$) at postassessment. However, there were no significant differences in effectiveness between the 2 groups (P ranged from .56 to .92). As hypothesized a priori, a greater number of days spent actively engaging with the conversational agent was associated with larger improvements in well-being at postassessment among participants in the intervention group ($\beta=.109$, $P=.04$).

Conclusions: The findings from this study suggest that the free conversational agent wellness self-help program was as effective as evidence-based web resources. Further research should explore strategies to increase participant engagement over time, as only a portion of participants were actively involved, and higher engagement was linked to greater improvements in well-being. Long-term follow-up studies are also necessary to assess whether these effects remain stable over time.

Trial Registration: ClinicalTrials.gov NCT06208566; https://clinicaltrials.gov/ct2/show/NCT06208566; OSF Registries osf.io/ahe2r; https://doi.org/10.17605/osf.io/ahe2r

(J Med Internet Res 2024;26:e53829) doi:10.2196/53829

KEYWORDS

well-being; chatbot; randomized controlled trial; prevention; flourishing

XSL•FO
RenderX

Introduction

Background

Subjective well-being is relevant not only to psychological health but also to social relationships, work productivity, job satisfaction, and other health domains [1,2]. Fortunately, a large body of scientific knowledge exists on methods for promoting subjective well-being [3]. Health promotion and growth-based interventions can effectively enhance individuals' life satisfaction [4].

Despite the promise of these interventions, they are rarely disseminated or implemented at the population level. This limited reach is due to inadequacies within health care systems, economic barriers, perceived stigma, and low accessibility of evidence-based health promotion tools [5,6]. A positive, growth-based approach focused on building self-help skills, rather than addressing problems, may help reduce stigma at the community level. Skills essential for well-being—such as emotion regulation, problem-solving, and other cognitive behavioral techniques [7,8]—can be taught early and may help prevent mental health issues later in life.

Internet-Based or App-Based Programs

Several internet-based interventions for subjective well-being are supported by evidence from randomized controlled trials [9-11]. Delivering prevention programs online may help reduce barriers such as appointment scheduling, time commitments, travel, and intervention costs [12]. Additionally, offering these interventions in a private, self-help format may lower the perceived stigma associated with psychological self-improvement. Previous research has found that chatbot technology can be a valuable tool for reducing the stigma associated with mental illness and improving access to mental health resources [13]. Another common reason individuals avoid formal services is a preference to manage their problems independently [14]. Self-help programs offered in a self-paced format may foster a greater sense of personal improvement.

The Next Step for Prevention

Despite their potential, many online interventions remain static and lack personalization to address individual needs, resulting in low dissemination and high dropout rates in trials [15,16]. Additionally, the limited availability of free online interventions reduces their public health impact. Automated conversational agents could provide a more tailored approach to enhance engagement and accessibility. A growing body of research explores conversational agent–based interventions in mental health, utilizing diverse study designs (eg, pilot feasibility studies, randomized controlled trials, quasi-experimental studies), technologies (eg, conversational agents accessed through apps, online platforms, offline computer programs, virtual reality headsets), intervention approaches (eg, cognitive behavioral therapy, psychoeducation), and mental health outcomes (eg, depression, anxiety, psychological distress, well-being) [17-20].

One of the earliest studies in this area tested a conversational agent named Woebot in a small sample of students (N=70) with elevated depressive and anxiety symptoms, yielding promising results [21]. Participants reported high satisfaction and showed improvements in depressive symptoms. However, the relevance of large-scale prevention remains uncertain, as the sample was selected based on elevated symptoms and consisted of a convenience sample of university students. In recent years, several other conversational agents targeting well-being or related constructs have been introduced, such as Shim [22], EMMA [23], Vivibot [24], and the 21-Day Stress Detox [25]. Similar to Woebot, their relevance and effectiveness for universal prevention remain unclear, often due to small sample sizes—ranging from 28 [22] to 64 participants [25]—or the lack of control groups [26].

While there are promising conversational agent–based interventions designed to promote mental health and well-being, fewer studies assess their impact on well-being outcomes compared with other mental health outcomes [20]. Furthermore, previous reviews highlight a lack of robust experimental designs, particularly randomized controlled trials using nonclinical samples, to demonstrate the efficacy of these conversational agent interventions for mental health and well-being [17,27]. Additionally, a significant number of studies examining conversational agents have used various psychological interventions without a solid theoretical foundation [17] or lack standardized measurement outcomes [19]. Finally, fostering greater interdisciplinary collaboration between computer science and mental health disciplines may help facilitate progress in this area [18].

In this study, we aimed to develop and test the effectiveness of a new conversational agent prevention program designed to enhance well-being and promote psychological growth. This project focuses not only on assessing the program's effectiveness on well-being compared with an online active control condition in a randomized controlled trial but also on increasing accessibility for individuals who may benefit from such an intervention. Thus, the study uses brief, low-burden assessments instead of a full battery typical of psychological studies to enhance acceptability, enrollment rates, and retention rates. We used the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) implementation science model [28] as a framework for this study. By adapting the model for our digital intervention, we address the following components in this paper: recruitment strategies, response rates, and enrollment rates (reach); changes in outcomes (effectiveness); and user satisfaction and participation/engagement (implementation).

Hypotheses

- H1: The intervention group will show significant improvements in well-being and psychological flourishing at the 1-month postassessment compared with the control group.
- H2: The effectiveness of the intervention will be similar across gender and age after controlling for participation/engagement and user satisfaction.
- H3: The intervention group will have higher retention rates and report greater active participation/engagement than the control group.

XSL•FO
RenderX

- H4: Within the intervention group, higher user satisfaction and greater active participation/engagement will be associated with improved well-being.
- H5: Within the intervention group, higher user satisfaction while using the conversational agent will lead to increased participation/engagement over time.

Methods

Ethics Approval

The study was approved by the Western Institutional Review Board (WCG) in the United States (approval number 1289179). All hypotheses and planned analyses for this study were registered on the Open Science Framework platform [29] and in ClinicalTrials.gov. A priori power analysis is detailed in our study preregistration, indicating that the minimal sample size required to detect small effects in overall analyses was 216, with $\alpha<.05$ and 95% power. The minimal sample size required for this analysis to detect small effects is 328, based on 2 assessment time points ($\alpha<.05$ and 95% power). Copies of the informed consent forms can be requested by contacting the first author (HMF).

Participants

Participants were recruited through advertising campaigns on Facebook (Meta Platforms, Inc.) and Instagram (Meta Platforms, Inc.) from July 2022 to April 2023. Eligibility criteria included being at least 18 years of age, English speaking, and residing in the United States with online access. After providing online informed consent, participants completed a brief questionnaire to assess their eligibility, demographics, and well-being. Participants were randomly assigned to either the conversational agent intervention or a selection of 3 evidence-based wellness resources linked from the study website. No financial compensation was provided for participation. Safety and security procedures were implemented, and participants were given referrals throughout the study. The procedures for data protection were detailed in the informed consent forms.

The registration process was initiated 2054 times, which included an eligibility check, informed consent, and the preassessment survey, followed by randomization into one of the groups. A total of 1349 participants provided consent and completed the preassessment. Four participants were lost due to a technical error during the preassessment, resulting in a sample of 1345 participants who completed the preassessment and were available for intent-to-treat analyses. The study flowchart is presented in the Results section.

Study Design

This study uses a 2 (group: intervention and active control) × 2 (time: preassessment and 1-month postassessment) design. An overview of the full study is available in the clinical trials registry (NCT06208566). The intervention consisted of a conversational agent universal prevention program designed to enhance well-being. The control group received referrals to a selection of web-based, empirically supported mental health information. Completion of the interventions was expected to take approximately 1 month, depending on the individual level of participation and engagement.

Randomization

Participants were randomized using a 1:1 allocation ratio with a block size of 10 from June 4, 2022, to July 29, 2022. A scientist not involved in the intervention conducted the randomization using a code developed specifically for this study. Participants were automatically assigned to 1 group after providing informed consent and completing the preassessment. After reaching the sample size necessary to detect effects in effectiveness based on a priori power analysis, and due to difficulties in recruitment, the randomization ratio was adjusted to 3:1 in favor of the intervention group on July 29, 2022. This change allowed for the recruitment of a larger sample for within-condition analyses (H4 and H5) while maintaining sufficient power to detect effects in the between-condition analyses. Participants were not blinded to the intervention to which they were assigned; however, they were aware of whether they were assigned to the intervention of interest or the comparator.

Intervention Conditions

Intervention Condition: Conversational Agent Zenny

The intervention group consists of a newly developed automated conversational agent self-care program, which includes 40 separate modules based on empirically supported techniques. The modules cover the following core sections: cognitive behavioral skills, interpersonal relationships, positive psychological growth, relaxation, goal setting, and emotional regulation skills. The modules emphasize cognitive behavioral skills aimed at reducing emotional distress, enhancing behavioral activation and healthy behaviors, and improving problem-solving skills. The program's goal is to introduce psychological concepts that have been shown to effectively manage emotions and improve personal efficiency to a broader public audience. Additionally, it aims to help individuals connect with widely available resources related to well-being. Individuals interested in self-help interventions often struggle to find appropriate resources, and the interventions they do find (eg, apps) may lack scientific validity. Providing personalized recommendations for resources tailored to an individual's needs can help reduce this barrier to care. The intervention was delivered through a decision tree–based conversational agent on WhatsApp Messenger (Meta Platforms, Inc.), with participants contacted daily for 1 month. Participants received messages to encourage regular engagement with new modules or prompts to revisit previously completed ones. The conversational agent allowed participants to choose which topics to discuss and which branches of those topics to explore in greater depth. It also provided reminders about the content that participants themselves selected. Before the study, revisions were made based on detailed feedback from various team members and experts. After completing the registration, participants received an access code to connect with the conversational agent named *Zenny* at no cost.

Active Control Condition: Web-Based Wellness Resources

The control group received referrals to a menu of freely available evidence-based wellness resources linked from the

XSL•FO
RenderX

study website, including *Your Healthiest Self: Wellness Toolkits* [30], *Doing What Matters in Times of Stress* [31], and *How Right Now* [32]. Participants in this group could self-select relevant modules aimed at improving mood, reducing stress and anxiety, addressing relationship issues, and exploring other wellness topics. These programs required only access to the websites and could be completed in a self-help format at the participants' own pace. The resources were selected through a review of freely available online materials covering similar topics to those addressed by *Zenny*, with the aim of promoting well-being. They were chosen based on the criteria of being evidence-informed and sourced from credible organizations. Instead of offering a single resource option, three were selected to better reflect real-world scenarios, where individuals can choose from various online options.

Measures

Assessments were conducted online through LimeSurvey (LimeSurvey GmbH) before randomization (preassessment) and again after 1 month for the postassessment (mean 39.80 days, SD 15.63 days). Participants received 3 email reminders to encourage their participation in the postassessment. Additionally, implementation data were collected exclusively for the conversational agent condition to further understand the impact of participation and engagement on outcomes. Demographic characteristics, including age and gender, were assessed for all participants during the preassessment. Qualitative data were collected at the end of the postassessment to gather feedback on participants' study experiences. The following outcomes were included based on the RE-AIM framework.

Measures of Reach

Recruitment Strategies

During the preassessment, participants were asked how they learned about the study (eg, through a group, advertisement, friend referral, or website). This information was collected to identify which recruitment strategies yielded the highest number of participants.

Enrollment Rate

This is calculated by dividing the number of participants who completed the preassessment by the total number of individuals who accessed the survey page during the recruitment period. The response rate, as defined in the study protocol, is the number of eligible participants who agreed to participate divided by the total number of visitors to the study page. However, to respect privacy, we did not collect data on the number of visitors and cannot identify unique visitors to report this information.

Retention Rate

This is defined as the number of participants who completed the 1-month postassessment.

Primary Outcomes Measures of Effectiveness

Well-Being

The 5-item World Health Organization Well-Being Scale (WHO-5) [33] is a widely used self-reported measure of current subjective well-being. This measure is well-validated and commonly utilized in a web-based format. Participants are asked to rate their well-being over the past 2 weeks using a Likert scale from 0 to 5, with responses ranging from "0=At no time" to "5=All of the time" (eg, "My daily life has been filled with things that interest me"). Raw scores on the WHO-5 range from 0 to 25 and are then multiplied by 4, resulting in a final score range of 0 to 100, where higher scores indicate higher well-being. A systematic review has demonstrated that the measure possesses adequate validity for assessing well-being over time in nonclinical samples [34]. In this study, the internal consistency of the WHO-5 was high, with a Cronbach α of 0.89 and 0.91 at pre- and postassessment.

Psychosocial Flourishing

The Flourishing Scale (FS) developed by Diener et al [35] is an 8-item measure of positive human functioning. It assesses aspects of flourishing, including positive relationships, feelings of competence, a sense of meaning and purpose in life, and engagement in daily activities. Participants rate each statement on a 7-point scale, ranging from 1 (strongly disagree) to 7 (strongly agree). Total scores can range from 8 to 56, with higher scores indicating greater well-being. In this study, the internal consistency of the FS was excellent, with a Cronbach α of 0.91 and 0.93 at pre- and postassessment.

Positive Psychological Health

The Mental Health Continuum-Short Form (MHC-SF) by Keyes [36] assesses positive psychological health through 14 items organized into 3 subscales: Emotional Well-Being, Social Well-Being, and Psychological Well-Being. Previous research indicates that the Psychological Well-Being subscale is highly correlated with the overall well-being factor, often yielding similar scores to the total scale [37]. Therefore, this study utilized the 6-item Psychological Well-Being subscale (PWB), which includes items focused on self-acceptance, environmental mastery, positive relationships with others, personal growth, autonomy, and purpose in life. Respondents rate the frequency of each feeling experienced in the past month using a 6-point Likert scale, where 1 represents "never" and 6 indicates "every day." In our sample, the Psychological Well-Being subscale of the MHC-SF demonstrated good internal consistency, with a Cronbach α of 0.86 and 0.89 at pre- and postassessment.

Postassessment Measures of Implementation

User Satisfaction

This was assessed during the postassessment for both conditions using 4 items. Satisfaction with the program was evaluated across dimensions of satisfaction, usefulness, relevance, and helpfulness. Participants rated these aspects on a 4-point Likert scale, ranging from 1 (strongly disagree) to 4 (strongly agree). The item regarding usefulness was rated on a 5-point Likert scale. A sum score was calculated by aggregating all individual item scores, resulting in a range from 4 to 17, where higher scores indicate greater satisfaction with the content. The scale demonstrated excellent internal consistency within our sample (Cronbach α=0.92). The user satisfaction scale was modified to include only 4 items instead of 6 to reduce redundancy and shorten the survey length. This modified version was submitted for ethical review and included in the study protocol before its

XSL•FO
RenderX

initiation, although it deviated from the original version registered on the Open Science Framework [29].

Participation/Engagement

This was assessed based on the reported frequency of using the intervention or control condition resources at the postassessment (2 items, r=0.79). In the first item, participants rated how often they participated in the last month on a 6-point Likert scale, ranging from “1=not at all” to “6=daily”. In the second item, participants rated their perceived level of activity on a 4-point Likert scale, ranging from “1=not at all active” to “4=very active.” The scores for both items were combined into a sum score, with higher scores indicating greater participation and engagement.

Intervention Conversational Agent Measures of Implementation

For the intervention within-condition analyses, the implementation measures listed in Textbox 1 were assessed.

Textbox 1. Implementation measures for the within-condition analyses.

1. Module rating of user satisfaction
   - This was assessed within the intervention group following each completed module. Users rated their satisfaction on a scale from 1 (not at all) to 5 (extremely). An average satisfaction rating across modules over the 30-day period was computed. Additionally, daily module ratings were utilized for within-condition analyses (H5).
2. Days chatted with the bot
   - Engagement with the conversational agent was quantified by counting the number of days each user interacted with the bot over the past 30 days, ranging from 0 to 30 days.
3. Total messages
   - In the intervention group, the total number of messages sent by users to the conversational agent was measured as an objective indicator of engagement. This metric includes any instance where the user initiated interaction with the conversational agent or responded to its messages within the past 30 days, with counts ranging from 0 to 30 days.
4. Modules started
   - This was the total number of modules that were initiated by the participant within the first 30 days, whether they were completed or not (ie, the sum of finished and unfinished modules).
5. Modules completed
   - This is the number of modules that were completed within the first 30 days (ie, finished modules).
6. Modules started not completed
   - This is the number of modules that were started but not completed within the first 30 days (ie, unfinished modules).

Data Analytical Strategy

Data from all participants who were randomly assigned at preassessment were analyzed using an intention-to-treat approach. Before analysis, data underwent thorough quality assurance checks to ensure accuracy and reliability. As a result of the required response format during the preassessment, no missing data were identified for the outcome measures. A small amount of data was missing for age (22/1345, 1.64%) due to data entry errors (involving 3-digit ages) and for gender (9/1345, 0.67%) because identifying gender was not a mandatory response for participation. Before testing the key hypotheses, baseline differences between the intervention and control groups were assessed using chi-square tests and t tests (2-tailed and unpaired), as appropriate.

For analyses of effectiveness (H1), separate analyses were conducted to assess both within-group and between-group effects, evaluating the magnitude of change in the outcome variables at postassessment (WHO-5, PWB, and FS). Full information maximum likelihood estimation with robust statistics was applied using Mplus 8.6 [38]. Within-group effects were assessed using a Wald chi-square test to evaluate changes in outcome variables at postassessment for both the intervention and control groups. Between-group effects were estimated through an analysis of changes approach, with preassessment data adjusted [39]. Additionally, to investigate whether between-group effects varied by gender and age, interaction effects were included in the models for age × group and gender × group (H2), while controlling for participation/engagement and user satisfaction. For our third hypothesis (H3) concerning differences between conditions, we compared retention rates across groups using chi-square tests. Additionally, self-reported participation and engagement at postassessment were analyzed using linear regression models, consistent with the previously outlined analytical approaches.

For the fourth hypothesis (H4), we conducted within-group analyses of implementation outcomes specifically for the intervention group. Path analyses were performed to investigate whether greater changes in well-being over time, as measured by change scores in the WHO-5, PWB, and FS, could be predicted by engagement metrics. These metrics included module ratings, the number of days users interacted with the

XSL•FO
RenderX

bot, and the total number of messages sent to the conversational agent. Additionally, self-reported user satisfaction, participation, and engagement measured at postassessment were included in the analysis. The models also controlled for age, gender, and preassessment levels. To evaluate whether higher user satisfaction, as indicated by module ratings, predicted subsequent increases in participation and engagement over the first 30 days after enrollment, we developed a multilevel model. This model accounted for elapsed time since enrollment and was implemented using Mplus statistical software, applying 2-level maximum likelihood estimation with robust statistics. Associations between average user satisfaction, as indicated by module ratings, and average participation/engagement over the 30 days are also reported. A significance level of $P<.05$ was utilized to determine statistical significance across all analyses. Additionally, Cohen *d* was presented as a measure of effect size, providing a standardized metric for evaluating the magnitude of group differences in the study. Larger Cohen *d* values indicate greater effect sizes [40].

Results

Recruitment Strategies and Enrollment Rate

To assess reach, recruitment strategies are summarized in Table 1, with Facebook ads identified as the most frequently reported source of enrollment. The enrollment rate among those who accessed the study survey page was 65% (1349/2054; Figure 1). The retention rate at 1 month postassessment for participants randomized at preassessment was 42% (576/1345). Results are reported in compliance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [41] (see Multimedia Appendix 1).

Descriptive statistics for the sample at preassessment are presented in Table 1. Preassessment levels of all study variables did not significantly differ based on randomization to the intervention and control groups (Table 1). Additionally, there were no statistically significant differences between participants who completed both the pre- and postassessments and those who only completed the preassessment (Table 2).

Table 1. Characteristics of the intervention and control groups at preassessment (N=1345).

Characteristics	Intervention group (n=949)	Active control group (n=396)	*P* value[a]
Age (years), mean (SD)[b]	47.22 (16.46)	47.11 (16.68)	.91
Gender, n (%)			.85
Women	764 (80.5)	317 (80.1)	
Men	164 (17.3)	70 (17.7)	
Other	21 (2.2)	9 (2.3)	
Primary outcomes, mean (SD)			
Well-being[c]	42.82 (21.80)	43.96 (21.77)	.38
Psychosocial flourishing[d]	39.04 (9.82)	38.43 (10.10)	.30
Positive psychological health[e]	17.71 (6.76)	17.27 (7.02)	.28
Recruitment strategy, n (%)			.75
Facebook group	120 (12.6)	56 (14.1)	
Facebook ad	736 (77.6)	302 (76.3)	
Referral from a friend	11 (1.2)	5 (1.3)	
Not available	82 (8.6)	33 (8.3)	

[a]Assessed with independent sample *t* tests (ie, 2-tailed and unpaired) or χ^2 analysis.

[b]n=933 for the intervention group and n=390 for the active control group.

[c]Measured using the 5-item World Health Organization Well-Being Scale.

[d]Measured using the Flourishing Scale.

[e]Measured using the Mental Health Continuum-Short Form.

XSL•FO
RenderX

Figure 1. Study Sample Flow Chart.

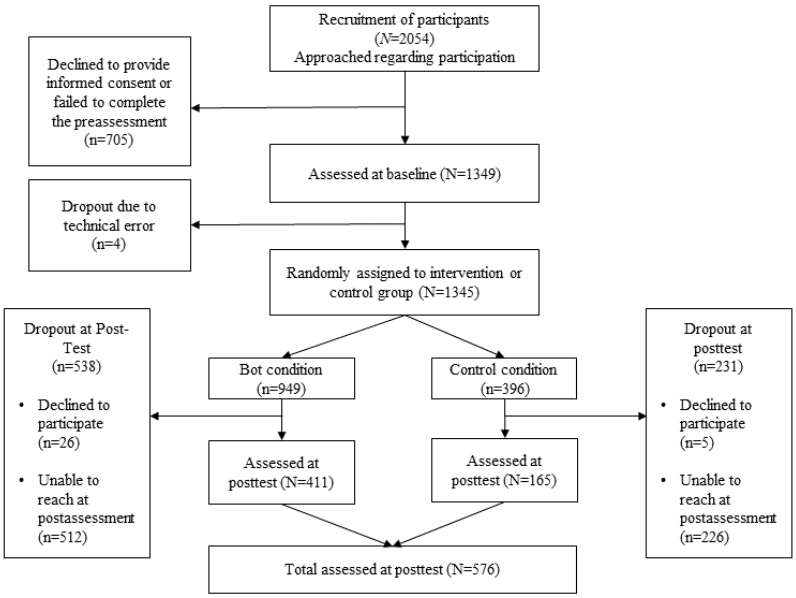

Table 2. Comparison of participants who completed postassessment versus those who did not complete postassessment (N=1345).[a]

Characteristics	Completed postassessment (n=576)	No postassessment (n=769)	*P* value[b]
Age (years), mean (SD)	47.01 (16.25)	47.32 (16.73)	.74
Gender, n (%)			.35
Female	467 (81.1)	614 (79.8)	
Male	100 (17.4)	134 (17.4)	
Other	9 (1.6)	21 (2.7)	
Primary outcomes, mean (SD)			
Well-being[c]	43.29 (22.78)	43.06 (21.03)	.85
Psychosocial flourishing[d]	39.02 (10.46)	38.73 (9.47)	.59
Positive psychological health[e]	17.93 (6.98)	17.32 (6.72)	.10
Recruitment source, n (%)			.49
Facebook group	77 (13.4)	99 (12.9)	
Facebook ad	443 (76.9)	595 (77.4)	
Referral from a friend	4 (0.7)	12 (1.6)	
Not available	52 (9.0)	63 (8.2)	

[a]For age n=569 for those who completed postassessment and n=754 for those with no postassessment.

[b]Assessed with independent sample *t* tests or χ^2 analysis.

[c]Measured using the 5-item World Health Organization Well-Being Scale.

[d]Measured using the Flourishing Scale.

[e]Measured using the Mental Health Continuum-Short Form.

Primary Outcomes

Results for the within-group changes for both the intervention and control groups are presented in Table 3. Significant improvements were observed in the WHO-5 (intervention: Cohen *d*=0.26, *P*<.001; control: Cohen *d*=0.24, *P*<.001), FS (intervention: Cohen *d*=0.19, *P*<.001; control: Cohen *d*=0.18, *P*<.001), and PWB (intervention: Cohen *d*=0.17, *P*=.001; control: Cohen *d*=0.24, *P*<.001) scales when measuring changes from pre- to postassessment. Between-group effects were estimated using an analysis of changes approach, controlling for preassessment scores in the models (H1). Although both groups showed significant improvements at postassessment, no significant between-group differences were found in the changes for WHO-5, PWB, and FS from pre- to postassessment (Table 4).

Next, we examined whether the effectiveness of the intervention varied by gender or age after controlling for

XSL•FO
RenderX

participation/engagement and user satisfaction (H2). The models included the main effects (gender or age and group) as well as the interaction effects between gender and group or age and group. We found no significant interaction effects of age or gender by group on the change scores of the primary outcomes (*P* values for interaction effects ranged from .46 to .94; full results are available upon request). This indicates that the results for the intervention and active control groups did not significantly differ based on age or gender.

In the next analysis (H3), we compared group differences in retention and participation/engagement. There was no significant difference in retention rates between the intervention and control groups, χ_1^2 (N=1345)=0.30, *P*=.57. Additionally, there was no statistically significant difference in self-reported participation/engagement at postassessment between the 2 groups (β=–.007, *P*=.92).

Table 3. Within-group effects for the intervention and active control groups with the intention-to-treat analysis (N=1345).[a]

Outcomes	Preassessment, mean (SD)	Postassessment, mean (SD)	Pre-post		
			Wald	*P* value	Cohen *d*
Intervention group (n=949)					
Well-being[b]	42.82 (21.80)	48.74 (23.26)	25.23	<.001	0.26
Psychosocial flourishing[c]	39.04 (9.82)	40.98 (10.69)	13.18	<.001	0.19
Positive psychological health[d]	17.72 (6.76)	18.91 (7.16)	10.57	.001	0.17
Active control group (n=396)					
Well-being[b]	43.97 (21.75)	49.39 (23.62)	22.19	<.001	0.24
Psychosocial flourishing[c]	38.43 (10.10)	40.20 (9.89)	19.42	<.001	0.18
Positive psychological health[d]	17.28 (7.01)	18.94 (6.69)	29.45	<.001	0.24

[a]Results indicate that in the intent-to-treat analyses, accounting for missing data at post, there is a significant improvement in well-being and flourishing for participants in the intervention group (Cohen *d*=0.17-0.26) and the active control group (Cohen *d*=0.18-0.24).

[b]Measured using the 5-item World Health Organization Well-Being Scale.

[c]Measured using the Flourishing Scale.

[d]Measured using the Mental Health Continuum-Short Form.

Table 4. Between-group effects comparing the intervention and active control groups with the intention-to-treat analysis (N=1345).[a]

Between-group effects	Mean post-pre change (SD), intervention group	Mean post-pre change (SD), active control group	β	SE	Estimated (SE)	*P* value
Well-being[b]	5.93 (16.00)	5.58 (17.64)	.004	0.041	0.078 (0.787)	.92
Psychosocial flourishing[c]	1.95 (6.94)	1.78 (6.23)	.019	0.038	0.147 (0.297)	.62
Positive psychological health[d]	1.15 (5.15)	1.53 (4.84)	–.022	0.038	–0.132 (0.227)	.56

[a]Based on the estimated means with the intent-to-treat analyses, accounting for missing data at post, there are no significant differences between the intervention groups in effects on primary outcomes. Group is coded –1 for control and 1 for intervention, positive β values indicate more change in the intervention group compared with the control group. There is a trend that the intervention group shows more improvements for all measures over time but not for flourishing between pre- and postassessment; however, all differences were not statistically significant.

[b]Measured using the 5-item World Health Organization Well-Being Scale.

[c]Measured using the Flourishing Scale.

[d]Measured using the Mental Health Continuum-Short Form.

Intervention Group Analyses

Descriptive statistics for the conversational agent interaction variables are reported in Table 5. Regarding H4, we examined whether user satisfaction, participation/engagement, and the intervention's conversational agent measures of engagement were associated with greater improvements in well-being as measured by the WHO-5, FS, and PWB scales. Regression analyses were conducted separately for each predictor of outcomes in the intervention group (Table 6). User satisfaction and participation/engagement at postassessment were significantly associated with greater improvements in all primary outcomes (*P*=.04 to <.001). Participants with more unfinished modules (modules started but not completed) showed less improvement in positive psychological health. Conversely, participants who engaged with the conversational agent on more days experienced significantly greater changes in well-being (*P*<.04). The results indicate that lower well-being at preassessment (WHO-5) was associated with more days of active messaging within the intervention group, which, in turn, was linked to larger improvements in well-being over time

XSL•FO
RenderX

(Figure 2). This indicates that the level of participation and engagement played a crucial role in driving change.

To examine the association between user satisfaction ratings and engagement in terms of messages with the conversational agent in the intervention group (H5), we conducted multilevel analyses that included elapsed time since enrollment, covering the first 30 days, in a 2-level model. Engagement was found to be negatively associated with time, indicating that there was greater engagement earlier in the 30-day period (n=435, β=–08, SE=0.03, *t*=–2.60, *P*=.009). Among the participants who provided user satisfaction ratings, the average number of ratings was 3.36 (SD 1.15) over the 30 days (n=211). The smaller sample size is attributed to some users not accessing the conversational agent. The module rating measure of user satisfaction significantly predicted the number of days messages were sent over the 30 days (n=211, β=.14, SE 0.03, *t*=4.31, *P*<.001). The association between average user satisfaction ratings and average participation/engagement across the 30 days approached statistical significance (*r*=0.13, *P*=.06).

Table 5. Descriptive statistics of conversational agent interaction data (implementation).[a]

Variables	Mean (SD)	Range
Days chatted with the bot[b]	4.26 (4.76)	1-29
Modules started[c]	3.68 (5.05)	0-32
Modules completed[d]	2.78 (4.69)	0-32
Modules started not completed[e]	0.90 (1.11)	0-10
Total messages[f]	51.09 (101.77)	0-740
Module rating of user satisfaction[g]	3.21 (1.00)	1-5

[a]Participants who engaged with the bot within the first 30 days: N=375, except for “days chatted with the bot” (n=339) and “module rating of user satisfaction” (n=211).

[b]Number of days the user messaged the conversational agent within the first 30 days.

[c]Number of modules started (finished and unfinished modules) within the first 30 days.

[d]Number of modules completed (finished modules) within the first 30 days.

[e]Number of modules started but not completed (unfinished modules) within the first 30 days.

[f]Count of messages sent by the user to the conversational agent within the first 30 days.

[g]Average rating of modules within the first 30 days.

XSL•FO
RenderX

Table 6. Implementation variables predicting well-being change outcomes from conversational agent and survey responses (n=949).

Variables	Well-being[a]		Psychosocial flourishing[b]		Positive psychological health[c]	
	β	*P* value	β	*P* value	β	*P* value
Postassessment measures						
User satisfaction[d]	.296	<.001[e]	.222	.002[f]	.206	.001[f]
Participation/engagement[g]	.236	<.001[e]	.153	.04[h]	.146	.04[h]
Intervention conversational agent measures						
Total messages[i]	.068	.15	–.006	.89	–.020	.66
Days chatted with the bot[j]	.109	.04[h]	–.014	.78	–.024	.62
Modules completed[k]	.064	.18	–.001	.98	–.019	.67
Modules started[l]	.066	.18	–.007	.88	–.038	.41
Modules started not completed[m]	.032	.56	–.033	.50	–.116	.07
Module rating of user satisfaction[n]	.152	.13	.185	.08	.134	.13

[a]Measured using the 5-item World Health Organization Well-Being Scale.
[b]Measured using the Flourishing Scale.
[c]Measured using the Mental Health Continuum-Short Form.
[d]User satisfaction scale with 4 items on satisfaction, usefulness, relevance, and helpfulness.
[e]*P*<.001.
[f]*P*<.01.
[g]Reported frequency spent using the program or control condition at the postassessment (2 items).
[h]*P*<.05.
[i]Count of messages sent by the user to the conversational agent within the first 30 days.
[j]Number of days the user messaged the conversational agent within the first 30 days (days active).
[k]Number of modules completed within the first 30 days.
[l]Number of modules started within the first 30 days.
[m]Number of modules started but not completed (unfinished modules) within the first 30 days.
[n]Average rating of modules within the first 30 days.

Figure 2. Days chatted with the conversational agent and associations with change in well-being (N=949). WHO-5 change indicates change in well-being. WHO-5: 5-item World Health Organization Well-Being Scale.

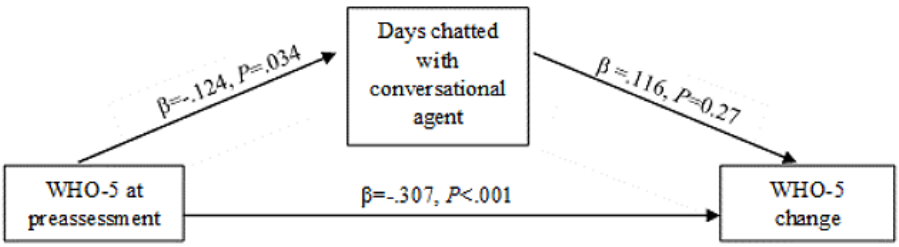

Discussion

Principal Findings

The primary aim of the study was to evaluate the effectiveness of a newly developed conversational agent–based self-care program (intervention group) compared with existing evidence-based wellness resources (active control group) on the primary outcomes of well-being (WHO-5), psychosocial flourishing (FS), and PWB scales. As anticipated, both groups showed significant improvements from pre- to postassessment in well-being (WHO-5 and PWB) and flourishing (FS). This indicates that with intention-to-treat analyses, which accounted for missing data at postassessment and adjusted for preassessment levels in the analytical model, significant improvements were observed in both well-being measures and psychosocial flourishing for participants in both the intervention and active control groups, with small effect sizes. However, contrary to our hypothesis, there were no significant differences between the groups when comparing changes from pre- to postassessment on the outcomes, as both groups demonstrated improvements. Taken together, participants in both groups—the

XSL•FO
RenderX

conversational agent self-care program (intervention) and the evidence-based wellness resources (active control)—showed improvements at postassessment. This suggests that both interventions may offer public health benefits for prevention, especially given their no-cost format. However, future studies should compare these interventions with other active intervention conditions and evaluate how they perform against no-intervention conditions.

This is in line with previous studies. A randomized controlled trial in Poland tested an agent-guided cognitive behavioral therapy (Fido) for reducing symptoms of anxiety and depression, comparing it with an active control group that used self-help books. The authors aimed to replicate findings from earlier studies, such as those involving Woebot. Their results showed that both groups (intervention and control) demonstrated reductions in anxiety and depressive symptoms postintervention [42]. By contrast, the Woebot study found a reduction in depressive symptoms only in the intervention group, not in the control group [21]. Similarly, a systematic review and meta-analysis found that conversational agent–based interventions yield promising results in reducing symptoms of depression and psychological distress, but no significant improvement in well-being was observed [20]. One possible explanation for the inconsistent findings regarding well-being, as suggested by Li and colleagues [20], is that measures of psychological distress tend to be more sensitive to change than measures of well-being. This implies that achieving significant improvements in well-being may require sustained, long-term engagement, or that different measures may be needed to detect small changes that could still hold public health significance.

No significant differences were found between groups regarding implementation outcomes, such as retention and engagement/participation, nor by gender or age. While the overall results did not indicate differences between groups, analyses within the intervention group revealed that user satisfaction and engagement with *Zenny* were related to changes in well-being outcomes. These findings underscore the importance of examining both implementation and effectiveness outcomes together and highlight the need to focus on enhancing user satisfaction and engagement in digital interventions to improve well-being.

Thus, while conversational agent–based interventions show promise for promoting mental health, most studies demonstrating improvements in mental health outcomes either lacked a control group or used inactive control groups. Furthermore, when comparing conversational agent–based interventions with an active control group, many studies either did not achieve significant between-group effects [17] or found significant results only for depressive symptoms and psychological distress, with no notable improvements in other outcomes, such as well-being [20]. Additionally, most interventions that included an active control group showed significant short-term effects but lacked long-term efficacy [18]. Research also indicates that agent-based interventions are more effective in clinical and subclinical settings than in nonclinical contexts [20]. This aligns with earlier findings that psychological interventions tend to be more effective for individuals with mental or physical health issues compared with those in community samples [43]. Therefore, further research is needed that includes active control groups and investigates the long-term efficacy of conversational agent–based interventions across various mental health outcomes, including well-being, using a universal prevention approach.

Strengths and Limitations

The use of WhatsApp posed a barrier to participation in the intervention group but not in the active control group, which may have affected the results by reducing exposure to the intervention. Consequently, implementation data (eg, module ratings, total messages) were available only for the intervention group. A potential solution would be to conduct the next study using WhatsApp for both groups to control for this difference or to implement the conversational agents on other platforms, such as a web browser.

Our study featured a newly developed conversational agent, which should be taken into account when interpreting the findings. While much of the qualitative feedback on the agent was positive—participants found it helpful, easy to access, and effective in identifying areas for personal change—some participants in the intervention group also provided negative feedback. This included issues such as reminders not working and occasionally slow responses. Additionally, recent research contradicts earlier findings [44] that advocated for the use of engagement reminders in digital interventions. A recent meta-analysis [18] revealed that not including automatic reminders in conversational agent–based interventions had a stronger positive effect on reducing depressive symptoms. This suggests that frequent reminders may hinder user interest rather than motivate them. According to this meta-analysis, future studies should prioritize personalization and empathetic responses, as these factors are strongly linked to the effectiveness of conversational agent–based interventions.

Furthermore, the conversational agent was still under development during data collection (eg, bug fixes), indicating that the actual effect size of the intervention should be interpreted with caution. While we selected resources that provided similar evidence-based self-help content for common risk factors affecting well-being in the active control condition, we did not systematically collect process data to compare the intervention content. This presents a potential direction for future research. Future studies should focus on strategies to enhance engagement in the intervention conditions and evaluate engagement with different types of content. Additionally, while participants were not aware of the content in the other condition, it was not possible to blind them during the intervention. Lastly, men and participants under the age of 35 years are significantly underrepresented in the sample. Furthermore, we did not assess other demographic variables, such as education, marital status, or employment status, to keep the participant survey brief. As a result, we cannot determine whether the findings are generalizable based on these characteristics.

Despite these limitations, the study has several strengths. It reports the initial results of a new conversational agent with a relatively large sample size, utilizing a randomized controlled trial design and including both men and women across a wide age range. The effect sizes observed for both the conversational

XSL•FO
RenderX

agent and the web-based resource condition, although small, may still hold public health relevance due to their use of freely available interventions.

Conclusion and Implications

This study compared 2 digital self-help programs aimed at promoting subjective well-being. Many self-help programs are available with limited evidence of their effectiveness, and this study helps address that gap by conducting a randomized trial of 2 such programs. Both the conversational agent and the evidence-based web resources demonstrated small improvements in well-being, but did not significantly differ in effectiveness. Additionally, the study highlighted the importance of engagement for achieving change and emphasized the need for future research focused on enhancing reach, engagement, and effectiveness in freely accessible self-help programs.

Acknowledgments

This research was funded by Meta Fundamental AI Research (FAIR) and research contracts awarded to HMF and through the University of Klagenfurt by Meta Fundamental AI Research (FAIR).

Authors' Contributions

The content of Zenny was developed by one of the authors (HMF), a licensed US psychologist, and other coauthors with psychology backgrounds (YB, JM, and CK). Implementation of the content and programming of the conversational agent were developed by coauthors (YB, SP, MU, EMS, ML, and AA).

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 635 KB - jmir_v26i1e53829_app1.pdf]

References

1. Huppert FA. Psychological well - being: evidence regarding its causes and consequences. Applied Psych Health & Well 2009 Jun 05;1(2):137-164. [doi: 10.1111/j.1758-0854.2009.01008.x]
2. Diener E, Biswas-Diener R. Happiness: Unlocking the Mysteries of Psychological Wealth. Hoboken, NJ: John Wiley & Sons; 2011.
3. Diener E, Ryan K. Subjective well-being: a general overview. South African Journal of Psychology 2009 Dec 01;39(4):391-406. [doi: 10.1177/008124630903900402]
4. Sin NL, Lyubomirsky S. Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. J Clin Psychol 2009 May;65(5):467-487. [doi: 10.1002/jclp.20593] [Medline: 19301241]
5. Barney LJ, Griffiths KM, Jorm AF, Christensen H. Stigma about depression and its impact on help-seeking intentions. Aust N Z J Psychiatry 2006 Jan;40(1):51-54. [doi: 10.1080/j.1440-1614.2006.01741.x] [Medline: 16403038]
6. Eisenberg D, Downs MF, Golberstein E, Zivin K. Stigma and help seeking for mental health among college students. Med Care Res Rev 2009 Oct 19;66(5):522-541. [doi: 10.1177/1077558709335173] [Medline: 19454625]
7. Berking M, Wupperman P. Emotion regulation and mental health. Current Opinion in Psychiatry 2012;25(2):128-134. [doi: 10.1097/yco.0b013e3283503669]
8. Malouff JM, Thorsteinsson EB, Schutte NS. The efficacy of problem solving therapy in reducing mental and physical health problems: a meta-analysis. Clin Psychol Rev 2007 Jan;27(1):46-57. [doi: 10.1016/j.cpr.2005.12.005] [Medline: 16480801]
9. Donker T, Bennett K, Bennett A, Mackinnon A, van Straten A, Cuijpers P, et al. Internet-delivered interpersonal psychotherapy versus internet-delivered cognitive behavioral therapy for adults with depressive symptoms: randomized controlled noninferiority trial. J Med Internet Res 2013 May 13;15(5):e82 [FREE Full text] [doi: 10.2196/jmir.2307] [Medline: 23669884]
10. Powell J, Hamborg T, Stallard N, Burls A, McSorley J, Bennett K, et al. Effectiveness of a web-based cognitive-behavioral tool to improve mental well-being in the general population: randomized controlled trial. J Med Internet Res 2012 Dec 31;15(1):e2 [FREE Full text] [doi: 10.2196/jmir.2240] [Medline: 23302475]
11. Twomey C, O'Reilly G, Byrne M, Bury M, White A, Kissane S, et al. A randomized controlled trial of the computerized CBT programme, MoodGYM, for public mental health service users waiting for interventions. Br J Clin Psychol 2014 Nov;53(4):433-450. [doi: 10.1111/bjc.12055] [Medline: 24831119]
12. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2013 Aug 09;44(6):1303-1317. [doi: 10.1017/s0033291713001943]

XSL•FO
RenderX

13. Kosyluk K, Baeder T, Greene KY, Tran JT, Bolton C, Loecher N, et al. Mental distress, label avoidance, and use of a mental health chatbot: results from a US survey. JMIR Form Res 2024 Apr 12;8:e45959 [FREE Full text] [doi: 10.2196/45959] [Medline: 38607665]
14. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
15. Barak A, Grohol JM. Current and future trends in internet-supported mental health interventions. Journal of Technology in Human Services 2011 Jul;29(3):155-196. [doi: 10.1080/15228835.2011.616939]
16. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
17. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
18. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862 [FREE Full text] [doi: 10.2196/43862] [Medline: 37115595]
19. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548 [FREE Full text] [doi: 10.2196/44548] [Medline: 37074762]
20. Li H, Zhang R, Lee Y, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
21. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
22. Ly KH, Ly A, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interventions 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
23. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. EMMA: an emotion-aware wellbeing chatbot. New York, NY: IEEE; 2019 Sep Presented at: The 8th International Conference on Affective Computing and Intelligent Interaction (ACII); September 3-6, 2019; Cambridge, UK p. 1-7. [doi: 10.1109/acii.2019.8925455]
24. Greer S, Ramo D, Chang Y, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
25. Williams R, Hopkins S, Frampton C, Holt-Quick C, Merry SN, Stasiak K. 21-day stress detox: open trial of a universal well-being chatbot for young adults. Social Sciences 2021 Oct 30;10(11):416. [doi: 10.3390/socsci10110416]
26. Vertsberger D, Naor N, Winsberg M. Adolescents' well-being while using a mobile artificial intelligence-powered acceptance commitment therapy tool: evidence from a longitudinal study. JMIR AI 2022 Nov 29;1(1):e38171 [FREE Full text] [doi: 10.2196/38171] [Medline: 38875600]
27. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
28. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
29. Foran H, Boureau YL, Kubb C, Müller J, Ung M, Poff S, et al. A randomized controlled trial of a bot-based self-help program. OSF Registries. URL: https://osf.io/ahe2r/resources [accessed 2024-11-14]
30. National Institutes of Health. Your healthiest self: wellness toolkits. National Institutes of Health. 2022. URL: https://www.nih.gov/health-information/your-healthiest-self-wellness-toolkits [accessed 2023-09-04]
31. Moltu C, Stefansen J, Svisdahl M, Veseth M. Negotiating the coresearcher mandate - service users' experiences of doing collaborative research on mental health. Disabil Rehabil 2012;34(19):1608-1616. [doi: 10.3109/09638288.2012.656792] [Medline: 22489612]
32. Centers for Disease Control and Prevention (CDC). How right now. CDC. 2022. URL: https://www.cdc.gov/howrightnow/index.html [accessed 2023-09-04]
33. World Health Organization (WHO). Wellbeing measures in primary health care/the DepCare Project: report on a WHO meeting. WHO. Geneva, Switzerland: WHO; 1998. URL: https://apps.who.int/iris/handle/10665/349766 [accessed 2023-09-04]
34. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]

XSL•FO
RenderX

35. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]
36. Keyes C. Brief description of the Mental Health Continuum Short Form (MHC-SF). PEP Lab. 2009. URL: https://peplab.web.unc.edu/wp-content/uploads/sites/18901/2018/11/MHC-SFoverview.pdf [accessed 2023-09-04]
37. Żemojtel-Piotrowska M, Piotrowski JP, Osin EN, Cieciuch J, Adams BG, Ardi R, et al. The mental health continuum-short form: the structure and application for cross-cultural studies-A 38 nation study. J Clin Psychol 2018 Jun 30;74(6):1034-1052. [doi: 10.1002/jclp.22570] [Medline: 29380877]
38. Muthén LK, Muthén BO. Mplus (Version 8.6). Mplus. 2021. URL: https://www.statmodel.com/ [accessed 2023-09-04]
39. Twisk J, Bosman L, Hoekstra T, Rijnhart J, Welten M, Heymans M. Different ways to estimate treatment effects in randomised controlled trials. Contemp Clin Trials Commun 2018 Jun;10:80-85 [FREE Full text] [doi: 10.1016/j.conctc.2018.03.008] [Medline: 29696162]
40. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed). Hillsdale, NJ: L. Erlbaum Associates; 1988.
41. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
42. Karkosz S, Szymański R, Sanna K, Michałowski J. Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. JMIR Form Res 2024 Mar 20;8:e47960 [FREE Full text] [doi: 10.2196/47960] [Medline: 38506892]
43. van Agteren J, Iasiello M, Lo L, Bartholomaeus J, Kopsaftis Z, Carey M, et al. A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. Nat Hum Behav 2021 May 19;5(5):631-652. [doi: 10.1038/s41562-021-01093-w] [Medline: 33875837]
44. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 01;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
FS: Flourishing Scale
MHC-SF: Mental Health Continuum-Short Form
PWB: Psychological Well-Being
WHO-5: 5-item World Health Organization Well-Being Scale

Edited by T McCall; submitted 20.10.23; peer-reviewed by D Dinh, K Kaphingst; comments to author 12.02.24; revised version received 30.04.24; accepted 15.08.24; published 06.12.24.

Please cite as:
Foran HM, Kubb C, Mueller J, Poff S, Ung M, Li M, Smith EM, Akinyemi A, Kambadur M, Waller F, Graf M, Boureau YL
An Automated Conversational Agent Self-Help Program: Randomized Controlled Trial
J Med Internet Res 2024;26:e53829
URL: https://www.jmir.org/2024/1/e53829
doi:10.2196/53829
PMID:39641985

©Heather M Foran, Christian Kubb, Janina Mueller, Spencer Poff, Megan Ung, Margaret Li, Eric Michael Smith, Akinniyi Akinyemi, Melanie Kambadur, Franziska Waller, Mario Graf, Y-Lan Boureau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of an Internet-Based Acceptance and Commitment Therapy Intervention for Reducing Psychological Distress in Health Care Professionals: Randomized Controlled Trial

Lin Zhang[1], MM; Shuang Huang[1], MPH; Sha Liu[1], BS; Yuanxiu Huang[1], MPH; Shan Chen[1], BM; Jinsong Hu[1], BPH; Mingzhong Xu[1], BM

Changsha Municipal Center for Disease Control and Prevention, Changsha, China

Corresponding Author:
Mingzhong Xu, BM
Changsha Municipal Center for Disease Control and Prevention
509 Wanjiali North Road
Changsha, 410004
China
Phone: 86 13873184830
Email: cscdcbm@126.com

Abstract

Background: Psychological distress is prevalent among health care professionals and can lead to poor-quality patient care. Internet-based acceptance and commitment therapy (iACT) is a promising intervention for improving mental health due to its low cost and easy access. However, there is limited evidence of its effectiveness in reducing health care professionals' psychological distress.

Objective: This study aims to examine the effects of iACT on psychological distress (stress, anxiety, and depression) among health care professionals in China.

Methods: From October 2022 to February 2023, a total of 108 health care professionals were recruited via WeChat and randomized into a 6-week iACT intervention program with therapist support (n=54) or waitlist control group (n=54). The intervention included 21 self-guided sessions combining teaching videos, mindfulness practices, and journal writing, followed by 7 live conferences to share experiences and discuss questions, all conducted via WeChat. Primary outcomes (stress, anxiety, and depression) and secondary outcomes (burnout and psychological flexibility) were collected using the Sojump platform, the most popular web-based survey platform in China. Generalized estimating equations were used to compare the outcomes between groups and assess the effects of group, time, and group-by-time interaction. Subgroup and sensitive analyses were performed to test the robustness of our findings across various groups.

Results: Among the 108 health care professionals, 68 (63%) completed the follow-up assessment at week 10, including 35 (64.8%) in the iACT group and 33 (61.1%) in the waitlist control group. Of the 54 participants in the iACT group, all attended at least 2 sessions, and 25 attended all 28 sessions. On average, participants attended 20 (71%) sessions. The iACT group showed significant improvement in the Depression Anxiety and Stress Scales-21 total score (d=0.82, 95% CI 0.39-1.26), and the effects were sustained for 4 weeks after the intervention (d=1.08, 95% CI 0.57-1.59). Compared to the control group, the iACT group showed significantly lower scores in burnout at week 6 (d=1.42, 95% CI 0.95-1.89) and week 10 (d=1.52, 95% CI 0.98-2.06). The iACT group showed significantly higher psychological flexibility at week 6 (d=1.23, 95% CI 0.77-1.69) and week 10 (d=1.15, 95% CI 0.64-1.66).

Conclusions: The iACT effectively decreased health care professionals' psychological distress and burnout and improved their psychological flexibility. Our findings provide implications and guidance for the development and broad implementation of iACT in health care settings to improve the mental health of health care professionals.

Trial Registration: Chinese Clinical Trial Register ChiCTR2400093584; https://tinyurl.com/38werwsk

(J Med Internet Res 2024;26:e59093) doi:10.2196/59093

XSL•FO
RenderX

KEYWORDS

acceptance and commitment therapy; internet-based intervention; stress; anxiety; depression; burnout; health care professionals; randomized controlled trial

Introduction

Worldwide, health care professionals experience significant psychological distress due to the demanding health care environment, heavy clinical workload, zero tolerance for mistakes, and conflicting roles, coupled with the shortage of resources in low-income countries [1]. Psychological distress refers to a state of emotional distress due to exposure to a stressful event that threatens one's physical or mental health and the inability to cope effectively with this stressor, which eventually leads to a range of emotional turmoil, such as stress, depression, and anxiety [2]. The COVID-19 pandemic has further aggravated the psychological distress of health care professionals due to the double burden of high infection risk and the demands of dealing with a public health emergency [3]. A literature review and meta-analysis reported a prevalence range of 24.1% to 67.55% for anxiety, 12.1% to 55.89% for depression, and 29.8% to 62.99% for stress among health care professionals during the COVID-19 pandemic [4]. These numbers were much higher than those reported before the pandemic. For instance, Lv et al [5] surveyed 8028 Chinese health care professionals from 8 provinces of China and found that the rates of anxiety, depression, and sleep disorders were obviously higher during COVID-19 compared with before the pandemic.

A major driver of health care professionals' psychological distress is burnout [6], a work-related stress syndrome resulting from continuous job pressures that individuals cannot effectively manage [7]. Burnout consists of 3 qualitative components: emotional exhaustion, depersonalization, and low personal accomplishment [7]. Health care professionals are at increased risk of burnout due to numerous stressful situations, such as witnessing the distress and death of patients; overloaded clinical work; time and role conflicts; and challenging interpersonal relationships with patients, colleagues, and leaders [8]. Approximately 1 in 3 physicians is experiencing burnout at any given time [9]. Burnout is estimated to have affected 10% to 70% of nurses and 30% to 50% of physicians and other health care professionals before COVID-19 [10]. A meta-analysis including 148 studies with 159,194 Asian health care professionals showed that the pooled prevalence for burnout was 68.3% during COVID-19 [11]. The positive link between burnout and psychological distress among health care professionals has been well established, with numerous studies showing that health care professionals with higher levels of burnout are also at increased risk of experiencing stress, depression, and anxiety [7,12,13].

Despite the high risk of burnout, not every health care professional experience severe psychological distress, which may be explained by the positive adaptation facing stress, such as psychological reflexibility. Psychological flexibility can be defined as "the ability to persist or to change behavior in a setting of competing psychological influences, guided by values and goals dependent on what the situation at hand affords" [14]. Psychological flexibility is an essential concept in the field of positive psychology and aligns well with the cognitive behavioral therapy framework [15]. A growing body of evidence suggests that psychological flexibility can mitigate the adverse impacts of burnout on psychological distress (ie, anxiety, depression, and stress) among health care professionals [16-18]. Therefore, psychological flexibility serves as a crucial psychological resilience factor against adverse mental health outcomes to improve health care professionals' general well-being.

Health care professionals' psychological distress and burnout may adversely affect the quality of care and patient safety, leading to increased medical errors and complaints, health care–associated infections, patient mortality, and decreased patient satisfaction [19]. On the system level, they may lead to low workforce productivity, poor work performance, high intention to leave, high turnover rate, and increased health care costs [20]. On a personal level, health care professionals with high levels of psychological distress and burnout are more likely to have risk behaviors such as substance abuse and even commit suicide [21]. Considering the high prevalence and severe outcomes of psychological distress and burnout among health care professionals, it is both urgent and critical to conduct effective interventions to address these problems among health care professionals.

Acceptance and commitment therapy (ACT) is a third-generation cognitive behavioral therapy aimed at fostering psychological flexibility through a range of mindfulness and values-based exercises while accepting difficult thoughts and feelings [22]. It involves 6 core therapeutic processes: acceptance, cognitive defusion, being present, self as context, values, and committed action [23]. ACT aims to enhance people's psychological flexibility to help them accept and live their lives in a meaningful way instead of changing their life conditions [22]. ACT encourages people to maintain a mindful and accepting attitude, free themselves from mental constraints, and evaluate themselves situationally. It also helps them develop psychological flexibility, uncover their values, and form goals and action plans to achieve fundamental changes [23].

Ever since the first study on ACT was published [24], hundreds of randomized controlled trials (RCTs) have been conducted to test its effectiveness among various populations across various countries [25]. There is strong evidence base supporting the effectiveness of ACT as a transdiagnostic treatment in the clinical and general population with multiple conditions. Gloster et al [26] reviewed 20 meta-analyses of ACT effectiveness, reporting 100 controlled effect sizes across 12,477 participants grouped by target conditions and comparison group. They found that ACT was effective in various conditions, including anxiety, depression, substance use, pain, and transdiagnostic groups [26]. In addition, ACT demonstrated advantages over inactive controls, treatment as usual, and most active intervention conditions [26]. Furthermore, ACT has also been shown to be

XSL•FO
RenderX

effective in improving the well-being of health care professionals. A recent meta-analysis of 10 RCTs revealed that ACT outperformed pooled control conditions in decreasing both general distress and work-related distress among health care professionals [27].

However, the application of ACT in health care professionals is limited due to the heavy workload and limited free time of health care professionals, making it challenging for them to attend face-to-face ACT interventions. These limitations could be addressed by internet-delivered ACT (iACT), which has gained increasing popularity due to its easy accessibility, greater convenience, and cost-effectiveness. Compared to face-to-face group interventions, iACT provides more convenience and reflexibility without the restriction of time, geography, or venue [28]. In addition, iACT offers greater privacy and can attract more health care professionals who have mental health problems but are reluctant to seek direct help due to their professional identities and concerns about privacy [29]. Furthermore, iACT is highly accessible due to the increasing number of internet users. Finally, iACT is less costly since it can be self-guided without relying heavily on psychotherapists as traditional face-to-face treatment [30]. Brown et al [31] conducted a systematic review and meta-analysis of the effectiveness of iACT. They found that iACT significantly decreased common mental disorders such as depression and anxiety and improved well-being [31]. In addition, they also observed high adherence rates, indicating the high acceptability of iACT in both patients and the public [31].

However, there is limited evidence of its effectiveness in improving health care professionals' mental health in China. Thus, this study aims to examine the effectiveness of an iACT intervention program in reducing psychological distress (stress, anxiety, and depression) and burnout and improving psychological reflexibility among health care professionals in China.

Methods

Study Design and Participants

A parallel RCT with 4 repeated measures was conducted among health care professionals in Hunan province between October 2022 and February 2023. The study protocol was registered with the Chinese Clinical Trial Registry (ChiCTR2400093584).

The inclusion criteria of participants included (1) aged 18 years and older; (2) registered health care professionals with formal licenses issued by the National Health Department and with at least 1 year of working experience; and (3) able to use the internet and WeChat (Tencent). The exclusion criteria were (1) participants who were off duty due to sick leave, maternal leave, or other reasons during the study period; (2) participants who had a prior diagnosis of severe mental illness; (3) participants with severe suicidal ideation as assessed by the last item score of the Depression Anxiety and Stress Scales-21 (DASS-21)≥3 or suicidal behaviors; and (4) participants who had received oral psychotropic drug or psychotherapy treatment within the past three months. All participants provided informed consent before enrollment.

Sample Size

The sample size was estimated using G*Power 3.1 (Heinrich-Heine-Universitat Dusseldorf, Dusseldorf, Germany). We selected "*F* test" and the test family "ANOVA: repeated measures, between factors" as the statistical test, and "DASS-21" as the key outcome variable when estimating the required sample size to achieve the minimal clinical difference. A previous meta-analysis showed that the average improvement effect of iACT intervention on depression and anxiety symptoms could be small, with an effect size of 0.24 to 0.38 and adherence rates between 53.33 and 97.33% [10]. The effect size was set at 0.24, α was set at .05, and the power was set at 0.80. Given that the surveys were to be repeated four times, assuming a repeated-measures correlation of 0.5 and an attrition rate of 22%, the required sample size was 107.

Procedure

Potential participants were recruited through advertisements on WeChat, and those interested could scan a QR code to complete a web-based screening questionnaire to determine their eligibility. WeChat is a multifunctional messaging app that offers messaging, voice calls, video calls, mobile payments, miniprograms, and a wide range of other features for daily use [32]. Those who passed the screening criteria were assessed via telephone to confirm inclusion and exclusion criteria. Detailed information on the study's purpose, procedure, benefits, and risks was provided to all participants. After providing electronic informed consent, the participants were randomly allocated 1:1 to the iACT group or the waitlist control (WLC) group. A simple randomization method was used to randomize the participants. Researchers who were blinded to the study generated random sequences on a website and assigned participants to the appropriate group. In addition, data collectors were unaware of the group assignments throughout the study period.

All assessments were conducted through questionnaires on the Sojump website [33]. Sojump, also named Wenjuanxing in Chinese, is one of the largest web-based survey platforms in China [34]. It provides multiple free functions related to surveys, such as questionnaire development and distribution, data analysis, and result reporting, similar to Qualtrics or SurveyMonkey in the United States [34]. Four questionnaires were administered at the following 4 time points: baseline (T0), 2 weeks into the program (T1: week 2), right after the intervention (T2: week 6), and 4 weeks after the intervention (T3: week 10). Participants were compensated with a gift (approximately US $6) if they completed all 4 assessments.

Ethical Considerations

The study protocol was approved by the ethics committees of Changsha Center for Disease Control and Prevention (2021003). Participants were fully informed of the study's objective, procedures, benefits, and potential risks. They were guaranteed privacy and anonymity when participating in the study. All information and data were used for research purposes and would not be shared with a third party. Participation in the study was totally voluntary, and participants could withdraw from the study at any time, which would not affect their work and life. All participants provided electronic informed consent before

XSL•FO
RenderX

participation. Participants were compensated with a gift (approximately US $6) if they completed all 4 assessments.

Intervention

The intervention was adapted from Hayes' [35] protocols, and the content was based on the 6 core processes of ACT: acceptance, cognitive defusion, being present, self as context, values, and committed action [36]. The components (ie, dosage, frequency, and duration) of the iACT were determined based on an extensive literature review and consultation with experts in iACT. Previous studies have shown that interventions with more sessions can produce more significant and long-lasting positive effects, presenting a dose-response relationship [37]. Therefore, we designed 28 sessions to maximize the intervention effects. Considering the busy clinical work of health care professionals, we limited each intervention session duration to 15-20 minutes to facilitate fragmented self-learning for health care professionals and minimize their participation burden. The intervention was tested among the general population in the early stages, and positive feedback was received.

The 6-week iACT intervention was offered through the Little Program on WeChat, and participants need to log in to the Little Program before using it. The flowchart of the specific intervention module and screenshots of the WeChat Mini Program are shown in Multimedia Appendix 1. The iACT program contained 28 sessions. The first 21 self-guided sessions were recorded by a psychologist and a meditation instructor, with each session unlocked every other day. Each session included a 10-15-minute recorded teaching video of a therapist, a 5-8-minute audio recording of guided mindfulness practice, and a journal writing task. The journals can be public or private between the "instructor" and the "student." Participants received a WeChat message on the morning of practice day about the assigned session and practice assignment, which needed to be completed within 48 hours. After watching the video, listening to the audio, and submitting the journal, the participants were considered to have completed 1 session. After that, 7 web-based live conferences were held on days 6, 12, 18, 24, 30, 36, and 42 by an ACT therapist to review lessons, answer questions, and discuss difficult experiences. The number of sessions completed by participants is recorded in the background. Details about the sessions are shown in Table 1.

XSL•FO
RenderX

Table 1. Session themes.

Day	Self-guided session themes			Live conferences
	iACT[a] teaching video	Mindfulness practices audio	Journal writing	
1	Understanding human suffering and the meaning of suffering	Understanding mindfulness	What might be holding you back from living the life you want?	
3	Search for the source of pain: our perspective on those difficulties	Serenity of the sea	What are your sensations and thoughts?”	
5	Escaping pain causes problems	Mountain meditation	Write down your coping strategies and their effects	
6	N/A[b]	N/A	N/A	✓
7	Brain mechanism of avoidance	What took you away	Practice self-caring behavior	
9	Acceptance: facing our painful emotions, thoughts, feelings, and impulses with an open attitude	Back to anchor	Make a list of your emotional events: “I (experienced)... I feel...”	
11	Cultivate a positive attitude	Naming Emotions	Positive mindset toolkit	
12	N/A	N/A	N/A	✓
13	Get out of your mind and observe your thoughts	Finding inherent patterns	What are you thinking about right now?	
15	Introduce cognitive fusion	Observe the thinking train	What are your emotions, sensations, and thoughts?	
17	Introduce cognitive defusion and defusion techniques	Leaves on running water	What were my emotions, sensations, and thoughts during this meditation?	
18	N/A	N/A	N/A	✓
19	The 3 senses of self	Metta Meditation	Retell your story	
21	Being the observing self	Deepening of Metta meditation: self-care	Describe your sensation, experience, and realization of the practice	
23	Introduce mindfulness	Shelter of mercy	Pay attention to 5 things	
24	N/A	N/A	N/A	✓
25	Experience mindfulness practice	Mindfulness body scan	Your sensation and experience	
27	Sitting meditation	Sitting meditation	Your experience	
29	Practice mindful eating	Practice eating a raisin mindfully	What do you feel?	
30	N/A	N/A	N/A	✓
31	Values as chosen life directions	Let go of your worries	Imagine you are at your 80th birthday party and your friend is giving a speech. Write down what you hear.	
33	What values are and are not	Compassion and value	10 areas of value	
35	Choose your values	Lake and pebbles—value discovery	Rank and test your value	
36	N/A	N/A	N/A	✓
37	Make an action plan	Accurately identify risks and resources	Target exercise table	
39	Build patterns of effective action	Fulfillment Exercises—empower action	Write down the new behavior patterns and action plans you want to create	
41	Applying values to daily life	Realize value through action	What thoughts come to your mind during a mindfulness practice?	
42	N/A	N/A	N/A	✓

[a]iACT: internet-based acceptance and commitment therapy.

[b]N/A: not applicable.

XSL•FO
RenderX

Interventionists

The interventionists included a psychologist, who was a psychotherapy supervisor with extensive experience in ACT; a meditation instructor, who was also an ACT therapist; and an ACT therapist. The psychologist and meditation instructor were responsible for recording the first 21 self-guided sessions since these sessions included highly standardized teaching modules that required high qualifications. In contrast, the ACT therapist was in charge of the last 7 web-based live conferences, which required less qualification but more flexibility in coordinating and responding to various needs. All interventionists received standard training in ACT and had rich experience in delivering iACT. They provided the intervention following a highly standardized treatment protocol and were supervised during the intervention.

WLC Group

Participants in the WLC group received no intervention but could access other forms of care at their freedom, the same way as the intervention group. They were instructed that they could seek formal and informal help whenever they encountered psychological problems during the study, as they normally would. Four weeks after the study completion, the WLC group could start the iACT of their choice.

Measures

Demographic Information

Demographic information such as age, sex, marital status, and education was collected at baseline.

Primary Outcome Measure

The DASS-21 [38] was used to assess health care professionals' psychological distress with 3 subscales: depression, anxiety, and stress. The DASS-21 contains 21 items (7 items in each subscale), and each item is rated on a 4-point Likert-type scale ranging from 0=did not apply to me at all to 3=applied to me very much or most of the time. The total score ranges from 0 to 63, with a higher score indicating higher psychological distress. The DASS-21 has shown good internal consistency and convergent and discriminant validity in Chinese samples [39,40]. In this study, the DASS-21 demonstrated good internal consistency with a Cronbach α of 0.91 for the total scale, 0.87 for the depression subscale, 0.78 for the anxiety subscale, and 0.80 for the stress subscale.

Secondary Outcome Measure

Burnout was measured by the Maslach Burnout Inventory-General Survey (MBI-GS) developed by Maslach et al [41] and adapted by Li and Shi [42]. The Chinese version of MBI-GS [43] contains 15 items under 3 dimensions: exhaustion (5 items), cynicism (4 items), and reduced personal accomplishment (6 items, reverse scored). Each item is rated on a 7-point scale ranging from 0 to 6 according to the frequency of symptoms. The weighted sum score = emotional exhaustion score×0.4+ cynicism score×0.3+ reduced personal accomplishment×0.3, with a cutoff of 1.5 distinguishing between those with (≥1.5) and those without burnout (<1.5). In our study, the MBI-GS demonstrated good internal consistency with a Cronbach α of 0.833 for the total scale.

Psychological flexibility was assessed using the Comprehensive Assessment of ACT Processes (CompACT) as a process measurement for ACT. The CompACT [44] consists of 23 items under 3 subscales: openness to experience (10 items), behavioral awareness (5 items), and valued action (8 items). Each item is rated on a 7-point Likert scale from 0 to 6. The total score ranges from 0 to 138, with a lower total score indicating greater psychological flexibility. The CompACT demonstrated good internal consistency in this study with a Cronbach α of 0.893.

Data Analysis

All data were analyzed via IBM SPSS statistics (version 20.0). Continuous variables were expressed as means and SDs, while categorical variables were presented as numbers and proportions. The independent samples 2-tailed *t* tests and chi-square tests were conducted to assess the homogeneity of the pretest data between the intervention and control groups. Generalized estimating equations (GEEs) were used to compare the outcomes (DASS-21, MBI-GS, and CompACT score changes) between the 2 groups and assess the group, time, and group-by-time interaction effects. Assessments conducted at baseline (T0), week 2 (T1), week 6 (T2), and week 10 (T3) were included as outcomes in the GEE analysis, with 0.2, 0.5, and 0.8 representing a small, medium, and large effect, respectively. The effect size was conducted through a web-based effect size calculator [45]. A subgroup analysis was performed using the independent samples 2-tailed *t* test to explore the sex differences in the intervention effects. In order to test whether participants with more severe psychological distress benefit more from the iACT program, we conducted a sensitive analysis on participants with a DASS-21 score≥20 (65th percentile) at baseline. The significance level was set at P=.05; all tests were 2-tailed.

Results

Participant Recruitment and Baseline Characteristics

A total of 173 health care professionals were recruited, among whom 108 (n=42, 38.9% physicians; n=58, 53.7%, nurses; and n=8, 7.4%, support staff) were included and randomized into the iACT group (n=54) and the WLC group (n=54). Among the 108 participants, 99 (91.7%) completed at least 1 assessment, and 68 (63%) completed the follow-up assessment at week 10, including 35 in the iACT group and 33 in the WLC group (Figure 1). Among the final 68 participants, there were 52 (76.5%) female participants and 16 (23.5%) male participants. There were no significant differences in the attrition rates between the intervention group and the control group (χ^2_1=0.159; P=.690). Of the 54 participants in the iACT group, all attended at least 2 sessions (including self-guided sessions and the live sessions), 25 attended all 28 sessions, and 44 attended more than half of the total sessions. On average, participants attended 20 (71%) of the total sessions.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant recruitment and retention in the intervention. iACT: internet-based acceptance and commitment therapy; ITT: intention-to-treat; WLC: waitlist control group.

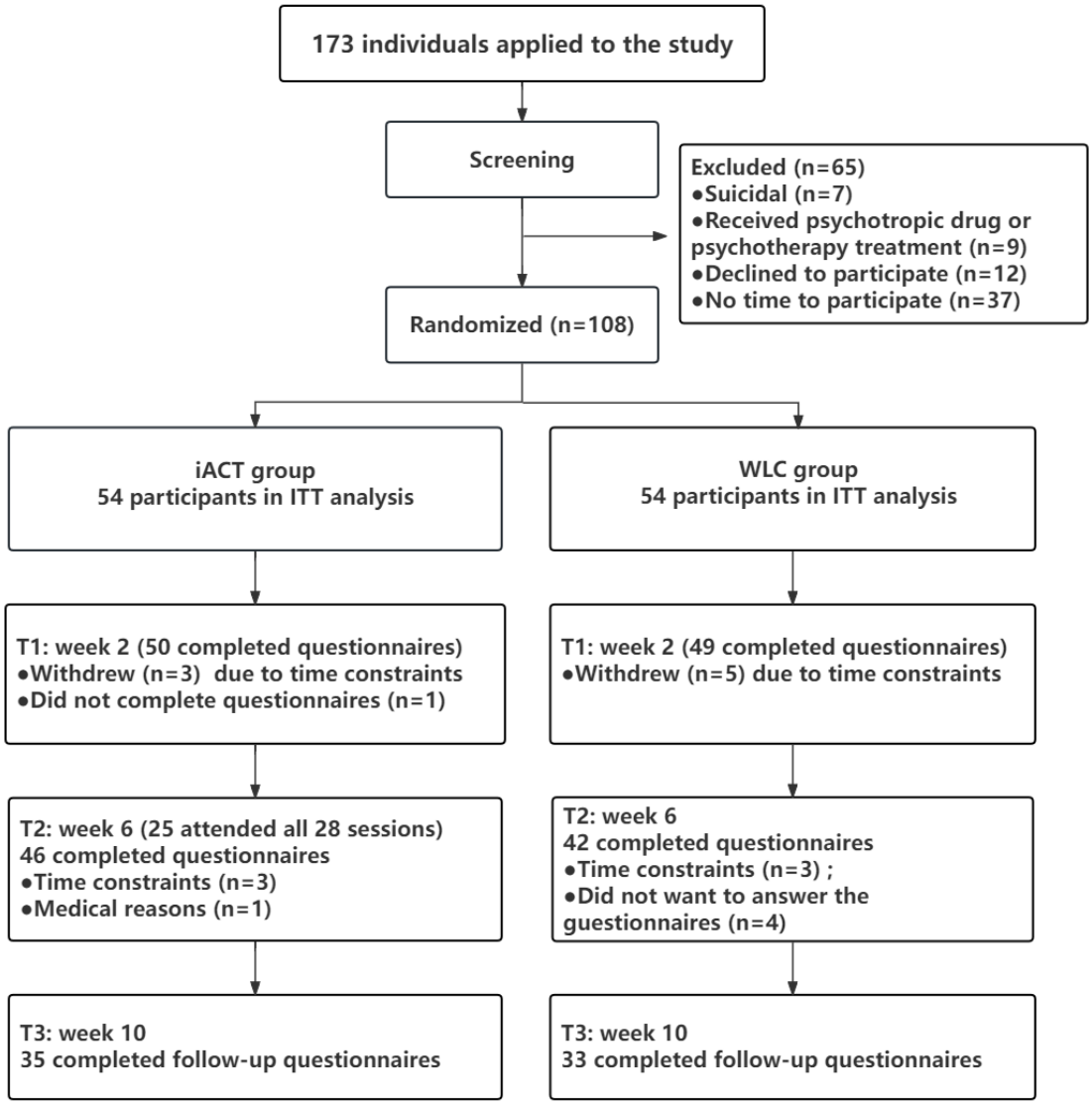

At baseline, the participants had a mean age of 38.45 (SD 7.26) years, and 81.5% (88/108) were female. Their mean scores for depression, anxiety, stress, burnout, and psychological flexibility were 5.56 (SD 3.53), 4.78 (SD 3.10), 7.43 (SD 3.29), 2.49 (SD 0.77), and 31.34 (SD 9.08), respectively (Table 1). There were no significant differences in baseline characteristics between the intervention group and the control group (Table 2) and between the completed sample and the missing sample (Multimedia Appendix 2).

XSL•FO
RenderX

Table 2. Baseline characteristics of participants.

Variables	Total sample (n=108)	iACT[a] group (n=54)	WLC[b] group (n=54)
Age (years), mean (SD)	38.45 (7.26)	38.35 (6.85)	38.56 (7.70)
Sex, n (%)			
Male	20 (18.5)	11 (20.4)	9 (16.7)
Female	88 (81.5)	43 (79.6)	45 (83.3)
Marital status, n (%)			
Single	13 (12.1)	8 (14.8)	5 (9.3)
Married	90 (83.3)	44 (81.5)	46 (85.2)
Others	5 (4.6)	2 (3.7)	3 (5.6)
Occupational role, n (%)			
Physician	37 (34.3)	21 (38.9)	19 (35.2)
Nurse	64 (59.3)	26 (48.1)	32 (59.3)
Other	7 (6.5%)	7 (13.0%)	3 (5.6%)
Education, n (%)			
Senior high school	3 (2.8%)	1 (1.9%)	2 (3.7%)
College degree or higher	105 (97.2%)	53 (98.1%)	52 (96.3%)
DASS-21[c], mean (SD)	17.76 (8.56)	17.24 (9.85)	18.28 (7.10)
Depression	5.56 (3.53)	5.30 (3.99)	5.81 (3.01)
Anxiety	4.78 (3.10)	4.74 (3.37)	4.81 (2.84)
Stress	7.43 (3.29)	7.20 (3.66)	7.65 (2.89)
MBI-GS[d], mean (SD)	2.49 (0.77)	2.43 (0.82)	2.56 (0.71)
CompACT[e], mean (SD)	31.34 (9.08)	32.02 (9.20)	30.67 (9.00)
Openness to experience	11.10 (5.51)	11.04 (5.63)	11.17 (5.43)
Behavioral awareness	9.41 (5.16)	9.89 (5.16)	8.93 (5.15)
Valued action	10.83 (5.15)	11.09 (5.53)	10.57 (4.77)

[a]iACT: internet-based acceptance and commitment therapy.

[b]WLC: waitlist control.

[c]DASS-21: Depression Anxiety and Stress Scales-21.

[d]MBI-GS: Maslach Burnout Inventory-General Survey.

[e]CompACT: Comprehensive Assessment of ACT Processes.

Primary Outcomes

As shown in Table 3, the GEE model revealed a statistically significant group effect (P<.001) and group-by-time interaction effect (P=.04) on the DASS-21 score. The iACT group showed more substantial changes in DASS-21 score at weeks 2, 6, and 10 (β=–3.85, –6.36, –7.50, respectively) than the WLC group. At weeks 2, 6, and 10, the iACT group consistently showed significantly lower scores in the subscales of depression (β=–1.24, –2.42, –2.87, respectively), anxiety (β=–1.25, –2.00, –2.12, respectively) and stress (β=–1.42, –1.98, –2.59, respectively) than the WLC group (all P<.05). The intergroup effect sizes for psychological distress at week 6 and week 10 were 0.82 (95% CI 0.39-1.26) and 1.08 (95% CI 0.57-1.59), respectively.

XSL•FO
RenderX

Table 3. Between-group comparisons of changes in outcome variables over time.

Outcome and time	β (95% CI)	*P* value	Cohen *d* (95% CI)	Group effect		Time effect		Group × time effect	
				Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value
DASS-21[a]				40.29 (1)	<.001[b]	0.298 (2)	.86	6.622 (2)	.04[b]
T0[c]	–1.04 (–4.31 to 2.24)	.53	Reference[d]						
T1[e]	–3.85 (–5.82 to –1.88)	<.001[b]	0.65 (0.24-1.05)						
T2[f]	–6.36 (–9.10 to –3.62)	<.001[b]	0.82 (0.39-1.26)						
T3[g]	–7.50 (–10.18 to –4.82)	<.001[b]	1.08 (0.57-1.59)						
Depression				35.66 (1)	<.001[b]	0.593 (2)	.74	11.43 (2)	.003[b]
T0	–0.52 (–1.87 to 0.83)	.22	Reference						
T1	–1.24 (–2.08 to –0.40)	.004[b]	0.45 (0.05-0.85)						
T2	–2.42 (–3.34 to –1.50)	<.001[b]	0.84 (0.40-1.28)						
T3	–2.87 (–3.81 to –1.92)	<.001[b]	1.03 (0.52-1.53)						
Anxiety				31.96 (1)	<.001[b]	0.67 (2)	.71	3.18 (2)	.20
T0	–0.07 (–1.26 to 1.12)	.60	Reference						
T1	–1.25 (–1.87 to –0.63)	<.001[b]	0.68 (0.28-1.09)						
T2	–2.00 (–2.89 to –1.10)	<.001[b]	0.77 (0.34-1.20)						
T3	–2.12 (–3.11 to –1.14)	<.001[b]	0.85 (0.35-1.35)						
Stress				24.75 (1)	<.001	5.16 (2)	.08	4.14 (2)	.13
T0	–0.44 (–1.70 to 0.81)	0.46	Reference						
T1	–1.42 (–2.26 to –0.58)	.001[b]	0.54 (0.14-0.94)						
T2	–1.98 (–3.24 to –0.72)	.002[b]	0.55 (0.12-0.98)						
T3	–2.59 (–3.65 to –1.53)	<.001[b]	0.93 (0.43-1.44)						
MBI-GS[h]				63.49 (1)	<.001[b]	26.29 (2)	<.001[b]	19.84 (2)	<.001[b]
T0	–0.13 (–0.42 to 0.16)	.38	Reference						
T1	–0.53 (–0.80 to –0.27)	<.001[b]	0.57 (0.17-0.97)						
T2	–0.92 (–1.12 to –0.71)	<.001[b]	1.42 (0.95-1.89)						
T3	–1.08 (–1.33 to –0.83)	<.001[b]	1.52 (0.98-2.06)						
CompACT[i]				40.95 (1)	<.001[b]	34.57 (2)	<.001[b]	15.8 (2)	<.001[b]

XSL•FO
RenderX

Outcome and time	β (95% CI)	*P* value	Cohen *d* (95% CI)	Group effect		Time effect		Group × time effect	
				Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value
T0	1.35 (–2.12 to 4.82)	0.44	Reference						
T1	–3.64 (–5.80 to –1.48)	.001[b]	0.69 (0.28-1.09)						
T2	–7.64 (–9.98 to –5.31)	<.001[b]	1.23 (0.77-1.69)						
T3	–8.18 (–10.88 to –5.49)	<.001[b]	1.15 (0.64-1.66)						
Openness to experience				6.94 (1)	.008[b]	21.86 (2)	<.001[b]	6.84 (2)	.03[b]
T0	–0.13 (–2.24 to 1.98)	.72	Reference						
T1	–0.81 (–2.17 to 0.56)	.248	0.21 (–0.18 to 0.61)						
T2	–2.16 (–3.55 to –0.77)	.002[b]	0.49 (0.06-0.91)						
T3	–2.20 (–3.75 to –0.65)	.005[b]	0.44 (–0.04 to 0.92)						
Behavioral awareness				5.67 (1)	0.02[b]	13.33 (2)	.001[b]	6.97 (2)	.03[b]
T0	0.96 (–1.00 to 2.93)	.30	Reference						
T1	–0.23 (–1.65 to 1.18)	.749	0.24 (–0.16 to 0.63)						
T2	–1.79 (–2.99 to –0.60)	.003[b]	0.70 (0.27-1.13)						
T3	–2.07 (–3.48 to –0.66)	.004[b]	0.69 (0.20-1.18)						
Valued action				30.39 (1)	<.001[b]	15.48 (2)	<.001[b]	8.66 (2)	.01[b]
T0	0.52 (–1.45 to 2.49)	.74	Reference						
T1	–2.54 (–3.81 to –1.27)	<.001[b]	0.66 (0.26-1.07)						
T2	–3.58 (–4.81 to –2.35)	<.001[b]	0.92 (0.48-1.36)						
T3	–3.81 (–5.15 to –2.47)	<.001[b]	0.88 (0.38-1.38)						

[a]DASS-21: Depression Anxiety and Stress Scales-21.
[b]The *P* value was statistically significant.
[c]T0: baseline.
[d]Effect size of the intervention was 0.
[e]T1: week 2.
[f]T2: week 6.
[g]T3: week 10.
[h]MBI-GS: Maslach Burnout Inventory-General Survey.
[i]CompACT: Comprehensive Assessment of ACT Processes.

Secondary Outcomes

Table 3 also showed statistically significant group effect, time effect, and group-by-time interaction effects on burnout and psychological flexibility (all *P*<.05). Compared to the WLC group, the iACT group had significantly greater changes in burnout score from T0 to T3 (β=–1.08), from T0 to T2 (β=–.92), and from T0 to T1 (β=–.57). The intergroup effect sizes for

XSL•FO
RenderX

burnout at week 6 and week 10 were 1.42 (95% CI 0.95-1.89) and 1.52 (95% CI 0.98-2.06), respectively. The ACT group showed more substantial changes in CompACT score at weeks 2, 6, and 10 (β=–3.64, –7.64, –8.18, respectively) than the WLC group, also reflected in its 3 subscales of openness to experience, behavioral awareness, and valued action. The intergroup effect sizes for psychological flexibility at week 6 and week 10 were 1.23 (95% CI 0.77-1.69) and 1.15 (95% CI 0.64-1.66), respectively.

Subgroup Analysis

Considering the large proportion of female individuals in the final sample, we further conducted a sex-based subgroup analysis to test if there were any significant sex differences in the intervention effects. Table 4 shows the results of the independent samples 2-tailed *t* tests to assess the primary and secondary outcomes. There were no statistical sex differences in all outcomes, including the DASS-21, MBI-GS, and CompACT scores (all *P*>.05).

Table 4. Subgroup analysis of the primary and secondary outcomes at week 10 (n=35).

	Male (n=9), mean (SD)	Female (n=26), mean (SD)	*t* test (*df*)	*P* value
DASS-21[a] (T3[b]-T0[c])	–2.33 (10.51)	–6.04 (7.90)	1.11 (33)	.27
MBI-GS[d] (T3-T0)	–1.19 (0.66)	–0.87 (0.83)	–1.05 (33)	.30
CompACT[e] (T3-T0)	–7.11 (4.26)	–9.19 (10.61)	0.83 (32)	.42

[c]DASS-21: Depression Anxiety and Stress Scales-21.

[b]T3: week 10.

[c]T0: baseline.

[d]MBI-GS: Maslach Burnout Inventory-General Survey.

[e]CompACT: Comprehensive Assessment of ACT Processes.

Sensitivity Analysis

To examine the effect of the interventions on participants who exhibited higher levels of distress, we conducted a sensitivity analysis on the participants with a DASS-21 score ≥20 (65th percentile) at baseline (21 in the iACT group and 20 in the WLC group). As shown in Table 5, the scores of the DASS-21 total scale and its 3 subscales increased in the iACT group but decreased in the WLC group. The iACT group showed more substantial changes in DASS-21 score at weeks 2, 6, and 10 (β=–7.91, –8.44, –11.11, respectively) than the WLC group. At follow-up, the ACT group had more participants who no longer met the DASS-21 criterion for significant distress than the WLC group (77% vs 64%). Regarding burnout, compared to the WLC group, the iACT group showed more remarkable changes in burnout scores from T0 to T3 (β=–0.74), from T0 to T2 (β=–0.74), and from T0 to T1 (β=–0.48). As for psychological flexibility, the ACT group showed more substantial changes in CompACT score at weeks 2, 6, and 10 (β=–6.41, –9.35, –8.30, respectively), also reflected in its 3 dimensions of openness to experience, behavioral awareness, and valued action.

XSL•FO
RenderX

Table 5. Changes in outcome variables over time for participants with DASS-21 total scores ≥20 at baseline (n=41).

Outcome and time	β (95% CI)	*P* value	Cohen *d* (95% CI)	Group effect		Time effect		Group × time effect	
				Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value
DASS-21[a]				39.72 (1)	<.001[b]	0.03 (2)	.99	1.52 (2)	.47
T0[c]	1.00 (–2.86 to 4.86)	.66	Reference[d]						
T1[e]	–7.91 (–10.79 to –5.02)	<.001[b]	1.64 (0.91-2.38)						
T2[f]	–8.44 (–13.68 to –3.19)	.002[b]	0.83 (0.09-1.57)						
T3[g]	–11.11 (–16.41 to –5.81)	<.001[b]	1.59 (0.55-2.64)						
Depression				32.63 (1)	<.001[b]	0.36 (2)	.84	2.16 (2)	.34
T0	0.36 (–1.67 to 2.39)	.72	Reference						
T1	–2.37 (–3.63 to –1.11)	<.001[b]	1.02 (0.34-1.70)						
T2	–2.84 (–4.28 to –1.40)	<.001[b]	0.98 (0.22-1.73)						
T3	–3.49 (–5.24 to –1.75)	<.001[b]	1.25 (0.25-2.24)						
Anxiety				41.66 (1)	<.001[b]	0.36 (2)	.83	1.17 (2)	.56
T0	0.12 (–1.67 to 1.90)	.47	Reference						
T1	–2.66 (–3.63 to –1.68)	<.001[b]	0.79 (0.005-1.53)						
T2	–2.92 (–4.72 to –1.12)	.001[b]	1.53 (0.50-2.57)						
T3	–3.86 (–5.73 to –1.98)	<.001[b]	1.20 (0.21-2.18)						
Stress				20.09 (1)	<.001[b]	0.74 (2)	.69	1.66 (2)	.44
T0	0.52 (–1.23 to 2.28)	.49	Reference						
T1	–2.79 (–4.24 to –1.35)	<.001[b]	1.10 (0.42-1.78)						
T2	–2.66 (–5.14 to –0.17)	.04	0.50 (–0.23 to 1.22)						
T3	–4.11 (–6.25 to –1.97)	<.001[b]	1.45 (0.43-2.47)						
MBI-GS[h]				15.68 (1)	<.001[b]	8.22 (2)	.02[b]	2.47 (2)	.29
T0	–0.12 (–0.59 to 0.34)	.59	Reference						
T1	–0.48 (–0.91 to –0.05)	.03[b]	0.44 (–0.21 to 1.08)						
T2	–0.74 (–1.07 to –0.42)	<.001[b]	1.18 (0.41-1.95)						
T3	–0.74 (–1.16 to –0.32)	.001[b]	1.12 (0.14-2.10)						
CompACT[i]				27.15 (1)	<.001[b]	7.40 (2)	.03[b]	5.04 (2)	.08

XSL•FO
RenderX

Outcome and time	β (95% CI)	*P* value	Cohen *d* (95% CI)	Group effect		Time effect		Group × time effect	
				Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value	Wald chi-square (*df*)	*P* value
T0	2.12 (–3.07 to 7.31)	.41	Reference						
T1	–6.41 (–9.52 to –3.29)	<.001[b]	1.25 (0.55-1.94)						
T2	–9.35 (–12.62 to –6.07)	<.001[b]	1.58 (0.77-2.39)						
T3	–8.30 (–12.54 to –4.06)	<.001[b]	1.29 (0.29-2.29)						
Openness to experience				9.85 (1)	.002[b]	6.67 (2)	.04[b]	9.20 (2)	.01[b]
T0	0.13 (–2.0 to 2.26)	.82	Reference						
T1	–2.56 (–4.60 to 0.52)	.01[b]	0.74 (0.08-1.39)						
T2	–4.23 (–6.26 to –2.19)	<.001[b]	0.99 (0.24-1.75)						
T3	–3.20 (–5.85 to –0.56)	.02[b]	0.70 (–0.24 to 1.64)						
Behavioral awareness				9.32 (1)	.002[b]	3.11 (2)	.21	0.53 (2)	.77
T0	–0.15 (–3.00 to 2.71)	.92	Reference						
T1	–1.73 (–3.85 to 0.38)	.11	0.44 (–0.20 to 1.09)						
T2	–2.27 (–3.88 to –0.65)	.006[b]	0.77 (0.03-1.51)						
T3	–2.61 (–4.42 to –0.79)	.005[b]	1.16 (0.17-2.14)						
Valued action				16.46 (1)	<.001[b]	1.64 (2)	.44	1.61 (2)	.45
T0	2.14 (–0.21 to 4.50)	.07	Reference						
T1	–1.57 (–2.78 to –0.36)	.01[b]	0.95 (0.28-1.62)						
T2	–2.31 (–3.45 to –1.17)	<.001[b]	1.21 (0.43-1.98)						
T3	–2.29 (–3.71 to –0.87)	.002[b]	0.82 (–0.13 to 1.76)						

[a]DASS-21: Depression Anxiety and Stress Scales-21.

[b]The *P* value was statistically significant.

[c]T0: baseline.

[d]Effect size of the intervention was 0.

[e]T1: week 2.

[f]T2: week 6.

[g]T3: week 10.

[h]MBI-GS: Maslach Burnout Inventory-General Survey.

[i]CompACT: Comprehensive Assessment of Acceptance and Commitment Therapy Processes.

XSL•FO
RenderX

Discussion

Principal Findings

Our results demonstrated the effectiveness of the iACT intervention in reducing health care professionals' burnout and psychological distress, including stress, anxiety, and depression, as well as improving their psychological flexibility. The effect was more substantial among health care professionals with higher baseline psychological distress. Following 6 weeks of iACT intervention, the iACT group had more significant decreases in psychological distress and burnout and a more significant increase in psychological flexibility than the WLC group. The effects persisted at follow-up assessments 1 month later. These results are consistent with a large amount of evidence demonstrating the efficacy of iACT in individuals with different psychological conditions [46-49]. A meta-analysis found that iACT demonstrated small yet significant and enduring effects on depression and anxiety compared to control groups [50].

The level of psychological distress among health care professionals in our study was comparable to that reported in previous studies in China using the same assessment tool [51]. Deng et al [51] conducted a meta-analysis including 34 studies to evaluate and compare the prevalence of psychological distress among health care professionals and the general public before and after COVID-19. Among health care professionals, the prevalence of depression and anxiety was 40% and 38% before COVID-19, 31% and 40% during COVID-19, and decreased to 22% and 22% after COVID-19, respectively [51]. Among the general public, the prevalence of depression and anxiety was 33% and 24% before COVID-19, 26% and 22% during COVID-19, and increased to 62% and 44% after COVID-19, respectively [51]. These findings suggested that the COVID-19 epidemic affected health care professionals more severely than the public. However, health care professionals tend to recover more quickly after the pandemic, possibly due to better professional knowledge, cognition, experience, and preparation in response to the pandemic [52].

The DASS-21 total score of the iACT group at baseline in our study was similar to that of undergraduate first-year students in the Levin RCT study [53]. However, the change in the DASS-21 total score from baseline to post-assessment (–6.36) in our study was more significant than in the Levin study (–4.59). One explanation may be related to our study's longer duration and more courses. The intervention program in our research spanned 6 weeks and consisted of 28 sessions, while the Levin program consisted of 2 web-based multimedia lessons and supplementary tailored emails. Most previous studies had a duration of intervention between 3 and 13 weeks. A meta-analysis including 10 RCT studies indicated that studies with more treatment sessions were associated with larger effect sizes for psychological distress [27].

In general, our study showed positive effects of iACT in alleviating depressive symptoms, which is consistent with the bulk of previous evidence. For instance, Lu et al [54] conducted an RCT and randomly assigned 145 nurses to either the 5-week fully automated intervention or the waiting group. The intervention group showed significant improvement in anxiety and depressive symptoms, which sustained even 3 months after the intervention [54]. Similar positive findings have also been reported by other studies, which consistently showed robust evidence to support the efficacy of iACT in benefiting individuals with depressive symptoms [36,55-58]. Thompson et al [36] conducted a systematic review and meta-analysis across 25 studies and found that iACT was effective in improving all outcomes at post-assessment and maintained at follow-up across diverse populations.

Our study showed that the iACT intervention significantly improved psychological flexibility, which was reflected in the more substantial decreases in the CompACT scale and its three subscales in the iACT group than in the WLC group. The low level of psychological flexibility is a significant risk factor for many psychological and behavioral problems [59]. Previous research found that psychological flexibility mediated or partially mediated the improvement of psychopathological symptoms such as anxiety and depression through ACT intervention [54,60,61]. For instance, Zhao et al [61] conducted an RCT and randomly assigned 124 Chinese university students to either the iACT intervention consisting of 6 weekly 30-minute units or the WLC group. They found that psychological flexibility significantly mediated the associations between iACT and both depressive symptoms and positive mental health [61]. These findings are in line with the framework of ACT, which focuses on using multiple therapeutic methods to improve psychological flexibility as a means to improve well-being. Psychological flexibility includes 3 overlapping subcomponents: openness (the willingness to have unwanted thoughts and feelings), awareness (the ability to notice one's experiences mindfully), and engagement (actions that enable progress on one's goals and values) [62]. All these subcomponents showed significant improvement after the iACT intervention, indicating their similar and equally important roles in mediating the intervention effects.

Noteworthily, we observed growing intervention effects with each subsequent week, which persisted even 4 weeks after the intervention, indicating the sustainability of the improvement effects. The improved effects may be explained by the dose-response relationship between the intervention and the effects and the time-accumulating effects [37]. As the intervention goes on, its health-promoting effects also accumulate over time, leading to persistent improvement that may last for a certain period of time after the intervention. However, the health benefits of the intervention may not last over an extended period after the intervention is terminated, raising concerns about sustainability [63]. The lack of sustainability is a common phenomenon in most intervention studies and a major contributor to the research-to-practice gap [64]. For instance, it has been estimated that it may take about 17 years to implement only 14% of evidence-based research outcomes in real-world settings [65]. Therefore, considering the temporal dynamics of the intervention effects, it is suggested that a long-term maintenance program after the iACT intervention may be necessary for the long-term well-being of health care professionals. Compared to the traditional offline intervention, iACT has the advantages of easy access, low cost,

XSL•FO
RenderX

and flexible schedule without time or geography restrictions [54], which paves the way for the development of more affordable and cost-effective long-term maintenance booster programs in the future.

The iACT intervention program in our study combined self-management with therapist guidance. Despite the web-based nature of the intervention, participants attended 71% of sessions of the iACT programs; this is consistent with other studies showing a compliance rate ranging from 53.33% to 97.33% [66-68]. The adherence rate is a crucial indicator reflecting how well an intervention aligns with individuals' needs and expectations and achieves its intended benefits, especially in the context of the psychological well-being of health care professionals [54]. The high adherence rate observed in our study indicates the high feasibility, accessibility, and acceptability of iACT among health care professionals, as well as the positive experiences of health care professionals related to iACT [54]. It seems that interventions with therapist guidance had greater effectiveness in reducing psychological distress and burnout and improving psychological flexibility than non-guided iACT [50]. Further studies are needed to determine the optimal therapist involvement that balances efficacy and costs.

Strengths and Limitations

Our study contributes to the literature by providing comprehensive and robust evidence on the effectiveness of iACT in decreasing stress, anxiety, depression, and burnout, as well as improving psychological flexibility among health care professionals in China. The study has several strengths, including a high adherence rate suggestive of high acceptance of the intervention by the participants, a long intervention duration with more treatment sessions to ensure a large effect size, and intensive follow-up evaluation with multiple time points encompassing pre-, mid-, and postintervention. Furthermore, the intervention is delivered through WeChat, the most commonly used social media in China. Health care professionals can easily access it without downloading any additional software, which can dramatically save time and energy for health care professionals who are already overwhelmed by heavy clinical workloads and limited free time. All these factors greatly enhance the effectiveness of the iACT intervention. In addition, we used multiple analysis strategies, including GEE, 2-tailed *t* tests, and chi-square tests, combined with subgroup analysis and sensitivity analysis to ensure the robustness of our results.

However, our findings should also be interpreted with caution due to several limitations. First, the sample was predominantly female, which may limit the generalization of the findings to males. However, subsequent sex-based subgroup analysis showed no significant differences. Future studies may consider recruiting a more sex-balanced sample to validate our results. Second, potential sampling bias may exist due to the self-selective nature of the web-based recruitment strategy. Health care professionals with higher levels of psychological distress who may benefit more from the intervention may be more likely to actively participate in the study, leading to more significant observed intervention effects. While large effect sizes were observed for decreases in general distress in participants and in those who exhibited higher levels of distress, only tentative conclusions can be made given the potential sampling bias. Third, all outcome variables were assessed subjectively based on self-report and are subject to bias such as recall bias and social desirability bias. Future studies may consider combining some objective measurement indicators to get a more accurate outcome assessment. Fourth, although there were no statistical differences in the baseline characteristics between participants who completed the intervention and those who dropped out, attrition may introduce bias and limit the generalizability of the findings. Future studies should explore multiple strategies to improve participant retention. Fifth, participants in this study were followed up for only 4 weeks after the intervention, which may limit the external validity of the results. More RCT studies with more extended follow-up periods are warranted to evaluate the long-term effectiveness of iACT in the future. Finally, we did not assess the interventionists' abilities and the intervention fidelity. However, the interventionists were highly qualified professionals with extensive experience in ACT, and they followed a highly standardized treatment protocol. Future studies should employ standard strategies, such as the recommendations from the Behavior Change Consortium established by The National Institutes of Health [69], to evaluate the intervention process and test the intervention fidelity.

Challenges and Implications

While our study provides promising evidence of the effectiveness of iACT in improving the well-being of health care professionals, challenges may exist in scaling up this program to the real-world practice of hospitals and health care professionals. It has been widely acknowledged that interventions that have been shown effective in research settings may not equally be applied and widely adopted in practice and community-based settings, also known as the research-to-practice gap [70]. Contextual barriers (such as lack of supporting policies and regulations), health system barriers (such as the hospital’s lack of administrative support and resources to build up and maintain the program), and individual barriers (such as health care professionals’ lack of time, interest, and motivation in engaging in the program) may hinder the widespread adoption of the program [71].

Therefore, to address the research-to-practice gap and improve the adoption of the iACT program among health care professionals, we need to overcome those barriers using a multilevel approach. For instance, organizational governance should be strengthened to promote awareness of the iACT in health care organizations, and financial and physical resources should be assigned to these organizations to facilitate the implementation of the iACT. For health care professionals, it is necessary to conduct specialized education and training to help them better understand the advantages of iACT and enable them to be better consumers and users of the program. In addition, cultivating scholar-practitioners, who are both producers and consumers of the intervention, and promoting practice-based research are also promising approaches to integrating research and practice to ensure the wide dissemination and adoption of the iACT among health care professionals [64].

XSL•FO
RenderX

Conclusions

In conclusion, this study adds empirical evidence of the efficacy of iACT training in decreasing health care professionals' psychological distress and burnout and improving their psychological flexibility. Our findings provide implications and guidance for the development and wide implementation of iACT in health care settings to improve the mental health of health care professionals.

Acknowledgments

This research was supported by the Science Foundation of Changsha City (kq2202039) and the Changsha City Health Medicine Research Project (KJ-B2023100).

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available as they contain participant information but are available from the corresponding author on reasonable request.

Authors' Contributions

LZ and SH contributed to the investigation and data curation. LZ contributed to the conceptualization, formal analysis, and writing of the original draft. SL, YH, and SC contributed to the methodology, writing, reviewing, and editing. JH and MX contributed to funding acquisition, resources, supervision, validation, and editing. All authors approved the final draft of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
the WeChat Mini Program.
[PNG File , 1032 KB - jmir_v26i1e59093_app1.png]

Multimedia Appendix 2
Comparison of baseline general information for completed and missed samples.
[DOCX File , 16 KB - jmir_v26i1e59093_app2.docx]

Multimedia Appendix 3
CONSORT-EHEALTH checklist for RCTs.
[PDF File (Adobe PDF File), 161 KB - jmir_v26i1e59093_app3.pdf]

References

1. Siau CS, Wee LH, Ibrahim N, Visvalingam U, Yeap LLL, Yeoh SH, et al. Predicting burnout and psychological distress risks of hospital healthcare workers. Malays J Public Health Med 2018;1:125-136 [FREE Full text]
2. Drapeau A, Marchand A, Beaulieu-Prévost D. Epidemiology of psychological distress. In: LAbate L, editor. Epidemiology of Psychological Distress, Mental Illnesses—Understanding, Prediction and Control. Rijeka, Croatia: InTech; 2012.
3. Pappa S, Ntella V, Giannakas T, Giannakoulis VG, Papoutsi E, Katsaounou P. Prevalence of depression, anxiety, and insomnia among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. Brain Behav Immun 2020;88:901-907 [FREE Full text] [doi: 10.1016/j.bbi.2020.05.026] [Medline: 32437915]
4. Li Y, Scherer N, Felix L, Kuper H. Prevalence of depression, anxiety and post-traumatic stress disorder in health care workers during the COVID-19 pandemic: a systematic review and meta-analysis. PLoS One 2021;16(3):e0246454 [FREE Full text] [doi: 10.1371/journal.pone.0246454] [Medline: 33690641]
5. Lv Y, Zhang Z, Zeng W, Li J, Wang X, Luo GH. Anxiety and depression survey of Chinese medical staff before and during COVID-19 defense. SSRN Journal. Preprint posted online on March 30, 2020 2020 [FREE Full text] [doi: 10.2139/ssrn.3551350]
6. Martins V, Serrão C, Teixeira A, Castro L, Duarte I. The mediating role of life satisfaction in the relationship between depression, anxiety, stress and burnout among Portuguese nurses during COVID-19 pandemic. BMC Nurs 2022;21(1):188 [FREE Full text] [doi: 10.1186/s12912-022-00958-3] [Medline: 35850892]
7. Edú-Valsania S, Laguía A, Moriano J. Burnout: a review of theory and measurement. Int J Environ Res Public Health 2022;19(3):1780 [FREE Full text] [doi: 10.3390/ijerph19031780] [Medline: 35162802]
8. Zou G, Shen X, Tian X, Liu C, Li G, Kong L, et al. Correlates of psychological distress, burnout, and resilience among Chinese female nurses. Ind Health 2016;54(5):389-395 [FREE Full text] [doi: 10.2486/indhealth.2015-0103] [Medline: 27021058]

XSL•FO
RenderX

9. de Hert S. Burnout in healthcare workers: prevalence, impact and preventative strategies. Local Reg Anesth 2020;13:171-183. [doi: 10.2147/lra.s240564] [Medline: 33149664]
10. Bridgeman PJ, Bridgeman MB, Barone J. Burnout syndrome among healthcare professionals. Am J Health Syst Pharm 2018;75(3):147-152. [doi: 10.2146/ajhp170460] [Medline: 29183877]
11. Ching SM, Ng KY, Lee KW, Yee A, Lim PY, Ranita H, et al. Psychological distress among healthcare providers during COVID-19 in Asia: systematic review and meta-analysis. PLoS One 2021;16(10):e0257983 [FREE Full text] [doi: 10.1371/journal.pone.0257983] [Medline: 34648526]
12. Shenoi AN, Kalyanaraman M, Pillai A, Raghava PS, Day S. Burnout and psychological distress among pediatric critical care physicians in the United States. Crit Care Med 2018;46(1):116-122. [doi: 10.1097/CCM.0000000000002751] [Medline: 29016364]
13. Xiao Y, Zhang H, Li Q, Xiao S, Dai T, Guo J, et al. Role stress and psychological distress among Chinese nurses during the COVID-19 pandemic: a moderated mediation model of social support and burnout. Front Psychiatry 2022;13:812929 [FREE Full text] [doi: 10.3389/fpsyt.2022.812929] [Medline: 35370815]
14. McCracken LM. Committed action: an application of the psychological flexibility model to activity patterns in chronic pain. J Pain 2013;14(8):828-835 [FREE Full text] [doi: 10.1016/j.jpain.2013.02.009] [Medline: 23651881]
15. Doorley JD, Goodman FR, Kelso KC, Kashdan TB. Psychological flexibility: what we know, what we do not know, and what we think we know. Social Personal Psych 2020;14(12):1-11. [doi: 10.1111/spc3.12566]
16. Chong YY, Frey E, Chien WT, Cheng HY, Gloster AT. The role of psychological flexibility in the relationships between burnout, job satisfaction, and mental health among nurses in combatting COVID-19: a two-region survey. J Nurs Scholarsh 2023;55(5):1068-1081. [doi: 10.1111/jnu.12874] [Medline: 36610054]
17. Holmberg J, Kemani MK, Holmström L, Öst LG, Wicksell RK. Psychological flexibility and Its relationship to distress and work engagement among intensive care medical staff. Front Psychol 2020;11:603986 [FREE Full text] [doi: 10.3389/fpsyg.2020.603986] [Medline: 33250832]
18. Ramaci T, Bellini D, Presti G, Santisi G. Psychological flexibility and mindfulness as predictors of individual outcomes in hospital health workers. Front Psychol 2019;10:1302 [FREE Full text] [doi: 10.3389/fpsyg.2019.01302] [Medline: 31249541]
19. Dyrbye LN, Shanafelt TD, Sinsky CA, Cipriano PF, Bhatt J, Ommaya A, et al. Burnout among health care professionals: a call to explore and address this underrecognized threat to safe, high-quality care. NAM Perspect 2017;7(7). [doi: 10.31478/201707b]
20. de Vries N, Boone A, Godderis L, Bouman J, Szemik S, Matranga D, et al. The race to retain healthcare workers: a systematic review on factors that impact retention of nurses and physicians in hospitals. Inquiry 2023;60:469580231159318 [FREE Full text] [doi: 10.1177/00469580231159318] [Medline: 36912131]
21. Sullivan V, Hughes V, Wilson DR. Nursing burnout and its impact on health. Nurs Clin North Am 2022;57(1):153-169. [doi: 10.1016/j.cnur.2021.11.011] [Medline: 35236605]
22. Hayes SC, Strosahl KD, Wilson KG. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change. New York, NY: Guilford Press; 2011.
23. Hayes SC. Acceptance and commitment therapy: towards a unified model of behavior change. World Psychiatry 2019;18(2):226-227 [FREE Full text] [doi: 10.1002/wps.20626] [Medline: 31059616]
24. Zettle RD, Hayes SC. Dysfunctional control by client verbal behavior: the context of reason-giving. Anal Verbal Behav 1986;4:30-38 [FREE Full text] [doi: 10.1007/BF03392813] [Medline: 22477508]
25. Hayes SC. State of the ACT evidence. Association for Contextual Behavioral Science. URL: http://contextualscience.org/state_of_the_act_evidence [accessed 2024-10-01]
26. Gloster AT, Walder N, Levin ME, Twohig MP, Karekla M. The empirical status of acceptance and commitment therapy: a review of meta-analyses. J Contextual Behav Sci 2020;18:181-192. [doi: 10.1016/j.jcbs.2020.09.009]
27. Prudenzi A, Graham CD, Clancy F, Hill D, O'Driscoll R, Day F, et al. Group-based acceptance and commitment therapy interventions for improving general distress and work-related distress in healthcare professionals: a systematic review and meta-analysis. J Affect Disord 2021;295:192-202 [FREE Full text] [doi: 10.1016/j.jad.2021.07.084] [Medline: 34479127]
28. Marcu G, Ondersma SJ, Spiller AN, Broderick BM, Kadri R, Buis LR. The perceived benefits of digital interventions for behavioral health: qualitative interview study. J Med Internet Res 2022;24(3):e34300 [FREE Full text] [doi: 10.2196/34300] [Medline: 35353045]
29. Peterson AL. Experiencing stigma as a nurse with mental illness. J Psychiatr Ment Health Nurs 2017 Jun;24(5):314-321. [doi: 10.1111/jpm.12306] [Medline: 28543870]
30. Renton T, Tang H, Ennis N, Cusimano MD, Bhalerao S, Schweizer TA, et al. Web-based intervention programs for depression: a scoping review and evaluation. J Med Internet Res 2014;16(9):e209 [FREE Full text] [doi: 10.2196/jmir.3147] [Medline: 25249003]
31. Brown M, Glendenning A, Hoon AE, John A. Effectiveness of web-delivered acceptance and commitment therapy in relation to mental health and well-being: a systematic review and meta-analysis. J Med Internet Res 2016;18(8):e221 [FREE Full text] [doi: 10.2196/jmir.6200] [Medline: 27558740]

XSL•FO
RenderX

32. Montag C, Becker B, Gan C. The multipurpose application WeChat: a review on recent research. Front Psychol 2018;9:2247 [FREE Full text] [doi: 10.3389/fpsyg.2018.02247] [Medline: 30618894]
33. Sojump. URL: http://www.sojump.com/ [accessed 2024-12-10]
34. Del Ponte A, Li L, Ang L, Lim N, Seow WJ. Evaluating SoJump.com as a tool for online behavioral research in China. J Behav Exp Finance 2024 Mar;41:100905. [doi: 10.1016/j.jbef.2024.100905]
35. Hayes SC, Smith S. Get Out of Your Mind and Into Your Life : The New Acceptance and Commitment Therapy. Oakland, CA: New Harbinger Publications; 2005.
36. Thompson EM, Destree L, Albertella L, Fontenelle LF. Internet-based acceptance and commitment therapy: a transdiagnostic systematic review and meta-analysis for mental health outcomes. Behav Ther 2021;52(2):492-507. [doi: 10.1016/j.beth.2020.07.002] [Medline: 33622516]
37. Levy HC, Worden BL, Davies CD, Stevens K, Katz BW, Mammo L, et al. The dose-response curve in cognitive-behavioral therapy for anxiety disorders. Cogn Behav Ther 2020 Nov;49(6):439-454. [doi: 10.1080/16506073.2020.1771413] [Medline: 32631134]
38. Clara IP, Cox BJ, Enns MW. Confirmatory factor analysis of the depression-anxiety-stress scales in depressed and anxious patients. J Psychopathol Behav Assess 2023;23:61-67. [doi: 10.1023/A:1011095624717]
39. Jiang LC, Yan YJ, Jin ZS, Hu ML, Wang L, Song Y, et al. Corrigendum: The Depression Anxiety Stress Scale-21 in Chinese hospital workers: reliability, latent structure, and measurement invariance across genders. Front Psychol 2020;11:741 [FREE Full text] [doi: 10.3389/fpsyg.2020.00741] [Medline: 32351436]
40. Cao CH, Liao XL, Jiang XY, Li XD, Chen IH, Lin CY. Psychometric evaluation of the Depression, Anxiety, and Stress Scale-21 (DASS-21) among Chinese primary and middle school teachers. BMC Psychol 2023;11(1):209 [FREE Full text] [doi: 10.1186/s40359-023-01242-y] [Medline: 37452365]
41. Maslach C, Jackson SE, Leiter MP, Schaufeli WB, Schwab RL. Maslach Burnout Inventory. Palo Alto, CA: Consulting Psychologists Press; 1986.
42. Li CP, Shi K. The influence of distributive justice and procedural justice on job burnout. Acta Psychologica Sinica 2003;35(05):677-684 [FREE Full text]
43. Pei P, Lin G, Li G, Zhu Y, Xi X. The association between doctors' presenteeism and job burnout: a cross-sectional survey study in China. BMC Health Serv Res 2020;20(1):715 [FREE Full text] [doi: 10.1186/s12913-020-05593-9] [Medline: 32746808]
44. Călinici MS, Călinici T. Comprehensive assessment of ACT processes CompACT. Romanian adaptation and short form validation. J Evid Based Psychother 2021;21(2):209-226. [doi: 10.24193/jebp.2021.2.19]
45. Practical meta analysis effect size calculator. Campbell Collaboration. URL: https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html [accessed 2024-11-26]
46. Carlbring P, Hägglund M, Luthström A, Dahlin M, Kadowaki A, Vernmark K, et al. Internet-based behavioral activation and acceptance-based treatment for depression: a randomized controlled trial. J Affect Disord 2013;148:331-337. [doi: 10.1016/j.jad.2012.12.020] [Medline: 23357657]
47. Dahlin M, Andersson G, Magnusson K, Johansson T, Sjögren J, Håkansson A, et al. Internet-delivered acceptance-based behaviour therapy for generalized anxiety disorder: a randomized controlled trial. Behav Res Ther 2016;77:86-95. [doi: 10.1016/j.brat.2015.12.007] [Medline: 26731173]
48. Ivanova E, Lindner P, Ly KH, Dahlin M, Vernmark K, Andersson G, et al. Guided and unguided acceptance and commitment therapy for social anxiety disorder and/or panic disorder provided via the internet and a smartphone application: a randomized controlled trial. J Anxiety Disord 2016;44:27-35. [doi: 10.1016/j.janxdis.2016.09.012] [Medline: 27721123]
49. Lappalainen P, Langrial S, Oinas-Kukkonen H, Muotka J, Lappalainen R. ACT for sleep—internet-delivered self-help ACT for sub-clinical and clinical insomnia: a randomized controlled trial. J Contextual Behav Sci 2019;12:119-127. [doi: 10.1016/j.jcbs.2019.04.001]
50. Hayes SC, Pistorello J, Levin ME. Acceptance and commitment therapy as a unified model of behavior change. Counsel Psychol 2012;40(7):976-1002. [doi: 10.1177/0011000012460836]
51. Deng Y, Chen Y, Zhang B. Different prevalence trend of depression and anxiety among healthcare workers and general public before and after the peak of COVID-19 occurred in China: a meta-analysis. Asian J Psychiatr 2021;56:102547 [FREE Full text] [doi: 10.1016/j.ajp.2021.102547] [Medline: 33465749]
52. Simione L, Gnagnarella C. Differences between health workers and general population in risk perception, behaviors, and psychological distress related to COVID-19 spread in Italy. Front Psychol 2020;11:2166 [FREE Full text] [doi: 10.3389/fpsyg.2020.02166] [Medline: 33013555]
53. Levin ME, Pistorello J, Seeley JR, Hayes SC. Feasibility of a prototype web-based acceptance and commitment therapy prevention program for college students. J Am Coll Health 2014;62(1):20-30 [FREE Full text] [doi: 10.1080/07448481.2013.843533] [Medline: 24313693]
54. Lu Y, Li Y, Huang Y, Zhang X, Wang J, Wu L, et al. Effects and mechanisms of a web- and mobile-based acceptance and commitment therapy intervention for anxiety and depression symptoms in nurses: fully decentralized randomized controlled trial. J Med Internet Res 2023;25:e51549 [FREE Full text] [doi: 10.2196/51549] [Medline: 38010787]

XSL•FO
RenderX

55. Gaudiano BA, Davis CH, Miller IW, Uebelacker LA. Development of a storytelling video self-help intervention based on acceptance and commitment therapy for major depression: open trial results. Behav Modif 2019;43(1):56-81 [FREE Full text] [doi: 10.1177/0145445517738932] [Medline: 29090593]
56. Lappalainen P, Granlund A, Siltanen S, Ahonen S, Vitikainen M, Tolvanen A, et al. ACT Internet-based vs face-to-face? A randomized controlled trial of two ways to deliver acceptance and commitment therapy for depressive symptoms: an 18-month follow-up. Behav Res Ther 2014;61:43-54. [doi: 10.1016/j.brat.2014.07.006] [Medline: 25127179]
57. Lappalainen P, Langrial S, Oinas-Kukkonen H, Tolvanen A, Lappalainen R. Web-based acceptance and commitment therapy for depressive symptoms with minimal support: a randomized controlled trial. Behav Modif 2015;39(6):805-834. [doi: 10.1177/0145445515598142] [Medline: 26253644]
58. Pots WT, Fledderus M, Meulenbeek PA, ten Klooster PM, Schreurs KM, Bohlmeijer ET. Acceptance and commitment therapy as a web-based intervention for depressive symptoms: randomised controlled trial. Br J Psychiatry 2016;208(1):69-77. [doi: 10.1192/bjp.bp.114.146068] [Medline: 26250745]
59. Hayes SC, Luoma JB, Bond FW, Masuda A, Lillis J. Acceptance and commitment therapy: model, processes and outcomes. Behav Res Ther 2006;44(1):1-25. [doi: 10.1016/j.brat.2005.06.006] [Medline: 16300724]
60. Gloster AT, Klotsche J, Chaker S, Hummel KV, Hoyer J. Assessing psychological flexibility: what does it add above and beyond existing constructs? Psychol Assess 2011;23(4):970-982. [doi: 10.1037/a0024135] [Medline: 21767025]
61. Zhao C, Ren Z, Jiang G, Zhang L. Mechanisms of change in an internet-based ACT study for depression in China. J Contextual Behav Sci 2022;24:51-59. [doi: 10.1016/j.jcbs.2022.03.004]
62. Hayes SC, Wilson KG, Gifford EV, Follette VM, Strosahl K. Experimental avoidance and behavioral disorders: a functional dimensional approach to diagnosis and treatment. J Consult Clin Psychol 1996;64(6):1152-1168. [doi: 10.1037//0022-006x.64.6.1152] [Medline: 8991302]
63. Hailemariam M, Bustos T, Montgomery B, Barajas R, Evans LB, Drahota A. Evidence-based intervention sustainability strategies: a systematic review. Implement Sci 2019;14(1):57 [FREE Full text] [doi: 10.1186/s13012-019-0910-6] [Medline: 31171004]
64. Smith LS, Wilkins N. Mind the gap: approaches to addressing the research-to-practice, practice-to-research chasm. J Public Health Manag Pract 2018:S6-S11 [FREE Full text] [doi: 10.1097/PHH.0000000000000667] [Medline: 29189499]
65. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
66. Levin ME, Hayes SC, Pistorello J, Seeley JR. Web-based self-help for preventing mental health problems in universities: comparing acceptance and commitment training to mental health education. J Clin Psychol 2016;72(3):207-225. [doi: 10.1002/jclp.22254] [Medline: 26784010]
67. Sagon AL, Danitz SB, Suvak MK, Orsillo SM. The mindful way through the semester: evaluating the feasibility of delivering an acceptance-based behavioral program online. J Contextual Behav Sci 2018;9:36-44. [doi: 10.1016/j.jcbs.2018.06.004]
68. French K, Golijani-Moghaddam N, Schröder T. What is the evidence for the efficacy of self-help acceptance and commitment therapy? A systematic review and meta-analysis. J Contextual Behav Sci 2017;6(4):360-374 [FREE Full text] [doi: 10.1016/j.jcbs.2017.08.002]
69. Bellg AJ, Borrelli B, Resnick B, Hecht J, Minicucci DS, Ory M, et al. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. Health Psychol 2004;23(5):443-451. [doi: 10.1037/0278-6133.23.5.443] [Medline: 15367063]
70. Rycroft‐Smith L. Knowledge brokering to bridge the research‐practice gap in education: where are we now? Rev Educ 2022;10(1):e3341. [doi: 10.1002/rev3.3341]
71. Correa VC, Lugo-Agudelo LH, Aguirre-Acevedo DC, Contreras JAP, Borrero AMP, Patiño-Lugo DF, et al. Individual, health system, and contextual barriers and facilitators for the implementation of clinical practice guidelines: a systematic metareview. Health Res Policy Syst 2020;18(1):74 [FREE Full text] [doi: 10.1186/s12961-020-00588-8] [Medline: 32600417]

Abbreviations

ACT: acceptance and commitment therapy
CompACT: Comprehensive Assessment of Acceptance and Commitment Therapy Processes
DASS-21: Depression Anxiety and Stress Scales-21
GEE: generalized estimating equation
iACT: internet-based acceptance and commitment therapy
MBI-GS: Maslach Burnout Inventory-General Survey
RCT: randomized controlled trial
WLC: waitlist control

XSL•FO
RenderX

Edited by A Schwartz; submitted 01.04.24; peer-reviewed by B Breznen, Y Sun, M Brown, H Li; comments to author 14.08.24; revised version received 05.10.24; accepted 11.11.24; published 18.12.24.

Please cite as:
Zhang L, Huang S, Liu S, Huang Y, Chen S, Hu J, Xu M
Effectiveness of an Internet-Based Acceptance and Commitment Therapy Intervention for Reducing Psychological Distress in Health Care Professionals: Randomized Controlled Trial
J Med Internet Res 2024;26:e59093
URL: https://www.jmir.org/2024/1/e59093
doi:10.2196/59093
PMID:39693127

©Lin Zhang, Shuang Huang, Sha Liu, Yuanxiu Huang, Shan Chen, Jinsong Hu, Mingzhong Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Approach for Addressing Suicidal Ideation and Behaviors in Youth Mental Health Services: Observational Study

Min K Chong[1], BSc (Hons); Ian B Hickie[1], AM, MD; Antonia Ottavio[2], Grad Dip, MCSL; David Rogers[3], BSN; Gina Dimitropoulos[4], MSW, PhD, RSW, RMFT; Haley M LaMonica[1], BSc, MA, PhD, ABPP-CN; Luke J Borgnolo[1], BPsych (Hons); Sarah McKenna[1], BSc, BA, PhD; Elizabeth M Scott[1,5], MD; Frank Iorfino[1], MBMSc, PhD

[1]Brain and Mind Centre, The University of Sydney, Camperdown, Australia
[2]headspace Bondi Junction, Sydney, Australia
[3]headspace Port Macquarie Youth Services, Port Macquarie, Australia
[4]Faculty of Social Work, University of Calgary, Calgary, AB, Canada
[5]School of Medicine, University of Notre Dame, Sydney, Australia

Corresponding Author:
Min K Chong, BSc (Hons)
Brain and Mind Centre
The University of Sydney
94 Mallett Street
Camperdown, 2050
Australia
Phone: 61 (02) 9114 2199
Email: min.chong@sydney.edu.au

Abstract

Background: Long wait times for mental health treatments may cause delays in early detection and management of suicidal ideation and behaviors, which are crucial for effective mental health care and suicide prevention. The use of digital technology is a potential solution for prompt identification of youth with high suicidality.

Objective: The primary aim of this study was to evaluate the use of a digital suicidality notification system designed to detect and respond to suicidal needs in youth mental health services. Second, the study aimed to characterize young people at different levels of suicidal ideation and behaviors.

Methods: Young people aged between 16 and 25 years completed multidimensional assessments using a digital platform, collecting demographic, clinical, social, functional, and suicidality information. When the suicidality score exceeded a predetermined threshold, established based on clinical expertise and service policies, a rule-based algorithm configured within the platform immediately generated an alert for treating clinicians. Subsequent clinical actions and response times were analyzed.

Results: A total of 2021 individuals participated, of whom 266 (11%) triggered one or more high suicidal ideation and behaviors notification. Of the 292 notifications generated, 76% (222/292) were resolved, with a median response time of 1.9 (range 0-50.8) days. Clinical actions initiated to address suicidality included creating safety plans (60%, 134/222), conducting safety checks (18%, 39/222), psychological therapy (8%, 17/222), transfer to another service (3%, 8/222), and scheduling of new appointments (2%, 4/222). Young people with high levels of suicidality were more likely to present with more severe and comorbid symptoms, including low engagement in work or education, heterogenous psychopathology, substance misuse, and recurrent illness.

Conclusions: The digital suicidality notification system facilitated prompt clinical actions by alerting clinicians to high levels of suicidal ideation and behaviors detected among youth. Further, the multidimensional assessment revealed complex and comorbid symptoms exhibited in youth with high suicidality. By expediting and personalizing care for those displaying elevated suicidality, the digital notification system can play a pivotal role in preventing rapid symptom progression and its detrimental impacts on young people's mental health.

(J Med Internet Res 2024;26:e60879) doi:10.2196/60879

XSL•FO
RenderX

KEYWORDS

mental health service; youth mental health; suicide management; clinical decision support; primary care; personalization; suicide; suicidal; youth; mental health; mental health care; suicide prevention; digital technology; online assessment; clinician; digital health; health informatics; clinical information

Introduction

The high prevalence of suicidal ideation and behaviors among youth is alarming [1]. While the biological and psychosocial factors associated with the development of suicidality are complex [2] and alone cannot predict the risk of suicide, it is strongly associated with long-term severe mental illnesses [3], poor social functioning, [1], and future suicidal behaviors [4,5]. Therefore, effective identification and management of suicidality in youth mental health services is crucial for detecting individuals with complex needs and preventing further progression of symptoms [6,7].

Conducting a thorough and efficient assessment of suicidality poses time and labor constraints on services with high demand [8]. Such barriers often lead to delays in screening, timely responses, and the delivery of appropriate interventions. Extended wait times can be especially unsafe for young people experiencing intense levels of suicidal ideation and behaviors [9,10].

Digital technology is a potential solution for operationalizing timely assessments in mental health services [7]. Further, a digital notification system that can alert severe or sudden changes in a client's mental health symptoms can facilitate prompt care coordination and allocation. The clinical use and feasibility of such systems have been generally well-accepted in other areas of medicine [11-13].

In 2017, a digital notification system that detects suicidal ideation and behaviors in youth mental health services was developed and evaluated [14]. This pilot study demonstrated its effectiveness for triaging young people with high levels of suicidality into appropriate care in large general practitioner practices and youth mental health services [14,15]. However, the longitudinal feasibility of this system has not yet been explored.

Therefore, this study aims to evaluate the use of a digital suicidality notification system in youth mental health services over a 5-year period in a naturalistic setting. The current study reports the rate and type of clinical responses initiated by suicidality notifications and identifies the clinical, social, and functional needs of young people at different levels of suicidality.

Methods

Recruitment

Young people aged between 16 and 25 years, who presented to primary (headspace [16]) and community mental health services between November 2018 and October 2023 were invited to participate. All young people who were within the age range and used a digital platform, Innowell (University of Sydney and PwC), were eligible to participate in this study.

Participant Procedures

All participants filled out self-reported, multidimensional assessments before their initial consultation as part of the intake process at participating services. After the mandatory initial assessment, reassessment was encouraged for symptom monitoring throughout care. The questionnaire was administered through the Innowell platform, a web-based platform that supports individuals' mental health by facilitating regular symptom monitoring, and client-clinician feedback [17].

The multidimensional assessments collected demographics, mental health (eg, depressive, anxiety, mania, psychosis, and eating disorder symptoms), suicidal thoughts and behaviors, social and occupational functioning (eg, education, employment, and social connectedness), and substance use information of participants. A detailed description of the assessments is provided in Multimedia Appendix 1.

Measures for Suicidal Thoughts and Behaviors

Suicidal ideation and behaviors were assessed using 2 self-reported measures. Suicidal ideation over the past month was measured using the Suicidal Ideation Attributes Scale (SIDAS, [18]). This scale is comprised of 5 items, assessing the frequency of suicidal thoughts, the individual's ability to control these thoughts and closeness of attempting a suicide, related distress, and the extent of their impact on performing daily tasks. Each item was rated using a 10-point Likert scale where a higher value indicated more severe suicidal ideation. The SIDAS has been validated in an online survey of community-based Australian adults [18]. In addition, the Columbia-Suicide Severity Rating Scale (C-SSRS, [19]) examined suicide intent and planning over the past month and the history of suicidal attempts in the lifetime. The internal validity and consistency of C-SSRS have been validated in a multisite study with adolescents and adults [19].

Digital Suicidality Notification System

The digital suicidality notification system is a configurable system embedded in the Innowell platform. When a young person completes the SIDAS and C-SSRS assessments, the system categorizes them into no, low, and high suicidality groups based on their results (Figure 1A). Then the system triggers appropriate actions based on their suicidality group. Thresholds for different levels of suicidality were developed by expert psychiatrists (IBH and EMS) and adjusted to existing suicidal risk management policies at services when necessary (development and thresholds of suicidality levels are provided in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 1. Workflow of the digital suicidality notification system.

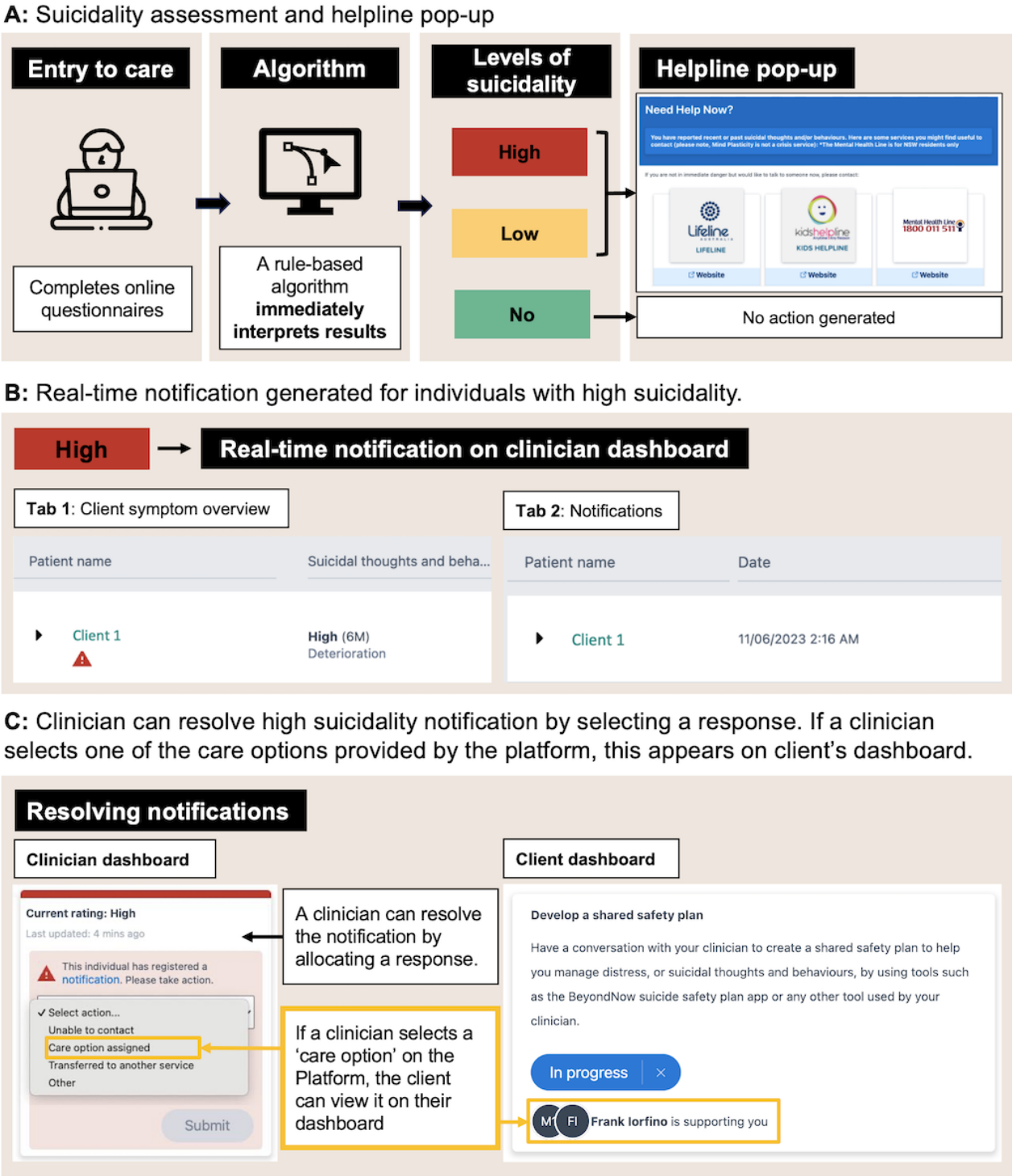

After categorization, participants in low or high suicidality groups immediately receive a pop-up message with contact information for 24-hour crisis support services, such as Lifeline, Kids Helpline, and Mental Health Helpline. For individuals in the high suicidality group, the system also sends an automatic notification to treating clinicians, prompting immediate clinical actions. This notification appears as a red triangle next to the person's name on the clinicians' Innowell dashboard (Figure 1B).

Clinicians can resolve the notification by logging clinical actions taken in response to the suicidality detected. There are 4 available options, which are "assigned a care option," "transferred to another service," "unable to contact," or "other." Under "assigned a care option," clinicians could further specify their clinical action by selecting one of the options provided on Innowell: "safety plan," "safety check," "psychological intervention," "scheduling a new appointment," or "other." Resolving the notification removes the red triangle (Figure 1C).

It was not mandatory for services to resolve these notifications. Instead, the research team recommended managing notifications according to the services' own safety and governance protocol.

XSL•FO
RenderX

This flexible adaptation approach was taken for smooth integration of the system into existing clinical workflows.

Further, participants were provided with clear expectations regarding how notifications would be responded to. During onboarding, they were informed about how the data collected would be used by services, and the international standard data privacy and security measures in place to ensure data safety. Staff communicated that these services did not provide 24-hour crisis support and that the platform would only be monitored during clinical hours (eg, Monday to Friday 8 AM to 5 PM). However, automated pop-up messages provided information on external crisis resources at all times.

Statistical Analysis

All statistical analyses were performed using R (version 4.3.1; R Core Team). In total, 3 pairwise comparisons were conducted between the suicidality groups: (1) no versus low suicidality, (2) low versus high suicidality, and (3) no versus high suicidality. Chi-square tests were used for categorical variables and Kruskal-Wallis tests were used for continuous variables. Post hoc pairwise comparisons were performed using Dunn tests, with a Bonferroni correction applied to address the 3 comparisons made for each variable and to counteract the family-wise error rate ($P<.02$). Analyses were performed with complete observations for each variable and missing values were regarded as missing at random.

In addition, an opinion statement was provided by coauthors, who are service staff from the participating services, to reflect on their experiences of using the digital notification system in youth mental health services.

Ethical Consideration

The Northern Sydney Local Health District Human Research Ethics Committees approved this study (HREC/17/HAWKE/480), and all participants gave online informed consent (through an opt-out process). In addition, to ensure confidentiality and privacy of participant information, all data stored in the research database were deidentified and could only be accessed by researchers in this study for research purposes.

Results

Sample Characteristics

A total of 2021 young people completed the intake multidimensional assessment, and 226 (11%) individuals triggered a high suicidality notification. The mean age of the cohort was 20.2 (SD 2.6) years and 72% (1463/2021) were female. Approximately a quarter of participants were receiving government benefits (26%, 523/2021), and two-thirds reported that they could not independently support themselves (67%, 1361/2021). One-tenth of individuals were not in employment, education, or training (NEET; 11%, 211/1902). On average, participants reported that they were out of work for 7.4 days (SD 7.9) out of the past 30 days. Further information on demographic, clinical, social, and functional characteristics of participants has been provided in Table 1.

XSL•FO
RenderX

Table 1. Baseline characteristics of participants and comparisons between different levels of suicidality.

	No suicidality	Low suicidality	High suicidality	No versus low, *P* value	Low versus high, *P* value	No versus high, *P* value	Missing values, n
Participants, n (%)	1002 (50)	793 (39)	226 (11)	__[a]	—	—	—
Mean age (years), (SD)	20 (2.5)	20.5 (2.6)	19.8 (2.6)	<.001	<.001	—	—
Females, n (%)	710 (71)	593 (75)	160 (71)	—	—	—	—
With disability, n (%)	50 (5)	86 (11)	25 (11)	<.001	—	.01	—
English speaking, n (%)	925 (92)	738 (93)	200 (89)	—	—	—	—
Indigenous, n (%)	51 (5)	34 (4)	11 (5)	—	—	—	—
Living alone, n (%)	62 (6)	74 (9)	23 (10)	—	—	—	—
Single, n (%)	613 (61)	497 (63)	152 (67)	—	—	—	—
Education, n (%)							8
Secondary	720 (72)	531 (67)	160 (71)	—	—	—	—
Tertiary	275 (27)	261 (33)	66 (29)	—	—	—	—
Social and occupational function							
NEET[b], n (%)	117 (12)	65 (8)	29 (13)	—	—	—	119
Days out of role in last 30 days, mean (SD)	6.7 (7.8)	7.3 (7.6)	10.5 (8.9)	—	<.001	<.001	—
Receiving government benefit, n (%)	261 (26)	202 (26)	60 (27)	—	—	—	—
Dependent support level, n (%)	663 (66)	536 (68)	162 (72)	—	—	—	—
Everyday functioning (WSAS)[c], mean (SD)	16.9 (8.5)	19.2 (8.3)	24.2 (8.0)	<.001	<.001	<.001	81
Social connectedness (SSSS)[d], mean (SD)	7.1 (3.4)	7.2 (3.3)	8.7 (3.8)	—	<.001	<.001	254
Personal mental health history							
Any family mental health history, n (%)	550 (55)	471 (59)	128 (57)	—	—	—	441
Mental illness history, n (%)	511 (51)	566 (71)	185 (82)	<.001	—	<.001	295
Mental health-related hospitalization history, n (%)	64 (6)	132 (17)	73 (32)	<.001	<.001	<.001	295
Sought prior treatment, n (%)	542 (54)	570 (72)	170 (75)	<.001	—	.001	293
Age when first sought help (y), mean (SD)	15.6 (3.6)	15.4 (3.7)	14.8 (3.9)	—	—	—	710
Experienced traumatic event, n (%)	350 (35)	333 (42)	107 (47)	—	—	—	388
Physical comorbidity, n (%)							
Major physical illness	255 (25)	232 (29)	64 (28)	—	—	—	388
Clinical presentation, mean (SD)							
Psychological distress (K-10)[e]	29.9 (8.7)	31.6 (8.4)	36.1 (6.8)	<.001	<.001	<.001	99
Depression (QIDS)[f]	12.6 (5.2)	14 (4.8)	17.7 (4.5)	<.001	<.001	<.001	304
Anxiety (OASIS)[g]	8.8 (4.2)	9.5 (4.1)	11.5 (4.5)	<.001	<.001	<.001	299
Psychosis (PQ-16)[h]	4.6 (3.8)	5 (3.5)	6.6 (4.1)	.003	<.001	<.001	381
Mania (ASRM)[i]	2.9 (2.9)	3 (2.9)	2.7 (2.7)	—	—	—	374

XSL•FO
RenderX

	No suicidality	Low suicidality	High suicidality	No versus low, *P* value	Low versus high, *P* value	No versus high, *P* value	Missing values, n
Eating disorder (EDE-Q)[j]	4.5 (3.0)	4.8 (3)	5.5 (3.0)	—	.001	<.001	349
At-risk mental states, n (%)							
Mania-like experience	169 (17)	178 (22)	47 (21)	<.001	—	<.001	374
Psychosis-like experience	269 (27)	270 (34)	100 (44)	<.001	—	<.001	—
Circadian disturbance	382 (38)	359 (45)	127 (56)	.003	.005	<.001	349
Suicidal thoughts and behaviors							
Self-harm history, n (%)	341 (34)	465 (59)	159 (70)	<.001	.002	<.001	346
Suicidal ideation (SIDAS)[k], mean (SD)	5.1 (9.5)	9.3 (9)	26 (12.4)	<.001	<.001	<.001	179
Prior suicide attempt, n (%)	192 (19)	369 (47)	179 (79)	<.001	<.001	<.001	216
Alcohol and substance misuse, mean (SD)							
Alcohol (AUDIT-C)[l]	4.4 (2.3)	4.6 (2.5)	5.2 (2.6)	—	—	.005	699
Cannabis (ASSIST)[m]	1.5 (2.8)	1.8 (3.1)	2.3 (3.4)	—	.007	<.001	354
Tobacco (ASSIST)	1.3 (2.1)	1.7 (2.3)	2.2 (2.4)	.003	<.001	<.001	352

[a]Not applicable.
[b]NEET: not in education, employment, or training.
[c]WSAS: Work and Social Adjustment Scale.
[d]SSSS: Schuster's Social Support Scale.
[e]K-10: Kessler-10.
[f]QIDS: Quick Inventory of Depressive Symptomatology.
[g]OASIS: Overall Anxiety Severity and Impairment Scale.
[h]PQ-16: Prodromal Questionnaire.
[i]ASRM: Altman Self-Rating Mania Scale.
[j]EDE-Q: Eating Behaviors and Body Image Questionnaire
[k]SIDAS: Suicidal Ideation Attributes Scale.
[l]AUDIT-C: Alcohol Use Disorders Identification Test.
[m]ASSIST: Alcohol, Smoking and Substance Involvement Screening.

Real-Time Helpline Pop-Up, Notification, and Response Characteristics

A total of 1828 helpline pop-up messages were generated, and 16% (292/1828) were accompanied by a real-time high suicidality notification. Out of the 292 notifications generated, 222 (76%) were resolved, and the median response time was 1.9 (range 0-50.8, mean 3.5, SD 5.4) days (Figure 2). Out of the 226 individuals who triggered a notification, 36 participants generated multiple notifications (mean 2.8, SD 1.4, range=2-6).

XSL•FO
RenderX

Figure 2. Distribution of time taken to resolve high suicidality notifications. The dashed line indicates the median response time (median 1.9 days, IQR 0-50.8 days). Outliers have been excluded in this graph for readability.

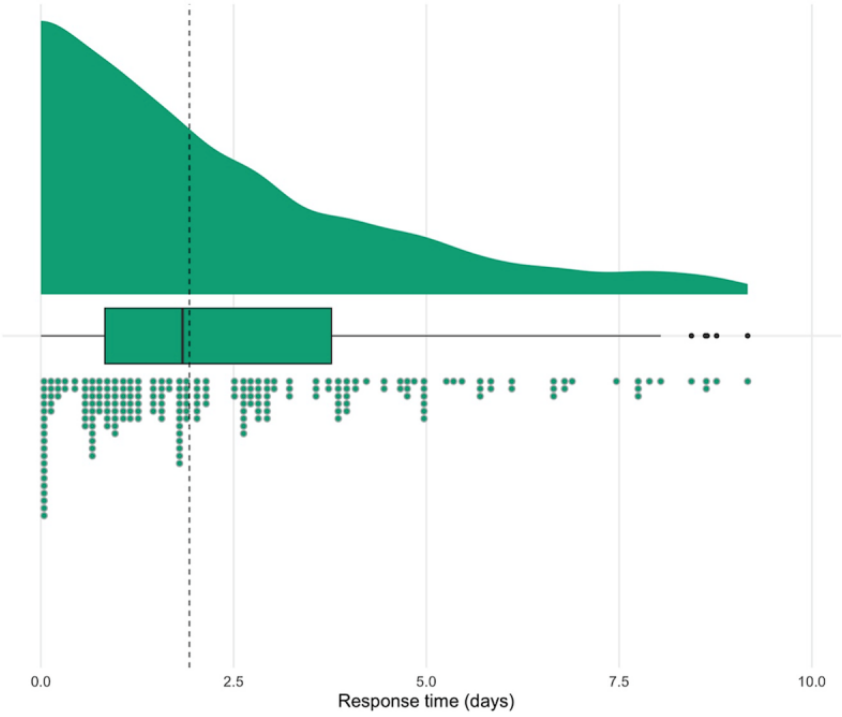

Most notifications were resolved by building a safety plan (134/222, 60%), followed by performing a safety check (39/222, 18%), initiating psychological therapies (17/222, 8%), transferring to another service (8/222, 4%; eg, general practitioner), and scheduling a new appointment with a mental health professional (4/222, 2%). Services were unable to contact 9% of individuals (20/222, Table 2).

Table 2. Clinical responses initiated to resolve high suicidality notifications.

Clinical response	Values, n (%)
Safety plan	134 (60)
Safety check	39 (18)
Unable to contact	20 (9)
Psychological therapy	17 (8)
Transferred to another service	8 (4)
New appointment offered	4 (2)

Differences in Clinical and Functional Characteristics Between Suicidality Groups

Pairwise comparisons were conducted between no, low, and high suicidality groups (Table 1). The low suicidality group was older than the high suicidality group (*z* score=3.5; *P*<.001) and no suicidality group (*z* score=4, *P*<.001). Further, disability was more prevalent in the high (χ^2_1=10.8; *P*=.001) and low (χ^2_1=20.8; *P*<.001) groups compared to the no suicidality group.

Several significant differences were observed between different suicidality groups in the social and occupational functioning domains. While there were no significant differences in the NEET status between the suicidality groups, the high suicidality group spent more days out of role compared with low (*z* score=5; *P*<.001) and no suicidality groups (*z* score=6.5, *P*<.001) within the last 30 days. Similarly, the high suicidality group experienced more social disconnection compared with the low (*z* score=5.0; *P*<.001) and no suicidality groups (*z* score=5.5; *P*<.001). Also, everyday functioning was most impaired in the high suicidality group (no vs high: *z* score=11.1, *P*<.001; low vs high: *z* score=7.3, *P*<.001; no vs low: *z* score=5.6, *P*<.001).

History of mental illness was least frequently reported in the no suicidality group (no vs low: χ^2_1=48.1, *P*<.001; no vs high: χ^2_1=38.4, *P*<.001), as well as mental health-related hospitalization history (no vs low: χ^2_1=37, *P*<.001; no vs high: χ^2_1=96.1, *P*<.001; low vs high: χ^2_1=21.5, *P*<.001), and previous help-seeking history (no vs low: χ^2_1=33.9, *P*<.001; no vs high: χ^2_1=10.7, *P*<.001) compared with 2 other groups. The severity of psychological distress, depressive, and anxiety symptoms increased in the order of no, low to high suicidality groups. However, only the high suicidality group had higher levels of

XSL•FO
RenderX

eating disorder behaviors (no vs high: z score=4.5, P<.001; low vs high: z score=3.2, P<.001).

The high suicidality group had the highest suicidal ideation compared with low (z score=13; P<.001) and no suicidality groups (z score=22.5; P<.001), more frequent history of self-harm (no vs high: χ^2_1=85.7, P<.001; low vs high: χ^2_1=9.5, P=.002) and higher rates of previous suicide attempts (no vs high: χ^2_1=251, P<.001; low vs high: χ^2_1=65.1, P<.001).

Reflections on the Digital Notification System in Youth Mental Health Services

Two service managers at participating centers highlighted the strengths of using the digital notification system to manage suicidality in youth mental health services (Textbox 1). The key strengths identified were service triage and wait time management, client engagement, facilitation of online resource use, and clinical decision-making support.

Textbox 1. Strengths of the digital notification system – insights from clinical practice.

Service triage and wait time management

- The system supports service triage by reducing wait times for young people with high suicidality, ensuring they receive care at the time of their need. The support of a digital system is especially helpful during periods of low staffing for maintaining the quality of care and responsiveness.

Client engagement

- Reviewing assessment results from the digital platform (Innowell) can start conversations for young people who might have not discussed their suicidality previously. However, achieving this requires a service-wide adoption of the digital technology. All service staff and clinicians need to understand and consistently use the platform throughout the client journey to maximize its potential as an engagement and communication tool.

Facilitation of online resource use

- Many young people are willing to engage in safety planning on the platform when guided by a clinician. This interaction not only helps to manage suicidality but also builds rapport and familiarizes young people with using the platform effectively.

Clinical decision-making support

- The care options provided on the digital platform offer guidance to clinicians, particularly those with limited experience to build confidence in addressing suicidality in young people.

Discussion

Principal Results

This study demonstrates that a digital suicidality notification system successfully initiated subsequent clinical responses for young people who expressed high suicidal ideation and behaviors. Of the high suicidality notifications, 76% (222/292) were resolved, and the median response time was 1.9 days (range 0-50.8 days). Building on its initial evaluation in 2017 [11], the high rate of resolved notifications over a 5-year period suggests that the system was accepted and used in real-world clinical settings.

Further, the study reveals the multidimensional and complex symptom presentations of young people with low or high suicidality levels. Compared with the no suicidality group, the low and high groups exhibited higher levels of clinical symptoms, more severe social and occupational impairments, and circadian disturbances. A history of mental health diagnosis and related hospitalization were more frequently reported in these groups.

Together, this study demonstrates the effectiveness of the digital notification system to detect suicidal needs, reveal associated symptom complexity, and trigger appropriate responses in youth mental health services.

Digital Systems for Rapid and Personalized Management of Suicidality

This evaluation demonstrated the potential of the digital notification system to facilitate stratification and personalization of care. The results show that most young people who elicited a notification received evidence-based treatment, including brief interventions (ie, safety checks and safety plan [20-23]), and long-term psychological interventions [24-26]. Alternatively, the system facilitated care stratification by expediting existing appointments and initiating care coordination, such as service transfers. Such cascades of clinical actions would have been delayed without this system. Research indicates that reducing wait times for individuals with high levels of suicidality could be critical due to its strong associations with severe mental illnesses [27], the risk of recurrence of suicidal behaviors [28], and comorbidity [29]. Further, there is a high likelihood that individuals present to health services before suicide [30]. Therefore, expedited clinical responses facilitated by the system may help reduce suicide rates through early detection and proactive care [5,31]. Addressing broader clinical and psychosocial needs associated with suicidal thoughts and behaviors could mitigate severe outcomes, such as suicide attempts [32]. and emergency department presentations [33].

While the tool led to appropriate clinical actions, a 2-day median response rate may raise concerns about its timeliness for supporting youth who are highly suicidal. However, it is important to recognize that these participating services did not provide 24-hour crisis support, and young people were informed

XSL•FO
RenderX

about the system's limited monitoring hours. A potential future use of this system could include automatic transfer of notification information to 24-hour crisis support services outside of service operating hours, to help young people receive immediate help.

Another research direction of this notification system is its potential to be reconfigured to detect different combinations of symptoms that indicate the need for rapid intervention, including psychosis-like symptoms and severe social impairments. In addition, the system could be enhanced by embedding machine learning tools to detect intraindividual changes in symptom trajectories [34,35], screen complex symptom presentations [36], or analyze linguistic patterns to detect suicidality based on user-created content [37,38]. Integrating these tools into the notification system could facilitate services to quickly understand young people's needs and take subsequent actions. This can be done by either providing direct care or coordinating with services that can deliver the required treatments.

Multidimensional Needs of Youth With High Suicidality

Notably, this study highlights the concurrent multidimensional impairments found in individuals with high levels of suicidality. Those who triggered a notification exhibited greater clinical, social, and functional impairments compared to groups with lower suicidality. These findings align with previous suicide research, where mental illnesses [39], psychosis-like experiences [40], irregular sleep-wake cycles [41], and eating disorders [42] have been associated with an increased likelihood of a suicide attempt. In particular, the association between high suicidality and severe social and occupational impairments, such as more days out of work and greater social disconnection, supports the causal relationship between these factors previously proposed [43-46].

These findings emphasize the importance of multidimensional assessments in addressing suicidality. Considering the recent critique of the marginally positive effect sizes of common suicide prevention interventions [25,47], deploying multidimensional measures could improve the identification of individuals' underlying mechanisms and support the formulation of personalized treatments. Digital platforms offer several advantages in this context, offering increased accessibility, privacy for clients to complete assessments, and reduced time and labor demand for services [48,49].

Implications in Youth Mental Health Services

A successful integration of the notification system into service workflow is crucial to maximize its benefits, as outlined in Textbox 1. Key strategies to facilitate implementation are the integration of the digital platform into clinical governance and the designation of key individuals to monitor notifications.

Integration of the digital system into clinical governance involves establishing organizational protocols that explicitly outline its use within the existing workflow. These protocols should be codesigned with clinicians and service staff and regularly reviewed to ensure acceptability and feasibility. Training sessions and modeling its use by senior clinicians can promote implementation by emphasizing its clinical utility [50,51]. Further, addressing common barriers such as data privacy concerns through transparent communication of data confidentiality would be critical for building trust in the system [52].

In addition, designating service staff to monitor and support digital platform use can facilitate the escalation of notifications into clinical actions. The concept of a digital navigator is one such example [53,54]. A digital navigator in services can alleviate the additional burden placed on clinicians by constantly monitoring the notification system and ensuring timely escalation of notifications. They can also be trained to deliver brief interventions, such as safety planning and safety checks, to help manage service demands [50,55].

Limitations

There are several limitations that should be taken into consideration. First, a higher proportion of participants in this study reported a history of self-harm or suicide attempts compared with previous studies [56,57]. This characteristic might influence the types of notifications generated and the subsequent responses. However, the primary objective of this study was to demonstrate the feasibility and acceptability of a digital notification system within a mental health service setting. Therefore, the system's high response rate, despite the frequent generation of alerts, strongly suggests its potential utility in managing cases with high clinical needs. Further research is required to assess its generalizability to populations with different characteristics and in various health care settings. Second, some may suggest that the threshold for a high suicidality notification is too sensitive, capturing individuals who may not have acute needs. However, the primary aim of the notification system is not to identify individuals with immediate suicidal risk. Rather, the notification is designed to inform clinicians about an individual's needs that require response within a reasonable time frame, due to the long-term implications for mental health. This notification system is a tool to facilitate personalized care. A part of enhancing tailored care could involve the use of an advanced machine learning tool that provides a more comprehensive understanding of the volatility of one's suicidal ideation and behaviors using their longitudinal data [34,58]. Third, the self-reported measures used can be subject to bias or underreporting. While the self-reported data should be supplemented with clinician reports, the measures used have been well-validated for detecting suicidality. Further, when data is collected longitudinally, it becomes a powerful tool for detecting intraindividual changes [59]. Finally, this study does not assess the impact of this notification system on future illness trajectories of youth. Future studies should evaluate the long-term effectiveness of the interventions facilitated by notifications, individual variability in treatment response, and its impact on overall suicidality expressed among youth.

Conclusions

This study demonstrates the effectiveness of a digital suicidality notification system in initiating timely clinical actions for young people expressing high suicidality. Further, it highlights that using multidimensional assessments for screening and monitoring of symptoms can be critical for providing personalized treatments. Future research should focus on

XSL•FO
RenderX

exploring client and clinician perspectives on using the system to improve its implementation. In addition, examining the longitudinal outcomes of clinical actions initiated by the system would provide insights into their impact on mental health trajectories and suicide rates. Ultimately, this study demonstrates the system's capacity to enhance service responses to suicidality detected in youth through efficient screening, continuous monitoring, and personalized interventions.

Acknowledgments

The funding sources of this study have had no input into the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. This work was supported by the Medical Research Future Fund Applied Artificial Intelligence in Health Care grant (MRFAI000097). MKC was supported by the Australian Government RTP Scholarship. IBH was supported by an NHMRC Research Fellowship (511921). FI was supported by the Bill and Patricia Richie Foundation and NHMRC EL1 Investigator Grant (GNT2018157). SM was supported by the Cottle Family Fellowship.

We thank all the young people who participated in this study and all the staff in the Youth Mental Health Team at the University of Sydney's Brain and Mind Centre, past and present, who contributed to this work.

Authors' Contributions

MKC wrote the manuscript. MKC conducted the formal data analyses. All authors (MKC, IBH, DR, GD, AO, LJB, SM, HML, EMS, and FI) substantively edited and revised the manuscript, approved the final version of the manuscript and agreed both be personally accountable for the author's own contributions and ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Conflicts of Interest

IBH is the Codirector, of Health and Policy at the Brain and Mind Centre (BMC) University of Sydney, Australia. The BMC operates early-intervention youth services at Camperdown under contract to Headspace. Professor Hickie has previously led community-based and pharmaceutical industry-supported (Wyeth, Eli Lily, Servier, Pfizer, AstraZeneca, Janssen Cilag) projects focused on the identification and better management of anxiety and depression. He is the Chief Scientific Advisor to, and a 3.2% equity shareholder in, InnoWell Pty Ltd which aims to transform mental health services through the use of innovative technologies. EMS is a Principal Research Fellow at the BMC at the University of Sydney. She is a Discipline Leader of Adult Mental Health at the School of Medicine, University of Notre Dame, and a Consultant Psychiatrist. She was the Medical Director of the Young Adult Mental Health Unit at St Vincent's Hospital Darlinghurst until January 2021. She has received honoraria for educational seminars related to the clinical management of depressive disorders supported by Servier, Janssen, and Eli-Lilly Pharmaceuticals. She has participated in a national advisory board for the antidepressant compound Pristiq, manufactured by Pfizer. She was the National Coordinator of an antidepressant trial sponsored by Servier. No other competing interests were reported.

Multimedia Appendix 1
A complete list of measures used in the Innowell Platform.
[DOCX File , 25 KB - jmir_v26i1e60879_app1.docx]

Multimedia Appendix 2
Thresholds for suicidality stratification.
[DOCX File , 26 KB - jmir_v26i1e60879_app2.docx]

References

1. Biswas T, Scott JG, Munir K, Renzaho AMN, Rawal LB, Baxter J, et al. Global variation in the prevalence of suicidal ideation, anxiety and their correlates among adolescents: a population based study of 82 countries. EClinicalMedicine 2020;24:100395 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100395] [Medline: 33015595]
2. Robinson J. Early intervention and suicide prevention. Early Interv Psychiatry 2008;2(3):119-121. [doi: 10.1111/j.1751-7893.2008.00070.x] [Medline: 21352146]
3. Abascal-Peiró S, Alacreu-Crespo A, Peñuelas-Calvo I, López-Castromán J, Porras-Segovia A. Characteristics of single vs. multiple suicide attempters among adult population: a systematic review and meta-analysis. Curr Psychiatry Rep 2023;25(11):769-791. [doi: 10.1007/s11920-023-01461-5] [Medline: 37740850]
4. Horwitz AG, Czyz EK, King CA. Predicting future suicide attempts among adolescent and emerging adult psychiatric emergency patients. J Clin Child Adolesc Psychol 2015;44(5):751-761 [FREE Full text] [doi: 10.1080/15374416.2014.910789] [Medline: 24871489]

XSL•FO
RenderX

5. Glenn CR, Lanzillo EC, Esposito EC, Santee AC, Nock MK, Auerbach RP. Examining the course of suicidal and nonsuicidal self-injurious thoughts and behaviors in outpatient and inpatient adolescents. J Abnorm Child Psychol 2017;45(5):971-983 [FREE Full text] [doi: 10.1007/s10802-016-0214-0] [Medline: 27761783]
6. Brodsky BS, Spruch-Feiner A, Stanley B. The zero suicide model: applying evidence-based suicide prevention practices to clinical care. Front Psychiatry 2018;9:33 [FREE Full text] [doi: 10.3389/fpsyt.2018.00033] [Medline: 29527178]
7. Stanley B, Mann JJ. The need for innovation in health care systems to improve suicide prevention. JAMA Psychiatry 2020;77(1):96-98. [doi: 10.1001/jamapsychiatry.2019.2769] [Medline: 31577340]
8. Davis M, Siegel J, Becker-Haimes EM, Jager-Hyman S, Beidas RS, Young JF, et al. Identifying common and unique barriers and facilitators to implementing evidence-based practices for suicide prevention across primary care and specialty mental health settings. Arch Suicide Res 2023;27(2):192-214 [FREE Full text] [doi: 10.1080/13811118.2021.1982094] [Medline: 34651544]
9. Deisenhammer EA, Ing CM, Strauss R, Kemmler G, Hinterhuber H, Weiss EM. The duration of the suicidal process. J. Clin. Psychiatry 2008;70(1):19-24. [doi: 10.4088/jcp.07m03904]
10. Millner AJ, Lee MD, Nock MK. Describing and measuring the pathway to suicide attempts: a preliminary study. Suicide Life Threat Behav 2017;47(3):353-369. [doi: 10.1111/sltb.12284] [Medline: 27477787]
11. Cherny NI, Parrinello CM, Kwiatkowsky L, Hunnicutt J, Beck T, Schaefer E, et al. Feasibility of large-scale implementation of an electronic patient-reported outcome remote monitoring system for patients on active treatment at a community cancer center. JCO Oncol Pract 2022;18(12):e1918-e1926 [FREE Full text] [doi: 10.1200/OP.22.00180] [Medline: 36240475]
12. Garcia SF, Wortman K, Cella D, Wagner LI, Bass M, Kircher S, et al. Implementing electronic health record-integrated screening of patient-reported symptoms and supportive care needs in a comprehensive cancer center. Cancer 2019;125(22):4059-4068 [FREE Full text] [doi: 10.1002/cncr.32172] [Medline: 31373682]
13. Warrington L, Absolom K, Holch P, Gibson A, Clayton B, Velikova G. Online tool for monitoring adverse events in patients with cancer during treatment (eRAPID): field testing in a clinical setting. BMJ Open 2019;9(1):e025185 [FREE Full text] [doi: 10.1136/bmjopen-2018-025185] [Medline: 30782751]
14. Iorfino F, Davenport TA, Ospina-Pinillos L, Hermens DF, Cross S, Burns J, et al. Using new and emerging technologies to identify and respond to suicidality among help-seeking young people: a cross-sectional study. J Med Internet Res 2017;19(7):e247 [FREE Full text] [doi: 10.2196/jmir.7897] [Medline: 28701290]
15. LaMonica HM, Iorfino F, Lee GY, Piper S, Occhipinti J, Davenport TA, et al. Informing the future of integrated digital and clinical mental health care: synthesis of the outcomes from project synergy. JMIR Ment Health 2022;9(3):e33060 [FREE Full text] [doi: 10.2196/33060] [Medline: 34974414]
16. McGorry PD, Tanti C, Stokes R, Hickie IB, Carnell K, Littlefield LK, et al. headspace: Australia's national youth mental health foundation--where young minds come first. Med J Aust 2007;187(S7):S68-S70. [doi: 10.5694/j.1326-5377.2007.tb01342.x] [Medline: 17908032]
17. Iorfino F, Cross SP, Davenport T, Carpenter JS, Scott E, Shiran S, et al. A digital platform designed for youth mental health services to deliver personalized and measurement-based care. Front Psychiatry 2019;10:595 [FREE Full text] [doi: 10.3389/fpsyt.2019.00595] [Medline: 31507465]
18. van Spijker BA, Batterham PJ, Calear AL, Farrer L, Christensen H, Reynolds J, et al. The suicidal ideation attributes scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. Suicide Life Threat Behav 2014;44(4):408-419. [doi: 10.1111/sltb.12084] [Medline: 24612048]
19. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011;168(12):1266-1267 [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
20. Milner AJ, Carter G, Pirkis J, Robinson J, Spittal MJ. Letters, green cards, telephone calls and postcards: systematic and meta-analytic review of brief contact interventions for reducing self-harm, suicide attempts and suicide. Br J Psychiatry 2015;206(3):184-190. [doi: 10.1192/bjp.bp.114.147819] [Medline: 25733570]
21. Bryan CJ, Mintz J, Clemans TA, Leeson B, Burch TS, Williams SR, et al. Effect of crisis response planning vs. contracts for safety on suicide risk in U.S. army soldiers: a randomized clinical trial. J Affect Disord 2017;212:64-72 [FREE Full text] [doi: 10.1016/j.jad.2017.01.028] [Medline: 28142085]
22. Stanley B, Brown GK, Brenner LA, Galfalvy HC, Currier GW, Knox KL, et al. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry 2018;75(9):894-900 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1776] [Medline: 29998307]
23. Miller IW, Camargo CA, Arias SA, Sullivan AF, Allen MH, Goldstein AB, et al. Suicide Prevention in an emergency department population: the ED-SAFE study. JAMA Psychiatry 2017;74(6):563-570 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0678] [Medline: 28456130]
24. Calati R, Courtet P. Is psychotherapy effective for reducing suicide attempt and non-suicidal self-injury rates? meta-analysis and meta-regression of literature data. J Psychiatr Res 2016;79:8-20. [doi: 10.1016/j.jpsychires.2016.04.003] [Medline: 27128172]

XSL•FO
RenderX

25. Harris LM, Huang X, Funsch KM, Fox KR, Ribeiro JD. Efficacy of interventions for suicide and self-injury in children and adolescents: a meta-analysis. Sci Rep 2022;12(1):12313 [FREE Full text] [doi: 10.1038/s41598-022-16567-8] [Medline: 35853962]
26. D'Anci KE, Uhl S, Giradi G, Martin C. Treatments for the prevention and management of suicide: a systematic review. Ann Intern Med 2019;171(5):334-342 [FREE Full text] [doi: 10.7326/M19-0869] [Medline: 31450239]
27. Arnone D, Karmegam SR, Östlundh L, Alkhyeli F, Alhammadi L, Alhammadi S, et al. Risk of suicidal behavior in patients with major depression and bipolar disorder - a systematic review and meta-analysis of registry-based studies. Neurosci Biobehav Rev 2024;159:105594 [FREE Full text] [doi: 10.1016/j.neubiorev.2024.105594] [Medline: 38368970]
28. Leza L, Haro B, López-Goñi JJ, Fernández-Montalvo J. Substance use disorder and lifetime suicidal behaviour: a scoping review. Psychiatry Res 2024;334:115830 [FREE Full text] [doi: 10.1016/j.psychres.2024.115830] [Medline: 38432115]
29. Palma-Álvarez RF, Daigre C, Ros-Cucurull E, Perea-Ortueta M, Ortega-Hernández G, Ríos-Landeo A, et al. Clinical features and factors related to lifetime suicidal ideation and suicide attempts in patients who have had substance-induced psychosis across their lifetime. Psychiatry Res 2023;323:115147 [FREE Full text] [doi: 10.1016/j.psychres.2023.115147] [Medline: 36913874]
30. Ahmedani BK, Westphal J, Autio K, Elsiss F, Peterson EL, Beck A, et al. Variation in patterns of health care before suicide: a population case-control study. Prev Med 2019;127:105796 [FREE Full text] [doi: 10.1016/j.ypmed.2019.105796] [Medline: 31400374]
31. Coppersmith DDL, Dempsey W, Kleiman EM, Bentley KH, Murphy SA, Nock MK. Just-in-time adaptive interventions for suicide prevention: promise, challenges, and future directions. Psychiatry 2022;85(4):317-333. [doi: 10.1080/00332747.2022.2092828] [Medline: 35848800]
32. Miguel C, Cecconi J, Harrer M, van Ballegooijen W, Bhattacharya S, Karyotaki E, et al. Assessment of suicidality in trials of psychological interventions for depression: a meta-analysis. The Lancet Psychiatry 2024;11(4):252-261. [doi: 10.1016/s2215-0366(24)00027-0]
33. Hawton K, Arensman E, Townsend E, Bremner S, Feldman E, Goldney R, et al. Deliberate self harm: systematic review of efficacy of psychosocial and pharmacological treatments in preventing repetition. BMJ 1998;317(7156):441-447 [FREE Full text] [doi: 10.1136/bmj.317.7156.441] [Medline: 9703526]
34. Varidel M, Hickie IB, Prodan A, Skinner A, Marchant R, Cripps S, et al. Dynamic learning of individual-level suicidal ideation trajectories to enhance mental health care. Npj Ment Health Res 2024;3(1):26 [FREE Full text] [doi: 10.1038/s44184-024-00071-0] [Medline: 38849429]
35. Capon W, Hickie IB, Fetanat M, Varidel M, LaMonica HM, Prodan A, et al. A multidimensional approach for differentiating the clinical needs of young people presenting for primary mental health care. Compr Psychiatry 2023;126:152404 [FREE Full text] [doi: 10.1016/j.comppsych.2023.152404] [Medline: 37524044]
36. Chong MK, Hickie IB, Cross SP, McKenna S, Varidel M, Capon W, et al. Digital application of clinical staging to support stratification in youth mental health services: validity and reliability study. JMIR Form Res 2023;7:e45161 [FREE Full text] [doi: 10.2196/45161] [Medline: 37682588]
37. Pan W, Wang X, Zhou W, Hang B, Guo L. Linguistic analysis for identifying depression and subsequent suicidal ideation on weibo: machine learning approaches. Int J Environ Res Public Health 2023;20(3):2688 [FREE Full text] [doi: 10.3390/ijerph20032688] [Medline: 36768053]
38. Safa R, Edalatpanah S, Sorourkhah A. Deep learning in personalized healthcare and decision suppor. In: Predicting Mental Health Using Social Media: A Roadmap for Future Development. Amsterdam: Elsevier; 2023:285-303.
39. Too LS, Spittal MJ, Bugeja L, Reifels L, Butterworth P, Pirkis J. The association between mental disorders and suicide: a systematic review and meta-analysis of record linkage studies. J Affect Disord 2019;259:302-313 [FREE Full text] [doi: 10.1016/j.jad.2019.08.054] [Medline: 31450139]
40. Zhou R, Foo JC, Nishida A, Ogawa S, Togo F, Sasaki T. Longitudinal relationships of psychotic-like experiences with suicidal ideation and self-harm in adolescents. Eur Child Adolesc Psychiatry 2024;33(6):1977-1985 [FREE Full text] [doi: 10.1007/s00787-023-02299-1] [Medline: 37740799]
41. Rumble ME, McCall WV, Dickson DA, Krystal AD, Rosenquist PB, Benca RM. An exploratory analysis of the association of circadian rhythm dysregulation and insomnia with suicidal ideation over the course of treatment in individuals with depression, insomnia, and suicidal ideation. J Clin Sleep Med 2020;16(8):1311-1319 [FREE Full text] [doi: 10.5664/jcsm.8508] [Medline: 32329435]
42. Amiri S, Khan MA. Prevalence of non-suicidal self-injury, suicidal ideation, suicide attempts, suicide mortality in eating disorders: a systematic review and meta-analysis. Eat Disord 2023;31(5):487-525. [doi: 10.1080/10640266.2023.2196492] [Medline: 37021980]
43. Iorfino F, Varidel M, Marchant R, Cripps S, Crouse J, Prodan A, et al. The temporal dependencies between social, emotional and physical health factors in young people receiving mental healthcare: a dynamic bayesian network analysis. Epidemiol Psychiatr Sci 2023;32:e56 [FREE Full text] [doi: 10.1017/S2045796023000616] [Medline: 37680185]
44. Skinner A, Osgood ND, Occhipinti J, Song YJC, Hickie IB. Unemployment and underemployment are causes of suicide. Sci Adv 2023;9(28):eadg3758 [FREE Full text] [doi: 10.1126/sciadv.adg3758] [Medline: 37436996]

XSL•FO
RenderX

45. Iorfino F, Ho N, Carpenter JS, Cross SP, Davenport TA, Hermens DF, et al. Predicting self-harm within six months after initial presentation to youth mental health services: a machine learning study. PLoS One 2020;15(12):e0243467 [FREE Full text] [doi: 10.1371/journal.pone.0243467] [Medline: 33382713]
46. Donnelly HK, Han Y, Kim S, Lee DH. Predictors of suicide ideation among South Korean adolescents: a machine learning approach. J Affect Disord 2023;329:557-565. [doi: 10.1016/j.jad.2023.02.079] [Medline: 36828148]
47. Fox KR, Huang X, Guzmán EM, Funsch KM, Cha CB, Ribeiro JD, et al. Interventions for suicide and self-injury: a meta-analysis of randomized controlled trials across nearly 50 years of research. Psychol Bull 2020;146(12):1117-1145. [doi: 10.1037/bul0000305] [Medline: 33119344]
48. Cohen AS, Fedechko T, Schwartz EK, Le TP, Foltz PW, Bernstein J, et al. Psychiatric risk assessment from the clinician's perspective: lessons for the future. Community Ment Health J 2019;55(7):1165-1172. [doi: 10.1007/s10597-019-00411-x] [Medline: 31154587]
49. Bradford S, Rickwood D. Acceptability and utility of an electronic psychosocial assessment (myAssessment) to increase self-disclosure in youth mental healthcare: a quasi-experimental study. BMC Psychiatry 2015;15:305 [FREE Full text] [doi: 10.1186/s12888-015-0694-4] [Medline: 26627041]
50. Victor SE, Salk RH, Porta G, Hamilton E, Bero K, Poling K, et al. Measurement-based care for suicidal youth: outcomes and recommendations from the services for teens at risk (STAR) center. PLoS One 2023;18(4):e0284073 [FREE Full text] [doi: 10.1371/journal.pone.0284073] [Medline: 37023038]
51. Freedman DE, Waddell AE, Bourdon A, Lam HT, Wang K. Educating mental health trainees about measurement-based care: a scoping review. Acad Psychiatry 2023;47(2):187-195. [doi: 10.1007/s40596-023-01749-x] [Medline: 36829099]
52. Solstad SM, Kleiven GS, Moltu C. Complexity and potentials of clinical feedback in mental health: an in-depth study of patient processes. Qual Life Res 2021;30(11):3117-3125. [doi: 10.1007/s11136-020-02550-1] [Medline: 32556824]
53. Wisniewski H, Gorrindo T, Rauseo-Ricupero N, Hilty D, Torous J. The role of digital navigators in promoting clinical care and technology integration into practice. Digit Biomark 2020;4(1):119-135 [FREE Full text] [doi: 10.1159/000510144] [Medline: 33442585]
54. Meyer A, Wisniewski H, Torous J. Coaching to support mental health apps: exploratory narrative review. JMIR Hum Factors 2022;9(1):e28301 [FREE Full text] [doi: 10.2196/28301] [Medline: 35258468]
55. Hans PK, Gray CS, Gill A, Tiessen J. The provider perspective: investigating the effect of the electronic patient-reported outcome (ePRO) mobile application and portal on primary care provider workflow. Prim Health Care Res Dev 2017;19(02):151-164. [doi: 10.1017/s1463423617000573]
56. Carpenter JS, Iorfino F, Cross S, Nichles A, Zmicerevska N, Crouse JJ, et al. Cohort profile: the Brain and mind Centre cohort: tracking multidimensional outcomes in young people presenting for mental healthcare. BMJ Open 2020;10(3):e030985 [FREE Full text] [doi: 10.1136/bmjopen-2019-030985] [Medline: 32229519]
57. Hawton K, Saunders KE, O'Connor RC. Self-harm and suicide in adolescents. Lancet 2012;379(9834):2373-2382. [doi: 10.1016/S0140-6736(12)60322-5] [Medline: 22726518]
58. Bryan CJ, Butner JE, May AM, Rugo KF, Harris JA, Oakey DN, et al. Nonlinear change processes and the emergence of suicidal behavior: a conceptual model based on the fluid vulnerability theory of suicide. New Ideas Psychol 2020;57:100758 [FREE Full text] [doi: 10.1016/j.newideapsych.2019.100758] [Medline: 32123464]
59. Lutz W, Schwartz B, Delgadillo J. Measurement-based and data-informed psychological therapy. Annu Rev Clin Psychol 2022;18(1):71-98. [doi: 10.1146/annurev-clinpsy-071720-014821] [Medline: 34910567]

Abbreviations

C-SSRS: Columbia-Suicide Severity Rating Scale
NEET: Not in employment, education, or training
SIDAS: Suicidal Ideation Attributes Scale

Edited by T de Azevedo Cardoso; submitted 27.05.24; peer-reviewed by V Astha, U Sinha, R Safa; comments to author 06.08.24; revised version received 23.08.24; accepted 25.09.24; published 18.12.24.

Please cite as:
Chong MK, Hickie IB, Ottavio A, Rogers D, Dimitropoulos G, LaMonica HM, Borgnolo LJ, McKenna S, Scott EM, Iorfino F
A Digital Approach for Addressing Suicidal Ideation and Behaviors in Youth Mental Health Services: Observational Study
J Med Internet Res 2024;26:e60879
URL: https://www.jmir.org/2024/1/e60879
doi:10.2196/60879
PMID:39693140

XSL•FO
RenderX

©Min K Chong, Ian B Hickie, Antonia Ottavio, David Rogers, Gina Dimitropoulos, Haley M LaMonica, Luke J Borgnolo, Sarah McKenna, Elizabeth M Scott, Frank Iorfino. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media

Feipeng Guo[1,2], PhD; Zixiang Liu[1], MSc; Qibei Lu[3], PhD; Shaobo Ji[4], PhD, Prof Dr; Chen Zhang[5], MSE

[1]Modern Business Research Center, Zhejiang Gongshang University, Hangzhou, China
[2]School of Management and E-Business, Zhejiang Gongshang University, Hangzhou, China
[3]School of International Business, Zhejiang International Studies University, Hangzhou, China
[4]Sprott School of Business, Carleton University, Ottawa, ON, Canada
[5]General Manager's Office, Hangzhou Gaojin Technology Co, Ltd, Hangzhou, China

Corresponding Author:
Qibei Lu, PhD
School of International Business
Zhejiang International Studies University
299 Liuhe Road
Xihu District
Hangzhou, 310030
China
Phone: 86 13857659112
Email: luqibei@zisu.edu.cn

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e57233

Abstract

Background: The COVID-19 pandemic raised wide concern from all walks of life globally. Social media platforms became an important channel for information dissemination and an effective medium for public sentiment transmission during the COVID-19 pandemic.

Objective: Mining and analyzing social media text information can not only reflect the changes in public sentiment characteristics during the COVID-19 pandemic but also help the government understand the trends in public opinion and reasonably control public opinion.

Methods: First, this study collected microblog comments related to the COVID-19 pandemic as a data set. Second, sentiment analysis was carried out based on the topic modeling method combining latent Dirichlet allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT). Finally, a machine learning logistic regression (ML-LR) model combined with a sparse matrix was proposed to explore the evolutionary trend in public opinion on social media and verify the high accuracy of the model.

Results: The experimental results show that, in different stages, the characteristics of public emotion are different, and the overall trend is from negative to positive.

Conclusions: The proposed method can effectively reflect the characteristics of the different times and space of public opinion. The results provide theoretical support and practical reference in response to public health and safety events.

(J Med Internet Res 2024;26:e47508) doi:10.2196/47508

KEYWORDS

COVID-19; social media public opinion; microblog; sentiment analysis; topic modeling

XSL•FO
RenderX

Introduction

Due to the global influence of COVID-19, it has recently become a research hot spot in many fields [1]. Public opinion on social media refers to the public opinion expressed on social media software as the communication platform and social issues as the main topic, including public cognition and attitude [2]. Public opinion generated by public security and health emergencies such as infectious diseases spreads rapidly and is of an explosive nature [3]. Reasonable control of public opinion played an important role in curbing public panic during the COVID-19 pandemic [4].

In recent years, social media platforms have gradually become the main information dissemination medium for the public [5]. The use of social media to measure public attention has also been gradually applied to the research of acute infectious diseases, such as influenza A (H1N1) [6]. Ahmad et al [7] demonstrated a correlation between the daily number of COVID-19–related tweets and the daily number of COVID-19 cases and deaths in Iran and Turkey. Public knowledge was limited due to a lack of information about, inexperience with, and poor awareness of the nature of the virus. Text-based data on social media platforms contain rich information about public opinion and sentiment [8]. As of September 2022, the number of monthly active users of microblogging platforms reached 584 million, with mobile use accounting for 95%, and the number of daily active users reached 253 million. Although social media platforms disseminate information, they also present changes in public emotions [9,10]. Text-based data on social media provide accurate geolocations, rich emotional information, and distinct topics; the different topic information within social media text data is an important data source for public opinion research [11].

Most traditional spatiotemporal studies of public opinion use the administrative region as the analysis unit or the coordinates of users' blogs or comments to analyze the geographic density [12]. Box-Steffensmeier and Moses [13] explored the level of sentiment and severity of the pandemic in each state in the United States during the COVID-19 pandemic; however, they used only the administrative region as the unit of analysis, and the scope of analysis was limited. Liu et al [14] conducted spatiotemporal analysis of the evolution of public opinion during the COVID-19 pandemic based on microblog data and visualized public sentiment and the number of microblogs in provinces and cities across the country. Although the research scope was expanded to the whole country, the selected spatial units were relatively concentrated and were not able to reflect the differences in public opinion at different scales. The spread of the pandemic has obvious regional differences between cities. Liu and Liu [15] used the GeoSEn geoparser to obtain geographic location information from text data and converted it into coordinates for emotional spatial analysis, but the monitoring location was not accurate, which meant the analytic results deviated from the spatial standard. In particular, the results from administrative units at different levels are quite different, while there is a small number of microblogs with user location coordinates, which leads to further errors in the results. In addition, the propagation of public opinion on social media is regionally correlated. The scope of public concern not only is limited to local areas but also includes information from other adjacent areas, which cannot be effectively mined by traditional public opinion analysis methods [16].

Social media data published by users contain rich geographic information, and many scholars have used social media to simulate and predict real events, such as predicting the spread of influenza, detecting earthquakes, and monitoring air quality. To monitor public opinion, Hu et al [17] proposed a model combining latent Dirichlet allocation (LDA) based on the document generation model and the k-means text classification algorithm based on genetic optimization; this combination can improve the accuracy of the clustering algorithm to monitor public opinion. Specifically, this method can identify and track topics related to public opinion on Twitter as well as mine the public's emotional characteristics in different fields. Zhang et al [18] studied algorithms to monitor a public opinion network, compared and analyzed the advantages and disadvantages of different text classification and emotional tendency algorithms, and discussed future trends in algorithm development to monitor public panic during the COVID-19 pandemic in different geographical locations. Microblogs contain emotional information from social users. Researchers such as Jang et al [19] have analyzed microblog data to determine the spatial and temporal variations in emotions expressed by Twitter users in North America. They also analyzed values given to emotions and the relationships between the influencing factors, which highlighted the study area index and emotional score. In their study, the spatiotemporal variation in the public's attention to the pandemic and emotion were discussed. Storey and O'Leary [20] constructed a global sentiment map of Twitter users and analyzed the spatiotemporal characteristics of urban sentiment by calculating the sentiment information from social media text-based data. The risk of pandemic spread at multiple time nodes was evaluated by converting geographic information from text into coordinates for emotional spatial analysis. Alhashmi et al [21] proposed a sentiment classification network model combining part of a speech attention mechanism and long short-term memory network, which can fully mine the relationship between emotional polarity words and emotional target words of sentences and analyze the correlation between the spatial distribution of online public opinion and high-risk areas of the pandemic through spatiotemporal migration.

This study aimed to mine social media comments on COVID-19 topics and explore their spatiotemporal distribution characteristics. The specific research period was from July 2020 to June 2021. More than 100,000 comments from 12 People's Daily microblogs were obtained through the microblog application programming interface (API) interface, and 2000 samples were screened for feature extraction and sentiment analysis. On this basis, we proposed 3 stages of the pandemic covering the changes in public opinion. Meanwhile, through topic modeling, 12 core topics were identified to map the potential drivers of these changes. The purpose of this study was to help government and public health security departments understand the public's emotions and views on the COVID-19 pandemic, formulate more reasonable pandemic prevention

XSL•FO
RenderX

policies, effectively respond to the COVID-19 pandemic, and maintain social stability.

The rest of the paper is structured as follows: The Methods section introduces the research ideas and key algorithms used in this study. The Results section extends the analysis of the experimental results. The Discussion section concludes the paper with contributions, limitations, and directions for future research.

Methods

Overview

First, based on the API interface for the Sina microblog, data from COVID-19–related microblog comments from July 1, 2020, to June 1, 2021, which was in the middle of the COVID-19 pandemic, were obtained and preprocessed. Second, we considered the characteristic of low granularity as well as short words in the microblog comments [22]. Based on the SnowNLP Chinese sentiment vocabulary ontology, the Dalian University of Technology sentiment classification dictionary [23] is optimized to identify sentiment features, and a multidimensional analysis model of public opinion on social media during the pandemic was constructed from the perspective of spatiotemporal correlation. Third, according to the multiscale division method of public opinion, the term frequency-inverse document frequency (TF-IDF) method was used to explore the semantic features of public opinion. Fourth, according to the difference in public opinion data characteristics, a model combining LDA and Bidirectional Encoder Representations from Transformers (BERT) [24] was used to classify the topics. The number of topics was determined according to minimum perplexity, and the topic weight was established using chi-square test results. Finally, a machine learning logistic regression (ML-LR) sparse matrix fusion model was used to analyze the evolutionary trend in public opinion combined with the monthly pandemic notification information. The technical route and structure of this paper are shown in Figure 1.

Figure 1. The system architecture of this study. LDA: latent Dirichlet allocation; ML-LR: machine learning logistic regression; TF-IDF: term frequency-inverse document frequency.

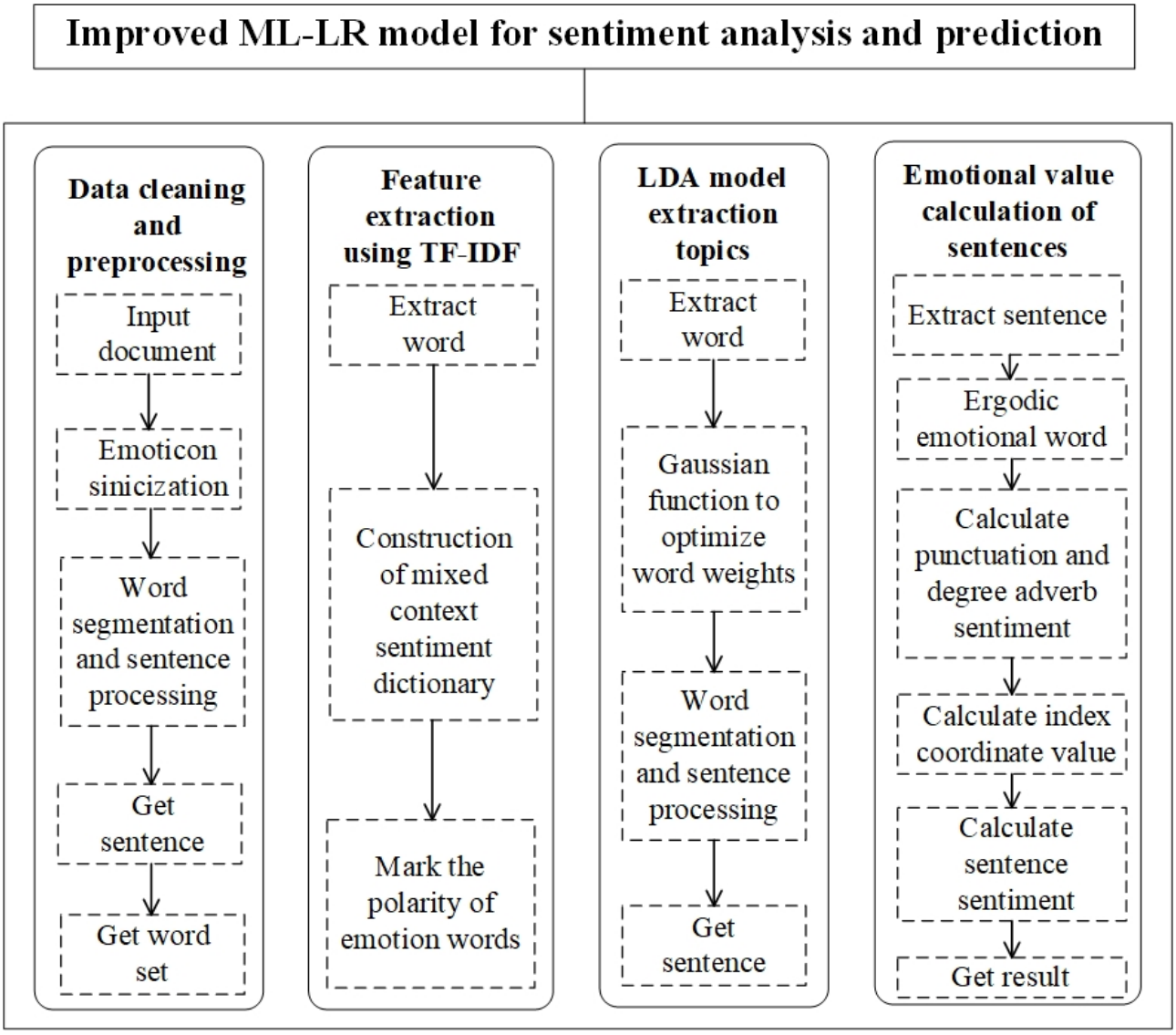

XSL•FO
RenderX

Data Extraction

This study collected data for topics related to COVID-19 in the monthly hot topic search of the People's Daily microblogs from July 2020 to June 2021, including the hot topic name, topic search volume, and date of the blog post with the most searched topic, and selected more than 100,000 comments from 12 microblogs. These 12 blogs have a high degree of heat (eg, likes, comments, and retweets), and the blog topics represent the hot events of the corresponding timeline, which can be regarded as a microcosm of the different development stages of the COVID-19 pandemic in China. The specific data from these 12 microblogs are shown in Table 1. Based on the ArcGIS interface of the geocoding package Geocoder, the geographic latitude and longitude of the check-in location coordinate system were then obtained, and the IP address geographic location information of the user was calculated. Through the Requests library of the Python language, the data interface was requested and parsed, and 2000 comments were screened for sentiment analysis. An example of the crawled microblog data is shown in Table 2.

Table 1. Data information statistics of 12 representative blogs about the COVID-19 pandemic in China from July 2020 to June 2021.

Topic	Date	Likes, n	Comments, n	Forwards, n
127 new cases were confirmed in 31 provinces	July 31, 2020	113,000	11,000	1453
Three new imported cases were reported in Wuhan	August 2, 2020	62,000	4987	1015
Yunnan released details of a new asymptomatic case	September 1, 2020	56,000	2973	811
Qingdao has collected 1.03 million samples for nucleic acid testing	October 12, 2020	61,000	3257	1580
Wu Zunyou said cold chain imports or become the source of the pandemic in China	November 11, 2020	30,000	2464	3072
A COVID-19 vaccine has been approved for marketing in China	December 31, 2020	720,000	40,000	93,000
Zhong Nanshan thanks you for not going home during the Spring Festival	January 25, 2021	113,000	8871	6025
China's medium and high-risk areas are cleared today	February 22, 2021	8935	1093	1498
Zhong Nanshan says there are dangers in not getting vaccinated in time	March 31, 2021	453,000	37,000	13,000
Cluster infection occurred among Chinese students	April 30, 2021	6697	673	235
The confirmed cases in Liwan, Guangzhou were infected with Indian variant strain	May 23, 2021	270,000	10,000	5009
Zhong Nanshan said the concept of close connection should be updated	June 26, 2021	101,000	4134	15,000

Table 2. Sample comment information about the COVID-19 pandemic in China crawled from a microblog platform from July 2020 to June 2021.

ID	Date	Review	IP
User 1	October 15, 2020	It is hoped that the test reagents can be distributed to every city in the country, whether developed or backward. The more backward provinces are the bigger the crisis, the news is closed, and the awareness of prevention is weak	Jilin Province
User 2	January 24, 2021	The most frightening thing is that the virus is still in its initial stage, and how it will mutate in the future is uncertain. What if it becomes more deadly	Guangdong Province
User 3	April 8, 2021	We call on the government to strictly control the price of masks and crack down on the bad behavior of price gouging by bad businesses in special periods	Beijing City
User 4	June 17, 2021	With such attention and rapid action and scientific prevention and treatment, the country will be able to effectively control and defeat the COVID-19 pandemic	Sichuan Province

Data Preprocessing

Issues with the crawled text data included missing column values, numerical anomalies, and special symbols, which need to be supplemented and filtered. First, the Jieba word segmentation tool was used to process the text segmentation, and a custom dictionary was established for the high-frequency words and proper nouns related to the pandemic. Second, the Trie tree structure was used for efficient word graph scanning, and a stop word list was established to filter out the noisy data from the text. Third, the text information was converted into vectors (ie, feature engineering transformation). The model can only input numbers (like vectors), not text; therefore, before running the model, any signal needed to be converted into a digital signal (eg, number, vector, matrix, tensor) that can be recognized by the model. The regular word segmentation result was used as the input for the model. The data preprocessing results are shown in Table 3.

XSL•FO
RenderX

Table 3. Comparison of a sample of annotated texts about COVID-19 in China before and after data preprocessing from July 2020 to June 2021.

Number	Review	
	Before pretreatment	After pretreatment
1	Now the pandemic is spreading, the whole network selling masks are doubling the price! Really no conscience!	[pandemic], [spread], [sell], [mask], [price rise], [no], [conscience]
2	We should praise the medical staff and scientific researchers who are still struggling in the front line!	[for], [struggle], [medical care], [scientific research], [personnel], [praise]
3	You don't need to panic, but you do need to be vigilant. If you notice suspicious symptoms, go to the hospital.	[don't], [in], [but], [vigilant], [if], [notice], [symptoms], [to], [hospital], [medical]

The specific text data preprocessing process included (1) removing duplicate data, (2) removing useless characters, (3) using Jieba word segmentation, (4) removing stop words, (5) removing low-frequency words, (6) removing any tags such as HTML, and (7) using the decision tree method to supplement the missing column values and integrate the data.

Sentiment Classification

To reflect the characteristics of public opinion, a method based on dictionary and thesaurus matching was used. The Dalian University of Technology sentiment dictionary was used to scan the strings in the dictionary one by one and provide auxiliary annotations for sentiment classification. This method combines the characteristics of the topic and Chinese grammar and expands the original dictionary from the dimensions of “word,” “part of speech,” “word sense number,” “sentiment classification,” “intensity,” and “polarity.”

The basic dictionary selected for this paper was the emotional vocabulary library of the Dalian University of Technology, which divides emotions into 7 categories (“anger,” “disgust,” “fear,” “sadness,” “surprise,” “good,” “happy”) and 21 subcategories. The initial emotional intensity was set to 5 levels (1, 3, 5, 7, 9), which is more detailed than other dictionaries. To facilitate sentiment calculation by a computer, we divided the polarity of sentiment words into 2 categories: positive (1) and negative (0). The formula for word sentiment is given in Equation 1:

$$s(w)=v(w)p(w) \quad \textbf{(1)}$$

where $s(w)$ represents the sentiment value of the term, $v(w)$ is the sentiment intensity of the word, and $p(w)$ denotes the sentiment polarity of the word.

First, we needed to calculate the TF-IDF values of the words in the text to obtain the feature matrix. The TF-IDF value of the feature item w_{ij} was calculated as shown in Equations 2-5:

Here, w_{ij} represents the j_{th} word appearing in the microblog text-based data from the i_{th} day. $C(w_{ij})$ is the number of occurrences of the term w_{ij}. D is the total number of documents in the daily microblog text-based data $|D_i|$ for the number of words in a document D_i. denotes the bag-of-words vector of the document D_i. The function $I(w_{ij},D_i)$ takes the value 1 or 0, where 1 means that the document D_i contains the word w_{ij}; otherwise, it takes the value 0. The part-of-speech and sentiment classification of the words are shown in Table 4.

XSL•FO
RenderX

Table 4. Part-of-speech and sentiment classification of words in the text of a sample of microblog comments with the topic of COVID-19 pandemic behavior in China from July 2020 to June 2021.

Word	Part of speech	Sentiment classification	Intensity	Polarity	Auxiliary classification
Anger	Adjective	NAU[a]	7	–1	NaN[b]
Fear	Adjective	NI[c]	5	–1	NG[d]
Grief	Verb	NB[e]	7	–1	NJ[f]
Surprise	Adjective	PC[g]	6	1	NaN
Hope	Noun	PD[h]	4	1	PH[i]
Pleasure	Verb	PA[j]	4	1	PE[k]

[a]NAU: anger.
[b]NaN: null value.
[c]NI: panic.
[d]NG: shame.
[e]NB: sadness.
[f]NJ: disappointment.
[g]PC: surprise.
[h]PD: respect.
[i]PH: praise.
[j]PA: happiness.
[k]PE: peace of mind.

Feature Extraction

This study used a topical term extraction method based on the combination of LDA and BERT. First, the review text was segmented into words, and the word vector was generated based on the hybrid pretrained model. The k-means clustering method was then used to cluster the word vectors, and the words with higher word frequencies were selected as representative words in each class. At the same time, other words in the same class were replaced by representative words to reduce the amount of input data. Finally, the biterm topic model (BTM) for short text was used to extract the topics.

The BTM is aimed at the characteristics of short text and extracts more informative topics using co-occurrence patterns of word pairs in the whole corpus. The modeling process of the BTM generates a corpus of word pairs. Model training and model parameter inference are then performed based on the generated corpus. Finally, the topic distribution and word distribution on the corpus are obtained. If we suppose there are M feature words, $|B|$ word pairs, and K topics in the corpus, then the corpus-level topic distribution is denoted by θ, the distribution of topic K is denoted by θ_k, and the word distribution is denoted by Ø, as shown in Equation 6.

where z represents a topic extracted, $p(z=k)$ is the probability of occurrence of topic k, and $p(z=k)$ is the probability of occurrence of words under topic k.

Perplexity is widely used to measure the fitting effect of an LDA model. Perplexity is a probabilistic model evaluation metric that evaluates the predictive ability of a trained model on new unseen data. LDA perplexity is the perplexity when the model predicts new unseen text. After understanding the calculation of the statement probability, the perplexity of the statement $s=w_1, w_2, w_3, ..., w_n$ can be defined as seen in Equation 7:

To calculate the perplexity of LDA, we needed to utilize the generation process of the model. Specifically, we needed to split the text data into training and test sets. The training set was used to build the LDA model, and the test set was used to calculate the perplexity of the model. In the test set, we treated each document as a sequence, taking one word at a time from the sequence and inferring the topic of that word based on the current model. Once the topic of the word was predicted, the model parameters were recalculated. In this process, we needed to calculate the perplexity of the text using the topic of the predicted word and the frequency of all the words.

Sentiment Analysis

The microblog comment text was divided into positive and negative sentiment, and the value represented the probability that the text contained that sentiment. The values ranged from 0 to 1; closer to 1 tended to be positive, and closer to 0 tended to be negative. To improve the accuracy of sentiment prediction, the sentiment classification model needed to be retrained. The main steps are described in the following paragraphs.

First, the training method of the Bayesian model was used to train the sentiment classifier, and the classification method in the Bayesian classification was then used to predict the sentiment classification and test the accuracy of the model. Finally, the newly trained model was saved.

XSL•FO
RenderX

The daily microblog text set $D=\{d_1, d_2, ..., d_m\}$. The variable m represents the number of microblog crawls per day. The sentiment score [illegible] of each microblog was obtained using the newly trained sentiment analysis model. In the actual judgment, to make the visualization results more intuitive, the value of [illegible] was lowered by 1. That is, when the return value was in the interval of –1 to 0, the sentiment probability value was negative; when the return value was between 0 and 1, the sentiment probability value was positive. The sentiment score S_D was calculated as shown in Equation 8:

[illegible]

First, the feature words were inputted, and the sentiment polarity was classified as 1 ("positive") or 0 ("negative"). Second, the TF-IDF calculation converted the words in the text into a word frequency matrix, and the matrix element $a[i][j]$ represented the word frequency of j words in type i text. *MemoryError* was used to control the parameter. The transformer was then used to calculate the TF-IDF weight of each word. The first *fit_transform* was used to calculate the TF-IDF, and the second *fit_transform* was used to convert the text into a term frequency matrix to obtain all the words in the bag-of-words model. The TF-IDF matrix was then extracted. The element $w[i][j]$ represented the TF-IDF weight of j word in the i type text, and the data were divided by combining the sparse matrix. Finally, the logistic regression classification method was used to calculate the model benchmark indicators. The binary classification ML-LR model is given in Equations 9 and 10:

[illegible]

where w is the weight, b is the bias (b can be 0), and we can expand the weight and input vector a little, as shown in Equation 11:

$$w=(w^{(1)}, w^{(2)}, ..., w^{(m)}, b)^T, X=(x^{(1)}, x^{(2)}, ..., x^{(m)}, 1)^T \quad \textbf{(11)}$$

Ethical Considerations

This study did not require ethics approval because all data collected were publicly available. There is no means within this paper or its supporting materials to establish the identification of users and their corresponding tweets.

Results

Analysis of the Change in Public Opinion Based on LDA Topic Modeling

We identified 12 key dates as turning points for mood scores during this period. Figure 2 shows the general trend in the monthly sentiment changes of microblog users during the COVID-19 pandemic. In general, the overall classification results show obvious clustering according to the time series.

In the first stage (July 2020 to October 2020), strong infectious tendencies and unknown virus characteristics exacerbated public panic, and user sentiment ratings were generally low. In the second phase (November 2020 to February 2021), changes in sentiment scores were relatively stable, except for a large decrease in January 2021 and February 2021, possibly due to material shortages and population movement that accelerated virus transmission. In the third phase (March 2021 to June 2021), the proportion of positive emotions gradually increased thanks to the successful development and widespread distribution of vaccines. The data show that the changes in the number of infected people and the effectiveness of pandemic prevention measures were important factors affecting the change in the polarity of public sentiment.

In the face of the pandemic, people's instinctive psychological reaction was panic and anxiety, and these negative emotions were particularly obvious in the first stage. In the second stage, government departments strengthened the control of public opinion on social media. As a result, the public's confidence about fighting the pandemic was significantly enhanced, and the emotional polarity gradually improved. In the third stage, the proportion of positive emotions increased significantly, and negative emotions gradually turned into affirmation and support for antipandemic actions. Figure 3 shows the keywords in the monthly hot topic searches on the microblogs during the COVID-19 pandemic.

The SnowNLP tool was used to calculate the sentiment value, and the text was classified according to the threshold. The LDA model was then used to mine the topics. The hyperparameters of the model were set as $\alpha=50/K$ and $\gamma=0.01$. The number of iterations was set at 1000. The number of potential topics was determined as $K=12$ according to the perplexity. The same weight was given to the topics with a monthly cycle by combining the change in the intensity of the influence of the propagation of the microblog information.

We obtained 12 topics from the feature extraction of microblog topics during the COVID-19 pandemic (see Table 5). Among them, pandemic prevention and people's livelihoods were the focus of public attention.

XSL•FO
RenderX

Figure 2. Monthly distribution of the sentiment polarity in microblog comments about the COVID-19 pandemic in China from July 2020 to June 2021.

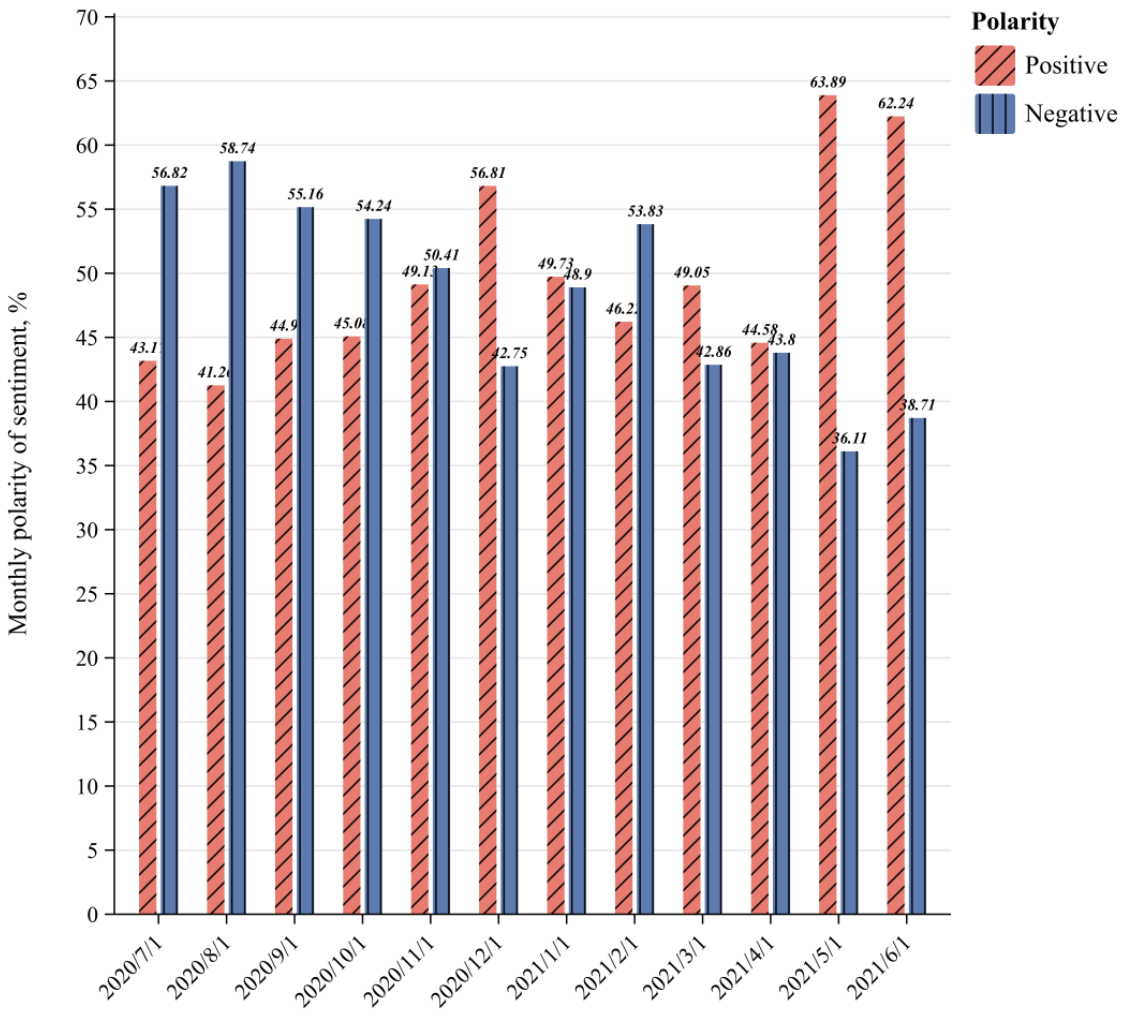

Figure 3. Monthly statistics of the keywords in the hot topic searches regarding COVID-19 pandemic behavior in China on microblogs from July 2020 to June 2021.

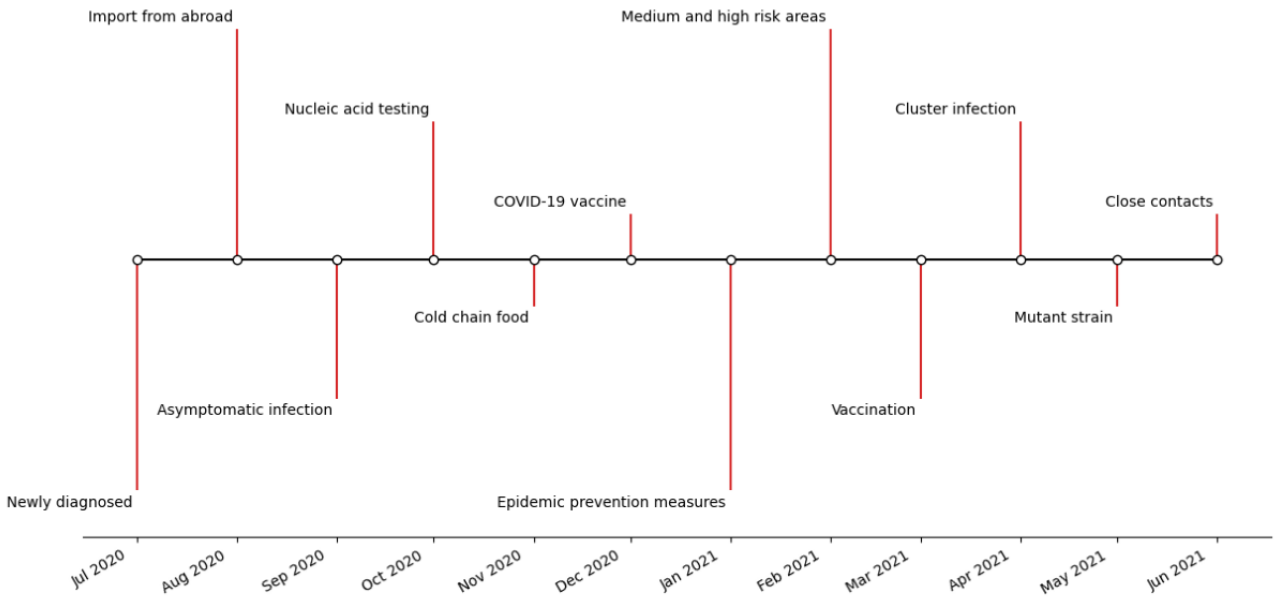

XSL•FO
RenderX

Table 5. Topic feature extraction analysis of microblog comments about the COVID-19 pandemic at different stages in China from July 2020 to June 2021.

Stage and subject number	Subject content	Keywords extracted by the LDA[a] model
Phase 1		
T1	New confirmed cases	Pandemic rebound, death, persistence, fear, migration, virus spread, contagion risk, pandemic prevention
T2	Import from abroad	Disinfect, detect, security, risk level, classification, source control, viral activity, personnel contact
T3	Asymptomatic infection	Isolation, therapy, monitor, incubation personal, close contacts, medical observe, infectiousness, immunity
T4	Nucleic acid testing (NAT)	24 hours, region, sampling method, screening, antigen detection, sampling point, information input, detection result
Phase 2		
T5	Cold chain food	Traceability, cheek, disinfect, safety, risk, touch, nucleic acid testing, virus
T6	COVID-19 vaccine	Clinical trials, safety, side effect, output, validity, develop, price, inoculate
T7	Pandemic prevention measures	Face mask, health, disinfect, body temperature, nucleic acid testing, immunity, quarantine, vaccine
T8	Medium or high-risk areas	Cases, asymptomatic infection, spread risk, medical insurance, living necessities, regional control, nucleic acid testing, health monitoring
Phase 3		
T9	Vaccination	Vaccination rate, necessity, population immunity, adverse reactions, antibody, allergy, personal protection, inoculation contraindication
T10	Cluster infection	Source of infection, chain of transmission, close contacts, gathering activity, nucleic acid testing, pandemic prevention awareness, route of transmission, policy control
T11	Variant strain	Omicron, Delta, critically patients, panic, contagious, pathogenic, mortality rate, drug resistance
T12	Close contact	Centralized isolation, medical observation, nucleic acid testing, body temperature, abnormal signs, incubation period, clinical symptoms, movement tracking

[a]LDA: latent Dirichlet allocation.

Multimedia Appendix 1 shows the results of topic modeling. In the first stage, the main topics were mainly focused on pandemic prevention and control. In the second stage, in addition to vaccines, food safety became the focus. The core topics of the third stage were COVID-19 vaccination and variants.

ML-LR Public Opinion Sentiment Analysis With a Sparse Matrix

To explore the spatial distribution of public emotions, we subtracted the same types of emotions from provinces and regions in adjacent stages according to regional differences and obtained the spatiotemporal changes in negative emotions, as shown in Figure 4. Taking provincial administrative regions as the basic statistical unit and the 3 evolution stages as fixed time intervals, the natural break point method was used to divide the proportion of negative emotions.

Neighboring provinces had similar emotional characteristics; that is, the spatiotemporal similarity of emotions was high. Moreover, the proportion of negative emotions was related to the severity of the local pandemic, and a high value for negative emotions had a spatial distribution similar to that of areas with a more severe pandemic situation. In addition, provinces with many confirmed cases had a high proportion of negative emotions, while provinces with a small number of confirmed cases had a relatively low proportion of negative emotions.

Combined with China's daily real-time notifications regarding the pandemic, the first stage of the pandemic in China occurred during the outbreak period, with the number of confirmed cases reaching a peak in August 2020 and the confirmed case rate gradually slowing down since then. As can be seen in Figure 5, during the period from July 2020 to January 2021, anger, disgust, fear, sadness, and other negative emotions prevailed. When combined with topic digging, this result shows that people's concerns regarding the COVID-19 pandemic included fear of human-to-human transmission, a domestic outbreak, and imported cases from other countries.

During the second stage (February 2021 to June 2021), the sentiment scores basically showed positive sentiment, indicating that the government adopted effective response policies and that public sentiment tended to stabilize again. Positive emotions such as surprise, good, and happy rapidly increased, indicating that the government invested considerable human and material resources at any cost. Therefore, the public was able to observe the determination and effectiveness of the country's antipandemic efforts. In addition, with the arrival of the inflection point during the third phase of the pandemic, the mass distribution of COVID-19 vaccines increased public confidence

XSL•FO
RenderX

that the pandemic could be prevented and controlled and that production, work, and prospects could resume after the pandemic.

To visually present the words occurring at a high frequency for each emotion, we used a word cloud library to generate cloud maps of 70 words for 7 emotions corresponding to positive and negative emotions in the 3 stages, as shown in Table 6. At the same time, high-frequency words were combined with the corresponding current affairs hot spots to explore the differences in and connections of different semantic features in different sentiment classifications.

Figure 4. Geographical distribution of posts with negative sentiment by Chinese microblog users from July 2020 to June 2021: (A) phase 1, (B), phase 2, (C) phase 3.

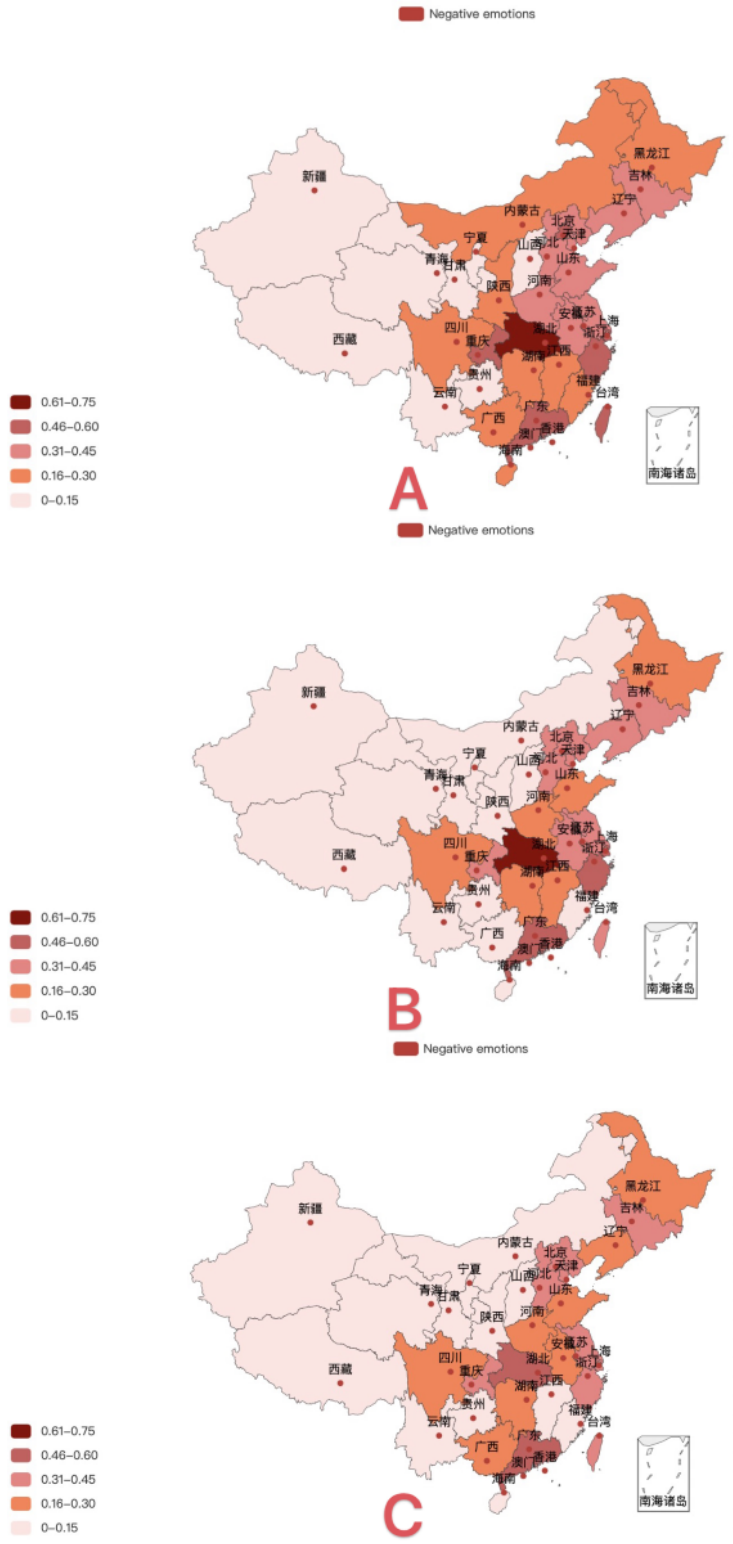

XSL•FO
RenderX

Figure 5. Bubble plot of the temporal trends in 7 sentiments based on microblog comments during the COVID-19 pandemic in China from July 2020 to June 2021.

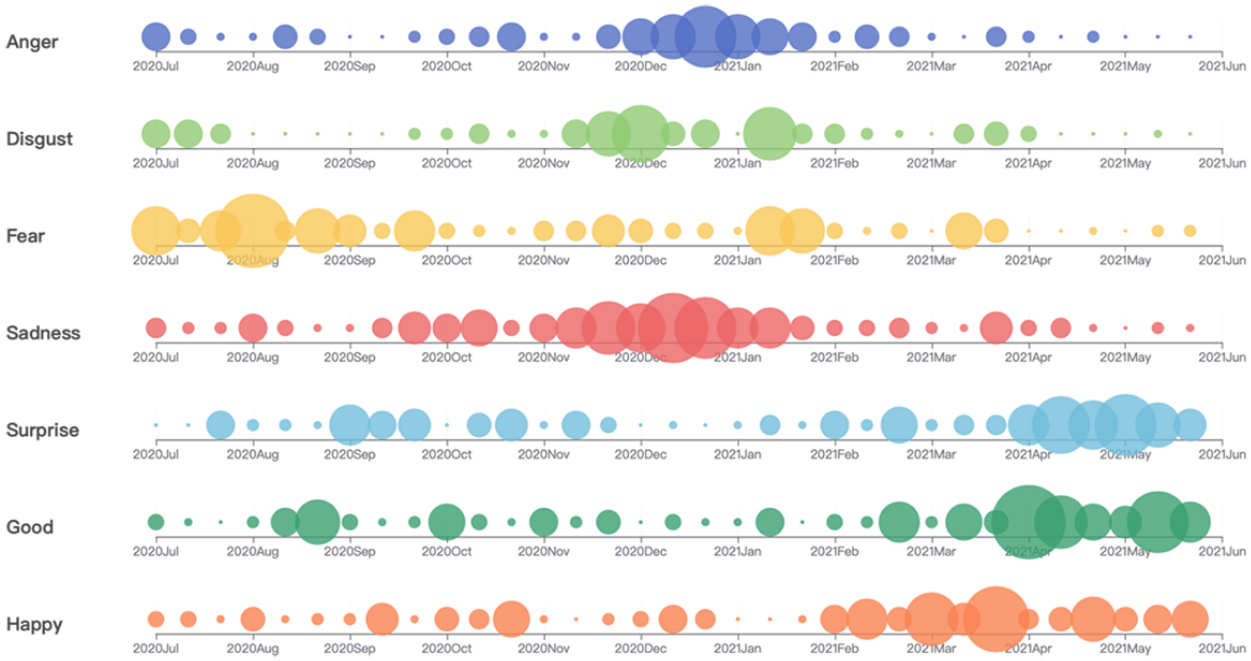

XSL•FO
RenderX

Table 6. The 7 emotional keywords based on COVID-19 pandemic topics found in microblog comment data in China from July 2020 to June 2021.

Topic and keywords	Number of times used
Topic 1: happy	
Sanguine	474
Hope	399
Happiness	307
Relieved	301
Believe	268
Pray	218
Topic 2: good	
Health	535
Security	490
Safety	462
Recovery	391
Bless	307
Great	274
Topic 3: surprise	
Unforeseen	367
Vaccine	350
Diagnose	329
Outbreak	274
Pathogen	241
Cure	207
Topic 4: sadness	
Epidemic	728
Collapse	597
Hopelessness	582
Despair	487
Negative	424
Chill	278
Topic 5: fear	
Virus	624
Infect	619
Danger	481
Trepidation	447
Horror	410
Helpless	368
Topic 6: disgust	
Rumor	394
Oppose	350
Keep away	311
Shirk	263
Back off	231
Slackness	196

XSL•FO
RenderX

Topic and keywords	Number of times used
Topic 7: anger	
Hurt	368
Indignation	359
Conceal	301
Forbid	248
Malice	217
Ferocity	145

In the initial stage of the pandemic, the words occurring at a high frequency included “hopeless,” “danger,” “infect,” and other words as well as “negative,” “rumor,” “helpless,” and many other words that express emotions. In the face of the sudden COVID-19 pandemic, the public was in an extremely unstable emotional state.

In the second stage of the pandemic, the words occurring at a high frequency included “keep away,” “pandemic,” and “pathogen.” Microblog topics focused on orderly prevention and control measures and a joint fight against the pandemic. The positive sentiment of netizens gradually increased, turning from fear to solidarity and cooperation.

In the third stage of the pandemic, words such as “vaccine,” “recovery,” “hope,” and “sanguine” appeared frequently on the microblog. Most people were full of confidence and believed that China could win the battle against the pandemic, expressing positive feelings of positivity, unity, and hope for the future.

Model Performance Analysis

To verify that the proposed model and method had high accuracy, we calculated accuracy, precision, recall, and the F_1-score, as shown in Equations 12-15:

The words in the text were converted into a word frequency matrix, and the TF-IDF weight of each word was counted. The TF-IDF matrix was then extracted, and its weight was calculated. Finally, the effectiveness of the model was evaluated by extracting parameters (precision, recall, accuracy) from the confusion matrix, as shown in Figure 6.

First, the word frequency matrix was generated using the text-based data after Chinese word segmentation and data cleaning. The *CountVectorizer* class was called to calculate the word frequency matrix, and the generated matrix was *X*. Second, the *TfidfTransformer* class was called to calculate the TF-IDF value of the term frequency matrix *X*, and the *Weight* matrix was obtained. The *Sklearn* machine learning package was then called to perform the classification operation, the *fit ()* function was called to train, and the predicted class labels were assigned to the *pre* array. Finally, we called the *PCA ()* function of *Sklearn* to reduce these features to 2 dimensions corresponding to the *X* and *Y* axes, from which we could evaluate the algorithm.

Based on the study of the traditional double classification confusion matrix, we divided the confusion matrix into 7 emotions according to the emotional characteristics. Very few data points were misclassified from the matrix by the model. Moreover, the performance of the model in terms of accuracy, recall, and the F_1-score of the weighted average of the 7 emotions “anger,” “disgust,” “fear,” “sad,” “surprise,” “good,” and “happy” was good, as shown in Table 7.

XSL•FO
RenderX

Figure 6. Multiclass confusion matrix for sentiment scoring based on COVID-19 pandemic-related topics in microblog comments in China from July 2020 to June 2021.

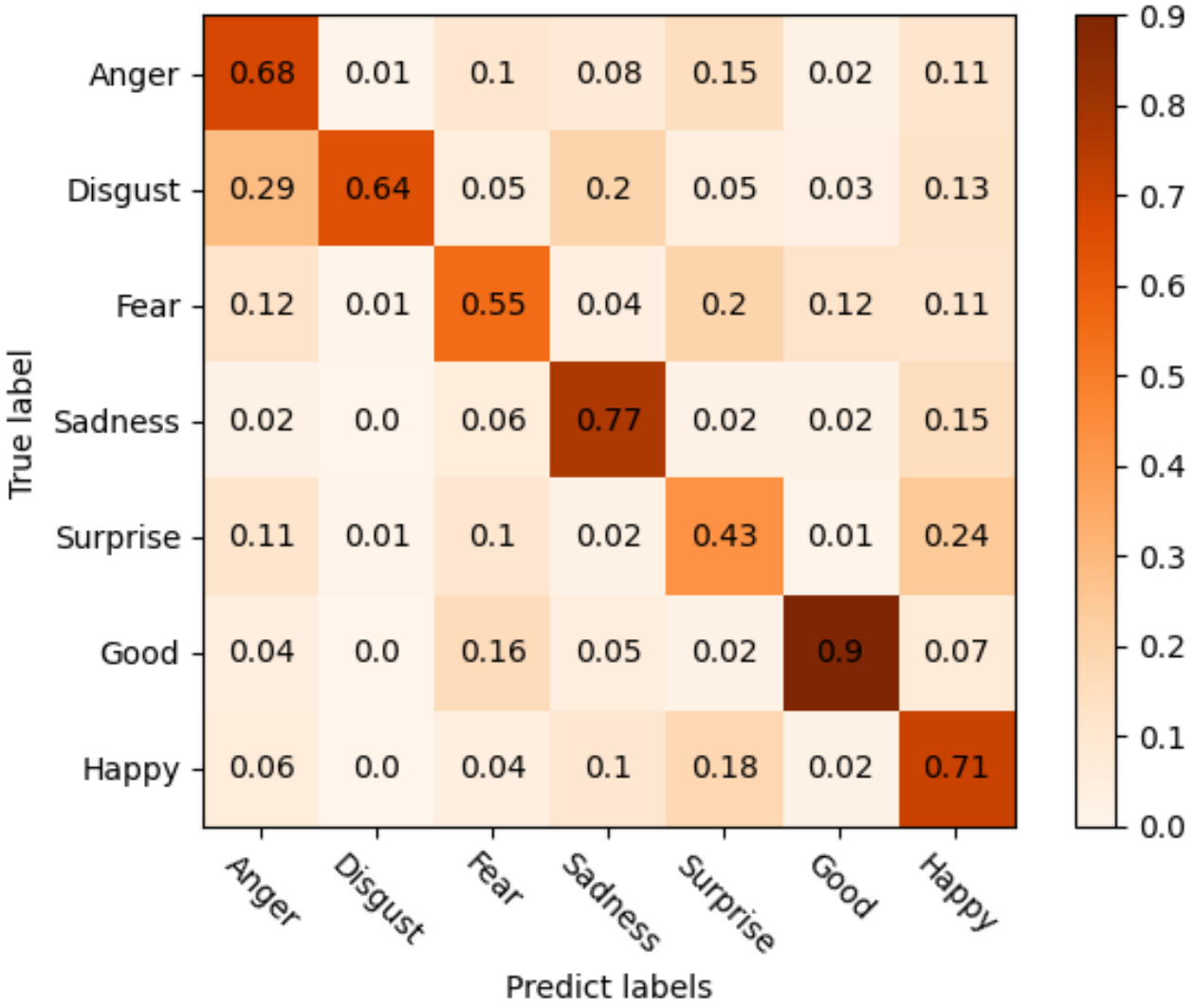

Table 7. Performance analysis of a machine learning logistic regression (ML-LR) language processing model fusing a sparse matrix based on Chinese COVID-19 pandemic topics from July 2020 to June 2021.

Measures	Precision	Recall	F_1-score	Support
Negative	0.8447	0.9525	0.9189	175
Positive	0.8158	0.8961	0.8344	428
Accuracy	N/A[a]	N/A	0.8581	603
Macro average	0.8579	0.6138	0.6440	603
Weighted average	0.7982	0.7181	0.7811	603

[a]N/A: not applicable.

Discussion

Principal Findings

This study constructed a multidimensional analysis model of public opinion on social media in China during the COVID-19 pandemic from the perspective of spatiotemporal correlations. We proposed an ML-LR model with a sparse matrix to analyze the evolutionary process of public opinion and interpret the dynamic relationship and multidimensional emotional characteristics of public opinion in different spatiotemporal environments. The results show that, due to the different trends in the pandemic situation and prevention and control efforts in different regions, there were differences in the emotional characteristics of public opinion. The amount of public opinion data at many different levels was similar in the temporal and spatial distributions, and the amount of public opinion data was positively correlated with the number of new cases. With the

XSL•FO
RenderX

rapid spread of the COVID-19 pandemic, the monthly amount of public opinion data increased. As the COVID-19 pandemic was gradually controlled, the amount of monthly public opinion data showed a downward "zigzag" trend. The spatial distribution of the amount of public opinion data was positively correlated with the distribution of COVID-19 spread, and the provinces with more public opinion data were mostly those areas with a more serious COVID-19 situation. This study can provide theoretical support and a practical reference for government and public health safety departments to deal with public health emergencies.

Comparison With Prior Work

In traditional research of public opinion on social media, the topic model method is often combined with a text clustering algorithm [12,23,25], but this approach is limited by the problem of insufficiently labeled data, making it difficult to reflect the changes in public mood. This paper proposes an ML-LR model with a fusion coefficient matrix to overcome these problems. In addition, social media data embedded with geographical location information provides valuable evaluation indicators for the study of public opinion characteristics, but the time span and geographical location range selected by traditional research of public opinion on social media are of limited size or are not representative [4,17]. This study captures the dynamic changes in public sentiment characteristics through multidimensional, spatiotemporal analysis of public opinion.

Limitations and Future Work

We will conduct further research regarding 2 aspects in the future. On one hand, when analyzing public sentiment during a pandemic, user comments may come from multiple social media platforms [25,26]. Subsequent research will consider adding data from other social media platforms, strive to include a more comprehensive user group, and describe more appropriate regional characteristics of public opinion. On the other hand, considering that the spread of the pandemic has certain geospatial heterogeneity under the joint influence of geographical proximity, transportation network, and pandemic prevention measures, taking provincial administrative regions as geospatial measurement units has certain limitations [27-29].

In future research, we will discuss how to obtain a multiscale visual analysis unit of the pandemic and public opinion according to the superposition of the transmission mode of the COVID-19 pandemic and multiscale geographic space. We will also mine the process of propagation during the pandemic from a more granular scale in terms of time and space.

Conclusions

The analytic method of public opinion on social media proposed in this study can effectively reflect the characteristics of public opinion in different regions and different periods of time. It can also provide theoretical support and a practical reference for the analysis of public opinion in major public health events, as well as provide correct guidance for government departments and effective control of the propagation of network public opinion.

Acknowledgments

This work was supported by the National Social Science Fund of China (No. 23BGL306，20BTQ067); the Basic and Public Welfare Research Project of Zhejiang Province, China (No. LGJ21G010001); the Soft Science Research Program of Zhejiang, China (No. 2022C35018); Key R&D Project in Zhejiang Province of China (No. 2022C01005, No. 2022C01144); and Major Project of Key Research Base of Humanities and Social Sciences of Ministry of Education (No. 22JJD790082).

Data Availability

The data sets containing confirmed COVID-19 cases and microblog comments are publicly available on the official websites of the Chinese Center for Disease Control and Prevention and the Sina microblog.

The coded data and filtered files are made available on request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Visual analysis of the COVID-19 topic modeling results based on microblog comments from July 2020 to June 2021 in China.
[PNG File , 1905 KB - jmir_v26i1e47508_app1.png]

References

1. Zhuang M, Li Y, Tan X, Xing L, Lu X. Analysis of public opinion evolution of COVID-19 based on LDA-ARMA hybrid model. Complex Intell Systems 2021 Sep 04;7(6):3165-3178 [FREE Full text] [doi: 10.1007/s40747-021-00514-7] [Medline: 34777976]

XSL•FO
RenderX

2. Priyadarshini I, Mohanty P, Kumar R, Sharma R, Puri V, Singh PK. A study on the sentiments and psychology of twitter users during COVID-19 lockdown period. Multimed Tools Appl 2022 Jun 14;81(19):27009-27031 [FREE Full text] [doi: 10.1007/s11042-021-11004-w] [Medline: 34149302]
3. Lopez CE, Gallemore C. An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. Soc Netw Anal Min 2021 Oct 20;11(1):102 [FREE Full text] [doi: 10.1007/s13278-021-00825-0] [Medline: 34697560]
4. Zhu B, Zheng X, Liu H, Li J, Wang P. Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. Chaos Solitons Fractals 2020 Nov;140:110123 [FREE Full text] [doi: 10.1016/j.chaos.2020.110123] [Medline: 32834635]
5. Lu Z. Analysis model of college students' mental health based on online community topic mining and emotion analysis in novel coronavirus epidemic situation. Front Public Health 2022 Sep 13;10:1000313 [FREE Full text] [doi: 10.3389/fpubh.2022.1000313] [Medline: 36187685]
6. Ogbuokiri B, Ahmadi A, Bragazzi NL, Movahedi Nia Z, Mellado B, Wu J, et al. Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts. Front Public Health 2022 Aug 12;10:987376 [FREE Full text] [doi: 10.3389/fpubh.2022.987376] [Medline: 36033735]
7. Ahmad W, Wang B, Xu H, Xu M, Zeng Z. Topics, sentiments, and emotions triggered by COVID-19-related tweets from Iran and Turkey official news agencies. SN Comput Sci 2021 Jul 29;2(5):394 [FREE Full text] [doi: 10.1007/s42979-021-00789-0] [Medline: 34341778]
8. Sun K, Wang H, Zhang J. The impact factors of social media users' forwarding behavior of COVID-19 vaccine topic: Based on empirical analysis of Chinese Weibo users. Front Public Health 2022;10:871722 [FREE Full text] [doi: 10.3389/fpubh.2022.871722] [Medline: 36187642]
9. Albahli S. Twitter sentiment analysis: An Arabic text mining approach based on COVID-19. Front Public Health 2022 Oct 10;10:966779 [FREE Full text] [doi: 10.3389/fpubh.2022.966779] [Medline: 36299761]
10. Sarirete A. A bibliometric analysis of COVID-19 vaccines and sentiment analysis. Procedia Comput Sci 2021;194:280-287 [FREE Full text] [doi: 10.1016/j.procs.2021.10.083] [Medline: 35013686]
11. Elsaka T, Afyouni I, Hashem I, Al Aghbari Z. Spatio-temporal sentiment mining of COVID-19 Arabic social media. IJGI 2022 Sep 02;11(9):476. [doi: 10.3390/ijgi11090476]
12. Johnson AK, Bhaumik R, Nandi D, Roy A, Mehta SD. Sexually transmitted disease-related Reddit posts during the COVID-19 pandemic: latent Dirichlet allocation analysis. J Med Internet Res 2022 Oct 31;24(10):e37258 [FREE Full text] [doi: 10.2196/37258] [Medline: 36219757]
13. Box-Steffensmeier JM, Moses L. Meaningful messaging: Sentiment in elite social media communication with the public on the COVID-19 pandemic. Sci Adv 2021 Jul 16;7(29):1 [FREE Full text] [doi: 10.1126/sciadv.abg2898] [Medline: 34261655]
14. Liu J, Liu L, Tu Y, Li S, Li Z. Multi-stage Internet public opinion risk grading analysis of public health emergencies: An empirical study on Microblog in COVID-19. Inf Process Manag 2022 Jan;59(1):102796 [FREE Full text] [doi: 10.1016/j.ipm.2021.102796] [Medline: 34744256]
15. Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. Vaccine 2021 Sep 15;39(39):5499-5505 [FREE Full text] [doi: 10.1016/j.vaccine.2021.08.058] [Medline: 34452774]
16. Yousefinaghani S, Dara R, Mubareka S, Papadopoulos A, Sharif S. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. Int J Infect Dis 2021 Jul;108:256-262 [FREE Full text] [doi: 10.1016/j.ijid.2021.05.059] [Medline: 34052407]
17. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with Twitter data in the United States: spatiotemporal perspective. J Med Internet Res 2021 Sep 10;23(9):e30854 [FREE Full text] [doi: 10.2196/30854] [Medline: 34346888]
18. Zhang X, Yang Q, Albaradei S, Lyu X, Alamro H, Salhi A, et al. Rise and fall of the global conversation and shifting sentiments during the COVID-19 pandemic. Humanit Soc Sci Commun 2021 May 17;8(1):1. [doi: 10.1057/s41599-021-00798-7]
19. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021 Feb 10;23(2):e25431 [FREE Full text] [doi: 10.2196/25431] [Medline: 33497352]
20. Storey VC, O'Leary DE. Text analysis of evolving emotions and sentiments in COVID-19 Twitter communication. Cognit Comput 2022 Jul 28:1-24 [FREE Full text] [doi: 10.1007/s12559-022-10025-3] [Medline: 35915743]
21. Alhashmi SM, Khedr AM, Arif I, El Bannany M. Using a hybrid-classification method to analyze Twitter data during critical events. IEEE Access 2021;9:141023-141035. [doi: 10.1109/access.2021.3119063]
22. Gao H, Guo D, Wu J, Li L. Weibo users' emotion and sentiment orientation in traditional Chinese medicine (TCM) during the COVID-19 pandemic. Disaster Med Public Health Prep 2022 Oct;16(5):1835-1838 [FREE Full text] [doi: 10.1017/dmp.2021.259] [Medline: 34369351]
23. Xie R, Chu SKW, Chiu DKW, Wang Y. Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis. Data Inf Manag 2021 Jan 01;5(1):86-99 [FREE Full text] [doi: 10.2478/dim-2020-0023] [Medline: 35402850]

XSL•FO
RenderX

24. Xie Q, Zhang X, Ding Y, Song M. Monolingual and multilingual topic analysis using LDA and BERT embeddings. Journal of Informetrics 2020 Aug;14(3):101055. [doi: 10.1016/j.joi.2020.101055]
25. Ghasiya P, Okamura K. Investigating COVID-19 news across four nations: a topic modeling and sentiment analysis approach. IEEE Access 2021;9:36645-36656. [doi: 10.1109/access.2021.3062875]
26. Hou K, Hou T, Cai L. Public attention about COVID-19 on social media: An investigation based on data mining and text analysis. Pers Individ Dif 2021 Jun;175:110701 [FREE Full text] [doi: 10.1016/j.paid.2021.110701] [Medline: 33536695]
27. Rahman MM, Khan NI, Sarker IH, Ahmed M, Islam MN. Leveraging machine learning to analyze sentiment from COVID-19 tweets: A global perspective. Eng Rep 2022 Sep 18;5(3):e12572 [FREE Full text] [doi: 10.1002/eng2.12572] [Medline: 36247344]
28. Naseem U, Razzak I, Khushi M, Eklund PW, Kim J. COVIDSenti: a large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE Trans. Comput. Soc. Syst 2021 Aug;8(4):1003-1015. [doi: 10.1109/tcss.2021.3051189]
29. Guo F, Zhou W, Lu Q, Zhang C. Path extension similarity link prediction method based on matrix algebra in directed networks. Computer Communications 2022 Apr;187:83-92. [doi: 10.1016/j.comcom.2022.02.002]

Abbreviations

API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
BTM: biterm topic model
LDA: latent Dirichlet allocation
ML-LR: machine learning logistic regression
TF-IDF: term frequency-inverse document frequency

Edited by A Mavragani; submitted 23.03.23; peer-reviewed by H Mondal, O Cohen, H Liu, Q Hongliang; comments to author 20.08.23; revised version received 09.09.23; accepted 20.12.23; published 31.01.24.

Please cite as:
Guo F, Liu Z, Lu Q, Ji S, Zhang C
Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media
J Med Internet Res 2024;26:e47508
URL: https://www.jmir.org/2024/1/e47508
doi:10.2196/47508
PMID:38294856

©Feipeng Guo, Zixiang Liu, Qibei Lu, Shaobo Ji, Chen Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceptions of the Agency and Responsibility of the NHS COVID-19 App on Twitter: Critical Discourse Analysis

Dan Heaton[1], BA, MA; Elena Nichele[1,2], PhD; Jérémie Clos[1], PhD; Joel E Fischer[1], PhD

[1]School of Computer Science, University of Nottingham, Nottingham, United Kingdom

[2]Lincoln International Business School, University of Lincoln, United Kingdom

Corresponding Author:
Dan Heaton, BA, MA
School of Computer Science
University of Nottingham
Wollaton Road
Nottingham, NG8 1BB
United Kingdom
Phone: 44 115 951 5151
Email: daniel.heaton@nottingham.ac.uk

Abstract

Background: Since September 2020, the National Health Service (NHS) COVID-19 contact-tracing app has been used to mitigate the spread of COVID-19 in the United Kingdom. Since its launch, this app has been a part of the discussion regarding the perceived social agency of decision-making algorithms. On the social media website Twitter, a plethora of views about the app have been found but only analyzed for sentiment and topic trajectories thus far, leaving the perceived social agency of the app underexplored.

Objective: We aimed to examine the discussion of social agency in social media public discourse regarding algorithm-operated decisions, particularly when the artificial intelligence agency responsible for specific information systems is not openly disclosed in an example such as the COVID-19 contact-tracing app. To do this, we analyzed the presentation of the NHS COVID-19 App on Twitter, focusing on the portrayal of social agency and the impact of its deployment on society. We also aimed to discover what the presentation of social agents communicates about the perceived responsibility of the app.

Methods: Using corpus linguistics and critical discourse analysis, underpinned by social actor representation, we used the link between grammatical and social agency and analyzed a corpus of 118,316 tweets from September 2020 to July 2021 to see whether the app was portrayed as a social actor.

Results: We found that active presentations of the app—seen mainly through personalization and agency metaphor—dominated the discourse. The app was presented as a social actor in 96% of the cases considered and grew in proportion to passive presentations over time. These active presentations showed the app to be a social actor in 5 main ways: informing, instructing, providing permission, disrupting, and functioning. We found a small number of occasions on which the app was presented passively through backgrounding and exclusion.

Conclusions: Twitter users presented the NHS COVID-19 App as an active social actor with a clear sense of social agency. The study also revealed that Twitter users perceived the app as responsible for their welfare, particularly when it provided instructions or permission, and this perception remained consistent throughout the discourse, particularly during significant events. Overall, this study contributes to understanding how social agency is discussed in social media discourse related to algorithmic-operated decisions This research offers valuable insights into public perceptions of decision-making digital contact-tracing health care technologies and their perceptions on the web, which, even in a postpandemic world, may shed light on how the public might respond to forthcoming interventions.

(J Med Internet Res 2024;26:e50388) doi:10.2196/50388

KEYWORDS

COVID-19; information system; automated decisions; agency metaphor; corpus linguistics; decision-making algorithm; transitivity

XSL•FO
RenderX

Introduction

Background

The agency of automated decision-making algorithms is a long-standing debate in academic research [1-3]. Decision-making algorithms can mitigate human errors or inaccuracies [4-7]. However, when they operate decisions in lieu of individuals, the algorithms can be seen to develop a social agency and be perceived as having humanlike characteristics [8,9]. When decision-making algorithms do not perform their assigned tasks as expected, investigating algorithmic agency can avoid additional problems, such as reinforcing bias or outcomes and undermining trust in automated decision-making systems [10-13].

An exemplary decision-making algorithm that has had significant societal impact is the National Health Service (NHS) COVID-19 app [14,15], which was used to mitigate the spread of COVID-19 in the United Kingdom by tracing contact with infected individuals and notifying people to self-isolate. Launched in September 2020, that is, 6 months after COVID-19 began to spread in the United Kingdom, this app gained public attention due to an array of issues and concerns [16-18], with social media websites such as Twitter voicing the views of many users. Several studies have investigated the sociological and epidemiological impact of the app [15,19-21], yet a gap persists with regard to how the app has been discussed on Twitter, specifically.

Agency expressed on social media can be investigated in multiple ways, such as through interview or observation [22,23]. We purposefully used corpus linguistics (CL)–informed and critical discourse analysis (CDA)–informed approaches to examine the relationship between grammatical agency and social agency [24]. Grammatical agency, or transitivity, can show whether an entity is presented actively performing an action or passively having an action performed onto them [25]. Deconstructing the agency of decision-making algorithms in the discourse can shed light on the perceived power relations between entities [26] and how these can ultimately indicate social actors in discourses [27].

To address the lack of examination of grammatical agency and transitivity in social media discourse, we uncovered the social agency of decision-making algorithms by examining tweets mentioning the NHS COVID-19 app. Specifically, we contributed to answering this overall question: How is social agency discussed in social media public discourse dealing with algorithmic-operated decisions when the artificial intelligence (AI) agency behind specific information systems is not openly disclosed? We added to the discussion around this answer by focusing on the following two research subquestions:

1. How is the NHS COVID-19 app presented on Twitter, especially with regard to social agency, and how society has been impacted by the deployment of this system?
2. What do these presentations of social agents communicate about the NHS COVID-19 app's responsibility to process information as perceived by the authors of the social media public discourse (ie, Twitter data) analyzed in this study?

By applying 2 methodological approaches, CL and CDA, underpinned by social actor representation (SAR) [27], we examined the use of the word "app" in context, which can be understood, as we will show, as a common grammatical subject of perceived agency. Thus, our work shows how the social agency of the app is implied or established through how users present it via grammatical constructions. Therefore, this study aimed to investigate the agency of the NHS COVID-19 app and how it was perceived by the public via its presentation on Twitter. Focusing on the relationship between grammatical and social agency and its impact on society, this study sheds light on the responsibility and blame attributed to the app in processing health care–related information.

Prior Work

To see whether Twitter users perceive the app to be a social actor, in its own right, or whether it is passively controlled, by the algorithmic-based AI system on which it operates, this section reviews the existing literature on the relationship between grammatical and social agency and agency and decision-making algorithms and gives an overview of the NHS COVID-19 app.

Grammatical and Social Agency

When defining agency, Gallagher [28] stated that it is a clear feeling of control and suggested that it impacts human consciousness. Silver et al [29] stated that a sense of agency also encompassed the responsibility felt due to actions undertaken and the effects they have. Therefore, social agency could be uncovered by examining grammatical agency [24-26].

Grammatically, Leslie [25] defined an agent as an entity with an internal source of energy through which it exerted force to carry out the activities referred to in the text. Expanding on this, Richardson et al [24] stated that agency in linguistics is often explored by examining how it is emphasized, manipulated, or concealed. As such, transitivity analysis—the examination of agency in the text—examined the use of active and passive voice or nominalization, where verbs were the word class converted to nouns. Accordingly, choices revealed the attitude and ideology of the language user or perceived agent [30]. In addition, research has shown that passive constructions tend to remove agency from the subject or dilute its impact [31]. Especially, when the subject was absent from the clause, the implied responsibility shifted [26]. Arguably, this refers to decision-making power, which was investigated in this study.

Alternatively, the agency can be conveyed through lexical choices. For instance, Morris et al [32] suggest that "people believed acceding trajectory evokes the impression of high animacy, which would be caused by enduring internal property, i.e. the volitional action" (eg, "the NASDAQ fought its way upward"). In contrast, "the descending trajectory suggests inanimacy, as a result of lack of external forces" (eg, "stocks drifted higher"). This phenomenon, called the agency metaphor, constituted the focus of this analysis, as well as transitivity, as they both communicate the capacity or potential to finalize decisions.

XSL•FO
RenderX

Agency and Decision-Making Algorithms

The last decade has witnessed an increased focus on the perceived social agency that decision-making algorithms have [1]. During this period, decision-making algorithms were deployed to solve societal problems, such as corruption, unethical practices, different interpretations, inaccuracies in assessments, and, perhaps most notably, inefficient decision-making [4]. Although the legal treatment of decision-making algorithms received significant attention in the literature [33-35], the social agency of these algorithms still seems underexplored. Nevertheless, given the increasingly crucial role that algorithms play in public life, it is important to unveil the relationship tying algorithms, agency, and autonomy together [36].

Among researchers who looked into the specifics of social agency and whether decision-making algorithms are perceived to have it, Rubel et al [36] defined several key ways in which issues of agency, autonomy, and respect for humans may be at odds with algorithmic decision-making. More specifically, algorithms may create rules that are impossible to follow and may not provide a rationale for decisions. Moreover, they may fail to provide room to appeal to outcomes, may not respect interpersonal boundaries, and may allow those deploying them to avoid repercussions in terms of finalizing decisions.

Other studies specifically examined the impact of decision-making algorithms and their perceived autonomy. For example, Lamanna and Byrne [37] focused on health care–related applications for decision-making algorithms and found that automating decision-making processes would be dehumanizing but argued that including algorithms would aid in stressful decision-making for medical professionals. However, this idea is not only applicable to medical studies. For instance, Riegler [38] investigated the impact of using decision-making algorithms to aid in stressful situations, but within the context of autonomous vehicles.

The social power that algorithms hold was investigated by Beer [11], who argued that although concerns regarding agency can be somewhat complex, algorithmic power is generally assumed to imply some sort of agency. Studies looking specifically at concerns surrounding the merging of human and algorithmic agencies [8,9] have questioned whether to make algorithms more humanlike or humans who are similar to algorithms [39].

Interestingly, humans may oppose the decision-making algorithms. According to Mahmud et al [40], the reasons behind algorithm aversion can be extracted from the existing literature and categorized as high level (societal), algorithm related, and factors concerning an individual. Within the algorithm factors category, one was the anthropomorphic presentation of an algorithm, along with its complexity, understandability, accuracy, and ability to learn. Similarly, Grange [41] indicated the use of machine learning or “black-box” techniques as a tool to automate decisions. Because humans may be unaware of how algorithms are designed to operate, it is critical that the perceptions of those impacted by algorithmic-based decisions are taken on board throughout the design, implementation, and review processes.

Despite the explanations provided, an agent may be systematically judged differently when they are perceived as artificial rather than human. For example, Feier et al [11] found that decision makers can rid themselves of guilt more easily by delegating to machines than to other people, thus showing that the availability of artificial agents could provide stronger incentives for human decision makers to delegate morally sensitive decisions. Therefore, decision-making algorithms can deflect responsibility, and thus blame, from human decision makers to artificial intelligence–operated systems.

Considering this, it may be worth questioning whether an algorithm can be presented as having the same agency as humans. In addition, the impact that this sort of presentation might have can provide complementary insights into this topic.

The NHS COVID-19 App

The NHS COVID-19 app, the contact-tracing algorithmic-based system created by Serco on behalf of the UK government to track active cases of COVID-19, has impacted the United Kingdom on multiple levels since its launch [14]. The app is available on mobile phones and uses exposure logging, as developed by Apple and Google [42]. This technology allows the app to send alerts using a randomly generated ID number when the user is close to another app user who has logged a positive COVID-19 test. Despite its scientific-based intended functionality, its users reported issues regarding backward incompatibility, incorrect alerts, and false-positive tests. Such unexpected technical problems meant that users had to self-isolate for 10 days even when the result was incorrect, with inadvertent consequences on their income and well-being [15].

Despite the UK government encouraging its adoption, the uptake of the app was less than expected at 20.9 million downloads between September and December 2021, with 1.7 million notifications being sent out in England and Wales [43]. According to Wymant et al [19], every 1% increase in the number of app downloads leads to a 0.8% to 2.3% reduction in the number of COVID-19 infections, with their findings suggesting that anywhere between 100,000 and 900,000 cases were averted because of the information inputted by users into the system. However, Mbwogge [17] claimed that a symptom-based contact-tracing system failed to meet the testing and tracing needs in the United Kingdom, which is further evidenced by the fact that cases and deaths relating to COVID-19 increased to be the highest in Europe.

Perhaps because of its technical challenges, a growing number of research projects have investigated the public attitudes toward digital contact tracing in the United Kingdom. Williams et al [44] interviewed 27 participants using web-based videoconferencing before the release of the COVID-19 app in the United Kingdom and found the response to be mixed and heavily influenced by moral reasoning. The analysis revealed 5 themes: lack of information and misconceptions surrounding COVID-19 contact-tracing apps; concerns over privacy; concerns over stigma; concerns over uptake; and contact tracing as the “greater good.” Samuel et al [45] conducted 35 semistructured qualitative interviews in April 2020, showing interviewees’ views on the potential of the app for contact tracing. The participants showcased a range of misconceptions

XSL•FO
RenderX

and worries. However, as there was no follow-up to this study, it was impossible to discover which participants would then choose to download the app once it was launched in September 2020. These insights shall inform our investigation of the impact of the NHS COVID-19 App in British society throughout the pandemic and the perceptions of this system by its (intended, actual, or former) users.

This possible evolution of attitudes toward the app was instead monitored by Dowthwaite et al [16], who surveyed 1001 UK adults and found that half of the participants had installed the app, with 60% of them claiming to comply with it on a regular basis. They also found that there were issues surrounding trust and understanding that hindered the effective adoption of the app. A follow-up analysis showed that there were statistically significant correlations between lower trust among nonusers, many aspects of the app, and the wider social and societal context [46]. A year after the app was launched, Pepper et al [2] identified 5 main themes during follow-up interview discussions: flaws in the app, usefulness and functionality affecting trust in the app, low trust in the UK government, varying degrees of trust in other stakeholders, and public disinterest. According to the study results, these factors contributed to a drop in compliance over time. Similar findings were proposed by Paucar et al [18], who stated that responsibility and trust made the app better accepted by the public. Even though these were always relevant, other factors, such as fear of infection, were contextual and time dependent. Arguably, this will be relevant when examining the presentations of social agents who tweeted about such an app system and its functionality as perceived or evaluated by its self-proclaimed users or experts.

For instance, in July 2021, when the relaxation of government restrictions led to an increased number of positive COVID-19 cases in the United Kingdom, media scrutiny of the app intensified because of the numerous notifications sent through the app [47]. As a result, this impacted the public's perception of the app and the pejorative blend "pingdemic" was coined [48]. This exemplifies the considerable impact that the deployment of the NHS COVID-19 App has had on British society and how this was reflected by the media and social media and the terminology they used.

To date, one study has examined tweets relating specifically to the app in the United Kingdom. Heaton et al [49] found, using a mix of computational linguistic tools, a general sentiment trend saw positivity detected within tweets from September to November 2020 only for tweets categorized as more negative in December 2020 and January 2021. Positive sentiment rose again in February and March 2020, dipping slightly in April, but rising again in May and June. Tweets were deemed less positive in July. Prominent topics included how the app works, obtaining the app, and the development of the app. Trust and fear were the most frequently detected emotions. However, what could not be ascertained through this research was how Twitter users specifically presented the app in terms of its agency or impact. Although it has been established that there is research interest in digital contact tracing from a sociological and epidemiological standpoint [2,16,18,20,21], a gap in the presentation of the app itself was detected, which our contribution aims to address.

Methods

Data

Data were collected with the aid of Twitter for Academic Purposes Application Programming Interface. Twitter was chosen as a data source because of the large amount of real-time data available [50]. In addition, Twitter data could be preprocessed before analysis [51], lending itself well to exploratory analysis principles [52].

Following the best practices recommended in social media research literature, we did not include any screenshots of tweets that may later identify their author, without obtaining consent. Instead, as part of the data-cleaning process, tweets were anonymized and only short extracts from tweets were reported verbatim (therefore, including typographical or grammatical inaccuracies). The project design was approved by the ethics committee of the university department. The data were pseudonymized during extraction, with a unique number generated to refer to each tweet. Stopwords were removed from the data set using Gensim, along with the removal of all long and short URLs and the indication "RT" (retweet) at the beginning of any tweet. Twitter handles that appeared within the tweets were also redacted using Gensim, for anonymity.

With permission, this study used the same data set as Heaton et al [49], as data had already been collected relating to this topic. Data extraction was performed using the Tweepy module in the Python programing language [53]. The key search criteria for this were tweets containing "@NHSCovid19App," which is the official Twitter handle for the United Kingdom's contact-tracing app and the related hashtag "#NHSCovid19App." The reason for this choice was to ensure that tweets were directly related to the experience of the contact-tracing app itself, rather than to the wider NHS Test and Trace system or the COVID-19 pandemic generally. Although key parts of the discourse may not be revealed through this search term alone, it provides a starting point for investigating the views expressed regarding the app.

In total, 180,281 tweets (1,797,052 words) were collected from September 23, 2020, the day before the app was launched in the United Kingdom, to July 31, 2021. Furthermore, a second data set was collected using the search term "pingdemic" to capture relevant tweets relating to the surge in self-isolation notifications in July 2021. This data set contained 36,022 tweets (831,579 words). Subsequently, tweets were condensed to remove advertisements from the data set, resulting in a final corpus of 118,316 tweets over an 11-month period. The data were sourced from the United Kingdom, and only tweets in English were selected. Therefore, the analysis investigated views expressed only in English.

Ethical Considerations

Despite these advantages, complex ethical considerations are imperative when scraping Twitter data for analysis. Although tweets are public (by default), Twitter "data" is not intentionally provided by users for the purposes of research, yet gaining

XSL•FO
RenderX

explicit informed consent to use their tweets for research is practically infeasible [54]. To align with the best practices in social media research, we anonymized tweets during data cleaning and shared only concise verbatim excerpts. The project design was approved by the ethics committee of School of Computer Science at the University of Nottingham (approval number CS-2020-R33).

Corpus Linguistics

CL was deemed as a suitable approach to answer our research question because of its ability to analyze large data sets, using specialized pieces of software to uncover patterns' relative efficiency [55]. CL takes the idea of further investigating the corpus (ie, the body of data collected to analyze) through a multitude of different language-focused perspectives. For example, CL can facilitate diachronic comparisons across corpora by focusing on lexical use [56]. Because of its capacity to identify language patterns in data sets containing hundreds of words [57-59], CL has been frequently deployed to analyze social media [60-62]. We also chose CL as it allows for the comparison of multiple corpora, identifying trends and patterns that distinguish multiple data sets. Thus, CL is particularly helpful when comparing data from different periods, as in this study.

With the aid of CL-computerized tools, we focused on collocation, that is, the co-occurrence of 2 or more words within a predefined word span [55]. When using frequency as the sole measure, Baker [63] stated that it might not be possible to verify whether a co-occurrence is a true reflection of a semantic relationship, that is, a connection based on word meaning, or whether chance played a part. Instead, statistical significance measures, such as LogDice or Log Likelihood, are useful indicators of lexical and grammatical associations between textual elements as well as themes [64]. In this sense, concordances help identify collocations, as they can show how adjacent or in close vicinity the related words are together [65].

The CL software used to undertake this analysis was The Sketch Engine [66], which was chosen for practical and analytic reasons. Indeed, it is freely available to many academics, it allows the upload of ad hoc corpora, and it provides The Sketch Engine a series of reference corpora that can be used for comparisons.

The analyses performed in this study were conducted in different stages. First, keyword analysis was used to distinguish keywords in the discourse using the embedded English Web 2020 (enTenTen20) [67] as the term for comparison. EnTenTen20 has over 36 billion words of specifically internet texts, including social media, and so acts as a suitable reference corpus. In addition, keyness scores were generated by comparing the frequency of words in the target corpus to the frequency of words in the reference corpus. This allowed us to examine the key characteristics of the corpus compiled, proving an overview of the tweets collected for the analysis.

Second, concordance lines featuring "app," as a potential social actor, were examined to prompt the collocation analysis, which was accomplished following different directions. To ascertain active constructions, our collocation criteria were "app" and one verb to the right (R1). To ascertain passive constructions, our search criteria was "by the app" and one verb to the left (L1). Passive constructions, such as "the app was used by many" and "the app has not been created" were included in the examination of verbs to the right (R1). Specifically, as they are passive constructions where the app is the object, they were removed from the active constructions and added to the passive construction category manually.

LogDice was considered as the statistical measure of collocational strength. LogDice compares the observed co-occurrence of words with their expected co-occurrence based on their individual frequencies. A high LogDice score indicates a strong association between 2 words, suggesting that they often appear together, whereas a low score implies a weaker association. LogDice was included as it not only measures the statistical significance of a collocation but also factors in the size of the subcorpus, making comparisons between subcorpora of different sizes easier.

To take advantage of this capacity, we split the corpus into 5 subcorpora that reflected the key moments in the evolution of the pandemic in the United Kingdom in chronological order:

- Period 1: app launch (September 2020)
- Period 2: early months (October-December 2020)
- Period 3: second national lockdown (January-February 2021)
- Period 4: later months (March-June 2021)
- Period 5: "Pingdemic" (July 2021)

We report the strongest collocates for each period, ranked by LogDice score.

We used a minimum threshold of 3 occurrences for the collocate to be significant enough to report; hence, the variation in collocates at each time.

Critical Discourse Analysis

Finally, from this, CDA was applied to examine agency and social action as expressed in the (concordance) lines, where "app" appeared as a keyword in context. In this study, CDA was used complementarily to The Sketch Engine CL analysis tool [66] to pinpoint different perspectives and meaning shades, referring to the various subtle nuances or different shades of meaning that a word or phrase can convey [68]. Therefore, these approaches were deemed especially effective together, as accomplished in studies [69-71] with similar purposes to this analysis.

CDA is an interpretative qualitative approach to text analysis that draws on related theoretical frameworks [72-74]. Overall, it can be used as a tool to better understand meanings implied through textual context [75-77]. Several studies have demonstrated the benefits of using CDA on Twitter discourses [78-80]. This fits with our data-driven approach to analysis in an attempt to answer our research questions regarding the presentation of the app and its impact on British society.

To successfully analyze the relationship between grammatical and social agency, we underpin our use of CDA within the theoretical framework of SAR, which is drawn from Social Action Theory. According to Social Action Theory, people

XSL•FO
RenderX

create society, institutions, and structures [81]; hence, examining social actions can provide an explanation for human behavior and societal change [82], including the app users' perception, which we intend to focus on in this study.

More specifically, SAR examines how grammatical structures convey social agency; for example, active or passive constructions and transitivity structures can be used to communicate who social actors are in a discourse of interest [27]. Similarly, syntactical features, sentence structures, and verbs within tweets give us an indication of how users perceive the app to be responsible for processing their information.

Multiple concepts can be considered as key when it comes to sociosemantic categories for analyzing social actors [27]. Among them, the removal of grammatical agents is called excluding. Conversely, when clues are left as to who the agent is, this is called backgrounding. In addition, actors can be personalized through word choices pertaining to the semantic nature of being "human" or impersonalized. Moreover, examining agency metaphors, previously outlined in this paper [32], can signal further personification of nonhuman entities. All of these aspects are especially relevant to this study because they imply humanlike perception, possibly indicating whether responsibility for consequences is attributed.

At the same time, social actors could be a group of people (genericized) or represented as single individuals (specified). In this sense, indetermination occurs when social actors are not specified (such as "someone"), whereas determination occurs when their identity is made known. All of these representation structures play a role in indicating the social and power dynamics within discourse, as shown in other Twitter case studies that used CL, CDA, and SAR [83-85]. In this analysis, tweets were examined in terms of the determination and indetermination they carry, as to how they deflect or attribute responsibility.

By analyzing all these characteristics in this Twitter discourse, we intended to identify common presentations of the NHS COVID-19 App, ultimately displaying how power relations are communicated in real-life data dealing with algorithmic-operated decisions, even when the mechanisms are not fully clear. After establishing these, we identified similar semantically related thematic groups (as seen previously in Razis et al [86] and Kitishat et al [87]) that could aid in the analysis of the presentation and perceptions of the app over time.

Results

Keyword Analysis

Table 1 shows the top 10 words with the highest keyness score when compared with EnTenTen, 2020 (all scores to 2 decimal places). The word with the highest keyness score was app, which supported our initial thinking that this would play a dominant role in the discourse because all tweets collected for this study included the expression "NHSCovid9App" to intentionally focus on discussions revolving around this system. With this in mind, we proceeded with the analysis as planned.

Table 1. The top 10 words with the highest keyness score.

Item	Relative frequency (per million)		Score
	Focus corpus	Reference corpus	
app	13,387.27	63.94	82.27
nhs	5427.78	15.62	47.81
download	3597.18	61.05	22.96
covid	1846.74	4.39	18.65
serco	1556.30	0.26	16.52
trace	1613.03	94.14	14.84
isolate	1355.90	4.08	13.99
test	3302.05	159.01	13.14
qr	1236.71	2.22	13.08
downloaded	1351.74	11.42	13.03

Presentation of the App: Timeline Overview

This section presents a timeline of the changes in the grammatical presentation of the app along with the potential social implications that this had. As part of these results, we found that "be" and "have" were frequently occurring collocates of the "app." Upon manual inspection of the tweets containing them, most were found to be auxiliary verbs. Whenever this was the case, they were considered as multiword expressions and analyzed on the basis of their overall meaning because they conveyed links between agency and responsibility together. First, we looked at the frequency of active and passive verbal constructions including "app." This overview is shown in Table 2, where the information system features actively in 97% of clauses. However, active and passive constructions alone do not necessarily provide a full account of how the app is presented in the discourse. For example, the app could be presented actively, yet it could carry limited social agency (eg, "you self isolate when the app pings you even though you don't have to"). To avoid misinterpretation, we combined CL and CDA.

XSL•FO
RenderX

Table 2. Frequency of active and passive presentation of app.

Period	Active	Passive	Total
September 2020	3700	138	3838
October-November 2020	1935	65	2000
January-February 2021	1396	36	1432
March-June 2021	740	5	745
July 2021	6108	122	6230

Period 1: App Launch (September 2020)

The launch month of the app saw 17,759 instances of the word "app," which was the highest engagement recorded across any month included in the corpus.

Active Presentations in September 2020

In September 2020, there were 3700 instances of active app presentations. The strongest 20 collocates are presented in Table 3. Many of these active presentations evaluate the app as underperforming, especially constructions containing "do" (LogDice: 9.42). Tweets stated that the app "doesn't do its job" or "doesn't work." In this sense, such instances reflected public perceptions of the app, which were frequently deemed dysfunctional.

In September itself, "app" and "say" (LogDice: 8.07) frequently co-occur in tweets which discuss the app presenting information that users struggle to understand. For example, one tweet questioned why the app "says" they "are in a medium risk area." Another tweet stated that, despite going elsewhere, the app "said" they were "still at home." Occasionally, users complained that it "says nothing." Tweet authors' use of the verb "say" suggests that the app behaved (or was perceived as behaving) like a human; hence, it is an illustration of personalization.

Another strong collocate of "app" was "tell"(LogDice: 7.90). This had semantics similar to "say." However, "tell" was mainly used to express that the app was instructing a user to self-isolate, in both actual scenarios (ie, upon entering test results into the app, the app "tells" them "to isolate and get a test") and hypothetical scenarios (eg, the app "told" them they "had to isolate even though their boss would not allow them to without symptoms"). In other instances containing "tell," users questioned the reliability of the app, for instance asking whether anyone believed "a word this app tells u?" Similarly, another user was confused about "what this app is telling [them]??" Although comparable with "say" semantically, it could be argued that the pragmatics of "tell" were different. For instance, "tell" acted imperatively when "telling" users "to stay at home." This constituted the personalization of the app. These examples could also be considered agency metaphors, as "tell" implied more volitional action as an imperative compared with "say."

Another way in which the app was presented actively was when users wrote that it "needs" something (LogDice: 7.50). In this month, users frequently tweeted about the operating system requirements for the app to function on mobile devices, whereby the app "needs ios 13.5" and "needing current ios updates or [it] won't work." The "needs" of the app not only presents it actively but also gives it humanlike characteristics, providing other examples of personalization, and hence a fuller account of the app's public image.

"Allow" also strongly collocates with "app" (LogDice: 7.35). In these instances, users discussed the function of the app and the permissions that the app granted. For example, "the app allows [them] to enter one postcode only," causing issues to people living and working in different areas. Interestingly, the user directed this grievance to the app itself, giving the impression that the app had social agency. Other occurrences of "allow" involved questions, for instance, asking whether the app will "allow" users to report themselves as testing positive, even when they are not. Another questioned if the app "allow[ed] for manual check in." These questions from Twitter users reiterated concern for the app's implied agency and could potentially be seen as additional examples of personalization, expressing public attitudes of uncertainty and worry about not being able to use the app.

"Think" is another strong collocate (LogDice: 7.24). Occurrences of this active presentation complained about the app's performance and accuracy. For example, when visiting different places, the app could "think" that users were "still at home." Another Twitter user reported deleting the app after getting a negative test back as there was no code to input the test and the app "thought" they "still had to isolate." One question was whether the app "thought" they had been "at the old venue for all that time" when they checked into a new venue after several days. Despite these open queries and concerns, the users presented the app as being able to think and act for themselves. In this case, personalization conveyed agency.

XSL•FO
RenderX

Table 3. Top 20 words ranked by collocational strength of "app" + one verb to the right (R1) in September 2020.

Rank	Collocate	Frequency	Collocate frequency	LogDice
1	do	567	23,295	9.42
2	be	1565	80,974	9.24
3	use	111	6766	8.43
4	work	108	7093	8.34
5	require	39	655	8.17
6	have	297	30,674	8.14
7	say	64	4011	8.07
8	tell	52	3357	7.9
9	fix	30	663	7.79
10	launch	25	724	7.5
11	need	47	4738	7.5
12	allow	23	815	7.35
13	know	44	5289	7.31
14	think	32	3154	7.24
15	install	21	871	7.2
16	seem	20	1221	7.03
17	develop	15	412	6.87
18	go	30	4620	6.87
19	let	19	1563	6.86
20	delete	17	1018	6.86

Passive Presentations in September 2020

The app is also presented passively on 3.59% (138/3838) of occasions this month, with collocates as shown in Table 4. These included discussions about whoever created the app. "Be" + "develop" was a multiword collocate of "app" (LogDice: 10.00). Some constructions were questions such as whether the app has "been developed by the nhs?" or the "app was developed by serco and [...] not the nhs." These tweets are examples of the backgrounding of the entities that (supposedly) created the app. Instead, this presents the app as passive, yet important in the construction, as, despite the lack of grammatical agency, the focus is still on the app. This is closely linked to "be" + "design" (LogDice: 9.4). This discussion around the app's intended function was subverted in some tweets; for example, "this app is designed to control sheeple." However, unlike the previously mentioned examples, these represented instances of exclusion and removed the agent from the construction. In these passive structures, Twitter users still discussed the app in a negative way, highlighting that the app's functionality was deemed unsatisfactory by its (self-declared) users.

Similarly, the "app" collocated strongly with "be" + "run," resulting in constructions containing the passivization of Serco. Examples included "but then the app is run by a private company" and "have heard this nhs app is run by serco?!" This again indicated that despite the passive presentation, users were still dissatisfied with the app.

Comparable instances portrayed the app passively, through a "has been" + verb construction (LogDice: 4.88) to state that the app "has been launched in england and wales after months of delay" or that "the government's app has been designed by a dog." Neither of these constructions indicated who was responsible for the launch or design of the app, thus exemplifying exclusion and using it to reiterate user dissatisfaction.

Other passive constructions, delivering a similar meaning, combined a verb and "by the app" and accounted for 44 of 138 occurrences in September 2020. "Recognize" (LogDice: 8.70) was mainly used regarding the inputted test results into the app. For instance, one user asked whether "only private tests will be recognised by the app" and another stated the simplicity of setting up code "recognised by the app." In both short extracts, the app's passive presentation removes agency and places it more with app developers. In terms of agency, "accept" (LogDice: 7.36) is similar to the data analyzed. For example, one user complained that the incorrectly formatted code was "not accepted by the app." All of these instances reflected the app's perceived lack of functionality, causing public criticism despite the passive presentation.

"Isolate" is another strong collocate of "by the app" (LogDice: 5.36). Tweet authors complained about being told to self-isolate using the app. One user, for instance, questioned liability if "notified of contact/need to isolate by the app." Thus, the app appears to be less of a focus in the structure, and agency is removed through passivization and backgrounding. Therefore,

XSL•FO
RenderX

the responsibility could possibly be transferred from the app to the user.

"Tell" (LogDice: 4.85), featured in complaints about people that they were being instructed by the app. An example was one user discussing a "person at my work" who had "just been told by the app to self-isolate and get a test." This presented the app passively and paid limited attention to it, with the "person" being the central figure, although indetermined and genericized. As in the previous case, responsibility seemed to be deflected to the user "by" the app.

In summary, in September 2020, many tweets actively presented the app, especially when uncertain about how the app functioned or could assist its users. This was mainly accomplished through personalization, portraying the app as if it were human. In these cases, passive presentations prominently discussed the development of the app and attributed it to Serco, the NHS, or the UK government, deflecting the responsibility from the app to these organizations or app users.

Table 4. Top 6 words ranked by collocational strength of "by the app" + one verb to the left (L1) in September 2020.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	recognise	5	350	8.7
2	accept	3	554	7.36
3	isolate	4	3137	5.36
4	track	3	2653	5.19
5	tell	3	3357	4.85
6	use	4	6766	4.27

Period 2: Early Months (October-December 2020)

The first 3 full months after the app was launched saw 6237 tweets using the word "app."

Active Presentations in October to December 2020

Active presentations of the app were seen 1935 of 2000 times, with the R1 collocates reported in Table 5.

"Use" (LogDice: 10.42) appeared most frequently in a duplicated tweet that had been sent from different regional NHS accounts. The text in question contained the structure "the app uses an algorithm to filter out false alarms." Therefore, the NHS promoted the app as a positive social actor in contrast to the negative presentations put forward by several members of the public, as detailed earlier.

Similarly, many of the tweets using "say" (LogDice: 8.58), released over these 3 months (October-December 2020), were comparable to those published at the time of the app launch (September 2020). Among others, one user tweeted about discrepancies between the supposed ending to their self-isolation period, stating that their app "said that [their] self isolation will be ended on 25 dec 2020 at 23.59," which was "different from what [they] have been told on text message and nhs website." Here, the app was presented as actively informing the user, which constituted another example of personalization. Interestingly, the same user states that they have been "told on the text message and nhs website," rather than be told by the message or by the website. This distinguished the app actively presented and other technological systems, appearing as vessels of information, rather than agents. These different presentations reaffirmed that the app was a social actor in this context.

"Tell" (LogDice: 8.55) was used in a similar way to "say," similar to the tweets found in September 2020. An example of this included one user tweeting that their child had "received a notification on the track & trace app telling her to self-isolate" yet only for 2 days. Another user stated that the app "tells" them their "home is medium risk" despite living in a rural area with low COVID-19 infection rates. These examples indicated that the app was providing instructions and thus had a social agency.

Another strong collocate was "have"(LogDice: 7.02). Although used as an auxiliary verb in most constructions, there were occasions when it acted as the main verb to indicate possession (or a lack of possession). For example, one tweet discussed that their relative was recovering from cancer and expressed frustration that the app had not notified them even though they had been in contact with a positive case. Accordingly, the app "has one job" to keep their relative safe. This is a clear example of personalization due to the idea that the app is able to perform a job yet responsible for the safety and welfare of their relative. Similar active presentations featured "have" as an auxiliary verb, as when a user joked that the app "has decided to turn off contact tracing," implying its autonomy and control.

When the app was presented as performing the opposite of its desired function, negation was used. A user complained that "the app has not alerted [them]" despite "living with someone who had tested positive" for the virus. Another user stated that their app "has not conducted exposure checks since 29 december." Both examples placed the agency with the app, alluding that the app was responsible for its own shortcomings. With "someone," this is an example of an indetermined construct, which further removed agency from the humans and placed it with the app.

On other occasions, where users wrote that the app "gives" them something (LogDice: 6.88), one complained that the app "gives [them] notification about people passing by [their] house," while another joked that the app gave them "a 3 day stay at home order." Another mused that the app was "giving the govt more control over our everyday lives." In all of these occurrences, the app was presented actively through personalization,

XSL•FO
RenderX

showcasing the perceived responsibility of the app for controlling users' lives.

Table 5. Top 20 word ranked by collocational strength of "app" + one verb to the right (R1) in October, November and December 2020.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	use	365	6766	10.42
2	say	70	4011	8.58
3	tell	61	3357	8.55
4	work	84	7093	8.25
5	be	703	80,974	8.12
6	do	154	23,295	7.64
7	have	129	30,674	7.02
8	show	12	1119	6.99
9	give	14	1920	6.88
10	allow	9	815	6.73
11	send	10	1134	6.72
12	install	9	871	6.7
13	seem	10	1221	6.68
14	store	15	2838	6.68
15	update	9	1443	6.43
16	develop	6	412	6.37
17	fail	6	446	6.35
18	keep	10	2083	6.34
19	crash	5	83	6.32
20	ask	8	1685	6.17

Passive Presentations in October to December 2020

When examining passive constructions, as shown in Table 6, passive presentations similar to the previous month can be seen. When focusing on "notify" (LogDice:7.87), tweets focused on hypothetical scenarios, with one tweet stating that they were not entitled to support should one be "notified by the app" as "they can't identify you" and another that questioned the legal ramifications if one was "only notified by the app" and not test and trace as a whole. These examples discussed the legal and financial implications of the app directing someone to self-isolate. In both instances, the app was not a prominent part, hence, passivization, and the central focus was on the impact rather than the app.

In contrast, when "tell" was used in passive constructions (LogDice: 5.59), many of these accounts were direct first-person narratives by app users. For example, one "got told by the app [...] to isolate for 12 days." Another explained they have "not been told by the app to isolate," even after their family member tested positive. In these cases, the authors recounted that they were provided with a service by the app, backgrounding the importance of the system in the process. Instead, these accounts tended to focus on obtaining answers from humans that the app could not provide.

Some "have" constructions were passive too. For example, one user wrote that they were at risk as the app "has not been created to include old smartphones." This passive construction implied that the app had been created by an unknown agent, thus exclusion. Although this passive construction removed some agency from the app, the fact that IT was mentioned explicitly in the tweet could still foreground the system as a social actor.

Table 6. Top 3 words ranked by collocational strength of "by the app" + one verb to the left (L1) in October, November, and December 2020.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	notify	3	385	7.87
2	isolate	5	3137	5.69
3	tell	5	3357	5.59

XSL•FO
RenderX

Period 3: Second National Lockdown (January-February 2021)

Active Presentations in January to February 2021

During this period, there were 1396 active presentations of the word "app," as shown in Table 7. "Ping" (LogDice: 7.98) was used actively to mean notify, with examples such as one user stating that "everyone knows it was your app pinging" and another writing that the app "pings you because you walk past someone in the street." These tweets suggested that the app was acting autonomously and had its own agency through personalization and, in the case of "someone," an indeterminism.

Additional instances of the app "telling" (LogDice: 7.63) recounted personal experiences and fewer reported hypothetical scenarios. Examples included one user stating that the app "tells" them they "have to isolate" from 10 days after the initial encounter date. Another question asked why the app was "telling" them "to isolate for 14 days" when they believed it was 10 days instead. Overall, the app was actively presented in these scenarios. Therefore, should someone be affected by COVID-19, the app may be more likely to be presented actively.

Hypothetical instances questioned the legitimacy of the app, such as one user hypothesizing why other individuals were self-isolating when they had no symptoms because "an app told you to." This presents the app as an implicated social actor. This could be seen to lessen the impact of the app, although presented actively, and may doubt the functionality of the system as a whole. Other active presentations that implied that the app had social agency removed the idea of instructing people to self-isolate. For instance, one author tweeted about the app "is creating a notification that has been stuck" on their screen for a long time. The idea that the app was "creating" a notification may further position it as a social actor. Instead of using the verb "notify," the author word class converted this to the noun "notification," using it in conjunction with a more personalized verb, "create." Therefore, this clearly indicated agency and placed responsibility on the app to self-regulate through the agency metaphor.

The authors using "do" (LogDice: 5.16) discussed the app's failed expectations. An example included one user writing that, despite their partner testing positive, "the so called world beating app didn't alert [them]." This active construction indicated that the app was perceived to be responsible for their safety.

Table 7. Top 18 words ranked by collocational strength of "app" + one verb to the right (R1) in January and February 2021.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	ping	7	455	7.98
2	tell	23	3357	7.63
3	cost	4	323	7.4
4	notify	4	385	7.29
5	say	18	4011	7.05
6	state	3	366	6.91
7	use	24	6766	6.77
8	work	22	7093	6.58
9	alert	7	2050	6.52
10	show	4	1119	6.38
11	be	189	80,974	6.25
12	keep	5	2083	6.01
13	seem	3	1221	5.88
14	have	43	30,674	5.5
15	need	7	4738	5.47
16	do	26	23,295	5.16
17	track	3	2653	4.98
18	store	3	2838	4.9

Passive Presentations in January to February 2021

When considering passive presentations, shown in Table 8, many tweets released in January and February 2020 were concerned with an individual being literally or hypothetically instructed by the app, for example, "ping" (LogDice: 8.90). Being "pinged by the app" was "as reliable as a handbrake on a canoe." According to another user, they had to isolate for 10 days after they "got pinged by the app." Other collocates such as "alert," "told," and "isolate" also followed similar patterns. This culminated in Twitter users potentially seeing the app as exemplifying unreliable government handling of the COVID-19 pandemic, shifting responsibility from the app to these organizations.

The app was also presented passively in conversations about its producers. For example, one tweeted that they resided and

XSL•FO
RenderX

worked in an area where the infection rate was high, yet "the app has been triggered once in its 4/5 months existence." Although this tweet presented the app passively, it placed the blame on the creators of the app, without even mentioning them, applying a reverse-exclusion strategy.

Table 8. Top 4 words ranked by collocational strength of "by the app" + one verb to the left (L1) in January and February 2021.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	ping	7	455	8.9
2	alert	3	2050	5.57
3	isolate	3	3137	4.96
4	tell	3	3357	4.86

Period 4: Later Months (March-June 2021)

Active Presentations in March to June 2021

Between March and June 2021, a total of 740 active presentations of "app" were found in the data set collected, as shown in Table 9, numerous presentations of which presented the app as a social actor.

One of the strongest collocates from these months, "provide" (LogDice: 7.22), described the app as helpful. An example of this was a tweet that stated that the app "provides anonymous information including risk alerts by postcode, a symptom checker, and test booking," which came from a devolved local NHS Twitter account. The app was presented as a social actor, supporting the idea of system confidentiality and positively evaluated. Until now, when the app had been actively portrayed, it had usually been negatively connotated. However, this was not the case for all instances of "provide," with other examples including one user that questioned why the app did not "provide update information" about local infection levels, whereas another user stated that "the app provides little to no information," thus indicating dissatisfaction with the app's performance.

Similar to most "provide" occurrences, "help" (LogDice: 5.22) was mainly seen in advertisements from devolved NHS Twitter accounts. In these cases, tweets contained constructions such as the app "helps stop the spread of the virus." Therefore, this presented the app as having a positive impact on society.

Although not a significant enough collocate to meet the minimum threshold, authors used "tell" in conjunction with "be," when discussing the app. Instances included that the app was "telling [them] 10 days from the 26th instead 20th," and another wondering how long they needed to isolate for, particularly if it was "just the 2 [days] that the app is telling [them]." Both of these examples could be categorized as a query about the lack of clarity that the app reflected as the rules about self-isolation were changing. As both constructions showed the app to be active, this not only added to the evidence of the app being presented as a social actor but also contributed to the discourse surrounding questions over the functionality of the app itself.

Authors used "have" (LogDice: 4.09) to present the system in an active way, with examples of tweets including "not only has the app failed me [...] it has created a problem for me," indicating that responsibility is attributed to the app. Another interesting presentation discussed the app as only guidance, as it "has no legal force." Here, the app is presented actively, yet the content of the structure could be argued to mitigate or remove social agency. This suggested a decrease in the system's responsibility and control.

Table 9. Top 6 words ranked by collocational strength of "app" + one verb to the right (R1) in March, April, May, and June 2021.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	provide	3	525	7.22
2	help	5	4271	5.22
3	be	68	80,974	4.78
4	work	6	7093	4.77
5	have	16	30,674	4.09
6	do	8	23,295	3.48

Passive Presentations in March to June 2021

Owing to the small number of passive presentations (5/745, 0.7%) from March to June 2021, collocation analysis would not be meaningful. However, upon manual inspection, these constructions were concerned with scenarios that did not involve the tweeting authors. For instance, one discussed a friend who "has been told by the app to stay in for 3 days." These tweets foregrounded the importance of the experience of the public by genericizing indetermining and backgrounding the app.

Period 5: "Pingdemic" (July 2021)

Active Presentations in July 2021

There were 374 active occurrences of "app" in the "NHSCovid19App" data set and a further 5734 occurrences in

XSL•FO
RenderX

the "pingdemic" data set (total 6108/6230, 98.04%). The collocations are listed in Table 10.

The strongest collocate was "disagree" (LogDice: 10.39). However, upon manual inspection, this was a headline that had been quote-tweeted multiple times. The variations in the headline read "U.K. Leaders Hail a Return to Normal; Their Phone App Disagrees" and "Britain's contact-tracing phone app disagrees, telling huge numbers of people to self-isolate." The idea that the app disagreed with powerful human entities exemplified personalization. Despite this coming from only 2 sources, the high number of shares indicates that others engaged with the idea.

Another strong collocate was "send" (LogDice: 9.38), in reference to the app sending a total of approximately 600,000 notifications to self-isolate. One tweet stated that the app "sending too many spurious notifications will reduce compliance." In this instance, the author presented the app as an active social actor because of the cause-and-effect relationship between the app and the members of the public, further implicating the app as an agent of change, showcasing it as a perceived responsible actor by the users.

The recorded resurgence of "ping" (LogDice: 8.54) in July 2021 is likely due to a new blended term for the increase in exposure notifications. In these instances, the app was presented actively, as performing actions ranging from matter-of-fact reporting ("NHS Covid app pinged 600,000 more people") to the nonsensical ("Every time a Covid app pings Boris Johnson loses one of his wingdings"). In each of these occurrences, the app was still presented as having agency and being a social actor through personalization, and hence depicted as causing frequent disruptions.

One occurrence where the app was presented as actively, "pinging" expressed disdain toward the members of the public who "self isolate when the app pings [them] even though [they] don't have to," hence suggesting that they "will blame the government for [their] own decisions." Despite the active presentation of the app, its impact as a potential social actor was mitigated through the author's sarcastic tone. Arguably, members of the public who use the app should be accountable for their own actions rather than blaming the app. Other tweets appeared to support this view, such as one stating that "it's not a pingdemic" as the app was "pinging ppl correctly," and another that detailed the app was "pinging" because "it is doing its job.". All these instances illustrate different ways in which responsibility can be attributed to entities other than the app.

"Tell" (LogDice: 9.15) once again revolved around instruction to self-isolate. One user wrote about how the app "is telling people to self-isolate" because of higher infection rates. This contrasted with other experiences, such as another user asking whether the app can "tell" them when they are "supposed to have been near an infected person." Another user wrote "the app told something like 700,000 to isolate," which resulted in allegedly instructing supermarket staff to isolate as they had mobile phones at work. This presented the app as a social actor, and perhaps as if it had a humanlike agency through personalization. Active presentation of the app was clear in these cases, as it demonstrated the system's capacity to instruct, thus having a social impact.

The app caused disruption through many other active presentations. "Wreak" (LogDice: 6.52) was used when users said the app "wreaks havoc." Similarly, "cripple" (LogDice: 5.70) featured in constructions like the "app cripples Britain." In addition, "threat" (LogDice: 5.50) featured in a tweet stating that the app "threatened to bring parts of the economy to a standstill." All these constructions presented the app as destructive and capable of creating harm, thus being responsible for social disruptions.

XSL•FO
RenderX

Table 10. Top 20 words ranked by collocational strength of "app" + one verb to the right (R1) in July 2021.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	disagree	76	96	10.39
2	send	45	451	9.38
3	tell	57	1535	9.15
4	be	845	47,226	9.14
5	have	165	12,256	8.59
6	ping	58	3351	8.54
7	do	93	6490	8.53
8	beg	19	53	8.43
9	work	36	1752	8.39
10	fail	13	478	7.57
11	delete	16	1057	7.54
12	install	10	64	7.49
13	start	13	917	7.32
14	cause	19	2259	7.28
15	alert	10	585	7.13
16	say	19	2778	7.1
17	use	12	1352	6.98
18	force	9	672	6.92
19	go	15	2327	6.91
20	design	7	223	6.86

Passive Presentations in July 2021

Passive constructions were more frequent in July (122) than previously (5). The collocational strength of these is shown in Table 11. However, due to the greater volume of tweets in this part of the discourse, this was proportionally lower than that in September 2020. "Ping" (LogDice: 8.30) showed users speaking hypothetically once again. One user questioned how society would cope "if everyone pinged by the app asked for a PCR test," while another user stated that "if you get pinged by the app you shouldn't need to self isolate." This may suggest that the focus was on the humans affected by the app rather than the app itself. This seemed to limit the system's social agency through the background. Similarly, collocates "alert," "contact," and "isolate" were found in tweets surrounding with the same idea.

Conversely, another strong collocate, "cause" (LogDice: 6.98), was used differently. Although the app was seen to "cause" damage and chaos during the "pingdemic," in active constructions, the passive presentations removed agency from the app. Examples included one user writing that "staff shortages have NOT been caused by the App" and another stating that the United Kingdom government was to blame, hence "it's not a 'pingdemic' caused by the app." Finally, another tweet built on this and criticized the media outlet The Daily Star for "adopting the right-wing press's line that the 'pingdemic' is caused by the app." This not only removed grammatical agency from the app but also mitigated its social agency by making other agencies appear more responsible.

Table 11. Top 6 words ranked by collocational strength of "by the app" + one verb to the left (L1) in July 2021.

Rank	Collocate	Frequency	Collocate frequency	logDice
1	ping	33	3351	7.02
2	alert	5	585	6.82
3	contact	6	1084	6.19
4	cause	9	2259	5.72
5	isolate	4	2063	4.68

XSL•FO
RenderX

Summary of Results

These results suggest that the app was presented in a predominantly active way (97.43% of occurrences—13,879/14,245 constructions), although some active presentations gave the app more social agency than others did. Approximately 100 participants carried less agency by mitigating activity in either verb constructions or other contextual information. This indicates that the app was presented as a social actor in approximately 96.73% (13,779/14,245) of the cases. This examination showed that the 13,779 active presentations, where the app constitutes a social actor, can be split into 5 broadly recurring themes: app informing (21.47%), app instructing (15.33%), app giving permission (9.1%), app disrupting (5.02%), and app functioning or not functioning (49.07%; Table 12). To answer our research question, the discussion will elaborate on the links between these constructs and the relationship between what is present in the discourse and what is present in the literature.

Table 12. Comparison of the percentage of each theme found when the app is presented as a social actor.

Theme	Values (%)
Informing	21.47
Instructing	15.33
Giving permission	9.1
Disrupting	5.02
Malfunctioning	49.07

Discussion

Overview

As mentioned earlier, our analysis showed that the app was presented actively as a social actor on approximately 96% of occurrences and unearthed 5 main categories for these active presentations: informing, instructing, providing permission, disrupting, and functioning. These categories revealed the personalized and independent decision-making role of the app. As for passive presentations, instances of backgrounding, where developers and the public were foregrounded, obscured the app's agency, potentially reflecting a small number of Twitter users' beliefs about the app's role and responsibility. This section explores the implications of these results.

Principal Findings: Trends of Active Agency

Overview

Through the analysis of transitivity in the 14,425 concordance lines considered, the collocations of "app" and "by the app" and CDA-informed analysis of agency and responsibility, underpinned by SAR, we identified 5 main categories that the active presentations of the app fall into: informing (21.47%), instructing (15.33%), providing permission (9.1%), disrupting (5.02%), and functioning (49.07%). The first 4 categories show the app to be personalized [27] and to make decisions independently [24]. Meanwhile, functioning included instances of the app acting autonomously but also simply functioning as intended or designed to do. In addition, this category included tweets where the app was functioning appropriately, but also contained tweets where it was presented as not functioning as desired. This may explain the large percentage of tweets in this category.

Informing

The app was presented actively (21.47% of occurrences) when providing information to its users through "saying" and "pinging." This happened especially at the start of the discourse, as the app was "saying" information that was difficult to understand (LogDice: 8.07). Similarly, users complained about the app informing them (41/64, 64% occurrences), determining a trend that followed other areas of discourse. An example of this was in the "early months" part of the timeline, where the app communicated about the status of their self-isolation period (LogDice: 8.58). In this sense, the idea of presenting information linked to the findings of the study by Williams et al [44], while also pointing out that the app proving information to users was deemed a core responsibility of the app, hence the questions and negative reactions when the app that "failed."

The app was presented actively informing also through the surge in "ping," a frequent collocate of "app" between January and February 2021 (LogDice: 7.98). The tweets containing these collocates depicted the app as acting autonomously, once again leaning toward personalization. Although some instances later in the discourse presented the app as providing useful information, most presentations still remained negative when it came to the information given—or not—to users (LogDice of "provide" in March-June 2021 being 7.22). "Ping" clearly continued being a verbal trend into July 2021, with many tweet authors discussing the impact that the app informing them of a COVID-19 positive or isolation status (LogDice: 8.54). This is complemented by strong collocations of "tell" (LogDice: 9.15) and "say" (LogDice: 7.10). Therefore, it could be inferred that the app's active presentation, using "ping," and the perception that it might have provided incorrect information contrasted with the rationale for having a decision-making algorithm in the first place [4].

Instructing

The app was also seen as actively providing instructions to users and the wider public (15.33% of occurrences). "Tell" was a frequent collocate of "app" throughout the discourse (LogDice scores of 7.90, 8.55, 7.63, and 9.15), and users presented this as the app instructing them to take action, most notably, to self-isolate (232/270, 85.9% occurrences). At the beginning of the discourse, 30 of 52 occurrences were hypothetical, likely due to the app being newly launched. Twitter users also questioned the instructions provided by the app (12/52, 23%

XSL•FO
RenderX

occurrences). This imperative tone continued in the final months of 2020, with users stating that the app instructed them to self-isolate (44/61, 72% occurrences). However, what became more apparent in this section of the discourse was that, although the app was presented as a social actor through personalization, the impact of this system was ridiculed through humorous additions to tweets or sarcastic tone (9/61, 15% occurrences). This likely softened the instructional impact that the app had, while still presenting it as a social actor.

The app continued to be presented as actively instructing in later parts of the discourse too, with direct first person accounts of experiences when the app "is telling" users (LogDice: 9.15), as well as reports of hypothetical scenarios (14/23, 61% occurrences). These constructions exhibited human-level agency through personalization. The other constructions seen in July 2021, such as those that contain "wreak," "cripple," and "threat," implied equally a significant level of agency as if the app's instructions could only result in negative consequences, although this will be explored in more detail in the "disrupting" section. The presentation of the app in this way intersected with concerns about the merging of algorithmic and human agency [8,9,39,88] because the app is presented as performing the job of a human. In particular, the app was featured in constructions where users were frustrated with its instructions and users or their lack, ultimately providing insights into the agency of the algorithm's perceived role and responsibility.

Giving Permission

Although less common than the previous 2 categories, the app was also presented actively when providing users or general public permission (9.1% of occurrences). This is most prominently seen in users stating that the app is "allowing." More present at the start of the discourse (LogDice: 7.35 for September 2020 and 6.73 for October-December 2020), due to the questions being asked about the app, this was less frequently discussed as time passed. Sets of tweets in the discourse pointed to the app providing permission. For example, the app "gives" notification of self-isolation periods at the end of 2020. This recalled some permission concerns found by Dowthwaite et al [16].

It could also be argued that "need," a strong collocate at the beginning of the discourse (LogDice: 7.50), intersected this theme and "functioning." The idea that the app needed to provide permission to humans was an occurrence of personalization, providing further insight into the idea that the app was not only given agency but also showcasing its necessity to process information systematically.

Disrupting

The app was presented as disrupting users' lives through active constructions. This is seen early in the discourse, when users commented on the app making disruptive decisions autonomously, such as turning off contact-tracing functionalities (88/495, 17.8% occurrences). This continued throughout the discourse, with users describing the problems the app has caused them. However, most tweets that suggested the app was actively disrupting the lives of the public appeared more toward the end of the sampled period, when the "pingdemic" occurred (456/845, 54% occurrences). Examples of these included instances when the app was defined as "wreak havoc," "cripple Britain," and "threaten the economy." This relates to the idea that the system failed to meet the needs and expectations of users [15,17]. It also supported the findings of Lamanna and Byrne [37] and Riegler [38], according to whom humans could be perceived as "at odds" with the decisions made by this system.

Functioning

One final category (49.07% of occurrences) was the app being presented as independently undertaking (or attempting to undertake) functional activities that were integral to its running. Tweets in the later months of 2020 stated that the app had a job, a clear personalization (LogDice: 7.02). Tweets in this discourse indicated that one of the intended primary functions of the app was to help or assist users. When this was perceived as not happening, the app was not fulfilling its (supposed) rationale for existing. This is particularly prevalent when the app was said to not be "helping," at the start of 2021 (LogDice: 5.22), and perceived as failing to keep users safe. The fact that Twitter authors saw the app as being responsible for their safety and welfare showed its prominence and influence as a social actor, similar to the findings of Kent [15] and Mbwogge [17]. In the later parts of the discourse, the app was occasionally presented as having limited legal power or obligation over users (48/283, 17% occurrences), providing insight into how the app was perceived as responsible for its users.

During the "pingdemic" part of the discourse, the app was said to "send" (LogDice: 9.38) many notifications, suggesting that the app was designed for this. Many of these tweets indicated that, although the app was not necessarily instructing users, it encouraged noncompliance with too many notifications. This recalled the findings from the follow-up study by Pepper et al [2]. In addition, it may also indicate that the app "pinging" was perceived as more invasive than simply "saying." Despite users wanting the app to function properly, they appeared to find "pinging" overbearing.

In addition, personalization was observed when looking at the app's perceived functionality. For example, the collocate "think" (LogDice: 7.24), seen throughout the discourse, regularly presented the app as stating something that was incorrect (30/39, 77% occurrences). This was related to the app not working as perceived by the Twitter user. This indicated that the app was perceived as having the capacity to think or act autonomously, leading to the opposition of system use [40,41,88].

Principal Findings: Trends of Passivization

Instances of backgrounding were found throughout the discourse. Examples of this included the way in which the developers of the app (Serco, the NHS, and the UK government) were foregrounded, especially at the start of the discourse, and how the public, affected by COVID-19 and isolation requirements, became a focus over time. This meant that the app was backgrounded according to the principles of SAR [27], with its agency obscured [26]. The app was still discussed negatively in these constructions, despite not being an overt social actor, due to its reduced agency. This presentation intersected part of the work outlined by Feier et al [11], who

XSL•FO
RenderX

suggested that decision-making algorithms may deflect blame from more responsible players. In addition, this portrayal may be a reflection of the perceived attitudes of Twitter users; some (14/42, 33% occurrences) believed that, while the app played an important role, the responsibility remained with the developers of the app or with the humans that used the app at their own discretion. The removal of the agency diluted the impact [31]. That said, the proportion of passive presentations of the app was very small (approximately 4% of all constructions) in comparison with active presentations.

However, considering verb choices, such as "tell" and "cause," passive constructions were still present. Thus, a small portion of tweets (32/366, 8.7% occurrences) using passive constructions appeared to imply that the app still has some agential power, which may be labeled as agency metaphors [32]. Such agential power could be considered as impacting the app; hence, the app was still deemed to have some responsibility for processing information.

Limitations

With a corpus of 118,316 tweets, it would have been practically impossible to manually examine each [89]. Hence, CL was used to filter the data set collected and identify relevant potential social actors through the analysis of the keyword "app" and its 14,245 collocates in the corpus, which were examined through concordance grids and LogDice. This methodological approach was intended to mitigate this issue of infeasibility.

Another limitation posed by CDA was subjective biases, impacting the interpretation of instances of sarcasm and humor, especially those that were less explicit. In this sense, this challenge is not new to researchers [90,91]. Nonetheless, the combination of CDA with computationally aided techniques was intended to reduce the impact of this difficulty and may benefit future research.

As "app" was the key-term searched, this work disregarded most instances where exclusion masked the app in constructions and the actual word did not feature. Nonetheless, this could constitute an interesting future research focus that encompassed explicitly excluded constructions. In addition, the system may have been discussed in tweets without specific reference to "app." Although other social actors replacing "app" would be hard to find a large corpus, a good starting point may be synonyms of "app" in this specific context, such as "(information) system," "application," "tool," "program" or "software." Similarly, related field-specific words may also offer relevant research insights, such as "functionality," "function(s)," "operation(s)," "spread(ing)," "track(ing)," or "trace/tracing."

Due to the brevity of Twitter discourse, which is limited by a 280 character limit, the app may have been presented actively to facilitate conciseness. For example, "the app told me to isolate" (26 characters) contained 6 fewer characters than "i was told to isolate by the app" (32 characters), which could have been an equally valuable semantic alternative. Such a possibly increased number of active presentations is likely to have affected the number of times the app was presented clearly as a social actor. Consequently, future work may involve examining other social media or text-sharing platforms that do not limit characters in posts or content to see if the most active agential presentations are comparable.

Conclusions

According to our CL-, CDA-, and SAR-based examination of agency and transitivity in tweets containing the word "app," published between September 2020 and July 2021, Twitter users presented the NHS COVID-19 App as a social actor and with a clear sense of social agency, addressing our first research question concerning how society had been impacted by the deployment of this system (ie, the NHS COVID-19 App).

Specifically, the app was predominantly presented actively by Twitter users in 96% of the cases, using various techniques, most notably personalization, but also including determination, agency metaphor, and genericism. Indeed, we found that these active presentations, which implied social agency, primarily conveyed the idea of app informing (21.47%), instructing (15.33%), providing permission (9.1%), disrupting (5.02%), and functioning or failing to (49.07%).

The app was also presented passively on occasions (approximately 3%), although this decreased as the discourse continued, reaching a maximum impact of 4% in September 2020 and decreasing to a minimum of 2% in July 2021. In such instances, the app was often backgrounded to make the developers or operators of the app more apparent or responsible. On occasion, the focus was on the members of the public affected by the app malfunctioning rather than the app itself. Comparable instances, when the impact of the app as a social actor was limited, the app was presented actively but simultaneously ridiculed.

The implications for this study, with regard to our second research question concerning the app's perceived responsibility to process information, are that Twitter users presented the app as responsible for their own welfare through various active presentations, especially when the app instructed them or provided permission. According to the tweets examined, the perceived responsibility to process information remained in the app throughout the discourse. Such a perception was especially pronounced when significant events prompted further questioning of the app's capabilities (ie, during the app's launch in September 2020, the second lockdown in January 2021 and the "pingdemic" phase in July 2021).

In addition to offering insights into web-based responses to this specific event, this contribution holds the potential for broader implications in the context of decision-making algorithms. Although the disruption caused by the pandemic has waned in the United Kingdom, the findings of this study shed light on how the public might respond to forthcoming decision-making algorithm interventions. This insight is particularly valuable in the context of health care or digital contact-tracing initiatives, shining light on barriers to adoption. Therefore, even in a postpandemic world, the findings of this study remain important.

Overall, this study has provided insights into how social agency communicated via social media public discourse dealing with algorithmic-operated decisions when the AI agency behind those information systems is not openly disclosed. Such a relationship

XSL•FO
RenderX

was exemplified by that between the NHS app, grammatical agency, and social agency, building on existing work on the social agency of decision-making algorithms [33,36,37,40]. Therefore, our study contributes to the investigation of the social impact of the NHS COVID-19 App, in particular, showcased through the combination of CL and CDA underpinned by SAR. Briefly, our research argues that the views expressed on social media indicate that the app was presented as having a perceived high level of responsibility for the welfare and safety of its users according to tweets that explicitly referred to the app.

Acknowledgments

All authors were supported by funder 1 (UKRI grant EP/V00784X/1). DH was supported by funder 2 (UKRI grant EP/S023305/1).

Data Availability

All data generated or analyzed during this study are included in this paper and its supplementary information files.

Authors' Contributions

DH was responsible for conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing the original draft, and reviewing and editing. EN was responsible for conceptualization, formal analysis, methodology, supervision, writing the original draft, and reviewing and editing. JC was responsible for conceptualization, supervision, and reviewing and editing. JEF was responsible for conceptualization, supervision, and reviewing and editing.

Conflicts of Interest

None declared.

References

1. Wagner B. Liable, but not in control? Ensuring meaningful human agency in automated decision‐making systems. Policy Internet 2019 Jan 24;11(1):104-122. [doi: 10.1002/poi3.198]
2. Pepper C, Reyes-Cruz G, Pena AR, Dowthwaite L, Babbage CM, Wagner H, et al. Understanding trust and changes in use after a year with the NHS COVID-19 contact tracing app in the United Kingdom: longitudinal mixed methods study. J Med Internet Res 2022 Oct 14;24(10):e40558 [FREE Full text] [doi: 10.2196/40558] [Medline: 36112732]
3. Araujo T, Helberger N, Kruikemeier S, de Vreese CH. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc 2020 Jan 01;35(3):611-623. [doi: 10.1007/s00146-019-00931-w]
4. Busch PA, Henriksen HZ. Digital discretion: a systematic literature review of ICT and street-level discretion. Inf Polity 2018 Feb 11;23(1):3-28. [doi: 10.3233/ip-170050]
5. Panagiotopoulos P, Klievink B, Cordella A. Public value creation in digital government. Gov Inf Q 2019 Oct;36(4):101421. [doi: 10.1016/j.giq.2019.101421]
6. Bullock JB. Artificial intelligence, discretion, and bureaucracy. Am Rev Public Adm 2019 Jun 18;49(7):751-761. [doi: 10.1177/0275074019856123]
7. Young MM, Bullock JB, Lecy JD. Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. Perspect Public Manag Gov 2019 Dec;2(4):301-313. [doi: 10.1093/ppmgov/gvz014]
8. Crang M, Graham S. SENTIENT CITIES ambient intelligence and the politics of urban space. Inf Commun Soc 2007 Dec;10(6):789-817. [doi: 10.1080/13691180701750991]
9. Ziewitz M. Governing algorithms: myth, mess, and methods. Sci Technol Hum Values 2015 Sep 30;41(1):3-16. [doi: 10.1177/0162243915608948]
10. Olhede S, Wolfe PJ. Blame the algorithm? Significance 2020 Oct;17(5):12. [doi: 10.1111/1740-9713.01441]
11. Feier T, Gogoll J, Uhl M. Hiding behind machines: when blame is shifted to artificial agents. arXiv. Preprint posted online January 27, 2021 2023.
12. Peeters R. The agency of algorithms: understanding human-algorithm interaction in administrative decision-making. Inf Polity 2020 Dec 04;25(4):507-522. [doi: 10.3233/ip-200253]
13. Velkova J, Kaun A. Algorithmic resistance: media practices and the politics of repair. Inf Commun Soc 2019 Aug 26;24(4):523-540. [doi: 10.1080/1369118x.2019.1657162]
14. Kretzschmar ME, Rozhnova G, Bootsma MC, van Boven M, van de Wijgert JH, Bonten MJ. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health 2020 Aug;5(8):e452-e459. [doi: 10.1016/s2468-2667(20)30157-2]
15. Kent C. UK contact-tracing app problems: building trust in tempramental technology. Medical Device Network. 2020 Oct 6. URL: https://www.medicaldevice-network.com/features/uk-contact-tracing-app-problems/ [accessed 2023-06-28]

XSL•FO
RenderX

16. Dowthwaite L, Fischer J, Perez Vallejos E, Portillo V, Nichele E, Goulden M, et al. Public adoption of and trust in the NHS COVID-19 contact tracing app in the United Kingdom: quantitative online survey study. J Med Internet Res 2021 Sep 17;23(9):e29085 [FREE Full text] [doi: 10.2196/29085] [Medline: 34406960]
17. Mbwogge M. Mass testing with contact tracing compared to test and trace for the effective suppression of COVID-19 in the United Kingdom: systematic review. JMIRx Med 2021 Apr 12;2(2):e27254 [FREE Full text] [doi: 10.2196/27254] [Medline: 33857269]
18. Paucar LH, Bencomo N, Sutcliffe A, Sawyer P. A Bayesian network-based model to understand the role of soft requirements in technology acceptance: the case of the NHS COVID-19 test and trace app in England and Wales. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. 2022 Presented at: SAC '22; April 25-29, 2022; Virtual Event. [doi: 10.1145/3477314.3507147]
19. Wymant C, Ferretti L, Tsallis D, Charalambides M, Abeler-Dörner L, Bonsall D, et al. The epidemiological impact of the NHS COVID-19 app. Nature 2021 Jun 12;594(7863):408-412. [doi: 10.1038/s41586-021-03606-z] [Medline: 33979832]
20. Marsh K, Griffiths E, Young JJ, Gibb CA, McMenamin J. Contributions of the EURO 2020 football championship events to a third wave of SARS-CoV-2 in Scotland, 11 June to 7 July 2021. Euro Surveill 2021 Aug;26(31):2100707 [FREE Full text] [doi: 10.2807/1560-7917.ES.2021.26.31.2100707] [Medline: 34355691]
21. Smith JA, Hopkins S, Turner C, Dack K, Trelfa A, Peh J, et al. Public health impact of mass sporting and cultural events in a rising COVID-19 prevalence in England. Epidemiol Infect 2022 Jan 31;150:e42. [doi: 10.1017/s0950268822000188]
22. Ahearn LM. Agency. J Linguistic Anthropol 2008 Jun 28;9(1-2):12-15. [doi: 10.1525/jlin.1999.9.1-2.12]
23. Grillitsch M, Rekers JV, Sotarauta M. Investigating agency: methodological and empirical challenges. In: Sotarauta M, Beer A, editors. Handbook on City and Regional Leadership. Cheltenham, UK: Edward Elgar Publishing; Feb 22, 2021.
24. Richardson P, Mueller CM, Pihlaja S. Cognitive Linguistics and Religious Language: An Introduction. Milton Park, UK: Taylor & Francis; 2021.
25. Leslie AM. A theory of agency. In: Causal Cognition: A Multidisciplinary Debate. Oxford, UK: Oxford University Press; 1995.
26. Clark WR. Agents and structures: two views of preferences, two views of institutions. Int Stud Q 1998 Jun;42(2):245-270. [doi: 10.1111/1468-2478.00081]
27. Van Leeuwen T. Discourse and Practice: New Tools for Critical Discourse Analysis. Oxford, UK: Oxford University Press; 2008.
28. Gallagher S. Philosophical conceptions of the self: implications for cognitive science. Trends Cogn Sci 2000 Jan;4(1):14-21. [doi: 10.1016/s1364-6613(99)01417-5] [Medline: 10637618]
29. Silver CA, Tatler BW, Chakravarthi R, Timmermans B. Social agency as a continuum. Psychon Bull Rev 2021 Apr 07;28(2):434-453 [FREE Full text] [doi: 10.3758/s13423-020-01845-1] [Medline: 33289061]
30. Oktar L. The ideological organization of representational processes in the presentation of us and them. Discourse Soc 2016 Jul 25;12(3):313-346. [doi: 10.1177/0957926501012003003]
31. Comrie B. In defense of spontaneous demotion: the impersonal passive. In: Grammatical Relations. Leiden, The Netherlands: Brill Publishers; 1977:47-58.
32. Morris MW, Sheldon OJ, Ames DR, Young MJ. Metaphors and the market: consequences and preconditions of agent and object metaphors in stock market commentary. Organ Behav Hum Decis Process 2007 Mar;102(2):174-192. [doi: 10.1016/j.obhdp.2006.03.001]
33. Zarsky T. The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci Technol Hum Values 2015 Oct 14;41(1):118-132. [doi: 10.1177/0162243915605575]
34. Barocas S, Selbst AD. Big data's disparate impact. Calif Law Rev 2016 Jun;104(3):671-732. [doi: 10.2139/ssrn.2477899]
35. Bodo B, Helberger N, Irion K, Zuiderveen Borgesius F, van de Velde B, Bol N, et al. Tackling the algorithmic control crisis – the technical, legal, and ethical challenges of research into algorithmic agents. Yale J Law Technol 2017;19:133-180.
36. Rubel A, Castro C, Pham A. Algorithms, agency, and respect for persons. Soc Theory Pract 2020 Jul;46(3):547-572. [doi: 10.5840/soctheorpract202062497]
37. Lamanna C, Byrne L. Should artificial intelligence augment medical decision making? The case for an autonomy algorithm. AMA J Ethics 2018 Sep 01;20(9):E902-E910 [FREE Full text] [doi: 10.1001/amajethics.2018.902] [Medline: 30242824]
38. Riegler C. The moral decision-making capacity of self-driving cars: socially responsible technological development, algorithm-driven sensing devices, and autonomous vehicle ethics. Contemp Read Law Soc Justice 2019;11(1):15-20. [doi: 10.22381/CRLSJ11120192]
39. Meisner C, Duffy BE, Ziewitz M. The labor of search engine evaluation: making algorithms more human or humans more algorithmic? New Media Soc 2022 Jan 27. [doi: 10.1177/14614448211063860]
40. Mahmud H, Islam AK, Ahmed SI, Smolander K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technol Forecast Soc Change 2022 Feb;175:121390. [doi: 10.1016/j.techfore.2021.121390]
41. Grange C. Algorithmically controlled automated decision-making and societal acceptability: does algorithm type matter? In: Proceedings of the 55th Hawaii International Conference on System Sciences. 2022 Presented at: 55th Hawaii International Conference on System Sciences; January 4-7, 2022; Maui, Hawaii URL: https://scholarspace.manoa.hawaii.edu/items/d4f8bec4-b3e7-45c0-9677-03daea9e97e8 [doi: 10.24251/hicss.2022.629]

XSL•FO
RenderX

42. Weekly statistics for NHS test and trace (England). United Kingdom Government. 2020 Jun 18. URL: https://www.gov.uk/government/collections/nhs-test-and-trace-statistics-england-weekly-reports [accessed 2023-06-28]
43. Pandit JA, Radin JM, Quer G, Topol EJ. Smartphone apps in the COVID-19 pandemic. Nat Biotechnol 2022 Jul;40(7):1013-1022. [doi: 10.1038/s41587-022-01350-x] [Medline: 35726090]
44. Williams SN, Armitage CJ, Tampe T, Dienes K. Public attitudes towards COVID-19 contact tracing apps: a UK-based focus group study. Health Expect 2021 Apr;24(2):377-385 [FREE Full text] [doi: 10.1111/hex.13179] [Medline: 33434404]
45. Samuel G, Roberts SL, Fiske A, Lucivero F, McLennan S, Phillips A, et al. COVID-19 contact tracing apps: UK public perceptions. Crit Public Health 2022 Jan 01;32(1):31-43 [FREE Full text] [doi: 10.1080/09581596.2021.1909707] [Medline: 35221546]
46. Dowthwaite L, Wagner HG, Babbage CM, Fischer JE, Barnard P, Nichele E, et al. The relationship between trust and attitudes towards the COVID-19 digital contact-tracing app in the UK. PLoS One 2022 Oct 27;17(10):e0276661 [FREE Full text] [doi: 10.1371/journal.pone.0276661] [Medline: 36301881]
47. Abbasi K. Covid-19: the UK's political gamble that bodes ill for health and the health service. BMJ 2021 Jul 22;374:n1848. [doi: 10.1136/bmj.n1848]
48. Rimmer A. Sixty seconds on . . . the pingdemic. BMJ 2021 Jul 19;374:n1822. [doi: 10.1136/bmj.n1822] [Medline: 34281909]
49. Heaton D, Clos J, Nichele E, Fischer J. Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. PeerJ Comput Sci 2023 Jan 30;9:e1211 [FREE Full text] [doi: 10.7717/peerj-cs.1211] [Medline: 37346687]
50. Kumar S, Morstatter F, Liu H. Twitter Data Analytics. New York, NY: Springer; 2014.
51. Jianqiang Z. Pre-processing boosting Twitter sentiment analysis? In: Proceedings of the 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity). 2015 Presented at: IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity); December 19-21, 2015; Chengdu, China. [doi: 10.1109/smartcity.2015.158]
52. Chong WY, Selvaretnam B, Soon LK. Natural language processing for sentiment analysis: an exploratory analysis on Tweets. In: Proceedings of the 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology. 2014 Presented at: 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology; December 3-5, 2014; Kota Kinabalu, Malaysia. [doi: 10.1109/icaiet.2014.43]
53. Woodfield K, Morrell G, Metzler K, Blank G, Finnegan J, Lucraft M. Blurring the boundaries? New social media, new social research: developing a network to explore the issues faced by researchers negotiating the new research landscape of online social media platforms. National Centre for Research Methods. 2013. URL: https://www.homelesshub.ca/resource/blurring-boundaries-new-social-media-new-social-research-developing-network-explore-issues [accessed 2023-12-25]
54. Roesslein J. tweepy documentation: release 3.6.0. Read the Docs. 2018 Mar 2. URL: https://buildmedia.readthedocs.org/media/pdf/tweepy/v3.6.0/tweepy.pdf [accessed 2023-06-28]
55. Jaworska S. Corpus approaches: investigating linguistic patterns and meanings. In: The Routledge Handbook of Language and Media. Milton Park, UK: Routledge; 2017.
56. Baker P. Sociolinguistics and Corpus Linguistics. Edinburgh, UK: Edinburgh University Press; 2010.
57. Hunston S. How can a corpus be used to explore patterns? In: The Routledge Handbook of Corpus Linguistics. Milton Park, UK: Routledge; 2022.
58. Anthony L. A critical look at software tools in corpus linguistics. Linguistic Res 2013;30(2):141-161 [FREE Full text] [doi: 10.17250/khisli.30.2.201308.001]
59. Kopaczyk J, Tyrkkö J. Applications of Pattern-driven Methods in Corpus Linguistics. Amsterdam, The Netherlands: John Benjamins Publishing Company; Mar 2018.
60. Nugraha IS, Sujatna ET, Mahdi S. Corpus linguistic study of tweets using #charliehebdo hashtag. JALL 2021 Feb;5(1):54-70. [doi: 10.25157/jall.v5i1.4965]
61. Nichele E, Kopf S. Es-Tu Charlie? Doing politics on Wikipedia. In: Proceedings of the Political Discourses: Multidisciplinary Approaches. 2015 Presented at: Political Discourses: Multidisciplinary Approaches; June 26-27, 2012; London, UK. [doi: 10.3109/09638288.2012.656792]
62. Russo KE, Grasso A. Coping with dis/ableism in Twitter discourse: a corpus-based critical appraisal analysis of the Hidden Disabilities Sunflower Lanyard case. Int J Language Stud 2022;16(4):65-94 [FREE Full text]
63. Baker P. Using Corpora in Discourse Analysis. London, UK: Continuum International Publishing Group; 2006.
64. Mautner G. Mining large corpora for social information: the case of elderly. Lang Soc 2007 Jan 26;36(01):51-72. [doi: 10.1017/s0047404507070030]
65. Hoey M. Text, Discourse and Corpora: Theory and Analysis. London, UK: Bloomsbury Academic; 2007.
66. Kilgarriff A, Rychly P, Smrz P, Tugwell D. The sketch engine. In: Proceedings of Euralex. 2004 Presented at: Euralex 2004 Proceedings; July 6-10, 2004; Lorient, France URL: https://www.researchgate.net/publication/260387608_ITRI-04-08_the_sketch_engine/> [doi: 10.1093/oso/9780199292332.003.0020]
67. Suchomel V. Better web corpora for corpus linguistics and NLP. Masaryk University. 2020. URL: https://is.muni.cz/th/u4rmz/Better_Web_Corpora_For_Corpus_Linguistics_And_NLP.pdf [accessed 2023-12-25]
68. Cruse DA. Lexical Semantics. Cambridge, UK: Cambridge University Press; 1986.

XSL•FO
RenderX

69. Baker P. Acceptable bias? Using corpus linguistics methods with critical discourse analysis. Crit Discourse Stud 2012 Aug;9(3):247-256. [doi: 10.1080/17405904.2012.688297]
70. Nartey M, Mwinlaaru IN. Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis. Corpora 2019 Aug;14(2):203-235. [doi: 10.3366/cor.2019.0169]
71. Abbas A, Zahra T. Corpus driven critical discourse analysis of 2020 presidential election campaign tweets of Donald Trump and Joe Biden. Hayatian J Linguist Lit 2021;5(1):26-47.
72. Johnson M, McLean E. Discourse analysis. In: Kobayashi A, editor. International Encyclopedia of Human Geography. Amsterdam, The Netherlands: Elsevier Science; 2019.
73. Hart C. Critical discourse analysis and metaphor: toward a theoretical framework. Crit Discourse Stud 2008 Apr 17;5(2):91-106. [doi: 10.1080/17405900801990058]
74. Kendall G. What is critical discourse analysis? Forum Qual Sozialforschung 2007;8(2). [doi: 10.17169/fqs-8.2.255]
75. van Dijk TA. What is political discourse analysis? Belgian J Linguistics 1997 Jan;11(1):11-52. [doi: 10.1075/bjl.11.03dij]
76. Tenorio EH. Critical discourse analysis, an overview. Nord J Engl Stud 2011;10(1):183-210. [doi: 10.35360/njes.247]
77. Bloor M, Bloor T. The Practice of Critical Discourse Analysis: An Introduction. Milton Park, UK: Taylor & Francis; 2013.
78. Aljarallah R. A critical discourse analysis of Twitter posts on the perspectives of women driving in Saudi Arabia. Arizona State University. 2017 Aug. URL: https://keep.lib.asu.edu/items/155782/view [accessed 2023-12-25]
79. Sveinson K, Allison R. “Something seriously wrong with U.S. soccer”: a critical discourse analysis of consumers’ Twitter responses to U.S. soccer’s girls’ apparel promotion. J Sport Manage 2021 Dec 21;36(5):446-458. [doi: 10.1123/jsm.2021-0127]
80. Kreis R. #refugeesnotwelcome: anti-refugee discourse on Twitter. Discourse Commun 2017 Jul 04;11(5):498-514. [doi: 10.1177/1750481317714121]
81. Weber M. Max Weber: Selections in Translation. Cambridge, UK: Cambridge University Press; 1978.
82. Engeström Y, Miettinen R, Punamäki RL. Perspectives on Activity Theory. Cambridge, UK: Cambridge University Press; 1999.
83. McGlashan M. Collective identity and discourse practice in the followership of the Football Lads Alliance on Twitter. Discourse Soc 2019 Nov 24;31(3):307-328. [doi: 10.1177/0957926519889128]
84. Fadanelli SB, Pozzo DF, Fin CC. The representation of social actors in the tweets of Jair Messias Bolsonaro. Antares 2020;12(25):74-99. [doi: 10.18226/19844921.v12.n25.04]
85. Bernard T. The discursive representation of social actors in the corporate social responsibility (CSR) and integrated annual (IA) reports of two South African mining companies. Crit Approach Discourse Anal Discip 2018;10(1):81-87.
86. Razis G, Anagnostopoulos I, Saloun P. Thematic labeling of Twitter accounts using DBpedia properties. In: Proceedings of the 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). 2016 Presented at: 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP); October 20-21, 2016; Thessaloniki, Greece URL: https://ieeexplore.ieee.org/document/7753393 [doi: 10.1109/smap.2016.7753393]
87. Kitishat AR, Kayed MA, Al-Ajalein M. A corpus-assisted critical discourse analysis of the Syrian refugee crisis in Jordanian newspapers. Int J Engl Linguist 2020 Sep 15;10(6):195. [doi: 10.5539/ijel.v10n6p195]
88. Beer D. The social power of algorithms. Inf Commun Soc 2016 Aug 08;20(1):1-13. [doi: 10.1080/1369118x.2016.1216147]
89. Wetherell M, Potter J. Discourse analysis and the identification of interpretive repertoires. In: Antaki C, editor. Analysing Everyday Explanation: A Casebook of Methods. Newbury Park, CA: SAGE Publications; 1988:168-183.
90. Gaskell G, Bauer MW, Atkinson P. Qualitative Researching with Text, Image and Sound: A Practical Handbook for Social Research. Thousand Oaks, CA: SAGE Publications; 2000.
91. Morgan A. Discourse analysis: an overview for the neophyte researcher. J Health Soc Care Improv 2017;2010.

Abbreviations

AI: artificial intelligence
CDA: critical discourse analysis
CL: corpus linguistics
NHS: National Health Service
SAR: social actor representation

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 29.06.23; peer-reviewed by E Jaafar, YK Suh; comments to author 19.10.23; revised version received 20.10.23; accepted 20.10.23; published 01.02.24.

Please cite as:
Heaton D, Nichele E, Clos J, Fischer JE
Perceptions of the Agency and Responsibility of the NHS COVID-19 App on Twitter: Critical Discourse Analysis
J Med Internet Res 2024;26:e50388
URL: https://www.jmir.org/2024/1/e50388
doi:10.2196/50388
PMID:38300688

©Dan Heaton, Elena Nichele, Jérémie Clos, Joel E Fischer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Analyzing Reddit Forums Specific to Abortion That Yield Diverse Dialogues Pertaining to Medical Information Seeking and Personal Worldviews: Data Mining and Natural Language Processing Comparative Study

Danny Valdez[1], PhD; Lucrecia Mena-Meléndez[1], PhD; Brandon L Crawford[1], PhD; Kristen N Jozkowski[1], PhD

Department of Applied Health Science, Indiana University School of Public Health, Bloomington, IN, United States

Corresponding Author:
Danny Valdez, PhD
Department of Applied Health Science
Indiana University School of Public Health
1025 E 7th Street
Bloomington, IN, 47403
United States
Phone: 1 8128038955
Email: danvald@iu.edu

Abstract

Background: Attitudes toward abortion have historically been characterized via dichotomized labels, yet research suggests that these labels do not appropriately encapsulate beliefs on abortion. Rather, contexts, circumstances, and lived experiences often shape views on abortion into more nuanced and complex perspectives. Qualitative data have also been shown to underpin belief systems regarding abortion. Social media, as a form of qualitative data, could reveal how attitudes toward abortion are communicated publicly in web-based spaces. Furthermore, in some cases, social media can also be leveraged to seek health information.

Objective: This study applies natural language processing and social media mining to analyze Reddit (Reddit, Inc) forums specific to abortion, including r/Abortion (the largest subreddit about abortion) and r/AbortionDebate (a subreddit designed to discuss and debate worldviews on abortion). Our analytical pipeline intends to identify potential themes within the data and the affect from each post.

Methods: We applied a neural network–based topic modeling pipeline (BERTopic) to uncover themes in the r/Abortion (n=2151) and r/AbortionDebate (n=2815) subreddits. After deriving the optimal number of topics per subreddit using an iterative coherence score calculation, we performed a sentiment analysis using the Valence Aware Dictionary and Sentiment Reasoner to assess positive, neutral, and negative affect and an emotion analysis using the Text2Emotion lexicon to identify potential emotionality per post. Differences in affect and emotion by subreddit were compared.

Results: The iterative coherence score calculation revealed 10 topics for both r/Abortion (coherence=0.42) and r/AbortionDebate (coherence=0.35). Topics in the r/Abortion subreddit primarily centered on information sharing or offering a source of social support; in contrast, topics in the r/AbortionDebate subreddit centered on contextualizing shifting or evolving views on abortion across various ethical, moral, and legal domains. The average compound Valence Aware Dictionary and Sentiment Reasoner scores for the r/Abortion and r/AbortionDebate subreddits were 0.01 (SD 0.44) and −0.06 (SD 0.41), respectively. Emotionality scores were consistent across the r/Abortion and r/AbortionDebate subreddits; however, r/Abortion had a marginally higher average *fear* score of 0.36 (SD 0.39).

Conclusions: Our findings suggest that people posting on abortion forums on Reddit are willing to share their beliefs, which manifested in diverse ways, such as sharing abortion stories including how their worldview changed, which critiques the value of dichotomized abortion identity labels, and information seeking. Notably, the style of discourse varied significantly by subreddit. r/Abortion was principally leveraged as an information and outreach source; r/AbortionDebate largely centered on debating across various legal, ethical, and moral abortion domains. Collectively, our findings suggest that abortion remains an opaque yet politically charged issue for people and that social media can be leveraged to understand views and circumstances surrounding abortion.

(J Med Internet Res 2024;26:e47408) doi:10.2196/47408

XSL•FO
RenderX

KEYWORDS

abortion; social media; Reddit; natural language processing; NLP; neural networks

Introduction

Background

Although the abortion debate is often framed along strict proabortion or antiabortion stances (eg, prochoice versus prolife—terms common in the United States, Ireland, and other English-speaking countries; *pro-elección* versus *provida* and *pro-aborto* versus *anti-aborto*—terms used in Mexico, Argentina, and other Spanish-speaking countries), actual abortion beliefs are complex, contextual, and at times contradictory [1-4]. Notably, despite media characterizations of these 2 oppositional perspectives—for people ascribing to either proabortion or prochoice labels (ie, broad abortion support) or antiabortion or prolife labels (ie, broad abortion opposition)—there exist circumstances in which people's views diverge from the dichotomy [5]. These circumstances include, for example, the gestation period of pregnancy [6], the context for seeking abortion [7], and whether people consider abortion as a legal versus moral issue [8]. In addition, attitudes toward abortion also vary across some demographic characteristics such as age, educational attainment, political affiliation, and race or ethnicity of a person or groups of people participating in a survey [1,9,10].

Beyond context-specific or cultural considerations that may predict complex abortion views, personal accounts, narratives, and discussions about abortion may similarly reveal the extent to which abortion views depart from a support or opposition dichotomy, including extreme abortion circumstances or personal experience with an abortion. Evidence suggests that these considerations are not ethnocentric but shared globally. Research comparing abortion beliefs between English-speaking US residents and Spanish-speaking US residents of diverse nations of origin demonstrates that clear general differences exist in abortion beliefs. Following investigations of the abovementioned considerations, we suggest that further research may yield more precise insights into evolving views on abortion [11].

Contextual, contradictory, and, in some cases, changing beliefs on abortion make it difficult to accurately assess global and US abortion climates beyond rote and dichotomized categories [12]. However, evidence strongly suggests the US and global populations hold views that depart from these 2 categories, reflecting abortion attitude complexity [1,10]. Although survey data have quantitatively supported the idea of abortion attitude complexity—qualitative data, broadly defined as any type of open-ended text, audio, visual, or language data, may add additional nuance to suggest where and how complexity may emerge. For example, interviews about abortion reveal specific circumstances that contribute to variability in people's views on abortion [4] or reveal how current events and news cycles, in turn, shape social beliefs and attitudes [13]. Qualitative data can also inform how people contextualize assistance-related resources such as those found on social media.

Social media posts, as a novel form of qualitative data, may similarly reveal how people view abortion and the associated complexity of belief systems at a population-level scale. Notably, the inescapable role of social media in the public lexicon has evolved over time into an outlet for community building and information dissemination that can connect users over shared interests disregarding location [14]. For example, the Pew Research Center contends that more than three-quarters of the US adult population regularly use at least 1 social media platform [15], and half of all the users have actively maintained at least 1 account for more than a decade. Because social media data are part of the public domain, longitudinal tracking of such data can represent an open-access running diary of thoughts, perspectives, and affective indicators—particularly for issues deemed controversial or contentious, including COVID-19 vaccination status, marriage equality, transgender sports bans, and abortion [16,17]. Furthermore, social media data are also global, implying that shared languages, regardless of geographic constraint, can contribute to discourses about abortion and associated beliefs therein.

Research has documented that people use social media to share their opinions and views and engage in debates on various topics, as well as to seek help and information and solicit personal advice that pertains to their situation or to something they are going through in life [18-20]. These web-based interactions vary widely across social media platforms and topics but may include discussions about substance use disorders [21], mental health [22], sexual assault [23], and managing HIV treatment [24], among a wide range of other topics. Furthermore, some more limited research has explored social media users' engagement and interactions as part of sharing personal experiences, soliciting help, and requesting information pertaining to abortion. This research has focused particularly on assessing how social media users rely on each other to discuss cost-related barriers to abortion care [25], to discuss decision-making processes regarding abortion methods [26], and to seek support to make abortion decisions when they may lack familial and medical support otherwise [27].

Reddit (Reddit, Inc) is a social networking website, which is defined by its structure that allows users to subscribe to forums on diverse topics, both controversial and noncontroversial. Their approach to topic discussion is distinct from other social media platforms in that users can opt into conversations with variably different foci depending on needs and interests. For example, previous research has demonstrated that Reddit can serve as a social connection metric, information-sharing tool, and outreach resource [28] for controversial or contentious social topics, including sexual assault [29], abortion [30], and addiction and recovery [21]. For most, Reddit forums are a source of information on these topics. However, many of these same topics, particularly those with political contexts, can also be discussed on different Reddit forums in more social commentary or debate-style perspectives. Abortion is one example of a contentious social topic with ranging subreddits pertaining to

XSL•FO
RenderX

different aspects of abortion, including as a social connection and information-sharing tool and debate platform.

Analyzing different facets of the same topic through various subreddits could yield nuanced aspects regarding crucial health topics unique from other quantitative and qualitative abortion research. Notably, as of December 2022, Reddit was the 20th most accessed website globally (sixth in the United States), and 50% of all Reddit users reside in the United States, with Canada, Australia, and the United Kingdom comprising approximately 20% of the total Reddit users. Reddit data can principally serve as a window into views on abortion in the United States; however, because not all English language data originate in the United States, it is also possible to observe abortion attitude complexity in a more Westernized, but global context or global reactions to news related to abortion in the United States.

Objectives

Advances in computational data mining have made it feasible to extract, analyze, and interpret these data en masse. This study used natural language processing (NLP) and data mining methods to identify and visualize latent themes across 2 distinct subreddits specific to abortion: r/Abortion and r/AbortionDebate. As a comparative study, we aimed to compare the semantic and content differences across these subreddits to gain a comprehensive social media portrait of abortion dialogue on Reddit. This study was guided by three research questions:

1. What themes emerge in a corpus of Reddit posts in r/Abortion, the largest subreddit dedicated to abortion social support and outreach?
2. What themes emerge in a corpus of Reddit posts in r/AbortionDebate?
3. What do similarities and differences by subreddits implicate regarding social media–derived beliefs and ideologies on abortion?

Methods

Data

Data for this study were collected over 5 months (ie, from January 2020 to May 2022) from the social networking website Reddit. Reddit represents an open network of communities where users can engage and connect with others over shared interests, hobbies, or personal experiences. Unlike other popular social media websites used for computational analyses, including X (X Corp, formerly known as Twitter), Reddit is unique in that users can create specific channels to form communities with other interested parties on diverse issues or topics. These channels, otherwise known as subreddits, comprise people with shared identities who find, subscribe to, and post within these channels. For instance, people interested in gaming can join the r/Gaming subreddit and people with depression can join the r/Depression subreddit.

We leveraged the PRAW (Python Reddit Application programming interface Wrapper) [31], a third-party application programming interface (API), to collect data for this study and specifically to isolate and download content posted into subreddits in English germane to abortion—we queried the API to allocate similar subreddits also spanning abortion-related topics. This query returned 1 additional subreddit: r/AbortionDebate. Given observable differences in framing (ie, people's abortion experiences vs debates about abortion), we included this subreddit in our study as an additional but mutually distinct unit of analysis; that is, we collected and stored data for r/Abortion and r/AbortionDebate as separate corpora intended for separate analyses. All data collected for this study were in English, which we selected for 2 reasons: first, >70% of all Reddit users originated from English-speaking countries, and second, at the time of data collection, Reddit posts originating in languages other than English were insufficient for analysis. In Spanish, for example, r/Aborto contained only 5 members, with no activity since 2019; similarly, we observed <50 Spanish-language posts in either r/Abortion or r/AbortionDebate.

Once we identified our subreddits of interest, we queried the API to collect new posts and top posts from the r/Abortion and r/AbortionDebate subreddits. After filtering our data for duplicates and accounting for API data scraping limits, our final sample size comprised 4966 posts, divided into 2 corpora: 56.68% (2815/4966) of r/AbortionDebate posts and 43.31% (2151/4966) of r/Abortion posts.

Analyses

We aimed to use NLP to identify salient categories in the r/Abortion and r/AbortionDebate subreddits. In numerous studies, latent Dirichlet allocation (LDA) topic models have been predominantly used for this purpose. LDA is a well-regarded unsupervised probabilistic model that evaluates word co-occurrence patterns using an iterative Gibbs sampling method [32]. Although LDA is often considered the gold standard within many academic and professional communities, advancements in NLP, artificial intelligence, and neural networks have introduced innovative topic modeling methods that can more closely approximate the potential meaning in these categories [33].

For this study, we applied one such advancement, the Bidirectional Encoder Representations from Transformers (BERT) topic modeling tool, BERTopic. BERTopic is an NLP topic modeling approach used to identify latent themes or topics within a collection of interrelated documents [34]. Unlike LDA, which uses probabilistic modeling to identify latent topics, BERTopic leverages pretrained embeddings from one of many transformer models, a type of neural network architecture in which an input sequence is compared against large-scale language models to calculate embeddings [35]. Embeddings are used to convert unstructured data, including words and sentences, into fixed-length continuous vectors. These vectors enable mathematical operations to capture semantic meanings, relationships, and other properties related to natural human language.

The vectors calculated using this approach tend to be highly dimensional and difficult to interpret. To reduce dimensionality while maintaining the integrity of our data, we applied a principal component analysis, which is commonly applied in NLP approaches for general dimensionality reduction purposes [36]. This analysis allowed us to extract and more easily interpret a range of possible clusters or topics in both the

XSL•FO
RenderX

r/Abortion and r/AbortionDebate subreddit data. Once we reduced the dimensionality of our vectors, we applied a Hierarchical Density-Based Spatial Clustering of Applications with Noise to identify latent clusters or topics [37], CountVectorizer to tokenize each topic, and class term frequency–inverse document frequency to extract topic words for each cluster [38].

Furthermore, to gauge the emotional tone or mood represented in each post from the studied corpora, we applied a Valence Aware Dictionary and Sentiment Reasoner (VADER), a rule-based sentiment analysis tool [39], and Text2Emotion, a rule-based emotion analysis tool [40]. VADER sentiment analysis is an algorithm and analysis that examines the polarity of words within each social media post. Posts are fed through a lexicon or web-based dictionary, which is precoded with values for all positive and negative words in the English language. When posts are run through the VADER lexicon, they receive a composite score. Negative VADER values denote lower affect (ie, −0.99 to −0.01), and positive values denote higher affect (ie, 0.01 to 0.99). Although an older tool, VADER is commonly used to assess content affect and emotional affect. In contrast, the Text2Emotion tool for emotion analysis scans each entry for key phrases and terms denoting one of four base emotions: (1) happy, (2) surprise, (3) fear, and (4) sadness. Collectively, these 2 tools can identify potential tonal differences in each post, again implicating the different uses of each subreddit included in the analysis. Both tools have been applied extensively in computational public health studies owing to their ease of access, replicability, and numerous validation studies [16,21,41,42].

Procedure

Our workflow is depicted in Figure 1. First, we queried the Reddit API to archive top and new posts from the r/Abortion and r/AbortionDebate subreddits. Data collected from the r/Abortion and r/AbortionDebate subreddits were saved as separate data files. After removing duplicate and non-English posts in either data file, we applied standard preprocessing steps to remove parts of speech that would detract from the clarity of our models, including articles, prepositions, punctuation, abbreviations, and numbers [43]. Once the data were cleaned, we tokenized our data at the sentence level before calculating embeddings. Once the data were preprocessed and tokenized, we proceeded with our BERTopic pipeline. First, to calculate embeddings in our data we applied all-MiniLM-L6_v2 [44], a transformer-based model developed by Microsoft Corp. This model is designed to be a smaller and more efficient transformer model than larger models, including a generative pretrained transformer or T5, which may make it more appropriate for smaller data sets; however, more research is needed to confirm this notion. Once we calculated embeddings for all sentences in each corpus, we applied a principal component analysis to reduce dimensionality in our data, retaining 5 components. We then ran an iterative topic model ranging from 10 to 80 topics and calculated coherence scores [45] to identify an optimal number of topics, retaining a topic solution with the highest coherence score. For both r/Abortion and r/AbortionDebate, the optimal solution was 10 topics, yielding a respective coherence score of 0.42 and 0.35, which indicates a marginal fit. After we extracted key terms per topic, we applied a sorting function to examine key terms in each entry. Each entry was then classified into one of 10 possible topics in either corpus. We lastly performed a VADER sentiment analysis and Text2Emotion emotion analysis for each entry in both corpora.

XSL•FO
RenderX

Figure 1. Workflow depicting the study's methodological pipeline. API: application programming interface; cTF-IDF: class term frequency–inverse document frequency; HBDSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise; PCA: principal component analysis; VADER: Valence Aware Dictionary and Sentiment Reasoner.

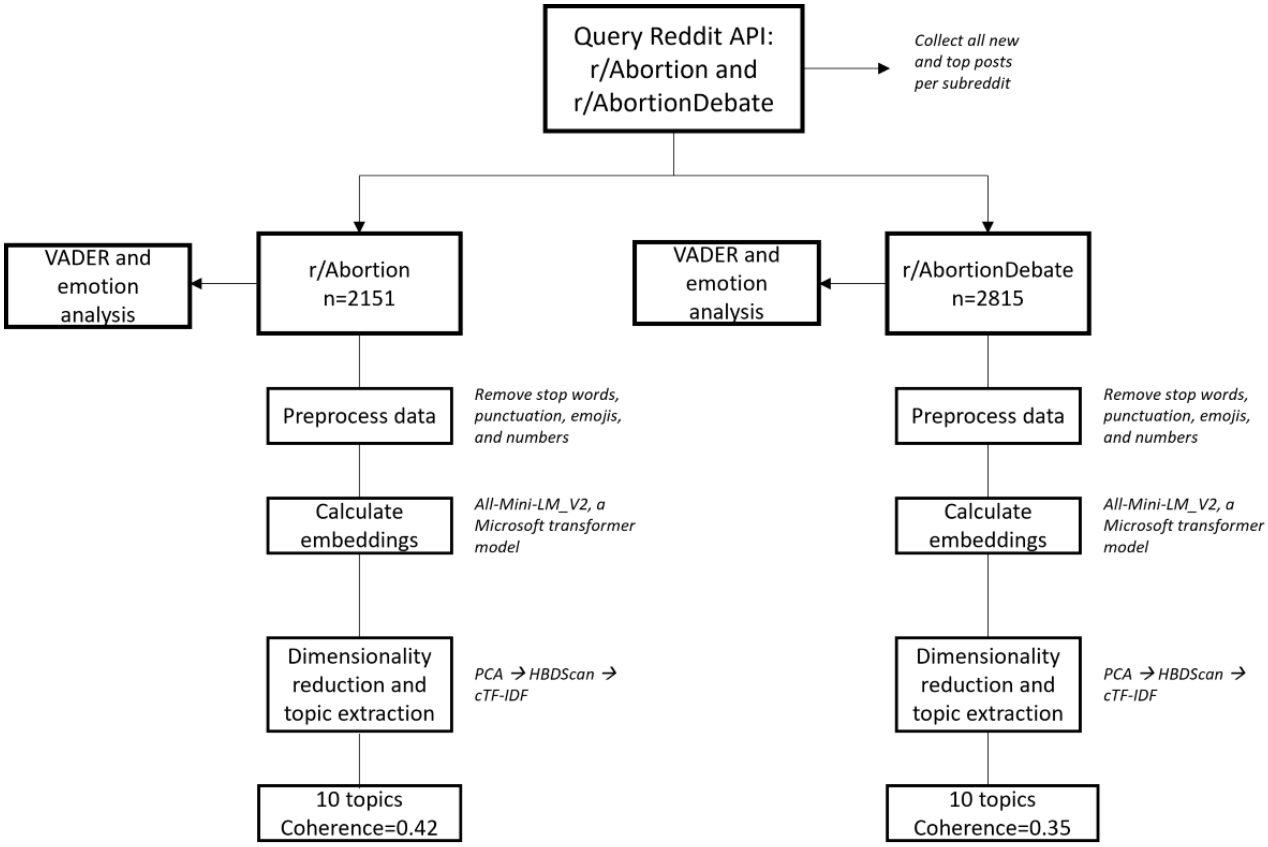

Ethical Considerations

This study involved a secondary analysis of deidentified and anonymized Reddit posts collected between January and May 2022. As this was an observational study with no contact between human subjects and no possible way to trace posts to any individual author, this study was exempt from Institutional Review Board review.

Results

Our study applied computational tools to collect and analyze subreddits specific to abortion. We aimed to examine how abortion was discussed on the social media platform Reddit, both as an information-sharing tool and as a platform for debating worldviews.

What Themes Emerged in a Corpus of Reddit Posts About r/Abortion, the Largest Subreddit Dedicated to Abortion Social Support and Outreach?

Our coherence score analysis indicated a 10-topic solution for the r/Abortion subreddit. Table 1 outlines each topic by keywords, the number of sentences belonging to each topic, and the percentage of each topic relative to the larger corpus. Names for each topic were derived by reviewing a small excerpt of Reddit data that were sorted into one of 10 topics by a sorting function using keywords.

The r/Abortion subreddit analysis revealed numerous ways in which abortion was discussed in a social support context. The most prominent topic of our study, *topic 1: sharing support*, comprised the bulk of the conversation with >18% total representation. Social support was commonly manifested by people sharing their own experiences with abortion or by friends and family members who may have experienced abortion. This was further evident by multiple topics containing information-sharing content: *abortion experience* (topics 3, 7, and 10). Beyond social support, several of our topics also appeared to discuss abortion in a neutral and educational information-sharing context: *general abortion* (topic 5) and *general pregnancy* (topic 8).

Further review of the topics added context to our findings. Table 2 provides a summary of each topic and the key excerpts that denote additional meaning. As shown in Table 2, there was little content indicative of debate or questioning one's position on abortion. Instead, we observed personal experience sharing, including narrative accounts of one's experience with abortion generally, miscarriage, and medication abortion specifically.

XSL•FO
RenderX

Table 1. Topic names and key terms pertaining to r/Abortion subreddit posts (n=2151).

Index	Name	Key terms	Sentences (n=42,732), n (%)
1	Sharing support	want, wanted, told, didnt, say, asked, said, talk, okay, dont	7970 (18.65)
2	Postabortion emotion	feel, sadness, feeling, depressed, regret, depression, emotionally, guilt, grieving, emotional	4502 (10.54)
3	Abortion experience	slept, pad, bed, hour, rest, test, waited, room, home, minute	4361 (10.21)
4	Social support	therapy, supportive, support, struggling, advice, experience, situation, care, family, talked	4215 (9.86)
5	General abortion	abortion, pregnancy, aborted, miscarriage, fetus, pregnant, abort, parenthood, birth, pro-choice	4093 (9.58)
6	Clinic experience	appointment, clinic, pill, prescription, sedation, patient, consultation, doctor, iud, obgyn	3770 (8.82)
7	Abortion experience	anesthesia, sedation, anxiety, painkiller, pain, ibuprofen, relief, anxious, painful, uncomfortable	3584 (8.39)
8	General pregnancy	pregnancy, pregnant, gestational, miscarriage, condom, positive, placenta, fetus, parenthood, pill	3441 (8.05)
9	Abortion alternatives	pregnant, pregnancy, adoption, child, parent, selfish, baby, kid, mother, marriage	3433 (8.03)
10	Abortion experience	cramp, crampy, cramping, bleeding, bleed, clot, ibuprofen, spotting, period, nausea	3363 (7.87)

XSL•FO
RenderX

Table 2. r/Abortion subreddit summaries with example excerpts.

Topic name	Summary	Excerpts
Sharing support	This cluster contained diverse content pertaining to lending and sharing abortion-related support.	• "I just wanted to say that I am thinking of you during this time" • "Do not let the extremist propaganda get to your head." • "This subreddit (r/Abortion) is a (mostly minus the trolls) community."
Postabortion emotion	This cluster contained content pertaining to sharing emotions associated with abortion decision-making and postabortion procedures.	• "I know I made the right decision; but I will always grieve you." • "I just want to vent my frustrations and emotions here." • "I still have my moments of sadness but overall, I would make the same decisions again if I needed to."
Abortion pre- and postexperience	This cluster contained content pertaining to pre- and postabortion experiences, including decision-making and outcomes.	• "When I waited I could not believe the amount of protestors outside." • "I'm panicking in my dreams and then I wake up and realize I still have six days to go." • "When she (my girlfriend) found out two weeks ago, she was around 11 weeks along"
Social support	This cluster contained content pertaining to social support related to abortion care.	• "I would like to offer a free ride to such locations (counselors) for anyone in need." • "I also want to thank the volunteer escorts who helped me to and from my car." • "If you are a volunteer escort, your actions and help are so important."
General abortion	This cluster contained general abortion content, usually in a supportive context.	• "Women who seek abortions are often vilified whereas women who give birth are celebrated." • "Abortion is health care, abortion is birth control." • "Since my abortion I have graduated high school, graduated an associates program, a bachelors program, and am in grad school."
Clinic experience	This cluster contained content pertaining to individual experiences at abortion care and general health care facilities.	• "My ultrasound was this morning and everyone at the clinic was rude." • "I had a good experience at the clinic, the staff were very considerate and kind." • "At the location I went to they had setup a fake clinic across the street and bought the house next door."
Abortion experience	This cluster contained content pertaining to pre- and postabortion experiences, including anticipating an abortion procedure.	• "However the protestors yelling at you while you walk into the clinic is ridiculous." • "I am also terrified about the pain- they do not offer anesthesia, just pain meds." • "I ended up hemorrhaging pretty badly and had to be taken to the hospital via ambulance."
General pregnancy	This cluster contained content pertaining to general pregnancy discussions, including accidental pregnancy and miscarriage.	• "10 weeks into the pregnancy, I miscarried." • "15 years later and I intentionally became pregnancy." • "Six months later, I intentionally stick my foot through the pregnancy door again."
Abortion alternatives	This cluster contained diverse content pertaining to abortion alternatives.	• "I considered adoption, but abortion was right for me at that time." • "Forced births are not the answer." • "Can anyone logically think of true alternatives to abortion?"
Abortion experience	This cluster contained content pertaining to abortion experiences through a medication abortion lens.	• "Did you experience different side effects, was the bleeding/clots different?" • "My paper says tramadol every 2 hours, ibuprofen every 6 hours, and Tylenol every 4." • "It took me about 16 hours after the misoprostol to start bleeding and it was extremely heavy and clotty."

Perhaps one of the most recurring patterns in our data was frank discussions about postabortion feelings in a clinical setting, ("I felt so nauseous in that waiting room, I was not sure I could go through with it"), a postabortion setting ("It took me a few days to finally feel like myself again post-abortion"), or a medication abortion context ("The mifepristone caused some pretty intense clotting after I took the pill"). The medication abortion narratives were sometimes framed as someone explaining their decision ("I chose the pill because where I live you cannot have someone with you when getting an abortion due to COVID-19").

What Themes Emerged in a Corpus of Reddit Posts in the r/AbortionDebate Subreddit?

Our coherence score analysis indicated a 10-topic solution for the r/AbortionDebate subreddit. Table 3 outlines each topic by keywords, the number of sentences belonging to each topic, and the percentage of each topic relative to the larger corpus. Names for each topic were derived by reviewing a small excerpt of Reddit data that was sorted into one of 10 topics by a sorting function based on keywords.

Unlike the r/Abortion subreddit, which we determined seemed to be used in a social support and information-sharing context,

XSL•FO
RenderX

the r/AbortionDebate subreddit comprised conversations dedicated to critically assessing abortion from legal, moral, and ethical perspectives. The topic with the greatest representation was *topic 1: Reddit forum rules and regulations*. In topic 1, we observed several posts directly from moderators explicitly warning against outright attacks, misinformation, and vitriol targeted at people with opposing views on abortion; this topic was absent completely in the r/Abortion topic model. The second most prominent topic, *topic 2: abortion morality*, was centered on debating abortion from a moral perspective. The topic with the smallest representation was topic 5, pertaining to *general pregnancy*. At face value, we did not observe a great overlap in topic content in the r/AbortionDebate subreddit compared with the r/Abortion subreddit. However, we reviewed additional excerpts to ascribe a deeper meaning to these topics to examine precisely how abortion debates manifested on these forums.

Table 4 outlines additional information about each topic, including a summary and key excerpts that implicate deeper meaning. This additional analysis allowed us to examine more precise moral, legal, and ethical arguments pertaining to people's expressed views on abortion.

Table 3. Topic names and key terms pertaining to the r/AbortionDebate subreddit posts (n=2815).

Index	Topic name	Key terms	Sentences (n=30,031), n (%)
1	Forum rules	post, rule, problem, posted, following, response, sorry, line, mod, issue	5847 (19.47)
2	Abortion morality	morality, debating, discussion, arguing, debate, argue, reasoning, argument, premise	3253 (10.83)
3	Abortion legality	law, enforcement, violence, consent, crime, autonomy, consenting, freedom, murder, bodily	3240 (10.79)
4	Personhood	abortion, parenthood, fetus, unborn, pregnancy, aborted, contraception, maternal, abort	2839 (9.45)
5	General pregnancy	prenatal, pregnancy, childbirth, fetus, maternal, pregnant, fetal, gestational, placenta	2754 (9.17)
6	Abortion debate	abortion, debate, fetus, prolife, argument, arguing, morality, advocate, prolifers	2700 (8.99)
7	Humanity and life	human, humanity, biological, personhood, sentient, moral, consciousness, concept, brain	2573 (8.57)
8	Euthanasia	euthanasia, killing, death, dying, murder, lethal, suicide, kill, die, killed	2325 (7.74)
9	Abortion alternatives	pregnancy, parenthood, contraception, abstinence, pregnant, contraceptive, fetus, unborn	2320 (7.72)
10	General pregnancy	fetus, foetus, unborn, fetal, conception, womb, embryo, pregnancy, childbirth, pregnant	2185 (7.27)

XSL•FO
RenderX

Table 4. r/AbortionDebate subreddit summaries with example excerpts.

Topic name	Summary	Excerpts
Forum rules	This cluster contained content from moderators reminding posters to mind rules and regulations.	• "As a courtesy, be respectful to others." • "Rules on this forum are non-negotiable." • "We will remove content that does not ascribe to our rules."
Abortion morality	This cluster contained diverse content on abortion morality, including acceptability and permissibility.	• "Title is kind of a joke but the question is serious: I'm wondering what you all think are the moral arguments for your 'side' of the debate." • "Do you have a clear point where you draw the line?" • "Real-time discussion will hopefully lead to better overall conversation about right versus wrong."
Abortion legality	This cluster contained content on abortion laws, legality, and future legal maneuvers to restrict abortion.	• "Creating a law that only affects 50% of the population is discrimination no matter how you look at it." • "Similar measures have in the past been filed by state Rep. Tony Tinderholt, R-Arlington, who received death threats and was placed under the protection of the Texas Department of Public Safety after he introduced the bill in 2017." • "And the Texas law is going to be copied across the country."
Personhood	This cluster contained discussions about personhood and pregnancy.	• "In early stages, can we really say a zygote is a person?" • "These laws are not restrictive to the fetus, if you think about it?" • "But really, at what point is abortion the murder of a person?"
General pregnancy	This cluster contained diverse content pertaining to pregnancy, including personal experiences and general facts.	• "The average newborn head is 11.1 cm in diameter." • "The first week after conception is pretty much taken up with travel plans and implantation." • "Within 8 years teen pregnancy dropped 54%."
Abortion debate	This cluster contained content pertaining to general abortion debates and dialogues.	• "I agree that abortion is murder but it is not okay to harass women outside planned parenthood!" • "I disagree that abortion is murder, but there is a limit." • "It's important to have an honest conversation about abortion to understand all sides."
Humanity and life	This cluster contained content pertaining to humanity and life with some insights into religion and spirituality.	• "There is no sin in keeping your religious beliefs to yourself while allowing others their right to choose." • "I know most of you right-wing Christians work hard for a living." • "To me an essential part of being a human being is that that human being is something/someone I can hold in my hands."
Euthanasia	This cluster contained content that used euthanasia as a talking point when discussing abortion beliefs.	• "Euthanasia/assistant suicide the same, the military most people go into that with the idea that they might have to kill a person, so I think that check out." • "I meet and argued with a lot of pro-lifers, I brought up both the death penalty and euthanasia!" • "Not once can I recalled a time where a pro-lifer have use it's a separate dna to attack the death penalty or euthanasia."
Abortion alternatives	This cluster discussed content that implicated the complexity of alternatives to abortion, including adoption and the foster system.	• "The foster care and adoption systems are already overfilled and refusing to accept anyone else." • "You could just put it up for adoption. That'd be fine if the problem was just parenthood, but it's not." • "Which when the woman don't want the offspring for various reasons (personal cost of altruism), they decide the personal cost of the act of parental altruism is too high."
General pregnancy	This cluster contained content that used pregnancy as a discussion point about abortion rights, morality, and legality.	• "Unborn humans literally live inside of someone else." • "On a post like a week ago I read that some PC folks are frustrated with the test tube of embryos vs infant in a burning building argument." • "If you are pro-choice, do you also adopt the anti-natalist perspective?"

For example, we observed that the abortion morality topic typically contained content related to drawing lines about abortion permissibility ("Where do [people] draw the line between acceptable and not acceptable"). This style of discussion was mirrored in conversations about fetal personhood ("Who here honestly believes a zygote is a person with rights?") and the role spirituality plays in moral arguments about abortion ("But what do Catholics really think on this issue?"). Discussions and arguments about abortion morality were notably similar to the content in *topic 3: abortion legality*. Content on this topic typically discussed new abortion-related laws, the merits of those laws, and opinions about their relative

XSL•FO
RenderX

effectiveness ("Texas passed a very restrictive law and it will serve as a benchmark for other states, watch"). Importantly, and across topics, we observed that people declared their abortion views ("I am pro-choice and I will always be") and, in some cases, discussed how their abortion views evolved over time ("I am pro-life, but we should be discussing the merits of abortion as a life-saving tool here"). Here, we observed more opinions than the outright support articulated in the r/Abortion subreddit.

What Do Similarities and Differences by Subreddit Implicate About Social Media–Derived Abortion Beliefs and Ideologies?

Figure 2 visually represents data from each subreddit, where dense, overlapping clusters signify similar topics (or higher collinearity) and nonoverlapping circles indicate dissimilar topics (or lower collinearity).

In both the r/Abortion and r/AbortionDebate subreddits, the intertopic distance maps depict mutual exclusivity in general abortion and pregnancy topics, distinguished by basic sharing of language and specific information related to pregnancy and abortion ("Sometimes a pregnancy can end without warning or reason"; "abortion is a women's health issue"). Beyond these statements, however, other conversations exhibit a richer and more nuanced discourse about abortion, overlapping between topics and offering deeper insights into an individual's worldview on abortion, and portraying how various co-occurring factors influence one's beliefs and worldviews ("Laws are one thing but have you considered the humanistic side of it all?").

We used VADER and Text2Emotion tools to discern affective differences between r/Abortion and r/AbortionDebate subreddits. The r/Abortion subreddit displayed a compound VADER score of 0.10, reflecting overall neutral content, whereas the r/AbortionDebate subreddit displayed a score of −0.06, denoting neutral to slightly negative content. The emotion analysis findings for the r/Abortion subreddit were as follows: *happy* (mean 0.06, SD 0.19), *angry* (mean 0.20, SD 0.31), *surprise* (mean 0.12, SD 0.26), *sad* (mean 0.20, SD 0.31), and *fear* (mean 0.36, SD 0.39). The emotion analysis findings for r/AbortionDebate subreddit were as follows: *happy* (mean 0.12, SD 0.27), *angry* (mean 0.05, SD 0.18), *surprise* (mean 0.11, SD 0.25), *sad* (mean 0.22, SD 0.35), *fear* (mean 0.28, SD 0.35).

Figure 2. Intertopic distance map indicating relative similarity and difference per topic per corpus. D1 and D2 refer to the 2 retained components for visualization purposes and are not meant to be interpreted.

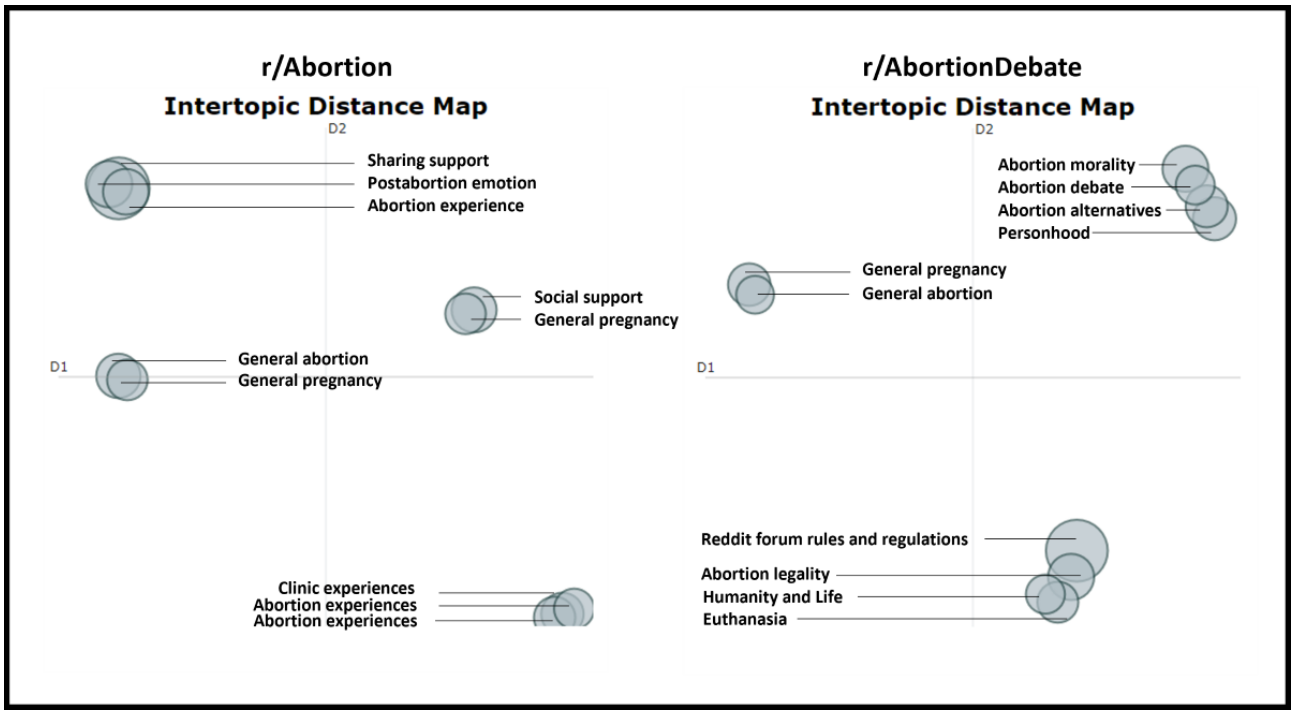

Furthermore, the Text2Emotion variable *fear* was prominent in r/Abortion, whereas *happy* was slightly more elevated in r/AbortionDebate. These observed differences are likely attributed to the differing nature and scope of the subreddits. For instance, the manifestation of *fear* may be more related to personal abortion narratives in r/Abortion, whereas happiness may arise from occasional friendly exchanges of views in r/AbortionDebate.

Despite their different foci, both subreddits contain myriad conversation topics, allowing for civil and enlightening discussions on evolving abortion views and ideologies. The discourse in these forums sometimes hints at the evolution of individual ideologies with time, reflecting the dynamic nature of personal beliefs and the influences shaping them (adapted excerpt: "I guess I just don't know my views"; excerpt: "My opinion changed over time, growing up in a Christian household I was always against abortion...until I needed one myself"). This phenomenon underscores the essential role of such platforms in fostering understanding and dialogue on the multifaceted issue of abortion.

XSL•FO
RenderX

Discussion

Our study leveraged Reddit data as a novel, big data form of qualitative data to examine abortion discourse on r/Abortion and r/AbortionDebate subreddits. We observed several important themes, including evidence of complexity in abortion-related social media posts, which warrant further discussion.

The r/Abortion Subreddit as an Information-Seeking or Information-Sharing Platform for People With Questions About Their Abortion Experiences

Within the r/Abortion subreddit, we noticed posters using this platform to discuss abortion in diverse, sometimes overlapping contexts. However, each topic emerging from r/Abortion typically involved a degree of information sharing, whether through the provision of available resources or sharing personal narratives and experiences with abortion. We primarily observed these types of posts in *topic 1: sharing support*; *topic 2: postabortion emotions*; *topics 3, 7, and 10: abortion experience*; and *topic 5: clinical experience*. The content within these topics typically involved direct sharing of one's experiences related to abortion or posing highly specific questions about access (eg, excerpt: "Is abortion legal past 6 weeks gestation in Oklahoma?") and medication abortion (eg, excerpt: "Abortion is legal here; can I get abortion pills by mail?"). Within the medication abortion topic, the content was both informative and supportive, with some posters sharing their experience in solidarity with others facing a similar choice. Notably, we did not observe any critiques against anyone's abortion narratives; rather, the tone and structure, as also evident in this study's VADER and emotion analysis, are largely informative and overall supportive of abortion. Given that the rules and guidelines established this subreddit as a place of nonconfrontational discussion, perhaps people advocating for other reproductive choices may have shared their perspectives in other subreddits, such as those related to adoption.

We acknowledge the possible connection between personal tendencies to share intimate information and the continually evolving role of the internet as a medium for social connection and information acquisition [46]. Notably, for the past 3 decades, the internet has become the most influential medium for information-seeking globally. The Pew Research Center indicates that approximately 80% of the adult population in the United States regularly use the internet to acquire general information or understand unfamiliar topics [47]. For example, an individual contemplating an abortion might opt to seek guidance in web-based forums to avoid potential ostracism from friends and family. Similarly, a friend or family member of someone considering an abortion might turn to web-based forums to secure advice or perspectives on assisting their loved one. Discourse on such platforms is crucial, especially when addressing sensitive topics that many may feel uneasy to discuss openly. This emphasizes the significance of the internet as a confidential and reliable resource for information and advice. Importantly, this also supports Reddit as a source of information for people needing abortion-related counseling.

These excerpts, and others in our composite sample, illustrate that social networking websites serve as a potentially crucial source of information for some [48], offering insights and details that may be otherwise unavailable, including local and state resources for abortion. This finding becomes particularly salient in light of the overturning of *Roe v. Wade*, which marked the end of federal protections for abortion until viability [49]. In the wake of this decision, 24 states enacted bans with limited exceptions or additional restrictions on abortion—generally earlier in terms of weeks of gestation than previously occurring under *Roe v. Wade* [50]. For those residing in states where abortion transitioned from being broadly legal to almost entirely illegal, web-based resources may have played a pivotal role during instances of unplanned pregnancy, as observed previously [51]. Further research is imperative to assess the efficacy of Reddit and other social networking sites in offering support and resources on this and other health-related topics. Notably, this subreddit contained little to no expression about personal abortion beliefs and ideologies.

The r/AbortionDebate Subreddit and Discussions of Abortion Identity and Changing Views Over Time

We did not observe much information or support sharing in the r/AbortionDebate subreddit. Rather, content in this subreddit discussed values and beliefs about abortion across many domains, including ethical, moral, legal, and humanistic. In several circumstances, we observed complex and nuanced abortion perspectives that do not correspond neatly to prochoice or prolife frameworks—2 commonly used but contested abortion identity labels used to outline personal abortion beliefs. For example, as many as half of the topics uncovered by r/AbortionDebate contained contradictory expressions regarding abortion and how the abortion debate was framed. These posts were broadly delineated as those deconstructing or debating prochoice and prolife movements and others explaining how circumstances contributed to moral and ethical shifts in abortion views, for example, in the following excerpts: "I was and will always be pro-choice, but my reaction was absolutely not [to abort a fetus with serious birth defects] even though I knew it was the right answer" and "I was pro-life and never thought I'd need Planned Parenthood until I did. My experience changed my opinion of them, but [I still wish] they didn't primarily exist to perform abortions." Here, the emphasis is far less on information or support sharing, rather the purpose is to articulate personal views about abortion and defend them accordingly. These findings align with ongoing abortion attitude research citing complex or nuanced abortion views that do not neatly fit into a singular label [52-54].

In addition to discussing and debating abortion values, we observed more combative content in the r/AbortionDebate subreddit. This is likely by design, namely to parse out people seeking information about abortion versus people looking to debate abortion [55]. Such differences between the r/Abortion and r/AbortionDebate subreddits were particularly evident in our sentiment and emotion analyses. For example, r/AbortionDebate yielded slightly more negative VADER affect scores and decreased emotion analysis scores for *fear*. We attributed more negative VADER scores to the often contentious exchanges among users (excerpt: "All these pro-choicers in here trying to lump as all as anti-women bigots"). We attributed lower *fear* scores to the apparent use of r/AbortionDebate as a

XSL•FO
RenderX

forum to discuss abortion views and not for sharing information or narrative accounts about abortion. In other words, negative language was reflected via discourse in the r/AbortionDebate subreddit, as opposed to expressing personal fears or concerns about abortion, which may have surfaced more in the r/Abortion subreddit. In this context, the r/AbortionDebate subreddit may be more useful for mining insights into abortion ideologies, particularly when examining precise factors about abortion, including moral and legal arguments, gestational limits, and others. However, to gain insights into how abortion, as a medical procedure, is communicated from a decision-making perspective, r/Abortion may be more informative.

We identified 2 main implications from the content differences observed in r/Abortion and r/AbortionDebate. First, opting for the right Reddit forum is critically important. Reddit's structure—where users select forums based on interests or needs—is different from other social networking sites. For people looking for ideally accurate, impartial information about abortion, r/Abortion or similar subreddits are suitable. Meanwhile, r/AbortionDebate is better for those wanting to discuss and ponder the ethical aspects of abortion. However, this choice is dependent on knowing how Reddit works. We project that a significant proportion of people may join the wrong forum and get exposed to unintended outcomes and viewpoints owing to a lack of preexisting knowledge about Reddit and its operations. Second, our observations support the idea that Reddit's higher moderation levels make it a valuable tool for social science research. Historically, Reddit has carried the reputation of fostering trolls and hate speech. However, for health content, subreddits tend to be more effectively moderated by content experts. As evidenced in our data, both subreddits seemed relatively free from hate speech and trolling because of this moderation, which is unique to Reddit compared with other social media platforms. Therefore, Reddit remains a fairly reliable platform for both users and researchers, especially in the wake of recent changes in APIs and data access on other platforms, including X (X Corp, formerly known as Twitter).

Social Media as a Resource and Triangulation Tool to Support Ongoing Quantitative and Qualitative Research on Abortion

Our findings, particularly those critiquing abortion identity labels or people explaining their contextual abortion beliefs, support extant research demonstrating that people's attitudes toward abortion are complex. Notably, this larger body of research argues that abortion attitudes are not unidimensional or polar but rather vary along legal, moral, social, and other similar domains [2,3,56,57]. This work is composed of both quantitative (surveys) and qualitative (interviews) data collections, which collectively yield deep insights into social attitude formation in the United States and how beliefs vary based on context and other dimensions. Consistent with these studies, our results support the notion that abortion attitudes and abortion decision-making are not unidimensional but involve multiple co-occurring considerations.

The novel nature of social media as data adds additional validity to previous abortion attitude research. This is particularly salient regarding how our findings triangulate or corroborate previous research on abortion attitude complexity. Notably, by mining Reddit abortion forums, we observed at least two principal uses of these forums: (1) as a space to share narratives and resources about abortion and (2) as a dedicated channel to debate abortion views. For many, Reddit forums could be a place where some people feel comfortable sharing or debating abortion views, although we acknowledge that more research on this area is needed. Furthermore, Reddit offers a somewhat anonymous space where people can gather the information they need about abortion or inform their perspectives on abortion. These shared Reddit perspectives, which are generally top of the mind, spontaneous, and unprompted [58], may provide a window into collective abortion beliefs that support or refute previous findings from other conventional forms of data collection. Similar uses of social media data, namely to corroborate findings on social issues, including gun control [59], marriage equality [60], and vaccination mandates [61], have been similarly leveraged. Therefore, we argue that social media can be a valuable source of data to help elucidate people's opinions on relevant social issues.

Furthermore, we argue that national surveys, strategic qualitative interviews, and mass social media scrapes as data sources yield specific outcomes that, when combined, provide a robust and comprehensive portrait of social issues. Survey data, which are strengthened when participants are identified via probability-based sampling protocols [62], reveal nationwide associations between demographic variables and other variables of interest. Qualitative data can reveal insights into highly specific research questions, for example, whether changing auxiliary verbs leads to diverging responses about abortion beliefs [63]. Social media data scrapes can offer population-level insights that support or contradict findings from previous studies at the population-level scope and scale [41]. Our Reddit data support previous findings from surveys and qualitative research, demonstrating how social media data can serve as a triangulation tool. We contend that further strategic applications of social media mining with traditional quantitative and qualitative research can provide highly accurate portrayals of social views in the United States.

Limitations and Future Research

This study has several limitations that we hope to address in future research. First, although Reddit posts can be construed as qualitative data, we did not perform a formal qualitative analysis using these data. Owing to the scope of this study, we instead leveraged NLP algorithms to categorize and visualize all data simultaneously. In the future, researchers could perform detailed qualitative inquiries with these data, which can occur with the entire data set or among one or several clusters depending on the scope and research questions. Second, our study was limited to exploratory analyses. Although more refined algorithms could more effectively annotate and classify our data, we believed that these approaches would better serve as a follow-up to our exploratory approach to mining Reddit data. Future studies should consider using our data for more refined machine learning–driven or artificial intelligence-driven tasks. Finally, our study was limited by its relatively small timeframe (5 months). It is likely that collecting data for an even longer period may have yielded more nuanced findings.

XSL•FO
RenderX

Conclusions

With the decision in *Dobbs v. Jackson Women's Health Organization* overturning *Roe v. Wade*, there is renewed attention to abortion as a contentious political and social issue. Despite abortion being an exceedingly complex topic, political debate and discussions about abortion are generally framed dichotomously as a support or opposition, or prolife or prochoice issue. However, extensive research indicates that public opinion about abortion does not ascribe neatly to that dichotomy and that circumstances beyond a person's control may lead to shifts in views of abortion over time. Our research corroborates such findings that detail the myriad ways in which abortion attitudes are complex and contextual, beyond simple information-seeking. Furthermore, our findings provide evidence that social media data can be a helpful triangulation tool for public opinion survey research.

Data Availability

The data are currently stored in a secure GitHub repository and are available for further analysis upon request.

Conflicts of Interest

None declared.

References

1. Jozkowski KN, Crawford BL, Hunt ME. Complexity in attitudes toward abortion access: results from two studies. Sex Res Soc Policy 2018 Mar 10;15(4):464-482. [doi: 10.1007/s13178-018-0322-4]
2. Jozkowski KN, Crawford BL, Turner RC, Lo WJ. Knowledge and sentiments of Roe v. Wade in the wake of justice Kavanaugh's nomination to the U.S. Supreme Court. Sex Res Soc Policy 2019 May 31;17(2):285-300. [doi: 10.1007/s13178-019-00392-2]
3. Jozkowski KN, Crawford BL, Willis M. Abortion complexity scores from 1972 to 2018: a cross-sectional time-series analysis using data from the general social survey. Sex Res Soc Policy 2020 Mar 09;18(1):13-26. [doi: 10.1007/s13178-020-00439-9]
4. Maier JM, Jozkowski KN, Valdez D, Crawford BL, Turner RC, Lo WJ. Applicability of a salient belief elicitation to measure abortion beliefs. Am J Health Behav 2021 Jan 01;45(1):81-94. [doi: 10.5993/AJHB.45.1.7] [Medline: 33402240]
5. Hans JD, Kimberly C. Abortion attitudes in context: a multidimensional vignette approach. Soc Sci Res 2014 Nov;48:145-156. [doi: 10.1016/j.ssresearch.2014.06.001] [Medline: 25131281]
6. Crawford BL, LaRoche KJ, Jozkowski KN. Examining abortion attitudes in the context of gestational age. Soc Sci Q 2022 May 16;103(4):855-867. [doi: 10.1111/ssqu.13157]
7. Smith TW. An evaluation of Spanish questions on the 2006 general social survey. NORC/University of Chicago. 2007 Mar. URL: https://gss.norc.org/Documents/reports/methodological-reports/MR109%20An%20Evaluation%20of%20Spanish%20Questions%20on%20the%202006%20General%20Social%20Survey.pdf [accessed 2024-01-29]
8. Bowman K, Goldstein S. Attitudes about abortion: a comprehensive review of polls from the 1970s to today. American Enterprise Institute. 2021 Nov 2. URL: https://www.aei.org/research-products/report/attitudes-about-abortion-a-comprehensive-review-of-polls-from-the-1970s-to-today/ [accessed 2022-07-21]
9. Doherty D. What can conjoint experiments tell us about Americans' abortion attitudes? Am Politics Res 2022 Jan 21;50(2):147-156. [doi: 10.1177/1532673x211053211]
10. Jelen TG, Wilcox C. Causes and consequences of public attitudes toward abortion: a review and research agenda. Polit Res Q 2016 Jul 02;56(4):489-500. [doi: 10.1177/106591290305600410]
11. Buyuker BE, LaRoche KJ, Bueno X, Jozkowski KN, Crawford BL, Turner RC, et al. A mixed-methods approach to understanding the disconnection between perceptions of abortion acceptability and support for Roe v. Wade among US adults. J Health Polit Policy Law 2023 Aug 01;48(4):649-678. [doi: 10.1215/03616878-10449896] [Medline: 36693179]
12. Friedersdorf C. There are more than two sides to the abortion debate. The Atlantic. 2021 Dec 10. URL: https://www.theatlantic.com/ideas/archive/2021/12/there-are-more-than-two-sides-to-the-abortion-debate/620978/ [accessed 2022-05-27]
13. Adamo C, Carpenter J. Sentiment and the belief in fake news during the 2020 presidential primaries. Oxf Open Econ 2023;2:odad051. [doi: 10.1093/ooec/odad051]
14. Milakovich ME, Wise JM. Internet technology as a global connector. In: Digital Learning. Cheltenham, UK: Edward Elgar Publishing; 2019. [doi: 10.4337/9781788979467.00008]
15. Perrin A. Social media usage: 2005-2015. Pew Research Center. 2015 Oct 8. URL: https://www.pewresearch.org/internet/2015/10/08/social-networking-usage-2005-2015/ [accessed 2024-01-29]
16. Bathina KC, Ten Thij M, Valdez D, Rutter LA, Bollen J. Declining well-being during the COVID-19 pandemic reveals US social inequities. PLoS One 2021 Jul 8;16(7):e0254114 [FREE Full text] [doi: 10.1371/journal.pone.0254114] [Medline: 34237087]
17. Zafarani R, Abbasi MA, Liu H. Social Media Mining: An Introduction. Cambridge, MA: Cambridge University Press; 2014. URL: http://www.socialmediamining.info/SMM.pdf

XSL•FO
RenderX

18. Jacques L, Valley T, Zhao S, Lands M, Rivera N, Higgins JA. "I'm going to be forced to have a baby": a study of COVID-19 abortion experiences on Reddit. Perspect Sex Reprod Health 2023 Jun 11;55(2):86-93. [doi: 10.1363/psrh.12225] [Medline: 37167095]
19. Priya S, Sequeira R, Chandra J, Dandapat SK. Where should one get news updates: Twitter or Reddit. Online Soc Netw Media 2019 Jan;9:17-29. [doi: 10.1016/j.osnem.2018.11.001]
20. Ong E, Davis L, Sanchez A, Stohl HE, Nelson AL, Robinson N. A review of women's unanswered questions following miscarriage on different social media platforms [A207]. Obstet Gynecol 2022 May;139:60S. [doi: 10.1097/01.AOG.0000826104.75828.4f]
21. Valdez D, Patterson MS. Computational analyses identify addiction help-seeking behaviors on the social networking website Reddit: insights into online social interactions and addiction support communities. PLOS Digit Health 2022 Nov;1(11):e0000143 [FREE Full text] [doi: 10.1371/journal.pdig.0000143] [Medline: 36812569]
22. Sit M, Elliott SA, Wright KS, Scott SD, Hartling L. Youth mental health help-seeking information needs and experiences: a thematic analysis of Reddit posts. Youth Soc 2022 Oct 29;56(1):24-41. [doi: 10.1177/0044118x221129642]
23. Abavi R, Branston A, Mason R, Du Mont J. An exploration of sexual assault survivors' discourse online on help-seeking. Violence Vict 2020 Feb 03;35(1):126-140. [doi: 10.1891/0886-6708.vv-d-18-00148]
24. Ayers JW, Zhu Z, Harrigian K, Wightman GP, Dredze M, Strathdee SA, et al. Managing HIV during the COVID-19 pandemic: a study of help-seeking behaviors on a social media forum. AIDS Behav 2023 Jul 21 (forthcoming). [doi: 10.1007/s10461-023-04134-9] [Medline: 37479919]
25. Higgins J, Lands M, Valley T, Carpenter E, Jacques L. Real-time effects of payer restrictions on reproductive healthcare: a qualitative analysis of cost-related barriers and their consequences among U.S. abortion seekers on Reddit. Int J Environ Res Public Health 2021 Aug 26;18(17):9013 [FREE Full text] [doi: 10.3390/ijerph18179013] [Medline: 34501602]
26. Jacques L, Carpenter E, Valley T, Alvarez B, Higgins J. Medication or surgical abortion? An exploratory study of patient decision making on a popular social media platform. Am J Obstet Gynecol 2021 Sep;225(3):344-347 [FREE Full text] [doi: 10.1016/j.ajog.2021.05.011] [Medline: 34022196]
27. Richards NK, Masud A, Arocha J. P28 Breaking down abortion barriers: Reddit users' empowerment in absence of parental and medical support. Contraception 2020 Oct;102(4):286. [doi: 10.1016/j.contraception.2020.07.047]
28. Sawicki J, Ganzha M, Paprzycki M, Watanobe Y. Reddit CrosspostNet—studying Reddit communities with large-scale Crosspost graph networks. Algorithms 2023 Sep 04;16(9):424. [doi: 10.3390/a16090424]
29. Lanthier S, Mason R, Logie CH, Myers T, Du Mont J. "Coming out of the closet about sexual assault": intersectional sexual assault stigma and (non) disclosure to formal support providers among survivors using Reddit. Soc Sci Med 2023 Jul;328:115978. [doi: 10.1016/j.socscimed.2023.115978] [Medline: 37276773]
30. Richards NK, Masud A, Arocha JF. Online abortion empowerment in absence of parental and medical support: a thematic analysis of a reddit community's contributions to decision-making and access. Research Square. Preprint posted online May 24, 2021 2024 [FREE Full text]
31. Madan P. Web scraping Reddit with python: a complete guide with code. GoLogin. 2023 Mar 23. URL: https://gologin.com/blog/web-scraping-reddit [accessed 2023-09-26]
32. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022. [doi: 10.7551/mitpress/1120.003.0082]
33. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, Boyd-Graber J. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2015 Presented at: 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015; Denver, CO. [doi: 10.3115/v1/w15-1212]
34. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022 May 6;7:886498 [FREE Full text] [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
35. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-demos.6]
36. Drikvandi R, Lawal O. Sparse principal component analysis for natural language processing. Ann Data Sci 2020 May 18;10(1):25-41. [doi: 10.1007/s40745-020-00277-x]
37. Stewart G, Al-Khassaweneh M. An implementation of the HDBSCAN* clustering algorithm. Appl Sci 2022 Feb 25;12(5):2405. [doi: 10.3390/app12052405]
38. Kim SW, Gil JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum Cent Comput Inf Sci 2019 Aug 26;9:30. [doi: 10.1186/s13673-019-0192-7]
39. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May 16;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]

XSL•FO
RenderX

40. Aslam N, Rustam F, Lee E, Washington PB, Ashraf I. Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model. IEEE Access 2022;10:39313-39324 [FREE Full text] [doi: 10.1109/ACCESS.2022.3165621]
41. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020 Dec 14;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
42. Adarsh R, Patil A, Rayar S, Veena KM. Comparison of VADER and LSTM for sentiment analysis. Int J Recent Technol Eng 2019 Mar;7(6):543 [FREE Full text]
43. Nesca M, Katz A, Leung C, Lix L. A scoping review of preprocessing methods for unstructured text data to assess data quality. Int J Popul Data Sci 2022;7(1). [doi: 10.23889/ijpds.v7i1.1757]
44. Hertling S, Portisch J, Paulheim H. KERMIT -- a transformer-based approach for knowledge graph matching. arXiv. Preprint posted online April 29, 2022 2024 [FREE Full text] [doi: 10.48550/arXiv.2204.13931]
45. O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Syst Appl 2015 Aug;42(13):5645-5657. [doi: 10.1016/j.eswa.2015.02.055]
46. Szymkowiak A, Melović B, Dabić M, Jeganathan K, Kundi GS. Information technology and Gen Z: the role of teachers, the internet, and technology in the education of young people. Technol Soc 2021 May;65:101565. [doi: 10.1016/j.techsoc.2021.101565]
47. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 7. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2023-03-20]
48. Frey E, Bonfiglioli C, Brunner M, Frawley J. Parents' use of social media as a health information source for their children: a scoping review. Acad Pediatr 2022 May;22(4):526-539 [FREE Full text] [doi: 10.1016/j.acap.2021.12.006] [Medline: 34906742]
49. Crawford BL, Simmons MK, Turner RC, Lo WJ, Jozkowski KN. Perceptions of abortion access across the United States prior to the Dobbs v. Jackson Women's Health Organization decision: results from a national survey. Perspect Sex Reprod Health 2023 Sep 20;55(3):153-164. [doi: 10.1363/psrh.12238] [Medline: 37475195]
50. Tracking abortion bans across the country. The New York Times. URL: https://www.nytimes.com/interactive/2022/us/abortion-laws-roe-v-wade.html [accessed 2023-09-26]
51. Reis BY, Brownstein JS. Measuring the impact of health policies using internet search patterns: the case of abortion. BMC Public Health 2010 Aug 25;10:514 [FREE Full text] [doi: 10.1186/1471-2458-10-514] [Medline: 20738850]
52. Kim T, Steinberg JR. Individual changes in abortion knowledge and attitudes. Soc Sci Med 2023 Mar;320:115722. [doi: 10.1016/j.socscimed.2023.115722] [Medline: 36709693]
53. Bueno X, Asamoah NA, LaRoche KJ, Dennis B, Crawford BL, Turner RC, et al. People's perception of changes in their abortion attitudes over the life course: a mixed methods approach. Adv Life Course Res 2023 Sep;57:100558. [doi: 10.1016/j.alcr.2023.100558] [Medline: 38054859]
54. Jozkowski KN, Mena-Meléndez L, Crawford BL, Turner RC. Abortion stigma: attitudes toward abortion responsibility, illegal abortion, and perceived punishments of "illegal abortion". Psychol Women Q 2023 Jul 04;47(4):443-461. [doi: 10.1177/03616843231181350]
55. Shen Q, Rosé CP. A tale of two subreddits: measuring the impacts of quarantines on political engagement on Reddit. Proc IntAAAI Conf Web Soc Media 2022 May 31;16(1):932-943. [doi: 10.1609/icwsm.v16i1.19347]
56. Crawford BL, Jozkowski KN, Turner RC, Lo WJ. Examining the relationship between Roe v. Wade knowledge and sentiment across political party and abortion identity. Sex Res Soc Policy 2021 May 28;19(3):837-848. [doi: 10.1007/s13178-021-00597-4]
57. LaRoche KJ, Jozkowski KN, Crawford BL, Haus KR. Attitudes of US adults toward using telemedicine to prescribe medication abortion during COVID-19: a mixed methods study. Contraception 2021 Jul;104(1):104-110 [FREE Full text] [doi: 10.1016/j.contraception.2021.04.001] [Medline: 33848466]
58. Kulkarni V, Kern ML, Stillwell D, Kosinski M, Matz S, Ungar L, et al. Latent human traits in the language of social media: an open-vocabulary approach. PLoS One 2018 Nov 28;13(11):e0201703 [FREE Full text] [doi: 10.1371/journal.pone.0201703] [Medline: 30485276]
59. Dowler K. Media influence on attitudes toward guns and gun control. Am J Crim Just 2002 Mar;26(2):235-247. [doi: 10.1007/bf02887829]
60. O'Connor C. 'Appeals to nature' in marriage equality debates: a content analysis of newspaper and social media discourse. Br J Soc Psychol 2017 Sep 27;56(3):493-514. [doi: 10.1111/bjso.12191] [Medline: 28239878]
61. Chen L, Ling Q, Cao T, Han K. Mislabeled, fragmented, and conspiracy-driven: a content analysis of the social media discourse about the HPV vaccine in China. Asian J Commun 2020 Sep 08;30(6):450-469. [doi: 10.1080/01292986.2020.1817113]
62. Catania JA, Dolcini MM, Orellana R, Narayanan V. Nonprobability and probability-based sampling strategies in sexual science. J Sex Res 2015;52(4):396-411. [doi: 10.1080/00224499.2015.1016476] [Medline: 25897568]
63. Maier JM, Jozkowski KN, Montenegro MS, Willis M, Turner RC, Crawford BL, et al. Examining auxiliary verbs in a salient belief elicitation. Health Behav Policy Rev 2021 Jul;8(4):374-393. [doi: 10.14485/HBPR.8.4.9]

XSL•FO
RenderX

Abbreviations

API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
LDA: latent Dirichlet allocation
NLP: natural language processing
PRAW: Python Reddit application programming interface Wrapper
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by A Mavragani; submitted 18.03.23; peer-reviewed by L Jacques, T Zhang; comments to author 27.07.23; revised version received 27.09.23; accepted 20.12.23; published 14.02.24.

Please cite as:
Valdez D, Mena-Meléndez L, Crawford BL, Jozkowski KN
Analyzing Reddit Forums Specific to Abortion That Yield Diverse Dialogues Pertaining to Medical Information Seeking and Personal Worldviews: Data Mining and Natural Language Processing Comparative Study
J Med Internet Res 2024;26:e47408
URL: https://www.jmir.org/2024/1/e47408
doi:10.2196/47408
PMID:38354044

©Danny Valdez, Lucrecia Mena-Meléndez, Brandon L Crawford, Kristen N Jozkowski. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study

Michael S Deiner[1,2], PhD; Natalie A Deiner[3]; Vagelis Hristidis[4], PhD; Stephen D McLeod[1,2,5], MD; Thuy Doan[1,2,6], MD, PhD; Thomas M Lietman[1,2,6], MD; Travis C Porco[1,2,6], MPH, PhD

[1]Department of Ophthalmology, University of California, San Francisco, San Francisco, CA, United States
[2]Francis I. Proctor Foundation for Research in Ophthalmology, University of California, San Francisco, San Francisco, CA, United States
[3]College of Letters and Science, University of California, Santa Barbara, Santa Barbara, CA, United States
[4]Department of Computer Science and Engineering, University of California, Riverside, Riverside, CA, United States
[5]American Academy of Ophthalmology, San Francisco, CA, United States
[6]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Travis C Porco, MPH, PhD
Francis I. Proctor Foundation for Research in Ophthalmology
University of California, San Francisco
490 Illinois St., Box 0944
San Francisco, CA, 94143-0944
United States
Phone: 1 415 476 0527
Email: travis.porco@ucsf.edu

Abstract

Background: Previous work suggests that Google searches could be useful in identifying conjunctivitis epidemics. Content-based assessment of social media content may provide additional value in serving as early indicators of conjunctivitis and other systemic infectious diseases.

Objective: We investigated whether large language models, specifically GPT-3.5 and GPT-4 (OpenAI), can provide probabilistic assessments of whether social media posts about conjunctivitis could indicate a regional outbreak.

Methods: A total of 12,194 conjunctivitis-related tweets were obtained using a targeted Boolean search in multiple languages from India, Guam (United States), Martinique (France), the Philippines, American Samoa (United States), Fiji, Costa Rica, Haiti, and the Bahamas, covering the time frame from January 1, 2012, to March 13, 2023. By providing these tweets via prompts to GPT-3.5 and GPT-4, we obtained probabilistic assessments that were validated by 2 human raters. We then calculated Pearson correlations of these time series with tweet volume and the occurrence of known outbreaks in these 9 locations, with time series bootstrap used to compute CIs.

Results: Probabilistic assessments derived from GPT-3.5 showed correlations of 0.60 (95% CI 0.47-0.70) and 0.53 (95% CI 0.40-0.65) with the 2 human raters, with higher results for GPT-4. The weekly averages of GPT-3.5 probabilities showed substantial correlations with weekly tweet volume for 44% (4/9) of the countries, with correlations ranging from 0.10 (95% CI 0.0-0.29) to 0.53 (95% CI 0.39-0.89), with larger correlations for GPT-4. More modest correlations were found for correlation with known epidemics, with substantial correlation only in American Samoa (0.40, 95% CI 0.16-0.81).

Conclusions: These findings suggest that GPT prompting can efficiently assess the content of social media posts and indicate possible disease outbreaks to a degree of accuracy comparable to that of humans. Furthermore, we found that automated content analysis of tweets is related to tweet volume for conjunctivitis-related posts in some locations and to the occurrence of actual epidemics. Future work may improve the sensitivity and specificity of these methods for disease outbreak detection.

(J Med Internet Res 2024;26:e49139) doi:10.2196/49139

KEYWORDS

conjunctivitis; microblog; social media; generative large language model; Generative Pre-trained Transformers; GPT-3.5; GPT-4; epidemic detection; Twitter; X formerly known as Twitter; infectious eye disease

XSL•FO
RenderX

Introduction

Background

Conjunctivitis, while usually self-limiting, results in substantial societal costs [1,2] and can give rise to large outbreaks [2-7]. The detection of conjunctivitis epidemics can help reduce societal burden, prevent impacts on eye health, and act as a warning sign for emerging outbreaks of higher-risk systemic infectious diseases such as COVID-19. Recently, a study of a COVID-19 variant identified in 2023 that caused febrile illness and respiratory symptoms in children found conjunctivitis in 42.8% of the individuals who were affected [8,9].

The usual process of monitoring conjunctivitis outbreaks through individual case identification is costly; moreover, conjunctivitis is not, in general, a reportable disease in the United States (although gonococcal cause is reportable [10]). Low-cost digital approaches using public search and social media big data for surveillance could help fill this and other information gaps in eye health [11] by providing real-time information [11-13]. Previously, we found that an analysis of Google time series for relative search volume for conjunctivitis can identify outbreaks of conjunctivitis with differing ability based on the keywords, the country, and the size of the outbreak [14]; we also found that social media posts have been correlated with the clinical occurrence of conjunctivitis [15] and have reflected the seasonal occurrence of allergic and infectious conjunctivitis [16]. This suggested that a future system based on an analysis of web-based search frequency could be automated, reporting potential outbreaks worldwide.

By analyzing the content of social media posts during these detected candidate epidemics, we have observed that spikes in conjunctivitis-related search data can be caused by many factors besides outbreaks. These causes include media coverage, celebrity affliction, movie titles, artist names, and other factors not specific to infectious conjunctivitis. Any automated system aiming to detect, and alert about, potential epidemics based on search data would still require the monitoring of content during any suspected epidemic period to improve specificity. Previous research has suggested that web-based content can be useful in infectious disease surveillance [12,13,17-29]. Unfortunately, manual content analysis can be time-consuming, but available generative large language models (LLMs) could be assessed for their potential to assist with such a task in an automated fashion.

The Aim of This Study

In this study, we investigated whether the analysis of geolocated time-series social media content [30] using LLMs could be used to accurately summarize the content of posts regarding conjunctivitis in general. To help refine our assessment of potential conjunctivitis outbreaks detected from search data in an automated fashion, we also investigated whether LLMs could assign a useful probability that a post's content is specifically about a conjunctivitis outbreak [31]. We obtained tweets from 9 of the countries assessed in our previous study [14] and presented these to GPT-3.5 and GPT-4 (OpenAI) [32], which are transformers-based LLMs. We tested the hypotheses that automated content analysis using these models can yield a time series of elicited outbreak probabilities and that these probabilities are correlated with tweet frequency and the occurrence of known epidemics.

Methods

Data Collection

On the basis of our previous analysis [14], we chose 9 countries for which we knew the dates of conjunctivitis epidemics. We chose these to include both small countries and island territories as well as large countries; results for no other countries were analyzed. For these countries, we collected tweets from the Twitter microblogging service (subsequently rebranded as X) using the Brandwatch interface. To obtain posts about conjunctivitis, we used a Boolean query containing words in multiple languages representing conjunctivitis (eg, "conjuctivitis," "conjuntivitis," "conjuntivite," and "pink eye"). We tailored this to exclude irrelevant content, such as that related to animals, and confounding content (such as celebrities having pink eye).

The full Boolean query is provided in Multimedia Appendix 1. Only tweets geolocated to each country were exported. The data cutoff window began on January 1, 2012 (January 1, 2013, for India), and ended on March 13, 2023. The data were exported on March 13, 2023, and the counts are summarized in Table 1. The corresponding epidemic start dates, presented in our previous study [14], are also included in Table 1.

XSL•FO
RenderX

Table 1. Summary of available tweets and known epidemics in the 9 study countries.

Location	Tweets, n	Epidemic start dates
India	4999	August 9, 2012; July 25, 2013; November 15, 2013; September 4, 2014; and April 9, 2017
Guam (United States)	282	May 15, 2014
Martinique (France)	336	May 14, 2017
Philippines	3976	August 27, 2015
American Samoa (United States)	68	April 1, 2014
Fiji	142	March 15, 2016
Costa Rica	1494	June 30, 2017
Haiti	512	May 15, 2017
Bahamas	385	May 15, 2017

Data Analysis

Automated Content Analysis

We used the OpenAI LLMs GPT-3.5 (gpt-3.5-turbo-0301) and GPT-4 (gpt-4-0314), accessed through the application programming interface [33]. Another potentially comparable LLM, Google Bard, was not available through an application programming interface at the time we conducted our study. GPT-4 was available in limited beta release and was only used for prompt 1. We chose to use the less expensive GPT-3.5 as well as the newer, potentially more advanced GPT-4.

First, for each tweet, we directly elicited a probability that the tweet indicated a conjunctivitis outbreak. For this, we used prompt 1:

> *How certain are you that the single Tweet provided below is about a large multiperson outbreak of pink eye occurring at the time the tweet was posted? A single case with no other evidence of spread or other infected people should correspond to a somewhat low probability. Respond in the form of "Tweet: x%," on a scale of 0% to 100%, and then provide a brief explanation of your answer. Given Tweet: <direct quote>*

Second, we asked the model to simply assess the occurrence of an epidemic, based on the content of the tweet. This was prompt 2:

> *Answer if the tweet below is about a large multiperson outbreak of conjunctivitis, occurring at the time the tweet was made. A single case with no other evidence of spread or other infected people should correspond to a somewhat low probability. The response choices are: NO, not conjunctivitis outbreak (the tweet is irrelevant or indicates 0-1 cases of conjunctivitis max, not spreading or not occurring at the time the tweet was made); MAYBE conjunctivitis outbreak (uncertain, the tweet indicates maybe 2 or more cases of conjunctivitis, maybe spreading); YES conjunctivitis outbreak (the tweet indicates more than 1 case of conjunctivitis and/or spreading, and occurring at the time the tweet was made). For your answer, respond first with one of the three choices (NO, "not conjunctivitis outbreak," MAYBE, "conjunctivitis outbreak," YES, "conjunctivitis outbreak") and then provide a brief explanation for your choice, including the type of disease if you say YES, "conjunctivitis outbreak." Given Tweet: <direct quote>*

Although the use of a continuous variable (elicited probability) from prompt 1 maximizes statistical power [34] compared with dichotomized data, we also included the results of the conceptually simpler prompt 2 along with the results of prompt 1 for comparison.

For both prompts 1 and 2, we replaced *<direct quote>* with each of the 12,194 tweets in turn, collecting all responses. For all queries, we used a *top_p* of 0.9 (the default value) and a temperature of 0.

To provide illustrative examples, we divided the tweets into groups with GPT-3.5–derived percentages of 0%, between 0% and 70% (exclusive), and >70% and randomly selected 3 tweets from each group. We removed specific identifying information from each tweet and lightly edited them to reduce discoverability [35]; we note that these tweets were public. Samples of these redacted tweets and LLM responses to prompt 1 for them were prepared solely for the illustration of LLM replies to the 2 prompts. Only replies to the original unredacted tweets were used in all analyses presented in this study.

Human Rater Validation of GPT Classification and Scoring

To validate the resulting conjunctivitis epidemic probabilities and classifications by GPT-3.5 and GPT-4 of the tweets, 2 human raters participated in a modified Delphi session. During the session, the raters manually reviewed a random sample of tweets, classified them into the same categories as the GPT models ("NO," "YES," and "MAYBE" conjunctivitis outbreak), and assigned a conjunctivitis epidemic probability score (0%-100%) to each. The human and GPT categorizations and scores were then compared.

We asked the 2 human raters to independently read each tweet, using the same prompts that were provided to the LLMs. For the testing set used, a random selection of tweets was stratified

XSL•FO
RenderX

by country and by the elicited probabilities from GPT-3.5 to ensure that as close to a maximum of 7 tweets that scored >50% and 7 tweets that scored <50% were included from each country (126 tweets in total). The sample size was chosen to provide a CI half-width of approximately 0.05 for estimated proportions near 0.5. Similarly, separate training sets of independent tweets were generated (18 per set). Only English-language tweets were used in validation. Training and testing sets were used as described in the following paragraph.

The raters first trained together on the first training set, assigning classification and probability scores via a Qualtrics survey (Qualtrics International Inc). A facilitated group discussion for the raters then followed, to reconcile disagreements in the categorization and scores as well as to gain familiarity with the discussion on Twitter (ie, to become aware of the language and components, such as hashtags and sarcasm, used in these posts). The raters subsequently completed a second iteration of the training with the second training set, followed by a similar brief discussion as before so that a general agreement was reached. We then provided the testing set in a separate Qualtrics survey (excluding any tweets used in training the raters) to the raters. Each rater assigned classification and probability scores to each post in the testing set, masked to the results of other raters and that of the machine and without any discussion.

Statistical Analysis

In time-series data of tweet volume about a disease, we expect an increase in the weekly count of posts about the disease during an epidemic compared with nonepidemic periods [36]. Therefore, as an assessment of the ability of the GPT models to assign higher probabilities to tweets in weeks where there may be more likely to be an epidemic (higher counts of total tweets per week) as well as to assign lower probabilities to weeks less likely to have an epidemic (low total counts of tweets per week), we asked whether the weekly count of posts about pink eye correlated with the scores assigned to that week by the LLMs. To calculate weekly values from the elicited probabilities from each of the GPT models, we first removed highly repetitive tweets as follows: we removed usernames beginning with @ from the content and then removed all tweets with duplicated content. From the remaining tweets, we averaged all those values for each week. Weeks with no tweets at all were scored as having a mean of 0. Elicited percentages were treated as continuous variables in statistical analyses [34]. We converted prompt 2 results to numerical values (to allow for correlation analysis) by assigning values: 0="NO," 1="MAYBE," and 2="YES."

For each country in Table 1, we constructed an indicator variable, which was 1 for any week an epidemic was believed to have started and for 3 weeks after. We then calculated the Pearson correlation between the number of tweets per week and the mean LLM-derived conjunctivitis outbreak probability score (prompt 1) or outbreak classification (prompt 2) for each week. We used time-series bootstrap [37] with a width of 4 to calculate CIs. We also correlated the weekly epidemic indicator variable with the mean elicited percentages obtained from GPT-3.5 and GPT-4 in response to prompt 1; we note that because we have no complete listing of all epidemics, the correlation will unavoidably underestimate the true relation between the occurrence of an epidemic and any content analysis measure. In addition, we included the volume of tweets matching our query as an outcome variable when comparing with epidemic indicator variables. As a sensitivity analysis, we conducted analysis without removing duplicates and without considering weeks with 0 tweets as having a mean of 0. Estimated Pearson correlations for the time series were compared between GPT-3.5 and GPT-4 by subtracting the estimates for each country and conducting the Wilcoxon signed rank test.

Ethical Considerations

The work carried out for this study was conducted after receiving approval from the University of California San Francisco Institutional Review Board (14-14743).

Results

Descriptive Summary of GPT Model Responses

The overall ability of the 2 LLMs to respond in the format that we desired was high. For prompt 1, GPT-3.5 provided usable responses 99.66% (12,153/12,194) of the time, whereas GPT-4 provided usable responses 100% (12,194/12,194) of the time. For prompt 2, GPT-3.5 provided usable responses 98.67% (12,032/12,194) of the time. We observed that the ability to summarize the content of the tweets seemed quite reliable and surprisingly nuanced; selected examples of prompt 1 responses, based on the redacted tweets described previously, are provided in Table 2. For comparison, prompt 2 responses to these same tweets are provided in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Selected actual GPT-3.5 and GPT-4 responses to prompt 1 for typical tweets from our data set (for this table, actual usernames, locations, and other identifiers were modified to avoid presenting identifying information).

Tweet	GPT-3.5 response	GPT-4 response
All undergrads under me got conjunctivitis. So not allowing to theatre and class. Some who r nt effected also act like they got. Managing class all alone. Im jealous seeing them enjoying.	• Tweet: 70% • Explanation: The tweet mentions that "All undergrads under me got conjunctivitis" which suggests a large multiperson outbreak. Additionally, the author mentions not allowing them to attend class or theater, which implies that the outbreak is significant enough to warrant such measures. However, there is no mention of any official reports or confirmation of the outbreak, so there is still some uncertainty.	• Tweet: 60% • Explanation: The tweet suggests that multiple undergraduates have contracted conjunctivitis (pink eye), which indicates a possible outbreak. However, the tweet also mentions that some students are pretending to have the condition, which makes it difficult to determine the true extent of the outbreak. Additionally, this is just one tweet and not enough to confirm a large multiperson outbreak.
Pharmacies ran out of eye drops for conjunctivitis as those infected continue to go to work, school, gatherings, and so on. They affect others–urgh!	• Tweet: 80% • Explanation: The tweet suggests that there is a shortage of eye drops for conjunctivitis, which implies that there is a high demand for them. The fact that those infected continue to go to work, school, and gatherings suggests that there is a high likelihood of transmission to others, leading to a large multiperson outbreak. However, without additional information, it is difficult to determine the exact size of the outbreak, hence the probability of 80%.	• Tweet: 80% • Explanation: The tweet suggests a higher probability of a large multiperson outbreak of pink eye, as it mentions pharmacies running out of eye drops for conjunctivitis and infected individuals continuing to attend work, school, and gatherings, which could lead to the spread of the infection. However, without further information or confirmation, there is still some uncertainty.
@user conjunctivitis??	• Tweet: 50% • Explanation: The tweet alone does not provide enough information to determine the certainty of a large multiperson outbreak of pink eye. The tweet only mentions one case of conjunctivitis and does not provide any additional information about the spread or other infected people. Therefore, there is a moderate probability of a large multiperson outbreak of pink eye occurring at the time the tweet was posted.	• Tweet: 10% • Explanation: The tweet only mentions "conjunctivitis" with a question mark, which provides very little information about the context or any potential outbreak. There is no mention of multiple people being affected or any indication of a large-scale event. Therefore, the certainty of this tweet being about a large multiperson outbreak of pink eye is quite low.
１日家にいたーひさびさのひきこもりやっぱ結膜炎気になって つきっきりしてたー だいぶ寝たー	• Tweet: 10% • Explanation: The tweet is in Japanese, but based on the use of the word "結膜炎" (conjunctivitis or pink eye) and the fact that the person stayed at home all day and rested, it is possible that they had a case of pink eye. However, there is no mention of other people being infected or any indication that this is a large multiperson outbreak. Therefore, the probability of this tweet being about a large multiperson outbreak of pink eye is relatively low at 10%.	• Tweet: 10% • Explanation: The tweet is in Japanese and translates to "I stayed home all day. It's been a while since I've been a shut-in. I was worried about conjunctivitis (pink eye) and stayed close. I slept a lot." This tweet only mentions a single case of pink eye and does not provide any information about a large multiperson outbreak. Therefore, the probability is low at 10%.
do i have pink eye or could it be my allergies again?	• Tweet: 0% • Explanation: The given tweet does not provide any evidence of a large multiperson outbreak of pink eye. The tweet only mentions a personal experience of the individual and does not suggest any spread or other infected people. Therefore, the probability of the tweet being about a large multiperson outbreak of pink eye is very low, at 0%.	• Tweet: 5% • Explanation: The tweet only mentions the possibility of the user having pink eye or allergies, but there is no mention of a large multiperson outbreak or any other infected people. Therefore, the probability is very low.
Did you know that the first case of viral conjunctivitis was found in 1902? It is 2020 and "science" still has no "cure" for it!	• Tweet: 0% • Explanation: The given tweet does not provide any evidence or mention of a large multiperson outbreak of pink eye occurring at the time the tweet was posted. It only talks about the history of viral conjunctivitis and the lack of a cure for it. Therefore, the probability of the tweet being about a large multiperson outbreak of pink eye is very low, at 0%.	• Tweet: 0% • Explanation: The tweet provided does not mention any current outbreak of pink eye or any indication of multiple infected people. It only discusses the first case of viral conjunctivitis found in 1902 and the lack of a cure in 2020.

XSL•FO
RenderX

Human Rater Validations of GPT-Assigned Scores and Classifications That a Tweet Was About Conjunctivitis Outbreaks

We found that, overall, the LLMs were less likely than the human raters to assign the extreme values of 0% or 100%. Moreover, GPT-3.5 chose larger values than GPT-4 for the validation sample. A descriptive summary of the validation sample is presented in Table 3.

Next, we computed correlations between the elicited percentages of the 2 human raters and the responses of GPT-3.5 and GPT-4 to prompt 1. Overall, the percentages derived from the replies of both GPT-3.5 and GPT-4 had a correlation coefficient of at least 0.6 with those of the human raters, although the responses of GPT-4 were more correlated with those of the human raters than those of GPT-3.5. The correlation of the percentages of GPT-4 were roughly as correlated with those of the human raters as the results of the human raters were with each other. These validation set results are summarized in Table 4. As a measure of interrater reliability for prompt 2, the estimated unweighted Cohen κ value was 0.64 (P<.001) for a comparison of the 2 human raters. The Cohen κ value for a comparison of the results of rater 1 with those of GPT-3.5 for prompt 2 was 0.51 (P<.001), and for a comparison of the results of rater 2 with those of GPT-3.5 for prompt 2, the Cohen κ value was 0.48 (P<.001).

Table 3. Validation with human raters: summary of grading. The proportion of the total corpus of validation testing set tweets (n=126) assigned by human raters and the GPT models to 0% and 100% probability that a tweet is about an outbreak is shown, along with the median (IQR) percentage assigned.

Measurement	Rater 1	Rater 2	GPT-3.5, prompt 1	GPT-4, prompt 1
Ratings of 0%, n (%)	4 (0.3)	5 (0.4)	1 (0.1)	0 (0)
Rating (%), median (IQR)	30 (0-90)	10 (0-100)	55 (10-70)	10 (10-30)
Ratings of 100%, n (%)	3 (0.2)	4 (0.3)	0 (0)	0 (0)

Table 4. Validation with human raters: correlation (Pearson r). Correlation matrix of elicited percentages from human raters and GPT models using prompt 1, based on the validation set (n=126).

Variable	Human 1	Human 2	GPT-3.5, prompt 1	GPT-4, prompt 1
Human 1, *r* (95% CI)	1	0.77 (0.68-0.83)	0.60 (0.47-0.70)	0.73 (0.64-0.80)
Human 2, *r* (95% CI)	0.77 (0.68-0.83)	1	0.53 (0.40-0.65)	0.77 (0.68-0.83)
GPT-3.5, prompt 1, *r* (95% CI)	0.60 (0.47-0.70)	0.53 (0.40-0.65)	1	0.77 (0.68-0.83)
GPT-4, prompt 1, *r* (95% CI)	0.73 (0.64-0.80)	0.77 (0.68-0.83)	0.77 (0.68-0.83)	1

Descriptive Summaries of GPT-3.5 and GPT-4 Probabilities and Classifications

For each of the 9 countries, summaries of the elicited percentages for the full set of tweets using GPT-3.5 and GPT-4 are shown in Figure 1 and Figure 2, respectively. The models provided low percentages (≤20%) for most of the tweets (7922/12194, 65.0% for GPT-3.5; 11070/12194, 90.8% for GPT-4) in all countries. Of the 12,194 tweets, 677 (5.55%) were removed because they were highly repetitive. From the remaining 11,517 tweets, the overall mean percentage elicited was 21%, with a median percentage of 10% (IQR: 5-50%). For prompt 1, neither GPT-3.5 nor GPT-4 provided any elicited percentages of 100%. Both showed profound final digit preference; in only 1 case did GPT-3.5 provide a percentage that did not end in 0 or 5, and all from GPT-4 ended in 0 or 5.

In response to prompt 2, where we simply asked the LLM to classify each tweet as “YES,” “NO,” or “MAYBE” regarding an outbreak of conjunctivitis, the distribution of classifications assigned to each tweet by GPT-3.5 is shown in Figure 3. Of note, in 162 (1.41%) of the 11,517 tweets, the LLM’s response did not begin with 1 of the 3 requested words, and we treated these as missing (although in all cases, the LLM responded with an explanation of why it was difficult to be sure of the answer and therefore did not choose 1 of the 3 response options).

XSL•FO
RenderX

Figure 1. Distributions of elicited epidemic probability assigned to tweets by GPT-3.5 for each country (12,194 tweets in total) using prompt 1. Light blue: 0% to 20%, green: 21% to 40%, yellow: 41% to 60%, orange: 61% to 80%, and red: 81% to 100%. The total count is placed at the top of each bar.

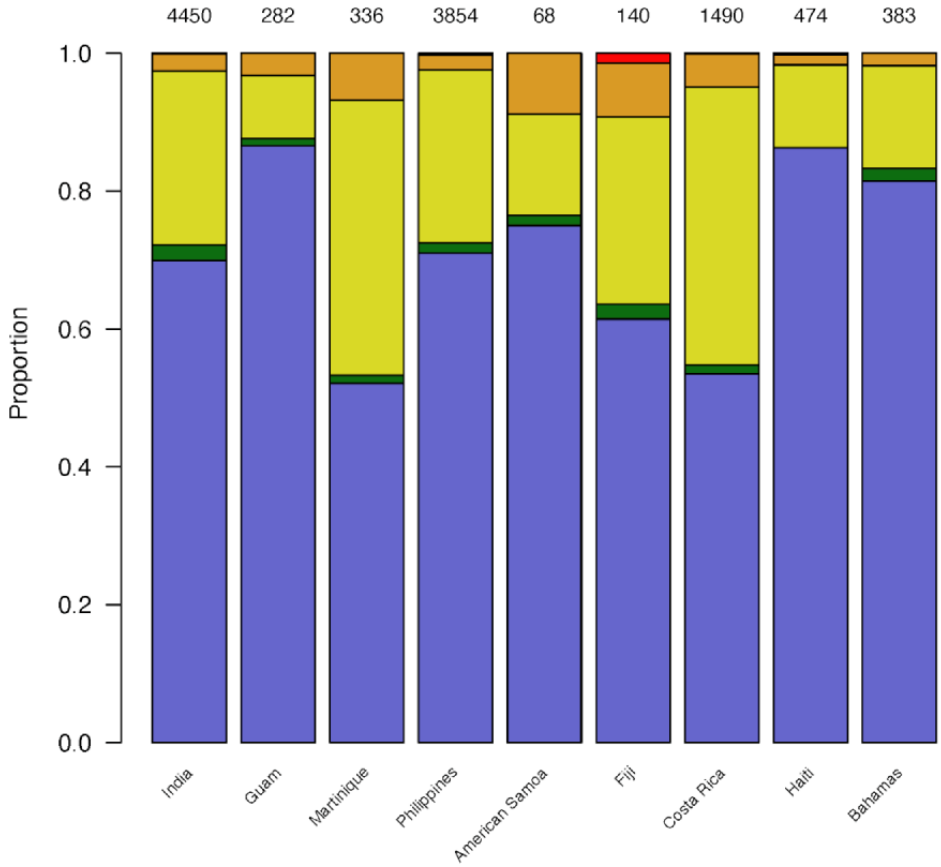

XSL•FO
RenderX

Figure 2. Distributions of elicited epidemic probability assigned to tweets by GPT-4 for each country (12,194 tweets in total) using prompt 1. Light blue: 0% to 20%, green: 21% to 40%, yellow: 41% to 60%, orange: 61% to 80%, and red: 81% to 100%. The total count is placed at the top of each bar.

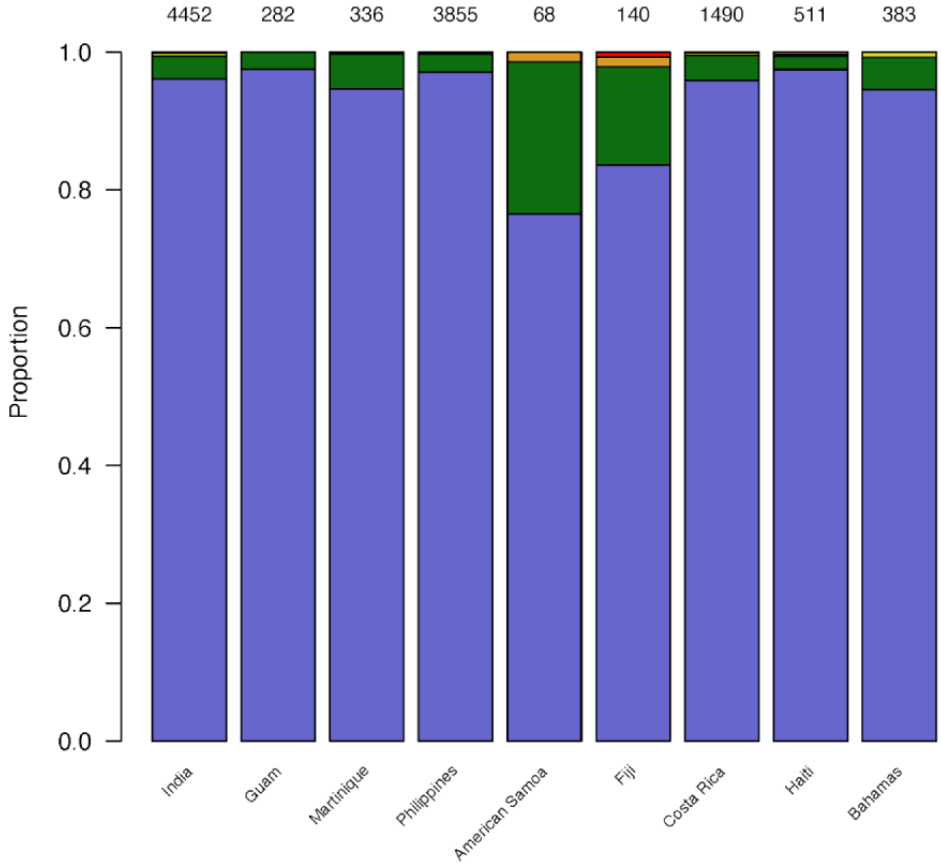

XSL•FO
RenderX

Figure 3. Distributions of elicited epidemic classifications assigned to tweets for GPT-3.5 for each country (12,194 tweets in total) using prompt 2. Light blue: "NO," yellow: "MAYBE," red: "YES," white: "missing." The total count is placed at the top of each bar.

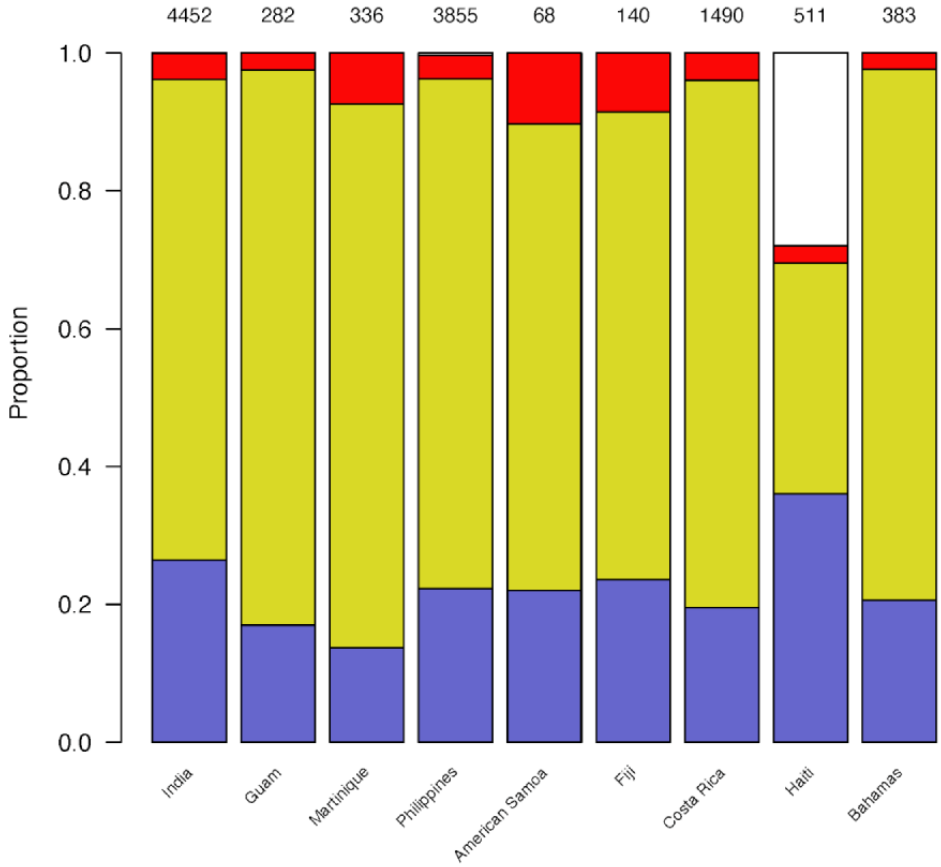

Correlation of Results for Tweets Between Models and Between Prompt 1 and Prompt 2

At the level of individual tweets, the probabilities assigned by GPT-3.5 and GPT-4 based on prompt 1 were highly correlated, with a Pearson *r* value of 0.42 (95% CI 0.41-0.44). To compare the results elicited from GPT-3.5 for prompts 1 and 2 per tweet, we converted the elicitations from prompt 2 to numerical values. Specifically, we assigned the following values: 0="NO," 1="MAYBE, and 2="YES." We found a correlation of 0.45 (95% CI 0.43-0.46) between the prompt 1–elicited probabilities and the prompt 2–elicited classifications.

Comparisons of Elicited Epidemic Probability and Epidemic Classification Results per Tweet Between Models and Between Prompt 1 and Prompt 2

We next compared the elicited epidemic probabilities from the LLMs with weekly tweet volume based on our search, as described in the *Methods* section. We computed the Pearson correlation of the number of tweets meeting the search criteria, as well as the mean elicited percentages for GPT-3.5 and GPT-4 in response to prompt 1. We also used a binary indicator of whether GPT-3.5 responded "YES" to prompt 2. The estimated correlations for GPT-3.5 using prompt 1 ranged from 0.10 (India) to 0.53 (American Samoa [United States]); for GPT-4 using prompt 1, the estimated correlations ranged from 0.18 (India) to 0.64 (Guam [United States]), with broadly higher correlations seen in GPT-4 (*P*=.004, Wilcoxon signed rank test). The results for each of the 9 countries are shown in Table 5. When weeks containing 0 tweets were excluded, the results were similar (refer to Table S2 in Multimedia Appendix 1). Similarly, when we did not exclude duplicated or highly repetitive tweets, the results were similar (although slightly lower; results not shown).

XSL•FO
RenderX

Table 5. Correlation of weekly tweet volume and elicited outbreak percentages by the GPT models (refer to the text for details).

Country	GPT-3.5, prompt 1, weekly mean, *r* (95% CI)	GPT-4, prompt 1, weekly mean, *r* (95% CI)	GPT-3.5 "YES," prompt 2, weekly mean, *r* (95% CI)
India	0.10 (−0.00 to 0.29)	0.18 (0.09 to 0.37)	0.04 (−0.01 to 0.13)
Guam (United States)	0.42 (0.34 to 0.57)	0.64 (0.55 to 0.79)	0.08 (0.04 to 0.18)
Martinique (France)	0.36 (0.31 to 0.66)	0.45 (0.40 to 0.81)	0.13 (0.08 to 0.26)
Philippines	0.14 (0.07 to 0.21)	0.23 (0.13 to 0.32)	0.05 (−0.02 to 0.13)
American Samoa (United States)	0.53 (0.39 to 0.89)	0.60 (0.52 to 0.94)	0.20 (0.09 to 0.85)
Fiji	0.33 (0.30 to 0.67)	0.42 (0.37 to 0.81)	0.18 (0.13 to 0.59)
Costa Rica	0.17 (0.13 to 0.33)	0.22 (0.16 to 0.43)	0.08 (0.04 to 0.16)
Haiti	0.12 (0.08 to 0.37)	0.29 (0.24 to 0.66)	0.05 (0.03 to 0.12)
Bahamas	0.41 (0.36 to 0.50)	0.58 (0.52 to 0.71)	0.06 (0.03 to 0.11)

Comparisons of Elicited Epidemic Probabilities With Known Epidemics

We next calculated the Pearson correlation of the weekly indicator variable with the mean elicited percentage for GPT-3.5 and GPT-4 in response to prompt 1. We note that because conjunctivitis is not typically reportable (except under special circumstances), no comprehensive set of known epidemics is available—weeks coded as *not epidemic related* could have contained epidemics. As conjunctivitis outbreaks no longer seem to be reported on the Program for Monitoring Emerging Diseases (ProMED) system, we restricted the analysis to the same time period as our earlier report [14]. The correlations with these epidemic indicators were smaller than those with the tweet counts and were effectively 0 in India, Costa Rica, Martinique (France), and the Philippines; the correlations were substantial for American Samoa. Smaller but nonetheless indicative results were found for Fiji, Guam, and Haiti (for GPT-4). For large nations, we found correlations that were lower than those for small countries or island territories, as expected based on our earlier findings [14]. As before, broadly higher correlations were found for GPT-4 (P=.004, Wilcoxon signed rank test). A summary of these correlations is presented in Table 6.

In Table 7, for each country, we computed the average of available elicited probabilities for the months containing a known epidemic start date and for the months without. For 8 (89%) of the 9 countries, this average was larger in the months with an epidemic start date than in the months without. To potentially improve specificity, we also calculated the mean of only those elicited probabilities that were ≥51% (in an unprespecified analysis). These findings are shown in the last 2 columns of Table 7; of the 9 countries, 5 (56%) had a much higher difference between the epidemic and nonepidemic months.

Figure 4 shows weekly mean elicited probabilities compared with epidemic and nonepidemic weeks and with weekly tweet volume for 3 selected countries. In the left column, we show in red the weekly mean of all available GPT-4–derived percentage likelihoods (with 0 when there are none); in the right column, we show the mean of all GPT-4 percentage likelihoods that are ≥51%. The green bands indicate epidemic periods (4 weeks before through 6 weeks after the reported start date of known epidemics). Not all conjunctivitis outbreaks are known and reported. For American Samoa, high weekly likelihood values corresponded with the peak in tweet volume and the known outbreak, whereas for some larger countries, such as India, this was not as apparent. In general, plots of weekly means of all likelihoods >51% provide a potentially more useful visualization of likely epidemics.

XSL•FO
RenderX

Table 6. Correlation of weekly mean GPT-3.5– and GPT-4–elicited epidemic probabilities with a weekly epidemic indicator (a time series taking the value 1 for the first 4 weeks of known reported outbreaks and 0 otherwise).

Country	GPT-3.5, prompt 1, weekly mean, *r* (95% CI)	GPT-4, prompt 1, weekly mean, *r* (95% CI)	GPT-3.5 "YES," prompt 2, weekly mean, *r* (95% CI)
India	−0.03 (−0.13 to 0.08)	−0.00 (−0.16 to 0.12)	−0.00 (−0.05 to 0.07)
Guam (United States)	0.05 (0.01 to 0.10)	0.13 (0.07 to 0.22)	−0.01 (−0.02 to −0.00)
Martinique (France)	0.03 (−0.02 to 0.09)	0.05 (−0.02 to 0.11)	0.02 (−0.01 to 0.07)
Philippines	0.06 (0.01 to 0.10)	0.07 (0.02 to 0.13)	0.01 (0.00 to 0.03)
American Samoa (United States)	0.40 (0.16 to 0.81)	0.60 (0.29 to 0.85)	0.20 (−0.00 to 0.94)
Fiji	0.13 (−0.01 to 0.26)	0.24 (0.08 to 0.42)	0.08 (−0.01 to 0.57)
Costa Rica	0.05 (0.03 to 0.10)	0.06 (0.03 to 0.12)	0.07 (0.04 to 0.15)
Haiti	0.01 (−0.02 to 0.07)	0.09 (0.03 to 0.16)	−0.01 (−0.02 to −0.00)
Bahamas	0.04 (0.01 to 0.09)	0.07 (0.01 to 0.14)	0.01 (−0.01 to 0.04)

Table 7. Average monthly mean values for GPT-3.5 and GPT-4 of all elicited epidemic probabilities for months not containing a known epidemic start date (column 2) compared with months with a known epidemic start date (column 3). Monthly means of only those elicited probabilities that were ≥51% are shown for months not containing a known epidemic start date (column 4) compared with months with a known epidemic start date (column 5).

Country	Monthly mean (SD) of all probabilities, nonepidemic months, GPT-3.5	Monthly mean (SD) of all probabilities, epidemic months, GPT-4	Monthly mean (SD) of probabilities >51%, nonepidemic months, GPT-3.5	Monthly mean (SD) of probabilities >51%, epidemic months, GPT-4
India	8.4 (4.1)	9.3 (5.3)	7.2 (20.8)	36 (32.8)
Guam (United States)	7.6 (4.4)	13.6 (N/A[a])	0 (0)	0 (N/A)
Martinique (France)	6.2 (5.2)	10 (N/A)	0.9 (7.3)	0 (N/A)
Philippines	9.8 (1.1)	13.7 (N/A)	8.2 (21.0)	80 (N/A)
American Samoa (United States)	1.1 (3.8)	16.4 (N/A)	0 (0)	80 (N/A)
Fiji	4.1 (6.5)	19.1 (N/A)	0 (0)	75 (N/A)
Costa Rica	10.5 (1.4)	11.9 (N/A)	2.7 (12.5)	0 (N/A)
Haiti	7.1 (4.9)	9.5 (N/A)	1.1 (9.1)	60 (N/A)
Bahamas	8.2 (4.3)	5 (N/A)	0.9 (7.3)	0 (N/A)

[a]N/A: not applicable.

XSL•FO
RenderX

Figure 4. Representative time series plots of average weekly elicited likelihood scores for GPT-4 for 3 countries (American Samoa, India, and Costa Rica) using either all output results (column 1) or only values >50% (column 2). Red points: weekly average elicited probability for any week with at least 1 post for all values (column 1) or for only values >50% (column 2), white bars: number of tweets per week, and green band: reported epidemic period.

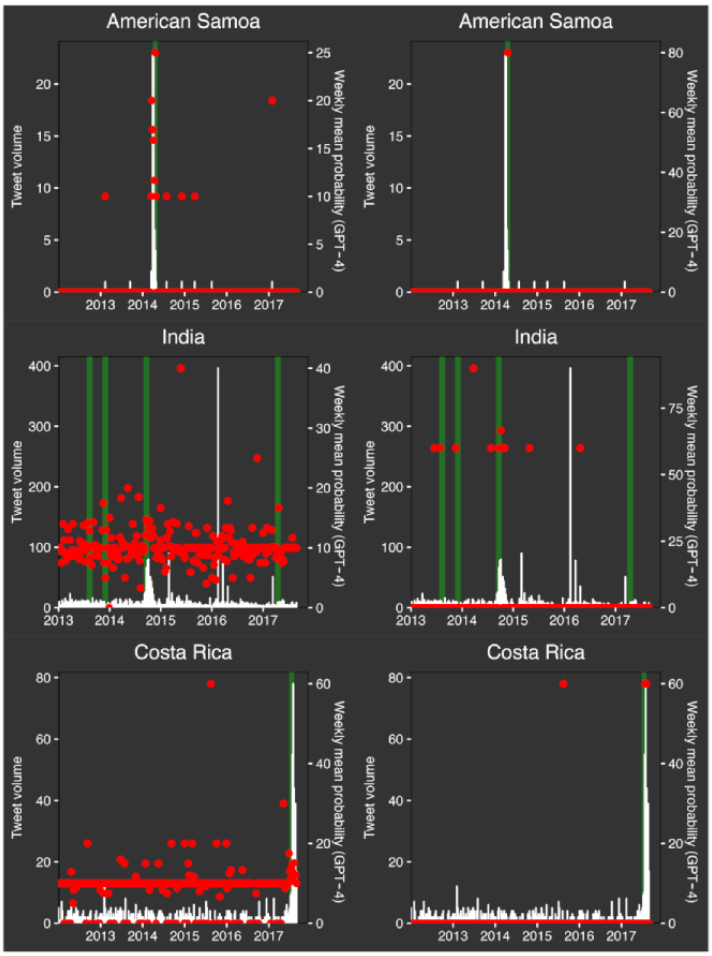

Discussion

Principal Findings

Our main findings, with regard to the objectives and hypotheses stated in the *Introduction* section, are as follows.

1. We found that LLMs can be used to assess Twitter content related to conjunctivitis in general and in relation to infectious outbreaks of conjunctivitis. We found that we could elicit percentages representing the probability of an outbreak on a regional basis (in the sense of quantifying an uncertain judgment).
2. The 2 LLMs we examined (GPT-3.5 and GPT-4) showed substantial correlation with each other's assessments of the likelihood of a conjunctivitis outbreak, as well as with the assessments of the 2 human raters.
3. We also found that these correlated with the results of other conjunctivitis-related prompts. In addition, we found evidence that the mean elicited percentages positively correlated with conjunctivitis-related tweet volume.
4. We also found evidence that these percentages correlate with known epidemics, particularly in selected small countries or island territories.

Our results suggest that our approach using a generative LLM (GPT-3.5 or GPT-4) could be used to both thematically define the contents of eye health–related tweets and assign Bayesian probability scores and classifications to help identify if a tweet is mentioning an eye disease outbreak. In view of the better performance of GPT-4 in benchmarks and tests [38-41], it is reassuring that the results from GPT-4 yielded higher correlations with tweet volume than GPT-3.5 with the same prompt. This study adds to a growing literature regarding the use of LLMs for analysis of social media posts related to health [42-45] (in our case, the assigning of a measure of health risk in addition to the interpretation of content). Future studies could explore the potential of the use of LLMs to assess the weekly content of posts about infectious eye cases to score the probability of an outbreak on a regional basis or as a low-cost weekly surveillance approach to help detect ocular epidemics. This could also validate suspected ocular epidemics determined from other web-based data sources [14]. This approach could also be applicable, in concert with topic modeling, to thematically define the content of posts regarding eye health risk. Such methods could allow for scalable thematic assessment of large sets of posts (eg, inductive content analysis [46] beyond the scope practical for time-consuming human analysis) to characterize current and emerging eye health topics of interest

XSL•FO
RenderX

to the public with specific eye conditions. Topics could be scored for factors such as toxicity in an unbiased fashion.

Future studies should assess the ability of our model to use other sources of data (web-based discussion groups, forums, or blogs) to interpret and assess the likelihood of eye disease outbreaks or other emerging eye health risks. In addition, we could explore the ability of these models to classify other key informative features of an outbreak, such as health severity, etiology, or size. Although we have chosen conjunctivitis as a model (and certainly conjunctivitis outbreaks can act as a harbinger of a systemic and higher-risk disease), the principles used to develop this model can be applied to identify outbreaks of symptoms associated with a wide range of localized or systemic diseases that pose severe population health risk or threaten a pandemic [8,9,47], especially when these symptoms may be nonreportable.

This study highlights a relatively new use of LLMs for infodemiology and suggests potential for more efficient assessment of social media than in prior works; for example, scalable thematic assessment of large sets of posts could be completed by LLMs with less manual effort required than in prior studies [46]. As LLMs continue to be developed, we anticipate that the quality of such assessments by LLMs will continue to improve and that costs will fall. In addition, new discoveries about improving methods of prompting LLMs for better results are steadily emerging. Investments in automated content screening of microblog posts, as well as other public social media, blog, and forum data, may be warranted as an additional channel of potentially useful information for disease outbreak surveillance. Such methods could be particularly useful for other nonreportable conditions.

Limitations

Our findings are subject to several limitations beyond those inherent in the selection of our 9 countries. Some relevant tweets may have been omitted because of our efforts to remove cinematic, celebrity-related, and other irrelevant content, and we note that an important potential application of LLMs is to help identify such content for elimination. It is also possible that our original query was missing some conjunctivitis-related keywords for some of the languages used in the countries included in our study, leading us to obtain low counts of posts about conjunctivitis in some languages. Future studies could further explore and expand keywords in other languages to improve our data signal for use in LLM analyses. Our prompts could be further optimized for the elicitation of probability scores from the LLMs with improved results [48]. Another limitation we found was that the LLM-elicited percentages did not correlate as well with known epidemics in large countries as they did with known epidemics in selected small countries or island territories. A possible reason for this could be that small disease outbreaks in large countries may occur frequently but go undetected when analyzing content for the entire country—this suggests that analysis of posts geolocated to smaller regions may prove more useful for detecting disease outbreaks in large countries.

In addition, tweets from some of the countries (eg, India, Martinique, Haiti, and Costa Rica) contained substantial content in other languages, and the current generations of the GPT models are somewhat less skillful in non-English languages. We note, however, that the models were entirely capable of translating and explaining content in many languages, which included Japanese, Marathi, and others in our sample [49,50], although we note a higher fraction of unusable replies for Haiti for GPT-3.5, prompt 2. Additional sources of social media data beyond Twitter could improve coverage and sensitivity. We also note that although the current LLMs were capable of replying with probabilities (expressed as percentages) seemingly indicating a degree of belief—with such values correlated with those of human raters—we have no evidence that these probabilities are calibrated (in the sense that the empirical relative frequency of true epidemics among tweets classified as probability X is, in fact, X). Finally, no complete database exists for known conjunctivitis outbreaks; therefore, it is not possible to precisely evaluate the sensitivity or specificity of our methods at this time.

Conclusions

Our findings suggest that GPT prompting can efficiently assess the content of social media posts and possible disease outbreaks to a degree of accuracy comparable to that of humans. Furthermore, we found that the results of our automated content analysis of tweet content is related to tweet volume for conjunctivitis-related posts in some locations as well as to the occurrence of actual epidemics. Future work may improve the sensitivity and specificity of these methods. The approaches presented in this manuscript suggest the potential to leverage LLMs to assess social media or forum posts not only for automated and highly efficient identification of infectious eye disease outbreaks and other emerging eye health risks but also to detect outbreaks of high-risk diseases or classify key epidemiological characteristics of cases during outbreaks. This could improve timely identification of the most severe disease outbreaks, enabling localized action for mitigating impact on human health.

Acknowledgments

This work was supported in part by a grant from the National Eye Institute of the National Institutes of Health (1R01EY024608-01A1; Principal Investigator [PI]: TML), a Core Grant for Vision Research from the National Eye Institute (EY002162; PI: Erik M Ullian), and an Unrestricted Grant from Research to Prevent Blindness (PI: JLD). The funding organizations had no role in the design or conduct of this research. The authors gratefully acknowledge the contributions of the 2 human raters used in the validation of the large language model responses. The subject of the paper concerns analysis of the results of generative artificial intelligence (AI)–where we studied the use of the generative AI tools GPT-3.5 and GPT-4 (OpenAI) for LLM-based analyses of social media posts. Generative AI was not used in ideation, or manuscript writing, or preparation, other than for assistance from GPT-3.5 in helping us to refine the R code to improve the formatting of the figures and some tables.

XSL•FO
RenderX

Data Availability

The primary social media data sets analyzed during this study are not publicly available owing to our terms of use agreement with the Brandwatch platform but are available from the corresponding author using a data sharing agreement on reasonable request. We have placed the posts in the Qualitative Data Repository via the University of California San Francisco.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Boolean query and additional tables.
[PDF File (Adobe PDF File), 259 KB - jmir_v26i1e49139_app1.pdf]

References

1. Pepose JS, Sarda SP, Cheng WY, McCormick N, Cheung HC, Bobbili P, et al. Direct and indirect costs of infectious conjunctivitis in a commercially insured population in the United States. Clin Ophthalmol 2020;14:377-387 [FREE Full text] [doi: 10.2147/OPTH.S233486] [Medline: 32103884]
2. Filleul L, Pagès F, Wan GC, Brottet E, Vilain P. Costs of conjunctivitis outbreak, Réunion Island, France. Emerg Infect Dis 2018 Jan;24(1):168-170 [FREE Full text] [doi: 10.3201/eid2401.170916] [Medline: 29260662]
3. Lernout T, Maillard O, Boireaux S, Collet L, Filleul L. A large outbreak of conjunctivitis on Mayotte Island, France, February to May 2012. Euro Surveill 2012 Jun 07;17(23):20192 [FREE Full text] [Medline: 22720741]
4. Centers for Disease ControlPrevention (CDC). Adenovirus-associated epidemic keratoconjunctivitis outbreaks--four states, 2008-2010. MMWR Morb Mortal Wkly Rep 2013 Aug 16;62(32):637-641 [FREE Full text] [Medline: 23945769]
5. Sié A, Diarra A, Millogo O, Zongo A, Lebas E, Bärnighausen T, et al. Seasonal and temporal trends in childhood conjunctivitis in Burkina Faso. Am J Trop Med Hyg 2018 Jul;99(1):229-232 [FREE Full text] [doi: 10.4269/ajtmh.17-0642] [Medline: 29761759]
6. Prajna NV, Lalitha P, Teja GV, Gunasekaran R, Sharma SS, Hinterwirth A, et al. Outpatient human coronavirus associated conjunctivitis in India. J Clin Virol 2022 Dec;157:105300 [FREE Full text] [doi: 10.1016/j.jcv.2022.105300] [Medline: 36209621]
7. Lalitha P, Prajna NV, Gunasekaran R, Teja GV, Sharma SS, Hinterwirth A, et al. Deep sequencing analysis of clinical samples from patients with acute infectious conjunctivitis during the COVID-19 delta surge in Madurai, India. J Clin Virol 2022 Dec;157:105318 [FREE Full text] [doi: 10.1016/j.jcv.2022.105318] [Medline: 36242841]
8. With new COVID-19 strain confirmed in Los Angeles county, residents advised to be aware of symptoms, take precautions. County of Los Angeles, Department of Public Health. URL: http://ph.lacounty.gov/phcommon/public/media/mediapubhpdetail.cfm?prid=4372 [accessed 2024-01-29]
9. Vashishtha VM, Kumar P. Preliminary clinical characteristics of pediatric COVID-19 cases during the ongoing Omicron XBB. medRxiv. Preprint posted online April 20, 2023 2023 [FREE Full text] [doi: 10.1101/2023.04.18.23288715]
10. National Notifiable Diseases Surveillance System (NNDSS). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nndss/index.html [accessed 2024-01-29]
11. Tsui E, Rao RC. Navigating social media in #Ophthalmology. Ophthalmol 2019 Jun;126(6):779-782 [FREE Full text] [doi: 10.1016/j.ophtha.2019.02.015] [Medline: 31122357]
12. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41:101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
13. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284 [FREE Full text] [doi: 10.2196/19284] [Medline: 32501804]
14. Deiner MS, McLeod SD, Wong J, Chodosh J, Lietman TM, Porco TC. Google searches and detection of conjunctivitis epidemics worldwide. Ophthalmology 2019 Sep;126(9):1219-1229 [FREE Full text] [doi: 10.1016/j.ophtha.2019.04.008] [Medline: 30981915]
15. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. JAMA Ophthalmol 2016 Sep 01;134(9):1024-1030 [FREE Full text] [doi: 10.1001/jamaophthalmol.2016.2267] [Medline: 27416554]
16. Deiner MS, McLeod SD, Chodosh J, Oldenburg CE, Fathy CA, Lietman TM, et al. Clinical age-specific seasonal conjunctivitis patterns and their online detection in twitter, blog, forum, and comment social media posts. Invest Ophthalmol Vis Sci 2018 Feb 01;59(2):910-920 [FREE Full text] [doi: 10.1167/iovs.17-22818] [Medline: 29450538]
17. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]

XSL•FO
RenderX

18. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. J Biomed Inform 2020 Aug;108:103500 [FREE Full text] [doi: 10.1016/j.jbi.2020.103500] [Medline: 32622833]
19. Kass-Hout TA, Alhinnawi H. Social media in public health. Br Med Bull 2013;108:5-24. [doi: 10.1093/bmb/ldt028] [Medline: 24103335]
20. Pilipiec P, Samsten I, Bota A. Surveillance of communicable diseases using social media: a systematic review. PLoS One 2023;18(2):e0282101 [FREE Full text] [doi: 10.1371/journal.pone.0282101] [Medline: 36827297]
21. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. J Med Internet Res 2014 Oct 20;16(10):e236 [FREE Full text] [doi: 10.2196/jmir.3416] [Medline: 25331122]
22. Sano Y, Hori A. 12-year observation of tweets about rubella in Japan: a retrospective infodemiology study. PLoS One 2023;18(5):e0285101 [FREE Full text] [doi: 10.1371/journal.pone.0285101] [Medline: 37155655]
23. Aslam AA, Tsou M, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. J Med Internet Res 2014 Nov 14;16(11):e250 [FREE Full text] [doi: 10.2196/jmir.3532] [Medline: 25406040]
24. Zhang Y, Chen K, Weng Y, Chen Z, Zhang J, Hubbard R. An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. Expert Syst Appl 2022 Jul 15;198:116882 [FREE Full text] [doi: 10.1016/j.eswa.2022.116882] [Medline: 35308584]
25. Al-Garadi MA, Khan MS, Varathan KD, Mujtaba G, Al-Kabsi AM. Using online social networks to track a pandemic: a systematic review. J Biomed Inform 2016 Aug;62:1-11 [FREE Full text] [doi: 10.1016/j.jbi.2016.05.005] [Medline: 27224846]
26. Gao W, Li L, Tao X, Zhou J, Tao J. Identifying informative tweets during a pandemic via a topic-aware neural language model. World Wide Web 2023;26(1):55-70 [FREE Full text] [doi: 10.1007/s11280-022-01034-1] [Medline: 35308294]
27. Wakamiya S, Kawai Y, Aramaki E. Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. JMIR Public Health Surveill 2018 Sep 25;4(3):e65 [FREE Full text] [doi: 10.2196/publichealth.8627] [Medline: 30274968]
28. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. JMIR Public Health Surveill 2016 Oct 20;2(2):e161 [FREE Full text] [doi: 10.2196/publichealth.5901] [Medline: 27765731]
29. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
30. Canaparo M, Ronchieri E, Scarso L. A natural language processing approach for analyzing COVID-19 vaccination response in multi-language and geo-localized tweets. Healthc Anal (N Y) 2023 Nov;3:100172 [FREE Full text] [doi: 10.1016/j.health.2023.100172] [Medline: 37064254]
31. Jungwirth D, Haluza D. Artificial intelligence and public health: an exploratory study. Int J Environ Res Public Health 2023 Mar 03;20(5):4541 [FREE Full text] [doi: 10.3390/ijerph20054541] [Medline: 36901550]
32. ChatGPT. OpenAI. URL: https://openai.com/chatgpt [accessed 2024-02-14]
33. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online May 28, 2020 2020 [FREE Full text]
34. Cumberland PM, Czanner G, Bunce C, Doré CJ, Freemantle N, García-Fiñana M, et al. Ophthalmic statistics note: the perils of dichotomising continuous variables. Br J Ophthalmol 2014 Jun;98(6):841-843 [FREE Full text] [doi: 10.1136/bjophthalmol-2014-304930] [Medline: 24682179]
35. Mason S, Singh L. Reporting and discoverability of “tweets” quoted in published scholarship: current practice and ethical implications. Res Ethics 2022 Feb 03;18(2):93-113 [FREE Full text] [doi: 10.1177/17470161221076948]
36. Joshi A, Sparks R, McHugh J, Karimi S, Paris C, MacIntyre CR. Harnessing tweets for early detection of an acute disease event. Epidemiol 2020 Jan;31(1):90-97 [FREE Full text] [doi: 10.1097/EDE.0000000000001133] [Medline: 31651659]
37. Kreiss J, Lahiri SN. Bootstrap methods for time series. In: Rao TS, Rao SS, Rao CR, editors. Time Series Analysis: Methods and Applications. Amsterdam, The Netherlands: North-Holland (Elsevier); 2012.
38. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023 Nov 22;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
39. Ateia S, Kruschwitz U. Is ChatGPT a biomedical expert? -- exploring the zero-shot performance of current GPT models in biomedical tasks. arXiv. Preprint posted online on June 28, 2023 2023 [FREE Full text]
40. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on March 20, 2023 2023 [FREE Full text]
41. Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and bard on the Japanese national dentist examination: a comparison study. Cureus 2023 Dec;15(12):e50369 [FREE Full text] [doi: 10.7759/cureus.50369] [Medline: 38213361]
42. Zhang L, Fan H, Peng C, Rao G, Cong Q. Sentiment analysis methods for HPV vaccines related tweets based on transfer learning. Healthcare (Basel) 2020 Aug 28;8(3):307 [FREE Full text] [doi: 10.3390/healthcare8030307] [Medline: 32872330]

XSL•FO
RenderX

43. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. JMIR Med Educ 2023 Mar 08;9:e46876 [FREE Full text] [doi: 10.2196/46876] [Medline: 36867743]
44. Saito R, Haruyama S. Estimating time-series changes in social sentiment @Twitter in U.S. metropolises during the COVID-19 pandemic. J Comput Soc Sci 2023;6(1):359-388 [FREE Full text] [doi: 10.1007/s42001-022-00186-4] [Medline: 36405087]
45. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
46. Honcharov V, Li J, Sierra M, Rivadeneira NA, Olazo K, Nguyen TT, et al. Public figure vaccination rhetoric and vaccine hesitancy: retrospective Twitter analysis. JMIR Infodemiology 2023;3:e40575 [FREE Full text] [doi: 10.2196/40575] [Medline: 37113377]
47. Outbreak of extensively drug-resistant Pseudomonas aeruginosa associated with artificial tears. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hai/outbreaks/crpa-artificial-tears.html [accessed 2024-01-29]
48. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on January 28, 2022 2022 [FREE Full text]
49. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on March 27, 2023 2023 [FREE Full text]
50. Jiao W, Wang W, Huang JT, Wang X, Shi S, Tu Z. Is ChatGPT a good translator? yes with GPT-4 as the engine. arXiv. Preprint posted online on January 20, 2023 2023 [FREE Full text]

Abbreviations

API: application programming interface
GPT: Generative Pre-trained Transformers
LLM: large language model
ProMED: Program for Monitoring Emerging Diseases

Edited by G Eysenbach, A Mavragani; submitted 19.05.23; peer-reviewed by S Pesälä, G Gonzalez-Hernandez; comments to author 21.11.23; revised version received 20.12.23; accepted 19.01.24; published 01.03.24.

Please cite as:
Deiner MS, Deiner NA, Hristidis V, McLeod SD, Doan T, Lietman TM, Porco TC
Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study
J Med Internet Res 2024;26:e49139
URL: https://www.jmir.org/2024/1/e49139
doi:10.2196/49139
PMID:38427404

©Michael S Deiner, Natalie A Deiner, Vagelis Hristidis, Stephen D McLeod, Thuy Doan, Thomas M Lietman, Travis C Porco. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Traditional and New Media on Smoking Intentions and Behaviors: Secondary Analysis of Tasmania's Tobacco Control Mass Media Campaign Program, 2019-2021

James Kite[1], MPH, PhD; Anne Grunseit[1], PhD; Glenn Mitchell[2], DBM; Pip Cooper[2], MSocSc; Lilian Chan[1], PhD; Bo-Huei Huang[3], PhD; Margaret Thomas[1], DrPH; Blythe O'Hara[1], PhD; Abby Smith[1,2], MS

[1]Prevention Research Collaboration, Sydney School of Public Health and Charles Perkins Centre, The University of Sydney, Camperdown, NSW, Australia

[2]Quit Tasmania, Cancer Council Tasmania, Hobart, Australia

[3]School of Public Health, Faculty of Health, University of Technology Sydney, Sydney, Australia

Corresponding Author:
James Kite, MPH, PhD
Prevention Research Collaboration
Sydney School of Public Health and Charles Perkins Centre
The University of Sydney
L6 Charles Perkins Centre (D17)
Camperdown, NSW, 2006
Australia
Phone: 61 286270838
Email: james.kite@sydney.edu.au

Abstract

Background: Tasmania, the smallest state by population in Australia, has a comprehensive tobacco control mass media campaign program that includes traditional (eg, television) and "new" channels (eg, social media), run by Quit Tasmania. The campaign targets adult smokers, in particular men aged 18-44 years, and people from low socioeconomic areas.

Objective: This study assesses the impact of the 2019-2021 campaign program on smokers' awareness of the campaign program, use of Quitline, and smoking-related intentions and behaviors.

Methods: We used a tracking survey (conducted 8 times per year, immediately following a burst of campaign activity) to assess campaign recall and recognition, intentions to quit, and behavioral actions taken in response to the campaigns. The sample size was approximately 125 participants at each survey wave, giving a total sample size of 2000 participants over the 2 years. We merged these data with metrics including television target audience rating points, digital and Facebook (Meta) analytics, and Quitline activity data, and conducted regression and time-series modeling.

Results: Over the evaluation period, unprompted recall of any Quit Tasmania campaign was 18%, while prompted recognition of the most recent campaign was 50%. Over half (52%) of those who recognized a Quit Tasmania campaign reported that they had performed or considered a quitting-related behavioral action in response to the campaign. In the regression analyses, we found having different creatives within a single campaign burst was associated with higher campaign recall and recognition and an increase in the strength of behavioral actions taken. Higher target audience rating points were associated with higher campaign recall (but not recognition) and an increase in quit intentions, but not an increase in behavioral actions taken. Higher Facebook advertisement reach was associated with lower recall among survey participants, but recognition was higher when digital channels were used. The time-series analyses showed no systematic trends in Quitline activity over the evaluation period, but Quitline activity was higher when Facebook reach and advertisement spending were higher.

Conclusions: Our evaluation suggests that a variety of creatives should be used simultaneously and supports the continued use of traditional broadcast channels, including television. However, the impact of television on awareness and behavior may be weakening. Future campaign evaluations should closely monitor the effectiveness of television as a result. We are also one of the first studies to explicitly examine the impact of digital and social media, finding some evidence that they influence quitting-related outcomes. While this evidence is promising for campaign implementation, future evaluations should consider adopting rigorous methods to further investigate this relationship.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47128) doi:10.2196/47128

KEYWORDS

mass media campaign; tobacco control; evaluation; social media campaign; social media; digital platform; tobacco; smoking; survey

Introduction

Tobacco smoking is a public health priority in Tasmania, Australia, as it causes a high burden of disease and death [1], and this state has a higher smoking prevalence compared to the national average [2]. As such, the state committed to reducing the burden of tobacco smoking through the Tasmanian Tobacco Control Plan, 2017-2021 [3]. Under that plan, Quit Tasmania, part of Cancer Council Tasmania, has responsibility for delivering and evaluating core components of the state's comprehensive tobacco control program that reinforce and complement each other, including the Quitline telephone counseling service, community interventions to increase quitting activity and use of Quitline, and the mass media campaign program (MMCP). The MMCP aims to motivate smokers to quit smoking and support recent quitters to maintain abstinence, as well as discourage nonsmokers from starting. The program includes television-led mass media campaigns, which are the focus of this evaluation.

Tasmania's MMCP aligns with international evidence for tobacco control mass media campaigns, showing that campaigns are a vital part of comprehensive tobacco control strategies [4]. However, the evidence predominantly comes from larger markets. Tasmania is Australia's smallest state by population with approximately 571,000 residents [5], meaning that the primary audience pool for the MMCP (smokers) is relatively small at 70,500 people. This creates unique challenges for the design, implementation, and evaluation of the MMCP, including difficulties in audience segmentation and recruitment for evaluation surveys. Comprehensive evaluations, such as this one, are therefore useful in expanding the evidence base.

The MMCP has 3 target audiences: adult smokers aged 18+ years (the primary target audience for all campaign activity); men aged 18-44 years; and those from low socioeconomic areas. These latter audiences reflect higher smoking prevalence in these groups [6]. In addition to television, the MMCP includes channels such as radio, out-of-home, digital (eg, Google Adwords and digital display advertising), social media (particularly Facebook), community events, and public relations. The television-led campaigns usually run 8 times a year in 4-week bursts (a total of 32 weeks a year), aiming to deliver 700 target audience rating points (TARPs; an estimate of target audience reach and frequency of exposure to the campaigns) for the aged 18+ years audience per month. Digital and social media activity run continuously. Campaign creative includes both "why to quit" (usually graphic or testimonial-style messages) and "how to quit" messages (those that promote quit support information and services). While campaign creatives on social media were usually independent of the television-led campaigns, a common call-to-action in the MMCP activities is to encourage smokers to contact the Quitline. For example, Facebook advertisement goals were set up to enhance reach and engagement, with content designed to target men aged 18-44 years and people from low socioeconomic areas. Facebook users who visited the Quit Tasmania website were retargeted with Facebook advertisements aimed to direct users to specific actions such as ordering a quit pack (click on the website) or contacting the Quitline.

This study assessed the impact of Tasmania's tobacco control MMCP's public education campaigns from 2019 to 2021 on adult smokers' and recent quitters' intentions and behaviors relating to smoking. Specifically, we asked the following questions:

- To what extent were the target audiences aware of the MMCP's campaigns?
- What level of use and engagement has there been with Quitline during 2019-2021?
- Has there been a change in target audiences' intentions to quit smoking or smoking behavior, including quitting salience, quit attempts, and quit successes?

The findings of this study will inform the future use of MMCP components by Quit Tasmania and provide comparative evidence of the different communication channels for tobacco control more broadly.

Methods

Ethical Considerations

This secondary analysis project was approved by The University of Sydney Human Research Ethics Committee (2021/779). Participants were asked to provide informed consent before commencing the research. All data were anonymous. No compensation was provided to participants for their time.

Data Sources

Evaluation of the MMCP centers on a tracking survey was conducted 8 times per year by the research company, EMRS. Each survey wave ran for 2 weeks, with the first week overlapping with the final week of each campaign burst. The surveys were conducted in August, September-October, November, and December 2019; January-February, March, April-May, June, August, September, October, and November-December 2020; and January, March, April, and June 2021. Approximately 125 adult smokers (having smoked at least 100 cigarettes in their lifetime self-identified as "current smokers") and recent quitters (having smoked at least 100 cigarettes in their lifetime and quit smoking in the last 12 months) were interviewed at each wave. Data are collected on campaign awareness, understanding of campaign messages, and behavioral actions taken relating to the television-led campaigns (only).

Interviews in 2019-2021 were initially carried out via face-to-face intercept surveys at 3 regional centers in Tasmania:

XSL•FO
RenderX

Hobart, Launceston, and Devonport. Interviewers read the survey questions to participants and recorded their responses on tablets. During the COVID-19 pandemic (March 2020-June 2021), recruitment and data collection switched to be predominantly by computer-assisted-telephone interviewing and web-based surveys. Recruitment for the telephone and web-based surveys included targeted calls to people on an EMRS panel who had previously expressed interest in being part of research, cold calling, inviting EMRS's web-based panel members, as well as an external panel run by Cint, to complete the survey in exchange for rewards, and advertising via Facebook and Instagram (Meta Platforms). Recruitment via social media was capped at a maximum of 10% of the sample for each year of the survey (ie, 2019-2020 and 2020-2021).

For face-to-face and telephone recruitment methods, interviewers explained the nature and purpose of the research to potential participants before seeking consent to proceed with the survey. A similar introductory text was provided to those invited to complete the web-based survey, including advice that completing the survey would be taken as providing informed consent. Sampling quotas for smoking status and age were applied to ensure that the sample included sufficient numbers of smokers and recent quitters and in the aged 18-44 years age group (to align with one of the secondary target audiences).

Data from the tracking survey were merged with channel and campaign metrics associated with the most recent campaign so we could examine the relationship between the most recent campaign's components and participants' outcomes. Campaign-related metrics included paid and bonus TARPs for the aged 18+ years audience, Facebook, website, and digital media analytics, and relevant placement and exposure metrics for out-of-home advertising, along with Quitline activity data.

Outcome Measures

From the tracking survey, we used campaign recall (unprompted recall of the relevant Quit Tasmania campaign) and campaign recognition (prompted recognition of the relevant Quit Tasmania campaign) to assess awareness. We combined reported "planning to quit in the next 30 days," "currently trying to quit," and "have tried to quit in last 6 months but am back smoking again" to identify people currently smoking who had tried, or were considering, quitting. Finally, participants who reported having considered or taken any behavioral action (quit, cut down, contacted Quitline, visited a quit smoking website, or contacted a health professional or service) in response to a Quit Tasmania campaign were classified as having "taken behavioural action." A behavioral action score was also created by summing across all actions with those who reported "considering" an action (eg, considered cutting down cigarettes) receiving a 1 and those who reported they had "performed" that action receiving a 2 to reflect the strength of action. From 5 behaviors therefore a minimum score of 0 and a maximum of 10 were possible. All survey items used in analyses are listed in full in Multimedia Appendix 1.

From the Quitline data, we extracted the total number of contacts per week (including counseling requests and quit pack-only requests) and the number of quit pack-only requests.

Statistical Analyses

We used generalized linear models with a binomial distribution and log link regression to analyze the tracking survey data for awareness (unprompted recall and prompted recognition) and whether the participant called the Quitline in response to the campaign (among those who recognized the most recent campaign) and had intentions to quit (among current smokers). We used Poisson regression for the total number of behavioral actions taken. The results are expressed as prevalence ratios or incident rate ratios (for behavioral action score) depending on the model and described as a percentage change in the outcome per unit increase in the predictor.

We also undertook time-series analyses (autoregressive integrated moving average models) of the Quitline data to examine trends over time and the association between campaign activity and Quitline engagement outcomes (quit pack requests and Quitline activity). The results are expressed as a change in the weekly measure of the outcome (bus advertisements were dropped as a predictor from the model of weekly quit-pack requests due to collinearity).

Covariates

Demographic data collected in the survey included gender identity, age, and socioeconomic status according to Socio-Economic Indexes for Areas quintiles of disadvantage [7]. Gender identity was limited to a binary response option (ie, man or woman) and, as this was a secondary analysis and the data were collected by a third-party research agency, we are unable to determine how nonbinary and gender-diverse responses were categorized.

Campaign information included in each generalized linear model were smoking status (recent quitter vs current smoker), men aged 18-44 years (vs men aged older than 44 years and all women), socioeconomic status (quintiles 1-2 vs 3-5), paid and bonus TARPs (in hundreds), Facebook reach (in thousands), Facebook or Google Adwords impressions (in thousands), use of out-of-home advertising (yes or no), resources (eg, posters and other marketing collateral such as stickers and sidewalk stencils; yes or no), displays (yes or no), digital (yes or no), and the number of creative executions that run in a campaign burst (2 versus 1). To emphasize that these were the target audiences, we have labelled men aged 18-44 years as "audience 1" and low socioeconomic status (Socio-Economic Indexes for Areas quintiles 1-2) as "audience 2" in the results tables.

Campaign information included in the time-series models were weekly measures of paid and bonus TARPs, organic and paid Facebook reach (in thousands), organic and paid Facebook impressions (in thousands), Facebook total spend, presence of digital campaign activity (yes or no), use of radio advertisements (yes or no), billboards (yes or no), bus advertisements (yes or no), and "other" channels (yes or no).

Results

Tracking Survey Sample Characteristics

Approximately one-fifth (21.6%) of the tracking survey sample were men aged 18-44 years, while 50% were from areas of the

XSL•FO
RenderX

most socioeconomic disadvantage (Table 1). Additionally, 84.1% were current smokers and 15.9% were recent quitters.

Table 1. Sample characteristics (unweighted) for campaign tracking survey of adult smokers and recent quitters in Tasmania, Australia, 2019-2021 (N=2000).

Characteristic	n (%)
Gender	
Men	871 (43.6)
Women	1129 (56.5)
Age (years)	
18-34	623 (31.2)
35-54	675 (33.8)
55+	702 (35.1)
Men 18-44 years (audience 1)	431 (21.6)
Socioeconomic status[a]	
Low SES[b] (audience 2)	1000 (50)
High SES	1000 (50)
Smoking status	
Current smoker	1682 (84.1)
Recent quitter	318 (15.9)

[a]Socio-Economic Indexes for Areas quintiles of disadvantage: low SES—quintiles 1-2 and high SES—quintiles 3-5.

[b]SES: socioeconomic status.

Awareness of MMCP

Over the evaluation period, 18% of survey participants recalled unprompted any Quit Tasmania campaign, while just over half (50.3%) recognized the most recent campaign (Table 2). Adjusted for all other variables, unprompted recall and prompted recognition were 48% and 35% lower respectively among men aged 18-44 years, compared to all other participants, and 87% and 34% higher respectively when 2 creative executions were run in a single burst, compared to when only one was run. Recall, but not recognition, was also 19% higher for each additional 100 bonus TARPs and 67% higher when out-of-home channels were used. However, recall was 2% lower for every 1000 additional Facebook users reached through Facebook advertisements. Recognition, but not recall, was 14% higher among recent quitters and 17% higher when digital channels were used, but it was 3% lower for every 1000 additional Google Adwords impressions and 28% lower when display advertising was used.

XSL•FO
RenderX

Table 2. Adjusted prevalence ratios for awareness of Quit Tasmania campaigns, 2019-2021 (campaign tracking survey data from adult smokers and recent quitters in Tasmania, Australia, N=2000). Percentages are weighted.

Characteristic	Recall, %	Adjusted prevalence ratio, 95% CI	Recognition, %	Adjusted prevalence ratio, 95% CI
Overall	18.1	__[a]	50.3	—
Audience 1				
Men aged older than 44 years and all women	21.5	Reference	54.3	Reference
Men aged 18-44 years	11.3	0.52 (0.37-0.72)	34.4	0.65 (0.55-0.76)
Audience 2				
High SES[b]	18.9	Reference	48.4	Reference
Low SES	17.2	0.93 (0.76-1.14)	46.9	0.97 (0.88-1.07)
Smoking status				
Current smokers	17.3	Reference	45.7	Reference
Recent quitters	20.5	1.21 (0.94-1.56)	53.7	1.14 (1.01-1.28)
Paid TARPs[c]/100	—	1.00 (0.90-1.10)	—	0.99 (0.95-1.03)
Bonus TARPs/100	—	1.19 (1.10-1.28)	—	1.03 (1.00-1.07)
FB[d] reach per 1000	—	0.98 (0.97-0.99)	—	0.99 (0.99-1.00)
Google Adwords impressions/1000	—	0.99 (0.96-1.02)	—	0.97 (0.95-0.98)
Out-of-home				
No	14.6	Reference	48	Reference
Yes	20.3	1.67 (1.26-2.21)	47.5	1.02 (0.89-1.17)
Resources				
No	23.8	Reference	32.4	Reference
Yes	17.8	0.54 (0.23-1.25)	48.7	1.06 (0.67-1.67)
Display				
No	18.4	Reference	52.2	Reference
Yes	17.9	0.86 (0.62-1.20)	43.2	0.72 (0.60-0.86)
Digital				
No	19.4	Reference	48.9	Reference
Yes	16.5	1.16 (0.82-1.65)	46.1	1.17 (1.01-1.36)
Creative executions per burst				
1	16.1	Reference	49.7	Reference
2	20.2	1.87 (1.39-2.53)	45.7	1.34 (1.15-1.55)

[a]—not available.

[b]SES: socioeconomic status.

[c]TARP: target audience rating point.

[d]FB: Facebook.

Use and Engagement With Quitline

On average, Quitline received 48 (SD 17.6) contacts per week between June 2019 and June 2021, with approximately half (26, SD 8.2) being requests for counseling. We found no significant trend in Quitline activity over the evaluation period (Table 3). Higher paid Facebook reach and higher Facebook advertisement spending were both associated with increased Quitline activity. Conversely, higher-paid Facebook impressions were associated with lower Quitline activity. We also found that Quitline activity declined when radio was used as a channel but increased when billboards were used.

XSL•FO
RenderX

Table 3. Facebook analytics and Quitline interactions time-series analysis results, 2019-2021. Percentages are weighted.

Variable	Adjusted beta, 95% CI	
	Quitline contacts	Quitline quit pack only
Trend	–0.13 (–0.33 to 0.07)	–0.07 (–0.36 to 0.22)
Bonus TARPs[a]	0.03 (–0.03 to 0.09)	0.004 (–0.06 to 0.06)
Paid TARPs	–0.02 (–0.06 to 0.01)	0.003 (–0.04 to 0.03)
Organic reach/1000	–9.92 (–25.68 to 5.84)	–3.37 (–15.16 to 8.42)
Paid reach/1000	1.59 (0.77 to 2.41)	1.04 (0.41 to 1.68)
Organic impressions/1000	8.9 (–5.25 to 23.05)	2.84 (–8.11 to 13.78)
Paid impressions/1000	–1.19 (–1.87 to –0.51)	–0.76 (–1.29 to –0.22)
FB[b] advertisement spend/AUS $100[c]	8.58 (4.69 to 12.47)	6.75 (3.19 to 10.32)
Digital	3.21 (–7.79 to 14.21)	4.83 (–7.60 to 17.26)
Radio	–25.89 (–41.12 to –10.65)	–17.7 (–30.82 to –4.58)
Billboards	18.27 (4.47 to 32.08)	14.46 (–4.84 to 33.76)
Other[d]	6.48 (–11.27 to 24.24)	4.12 (–9.1 to 17.34)

[a]TARP: target audience rating point.

[b]FB: Facebook.

[c]AUS $1 = US $0.65 at the time of this study.

[d]For example, butt bins, messenger bots, supercars, and grassroots.

Intentions and Behavioral Responses to Campaigns

With regards to quitting intentions, each additional 100 paid TARPs was associated with an 8% increase in participants reporting that they were considering quitting (Table 4). The use of resources during campaign periods was associated with an over 2-fold increase in participants reporting that they were considering quitting, although the wide CI means that this result should be interpreted with caution. Participants who recognized the most recent campaign were over twice as likely to report that they had called the Quitline in the last 6 months. Facebook advertisement reach was associated with a 1% decline in participants considering quitting and a 4% decline in participants reporting that they had called the Quitline in the past 6 months for every 1000 Facebook users reached.

XSL•FO
RenderX

Table 4. Adjusted prevalence ratios for quitting intentions and calls to Quitline among current smokers (from campaign tracking survey of adult smokers in Tasmania, Australia, n=1682), 2019-2021[a].

Characteristic	Considered quitting[b], %	Adjusted prevalence ratio, 95% CI	Called Quitline in the last 6 months, %	Adjusted prevalence ratio, 95% CI
Audience 1				
Men aged older than 44 years and all women	35.7	Reference	7.1	Reference
Men aged 18-44 years	31.3	0.88 (0.73-1.07)	6.7	1.14 (0.72-1.80)
Audience 2				
High SES[c]	35.8	Reference	7.2	Reference
Low SES	32.4	0.90 (0.78-1.04)	6.7	0.93 (0.64-1.35)
Paid TARPs[d]/100	—[e]	1.08 (1.01-1.15)	—	1.20 (1.00-1.43)
Bonus TARPs/100	—	1.02 (0.97-1.08)	—	1.05 (0.92-1.19)
FB[f] reach/1000	—	0.99 (0.98-1.00)	—	0.96 (0.93-0.99)
Google Adwords impressions/1000	—	1.01 (0.99-1.03)	—	1.02 (0.96-1.07)
Out-of-home				
No	38.2	Reference	10.1	Reference
Yes	31.8	1.04 (0.86-1.25)	5	0.96 (0.58-1.60)
Resources				
No	26.9	Reference	2.5	Reference
Yes	34.7	2.12 (1.14-3.94)	7.2	7.58 (1.35-42.67)
Display				
No	37.7	Reference	9.5	Reference
Yes	30.1	1.02 (0.8-1.31)	4.4	0.84 (0.43-1.63)
Digital				
No	36.8	Reference	9	Reference
Yes	31	0.98 (0.77-1.25)	4.3	0.84 (0.46-1.54)
Creative executions per burst				
1	37.4	Reference	9	Reference
2	31	0.94 (0.76-1.17)	4.9	0.86 (0.48-1.56)
Recognized most recent advertisement				
No	32.8	Reference	4.4	Reference
Yes	35.9	1.03 (0.89-1.21)	9.9	2.15 (1.40-3.31)

[a]Percentages are weighted.

[b]Defined as planning to quit in the next 30 days OR trying to quit at the moment OR have tried to quit in the last 6 months but back to smoking (vs not thinking about quitting OR thinking about quitting in next 6 months).

[c]SES: socioeconomic status.

[d]TARP: target audience rating point.

[e]Not available.

[f]FB: Facebook.

Of those who recognized a Quit Tasmania campaign, 52% reported that they had considered or performed a quitting-related behavioral action in response to a campaign (Table 5). Recent quitters had a behavioral action score 1.71 times higher than current smokers. Additionally, the use of more than one creative execution in a single burst was associated with a 41% increase in the number of behavioral actions taken, while the increase was 1% per 1000 additional people reached on Facebook. However, behavioral action scores declined by 29% and 47% when digital and out-of-home channels respectively were used.

XSL•FO
RenderX

Table 5. Incidence rate ratios from the Poisson model of total number of behavioral actions taken or considered following seeing the most recent campaign (only for those in the campaign tracking survey of adult smokers and recent quitters in Tasmania, Australia, who recognized the most recent campaign n=1007), 2019-2021[a].

Characteristic	Took or considered behavioral action, %	Incidence rate ratio
Overall	51.7	__[b]
Audience 1		
Men aged older than 44 years and all women	51.2	Reference
Men aged 18-44 years	56	1.14 (0.90-1.44)
Audience 2		
High SES[c]	52.1	Reference
Low SES	52.7	1.10 (0.93-1.30)
Smoking status		
Current smokers	50.1	Reference
Recent quitters	58.1	1.71 (1.43-2.05)
Paid TARPs[d]/100	__[e]	0.99 (0.91-1.07)
Bonus TARPs/100	—	1.05 (0.99-1.11)
FB[f] reach (per 1000)	—	1.01 (1.00-1.03)
Google Adwords impressions/1000	—	0.98 (0.96-1.01)
Out-of-home		
No	63.2	Reference
Yes	45.8	0.53 (0.41-0.67)
Resources		
No	55.6	Reference
Yes	52.2	0.65 (0.28-1.52)
Display		
No	54.8	Reference
Yes	49.4	0.81 (0.56-1.19)
Digital		
No	59.8	Reference
Yes	42.2	0.71 (0.50-1.00)
Creative executions per burst		
1	54.5	Reference
2	50	1.41 (1.01-1.95)

[a]Percentages are weighted.
[b]—: not available.
[c]SES: socioeconomic status.
[d]TARP: target audience rating point.
[e]Not available.
[f]FB: Facebook.

Discussion

Our results provide important insights into implementing tobacco control public education campaigns in smaller markets. In particular, traditional broadcast channels, such as television, should continue to be used as channels in tobacco control campaigns as they seem to have a role in generating behavioral actions such as quit attempts. They also suggest that using a variety of creative executions is most effective in encouraging behavior change. Equally, though, our findings highlight some areas for further research, especially about how best to use traditional channels and digital and social channels together.

The MMCP exceeded its targeted 700 TARPs per month in most months during the evaluation period. This figure aligns

XSL•FO
RenderX

with available evidence on ideal TARP weights for television campaigns [8-10]. However, studies examining ideal TARP weights were largely conducted before the rapid growth of social media and streaming services, meaning that the ongoing relevance of this evidence is unclear. We did find that higher bonus TARPs were associated with higher recall and that higher paid TARPs were associated with increased quitting intentions, which collectively provide some support for the ongoing relevance of existing evidence on appropriate TARP levels. However, we also note that higher paid TARPs were not associated with increased recall or recognition, higher bonus TARPs were not associated with increased quit intentions or Quitline contacts, and there was no apparent relationship between TARPs and behavioral actions in response to a campaign. While our results provide some support for the ongoing use of television as a major channel for tobacco control campaigns, they also suggest that campaign evaluations should closely monitor the effectiveness of television as a channel, as the relationship between TARPs and behavioral outcomes may be weakening. This is especially important given how resource-intensive television-led campaigns are to produce and implement.

Our study is one of the few to look at the association between key campaign outcomes and channels other than television. Notably, there were some unexpected and conflicting relationships between digital and social media channels and campaign outcomes. This included Facebook advertisement reach being associated with lower recall and reduced intention to quit but increased Quitline activity and stronger behavioral action as a result of seeing the campaigns. Similarly, digital channels were associated with increased recognition but also reduced reporting of behavioral actions. Further, 1 possible explanation for these results is that the MMCP content on Facebook was usually unrelated to the television-led campaigns. Another possibility is there are significant demographic differences between our survey participants and Facebook users, although this risk seems small given 89% of Australians have an active Facebook account [11]. Greater synergy between the different channels may therefore be worth exploring in future campaigns. It is also possible that unmeasured factors, such as the messaging and seasonality, may have affected outcomes in unexpected ways. Our results also highlight the difficulty of interpreting campaign effects when so many channels are used, reinforcing the need to focus on digital and social media channels in future evaluations, including implementing them separately as the different channels may have different campaign effects. It will be especially important to examine the relationship between these channels and behavioral outcomes. Much of the existing literature evaluating digital campaigns focuses on generating engagement (eg, views, likes, and shares), but the question remains as to the value of engagement in achieving campaign outcomes [12,13]. While our results provide some insights into this question, more campaign evaluations need to specifically examine this if we are to understand the role of digital and social media channels in campaigns.

We found that running a variety of creative executions simultaneously was associated with higher recall and recognition and increased the strength of behavioral action, compared to when only 1 campaign was run. To our knowledge, this issue has not been explored before so it is difficult to know the precise mechanism for this association. We speculate that the variety facilitates wider appeal of the campaign and hence higher awareness of the campaigns and a greater likelihood of behavioral impacts. Many campaign evaluations focus on a single creative concept [14-16] so this finding is useful for other jurisdictions as it highlights that there may be value in varying message, tone, and style of the creative, especially as the TARP weightings did not vary significantly between campaign bursts. It should be noted that previous evaluations found that graphic and emotional “why to quit” executions are more effective at generating recall and increased Quitline contacts [17,18], while other evaluations have found that campaigns that evoke different emotions have different effects on quitting thoughts and quit attempts [19,20]. However, our study did not directly compare different creative styles and, as noted above, the television-led campaigns and social media campaigns did not always align. This means that future comprehensive evaluations should consider this question to identify the best approach to generating quitting-related outcomes.

Findings related to Quitline activity may have been subject to potential confounders such as community interventions to increase the use of Quitline (eg, free nicotine replacement therapy and referral pathways with community service and health services) that were implemented concurrently with the MMCP over the evaluation period. The evaluation of the MMCP was limited by the minimal “downtime” between campaign bursts which meant that it was not possible to explore quitting-related outcomes in the absence of campaign activity. Additionally, the mixed recruitment methods due to COVID-19 could have affected survey responses as people with different characteristics may have been more or less likely to respond to the survey when it was offered in 1 mode rather than another, and the way that questions are interpreted and answered may have differed by recruitment method. We were unable to explore what effects, if any, the mixed recruitment methods had on the survey data due to the recruitment method only being recorded for the surveys conducted between July 2020 to June 2021 (ie, half the sample). While other campaign evaluations have observed important differences such as higher unprompted recall for telephone respondents and higher prompted awareness for web-based respondents [21], we are confident in our conclusions because the campaign implementation was not related to the survey administration and thus there is no mechanism whereby the survey administration mode could influence the relationships between the campaign and the outcomes we tested. Further, recognition was only asked of the creative with the highest weighting in terms of campaign spending when multiple creatives ran simultaneously, meaning recognition was likely underestimated for some campaigns. While not ideal for the evaluation, this decision was made by Quit Tasmania to keep the costs of the survey down. The results therefore may underestimate the impact of awareness of the campaign as generated by all channels. The collection of multiple measures for cognitive and behavioral responses to campaigns and comprehensive reporting and examination of multiple campaign channels, however, is a strength of this evaluation.

XSL•FO
RenderX

The evaluation of the MMCP suggests that there is still significant benefit to be derived from running tobacco control mass media campaigns in smaller markets. Future campaigns should continue to make use of a mix of traditional channels and new media for campaigns, although these should be closely monitored and evaluated as the media market is dynamic and consumption habits are changing rapidly.

Acknowledgments

We would like to thank Rory Wilson for his role in overseeing and commenting on the evaluation. We also thank the Tasmanian Quitline Team for providing Quitline activity data. This work was supported by the Department of Health Tasmania and Cancer Council Tasmania.

Data Availability

The data sets analyzed are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to this study design. JK and AG led the analysis. JK led the interpretation and wrote the initial draft. All authors reviewed and edited this paper.

Conflicts of Interest

JK, AG, LC, BHH, MT, and BO were paid by Quit Tasmania to undertake the evaluation.

Multimedia Appendix 1
Tracking survey items used in analyses.
[DOCX File , 31 KB - jmir_v26i1e47128_app1.docx]

References

1. Burden of tobacco use in Australia: Australian Burden of Disease Study 2015. Australian Institute of Health and Welfare. 2019. URL: https://www.aihw.gov.au/reports/burden-of-disease/burden-of-tobacco-use-in-australia/summary [accessed 2024-02-07]
2. Greenhalgh EM, Bayly M, Scollo M. 1.13 Smoking by Australian states and territories. In: Greenhalgh EM, Scollo MM, Winstanley MH, editors. Tobacco in Australia: Facts and Issues. Melbourne: Cancer Council Victoria; 2021.
3. Tasmanian Tobacco Control Plan 2017-21. Department of Health and Human Services, Tasmania. 2017. URL: https://www.health.tas.gov.au/sites/default/files/2021-11/Tobacco_Control_Plan_2017-2021_DoHTasmania.docx [accessed 2024-02-07]
4. Durkin S, Brennan E, Wakefield M. Mass media campaigns to promote smoking cessation among adults: an integrative review. Tob Control 2012;21(2):127-138 [FREE Full text] [doi: 10.1136/tobaccocontrol-2011-050345] [Medline: 22345235]
5. National, state and territory population. Australian Bureau of Statistics. 2022. URL: https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/mar-2022 [accessed 2024-02-07]
6. National drug strategy household survey 2019. Australian Institute of Health and Welfare. Canberra, Australia; 2020. URL: https://www.aihw.gov.au/about-our-data/our-data-collections/national-drug-strategy-household-survey/2019-ndshs [accessed 2024-02-07]
7. Socio-Economic Indexes For Areas (SEIFA) 2016. Australian Bureau of Statistics. Canberra, Australia; 2018. URL: https://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001 [accessed 2024-02-07]
8. Wakefield MA, Spittal MJ, Yong HH, Durkin SJ, Borland R. Effects of mass media campaign exposure intensity and durability on quit attempts in a population-based cohort study. Health Educ Res 2011;26(6):988-997 [FREE Full text] [doi: 10.1093/her/cyr054] [Medline: 21730252]
9. Davis KC, Patel D, Shafer P, Duke J, Glover-Kudon R, Ridgeway W, et al. Association between media doses of the tips from former smokers campaign and cessation behaviors and intentions to quit among cigarette smokers, 2012-2015. Health Educ Behav 2018;45(1):52-60 [FREE Full text] [doi: 10.1177/1090198117709316] [Medline: 28497703]
10. Brennan E, Jeong M, Momjian-Kybert A, Hornik R. Preventing and reducing tobacco use among youth and young adults: a systematic review of the effectiveness of mass media interventions, 2008-2013. University of Pennsylvania. 2016. URL: http://repository.upenn.edu/asc_papers/471 [accessed 2024-02-07]
11. Yellow social media report 2020: consumer statistics. Yellow. 2020. URL: https://www.yellow.com.au/wp-content/uploads/sites/2/2020/07/Yellow-Social-Media-Report-2020-Consumer-Statistics.pdf [accessed 2024-02-07]
12. Chan L, O'Hara B, Phongsavan P, Bauman A, Freeman B. Review of evaluation metrics used in digital and traditional tobacco control campaigns. J Med Internet Res 2020;22(8):e17432 [FREE Full text] [doi: 10.2196/17432] [Medline: 32348272]

XSL•FO
RenderX

13. Kite J, Chan L, MacKay K, Corbett L, Reyes-Marcelino G, Nguyen B, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023;25:e46345 [FREE Full text] [doi: 10.2196/46345] [Medline: 37450325]
14. Duke JC, Farrelly MC, Alexander TN, MacMonegle AJ, Zhao X, Allen JA, et al. Effect of a national tobacco public education campaign on youth's risk perceptions and beliefs about smoking. Am J Health Promot 2018;32(5):1248-1256. [doi: 10.1177/0890117117720745] [Medline: 28759999]
15. Neff LJ, Patel D, Davis K, Ridgeway W, Shafer P, Cox S. Evaluation of the national tips from former smokers campaign: the 2014 longitudinal cohort. Prev Chronic Dis 2016;13:E42 [FREE Full text] [doi: 10.5888/pcd13.150556] [Medline: 27010845]
16. Nagelhout GE, Osman A, Yong HH, Huang LL, Borland R, Thrasher JF. Was the media campaign that supported Australia's new pictorial cigarette warning labels and plain packaging policy associated with more attention to and talking about warning labels? Addict Behav 2015;49:64-67 [FREE Full text] [doi: 10.1016/j.addbeh.2015.05.015] [Medline: 26050643]
17. Nonnemaker JM, Farrelly MC, Kamyab K, MacMonegle AJ. Do different styles of antismoking ads influence the types of smokers who call quitlines? Health Educ Res 2013;28(1):23-30 [FREE Full text] [doi: 10.1093/her/cys080] [Medline: 22843327]
18. Dunlop SM, Perez D, Cotter T. The natural history of antismoking advertising recall: the influence of broadcasting parameters, emotional intensity and executional features. Tob Control 2014;23(3):215-222. [doi: 10.1136/tobaccocontrol-2011-050256] [Medline: 23143868]
19. Durkin S, Bayly M, Brennan E, Biener L, Wakefield M. Fear, sadness and hope: which emotions maximize impact of anti-tobacco mass media advertisements among lower and higher SES groups? J Health Commun 2018;23(5):445-461. [doi: 10.1080/10810730.2018.1463320] [Medline: 29702038]
20. O'Hara BJ, Owen KB, Bauman AE, Dunlop S, Phongsavan P, Furestad E, et al. Hope and sadness: balancing emotions in tobacco control mass media campaigns aimed at smokers. Health Promot J Austr 2023;34(4):856-866 [FREE Full text] [doi: 10.1002/hpja.683] [Medline: 36444612]
21. Hollier LP, Pettigrew S, Slevin T, Strickland M, Minto C. Comparing online and telephone survey results in the context of a skin cancer prevention campaign evaluation. J Public Health (Oxf) 2017;39(1):193-201 [FREE Full text] [doi: 10.1093/pubmed/fdw018] [Medline: 26968482]

Abbreviations

MMCP: mass media campaign program
TARP: target audience rating point

Edited by A Mavragani; submitted 09.03.23; peer-reviewed by S Beasley, L Pederson; comments to author 07.07.23; revised version received 09.08.23; accepted 29.12.23; published 05.03.24.

Please cite as:
Kite J, Grunseit A, Mitchell G, Cooper P, Chan L, Huang BH, Thomas M, O'Hara B, Smith A
Impact of Traditional and New Media on Smoking Intentions and Behaviors: Secondary Analysis of Tasmania's Tobacco Control Mass Media Campaign Program, 2019-2021
J Med Internet Res 2024;26:e47128
URL: https://www.jmir.org/2024/1/e47128
doi:10.2196/47128
PMID:38441941

©James Kite, Anne Grunseit, Glenn Mitchell, Pip Cooper, Lilian Chan, Bo-Huei Huang, Margaret Thomas, Blythe O'Hara, Abby Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Agendas on Nursing in South Korea Media: Natural Language Processing and Network Analysis of News From 2005 to 2022

Daemin Park[1], PhD; Dasom Kim[2], PhD; Ah-hyun Park[3], PhD

[1]School of Media & Communication, Sunmoon University, Chungcheongnam-do, Republic of Korea

[2]Home Visit Healthcare Team, Expert Group on Health Promotion for Seoul Metropolitan Government, Seoul, Republic of Korea

[3]Tobacco Control Team, Expert Group on Health Promotion for Seoul Metropolitan Government, Seoul, Republic of Korea

Corresponding Author:
Dasom Kim, PhD
Home Visit Healthcare Team
Expert Group on Health Promotion for Seoul Metropolitan Government
#410, Life Science Building.Annex, 120, Neungdong-ro, Gwangjin-gu
Seoul, 05029
Republic of Korea
Phone: 82 1072040418
Email: dudurdaram@naver.com

Abstract

Background: In recent years, Korean society has increasingly recognized the importance of nurses in the context of population aging and infectious disease control. However, nurses still face difficulties with regard to policy activities that are aimed at improving the nursing workforce structure and working environment. Media coverage plays an important role in public awareness of a particular issue and can be an important strategy in policy activities.

Objective: This study analyzed data from 18 years of news coverage on nursing-related issues. The focus of this study was to examine the drivers of the social, local, economic, and political agendas that were emphasized in the media by the analysis of main sources and their quotes. This analysis revealed which nursing media agendas were emphasized (eg, social aspects), neglected (eg, policy aspects), and negotiated.

Methods: Descriptive analysis, natural language processing, and semantic network analysis were applied to analyze data collected from 2005 to 2022. BigKinds were used for the collection of data, automatic multi-categorization of news, named entity recognition of news sources, and extraction and topic modeling of quotes. The main news sources were identified by conducting a 1-mode network analysis with SNAnalyzer. The main agendas of nursing-related news coverage were examined through the qualitative analysis of major sources' quotes by section. The common and individual interests of the top-ranked sources were analyzed through a 2-mode network analysis using UCINET.

Results: In total, 128,339 articles from 54 media outlets on nursing-related issues were analyzed. Descriptive analysis showed that nursing-related news was mainly covered in social (99,868/128,339, 77.82%) and local (48,056/128,339, 48.56%) sections, whereas it was rarely covered in economic (9439/128,339, 7.35%) and political (7301/128,339, 5.69%) sections. Furthermore, 445 sources that had made the top 20 list at least once by year and section were analyzed. Other than "nurse," the main sources for each section were "labor union," "local resident," "government," and "Moon Jae-in." "Nursing Bill" emerged as a common interest among nurses and doctors, although the topic did not garner considerable attention from the Ministry of Health and Welfare. Analyzing quotes showed that nurses were portrayed as heroes, laborers, survivors of abuse, and perpetrators. The economic section focused on employment of youth and women in nursing. In the political section, conflicts between nurses and doctors, which may have caused policy confusion, were highlighted. Policy formulation processes were not adequately reported. Media coverage of the enactment of nursing laws tended to relate to confrontations between political parties.

Conclusions: The media plays a crucial role in highlighting various aspects of nursing practice. However, policy formulation processes to solve nursing issues were not adequately reported in South Korea. This study suggests that nurses should secure policy compliance by persuading the public to understand their professional perspectives.

(J Med Internet Res 2024;26:e50518) doi:10.2196/50518

XSL•FO
RenderX

KEYWORDS

nurses; news; South Korea; natural language processing; NLP; network analysis; politicization

Introduction

Controversy Over Nursing Legislation in South Korea

The COVID-19 pandemic has renewed the focus on the importance of nurses in South Korea's health care system with a resultant increase in the interest in improving the nursing workforce structure, working conditions, and treatment [1]. Although the pandemic has fueled a movement to improve nursing policies in South Korea, nurses face difficulties in nursing-related policy activities because of various stakeholders and policy environments. As the nursing profession is regulated by the Medical Service Act enacted in 1951, rather than an independent law such as the Nursing Bill, there are concerns about the inadequacy of the legal status or rights of nurses and, consequently, their diminished role in the health care system.

In South Korea, the debate over nursing legislation has been ongoing for more than a decade. In 2005, Kim Sun-mi, a representative of the ruling Uri Party, first proposed nursing legislation—the Nursing Bill—in the Korean National Assembly; however, the bill was abandoned. In 2018, the bill was reintroduced and rejected and was eventually passed in the National Assembly in April 2023 as the Nursing Bill, shortly after the COVID-19 pandemic ended. However, President Yoon Suk Yeol vetoed the Nursing Bill, and the bill was rejected again in May 2023.

Compared with the early days of the COVID-19 pandemic, it is presently difficult to obtain a favorable public opinion among the public and other professions at the policy level. This may be because of the differences between the general public's interests in nurses' workforce structure, work, and treatment and the need for nursing-related policies [2,3].

Research Context: The Importance of Media in Health Policy

Media coverage has an important impact on shaping public opinion on certain issues. The policy agenda–setting theory suggests that media coverage of an issue can influence the public agenda, which in turn can influence the policy agenda [4,5]. When the media continues to report on an issue, it draws public attention, which can lead to an increased public demand for action from policymakers. Thus, media coverage can be viewed as a form of valuable large-scale data that reflects the flow of social interest and changes in a particular phenomenon or issue. These data can be used to identify paradigm shifts in policies related to the issue.

With the aging population and increasing prevalence of chronic diseases, the medical burden would increase in South Korea. Furthermore, changes in the social environment and policies, such as the introduction of integrated nursing care services, affect the demand for nursing care. In this context, it is important for the public and policymakers to understand nursing care accurately. Media coverage has a significant impact on public perception of medical issues [6,7]. However, research on media coverage in nursing is still limited.

Previous studies focused on identifying the images of nurses in the media or analyzing individual issues [1,8-11]. Both the original Woodhull study conducted in 1997 and a subsequent study conducted in 2017 examined the representation of nurses in health-related news coverage. However, both studies had limitations in that the research involved analysis of news from a single year and focused solely on the health-related context without exploring nurses' representation as news sources in all news sections [12,13].

Methodologically, several studies have used natural language processing (NLP) of news data, topic modeling, and network analysis [8-11,14]. However, most studies analyzed articles over a short period of less than a year [8-11,14]. This strategy has limited the scope of these studies and prevented them from capturing the entire range of issues faced by the nursing community. Additionally, the short timeframe of these studies makes it difficult to capture the long-term development and context of policymaking.

For collecting and NLP of data, we used BigKinds, the largest public news database in South Korea, comprising over 80 million news articles published starting in 1990 [15,16]. It was developed by the Korea Press Foundation, a governmental organization, to support the copyright business of news media. BigKinds is widely used for academic research of news analysis in various fields including nursing studies [9,11]. This study also conducted network analysis and qualitative analysis of quotes.

This study examined how nursing-related issues have changed over time by analyzing, by year and section, approximately 130,000 articles that were published in 54 major Korean media outlets over an 18-year period.

This study focuses on the sources of news articles and their quotes. Sources are key informants whom a reporter has heard directly from and therefore cited in the news for a direct quote [17]. In communication studies, sources are the most important elements of news writing. News is a description of what a source has said [18]. Thus, citing sources is a core factual practice in journalism, along with presenting figures and being present at the scene for direct observation [19].

This study aimed to examine the drivers of the social, local, economic, and political agendas emphasized in the media by the analysis of main sources and their quotes. This analysis revealed which nursing media agendas were emphasized (eg, social aspects), neglected (eg, policy aspects), and negotiated. Furthermore, this study determined the way in which policy formulation processes were not adequately reported.

Research Questions

The research questions (RQs) of this study are as follows:

- RQ1: How many articles were reported by year and section?
- RQ2: Who have been the main sources of media coverage of nurses since 2005, by year and section?

XSL•FO
RenderX

- RQ3: What concerns did the health care community, in particular, nurses, medical doctors, and the Ministry of Health and Welfare (MOHW), pay more attention to and what were their common concerns?
- RQ4: What are the key issues of nursing controversies in the press in quotes?

RQ1 was researched by descriptive analysis; RQ2 was researched by analyzing the news source network; was researched RQ3 by analyzing the news source-topic network; and RQ4 was researched by an analysis of quotes from key sources. This strategy will allow us to see who has led the agenda-setting of nurse-related issues and to identify the agenda that has been emphasized by each section in the media.

Methods

Study Design

Descriptive analysis, NLP, and semantic network analysis were applied to analyze data collected from 2005 to 2022 using BigKinds (Korea Press Foundation), SNAnalyzer, and UCINET. First, BigKinds were used for the collection of data, descriptive analysis, automatic multi-categorization of news, named entity recognition (NER) of news sources, and extraction of quotes from news articles. Second, the main news sources were identified by conducting a 1-mode network analysis with SNAnalyzer. Third, the main agendas of nursing-related news coverage were examined by conducting the qualitative analysis of major sources' quotes by section. Fourth, the common and individual interests of the top-ranked sources were analyzed through a 2-mode network analysis using UCINET.

Data Collection

The collection of news data with NLP was conducted using BigKinds. We used 1 search term "간호사," which means "nurse" in Korean. It is rare that an article about nurses does not mention "nurses"; rather, there may be an article that contains the word "nurse" but indirectly deals with nurses. Articles wherein nurses were peripherally featured were not excluded as they could provide context for nurse coverage.

This study collected data from sources and quotes in nurse articles published in South Korea from January 1, 2005, to December 31, 2022. In 2005, an independent nursing law was proposed for the first time in South Korea by a member of the National Assembly.

The news articles were collected from 54 major domestic media outlets, including national, economic, local, broadcasting, and specialized newspapers, and provided by BigKinds.

Natural Language Processing

BigKinds provides various advanced NLP functions, including automatic article classification, NER, quote extraction, mapping of sources and quotes through semantic analysis, and auto-tagging and ranking of topics in each quote through topic modeling.

The news analyzed was selected from 4 sections—social, local, economic, and political—which are highly related to the policy aspects of nurses and contain numerous articles. The sections are automatically classified by BigKinds. Automatic classification overcomes the differences in naming sections between years and media outlets. Each article can be categorized into a maximum of 3 different sections. For instance, a political article dealing with regional issues can be classified into both political and local sections simultaneously.

Source extraction is accomplished through NER for names, affiliations, and job titles. According to the manual, BigKinds' NER uses an algorithm that combines structured support vector machine and Bidirectional Encoder Representations from Transformers (BERT) with an F_1-score of 91.5%.

We extracted quotes from BigKinds and mapped sources to their quotes through semantic analysis. This study focused on major sources' quotes.

Cleansing

Data cleansing was done in 2 stages. In the first stage, we removed incorrect sources based on character count, part of speech, and so on. Furthermore, we removed anonymous sources and remaining institutional sources. We kept sources who were labeled as a profession or group, such as "nurses" or "hospitals," and who were closely related to the research topics.

The second round of cleansing was conducted after ranking the sources according to the news source network analysis described later. Given the nature of semantic networks, incorrect results in NLP tend to be marginalized by low rankings. Therefore, rather than refining the entire data set from the beginning, it is more efficient to refine only the top-ranked sources who are candidates for the analysis.

In this study, we first performed a second round of cleansing on the top 100 sources. If there was a synonym for the name of an organization due to using abbreviations or anaphoras, we kept the higher-ranked source and unified them with the same name. Through this process, the top 20 news sources by section and year were selected. All the different sources that were tied around the 20th position of the cutoff were included. Finally, 445 sources that had made the top 20 list at least once by year and section were selected. We present the top 20 sources in each section based on cumulative yearly degree centrality, summarized briefly.

Descriptive Analysis

BigKinds provides the number of articles related to searching keywords by year and section. The total number of articles is deduplicated but the sum of the number of articles by section can exceed the total number of articles because each article is categorized into up to 3 sections.

Semantic Network Analysis

In this study, thousands of sources were cited in the articles related to nursing for decades. Thus, semantic network analysis was performed to rank sources by year and section and determine the relationship between sources and topics. Specifically, news source network analysis and news source-topic network analysis were conducted.

The news source network is an undirected 1-mode network with sources as nodes and article cooccurrences as edges. The

XSL•FO
RenderX

importance of each source is evaluated based on the degree centrality of the news source network. A source with a high degree centrality appears in many articles and is discussed in many agendas with many different sources [16]. We also provide descriptive information (size, number of edges, and density) of networks.

A program called SNAnalyzer was used for network analysis [16,20]. SNAnalyzer analyzes multiple Excel files simultaneously with folder-to-folder input and output and provides data cleansing, file name standardization, degree centrality, tie strength, ranking, and descriptive statistics for up to 1,048,575 nodes per file.

NetDraw of UCINET software was used to visualize the news source-topic network. News source-topic network analysis is a 2-mode undirected network with sources and quote topics as nodes and quote cooccurrences as edges. The strength of networks shows the number of times the sources mentioned each topic. News source–topic network analysis can be used to determine which topics are of common interest to sources, and which are of interest only to specific sources. This study focused on 3 groups of sources: nurses, doctors, and the MOHW. Each group was merged according to their affiliations to compare interests at the organizational level.

Quote Analysis

This study also performed quote analysis as a qualitative analysis to deeply understand the context of nursing in news media. In quote analysis, we focused on quotes uttered by the top 445 sources of each year and each section and contained the word "nursing" as a topic in the quotes. Because most top sources were highly cited, analyzing the quotes alone from the top sources resulted in analyzing a substantial number of citations. The number of quotes containing "간호 (nursing)" was 26,926. The total number of deduplicated quotes was 162,316.

Ethical Considerations

Due to the nature of the research involving nonhuman participants, this study was exempt from review by the institutional review board of SM University (202302-001-1). Researchers collected publicly available existing data.

Results

Descriptive Analysis

The total number of articles was 128,339, excluding duplicate articles. The number of articles per section was 99,868 for the social section, 48,056 for the local section, 9439 for the economy section, and 7301 for the political section. A given article was classified as a duplicate in up to 3 sections.

Figure 1 shows the frequency distribution of the articles by year and section. The years 2009, 2015, and 2020 saw surges in news articles corresponding to the swine flu, Middle East respiratory syndrome (MERS), and COVID-19 epidemics, respectively. In 2018, a "burning" incident—a metaphor for bullying in the workplace—occurred among nurses at the Asan Medical Center in Seoul. This incident led to a rapid increase in the number of articles that focused on the nursing culture. These results suggest that the media paid considerable attention to the response to new infectious diseases and accidents in the nursing profession.

Figure 1. Number of articles by year and news section.

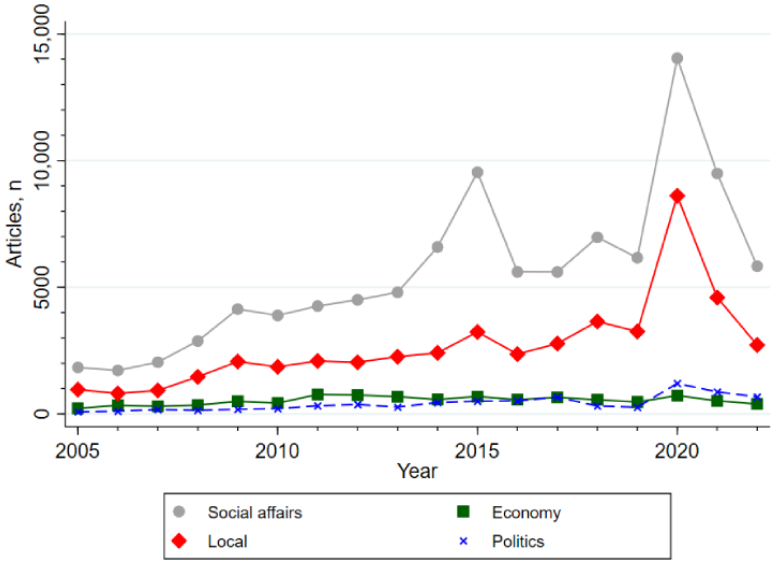

Semantic Network Analysis: News Source Network Analysis

Overview

There were 72 news source networks. We present the size, number of edges, and density of these networks as network-level descriptive statistics in Multimedia Appendix 1.

Multimedia Appendices 2-5 present the top 20 sources by centrality value, which represents the level of connectivity in terms of the years and sections. Each source's rank is determined as the sum of the links. Sources related to the social aspect include incidents and accidents, medical and health issues, education, labor and welfare, trade unions, and activities in civil society. Local governments, such as the Seoul metropolitan government, police, and courts, serve as major sources for the

XSL•FO
RenderX

social departments of media companies. Therefore, nurses and police are recognized as important sources in the social sector.

The main sources representing each section are briefly presented in the following sections with tables in the Multimedia Appendices. The interests of main sources are discussed in more detail in the sections on news source-topic network analysis and quote analysis.

Social Section

"Nurses (the total sum of the degree centrality is 2091)" were the most important sources in the social section of the media, especially in crucial cases such as the 2009 H1N1 influenza epidemic (degree centrality in 2009 is 65), the 2013 Jinju Medical Center incident (sum of degree=74), the 2015 MERS incident (sum of degree=117), and the 2016 nurse workplace bullying and medical malpractice controversy (sum of degree=130). The above-described importance persisted through 2018 (sum of degree=176) and was further emphasized by the COVID-19 pandemic, which began in 2020 (sum of degree=351) in a different context. During the COVID-19 pandemic, governmental sources, such as Central Disaster Management (242 in 2021), came to the forefront.

The period when the "hospital" source (the total sum of the degree centrality is 1790) was considered important coincided with the period when the "nurse" source served as an important source. This was mainly because the hospital was both a nurse's workplace and a medical malpractice site. The importance of the "police" source significantly increased in 2018 (sum of degree=158) when the issue of nurse suicide by bullying and medical malpractice gained social prominence.

Local Section

"Hospital" (577 in total) is the most important source in the local section, especially in 2009 (sum of degree=42), 2014-2015 (sum of degree=34 and 38), and 2017-2021 (sum of degree=54, 36, 39, 103, and 70). During these periods, various impertinent issues were highlighted, including overwork, poor working conditions, and subsequent turnover within the hospital setting. Trade unions have emerged as crucial sources and have thereby contributed to the discourse surrounding the identified issues.

Related to the "labor union" (207 in total), specific incidents that garnered significant attention included the decision to close the Jinju Medical Center in South Gyeongsang Province in 2013 (sum of degree=37), which was criticized for the perceived moral hazard potential. Additionally, the general strike organized by the health care union in 2021 (sum of degree=36) in regard to the expansion of doctor numbers and the adjustment of the patient-to-nurse ratio became a topic of extensive discussion.

The media have played a vital role in disseminating information and shaping public opinion during the COVID-19 pandemic. Of particular interest were the decisions made by the local governments and their respective heads with regard to quarantine measures. For example, during the COVID-19 outbreak in Daegu in early 2020, Daegu's health care and quarantine systems were quickly overwhelmed, and media coverage of the crisis was extensive. "Kwon Young-jin" (135 in 2020), the mayor of Daegu Metropolitan City, appeared in the news, as well as "Gwangju Metropolitan Government" (26 in 2020), which helped Daegu during its time of need. In July 2020, Daegu extended its helping hand to Gwangju when the latter faced a shortage of hospital beds due to the surge in COVID-19 cases. This act of solidarity and collaboration exemplifies the importance of interregional cooperation during public health emergencies.

Furthermore, the enactment of local government ordinances aimed at supporting essential workers, including nurses, in a non–face-to-face environment, as necessitated by the COVID-19 pandemic, is noteworthy. In this context, "Chong Won-o" (76 in total), head of the Seongdong-gu District Office in Seoul, enacted ordinances on safety measures, web-based work-support systems, and psychological counseling for essential workers such as nurses.

Economic and Political Section

The main sources of the economic news were related to free trade agreements (FTA). Unions for youth and women employment and corporate sources were also prominently featured. President "Park Geun-hye" (24 in total), who emphasized economic development by Korean nurses dispatched to Germany in the 1960s and 1970s, played a pivotal role as the primary source of economic development. During the COVID-19 pandemic, local governments became important in the economic sector.

Regarding the political section, the main sources have traditionally been the president and politicians affiliated with major political parties (ie, Ha Tae-keung, 79 in total). In the context of the COVID-19 pandemic, these sources emerged as a major source, as President "Moon Jae-in" (577 in total) engaged in a verbal dispute over a post on Facebook that praised the diligent efforts of nurses.

Semantic Network Analysis: News Source-Topic Network Analysis

A visualization of the news source-topic network for the top 50 topic words by health care source is shown in Figure 2. The analysis revealed 15 common topic words that were prominent among nurses, doctors, and the MOHW. The top common terms encompassed general health-related topics such as "nurses," "patients," "medical staff," and "hospitals." Notably, pandemic-related topics like "COVID-19," "MERS" "confirmed cases," and "infectious diseases" also featured prominently as shared interests.

XSL•FO
RenderX

Figure 2. Topic network analysis by news sources. CCTV: closed-circuit television; MERS: Middle East respiratory syndrome.

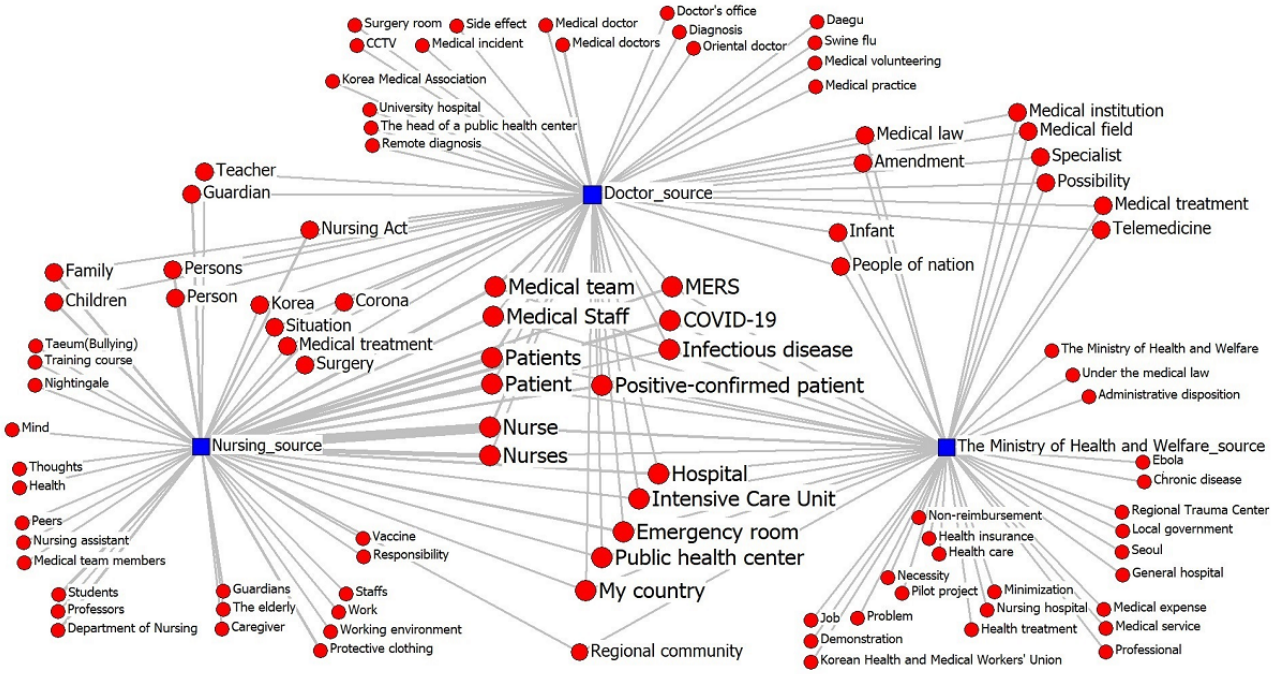

"Nursing Bill" emerged as an issue of high interest among sources from the nursing and doctor groups although the topic did not garner significant attention from MOHW sources. Interestingly, both doctors and MOHW sources exhibited common interests in "the Medical Act" and "amendments," although nurses mentioned the Medical Act less frequently. Moreover, the nurse group and MOHW sources shared a major concern with regard to the "community," indicating a consensus on the importance of expanding public health care in the community.

Nurses demonstrated a distinct interest in education-related themes, such as "students" and "nursing department," as well as concerns regarding poor working conditions, including "protective clothing" and "working environment." Additionally, professional pride was evident through terms like "sense of mission" and "nightingale." Unfortunately, workplace harassment, represented by the term "burning," has emerged as an important issue for nurses.

Conversely, physician sources displayed a particular interest in "operating room" and "closed-circuit television (CCTV)." Doctors voiced opposition to the implementation of closed-circuit television surveillance in operating rooms to prevent medical errors. Furthermore, they expressed concerns about "telemedicine" and "Oriental medicine," fearing potential infringement of doctors' status. Their apprehension lies in the possibility that nondoctor medical practitioners, such as nurses and herbalists, may perform medical treatments and thereby exclude doctors through telemedicine platforms.

Quote Analysis

Social Sections

The timeline and main topics of interest in articles related to nurses are detailed in Multimedia Appendix 6. Nurses have been portrayed in many ways in the social sector, and these portrayals have encompassed various roles that range from heroes to workers to survivors of abuse and even to perpetrators.

Nurses have been portrayed as heroes and dedicated nightingales who have been fighting at the frontlines of epidemic prevention to stop the spread of various infectious diseases—from swine flu and MERS to COVID-19. Such portrayals frequently emphasize courage, compassion, and dedication [1]. Besides their role in epidemic prevention, nurses' heroic images extend beyond this domain and encompass their initial responses to various mass casualty incidents. An illustrative example is the Itaewon tragedy of 2022, wherein the remarkable efforts of nurses were at the forefront. One nurse said, "I think I performed CPR so extensively that I lost count of the number of individuals I attended to" [21].

Beyond this heroic image, other reports have emerged which indicate that nurses face unfavorable working conditions like those of laborers. These conditions are characterized by excessive workloads, inadequate compensation, and lack of recognition of their invaluable contributions. Nurses' voices in the social media coverage extend beyond merely conveying the difficulties associated with nursing work. Furthermore, nurses serve as valuable sources who articulate calls for policy alternatives or provide evaluations of existing policies. This multifaceted portrayal of nurses, which highlights their roles as workers, experts, and heroes, is observed both in the national and international media [22,23].

Furthermore, nurses appeared in the media as suspects or perpetrators of medical accidents, such as the death of a newborn at Ewha Womans University Mokdong Hospital in 2017, and as key sources in cases involving surrogate doctors. However, nurses are not responsible for medical errors and are described as enforcers who function under the supervision of hospitals and doctors.

XSL•FO
RenderX

Nurses have progressed beyond communicating difficulties in the nursing workplace and whistleblowing incidents to initiating calls for policy alternatives and evaluating ongoing policies in media reports. In particular, the Korean Nursing Association emerged as a major source, besides nurses, when discussions of the Nursing Care Act began to be earnestly covered in the media. Simultaneously, the Korean Medical Association strongly opposed the nursing law.

Local Sections

In the local section, there was further emphasis on poor health care and working conditions, which were revealed in the social section. Primary sources associated with "hospitals" have voiced concern with regard to the challenges that are encountered within local health care facilities, particularly in socioeconomically disadvantaged regions. Throughout different infectious disease outbreaks, including the pandemic, MERS, and COVID-19, the focal points of the discussion included the burdensome workloads imposed on nurses, inadequate staffing levels, and insufficient resources, including facilities and equipment.

This situation underscores the need for an expanded role of nurses within the community, greater recognition of the nursing profession, and enhanced public awareness of health care issues. Public and governmental sources, such as the MOHW, nursing organizations, and unions, have advocated the need for increased investment in public health care. The provision of health care services based solely on market principles is limited to regions characterized by smaller populations and relatively lower incomes. Consequently, the government has endeavored to fortify local public health care by emphasizing national university hospitals and public health centers as central hubs.

Various measures have been pursued to achieve the above-described goal. These measures include legislative initiatives for the establishment of community and professional nurse-systems, as well as the expansion and reinforcement of the nursing profession's role through mechanisms such as the nursing grade system and the expansion of nursing schools. Moreover, public support has been provided to bolster these efforts. Although it is acknowledged that challenges, including trial and error, disagreements within the nursing profession, and conflicts with other health care organizations, such as the Korean Medical Association, have emerged during this process, the media generally presents the strengthening of local public health care as being a step toward aligning with the enhanced status of the nursing profession.

In the case of the closure of Jinju Public Medical Center in 2013 and the health care union's call for a general strike in 2021, "unions" appear as the main source of information, calling for the strengthening of public health care. The nursing profession, along with various sources from government departments, such as the MOHW, emphasizes the need to bolster public health care in local media discourse. Media reports portray the government as actively seeking to fortify local public health care with a focus on local university hospitals and public health centers. Specifically, various policies were mentioned, such as the introduction of nurse practitioners, visiting nurses, dispatch nurses, nurses in training, part-time nurses, comprehensive nursing services, a nursing grading system, more nursing schools, and public health nurses. These policies aim to improve the local health care environment by expanding and strengthening the roles of nurses and providing public support.

Economy Sections

Between 2005 and 2007, the opening of the health care market emerged as a significant economic agenda within the context of the Korea-US FTA. Consequently, notable figures such as Kim Jong-hoon, who was the Korea's chief representative in the US-Korea FTA; Kim Hyun-Jong, who served as the head of the Trade Negotiations Division at the Ministry of Foreign Affairs; and Wendy Cutler, who was the US chief representative, became a key source in the economy section of the media.

Unions represent another significant source in the economic sector, particularly in relation to the interests of nurses. Various unions, including major hospital unions, the National Healthcare Workers' Union, the Democratic Trade Union Confederation (to which the National Healthcare Workers' Union is affiliated), and the Public Transport Workers' Union Medical Solidarity Headquarters (affiliated to the Public Transport Workers' Union) have been actively engaged in shaping the discourse on nurses' issues within the broader framework of labor concerns. These unions shed light on key topics, such as insufficient nurse recruitment, the growing trend of casualization through the introduction of part-time work, and substandard working conditions for women employees. The sources from the "union" sector on the economy section of the media provide valuable insights that reflect a broader sense of solidarity with the collective of nurses.

In addition to the heroic portrayal of nurses during the pandemic, the use of nurse imagery in economic development is observed, as exemplified by the depiction of the "Korean nurses in West Germany" as an economic development hero. President Park Geun-hye has associated the concepts of "Korean nurses to West Germany," along with "Korean miners to West Germany" and "Middle Eastern construction workers," with overseas employment initiatives that are aimed at combatting youth unemployment and interlinking nurses within the economic policy direction.

Political Sections

Politicians from major parties and heads of government ministries, including Moon Jae-in, have emerged as key sources. During the COVID-19 pandemic in 2020 and 2021, President Moon found himself amid a controversy. The president is expected to express gratitude to the health care workers, particularly nurses, doctors, and pharmacists.

His emphasis on the nurses' hard work became a point of political contention. A notable example was the 2020 controversy surrounding the division of medical staff at a time when doctors were on strike to protest the government's plans to increase the number of seats in public medical schools and establish new public medical schools. On September 2, 2020, President Moon wrote a Facebook post stating, "Nurses are silently guarding medical sites where doctors, including specialists, have left," and highlighting that the majority of medical staff referred to as "doctors" were actually nurses. These statements drew criticism from members of the opposite

XSL•FO
RenderX

People's Power Party, such as the representative Ha Tae-kyung and spokesperson Kim Eun-hye, who accused President Moon of dividing the medical staff based on their profession. In contrast, the ruling Democratic Party of Korea and representative Ko Goo-jung refuted the division of the medical staff frame and countered that it was the People's Power Party that created a division between the government and medical staff. Furthermore, debates surrounding the authorship of Facebook posts added another layer of controversy to the situation.

In the political section, the news articles featured a range of stakeholders, including the nursing community, who voiced their perspectives on the Nursing Bill. However, rather than providing an in-depth analysis of the policy's specific content, media coverage has primarily emphasized the conflict between nurses and doctors with regard to the enactment of the bill, as well as the debates between President Moon Jae-in and the opposition parties.

The increased media attention on the nursing law enactment began in earnest under President Moon Jae-in's administration after the COVID-19 pandemic, and the controversy surrounding the enactment of the nursing law that represented the ongoing conflict between the Korean Nurses Association and the Korean Medical Association has also received significant attention. The conflict over the Nursing Bill was manifested in various ways, such as the resolution of a general strike by nurses in 2021 and the boycott of the national nursing examination in 2022. The strategies used by nurses in their struggle, which encompassed labor disputes with unions and profession-wide refusal to participate in the national examination, reflect the dual nature of their roles as both workers and professionals.

Following the public's widespread recognition and support for the dedicated efforts of nurses during the COVID-19 pandemic, the Nursing Bill has resurfaced as an important issue that has led both the major presidential candidates in the 2022 election to pledge their commitment toward enacting nursing legislation. However, a divergence in approach has become apparent, with the newly elected People's Power Party expressing caution following internal disagreements. Contrarily, the Democratic Party affirmed its intent to pursue the enactment of the nursing legislation. This eventually resulted in the president vetoing the legislation that had been passed by the opposition-led National Assembly in 2023.

Discussion

Principal Findings

This study focused on analyzing media coverage of the nursing agenda in South Korea over an 18-year period starting in 2005 and examined the coverage patterns, sources responsible for reporting, and quotes in articles, both annually and across different sections. To achieve this objective, a large data set comprising nurse-related articles from BigKinds was used.

In sum, 128,339 articles from 54 media outlets on nursing-related issues were analyzed. The news was mainly covered in social and local sections. According to news source network analysis, 445 main sources who had made the top 20 list at least once by year and section were selected. Multimedia Appendices 2-5 show the top 20 sources by year and section.

The news source-topic network analysis highlights nursing legislation as a common concern for intense conflict between doctors and nurses. The "Nursing Bill" emerged as a common interest among nurses and doctors although the topic did not garner considerable attention from the MOHW.

This study aimed to examine the drivers of the social, local, economic, and political agendas emphasized in the media by the analysis of main sources and their quotes. In the social section, various issues were covered, including the COVID-19 response, workplace bullying, and nursing bills. On the local level, poor working conditions and strengthening public health care were key issues. Considering the economic aspect, labor issues and overseas work were hot topics. On the political aspect, conflicts between nurses and doctors over the nursing bills and the resulting controversy were crucial. This study revealed the media's evolving portrayal of nurses over time. Nurses are portrayed in various roles, such as heroes, workers, survivors of abuse, and even perpetrators, across different topics covered by the media. Nonetheless, in the economic and political spheres, nurses' voices tend to diminish. Media coverage of the enactment of nursing laws tends to move on confrontations between political parties.

Comparison to Prior Work

Aber and Hawkins [24] undertook a comprehensive examination of advertisements featured in medical and nursing journals. Their findings elucidated a prevailing depiction of nurses in print media, where they were predominantly portrayed as ornamental figures, sexualized objects, or ancillaries to doctors. Hoyle et al [25] embarked on a comprehensive exploration of nurses' perceptions regarding the media's influence on the public understanding of the nursing profession. They discerned recurrent themes, notably the media's negative portrayal of the nursing profession—a sentiment echoed in other studies [26]. In contrast, the severe acute respiratory syndrome crisis served as a crucial juncture in the media representation of nurses, akin to the role of the COVID-19 pandemic in South Korea [9,10,27]. These preceding studies have demonstrated the media's paramount influence in shaping societal perceptions of nursing.

Our study underscores a more dynamic evolution of media portrayals pertaining to nurses. During health crises, the media consistently lionized nurses, portraying them as the unsung heroes on the frontlines. However, media portrayals of nurses have undergone fundamental transformations, mirroring the societal evolutions and the multifaceted challenges encountered by the nursing fraternity. With South Korea confronting the challenges of an aging demographic, there is an evident amplification in nurses' community-based responsibilities. This evolution signifies a transition from a traditionally passive representation to a more proactive and central role transcending their conventional subordinate position relative to doctors.

Media coverage can influence not only the image of nurses but also their working conditions, the structure of the health care workforce, and the legislative process [8,26,28]. The paradigm shift in the media agenda has precipitated heightened political

XSL•FO
RenderX

and juridical tensions between the medical and nursing professions. Conflicts between health care professionals, as evidenced by the political dimension analysis, can impede policymaking and result in policy confusion due to politicization. Previous instances of policy confusion within the health care community, including health insurance integration, medical division of labor, insurance finance policy, and reimbursement policy, have been attributed to conflicting interests and backlash from targeted groups [29,30]. Moreover, these factors have been reflected in the media representations.

Limitations

This study provides a comprehensive analysis of media coverage of nurses over an 18-year period since 2005, with a focus on the sources. However, this study has several limitations.

First, the study analyzed only major newspapers and broadcast media that are listed on BigKinds, which means that data from health care journals that were not included in BigKinds could not be analyzed. Future research could compare agenda-related differences between the mass media and health care professional journals to explore how variations in general and public agendas influence policy agendas. Additionally, it would be valuable to examine discussions on platforms, such as Twitter and blogs, to capture a broader range of perspectives.

Second, as this study's analysis was based on all articles containing the keyword "nurses," this strategy limited the ability to descriptively delve into specific agendas. A follow-up study could focus on analyzing a specific agenda, such as the Nursing Bill, over an extended timeframe to gain a more nuanced understanding of policymaking through media channels.

Third, this study implies that health care policy is not solely based on scientific advancements or socioeconomic justifications but is inherently political. Strategic media outreach and policy research are necessary to elucidate policy directions that nursing communities aim to pursue.

Conclusions

This study used a combination of NLP and semantic network analyses to examine how nurse-related issues were covered by the media, specifically focusing on sources and quotes. The media plays a crucial role in highlighting various aspects of nursing practice and nurses, thereby contributing to political engagement and policy activism. However, policy formulation processes to solve nursing issues have not been adequately reported in South Korea.

When engaging with the media to promote nursing policy, it is crucial to navigate the politicization of issues and focus on how policy matters are prioritized on the agenda. As advocates for public health, nurses are responsible for actively engaging in legislative and policymaking processes from a political standpoint [31-33]. However, nurses may not fully realize the extent to which nursing practice relies on public policy decisions and their potential to shape those decisions [34,35]. Nurses should participate in various activities at different levels, such as attending public forums and giving media interviews, to voice their professional views and positions as advocates [36,37]. Nurses need political competence to address the broader determinants of health, effectively intervene in culturally diverse societies, collaborate in developing humane health care systems, and bring nursing values to policy discussions [38].

Acknowledgments

This study was supported by the SunMoon University Research Grant in 2021 (Number 2022-007). The authors used ChatGPT and DeepL for the initial English translation.

Data Availability

Data are not available to share due to Department of Veterans Affairs and ethical retractions; however, annotation instructions and vocabulary are available upon request..

Authors' Contributions

Conceptualization and methodology were developed by DP. Data curation and analysis tasks were performed by DP. DP secured funding for the project. The investigation involved DP, AHP, and DK. Project administration and supervision were handled by DP, AHP, and DK. Resources and software management were under the purview of DP. Validation was conducted by DP, AHP, and DK. Visualization tasks were executed by DP. Writing of the original draft and review and editing were collaborative efforts involving DP, AHP, and DK.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The size, number of edges, and density of 72 news source networks.
[DOCX File , 21 KB - jmir_v26i1e50518_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Top 20 news sources by year in the social section.
[DOCX File , 25 KB - jmir_v26i1e50518_app2.docx]

Multimedia Appendix 3
Top 20 news sources by year in the local section.
[DOCX File , 24 KB - jmir_v26i1e50518_app3.docx]

Multimedia Appendix 4
Top 20 news sources by year in the economy section.
[DOCX File , 23 KB - jmir_v26i1e50518_app4.docx]

Multimedia Appendix 5
Top 20 news sources by year in the political section.
[DOCX File , 25 KB - jmir_v26i1e50518_app5.docx]

Multimedia Appendix 6
The timeline and main agenda in nurse articles.
[DOCX File , 17 KB - jmir_v26i1e50518_app6.docx]

References

1. An J. Analysis of image of nurses on articles in Korean daily newspapers after Corona pandemic. J Convergence Cult Technol 2021;7:66-74 [FREE Full text] [doi: 10.17703/JCCT.2021.7.1.66]
2. Park K, Kwon M. The awareness of nursing students on nursing law legislation in Korea. J Korean Assn Learn Cent Curric Instr 2018;18(11):1-18 doi: 10.22251/jlcci.2018.18.11.1. [doi: 10.22251/jlcci.2018.18.11.1]
3. Lee HJ, Kang HS. The perceptions of nurses and nursing students on nursing law legislation in Korea. J Korean Acad Nurs Adm 2006;12:355-364 eISSN: 2288-4955 [FREE Full text]
4. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Opin Quart 1972;36:176-187. [doi: 10.1086/267990]
5. Iyengar S, Simon A. News coverage of the gulf crisis and public opinion. Commun Res 1993;20(3):365-383. [doi: 10.1177/009365093020003002]
6. Kato M, Yoshimatsu F, Saito T. Trends in media coverage during the monkeypox outbreak: content analysis. J Med Internet Res 2023;25:e45787 [FREE Full text] [doi: 10.2196/45787] [Medline: 37335596]
7. Huang M, ElTayeby O, Zolnoori M, Yao L. Public opinions toward diseases: infodemiological study on news media data. J Med Internet Res 2018;20(5):e10047 [FREE Full text] [doi: 10.2196/10047] [Medline: 29739741]
8. Chin YR, Kwon MH. Contents analysis on the media about the working conditions of nurses. J Korea Soc Comput Inf 2021;26:109-117 [FREE Full text] [doi: 10.9708/jksci.2021.26.02.109]
9. Park MY, Jeong SH, Kim HS, Lee EJ. [Images of nurses appeared in media reports before and after outbreak of COVID-19: text network analysis and topic modeling]. J Korean Acad Nurs 2022 Jun;52(3):291-307. [doi: 10.4040/jkan.22002] [Medline: 35818878]
10. Baek SM, Park M. Analysis of headline news about nurses before and after the COVID-19 pandemic. J Korean Acad Nurs Adm 2022;28(4):319-330 [FREE Full text] [doi: 10.11111/jkana.2022.28.4.319]
11. Yun EK, Kim JO, Byun HM, Lee GG. [Topic modeling and keyword network analysis of news articles related to nurses before and after "the Thanks to You Challenge" during the COVID-19 pandemic]. J Korean Acad Nurs 2021 Aug;51(4):442-453. [doi: 10.4040/jkan.20287] [Medline: 34497253]
12. Mason DJ, Nixon L, Glickstein B, Han S, Westphaln K, Carter L. The Woodhull Study revisited: nurses' representation in health news media 20 years later. J Nurs Scholarsh 2018;50(6):695-704 [FREE Full text] [doi: 10.1111/jnu.12429] [Medline: 30260071]
13. Sigma Theta Tau International. The Woodhull Study on Nursing and The Media: Health Care's Invisible Partner. Indianapolis, IN: Center Nursing Press; 1997.
14. de Melo T, Figueiredo CMS. Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach. JMIR Public Health Surveill 2021;7(2):e24585 [FREE Full text] [doi: 10.2196/24585] [Medline: 33480853]
15. BigKinds. URL: https://www.bigkinds.or.kr/ [accessed 2023-03-05]
16. Park D, Lee H, Jeong SH. Production and correction of misinformation about fine dust in the Korean news media: a big data analysis of news from 2009 to 2019. American Behavioral Scientist 2022 Aug 26:000276422211182. [doi: 10.1177/00027642221118287]
17. Stephens L. Book review: deciding what's news: a study of CBS evening news, NBC nightly news, newsweek and time. Newspaper Res J 2016 Aug 01;2(2):72-73 ISBN 10: 0810122375 [FREE Full text] [doi: 10.1177/073953298100200209]

XSL•FO
RenderX

18. Blanchard RO, Cantor MG, Sigal LV. Reporters and officials: the organization and politics of newsmaking. Contemp Sociol 1976 Jan;5(1):47. [doi: 10.2307/2062646]
19. Van Dijk TA. News as Discourse. New York, NY: Routledge; 2013.
20. SNAnalyzer. URL: https://bit.ly/47esex6 [accessed 2023-03-05]
21. Baek M. Carrying the patient, CPR all night until the knee bleeds... "little heroes save lives". Seoul Shinmun. 2022. URL: https://www.seoul.co.kr/news/newsView.php?id=20221031004006 [accessed 2023-07-05]
22. Mason DJ, Glickstein B, Westphaln K. Journalists' experiences with using nurses as sources in health news stories. Am J Nurs 2018 Oct;118(10):42-50. [doi: 10.1097/01.NAJ.0000546380.66303.a2] [Medline: 30260886]
23. Mohammed S, Peter E, Killackey T, Maciver J. The "nurse as hero" discourse in the COVID-19 pandemic: a poststructural discourse analysis. Int J Nurs Stud 2021 May;117:103887 [FREE Full text] [doi: 10.1016/j.ijnurstu.2021.103887] [Medline: 33556905]
24. Aber CS, Hawkins JW. Portrayal of nurses in advertisements in medical and nursing journals. Image J Nurs Sch 1992;24(4):289-293. [doi: 10.1111/j.1547-5069.1992.tb00735.x] [Medline: 1452183]
25. Hoyle LP, Kyle RG, Mahoney C. Nurses' views on the impact of mass media on the public perception of nursing and nurse–service user interactions. J Res Nurs 2017 Nov 15;22(8):586-596. [doi: 10.1177/1744987117736363]
26. Girvin J, Jackson D, Hutchinson M. Contemporary public perceptions of nursing: a systematic review and narrative synthesis of the international research evidence. J Nurs Manag 2016 Nov;24(8):994-1006. [doi: 10.1111/jonm.12413] [Medline: 27406529]
27. Hall LM, Angus J, Peter E, O'Brien-Pallas L, Wynn F, Donner G. Media portrayal of nurses' perspectives and concerns in the SARS crisis in Toronto. J Nurs Scholarsh 2003;35(3):211-216 [FREE Full text] [doi: 10.1111/j.1547-5069.2003.00211.x] [Medline: 14562487]
28. Gill J, Baker C. The power of mass media and feminism in the evolution of nursing's image: a critical review of the literature and implications for nursing practice. J Med Humanit 2021 Sep;42(3):371-386. [doi: 10.1007/s10912-019-09578-6] [Medline: 31713004]
29. Kim YJ, Park GJ, Moon JW, Lee HYMC. Policy in Nursing and Healthcare. Seoul, South Korea: Gyechuk Munhwasa; 2018.
30. Campos PA, Reich MR. Political analysis for health policy implementation. Health Syst Reform 2019;5(3):224-235 [FREE Full text] [doi: 10.1080/23288604.2019.1625251] [Medline: 31390295]
31. Moon J. Health Policy. 3rd Edition. Seoul, South Korea: Gyechuk Munwhasa; 2013:299-321.
32. Boswell C, Cannon S, Miller J. Nurses' political involvement: responsibility versus privilege. J Prof Nurs 2005;21(1):5-8. [doi: 10.1016/j.profnurs.2004.11.005] [Medline: 15682155]
33. Des Jardin K. Political involvement in nursing--politics, ethics, and strategic action. AORN J 2001 Nov;74(5):614-8, 621. [Medline: 11725441]
34. Salvage J, White J. Nursing leadership and health policy: everybody's business. Int Nurs Rev 2019 Jun;66(2):147-150 [FREE Full text] [doi: 10.1111/inr.12523] [Medline: 31124127]
35. Buck-McFadyen E, MacDonnell J. Contested practice: political activism in nursing and implications for nursing education. Int J Nurs Educ Scholarsh 2017 Jul 27;14(1):20160026. [doi: 10.1515/ijnes-2016-0026] [Medline: 28749781]
36. Wilson DM, Underwood L, Kim S, Olukotun M, Errasti-Ibarrondo B. How and why nurses became involved in politics or political action, and the outcomes or impacts of this involvement. Nurs Outlook 2022;70(1):55-63. [doi: 10.1016/j.outlook.2021.07.008] [Medline: 34493399]
37. Baumann SL, Shaw HK. Improving global health and nursing's international influence. Nurs Sci Q 2022 Jul;35(3):368-373. [doi: 10.1177/08943184221092437] [Medline: 35762058]
38. Suhd-Brondstatter J. From nurse to politician: nurses who left the bedside for public policy [thesis]. University of California, Davis. 2021. URL: https://escholarship.org/uc/item/58b1b3gh [accessed 2023-07-05]

Abbreviations

BERT: bidirectional encoder representations from transformers
FTA: free trade agreements
IRB: institutional review board
MERS: Middle East respiratory syndrome
MOHW: Ministry of Health and Welfare
NER: named entity recognition
NLP: natural language processing
RQ: research question

XSL•FO
RenderX

Edited by T Leung; submitted 03.07.23; peer-reviewed by J Delgado-Ron, C Zhao, S Pesälä; comments to author 19.10.23; revised version received 09.11.23; accepted 21.02.24; published 19.03.24.

Please cite as:
Park D, Kim D, Park AH
Agendas on Nursing in South Korea Media: Natural Language Processing and Network Analysis of News From 2005 to 2022
J Med Internet Res 2024;26:e50518
URL: https://www.jmir.org/2024/1/e50518
doi:10.2196/50518
PMID:38393293

©Daemin Park, Dasom Kim, Ah-hyun Park. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study

Annika Molenaar[1], BNutrSc (Hons); Dickson Lukose[2], PhD; Linda Brennan[3], PhD; Eva L Jenkins[1], BNutrSc (Hons); Tracy A McCaffrey[1], PhD

[1]Department of Nutrition, Dietetics and Food, Monash University, Notting Hill, Australia
[2]Tabcorp Holdings Ltd, Melbourne, Australia
[3]School of Media and Communication, RMIT University, Melbourne, Australia

Corresponding Author:
Tracy A McCaffrey, PhD
Department of Nutrition, Dietetics and Food
Monash University
Level 1
264 Ferntree Gully Road
Notting Hill, 3168
Australia
Phone: 61 3 9905 6862
Email: tracy.mccaffrey@monash.edu

Abstract

Background: Social media has the potential to be of great value in understanding patterns in public health using large-scale analysis approaches (eg, data science and natural language processing [NLP]), 2 of which have been used in public health: sentiment analysis and topic modeling; however, their use in the area of food security and public health nutrition is limited.

Objective: This study aims to explore the potential use of NLP tools to gather insights from real-world social media data on the public health issue of food security.

Methods: A search strategy for obtaining tweets was developed using food security terms. Tweets were collected using the Twitter application programming interface from January 1, 2019, to December 31, 2021, filtered for Australia-based users only. Sentiment analysis of the tweets was performed using the Valence Aware Dictionary and Sentiment Reasoner. Topic modeling exploring the content of tweets was conducted using latent Dirichlet allocation with BigML (BigML, Inc). Sentiment, topic, and engagement (the sum of likes, retweets, quotations, and replies) were compared across years.

Results: In total, 38,070 tweets were collected from 14,880 Twitter users. Overall, the sentiment when discussing food security was positive, although this varied across the 3 years. Positive sentiment remained higher during the COVID-19 lockdown periods in Australia. The topic model contained 10 topics (in order from highest to lowest probability in the data set): “Global production,” “Food insecurity and health,” “Use of food banks,” “Giving to food banks,” “Family poverty,” “Food relief provision,” “Global food insecurity,” “Climate change,” “Australian food insecurity,” and “Human rights.” The topic “Giving to food banks,” which focused on support and donation, had the highest proportion of positive sentiment, and “Global food insecurity,” which covered food insecurity prevalence worldwide, had the highest proportion of negative sentiment. When compared with news, there were some events, such as COVID-19 support payment introduction and bushfires across Australia, that were associated with high periods of positive or negative sentiment. Topics related to food insecurity prevalence, poverty, and food relief in Australia were not consistently more prominent during the COVID-19 pandemic than before the pandemic. Negative tweets received substantially higher engagement across 2019 and 2020. There was no clear relationship between topics that were more likely to be positive or negative and have higher or lower engagement, indicating that the identified topics are discrete issues.

Conclusions: In this study, we demonstrated the potential use of sentiment analysis and topic modeling to explore evolution in conversations on food security using social media data. Future use of NLP in food security requires the context of and interpretation by public health experts and the use of broader data sets, with the potential to track dimensions or events related to food security to inform evidence-based decision-making in this area.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47826) doi:10.2196/47826

KEYWORDS

food security; food insecurity; public health; sentiment analysis; topic modeling; natural language processing; infodemiology

Introduction

Background

Social media has become ubiquitous for people creating and sharing information, news, and experiences in real time, including communicating about issues such as health and nutrition. This engagement on social media creates a vast amount of information that is continually being updated—all day, every day. Deciphering large volumes of information, such as that from social media, can help inform future public health practices based on the current state of affairs, track disease outbreaks, reduce health misinformation, encourage social mobilization by understanding what is important to the public, and highlight future directions in health care [1]. This study used food security as an example of a complex and prevalent public health issue.

Food security can be defined as the availability of and physical, social, and financial access to sufficient, safe, culturally appropriate, and nutritionally adequate food [2,3]. Data science and machine learning techniques (Multimedia Appendix 1 [4-15]) present opportunities to analyze and interpret large-scale public health data to gain an understanding of what is being discussed about food security, in what way, and by whom. Machine learning can classify real-world data such as discussions on social media about food security through statistical models and algorithms built from the analyzed data [16]. One area of data science and machine learning of particular interest in social media analysis is natural language processing (NLP). NLP techniques are able to learn and understand human language [4] and, therefore, can explore the opinions and real-life experiences of social media users through their web-based conversations related to public health issues such as food security [17].

At the public health level, the use of electronic media such as social media for information gathering to understand and inform public health is known as infodemiology [5,18]. One of the goals of infodemiology is to collect and evaluate information on the web (often using data science techniques) that is related to public health, including public communication patterns and behaviors related to a public health issue [5]. Alongside infodemiology is infoveillance, which refers to the use of web-based information for surveillance purposes such as tracking public health events [5]. Infodemiology and infoveillance were key techniques used during the COVID-19 pandemic and vaccination rollout [19]. For example, infodemiology and infoveillance were used to classify and explore misinformation about COVID-19 [20], explore public discourse on COVID-19 and vaccinations [21,22], and track COVID-19 cases and deaths [23]. COVID-19 also highlighted the issue of misinformation and the emergence of an infodemic (Multimedia Appendix 1), with users having access to vast amounts of information, misinformation, and disinformation during the pandemic [24]. Public health professionals, alongside data scientists and behavior change experts, play a role in understanding the theories regarding misinformation and the strategies that can be used to monitor and mitigate the spread of health misinformation, particularly using digital technologies and social media [25].

A commonly used NLP technique to interpret social media data in infodemiology is sentiment analysis (Multimedia Appendix 1), which enables understanding of the discourse on a topic [26]. Sentiment analysis—sometimes referred to as "emotion analysis," "subjectivity analysis," or "opinion mining"—analyzes the opinions, sentiments, attitudes, and emotions embodied within written forms of natural language (eg, social media data) [6]. One review found that sentiment analysis was used in 12 studies in the area of health care to analyze Twitter data with different sentiment analysis tools ranging from open-source publicly available tools to tools produced specifically for the study [27]. Sentiment analysis was also used in 86 studies in the areas of health and well-being. These studies used data from social networking sites and web-based retail platforms and covered a wide range of topics, for example, health conditions, health treatments, mental health, and quality of life [28]. Previous research on social media related to nutrition has largely focused on engagement (eg, likes, shares, and comments) on a small scale (between 9 social media profile pages and 736 social media posts) using manual analysis by topic experts [29-31] and has less frequently explored the breadth of the public's opinions and emotions expressed in social media posts. More recently, sentiment analysis tools, along with additional data science techniques such as topic modeling and social network analysis, were used to explore many nutrition-related topics on social media across 37 studies [32]. Using sentiment analysis alongside other NLP techniques enables researchers to gain a more in-depth understanding of large data sets such as those created in social media, thus providing further insights into potential implications for public health.

Topic modeling is an NLP process that is able to sort textual data (eg, social media data) into different themes or categories of topics using probabilistic algorithms [7,33]. One goal of infodemiology is to explore co-occurrences of different concepts of real-world social media data [5]; this is achieved through topic modeling, which groups text-based data into themes through co-occurrences of words and concepts. Topic modeling can use large data sets to explore relationships between themes of conversation and changes over time through topic evolution [33]. Topic modeling can also be used to track the evolution of discussions across time, taking snapshots of data at different time points and comparing the sentiment, emotion, or topic analyses at each time point. Topic modeling has been used to characterize specific areas of health that social media users commonly discuss on social media platforms [34,35], and recently, topic modeling has been the focus of a great deal of research exploring the discourse on the COVID-19 pandemic

XSL•FO
RenderX

through news and social media sources [36-38]. As a social media analysis tool, topic modeling has the potential to categorize and explore real-time opinions, beliefs, and attitudes in a real-world public health context.

Thus far, infodemiology and infoveillance studies have primarily focused on disease states, outbreaks and epidemics, health care, drugs, smoking, alcohol, and mental health, with less focus on nutrition-related public health issues [26]. Creating methodological processes for gathering information to inform practice or policy has been an urgent focus in research on communicable diseases such as COVID-19. However, the focus on such NLP processes does not exist in areas of complex, multifaceted public health issues such as food security despite its importance to overall health and well-being and the pervasiveness across different population groups.

The concept of food security is underpinned by different dimensions related to access to food and the stability of these dimensions, a population or individual's food access and availability, the ability to use the nutrition from the food [3], agency to influence the food system, and the sustainability of the food from both a social and ecological perspective [39]. The term "food security" refers to when the dimensions have been achieved, and the term "food insecurity" refers to when all these dimensions have not been achieved. The prevalence of food insecurity and subsequent malnutrition worldwide has been increasing [40], with most undernourished people being from low- and middle-income countries in Asia, where 381 million people experience food insecurity, and Africa, where >250 million people experience food insecurity [41]. In high-income countries, the health effects of food insecurity are varied; in adults, they include the development of chronic diseases and obesity [42,43], mental illness, and social isolation [44,45], and in children, they include poor physical and academic development and behavioral issues [46]. Owing to its prominence and the effects it has on nutrition and health, food security is the focus of one of the United Nations Sustainable Development Goals, that is, the goal of ending hunger, achieving food security, improving nutrition, and promoting more sustainable agriculture by 2030 [40].

As with most public health issues, high-income countries are not immune to the effects of food insecurity; for example, it is estimated that 4% to 13% of Australians are affected by food insecurity [47]. Unfortunately, in Australia, there is a lack of a national coordinated response to address food insecurity, with most interventions being at the level of the state or local council area [48]. In addition, the current focus for addressing food insecurity relies on food relief and food banks and, therefore, does not address or seek to further understand the systemic causes of food insecurity [49,50] or the changing issues related to food access arising from events such as the COVID-19 pandemic [51]. This makes information and insight gathering potentially difficult, with widespread and differing reporting on the prevalence of as well as response to the issue of food security. Given that the response to food insecurity in Australia is potentially falling short owing to the overall prevalence and lack of coordinated action, there is a need for new strategies. To gather real-world insights to help support and inform such strategies and decision-making in the area, new data sources, including those of large scale and with real-time updates, should be explored. This has been done in previous research that used artificial intelligence and NLP to use data to predict crop yield and, therefore, assist with cropland mapping to enhance food production and improve access, which is one dimension of food security [52,53]. Other research has highlighted how machine learning can assist in exploring complex socioeconomic parameters related to food security and the interactions among key agents such as climate change, food price dynamics, social networks, and food markets [53].

Objectives

Given that the issue of food security is broad, complex, and multifaceted, this study explored all dimensions of food security with a focus on the Australian context. To be able to capture real-world opinions through social media related to a complex public health issue such as food security, it is crucial to have a deep understanding of the necessary methodological processes. As previous research has not used both the NLP techniques of sentiment and topic analysis on social media data related to food security, the research objective was to gather insights into the potential of these methods in this area. Therefore, the aim of this study was to explore the value of using NLP tools to gather insight from real-world social media data on food security. This study constitutes one of the steps toward using real-world data sources to build infoveillance in public health areas such as food security, with the ultimate aim of enabling evidence-based decision-making for public health professionals. Infoveillance-informed decision-making regarding food security has the potential to create interventions that can keep up with real-time changes in the area and are informed by a broad range of stakeholders, including people experiencing food insecurity.

Methods

Search Term Development

Twitter was chosen as the data collection platform in January 2021 because of its text-based nature, which is suitable for NLP, and the large amounts of publicly available data that can be accessed through the Twitter application programming interface (API; Multimedia Appendix 1). It is acknowledged that Twitter represents only a subset of the population, and therefore, the analysis will not be representative of the broader population of social media users. Nonetheless, the procedures outlined in this paper can be applied to other large data sets from social media.

An iterative process of search term development was used to identify a Twitter search term strategy that retrieved tweets relevant to the topic of food security. Publicly available social media posts related to food security were mined from Twitter using the Twitter API. The initial search terms included words related to food security, food banks, and food relief and relevant hashtags such as #zerohunger, #feedthehungry, and #foodforall, which were determined through manual identification of key tweets in the area of food security. Tweets from users indicating that they were from Australia (ie, location in their Twitter biography) were collected for a week using each iteration of search terms. This data set with a week of tweets containing the search term list was manually assessed for relevance to the topic of food security. The search terms were subsequently refined

XSL•FO
RenderX

based on search terms that produced irrelevant tweets, and additional terms were included to gather more relevant tweets that were not captured using other search terms. Search terms for data collection went through 5 iterations, with manual relevance coding of between 300 and 535 tweets posted from the previous week over 5 different weeks, to ascertain terms that were included in relevant tweets. The following search terms were in the final list included in the API call for data collection: "food security," "food insecurity," "foodbank," "food bank," "food relief," "food insecure," "food secure," "food shelter," #foodsecurity, #foodinsecurity, #foodinsecure, #foodsecure, #foodequity, #zerohunger, #endhunger, #foodforall, #feedthehungry, and #foodbank.

Data Collection

Data from 2019, 2020, and 2021 were collected using the Twitter API from January 1, 2019, to December 31, 2021 (Figure 1). The Twitter Advanced Search API (rather than scraping) was used to extract all publicly available global original tweets, retweets, and tweet replies that contained at least one of the relevant search terms outlined previously. Another inclusion criterion was English-language tweets. Owing to the rate limits of the API, the data were collected in retrospect over a 1-month period. There were 500 tweets collected in each API call with a wait time of 2 seconds between each call. The process consisted of using the search terms in the "Ingest Tweets" function of the engine to collect up to 500 tweets. These were subsequently written into a JSON file. If there was a next token, the engine would pause for 2 seconds before ingesting another set of tweets and writing another JSON file. This process continued until there was no next token. From the Twitter API, the cross-sectional data from these tweets were processed into a data lake in JSON format, which required further processing for final use.

Figure 1. Flow diagram of Twitter data collection, preprocessing, and natural language processing analysis. API: application programming interface; ASCII: American Standard Code for Information Interchange; LDA: latent Dirichlet allocation; NLTK: Natural Language Toolkit; VADER: Valence Aware Dictionary and Sentiment Reasoner.

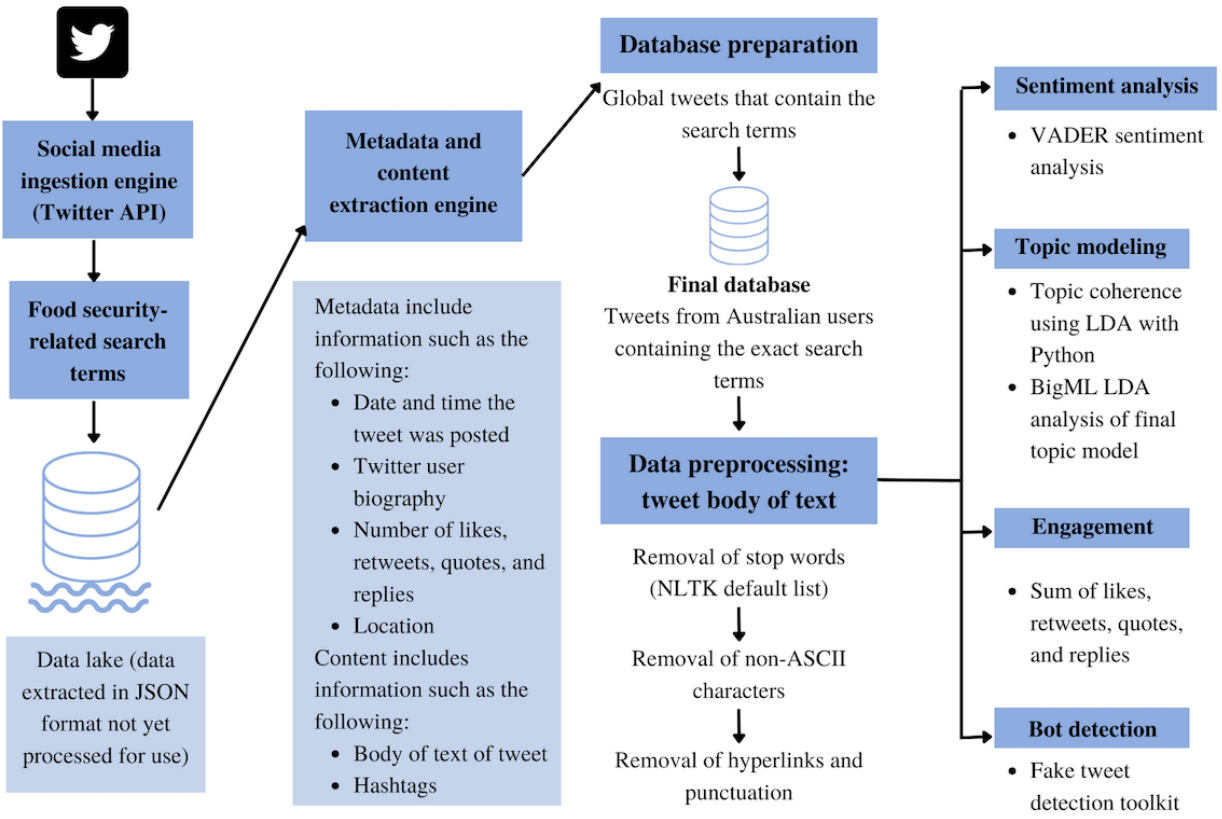

Data Processing

Data cleaning and processing were performed using the Python software (Python Software Foundation) [54]. Data were processed from the JSON format (Multimedia Appendix 1) using a metadata and content extraction engine built by the Monash Data Futures Institute to process metadata, such as the date and time the tweet was posted; engagement data, including the number of likes, retweets, quotes, and replies; location of Twitter users; Twitter user biography and verification status (Multimedia Appendix 1); body of text of the tweet; presence of any media, such as photos, videos, or Graphics Interchange Format images; and source from which the tweet was shared. Although global data were collected, the analysis only used data from Twitter accounts from Australia. That is, tweets sent only by Twitter users who had an Australian state or territory listed on their Twitter biography were used in the final database. As geolocation data were not available for all tweets, the location of Twitter users was determined using a filter that identified Australian location names in the text. These location names and common abbreviations (eg, *SYD* is a common abbreviation for Sydney, New South Wales, Australia) were sourced from the Australian Bureau of Statistics [55]. Using only Australian data allowed the authors to gather a more nuanced picture of the

XSL•FO
RenderX

discussion on food security that was occurring at the time of more specific events and the state of food security and the COVID-19 pandemic specifically in Australia. The final database also consisted of tweets that contained the exact phrase occurring simultaneously for any 2-word or bigram search term (ie, "food security") to increase the relevance and specificity of the included tweets.

Data preprocessing for sentiment analysis and topic modeling included the removal of stop words from the body of the tweet text to leave only potentially meaningful words. The Natural Language Toolkit (Team NLTK) default stop word list was used [56]. The American Standard Code for Information Interchange, which contains 128 characters including the numbers 0 to 9, the English letters A to Z, and some special characters, was used to remove all words that were not American Standard Code for Information Interchange characters [57]. Punctuation and hyperlinks were also removed from the tweet text.

With the emergence of the large-scale use of text-generative models (eg, generative pretrained transformer-based content generators), the proliferation of "fake" (or non–human-generated) social media content is ever increasing. Typical misuses of text-generative models include fake news generation, fake product review generation, and spamming or phishing. Eliminating tweets by nonhuman entities (eg, social bots) is a challenge in itself given that 9% to 15% of Twitter accounts are bot accounts (equivalent to almost 48 million Twitter accounts), and these bots generate almost 35% of Twitter content [58]. To eliminate these fake tweets, we used a fake tweet detection toolkit developed by the Monash Data Futures Institute. This toolkit was applied to the original tweet data before preprocessing.

A measure for "engagement" with the Twitter posts was created using the sum of the number of likes, retweets, quotes, and replies each tweet received. The same tweet could occur multiple times across the data set whenever it was retweeted by an eligible Twitter account (ie, from Australia). These retweets were treated as unique tweets in the data set with their own number of likes, quotes, and replies. The tweet source was refined to create an "other" category including categories that accounted for <1% of instances in the data set. This comprised 142 different platforms, including Instagram, LinkedIn, Facebook, and WordPress.

Data Analysis

Sentiment Analysis

Sentiment analysis was conducted using the Valence Aware Dictionary and Sentiment Reasoner (VADER), which is an open-source linguistic rule and lexicon-based sentiment analysis tool [59]. The tool is based on grammatical and syntactical rules that describe word order–sensitive relationships. This includes degree modifiers that affect the intensity of the sentiment of a sentence. That is, "the service is very good" has a higher positive sentiment than "the service is good" because of the addition of "very" [59]. VADER also uses a lexicon with words assigned to a polarity on a scale of –1 (very negative) to +1 (very positive) based on the average polarity score of the words within the lexicon assigned by 10 independent human raters [59]. This lexicon was specifically designed to analyze social media content, including the sentiment of emojis [59].

The lexicon of the VADER sentiment analyzer (ie, the allocation of a sentiment to each non–stop word) was reviewed to ascertain whether there was agreement between the top 100 most frequent non–stop words in the development data set and their assigned sentiment (ie, very positive, positive, negative, very negative, or neutral). After the data were cleaned and preprocessed, the text of each tweet was processed using the VADER sentiment engine. VADER applies the polarity score to each word present in the tweet text to create 5 outputs. All the positive words create a positive score, the negative words create a negative score, and the neutral words create a neutral score [59]. These 3 scores are summed to create a compound score, which is then normalized between –1 (most negative) and +1 (most positive) [59]. The compound score is then classified into sentiment categories: very negative, negative, neutral, positive, or very positive [59], as outlined in Textbox 1.

Textbox 1. Examples of sentiment categories.

Example tweet text extracts with corresponding sentiment

- Example positive tweet (compound sentiment score=0.76): "A big thanks to all those who donated today to our food bank."
- Example neutral tweet (compound sentiment score=0): "Roughly 25,000 people [in] New Brunswick used food bank services including soup kitchens."
- Example negative tweet (compound sentiment score=–0.67): "Economic growth sub par years high unemployment casualisation record underemployment wages stagnant food insecurity house prices dropping LNP [Liberal National Party] blame next Labor Govt #auspol."

Topic Modeling

The tweet text was used to develop themes or topics through the process of topic modeling using latent Dirichlet allocation (LDA). LDA is a probabilistic algorithm that groups similar text-based data that commonly occur together within a data set into themes [7,33]. As LDA topic modeling is an unsupervised machine learning process, the model created is based on the data themselves and the relationships found within the words present in the textual data. Given its unsupervised nature, LDA topic modeling does not specify the number of topics that are most appropriate or representative of the themes for the data set and, therefore, requires the specification of the number of topics to be used in the model. Consequently, we used a measure known as coherence to help determine the most appropriate number of topics to include. Coherence measures have been shown to correspond well to a human interpretation of the topics from a data set [60]. Topic coherence can provide a score for a

XSL•FO
RenderX

single topic by measuring the degree to which the high-scoring words within that topic are semantically similar to each other [61]. This coherence score is then used to determine whether a set number of topics for that data set are semantically interpretable rather than being related only because of statistical inference [61].

For this study, we used the Gensim (RARE Technologies Ltd) implementation of LDA topic coherence [62], which is an implementation of a 4-stage topic coherence specified by Röder et al [63]. The results of coherence testing revealed that the models with the highest coherence scores were for a topic model with 19 topics (coherence score=0.481) and 10 topics (coherence score=0.478). To determine semantic coherence at the human level, a model including the 10 and 19 topics was created using BigML (BigML, Inc). Manual coherence testing involved visually inspecting the models created through BigML, including the spread of the topics, coherence of the top 10 terms within each topic, and distinction among the different sets of top-10 words within each topic. BigML displays the topics in a topic map (Figure 2) by plotting topics as circles, with the size of the circle representing the topic probability [64]. The position of the topics in the 2D plane of the map is defined by the thematic closeness among the different topics [64]. The topic model with 19 topics (Figure S1 in Multimedia Appendix 2) consisted of topics with a high crossover of key terms and little semantic differentiation. From this manual coherence testing, it was determined that 10 topics (Figure 2) had the greatest semantic coherence.

Figure 2. Topic model distribution of Twitter food security data as visualized on the topic map from BigML (BigML, Inc). The size of the circle represents the probability of that topic being discussed within the data set, and the position of the circles represents the thematic closeness of the topics. There are no axes to consider when interpreting this figure.

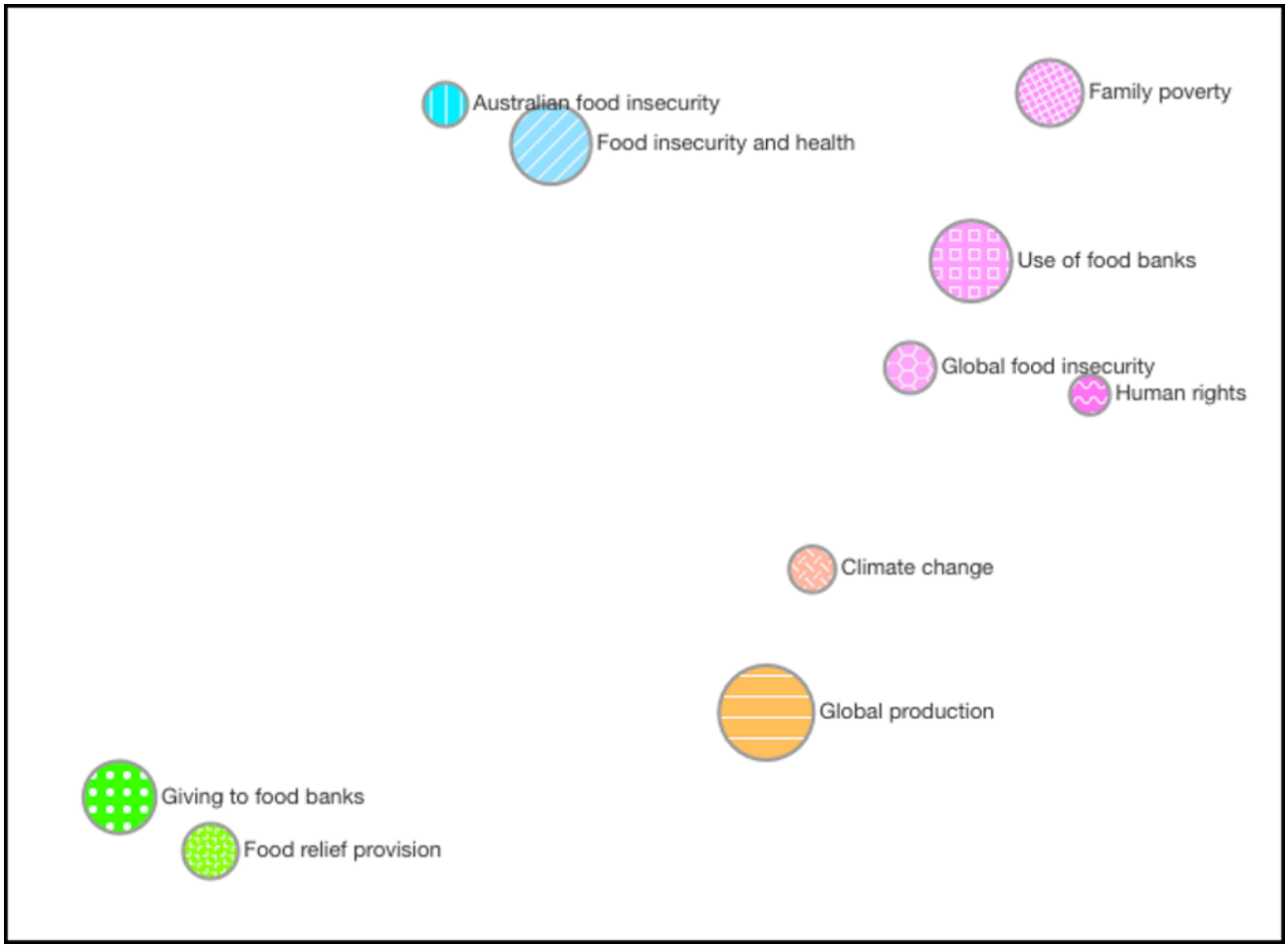

The final topic model was created using the BigML web machine learning platform, which uses an implementation of LDA [64]. The following settings were used for the topic modeling: 10 topics (as determined during coherence testing); 10 top terms per topic; stemming of words, which reduces words to their word stem (eg, the words “agriculture” and “agricultural” would be grouped together and treated as 1 word); and use of bigrams, which allowed for the inclusion of 2-word phrases as 1 term (eg, food security). The batch topic distribution function in BigML was then used to determine, for each tweet, the probability that it discussed each of the 10 topics [64]. Each tweet within the data set had a probability value of that tweet being categorized into each of the 10 topics. The topic with the highest probability was then assigned as the topic for that tweet. For the final topic model, the names were based on the top 10 terms within each topic. In total, 3 authors (AM, LB, and TAM) independently determined names before coming together to triangulate and finalize the most appropriate topic names.

Statistical Testing

Normality testing was conducted using the Kolmogorov-Smirnov test, histograms, and *Q*-*Q* plots using the

XSL•FO
RenderX

Python software SciPy statistics program [65]. The data were found to be not normally distributed, and therefore, median and percentiles and nonparametric tests were used where applicable. The Kruskal-Wallis test was conducted to explore differences in engagement for tweets with different topic and sentiment classifications. Post hoc Dunn tests were used when the Kruskal-Wallis test was significant at $P<.05$. The chi-square test for independence was used to determine the differences between categorical variables.

Visualization and Interpretation

The Python library *matplotlib* was used to visualize the data [66]. This included observing the evolution of sentiment and topics across time, which involved plotting changes in sentiment and topic across the years, months, and quarters. To help with the interpretation of sentiment and topic evolution across time, the COVID-19 pandemic situation in Australia at the time of tweet collection was considered. The COVID-19 pandemic and related lockdowns had substantial effects on food security in Australia and worldwide [67]. Therefore, it is important to consider the effect of the pandemic when assessing tweets related to food security during this period. To explore any associations with COVID-19 lockdowns in Australia, the sentiment and topic evolution graphs were plotted with shading for times when a state or territory of Australia was in a COVID-19 lockdown [68,69]. If that month had more than a week of lockdowns in total across the Australian states and territories, it was shaded in gray. Given that the data included tweets from users from anywhere in Australia, it was decided to shade any lockdowns occurring in Australia even when only 1 state or territory had an active lockdown. The lockdowns differed in their level of restrictions, with some states or territories imposing travel limits (ie, not being permitted to travel >5 km from the place of residence) during some periods. The lockdowns primarily imposed restrictions on retail businesses that people were permitted to access. The businesses to which access was permitted were generally only supermarkets, take-away food restaurants, and pharmacies, and people were also able to undertake other essential travel, such as seeking medical treatment and going to work when it was deemed essential to attend the worksite in person.

In addition, Australia introduced COVID-19 support payments for those who were unable to work or lost their jobs because of the COVID-19 pandemic [70]. The date when this was introduced in March 2020 and when it was removed in March 2021 [70] were also plotted on the sentiment and topic evolution graphs. Alongside these COVID-19 food security–related events in Australia, key news events related to food security and the topics identified through topic modeling were tracked for each month for qualitative discussion within the results [71]. During manual exploration of the tweets, it was evident that the issues discussed included topics broader than the state of food security in Australia alone, which was also highlighted in the topic modeling. Therefore, it was decided to examine news headlines of global as well as Australian-based events related to food security and the topics from topic modeling [71]. This was a broad comparison with news headlines from one source and, therefore, explored associations with food security events and the potential usefulness of examining news headlines in this way but did not determine the specific reasons behind the sentiment and topic trends.

Ethical Considerations

The data for this study were collected through public Twitter profiles, and we adhered to the privacy policies, terms of use, and terms and conditions of Twitter; we aggregated only anonymized data without displaying user identification. Ethics approval for this study was granted by the Monash University Human Research Ethics Committee (approval 27376).

Results

Tweet Descriptives

After filtering the Twitter data by search terms across the 3-year period, there were 38,070 tweets from 14,880 unique Australia-based Twitter users. Of the Twitter users included in our study, most were not verified users (36,302/38,070, 95.36%; Table 1), which indicates fewer organizational accounts or well-known figures discussing the issues. There was a larger proportion of tweets in 2020 than in 2019 and 2021, with 2021 having the second highest number of tweets and 2019 having the lowest number of tweets. Retweets were the most common tweet type, accounting for >70% (28,062/38,070, 73.71%) of the sample and remaining the most common across all years of the study period. The tweets generally consisted of text only (34,230/38,070, 89.91%), which is most appropriate for text-based NLP. This was consistent in 2020 and 2021, but in 2019, there were a higher number of tweets with photos. The fake tweet detection toolkit estimated that a large majority of tweets (34,895/38,070, 91.66%) did not come from suspected bot accounts; however, the proportion of tweets from suspected bot accounts increased across the years from 6.09% (643/10,562) in 2019 to 11.3% (1379/12,206) in 2021.

XSL•FO
RenderX

Table 1. Food security tweet descriptive data over the study period (2019 to 2021; N=38,070).

	Total, n (%)	2019 (n=10,562), n (%)	2020 (n=15,302), n (%)	2021 (n=12,206), n (%)
Tweet type[a]				
Original tweet	4121 (10.82)	1218 (11.53)	1612 (10.53)	1291 (10.58)
Retweet	28,062 (73.71)	7800 (73.85)	11,268 (73.64)	8994 (73.69)
Reply	3918 (10.29)	954 (9.03)	1595 (10.42)	1369 (11.22)
Quote	1969 (5.17)	590 (5.59)	827 (5.4)	552 (4.52)
Tweet format[a]				
Text only	34,230 (89.91)	9199 (87.1)	13,851 (90.52)	11,180 (91.59)
Photo	3747 (9.84)	1336 (12.65)	1416 (9.25)	995 (8.15)
Animated GIF	66 (0.17)	17 (0.16)	33 (0.22)	16 (0.13)
Video	2 (0.01)	10 (0.09)	2 (0.01)	15 (0.1)
Twitter user verification[a]				
Verified	1768 (4.64)	427 (4.04)	822 (5.37)	519 (4.25)
Not verified	36,302 (95.36)	10,135 (95.96)	14,480 (94.63)	11,687 (95.75)
Authenticity[a]				
Suspected bot account	3175 (8.34)	643 (6.09)	1153 (7.53)	1379 (11.3)
Not suspected bot account	34,895 (91.66)	9919 (93.91)	14,149 (92.47)	10,827 (88.7)

[a]Chi-square test of independence significant at $P<.001$ for differences between the years.

Sentiment Analysis and Sentiment Evolution

Overview

The most predominant sentiment among the data set was positive (14,966/38,070, 39.31%), followed by negative sentiment (11,638/38,070, 30.57%; Table S1 in Multimedia Appendix 3). The extreme ends of the sentiment scale, that is, very positive and very negative, were the least common. There was a significant difference ($P<.001$; Table S1 in Multimedia Appendix 3) in the sentiment proportions across the 3-year period, with negative sentiment being slightly more predominant in 2019 and neutral and positive sentiment being slightly more predominant in 2021.

Sentiment Evolution in Relation to News Headlines

When examining sentiment evolution by month (Figure 3), positive sentiment was high in July 2020, which was dominated by events related to COVID-19 in Australia. In addition, positive sentiment was high in June 2021, when the Australian Government announced a new COVID-19 disaster payment to replace the earlier support payments for people without work. This month, the World Food Programme also warned of famine across Africa arising from the COVID-19 pandemic. Negative sentiment was highest in October 2019 and January 2020, a period that was characterized by the Australian bushfire season dubbed “Black Summer” in which many homes were destroyed and families were displaced and required emergency food relief. Another period in which positive sentiment was high was October 2021, which was also when Madagascar faced a food crisis and, in Australia, COVID-19 lockdown restrictions were beginning to ease after the longest run of lockdowns across the country. Very positive sentiment was high in June 2020, which was a month with no COVID-19 lockdowns. From July 2020 to September 2020, positive sentiment was high despite the COVID-19 lockdowns across Australia. During the subsequent lockdowns in 2021, positive sentiment continued to be dominant despite public discontent throughout the lockdown periods. Neutral sentiment rarely had the highest proportion, only being greater than positive or negative sentiment in August 2021, which demonstrates that the conversation was generally polarized across the years.

XSL•FO
RenderX

Figure 3. Proportion of sentiment by month for Twitter food security data. The black lines represent the start and end of COVID-19 support payments by the Australian Government; the gray-shaded sections represent periods when there was a COVID-19 lockdown in an Australian state or territory that lasted >1 week in total within the month.

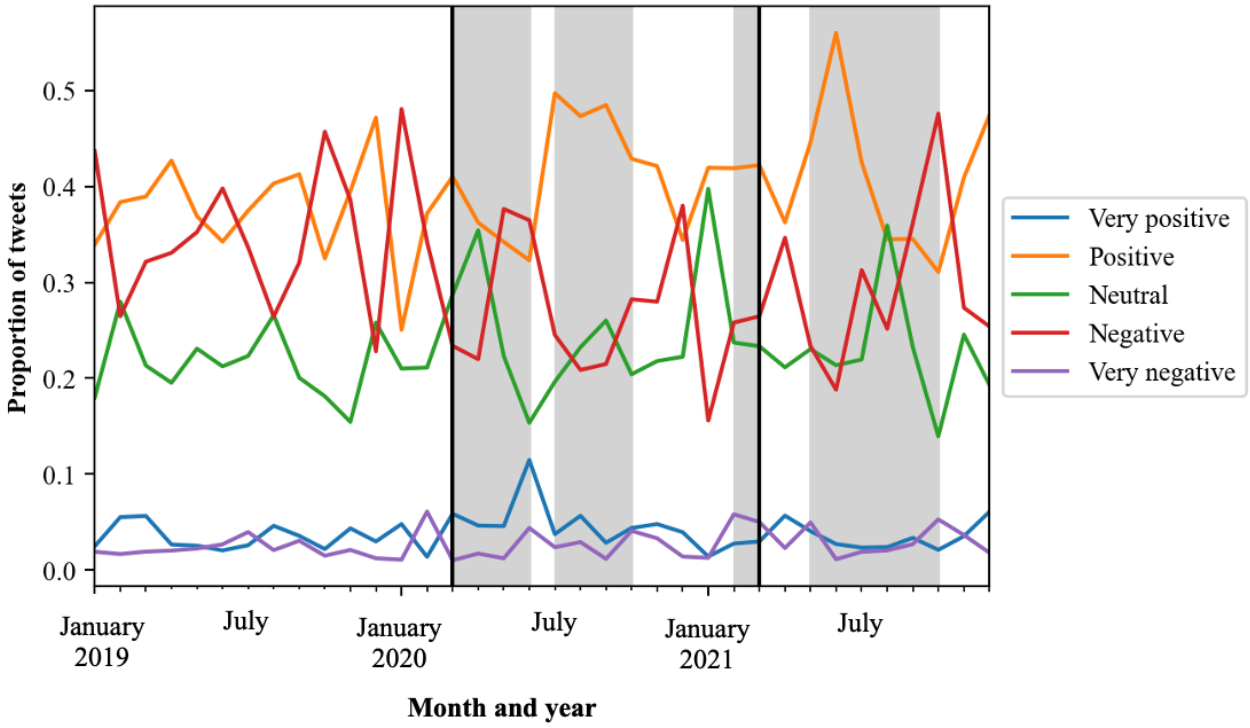

Topic Modeling

The topic model distribution with the highest topic coherence score comprised 10 topics. “Global production” considered food production and agriculture and was the most predominant topic in the data set with the highest probability (Table S2 in Multimedia Appendix 3 and Figure 2). “Global production” was clustered on its own, and its closest relationship was to “Climate change,” which also discussed issues regarding agriculture because of its focus on the environment (Figure 2). “Food insecurity and health” was the second most probable topic within the data set, characterized food insecurity as a public health issue, and was clustered with “Australian food insecurity.” There was a cluster of 4 topics that included “Use of food banks,” “Family poverty,” “Global food insecurity,” and “Human rights.” Finally, a cluster of 2 topics with the furthest distance discussed “Giving to food banks,” including donation and volunteering, and “Food relief provision” at the national level.

Topic Evolution Across Time

Overview

Within the data set, a probability value was calculated for individual tweets across all 10 topics. When examining the topic with the highest probability across the years, most tweets in both 2019 and 2020 discussed “Global production” (2246/10,562, 21.26% and 2810/15,302, 18.36%, respectively; Table S2 in Multimedia Appendix 3 and Figure 4). In 2021, “Food relief provision” accounted for the highest proportion (1762/12,206, 14.44%) of tweet topics, followed by “Use of food banks” (1641/12,206, 13.44%).

When examining the evolution of topic discussion across individual months, “Global production” had the highest probability for 50% (18/36) of the months, and “Food insecurity and health” had the second highest probability for 17% (6/36; Figure 5) of the months. Topic evolution by quarter (Figure S2 in Multimedia Appendix 2 and Table S3 in Multimedia Appendix 3) was not as nuanced, and therefore, the changes by month were used for further exploration.

XSL•FO
RenderX

Figure 4. Topic evolution by year based on the proportion of food security tweets assigned to each topic created through topic modeling.

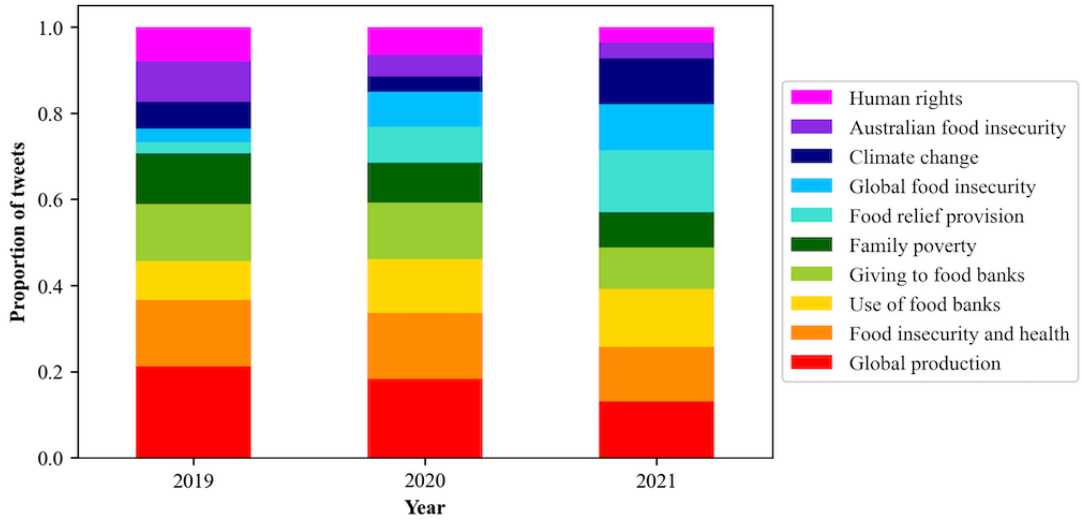

Figure 5. Proportion of highest-probability topics created through topic modeling by month. (A) Topics 1 to 5; (B) Topics 6 to 10. The black lines represent the start and end of COVID-19 support payments by the Australian Government; the gray-shaded sections represent periods in which there was a COVID-19 lockdown in an Australian state or territory that lasted >1 week in total within the month.

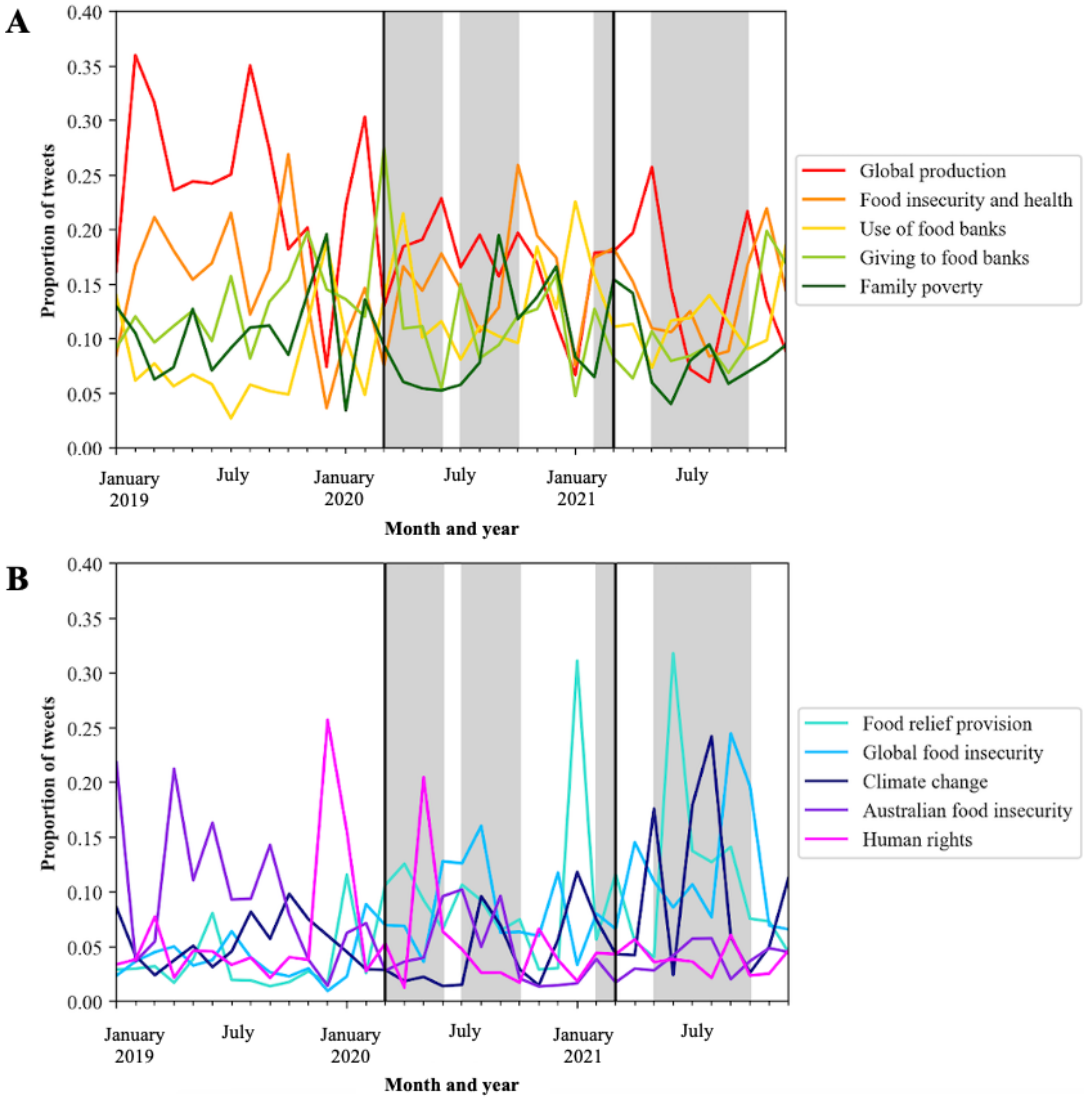

XSL•FO
RenderX

Topic Evolution in Relation to News Headlines

The proportion of predominant topics among the tweets varied across the months (Figure 5). "Global production" remained a largely discussed topic across time, with dominant periods in February 2019, August 2019, and February 2020, with February 2020 including an announcement that Europe had experienced its warmest January on record and the beginning of the rise of COVID-19 outbreaks and deaths across the world. "Food relief provision" also had high periods in January 2021 and June 2021, with a high proportion of positive sentiment also in June 2021. In June 2021, the World Food Programme warned of famine across Africa owing to the COVID-19 pandemic. In addition, the Australian Government announced a new disaster payment that replaced earlier support payments for casual workers who lost work because of the lockdowns and had no support payments during such times.

Other notable high-proportion periods of topics by month included "Giving to food banks" in March 2020, which corresponded to the beginning of the COVID-19 lockdowns across Australia. "Family poverty" was at its highest proportion in December 2019 during the Australian bushfire season (dubbed "Black Summer," in which many homes were destroyed and families were displaced) and in September 2020, when Australia officially entered a recession for the first time since 1991. Despite these events, positive sentiment was high in September 2020 (Figure 3). "Global food insecurity" was at its highest proportion in September 2021, which was when there were news headlines of war in Tigray, Ethiopia, with people facing severe food insecurity, and a United Nations Children's Fund report was released that highlighted that, in 91 countries, most infants experience malnourishment and food insecurity [72]. "Climate change" was at its highest proportion in August 2021 during a heat wave and ongoing wildfires across Europe. The topic of "Human rights" was at its highest proportion in December 2019, with many civilians fleeing Syria during the Syrian civil war, and also in May 2020. Topics specifically related to the prevalence of food insecurity, poverty, and food relief in Australia were not consistently more prominent during the COVID-19 pandemic than during the prepandemic period.

Sentiment and Topic

There was a significant difference between the proportion of sentiment categories for each topic ($P<.001$; Table S4 in Multimedia Appendix 3), as shown in Figure 6.

Figure 6. Proportion of food security tweets in sentiment categories by topic created through topic modeling.

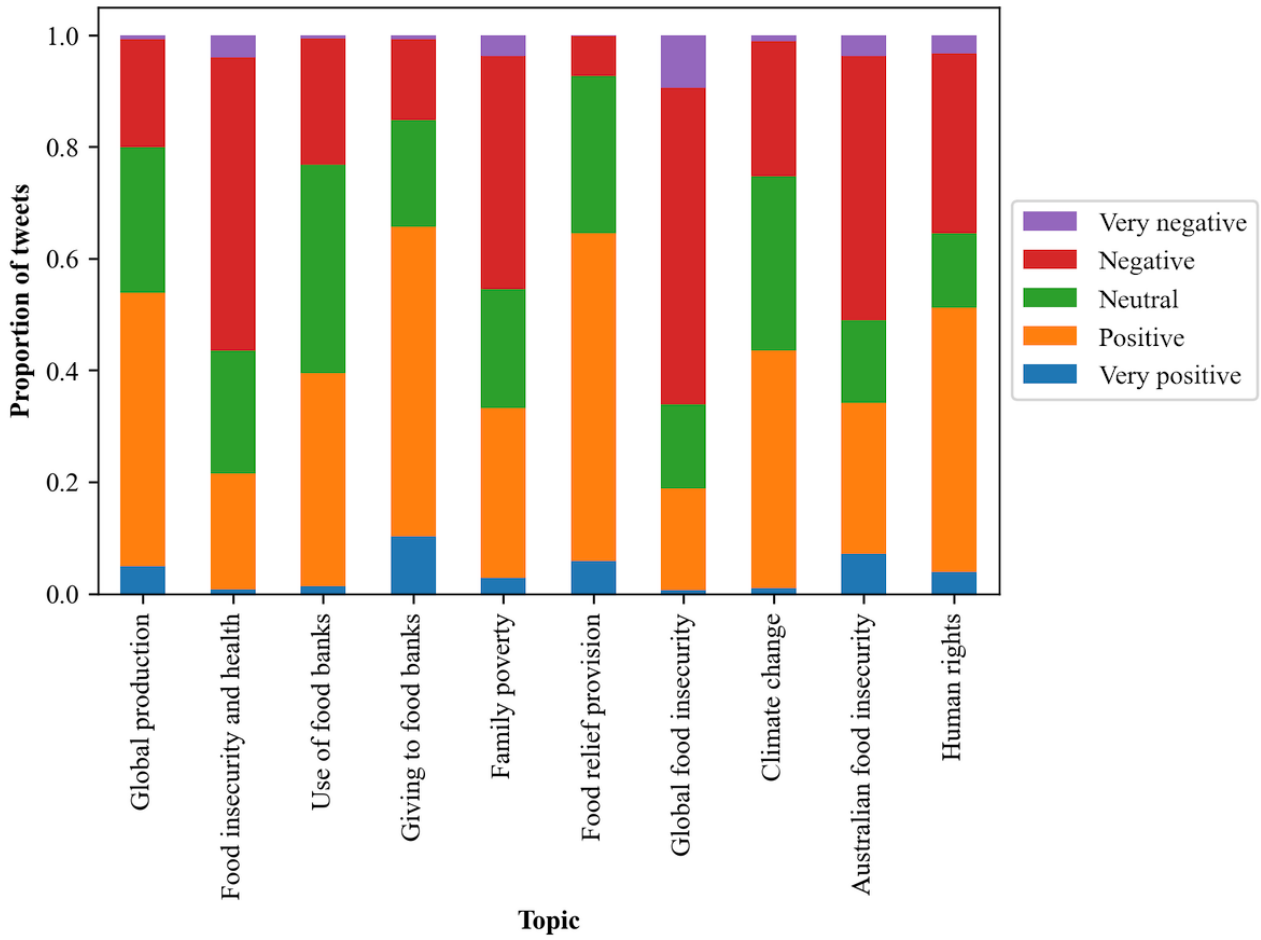

"Global production" of food was discussed with a generally positive sentiment (3262/6656, 49.01%). "Food insecurity and health" and "Family poverty" were more negative in their sentiment (2885/5487, 52.58% and 1530/3670, 41.69%, respectively). "Use of food banks" by individuals had the highest proportion of neutral sentiment (1692/4534, 37.32%) across the topics and was otherwise largely positive (1727/4534, 38.09%). "Giving to food banks," which centered on donations and

XSL•FO
RenderX

charitable support, had the highest proportion of positive sentiment (2526/4561, 55.38% positive and 467/4561, 10.24% very positive). The topic with the highest proportion of positive sentiment (1944/3318, 58.59%) was "Food relief provision." Both "Global food insecurity" and "Australian food insecurity," which discussed food insecurity prevalence, were predominantly negative (1620/2859, 56.66% and 1052/2223, 47.32%, respectively), and "Global food insecurity" had the highest proportion of negative and very negative sentiment (270/2859, 9.44%). "Climate change" in relation to food security was discussed with generally positive (1067/2506, 42.58%) or neutral (780/2506, 31.13%) sentiment. When discussing "Human rights" topics such as access to shelter, food, and water, a positive sentiment was generally expressed (1067/2256, 47.3%). Topics with a predominantly negative sentiment contained more negative key terms such as "insecurity" and "poverty," whereas some predominantly positive topics had key terms with positive sentiment such as "support" and "help" (Table S2 in Multimedia Appendix 3).

Tweet Engagement

The overall engagement score for the tweets was significantly higher in 2021 than in 2019 and 2020 ($P<.001$; Table 2). In the overall sample, very negative and negative tweets received substantially higher engagement (Table 2). When comparing across years, this was also true for 2019 and 2020; however, engagement was significantly higher for neutral tweets in 2021 ($P<.001$). Despite "Global production" being the most frequently discussed topic, engagement was significantly lower than for other topics overall and remained low across the years ($P<.001$). Overall, "Climate change," "Human rights," and "Family poverty" received the most engagement (median 142, 126, and 111, respectively). In 2019, only "Human rights" and "Family poverty" remained the topics that received the highest level of engagement. In 2020, the topic with the highest level of engagement was also "Human rights," whereas in 2021, this topic had a much lower engagement rate (median 381 and 4, respectively). "Climate change" engagement was at its highest level in 2021 (median 200). Topics that received consistently lower engagement across the years were "Global production," "Food insecurity and health," and "Giving to food banks." There was no clear relationship between the predominant sentiment of a topic and engagement with that topic, with 2 of the top 3 topics with the highest engagement having an overall positive sentiment and the third having a negative sentiment.

XSL•FO
RenderX

Table 2. Twitter engagement of food security tweets with different highest-probability topics created through topic modeling and sentiment analysis[a] (superscripted letters denote statistical significance).

	Engagement total, median (IQR)	Engagement in 2019 (n=10,562), median (IQR)	Engagement in 2020 (n=15,302), median (IQR)	Engagement in 2021 (n=12,206), median (IQR)
Engagement (sum of likes, replies, quotes, and retweets)	11 (2-165)	10 (2-141)[b]	11 (2-123)[b]	15 (3-270)[c]
Sentiment				
Very negative	18 (2-205)[d]	20.5 (4-560)[b]	38 (4-126)[b]	10 (1-209)[b,e]
Negative	22 (3-238)[d]	25 (4-244)[b]	28 (4-381)[b]	13.5 (2-128)[b]
Neutral	11 (2-304)[f]	7 (2-86)[c]	9 (2-99)[c]	34 (3-930)[c]
Positive	9 (2-100)[g]	7 (2-102)[c]	7 (2-46)[h]	13 (3-247)[h]
Very positive	6 (2-22)[i]	4 (2-9)[h]	9 (3-94)[c,h]	5 (2-14)[e]
Topics and predominant sentiment				
Global production				
Positive	5 (2-14)[d]	4 (1-10)[b]	6 (2-18)[b]	5 (2-15)[b]
Food insecurity and health				
Negative	7 (2-25)[f]	7 (2-19)[c]	8 (2-34)[c]	6 (2-22)[b]
Use of food banks				
Positive and neutral	37 (3-501)[g]	85 (5-1715)[e,h]	25 (2-361)[e,h]	48 (3-501)[c]
Giving to food banks				
Positive	8 (2-79)[i]	8 (2-58)[j,k]	8 (2-73)[c]	7 (2-131)[h]
Family poverty				
Negative	111 (5-2167)[l]	285 (11-4033)[m]	125 (4-2214)[j,n]	42 (3-1288)[c]
Food relief provision				
Positive	12 (3-171)[o]	4 (1-10)[b,c]	6 (2-18)[b]	51 (5-930)[c]
Global food insecurity				
Negative	13 (2-129)[i,o]	6 (1-21)[b,c,j]	38 (4-238)[h]	6 (1-86)[b]
Climate change				
Positive	142 (14-581)[l]	78 (5-244)[h,n]	25 (3-345)[e,h,j,m]	200 (43-609)[e]
Australian food insecurity				
Negative	19 (3-141)[o]	23 (2-141)[k,n]	18 (3-121)[e,m]	11 (3-247)[h]
Human rights				
Positive	126 (3-1053)[g,l]	1053 (5-1053)[e,m]	381 (3-8509)[n]	4 (1-127)[b,h]

[a]Predominant sentiment refers to the sentiment with the highest proportion for each topic, as shown in Figure 6 and Table S4 in Multimedia Appendix 3. $P<.001$ Kruskal-Wallis test for differences between topic and year and differences between sentiment categories.

[b-o]Values within topic overall, sentiment overall, and topic and sentiment by each year with different superscript letters are significantly different from each other using the post hoc Dunn test and Bonferroni correction.

Discussion

Principal Results

This study used the NLP techniques of sentiment analysis and topic modeling to explore the conversation around food security on Twitter in Australia. The key findings indicate that the overall sentiment of the tweets related to food security was positive, although this varied when assessed by month across the 3-year study period. Positive sentiment remained higher than other sentiment categories during the COVID-19 lockdown periods in Australia. Extremes of sentiment (ie, very negative and very positive) were not common, and neutral sentiment remained lower than both positive and negative sentiment throughout the study period except for August 2021. A topic model with 10 topics related to food security was created based on high topic

XSL•FO
RenderX

coherence. The most predominant topic, "Global production," was related to food production and agriculture, which clustered semantically on its own, with the closest related topic being "Climate change." There were several topics related to food relief and food banks, with different focuses related to public health, volunteering and donation, government support of food banks, and use of food banks by families. When comparing predominant sentiment in the topics, "Giving to food banks," which focused on support and donation to food banks, had the highest proportion of tweets with positive sentiment, and "Global food insecurity," which refers to the prevalence of food insecurity worldwide, had the highest proportion of tweets with negative sentiment. Negative tweets received substantially higher engagement in 2019 and 2020 than in 2021. Despite being the most frequent topic, "Global production" received substantially lower engagement. There was no clear relationship between the predominant sentiment of topics and the engagement rate.

Comparison With Prior Work

Infodemiology involves a specific way of developing knowledge through web-based data sources and can explore discussions and potential influences on health; however, its application in the area of food security as a public health issue is still emerging. Previous studies have used a variety of NLP techniques as part of an infodemiological process to explore different areas of food security for different purposes; however, they have not combined sentiment and topic analysis across time and in relation to social media engagement, as in this study. Studies have used NLP to predict food deserts using the sentiment and nutritional value of the foods mentioned on social media within different geographic regions [73] and to summarize the academic literature on community gardens [74]. Research similar to this study includes a study by Mayasari et al [75] that used Google Trends to explore food security and dietary and lifestyle behaviors during the COVID-19 pandemic. Our study found an increase in the popularity of food security conversations at the beginning of the COVID-19 pandemic, with similar findings related to frequency of discussion to those of Mayasari et al [75], who also found that Australia was among the top countries in search frequency regarding food security. Similarly, Martin et al [76] used topic modeling of tweets related to food security during the COVID-19 pandemic and found an increase in posts on food security, particularly related to food banks and food relief. Martin et al [76] also found topics similar to those in our study, with food assistance, needs, and resources found to be the most dominant topic category, and these topics similarly discussed free food, donation, government assistance, food systems, and food banks. Martin et al [76] also highlighted the topic of emergency preparedness, which covered individual family crises, COVID-19–related food insecurity, and emergency aid in Tigray, all of which were discussed in tweets in our study. The time-series topic analysis by Benites-Lazaro et al [77] highlighted discussions on ethanol production (the topic of the study) and food security dominated by government bodies in comparison with nongovernment and media and peaking from 2007 to 2009, which aligned with the world food crisis. *Government* was a keyword in 2 topics in this study: "Food relief provision" and "Human rights"; however, as Twitter was the data source, there would not likely be the same amount of information coming from government sources as that in the study by Benites-Lazaro et al [77], which used government documents as one of its data sources.

Using a manual analysis approach to infodemiology, research has also examined Facebook posts in community groups in the Pacific Northwest of the United States and found that food assistance and free meals were the most commonly discussed topics in relation to food security [78]. Although conducting only manual analysis, Nguyen et al [78] also found that Facebook posts about community gratitude and those that incorporated culture received more engagement than other posts. These previous studies and our study highlight the potential for computational analysis of infodemiology (eg, sentiment analysis and topic modeling), which allows for the exploration of a broader range of information to gather insights into food security and public health.

Sentiment analysis techniques have been used to achieve various goals in the public health arena of food security. On a broad scale, sentiment analysis has been used to assess the academic literature on food security worldwide [79]. Masih et al [79] found that predictors of positive sentiment included empowerment, farming, and certain government policies and interventions, whereas predictors of negative sentiment included climate change and other government policies and interventions [79]. In Australia, climate change has presented a major challenge and is a predictor of negative sentiment and dissatisfaction with the government's response to climate change and natural disasters [79]. In comparison, our study found that discussions on global production and farming had a positive sentiment and that the topic of climate change overall also had a positive rather than negative sentiment. In addition, our study found that the government was primarily discussed in relation to food relief and food security rather than in relation to climate change or agriculture. A study with a different focus used sentiment analysis to predict sentiment on agriculture by farming communities on Twitter given that crop yields are a measure of food security in different geographical areas [80]. Sentiment analysis has found that, when discussing staple foods in Indonesia as a measure of food security, price volatility and the inability to purchase staple foods at current prices are predictors of negative sentiment [81].

Although this study examined food security more broadly, sentiment analysis has also been used to investigate specific areas of food security. For example, a study assessing the sentiment of web-based conversations specifically related to local food and food banks found that the net sentiment was negative [82]. In contrast, the Twitter posts in this study classified in topics related to food banks were largely positive. This is likely due to the type of discussion in the study by Jung et al [82] being dominated by negative words such as "struggle," "difficult," and "desperate," whereas this study discussed more positive aspects, such as donations, support of food banks, and having access to these services. In addition, Scott et al [83] examined the Supplemental Nutrition Assistance Program (SNAP) in the United States. These researchers found that news articles discussing the SNAP with extreme right media bias were more likely to score on either side of extreme sentiment

XSL•FO
RenderX

[83]. When events such as budget cuts to the SNAP occurred, news articles were more likely to have negative sentiment [83], highlighting the potential of sentiment analysis as an efficient indicator of the state of a topic at different time points or during different periods. However, on its own, sentiment analysis can provide only limited detail on text-based data, and therefore, other data analysis methods such as topic modeling can be used to capture further meaning.

The topic modeling findings of this study can be used in data triangulation with data related to food security from other sources, including sources that are not social media. Previous research has used topic modeling of YouTube and newspaper data and found some consistency between the regions discussing food security and household survey data on the food security risk of these same regions, potentially serving as early warning signals for at-risk areas [84]. Our study found that the topics in food security social media data reflected all the dimensions of food security as defined by the Food and Agriculture Organization [3] and Clapp et al [39] to different extents. The topics most commonly covered the dimensions of access and stability, whereas availability, sustainability, and use of nutrition and agency were less commonly the focus of the themes. Despite the tweets being from Australian users only, there was a range of tweets covering food security issues worldwide from both the public health and economic perspectives. This included food insecurity prevalence, with the topic "Global food insecurity" increasing in number of tweets across the 3-year period and being predominantly negative in sentiment. These findings align with the increase in people experiencing food insecurity worldwide during this time and the associated health issues of undernourishment, particularly during the COVID-19 pandemic [40]. The topic of "Global production" covered worldwide issues related to agriculture and food production and had the greatest probability of being discussed despite having low engagement. This low level of engagement may indicate the distance between the issue being on a global scale and the individual feeling that they are not personally responsible, therefore making them less likely to engage with the topic. Other research findings highlighted issues regarding production, with food imports and exports slowing and sometimes even stopping throughout the pandemic because of factors such as shortages of labor in agriculture and food production and lack of agricultural supplies having major effects on the food supply chain worldwide [85]. Apart from the COVID-19 pandemic, climate change was also a global threat to food security identified in the topic model. Previous research has demonstrated that adverse weather events, including droughts, flooding, and cyclones, have a major impact on agriculture and food production, with reduced crop yields and subsequent instability in food prices [86,87]. However, sentiment analysis of the topic of "Climate change" showed that tweets on this topic were primarily positive, indicating that the discussion on the effects of climate change on food security was framed more positively and potentially more toward opportunities or solutions in this area.

Of the topics covered in our data set of tweets, 3 were clearly related to food security issues specific to Australia, which was evident in the inclusion of Australia-specific key terms (eg, *Australia* and *auspol*). "Food insecurity and health" was largely related to food security in the wake of the COVID-19 pandemic in Australia. This topic aligned with the response to COVID-19 in Australia, which included lockdowns across individual states and territories from March 2020 to October 2021. Although these lockdowns kept case numbers relatively low, they resulted in substantial income loss [88]. To address income loss because of the closure of many industries, the Australian Government introduced a wage subsidy scheme and increase to welfare [88]. Although this is estimated to have reduced total job losses [89], there was an increase in people experiencing food insecurity in Australia up to an estimated 19% to 26% [90,91]. People receiving these payments were up to 3.5 times more likely to experience food insecurity than employed individuals [91]. In this study, the topic "Australian food insecurity" included discussion of a call to action for the Australian Government to address the situation of people experiencing food insecurity marked by a negative sentiment, which may indicate dissatisfaction with the government action or an increased prevalence of experiencing food insecurity and subsequent health issues.

Data collected during the COVID-19 pandemic highlighted an increased reliance on food relief during this period in Australia [92]. This increase was reflected in the topics related to food security in the social media data collected in this study, with 3 different topics specifically related to aspects of food relief and food banks. The predominant response to food insecurity in Australia remains the provision of food relief [48], and this was evident throughout the years of the COVID-19 pandemic [93]. However, in this study, topics related to food insecurity prevalence and food relief within Australia were not consistently more prevalent throughout the pandemic than they were the year before the pandemic, suggesting that these issues were of key interest before the pandemic. In addition, the topic model highlighted key groups that were accessing food relief (eg, students) in the topic "Use of food banks." In Australia, through the demand for emergency food relief during the pandemic, students, particularly international students, were highlighted as a group of people experiencing food insecurity and the related mental health consequences for the first time [93]. This study also highlights key events that may have influenced how the topic area of food security was reported. This includes high proportions of negative sentiment and the topic of "Family poverty" during the bushfires in the Australian summer of 2019 to 2020 (Black Summer). This bushfire season saw many homes destroyed and people displaced as well as loss of livestock and disruption of agricultural land [94]. Although it is not certain that this event influenced the negative sentiment, topic modeling and sentiment analysis can help explore issues that may be driving conversations across time.

Suggestions for Future Research

Future research on food security could focus on furthering infodemiology techniques through interdisciplinary teams that can use NLP techniques in a health-specific context. This study highlighted the value of the information interpretation techniques of topic modeling and sentiment analysis, which can be used to interpret social media data in the areas of food security. The topics highlighted in the data set had some alignment with events and topics of interest in the food security domain identified

XSL•FO
RenderX

through different sources. However, future use of broader data sets including information such as stakeholder opinions could enhance the understanding of associations among sentiment, topic, and news events. The importance of a topic-specific lexicon created by domain experts to gather relevant information was apparent, with other research bodies working toward a comprehensive lexicon in the food security domain [95]. Future research should use the techniques applied in this study as one step within broader infodemiology and infoveillance efforts in public health areas such as food security. Infoveillance could be used to track events that could affect food security at the national or global level, such as climate change, food production, food supply, and government policies related to food security. These techniques could also include using multiple platforms of information—from news articles to different social media platforms and using search terms beyond hashtags—to capture a wider understanding. Infoveillance techniques also have the potential to track trends in the prevalence of food insecurity, thereby enabling the public health sector to improve some of the major effects of food insecurity on health status in a more proactive way by detecting early warning signals [84]. Currently, the prevalence of food insecurity is not completely understood or able to be tracked because of the difficulty in obtaining data and the use of different tools that do not measure all the dimensions of food security [96,97]. This research highlights that, although some topics related to food insecurity are commonly discussed on Twitter, they do not necessarily receive a great deal of engagement from other users. To translate these findings into action, there is a need to engage and connect those talking about these different areas of food security to create collective action.

Limitations

This study has several limitations that should be acknowledged. First, it used Twitter data as the only social media source. Although Twitter is the most commonly used source of information for infodemiology studies [26], there are limitations to its use as the only source. Individuals who use Twitter have been found to be younger than the general population, more likely to be male, and generally more educated and politically attentive than nonusers of social media [98]. Therefore, the users and the information they post are not necessarily representative of the viewpoints of the wider population. A further limitation of using Twitter could be that it has recently experienced a decline in popularity [99]; however, the study collected the data before this decline. Second, the study used VADER sentiment analysis, which lacks the ability to apply context. VADER cannot consider the context of the words and, therefore, cannot consider the potentially different meanings of words within the context of nutrition and food security. Third, in relation to the examination of engagement, tweets that were posted toward the end of data collection in 2021 had less time to receive engagement on Twitter and, therefore, may not have reached their full potential for engagement. However, because of the fast-paced nature of Twitter, most engagement is likely to occur soon after the tweet is posted [100]. Fourth, inferences about the association between news articles covering key events and different sentiment and topic occurrences are not certain. These events are not necessarily the reason for these occurrences as it cannot be determined among the large data set of this study whether these specific events were the most discussed during these periods.

Conclusions

There is potential for the use of NLP techniques to explore social media data to further understand complex areas of public health such as food security. In this study, we demonstrated the value of sentiment analysis and topic modeling in exploring changes in sentiment and key topics discussed in Australia. Topic modeling highlighted the focus on food relief and food banks in the context of Australia and on broader food security themes of global production and supply of food and the effect of climate change on food security. The food security data overall had a slightly more positive sentiment; however, posts with negative sentiment received higher engagement, suggesting the tone of discussion in this topic area that may gather the most attention. Key topics of interest and sentiment evolved throughout the 3-year period, including during the COVID-19 pandemic; however, positive sentiment when discussing food security remained high even throughout lockdowns and subsequent food security crises in Australia. However, because of the discrepancies in associations among sentiment, topic, and news events, there is a need for the use of broader data sets covering more aspects of food security, including different stakeholders. Future use of NLP in food security and public health requires the context of and interpretation by public health experts, with the potential to track dimensions or events related to food security to inform evidence-based decision-making in the public health area of food security.

Acknowledgments

The authors would like to acknowledge and thank the individuals from the Monash Data Futures Institute for their previous work in developing the metadata, content extraction engine, and fake tweet detection toolkit used in this study. AM and ELJ were supported by Australian Government Research Training Program scholarships. The funders had no role in the conduct of the study or the approval of the manuscript for publication.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request. To comply with ethical requirements, the data are not publicly available to ensure anonymity.

XSL•FO
RenderX

Authors' Contributions

AM was involved in the conceptualization of the study, development of the methodology, implementation of the computer code, formal analysis, data curation, visualization, and writing of the original draft of the manuscript. DL was involved in the development of the methodology, implementation of the computer code, data curation, supervision, and review and editing of the manuscript. LB was involved in the conceptualization of the study, development of the methodology, supervision, and review and editing of the manuscript. ELJ was involved in the conceptualization of the study, development of the methodology, and review and editing of the manuscript. TAM was involved in the conceptualization of the study, development of the methodology, supervision, and review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Glossary of terms.
[DOCX File , 16 KB - jmir_v26i1e47826_app1.docx]

Multimedia Appendix 2
Additional figures including additional topic model distribution and proportion of topics by quarter.
[DOCX File , 376 KB - jmir_v26i1e47826_app2.docx]

Multimedia Appendix 3
Additional tables including sentiment of tweets across study period, topic analysis description, topic analysis by quarter and the sentiment of topics.
[DOCX File , 22 KB - jmir_v26i1e47826_app3.docx]

References

1. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
2. Pinstrup-Andersen P. Food security: definition and measurement. Food Sec 2009 Jan 21;1(1):5-7. [doi: 10.1007/s12571-008-0002-y]
3. An introduction to the basic concepts of food security. Food and Agriculture Organization. 2008. URL: http://www.fao.org/3/al936e/al936e.pdf [accessed 2021-06-18]
4. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
5. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
6. Liu B. Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 2012 May 23;5(1):1-167. [doi: 10.2200/s00416ed1v01y201204hlt016]
7. Li S. Topic modeling and Latent Dirichlet Allocation (LDA) in Python. Medium. 2018. URL: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 [accessed 2021-08-31]
8. Application Programming Interface (API). IBM Cloud Education. 2020. URL: https://www.ibm.com/topics/api [accessed 2021-04-13]
9. What is data science? IBM. URL: https://www.ibm.com/au-en/topics/data-science [accessed 2021-04-13]
10. Introducing JSON. JSON. URL: https://www.json.org/json-en.html [accessed 2023-04-03]
11. Guthrie L, Pustejovsky J, Wilks Y, Slator BM. The role of lexicons in natural language processing. Commun ACM 1996 Jan;39(1):63-72. [doi: 10.1145/234173.234204]
12. Machine learning: what it is and why it matters. SAS. 2021. URL: https://www.sas.com/en_au/insights/analytics/machine-learning.html [accessed 2023-02-17]
13. Sloan L, Quan-Haase A. The SAGE Handbook of Social Media Research Methods. Thousand Oaks, CA: Sage Publications; 2017.
14. Stemming and lemmatization. Cambridge University Press. 2008. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html [accessed 2023-02-17]
15. How to get the blue checkmark on Twitter. Twitter. URL: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts [accessed 2023-02-17]
16. What is machine learning? IBM Corp. URL: https://www.ibm.com/au-en/cloud/learn/machine-learning [accessed 2021-08-09]
17. Farzindar A, Inkpen D. Natural Language Processing for Social Media. 2nd Edition. Cham, Switzerland: Springer; 2018.

XSL•FO
RenderX

18. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002 Dec 15;113(9):763-765. [doi: 10.1016/s0002-9343(02)01473-0] [Medline: 12517369]
19. Calleja N, AbdAllah A, Abad N, Ahmed N, Albarracin D, Altieri E, et al. A public health research agenda for managing infodemics: methods and results of the first WHO infodemiology conference. JMIR Infodemiology 2021 Sep 15;1(1):e30979 [FREE Full text] [doi: 10.2196/30979] [Medline: 34604708]
20. Cuan-Baltazar JY, Muñoz-Perez MJ, Robledo-Vega C, Pérez-Zepeda MF, Soto-Vega E. Misinformation of COVID-19 on the internet: infodemiology study. JMIR Public Health Surveill 2020 Apr 09;6(2):e18444 [FREE Full text] [doi: 10.2196/18444] [Medline: 32250960]
21. Rovetta A, Bhagavathula AS. COVID-19-related web search behaviors and infodemic attitudes in Italy: infodemiological study. JMIR Public Health Surveill 2020 May 05;6(2):e19374 [FREE Full text] [doi: 10.2196/19374] [Medline: 32338613]
22. Hou Z, Tong Y, Du F, Lu L, Zhao S, Yu K, et al. Assessing COVID-19 vaccine hesitancy, confidence, and public engagement: a global social listening study. J Med Internet Res 2021 Jun 11;23(6):e27632 [FREE Full text] [doi: 10.2196/27632] [Medline: 34061757]
23. Mavragani A. Tracking COVID-19 in Europe: infodemiology approach. JMIR Public Health Surveill 2020 Apr 20;6(2):e18941 [FREE Full text] [doi: 10.2196/18941] [Medline: 32250957]
24. Tangcharoensathien V, Calleja N, Nguyen T, Purnat T, D'Agostino M, Garcia-Saiso S, et al. Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. J Med Internet Res 2020 Jun 26;22(6):e19659 [FREE Full text] [doi: 10.2196/19659] [Medline: 32558655]
25. Vital BD, Miranda D. Local public health professionals contribution to infodemic management. Eur J Public Health 2021;31(Supplement_3):860. [doi: 10.1093/eurpub/ckab164.860]
26. Mavragani A. Infodemiology and infoveillance: scoping review. J Med Internet Res 2020 Apr 28;22(4):e16206 [FREE Full text] [doi: 10.2196/16206] [Medline: 32310818]
27. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care Tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
28. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 2020 Jan 28;8(1):e16023 [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
29. Ramachandran D, Kite J, Vassallo AJ, Chau JY, Partridge S, Freeman B, et al. Food trends and popular nutrition advice online - implications for public health. Online J Public Health Inform 2018;10(2):e213 [FREE Full text] [doi: 10.5210/ojphi.v10i2.9306] [Medline: 30349631]
30. Barklamb AM, Molenaar A, Brennan L, Evans S, Choong J, Herron E, et al. Learning the language of social media: a comparison of engagement metrics and social media strategies used by food and nutrition-related social media accounts. Nutrients 2020 Sep 16;12(9):2839 [FREE Full text] [doi: 10.3390/nu12092839] [Medline: 32948033]
31. Klassen KM, Borleis ES, Brennan L, Reid M, McCaffrey TA, Lim MS. What people "like": analysis of social media strategies used by food industry brands, lifestyle brands, and health promotion organizations on Facebook and Instagram. J Med Internet Res 2018 Jun 14;20(6):e10227 [FREE Full text] [doi: 10.2196/10227] [Medline: 29903694]
32. Molenaar A, Jenkins EL, Brennan L, Lukose D, McCaffrey TA. The use of sentiment and emotion analysis and data science to assess the language of nutrition-, food- and cooking-related content on social media: a systematic scoping review. Nutr Res Rev 2023 Mar 30:1-36. [doi: 10.1017/S0954422423000069] [Medline: 36991525]
33. Blei DM. Probabilistic topic models. Commun ACM 2012 Apr 01;55(4):77-84. [doi: 10.1145/2133806.2133826]
34. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PLoS One 2014;9(8):e103408 [FREE Full text] [doi: 10.1371/journal.pone.0103408] [Medline: 25084530]
35. Asghari M, Sierra-Sosa D, Elmaghraby A. Trends on health in social media: analysis using twitter topic modeling. In: Proceedings of the 2018 IEEE International Symposium on Signal Processing and Information Technology. 2018 Presented at: ISSPIT '18; December 6-8, 2018; Louisville, KY p. 558-563 URL: https://ieeexplore.ieee.org/document/8642679 [doi: 10.1109/isspit.2018.8642679]
36. Chipidza W, Akbaripourdibazar E, Gwanzura T, Gatto NM. Topic analysis of traditional and social media news coverage of the early COVID-19 pandemic and implications for public health communication. Disaster Med Public Health Prep 2022 Oct 03;16(5):1881-1888 [FREE Full text] [doi: 10.1017/dmp.2021.65] [Medline: 33653437]
37. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
38. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021 Feb 10;23(2):e25431 [FREE Full text] [doi: 10.2196/25431] [Medline: 33497352]
39. Clapp J, Moseley WG, Burlingame B, Termine P. Viewpoint: the case for a six-dimensional food security framework. Food Policy 2022 Jan;106:102164. [doi: 10.1016/j.foodpol.2021.102164]
40. Goal 2: end hunger, achieve food security and improved nutrition and promote sustainable agriculture. United Nations. 2021. URL: https://sdgs.un.org/goals/goal2 [accessed 2021-06-18]
41. Food security and nutrition around the world in 2020. Food and Agriculture Organization. URL: http://www.fao.org/3/ca9692en/online/ca9692en.html#chapter-1_1 [accessed 2021-06-18]

XSL•FO
RenderX

42. Nagata JM, Palar K, Gooding HC, Garber AK, Bibbins-Domingo K, Weiser SD. Food insecurity and chronic disease in US young adults: findings from the national longitudinal study of adolescent to adult health. J Gen Intern Med 2019 Dec 01;34(12):2756-2762 [FREE Full text] [doi: 10.1007/s11606-019-05317-8] [Medline: 31576509]
43. Weaver LJ, Fasel CB. A Systematic Review of the Literature on the Relationships between Chronic Diseases and Food Insecurity. Food Nutr Sci 2018;09(05):519-541 [FREE Full text] [doi: 10.4236/fns.2018.95040]
44. Elgar FJ, Pickett W, Pförtner TK, Gariépy G, Gordon D, Georgiades K, et al. Relative food insecurity, mental health and wellbeing in 160 countries. Soc Sci Med 2021 Jan;268:113556. [doi: 10.1016/j.socscimed.2020.113556] [Medline: 33293171]
45. Martin M, Maddocks E, Chen Y, Gilman S, Colman I. Food insecurity and mental illness: disproportionate impacts in the context of perceived stress and social isolation. Public Health 2016 Mar;132:86-91. [doi: 10.1016/j.puhe.2015.11.014] [Medline: 26795678]
46. Shankar P, Chung R, Frank DA. Association of food insecurity with children's behavioral, emotional, and academic outcomes: a systematic review. J Dev Behav Pediatr 2017;38(2):135-150. [doi: 10.1097/dbp.0000000000000383] [Medline: 28134627]
47. Bowden M. Understanding food insecurity in Australia. Australian Institute of Family Studies. 2020. URL: https://aifs.gov.au/resources/policy-and-practice-papers/understanding-food-insecurity-australia [accessed 2024-01-29]
48. Yii V, Palermo C, Kleve S. Population‐based interventions addressing food insecurity in Australia: a systematic scoping review. Nutr Diet 2019 Oct 08;77(1):6-18. [doi: 10.1111/1747-0080.12580]
49. Rizvi A, Wasfi R, Enns A, Kristjansson E. The impact of novel and traditional food bank approaches on food insecurity: a longitudinal study in Ottawa, Canada. BMC Public Health 2021 Apr 22;21(1):771 [FREE Full text] [doi: 10.1186/s12889-021-10841-6] [Medline: 33882881]
50. Tarasuk V, Dachner N, Hamelin AM, Ostry A, Williams P, Bosckei E, et al. A survey of food bank operations in five Canadian cities. BMC Public Health 2014 Nov 28;14(1):1234 [FREE Full text] [doi: 10.1186/1471-2458-14-1234] [Medline: 25432209]
51. The state of food security and nutrition in the world 2021. Food and Agriculture Organization. 2021. URL: https://www.fao.org/state-of-food-security-nutrition/2021/en/ [accessed 2024-01-29]
52. Shoaib MR, Emara HM, Zhao J. Revolutionizing global food security: empowering resilience through integrated AI foundation models and data-driven solutions. arXiv. Preprint posted online October 21, 2023 2023:1-32 [FREE Full text]
53. Jahan N, Akilan T, Phalke AR. Machine learning for global food security: a concise overview. In: Proceedings of the 2022 IEEE International Humanitarian Technology Conference. 2022 Presented at: IHTC '18; Septemebr 8-11, 2022; Ottawa, ON p. 63-68 URL: https://ieeexplore.ieee.org/document/9998400 [doi: 10.1109/ihtc56573.2022.9998400]
54. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
55. Data by region. Australian Bureau of Statistics. URL: https://dbr.abs.gov.au/ [accessed 2023-11-17]
56. Loper E, Bird S. Nltk: the natural language toolkit. arXiv. Preprint posted online May 17, 2002 2002:1-8 [FREE Full text] [doi: 10.3115/1118108.1118117]
57. ASCII table according to Windows-1252. American Standard Code for Information Interchange. URL: https://www.ascii-code.com/ [accessed 2023-02-07]
58. Varol O, Ferrara E, Davis CA, Menczer F, Flammini A. Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. 2017 May 03 Presented at: ICWSM '17; May 15-18, 2017; Montreal, QC p. 280-289 URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14871/14721 [doi: 10.1609/icwsm.v11i1.14871]
59. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May 16;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]
60. Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: EMNLP '11; July 27-31, 2011; Edinburgh, UK p. 262-272 URL: https://dl.acm.org/doi/10.5555/2145432.2145462
61. Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D. Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012 Presented at: EMNLP-CoNLL '12; July 12-14, 2012; Jeju Island, Korea p. 952-961 URL: https://dl.acm.org/doi/10.5555/2390948.2391052
62. Kapadia S. Evaluate topic model in Python: latent Dirichlet allocation. GitHub. URL: https://github.com/kapadias/medium-articles/blob/master/natural-language-processing/topic-modeling/Evaluate%20Topic%20Models.ipynb [accessed 2022-11-07]
63. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 2015 Presented at: WSDM '15; February 2-6, 2015; Shanghai, China p. 399-408 URL: https://dl.acm.org/doi/10.1145/2684822.2685324 [doi: 10.1145/2684822.2685324]
64. Topic modeling. BigML. 2022. URL: https://bigml.com/features/topic-model [accessed 2022-09-13]
65. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Feb 24;17:261-272 [FREE Full text] [doi: 10.1038/s41592-020-0772-5] [Medline: 32094914]

XSL•FO
RenderX

66. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/mcse.2007.55]
67. Laborde D, Martin W, Swinnen J, Vos R. COVID-19 risks to global food security. Science 2020 Jul 31;369(6503):500-502. [doi: 10.1126/science.abc4765] [Medline: 32732407]
68. Impact of lockdowns on household consumption - insights from alternative data sources. Australian Bureau of Statistics. URL: https://www.abs.gov.au/articles/impact-lockdowns-household-consumption-insights-alternative-data-sources [accessed 2022-11-01]
69. COVID timeline. Australian Government Department of Employment and Workplace Relations. URL: https://www.dese.gov.au/download/13105/ecec-covid-19-timeline/25163/ecec-covid-19-timeline/xlsx [accessed 2022-11-01]
70. JobKeeper payment. Australian Government Treasury. URL: https://treasury.gov.au/coronavirus/jobkeeper#:~:text=From%204%20January%202021%20to,average%20in%20the%20reference%20period [accessed 2022-11-01]
71. Portal:current events. Wikipedia. URL: https://en.wikipedia.org/wiki/Portal:Current_events [accessed 2023-01-30]
72. Fed to Fail: the crisis of children's diets in early life. United Nations Children's Fund. 2021. URL: https://data.unicef.org/resources/fed-to-fail-2021-child-nutrition-report/#:~:text=UNICEF's%20Flagship%20Child%20Nutrition%20Report,safe%20and%20age%20appropriate%20diets [accessed 2024-01-29]
73. Sigalo N, St Jean B, Frias-Martinez V. Using social media to predict food deserts in the united states: infodemiology study of Tweets. JMIR Public Health Surveill 2022 Jul 05;8(7):e34285 [FREE Full text] [doi: 10.2196/34285] [Medline: 35788108]
74. Gregis A, Ghisalberti C, Sciascia S, Sottile F, Peano C. Community garden initiatives addressing health and well-being outcomes: a systematic review of infodemiology aspects, outcomes, and target populations. Int J Environ Res Public Health 2021 Feb 17;18(4):1943 [FREE Full text] [doi: 10.3390/ijerph18041943] [Medline: 33671320]
75. Mayasari NR, Ho DK, Lundy DJ, Skalny AV, Tinkov AA, Teng IC, et al. Impacts of the COVID-19 pandemic on food security and diet-related lifestyle behaviors: an analytical study of google trends-based query volumes. Nutrients 2020 Oct 12;12(10):3103 [FREE Full text] [doi: 10.3390/nu12103103] [Medline: 33053656]
76. Martin NM, Poirier L, Rosenblum AJ, Reznar MM, Gittelsohn J, Barnett DJ. Enhancing artificial intelligence for twitter-based public discourse on food security during the COVID-19 pandemic. Disaster Med Public Health Prep 2022 Aug 04:1-25. [doi: 10.1017/dmp.2022.207] [Medline: 35924366]
77. Benites-Lazaro L, Giatti L, Giarolla A. Topic modeling method for analyzing social actor discourses on climate change, energy and food security. Energy Res Soc Sci 2018 Nov;45:318-330. [doi: 10.1016/j.erss.2018.07.031]
78. Nguyen CJ, Pham C, Jackson AM, Ellison NL, Sinclair K. Online food security discussion before and during the COVID-19 pandemic in native Hawaiian and pacific islander community groups and organizations: content analysis of Facebook posts. Asian Pac Isl Nurs J 2022 Sep 30;6(1):e40436 [FREE Full text] [doi: 10.2196/40436] [Medline: 36212246]
79. Masih J, Sharma A, Patel L, Gade S. Indicators of food security in various economies of world. J Agric Sci 2017 Feb 13;9(3):254. [doi: 10.5539/jas.v9n3p254]
80. Dunnmon J, Ganguli S, Hau D, Husic B. Predicting us state-level agricultural sentiment as a measure of food security with tweets from farming communities. arXiv. Preprint posted online February 13, 2019 2019:1-10 [FREE Full text] [doi: 10.48550/arXiv.1902.07087]
81. Surjandari I, Naffisah MS, Prawiradinata MI. Text mining of Twitter data for public sentiment analysis of staple foods price changes. J ind intell inf 2015;3(3):253-257. [doi: 10.12720/jiii.3.3.253-257]
82. Jung J, Tao J, Widmar NO. Quantifying "local food" online and social media in the United States for 2018-2021. Agric Food Secur 2022 Dec 16;11(59):1-13 [FREE Full text] [doi: 10.1186/s40066-022-00397-y] [Medline: 36540662]
83. Scott D, Oh J, Chappelka M, Walker-Holmes M, DiSalvo C. Food for thought: analyzing public opinion on the supplemental nutrition assistance program. J Technol Hum Serv 2018 Jan 15;36(1):37-47. [doi: 10.1080/15228835.2017.1416514]
84. Ba CT, Choquet C, Interdonato R, Roche M. Explaining food security warning signals with YouTube transcriptions and local news articles. In: Proceedings of the 2022 ACM Conference on Information Technology for Social Good. 2022 Presented at: GoodIT '22; September 7-9, 2022; Limassol, Cyprus p. 315-322 URL: https://dl.acm.org/doi/10.1145/3524458.3547240 [doi: 10.1145/3524458.3547240]
85. Workie E, Mackolil J, Nyika J, Ramadas S. Deciphering the impact of COVID-19 pandemic on food security, agriculture, and livelihoods: a review of the evidence from developing countries. Curr Res Environ Sustain 2020 Dec;2:100014 [FREE Full text] [doi: 10.1016/j.crsust.2020.100014] [Medline: 34977605]
86. Brown ME, Funk CC. Climate. Food security under climate change. Science 2008 Feb 01;319(5863):580-581. [doi: 10.1126/science.1154102] [Medline: 18239116]
87. Wheeler T, von Braun J. Climate change impacts on global food security. Science 2013 Aug 02;341(6145):508-513. [doi: 10.1126/science.1239402]
88. Stobart A, Duckett S. Australia's Response to COVID-19. Health Econ Policy Law 2022 Jan;17(1):95-106 [FREE Full text] [doi: 10.1017/S1744133121000244] [Medline: 34311803]
89. Bishop J, Day I. How many jobs did JobKeeper keep? Reserve bank of Australia. 2020. URL: https://www.rba.gov.au/publications/rdp/2020/2020-07/full.html [accessed 2023-02-23]

XSL•FO
RenderX

90. Kleve S, Bennett CJ, Davidson ZE, Kellow NJ, McCaffrey TA, O'Reilly S, et al. Food insecurity prevalence, severity and determinants in Australian households during the COVID-19 pandemic from the perspective of women. Nutrients 2021 Nov 26;13(12):4262 [FREE Full text] [doi: 10.3390/nu13124262] [Medline: 34959816]
91. Kent K, Murray S, Penrose B, Auckland S, Visentin D, Godrich S, et al. Prevalence and socio-demographic predictors of food insecurity in Australia during the COVID-19 pandemic. Nutrients 2020 Sep 02;12(9):2682 [FREE Full text] [doi: 10.3390/nu12092682] [Medline: 32887422]
92. Foodbank hunger report 2020. Foodbank. 2020. URL: https://www.foodbank.org.au/wp-content/uploads/2020/10/FB-HR20.pdf [accessed 2023-02-23]
93. Louie S, Shi Y, Allman-Farinelli M. The effects of the COVID-19 pandemic on food security in Australia: a scoping review. Nutr Diet 2022 Feb;79(1):28-47 [FREE Full text] [doi: 10.1111/1747-0080.12720] [Medline: 35233911]
94. Filkov AI, Ngo T, Matthews S, Telfer S, Penman TD. Impact of Australia's catastrophic 2019/20 bushfire season on communities and environment. Retrospective analysis and current trends. J Saf Sci Resil 2020 Sep;1(1):44-56. [doi: 10.1016/j.jnlssr.2020.06.009]
95. Roche M, Lindsten A, Lundén T, Helmer T. LEAP4FNSSA lexicon: towards a new dataset of keywords dealing with food security. Data Brief 2022 Dec;45:108680 [FREE Full text] [doi: 10.1016/j.dib.2022.108680] [Medline: 36425989]
96. Ashby S, Kleve S, McKechnie R, Palermo C. Measurement of the dimensions of food insecurity in developed countries: a systematic literature review. Public Health Nutr 2016 May 24;19(16):2887-2896. [doi: 10.1017/s1368980016001166]
97. Kleve S, Gallegos D, Ashby S, Palermo C, McKechnie R. Preliminary validation and piloting of a comprehensive measure of household food security in Australia. Public Health Nutr 2018 Feb;21(3):526-534 [FREE Full text] [doi: 10.1017/S1368980017003007] [Medline: 29125096]
98. Mellon J, Prosser C. Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. Polit Res 2017 Jul 13;4(3):205316801772000. [doi: 10.1177/2053168017720008]
99. Dang S. Exclusive: Twitter is losing its most active users, internal documents show. Reuters. URL: https://www.reuters.com/technology/exclusive-where-did-tweeters-go-twitter-is-losing-its-most-active-users-internal-2022-10-25/ [accessed 2023-02-23]
100. Bae Y, Ryu P, Kim H. Predicting the lifespan and retweet times of tweets based on multiple feature analysis. ETRI J 2014 Jun 01;36(3):418-428. [doi: 10.4218/etrij.14.0113.0657]

Abbreviations

API: application programming interface
LDA: latent Dirichlet allocation
NLP: natural language processing
SNAP: Supplemental Nutrition Assistance Program
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by A Mavragani; submitted 13.04.23; peer-reviewed by M Chatzimina, N Yiannakoulias; comments to author 27.08.23; revised version received 05.12.23; accepted 20.12.23; published 21.03.24.

Please cite as:
Molenaar A, Lukose D, Brennan L, Jenkins EL, McCaffrey TA
Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study
J Med Internet Res 2024;26:e47826
URL: https://www.jmir.org/2024/1/e47826
doi:10.2196/47826
PMID:38512326

©Annika Molenaar, Dickson Lukose, Linda Brennan, Eva L Jenkins, Tracy A McCaffrey. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Bridging and Bonding Social Capital by Analyzing the Demographics, User Activities, and Social Network Dynamics of Sexual Assault Centers on Twitter: Mixed Methods Study

Jia Xue[1,2], PhD; Qiaoru Zhang[3*], BA; Yun Zhang[3*], MI; Hong Shi[3*], MI; Chengda Zheng[3*], MI; Jingchuan Fan[1], MSW, MPH; Linxiao Zhang[1], MA, MSc; Chen Chen[3], PhD; Luye Li[4], PhD; Micheal L Shier[1], PhD

[1]Factor Inwentash Faculty of Social Work, University of Toronto, Toronto, ON, Canada

[2]Faculty of Information, University of Toronto, Toronto, ON, Canada

[3]Artificial Intelligence for Justice Lab, University of Toronto, Toronto, ON, Canada

[4]Department of Sociology, Anthropology, Social Work, and Criminal Justice, Seton Hall University, South Orange, NJ, United States

[*]these authors contributed equally

Corresponding Author:
Jia Xue, PhD
Factor Inwentash Faculty of Social Work
University of Toronto
246 Bloor Street West
Toronto, ON, M5S 1V4
Canada
Phone: 1 416 946 5429
Email: jia.xue@utoronto.ca

Abstract

Background: Social media platforms have gained popularity as communication tools for organizations to engage with clients and the public, disseminate information, and raise awareness about social issues. From a social capital perspective, relationship building is seen as an investment, involving a complex interplay of tangible and intangible resources. Social media–based social capital signifies the diverse social networks that organizations can foster through their engagement on social media platforms. Literature underscores the great significance of further investigation into the scope and nature of social media use, particularly within sectors dedicated to service delivery, such as sexual assault organizations.

Objective: This study aims to fill a research gap by investigating the use of Twitter by sexual assault support agencies in Canada. It seeks to understand the demographics, user activities, and social network structure within these organizations on Twitter, focusing on building social capital. The research questions explore the demographic profile, geographic distribution, and Twitter activity of these organizations as well as the social network dynamics of bridging and bonding social capital.

Methods: This study used purposive sampling to investigate sexual assault centers in Canada with active Twitter accounts, resulting in the identification of 124 centers. The Twitter handles were collected, yielding 113 unique handles, and their corresponding Twitter IDs were obtained and validated. A total of 294,350 tweets were collected from these centers, covering >93.54% of their Twitter activity. Preprocessing was conducted to prepare the data, and descriptive analysis was used to determine the center demographics and age. Furthermore, geolocation mapping was performed to visualize the center locations. Social network analysis was used to explore the intricate relationships within the network of sexual assault center Twitter accounts, using various metrics to assess the network structure and connectivity dynamics.

Results: The results highlight the substantial presence of sexual assault organizations on Twitter, particularly in provinces such as Ontario, British Columbia, and Quebec, underscoring the importance of tailored engagement strategies considering regional disparities. The analysis of Twitter account creation years shows a peak in 2012, followed by a decline in new account creations in subsequent years. The monthly tweet activity shows November as the most active month, whereas July had the lowest activity. The study also reveals variations in Twitter activity, account creation patterns, and social network dynamics, identifying influential *social queens* and marginalized entities within the network.

Conclusions: This study presents a comprehensive landscape of the demographics and activities of sexual assault centers in Canada on Twitter. This study suggests that future research should explore the long-term consequences of social media use and

XSL•FO
RenderX

examine stakeholder perceptions, providing valuable insights to improve communication practices within the nonprofit human services sector and further the missions of these organizations.

(J Med Internet Res 2024;26:e50552) doi:10.2196/50552

KEYWORDS

social media; Twitter; sexual assault; nonprofits; Canada; violence; geolocation; communication

Introduction

Use of Social Media by Nonprofit Organizations

Social media platforms, including Twitter (subsequently rebranded as X [X Corp]), have gained popularity among nonprofit advocacy organizations as essential tools for communication and public engagement [1,2]. Nonprofit organizations are increasingly recognizing the strategic value of social media in fostering public engagement, securing donations, disseminating information, recruiting volunteers, and raising awareness about social issues [3-8]. Today, most large and mid-sized nonprofit organizations actively maintain at least 1 social media account, underscoring the extensive use of social media within the nonprofit realm [9].

Twitter, for instance, offers nonprofit organizations a platform to create profiles, establish networks, and engage socially through features such as tweeting, sharing multimedia content, replying, and retweeting [10]. Recognized as a cost-effective means of consistently reaching a broader audience [11,12], Twitter proves especially valuable for nonprofit organizations, often facing limited financial resources and dedicated communication staff [12], including sexual assault centers, to actively engage with key stakeholders and spark meaningful conversations [13,14]. Moreover, engaging in dialogues with other Twitter users forms a central aspect of communication for these organizations, facilitating increased supporter involvement, knowledge dissemination, and the creation of supportive communities [15,16]. Nonprofit organizations have successfully captured their followers' attention by regularly tweeting, responding to specific tweets, and retweeting other users' content [2]. This social media engagement can be harnessed by nonprofit organizations to share educational information and advocate for social causes [17].

Bridging and Bonding Social Capital

Social capital plays a critical role in understanding the effectiveness of nonprofit organizations, as it is embedded within their networks, enabling them to enhance their adaptive capabilities by consolidating shared interests and harnessing diverse resources [18,19]. Within the context of social capital, Putnam [20] distinguishes between 2 fundamental forms: bridging and bonding social capital. Bridging social capital encompasses the distant and weak connections between individuals from diverse backgrounds, facilitating information flow. This often manifests as nonmutual following relationships between organizations and a diverse public. In contrast, bonding social capital revolves around preexisting and robust ties that reinforce homogeneity among groups, fostering emotional and social support. An illustrative example of this is the mutual following relationships observed between similar organizations [21,22].

From a social capital perspective, relationship building is seen as an investment. It involves a complex interplay of tangible and intangible resources, both embedded within existing relationships and generated through the act of forging new ones [23,24]. The success of nonprofit organizations relies significantly on their capacity to establish high-quality relationships with key stakeholders, including donors, clients, grant makers, seekers, and the broader public [25,26]. The social capital of nonprofits comprises the wealth of resources intricately embedded within these strategic alliances and stakeholder relationships [27]. Xu and Saxton [26] propose that the effective acquisition of social capital, at an elevated level, relies on the scope and quality of stakeholder connections. Their study introduces and demonstrates the significance of 2 primary stakeholder engagement strategies: content-based and connection-based strategies. This study underscores that the attainment of social capital is less about the number of stakeholder engagements and more about the breadth of those engagements. This breadth includes diverse stakeholder connections.

Social Media–Based Social Capital

Social media–based social capital signifies the diverse social networks that organizations can foster through their engagement on social media platforms [26]. The potential of social media to nurture and sustain web-based–offline social capital is substantial, although its effectiveness varies across platforms and strategies [28]. Platforms such as Facebook (Meta Platforms), Twitter, and Instagram (Meta Platforms) offer distinctive usability features that influence the dynamics of bridging and bonding social capital among their users. An important study conducted by Phua et al [22] examined the impact of 4 major social networking sites (Facebook, Twitter, Instagram, and Snapchat [Snap Inc]) on the development of web-based bridging and bonding social capital among 297 users. Their findings indicate that Twitter users exhibit the highest levels of bridging social capital, followed by Instagram, Facebook, and Snapchat. Conversely, when it comes to bonding social capital, Snapchat users demonstrate the highest levels, followed by Facebook, Instagram, and Twitter. Furthermore, research suggests a direct correlation between the number of followers and the development of bonding social capital [21]. Another study by Xu and Saxton [26], focusing on 198 community foundations, reinforces the importance of social media engagement strategies tailored to multiple intersectoral stakeholders and diverse communication patterns, which substantially contribute to the development of social media–based social capital.

XSL•FO
RenderX

In the context of nonprofit organizations, studies by Henry and Bosman [29] and Lee and Shon [30] underscore the positive impact of web-based social capital generated through social networking sites on charitable outcomes. These studies reveal that the quantity of Twitter followers is linked positively with personal contributions, although not necessarily with full-time equivalent volunteers [30]. Moreover, Xu and Saxton [26], drawing from Twitter data consisting of 198 community foundations, highlight the pivotal role of stakeholder engagement diversity over connection quantities. They emphasize the significance of using multiple communicative cues, such as message elements, and targeting intersectoral and interregional stakeholders in the successful acquisition of social capital through social networking sites. Leveraging social media platforms offers numerous advantages to nonprofit organizations, including the engagement of a donor base within the general population [1,31], the facilitation of communication strategies through the dissemination of information to a broader global audience [32], and the support of advocacy efforts for social change and community mobilization [17]. Svensson et al [33] examined the Twitter use of sport-for-development organizations and identified varying levels of engagement across different entities, potentially limiting the cultivation of social media–based social capital within this sector. Investigating the extent of social media use serves as a valuable tool to inform recommendations aimed at enhancing nonprofits' web-based presence and fostering social media–based social capital [2,4,34].

These findings underscore the great significance of further investigation into the scope and nature of social media use, particularly within sectors dedicated to service delivery, such as sexual assault programs and organizations, which share a common mission and focus of their efforts. The endeavor to augment social capital through social media within a given sector has the potential to expand the donor and volunteer base, engage the community in matters affecting everyone, and catalyze broader social change at the policy level by mobilizing concerned citizens.

Aim of the Study

This study aims to address the existing research gap surrounding the use of social media platforms such as Twitter by specific organizations, such as sexual assault support agencies. This study intends to investigate user activities, demographics, and social network structures within these organizations on Twitter. By doing so, we aim to contribute to a better understanding of the current state of social media adoption and social network structures within sexual assault organizations in Canada. In addition, this study provides valuable insights and recommendations for building social capital among the sexual assault organizations on social media. To achieve these goals, we formulated the following research questions (RQs):

- RQ1: What is the demographic profile of sexual assault centers in Canada with official Twitter accounts, including their geographical distribution, years of establishment, and age of the organizations?
    - RQ1a: How prevalent are sexual assault centers in Canada with official Twitter accounts, and which provinces and territories have the highest number of centers actively using Twitter?
    - RQ1b: What are the geographic locations of sexual assault centers with official Twitter accounts in Canada?
    - RQ1c: In what years were the Twitter accounts of sexual assault centers in Canada established, and are there any differences in account creation among provinces and territories?
    - RQ1d: What is the average age of these centers since establishing their official Twitter accounts, and do any differences in account creation exist among provinces and territories?
- RQ2: What is the user activity of sexual assault centers in Canada with official Twitter accounts, including their posting frequency by month and year?
    - RQ2a: How many sexual assault centers maintain an active Twitter account each year in each province or territory?
    - RQ2b: What are the Twitter activity and posting patterns of sexual assault centers while they are active on Twitter?
    - RQ2c: How do the Twitter activity and posting patterns vary across provinces and territories?
- RQ3: What are the social network dynamics of bridging and bonding the social capital of sexual assault centers in Canada?
    - RQ3a: What are the variations in network size, specifically in terms of followers and followings, among sexual assault centers in different provinces and territories in Canada?
    - RQ3b: What is the relationship between followers and followings of these organizations on Twitter?
    - RQ3c: What insights can be gained from the social network structure of sexual assault centers on Twitter?

Methods

Sampling

This study used purposive sampling to select sexual assault centers in Canada. Our sampling frame was developed by combining lists of sexual assault centers by province and territory from 2 sources: the Canadian Association of Sexual Assault Centres website and the Sexual Assault Centres, Crisis Lines, and Support Services directory. After removing duplicates, our sample frame consisted of 350 sexual assault centers across 10 provinces and 3 territories, providing basic information such as center name, phone number, email, and website. Our inclusion criteria were that the sexual assault center had a Twitter account and had posted at least 1 tweet. To confirm eligibility, a research assistant manually searched the home page of these centers and Twitter pages and conducted Google searches. We determined that 127 organizations had Twitter accounts, but 3 of them had never tweeted anything. As a result, our final sample consisted of 124 Twitter accounts belonging to sexual assault centers across 9 provinces and the Yukon and Northwest Territories. It should be noted that there

XSL•FO
RenderX

were no sexual assault centers in Prince Edward Island and Nunavut that used Twitter.

Twitter Handles' Acquisition

We collected the Twitter account name, location (eg, Toronto, Ontario), and Twitter handle (eg, @ABCD) for each sexual assault center's Twitter account. The Twitter handle represented as "@name" is used by followers when replying to, mentioning, and sending direct messages to an account. We identified 22 duplicate Twitter handles among the sampled centers. As a result, our final sampling list consisted of 113 unique Twitter handles obtained from 124 centers. We gathered this information directly from the home page of each sexual assault center's Twitter account.

Data Collection

Acquisition of Twitter IDs

To collect the data necessary for this study, we obtained Twitter IDs for the 113 unique Twitter handles in our sample. A Twitter ID (eg, 12345678) is a unique numeric value associated with each Twitter handle, and it cannot be changed. We converted each Twitter handle to its corresponding Twitter ID. To ensure the accuracy of our conversions, 2 research assistants verified the results using 3 different websites: TweeterID, CodeOfaNinja, and Comment Picker.

Collection of Tweets

We used the 113 Twitter IDs associated with the sampled 124 sexual assault centers to collect their corresponding tweets. To accomplish this, we used Twitter's academic search application programing interface (API) full archive end point and timeline end point, which allowed us to retrieve tweets published as early as 2006 [35]. We accessed the Twitter API using the native rest API requests. Our data collection process was conducted on March 15, 2023. We downloaded all tweets posted by the sampled centers from the date of each account's establishment to March 15, 2023. Our data set included 294,350 tweets from 124 sexual assault centers, crisis lines, or support services. We obtained a substantial portion of the total number of tweets published by each Twitter ID on Twitter, specifically, >93.54%.

Data Features

We collected several features for each individual tweet message, including the user ID (user_id_str), user account creation date (user_created_at), user location (user_location), username (user_name), user screen name (user_screen_name), tweet creation time (tweet_created_time), full text of the tweet (full_text), and full text of any retweeted status (retweeted_status_full_text).

On Twitter, users commonly use functions such as retweets, replies, mentions, and hashtags. Retweets refer to publicly shared tweets between users and their followers. Users can also add their own comments and media before retweeting. In addition, users can participate in conversations on Twitter by replying to other users and mentioning them in their tweets. Finally, hashtags allowed users to easily follow and search for topics of interest.

Data Analysis

Preprocessing of Raw Data

To address our RQs, we preprocessed the raw data using the following steps:

1. We removed URLs from the tweets.
2. We removed all punctuation marks, with the exception of apostrophes, which are important for contextual meaning in certain words (eg, "We're").
3. We removed any bigrams from the set if either of its elements belonged to the list of stop words. For instance, the phrase "The increasing awareness about sexual assault" would generate the bigrams "the increasing," "increasing awareness," "awareness about," "about sexual," and "sexual assault." In this case, the stop words "the" and "about" would be removed from the bigrams "the increasing," "awareness about," and "about sexual," leaving the bigrams "increasing awareness" and "sexual assault" in the set.

Descriptive Analysis

Descriptive analysis was used to calculate the number of sexual assault centers in each province, the number of centers created each year, and their average age. The age of each sexual assault center was determined by dividing the month of March 2023 by the establishment date of that center. For example, we used R's *difftime* method (R Core Team) to calculate the age of center A's Twitter account, which was created on March 12, 2009. By subtracting "2009-03-12" from "2023-03-15" to obtain the time difference, we determined that this center has an age of 14 years.

Geolocation Mapping of Sexual Assault Centers

We used the Twitter accounts' IDs and sexual assault center locations to determine the actual locations of each tweet sent by the 124 centers. The locations were plotted and visualized on a map of Canada. One research assistant manually (QZ) retrieved center location information, including the city, region, and province, from the centers' official websites and obtained the longitudes and latitudes of the cities where the centers were located. The Google API was used to calibrate the geolocations if the absolute distance discrepancy between manually identified geolocations and Google map geolocations was >6 km. Finally, we developed a Python script to automatically generate D3.js for mapping all the centers with their latitudes and longitudes (the script is available upon request). We used Figma to indicate the cities on the map [36].

Social Network Analysis

Social network analysis is one of the most effective techniques for visualizing and assessing network connectivity dynamics, offering insights into patterns of connection and disconnection among participants at a given moment. In our study, we used social network analysis to construct networks from Twitter accounts, where nodes represented accounts and directed edges symbolized follower relationships. We used *Pyvis* [37] and *NetworkX* in Python to create the network, resulting in 111 nodes and 995 edges, which is a visual representation of the relationships within the sexual assault center's Twitter accounts. To delve deeper, we applied a range of metrics: (1) density, which quantifies the percentage of actual connections within

XSL•FO
RenderX

the network; (2) degree centrality, which evaluates a node's significance by examining its connections, distinguishing between incoming (in-degree) and outgoing (outdegree) connections [38]; (3) eigenvector centrality, which measures a node's influence in a network by considering the relative scores to connected nodes and is based on the concept that connections to nodes with higher scores exert a more significant influence on determining the node's score, in contrast to connections with nodes having lower scores [39]; (4) modularity, indicating the network's community organization strength through clustering [40,41]; (5) betweenness centrality, evaluating an individual's role as a bridge between unconnected entities, fostering vital connections among clusters, communities, and organizations [38]; and (6) closeness, gauging a node's centrality in a connected graph by summing the shortest path lengths to all other nodes [42]. These metrics collectively provided a comprehensive understanding of the social network's structure, shedding light on its various facets and opportunities for relationship cultivation and network analysis.

Ethical Considerations

The data set and analyses relied on publicly accessible secondary Twitter data; thus, no ethics approval or organizational consent was necessary. The study data presented in this manuscript were subjected to anonymization and deidentification procedures. All personally identifiable information, including but not limited to individual organizations' identities, pictures, user-specific data, or tweets that have not been rephrased, have been meticulously removed from the data set to ensure complete anonymity.

Results

Demographic Profile of Sexual Assault Centers on Twitter

Prevalence of Sexual Assault Centers on Twitter

We found that 124 (35.4%) centers out of the 350 sampled sexual assault centers have an official Twitter account and have posted at least 1 tweet since their establishment. We investigated the locations of the 124 sexual assault centers in Canada that have official Twitter accounts. The province with the highest number of sexual assault centers was Ontario (n=34), followed by British Columbia (BC; n=24) and Quebec (n=23). These 3 provinces accounted for two-thirds (81/124, 65.3%) of all sampled sexual assault centers in Canada. Newfoundland and Labrador and Yukon have only 1 sexual assault center each. Figure 1 shows the prevalence of sexual assault centers in Canada that have posted tweets.

XSL•FO
RenderX

Figure 1. Sexual assault centers in Canada. AB: Alberta; BC: British Columbia; MB: Manitoba; NB: New Brunswick; NL: Newfoundland and Labrador; NS: Nova Scotia; NT: Northwest Territories; NU: Nunavut; ON: Ontario; PE: Prince Edward Island; QC: Quebec; SK: Saskatchewan; YT: Yukon.

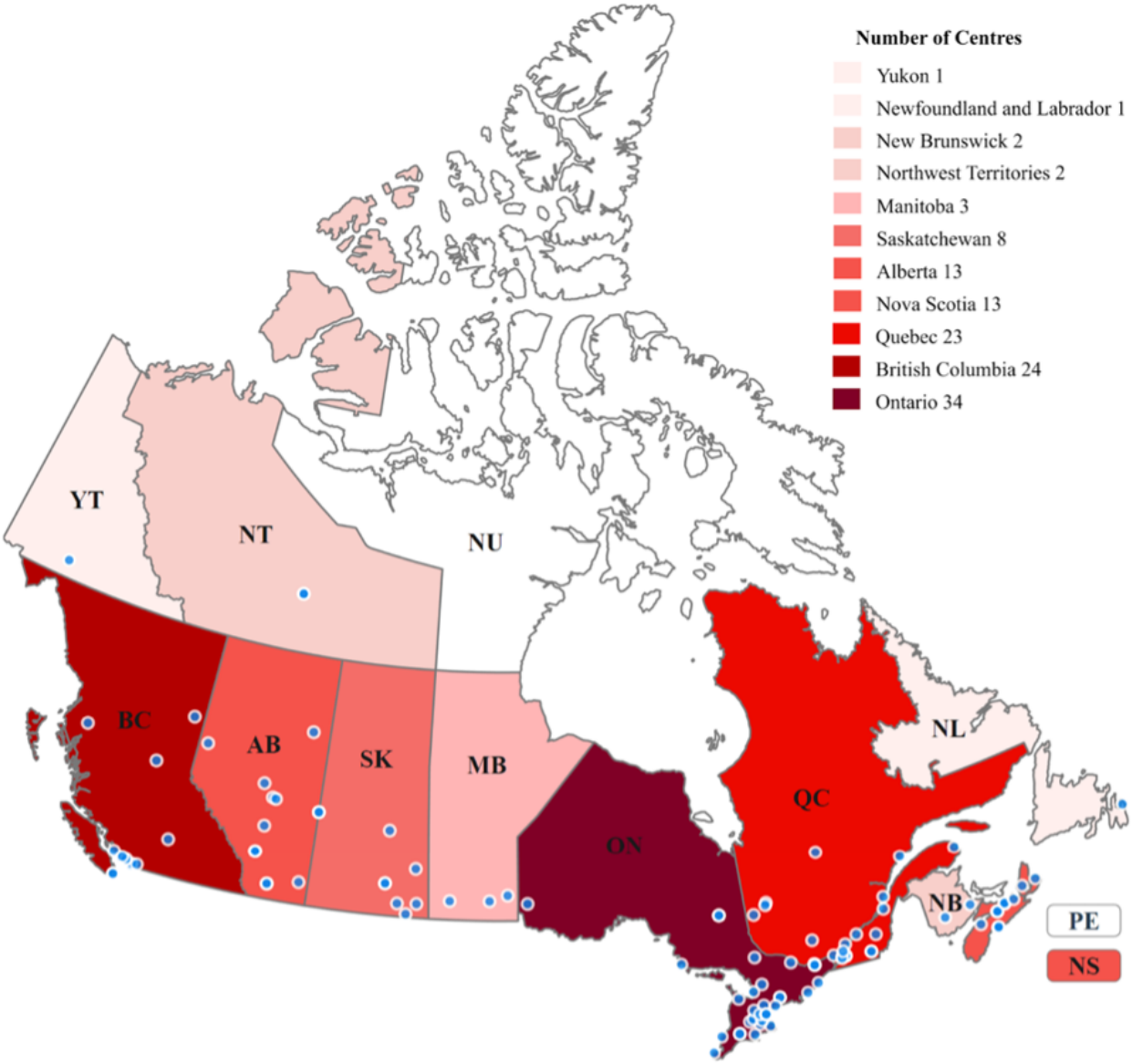

The Geographic Distribution of Sexual Assault Centers on Twitter

To visualize the distribution of sexual assault centers with Twitter accounts, we created geographic distribution maps for Ontario, BC, and Quebec, which had the highest number of centers in our sample. Additional geographic distributions of Twitter accounts in the remaining provinces and territories are presented in Multimedia Appendix 1.

Ontario

Our sample included 34 sexual assault centers from 27 cities in Ontario, shown in Figure 2. Our analysis revealed that most of these centers were concentrated in the southeast region of the province, which is also where most of Ontario's population resides [43]. Specifically, many centers were found in the cities of Toronto, Ottawa, Peterborough, Timmins, Brampton, and London, which have larger populations in Ontario.

XSL•FO
RenderX

Figure 2. Sexual assault centers in Ontario (ON; 34 centers). QC: Quebec.

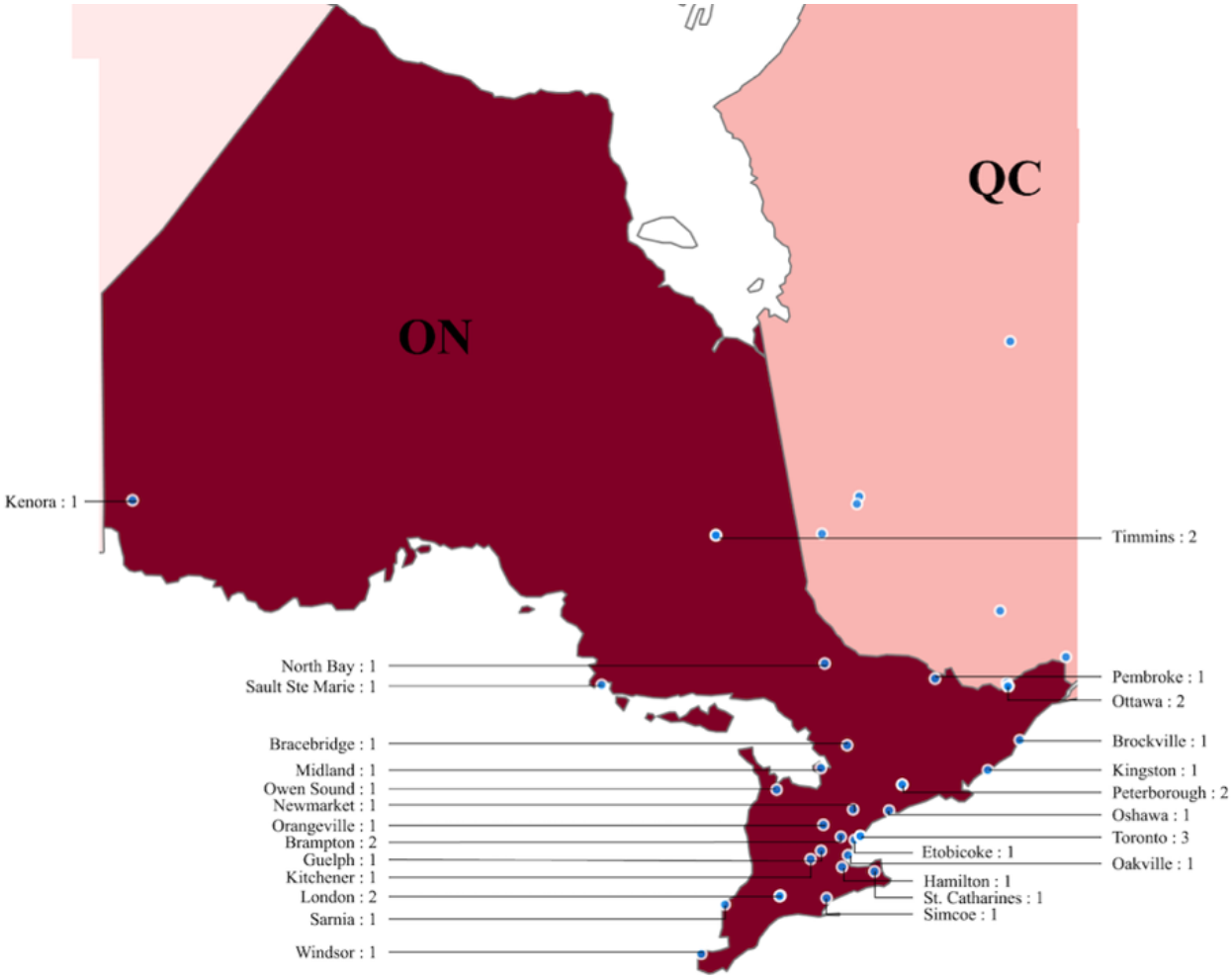

British Columbia

A total of 24 sexual assault centers were identified in BC, spread across 15 cities, shown in Figure 3. Vancouver had the highest number of centers in the province, followed by Surrey. As per population distribution, most of the population in BC resides in the southern part of the province [44], and similarly, most of the sampled centers are located in the southern region of BC.

XSL•FO
RenderX

Figure 3. Sexual assault centers in British Columbia (BC: 24 centers).

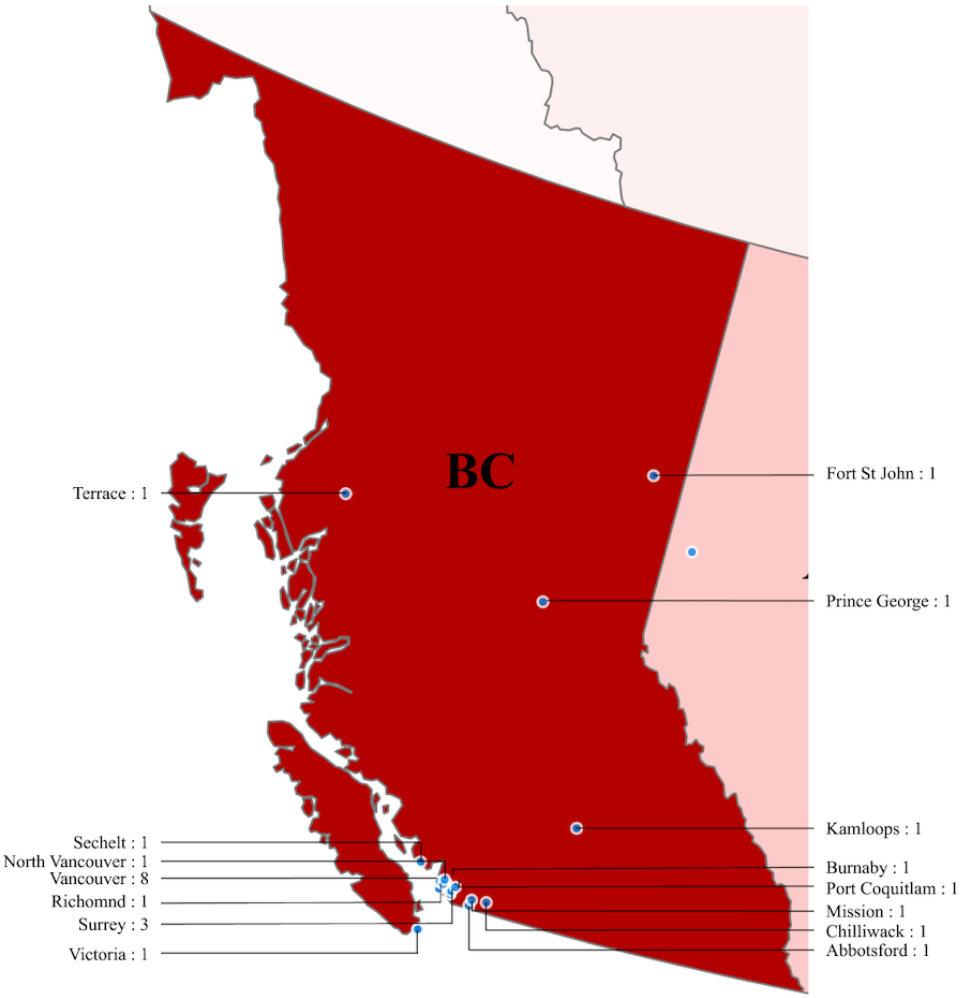

Quebec

In Quebec, there are 23 sexual assault centers located in 17 cities, with 5 centers located in Montreal, shown in Figure 4. We found that nearly all the centers are situated in the southern region of Quebec, where most of the province's population resides [45]. Notably, no centers were located in the northern region of the province.

XSL•FO
RenderX

Figure 4. Sexual assault centers in Quebec (QC; 23 centers).

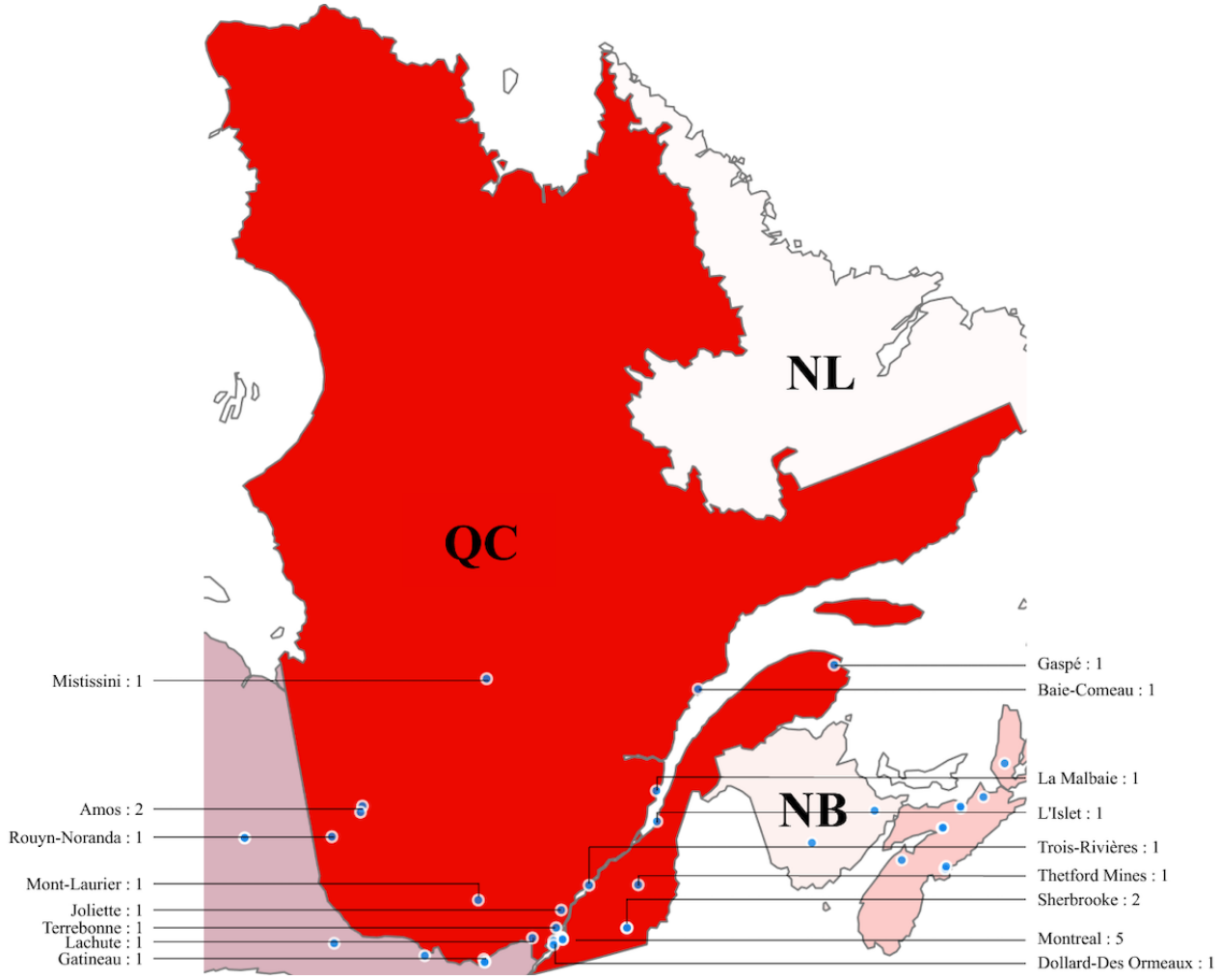

Twitter Account Creation Year by Sexual Assault Centers on Twitter: Provincial and Territorial Differences

We analyzed the year of establishment of the Twitter accounts and assessed whether there were any differences in account creation among provinces and territories. Figure 5 shows our analysis results of the sampled sexual assault centers' Twitter account creation year. The first Twitter account was created in 2009, and the number of sexual assault centers gradually increased from 2009 to 2012, peaking in 2012 with 23 new accounts. However, from 2013 to 2017, the number of Twitter accounts created by sexual assault centers decreased. In 2019, only 1 sexual assault center created Twitter accounts, and there were no new accounts in 2021. In 2022, we identified 3 centers that had established new Twitter accounts. Notably, all 3 centers that created new Twitter accounts had been in operation for >30 years, as confirmed by our examination of their official profiles.

XSL•FO
RenderX

Figure 5. Establishment years of Twitter accounts for sexual assault centers across provinces and territories. The Lloydminster Sexual Assault and Information Centre serves both Alberta and Saskatchewan, so we counted it for both provinces.

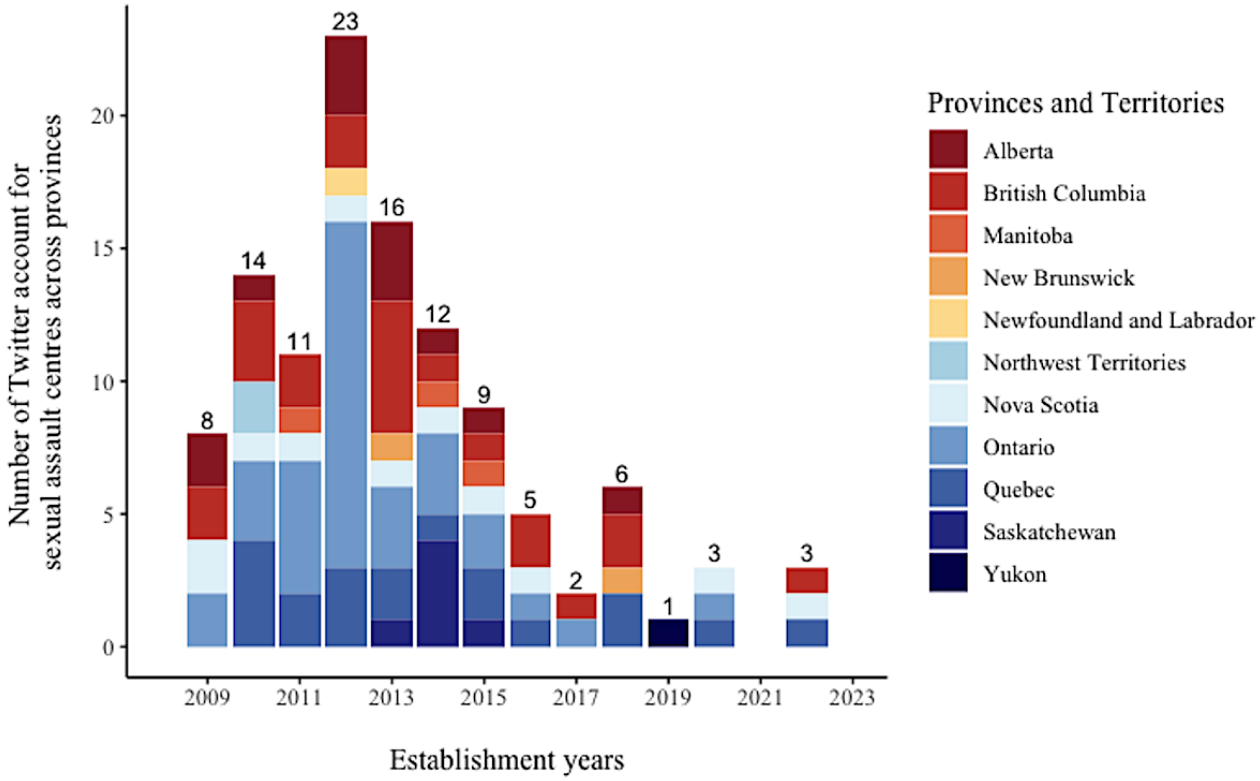

In addition, we analyzed the distribution of the Twitter accounts created across provinces from 2009 to 2023 and found provincial differences. A total of 8 sexual assault centers from 4 provinces, including Alberta, BC, Nova Scotia, and Ontario, created their Twitter accounts in 2009. From 2009 to 2017, we observed a recurring trend among centers located in Ontario and BC, the 2 provinces with the highest number of centers in Canada, where they established new Twitter accounts on an annual basis. Ontario contributed to the newest Twitter accounts created in 2011 (n=5) and 2012 (n=13). In Quebec, the third-largest province in terms of the number of centers with Twitter accounts, all sexual assault centers began using Twitter after 2010, and remarkably, 7 new centers were created in that year alone. We also noted that most centers in Saskatchewan established their Twitter accounts in 2014. In addition, 3 sexual assault centers located in the Canadian territories also created their Twitter accounts. Specifically, in 2019, one organization in Yukon established a Twitter account, whereas in 2010, two centers located in the Northwest Territories created their Twitter accounts.

Average Age of Sexual Assault Centers on Twitter by Province and Territory

We calculated the age of each sexual assault center's Twitter account since its creation. We determined the duration of time that each sexual assault center had its official Twitter account by subtracting the most recent month of collected tweets (March 2023) from the account creation date. Then, we computed the average length of time for all sexual assault centers in each province and presented the results in Table 1.

XSL•FO
RenderX

Table 1. Descriptive statistics for sexual assault centers' Twitter accounts by province and territory.

Province or territory	Age (y), mean (SD; range)	Portion (%)	Tweets, n	Followers, n	Following, n	Favorites, n	Listed, n
Alberta (n=13)	10.26 (2.39; 5-14)	10.48	39,021	14,115	8167	39,525	296
British Columbia (n=24)	9.33 (3.23; 1-14)	19.35	73,602	29,937	16,635	39,210	730
Manitoba (n=3)	9.30 (1.99; 8-12)	2.42	5691	3004	1421	3966	54
New Brunswick (n=2)	7.31 (3.09; 5-10)	1.61	355	428	480	440	3
Newfoundland and Labrador (n=1)	10.52 (0.00; 11-11)	0.81	1810	1669	504	904	22
Nova Scotia (n=13)	9.55 (4.14; 0-14)	1.61	22,673	14,032	6118	19,016	300
Northwest Territories (n=2)	12.77 (0.23; 13-13)	10.48	1043	659	212	172	34
Ontario (n=34)	10.33 (2.20; 3-14)	27.42	134,467	48,263	27,775	59,601	1111
Quebec (n=23)	9.34 (3.36; 1-12)	18.55	8801	8007	4805	4525	179
Saskatchewan (n=8)	8.75 (0.60; 8-10)	6.45	5899	1826	1915	1859	14
Yukon (n=1)	3.34 (0.00; 3-3)	0.81	1551	187	484	1314	1
Total (n=124)	9.17 (1.29; 0-14)	100	294,913	122,127	68,516	170,532	2744

Table 1 displays the average length of time, SD, and range of the duration of years after the establishment of Twitter accounts by sexual assault centers in each province. For example, 13 sexual assault centers in Alberta created Twitter accounts, and the average number of years since their accounts' establishment was 10.26 (SD 2.39) years. The range of 5 to 14 indicated that the earliest Twitter account was created in 2009 (March 2023: 14 y), whereas the most recent account was established in 2018 (March 2023: 5 y). These 13 sexual assault centers accounted for 10.5% (n=124) of the total 124 sexual assault centers in Canada.

To obtain a comprehensive understanding of Twitter presence and engagement, we used the metadata of Twitter users related to sexual assault centers in each province, which we obtained from the Twitter timeline API. Specifically, we extracted data such as "followers_count," "friends_count," "favorites_count," and "listed_count" to determine the total number of followers, following, favorites, and listed users, respectively, for each province. We also aggregated the collected data by province to calculate the total number of tweets posted for each province.

User Activity of Sexual Assault Centers on Twitter

Active Twitter Accounts in Canadian Sexual Assault Centers by Province and Territory

We analyzed the data to determine the number of active Twitter accounts maintained by sexual assault centers in each Canadian province and territory each year. We defined an active account as one that posted at least 1 tweet in a given year. The results indicated a steady increase in the number of active Twitter accounts in Ontario and BC since 2009, as shown in Figure 6. In Quebec, there was an increase in the number of centers from 3 in 2011 to 9 in 2015, but this trend reversed in the following years, indicating that many Twitter accounts became inactive. Although there was an increase in the number of registered Twitter accounts from 13 in 2015 to 16 in 2018, only 4 of these accounts remained active in 2020, down from 9 in 2015. Alberta ranks third or fourth in terms of active Twitter accounts.

XSL•FO
RenderX

Figure 6. The yearly count of active Twitter accounts maintained by sexual assault centers in each Canadian province or territory (Alberta [AB]: 13 centers; British Columbia [BC]: 24 centers; Manitoba [MB]: 3 centers; New Brunswick [NB]: 2 centers; Newfoundland and Labrador [NL]: 1 center; Nova Scotia [NS]: 2 centers; Northwest Territories [NT]: 2 centers; Nunavut [NU]: 0 centers; Ontario [ON]: 34 centers; Prince Edward Island [PE]: 0 centers; Quebec [QC]: 23 centers; Saskatchewan [SK]: 8 centers; Yukon [YT]: 1 center).

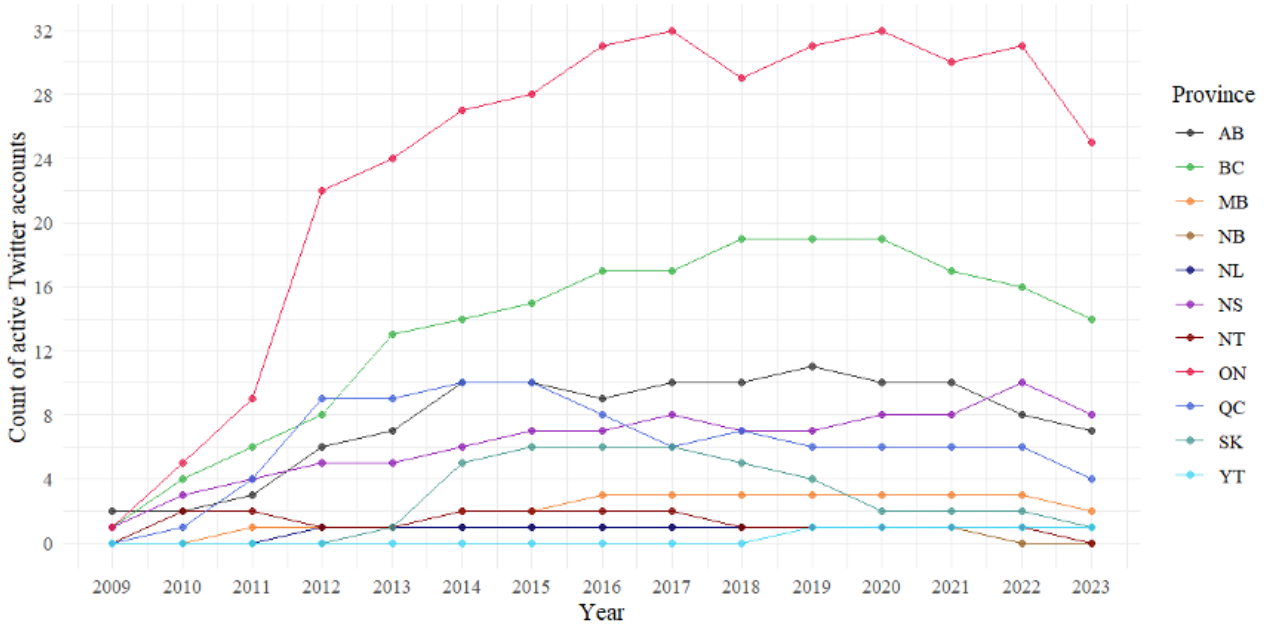

In Manitoba, 2 sexual assault centers had active Twitter accounts until 2020. Only 1 center in New Brunswick registered a Twitter account in 2013 but was inactive in 2018 and 2020 while maintaining activity in other years. In Newfoundland and Labrador, the only sexual assault center remained active on Twitter from 2012 to 2020. Nova Scotia consistently showed an increasing trend in the number of active Twitter accounts from 2009 to 2020. Meanwhile, the only center in the Northwest Territories maintained an active Twitter account since 2010. In Saskatchewan, the number of active Twitter accounts increased to 6 in 2015 but decreased to 3 in 2020.

Twitter Activity and Posting Patterns of Sexual Assault Centers

We also examined the Twitter activity of sexual assault centers in Canada, investigating the popular times for tweeting and the number of tweets posted per month. Figure 7 shows the total number of tweets posted by all sampled centers aggregated by month. Over a 12-year period, these centers posted an average of 12,849 tweets per month. The most active month was November, with a total of 16,239 tweets, whereas the least active was July, with only 9079 tweets. March and May were the peak tweeting months, with >15,000 tweets, whereas August had the fewest tweets, with approximately 9500 tweets.

XSL•FO
RenderX

Figure 7. Total count of monthly tweets by all sexual assault centers in Canada.

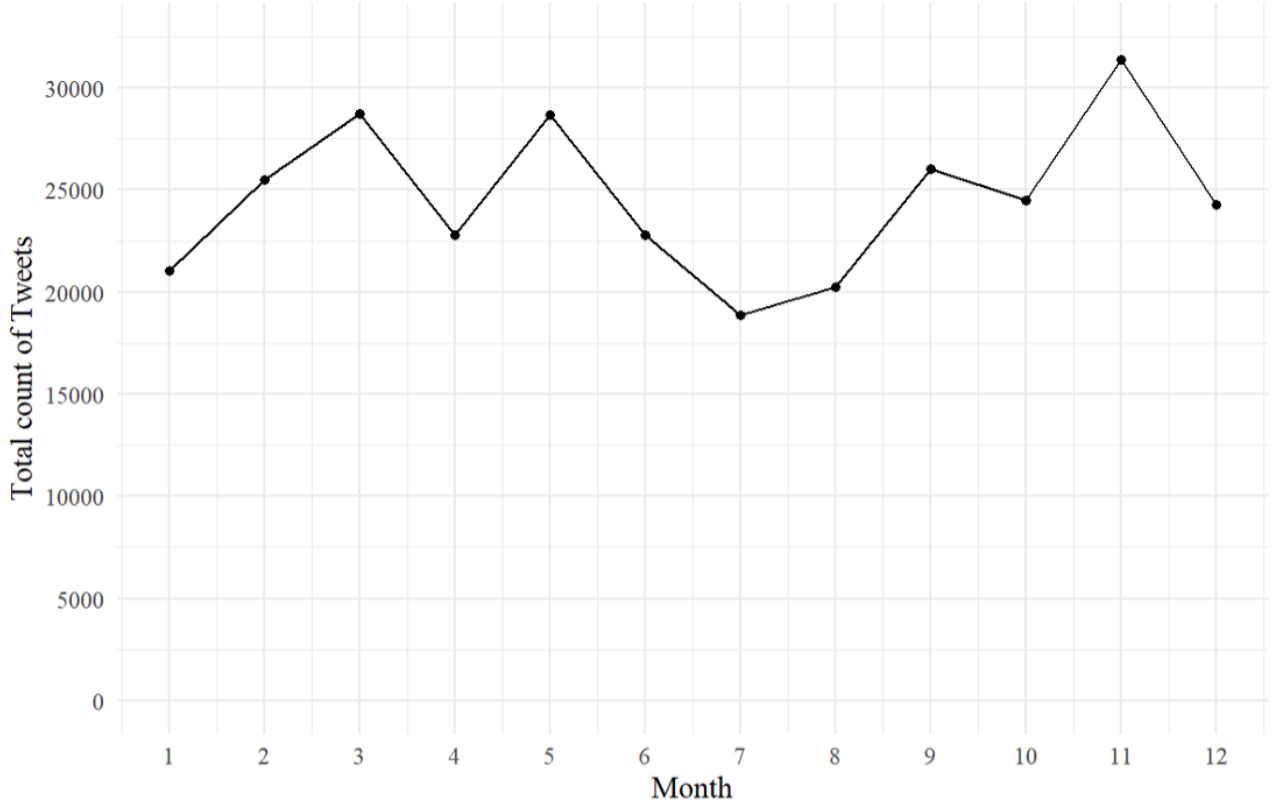

We analyzed the monthly average tweet count of sexual assault centers during their active status (n=92). We computed the average number of tweets sent by each center per month from the first tweet after creating their account to the last tweet during the data collection period. Results showed that a large majority of the centers have a relatively low tweeting frequency, with the highest frequency of centers (over 20) averaging between 0 to 10 tweets per month. The distribution is right-skewed, showing that as the average monthly tweeting volume increases, the number of centers engaging at the level decreases. A small number of centers tweet between 10 and 30 times per month, and very few centers exceed this range. There are occasional outliers, with one center in particular averaging a significantly higher number of tweets per month, at around 180. This center is an extreme outlier in comparison to the rest of the data set. Overall, our findings suggest that sexual assault centers tend to use Twitter moderately, with the bulk of them tweeting less than 20 times per month, and a very exceptional few tweeting much more frequently.

Comparative Analysis of Tweet Activity Across Provinces and Territories

To answer RQ5, we further analyzed the total number of tweets posted by centers in different provinces and territories each month and compared the number of tweets posted across provinces and territories each year. Figure 8 shows the annual tweet activity generated from all sexual assault center accounts across provinces and territories, with data collected until March 15, 2023. Our analysis of the tweet activity revealed a gradual increase in tweeting volume across provinces and territories until 2017. Notably, Ontario exhibited the highest frequency of tweet activity in 2016, with approximately 17,000 tweets. However, with the exception of accounts in BC and Nova Scotia, the tweet activity gradually declined from 2017 to March 2023, returning to activity levels last observed in 2013 or 2014. It is worth noting that some provinces and territories, such as New Brunswick and Northwest Territories, had <100 tweets in the peak month, and hence, they were not included in our figure. The findings suggest a potential decrease in Twitter activity among sexual assault centers in recent years.

XSL•FO
RenderX

Figure 8. Tweet volume of sexual assault centers by provinces and years. Note: we did not list the NB and NT provinces, which had <100 tweets in the peak month (Alberta [AB]: 13 centers; British Columbia [BC]: 24 centers; Manitoba [MB]: 3 centers; New Brunswick [NB]: 2 centers; Newfoundland and Labrador [NL]: 1 center; Nova Scotia [NS]: 2 centers; Northwest Territories [NT]: 2 centers; Nunavut [NU]: 0 centers; Ontario [ON]: 34 centers; Prince Edward Island [PE]: 0 centers; Quebec [QC]: 23 centers; Saskatchewan [SK]: 8 centers; Yukon [YT]: 1 center).

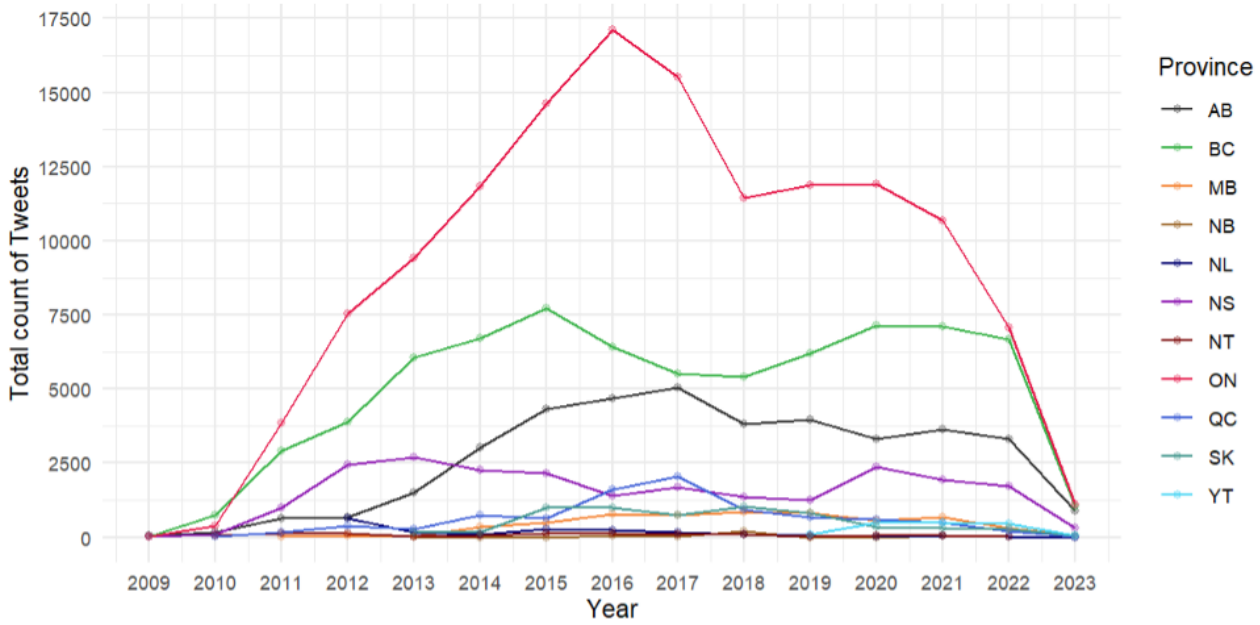

Social Network Dynamics of Sexual Assault Centers in Canada

Network Size

The sexual assault organizations under investigation had an average follower count of 1543 (SD 1555) and an average following count of 819 (SD 876). The range of follower counts was quite diverse, starting at a minimum of 8 followers for one organization and reaching a maximum of 7228 followers for another organization. Similarly, the following counts also showed significant variation, with one center having the lowest count of 1 following, whereas another center had the highest following count of 4458. More detailed information about the Followers, followings, and measurement of social network analysis are in Multimedia Appendix 2.

Sexual assault organizations across 11 provinces exhibited varying average follower counts, ranging from as low as 1 follower in New Brunswick to 17 followers in Ontario, as shown in Table 2. The organization located in Ontario had the highest number of followers, totaling 46. In contrast, some sexual assault organizations in provinces such as Alberta and BC had no followers at all. Similarly, the average number of followings by these organizations across the 11 provinces and territories ranged from 0 in New Brunswick to 18 in Ontario. The organization with the most followings was also situated in Ontario, with a total of 38 followings, whereas several organizations in provinces such as New Brunswick and Quebec had no followings.

XSL•FO
RenderX

Table 2. Descriptive statistics of followers and followings from 11 provinces and territories.

Province or territory	Followers, mean (SD; range)	Followings, mean (SD; range)
Alberta	7.55 (6.83; 0-25)	7.91 (6.25; 0-21)
British Columbia	6.48 (8.70; 0-30)	5.52 (5.06; 0-20)
Manitoba	2.67 (2.08; 1-5)	3 (2.65; 1-6)
Manitoba	1 (N/A[a]; 1-1)	0 (N/A; 0-0)
Newfoundland and Labrador	7 (N/A; 7-7)	5 (N/A; 5-5)
Northwest Territories	1.50 (0.71; 1-2)	3 (1.41; 2-4)
Nova Scotia	4.91 (3.86; 0-15)	5.18 (5.95; 0-21)
Ontario	17.20 (11.55; 0-46)	17.60 (10.22; 1-38)
Quebec	4.20 (3.71; 0-13)	4.30 (4.14; 0-14)
Saskatchewan	2 (1.83; 0-4)	1.75 (0.96; 1-3)
Yukon	2.50 (2.12; 1-4)	1.50 (2.12; 0-3)

[a]N/A: not applicable.

Relationship Between Followers and Followings on Twitter

We analyzed the Twitter user network by exploring the connections between the followers and the following lists. Figure 9 shows the log-log plot of the correlation between followers and followings. Each point on the graph represents an individual user, with the x-axis representing the user's followers and the y-axis indicating their following count. The plot demonstrates that as the number of followers continued to increase, the followings also indicated an increasing trend. At the middle of the plot, users with a medium number of followers have a high number of followings.

XSL•FO
RenderX

Figure 9. Relationship between followers and followings on Twitter.

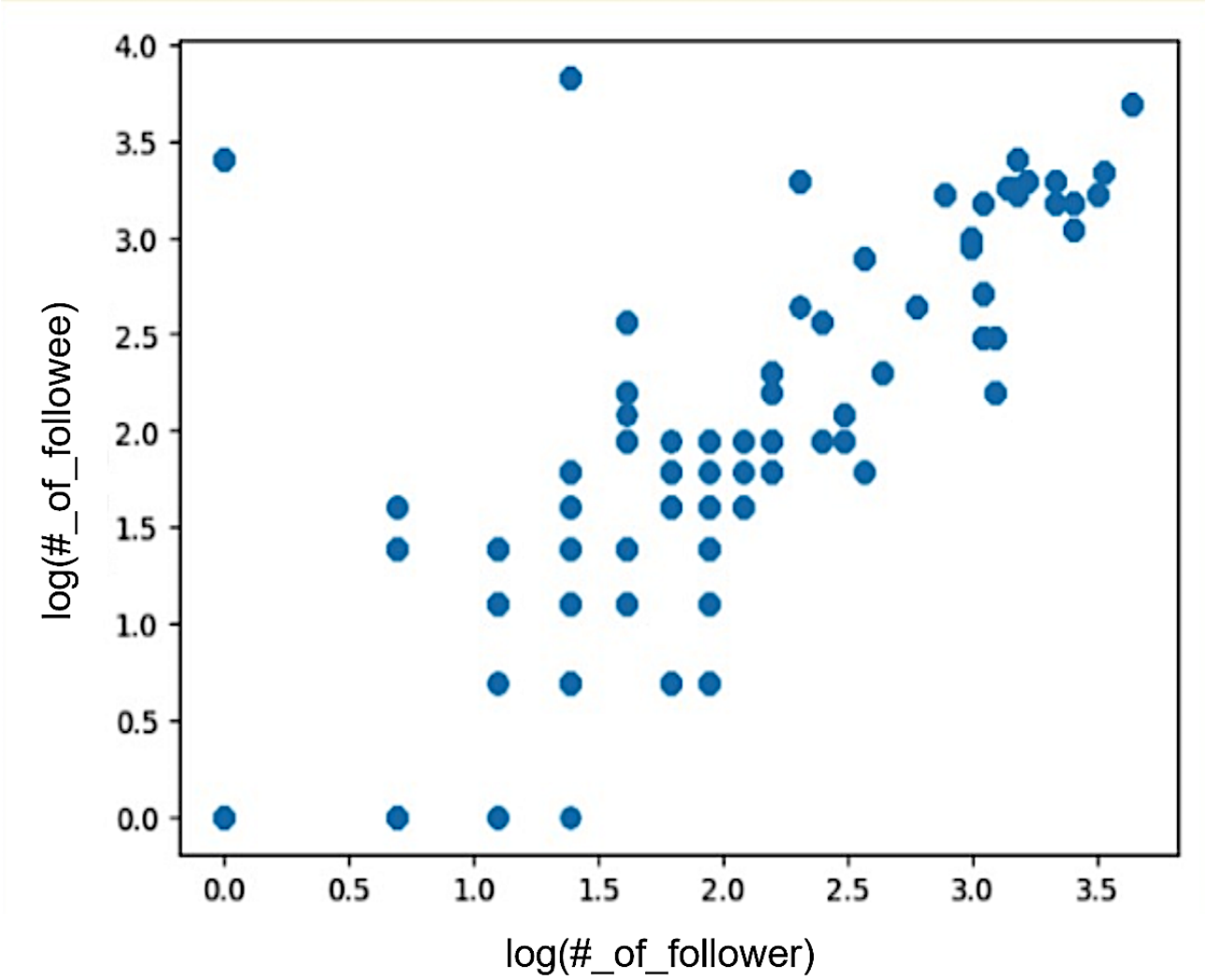

Analysis of Twitter Social Network Structure and Node Categorization

Figure 10 presents a full network map, illustrating the relationships between followers and followings among 111 sexual assault centers on Twitter. This graph features 111 nodes and 995 edges. Each node represented a sexual assault center on Twitter, and each edge is directional, with arrows symbolizing "x follows y" relationship. In this context, Y serves as the following node, implying that it is followed by other nodes, whereas X acts as the follower node, signifying that it follows other nodes.

XSL•FO
RenderX

Figure 10. Social network structure.

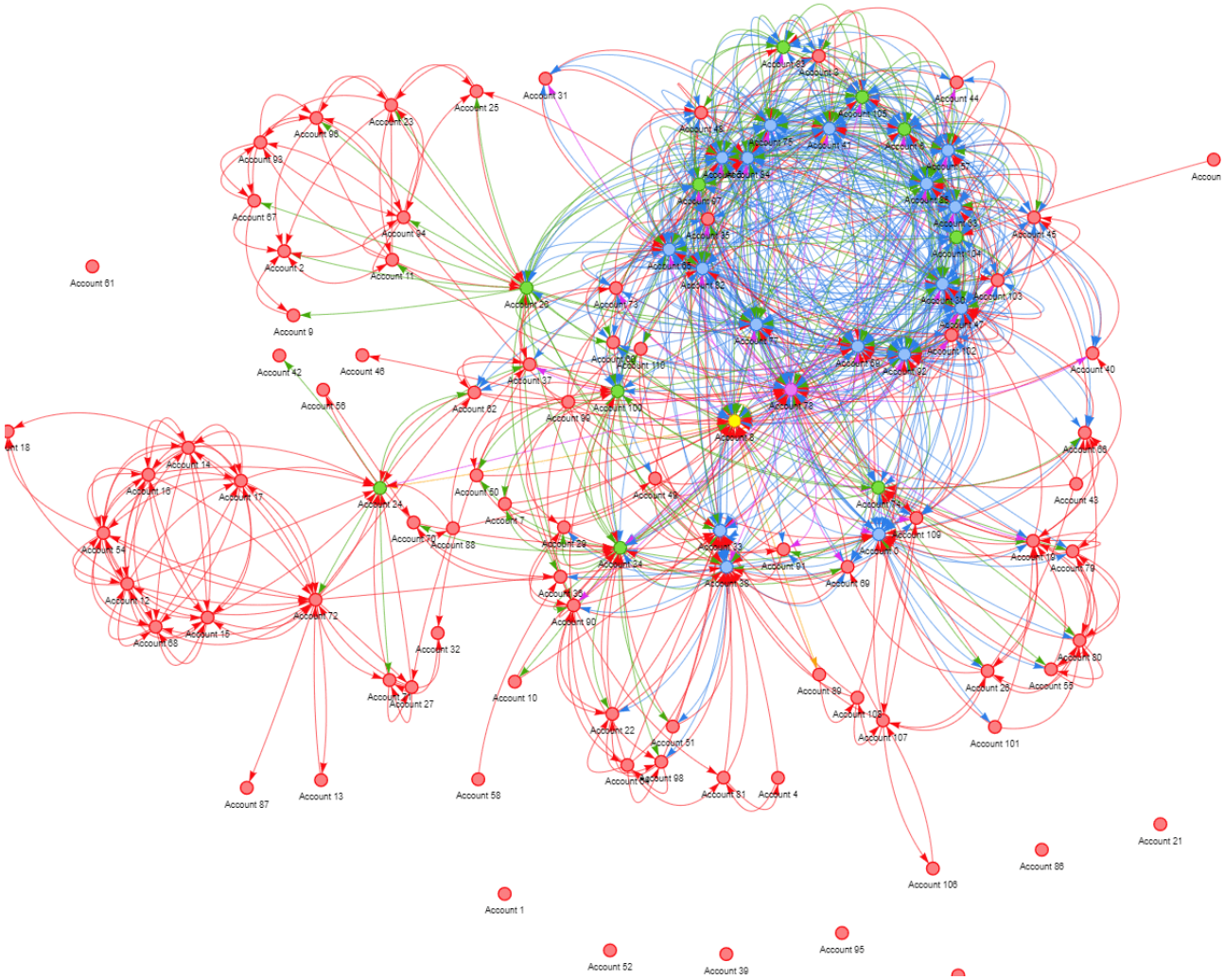

Among the 111 sexual assault centers on Twitter, the average count of both followers and followings was approximately 9. There is a notable variability in these counts, with a SD of 9.15 for followings and 9.85 for followers. The maximum number of followings observed was 38, whereas the maximum number of followers reached was 46, with a minimum count of 0.

We categorized them into different categories to account for the variability in the number of followers. The graph illustrates nodes of various colors, each representing a specific range of followers. Red denotes those with <11 followers, green indicates 11 to 21 followers, blue represents 21 to 31 followers, purple signifies 31 to 41 followers, and yellow indicates those with ≥41 followers. Notably, there is a single green node, which stands out with 46 followers. The in-degree centrality is 0.42, which is the highest value among all the nodes, underscoring their significance. In addition, they possess a closeness score of 0.49, ranking them within the top 1% among all nodes, implying their high level of proximity to other nodes and less dependency on others for information transmission. Furthermore, their betweenness score was 0.076, signifying their involvement in a substantial number of shortest paths and positioning them among the top 7 nodes in the betweenness ranking. We classify the sexual assault centers that meet the criteria of high closeness, in-degree centrality, low betweenness, and >41 followers and followings as *social queens*.

The figure also draws attention to a cluster of sexual assault centers characterized by a smaller number of followers and followings. We classified these nodes as marginalized entities. These centers have a limited impact on information transmission on Twitter, leading to their closeness and other metrics registering at 0.

Identification of Modular Patterns and Key Nodes in Twitter Network

Figure 11 is a section from Figure 10, in which distinct patterns emerge as certain nodes cluster together into modularities. As an example, we can examine a particular section of Figure 10 where a modularity forms—a small cluster consisting of several nodes that mutually follow each other. We incorporated eigenvector centrality to interpret this modularity. Within the bottom-right corner of this modularity, we find that accounts 24 and 72 (anonymous Twitter handle_names) exhibit relatively high values of eigenvector centrality. A higher eigenvector centrality score implies greater significance of the node when compared with its neighboring points. Furthermore, the importance of the node itself is directly linked to the significance of the neighboring nodes connected to it. Consequently, this

XSL•FO
RenderX

specific node can be regarded as the "social queen" within this particular modularity.

Figure 11. Identification of modular patterns and key nodes in Twitter network.

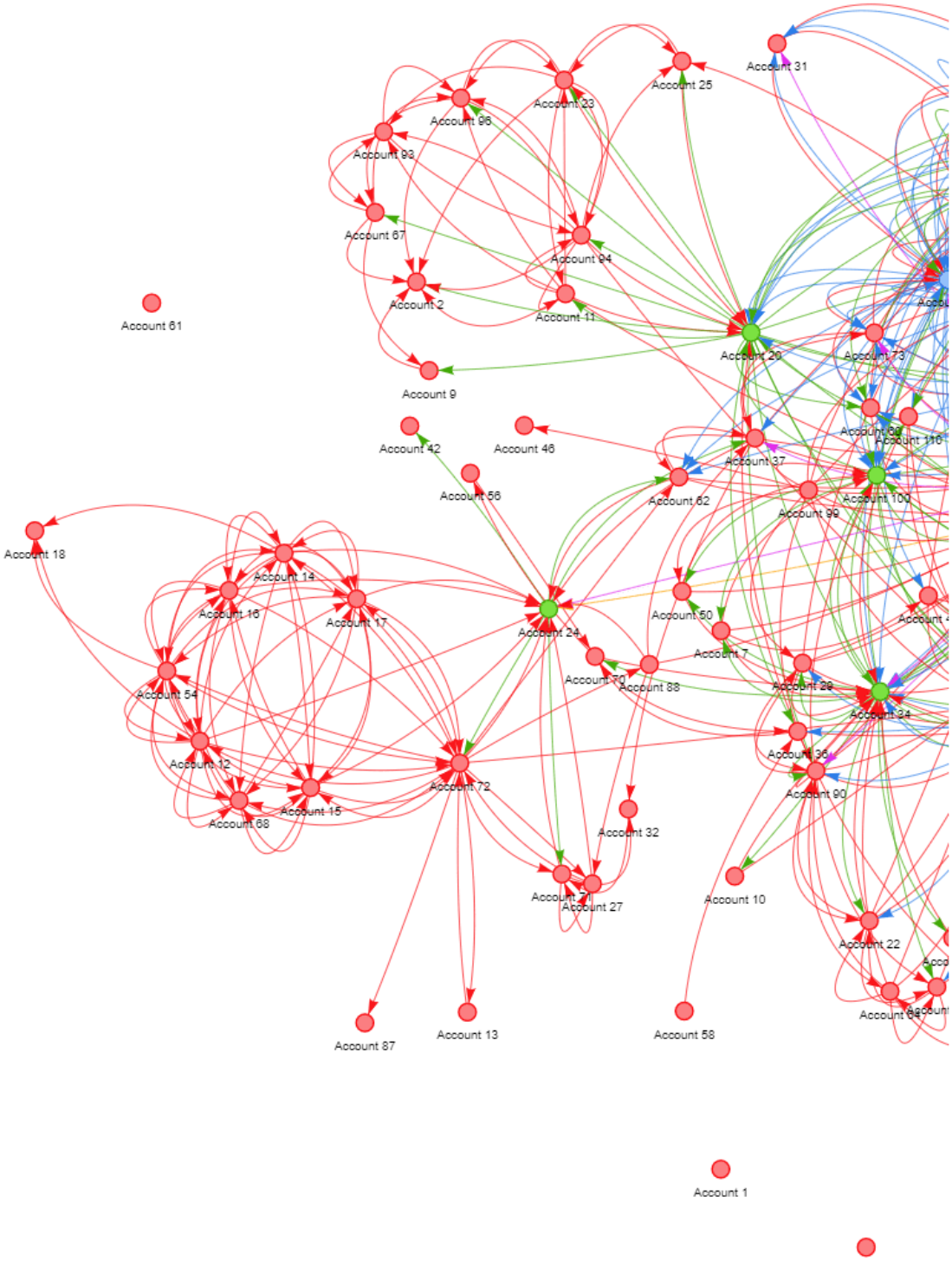

Discussion

Principal Findings

This study represents a pioneering effort to conduct a comprehensive analysis of Twitter's social network, user activities, and demographics within the context of sexual assault centers in Canada. By mapping and analyzing their Twitter practices, this research contributes to a better understanding of the social media landscape of sexual assault support organizations in Canada. The findings underscore the potential of Twitter as a platform for sexual assault organizations to build social capital, enhance their influence, and expand their reach. Moreover, it highlights the need for tailored engagement strategies that consider regional disparities and the unique

XSL•FO
RenderX

characteristics of each province and territory. Our findings align with the broader literature on social capital, specifically bridging and bonding social capital. Among the various social media platforms, Twitter emerges as a valuable data set to study sexual violence [46] as well as a notable facilitator of bridging social capital, consistent with previous research that has underscored Twitter's ability to connect organizations with a diverse and expansive audience [22].

Demographic Profile of Sexual Assault Centers on Twitter

The results of this study reveal a substantial presence of sexual assault centers in Canada on Twitter, signifying their acknowledgment of Twitter as a valuable communication and social capital development tool. Out of the 350 sampled centers, 124 (35.4%) maintain an active Twitter presence, highlighting the significant proportion of sexual assault organizations in Canada that recognize Twitter's efficacy as a communication medium for engaging with their stakeholders and the public. This trend aligns with the broader nonprofit sector, where most large and mid-sized nonprofits maintain at least 1 social media account [9]. However, it is worth noting that effective communication on Twitter may be hindered by a reliance on broadcasting rather than engaging in dialogue, as observed in previous research [47]. The success of acquiring social capital through social media appears to depend on the extent and quality of stakeholder connections, emphasizing the importance of diverse engagement strategies and the diversity and complexity of message elements [26]. In the context of sexual assault centers and support services, social media–based social capital holds significant potential for increasing the donor and volunteer base, engaging with the community on social issues, and promoting wider social change [33].

The geographic distribution of sexual assault centers with Twitter accounts highlights regional variations in Twitter use and engagement. Ontario, BC, and Quebec emerged as the provinces with the highest number of centers using Twitter, collectively accounting for two-thirds of all sampled centers in Canada, indicating higher levels of social capital in those regions owing to increased opportunities for information sharing, emotional kinship, trust, and social support [20]. The concentration of centers in these provinces aligns with their higher population densities and emphasizes the importance of social media platforms, such as Twitter, in reaching a broader audience consistently [43]. The southeast regions of Ontario and Quebec as well as the southern region of BC showed a higher concentration of sexual assault centers with Twitter accounts, likely reflecting the higher population densities in these areas. Furthermore, spatial distribution may influence the topics and issues addressed in their tweets. In BC, the distribution of sexual assault centers was also concentrated in specific regions, such as Vancouver. The content of tweets from centers in these regions may reflect local concerns and initiatives. It is essential to consider the unique characteristics and needs of each region when developing communication strategies and leveraging social media platforms for social capital enhancement.

The average age of sexual assault centers' Twitter accounts was calculated to determine the duration of their presence on the platform. The findings showed that the average duration varied across provinces and territories, ranging from 5 to 14 years. The Northwest Territories had the longest average duration of 12.77 years, indicating a relatively early adoption and subsequent use of Twitter among sexual assault centers in the territory. In contrast, provinces or territories such as Yukon and New Brunswick had a shorter average duration. These variations in account age reflect differences in the timing of adoption and highlight the diverse trajectories of Twitter use among sexual assault centers across Canada. These differences may be influenced by organizational factors, regional context, or resource availability. Centers with older Twitter accounts may have accumulated more followers and established stronger web-based communities, whereas newer accounts may need to focus on building and expanding their web-based presence.

User Activity of Sexual Assault Centers on Twitter

Patterns of account creation offer insights into the temporal dynamics of engagement and emphasize the need for continuous and consistent communication efforts. The recent decline in the creation of new Twitter accounts by sexual assault centers in recent years may signal a saturation point, where most centers have already established their Twitter presence. Alternatively, it could be attributed to factors such as resource constraints, changing organizational priorities, limited staff dedicated to communication practices, or a shift in focus to other social media platforms.

The examination of the average monthly tweet count for active sexual assault centers provides valuable insights into their Twitter activity levels. The results indicate variations in tweet frequency across provinces, with Ontario and BC consistently demonstrating higher tweet volumes compared with other provinces. This observation aligns with the higher number of active Twitter accounts and underscores the importance of ongoing engagement and dialogue with stakeholders through regular tweets.

Social Network Dynamics of Sexual Assault Centers in Canada

The findings related to social network dynamics reveal the landscape of Twitter engagement among sexual assault organizations. On average, these organizations have amassed approximately 1543 followers, demonstrating their capacity to reach a substantial audience. Simultaneously, they follow an average of 819 other accounts, indicating their active involvement within the Twitter community. This indicates the potential for these organizations to disseminate information, provide support, and raise awareness about their critical missions. This study aligns with the perspective emphasizing the importance of bridging social capital facilitated by Twitter's ability to connect with various stakeholders, including service recipients, donors, and the general public [26]. It highlights the potential and disparities in bridging social capital among sexual assault centers across provinces and territories in Canada, suggesting that these organizations can better leverage Twitter to establish connections beyond their immediate constituencies. This aligns with the notion that social media platforms such as

XSL•FO
RenderX

Twitter can extend an organization's reach and promote the flow of information across various stakeholders [22].

The findings also uncovered regional disparities in Twitter engagement among sexual assault organizations in Canada. Sexual assault organizations across Canada's provinces exhibited varying degrees of Twitter activity. Although some provinces, such as Ontario, displayed robust engagement, others, such as New Brunswick, had limited presence and following. Our observation resonates with previous studies (eg, [9]) that have emphasized the role of regional context in shaping nonprofit organizations' social media use. These regional disparities suggest the need for tailored strategies to maximize the impact of Twitter engagement, considering the unique characteristics and needs of each province.

Within the context of nonprofit organizations, research has indicated a positive relationship between follower count and bonding social capital [21]. Our study aligns with this perspective by demonstrating a positive association between followers and followings. As follower counts increase, there is a corresponding increase in followings, indicating a proactive approach by organizations to engage with their audience. This observation underscores the importance of reciprocity and interaction on Twitter. This suggests that a larger number of followers on Twitter can contribute to increased financial support, in line with the positive impact of web-based social capital generated through social networking sites on charitable outcomes [29,30].

Within the intricate network of sexual assault centers on Twitter, we identified nodes with distinct characteristics. Some organizations emerged as "social queens," characterized by high in-degree centrality, closeness scores, and low betweenness, coupled with substantial followers and followings. These "social queens" play pivotal roles in information transmission, networking, and community building. This finding suggests that organizations can strategically use Twitter to enhance their influence and reach within their fields of operation. However, this study also highlights the presence of marginalized entities with limited follower counts, which may face challenges in impacting information transmission on Twitter. This underscores the importance of proactive engagement strategies for organizations seeking to maximize their impact through social media. Sexual assault organizations can benefit from a comprehensive understanding of their social network structures, enabling them to identify opportunities to strengthen their social capital, expand their donor base, and effectively engage the community.

Limitations

This study had some limitations. First, the findings are specific to sexual assault centers in Canada and may not be applicable to other countries or regions owing to cultural, social, and organizational differences. Each country or region may have unique characteristics that influence the use of Twitter and other social media platforms by sexual assault organizations. Second, the study focuses solely on Twitter data and does not consider the use of other platforms such as Facebook, Instagram, or Snapchat, which could provide additional insights into communication strategies and social media practices. Therefore, the findings of this study may not provide a comprehensive understanding of the organizations' overall social media use. Third, the study is cross-sectional, providing a snapshot of Twitter use at a specific time, and does not capture longitudinal changes or trends. A longitudinal study would offer more detailed insights into the evolution of Twitter practices and the effectiveness of communication strategies used by these organizations. For example, we lack information about growth and changes in the number of followers over time. The follower counts remained static at the time of data collection. Future studies may need to explore how follower counts evolve dynamically to gain deeper insights. Fourth, the study primarily focused on describing Twitter use at the organizational level rather than evaluating the effectiveness or outcomes of the communication strategies used. Further research is required to assess the impact and outcomes of social media use in this context. Fifth, this study did not have access to demographic information related to the organizational size of the nonprofits, which typically includes factors such as the number of employees, volunteers, or annual budget. Unfortunately, Twitter does not provide access to such data, resulting in its absence from this study. Finally, the study does not delve deeper into the content of tweets posted by sexual assault organizations on Twitter. Future studies could explore the thematic analysis of tweets, sentiment analysis to understand the emotional tone of their messages, and the effectiveness of specific content strategies used by these organizations to engage their audience and advocate for their cause.

Conclusions

In conclusion, this study provides valuable insights into the current use and social structure of Twitter by sexual assault centers, crisis lines, and support services in Canada. The findings highlight the widespread adoption of Twitter among these organizations and the potential for leveraging social media platforms to build social capital. By recognizing regional disparities, identifying key players, and understanding the dynamics of followers and followings, sexual assault organizations can better navigate the Twitter landscape to further their missions of promoting awareness and support for survivors of sexual assault. Further research in this area can explore the long-term impact of social media use on organizational outcomes and stakeholder perceptions into enhancing social capital within the nonprofit sector and beyond, providing additional guidance for effective communication practices in the nonprofit human services sector and ultimately contributing to the broader goals of these organizations.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Sexual assault centers with official Twitter accounts in Alberta, New Brunswick, Newfoundland and Labrador, Nova Scotia, Manitoba, Saskatchewan, Northwest Territories, and Yukon.
[DOCX File , 654 KB - jmir_v26i1e50552_app1.docx]

Multimedia Appendix 2
Followers, followings, and measurement of social network analysis.
[DOCX File , 35 KB - jmir_v26i1e50552_app2.docx]

References

1. Guo C, Saxton GD. Getting attention: an organizational-level analysis. In: The Quest for Attention. Redwood City, CA: Stanford University Press; 2020.
2. Guo C, Saxton GD. Speaking and being heard: how nonprofit advocacy organizations gain attention on social media. Nonprofit Volunt Sect Q 2018;47(1):5-26. [doi: 10.1177/0899764017713724]
3. Bhati A, McDonnell D. Success in an online giving day: the role of social media in fundraising. Nonprofit Volunt Sect Q 2020;49(1):74-92. [doi: 10.1177/0899764019868849]
4. Campbell DA, Lambright KT. Are you out there? internet presence of nonprofit human service organizations. Nonprofit Volunt Sect Q 2019;48(6):1296-1311. [doi: 10.1177/0899764019852673]
5. Jung K, Valero JN. Assessing the evolutionary structure of homeless network: social media use, keywords, and influential stakeholders. Technol Forecast Soc Change 2016 Sep;110:51-60. [doi: 10.1016/j.techfore.2015.07.015]
6. Maxwell SP, Carboni JL. Social media management. Nonprofit Manag Leadersh 2016 Sep 19;27(2):251-260. [doi: 10.1002/nml.21232]
7. Zhou H, Ye S. Legitimacy, worthiness, and social network: an empirical study of the key factors influencing crowdfunding outcomes for nonprofit projects. Voluntas 2018 Jun 4;30:849-864. [doi: 10.1007/s11266-018-0004-0]
8. Zhou H, Ye S. Fundraising in the digital era: legitimacy, social network, and political ties matter in China. Voluntas 2019 Apr 01;32:498-511. [doi: 10.1007/s11266-019-00112-9]
9. Nah S, Saxton GD. Modeling the adoption and use of social media by nonprofit organizations. New Media Soc 2013;15(2):294-313. [doi: 10.1177/1461444812452411]
10. Ellison NB, Vitak J, Gray R, Lampe C. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. J Comput Mediat Commun 2014 Jul 01;19(4):855-870. [doi: 10.1111/jcc4.12078]
11. Li C, Bernoff J. Groundswell: Winning in a World Transformed by Social Technologies. Brighton, MA: Harvard Business Review Press; 2011.
12. Park H, Rodgers S, Stemmle J. Analyzing health organizations' use of Twitter for promoting health literacy. J Health Commun 2013;18(4):410-425. [doi: 10.1080/10810730.2012.727956] [Medline: 23294265]
13. Lovejoy K, Waters RD, Saxton GD. Engaging stakeholders through Twitter: how nonprofit organizations are getting more out of 140 characters or less. Public Relat Rev 2012 Jun;38(2):313-318. [doi: 10.1016/j.pubrev.2012.01.005]
14. Xue J, Chen J, Chen C, Hu R, Zhu T. The hidden pandemic of family violence during COVID-19: unsupervised learning of tweets. J Med Internet Res 2020 Nov 06;22(11):e24361 [FREE Full text] [doi: 10.2196/24361] [Medline: 33108315]
15. Lovejoy K, Saxton GD. Information, community, and action: how nonprofit organizations use social media. J Comput Mediat Commun 2012 Apr 01;17(3):337-353. [doi: 10.1111/j.1083-6101.2012.01576.x]
16. Xue J, Macropol K, Jia Y, Zhu T, Gelles RJ. Harnessing big data for social justice: an exploration of violence against women‐related conversations on Twitter. Human Behav Emerg Technol 2019 Jul 26;1(3):269-279. [doi: 10.1002/hbe2.160]
17. Guo C, Saxton GD. Tweeting social change: how social media are changing nonprofit advocacy. Nonprofit Volunt Sect Q 2014;43(1):57-79. [doi: 10.1177/0899764012471585]
18. Kapucu N, Demiroz F. A social network analysis approach to strengthening nonprofit collaboration. J Appl Manag Entrep 2015 Jan;20(1):87-101. [doi: 10.9774/GLEAF.3709.2015.ja.00007]
19. Putnam RD. Tuning in, tuning out: the strange disappearance of social capital in America. PS Political Sci Politics 1995;28(4):664-683 [FREE Full text] [doi: 10.2307/420517]
20. Putnam RD. Bowling Alone: The Collapse and Revival of American Community. New York, NY: Simon & Schuster; 2001.
21. Hofer M, Aubert V. Perceived bridging and bonding social capital on Twitter: differentiating between followers and followees. Comput Hum Behav 2013 Nov;29(6):2134-2142. [doi: 10.1016/j.chb.2013.04.038]
22. Phua J, Jin SV, Kim J. Uses and gratifications of social networking sites for bridging and bonding social capital: a comparison of Facebook, Twitter, Instagram, and Snapchat. Comput Hum Behav 2017 Jul;72:115-122. [doi: 10.1016/j.chb.2017.02.041]
23. Bourdieu P. Distinction: a social critique of the judgment of taste. In: Food and Culture, 4th Edition. Milton Park, UK: Routledge; 2017.
24. Lin N. Building a network theory of social capital. In: Social Capital, 1st Edition. Milton Park, UK: Routledge; 2001.

XSL•FO
RenderX

25. Lai CH, Fu JS. Humanitarian relief and development organizations' stakeholder targeting communication on social media and beyond. Voluntas 2020 Mar 05;32:120-135. [doi: 10.1007/s11266-020-00209-6]
26. Xu W, Saxton GD. Does stakeholder engagement pay off on social media? a social capital perspective. Nonprofit Volunt Sect Q 2018 Aug 02;48(1):28-49 [FREE Full text] [doi: 10.1177/0899764018791267]
27. Doerfel ML, Atouba Y, Harris JL. (Un)obtrusive control in emergent networks: examining funding agencies' control over nonprofit networks. Nonprofit Volunt Sect Q 2017;46(3):469-487. [doi: 10.1177/0899764016664588]
28. Williams JR. The use of online social networking sites to nurture and cultivate bonding social capital: a systematic review of the literature from 1997 to 2018. New Media Soc 2019;21(11-12):2710-2729. [doi: 10.1177/1461444819858749]
29. Henry R, Bosman L. Strategic management and social media: an empirical analysis of electronic social capital and online fundraising. In: Olivas-Luján MR, Bondarouk T, editors. Social Media in Strategic Management. Bingley, UK: Emerald Group Publishing Limited; 2013:43-62.
30. Lee YJ, Shon J. Nonprofits' online social capital and charitable support. J Nonprofit Public Sect Mark 2023;35(3):290-307. [doi: 10.1080/10495142.2021.1982112]
31. Saxton GD, Wang L. The social network effect: the determinants of giving through social media. Nonprofit Volunt Sect Q 2013 Apr 24;43(5):850-868. [doi: 10.1177/0899764013485159]
32. Seo H, Vu HT. Transnational nonprofits' social media use: a survey of communications professionals and an analysis of organizational characteristics. Nonprofit Volunt Sect Q 2020 Feb 28;49(4):849-870. [doi: 10.1177/0899764020908340]
33. Svensson PG, Mahoney TQ, Hambrick ME. Twitter as a communication tool for nonprofits: a study of sport-for-development organizations. Nonprofit Volunt Sect Q 2014 Oct 16;44(6):1086-1106. [doi: 10.1177/0899764014553639]
34. Yang A, Liu W. Coalition networks for the green new deal: nonprofit public policy advocacy in the age of social media. Nonprofit Volunt Sect Q 2022 Sep 20;52(5):1284-1307. [doi: 10.1177/08997640221123341]
35. Makice K. Twitter API: Up and Running. Sebastopol, CA: O'Reilly Media, Inc; Mar 2009.
36. Amr T, Stamboliyska R. Practical D3.js. New York, NY: Apress; 2016.
37. Interactive network visualizations. Pyvis. URL: https://pyvis.readthedocs.io/en/latest/ [accessed 2024-03-01]
38. Conducting a social network analysis. Converge. URL: https://www.converge.net/toolkit/conducting-a-social-network-analysis [accessed 2024-03-01]
39. Bonacich P. Some unique properties of eigenvector centrality. Soc Netw 2007 Oct;29(4):555-564. [doi: 10.1016/j.socnet.2007.04.002]
40. McCurdie T, Sanderson P, Aitken LM. Applying social network analysis to the examination of interruptions in healthcare. Appl Ergon 2018 Feb;67:50-60. [doi: 10.1016/j.apergo.2017.08.014] [Medline: 29122200]
41. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 2004 Feb;69(2 Pt 2):026113. [doi: 10.1103/PhysRevE.69.026113] [Medline: 14995526]
42. Closeness centrality (centrality measure). GeeksforGeeks. URL: https://www.geeksforgeeks.org/closeness-centrality-centrality-measure/ [accessed 2024-03-01]
43. Canada Ontario Density 2016.png. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:Canada_Ontario_Density_2016.png#file [accessed 2024-03-01]
44. McPhee A. Canada British Columbia Density 2016.png. Wikimedia Commons. 2019. URL: https://commons.wikimedia.org/wiki/File:Canada_British_Columbia_Density_2016.png
45. McPhee A. Canada Quebec Density 2016.png. Wikimedia Commons. 2019. URL: https://commons.wikimedia.org/wiki/File:Canada_Quebec_Density_2016.png
46. Xue J, Zhang B, Zhang Q, Hu R, Jiang J, Liu N, et al. Using Twitter-based data for sexual violence research: scoping review. J Med Internet Res 2023 May 15;25:e46084 [FREE Full text] [doi: 10.2196/46084] [Medline: 37184899]
47. Stabile B, Grant A, Purohit H, Sharan Bonala S. Take back the tweet: social media use by anti‐gender‐based violence organizations. Sex Gend Policy 2021 Apr 29;4:38-56. [doi: 10.1002/sgp2.12029]

Abbreviations

API: application programing interface
BC: British Columbia
RQ: research question

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.07.23; peer-reviewed by W Ceron; comments to author 04.12.23; revised version received 23.12.23; accepted 13.02.24; published 27.03.24.

Please cite as:
Xue J, Zhang Q, Zhang Y, Shi H, Zheng C, Fan J, Zhang L, Chen C, Li L, Shier ML
Bridging and Bonding Social Capital by Analyzing the Demographics, User Activities, and Social Network Dynamics of Sexual Assault Centers on Twitter: Mixed Methods Study
J Med Internet Res 2024;26:e50552
URL: https://www.jmir.org/2024/1/e50552
doi:10.2196/50552
PMID:38536222

©Jia Xue, Qiaoru Zhang, Yun Zhang, Hong Shi, Chengda Zheng, Jingchuan Fan, Linxiao Zhang, Chen Chen, Luye Li, Micheal L Shier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Medical Misinformation in Polish on the World Wide Web During the COVID-19 Pandemic Period: Infodemiology Study

Małgorzata Chlabicz[1,2], PhD; Aleksandra Nabożny[3*], PhD; Jolanta Koszelew[4], PhD; Wojciech Łaguna[5], MSc; Anna Szpakowicz[6], PhD; Paweł Sowa[1], PhD; Wojciech Budny[7], PhD; Katarzyna Guziejko[8], PhD; Magdalena Róg-Makal[2], PhD; Sławomir Pancewicz[9], Prof Dr Med; Maciej Kondrusik[9], PhD; Piotr Czupryna[9], PhD; Beata Cudowska[10], PhD; Dariusz Lebensztejn[10], Prof Dr Med; Anna Moniuszko-Malinowska[9], Prof Dr Med; Adam Wierzbicki[11*], Prof Dr; Karol A Kamiński[1], Prof Dr Med

[1]Department of Population Medicine and Lifestyle Diseases Prevention, Medical University of Białystok, Białystok, Poland
[2]Department of Invasive Cardiology, Medical University of Białystok, Białystok, Poland
[3]Department of Software Engineering, Gdańsk University of Technology, Gdańsk, Poland
[4]R&D Department, Science4People Limited Liability Company, Szczecin, Poland
[5]Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland
[6]Department of Cardiology, Medical University of Bialystok, Białystok, Poland
[7]Department of Allergology and Internal Medicine, Medical University of Białystok, Białystok, Poland
[8]2nd Department of Lung Diseases and Tuberculosis, Medical University of Białystok, Białystok, Poland
[9]Department of Infectious Diseases and Neuroinfection, Medical University of Białystok, Białystok, Poland
[10]Department of Pediatrics, Gastroenterology, Hepatology, Nutrition, Allergology and Pulmonology, Medical University of Bialystok, Białystok, Poland
[11]Department of Computer Science, Polish-Japaneese Academy of Information Technology, Warsaw, Poland
[*]these authors contributed equally

Corresponding Author:
Aleksandra Nabożny, PhD
Department of Software Engineering
Gdańsk University of Technology
ul. Gabriela Narutowicza 11/12
Gdańsk, 80-233
Poland
Phone: 48 58 348 67 00
Email: alenaboz@pg.edu.pl

Abstract

Background: Although researchers extensively study the rapid generation and spread of misinformation about the novel coronavirus during the pandemic, numerous other health-related topics are contaminating the internet with misinformation that have not received as much attention.

Objective: This study aims to gauge the reach of the most popular medical content on the World Wide Web, extending beyond the confines of the pandemic. We conducted evaluations of subject matter and credibility for the years 2021 and 2022, following the principles of evidence-based medicine with assessments performed by experienced clinicians.

Methods: We used 274 keywords to conduct web page searches through the BuzzSumo Enterprise Application. These keywords were chosen based on medical topics derived from surveys administered to medical practitioners. The search parameters were confined to 2 distinct date ranges: (1) January 1, 2021, to December 31, 2021; (2) January 1, 2022, to December 31, 2022. Our searches were specifically limited to web pages in the Polish language and filtered by the specified date ranges. The analysis encompassed 161 web pages retrieved in 2021 and 105 retrieved in 2022. Each web page underwent scrutiny by a seasoned doctor to assess its credibility, aligning with evidence-based medicine standards. Furthermore, we gathered data on social media engagements associated with the web pages, considering platforms such as Facebook, Pinterest, Reddit, and Twitter.

Results: In 2022, the prevalence of unreliable information related to COVID-19 saw a noteworthy decline compared to 2021. Specifically, the percentage of noncredible web pages discussing COVID-19 and general vaccinations decreased from 57% (43/76) to 24% (6/25) and 42% (10/25) to 30% (3/10), respectively. However, during the same period, there was a considerable

XSL•FO
RenderX

uptick in the dissemination of untrustworthy content on social media pertaining to other medical topics. The percentage of noncredible web pages covering cholesterol, statins, and cardiology rose from 11% (3/28) to 26% (9/35) and from 18% (5/28) to 26% (6/23), respectively.

Conclusions: Efforts undertaken during the COVID-19 pandemic to curb the dissemination of misinformation seem to have yielded positive results. Nevertheless, our analysis suggests that these interventions need to be consistently implemented across both established and emerging medical subjects. It appears that as interest in the pandemic waned, other topics gained prominence, essentially "filling the vacuum" and necessitating ongoing measures to address misinformation across a broader spectrum of health-related subjects.

(J Med Internet Res 2024;26:e48130) doi:10.2196/48130

KEYWORDS

infodemic; fake news; information credibility; online health information; evidence based medicine; EBM; false; credibility; credible; health information; online information; information quality; infoveillance; infodemiology; misinformation; disinformation

Introduction

Background

Although disinformation has been recognized since the early days of the printing press in previous centuries, its significance has grown substantially in the era of the internet. Technology historian Melvin Kranzberg aptly remarked, "Technology is neither inherently good nor bad; nor is it neutral." [1]. Similarly, the internet, with its free access to information through the World Wide Web (WWW), not only offers numerous advantages but also facilitates the rapid spread of noncredible or fake stories. Disinformation poses a threat to society across various aspects of life, with particular significance in the realm of online health information. The dissemination of false health-related information can significantly impact individuals' well-being and potentially result in long-term consequences for their health [2]. This study aims to identify medicine-related topics that contribute to the proliferation of unreliable content on the web, such as those related to the COVID-19 pandemic, vaccination, statins, and diet. Additionally, we sought to observe the dynamics of these topics over a 2-year period (2021-2022).

The years 2020, 2021, and 2022 were largely overshadowed by the SARS-CoV-2 pandemic in the realm of medicine-related online health information. Consequently, it is unsurprising that investigations into the dissemination of medical disinformation during this period predominantly centered around topics associated with COVID-19. Furthermore, the research placed a strong emphasis on the perspective of the general public, rather than that of medical practitioners who grappled with the consequences of the widespread dissemination of disinformation. Our study endeavored to address 2 key challenges: first, to comprehend online health information beyond the scope of the pandemic, and second, to investigate noncredible online health information through the lens of medical practitioners.

The inception of our study involved gathering surveys from doctors, outlining the topics they encountered in their clinical practice. Consequently, we obtained information about false content that significantly influenced patients' attitudes toward medical recommendations. Additionally, we collected data on disinformation concerning the COVID-19 pandemic and 6 other medical topics, presenting a comprehensive overview of the overall dynamics of the spread of medical online disinformation. Our approach involved a systematic search for the most widely read articles and social media engagements pertaining to selected topics in 2021 and 2022. Subsequently, a team of experienced clinicians assessed the credibility of these sources, adhering to the principles of evidence-based medicine. This methodology enabled us to pinpoint crucial areas for enhancing interventions aimed at curbing the dissemination of noncredible online health information.

Literature Review

Although various definitions of information credibility [3] and noncredible information, including disinformation or misinformation, exist [4], there remains a lack of a clear definition for false information. In some instances, the literature uses the term "medical fake news" [5], which is defined as "news that is intentionally false and could mislead readers" and "information discrepant with medical knowledge" [6]. The primary objective of such misinformation is to manipulate and shape public opinion. Numerous publications highlight the adverse repercussions of disinformation spreading on the internet within communities [7,8]. An example is the study conducted by Betsch et al [9], which revealed that spending just 5-10 minutes on vaccine-critical websites heightened the perception of vaccine risk. This, in turn, led to negative perceptions of vaccination risks and a decreased intention to vaccinate.

Evidence-based medicine furnishes factual information that is not only relevant to everyday medical practice but also invaluable for experts endeavoring to refute medical fake news. Coined in 1991 [10], evidence-based medicine is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [11]. Evidence-based medicine aims to integrate the clinician's experience, the patient's values, and the best available scientific information to inform clinical management decisions. The principles of evidence-based medicine apply not only to individual medical practice but also to institutions and the health care system at large.

The novel coronavirus, known as SARS-CoV-2, gave rise to COVID-19. Beyond causing the COVID-19 pandemic, it served as a catalyst for extensive medical disinformation campaigns [12-14]. The internet became inundated with millions of posts about COVID-19 [15]. The challenges associated with developing accurate diagnostic tests, therapeutic protocols, and

XSL•FO
RenderX

methods for preventing SARS-CoV-2 infections have created fertile ground for manipulation and disinformation, particularly in health decision-making [13-15]. False information about the pandemic proliferated through mass media, social platforms, and text messaging applications, leading to various adverse effects such as (1) reducing patients' willingness to vaccinate, (2) obstructing measures to contain disease outbreaks, and (3) instigating the physical interruption of access to health care, among others [16].

The proliferation of fake news about COVID-19 has significantly contributed to the dissemination of disinformation in related fields, notably in the context of vaccination, putting scientific knowledge to the test [17,18]. While the antivaccination movement had already propagated conspiracy theories, such as the notion that large pharmaceutical companies and other vested interests exaggerate vaccination benefits while concealing risks or dangers [19], the pandemic has heightened the phenomenon of skepticism toward immunization. Vaccine hesitancy was previously noted in the cases of well-known vaccinations such as the MMR (measles, mumps, and rubella) vaccine [20] and the flu vaccine [21]. Research has indicated that attitudes and conspiracy beliefs, which can be viewed as a particular set of behaviors, are correlated with the willingness to vaccinate. For instance, Dubé et al [22] demonstrated that parents who chose to vaccinate their children were more likely to receive information about vaccines from health care professionals and less inclined to seek information on the internet. Additionally, Sommariva et al [23] found that disinformation circulating on social networks could impede disease prevention efforts. Regarding COVID-19 vaccines, new conspiracy theories have surfaced, such as microchip injection, unchecked adverse effects, absence of safety assessment before vaccine distribution, and unproven vaccine effectiveness. Studies indicate that exposure to fake news diminishes the intention to receive the COVID-19 vaccine [24]. Even brief exposure to false information has been shown to potentially modify or influence human behavior [22]. As a result, conspiracy beliefs, trust in information, and brief exposure to online news can impact individuals' vaccination choices [24]. Findings from a randomized controlled trial by Loomba et al [25] indicated that exposure to online disinformation about COVID-19 vaccines led to a significant decline in vaccination intent, in the samples from both the United States (6.4% points) and the United Kingdom (6.2% points).

The issue of disinformation in cardiology, particularly regarding statins, has been previously documented [26,27]. Although overshadowed during the COVID-19 pandemic, it is resurfacing. Waszak et al [28] conducted an analysis of topics drawing public attention between 2012 and 2017. The primary subjects were cancer, vaccination, heart attack, stroke, HIV/AIDS, hypertension, and diabetes. Vaccination emerged as the topic most affected by fake news, with a staggering 90%, followed by hypertension and HIV/AIDS, both at 70%.

The literature has delved into the analysis of medical information, particularly those related to nutrition, that has been deemed false or misleading [7,29]. While discussions regarding the impact of disinformation and misinformation on knowledge levels and nutritional decision-making are commonplace, there is a relatively infrequent analysis of the extent of such content on the internet [30]. Several efforts have been made to quantify the phenomenon of nutrition disinformation during the pandemic period. For instance, a Turkish study estimated that approximately 20% of analyzed YouTube videos on nutrition topics contained fake news [31], while a Korean study reported as much as 37% [32]. From a qualitative perspective, nutrition misinformation during the COVID-19 pandemic era encompassed topics related to food immunity boosters, purportedly designed to reduce the risk or severity of infection. Studies have explored the impact of supplementation with zinc, vitamin C, vitamin A, herbs, or nutraceuticals on mitigating the mentioned risks [33,34]. An Indian study observed that over 70% of respondents reflected immunity-boosting food in their dietary decisions [8]. The pandemic led to a surge in the circulation of fake news concerning the supposed protective effects of alcohol in the context of SARS-CoV-2 infection, including the unfounded hypothesis that strong alcohol can kill the virus. Numerous articles perpetuated the notion that light alcohol such as beer or wine could stimulate the immune system to combat the coronavirus [7,35,36]. This misinformation had severe consequences, as evidenced by the 60 people who lost their sight and nearly 6000 individuals hospitalized after consuming methanol as a "preventive" measure [7,36]. While a study from Hong Kong suggested that around 19% of respondents encountered content suggesting a beneficial link between alcohol consumption and the prevention of SARS-CoV-2 infection [37], there is a scarcity of scientific reports estimating the overall prevalence of fake news on this subject. It is worth noting that for the topic of dietary interventions, the term "misinformation" is used instead of "disinformation." This is because some of the recommendations might have been well-intentioned but were ineffective or misguided self-help strategies. Hence, such information does not fall into the category of intentionally misleading (disinformation).

In summary, while there are numerous reports highlighting the harmful effects of medical disinformation and misinformation on public health, there remains a dearth of studies focusing on the dynamics of the dissemination of unreliable online health information.

Methods

Preliminary Questionnaires

The initial phase of data collection involved identifying areas where topics were likely to generate unreliable content. We began by pinpointing myths and false beliefs commonly propagated by patients in doctors' offices and hospital wards, recognizing that these beliefs realistically impede doctors from effectively treating patients. To facilitate identification, a questionnaire was developed and administered for this purpose (Multimedia Appendix 1). The questionnaires were gathered at the Clinical Hospital of the Medical University of Bialystok in Poland and through an online survey. The online survey was made possible through collaborations with institutions that have signed letters of intent with the Medical University of Bialystok, including the College of Family Physicians and the Polish

XSL•FO
RenderX

Society of Cardiology. Data collection spanned from November 29, 2021, to December 26, 2021 (refer to Multimedia Appendix 1). The questionnaire encompassed inquiries about the physician's age, gender, specialization, length of service in the profession, and place of work. Additionally, an open-ended question was included: "What false medical information do you encounter in your professional practice when talking to patients?," with a total of 121 physicians responding. The average age of the doctors was 41.1 (SD 11.4) years, with women constituting 64.5% (78/121). The average duration of employment in the profession was 15.2 (SD 11.4) years. The physicians had various specialties, including internal medicine (n=41), cardiology (n=34), pediatrics (n=21), family medicine (n=18), surgery (n=8), hematology (n=6), allergology (n=6), ophthalmology (n=4), rheumatology (n=1), pulmonology (n=1), emergency medicine (n=1), and doctors in training (n=3). Some physicians held multiple medical specialties. In terms of workplace distribution, the majority of doctors worked exclusively in hospitals (n=70), while others worked in a combination of hospitals, public, and private clinics (n=8); hospitals and public clinics (n=8); or hospitals and private clinics (n=19). Additionally, some respondents were exclusively employed in public clinics (n=11) or private clinics (n=5). The physicians provided descriptions of all the topics they had encountered in their professional practice and quoted specific questions posed by patients. Through the analysis of the questionnaires, we identified common problems and subsequently categorized them into 9 groups (topics). These topics were delineated based on their consistent scope and association with false medical information.

Web Page and Social Media Engagements Retrieval Using BuzzSumo

Subsequently, queries were formulated based on the identified topics (problem groups). These topics were then assigned to each member of another team of physicians, who conducted searches using the DuckDuckGo search engine. The queries comprised single words or several words forming a specific meaningful phrase. Physicians considered only queries that resulted in unreliable websites. In total, 274 queries were collected through this method.

Subsequently, the BuzzSumo Enterprise Application (BuzzSumo Limited) was used [38]. BuzzSumo, a social media analytics tool, searches the internet for web documents based on queries and generates engagement reports, encompassing comments, likes, and shares on social media. This tool collects data from various platforms, including Facebook, Pinterest, Reddit, and Twitter, to compile a list of article web pages with the highest online engagement. *Engagement*, in this context, is defined as the cumulative number of interactions users have with a specific article link, encompassing actions such as "liking," "sharing," and "commenting." Upon entering a query, the application provides statistics for the linked pages. The subsequent steps of the web search process are illustrated in Figure 1.

Figure 1. Four subsequent stages of web search research. The figure includes a step description, the product of the step, and an example.

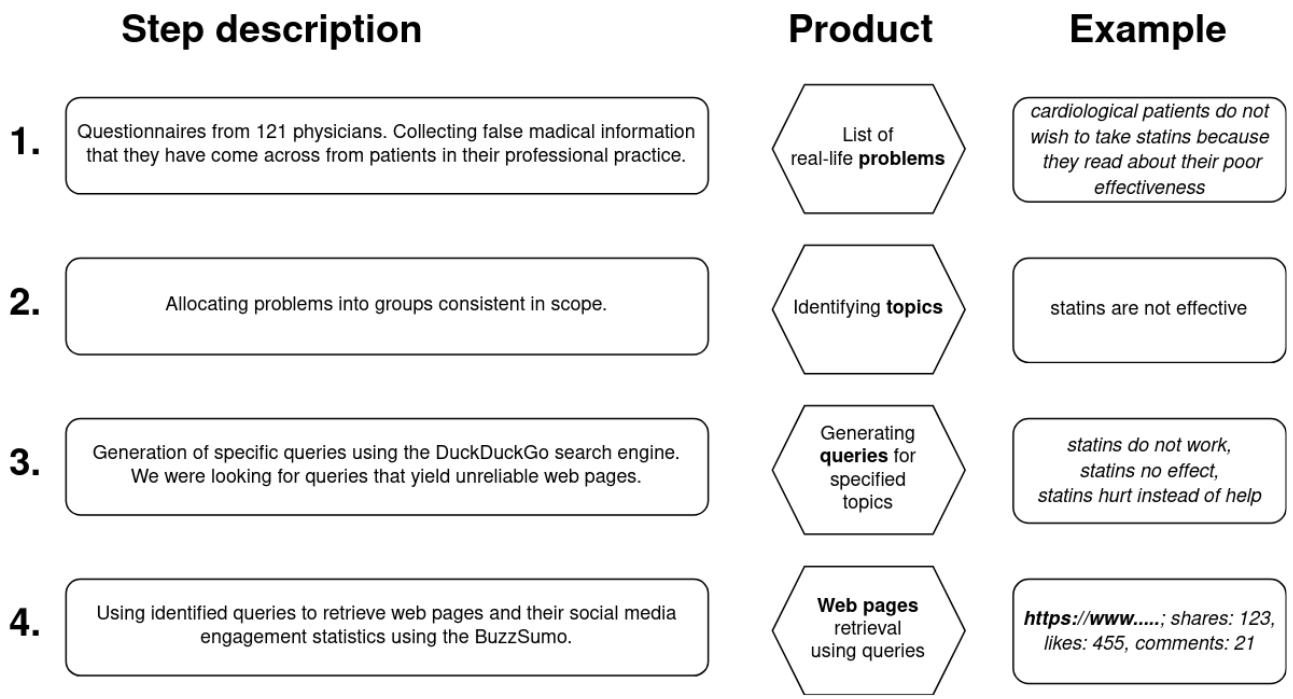

The search was restricted to 2 specific periods: (1) January 1, 2021, to December 31, 2021; (2) January 1, 2022, to December 31, 2022. Searches were filtered by time and language, specifically Polish. The analysis occurred in 2 consecutive phases. The first phase covered the entirety of 2021 and took place in January 2022. The second phase occurred in January 2023 and encompassed the entirety of 2022. This approach allowed us to evaluate current websites, recognizing that the online environment is dynamic, and web pages are frequently deactivated or cease to generate traffic. This observation is further supported by our findings. Even though identical queries were used to search websites during both periods, there was no repetition of websites selected in 2021 during the search in 2022. This lack of overlap underscores the dynamic nature of the

XSL•FO
RenderX

online environment, where web pages frequently change or cease to be relevant over time.

For inclusion in the analysis, only pages that garnered at least ten social media engagements and contained text were considered; videos were not evaluated. Subsequently, the 10 most popular pages per query were analyzed, with "popular" referring to those with the highest total engagements. Each web page underwent review by an experienced doctor, holding at least five years of professional experience, who is also a scientist at the Medical University of Bialystok. To facilitate clarity in assessing the web pages, categories were introduced based on their content: medical, nonmedical, and difficult to evaluate. Of the 274 queries, only 58 yielded medical content in 2021, and 32 in 2022, as illustrated in Figure 2.

Figure 2. The figure contains a detailed description of the subsequent stages of the web search research, with the year 2021 as an example. Topic identification, query generation, web page retrieval, and the subsequent web page filtration are depicted.

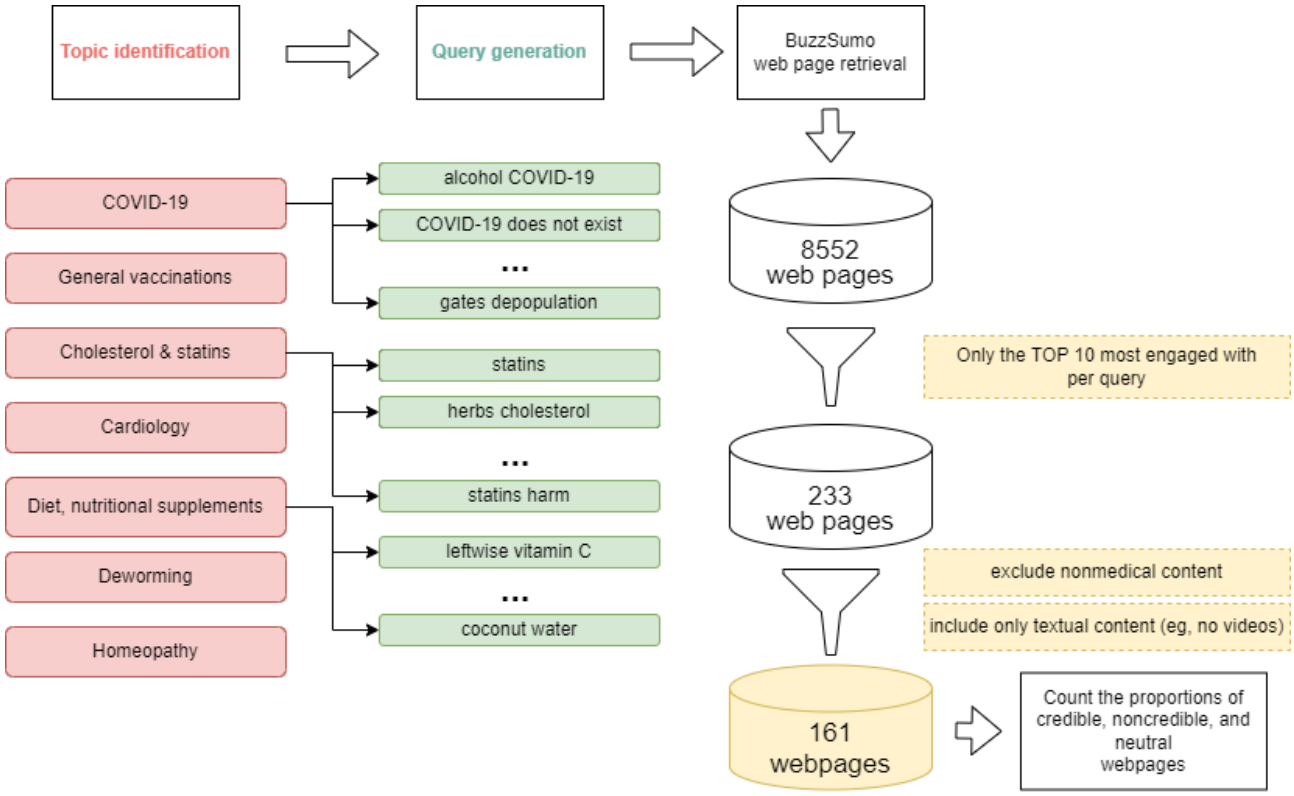

Regarding web pages categorized as medical, the content was evaluated in terms of credibility, classified as credible, noncredible, or neutral. Specialists in the pertinent medical field assessed websites relevant to their expertise, such as cardiologists for topics on statins and infectious disease specialists for COVID-19 topics. The assessment of health information credibility in the news involved comparing it with current textbooks, guidelines, and scientific articles in the respective field. The evaluation adhered to evidence-based medicine standards, and the criteria used for assessment are detailed in Table 1.

XSL•FO
RenderX

Table 1. Content categories.

Category	Content
Credible medical websites	• Contains verifiable/up-to-date medical information and does not contain information intended to mislead the viewer/reader.
Noncredible medical websites	• Contains false medical information. • Contains unverified medical information. • Contains outdated medical information that is inconsistent with current standards. • Contains medical information consistent with evidence-based medicine, but the implications of the fact(s) presented are falsely exaggerated. • Contains partially not credible medical information: miscalculations or erroneous conclusions from true analyses; a mixture of credible and not credible information. • Contains noncredible medical information or information inconsistent with evidence-based medicine, but part of the text softens the implication of the sentence.
Neutral medical websites	• Contains information related to medicine, but describes a story related to a disease, reports on a medical encounter, and presents regulations in the health care sphere.
Nonmedical website	• Does not contain information related to medicine.

The data were collected and analyzed using Microsoft Excel (Microsoft, Inc.). Descriptive statistics were computed for categorical variables, presenting the frequencies of the variables under investigation in this scientific study. The chi-square test was used to assess the distribution of medical web pages categorized as credible, noncredible, or neutral between the years 2021 and 2022. The chi-square test was used to ascertain whether there was a statistically significant ($P<.05$) difference in the distribution of both the number and popularity of these web pages across the mentioned categories. The aim was to investigate whether there were any shifts in the credibility classification of web pages and their corresponding popularity over the 2 years. In cases where expected values were less than 5, the Fisher exact test was used. Considering the least represented data, the popularity of neutral medical web pages for 2021 and 2022 was excluded from the analysis due to computational constraints associated with the Fisher exact test.

Ethical Considerations

The study used only anonymized or de-identified data. Data was collected using BuzzSumo, which uses aggregated, anonymous statistics and does not access personal data. The Bioethics Committee of the Medical University of Bialystok deemed that this study does not meet the definition of a medical experiment, as it is not carried out on human beings or biological material taken from a human being for scientific purposes. Therefore, the approval of the Bioethics Committee was not required.

Results

Preliminary Questionnaires

The most common topics indicated by doctors were COVID-19 infection, COVID-19 vaccination, vaccination against other infectious diseases, hypercholesterolemia and statins, hypertension, nutritional supplements, and antibiotics. A list of topics is shown in Table 2.

Table 2. The list of topics (problem groups) indicated by doctors in questionnaires, the number of times the topic was mentioned, and the resulting number of queries generated from this topic for data collection purposes.

Number	Topics	Times the topic was indicated by doctors, n	Queries, n
1	Coronavirus disease 2019 (COVID-19)	74	85
2	General vaccinations	64	18
3	Cholesterol and statins	52	45
4	Cardiology	27	68
5	Diet, nutritional supplements, and alcohol	26	32
6	Antibiotics	12	12
7	Deworming	4	9
8	Lyme disease	2	1
9	Homeopathy	1	4

XSL•FO
RenderX

Web Page and Social Media Engagements Retrieval Using BuzzSumo

In total, 8552 web pages were collected in 2021, decreasing to 6404 in 2022. COVID-19 topics held the top spot in popularity during 2021, whereas cholesterol and statins took the lead in 2022. Detailed topic-specific data are provided in Table 3.

After following the steps outlined in Figure 1, a total of 161 web pages from 2021 and 105 from 2022 underwent analysis. The filtering process is illustrated in Table 4.

Table 3. The number of all web pages based on the queries, retrieved by BuzzSumo (2021-2022).

Topics	All web pages retrieved by BuzzSumo in 2021, n	All web pages retrieved by BuzzSumo in 2022, n
Coronavirus disease 2019 (COVID-19)	5448	1495
General vaccinations	468	59
Cholesterol and statins	1152	3282
Cardiology	1326	1456
Diet, nutritional supplements, and alcohol	150	23
Deworming	3	8
Homeopathy	8	80
Total	8552	6404

Table 4. Subsequent stages of web search research according to Figure 1.

Stages	Web pages in 2021, n	Web pages in 2023, n
Web pages retrieved by BuzzSumo using selected queries	8552	6404
Web pages that provoked engagements at least ten times (videos were not evaluated)	233	137
Web pages considered by experts as containing medical content	161	105

In 2021, web pages related to the COVID-19 topic were the most popular, accumulating 221,301 engagements, followed by the general vaccinations topic, which scored 22,414 engagements. Similarly, in 2022, the most popular web pages were those associated with the COVID-19 topic, generating 18,264 engagements, followed by cholesterol and statins, which garnered 16,844 engagements. The topics most tainted with fake or noncredible content in 2021 were COVID-19 (43/76, 57%), followed by general immunization (10/25, 40%). In 2022, diet, nutritional supplements, and alcohol took the lead in content pollution (3/5, 60%), followed by general vaccinations (3/10, 30%). The percentage of fake or noncredible web pages related to COVID-19 and general vaccinations decreased from 57% (43/76) to 25% (6/24) and 42% (10/25) to 30% (3/10), respectively. Conversely, the percentage of fake or noncredible web pages related to cholesterol/statins and cardiology increased from 11% (3/28) to 26% (9/35) and 13% (5/38) to 26% (6/23), respectively (Figure 3).

XSL•FO
RenderX

Figure 3. The figure shows the differences in the prevalence and interactions on social media of non-credible, credible, and neutral web pages in the two periods: (1) January 1, 2021 to December 31, 2021; (2) January 1, 2022 - December 31, 2022. The bars indicate the number of websites that exceeded the social media engagement threshold (top 10 web pages mostly interacted with per search query) in a given period related to a given medical topic. Pie charts show the number of interactions (shares, likes, comments) with non-credible, credible, and neutral web pages over a given period, on a given medical topic. The aggregated social media engagement data was retrieved using the BuzzSumo Enterprise Application.

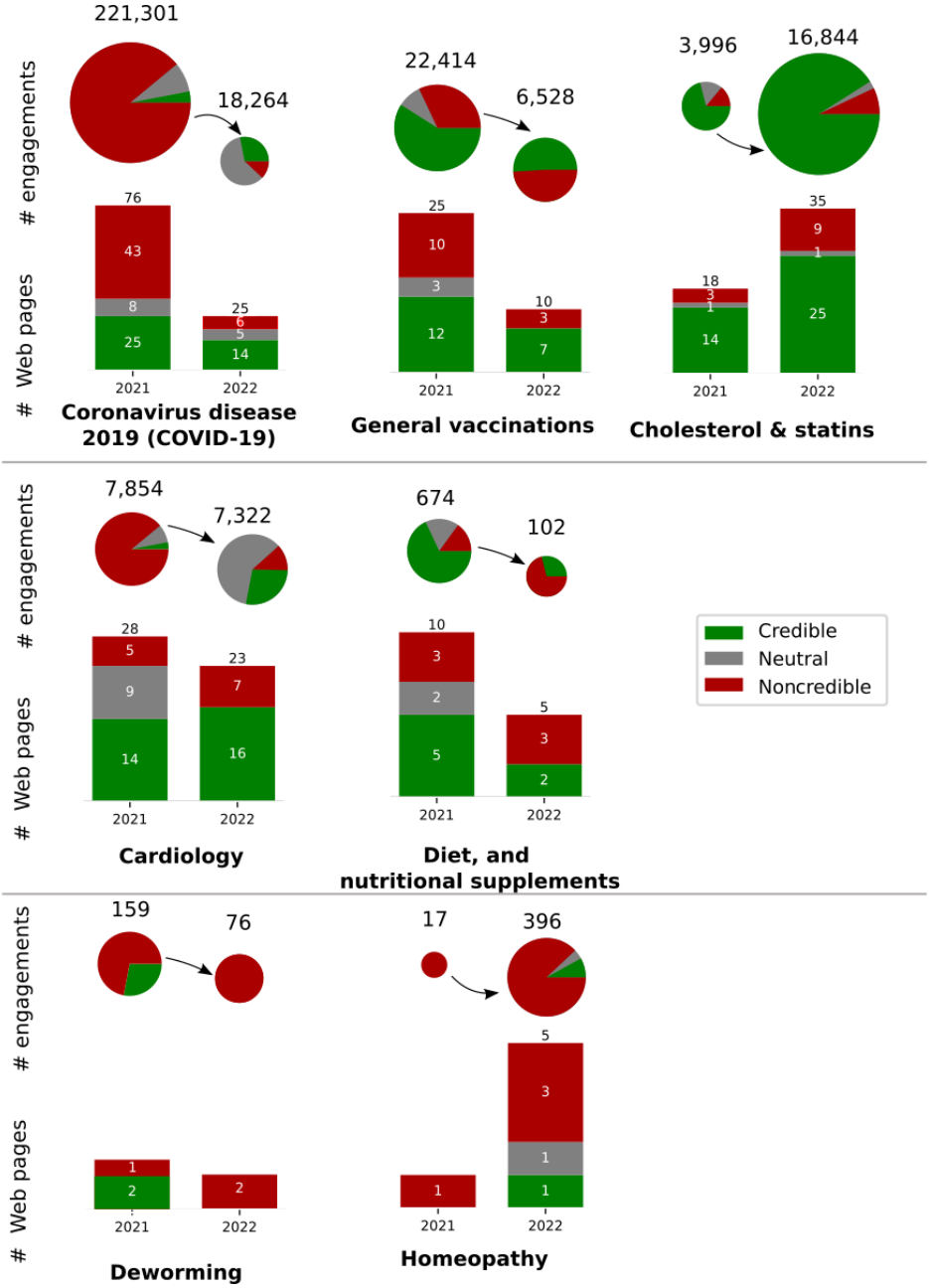

In Table 5, the count of medical web pages, as assessed by medical experts, is categorized into credible, noncredible, or neutral. Web pages labeled as medical neutral did not contain diagnostic or treatment-related medical information but mainly focused on health stories or described changes in medical regulations or conferences. In 2021, the count of medical web pages rated as credible, noncredible, and neutral was 72, 66, and 23, respectively. In 2022, these figures changed to 66, 32, and 7, respectively. Notably, the percentage of noncredible content from all web pages in 2022 was much lower at 30.5% (32/105), compared with 41% (66/161) in 2021.

In 2021, credible web pages garnered 24,398 engagements, increasing to 30,301 engagements in 2022. Web pages featuring fake news generated 210,018 engagements in 2021 and 7942 engagements in 2022. Further details are presented in Table 6.

XSL•FO
RenderX

Table 5. The number of medical web pages, divided into credible, noncredible, and neutral (2021-2022).

Topics	2021			2022			*P* value
	Credible medical web pages, n	Noncredible medical web pages, n	Neutral medical web pages, n	Credible medical web pages, n	Noncredible medical web pages, n	Neutral medical web pages, n	
Coronavirus disease 2019 (COVID-19)	25	43	8	14	6	5	.02[a]
General vaccinations	12	10	3	7	3	0	.47[b]
Cholesterol and statins	14	3	1	25	9	1	.76[b]
Cardiology	14	5	9	17	6	0	.007[b]
Diet, nutritional supplements, and alcohol	5	3	2	2	3	0	.62[b]
Deworming	2	1	0	0	2	0	.40[b]
Homeopathy	0	1	0	1	3	1	>.99[b]
Total	72	66	23	66	32	7	.01[a]

[a]Chi-square test.

[b]Fisher exact test.

Table 6. The number of aggregated engagements of medical web pages, divided into credible, fake news, and neutral (2021-2022).

Topics	2021, n			2022, n			*P* value
	Credible medical web page popularity, n	Noncredible medical web page popularity, n	Neutral medical web page popularity, n	Credible medical web page popularity, n	Noncredible medical web page popularity, n	Neutral medical web page popularity, n	
Coronavirus disease 2019 (COVID-19)	6739	197,100	17,462	5138	2145	10,981	<.001[a]
General vaccinations	13,234	7226	1954	3325	3203	0	<.001[b]
Cholesterol and statins	2831	579	586	15,324	1228	292	<.001[a]
Cardiology	1090	4883	1881	6454	868	0	<.001[b]
Diet, nutritional supplements, and alcohol	459	99	116	30	72	0	<.001[b]
Deworming	45	114	0	0	76	0	<.001[b]
Homeopathy	0	17	0	30	350	16	.33[b]
Total	24,398	210,018	21,999	30,301	7942	11,289	<.001[a]

[a]Chi-square test.

[b]Fisher exact test.

Discussion

Principal Findings

Numerous researchers have endeavored to develop algorithms for identifying disinformation and misinformation on the web. Machine learning protocols and artificial intelligence have been used to detect fake news related to SARS-CoV-2 infection on social media [13,14,39]. Iwendi et al [14] introduced new features to accurately recognize COVID-19–related fake news in online information, achieving an accuracy rate of over 86%. The implementation of machine learning–supported measures during the COVID-19 pandemic to curb the dissemination of fake news appears to have proven effective. Nevertheless, it is imperative to implement such interventions continuously, particularly in emerging medical fields vulnerable to generating unreliable content, given the significant impact on the community.

XSL•FO
RenderX

This study primarily addressed the perspective of the layperson, recognizing the challenges they face in comprehending medical information and identifying reliable sources. Laypersons are consequently at risk of coming across noncredible information [9,25]. By contrast, health care providers rely on established, reliable sources of medical knowledge grounded in evidence-based medicine [10].

In the last 2 years, COVID-19 emerged as the most commonly reported topic generating "fake news" in doctors' daily clinical practice. Despite the availability of evidence-based medicine guidelines for the treatment of COVID-19 [40], our study revealed that in Poland, the most popular web pages still centered around the use of alternative methods to treat SARS-CoV-2 infection, such as amantadine. Notably, the proportion of noncredible web pages in the selected sample regarding COVID-19 was 0.57 in 2021, which reduced to 0.24 in 2022. It is noteworthy that credible web pages were less prevalent on the web, with a considerably lower number of engagements compared with noncredible ones in both years (6739 vs 221,301 and 5138 vs 18,264). This suggests that false information about COVID-19 remained prominent on the web in 2021. However, the significant decrease in total engagements related to noncredible content may indicate that interventions to detect, remove, and correct disinformation about COVID-19, implemented by major tech companies such as Google and Facebook, had a positive impact on curtailing the spread of fake news. On-platform interventions comprised highlighting, surfacing, and prioritizing content from authoritative sources, as well as cooperating with fact-checkers and health authorities to flag and remove disinformation [41]. Another conceivable factor contributing to the diminished spread of COVID-19–related fake news could be a waning interest in the pandemic. Possible reasons are people growing accustomed to it, the easing of restrictions, and the emergence of another global crisis in 2022, namely, the war in Ukraine. Nevertheless, our study has identified a decrease in interest in noncredible information about COVID-19 from 2021 to 2022. Interestingly, the popularity of other topics, particularly cholesterol/statins and cardiology, has seen an increase. It appears that new topics are effectively "filling up the vacuum" created by the declining interest in the virus.

In the case of the topic "cholesterol and statins," social media witnessed a substantial surge in interest in this field between 2021 and 2022. The number of aggregated total engagements grew more than fourfold, escalating from around 4000 to approximately 16,800. This surge coincided with an increased interest in other cardiology topics and a decline in interest in COVID-19. While most web page contents were evaluated as credible, noncredible content accounted for 11% (3/28) in 2021, rising to 26% (9/35) in 2022. Notably, during the period of heightened interest in the topic of "cholesterol and statins," the percentage of content classified as noncredible also experienced an increase. This phenomenon was corroborated in the other studied categories. Hypercholesterolemia is a condition that does not manifest clinical symptoms directly. It remains asymptomatic for an extended period, impacting life comfort only when complications such as stroke or myocardial infarction arise. Patients often visit doctor's offices for unrelated reasons, only to receive a diagnosis of hypercholesterolemia. Subsequently, doctors assess the need for pharmacological treatment. As a consequence of this unique scenario, doctors' competence in this field is frequently called into question. Given the general low trust in the medical profession in Poland, coupled with heightened fears of side effects and distant prospects of clinical benefits, patients often seek alternative information on the internet and social media. The increased interest in the subject serves as fertile ground for the creation of fake news content.

Based on the results obtained, the topic of deworming generated the least engagement. However, this issue was identified as prevalent among patients in our initial questionnaire. Surprisingly, the number of engagements for noncredible web pages concerning this topic surged from 71.7% (114/159) in 2021 to a staggering 100% (76/76) in 2022. While the absolute numbers of web pages are modest (3 and 2 for 2021 and 2022, respectively), the absence of any credible source of information about this topic in 2022 should raise alarms within the medical community. Qualitative insights uncovered the influential role of social media influencers in shaping the narratives surrounding deworming. These findings suggest that deworming topics might not generate a substantial number of clicks, but they carry a potent and convincing message for the layperson. A study reveals that the analysis of internet searches accurately predicts seasonal peaks in hairworm numbers, identifying host characteristics that align precisely with scientific data [42]. Drawing from this study, we can infer that lay internet users frequently search for information about parasites but seldom land on reliable sources.

Throughout the pandemic, we all encountered the phenomenon known as an "infodemic," an influx of an extensive amount of information, including false and fake news, during the outbreak of a particular disease [15]. While COVID-19 serves as a vivid illustration of how harmful and contagious medical disinformation can be, there remain other health topics where online disinformation is rampant. These topics were already sources of disinformation before the COVID-19 pandemic and continue to pose a significant threat to society. One of the strengths of our study is that the analysis of most unreliable topics is grounded in clinical practice. Additionally, the evaluation of the websites was conducted by an experienced team of clinicians.

Limitations

The initial questionnaire had a limited sample size with only 121 respondents. Unfortunately, during the COVID-19 pandemic, doctors were facing overwhelming workloads. Nevertheless, the respondents represented a diverse group of doctors from various specialties, working in different settings—both inpatient and outpatient—and spanning different age groups. Notably, these doctors provided comprehensive descriptions of all the problems they had encountered with disinformation in their professional work. The questionnaire did not prescribe specific topics for discussion. Doctors, based on their clinical experience, independently described these phenomena. Despite the relatively small number of respondents, this approach ensured a comprehensive exploration of the topics.

XSL•FO
RenderX

Examining only the first 10 websites selected with BuzzSumo using the queries might not provide a complete picture of the spread of medical fake news across the web. Nevertheless, given our limited resources, primarily the time of the experts, we endeavored to capture the dynamics of the most popular content as accurately as possible. Another limitation is the selection of websites in both time intervals using precisely the same queries. In the second time slot, there is a possibility that new medical topics were not included in the selection. To address this limitation, we plan to reintroduce the questionnaire and assess the dynamics to identify potential new and emerging topics generating significant medical noncredible content on the internet.

Conclusions

Our analysis underscores several critical areas for improvement in interventions aimed at halting the spread of noncredible online health information. These can be summarized as follows:

- While interventions implemented by big tech companies have proven effective in combating COVID-19 disinformation, the surge in the dissemination of noncredible content between 2021 and 2022 in other topics underscores the need for these interventions to extend beyond COVID-19–related issues.
- Beyond COVID-19, the topics of “cholesterol and statins” and “cardiology” stand out as significant generators of unreliable content. Initiating interventions with a focus on these topics may be a worthwhile consideration.
- The “deworming” topic illustrates that online activity does not consistently mirror the actual spread of disinformation. Despite being frequently indicated in questionnaires, this topic generated only a few web pages, with minimal total engagements. It is noteworthy that all web pages found in 2022 related to this topic were noncredible. This underscores the significance of periodically repeating questionnaires among physicians.
- The “deworming” case also brings to light the inadequacy of coverage for certain medical topics by credible sources of online health information. There should be a special emphasis on addressing such topics.
- Continuous monitoring of the dynamics in the dissemination of disinformation for topics that are frequently shared, commented on, or liked is essential. This proactive approach allows for timely reactions to potential future infodemics.

Acknowledgments

The research was carried out as part of the research work number INFOSTRATEG-I/0010/2021-00 at the Medical University of Bialystok (Poland), Polish-Japanese Academy of Information Technology in Warsaw (Poland), and Science4People sp. z o.o. in Szczecin (Poland) financed from a subsidy provided by the National Centre for Research and Development, Poland.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The preliminary questionnaire body.
[PNG File , 39 KB - jmir_v26i1e48130_app1.png]

References

1. Kranzberg M. History of technology. Science 1960 Apr 22;131(3408):1246-1247. [doi: 10.1126/science.131.3408.1246]
2. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 02;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
3. Wierzbicki A. Web Content Credibility. Cham, Switzerland: Springer International Publishing; 2018.
4. Baines D, Elliott R. Defining misinformation, disinformation and malinformation: an urgent need for clarity during the COVID-19 infodemic (Working Paper). University of Birmingham. 2020. URL: https://repec.cal.bham.ac.uk/pdf/20-06.pdf [accessed 2024-01-01]
5. Packer M. Are meta-analyses a form of medical fake news? Thoughts about how they should contribute to medical science and practice. Circulation 2017 Nov 28;136(22):2097-2099. [doi: 10.1161/CIRCULATIONAHA.117.030209] [Medline: 29180491]
6. Mamak K. Do we need the criminalization of medical fake news? Med Health Care Philos 2021 Jun;24(2):235-245 [FREE Full text] [doi: 10.1007/s11019-020-09996-7] [Medline: 33398487]
7. Islam MS, Sarkar T, Khan SH, Mostofa Kamal A, Hasan SMM, Kabir A, et al. COVID-19-related infodemic and its impact on public health: a global social media analysis. Am J Trop Med Hyg 2020 Oct;103(4):1621-1629 [FREE Full text] [doi: 10.4269/ajtmh.20-0812] [Medline: 32783794]
8. Gavaravarapu SM, Seal A, Banerjee P, Reddy T, Pittla N. Impact of 'infodemic in pandemic' on food and nutrition related perceptions and practices of Indian internet users. PLoS One 2022;17(4):e0266705 [FREE Full text] [doi: 10.1371/journal.pone.0266705] [Medline: 35446865]
9. Betsch C, Renkewitz F, Betsch T, Ulshöfer C. The influence of vaccine-critical websites on perceiving vaccination risks. J Health Psychol 2010 Mar 26;15(3):446-455. [doi: 10.1177/1359105309353647] [Medline: 20348365]

XSL•FO
RenderX

10. Guyatt G, Cook D, Haynes B. Evidence based medicine has come a long way. BMJ 2004 Oct 30;329(7473):990-991 [FREE Full text] [doi: 10.1136/bmj.329.7473.990] [Medline: 15514320]
11. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996 Jan 13;312(7023):71-72 [FREE Full text] [doi: 10.1136/bmj.312.7023.71] [Medline: 8555924]
12. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. JAMA 2020 Aug 25;324(8):782-793. [doi: 10.1001/jama.2020.12839] [Medline: 32648899]
13. Pennycook G, McPhetres J, Zhang Y, Lu JG, Rand DG. Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. Psychol Sci 2020 Jul;31(7):770-780 [FREE Full text] [doi: 10.1177/0956797620939054] [Medline: 32603243]
14. Iwendi C, Mohan S, Khan S, Ibeke E, Ahmadian A, Ciano T. Covid-19 fake news sentiment analysis. Comput Electr Eng 2022 Jul;101:107967 [FREE Full text] [doi: 10.1016/j.compeleceng.2022.107967] [Medline: 35474674]
15. World Health Organization (WHO). Infodemic. WHO. Geneva, Switzerland: World Health Organization URL: https://www.who.int/health-topics/infodemic#tab=tab_1 [accessed 2023-04-01]
16. Borges do Nascimento IJ, Pizarro AB, Almeida JM, Azzopardi-Muscat N, Gonçalves MA, Björklund M, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022 Sep 01;100(9):544-561 [FREE Full text] [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
17. Montagni I, Ouazzani-Touhami K, Mebarki A, Texier N, Schück S, Tzourio C, CONFINS group. Acceptance of a Covid-19 vaccine is associated with ability to detect fake news and health literacy. J Public Health (Oxf) 2021 Dec 10;43(4):695-702 [FREE Full text] [doi: 10.1093/pubmed/fdab028] [Medline: 33693905]
18. Sowa P, Kiszkiel, Laskowski PP, Alimowski M, Szczerbiński, Paniczko M, et al. COVID-19 vaccine hesitancy in Poland-multifactorial impact trajectories. Vaccines (Basel) 2021 Aug 07;9(8):876 [FREE Full text] [doi: 10.3390/vaccines9080876] [Medline: 34452001]
19. Jolley D, Douglas KM. The effects of anti-vaccine conspiracy theories on vaccination intentions. PLoS One 2014;9(2):e89177 [FREE Full text] [doi: 10.1371/journal.pone.0089177] [Medline: 24586574]
20. Toll M, Li A. Vaccine sentiments and under-vaccination: attitudes and behaviour around measles, mumps, and rubella vaccine (MMR) in an Australian cohort. Vaccine 2021 Jan 22;39(4):751-759. [doi: 10.1016/j.vaccine.2020.11.021] [Medline: 33218781]
21. Borah P, Hwang J. Trust in doctors, positive attitudes, and vaccination behavior: the role of doctor-patient communication in H1N1 vaccination. Health Commun 2022 Oct;37(11):1423-1431. [doi: 10.1080/10410236.2021.1895426] [Medline: 33685304]
22. Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger J. Vaccine hesitancy: an overview. Hum Vaccin Immunother 2013 Aug;9(8):1763-1773 [FREE Full text] [doi: 10.4161/hv.24657] [Medline: 23584253]
23. Sommariva S, Vamos C, Mantzarlis A, Đào LU, Martinez Tyson D. Spreading the (fake) news: exploring health messages on social media and the implications for health professionals using a case study. American Journal of Health Education 2018 Jun 07;49(4):246-255. [doi: 10.1080/19325037.2018.1473178]
24. Pakalniškienė V, Kairys A, Jurkuvėnas V, Mikuličiūtė V, Ivleva V. Could belief in fake news predict vaccination behavior in the elderly? Int J Environ Res Public Health 2022 Nov 12;19(22):14901 [FREE Full text] [doi: 10.3390/ijerph192214901] [Medline: 36429616]
25. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
26. Catapano AL, Wiklund O, European Atherosclerosis Society. Think again about cholesterol survey. Atheroscler Suppl 2015 Dec;20:1-5. [doi: 10.1016/S1567-5688(15)30011-8] [Medline: 26671304]
27. Hill JA. Medical misinformation. Thorac Cardiovasc Surg 2019 Mar;67(2):80-82 [FREE Full text] [doi: 10.1055/s-0039-1678544] [Medline: 30695791]
28. Waszak PM, Kasprzycka-Waszak W, Kubanek A. The spread of medical fake news in social media – the pilot quantitative study. Health Policy and Technology 2018 Jun;7(2):115-118. [doi: 10.1016/j.hlpt.2018.03.002]
29. Reyneke G, Hughes J, Grafenauer S. Consumer understanding of the Australian Dietary Guidelines: recommendations for legumes and whole grains. Nutrients 2022 Apr 22;14(9):1753 [FREE Full text] [doi: 10.3390/nu14091753] [Medline: 35565721]
30. Ruani MA, Reiss MJ. Susceptibility to COVID-19 nutrition misinformation and eating behavior change during lockdowns: an international web-based survey. Nutrients 2023 Jan 14;15(2):451 [FREE Full text] [doi: 10.3390/nu15020451] [Medline: 36678321]
31. Inan-Eroglu E, Buyuktuncer Z. YouTube: is it a reliable source of nutrition information on COVID-19 pandemic? Healthcare (Basel) 2022 Sep 29;10(10):1911 [FREE Full text] [doi: 10.3390/healthcare10101911] [Medline: 36292358]
32. Moon H, Lee GH. Evaluation of Korean-language COVID-19-related medical information on YouTube: cross-sectional infodemiology study. J Med Internet Res 2020 Aug 12;22(8):e20775 [FREE Full text] [doi: 10.2196/20775] [Medline: 32730221]

XSL•FO
RenderX

33. Ahmed I, Hasan M, Akter R, Sarkar BK, Rahman M, Sarker MS, et al. Behavioral preventive measures and the use of medicines and herbal products among the public in response to Covid-19 in Bangladesh: a cross-sectional study. PLoS One 2020;15(12):e0243706 [FREE Full text] [doi: 10.1371/journal.pone.0243706] [Medline: 33306725]
34. Hamulka J, Jeruszka-Bielak M, Górnicka M, Drywień ME, Zielinska-Pukos MA. Dietary supplements during COVID-19 outbreak. Results of Google Trends analysis supported by PLifeCOVID-19 online studies. Nutrients 2020 Dec 27;13(1):54 [FREE Full text] [doi: 10.3390/nu13010054] [Medline: 33375422]
35. Lebin JA, Mudan A, Wu AHB. Chronic alcohol use does not protect against COVID-19 infection. Am J Emerg Med 2021 Jul;45:640-641 [FREE Full text] [doi: 10.1016/j.ajem.2020.11.024] [Medline: 33221114]
36. Aghababaeian H, Hamdanieh L, Ostadtaghizadeh A. Alcohol intake in an attempt to fight COVID-19: a medical myth in Iran. Alcohol 2020 Nov;88:29-32 [FREE Full text] [doi: 10.1016/j.alcohol.2020.07.006] [Medline: 32693023]
37. Luk TT, Zhao S, Weng X, Wong JY, Wu YS, Ho SY, et al. Exposure to health misinformation about COVID-19 and increased tobacco and alcohol use: a population-based survey in Hong Kong. Tob Control 2021 Nov;30(6):696-699. [doi: 10.1136/tobaccocontrol-2020-055960] [Medline: 32855353]
38. BuzzSumo. URL: http://www.buzzsumo.com/ [accessed 2023-04-01]
39. Khan S, Hakak S, Deepa N, Prabadevi B, Dev K, Trelova S. Detecting COVID-19-related fake news using feature extraction. Front Public Health 2021;9:788074 [FREE Full text] [doi: 10.3389/fpubh.2021.788074] [Medline: 35059379]
40. Flisiak R. Zalecenia dotyczące postępowania w zakażeniach SARS-CoV-2 Polskiego Towarzystwa Epidemiologów i Lekarzy Chorób Zakaźnych z 23 lutego. Medycyna Praktyczna 2022;2022(3):49-77 [FREE Full text]
41. Organisation for Economic Co-operation and Development (OECD). Combatting COVID-19 disinformation on online platforms. OECD. Paris, France: OECD; 2020 Jul 3. URL: https://www.oecd.org/coronavirus/policy-responses/combatting-covid-19-disinformation-on-online-platforms-d854ec48/ [accessed 2024-01-01]
42. Doherty J, Filion A, Bennett J, Raj Bhattarai U, Chai X, de Angeli Dutra D, et al. The people science: can passively crowdsourced internet data shed light on host-parasite interactions? Parasitology 2021 Sep;148(11):1313-1319. [doi: 10.1017/S0031182021000962] [Medline: 34103103]

Abbreviations

MMR: measles, mumps, and rubella
WWW: World Wide Web

Edited by G Eysenbach; submitted 14.04.23; peer-reviewed by S Neely, F Medina; comments to author 04.08.23; revised version received 26.08.23; accepted 29.11.23; published 29.03.24.

Please cite as:
Chlabicz M, Nabożny A, Koszelew J, Łaguna W, Szpakowicz A, Sowa P, Budny W, Guziejko K, Róg-Makal M, Pancewicz S, Kondrusik M, Czupryna P, Cudowska B, Lebensztejn D, Moniuszko-Malinowska A, Wierzbicki A, Kamiński KA
Medical Misinformation in Polish on the World Wide Web During the COVID-19 Pandemic Period: Infodemiology Study
J Med Internet Res 2024;26:e48130
URL: https://www.jmir.org/2024/1/e48130
doi:10.2196/48130
PMID:38551638

©Małgorzata Chlabicz, Aleksandra Nabożny, Jolanta Koszelew, Wojciech Łaguna, Anna Szpakowicz, Paweł Sowa, Wojciech Budny, Katarzyna Guziejko, Magdalena Róg-Makal, Sławomir Pancewicz, Maciej Kondrusik, Piotr Czupryna, Beata Cudowska, Dariusz Lebensztejn, Anna Moniuszko-Malinowska, Adam Wierzbicki, Karol A Kamiński. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Influence of Joe Wicks on Physical Activity During the COVID-19 Pandemic: Thematic, Location, and Social Network Analysis of X Data

Wasim Ahmed[1], BA, MSc, PhD; Opeoluwa Aiyenitaju[2], BE, MSc, PhD; Simon Chadwick[3], PhD; Mariann Hardey[4], BA, MA, PhD; Alex Fenton[5], BSc, MA, PhD

[1]Management School, University of Stirling, Stirling, United Kingdom
[2]Business School, Manchester Metropolitan University, Manchester, United Kingdom
[3]School of Knowledge Economy and Management, Paris, France
[4]Business School, University of Durham, Durham, United Kingdom
[5]Business School, University of Chester, Chester, United Kingdom

Corresponding Author:
Wasim Ahmed, BA, MSc, PhD
Management School
University of Stirling
Airthrey Road
Stirling, FK9 4LA
United Kingdom
Phone: 44 1482 346311
Email: w.ahmed@hull.ac.uk

Abstract

Background: Social media (SM) was essential in promoting physical activity during the COVID-19 pandemic, especially among people confined to their homes. Joe Wicks, a fitness coach, became particularly popular on SM during this time, posting daily workouts that millions of people worldwide followed.

Objective: This study aims to investigate the influence of Joe Wicks on SM and the impact of his content on physical activity levels among the public.

Methods: We used NodeXL Pro (Social Media Research Foundation) to collect data from X (formerly Twitter) over 54 days (March 23, 2020, to May 15, 2020), corresponding to the strictest lockdowns in the United Kingdom. We collected 290,649 posts, which we analyzed using social network analysis, thematic analysis, time-series analysis, and location analysis.

Results: We found that there was significant engagement with content generated by Wicks, including reposts, likes, and comments. The most common types of posts were those that contained images, videos, and text of young people (school-aged children) undertaking physical activity by watching content created by Joe Wicks and posts from schools encouraging pupils to engage with the content. Other shared posts included those that encouraged others to join the fitness classes run by Wicks and those that contained general commentary. We also found that Wicks' network of influence was extensive and complex. It contained numerous subcommunities and resembled a broadcast network shape. Other influencers added to engagement with Wicks via their networks. Our results show that influencers can create networks of influence that are exhibited in distinctive ways.

Conclusions: Our study found that Joe Wicks was a highly influential figure on SM during the COVID-19 pandemic and that his content positively impacted physical activity levels among the public. Our findings suggest that influencers can play an important role in promoting public health and that government officials should consider working with influencers to communicate health messages and promote healthy behaviors. Our study has broader implications beyond the status of fitness influencers. Recognizing the critical role of individuals such as Joe Wicks in terms of health capital should be a critical area of inquiry for governments, public health authorities, and policy makers and mirrors the growing interest in health capital as part of embodied and digital experiences in everyday life.

(J Med Internet Res 2024;26:e49921) doi:10.2196/49921

XSL•FO
RenderX

KEYWORDS

social media; social network analysis; COVID-19; influencers; public health; social network; physical activity; promotion; fitness; exercise; workout; Twitter; content creation; communication

Introduction

Background

Numerous countries across the world are encountering significant health problems, which are often considered to be a crisis [1]. These problems range from increasing obesity [2] and a high incidence of diabetes [3] to heart problems [4]. The acuteness of such problems is thought to be the consequence of factors including, for example, poor diet [5]. However, they are also identified as being associated with the increasingly sedentary lifestyles that some people lead [6]. A sedentary lifestyle is characterized as one that involves extended engagement in behaviors characterized by minimal movement, low energy expenditure, and significant periods of rest [7].

This study aims to examine the role of a fitness influencer on the social media (SM) platform known as "X." X serves as a platform for the real-time dissemination of information, public engagement, and discourse. Rebranded in 2023 from "Twitter" (which was the name of the platform when the study was conducted), X allows users to post brief messages, now called "posts," which can be easily shared through "reposts" and interacted with by a global audience. The fitness influencer under study is Joe Wicks, who is a British fitness coach, TV presenter, and author. Wicks gained considerable public attention during the COVID-19 pandemic through his web-based workout sessions designed to promote physical exercise during lockdowns.

During the first quarter of 2020, most countries globally were affected by the onset of the COVID-19 pandemic [8]. A feature of lockdowns across the world was the compulsory closure of schools. This further heightened worries about physical inactivity during the pandemic, as schools are often the only means some children can engage in physical activity [9]. This often resulted in the physical aspects of school curriculums either being ignored or neglected [10]. Hence, during the early days of the first lockdowns taking place worldwide, there was considerable societal and parental concern about maintaining activity levels. It is within this context that a British fitness coach, Joe Wicks, took the unilateral decision to become the nation's self-appointed "school PE teacher" (PE refers to physical education). At the time, Wicks had already become a prominent figure on SM, sharing fitness tips and training programs via his YouTube, Instagram, and X accounts.

Given the profound impact that Wicks has had on people's attitudes and behaviors, we undertook to examine the nature of his SM influence. Specifically, we aim to highlight how his content has moved around a SM network, identify significant other users, determine the shape of Wicks' network, analyze how people interacted with the content, and identify the geographical spread of content. In so doing, this study's overall contribution is to how individuals such as Wicks influence others to engage in physical activity.

The policy implications of this study are significant, with health organizations showing interest in this line of work. Although there is some evidence from the literature to indicate that spokespeople and endorsers have played an important role in state-led communications campaigns [11], SM has, therefore, changed the dynamic of relations between governments and populations [12]. We thus assert that this study has significant implications for public health messaging and the way such communications are delivered by governments.

Overview of Existing Research

In this section, we will review discussions on the formation of social networks, SM, and SM influencers in the context of public health. Furthermore, we highlight research gaps in SM engagement in public health. SM can be described as the collection of digital channels and tools such as Facebook, YouTube, and X with defining characteristics to facilitate engagement [13,14] and encourage user-generated content, social interaction, and real-time collaboration. Specifically, SM use is one of the most popular web-based activities, with an estimated 3.6 billion users worldwide as of 2020, a number that is projected to increase to approximately 4.41 billion by 2025 [15]. Facebook, X, Instagram, WhatsApp, and Google are some of the SM platforms being used every day by millions of people worldwide for accessing and disseminating information [16].

Heinze et al [17] describe SM influencers as individuals who influence behavior to a lesser or greater extent. They can be classed as micro, macro, meso, or mega influencers in terms of their scale of influence, number of followers, and characteristics. SM also enables people to interact from diverse geographical locations. It also provides the opportunity to learn about an audience through conversations. SM not only shapes lives but also influences professional environments [18,19]. The increasing use of SM in health care has enabled its users to access and share information regarding health topics. As such, it has empowered users to take control of their health care needs [20].

Public health organizations need to further harness SM channels to identify what people care about and are discussing to ensure an informed communication strategy [21]. Despite the fact that the various conversations on the possible relevance SM could have in influencing health behaviors and outcomes [18,22], oftentimes, public health organizations use SM as a 1-way broadcast and mass dissemination of information rather than engaging audiences in interactive information-sharing dialogue [14,23,24]. The increased use of SM is prevalent, "People have discovered ways to use SM to impact their health, their community's health, and public health, much of which has gone entirely unstudied" [25]. Bruno Latour has produced significant work on the nature of truth, particularly in science and public health. Latour argues that truth is not a fixed or objective thing but rather is constructed through social interactions. This means that what we consider true is a matter of facts and the relationships between people, objects, and ideas. Latour's work

XSL•FO
RenderX

has important implications for the study of influencers—particularly the impact of digital health contexts [26]. If truth is not fixed, then the influence of an influencer may not be solely based on their expertise or trustworthiness. Rather, it may also be influenced by factors such as the influencer's social network, their ability to connect with their audience, and how they frame their messages.

SM is useful for identifying the key public health influencers and driving web-based conversations on health and well-being topics that are valuable and meaningful to them. Influencers can be classed as both organizations and individuals [27,28] with qualities such as credibility, expertise, and enthusiasm [29]. Influencers are also persistent in convincing consumers or driving conversations on a topic. Among other interesting descriptions of SM influencers, they are viewed as those who have built a sizeable social network of followers or as a third-party endorser that shapes audience attitudes using SM such as blogs, posts, and so on [30-32]. In their early assessment of SM's impact on health research, Moorhead et al [33] discovered that SM can improve health care by providing a platform for public engagement, patient interaction, and health-professional communication regarding health-related issues. These interactions have the potential to promote health. In addition, the authors speculate how SM can enhance health communication. However, the authors caution that it is essential to carefully monitor the quality and veracity of information shared on SM platforms while protecting the privacy of users. These results align with those of studies conducted on health determinants.

The literature discusses various challenges in SM engagement for public health. The need for studies with larger sample sizes and more robust methodologies to study SM algorithms and behavior is crucial [33-35]. A literature gap also exists relating to the effectiveness of SM engagement on public health outcomes and how these impacts change individual behavior, altering public attitudes and influencing public health policy [14,25,36]. Increasingly, and certainly, in response to national lockdowns for COVID-19, SM serves as a conduit for health-related information and support. Such digital use by health consumers, some of whom are motivated to join online community groups to help cope with long-term health conditions, has been linked to increases in social support and coping and support for mental health [37,38]. For example, Hardey's [39] and Hjorth and Lupton's [40] research has demonstrated how the combination of health information on SM and the use of self-tracking technologies for managing chronic conditions is associated with higher levels of health sharing between intimate partners and new forms of social support, as well as improved engagement with health practitioners. In addition, Weinstein et al [41] found for some young men that the use of SM support groups led to reductions in feelings of loneliness and suicidal thoughts. Finally, Dowrick et al's [42] study of digital connections between UK health care workers during the COVID-19 pandemic showed the potential use of SM to support care interactions and increase patient social capital. Patient social capital refers to the resources that a patient has access to through their social network, such as emotional support, financial assistance, practical help, and information [43].

While this body of work paints an inspiring image of the impact of digital networks on health consumers, 2 significant areas in the literature require further investigation. The first involves prior studies focusing on the influence of digital networks on individual health consumers or specific health concerns and evaluating long-term perceptions of change in response to certain types of information (including well-known individuals). Monitoring this type of health behavior change covers broader networks, such as caregivers, who require assistance in caring for and coping with patients [44]. The second area that deserves investigation is the interaction between digital media consumption, social capital [45,46], and intended health outcomes [47]. Despite the fact that there has been little research on the effectiveness of health influencers' reach and the means by which they might promote health change, a number of studies have looked at how digital content use affects social capital—and, as a result, health outcomes. Natural language processing was used in 1 study to analyze X text in order to assess how fitness affects the development of social capital [48]. Another study compared multiple SM platforms (YouTube and Instagram) with health care professionals, fitness influencers, and healthy food influencers in order to quantify engagement [49].

Previous studies have hypothesized that how consumers interact with or have access to digital health content improves their sense of social support, which influences well-being, which affects social capital and health outcomes [50]. This study aims to establish the extent to which Joe Wicks had an impact on health during the pandemic. By analyzing network reach and themes, we hope to analyze representations of health and pictures of fitness as interpreted through digital engagement. By doing so, we can better examine digital manifestations of "health" in everyday life—the components of and settings for embodied fitness activities. Furthermore, by determining Wicks' influence, we aim to assess whether this information was effective in persuading people to engage in physical exercise during the pandemic. This study adds to the growing body of research on governments' neoliberal health objectives and the depiction of fitness personalities in the perception and engagement with health and fitness. The study adds empirical understandings to the study of fitness and health influencers while also engaging critically with literature. To make SM engagement more effective in a crisis, influencers or public health communicators should be able to respond and engage with the audience in real time [51].

Methods

Data Retrieval

The X Academic Track application programming interface was used to retrieve posts from March 23, 2020 (T00:00:00Z) to May 15, 2020 (T23:59:00Z). This time period covered a time that included some of the strictest lockdown policies during the pandemic in the United Kingdom. In total, 290,649 posts were retrieved. The keywords "Joe Wicks" or "thebodycoach" were used to retrieve posts, and this included users who were replied

XSL•FO
RenderX

to or mentioned in posts. The term “thebodycoach” was used as it is the X (Twitter) user handle of Joe Wicks. X is the most open network and, therefore, lends itself well to social network analysis (SNA) and SM research [52]. The method has been widely used in recent research to study COVID-19 disinformation [53-55].

Data Analysis (SNA and Thematic Analysis)

SNA

A systematic random sample of data (5%) was entered into NodeXL Pro for further SNA analysis. Numerous research studies have successfully used NodeXL Pro for analyzing SM data using SNA to develop new knowledge [56]. Moreover, previous empirical research has identified 6 types of X network structures that topics on X follow, as shown in Figure 1 [57]. This study used an SNA approach to answer the research questions.

The information flow within the SM space involves a few types of engagements of users with content such as following, subscribing, sharing, and reposting [58]. SM users form a pattern of connections viewed as social networks. For instance, on X, connections are formed when users mention, like, reply, or repost other users, which reveal the shape of their network [59]. The structure of a network captures patterns of information flow, such as measurements of centrality, degree, and betweenness [60,61]. Centralization captures the degree of aggregated connections around a few actors in a network. Centrality depicts the hierarchy of information flow and allows the identification of influential users in a network. Research on X has found that these information flows revolve around 6 kinds of X SM network structures: polarized crowds (divided), tight crowds (unified), brand clusters (fragmented), community clusters (clustered), broadcast networks (in-hub and spoke), and support networks (out-hub and spoke) [57], which were provided above.

Figure 1. Six types of X networks (simplified form). Sources: Smith et al [57] and the Social Media Research Foundation.

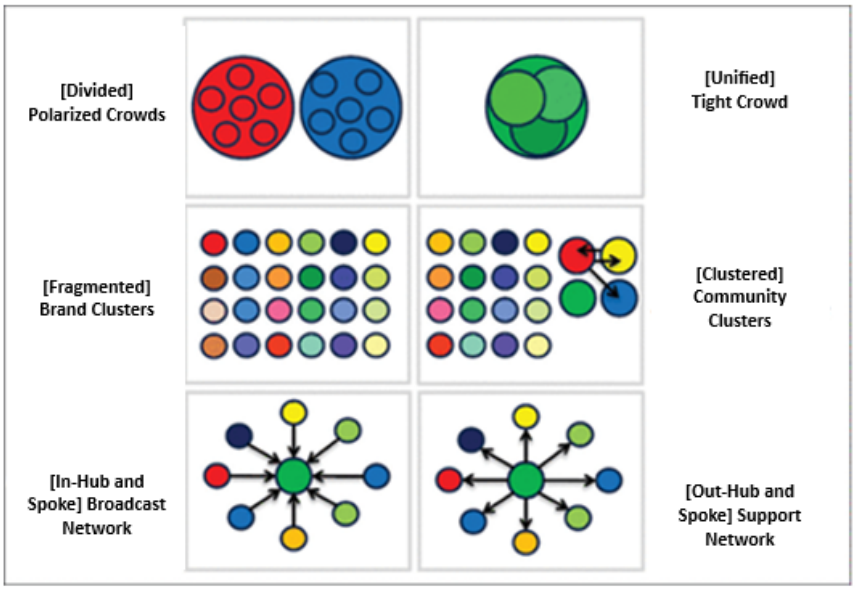

Thematic Analysis

Data were then analyzed using thematic analysis [62] until thematic saturation occurred [63]. A sample of data (n=400) was then used to project the frequency of themes across the data which would reflect the sample overall. Braun and Clarke’s [62] thematic analysis consists of 6 main steps aimed at identifying and analyzing patterns or themes. The process began with becoming familiar with the data through repeated reading and initial note-taking. The second step was to generate initial codes from the data. Coding is the process of tagging or highlighting segments of the data that appear interesting or relevant to the research question. In the third step, the focus was on searching for themes by collating all the coded data. The fourth step was to review these potential themes to ensure they have enough data to support them and that they are distinct from one another. The fifth step was to define and name themes. This involved developing a detailed analysis of each theme, determining what aspect of the data each theme captured and what was interesting about them. Finally, the sixth step was to write up the results.

Betweenness Centrality

The betweenness centrality metric can be used to detect influence within a network. We can express the betweenness centrality of a node “v” as follows:

In relation to the above, σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the occurrence of those paths that pass through v (but not where v is the end point). The metric is commonly applied to SM platforms to detect influential users [64]. This is because betweenness centrality can identify the important users within a network based on how much they act as a bridge between different parts of a network. We can think of this in another way. That is, if we were to remove select users with a high betweenness centrality, this would disrupt the entire network. This is further represented in Figure 2 below, which

XSL•FO
RenderX

shows “D” in red as the most influential node because it has a unique link to node “E.”

Figure 2. Betweenness centrality example.

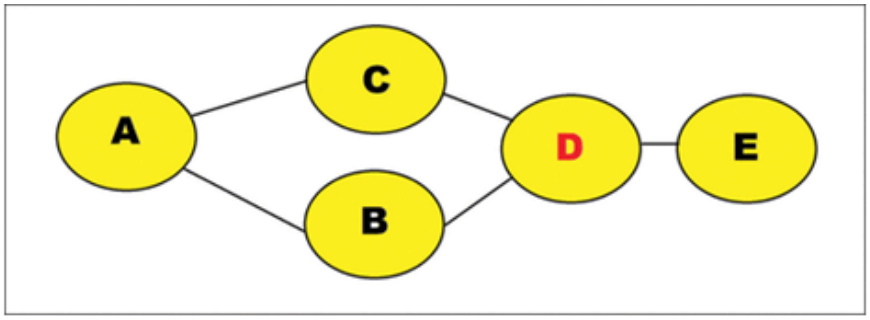

Ethical Considerations

The study gained ethical approval from Newcastle University (Ref: 26055/2022). The study was observational using SM data. However, care was taken to not draw attention to individual users, and SM posts and user handles were anonymized within the manuscript.

Results

Results of SNA

Figure 3 provides an overview of the network’s top 10 groups, and each group is labeled from the largest (group 1) to the smallest (group 10). The network shows Wicks’ influence across the top 10 groups of users. We can see that Wicks’ reach stretched across several groups due to the number of lines pointing out the group to which Joe Wicks belongs, as this indicates users in other groups were reposting and mentioning Wicks’ content. Groups 2 to 10 highlight how several smaller subcommunities have formed around Wicks’ content, led by other influencers that are listed in Table 1.

Figure 3. Examining the top 10 largest groups on X posting about Joe Wicks (March 23, 2020, to May 15, 2020).

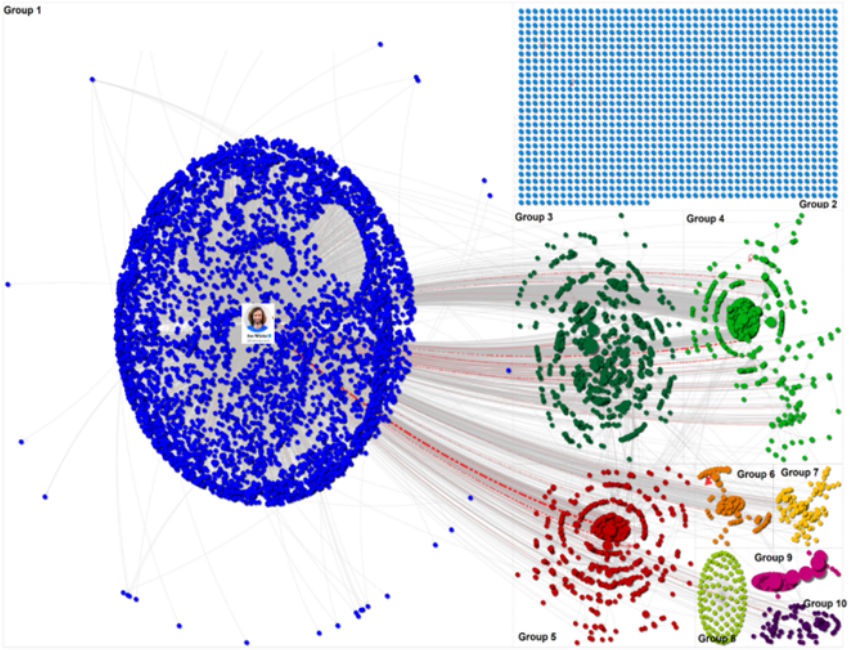

XSL•FO
RenderX

Table 1. Overview of the key influential users.

Rank	User	Betweenness centrality	Followers	Location
1	Joe Wicks	245,642,386	421,565	England
2	James Corden	2,746,155	10,959,093	California
3	Citizen	2,715,772	797	England
4	Citizen	2,559,876	2158	Derbyshire
5	Citizen	2,528,820	4875	United Kingdom and South Africa
6	Freelance writer, editor, and author	2,504,078	10,620	N/A[a]
7	Citizen	1,749,337	9	N/A
8	Citizen	1,402,979	182	England
9	Math teacher	1,311,087	4053	England
10	Citizen	1,284,940	532	N/A

[a]N/A: not applicable.

Results of Betweenness Centrality

Table 1 provides an overview of the most influential users (ranked by betweenness centrality) and provides insights into the rank, user type, betweenness centrality score, number of followers, and location.

Joe Wicks was among the most impactful, but this is not surprising and potentially an artifact of the data collection (as the user-handle of Joe Wicks and search string was used to obtain data). James Corden is in second place, with 3.6 million followers. James Corden is an English actor, comedian, singer, writer, producer, and television host. He is known in the United Kingdom for cowriting and appearing in the BBC sitcom Gavin and Stacey. In the United States, Corden gained prominence for hosting The Late Show with James Corden, a late-night television program broadcast on CBS from 2015 to 2023.

The list of users also includes several individuals labeled as “Citizens” who were ordinary public members located in areas like England and Derbyshire. One such user has ties to both the United Kingdom and South Africa. These users had varying numbers of followers, ranging from 9 to 4875. Other users included a freelance writer, editor, and author with 10,620 followers, as well as a math teacher from England with 4053 followers. Both have managed to accumulate a noticeable following despite their different professions. The users who are not celebrities, such as the freelance writer as well as the math teacher from England, have still managed to garner a sizable following. This suggests that they have been posting engaging content that has captured the attention and imagination of their audience.

Time Series Analysis

Figure 4 provides a time series analysis of posting across the 54-day period. Most of the traffic appears to peak during the first phases of the lockdown, where Wicks was actively promoting his live streams.

Figure 4. Time series analysis of X activity related to Joe Wicks (March 23, 2020, to May 15, 2020).

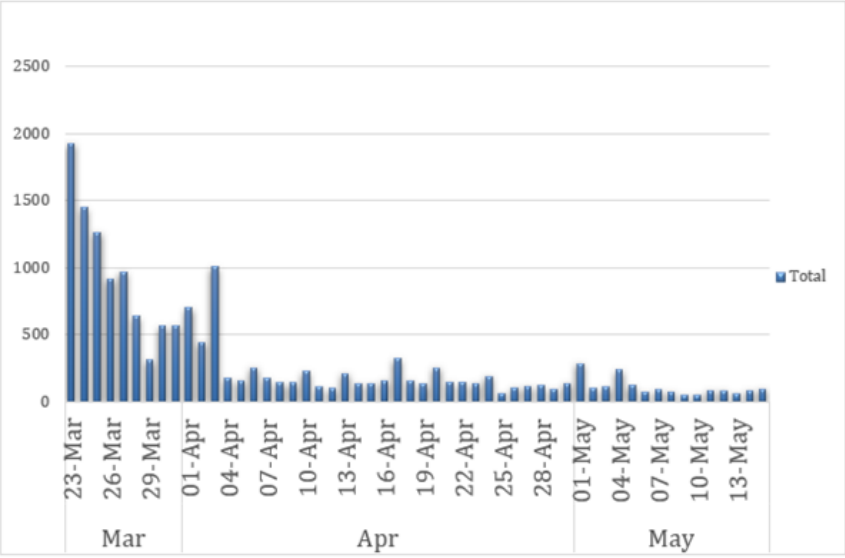

XSL•FO
RenderX

Results of Thematic Analysis

Overview of Themes

To set the context for the thematic analysis, a key post was sent by Joe Wicks on March 19, 2020, that kick-started the discussions, the post noted:

> *PE WITH JOE starting Monday morning at 9am on my YOUTUBE channel: The Body Coach TV Please please share this with as many people as you can Our kids need this more than ever. Share it on your stories, your wall, your Twitter, whatsapp and school newsletters*

This post called for and encouraged a YouTube fitness session run by Joe Wicks to be shared by others. Wicks highlighted that children would benefit the most from the videos and encouraged users to share the information widely, including school newsletters. The post above was quoted and reposted across March, and based on the number of reposts and likes, it appears this was a very successful approach in spreading the word of the online class.

Table 2 provides an overview of the key themes, their description, and examples of posts for each thematic heading. The most frequent type of posts was sent by parents who were showcasing how their children were engaging with the exercise activities from Wick live streams.

Table 2. Overview of key themes and description.

Theme	Description	Examples	Frequency, n (%)
Children taking part	Parents shared posts with images, text, or video showing children taking part and showing appreciation to Joe Wicks.	My children are absolutely LOVING their workouts! [VIDEO]	150 (37.5)
Other commentary and news articles	News articles on Joe Wicks and other general nonspecific commentary.	I wonder if there could be a workout for elderly and disabled who can't move as much?	110 (27.5)
Encouraging others to join in	Users encouraged others to take place by posting, reposting, and activity mentioning other users.	Hey, everyone! Be sure to get involved with this! [LINK]	90 (22.5)
Schools	Schools would post to encourage pupils to take part in the exercises.	PE classes just like school! Be sure to join in! [LINK]	50 (12.5)

Theme 1: Children Taking Part

In this theme, parents would share posts with images, text, or videos showing children taking part in Joe Wicks–related activities to express appreciation for Joe Wicks.

For instance, one user noted:

> *My children are absolutely LOVING their workouts!* [Link to VIDEO of children exercising]

Another user would go on to note:

> *I'm super proud of my children for taking part in Joe Wicks for their daily exercises* [Image of Children]

The theme showed that parents and caregivers observed the positive effects of physical activity and routine in promoting a sense of mindfulness among families and households that were struggling to manage their time effectively during the lockdown. This is consistent with the instant interconnectedness that SM facilitates and the influence of prominent public figures like Joe Wicks in promoting a sense of collective unity during the pandemic. This finding is consistent with the observations made by Zhu et al [65] and Durau et al [66] about the use of SM as a platform for disseminating narrative-based health content and amplifying health communication.

Theme 2: Other Commentary and News Articles

This theme largely contained news articles on Joe Wicks and other general nonspecific commentary. For instance, 1 comment from a user was as follows:

> *I wonder if there could be a workout for the elderly and disabled who can't move as much?*

There was a wide range of news articles commenting on the overall popularity of Joe Wicks, and additionally, media began to report on future ambitions; for instance, 1 post linked to a Daily Mail article that noted:

> *Joe Wicks Targets the US! The Highly Driven Body Coach Celebrity Plans to Conquer America*

The theme reinforced the public interest in health by exploring how the media reported this. More broadly, this theme engages with the proliferation of media coverage and speculation about the impact of this on health. The literature has largely focused on misleading media coverage of COVID-19, particularly discrimination and its impact on human rights [67]. However, alongside these important global social issues, there is also a sense of frivolity in the reporting and reaction to such articles on SM, which seeks to underscore the notoriety of public figures like Wicks in order to establish a global health narrative around exercise.

Theme 3: Encouraging Others to Join in

In this theme, users encouraged others to take place by posting, reposting, and actively mentioning other users. For instance, a user posted:

> *Hey, everyone! Be sure to get involved with this!* [LINK to Joe Wicks Content]

Other users were more lively and shared their own images; for instance, one user posted:

XSL•FO
RenderX

I've prepared my disco diva attire for the Joe Wicks gathering tomorrow. Have you chosen your ensemble? [Image of user in disco attire]

The theme demonstrated the interconnectedness within households in response to Wicks' exercise activities. The literature suggests that establishing connections can positively impact home dynamics, facilitate individuals' willingness to discuss personal health concerns, and initiate sustained improvements in individuals' health behaviors [68,69].

Theme 4: Schools

In this theme, schools would post to encourage pupils to participate in the exercises. For instance, 1 school account would post:

PE classes just like school! Be sure to join in! [LINK]

Another school posted:

Anyone for Physical Exercise featuring Joe Wicks?

The above post received several replies from parents who would share images of their children taking part and interacting with the school.

The theme highlighted the importance of health interactions and education, as schools played a vital role in advocating for children during the COVID-19 pandemic. Other researchers have also provided evidence of the transformative benefits and impact of education on health. Additionally, there is ongoing speculation about the global consequences of school closures and the effectiveness of information-seeking behaviors and public health messages following the reopening of schools [10,70,71].

Beyond the risks of COVID-19, the themes above also underscore the interest in more light-hearted and enjoyable aspects of health, particularly activities that households can do together. Common across the themes was the desire for connection and the impact of global interactions in supporting public health initiatives.

Results of Location Analysis

Figures 5-7 show a range of maps that document X activity around the world. The cyan-colored dots represent an X interaction (such as a post, repost, or quote post). Figure 5 provides a visual overview of the range of locations that were posted during this time. Due to the borderless reach of X, users around the world were engaging with content shared by Wicks.

Figure 6 provides an overview of X activity in the United Kingdom. It shows activity across the United Kingdom, with users clustered in Scotland and England.

Figure 7 provides insight into activity from X in the United States. It shows a geographical spread of X users across various states.

Figure 5. Geographical overview of users on X related to Joe Wicks (March 23, 2020, to May 15, 2020).

XSL•FO
RenderX

Figure 6. Posts across the United Kingdom related to Joe Wicks (March 23, 2020, to May 15, 2020).

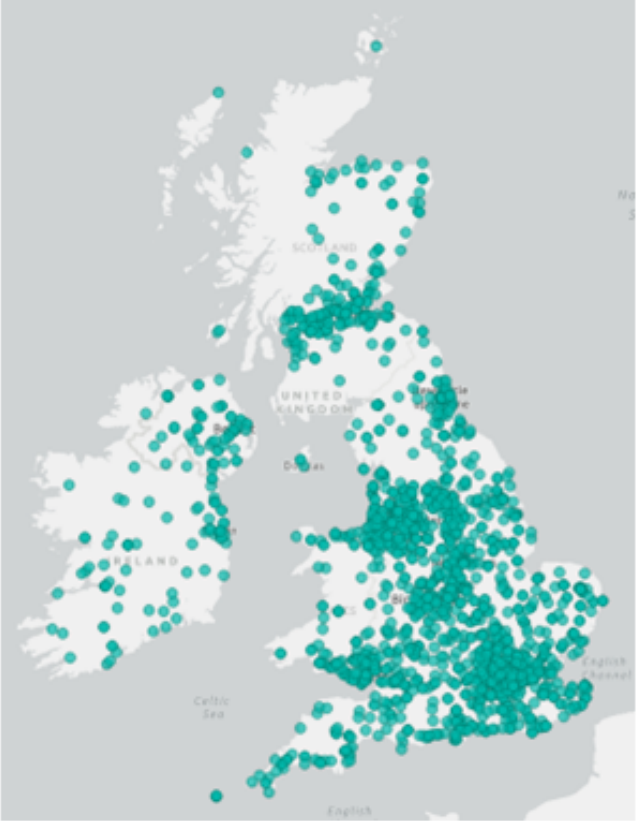

Figure 7. Posts across the United States related to Joe Wicks (March 23, 2020, to May 15, 2020).

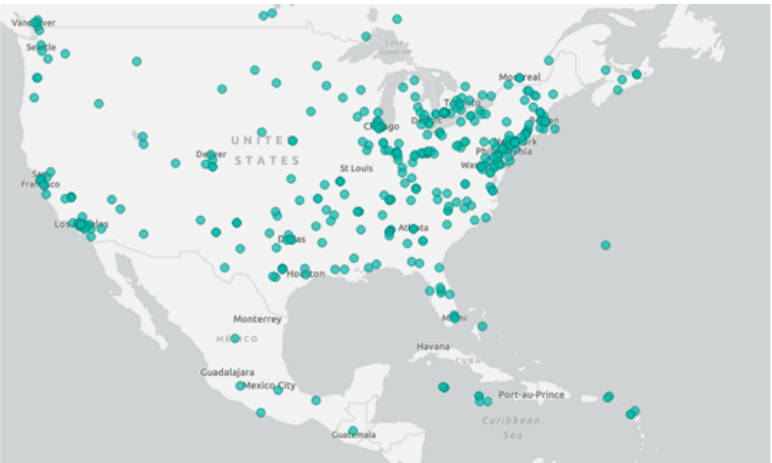

Discussion

Principal Findings and Implications

This study examined how much Joe Wicks, a fitness influencer, impacted health and health routines during the COVID-19 pandemic. We found that Wicks was a key influencer in motivating people to stay active and fit during this time. His network of influence was extensive and complex and included numerous subcommunities. We found that fitness influencers can create networks of influence exhibited in distinctive ways.

Wicks' SM accounts were being viewed by millions of people—domestically, internationally, and globally [72]. Indeed, reports of his workout sessions commonly began referring to him as "the nation's PE teacher." What is clear is that throughout the pandemic lockdown, Wicks had a profound effect on attitudes and behavior toward physical fitness, influencing large numbers of people to engage with and in physical activity to the extent that Britain has never seen before (at least in recent times). Wicks was awarded an MBE (Member of the Order of the British Empire) in 2020 for his live streams during the pandemic and a Guinness World Record for the most views on a fitness workout live on YouTube, with over 950,000 viewers [73].

Our findings support the hypothesis that Wicks positively impacted public health during the pandemic. His content encouraged people to engage in physical activity, an essential component of a healthy lifestyle. Such nudges toward health activities are especially important during times of crisis, when people may be more likely to neglect their health.

XSL•FO
RenderX

We found that ordinary citizens sharing content created by Joe Wicks were the key influencers and bridges in his network. Our results show that Wicks' network of influence resembled a broadcast network with several smaller subcommunities. A qualitative analysis of posts further highlighted how Wicks' content had encouraged younger people to continue with physical activity, which likely improved their overall health and well-being.

Our analysis shows that Joe Wicks was likely to have influenced physical exercise and activity across the world during the pandemic. Hence, the underpinning central tenant of this study is confirmed. As such, we conclude that SM, in this case, X, can serve both as an influencing platform in its own right and as a channel through which individuals can influence [74]. The relevance of social capital in relation to health is now well recognized in the social sciences, and theoretically informed studies on fitness influencers' promotion and portrayal of "healthy" bodies is a developing area of interest for health sociology and ill health studies [75,76]. With its long-standing empirical interest in social capital, health sociology is beginning to pay attention to the personalities and figures who have emerged because of the shift to neoliberal health policy. However, because this work is relatively new [77], important themes and areas of investigation are still being developed.

Our study has broader implications beyond the status of fitness influencers. Recognizing the critical role of individuals such as Joe Wicks in terms of health capital should be a critical area of inquiry for governments, public health authorities, and policy makers and mirrors the growing interest in health capital as part of embodied and digital experiences in everyday life. This is an especially important observation to make in Western cultural settings; given cynicism about government and the role of politicians, trust between them and the population is in question. Indeed, we suggest that such is this cynicism that some populations choose not to believe or engage with government messaging. On this basis, we contend that this study is an especially important one because it highlights the opportunities for governments, health officials, and policy makers to effect behavioral change. That is, rather than state entities seeking to communicate with and engage populations, we believe that SM influencers could be a more effective way of inducing behavioral change (on behalf of states and countries).

In addition to influencers and their influence, this study also highlights the role of significant others in SM networks. If Wicks has been the SM engine driving influence, then a community of other SM users has been the fuel that has built and sustained Wicks' influence. Given the scale of the SM ecosystem, we speculate that only a small number of what we label "ultra-influencers" are likely to have the reach and power to affect attitudinal and behavioral change (the likes of Hollywood actors and international music stars come to mind). As such, while we acknowledge the reach and power of Wicks, we nevertheless believe that his overall influence is an outcome of the aggregated influence of several others. Moreover, there is scope here to combine the ultrareputation of influencers such as Wicks with health technologies to further help individuals and households to make and sustain changes [78].

Limitations and Future Research

Our mixed methods approach offers significant analytic strength, providing a more robust understanding of Wicks' presentation and influence. SNA reveals how Wicks' presentation took hold within health narratives, while thematic analysis provides nuance to its nature.

Wicks' fitness influencer status has 2 leading roles: educator and motivator for health activities. The pandemic's national and global policy changes prevented a more robust longitudinal design with the thematic data. Future research could investigate how other fitness influencers adapt their presentation across multiple platforms and situations over time. Additionally, Wicks' daily fitness routines encapsulate his health strategy during a crisis rather than more robust health resources outside the pandemic. Therefore, while the analysis of Wicks' daily fitness routines allows us to deduce different types of engagement, we cannot be sure of the levels of interest in the depiction of fitness and how health is presented in its own right when it is part of influencer messaging.

We focused our research on Wicks' influencer-style performance of fitness in his shared content. Our findings suggest that the content allowed Wicks to successfully engage in a unique form of impression management applied to fitness and health strategies. However, Wicks' followers can also perform in healthy ways; this study provides a limited overview of how this manifested among Wicks' followers. We suspect that followers of fitness influencers would likely be qualitatively different. Because fitness tropes and health societal structures are embedded in individuals' lived experiences, how fitness is experienced would differ toward different aims and be expressed differently. We suggest this is a fruitful area for future research. Additionally, our mixed methods approach does not allow us to directly question followers about the broader impacts of their engagement with Wicks' content. Future interview-based research might seek to deepen the meanings and themes here. Future research might also investigate the role of fitness influencers in shaping the health performance of followers on SM.

A further limitation of the study is the timeframe examined. The research captures conversations on X related to a specific snapshot in time. Therefore limiting the ability to reflect changes in public behavior or opinion over a longer period. Consequently, the findings may not be generalizable over different periods. However, they would provide a useful comparison point for future studies examining other periods. Another limitation involves the challenges of conducting a qualitative analysis of brief posts. X's character limit at the time (280 characters) means that deep and nuanced insights may not emerge compared with interview-based data. Furthermore, the brevity of posts means that users often draw upon abbreviations, hashtags, and other condensed forms of communication that are not easily amenable to detailed qualitative interpretation. Furthermore, the study focused solely on X data, which could introduce biases because it may not accurately represent a larger population. X users may have specific demographic characteristics that are not fully representative of the population being studied. A further limitation of the study is the focus on

XSL•FO
RenderX

X interactions as the primary metric for gauging the effectiveness of motivating physical exercise during the pandemic. While this approach offers insights into public engagement, it may not fully capture the nuanced behaviors and long-term commitments of individuals who were influenced but did not express this online. For instance, some individuals may have been motivated to exercise regularly without posting their experiences on SM, while others may have expressed intent online but did not follow through offline. Future research could benefit from a multidimensional approach that combines SM analytics with in-person or survey-based methods to provide a more comprehensive understanding of the impact.

Conclusion

The fact that 1 person can influence the behavior of another is nothing new. What has changed in the last 2 decades is the rise of SM platforms like X. As a result, the nature of influence and the role that individuals and groups play as influencers has shifted. As a result, there is developing literature on this topic. We have contributed to this body of work by analyzing Joe Wicks' influence. This study demonstrates that individuals like Wicks can influence behavioral change. This alone justifies additional research in both related and unrelated circumstances concerning the impact of key persons of influence on health behavior change and fitness. Platforms such as X and popular accounts like "@thebodycoach," find considerable reach and influence through strong reputational links (for instance, increased user engagement or even trending topics).

There are various methods for gaining influence through SM platforms; in this study, Wicks stood out as a central node in a broadcast network. In turn, his impact has been maintained by the second wave of influencers. This influence hierarchy creates exciting research opportunities in this and other contexts. Additionally, meta-comments on health and fitness in this example reflect a shift in user motivation, driven in part by the feeling of digital engagement but also by the ability to impact real-time health behavior change during the Joe Wicks live-streamed training sessions. The collectivized activity of this type is an intriguing and significant area for research, combining digital and offline activities aimed at behavior modification and, we argue, social capital, but it requires additional investigation, and we conclude with such a call to action for other researchers working in this field.

Finally, we have seen how Wicks has been able to influence behavior in ways that governments and health officials frequently cannot. As a result, we urge that researchers and practitioners focus more on influence and influencers in order to improve the effectiveness of public health policies and communications. After all, if an online fitness instructor can persuade millions to exercise on a daily basis, other such influencers should be able to have similar effects in a variety of other contexts.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Gostin LO, Hodge JG, Bloom BR, El-Mohandes A, Fielding J, Hotez P, et al. The public health crisis of underimmunisation: a global plan of action. Lancet Infect Dis 2020;20(1):e11-e16. [doi: 10.1016/S1473-3099(19)30558-4] [Medline: 31706795]
2. Campbell N, Browne S, Claudy M, Mialon M, Hercberg S, Goiana-da-Silva F, et al. The gift of data: industry-led food reformulation and the obesity crisis in Europe. J Public Policy Mark 2021;40(3):389-402. [doi: 10.1177/0743915620983842]
3. Rowley WR, Bezold C, Arikan Y, Byrne E, Krohe S. Diabetes 2030: insights from yesterday, today, and future trends. Popul Health Manag 2017;20(1):6-12 [FREE Full text] [doi: 10.1089/pop.2015.0181] [Medline: 27124621]
4. Stiles MK, Wilde AAM, Abrams DJ, Ackerman MJ, Albert CM, Behr ER, et al. 2020 APHRS/HRS expert consensus statement on the investigation of decedents with sudden unexplained death and patients with sudden cardiac arrest, and of their families. Heart Rhythm 2021;18(1):e1-e50 [FREE Full text] [doi: 10.1016/j.hrthm.2020.10.010] [Medline: 33091602]
5. Russell JC, Flood VM, Yeatman H, Wang JJ, Mitchell P. Food insecurity and poor diet quality are associated with reduced quality of life in older adults. Nutr Diet 2016;73(1):50-58. [doi: 10.1111/1747-0080.12263]
6. Tudor-Locke C, Craig CL, Thyfault JP, Spence JC. A step-defined sedentary lifestyle index: less than 5000 steps/day. Appl Physiol Nutr Metab 2013 Feb;38(2):100-114 [FREE Full text] [doi: 10.1139/apnm-2012-0235] [Medline: 23438219]
7. Medeiros SSDE, Enes CC, Nucci LB. Association of weight perception, body satisfaction, and weight loss intention with patterns of health risk behaviors in adolescents with overweight and obesity. Behav Med 2023;49(3):302-311. [doi: 10.1080/08964289.2022.2057407] [Medline: 35465848]
8. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed 2020;91(1):157-160 [FREE Full text] [doi: 10.23750/abm.v91i1.9397] [Medline: 32191675]
9. Huberty J, Dinkel D, Coleman J, Beighle A, Apenteng B. The role of schools in children's physical activity participation: staff perceptions. Health Educ Res 2012;27(6):986-995 [FREE Full text] [doi: 10.1093/her/cys071] [Medline: 22717940]

XSL•FO
RenderX

10. Van Lancker W, Parolin Z. COVID-19, school closures, and child poverty: a social crisis in the making. Lancet Public Health 2020;5(5):e243-e244 [FREE Full text] [doi: 10.1016/S2468-2667(20)30084-0] [Medline: 32275858]
11. Sanders K, Crespo MJC, Holtz-Bacha C. Communicating governments: a three-country comparison of how governments communicate with citizens. Int J Press/Politics 2011;16(4):523-547. [doi: 10.1177/1940161211418225]
12. Wukich C, Mergel I. Closing the citizen-government communication gap: content, audience, and network analysis of government tweets. J Homel Secur Emerg Manag 2015;12(3):707-735.
13. Lyu K, Kim H. Sentiment analysis using word polarity of social media. Wireless Pers Commun 2016;89(3):941-958. [doi: 10.1007/s11277-016-3346-1]
14. Heldman AB, Schindelar J, Weaver JB. Social media engagement and public health communication: implications for public health organizations being truly "social". Public Health Rev 2013;35(1):13. [doi: 10.1007/bf03391698]
15. Number of social media users worldwide from 2017 to 2027. Statista. 2023. URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [accessed 2023-06-13]
16. González-Padilla DA, Tortolero-Blanco L. Social media influence in the COVID-19 pandemic. Int Braz J Urol 2020;46(suppl.1):120-124 [FREE Full text] [doi: 10.1590/S1677-5538.IBJU.2020.S121] [Medline: 32550706]
17. Heinze A, Fletcher G, Cruz A, Rashid T. Digital and Social Media Marketing: A Results-Driven Approach, Second Edition. London, England: Routledge; 2020.
18. George DR, Rovniak LS, Kraschnewski JL. Dangers and opportunities for social media in medicine. Clin Obstet Gynecol 2013;56(3):453-462 [FREE Full text] [doi: 10.1097/GRF.0b013e318297dc38] [Medline: 23903375]
19. Etter M, Ravasi D, Colleoni E. Social media and the formation of organizational reputation. Acad Manag 2019;44(1):28-52. [doi: 10.5465/amr.2014.0280]
20. Staccini P, Lau AYS, Section Editors for the IMIA Yearbook Section on Consumer Health Informatics and Education. Social media, research, and ethics: does participant willingness matter? Yearb Med Inform 2020;29(1):176-183 [FREE Full text] [doi: 10.1055/s-0040-1702022] [Medline: 32823313]
21. Nguyen A, Catalan-Matamoros D. Digital mis/disinformation and public engagement with health and science controversies: fresh perspectives from Covid-19. Media Commun 2020;8(2):323-328 [FREE Full text] [doi: 10.17645/mac.v8i2.3352]
22. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015;22(1):243-256 [FREE Full text] [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
23. Neiger BL, Thackeray R, Burton SH, Giraud-Carrier CG, Fagen MC. Evaluating social media's capacity to develop engaged audiences in health promotion settings: use of Twitter metrics as a case study. Health Promot Pract 2013;14(2):157-162. [doi: 10.1177/1524839912469378] [Medline: 23271716]
24. Chou WYS, Prestin A, Lyons C, Wen KY. Web 2.0 for health promotion: reviewing the current evidence. Am J Public Health 2013;103(1):e9-e18 [FREE Full text] [doi: 10.2105/AJPH.2012.301071] [Medline: 23153164]
25. Pagoto S, Waring ME, Xu R. A call for a public health agenda for social media research. J Med Internet Res 2019;21(12):e16661 [FREE Full text] [doi: 10.2196/16661] [Medline: 31855185]
26. Henwood F, Marent B. Understanding digital health: productive tensions at the intersection of sociology of health and science and technology studies. Sociol Health Illn 2019;41(Suppl 1):1-15 [FREE Full text] [doi: 10.1111/1467-9566.12898] [Medline: 31599984]
27. Booth N, Matic JA. Mapping and leveraging influencers in social media to shape corporate brand perceptions. Corp Commun Int J 2011;16(3):184-191. [doi: 10.1108/13563281111156853]
28. Biran O, Rosenthal S, Andreas J, McKeown K, Rambow O. Detecting influencers in written online conversations. Columbia University. 2012. URL: https://academiccommons.columbia.edu/doi/10.7916/D88341DV [accessed 2024-03-13]
29. Khamis S, Ang L, Welling R. Self-branding, 'micro-celebrity' and the rise of social media influencers. Celebrity Studies 2016;8(2):191-208. [doi: 10.1080/19392397.2016.1218292]
30. Gorry GA, Westbrook RA. Winning the internet confidence game. Corp Reputation Rev 2009;12(3):195-203. [doi: 10.1057/crr.2009.16]
31. Freberg K, Graham K, McGaughey K, Freberg LA. Who are the social media influencers? a study of public perceptions of personality. Public Relat Rev 2011;37(1):90-92. [doi: 10.1016/j.pubrev.2010.11.001]
32. De Veirman M, Cauberghe V, Hudders L. Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude. Int J Advert 2017;36(5):798-828. [doi: 10.1080/02650487.2017.1348035]
33. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
34. Jürgens P. Communities of communication: making sense of the "social" in social media. In: Bredl K, Hunniger J, Jensen JL, editors. Methods for Analyzing Social Media. London: Routledge; 2017:45-59.
35. Ghani NA, Hamid S, Hashem IAT, Ahmed E. Social media big data analytics: a survey. Comput Human Behav 2019;101:417-428. [doi: 10.1016/j.chb.2018.08.039]

XSL•FO
RenderX

36. Valkenburg PM, Koutamanis M, Vossen HGM. The concurrent and longitudinal relationships between adolescents' use of social network sites and their social self-esteem. Comput Human Behav 2017;76:35-41 [FREE Full text] [doi: 10.1016/j.chb.2017.07.008] [Medline: 29104364]
37. Allen C, Vassilev I, Kennedy A, Rogers A. The work and relatedness of ties mediated online in supporting long-term condition self-management. Sociol Health Illn 2020;42(3):579-595 [FREE Full text] [doi: 10.1111/1467-9566.13042] [Medline: 31769045]
38. Gillett J. Media activism and internet use by people with HIV/AIDS. Sociol Health Illn 2003;25(6):608-624 [FREE Full text] [doi: 10.1111/1467-9566.00361] [Medline: 12919448]
39. Hardey MM. On the body of the consumer: performance-seeking with wearables and health and fitness apps. Sociol Health Illn 2019;41(6):991-1004 [FREE Full text] [doi: 10.1111/1467-9566.12879] [Medline: 30834568]
40. Hjorth L, Lupton D. Digitised caring intimacies: more-than-human intergenerational care in Japan. Int J Cult Stud 2021;24(4):584-602. [doi: 10.1177/1367877920927427]
41. Weinstein E, Kleiman EM, Franz PJ, Joyce VW, Nash CC, Buonopane RJ, et al. Positive and negative uses of social media among adolescents hospitalized for suicidal behavior. J Adolesc 2021;87:63-73. [doi: 10.1016/j.adolescence.2020.12.003] [Medline: 33493982]
42. Dowrick A, Mitchinson L, Hoernke K, Symmons SM, Cooper S, Martin S, et al. Re-ordering connections: UK healthcare workers' experiences of emotion management during the COVID-19 pandemic. Sociol Health Illn 2021;43(9):2156-2177 [FREE Full text] [doi: 10.1111/1467-9566.13390] [Medline: 34706107]
43. Beaudoin CE, Tao CC. Benefiting from social capital in online support groups: an empirical study of cancer patients. Cyberpsychol Behav 2007;10(4):587-590. [doi: 10.1089/cpb.2007.9986] [Medline: 17711369]
44. Brewer RN, Schoenebeck S, Lee K, Suryadevara H. Challenging passive social media use: older adults as caregivers online. Proc ACM Hum-Comput Interact 2021;5(CSCW1):1-20. [doi: 10.1145/3449197]
45. Nguyen MH, Hunsaker A, Hargittai E. Older adults' online social engagement and social capital: the moderating role of internet skills. Inf Commun Soc 2020;25(7):942-958. [doi: 10.1080/1369118x.2020.1804980]
46. Fenton A, Gillooly L, Vasilica CM. Female fans and social media: micro-communities and the formation of social capital. Eur Sport Manag Q 2021;23(2):370-390. [doi: 10.1080/16184742.2020.1868546]
47. Pitas N, Ehmer C. Social capital in the response to COVID-19. Am J Health Promot 2020;34(8):942-944 [FREE Full text] [doi: 10.1177/0890117120924531] [Medline: 32394721]
48. Auxier B, Buntain C, Golbeck J. Analyzing sentiment and themes in fitness influencers' twitter dialogue. In: Taylor NG, Christian-Lamb C, Martin MH, Nardi B, editors. Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings. Cham: Springer International Publishing; 2019:429-435.
49. Puntha P, Jitanugoon S, Lee PC. Engagement on social networks during the COVID-19 pandemic: a comparison among healthcare professionals, fitness influencers, and healthy food influencers. New York, NY, US: Association for Computing Machinery; 2021 Presented at: Proceedings of the 8th Multidisciplinary International Social Networks; November 15-17, 2021; Bergen, Norway p. 5-14. [doi: 10.1145/3504006.3504008]
50. Fortuna KL, Brooks JM, Umucu E, Walker R, Chow PI. Peer support: a human factor to enhance engagement in digital health behavior change interventions. J Technol Behav Sci 2019;4(2):152-161 [FREE Full text] [doi: 10.1007/s41347-019-00105-x] [Medline: 34337145]
51. Merchant RM, Elmer S, Lurie N. Integrating social media into emergency-preparedness efforts. N Engl J Med 2011;365(4):289-291 [FREE Full text] [doi: 10.1056/NEJMp1103591] [Medline: 21793742]
52. Williams HTP, McMurray JR, Kurz T, Lambert FH. Network analysis reveals open forums and echo chambers in social media discussions of climate change. Glob Environ Change 2015;32:126-138 [FREE Full text] [doi: 10.1016/j.gloenvcha.2015.03.006]
53. Ahmed W, Önkal D, Das R, Krishnan S, Olan F, Hardey M, et al. Developing techniques to support technological solutions to disinformation by analyzing four conspiracy networks during COVID-19. IEEE Trans Eng Manage 2023:1-18. [doi: 10.1109/tem.2023.3273191]
54. Ahmed W, Vidal-Alaball J, Downing J, Seguí FL. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. J Med Internet Res 2020;22(5):e19458 [FREE Full text] [doi: 10.2196/19458] [Medline: 32352383]
55. Ahmed W, Seguí FL, Vidal-Alaball J, Katz MS. COVID-19 and the "film your hospital" conspiracy theory: social network analysis of Twitter data. J Med Internet Res 2020;22(10):e22374 [FREE Full text] [doi: 10.2196/22374] [Medline: 32936771]
56. Ahmed W, Lugovic S. Social media analytics: analysis and visualisation of news diffusion using NodeXL. Online Inf Rev 2019;43(1):149-160. [doi: 10.1108/oir-03-2018-0093]
57. Smith MA, Rainie L, Shneiderman B, Himelboim I. Mapping Twitter topic networks: from polarized crowds to community clusters. Pew Research Center. 2014. URL: https://policycommons.net/artifacts/620393/mapping-twitter-topic-networks/1601593/ [accessed 2022-10-25]
58. Himelboim I, Smith MA, Rainie L, Shneiderman B, Espina C. Classifying Twitter topic-networks using social network analysis. Soc Media Soc 2017;3(1):1-13 [FREE Full text] [doi: 10.1177/2056305117691545]

XSL•FO
RenderX

59. Matei S. Analyzing social media networks with NodeXL: insights from a connected world by Derek Hansen, Ben Shneiderman, and Marc A. Smith. Int J Hum-Comput Interact 2011;27(4):405-408. [doi: 10.1080/10447318.2011.544971]
60. Borgatti SP. Centrality and network flow. Soc Netw 2005;27(1):55-71. [doi: 10.1016/j.socnet.2004.11.008]
61. Burt RS. Brokerage and Closure: An Introduction to Social Capital. Oxford, UK: Oxford University Press; 2007.
62. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
63. Ando H, Cousins R, Young C. Achieving saturation in thematic analysis: development and refinement of a codebook. Compr Psychol 2014;3:1-7 [FREE Full text] [doi: 10.2466/03.CP.3.4]
64. Heinze A, Fletcher G, Rashid T, Cruz A. Digital and Social Media Marketing: A Results-Driven Approach, Second Edition. London, England: Routledge; 2020.
65. Zhu C, Xu X, Zhang W, Chen J, Evans R. How health communication via Tik Tok makes a difference: a content analysis of Tik Tok accounts run by Chinese provincial health committees. Int J Environ Res Public Health 2019;17(1):192 [FREE Full text] [doi: 10.3390/ijerph17010192] [Medline: 31892122]
66. Durau J, Diehl S, Terlutter R. Motivate me to exercise with you: the effects of social media fitness influencers on users' intentions to engage in physical activity and the role of user gender. Digit Health 2022;8:20552076221102769 [FREE Full text] [doi: 10.1177/20552076221102769] [Medline: 35615268]
67. Wen J, Aston J, Liu X, Ying T. Effects of misleading media coverage on public health crisis: a case of the 2019 novel coronavirus outbreak in China. Anatolia 2020;31(2):331-336. [doi: 10.1080/13032917.2020.1730621]
68. Hardey MM. Tracking the trackers: self-tracking in households as social practice. Digit Health 2022;8:20552076221093131 [FREE Full text] [doi: 10.1177/20552076221093131] [Medline: 35433023]
69. Hardey M. Household Self-Tracking During a Global Health Crisis: Shaping Bodies, Lives, Health and Illness. Bingley, UK: Emerald Publishing Limited; 2022.
70. Singu S, Acharya A, Challagundla K, Byrareddy SN. Impact of social determinants of health on the emerging COVID-19 pandemic in the United States. Front Public Health 2020;8:406 [FREE Full text] [doi: 10.3389/fpubh.2020.00406] [Medline: 32793544]
71. Masonbrink AR, Hurley E. Advocating for children during the COVID-19 school closures. Pediatrics 2020;146(3):e20201440 [FREE Full text] [doi: 10.1542/peds.2020-1440] [Medline: 32554517]
72. Windsor S. Joe Wicks got children moving—how to keep them active as lockdown ends. The Conversation. 2020. URL: http://theconversation.com/joe-wicks-got-children-moving-how-to-keep-them-active-as-lockdown-ends-140591 [accessed 2023-06-13]
73. Joe Wicks' PE with Joe smashes YouTube livestream record. Guinness World Records. 2020. URL: https://www.guinnessworldrecords.com/news/2020/4/joe-wicks-pe-with-joe-smashes-youtube-livestream-record-614934 [accessed 2023-06-13]
74. Sokolova K, Perez C. You follow fitness influencers on YouTube. but do you actually exercise? how parasocial relationships, and watching fitness influencers, relate to intentions to exercise. J Retail Consum Serv 2021;58:102276. [doi: 10.1016/j.jretconser.2020.102276]
75. Arena R, Myers J, Ozemek C, Hall G, Severin R, Laddu D, et al. An evolving approach to assessing cardiorespiratory fitness, muscle function and bone and joint health in the COVID-19 era. Curr Probl Cardiol 2022;47(1):100879 [FREE Full text] [doi: 10.1016/j.cpcardiol.2021.100879] [Medline: 34103194]
76. Enamela P, Prybutok GL, Dake JL, Prybutok VR. A case study on the influence of fitness trackers on older adults' health and fitness activities. Int J Electron Healthc 2021;11(4):348-363. [doi: 10.1504/ijeh.2021.117828]
77. Goodyear VA, Boardley I, Chiou SY, Fenton SAM, Makopoulou K, Stathi A, et al. Social media use informing behaviours related to physical activity, diet and quality of life during COVID-19: a mixed methods study. BMC Public Health 2021;21(1):1333 [FREE Full text] [doi: 10.1186/s12889-021-11398-0] [Medline: 34229651]
78. Beh PK, Ganesan Y, Iranmanesh M, Foroughi B. Using smartwatches for fitness and health monitoring: the UTAUT2 combined with threat appraisal as moderators. Behav Inf Technol 2019;40(3):282-299. [doi: 10.1080/0144929x.2019.1685597]

Abbreviations

NLP: natural language processing
SM: social media
SNA: social network analysis

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 13.06.23; peer-reviewed by K Sellen, W Al-Chetachi, D Dinh; comments to author 13.09.23; revised version received 15.09.23; accepted 06.03.24; published 29.03.24.

Please cite as:
Ahmed W, Aiyenitaju O, Chadwick S, Hardey M, Fenton A
The Influence of Joe Wicks on Physical Activity During the COVID-19 Pandemic: Thematic, Location, and Social Network Analysis of X Data
J Med Internet Res 2024;26:e49921
URL: https://www.jmir.org/2024/1/e49921
doi:10.2196/49921
PMID:38551627

©Wasim Ahmed, Opeoluwa Aiyenitaju, Simon Chadwick, Mariann Hardey, Alex Fenton. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Discourse, User Reactions, and Conspiracy Theories on the X Platform About HIV Vaccines: Data Mining and Content Analysis

Jueman M Zhang[1], PhD; Yi Wang[2], PhD; Magali Mouton[3]; Jixuan Zhang[4]; Molu Shi[5], PhD

[1]Harrington School of Communication and Media, University of Rhode Island, Kingston, RI, United States
[2]Department of Communication, University of Louisville, Louisville, KY, United States
[3]School of Rehabilitation Sciences, University of Ottawa, Ottawa, ON, Canada
[4]Polk School of Communications, Long Island University, Brooklyn, NY, United States
[5]College of Business, University of Louisville, Louisville, KY, United States

Corresponding Author:
Jueman M Zhang, PhD
Harrington School of Communication and Media
University of Rhode Island
10 Ranger Road
Kingston, RI, 02881
United States
Phone: 1 401 874 2110
Email: jueman.zhang@uri.edu

Abstract

Background: The initiation of clinical trials for messenger RNA (mRNA) HIV vaccines in early 2022 revived public discussion on HIV vaccines after 3 decades of unsuccessful research. These trials followed the success of mRNA technology in COVID-19 vaccines but unfolded amid intense vaccine debates during the COVID-19 pandemic. It is crucial to gain insights into public discourse and reactions about potential new vaccines, and social media platforms such as X (formerly known as Twitter) provide important channels.

Objective: Drawing from infodemiology and infoveillance research, this study investigated the patterns of public discourse and message-level drivers of user reactions on X regarding HIV vaccines by analyzing posts using machine learning algorithms. We examined how users used different post types to contribute to topics and valence and how these topics and valence influenced like and repost counts. In addition, the study identified salient aspects of HIV vaccines related to COVID-19 and prominent anti–HIV vaccine conspiracy theories through manual coding.

Methods: We collected 36,424 English-language original posts about HIV vaccines on the X platform from January 1, 2022, to December 31, 2022. We used topic modeling and sentiment analysis to uncover latent topics and valence, which were subsequently analyzed across post types in cross-tabulation analyses and integrated into linear regression models to predict user reactions, specifically likes and reposts. Furthermore, we manually coded the 1000 most engaged posts about HIV and COVID-19 to uncover salient aspects of HIV vaccines related to COVID-19 and the 1000 most engaged negative posts to identify prominent anti–HIV vaccine conspiracy theories.

Results: Topic modeling revealed 3 topics: HIV and COVID-19, mRNA HIV vaccine trials, and HIV vaccine and immunity. HIV and COVID-19 underscored the connections between HIV vaccines and COVID-19 vaccines, as evidenced by subtopics about their reciprocal impact on development and various comparisons. The overall valence of the posts was marginally positive. Compared to self-composed posts initiating new conversations, there was a higher proportion of HIV and COVID-19–related and negative posts among quote posts and replies, which contribute to existing conversations. The topic of mRNA HIV vaccine trials, most evident in self-composed posts, increased repost counts. Positive valence increased like and repost counts. Prominent anti–HIV vaccine conspiracy theories often falsely linked HIV vaccines to concurrent COVID-19 and other HIV-related events.

Conclusions: The results highlight COVID-19 as a significant context for public discourse and reactions regarding HIV vaccines from both positive and negative perspectives. The success of mRNA COVID-19 vaccines shed a positive light on HIV vaccines. However, COVID-19 also situated HIV vaccines in a negative context, as observed in some anti–HIV vaccine conspiracy theories

XSL•FO
RenderX

misleadingly connecting HIV vaccines with COVID-19. These findings have implications for public health communication strategies concerning HIV vaccines.

(J Med Internet Res 2024;26:e53375) doi:10.2196/53375

KEYWORDS

HIV; vaccine; Twitter; X platform; infodemiology; machine learning; topic modeling; sentiment; conspiracy theory; COVID-19

Introduction

Background

Vaccination has long been recognized as a crucial preventive measure against diseases and infections, but opposition to vaccines has endured [1]. HIV vaccination has been regarded as a potential preventive measure to help combat the HIV epidemic in the United States, with research and progress dating back to the mid-1980s but without success thus far [2]. An estimated 1.2 million people were living with HIV in the United States by the end of 2021, with 36,136 new HIV diagnoses reported in 2021 [3].

On January 27, 2022, the biotechnology company Moderna announced the initiation of clinical trials for an HIV vaccine using messenger RNA (mRNA) technology [4]. In March 2022, the National Institutes of Health announced the start of clinical trials for 3 mRNA HIV vaccines [5]. The mRNA technology had previously been used in the Pfizer-BioNTech and Moderna COVID-19 vaccines, which protected individuals against severe symptoms and fatalities during the pandemic [6]. Following the successes of mRNA COVID-19 vaccines, which led to the Nobel Prize in Physiology or Medicine being awarded to 2 scientists in October 2023 [7], researchers have been investigating the potential of mRNA vaccines for various other diseases, including HIV [8,9]. The announcements of clinical trials for mRNA HIV vaccines revived public discussion on the prospect of vaccines to combat HIV [9] despite >3 decades of unsuccessful research [2]. Meanwhile, these announcements were made against the backdrop of intense vaccine debates during the COVID-19 pandemic, with misinformation and conspiracy theories fueling vaccine hesitancy [10-12].

The X platform, formerly known as Twitter, has been a significant social media platform and a vital source for text-based public discourse. Posts on X have been studied to understand public discourse about vaccines in general [13-15] and about specific vaccines, such as COVID-19 vaccines in recent years [12,16,17]. However, there is a dearth of research about public discourse on HIV vaccines on social media. Given the advancement in mRNA technology in COVID-19 vaccines and heated vaccine debates, it has become especially important to gain insights into public discourse and reactions regarding potential new vaccines.

This study is grounded in the growing field of infodemiology and infoveillance, which investigates the "distribution and determinants of information in an electronic medium," specifically on the web, by analyzing unstructured text with the aim of informing public health practices or serving surveillance objectives [18]. In recent infodemiology and infoveillance studies, machine learning algorithms have been increasingly used to examine substantial amounts of social media content, such as posts on X related to COVID-19 vaccines [12,16,17] and HIV prevention [19], to extract insights into public discourse and reactions. These algorithms automatically analyze extensive volumes of posts and capture latent textual information such as topics and sentiments. This study aimed to investigate how users used different post types to contribute original content to topics and valence identified through machine learning algorithms and how these topics and valence affected user reactions on X regarding HIV vaccines. In addition, by manually coding the most engaged posts, similar to an approach used in previous infodemiology and infoveillance research [20], the study intended to identify salient aspects of HIV vaccines related to COVID-19 as well as prominent anti–HIV vaccine conspiracy theories. Analyzing posts on X about HIV vaccines can shed light on public discourse and information diffusion. These findings have implications for shaping public health communication strategies about HIV vaccines [18]. Furthermore, the findings may help in understanding the acceptability of the HIV vaccine upon its successful development in comparison with adherence to existing HIV prevention measures. Previous infodemiology and infoveillance research effectively increased the forecast accuracy of COVID-19 vaccine uptake by leveraging insights derived from posts on X [21].

Literature Review

Public Discourse About HIV Prevention on X

Social media platforms have become important channels for HIV communication, enabling the dissemination of and engagement with content encompassing a wide array of issues related to HIV prevention, treatment, coping, and available resources [22,23]. An earlier infodemiology study examined 69,197 posts on the X platform containing the hashtag #HIVPrevention between 2014 and 2019 and categorized these posts into 10 identified topics concerning HIV prevention [19]. Among them, pre-exposure prophylaxis had the highest representation with 13,895 posts, followed by HIV testing; condoms; harm reduction; gender equity and violence against women; voluntary medical male circumcision; sex work; postexposure prophylaxis; elimination of mother-to-child transmission of HIV; and abstinence, which had the lowest representation with 180 posts. Furthermore, that study suggested a consistency between the volume of posts related to specific HIV prevention measures on X over time and the temporal trends in the uptake of those measures [19]. It is noteworthy that the topic of HIV vaccines was absent, which suggested minimal public discourse on the topic during these years. This may be associated with the extensive history of unsuccessful research in this area [2].

XSL•FO
RenderX

Despite the availability of current HIV prevention measures, efforts have been made to develop HIV vaccines, which are considered necessary to bridge the gap between the challenges in adhering to existing HIV prevention measures and the ambitious goal set by United Nations member states to end the HIV epidemic by 2030 [24,25]. The surge in public discussion about HIV vaccines, possibly elicited by the clinical trials for mRNA HIV vaccines [9], presented an optimal opportunity to investigate public discourse and reactions regarding HIV vaccines. To the best of our knowledge, this is the first study to analyze posts on X about HIV vaccines.

Public Discourse and Post Types on X

On the X platform, public discourse featuring original content can be observed through 3 post types: self-composed posts, quote posts, and replies [26]. X users can compose a post. They can also create a quote post, which entails reposting a post while adding their comments. In addition, they can reply to a post to share their comments [26]. While self-composed posts initiate new conversations, quote posts and replies enable users to join existing conversations by contributing their own comments [27]. The Pew Research Center's analysis of survey respondents' posts on X from October 2022 to April 2023 revealed the composition of different types of posts. Regarding the 3 types of posts containing original content, replies accounted for the highest proportion at 40%, followed by self-composed posts at 15% and quote posts at 9%. The remaining 35% were reposts [28].

Machine learning algorithms have been increasingly used in recent years to identify latent message features, including textual topics and sentiment valence, among vast numbers of social media posts, as exemplified by previous research analyzing posts on X about COVID-19 vaccines [12,16,17] and HIV prevention [19]. However, the patterns of public discourse in social media conversations are unclear. Specifically, there is a scarcity of research on how people contribute their original content to topics and valence related to a public health issue. This study aimed to address this gap by examining the relationship between post types and message features, specifically topics and valence uncovered using machine learning algorithms, with a focus on HIV vaccines as the subject matter. The findings will advance our knowledge of user contributions to social media conversations about HIV vaccines.

Message Features Influencing User Reactions on X

Examining message diffusion on social media has been a multifaceted challenge, especially with vaccines being a contentious issue debated fervently during the COVID-19 pandemic [16]. Another contribution of this study is to advance this research area by using machine learning to investigate the synergistic impact of content and account features on user reactions regarding a potential new vaccine amid the context of intense vaccine debates.

The extent to which a message results in optimal diffusion on social media can be gauged by user reactions [16,29-31]. On X, a user can engage with posts—be it a self-composed post, quote post, or reply—in 2 primary 1-click reactions: liking and reposting [26]. An X user can like a post to show appreciation for it or repost it to share it publicly. Compared to liking, reposting is a more social behavior [16,32]. Unlike X's old timeline, which mostly displayed posts from accounts that a user followed, its current "For you" timeline also shows posts that those accounts have engaged with along with other posts recommended based on user reactions [33]. The nature of promoting posts based on user reactions makes it more important to investigate the factors that influence user reactions.

This study investigated 2 categories of message-level features that, according to previous research, can drive user interactions: content features in terms of topics and valence and account features in terms of user verification and follower count. Post topics affect likes and reposts on X [16,30,34]. Previous research on COVID-19 vaccine posts on X has indicated that posts containing useful information garner more likes and reposts [16]. This is likely because information utility fills people's knowledge gaps and serves their utilitarian needs in the face of health risks [16,32,34-36]. In addition, previous studies have suggested that the novelty of useful information further facilitates sharing of digital health information [32,36], such as updates about COVID-19 vaccine development [12]. Given the initial success of mRNA technology in COVID-19 vaccines, mRNA HIV vaccine candidates may possess the inherent features of prospective usefulness and ongoing novelty. As a result, posts presenting pertinent information have the potential to generate more likes and reposts. Meanwhile, the announcements of clinical trials for mRNA HIV vaccines were made amidst intense vaccine debates during the COVID-19 pandemic [12]. Previous research has shown that perceived controversiality in health information increases viewership but not sharing on social media [32]. In the context of the heated controversy surrounding vaccines, it is crucial to understand user reactions to new potential vaccines.

In addition to post topics, post valence can play a role in user reactions [34]. Past research has generally revealed that there are more positive than negative posts on X about vaccines in general [13-15] and, more recently, about COVID-19 vaccines in particular [12,16,17]. However, the influence of post valence on user reactions remains unclear. One study on COVID-19 vaccines showed that positive posts on X received more likes but not more reposts [16]. Another study on vaccines regardless of their type revealed that antivaccine posts garnered more reposts than provaccine posts on X [13]. A psychological rationale supporting the social transmission of positive content is the motivation of individuals to present themselves positively and shape their self-identity [35,37]. In comparison, social transmission of negative content can be attributed to the idea that certain negative content triggers activation, which drives user reactions [35].

Furthermore, previous research has shown that account features such as verification status and follower count affect user reactions on social media [13,16,34]. Given the vast amounts of information available in the digital age, the authenticity of user accounts becomes crucial in the diffusion of health information. One study revealed that account verification enhanced the number of likes and reposts for posts about COVID-19 vaccines on X [16]. Another study indicated that

XSL•FO
RenderX

follower counts increased the number of reposts for posts about vaccines on X regardless of vaccine type [13].

Conspiracy Theories

A conspiracy theory refers to the belief that a coalition of powerholders forms secret agreements with malevolent intentions [38,39]. It differs from other types of misinformation by hypothesizing a pattern in which people, objects, or events are interconnected in a causal manner [39]. Previous research has revealed conspiracy theories as a salient theme in antivaccine discourse on social media, along with other themes such as side effects and inefficacy [40,41]. For HIV vaccines, conspiracy theories are crucial in understanding public discourse against them given the limited information about side effects and inefficacy until future success. An additional contribution of this study is the identification of prominent anti–HIV vaccine conspiracy theories through manual coding of the most engaged with negative posts.

Antivaccine conspiracy theories contribute to vaccine hesitancy [42-44], as observed recently with COVID-19 vaccines [10,11]. Understanding the themes and reasoning behind antivaccine conspiracy theories will provide vital implications for deploying evidence-based and logic-driven strategies to counter them [45-47]. A systematic review of antivaccine discourse on social media from 2015 to 2019 revealed pre–COVID-19 conspiracy theories [41]. These theories claimed that powerholders promoted vaccines for self-serving interests, including hiding vaccine side effects for financial gain and controlling society and the population [40,41]. During the COVID-19 pandemic, antivaccine conspiracy theories thrived on social media. Some theories claimed that the pandemic was invented for pharmaceutical companies' profit from vaccines [44], whereas others linked mRNA COVID-19 vaccines to infertility and population control [10,11,44,48,49]. Another conspiracy theory claimed that Bill Gates and the US government aimed to implant trackable microchips into people through mass vaccination [11,27,49]. This aligns with conspiracy theories from earlier years. In particular, the Big Pharma conspiracy theory claims that pharmaceutical companies, together with politicians and other powerholders, conspire against the public interest [50]. The New World Order conspiracy theory alleges that a power elite with a globalization agenda colludes to rule the world [51]. Conspiracy theories have also linked other vaccines, such as poliovirus vaccines in the past [52,53] and COVID-19 vaccines in recent years, to HIV infection [54,55]. These conspiracy theories were based on the claims that alleged vaccines contained HIV.

Research Questions

To understand public discourse and reactions surrounding HIV vaccines on the X platform, we put forward the following research questions (RQs):

1. What are the topics of the posts about HIV vaccines? (RQ 1)
2. What is the valence of the posts about HIV vaccines? (RQ 2)
3. How do topics and valence vary across different types of posts? (RQ 3)
4. How do content features (topics and valence) and account features (verification status and follower count) affect 1-click reactions in terms of likes and reposts, respectively? (RQ 4)
5. What are the prominent anti–HIV vaccine conspiracy theories that receive the most reactions? (RQ 5)

Methods

Data Source

We collected English-language original posts about HIV vaccines on the X platform from January 1, 2022, to December 31, 2022, using Netlytic [56]. The selected time frame began in January 2022 with the initiation of mRNA HIV vaccine clinical trials fueling public discussion and concluded in December 2022, a significant month for HIV and AIDS awareness marked by World AIDS Day on the first day of the month. Posts, excluding reposts, that contained both keywords (case insensitive)—"HIV" and "vaccine"—were extracted, resulting in a total of 36,424 posts across 365 days. Posts were collected weekly. Posts published from the last ending time point to at least 24 hours before each collection time point were included in the data set, allowing for a substantial reaction time.

Procedure

The unit of analysis was a post. For each post, automated extraction produced data for user reactions (the number of likes and reposts) as well as account features (account verification status and follower count). All 36,424 posts underwent topic modeling using latent Dirichlet allocation (LDA) to identify latent topics, as well as sentiment analysis using Valence Aware Dictionary and Sentiment Reasoner (VADER) to access valence. LDA generated topic-specific loadings and identified the dominant topic for each post. VADER generated a valence compound score for each post, which was also categorized as positive, neutral, or negative based on standard VADER classification values.

LDA revealed 3 topics. As the topic of HIV and COVID-19 dominated in a large proportion of posts, we manually coded the 1000 most engaged posts containing the words "HIV" and "COVID" to uncover the salient aspects of HIV vaccines related to COVID-19. To develop coding for subtopics, 2 researchers initially reviewed and coded the top 200 posts with the most reactions. Subtopics were categorized by adapting existing categories from the literature [16,34] and integrating newly identified subtopics from the posts. The Scott π was 0.80 for categorizing subtopics. Subsequently, each researcher independently coded half of the remaining 800 posts.

We then conducted cross-tabulation analyses among all posts to examine the distribution of topics and valence among different types of posts. Furthermore, we conducted linear regression analyses among all posts to assess the influence of content and account features on these 1-click reactions. Of all 36,424 posts, 19,284 (52.94%) received ≥1 like, and 9155 (25.13%) received ≥1 repost. We added a constant value of 1 to all data points for likes and reposts before applying the natural logarithm. This was done to include posts with 0 likes or reposts and to mitigate the skewness of the data distribution.

XSL•FO
RenderX

Of the 28,439 posts that received likes or reposts, 6176 (21.72%) were negative. We manually coded the top 1000 negative posts with the most reactions to uncover prominent anti–HIV vaccine conspiracy theories. To develop coding for conspiracy theories, 2 researchers initially reviewed and coded the top 200 negative posts that received the most reactions. Posts containing conspiracy theories were identified based on expressions of postulated causal connections between people, objects, or events with malevolent intent [38,39]. Conspiracy theories were then classified based on the existing ones from the literature [50,51] and the emerging ones observed in the posts. Coding discrepancies were resolved through a further review of questionable posts and refinement of the conspiracy theories following the approach used in previous social media content analyses [40,57]. The procedure identified conspiracy theories and established intercoder reliability. The Scott π was 0.83 for identifying conspiracy theories and 0.81 for categorizing them. Each researcher then independently coded half of the remaining 800 negative posts.

Measures

User Reactions

One-click reactions were measured by the number of likes and reposts, which were automatically extracted. Because a small number of posts garnered significant 1-click reactions, the distribution of likes and reposts was right skewed. To reduce right skewness, we used the natural logarithm of the number of likes and reposts in linear regression analyses, as done in previous research [16,30,34].

Post Topics

All posts underwent topic modeling using LDA [58]. Topic modeling is a commonly used unsupervised learning method that generates a probabilistic model for a corpus of text data [59]. As a widely used topic model [59], LDA has been applied to discover topics within rich sources of digital health information, such as electronic health records [60], reviews on the web [61], and posts on X [16,34].

LDA relies on 2 matrices to define the underlying topical structure: the word-topic matrix and the document-topic matrix [62]. In this study, a post was considered a document. The general idea is that a post is represented by a Dirichlet distribution of latent topics, with each latent topic being represented by a Dirichlet distribution of words [59]. In the word-topic matrix, where the rows represent words and the columns represent topics, each element reveals the conditional probability of a word appearing within a topic [62]. A topic can be interpreted by examining a list of the most probable words ranked by their frequencies within a given topic using 3 to 30 words [63]. In the document-topic matrix, where rows represent posts and columns represent topics, each element reveals the conditional probability of a topic underlying a post [62]. In other words, it reveals the topic-specific loadings for each post.

When interpreting each topic, we reviewed the word-topic matrix as well as sample posts with high topic-specific loadings and significant reactions. LDA generated topic-specific loadings for each post ranging from 0 to 1, with values closer to 1 indicating a higher probability of a topic being associated with a post. Furthermore, LDA determined the dominant topic for each post by selecting the topic with the highest topic-specific loading among all topics. In the cross-tabulation analysis examining the distribution of topics across post types, the dominant topic for each post was entered for analysis. In the linear regression models assessing message-level drivers of user reactions, topic-specific loadings for each post were entered as topic values following previous research [16,34].

Post Valence

We used VADER to analyze the sentiment valence of each post. VADER is a rule-based model specifically attuned for assessing sentiments expressed in social media text [64]. VADER generated a compound valence score for each post ranging from –1 to 1, with a value of –1 indicating the most negative sentiment and a value of 1 indicating the most positive sentiment [65]. The standard VADER compound value thresholds for classifying valence categories are as follows: 0.05 to 1 for positive, −0.05 to 0.05 for neutral, and −0.05 to −1 for negative [65]. In the cross-tabulation analysis examining the distribution of valence among post types, the valence category for each post was entered for analysis. In the linear regression models assessing message-level drivers of user reactions, the VADER compound valence score for each post was used.

Post Type

This study collected original posts excluding reposts. For each original post, it was automatically extracted whether it was a self-composed post, a quote post with comments, or a reply.

Conspiracy Theories

In total, 2 researchers manually coded the top 1000 out of 6176 negative posts with the highest total number of likes and reposts to uncover highly engaged conspiracy theories. They distinguished conspiracy theories from other types of negative information, particularly other types of misinformation, by recognizing the presence of a hypothesized pattern of causal connections between people, objects, or events for malicious intent [38,39]. Conspiracy theories were then categorized based on the existing ones from the literature and the emerging ones observed in the posts.

As an example, consider a post paraphrased as follows:

> *Image using condoms consistently, only to contract HIV from a COVID vaccine.*

It was posted on February 9, 2022, and received 783 likes and 296 reposts. This post was not coded as displaying a conspiracy theory as it only presented misinformation suggesting that COVID-19 vaccines caused HIV. In comparison, another post was paraphrased as follows:

> *The COVID vaccine contained a spike protein derived from HIV. I was banned from saying this and ridiculed for months. Also, pharmacies stock up HIV self-tests.*

It was posted on February 8, 2022, with 147 likes and 48 reposts. This post was coded as displaying a conspiracy theory. It was further classified within the category of conspiracy theories linked to COVID-19 vaccines containing, causing, or increasing HIV. This post suggested a hypothesized pattern of maliciously

XSL•FO
RenderX

intended causal connections between the claim that the COVID-19 vaccine contained HIV and the stocking of HIV self-tests in pharmacies. As another example, a post was paraphrased as follows:

> *Scientists uncover a "highly virulent" strain of HIV in the Netherlands.*

It was posted on February 12, 2022, and received 11 likes and 11 reposts. This post conveyed negative information but did not present a conspiracy theory. In comparison, another post was paraphrased as follows:

> *By coincidence again, the development of a new mRNA HIV vaccine began just before the emergence of the new HIV strain.*

It was posted on February 8, 2022, and received 102 likes and 4 reposts. This post was coded as presenting a conspiracy theory and further classified into the category of conspiracy theories linked to the identification of a new highly virulent HIV strain. This post emphasized the speculative timing of the discovery of the new highly virulent HIV strain occurring shortly after the announcement of the development of a new mRNA HIV vaccine.

Account Features

For each post, the posting account's verification status and follower count were automatically extracted.

Data Analysis

We used cross-tabulation analyses to investigate the distribution of topics and valence across different post types, in which the dominant topic and valence category for each post were entered, respectively, alongside the post type. We used linear regression models to examine the message-level drivers of user reactions among posts that received likes or reposts. In the linear regression models, a constant value of 1 was added to all data points of like and repost counts. The natural log-transformed values for each post were then regressed on 3 topic-specific loadings generated from LDA, the valence compound score generated from VADER, and 2 autoextracted account features—account verification status and follower count. The "plus one" technique was used to include posts that received 0 likes or reposts and to address the skewness of the data distribution.

Ethical Considerations

Following Long Island University's institutional review board determination process, an institutional review board review was deemed unnecessary for this study, which collected and analyzed publicly available social media data. All referenced posts were paraphrased to avoid association with any particular user on the X platform.

Results

Post Topics

RQ 1 asked about the topics present in all the posts. We trained a topic model using LDA exploring topic numbers ranging from 2 to 20. The optimal number of topics (k) was selected considering both the coherence score (C_v) and the topic model visualization in a Python library called *pyLDAvis*, as done in previous research [16,66]. C_v is a metric that reflects the semantic coherence of topics by evaluating the word co-occurrence likelihood within topics [67]. A higher C_v indicates a better classification achieved by the topic model. In this study, the model with 2 topics (k=2) yielded the highest C_v (0.42), whereas the model with 3 topics (k=3) yielded the second highest C_v (0.35). The *pyLDAvis* chart depicts each topic as a circle. Overlapping areas between circles suggest similarities in topics. Thus, a chart without overlapping circles is preferable for k. The *pyLDAvis* chart for this study showed that, when the value of k was 2 or 3, the circles did not overlap. However, when k reached 4, the circles began to overlap, and overlapping circles persisted for values of k ranging from 4 to 20. Between the k values of 2 and 3, we opted for a model comprising 3 topics (k=3) considering that a smaller number of topics tends to result in overly broad meanings for each topic [68].

Table 1 summarizes the 3 topics and lists their representative posts. Each topic was interpreted by examining the top 10 probable words ranked by frequency, along with sample posts exhibiting high topic-specific loadings and 1-click reactions. Topic 1 was HIV and COVID-19, covering 78% of the tokens [69] and dominating in 92.46% (33,678/36,424) of the posts. Topic 2 was mRNA HIV vaccine trials, covering 14% of the tokens and dominating in 5.91% (2151/36,424) of the posts. Topic 3 was HIV vaccine and immunity, covering 8% of the tokens and dominating in 1.63% (595/36,424) of the posts.

Figure 1 illustrates the daily numbers of original posts about HIV vaccines throughout 2022, in total and categorized into 3 topics. Moderna's announcement of clinical trials for its first mRNA HIV vaccine on January 27, 2022, likely triggered the initial surge, culminating in a daily peak when the number of posts reached 805 on January 29, 2022. The daily number of posts about mRNA HIV vaccine trials (topic 2) in the week following Moderna's announcement was higher than on other days throughout the year. Nevertheless, even during that week, there were higher daily numbers of posts about HIV and COVID-19 (topic 1), which remained dominant among the 3 topics during the entire year. The year's second and highest daily peak occurred on February 8, 2022, recording a total of 1603 posts, most of which focused on HIV and COVID-19 (topic 1). This could be attributed to the emergence of new HIV-related events in early February 2022, including the promotion of HIV tests by public figures [64] and the discovery of a new highly virulent HIV strain [65]. The third highest daily peak, comprising 1085 posts, occurred on May 18, 2022, which has marked HIV Vaccine Awareness Day since 1998. Most of the posts centered on HIV and COVID-19 (topic 1). The remainder of the year did not reach such high peaks, with the largest daily volume of 205 posts occurring on December 2, 2022, the day following World AIDS Day, observed since 1988. Similar to previous daily peaks, most of the posts revolved around HIV and COVID-19 (topic 1).

The results revealed the dominance of HIV and COVID-19 (topic 1) in 92.46% (33,678/36,424) of the posts, with *HIV* as the most frequent word and *COVID* as the fourth most frequent word. To gain a deeper understanding of salient aspects of HIV

XSL•FO
RenderX

vaccines related to COVID-19, we manually coded the top 1000 posts with the highest total number of likes and reposts that contained both *HIV* and *COVID*. Table 2 summarizes the subtopics and their representative posts with like and repost counts.

The first major subtopic, comprising 24% (240/1000) of the posts, focused on the reciprocal influence of HIV vaccines and COVID-19 vaccines on each other's development. Years of HIV vaccine research facilitated the rapid development of mRNA COVID-19 vaccines, and the success of COVID-19 vaccines might accelerate the development of mRNA HIV vaccines. The second major subtopic, comprising 17.6% (176/1000) of the posts, involved comparisons between HIV and COVID-19 in various aspects. Specifically, the development speed of HIV vaccines compared to COVID-19 vaccines was a major point of comparison. In addition, some posts questioned whether potential HIV vaccines could be comparable to COVID-19 vaccines in terms of cost and accessibility during rollout. Others raised concerns about efficacy, safety, and inequality for both vaccines. The third major subtopic, comprising 26.5% (265/1000) of the posts, connected COVID-19 vaccines with HIV. One issue discussed was whether COVID-19 vaccines contained, caused, or increased HIV. Another issue raised was distinguishing between HIV symptoms and COVID-19 vaccine side effects, such as a fabricated condition called *VAIDS*, short for vaccine-acquired immunodeficiency syndrome. The fourth major subtopic, comprising 13.6% (136/1000) of the posts, featured conspiracy theories that presented hypothesized patterns linking COVID-19, HIV, and their vaccines with malicious intent. Prominent conspiracy theories in this subtopic included connecting misinformation that COVID-19 vaccines contain, cause, or increase HIV with the ongoing development of HIV vaccines; associating HIV and AIDS symptoms with side effects of COVID-19 vaccines; and claiming that COVID-19 originated from unsuccessful HIV vaccine research. As this study also manually coded the 1000 most engaged negative posts to identify prominent conspiracy theories, additional results pertaining to conspiracy theories will be discussed further in another subsection. The remaining posts related to HIV and COVID-19 included those that generally mentioned research on them or made connections without specifying details.

XSL•FO
RenderX

Table 1. Summary of topics with representative paraphrased posts (N=36,424).

Topic number and label	Top 10 words by frequency	Topic proportion (% of tokens)	Posts, n (%)	Valence, mean (SD)	Representative paraphrased posts
Topic 1: HIV and COVID-19	• Vaccine • HIV • Use • Covid • Aids • Year • Get • People • Test • Make	77.9	33,678 (92.46)	0.055 (0.480)	• It's astonishing that some people are opposed to vaccinations, when COVID vaccines have paved the way for HIV vaccine trials. (January 29, 2022; topic-specific loading: 0.862; likes: 30,591; reposts: 370) • Let me clarify. A small portion of the HIV virus was added in the Covid vaccine as to stabilize the protein. This vaccine was given to people. Now, a year later, they are indicating a rise in HIV cases, urging people to get HIV tests while developing an HIV vaccine. (February 8, 2022; topic-specific loading: 0.807; likes: 3207; reposts: 1456)
Topic 2: mRNA[a] HIV vaccine trials	• Trial • mRNA • Moderna • Technology • Clinical • Begin • Experimental • First • Launch • News	14.4	2151 (5.91)	0.040 (0.363)	• I am thrilled to share exciting news. The first patient has received a dose in the phase 1 trial of Moderna's mRNA-1644, an HIV vaccine candidate, which utilizes the same mRNA technology as our COVID-19 vaccines! (January 27, 2022; topic-specific loading: 0.521; likes: 18,895; reposts: 881) • I have some exciting news to share. The first patient has received a dose in the phase 1 trial of Moderna's and National Institutes of Health HIV trimer vaccine candidate, mRNA-1574, which utilizes the same mRNA technology as our COVID-19 vaccines! (March 21, 2022; topic-specific loading: 0.585; likes: 7056; reposts: 1224)
Topic 3: HIV vaccine and immunity	• Response • Immune • Risk • Protein • Antibody • Immunity • Spike_protein • High • Add • GT	7.8	595 (1.63)	−0.008 (0.445)	• New study: Analyzing two cooperating antibodies reveals immune pressure on HIV Env to elicit a V3-glycan supersite broadly neutralizing antibody lineage. The goal of vaccine is to stimulate broadly neutralizing antibodies. (September 26, 2022; topic-specific loading: 0.602; likes: 11; reposts: 4) • QTNSPRRAR is the t-cell CD8 activator essential in establishing a durable t-cell response to ACE2 receptor binding. This creates a sustained t-cell response to binding and then the antibody response to HIV like proteins, which is like all other t-cell vaccine adjuvants. (April 25, 2022; topic-specific loading: 0.506; likes: 12; reposts: 1)

[a]mRNA: messenger RNA.

XSL•FO
RenderX

Figure 1. Number of original posts about HIV vaccines in total and by topic in 2022. mRNA: messenger RNA.

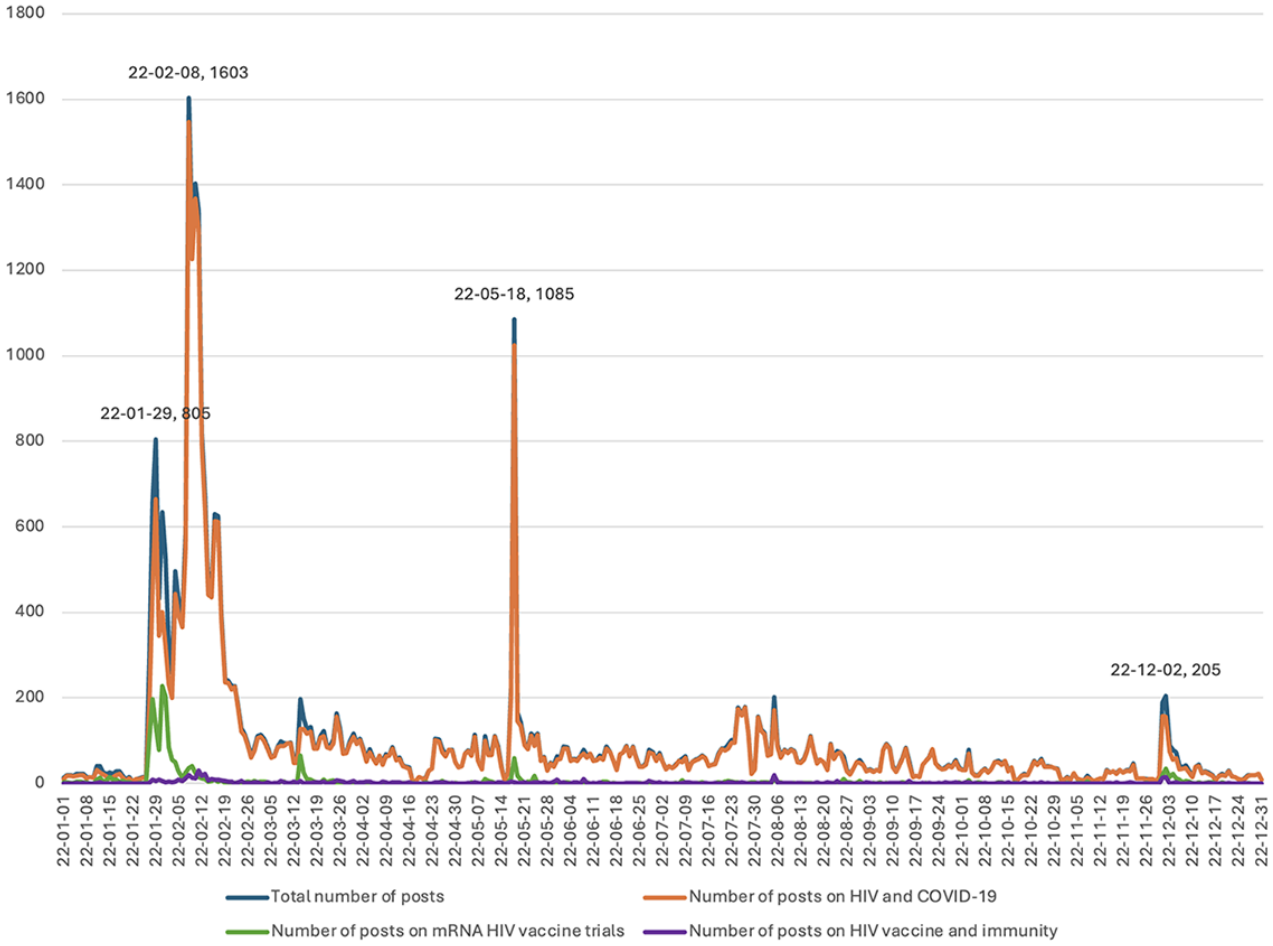

XSL•FO
RenderX

Table 2. Subtopics of HIV and COVID-19 and representative posts with a high number of reactions (N=1000).

Subtopic	Posts, n (%)	Paraphrased posts with a high number of reactions in each subtopic	Reaction count[a] (likes and reposts)
Reciprocal impact (n=240)			
COVID-19 vaccine research aiding HIV vaccines	211 (87.9)	The progress of items like PrEP[b] shouldn't be underestimated. The medical advancements with COVID boosted HIV vaccine development. The future is promising. (January 28, 2022)	4342 (4100 and 242)
HIV vaccine research aiding COVID-19 vaccines	29 (12.1)	Great presentation, addressing HIV and COVID vaccines. Despite 40 years of dedicated work, an HIV vaccine remains elusive, but the work was crucial in developing effective COVID vaccines. (February 13, 2022)	80 (62 and 18)
Comparisons (n=176)			
HIV and COVID-19 vaccines	61 (34.7)	Will this new HIV "vax" work similarly to the COVID vaccine? Does it not prevent transmission, acquisition, and have a short duration of efficacy? "It's just a slight case of AIDS"? (February 16, 2022)	225 (173 and 52)
Development speed of HIV vs COVID-19 vaccines	32 (18.2)	HIV was first identified 40 years ago, yet we still lack a vaccine, with only 3 cures. COVID emerged just over 2 years ago, and highly effective vaccines have saved over a million American lives. (February 16, 2022)	3058 (2710 and 348)
Vaccine inequality	19 (10.8)	Ready for the widespread distribution of an HIV vaccine? We are not prepared, considering the experience of the COVID vaccine rollout. It's crucial to plan for equal access to a future "people's vaccine." #HIVVaccineAwarenessDay (May 18, 2022)	31 (23 and 8)
HIV and COVID-19 pandemics	34 (19.3)	HIV+ kids were medically neglected and lost their lives in the late '90s, even as adults benefited from life-saving ART. Now the same situation applies to kids with COVID. No paxlovid or sotrovimab for kids under 12, and no vaccine for kids under 5. Over 1200 kids have died. Regulations are set. FDA, it's time to act. (February 16, 2022)	3210 (2579 and 631)
HIV and COVID-19 symptoms	15 (8.5)	People not seroconverting still develop #LongCOVID. An initial weak immune response could result in viral persistence, as is seen with HIV. A higher pathogen load tends to lead to more severe outcomes. Earlier vaccine or therapy tend to yield better outcomes. (February 15, 2022)	90 (67 and 23)
Rights of the COVID-19 unvaccinated vs HIV-positive individuals	15 (8.5)	HIV positive individuals can join the military, but current service members are being discharged for refusing a COVID vaccine, even though the military has invested a significant amount of money on them. (September 27, 2022)	250 (181 and 69)
COVID-19 vaccines and HIV (n=265)			
COVID-19 vaccines containing, causing, or increasing HIV	185 (69.8)	There's talk about the COVID vaccine potentially causing HIV. Shouldn't those vaccinated still practice safe sex? (February 6, 2022)	439 (382 and 57)
HIV or COVID-19 injuries such as VAIDS[c]	31 (11.7)	Is the next big thing vaccine-induced AIDS, which is also known as COVID vaccines? (February 9, 2022)	281 (176 and 105)
COVID-19 vaccines for HIV-positive individuals	49 (18.5)	Sputnik V is the world's first COVID vaccine to proven effective in HIV patients (March 29, 2022)	206 (169 and 17)
Conspiracy theories (n=136)			
Linked to COVID-19 vaccines containing, causing, or increasing HIV	69 (50.7)	Yesterday: Certain COVID vaccines might raise the risk of HIV infection. Today: Trials have begun for an HIV vaccine. (February 8, 2022)	5255 (3822 and 1433)
Linked to HIV or COVID-19 vaccine injuries such as VAIDS	9 (6.6)	Testing positive for HIV antibodies, as a response to a COVID vaccine containing HIV proteins, or its more severe and unexpected consequence, doesn't imply you have HIV. But it provides a convenient cover for vaccine injuries. Blame it on HIV! (February 11, 2022)	43 (36 and 7)
HIV vaccine research causing COVID-19	4 (2.9)	Why are COVID patients treated with an HIV medication? COVID-19 has been involved in HIV vaccine studies, and somehow components of the HIV envelope have found their way into SARS-CoV-2, resulting in coronavirus-HIV hybrids. (August 14, 2022)	47 (32 and 15)
Other conspiracy theories	54 (39.7)	__[d]	—

XSL•FO
RenderX

Subtopic	Posts, n (%)	Paraphrased posts with a high number of reactions in each subtopic	Reaction count[a] (likes and re-posts)
Debunking conspiracy theories (n=31)	31 (3.1)	How antivaxxers believe time travel works so that a covid vaccine that was never distributed mystically caused a new HIV strain in 1992. (8 Feb 2022)	31 (29 and 2)
Research on COVID-19 and HIV (n=31)	31 (3.1)	Exited to work as a researcher at the intersection of TB, HIV and COVID prevention science, global health equity, and clinical ID. (June 2, 2022)	86 (83 and 3)
COVID-19, HIV, and vaccines (n=12)	12 (1.2)	The connection between COVID, COVID vaccine, HIV and AIDS. [an article link]. (May 16, 2022)	67 (32 and 35)
Other (n=109)	109 (10.9)	—	—

[a]The reaction count is the total number of likes and reposts.

[b]PrEP: pre-exposure prophylaxis.

[c]VAIDS: vaccine-acquired immunodeficiency syndrome.

[d]The categories labeled as "other" contain various topics. Thus, no representative post is displayed.

Post Valence

RQ 2 asked about the sentiment valence present in all the posts. According to the standard VADER classification values, valence is categorized by compound scores as follows: positive (0.05 to 1), neutral (−0.05 to 0.05), and negative (−0.05 to −1) [65]. On average, all posts had a marginally positive score of 0.053. HIV and COVID-19 (topic 1) had a slightly positive average score of 0.055. The mRNA HIV vaccine trials (topic 2) had a neutral average score of 0.040, leaning toward the positive side. HIV vaccine and immunity (topic 3) had a more neutral average score of −0.0008. Moreover, 42.78% (15,584/36,424) of the posts were positive, 25.64% (9338/36,424) of the posts were neutral, and 31.58% (11,502/36,424) of the posts were negative.

Topics and Valence Across Post Types

Of the 36,424 posts, 18,580 (51.01%) were replies, making up over half of the overall count. Self-composed posts totaled 41.6% (15,151/36,424), whereas the remaining 7.39% (2693/36,424) were quote posts. RQ 3 asked about the distribution of topics and valence among the 3 post types. As Table 3 shows, the distribution of topics varied by post type (N=36,424, χ^2_4=2511.4, P<.001). Of the self-composed posts, 85.36% (12,933/15,151) focused on HIV and COVID-19 (topic 1) and 13.21% (2001/15,151) focused on mRNA HIV vaccine trials (topic 2). In comparison, quote posts and replies exhibited a different pattern, in each case >97% of posts centering on HIV and COVID-19 (topic 1; 2616/2693, 97.14% and 18,129/18,580, 97.57%, respectively).

As Table 4 shows, the distribution of valence also varied by post type (N=36,424, χ^2_4=911.7, P<.001). The proportion of positive posts was slightly higher among self-composed posts at 44.95% (6810/15,151) compared to replies at 41.09% (7634/18,580) and quote posts at 42.33% (1140/2693). Self-composed posts had a smaller proportion of negative posts at 23.56% (3570/15,151) compared to replies at 37.64% (6994/18,580) and quote posts at 34.83% (938/2693). The proportion of neutral posts was larger for self-composed posts at 31.49% (4771/15,151) compared to quote posts at 22.84% (615/2693) and replies at 21.27% (3952/18,580).

Regarding the distribution of topics and valence among the 3 types of posts, quote posts and replies displayed similarities, whereas self-composed posts diverged. Compared to self-composed posts, which initiate new conversations, there was a higher proportion of HIV and COVID-19-related posts (topic 1) and a greater proportion of negative posts among quote posts and replies, which contribute to existing conversations.

Table 3. Cross-tabulation of topics across post types (N=36,424)[a].

Topic number and label	Self-composed posts (n=15,151), n (%)	Quote posts with comments (n=2693), n (%)	Replies (n=18,580), n (%)	Total in topic (N=36,424), n (%)
Topic 1: HIV and COVID-19	12,933 (85.36)	2616 (97.14)	18,129 (97.57)	33,678 (92.46)
Topic 2: mRNA[b] HIV vaccine trials	2001 (13.21)	52 (1.93)	98 (0.53)	2151 (5.91)
Topic 3: HIV vaccine and immunity	217 (1.43)	25 (0.93)	353 (1.9)	595 (1.63)
Total in post type (N=36,424)	15,151 (41.6)	2693 (7.39)	18,580 (51.01)	36,424 (100)

[a]N=36,424, χ^2_4=2511.4, P<.001.

[b]mRNA: messenger RNA.

XSL•FO
RenderX

Table 4. Cross-tabulation of valence across post types (N=36,424)[a].

Valence	Self-composed posts (n=15,151), n (%)	Quote posts with comments (n=2693), n (%)	Replies (n=18,580), n (%)	Total in valence (N=36,424), n (%)
Positive	6810 (44.95)	1140 (42.33)	7634 (41.09)	15,584 (42.78)
Neutral	4771 (31.49)	615 (22.84)	3952 (21.27)	9338 (25.64)
Negative	3570 (23.56)	938 (34.83)	6994 (37.64)	11,502 (31.58)
Total in post type (N=36,424)	15,151 (41.6)	2693 (7.39)	18,580 (51.01)	36,424 (100)

[a]N=36,424, X^2_4=911.7, P<.001.

Content and Account Features Influencing User Reactions

RQ 4 asked about the influence of content and account features on likes and reposts.

Liking is more common than reposting. While 52.94% (19,284/36,424) of posts received an average of 24.83 likes, ranging from 1 to 102,843, a total of 25.13% (9155/36,424) posts received an average of 11.38 reposts, ranging from 1 to 10,572. Table 5 reveals the influence of content features (topics and valence) and account features (verification status and follower count) on the natural log-transformed number of likes and reposts. Both linear regression models were significant at P<.001. The adjusted R^2 was 0.072 for the like model and 0.090 for the repost model.

Among the 3 topics identified using LDA, HIV and COVID-19 (topic 1) did not affect like counts but decreased repost counts. In comparison, mRNA HIV vaccine trials (topic 2) decreased like counts while increasing repost counts. Positive valence increased like and repost counts. Account verification status and follower count increased like and repost counts.

Table 5. Linear regression models on predictors of likes and reposts (N=36,424).

Variables	Logarithm			
	Like count+1[a,b]		Repost count+1[a,c]	
	Standard β	*P* value	Standard β	*P* value
Content features				
Post topic				
Topic 1: HIV and COVID-19	−.003	.68	−.042	<.001
Topic 2: mRNA[d] HIV vaccine trials	−.039	<.001	.018	.02
Topic 3: HIV vaccine and immunity	−.910[e]	<.001	−.960[e]	<.001
Post valence	.034	<.001	.033	<.001
Account features				
Account verification status	.234	<.001	.239	<.001
Follower count	.095	<.001	.114	<.001

[a]The natural logarithm, ln (Y_i+1), was calculated on like and repost counts. This transformation was conducted to include posts receiving 0 likes and reposts, as well as to account for the skewness of the data distribution.

[b]F (model significance): P<.001; adjusted R^2=0.072.

[c]F (model significance): P<.001; adjusted R^2=0.090.

[d]mRNA: messenger RNA.

[e]The models excluded topic 3 on HIV vaccine and immunity to address multicollinearity issues arising from its correlations with topics 1 and 2. The reported standard β for topic 3 represents a possible β value if it had been included in the models.

Posts With Most Reactions

Table 6 summarizes posts ranked within the top 5 for the number of likes and reposts presented in chronological order. It is worth noting that all posts in the top 5 for likes and reposts were self-composed. One particular post, which garnered the most likes (n=102,843) and reposts (n=10,572), expressed the incredible feeling of witnessing the development of an HIV vaccine within our lifetimes. It was posted by an unverified account on January 28, 2022, the day after Moderna's announcement of clinical trials for its first mRNA HIV vaccine.

XSL•FO
RenderX

Table 6. Posts ranked in the top 5 for likes and reposts, paraphrased and in chronological order[a].

Paraphrased posts in chronological order	Like count and rank	Repost count and rank
I am thrilled to share exciting news. The first patient has received a dose in the phase 1 trial of Moderna's mRNA-1644, an HIV vaccine candidate, which utilizes the same mRNA technology as our COVID-19 vaccines! (January 27, 2022)	18,895; third	10,572; first
We are witnessing the development of an HIV vaccine within our lifetimes. How amazing is that? (January 28, 2022)	102,843; first	3881; fourth
It's astonishing that some people are opposed to vaccinations, when Covid vaccines have paved the way for HIV vaccine trials. (January 29, 2022)	32,115; second	4370; second
I have some exciting news to share. The first patient has received a dose in the phase 1 trial of Moderna's and National Institutes of Health HIV trimer vaccine candidate, mRNA-1574, which utilizes the same mRNA technology as our COVID-19 vaccines! (March 21, 2022)	7056; fifth	1224
Many medical and pharmaceutical achievements including HIV vaccine in progress. What an incredible time to be living in!" (June 3, 2022)	5034	2144; fifth
I am overjoyed to share exciting news. Despite being in an early phase 1 trial, results for HIV vaccine candidate eOD-GT8 60mer have revealed 97% of participants (with only one exception) produced an antibody response against HIV! (December 3, 2022)	18,109; fourth	3906; third

[a]Ranks beyond the fifth were not indicated.

Anti–HIV Vaccine Conspiracy Theories

RQ 5 asked about prominent anti–HIV vaccine conspiracy theories. Of the 1000 negative posts that received the most reactions, 227 (22.7%) contained conspiracy theories. As Table 7 shows, we classified these prominent anti–HIV vaccine conspiracy theories into 4 categories and presented their representative posts and the number of reactions.

The first category, comprising 44.9% (102/227) of the posts, formulated conspiracy theories by connecting COVID-19, COVID-19 vaccines, HIV, and HIV vaccines. For instance, 52.9% (54/102) of these posts connected the misinformation regarding COVID-19 vaccines containing, causing, or increasing HIV with the ongoing efforts to develop HIV vaccines. This misinformation may have arisen from past occurrences resurfacing following Moderna's initiation of its mRNA HIV vaccine trials. One incident occurred at the end of 2020, when an Australian COVID-19 vaccine, which used a small fragment of protein from HIV to clamp SARS-CoV-2's spike proteins, was abandoned due to false HIV-positive results [70]. Another incident occurred in October 2020, when 4 researchers sent a letter to a medical journal expressing concerns about the potential increased risk of HIV acquisition among men receiving COVID-19 vaccines using adenovirus type-5 vectors without supporting data from COVID-19 vaccines [71]. The misinformation typically interpreted the incidents out of context and generally suggested that COVID-19 vaccines contained, caused, or increased HIV without specifying details. In addition, there were conspiracy theories linking HIV and AIDS to COVID-19 vaccine side effects, including a fabricated condition known as VAIDS. VAIDS falsely suggests that COVID-19 vaccines caused immune deficiency [72]. Furthermore, there were claims that COVID-19 originated from unsuccessful HIV vaccine research.

The second category, comprising 38.3% (87/227) of the posts, suggested that the alignment of concurrent events with Moderna's start of mRNA HIV vaccine trials in late January 2022 was intentional to manipulate the market for HIV vaccines. These events included the rising HIV discussion and fear; promotion of HIV tests by public figures [73]; the discovery of a new highly virulent HIV strain [74]; and the passing away of HIV researchers, including Luc Montagnier, codiscoverer of HIV with an antivaccine stance during the COVID-19 pandemic [75], all occurring in early February 2022.

The third category, with 11.5% (26/227) of the posts, revealed conspiracy theories based on the distrust of powerholders [76]. Some posts extended existing conspiracy theories, such as the Big Pharma conspiracy theory [50] and the New World Order conspiracy theory [51], into the context of HIV vaccines, emphasizing the intent of powerholders, including major pharmaceutical companies and governments, behind vaccine promotion for financial profits and society control. Other posts created conspiracy theories about the government's research on HIV vaccines. The remaining posts generally stated that HIV vaccines were a scam. The final category comprised the remaining 5.3% (12/227) of the posts with other conspiracy theories.

It is worth noting that, of the 227 posts containing conspiracy theories, 39 (17.2%) were posted by accounts that had already been suspended at the time of manual coding. For these posts, the X platform displays the following message—"This post is from a suspended account"—and the content of the post is not visible. The X platform suspends accounts that violate its rules [77]. However, specific details of the violations are not accessible on the platform. The invisibility of these posts halted their spread when the suspension was enacted. For our manual coding of these posts, we used the text obtained during the data collection process.

XSL•FO
RenderX

Table 7. Anti–HIV vaccine conspiracy theories and representative posts with a high number of reactions (N=227).

Conspiracy theory	Post count, n (%)	Paraphrased posts with a high number of reactions for each conspiracy theory	Reaction count[a] (likes and reposts)
HIV and COVID- 19 (n=102)			
Linked to COVID-19 vaccines containing, causing, or increasing HIV	54 (52.9)	The COVID vaccine contained a spike protein derived from HIV. I was banned for saying this and ridiculed for months. Also, pharmacies stock up HIV self-tests. (February 8, 2022)[b]	195 (147 and 48)
Linked to whether HIV or COVID-19 vaccines cause injuries such as VAIDS[c]	38 (37.3)	Not about HIV or AIDS. It's deliberate immune system destruction called VAIDS, Vaccine Acquired Immunodeficiency Syndrome, engineered into vaccines before the pandemic by Gates and Fauci. Coordinated genocide. Stop. (February 12, 2022)[b]	992 (617 and 375)
HIV vaccine research causing COVID-19	10 (9.8)	Not surprisingly, this leads to the theory that SARS-CoV-2 was developed as vaccine, possibly against HIV. And it may have leaked. (September 20, 2022)	132 (96 and 36)
Linked to HIV-related events (n=87)			
Rising HIV discussion and fear	29 (33.3)	Are Canadians overlooking the connection between recent media coverage on an HIV epidemic, development of a new HIV vaccine, and their last vaccines? (February 7, 2022)	7766 (5887 and 1879)
HIV tests urged by public figures	28 (32.2)	Three Members of Parliament announced their recent HIV test results in the past few days. Oxford University is developing an HIV vaccine. When will HIV vaccines become mandatory? (February 7, 2022)	1182 (890 and 292)
New highly virulent HIV strain found	20 (23)	By coincidence again, the development of a new mRNA HIV vaccine began just before the emergence of the new HIV strain. (February 8, 2022)	106 (102 and 4)
HIV researchers passing away	10 (11.5)	Prominent HIV researchers are facing assassinations. Why are they being targeted after data shows Covid vaccine causes HIV symptoms? There is an effort to silence those with revealing knowledge. (February 10, 2022)	630 (501 and 129)
Powerholders' interests (n=26)			
Big Pharma's profits	6 (23.1)	Big Pharma stages a fake pandemic and profits from drugs. They add HIV to drugs to trigger an AIDS crisis and then promote an HIV vaccine. (February 8, 2022)	23 (16 and 7)
Depopulation and New World Order	4 (15.4)	How can we reclaim our nation from destructive forces and establish a New World Order? Covid hoax, HIV vaccine depopulation, staged riots, election manipulation. WEF destroys countries globally. (December 16, 2022)	23 (14 and 9)
Linked to government HIV vaccine research	8 (30.8)	The highly revered person is one accountable. If he could do it then, what did he do with mRNA? Revealed: How vulnerable children was treated with hardship in Fauci's obsessive pursuit of an HIV vaccine? (December 21, 2022)	90 (59 and 31)
HIV vaccine as a scam	8 (30.8)	I sense the emergency of an HIV vaccine scam. (February 9, 2022)	588 (518 and 70)
Other	12 (5.3)	__[d]	—

[a]The reaction count is the total number of likes and reposts.

[b]The posts were from suspended accounts.

[c]VAIDS: vaccine-acquired immunodeficiency syndrome.

[d]The categories labeled as "other" contain various conspiracy theories. Thus, no representative post is displayed.

Discussion

Principal Findings

This study investigated the patterns of public discourse and the message-level drivers of user reactions on the X platform regarding HIV vaccines through the analysis of posts using machine learning algorithms. We examined the distribution of topics and valence across different post types and assessed the influence of content features (topics and valence) and account features (account verification status and follower count) on like and repost counts. In addition, we manually coded the 1000 most engaged posts about HIV and COVID-19 to understand the salient aspects of HIV vaccines related to COVID-19 and the 1000 most engaged negative posts to identify prominent anti–HIV vaccine conspiracy theories.

XSL•FO
RenderX

The results revealed that COVID-19 plays a substantial role as a context for public discourse and reactions regarding HIV vaccines. Of the 3 topics identified using LDA, the leading topic was HIV and COVID-19, covering 78% of tokens and dominating in 92.46% (33,678/36,424) of the posts. Furthermore, on each of the top 4 days with the highest post counts, most of the posts were about HIV and COVID-19. This comprehensive topic included important subtopics that linked HIV vaccines with COVID-19 vaccines, as demonstrated through the manual coding of the 1000 most engaged posts about HIV and COVID-19. These subtopics encompassed the reciprocal influence of HIV vaccines and COVID-19 vaccines in advancing each other's development; comparisons in their development speed; inquiries about the possible alignment of HIV vaccines with COVID-19 vaccines in terms of cost and accessibility during distribution; and concerns about efficacy, safety, and equality for both vaccines.

COVID-19 positioned HIV vaccines in both a positive and negative context. On the one hand, the success of mRNA technology in COVID-19 vaccines [6] potentially cast mRNA HIV vaccines in a positive light. The topic of HIV and COVID-19 had a marginally positive valence score of 0.055. Moreover, 3 (60%) out of the 5 most liked posts and 2 (40%) out of the 5 most reposted posts expressed excitement about advancements in HIV vaccines that were based on the experience with COVID-19 vaccines. On the other hand, antivaccine discourse, including conspiracy theories, heated up during the COVID-19 pandemic [10,11,27,44,48,49], which posed challenges to HIV vaccines. Of the 1000 most engaged posts about HIV and COVID-19, a total of 136 (13.6%) featured conspiracy theories. Of the 1000 most engaged negative posts, 227 (22.7%) contained conspiracy theories, with 102 (44.9%) of them revolving around HIV and COVID-19. For instance, a prominent conspiracy theory connected the misinformation about COVID-19 vaccines containing, causing, or increasing HIV infection [55] with the initiation of clinical trials for mRNA HIV vaccines [4,5], implying a malevolent intent behind the deliberate connection. The results indicate that conspiracy theories tend to elicit an approach-oriented response, as evidenced by people engaging in liking and reposting, as opposed to an avoidance-oriented approach [39]. This underscores the need to intensify efforts to counter conspiracy theories in public health communication about HIV vaccines.

According to a study conducted by the Pew Research Center, irrespective of the subject matter, replies constituted the largest portion of original posts on X, followed by self-composed and quote posts [28]. Specifically, the number of replies was 3 times greater than that of self-composed posts. In this study, although replies constituted slightly more than half (18,580/36,424, 51.01%) of the posts, it is worth noting that the subject of HIV vaccines elicited a higher proportion of self-composed posts at 41.6% (15,151/36,424). Specifically, the number of replies was 23% higher than that of self-composed posts. Moreover, the topic of mRNA vaccine trials was most evident in self-composed posts compared to replies and quote posts. In comparison, there was a higher proportion of focus on the topic of HIV and COVID-19 and a greater proportion of negative posts among quote posts and replies, which contribute to existing conversations. This suggests that users were more likely to initiate new conversations rather than joining existing conversations about mRNA HIV vaccines. In contrast, they were more likely to join existing conversations rather than starting new conversations about HIV and COVID-19. In addition, users were less likely to initiate new conversations negatively but more likely to contribute negatively to existing ones.

As the primary topic, HIV and COVID-19 had no impact on like counts but had a negative impact on repost counts. In comparison, the topic of mRNA HIV vaccine trials had a negative impact on like counts and a positive impact on repost counts. The results should be interpreted while considering that, as revealed in previous research [16,34] and this study, most posts on the X platform are unlikely to receive likes and even less likely to receive reposts. In this study, among the total of 36,424 posts, approximately half (n=19,284, 52.94%) received likes, and approximately one-quarter (n=9155, 25.13%) received reposts. To include all posts and mitigate the data distribution skewness in the linear regression analysis, we applied the "plus one" technique. This involved adding a constant value of 1 to all like and repost data points before taking the natural logarithm. Although most posts were not liked or reposted, it is noteworthy that the topic of mRNA HIV vaccines led to an increase in repost counts, highlighting its positive influence on social sharing. In addition, 2 (40%) out of the 5 most reposted posts were about mRNA HIV vaccine trials. These results correspond to the findings of previous research that suggested the diffusion of novel useful information [12,16,32,36].

The overall valence of the posts about HIV vaccines was marginally positive. The positivity aligns with the positive sentiment found in posts on X about vaccines in general [13-15] and COVID-19 vaccines in particular [12,16,17]. However, the positivity about HIV vaccines was not apparent as the average score of 0.053 placed it on the edge of the neutral range, which goes from −0.05 to 0.05 according to the standard VADER classification values. Positive sentiment had a favorable impact on like and repost counts, partially consistent with findings of previous research on COVID-19 vaccines [16]. The post that achieved the most likes conveyed the incredible feeling of witnessing the development of an HIV vaccine in our lifetimes. This could be attributed to the psychological rationale that social transmission of positive content fulfills people's motivation to present a positive image [35,37]. In alignment with the findings of previous research [13,16,34], account verification status and follower count increased like and repost counts.

This study has implications for public health communication related to HIV vaccines and potentially other vaccines. Given the massive scale of the COVID-19 vaccination campaign, it is understandable that people will draw comparisons with other vaccines. Topic modeling identified HIV and COVID-19 as the primary topic, and manual coding revealed various intertwined aspects. Leveraging the advantages observed in the COVID-19 vaccine campaign, such as its widespread accessibility, could be valuable. Furthermore, addressing common concerns such as efficacy, safety, and inequality could also prove beneficial.

XSL•FO
RenderX

In the case of HIV vaccines, it is essential to tackle concerns associated with COVID-19 vaccines, especially those related to HIV vaccines. A major subtopic of HIV and COVID-19 involved suspicions about COVID-19 vaccines containing, causing, or increasing HIV. Another major subtopic was the confusion between HIV symptoms and the alleged side effects of COVID-19 vaccines, such as VAIDS. Misinformation concerning both subtopics has been woven into conspiracy theories, further complicating this situation. To combat misinformation and conspiracies that have these elements, efforts could focus on promoting evidence-based factual information [45-47].

Another notable technique in the conspiracy theories was linking concurrent COVID-19 and other HIV-related events in unsubstantiated relationships to create false perceptions, suggesting that these events were intentional to manipulate the market for HIV vaccines. These HIV-related events included rising HIV discussion and fear, promotion of HIV tests by public figures [73], the discovery of a new highly virulent HIV strain [74], and the passing away of HIV researchers, all occurring in early February 2022. These findings suggest that refuting false connections among such concurrent events can be an effective strategy to counter these conspiracy theories [45-47]. These occurrences, frequently entwined within conspiracy theories, could be specifically addressed in public health communication efforts.

Limitations

This study has several limitations. Because we used autoidentified content features (topics and valence) and autoextracted account features (verification status and follower count) in the regression models to predict the autoextracted number of user reactions (likes and reposts), the results were mostly limited to the examined autoidentified and autoextracted factors. For instance, political polarization, which manifested in a wide range of issues, including response to vaccines [78], could be a factor worth investigating in future studies. Furthermore, manual coding of conspiracy theories revealed a prevalent technique of twisting concurrent events into false relationships. This underscores the significance of refuting unfounded associations among these incidents to counter such conspiracy theories. It will be interesting for future research to assess the impact of this technique on user reactions to conspiracy theories. These findings could provide further insights into public health communication strategies to combat conspiracy theories.

Conclusions

The results highlight COVID-19 as a significant backdrop for public discourse and reactions on the X platform regarding HIV vaccines. COVID-19 situated HIV vaccines in both a positive and negative context. The success of mRNA COVID-19 vaccines shed a positive light on HIV vaccines. However, COVID-19 also situated HIV vaccines in a negative context, as evident in anti–HIV vaccine conspiracy theories falsely linking HIV vaccines to COVID-19. The findings provide implications for public health communication strategies concerning HIV vaccines.

Acknowledgments

This study was supported in part by the College of Arts and Sciences and the Harrington School of Communication and Media at the University of Rhode Island. The authors express their appreciation for the support. The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The data sets collected and analyzed during this study are available from the corresponding author upon request.

Conflicts of Interest

None declared.

References

1. Hussain A, Ali S, Ahmed M, Hussain S. The anti-vaccination movement: a regression in modern medicine. Cureus 2018 Jul 03;10(7):e2919 [FREE Full text] [doi: 10.7759/cureus.2919] [Medline: 30186724]
2. History of HIV vaccine research. National Institute of Allergy and Infectious Diseases. URL: https://www.niaid.nih.gov/diseases-conditions/hiv-vaccine-research-history [accessed 2023-09-20]
3. HIV surveillance report. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-34/index.html [accessed 2023-09-20]
4. IAVI and Moderna launch trial of HIV vaccine antigens delivered through mRNA technology. Moderna. 2022 Jan 27. URL: https://investors.modernatx.com/news/news-details/2022/IAVI-and-Moderna-Launch-Trial-of-HIV-Vaccine-Antigens-Delivered-Through-mRNA-Technology/default.aspx [accessed 2023-09-20]
5. NIH launches clinical trial of three mRNA HIV vaccines. National Institutes of Health. 2022 Mar 14. URL: https://www.nih.gov/news-events/news-releases/nih-launches-clinical-trial-three-mrna-hiv-vaccines [accessed 2024-09-20]

XSL•FO
RenderX

6. Tenforde MW, Self WH, Gaglani M, Ginde AA, Douin DJ, Talbot HK, et al. Effectiveness of mRNA vaccination in preventing COVID-19-associated invasive mechanical ventilation and death - United States, March 2021-January 2022. MMWR Morb Mortal Wkly Rep 2022 Mar 25;71(12):459-465 [FREE Full text] [doi: 10.15585/mmwr.mm7112e1] [Medline: 35324878]
7. Press release. Nobel Prize Outreach. URL: https://www.nobelprize.org/prizes/medicine/2023/press-release/ [accessed 2023-10-04]
8. May M. After COVID-19 successes, researchers push to develop mRNA vaccines for other diseases. Nat Med 2021 Jun;27(6):930-932 [FREE Full text] [doi: 10.1038/s41591-021-01393-8] [Medline: 34059823]
9. mRNA vaccine technology: a promising idea for fighting HIV. National Institute of Health. 2023 Feb 20. URL: https://covid1nih.gov/news-and-stories/mrna-vaccine-technology-promising-idea-fighting-hiv [accessed 2023-09-20]
10. Abbasi J. Widespread misinformation about infertility continues to create COVID-19 vaccine hesitancy. JAMA 2022 Mar 15;327(11):1013-1015. [doi: 10.1001/jama.2022.2404] [Medline: 35191947]
11. Islam MS, Kamal AH, Kabir A, Southern DL, Khan SH, Hasan SM, et al. COVID-19 vaccine rumors and conspiracy theories: the need for cognitive inoculation against misinformation to improve vaccine adherence. PLoS One 2021 May 12;16(5):e0251605 [FREE Full text] [doi: 10.1371/journal.pone.0251605] [Medline: 33979412]
12. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021 Jun 29;23(6):e24435 [FREE Full text] [doi: 10.2196/24435] [Medline: 34115608]
13. Blankenship EB, Goff ME, Yin J, Tse ZT, Fu KW, Liang H, et al. Sentiment, contents, and retweets: a study of two vaccine-related Twitter datasets. Perm J 2018 Sep 1;22:17-138. [doi: 10.7812/tpp/17-138]
14. Gunaratne K, Coomes EA, Haghbayan H. Temporal trends in anti-vaccine discourse on Twitter. Vaccine 2019 Aug 14;37(35):4867-4871. [doi: 10.1016/j.vaccine.2019.06.086] [Medline: 31300292]
15. Love B, Himelboim I, Holton A, Stewart K. Twitter as a source of vaccination information: content drivers and what they are saying. Am J Infect Control 2013 Jun;41(6):568-570. [doi: 10.1016/j.ajic.2012.10.016] [Medline: 23726548]
16. Zhang J, Wang Y, Shi M, Wang X. Factors driving the popularity and virality of COVID-19 vaccine discourse on Twitter: text mining and data visualization study. JMIR Public Health Surveill 2021 Dec 03;7(12):e32814 [FREE Full text] [doi: 10.2196/32814] [Medline: 34665761]
17. Hussain A, Tahir A, Hussain Z, Sheikh Z, Gogate M, Dashtipour K, et al. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. J Med Internet Res 2021 Apr 05;23(4):e26627 [FREE Full text] [doi: 10.2196/26627] [Medline: 33724919]
18. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
19. Burgess R, Feliciano JT, Lizbinski L, Ransome Y. Trends and characteristics of #HIVPrevention Tweets posted between 2014 and 2019: retrospective infodemiology study. JMIR Public Health Surveill 2022 Aug 11;8(8):e35937 [FREE Full text] [doi: 10.2196/35937] [Medline: 35969453]
20. Tran HT, Lu SH, Tran HT, Nguyen BV. Social media insights during the COVID-19 pandemic: infodemiology study using big data. JMIR Med Inform 2021 Jul 16;9(7):e27116 [FREE Full text] [doi: 10.2196/27116] [Medline: 34152994]
21. Sigalo N, Awasthi N, Abrar SM, Frias-Martinez V. Using COVID-19 vaccine attitudes on Twitter to improve vaccine uptake forecast models in the United States: infodemiology study of Tweets. JMIR Infodemiology 2023 Aug 21;3:e43703 [FREE Full text] [doi: 10.2196/43703] [Medline: 37390402]
22. Taggart T, Grewe ME, Conserve DF, Gliwa C, Roman Isler M. Social media and HIV: a systematic review of uses of social media in HIV communication. J Med Internet Res 2015 Nov 02;17(11):e248 [FREE Full text] [doi: 10.2196/jmir.4387] [Medline: 26525289]
23. Chan MP, Morales A, Zlotorzynska M, Sullivan P, Sanchez T, Zhai C, et al. Estimating the influence of Twitter on pre-exposure prophylaxis use and HIV testing as a function of rates of men who have sex with men in the United States. AIDS 2021 May 01;35(Suppl 1):S101-S109. [doi: 10.1097/QAD.0000000000002838] [Medline: 33867493]
24. The world needs an HIV vaccine. International AIDS Vaccine Initiative. URL: https://www.iavi.org/wp-content/uploads/2024/02/iavi_fact_sheet_need-for-hiv-vaccine.pdf [accessed 2024-03-15]
25. AIDS and the sustainable development goals. United Nations Programme on HIV/AIDS. URL: https://www.unaids.org/en/AIDS_SDGs [accessed 2023-09-20]
26. Using X. Twitter. URL: https://help.twitter.com/en/using-x [accessed 2024-09-20]
27. Trevisan M, Vassio L, Giordano D. Debate on online social networks at the time of COVID-19: an Italian case study. Online Soc Netw Media 2021 May;23:100136 [FREE Full text] [doi: 10.1016/j.osnem.2021.100136] [Medline: 36570036]
28. Chapekis A, Smith A. How U.S. adults on Twitter use the site in the Elon Musk era. Pew Research Center. 2023 May 17. URL: https://www.pewresearch.org/short-reads/2023/05/17/how-us-adults-on-twitter-use-the-site-in-the-elon-musk-era/ [accessed 2023-09-20]
29. Lee J, Hong IB. Predicting positive user responses to social media advertising: the roles of emotional appeal, informativeness, and creativity. Int J Inf Manage 2016 Jun;36(3):360-373. [doi: 10.1016/j.ijinfomgt.2016.01.001]

XSL•FO
RenderX

30. Pancer E, Poole M. The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 U.S. presidential nominees' tweets. Soc Influ 2016;11(4):259-270. [doi: 10.1080/15534510.2016.1265582]
31. Boulianne S, Larsson AO. Engagement with candidate posts on Twitter, Instagram, and Facebook during the 2019 election. New Media Soc 2023;25(1):119-140. [doi: 10.1177/14614448211009504]
32. Kim HS. Attracting views and going viral: how message features and news-sharing channels affect health news diffusion. J Commun 2015 Jun 01;65(3):512-534 [FREE Full text] [doi: 10.1111/jcom.12160] [Medline: 26441472]
33. About your For you timeline on X. Twitter. URL: https://help.twitter.com/en/using-twitter/twitter-timeline [accessed 2023-09-20]
34. Nanath K, Joy G. Leveraging Twitter data to analyze the virality of Covid-19 tweets: a text mining approach. Behav Inf Technol 2023;42(2):196-214. [doi: 10.1080/0144929x.2021.1941259]
35. Berger J, Milkman KL. What makes online content viral? J Market Res 2012;49(2):192-205. [doi: 10.1509/jmr.10.0353]
36. Tellis GJ, MacInnis DJ, Tirunillai S, Zhang Y. What drives virality (sharing) of online digital content? the critical role of information, emotion, and brand prominence. J Market 2019;83(4):1-20. [doi: 10.1177/0022242919841034]
37. Wojnicki AC, Godes D. Word-of-mouth as self-enhancement. SSRN J 2006 Jun 16. [doi: 10.2139/ssrn.908999]
38. Pummerer L, Böhm R, Lilleholt L, Winter K, Zettler I, Sassenberg K. Conspiracy theories and their societal effects during the COVID-19 pandemic. Soc Psychol Pers Sci 2022;13(1):49-59. [doi: 10.1177/19485506211000217]
39. van Prooijen JW, van Vugt M. Conspiracy theories: evolved functions and psychological mechanisms. Perspect Psychol Sci 2018 Nov;13(6):770-788 [FREE Full text] [doi: 10.1177/1745691618774270] [Medline: 30231213]
40. Massey PM, Kearney MD, Hauer MK, Selvan P, Koku E, Leader AE. Dimensions of misinformation about the HPV vaccine on Instagram: content and network analysis of social media characteristics. J Med Internet Res 2020 Dec 03;22(12):e21451 [FREE Full text] [doi: 10.2196/21451] [Medline: 33270038]
41. Wawrzuta D, Jaworski M, Gotlib J, Panczyk M. Characteristics of antivaccine messages on social media: systematic review. J Med Internet Res 2021 Jun 04;23(6):e24564 [FREE Full text] [doi: 10.2196/24564] [Medline: 34085943]
42. Garett R, Young SD. Online misinformation and vaccine hesitancy. Transl Behav Med 2021 Dec 14;11(12):2194-2199 [FREE Full text] [doi: 10.1093/tbm/ibab128] [Medline: 34529080]
43. Jolley D, Douglas KM. The effects of anti-vaccine conspiracy theories on vaccination intentions. PLoS One 2014 Feb 20;9(2):e89177 [FREE Full text] [doi: 10.1371/journal.pone.0089177] [Medline: 24586574]
44. Ullah I, Khan KS, Tahir MJ, Ahmed A, Harapan H. Myths and conspiracy theories on vaccines and COVID-19: potential effect on global vaccine refusals. Vacunas 2021;22(2):93-97 [FREE Full text] [doi: 10.1016/j.vacun.2021.01.001] [Medline: 33727904]
45. Chan MP, Jones CR, Hall Jamieson K, Albarracín D. Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. Psychol Sci 2017 Nov;28(11):1531-1546 [FREE Full text] [doi: 10.1177/0956797617714579] [Medline: 28895452]
46. Ecker UK, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, et al. The psychological drivers of misinformation belief and its resistance to correction. Nat Rev Psychol 2022 Jan 12;1:13-29. [doi: 10.1038/s44159-021-00006-y]
47. Jolley D, Douglas KM. Prevention is better than cure: addressing anti‐vaccine conspiracy theories. J Appl Soc Psychol 2017 Aug;47(8):459-469. [doi: 10.1111/jasp.12453]
48. Dotto C. Vaccine infertility claims on YouTube sweep across fringe platforms. First Draft News. 2021 Mar 24. URL: https://firstdraftnews.org/articles/vaccine-infertility-claims-youtube-fringe/ [accessed 2023-09-20]
49. Hamei L, Lopes L, Kirzinger A, Sparks G, Stokes M, Brodie M. KFF COVID-19 vaccine monitor: media and misinformation. Kaiser Family Foundation. 2021 Nov 8. URL: https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-media-and-misinformation/ [accessed 2023-09-20]
50. Blaskiewicz R. The Big Pharma conspiracy theory. Med Writ 2013;22(4):259-261. [doi: 10.1179/2047480613z.000000000142]
51. Spark A. Conjuring order: the new world order and conspiracy theories of globalization. Sociol Rev 2000;48(2_suppl):46-62. [doi: 10.1111/j.1467-954x.2000.tb03520.x]
52. Aylward RB, Heymann DL. Can we capitalize on the virtues of vaccines? insights from the polio eradication initiative. Am J Public Health 2005 May;95(5):773-777. [doi: 10.2105/AJPH.2004.055897] [Medline: 15855451]
53. Jegede AS. What led to the Nigerian boycott of the polio vaccination campaign? PLoS Med 2007 Mar;4(3):e73 [FREE Full text] [doi: 10.1371/journal.pmed.0040073] [Medline: 17388657]
54. Fact check: an Australian vaccine trial did not give trial participants HIV. Reuters. 2020 Dec 18. URL: https://www.reuters.com/article/idUSKBN28R2WM/ [accessed 2023-09-20]
55. COVID-19 vaccines do not cause HIV or AIDS. Reuters. 2022 Feb 21. URL: https://www.reuters.com/article/factcheck-vaccines-hiv/fact-check-covid-19-vaccines-do-not-cause-hiv-or-aids-idUSL1N2UW10H [accessed 2023-09-20]
56. Gruzd A. Netlytic: software for automated text and social network analysis. Netlytic. 2016. URL: https://netlytic.org/home/?page_id=49 [accessed 2023-09-20]
57. Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. J Med Internet Res 2016 Dec 05;18(12):e318 [FREE Full text] [doi: 10.2196/jmir.6670] [Medline: 27919863]

XSL•FO
RenderX

58. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text]
59. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, et al. Text summarization techniques: a brief survey. Int J Adv Comput Sci Appl 2017;8(10):397-405. [doi: 10.14569/IJACSA.2017.081052]
60. Fohner AE, Greene JD, Lawson BL, Chen JH, Kipnis P, Escobar GJ, et al. Assessing clinical heterogeneity in sepsis through treatment patterns and machine learning. J Am Med Inform Assoc 2019 Dec 01;26(12):1466-1477 [FREE Full text] [doi: 10.1093/jamia/ocz106] [Medline: 31314892]
61. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc 2014;21(6):1098-1103 [FREE Full text] [doi: 10.1136/amiajnl-2014-002711] [Medline: 24918109]
62. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, et al. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun Methods Meas 2018;12(2-3):93-118. [doi: 10.1080/19312458.2018.1430754]
63. Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014 Presented at: ILLBI 2014; June 27, 2014; Baltimore, MD. [doi: 10.3115/v1/w14-3110]
64. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May 16;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
65. VADER-sentiment-analysis. GitHub. URL: https://github.com/cjhutto/vaderSentiment [accessed 2023-09-20]
66. Feng Y, Chen H, Kong Q. An expert with whom i can identify: the role of narratives in influencer marketing. Int J Advert 2021;40(7):972-993. [doi: 10.1080/02650487.2020.1824751]
67. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015 Presented at: WSDM '15; February 2-6, 2015; Shanghai, China. [doi: 10.1145/2684822.2685324]
68. Gan J, Qi Y. Selection of the optimal number of topics for LDA topic model-taking patent policy analysis as an example. Entropy (Basel) 2021 Oct 03;23(10):1301 [FREE Full text] [doi: 10.3390/e23101301] [Medline: 34682025]
69. API documentation. pyLDAvis. URL: https://pyldavis.readthedocs.io/en/latest/modules/API.html [accessed 2023-09-20]
70. Horizon special: the vaccine. British Broadcasting Corporation. 2021. URL: https://www.bbc.co.uk/programmes/m000x2tf [accessed 2023-09-20]
71. Buchbinder SP, McElrath MJ, Dieffenbach C, Corey L. Use of adenovirus type-5 vectored vaccines: a cautionary tale. Lancet 2020 Oct 31;396(10260):e68-e69 [FREE Full text] [doi: 10.1016/S0140-6736(20)32156-5] [Medline: 33091364]
72. Kelety J. COVID-19 vaccines don't contain HIV. AP News. 2023 Jan 30. URL: https://apnews.com/article/fact-check-coronavirus-vaccine-hiv-185375755407 [accessed 2023-09-20]
73. Davies C. Prince Harry: get tested for HIV to protect others in same way as for Covid. Guardian News and Media Limited. 2022 Feb 10. URL: https://www.theguardian.com/uk-news/2022/feb/10/prince-harry-get-tested-for-hiv-to-protect-others-in-same-way-as-covid-diana-princess-of-wales [accessed 2024-03-06]
74. Wymant C, Bezemer D, Blanquart F, Ferretti L, Gall A, Hall M, et al. A highly virulent variant of HIV-1 circulating in the Netherlands. Science 2022 Feb 04;375(6580):540-545 [FREE Full text] [doi: 10.1126/science.abk1688] [Medline: 35113714]
75. Epstein RH. Luc Montagnier, nobel-winning co-discoverer of H.I.V., dies at 89. The New York Times. 2022 Feb 10. URL: https://www.nytimes.com/2022/02/10/science/luc-montagnier-dead.html [accessed 2023-09-20]
76. Bunting H, Gaskell J, Stoker G. Trust, mistrust and distrust: a gendered perspective on meanings and measurements. Front Polit Sci 2021 Jul 6;3(642129):1-15. [doi: 10.3389/fpos.2021.642129]
77. About suspended accounts. Twitter. URL: https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts [accessed 2023-09-20]
78. Fridman A, Gershon R, Gneezy A. COVID-19 and vaccine hesitancy: a longitudinal study. PLoS One 2021 Apr 16;16(4):e0250123 [FREE Full text] [doi: 10.1371/journal.pone.0250123] [Medline: 33861765]

Abbreviations

LDA: latent Dirichlet allocation
mRNA: messenger RNA
RQ: research question
VADER: Valence Aware Dictionary and Sentiment Reasoner
VAIDS: vaccine-acquired immunodeficiency syndrome

XSL•FO
RenderX

Edited by G Eysenbach; submitted 04.10.23; peer-reviewed by X Ma, J Zhang; comments to author 18.10.23; revised version received 08.11.23; accepted 28.02.24; published 03.04.24.

Please cite as:
Zhang JM, Wang Y, Mouton M, Zhang J, Shi M
Public Discourse, User Reactions, and Conspiracy Theories on the X Platform About HIV Vaccines: Data Mining and Content Analysis
J Med Internet Res 2024;26:e53375
URL: https://www.jmir.org/2024/1/e53375
doi:10.2196/53375
PMID:38568723

©Jueman M Zhang, Yi Wang, Magali Mouton, Jixuan Zhang, Molu Shi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Emerging Trends in Information-Seeking Behavior for Alpha-Gal Syndrome: Infodemiology Study Using Time Series and Content Analysis

Jamie L Romeiser[1], PhD, MPH; Nicole Jusko[1], MPH; Augusta A Williams[1], SCD, MPH

Department of Public Health and Preventive Medicine, Upstate Medical University, Syracuse, NY, United States

Corresponding Author:
Jamie L Romeiser, PhD, MPH
Department of Public Health and Preventive Medicine
Upstate Medical University
766 Irving Ave
Syracuse, NY, 13210
United States
Phone: 1 315 464 6897
Email: RomeiseJ@upstate.edu

Abstract

Background: Alpha-gal syndrome is an emerging allergy characterized by an immune reaction to the carbohydrate molecule alpha-gal found in red meat. This unique food allergy is likely triggered by a tick bite. Cases of the allergy are on the rise, but prevalence estimates do not currently exist. Furthermore, varying symptoms and limited awareness of the allergy among health care providers contribute to delayed diagnosis, leading individuals to seek out their own information and potentially self-diagnose.

Objective: The study aimed to (1) describe the volume and patterns of information-seeking related to alpha-gal, (2) explore correlations between alpha-gal and lone star ticks, and (3) identify specific areas of interest that individuals are searching for in relation to alpha-gal.

Methods: Google Trends Supercharged-Glimpse, a new extension of Google Trends, provides estimates of the absolute volume of searches and related search queries. This extension was used to assess trends in searches for alpha-gal and lone star ticks (*lone star tick*, *alpha gal*, and *meat allergy*, as well as *food allergy* for comparison) in the United States. Time series analyses were used to examine search volume trends over time, and Spearman correlation matrices and choropleth maps were used to explore geographic and temporal correlations between alpha-gal and lone star tick searches. Content analysis was performed on related search queries to identify themes and subcategories that are of interest to information seekers.

Results: Time series analysis revealed a rapidly increasing trend in search volumes for alpha-gal beginning in 2015. After adjusting for long-term trends, seasonal trends, and media coverage, from 2015 to 2022, the predicted adjusted average annual percent change in search volume for alpha-gal was 33.78%. The estimated overall change in average search volume was 627%. In comparison, the average annual percent change was 9.23% for lone star tick, 7.34% for meat allergy, and 2.45% for food allergy during this time. Geographic analysis showed strong significant correlations between alpha-gal and lone star tick searches especially in recent years (ρ=0.80; P<.001), with primary overlap and highest search rates found in the southeastern region of the United States. Content analysis identified 10 themes of primary interest: diet, diagnosis or testing, treatment, medications or contraindications of medications, symptoms, tick related, specific sources of information and locations, general education information, alternative words for alpha-gal, and unrelated or other.

Conclusions: The study provides insights into the changing information-seeking patterns for alpha-gal, indicating growing awareness and interest. Alpha-gal search volume is increasing at a rapid rate. Understanding specific questions and concerns can help health care providers and public health educators to tailor communication strategies. The Google Trends Supercharged-Glimpse tool offers enhanced features for analyzing information-seeking behavior and can be valuable for infodemiology research. Further research is needed to explore the evolving prevalence and impact of alpha-gal syndrome.

(J Med Internet Res 2024;26:e49928) doi:10.2196/49928

XSL•FO
RenderX

KEYWORDS

alpha-gal; alpha gal; alpha-gal syndrome; lone star tick; infodemiology; time series; content analysis; Google Trends; allergy; allergic; immune; immunology; immunological; information behavior; information behaviour; information seeking; geographic

Introduction

Food allergy is a growing public health concern. The global burden of food allergy is increasing, with 220 million individuals worldwide estimated to be affected by at least 1 food allergy [1-3]. In the United States, it is estimated that 26 million, or about 6.2%, of US adults and children have a food allergy [4]. Food allergies can induce a hefty economic burden on individuals and families due to lost labor, productivity, out-of-pocket costs, and opportunity costs [5] and can decrease the quality of life of patients in impacted physical health and mental health [6,7].

One somewhat unconventional food allergy on the rise is the alpha-gal allergy, also known as alpha-gal syndrome, red meat allergy, or tick bite meat allergy [8]. Alpha-gal syndrome is a type of allergy that is characterized by an immune reaction to the carbohydrate molecule galactose-alpha-1,3-galactose (alpha-gal), which is found in most mammalian or “red meat” [9]. This immune reaction can result in a variety of symptoms including hives, swelling of the face or facial features, shortness of breath, abdominal pain, gastrointestinal issues, anaphylaxis, or even fatality [10]. Whereas conventional food allergies usually involve an immediate immune response, alpha-gal allergy is characterized by the delayed onset of symptoms (ie, 3-8 hours following exposure) [8,9,11,12]. This delayed response, compounded with the varying clinical presentation, can make the diagnosis of the allergy challenging.

The alpha-gal allergy stands out not only due to its delayed nature but also because of the unconventional way most individuals develop the allergy. In most cases, it is thought that transmission occurs through the bite of a tick, which injects alpha-gal into the bloodstream. This sensitizes individuals to the molecule and can lead to an immune response when consuming mammalian meat [13]. Alpha-gal cases have been reported worldwide and are associated with various tick species. In the United States, however, growing evidence suggests that alpha-gal allergy is primarily linked to lone star ticks [14]. The connection between alpha-gal allergy and lone star ticks was initially described in 2011 [15] and has since been supported by subsequent studies [12,13,16-18]. Recent evidence from a case-control study further strengthens this association, revealing significantly higher odds of tick exposures among individuals with alpha-gal syndrome compared to controls [16].

Lone star ticks are traditionally found in the Southeastern region of the United States. However, the geographic range of lone star ticks has expanded to the Northeast [19] and Midwest in part due to climate change and sensitivity to microclimate conditions [18]. Indeed, exposure to and cases of alpha-gal allergy are also expanding beyond the Southeastern United States [20]. Recognition and general awareness of alpha-gal allergy are growing [21], but health care providers’ knowledge of the connection between lone star ticks and alpha-gal allergy may be lagging. A recent survey conducted among clinicians in Illinois revealed a lack of awareness regarding the link between lone star tick bites and alpha-gal allergy, as well as limited familiarity with diagnostic testing for the condition [22].

Because there remains considerable variation in health care providers’ knowledge of the condition [23], patients often play a critical role in driving their own diagnoses [24]. As 1 study reported, there is often a lengthy amount of time between symptom onset and diagnosis, with patients visiting multiple health care settings and receiving numerous referrals before obtaining the correct diagnosis [23]. In general, it is likely that many patients rely on finding their own sources of information to learn about the allergy and potentially self-diagnose it before receiving an official diagnosis.

The alpha-gal syndrome is not a reportable or notifiable condition and was only granted a separate International Classification of Diseases *10th Revision* (*ICD-10*) code for diagnostic identification in 2022. Therefore, little data exist on the prevalence of the allergy [8]. In the absence of incidence or prevalence data, infoveillence and infodemiology techniques can be used to help describe the overall population interest in the topic of alpha-gal [25], as well as correlations with potentially related search terms. Many studies over the past decade have used Google Trends as a tool to explore patterns in health information–seeking behavior [26,27]. In fact, 1 prior study using Google Trends found a high correlation between alpha-gal and lone star tick, including a general upward trend in relative search volume (RSV) [21]. Data were examined from 2004 to 2019, but it is likely that both awareness and diagnosis of the allergy have grown in the past 3 years [20].

Further, new tools have emerged that offer enhanced features for analyzing various aspects of information-seeking behavior. One such tool is Glimpse, which is an extension app of Google Trends [28]. Google Trends has been both praised and criticized for reporting all searches as RSVs [29]. RSV is a query share of a search term within a specific geography and time range, which is then normalized to the highest point of popularity within that time period [26]. A notable drawback of this approach is that the absolute volume is not provided; therefore, there is no way to track an estimated number of queries for a particular topic [30]. Because the RSV is indexed to the highest point of popularity for a term within a time period, reproducibility of research results can be challenging [26,29]. To address this issue, the Supercharged-Glimpse extension offers an estimate of the absolute search volume, as well as a dashboard of additional information such as related search terms. The absolute volume numbers are overlayed on the traditional RSV index graph. This tool could be useful for understanding the specific content topics and themes that people search for health information.

Additional quantification of the changing volume of information seeking for alpha-gal can provide valuable evidence regarding shifts in awareness levels and potential changes in alpha-gal prevalence. Moreover, analyzing the evolving

XSL•FO
RenderX

information-seeking patterns for both alpha-gal and lone star ticks can offer insights into the changing public interest across different geographic areas. Finally, health care providers and public health officials could benefit from understanding the topics that are most relevant to their patients and the public and adjust their communication strategies for alpha-gal accordingly. Therefore, using the Google Trends Supercharged-Glimpse extension, our objectives were to (1) describe the basic volume and patterns of information seeking related to alpha-gal; (2) explore further correlations between alpha-gal and lone star ticks; and (3) understand and identify specific questions, concerns, and areas of interest that individuals are searching for in relation to alpha-gal.

Methods

Google Trends Supercharged-Glimpse

Google Trends Supercharged-Glimpse is a new web-based tool available as an extension of Google Trends [28]. Glimpse functions in parallel with Google Trends and uses a similar random sampling approach to produce reports containing multiple components of trends in web-based information–seeking behavior. Searches are conducted on the Google Trends platform, with options to specify a particular geographic location and time period (spanning from 2004 to the present day). With the Glimpse extension, estimated search trends over time are provided as an absolute volume, rather than RSV. This allows for direct comparisons of volume from separate terms conducted in separate searches. The extension also produces a list of the highest volume of keywords and questions related to the search term, which is akin to a listing of terms that information seekers use in addition to the main keyword. Reports also contain the RSV for geographic areas from Google Trends.

Search Strategy

To begin our search, we implemented a methodological framework based upon previous strategies aimed to establish a consistent approach to conducting Google Trends research [31,32] and adhered to a suggested checklist for documenting and reporting our search strategy [26]. We report keywords in italics and the classification of the keyword in parentheses.

Google Trends Supercharged-Glimpse was used to determine search trends for 2 alpha-gal keywords: *alpha gal* (search term) and *meat allergy* (search term). Additional keywords related to alpha-gal were explored (*alpha-gal*, *alpha-gal syndrome*, and *alpha gal syndrome*), but these terms were visibly unstable, indicating a low search volume (with most months returning less than 1000 searches) and lower likelihood of using these terms when seeking information. Because these spelling variations produced very low volumes of searches and because the Supercharged tool does not allow modifiers (eg, "+" to add the terms together), we proceeded with the analysis without aggregating these data [31] but recognized these as potential limitations. We selected *lone star tick* (animal) as a keyword to explore objective 2 and *food allergy* (search term) as a broad comparison term. All keyword searches were compared with and without the use of quotations and produced similar results in terms of volume. Data were limited to the United States due to the context of the alpha-gal and lone star tick exploration, and the time frame ranged to include all available data (from January 1, 2004, to March 1, 2023). All keyword searches were conducted separately, without combining keywords. Similar to Google Trends, the Glimpse extension performs a sampling approach to estimate the absolute search volume. As with all sampling approaches, there is a degree of variability each time the data are queried. Further, there is a degree of caching that occurs; therefore, the same search conducted in a short time period (eg, within 10 minutes) may have identical numerical results. To produce a better estimate of the absolute volume of searches for our selected search terms, data were collected at 2 PM daily for a period of 10 days (March 2-12, 2023). Data for each term were compiled and averaged for search volume over time (absolute volume), related search queries (qualitative lists), and geographic interest (RSV). To demonstrate the overlap and correlation between the traditional RSV index values from Google Trends and the absolute volume estimates from the Supercharged-Glimpse extension, the averaged absolute volume and the averaged RSV index values for *alpha gal* (search term) are plotted in Figure S1 in Multimedia Appendix 1. Absolute volume over time for all terms is presented in the *Results* section.

Statistical Analysis

Objective 1: Trends Over Time

Time series analyses were used to assess the trends in all 4 keywords over the study period. A quasi-Poisson regression was fit to account for the overdispersion of the count outcome variables. The regression used a penalized spline on year to account for long-term trends in these various allergies and related searches over time. We opted to use a penalized spline because the data appeared to be nonlinear in the raw data plots. Using penalized splines on the year allowed for the flexibility to account for the complex patterns seen in the raw data plots while not overfitting the model. The seasonality of web-based searches is likely to mirror the seasonality of cases of disease [31]. Therefore, we expected to see seasonal variation in search patterns for all 4 search terms, and indeed, this is what was revealed in graphing the monthly search volumes. There are seasonal variations in tick behaviors and tick-host encounters, which both peak during the warm season [33,34]. Additionally, food-related anaphylaxis has been found to increase during the warm season due to oral allergies (eg, pollen-food allergy syndrome) [35]. Therefore, a binary variable was created to indicate the warm season (May-September), which was then controlled via a linear term to account for seasonal trends. Google trends data for rare diseases can be influenced by the media [29]; therefore, the dates of national media stories on alpha-gal were identified from 2012 through 2023, and this (binary variable for the months that contained national media coverage) was controlled using a linear term. The predicted absolute search volume values are shown in the *Results* section. The average annual percent change (AAPC) in search volume was calculated based on this adjusted model for all 4 terms [36].

Objective 2: Alpha-Gal and Lone Star Tick

While our original intent was to describe how information-seeking behavior for both *alpha gal* and *lone star*

XSL•FO
RenderX

tick changed geographically and temporally from 2004 to 2022, initial data extraction efforts revealed that the RSV geographic metrics were unstable and unreliable from 2004 to 2013. This is likely due to a lower volume of searches conducted during this time period, which would increase sampling variability. To produce more stable and reproducible results for the geographic analysis, the geographic interest index values were queried and exported in 3-year time period increments (2014-2016, 2017-2019, and 2020-2022) for the *alpha gal* (search term) and *lone star tick* (*animal*) keywords. Choropleth maps were produced. The distribution of the data was found to be nonnormally distributed (Shapiro-Wilk test *P* values <.05 for all variables); therefore, a Spearman correlation matrix was generated to examine geographic and temporal correlations for *alpha gal* and *lone star tick* using the RSV index values for the 3 time periods (2014-2016, 2017-2019, and 2020-2022).

Objective 3: Content Analysis

A conceptual content analysis was performed on the related search query data for alpha-gal. Lists from the 10 separate days were examined and found to be nearly identical. Lists were then combined and duplicates were removed, leaving a total of 371 related searches. All 371 related search queries contained the words “alpha gal” either before or after the additional keywords (eg, *alpha gal hives*). For the qualitative content analysis, an inductive coding strategy was first implemented to identify overarching themes based on the content of the data itself. We identified 10 main themes: diet, diagnosis or testing, treatment, medications or contraindications of medications, symptoms, tick related, specific sources of information and locations, general education information, alternative words for alpha-gal, and unrelated or other. Two authors (JR and NJ) independently coded the list of 371 words based on the main concept themes. The results from the coders were compared, and discrepancies were resolved for a final theme designation. This process was repeated within each theme to further identify theme-based subcategories of interest. The total number of subthemes identified was 41. The frequency of keywords in each search theme and theme-based subcategories was compiled and described. All analyses were performed using SAS (version 9.4; SAS Institute) and R Studio (R Foundation for Statistical Computing).

Ethical Considerations

All data were publicly available and unidentifiable search engine metadata. Data are not used at the individual level and do not involve human subjects; therefore, institutional review board approval was not required for this study.

Results

Time Series Analyses

The average absolute search volumes from 2004 to March 2023 were plotted for *alpha gal*, *meat allergy*, *lone star tick*, and *food allergy* (Figure 1). The comparison of the traditional RSV index values and the absolute volume estimates demonstrated near-perfect overlap, with a Spearman correlation coefficient of ρ=0.99 (*P*<.001; Figure S1 in Multimedia Appendix 1). In the time series analysis, the estimated absolute search volume values over time were nonlinear for all 4 search terms (Figure 2). The binary indicator for the warm season was a significant predictor for all 4 search terms. While all 4 search terms demonstrated this significant seasonal trend, the magnitude of the warm season term was greatest for *lone star tick* and lowest for *food allergy.* Media coverage was only a significant predictor for *alpha gal*, *lone star tick*, and *meat allergy*. National media coverage of alpha-gal did not significantly impact *food allergy* searches. Based on the expected search volumes after adjusting for seasonality and media events in the time series analysis, the AAPC from 2004 to 2022 was 18.76%, 10.87%, 7.52%, and 1.09% for *alpha gal*, *meat allergy*, *lone star tick*, and *food allergy*, respectively. Interestingly, search trend volume began to noticeably increase starting around 2015 for *alpha gal.* From January 2015 to January 2022, the AAPC was 33.78%, 9.23%, 7.34%, and 2.45% for *alpha gal*, *meat allergy*, *lone star tick*, and *food allergy*, respectively. The estimated total increase during this time period was 627%, 81%, 61%, and 25% for *alpha gal*, *meat allergy*, *lone star tick*, and *food allergy* searches.

XSL•FO
RenderX

Figure 1. Absolute search volume trends for *alpha gal* (red), *food allergy* (green), *lone star tick* (blue), and *meat allergy* (purple), 2004-2023.

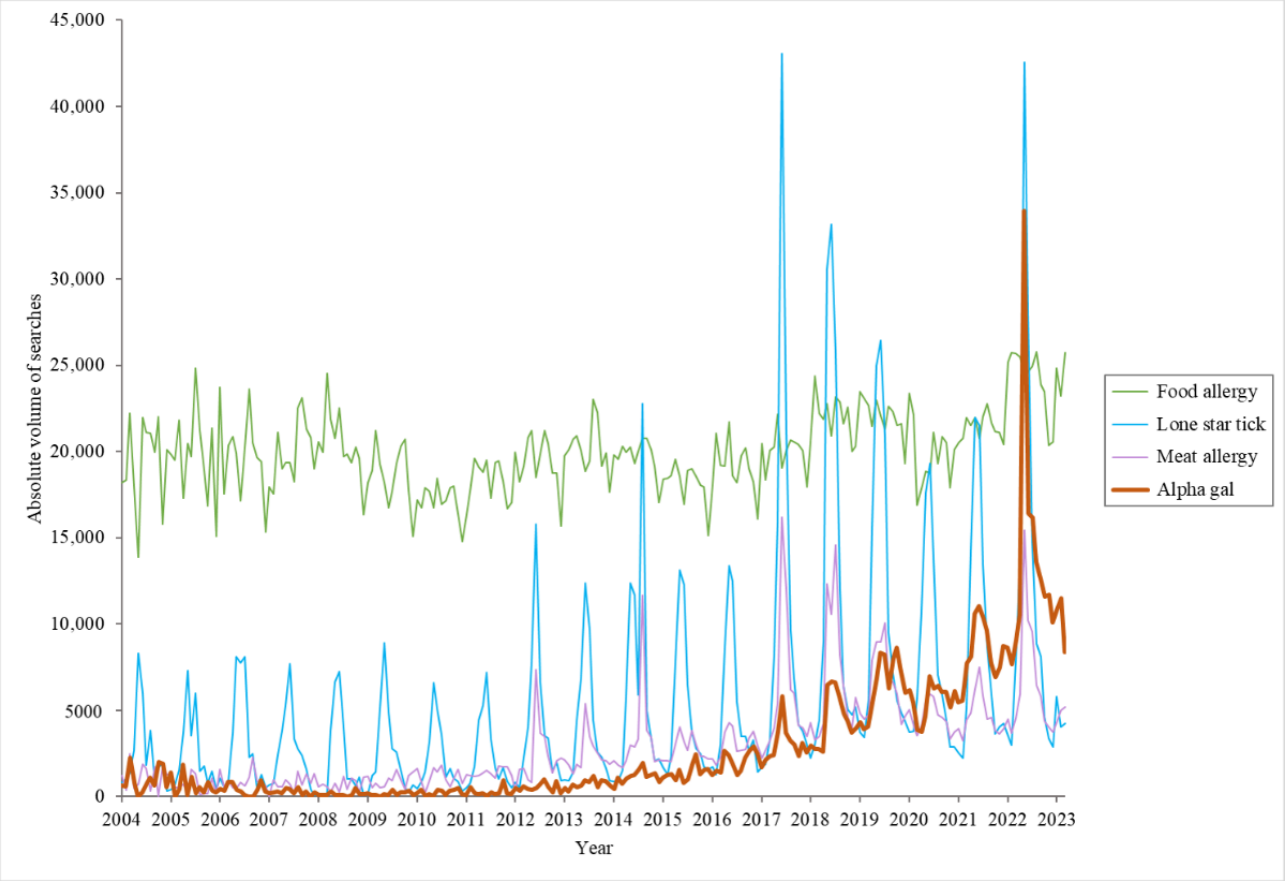

Figure 2. Estimated number of monthly searches for *alpha gal* (red), *food allergy* (green), *lone star tick* (blue), and *meat allergy* (purple) after adjustment for long-term trends, season, and media reports. The mean number of searches for each term is shown by a solid line, and 95% CIs are shaded around each respective line.

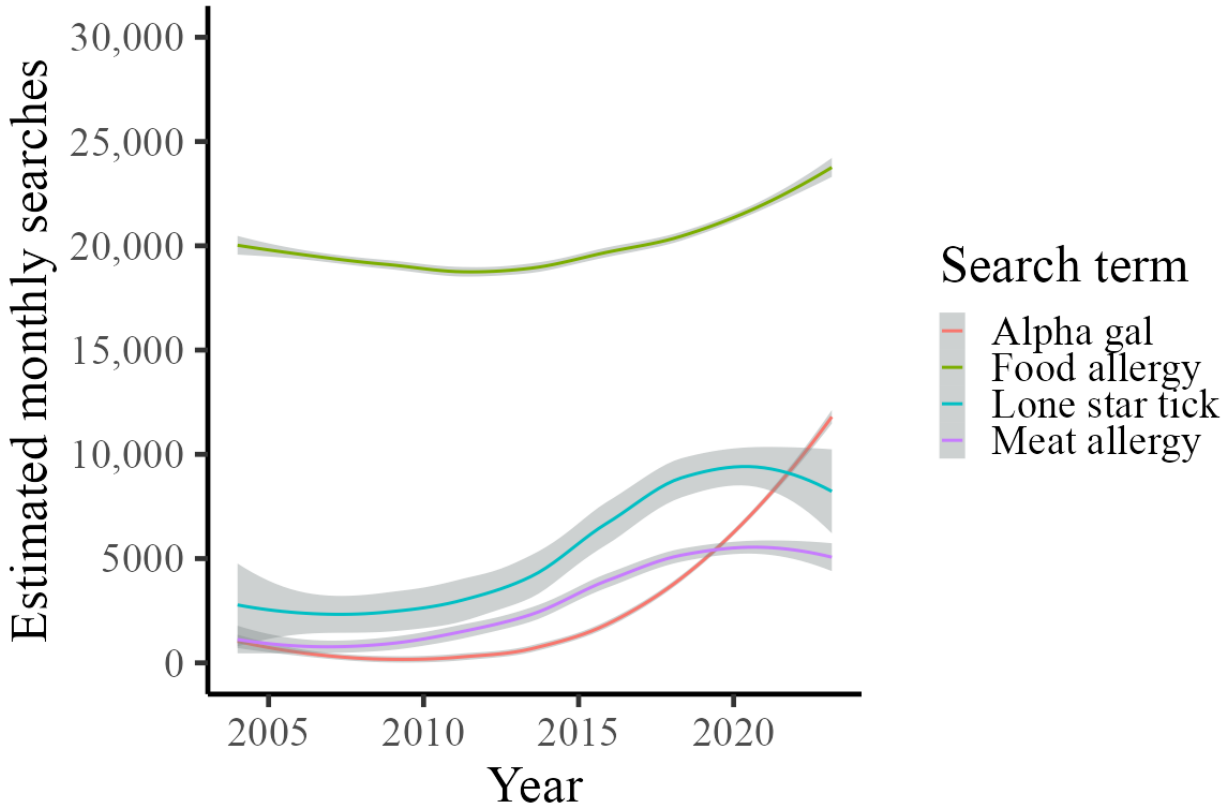

XSL•FO
RenderX

Alpha-Gal and Lone Star Tick

Choropleth maps of the RSV indices for *alpha gal* and *lone star tick*, as well as the Spearman correlation coefficient for each time period, are presented in Figure 3. All correlations were significant at $P<.001$. State RSV indices for *alpha gal* and *lone star tick* were moderately correlated in 2014-2016 (ρ=0.59) but became more strongly correlated in the latter 2 time periods (ρ=0.82; ρ=0.80). These correlations can be visualized by the noticeable geographic overlap observed over time between states with a high information-seeking interest in *alpha gal* and *lone star ticks*.

Figure 3. Spatial and temporal correlations between alpha gal and lone star tick information seeking over time. A. All Spearman ρ correlations for state-based RSV values were significant at the $P<.001$ level. Orange values indicate the 3 primary correlations of interest. B. Correlations can be further visualized in the choropleth maps. Dark blue states on the map indicate high RSV values. RSV: relative search volume.

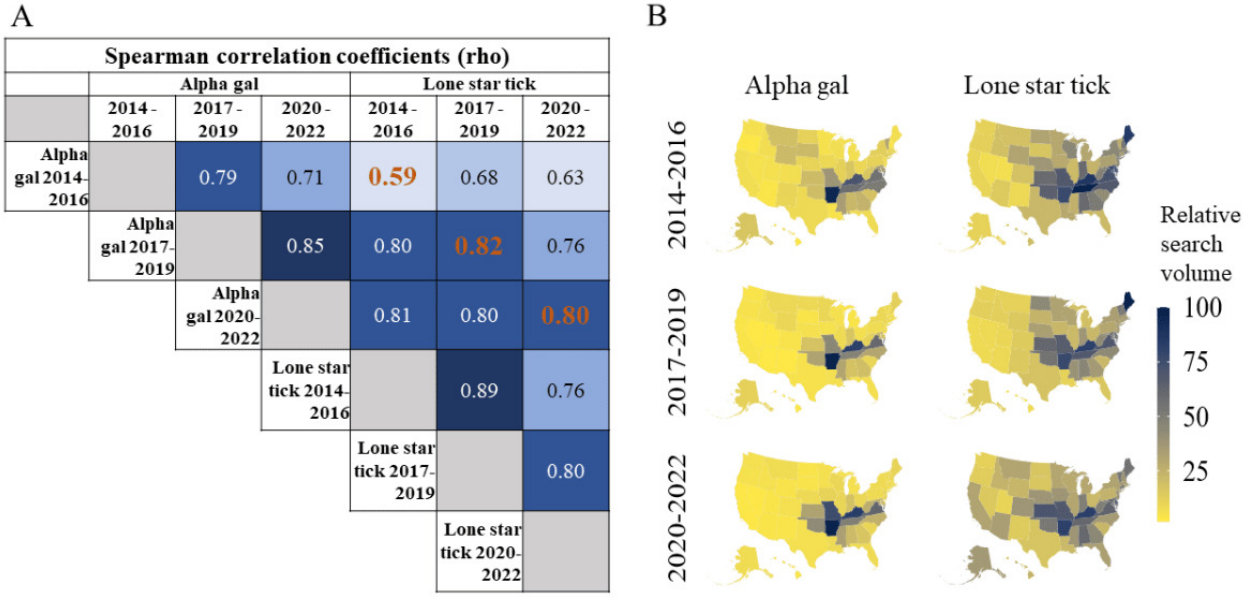

Content Analysis

Overview

The content analysis revealed 10 overarching themes (Figure 4). Of the 10 overarching themes, 7 were further divided into subcategories.

XSL•FO
RenderX

Figure 4. Content analysis of related search queries. Ten primary themes were identified, and multiple subcategories within the themes revealed diverse information-seeking behavior. ICD-10: International Classification of Diseases 10th Revision.

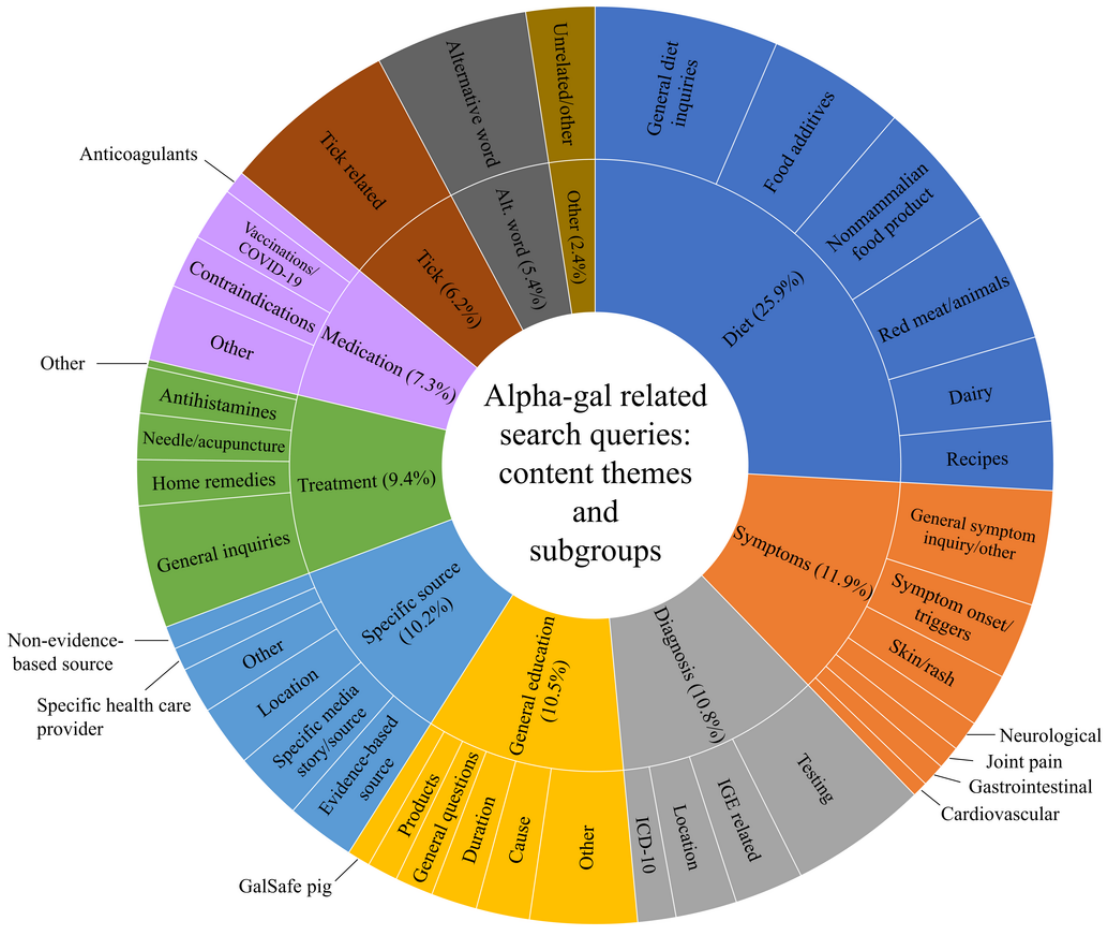

Diet

The most diverse search theme was diet (96/371, 25.9%), which could be further divided into 6 subthemes. The most popular subtheme was composed of general inquiries (eg, *what foods to avoid with alpha gal?* and *what foods are safe with alpha gal?*), followed by food additives (eg, *gelatin*, *glycerin*, and *magnesium stearate*), red meat (ie, *specific mammal inquiries*), nonmammalian food options (eg, *ostrich*, *eggs*, and *turkey*), dairy products (eg, *butter*, *milk*, and *cheese*), and searches for recipes.

Symptoms

Symptom-related searchers accounted for 11.9% (44/371) of the total related search queries. General symptom–related questions or inquiries (eg, *what are the symptoms?*) were most prominent, followed by symptom onset or triggers (eg, *reaction time*, *onset*, and *exercise*). The other 5 specific subtheme symptoms that emerged were skin or rash (7/44, 15.9%), neurological (4/44, 9.1%), joint pain (3/44, 6.8%), gastrointestinal (3/44, 6.8%), and cardiovascular (2/44, 4.5%) symptoms.

Diagnosis

Diagnosis-related searchers accounted for 10.8% (40/371) of the total related searches, including 4 subthemes of general searches for testing (eg, *blood test* and *test results*), specific searches for IgE (eg, *IgE numbers* and *levels*), specific laboratories that perform testing, and searches for *ICD-10* codes.

General Education

General education and information seeking accounted for 10.5% (39/371) of the related searches. This theme was further broken down into causes (eg, *what is the cause of alpha gal?*, *is alpha gal genetic?*, *is alpha gal contagious?*), duration (eg, *is alpha gal permanent?*, *does alpha gal go away?*), general questions (eg, *what is alpha gal?*), products, information seeking on the knockout pig, and other (eg, *alpha gal and COVID-19*, *alpha-gal while pregnant*, and *prevalence*).

Specific Sources

Specific sources of information or location accounted for 10.2% (38/371) of searches. Around 24% (9/38) of these were media-specific searches (eg, *alpha-gal radiolab* and *nytimes*), 24% (9/38) were evidence-based sources (eg, *NIH*, *UpToDate*, *CDC*, and *Mayo Clinic*), and 21% (8/38) were location or place specific (eg, *Kentucky*, *LynchburgVA*, and *UVA*). Other subgroups included nonevidence-based sources, specific clinician or health care provider searches, and others.

Treatment

Treatment-related searches accounted for 9.4% (35/371). Most searches were general (eg. *how to treat alpha gal* and *is it curable?*), while others were specifically related to acupuncture,

XSL•FO
RenderX

antihistamines, and home remedies (eg, *natural remedies* and *essential oils*).

Medication

Medication or contraindication-related searches accounted for 7.8% (27/371) of related searches. Roughly 25% (7/27) of these searches were vaccine related (eg, *flu shot* and *COVID-19 vaccine*). About 25% (7/27) were inquiries about medications known to contain alpha-gal (eg, heparin), while the majority (37%, 16/27) were general inquiries on what medications were necessary to avoid (eg, *alpha gal in medications*, *drugs to avoid*, *alpha gal medication list*, and *alpha gal* and *anesthesia*).

Discussion

Principal Findings

To our knowledge, we believe this is one of the first studies to investigate absolute search volume trends over time using the Google Supercharged-Glimpse extension app. Our analysis revealed several important trends and patterns regarding public interest and information seeking for alpha-gal allergy.

First, over the past several years, the volume of searches for the alpha-gal allergy is significantly increasing at a rapid rate. After adjusting for seasonality and national media stories, the overall increase for *alpha gal* searches over time was at least 6 times that of the other search terms. Interestingly, *alpha gal* search term volume surpassed that of *meat allergy* in 2019, which likely indicates an increase in the awareness of the correct terminology for the allergy. Further, while *lone star tick* and *meat allergy* search term volumes seemed to level off from 2020 to 2022, *alpha gal* search term interest continued to grow. National media stories were found to have a significant impact on the search volume for *alpha gal*, *meat allergy*, and *lone star tick* terms, highlighting the role that media can play in engaging public interest and awareness of these topics.

Geographic popularity information-seeking for the term *alpha gal* did not vary greatly over time and remained concentrated in the southeast regions of the United States. While this finding was similar to a study using 2019 data [21], it was somewhat unexpected. Diagnoses of alpha-gal syndrome have occurred in other geographic regions including the Northeast and Midwest [20], and cases are increasing in those areas [8]. It is possible that this finding might be reflective of how the RSV index is generated. States with larger populations could have a greater absolute volume of interest, but this might not be reflected in the RSV index. There were no states throughout time with an RSV index of 0, which indicates that people are seeking information for the alpha-gal allergy in all 50 states. This finding is similar to laboratory-based studies that show documented cases in most states [20].

Geographic searches for the terms *alpha gal* and *lone star tick* strongly overlap. States have similarly ranked RSV indices for both *alpha gal* and *lone star tick* terms, and these have strengthened over time. Of note, geographic search interest in lone star ticks appears to be expanding. Search interest maps from 2014 to 2016 and 2017 to 2019 strongly overlap with a 2016 study that documented expansions of the lone star tick range [19]. Furthermore, search interest maps from 2020 to 2022 show a similar overlap with predicted lone star tick habitat expansion models developed in 2021 [37]. It is possible, therefore, that future interest and prevalence maps for alpha-gal may expand in similar ways. This serves as a forewarning, highlighting the urgent need to expand education efforts for both the general public and health care providers.

Finally, people are seeking a broad array of topics related to alpha-gal. The largest subgroup of content searched for in conjunction with alpha-gal was tick related. Encouragingly, national media stories and evidence-based sources comprised almost 50% of the specific sources that people were searching for. Unsurprisingly, diet was the largest theme, and interest in concealed sources of exposure and food additives remains a large concern [11]. The ubiquitous presence of animal products in food, medications, and other products can make it extremely difficult for individuals to know if what they ingest is safe [8]. Given the challenges of identifying safe foods and products, there is a clear need for improved food and product labeling. Additionally, information seeking for specific symptoms highlighted the broad array of ways that alpha-gal can manifest within an individual. Symptom variety remains one of the major reasons alpha-gal remains underdiagnosed. Health care professionals should not only be familiar with the most commonly identifiable symptoms of alpha-gal such as urticaria or anaphylaxis [18] but also be knowledgeable about symptom manifestations like joint pain [8], gastrointestinal symptoms [12,18,38], dizziness, or heart palpitations [10].

Limitations

There are several limitations to this study. First, we were only able to capture information seeking on 1 search engine platform. Second, absolute search volume estimates were not available geographically, and the RSV geographic data were unstable when examined yearly. We chose to sacrifice the granularity of a year-by-year analysis in favor of more reliable results. Third, the absolute volume was also not provided for each of the related searches. Therefore, the proportional composition of the content analysis themes represents the diversity within the theme, not necessarily the popularity of the theme. It is possible that tick-related searches encompassed the majority of the volume of the related alpha-gal searches, but we cannot identify this based on the data. Finally, at the time of the initial data analysis and submission of this study, alpha-gal syndrome was only available as a search term and not a disease. Since that time, the term has become searchable on the Google Trends platform as a designated syndrome. Absolute search volume trends over time for both *alpha gal* (search term) and *alpha-gal syndrome* (syndrome) are remarkably similar, indicating there is likely direct overlap between these 2 searches. It is unknown to what degree the addition of this condition as a designated syndrome may affect the volume reported for the *alpha gal* (search term) in the future.

Conclusions

Information seeking for alpha-gal syndrome is rapidly increasing. Geographic overlap with lone star tick searches might suggest future expansions in alpha-gal interest and prevalence, thereby emphasizing the urgency for increased education efforts. The diverse range of topics and symptoms

XSL•FO
RenderX

searched for highlights the ongoing challenges faced by individuals affected by alpha-gal. However, the content and subthemes identified can serve as a valuable guide to facilitate public health outreach and effective patient-physician communication.

Acknowledgments

We would like to thank Adam Starke for assistance in generating the choropleth maps.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of absolute search volume and relative search volume (RSV) index over time for *alpha gal* (search term).
[PNG File , 162 KB - jmir_v26i1e49928_app1.png]

References

1. Dunlop JH, Keet CA. Epidemiology of food allergy. Immunol Allergy Clin North Am 2018;38(1):13-25. [doi: 10.1016/j.iac.2017.09.002] [Medline: 29132669]
2. De Martinis M, Sirufo MM, Suppa M, Ginaldi L. New perspectives in food allergy. Int J Mol Sci 2020;21(4):1474 [FREE Full text] [doi: 10.3390/ijms21041474] [Medline: 32098244]
3. Fiocchi A, Risso D, DunnGalvin A, Díaz SNG, Monaci L, Fierro V, et al. Food labeling issues for severe food allergic patients. World Allergy Organ J 2021;14(10):100598 [FREE Full text] [doi: 10.1016/j.waojou.2021.100598] [Medline: 34703523]
4. Ng AE, Boersma P. Diagnosed allergic conditions in adults: United States, 2021. NCHS Data Brief 2023(460):1-8 [FREE Full text] [Medline: 36700875]
5. Gupta R, Holdford D, Bilaver L, Dyer A, Holl JL, Meltzer D. The economic impact of childhood food allergy in the United States. JAMA Pediatr 2013;167(11):1026-1031 [FREE Full text] [doi: 10.1001/jamapediatrics.2013.2376] [Medline: 24042236]
6. Blok BMJF, Dubois AEJ, Vlieg-Boerstra BJ, Elberink JNGO, Raat H, DunnGalvin A, et al. Health-related quality of life of food allergic patients: comparison with the general population and other diseases. Allergy 2010;65(2):238-244 [FREE Full text] [doi: 10.1111/j.1398-9995.2009.02121.x] [Medline: 19796214]
7. DunnGalvin A, Dubois AEJ, Blok BMJF, Hourihane JOB. The effects of food allergy on quality of life. Chem Immunol Allergy 2015;101:235-252 [FREE Full text] [doi: 10.1159/000375106] [Medline: 26022884]
8. Commins SP. Diagnosis and management of alpha-gal syndrome: lessons from 2,500 patients. Expert Rev Clin Immunol 2020;16(7):667-677 [FREE Full text] [doi: 10.1080/1744666X.2020.1782745] [Medline: 32571129]
9. Carson AS, Gardner A, Iweala OI. Where's the Beef? Understanding allergic responses to red meat in alpha-gal syndrome. J Immunol 2022;208(2):267-277 [FREE Full text] [doi: 10.4049/jimmunol.2100712] [Medline: 35017216]
10. Alpha-gal syndrome. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/ticks/alpha-gal/index.html [accessed 2023-06-07]
11. Macdougall JD, Thomas KO, Iweala OI. The meat of the matter: understanding and managing alpha-gal syndrome. Immunotargets Ther 2022;11:37-54 [FREE Full text] [doi: 10.2147/ITT.S276872] [Medline: 36134173]
12. Wilson JM, Schuyler AJ, Workman L, Gupta M, James HR, Posthumus J, et al. Investigation into the α-gal syndrome: characteristics of 261 children and adults reporting red meat allergy. J Allergy Clin Immunol Pract 2019;7(7):2348-2358.e4 [FREE Full text] [doi: 10.1016/j.jaip.2019.03.031] [Medline: 30940532]
13. Crispell G, Commins SP, Archer-Hartman SA, Choudhary S, Dharmarajan G, Azadi P, et al. Discovery of alpha-gal-containing antigens in North American tick species believed to induce red meat allergy. Front Immunol 2019;10:1056 [FREE Full text] [doi: 10.3389/fimmu.2019.01056] [Medline: 31156631]
14. Mitchell CL, Lin FC, Vaughn M, Apperson CS, Meshnick SR, Commins SP. Association between lone star tick bites and increased alpha-gal sensitization: evidence from a prospective cohort of outdoor workers. Parasit Vectors 2020;13(1):470 [FREE Full text] [doi: 10.1186/s13071-020-04343-4] [Medline: 32928302]
15. Commins SP, James HR, Kelly LA, Pochan SL, Workman LJ, Perzanowski MS, et al. The relevance of tick bites to the production of IgE antibodies to the mammalian oligosaccharide galactose-α-1,3-galactose. J Allergy Clin Immunol 2011;127(5):1286-1293.e6 [FREE Full text] [doi: 10.1016/j.jaci.2011.02.019] [Medline: 21453959]

XSL•FO
RenderX

16. Kersh GJ, Salzer J, Jones ES, Binder AM, Armstrong PA, Choudhary SK, et al. Tick bite as a risk factor for alpha-gal-specific immunoglobulin E antibodies and development of alpha-gal syndrome. Ann Allergy Asthma Immunol 2023;130(4):472-478 [FREE Full text] [doi: 10.1016/j.anai.2022.11.021] [Medline: 36574585]
17. Sharma SR, Karim S. Tick saliva and the alpha-gal syndrome: finding a needle in a haystack. Front Cell Infect Microbiol 2021;11:680264 [FREE Full text] [doi: 10.3389/fcimb.2021.680264] [Medline: 34354960]
18. Young I, Prematunge C, Pussegoda K, Corrin T, Waddell L. Tick exposures and alpha-gal syndrome: a systematic review of the evidence. Ticks Tick Borne Dis 2021;12(3):101674 [FREE Full text] [doi: 10.1016/j.ttbdis.2021.101674] [Medline: 33529984]
19. Molaei G, Little EAH, Williams SC, Stafford KC. Bracing for the worst - range expansion of the lone star tick in the Northeastern United States. N Engl J Med 2019;381(23):2189-2192 [FREE Full text] [doi: 10.1056/NEJMp1911661] [Medline: 31800982]
20. Binder AM, Commins SP, Altrich ML, Wachs T, Biggerstaff BJ, Beard CB, et al. Diagnostic testing for galactose-alpha-1,3-galactose, United States, 2010 to 2018. Ann Allergy Asthma Immunol 2021;126(4):411-416.e1 [FREE Full text] [doi: 10.1016/j.anai.2020.12.019] [Medline: 33422649]
21. Iglesia EGA, Stone CA, Flaherty MG, Commins SP. Regional and temporal awareness of alpha-gal allergy: an infodemiological analysis using Google Trends. J Allergy Clin Immunol Pract 2020;8(5):1725-1727.e1 [FREE Full text] [doi: 10.1016/j.jaip.2019.12.003] [Medline: 31857260]
22. Carson DA, Kopsco H, Gronemeyer P, Mateus-Pinilla N, Smith GS, Sandstrom EN, et al. Knowledge, attitudes, and practices of illinois medical professionals related to ticks and tick-borne disease. One Health 2022;15:100424 [FREE Full text] [doi: 10.1016/j.onehlt.2022.100424] [Medline: 36277108]
23. Flaherty MG, Kaplan SJ, Jerath MR. Diagnosis of life-threatening alpha-gal food allergy appears to be patient driven. J Prim Care Community Health 2017;8(4):345-348 [FREE Full text] [doi: 10.1177/2150131917705714] [Medline: 28447914]
24. Flaherty MG, Threats M, Kaplan SJ. Patients' health information practices and perceptions of provider knowledge in the case of the newly discovered alpha-gal food allergy. J Patient Exp 2020;7(1):132-139 [FREE Full text] [doi: 10.1177/2374373518808310] [Medline: 32128382]
25. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
26. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of Google Trends in health care research: a systematic review. PLoS One 2014;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
27. Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (infoveillance): systematic review. J Med Internet Res 2020;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
28. Glimpse home page. Glimpse. 2023. URL: https://meetglimpse.com/extension/ [accessed 2023-03-02]
29. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health 2017;7(3):185-189 [FREE Full text] [doi: 10.1016/j.jegh.2017.06.001] [Medline: 28756828]
30. Alibudbud R. Google Trends for health research: its advantages, application, methodological considerations, and limitations in psychiatric and mental health infodemiology. Front Big Data 2023;6:1132764 [FREE Full text] [doi: 10.3389/fdata.2023.1132764] [Medline: 37050919]
31. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical approaches in Google Trends research: systematic review. J Med Internet Res 2018;20(11):e270 [FREE Full text] [doi: 10.2196/jmir.9366] [Medline: 30401664]
32. Neumann K, Mason SM, Farkas K, Santaularia NJ, Ahern J, Riddell CA. Harnessing google health trends data for epidemiologic research. Am J Epidemiol 2023;192(3):430-437 [FREE Full text] [doi: 10.1093/aje/kwac171] [Medline: 36193858]
33. Jordan RA, Egizi A. The growing importance of lone star ticks in a lyme disease endemic county: passive tick surveillance in Monmouth County, NJ, 2006 - 2016. PLoS One 2019;14(2):e0211778 [FREE Full text] [doi: 10.1371/journal.pone.0211778] [Medline: 30753233]
34. Preventing tick bites. Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/ticks/avoid/on_people.html#:~:text=Tick [accessed 2024-02-22]
35. Lam HCY, Turner PJ, Hemming D, Jarvis DL. Seasonality of food-related anaphylaxis admissions and associations with temperature and pollen levels. J Allergy Clin Immunol Pract 2021;9(1):518-520.e2 [FREE Full text] [doi: 10.1016/j.jaip.2020.07.032] [Medline: 32745700]
36. Clegg LX, Hankey BF, Tiwari R, Feuer EJ, Edwards BK. Estimating average annual per cent change in trend analysis. Stat Med 2009;28(29):3670-3682 [FREE Full text] [doi: 10.1002/sim.3733] [Medline: 19856324]
37. Alkishe A, Raghavan RK, Peterson AT. Likely geographic distributional shifts among medically important tick species and tick-associated diseases under climate change in North America: a review. Insects 2021;12(3):225 [FREE Full text] [doi: 10.3390/insects12030225] [Medline: 33807736]

XSL•FO
RenderX

38. Richards NE, Richards RD. Alpha-gal allergy as a cause of intestinal symptoms in a gastroenterology community practice. South Med J 2021;114(3):169-173 [FREE Full text] [doi: 10.14423/SMJ.0000000000001223] [Medline: 33655311]

Abbreviations

AAPC: average annual percent change
ICD-10: International Classification of Diseases 10th Revision
RSV: relative search volume

Edited by A Mavragani; submitted 14.06.23; peer-reviewed by JR Medina, A Allam, S Wei; comments to author 08.02.24; revised version received 28.02.24; accepted 23.03.24; published 08.05.24.

Please cite as:
Romeiser JL, Jusko N, Williams AA
Emerging Trends in Information-Seeking Behavior for Alpha-Gal Syndrome: Infodemiology Study Using Time Series and Content Analysis
J Med Internet Res 2024;26:e49928
URL: https://www.jmir.org/2024/1/e49928
doi:10.2196/49928
PMID:38717813

©Jamie L Romeiser, Nicole Jusko, Augusta A Williams. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

News Coverage of the COVID-19 Pandemic on Social Media and the Public's Negative Emotions: Computational Study

Hanjing Wang[1*], MSc; Yupeng Li[2*], PhD; Xuan Ning[3], PhD

[1]School of Communication, Hong Kong Baptist University, Hong Kong, China (Hong Kong)

[2]Department of Interactive Media, Hong Kong Baptist University, Hong Kong, China (Hong Kong)

[3]Department of Social Sciences, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

[*]these authors contributed equally

Corresponding Author:
Yupeng Li, PhD
Department of Interactive Media
Hong Kong Baptist University
Kowloon Tong, Kowloon
Hong Kong, 000000
China (Hong Kong)
Phone: 852 3411 8263
Email: ivanypli@gmail.com

Abstract

Background: Social media has become an increasingly popular and critical tool for users to digest diverse information and express their perceptions and attitudes. While most studies endeavor to delineate the emotional responses of social media users, there is limited research exploring the factors associated with the emergence of emotions, particularly negative ones, during news consumption.

Objective: We aim to first depict the web coverage by news organizations on social media and then explore the crucial elements of news coverage that trigger the public's negative emotions. Our findings can act as a reference for responsible parties and news organizations in times of crisis.

Methods: We collected 23,705 Facebook posts with 1,019,317 comments from the public pages of representative news organizations in Hong Kong. We used text mining techniques, such as topic models and Bidirectional Encoder Representations from Transformers, to analyze news components and public reactions. Beyond descriptive analysis, we used regression models to shed light on how news coverage on social media is associated with the public's negative emotional responses.

Results: Our results suggest that occurrences of issues regarding pandemic situations, antipandemic measures, and supportive actions are likely to reduce the public's negative emotions, while comments on the posts mentioning the central government and the Government of Hong Kong reveal more negativeness. Negative and neutral media tones can alleviate the rage and interact with the subjects and issues in the news to affect users' negative emotions. Post length is found to have a curvilinear relationship with users' negative emotions.

Conclusions: This study sheds light on the impacts of various components of news coverage (issues, subjects, media tone, and length) on social media on the public's negative emotions (anger, fear, and sadness). Our comprehensive analysis provides a reference framework for efficient crisis communication for similar pandemics at present or in the future. This research, although first extending the analysis between the components of news coverage and negative user emotions to the scenario of social media, echoes previous studies drawn from traditional media and its derivatives, such as web newspapers. Although the era of COVID-19 pandemic gradually brings down the curtain, the commonality of this research and previous studies also contributes to establishing a clearer territory in the field of health crises.

(J Med Internet Res 2024;26:e48491) doi:10.2196/48491

KEYWORDS

web news coverage; emotions; social media; Facebook; COVID-19

XSL•FO
RenderX

Introduction

Background

The COVID-19 pandemic was one of the most unprecedented global crises in human history. Researchers have been groping for direct countermoves to the COVID-19 disease since its initial occurrence, with massive collaborative efforts committed. Notwithstanding, the pandemic posed evolving challenges to public health, politics, the economy, and personal well-being. The uncertainty that comes along with pandemics greatly disturbed people's current lives and changed their expectations for the future. The public health situation in Hong Kong was once severe, with >10,000 new confirmed cases on average being reported each day during the fifth wave of the COVID-19 pandemic [1]. It is astonishing that Hong Kong ranked the highest fatality rate in the world given its first-class medical care quality [2].

Pandemics usually bring negative emotions, such as anger, sadness, fear, etc [3]. Before the rebound of the pandemic, Hong Kong had been expecting the border to reopen with mainland China after rounds of negotiations as proof of efforts to keep the pandemic situation under control. The suddenly acute situation rapidly crashed people's dreams of reuniting with family or regular economic exchanges, making them feel at sea, along with the unavoidable increased level of depression [4,5]. It is also reported that citizens panic bought daily necessities amid the possibility of a citywide lockdown [6], when a substantial number of residents experienced anxiety, loneliness, and even mental health issues [7]. The negative feelings of individuals, such as fear, anger, and sadness, which are bound to impair people's state of mind, call for close attention because they will affect people's cognitive as well as their physical health [8]. When it comes to communities, the collective sentiment is also devoted to shaping the distribution of formal social control measures, which is an important expression of the collective sentiment [9]. The negative emotions, if constantly diffused, could contribute to the evolution of a societal negative emotional climate, which might lead to changes in personal values and further political consequences and normative implications [10].

The primary and authoritative source that informs citizens of pandemic situations and public health responses is news media, which is an institutionalized channel that broadcasts crisis information to a broad audience and is an inseparable part of political crisis management [11,12]. Meanwhile, the internet offers a capacious playground for official news accounts to play their roles while meeting the goal of being timely. Among the digital organisms that rejuvenate news media, social media is a representative information carrier for content consumption during crises [13]. News consumption on social media has been a prevalent practice for audiences to establish connections with the uncharted world, thereby affecting their understanding of news subjects and events unconsciously. This is particularly true for those who are under quarantine and obeying social distancing rules during the urgent level of a pandemic. When web news coverage of a crisis triggers the audience's emotions, the collective sentiment concerning cultural, social, political, or economic events can be directly observed via user interactions on the same platform. Hence, the investigation of news coverage on social media may directly enable the news media to play a better role as a gatekeeper in crisis communication.

Most of the COVID-19–related research on web-based emotional expressions focuses on describing components and intensity instead of attempting to explain the reason for the emergence of the emotions [14]. Specifically, we have limited systematic knowledge of the connection between crisis coverage on social media and users' emotionality [15]. Moreover, existing studies usually set aggregate emotionality (ie, the sum of all the specific emotions) as their target instead of focusing on the negative emotions, which are worth extra attention. A deeper understanding of emotional responses that might deteriorate mental health is crucial for the well-being of individuals caught in a crisis. Besides, negative emotions and positive emotions amid a crisis function oppositely regarding perceptions; for example, the former may foster a greater perceived threat and can contribute to far-reaching impacts on individuals, such as their shift in values [10]. Thus, we devoted our efforts to filling the identified research gap by endeavoring to explain the negative emotions of the audience from the perspective of news coverage on social media. Given our research context of Hong Kong, we chose to examine COVID-19–related news posts and their associated comments on Facebook, the most popular and widely used social platform in Hong Kong [16].

Accordingly, our study proposed two research questions (RQs):

- RQ1: what and how are COVID-19–related issues and subjects of web news coverage by official news accounts associated with negative emotions in Facebook users' comments in Hong Kong?
- RQ2: what are the effects of media tone and the interaction effects of media tone with issues or subjects on negative emotions in Facebook users' comments in Hong Kong?

In this regard, we crawled Facebook COVID-19–related posts published by representative Hong Kong news organizations and the corresponding comments to investigate how news coverage on social media affects the public's negative emotions amid the fifth wave of the pandemic. We situated the examination of negative emotional responses and crucial news elements of subjects, issues, media tone, and length in the framework of appraisal theory. Specifically, we first used seed words, such as "COVID-19," "coronavirus," "Omicron," etc, to filter out the irrelevant, that is, not COVID-19–related, posts and identify other related words for post selections. Afterward, we managed to locate the focal points of societal attention under the threat of COVID-19 through frequent words and topic models, and we mapped keywords to identify issues and subjects. Then, we leveraged state-of-the-art machine learning techniques to reveal the emotion dynamics and sentiment embedded in the news posts. Finally, we used regression analysis with fixed effects to better understand and explain users' negative emotional responses.

XSL•FO
RenderX

Research Framework and Hypotheses

Overview

We situated our research framework in appraisal theory [17-19]. Our motivation for applying appraisal theory to analyze web news coverage was that emotions emerge through the evaluative process of particular circumstances, eliciting distinct responses among individuals [20]. The appraisal theory laid the foundation for sentiment extraction and thus provided insights for understanding the emotions of social media posts, such as content created by microbloggers to explore the minds of authors [20]. In our paper, we constructed the examination of subjects and issues studied by previous studies [15], together with other identified crucial news elements, media tone, and length, on the basis of the framework of appraisal theory. We studied the appraisal expressions from the aspect of attitudes [21], that is, the negative emotions (anger, fear, and sadness) of the appraiser, that is, the audiences or the commenters of the news posts. We sought to develop a research model based on the effect of news coverage from the perspectives of issues, subjects, media tone, and news length on users' negative emotions. We put forth each hypothesis after discussing the theoretical basis. Our research model is presented in Figure 1.

Figure 1. Theoretical framework. HK: Hong Kong; RQ: research question.

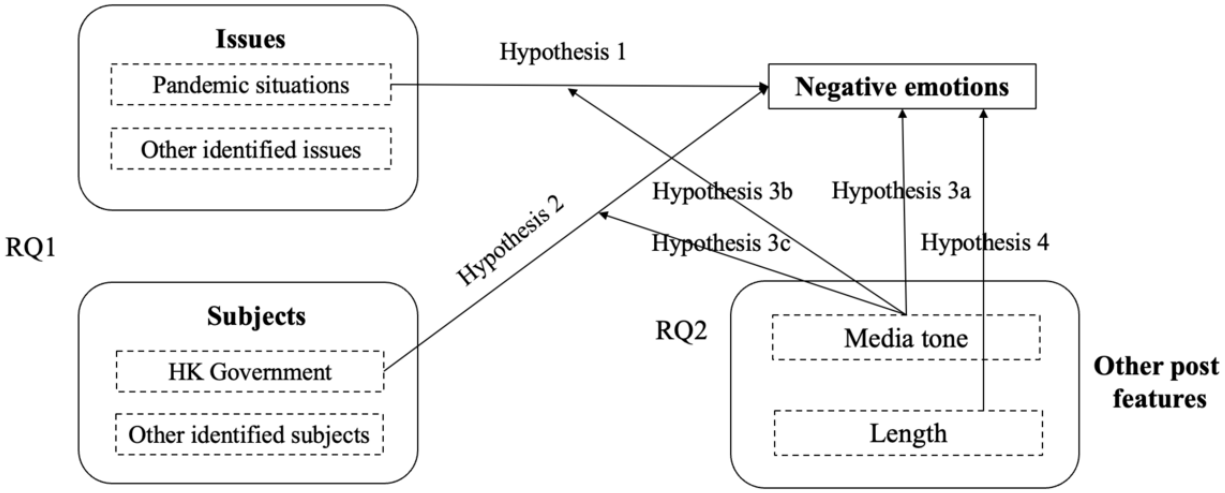

Issues and Subjects

According to appraisal theory, emotions are prompted based on subjective assessments of events and situations according to personal significance [18]. The elements of news coverage may arouse the feelings of the audience, and the individual differences in the appraisal of news content may be illustrated in variations of audience responses [22], which is also applicable in crisis communication [23,24]. Prior studies identified news topics and the content features per se as critical factors in explaining the engagement of commenting behavior [25]. Users that are aroused by the covered topics may desire to express their opinions or attitudes and thus have the motivation to comment with emotions [26,27]. Existing findings explore the different effects of news stories' framing elements on readers' responses. It is suggested that news topics such as political conflicts have more comments but also more toxicity, while topics such as science and technology induce fewer toxic comments [28,29]. Specific subtopics of health in web news are also found to contribute to different idiosyncratic reader responses [30]. Scholars have applied categorized issues and local political actors to analyze the web news coverage of websites in the early stage of COVID-19 [15].

In this study, drawing on appraisal theory, we intended to postulate the impacts of issues and subjects covered in news posts on the negative emotions in users' comments, which refer to the mentioned important combined subtopics and the subjects in the posts, respectively. Specifically, we decided to examine the role of issues and subjects that are characterized as ubiquitously dramatic and evocative for a lasting period, which also received adequate attention from mainstream media [31]. On the basis of these criteria, one of the most representative and worldwide universal issues is pandemic situations [32-34], in which information such as confirmed cases and regional risk levels is presented. As for subjects, the government, which acts as a crisis manager, plays an essential role in crisis management [15]. The performance of the government in coping with the crisis may arouse public emotions [35,36], and the clues of government inefficiency in problem-solving are prone to stimulating negative feedback [25,37].

Hence, regarding RQ1, we proposed the following hypotheses:

- Hypothesis 1: the occurrence of the issue about pandemic situations in the COVID-19–related news post would be associated with a higher level of negative emotions in user comments.
- Hypothesis 2: the occurrence of the subject about the Government of Hong Kong in the COVID-19–related news post would be associated with a higher level of negative emotions in user comments.

XSL•FO
RenderX

Media Tone

The news coverage sometimes manifests a preferred tone selectively. Mass communication has priming effects on the general audience given the selective attention of people [38,39]. The evaluative tone of media coverage is conceptualized as a composition of second-level agenda setting [40]. People routinely draw upon particularly salient information, which is shaped considerably by mass media. Media tone provides us a more detailed understanding of incidents together with subsequent attitudes and opinions [41], public's perceptions regarding the economy [42], or even influences observable behavior [39]. Research on economic news coverage suggests that newspapers exhibit a negativity bias while Twitter exhibits a positivity bias toward the same undesirable event [43]. The media tone may directly act on the readers as the message valence may affect the emotional and cognitive responses of audience members [44]. Multifarious studies have shown that the emotional valence of crisis media content has a significant impact on user engagement [45,46], but whether the conclusions are safely applicable to emotional responses is still in the darkness. Eisele et al [15] found that the negativity and positivity of web news articles both have limited effects on the emotionality of commenting; meanwhile, the contrasting evidence shows that stronger audience reactions often accompany negative news [47,48]. The audience's bias may lead to asymmetrical effects on the audience under positive and negative tones; additionally, the latter could have polarizing power on the news audience [49] and a greater effect on message reception [50]. The negatively presented crisis may arouse pessimism. Moreover, the audience would blame an undesirable event in negatively framed news on the government [51].

With reference to RQ2, we formulated the following hypothesis:

- Hypothesis 3a: the negative tone of a COVID-19–related news post would be more likely to increase the level of negative emotions in user comments than a positive tone.
- Hypothesis 3b: the interaction effect of a negative tone in the COVID-19–related news post and the occurrence of the issue of pandemic situations would increase the level of negative emotions in user comments.
- Hypothesis 3c: the interaction effect of a negative tone in the COVID-19–related news post and the occurrence of the subject of the Government of Hong Kong would increase the level of negative emotions in user comments.

Post Length

Length is one of the crucial variables that characterize content [52], whose impacts on user engagement are often discussed [53,54]. Existing studies consider content length as one of the dimensions of reviewers' perception and evaluation based on cognitive overload theory [55]. When individuals are experiencing information processing, sometimes not all the inputs can be fully digested and used, which refers to cognitive overload [56,57]. For the extrinsic load, which means the individuals are free of the inherent trouble with information processing, the load is induced by how the content is organized [58-60]. Prior studies have discovered that a total lack of text or overabundant text is inefficient in provoking engagement, in contrast to an appropriate text scale [55,61]. The extra cognitive load is also found to have positive effects on negative emotions, which suggests that the information contained within the content once exceeds the scope of proper readability would induce negativeness in readers [62,63]. We conceptualized length as a crucial news feature because the size of the news can affect the diversity of content, the thinking depth of the audience, and the attractiveness of the news [64,65]. Therefore, we developed a hypothesis as follows:

- Hypothesis 4: the length of the COVID-19–related news post would be curvilinearly related to the level of negative emotions in user comments.

Methods

Overview

We chose Facebook as the target social platform for our research for the following reasons. First, Facebook is one of the most influential and popular social media platforms globally and the most used platform in Hong Kong [16]. Facebook users are concentrated in the age group of 18 to 44 years, which accounts for more than half the population in Hong Kong [66]. In addition, news organizations in Hong Kong use Facebook as an integrated channel for publishing up-to-date news, self-promotion, and connecting with readers, which allows researchers to observe their interactions from public pages directly.

For variable construction, we leveraged 2 fine-tuned models of Bidirectional Encoder Representations from Transformers (BERT) [67,68], which is considered the best-performing computational method [69], to classify the emotions revealed in user comments and the sentiment of the posts, respectively. For news coverage, we mainly applied the topic model for subtopics and references in dictionary development and then mapped the preprocessed news pieces into issues and subjects. Finally, we adopted negative binomial regression for data analysis. The measurements of variables were conducted in Python (Python Software Foundation), and the statistical analysis was implemented in RStudio (Posit). The analytical flow is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Analytical flow. BERT: Bidirectional Encoder Representations from Transformers.

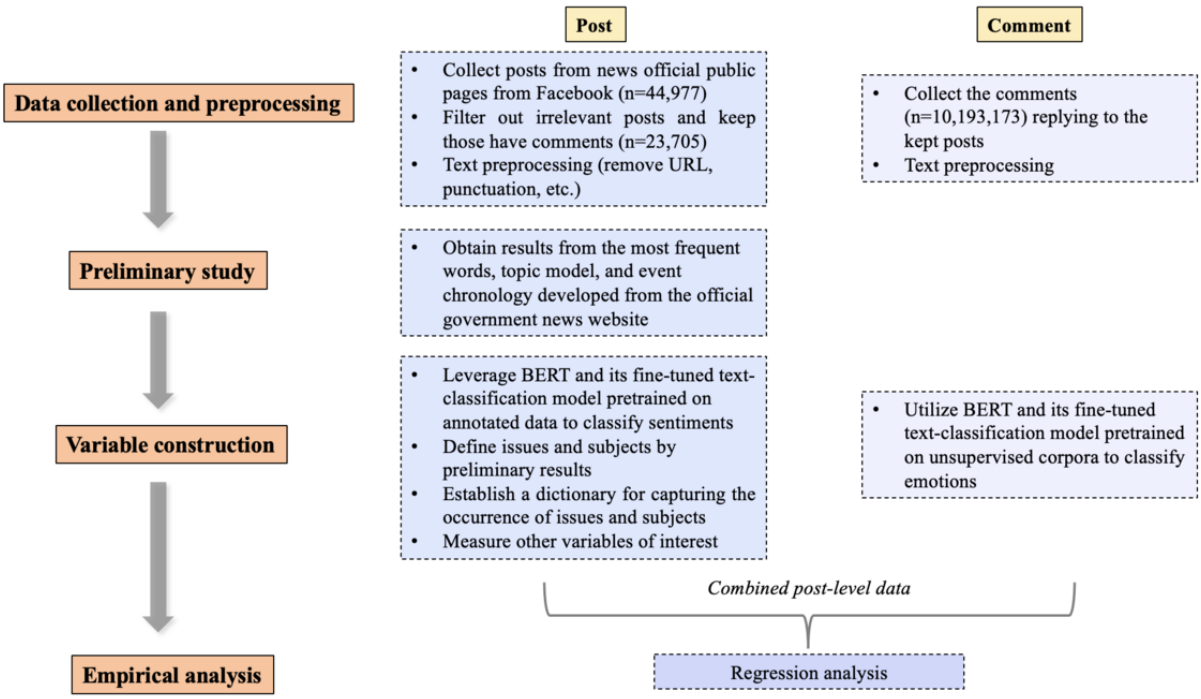

Data Collection and Preprocessing

To shed light on the emotional dynamics, we crawled posts and comments on Facebook. Specifically, we first determined the public pages of the major mainstream news accounts (refer to Table S1 in Multimedia Appendix 1), with about 505,371 (SD 364,386) followers on average. The collection period ranges from January 2022 to May 2022, which almost covered one complete circle of the fifth wave of COVID-19 in Hong Kong [70]. To filter the COVID-19–related posts, we developed a list of keywords, as shown in Textbox 1.

We started by searching with the seed words, such as "COVID-19," "Omicron," and "fifth wave," that commonly occur in public view, and we added words of posts in the searching results that both concurred with the seed words and are directly related to the pandemic. The process would be repeated until the keywords list remained unchanged for an extra 50 posts. The posts that merely hit the keywords but were concerned about irrelevant issues, such as the promotion slogan of news organization apps, were left out. The comments on the filtered posts were collected correspondingly. The filtering criteria reached the agreement of all the coauthors. After filtering out those irrelevant posts, we collected 40,701 posts, of which 23,705 posts have 1,019,317 comments in total.

Before the measurement of variables, we first cleaned the texts for subsequent analysis. The messages on social platforms are usually presented informally, which might influence the machine learning model performance [71]; thus, text preprocessing is an essential prerequisite before the model training [72-74]. To clean the data, the posts and comments were processed by removing the URLs as well as redundant punctuation and normalizing the whitespaces. We then leveraged the OpenCC library [75] to convert simplified Chinese characters to Hong Kong standard traditional Chinese characters.

Textbox 1. Keywords for extracting COVID-19–related posts.

Original keywords

- *Covid*, 疫苗, 新型冠狀病毒, 新冠, 疫情, *Omicron*, 奥密克戎, 抗疫, 快測申報, 檢測, 確診, 無症狀, 陽性, 防疫, 隔離, 禁足, 強制檢測, 方艙, 封城, 第五波, 快速檢測, 動態清零

Translation

- *Covid, vaccine, COVID-19, pandemic situation, Omicron, anti-pandemic, rapid diagnostic test (report), confirmed cases, asymptomatic infection, tested positive, pandemic prevention, quarantine, lockdown, compulsory testing, mobile cabin hospital, fifth wave, dynamic clearing.*

Ethical Considerations

This study is a computational analysis that uses publicly available data, which do not include personally identifiable information or sensitive individual data. Our research complies with the principles of ethical conduct in research involving data without interaction with human subjects and identifiable private information. As such, the nature of this study did not necessitate the application for ethics board review or approval. All the data are securely stored and only accessed by personnel involved in the research team.

XSL•FO
RenderX

Operation of Variables

Dependent Variable

Our dependent variable was the negative emotions conveyed by the user comments, which was measured by the aggregate number of comments labeled as negative emotions under each post. Automatic recognition of emotional information in social media texts based on natural language processing technology can help understand the attitude of netizens toward various events and portray people's emotional fluctuations on a large scale over time. To this end, we adopted an effective method [76] to fine-tune a pretrained language model called BERT [67], which has proven to be empirically powerful in natural language processing. In carrying out super large-scale pretraining on unsupervised corpora, the fine-tuned BERT model adopts the transformer encoder as the language model and infers the relationship between input and output completely through the attention mechanism [68]. The model's architecture is provided in Figure 3.

We trained the deep learning model on the SMP2020-EWECT COVID-19 pandemic Weibo data set [77]. We adopted the data set from Weibo because Hong Kong and the Chinese mainland share a generally similar sociocultural context. While the model was trained on a simplified Chinese data set, preliminary experiments show that models trained on simplified Chinese data sets can be directly used to infer Cantonese texts. The data set consists of the COVID-19–related microblogs since the outbreak of the COVID-19 pandemic, that is, the user-generated postings, on Weibo, which is one of the largest Chinese social media platforms. Each microblog has been previously labeled in the categories "happy," "surprise," "neutral," "sad," "fear," and "angry," and the last 3 categories of emotions belong to negative emotions. The split of the training set, validation set, and test set was predetermined to be around 8:2:3. We used microblogs in the training set for model training and the predicted results of the labeled microblogs for calculating cross-entropy loss. The evaluation of the test set suggested that the model can achieve an accuracy of 0.772 and a macroaveraged F_1-score of 0.739, which has a better performance than classifiers of similar tasks [78]. Since the recent development of large language models, ChatGPT (OpenAI) [79] has achieved comparable performance compared with BERT in sentiment analysis [80], and we evaluated the performance of BERT and ChatGPT 3.5 and validated it by human assessment of random samples. Overall, BERT performed better than ChatGPT with an accuracy of 0.73 versus 0.48 on the basis of human assessment. The detailed assessment of the models is shown in Table 1.

The output of the emotion detection model applied to our collected data was one-hot encoded. In other words, if a comment is recognized as one kind of emotion, for instance, "angry," the value of that comment under the category "angry" would be "1"; otherwise, it would be "0." As our dependent variable was the negative emotions, we only focused on the emotions "angry," "fear," and "sad." The number of emotions belonging to the comments under the same post were aggregated. The examples of 6 emotions and their average amount per post are presented in Table 2.

Figure 3. Analytical flow. BERT: Bidirectional Encoder Representations from Transformers. CLS: the special token "classification".

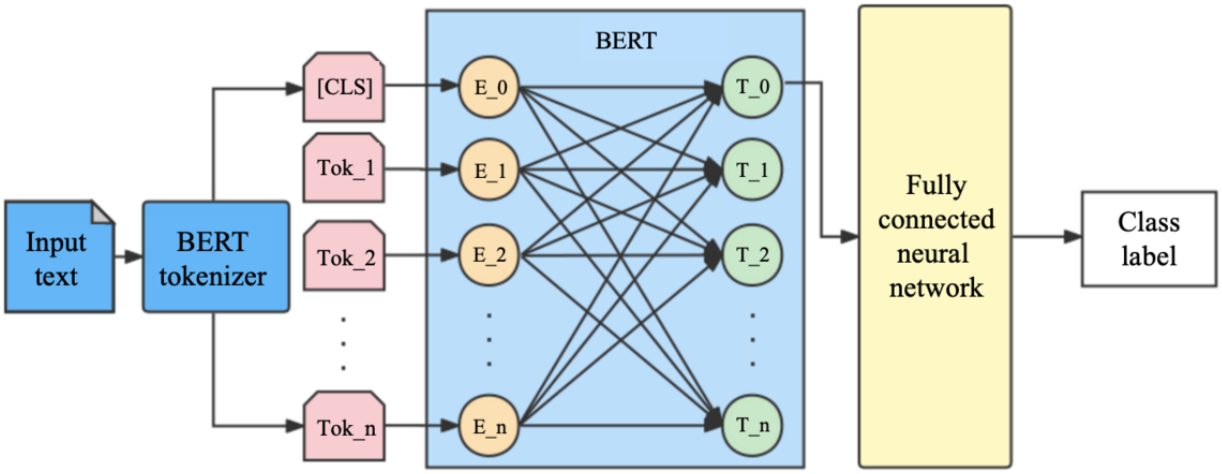

XSL•FO
RenderX

Table 1. Comparison of performance metrics of Bidirectional Encoder Representations from Transformers (BERT) and ChatGPT in emotion detection.

Model and emotion	Precision	Recall	F_1-score
BERT			
Angry	0.40	0.89	0.55
Sad	0.99	0.79	0.88
Fear	0.36	0.73	0.48
Neutral	0.96	0.51	0.66
Happy	0.80	0.86	0.83
Surprise	0.85	0.87	0.86
ChatGPT			
Angry	0.37	0.44	0.40
Sad	0.75	0.44	0.55
Fear	0.48	0.49	0.48
Neutral	0.42	0.69	0.53
Happy	0.68	0.39	0.49
Surprise	0.35	0.21	0.27

Table 2. Examples of emotions and average number per post.

Emotion	Examples	Number per post, mean (SD)
Happy	"A thought on the wrong path, thankfully no misfortune has happened! The hard times will eventually pass, hang on!"	6.28 (67.84)
Angry	"With so much money spent on building cabins, is there a place for the homeless?"	24.13 (51.73)
Sad	"They are all poor people, so desperate that they have to rob, but the shopkeepers already barely make a living."	0.83 (2.39)
Fear	"I'll be scared to death if not die of sickness. I'm extremely anxious!"	1.57 (3.94)
Surprise	"Losing millions every month? Unbelievable"	0.69 (1.94)
Neutral	"Put more money into medical care. Focus on the elderly and children."	10.95 (67.20)

Independent Variables and Control Variables

Issues and Subjects

As the pandemic evolution, policies, and terminologies are highly varied among countries and regions, there is no universally available or applicable taxonomy for us to map the news coverage of COVID-19 in Hong Kong on Facebook. Thereby, we applied the philosophy of grounded theory [81] to learn the structure of content and establish the foundation for the mapping of news coverage. To better grasp the overview of our post data, we first collected the frequent words of monthly posts and the total posts to locate those words that are consistent across months or time varying to have an initial understanding of our post data. We limited the frequent words to the top 50 words for consideration of relevance and representativeness after comparison. The results from comparison indicated that pandemic situations, testing, confirmed cases, vaccines, and economic consequences were the most outstanding and commonly shared words across months.

Then, to layout the topics for the preparation of the topics and subjects, we used the latent Dirichlet allocation topic model [82]. Latent Dirichlet allocation does well at discovering latent topics from given words and documents, which enables us to encode many different kinds of prior knowledge [83]. To determine the optimal number of topics, we calculated the perplexity and coherence scores of models with varied numbers of topics separately in Figures 4 and 5. A lower perplexity score and a higher coherence score suggest a topic model with better performance [84]. Taking the perplexity and coherence scores into account, we decided 8 to be the topic numbers in cases of overfitting issues. The topics along with their top words are shown in Table S2 in Multimedia Appendix 1. The output of our topic model divided the content into 8 subtopics, which required further inspection because we intended to disentangle the issues and the subjects that may intersect among the topics.

Afterward, we randomly sampled 50 posts each from the collected data of news organization accounts to identify categories of issues and subjects based on preliminary work, to which we constantly compared the posts and assigned them accordingly. We determined the categorization rubrics of issues according to two criteria: (1) each post is supposed to mention at least one of the identified issues and (2) the keywords of issues and the subjects should be mutually exclusive. The categorization of subjects was similar, but the posts may not

XSL•FO
RenderX

necessarily mention any of the subjects. After defining the categories, we set up unique keywords for each issue and sector for labeling the posts. Aside from the data we acquired, we also developed a highlighted-event chronology from the Government of Hong Kong's official news website [85] to help validate the process of establishing dictionaries and ensure the credibility of the construction of news coverage. The keywords were selected through initial screening based on findings from the most frequent words and topic model and then modified through rounds of validation on a random set of 50 posts each time. The adjustment to keywords was repeated until the dictionaries did not change further. The categories of issues and subjects with examples and times of occurrences are presented in Table 3.

Figure 4. Perplexity scores under different numbers of topics.

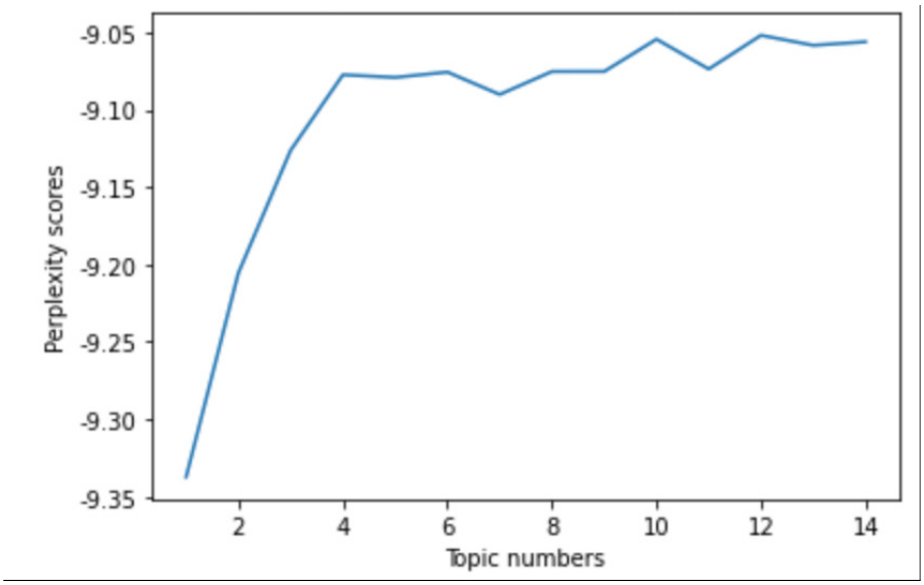

Figure 5. Coherence scores under different numbers of topics.

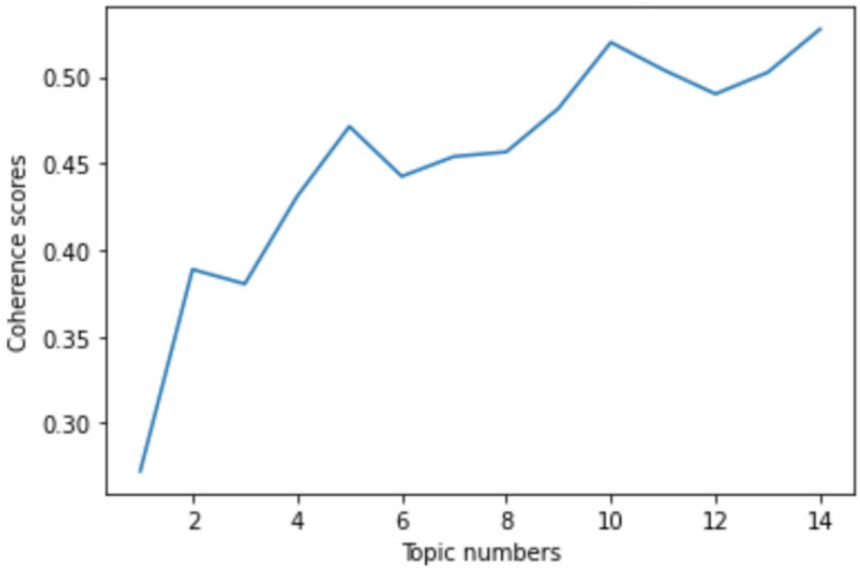

XSL•FO
RenderX

Table 3. Examples and occurrences of the categorized issues and subjects.

Category	Examples	Occurrences, n
Issues		
Pandemic situations	"Five members of the Cathay Pacific crew were infected with the Omicron variant virus."	18,929
Supportive actions	"The sixth round of the AEF comprises 48 schemes which are estimated to benefit 67,000 businesses, operators of 40,000 transport tools and 750,000 individuals."	1443
Information and advice	"The WHO does not consider the Omicron variant IHU to be a significant threat at this time."	8596
Antipandemic measures	"In view of the rapidly worsening global pandemic situation due to the Omicron variant, passenger flights from eight countries will not be permitted to land in Hong Kong from January 8."	12,523
Subjects		
Government of Hong Kong	"Chief Executive: surrendering to the virus is not an option."	7992
Central government	"The Leader of the Mainland Chinese medicine expert group of the Central Authorities, Mr Tong Xiaolin, today led a delegation of Chinese medicine experts from the Mainland to Hong Kong for field visits and offering support in Hong Kong's fight against the pandemic."	1808
Functional sectors	"The Hospital Authority said its main strategy for adjusting some nonemergency services is to safeguard the public."	1795

Media Tone

To determine the sentiments of posts, we applied the data set of sentiment detection of Weibo users during the COVID-19 pandemic [86] to fine-tune a BERT-based text classification model for sentiment analysis. The data set includes 1,000,000 Weibo posts from January 1, 2020, to February 20, 2020, and 10% of them were previously labeled with 1 (positive), 0 (neutral), and −1 (negative). The model architecture for media tone detection is similar to that in Figure 3 to infer the sentiment of the collected posts. It turns out that COVID-19 coverage on social media in Hong Kong was dominated by a neutral tone, with 226 posts with negative tone and 461 with positive tone. The detailed assessment of model metrics is listed in Table 4.

Table 4. Comparison of performance metrics of Bidirectional Encoder Representations from Transformers (BERT) and ChatGPT in sentiment analysis.

Model and tone	Precision	Recall	F_1-score
BERT			
Positive tone	0.72	0.96	0.82
Negative tone	0.97	0.95	0.96
Neutral tone	0.96	0.78	0.86
ChatGPT			
Positive tone	0.57	0.87	0.69
Negative tone	0.67	0.94	0.78
Neutral tone	0.76	0.26	0.39

Length

We used Python to measure the length of the string of post content; in other words, the number of words each post contains is treated as the numerical value of the variable length (mean 212.44, SD 152.24).

Besides, extant studies found a rise in negative emotional responses with the increase in COVID-19 cases [87], so we controlled the *daily confirmed cases* to indicate the severity of the pandemic at the time of posting, which is the number of new COVID-19 cases within the same day [88]. We also controlled the *fixed effects of the posting accounts of news organizations* in the form of dummy variables to control for variables that are not observable or measurable across the accounts [89,90] and *the fixed effects of time by adding dummy variables for months in the regression models* [55,91].

Data Analysis

Considering that most of our variables are dummy variables and categorical variables, we used the negative binomial model, which is suitable for modeling count data with overdispersion and is commonly applied in social media research [92-94]. To examine the conditions for the negative binomial model, we first calculated the ratio of residual deviance and its *df* in a Poisson regression model, which was 35.657 (much greater than 1). Moreover, we calculated the variance inflation factors in a multiple regression model to examine the multicollinearity [95]. As Table S3 in Multimedia Appendix 1 shows, the variance

XSL•FO
RenderX

inflation factors of all the variables were not unusually >1.0, which means there is no concern for multicollinearity. Consequently, we chose the negative binomial regression on postlevel data to explore the characteristics that are supposed to explain the negative emotions revealed in the comments. All the regression analyses were conducted using the package *MASS* [96] in RStudio.

Results

Descriptive Summary

Figure 6 illustrates the temporal dynamics of each type of emotion. The dashed lines represent each emotion, while the solid line represents the daily confirmed cases. Overall, anger occupied the largest proportion of all emotions. From the selected events labeled in Figure 6, we can tell that the temporary peaks of the negative emotions coincided with the representative events we identified from the aforementioned work of event chronology and distinct terms of the most frequent words. The exploratory findings demonstrated that users' negative emotions may be aroused by particular events.

Figure 6. Emotion dynamics of detailed categories and daily confirmed cases.

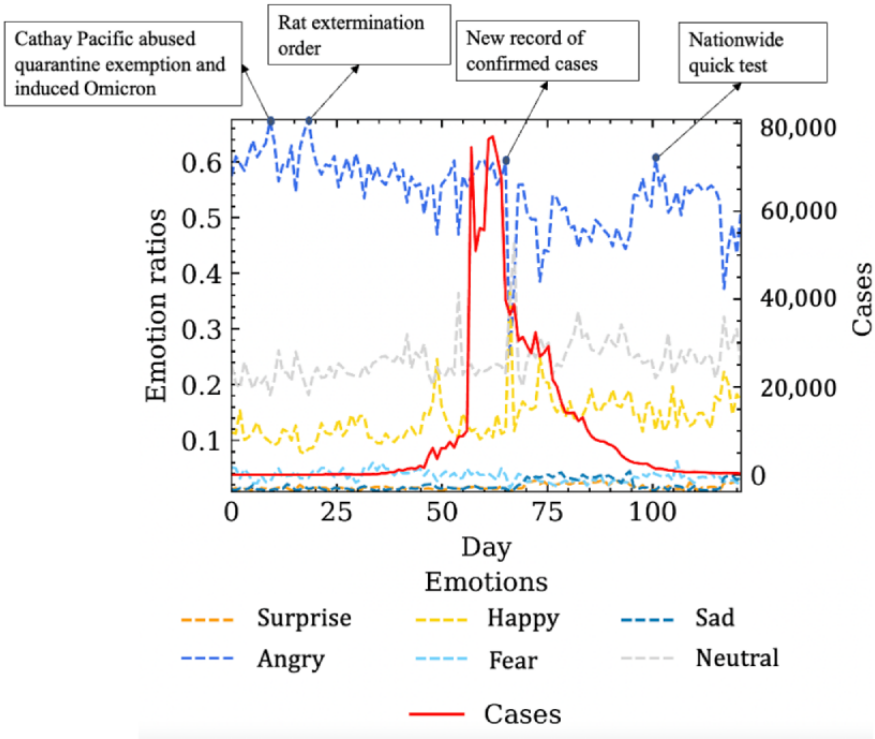

Regression

Aside from the dummy variables, we normalized the rest of the variables because of their differentiated magnitudes. We applied R (version 4.1.2; R Foundation for Statistical Computing) to implement data analysis. Table 5 depicts the statistics of regression, in which we tested the effects of issues, subjects, and other content characteristics of news posts on the users' negative emotions.

With the fixed effect of publishing account and months controlled, we first included the main variables of interest in model 1 (ie, issues, subjects, and media tones), then we added the interaction effects of media tone (ie, media tone×Government of Hong Kong and media tone×pandemic situation) in model 2, and we finally added squared variables (ie, squared terms of length and case number) in model 3.

The salience of issues in news coverage reflected the primary concerns of the citizens facing the pandemic crisis. Among the 4 issues, antipandemic measures (B=–0.050; P=.006), pandemic situations (B=–0.261; P<.001), and supportive actions (B=–0.083; P=.02) had a significant negative influence on the negative emotions; that is, the occurrence of these 3 issues could decrease the log count of negative user emotions. The effect of the issue on information and advice, although not significant and minuscule (B=0.005; P=.80), illustrated a positive impact on the log count of negative emotions. The prominent subjects, on the other hand, demonstrated those whom the public considers to play a crucial part in crisis administration. Both the governments of the Chinese mainland (B=0.120; P<.001) and Hong Kong (B=0.324; P<.001) had positive impacts on users' negative emotions. Besides, the results illustrated that the existence of functional sectors had a negative effect, although not that significant, on negative emotional responses (B=–0.058; P=.09).

Compared with the positive tone embedded in the news posts, the neutral tone (B=–0.338; P<.001) and negative tone (B=–0.496; P<.001) had negative effects on the log count of negative emotions. In other words, the neutral sentiment and negative sentiment manifested in posts would both decrease audience's negative emotions. This finding may be due to the

XSL•FO
RenderX

backfire effects of the out-of-place intentional positive media tone amid the dire pandemic situation. Similar effects appear in the processes involved in the forced declaration of preferred pronouns, ensuring the persistence of public opinion by an authority's commands [97], and the perception of corporate reputation in response to promotional tone [98] in that the audience tends to lack trust in the sincerity of such commitment and negative reactions emerge subsequently [99].

As for the interaction effects of the media tone with the subject and the issue, we paid extra attention to the representative components, the Government of Hong Kong, and the pandemic situations. In the case of a positive tone as the reference group, the post with a negative tone and mentioning the Government of Hong Kong had a statistically significant positive impact on the negative emotions (B=0.896; P<.001). The interaction effect of neutral tone and the occurrence of the Government of Hong Kong was positive as well, but it was not significant (B=0.027; P=.90). When it comes to the issue, both the negative (B=0.950; P<.001) and neutral tone (B=0.391; P=.07) of the news post had a positive interaction effect with the mention of pandemic situations, respectively, on the negative user emotions, and still, the latter was not significant enough. We plotted the interaction diagram in Figures 7 and 8 to illustrate the interaction effects straightforwardly, from which we can tell that the negative tone of the post, if meeting the subject of the Government of Hong Kong or the issue of pandemic situations, would increase the user's negative emotions notably.

We also explored the relationship between the length of the post and its quadratic form with negative emotions separately. The length of the post had a significant negative effect on the negative emotions (B=–0.614; P=.003), while the squared term of length indicated a significant positive effect (B=0.651; P=.02). It suggested that the negative effect of post length on the count of negative emotions increases with higher values of length. Table 6 summarizes our hypothesis-testing results.

XSL•FO
RenderX

Table 5. Regression results[a].

Characteristics	*Dependent variable:* negative emotions			
	Model 1, *B* (SE)	Model 2, *B* (SE)	Model 3, *B* (SE)	Model 4, *B* (SE)
Neutral tone	−0.338[b] (0.084)	−0.339[b] (0.094)	−0.648[c] (0.198)	−0.662[c] (0.205)
Negative tone	−0.496[b] (0.103)	−0.791[b] (0.117)	−1.228[b] (0.227)	−1.397[b] (0.234)
Antipandemic measures	−0.050[c] (0.018)	−0.046[b] (0.018)	−0.049[c] (0.018)	−0.042[d] (0.018)
Pandemic situations	−0.261[b] (0.022)	−0.261[b] (0.021)	−0.659[c] (0.217)	−0.679[c] (0.217)
Supportive actions	−0.083[d] (0.035)	−0.093[c] (0.035)	−0.087[d] (0.035)	−0.099[c] (0.035)
Info	0.005 (0.018)	0.008 (0.018)	0.004 (0.018)	0.010 (0.018)
Central government	0.120[b] (0.033)	0.102[c] (0.033)	0.119[b] (0.033)	0.095[c] (0.033)
HKgov[e]	0.324[b] (0.019)	0.285 (0.209)	0.323[b] (0.019)	0.269 (0.209)
Sector	−0.058 (0.035)	−0.055 (0.035)	−0.057 (0.035)	−0.058 (0.035)
Length	−0.166[d] (0.081)	−0.206[d] (0.081)	−0.175[d] (0.081)	−0.614[c] (0.200)
Case	0.430[b] (0.041)	0.428[b] (0.040)	0.430[b] (0.041)	1.142[b] (0.153)
Length2	N/A[f]	N/A	N/A	0.651[d] (0.285)
Case2	N/A	N/A	N/A	−0.730[b] (0.152)
Neutral tone:HKgov	N/A	0.027 (0.209)	N/A	0.041 (0.209)
Negative tone:HKgov	N/A	0.896[b] (0.248)	N/A	0.815[c] (0.249)
Neutral tone:pandemic situations	N/A	N/A	0.391 (0.218)	0.415 (0.218)
Negative tone:pandemic situations	N/A	N/A	0.950[b] (0.255)	0.824[c] (0.256)
Constant	2.647[b] (0.094)	2.658[b] (0.103)	2.966[b] (0.201)	2.998[b] (0.208)
Theta	0.687[b] (0.006)	0.688[b] (0.006)	0.688[b] (0.006)	0.689[b] (0.006)

[a]n=23,705 observations were used in this regression analysis. Poster=yes for all models and month=yes for all models.

[b]$P<.001$.

[c]$P<.01$.

[d]$P<.05$.

[e]HKgov: Hong Kong Government.

[f]N/A: not applicable.

XSL•FO
RenderX

Figure 7. Interaction effect between media tone and occurrence of the Hong Kong (HK) Government.

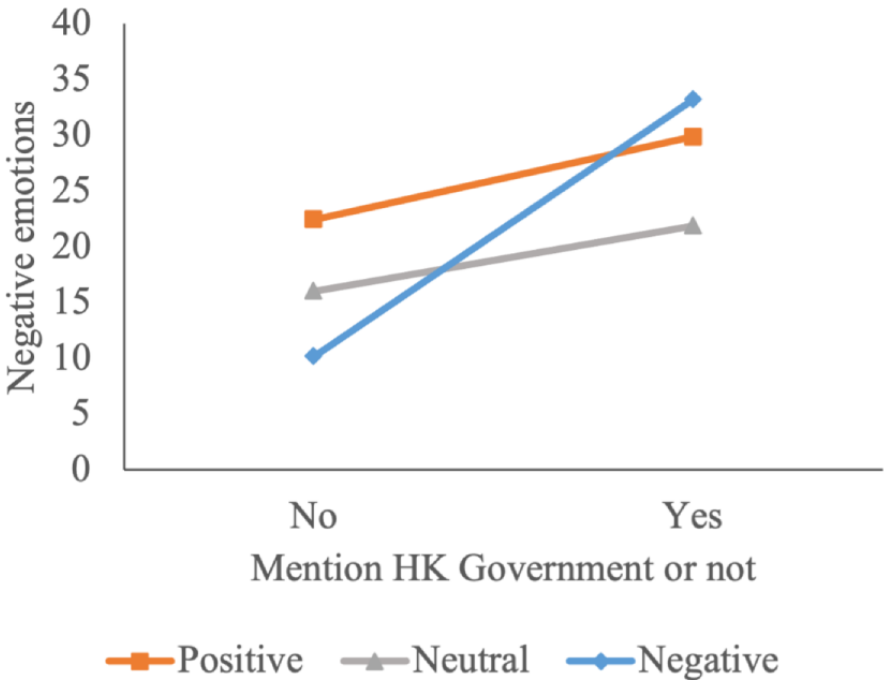

Figure 8. Interaction effect between media tone and occurrence of pandemic situations.

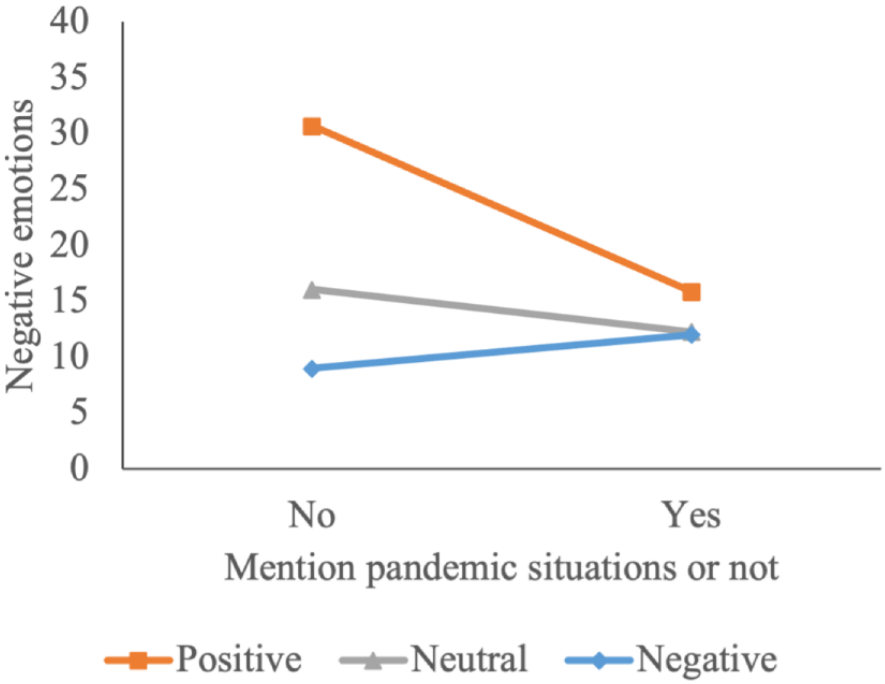

Table 6. Summary of hypothesis-testing results.

Hypothesis	Result
1	Unsupported
2	Supported
3a	Unsupported
3b	Supported
3c	Supported
4	Supported

XSL•FO
RenderX

Discussion

Principal Findings

This paper systematically presents the primary concern of Hong Kong citizens amid the fifth wave of the COVID-19 pandemic. We fill the research gap by not only depicting but also excavating the factors that might explain the elicitation of negative emotions from the perspective of coverage by news organizations on social media. Besides, we use Facebook, the most widely used social platform in Hong Kong, to shed light on the negativeness of the audience and thus provide references for stakeholders to grasp public opinion, improve residents' psychological well-being, and implement crisis communication more effectively under similar circumstances. We find that the occurrences of issues regarding pandemic situations, antipandemic measures, and supportive actions are likely to reduce the public's negative emotions, whereas comments on the posts mentioning the central government and the Government of Hong Kong reveal more negativeness. Besides, media tone can interact with the subjects and the issues in the news to affect users' negative emotions. Our findings about the prominent news elements are consistent with previous research studying emotional responses to news coverage by newspapers [15]. The relationship between post length and negative emotions is nonlinear. As the length of a post increases, there tends to be a decrease in the count of negative emotions experienced by individuals, but those more verbose posts will trigger more negative responses, echoing other research on article length and user engagement [55]. In the following sections, we discuss our theoretical contributions and practical implications, followed by reflections on limitations and future research.

Theoretical Contributions

Our findings about the key elements of news coverage situate well with the appraisal theory from which we developed the main research framework. To our best knowledge, our work is the first to examine the impacts of news coverage regarding the COVID-19 pandemic on user emotions published by official news organization accounts on social media, carving out a new space for research on the emotional influences of web news coverage. Among the issues and subjects identified from scratch, we recognize the components that may either arouse or relieve the users' negative emotions. Research regarding news coverage on the COVID-19 pandemic often places much emphasis on pandemic updates and public health efforts, which is reasonable because it is crucial to inform and educate the public under a crisis of great uncertainty [100]. Our study is consistent with previous research that mentions that issues regarding antipandemic measures could reduce negative emotional responses. Counterintuitively, the news posts updating pandemic situations are accompanied by a lower level of negative emotions. This finding may be because the public's negative emotions are more likely to be elicited by a high level of uncertainty [48,101,102] than an informative environment. People have the innate intention to know the progress of ongoing events and need reassurance that the situation is under proper settlement [103]. Timely updates, such as the current situation, investigation of the chain of transmission, and prediction about a future trend, may reduce uncertainty, which assists people in getting over the crisis psychologically [103]. Aside from the informative coverage on combating the pandemic, we argue that supportive actions also have a place in relieving the negative emotions of the news audience, although our data reveal that the media attention assigned to this issue is not that outstanding compared with other issues. When it comes to the subjects, we find that the occurrence of political crisis managers, that is, the central government of the Chinese mainland and the Government of Hong Kong, both increase the negative emotional responses, which is in accordance with prior research on how appearances of political actors in news affect user emotionality [15,87]. The government may inevitably become the target of public censure, but there still exists ways to improve community resilience for the rehabilitation from COVID-19. The negative relationship between the occurrence of functional sectors and negative emotions, although not significant, suggests the potential capability of the efforts of the refined and specialized functional sectors to be acknowledged by the general audience. This finding about supportive actions and functional sectors might support the establishment of community resilience on social media from the angles of organizational linkages, cooperation, and collective efficacy empowerment [104,105]. Apart from the main elements of news coverage, the text length manifests a curvilinear relationship with the negative emotional responses because of the possible cognitive overload. The message may be informative with the increased length at the beginning but add unnecessary load later, suggesting that the news length is not supposed to be merely a control variable as in previous research but an element to be considered in analyzing news coverage.

In addition, we contribute to expanding research on the priming effects of media tone. Our results concerning interaction effects demonstrate that the negative media tone, when co-occurring with the representative issue and subject, that is, the Government of Hong Kong and the pandemic situation, would be more likely to ignite negative emotions in user commenting. However, the main effect of the negative and neutral media tone would decrease the negative emotions compared with the positive media tone. The backfire effects might explain this finding because the backfiring may occur in the audience's understanding process involved in forced consensus [97,106], and the positive tone in a nonoptimistic crisis environment might induce unanticipated and undesirable consequences.

Practical Implications

Our work also provides practical implications from 3 perspectives. First, we suggest crisis communication strategies for the government to play a better role as a crisis manager to fulfill the public's expectations and thus reduce negative emotions and boost community resilience. A crisis might be a turning point for better or worse, in which shrewd crisis management plays a key role [107]. The public's disappointment at the government is likely to rise when the latter part's promises fail to be fulfilled, yet the flexibility and transparency in crisis communication are effective in dealing with uncertainty and a lack of trust, especially during the unexpected burst of pandemics [108]. Transparency, although likely to initially cause uneasiness, fear, and distrust, eventually would contribute

XSL•FO
RenderX

to regaining public trust and reduce redundant panic in the long term [109,110]. Meanwhile, it is noteworthy that functional sectors deserve much more attention. As community resilience increasingly enters the view of the public and administrative sectors, it is vital to engage local people in the entire process of mitigation, build organizational connections, and preserve social supports [104]. Emphasis on the existence of these specialized functional sectors might be a shot in the arm for the restless people.

Moreover, our study offers suggestions for news organizations on coverage in social media to improve the audience's psychological well-being and maintain a harmonious web-based public opinion environment. Our work suggests 2 criteria in view of the news coverage components that reduce negative emotions on social media. The first is extra immediacy, which is essential for mass communication amid the pandemic [111]. The issues of antipandemic measures and pandemic situations both require up-to-date messages and quick responses in terms of those constantly updating information [112]. The timely essence of the measures coping with the COVID-19 pandemic puts forth the same pace as the news coverage [113]. Another criterion is effectiveness, which ensures proper information management across the stages of a pandemic. Public health efforts to keep an outbreak under control and update on confirmed cases usually occupy people's attention during the initial stages of a pandemic. However, when the public has better knowledge as the pandemic unfolds, additional information is anticipated [100], in which case the previous pattern of coverage is no longer satisfactory. Issues such as supportive actions are able to convey collective efficacy, which signifies the postcrisis period and is beneficial to rebuilding community resilience [105]. News organizations should also be careful that an excessively detailed news post on social media is not necessarily the most effective way to be informative. Endeavors aligning with these findings may reduce the elicited user emotions when they rely on social media for news consumption and consequently facilitate a harmonious web environment [114].

Last but not least, through an examination of the emotional responses elicited by news coverage by news organizations on social media, the findings of this study provide insights into the possible change in attitudes toward health behavior, considering the role of emotions in attitudinal change [115]. Individuals may be accommodated to be more receptive to health-related information conveyed through news reports with strategies that avoid negative emotional stimulators.

Conclusions

Understanding the relationship between the negative emotions of commenters and news coverage on social media is of great significance not only for prompting rapid and efficient public health responses but also for shaping and maintaining individuals' values. By using machine learning techniques, this study sheds light on the impacts of various components of news coverage (issues, subjects, media tone, and length) on social media on the public's negative emotions (anger, fear, and sadness). Our comprehensive analysis culminates in results that will provide a reference framework for efficient crisis communication in the case of similar pandemics at present or in the future. This research, although discovering new findings by first extending the analysis between the components of news coverage and negative user emotions to the scenario of social media, echoes previous studies drawn from traditional media and its derivatives, such as web newspapers. Although the era of the COVID-19 pandemic gradually brings down the curtain, the commonality of this research and previous studies also contributes to establishing clearer territory in the field of health crises.

Besides the aforementioned theoretical and practical implications, this study still has several limitations. First, even though we applied state-of-the-art machine learning techniques to reveal emotion dynamics and measure the media tone, it is unavoidable that misclassifications still occur in text processing procedures. BERT-based models are among the top natural language processing models for emotion and sentiment analysis, but as with other machine learning models, they cannot achieve 100% accuracy, such as the difficulty in understanding informal language, which can be quite common in web-based communication and social media to express their emotions [116]. The misclassifications may be unavoidable in research involving automatic detection with large-scale data, and continuous improvement and customization are needed in future research [78]. Future research may apply methods to address this prevalent problem caused by using machine learning classification outcome as a variable in regression [117]. Besides, although we control the individual fixed effects of the poster as well as the time fixed effects, and most of the effects in our model are significant, our findings unravel the correlations of variables instead of causal relationships. We plan to explore more causality between news coverage and user emotions in future research. In addition, the research context of our study is limited to Hong Kong in the context of the fifth wave of the COVID-19 pandemic and Facebook due to the representativeness of the platform and the pandemic situation. Hence, we acknowledge that the findings are not necessarily generalizable to other countries, other platforms, or other waves of the COVID-19 pandemic in Hong Kong. Although we fill the research gap in revealing and explaining the emotion dynamics of the fifth wave of the COVID-19 pandemic through the lens of news coverage on social media, we welcome further evidence to prove its adaptability from diverse perspectives.

Acknowledgments

This study was supported in part by the National Natural Science Foundation of China under grant 72304041, in part by the Guangdong Basic and Applied Basic Research Foundation under grant 2023A1515011562, in part by the National Natural Science Foundation of China under grant 62202402, in part by the Guangdong Basic and Applied Basic Research Foundation under grant

XSL•FO
RenderX

2022A1515011583, in part by Beijing Normal University-Hong Kong Baptist University United International College Start-up Research Fund (reference UICR0700026-22), in part by Start-up research fund of Beijing Normal University-Hong Kong Baptist University United International College (UICR0700026-22), in part by the One-Off Tier 2 Start-Up grant (2020/2021) of Hong Kong Baptist University under grant RC-OFSGT2/20-21/COMM/002, in part by the Startup grant (Tier 1) for New Academics of Hong Kong Baptist University under grant AY2020/21, in part by the AI-Info Communication Study Scheme 2021/22 (reference AIS 21-22/06), and in part by the Digital Scholarship grant 2022/23 (first round; reference DSG/2223/108).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive details of materials and analysis.
[DOC File , 76 KB - jmir_v26i1e48491_app1.doc]

References

1. News at a glance: Hong Kong's high COVID-19 death rate, Biden's 'test and treat,' and a ResearchGate lawsuit. Science News. URL: https://www.science.org/content/article/news-glance-hong-kong-s-high-covid-19-death-rate-biden-s-test-and-treat-and [accessed 2022-09-05]
2. Latest situation of local situation of COVID-19 by March 11th 2022. Centre for Health Protection. 2022. URL: https://www.chp.gov.hk/tc/index.html [accessed 2022-09-05]
3. Taylor S. The psychology of pandemics. Annu Rev Clin Psychol 2022 May 09;18(1):581-609. [doi: 10.1146/annurev-clinpsy-072720-020131] [Medline: 34780260]
4. Choi EP, Hui BP, Wan EY. Depression and anxiety in Hong Kong during COVID-19. Int J Environ Res Public Health 2020 May 25;17(10):3740 [FREE Full text] [doi: 10.3390/ijerph17103740] [Medline: 32466251]
5. Salari N, Hosseinian-Far A, Jalali R, Vaisi-Raygani A, Rasoulpoor S, Mohammadi M, et al. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. Global Health 2020 Jul 06;16(1):57 [FREE Full text] [doi: 10.1186/s12992-020-00589-w] [Medline: 32631403]
6. Hutton M. COVID-19: panic buying leaves Hong Kong's refugees with little to eat, as fifth wave sends food prices soaring. Hong Kong Free Press. 2022. URL: https://hongkongfp.com/2022/03/04/covid-19-panic-buying-leaves-hong-kongs-refugees-with-little-to-eat-as-fifth-wave-sends-food-prices-soaring/ [accessed 2022-09-05]
7. Yeo R. Post-traumatic stress, anxiety, depression: how fifth wave of COVID-19 infections has affected Hongkongers' mental health. South China Morning Post. URL: https://www.scmp.com/news/hong-kong/society/article/3169733/post-traumatic-stress-anxiety-depression-how-fifth-wave [accessed 2022-09-05]
8. Porcelli P. Fear, anxiety and health-related consequences after the COVID-19 epidemic. Clin Neuropsychiatry 2020 Apr;17(2):103-111 [FREE Full text] [doi: 10.36131/CN20200215] [Medline: 34908979]
9. Baumer EP, Martin KH. Social organization, collective sentiment, and legal sanctions in murder cases. Am J Sociol 2013 Jul;119(1):131-182. [doi: 10.1086/672082]
10. Steinert S. Corona and value change. The role of social media and emotional contagion. Ethics Inf Technol 2021 Jul 21;23(Suppl 1):59-68 [FREE Full text] [doi: 10.1007/s10676-020-09545-z] [Medline: 32837288]
11. Davidson AE, Wallack L. A content analysis of sexually transmitted diseases in the print news media. J Health Commun 2004 Mar;9(2):111-117. [doi: 10.1080/10810730490425268] [Medline: 15204822]
12. Reynolds R, W Seeger M. Crisis and emergency risk communication as an integrative model. J Health Commun 2005 Feb 23;10(1):43-55. [doi: 10.1080/10810730590904571] [Medline: 15764443]
13. Westerman D, Spence PR, van der Heide B. Social media as information source: recency of updates and credibility of information. J Comput Mediat Comm 2013 Nov 08;19(2):171-183. [doi: 10.1111/jcc4.12041]
14. Jones R, Mougouei D, Evans SL. Understanding the emotional response to COVID-19 information in news and social media: a mental health perspective. Hum Behav Emerg Technol 2021 Dec 28;3(5):832-842 [FREE Full text] [doi: 10.1002/hbe2.304] [Medline: 34901769]
15. Eisele O, Litvyak O, Brändle VK, Balluff P, Fischeneder A, Sotirakou C, et al. An emotional rally: exploring commenters' responses to online news coverage of the COVID-19 crisis in Austria. Digit Journal 2021 Dec 01;10(6):952-975. [doi: 10.1080/21670811.2021.2004552]
16. Hong Kong digital marketing 2020. AsiaPac Net Media Limited. 2020. URL: https://www.asiapacdigital.com/digital-marketing-insight/hk-digital-marketing-2020 [accessed 2022-09-05]

XSL•FO
RenderX

17. Roseman IJ, Smith CA. Appraisal theory. In: Scherer KR, Schorr A, Johnstone T, editors. Appraisal Processes in Emotion: Theories, Methods, Research. Oxford, UK: Oxford University Press; 2001:3-19.
18. Scherer KR. Appraisal processes in emotion. In: Davidson RJ, Sherer KR, Goldsmith HH, editors. Handbook of Affective Sciences. Oxfordshire, UK: Oxford University Press; 2003:572-595.
19. Scherer KR. Appraisal theory. In: Dalgleish T, Power MJ, editors. Handbook of Cognition and Emotion. Hoboken, NJ: John Wiley & Sons; 1999.
20. Korenek P, Šimko M. Sentiment analysis on microblog utilizing appraisal theory. World Wide Web 2013 Aug 13;17(4):847-867. [doi: 10.1007/s11280-013-0247-z]
21. Soo'Guan KC, Nourbakhsh A, Na J. Sentiment analysis of online news text: a case study of appraisal theory. Online Inf Rev 2012;36(6):858-878. [doi: 10.1108/14684521211287936]
22. Pahayahay A, Khalili-Mahani N. What media helps, what media hurts: a mixed methods survey study of coping with COVID-19 using the media repertoire framework and the appraisal theory of stress. J Med Internet Res 2020 Aug 06;22(8):e20186 [FREE Full text] [doi: 10.2196/20186] [Medline: 32701459]
23. Robertson CE, Pröllochs N, Schwarzenegger K, Pärnamets P, van Bavel JJ, Feuerriegel S. Negativity drives online news consumption. Nat Hum Behav 2023 May 16;7(5):812-822 [FREE Full text] [doi: 10.1038/s41562-023-01538-4] [Medline: 36928780]
24. Kim HJ, Cameron GT. Emotions matter in crisis: the role of anger and sadness in the publics' response to crisis news framing and corporate crisis response. Commun Res 2011 Jan 10;38(6):826-855. [doi: 10.1177/0093650210385813]
25. Ksiazek TB. Commenting on the news. Journal Stud 2016 Aug 02;19(5):650-673. [doi: 10.1080/1461670x.2016.1209977]
26. Barnes R. Understanding the affective investment produced through commenting on Australian alternative journalism website New Matilda. New Media Soc 2013 Nov 13;17(5):810-826. [doi: 10.1177/1461444813511039]
27. Canter L. The misconception of online comment threads. Journal Pract 2013 Oct;7(5):604-619. [doi: 10.1080/17512786.2012.740172]
28. Coe K, Kenski K, Rains SA. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. J Commun 2014 Jun 16;64(4):658-679. [doi: 10.1111/jcom.12104]
29. Salminen J, Sengün S, Corporan J, Jung S, Jansen BJ. Topic-driven toxicity: exploring the relationship between online toxicity and news topics. PLoS One 2020 Feb 21;15(2):e0228723 [FREE Full text] [doi: 10.1371/journal.pone.0228723] [Medline: 32084164]
30. Suran M, Holton AE, Coleman R. Topical punch: health topics as drivers of idiosyncratic reader responses to online news. Journal Mass Commun Q 2014 Sep 18;91(4):725-739. [doi: 10.1177/1077699014550093]
31. Cho J, Boyle MP, Keum H, Shevy MD, McLeod DM, Shah DV, et al. Media, terrorism, and emotionality: emotional differences in media content and public reactions to the September 11th terrorist attacks. J Broadcast Electron Media 2003 Sep;47(3):309-327. [doi: 10.1207/s15506878jobem4703_1]
32. Yu S, He S, Cai Z, Lee I, Naseriparsa M, Xia F. Exploring public sentiment during COVID-19: a cross country analysis. IEEE Trans Comput Soc Syst 2023 Jun;10(3):1083-1094. [doi: 10.1109/tcss.2022.3146633]
33. Sesagiri Raamkumar A, Tan SG, Wee HL. Measuring the outreach efforts of public health authorities and the public response on Facebook during the COVID-19 pandemic in early 2020: cross-country comparison. J Med Internet Res 2020 May 19;22(5):e19334 [FREE Full text] [doi: 10.2196/19334] [Medline: 32401219]
34. Liang G, Zhao J, Lau HY, Leung CW. Using social media to analyze public concerns and policy responses to COVID-19 in Hong Kong. ACM Trans Manage Inf Syst 2021 Sep 08;12(4):1-20. [doi: 10.1145/3460124]
35. Jin Y, Pang A, Cameron GT. Toward a publics-driven, emotion-based conceptualization in crisis communication: unearthing dominant emotions in multi-staged testing of the integrated crisis mapping (ICM) model. J Public Relat Res 2012 Jun;24(3):266-298. [doi: 10.1080/1062726x.2012.676747]
36. Lin J, Carter L, Liu D. Privacy concerns and digital government: exploring citizen willingness to adopt the COVIDSafe app. Eur J Inf Syst 2021 May 05;30(4):389-402. [doi: 10.1080/0960085x.2021.1920857]
37. Chon MG, Park H. Predicting public support for government actions in a public health crisis: testing fear, organization-public relationship, and behavioral intention in the framework of the situational theory of problem solving. Health Commun 2021 Apr 09;36(4):476-486. [doi: 10.1080/10410236.2019.1700439] [Medline: 31818134]
38. Fiske ST, Taylor SE. Social Cognition. 2nd edition. New York, NY: McGraw-Hill Humanities; 1991.
39. McCombs ME. The agenda-setting role of the mass media in the shaping of public opinion. University of Texas at Austin. 2002. URL: https://sticerd.lse.ac.uk/dps/extra/McCombs.pdf [accessed 2024-04-29]
40. Sheafer T. How to evaluate it: the role of story-evaluative tone in agenda setting and priming. J Commune 2007;57(1):21-39. [doi: 10.1111/j.0021-9916.2007.00327.x]
41. Willnat L, Zhu J. Newspaper coverage and public opinion in Hong Kong: a time‐series analysis of media priming. Pol Comm 1996 Apr;13(2):231-246. [doi: 10.1080/10584609.1996.9963109]
42. Blood DJ, Phillips PC. Economic headline news on the agenda: new approaches to understanding causes and effects. In: Blood DJ, Phillips PC, editors. Communication and Democracy. New York, NY: Routledge; 1997.
43. Soroka S, Daku M, Hiaeshutter-Rice D, Guggenheim L, Pasek J. Negativity and positivity biases in economic news coverage: traditional versus social media. Commun Res 2017 Aug 21;45(7):1078-1098. [doi: 10.1177/0093650217725870]

XSL•FO
RenderX

44. Bolls PD, Lang A, Potter RF. The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. Commun Res 2016 Jun 29;28(5):627-651. [doi: 10.1177/009365001028005003]
45. Zavattaro SM, French PE, Mohanty SD. A sentiment analysis of U.S. local government tweets: the connection between tone and citizen involvement. Gov Inf Q 2015 Jul;32(3):333-341. [doi: 10.1016/j.giq.2015.03.003]
46. Xu WW, Zhang C. Sentiment, richness, authority, and relevance model of information sharing during social crises—the case of #MH370 tweets. Comput Human Behav 2018 Dec;89:199-206 [FREE Full text] [doi: 10.1016/j.chb.2018.07.041]
47. Soroka S, McAdams S. News, politics, and negativity. Polit Commun 2015 Feb 03;32(1):1-22. [doi: 10.1080/10584609.2014.881942]
48. Ziegele M, Breiner T, Quiring O. What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. J Commun 2014 Oct 16;64(6):1111-1138. [doi: 10.1111/jcom.12123]
49. Wu Y, Shen F. Negativity makes us polarized: a longitudinal study of media tone and opinion polarization in Hong Kong. Asian J Commun 2020 Jun 23;30(3-4):199-220. [doi: 10.1080/01292986.2020.1784968]
50. Doyle E, Lee Y. Context, context, context: priming theory and attitudes towards corporations in social media. Public Relat Rev 2016 Dec;42(5):913-919. [doi: 10.1016/j.pubrev.2016.09.005]
51. Hooghe M, Okolikj M. The long-term effects of the economic crisis on political trust in Europe: is there a negativity bias in the relation between economic performance and political support? Comp Eur Polit 2020 May 27;18(6):879-898. [doi: 10.1057/s41295-020-00214-5]
52. Schreiner M, Fischer T, Riedl R. Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda. Electron Commer Res 2019 May 10;21(2):329-345. [doi: 10.1007/S10660-019-09353-8]
53. Schultz CD. Driving likes, comments, and shares on social networking sites: how post characteristics affect brand interactions in apparel retailing. In: Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World. 2016 Presented at: ICEC '16; August 17-19, 2016; Suwon, Republic of Korea p. 1-9 URL: https://dl.acm.org/doi/10.1145/2971603.2971612 [doi: 10.1145/2971603.2971612]
54. de Vries L, Gensler S, Leeflang PS. Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. J Interact Mark 2012 May;26(2):83-91. [doi: 10.1016/j.intmar.2012.01.003]
55. Gupta RK, Kaushik K. Linking text characteristics of ideas to their popularity in online user innovation communities. Comput Human Behav 2022 Nov;136:107382. [doi: 10.1016/j.chb.2022.107382]
56. Jones Q, Ravid G, Rafaeli S. Information overload and the message dynamics of online interaction spaces: a theoretical model and empirical exploration. Inf Syst Res 2004 Jun;15(2):194-210. [doi: 10.1287/isre.1040.0023]
57. Kirsh D. A few thoughts on cognitive overload. Intellectica 2000;30(1):19-51. [doi: 10.3406/intel.2000.1592]
58. Burmeister A, Alterman V, Fasbender U, Wang M. Too much to know? The cognitive demands of daily knowledge seeking and the buffering role of coworker contact quality. J Appl Psychol 2022 Aug;107(8):1303-1322. [doi: 10.1037/apl0000925] [Medline: 34014685]
59. Sweller J. Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ Psychol Rev 2010 Apr 23;22(2):123-138. [doi: 10.1007/s10648-010-9128-5]
60. van Merriënboer JJ, Sweller J. Cognitive load theory and complex learning: recent developments and future directions. Educ Psychol Rev 2005 Jun;17(2):147-177 [FREE Full text] [doi: 10.1007/s10648-005-3951-0]
61. Wagner T, Baccarella C, Voigt KI. Antecedents of brand post popularity in Facebook: the influence of images, videos, and text. In: Proceedings of 15th International Marketing Trends Conference. 2016 Presented at: IMTC '16; January 21-23, 2016; Venice, Italy URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2697458
62. Li X, Chan M. Smartphone uses and emotional and psychological well-being in China: the attenuating role of perceived information overload. Behav Inform Technol 2021 May 20;41(11):2427-2437. [doi: 10.1080/0144929x.2021.1929489]
63. Tzafilkou K, Perifanou M, Economides AA. Negative emotions, cognitive load, acceptance, and self-perceived learning outcome in emergency remote education during COVID-19. Educ Inf Technol (Dordr) 2021 Jun 12;26(6):7497-7521 [FREE Full text] [doi: 10.1007/s10639-021-10604-1] [Medline: 34149299]
64. Bernstein J. TLDR: so just how short should your online article be? The Guardian. URL: https://www.theguardian.com/media-network/2015/jul/15/tldr-quartz-associated-press-article-length [accessed 2023-12-19]
65. Masini A, van Aelst P, Zerback T, Reinemann C, Mancini P, Mazzoni M, et al. Measuring and explaining the diversity of voices and viewpoints in the news. Journalism Stud 2017 Jul 13;19(15):2324-2343. [doi: 10.1080/1461670x.2017.1343650]
66. Empowering Organizations to Scale Globally. URL: https://oosga.com/social-media/hkg/ [accessed 2023-12-19]
67. Sun F, Liu J, Wu J, Pei C, Lin X, Ou W, et al. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019 Presented at: CIKM '19; November 3-7, 2019; Beijing, China p. 1441-1450 URL: https://dl.acm.org/doi/10.1145/3357384.3357895 [doi: 10.1145/3357384.3357895]
68. Vaswani A, Shazeer N, Parmar N. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 1-11 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

XSL•FO
RenderX

69. Nissen IA, Walter JG, Charquero-Ballester M, Bechmann A. Digital infrastructures of COVID-19 misinformation: a new conceptual and analytical perspective on fact-checking. Digit Journal 2022 Feb 08;10(5):738-760. [doi: 10.1080/21670811.2022.2026795]
70. Moving along the path to normalcy. Government of Hong Kong. URL: https://www.news.gov.hk/chi/2022/05/20220530/20220530_101943_247.html [accessed 2022-09-05]
71. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. PLoS One 2020 May 1;15(5):e0232525 [FREE Full text] [doi: 10.1371/journal.pone.0232525] [Medline: 32357164]
72. Bao Y, Quan C, Wang L, Ren F. The role of pre-processing in Twitter sentiment analysis. In: Proceedings of the 10th International Conference on Intelligent Computing Methodologies. 2014 Presented at: ICIC '14; August 3-6, 2014; Taiyuan, China p. 615-624 URL: https://link.springer.com/chapter/10.1007/978-3-319-09339-0_62 [doi: 10.1007/978-3-319-09339-0_62]
73. D'Andrea E, Ducange P, Bechini A, Renda A, Marcelloni F. Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Syst Appl 2019 Feb;116:209-226. [doi: 10.1016/j.eswa.2018.09.009]
74. Hair Zaki UH, Ibrahim R, Abd HS, Kamsani II. Text detergent: the systematic combination of text pre-processing techniques for social media sentiment analysis. In: Saeed F, Mohammed F, Ghaleb F, editors. Advances on Intelligent Informatics and Computing: Health Informatics, Intelligent Systems, Data Science and Smart Computing. Cham, Switzerland: Springer; 2022:50-61.
75. BYVoid: conversion between traditional and simplified Chinese. GitHub. URL: https://github.com/BYVoid/OpenCC [accessed 2024-04-05]
76. BrownSweater/BERT_SMP2020-EWECT. GitHub. URL: https://github.com/BrownSweater/BERT_SMP2020-EWECT [accessed 2020-09-05]
77. SMP2020 Weibo emotion classification evaluationSMP2020 -EWECT. GitHub. URL: https://smp2020ewect.github.io/ [accessed 2024-04-05]
78. Wang X, Chao F, Yu G, Zhang K. Factors influencing fake news rebuttal acceptance during the COVID-19 pandemic and the moderating effect of cognitive ability. Comput Human Behav 2022 May;130:107174 [FREE Full text] [doi: 10.1016/j.chb.2021.107174] [Medline: 35002055]
79. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2024-05-15]
80. Zhong Q, Ding L, Liu J, Du B, Tao D. Can ChatGPT understand too? a comparative study on ChatGPT and fine-tuned BERT. arXiv Preprint posted online on February 19, 2023.
81. Glaser BG. Conceptualization: on theory and theorizing using grounded theory. Int J Qual Methods 2016 Nov 30;1(2):23-38. [doi: 10.1177/160940690200100203]
82. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022. [doi: 10.7551/mitpress/1120.003.0082]
83. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2018 Nov 28;78(11):15169-15211. [doi: 10.1007/s11042-018-6894-4]
84. Xie R, Chu SK, Chiu DK, Wang Y. Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis. Data Inf Manag 2021 Jan 01;5(1):86-99 [FREE Full text] [doi: 10.2478/dim-2020-0023] [Medline: 35402850]
85. News of government of Hong Kong. Government of Hong Kong. 2022. URL: https://www.news.gov.hk/ [accessed 2022-09-05]
86. Sentiment detection of Weibo users during COVID-19 Data set. Government of Beijing. 2020. URL: https://www.datafountain.cn/competitions/423/datasets [accessed 2022-09-05]
87. Basile V, Cauteruccio F, Terracina G. How dramatic events can affect emotionality in social posting: the impact of COVID-19 on Reddit. Future Internet 2021 Jan 27;13(2):29. [doi: 10.3390/fi13020029]
88. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534. [doi: 10.1016/s1473-3099(20)30120-1]
89. Claussen J, Kretschmer T, Mayrhofer P. The Effects of Rewarding User Engagement: The Case of Facebook Apps. Inform Syst Res 2013 Mar;24(1):186-200 [FREE Full text] [doi: 10.1287/isre.1120.0467]
90. Yoon G, Li C, Ji YG, North M, Hong C, Liu J. Attracting comments: digital engagement metrics on Facebook and financial performance. J Advert 2018 Jan 24;47(1):24-37. [doi: 10.1080/00913367.2017.1405753]
91. Lee HM, van Dolen W. Creative participation: collective sentiment in online co-creation communities. Inf Manag 2015 Dec;52(8):951-964. [doi: 10.1016/j.im.2015.07.002]
92. Guo M, Sun FS. Like, comment, or share? Exploring the effects of local television news Facebook posts on news engagement. J Broadcast Electron Media 2020 Dec 10;64(5):736-755. [doi: 10.1080/08838151.2020.1851125]
93. Heiss R, Schmuck D, Matthes J. What drives interaction in political actors' Facebook posts' profile and content predictors of user engagement and political actors' reactions. Inf Commun Soc 2018 Mar 08;22(10):1497-1513. [doi: 10.1080/1369118x.2018.1445273]

XSL•FO
RenderX

94. Kite J, Foley BC, Grunseit AC, Freeman B. Please like me: Facebook and public health communication. PLoS One 2016;11(9):e0162765 [FREE Full text] [doi: 10.1371/journal.pone.0162765] [Medline: 27632172]
95. Mansfield ER, Helms BP. Detecting multicollinearity. Am Stat 1982 Aug;36(3a):158-160. [doi: 10.1080/00031305.1982.10482818]
96. Venables WN, Ripley BD. Modern Applied Statistics with S-PLUS. New York, NY: Springer; 2013.
97. Conway III LG, Schaller M. When authorities' commands backfire: attributions about consensus and effects on deviant decision making. J Pers Soc Psychol 2005 Sep;89(3):311-326. [doi: 10.1037/0022-3514.89.3.311] [Medline: 16248716]
98. Dauphinais MM. CSR communication strategies: how to steer away from consumer skepticism. University of New Hampshire. 2020. URL: https://scholars.unh.edu/cgi/viewcontent.cgi?article=1534&context=honors [accessed 2024-04-05]
99. Kim S. The process model of corporate social responsibility (CSR) communication: CSR communication and its relationship with consumers' CSR knowledge, trust, and corporate reputation perception. J Bus Ethics 2017 Jan 17;154(4):1143-1159. [doi: 10.1007/s10551-017-3433-6]
100. Chipidza W, Akbaripourdibazar E, Gwanzura T, Gatto NM. Topic analysis of traditional and social media news coverage of the early COVID-19 pandemic and implications for public health communication. Disaster Med Public Health Prep 2022 Oct;16(5):1881-1888 [FREE Full text] [doi: 10.1017/dmp.2021.65] [Medline: 33653437]
101. Berkowitz D. Non-routine news and newswork: exploring a what-a-story. J Commun 1992;42(1):82-94. [doi: 10.1111/j.1460-2466.1992.tb00770.x]
102. Taha S, Matheson K, Cronin T, Anisman H. Intolerance of uncertainty, appraisals, coping, and anxiety: the case of the 2009 H1N1 pandemic. Br J Health Psychol 2014 Sep 09;19(3):592-605. [doi: 10.1111/bjhp.12058] [Medline: 23834735]
103. Holladay SJ. Are they practicing what we are preaching? An investigation of crisis communication strategies in the media coverage of chemical accidents. In: The Handbook of Crisis Communication. New York, NY: Wiley-Blackwell; 2010:159-180.
104. Norris FH, Stevens SP, Pfefferbaum B, Wyche KF, Pfefferbaum RL. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. Am J Community Psychol 2008 Mar 22;41(1-2):127-150. [doi: 10.1007/s10464-007-9156-6] [Medline: 18157631]
105. Xie L, Pinto J, Zhong B. Building community resilience on social media to help recover from the COVID-19 pandemic. Comput Human Behav 2022 Sep;134:107294 [FREE Full text] [doi: 10.1016/j.chb.2022.107294] [Medline: 35431426]
106. Conway LG, Salcido A, Gornick LJ, Bongard KA, Moran MA, Burfiend C. When self-censorship norms backfire: the manufacturing of positive communication and its ironic consequences for the perceptions of groups. Basic Appl Soc Psych 2009 Nov 18;31(4):335-347. [doi: 10.1080/01973530903317169]
107. Sanjeev MA, Pande N, P K SK. Role of effective crisis communication by the government in managing the first wave COVID-19 pandemic - a study of Kerala government's success. J Public Aff 2021 Nov;21(4):e2721 [FREE Full text] [doi: 10.1002/pa.2721] [Medline: 34512184]
108. Moon MJ. Fighting COVID-19 with agility, transparency, and participation: wicked policy problems and new governance challenges. Public Adm Rev 2020 May 20;80(4):651-656 [FREE Full text] [doi: 10.1111/puar.13214] [Medline: 32836434]
109. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
110. Ford N, Thomas R, Grove J. Transparency: a central principle underpinning trustworthy guidelines. J Clin Epidemiol 2022 Feb;142:246-248 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.11.025] [Medline: 34798288]
111. Navarro ER, Guerra-Ayala MJ, Casimiro-Urcos WH, Vélez-Jiménez D, Casimiro-Urcos NC, Salazar-Montoya EO, et al. Relevance of the mass media in prevention, education and contextual management of COVID-19. World J Educ Technol Curr Issues 2021 Jan 30;13(1):129-146. [doi: 10.18844/wjet.v13i1.5423]
112. Subbian V, Solomonides A, Clarkson M, Rahimzadeh VN, Petersen C, Schreiber R, et al. Ethics and informatics in the age of COVID-19: challenges and recommendations for public health organization and public policy. J Am Med Inform Assoc 2021 Jan 15;28(1):184-189 [FREE Full text] [doi: 10.1093/jamia/ocaa188] [Medline: 32722749]
113. Jindal C, Kumar S, Sharma S, Choi YM, Efird JT. The prevention and management of COVID-19: seeking a practical and timely solution. Int J Environ Res Public Health 2020 Jun 04;17(11):3986 [FREE Full text] [doi: 10.3390/ijerph17113986] [Medline: 32512826]
114. Lu Y, Gall Myrick J. Cross-cutting exposure on Facebook and political participation: unraveling the effects of emotional responses and online incivility. J Media Psychol 2016 Jul;28(3):100-110. [doi: 10.1027/1864-1105/a000203]
115. Petty RE, Briñol P. Emotion and persuasion: cognitive and meta-cognitive processes impact attitudes. Cogn Emot 2015 Oct 10;29(1):1-26. [doi: 10.1080/02699931.2014.967183] [Medline: 25302943]
116. Talaat AS. Sentiment analysis classification system using hybrid BERT models. J Big Data 2023 Jun 23;10(1):110. [doi: 10.1186/s40537-023-00781-w]
117. Zhang H. How using machine learning classification as a variable in regression leads to attenuation bias and what to do about it. SocArXiv 2021 May 28 [FREE Full text] [doi: 10.31235/osf.io/453jk]

XSL•FO
RenderX

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
RQ: research question

Edited by A Mavragani; submitted 26.04.23; peer-reviewed by X Zhou, M Gasmi , A Chang; comments to author 11.10.23; revised version received 28.12.23; accepted 27.02.24; published 06.06.24.

Please cite as:
Wang H, Li Y, Ning X
News Coverage of the COVID-19 Pandemic on Social Media and the Public's Negative Emotions: Computational Study
J Med Internet Res 2024;26:e48491
URL: https://www.jmir.org/2024/1/e48491
doi:10.2196/48491
PMID:38843521

©Hanjing Wang, Yupeng Li, Xuan Ning. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Monitoring Adverse Drug Events in Web Forums: Evaluation of a Pipeline and Use Case Study

Pierre Karapetiantz[1], PhD; Bissan Audeh[1], PhD; Akram Redjdal[1], PhD; Théophile Tiffet[2,3], MD; Cédric Bousquet[1,2], PhD, PharmD; Marie-Christine Jaulent[1], PhD

[1]Inserm, Sorbonne Université, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006, Paris, France

[2]Service de santé publique et information médicale, CHU de Saint Etienne, 42000 Saint-Etienne, France

[3]Institut National de la Santé et de la Recherche Médicale, Université Jean Monnet, SAnté INgéniérie BIOlogie St-Etienne, SAINBIOSE, 42270 Saint-Priest-en-Jarez, France

Corresponding Author:
Marie-Christine Jaulent, PhD
Inserm
Sorbonne Université
université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006
15 rue de l'école de Médecine
Paris, 75006
France
Phone: 33 144279108
Email: marie-christine.jaulent@inserm.fr

Abstract

Background: To mitigate safety concerns, regulatory agencies must make informed decisions regarding drug usage and adverse drug events (ADEs). The primary pharmacovigilance data stem from spontaneous reports by health care professionals. However, underreporting poses a notable challenge within the current system. Explorations into alternative sources, including electronic patient records and social media, have been undertaken. Nevertheless, social media's potential remains largely untapped in real-world scenarios.

Objective: The challenge faced by regulatory agencies in using social media is primarily attributed to the absence of suitable tools to support decision makers. An effective tool should enable access to information via a graphical user interface, presenting data in a user-friendly manner rather than in their raw form. This interface should offer various visualization options, empowering users to choose representations that best convey the data and facilitate informed decision-making. Thus, this study aims to assess the potential of integrating social media into pharmacovigilance and enhancing decision-making with this novel data source. To achieve this, our objective was to develop and assess a pipeline that processes data from the extraction of web forum posts to the generation of indicators and alerts within a visual and interactive environment. The goal was to create a user-friendly tool that enables regulatory authorities to make better-informed decisions effectively.

Methods: To enhance pharmacovigilance efforts, we have devised a pipeline comprising 4 distinct modules, each independently editable, aimed at efficiently analyzing health-related French web forums. These modules were (1) web forums' posts extraction, (2) web forums' posts annotation, (3) statistics and signal detection algorithm, and (4) a graphical user interface (GUI). We showcase the efficacy of the GUI through an illustrative case study involving the introduction of the new formula of Levothyrox in France. This event led to a surge in reports to the French regulatory authority.

Results: Between January 1, 2017, and February 28, 2021, a total of 2,081,296 posts were extracted from 23 French web forums. These posts contained 437,192 normalized drug-ADE couples, annotated with the Anatomical Therapeutic Chemical (ATC) Classification and Medical Dictionary for Regulatory Activities (MedDRA). The analysis of the Levothyrox new formula revealed a notable pattern. In August 2017, there was a sharp increase in posts related to this medication on social media platforms, which coincided with a substantial uptick in reports submitted by patients to the national regulatory authority during the same period.

Conclusions: We demonstrated that conducting quantitative analysis using the GUI is straightforward and requires no coding. The results aligned with prior research and also offered potential insights into drug-related matters. Our hypothesis received partial confirmation because the final users were not involved in the evaluation process. Further studies, concentrating on ergonomics and the impact on professionals within regulatory agencies, are imperative for future research endeavors. We

XSL•FO
RenderX

emphasized the versatility of our approach and the seamless interoperability between different modules over the performance of individual modules. Specifically, the annotation module was integrated early in the development process and could undergo substantial enhancement by leveraging contemporary techniques rooted in the Transformers architecture. Our pipeline holds potential applications in health surveillance by regulatory agencies or pharmaceutical companies, aiding in the identification of safety concerns. Moreover, it could be used by research teams for retrospective analysis of events.

(J Med Internet Res 2024;26:e46176) doi:10.2196/46176

KEYWORDS

pharmacovigilance; social media; scraper; natural language processing; signal detection; graphical user interface

Introduction

Social Media as a Complementary Data Source for Pharmacovigilance

One primary mission of regulatory agencies such as the FDA (Food and Drug Administration) or the EMA (European Medicines Agency) is to monitor drug usage and adverse drug events (ADEs) to mitigate the risks associated with drugs within the population. This task entails analyzing diverse data sources, including clinical trials, postmarketing surveillance, spontaneous reporting systems, and published scientific literature. Despite the wealth of available data, some ADEs are not always detected promptly, largely because of underreporting. In France, for instance, underreporting was estimated to range between 78% and 99% from 1997 to 2002 [1]. To tackle this challenge, several countries have implemented systems allowing patients to report ADEs.

Additional sources for detecting ADEs have been under exploration, such as electronic patient records [2-4] and social media platforms [5-9]. While some argue that social media alone cannot serve as a primary source for signal detection [10], it can be viewed as a valuable secondary source for monitoring emerging adverse drug reactions or reinforcing signals previously identified through spontaneous reports stored in traditional pharmacovigilance databases [11]. In a prior study by the authors, patient profiles and reported ADEs found in web forums were compared with those in the French Pharmacovigilance Database (FPVD). The forums tended to represent younger patients, more women, less severe cases, and a higher incidence of psychiatric disorder–related ADEs compared with the FPVD [12]. Moreover, forums reported a greater number of unexpected ADEs. Over the past decade, several tools for evaluating social media posts have been described in the literature [13]. Specifically, effective ADE detection in social media necessitates both quantitative and qualitative analyses of data [14].

Qualitative Approach for Individual Assessment of Posts

Qualitative assessment entails evaluating whether users' messages contain pertinent information for an assessment akin to a pharmacovigilance case report. This includes details such as the patient's age and gender, the severity of the case, the expectedness and timeline of the adverse event, time-to-onset, dechallenge (outcome upon drug withdrawal), and rechallenge (outcome upon drug reintroduction). For instance, GlaxoSmithKline Inc. implemented the qualitative approach Insight Explorer, which facilitates the collection of extensive data for causality and quality assessment. Users can input data including personal information (eg, age range, gender) and product details (eg, name, route of administration, duration of use, dosage). This approach was adapted for the WEB-RADR (Recognizing Adverse Drug Reactions) project to manually construct a gold standard of curated patient-authored text [15].

Quantitative Approach for Monitoring Adverse Drug Events on Social Media

Quantitative evaluation involves analyzing extracted data using descriptive and analytical statistics, such as signal detection and change-point analysis. Numerous projects have been undertaken to monitor ADEs on social media. One of the earliest projects is the PREDOSE (Prescription Drug Abuse Online Surveillance and Epidemiology) project [5], which investigates the illicit use of pharmaceutical opioids reported in web forums. While the PREDOSE project showcased the potential of leveraging social media for opioid monitoring, notable limitations are the lack of deidentification and signal detection methods. MedWatcher Social, a monitoring platform for health-related web forums, Twitter, and Facebook, represents a prototype application developed in 2014 [16]. Yeleswarapu et al [6] outlined a semiautomatic pipeline that applies natural language processing (NLP) tasks to extract ADEs from MEDLINE abstracts and user comments from health-related websites. However, this pipeline was not intended for routine use.

The Domino's interface [17], developed in 2018 by the University of Bordeaux in France and funded by the French Medicines Agency (Agence nationale de sécurité du médicament et des produits de santé [ANSM]), was designed to analyze drug misuses in health-related web forums using NLP methods and the summary of product characteristics. Initially tailored for antidepressant drugs, this tool does not primarily focus on ADE surveillance.

Another pipeline, described by Nikfarjam et al in 2019 [7], used a neural network–based named entity recognition system specifically designed for user-generated content in social media. This platform is dedicated to identifying the association of cutaneous ADEs with cancer therapy drugs. The study focused on a selection of drugs and only examined 8 ADEs.

Magge et al [8] described a pipeline aimed at the extraction and normalization of adverse drug mentions on Twitter. Their pipeline consisted of an ADE classifier designed to identify tweets mentioning an ADE, which were then mapped to a MedDRA (Medical Dictionary for Regulatory Activities Terminology) code. However, the normalization process was

XSL•FO
RenderX

confined to the ADEs present in the training set. Neither Nikfarjam's nor Magge's pipeline provides a graphical user interface.

Some private companies also offer tools for analyzing social media for pharmacovigilance purposes. For instance, the DETECT platform was developed as part of a collaborative project in France by Kappa Santé [18]. This system enabled the labeling of posts with known controlled vocabulary concepts, and signal detection was conducted [19]. Within the scope of this project, Expert System Company implemented BIOPHARMA Navigator to extract web forum posts, while the Luxid Annotation Server provided web services for the automatic annotation of posts.

An important finding from the studies of the last decade is that while regulatory agencies have begun using data sources beyond spontaneous reports, social media has yet to be fully leveraged in real-world settings due to the immaturity of available solutions. Primarily, these solutions are essentially proofs of concept that lack scalability and are challenging for experts to evaluate routinely, primarily due to the absence of a graphical user interface to present information.

Our aim was to assess the potential of integrating social media into pharmacovigilance and enhancing decision-making with this novel data source. To achieve this, our objective was to develop and assess a pipeline that processes data from the extraction of web forum posts to the generation of indicators and alerts within a visual and interactive environment. The goal was to create a user-friendly tool that enables regulatory authorities to make better-informed decisions effectively.

This article presents the design and implementation of our pipeline dedicated to harnessing posts from social media. In addition, we showcase the use of the pipeline through a specific use case, emphasizing the importance of monitoring drugs in social media to better address patients' expectations.

Methods

Overview

The PHARES project (Pharmacovigilance in Social Networks), funded from 2017 to 2019 by the French ANSM, aimed to develop a software suite (a pipeline) enabling pharmacovigilance users to analyze social networks, particularly messages posted on forums. The objective of the pipeline is to facilitate routine use through continuous post extraction and quantitative data analysis from web forums, specifically tailored for the French language.

The pipeline is made up of 4 modules, each referring to its own methods (Figure 1):

The Scraper module, which extracts posts from forums using a previously developed tool, Vigi4Med (V4M) scraper [9], and produces a comma-separated values (CSV) file filled with the texts extracted.

The Annotation module, which extracts elements of interest from the posts and registers annotations in CSV files, with each line representing an annotation of an ADE or a drug. When a causality relationship is identified, both an ADE and a drug are annotated on the same line.

The Statistical module, which performs quantitative analysis on the annotated posts, generating numerical data, tables, or figures.

Figure 1. Structure of the PHARES' pipeline, with modules in blue and data information in red. CSV: comma-separated values; PHARES: Pharmacovigilance in Social Networks.

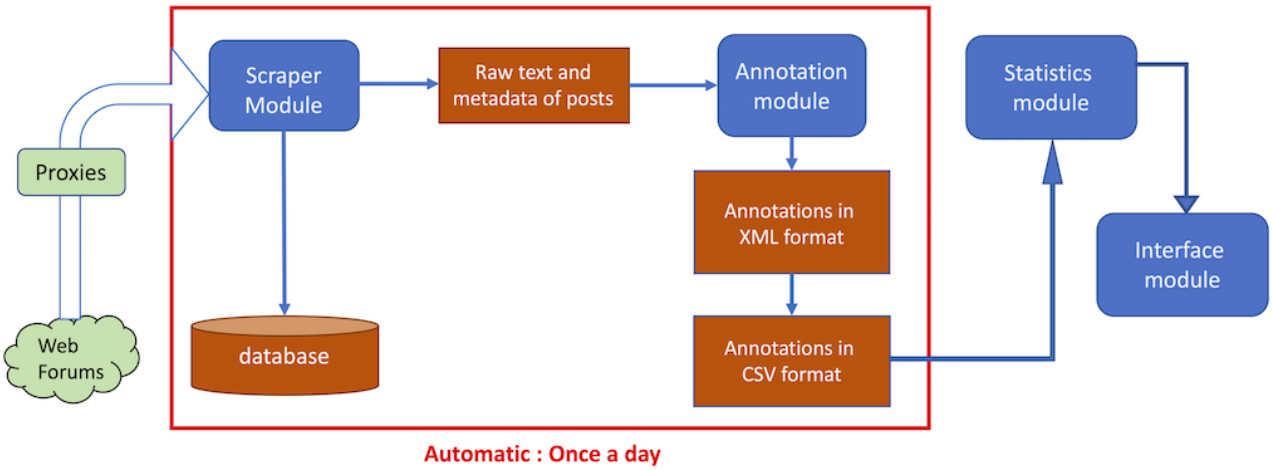

The Interface module, which supports query definition and visualization of results.

The methodology used to evaluate the PHARES pipeline involved comparing its performance with existing platforms mentioned above, in accordance with a set of criteria established with prospective PHARES users. The criteria, specific to each module, are as follows:

- General level: focus on ADEs, designed for routine usage.
- Scraper: collects all posts of a selected website, performs deidentification, allows to extract posts from web forums, and is open source.
- Statistics: the temporal evolution of posts or annotations is displayed and a change-point analysis (detecting breakpoints) is possible.

XSL•FO
RenderX

- Signal detection: allows to apply at least one signal detection method, displays the temporal evolution of the proportional reporting ratio (PRR), and allows to perform a logistic regression–based signal detection method.
- Graphical user interface: has an interface for users.

Scraper Module

V4M Scraper is an open-source tool designed for data extraction from web forums [9]. Its primary functions are optimizing scraping time, filtering out posts primarily focused on advertisements, and structuring the extracted data semantically. The module operates by taking a configuration file as input, which contains the URL of the targeted forum. The algorithm navigates through forum pages and generates resource description framework (RDF) triplets for each extracted element, allowing for potential alignment with external semantic resources. A caching mechanism has been integrated into this tool to maintain a local copy of previously visited pages, thereby avoiding redundant requests to websites for already scraped web pages, particularly in cases of errors or testing, for example. Vigi4Med V4M Scraper was customized for the PHARES project, as indicated by the red elements in Figure S1 in Multimedia Appendix 1. The database format (Figure S2 in Multimedia Appendix 1) was implemented to enhance interaction with the interface. Specifically, the main scraping script was adjusted to produce a simplified tabular format (CSV) of the extracted data and to store these data in a database. This modification aims to facilitate input to the subsequent module of the pipeline (annotation). V4M Scraper was customized to enable a continuous scraping routine, wherein data extracted from web forums are automatically and regularly annotated and registered. A log file was integrated into the scraper structure to maintain a record of the last scraped element. This log file ensures that the daily routine scraping always begins from the last scraped point. An automation tool (crontab) is used to schedule the execution of the pipeline for each forum on a daily basis at a specific time.

A total of 23 public French health-related web forums were selected through a combination of Google searches and from a list of certified health websites provided by the HON Foundation, in collaboration with the French National Health Authority (HAS). The selection criteria included the requirement for websites to be hosted in France, feature a discussion board or space for sharing experiences, and have more than 10 patient contributions. Furthermore, Twitter posts are collected and analyzed by the pipeline. This is achieved using the Twitter API for data collection, followed by employing the same modules used for processing web forum posts.

Annotation Module

Entities corresponding to drugs and pathological conditions in social media were identified and annotated using an NLP pipeline [20]. Initially, conditional random fields were used to account for global dependencies [21]. Specifically, the model considers the entire sequence when making predictions for individual tokens. This approach is advantageous for entity extraction tasks, as the presence of an entity in one part of the text can influence the likelihood of other entities in the vicinity. Second, a support vector machine is used to predict the causality relationship between an entity identified as a drug and another entity identified as an ADE. The annotation method used in this module was implemented at an early stage of the pipeline's design. Currently, the named entity recognition task of this module is undergoing revision to incorporate more recent advancements in NLP algorithms [22-26].

In a third step, the detected annotations were normalized using codes from the MedDRA and the Anatomical Therapeutic Classification (ATC) to ensure they were suitable for signal detection purposes.

MedDRA is an international medical hierarchical terminology comprising 5 levels used to code potential ADEs in pharmacovigilance. The highest level is the system organ class, which is further divided into high-level group terms, then into high-level terms, preferred terms (PTs), and finally lowest level terms. Typically, the PT level is used in pharmacovigilance signal detection.

The ATC classification system is a drug classification used in France for pharmacovigilance purposes. It categorizes the active ingredients of drugs based on the organ system they primarily affect. The classification comprises 5 levels: the anatomical main group (consisting of 14 main groups), the therapeutic subgroup, the therapeutic/pharmacological subgroup, the chemical/therapeutic/pharmacological subgroup, and the chemical substance. Typically, the fifth level (chemical substance) is used in pharmacovigilance signal detection.

The outputs of the annotation module are CSV files with the following variables:

- Concerning the post: forum name, post ID, and date
- Concerning the ADE: verbatim, normalized term, unified medical language system's concept unique identifier, and MedDRA code
- Concerning the drug: verbatim, normalized term, active ingredient, and ATC code

In these CSV files, each line can consist of either an adverse event (ADE) annotation, a drug annotation, or both when a causality relationship has been identified between the drug and the ADE. Table 1 provides a sample of the database.

In a prior study, we selected posts where at least one ADE associated with 6 drugs (agomelatine, baclofen, duloxetine, exenatide, strontium ranelate, and tetrazepam) had been detected by this algorithm. A manual review revealed that among 5149 posts, 1284 (24.94%) were validated as pharmacovigilance cases [12]. The fundamental metrics used to assess the performance of the annotation module were precision (P), recall (R), and their harmonic mean F_1-score. To calculate these metrics, it is necessary to evaluate false negatives for nonrecognition of relevant terms, false positives for irrelevant recognitions, and true positives for correct recognitions. Precision, recall, and F_1-score are defined as follows:

Precision = (true positive)/(true positive + false positive); recall = (true positive)/(true positive + false negative); F_1-score = (2 × precision × recall)/(precision + recall) **(1)**

XSL•FO
RenderX

In the "Results" section, we present a comparison of the performance of the annotation module with the performance of state-of-the-art methods [8,22,25,26].

Table 1. Sample of the database after annotation and normalization; 8 lines corresponding to 8 annotated couples in the same post.

Forum name	Post ID	Date	Time	ADE[a] verbatim	ADE normalized	Concept unique identifier	Drug verbatim	Drug normalized	Active ingredient	MedDRA[b] code	ATC[c] code
Atoute	7354	October 8, 2018	21:37:00	Maux de tête	Céphalée	C0018681	Lévothyrox	LEVOTHYROX	Levothyroxine sodique	__[d]	H03AA01
Atoute	7354	October 8, 2018	21:37:00	Maux de tête	Céphalée	C0018681	Calcium	—	—	—	—
Atoute	7354	October 8, 2018	21:37:00	Nodules cancereux	—	—	Lévothyrox	LEVOTHYROX	Levothyroxine sodique	—	H03AA01
Atoute	7354	October 8, 2018	21:37:00	Nodules cancereux	—	—	Calcium	—	—	—	—
Atoute	7354	October 8, 2018	21:37:00	Fatigue	Fatigue	C0015672	Lévothyrox	LEVOTHYROX	Levothyroxine sodique	10016256	H03AA01
Atoute	7354	October 8, 2018	21:37:00	fatigue	Fatigue	C0015672	Calcium	—	—	10016256	—
Atoute	7354	October 8, 2018	21:37:00	Perte de poids	Poids diminué	C0043096	Lévothyrox	LEVOTHYROX	Levothyroxine sodique	10048061	H03AA01
Atoute	7354	October 8, 2018	21:37:00	Perte de poids	Poids diminué	C0043096	Calcium	—	—	10048061	—

[a]ADE: adverse event.

[b]MedDRA: Medical Dictionary for Regulatory Activities Terminology.

[c]ATC: Anatomical Therapeutic Classification.

[d]No data are available for this slot.

Statistical Module

This module generates general statistics and diagrams for web forums or Twitter. It provides data such as the number of annotated posts (related to the drug, the ADE, or both), the count of drug-ADE pairs identified, and the distribution of ADEs' MedDRA-PTs. In addition, a change-point analysis method was used to detect significant changes over time in the mean number of posts mentioning the drug and ADE [27].

Besides, several statistical signal detection methods were implemented to generate potential signals. Safety signals, which provide information on adverse events that may potentially be caused by a medicine, were further evaluated by pharmacovigilance experts to determine the causal relationship between the medicine and the reported adverse event.

The statistical module implements 3 signal detection methods, including 2 well-known and frequently used disproportionality signal detection methods: the PRR [28] and the reporting odds ratio (ROR) [29]. In addition, a complementary method, a logistic regression–based signal detection method known as the class imbalanced subsampling lasso [30], was used.

PRR and ROR are akin to a relative risk and an odds ratio, respectively. However, they differ in their denominators: as the number of exposed patients is typically unknown in pharmacovigilance databases, the denominator in PRR and ROR calculations is the number of cases reported in the pharmacovigilance database.

PRR and ROR are specific to each drug-ADE pair and can be directly computed from the contingency table (Table 2).

Table 2. Contingency table for disproportionality analysis.

	Adverse drug event of interest	Other adverse drug events
Drug of interest	*A*	*b*
Other drugs	*C*	*d*

The PRR compares the proportion of an ADE among all the ADEs reported for a specific drug with the same proportion for all other drugs in the database (Equation 2). A PRR significantly greater than 1 suggests that the ADE is more frequently reported for patients taking the drug of interest, while a PRR equal to 1 suggests independence between the 2 variables.

$$PRR = [a/(a + b)]/[c/(c + d)] \quad \textbf{(2)}$$

The ROR quantifies the strength of the association between drug administration and the occurrence of the ADE. It represents the ratio of the odds of drug administration when the ADE is present to the odds of drug administration when the ADE is absent (Equation 3). When the 2 events are independent, the ROR equals 1. An ROR significantly greater than 1 suggests

XSL•FO
RenderX

that drug administration is associated with the presence of the ADE.

$$ROR = ad/bc \quad \textbf{(3)}$$

We considered events over posts for the calculation of disproportionality statistics. If the same drug-ADE pair was identified multiple times within a post, the pair was counted as many times as it occurred in the calculation.

Disproportionality analysis has certain limitations, including the confounding effect resulting from coreported drugs and the masking effect, where the background relative reporting rate of an ADE is distorted by extensive reporting on the ADE with a specific drug or drug group. Caster et al [31] demonstrated through 2 real case examples how multivariate regression–based approaches can address these issues. Harpaz et al also suggested that logistic regression could be used for safety surveillance [32]. Initially designed for pharmacovigilance case reports, we hypothesize that they may also be applicable to posts.

The logistic regression model specifically focuses on a particular ADE or a group of ADEs. It involves creating a vector that represents the presence (1) or absence (0) of the ADE of interest in the pharmacovigilance case (in our case, in the post). Additionally, a matrix is generated to represent the administration or nonadministration of all drugs in the database by the patient (1 for administration and 0 for nonadministration). Figure S3 in Multimedia Appendix 1 illustrates an example of using logistic regression. In our case, we assumed that if a drug was annotated in the post, it was taken by the patient. The logistic regression aims to predict the probability of the presence of the ADE (ADE=1) of interest based on the presence of all (N_m) drugs in the database (Equation 4), where X represents the distribution of the presence/absence of the drugs. The adjusted factors included only concomitant medications, as patient-related factors are often missing in web forums' posts. Therefore, we did not need to address the impact of missing data, which should be evaluated when necessary.

$$\ln([P(X|ADE=1)]/[P(X|ADE=0)]) = a + b1 \times Drug1 + \cdots + bi \times Drug_i + \cdots + bNm \times Drug_{Nm} \quad \textbf{(4)}$$

The selection of the drugs depends on the parameter b_i. If $b_i<0$, the drug i decreases the risk of the ADE, and if $b_i>0$, the drug i increases the risk of the ADE.

Then, 2 sets are defined:

- S_1: set of n_1 posts with an annotation of the ADEs of interest.
- S_0: set of n_0 posts without an annotation of the ADEs of interest.

In our case $n_0>>n_1$, indicating a significant imbalance toward posts lacking annotations of the ADEs of interest. To address this issue, we took a subsample with a more favorable ratio of posts with annotated ADEs versus those without. Additionally, to enhance result stability, we conducted multiple draws instead of just one.

In practice, we generated B subsamples. Each subsample was constructed by randomly drawing, with replacement, n_1 posts from S_1 and R posts from S_0, where R=max($4n_1$, $4N_m$). The choice of $4n_1$ was inspired by case-control studies, while $4N_m$ was included to ensure an adequate number of observations considering the multitude of predictors.

The maximum number of drug predictors is set to 50 and the method is then applied on B=250 drawings. Finally, the distribution of interest is the distribution of the number of times the drug was selected as a predictor ($b_i>0$). The drugs retained as final predictors are those in which the [illegible] quantile of this distribution is superior to 0. [illegible] can be equal to 5, 10, or 15.

We implemented a change-point analysis method described in [27] to detect whether there was a change in the evolution over time of a chosen statistic, such as the number of a specific drug-ADE pair, the number of ADEs associated with a specific drug, or the number of drugs associated with a specific ADE. The method uses the Cumulative Sum (CUSUM) algorithm to analyze the evolution of statistics over time, comparing current values with the period mean. It identifies breakpoints by calculating the highest difference in statistical values and comparing it with random samples. The process repeats for periods before and after detected breakpoints until no more are found.

User Interface Module

The user interface module facilitates user interaction with the pipeline in a user-friendly manner. The interface comprises a dashboard divided into 2 main parts. The left dark column (Figure 2) serves as a control sidebar, where users can select parameters to filter the data, including the forum, period, drug(s) according to the ATC classification, and ADE(s) according to a level in the MedDRA hierarchy. On the right side of the interface, various visualizations are available, organized into several tabs such as "Forum Statistics" and "Consultation of Posts," with additional tabs for statistics that become active upon querying.

Before applying a specific query, the interface provides general information about the currently available data (Figure 2), including the total annotated posts since 2017 (n=2,081,296) and total annotations since 2017 (n=2,454,310). In addition, a "Consultation of Tweets" tab (not visible in the figure) displays the total annotated tweets since March 2020 (n=46,153).

Furthermore, several tabs corresponding to different types of statistics, including "Forums Statistics" and "Twitter Statistics," provide general statistics and diagrams for web forums and Twitter. Examples of these are pie charts showing forum distribution, line charts depicting the evolution of drug and ADE mentions, histograms displaying ADE distribution by system organ class, and line charts illustrating the temporal trend of posts containing the drug and an ADE, as shown in Figures 3 and 4. The "Annotations Plot" tab displays annotations of drugs and adverse effects selected by the user, along with forum information, PTs, high-level terms, high-level group terms, dates, and hours. The "Logistic Regression" tab allows users to choose parameters for applying logistic regression. In the "Disproportionality" tab, users can choose between the PRR and ROR methods, with the time evolution of the chosen method

XSL•FO
RenderX

displayed. The “Change-Point” tab enables analysis of temporal evolution, with identified breakpoints indicated. The “Consultation of Posts” and “Consultation of Tweets” tabs provide details on annotated posts/tweets, including downloadable tables. The statistical module performs calculations based on user queries, updating the interface accordingly. If multiple drugs or adverse events are selected, they are treated as new entities for analysis.

The interface was implemented using the R language and environment (R Foundation) for statistical computing and graphics [33], leveraging the Shiny package [34] for development.

Figure 2. Home of the interface before a query. The section on the left allows to perform a query, while the central section shows the total number of annotated posts since 2017 (n=2,081,296) and the total number of annotations since 2017 (2,454,310).

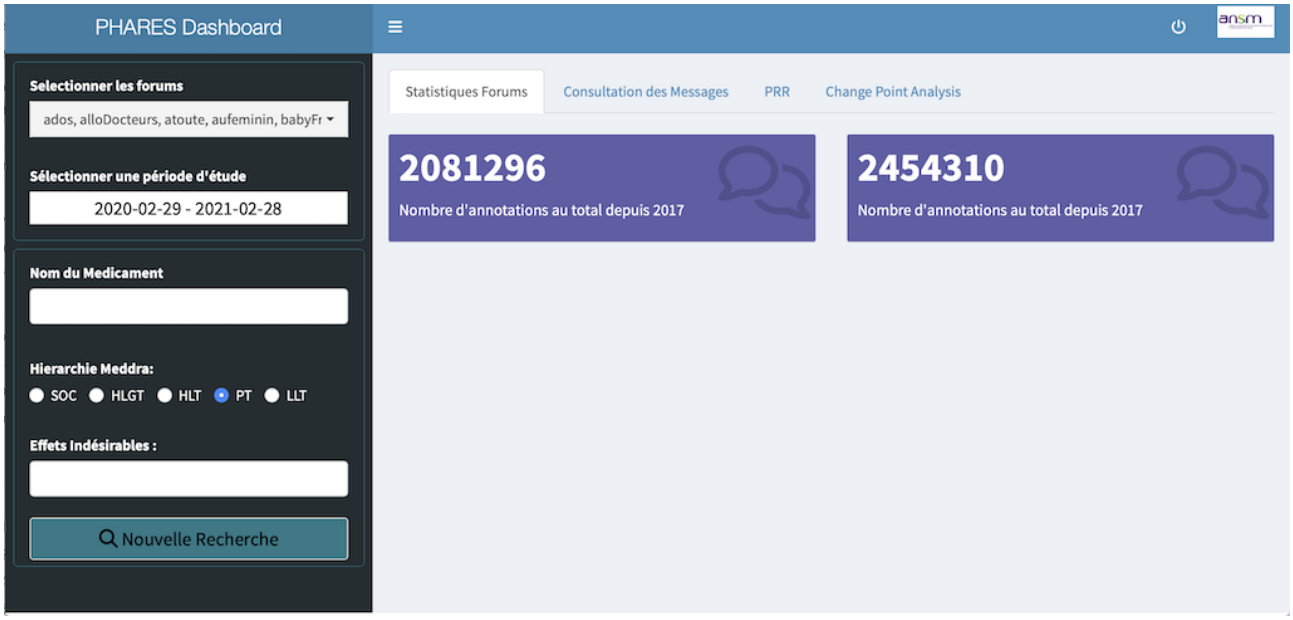

Figure 3. The Statistic forum tab displaying the central section's upper part after a query (Paracetamol and dizziness). The image displays the number of posts related to the selected couple, the number of times the selected couples is annotated, a pie chart with the distribution of web forums, and a line chart with the evolution of the mentions of the drug, the ADE, and the couple. ADE: adverse drug event.

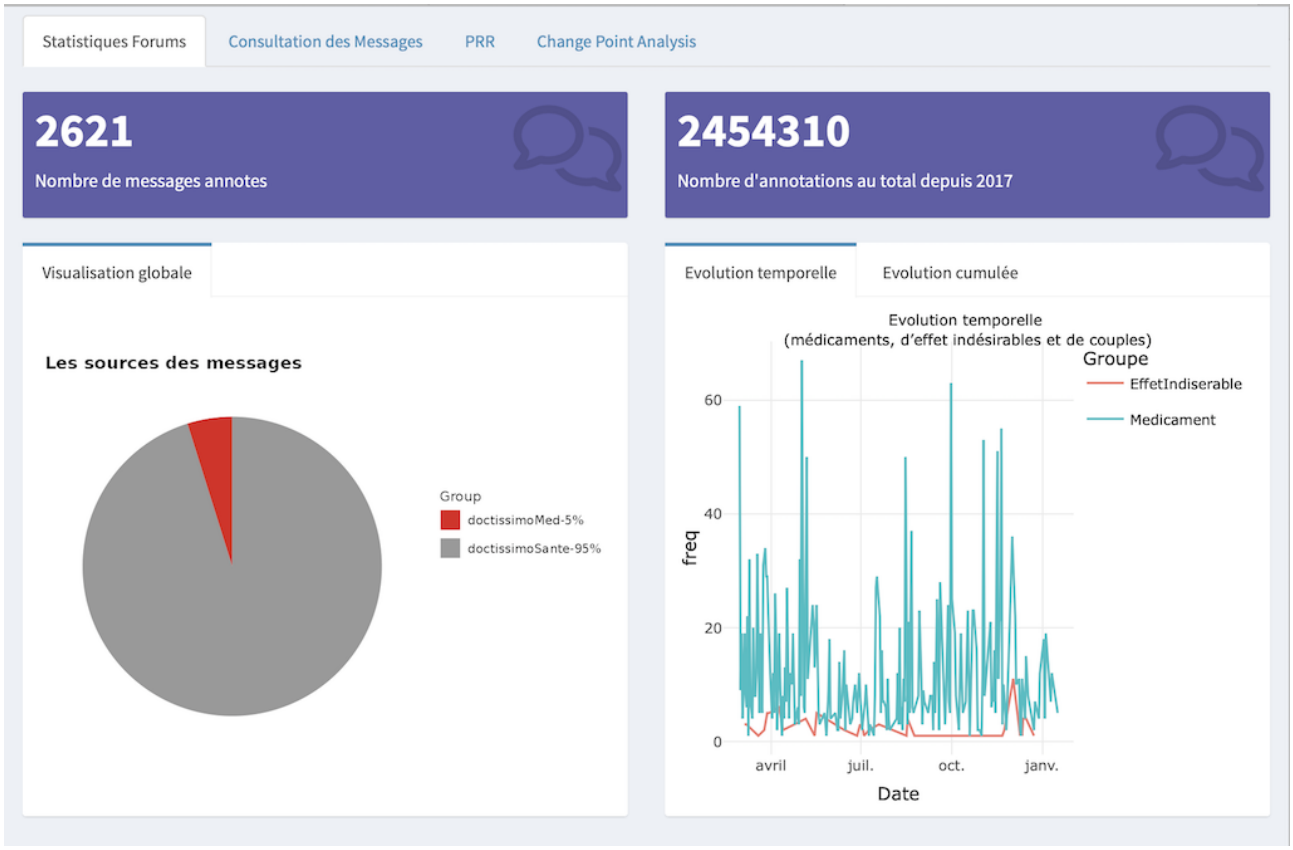

XSL•FO
RenderX

Ethical Considerations

A statement by an Institutional Review Board was not required because we used only publicly available data that do not necessitate Institutional Review Board review.

This study complied with the European General Data Protection Regulation (GDPR), which has been in force since 2018 in Europe [35]. The GDPR enhances the protection of individuals by introducing the right to be informed about the processing of personal data. However, informing each user individually may be impractical. Therefore, the GDPR introduces 2 legal conditions where informed consent is not mandatory, which can be interpreted as supporting the processing of web forum posts for pharmacovigilance (Article 9): "(e) processing relates to personal data which are manifestly made public by the data subject; [. . .] (i) processing is necessary for reasons of public interest in the area of public health, such as [. . .] ensuring high standards of quality and safety of health care and of medicinal products . . .." The GDPR also requires data processing to "not permit or no longer permits the identification of data subjects" (Article 89). Deidentification was conducted during the extraction of posts from web forums to ensure privacy [9]. User identifiers in the main RDF file were encrypted using the SHA1 algorithm [36]. The correspondence between these encrypted identifiers and the original keys is presented in RDF triplets in a separate file, referred to as the "keys file." Therefore, the only way to retrieve the original authors' identities is by concatenating the main RDF containing the encrypted data with the keys file, which is kept in a secured location. Moreover, all our data processing was carried out on a secured server with restricted access.

Results

General Results About the Pipeline

The primary outcome of this study is the operational PHARES pipeline itself. Daily extraction and annotation of posts are initiated and imported into the database linked to the user interface. In this paper, the platform's use will be demonstrated through a specific use case on the analysis of Levothyrox ADE mentions in forums (discussed later). In addition, we conducted a comparative analysis of the PHARES pipeline with the existing platforms mentioned in the "Introduction" section, based on the criteria listed in the "Methods" section.

Of the 10 identified pipelines, half were public and half were private. While 8 out of 10 focused on ADEs, only 4 were designed for routine usage. Five scrapers were open source, and all posts from considered websites were extracted by only 6 of the scrapers (with others extracting posts under certain conditions). Six scraped web forum posts, but only 3 performed deidentification. Additionally, 4 pipelines focused on the French language. A total of 6 pipelines displayed the temporal evolution of the number of posts, but only 1 conducted a change-point analysis. Signal detection methods were performed by only 4 of them, with none displaying the temporal evolution of the PRR nor a logistic regression–based method. Finally, 6 of them had an interface (Table 3).

XSL•FO
RenderX

Table 3. PHARES[a] and identified pipelines' characteristics match with the identified evaluation criteria[b].

Pipeline	General			Scraper				Annotation	Statistics		Signal detection			
	Focus on ADEs[c]	Routine usage	Public/private	All posts	Deidentification	Web forums	Open source	French language	Temporal evolution	Change-point analysis	Signal detection	PRR[d] temporal evolution	Logistic regression	Interface
PREDOSE[e]	X	✓	Public	✓	X	✓	✓	X	✓	X	X	X	X	✓
Insight Explorer	✓	X	Private	X	X	X	✓	X	X	X	X	X	X	✓
MedWatcher Social	✓	✓	Public	X	X	✓	✓	X	✓	X	✓	X	X	✓
Yeleswarapu et al [6]	✓	X	Private	X	X	X	X	X	X	X	✓	X	X	X
Domino	X	✓	Public	✓	X	✓	✓	✓	✓	X	X	X	X	✓
Nikfarjam et al [7]	✓	X	Public and Private	X	X	X	X	X	X	X	X	X	X	X
Magge et al [8]	✓	X	Public	✓	X	X	✓	X	✓	X	X	X	X	X
ADR-PRISM[f]	✓	X	Public and Private	✓	✓	✓	X	✓	✓	X	✓	X	X	✓
Kappa Santé	✓	✓	Private	✓	✓	✓	X	✓	✓	✓	✓	X	X	✓
Expert System	✓	X	Private	✓	✓	✓	X	✓	X	X	X	X	X	✓

[a]PHARES: Pharmacovigilance in Social Networks.

[b]The X symbol means that the characteristic is missing and the symbol ✓ means the characteristic is fulfilled.

[c]ADE: adverse drug event.

[d]PRR: proportional reporting ratio.

[e]PREDOSE: Prescription Drug Abuse Online Surveillance and Epidemiology.

[f]ADR-PRISM: Adverse Drug Reaction from Patient Reports in Social Media.

Annotation Module's Comparison With Up-to-Date State-of-the-Art Methods

We also compared the performance of our annotation process with those of up-to-date state-of-the-art methods (Table 4).

While the annotation module demonstrated good performance for named entity recognition (F_1-score=0.886), it remains slightly below the state of the art. Presently, in medical texts, the best performances are achieved by Hussain et al [25] and Ding et al [26] for the named entity recognition task, and by Xia [22] for the relationship extraction task. On Twitter, known for its notably more complex data, Hussain et al [25] achieved slightly better results than our annotator, while Ding et al [26] achieved slightly worse results.

XSL•FO
RenderX

Table 4. Comparison of our annotation process' performances with up-to-date state-of-the-art methods. Performances are given as precision, recall, and F1-score and are divided into 2 categories[a].

Annotator	Language	Data	Natural language processing method	Named entity recognition (precision; recall; F_1-score)	Relationship extraction (precision; recall; F_1-score)
PHARES[b]	French	Patient's web drug review	Conditional random fields and support vector machines	0.926; 0.845; 0.886	0.683; 0.956; 0.797
Magge et al [8]	English	Twitter	BERT[c] neural networks	0.82; 0.76; 0.78	—[d]
Xia [22]	English	Medical texts	HAMLE[e] model	—	0.929; 0.914; 0.921
Hussain et al [25]	English	Medical texts (PubMed) and Twitter	BERT	0.982; 0.964; 0.976 (PubMed) and 0.840; 0.861; 0.896 (X/Twitter)	—
Ding et al [26]	English	Medical texts (PubMed) and Twitter	BGRU[f] + char LSTM[g] attention + auxiliary classifier	0.867; 0.948; 0.906 (PubMed) and 0.785; 0.914; 0.844 (Twitter)	—

[a]The 2 categories are entity recognition, which is the detection of a drug or ADE mention, and relationship extraction, which is the detection of a relation between a drug and an ADE.

[b]PHARES: Pharmacovigilance in Social Networks.

[c]BERT: Bidirectional Encoder Representations from Transformer.

[d]Not available.

[e]HAMLE: Historical Awareness Multi-Level Embedding.

[f]BGRU: Bidirectional Gated Recurrent Unit.

[g]LSTM: Long-Short-Term-Memory.

Summary of the Result

From January 1, 2017, to February 28, 2021, a total of 2,081,296 posts were extracted from 23 French web forums (Table 5). We obtained 713,057 normalized annotations of drugs, 1,527,004 normalized annotations of ADEs, and 437,192 annotations of normalized drug-ADE couples. The number of posts annotated with at least one normalized drug-ADE couple was equal to 125,279 (6.02%). Table 4 summarizes the number of posts extracted per forum, the publication dates, and the description of the web forum. For 1 forum, the publication dates were not available. A total of 9 were generalist health forums, 3 were specialized for parents of a young baby, 2 for families, 3 for mothers, 2 specialized in thyroid issues, 1 for pregnant women, 1 for women, 1 for parents of a teenager or for teenagers, 1 for sports persons, and 1 specialized in rare diseases.

XSL•FO
RenderX

Table 5. Number of extracted posts per forum, publication dates of the first and last extracted posts, and forums' descriptions.

Forum	Extracted posts, n	Publication date of the first extracted post	Publication date of the last extracted post	Description
thyroideNEW	451,253	February 15, 2001	February 25, 2021	Specialized in thyroid issues
doctissimoSante	248,691	March 19, 2003	January 16, 2021	Generalist health forum
doctissimoNutrition	183,730	December 30, 2002	January 16, 2021	Specialized in nutrition
infoBebe	127,341	November 30, 2000	March 08, 2019	Specialized for parents of a young baby
atoute	118,415	February 05, 2005	February 28, 2021	Generalist health forum
notreFamille	97,098	March 16, 2000	October 26, 2017	Specialized for families
magicMaman	96,713	June 14, 1999	February 22, 2021	Specialized for mothers
doctissimoMed	95,531	August 05, 2002	January 15, 2021	Generalist health forum
doctissimoGrossesse	93,449	November 09, 2006	January 15, 2021	Specialized for pregnant women
thyroide	73,376	September 25, 2001	January 07, 2019	Specialized in thyroid issues
aufeminin	72,732	April 05, 2001	January 09, 2020	Specialized for women
mamanVie	69,167	June 07, 2006	April 10, 2019	Specialized for mothers
onmeda	61,428	July 25, 2001	February 24, 2021	Generalist health forum
ados	58,181	June 20, 2006	March 08, 2019	Specialized for parents of a teenager or for teenagers
carenity	52,659	May 16, 2011	August 29, 2020	Generalist health forum
famili	51,844	November 06, 2000	November 17, 2019	Specialized for families
babyFrance	43,806	January 20, 2003	April 30, 2018	Specialized for parents of young baby
bebeMaman	38,450	__[a]	—	Specialized for mothers of young baby
alloDocteurs	15,833	June 15, 2009	February 09, 2021	Generalist health forum
reboot	9383	May 04, 2016	February 25, 2021	Generalist health forum
futura	6765	May 12, 2003	February 22, 2021	Generalist health forum
sportSante	6350	May 10, 2011	January 14, 2020	Specialized for sportsperson
maladieRares	4827	October 09, 2012	May 14, 2020	Specialized in rare diseases
queChoisir	4250	June 16, 2003	February 11, 2021	Generalist health forum

[a]Not available.

Use Case: Analysis of Levothyrox ADE Mentions in Forums

To demonstrate the usage of the pipeline, we chose to focus on Levothyrox as a case study. Levothyrox is a drug prescribed in France since 1980 for hypothyroidism and circumstances where it is necessary to limit the thyroid-stimulating hormone. In 2017, a new formula of Levothyrox, differing from the 30-year-old drug at the excipient level (with lactose being replaced by mannitol and citric acid in the new formula), was marketed with widespread media coverage. In parallel, an unexpected increase in notifications of ADEs for this drug was detected. Viard et al [37] were unable to find any pharmacological rationale to explain that signal. Approximately 32,000 adverse effects were reported by patients in France in 2017, representing 42% of all the ADEs collected yearly [38]. Most of these notifications concerned the new formulation of Levothyrox and led to the "French Levothyrox crisis." In 2017, 1664 notifications of ADEs were spontaneously reported by patients to the Pharmacovigilance Center of Nice. Among the 1544 reviewed notifications, 1372 concerned Levothyrox while only 172 concerned other drugs [37].

In this use case, the study period was from January 1, 2017, to February 28, 2021, and the drugs included were 2 drugs from the "H03AA Thyroid hormones" ATC class: "Levothyroxine sodium" and "associations of levothyroxine and liothyronine." A total of 17 forums were selected as they included at least one post with information about these drugs. Posts were extracted, annotated, and analyzed through the pipeline from several forums (Table 6). Signal detection methods were applied to an ADE chosen as it frequently appeared with Levothyrox in our data: "tiredness." A signal can be detected when the lower bound of the 95% CI of the logarithm of the PRR is greater than 0. For logistic regression, we applied the tenth quantile. A total of 11,340 posts contained an annotation concerning the drugs of interest. Figure S4 in Multimedia Appendix 1 illustrates the source and evolution over time of these posts. Out of a total of 50,127 annotations of Levothyrox, they principally originated from the Vivre sans thyroïde forum and were mostly posted in

XSL•FO
RenderX

mid-2017 (Figure 4, Table 6). The results of the statistical analysis were displayed by the user interface.

ADEs annotated with Levothyrox were mainly from system organ classes: general disorders and administration site conditions (29.6%), metabolism and nutrition disorders (11.6%), and endocrine disorders (11.4%). The PTs mostly found in association with Levothyrox are listed in Table 7. All this information is accessible in the interface module (Figure S5 in Multimedia Appendix 1).

We chose the PT "tiredness" for the signal detection analysis. A total of 85,976 posts were annotated with either one of the drugs of interest or the ADE tiredness. Among them, 1841 Levothyrox-tiredness couples were found, mostly in 2017 (Table 7).

Figure 5 illustrates the time evolution of the PRR for the Levothyrox-tiredness couple. Figure S6 in Multimedia Appendix 1 displays the source and evolution over time of French web forums' posts for this couple. A signal is consistently generated throughout the period as the logarithm of the PRR is always greater than 0.

Figure 4. The Statistic forum tab showing the central section's lower part after a query (Paracetamol and dizziness). The image displays an histogram of the distribution of ADEs grouped under "Preferred Terms." ADE: adverse drug event.

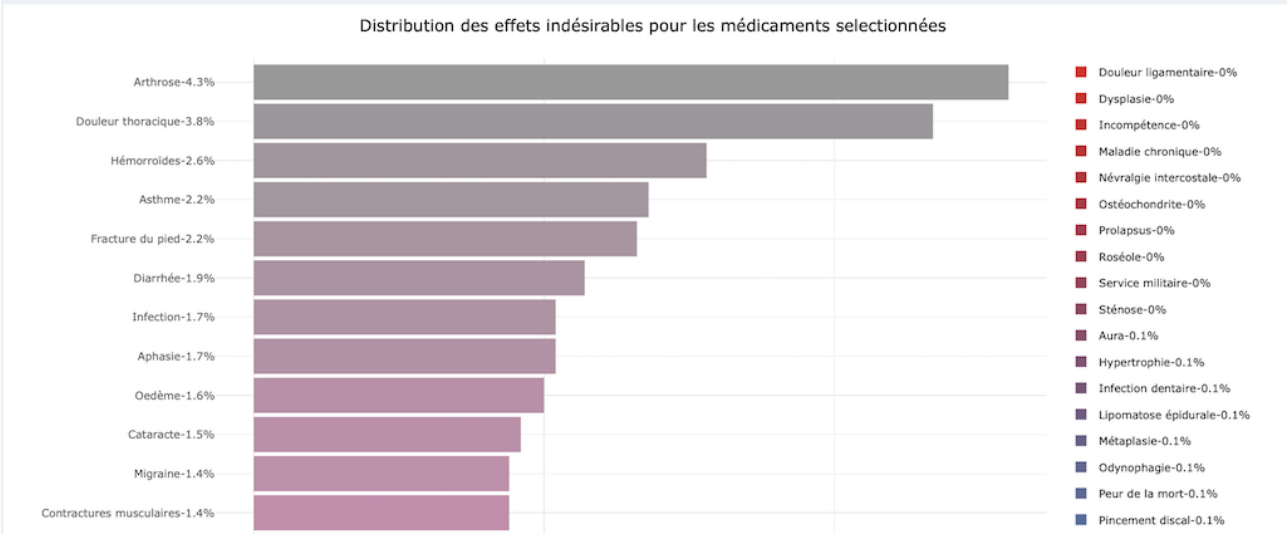

Table 6. Number of annotations of Levothyrox per forum in 11,340 posts from French web forums from January 1, 2017, to February 28, 2021.

Forum	Value, n	Cumulative frequency, %
Vivre sans thyroïde	41,211	82.21
Doctissimo Santé	4230	90.65
Doctissimo Grossesse	1476	93.60
Doctissimo Nutrition	1177	95.94
Carenity	863	97.67
Allo docteurs	502	98.67
Atoute	170	99.01
Doctissimo medicaments	166	99.34
Que choisir	85	99.51
Maladie rares	76	99.66
Au feminin	58	99.77
Sport santé	50	99.87
Onmeda	48	99.97
Famili	7	99.98
Futura	5	99.99
Maman vie	2	100.00
Magic maman	1	100.00

XSL•FO
RenderX

Table 7. The 20 preferred terms most frequently found with Levothyrox in 11,340 posts from French web forums from January 1, 2017, to February 28, 2021.

Preferred terms	Values, n
Pain	1882
Tiredness	1841
Faintness	1267
Hypothyroidism	1110
Dizziness	912
Insomnia	627
Palpitations	571
Hyperthyroidism	568
Malignant tumor	560
Anxiety	498
Overdose	490
Nervous tension	484
Myalgia	409
Nausea	388
Stress	380
Diarrhea	354
Tachycardia	322
Muscle spasms	321
Convulsions	302
Arthralgia	276

Figure 5. Evolution of the monthly cumulated PRR over time for the annotated couples of Levothyrox-tiredness from 11,340 French web forum posts from January 1, 2017, to February 28, 2021. PRR: proportional reporting ratio.

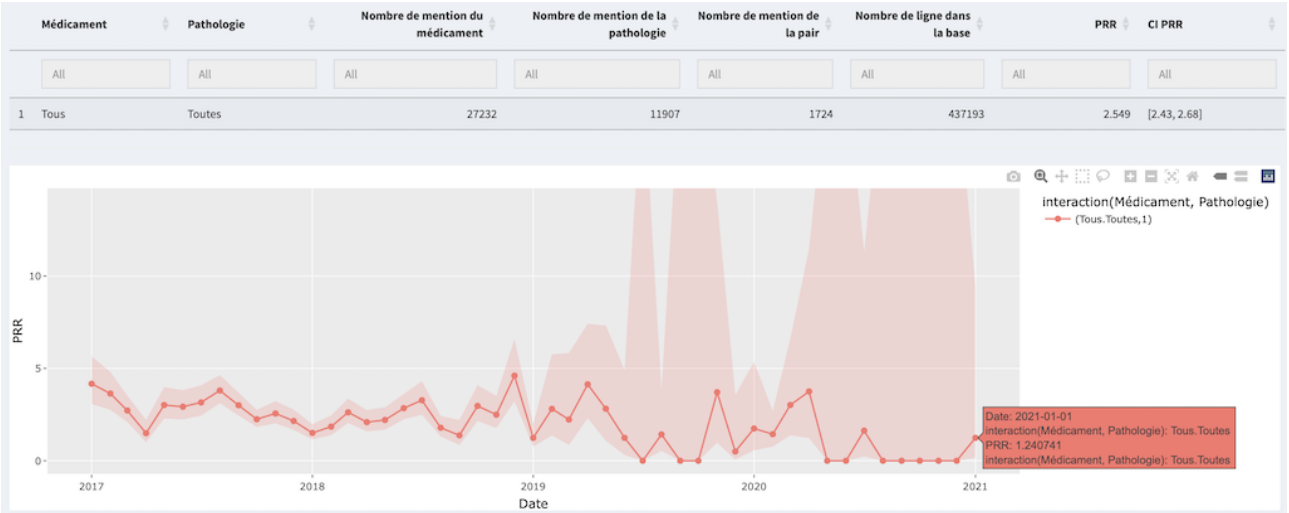

A total of 11 drugs were found to be associated with tiredness using logistic regression: paclitaxel, pegfilgrastim, Levothyrox, glatiramer acetate, escitalopram ferrous sulfate, the combination of Levothyrox and liothyronine, secukinumab, methotrexate, bismuth potassium, tetracycline, and metronidazole.

Change-point analysis was conducted on the monthly evolution of the number of Levothyrox-ADE couples detected in web forums. Six breakpoints were identified (Figure 6), and 3 of them correlated with an increase in the number of ADEs found with Levothyrox on web forums. These increases occurred in August 2017 and in September and December 2018.

This use case demonstrates that the results obtained through the pipeline, particularly in the context of Levothyrox, align with findings in the literature derived from more traditional data sources such as case reports in pharmacovigilance (see the "Discussion" section). It underscores the potential of leveraging

XSL•FO
RenderX

such a pipeline to monitor a drug, not only retrospectively but also in real time using social media. Consequently, PHARES has the capability to potentially uncover new signals in pharmacovigilance.

Figure 6. Breakpoints identified with the CPA method (vertical red lines) in the temporal evolution of the number of Levothyrox-ADE couples (horizontal red line) annotated in 11,340 French web forums posts from January 1, 2017, to February 28, 2021. ADE: adverse drug event; CPA: change-point analysis.

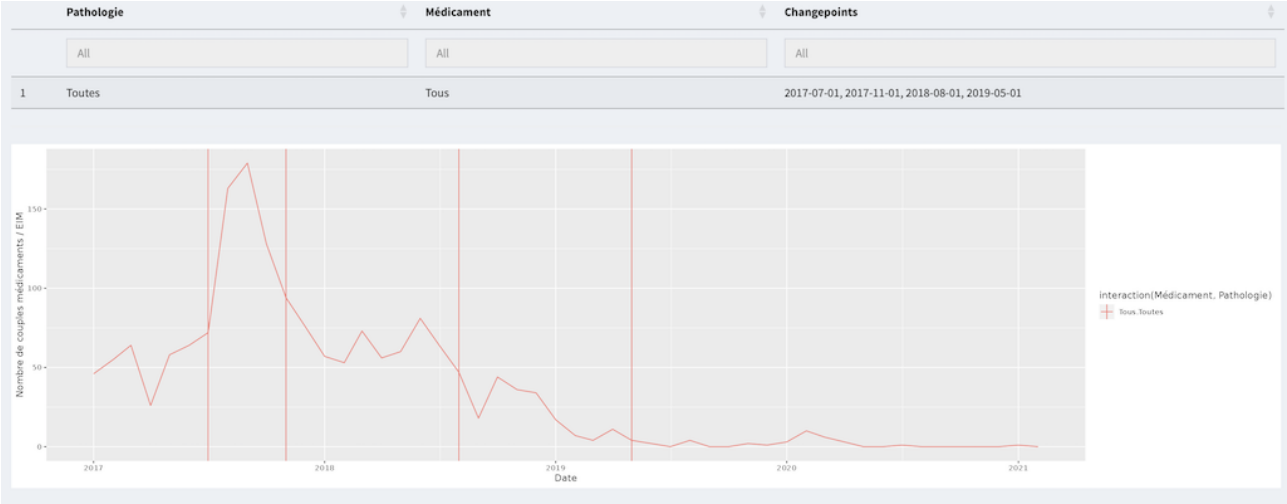

Discussion

Principal Findings

To align with our objective, we implemented and evaluated a pipeline that processes data from the extraction of web forum posts to the generation of indicators and alerts within a visual and interactive environment. Through this pipeline, we demonstrated that quantitative analysis can be conducted through the interface without requiring the user to code. We discovered the feasibility of acquiring information akin to the literature regarding a drug's ADEs, as well as unexpected ADEs and significant event dates related to a drug. This underscores the relevance and utility of such a pipeline.

A conceptual contribution of this research was the proposal of a methodology for designing a pipeline to facilitate pharmacovigilance studies on web forums. This involved describing 4 independent modules and outlining their interactions. Additionally, another contribution was the adaptation of certain pharmacovigilance analysis methods for the examination of data extracted from web forum posts. The logistic regression–based method presented in this article was originally tailored for pharmacovigilance cases to consider co-prescriptions of drugs. We have adapted it to suit the analysis of pharmacovigilance data extracted from web forum posts.

Comparison With Prior Work

The PHARES pipeline offers added value compared with previous pipelines in terms of the criteria set, which reflects an analysis of experts' needs for routine monitoring of ADEs in social media. Unlike previous approaches, the scrapers used in PHARES routinely perform deidentification, and the inclusion of change-point analysis, the evolution of PRRs over time, and a logistic regression–based signal detection method were previously unavailable. The temporal evolution of the number of posts and a signal detection method are also seldom supported. Designed for routine usage and focused on ADEs, all posts from selected web forums are scraped and deidentified using an open-source scraper.

The period and selected web forums differed between both studies: Audeh et al [38] covered the period from January 2015 to December 2017, while our study spanned from January 2017 to February 2021. Additionally, Audeh et al [38] included only 1 web forum specialized in thyroid issues, whereas we incorporated this specific forum along with 16 others. The main ADEs associated with Levothyrox in our study align with those found by Audeh et al [38] on similar data, albeit without using the interface. In our study, the 10 most frequent symptoms were pain, tiredness, faintness, hypothyroidism, dizziness, insomnia, palpitations, hyperthyroidism, malignant tumor, and anxiety. By contrast, Audeh et al [38] reported tiredness, weight gain, pain, ganglions, hot flush, chilly, inflammation, faintness, weight loss, and discomfort.

Furthermore, the PHARES pipeline surpasses previous efforts, particularly regarding several criteria. These include the annotation tool, where only 4 pipelines were identified using a French annotator tool. In terms of available statistics, only 1 pipeline met both criteria we identified. Regarding signal detection, among the 3 criteria identified, 5 pipelines matched with only 1, while the remaining 5 matched with none.

In the use case, a notable increase in the number of ADEs associated with Levothyrox was detected using the change-point analysis method a few months after the introduction of the new formula in March 2017, specifically in August 2017. This surge coincided with the initial declaration to the pharmacovigilance network and a petition initiated by patients to reintroduce the former formula in June 2017. We compared these findings with results from a pharmacovigilance study based on spontaneous reporting. Out of 1554 notifications spontaneously addressed by patients to the Pharmacovigilance Center of Nice from January 1, 2017, to December 31, 2017, 1372 were related to the new formula of Levothyrox, representing 7342 ADEs. Our comparison with these data clarified our findings. The 10 most

XSL•FO
RenderX

frequently reported ADEs in these notifications closely resembled our own results [37]. These were asthenia, headache, dizziness, hair loss, insomnia, cramps, weight gain, nausea, muscle pain, and irritability. Consequently, our results demonstrate coherence with the existing literature. This study illustrates the feasibility of identifying the date of significant events related to a drug. However, it is noteworthy that the detection of such events is not necessarily expedited through social media compared with the traditional pharmacovigilance system.

Limitations

The method used in our annotation process was integrated at an early stage during the pipeline's design. Regarding the identification of drugs and symptoms, our annotation process exhibited the following performances: precision=0.926, recall=0.845, and F_1-score=0.886 [20]. Similarly, for discerning the relationship between the drug and the ADEs, the performances were precision=0.683, recall=0.956, and F_1-score=0.797 [20]. This study marked the inaugural publication on using NLP methods to identify ADEs in French-language web forums. The annotation process was thus developed using contemporary state-of-the-art methodologies at the time. However, it would now stand to gain from the integration of more recent NLP algorithms for named entity recognition [8,23,24]. These newer algorithms offer comparable performances while effectively handling more complex data, thereby enhancing the efficacy of NLP analysis. However, because of our emphasis on the genericity of the approach and the interoperability between the different modules rather than solely focusing on the performance of each module, we opted not to use these algorithms. Nevertheless, contemporary state-of-the-art methods for annotating ADEs from social media posts encompass convolutional neural networks trained on top of pretrained word vectors for sentence-level classification [24] and transformers using the bidirectional encoder representations from transformers (BERT) language model [39]. Hussain et al [25] introduced a multitask neural network based on BERT with hyperparameter optimization capable of sentence classification and named entity recognition. This model achieved performances of precision=0.840, recall=0.861, and F_1-score=0.896 on the Twitter (X)-TwiMed data set. Additionally, Magge et al [8] presented a pipeline consisting of 3 BERT neural networks designed to classify sentences, extract named entities, and normalize those entities to their respective MedDRA concepts. The performances of this model were as follows: precision=0.82, recall=0.76, and F_1-score=0.78 on the SMM4H-2020 data set (Twitter/X). Thanks to our modular design, it will be straightforward to substitute our current annotation process with an enhanced model in the future.

Several limitations should be acknowledged for future work. First, the scraper relies on the HTML structure of web forums, necessitating updates to its configuration files if a forum alters its page design. Additionally, our interface lacks the capability to incorporate alternate identifiers for drugs or ADEs. For instance, patients may commonly refer to the drug "baclofen" as "baclo" on social media platforms. Consequently, the number of posts pertaining to a drug or ADE could potentially be underestimated.

Forums must be selected before query execution to mitigate calculation time. However, selecting forums based on the presence of information related to a particular drug or ADE can introduce bias into signal detection methods, particularly in disproportionality analysis, where the drug-ADE pair may be overrepresented. Another limitation in qualitative analysis of posts is the inability of users to edit annotations or record typical pharmacovigilance qualitative data.

The assumption that all drugs mentioned in a post were consumed simultaneously by the user, as applied in the logistic regression–based method, introduces an evident bias.

One limitation associated with the use of social media data pertains to fraudulent posts. The pseudonymity inherent in these platforms provides malevolent individuals with the opportunity to disseminate false rumors. Additionally, patients might post identical or similar messages across multiple discussion boards, or even multiple times on the same board. Thus, it is crucial to consider these factors to mitigate biases in signal detection.

Perspectives

In the short to medium term, our objectives are updating the annotation module to enhance accuracy, improving the qualitative analysis by enabling users to edit and correct annotations, and expanding the range of signal detection methods available in the statistics module.

This method could indeed be beneficial for identifying potential drug misuse and unknown ADEs [40]. By categorizing pathological terms found in web forums based on their presence in the summary of product characteristics, we can distinguish between indications, known ADEs, and potential instances of drug misuse or unexpected ADEs. However, it is important to note that considering all pathological terms found in the summary of product characteristics as indications might obscure cases of drug inefficiency. Therefore, a nuanced approach is necessary to ensure comprehensive and accurate analysis.

We next tested our pipeline from the perspective of end users. However, the hypothesis was only partially confirmed, indicating the need for further studies. These studies should include evaluations with ergonomic criteria.

In the long term, our vision is to expand this tool to encompass other languages and themes beyond pharmacovigilance. This includes areas such as drug misuse, the consumption of food supplements, and the use of illegal drugs. French web forums dedicated to recreational drug use already exist, providing a valuable source of data for such endeavors.

Conclusions

Our hypothesis focused on the challenge encountered by regulatory agencies in using social media, primarily because of the lack of appropriate decision-making tools. To tackle this challenge, we devised a pipeline consisting of 4 editable modules aimed at effectively analyzing health-related French web forums for pharmacovigilance purposes. Using this pipeline and its user-friendly interface, we successfully demonstrated

XSL•FO
RenderX

the feasibility of conducting quantitative analyses without the need for coding. This approach yielded coherent results and holds the potential to reveal new insights about drugs.

A practical implication of our pipeline is its potential application in health surveillance by regulatory agencies such as the ANSM or pharmaceutical companies. It can be instrumental in detecting issues related to drug safety and efficacy in real time. Furthermore, research teams can leverage this tool to retrospectively analyze events and gain valuable insights into pharmacovigilance trends.

Acknowledgments

The annotation module was developed by François Morlane-Hondère, Cyril Grouin, Pierre Zweigenbaum, and Leonardo Campillos-Llanos from the Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI). Code review for the graphical user interface in R language was performed by Stevenn Volant under a contract with the Stat4Decision company. Stat4Decision was not involved in designing the study and writing this article. This work was funded by the Agence nationale de sécurité du médicament et des produits de santé (ANSM) through Convention No. 2016S076 and was supported by a PhD contract with Sorbonne Université.

Data Availability

Our data were extracted from web forums that do not allow data sharing. Thus, as we are not the owners of the data we cannot make the data available. The scrapper we developed to extract these data is open source and can be used to extract data from web forum posts. The tool as well as full documentation (in English and French) of the code and configuration file are available online [41].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Vigi4Med Scraper structure, PHARES database structure, example of data representation, and source and evolution over time of web forum posts. PHARES: Pharmacovigilance in Social Networks.
[DOCX File , 931 KB - jmir_v26i1e46176_app1.docx]

References

1. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions : a systematic review. Drug Saf 2006;29(5):385-396. [doi: 10.2165/00002018-200629050-00003] [Medline: 16689555]
2. Liu F, Jagannatha A, Yu H. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. Drug Saf 2019 Jan;42(1):95-97 [FREE Full text] [doi: 10.1007/s40264-018-0766-8] [Medline: 30649734]
3. Warrer P, Hansen EH, Juhl-Jensen L, Aagaard L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. Br J Clin Pharmacol 2012 May;73(5):674-684 [FREE Full text] [doi: 10.1111/j.1365-2125.2011.04153.x] [Medline: 22122057]
4. Black C, Tagiyeva - Milne N, Helms P, Moir D. Pharmacovigilance in children: detecting adverse drug reactions in routine electronic healthcare records. A systematic review. Brit J Clinical Pharma 2015 May 28;80(4):844-854. [doi: 10.1111/bcp.12645] [Medline: 25819310]
5. Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, et al. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. Journal of Biomedical Informatics 2013 Dec;46(6):985-997. [doi: 10.1016/j.jbi.2013.07.007] [Medline: 23892295]
6. Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Med Inform Decis Mak 2014 Feb 24;14(1):1-16. [doi: 10.1186/1472-6947-14-13]
7. Nikfarjam A, Ransohoff JD, Callahan A, Jones E, Loew B, Kwong BY, et al. Early detection of adverse drug reactions in social health networks: a natural language processing pipeline for signal detection. JMIR Public Health Surveill 2019 Jun 03;5(2):e11264. [doi: 10.2196/11264] [Medline: 31162134]
8. Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. J Am Med Inform Assoc 2021 Sep 18;28(10):2184-2192. [doi: 10.1093/jamia/ocab114] [Medline: 34270701]
9. Audeh B, Beigbeder M, Zimmermann A, Jaillon P, Bousquet C. Vigi4Med scraper: a framework for web forum structured data extraction and semantic representation. PLoS One 2017 Jan 25;12(1):e0169658. [doi: 10.1371/journal.pone.0169658] [Medline: 28122056]

XSL•FO
RenderX

10. Caster O, Dietrich J, Kürzinger ML, Lerch M, Maskell S, Norén GN, et al. Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project. Drug Saf 2018 Dec;41(12):1355-1369 [FREE Full text] [doi: 10.1007/s40264-018-0699-2] [Medline: 30043385]
11. Bousquet C, Audeh B, Bellet F, Lillo-Le Louët A. Comment on "Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project". Drug Saf 2018 Dec 19;41(12):1371-1373. [doi: 10.1007/s40264-018-0747-y] [Medline: 30341678]
12. Karapetiantz P, Bellet F, Audeh B, Lardon J, Leprovost D, Aboukhamis R, et al. Descriptions of adverse drug reactions are less informative in forums than in the French pharmacovigilance database but provide more unexpected reactions. Front Pharmacol 2018 May 1;9:439-411 [FREE Full text] [doi: 10.3389/fphar.2018.00439] [Medline: 29765326]
13. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: a scoping review. J Med Internet Res 2015 Jul 10;17(7):e171 [FREE Full text] [doi: 10.2196/jmir.4304] [Medline: 26163365]
14. Karapetiantz P, Audeh B, Faille J, Lillo-Le Louët A, Bousquet C. Qualitative and quantitative analysis of web forums for adverse events detection: "strontium ranelate" case study. Stud Health Technol Inform 2019 Aug 21;264:964-968. [doi: 10.3233/SHTI190367] [Medline: 31438067]
15. Casperson T, Painter J, Dietrich J. Strategies for distributed curation of social media data for safety and pharmacovigilance. 2016 Presented at: International Conference on Data Science (ICDATA); October 1, 2016; Barcelona, Spain p. 118-124.
16. Freifeld CC. Digital pharmacovigilance: The medwatcher system for monitoring adverse events through automated processing of internet social media and crowdsourcing. OpenBU Libraries. Boston University: OpenBU; 2014. URL: https://open.bu.edu/handle/2144/10995
17. Cossin S, Lebrun L, Lobre G, Loustau R, Jouhet V, Griffier R, et al. Romedi: an open data source about French drugs on the semantic web. Stud Health Technol Inform 2019 Aug 21;264:79-82. [doi: 10.3233/SHTI190187] [Medline: 31437889]
18. Abdellaoui R, Schück S, Texier N, Burgun A. Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help? JMIR Public Health Surveill 2017 Jun 22;3(2):e36 [FREE Full text] [doi: 10.2196/publichealth.6577] [Medline: 28642212]
19. Bousquet C, Dahamna B, Guillemin-Lanne S, Darmoni SJ, Faviez C, Huot C, et al. The adverse drug reactions from patient reports in social media project: five major challenges to overcome to operationalize analysis and efficiently support pharmacovigilance process. JMIR Res Protoc 2017 Sep 21;6(9):e179 [FREE Full text] [doi: 10.2196/resprot.6463] [Medline: 28935617]
20. Morlane-Hondère F, Grouin C, Zweigenbaum P. Identification of drug-related medical conditions in social media. 2016 May Presented at: The Tenth International Conference on Language Resources and Evaluation (LREC'16); May 2, 2016; Portoroz, Slovenia p. 2022-2028.
21. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. San Francisco, CA: Morgan Kaufmann Publishers; 2001 Jun 28 Presented at: Eighteenth International Conference on Machine Learning (ICML 2001); June 28, 2001 to July 1, 2001; Williamstown, MA p. 282-289.
22. Xia L. Historical profile will tell? A deep learning-based multi-level embedding framework for adverse drug event detection and extraction. Decision Support Systems 2022 Sep;160:113832. [doi: 10.1016/j.dss.2022.113832]
23. Yu D, Vydiswaran VGV. An assessment of mentions of adverse drug events on social media with natural language processing: model development and analysis. JMIR Med Inform 2022 Sep 28;10(9):e38140 [FREE Full text] [doi: 10.2196/38140] [Medline: 36170004]
24. Rezaei Z, Ebrahimpour-Komleh H, Eslami B, Chavoshinejad R, Totonchi M. Adverse drug reaction detection in social media by deep learning methods. Cell J 2020 Oct;22(3):319-324 [FREE Full text] [doi: 10.22074/cellj.2020.6615] [Medline: 31863657]
25. Hussain S, Afzal H, Saeed R, Iltaf N, Umair MY. Pharmacovigilance with transformers: a framework to detect adverse drug reactions using BERT fine-tuned with farm. Comput Math Methods Med 2021;2021:5589829 [FREE Full text] [doi: 10.1155/2021/5589829] [Medline: 34422092]
26. Ding P, Zhou X, Zhang X, Wang J, Lei Z. An attentive neural sequence labeling model for adverse drug reactions mentions extraction. IEEE Access 2018;6:73305-73315. [doi: 10.1109/access.2018.2882443]
27. Xu Z, Kass-Hout T, Anderson-Smits C, Gray G. Signal detection using change point analysis in postmarket surveillance. Pharmacoepidemiol Drug Saf 2015 Jun 22;24(6):663-668 [FREE Full text] [doi: 10.1002/pds.3783] [Medline: 25903221]
28. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001 Dec 10;10(6):483-486. [doi: 10.1002/pds.677] [Medline: 11828828]
29. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002 Feb 06;11(1):3-10. [doi: 10.1002/pds.668] [Medline: 11998548]
30. Ahmed I, Pariente A, Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. Stat Methods Med Res 2018 Mar 25;27(3):785-797. [doi: 10.1177/0962280216643116] [Medline: 27114328]
31. Caster O, Norén GN, Madigan D, Bate A. Large‐scale regression‐based pattern discovery: the example of screening the WHO global drug safety database. Statistical Analysis 2010 Jul 20;3(4):197-208. [doi: 10.1002/sam.10078]

XSL•FO
RenderX

32. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther 2013 Jun 11;93(6):539-546 [FREE Full text] [doi: 10.1038/clpt.2013.24] [Medline: 23571771]
33. Team R. The R Project for Statistical Computing. R Foundation. URL: http://www.R-project.org/ [accessed 2024-04-26]
34. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y. shiny: web application framework for R. Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=shiny [accessed 2023-01-30]
35. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. URL: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016R0679 [accessed 2024-04-26]
36. SHA-1. Wikipedia. 2023. URL: https://en.wikipedia.org/w/index.php?title=SHA-1&oldid=1135933131 [accessed 2023-01-30]
37. Viard D, Parassol-Girard N, Romani S, Van Obberghen E, Rocher F, Berriri S, et al. Spontaneous adverse event notifications by patients subsequent to the marketing of a new formulation of Levothyrox amidst a drug media crisis: atypical profile as compared with other drugs. Fundam Clin Pharmacol 2019 Aug 07;33(4):463-470. [doi: 10.1111/fcp.12446] [Medline: 30575110]
38. Audeh B, Grouin C, Zweigenbaum P, Bousquet C, Jaulent M, Benkhebil M. French Levothyrox® crisis: retrospective analysis of social media. Bogota, Colombia: Springer International Publishing; 2019 Oct 26 Presented at: Conference ISOP - International Society of Pharmacovigilance; October 1, 2019; Bogota, Colombie URL: https://hal.archives-ouvertes.fr/hal-02411632
39. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019. 2019 Jun 07 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf
40. Campillos-Llanos L, Grouin C, Lillo-Le Louët A, Zweigenbaum P. Initial experiments for pharmacovigilance analysis in social media using summaries of product characteristics. Stud Health Technol Inform 2019 Aug 21;264:60-64. [doi: 10.3233/SHTI190183] [Medline: 31437885]
41. Vigi4Med Scraper. GitHub. URL: https://github.com/bissana/Vigi4Med-Scraper [accessed 2024-04-22]

Abbreviations

ADE: adverse drug event
ANSM: Agence nationale de sécurité du médicament et des produits de santé
ATC: Anatomical Therapeutic Classification
BERT: Bidirectional Encoder Representations from Transformer
CSV: comma-separated values
CUSUM: Cumulative Sum
EMA: European Medicines Agency
FDA: Food and Drug Administration
FPVD: French Pharmacovigilance Database
GDPR: General Data Protection Regulation
HAS: French National Health Authority
MedDRA: Medical Dictionary for Regulatory Activities Terminology
NLP: natural language processing
PHARES: Pharmacovigilance in Social Networks
PREDOSE: Prescription Drug Abuse Online Surveillance and Epidemiology
PRR: proportional reporting ratio
PT: preferred term
RDF: resource description framework
ROR: reporting odds ratio
WEB-RADR: Recognizing Adverse Drug Reactions

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.02.23; peer-reviewed by S Matsuda, L Shang; comments to author 06.07.23; revised version received 20.10.23; accepted 12.03.24; published 18.06.24.

Please cite as:
Karapetiantz P, Audeh B, Redjdal A, Tiffet T, Bousquet C, Jaulent MC
Monitoring Adverse Drug Events in Web Forums: Evaluation of a Pipeline and Use Case Study
J Med Internet Res 2024;26:e46176
URL: https://www.jmir.org/2024/1/e46176
doi:10.2196/46176
PMID:38888956

©Pierre Karapetiantz, Bissan Audeh, Akram Redjdal, Théophile Tiffet, Cédric Bousquet, Marie-Christine Jaulent. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Nutrients and Foods Recommended for Blood Pressure Control on Twitter in Japan: Content Analysis

Marina Terada[1], MPH; Tsuyoshi Okuhara[2], MPH, PhD; Rie Yokota[3], MPH, PhD; Takahiro Kiuchi[2], MD, PhD; Kentaro Murakami[4], PhD

[1]Department of Health Communication, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[2]Department of Health Communication, School of Public Health, Graduate School of Medicine, The University of Tokyo, Japan
[3]Department of Medical Communication, School of Pharmacy and Pharmaceutical Sciences, Hoshi University, Tokyo, Japan
[4]Department of Social and Preventive Epidemiology, School of Public Health, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Marina Terada, MPH
Department of Health Communication
Graduate School of Medicine
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 358006549
Fax: 81 356890726
Email: marimomo@g.ecc.u-tokyo.ac.jp

Abstract

Background: Management and prevention of hypertension are important public health issues. Healthy dietary habits are one of the modifiable factors. As Twitter (subsequently rebranded X) is a digital platform that can influence public eating behavior, there is a knowledge gap regarding the information about foods and nutrients recommended for blood pressure control and who disseminates them on Twitter.

Objective: This study aimed to investigate the nature of the information people are exposed to on Twitter regarding nutrients and foods for blood pressure control.

Methods: A total of 147,898 Japanese tweets were extracted from January 1, 2022, to December 31, 2022. The final sample of 2347 tweets with at least 1 retweet was manually coded into categories of food groups, nutrients, user characteristics, and themes. The number and percentage of tweets, retweets, and themes in each category were calculated.

Results: Of the 2347 tweets, 80% (n=1877) of tweets mentioned foods, which were categorized into 17 different food groups. Seasonings and spices, including salt, were most frequently mentioned (1356/1877, 72.2%). This was followed by vegetable and fruit groups. The 15 kinds of nutrients were mentioned in 1566 tweets, with sodium being the largest proportion at 83.1% (n=1301), followed by potassium at 8.4% (n=132). There was misinformation regarding salt intake for hypertension, accounting for 40.8% (n=531) of tweets referring to salt, including recommendations for salt intake to lower blood pressure. In total, 75% (n=21) of tweets from “doctors” mentioned salt reduction is effective for hypertension control, while 31.1% (n=74) of tweets from “health, losing weight, and beauty-related users,” 25.9% (n=429) of tweets from “general public,” and 23.5% (n=4) tweets from “dietitian or registered dietitian” denied salt reduction for hypertension. The antisalt reduction tweets accounted for 31.5% (n=106) of the most disseminated tweets related to nutrients and foods for blood pressure control.

Conclusions: The large number of tweets in this study indicates a high interest in nutrients and foods for blood pressure control. Misinformation asserting antisalt reduction was posted primarily by the general public and self-proclaimed health experts. The number of tweets from nutritionists, registered dietitians, and doctors who were expected to correct misinformation and promote salt reduction was relatively low, and their messages were not always positive toward salt reduction. There is a need for communication strategies to combat misinformation, promote correct information on salt reduction, and train health care professionals to effectively communicate evidence-based information on this topic.

(J Med Internet Res 2024;26:e49077) doi:10.2196/49077

XSL•FO
RenderX

KEYWORDS

Twitter; food; nutrition; misinformation; salt; content analysis; hypertension; blood pressure; sodium; salt reduction

Introduction

Hypertension is a modifiable risk factor for cardiovascular disease and premature death [1,2]. In 2019, the global age-standardized prevalence of hypertension in adults aged 30-79 years was reported to be 32% in women and 34% in men [3]. In Japan, the prevalence of hypertension in adults aged 20 years or older was 24.9% in women and 29.9% in men in 2019 [4]. As more than 20% of adults are estimated to have hypertension, its prevention and management remain important public health issues [3].

One of the modifiable preventive factors for hypertension is healthy dietary habits such as reduced salt intake and increased potassium intake [1,5], in addition to other factors such as engaging in moderate physical activity and maintaining normal body weight [5]. The National High Blood Pressure Education Program [5] and the Japanese Society of Hypertension [6] have established dietary and nutrient intake guidelines to prevent and manage hypertension. However, the amount of salt and vegetable intake recommended for blood pressure reduction in Japanese adults has not been achieved in adults aged 20 years or older in 2019 [4].

Our previous study found that “nutrient and hypertension” is one of the most frequently searched keywords in Google among information related to nutrients from 2017 to 2022 [7], indicating high public interest in digital information on diet and hypertension. Furthermore, in a Japanese national survey, between 15.4% and 39.3% of women aged 20-49 years and 10.8%-17.6% of men in the same age group reported that social media was a source of information that influenced their eating behavior in 2019 [4]. Social media is a platform for information on which people post food and diets [8]. Specifically, Japan ranks as the second largest active Twitter (subsequently rebranded X) user, with 69 million users as of 2024 [9]. Twitter is an important communication channel, as more than 40% of the Japanese population uses Twitter [10], and 38.2% of individuals in Japan with hypertension or high blood pressure engage with social media in 2020 [11]. Therefore, we speculate that Twitter users are exposed to information that may influence their eating behaviors.

In a study assessing 550,338 English tweets related to cardiovascular diseases from 2009 to 2015, there were 23,459 tweets regarding hypertension [12], and 43% of 500 manually coded tweets were about treatment and control of hypertension, including diet [12]. A random sample of 1251 tweets from 2009 to 2015 regarding cardiovascular disease also reported that 8% of the tweets mentioned prevention including dietary advice for blood pressure and diabetes control [13]. Studies analyzing information related to hypertension on social media platforms such as Twitter and Facebook reported that common themes included “awareness,” “risk factors,” and “treatment and management” [12-14]. However, it is unclear what types of nutrients and foods are recommended, particularly on Twitter, for the prevention and control of hypertension. A content analysis of 298 tweets related to heart failure and nutrients found tweets mentioning general nutrients, sodium restriction, energy needs, dietary supplements, alcohol, and proteins for heart failure [15]. This study only collected tweets that contained the “#heartfailure” hashtag, which resulted in a limited number of tweets and did not provide a comprehensive understanding of tweets related to nutrients and foods for blood pressure control [15]. Furthermore, these studies did not investigate the user characteristics that share information about nutrients and foods for managing hypertension.

Content related to hypertension, nutrients, and food is highly prevalent on Twitter, indicating that this topic interests the public. To our knowledge, no study has analyzed the content of tweets related to nutrients and foods for blood pressure control on Twitter. Therefore, this study aimed to investigate the nature of the information people are exposed to on Twitter regarding nutrients and foods for blood pressure control. Furthermore, we discuss the potential impact of this information on Twitter users’ dietary behaviors. The research questions (RQs) are as follows:

- RQ1: What types of nutrients and foods are recommended on Twitter for blood pressure control?
- RQ2: Do the content of tweets related to blood pressure control, nutrients, and food vary according to user characteristics?
- RQ3: What types of foods, nutrients, and themes are commonly featured in the highly retweeted tweets related to blood pressure control, nutrients, and foods?

Methods

Data Extraction

We extracted 147,898 Japanese tweets posted from January 1, 2022, to December 31, 2022, matching keywords related to hypertension, nutrients, and food consumption. The keywords used in the search are listed in Multimedia Appendix 1, as we used complex Japanese terms. The keywords were selected based on previous hypertension studies on Facebook [14] and a content analysis of tweets related to heart failure and nutrients on Twitter [15]. The data were obtained from Social Insight [16], a subscription Twitter aggregation service that allows the purchase of tweets and their associated data, such as the number of retweets likes, and followers. We obtained tweet texts, user descriptions, usernames, number of retweets, likes, followers, and posted times. As Figure 1 shows, retweets were excluded from 147,898 tweets, leaving 57,635 eligible for analysis. The Pearson correlation coefficient (r) between the daily number of tweets and the daily number of retweets was calculated to be 0.28 (95% CI 0.19-0.37), indicating a low correlation between the number of tweets and their diffusion to the public, as measured by retweet counts. Accordingly, we assumed that sampling based on tweet counts for a certain period might not capture tweets with a high diffusion index, as measured by retweet counts. Therefore, to analyze the tweets more likely to have been exposed to a larger audience, we conducted purposive

XSL•FO
RenderX

sampling by focusing on tweets with high retweet counts in our study sample. After calculating the retweet counts of the 57,635 tweets, we found that the lower quartile value for retweet was 0 up to the 75th percentile. Limiting the analysis to tweets with at least 1 retweet (ie, retweet counts above the 75th percentile), we extracted 4068 tweets. Next, we manually screened 4068 tweets based on the following inclusion and exclusion criteria: the inclusion criteria were tweets recommending specific nutrients or foods for blood pressure control. The exclusion criteria were tweets unrelated to hypertension, nutrients, or foods, or tweets that did not directly recommend nutrients or foods for blood pressure control (ie, personal food records). We included 2347 tweets in the final analysis.

Figure 1. Flow diagram for inclusion and exclusion of tweets posted from January 1, 2022, to December 31, 2022.

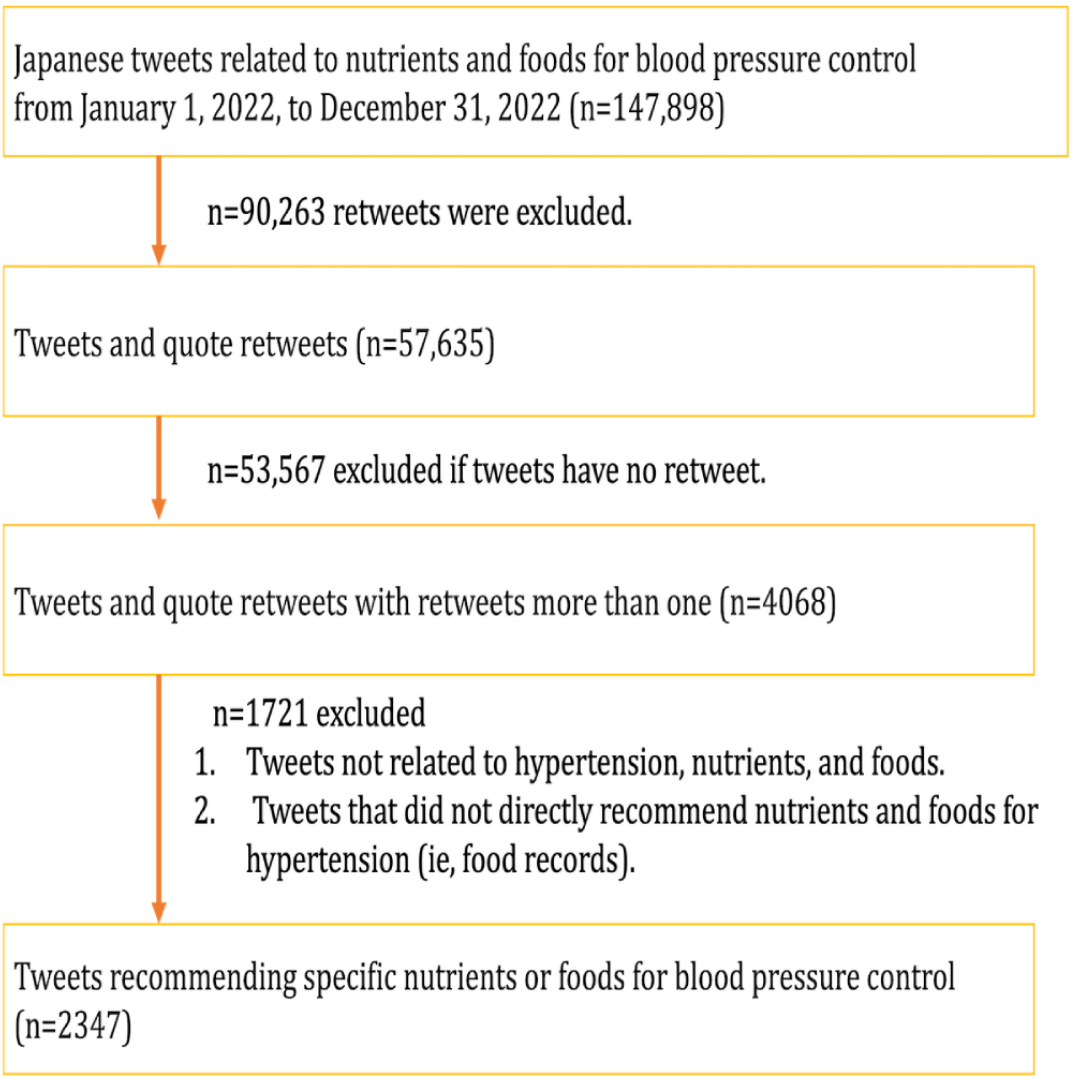

Coding Procedures

Procedure and Data Analysis

The first author (MT) screened 4068 tweets with the aforementioned exclusion criteria and created a coding manual. Subsequently, 2347 tweets were deemed eligible for final coding. Then, MT conducted a training session using a manual with the third author (RY). MT and RY coded 469 tweets (20% of the final sample of 2347 tweets) that were randomly sampled to assess interrater reliability for all categories. If there were discrepancies between the coders, a second author (TO) joined the discussion until a consensus was reached. Excel (version 16.70; Microsoft Corp) was used for coding.

User Characteristics

Based on the usernames and self-descriptions, we coded users as “health, losing weight, and beauty-related users,” “food manufacturer,” “IT company and mass media,” “government and academic institution,” “medical institution,” and “pharmaceutical manufacturer” based on our previous study [7]. Subsequently, we inductively added user characteristics [17]. If a user matched 2 or more attributes, they were categorized into all applicable categories (eg, 1 user could be a doctor and food and nutrition-related users). Users not matching all the listed attributes were classified as “general public.”

Themes of Tweets

Tweets containing recommendations for foods that can help control blood pressure were categorized into food groups using the Food Composition Database provided by the Japanese

XSL•FO
RenderX

Ministry of Education, Culture, Sports, Science, and Technology [18]. In a previous study, there have been reports suggesting the existence of misleading information claiming that salt reduction has no effect on blood pressure management or advocating alternative treatments not recommended by current guidelines [19]. Building upon these findings, this study also categorized misleading information that denies the relationship of salt reduction for hypertension or advocating alternative measures not recommended by current guidelines. The nutrients were added inductively. Each food or nutrient mentioned in the tweets was added to its respective category. Because this study aimed to analyze the content of the tweets, we did not code the web pages linked to the URLs in the tweets. Images (eg, photographs) containing information about foods or nutrients recommended for blood pressure control were considered for analysis. As an indicator of credibility, we coded the presence or absence of citations to support the information provided in the tweet. To code the citation sources, we adopted a deductive approach for scientific papers, books, dietary reference intakes, the Ministry of Health, Labour and Welfare, the Japanese Society of Hypertension, the Japanese Circulation Society, and the Dietary Approaches to Stop Hypertension, as described in our previous study [7]. We also used an inductive approach to include other sources, such as specific health food products approved by the Consumer Affairs Agency in Japan and news articles. We classified tweets as advertisements if they promoted a specific brand, product, or store for advertising purposes, such as recommending a particular brand or product for blood pressure control.

Statistical Analysis

The number and percentage of tweets, retweets, and tweets in each category were calculated. To evaluate the characteristics of the highly diffused tweets, they were classified into 4 quartiles based on the number of retweets. The number and percentage of tweets in each category are shown. The interrater reliability coefficient was calculated using the Gwet AC1 statistic, which is more consistent with a lower prevalence than Cohen κ statistics [20-22]. The analysis was conducted using R for macOS (version 4.2.0; The R Foundation).

Ethical Considerations

This study was approved by the ethics committee of the Graduate School of Medicine at the University of Tokyo (2022288NI). Inform consent could not be obtained due to the nature of the data. This report does not contain any identifying information or direct quotes in accordance with published guidelines and recommendations to support anonymity [23].

Results

Sample Description

The final sample consisted of 2347 Japanese tweets that met the inclusion criteria, posted from January 1, 2022, to December 31, 2022, with more than 1 retweet. The number of unique users was 1497 after removing duplicate accounts.

Nutrients and Foods Recommended for Blood Pressure Control

Table 1 shows the types of food groups and nutrients recommended for blood pressure control in the tweets. The interrater reliability was from 0.80 (95% CI 0.75-0.86) to 1.00 (95% CI 1.00-1.00) for these categories. In the analysis of 2347 tweets, 80% (n=1877) of tweets mentioned foods, which were categorized into 17 different food groups. Among food groups, seasonings and spices accounted for the largest proportion of tweets at 72.2% (n=1356), of which, tweets mentioning salt accounted for a large proportion at 69.3% (n=1301), representing 55.4% (n=1301) of the total tweets analyzed. A food group of vegetables accounted for 13.6% (n=256), followed by fruits at 4.4% (n=83) and seafood at 2.9% (n=55). Of 2347 tweets analyzed, 1566 (66.7%) mentioned nutrients, specifically 15 different types of nutrients recommended for blood pressure control. Sodium was the most frequently discussed nutrient at 83.1% (n=1301), followed by potassium at 8.4% (n=132) and minerals at 4.7% (n=73). Among the nutrients mentioned, 1.3% (n=21) referenced dietary supplements, with 71.4% (n=15) of these mentions specifically relating to γ-aminobutyric acid (Multimedia Appendix 2).

XSL•FO
RenderX

Table 1. The types of foods and nutrients recommended for blood pressure control in tweets posted from January 1, 2022, to December 31, 2022 (N=2347)[a].

	Values, n (%)
Food groups (n=1877)	
Seasonings and spices	1356 (72.2)
Vegetables	256 (13.6)
Fruits	83 (4.4)
Fish and shellfish	55 (2.9)
Beverages	43 (2.3)
Cereals	40 (2.1)
Potatoes	29 (1.5)
Pulses	28 (1.5)
Dairy products	26 (1.4)
Nuts and seeds	24 (1.3)
Mushrooms	13 (0.7)
Algae	12 (0.6)
Meat	9 (0.5)
Confectioneries	8 (0.4)
Fats and oils	3 (0.2)
Sugars and sweeteners	3 (0.2)
Ready-to-eat foods	119 (6.3)
Nutrients (n=1566)	
Sodium[b]	1301 (83.1)
Potassium	132 (8.4)
Minerals	73 (4.7)
Protein	61 (3.9)
Dietary fiber	51 (3.3)
Magnesium	49 (3.1)
GABA[c]	28 (1.8)
Citric acid	25 (1.6)
Sodium bicarbonate	22 (1.4)
Calcium	22 (1.4)
EPA[d]	16 (1)
Omega-3 fatty acids	14 (0.9)
DHA[e]	12 (0.8)
Iron	12 (0.8)
Polyphenols	10 (0.6)
Others (eg, zinc and linoleic acid)	11 (0.7)

[a]A single tweet can be classified into multiple foods and nutrients, and the total number of tweets may exceed the total number of unique tweets. Food group of eggs was excluded, as they were not recommended for blood pressure control. Carbohydrates were mentioned in 66 tweets but were excluded because they were not recommended for blood pressure control.

[b]Salt was included with sodium as it is commonly tweeted about inseparably and interchangeably.

[c]GABA: γ-aminobutyric acid.

[d]EPA: eicosatetraenoic acid.

[e]DHA: docosahexaenoic acid.

XSL•FO
RenderX

Themes of Tweets and User Characteristics

Table 2 presents the proportion of tweets related to each theme based on user characteristics, definitions, and the total number of tweets associated with each user characteristic. The interrater reliability was from 0.84 (95% CI 0.79-0.88) to 1.00 (95% CI 1.00-1.00) for these categories. The users were classified into 12 categories. Except for the general public at 70.6% (n=1658), advertisement users, such as e-commerce users, accounted for the largest proportion of tweets at 10.3% (n=242), followed by "health, losing weight, and beauty-related users" at 10.1% (n=238), "food and nutrition-related users" at 6.9% (n=163), 69%-93.6% of users recommended any food for blood pressure control.

Table 2. The user characteristics, the number of tweets and themes, posted from January 1, 2022, to December 31, 2022.

User characteristics	Definition	Tweets (n=2347), n (%)[a,b]	Recommendation for blood pressure control		Themes of tweets			
			Foods (n=1877), n (%)[c]	Nutrients (n=1566), n (%)[c]	Prosalt reduction (n=770), n (%)[c]	Antisalt reduction (n=531), n (%)[c]	Advertisement (n=242), n (%)[c]	Reference (n=122), n (%)[c]
Nutritionist or registered dietitian	Nutritionist or registered dietitian	17 (0.7)	15 (88.2)	12 (70.6)	5 (29.4)	4 (23.5)	1 (5.9)	2 (11.8)
Doctors	Doctors	28 (1.2)	23 (82.1)	24 (85.7)	21 (75)	2 (7.1)	2 (7.1)	8 (28.6)
Other medical professionals	Veterinarian, dentist, dental hygienist, nurse, midwife, public health nurse, pharmacist, clinical psychologist, caregiver welfare worker, acupuncturist, and Chinese herbalist	58 (2.5)	40 (69)	33 (56.9)	11 (19)	8 (13.8)	2 (3.4)	1 (1.7)
Health, losing weight, and beauty-related users	Users who clearly indicate that they tweet about health, hypertension, losing weight, and beauty	238 (10.1)	196 (82.4)	150 (63.0)	26 (10.9)	74 (31.1)	26 (10.9)	8 (3.4)
Food and nutrition-related users	Users who clearly indicate that they tweet about cooking, meals, and nutrition or those who described themselves as food coordinator and supplement adviser	163 (6.9)	135 (82.8)	77 (47.2)	30 (18.4)	12 (7.4)	25 (15.3)	16 (9.8)
Food manufacturer	Food manufacturers such as soba food companies	47 (2)	44 (93.6)	28 (59.6)	22 (46.8)	1 (2.1)	19 (40.4)	5 (10.6)
IT company and mass media	Official news media and digital news media	42 (1.8)	31 (73.8)	24 (57.1)	20 (47.6)	2 (4.8)	1 (2.4)	7 (16.7)
Government, organizations, and university institutions	Government, organizations, and university institutions	31 (1.3)	28 (90.3)	26 (83.9)	21 (67.7)	3 (9.7)	2 (6.5)	0 (0)
Medical institutions	Hospitals and clinics	15 (0.6)	12 (80)	12 (80)	10 (66.7)	1 (6.7)	0 (0)	6 (40)
Pharmaceutical manufacturer	Pharmaceutical manufacturers	40 (1.7)	35 (87.5)	25 (62.5)	9 (22.5)	4 (10)	2 (5)	2 (5)
Advertisement	Digital shopping companies and other accounts that promote their products for advertisement purposes	242 (10.3)	219 (90.5)	163 (67.4)	118 (48.8)	8 (3.3)	91 (37.6)	20 (8.3)
General public	Users could not be classified into any of the above	1658 (70.6)	1299 (78.3)	1115 (67.2)	531 (32)	429 (25.9)	276 (16.6)	67 (4)

[a]An account can be classified into multiple user characteristics, and the total number of tweets in each column may exceed the total number of unique tweets.

[b]The percentage in the tweet count column is based on 2347 tweets and shows the distribution of each user attribute in the total number of tweets.

[c]Percentage is based on the total tweet count for each user and shows the distribution of themes for each user.

Of the total tweets analyzed, 1301 (55.4%) referred to salt. These tweets included 2 opposing views on reducing salt intake in hypertension control. The "prosalt reduction theme" recommended reducing salt intake for hypertension control, while the "antisalt reduction theme" claimed that reducing salt intake was not related to higher blood pressure or that consuming salt was beneficial for lowering blood pressure. Among tweets referred salt, 59.2% (n=770) was the "prosalt reduction theme" and 40.8% (n=531) was the "antisalt reduction theme." Regarding the relationship between user characteristics and the theme of salt reduction, 75% (n=21) of tweets from "doctors" and 66.7% (n=10) from "medical institution"

XSL•FO
RenderX

mentioned salt reduction is effective for hypertension control. In contrast, 31.1% (n=74) of tweets from "health, losing weight, and beauty-related users," 25.9% (n=429) of tweets from "general public," and 23.5% (n=4) tweets from "dietitian or registered dietitian" denied salt reduction for hypertension control.

Regarding advertising tweets, 40.4% (n=19) of "food manufacturer," 37.6% (n=91) of "advertisements users," 16.6% (n=276) of "general public," and 15.3% (n=25) of "food and nutrition-related users" had posted tweets for advertising purposes. Regarding the inclusion of citations, 40% (n=6) of medical institutions and 28.6% (n=8) of doctors included references in their tweets.

Number of Retweets and Themes

The median number of retweets was 2.0 (IQR 1.0-7.0), with a maximum of 7292. Among the 2347 tweets, the 0th percentile and 25th percentile of retweets were both equal to 1. Consequently, tweets with only 1 retweet were excluded to categorize tweet counts based on the quartiles of retweet numbers, as shown in Table 3. Among tweets with more than 2 retweets (n=1372), there were 530 tweets in the first quartile (1Q), 184 tweets in the second quartile (2IQR), 321 tweets in the third quartile (3IQR), and 337 tweets in the fourth quartile (4IQR). Among tweets in the lowest 1Q of retweet numbers, the percentage of "prosalt reduction theme" for hypertension was 31.7% (n=168); however, this percentage decreased to 24.9% (n=84) among tweets in the highest 4IQR of retweet numbers. Conversely, among tweets in the lowest 1Q of retweet numbers, the percentage of "antisalt reduction theme" for hypertension was 22.6% (n=120); this percentage increased to 31.5% (n=106) among tweets in the highest 4IQR of retweet numbers.

Table 3. The distribution of each category in tweets classified by retweet quartiles posted from January 1, 2022, to December 31, 2022 (n=1372)[a].

Retweet	Median (IQR)	Range	Unique tweets, n	Recommendation for blood pressure control		Themes of tweets			
				Foods, n (%)	Nutrients, n (%)	Prosalt reduction, n (%)	Antisalt reduction, n (%)	Advertisement, n (%)	Reference, n (%)
1Q	2.0 (2.0-3.0)	2.0-3.0	530	418 (78.9)	342 (64.5)	168 (31.7)	120 (22.6)	55 (10.4)	31 (5.8)
2IQR	4.0 (4.0-5.0)	4.0-5.0	184	147 (79.9)	120 (65.2)	57 (31)	47 (25.5)	12 (6.5)	10 (5.4)
3IQR	9.0 (7.0-12.0)	6.0-17.0	321	237 (73.8)	186 (57.9)	77 (24)	77 (24)	30 (9.3)	27 (8.4)
4IQR	57.0 (28.0-166.0)	18.0-7292.0	337	288 (85.5)	238 (70.6)	84 (24.9)	106 (31.5)	22 (6.5)	16 (4.7)

[a]Because of the possibility of each tweet being classified into more than 1 category, the total may not add up to the total number of tweets analyzed. The percentage denominator represents the unique number of tweets in each quartile category and shows the distribution of themes in each quartile.

Multimedia Appendix 3 shows the tweet content of the top 15 tweets with the most retweets among the total sample of 2347 tweets. The first 6 tweets were disseminated by users categorized as food company and advertisement user, promoting their products for blood pressure control. The 8th, 12th, and 15th tweets contained antisalt reduction content, endorsing natural salt or asserting that increased salt intake is beneficial for lowering blood pressure. The 9th and 10th tweets attempted to debunk such antisalt reduction content. Other tweets in Multimedia Appendix 3 promoted various products or specific vegetables.

Discussion

Principal Results

This study obtained 147,898 tweets that matched keywords related to hypertension, nutrients, and foods, including retweets, posted on Twitter in 2022. Of these, 2347 tweets were analyzed after purposive sampling, and the recommended nutrients and foods for blood pressure control, themes, and user characteristics were identified. The large number of tweets in this study indicates a high level of interest in nutrients and foods for blood pressure control. However, in this study, 40.8% (n=531) of tweets asserted that salt reduction has no effect on blood pressure management or endorsed alternative treatments not recommended by current guidelines. Moreover, more than 30% (106/337) of the most disseminated tweets were antisalt reduction, suggesting the spread of misinformation of salt for blood pressure control.

Nutrients and Foods Recommended for Blood Pressure Control

In total, 80% (n=1877) of all tweets mentioned foods for hypertension control. The most mentioned food groups were seasonings and spices, followed by vegetables, fruit, and seafood. Salt was the most frequently mentioned item in both nutrients and food groups, accounting for more than half of all tweets in the analysis. These results are consistent with scientifically recommended food groups for preventing and controlling hypertension [5,6,24]. Potassium is the second frequently mentioned nutrient and is consistent with the recommended nutrients for blood pressure control [5]. There is less mention of recommended nutrients for blood pressure control, such as reducing saturated and total fats [5,6,25] than minerals, which were not specified in the tweets.

Themes of Tweets and User Characteristics

The majority of tweets (n=1658, 70.6%) were from the "general public," which could not be classified, showing the anonymity

XSL•FO
RenderX

of Twitter. This was followed by "advertisement users," "health, losing weight, and beauty-related users," and "food and nutrition-related users." This study found that users who identified themselves as health gurus or health or diet experts, as previously noted in other studies [26], were among the most frequent sources of highly disseminated tweets regarding food and nutrients for blood pressure control in addition to general public. In other words, people are exposed to information about nutrients and foods for blood pressure control on Twitter primarily from users who identify themselves as health gurus or health or diet experts rather than from medical professionals, registered nutritionists, or medical institutions. Furthermore, more than 30% (n=74) of "health, losing weight, and beauty-related users" promoted increased salt intake for lowering blood pressure. A reduction in dietary salt intake reduces blood pressure [27,28], which is dose-dependent and has a linear relationship rather than a J-shaped one [29]. However, this study found that self-proclaimed experts on Twitter [26] spread misleading information on nutrients and foods for blood pressure control. In recent years, several studies suggesting that reducing salt intake is ineffective in treating hypertension have been published in a single journal. Several biases have been pointed out, such as funding from the food industry, raising concerns about the reliability, and ethics of science [30]. This study further suggests that misinformation about salt reduction for hypertension is disseminated on social media platforms such as Twitter. Additionally, although registered dietitians and nutritionists are expected to combat misinformation about nutrition [26], in this study, the number of tweets from registered dietitians or nutritionists was only 0.7% (n=17). Furthermore, 23.5% (n=4) of them disseminated misleading information that denied an association between excessive salt intake and hypertension, contradicting the guidelines for hypertension and diet [5,6]. Conversely, 75% (n=21) of doctors recommended reducing salt intake, but the number of tweets from doctors was only up to 1.2% (n=28); it is difficult to evaluate whether they could combat misinformation about the controversial discussion about salt reduction. Cappuccio et al [30] have provided a matrix of facts that refute myths skeptical of reducing salt intake, such as "our body needs salt and restriction is a mistake" or "natural salt is healthier than refined salt," which aligns with the content in this study. Therefore, nutritionists, registered dietitians, and doctors should use this matrix to combat misinformation and proactively provide accurate information on salt intake and hypertension control.

Although "medical institutions" and "doctors" cited references frequently, only 5.2% (n=122) of tweets overall included citations. This may be because of the limit of 140 Japanese characters per tweet. Previous research showed that at least 40% of nutrition-related websites included 1 or more citations [7], suggesting that the low citations on Twitter make it difficult to select reliable information for lay individuals.

Number of Retweets and Themes

This study found that antisalt reduction tweets accounted for 31.5% (n=106) of the most disseminated tweets related to nutrients and food for blood pressure control, whereas only 24.9% (n=84) were prosalt reduction tweets. This result is consistent with previous studies showing that fake news is more likely to spread than true news [31], and misleading videos on hypertension have been found to have high user engagement [19]. However, studies examining the credibility of tweets have reported that the more retweets a tweet has, the lower the reader's trust in it [32-34]. Therefore, there is a possibility that the credibility of antisalt reduction with a high number of retweet is low, although these studies did not evaluate the credibility of tweets related to nutrition and hypertension [35]; it is not clear how the number of retweet for nutrition-related tweets is related to their credibility by Twitter users, which requires further investigation. Furthermore, while retweets are commonly used as a measure of diffusion [31,36], they also serve as a means to "argue" against the content being retweeted [37,38]. Hence, it is crucial to understand that retweets against antisalt reduction might also involve sharing content to counteract it. The dissemination of misinformation through tweets with high retweet numbers and greater exposure is a cause for concern. Evaluating the details of misinformation regarding salt reduction on Twitter and its potential impact on public dietary habits is necessary.

Limitations

This study had some limitations. First, user characteristics were obtained from the self-description profile and usernames and may not represent the actual status. Next, blood pressure increases with age [39,40], a trend that may not align with the predominantly young age group using Twitter [10]. However, hypertension should be addressed through a life-course approach that includes younger age groups [41]. Therefore, analyzing and evaluating information related to nutrients and foods as modifiable factors of hypertension on Twitter, which is a widely used information channel in Japan for young and older people, are important. Third, this study used the number of retweets to indicate dissemination and influence without considering the number of followers or favorites. Fourth, because this study is a content analysis that enables the evaluation of communication symbols, we could not evaluate the influence of exposure to tweets on public perception and actual eating behavior. Finally, the tweets analyzed in this study were limited to Japanese; information from other languages was not investigated. Japanese cultural foods, such as miso, which are unique to Japan, were recommended for preventing hypertension in this study, indicating cultural characteristics. In future research, it will be necessary to conduct studies that include multiple languages and examine the kinds of information related to nutrients and foods recommended for hypertension control that are being disseminated in the context of different cultures.

Conclusions

This study identified the recommended nutrients and foods for blood pressure control and user characteristics on Twitter in 2022. The food groups and nutrients recommended in the tweets were consistent with the guidelines, but misinformation related to salt reduction accounted for more than 40.8% (n=531) of the tweets referred salt for blood pressure control. Most of the misinformation was tweeted by users including general public and self-proclaimed health experts. The number of tweets from nutritionists, registered dietitians, and doctors who were expected to correct misinformation and promote salt reduction

XSL•FO
RenderX

was relatively low, and their messages were not always positive toward salt reduction. There is a need for communication strategies to combat misinformation, promote correct information on salt reduction, and train health care professionals to effectively communicate evidence-based information on this topic.

Acknowledgments

This work was supported by a grant (22FA1022) from the Ministry of Health, Labour and 531 Welfare, Japan. The Ministry of Health, Labour and Welfare had no role in the design, analysis, or writing of this paper.

Data Availability

The data sets generated and analyzed during this study are not publicly available as they were purchased from the Social Insights. However, they are available from the corresponding author on reasonable request with permission from the Social Insights.

Authors' Contributions

MT contributed to the concept and design of the study and data collection and management, formulated the research, analyzed and interpreted the data, prepared the first draft of the manuscript, and had primary responsibility for the final content. TO contributed to the concept and design of the study and data collection and management and provided critical input into the final draft of the manuscript. RY contributed to coding the data and management and provided critical input into the final draft of the manuscript. TK provided critical input into the final draft of the manuscript. KM contributed to the concept and design of the study and data collection and management provided critical input into the final draft of the manuscript. All authors have read and agreed to the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The keywords selected for searching tweets related to nutrients and foods for blood pressure control.
[DOCX File , 19 KB - jmir_v26i1e49077_app1.docx]

Multimedia Appendix 2
The proportions of tweets about nutrients that include supplements posted from January 1, 2022, to December 31, 2022.
[DOCX File , 20 KB - jmir_v26i1e49077_app2.docx]

Multimedia Appendix 3
Top 15 tweets with the most retweets posted from January 1, 2022, to December 31, 2022.
[DOCX File , 19 KB - jmir_v26i1e49077_app3.docx]

References

1. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nat Rev Nephrol 2020;16(4):223-237 [FREE Full text] [doi: 10.1038/s41581-019-0244-2] [Medline: 32024986]
2. Roth GA, Nguyen G, Forouzanfar MH, Mokdad AH, Naghavi M, Murray CJL. Estimates of global and regional premature cardiovascular mortality in 2025. Circulation 2015;132(13):1270-1282. [doi: 10.1161/CIRCULATIONAHA.115.016021] [Medline: 26408271]
3. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet 2021;398(10304):P957-P980 [FREE Full text] [doi: 10.1016/S0140-6736(21)01330-1] [Medline: 34450083]
4. The National Health and Nutrition Survey (NHNS) Japan, 2019 Summary. The Ministry of Health, Labor and Welfare. 2019. URL: https://www.nibiohn.go.jp/eiken/kenkounippon21/download_files/eiyouchousa/2019.pdf [accessed 2023-04-10]
5. Whelton PK, He J, Appel LJ, Cutler JA, Havas S, Kotchen TA, et al. Primary prevention of hypertension: clinical and public health advisory from The National High Blood Pressure Education Program. JAMA 2002;288(15):1882-1888. [doi: 10.1001/jama.288.15.1882] [Medline: 12377087]
6. Guidelines for the management of hypertension 2019. The Japanese Society of Hypertension. 2019. URL: https://www.jpnsh.jp/data/jsh2019/JSH2019_noprint.pdf [accessed 2023-04-10]
7. Murakami K, Shinozaki N, Kimoto N, Onodera H, Oono F, McCaffrey TA, et al. Web-based content on diet and nutrition written in Japanese: infodemiology study based on Google Trends and Google Search. JMIR Form Res 2023;7:e47101 [FREE Full text] [doi: 10.2196/47101] [Medline: 37971794]

XSL•FO
RenderX

8. Lewis T, Phillipov M. Food/media: eating, cooking, and provisioning in a digital world. Commun Res Pract 2018;4(3):207-211. [doi: 10.1080/22041451.2018.1482075]
9. Leading countries based on number of X (formerly Twitter) users as of April 2024. statista. 2024. URL: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ [accessed 2024-06-05]
10. Survey report on usage time and information behavior of information and communication media. Ministry of Internal Affairs and Communications, Japan. 2020. URL: https://www.soumu.go.jp/main_content/000708015.pdf [accessed 2023-03-04]
11. Mitsutake S, Takahashi Y, Otsuki A, Umezawa J, Yaguchi-Saito A, Saito J, et al. Chronic diseases and sociodemographic characteristics associated with online health information seeking and using social networking sites: nationally representative cross-sectional survey in Japan. J Med Internet Res 2023;25:e44741 [FREE Full text] [doi: 10.2196/44741] [Medline: 36862482]
12. Sinnenberg L, DiSilvestro CL, Mancheno C, Dailey K, Tufts C, Buttenheim AM, et al. Twitter as a potential data source for cardiovascular disease research. JAMA Cardiol 2016;1(9):1032-1036 [FREE Full text] [doi: 10.1001/jamacardio.2016.3029] [Medline: 27680322]
13. Yang Q, Tufts C, Ungar L, Guntuku S, Merchant R. To retweet or not to retweet: understanding what features of cardiovascular tweets influence their retransmission. J Health Commun 2018;23(12):1026-1035 [FREE Full text] [doi: 10.1080/10810730.2018.1540671] [Medline: 30404564]
14. Al Mamun M, Ibrahim HM, Turin TC. Social media in communicating health information: an analysis of Facebook groups related to hypertension. Prev Chronic Dis 2015;12:140265 [FREE Full text] [doi: 10.5888/pcd12.140265] [Medline: 25633486]
15. Hand RK, Kenne D, Wolfram TM, Abram JK, Fleming M. Assessing the viability of social media for disseminating evidence-based nutrition practice guideline through content analysis of Twitter messages and health professional interviews: an observational study. J Med Internet Res 2016;18(11):e295 [FREE Full text] [doi: 10.2196/jmir.5811] [Medline: 27847349]
16. Social Insights. URL: https://sns.userlocal.jp/ [accessed 2023-03-07]
17. Kondracki NL, Wellman NS, Amundson DR. Content analysis: review of methods and their applications in nutrition education. J Nutr Educ Behav 2002;34(4):224-230. [doi: 10.1016/s1499-4046(06)60097-3] [Medline: 12217266]
18. Ministry of Education, Culture, Sports, Science and Technology. Food Composition Database. URL: https://fooddb.mext.go.jp/ [accessed 2023-03-07]
19. Kumar N, Pandey A, Venkatraman A, Garg N. Are video sharing web sites a useful source of information on hypertension? J Am Soc Hypertens 2014;8(7):481-490. [doi: 10.1016/j.jash.2014.05.001] [Medline: 25064770]
20. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013;13:61 [FREE Full text] [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
21. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
22. Nishiura H. A robust statistic AC_1 for assessing inter-observer agreement in reliability studies. J Jpn Soc Radiat Technol 2010;66(11):1485-1491 [FREE Full text] [doi: 10.6009/jjrt.66.1485] [Medline: 21099180]
23. Ford E, Shepherd S, Jones K, Hassan L. Toward an ethical framework for the text mining of social media for health research: a systematic review. Front Digit Health 2021;2:592237 [FREE Full text] [doi: 10.3389/fdgth.2020.592237] [Medline: 34713062]
24. Appel LJ, Brands MW, Daniels SR, Karanja N, Elmer PJ, Sacks FM, American Heart Association. Dietary approaches to prevent and treat hypertension: a scientific statement from the American Heart Association. Hypertension 2006;47(2):296-308. [doi: 10.1161/01.HYP.0000202568.01167.B6] [Medline: 16434724]
25. Appel LJ, Moore TJ, Obarzanek E, Vollmer WM, Svetkey LP, Sacks FM, et al. A clinical trial of the effects of dietary patterns on blood pressure. N Engl J Med 1997;336(16):1117-1124. [doi: 10.1056/NEJM199704173361601] [Medline: 9099655]
26. Diekman C, Ryan CD, Oliver TL. Misinformation and disinformation in food science and nutrition: impact on practice. J Nutr 2023;153(1):3-9 [FREE Full text] [doi: 10.1016/j.tjnut.2022.10.001] [Medline: 36913465]
27. Aburto NJ, Ziolkovska A, Hooper L, Elliott P, Cappuccio FP, Meerpohl JJ. Effect of lower sodium intake on health: systematic review and meta-analyses. BMJ 2013;346:f1326 [FREE Full text] [doi: 10.1136/bmj.f1326] [Medline: 23558163]
28. He FJ, MacGregor GA. Salt reduction lowers cardiovascular risk: meta-analysis of outcome trials. Lancet 2011;378(9789):380-382. [doi: 10.1016/S0140-6736(11)61174-4] [Medline: 21803192]
29. Filippini T, Malavolti M, Whelton PK, Naska A, Orsini N, Vinceti M. Blood pressure effects of sodium reduction: dose-response meta-analysis of experimental studies. Circulation 2021;143(16):1542-1567 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.050371] [Medline: 33586450]
30. Cappuccio FP, Campbell NRC, He FJ, Jacobson MF, MacGregor GA, Antman E, et al. Sodium and health: old myths and a controversy based on denial. Curr Nutr Rep 2022;11(2):172-184 [FREE Full text] [doi: 10.1007/s13668-021-00383-z] [Medline: 35165869]

XSL•FO
RenderX

31. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
32. Lin X, Spence PR. Identity on social networks as a cue: identity, retweets, and credibility. Commun Stud 2018;69(5):461-482. [doi: 10.1080/10510974.2018.1489295]
33. Lin X, Spence PR. Others share this message, so we can trust It? An examination of bandwagon cues on organizational trust in risk. Inf Process Manag 2019;56(4):1559-1564. [doi: 10.1016/j.ipm.2018.10.006]
34. Lin X, Spence PR, Lachlan KA. Social media and credibility indicators: the effect of influence cues. Comput Hum Behav 2016;63:264-271. [doi: 10.1016/j.chb.2016.05.002]
35. Jenkins EL, Ilicic J, Barklamb AM, McCaffrey TA. Assessing the credibility and authenticity of social media content for applications in health communication: scoping review. J Med Internet Res 2020;22(7):e17296 [FREE Full text] [doi: 10.2196/17296] [Medline: 32706675]
36. Systematic scoping review on social media monitoring methods and interventions relating to vaccine hesitancy. European Centre for Disease Prevention and Control. 2020. URL: https://www.ecdc.europa.eu/sites/default/files/documents/vaccine-hesitancy-systematic-scoping-review-social-media.pdf [accessed 2023-06-06]
37. Majmundar A, Allem JP, Boley Cruz T, Unger JB. The Why We Retweet scale. PLoS One 2018;13(10):e0206076 [FREE Full text] [doi: 10.1371/journal.pone.0206076] [Medline: 30335827]
38. Gruber H. Quoting and retweeting as communicative practices in computer mediated discourse. Discourse Context Media 2017;20:1-9. [doi: 10.1016/j.dcm.2017.06.005]
39. Singh GM, Danaei G, Pelizzari PM, Lin JK, Cowan MJ, Stevens GA, et al. The age associations of blood pressure, cholesterol, and glucose: analysis of health examination surveys from international populations. Circulation 2012;125(18):2204-2211 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.111.058834] [Medline: 22492580]
40. Pinto E. Blood pressure and ageing. Postgrad Med J 2007 Feb;83(976):109-114 [FREE Full text] [doi: 10.1136/pgmj.2006.048371] [Medline: 17308214]
41. Olsen MH, Angell SY, Asma S, Boutouyrie P, Burger D, Chirinos JA, et al. A call to action and a lifecourse strategy to address the global burden of raised blood pressure on current and future generations: the Lancet Commission on hypertension. Lancet 2016;388(10060):2665-2712. [doi: 10.1016/S0140-6736(16)31134-5] [Medline: 27671667]

Abbreviations

1Q: first quartile
2IQR: second quartile
3IQR: third quartile
4IQR: fourth quartile
RQ: research question

Edited by A Mavragani; submitted 18.05.23; peer-reviewed by O Steichen, P Zhang; comments to author 01.03.24; revised version received 12.03.24; accepted 10.05.24; published 20.06.24.

Please cite as:
Terada M, Okuhara T, Yokota R, Kiuchi T, Murakami K
Nutrients and Foods Recommended for Blood Pressure Control on Twitter in Japan: Content Analysis
J Med Internet Res 2024;26:e49077
URL: https://www.jmir.org/2024/1/e49077
doi:10.2196/49077
PMID:

©Marina Terada, Tsuyoshi Okuhara, Rie Yokota, Takahiro Kiuchi, Kentaro Murakami. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing Message Deployment During Public Health Emergencies Through Social Media: Empirical Test of Optimizing Content for Effective Dissemination

Paola Pascual-Ferrá[1], PhD; Neil Alperstein[1], PhD; Julia Burleson[2], MSPH; Amelia M Jamison[2], MAA, MPH; Ananya Bhaktaram[2], MHS; Sidharth Rath[3], BA, IDS; Rohini Ganjoo[4], PhD; Satyanarayan Mohanty[5], MA, MBA; Daniel J Barnett[6], MD; Rajiv N Rimal[2], PhD

[1]Loyola University Maryland, Baltimore, MD, United States

[2]Health, Behavior and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[3]Swasthya Plus Network, Bhubaneswar, OD, India

[4]Biomedical Laboratory Sciences, George Washington University, Washington, DC, United States

[5]D-COR Consulting Pvt Ltd, Bhubaneswar, OD, India

[6]Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

Corresponding Author:
Paola Pascual-Ferrá, PhD
Loyola University Maryland
4501 North Charles Street
Baltimore, MD, 21210
United States
Phone: 1 4106175789
Email: ppascualferra@loyola.edu

Abstract

Background: During an infodemic, timely, reliable, and accessible information is crucial to combat the proliferation of health misinformation. While message testing can provide vital information to make data-informed decisions, traditional methods tend to be time- and resource-intensive. Recognizing this need, we developed the rapid message testing at scale (RMTS) approach to allow communicators to repurpose existing social media advertising tools and understand the full spectrum of audience engagement.

Objective: We had two main objectives: (1) to demonstrate the use of the RMTS approach for message testing, especially when resources and time are limited, and (2) to propose and test the efficacy of an outcome variable that measures engagement along a continuum of viewing experience.

Methods: We developed 12 versions of a single video created for a vaccine confidence project in India. We manipulated video length, aspect ratio, and use of subtitles. The videos were tested across 4 demographic groups (women or men, younger or older). We assessed user engagement along a continuum of viewing experience: obtaining attention, sustaining attention, conveying the message, and inspiring action. These were measured by the percentage of video watched and clicks on the call-to-action link.

Results: The video advertisements were placed on Facebook for over 4 consecutive days at the cost of US $450 and garnered a total of 3.34 million impressions. Overall, we found that the best-performing video was the shorter version in portrait aspect ratio and without subtitles. There was a significant but small association between the length of the video and users' level of engagement at key points along the continuum of viewing experience (N=1,032,888; χ^2_4=48,261.97; P<.001; V=.22). We found that for the longer video, those with subtitles held viewers longer after 25% video watch time than those without subtitles (n=15,597; χ^2_1=7.33; P=.007; V=.02). While we found some significant associations between the aspect ratio, the use of subtitles, and the number of users watching the video and clicking on the call-to-action link, the effect size for those were extremely small.

Conclusions: This test served as a proof of concept for the RMTS approach. We obtained rapid feedback on formal message attributes from a very large sample. The results of this test reinforce the need for platform-specific tailoring of communications. While our data showed a general preference for a short video in portrait orientation and without subtitles among our target audiences on Facebook, that may not necessarily be the case in other social media platforms such as YouTube or TikTok, where

XSL•FO
RenderX

users go primarily to watch videos. RMTS testing highlights nuances that communication professionals can address instead of being limited to a “one size fits all” approach.

(J Med Internet Res 2024;26:e50871) doi:10.2196/50871

KEYWORDS

message testing; web-based communication; user engagement; vaccine communication; methodology; Meta; Facebook; advertising; infodemic; communication; infodemiology; social media advertising tool; social media; audience; engagement; rapid message testing at scale; mobile phone

Introduction

Background

Message testing broadly refers to the practice of using data to make informed decisions about what is the most effective way of conveying a message to a specific target audience in a particular geographical area and cultural context to elicit a desired response. The World Health Organization recommends that health communicators engage in message testing as an important step to increase the efficacy of health campaigns, particularly during public health emergencies [1].

During the COVID-19 pandemic and accompanying infodemic, a deluge of false, misleading, and unreliable health claims obscured credible public health messaging, creating a “simultaneous crises of epidemics and misinformation” [2]. The term infodemic is defined as “an information epidemic that can lead to engaging in dangerous behavior” during a public health emergency [3]. For example, Hu et al [4] found that during the COVID-19 infodemic, there was a “global profusion of monikers and hashtags for COVID-19...[that] contributed to a backlash against China and the Chinese people.” The increased speed and reach with which misinformation spread via social media platforms exacerbated the infodemic. To that end, Hang et al [5] concluded the pandemic brought on an “unprecedented infodemic of false or misleading information that spread rapidly through online social networks.” In a joint statement, the World Health Organization along with the United Nations, UNICEF (United Nations Children’s Fund), and other international organizations emphasized that it is essential for public health messaging during an infodemic to be timely, accurate, and resonant to counter misinformation and disinformation [6].

Message testing can be a vital tool to help communicators counter misinformation, particularly as evidence emerges to support the relationship between quality health communication and increased public trust [7,8]. Message testing can be accomplished using both traditional quantitative (eg, controlled message tests) and qualitative (eg, focus groups) methods. However, acknowledging that falsehoods can often go viral, traditional message testing methods may not generate evidence fast enough to compete in an increasingly crowded and rapidly evolving information ecosystem [9]. Given the resource-intensive nature of traditional message testing and the urgency to respond quickly during public health emergencies, our team recognized the need for more accessible methods for public health practitioners. Rapid message testing at scale (RMTS) is an approach that can help public health professionals and communicators optimize content through rapid testing that is both inexpensive and uses large samples of data to promote evidence-based decision-making.

The RMTS approach works by using data from social media platforms and repurposing them to make evidence-based communication decisions. Furthermore, it recognizes that viewer attention is not binary but falls along a spectrum. We developed a measure of the continuum of viewing experience measures to assess audience engagement and test the RMTS approach. This allowed us to assess differences in user engagement related to 3 content-agnostic metrics: aspect ratio, video length, and subtitle use across different audience types. In addition to testing the RMTS approach, the results of this analysis produced valuable information for users looking to improve advertisement performance on social media more generally, particularly when adapting content from other media formats such as television advertisements, thus helping organizations increase their return on investment.

Web-Based Message Testing

Due to the low barrier to entry and potential for immediate and widespread dissemination, social media is often viewed as a cost-effective vehicle for a communication campaign, especially when time and reach are of the essence, such as during an infodemic. Platforms such as Google and Meta have created user-friendly interfaces that make low-cost digital advertising accessible to almost anyone with a credit card. Many of these services offer data for users to make sense of audience engagement and select specific advertisement features. While placing video advertisements on social media can be less expensive than paying for television advertisements, the number of available platforms—each with its own features and formats—makes designing messages quite complex. For example, on some platforms, a horizontal aspect ratio for viewing videos is preferred (YouTube, Google LLC), while on others vertical is preferred (Instagram, Meta Platforms, Inc, or Facebook, Meta Platforms, Inc). This may seem trivial at first glance, but it is important to recognize that there is agency in viewing videos on digital media, making it easy to move on to the next video. In other words, the format may impact whether the video is watched at all, viewed for a few seconds or longer, or results in repeat viewing. Video length is another factor to be considered. Traditional public service announcements on television may last up to 2 minutes. However, in the digital age, many people watch only a few seconds of a video before moving on to other content; thus, structuring a video to grab and keep the attention of the viewer is paramount. This is not to negate the importance of compelling and theoretically developed content. Rather, we propose that, in a digital environment,

XSL•FO
RenderX

format attributes and formal features can make a difference in terms of sustaining attention and user engagement.

With video content in particular, the issue of viewer abandonment—abandoning a video before watching it in its entirety—is one of the greatest challenges for communicators. Researchers who studied watch time for short videos (5 minutes or less) on YouTube found that on average, 20% of viewers abandoned the video within the first 10 seconds, 33% abandoned the video within the first 30 seconds, and 44% of viewers did so within the first 60 seconds [10]. This is an important consideration when designing and creating video content for social media.

Another challenge when placing advertisements on social media platforms is banner blindness, a form of practiced avoidance in which individuals turn away from advertising content that appears on the screen [11]. It is all too easy to equate exposure to social media with awareness. However, distracted viewing or watching a message with split attention—when users' eyes are on the screen but their minds are elsewhere—has become commonplace in the digital world. For that matter, the amount of views and likes a video receives are considered superficial or vanity metrics that may not accurately reflect the quality of user engagement [12]. The issue of attention span extends beyond the user to consider the platform itself and how its algorithms' video length preferences affect what content is suggested to viewers [13]. The viewing experience on digital platforms is complex as viewing situations are different for different media. For instance, while sound, sound effects, and music may be important on a specific medium or in a private (home) viewing situation, in other more public viewing situations, a video whose message comes through without sound may be the preferred option [14,15]. Multimedia Appendix 1 provides a glossary of terms relevant to our study.

Assessing User Engagement

With the RMTS approach, we sought to move beyond vanity metrics to develop a model of more nuanced metrics based on a continuum of viewing experience. The approach is based on A/B testing premises, a standard in marketing practice that works well within the bounds of digital media. The practice is commonly used to test which 1 of 2 options (in our case, video advertisements) performs better among specific target audiences to make an evidence-based decision before investing significant advertising resources. Public health communicators are beginning to adopt this marketing industry approach by using A/B testing [16]. For instance, Facebook teamed up with UNICEF during the COVID-19 pandemic to test the effectiveness of different messages in building vaccine confidence using their social media platform [17]. Their research formed the basis for the Vaccine Demand Observatory, a collaboration between UNICEF, The Public Goods Projects, and Yale University's Institute for Global Health to help counter vaccine misinformation using social listening analytics [18]. Their research confirmed the ability to effectively microtarget vaccine promotion messaging to specific audiences through A/B testing [19].

With the RMTS approach, we move beyond A/B testing's focus on "which advertisement pulls best" to a continuum of viewing experience that begins with a view or simple exposure as a baseline (obtaining attention), then progresses on to consider the percentage of a video watched—5%, 25%, 75%, or 90% (sustaining attention to conveying a message), and counts clicks on a call-to-action link (inspiring action) as a more meaningful way to measure user engagement. There is flexibility in this model to apply weighting to the viewing experiences of different audiences. For example, questions can be raised about whether younger viewers watch more or less of a video compared to older viewers, or whether younger people are more likely to click on the call-to-action link than older people.

In addition to conceptualizing age as a binary, we can also apply the same viewing experiences to gender or look at both gender and the gender-by-age interaction to gauge viewer experiences. With this data we can segment the audience based on their viewing experiences, a novel approach that not only delivers the optimized message to the audience most receptive to it but also considers viewing platforms or viewing style based on the user's preference. Such an approach allows the content creator to choose the optimal combination of specific format attributes and formal features that can help the message resonate best among specific target audiences.

In our test of the RMTS, we were interested in exploring the impact of (1) the video's aspect ratio (ie, rectangular, square, portrait), (2) the video's length (short or long), and (3) the use of subtitles (present or absent) on the continuum of the viewing experience. Specifically, we raised the following research questions (RQs):

- RQ1: What is the impact of changing the aspect ratio of a video—a format attribute—on watch time and clicks on the call-to-action link?
- RQ2: How do formal features such as video length and subtitles impact watch time and clicks?
- RQ3: How do watch time and clicks of optimized videos compare to those of videos that have not been optimized?

We were also interested in examining the combination of video format attributes and formal features that led to higher abandonment rates for each of the 4 demographic segments in our study.

In summary, the purpose of this study was two-fold: (1) to demonstrate the use of the RMTS approach for message testing, especially when time and resources are limited such as during an infodemic, and (2) to propose and test the efficacy of an outcome variable that measures engagement along a continuum of viewing experience.

Methods

Study Design

This study is part of a larger vaccine confidence project we conducted in Odisha, India [20]. As of 2024, India is the country with the highest number of Facebook users in the world, with over 1 in 5 people using the platform [21]. As of February 2024, the largest Facebook user group in India were people aged 25-44 years, followed by people aged 18-24 years, while the majority of users (68%) identified as male [22].

XSL•FO
RenderX

Intervention

For this study, we created 48 paid video advertisements on Facebook. The videos were all variations of one of the videos that we had produced for the larger vaccine confidence project to promote uptake of the COVID-19 vaccine, specifically the second dose. The original video length was 1 minute and 55 seconds and used narrative storytelling to address vaccine hesitancy. The story was set during a child's birthday celebration. Mini, the protagonist, arrives late to the party and shares that she needs to leave early to go get the second dose of the COVID-19 vaccine. The video shows the conversation that ensues among family members. Mini explains the benefits of being vaccinated and counters the arguments made by Somnath, who plays the role of the vaccine skeptic, using a collective health frame. The video promotes the importance of being fully vaccinated against COVID-19 and includes a link that the viewer can click on to learn more and schedule an appointment to be vaccinated.

For this study, we created a shorter 58-second version of the video. For both the short and long versions of the video, we created a version that included subtitles in Odia, the language spoken in the video. By subtitles, we refer to captions or the accompanying text in the same language spoken in the videos that are shown near the bottom of the frame, imposed over a black gradient background to make them easier to read. The subtitles present the dialogue in text so that users who have their audio turned off, which is a popular behavior for users consuming content on their phones [14], can also follow the dialogue. The use of subtitles in our study is akin to the practice of close captioning, which has the additional benefit of making the content more accessible for Odia audiences who have difficulty processing audio content.

Variables

The independent variables we looked at in our study were aspect ratio, subtitles, and length of videos. We took each of the 4 video variations (long or short, with or without subtitles) and configured them in 3 different aspect ratios: 16:9 (rectangle), 1:1 (square), and 4:5 (portrait) [23]. Table 1 shows the 12 video variations we tested in this study. We manipulated the length of the video, the aspect ratio, and the use (or lack of) subtitles. The original video was 1 minute 55 seconds long. The short version was 58 seconds long. We used the following Facebook specifications for testing the aspect ratio variable: rectangle (16:9), square (1:1), and portrait (4:5). All video advertisements included a call-to-action link that viewers could click on to find more information about vaccination.

Figure 1 shows how different aspect ratios affect the extent of the original image that is visible to the viewer: rectangular (the entire image), square (the portion of the image shown between the solid white lines), and portrait (the portion of the image shown within the dashed lines). The original video was shot in landscape orientation. The still captures the moment when Somnath (male, left), who played the role of the vaccine skeptic, gestures that he does not intend to receive the vaccine. The images include subtitles in Odia, the language spoken in the video [24]. In the portrait aspect ratio version of the video, the subtitles appear centered at the bottom of the screen.

In terms of demographic variables, the 12 variations we created for this study were tested with different audience demographics within the Odisha, India region. We used 4 age or gender combinations (ie, women or men, younger or older) equaling 48 unique advertisements. In our study, "younger" refers to users aged 18-22 years, and "older" refers to users aged 30-35 years, both falling within our age demographic of interest. We chose these 2 variables because age and gender are two of the most basic (and identifiable) demographic elements; these are the go-to variables as public health agencies would want to know how younger and older people, as well as men and women, differ. Each Facebook user within our target demographic was exposed to only 1 version of the video.

Table 1. Video variations by aspect ratio, presence or absence of subtitles, and video length.

Video variation number	Aspect ratio	Subtitles	Video length
1	Rectangle	No	Original
2	Rectangle	No	Short
3	Rectangle	Yes	Original
4	Rectangle	Yes	Short
5	Square	No	Original
6	Square	No	Short
7	Square	Yes	Original
8	Square	Yes	Short
9	Portrait	No	Original
10	Portrait	No	Short
11	Portrait	Yes	Original
12	Portrait	Yes	Short

XSL•FO
RenderX

Figure 1. Impact of aspect ratio on image area visible to viewers.

Measures

In our study, we were interested in measuring engagement along a continuum of viewing experience as our main outcome variable. Engagement was assessed as the number of seconds that the viewer watched the video and whether or not they clicked on the call-to-action link. We compared the results based on a continuum of viewing experience (ie, obtaining attention, sustaining attention, conveying a message, and inspiring action) to determine which approach would likely be more successful based on the criteria established for the test. We operationalized each step along the continuum of viewing experience as follows: obtaining attention (user watched 5% of the total video length), sustaining attention (user watched 25%-75% of the total video length), conveying a message (user watched 90% of the total video length), and inspiring action (user clicked on the call-to-action link). We compared the level of engagement garnered by the short video to the level of engagement garnered by the longer video. We were interested in learning how long viewers watched the video (watch time) and whether they clicked on the call-to-action link. We calculated the abandonment rate (AR) as follows:

where VT_1 represents the number of views at the time stamp closer to the start of the video and VT_2 represents the number of views at the time stamp further from the start of the video. We were also interested in whether viewers clicked on the call-to-action link shown in the video as a measure of behavior related to the viewing experience.

Statistical Analysis

For the statistical analysis, we ran a 3 (aspect ratio: rectangular, square, or portrait) × 2 (length: original or shortened) × 2 (presence or absence of subtitles) between-subjects experiment across 2 (age: younger or older) × 2 (sex: male or female) groups (for a total of 48 groups), using Meta's A/B testing platform [25]. We used χ^2 tests of independence to see if there were significant differences in the values observed across the variables of interest. In terms of sample size, N represents the total number of views and clicks for the 12 video variations, while n is used for the specific number of users watching the videos at 5%, 25%, 75%, and 90% watch time and the specific number of users who clicked on the call-to-action link. The results were considered significant when the probability or *P* value was .05 or less. Effect sizes were calculated using Cramér *V*, which is recommended for measuring the strength of association between nominal variables when the variables have more than 2 categories (the contingency table is larger than 2 × 2). Effect sizes were interpreted using Cohen's [26] conventions for interpreting equivalent ϕ values. All statistics were calculated in Excel.

Ethical Considerations

This study was approved by both the Institutional Review Board (IRB) at Johns Hopkins University in the United States (#18543) and the Sigma IRB in India (#10075). All data used in this study were collected through Meta's A/B platform, which does not provide individual-level data. Therefore, the data are anonymized and aggregated upon receipt, without any personally identifiable information, and consistent with Meta's data privacy and use policies.

Results

The test was conducted over four consecutive days beginning in early August 2022 and cost US $450. The 48 video advertisements placed on Facebook garnered a total of 3.34 million impressions. The videos received a total of 700,239 views lasting at least 3 seconds (a 21% conversion rate from impressions); 182,890 views lasting at least 15 seconds; and a total of 25,270 clicks on the call-to-action link (3.6% of all 3-second views). The majority of viewers at 3 seconds of video play were men (399,703/700,239, 57%) and users aged between

XSL•FO
RenderX

30 and 35 years (444,164/700,239, 63%). Multimedia Appendix 2 contains the aggregate data from our study. We created a prototype for a web-based dashboard that would allow content creators to make evidence-based decisions based on the demographic segment (Figure 2). We have included a link to the web-based dashboard in the references section [27].

Table 2 shows the results of the χ^2 tests by variable. In terms of RQ1, we observed that the video advertisements shown in the rectangle aspect ratio consistently experienced the highest abandonment rate, while the opposite was true for video advertisements shown in the portrait aspect ratio. In other words, the portrait aspect ratio seemed to retain viewers longer than other aspect ratios. The same was true for the numbers of clicks on the call-to-action link, with videos in the portrait aspect ratio yielding the greatest number of clicks, and videos in the rectangular aspect ratio yielding the least number of clicks.

In terms of RQ2, we observed that the short video, including both with and without subtitles, consistently outperformed the original (long) video across all audience segments. In fact, there was a strong and significant negative correlation between watch time and number of users watching, $r(8)=-0.82$, $P=.01$. In other words, as video play length increased, the number of users watching decreased. The abandonment rate for the original (long) video, both with and without subtitles, was almost 99% from 5% to 90% watch time, compared to 88% for the short video.

In terms of RQ3, the short video *without* subtitles attracted and retained more views longer than the short version *with* subtitles. We found that what we initially thought would be the optimal version (shorter, with subtitles and using the portrait aspect ratio) was not the best-performing version for all audience segments. We also found that, overall, the long video version with subtitles retained more views than the long version without them after the 25% watch time drop-off (n=15,597; $\chi 21=7.33$; $P=.007$; V=.02). This was true for all audience segments except younger women, for whom the short video version shown in portrait orientation and without subtitles performed better across the board.

Figure 2. Screenshot of web-based dashboard comparing video variations.

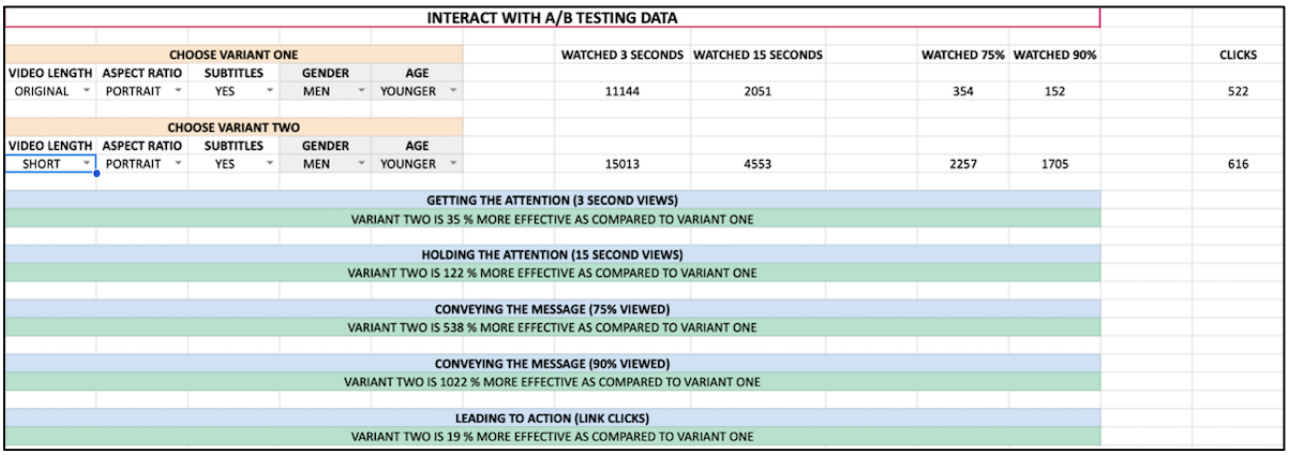

XSL•FO
RenderX

Table 2. $\chi 2$ test results by variable. All results were significant (P<.001).

Variable	Views by percentage of video watched, n				Clicks, n	Chi-square (*df*)	Cramér *V*
	5%	25%	75%	90%			
Gender						3999.68 (4)	0.06
Women	300,536	89,090	37,162	26,136	11,831		
Men	399,703	93,800	36,228	24,963	13,439		
Age range (years)						833.17 (4)	0.03
18-22	256,075	68,710	26,370	18,227	11,358		
30-35	444,164	114,180	47,020	32,872	13,912		
Subtitles						27.33 (4)	0.01
Yes	347,714	89,763	36,402	25,129	12,328		
No	352,525	93,127	36,988	25,970	12,942		
Video length						48,261.97 (4)	0.22
Original	306,126	58,211	11,093	4504	11,736		
Short	394,113	124,679	62,297	46,595	13,534		
Aspect ratio						3019.08 (8)	0.04
Rectangle	227,017	54,638	19,610	13,387	7097		
Square	233,754	59,997	23,967	16,524	8687		
Portrait	239,468	68,255	29,813	21,188	9486		

Discussion

Principal Findings

Our main goal in this study was to introduce and apply a new approach for rapid message testing at scale that could be used by public health professionals to test and optimize content for public dissemination quickly and effectively during a public health crisis. The RMTS approach allowed us to test 48 social media advertisements on Facebook quickly and at a large scale. Overall, our results indicate that, regardless of aspect ratio manipulations and the presence or absence of subtitles, shorter videos, in portrait aspect ratio and without subtitles performed better across the board than other versions of the video. Our results are not surprising given what we know about how users tend to access social media on their mobile devices. The preference for portrait orientation could be because most users access Facebook on their smartphones, and most smartphone users hold their phones upright [28]. In that orientation, portrait aspect videos fill the screen and may create a more pleasant viewing experience without the user needing to expend additional energy tilting the device horizontally to view in landscape mode.

At the same time, it is not surprising that the shorter video version was able to retain more views than the original (longer) video version. Given what we know about distracted viewing and split attention among internet users, we do not interpret measures of viewing attention that are less than 100% as necessarily negative. While video producers and message sponsors strive to achieve 100% video completion rates, increased video abandonment rates have been a pesky issue for social media marketers. However, a video that is viewed 25% of the way through may not be unsuccessful, as the viewer may have gotten out of the video all that they wanted; the viewer perhaps clicked on the “learn more” button before the end of the video. However, it is also possible within this construct that a viewer decides to move on to other content after a few seconds of viewing.

In terms of subtitles, we were surprised to find that the video variations without subtitles performed better than those with subtitles. What we initially thought would be the optimal version (shorter, with subtitles and using the portrait aspect ratio) was not the best-performing version for all audience segments. Instead, the short video version shown in portrait orientation and without subtitles received the highest number of clicks across all audience segments. This is contrary to what we expected to find given industry reports citing the trend of users watching media on their mobile devices with the sound turned off. We thought the addition of subtitles would be beneficial to those users, but videos with subtitles had a higher abandonment rate at 25% watch time than those without. However, subtitles seemed to have helped retain more viewers and sustain their attention past the 25% watch time mark, except for younger women. Further research focusing on qualitative methodologies might shed light on these findings.

Our data suggests that on Facebook, shorter videos (<1 minute) shown in portrait orientation and without subtitles can lead to greater viewer engagement, but that subtitles may be beneficial in retaining views on longer videos (>1 minute). These findings may inform the construction of videos for dissemination on social media—editing the video for a shorter or perhaps longer beginning if, for example, the goal is to keep attention for as long as possible. At the end of the continuum is the behavioral outcome or action, which would take the form of a click on a

XSL•FO
RenderX

"learn more" button embedded either in a video or video description. Further, while the ultimate measure may be whether the viewer clicks on the call-to-action link or not, we would not want to discount the other measures of attention offered by this continuum of viewing experience model based on viewer demographics of age and gender.

Strengths and Limitations

The proposed RMTS approach satisfies some of the limitations present in A/B testing about the length of the test and the size of the sample. The issue regarding how long one should run an A/B test depends on the time and resources of the organization. Meta recommends that users of their A/B testing platform run their A/B tests for at least 2 weeks to 30 days [29]. However, depending on how many variations of an advertisement one is testing, this length of time can quickly add up. Furthermore, in a public health emergency, when time is of the essence, spending 2 weeks to a month just on testing is impractical. For all these reasons and working with a budget that we believed would be affordable (under US $500), we selected a 4-day test. During this 4-day interval, we garnered over 3 million impressions and over 1.3 million views beyond the first 3 seconds of the advertisements.

Our research study has several limitations. First, our research was limited to one social media platform, Facebook, and to the metrics provided by their A/B testing platform. We used Facebook because of the extent of its use among our target audience in India and because Meta's A/B testing platform allowed us to test different variations of 1 video conveniently and cost-effectively. Our study looks at the demographic variables of age (young and old) and gender (male and female) only and does not integrate more diverse demographic variables that could offer a more nuanced understanding of viewer behavior. Unfortunately, these are the only demographic variables that we have access to through the Facebook A/B testing platform as part of their user privacy protections. We are also mindful of the fact that each platform has its own algorithmic bias and that the viewing experience is different depending on the platform. Prior research has shown the benefit of adopting a multi-platform approach that incorporates data from at least 2 social media platforms [30]. Given this limitation, we caution readers from extrapolating our Facebook findings to other social media platforms and recommend that future research incorporates a multi-platform approach. At the same time, the formats that we studied can be applied to other social media platforms such as Instagram, YouTube, and TikTok. We believe this approach can help organizations quickly and efficiently adapt messages across platforms without having to change the content to better target different audience segments, especially in time-sensitive situations when messages need to be disseminated quickly.

Another limitation is that when we shorten videos for testing as we did in this study, the longer and shorter versions vary on numerous dimensions and content, and hence a direct comparison might be difficult. This speaks to internal validity, as it affects our ability to make causal inferences. On the other hand, content creators also look for recommendations regarding the ideal length of video. Hence, if the recommendation is to make shorter videos, the underlying premise, of course, is that we have sacrificed content that would have been available in the longer format; that is the price we have to pay in comparing short and long videos.

Finally, while we found some significant associations between the length of the video, the use of subtitles, and the number of users watching the video at key points along the continuum of viewing experience, the effect size for those were overall very small. This means that our results have limited practical applications when it comes to the variables we tested. Maybe researchers looking at other forms of content beyond video advertisements on Facebook, and content variables other than aspect ratio and the use of subtitles, may find different and perhaps larger effect sizes. However, when it comes to population health interventions, Matthay et al [31] have noted that "even studies of interventions with small effect sizes can offer valuable evidence to inform population health if such interventions can be implemented broadly."

Future Research

Based on our study's findings, we have several recommendations for future research. Given the reality of distracted viewing and attention spans on social media platforms that we mentioned, future research could assess the sustained impact of video advertisements on viewer behavior over time beyond initial engagement metrics. Our study focused on only 1 message to address vaccine hesitancy, with 1 protagonist (female), and using a collectivist appeal. We experimented with video length, aspect ratio, and subtitles. Future research could look at the impact of additional content variables such as different messages to address vaccine hesitancy, tone of voice, other forms of appeal, the identity of the protagonist, and cues such as music, to name some, that could also significantly impact viewer engagement. Our study was also limited to 1 social media platform, Facebook, which has its own unique algorithmic bias. At the most basic level, Meta's A/B testing platform results we had access to were limited to binary gender data, excluding the perspectives of those who do not conform to those binary norms. Future research should look at incorporating other forms of content (beyond video advertisements) and applying the RMTS approach using other social media platforms.

This study is also limited by its quantitative methodology and its cultural specificity. To understand users' more nuanced experience, we recommend that future research include a qualitative component in the methodology that explores how users from different backgrounds and identities consume videos, not just those whom Facebook identifies as young or old, female or male. Furthermore, our study was limited to Facebook users in Odisha, India. The content of the video itself was developed in collaboration with a Community Advisory Board in Odisha, and so, in a way, could be considered culturally specific. Future research should explore the impact that cultural differences have on message reception and engagement, especially given the global nature of social media platforms [32].

Finally, using social media for research can raise ethical considerations that researchers should carefully navigate, ranging from informed consent, privacy, data security,

XSL•FO
RenderX

transparency, and accountability. For this study, we mentioned one of the limitations related to our limited demographic data. That, in itself, is a function of the process we used to collect the data. We worked exclusively through Meta's A/B testing platform, which provides only anonymized and aggregated quantitative data in a way that conforms to Meta's data policy. In other words, we have no way of identifying individuals based on the data we received. However, researchers looking at conducting studies with a strong qualitative component (eg, using comments and personally-identifying data posted on social media that is not meant to be shared outside that group or platform) should comply with their institutions' IRB approval process, prioritize seeking informed consent from users, and afford the appropriate protections to participants following best practices for conducting qualitative research using social media [33-35]. In addition, special care needs to be taken when conducting tests during an infodemic, where there is already an overwhelming amount of information, some of which may be mis- or disinformation, competing for users' attention and which could take away from the efforts of public health agencies. We believe this has to be balanced with conducting research whose findings can be critical in how subsequent crises are handled and for which communication guidelines are prepared. We hope that professionals working in public health will be able to use this RMTS approach to effectively respond and reach different publics during an unfolding public health crisis.

Acknowledgments

This research was funded by Global Impact (VCF-019) to Johns Hopkins University (Rajiv N Rimal, principal investigator). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of Global Impact.

Data Availability

The data used in this study is included in Multimedia Appendix 2 as a .docx file. Please contact the corresponding author if you would prefer a .csv file.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Glossary of terms related to the continuum of viewing experience.
[DOCX File , 29 KB - jmir_v26i1e50871_app1.docx]

Multimedia Appendix 2
Number of users who watched each video variation by age and gender, percentage of total video length played, and number of clicks (data file).
[DOCX File , 34 KB - jmir_v26i1e50871_app2.docx]

References

1. The Collective Service EPI-WIN Team. Message testing during an epidemic or pandemic. World Health Organization. URL: https://www.who.int/teams/epi-win/the-collective-service/message-testing#:~:text=During%20these%20challenging%20times%2C%20it%27s,what%20will%20work%20and%20why [accessed 2023-06-30]
2. Tan CW, Yu P. Contagion source detection in epidemic and infodemic outbreaks: mathematical analysis and network algorithms. Found Trends Networking 2023;13(2-3):107-251. [doi: 10.1561/1300000068]
3. Rovetta A, Castaldo L. Are we sure we fully understand what an infodemic is? a global perspective on infodemiological problems. JMIRx Med 2022;3(3):e36510 [FREE Full text] [doi: 10.2196/36510] [Medline: 36409169]
4. Hu Z, Yang Z, Li Q, Zhang A. The COVID-19 infodemic: infodemiology study analyzing stigmatizing search terms. J Med Internet Res 2020;22(11):e22639 [FREE Full text] [doi: 10.2196/22639] [Medline: 33156807]
5. Hang CN, Yu P, Chen S, Tan CW, Chen G. MEGA: machine learning-enhanced graph analytics for infodemic risk management. IEEE J Biomed Health Inform 2023;27(12):6100-6111. [doi: 10.1109/JBHI.2023.3314632] [Medline: 37713230]
6. Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation. World Health Organization (WHO). URL: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation [accessed 2023-06-30]
7. Hyland-Wood B, Gardner J, Leask J, Ecker UKH. Toward effective government communication strategies in the era of COVID-19. Humanit Soc Sci Commun 2021;8(1):30. [doi: 10.1057/s41599-020-00701-w]

XSL•FO
RenderX

8. Walter N, Brooks JJ, Saucier CJ, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. Health Commun 2021;36(13):1776-1784. [doi: 10.1080/10410236.2020.1794553] [Medline: 32762260]
9. Bak-Coleman JB, Kennedy I, Wack M, Beers A, Schafer JS, Spiro ES, et al. Combining interventions to reduce the spread of viral misinformation. Nat Hum Behav 2022;6(10):1372-1380 [FREE Full text] [doi: 10.1038/s41562-022-01388-6] [Medline: 35739250]
10. Cutler M. Why YouTube viewers have ADD and how to stop it. AdAge. URL: https://adage.com/article/digitalnext/marketing-online-video-viewers-quit-30-seconds/146218 [accessed 2023-06-30]
11. Pernice K. Banner blindness revisited: users dodge ads on mobile and desktop. Nielsen Norman Group. URL: https://www.nngroup.com/articles/banner-blindness-old-and-new-findings/ [accessed 2023-06-30]
12. Rogers R. Otherwise engaged: social media from vanity metrics to critical analytics. Int J Commun 2018;12:450-472 [FREE Full text]
13. Rieder B, Coromina Ò, Matamoros-Fernández A. Mapping YouTube: a quantitative exploration of a platformed media system. First Monday 2020;25(8). [doi: 10.5210/fm.v25i8.10667]
14. Lafayette J. Mobile videos often watched without audio, study finds. Next TV. URL: https://www.nexttv.com/news/mobile-videos-often-watched-without-audio-study-finds [accessed 2023-06-30]
15. Best practices for video mobile ads. Meta Business Help Center. URL: https://www.facebook.com/business/help/144240239372256?id=603833089963720 [accessed 2023-06-30]
16. Brown KV. To reach vaccine holdouts, scientists take a page from digital marketing. Bloomberg. URL: https://www.bloomberg.com/news/articles/2021-09-25/to-reach-vaccine-holdouts-scientists-take-a-page-from-digital-marketing [accessed 2023-06-30]
17. Khan K, McGorman L. Building vaccine confidence amid COVID-19 with UNICEF. Meta. 2021. URL: https://about.fb.com/news/2021/04/building-vaccine-confidence-amid-covid-19-with-unicef/ [accessed 2023-06-30]
18. Sidhu S. Vaccination demand observatory launched to strengthen local communication programmes to address vaccine misinformation. UNICEF. URL: https://www.unicef.org/eap/press-releases/vaccination-demand-observatory-launched-strengthen-local-communication-programmes [accessed 2023-06-30]
19. Winters M, Christie S, Lepage C, Malik AA, Bokemper S, Abeyesekera S, et al. Persuasive COVID-19 vaccination campaigns on Facebook and nationwide vaccination coverage in Ukraine, India, and Pakistan. PLOS Glob Public Health 2023;3(9):e0002357 [FREE Full text] [doi: 10.1371/journal.pgph.0002357] [Medline: 37756298]
20. Bhaktaram A, Ganjoo R, Jamison AM, Burleson J, Pascual-Ferra P, Alperstein N, et al. Creation, dissemination, and evaluation of videos to promote COVID-19 vaccination in India: a research protocol. Gates Open Res 2022;6:82 [FREE Full text] [doi: 10.12688/gatesopenres.13628.2] [Medline: 36415882]
21. Digital 2024 April global statshot report. DataReportal. URL: https://datareportal.com/reports/digital-2024-april-global-statshot [accessed 2024-05-04]
22. Facebook users in India. NapoleonCat Team. 2024. URL: https://napoleoncat.com/stats/facebook-users-in-india/2024/02/ [accessed 2024-05-04]
23. Meta. Best practices for aspect ratios. Meta Business Help Center. URL: https://www.facebook.com/business/help/103816146375741?id=271710926837064 [accessed 2023-06-30]
24. Burleson J, Ganjoo R, Rath S, Rath N, Bhaktaram A, Jamison AM, et al. Production of vaccination videos in India: learnings from a science-art partnership. BMC Public Health 2023;23(1):736 [FREE Full text] [doi: 10.1186/s12889-023-15607-w] [Medline: 37085863]
25. Meta. About A/B testing. Meta Business Help Center. URL: https://www.facebook.com/business/help/1738164643098669?id=445653312788501&ref=rooster [accessed 2023-06-30]
26. Cohen H. Statistical Power Analysis for Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
27. Web-based dashboard: how public health videos for boosting vaccine confidence perform on Facebook. Swasthya Plus Network. URL: https://swasthyaplus.com/dashboard-insights-vcf/ [accessed 2023-06-30]
28. Mulier L, Slabbinck H, Vermeir I. This way up: the effectiveness of mobile vertical video marketing. J Interact Market 2021;55:1-15. [doi: 10.1016/j.intmar.2020.12.002]
29. Find your winning campaign strategy with A/B testing. Meta Business Help Center. URL: https://www.facebook.com/business/ads/ab-testing-advanced# [accessed 2024-05-04]
30. Pascual-Ferrá P, Alperstein N, Barnett DJ. A multi-platform approach to monitoring negative dominance for COVID-19 vaccine-related information online. Disaster Med Public Health Prep 2022;16(6):2471-2479 [FREE Full text] [doi: 10.1017/dmp.2021.136] [Medline: 33938423]
31. Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler N, et al. Powering population health research: considerations for plausible and actionable effect sizes. SSM Popul Health 2021;14:100789 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100789] [Medline: 33898730]
32. Airhihenbuwa C, Iwelunmor J, Munodawafa D, Ford C, Oni T, Agyemang C, et al. Culture matters in communicating the global response to COVID-19. Prev Chronic Dis 2020;17:200245 [FREE Full text] [doi: 10.5888/pcd17.200245] [Medline: 32644918]

XSL•FO
RenderX

33. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
34. Samuel G, Buchanan E. Guest editorial: ethical issues in social media research. J Empir Res Hum Res Ethics 2020;15(1-2):3-11. [doi: 10.1177/1556264619901215] [Medline: 31959061]
35. Lathan HS, Kwan A, Takats C, Tanner JP, Wormer R, Romero D, et al. Ethical considerations and methodological uses of Facebook data in public health research: a systematic review. Soc Sci Med 2023;322:115807 [FREE Full text] [doi: 10.1016/j.socscimed.2023.115807] [Medline: 36889221]

Abbreviations

IRB: Institutional Review Board
RMTS: rapid message testing at scale
RQ: research question
UNICEF: United Nations Children's Fund

Edited by A Mavragani; submitted 14.07.23; peer-reviewed by F Medina, CN Hang; comments to author 09.03.24; revised version received 09.05.24; accepted 11.06.24; published 26.07.24.

Please cite as:
Pascual-Ferrá P, Alperstein N, Burleson J, Jamison AM, Bhaktaram A, Rath S, Ganjoo R, Mohanty S, Barnett DJ, Rimal RN Assessing Message Deployment During Public Health Emergencies Through Social Media: Empirical Test of Optimizing Content for Effective Dissemination
J Med Internet Res 2024;26:e50871
URL: https://www.jmir.org/2024/1/e50871
doi:10.2196/50871
PMID:38861266

©Paola Pascual-Ferrá, Neil Alperstein, Julia Burleson, Amelia M Jamison, Ananya Bhaktaram, Sidharth Rath, Rohini Ganjoo, Satyanarayan Mohanty, Daniel J Barnett, Rajiv N Rimal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cancer Prevention and Treatment on Chinese Social Media: Machine Learning–Based Content Analysis Study

Keyang Zhao[1*], DPhil; Xiaojing Li[1,2*], Prof Dr; Jingyang Li[3], DPhil

[1]School of Media & Communication, Shanghai Jiao Tong University, Shanghai, China

[2]Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China

[3]School of Software, Shanghai Jiao Tong University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Xiaojing Li, Prof Dr
School of Media & Communication
Shanghai Jiao Tong University
800 Dongchuan Rd.
Minhang District
Shanghai, 200240
China
Phone: 86 13918611103
Fax: 86 21 34207088
Email: lixiaojing@sjtu.edu.cn

Abstract

Background: Nowadays, social media plays a crucial role in disseminating information about cancer prevention and treatment. A growing body of research has focused on assessing access and communication effects of cancer information on social media. However, there remains a limited understanding of the comprehensive presentation of cancer prevention and treatment methods across social media platforms. Furthermore, research comparing the differences between medical social media (MSM) and common social media (CSM) is also lacking.

Objective: Using big data analytics, this study aims to comprehensively map the characteristics of cancer treatment and prevention information on MSM and CSM. This approach promises to enhance cancer coverage and assist patients in making informed treatment decisions.

Methods: We collected all posts (N=60,843) from 4 medical WeChat official accounts (accounts with professional medical backgrounds, classified as MSM in this paper) and 5 health and lifestyle WeChat official accounts (accounts with nonprofessional medical backgrounds, classified as CSM in this paper). We applied latent Dirichlet allocation topic modeling to extract cancer-related posts (N=8427) and identified 6 cancer themes separately in CSM and MSM. After manually labeling posts according to our codebook, we used a neural-based method for automated labeling. Specifically, we framed our task as a multilabel task and utilized different pretrained models, such as Bidirectional Encoder Representations from Transformers (BERT) and Global Vectors for Word Representation (GloVe), to learn document-level semantic representations for labeling.

Results: We analyzed a total of 4479 articles from MSM and 3948 articles from CSM related to cancer. Among these, 35.52% (2993/8427) contained prevention information and 44.43% (3744/8427) contained treatment information. Themes in CSM were predominantly related to lifestyle, whereas MSM focused more on medical aspects. The most frequently mentioned prevention measures were early screening and testing, healthy diet, and physical exercise. MSM mentioned vaccinations for cancer prevention more frequently compared with CSM. Both types of media provided limited coverage of radiation prevention (including sun protection) and breastfeeding. The most mentioned treatment measures were surgery, chemotherapy, and radiotherapy. Compared with MSM (1137/8427, 13.49%), CSM (2993/8427, 35.52%) focused more on prevention.

Conclusions: The information about cancer prevention and treatment on social media revealed a lack of balance. The focus was primarily limited to a few aspects, indicating a need for broader coverage of prevention measures and treatments in social media. Additionally, the study's findings underscored the potential of applying machine learning to content analysis as a promising research approach for mapping key dimensions of cancer information on social media. These findings hold methodological and practical significance for future studies and health promotion.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55937) doi:10.2196/55937

KEYWORDS

social media; cancer information; text mining; supervised machine learning; content analysis

Introduction

In 2020, 4.57 million new cancer cases were reported in China, accounting for 23.7% of the world's total [1]. Many of these cancers, however, can be prevented [2,3]. According to the World Health Organization (WHO), 30%-50% of cancers could be avoided through early detection and by reducing exposure to known lifestyle and environmental risks [4]. This underscores the imperative to advance education on cancer prevention and treatment.

Mass media serves not only as a primary channel for disseminating cancer information but also as a potent force in shaping the public health agenda [5,6]. Previous studies have underscored the necessity of understanding how specific cancer-related content is presented in the media. For example, the specific cancer types frequently mentioned in news reports have the potential to influence the public's perception of the actual incidence of cancer [7].

Nowadays, social media plays an essential role in disseminating health information, coordinating resources, and promoting health campaigns aimed at educating individuals about prevention measures [8]. Additionally, it influences patients' decision-making processes regarding treatment [9]. A study revealed that social media use correlates with increased awareness of cancer screening in the general population [10]. In recent years, there has been a notable surge in studies evaluating cancer-related content on social media. However, previous studies often focused on specific cancer types [11] and limited aspects of cancer-related issues [12]. The most recent comprehensive systematic content analysis of cancer coverage, conducted in 2013, indicated that cancer news coverage has heavily focused on treatment, while devoting very little attention to prevention, detection, or coping [13].

Evaluating cancer prevention information on social media is crucial for future efforts by health educators and cancer control organizations. Moreover, providing reliable medical information to individuals helps alleviate feelings of fear and uncertainty [14]. Specifically, patients often seek information online when making critical treatment decisions, such as chemotherapy [15]. Therefore, it is significant to comprehensively evaluate the types of treatment information available on social media.

Although many studies have explored cancer-related posts from the perspectives of patients with cancer [16] and caregivers [17], the analysis of posts from medical professionals has been found to be inadequate [18]. This paradox arises from the expectation that medical professionals, given their professional advantages, should take the lead in providing cancer education on social media. Nevertheless, a significant number of studies have highlighted the prevalence of unreliable medical information on social media [19]. A Japanese study highlighted a concerning phenomenon: despite efforts by medical professionals to promote cancer screening online, a significant number of antiscreening activists disseminated contradictory messages on the internet, potentially undermining the effectiveness of cancer education initiatives [20]. Hence, there is an urgent need for the accurate dissemination of health information on social media, with greater involvement from scientists or professional institutions, to combat the spread of misinformation [21]. Despite efforts to study professional medical websites [22] and apps [23], there remains a lack of comprehensive understanding of the content posted on medical social media (MSM). Further study is thus needed to compare the differences between cancer information on social media from professional medical sources and nonprofessional sources to enhance cancer education.

For this study, we defined social media as internet-based platforms characterized by social interactive functions such as reading, commenting, retweeting, and timely interaction [24]. Based on this definition, we further classified 2 types of media based on ownership, content, and contributors: common social media (CSM) and MSM. MSM refers to social media platforms owned by professional medical institutions or organizations. It primarily provides medical and health information by medical professionals, including medical-focused accounts on social media and mobile health apps. CSM refers to social media owned or managed by individuals without medical backgrounds. It mainly provides health and lifestyle content.

Similar to Facebook (Meta Platforms, Inc.), WeChat (Tencent Holdings Limited) is the most popular social media platform in China, installed on more than 90% of smartphones. Zhang et al [25] has indicated that 63.26% of people prefer to obtain health information from WeChat. Unlike other Chinese social media platforms, WeChat has a broader user base that spans various age groups [26]. WeChat Public Accounts (WPAs) operate within the WeChat platform, offering services and information to the public. Many hospitals and primary care institutions in China have increasingly registered WPAs to provide health care services, medical information, health education, and more [27]. Therefore, this study selected WPA as the focus of research.

Based on big data analytics, this study aims to comprehensively map the characteristics of cancer treatment and prevention information on MSM and CSM, which could significantly enhance cancer coverage and assist patients in treatment decision-making. To address the aforementioned research gaps, 2 research questions were formulated.

- Research question 1: What are the characteristics of cancer prevention information discussed on social media? What are the differences between MSM and CSM?
- Research question 2: What are the characteristics of cancer treatment information discussed on social media? What are the differences between MSM and CSM?

XSL•FO
RenderX

Methods

Data Collection and Processing

We selected representative WPAs based on the reports from the "Ranking of Influential Health WeChat Public Accounts" [28] and the "2021 National Rankings of Best Hospitals by Specialty" [29]. In this study, we focused on 4 medical WPAs within MSM: Doctor Dingxiang (丁香医生), 91Huayi (华医网), The Cancer Hospital of Chinese Academy of Medical Sciences (中国医学科学院肿瘤医院), and Fudan University Shanghai Cancer Center (复旦大学附属肿瘤医院). We also selected 5 health and lifestyle WeChat Official Accounts classified as CSM for this study: Health Times (健康时报), Family Doctor (家庭医生), CCTV Lifestyle (CCTV 生活圈), Road to Health (健康之路), and Life Times (生命时报).

We implemented a Python-based (Python Foundation) crawler to retrieve posts from the aforementioned WPAs. Subsequently, we implemented a filtration process to eliminate noisy and unreliable data. Note that our focus is on WPAs that provide substantial information, defined as containing no fewer than a certain number of characters. We have deleted documents that contain less than 100 Chinese characters. Furthermore, we have removed figures and videos from the remaining documents. Eventually, we conducted an analysis at the paragraph level. According to our findings from random sampling, noise in articles from WPAs mostly originates from advertisements, which are typically found in specific paragraphs. Therefore, we retained only paragraphs that did not contain advertising keywords. In total, we collected 60,843 posts from these WPAs, comprising 20,654 articles from MSM and 40,189 articles from CSM.

The workflow chart in Figure 1 depicts all procedures following data collection and preprocessing. After obtaining meaningful raw documents, we performed word-level segmentation on the texts. We then removed insignificant stopwords and replaced specific types of cancers with a general term to facilitate coarse-grained latent Dirichlet allocation (LDA)–based filtering. Subsequently, we conducted fine-grained LDA topic modeling on the filtered documents without replacing keywords to visualize the topics extracted from the WPAs. Furthermore, we utilized a manually labeled codebook to train a long short-term memory (LSTM) network for document classification into various categories. Finally, we performed data analysis using both the topic distribution derived from fine-grained LDA and the classified documents.

XSL•FO
RenderX

Figure 1. Workflow chart.

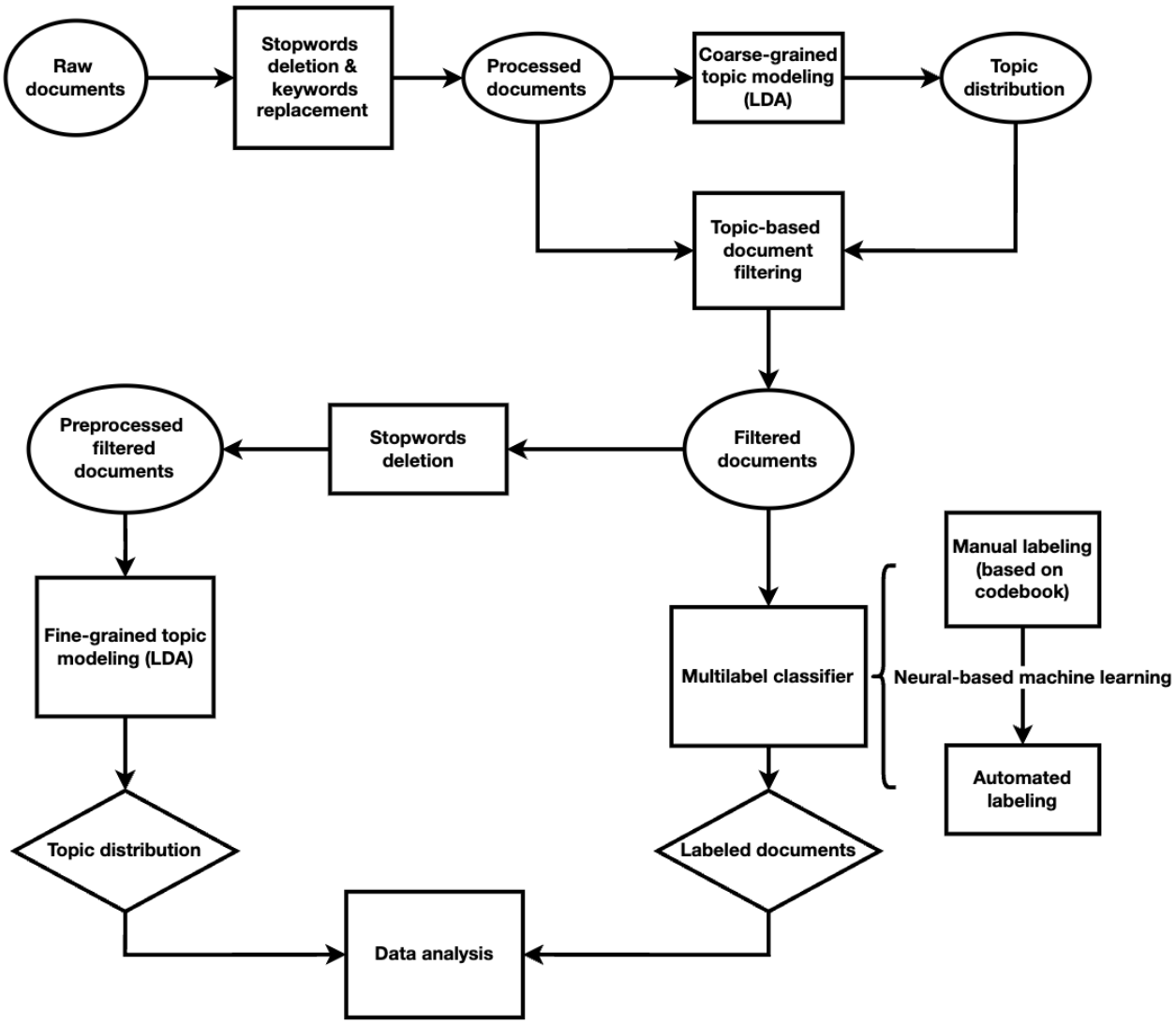

Latent Dirichlet Allocation Topic Modeling

Overview

LDA is a generative statistical model that explains sets of observations by latent groups, revealing why some parts of the data are similar [30]. The LDA algorithm can speculate on the topic distribution of a document.

When comparing LDA with other natural language processing methods such as LSTM-based deep learning, it is worth noting that LDA stands out as an unsupervised learning algorithm. Unlike its counterparts, LDA has the ability to uncover hidden topics without relying on labeled training data. Its strength lies in its capability to automatically identify latent topics within documents by analyzing statistical patterns of word co-occurrences. In addition, LDA provides interpretable outcomes by assigning a probability distribution to each document, representing its association with various topics. Similarly, it assigns a probability distribution to each topic, indicating the prevalence of specific words within that topic. This feature enables researchers to understand the principal themes present in their corpus and the extent to which these themes are manifested in individual documents.

The foundational principle of LDA involves using probabilistic inference to estimate the distribution of topics and word allocations. Specifically, LDA assumes that each document is composed of a mixture of a small number of topics, and each word's presence can be attributed to one of these topics. This approach allows for overlapping content among documents, rather than strict categorization into separate groups. For a deeper understanding of the technical and theoretical aspects of the LDA algorithm, readers are encouraged to refer to the research conducted by Blei et al [30]. In this context, our primary focus was on the application of the algorithm to our corpus, and the procedure is outlined in the following sections.

Document Selection

Initially, document selection involves using a methodological approach to sample documents from the corpus, which may include random selection or be guided by predetermined criteria such as document relevance or popularity within the social media context.

Topic Inference

Utilizing LDA or a similar topic modeling technique, we infer the underlying topical structure within each document. This

XSL•FO
RenderX

involves modeling documents as mixtures of latent topics represented by a Dirichlet distribution, from which topic proportions are sampled.

Topic Assignment to Words

After determining topic proportions, we proceed to assign topics to individual words in the document. Using a multinomial distribution, each word is probabilistically associated with one of the inferred topics based on the previously derived topic proportions.

Word Distribution Estimation

Each topic is characterized by a distinct distribution over the vocabulary, representing the likelihood of observing specific words within that topic. Using a Dirichlet distribution, we estimate the word distribution for each inferred topic.

Word Generation

Finally, using the multinomial distribution again, we generate words for the document by sampling from the estimated word distribution corresponding to the topic assigned to each word. This iterative process produces synthetic text that mirrors the statistical properties of the original corpus.

To filter out noncancer-related documents in our case, we replaced cancer-related words with “癌症” (cancer or tumor in Chinese) in all documents. We then conducted an LDA analysis to compute the topic distribution of each document and retained documents related to topics where “癌症” appears among the top 10 words.

In our study, we used Python packages such as jieba and gensim for document segmentation and extracting per-topic-per-word probabilities from the model. During segmentation, we applied a stopword dictionary to filter out meaningless words and transformed each document into a cleaned version containing only meaningful words.

During the LDA analysis, to determine the optimal number of topics, our main goal was to compute the topic coherence for various numbers of topics and select the model that yielded the highest coherence score. Coherence measures the interpretability of each topic by assessing whether the words within the same topic are logically associated with each other. The higher the score for a specific number *k*, the more closely related the words are within that topic. In this phase, we used the Python package pyLDAvis to compare coherence scores with different numbers of topics. Subsequently, we filtered and retained only the documents related to cancer topics, resulting in 4479 articles from MSM and 3948 articles from CSM.

Among the filtered articles, we conducted another LDA analysis to extract topics from the original articles without replacing cancer-related words. Using pyLDAvis, we calculated the coherence score and identified 6 topics for both MSM and CSM articles.

To visualize the topic modeling results, we created bar graphs where the y-axis indicates the top 10 keywords associated with each topic, and the x-axis represents the weight of each keyword (indicating its contribution to the topic). At the bottom of each graph (Figures 2 and 3), we generalized and presented the name of each topic based on the top 10 most relevant keywords.

Figure 2. Cancer topics on medical social media (MSM).

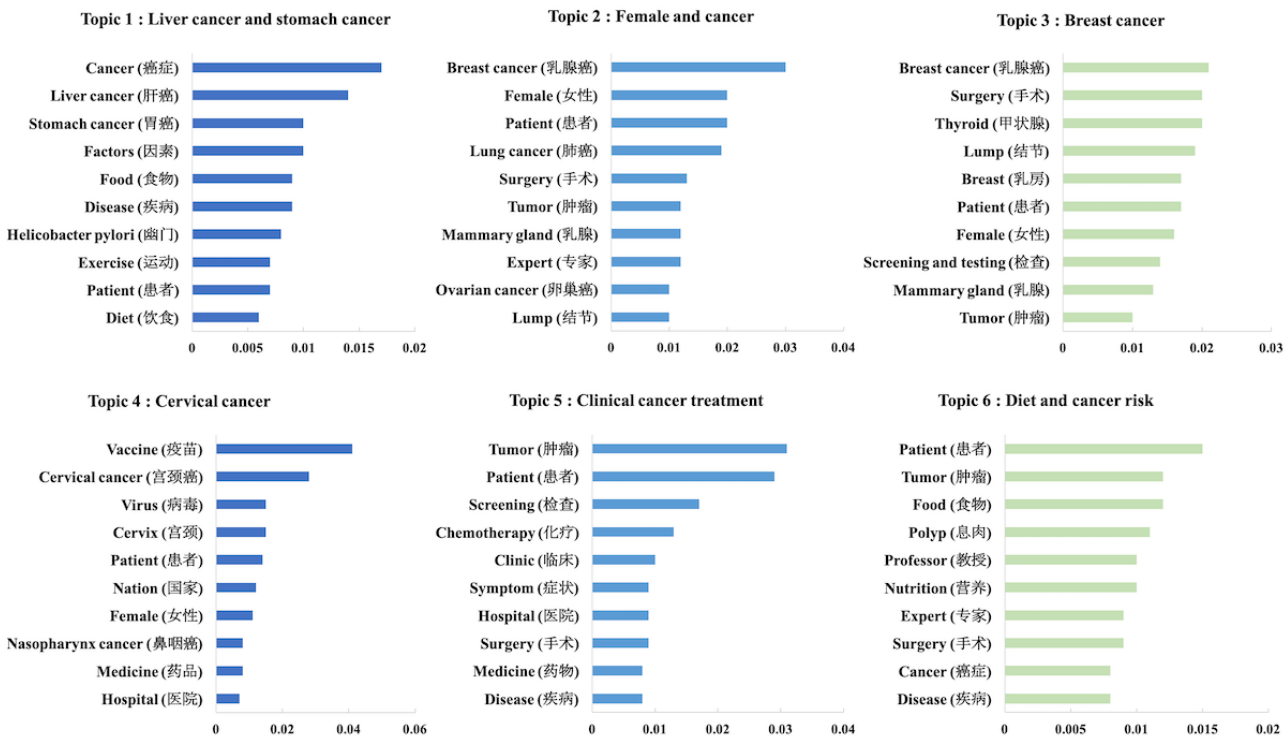

XSL•FO
RenderX

Manual Content Analysis: Coding Procedure

Based on the codebook, 2 independent coders (KZ and JL) engaged in discussions regarding the coding rules to ensure a shared understanding of the conceptual and operational distinctions among the coding items. To ensure the reliability of the coding process, both coders independently coded 100 randomly selected articles. Upon completion of the pilot coding, any disagreements were resolved through discussion between the 2 coders.

For the subsequent coding phase, each coder was assigned an equitable proportion of articles, with 10% of the cancer-related articles randomly sampled from both MSM samples (450/4479) and CSM samples (394/3948). Manual coding was performed on a total of 844 articles, which served as the training data set for the machine learning model. The operational definitions of each coding variable are detailed in Multimedia Appendix 1.

Coding Measures

Cancer Prevention Measures

Coders identified whether an article mentioned any of the following cancer prevention measures [31-35]: (1) avoid tobacco use, (2) maintain a healthy weight, (3) healthy diet, (4) exercise regularly, (5) limit alcohol use, (6) get vaccinated, (7) reduce exposure to ultraviolet radiation and ionizing radiation, (8) avoid urban air pollution and indoor smoke from household use of solid fuels, (9) early screening and detection, (10) breastfeeding, (11) controlling chronic infections, and (12) other prevention measures.

Cancer Treatment Measures

Coders identified whether an article mentioned any of the following treatments [36]: (1) surgery (including cryotherapy, lasers, hyperthermia, photodynamic therapy, cuts with scalpels), (2) radiotherapy, (3) chemotherapy, (4) immunotherapy, (5) targeted therapy, (6) hormone therapy, (7) stem cell transplant, (8) precision medicine, (9) cancer biomarker testing, and (10) other treatment measures.

Neural-Based Machine Learning

In this part, we attempted to label each article using a neural network. As mentioned earlier, we manually labeled 450 MSM articles and 394 CSM articles. We divided the labeled data into a training set and a test set with a ratio of 4:1. We adopted the pretrained Bidirectional Encoder Representations from Transformers (BERT) model. As BERT can only accept inputs with fewer than 512 tokens [37], we segmented each document into pieces of 510 tokens (accounting for BERT's automatic [CLS] and [SEP] tokens, where [CLS] denotes the start of a sentence or a document, and [SEP] denotes the end of a sentence or a document) with an overlap of 384 tokens between adjacent pieces. We began by utilizing a BERT-based encoder to encode each piece and predict its labels using a multioutput decoder. After predicting labels for each piece, we pooled the outputs for all pieces within the same document and used an LSTM network to predict final labels for each document.

Ethical Considerations

This study did not require institutional research board review as it did not involve interactions with humans or other living entities, private or personally identifiable information, or any pharmaceuticals or medical devices. The data set consists solely of publicly available social media posts.

Results

Cancer Topics on Social Media

Applying LDA, we identified 6 topics each for MSM and CSM articles. The distribution of topics among MSM and CSM is presented in Table 1, while the keyword weights for each topic are illustrated in Figures 2 and 3.

XSL•FO
RenderX

Table 1. Distribution of topics on medical social media and common social media (N=8427).a,b

Media type and topic number	Topic description	Articles, n (%)	Top 10 keywords
Medical social media			
Topic 1	Liver cancer and stomach cancer	1519 (18.03)	Cancer (癌症), liver cancer (肝癌), stomach cancer (胃癌), factors (因素), food (食物), disease (疾病), *Helicobacter pylori* (幽门), exercise (运动), patient (患者), and diet (饮食)
Topic 2	Female and cancer	1611 (19.12)	Breast cancer (乳腺癌), female (女性), patient (患者), lung cancer (肺癌), surgery (手术), tumor (肿瘤), mammary gland (乳腺), expert (专家), ovarian cancer (卵巢癌), and lump (结节)
Topic 3	Breast cancer	1093 (12.97)	Breast cancer (乳腺癌), surgery (手术), thyroid (甲状腺), lump (结节), breast (乳房), patient (患者), female (女性), screening and testing (检查), mammary gland (乳腺), and tumor (肿瘤)
Topic 4	Cervical cancer	1019 (12.09)	Vaccine (疫苗), cervical cancer (宫颈癌), virus (病毒), cervix (宫颈), patient (患者), nation (国家), female (女性), nasopharynx cancer (鼻咽癌), medicine (药品), and hospital (医院)
Topic 5	Clinical cancer treatment	2548 (30.24)	Tumor (肿瘤), patient (患者), screening (检查), chemotherapy (化疗), clinic (临床), symptom (症状), hospital (医院), surgery (手术), medicine (药物), and disease (疾病)
Topic 6	Diet and cancer risk	1741 (20.66)	Patient (患者), tumor (肿瘤), food (食物), polyp (息肉), professor (教授), nutrition (营养), expert (专家), surgery (手术), cancer (癌症), and disease (疾病)
Common social media			
Topic 1	Cancer-causing substances	1136 (13.48)	Foods (食物), nutrition (营养), carcinogen (致癌物), food (食品), ingredient (含量), vegetable (蔬菜), cancer (癌症), body (人体), lump (结节), and formaldehyde (甲醛)
Topic 2	Cancer treatment	1319 (15.65)	Patient (患者), cancer (癌症), hospital (医院), lung cancer (肺癌), tumor (肿瘤), medicine (药物), disease (疾病), professor (教授), surgery (手术), and clinic (临床)
Topic 3	Female and cancer risk	1599 (18.97)	Screening and testing (检查), female (女性), disease (疾病), breast cancer (乳腺癌), cancer (癌症), lung cancer (肺 癌), patient (患者), body (身体), tumor (肿瘤), and risk (风险)
Topic 4	Exercise, diet, and cancer risk	1947 (23.10)	Cancer (癌症), exercise (运动), food (食物), risk (风险), body (身体), disease (疾病), suggestion (建议), patient (患者), fat (脂肪), and hospital (医院)
Topic 5	Screening and diagnosis of cancer	1790 (21.24)	Screening and testing (检查), disease (疾病), hospital (医院), stomach cancer (胃癌), symptom (症状), patient (患者), cancer (癌症), liver cancer (肝癌), female (女性), and suggestion (建议)
Topic 6	Disease and body parts	869 (10.31)	Disease (疾病), intestine (肠道), food (食物), hospital (医院), oral cavity (口腔), patient (患者), teeth (牙齿), cancer (癌症), ovary (卵巢), and garlic (大蒜)

[a]In each article, different topics may appear at the same time. Therefore, the total frequency of each topic did not equate to the total number of 8427 articles.

[b]To ensure the accuracy of the results, directly translating sampled texts from Chinese into English posed challenges due to differences in semantic elements. In English, cancer screening refers to detecting the possibility of cancer before symptoms appear, while diagnostic tests confirm the presence of cancer after symptoms are observed. However, in Chinese, the term “检查” encompasses both meanings. Therefore, we translated it as both screening and testing.

XSL•FO
RenderX

Figure 3. Cancer topics on common social media (CSM).

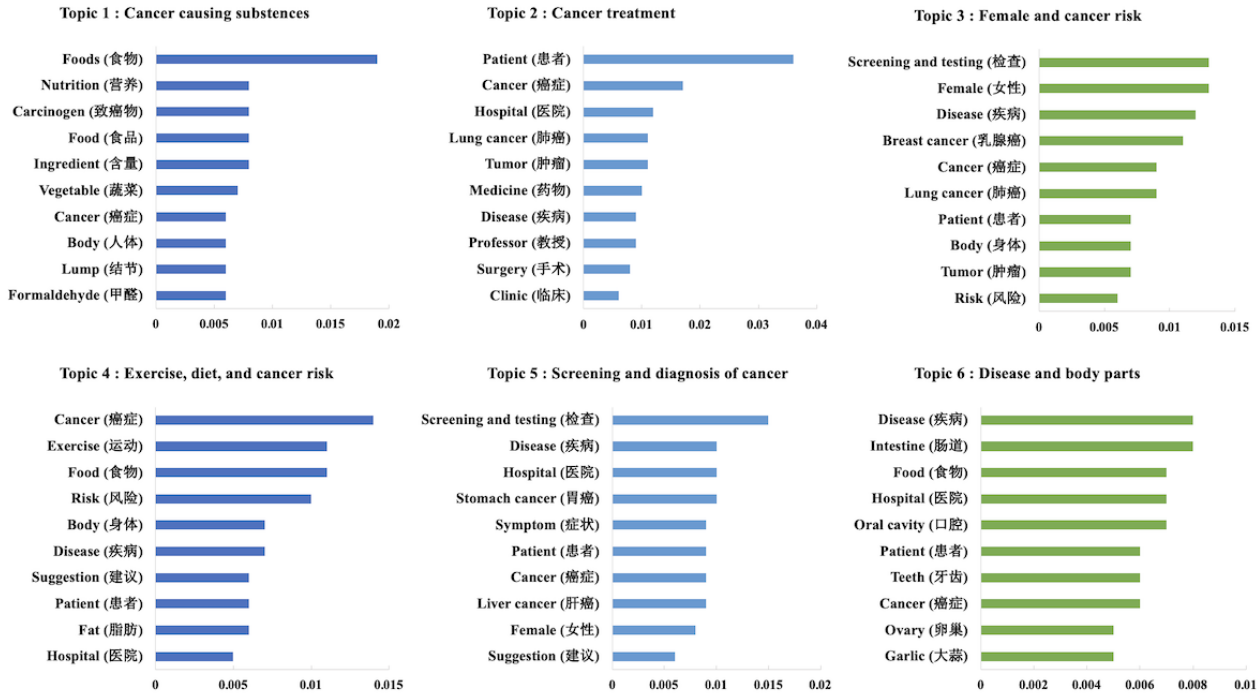

Among MSM articles, topic 5 was the most frequent (2548/8427, 30.24%), followed by topic 6 (1741/8427, 20.66%) and topic 2 (1611/8427, 19.12%). Both topics 5 and 6 focused on clinical treatments, with topic 5 specifically emphasizing cancer diagnosis. The keywords in topic 6, such as “polyp,” “tumor,” and “surgery,” emphasized the risk and diagnosis of precancerous lesions. Topic 2 primarily focused on cancer surgeries related to breast cancer, lung cancer, and ovarian cancer. The results indicate that MSM articles concentrated on specific cancers with higher incidence in China, including stomach cancer, liver cancer, lung cancer, breast cancer, and cervical cancer [10].

On CSM, topic 4 (1947/8427, 23.10%) had the highest proportion, followed by topic 5 (1790/8427, 21.24%) and topic 3 (1599/8427, 18.97%). Topic 6 had the smallest proportion. Topics 1 and 4 were related to lifestyle. Topic 1 particularly focused on cancer-causing substances, with keywords such as “food,” “nutrition,” and “carcinogen” appearing most frequently. Topic 4 was centered around exercise, diet, and their impact on cancer risk. Topics 3 and 5 were oriented toward cancer screening and diagnosis. Topic 3 specifically focused on female-related cancers, with discussions prominently featuring breast cancer screening and testing. Topic 5 emphasized early detection and diagnosis of stomach and lung cancers, highlighting keywords such as “screening” and “symptom.”

Cancer Prevention Information

Our experiment on the test set showed that the machine learning model achieved F_1-scores above 85 for both prevention and treatment categories in both MSM and CSM. For subclasses within prevention and treatment, we achieved F_1-scores of at least 70 for dense categories (with an occurrence rate >10%, ie, occurs in >1 of 10 entries) and at least 50 for sparse categories (with an occurrence rate <10%, ie, occurs in <1 of 10 entries). Subsequently, we removed items labeled as “other prevention measures” and “other treatment measures” due to semantic ambiguity.

Table 2 presents the distribution of cancer prevention information across MSM (n=4479) and CSM (n=3948).

XSL•FO
RenderX

Table 2. Distribution of cancer prevention information on MSMa and CSMb.

Type of cancer prevention measures	Number of articles on MSM (n=4479), n (%)	Number of articles on CSM (n=3948), n (%)
Articles containing prevention information	1137 (25.39)	1856 (47.01)
Early screening and testing	737 (16.45)	1085 (27.48)
Healthy diet	278 (6.21)	598 (15.15)
Get vaccinated	261 (5.83)	113 (2.86)
Avoid tobacco use	186 (4.15)	368 (9.32)
Exercise regularly	135 (3.01)	661 (16.74)
Limit alcohol use	128 (2.86)	281 (7.12)
Avoid urban air pollution and indoor smoke from household use of solid fuels	19 (0.42)	64 (1.62)
Maintain a healthy weight	18 (0.40)	193 (4.89)
Practice safe sex	12 (0.27)	4 (0.10)
Controlling chronic infections	3 (0.07)	32 (0.81)
Reduce exposure to radiation	2 (0.04)	1 (0.03)
Breastfeeding	1 (0.02)	1 (0.03)

[a]MSM: medical social media.

[b]CSM: common social media.

Cancer Prevention Information on MSM

The distribution of cancer prevention information on MSM (n=4479) is as follows: articles discussing prevention measures accounted for 25.39% (1137/4479) of all MSM cancer-related articles. The most frequently mentioned measure was “early screening and testing” (737/4479, 16.45%). The second and third most frequently mentioned prevention measures were “healthy diet” (278/4479, 6.21%) and “get vaccinated” (261/4479, 5.83%). The least mentioned prevention measures were “controlling chronic infections” (3/4479, 0.07%), “reduce exposure to radiation” (2/4479, 0.04%), and “breastfeeding” (1/4479, 0.02%), each appearing in only 1-3 articles.

Cancer Prevention Information on CSM

As many as 1856 out of 3948 (47.01%) articles on CSM referred to cancer prevention information. Among these, “early screening and testing” (1085/3948, 27.48%) was the most commonly mentioned prevention measure. “Exercise regularly” (661/3948, 16.74%) and “healthy diet” (598/3948, 15.15%) were the 2 most frequently mentioned lifestyle-related prevention measures. Additionally, “avoid tobacco use” accounted for 9.32% (368/3948) of mentions. Other lifestyle-related prevention measures were “limit alcohol use” (281/3948, 7.12%) and “maintain a healthy weight” (193/3948, 4.89%). The least mentioned prevention measures were “practice safe sex” (4/3948, 0.10%), “reduce exposure to radiation” (1/3948, 0.03%), and “breastfeeding” (1/3948, 0.03%), each appearing in only 1-4 articles.

Cancer Prevention Information on Social Media

Table 3 presents the overall distribution of cancer prevention information on social media (N=8427). Notably, CSM showed a stronger focus on cancer prevention (1856/3948, 47.01%) compared with MSM (1137/8427, 13.49%). Both platforms highlighted the importance of early screening and testing. However, MSM placed greater emphasis on vaccination as a prevention measure. In addition to lifestyle-related prevention measures, both CSM and MSM showed relatively less emphasis on avoiding exposure to environmental carcinogens, such as air pollution, indoor smoke, and radiation. “Breastfeeding” was the least mentioned prevention measure (2/3948, 0.05%) on both types of social media.

XSL•FO
RenderX

Table 3. Distribution of cancer prevention information on social media.

Type of cancer prevention measures	Number of articles on MSM[a], n (%)	Number of articles on CSM[b], n (%)	Number of articles overall (N=8427), n (%)
Articles containing prevention information	1137 (13.49)	1856 (22.02)	2993 (35.52)
Early screening and testing	737 (8.75)	1085 (12.88)	1822 (21.62)
Healthy diet	278 (3.30)	598 (7.10)	876 (10.40)
Get vaccinated	261 (3.10)	113 (1.34)	374 (4.44)
Avoid tobacco use	186 (2.21)	368 (4.37)	554 (6.57)
Exercise regularly	135 (1.60)	661 (7.84)	796 (9.45)
Limit alcohol use	128 (1.52)	281 (3.33)	409 (4.85)
Avoid urban air pollution and indoor smoke from household use of solid fuels	19 (0.23)	64 (0.76)	83 (0.98)
Maintain a healthy weight	18 (0.21)	193 (2.29)	211 (2.50)
Practice safe sex	12 (0.14)	4 (0.05)	16 (0.19)
Controlling chronic infections	3 (0.04)	32 (0.38)	35 (0.42)
Reduce exposure to radiation	2 (0.02)	1 (0.01)	3 (0.04)
Breastfeeding	1 (0.01)	1 (0.01)	2 (0.02)

[a]MSM: medical social media.
[b]CSM: common social media.

Cancer Treatment Information

Table 4 presents the distribution of cancer treatment information on MSM (n=4479) and CSM (n=3948).

Table 4. Distribution of cancer treatment information on MSMa and CSMb.

Type of cancer treatment measures	Number of articles on MSM (n=4479), n (%)	Number of articles on CSM (n=3948), n (%)
Articles containing treatment information	2966 (66.22)	778 (19.71)
Surgery	2045 (45.66)	419 (10.61)
Chemotherapy	1122 (25.05)	285 (7.22)
Radiation therapy	1108 (24.74)	232 (5.88)
Cancer biomarker testing	380 (8.48)	55 (1.39)
Targeted therapy	379 (8.46)	181 (4.58)
Immunotherapy	317 (7.08)	22 (0.56)
Hormone therapy	47 (1.05)	14 (0.35)
Stem cell transplantation therapy	5 (0.11)	0 (0)

[a]MSM: medical social media.
[b]CSM: common social media.

Cancer Treatment Information on MSM

Cancer treatment information appeared in 66.22% (2966/4479) of MSM posts. “Surgery” was the most frequently mentioned treatment measure (2045/4479, 45.66%), followed by “chemotherapy” (1122/4479, 25.05%) and “radiation therapy” (1108/4479, 24.74%). The proportions of “cancer biomarker testing” (380/4479, 8.48%), “targeted therapy” (379/4479, 8.46%), and “immunotherapy” (317/4479, 7.08%) were comparable. Only a minimal percentage of articles (47/4479, 1.05%) addressed “hormone therapy.” Furthermore, “stem cell transplantation therapy” was mentioned in just 5 out of 4479 (0.11%) articles.

Cancer Treatment Information on CSM

Cancer treatment information accounted for only 19.71% (778/3948) of CSM posts. “Surgery” was the most frequently mentioned treatment measure (419/3948, 10.61%), followed by “chemotherapy” (285/3948, 7.22%) and “radiation therapy” (232/3948, 5.88%). Relatively, the frequency of “targeted therapy” (181/3948, 4.58%) was similar to that of the first 3 types. However, “cancer biomarker testing” (55/3948, 1.39%),

XSL•FO
RenderX

"immunotherapy" (22/3948, 0.56%), and "hormone therapy" (14/3948, 0.35%) appeared rarely on CSM. Notably, there were no articles on CSM mentioning stem cell transplantation.

Cancer Treatment Information on Social Media

Table 5 shows the overall distribution of cancer treatment information on social media (N=8427). A total of 44.43% (3744/8427) of articles contained treatment information. MSM (2966/8427, 35.20%) discussed treatment information much more frequently than CSM (778/8427, 9.23%). Furthermore, the frequency of all types of treatment measures mentioned was higher on MSM than on CSM. The 3 most frequently mentioned types of treatment measures were surgery (2464/8427, 29.24%), chemotherapy (1407/8427, 16.70%), and radiation therapy (1340/8427, 15.90%). Relatively, MSM (380/8427, 4.51%) showed a higher focus on cancer biomarker testing compared with CSM (55/8427, 0.65%).

Table 5. Distribution of cancer treatment information on social media.

Type of cancer treatment measures	Number of articles on MSM[a], n (%)	Number of articles on CSM[b], n (%)	Number of articles overall (N=8427), n (%)
Articles containing treatment information	2966 (35.20)	778 (9.23)	3744 (44.43)
Surgery	2045 (24.27)	419 (4.97)	2464 (29.24)
Radiation therapy	1108 (13.15)	232 (2.75)	1340 (15.90)
Chemotherapy	1122 (13.31)	285 (3.38)	1407 (16.70)
Immunotherapy	317 (3.76)	22 (0.26)	339 (4.02)
Targeted therapy	379 (4.50)	181 (2.15)	560 (6.65)
Hormone therapy	47 (0.56)	14 (0.17)	61 (0.72)
Stem cell transplant	5 (0.06)	0 (0.00)	5 (0.06)
Cancer biomarker testing	380 (4.51)	55 (0.65)	435 (5.16)

[a]MSM: medical social media.

[b]CSM: common social media.

Discussion

Cancer Topics on MSM and CSM

In MSM, treatment-related topics constituted the largest proportion, featuring keywords related to medical examinations. Conversely, in CSM, the distribution of topics appeared more balanced, with keywords frequently associated with cancer risk and screening. Overall, the distribution of topics on MSM and CSM revealed that CSM placed greater emphasis on lifestyle factors and early screening and testing. Specifically, CSM topics focused more on early cancer screening and addressed cancer types with high incidence rates. By contrast, MSM topics centered more on clinical treatment, medical testing, and the cervical cancer vaccine in cancer prevention. Additionally, MSM focused on types of cancers that are easier to screen and prevent, including liver cancer, stomach cancer, breast cancer, cervical cancer, and colon cancer.

Cancer Prevention Information on MSM and CSM

Through content analysis, it was found that 35.52% (2993/8427) of articles on social media contained prevention information, and 44.43% (3744/8427) contained treatment information. Compared with MSM (1137/8427, 13.49%), CSM (2993/8427, 35.52%) focused more on prevention.

Primary prevention mainly involves adopting healthy behaviors to lower the risk of developing cancer, which has been proven to have long-term effects on cancer prevention. Secondary prevention focuses on inhibiting or reversing carcinogenesis, including early screening and detection, as well as the treatment or removal of precancerous lesions [38]. Compared with cancer screening and treatment, primary prevention is considered the most cost-effective approach to reducing the cancer burden.

From our results, "early screening and testing" (1822/8427, 21.62%) was the most frequently mentioned prevention measure on both MSM and CSM. According to a cancer study from China, behavioral risk factors were identified as the primary cause of cancer [10]. However, measures related to primary prevention were not frequently mentioned. Additionally, lifestyle-related measures such as "healthy diet," "regular exercise," "avoiding tobacco use," and "limiting alcohol use" were mentioned much less frequently on MSM compared with CSM.

Furthermore, "avoiding tobacco use" (554/8427, 6.57%) and "limiting alcohol use" (409/8427, 4.85%) were rarely mentioned, despite tobacco and alcohol being the leading causes of cancer. In China, public policies on the production, sale, and consumption of alcohol are weaker compared with Western countries. Notably, traditional Chinese customs often promote the belief that moderate drinking is beneficial for health [39]. Moreover, studies indicated that the smoking rate among adult men exceeded 50% in 2015. By 2018, 25.6% of Chinese adults aged 18 and above were smokers, totaling approximately 282 million smokers in China (271 million males and 11 million females) [40]. These statistics align with the consistently high incidence of lung cancer among Chinese men [41]. Simultaneously, the incidence and mortality of lung cancer in Chinese women were more likely associated with exposure to second-hand smoke or occupation-related risk factors.

XSL•FO
RenderX

Although MSM (261/8427, 3.10%) mentioned vaccination more frequently than CSM (113/8427, 1.34%), vaccination was not widely discussed on social media overall (374/8427, 4.44%). The introduction of human papillomavirus vaccination in China has lagged for more than 10 years compared with Western countries. A bivalent vaccine was approved by the Chinese Food and Drug Administration in 2017 but has not been included in the national immunization schedules up to now [42].

According to the "European Code Against Cancer" [43], breastfeeding is recommended as a measure to prevent breast cancer. However, there were no articles mentioning the role of breastfeeding in preventing breast cancer on social media.

One of the least frequently mentioned measures was "radiation protection," which includes sun protection. Although skin cancer is not as common in China as in Western countries, China has the largest population in the world. A study showed that only 55.2% of Chinese people knew that ultraviolet radiation causes skin cancer [33]. Additional efforts should be made to enhance public awareness of skin cancer prevention through media campaigns.

Overall, our results indicate that social media, especially MSM, focused more on secondary prevention. The outcomes of primary prevention are challenging to identify in individuals, and studies on cancer education may partly explain why primary prevention was often overlooked [44].

Cancer Treatment Information on MSM and CSM

Compared with a related content analysis study in the United States, our findings also indicate that the media placed greater emphasis on treatment [45]. Treatment information on MSM was more diverse than on CSM, with a higher proportion of the 3 most common cancer treatments—surgery, chemotherapy, and radiation therapy—mentioned on MSM compared with CSM. Notably, CSM (232/8427, 2.75%) mentioned radiation therapy less frequently compared with MSM (1108/8427, 13.15%), despite it being one of the most common cancer treatment measures in clinical practice.

In addition to common treatment methods, other approaches such as targeted therapy (560/8427, 6.65%) and immunotherapy (339/8427, 4.02%) were rarely discussed. This could be attributed to the high costs associated with these treatments. A study revealed that each newly diagnosed patient with cancer in China faced out-of-pocket expenses of US $4947, amounting to 57.5% of the family's annual income, posing an unaffordable economic burden of 77.6% [46]. In 2017, the Chinese government released the National Health Insurance Coverage (NHIC) policy to improve the accessibility and affordability of innovative anticancer medicines, leading to reduced prices and increased availability and utilization of 15 negotiated drugs. However, a study indicated that the availability of these innovative anticancer drugs remained limited. By 2019, the NHIC policy had benefited 44,600 people, while the number of new cancer cases in China in 2020 was 4.57 million [47]. The promotion of information on innovative therapies helped patients gain a better understanding of their cancer treatment options [48].

Practical Implications

This research highlighted that MSM did not fully leverage its professional background in providing comprehensive cancer information to the public. In fact, MSM holds substantial potential for contributing to cancer education. The findings from the content analysis also have practical implications for practitioners. They provide valuable insights for experts to assess the effectiveness of social media, monitor the types of information available to the public and patients with cancer, and guide communication and medical professionals in crafting educational and persuasive messages based on widely covered or less attended content.

Limitations and Future Directions

This study had some limitations. First, we only collected 60,843 articles from 9 WPAs in China. Future research could broaden the scope by collecting data from diverse countries and social media platforms. Second, our manual labeling only extracted 10% (450/4479 for MSM and 394/3948 for CSM) of the samples; the accuracy of the machine learning model could be enhanced by training it with a larger set of labeled articles. Finally, our results only represented the media's presentation, and the impact of this information on individuals remains unclear. Further work could examine its influence on behavioral intentions or actions related to cancer prevention among the audience.

Conclusions

The analysis of cancer-related information on social media revealed an imbalance between prevention and treatment content. Overall, there was more treatment information than prevention information. Compared with MSM, CSM mentioned more prevention information. On MSM, the proportion of treatment information was greater than prevention information, whereas on CSM, the 2 were equal. The focus on cancer prevention and treatment information was primarily limited to a few aspects, with a predominant emphasis on secondary prevention rather than primary prevention. There is a need for further improvement in the coverage of prevention measures and treatments for cancer on social media. Additionally, the findings underscored the potential of applying machine learning to content analysis as a promising research paradigm for mapping key dimensions of cancer information on social media. These findings offer methodological and practical significance for future studies and health promotion.

Acknowledgments

This study was funded by The Major Program of the Chinese National Foundation of Social Sciences under the project "The Challenge and Governance of Smart Media on News Authenticity" (grant number 23&ZD213).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Definitions and descriptions of coding items.
[PDF File (Adobe PDF File), 41 KB - jmir_v26i1e55937_app1.pdf]

References

1. International Agency for Research on Cancer (IARC), World Health Organization (WHO). Cancer today: the global cancer observatory. IARC. Geneva, Switzerland: WHO; 2020. URL: https://gco.iarc.who.int/today/en [accessed 2023-12-25]
2. Yu S, Yang CS, Li J, You W, Chen J, Cao Y, et al. Cancer prevention research in China. Cancer Prev Res (Phila) 2015 Aug;8(8):662-674. [doi: 10.1158/1940-6207.CAPR-14-0469] [Medline: 26076697]
3. Xia C, Dong X, Li H, Cao M, Sun D, He S, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl) 2022 Feb 09;135(5):584-590 [FREE Full text] [doi: 10.1097/CM9.0000000000002108] [Medline: 35143424]
4. World Health Organization (WHO). Cancer. WHO. 2023. URL: https://www.who.int/news-room/facts-in-pictures/detail/cancer [accessed 2023-12-27]
5. Pagoto S, Waring ME, Xu R. A call for a public health agenda for social media research. J Med Internet Res 2019 Dec 19;21(12):e16661 [FREE Full text] [doi: 10.2196/16661] [Medline: 31855185]
6. Tekeli-Yesil S, Tanner M. Understanding the contribution of conventional media in earthquake risk communication. J Emerg Manag Disaster Commun 2024 Jun 01;05(01):111-133. [doi: 10.1142/s2689980924500052]
7. Jensen JD, Scherr CL, Brown N, Jones C, Christy K, Hurley RJ. Public estimates of cancer frequency: cancer incidence perceptions mirror distorted media depictions. J Health Commun 2014;19(5):609-624. [doi: 10.1080/10810730.2013.837551] [Medline: 24446814]
8. Banaye Yazdipour A, Niakan Kalhori SR, Bostan H, Masoorian H, Ataee E, Sajjadi H. Effect of social media interventions on the education and communication among patients with cancer: a systematic review protocol. BMJ Open 2022 Nov 30;12(11):e066550 [FREE Full text] [doi: 10.1136/bmjopen-2022-066550] [Medline: 36450436]
9. Wallner LP, Martinez KA, Li Y, Jagsi R, Janz NK, Katz SJ, et al. Use of online communication by patients with newly diagnosed breast cancer during the treatment decision process. JAMA Oncol 2016 Dec 01;2(12):1654-1656 [FREE Full text] [doi: 10.1001/jamaoncol.2016.2070] [Medline: 27468161]
10. Sun D, Li H, Cao M, He S, Lei L, Peng J, et al. Cancer burden in China: trends, risk factors and prevention. Cancer Biol Med 2020 Nov 15;17(4):879-895 [FREE Full text] [doi: 10.20892/j.issn.2095-3941.2020.0387] [Medline: 33299641]
11. Basch CH, Menafro A, Mongiovi J, Hillyer GC, Basch CE. A content analysis of YouTube videos related to prostate cancer. Am J Mens Health 2017 Jan;11(1):154-157 [FREE Full text] [doi: 10.1177/1557988316671459] [Medline: 27694554]
12. Vasconcelos Silva C, Jayasinghe D, Janda M. What can Twitter tell us about skin cancer communication and prevention on social media? Dermatology 2020;236(2):81-89 [FREE Full text] [doi: 10.1159/000506458] [Medline: 32097934]
13. Hurley RJ, Riles JM, Sangalang A. Online cancer news: trends regarding article types, specific cancers, and the cancer continuum. Health Commun 2014;29(1):41-50. [doi: 10.1080/10410236.2012.715538] [Medline: 23356571]
14. Mishel MH, Germino BB, Lin L, Pruthi RS, Wallen EM, Crandell J, et al. Managing uncertainty about treatment decision making in early stage prostate cancer: a randomized clinical trial. Patient Educ Couns 2009 Dec;77(3):349-359. [doi: 10.1016/j.pec.2009.09.009] [Medline: 19819096]
15. Brown P, Kwan V, Vallerga M, Obhi HK, Woodhead EL. The use of anecdotal information in a hypothetical lung cancer treatment decision. Health Commun 2019 Jun;34(7):713-719. [doi: 10.1080/10410236.2018.1433415] [Medline: 29393661]
16. Crannell WC, Clark E, Jones C, James TA, Moore J. A pattern-matched Twitter analysis of US cancer-patient sentiments. J Surg Res 2016 Dec;206(2):536-542. [doi: 10.1016/j.jss.2016.06.050] [Medline: 27523257]
17. Gage-Bouchard EA, LaValley S, Mollica M, Beaupin LK. Cancer communication on social media: examining how cancer caregivers use Facebook for cancer-related communication. Cancer Nurs 2017;40(4):332-338. [doi: 10.1097/NCC.0000000000000418] [Medline: 27442210]
18. Reid BB, Rodriguez KN, Thompson MA, Matthews GD. Cancer-specific Twitter conversations among physicians in 2014. JCO 2015 May 20;33(15_suppl):e17500. [doi: 10.1200/jco.2015.33.15_suppl.e17500]
19. Warner EL, Waters AR, Cloyes KG, Ellington L, Kirchhoff AC. Young adult cancer caregivers' exposure to cancer misinformation on social media. Cancer 2021 Apr 15;127(8):1318-1324 [FREE Full text] [doi: 10.1002/cncr.33380] [Medline: 33368215]
20. Okuhara T, Ishikawa H, Okada M, Kato M, Kiuchi T. Assertions of Japanese websites for and against cancer screening: a text mining analysis. Asian Pac J Cancer Prev 2017 Apr 01;18(4):1069-1075 [FREE Full text] [doi: 10.22034/APJCP.2017.18.4.1069] [Medline: 28547943]

XSL•FO
RenderX

21. Qin L, Zhang X, Wu A, Miser JS, Liu Y, Hsu JC, et al. Association between social media use and cancer screening awareness and behavior for people without a cancer diagnosis: matched cohort study. J Med Internet Res 2021 Aug 27;23(8):e26395 [FREE Full text] [doi: 10.2196/26395] [Medline: 34448708]
22. Denecke K, Nejdl W. How valuable is medical social media data? Content analysis of the medical web. Information Sciences 2009 May 30;179(12):1870-1880. [doi: 10.1016/j.ins.2009.01.025]
23. Bender JL, Yue RYK, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. J Med Internet Res 2013 Dec 23;15(12):e287 [FREE Full text] [doi: 10.2196/jmir.2661] [Medline: 24366061]
24. Li X, Liu Q. Social media use, eHealth literacy, disease knowledge, and preventive behaviors in the COVID-19 pandemic: cross-sectional study on Chinese netizens. J Med Internet Res 2020 Oct 09;22(10):e19684 [FREE Full text] [doi: 10.2196/19684] [Medline: 33006940]
25. Zhang X, Wen D, Liang J, Lei J. How the public uses social media wechat to obtain health information in China: a survey study. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):66 [FREE Full text] [doi: 10.1186/s12911-017-0470-0] [Medline: 28699549]
26. Elad B. WeChat statistics by device allocation, active users, country wise traffic, demographics and marketing channels, social media traffic. EnterpriseAppsToday. 2023. URL: https://www.enterpriseappstoday.com/stats/wechat-statistics.html [accessed 2023-12-26]
27. Liang X, Yan M, Li H, Deng Z, Lu Y, Lu P, et al. WeChat official accounts' posts on medication use of 251 community healthcare centers in Shanghai, China: content analysis and quality assessment. Front Med (Lausanne) 2023;10:1155428 [FREE Full text] [doi: 10.3389/fmed.2023.1155428] [Medline: 37378298]
28. NewRank. Ranking of influential health WeChat public accounts（中国健康类微信影响力排行榜）. NewRank（新榜）. 2018. URL: https://newrank.cn/public/info/rank_detail.html?name=health [accessed 2021-04-30]
29. Hospital Management Institute of Fudan University. 2021 National rankings of best hospitals by oncology specialty （2021年度肿瘤科专科声誉排行榜）. Hospital Management Institute of Fudan University. 2021. URL: https://rank.cn-healthcare.com/fudan/specialty-reputation/year/2021/sid/2 [accessed 2021-05-01]
30. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research 2003;3:993-1022 [FREE Full text]
31. World Health Organization (WHO). Health topic: cancer. WHO. URL: https://www.who.int/health-topics/cancer#tab=tab_2 [accessed 2023-12-27]
32. Moore SC, Lee I, Weiderpass E, Campbell PT, Sampson JN, Kitahara CM, et al. Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults. JAMA Intern Med 2016 Jun 01;176(6):816-825 [FREE Full text] [doi: 10.1001/jamainternmed.2016.1548] [Medline: 27183032]
33. Stephens P, Martin B, Ghafari G, Luong J, Nahar V, Pham L, et al. Skin cancer knowledge, attitudes, and practices among Chinese population: a narrative review. Dermatol Res Pract 2018;2018:1965674 [FREE Full text] [doi: 10.1155/2018/1965674] [Medline: 29973951]
34. International Agency for Research on Cancer (IARC). Agents classified by the IARC monographs, volumes 1–136. IARC. URL: https://monographs.iarc.who.int/agents-classified-by-the-iarc/ [accessed 2023-12-25]
35. Han CJ, Lee YJ, Demiris G. Interventions using social media for cancer prevention and management. Cancer Nurs 2018;41(6):E19-E31. [doi: 10.1097/ncc.0000000000000534]
36. National Institutes of Health (NIH), National Cancer Institute (NCI). Types of cancer treatment. NIH. URL: https://www.cancer.gov/about-cancer/treatment/types [accessed 2021-03-15]
37. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans Audio Speech Lang Process 2021;29:3504-3514. [doi: 10.1109/taslp.2021.3124365]
38. Loomans-Kropp HA, Umar A. Cancer prevention and screening: the next step in the era of precision medicine. NPJ Precis Oncol 2019;3:3 [FREE Full text] [doi: 10.1038/s41698-018-0075-9] [Medline: 30701196]
39. Tang Y, Xiang X, Wang X, Cubells JF, Babor TF, Hao W. Alcohol and alcohol-related harm in China: policy changes needed. Bull World Health Organ 2013 Jan 22;91(4):270-276. [doi: 10.2471/blt.12.107318]
40. Zhang M, Yang L, Wang L, Jiang Y, Huang Z, Zhao Z, et al. Trends in smoking prevalence in urban and rural China, 2007 to 2018: findings from 5 consecutive nationally representative cross-sectional surveys. PLoS Med 2022 Aug;19(8):e1004064 [FREE Full text] [doi: 10.1371/journal.pmed.1004064] [Medline: 36006870]
41. Li J, Wu B, Selbæk G, Krokstad S, Helvik A. Factors associated with consumption of alcohol in older adults - a comparison between two cultures, China and Norway: the CLHLS and the HUNT-study. BMC Geriatr 2017 Jul 31;17(1):172 [FREE Full text] [doi: 10.1186/s12877-017-0562-9] [Medline: 28760157]
42. Feng R, Zong Y, Cao S, Xu R. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? Cancer Commun (Lond) 2019 Apr 29;39(1):22 [FREE Full text] [doi: 10.1186/s40880-019-0368-6] [Medline: 31030667]
43. Scoccianti C, Key TJ, Anderson AS, Armaroli P, Berrino F, Cecchini M, et al. European code against cancer 4th Edition: breastfeeding and cancer. Cancer Epidemiol 2015 Dec;39 Suppl 1:S101-S106 [FREE Full text] [doi: 10.1016/j.canep.2014.12.007] [Medline: 26116994]

XSL•FO
RenderX

44. Espina C, Porta M, Schüz J, Aguado IH, Percival RV, Dora C, et al. Environmental and occupational interventions for primary prevention of cancer: a cross-sectorial policy framework. Environ Health Perspect 2013 Apr;121(4):420-426 [FREE Full text] [doi: 10.1289/ehp.1205897] [Medline: 23384642]
45. Jensen JD, Moriarty CM, Hurley RJ, Stryker JE. Making sense of cancer news coverage trends: a comparison of three comprehensive content analyses. J Health Commun 2010 Mar;15(2):136-151. [doi: 10.1080/10810730903528025] [Medline: 20390983]
46. Huang H, Shi J, Guo L, Zhu X, Wang L, Liao X, et al. Expenditure and financial burden for common cancers in China: a hospital-based multicentre cross-sectional study. The Lancet 2016 Oct;388:S10. [doi: 10.1016/S0140-6736(16)31937-7]
47. People's Daily. 17 Cancer drugs included in medical insurance at reduced prices, reducing medication costs by over 75% (17种抗癌药降价进医保减轻药费负担超75%). People's Daily. 2019. URL: http://www.gov.cn/xinwen/2019-02/13/content_5365211.htm [accessed 2023-12-25]
48. Fang W, Xu X, Zhu Y, Dai H, Shang L, Li X. Impact of the National Health Insurance Coverage Policy on the Utilisation and Accessibility of Innovative Anti-cancer Medicines in China: An Interrupted Time-Series Study. Front Public Health 2021;9:714127 [FREE Full text] [doi: 10.3389/fpubh.2021.714127] [Medline: 34422752]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CSM: common social media
GloVe: Global Vectors for Word Representation
LDA: latent Dirichlet allocation
LSTM: long short-term memory
MSM: medical social media
NHIC: National Health Insurance Coverage
WHO: World Health Organization
WPA: WeChat public account

Edited by S Ma; submitted 02.01.24; peer-reviewed by F Yang, D Wawrzuta; comments to author 20.03.24; revised version received 19.04.24; accepted 03.06.24; published 14.08.24.

Please cite as:
Zhao K, Li X, Li J
Cancer Prevention and Treatment on Chinese Social Media: Machine Learning–Based Content Analysis Study
J Med Internet Res 2024;26:e55937
URL: https://www.jmir.org/2024/1/e55937
doi:10.2196/55937
PMID:

©Keyang Zhao, Xiaojing Li, Jingyang Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigation of Public Acceptance of Misinformation Correction in Social Media Based on Sentiment Attributions: Infodemiology Study Using Aspect-Based Sentiment Analysis

Ning Ma[1], DM; Guang Yu[1], Prof Dr; Xin Jin[2], JSD

[1]School of Management, Harbin Institute of Technology, Harbin, China

[2]School of Social Sciences, Harbin Institute of Technology, Harbin, China

Corresponding Author:
Guang Yu, Prof Dr
School of Management
Harbin Institute of Technology
92 West Dazhi Street
Harbin, 150001
China
Phone: 86 139 3649 0774
Email: yug@hit.edu.cn

Abstract

Background: The proliferation of misinformation on social media is a significant concern due to its frequent occurrence and subsequent adverse social consequences. Effective interventions for and corrections of misinformation have become a focal point of scholarly inquiry. However, exploration of the underlying causes that affect the public acceptance of misinformation correction is still important and not yet sufficient.

Objective: This study aims to identify the critical attributions that influence public acceptance of misinformation correction by using attribution analysis of aspects of public sentiment, as well as investigate the differences and similarities in public sentiment attributions in different types of misinformation correction.

Methods: A theoretical framework was developed for analysis based on attribution theory, and public sentiment attributions were divided into 6 aspects and 11 dimensions. The correction posts for the 31 screened misinformation events comprised 33,422 Weibo posts, and the corresponding Weibo comments amounted to 370,218. A pretraining model was used to assess public acceptance of misinformation correction from these comments, and the aspect-based sentiment analysis method was used to identify the attributions of public sentiment response. Ultimately, this study revealed the causality between public sentiment attributions and public acceptance of misinformation correction through logistic regression analysis.

Results: The findings were as follows: First, public sentiments attributed to external attribution had a greater impact on public acceptance than those attributed to internal attribution. The public associated different aspects with correction depending on the type of misinformation. The accuracy of the correction and the entity responsible for carrying it out had a significant impact on public acceptance of misinformation correction. Second, negative sentiments toward the media significantly increased, and public trust in the media significantly decreased. The collapse of media credibility had a detrimental effect on the actual effectiveness of misinformation correction. Third, there was a significant difference in public attitudes toward the official government and local governments. Public negative sentiments toward local governments were more pronounced.

Conclusions: Our findings imply that public acceptance of misinformation correction requires flexible communication tailored to public sentiment attribution. The media need to rebuild their image and regain public trust. Moreover, the government plays a central role in public acceptance of misinformation correction. Some local governments need to repair trust with the public. Overall, this study offered insights into practical experience and a theoretical foundation for controlling various types of misinformation based on attribution analysis of public sentiment.

(J Med Internet Res 2024;26:e50353) doi:10.2196/50353

XSL•FO
RenderX

KEYWORDS

misinformation correction; sentiment attribution; public acceptance; public sentiments; aspect-based sentiment analysis; pretraining model

Introduction

Background

Scholarly interest in researching misinformation dissemination has been extensive [1,2], particularly given the exacerbation of this issue since the COVID-19 pandemic. Such misinformation encompasses a range of forms, including the spread of health misinformation [3], fake news [4], conspiracy theories [5], and other types of COVID-19 misinformation [6]. The spread of misinformation can significantly hinder individuals' access to accurate information and may exacerbate psychological distress [7]. In addition, the public may lose faith in the government, leading to increased fear and anxiety [8]. The resulting infodemic has heightened concerns about the adverse effects of misinformation [9,10] and prompted greater research attention toward misinformation and infodemic governance [11].

Research on how to mitigate the spread of misinformation has resulted in several recommended interventions [12]. The intervention and prevention strategies for misinformation are also key research points [13]. Fact checking was believed to be an effective approach in reducing the spread of misinformation, but it often requires extensive coordination to improve public resistance [14]. Applying warning tags to social media posts could also potentially reduce sharing behaviors [15], but the credibility of untagged posts could be potentially enhanced [16]. Inaccurate tagging can also aggravate the spread of misinformation, making warning tags ineffective [17]. Another effective approach is to post correction information that directly addresses false facts or claims [18], with authoritative experts playing a prominent role in correction communication [9,19,20]. Using social media platforms and the structure of social networks to block misinformation is also an effective approach [21], and correcting information in real time through social media platforms is another strategy [22].

Preventive measures are also necessary to combat misinformation among the public [12]. The most common framework for implementing preventive interventions is the psychological inoculation theory [23]; it had lasting effects in responding to misinformation about COVID-19 [24]. Furthermore, educational preventive measures are thought to be the most effective [25,26], and personal awareness of misinformation is equally significant [27]. To implement reliable preventive measures on a large scale, we need more professionals with advanced knowledge and convergence of many efforts to yield reliable results [12]. Therefore, government policy support is essential [28], and a clear legal framework on misinformation needs to be established promptly [29]. However, both debunking [30] and prebunking, which refers to prophylactic interventions [23], have limitations and may not be as effective as anticipated [12]. While scholars have proposed suggestions to improve the correction effect [31], sometimes misinformation correction can even backfire [32].

Although researchers have made concerted efforts to correct misinformation on social networks, the results have been suboptimal in many cases [33]. This not only perpetuates misinformation but also risks public resentment [34], undermining the persuasive power of subsequent corrections. Current research on evaluating the effectiveness of misinformation correction through social media has analyzed the impact of various post characteristics on correction outcomes [35]. As emotional information can provide valuable insights into understanding the public perception of misinformation [36-39], more research is needed on how public sentiment can be leveraged to analyze the reasons for poor correction effectiveness. Analyzing public sentiment can help identify public attitudes toward misinformation [34,40], including the effectiveness of correction. Considering the limitations regarding objectivity and timeliness of using questionnaires to assess public attitudes toward misinformation and correction [41], social media data can be used more objectively to analyze public acceptance of misinformation correction [42]. Governments can also use public sentiment analysis to combat misinformation and inappropriate sharing behaviors [43].

Objectives

It is vital to understand how to evaluate public acceptance of misinformation correction using public sentiment analysis on social media. To address this research goal, this study focused on misinformation correction during the COVID-19 pandemic and mining the key factors influencing public acceptance of correction. Specifically, the following three research questions were answered in this study:

1. How to characterize public acceptance of misinformation correction and different aspects of public sentiment? (Research question 1)
2. What are the differences in the impact of various public sentiment attributions on the correction of different types of misinformation? (Research question 2)
3. Which attribution of public sentiment is more critical to shaping public acceptance of correction of different types of misinformation? (Research question 3)

To comprehensively explore the aforementioned questions, this study collected public interaction data on the Weibo platform pertaining to important misinformation dissemination events related to COVID-19 since the outbreak. First, aspect-based sentiment analysis (ABSA) was performed to attribute public sentiments to multiple dimensions drawing on attribution theory. Second, the study explored how public acceptance of misinformation correction related to public sentiment attributions in different events from the public perspective. Finally, the significant degree of attribution results obtained through public sentiment response revealed the specific reasons for public attitudes. Consequently, this study provides theoretical support and practical evidence for developing correction strategies for misinformation, which could positively impact the formulation of correction communication measures.

XSL•FO
RenderX

Methods

Research Design and Framework

This study proposed a research framework, as depicted in Figure 1, based on attribution theory to evaluate public acceptance of misinformation correction by analyzing the attribution results of public sentiment response to correction.

Figure 1. Research and analysis framework of public acceptance of misinformation correction based on attribution theory. The framework includes the specific steps of data collection, model training, causality analysis, and other processes.

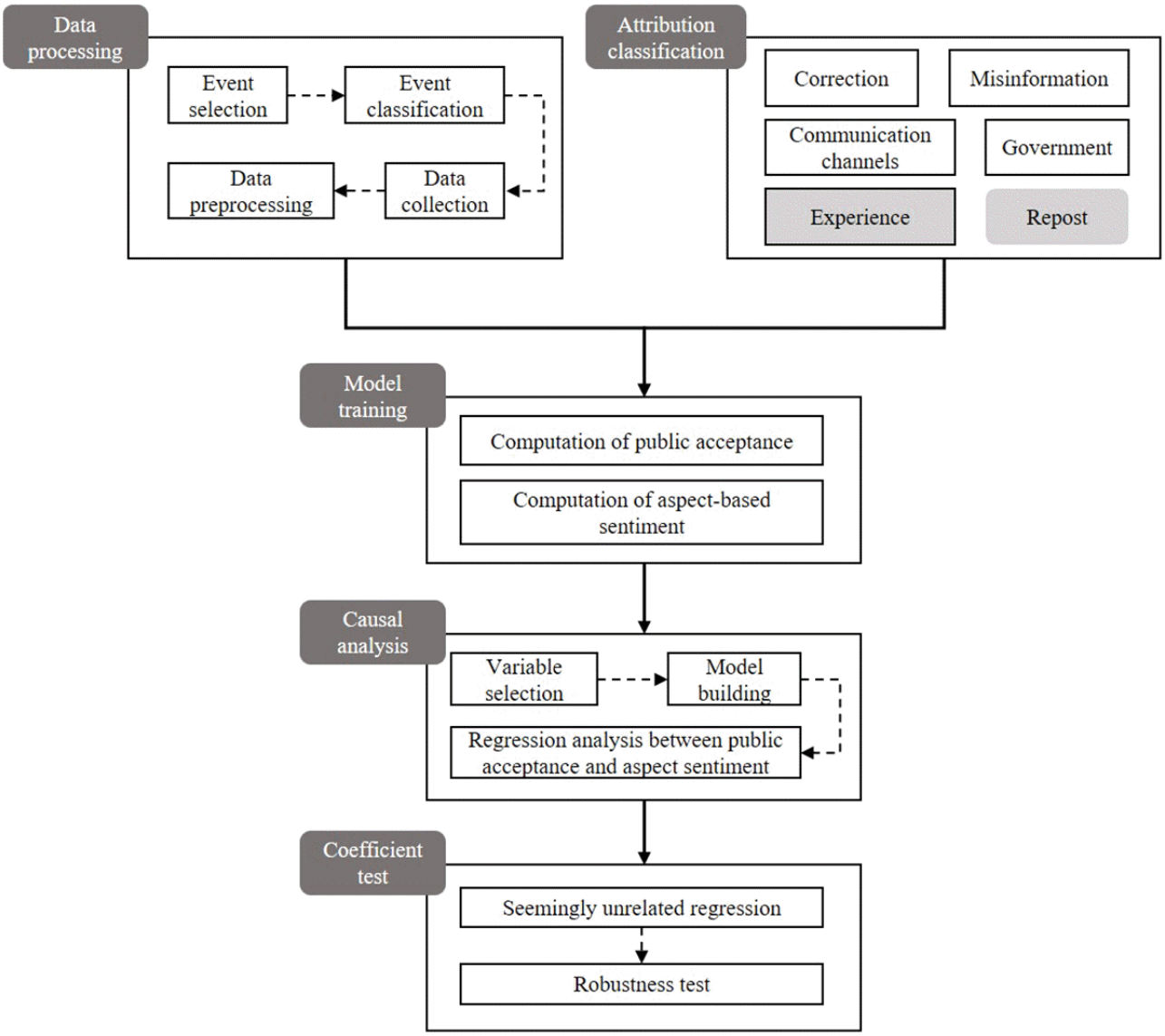

The underlying reasons for the ineffectiveness of misinformation correction as perceived by the public were elucidated by this framework. By examining various types of misinformation dissemination events, we categorized and analyzed key public sentiment attributions that contribute to public acceptance of misinformation correction in different situations. Furthermore, we considered public views that are not directly related to correction and whose role in analyzing the efficacy of correction cannot be ignored.

Event Selection and Data Collection

Weibo is a prominent social media platform in China that has been extensively studied in relation to public opinion [42]. One noteworthy feature of Weibo is the Weibo Refuting Rumor account, which serves to clarify and correct rumors that circulate on the web. This study used the posts from the Weibo Refuting Rumors account to screen 31 misinformation events related to COVID-19 that occurred between January 26, 2020, and April 14, 2022. The corresponding correction posts and their comments were obtained through data support from Zhiwei Data Sharing Platform. In total, the correction posts for the 31 events comprised 33,422 Weibo posts, and the corresponding Weibo comments amounted to 370,218. This study analyzed these comments to investigate the attributions that influenced public attitudes regarding misinformation correction.

Current research on misinformation, particularly related to COVID-19, focuses on categorizing it based on content. One such classification method through fact checking divided misinformation into 6 categories, which included treatment, conspiracy, government measures, vaccines, number of cases, and others [38]. Drawing from this categorization, this study categorized 31 misinformation events related to COVID-19 into

XSL•FO
RenderX

4 main categories through fact checking and examining the public opinion on social networks in China: prevention and treatment, conspiracy, government measures, and COVID-19 development. (Specific event information and classification can be found in Multimedia Appendix 1.)

Main Components of Sentiment Attributions in Misinformation Correction

Public sentiment responses to misinformation correction can originate from various sources [44]. According to attribution theory, public perception of misinformation correction is influenced by how they attribute the causes of their sentiment response. This attribution process involves 3 dimensions: locus, stability, and controllability [45]. When the public encounters misinformation correction, the attribution of public sentiment response ultimately falls on the issue of locus of causality. At this point, whether the public sentiment responses are attributed to internal or external attributions becomes the key of the research [46,47].

This study categorized the sources of public sentiment based on the key elements of misinformation into 4 categories: the background of the event, the channel of information, the information itself, and the recipient [34]. Regarding actual misinformation correction, the background of the event corresponds to the aspect of misinformation, the channel of information refers to the communication channel aspect, the information itself pertains to the correction aspect, and the recipient is the public. The internal attribution of public sentiment pertains to the public's own experience as the recipient, whereas other aspects of public sentiment fall under external attribution. Public sentiment responses attributed to these different aspects are associated with public acceptance of misinformation correction. Furthermore, the influence of the government on public sentiment is a crucial factor, and its response to misinformation also affects the public sentiment response [28]. According to the aforementioned theoretical analysis, public sentiments were divided into fine-grained sentiments in 6 aspects and 11 dimensions, which are detailed in Table 1 and reflect the attributions of public sentiment response.

Table 1. The specific description and classification of sentiment attribution into aspects.

Locus, aspect, and dimension	Specific description
External	
Correction	
MCB	Correction action against misinformation
MD	Debunker of misinformation
Misinformation	
MMO	The main object of misinformation events
MSB	The spread and manufacturing of misinformation
DOM	Disseminators and producers of misinformation
Communication channels	
CM	Communication media and celebrities
Government	
GA	Official government and public authority representative
LG	Local government
TM	Treatment and response measures in misinformation events
Internal	
Experience	
SP	Personal feelings and experience
No attribution	
Repost	
REP	Repost

Ethical Considerations

The comment data used for this study are publicly accessible on the Weibo platform, and none of the data contained any details that could be used to track individual users. The authors consulted with their respective institutional review boards and were informed that no approvals were necessary.

Computation of Public Acceptance and Public Sentiments

To evaluate public attitudes toward misinformation correction in the comments, it is crucial to extract public acceptance of the correction. As the text data used in this study were in Chinese, the Bidirectional Encoder Representations from Transformers (BERT) pretraining model based on the Chinese pretraining

XSL•FO
RenderX

model of the Social Computing and Information Retrieval Research Center from the Harbin Institute of Technology was used to assess public opinions [48,49]. Specifically, comments were labeled as *tag 1* if the public recognized the correction and as *tag 0* if the public considered the correction unreliable or ineffective. In the labeling process, 2 professional researchers randomly selected 10,000 comments from all the processed Weibo comments and labeled them based on whether the public recognized the correction. The tagged text was then cross-validated by 2 professional researchers. After labeling, the data were split into validation and training sets in a 1:9 ratio, with labels of 1 representing public acceptance and 0 representing public nonacceptance. The pretraining model was then adjusted to the use environment and used for model training. The model had a good performance, and it is presented in Table 2. It was used to assess public acceptance of the correction in the remaining comments.

Table 2. Model performance of Bidirectional Encoder Representations from Transformers (BERT) and aspect-based sentiment analysis (ABSA) on different indicators.

Model and effectiveness indicator	Performance (%)
BERT_Roberta	
Accuracy	83.21
Recall	81.55
F_1-score	80.71
ABSA	
APC_accuracy[a]	87.27
APC_f1_score[b]	87.18
ATE_accuracy[c]	89.08
ATE_f1_score[d]	88.87

[a]APC_accuracy: ABSA development has progressed effectively. The accuracy of aspect polarity category.

[b]APC_f1_score: The F_1-score of aspect polarity category.

[c]ATE_accuracy: The accuracy of aspect term extraction.

[d]ATE_f1_score: The F_1-score of aspect term extraction.

To identify the specific target objects of public sentiments in each text and analyze the attribution of public sentiment response, this study used the aspect-based sentiment computing method. ABSA is a method to analyze the sentiments of different aspects of an object in text. Unlike traditional sentiment classification, which only provides the comprehensive sentiment polarity of the text, ABSA focuses on both target extraction and target sentiment classification, making it a more fine-grained sentiment analysis method [50,51]. ABSA development has progressed effectively and has greatly improved the performance of the model, especially Chinese ABSA [52]. The most common application scenario for ABSA is in content analysis of restaurant and hotel reviews [53], where it completed the task of fine-grained sentiment analysis with 5 aspects and 18 dimensions [54]. Recently, ABSA has also been applied to extract themes and analyze relevant discourse on social media during COVID-19 [55], and other scholars have conducted aspect-based fine-grained sentiment analysis on user interaction data on social media about country image during the COVID-19 pandemic [56].

This study used ABSA involving 2 tasks: aspect term extraction and aspect polarity classification [52]. To adapt to the language environment of misinformation on Weibo, the corpus labeling rules were adjusted, and public sentiments were classified into negative and nonnegative categories. The texts were labeled according to the classification in Table 1. In total, 2 professional researchers labeled the corpus, resulting in a total of 17,629 corpora through cross-validation, which were divided into validation and training sets with a ratio of 1:9 for model training. The model performance is shown in Table 2. This model was then used to predict aspect sentiments in the remaining comments and investigate the specific attributions of public sentiments.

Causal Analysis of Public Acceptance of Correction

This study used logistic regression analysis to investigate the particular attributions that impact public acceptance of misinformation correction. The dependent variable for this analysis was public acceptance of misinformation correction, whereas the independent variables were the number of negative and nonnegative sentiments in the text across 11 dimensions. Moreover, the study controlled for gender, number of fans, number of likes, geographical location, and user authentication identity type by incorporating them as control variables in the following logistic regression model:

$$Z = \ln(p/[1 - p]) = b_0 + b_1\text{MCB_non} + b_2\text{MCB_neg} + b_3\text{MCB_non} + b_4\text{MD_neg} + b_5\text{MMO_non} + b_6\text{MMO_neg} + b_7\text{MSB_non} + b_8\text{MSB_neg} + b_9\text{DOM_non} + b_{10}\text{DOM_neg} + b_{11}\text{CM_non} + b_{12}\text{CM_neg} + b_{13}\text{GA_non} + b_{14}\text{GA_neg} + b_{15}\text{LG_non} + b_{16}\text{LG_neg} + b_{17}\text{TM_non} + b_{18}\text{TM_neg} + b_{19}\text{SP_non} + b_{20}\text{SP_neg} + b_{21}\text{REP} + b_{22}\text{GEN} + b_{23}\text{FAN} + b_{24}\text{LIKE} + b_{25}\text{LOC} + b_{26}\text{TYPE} + \varepsilon_0,$$

XSL•FO
RenderX

Where p was the probability that the public thought the correction was valid, $1 - p$ was the probability that the public thought the correction was invalid, b_i was the regression coefficient of each variable, b_0 was the constant term, and e_0 was the random disturbance term. The specific variable descriptions are listed in Textbox 1. As the aspect sentiment calculation for each comment was based on the number of negative or nonnegative sentiments present within that comment, each variable representing aspect sentiment included both negative and nonnegative dimensions.

Textbox 1. Variable names and description of logistic regression analysis.

Variable and description

- MCB_non: number of nonnegative sentiments about correction action against misinformation
- MCB_neg: number of negative sentiments about correction action against misinformation
- MD_non: number of nonnegative sentiments about debunker of misinformation
- MD_neg: number of negative sentiments about debunker of misinformation
- MMO_non: number of nonnegative sentiments about the main object of misinformation events
- MMO_neg: number of negative sentiments about the main object of misinformation events
- MSB_non: number of nonnegative sentiments about the spread and manufacturing of misinformation
- MSB_neg: number of negative sentiments about the spread and manufacturing of misinformation
- DOM_non: number of nonnegative sentiments about disseminators and producers of misinformation
- DOM_neg: number of negative sentiments about disseminators and producers of misinformation
- CM_non: number of nonnegative sentiments about communication media and celebrities
- CM_neg: number of negative sentiments about communication media and celebrities
- GA_non: number of nonnegative sentiments about the official government and public authority representatives
- GA_neg: number of negative sentiments about the official government and public authority representatives
- LG_non: number of nonnegative sentiments about local government
- LG_neg: number of negative sentiments about local government
- TM_non: number of nonnegative sentiments about treatment and response measures in misinformation events
- TM_neg: number of negative sentiments about treatment and response measures in misinformation events
- SP_non: number of nonnegative sentiments about personal feelings and experience
- SP_neg: number of negative sentiments about personal feelings and experience
- REP: repost
- GEN: gender
- FAN: number of fans
- LIKE: number of likes
- LOC: the geographical location of the user's registration
- TYPE: user authentication identity type

This study examined the primary public sentiment attributes that influenced public acceptance of misinformation correction posted on social media based on theoretical hypotheses and conceptual models. From a public sentiment perspective, this study extracted variables from various aspects while controlling for user-specific attributes such as gender, number of fans, number of likes, and geographical location. The aim was to analyze the attributions of public sentiments related to public acceptance of misinformation correction. Thus, this study sought to identify why the public perceived correction to be ineffective and provided theoretical support for relevant correction communication methods.

Differences in Public Sentiment Attributions Among the Corrections of Different Types of Misinformation

We constructed a regression model to analyze which variables had a significant impact on public acceptance of misinformation correction. However, it remained to be tested whether the influence of the same variable on different types of misinformation could be compared through the regression coefficients. In this study, we categorized the misinformation events into 4 types: prevention and treatment (type P), conspiracy (type C), government measures (type G), and COVID-19 development (type D). The 4 types of misinformation constituted 4 subsample groups, and we tested the differences in coefficients between groups for all variables except for the control variables using coefficient tests based on

XSL•FO
RenderX

the seemingly unrelated regression. Existing methods can only test the coefficient difference between 2 groups of samples at a single time, so we had to test the coefficient difference among the 4 subsample groups in pairs. This approach enabled us to assess the difference in the role of a certain aspect of public sentiment in different types of misinformation, providing strategic support for the implementation of correction communication.

Results

Multidimensional Public Sentiments in Misinformation Correction

We conducted statistical analysis on public acceptance of misinformation correction in all comments and public sentiment response to different aspects, including the control variables, for the entire sample data. Table 3 presents the descriptive statistics of the overall sample data. As *GEN*, *LOC*, and *TYPE* were categorical variables and *LOC* and *TYPE* had numerous categories, the specific count of each category is not listed in Table 3.

Simple repost aspect was considered as the repost behavior after recognizing the correction in this study. However, Table 3 reveals that the average level of public acceptance of misinformation correction was 0.508, indicating that slightly over half of the public perceived the correction as effective. This finding suggests that the impact of correction on the public was not sufficiently robust. In addition, the statistics of aspect sentiments demonstrated that the number of aspect sentiments expressed in each text varied. Negative sentiments toward the dimensions of government treatment and response measures could appear up to 12 times in a single comment. Some comments did not even express sentiments about a particular dimension and might have solely been reposts.

Upon examining public acceptance of misinformation correction by event type, we identified differences in the level of acceptance across different types of misinformation. As depicted in Figure 2, the correction of type P garnered the highest public acceptance, whereas the correction of types G and D had the lowest acceptance. However, public acceptance of correction of type C fluctuated greatly, suggesting that the effect of correction for type C was not consistent. Moreover, Figure 2 does not reveal any distinct pattern regarding public acceptance of correction based on the sequence of events. Therefore, overall, corrections did not yield fixed and effective outcomes.

The findings presented in Figure 3 indicate that the comments on correction contained more sentiments regarding the aspect of misinformation, creating a noticeable gap between this and the other aspects. Aside from simple reposts, the number of sentiments in the remaining aspects was nearly equal. Furthermore, the composition of sentiments in each aspect of correction significantly varied across different types of misinformation, as evidenced by the proportion of sentiments toward correction depicted in Figure 3. For instance, comments on the correction of type G displayed significantly more sentiment toward the government compared to comments on the correction of other types. These observations indicated that there were variations in public acceptance of the correction of different types of misinformation, with marked differences in the composition of sentiments across specific aspects. Hence, conducting differentiated analyses on the correction of different types of misinformation was necessary to identify the specific reasons for poor public acceptance of correction.

Figure 4 depicts the co-occurrence network of public aspect sentiments in the correction of different types of misinformation, which highlights the disparities in public expression. The co-occurrence network illustrates the co-occurrence of different dimensions of aspect sentiments in public expression. The Jaccard coefficient served as the index for drawing the co-occurrence network, where all edges with a Jaccard coefficient of >0.1 were displayed. Our analysis revealed that the most central aspect sentiment was not necessarily the most frequent among different types of misinformation, and the central aspect sentiment in each type of misinformation network differed. For instance, in type C, the aspect sentiment of the *LG* dimension was central, whereas the aspect sentiment of *MSB* was central in type D. The aspect sentiment of *MCB* was central in type G, and the aspect sentiment of *GA* was central in type P. These differences in co-occurrence networks suggested the need for differentiated analyses of comments to uncover the reasons why the correction was not more recognized by the public.

XSL•FO
RenderX

Table 3. Descriptive statistics of the variables.

Variable	Observations, n	Values, mean (SD; range)
Attitudes	370,218	0.508 (0.5; 0-1)
MCB_non[a]	370,218	0.028 (0.199; 0-8)
MCB_neg[b]	370,218	0.036 (0.213; 0-8)
MD_non[c]	370,218	0.007 (0.1; 0-3)
MD_neg[d]	370,218	0.002 (0.04; 0-3)
MMO_non[e]	370,218	0.116 (0.377; 0-5)
MMO_neg[f]	370,218	0.066 (0.286; 0-7)
MSB_non[g]	370,218	0.017 (0.135; 0-5)
MSB_neg[h]	370,218	0.063 (0.275; 0-5)
DOM_non[i]	370,218	0.007 (0.088; 0-3)
DOM_neg[j]	370,218	0.047 (0.244; 0-6)
CM_non[k]	370,218	0.034 (0.2; 0-7)
CM_neg[l]	370,218	0.035 (0.22; 0-6)
GA_non[m]	370,218	0.019 (0.154; 0-6)
GA_neg[n]	370,218	0.014 (0.127; 0-5)
LG_non[o]	370,218	0.013 (0.122; 0-6)
LG_neg[p]	370,218	0.009 (0.101; 0-7)
TM_non[q]	370,218	0.024 (0.167; 0-4)
TM_neg[r]	370,218	0.029 (0.181; 0-12)
SP_non[s]	370,218	0.042 (0.209; 0-3)
SP_neg[t]	370,218	0.026 (0.164; 0-4)
REP[u]	370,218	0.153 (0.362; 0-3)
GEN[v]	370,218	1.485 (0.516; 1-3)
FAN[w]	370,218	35,210.49 (1,458,389; 0-2.31×10^8)
LIKE[x]	370,218	7.264 (484.681; 0-173,209)
LOC[y]	370,218	20.388 (9.079; 1-36)
TYPE[z]	370,218	5.732 (4.674; 1-11)

[a]MCB_non: number of nonnegative sentiments about correction action against misinformation.
[b]MCB_neg: number of negative sentiments about correction action against misinformation.
[c]MD_non: number of nonnegative sentiments about debunker of misinformation.
[d]MD_neg: number of negative sentiments about debunker of misinformation.
[e]MMO_non: number of nonnegative sentiments about the main object of misinformation events.
[f]MMO_neg: number of negative sentiments about the main object of misinformation events.
[g]MSB_non: number of nonnegative sentiments about the spread and manufacturing of misinformation.
[h]MSB_neg: number of negative sentiments about the spread and manufacturing of misinformation.
[i]DOM_non: number of nonnegative sentiments about disseminators and producers of misinformation.
[j]DOM_neg: number of negative sentiments about disseminators and producers of misinformation.
[k]CM_non: number of nonnegative sentiments about communication media and celebrities.
[l]CM_neg: number of negative sentiments about communication media and celebrities.

XSL•FO
RenderX

[m]GA_non: number of nonnegative sentiments about the official government and public authority representatives.

[n]GA_neg: number of negative sentiments about the official government and public authority representatives.

[o]LG_non: number of nonnegative sentiments about local government.

[p]LG_neg: number of negative sentiments about local government.

[q]TM_non: number of nonnegative sentiments about treatment and response measures in misinformation events.

[r]TM_neg: number of negative sentiments about treatment and response measures in misinformation events.

[s]SP_non: number of nonnegative sentiments about personal feelings and experience.

[t]SP_neg: number of negative sentiments about personal feelings and experience.

[u]REP: repost.

[v]GEN: gender.

[q]FAN: number of fans.

[x]LIKE: number of likes.

[y]LOC: the geographical location of the user's registration.

[z]TYPE: user authentication identity type.

Figure 2. The proportion of public acceptance of correction. These 31 misinformation events related to COVID-19 occurred between January 26, 2020, and April 14, 2022. (A) Box plot of the proportion of public acceptance of correction in each type of misinformation. (B) Histogram of the proportion of public acceptance of correction in each misinformation event.

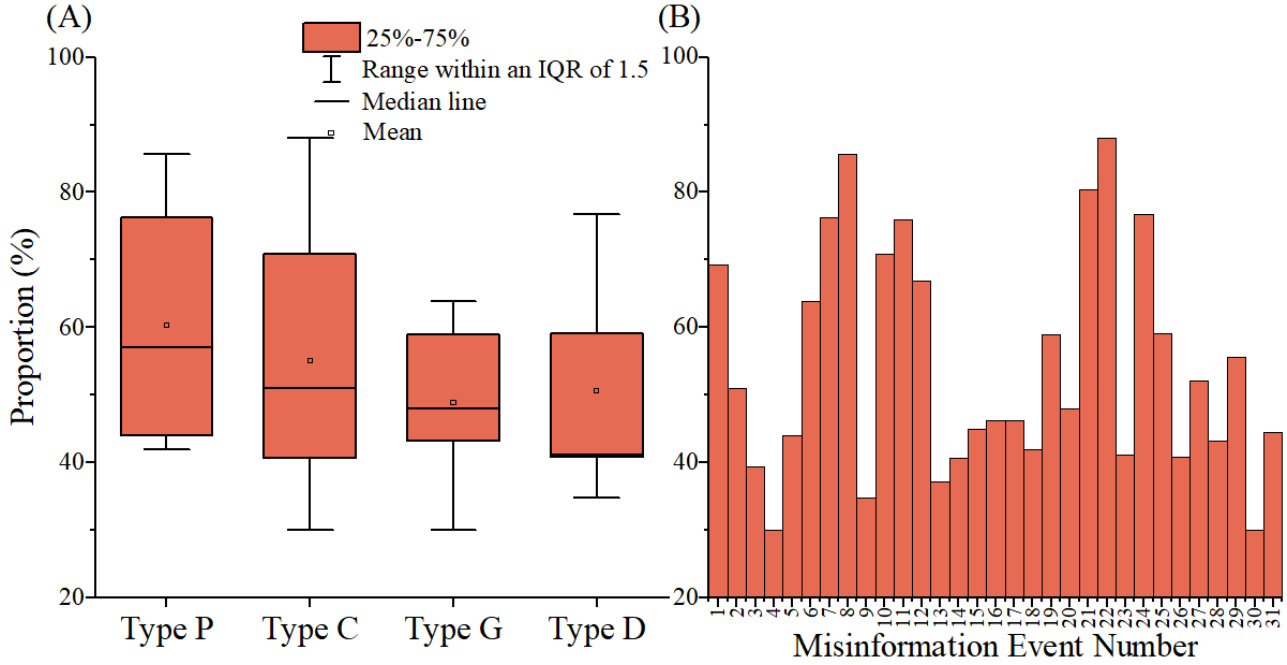

XSL•FO
RenderX

Figure 3. The composition of aspect sentiments in the correction of different types of misinformation. These 31 misinformation events were divided into 4 types: prevention and treatment (type P), conspiracy (type C), government measures (type G), and COVID-19 development (type D). (A) The number of public sentiments in different aspects. (B) The proportion of public sentiments in different aspects.

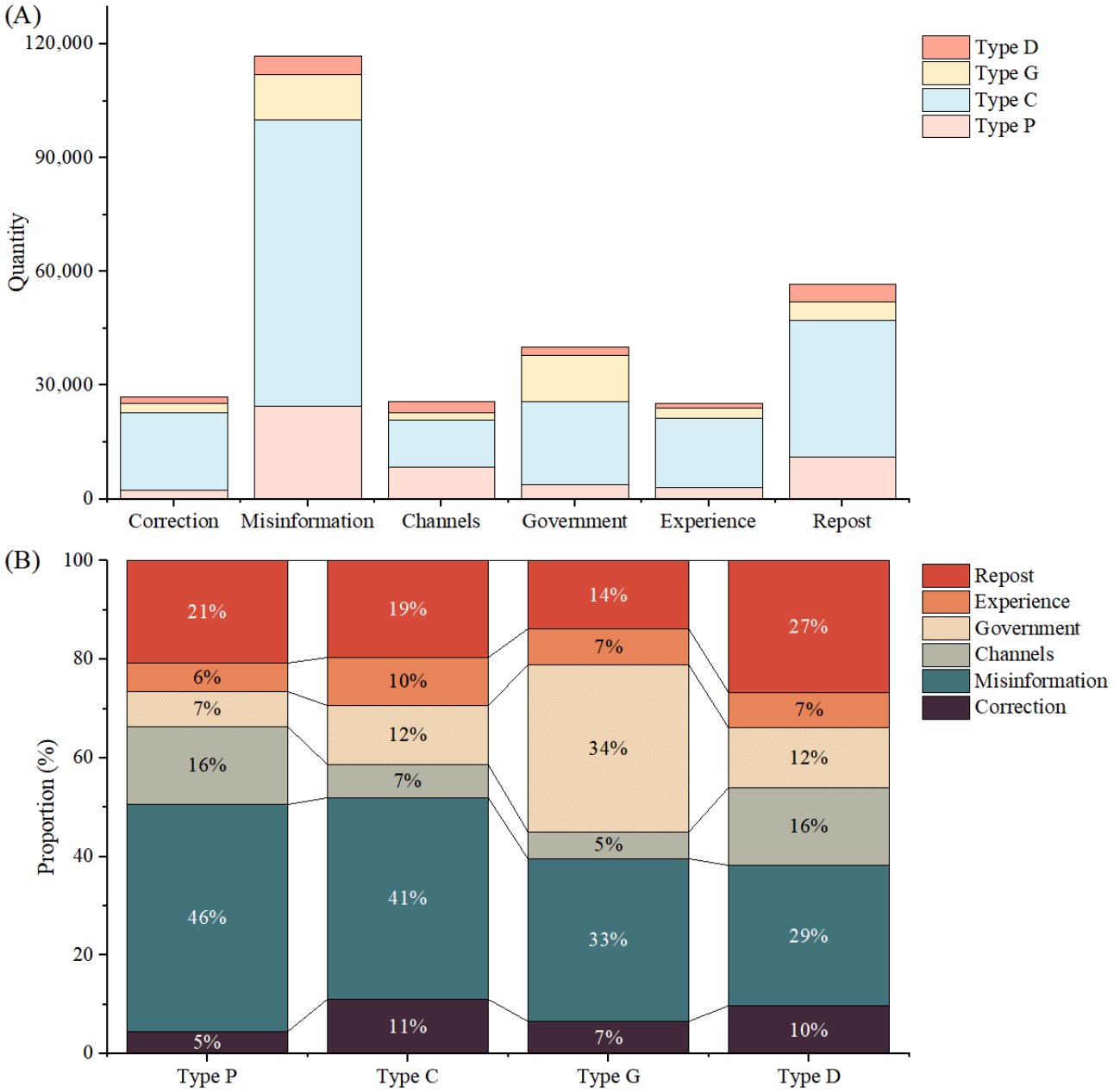

XSL•FO
RenderX

Figure 4. The attribution co-occurrence network of public aspect sentiment in the correction of different types of misinformation. (A) The co-occurrence network of type C. (B) The co-occurrence network of type D. (C) The co-occurrence network of type G. (D) The co-occurrence network of type P. CM: communication media and celebrities; DOM: disseminators and producers of misinformation; GA: official government and public authority representatives; LG: local government; MCB: correction action against misinformation; MMO: main object of misinformation events; MSB: spread and manufacturing of misinformation; SP: personal feelings and experience; TM: treatment and response measures in misinformation events.

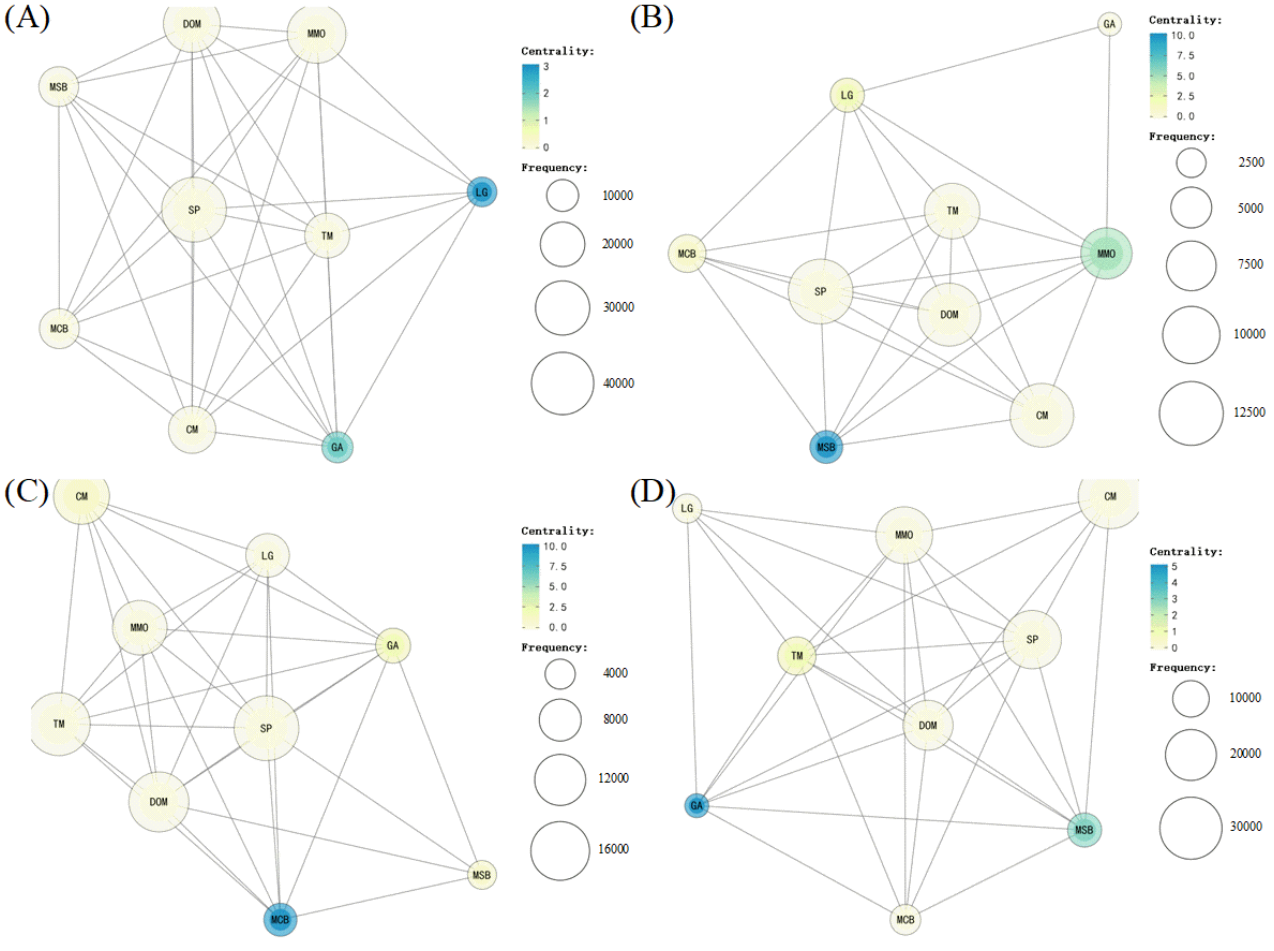

Causality Between Public Acceptance of Misinformation Correction and Attributions of Public Sentiment

The overall sample data were regressed at the event level, and the significance of aspect sentiments for each dimension of all events was calculated, as shown in Figure 5. The significance was normalized to obtain the relative significance intensity of sentiments in each dimension. Figure 5 shows that the repost aspect had a significant relationship with public acceptance of misinformation correction. Moreover, a certain degree of regularity in the relative significance intensity of different types of misinformation was revealed. Although there were some differences in the significance of sentiment in each dimension in each event, the results demonstrated that public sentiments in different dimensions were significantly related to whether the public perceived the correction as effective.

XSL•FO
RenderX

Figure 5. The relative significance intensity of different aspects of public sentiment. The redder the color block, the higher the relative significance intensity. (A) The relative significance intensity in each misinformation event. (B) The relative significance intensity in different types of misinformation. The relative significance intensity in different types of misinformation showed a unique distribution, so the research needed to consider the differences and common characteristics of the different types of misinformation. CM_neg: number of negative sentiments about communication media and celebrities; CM_non: number of nonnegative sentiments about communication media and celebrities; DOM_neg: number of negative sentiments about disseminators and producers of misinformation; DOM_non: number of nonnegative sentiments about disseminators and producers of misinformation; GA_neg: number of negative sentiments about the official government and public authority representatives; GA_non: number of nonnegative sentiments about the official government and public authority representatives; LG_neg: number of negative sentiments about local government; LG_non: number of nonnegative sentiments about local government; MCB_neg: number of negative sentiments about correction action against misinformation; MCB_non: number of nonnegative sentiments about correction action against misinformation; MD_neg: number of negative sentiments about debunker of misinformation; MD_non: number of nonnegative sentiments about debunker of misinformation; MMO_neg: number of negative sentiments about the main object of misinformation events; MMO_non: number of nonnegative sentiments about the main object of misinformation events; MSB_neg: number of negative sentiments about the spread and manufacturing of misinformation; MSB_non: number of nonnegative sentiments about the spread and manufacturing of misinformation; REP: repost; SP_neg: number of negative sentiments about personal feelings and experience; SP_non: number of nonnegative sentiments about personal feelings and experience; TM_neg: number of negative sentiments about treatment and response measures in misinformation events; TM_non: number of nonnegative sentiments about treatment and response measures in misinformation events.

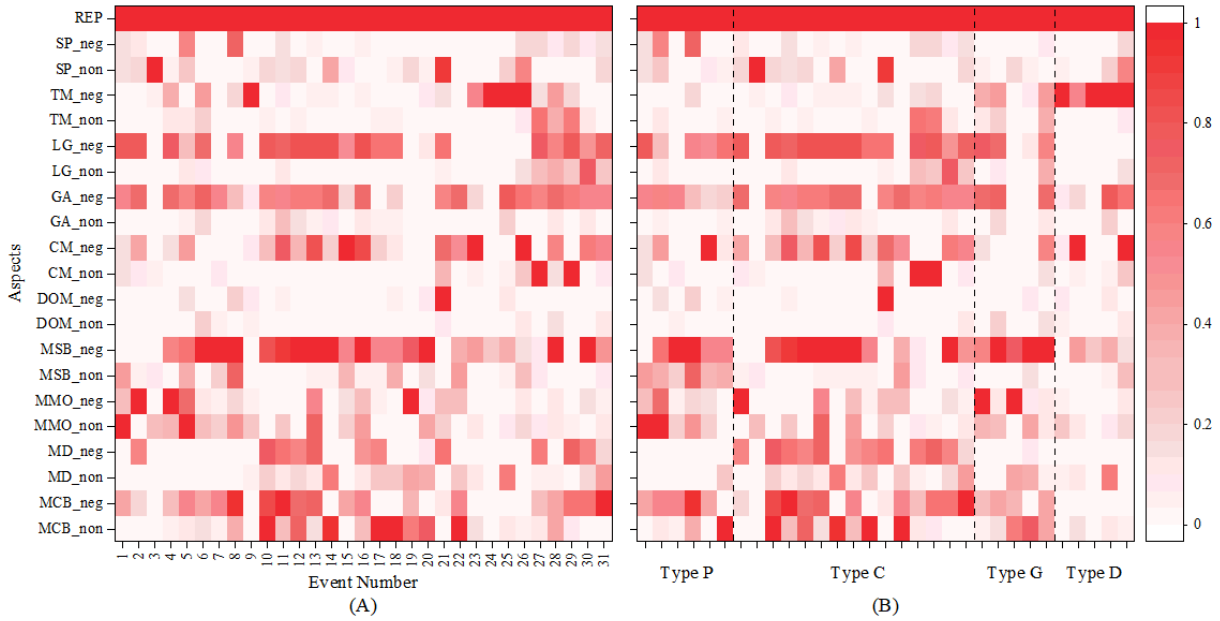

We randomly selected two-thirds of the overall sample data for grouped regression based on 4 different types. Table 4 shows the regression results of public acceptance of misinformation correction for various types of misinformation and each corresponding variable. The variables *LIKE*, *LOC*, and *TYPE* were included as control variables but were found to be insignificant across different types. While *FAN* was found to be significant, its coefficient was too small to be considered a reference, and only GEN had a certain impact on public acceptance. Therefore, the influence of gender was always controlled in the grouped regression experiment. In addition, we only considered the causality between public aspect sentiment and public acceptance of correction.

XSL•FO
RenderX

Table 4. Regression coefficients for different types of misinformation.

Variable	Type P regression coefficient	Type C regression coefficient	Type G regression coefficient	Type D regression coefficient
MCB_non[a]	1.232[b]	1.168[b]	1.15[b]	1.542[b]
MCB_neg[c]	–0.495[b]	–1.033[b]	–0.816[b]	–0.17[d]
MD_non[e]	1.855[b]	1.183[b]	1.523[b]	1.766[b]
MD_neg[f]	–0.191	–0.652[b]	–0.481	0.428
MMO_non[g]	–0.514[b]	–0.209[b]	–0.798[b]	–0.918[b]
MMO_neg[h]	–1.147[b]	–0.596[b]	–1.746[b]	–0.137
MSB_non[i]	–0.352[b]	–0.24[b]	–0.222[j]	–0.339[j]
MSB_neg[k]	2.011[b]	0.934[b]	1.636[b]	1.606[b]
DOM_non[l]	–0.48[b]	–0.418[b]	–0.477[b]	–0.223
DOM_neg[m]	0.162[b]	0.373[b]	0.103[j]	0.34[b]
CM_non[n]	–0.106[b]	0.255[b]	0.347[b]	0.1
CM_neg[o]	0.723[b]	0.651[b]	0.886[b]	1.693[b]
GA_non[p]	0.574[b]	1.093[b]	0.075	0.872[b]
GA_neg[q]	–0.601[b]	–0.982[b]	–0.915[b]	–0.902[b]
LG_non[r]	–0.042	–0.71[b]	–0.597[b]	0.052
LG_neg[s]	–1.219[b]	–1.46[b]	–1.508[b]	–0.243
TM_non[t]	–0.431[b]	–0.221[b]	–0.5[b]	–0.086
TM_neg[u]	–0.368[b]	–0.384[b]	–0.939[b]	–2.461[b]
SP_non[v]	0.276[b]	1.099[b]	0.347[b]	0.499[b]
SP_neg[w]	–1.064[b]	–0.389[b]	–0.003	–0.316[b]
REP[x]	2.933[b]	2.92[b]	2.729[b]	2.807[b]
GEN[y]	–0.2[b]	–0.017[d]	0.188[b]	0.301[b]
FAN[z]	0[b]	0[b]	0[b]	0[b]
LIKE[aa]	0	0[d]	0	0
LOC[ab]	–0.001	0.001	0	–0.002
TYPE[ac]	–0.003	0	–0.001	0.001
Constant	–0.005	–0.113[b]	–1.052[b]	–1.404[b]

[a]MCB_non: number of nonnegative sentiments about correction action against misinformation.
[b]$P<.01$.
[c]MCB_neg: number of negative sentiments about correction action against misinformation.
[d]$P<.10$.
[e]MD_non: number of nonnegative sentiments about debunker of misinformation.
[f]MD_neg: number of negative sentiments about debunker of misinformation.
[g]MMO_non: number of nonnegative sentiments about the main object of misinformation events.
[h]MMO_neg: number of negative sentiments about the main object of misinformation events.
[i]MSB_non: number of nonnegative sentiments about the spread and manufacturing of misinformation.
[j]$P<.05$.
[k]MSB_neg: number of negative sentiments about the spread and manufacturing of misinformation.
[l]DOM_non: number of nonnegative sentiments about disseminators and producers of misinformation.

XSL•FO
RenderX

[m]DOM_neg: number of negative sentiments about disseminators and producers of misinformation.

[n]CM_non: number of nonnegative sentiments about communication media and celebrities.

[o]CM_neg: number of negative sentiments about communication media and celebrities.

[p]GA_non: number of nonnegative sentiments about the official government and public authority representatives.

[q]GA_neg: number of negative sentiments about the official government and public authority representatives.

[r]LG_non: number of nonnegative sentiments about local government.

[s]LG_neg: number of negative sentiments about local government.

[t]TM_non: number of nonnegative sentiments about treatment and response measures in misinformation events.

[u]TM_neg: number of negative sentiments about treatment and response measures in misinformation events.

[v]SP_non: number of nonnegative sentiments about personal feelings and experience.

[w]SP_neg: number of negative sentiments about personal feelings and experience.

[x]REP: repost.

[y]GEN: gender.

[z]FAN: number of fans.

[aa]LIKE: number of likes.

[ab]LOC: the geographical location of the user's registration.

[ac]TYPE: user authentication identity type.

All aspects of sentiments in each dimension corresponded to human factors. Therefore, it can be proven that, regardless of the type of misinformation used in this study, public sentiments influenced by human factors affect public acceptance. For SP attributed to internal attribution, it influenced all types. Public acceptance of misinformation correction exhibited a positive correlation with positive public sentiments attributed to personal experience while displaying a negative correlation with negative public sentiments attributed to personal experience. However, its coefficient was not strong in all regression results. On the other hand, for the other dimensions of external attribution, sentiments played a significant role in the regression results. This finding indicated that public sentiments attributed to external attribution had a greater impact on public acceptance than public sentiments attributed to internal attribution regardless of the correction of different types of misinformation.

Regarding correction aspect, public acceptance of misinformation correction was positively correlated with the nonnegative sentiments of *MCB*, negatively correlated with the negative sentiments of *MCB*, and positively correlated with the nonnegative sentiments of *MD*. In addition, the negative sentiments of MD in type C were negatively correlated with public acceptance. For the sentiments of misinformation aspect, the regression coefficients of positive and negative sentiments of the 3 dimensions were integrated. The results supported that public acceptance exhibited a negative correlation with positive sentiments attributed to misinformation while displaying a positive correlation with negative sentiments attributed to misinformation. Regarding the sentiments of communication channels, only in type C and type G public acceptance of misinformation correction was shown to be positively correlated with the nonnegative sentiments of *CM*. However, in all types, public acceptance was positively correlated with the negative sentiments of *CM*. Finally, in the government aspect, public acceptance was only negatively correlated with the negative sentiments of *GA*, positively correlated with the nonnegative sentiments of *GA*, and negatively correlated with aspect sentiments on both of the other dimensions.

Different Effectiveness of Aspect Sentiments in the Correction of Different Types of Misinformation

To reveal which dimension of sentiments played a greater role in correcting different types of misinformation, it was necessary to compare the regression coefficients of the 4 subsample groups. Therefore, it was necessary to test the difference in coefficients between groups. After pairwise testing among the 4 subsample groups, the results are presented in Table 5. The differences in coefficients between the 2 groups that passed the test are marked with footnotes in the table. Most of these coefficients could be directly compared with the regression coefficients in the grouped regression results.

XSL•FO
RenderX

Table 5. Test for differences in coefficients between groups.

Variable	*P* value (type D and type C)	*P* value (type G and type C)	*P* value (type P and type C)	*P* value (type D and type G)	*P* value (type D and type P)	*P* value (type G and type P)
MCB_non[a]	.01[b]	.84	.46	.02[b]	.06[c]	.50
MCB_neg[d]	<.001[e]	.04[b]	<.001[e]	<.001[e]	.008[e]	.006[e]
MD_non[f]	.001[e]	.09[c]	.003[e]	.44	.69	.30
MD_neg[g]	.05[c]	.63	.15	.21	.40	.57
MMO_non[h]	<.001[e]	<.001[e]	<.001[e]	.28	<.001[e]	<.001[e]
MMO_neg[i]	<.001[e]	<.001[e]	<.001[e]	<.001[e]	<.001[e]	<.001[e]
MSB_non[j]	<.001[e]	<.001[e]	.13	.42	<.001[e]	<.001[e]
MSB_neg[k]	<.001[e]	<.001[e]	<.001[e]	.72	<.001[e]	<.001[e]
DOM_non[l]	.13	<.001[e]	.51	.001[e]	.09[c]	<.001[e]
DOM_neg[m]	.81	<.001[e]	<.001[e]	.02[b]	<.001[e]	<.001[e]
CM_non[n]	.15	.22	<.001[e]	.06[c]	.08[c]	<.001[e]
CM_neg[o]	<.001[e]	.001[e]	.12	<.001[e]	<.001[e]	.02[b]
GA_non[p]	.19	<.001[e]	<.001[e]	<.001[e]	.04[b]	<.001[e]
GA_neg[q]	.48	.38	.001[e]	.93	.17	.06[c]
LG_non[r]	<.001[e]	.17	<.001[e]	<.001[e]	.07[c]	<.001[e]
LG_neg[s]	<.001[e]	.81	.28	<.001[e]	<.001[e]	.27
TM_non[t]	.38	<.001[e]	.08[c]	.006[e]	.06[c]	.52
TM_neg[u]	<.001[e]	<.001[e]	.82	<.001[e]	<.001[e]	<.001[e]
SP_non[v]	<.001[e]	<.001[e]	<.001[e]	.18	<.001[e]	<.001[e]
SP_neg[w]	.64	<.001[e]	<.001[e]	.002[e]	<.001[e]	<.001[e]
REP[x]	<.001[e]	<.001[e]	.02[b]	.06[c]	<.001[e]	.009[e]

[a]MCB_non: number of nonnegative sentiments about correction action against misinformation.
[b]$P<.05$.
[c]$P<.10$.
[d]MCB_neg: number of negative sentiments about correction action against misinformation.
[e]$P<.01$.
[f]MD_non: number of nonnegative sentiments about debunker of misinformation.
[g]MD_neg: number of negative sentiments about debunker of misinformation.
[h]MMO_non: number of nonnegative sentiments about the main object of misinformation events.
[i]MMO_neg: number of negative sentiments about the main object of misinformation events.
[j]MSB_non: number of nonnegative sentiments about the spread and manufacturing of misinformation.
[k]MSB_neg: number of negative sentiments about the spread and manufacturing of misinformation.
[l]DOM_non: number of nonnegative sentiments about disseminators and producers of misinformation.
[m]DOM_neg: number of negative sentiments about disseminators and producers of misinformation.
[n]CM_non: number of nonnegative sentiments about communication media and celebrities.
[o]CM_neg: number of negative sentiments about communication media and celebrities.
[p]GA_non: number of nonnegative sentiments about the official government and public authority representatives.
[q]GA_neg: number of negative sentiments about the official government and public authority representatives.
[r]LG_non: number of nonnegative sentiments about local government.
[s]LG_neg: number of negative sentiments about local government.
[t]TM_non: number of nonnegative sentiments about treatment and response measures in misinformation events.

XSL•FO
RenderX

[u]TM_neg: number of negative sentiments about treatment and response measures in misinformation events.

[v]SP_non: number of nonnegative sentiments about personal feelings and experience.

[w]SP_neg: number of negative sentiments about personal feelings and experience.

[x]REP: repost.

Combining the analyses in Tables 4 and 5, we observed that, excluding the effect of repost, in type P, the aspect sentiment with the largest effect was the negative sentiments of *MSB* followed by the nonnegative sentiments of *MD* and the nonnegative sentiments of *MCB*. The negative sentiments of *LG* and the negative sentiments of *MMO* also played a significant role in determining whether the public recognized the correction. This finding suggested that, compared with the correction of other types of misinformation, public sentiments about the spread and manufacturing of misinformation in type P played a more significant role. In type C, the negative sentiments of *LG* were the most influential aspect sentiments, whereas *MCB* and *GA* had large regression coefficients, whether positive or negative. This phenomenon indicated that, in type C, public sentiments toward the correction aspect as well as the government aspect played a more significant role. In type G, the negative sentiments of *MMO* had a more significant effect. Public attention and discussion on the misinformation aspect had a significant impact on the correction of type G. Furthermore, the role of local government in this type of misinformation needed to be considered as public perception of local government also played a significant role. In type D, the most significant impact on misinformation correction was from the negative sentiments of *TM*, with the negative sentiments of *CM* also playing a vital role. Notably, the causality between public acceptance of misinformation correction and negative sentiments of *TM* was apparent. The regression coefficient of negative sentiments of *CM* was also the highest among the 4 types of misinformation. This showed that, for the correction of type D, public discussions about treatment and response measures of the government and communication channels better reflected public acceptance of correction.

Robustness Test

Finally, to ensure the robustness and reliability of the experimental results and conclusions, the remaining third of the sample data was randomly selected from the entire sample and regrouped according to the 4 different types of misinformation for further testing. The results of the robustness test are shown in Table 6.

The test results showed only minor changes in regression coefficients and significance, which did not alter the overall results and conclusions of the experiment. On the basis of this, it could be concluded that the experimental results of this study were robust and reliable.

XSL•FO
RenderX

Table 6. Robustness test results using different data.

Variable	Type P regression coefficient	Type C regression coefficient	Type G regression coefficient	Type D regression coefficient
MCB_non[a]	1.123[b]	1.194[b]	1.153[b]	1.493[b]
MCB_neg[c]	–0.567[b]	–1.02[b]	–0.854[b]	–0.212[d]
MD_non[e]	1.813[b]	1.158[b]	1.615[b]	1.773[b]
MD_neg[f]	–0.307	–0.629[b]	0.056	0.006
MMO_non[g]	–0.51[b]	–0.212[b]	–0.782[b]	–1.024[b]
MMO_neg[h]	–1.148[b]	–0.62[b]	–1.786[b]	–0.065
MSB_non[i]	–0.344[b]	–0.252[b]	–0.375[j]	0.204
MSB_neg[k]	1.941[b]	0.919[b]	1.518[b]	1.574[b]
DOM_non[l]	–0.492[b]	–0.456[b]	–0.483[b]	–0.039
DOM_neg[m]	0.137[b]	0.393[b]	0.1[d]	0.272[b]
CM_non[n]	–0.044	0.231[b]	0.407[b]	0.1
CM_neg[o]	0.705[b]	0.636[b]	0.916[b]	1.638[b]
GA_non[p]	0.556[b]	1.121[b]	0.024	0.653[b]
GA_neg[q]	–0.411[b]	–0.939[b]	–0.826[b]	–1.045[b]
LG_non[r]	–0.096	–0.686[b]	–0.424[b]	0.145
LG_neg[s]	–1.205[b]	–1.687[b]	–1.326[b]	–0.206
TM_non[t]	–0.496[b]	–0.238[b]	–0.515[b]	0.002
TM_neg[u]	–0.398[b]	–0.384[b]	–0.881[b]	–2.421[b]
SP_non[v]	0.266[b]	1.086[b]	0.342[b]	0.471[b]
SP_neg[w]	–1.054[b]	–0.373[b]	0.043	–0.194[d]
REP[x]	2.884[b]	2.906[b]	2.735[b]	2.818[b]
GEN[y]	–0.238[b]	–0.015	0.203[b]	0.325[b]
FAN[z]	0[j]	0[b]	0[b]	0[b]
LIKE[aa]	0	0	0	0
LOC[ab]	–0.001	0.001	0	–0.003
TYPE[ac]	–0.002	0.001	0	0.001
Constant	0.038	–0.12[b]	–1.058[b]	–1.398[b]

[a]MCB_non: number of nonnegative sentiments about correction action against misinformation.
[b]$P<.01$.
[c]MCB_neg: number of negative sentiments about correction action against misinformation.
[d]$P<.10$.
[e]MD_non: number of nonnegative sentiments about debunker of misinformation.
[f]MD_neg: number of negative sentiments about debunker of misinformation.
[g]MMO_non: number of nonnegative sentiments about the main object of misinformation events.
[h]MMO_neg: number of negative sentiments about the main object of misinformation events.
[i]MSB_non: number of nonnegative sentiments about the spread and manufacturing of misinformation.
[j]$P<.05$.
[k]MSB_neg: number of negative sentiments about the spread and manufacturing of misinformation.
[l]DOM_non: number of nonnegative sentiments about disseminators and producers of misinformation.

XSL•FO
RenderX

[m]DOM_neg: number of negative sentiments about disseminators and producers of misinformation.

[n]CM_non: number of nonnegative sentiments about communication media and celebrities.

[o]CM_neg: number of negative sentiments about communication media and celebrities.

[p]GA_non: number of nonnegative sentiments about the official government and public authority representatives.

[q]GA_neg: number of negative sentiments about the official government and public authority representatives.

[r]LG_non: number of nonnegative sentiments about local government.

[s]LG_neg: number of negative sentiments about local government.

[t]TM_non: number of nonnegative sentiments about treatment and response measures in misinformation events.

[u]TM_neg: number of negative sentiments about treatment and response measures in misinformation events.

[v]SP_non: number of nonnegative sentiments about personal feelings and experience.

[w]SP_neg: number of negative sentiments about personal feelings and experience.

[x]REP: repost.

[y]GEN: gender.

[z]FAN: number of fans.

[aa]LIKE: number of likes.

[ab]LOC: the geographical location of the user's registration.

[ac]TYPE: user authentication identity type.

Discussion

Different Types of Misinformation Require Different Correction Communication

The study findings indicated that there were differences in public acceptance of correction among different types of misinformation and the key attributions that affect public acceptance also differed among them. However, regardless of the number of public discussions or the regression results, external attributions were the most significant among the corrections of 4 types of misinformation. The experimental results indicated that personal feelings and experience had an impact but were not the most important. The attributions in the aspects of correction and misinformation most directly influenced public acceptance of correction. While the media and government were the clear targets of most public sentiment responses, public attitudes toward these 2 aspects of attributions were significantly different from those toward other attributions. Compared with existing literature on misinformation correction effectiveness [2], this study provided a new analytical perspective.

In the correction of *prevention and treatment*, the regression results indicated that the public disliked misinformation and expected timely and accurate correction, with the government playing a crucial role. Implementers of correction and clarification of misinformation should be chosen carefully. In the correction of *conspiracy*, the findings illustrated that the public was dissatisfied with both the correction communication method and the implementers. Especially when the government confirmed and corrected misinformation, the public expressed negative sentiments toward local governments more frequently, so the way and timing of government involvement in the correction process needs to be considered prudently. Representatives with greater credibility could play a better role in the correction process [20]. In the correction of *government measures*, public negative sentiments toward misinformation revealed their attention to specific government measures. This showed that, in addition to the negative impact of misinformation, the public also sought to eliminate various inconveniences caused by the pandemic, reflecting their expectations of the government. However, in the correction of *COVID-19 development*, the study found that the public paid more attention to government treatment and response measures. The public considered the impact of treatment measures on themselves or whether the measures met their physical or psychological needs. The spread of this type of misinformation also stemmed from public concern about the uncertainty risk [57].

However, previous studies have emphasized that corrections are different depending on the nature [2,58]. The current approach to correction communication is too simplistic and repetitive, leading to public apathy and backlash [33]. It is necessary to adopt different methods of correcting misinformation to prevent stress reactions or aversion to correction and, ultimately, prevent the public from being exposed to misinformation [34]. Public demand for correction measures varies depending on the differences in misinformation types. The investigation into the specific attributions of public acceptance of misinformation correction is valuable as it uses the aspect sentiment analysis method to not only determine public acceptance but also uncover the underlying attributions affecting it. Ultimately, the fundamental issue is that the public must trust the corrections they receive, which requires flexible and diverse methods of communication and correction channels that the public trusts.

The Collapse of Media Credibility

The media play a supporting role in public acceptance of misinformation correction, but the effectiveness of the media in this regard is currently questionable. Public distrust of communication channels, particularly the media, was evident across all types of misinformation. Negative sentiments about communication channels were positively correlated with public acceptance of correction, indicating that the media were not perceived as trustworthy by the public. Media failure to gain public trust had resulted in a negative impact on the effectiveness of corrections. Discussions involving authoritative experts and celebrities were the only aspects of communication channels that generated positive sentiments.

XSL•FO
RenderX

In contrast, the public was more likely to trust authoritative experts. Analysis of comments revealed that one of the main reasons for this distrust was the tendency of some media outlets to prioritize speed over verification, leading to the spread of rumors and misinformation. Therefore, the public expressed their distrust of both traditional media and "we-media." The collapse of media credibility has reduced the persuasive power of correction spread by the media, thereby diminishing the effectiveness of correction. Enhancing the role of communication channels in promoting public acceptance of corrections might require the support of authoritative experts [20]. Therefore, the media need to work toward rebuilding their image in the public impression to regain lost credibility and effectively support the spread and effect of misinformation correction.

The Crucial Role of the Government

The government plays a crucial role in public acceptance of misinformation correction. Public sentiments toward correction of different types of misinformation also raised significant concerns regarding government involvement. However, experimental findings on public aspect sentiments toward the government demonstrated that only positive sentiments of the *GA* dimension were positively correlated with public acceptance. Both positive and negative sentiments of the *LG* and *TM* dimensions were negatively correlated with public acceptance. This phenomenon suggested that, when assessing public sentiments regarding the government, only the *GA* dimension had a positive guiding effect.

Although the public still maintained trust in the official government image, expectations of the government persisted. Concerning local governments and treatment and response measures, public trust in some local governments was eroded due to the pandemic, aligning with previous research findings [8]. Many treatment and response measures also failed to meet public expectations. Nevertheless, there is no doubt that the government plays a decisive role in both managing the pandemic and mitigating the impact of misinformation on the public [28,59]. Consequently, to enhance communication efficacy and foster public government engagement, trust between the government and the public must be restored. The pivotal role of the government in shaping public acceptance of misinformation correction cannot be overstated as it also exerts a positive influence on other aspects. Nonetheless, based on the results of this study, efforts are needed to solidify the official government image, rebuild trust in local governments, and implement comprehensive response measures to fully leverage the government's crucial role.

Limitations and Directions for Future Research

It is important to note that the data used in this study solely originated from user interactions on the Weibo platform. As a result, the analysis of user behaviors and causality can only be applied within the context of Weibo. Therefore, further verification is necessary to generalize the research findings and conclusions to other platforms. Furthermore, the data used in this study specifically pertained to rumors surrounding the COVID-19 pandemic, and they did not encompass rumors unrelated to the pandemic. Consequently, the extension of conclusions and potential variations in other types of misinformation warrant exploration in future research.

Although this study incorporated several control variables and strived to focus on the role of aspect sentiments, it is important to acknowledge that there might be variables associated with personal characteristics that were not accessible within the existing data. This study primarily examined group behavior and might have limitations in accounting for the influence of individual traits. In addition, data on individuals who do not disclose their opinions cannot be obtained, which is a limitation of using social media data for research. As the data were from Chinese social media, more data are needed to verify the expansibility of relevant conclusions to countries with other cultural backgrounds in future research. This study will help identify effective methods for correcting misinformation that are not limited to the pandemic.

Conclusions

Misinformation correction often falls short in effectively reaching and persuading the public, necessitating a thorough investigation into the reasons influencing public acceptance of such efforts. This study identified crucial sentiment attributions that shape public acceptance of correction regarding different types of misinformation. It highlighted the role of aspect sentiments in analyzing public acceptance of misinformation correction.

The findings revealed that a flexible, issue-specific approach is required for correcting different types of misinformation. The correction strategies also require flexible communication tailored to public sentiment attribution. The decline in media credibility negatively impacted the effectiveness of misinformation correction. Regulating media behavior and rebuilding media image are essential for regaining public trust. Furthermore, the government plays a critical role; in particular, some local governments need to regain the trust of the public, enabling government-led correction communication to be more effective.

This study identified key reasons contributing to the difference in public acceptance of misinformation correction, focusing on public sentiment attributions. It provided an empirical and methodological foundation for addressing misinformation governance to some extent.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grants 72074060 and 72374056).

XSL•FO
RenderX

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

NM proposed the original conceptualization. All authors contributed to methodology, validation, and writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details and types of misinformation-spreading events.
[DOCX File , 19 KB - jmir_v26i1e50353_app1.docx]

References

1. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
2. Aikin KJ, Betts KR, O'Donoghue AC, Rupert DJ, Lee PK, Amoozegar JB, et al. Correction of overstatement and omission in direct-to-consumer prescription drug advertising. J Commun 2015 Jun 23;65(4):596-618. [doi: 10.1111/jcom.12167]
3. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
4. Domenico GD, Sit J, Ishizaka A, Nunan D. Fake news, social media and marketing: a systematic review. J Bus Res 2021 Jan;124:329-341. [doi: 10.1016/j.jbusres.2020.11.037]
5. van Mulukom V, Pummerer LJ, Alper S, Bai H, Čavojová V, Farias J, et al. Antecedents and consequences of COVID-19 conspiracy beliefs: a systematic review. Soc Sci Med 2022 May;301:114912 [FREE Full text] [doi: 10.1016/j.socscimed.2022.114912] [Medline: 35354105]
6. Gabarron E, Oyeyemi SO, Wynn R. COVID-19-related misinformation on social media: a systematic review. Bull World Health Organ 2021 Jun 01;99(6):455-63A [FREE Full text] [doi: 10.2471/BLT.20.276782] [Medline: 34108756]
7. Wang R, Zhang H. Who spread COVID-19 (mis)information online? Differential informedness, psychological mechanisms, and intervention strategies. Comput Human Behav 2023 Jan;138:107486 [FREE Full text] [doi: 10.1016/j.chb.2022.107486] [Medline: 36120514]
8. Pickles K, Cvejic E, Nickel B, Copp T, Bonner C, Leask J, et al. COVID-19 misinformation trends in Australia: prospective longitudinal national survey. J Med Internet Res 2021 Jan 07;23(1):e23805 [FREE Full text] [doi: 10.2196/23805] [Medline: 33302250]
9. Su Y. It doesn't take a village to fall for misinformation: social media use, discussion heterogeneity preference, worry of the virus, faith in scientists, and COVID-19-related misinformation beliefs. Telemat Inform 2021 May;58:101547 [FREE Full text] [doi: 10.1016/j.tele.2020.101547] [Medline: 36570475]
10. Freeman D, Waite F, Rosebrock L, Petit A, Causier C, East A, et al. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. Psychol Med 2020 May 21;52(2):251-263. [doi: 10.1017/s0033291720001890]
11. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb;395(10225):676. [doi: 10.1016/s0140-6736(20)30461-x]
12. van der Linden S. Misinformation: susceptibility, spread, and interventions to immunize the public. Nat Med 2022 Mar 10;28(3):460-467. [doi: 10.1038/s41591-022-01713-6] [Medline: 35273402]
13. Jiang M, Gao Q, Zhuang J. Reciprocal spreading and debunking processes of online misinformation: a new rumor spreading–debunking model with a case study. Phys A Stat Mech Appl 2021 Mar;565:125572. [doi: 10.1016/j.physa.2020.125572]
14. Chou WY, Gaysynsky A, Vanderpool RC. The COVID-19 misinfodemic: moving beyond fact-checking. Health Educ Behav 2021 Mar;48(1):9-13 [FREE Full text] [doi: 10.1177/1090198120980675] [Medline: 33322939]
15. Clayton K, Blair S, Busam JA, Forstner S, Glance J, Green G, et al. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Polit Behav 2019 Feb 11;42(4):1073-1095. [doi: 10.1007/S11109-019-09533-0]
16. Pennycook G, Bear A, Collins ET, Rand DG. The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Management Science 2020 Nov;66(11):4944-4957. [doi: 10.1287/mnsc.2019.3478]
17. Freeze M, Baumgartner M, Bruno P, Gunderson JR, Olin J, Ross MQ, et al. Fake claims of fake news: political misinformation, warnings, and the tainted truth effect. Polit Behav 2020 Feb 05;43(4):1433-1465. [doi: 10.1007/s11109-020-09597-3]

XSL•FO
RenderX

18. Jin Y, van der Meer TG, Lee YI, Lu X. The effects of corrective communication and employee backup on the effectiveness of fighting crisis misinformation. Public Relat Rev 2020 Sep;46(3):101910. [doi: 10.1016/j.pubrev.2020.101910]
19. Hameleers M, van der Meer TG. Misinformation and polarization in a high-choice media environment: how effective are political fact-checkers? Commun Res 2019 Jan 13;47(2):227-250. [doi: 10.1177/0093650218819671]
20. Vraga EK, Bode L, Tully M. Creating news literacy messages to enhance expert corrections of misinformation on twitter. Commun Res 2020 Jan 30;49(2):245-267. [doi: 10.1177/0093650219898094]
21. Hosni AI, Li K. Minimizing the influence of rumors during breaking news events in online social networks. Knowl Based Syst 2020 Apr;193:105452. [doi: 10.1016/j.knosys.2019.105452]
22. Walter N, Brooks JJ, Saucier CJ, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. Health Commun 2020 Aug 06:1-9. [doi: 10.1080/10410236.2020.1794553] [Medline: 32762260]
23. Lewandowsky S, van der Linden S. Countering misinformation and fake news through inoculation and prebunking. Eur Rev Soc Psychol 2021 Feb 22;32(2):348-384. [doi: 10.1080/10463283.2021.1876983]
24. Basol M, Roozenbeek J, Berriche M, Uenal F, McClanahan WP, van der Linden S. Towards psychological herd immunity: cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. Big Data Soc 2021 May 11;8(1):205395172110138. [doi: 10.1177/20539517211013868]
25. Dib F, Mayaud P, Chauvin P, Launay O. Online mis/disinformation and vaccine hesitancy in the era of COVID-19: why we need an eHealth literacy revolution. Hum Vaccin Immunother 2022 Dec 31;18(1):1-3 [FREE Full text] [doi: 10.1080/21645515.2021.1874218] [Medline: 33625960]
26. Maertens R, Roozenbeek J, Basol M, van der Linden S. Long-term effectiveness of inoculation against misinformation: three longitudinal experiments. J Exp Psychol Appl 2021 Mar;27(1):1-16. [doi: 10.1037/xap0000315] [Medline: 33017160]
27. Boudewyns V, Betts KR, Johnson M, Paquin RS, O'Donoghue AC, Southwell BG. Experimental evidence of consumer and physician detection and rejection of misleading prescription drug website content. Res Social Adm Pharm 2021 Apr;17(4):733-743. [doi: 10.1016/j.sapharm.2020.06.019] [Medline: 32792323]
28. Rodrigues UM, Xu J. Regulation of COVID-19 fake news infodemic in China and India. Media Int Aust 2020 Aug 17;177(1):125-131. [doi: 10.1177/1329878x20948202]
29. Marsden C, Meyer T, Brown I. Platform values and democratic elections: how can the law regulate digital disinformation? Comput Law Secur Rev 2020 Apr;36:105373. [doi: 10.1016/j.clsr.2019.105373]
30. Freiling I, Krause NM, Scheufele DA, Brossard D. Believing and sharing misinformation, fact-checks, and accurate information on social media: the role of anxiety during COVID-19. New Media Soc 2023 Jan 22;25(1):141-162 [FREE Full text] [doi: 10.1177/14614448211011451] [Medline: 36620434]
31. Wang X, Chao F, Yu G. Evaluating rumor debunking effectiveness during the COVID-19 pandemic crisis: utilizing user stance in comments on Sina Weibo. Front Public Health 2021;9:770111 [FREE Full text] [doi: 10.3389/fpubh.2021.770111] [Medline: 34926388]
32. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 02;41:433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
33. Trevors G, Duffy MC. Correcting COVID-19 misconceptions requires caution. Educ Res 2020 Sep 08;49(7):538-542. [doi: 10.3102/0013189x20953825]
34. Chen S, Xiao L, Kumar A. Spread of misinformation on social media: what contributes to it and how to combat it. Comput Human Behav 2023 Apr;141:107643. [doi: 10.1016/j.chb.2022.107643]
35. Zhang Y, Guo B, Ding Y, Liu J, Qiu C, Liu S, et al. Investigation of the determinants for misinformation correction effectiveness on social media during COVID-19 pandemic. Inf Process Manag 2022 May;59(3):102935 [FREE Full text] [doi: 10.1016/j.ipm.2022.102935] [Medline: 35400028]
36. Petz G, Karpowicz M, Fürschuß H, Auinger A, Stříteský V, Holzinger A. Computational approaches for mining user's opinions on the Web 2.0. Inf Process Manage 2014 Nov;50(6):899-908. [doi: 10.1016/j.ipm.2014.07.005]
37. Carrasco-Farré C. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. Humanit Soc Sci Commun 2022 May 09;9(1):1. [doi: 10.1057/s41599-022-01174-9]
38. Charquero-Ballester M, Walter JG, Nissen IA, Bechmann A. Different types of COVID-19 misinformation have different emotional valence on Twitter. Big Data Soc 2021 Sep 22;8(2):205395172110412. [doi: 10.1177/20539517211041279]
39. Jang H, Rempel E, Roe I, Adu P, Carenini G, Janjua NZ. Tracking public attitudes toward COVID-19 vaccination on tweets in Canada: using aspect-based sentiment analysis. J Med Internet Res 2022 Mar 29;24(3):e35016 [FREE Full text] [doi: 10.2196/35016] [Medline: 35275835]
40. Zhou C, Li K, Lu Y. Linguistic characteristics and the dissemination of misinformation in social media: the moderating effect of information richness. Inf Process Manage 2021 Nov;58(6):102679. [doi: 10.1016/j.ipm.2021.102679]
41. Tully M, Bode L, Vraga EK. Mobilizing users: does exposure to misinformation and its correction affect users' responses to a health misinformation post? Soc Media Soc 2020 Dec 10;6(4):205630512097837. [doi: 10.1177/2056305120978377]
42. Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 2021 Mar;3(3):e175-e194. [doi: 10.1016/s2589-7500(20)30315-0]

XSL•FO
RenderX

43. Rahmanti AR, Chien CH, Nursetyo AA, Husnayain A, Wiratama BS, Fuad A, et al. Social media sentiment analysis to monitor the performance of vaccination coverage during the early phase of the national COVID-19 vaccine rollout. Comput Meth Prog Bio 2022 Jun;221:106838 [FREE Full text] [doi: 10.1016/j.cmpb.2022.106838] [Medline: 35567863]
44. Ridderinkhof KR. Emotion in action: a predictive processing perspective and theoretical synthesis. Emot Rev 2017 Oct;9(4):319-325 [FREE Full text] [doi: 10.1177/1754073916661765] [Medline: 29098017]
45. Weiner B. An attributional theory of achievement motivation and emotion. Psychol Rev 1985;92(4):548-573. [doi: 10.1037/0033-295x.92.4.548]
46. Saleh MI. Attribution theory revisited: probing the link among locus of causality theory, destination social responsibility, tourism experience types, and tourist behavior. J Travel Res 2022 Sep 19;62(6):1309-1327. [doi: 10.1177/00472875221119968]
47. Weiner B. Attributional thoughts about consumer behavior. J Consum Res 2000 Dec 01;27(3):382-387. [doi: 10.1086/317592]
48. Che W, Feng Y, Qin L, Liu T. N-LTP: an open-source neural language technology platform for Chinese. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2021 Presented at: EMNLP '21; November 7-11, 2021; Virtual Event p. 42-49 URL: https://aclanthology.org/2021.emnlp-demo.6.pdf [doi: 10.18653/v1/2021.emnlp-demo.6]
49. Cui Y, Che W, Liu T, Qin B, Wang S, Hu G. Revisiting pre-trained models for Chinese natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual Event p. 657-668 URL: https://aclanthology.org/2020.findings-emnlp.58.pdf [doi: 10.18653/v1/2020.findings-emnlp.58]
50. Li X, Bing L, Li P, Lam W. A unified model for opinion target extraction and target sentiment prediction. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. 2019 Presented at: AAAI '19/IAAI '19/EAAI '19; January 27-February 1, 2019; Honolulu, HI p. 6714-6721 URL: https://dl.acm.org/doi/proceedings/10.5555/3508659 [doi: 10.1609/aaai.v33i01.33016714]
51. Li X, Bing L, Zhang W, Lam W. Exploiting BERT for end-to-end aspect-based sentiment analysis. arXiv Preprint posted online October 2, 2019. [FREE Full text] [doi: 10.18653/v1/d19-5505]
52. Yang H, Zeng B, Yang J, Song Y, Xu R. A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction. Neurocomput 2021 Jan;419:344-356. [doi: 10.1016/j.neucom.2020.08.001]
53. Zhang Y, Lu H, Jiang C, Li X, Tian X. Aspect-based sentiment analysis of user reviews in 5G networks. IEEE Network 2021 Jul;35(4):228-233. [doi: 10.1109/mnet.011.2000400]
54. Zhang Y, Du J, Ma X, Wen H, Fortino G. Aspect-based sentiment analysis for user reviews. Cogn Comput 2021 Jul 13;13(5):1114-1127. [doi: 10.1007/s12559-021-09855-4]
55. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021 Feb 10;23(2):e25431 [FREE Full text] [doi: 10.2196/25431] [Medline: 33497352]
56. Chen H, Zhu Z, Qi F, Ye Y, Liu Z, Sun M, et al. Country image in COVID-19 pandemic: a case study of China. IEEE Trans Big Data 2021 Mar 1;7(1):81-92. [doi: 10.1109/tbdata.2020.3023459]
57. Li MH, Chen Z, Rao LL. Emotion, analytic thinking and susceptibility to misinformation during the COVID-19 outbreak. Comput Human Behav 2022 Aug;133:107295 [FREE Full text] [doi: 10.1016/j.chb.2022.107295] [Medline: 35431427]
58. Aikin KJ, Southwell BG, Paquin RS, Rupert DJ, O'Donoghue AC, Betts KR, et al. Correction of misleading information in prescription drug television advertising: the roles of advertisement similarity and time delay. Res Social Adm Pharm 2017 Mar;13(2):378-388. [doi: 10.1016/j.sapharm.2016.04.004] [Medline: 27178746]
59. Yu W, Shen F, Min C. Correcting science misinformation in an authoritarian country: an experiment from China. Telemat Inform 2022 Jan;66:101749. [doi: 10.1016/j.tele.2021.101749]

Abbreviations

ABSA: aspect-based sentiment analysis
BERT: Bidirectional Encoder Representations from Transformers

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.06.23; peer-reviewed by B Southwell, A Holzinger, B Southwell; comments to author 27.05.24; revised version received 16.06.24; accepted 19.06.24; published 16.08.24.

Please cite as:
Ma N, Yu G, Jin X
Investigation of Public Acceptance of Misinformation Correction in Social Media Based on Sentiment Attributions: Infodemiology Study Using Aspect-Based Sentiment Analysis
J Med Internet Res 2024;26:e50353
URL: https://www.jmir.org/2024/1/e50353
doi:10.2196/50353
PMID:

©Ning Ma, Guang Yu, Xin Jin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Epidemiology of Prescription Drug References on X (Formerly Twitter): Neural Network Topic Modeling and Sentiment Analysis

Varun K Rao[1], MS; Danny Valdez[2], PhD; Rasika Muralidharan[3], MS; Jon Agley[2], MPH, PhD; Kate S Eddens[1], PhD; Aravind Dendukuri[3], MS; Vandana Panth[3], MS; Maria A Parker[2], PhD

[1]Department of Epidemiology & Biostatistics, School of Public Health Bloomington, Indiana University Bloomington, Bloomington, IN, United States
[2]Department of Applied Health Science, School of Public Health Bloomington, Indiana University Bloomington, Bloomington, IN, United States
[3]Luddy School of Informatics, Computing and Engineering, Indiana University Bloomington, Bloomington, IN, United States

Corresponding Author:
Maria A Parker, PhD
Department of Applied Health Science
School of Public Health Bloomington
Indiana University Bloomington
809 E. 9th St.
Bloomington, IN, 47405
United States
Phone: 1 812 856 5950
Email: map2@iu.edu

Abstract

Background: Data from the social media platform X (formerly Twitter) can provide insights into the types of language that are used when discussing drug use. In past research using latent Dirichlet allocation (LDA), we found that tweets containing “street names” of prescription drugs were difficult to classify due to the similarity to other colloquialisms and lack of clarity over how the terms were used. Conversely, “brand name” references were more amenable to machine-driven categorization.

Objective: This study sought to use next-generation techniques (beyond LDA) from natural language processing to reprocess X data and automatically cluster groups of tweets into topics to differentiate between street- and brand-name data sets. We also aimed to analyze the differences in emotional valence between the 2 data sets to study the relationship between engagement on social media and sentiment.

Methods: We used the Twitter application programming interface to collect tweets that contained the street and brand name of a prescription drug within the tweet. Using BERTopic in combination with Uniform Manifold Approximation and Projection and k-means, we generated topics for the street-name corpus (n=170,618) and brand-name corpus (n=245,145). Valence Aware Dictionary and Sentiment Reasoner (VADER) scores were used to classify whether tweets within the topics had positive, negative, or neutral sentiments. Two different logistic regression classifiers were used to predict the sentiment label within each corpus. The first model used a tweet’s engagement metrics and topic ID to predict the label, while the second model used those features in addition to the top 5000 tweets with the largest term-frequency–inverse document frequency score.

Results: Using BERTopic, we identified 40 topics for the street-name data set and 5 topics for the brand-name data set, which we generalized into 8 and 5 topics of discussion, respectively. Four of the general themes of discussion in the brand-name corpus referenced drug use, while 2 themes of discussion in the street-name corpus referenced drug use. From the VADER scores, we found that both corpora were inclined toward positive sentiment. Adding the vectorized tweet text increased the accuracy of our models by around 40% compared with the models that did not incorporate the tweet text in both corpora.

Conclusions: BERTopic was able to classify tweets well. As with LDA, the discussion using brand names was more similar between tweets than the discussion using street names. VADER scores could only be logically applied to the brand-name corpus because of the high prevalence of non–drug-related topics in the street-name data. Brand-name tweets either discussed drugs positively or negatively, with few posts having a neutral emotionality. From our machine learning models, engagement alone was not enough to predict the sentiment label; the added context from the tweets was needed to understand the emotionality of a tweet.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57885) doi:10.2196/57885

KEYWORDS

digital epidemiology; BERTtopic; Valence Aware Dictionary and Sentiment Reasoner; VADER; sentiment analysis; social media; prescription drugs; prescription; prescriptions; drug; drugs; drug use; platform X; Twitter; tweet; tweets; latent Dirichlet allocation; machine-driven; natural language processing; NLP; brand name; logistic regression; machine learning; health informatics

Introduction

Current State of Social Media for Public Health Surveillance

Social networking websites such as X (formerly Twitter), Facebook, and Instagram are often described as "digital town squares" [1], where people can openly and freely have conversations and discussions about nearly any topic or issue, including those that may not be legal, ethical, or socially acceptable. The broad use and open nature of these conversations have led researchers to use social media to monitor and surveil real-world issues pertaining to public health [2-5]. For example, previous studies have analyzed social media data to develop a real-time influenza surveillance dashboard [6]; monitor the language associated with stress, loneliness, and anxiety during the early months of the US COVID-19 outbreak [7,8]; and track public responses to critical news cycles [9], including recent shifts in US abortion legality [10]. These types of projects focus on extrapolating "real-world" data (such as prevalence rates of influenza or anxiety) from social media discourse. Importantly, numerous practical, analytic, and ethical issues remain to be studied and addressed around the use of social media data for projects that have the potential to directly or immediately impact public or personal health [3,11].

There is a subtle distinction between work seeking to estimate health-related factors such as disease prevalence rates from social media (ie, as described in the previous paragraph) and research specifically focused on understanding public conversations and discourse on social media. There are likely still biases inherent in such analyses (eg, nonindependence of data, platforms' algorithmic drivers of conversation, and trending topics) [3]. At the same time, analysis of discourse does not attempt to extrapolate secondary or tertiary data points outside of the dialogue itself. Instead, it approaches social media as a lens through which we can view naturally occurring conversations to provide insights about the "state of discourse" in the population of social media users. Such conversations have been studied around a diverse multitude of topics, such as national parks in South Africa [12], healthy diets [13], COVID-19 vaccines [5], and mental health during Mental Health Awareness Week [14].

Whose Conversations and Discourse Can Readily Be Studied?

A substantial majority of US persons aged 18 to 64 years use social media, as do nearly half of those aged ≥65 years [15]. Therefore, large-scale analyses of posts on social media can be used to infer how the general population might feel about specific issues (though with caution, as noted in the previous paragraph [3]). One important caveat, though, is that different platforms have different user demographics, and people use different platforms at various rates [16]. For instance, data from several years ago suggest that users of Facebook tend to be older (aged ≥50 years), while users of X (Twitter) tend to be younger (aged 18-29 years) [16,17]. These social media dynamics, which may change over time, can influence the types of content that users post and view.

According to a Pew survey on teenage social media use [18] in 2022, about 54% of all teens reported that it would be difficult to give up social media, and among teens who view social media use positively, 46% of teens reported that the main reason they use social media is to connect and socialize with others [19]. Previously, we noted that X (Twitter) is primarily used by younger populations. On X (Twitter), individuals can not only connect with one another but can also become part of web-based communities that discuss diverse topics.

Learning About Drug Use From Social Media Discourse

The United States is in the midst of a drug overdose epidemic that, in recent years, has claimed >100,000 lives every 12 months [20]. While major strides have been made in attenuating the harm from this crisis, such as increasing access to naloxone [21,22] and harm-reduction strategies [23], the persistence of harmful outcomes associated with drug use suggests that additional strategies and information are needed.

Typically, information about drug use is obtained from investigator-directed research studies [24] (eg, surveys and interviews), and such studies contribute meaningfully to this domain of knowledge. At the same time, such mechanisms rely on researchers' presuppositions about what questions to ask and what topics are important. In contrast, large-scale analyses of social conversations have the potential to elucidate aspects of drug use about which scholars are unaware, or less aware, but that may be important to facilitating harm reduction. This exploratory work can theoretically be used to identify new research strategies, approaches, and theories around drug use that are grounded in inductive analysis of discourse rather than deduced from existing theoretical frameworks. For example, understanding these informal communities can help public health officials better understand real drug use trends that they might see among younger populations. Similarly, learning about the emotional valence of the discussion of specific substances might help inform context-targeted communication strategies.

On the basis of user demographics, when collecting data in 2022, we perceived that X (Twitter) would be a useful source of observational data to understand how young people feel about several types of drugs as well as about drug use more generally. Until recently, X (Twitter) was the social media platform of choice for researchers in this space due to the large amount of short-form textual data available on the platform through its application programming interface (API) [2,6,11,25]. While

XSL•FO
RenderX

recent changes to the API have made research on the platform substantively more prohibitive, recent data sets collected before this change still offer excellent utility. This study used such a data set to better understand the themes, sentiment, and engagement levels for drug-related social media conversations. Principles identified through this work will have utility for natural language processing (NLP) analyses across multiple social media platforms.

Literature Review

Specific to drug use and outcomes, researchers have used data from X (Twitter) to identify adverse drug reactions using methods from machine learning [26-28], monitor population-level opioid abuse in real time [29-31], study user sentiment about specific types of drugs [32,33], and characterize how young people feel about certain drugs like cannabis and drug use more generally [34-36]. A study by Meng et al [37] using data from X (Twitter) found that the types of drugs people used varied by demographic characteristics as well as geographic characteristics. The authors collected 79.8 million tweets and analyzed 699,757 tweets that were related to drug use to find associations between the sentiment recorded in sentiment-related tweets and zip codes by analyzing which drugs were tweeted about the most often using hashtags. Another study by Stevens et al [36] identified which drugs are discussed by younger populations and identified specific themes indicating how young people discuss drug use on social media. Both studies sampled a large amount of data but qualitatively coded a subset of their data set. Taken together, these studies suggest that using social media as a barometer to understand public sentiment may be a fast way to ascertain public sentiment without having to use advanced surveying methodology while avoiding certain implicit assumptions that might be made in such surveys.

Our research builds on these previous studies using a similar-sized X (Twitter) data set to Meng et al [37]. However, our approach was distinct; we leveraged unsupervised machine learning techniques to computationally identify the main themes in our drug use tweet data set instead of manually analyzing tweets looking for mentions of specific drug names. In our prior analysis of this data set [38], we used latent Dirichlet allocation (LDA) to generate topics based on “street-name” tweets (eg, colloquial terms) or “brand-name” tweets (eg, trademarked or generic terms) [39]. Using that method, we found that tweets that fit into these 2 categories had different themes. Tweets that contained the brand or prescription name of a drug (eg, OxyContin, Vicodin, fentanyl, etc) had a higher likelihood of referring to the impact that drug use has in the context of US politics, political conversations, and society at large. This contrasts with tweets that referred to drugs via their street names (eg, Vikes, Oxys, etc), where individuals would, at times, openly and informally discuss their drug use.

Furthermore, in contrast to tweets using street names, LDA more clearly categorized tweets containing brand names of drugs into specific drug categories, and as noted, many such tweets contained discussion of political events. Tweets containing street names were more difficult to classify using LDA for 2 reasons. First, street names for drugs could often refer to other words with different meanings and contexts, leading to 2 tweets that could contain the same term, for example, “vike,” but refer to different things entirely (eg, Vicodin or the Minnesota Vikings). Second, compared with the brand-name data set, people appeared to use informal terms to discuss drug use in unique or different nonpolitical contexts, leading to more topics being needed to accurately understand the corpus. We concluded from our previous study [38] that unsupervised machine learning techniques could be leveraged to understand how the public perceives drug use on social media but that its utility for categorizing tweets using street names for drugs was lower than for tweets using brand-name drugs.

As noted, our previous work used LDA, which relies on probability distributions and word co-occurrences to determine latent topics. To expand on this work, we leveraged a neural network approach to topic modeling called BERTopic [40]. BERTopic relies on semantic word embeddings instead of word co-occurrence, so the algorithm can create coherent topics by understanding the context of each word from pretrained weights. In addition to using BERTopic to perform topic generation, we conducted sentiment analysis on the data that we had collected to understand the intensity and level of emotions associated with each tweet. As part of a larger discussion on digital surveillance of drug-related communication, we sought to expand our previous work by using a more advanced topic modeling tool, in addition to sentiment analysis, to add further context to the types of drug dialogues that may be occurring on the web and to find whether key differences are observed by the type of drug (ie, brand name vs street name). We used Valence Aware Dictionary and Sentiment Reasoner (VADER) scores [41] to characterize the intensity of emotions of each tweet and determined the mean VADER scores for each topic. BERTopic, a newer sentiment analysis tool, is widely viewed as a more accurate topic generator than LDA. Using these methods, our research was guided by three specific research questions:

1. Using a neural network approach to topic modeling, what key semantic and thematic differences are observed in a corpus of tweets pertaining to a drug’s brand name versus street name?
2. Using a lexicon-based sentiment analysis tool, what lexical differences in sentiment are observed in a corpus of tweets pertaining to a drug’s brand name compared with its street name?
3. Using logistic regression, can we accurately predict the VADER-generated sentiment label of a tweet (ie, positive, negative, or neutral) from a tweet’s engagement metrics?

Findings from this study stand to further refine our data by more clearly identifying content not pertaining to drug use or drug communication. The more refined corpus derived from such an approach, with reduced prevalence of extraneous content, can be further leveraged to construct a drug communication classifier that may better assist in analyzing larger, unstructured language data. Furthermore, by comparing results from LDA, a probabilistic approach to topic modeling, and Bidirectional Encoder Representations from Transformers (BERT), a neural network approach to topic modeling, our study stands to document the growing body of research supporting neural network topic modeling as the optimal choice for unsupervised

XSL•FO
RenderX

NLP tasks. Importantly, findings from this study can also inform an additional pipeline to construct a classifier pertaining to drug communication on the web.

Methods

Data Collection

Data for this study were collected from X, the social networking website formerly known as Twitter, between October and December 2022 before the discontinuation of its open-access API. To obtain the data relevant to this study, we leveraged the National Institute on Drug Abuse's list of commonly abused prescription drugs to create the brand-name corpus. See Textbox 1 for a list of all queried drugs, parsed by brand and street names.

Textbox 1. X (formerly Twitter) application programming interface queries by brand name and pseudonym (street name).

Brand-name queries (n=31)

- Xanax, Percocet, Oxycontin, Vicodin, Fentanyl, Opana, Kadian, Avinza, Adderall, Ritalin, Ambien, Sonata, Lunesta, Valium, Librium, Halcion, Ativan, Amytal, Nembutal, Seconal, Roxanol, Duramorph, Actiq, Duragesic, Sublimaze, Tylox, Percodan, Biphetamine, Dexedrine, Concerta, MDMA

Street-name queries (n=33)

- Hillbilly Heroin, Oxy, Oxy 80s, Rushbo, Blue Mollies, Black Mollies, Percs, Happy Pills, Barbs, Phennies, Tooies, Downers, Tranks, A-Minus, Zombie Pills, Skippy, The Smart Drug, Vitamin R, Benzos, Benzies, R-Ball, Crystal Meth, Pep Pills, Ludes, Hydros, Idiot Pills, Watson 387, Dexy, Dexies, Ampes, Super Jellies, Speed Pill, Uppers

Using this list as a reference point, we created bots to run strategic queries and Boolean phrases to collect tweets containing a reference to ≥1 prescription or street-drug names. As a comparative study, we triaged all tweets into one of the following two corpora: (1) a brand-name corpus that comprised all tweets with reference to prescription drugs, branded or technical names, and (2) a street-name corpus that comprised all tweets with reference to colloquial names for those drugs. Before cleaning, we had collected 362,216 (38.79%) tweets containing street-name references and 571,564 (61.21%) tweets that contained references to prescription brand-name drugs, totaling 933,780 tweets. After cleaning the data, which involved standardizing the text to identify and remove duplicates, the brand-name corpus contained 245,145 tweets and the street-name corpus contained 170,618 tweets, for a composite sample size of 415,763 (see Parker et al [38] for further insights into the development of this corpus).

Approaches

Overview

In this study, we combined a variety of NLP and machine learning tasks, including those pertaining to theme generation (neural network topic models), dimensionality reduction, and sentiment detection using VADER. We also used an informal qualitative review of our data and exploratory multinomial logistic regression. We explain each briefly below.

Neural Network Topic Modeling

Topic modeling refers to an NLP technique that uses a series of calculations to extract latent topics or themes from a collection of related documents or texts. We used a neural network topic modeling pipeline by generating topics using BERT vectors. BERT is a powerful, state-of-the-art transformer-based language retrain model that can understand the context and meaning of words and sentences by comparing input data against a large-scale, pretrained data set. BERTopic is a topic modeling technique that uses BERT vectors to extract latent topics from corpora using one of many pretrained transformer models [42]. BERT's ability to generate high-quality word embeddings with clustering techniques produces coherent and semantically and contextually meaningful topics from a corpus of documents. Because the meaning of a word can change depending on the context, this is particularly useful for textual data analysis.

Dimensionality Reduction

Calculating BERT embeddings generated for corpora is computationally expensive and requires substantial computing power to run effectively. Therefore, dimensionality reduction, the process of transforming high-dimensional data into lower-dimensional data while retaining key elements of the data, is a key component of the topic extraction process. To accomplish this, we used 2 approaches: Uniform Manifold Approximation and Projection, a dimensionality reduction tool that can better detect the complex relationships between tweets on the basis of their language, and k-means clustering (k-means), a popular algorithm used for classification, clustering, and topic modeling, which was used as a clustering algorithm to perform topic modeling on BERT embeddings of the corpus data. The fundamental principle of k-means is to split a data set into k-clusters by defining k-centroid values in feature space. These centroids are initially randomly assigned and used to define the clusters. Through iterative assignment, the centroids are updated on the basis of how the data points are placed in the feature space. The choice of "k," representing the number of clusters to consider, is a critical parameter that can be tuned to control the algorithm's sensitivity to local variations in the data.

To find the number of k-topics, we measured the coherence score of different topic configurations. A coherence score [43] is derived from an iterative analysis to identify the optimal number of topics for a given corpus. Coherence scores are a way to evaluate the efficacy of topic models by measuring how well our topics represent the text corpora they are based on. A

XSL•FO
RenderX

coherence score ranges from 0 to 1, and larger scores theoretically equate to more interpretable topics.

Sentiment Analysis

We used VADER [41] to analyze and score the emotionality of our text. VADER is a rule-based tool for sentiment analysis that uses a specialized lexicon to capture both the polarity (positive, negative, and neutral) and the intensity of the sentiments expressed in a text. Unlike traditional sentiment analysis, VADER focuses on context-dependent emotional tones and accounts for nuanced sentiment expressions. This makes VADER particularly useful in deciphering sentiment in social media text, customer reviews, and informal communication, where conventional sentiment analysis techniques might fall short. VADER uses a lexicon of words and phrases, each of which is assigned a sentiment score based on their emotional connotations. Then, from the word order and sentence structure of a document, the intensity of the sentiment changes. For example, a phrase such as "Yay. Another phone interview" has a different sentiment score from "Yay! Another phone interview!" due to the extra exclamation marks, which would result in an increase in the intensity of the score. Sentiment scores in VADER range from –1 (very high negative valence) to +1 (very high positive valence). The sentiment score associated with a tweet is calculated by adding the individual sentiment valence scores from each word that corresponds to a word in the VADER lexicon and considering the punctuation and capitalization of a tweet to adjust the score accordingly. That value is then normalized from –1 to +1. We refer to this as the normalized, weighted VADER compound score (or compound score more generally). Using this number, we can measure the strength of the emotions associated with a tweet. After finding the sentiment compound score, we then classify the score into 3 labels: positive, negative, or neutral. A *neutral sentiment* is any sentiment where the score is between, but does not include, –0.05 and 0.05 [44]. A *positive sentiment* is defined as any VADER score ≥0.05, while a *negative sentiment* is any score ≤–0.05. We then report the percentage of tweets that are positive, negative, or neutral in our corpus. Given our research questions, we extracted a compound VADER score (with a possible range of –0.99 to 0.99) and a label (positive, negative, or neutral) based on our cutoff criteria. Our use of VADER is strongly supported in computational health science research [45-48].

Informal Manual Review

After we extracted latent topics, we applied a sorting function in which tweets in our corpus were assigned to one of the k-corresponding topics on the basis of the presence of topic keywords. Once data in both corpora were sorted into topics, we briefly reviewed a select number of posts for each topic to add context to topic names and keywords. This process is standard for topic modeling analyses, as computers can only extract latent topics and cannot infer deeper meaning with unsupervised NLP methods.

Sentiment Label Prediction

All tweets in our study were collected with their engagement metrics, including likes, replies, and retweets. Previous research [49,50] suggests that certain facets of language including affect (or sentiment), tone, and content are associated with highly positive or negative sentiment content, which in turn is associated with higher engagement on social media. While different engagement metrics (likes, retweets, and replies) are associated with different meanings for people [51], individuals engage more with highly inflammatory content [52,53]. However, there is some disagreement about whether positive or negative content is engaged with more frequently [53]. Here, our objective was to determine whether we could predict the sentiment label of a tweet given its BERT-generated topic and the number of likes, retweets, and replies it has. The sentiment label of a tweet is +1, or 0, or –1, signifying a positive, neutral, or negative sentiment polarity for that tweet, respectively. From past research, we know that tweets with highly emotional language are retweeted more and generally receive more engagement [50,52,54]. We hypothesized that knowing the general content of a post (which is what the topic ID will tell us) and how engaged users are with a tweet would allow for accurate prediction of the sentiment label. To test that idea, for each corpus, we created a regression model to find whether labels can be predicted without needing the tweet text. These models contain covariates; engagement metrics (number of likes, retweets, and replies); and generated topic IDs. In addition, we compared this model with another model that used these variables and added the term-frequency–inverse document frequency (TF-IDF) vectorized clean-tweet text as a covariate to understand if word context was needed to accurately predict the sentiment label. TF-IDF vectorization [55,56] is a method to convert the textual information of a document to a numerical representation where each word in the document is converted to a number representing how important that word is in the corpus. This makes it easier to compare how similar 2 documents are in the corpus. In our exploratory regression models, we used the top 5000 features from each corpus based on the generated TF-IDF scores. By comparing these 2 models, we determined the effect that the context of a tweet has on predicting the emotionality associated with the tweet.

To predict the sentiment labels for each tweet, we used a multinomial multivariate logistic regression model. The purpose of this model was to classify tweets into one of the following three categories: positive (+1), negative (–1), or neutral (0) sentiment tweets. We implemented a classifier that used logistic regression to find the label for each tweet. Since we were interested in whether the label itself could be predicted using engagement metrics and the topic ID, we did not use any specific label type as a reference group and used the one-vs-rest heuristic method to classify labels. To evaluate the efficacy of our models, we used the F_1-score, precision, recall, and accuracy metrics to compare all models. The accuracy metric measures how often the predicted label from a model matches the true sentiment label, while the precision metric measures the proportion of true positives found by the model. The recall metric measures the proportion of true positives identified divided by the sum of true positives and false negatives, while the F_1-score can be defined as the harmonic mean of the recall and precision metrics. This score is the definitive measure of how well a model correctly predicts values since, unlike accuracy, it considers

XSL•FO
RenderX

how often the model classifies outcomes as false positives and false negatives. We used the macroversion of the F_1-score, recall, and precision metrics to account for label imbalance. These metrics are standard for this type of modeling procedure [57] (for more information on macrologistic regression with F_1-score, recall, and precision metrics, see Tarekegn et al [58] and Manning et al [59]). The *sklearn* package (scikit learn) was used to train and test the regression models, and VADER sentiment analysis tools were used from the VADER sentiment python package [41].

Ethical Considerations

The study data were collected using the formerly available Twitter API. All study data consisted of public "Tweets" on the Twitter or X platform. For the sake of this study, usernames and location data were not used for any part of the analysis. Collection and analyses of these data was designated by the Indiana University Institutional Review Board as Exempt (#18081).

Procedure

Data Collection

Over 3 months, we continuously collected data via the (formerly) openly accessible X (Twitter) API using the search terms outlined in Textbox 1. For all brand-specific queries (eg, Adderall, Vicodin, Percocet, etc), we created a singular composite data set, hereafter referred to as the brand corpus (n=245,145), after initially collecting 571,564 brand-related tweets. For all colloquial, slang, or other similar mentions of a drug (ie, Addies, Vikes, Perks, etc), we created a second composite data set, hereafter referred to as the street corpus (n=170,618) after initially collecting 362,216 tweets.

Data Cleaning for BERT and VADER Tasks

After collecting tweets, we began processing the data ahead of the BERT, VADER, and regression analysis. For each data set, we first created a new column named "clean_text," where we copied the nonpreprocessed text. From this new column, we then performed our cleaning operations using regular expressions. First, we removed any URLs, the mention symbol (@), emojis, numbers, punctuation, and special characters. Then, we removed any white space present in each tweet to create consistently spaced text. Next, we removed any unnecessary parts of speech using a lemmatizer in addition to stop words, which typically obfuscate the clarity of topic models. For the BERT analysis, we compositely analyzed the text that was entirely preprocessed, in line with standard topic modeling applications. For the VADER analysis, we analyzed the unprocessed text, in accordance with conventional VADER applications, to ensure that the context (including punctuation, adverbs, and adjectives) was considered in the final sentiment score.

Coherence Score Calculations

Once the data were preprocessed, we performed iterative topic models with coherence score calculation to identify optimal model fit, beginning with baseline recommendations outlined by Parker et al [38]. To perform an iterative BERT analysis, we tested a range of topic model solutions ranging from 10 to 60 topics, iterating by increments of 10 (eg, k=10, 20, 30...60 topics). For the brand-name corpus, we found that a smaller number of topics <10 would be needed to find the optimal coherence score. As such, we tested a range of topics from 5 to 20 in increments of 5 (ie, k=5, 10, 15, 20). After each iteration, we calculated a coherence score, which infers the degree to which a human can intuitively understand what a computer-generated topic represents. Higher coherence scores denote greater clarity; lower coherence scores denote lesser clarity. After running all iterations, we identified a different topic solution per corpus. We identified 5 topics (brand-name coherence=0.699) and 40 topics (street-name coherence=0.600) as the optimal topic fit for our data sets. Once we identified the optimal topic solution for the brand and street corpora, we then created a sorting function that triaged all data points into one of the k-respective topics based on keyword matching. After sorting the data, we performed an informal qualitative review to identify the primary topic themes, which were retrospectively named.

VADER Analysis

We ran the nonprocessed text through the VADER lexicon. For each entry, we calculated the normalized compound sentiment for each tweet. Then, we labeled tweets as having positive, negative, or neutral sentiments if the compound score for sentiment was ≥0.05, between but not inclusive of 0.05 and –0.05, and ≤0.05, respectively, for each label. This threshold value for sentiment is a common standard when using normalized VADER scores [41]. We reported the mean and SD of the compound sentiment score for both corpora. After labeling tweets as positive, negative, and neutral, we counted the number of tweets that contained each label and compared the percentage of positive, negative, and neutral tweets between corpora.

Regression Analysis

For the regression analysis, we used the sentiment labels from our VADER analysis, converting the labels from positive, neutral, and negative to +1, 0, and –1. The data set was split (80:20 ratio) for training and testing, respectively. First, we used logistic regression to predict sentiment labels based on the tweet's topic ID and specific engagement metrics (ie, likes, replies, or retweets). This was conducted separately for each engagement metric; combining them necessitated establishing a method to appropriately weigh the different engagement metrics, since each engagement behavior implies a different degree of "engagement" (eg, "liking" a tweet takes less effort than writing a reply). Next, we applied a multiclass logistic regression to predict sentiment labels, incorporating the topic ID, engagement metrics, and top 5000 features based on their TF-IDF vectorization. Finally, we applied the Limited-memory Broyden-Fletcher-Goldfarb-Shannon optimizer to optimize the weights in our model. We reported the macroaggregated precision [59], recall, accuracy, and F_1-score metrics among the multinomial models [60,61]. This specific type of aggregation was performed since the distribution of sentiment labels was fairly balanced.

XSL•FO
RenderX

Results

Research Question 1: Using a Neural Network Approach to Topic Modeling, What Key Semantic and Thematic Differences Are Observed in a Corpus of Tweets Pertaining to a Drug's Brand Name Versus a Street Name?

Overview

Our neural network topic modeling pipeline identified several noteworthy differences in the brand and street-name corpora. This includes optimal topic size in either corpus, scope of the topics, and relative clarity in the final models. Table 1 provides information about the 5 topics in the brand-name corpus (the optimal number of topics based on the coherence score measurement). In Table 2, we report on the themes of each cluster as reported by BERTopic. We contrast this with the findings in Table 3, where we searched for 40 topics in the street-name data set. We describe the top 10 words in each topic in Table 3; then, we summarize the meaning of the groups in Table 4. The groups were determined qualitatively in Table 4 by cross-referencing Figure 1, based on which topics were overlapping.

Table 1. Brand-name topic ID information, including key terms, count, and percentage of topic ID (n=245,145).

Topic ID	Top 10 search terms per topic	Document count, n (%)
0	Adderall, Ritalin, ADHD[a], amphetamine, stimulant, medication, prescription, drug, prescribed, meth	76,798 (31.33)
1	fentanyl, cartel, Biden, heroin, illegals, crisis, drug, trafficking, Bidens, epidemic	59,382 (24.22)
2	Psychedelics, LSD[b], shrooms, psychedelic, drug, ecstasy, weed, pill, cocaine, ketamine	40,001 (16.32)
3	Xanax, anxiety, Vicodin, drug, pill, prescribed, calm, bar, addicted, panic	37,048 (15.11)
4	Sonata, Beethoven, piano, symphony, Mozart, composer, concerto, allegro, Chopin, moonlight	31,916 (13.02)

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]LSD: lysergic acid diethylamide.

Table 2. Brand-name group information, including key terms, count, and percentage. The qualitative themes were generated based on the top 10 terms seen in Since only 5 topics were found from the BERTopic model, the topics and groups were able to be matched with each other easily (n=245,145).

Group ID	Overarching themes	Topic IDs in group (topics forming groups in Figure 2)	Document count, n (%)
A	Adderall, Ritalin, ADHD[a], stimulant use	0	76,798 (31.33)
B	music, concerts, posts unrelated to drug use	4	31,916 (13.02)
C	psychedelics, LSD[b], hallucinogens	2	40,001 (16.32)
D	Xanax, anxiety, depressants	3	37,048 (15.11)
E	fentanyl, overdose, US politics	1	59,382 (24.22)

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]LSD: lysergic acid diethylamide.

XSL•FO
RenderX

Table 3. Individual street topic information, including key terms, count, and percentage (n=170,618).

Topic ID	Top 10 search terms per topic	Document count, n (%)
0	skippy, skippys, skipp, skip, damned, damn, love, darn, f*ck, hell	10,703 (6.27)
1	barb, barbz, barbed, barbarian, beyhive, fav, time, lmao, stardust, bg	10,339 (6.06)
2	percs, perc, perk, perky, leave, n***a, shit, im, bruh, bro	9634 (5.65)
3	playoff, qbs, fumble, nfl, dallas, 49ers, afc, touchdown, offense, qb	8916 (5.23)
4	pill, happiness, smiling, antidepressant, mood, happy, joy, depression, smile, happiest	8480 (4.97)
5	meth, crystal, methamphetamine, crystalmeth, drug, cocaine, heroin, coke, methclouds, addict	8432 (4.94)
6	vikes, vikesbills, losing, winning, playoff, game, win, loss, lose, beat	7986 (4.68)
7	benzodiazepine, benzos, benzo, xanax, antidepressant, prescribing, antipsychotic, medication, ssri, anxiety	7754 (4.54)
8	barb, barbz, nicki, minaj, rapper, rap, lil, nickis, gang, grammy	7313 (4.29)
9	skippy, skippys, taxpayer, tory, cpc, government, trickle, labour, politician, govt	7200 (4.22)
10	cannabis, marijuana, weed, drug, heroin, psychedelics, shrooms, morphine, cocaine, lsd	6926 (4.06)
11	crackheads, perc, lean, crack, crackhead, shrooms, percs, coke, drug, weed	5611 (3.29)
12	upper, lower, higher, high, knockeruppers, taking, pickeruppers, like, hand, took	5401 (3.17)
13	percs, perc, pop, nigga, poppin, bitch, popping, lil, yo, dat	5112 (3)
14	skippy, skippys, fact, pathetic, racist, ignorance, hate, claim, troll, false	5024 (2.94)
15	torch, welder, welding, wgas, oxys, weld, profitable, ox, oxy, kit	4965 (2.91)
16	trading, stockmarket, market, stock, profit, investing, earnings, investment, marketbreadth, sector	4786 (2.81)
17	murdered, victim, peadophiles, murder, 911, twitter, social, room, downer, dont	4690 (2.75)
18	song, release, album, music, 2019, muddy, toe, tpne, weekend, forever	4431 (2.6)
19	janet, dorothy, barb, betty, robert, love, kitty, miss, rachel, dearest	3865 (2.27)
20	yellow, referee, ref, penalty, foul, match, fifa, fifaworldcup, england, worldcup	3813 (2.23)
21	eileen, dexy, dexys, dex, dexies, dexter, dexytools, dexy_buys, dexy_updates, dextools	3805 (2.23)
22	skol, vikes, vikesbites, skolvikes, gopher, game, team, win, hock, winning	3772 (2.21)
23	house, budget, buying, buy, home, fixerupper, fixer, buyer, #shopmycloset, renovation	3592 (2.11)
24	nsfwtwitte, leakedvideos, nsfwtwt, leakedvideo, nsfwtw, nsfwvid, nsfw, discord, skippyleaks, chastitylifestyle	3331 (1.95)
25	grove, downersgrove, hiring, retailjobs, suburb, downtown, downer, st, naperville, chicago	2604 (1.53)
26	debbiedowners, debbie, downer, debby, nancy, gue, karen, dah, boebert, owl	2488 (1.46)
27	peanut, butter, skippy, snack, jelly, jiffy, reeses, chocolate, nuttin, sandwich	1969 (1.15)
28	gain, daily, gme, reduce, wmt, totalday, mixed, sqqq, amp, pt	1874 (1.1)
29	central, basketball, varsity, halftime, chicago, tonight, livestream, tournament, illinois, east	1487 (0.87)
30	56mmuppers, rifle, firearm, ar15, 9mmuppers, 62x39uppers, blackoutuppers, receiver, armed, barrel	1418 (0.83)
31	rushbo, rushbos, rushie, rush, el, miss, limbaugh, limbaughs, bo, linda	619 (0.36)
32	jordanpeterson, peterson, jordan, shooter, manson, follower, serotonin, twitter, walmart, fan	520 (0.3)
33	spy, trader, chatroom, gden, roku, ccl, gmbl, rgr, rcl, wfc	410 (0.24)
34	pigeon, meth, prison, detained, correctional, backpack, carrying, caught, arrested, smuggle	390 (0.23)
35	blackoutuppers, grape, blackout, upper, receiver, stainless, 316, 300, defense, tactical	377 (0.22)
36	volume, callput, xle, plug, overview, 192, ratio, energy, xrxoxy101, total	167 (0.1)
37	oxy_usdt, wrx_usdt, oxy_usdtsuggested, aln_usdt, xyo_usdt, wncg_usdt, xprt_usdt, usdt, lamb_usdt, aioz_usdt	166 (0.1)
38	meth, jordanpeterson, peterson, serotonin, manson, stimulant, shooter, follower, jordan, cybermen	165 (0.1)
39	stock, group, chatroom, trade, trxc, amd, astx, mgm, gmbl, amzn	83 (0.05)

XSL•FO
RenderX

Table 4. Street name grouped topics including overarching themes, document count, and percentage. The qualitative themes were generated based on the top 10 terms seen in From the 40 topics, 8 groups were found from the overlapping topics seen in Figure 1. Each of the 8 groups has a unique theme associated with it, with differing numbers of topic per theme (n=170,618).

Group ID	Overarching themes	Topic IDs in group (topics that form groups in Figure 1)	Document count, n (%)
F	Group cluster pertaining to sports related topics and themes	3, 6, 20, 22, 29	25,974 (15.22)
G	Group cluster pertaining to pop culture fandoms (eg, the Barbz, a Nicki Minaj fanbase)	1, 8, 19, 31	22,136 (12.97)
H	Grouped cluster pertaining to firearm dialogue and online sales	30, 35	1795 (1.05)
I	Grouped cluster pertaining to stock exchanges (eg, Oxy)	15, 16, 32, 36, 37, 39	10,687 (6.26)
J	Grouped clusters pertaining to Percocet use and access	2, 11, 13, 33	20,767 (12.12)
K	Group of clusters comprising unclear, uncertain topics	12, 17, 18, 21, 23, 24, 25, 26	30,342 (17.78)
L	Grouped clusters pertaining to “Skippy” as a peanut butter brand, drug, and political figure	0, 9, 14, 27	24,896 (14.59)
M	Grouped cluster pertaining to assorted drug use, including meth, crack-cocaine, and others.	4, 5, 7, 10, 28, 34, 38	34,021 (19.94)

Figure 1. Street corpus intertopic distance map denoting topic overlap. The topics were generated from our BERTopic model, and the themes were decided from qualitative analysis of the posts within each topic. From our iterative BERTopic analysis, the number of topics with the highest coherence of 0.600 was 40.

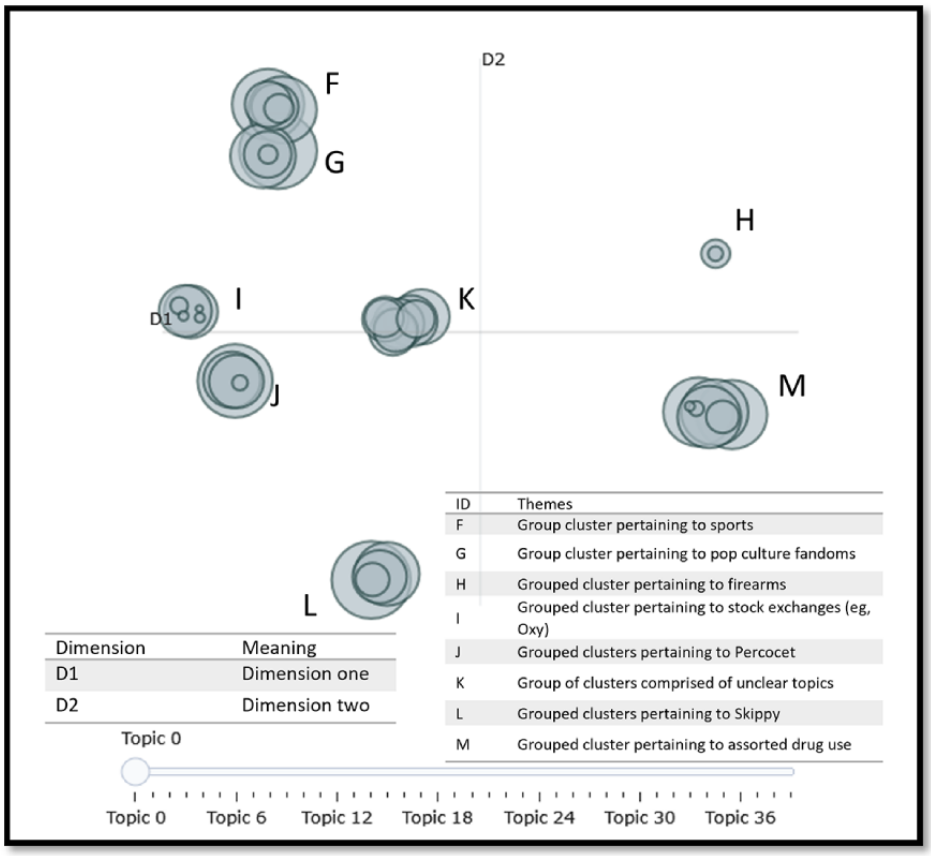

Brand Corpus

Our iterative BERTopic analysis yielded a 5-topic solution (coherence=0.699). Figure 2 provides a visualization of our data using an intertopic distance map. This map allows us to infer the relative similarity (or high correlation) and dissimilarity (or low correlation) of each topic relative to one another. From Figure 2, we can infer 5 mutually distinct topics, which is evidenced by the absence of overlap between clusters. When reviewing each cluster’s keywords, we further inferred that each topic pertained to an overarching drug class. Group A principally referred to stimulant use; group B referred to music, concerts, or tweets otherwise not pertaining to drug use; group C referred to psychedelics and hallucinogens; group D referred to depressants; and group E referred to fentanyl use and overdose.

XSL•FO
RenderX

Figure 2. Brand corpus intertopic distance map denoting topic overlap. The topics were generated from our BERTopic model, and the themes were decided from the qualitative analysis of the posts within each topic. From our iterative BERTopic analysis, the number of topics with the highest coherence of 0.699 was 5. ADHD: attention-deficit/hyperactivity disorder; LSD: lysergic acid diethylamide.

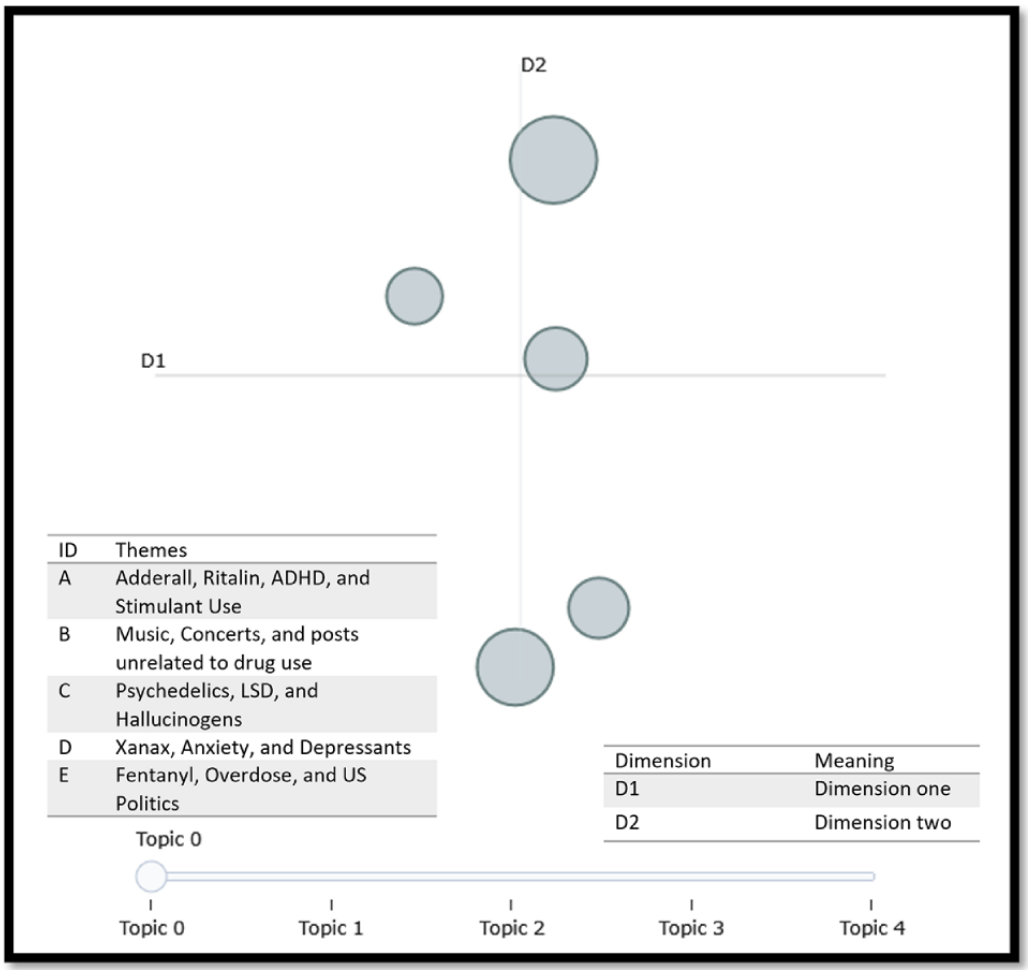

Table 1 offers further context regarding the distribution of topics, while Table 2 shows the relevant groupings and themes based on the topics in Table 1 and the clustering shown in Figure 2. We note that the group ID in Table 1 corresponds to the clusters labeled in Figure 2. The 2 groups with the greatest prominence were group A (76,798/245,145, 31.33%; *Adderall*, *Ritalin*, *ADHD*, and *stimulant use*) and group E (59,382/245,145, 24.22%; *fentanyl*, *overdose*, *US politics*), comprising >55% of the brand-name corpus. Regarding stimulant use, or group A, we observed a variety of different subthemes, including recreational use (tweet: "being on Adderall is so fun bc i just spent 30 minutes watching tik toks of snoopy dancing to different songs") and as a current events topic (tweet: "@JoeBiden what is your plan to fix the adderall shortage?"). The second most prominent theme, fentanyl, or group E, was largely centered on discussing the drug in a strongly political and current events context, often spanning overdose rates and the impact of immigration on fentanyl availability (tweet: "They were killed by people with guns. BTW, you also forgot 108,000 people killed by open borders fentanyl in the last year."). Notably, we did not observe much discussion about the recreational use of fentanyl in our data. Groups C (psychedelics) and D (depressants) largely covered recreational uses of these drugs. However, we did observe a body of tweets advertising the sale of hallucinogenic products in states where their use is ostensibly legal (tweet: "I love microdosing and I gladly recommend [redacted] on Instagram they got shrooms LSD dmt MDMA fast shipping and delivery"). We classified 13.02% (31,916/245,145) of our data into group B, which we qualitatively deemed to contain posts not specific to drug use. Recurring mentions in group B included music, concerts, and car brands (tweet: "The suspect fled the scene in a white, four-door, Hyundai Sonata with an obscured North Carolina temporary tag, according to police"; tweet: "Nice piece, devils trill sonata is a good choice ."), which may be explained by the name, Sonata, and its various associations.

Street Corpus

Our iterative BERTopic analysis yielded a 40-topic solution for the street corpus (coherence=0.600). Figure 1 visualizes our topics using an intertopic distance map where the overlap denotes high topic correlation, and sparsity indicates low topic correlation. Unlike the brand corpus, which contained 5 nonoverlapping topics that could be easily generalized into specific themes, the 40 topics associated with the street corpus had various degrees of overlap, which indicates highly similar, or correlated, topics. When reviewing the keywords associated with each of the 40 topics and associated distributions (Table 3), we categorized our data further along 8 overarching themes as further illustrated in Figure 1. More specifically, clusters

XSL•FO
RenderX

associated with group F were largely about sports, group G about pop culture fandoms, group H about firearms, group I about the stock exchange, and group J about Percocet, while group K contained unclear focus, group L contained a variety of tweets about "skippy" in various contexts, and group M contained posts about assorted drug use.

Table 4 offers further context regarding general topic distribution and group clustering. There were fewer topics pertaining exclusively to drugs and drug use in the street corpus. In place of such drug-related conversations, we instead observed a disjointed collection of topics that were either not clear (group K: 30,342/170,618, 17.78% representation) or more succinctly focused on non–drug-related topics including sports (group F: 25,974/170,618, 15.22% representation), pop culture fanbases (group G: 22,136/170,618, 12.97% representation), firearm dialogues and sales (group H: 1795/170,618, 1.05%), stock prices and sales (eg, OXY; group I: 10,687/170,618, 6.26% representation), and myriad uses for the term "skippy," (group L: 24,896/170,618, 14.59% representation). Importantly, these non–drug-related topics all contained the appropriate query name, yet the foci of the tweets were decisively not drug related. For example, tweets regarding sports referenced the Minnesota Vikings using their common nickname, "the vikes" (tweet: "Ya, the unknown clock. The vikes would get screwed on that one. I promise you that"). For fandom, we observed a substantive body of tweets about Nicki Minaj's fanbase, commonly referred to as "the barbz" (tweet: "Barbs weird always wanting Nicki to be friends with people who don't like her"). Barbs, or barbz, also refers to a common street name for barbiturates. For stock prices, tweets referenced Occidental Petroleum Corporation, listed on the US Stock Exchange, as "OXY" (tweet: "I'm also very bullish on $OXY stock"). Skippy often referenced a peanut butter brand (tweet: "id honestly put skippy peanut butter in my top five favorite foods") and also referenced Canadian politician Pierre Poilievre, leader of the Conservative Party of Canada [62] (tweet: "Yet another one that Skippy, nor the Conservatives have a solution to address. Just like when they voted against dental care for children."). However, despite the noise inherent to these conflated topics, we also observed numerous instances in which a tweet referenced a particular query and was, in fact, drug related.

After an informal qualitative review, we determined that approximately 32% of posts (groups J and M) pertained directly to drug use. In contexts where a post was about a specific kind of drug use, we observed more direct statements about recreational use. We also determined groups J and M largely, and nearly exclusively, referred to drug use in a recreational and often light-hearted context (tweet: "Honestly, most of the prosecutors I know were also coked out—it's refreshing to see a cop who loves downers so much"; tweet: "Ohh yeah ladies, I forgot to mention they had me on downers and I smoked pot.").

Contrast Between Corpora

We observed both obvious and nuanced differences between corpora. First, the BERT-identified optimal number of topics differed between the brand corpus and the street corpus, which may reflect the relative consistency of brand-related content and the broad diversity of the street-related content. Indeed, in the brand corpus, we observed consistent discussions of a drug in a recreational context. However, we also consistently observed how certain drugs, including fentanyl and Adderall, we often discussed in a current events context (ie, the nationwide Adderall shortage) or in a sociopolitical context (ie, immigration and its effects on fentanyl distribution along the southern border). These more formal pockets of conversation were almost entirely lacking in the street corpus where only a small portion of the tweets explicitly mentioned drug use; nevertheless, we acknowledge that a full review of each tweet was not undertaken. When it was apparent that a tweet contained an appropriate query but no mention of a drug, we observed the content pertaining to the term's other potential applications or uses. Unique to the street corpus seemed to be more positive mentions of a given drug, typically in a recreational use context or as a light-hearted exchange. Many tweets in the street corpus also had limited context, making it difficult for a computer or members of the study team to appropriately categorize (tweet: "OMG. I love the barbz so much"; tweet: "Gotta love my Vikes"). Thus, despite leveraging a more refined algorithm to conduct a topic modeling analysis (in contrast to our prior use of LDA), there was still an inherent messiness to these data that require further refinement and consideration.

Research Question 2: Using a Lexicon-Based Sentiment Analysis Tool, What Lexical Differences in Sentiment Are Observed in a Corpus of Tweets Pertaining to a Drug's Brand Name Compared With its Street Name?

In addition to content differences in the brand and street-name BERTopic analysis, we also identified affective similarities and differences using VADER, a lexicon-based sentiment analysis tool. For this analysis, we extracted the compound VADER score per tweet, which ranged from –0.99 to 0.99, and emphasized a tweet's valence intensity. We also extracted the sentiment label (positive, negative, or neutral) based on our cutoff criteria. We observed key differences by score and label. First, the mean VADER compound score for the brand corpus was between –0.05 and 0.05, showing that the mean sentiment was neutral, while the mean compound score for the street corpus was >0.05, indicating an inclination toward positive sentiment in the street corpus (mean brand compound score –0.0082 SD 0.477; mean street compound score 0.11 SD 0.478). However, the SD for the compound scores was large in both corpora (0.47), and this suggests that we cannot broadly generalize the sentiment in the street or brand corpus as being predominantly positive, negative, or neutral.

Table 5 shows the percentage of tweets in each corpus that fit within a specific sentiment label. These percentages do not account for the magnitude of a tweet's sentiment score. Tweets that were only slightly positive (eg, 0.051) were labeled as having positive sentiment and were categorized alongside extremely positive tweets (eg, 0.80). In other words, our findings indicate both the overall magnitude of sentiment across all tweets in a corpus (the aforementioned mean compound scores) as well as the prevalence of tweets classified with each sentiment label according to our established cutoff scores (Table 5).

One explanation for lower average VADER scores in the brand corpus may be the political nature of a substantive body of these

XSL•FO
RenderX

tweets. For example, tweets about fentanyl often emphasized overdose, border security, and other similarly tense political dynamics, which were largely absent from the street corpus. This distinction may also explain the greater presence of tweets tagged with a "positive" VADER value in the street corpus (77,543/170,618, 45.45%) versus the brand corpus (88,826/245,145, 36.23%). Other insights gleaned from VADER include a smaller number of tweets tagged as negative in the street corpus compared with the brand corpus (street: 47,603/170,618, 27.9%; brand: 86,586/245,145, 35.32%). Both corpora contained similar amounts of posts with a neutral sentiment.

Table 5. Sentiment percentages for brand and street corpus on the basis of computer-assigned sentiment labels (positive, negative, or neutral).

Corpus	Positive sentiment total	Negative sentiment total	Neutral sentiment total
Brand name	36.2	35.3	28.5
Street name	45.4	27.9	26.7

Research Question 3: Can We Accurately Predict the VADER-Generated Sentiment Label of a Tweet (ie, Positive, Negative, Neutral) From a Tweet's Engagement Metric?

When performing logistic regression with the brand-name corpus, the model excluding the text of a tweet as part of the features was 38.5% accurate on average across all engagement metrics compared with the model including tweet text as a feature where the model was 82.8% accurate on average. Similarly, the average model accuracy in the street-name corpus for the model excluding tweet text was 46.7% accurate, while the model including tweet text was 85.4% accurate. In both corpora, the models that included the text of a tweet as a feature when performing TF-IDF vectorization were more accurate by around 40%. We noted negligible differences in accuracy when comparing the likes, retweets, and replies models to each other within each corpus. The macro F_1-scores were even more different between the models that did not use the vectorized text (brand=0.231; street=0.214) compared with the models that did use the vectorized text (brand=0.828; street=0.854). Summary statistics are shown in Table 6.

Table 6. Summary statistics for regression models[a].

Corpus	Accuracy (%)	Precision (%)	Recall (%)	F_1-score (%)
Brand name				
Without text	38.5	58.3	33.5	23.1
With text	82.8	82.8	83	82.8
Street name				
Without text	46.7	35	33.4	21.4
With text	85.4	84.8	84.6	84.7

[a]For both the street- and brand-name corpus, we find that adding the text of the tweet as a feature to our regression model greatly improved the accuracy, precision, recall, and F_1-score compared with the model that did not incorporate this feature.

Discussion

Principal Findings

Overview

This study used a neural network approach to topic modeling (BERTopic) to examine 2 contemporaneous corpora of tweets selected for brand and street-name drug references. Interestingly, differences in the interpretability between the corpora that we first observed with LDA [38] remained salient with this more advanced approach. Then, using VADER, we identified that the street-name corpus has a larger inclination toward positive sentiment, while the brand-name corpus contains similar amounts of tweets labeled positive and negative. Finally, we combined the results from the topic model and sentiment analysis to create predictive models (logistic regression) to estimate sentiment labels from the topic ID and engagement metrics and compared the accuracy of the models that included the vectorized tweet text as a covariate and the models that did not.

Topic Analysis

BERTopic, in combination with Uniform Manifold Approximation and Projection and k-means clustering, yielded statistically coherent clustering of topics, although the outputs for the street-name corpus were more difficult to interpret and generalize. The tweets in the brand-name corpus discussed different drugs in the context of their intended uses, as well as how certain drugs were perceived to relate to ongoing political or social issues. The brand-name data set could be reduced to 5 major themes: broad discussion about fentanyl use and its discussion in a sociopolitical context; stimulant use (eg, Adderall, Ritalin, etc); discussion about music and car models related to the word "sonata"; psychedelic use; and discussion about anxiety-related medication (Xanax). The discourse about fentanyl was especially varied, with many topics containing posts relating to politics, immigration, border security, and, in some cases, actual use. This differed from how people discussed Adderall; in our data, people were concerned about the 2022 Adderall shortage [63] and were interested in how to use the drug safely. As we indicated in the *Results* section, Sonata, the

XSL•FO
RenderX

brand name of a sleep aid, tended to capture tweets about music and the Hyundai Sonata car model, and those tweets formed the only topic and category that was not drug related.

For the street-name corpus, the BERTopic model with the highest statistical coherence score produced 40 topics, many of which overlapped and were not necessarily related to drug use. Only 32.11% (54,788/170,618) of all tweets were sorted into topics that pertained primarily to drug use, allowing the inference that most posts pertained to nongermane topics. Observationally, this was because many street names for drugs can refer to a variety of real-world concepts or phenomena (eg, words do not necessarily refer to a drug without additional context). Previous research supports the idea that machine-based NLP approaches may struggle to parse content containing street names for drugs effectively [38,39,64]. In the street-name corpus, 6 of the 8 clusters were sorted around terms unrelated to drug use. Out of these 6 clusters for the street-name corpus, 4 (67%) clusters (60,592/170,618, 35.51% of all posts) contained themes relating to football, fandoms, firearms, and the stock market. The last 2 clusters were even more difficult to categorize: we could only find general themes relating to the word "Skippy" (sometimes used colloquially to refer to stimulants) for one, and the other did not appear to us (as human interpreters) to have a core theme, although the NLP approach had a computational reason for generating the topics and cluster.

Comparing the topics in the 2 corpora, 10 (25%) out of 40 topics in the street-name corpus contained <1% of all posts, whereas the brand-name corpus had only 5 topics total. The street-name corpus contained many niche discussion topics compared with the few general themes of the brand-name corpus. On the basis of our findings from the BERTopic output, we suspect that refining a complex data set of this size by eliminating content that is not drug specific would be arduous. However, in moving from LDA [38] to BERTopic (a more refined algorithm), we were better able to identify pockets of conversation that were not drug specific and were better able to tag them appropriately. Future research should consider additional work in data refining and classifier building with the street-specific data set.

Sentiment Analysis and Predictive Modeling

We used VADER to assess the sentiments of tweets and found that both corpora contained tweets with a wide range of sentiments. Interestingly, we found that the street-name corpus had a larger proportion of positively labeled tweets compared with the brand-name corpus. In our study, the terminology categorization for street-drug terms was complex, which may raise questions as to VADER's applicability. However, VADER's original validation study was particularly successful at classifying tweets or microblog text (vs other forms of text), outperforming even human raters, and the dictionary of lexical features was designed, in principle, to be domain agnostic [41]. This increases our confidence in the VADER-based assessment of the data. We hypothesize that the street-name corpus was made up of many topics that are unrelated to drug use. Therefore, we suspect that many positive tweets were support from fans, such as fans of Nicki Minaj (barbs) and the Minnesota Vikings (Vikes). However, this analysis pipeline was not able to directly link words and sentiment, so we cannot be sure whether that was the case.

Since the language features associated with emotionality were based on the VADER lexicon, we can know *what* kinds of things were scored as positive but not *why* those features were used to express a certain sentiment. Understanding the motivations behind positive communication is an important next step in understanding how individuals feel about drug use at scale. Arguendo, it might be the case that lexical features (eg, words, capitalization, context, punctuation, etc) associated with positive sentiment occur more often in drug discourse during events (eg, concerts) than drug discourse referring to isolated or solo use. To truly understand why individuals feel a certain way about different types of drug use would require additional deep qualitative methods and analysis. We used multinomial logistic regression to understand if we could predict the sentiment label or emotionality of a tweet using information about the tweet's topic and how engaged users are with the tweet. We tested permutations of regression models that either (1) included tweet text as a covariate or (2) did not include it. We found that the models including the tweet text as a covariate explained more variation in tweet sentiment (by approximately 60% according to the macro F_1-score) than the models that did not incorporate text as a feature. This result was consistent across both corpora, showing that the generated topic ID and engagement metrics were not sufficient to predict the sentiment of a given tweet. Given the variables to which we had access, the only way to accurately predict tweet sentiment was to use the language itself. This means that aspects *about* a tweet, such as what it discusses (its topic ID) and how engaged people are with a tweet (number of likes, replies, and retweets), cannot be used to accurately predict the emotionality of a given tweet. This speaks to the diversity of opinions within a topic and how difficult it is to understand the sentiment of a tweet without knowing the full context within a post. Without the full context, we cannot predict whether a tweet about drug use will have positive or negative sentiment, even if we know what drug is being discussed and how well engaged people are with a post.

Comparison to Previous Literature

In the peer review for a paper on our previous LDA model (Parker et al [38]), reviewers suggested that an appropriate next step would be the use of neural network modeling, which we performed here. The results of the BERTopic model support the conclusions from the LDA model. Specifically, the brand-name corpus was more easily categorized by a machine-based approach than the street-name corpus. As before, this difference seemed attributable to the fact that many of the words in the street corpus do not have a clear meaning outside of a narrow context. For instance, the word "Skippy" can refer to methylphenidate (eg, Ritalin), a brand of peanut butter, or a Canadian politician. In contrast, "fentanyl" has an unambiguous meaning even without context.

The most obvious difference between the models is the number of topics generated. In our prior work, the LDA model generated 20 topics for the brand-name data set, while in this paper, the optimal BERTopic model was able to use 5 topics to cluster all posts. In contrast, the harder-to-parse street-name corpus resulted

XSL•FO
RenderX

in more similar numbers of topics for LDA and BERTopic (35 and 40, respectively). The BERTopic analysis could more clearly delineate the different topics of discussion based on word context, allowing for an increased number of topics for the street-name data set and fewer topics in the brand-name corpus since discussion in the brand-name corpus is more homogenous and easily categorizable. The BERTopic model generated more cohesive themes than the LDA model due to pretrained BERT embeddings, which accurately captured the semantic relationships between words; thus, words with multiple meanings are better understood and categorized. In contrast, LDA uses word co-occurrence to generate topics for tweets, so LDA topic models might group documents together into the same topic that have the same word although this word is used in different contexts. As an example, "Adderall" can co-occur alongside other words like "anxiety" and "Ritalin." In Parker et al [38], the LDA model created 4 separate topics relating to Adderall use and 1 topic relating to the Adderall shortage. However, as we see from our BERTopic model, the more sophisticated algorithm was able to condense those same 4 topics into 1 topic relating to Adderall use, while discussion about the shortage was grouped into the topic relating to the intersection between politics and drug use.

Previous work by Nasralah et al [65] used LDA to better understand the most-discussed topics relating to the opioid epidemic by analyzing 503,830 tweets and filtering tweets via an evaluation matrix. Similar work [66,67] analyzing people's reactions to the opioid epidemic has been conducted using textual analysis algorithms to find themes in X (Twitter) data. A study by Tassone et al [68] used convolutional neural networks and other deep learning techniques to classify whether tweets about drug use were encouraging drug use (positive) or discouraging drug use (negative) and created synthetic tweets on drug use based on real tweets about drugs. While that approach also incorporated sentiment, our definition of a positive or negative tweet was dependent on the VADER classification instead of defining based on whether a tweet encourages or discourages drug use. In addition, we used a semisupervised technique (BERTopic) to classify tweets into general themes. Many studies [32,69,70] that identify themes for a collection of tweets pertaining to drug use using manual annotation methods, including inductive and deductive qualitative coding, have also been conducted. In 2022, Al-Garadi et al [71] used LDA and VADER scores to understand the different reasons for nonmedical prescription drug use. Cavazos-Rehg et al [34,35] focused on a single drug, marijuana, and how young people discuss marijuana use and react to popular accounts that discuss marijuana use. Both studies from Cavazos-Rehg et al [34,35] assessed sentiment using Twitter, but instead of analyzing sentiment using VADER, they used a crowd-sourcing service to code the sentiment of the tweets. In contrast, we used a classifier model across a wide variety of prescription drug conversations on Twitter rather than using human coders.

Strengths and Limitations

Our study's strength lies in the cohesive topics generated by BERTopic, which enabled a clear understanding of the general themes of discussion in the street- and brand-name corpora. However, there are some limitations to our study. First, we cannot distinguish *why* the amount of positive sentiment differed between the brand and street tweets. The VADER analysis that we performed was descriptive in nature, and although we found the sentiment label and compound score of each tweet, we could not summarize why X (Twitter) users expressed positive or negative sentiments about a drug. Some form of stance detection would have to be conducted to better understand how different users feel about specific drugs. From the VADER scores, we can only identify aggregate trends regarding sentiment and not make conclusions about how individuals feel about specific types of drugs.

In our text-comprehensive regression models, we classified the sentiment labels of tweets with a macro F_1-score of 82.8% in the brand-name corpus and 84.7% in the street-name corpus. Our modeling shows that sentiment labels can best be predicted using the cleaned text of a tweet as part of the feature set including engagement metrics and topic ID. However, without the text of a post, the F_1-score fell to 23.1% in the brand-name corpus and to 21.4% in the street-name corpus. This points to a limitation of topic modeling, that it is primarily an exploratory form of analysis that cannot tell us about the emotionality of a data set. Topic models can help researchers find the general ways how people are discussing a topic, but these topics can neither be used to predict the sentiment within the topic, nor, more obviously, allow deeper inferences about motivations and intentions.

We were also limited by VADER, which is a lexicon-based sentiment analysis tool. Although the use of VADER is widely supported in the literature, there are concerns that VADER scores could be biased due to the overrepresentation or absence of certain words in the lexicon. In our case, certain slang terms for prescription drugs such as "perc" or "fent" are not present in the VADER lexicon as well as certain prescription drug names like Adderall or Ritalin. For our work, we were more interested in the *context* around certain prescription drug names and slang terms. We wanted to understand the emotional affect around certain terms, not necessarily the affect of the term itself. For future work, more work could be done to expand the VADER lexicon to include slang terms in addition to prescription drug names.

One final limitation is the lack of generalizability in our study. From the time we collected our data, Twitter has been rebranded to X, and the number of active users, the way that users interact with the site, and the algorithm to show users' content have all changed. We are not able to replicate our study since acquiring the volume of data that was available in the past is not feasible. The "infoveillance" component of our analysis is also put under question since geotagging is no longer available. The future of this type of research must be found on other social networking platforms, such as Facebook, Instagram, and BlueSky, which offer first-party APIs to track their data, and through platforms like PushShift, which is a third-party API for Reddit data.

Implications

Our findings broadly illustrate the importance of using more advanced computational approaches to mine social media data for conversations mentioning prescription drugs. In this section,

XSL•FO
RenderX

we offer some practical implications of our study, including the importance of a refined data set for classifier construction and the need for more advanced sentiment analysis tools.

Our BERTopic model classified the street- and drug-name corpora into a coherent set of individual topics, leading to a higher number of topic clusters in the street-name data set and fewer (only 5) topics for the brand-name data set. By leveraging BERTopic and regression models, we were able to further refine our data set, capturing more nuanced topic meaning to create a future classifier pertaining to web-based communication about drug use. More importantly, we were able to further isolate extraneous content (ie, tweets about cars, fanbases, and sports teams), which, theoretically, would impede the ability to train an accurate classifier. We have taken the first steps to build this classifier by identifying extraneous content. The next step would be to begin a manual annotation process of the refined data set using qualitative expertise to "tag" our data and begin a test-retest approach with training and validation data.

Using VADER, we identified tweets as having positive, negative, or neutral sentiments. Then, we compared the percentages of positive, negative, and neutral tweets between the 2 corpora. This type of analysis allows us to characterize the sentiment in aggregate for the brand and street corpora. To further understand the sentiment that users on X (Twitter) have toward certain drugs, we need to perform more text filtering to find what specific words and phrases are used with certain drug-related words. The next steps include conducting an analysis to identify the lexicon surrounding the street and brand names of prescription drugs to form a better understanding of how certain drugs are discussed. With a more refined data set enhanced by qualitative coding, we may begin to build a training data set that could contain social media illicit drug use conversation data useful for designing health communication interventions.

Conclusions

This work has shown how data from X (Twitter) can be used to identify topical trends surrounding both informal and formal discussions of drug use among users on the platform. Our work combines topic modeling and sentiment analysis to give greater detail on how users on X (Twitter) feel about different types of prescription drugs. Consistent with Parker et al [38], we found that colloquialisms used in the street-name corpus disguise how people discuss drug use. The improved clustering offered by BERTopic allowed us to identify cohesive themes in the street- and brand-name corpora. The clear themes shown in the brand-name corpus contrast with the difficulties in parsing how individuals discuss street-name prescription drug use. From our literature review, we could not find many other works that captured the difficulties in trying to understand how individuals discuss street-name drug use. This points to a potential gap in the drug-discussion literature on how to analyze drugs when their street names are used. Furthermore, VADER analysis detected more positive sentiment among discussions in the street-name corpus compared with the brand-name corpus. Regression analysis of this classifier model determined that predicting the sentiment of drug use discussion is difficult without the full discussion context; topic and engagement metrics alone were insufficient to predict the sentiment of a street- or brand-name tweet.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Yeung D. The 'digital town square' problem. RAND Corporation. 2023 Jan 13. URL: https://www.rand.org/blog/2023/01/the-digital-town-square-problem.html [accessed 2024-04-29]
2. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/ajph.2016.303512]
3. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41(1):101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
4. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013 Dec 9;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
5. Guntuku SC, Buttenheim AM, Sherman G, Merchant RM. Twitter discourse reveals geographical and temporal variation in concerns about COVID-19 vaccines in the United States. Vaccine 2021 Jul 05;39(30):4034-4038 [FREE Full text] [doi: 10.1016/j.vaccine.2021.06.014] [Medline: 34140171]
6. Lee K, Agrawal A, Choudhary A. Real-time disease surveillance using Twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013 Presented at: KDD '13; August 11-14, 2013; Chicago, IL p. 1474-1477 URL: https://dl.acm.org/doi/10.1145/2487575.2487709 [doi: 10.1145/2487575.2487709]

XSL•FO
RenderX

7. Guntuku SC, Sherman G, Stokes DC, Agarwal AK, Seltzer E, Merchant RM, et al. Tracking mental health and symptom mentions on Twitter during COVID-19. J Gen Intern Med 2020 Sep 07;35(9):2798-2800 [FREE Full text] [doi: 10.1007/s11606-020-05988-8] [Medline: 32638321]
8. Clark E, Fredricks K, Woc-Colburn L, Bottazzi ME, Weatherhead J. Disproportionate impact of the COVID-19 pandemic on immigrant communities in the United States. PLoS Negl Trop Dis 2020 Jul 13;14(7):e0008484 [FREE Full text] [doi: 10.1371/journal.pntd.0008484] [Medline: 32658925]
9. Anwar A, Ilyas H, Yaqub U, Zaman S. Analyzing QAnon on Twitter in context of US elections 2020: analysis of user messages and profiles using VADER and BERT topic modeling. In: Proceedings of the 22nd Annual International Conference on Digital Government Research. 2021 Presented at: dg.o '21; June 9-11, 2021; Omaha, NE p. 82-88 URL: https://tinyurl.com/2a6b4h6r [doi: 10.1145/3463677.3463718]
10. Valdez D, Goodson P. Neutral or framed? A sentiment analysis of 2019 abortion laws. Sex Res Social Policy 2022 Jan 18;19(3):936-945 [FREE Full text] [doi: 10.1007/s13178-022-00690-2] [Medline: 35069923]
11. Weeg C, Schwartz HA, Hill S, Merchant RM, Arango C, Ungar L. Using Twitter to measure public discussion of diseases: a case study. JMIR Public Health Surveill 2015 Jun 26;1(1):e6 [FREE Full text] [doi: 10.2196/publichealth.3953] [Medline: 26925459]
12. Mangachena JR, Pickering CM. Implications of social media discourse for managing national parks in South Africa. J Environ Manage 2021 May 01;285:112159. [doi: 10.1016/j.jenvman.2021.112159] [Medline: 33631485]
13. Lynn T, Rosati P, Leoni Santos G, Endo PT. Sorting the healthy diet signal from the social media expert noise: preliminary evidence from the healthy diet discourse on Twitter. Int J Environ Res Public Health 2020 Nov 18;17(22):8557 [FREE Full text] [doi: 10.3390/ijerph17228557] [Medline: 33218105]
14. Makita M, Mas-Bleda A, Morris S, Thelwall M. Mental health discourses on Twitter during mental health awareness week. Issues Ment Health Nurs 2021 May 14;42(5):437-450. [doi: 10.1080/01612840.2020.1814914] [Medline: 32926796]
15. Social media use by age. Pew Research Center. URL: https://www.pewresearch.org/internet/chart/social-media-use-by-age/ [accessed 2024-04-29]
16. Atske S. Social media use in 2021. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-04-04]
17. Hughes A, Wojcik S. Sizing up Twitter users. Pew Research Center. 2019. URL: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ [accessed 2024-04-29]
18. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2024-04-29]
19. Rainie MA, Vogels EA, Perrin A, Rainie L. Connection, creativity and drama: teen life on social media in 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/11/16/connection-creativity-and-drama-teen-life-on-social-media-in-2022/ [accessed 2024-04-29]
20. U.S. overdose deaths decrease in 2023, first time since 2018. Centers for Disease Control and Prevention. 2024 May 15. URL: https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2024/20240515.htm [accessed 2024-04-29]
21. Janet W, Sean M, Czarina B. Expanding access to naloxone: a review of distribution strategies. Health Policy and Services Research. 2019. URL: https://digirepo.nlm.nih.gov/catalog/nlm:nlmuid-101750194-pdf [accessed 2024-04-29]
22. Penm J, MacKinnon NJ, Boone JM, Ciaccia A, McNamee C, Winstanley EL. Strategies and policies to address the opioid epidemic: a case study of Ohio. J Am Pharm Assoc (2003) 2017 Mar;57(2S):S148-S153 [FREE Full text] [doi: 10.1016/j.japh.2017.01.001] [Medline: 28189539]
23. Ellis JD, Dunn KE, Huhn AS. Harm reduction for opioid use disorder: strategies and outcome metrics. Am J Psychiatry 2024 May 01;181(5):372-380. [doi: 10.1176/appi.ajp.20230918] [Medline: 38706335]
24. National survey on drug use and health: summary of methodological studies, 1971–2014. Substance Abuse and Mental Health Services Administration. 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK519735/ [accessed 2024-04-29]
25. Cormier M, Cushman M. Innovation via social media - the importance of Twitter to science. Res Pract Thromb Haemost 2021 Mar;5(3):373-375 [FREE Full text] [doi: 10.1002/rth2.12493] [Medline: 33870022]
26. Bian J, Topaloglu U, Yu F. Towards large-scale Twitter mining for drug-related adverse events. In: Proceedings of the 2012 ACM International Workshop on Smart Health and Wellbeing. 2012 Oct 29 Presented at: SHB'12; 2012; Maui, HI p. 25-32 URL: https://europepmc.org/abstract/MED/28967001 [doi: 10.1145/2389707.2389713]
27. Masino AJ, Forsyth D, Fiks AG. Detecting adverse drug reactions on Twitter with convolutional neural networks and word embedding features. J Healthc Inform Res 2018 Jun 12;2(1-2):25-43 [FREE Full text] [doi: 10.1007/s41666-018-0018-9] [Medline: 35415401]
28. Moh M, Moh TS, Peng Y, Wu L. On adverse drug event extractions using twitter sentiment analysis. Netw Model Anal Health Inform Bioinforma 2017 Sep 18;6(1):18. [doi: 10.1007/s13721-017-0159-4]
29. Sarker A, O'Connor K, Ginn R, Scotch M, Smith K, Malone D, et al. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. Drug Saf 2016 Mar 9;39(3):231-240 [FREE Full text] [doi: 10.1007/s40264-015-0379-4] [Medline: 26748505]

XSL•FO
RenderX

30. Sarker A, Gonzalez-Hernandez G, Perrone J. Towards automating location-specific opioid toxicosurveillance from Twitter via data science methods. Stud Health Technol Inform 2019 Aug 21;264:333-337 [FREE Full text] [doi: 10.3233/SHTI190238] [Medline: 31437940]
31. Phan N, Chun SA, Bhole M, Geller J. Enabling real-time drug abuse detection in Tweets. In: Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering. 2017 Presented at: ICDE '17; April 19-22, 2017; San Diego, CA p. 1510-1514 URL: https://ieeexplore.ieee.org/document/7930118 [doi: 10.1109/icde.2017.221]
32. Tofighi B, Aphinyanaphongs Y, Marini C, Ghassemlou S, Nayebvali P, Metzger I, et al. Detecting illicit opioid content on Twitter. Drug Alcohol Rev 2020 Mar 22;39(3):205-208 [FREE Full text] [doi: 10.1111/dar.13048] [Medline: 32202005]
33. Chan B, Lopez A, Sarkar U. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. PLoS One 2015 Aug 7;10(8):e0135072 [FREE Full text] [doi: 10.1371/journal.pone.0135072] [Medline: 26252774]
34. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza RA, Bierut LJ. Twitter chatter about marijuana. J Adolesc Health 2015 Feb;56(2):139-145 [FREE Full text] [doi: 10.1016/j.jadohealth.2014.10.270] [Medline: 25620299]
35. Cavazos-Rehg P, Krauss M, Grucza R, Bierut L. Characterizing the followers and tweets of a marijuana-focused Twitter handle. J Med Internet Res 2014 Jun 27;16(6):e157 [FREE Full text] [doi: 10.2196/jmir.3247] [Medline: 24974893]
36. Stevens RC, Brawner BM, Kranzler E, Giorgi S, Lazarus E, Abera M, et al. Exploring substance use tweets of youth in the United States: mixed methods study. JMIR Public Health Surveill 2020 Mar 26;6(1):e16191 [FREE Full text] [doi: 10.2196/16191] [Medline: 32213472]
37. Meng HW, Kath S, Li D, Nguyen QC. National substance use patterns on Twitter. PLoS One 2017 Nov 6;12(11):e0187691 [FREE Full text] [doi: 10.1371/journal.pone.0187691] [Medline: 29107961]
38. Parker MA, Valdez D, Rao VK, Eddens KS, Agley J. Results and methodological implications of the digital epidemiology of prescription drug references among Twitter users: latent Dirichlet allocation (LDA) analyses. J Med Internet Res 2023 Jul 28;25:e48405 [FREE Full text] [doi: 10.2196/48405] [Medline: 37505795]
39. Katsuki T, Mackey TK, Cuomo R. Establishing a link between prescription drug abuse and illicit online pharmacies: analysis of Twitter data. J Med Internet Res 2015 Dec 16;17(12):e280 [FREE Full text] [doi: 10.2196/jmir.5144] [Medline: 26677966]
40. Xu WW, Tshimula JM, Dubé È, Graham JE, Greyson D, MacDonald NE, et al. Unmasking the Twitter discourses on masks during the COVID-19 pandemic: user cluster-based BERT topic modeling approach. JMIR Infodemiology 2022 Dec 9;2(2):e41198 [FREE Full text] [doi: 10.2196/41198] [Medline: 36536763]
41. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Weblogs Soc Media 2014 May 16;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]
42. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv Preprint posted online on March 11, 2022 [FREE Full text]
43. O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications 2015 Aug;42(13):5645-5657. [doi: 10.1016/j.eswa.2015.02.055]
44. About the scoring. VaderSentiment. 2021. URL: https://vadersentiment.readthedocs.io/en/latest/pages/about_the_scoring.html [accessed 2024-04-29]
45. Bathina KC, Ten Thij M, Valdez D, Rutter LA, Bollen J. Declining well-being during the COVID-19 pandemic reveals US social inequities. PLoS One 2021 Jul 8;16(7):e0254114 [FREE Full text] [doi: 10.1371/journal.pone.0254114] [Medline: 34237087]
46. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020 Dec 14;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
47. Gibbons J, Malouf R, Spitzberg B, Martinez L, Appleyard B, Thompson C, et al. Twitter-based measures of neighborhood sentiment as predictors of residential population health. PLoS One 2019 Jul 11;14(7):e0219550 [FREE Full text] [doi: 10.1371/journal.pone.0219550] [Medline: 31295294]
48. Valdez D, Patterson MS. Computational analyses identify addiction help-seeking behaviors on the social networking website Reddit: insights into online social interactions and addiction support communities. PLOS Digit Health 2022 Nov 9;1(11):e0000143 [FREE Full text] [doi: 10.1371/journal.pdig.0000143] [Medline: 36812569]
49. Rathje S, Van Bavel JJ, van der Linden S. Out-group animosity drives engagement on social media. Proc Natl Acad Sci USA 2021 Jun 29;118(26):e2024292118 [FREE Full text] [doi: 10.1073/pnas.2024292118] [Medline: 34162706]
50. Berger J. Arousal increases social transmission of information. Psychol Sci 2011 Jul;22(7):891-893 [FREE Full text] [doi: 10.1177/0956797611413294] [Medline: 21690315]
51. Sekimoto K, Seki Y, Yoshida M, Umemura K. The metrics of keywords to understand the difference between retweet and like in each category. In: Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. 2020 Presented at: WI-IAT '20; December 14-20, 2020; Melbourne, Australia p. 560-567 URL: https://www.computer.org/csdl/proceedings-article/wi-iat/2020/192400a560/1uHhvviLp6w [doi: 10.1109/wiiat50758.2020.00084]

XSL•FO
RenderX

52. Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ. Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci USA 2017 Jul 11;114(28):7313-7318 [FREE Full text] [doi: 10.1073/pnas.1618923114] [Medline: 28652356]
53. Tsugawa S, Ohsaki H. On the relation between message sentiment and its virality on social media. Soc Netw Anal Min 2017 May 11;7(1):19. [doi: 10.1007/s13278-017-0439-0]
54. Crockett MJ. Moral outrage in the digital age. Nat Hum Behav 2017 Nov 18;1(11):769-771. [doi: 10.1038/s41562-017-0213-3] [Medline: 31024117]
55. Grootendorst M. BERTopic: c-TF-IDF. GitHub. 2024. URL: https://maartengr.github.io/BERTopic/getting_started/ctfidf/ctfidf.html [accessed 2024-04-29]
56. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc 2004 Oct;60(5):503-520. [doi: 10.1108/00220410410560582]
57. Olson DL, Delen D. Performance evaluation for predictive modeling. In: Olson DL, Delen D, editors. Advanced Data Mining Techniques. Cham, Switzerland: Springer; 2008:137-147.
58. Tarekegn A, Ricceri F, Costa G, Ferracin E, Giacobini M. Predictive modeling for frailty conditions in elderly people: machine learning approaches. JMIR Med Inform 2020 Jun 04;8(6):e16678 [FREE Full text] [doi: 10.2196/16678] [Medline: 32442149]
59. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge, MA: Cambridge University Press; 2008.
60. Shi HY, Lee K, Lee H, Ho W, Sun D, Wang J, et al. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. PLoS One 2012 Apr 26;7(4):e35781 [FREE Full text] [doi: 10.1371/journal.pone.0035781] [Medline: 22563399]
61. Su L, Liu C, Li D, He J, Zheng F, Jiang H, et al. Toward optimal heparin dosing by comparing multiple machine learning methods: retrospective study. JMIR Med Inform 2020 Jun 22;8(6):e17648 [FREE Full text] [doi: 10.2196/17648] [Medline: 32568089]
62. Forrest M. 43 things to know about Justin Trudeau's new rival. Politico. 2022 Sep 10. URL: https://www.politico.com/news/2022/09/09/conservative-pierre-poilievre-in-his-own-words-00055343 [accessed 2024-04-29]
63. FDA announces shortage of Adderall. U.S. Food and Drug Administration. 2023. URL: https://www.fda.gov/drugs/drug-safety-and-availability/fda-announces-shortage-adderall [accessed 2024-04-29]
64. Adrover C, Bodnar T, Huang Z, Telenti A, Salathé M. Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter. JMIR Public Health Surveill 2015 Jul 27;1(2):e7 [FREE Full text] [doi: 10.2196/publichealth.4488] [Medline: 27227141]
65. Nasralah T, El-Gayar O, Wang Y. Social media text mining framework for drug abuse: development and validation study with an opioid crisis case analysis. J Med Internet Res 2020 Aug 13;22(8):e18350 [FREE Full text] [doi: 10.2196/18350] [Medline: 32788147]
66. Glowacki EM, Glowacki JB, Wilcox GB. A text-mining analysis of the public's reactions to the opioid crisis. Subst Abus 2018 Apr 01;39(2):129-133. [doi: 10.1080/08897077.2017.1356795] [Medline: 28723265]
67. Rajesh K, Wilcox G, Ring D, Mackert M. Reactions to the opioid epidemic: a text-mining analysis of tweets. J Addict Dis 2021 Oct 26;39(2):183-188. [doi: 10.1080/10550887.2020.1834812] [Medline: 33103608]
68. Tassone J, Yan P, Simpson M, Mendhe C, Mago V, Choudhury S. Utilizing deep learning and graph mining to identify drug use on Twitter data. BMC Med Inform Decis Mak 2020 Dec 30;20(Suppl 11):304 [FREE Full text] [doi: 10.1186/s12911-020-01335-3] [Medline: 33380324]
69. Black JC, Margolin ZR, Olson RA, Dart RC. Online conversation monitoring to understand the opioid epidemic: epidemiological surveillance study. JMIR Public Health Surveill 2020 Jun 29;6(2):e17073 [FREE Full text] [doi: 10.2196/17073] [Medline: 32597786]
70. Dekeseredy P, Sedney CL, Razzaq B, Haggerty T, Brownstein HH. Tweeting stigma: an exploration of Twitter discourse regarding medications used for both opioid use disorder and chronic pain. J Drug Issues 2021 Jan 12;51(2):340-357. [doi: 10.1177/0022042620986508]
71. Al-Garadi MA, Yang Y, Guo Y, Kim S, Love JS, Perrone J, et al. Large-scale social media analysis reveals emotions associated with nonmedical prescription drug use. Health Data Sci 2022 Apr 27;2022:9851989 [FREE Full text] [doi: 10.34133/2022/9851989] [Medline: 37621877]

Abbreviations

API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
LDA: latent Dirichlet allocation
NLP: natural language processing
TF-IDF: term-frequency–inverse document frequency
VADER: Valence Aware Dictionary and Sentiment Reasoner

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by S Matsuda, A Wani; comments to author 18.05.24; revised version received 12.06.24; accepted 01.07.24; published 23.08.24.

Please cite as:
Rao VK, Valdez D, Muralidharan R, Agley J, Eddens KS, Dendukuri A, Panth V, Parker MA
Digital Epidemiology of Prescription Drug References on X (Formerly Twitter): Neural Network Topic Modeling and Sentiment Analysis
J Med Internet Res 2024;26:e57885
URL: https://www.jmir.org/2024/1/e57885
doi:10.2196/57885
PMID:39178036

©Varun K Rao, Danny Valdez, Rasika Muralidharan, Jon Agley, Kate S Eddens, Aravind Dendukuri, Vandana Panth, Maria A Parker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessment of the Quality and Readability of Web-Based Arabic Health Information on Halitosis: Infodemiological Study

Khalid Aboalshamat[1], BDS, MPH, PhD

Preventative Dentistry Department, College of Dental Medicine, Umm Al-Qura University, Makkah, Saudi Arabia

Corresponding Author:
Khalid Aboalshamat, BDS, MPH, PhD
Preventative Dentistry Department
College of Dental Medicine
Umm Al-Qura University
Taif Road
Makkah, 21955
Saudi Arabia
Phone: 966 125270000
Email: ktaboalshamat@uqu.edu.sa

Abstract

Background: Halitosis, characterized by an undesirable mouth odor, represents a common concern.

Objective: This study aims to assess the quality and readability of web-based Arabic health information on halitosis as the internet is becoming a prominent global source of medical information.

Methods: A total of 300 Arabic websites were retrieved from Google using 3 commonly used phrases for halitosis in Arabic. The quality of the websites was assessed using benchmark criteria established by the *Journal of the American Medical Association*, the DISCERN tool, and the presence of the Health on the Net Foundation Code of Conduct (HONcode). The assessment of readability (Flesch Reading Ease [FRE], Simple Measure of Gobbledygook, and Flesch-Kincaid Grade Level [FKGL]) was conducted using web-based readability indexes.

Results: A total of 127 websites were examined. Regarding quality assessment, 87.4% (n=111) of websites failed to fulfill any *Journal of the American Medical Association* requirements, highlighting a lack of authorship (authors' contributions), attribution (references), disclosure (sponsorship), and currency (publication date). The DISCERN tool had a mean score of 34.55 (SD 7.46), with the majority (n=72, 56.6%) rated as moderate quality, 43.3% (n=55) as having a low score, and none receiving a high DISCERN score, indicating a general inadequacy in providing quality health information to make decisions and treatment choices. No website had HONcode certification, emphasizing the concern over the credibility and trustworthiness of these resources. Regarding readability assessment, Arabic halitosis websites had high readability scores, with 90.5% (n=115) receiving an FRE score ≥80, 98.4% (n=125) receiving a Simple Measure of Gobbledygook score <7, and 67.7% (n=86) receiving an FKGL score <7. There were significant correlations between the DISCERN scores and the quantity of words ($P<.001$) and sentences ($P<.001$) on the websites. Additionally, there was a significant relationship ($P<.001$) between the number of sentences and FKGL and FRE scores.

Conclusions: While readability was found to be very good, indicating that the information is accessible to the public, the quality of Arabic halitosis websites was poor, reflecting a significant gap in providing reliable and comprehensive health information. This highlights the need for improving the availability of high-quality materials to ensure Arabic-speaking populations have access to reliable information about halitosis and its treatment options, tying quality and availability together as critical for effective health communication.

(J Med Internet Res 2024;26:e54072) doi:10.2196/54072

KEYWORDS

halitosis; bad breath; malodor, Arabic web-based; infodemiological study; oral malodor; readability; infodemiology; health information; Arabic mouth medical information; reliable information; odor treatment

XSL•FO
RenderX

Introduction

Halitosis, bad breath or oral malodor, is a condition characterized by an undesirable mouth odor [1,2]. The primary factors contributing to 80%-90% of halitosis cases are intraoral in nature, which include inadequate oral hygiene practices, a coated tongue, and periodontal conditions [3]. Halitosis is a common problem worldwide, with a prevalence of 66% in China [4], 53.5% in northern Italy [5], and 32.5% in Switzerland [6]. Also, studies in Arab countries found a prevalence ranging from 6.2% to 68.5% in Saudi Arabia [7-9] and 78% in Jordan [10]. Halitosis is called a "social life killer" due to its effect on relationship breakups, social awkwardness, rejection of job and career opportunities, and even business problems [11]. Halitosis has been found to lower the quality of life, affecting marriage, intimate relationships, friendships, and daily social interactions [11]. These consequences can severely impact an individual's social life. Beyond the immediate social impact, persistent halitosis can signal underlying health issues, including oral infections and systemic conditions such as Parkinson disease, allergic rhinitis, *Helicobacter pylori* gastric infection, viral hepatitis B [3], and even COVID-19 [9].

There are different professional methods to deal with halitosis such as dental treatment, mouthwash, and toothpaste to eliminate the source of bacterial problems [12]. Probiotics have been found effective in managing halitosis despite the need for further clinical evidence [13] underscoring the importance of continuing research and clinical trials to validate their efficacy. Similarly, tongue scraping has been found to provide only short-term reduction in volatile sulfur compounds and halitosis [14]. Nevertheless, many patients tend to address halitosis with home remedies such as honey and anise [9], and other herbal remedies that are popular due to their low cost and minimal side effects [12]. People usually use the internet as a source of information to find home remedies [15].

In recent times, accessing medical information through the internet has become a prominent global trend [16-18]. Notably, a substantial percentage of patients (ranging from 45% to 85%) bring information sourced from their web-based searches when they consult with health care providers [19]. This inclination toward web-based health information-seeking has been further amplified by the recent COVID-19 pandemic, compelling individuals to turn to the internet to address their queries [20-22]. However, a critical issue that looms large is the proliferation of misinformation on the internet, a concern documented as a worldwide phenomenon [23,24]. This poses a significant threat to individuals' quality of life and even mortality [23], necessitating the implementation of stringent regulations, heightened public awareness, and enhanced dissemination of reliable health-related information [24].

As a result, numerous papers evaluating the quality and readability of web-based information on oral health conditions and diseases have been published in English. Examples of these papers include studies on halitosis [25], medication-related osteonecrosis of the jaw [26], postendodontic treatment selection [27], and treatment of the mouth in systemic sclerosis [28]. Analogous research was carried out in other languages, including Danish [29] and Spanish [30], suggesting the significance of additional languages. However, there is limited research available on oral health topics in Arabic, assessing general content to the public. A few studies have investigated oral cancer [31], dental implants [32], periodontal disease [33], and denture hygiene information [34]. The majority of these studies used comparable approaches, including the readability calculator tool, the DISCERN tool [35] to assess website quality, and the existence of the Health on the Net Foundation Code of Conduct (HONcode) [36]. Even though halitosis is considered the third most common reason for seeking dental treatment [11], no research has been done to evaluate the quality and readability of halitosis information in Arabic. In the literature, there is a single study that assesses the quality and readability of halitosis but only in English [25], which highlights the gap in knowledge. Additionally, there are 22 nations with Arabic as their official language, and more than 422 million people speak it [37]. Thus, this study aims to assess the quality and readability of web-based Arabic health information on halitosis. The alternative hypothesis for this study is that there is a significant association between the quality and readability of web-based Arabic health information on halitosis.

Methods

Search Strategy

This research falls under the domain of infodemiological investigation (study of web-based information's impact on health). The study was conducted using Google Chrome, specifically version 114.0.5735.6. The search was done in incognito mode. This approach was used to mitigate the impact of search history and tailor the search algorithms to customize the search results [38]. The search was carried out using widely used search engines such as Google. The search was conducted on May 7, 2023. Google was used exclusively because it is the world's leading search engine.

Three synonymous terms for halitosis, both in formal and informal Arabic, were used, which translate to "halitosis," "mouth breath," and "teeth smell" in English. These terms in Arabic are " ," " ," and " ." A total of 300 web pages were obtained by retrieving the first 100 websites for each phrase. All instances of duplicated websites were removed. The previous points outline the inclusion criteria for the website; however, the selection of websites for this research was determined based on the following exclusion criteria: sources in languages other than Arabic; dictionaries; Arabic books published before the 18th century; proprietary commercial merchandise available for purchase on e-commerce platforms such as Amazon or Noon platforms; social forums and social media websites; academic scientific papers or textbooks; websites that are prohibited or have restricted access, necessitating the use of identification and a password; sources with no information about halitosis or information presented only in hints; religious websites that ask for religious clarification "Fatwa" regarding halitosis; and content delivered exclusively through video, audio, or PowerPoint presentations. The researchers applied these criteria to ensure that only relevant

XSL•FO
RenderX

and trustworthy material was included in the study. This selection process is delineated in Figure 1.

Each website was categorized in terms of specialization, affiliation, and content type [39]. Specialization varied from being partial or exclusive to the topic. The affiliation had 5 different classes: nonprofit organization, university or medical center, government, commercial, and journalism. In terms of content types, the website could provide clinical trials, medical facts, human stories, or questions and answers. Additionally, each website was noted to include video, audio, and images on each website was documented.

Figure 1. Flowchart illustrating the hierarchical structure and interconnectedness of the websites discovered through the search method used.

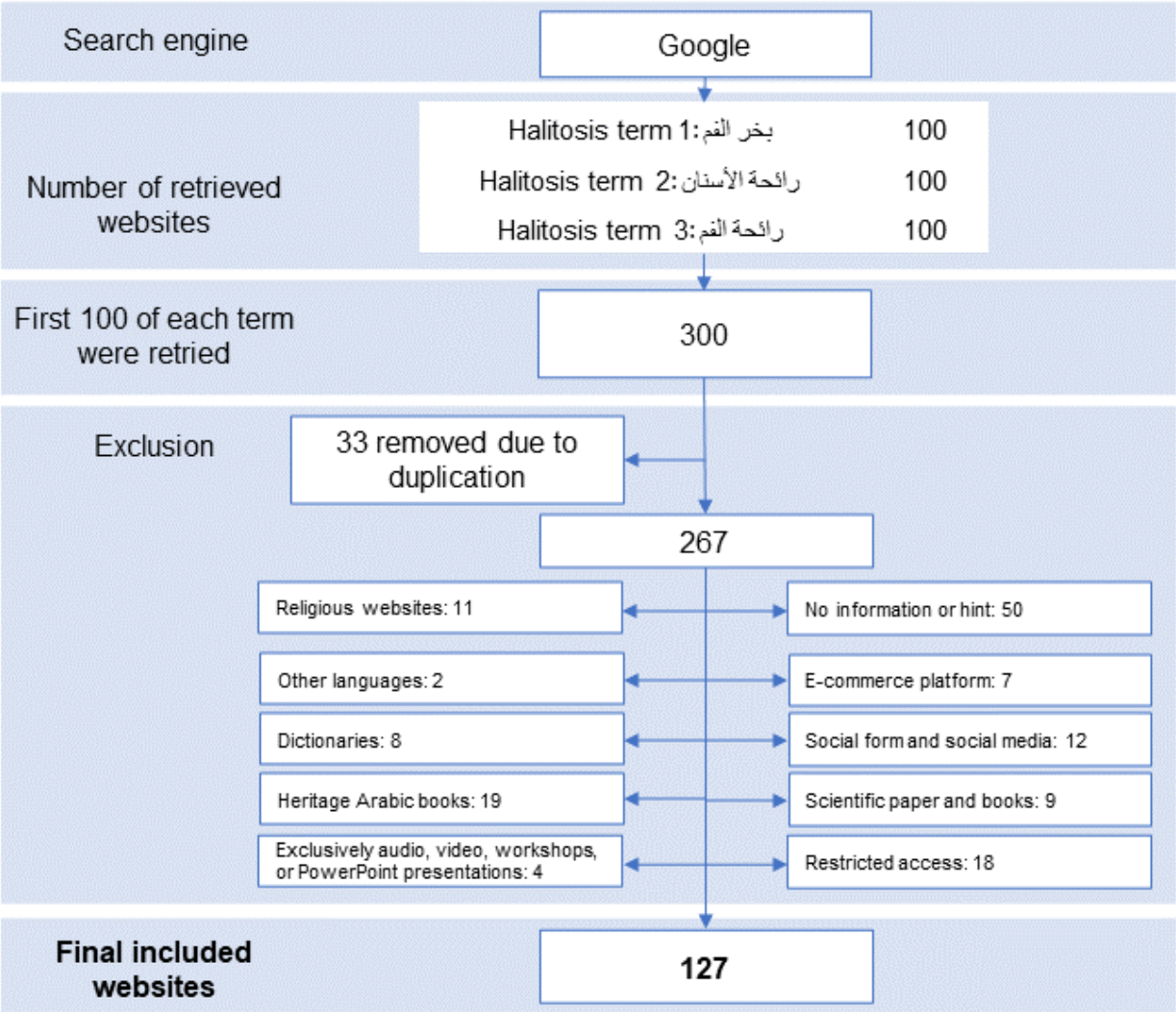

Quality Assessment

To evaluate website quality, a comprehensive evaluation was conducted by 2 assessors, denoted as RA and AA, both of whom are qualified professionals as dentists. They used the DISCERN [35] and *Journal of the American Medical Association* (*JAMA*) tools [40] in this assessment. The calibration process consisted of 2 parts. In the first stage, each assessor independently evaluated 10 websites. Any discrepancies in their assessments were subsequently addressed through discussions with the lead investigator. A further evaluation of 20 websites was carried out, and any disagreements were resolved through collaborative efforts with a discussion with the lead investigator. This was to ensure calibration. For the following website, any disagreements were verified by the lead investigator as a final decision. The DISCERN instrument comprises a set of 16 questions categorized into 3 main sections. The first section, questions 1-8, aims to assess the credibility of websites as sources of information for various medical topics. The next section, including questions 9-15, focuses on various aspects of therapeutic options. Question 16 serves as an overall evaluation criterion for quality. The theoretical score range of the DISCERN instrument spans from 16 to 80. Websites scoring between 16 and 32 are classified as having poor quality, those scoring between 33 and 64 are considered to have moderate quality, and websites that score 65 or greater are classified as high quality.

The assessment of each website was conducted using the *JAMA* criteria for website evaluation [40], the presence of the HONcode [36], and the DISCERN tool [35]. The *JAMA*

XSL•FO
RenderX

benchmarks encompass four primary criteria that must be met: (1) authorship, which involves the inclusion of authors' contributions, affiliations, and their relevant credentials; (2) attribution, which entails providing references and citations; (3) disclosure, which requires indicating any sponsorship, ownership, commercial funding, advertising, or potential conflicts of interest; and (4) currency, which necessitates the inclusion of the publication date and information on updates.

Each website received a score of 1 point if it met one of the criteria; otherwise, it received a score of 0. The *JAMA* benchmark score was calculated as the cumulative total of the preceding categories, which ranged from 0 to 4 points, with 4 being the maximum score attainable. The HONcode tool allows websites to obtain authorization to display a stamp, similar to a HON award badge, on their web page, contingent upon their adherence to the HONcode standards, which is valid for a period of 1 year.

Readability Assessment

The notion of readability refers to the systematic use of formulas to determine the level of reading comprehension required for an individual to understand written materials [41]. The readability of the content was evaluated using a freely available readability calculator [42], which is commonly used to measure the readability of English text. This calculator was the standard method of measurement by literature for similar studies to assess dental English websites [25,27,28] and Arabic dental websites [31-34]. This calculator uses 3 indices: the Flesch-Kincaid Grade Level (FKGL), the Simple Measure of Gobbledygook (SMOG), and the Flesch Reading Ease (FRE). The FKGL uses metrics such as average sentence length and average syllables per word to determine the level of reading complexity. In contrast, the SMOG Index calculates the ratio of words that have 3 or more syllables. As the score increases, the reading difficulty also increases. According to the previous literature [43,44], the FKGL and SMOG readability scores should ideally be around 7. The FRE algorithm computes a numerical score between 0 and 100 by analyzing the mean sentence length and the mean syllable count per word. A score of 80 or greater on the FRE scale suggests content that is easily comprehensible [43,44].

Statistical Analysis

The process of data entry and cleansing was conducted using the Microsoft Excel program. The data analysis was carried out using SPSS version 29 (IBM Corp). Descriptive statistics were used to analyze the data collected from websites, using measures such as count, percentages, mean, SD, median, minimum, maximum, and IQR. The Spearman correlation coefficient was used to evaluate the association between *JAMA*, DISCERN scores, and readability indices. This is because all the correlation tests conducted violated the normality assumption of the dependent variables (number of words, number of sentences, FKGL, SMOG, and FRE). The *P* values of the Shapiro-Wilk tests for the aforementioned variables were $P<.001$. A predetermined $P=.05$ was used as the threshold for determining statistical significance. Raw data are available in Multimedia Appendix 1.

Ethical Considerations

Given that this is an infodemiological study and does not involve data collection from humans or animals, no institutional review board approval or ethical application was required. Consequently, no submission was made to the ethics committee at Umm Al-Qura University, Saudi Arabia.

Results

Arabic Halitosis Websites

Of the 300 websites about halitosis, only 267 websites remained after removing the duplication (n=33, 11%), as seen in Figure 1. The excluded websites encompassed a diverse range of categories, including those in other languages (n=2, 0.6%), dictionaries (n=8, 2.6%), heritage Arabic books (n=19, 6.3%), e-commerce platforms (n=7, 2.3%), social forums and social media websites (n=12, 4%), academic scientific papers or textbooks focused on halitosis (n=9, 3%), websites with restricted access (n=18, 6%), as well as those offering either no information or mere hints (n=50, 16.6%). Additionally, a small number solely used video, audio, or Microsoft PowerPoint presentations (n=4, 1.3%), and some websites focused on religious inquiry "Fatwa" regarding halitosis (n=11, 3.6%). The classification of the Arabic halitosis websites in terms of content type, affiliation, specialization, and inclusion of image, video, and audio content is displayed in Table 1. Most of the Arabic websites on halitosis predominantly feature medical facts, with 126 (99.2%) sites focusing exclusively on this content type. Almost all sites (n=125, 98.4%) are exclusively related to halitosis. In terms of affiliation, journalism and commercial entities are the leading sources, comprising 41.7% (n=53) and 37.8% (n=48) of the websites, respectively. The inclusion of multimedia content is notable, with 81.1% (n=103) of the sites containing images, but videos (n=10, 7.8%) and audio (n=4, 3.15%) are less commonly featured.

XSL•FO
RenderX

Table 1. Classification of the Arabic halitosis websites in terms of content type, affiliation, specialization, and inclusion of image, video, and audio content.

Item	Websites, n (%)
Content type	
Medical facts	126 (99.2)
Human interest stories	1 (0.79)
Question and answers	0 (0)
Clinical trials	0 (0)
Specialization	
Exclusively related to halitosis	125 (98.4)
Partly related to halitosis	2 (1.5)
Affiliation	
Journalism	53 (41.7)
Commercial	48 (37.8)
University or medical center	19 (14.9)
Government	4 (3.1)
Nonprofit organization	3 (2.3)
Contain image	
Yes	103 (81.1)
Contain video	
Yes	10 (7.8)
Contain audio	
Yes	4 (3.1)

Assessment of the Quality of Arabic Halitosis Websites

When evaluating the *JAMA* benchmark criteria, it was found that only 2.3% (n=3) met the requirements for authorship, 3.1% met the criteria for attribution, 0.7% (n=1) met the standards for disclosure, and 9.4% (n=12) met the criteria for currency, as indicated in Table 2. The majority of websites (n=111, 87.4%) did not meet any of the criteria, 9.4% (n=12) met 1 criterion, and 3.1% (n=4) met 2 criteria. No website reached a score of 3 or the maximum of 4 based on the *JAMA* benchmark criteria. The mean for the *JAMA* benchmark criteria was 0.15 (SD 0.44), and the median was 0 (IQR 0-0). All the websites lacked HONcode certification.

According to the DISCERN score classification, none of the websites achieved a high-quality score (ie, 65 or greater). A total of 72 (56.6%) websites received a moderate quality score (between 33 and 64), while 55 (43.3%) obtained a low-quality score (between 16 and 32), as indicated by the scores for each item shown in Table 3. The mean for the total DISCERN score was 34.55 (SD 7.46). The criteria best addressed were relevance (Q3: 4.91, SD 0.43) and clarity of treatment choices (Q14: 4.01, SD 1.55), whereas the least addressed were treatment impact on quality of life (Q13: 1.02, SD 0.18) and addressing uncertainties (Q8: 1.03, SD 0.35).

XSL•FO
RenderX

Table 2. Assessment of the quality of Arabic halitosis websites using JAMA[a] benchmark criteria.

JAMA criteria	Website, n (%)
Authorship	3 (2.3)
Attribution	4 (3.1)
Disclosure	1 (0.7)
Currency	12 (9.4)
0 *JAMA* criteria	111 (87.4)
1 *JAMA* criterion	12 (9.4)
2 *JAMA* criteria	4 (3.1)
3 or 4 *JAMA* criteria	0 (0)

[a]JAMA: Journal of the American Medical Association.

Table 3. DISCERN assessment of Arabic halitosis websites.

Domain and question	Score, mean (SD)
Reliability	
Q1. Are the aims clear?	1.43 (0.86)
Q2. Does it achieve its aims?	1.72 (1.48)
Q3. Is it relevant?	4.91 (0.43)
Q4. Is it clear what sources of information were used to compile the publication?	1.83 (1.17)
Q5. Is it clear when the information used or reported in the publication was produced?	2.51 (1)
Q6. Is it balanced and unbiased?	3.68 (0.95)
Q7. Does it provide details of additional sources of support and information?	2.29 (1.7)
Q8. Does it refer to areas of uncertainty?	1.03 (0.35)
Treatment options	
Q9. Does it describe how each treatment works?	3.00 (1.51)
Q10. Does it describe the benefits of each treatment?	2.06 (1.15)
Q11. Does it describe the risks of each treatment?	1.28 (0.7)
Q12. Does it describe what would happen if no treatment is used?	1.16 (0.69)
Q13. Does it describe how the treatment choices affect the overall quality of life?	1.02 (0.18)
Q14. Is it clear that there may be more than 1 possible treatment choice?	4.01 (1.55)
Q15. Does it provide support for shared decision-making?	1.35 (0.95)
Q16. Overall quality of the publication as a source of information about Halitosis.	1.28 (0.83)

Readability of Arabic Halitosis Websites

Using the readability calculator, the median, IQR, minimum, and maximum were recorded for different items as shown in Table 4. The readability assessment showed a median of 3164 (IQR 2024-4618) characters (without spaces) and 669 (IQR 412-1025) words per paper, with content typically comprising 30 sentences. The readability assessments revealed that the majority of the materials (n=86, 67.7%) have an FKGL below 7, with 41 materials (32.2%) scoring 7 or more, indicating easier comprehension. Meanwhile, the SMOG and FRE scores suggest a high readability level, with 98.4% (n=125) of materials categorized below a SMOG score of 7 and 90.5% (n=115) having an FRE score of 80 or higher. There were only 1.5% (n=2) of the material categorized as equal or greater SMOG score of 7, and 9.4% (n=12) having an FRE score below 80.

Spearman correlation coefficients were computed to assess the relationships between the DISCERN score and various readability indices, including the number of words, number of sentences, FKGL, SMOG, and FRE, as presented in Table 5.

XSL•FO
RenderX

Table 4. Readability of Arabic halitosis websites.

Item	Score	
	Median (IQR)	Minimum-maximum
Number of characters (without spaces)	3164 (2024-4618)	416-32,898
Number of words	669 (412-1025)	86-6929
Number of sentences	30 (15-40)	2-325
Average number of characters per word	4.75 (4.6-4.8)	4.08-5.19
Average number of syllables per word	1 (1-1.01)	1-1.1
Average number of words per sentence	22.91 (19.1-29.6)	12.12-611.5
FKGL[a]	5.16 (3.8-7.7)	0.96-234.71
SMOG[b]	3.71 (3-4.3)	3-7.74
FRE[c]	98.64 (92-101.9)	8.05-498.58

[a]FKGL: Flesch-Kincaid Grade Level.
[b]SMOG: Simple Measure of Gobbledygook.
[c]FRE: Flesch Reading Ease.

Table 5. Assessing the relationship between halitosis websites' DISCERN score and readability indices using Spearman correlation.

	DISCERN	Number of words	Number of sentences	FKGL[a]	SMOG[b]	FRE[c]
DISCERN						
Spearman ρ	1	0.433	0.433	–0.123	0.176	0.092
P value	—[d]	<.001	<.001	.17	.047	.30
Number of words						
Spearman ρ	0.433	1	0.773	0.043	0.146	0.048
P value	<.001	—	<.001	.63	.10	.59
Number of sentences						
Spearman ρ	0.433	0.773	1	–0.482	0.063	0.402
P value	<.001	<.001	—	<.001	.49	<.001
FKGL						
Spearman ρ	–0.123	0.043	–0.482	1	0.151	–0.895
P value	.17	.63	<.001	—	.09	<.001
SMOG						
Spearman ρ	0.176	0.146	0.063	0.151	1	–0.208
P value	.047	.10	.49	.09	—	.02
FRE						
Spearman ρ	0.092	0.048	0.402	–0.895	–0.208	1
P value	.30	.59	<.001	<.001	.02	—

[a]FKGL: Flesch-Kincaid Grade Level.
[b]SMOG: Simple Measure of Gobbledygook.
[c]FRE: Flesch Reading Ease.
[d]Not applicable.

XSL•FO
RenderX

Discussion

Overview

While halitosis is one of the most common problems in dental care [11], the internet is considered to be a main source of information to the public [45] despite the quality and overwhelming quantity of information. This study aimed to assess the quality and readability of web-based Arabic health information on halitosis. Despite the high number of sites regarding halitosis in Arabic, our results revealed that the Arabic halitosis websites are highly readable but poor in quality from a professional and academic point of view.

Quality Assessment of Arabic Halitosis Websites

DISCERN score was approximately similar to a previous study about halitosis [25] and other previous Arabic infodemiological studies that investigated denture hygiene [34], dental implants [32], and periodontal diseases [33]. There were small noncrucial differences, and the scores ranged between the low and moderate categories of DISCERN. However, our DISCERN result was lower than the other infodemiological studies that assessed the quality and readability of third molar tooth websites in English [46]. In this study, not even a single website scored high according to the DISCERN cutoff. On the other hand, the *JAMA* score in this study was found to be lower than all the previous studies about English halitosis websites [25] or other Arabic dental infodemiological studies [32-34]. It should be noted that the previous study about English halitosis websites fulfilled the criteria of authorship, attribution, disclosure, and currency by 51.8%, 35.6%, 29%, and 48.5%, which is far higher than the percentages of such items in our studies. This leads to the conclusion that the quality of available material on Arabic websites on dental topics is generally not satisfactory, and the quality of websites about halitosis is much poorer in Arabic language than in English although both are unsatisfactory. This absence of high-quality Arabic websites shows the need for the development of comprehensive and reliable web-based dental resources. It is possible to use DISCERN items as guidelines to develop better Arabic content about halitosis. This is especially true for the items that scored low in the DISCERN assessment, which deals with uncertainty and on how treatment affects the quality of life. The English websites with higher scores in dental topics are not an odd finding. This is because DISCERN is available mainly for English websites. However, collaborative initiatives involving English content creators and dental associations in Arabic-speaking regions can be established to address these gaps.

There was no single website in this study that had the HONcode certification, which is somewhat similar to other infodemiological studies about dental websites in Arabic, which reported less than 2 websites having HONcode [32-34]. On the other hand, 14.9% of websites about halitosis in English had HONcode [25], with similar percentages in other infodemiological studies about dental issues in English [47-49]. One of the explanations to this is that HONcode might be familiar among websites in English, and that granting the certification might be more convenient to English websites rather than those in Arabic. In fact, the HONcode instruction is given in English only. It is recommended that HONcode certification be introduced to medical and dental website content creators or that a similar certificate be made available that is assessed in Arabic.

Readability Assessment of Arabic Halitosis Websites

With regard to readability, this study indicated that the majority of websites were readable for those possessing language skills beyond the sixth grade level. This aligned with other studies that assessed previous Arabic dental websites [32-34]. This might not be the case in the previous study of halitosis in English [25], and other dental topics in English also [28] found that the level of reading is more challenging. These results did not find a relation between FKGL, SMOG, and FRE with DISCERN scores. This refutes a proposed explanation that the websites were created simply and the quality of the websites was reduced. However, there are 2 possible explanations for this. First, the Arabic content creator wrote in simple language to reach a public audience compared to the English websites. Second, the language calculator was designed for the English language in the United States and might not be very accurate when used for other languages, despite being used by previous studies [32-34]. Such aspects need to be verified by future studies and include professional assessment using the common method to assess the readability of Arabic content.

Contents of Arabic Halitosis Websites

As the previous aspects investigated Arabic halitosis websites' contents and readability, the assessment did not include the verification of the contents' validity as done in other studies [50,51]. Merely observing the presence or absence of DISCERN quality items might not sufficiently judge the validity and robustness of the information provided. Referencing evidence-based resources about halitosis [52,53] can offer a better assessment of new information related to halitosis. Furthermore, leveraging artificial intelligence could be a promising future direction as a tool to provide deeper insights into the content of such websites [54]. Artificial intelligence and natural language processing technologies can aggregate and process content from specified websites, enabling researchers to categorize the information into accurate, uncertain, and incorrect categories. Following this classification, a detailed analysis of the website content can be conducted. However, implementing such a methodology necessitates a comprehensive and nuanced strategy.

One of the notable aspects of this study is that there are a considerable number of websites (n=19, 6.3%) displaying heritage Arabic books discussing halitosis that were published centuries ago. For example, there was a book that was published in 776 Hijri, that is, AD 1398, and another book that was published in AD 1505. Also, there were 11 websites that had religious clarifications (Fatwa) regarding halitosis. This can be understood, as Islamic teachings have some specific teachings regarding halitosis; for instance, Muslims are encouraged to use *Salvadora persica* (siwak) for good oral breath [55]. Also, people who have halitosis are discouraged from praying in a mosque. In fact, halitosis is a serious issue not only in Islam; for example, according to the Talmud, a holy book for Jews, halitosis can also be a cause of divorce [56]. Such aspects might

XSL•FO
RenderX

be important as a social-cultural dimension when crafting website content suitable for the audience.

In this study, social media results were excluded from the investigated websites. Social media often uses short text and might depend on continuous dialogue that does not belong to a single entity, which makes it very difficult to assess using the existing tools. However, further assessments can be conducted on social media using appropriate assessment tools.

This study might be one of the few studies that assessed the quality and readability of dental problems in Arabic, using a well-known validated tool. One of the limitations of this study is that the validity and the contents of the Arabic halitosis websites were not assessed in comparison to scientific evidence-based resources. It is not known according to the assessment if the current website provides sound or misleading information. The importance of such an aspect is accentuated in this era of massive health misinformation [23,24]. Future studies should be directed to conduct content analysis and verify the trustworthiness of the information on these websites. Also, there are studies that introduce different approaches to deal with halitosis [3,13] such as probiotics and others that should be included in the websites for the public in order to provide thorough content. Also, the website search was conducted over a desktop computer, which might be different if we conducted the search over a smartphone [57], especially in this new era where people search for information on mobile devices from anywhere and at any time.

Study Limitations

The websites retrieved in this study were taken from 1 search engine, unlike some of the previous studies. The reason might be because Google has been the main search engine for the majority of people worldwide over the last 10 years, with a large difference compared to other search engines [58]. However, this limitation should be considered in future research, and other search engines should be investigated, such as Bing and Yahoo. Finally, this study treated Arabic websites as a homogeneous group, not accounting for the geographical diversity among Arabic speakers that could influence digital health information. This diversity might result in digital content variations related to culture, region, and accessibility. Consequently, this could affect the generalizability of our findings.

Conclusions

While the readability of halitosis and other dental topic websites in Arabic was very good, their quality was poor. It is essential to make reliable information accessible for understanding halitosis and other dental topics, as well as treatment options, for Arabic-speaking populations. Collaborative efforts should be directed toward formulating such content, especially considering recent advancements in treatment modalities. Future studies should investigate the content of dental websites in Arabic to assess the validity of information using evidence-based approaches.

Acknowledgments

The author expresses gratitude to Alyaa Amer and Rawan Ammar for their assistance in the assessment of the websites for this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Raw data.
[XLSX File (Microsoft Excel File), 135 KB - jmir_v26i1e54072_app1.xlsx]

References

1. Quirynen M, Dadamio J, Van den Velde S, De Smit M, Dekeyser C, Van Tornout M, et al. Characteristics of 2000 patients who visited a halitosis clinic. J Clin Periodontol 2009;36(11):970-975. [doi: 10.1111/j.1600-051X.2009.01478.x] [Medline: 19811581]
2. Kayombo CM, Mumghamba EG. Self-reported halitosis in relation to oral hygiene practices, oral health status, general health problems, and multifactorial characteristics among workers in Ilala and Temeke municipals, Tanzania. Int J Dent 2017;2017:8682010 [FREE Full text] [doi: 10.1155/2017/8682010] [Medline: 28280509]
3. Memon MA, Memon HA, Muhammad FE, Fahad S, Siddiqui A, Lee KY, et al. Aetiology and associations of halitosis: a systematic review. Oral Dis 2023;29(4):1432-1438. [doi: 10.1111/odi.14172] [Medline: 35212093]
4. Du M, Li L, Jiang H, Zheng Y, Zhang J. Prevalence and relevant factors of halitosis in Chinese subjects: a clinical research. BMC Oral Health 2019;19(1):45 [FREE Full text] [doi: 10.1186/s12903-019-0734-4] [Medline: 30866896]
5. Aimetti M, Perotto S, Castiglione A, Ercoli E, Romano F. Prevalence estimation of halitosis and its association with oral health-related parameters in an adult population of a city in North Italy. J Clin Periodontol 2015;42(12):1105-1114. [doi: 10.1111/jcpe.12474] [Medline: 26477536]

XSL•FO
RenderX

6. Bornstein MM, Kislig K, Hoti BB, Seemann R, Lussi A. Prevalence of halitosis in the population of the city of Bern, Switzerland: a study comparing self-reported and clinical data. Eur J Oral Sci 2009;117(3):261-267. [doi: 10.1111/j.1600-0722.2009.00630.x] [Medline: 19583753]
7. AlSadhan SA. Self-perceived halitosis and related factors among adults residing in Riyadh, Saudi Arabia. A cross sectional study. Saudi Dent J 2016;28(3):118-123 [FREE Full text] [doi: 10.1016/j.sdentj.2016.06.001] [Medline: 27656078]
8. Bhat MYS, Alayyash AA. Social stigma related to halitosis in Saudi and British population: a comparative study. J Dent Res Rev 2016;3(2):65 [FREE Full text] [doi: 10.4103/2348-2915.184215]
9. Aboalshamat K, Farsi N, Merwass R, Alshubaili D, Filimban L. The impact of COVID-19 on halitosis among the general population in Saudi Arabia: a potential symptom. J Res Med Dent Sci 2021;9(11):45-52 [FREE Full text]
10. Hammad MM, Darwazeh AM, Al-Waeli H, Tarakji B, Alhadithy TT. Prevalence and awareness of halitosis in a sample of Jordanian population. J Int Soc Prev Community Dent 2014 Dec;4(Suppl 3):S178-S186 [FREE Full text] [doi: 10.4103/2231-0762.149033] [Medline: 25625076]
11. Azodo C, Osazuwa-Peters N, Omili M. Psychological and social impacts of halitosis: a review. J Soc Psychol 2010;3(1):74-92 [FREE Full text]
12. Froum SJ, Shi Y, Reis N, Asvaplungprohm T. A narrative review of the diagnosis, etiology, and treatment of halitosis over the past three decades. Compend Contin Educ Dent 2022;43(5):E5-E8. [Medline: 35523316]
13. López-Valverde N, López-Valverde A, Macedo de Sousa B, Rodríguez C, Suárez A, Aragoneses JM. Role of probiotics in halitosis of oral origin: a systematic review and meta-analysis of randomized clinical studies. Front Nutr 2021;8:787908 [FREE Full text] [doi: 10.3389/fnut.2021.787908] [Medline: 35127785]
14. Outhouse TL, Fedorowicz Z, Keenan JV, Al-Alawi R. A Cochrane systematic review finds tongue scrapers have short-term efficacy in controlling halitosis. Gen Dent 2006;54(5):352. [Medline: 17004573]
15. Parisius LM, Stock-Schröer B, Berger S, Hermann K, Joos S. Use of home remedies: a cross-sectional survey of patients in Germany. BMC Fam Pract 2014;15:116 [FREE Full text] [doi: 10.1186/1471-2296-15-116] [Medline: 24916173]
16. Fiksdal AS, Kumbamu A, Jadhav AS, Cocos C, Nelsen LA, Pathak J, et al. Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. J Med Internet Res 2014;16(10):e224 [FREE Full text] [doi: 10.2196/jmir.3341] [Medline: 25348028]
17. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]
18. AlGhamdi KM, Moussa NA. Internet use by the public to search for health-related information. Int J Med Inform 2012;81(6):363-373. [doi: 10.1016/j.ijmedinf.2011.12.004] [Medline: 22217800]
19. Murray E, Lo B, Pollack L, Donelan K, Catania J, Lee K, et al. The impact of health information on the internet on health care and the physician-patient relationship: national U.S. survey among 1.050 U.S. physicians. J Med Internet Res 2003;5(3):e17 [FREE Full text] [doi: 10.2196/jmir.5.3.e17] [Medline: 14517108]
20. An L, Russell DM, Mihalcea R, Bacon E, Huffman S, Resnicow K. Online search behavior related to COVID-19 vaccines: infodemiology study. JMIR Infodemiology 2021;1(1):e32127 [FREE Full text] [doi: 10.2196/32127] [Medline: 34841200]
21. Rajan A, Sharaf R, Brown RS, Sharaiha RZ, Lebwohl B, Mahadev S. Association of search query interest in gastrointestinal symptoms with COVID-19 diagnosis in the United States: infodemiology study. JMIR Public Health Surveill 2020;6(3):e19354 [FREE Full text] [doi: 10.2196/19354] [Medline: 32640418]
22. Bento AI, Nguyen T, Wing C, Lozano-Rojas F, Ahn Y, Simon K. Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. Proc Natl Acad Sci U S A 2020;117(21):11220-11222 [FREE Full text] [doi: 10.1073/pnas.2005335117] [Medline: 32366658]
23. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
24. Borges do Nascimento IJ, Pizarro AB, Almeida JM, Azzopardi-Muscat N, Gonçalves MA, Björklund M, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022;100(9):544-561 [FREE Full text] [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
25. Jo JH, Kim EJ, Kim JR, Kim MJ, Chung JW, Park JW. Quality and readability of internet-based information on halitosis. Oral Surg Oral Med Oral Pathol Oral Radiol 2018;125(3):215-222. [doi: 10.1016/j.oooo.2017.12.001] [Medline: 29325852]
26. Meade MJ. Medication-related osteonecrosis of the jaw: a cross-sectional survey assessing the quality of information on the internet. Oral Surg Oral Med Oral Pathol Oral Radiol 2022;133(4):e83-e90. [doi: 10.1016/j.oooo.2021.08.029] [Medline: 34753689]
27. Erdinç G, Özdemir O. Readability and quality assessment of web-based information concerning post-endodontic treatment selection. Eur J Ther 2023;29(3):534-541 [FREE Full text] [doi: 10.58600/eurjther1763]
28. Abdouh I, Porter S, Fedele S, Elgendy N, Ni Riordain R. Web-based information on the treatment of the mouth in systemic sclerosis. BMC Rheumatol 2020;4(1):61 [FREE Full text] [doi: 10.1186/s41927-020-00160-5] [Medline: 33292857]
29. Ferlias N, Ostapiuk MA, Diekema TN, Kristensen MG, Kristensen KD, Stoustrup P. Online information on orthodontic treatment in Denmark: a population-based quality assessment. AJO-DO Clin Companion 2021;1(2):119-126. [doi: 10.1016/j.xaor.2021.05.001]

XSL•FO
RenderX

30. Ayala Aguirre PE, Aníbal I, Strieder A, Lotto M, Lopes Rizzato V, Pereira Cruvinel AF, et al. Online quality and readability assessment of early childhood caries information available on websites from distinct countries: a cross-sectional study. Eur J Paediatr Dent 2022;23(1):15-20 [FREE Full text] [doi: 10.23804/ejpd.2022.23.01.03] [Medline: 35274537]
31. Alakhali MS. Quality assessment of information on oral cancer provided at Arabic speaking websites. Asian Pac J Cancer Prev 2020;21(4):961-966 [FREE Full text] [doi: 10.31557/APJCP.2020.21.4.961] [Medline: 32334456]
32. Halboub E, Al-Ak'hali MS, Alqahtani AS, Abdulghani EA, Kamil MA, Alhajj MN, et al. Quality of web-based Arabic health information on dental implants: an infodemiological study. BMC Oral Health 2023;23(1):232 [FREE Full text] [doi: 10.1186/s12903-023-02938-8] [Medline: 37081436]
33. Al-Ak'hali MS, Fageeh HN, Halboub E, Alhajj MN, Ariffin Z. Quality and readability of web-based Arabic health information on periodontal disease. BMC Med Inform Decis Mak 2021;21(1):41 [FREE Full text] [doi: 10.1186/s12911-021-01413-0] [Medline: 33541345]
34. Alhajj MN, Mashyakhy M, Ariffin Z, Ab-Ghani Z, Johari Y, Salim NS. Quality and readability of web-based Arabic health information on denture hygiene: an infodemiology study. J Contemp Dent Pract 2020;21(9):956-960. [Medline: 33568578]
35. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
36. Boyer C, Baujard V, Geissbuhler A. Evolution of health web certification through the HONcode experience. Stud Health Technol Inform 2011;169:53-57. [Medline: 21893713]
37. Binmahfoudh A. Improved deep learning sentiment analysis for Arabic. J Theor Appl Inf Technol 2023;101(3):1251-1260 [FREE Full text]
38. Hannak A, Sapiezynski P, Molavi KA, Krishnamurthy B, Lazer D, Mislove A. Measuring personalization of web search. 2013 Presented at: 22nd International Conference on World Wide Web; May 13-17, 2013; Rio de Janeiro, Brazil URL: https://dl.acm.org/doi/10.1145/2488388.2488435 [doi: 10.1145/2488388.2488435]
39. Ní Ríordáin R, McCreary C. Dental patients' use of the internet. Br Dent J 2009;207(12):583. [doi: 10.1038/sj.bdj.2009.1137] [Medline: 20019727]
40. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the internet: caveant lector et viewor—let the reader and viewer beware. JAMA 1997;277(15):1244-1245. [Medline: 9103351]
41. Albright J, de Guzman C, Acebo P, Paiva D, Faulkner M, Swanson J. Readability of patient education materials: implications for clinical practice. Appl Nurs Res 1996;9(3):139-143. [doi: 10.1016/s0897-1897(96)80254-0] [Medline: 8771859]
42. Tests document readability: readability calculator. Online Utility. URL: https://www.online-utility.org/english/readability_test_and_improve.jsp [accessed 2023-05-10]
43. Edmunds MR, Barry RJ, Denniston AK. Readability assessment of online ophthalmic patient information. JAMA Ophthalmol 2013;131(12):1610-1616. [doi: 10.1001/jamaophthalmol.2013.5521] [Medline: 24178035]
44. Kher A, Johnson S, Griffith R. Readability assessment of online patient education material on congestive heart failure. Adv Prev Med 2017;2017:9780317. [doi: 10.1155/2017/9780317] [Medline: 28656111]
45. Shinchuk LM, Chiou P, Czarnowski V, Meleger AL. Demographics and attitudes of chronic-pain patients who seek online pain-related medical information: implications for healthcare providers. Am J Phys Med Rehabil 2010;89(2):141-146. [doi: 10.1097/PHM.0b013e3181c56938] [Medline: 19966558]
46. Hanna K, Brennan D, Sambrook P, Armfield J. Third molars on the internet: a guide for assessing information quality and readability. Interact J Med Res 2015;4(4):e19 [FREE Full text] [doi: 10.2196/ijmr.4712] [Medline: 26443470]
47. Alshehri A, Alghofaili N, ALshunaiber R, Alkadi L. Quality and readability assessment of internet-based information on common prosthodontic treatments. Int J Prosthodont 2022;35(1):62-67. [doi: 10.11607/ijp.7063] [Medline: 33651024]
48. Wiriyakijja P, Fedele S, Porter S, Ni Riordain R. Web-based information on the treatment of oral leukoplakia - quality and readability. J Oral Pathol Med 2016 Sep;45(8):617-620. [doi: 10.1111/jop.12459] [Medline: 27233229]
49. Alnafea S, Fedele S, Porter S, Ni Riordain R. Online information on the treatment of burning mouth syndrome: quality and readability. J Oral Facial Pain Headache 2017;31(2):147-151 [FREE Full text] [doi: 10.11607/ofph.1717] [Medline: 28437511]
50. Raj S, Ghosh A, Pandiyan S, Chauhan D, Goel S. Analysis of YouTube content on substance use disorder treatment and recovery. Int J Soc Psychiatry 2023;69(8):2097-2109. [doi: 10.1177/00207640231190304] [Medline: 37650472]
51. Bragazzi NL, Prasso G, Re TS, Zerbetto R, Del Puente G. A reliability and content analysis of Italian language anorexia nervosa-related websites. Risk Manag Healthc Policy 2019;12:145-151 [FREE Full text] [doi: 10.2147/RMHP.S193088] [Medline: 31534378]
52. Szalai E, Tajti P, Szabó B, Kói T, Hegyi P, Czumbel LM, et al. Organoleptic and halitometric assessments do not correlate well in intra-oral halitosis: a systematic review and meta-analysis. J Evid Based Dent Pract 2023;23(3):101862 [FREE Full text] [doi: 10.1016/j.jebdp.2023.101862] [Medline: 37689445]
53. Szalai E, Tajti P, Szabó B, Hegyi P, Czumbel LM, Shojazadeh S, et al. Daily use of chlorine dioxide effectively treats halitosis: a meta-analysis of randomised controlled trials. PLoS One 2023;18(1):e0280377 [FREE Full text] [doi: 10.1371/journal.pone.0280377] [Medline: 36634129]

XSL•FO
RenderX

54. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 2022;24(8):e36823 [FREE Full text] [doi: 10.2196/36823] [Medline: 36006692]
55. Ramli H, Said S, Ismail A, Dom T. Siwak as a prophetic and evidence-based oral hygiene tool: a qualitative study among Islamic scholars. Islamiyyat 2023;45(2):77-92 [FREE Full text] [doi: 10.17576/islamiyyat-2023-4502-06]
56. Shifman A, Orenbuch S, Rosenberg M. Bad breath—a major disability according to the Talmud. Isr Med Assoc J 2002;4(10):843-845 [FREE Full text] [Medline: 12389360]
57. Mao J, Liu Y, Kando N, Luo C, Zhang M, Ma S. Investigating result usefulness in mobile search. 2018 Presented at: Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Proceedings: 10772 (Lecture Notes in Computer Science); March 26-29, 2018; Grenoble, France URL: http://www.thuir.cn/group/~YQLiu/publications/ECIR2018.pdf
58. Market share of leading desktop search engines worldwide from January 2015 to January 2024. Statista. 2024. URL: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines [accessed 2024-03-03]

Abbreviations

FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
HONcode: Health on the Net Foundation Code of Conduct
JAMA: Journal of the American Medical Association
SMOG: Simple Measure of Gobbledygook

Edited by A Mavragani; submitted 28.10.23; peer-reviewed by M Alhajj, R Sharka, J Lander; comments to author 26.02.24; revised version received 03.03.24; accepted 22.06.24; published 28.08.24.

Please cite as:
Aboalshamat K
Assessment of the Quality and Readability of Web-Based Arabic Health Information on Halitosis: Infodemiological Study
J Med Internet Res 2024;26:e54072
URL: https://www.jmir.org/2024/1/e54072
doi:10.2196/54072
PMID:39196637

©Khalid Aboalshamat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Normalization of Vaping on TikTok Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis: Mixed Methods Study

Sungwon Jung[1], MS; Dhiraj Murthy[1], PhD; Bara S Bateineh[2], MD, PhD; Alexandra Loukas[3], PhD; Anna V Wilkinson[2], PhD

[1]School of Journalism and Media, University of Texas at Austin, Austin, TX, United States

[2]University of Texas Health Science Center at Houston School of Public Health, Houston, TX, United States

[3]Department of Kinesiology and Health Education, University of Texas at Austin, Austin, TX, United States

Corresponding Author:
Sungwon Jung, MS
School of Journalism and Media
University of Texas at Austin
300 W Dean Keeton St
Austin, TX, 78712
United States
Phone: 1 512 749 3267
Email: sungwon.jung@utexas.edu

Abstract

Background: Social media posts that portray vaping in positive social contexts shape people's perceptions and serve to normalize vaping. Despite restrictions on depicting or promoting controlled substances, vape-related content is easily accessible on TikTok. There is a need to understand strategies used in promoting vaping on TikTok, especially among susceptible youth audiences.

Objective: This study seeks to comprehensively describe direct (ie, explicit promotional efforts) and indirect (ie, subtler strategies) themes promoting vaping on TikTok using a mixture of computational and qualitative thematic analyses of social media posts. In addition, we aim to describe how these themes might play a role in normalizing vaping behavior on TikTok for youth audiences, thereby informing public health communication and regulatory policies regarding vaping endorsements on TikTok.

Methods: We collected 14,002 unique TikTok posts using 50 vape-related hashtags (eg, *#vapetok* and *#boxmod*). Using the k-means unsupervised machine learning algorithm, we identified clusters and then categorized posts qualitatively based on themes. Next, we organized all videos from the posts thematically and extracted the visual features of each theme using 3 machine learning–based model architectures: residual network (ResNet) with 50 layers (ResNet50), Visual Geometry Group model with 16 layers, and vision transformer. We chose the best-performing model, ResNet50, to thoroughly analyze the image clustering output. To assess clustering accuracy, we examined 4.01% (441/10,990) of the samples from each video cluster. Finally, we randomly selected 50 videos (5% of the total videos) from each theme, which were qualitatively coded and compared with the machine-derived classification for validation.

Results: We successfully identified 5 major themes from the TikTok posts. *Vape product marketing* (1160/10,990, 8.28%) reflected direct marketing, while the other 4 themes reflected indirect marketing: *TikTok influencer* (3775/14,002, 26.96%), *general vape* (2741/14,002, 19.58%), *vape brands* (2042/14,002, 14.58%), and *vaping cessation* (1272/14,002, 9.08%). The ResNet50 model successfully classified clusters based on image features, achieving an average F_1-score of 0.97, the highest among the 3 models. Qualitative content analyses indicated that vaping was depicted as a normal, routine part of daily life, with TikTok influencers subtly incorporating vaping into popular culture (eg, gaming, skateboarding, and tattooing) and social practices (eg, shopping sprees, driving, and grocery shopping).

Conclusions: The results from both computational and qualitative analyses of text and visual data reveal that vaping is normalized on TikTok. Our identified themes underscore how everyday conversations, promotional content, and the influence of popular figures collectively contribute to depicting vaping as a normal and accepted aspect of daily life on TikTok. Our study provides valuable insights for regulatory policies and public health initiatives aimed at tackling the normalization of vaping on social media platforms.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55591) doi:10.2196/55591

KEYWORDS

electronic cigarettes; vaping; social media; natural language processing; computer vision

Introduction

Background

TikTok is a short video–sharing platform that has grown in popularity, primarily among youth and young adults, since its launch in 2016. Although TikTok restricts uploading videos containing "the depiction, promotion, or trade of drugs or other controlled substances," such content persists on the platform. The e-cigarette or vaping industry and provaping influencers are increasingly using social media platforms, including TikTok, to disseminate content that glamorizes and normalizes vaping [1,2]. This surge of promotional content is concerning because it shapes perceptions, especially those of younger people, suggesting that vaping is a desirable and socially acceptable behavior [3]. By depicting vaping in a humorous light and portraying it as cool, as well as associating it with a particular lifestyle, such content can indirectly send misleading messages, implying that vaping is an integral and normal aspect of daily life [4]. In addition, despite existing policies, some e-cigarette retailers continue to directly market e-cigarette products on social media in ways that are explicitly designed to appeal to specific targeted audiences, such as youth [5]. This includes offering exclusive deals, using attractive packaging, and introducing a variety of flavors. The underlying strategies of these marketing efforts are to desensitize people to the presence and use of e-cigarettes [6], making it seem that vaping is a routine part of everyday life [4]. This trend is alarming because it downplays the health risks associated with vaping and shapes the attitudes and behaviors of young people who are vulnerable and open to suggestion [7].

Previous work based on qualitative content analysis has examined how TikTok posts depict e-cigarettes and vaping [1,2]. These studies are based on manual, human coding, which requires a substantial number of labor hours, a method that does not effectively scale to the large volume of content currently produced and consumed on TikTok [8,9]. This study uses a hybrid method to analyze a large set of TikTok posts depicting e-cigarettes and vaping. A hybrid method combines computational and qualitative thematic analysis of textual and visual data, respectively, to ensure precision and reduce bias. This approach extends existing research by examining both linguistic (ie, text) and visual (ie, video) elements of e-cigarette content on TikTok. The purpose of this study is to identify direct or indirect themes promoting e-cigarettes on TikTok and to describe how these themes might play a role in normalizing vaping behavior on TikTok for youth audiences.

The Normalization of Vaping

Normalization in sociology refers to the process by which an individual adjusts their behavior, thoughts, beliefs, or feelings to conform to the social norms and expectations of their culture or group [10]. This process can occur through repetition, ideology, and propaganda [11]. The normalization of vaping refers to the process by which vaping becomes perceived as a typical, socially acceptable behavior within society [12,13]. The shift in perception and attitude typically occurs gradually through multiple factors that interact and reinforce each other, such as widespread advertising, media representation, cultural trends, peer influences, product accessibility, and perceived safety and harm reduction messages [14-17].

The marketing of vaping devices and flavor options across digital platforms and print media could play an important role in shaping public perceptions of vaping [14]. Similarly, the portrayal of vaping in movies, television shows, and web-based content contributes significantly to its normalization, especially when famous characters and models are depicted using e-cigarettes, subtly implying that vaping is acceptable and trendy [15]. Peer influence also has a strong impact on individuals' attitudes toward vaping; if vaping is common within a peer group, others in the group are more likely to vape to fit in [16]. The ease of accessing e-cigarettes, boosted by web-based sales and in the absence of strict regulations, further increases their use. Finally, the perception of vaping as a safer alternative to traditional smoking, driven by harm reduction messages, lowers barriers to experimentation and regular use [18]. All these factors collectively create an environment where vaping is increasingly normalized, especially among younger people.

The normalization of vaping carries considerable implications for public health, especially if the risks associated with vaping trends are not adequately addressed or recognized; for instance, research shows that the normalization of vaping, especially among youth, can increase dependence risk, normalize smoking behavior, and potentially lead to the initiation of traditional tobacco cigarettes, especially among those who might never have smoked otherwise [19,20].

Social Media Marketing Strategies and the Normalization of Vaping

Social media platforms have become a pivotal tool in shaping public behaviors in recent years [21,22]. TikTok, for instance, offers users the opportunity to share their thoughts, experiences, and preferences to a vast audience [23]. Recent data indicate a surge in content related to vaping use on these platforms. Notably, the majority of messages are positive and favorable toward vaping [24]. Content in these messages ranges from personal testimonials about the benefits of vaping over traditional smoking to endorsements by influencers of specific e-cigarette brands and flavors [25]. Such messages might lead to gradually establishing the belief that vaping is a normalized and socially acceptable behavior [26,27], regardless of any associated potential health risks or controversies [28,29].

Both direct (ie, explicit promotional efforts of e-cigarettes) and indirect (ie, subtler strategies presenting the promotion in a lifestyle-oriented manner) marketing through social media platforms, including TikTok, significantly influence the popularity of e-cigarette use [1,2]. Individuals who are exposed

XSL•FO
RenderX

to promotional or user-generated e-cigarette content have a lower perception of harm, have more positive attitudes toward vaping, and have a higher risk of initiating e-cigarette use [26,27]; for example, a recent longitudinal study found that adolescents who reported seeing e-cigarette posts at least 1 time a week on TikTok had a higher risk of lifetime and current e-cigarette use [24]. The findings also indicated that the frequency of TikTok use was a significant predictor of initiating e-cigarette use [24]. The findings from another study indicated that when young people on social media (friends, influencers, etc) are portrayed as happy and popular while using e-cigarettes, it creates a perception that vaping is a normative behavior within their social circles and, as such, is a behavior to emulate [30].

TikTok's community guidelines [31] regarding "drugs, controlled substances, alcohol, and tobacco" clearly state that users should not share content that involves "the purchase, sale, trade, or solicitation of drugs or other controlled substances, alcohol or tobacco products (including e-cigarettes)." However, e-cigarette content is prevalent on TikTok, often masked by indirect advertising methods [32]. Instead of traditional advertisements, e-cigarette brands collaborate with influencers to promote their products in their daily life posts and videos, making the promotion seem more organic. Influencers also link e-cigarette use with appealing lifestyles, such as wellness, party scenes, or luxury [2]. The repetitive e-cigarette content posted by influencers could potentially normalize the act of vaping and make users more receptive to vaping, especially if the content is presented without highlighting potential risks. This psychological phenomenon, often termed the "mere exposure effect," suggests that people tend to develop a preference for behaviors merely because they are familiar with them [33].

In addition, content featuring individuals from different backgrounds engaging with e-cigarettes has the potential to appeal to wider demographics. Individuals often connect more with content that mirrors their own experiences, identity, ethnicity, and age, which may prompt them to think about experimenting with e-cigarettes [34]. The dynamic interplay and use of these indirect marketing strategies on TikTok and other platforms contribute to the normalization of vaping by integrating it into popular culture, social practices, and lifestyle images among various backgrounds. Thus, in this study, we aimed to identify and describe themes—either direct or indirect—that capture how e-cigarettes are promoted on TikTok and understand how these themes might play a role in the perception and normalization of vaping among young people.

The Need to Use Computational Methods to Study Vaping on Social Media

Given the rising popularity of TikTok as a social media platform, several qualitative studies have successfully examined e-cigarette–related content on TikTok. Sun et al [30] found that the themes capturing e-cigarettes depicted in TikTok videos included humor and jokes, lifestyle, marketing, vaping tricks, nicotine and addiction, creativity, and caution. Purushothaman et al [35] concluded that more than two-thirds of the posts contained references to e-cigarettes. Although these studies advance our understanding of how e-cigarettes are depicted on social media, they are limited to the examination of a small number of posts because the content was analyzed manually. The volume of social media posts has rapidly risen and outpaced thematic analysis based on manual content analysis. Computational methods—both natural language processing and computer vision—can help researchers analyze e-cigarette–related content on social media at scale [36,37].

Media scholars increasingly use computational analytic methods (eg, developing a computational interface to facilitate the work of human coders) to manage ever-larger data sets [38]. Computational analytic methods have the potential to overcome many of the sampling and coding limitations of conventional content analysis. Early studies in computational social science focused on the TikTok platform have primarily concentrated on text analysis using natural language processing [39]. In the vast domain of health care, natural language processing serves as a powerful tool with diverse applications, offering insights and solutions to various challenges faced by clinicians, researchers, and health care providers [40]. Natural language processing techniques are instrumental in mining textual data from web-based health forums, social media platforms, and patient-generated content [41]. By analyzing conversations, posts, and comments, natural language processing algorithms can identify emerging health trends, public perceptions of diseases and treatments, adverse drug reactions, and even sentiments toward health care providers and services; for example, Sun et al [42] used topic modeling, a form of natural language processing and machine learning that automatically identifies topics or themes in a collection of documents [8], to examine COVID-19 vaccine–related posts on TikTok through textual data. Furthermore, using natural language processing, Bharti et al [43] created a multilingual conversational bot to deliver essential health care education, information, and advice to individuals with chronic conditions.

More recent social media research has extended the scope of computational analytic methods to encompass computer vision techniques, a method designed for analyzing social media images and videos [44,45]. Computational techniques rooted in deep learning have facilitated the automated examination of vaping-related visual content shared on social media platforms such as Instagram and TikTok [45]. Moreover, machine learning models based on computer vision have been found to effectively and precisely recognize e-cigarette–related content (eg, vaping devices and vapor clouds) on predominantly visually oriented social media platforms [44]. The transition from text-focused analysis to computer vision analysis signifies an evolution in TikTok research methodologies, offering a more comprehensive understanding of the platform's multifaceted content and societal implications. Nevertheless, there is still a dearth of studies applying computational content analysis methods to the context of vaping. Research is particularly lacking in the realm of combining computational methods with both text and images, especially to detect vaping-related themes.

The Fundamentals of Computational Image Clustering

Computational image clustering is a popular computer vision method that groups images based on visual similarity [46]. Computer vision involves developing algorithms and models to help computers interpret and understand images and videos

XSL•FO
RenderX

[47]. To mimic human vision, computer vision teaches machines to recognize objects, scenes, and patterns as well as understand visual inputs [48]. Computer vision performs object recognition and content categorization using image features, patterns, and structures [49]. The goal is to find patterns or similarities in a large number of images without knowing their content or categories. Each cluster comprises images with similar visuals. This procedure enhances visual data navigation and comprehension, particularly for large image collections. There are 2 steps to image clustering: feature extraction and feature-based clustering.

Several machine learning computational models are capable of performing feature extraction. In this step, visual attributes from the input images are extracted. Initially, visual elements such as edges, patterns, and colors may be included [46]. The ultimate goal is to identify objects or patterns in images and capture more complex and higher-level features. Image pixel data are converted into numerical representations for computational analysis in this process [50]. Clustering algorithms group similar images into coherent categories after extracting these visual features [51]. The algorithms categorize images into meaningful groups to facilitate retrieval, analysis, and interpretation.

A convolutional neural network (CNN) can extract features and cluster images. CNNs are artificial intelligence models that specialize in image classification [52]. The main steps in the CNN procedure are feature extraction and classification [53]. Vision transformers (ViTs) are a deep learning model architecture commonly used in computer vision tasks. Transformers are a deep learning model architecture designed for natural language processing tasks, characterized by their attention mechanism and ability to capture long-range dependencies [54]. Transformers' self-attention and parallelization make them good at sequential data processing [55]. K-means clustering is a popular unsupervised machine learning method for splitting a data set into nonoverlapping clusters. This algorithm determines data similarity [56]. Using the feature vectors from the previous step, k-means clustering groups similar images [57].

Using Computational Image Clustering on Vape-Related Content on Social Media

Multimodal content clustering has not been used to study the normalization of vaping on social media. However, image clustering using computational methods (ie, CNNs and k-means clustering together) has been used successfully to study visual content on social media; for example, Ketonen and Malik [58] used residual network (ResNet) and Visual Geometry Group CNN models with k-means clustering to analyze and describe Instagram posts related to vaping. The authors found 7 clusters—*e-liquid*, *e-liquids*, *e-cigarettes*, *product packages*, *people*, *statements*, and *miscellaneous*—with an accuracy of 90.76%. One limitation of the study was that it relied solely on posts under a single hashtag, which raises the possibility that not all posts were directly relevant to e-cigarettes [58]. Furthermore, the use of images from the identified clusters for training a classifier for predicting a category for each image in the data set may result in the inclusion of certain inaccuracies within the final group labels. Vassey et al [59] measured the influence of e-cigarettes on Instagram using the Inception v3 model, one of the CNN-based models developed by Google, which correctly classified >90% of the images containing men, women, e-juices (the liquid used in e-cigarettes), and mod system devices (a type of e-cigarette that allows users to customize their vaping experience) across all samples. However, the Inception v3 model's training was limited to recognizing a single class per image although specific images contained 2 distinct categories.

Setiawan and Purnama [60] applied the DarkNet19 CNN architecture to extract image features from tobacco leaf images. The authors used k-means to cluster these images (with an accuracy ranging from 0.81 to 0.93 for all 3 classes). Previous research has successfully integrated feature extraction techniques using CNN architectures based on machine learning and clustering methods. However, previous research was restricted to the use of posts under a single hashtag, and the model used needed to be updated for strong, robust performance to overcome the limitations of previous social media studies related to vaping by collecting large amounts of data using multiple hashtags and using a model with robust performance.

This study aimed to use a hybrid computational and qualitative approach to analyze a substantial volume of TikTok content related to e-cigarettes, using k-means clustering and pretrained ViT and CNN models. By identifying direct and indirect themes in the promotion of e-cigarettes on TikTok, we sought to elucidate how these themes contribute to the normalization of vaping behavior among youth audiences. This comprehensive analysis addresses gaps in existing research methodologies and provides insights into the multifaceted factors influencing the portrayal and perception of vaping on social media platforms. The findings from this study are necessary to inform the development of interventions targeting adolescents and young adults that counteract the glamorization of vaping, as evidenced by the strategies used by e-cigarette companies [61]. The findings may also be used to inform legislative actions, educational campaigns, and digital interventions aimed at denormalizing harmful behaviors and safeguarding public health, a need further emphasized by recent findings on the prevalence of provaping content on platforms such as TikTok [62].

Methods

Hybrid Approach

We used a hybrid approach—computational as well as qualitative methods—to derive vaping-related themes from TikTok posts (Figure 1).

We collected 18,946 TikTok posts published from 2018 to 2023 using a set of vape-related hashtags (eg, *#vapetok*, *#boxmod*, and *#vaper*; refer to Multimedia Appendix 1 for the complete list of hashtags). Although TikTok has a research application programming interface (API) for academics, it often returns inaccurate metadata; therefore, we used a data scraping approach, which obtained accurate data and metadata. Data scraping was performed in 2 steps. First, we used Zeeschuimer [63], a Firefox browser extension that facilitates the scraping of data from social media platforms, including Twitter,

XSL•FO
RenderX

Instagram, TikTok, and LinkedIn. Prior studies have demonstrated the efficacy of Zeeschuimer as a data mining tool [64,65]. In accordance with previous studies, we used the Zeeschuimer tool to obtain e-cigarette–related content from TikTok's social media data.

Figure 1. Project architecture (relevant figures and tables are referenced in parentheses). OpenCV: Open-Source Computer Vision Library. TF-IDF: term frequency–inverse document frequency.

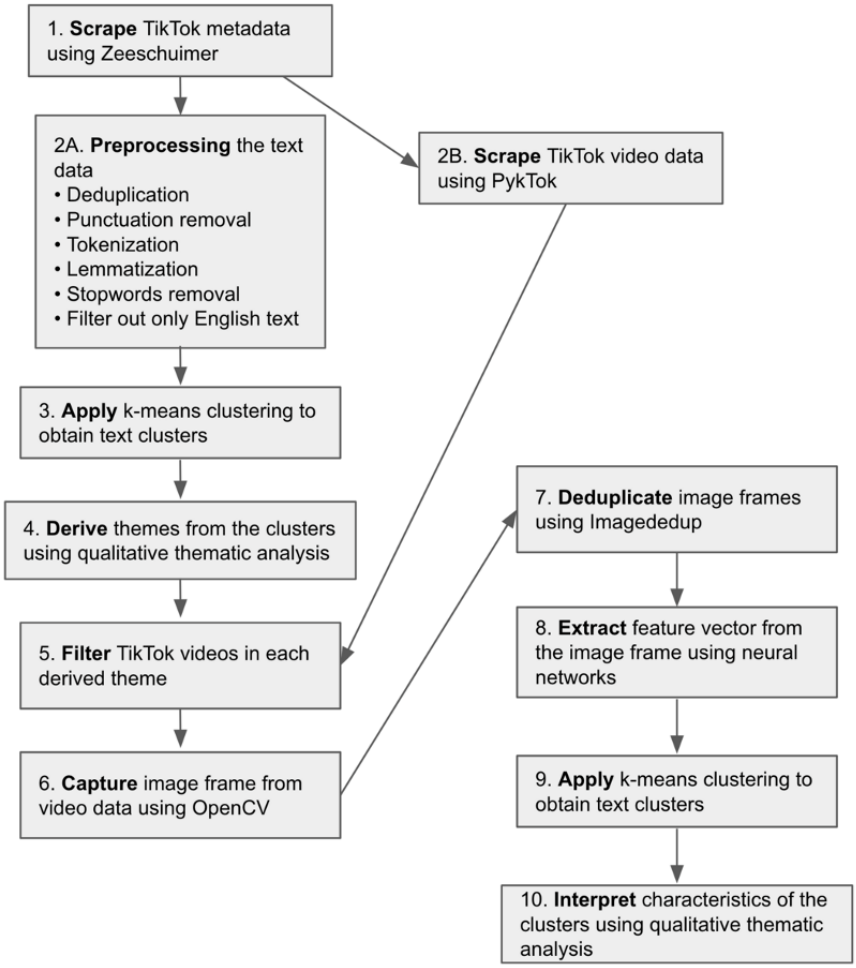

Zeeschuimer enables the export of the collected metadata in newline delimited JSON format. To scrape the data using Zeeschuimer, we scrolled down the TikTok hashtag URLs for each targeted hashtag. We collected 20,575 newline delimited JSON files for each hashtag. The metadata fields included fundamental video information such as *video_id*, *video_timestamp*, *video_duration*, *video_description*, *video_is_ad*, *video_height*, *video_duration*, *video_location*, and *author_name*, as well as user engagement elements such as *video_sharecount*, *video_commentcount*, and *video_playcount*. Next, we augmented our data set with an additional step: we collected TikTok videos using the *Pyktok* [66] package. *Pyktok* obtains data directly from the JSON objects embedded within TikTok web pages and from APIs that are not publicly documented. On the basis of the *video_id* collected using Zeeschuimer, we verified the valid *video_id* that is public and downloaded the TikTok videos. Accordingly, we obtained 18,946 data points for 50 hashtags (Figure 2).

Figure 2. Primary data set evolution for the analysis: a flowchart of the data processing pipeline, starting with metadata collection, followed by video ID deduplication, extraction of images, and finally image deduplication. FPS: frames per second.

Ethical Considerations

This study underwent review by the institutional review board at the University of Texas at Austin (STUDY00006075) and was determined not to constitute human participant research, as defined by regulations set forth by the Department of Health and Human Services and the Food and Drug Administration. Consequently, institutional review board review and approval were not required for this study.

As the data used in this study are publicly available, informed consent was not required. To ensure confidentiality and privacy, all data were anonymized. Specifically, we took measures to prevent the identification of the individuals whose posts were

XSL•FO
RenderX

used for analysis, including blurring identifiable information such as user IDs or TikTok IDs in any images used.

Text Clustering Methods

We combined the textual data from the *video_description* column. We then preprocessed the data using the Python *NLTK* library [67] to perform lemmatization (finding the basic form of the word inflected in multiple forms in a sentence), punctuation removal, and capitalization removal. All collected text was tokenized or divided into word tokens, which are grammatically indivisible language elements. After the basic text preprocessing, we used the Python package *langdetect* [68] to detect and filter out only English text. Subsequently, we deduplicated the data set, which resulted in 14,002 posts, to define the clusters. As the TikTok data set lacked correct answer labels, we created a cluster using the k-means algorithm, an unsupervised learning method. The document was first converted to a vector using the term frequency–inverse document frequency (TF-IDF) method. The TF-IDF weight is a statistical value used in information retrieval and text mining that indicates how important a word is in a specific document when there is a document group consisting of many documents [69]. We set the minimum document frequency to 5 as a parameter of the TF-IDF vectorizer. Words that appeared fewer than 5 times were excluded from the word dictionary. In addition, we set the *ngram_range* to (1, 2), which means that up to 2 groups of words were treated as if they were 1 word.

Next, we normalized the word score of each document vector to keep it between 0 and 1 so that vector clustering could be performed effectively. We used cosine similarity between vector distances to perform clustering. We used k-means clustering as a computational clustering method. This is a clustering method in which the sum of the squared distances between the average vector in each cluster and the vectors in the cluster is minimized. As the center point is chosen at random, the clustering result may be different or poor. To determine the optimal number of clusters (k), we used the elbow method, which has been proven to be an excellent technique [70]. The elbow method is selected after considering the section where the inertia value, representing the sum of the distances between clusters, drops rapidly as the optimal k.

After identifying the optimal number of text clusters, we conducted qualitative textual thematic analysis to assess the coherence and relevance of the clusters, further validating the robustness of our findings. We met twice as a group to evaluate the uniqueness of the themes derived from the elbow method, and the final themes were decided by consensus. We investigated the posts that belonged to each cluster in depth, analyzed the data composition of each cluster, and grouped the posts into themes.

Video Clustering Methods

After conducting the thematic analysis on the text corpus, we conducted image clustering on the derived themes. We examined whether the themes could be categorized according to the videos' visual elements. Within the themes, we also identified subclusters. The average length of the videos with themes was between 22 and 31 seconds. Theme 1 encompassed 2741 videos, theme 2 included 1272 videos, theme 3 had 1160 videos, theme 4 had the largest collection of 3775 videos, and theme 5 included 2042 videos.

We initially sampled every 25th video from the total data set. We discovered that if the video lasted longer than 10 seconds, most of the first few seconds were not devoted to vaping. Following this observation, to capture the vape-related moments in the TikTok posts, we chose a frame rate of 10 frames per second and extracted frames at regular intervals of 10 frames per second from each collected video. This approach was used to mitigate any potential bias that may arise from the researcher's subjective selection process.

Subsequently, the Python-based *Imagededup* [71] package was used to eliminate duplicate images. As mentioned previously, the *Imagededup* package has been widely used in computer vision research, demonstrating its efficacy in image deduplication [72,73]. The software package used a built-in CNN encoder for image deduplication, and the CNN encoder used the *mobilenet_v3_small* model to extract features from the images in the original image directory.

By computing the cosine similarities after the encoding of the images, the CNN algorithm selectively identified and removed only those image files duplicated from the original image directory. Cosine similarity is a mathematical metric used to compare the similarity of 2 vectors, extended to image comparison [74]. To preserve visual features, the CNN algorithm encoded each image into a vector. When these vectors were compared using the cosine similarity metric, the outcome indicated how similar the 2 images were regarding visual content. A cosine similarity score of 1 indicated that the images were identical, whereas a score of 0 indicated dissimilarity. The strict cosine similarity threshold score was set at 0.85, which means that only image pairs with a cosine similarity score of ≥0.85 would be recognized as duplicates. Images with a lower similarity score would be considered distinct.

After deduplication, we obtained 2520 (65.74%) images for theme 1 from 3833 images. The same procedure was replicated for themes 2, 3, 4, and 5. We obtained 1242 (57.18%) images after removing duplicates from a pool of 2172 images for theme 2. In addition, we obtained 1269 (69.69%) images for theme 3 from 1821 images, and 4276 (57.05%) images for theme 4 from 7495 images. Finally, we obtained 1773 (71.64%) images for theme 5 from 2475 images.

CNN-Based Clustering

The CNN model was trained on a large data set to learn visual features, which is referred to as pretraining [75]. To classify images, feature extraction uses CNN layers to recognize increasingly complex patterns and objects. CNNs can classify images by generating a probability distribution over output categories [75]; for example, CNNs can recognize e-cigarettes and other similar items in images. After pretraining, the CNN classification layers (which classify images) are removed, leaving only the feature extraction layers [76]. These feature extraction layers extract high-level attributes from images, converting input images into feature vectors [77]. These feature

XSL•FO
RenderX

vectors are useful for image clustering because they capture image shapes, textures, and object parts.

We used the *PyTorch* library [78] for the image clustering. We resized the images to a uniform size of 224×224 pixels. We converted the images to a tensor format, which is a multidimensional array or data structure that generalizes scalars, vectors, and matrices, suitable for deep learning. We used the *DataLoader* function in *PyTorch*, which creates batches of images from the data set to be used during model training and evaluation from the data set. The function shuffles the data to ensure randomness during training. Next, we loaded a pretrained CNN model (ie, Visual Geometry Group with 16 layers [VGG16] and ResNet with 50 layers [ResNet50]), which is a built-in deep neural network model that has been trained on a large data set for image classification tasks. We specifically opted for the VGG16 [79] and ResNet50 [80] models due to their widely recognized reputation for robustness and accuracy in tasks related to image recognition [81,82]. We removed the final fully connected layers from the model, leaving only the convolutional layers. These convolutional layers were used for feature extraction. The images were passed through the VGG16 network, and the output features were collected. These features capture high-level visual information about each image and were used for clustering.

Features extracted from the images were stored in the *features* list. We turned on the evaluation mode of the model and then iterated a loop through the batches of images in our data set. For each batch, the model processes images without gradient computation (no gradients are backpropagated) using *torch.no_grad()*, and the output features are appended to the features list. The extracted features were concatenated into a single tensor and reshaped for subsequent clustering. We then determined the optimal number of clusters for the features we obtained using the silhouette score. The silhouette score measures how similar each data point in 1 cluster is to other data points in the same cluster compared with data points in other clusters. Our process involved iterating through different numbers of clusters, ranging from 2 to 10, to explore various clustering configurations. For each iteration, we computed the silhouette score for the corresponding clustering number. The silhouette score ranges from –1 to 1, where a higher score indicates better-defined clusters.

By systematically evaluating the silhouette scores across different cluster numbers, we identified the number of clusters associated with the highest silhouette score, indicating the optimal clustering number. After determining the optimal number of clusters, we performed k-means clustering using this number of clusters. The k-means algorithm groups the image features into clusters based on their similarity. We later saved images belonging to each cluster in separate subdirectories. For each cluster, we also created a CSV file to store information about the images in this cluster for qualitative analysis.

ViT-Based Clustering

The ViT extends the transformer architecture to process images for computer vision by treating them as nonoverlapping patches or tiles, in contrast to CNNs, which process images as pixels. The ViT model treats each patch as a "word," using self-attention mechanisms such as natural language processing transformers [83]. This allows the model to capture image patch dependencies and relationships over time. In the ViT, positional embeddings aid the model in understanding the relative positions of patches in an image [84]. Maurício et al [85] discovered that ViTs perform similarly to traditional CNNs in image classification benchmarks. The ViT model was used in recent health studies and performed well [86,87]. The feature extraction process for ViTs is similar to that for CNN models. We extracted features from our data set images using pretrained ViT models (eg, ViT-base and ViT-large). When trained on large image data sets, these models can encode images into rich feature vectors. The ViT model is used to extract features from images. The model transforms each image into a high-dimensional feature vector (representation). This vector represents the content and appearance of the image. We used *Transformers* [88], a *PyTorch*-based Python library created by Hugging Face [89], to implement the ViT model. We specifically chose the ViT model due to its proven effectiveness as a transformer-based model in comprehending visual content through tokenization [90].

After data transformation (resizing and converting into tensors, as explained previously), we initialized the ViT model using the *ViTFeatureExtractor* and *ViTForImageClassification* functions. Due to its computing capabilities, we used the *vit-base-patch16-224-in21k* model for the pretrained model. The model's name suggests that it is moderately sized ("base"), processes images as 16×16 pixel patches, accepts input images of size 224×224 pixels, and has been trained on a data set containing 21,000 distinct image classification categories. The code iterates through the data set to extract image features using the ViT model, and the extracted features are stored in the features list and concatenated into a single tensor for further analysis. We replicated the same step from CNN-based image clustering for the feature-based clustering task. We computed the silhouette score to determine the optimal number of clusters.

Video Classification

We switched from clustering to classification to evaluate the quality of the clusters produced by the unsupervised k-means algorithm in each model. This approach was used to determine whether the features learned during the clustering phase are informative and can be used to predict cluster assignments accurately. We divided all features and cluster labels into training (80%) and testing (20%) data sets for classification. We used a simple neural network classification model with 2 linear layers and rectified linear unit activation in between. Rectified linear unit is an activation function commonly used in artificial neural networks, particularly in deep learning models. The model underwent training with cross-entropy loss and the Adam optimizer. Cross-entropy loss or log loss is a measure of how well a probabilistic classification model predicts the probability distribution of the true labels given the input data. The Adam optimizer efficiently adjusts the model's parameters during training by adapting learning rates and incorporating momentum that accelerates the learning in the relevant directions [91]. We trained the model for 10 epochs on each theme. We then calculated the F_1-score, which is a classification metric that combines precision and recall into a

XSL•FO
RenderX

single value and provides a balanced measure of a model's performance.

Qualitative Thematic Analysis

After obtaining the output clusters from 3 individual models (ie, ViT, VGG16, and ResNet50), we checked the samples from each cluster for each model to verify the clustering result. We determined the best model based on the performance; next, we performed qualitative thematic analysis on each theme after studying the final clustered images for each cluster on the themes. In validating the image clusters, we encountered challenges associated with interpreting visual similarities and patterns. To overcome this hurdle, we conducted a thorough examination of the clustered images, comparing them both within and across clusters to verify the consistency and accuracy of our clustering results. We initially picked 25% of the samples from each cluster, comprehensively analyzed each cluster's images to determine whether the images were correctly clustered, and examined the characteristics of the cluster. For the second phase, we randomly selected 50 videos from each of the 5 themes to compare the content against the theme descriptor.

Results

Text Clustering Results

Overview

The elbow method determined the optimal number of clusters in the textual data set to be 15 (Figure 3). Our final sample was 14,002 TikTok posts after deduplication, but 3012 (21.51%) posts were unrelated to our research scope and removed from our sample, leaving 10,990 (78.49%) e-cigarette–related posts (see Multimedia Appendix 2 for details), from which we successfully identified 5 themes (summarized in Table 1).

The computational textual analysis identified 5 distinct clusters, and our qualitative analysis indicated that these 5 clusters were organized by the following themes: *general vape*, *vaping cessation*, *vape product marketing*, *TikTok influencer*, and *vape brands*. These themes provided insights into the web-based discourse surrounding vaping, from general discussions to specific aspects such as cessation, influencers, and brands (Figure 4).

Figure 3. Elbow method result from k-means text clustering indicating that the optimal number of clusters is 15.

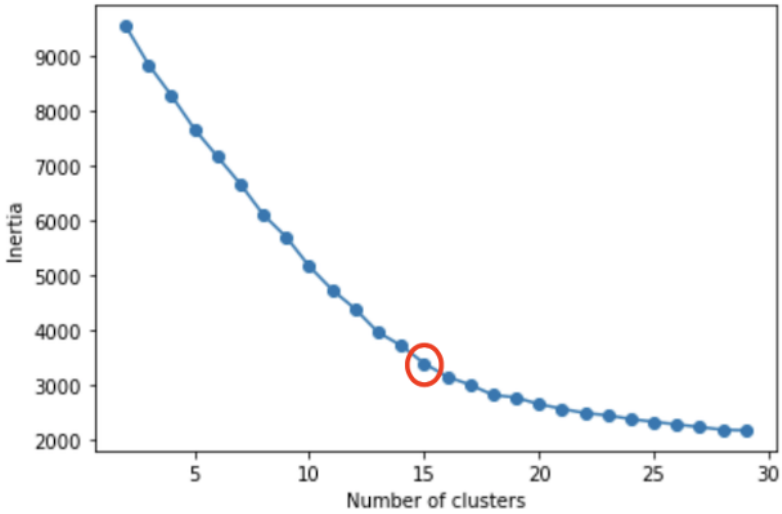

Table 1. Results of the textual thematic analysis.

Themes	Aggregated clusters (n=12), n (%)	Posts (n=10,990), n (%)	Indicators (words or hashtags)
General vape (including components, eg, coil, cartridges, and pod, as well as use, eg, blinker)	4 (33.33)	2741 (19.58)	vapetricks, cigarroeletronico, pod, vapelife, smoke, vapeo, vapor, vapemod, vapers, coil, vapepen, vapersontiktok, nicotine, know, funny, smoke, y'all, relatable, stop, vapeporn, vapefamily, wonderwaterdrip, girlvaper, vapersontiktok, carts, blinkersonlyfoo, fakecarts, blinkereyes
Vaping cessation	1 (8.33)	1272 (9.08)	vapelife, vapingislife, quit, addiction, quitvaping, health, vaping is bad, stop, stopvaping
Vape product marketing (including flavors)	1 (8.33)	1160 (8.28)	ultimatejuice, flavour, liquid, ultimatepuff, premiumliquid, eliquidshop
TikTok influencer (includes skate, gamer, vape, and other influencers)	3 (25)	3775 (26.96)	skateboarding, skater, nintendo, vapetiktoker, game, nintendo, retro, xbox, retrogrames, videogames, sega, console, games, gaming, thevideogamecollector, kingtinotazo (specific influencer)
Vape brands	3 (25)	2042 (14.58)	elf, geekbar, elfbars, pods, fresh mary, whereismymary, spookymary, scarymary, lostmarybm600 Juulgang, juullife, juulchallenge, juuling, juulsquad, mint, mango

XSL•FO
RenderX

Figure 4. An example of each of the themes derived from text clustering (selected representative images from actual posts corresponding to each theme).

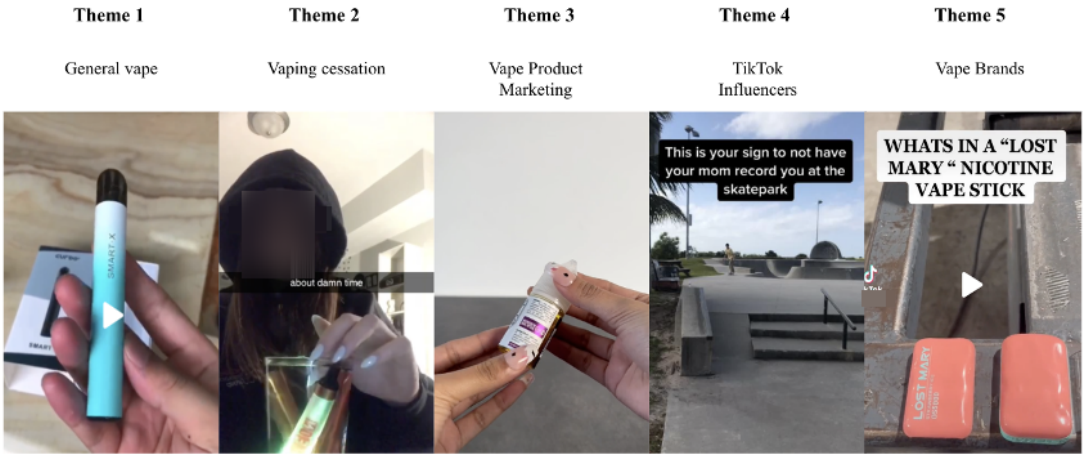

Theme 1: General Vape

This theme encompassed general discussions and content regarding vaping, such as its components and use. *General vape* consisted of a substantial number of posts (2741/14,002, 19.58%) and a variety of indicators, including *vapetricks*, *coil*, and *vapelife*. These posts seemed to focus on various aspects of vaping, from the technical, such as coils and pods, to the experiential, such as user experiences. In addition, terms such as *blinker*, which is a light that flashes on and off to prevent overheating of the vaping device, suggested discussions of vaping techniques and tricks. This theme reflected a diverse and active vaping community on TikTok, engaging in discussions about their vaping experiences and preferences and possibly promoting or sharing vaping-related knowledge.

Theme 2: Vaping Cessation

This theme had the second lowest number of posts (1272/14,002, 9.08%). This theme discussed quitting vaping, addressing addiction, and vaping's negative health effects. Indicators such as *quitvaping*, *stopvaping*, and *vaping is bad* suggested that a large number of individuals used social media platforms to seek support, share their experiences, and declare their intentions to quit vaping. This theme emphasized the awareness and concerns regarding the addictive nature of vaping, as well as the efforts of some vapers to overcome their addiction.

Theme 3: Vape Product Marketing

This theme pertained to the marketing of vape products, with a particular focus on flavors, and was characterized by a substantial post number (1160/14,002, 8.28%). The posts associated with this theme often included indicators such as *ultimatejuice*, *flavour*, *liquid*, *ultimatepuff*, *premiumliquid*, and *eliquidshop*. These indicators highlighted the prominence of discussions and content related to the marketing and promotion of various vape products, with an emphasis on the diverse range of flavors available in the vaping industry. This theme reflected the significance of flavor offerings in the marketing strategies of vape product manufacturers and highlighted the considerable attention and discourse surrounding this aspect within the vaping community or industry.

Theme 4: TikTok Influencer

This theme comprised a significant portion of the conversation (3775/14,002, 26.96%), entailing discussions regarding TikTok influencers and focusing on skateboarding, gaming, and vaping categories. The use of indicators such as *skateboarding*, *gamer*, and *vapetiktoker* emphasized the diversity of the discussed influencers. This theme indicated that influencers may have shared their content or discussed their impact on the platform. This theme reflected the prominent role of influencers in shaping web-based conversations, especially on platforms such as TikTok.

Theme 5: Vape Brands

This theme focused on specific vape brands and their products, accounting for 14.58% (2042/14,002) of the posts. The *vape brands* theme included terms such as *Juulgang*, *elfbars*, and *mint*. Users engaged in conversations regarding various e-cigarette brands, potentially discussing product preferences, flavors, and experiences. Specific brand names such as *Juul* indicated that many users were discussing this brand. This theme focused on the impact of brand loyalty and product diversity on the vaping community.

Video Clustering Results

Overview

We chose to use the ResNet50 model for the qualitative thematic analysis based on the F_1-score. Table 2 displays the ResNet50 model's consistently high performance on each theme (average F_1-score: 0.97), which is likely due to the improved performance of neural networks with more layers.

The image clustering results for each theme from the preceding textual thematic analysis revealed the normalization of vaping on TikTok; for example, TikTok influencers in domains such as skateboarding and gaming and TikTok vloggers uploading their daily shopping lists depict vaping as an ordinary component of daily life. Furthermore, the individuals in these videos are shown vaping in various locations (eg, at home or in a car). These findings suggest a concerning trend of normalizing vaping as an everyday practice on TikTok.

XSL•FO
RenderX

Table 2. Performance metrics for video classification for each model. The residual network with 50 layers (ResNet50) model performed consistently well across all 5 themes compared with the other 2 models.

	F_1-scores				
	General vape (theme 1)	Vaping cessation (theme 2)	Vape product marketing (theme 3)	TikTok influencer (theme 4)	Vape brands (theme 5)
ResNet50	0.974	0.972	0.960	0.975	0.975
VGG16[a]	0.932	0.986	0.953	0.955	0.980
ViT[b]	0.943	0.941	0.847	0.986	0.778

[a]VGG16: Visual Geometry Group model with 16 layers.

[b]ViT: vision transformer.

Theme 1: General Vape

The findings of our image clustering analysis indicated that the images in the *general vape* theme formed 3 distinct subclusters. We characterized each subcluster individually after carefully examining the images (Figures 5A-5C).

Figure 5. Sample image clustering results from theme 1 (general vape); 3 distinct subclusters were formed. (A) Videos portraying individuals actively participating in the act of vaping by using vaping devices. (B) Videos featuring individuals who choose not to use vaping devices, opting instead to partake in conversations that are unrelated to the act of vaping. (C) Videos showcasing individuals participating in conversations that cover a range of topics, including vaping-related subjects such as the demonstration of vaping devices as well as broader subjects such as personal health in which the creators seek to satirize antivaping messages.

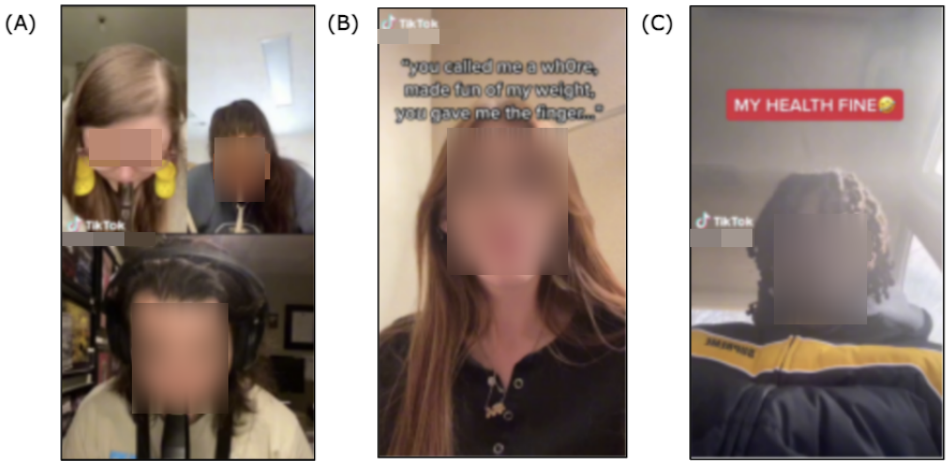

Theme 2: Vape Cessation

The findings of our image clustering analysis indicated that the images in the *vape cessation* theme formed 2 distinct subclusters: (1) experts or users with knowledge about vaping come on screen to inform video viewers about the link between vaping and potential diseases (eg, lung-related diseases) and urge them to quit vaping (Figure 6A), and (2) users who are actually quitting vaping or have quit vaping share their own experiences (Figure 6B).

XSL•FO
RenderX

Figure 6. Sample image clustering results from theme 2 (vape cessation); 2 subclusters were formed. (A) Experts or users with knowledge about vaping come on screen to inform video viewers about the link between vaping and potential diseases (eg, lung-related diseases) and urge them to quit vaping. (B) Users who are actually quitting vaping or have quit vaping share their own experiences.

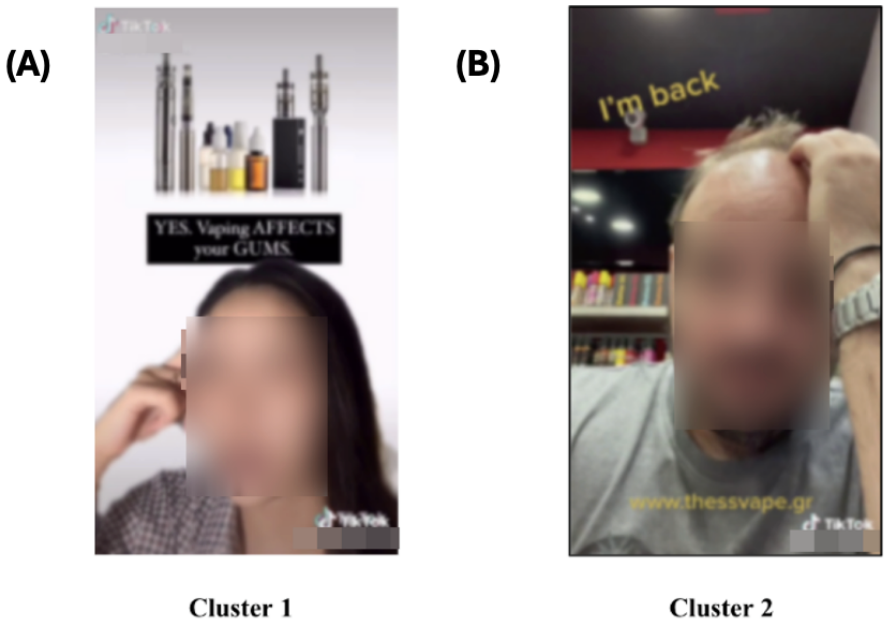

Theme 3: Vape Product Marketing

We identified 2 distinct subclusters within the *vape product marketing* theme: (1) videos that actively showcased and introduced vaping devices, illustrating the various aspects and functionalities associated with these products (Figure 7A); and (2) videos that featured local shops actively selling and advertising vaping devices, indicating a localized and direct promotional approach (Figure 7B). Despite TikTok's explicit guidelines opposing the direct promotion of vaping products, these subclustering results underscore the effectiveness of direct advertising or marketing of vaping devices at the microlevel, encompassing both individual users and local shops. This observation sheds light on the nuanced ways in which TikTok's platform is used for promotional activities that may, at times, circumvent platform guidelines to reach a targeted audience.

Figure 7. Sample image clustering results from theme 3 (vape product marketing); 2 subclusters were formed. (A) Videos that actively showcased and introduced vaping devices, illustrating the various aspects and functionalities associated with these products. (B) Videos that featured local shops actively selling and advertising vaping devices, indicating a localized and direct promotional approach.

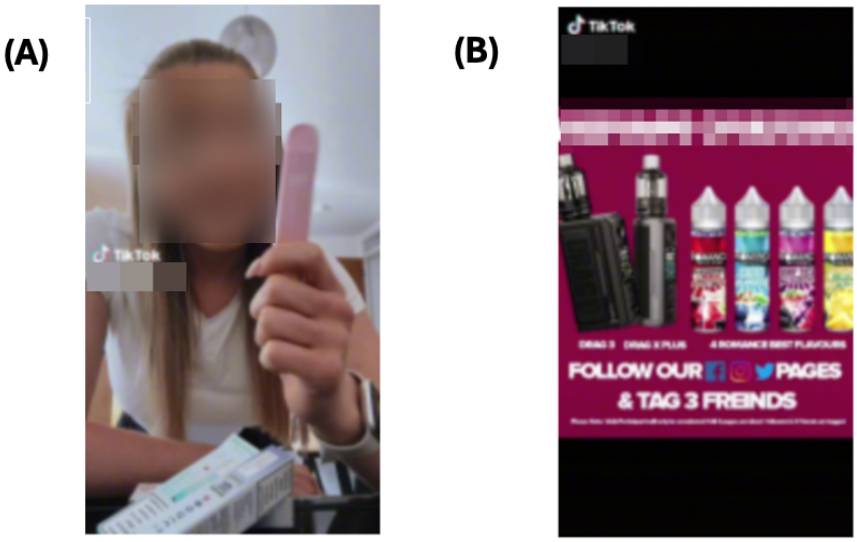

Theme 4: TikTok Influencer

We could not find differences between the images in each subcluster within the *TikTok influencer* theme. These findings demonstrate that the output of a black box computer algorithm cannot be wholly trusted. Instead, a human-in-the-loop process, in which researchers interpret the results and determine their significance, is required. Although a vaping device was visible in the video, it was not the focal point but a supplementary element. In other words, TikTok videos in the *TikTok influencer*

XSL•FO
RenderX

cluster were depicting vaping alongside daily or quotidian activities such as playing video games, driving, grocery shopping, and reading books (Figure 8). There were also vape influencers in the videos, but instead of actively promoting vape products, they flaunted adding new decorations to vapes or customizing cartridges. In other words, most images within the *TikTok influencer* theme depicted vaping as a natural part of daily life.

Figure 8. Sample image clustering results from theme 4 (TikTok influencer); no differences were found between the images within each subcluster.

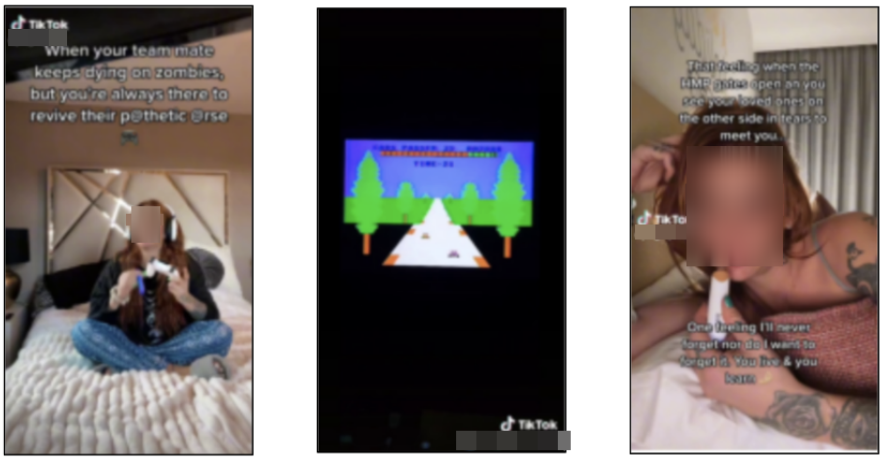

Theme 5: Vape Brands

We identified 2 distinct subclusters within the vape brands theme (Figures 9A and 9B). In other videos, point of view (POV) is frequently used as a text caption. POV is frequently used in captions and on-video text captions to indicate that the viewer should view the scene from their (the creator's) POV; for instance, a user uploaded a video with the text caption "POV: you discovered elf bar made a calendar" and sang along with the background music, which contains the lyrics "Oh, I want you so badly. It's my greatest wish." These findings demonstrate that TikTok users are actively participating in normalizing vaping as an everyday practice. This is exemplified by the platform's how-to videos and everyday vaping scenarios, which reflect a concerning trend in the platform's content.

Figure 9. Sample image clustering results from theme 5 (vape brands); 2 subclusters were formed. (A) Videos demonstrating how to use a vaping device after opening the packaging, reviewing the product, and vaping. (2) Vaping-related context and vape use in daily settings. A TikTok user uploaded a shopping haul–related video, in which she displays a new vape (Lost Mary) and declares, "TikTok, please don't ban me; I'm actually 21 this year.".

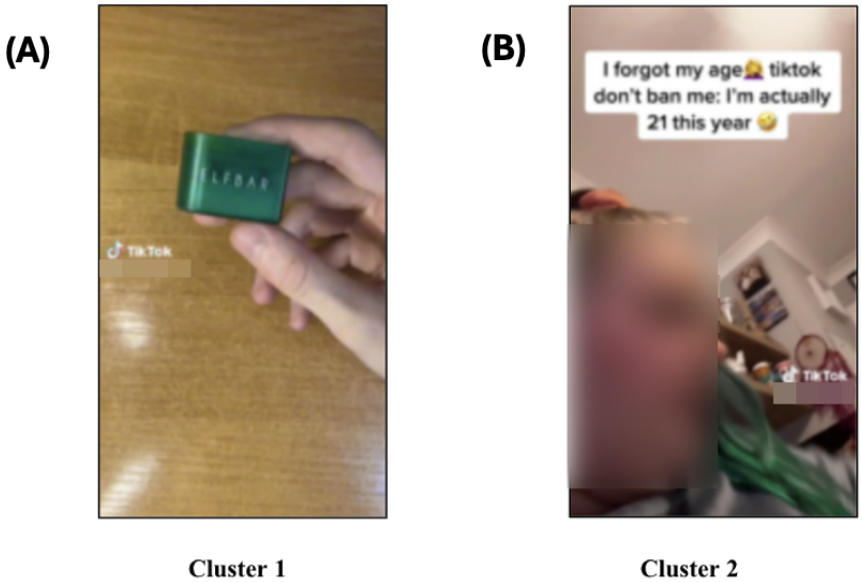

Qualitative Thematic Analysis Results

In our qualitative thematic analysis, 50 videos were randomly selected from each of the 5 themes. These videos were viewed, and the theme of individual videos was compared with the theme descriptor to validate the themes.

Of the 50 videos selected at random from theme 1 (*general vape*), 40 (80%) were vape related; the 10 (20%) videos that were not vape related included a concert in Kuala Lumpur, Malaysia (which featured a smoky environment; n=1, 10%); a public service announcement vehicle (which featured a blue blinking light; n=1, 10%); car-related videos featuring blinkers (n=3, 30%); a festive scrunchy (n=1, 10%); a disposable camera (n=2, 20%); an advertisement for JLCPCB, a Chinese company

XSL•FO
RenderX

known for manufacturing printed circuit boards (blinking circuits; n=1, 10%), and a calligraphy pen (n=1, 10%).

For theme 2 (*vaping cessation*), of the 50 videos selected at random, only 2 (4%) were not vape related (video promoting clothing: n=1, 50%; video promoting instant coffee: n=1, 50%). Of the 48 vape-related videos, 15 (31%) covered content directly related to vaping cessation, and 2 (4%) were related to quitting cigarettes (video suggesting that vaping was the best way to quit cigarettes: n=1, 50%). Of the 15 videos promoting cessation, 3 (30%) used dentists; several videos (4/15, 27%) provided tips on how to quit vaping; and others (8/15, 53%) shared information on the health and mental health benefits of quitting, such as feeling good after quitting vaping.

Of the 50 videos selected at random from theme 3 (*vape product marketing*), 13 (26%) were not vape related. The TikTok videos that were not vape related included an episode of a story (1/13, 8%), live music concerts that were filled with smoke (3/13, 23%), a girl bemoaning the loss of her boyfriend (1/13, 8%), discussions of video games (4/13, 31%), World War I Prussian rifle cartridges (1/13, 8%), a Wing Stop taste tester (who mentions a “hot box” meal; 1/13, 8%), a tattoo pen (1/13, 8%), and a calligraphy pen (1/13, 8%).

Only 1 (2%) of the 50 videos selected at random from theme 4 (*TikTok influencer*) was not vape related. Of the 49 vape-related videos, 6 (12%) conveyed an antivaping message, and 1 (2%) showed the aftermath of a vaping device exploding in a bedroom. The majority of the videos (42/49, 86%) shared images of vape pens, vape juice, and people in vape shops and presented positive images of young people vaping.

Finally, of the 50 videos selected from theme 5 (*vape brands*), 7 (14%) were not vape related. These included a video by “coolkidclaire” (1/7, 14%), boys dancing or playing football (which could have been mistaken for blinking; 3/7, 43%), a clothing store (1/7, 14%), food (1/7, 14%), and a car (another possible blinker-related error; 1/7, 14%).

Of the 250 videos reviewed, 33 (13.2%) were not vape related. However, most of these non–vape-related videos (8/33, 24%) featured a smoke-filled environment, included blinking in some form, or displayed the image of a pen. As such, most of the videos (13/33, 39%) included words or aspects of images that were similar in nature to vaping-related content and content that could easily be misclassified by a machine, highlighting the need for human interaction in this process. Moreover, only theme 2 included a variety of vaping-related images that were not related to the theme’s descriptor of *vaping cessation*. Although theme 4 (*TikTok influencer*) included some videos (7/50, 14%) providing tips on how to quit vaping and encouraging quitting, only 15 (30%) of the 50 videos from theme 2 (*vaping cessation*) conveyed this message. Again, in comparison with the other 4 themes in which most videos fit the general descriptor of the theme, this was less the case with TikTok videos related to quitting vaping. Thus, the only theme in which there was any noteworthy misclassification was in theme 2, *vaping cessation*, wherein only 15 (30%) of the 50 videos were specifically cessation related.

However, misclassification was not a failure of the model because (1) only 2 (4%) of the 50 videos in theme 2 were not directly vape related; and (2) classifying cessation-related content was challenging due to the complexity of the material, as well as the use of humor and double entendres by TikTok video creators. Many TikTok videos were characterized by a light and fun tone through the juxtaposition of ideas, words, and images, a style that was common in many of these vaping-related videos (149/250, 59.6%). Moreover, although relatively few in number (48/250, 19.2%), these vaping cessation videos are important because of the tone used by the creator, a tone that might resonate well with other TikTok users.

Finally, the nature of the misclassifications with theme 2 underscores the possibility that when humor and double entendres are an integral aspect of the message, computer-driven classification will benefit from maintaining a human-in-the-loop approach. While we can rely on machine learning algorithms to sort through thousands of images and text entries, improving the accuracy of the process requires humans to interpret images and text that are context specific and rely on socially inferred meaning.

Discussion

Principal Findings

This study is one of the first to use a combination of computational methods to analyze both text and image data from a large set of vaping-related content posted on TikTok between 2018 and 2023. Our analysis of the posts (n=10,990) identified 5 main themes: *general vape*, *vaping cessation*, *vape product marketing*, *TikTok influencer*, and *vape brands*. Notably, the majority of these themes (3/5, 60%) indirectly promoted vaping and may contribute to creating a web-based environment that normalizes e-cigarette use. Only one theme themes (1/5, 20%) conveyed antivaping messages; the theme identified as *vaping cessation* included more posts that indirectly promoted vaping and contributed to creating a web-based environment that normalizes e-cigarette use rather than supporting vaping cessation.

We found a relatively low proportion of posts (1160/10,990, 10.56%) explicitly engaged in direct marketing. It is noteworthy that when direct marketing was observed, it primarily focused on promoting various vaping brands and flavors. Our findings of the indirect promotion of vaping products on TikTok, particularly the promotion of vaping brands and flavors, align with those of the study by Vassey et al [45], which revealed a significant increase in the prevalence of e-juice flavor names and e-cigarette brand names on Instagram and TikTok. Of note, existing studies indicate that the variety of available flavors, ranging from fruity to sweet and other appealing tastes, is one of the top reasons cited by young people to experiment with e-cigarettes [92,93].

Further findings indicated that popular TikTok users or influencers primarily promoted vaping products indirectly. Influencers, often associated with youthfulness on social media, are skilled at subtly endorsing products, thereby making them more relatable than traditional advertisements. The findings

XSL•FO
RenderX

showed that influencers make vaping look like a natural part of their daily lives as they engage in activities such as gaming, partying, traveling, and so on, which may lead to indirectly normalizing the behavior. Brands partnering with influencers can effectively shape attitudes and behaviors because it does not trigger the same resistance as overt direct advertising [94]. Nonetheless, these findings align with and expand prior research documenting a notable presence of influencers' role in promoting vaping on social media platforms on behalf of tobacco brands [1]; for example, a recent study showed that 20% of the influencers exclusively posted about e-cigarette use on Instagram, while 80% of them posted about e-cigarettes and other topics [3]. Moreover, when influencers create content related to vaping, TikTok algorithms actively promote this content to a larger audience, significantly expanding its reach and impact [94].

Our study identified a substantial amount of content that violates TikTok's current policies, underscoring the need for stricter regulations in web-based advertising, particularly when the main companies use indirect methods through smaller sellers or influencers to promote their vaping products. Furthermore, the normalization of vaping through influencers or other strategies poses a significant concern for public health, especially among younger people who are more susceptible to the influences of such advertising [95].

Implications for Policy and Research

Given the findings of this study, social media platforms should consider implementing more robust measures to ensure that videos reach only their intended audiences; for example, social media platforms should (1) require age verification for users accessing any vaping content or engaging with vaping-related content, (2) require enhanced posts that highlight the potential health risks associated with vaping, and (3) prevent content that contains misleading information about the safety of vaping. Policy makers could develop regulations to enforce stricter advertising policies, explicitly prohibiting the promotion of vaping products and restricting content that indirectly promotes vaping [96,97].

Policy makers may also consider implementing advanced algorithms to detect and flag content that indirectly promotes vaping. These algorithms can identify visual patterns and keyword text associated with vaping-related content. Furthermore, policies should develop and enforce specific guidelines for influencers, including transparency requirements regarding affiliations with vaping brands and restrictions on promoting vaping to underage populations. In addition, investing in the development of prevention programs aimed at denormalizing vaping use among youth could be crucial in mitigating the spread of vaping culture and its associated risks.

Future Work and Limitations

We successfully implemented a novel approach that combines 2 machine learning methods—natural language processing and computer vision—alongside qualitative thematic analysis. Specifically, our approach leveraged computational methods, specifically k-means clustering for textual data and state-of-the-art ViT and CNN models for image data. A computational, machine learning–based methodology allowed us to analyze a large volume of textual and visual content efficiently. We uncovered key insights into the prevalence and acceptance of vaping on TikTok using these computational techniques. However, we acknowledge a few limitations. Incorporating video transcription software, which automatically transcribes spoken language into written text, could provide additional textual data for analysis in future work. This enhancement could significantly improve the text corpus for thematic analysis. In addition, due to our hashtag-based data scraping approach, we encountered limitations in extracting comprehensive vaping-related information; for instance, our objective was to scrape content from posts containing the hashtag *#blinker*, which denotes a visual indicator akin to the light-emitting diode lights found on a vaping device, including an e-cigarette or a vape pen. However, our scraped results included irrelevant content, such as a public service announcement vehicle with a blue blinking light and a calligraphy pen. The algorithmic flow of TikTok necessitates a human-in-the-loop approach, which we implemented successfully in this study.

Future work could benefit from incorporating video transcription software, further enriching textual data for even more comprehensive thematic analysis. Although our classification of vaping-related messages was successful (average F_1-score=0.97), our classification of vaping cessation messages presents opportunities for further improvement. Specifically, cessation-related messages are difficult for machines to understand because they contain subtle cues (eg, humor). These findings indicate that humans are vital to the classification process. Future work should focus on developing methods to enhance a machine's ability to interpret context-specific images and text and extract socially inferred meaning in complex areas such as humor and vaping cessation. Furthermore, exploring the integration of more advanced natural language processing models, such as Bidirectional Encoder Representations from Transformers [98], could further enhance the depth and accuracy of our text analysis. These models offer contextual understanding, capturing intricate relationships within text that may provide deeper insights into the themes and sentiments expressed in vaping-related content on TikTok.

Conclusions

This study conducted a thematic analysis of vaping-related content on TikTok, using a hybrid computational and qualitative approach to analyze textual and visual data, respectively, to ensure precision and reduce bias. We aimed to discern the patterns and themes within vaping-related content to understand the extent to which vaping is normalized in the context of social media discourse. We conducted a multifaceted investigation into user-generated content on social media platforms related to e-cigarette (vape) discussions, focusing specifically on TikTok. We found that vaping-related content on TikTok is marked by distinct themes, including those portraying vaping as a common activity in daily life, discussions on vaping cessation, TikTok influencers incorporating vaping into their content, and specific discussions related to vape brands and product marketing. These themes provided valuable insights into the normalization of vaping on the platform. Images within

XSL•FO
RenderX

each clustered theme highlighted the various facets of vaping, such as active vaping practices, conversations unrelated to vaping but with a vaping device visible in the video, and vaping as an integral part of daily life.

Consistent with these findings, the qualitative analysis revealed that our machine learning approach correctly recognized vape-related content by successfully parsing these data into smaller meaningful units (ie, the 5 themes). The cessation-related TikTok videos merit additional comment and research. They are few in number (1272/14,002, 9.08%), but this is unsurprising because they are created by TikTok users. However, their impact may be far greater than officially created content because the user-generated approach of cessation messaging might be far more appealing to other TikTok users than cessation messages created by professionals.

Overall, our computational methods with a qualitative human-in-the-loop approach allowed us to gain a deeper understanding of how vaping is represented and normalized on TikTok. Our study provided a broader perspective, emphasizing the normalization of vaping practices rather than focusing solely on the textual or visual content of prior research. Moving forward, our findings underscore the importance of ongoing surveillance of vaping-related content on social media platforms, highlighting the need for targeted interventions to mitigate the normalization of harmful behaviors among youth populations.

Acknowledgments

This research was supported by the National Institutes of Health (R01MD017280; AL and AVW).

Data Availability

The data sets and code generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of TikTok hashtags. We scraped the data using 50 hashtags.
[DOCX File , 13 KB - jmir_v26i1e55591_app1.docx]

Multimedia Appendix 2
Text clustering result. We identified 15 clusters using the k-means clustering method.
[DOCX File , 17 KB - jmir_v26i1e55591_app2.docx]

References

1. Sun T, Lim CC, Chung J, Cheng B, Davidson L, Tisdale C, et al. Vaping on TikTok: a systematic thematic analysis. Tob Control 2023 Mar 26;32(2):251-254. [doi: 10.1136/tobaccocontrol-2021-056619] [Medline: 34312317]
2. Morales M, Fahrion A, Watkins SL. #NicotineAddictionCheck: puff bar culture, addiction apathy, and promotion of e-cigarettes on TikTok. Int J Environ Res Public Health 2022 Feb 05;19(3):1820 [FREE Full text] [doi: 10.3390/ijerph19031820] [Medline: 35162846]
3. Vogel EA, Ramo DE, Rubinstein ML, Delucchi KL, Darrow SM, Costello C, et al. Effects of social media on adolescents' willingness and intention to use e-cigarettes: an experimental investigation. Nicotine Tob Res 2021 Mar 19;23(4):694-701 [FREE Full text] [doi: 10.1093/ntr/ntaa003] [Medline: 31912147]
4. Jeffares Y. From smoking to vaping : vapers' perspectives of policy related facilitators and barriers. Massey University. 2020. URL: https://mro.massey.ac.nz/items/73f9ceec-4781-437b-9914-3a226b047577 [accessed 2024-08-29]
5. Vogel EA, Hashemi R, Ramo DE, Darrow SM, Costello C, Prochaska JJ. Adolescents' perceptions of nicotine vaping-related social media content. Psychol Pop Media 2024;13(2):240-248. [doi: 10.1037/ppm0000452]
6. Collins L, Glasser AM, Abudayyeh H, Pearson JL, Villanti AC. E-cigarette marketing and communication: how e-cigarette companies market e-cigarettes and the public engages with e-cigarette information. Nicotine Tob Res 2019 Jan 01;21(1):14-24 [FREE Full text] [doi: 10.1093/ntr/ntx284] [Medline: 29315420]
7. Verhaegen A, Van Gaal L. Vaping and cardiovascular health: the case for health policy action. Curr Cardiovasc Risk Rep 2019 Nov 19;13:37. [doi: 10.1007/s12170-019-0634-9]
8. Lewis SC, Zamith R, Hermida A. Content analysis in an era of big data: a hybrid approach to computational and manual methods. J Broadcast Electron Media 2013 Jan;57(1):34-52. [doi: 10.1080/08838151.2012.761702]
9. Su LY, Cacciatore MA, Liang X, Brossard D, Scheufele DA, Xenos MA. Analyzing public sentiments online: combining human- and computer-based content analysis. Inf Commun Soc 2016 May 09;20(3):406-427. [doi: 10.1080/1369118x.2016.1182197]

XSL•FO
RenderX

10. Richards C, Glasgow L, Bidaisee S, Ramdass P, Oladimeji T, Loukas M, et al. White paper on vaping: electronic cigarettes use – a case for restrictive policies in Grenada, West Indies. Int Public Health J 2020;12(1):19-32 [FREE Full text]
11. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554. [doi: 10.1177/0038038509103208]
12. McEvoy R, Ballini L, Maltoni S, O'Donnell CA, Mair FS, Macfarlane A. A qualitative systematic review of studies using the normalization process theory to research implementation processes. Implement Sci 2014 Jan 02;9:2 [FREE Full text] [doi: 10.1186/1748-5908-9-2] [Medline: 24383661]
13. Linkenbach JW, Lubbers DT, Brandon JM, Ooms JD, Langenberg AJ, Kilmer JR. Assessing adolescent vaping norms and perceptions in a statewide multi-community project. Subst Use Misuse 2023 Jan 12;58(3):428-433. [doi: 10.1080/10826084.2023.2165413] [Medline: 36633293]
14. Lozano P, Arillo-Santillán E, Barrientos-Gutíerrez I, Reynales Shigematsu LM, Thrasher JF. E-cigarette social norms and risk perceptions among susceptible adolescents in a country that bans e-cigarettes. Health Educ Behav 2019 Apr 04;46(2):275-285 [FREE Full text] [doi: 10.1177/1090198118818239] [Medline: 30606077]
15. Carpenter CM, Wayne GF, Pauly JL, Koh HK, Connolly GN. New cigarette brands with flavors that appeal to youth: tobacco marketing strategies. Health Aff (Millwood) 2005 Nov;24(6):1601-1610. [doi: 10.1377/hlthaff.24.6.1601] [Medline: 16284034]
16. Payne JD, Orellana-Barrios M, Medrano-Juarez R, Buscemi D, Nugent K. Electronic cigarettes in the media. Proc (Bayl Univ Med Cent) 2016 Jul 11;29(3):280-283 [FREE Full text] [doi: 10.1080/08998280.2016.11929436] [Medline: 27365871]
17. Groom AL, Vu TH, Landry RL, Kesh A, Hart JL, Walker KL, et al. The influence of friends on teen vaping: a mixed-methods approach. Int J Environ Res Public Health 2021 Jun 24;18(13):6784 [FREE Full text] [doi: 10.3390/ijerph18136784] [Medline: 34202600]
18. Golan R, Muthigi A, Ghomeshi A, White J, Saltzman RG, Diaz P, et al. Misconceptions of vaping among young adults. Cureus 2023 Apr;15(4):e38202 [FREE Full text] [doi: 10.7759/cureus.38202] [Medline: 37252604]
19. Soneji S, Barrington-Trimis JL, Wills TA, Leventhal AM, Unger JB, Gibson LA, et al. Association between initial use of e-cigarettes and subsequent cigarette smoking among adolescents and young adults: a systematic review and meta-analysis. JAMA Pediatr 2017 Aug 01;171(8):788-797 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.1488] [Medline: 28654986]
20. Watkins SL, Glantz SA, Chaffee BW. Association of noncigarette tobacco product use with future cigarette smoking among youth in the population assessment of tobacco and health (PATH) study, 2013-2015. JAMA Pediatr 2018 Feb 01;172(2):181-187 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.4173] [Medline: 29297010]
21. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q 2014 Mar 06;92(1):7-33 [FREE Full text] [doi: 10.1111/1468-0009.12038] [Medline: 24597553]
22. Giustini D, Ali SM, Fraser M, Kamel Boulos MN. Effective uses of social media in public health and medicine: a systematic review of systematic reviews. Online J Public Health Inform 2018 Sep 21;10(2):e215 [FREE Full text] [doi: 10.5210/ojphi.v10i2.8270] [Medline: 30349633]
23. MacKinnon KR, Kia H, Lacombe-Duncan A. Examining TikTok's potential for community-engaged digital knowledge mobilization with equity-seeking groups. J Med Internet Res 2021 Dec 09;23(12):e30315 [FREE Full text] [doi: 10.2196/30315] [Medline: 34889739]
24. Vassey J, Galimov A, Kennedy CJ, Vogel EA, Unger JB. Frequency of social media use and exposure to tobacco or nicotine-related content in association with e-cigarette use among youth: a cross-sectional and longitudinal survey analysis. Prev Med Rep 2022 Dec;30:102055 [FREE Full text] [doi: 10.1016/j.pmedr.2022.102055] [Medline: 36531097]
25. McCausland K, Maycock B, Leaver T, Jancey J. The messages presented in electronic cigarette-related social media promotions and discussion: scoping review. J Med Internet Res 2019 Feb 05;21(2):e11953 [FREE Full text] [doi: 10.2196/11953] [Medline: 30720440]
26. Pokhrel P, Herzog TA, Fagan P, Unger JB, Stacy AW. E-cigarette advertising exposure, explicit and implicit harm perceptions, and e-cigarette use susceptibility among nonsmoking young adults. Nicotine Tob Res 2019 Jan 01;21(1):127-131 [FREE Full text] [doi: 10.1093/ntr/nty030] [Medline: 29444275]
27. Wang L, Chen J, Ho SY, Leung LT, Wang MP, Lam TH. Exposure to e-cigarette advertising, attitudes, and use susceptibility in adolescents who had never used e-cigarettes or cigarettes. BMC Public Health 2020 Sep 04;20(1):1349 [FREE Full text] [doi: 10.1186/s12889-020-09422-w] [Medline: 32887586]
28. Cao DJ, Aldy K, Hsu S, McGetrick M, Verbeck G, De Silva I, et al. Review of health consequences of electronic cigarettes and the outbreak of electronic cigarette, or vaping, product use-associated lung injury. J Med Toxicol 2020 Jul 16;16(3):295-310 [FREE Full text] [doi: 10.1007/s13181-020-00772-w] [Medline: 32301069]
29. Bataineh BS, Wilkinson AV, Sumbe A, Clendennen SL, Chen B, Messiah SE, et al. The association between tobacco and cannabis use and the age of onset of depression and anxiety symptoms: among adolescents and young adults. Nicotine Tob Res 2023 Jul 14;25(8):1455-1464 [FREE Full text] [doi: 10.1093/ntr/ntad058] [Medline: 37042355]
30. Sun T, Lim CC, Stjepanović D, Leung J, Connor JP, Gartner C, et al. Has increased youth e-cigarette use in the USA, between 2014 and 2020, changed conventional smoking behaviors, future intentions to smoke and perceived smoking harms? Addict Behav 2021 Dec;123:107073. [doi: 10.1016/j.addbeh.2021.107073] [Medline: 34364109]

XSL•FO
RenderX

31. Community guidelines. TikTok. URL: https://www.tiktok.com/community-guidelines?lang=en [accessed 2024-03-20]
32. Rutherford BN, Sun T, Lim CC, Chung J, Cheng B, Davidson L, et al. Changes in viewer engagement and accessibility of popular vaping videos on TikTok: a 12-month prospective study. Int J Environ Res Public Health 2022 Jan 20;19(3):1141 [FREE Full text] [doi: 10.3390/ijerph19031141] [Medline: 35162170]
33. Bornstein RF. Exposure and affect: overview and meta-analysis of research, 1968–1987. Psychol Bull 1989 Sep;106(2):265-289. [doi: 10.1037/0033-2909.106.2.265]
34. Kim H, Davis AH, Dohack JL, Clark PI. E-cigarettes use behavior and experience of adults: qualitative research findings to inform e-cigarette use measure development. Nicotine Tob Res 2017 Feb 13;19(2):190-196 [FREE Full text] [doi: 10.1093/ntr/ntw175] [Medline: 27613944]
35. Purushothaman V, McMann T, Nali M, Li Z, Cuomo R, Mackey TK. Content analysis of nicotine poisoning (Nic Sick) videos on TikTok: retrospective observational infodemiology study. J Med Internet Res 2022 Mar 30;24(3):e34050 [FREE Full text] [doi: 10.2196/34050] [Medline: 35353056]
36. Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, et al. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res 2015 Oct 27;17(10):e243 [FREE Full text] [doi: 10.2196/jmir.4969] [Medline: 26508089]
37. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. J Med Internet Res 2016 Feb 26;18(2):e41 [FREE Full text] [doi: 10.2196/jmir.4738] [Medline: 26920122]
38. Zamith R, Lewis SC. Content analysis and the algorithmic coder: what computational social science means for traditional modes of media analysis. Ann Am Acad Pol Soc Sci 2015 Apr 09;659(1):307-318. [doi: 10.1177/0002716215570576]
39. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearb Med Inform 2019 Aug;28(1):208-217 [FREE Full text] [doi: 10.1055/s-0039-1677918] [Medline: 31419834]
40. Elbattah M, Arnaud E, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. In: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies. 2021 Presented at: BIOSTEC 2021; February 11-13, 2021; Virtual Event. [doi: 10.5220/0010414508250832]
41. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 04;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
42. Sun S, Liu Z, Zhai Y, Wang F. COVID-19 vaccines on TikTok: a big-data analysis of entangled discourses. Int J Environ Res Public Health 2022 Oct 14;19(20):13287 [FREE Full text] [doi: 10.3390/ijerph192013287] [Medline: 36293868]
43. Bharti U, Bajaj D, Batra H, Lalit S, Lalit S, Gangwani A. Medbot: conversational artificial intelligence powered chatbot for delivering tele-health after COVID-19. In: Proceedings of the 5th International Conference on Communication and Electronics Systems. 2020 Presented at: ICCES 2020; June 10-12, 2020; Coimbatore, India. [doi: 10.1109/icces48766.2020.9137944]
44. Murthy D, Ouellette RR, Anand T, Radhakrishnan S, Mohan NC, Lee J, et al. Using computer vision to detect e-cigarette content in TikTok videos. Nicotine Tob Res 2024 Feb 15;26(Supplement_1):S36-S42. [doi: 10.1093/ntr/ntad184] [Medline: 38366342]
45. Vassey J, Kennedy CJ, Herbert Chang HC, Smith AS, Unger JB. Scalable surveillance of e-cigarette products on Instagram and TikTok using computer vision. Nicotine Tob Res 2024 Apr 22;26(5):552-560 [FREE Full text] [doi: 10.1093/ntr/ntad224] [Medline: 37947283]
46. Holm EA, Cohn R, Gao N, Kitahara AR, Matson TP, Lei B, et al. Overview: computer vision and machine learning for microstructural characterization and analysis. Metall Mater Trans A 2020 Sep 29;51(12):5985-5999. [doi: 10.1007/S11661-020-06008-4]
47. Szeliski R. Computer Vision: Algorithms and Applications. London, UK: Springer International Publishing; 2010.
48. Xu S, Wang J, Shou W, Ngo T, Sadick AM, Wang X. Computer vision techniques in construction: a critical review. Arch Comput Methods Eng 2020 Oct 19;28(5):3383-3397. [doi: 10.1007/s11831-020-09504-3]
49. Forsyth DA, Ponce J. Computer Vision: A Modern Approach International Edition. London, UK: Pearson Education; 2002.
50. Han K, Vedaldi A, Zisserman A. Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019 Presented at: ICCV 2019; October 27-November 02, 2019; Seoul, Korea (South). [doi: 10.1109/iccv.2019.00849]
51. Datta R, Joshi D, Li J, Wang JZ. Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 2008 May 08;40(2):1-60. [doi: 10.1145/1348246.1348248]
52. Blinn CE. Increasing the precision of forest area estimates through improved sampling for nearest neighbor satellite. Virginia Polytechnic Institute and State University. 2005 Jul 29. URL: https://vtechworks.lib.vt.edu/server/api/core/bitstreams/eeb2f9b4-ffbc-41a3-857a-5cd8a040bf0d/content [accessed 2024-08-21]
53. Liu C, Li M, Zhang Y, Han S, Zhu Y. An enhanced rock mineral recognition method integrating a deep learning model and clustering algorithm. Minerals 2019 Aug 26;9(9):516. [doi: 10.3390/min9090516]

XSL•FO
RenderX

54. Galassi A, Lippi M, Torroni P. Attention in natural language processing. IEEE Trans Neural Netw Learn Syst 2021 Oct;32(10):4291-4308. [doi: 10.1109/tnnls.2020.3019893]
55. Cheon M, Yoon SJ, Kang B, Lee J. Perceptual image quality assessment with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021 Presented at: CVPRW 2021; June 19-25, 2021; Nashville, TN. [doi: 10.1109/cvprw53098.2021.00054]
56. Murthy VS, Vamsidhar E, Kumar JN, Rao PS. Content based image retrieval using hierarchical and K-means clustering techniques. Int J Eng Sci Technol 2010;2(3):209-212.
57. Murthy D, Lee J, Dashtian H, Kong G. Influence of user profile attributes on e-cigarette-related searches on YouTube: machine learning clustering and classification. JMIR Infodemiology 2023 Apr 12;3:e42218 [FREE Full text] [doi: 10.2196/42218] [Medline: 37124246]
58. Ketonen V, Malik A. Characterizing vaping posts on Instagram by using unsupervised machine learning. Int J Med Inform 2020 Sep;141:104223 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104223] [Medline: 32623330]
59. Vassey J, Metayer C, Kennedy CJ, Whitehead TP. #Vape: measuring e-cigarette influence on Instagram with deep learning and text analysis. Front Commun (Lausanne) 2020 Jan 22;4:75 [FREE Full text] [doi: 10.3389/fcomm.2019.00075] [Medline: 35233388]
60. Setiawan W, Purnama A. Tobacco leaf images clustering using DarkNet19 and K-means. In: Proceedings of the 6th Information Technology International Seminar. 2020 Presented at: ITIS 2020; October 14-16, 2020; Surabaya, Indonesia. [doi: 10.1109/itis50118.2020.9321035]
61. Lavack AM. De-normalization of tobacco in Canada. Soc Market Q 2016 Aug 01;5(3):82-85. [doi: 10.1080/15245004.1999.9961068]
62. Jancey J, Leaver T, Wolf K, Freeman B, Chai K, Bialous S, et al. Promotion of e-cigarettes on TikTok and regulatory considerations. Int J Environ Res Public Health 2023 May 09;20(10):5761 [FREE Full text] [doi: 10.3390/ijerph20105761] [Medline: 37239490]
63. digitalmethodsinitiative / zeeschuimer. GitHub. URL: https://github.com/digitalmethodsinitiative/zeeschuimer [accessed 2023-06-30]
64. Navarro C, Peres-Neto L. “Hair for freedom” movement in Iran: interreligious dialogue in social media activism? Religions 2023 May 04;14(5):602. [doi: 10.3390/rel14050602]
65. Rozaki E. Reading between the likes: the influence of BookTok on reading culture. Utrecht University. 2023. URL: https://studenttheses.uu.nl/handle/20.500.12932/43888 [accessed 2024-08-21]
66. dfreelon / pyktok. GitHub. URL: https://github.com/dfreelon/pyktok [accessed 2023-06-30]
67. Natural language toolkit. NLTK Project. URL: https://www.nltk.org/ [accessed 2023-07-01]
68. langdetect 1.0.9. PyPI. URL: https://pypi.org/project/langdetect/ [accessed 2023-07-01]
69. Jones KS. A statistical interpretation of term specificity and its application in retrieval. In: Lewis HR, editor. Ideas That Created the Future: Classic Papers of Computer Science. Cambridge, MA: The MIT Press; 2021.
70. Et-taleby A, Boussetta M, Benslimane M. Faults detection for photovoltaic field based on K-means, elbow, and average silhouette techniques through the segmentation of a thermal image. Int J Photoenergy 2020 Dec 4;2020:1-7. [doi: 10.1155/2020/6617597]
71. idealo / imagededup. GitHub. URL: https://github.com/idealo/imagededup [accessed 2023-10-01]
72. Carlini N, Jagielski M, Zhang C, Papernot N, Terzis A, Tramer F. The privacy onion effect: memorization is relative. arXiv 2024 Preprint posted online June 21, 2022 [FREE Full text]
73. Rao VN, Korolova A. Discrimination through image selection by job advertisers on Facebook. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023 Presented at: FAccT '23; June 12-15, 2023; Chicago, IL. [doi: 10.1145/3593013.3594115]
74. Tao Z, Liu H, Fu H, Fu Y. Image cosegmentation via saliency-guided constrained clustering with cosine similarity. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017 Presented at: AAAI-17; February 4-9, 2017; San Francisco, CA. [doi: 10.1609/aaai.v31i1.11203]
75. Alelyani S, Tang J, Liu H. Feature selection for clustering: a review. In: Aggarwal CC, Reddy CK, editors. Data Clustering. Boca Raton, FL: CRC Press; 2014.
76. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput 2017;29(9):2352-2449. [doi: 10.1162/neco_a_00990]
77. Pandya B, Cosma G, Alani AA, Taherkhani A, Bharadi V, McGinnity TM. Fingerprint classification using a deep convolutional neural network. In: Proceedings of the 4th International Conference on Information Management (ICIM). 2018 Presented at: ICIM 2018; May 25-27, 2018; Oxford, UK. [doi: 10.1109/infoman.2018.8392815]
78. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. arXiv 2024 Preprint posted online December 03, 2019. [doi: 10.48550/arXiv.1912.01703]
79. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2024 Preprint posted online September 4, 2014 [FREE Full text]

XSL•FO
RenderX

80. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: CVPR 2016; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
81. Shaha M, Pawar M. Transfer learning for image classification. In: Proceedings of the Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). 2018 Presented at: ICECA 2018; March 29-31, 2018; Coimbatore, India. [doi: 10.1109/iceca.2018.8474802]
82. Yuan J, Fan Y, Lv X, Chen C, Li D, Hong Y, et al. Research on the practical classification and privacy protection of CT images of parotid tumors based on ResNet50 model. J Phys Conf Ser 2020 Jun 01;1576:012040. [doi: 10.1088/1742-6596/1576/1/012040]
83. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv 2024 Preprint posted online June 12, 2017.
84. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv 2024 Preprint posted online October 22, 2020 [FREE Full text]
85. Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: a literature review. Appl Sci 2023 Apr 28;13(9):5521. [doi: 10.3390/app13095521]
86. Park S, Kim G, Oh Y, Seo JB, Lee SM, Kim JH, et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. Med Image Anal 2022 Jan;75:102299 [FREE Full text] [doi: 10.1016/j.media.2021.102299] [Medline: 34814058]
87. Tsai PC, Lee TH, Kuo KC, Su FY, Lee TL, Marostica E, et al. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. Nat Commun 2023 Apr 13;14(1):2102 [FREE Full text] [doi: 10.1038/s41467-023-37179-4] [Medline: 37055393]
88. Transformers. Hugging Face. URL: https://huggingface.co/docs/transformers/index [accessed 2023-10-05]
89. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv 2024 Preprint posted online October 09, 2019 [FREE Full text]
90. Yuan J, Barmpoutis P, Stathaki T. Effectiveness of vision transformer for fast and accurate single-stage pedestrian detection. In: Proceedings of the 36th Conference on Neural Information Processing Systems. 2022 Presented at: NeurIPS 2022; November 28-December 9, 2022; New Orleans, LA.
91. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2024 Preprint posted online December 22, 2014 [FREE Full text] [doi: 10.5860/choice.45-0602]
92. Kong G, Morean ME, Cavallo DA, Camenga DR, Krishnan-Sarin S. Reasons for electronic cigarette experimentation and discontinuation among adolescents and young adults. Nicotine Tob Res 2015 Jul;17(7):847-854 [FREE Full text] [doi: 10.1093/ntr/ntu257] [Medline: 25481917]
93. Zare S, Nemati M, Zheng Y. A systematic review of consumer preference for e-cigarette attributes: flavor, nicotine strength, and type. PLoS One 2018 Mar 15;13(3):e0194145 [FREE Full text] [doi: 10.1371/journal.pone.0194145] [Medline: 29543907]
94. Glasman LR, Albarracín D. Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation. Psychol Bull 2006 Sep;132(5):778-822 [FREE Full text] [doi: 10.1037/0033-2909.132.5.778] [Medline: 16910754]
95. Vassey J, Valente T, Barker J, Stanton C, Li D, Laestadius L, et al. E-cigarette brands and social media influencers on Instagram: a social network analysis. Tob Control 2023 Aug 07;32(e2):e184-e191 [FREE Full text] [doi: 10.1136/tobaccocontrol-2021-057053] [Medline: 35131947]
96. Snell LM, Nicksic N, Panteli D, Burke S, Eissenberg T, Fattore G, et al. Emerging electronic cigarette policies in European member states, Canada, and the United States. Health Policy 2021 Apr;125(4):425-435 [FREE Full text] [doi: 10.1016/j.healthpol.2021.02.003] [Medline: 33663799]
97. O'Connell M, Kephart L. Local and state policy action taken in the United States to address the emergence of e-cigarettes and vaping: a scoping review of literature. Health Promot Pract 2022 Jan 07;23(1):51-63. [doi: 10.1177/1524839920963691] [Medline: 33164552]
98. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv 2024 Preprint posted online October 11, 2018 [FREE Full text]

Abbreviations

API: application programming interface
CNN: convolutional neural network
POV: point of view
ResNet50: residual network
TF-IDF: term frequency–inverse document frequency
VGG16: Visual Geometry Group with 16 layers
ViT: vision transformer

XSL•FO
RenderX

Edited by A Mavragani; submitted 18.12.23; peer-reviewed by J Vassey, H Shih, M Elbattah, N Baharom; comments to author 29.02.24; revised version received 07.04.24; accepted 20.05.24; published 11.09.24.

Please cite as:
Jung S, Murthy D, Bateineh BS, Loukas A, Wilkinson AV
The Normalization of Vaping on TikTok Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis: Mixed Methods Study
J Med Internet Res 2024;26:e55591
URL: https://www.jmir.org/2024/1/e55591
doi:10.2196/55591
PMID:

©Sungwon Jung, Dhiraj Murthy, Bara S Bateineh, Alexandra Loukas, Anna V Wilkinson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Uncovering the Complexity of Perinatal Polysubstance Use Disclosure Patterns on X: Mixed Methods Study

Dezhi Wu[1], PhD, PMP; Hannah Shead[2], MS; Yang Ren[3], MS; Phyllis Raynor[4], PhD; Youyou Tao[5], PhD; Harvey Villanueva[3]; Peiyin Hung[6], PhD; Xiaoming Li[6], PhD; Robert G Brookshire[1], PhD; Kacey Eichelberger[7,8], MD; Constance Guille[9], MD; Alain H Litwin[7,8], MD; Bankole Olatosi[6], PhD

[1]Department of Integrated Information Technology, University of South Carolina, Columbia, SC, United States
[2]Department of Mathematics, Augusta University, Augusta, GA, United States
[3]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, United States
[4]College of Nursing, University of South Carolina, Columbia, SC, United States
[5]Department of Information Systems and Business Analytics, Loyola Marymount University, Los Angeles, CA, United States
[6]Arnold School of Public Health, University of South Carolina, Columbia, SC, United States
[7]School of Medicine Greenville, University of South Carolina, Greenville, SC, United States
[8]Prisma Health, Greenville, SC, United States
[9]College of Medicine, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Dezhi Wu, PhD, PMP
Department of Integrated Information Technology
University of South Carolina
550 Assembly Street
Columbia, SC, 29208
United States
Phone: 1 8037774691
Email: dezhiwu@cec.sc.edu

Abstract

Background: According to the *Morbidity and Mortality Weekly Report*, polysubstance use among pregnant women is prevalent, with 38.2% of those who consume alcohol also engaging in the use of one or more additional substances. However, the underlying mechanisms, contexts, and experiences of polysubstance use are unclear. Organic information is abundant on social media such as X (formerly Twitter). Traditional quantitative and qualitative methods, as well as natural language processing techniques, can be jointly used to derive insights into public opinions, sentiments, and clinical and public health policy implications.

Objective: Based on perinatal polysubstance use (PPU) data that we extracted on X from May 1, 2019, to October 31, 2021, we proposed two primary research questions: (1) What is the overall trend and sentiment of PPU discussions on X? (2) Are there any distinct patterns in the discussion trends of PPU-related tweets? If so, what are the implications for perinatal care and associated public health policies?

Methods: We used X's application programming interface to extract >6 million raw tweets worldwide containing ≥2 prenatal health- and substance-related keywords provided by our clinical team. After removing all non–English-language tweets, non-US tweets, and US tweets without disclosed geolocations, we obtained 4848 PPU-related US tweets. We then evaluated them using a mixed methods approach. The quantitative analysis applied frequency, trend analysis, and several natural language processing techniques such as sentiment analysis to derive statistics to preview the corpus. To further understand semantics and clinical insights among these tweets, we conducted an in-depth thematic content analysis with a random sample of 500 PPU-related tweets with a satisfying κ score of 0.7748 for intercoder reliability.

Results: Our quantitative analysis indicates the overall trends, bigram and trigram patterns, and negative sentiments were more dominant in PPU tweets (2490/4848, 51.36%) than in the non-PPU sample (1323/4848, 27.29%). Paired polysubstance use (4134/4848, 85.27%) was the most common, with the combination *alcohol and drugs* identified as the most mentioned. From the qualitative analysis, we identified 3 main themes: *nonsubstance*, *single substance*, and *polysubstance*, and 4 subthemes to contextualize the rationale of underlying PPU behaviors: *lifestyle*, *perceptions of others' drug use*, *legal implications*, and *public health*.

XSL•FO
RenderX

Conclusions: This study identified underexplored, emerging, and important topics related to perinatal PPU, with significant stigmas and legal ramifications discussed on X. Overall, public sentiments on PPU were mixed, encompassing negative (2490/4848, 51.36%), positive (1884/4848, 38.86%), and neutral (474/4848, 9.78%) sentiments. The leading substances in PPU were *alcohol and drugs*, and the normalization of PPU discussed on X is becoming more prevalent. Thus, this study provides valuable insights to further understand the complexity of PPU and its implications for public health practitioners and policy makers to provide proper access and support to individuals with PPU.

(J Med Internet Res 2024;26:e53171) doi:10.2196/53171

KEYWORDS

polysubstance use; prenatal care; perinatal care; pregnant care; social media; Twitter; sentiment analysis

Introduction

Background

There is a growing concern regarding the rising number of pregnant women who use harmful substances [1]. According to the National Survey on Drug Use and Health, 8.5% of pregnant women in the United States reported illicit drug use, 7.1% reported marijuana use, and 11.5% consumed alcohol in the previous month [2]. These figures represent a significant increase since 2015, with a 4.7% rise in illicit drug use, a 3.4% increase in marijuana use, and a 9.3% surge in alcohol use [3]. It is even more alarming that the incidence of polysubstance use disorders (PUDs) increases during the perinatal period, with pregnant women consuming more than one harmful substance. The *Morbidity and Mortality Weekly Report* highlights that 38.2% of pregnant individuals who consumed alcohol also reported using other substances [1]. In addition, nearly half of pregnant individuals who reported opioid use also disclosed alcohol consumption [4]. Thus, perinatal polysubstance use (PPU) poses severe risks to both mothers and babies, resulting in a significant public health issue and adverse effects for both mother and baby health at large.

Clinically, use of 1 substance during pregnancy can cause adverse health outcomes; for example, pregnant individuals who use tobacco or alcohol can already cause premature birth or low birth weight, respectively. Worse, the combined use substantially increases the risk of negative fetal outcomes [5]. Similarly, the co-use of opioids and alcohol during the perinatal period can lead to neonatal withdrawal syndrome, which manifests in infants as tremors, fever, and seizures [6], leading to suboptimal pregnancy outcomes [7]. Recent trends show an alarming increase in harmful PPU associated with severe fetal health effects such as birth defects, delayed fetal development, and fetal demise [6,8]. These persistent maternal and fetal health challenges underscore the urgent need for public health policies, targeted interventions, and treatments designed to reduce and eliminate substance use during the pregnancy period and beyond [1,8].

Previous studies exploring substance use during pregnancy have primarily used interview data or survey data collected from national health samples [1,9-11]. However, a significant limitation of these traditional approaches is that pregnant women might withhold information due to feelings of shame or guilt, especially if they fear judgment from interviewers. For example, Paris et al [11] delved into interviews to understand why pregnant women conceal their substance use and what eventually prompts them to seek help. They found that shame, guilt, and societal stigma are primary factors leading women to hide their substance use from their providers when seeking prenatal care [11]. The stigma associated with substance use during pregnancy might prevent some participants from fully disclosing their reasons for continued drug use, which could exacerbate feelings of shame and both internal and external stigma [11]. These stigmas can reinforce secretive substance use patterns, delaying help-seeking behavior and potentially worsening health outcomes for both mother and child. In addition, the fear of facing critical charges may compel women to remain silent [12]. In this study context, compared to traditional data collected from interviews or surveys, social media could provide a more unfiltered platform for public discussions as individuals may choose to anonymously share their thoughts and concerns, potentially providing richer and more authentic data.

The widespread use of social media has transformed platforms such as X, formerly known as Twitter, into rich repositories of organic data that are not typically accessible through clinical trials, surveys, or interviews in the health care sector. As a result, these social media platforms have begun to function as public health surveillance tools, offering a new and vast source of big data for innovative research [13-15]. This shift has significantly enhanced our understanding of health issues and public opinions [13]. Social media data not only are massive but also offer the advantage of reducing biases that are often present in traditional interview-based studies [16]. Analyzing public opinions on social media has proven valuable to clinicians, public health officials, and policy makers, helping gauge the accuracy of health messages being disseminated and understand public perceptions of these messages, thereby informing more effective and relevant health communication strategies [14,15].

Extensive studies have explored the intersection of substance use disorder (SUD) and social media. For instance, research has identified a trend of normalization in tweets related to hookah use [17]. Tobacco-related content on social media propagates rapidly through social networks, especially when shared by popular users with large numbers of followers [18]. Furthermore, discussions related to health risks and warnings associated with nicotine products have been prominently featured on platforms such as X [19]. More specifically, social media has been proven to be an invaluable resource for gathering insights on substance use among pregnant individuals, providing a rich vein of user-generated content for further studies [20].

Despite the prevalence of social media for studying SUD, PPU patterns on social media are largely unexplored. This study

XSL•FO
RenderX

sought to address this significant research gap by exploring and identifying PPU patterns among pregnant individuals to inform public health policies and clinical implications for perinatal care. Specifically, we planned to collect data on PPU from X, a social media platform that health care researchers constantly use for health-related data extraction and exploring real-time public health surveillance, different diseases, and web-based communities. In this study, we focused on identifying the most commonly discussed types of polysubstance use in perinatal care, and initially understanding what combinations of substances are more prevalent, why, and their public sentiments.

To achieve this, we began with a systematic investigation of PPU by analyzing tweets from X. We then aimed to uncover emerging topics and trends surrounding PPU in pregnant individuals and analyze sentiments. We collected a sample of 4848 US-based PPU tweets on X and used natural language processing (NLP) for frequency, trend, bigram, trigram, and sentiment analyses. Using another sample as a control group with non-PPU tweets during the same time frame, we compared the PPU and non-PPU groups and identified significant differences, allowing us to assess the accuracy of machine learning (ML) methods in sentiment analysis. To explore more in-depth themes and subthemes discussed in the PPU tweets, we applied qualitative content analysis using a random sample of 500 PPU tweets and identified multiple PPU themes and subthemes. As one of the first studies of its kind, we expect that our findings will offer valuable and timely insights for enhancing maternal health– and substance-related public health policies and PPU interventions more customized for the perinatal population.

Objectives

In this study, we aimed to address the following two primary research questions (RQs):

1. What is the overall trend and sentiment of PPU discussions on X? (RQ 1)
2. Are there any distinct patterns in the discussion trends of PPU-related tweets? If so, what are the implications for perinatal care and associated public health policies? (RQ 2)

Methods

Data

Using the streaming application programming interface provided by the X Research Developer Platform, we initially extracted >6 million raw perinatal SUD–related tweets from around the world spanning May 1, 2019, to October 31, 2021. Subsequently, we refined this user-generated content by filtering out non–English-language tweets and tweets from non-US accounts and accounts without disclosed geolocations. In addition, we screened the data set to eliminate spam, advertisements, URLs, punctuation, stop words, and duplicate tweets. To be included in our analysis, each tweet had to contain at least 2 words from a predetermined data set of substance- and prenatal and child health–related keywords provided by our clinical team (detailed in Multimedia Appendix 1). After this rigorous filtering process, the final sample of relevant and clean tweets for this study comprised 4848 PPU-related tweets. To identify variances in our data, we also collected another set of 4848 non-PPU tweets on X during the same time frame as the control group for comparison analysis.

Ethical Considerations

The University of South Carolina's institutional review board reviewed our application (Pro00122484), social media data exaction, and analysis method for this study. This study was deemed non–human subjects research and thus exempted from review.

Study Design

Overview

This study used a mixed methods approach to analyze PPU data from X. The Quantitative Analysis Methods section introduces the quantitative analysis methods applied in this study, whereas the Qualitative Analysis Methods section details the qualitative methods used. For the quantitative analysis, 4848 tweets were analyzed. From this data set, 500 PPU tweets were randomly selected for more in-depth qualitative analysis.

Quantitative Analysis Methods

The quantitative analysis leveraged NLP techniques to elucidate the attributes and characteristics of the corpus of tweets. We used 2 key NLP techniques: frequency analysis to understand tweet trends, bigrams, trigrams, and sentiment analysis for the PPU-related sample of 4848 tweets, which we also used to compare their sentiment differences against a control group using a set of random 4848 non-PPU tweets extracted from X during the same period [21,22].

Before applying NLP techniques, proper cleaning was performed on the tweet data set to ensure the accuracy of the NLP outcomes. This data cleaning was solely performed for the quantitative analysis to apply various ML algorithms to analyze our tweets. The process involved tokenizing cleaned tweets into arrays in which each word represented an element. We removed stop words such as *a*, *is*, and *the*, which offer limited insights into tweet sentiment. To avoid the potential loss of sentiment intensity when lemmatizing words, we used the raw text as input for our sentiment analysis. In addition, non–American Standard Code for Information Exchange characters were removed from the text.

Sentiment analysis, which classifies the expressed sentiment in text, was conducted using the Valence Aware Dictionary and Sentiment Reasoner (VADER), which is designed to account for nuances such as punctuation, capitalization, and even emojis, thus providing a more accurate reflection of sentiment in the text. VADER, a lexicon and rule-based tool tailored for social media, assesses the polarity (positive, negative, and neutral) of each tweet and assigns a sentiment score ranging from –1 to 1. For instance, in the tweet "I hate and discourage the use of drugs while pregnant," it would highlight "hate" and "discourage" as negative, whereas in "I enjoy drugs and alcohol," it would identify "enjoy" as positive. On the basis of established thresholds, tweets with scores of >0.05 are deemed positive, tweets with scores of <–0.05 are deemed negative, and those in between are deemed neutral [23]. Sentiment analysis has broad

XSL•FO
RenderX

applications, including in marketing, politics, and health care, and is particularly effective in X-based health research [21,22]. Thus, in this study, we aimed to apply VADER sentiment analysis to identify the prevalent sentiments associated with the most frequently reoccurring substance combinations in PPU-related tweets.

Furthermore, we conducted frequency analyses, focusing on trend, bigram, and trigram analyses. Bigrams are sets of 2 words together, and trigrams are sets of 3 words together. This approach sheds light on prevalent discussion topics within the data set without necessitating a manual review of each tweet; as noted by Sarker et al [14], such analysis provides meaningful insights into the textual content.

Qualitative Analysis Methods

Qualitative analysis enabled a deeper exploration of the underlying semantics of individual tweets. By applying thematic content analysis, we gained deeper insights into PPU-related themes, subthemes, and user behaviors. This method, recently applied in the health care field, helps distill textual data into coherent themes. Thematic analysis emphasizes emergent themes and subjectivity, whereas content analysis focuses on explicit content and objectivity using either inductive or deductive approaches. Inductive content analysis derives concepts, categories, and themes directly from the data, whereas deductive analysis uses predefined frameworks [24,25].

For this analysis, we selected a random sample of 500 PPU tweets from our 4848 PPU-related tweets. In total, 2 independent researchers conducted the qualitative analysis using inductive coding techniques. The tweets were categorized into themes and subthemes using the NVivo software (version 12.0; Lumivero). The researchers started coding tweets together to ensure similar thought processes, followed by independent coding interspersed with periodic meetings to reconcile differences and jointly analyze more complex tweets. After coding all 500 tweets, we assessed interrater reliability using the κ score to determine the degree of agreement between the 2 coders and account for the probability of coding similarly by chance [26]. We computed 2 κ scores using NVivo's built-in functions: one for the tweets in the main themes and the other for substance combinations. The resulting average κ score was 0.6940 for differentiating tweets about single-substance use, polysubstance use, and non–substance use, achieving reasonable agreement between the 2 coders. Further investigation was conducted to find the average κ score between the combinations of substances. For substance combinations, the average κ score was 0.7748, reflecting a satisfying agreement given that previous studies have identified a κ score of 0.60 as moderate [26] and 0.79 as significant [14].

We established a codebook for this thematic content analysis. Figure 1 illustrates our coding structure with the first few layers of the multilayer codebook. Each tweet was first classified as non–substance use, single-substance use, or polysubstance use. Tweets regarding polysubstance use were further categorized by the specific combination of substances discussed. The final coding structure contains >3 levels or subbranches, so Figure 1 visually represents the general coding process for our qualitative analysis.

Figure 1. The general structure of the codebook. The 3 dots indicate more coding subbranches horizontally and vertically.

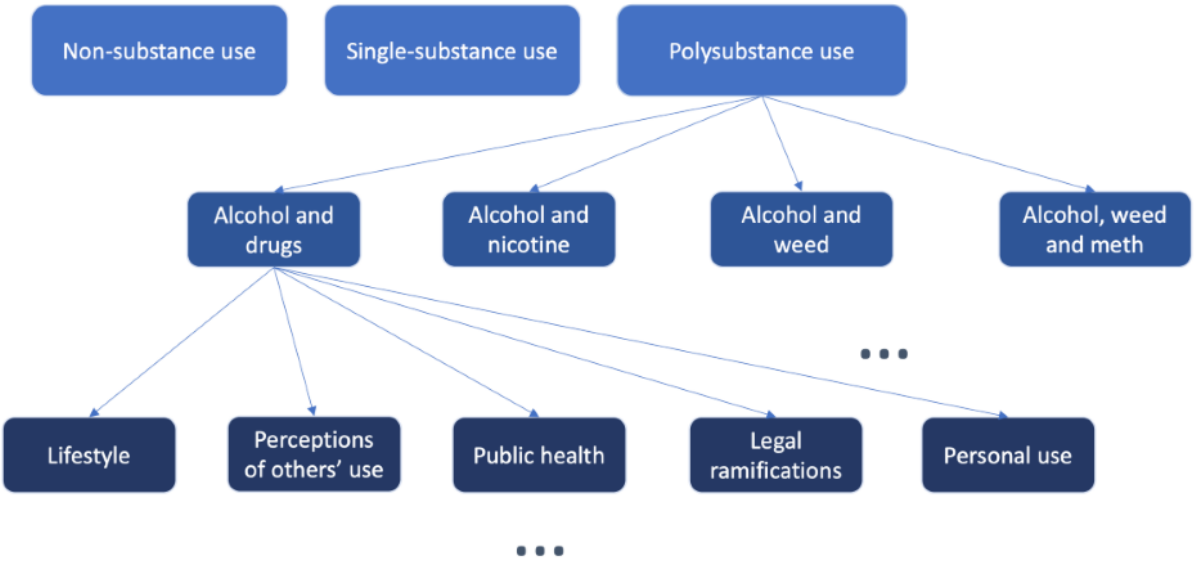

XSL•FO
RenderX

Results

Quantitative Analysis

Sentiment Analysis

To identify specific combinations of PPU, we analyzed commonly mentioned substances from the PPU tweets. We extracted combinations of substances that were both mentioned explicitly in a tweet; for instance, a tweet containing the words "alcohol" and "drug" was categorized into the "alcohol and drugs" combination. We then conducted sentiment analysis on the most frequently reoccurring substance combinations in our data. We further quantified sentiment differences using the chi-square test for the frequency counts of sentiment categories—positive, neutral, or negative—across PPU and non-PPU tweets. This statistical test was crucial for identifying significant variations in sentiment distributions between the 2 groups.

Table 1 displays the overall VADER sentiment analysis results for PPU tweets and random non-PPU tweets. To assess the sentiment distribution differences between PPU and non-PPU tweets, we applied the chi-square test and reported the results in Table 1. Notably, PPU tweets had a significantly lower percentage of positive sentiments (1884/4848, 38.86%) compared to the non-PPU tweet sample (2206/4848, 45.5%), with a *P* value of <.001, indicating that positive sentiments are less prevalent in PPU tweets. Conversely, negative sentiments were more common in PPU tweets (2490/4848, 51.36%) than in the non-PPU sample (1323/4848, 27.29%), with the statistical significance underscored by a *P* value of <.001. Neutral sentiments were significantly rarer in PPU tweets (474/4848, 9.78%) as opposed to in the control sample (1319/4848, 27.21%), also with a *P* value of <.001, suggesting a lower occurrence of neutral sentiments deemed useful for PPU analysis.

Table 1. Sentiment comparison analysis between perinatal polysubstance use (PPU) and non-PPU tweets (N=4848).

Sentiment	PPU tweets, n (%)	Random non-PPU tweets, n (%)	*P* value
Negative	2490 (51.36)	1323 (27.29)	<.001
Positive	1884 (38.86)	2206 (45.5)	<.001
Neutral	474 (9.78)	1319 (27.21)	<.001

Table 2 presents the sentiment analysis for the top 5 most reoccurring substance combinations. Regarding "alcohol and drugs," "meth and drugs," "weed and drugs," and "cocaine and drugs," the sentiment of over half of these substance pairs was negative, and the sentiment of <40% was positive, except for the sentiment of the "weed and meth" combination, which was mixed (positive: 70/160, 43.7%; negative: 76/160, 47.5%; neutral: 14/160, 8.8%).

Table 2. Sentiment analysis for the top 5 substance combinations.

Combination	Positive, n (%)	Neutral, n (%)	Negative, n (%)
Alcohol and drugs (n=861)	318 (36.9)	53 (6.2)	490 (56.9)
Meth and drugs (n=301)	122 (40.5)	14 (4.7)	165 (54.8)
Weed and drugs (n=202)	73 (36.1)	24 (11.9)	105 (52)
Weed and meth (n=160)	70 (43.7)	14 (8.8)	76 (47.5)
Cocaine and drugs (n=152)	51 (33.6)	11 (7.2)	90 (59.2)

Trend, Bigram, and Trigram Analyses

Next, we illustrate the overall tweet distribution of our main PPU tweets from May 2019 to October 2021 in Figure 2. We further present the frequency of the most popular bigrams and trigrams in PPU tweets (ie, 2- and 3-word sequences of words, respectively) in Table 3 and 4. Regarding the bigram frequencies listed in Table 3, the 2 apparent themes pertain to drug use and family dynamics. The most common bigrams were "drug alcohol," "drug use," and "smoke weed." These bigrams provided a glimpse of the major substance combinations that we were able to further identify through qualitative analysis, discussed in the next section. In addition, there was a multitude of bigrams alluding to domestic situations (eg, "family member" and "child abuse"). Regarding the trigrams listed in Table 4, the most common ones were related to drug abuse. A large quantity of the results were trigram permutations of the words "drug," "alcohol," "abuse," and "addiction." Table 4 presents these polysubstances containing the dominant substances.

XSL•FO
RenderX

Figure 2. Overall trend of perinatal polysubstance use tweets from May 1, 2019, to October 31, 2021.

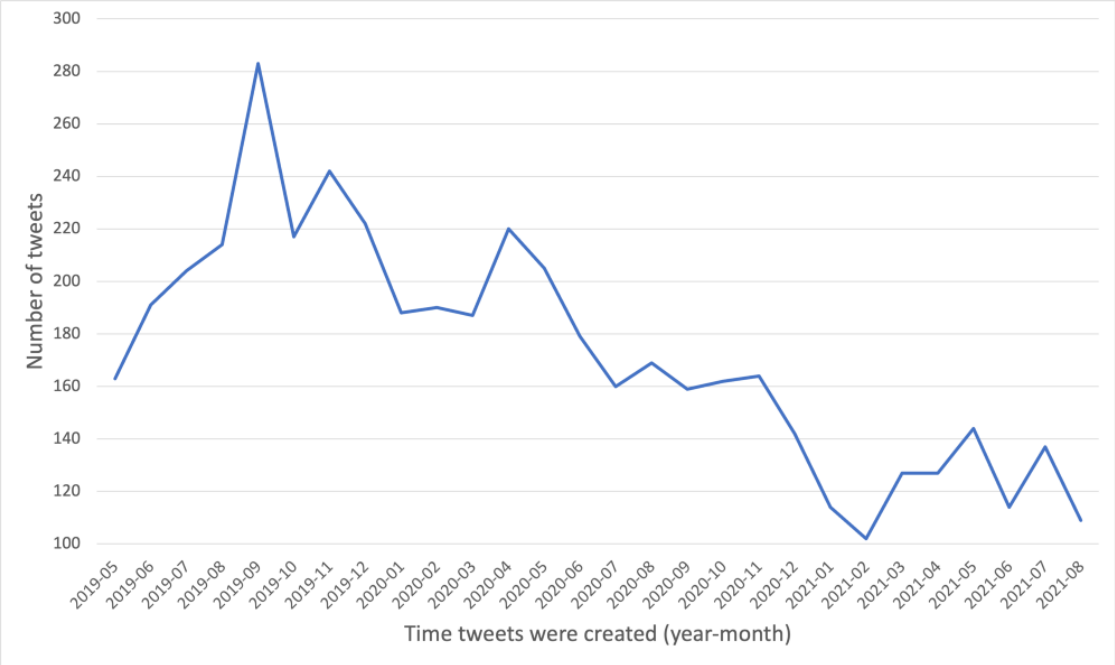

Table 3. Top 10 frequent bigrams in perinatal polysubstance use tweets.

Bigram	Frequency, n (%)
“Drug alcohol”	419 (23.92)
“Drug use”	356 (20.32)
“Smoke weed”	183 (10.45)
“Alcohol drug”	159 (9.08)
“Family member”	130 (7.42)
“Drug addict”	129 (7.36)
“Drug dealer”	102 (5.82)
“Drug user”	96 (5.48)
“Child abuse”	91 (5.19)
“Drug addiction”	87 (4.97)

Table 4. Top 10 frequent trigrams in perinatal polysubstance use tweets.

Trigram	Frequency, n (%)
“Drug alcohol abuse”	41 (17.52)
“Abuse drug alcohol”	24 (10.26)
“Alcoholic drug addict”	24 (10.26)
“Drug addict alcoholic”	23 (9.83)
“Child abuse drug”	22 (9.4)
“Mom pop vape”	22 (9.4)
“Drug alcohol addiction”	21 (8.97)
“Smell like weed”	20 (8.55)
“Alcoholism drug addiction”	19 (8.12)
“Fetal alcohol syndrome”	18 (7.69)

Overall, among the 4848 PPU tweets, we found a complicated polysubstance use among pregnant individuals. Of the 4848 tweets, 4134 (85.27%) tweets contained 2 substance keywords, 539 (11.12%) included 3 substance keywords, and 175 (3.61%)

XSL•FO
RenderX

mentioned >3 substances. Most PPU patterns turned out to be paired with 2-substance use. The top 5 substance combinations were *alcohol and drugs*, *meth and drugs*, *weed and drugs*, *cocaine and drugs*, and *alcohol and meth*, listed in Table 2. Table 5 further details the percentages of tweets ranked as the top 11 substance combinations. Figure S1 in Multimedia Appendix 2 visually presents the overall substance combinations in the PPU tweets. Table S1 in Multimedia Appendix 2 further summarizes the top 60 substance combinations included in the "other" category, including other identified double and triple substance combinations and their percentage distributions in this study. The *alcohol and drugs* combination was the most frequently reoccurring polysubstance group. This further supports the common perceptions about alcohol's popularity and its widespread acceptance, which has transformed general opinions to no longer consider it as a drug, as indicated in the following tweet quote, also indicating that legal substances tend to be discussed more openly and frequently on social media:

> *LMAO ALCOHOL AND TOBACCO IS LEGAL SO ARE YOU GONNA THINK THATS OKAY TO DO WHILE PREGNANT TOO?*

Table 5. Percentage of the top 11 substance pairs identified in the perinatal polysubstance use tweets (N=4848).

Keyword 1	Keyword 2	Tweets, n (%)
Alcohol	Drugs	861 (17.76)
Meth	Drugs	301 (6.21)
Weed	Drugs	202 (4.17)
Weed	Meth	160 (3.3)
Cocaine	Drugs	152 (3.14)
Alcohol	Meth	146 (3.01)
Alcohol	Weed	130 (2.68)
Heroin	Meth	117 (2.41)
Heroin	Drugs	102 (2.1)
Opioid	Drugs	92 (1.9)
Weed	Cigarette	81 (1.67)

Qualitative Analysis

Main Themes

We randomly selected a sample of 500 PPU tweets for a more in-depth thematic analysis. The tweets were initially coded into 3 categories depending on the substances mentioned (ie, *nonsubstance*, meaning that no substance was mentioned in the tweets; single substance, indicating that only 1 substance was discussed in the tweets; and *polysubstance*, representing multiple substances mentioned in tweets). PPU tweets were categorized based on the substance name and combination with another substance. We used 3 coding levels to provide an overall visual structure of our codebook illustrated in Figure 1 in the Qualitative Analysis Methods section, and we found more subthemes within each category to allow the researchers to capture the exact essence of each tweet.

Table 6 contains the definitions for each identified theme category with example tweets. Following our coding process (Figure 1), the 2 researchers independently coded and identified themes and subthemes for the sample of 500 PPU-related tweets. There may be some overlaps in the categories; for example, a tweet can start discussing single-substance use and then later discuss polysubstance use. We coded this type of tweets as polysubstance use because they discussed both drug and alcohol use.

XSL•FO
RenderX

Table 6. The 3 main themes discussed in the perinatal polysubstance use tweets.

Theme of substance use	Theme definition	Sample tweets
Non–substance use (n=36)	Content discussing the users' opinions explicitly not related to substance use	• "Sometimes I feel my fur child is a human! She drug her cover over to the couch, then got up and got under it https://link" • "My children from 26 -35 yrs old were exposed to glyphosate from birth. Many rural children were. I wish they hadn't been and we had not believed NZ and US weed scientists who assured it was safe enough to drink."
Single-substance use (n=223)	Content discussing single-substance use	• "I'm taking the Alcohol people out—are you pro-choice when it comes to abortion and drugs?" • "Babies born to mothers who smoke tobacco are at greater risk for ADHD. Children of mothers with high levels of nicotine exposure were more than twice as likely to develop ADHD than children born to mothers with low levels of exposure."
Polysubstance use (n=260)	Content discussing the use of multiple substances	• "I'm not going to share my story (science speaks for itself), but #youknowme for sure. I abused drugs and alcohol for years on & off—really nasty ones that do terrible damage to a fetus. Poor nutrition, no medical care, & could not stop using even when I suspected I was pregnant." • "As a case worker I have had children look me in the eyes & ask me why their mommy & daddy don't love them as much as they love drugs & alcohol. No child should have to search their entire life for someone to love them they way they deserved love from the beginning #ProChoice"

Substance Combinations

In addition to main-level categories or themes, we further coded the contents into subthemes to better illustrate the discussions on X. For the main PPU themes identified, we delved into specific drug combinations observed in the tweets. The most frequently reoccurring combinations were *alcohol and drugs*, *alcohol and cannabis*, and *alcohol and nicotine*. Table 7 presents a brief description of each, agreed upon by the 2 coders, along with example tweets. The κ score for the substance combinations was 0.7748.

XSL•FO
RenderX

Table 7. Definitions and examples of the most frequently reoccurring substance combinations in the perinatal polysubstance use tweets.

Combination subtheme	Definition	Sample tweets
Alcohol and drugs (n=120)	Discussions pertaining to the use of multiple substances or the mention of multiple substances in the same tweet	• "i literally hate today because my father abandoned me for drugs, alcohol & hookers he was NEVER in my life he was always in and out of jail he stole money from my mom he never cared but NOW he's coming back into my life & it sucks because i wish he were there when i needed him." • "The worst thing is watching someone you love throw their life away. Will never understand people who choose alcohol and drugs over their family and you can't help them anymore. Some people don't know how lucky they are."
Alcohol and cannabis (n=26)	Discussions pertaining to the use of multiple substances or the mention of multiple substances in the same tweet	• "I once read a forum where a guy was advocating weed to treat alcohol withdrawal. Can you imagine the last thing you think before you die a horrible death from delirium tremens being how much you want Arby's. Just shaky-poos and French Dips baby. F***** conk out to some Pink Floyd." • "I think she was high (from alcohol and weed) when she was making that Baby and she didn't remember the next day, poor thing."
Alcohol and nicotine (n=23)	Discussions pertaining to the use of multiple substances or the mention of multiple substances in the same tweet	• "Ain't nobody talking about giving pregnant women cigarettes and vodka I'm pro-choice and I'd cuss a woman out who was planning to carry a baby to term yet drinking alcohol I know all about birth defects my niece is a nurse my grandmother before her...That is not what we" • "This is no different tNhan taking my kids with me into Walgreens, CVS, the supermarket where #alcohol & #tobacco are in abundance. Or eating Mexican dinner with family with margaritas. And I'm still sick. I was refused service. I'm utterly disappointed & outraged over this!"
Nicotine and cannabis (n=10)	Discussions pertaining to the use of multiple substances or the mention of multiple substances in the same tweet	• "Smoking weed/cigarettes while pregnant was never said to be okay BUT moms do it anyway if they want to do THATS THEIR PROBLEM! STOP MAKING IT YOURS! If you got an opinion then okay but if someone else dont agree F*** IT! Why get so heated just cuz they feel different than you? ", • "There's no one consistent in this house like our family's middle child he wakes up smokes a cigarette, comes back and rolls a joint and goes to blaze it, he'll roll another one before lunch and blaze it and another one & blaze it before supper and before sleeping."

Subthemes

To extract more meaningful insights from each thematic area, we created additional subthemes within the identified main substance combinations. The most frequently reoccurring topics were lifestyle, perceptions of others' use, legal ramifications, and public health. These prominent subthemes were based on the frequency of tweets in each category without necessarily pertaining to PPU. Broad definitions of each of the leading subthemes can be found in Table 8 along with and example tweets.

There are instances in which the subtheme breakdowns contained additional themes within them. The subtheme *lifestyle* contained more tweets with positive sentiment than negative sentiment (14 vs 12 tweets). The positive subtheme illustrated support for substance use when X users expressed personal experiences and normalcy with substance use. Meanwhile, the negative subtheme discussed how others' use has affected the tweeters' lifestyle choices. *Lifestyle* was the most frequently occurring subtheme within the major substance combinations, and thus, individuals were comfortable discussing their substance use through social media. Many individuals discussed daily substance use, social substance use, and family's or friends' substance use as it pertained to lifestyle decisions.

XSL•FO
RenderX

Table 8. Definitions and examples of the most frequently occurring subthemes connected to polysubstance use.

Subtheme and substance combination	Example tweet
Lifestyle (n=38): content conveying certain habits and lifestyle choices pertaining to nicotine and alcohol	
Alcohol and nicotine	• "In Stoke, the young ladies switch to menthol cigarettes when they are confirmed pregnant. And they increase their alcohol intake, as they are drinking for two." • "I don't have debt...I don't smoke & seldom drink alcohol. It impairs my judgement. Cigarettes & alcohol are expensive. You have to launder clothing more often which wears them out prematurely plus they cause health problems in kids. My dad smoked. I was sick as infant & kid."
Alcohol and cannabis	• "Like 90% of my friends are married & starting families, instead of joining them in the baby making i've come to terms with that fact that i'm forever gonna be crazy aunt D with the cool tattoos that buys them alcohol & teaches them how to roll a perfect blunt & i'm ok with that."
Nicotine and cannabis	• "There's no one consistent in this house like our family's middle child he wakes up smokes a cigarette, comes back and rolls a joint and goes to blaze it, he'll roll another one before lunch and blaze it and another one & blaze it before supper and before sleeping."
Alcohol and drugs	• "My favorite thing about not being pregnant is drugs & alcohol "
Perceptions of others' use (n=33): discussions of substance use in others from witnessed events to assumed use	
Nicotine and cannabis	• "Smoking weed/cigarettes while pregnant was never said to be okay BUT moms do it anyway if they want to do THATS THEIR PROBLEM! STOP MAKING IT YOURS! If you got an opinion then okay but if someone else dont agree F*** IT! Why get so heated just cuz they feel different than you? "
Alcohol and nicotine	• "Ain't nobody talking about giving pregnant women cigarettes and vodka I'm pro-choice and I'd cuss a woman out who was planning to carry a baby to term yet drinking alcohol I know all about birth defects my niece is a nurse my grandmother before her...That is not what we"
Alcohol and drugs	• "...like I know everything, I just can't imagine how decestated [devastated] I would be personally to find out I had been unkowingly harming my baby. Whether it be pills/alcohol/illicit drugs babies aren't intended to be exposed to any of it and if that was me I would want to consider abortion."
Legal ramifications (n=28): discussions of legal situations, such as laws, regulations, and charges	
Alcohol and nicotine	• "If you can vote, die in battle, drive, & sadly enough have an abortion at 18 why not leave the tobacco law in place for 18 & possibly even lower alcohol age to 18 Just saying"
Alcohol and drugs	• "More talkin out of your backside because clearly you either haven't read or understand the bill! There's only a potential to be prosecuted for miscarriages if they are deliberate and willful like taking excess drugs and alcohol to deliberately cause a miscarriage."
Public health (n=23): content that looks to educate the public on substance use or offer support to users	
Alcohol and drugs	• "Are you a #family member of someone with a #drug or #alcohol #addiction? Do you support a #partner, a #child, #parent or #friend with #substance misuse? Do you need #help but don't know where to turn to? Call Michael on (phone #) for a friendly confidential #chat #carers #cope https://link"

Another common theme throughout *perception of others' use* was "use while pregnant." The total number of tweets in this subtheme of the major substance combinations discussed was 33. Additional themes allow for more accuracy and ensure that the exact topic discussed is being captured, further assisting the discovery of more specific recurring themes. The prevalence of the subsubtheme *perceptions of others' use* can be attributed to the intrinsic purpose of social media platforms such as X. Oftentimes, these platforms are used for individuals to voice their opinions of disagreement and concession. Some users stood up for the pregnant women while still holding their stance that substance use should be avoided:

> *Smoking weed/cigarettes while pregnant was never said to be okay BUT moms do it anyway if they want to do THATS THEIR PROBLEM! STOP MAKING IT YOURS! If you got an opinion then okay but if someone else dont agree **** IT! Why get so heated just cuz they feel different than you?*

Others chose to adhere to their beliefs and shame individuals who thought otherwise and even went as far as to suggest aborting the pregnancy if it was believed that substances were used while pregnant:

XSL•FO
RenderX

> *Whether it be pills/alcohol/illicit drugs babies aren't intended to be exposed to any of it and if that was me I would want to consider abortion.*

Discussions of abortion and miscarriage laws were the main topic behind the subtheme of *legal ramifications*. Criminal charges can be brought against a mother who uses substances while pregnant; some of those charges include child endangerment, attempted aggravated child abuse, child neglect, and even manslaughter [12]. Many of the tweets that fell under this subtheme were related to when a woman could be charged for a miscarriage if she ingested substances during her pregnancy. One discussion revolved around whether a woman would be held responsible for a miscarriage due to substance use if she was unaware of her pregnancy.

Users who are currently pregnant and struggling with substance abuse may find all the negativity and judgment discouraging, thus making them hesitant to reach out for help due to fear of persecution. Researchers who studied secrecy versus disclosure found several explanations for apprehension in pregnant women with substance use issues, specifically shame and fear of judgment as motives for not reaching out to family and physicians [11]. Mothers who use substances are ashamed of the harm they are potentially causing their children while they live in fear of criminal charges they could be subject to, along with the possibility of having their children taken away from them [12,27].

The following is one of many tweets discovered in our corpus that project negative opinions of women who partake in substance use during their pregnancy. It exemplifies the stigmas that surround the topic:

> *just really grossed out by girls who smoke weed, vape, and drink while they're pregnant do better.*

Discussion

Principal Findings

Quantitative Findings

The quantitative sentiment analysis revealed clear variations in user sentiments toward PPU. Tweets related to PPU presented a significantly lower percentage of positive sentiments compared to those unrelated to PPU. In addition, PPU-related tweets presented a significantly higher percentage of negative sentiments. Furthermore, neutral sentiments were significantly less frequent in PPU tweets—only 9.78% (474/4848) compared to 27.21% (1319/4848) in the non-PPU sample—highlighting a lower tendency to express neutral views on PPU. We also identified a set of complicated substance use combinations in the 4848 PPU tweets, among which paired substance use (4134/4848, 85.27%) dominated, with "alcohol and drugs," "meth and drugs," "weed and meth," and "cocaine and drugs" as the top 5 paired substance types. The combination of "alcohol and drugs" use was ranked as the most frequently reoccurring combination in PPU discussions on X. A total of 11.12% (539/4848) of the PPU tweets mentioned 3 substances, and 3.61% (175/4848) included >3 substances.

Qualitative Findings

The results from thematic content analysis illustrated that alcohol and drugs were the driving forces of the substance combinations. These substances were the most frequent in the combinations discovered through the qualitative analysis. Alcohol is a widely accepted substance due to its use in social gatherings and personal consumption as a legal substance if users' ages meet legal requirements, thus making it a leading substance in this study.

On the basis of the qualitative sentiment analysis, PPU had an overall negative sentiment in tweets, implying that substance use during pregnancy was mostly not supported in the gathered tweets. Except in specific situations, X users essentially perceived that PPU is dangerous and should be avoided. Some users even suggested abortion to avoid adverse outcomes for the infant. Consistent themes from this intersection include how others viewed the pregnant individual's PPU regarding the negative health outcomes for the child when substances are used during pregnancy and the significant stigma related to using substances while pregnant. The subtheme *public health* provided some knowledge to educate pregnant individuals on avoiding, reducing, and intervening in PPU and offer support to this vulnerable population.

Furthermore, a high volume of tweets pertaining to *legal ramifications* (n=28) reflected impassioned speakers on social media platforms. Users will tweet their opinions to corroborate, advertise, and support their views on the current state of US politics and policies. One major topic dealt with miscarriage charges for pregnant individuals who used substances during the pregnancy and how their behaviors may have led to the miscarriage. This can be connected to the fear of disclosing their substance use patterns while pregnant.

Comparison With Prior Work

Tweets containing shame and judgmental opinions of mothers who use substances were also found. This is consistent with previous work investigating the reasons for the secrecy of drug use in pregnant women [11]. Although the X platform is anonymous, these negative opinions expressed by X users may make it difficult for pregnant individuals to disclose their drug use and a desire for help for fear of being judged on a large social media platform. Public health agencies may want to ensure nonstigmatizing messaging with treatment resources to pregnant women on social media as they may routinely access the platform to gather information but may be less inclined to tweet about it. Our identified major substance combinations were consistent with those found in previous studies. However, while previous research has typically identified alcohol and cannabis or alcohol and tobacco as the most common substance combinations among pregnant individuals [1,28], our analysis revealed that alcohol and drugs were the most prevalent substances within the combinations studied. Through sentiment analysis, we were able to identify a narrow gap between positive sentiment and negative sentiment tweets; this can be due to the normalization of substance use as previous research on the use of alcohol and cannabis through X found a similar theme of normalization [26].

XSL•FO
RenderX

Limitations

One limitation of this study was the fact that there were not many tweets specifically discussing PPU in the perinatal population, coupled with the lack of first-person experiences being shared. It has been noted that many pregnant individuals are afraid to admit substance use due to the associated stigma [11]. While X is a large social media platform that is used to anonymously interact with others on the web, few individuals openly disclosed personal use during pregnancy due to stigma, negative health effects, and potential legal ramifications. Next, although our list of inclusion terms was rigorously selected by clinicians, it may not have been exhaustive, potentially missing some relevant terms and tweets. Moreover, with the recent X data access policy changes, without substantial funding and resources, public social media data such as those of X cannot be continuously accessible and affordable to social media researchers, including those who conduct health care social media research. In addition, it is challenging to study different types of X users, who are not explicitly categorized into organizations or any other user groups, without a high level of self-disclosure by users themselves.

Conclusions and Future Directions

Among the substance combinations discovered through the qualitative analysis, we identified alcohol and drugs as the most prevalent substances discussed in PPU tweets. The legality of alcohol use and the growing legalization of marijuana could be the reason for the increased frequency of pairing these specific substances. We found that the public opinion of X users was generally that pregnant individuals should avoid using single or multiple substances. Our quantitative and qualitative analyses supported more negative sentiment discussions of PPU on X, with additional communication exchanges discussing criminal charges for pregnant women who use drugs. It is promising to see public health messaging by individuals and public health organizations on the benefits of good prenatal care, recommending avoidance of substance use during pregnancy and offering support to the vulnerable perinatal population. In addition, substance treatment resource information was exchanged on where and how to seek assistance for individuals or family members who were addicted to drugs. Retweeting these types of positive communication exchanges could provide an opportunity for broader dissemination and education of individuals on ways to acquire these needed resources. As such, social media such as X could be further used by health organizations and web-based communities to educate the perinatal population and provide needed support and help anytime, anywhere, and at scale.

Overall, this research allowed us to gain a more holistic understanding of PPU patterns discussed on X, uncovering the complexity of the alarming PPU phenomenon and its important implications to PPU practitioners and associated public health policy makers regarding legal ramifications. Through a mixed methods approach, this study established a solid foundation that requires further investigations to understand the underlying mechanisms of PPU contextually, technically, and clinically. The methodologies used in this study can be adapted to analyze content on other social media platforms such as Reddit, Instagram, TikTok, Pinterest, and Weibo, where discussions about substance use also occur [29]. From a computational analysis perspective, we plan to train ML models to automatically detect the PPU tweets at scale, potentially identifying at-risk women who use substances and, thus, providing them with the proper knowledge, early intervention, treatment for PPU-related addiction, and necessary support that they might not have otherwise obtained. There remains a significant need to advocate for and implement PPU-related early intervention programs at the individual, community, clinical, and governmental levels. This study implies that providing relevant patient education through early intervention programs, access to addiction treatments and community support, and extended perinatal care services is critical to supporting the vulnerable perinatal population with PPU for better health outcomes for both mothers and babies. In addition, this study did not cover PPU for medical purposes, which can be our future research if we have more access to PPU-related clinical data to supplement our social media data to further enrich our knowledge and the clinical implications in this important health domain.

Acknowledgments

The authors would like to acknowledge the generous funding support provided by the University of South Carolina (USC), Columbia, South Carolina, United States (grant 80002838; principal investigator [PI]: DW), and from the USC Big Data Health Science Center, a USC Excellence Initiative program (grants BDHSC-2021-14 and BDHSC-2022; PI: DW), and a research grant from the USC Center for Advancing Chronic Care Outcomes through Research and Innovation (Seed Grant; PI: DW). The research reported in this publication was supported by the National Institute on Drug Abuse under award 1K23DA051626-01A1 (PI: PR) and the National Institutes of Health Big Data Health Science Scholar Program for Infectious Diseases (T35AI165252). The content expressed in this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Keywords provided by our clinical team for this research.

XSL•FO
RenderX

[DOCX File , 22 KB - jmir_v26i1e53171_app1.docx]

Multimedia Appendix 2
Overall substance combinations in the perinatal polysubstance use tweets and the top 60 substance combinations labeled as other.
[DOCX File , 204 KB - jmir_v26i1e53171_app2.docx]

References

1. England LJ, Bennett C, Denny CH, Honein MA, Gilboa SM, Kim SY, et al. Alcohol use and co-use of other substances among pregnant females aged 12-44 years - United States, 2015-2018. MMWR Morb Mortal Wkly Rep 2020 Aug 07;69(31):1009-1014 [FREE Full text] [doi: 10.15585/mmwr.mm6931a1] [Medline: 32759915]
2. Frazer Z, McConnell K, Jansson LM. Treatment for substance use disorders in pregnant women: motivators and barriers. Drug Alcohol Depend 2019 Dec 01;205:107652. [doi: 10.1016/j.drugalcdep.2019.107652] [Medline: 31704383]
3. Hand DJ, Short VL, Abatemarco DJ. Substance use, treatment, and demographic characteristics of pregnant women entering treatment for opioid use disorder differ by United States census region. J Subst Abuse Treat 2017 May;76:58-63. [doi: 10.1016/j.jsat.2017.01.011] [Medline: 28161143]
4. Clemans-Cope L, Lynch V, Howell E, Hill I, Holla N, Morgan J, et al. Pregnant women with opioid use disorder and their infants in three state medicaid programs in 2013-2016. Drug Alcohol Depend 2019 Feb 01;195:156-163. [doi: 10.1016/j.drugalcdep.2018.12.005] [Medline: 30677745]
5. Hamułka J, Zielińska MA, Chądzyńska K. The combined effects of alcohol and tobacco use during pregnancy on birth outcomes. Rocz Panstw Zakl Hig 2018;69(1):45-54 [FREE Full text] [Medline: 29517191]
6. McQueen K, Murphy-Oikonen J. Neonatal abstinence syndrome. N Engl J Med 2016 Dec 22;375(25):2468-2479. [doi: 10.1056/nejmra1600879]
7. Sebastiani G, Borrás-Novell C, Casanova MA, Pascual Tutusaus M, Ferrero Martínez S, Gómez Roig MD, et al. The effects of alcohol and drugs of abuse on maternal nutritional profile during pregnancy. Nutrients 2018 Aug 02;10(8):1008 [FREE Full text] [doi: 10.3390/nu10081008] [Medline: 30072661]
8. Jarlenski MP, Paul NC, Krans EE. polysubstance use among pregnant women with opioid use disorder in the United States, 2007-2016. Obstet Gynecol 2020 Sep;136(3):556-564 [FREE Full text] [doi: 10.1097/AOG.0000000000003907] [Medline: 32769641]
9. Kar P, Tomfohr-Madsen L, Giesbrecht G, Bagshawe M, Lebel C. Alcohol and substance use in pregnancy during the COVID-19 pandemic. Drug Alcohol Depend 2021 Aug 01;225:108760 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108760] [Medline: 34102507]
10. Jones SC, Telenta J, Shorten A, Johnson K. Midwives and pregnant women talk about alcohol: what advice do we give and what do they receive? Midwifery 2011 Aug;27(4):489-496. [doi: 10.1016/j.midw.2010.03.009] [Medline: 20471731]
11. Paris R, Herriott AL, Maru M, Hacking SE, Sommer AR. Secrecy versus disclosure: women with substance use disorders share experiences in help seeking during pregnancy. Matern Child Health J 2020 Nov;24(11):1396-1403. [doi: 10.1007/s10995-020-03006-1] [Medline: 33025236]
12. Angelotta C, Appelbaum PS. Criminal charges for child harm from substance use in pregnancy. J Am Acad Psychiatry Law 2017 Jun;45(2):193-203. [Medline: 28619859]
13. Myslín M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013 Aug 29;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
14. Sarker A, Chandrashekar P, Magge A, Cai H, Klein A, Gonzalez G. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. J Med Internet Res 2017 Oct 30;19(10):e361 [FREE Full text] [doi: 10.2196/jmir.8164] [Medline: 29084707]
15. Alshammari SM, Nielsen RD. Less is more: with a 280-character limit, Twitter provides a valuable source for detecting self-reported flu cases. In: Proceedings of the 2018 International Conference on Computing and Big Data. 2018 Presented at: ICCBD '18; September 8-10, 2018; Charleston, SC p. 1-6 URL: https://dl.acm.org/doi/10.1145/3277104.3277105 [doi: 10.1145/3277104.3277105]
16. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]
17. Allem JP, Dharmapuri L, Leventhal AM, Unger JB, Boley Cruz T. Hookah-related posts to Twitter from 2017 to 2018: thematic analysis. J Med Internet Res 2018 Nov 19;20(11):e11669 [FREE Full text] [doi: 10.2196/11669] [Medline: 30455162]
18. Chu KH, Majmundar A, Allem JP, Soto DW, Cruz TB, Unger JB. Tobacco use behaviors, attitudes, and demographic characteristics of tobacco opinion leaders and their followers: Twitter analysis. J Med Internet Res 2019 Jun 04;21(6):e12676 [FREE Full text] [doi: 10.2196/12676] [Medline: 31165716]

XSL•FO
RenderX

19. Allem JP, Dormanesh A, Majmundar A, Unger JB, Kirkpatrick MG, Choube A, et al. Topics of nicotine-related discussions on Twitter: infoveillance study. J Med Internet Res 2021 Jun 07;23(6):e25579 [FREE Full text] [doi: 10.2196/25579] [Medline: 34096875]
20. Oram D, Tzilos Wernette G, Nichols LP, Vydiswaran VV, Zhao X, Chang T. Substance use among young mothers: an analysis of Facebook posts. JMIR Pediatr Parent 2018 Dec 04;1(2):e10261 [FREE Full text] [doi: 10.2196/10261] [Medline: 31518312]
21. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 2020 Jan 28;8(1):e16023 [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
22. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care Tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
23. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with twitter data in the United States: spatiotemporal perspective. J Med Internet Res 2021 Sep 10;23(9):e30854 [FREE Full text] [doi: 10.2196/30854] [Medline: 34346888]
24. Albanese-O'Neill A, Schatz DA, Thomas N, Bernhardt JM, Cook CL, Haller MJ, et al. Designing online and mobile diabetes education for fathers of children with type 1 diabetes: mixed methods study. JMIR Diabetes 2019 Aug 06;4(3):e13724 [FREE Full text] [doi: 10.2196/13724] [Medline: 31389338]
25. Yabumoto M, Miller E, Rao A, Tabor HK, Ormond KE, Halley MC. Perspectives of rare disease social media group participants on engaging with genetic counselors: mixed methods study. J Med Internet Res 2022 Dec 21;24(12):e42084 [FREE Full text] [doi: 10.2196/42084] [Medline: 36542454]
26. Krauss MJ, Grucza RA, Bierut LJ, Cavazos-Rehg PA. “Get drunk. Smoke weed. Have fun.”: a content analysis of tweets about marijuana and alcohol. Am J Health Promot 2016 Nov 17;31(3):200-208. [doi: 10.4278/ajhp.150205-qual-708]
27. Paterno MT, Low M, Gubrium A, Sanger K. Mothers and mentors: exploring perinatal addiction and recovery through digital storytelling. Qual Health Res 2019 Mar 06;29(4):545-556. [doi: 10.1177/1049732318777474] [Medline: 29871558]
28. Qato DM, Zhang C, Gandhi AB, Simoni-Wastila L, Coleman-Cowger VH. Co-use of alcohol, tobacco, and licit and illicit controlled substances among pregnant and non-pregnant women in the United States: findings from 2006 to 2014 National Survey on Drug Use and Health (NSDUH) data. Drug Alcohol Depend 2020 Jan 01;206:107729. [doi: 10.1016/j.drugalcdep.2019.107729] [Medline: 31760250]
29. Rutherford BN, Lim CC, Johnson B, Cheng B, Chung J, Huang S, et al. #TurntTrending: a systematic review of substance use portrayals on social media platforms. Addiction 2023 Feb 08;118(2):206-217 [FREE Full text] [doi: 10.1111/add.16020] [Medline: 36075258]

Abbreviations

ML: machine learning
NLP: natural language processing
PPU: perinatal polysubstance use
PUD: polysubstance use disorder
RQ: research question
SUD: substance use disorder
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by A Mavragani; submitted 28.09.23; peer-reviewed by M Jenkins, T Deng, S Ramlawi; comments to author 13.10.23; revised version received 06.05.24; accepted 11.06.24; published 20.09.24.

Please cite as:
Wu D, Shead H, Ren Y, Raynor P, Tao Y, Villanueva H, Hung P, Li X, Brookshire RG, Eichelberger K, Guille C, Litwin AH, Olatosi B
Uncovering the Complexity of Perinatal Polysubstance Use Disclosure Patterns on X: Mixed Methods Study
J Med Internet Res 2024;26:e53171
URL: https://www.jmir.org/2024/1/e53171
doi:10.2196/53171
PMID:

©Dezhi Wu, Hannah Shead, Yang Ren, Phyllis Raynor, Youyou Tao, Harvey Villanueva, Peiyin Hung, Xiaoming Li, Robert G Brookshire, Kacey Eichelberger, Constance Guille, Alain H Litwin, Bankole Olatosi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

XSL•FO
RenderX

and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Racial Biases Associated With Pulse Oximetry: Longitudinal Social Network Analysis of Social Media Advocacy Impact

Wasim Ahmed[1], BA, MSc, PhD; Mariann Hardey[2], BA, MA, PhD; Bradford David Winters[3], MD, PhD; Aarti Sarwal[4], MD

[1]Marketing Management and Business Strategy, Hull University Business School, University of Hull, Kingston Upon Hull, United Kingdom
[2]Durham University Business School, Durham University, Durham, United Kingdom
[3]Anesthesiology and Critical Care, John Hopkins School of Medicine, Baltimore, MD, United States
[4]Neurology, Atrium Wake Forest School of Medicine, Winston-Salem, NC, United States

Corresponding Author:
Wasim Ahmed, BA, MSc, PhD
Marketing Management and Business Strategy
Hull University Business School
University of Hull
Cottingham Road
Kingston Upon Hull, HU6 7RX
United Kingdom
Phone: 44 1482 466914
Email: w.ahmed@hull.ac.uk

Abstract

Background: Pulse oximetry is a noninvasive method widely used in critical care and various clinical settings to monitor blood oxygen saturation. During the COVID-19 pandemic, its application for at-home oxygen saturation monitoring became prevalent. Further investigations found that pulse oximetry devices show decreased accuracy when used on individuals with darker skin tones. This study aimed to investigate the influence of X (previously known as Twitter) on the dissemination of information and the extent to which it raised health care sector awareness regarding racial disparities in pulse oximetry.

Objective: This study aimed to explore the impact of social media, specifically X, on increasing awareness of racial disparities in the accuracy of pulse oximetry and to map this analysis against the evolution of published literature on this topic.

Methods: We used social network analysis drawing upon Network Overview Discovery and Exploration for Excel Pro (NodeXL Pro; Social Media Research Foundation) to examine the impact of X conversations concerning pulse oximetry devices. Searches were conducted using the Twitter Academic Track application programming interface (as it was known then). These searches were performed each year (January to December) from 2012 to 2022 to cover 11 years with up to 52,052 users, generating 188,051 posts. We identified the nature of influencers in this field and monitored the temporal dissemination of information about social events and regulatory changes. Furthermore, our social media analysis was mapped against the evolution of published literature on this topic, which we located using PubMed.

Results: Conversations on X increased health care awareness of racial bias in pulse oximetry. They also facilitated the rapid dissemination of information, attaining a substantial audience within a compressed time frame, which may have impacted regulatory action announced concerning the investigation of racial biases in pulse oximetry. This increased awareness led to a surge in scientific research on the subject, highlighting a growing recognition of the necessity to understand and address these disparities in medical technology and its usage.

Conclusions: Social media platforms such as X enabled researchers, health experts, patients, and the public to rapidly share information, increasing awareness of potential racial bias. These platforms also helped connect individuals interested in these topics and facilitated discussions that spurred further research. Our research provides a basis for understanding the role of X and other social media platforms in spreading health-related information about potential biases in medical devices such as pulse oximeters.

(J Med Internet Res 2024;26:e56034) doi:10.2196/56034

XSL•FO
RenderX

KEYWORDS

social media; X; racial biases; pulse oximetry; advocacy; impact; awareness; racial; bias; biases; longitudinal study; information; dissemination; disparity; disparities; accuracy; social network analysis; Academic Track application programming interface; API

Introduction

Pulse oximetry is a noninvasive technique used extensively in critical care and other clinical environments to measure the oxygen saturation of blood. During the COVID-19 pandemic, the use of pulse oximetry for home oxygen saturation monitoring became widespread, allowing clinicians to track oxygen saturation levels, hence disease progression at home, with the goal of reducing avoidable emergency department visits [1]. Several applications of pulse oximetry, including self-monitoring at home, constant wireless monitoring in hospital facilities, and closed-loop oxygen delivery, were investigated during the pandemic [2]. With increasing monitoring, potential problems were noticed with pulse oximetry, such as accuracy concerns in skin colors and sources of error that could affect its use [3].

Subsequent investigations revealed that pulse oximetry measurement devices exhibit reduced accuracy when used on individuals with darker skin tones, a measurement issue that has been underappreciated for decades [4]. This phenomenon is attributed to the higher melanin content in individuals with dark skin, which leads to increased light absorption, impeding the accurate measurement of oxygenated blood levels by pulse oximeters [5]. Further investigations uncovered an adverse impact of pulse oximetry accuracy and evidence of statistical bias in pulse oximeter measurements, specifically in relation to race on patient care in black and colored patients, putting these patients at risk of hypoxemia and delayed escalation of care both in COVID-19– and non–COVID-19–related clinical scenarios [6-8].

Researchers called for more investigation into measures of bias, precision, and accuracy in the use of pulse oximetry to improve the interpretation of evolving research findings and facilitate the identification of potential solutions to these disparities [9]. Increasing awareness highlighted the racial bias in several medical devices, increased variation in monitoring data for Black patients, and its adverse impact on delivering evidence-based care, including pulse oximeter readings [10].

Health care professionals and the general public are now more aware of this issue due to campaign groups, advocacy, communities, and social media activism [3,11-13] and increasing awareness of disparities associated with pulse oximetry use in biomedical literature. With several hallmark social events around the time of the pandemic elaborating existing societal biases in health care that were disseminated globally through social media, we wondered about the role social media played in connecting social events to health care awareness on racial biases in a monitoring device such as pulse oximetry. We undertook this project to explore the impact of social media on increasing awareness of racial disparities in the accuracy of pulse oximetry and how information spreads among physicians. We used social network analysis to analyze the evolution of the conversations around pulse oximetry devices on X (previously known as Twitter) in relation to timelines of major social events raising health care disparities. Social network analysis has recently been used to explore and analyze the drivers of conspiracy theories [14,15].

We found that social media helped raise awareness of racial bias in pulse oximetry and should be considered a key communication channel alongside traditional methods such as medical publications and conferences. Our findings suggest that social media platforms have the potential to raise awareness about health inequities effectively and can be used to advocate for transformative regulatory action, adding to the growing body of evidence that social media can play a crucial role in advancing health equity. Our study is the first to analyze social media discussions on the specific topic of racial disparities in pulse oximetry accuracy.

Methods

Data Retrieval

The study adopted a retrospective review of social media activity, drawing upon data from X. Specifically, data were retrieved using the Academic Track application programming interface (API), as it was known then, using the search string: “pulseoximetry” OR “Pulse Oximetry” OR “Oximetry.” Searches were conducted using the X API each year (January to December) from 2012 to 2022 to cover an 11-year time period. This time period was selected as it generated the most sufficient data. We also examined literature indexed on PubMed between 1945 and 2020 and mapped it corresponding to 2012-2022. We captured data leading up to the end of 2022 when the US Food and Drug Administration (FDA) convened a group about ongoing concerns on the accuracy of pulse oximeters in individuals with darker skin pigmentations despite initial notification related to its use in the pandemic in February 2021.

Data Analysis

Data were analyzed drawing upon Network Overview Discovery and Exploration for Excel (NodeXL, version 1.0.1.514; Social Media Research Foundation) to perform social network analysis and generate social network visuals. The Harel-Koren Fast Multiscale layout algorithm, which is useful for visualizing large undirected networks [14], was used to lay out the report. The nodes, entities like people or organizations between which connections are being explored, were clustered, drawing on the Clauset-Newman-Moore algorithm [16]. This gathered data specific to 2020 provided the most influential organizations and individual users on X participating in the discussion involving pulse oximetry. We also extracted their number of followers. NodeXL was also used to analyze the content of tweets to generate “word pairs,” which are able to show the 2 most frequently co-occurring words. Our study also drew upon the betweenness centrality algorithm [17]. Betweenness centrality is a method for determining the degree to which a node affects the information flow in a graph. It is frequently used to identify

XSL•FO
RenderX

nodes that function as a connection between 2 sections of a graph. The algorithm calculates shortest paths between all pairs of nodes in a graph. We also categorized all posts based on the 6 types of X social network structures that can emerge [17], which are visually represented in Figure 1.

Figure 1. 6 types of network structures that can emerge on X (formerly Twitter; sourced from Smith et al [18] and the Social Media Research Foundation).

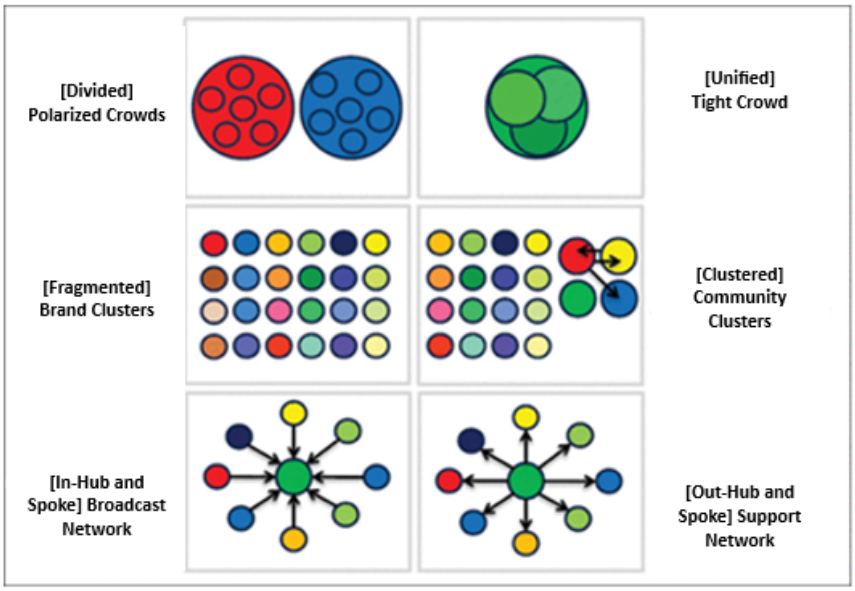

Polarized Crowds in X networks are formed by distinct, large groups discussing the same topic but with little to no interaction between them, often reflecting divergent views on contentious subjects like politics. Tight Crowds are formed around specialized topics or communities, such as professional interests, hobbies, or niche activities. In these networks, participants are highly interconnected, creating a dense web of interactions. Brand Clusters emerge when conversations on X revolve around specific brands or products without forming strong interconnections among the discussing participants. This results in many isolated individuals or small groups mentioning the brand but not engaging with one another, creating a scattered landscape of discussions. Community Clusters appear in conversations that encompass a broad topic with multiple subtopics or perspectives, forming several smaller, densely connected groups within the more considerable debate. The Broadcast Network structure is observed when a single, prominent account or a few accounts disseminate information to many followers, mirroring a hub-and-spoke model. Support Networks are identified by their hub-and-spoke configuration, where a central account, typically belonging to a company's customer service department, interacts with numerous individual users.

Social media platforms have become central hubs for civil society, serving as venues for sharing information, engaging in public discourse, and hosting debates and conflicts [19]. As modern equivalents of the public square, the discussions on social media hold significant importance, comparable to documenting any substantial public gathering [18,20,21]. Creating network maps of these social media dialogues on platforms like X can shed light on the societal impact of social media. These maps resemble overhead snapshots of a crowd, revealing approximate population size and diversity.

Like topographic maps display mountain ranges' peaks, network maps can reveal the most prominent points within a network landscape [18]. Specific individuals hold positions in these networks that are akin to strategically essential locations in physical terrain [18]. Using network metrics of "centrality," it is possible to pinpoint key individuals who are pivotal in discussion networks, acting as leaders in the conversation. The content generated by these central figures tends to gain considerable popularity and is often widely shared across the network, highlighting their influential role in shaping social media discourse.

Ethical Considerations

The study had ethical approval from Newcastle University (Reference 26055/2022), and we also took care to not draw attention to any individuals or accounts that were not already in the public domain. The majority of our analysis was conducted on an aggregate basis. During the editing process, we drew upon generative language models to improve readability and identify typographical errors of existing human-authored text with oversight from the authors, but these models were not used to generate any content of their own.

Results

Social Network and Interaction Analysis

The social network of pulse oximetry discussions on X from 2012 to 2022 is shown in Figure 2. We cut off in 2022 due to changes in the X (Twitter) API and are no longer able to access data through the Academic Track API. The social network analysis shows that the social network of pulse oximetry discussions on X has evolved significantly since 2019. In the early years (2012 to 2018), the largest group in each network consisted of a significant "isolates" group where users posted without being reposted and mentioning other users. Our analysis

XSL•FO
RenderX

suggests that the topics were broad-ranging, with many "islands" of users conversing with each other. There was also very little cross-collaboration, suggesting that the networks lacked content and engagement. In 2019, there was a transformation as a "broadcast network" appeared for the first time as the largest group in the network. This change resulted from the discussion around pulse oximetry becoming more centralized, with a few key users influencing many others. There was also some evidence of cross-collaboration, suggesting that the networks were becoming more integrated. The most significant change occurred in 2020, when the network had been completely transformed, with many densely connected groups of users and several broadcast networks appearing. Users also appeared to be conversing across groups. Our results show that the discussion around pulse oximetry had become much more active and engaged, with various perspectives being represented. A similar pattern emerged in 2021 and 2022, with even larger individual groups and more evidence of cross-collaboration. This suggests that the social network of pulse oximetry discussions on X continued to evolve and mature. Overall, the social network analysis shows that the X pulse oximetry discussion has become more active, engaged, and integrated over time. This change in engagement is likely due to several factors, including the COVID-19 pandemic, which has increased public awareness of pulse oximetry.

Figure 2. Social network analysis overview (2012 to 2022) of X posts and interactions (all forms of reposts and replies) on pulse oximetry.

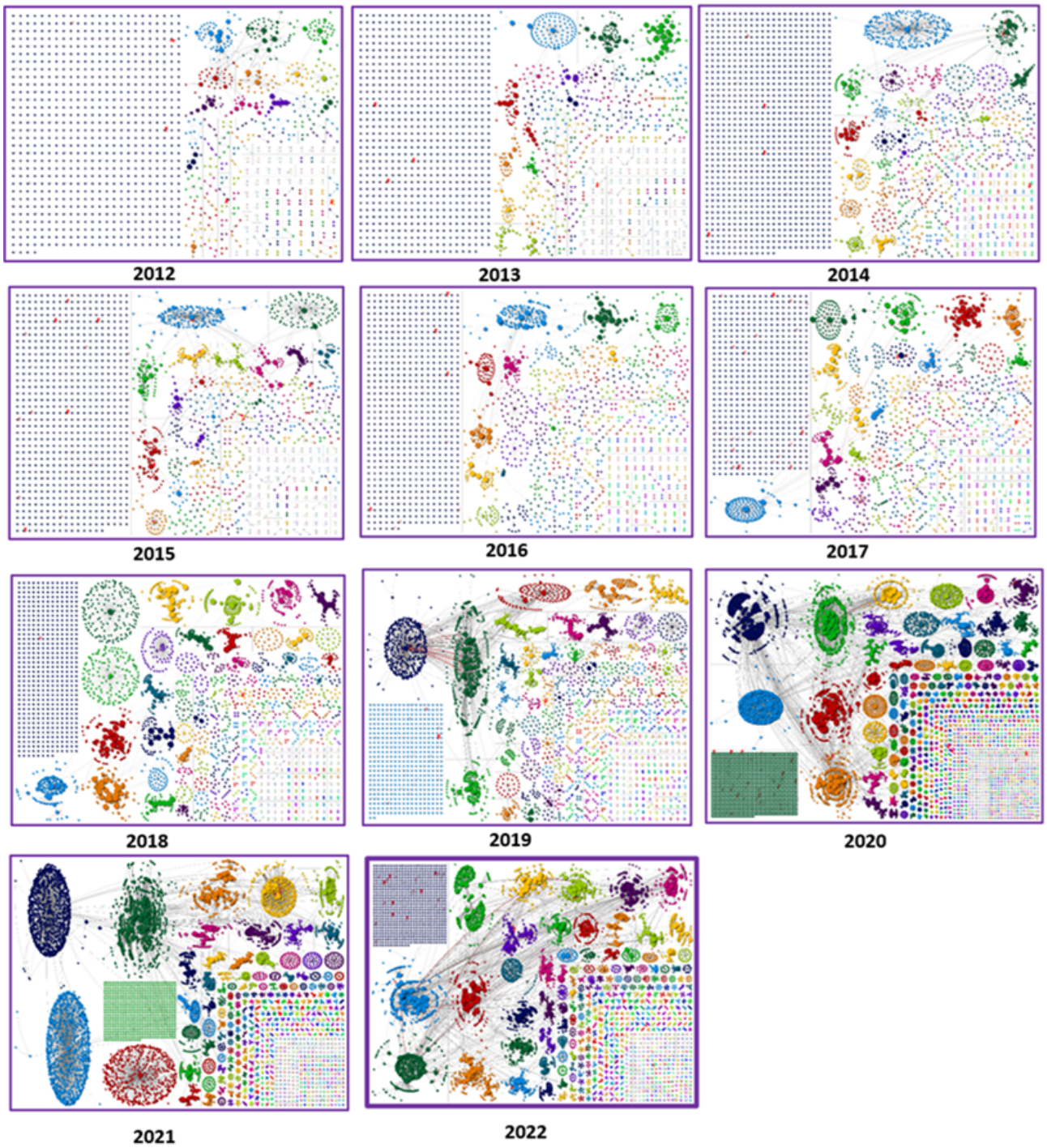

XSL•FO
RenderX

Figure 3 shows the volume of X interactions, that is, all posts, all forms of reposts, and replies. Figure 4 shows the number of users within the networks over this time. For many years (2012 to 2017), activity remained similar, with a range of posts sent between 4309 and 5533. There was a slight increase in 2018 to 8606 posts and a more significant increase in 2019 to 11,035 posts, with a sudden jump in 2020 to 40,508 and a peak of 77,178 in 2021. While biomedical publications have traditionally had access through professional gatekeepers, we found that social media serves as a platform for individuals to widen existing information resources and connect with a broader network of health influencers [22,23]. In this case, specific health influencers and users on platforms like X played a role in highlighting the bias and discrimination of pulse oximetry technology. They emphasized the lack of formal health information available on the issue, leading people to turn to news articles or seek clarification through discussion topics and trends on X [24].

Figure 3. Interactions on the X platform.

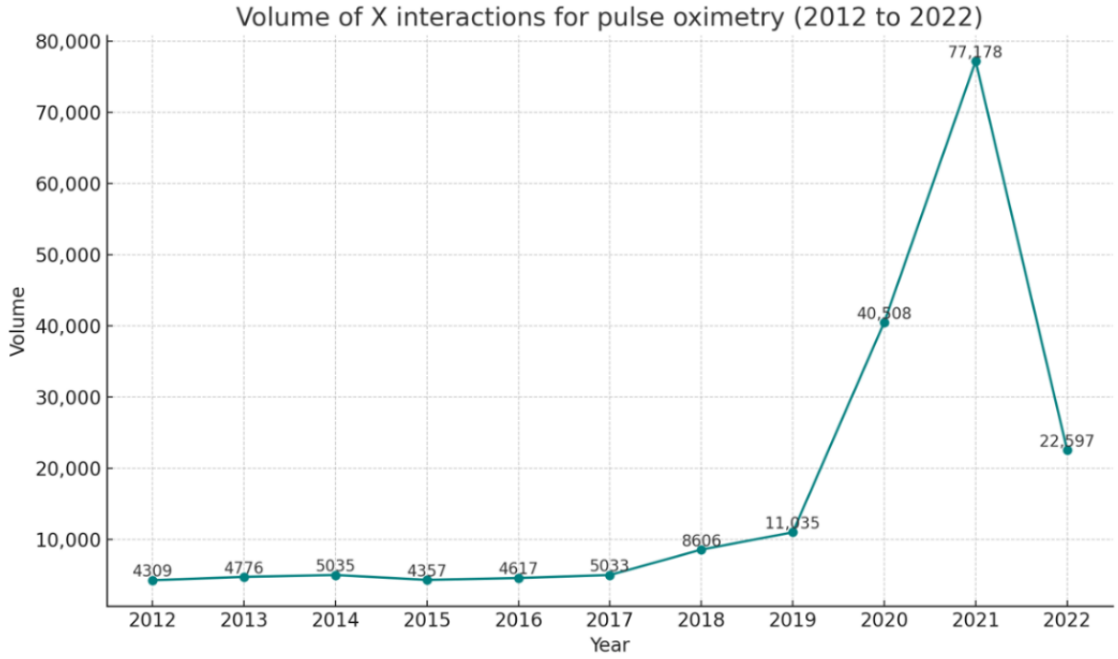

Figure 4. The volume of users posting.

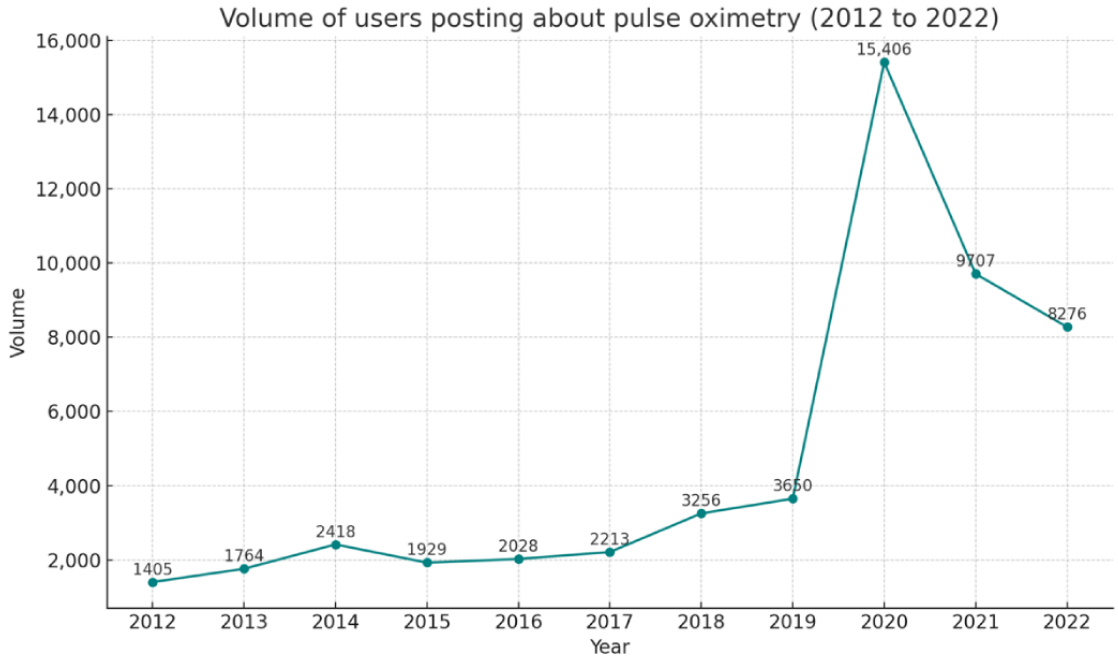

Figure 5 provides an overview of all literature indexed in PubMed for "Pulse Oximetry" between 1946 and 2020. When searching specifically for the search string of "pulse oximetry racial bias," PubMed highlights 2 papers published in 2020, a total of 10 papers in 2021, and 14 papers in 2022. These studies were widely shared and posted across X.

XSL•FO
RenderX

Figure 5. PubMed search for “pulse oximetry” (1946 to 2023).

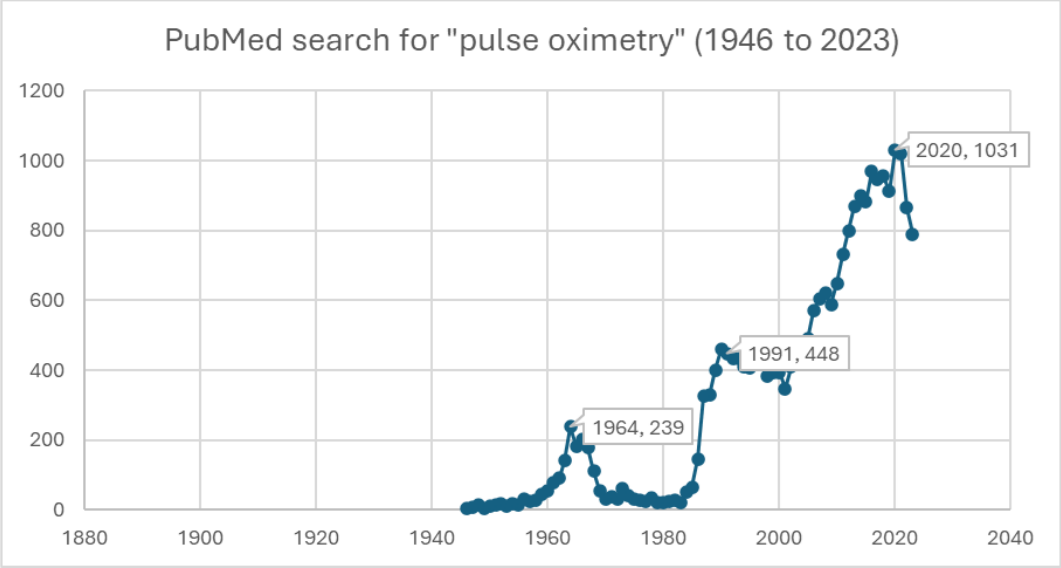

The relationship between pulse oximetry readings and skin pigmentation has been known within the medical community for an extended period, as evidenced by research studies such as Adler et al [25], Bickler et al, [26], and Sjoding et al [27]. However, the significant implications of this correlation on racial disparities in health care outcomes remained largely unexplored and unaddressed until a surge in social media discussions around racial issues in 2019-2020. While a limited number of studies have explored knowledge sharing and data in this field, Jamali et al [28] shed light on a critical issue underscoring the racial disparity in oxygen saturation measurements by pulse oximetry. Sudat et al [29] and Shi et al [30] report similar findings and racial disparities at the height of the COVID-19 pandemic. This heightened awareness coincided with an increase in scientific research focused on the topic, indicating a newfound recognition of the need to understand and reconcile these disparities in medical technology and its application. This shift marks an essential step in addressing and rectifying biases in medical diagnostics that disproportionately affect people of color. Figures 6 and 7, created by searching PubMed for published papers, highlight how the increase of peer-reviewed publications on pulse oximetry coincides with the increased social media activity on X.

Figure 6. PubMed searches for pulse oximetry and race-related keywords (1990 to 2023).

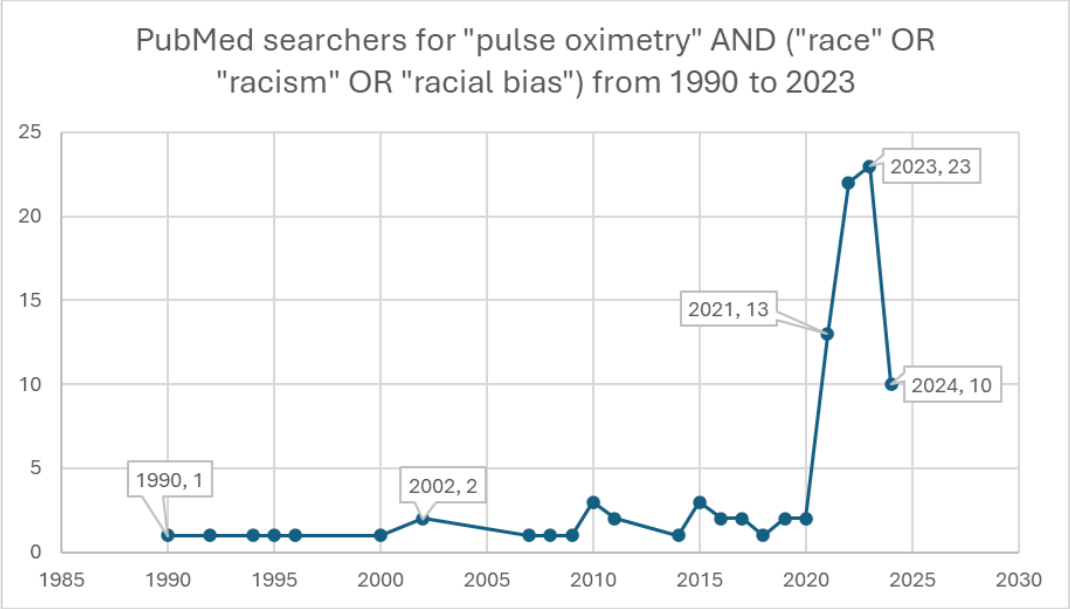

XSL•FO
RenderX

Figure 7. PubMed search for pulse oximetry and skin-related keywords (1987 to 2023).

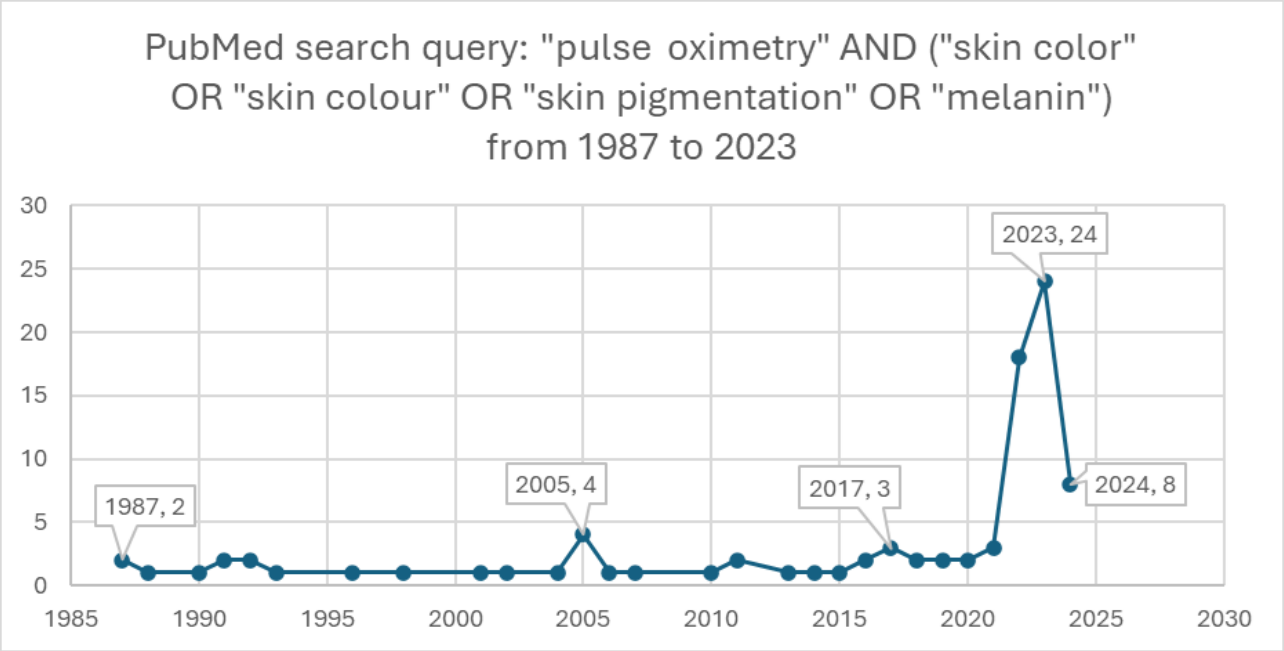

Examining Key Sources and Information Shared in 2020

Several key web sources were frequently shared during discussions surrounding racial bias in pulse oximetry, shown in Table 1. The table includes details of the title, publisher, and number of shares.

Table 1. Overview of key sources, publisher, and number of shares (year 2020).

Source title	Publisher	Number of shares
Racial Bias in Pulse Oximetry Measurement	New England Journal of Medicine	327
The Infection That's Silently Killing Coronavirus Patients	The New York Times	187
Takuo Aoyagi, an Inventor of the Pulse Oximeter, Dies at 84	The New York Times	57
Pulse oximetry to detect early deterioration of patients with COVID-19 in primary and community care settings	NHS[a] England	50
COVID Oximetry @home	NHS England	48

[a]NHS: National Health Service.

The paper titled "Racial Bias in Pulse Oximetry Measurement" [27], published by the New England Journal of Medicine, emerged as a leading reference with 327 shares. The New York Times was also featured prominently in the discourse, with its articles covering various newsworthy aspects around pulse oximetry and other general COVID-19–related articles. In addition, content from the National Health Service, England, that provided general information on pulse oximetry during the COVID-19 pandemic, was frequently cited. The 3 sources were the most repeatedly shared sources during discussions surrounding racial bias in pulse oximetry. Due to the frequency of the sources, they were considered the most authoritative and informative sources on the topic. The sources were identified through a thorough search of web-based discussions and social media posts about racial bias in pulse oximetry. The frequency of sharing was used as a metric to determine which sources were the most prominent and influential.

The discussions surrounding racial bias in pulse oximetry revealed certain recurrent word pairs summarized in Table 2 that shed light on the focal points of the conversation. Table 2 provides insight into the top word pairs used together alongside their frequency.

XSL•FO
RenderX

Table 2. Overview of most frequently occurring word pairs in 2020.

Top word pairs	Occurrences
covid,19	2462
darker,skin	1499
oximetry,key	1494
management,covid	1479
clinical,management	1478
instrument,clinical	1477
inaccurate,darker	1477
key,instrument	1477

The pairing "COVID-19" with pulse oximetry with 2462 occurrences reflects the increasing interest in the device in monitoring patients with COVID-19. The word combination "darker skin" at 1499 instances highlights concerns about the potential inaccuracies of pulse oximetry readings in individuals with darker skin tones. This is further supported by the pair "inaccurate, darker" with 1477 mentions, which directly addresses the issue of instrument precision with darker skin. Other word pairs appear to highlight the importance and centrality of pulse oximetry in certain contexts, hinting at its foundational role in medical procedures or monitoring. Other discourse appeared to tie the tool's clinical management capabilities, particularly in the management of patients with COVID-19. In the context of discussions on racial bias in pulse oximetry, several key influencers were identified based on betweenness centrality, which is summarized in Table 3. The table provides insight into the rank, name description, follower count, and location of the users who were most impactful.

Table 3. Overview of most influential users in 2020.

Influence rank	Name (on X)	Description	Number of followers	Location
1	Esther Choo, MD, MPH	Emergency physician and professor	201,890	United States
2	Matt Inada-Kim	National clinical director	2648	England
3	Ithan Peltan, MD, MSc	ICU[a] and pulmonary physician and associate professor	513	United States
4	Utibe R. Essien, MD, MPH	Physician-scientist and assistant professor of medicine	25,782	United States
5	Faheem Younus, MD	clinician and a certified physician executive	496,647	United States

[a]ICU: intensive care unit.

The most influential user was Dr Esther Choo (MD, MPH), an emergency physician and professor from the United States. With an impressive following of 201,890, she stands out as a dominant voice on the subject. Following her is Matt Inada-Kim, the national clinical director from England. Although he has a notably smaller audience, boasting only 2648 followers, his influence in the discourse remains significant. The third influencer is Ithan Peltan (MD, MSc), an intensive care unit and pulmonary physician and associate professor from the United States. He engages with a community of 513 followers.

Another notable figure is Dr Utibe R. Essien (MD, MPH), also from the United States. As a physician-scientist and assistant professor of medicine, he has garnered attention from 25,782 followers. Finally, Dr Faheem Younus (MD), a clinician and certified physician executive from the United States, brings a substantial audience into the fold with 496,647 followers. The presence and reach of these influencers shaped the narrative and discussions.

Discussion

Principal Findings

The preliminary findings showed that the use of general keywords such as "pulse oximetry," "Pulse Oximetry," or "Oximetry" yielded a substantial cohort of individuals who expressed concerns or shared information regarding the racial bias inherent in the output of the pulse oximeter device. This observation implies an escalating level of interest on this subject on the X platform. Such conversations will typically occur in biomedical research, but these conversations were occurring on social media among the public, highlighting that people turned to social media. In addition, it was observed that several notable users used their position to increase public consciousness regarding potential concerns associated with pulse oximeters which highlights the influence of leaders through engagement in social media. The rapid and global dissemination of these posts highlights the influential capacity of X as a platform for expeditious information sharing. As modern equivalents of the public square, the discussions on social media hold significant importance [18,20,21]. Our study identified and categorized the

XSL•FO
RenderX

primary clusters and users participating in this discourse. The tight shaping of the clusters demonstrates how users can form communities around topics of interest and that they can work together to help debunk issues and bias.

Whether the information on social media related to pulse oximetry and its racial bias was enough to change people's behaviors in increasing conversation and advocacy around the topic is a complex question. Understanding the shape and types of influence among health networks on social media is essential for identifying the challenges of using social media for health communication. Abroms et al [31] acknowledge the potential of social media for improving public health while raising concerns about its negative impact on physical and mental well-being and its role in spreading misinformation. Relatedly, there is growing interest in the influence of professional health workers posting on social media. Engebretsen [32] emphasizes the importance of health practitioners maintaining a balance between establishing social proximity and maintaining professional detachment while disseminating health guidance on social media platforms such as Instagram (Meta). In addition, engaging responsible professional health care practitioners in disseminating valid and accurate information through social networks is essential to addressing the concerns raised by Abroms et al [31].

Here, the lack of formal access to health policy has forced health advocacy movements to advocate on a global and cross-platform level. Many social media users typically prefer 1 or 2 platforms, which means they may risk missing crucial health information. Vance et al [33] acknowledged the growing importance of social media platforms as sources of health information, especially among adolescents and young adults. However, they also emphasized the drawbacks of anonymous authorship and the dissemination of subjective viewpoints as objective facts. Pradekso et al's [34] study highlights the importance of understanding the distinct health misinformation concerns linked with different types of social media.

Our findings highlight X's potential as a platform for raising awareness for important health issues, even among those who may not be traditionally engaged with scientific research. We emphasize how health experts can use platforms like X to target specific communities, partnering with influential users and tailoring content to trending topics to appeal to different audiences.

The pandemic represented a significant inflection point for discussion in the public domain about racial bias in pulse oximetry. In response to the early stages of the pandemic, there was an increase in scientific papers concerning oximetry. The recent surge in research on racial bias in pulse oximetry stems from several factors. First, the COVID-19 pandemic reignited concerns about racial disparities in health care outcomes. This renewed focus exposed biases embedded in the design of medical devices [35], particularly in home-based products like pulse oximeters [36]. Second, the increased availability of data on pulse oximetry readings and patient outcomes provided researchers with real-world evidence of these biases [30,37]. Finally, heightened public and scientific interests in the topic attracted additional funding for health research efforts [38]. With the increased number of studies on racial bias after 2020, we can speculate that the rigor of the studies also improved due to the availability of data, the development of new analytical methods, and the growing number of researchers working in this area.

Social media platforms like X allowed researchers, health experts, patients, and the general public to share information quickly and widely, raising awareness about the potential for racial bias, finding other people interested in the topics and generating discussions for further research. Our study, therefore, has some limitations. First, selection bias is limited to a single sample of posts on one platform. Further research encompassing different platforms and related terms is needed to confirm our findings. Second, we did not collect demographic data on the users in our sample, so we cannot say who is most likely to be exposed to information about racial bias in pulse oximeters. Third, we did not perform a sentiment analysis or qualitative analysis of the posts to show advocacy-related follow-up, so we cannot say how users respond to this information. Fourth, we could not objectively report how these could have tied into increasing scientific discourse on the topic or heightened the need for action for regulatory authorities such as the FDA. Fifth, a deeper understanding of the context in which race and oximetry are being discussed is not there, as our study looks at activity generated by posts. However, our future work aims to better understand the context of the discussions. Finally, the reliance on social media data presents another limitation. Platforms such as X, while offering rapid access to health information, have their own limitations. Concerns exist regarding potential biases within platform algorithms, as evidenced by Savolainen's (27) work on algorithmic governance and user struggles with platform expectations. This highlights users' potential difficulty in accessing valid health information on social media.

Notwithstanding these constraints, our research offers a crucial foundation for understanding the function of X and other social media platforms in disseminating health-related information about potential biases in health devices such as pulse oximeters. Social events are invariably discussed in social media–fueled scientific discourse on a topic known to the scientific community for decades but only researched in relation to health care disparities after social media raised awareness. Future research should aim to expand upon our study by using more comprehensive and diverse sampling techniques, gathering user demographic information, and using qualitative analysis of posts to gain a deeper understanding of user reactions. Furthermore, research by Philip et al [9] during the pandemic emphasized the need for a critical re-evaluation of pulse oximetry technology. Their work highlighted concerns about racial bias in pulse oximetry readings and proposed modifications to account for patient skin color variations. In response to data bias, Ferrari et al [39] advocated for implementing multi-wavelength pulse oximeters that incorporate melanin absorption correction to enhance accuracy, particularly in the context of the COVID-19 pandemic. These studies parallel the concerns we observed in social media discussions, further highlighting the need for continued research and technological improvements.

XSL•FO
RenderX

Conclusion

Our analysis of social media discussions revealed a significant increase in public interest surrounding racial disparities in pulse oximetry readings. These findings highlight the need for increased public awareness and education about pulse oximetry limitations. In addition, our study provides valuable insights for future research efforts focused on improving pulse oximetry technology and addressing identified biases. Ultimately, ongoing collaboration between health care professionals and the public is crucial for ensuring the fair and accurate use of pulse oximetry for all patients. Furthermore, this study demonstrates the potential of social media platforms, such as X, as a space for disseminating academic research about racial bias in pulse oximetry. We observed key influential users leveraging their platforms to raise awareness of these issues and rapidly disseminate early scientific findings. These results offer valuable insights for raising awareness of critical health issues and understanding health information propagation on social media.

Our analysis suggests that social media can significantly enhance health information sharing and advocacy strategies, particularly for nonhealth professionals. This study highlights the role of key influencers and network links on social media in shaping health information dissemination. These influencers and network connections crucially impact how health organizations use social media and how individuals engage in health social movements like advocacy. While traditional, offline health information channels aimed at patient education remain important, our findings suggest that influence in the health sphere is increasingly intertwined with, or even dependent upon, the presence of key actors on digital platforms.

Conflicts of Interest

AS serves as the social media Editor for Critical Care Medicine and Neurocritical care.

References

1. France MW, Porter AH, Bennett CJ, McCarthy KL. Use of home pulse oximetry to enhance remote COVID-19 monitoringvaluating a single centre experience. Search Life-Sciences Literature 2022. [doi: 10.1101/2022.11.07.22281915]
2. Michard F, Shelley K, L'Her E. COVID-19: pulse oximeters in the spotlight. J Clin Monit Comput 2021;35(1):11-14 [FREE Full text] [doi: 10.1007/s10877-020-00550-7] [Medline: 32578070]
3. Luks AM, Swenson ER. Pulse oximetry for monitoring patients with COVID-19 at home. Potential pitfalls and practical guidance. Ann Am Thorac Soc 2020;17(9):1040-1046 [FREE Full text] [doi: 10.1513/AnnalsATS.202005-418FR] [Medline: 32521167]
4. Valbuena VSM, Merchant RM, Hough CL. Racial and ethnic bias in pulse oximetry and clinical outcomes. JAMA Intern Med 2022;182(7):699-700. [doi: 10.1001/jamainternmed.2022.1903] [Medline: 35639397]
5. Cabanas AM, Martín-Escudero P, Shelley KH. Improving pulse oximetry accuracy in dark-skinned patients: technical aspects and current regulations. Br J Anaesth 2023;131(4):640-644. [doi: 10.1016/j.bja.2023.07.005] [Medline: 37544838]
6. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2020;383(25):2477-2478 [FREE Full text] [doi: 10.1056/NEJMc2029240] [Medline: 33326721]
7. Palmer SJ. Racial bias in pulse oximetry: a significant problem in healthcare. Br J Health Assist Mark Allen Group 2022;16(9):430-433. [doi: 10.12968/bjha.2022.16.9.430]
8. Patwari N, Huang D, Bonetta-Misteli K. Short: racial disparities in pulse oximetry cannot be fixed with race-based correction. In: Electrical Engineering and Systems Science Signal Processing. 2024 Presented at: 2023 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE); June 2023; Orlando, FL, USA URL: https://ieeexplore.ieee.org/document/10183754 [doi: 10.1145/3580252.3586986]
9. Philip KEJ, Tidswell R, McFadyen C. Racial bias in pulse oximetry: more statistical detail may help tackle the problem. BMJ 2021;372:n298. [doi: 10.1136/bmj.n298] [Medline: 33531354]
10. Plaisime MV. Invited commentary: undiagnosed and undertreated-the suffocating consequences of the use of racially biased medical devices during the COVID-19 pandemic. Am J Epidemiol 2023;192(5):714-719 [FREE Full text] [doi: 10.1093/aje/kwad019] [Medline: 36702767]
11. Cabanas AM, Fuentes-Guajardo M, Latorre K, León D, Martín-Escudero P. Skin pigmentation influence on pulse oximetry accuracy: a systematic review and bibliometric analysis. Sensors (Basel) 2022;22(9):3402 [FREE Full text] [doi: 10.3390/s22093402] [Medline: 35591092]
12. Colvonen PJ. Response to: investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2021;4(1):38 [FREE Full text] [doi: 10.1038/s41746-021-00408-5] [Medline: 33637822]
13. Sikka T. The colour of technology: COVID-19, race, and the pulse oximeter. In: Disrupted Knowledge. United Kingdom: Brill Press; 2023:55-77.
14. Ahmed W, Vidal-Alaball J, Downing J, López Seguí F. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. J Med Internet Res 2020;22(5):e19458 [FREE Full text] [doi: 10.2196/19458] [Medline: 32352383]
15. Ahmed W, López Seguí F, Vidal-Alaball J, Katz MS. COVID-19 and the "Film Your Hospital" conspiracy theory: social network analysis of Twitter data. J Med Internet Res 2020;22(10):e22374 [FREE Full text] [doi: 10.2196/22374] [Medline: 32936771]

XSL•FO
RenderX

16. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys 2004;70(6 Pt 2):066111. [doi: 10.1103/PhysRevE.70.066111] [Medline: 15697438]
17. Freeman LC. A set of measures of centrality based on betweenness. Sociometry JSTOR 1977;40(1):35. [doi: 10.2307/3033543]
18. Smith MA, Rainie L, Shneiderman B, Himelboim I. Mapping Twitter topic networks: From polarized crowds to community clusters. Pew Research Center. 2014. URL: https://www.pewresearch.org/internet/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/
19. Papacharissi Z. The virtual sphere: The internet as a public sphere. New Media & Society 2002;4(1):9-27. [doi: 10.1177/14614440222226244]
20. Cowls J, Ma C. The internet's town square? A critical discourse analysis of Parler's "Past Imaginary". Soc Media Soc 2023;9(2). [doi: 10.1177/20563051231181099]
21. Agudelo-Hernández F, Porras LFC, Álvarez ABG. Declaration of the town square: the urgency of speaking as one. J Hum Rights Pract 2024;16(2):624-637. [doi: 10.1093/jhuman/huae002]
22. Gupta S, Dash SB, Mahajan R. The role of social influencers for effective public health communication. OIR 2021;46(5):974-992. [doi: 10.1108/oir-01-2021-0012]
23. Lutkenhaus RO, Jansz J, Bouman MP. Tailoring in the digital era: stimulating dialogues on health topics in collaboration with social media influencers. Digit Health 2019;5:2055207618821521 [FREE Full text] [doi: 10.1177/2055207618821521] [Medline: 30729023]
24. Lim MSC, Molenaar A, Brennan L, Reid M, McCaffrey T. Young adults' use of different social media platforms for health information: insights from web-based conversations. J Med Internet Res 2022;24(1):e23656 [FREE Full text] [doi: 10.2196/23656] [Medline: 35040796]
25. Adler JN, Hughes LA, Vivilecchia R, Camargo CA. Effect of skin pigmentation on pulse oximetry accuracy in the emergency department. Acad Emerg Med 1998;5(10):965-970 [FREE Full text] [doi: 10.1111/j.1553-2712.1998.tb02772.x] [Medline: 9862586]
26. Bickler PE, Feiner JR, Severinghaus JW. Effects of skin pigmentation on pulse oximeter accuracy at low saturation. Anesthesiology 2005;102(4):715-719 [FREE Full text] [doi: 10.1097/00000542-200504000-00004] [Medline: 15791098]
27. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2021;385(26):2496. [doi: 10.1056/NEJMx210003] [Medline: 34936755]
28. Jamali H, Castillo LT, Morgan CC, Coult J, Muhammad JL, Osobamiro OO, et al. Racial disparity in oxygen saturation measurements by pulse oximetry: evidence and implications. Ann Am Thorac Soc 2022;19(12):1951-1964. [doi: 10.1513/AnnalsATS.202203-270CME] [Medline: 36166259]
29. Sudat SEK, Wesson P, Rhoads KF, Brown S, Aboelata N, Pressman AR, et al. Racial disparities in pulse oximeter device inaccuracy and estimated clinical impact on COVID-19 treatment course. Am J Epidemiol 2023;192(5):703-713 [FREE Full text] [doi: 10.1093/aje/kwac164] [Medline: 36173743]
30. Shi C, Goodall M, Dumville J, Hill J, Norman G, Hamer O, et al. The accuracy of pulse oximetry in measuring oxygen saturation by levels of skin pigmentation: a systematic review and meta-analysis. BMC Med 2022;20(1):267 [FREE Full text] [doi: 10.1186/s12916-022-02452-8] [Medline: 35971142]
31. Abroms LC, Gold RS, Allegrante JP. Promoting health on social media: the way forward. Health Educ Behav 2019;46(2_suppl):9-11. [doi: 10.1177/1090198119879096] [Medline: 31742461]
32. Engebretsen M. Communicating health advice on social media: a multimodal case study. MedieKultur: Journal of Media and Communication Research, 2023;39(74):164-184. [doi: 10.7146/mk.v39i74.134085]
33. Vance K, Howe W, Dellavalle RP. Social internet sites as a source of public health information. Dermatol Clin 2009;27(2):133-136, vi. [doi: 10.1016/j.det.2008.11.010] [Medline: 19254656]
34. Pradekso T, Manalu SR, Setyabudi D. Social Media Use and Exposure to Health Related False Information. 2020 Presented at: 4th International Conference on Indonesian Social and Political Enquiries; 21-22 October 2019; Semarang, Central Java, Indonesia. [doi: 10.4108/eai.21-10-2019.2294449]
35. Mousavi SS, Reyna MA, Clifford GD, Sameni R. A survey on blood pressure measurement technologies: addressing potential sources of bias. Sensors (Basel) 2024;24(6):1730 [FREE Full text] [doi: 10.3390/s24061730] [Medline: 38543993]
36. Alboksmaty A, Beaney T, Elkin S, Clarke JM, Darzi A, Aylin P, et al. Effectiveness and safety of pulse oximetry in remote patient monitoring of patients with COVID-19: a systematic review. Lancet Digit Health 2022;4(4):e279-e289 [FREE Full text] [doi: 10.1016/S2589-7500(21)00276-4] [Medline: 35337644]
37. Philip KEJ, Bennett B, Fuller S, Lonergan B, McFadyen C, Burns J, et al. Working accuracy of pulse oximetry in COVID-19 patients stepping down from intensive care: a clinical evaluation. BMJ Open Respir Res 2020;7(1):e000778 [FREE Full text] [doi: 10.1136/bmjresp-2020-000778] [Medline: 33361436]
38. Sauermann H, Vohland K, Antoniou V, Balázs B, Göbel C, Karatzas K, et al. Citizen science and sustainability transitions. Res Policy 2020;49(5):103978. [doi: 10.1016/j.respol.2020.103978]
39. Ferrari M, Quaresima V, Scholkmann F. Pulse oximetry, racial bias and statistical bias: further improvements of pulse oximetry are necessary. Ann Intensive Care 2022;12(1):19 [FREE Full text] [doi: 10.1186/s13613-022-00992-z] [Medline: 35190907]

XSL•FO
RenderX

Abbreviations

API: application programming interface
FDA: US Food and Drug Administration
NodeXL: Network Overview Discovery and Exploration for Excel

Edited by A Mavragani; submitted 03.01.24; peer-reviewed by K Sellen, A Mittal, M Katz; comments to author 29.02.24; revised version received 27.03.24; accepted 25.07.24; published 08.10.24.

Please cite as:
Ahmed W, Hardey M, Winters BD, Sarwal A
Racial Biases Associated With Pulse Oximetry: Longitudinal Social Network Analysis of Social Media Advocacy Impact
J Med Internet Res 2024;26:e56034
URL: https://www.jmir.org/2024/1/e56034
doi:10.2196/56034
PMID:

©Wasim Ahmed, Mariann Hardey, Bradford David Winters, Aarti Sarwal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Selling Misleading "Cancer Cure" Books on Amazon: Systematic Search on Amazon.com and Thematic Analysis

Marco Zenone[1], PhD; May van Schalkwyk[1], PhD; Greg Hartwell[1], PhD; Timothy Caulfield[2], LLB, LLM; Nason Maani[3], PhD

[1]Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, United Kingdom
[2]Health Law Institute, Faculty of Law, University of Alberta, Edmonton, AB, Canada
[3]Global Health Policy Unit, The University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Marco Zenone, PhD
Faculty of Public Health and Policy
London School of Hygiene & Tropical Medicine
Keppel St
London, WC1E 7HT
United Kingdom
Email: marco.zenone@lshtm.ac.uk

Abstract

Background: While the evidence base on web-based cancer misinformation continues to develop, relatively little is known about the extent of such information on the world's largest e-commerce website, Amazon. Multiple media reports indicate that Amazon may host on its platform questionable cancer-related products for sale, such as books on purported cancer cures. This context suggests an urgent need to evaluate Amazon.com for cancer misinformation.

Objective: This study sought to (1) examine to what extent are misleading cancer cure books for sale on Amazon.com and (2) determine how cancer cure books on Amazon.com provide misleading cancer information.

Methods: We searched "cancer cure" on Amazon.com and retrieved the top 1000 English-language book search results. We reviewed the books' descriptions and titles to determine whether the books provided misleading cancer cure or treatment information. We considered a book to be misleading if it suggested scientifically unsupported cancer treatment approaches to cure or meaningfully treat cancer. Among books coded as misleading, we conducted an inductive latent thematic analysis to determine the informational value the books sought to offer.

Results: Nearly half (494/1000, 49.4%) of the sampled "cancer cure" books for sale on Amazon.com appeared to contain misleading cancer treatment and cure information. Overall, 17 (51.5%) out of 33 Amazon.com results pages had 50% or more of the books coded as misleading. The first search result page had the highest percentage of misleading books (23/33, 69.7%). Misleading books (n=494) contained eight themes: (1) claims of efficacious cancer cure strategies (n=451, 91.3%), (2) oversimplifying cancer and cancer treatment (n=194, 39.3%), (3) falsely justifying ineffective treatments as science based (n=189, 38.3%), (4) discrediting conventional cancer treatments (n=169, 34.2%), (5) finding the true cause of cancer (n=133, 26.9%), (6) homogenizing cancer (n=132, 26.7%), (7) discovery of new cancer treatments (n=119, 24.1%), and (8) cancer cure suppression (n=82, 16.6%).

Conclusions: The results demonstrate that misleading cancer cure books are for sale, visible, and prevalent on Amazon.com, with prominence in initial search hits. These misleading books for sale on Amazon can be conceived of as forming part of a wider, cross-platform, web-based information environment in which misleading cancer cures are often given prominence. Our results suggest that greater enforcement is needed from Amazon and that cancer-focused organizations should engage in preemptive misinformation debunking.

(J Med Internet Res 2024;26:e56354) doi:10.2196/56354

KEYWORDS

cancer; Amazon; misinformation; e-commerce; cancer cure; cancer misinformation; misleading; cancer information; treatment; cancer treatment; thematic analysis; misleading; online information

XSL•FO
RenderX

Introduction

Following a cancer diagnosis, patients with cancer understandably turn to web-based sources to look for information on treatment, to seek emotional support [1], or to learn from the experiences of other patients with cancer [2]. Unfortunately, the web-based informational cancer environment they navigate is often polluted with inaccurate and harmful cancer information. A recent survey of patients with cancer found that over 55.9% reported seeing cancer misinformation on the web and 25% received advice on alternative cancer cures [3], which are associated with worse outcomes [4,5]. Cancer treatment misinformation is prevalent on social media platforms [6] such as YouTube [7,8], TikTok [9], Facebook [10], Twitter (rebranded as X) [11], and Pinterest [12] and has been found to be amplified by search algorithms [13]. Multiple studies suggest that misleading cancer information receives more engagement than factual sources [6,14] and that cancer misinformation is partially driven by financial interests and incentives [15,16]. Web-based cancer misinformation unfairly burdens patients with cancer and those close to them, who are forced to evaluate and navigate complex, "believable" yet false cancer information at a time of emotional and psychological strain [17].

Amazon is the world's largest e-commerce website. Each day, tens of thousands of products are sold on the platform to Amazon's 300+ million users [18]. Amazon assists its users to sell, advertise, or purchase products in an efficient and convenient process with few barriers [19]. As of October 2023, a search for "cancer" on Amazon (Amazon.com) returns over 100,000 search results [20]. Amazon purports to restrict what types of products can be sold to protect its users from deceptive advertising or products and comply with the laws and regulations of the countries in which it operates. For example, Amazon prohibits product claims for any disease that do not have the approval of the US Food and Drug Administration (FDA) [21]. This policy would likely disallow any unapproved products stating they can cure, prevent, or meaningfully treat cancer.

However, in spite of such policies existing, multiple media reports document sellers on Amazon listing dangerous and "misleading" [22] cancer products. Audits have found Amazon to promote antivaccination books [23]. Media reporting suggests that Amazon hosted COVID-19 conspiracy documentaries [24] and other COVID-19 misinformation [25]. Related to cancer, it has been suggested that Amazon gives recommendations for books advising unproven or disproven cancer cures [26,27]. At the time of writing, the number one best seller under "cancer" on Amazon is a book that describes how a person with cancer healed himself naturally after refusing chemotherapy following surgery [28]. Media reports note that the book presents a misleading narrative, in which the author opted out of curative chemotherapy when he in fact opted out of adjuvant chemotherapy [29,30]. The author of the book has been demonetized on YouTube [31] but still appears to earn commissions and sell cancer-related books, anticancer supplements, and immune support products on Amazon [32].

This context suggests an immediate need to evaluate Amazon for cancer misinformation. The tools offered by Amazon to its sellers may enable the sale of books promoting scientifically unsupported cancer treatments and cures. Evaluating alleged cancer cure books for sale on Amazon is important to determine the information patients with cancer may view and purchase. Misleading cancer cure information may lead to dangerous treatment decisions [33] and undermine patients' confidence in evidence-based treatments. Therefore, this study sought to (1) examine to what extent are misleading cancer cure books for sale on Amazon.com and (2) determine how cancer cure books on Amazon.com provide misleading cancer information. The results can provide valuable information to public health authorities and external regulators, to ensure that patients with cancer are not financially exploited to purchase books with harmful information.

Methods

Data Collection

We searched "cancer cure" on Amazon.com (10,000+ search results) and retrieved the top 1000 English-language book search results, including sponsored books, on August 3, 2023. Similar to other studies, we performed the search without logging into the platform to avoid the search biases of an existing profile [34]. The Amazon.com search function matches customers to products that fit their typed description. Our search mimics the searches of patients with cancer who may be seeking cancer cure information on Amazon.com. To collect the data, MZ created 2 Data Miner scrapers [35] that retrieved the product information of each result, including its title, author, description, rating score, number of ratings, price, year published, number of pages, publisher-supplied "about the author" information, and the page number of the Amazon.com search result that the book appeared in. Data Miner is a program that allows its user to create simple scrapers that automatically collect certain fields from a web page. The first Data Miner scraper collected the URLs of each result listed from the "cancer cure" Amazon.com search. The second Data Miner scraper was created to collect the aforementioned web page fields (eg, book name) from each of the URLs collected from the first scrape.

Designating a Book as Misleading

MZ reviewed the books' descriptions and titles to determine whether the books provided misleading cancer cure or treatment information. As the books are for sale and their titles and descriptions represent their value to a potential customer, we considered this information appropriate to determine whether the books for sale are misleading or not. We considered a book to be misleading if it suggested scientifically unsupported cancer treatment approaches that cure or meaningfully treat cancer. To determine which books are misleading, we consulted a list of alternative and natural treatments named in a study examining which alternative cancer treatments patients use [36], along with the National Cancer Institute's list of alternative and complementary cancer treatments [37]. While it is not possible to identify every unproven cancer cure, our sources contain the standard and most well-known examples. We then applied the lists against the results. We additionally reviewed books for misleading cues. This included overt statements that are contrary to scientific or clinical consensus and present conspiratorial

XSL•FO
RenderX

narratives. To ensure consistent inclusion decisions, TC blindly audited 10% of inclusion decisions. TC found no disagreements, demonstrating high agreement and consistent coding. TC additionally reviewed the code applications MZ marked as requiring a second opinion.

Analysis

We conducted a qualitative, inductive latent thematic analysis [38] to determine the informational value the books sought to offer to potential customers. We followed the thematic analysis steps outlined by Clarke and Braun [39], consisting of data familiarization, generation of initial codes, searching for themes, reviewing themes, defining and naming themes, and reporting. To increase the trustworthiness of the analysis, the authors adopted the recommendations of Nowell et al [40]. MZ, GH, TC, and NM reviewed the data, undertook data review, and iteratively drafted and redrafted a coding frame until a final frame captured themes. MZ coded the data. GH audited nearly half (236/494, 47.8%) of thematic book coding, finding agreement with 95.8% (1,808/1,888) of code decisions, signifying consistent and accurate coding. All disagreements were resolved through discussion. GH reviewed the thematic code applications MZ marked as requiring a second opinion. Our study followed the Standards for Qualitative Research Reporting guidelines (Multimedia Appendix 1) [41].

Ethical Considerations

This study did not require ethical approval because all data collected and analyzed are publicly available and posted without the expectation of privacy.

Results

Frequency of Misleading Cancer Cure Books

Our analysis found that 49.4% (494/1000) of the sampled “cancer cure” books for sale on Amazon.com contained misleading cancer treatment or cure information. Misleading books were consistent throughout the search result pages (Figure 1), with 17 (51.5%) out of 33 Amazon.com results pages having 50% or more of the books coded as misleading. Results in the first 3 pages contained the highest percentage of misleading books (first page: 23/33, 69.7%; second page: 23/33 69.7%; third page: 22/33, 66.7%). The lowest percentage of misleading books across any page found was 35.5% (11/31).

Figure 1. Percentage of misleading cancer books on Amazon.com by search result page number.

Misleading Cancer Book Themes

Overview

Misleading books (n=494) contained eight themes: (1) claims of efficacious cancer cure strategies (n=451, 91.3%), (2) oversimplifying cancer and cancer treatment (n=194, 39.3%), (3) falsely justifying ineffective treatments as science based (n=189, 38.3%), (4) discrediting conventional cancer treatments (n=169, 34.2%), (5) finding the true cause of cancer (n=133, 26.9%), (6) homogenizing cancer (n=132, 26.7%), (7) discovery of new cancer treatments (n=119, 24.1%), and (8) cancer cure suppression (n=82, 16.6%). Their frequencies and definitions are shown in Table 1.

XSL•FO
RenderX

Table 1. Misleading cancer books themes by frequency and illustrative example.

Theme	Frequency (n=494), n (%)	Definition
Claims of efficacious cancer cure strategies	451 (91.3)	• Treatment can or has cured cancer
Oversimplifying cancer and cancer treatment	194 (39.3)	• Cancer is not complicated and is simple to understand, treat, and cure. Cancer can be treated using simple foods or at-home remedies
Falsely justifying ineffective treatments as science based	189 (38.3)	• Ineffective cancer cures are scientifically proven to work • Justified by the author's own academic research • Contains unsupported conclusions or demonstrates a lack of understanding of scientific processes and evidence thresholds
Discrediting conventional cancer treatments	169 (34.2)	• Critical descriptions of chemotherapy, radiation, surgery, "Western medicine," and conventional treatments, including side effects, causing harm (poisoning, worsening cancer, and killing patients with cancer), and lack of effectiveness
Finding the true cause of cancer	133 (26.9)	• Treating cancer requires understanding the root cause of cause, why cancer occurs, and presentations of paradigm shifts on what's known to cause cancer
Homogenizing cancer	132 (26.7)	• One treatment can cure all types of cancer and/or other diseases
Discovery of new cancer treatments	119 (24.1)	• People with cancer, their family, or health practitioners finding new cancer cures and treatment options
Cancer cure suppression	82 (16.6)	• Cancer cures already exist but are hidden or banned because of financial interests from the pharmaceutical industry, legal battles with regulators or medical associations, and conspiracies to ruin reputations

Theme 1: Claims of Efficacious Cancer Cure Strategies

The most observed theme, claims of efficacious cancer cure strategies (n=451, 91.3%), referred to books giving cancer treatment information that is alleged to cure or meaningfully treat cancer. Here, language that directly posits efficacy for cancer treatment is used. For example, a cookbook with a title indicating that a specific diet can cure cancer states the following: "Approved, tested, and trusted recipes to reverse, prevent and cure cancer diseases completely." These books provided efficacy assurances demonstrating supposed effectiveness.

Theme 2: Oversimplifying Cancer and Cancer Treatment

How to treat, cure, and think about cancer was oversimplified in 194 (39.3%) books. Books claimed that cancer is not complicated to treat but rather is simple to understand, treat, and cure. Books suggested simple remedies capable of curing cancer. For example, a book description states that it is possible to cure cancer solely using carrot juice: "In 2013, [anonymized] cured stage 4 colon cancer without chemotherapy or radiation by drinking carrot juice…[anonymized] wanted to find out why such an apparently simple cancer cure – just carrots – works." Books also provided unrealistic timelines for how long simple cures or treatment approaches would take to cure cancer. For example, another book states that it is possible to cure cancer within a week to a month using anticancer remedies: "The ultimate cure for bone cancer is here, and it is not as complicated to be found as we have been made to think…cure your cancer with anticancer remedies within a month or maybe within a week." Books simplifying cancer present an inaccurate portrayal of complex diseases and treatment decisions, including consideration of a person's comorbidities and current medications.

Theme 3: Falsely Justifying Ineffective Treatments as Science Based

Relatedly, books used scientific-sounding language to justify that their outlined treatments are suitable and efficacious for primary cancer treatment (n=189, 38.3%). Authors described their own research and presented unsupported conclusions that demonstrate a lack of understanding or appreciation of the scientific process and evidence standards. For example, a book with a title indicating that their advocated cancer therapy is scientifically proven states the following: "The [anonymized book title] reveals a scientific, clinically proven natural therapy for healing cancer and creating optimal health." The misuse of scientific language and standards inaccurately portrays misleading treatment as being effective for cancer and misrepresents established scientific consensuses.

Theme 4: Discrediting Conventional Cancer Treatments

Perceived inadequacies of scientifically supported cancer treatments and alternatives to them were found in 169 (34.2%) results. Books provided critical descriptions of chemotherapy, radiation, surgery, "Western medicine," and other conventional cancer treatments, including their side effects; unsupported statements on the lack of effectiveness; and allegations that they cause unjustifiable harms such as poisoning, worsening of cancer, and death. For example, a book promoting alternative cancer treatments states the following: "cancer patients are still given questionable chemotherapies, which are proven to have no effect at all in the case of most cancer…why are cancer patients treated with a healing method that triggers this disease."

XSL•FO
RenderX

Books represented their treatments as filling the gaps of inadequate cancer care. For example, a book description states that "death is blamed on cancer [after] chemotherapy' rather than deception and inappropriate treatments…" Books alleging the inadequacies of conventional cancer treatments disparage best treatment practices and undermine health providers.

Theme 5: Finding the True Cause of Cancer

Books argued that curing cancer requires an understanding of the true cause of cancer and why cancer occurs (n=133, 26.9%). At times, books linked to previously established risk factors for cancers, such as processed foods, but misleading books suggested oversimplified, exaggerated, or unsupported reasons why cancer develops. For example, a book suggests that cancer may be a fungus or that cancer is caused by fungi: "It's been said that cancer is a fungus…while the verdict is still out cancer is a fungus, here is indisputable proof that some common fungi, found in foods and certain pharmaceuticals, cause cancer."

Theme 6: Homogenizing Cancer

The homogenization of cancer occurred in 132 (26.7%) books. One treatment was represented to cure all types of cancer and/or other diseases. For example, a book description states the following: "One of the key conclusions that we reached through our individual journeys was that whether you are a sufferer of superficial urinary bladder cancer, or any other condition, the same protocol that we used to heal will apply to you." In addition to homogenizing cancer types, homogenization of the cause of cancer and other illnesses also occurred. For example, a book giving advice on how to reverse breast cancer through kidney filtration illustrated both this and theme 5 together when stating the following: "We used the same protocols irrespective of condition name as we had found the root cause of all conditions/disease to be the same." Homogenizing cancer treatments for all cancers or other diseases provides misleading information and ignores the complexities of cancer care.

Theme 7: Discovery of New Cancer Treatments

Discovery of and creation of new alleged cancer cures were seen in 119 (24.1%) books. Here, authors described how they or the person they're writing about developed or found a new cancer treatment protocol. The motivation for the discovery was typically due to a personal situation, such as the author themselves or a family member having an incurable form of cancer, or a medical provider or scientist committing to finding a new cure. For example, a book telling the story of a man who found a tea that cures cancer states the following: "This is the extraordinary account of the cancer remission of a man told he would be dead in two or three months, what he decided to do, and the almost miraculous tea that he accidentally created, which paved the way for his cancer's disappearance, never to return." Cancer treatment discovery books provide narratives of persons in desperate situations that may be compelling for readers facing similar challenges.

Theme 8: Cancer Cure Suppression

The last theme, cancer cure suppression, refers to claims that cancer cures already exist but are suppressed due to the vested interests of various others (n=82, 16.6%). These books argue that they are hidden or banned from practice because of financial interests from the pharmaceutical industry, purposeful legal battles with regulators and medical associations, and conspiracies to ruin personal reputations. For example, a book description states the following: "This book exposes the cures for cancer that are discovered and then suppressed by corporate and political interests. Big medical companies are dependent on a continuing increase in cancer so that they can continue increasing their profits." Books also allege that certain figures, sometimes the author themselves or another public figure, are being cancelled or censored. For example, a book description states: the following "His [anonymized name] internet speech was banned by a federal judge in 2014 on behalf of the FDA in order to hide natural cures from the public." Books alleging that cancer cures are suppressed promote conspiracy theories and revise or misrepresent historical events and decisions.

Discussion

Principal Findings

These results demonstrate that misleading cancer cure and treatment books are for sale, visible, and seemingly prevalent on Amazon.com. Nearly half of the "cancer cure" books for sale offered misleading and potentially harmful cancer treatment misinformation. Misleading books directly claimed to have efficacious cures for cancer, undermined scientifically supported treatments, misapplied scientific reasoning, oversimplified cancer and cancer treatments, and promoted conspiracy theories. Books offered information that may delay or encourage patients to opt out of best-standard treatments and create false hope.

Our study contributes to the increasing research documenting the presence, spread, and mechanisms of medical misinformation on Amazon. Notably, Amazon's algorithm has been reported to amplify books with vaccine misinformation [23,42]. In other cases, Amazon has been found to host products selling fake autism cures [43] and promote COVID-19 misinformation [25]. Amazon's search results have been found to rank products with misinformation higher than sources debunking misinformation [34]. Concerningly, misinformation in Amazon-hosted products may be amplified by fake product reviews, which are associated with a causal increase in product sales [44,45].

Amazon effectively gives legitimacy affordances [46] to authors without medical credentials to freely write medical information and advice books. Previous studies have documented how other technology platforms, like Meta and Google, may inadvertently enable persons selling disproven or unproven cancer therapies to appear to be legitimate medical providers [10]. Many books in our study were self-published, thus undergoing limited or no external or publisher review [22]. Amazon may enable such books to be put up for sale, advertised, matched to user search queries depending on keywords, and displayed alongside books written by qualified authors. The blurred differentiation between books offering misleading or credible information creates unfair difficulties for patients with cancer evaluating sources of web-based information during a highly stressful time in their lives.

The presence of misleading cancer cure books for sale on Amazon occurs despite Amazon policies that prohibit

XSL•FO
RenderX

non–FDA-approved medical claims [21]. Presumably, this policy would include books whose market value is derived from the cancer treatment information they offer, although this is not immediately clear. The findings of the study may support Amazon to expand or revise its existing policies to protect its users. The presence of misleading cancer cure books also occurs after previous criticism of Amazon for allowing the sale of books with misleading information during the COVID-19 pandemic [47,48]. This context suggests that Amazon's disinformation and misinformation policies related to the sale of health-related books are at this time limited, inadequately protecting its users.

Amazon benefits from a distinct form of "platform power" [49] that complicates regulators' abilities and their political support to act upon health misinformation. Platform power refers to the "tacit allegiance" of a platform's consumers. Confronting the mechanisms that make Amazon successful, such as its algorithms, affordances to sellers, advertisements, and convenience, is likely not in the interest—and would feel beyond the power—of many Amazon users. This also limits the public support of political entities to act upon health misinformation on Amazon. Therefore, while receiving this distinct platform power, Amazon, like other big tech platforms, may not be incentivized [50] to act upon health misinformation until its users demand such actions or until its business interests are threatened. It has been argued that digital platforms, which are reliant on automated, algorithmically driven, paid advertising for income, are themselves important commercial determinants of health [50,51]. Furthermore, they may have an inherent conflict of interest, where meaningfully curbing exploitation might require greater resources committed to oversight and regulatory compliance or reductions in the profitability of certain platform features. In light of their scale, power, and contributions to the wider digital ecosystem, such platforms may have cumulative harmful effects on the spread and consistent appearance of such misinformation. This may also occur in ways that make the independent investigation of effective cancer treatment, by those who may have just received a diagnosis, highly challenging in inequitable ways. Inherent to these issues is the tension between the informational needs of the public and their right to be protected from the harms associated with the spread of misinformation on the one hand, and the interests of corporations like Amazon to maximize profits on the other hand.

Our results and those of other studies documenting harmful web-based ecosystems for patients with cancer suggest that cancer-focused organizations should engage in preemptive treatment misinformation debunking [52]. Dispelling cancer misinformation before patients are exposed to it can result in a form of psychological inoculation against ineffective treatments [53]. Patients with cancer, especially those with terminal prognoses, are inevitably susceptible to exploitive actions from bad actors selling false hope [54]. Web-based technology platforms, such as specific social media platforms, Amazon, or Google, inadvertently assist such bad actors in selling their products [10]. Health care providers and cancer support organizations are well placed to preemptively debunk such content, ensuring a first-line defense against cancer misinformation. Proactive surveillance of cancer treatment misinformation and health literacy initiatives can help curb misinformation effects.

Limitations

Our study has several important limitations. We collected and analyzed the publicly available information from a book's Amazon.com listing page. It was not feasible nor ethical to purchase and read the entire book's content. However, the product information collected, including the book's title and description, is suitable for analysis due to its conveyance of purpose and content. In deciding whether the book was misleading or not, we relied upon external lists and signifiers of alternative cancer treatments. We sought to mitigate this limitation by auditing misleading coding application decisions. Finally, as Amazon.com search results are personalized based on prior behavior, it is possible that recommendations or highlighted products may appear even more frequently than reported here for users who previously engaged with such products or those closely related to them.

Future Research

Future research into cancer misinformation on Amazon could include a focus on 3 areas in seeking to assess the potential dangers of such books: best sellers, advertising, and impact. Amazon has several lists for general and specific types of cancer, each of which has its own "best sellers" list that evolves over time [55]. It is prudent to track the best-selling, cancer-related "information" books for sale on Amazon for misinformation. Second, this study did not examine paid advertising of books containing cancer misinformation. Future research should determine the extent of paid advertising involving misleading cancer cure books on Amazon. Finally, the impact of misleading cancer books on their readers could be evaluated using survey or focus group methods. Other data sources, such as the reviews of misleading cancer cure books available on Amazon, could also be worthy of study for accuracy, provenance, and potential manipulation, to determine whether and how book reviews disseminate cancer misinformation.

Acknowledgments

The authors would like to thank the Public Health Agency of Canada, the Canadian Institutes of Health Research, and the Stem Cell Network for their generous support of ScienceUpFirst and the project Law, Public Policy and Social License for Next-Generation Regenerative. GH has a Fellowship (MH052) funded as part of the Three National Institute for Health and Care Research (NIHR) Schools Mental Health Programme. The views expressed are those of the author and not necessarily those of any funders. The funders had no role in study conceptualization, design, or interpretation.

XSL•FO
RenderX

Data Availability

The coded data and media files are available upon reasonable request to the corresponding author.

Authors' Contributions

MZ conceptualized the manuscript. MZ collected the data. All authors reviewed the data. MZ drafted a coding framework. All authors reviewed and provided feedback on the coding framework. MZ coded the data. All authors interpreted the results. MZ wrote the manuscript. All authors reviewed and agreed upon the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standards for Reporting Qualitative Research checklist.
[PDF File (Adobe PDF File), 115 KB - jmir_v26i1e56354_app1.pdf]

References

1. Holdren J, Surkan K, Downing A. Perspectives of people with cancer or hereditary cancer risk on the use and value of online peer support. J Patient Cent Res Rev 2023 Apr 18;10(2):58-67 [FREE Full text] [doi: 10.17294/2330-0698.1968] [Medline: 37091112]
2. Paulson AE, Stein A, Kendal JK, Bernthal NM, Wessel LE. Most patients with bone sarcomas seek emotional support and information about other patients' experiences: a thematic analysis. Clin Orthop Relat Res 2023 Jul 11;482(1):161-171. [doi: 10.1097/corr.0000000000002761]
3. Lazard AJ, Nicolla S, Vereen RN, Pendleton S, Charlot M, Tan H, et al. Exposure and reactions to cancer treatment misinformation and advice: survey study. JMIR Cancer 2023 Jul 28;9:e43749 [FREE Full text] [doi: 10.2196/43749] [Medline: 37505790]
4. Johnson SB, Park HS, Gross CP, Yu JB. Complementary medicine, refusal of conventional cancer therapy, and survival among patients with curable cancers. JAMA Oncol 2018 Oct 01;4(10):1375-1381 [FREE Full text] [doi: 10.1001/jamaoncol.2018.2487] [Medline: 30027204]
5. Johnson S, Park H, Gross C, Yu J. Use of alternative medicine for cancer and its impact on survival. J Natl Cancer Inst 2018 Jan 01;110(1):121-124. [doi: 10.1093/jnci/djx145] [Medline: 28922780]
6. Johnson SB, Parsons M, Dorff T, Moran MS, Ward JH, Cohen SA, et al. Cancer misinformation and harmful information on Facebook and other social media: a brief report. J Natl Cancer Inst 2022 Jul 11;114(7):1036-1039 [FREE Full text] [doi: 10.1093/jnci/djab141] [Medline: 34291289]
7. Abramson M, Feiertag N, Javidi D, Babar M, Loeb S, Watts K. Accuracy of prostate cancer screening recommendations for high-risk populations on YouTube and TikTok. BJUI Compass 2023 Mar;4(2):206-213 [FREE Full text] [doi: 10.1002/bco2.200] [Medline: 36816146]
8. Segado-Fernández S, Herrera-Peco I, Jiménez-Gómez B, Ruiz Núñez C, Jiménez-Hidalgo PJ, Benítez de Gracia E, et al. Realfood and cancer: analysis of the reliability and quality of YouTube content. Int J Environ Res Public Health 2023 Mar 13;20(6):5046 [FREE Full text] [doi: 10.3390/ijerph20065046] [Medline: 36981954]
9. Xu AJ, Taylor J, Gao T, Mihalcea R, Perez-Rosas V, Loeb S. TikTok and prostate cancer: misinformation and quality of information using validated questionnaires. BJU Int 2021 Oct 21;128(4):435-437. [doi: 10.1111/bju.15403] [Medline: 33811424]
10. Zenone M, Snyder J, Bélisle-Pipon JC, Caulfield T, van Schalkwyk M, Maani N. Advertising alternative cancer treatments and approaches on Meta social media platforms: content analysis. JMIR Infodemiology 2023 May 31;3:e43548 [FREE Full text] [doi: 10.2196/43548] [Medline: 37256649]
11. Yussof I, Ab Muin NF, Mohd M, Hatah E, Mohd Tahir NA, Mohamed Shah N. Breast cancer prevention and treatment misinformation on Twitter: an analysis of two languages. Digit Health 2023 Oct 06;9:20552076231205742 [FREE Full text] [doi: 10.1177/20552076231205742] [Medline: 37808244]
12. Modell SM, Ponte AH, Director HR, Pettersen SK, Kardia SL, Goltz HH. Breast cancer prevention misinformation on Pinterest: one side of a thick coin. Am J Public Health 2023 Mar;113(3):e1-e2. [doi: 10.2105/AJPH.2022.307203] [Medline: 36791357]
13. Yoon HY, You KH, Kwon JH, Kim JS, Rha SY, Chang YJ, et al. Understanding the social mechanism of cancer misinformation spread on YouTube and lessons learned: infodemiological study. J Med Internet Res 2022 Nov 14;24(11):e39571 [FREE Full text] [doi: 10.2196/39571] [Medline: 36374534]
14. Kureyama N, Terada M, Kusudo M, Nozawa K, Wanifuchi-Endo Y, Fujita T, et al. Fact-checking cancer information on social media in Japan: retrospective study using Twitter. JMIR Form Res 2023 Sep 06;7:e49452 [FREE Full text] [doi: 10.2196/49452] [Medline: 37672310]

XSL•FO
RenderX

15. Warner EL, Basen-Engquist KM, Badger TA, Crane TE, Raber-Ramsey M. The online cancer nutrition misinformation: a framework of behavior change based on exposure to cancer nutrition misinformation. Cancer 2022 Jul 01;128(13):2540-2548 [FREE Full text] [doi: 10.1002/cncr.34218] [Medline: 35383913]
16. Sissung TM, Figg WD. Facebook groups for alternative treatments for cancer: advertising masquerading as community support. Lancet Oncol 2021 Jan;22(1):25-26. [doi: 10.1016/S1470-2045(20)30457-5] [Medline: 33387494]
17. Trivedi N, Lowry M, Gaysynsky A, Chou WYS. Factors associated with cancer message believability: a mixed methods study on simulated Facebook posts. J Cancer Educ 2022 Dec 19;37(6):1870-1878 [FREE Full text] [doi: 10.1007/s13187-021-02054-7] [Medline: 34145508]
18. Curry D. Amazon statistics. Business of Apps. URL: https://www.businessofapps.com/data/amazon-statistics/ [accessed 2023-10-13]
19. Weigel M. What you don't know about Amazon. The New York Times. 2022 Apr 21. URL: https://www.nytimes.com/2022/04/21/opinion/amazon-product-liability.html [accessed 2023-11-16]
20. Search: "cancer". Amazon. URL: https://www.amazon.com/s?k=cancer&crid=O6XIQLF76X4X&sprefix=cance%2Caps%2C155&ref=nb_sb_noss_2 [accessed 2023-10-12]
21. Prohibited product claims. Amazon. URL: https://sellercentral.amazon.com/help/hub/reference/external/G202024200 [accessed 2023-10-13]
22. Dewey C. Popular on Amazon: wildly misleading self-published books about Ebola, by random people without medical degrees. Washington Post. 2014 Oct 2. URL: https://www.washingtonpost.com/news/the-intersect/wp/2014/10/02/popular-on-amazon-wildly-misleading-self-published-books-about-ebola-by-random-people-without-medical-degrees/ [accessed 2023-10-12]
23. Shin J, Valente T. Algorithms and health misinformation: a case study of vaccine books on Amazon. J Health Commun 2020 May 03;25(5):394-401. [doi: 10.1080/10810730.2020.1776423] [Medline: 32536257]
24. Gilbert D. Amazon is selling a bogus 'Plandemic' COVID conspiracy book in its 'Science' section. Vice. 2021 Jul 6. URL: https://www.vice.com/en/article/jg8vdp/amazon-is-selling-a-bogus-plandemic-covid-conspiracy-book-in-its-science-section [accessed 2023-10-13]
25. Schubert C. COVID-19 misinformation still featured on Amazon, months after inquiries from lawmakers. GeekWire. 2021 Dec 14. URL: https://www.geekwire.com/2021/covid-19-conspiracy-books-still-featured-on-amazon-months-after-inquiries-from-lawmakers/ [accessed 2023-10-12]
26. Matthews S. Amazon is selling bogus 'cancer cure' books that wrongly tell patients to drink carrot juice, take CBD oil or adopt an alkaline diet. Daily Mail. 2020 Jan 1. URL: https://www.dailymail.co.uk/health/article-7665239/Amazon-selling-hundreds-bogus-cancer-cure-books.html [accessed 2023-07-10]
27. Robinson R. Medical misinformation rife in Amazon bestsellers about public health. Coda Story. 2022 Sep 23. URL: https://www.codastory.com/newsletters/amazon-books-misinformation/ [accessed 2023-10-13]
28. Wark C. Chris Beat Cancer: A Comprehensive Plan for Healing Naturally. 2nd edition. Carlsbad, CA: Hay House Inc; 2021.
29. Gorski D. Yes, Chris beat cancer, but it wasn't quackery that cured him. Science-Based Medicine. 2013 Oct 15. URL: https://sciencebasedmedicine.org/chris-beat-cancer/ [accessed 2023-10-13]
30. Ohlheiser A. They turn to Facebook and YouTube to find a cure for cancer — and get sucked into a world of bogus medicine. The Washington Post. 2019 Jun 25. URL: https://www.washingtonpost.com/lifestyle/style/they-turn-to-facebook-and-youtube-to-find-a-cure-for-cancer--and-get-sucked-into-a-world-of-bogus-medicine/2019/06/25/6df3ddae-7cdc-11e9-a5b3-34f3edf1351e_story.html [accessed 2022-07-28]
31. Gorski D. Facebook, Google, and social media vs. medical misinformation: an update. Science-Based Medicine. 2019 Jul 8. URL: https://sciencebasedmedicine.org/facebook-google-and-social-media-vs-medical-misinformation-an-update/ [accessed 2023-10-13]
32. Chris Beat Cancer's Amazon page. Amazon. URL: https://www.amazon.com/shop/chrisbeatcancer [accessed 2023-10-13]
33. Johnson SB, Bylund CL. Identifying cancer treatment misinformation and strategies to mitigate its effects with improved radiation oncologist-patient communication. Pract Radiat Oncol 2023 Feb;13(4):282-285. [doi: 10.1016/j.prro.2023.01.007] [Medline: 36736620]
34. Juneja P, Mitra T. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. 2021 May 7 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-27. [doi: 10.1145/3411764.3445250]
35. Data Miner. URL: https://dataminer.io/login [accessed 2021-07-17]
36. Snyder J, Zenone M, Caulfield T. Crowdfunding for complementary and alternative medicine: what are cancer patients seeking? PLoS One 2020 Nov 20;15(11):e0242048 [FREE Full text] [doi: 10.1371/journal.pone.0242048] [Medline: 33216790]
37. CAM therapies: A-Z. National Cancer Institute. URL: https://cam.cancer.gov/health_information/cam_therapies_a-z.htm [accessed 2023-10-12]
38. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]

XSL•FO
RenderX

39. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
40. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 02;16(1):160940691773384. [doi: 10.1177/1609406917733847]
41. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
42. Hussein E, Eldardiry H. Investigating misinformation in online marketplaces: an audit study on Amazon. arXiv Preprint posted online on September 25, 2020. [doi: 10.48550/arXiv.2009.12468]
43. O'Kane C. Amazon removes books promoting autism "cures" and vaccine misinformation. CBS News. 2019 Mar 13. URL: https://www.cbsnews.com/news/amazon-removes-books-promoting-autism-cures-and-vaccine-misinformation/ [accessed 2023-10-12]
44. He S, Hollenbeck B, Proserpio D. The market for fake reviews. Marketing Science 2022 Oct 1;41(5):896-921. [doi: 10.2139/ssrn.3664992]
45. Ovide S. How fake reviews hurt us and Amazon. The New York Times. 2020 Nov 19. URL: https://www.nytimes.com/2020/11/19/technology/fake-reviews-amazon.html [accessed 2023-10-12]
46. Di Domenico G, Nunan D, Pitardi V. Marketplaces of misinformation: a study of how vaccine misinformation is legitimized on social media. Journal of Public Policy & Marketing 2022 Jul 12;41(4):319-335. [doi: 10.1177/07439156221103860]
47. Bond S. An anti-vaccine book tops Amazon's COVID search results. lawmakers call foul. NPR. 2021 Sep 9. URL: https://www.npr.org/2021/09/09/1035559330/democrats-slam-amazon-for-promoting-false-covid-cures-and-anti-vaccine-claims [accessed 2023-10-12]
48. Palmer A. Sen. Elizabeth Warren asks Amazon CEO Andy Jassy to explain why the company's algorithms recommend COVID misinformation. CNBC. 2021 Sep 8. URL: https://www.cnbc.com/2021/09/08/elizabeth-warren-asks-amazon-ceo-to-crack-down-on-covid-misinformation.html [accessed 2023-10-12]
49. Culpepper PD, Thelen K. Are we all Amazon primed? consumers and the politics of platform power. Comp Polit Stud 2019 Jun 11;53(2):288-318. [doi: 10.1177/0010414019852687]
50. Zenone M, Kenworthy N, Maani N. The social media industry as a commercial determinant of health. Int J Health Policy Manag 2023 Apr 27;12:6840 [FREE Full text] [doi: 10.34172/ijhpm.2022.6840] [Medline: 35490262]
51. Kenworthy N, Storeng KT, Zenone M. The digital technology sector and health systems. In: Maani N, Petticrew M, Galea S, editors. Commercial Determinants of Health. New York, NY: Oxford University Press; 2023.
52. Tay LQ, Hurlstone MJ, Kurz T, Ecker UKH. A comparison of prebunking and debunking interventions for implied versus explicit misinformation. Br J Psychol 2022 Aug 29;113(3):591-607. [doi: 10.1111/bjop.12551] [Medline: 34967004]
53. Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S. Psychological inoculation improves resilience against misinformation on social media. Sci Adv 2022 Aug 26;8(34):eabo6254 [FREE Full text] [doi: 10.1126/sciadv.abo6254] [Medline: 36001675]
54. Snyder J. Exploiting Hope: How the Promise of New Medical Interventions Sustains Us--and Makes Us Vulnerable. New York, NY: Oxford University Press; 2020.
55. Amazon best sellers: cancer. Amazon. URL: https://www.amazon.com/Best-Sellers-Books-Cancer/zgbs/books/171141/ref=zg_bs_unv_books_4_4673_2 [accessed 2023-10-12]

Abbreviations

FDA: Food and Drug Administration

Edited by A Mavragani; submitted 13.01.24; peer-reviewed by Y Zhu, J Mihanovic; comments to author 08.03.24; revised version received 26.04.24; accepted 06.08.24; published 08.10.24.

Please cite as:
Zenone M, van Schalkwyk M, Hartwell G, Caulfield T, Maani N
Selling Misleading "Cancer Cure" Books on Amazon: Systematic Search on Amazon.com and Thematic Analysis
J Med Internet Res 2024;26:e56354
URL: https://www.jmir.org/2024/1/e56354
doi:10.2196/56354
PMID:

©Marco Zenone, May van Schalkwyk, Greg Hartwell, Timothy Caulfield, Nason Maani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.10.2024. This is an open-access article distributed under the terms of the

XSL•FO
RenderX

Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Public Emotions on Obesity During the COVID-19 Pandemic Using Sentiment Analysis and Topic Modeling: Cross-Sectional Study

Jorge César Correia[1], MD; Sarmad Shaharyar Ahmad[2], MSc; Ahmed Waqas[3], MBBS, PhD; Hafsa Meraj[4], MBBS; Zoltan Pataky[1], MD

[1]Unit of Therapeutic Patient Education, WHO Collaborating Center, University Hospitals of Geneva and University of Geneva, Geneva, Switzerland
[2]School of Mathematics, Computer Science & Engineering, Liverpool Hope University, Liverpool, United Kingdom
[3]Department of Primary Care & Mental Health, Institute of Population Health, University of Liverpool, Liverpool, United Kingdom
[4]Greater Manchester Mental Health NHS Foundation Trust, Salford, United Kingdom

Corresponding Author:
Jorge César Correia, MD
Unit of Therapeutic Patient Education
WHO Collaborating Center
University Hospitals of Geneva and University of Geneva
Chemin Venel 7
Geneva, 1206
Switzerland
Phone: 41 (0)22 372 9722
Email: JorgeCesar.Correia@hcuge.ch

Abstract

Background: Obesity is a chronic, multifactorial, and relapsing disease, affecting people of all ages worldwide, and is directly related to multiple complications. Understanding public attitudes and perceptions toward obesity is essential for developing effective health policies, prevention strategies, and treatment approaches.

Objective: This study investigated the sentiments of the general public, celebrities, and important organizations regarding obesity using social media data, specifically from Twitter (subsequently rebranded as X).

Methods: The study analyzes a dataset of 53,414 tweets related to obesity posted on Twitter during the COVID-19 pandemic, from April 2019 to December 2022. Sentiment analysis was performed using the XLM-RoBERTa-base model, and topic modeling was conducted using the BERTopic library.

Results: The analysis revealed that tweets regarding obesity were predominantly negative. Spikes in Twitter activity correlated with significant political events, such as the exchange of obesity-related comments between US politicians and criticism of the United Kingdom's obesity campaign. Topic modeling identified 243 clusters representing various obesity-related topics, such as childhood obesity; the US President's obesity struggle; COVID-19 vaccinations; the UK government's obesity campaign; body shaming; racism and high obesity rates among Black American people; smoking, substance abuse, and alcohol consumption among people with obesity; environmental risk factors; and surgical treatments.

Conclusions: Twitter serves as a valuable source for understanding obesity-related sentiments and attitudes among the public, celebrities, and influential organizations. Sentiments regarding obesity were predominantly negative. Negative portrayals of obesity by influential politicians and celebrities were shown to contribute to negative public sentiments, which can have adverse effects on public health. It is essential for public figures to be mindful of their impact on public opinion and the potential consequences of their statements.

(J Med Internet Res 2024;26:e52142) doi:10.2196/52142

KEYWORDS

obesity; Twitter; infodemic; attitude; opinion; perception; perspective; obese; weight; overweight; social media; tweet; sentiment; topic modeling; BERT; Bidirectional Encoder Representations from Transformers; NLP; natural language processing; general public; celebrities

XSL•FO
RenderX

Introduction

Obesity is a complex, multifactorial, and relapsing disease that has become increasingly prevalent worldwide among people belonging to all age groups [1]. The age-standardized prevalence of obesity has increased from 4.6% in 1980 to 14% in 2019, with a rising trend globally [1]. It is associated with a high risk of multiple comorbidities, including type 2 diabetes; cardiovascular, respiratory, and digestive diseases; and multiple cancers [2-4]. Obesity also leads to important, negative psychological and social impacts, as well as a significant risk of depression, anxiety disorders, and other psychiatric diseases. Furthermore, obesity is linked to public and internalized stigma, with individuals experiencing low self-esteem, social isolation, and mental health problems [5,6]. People with obesity may also be exposed to bullying, discrimination, and negative stereotypes on social media platforms [7,8].

While primary research, including epidemiological and interventional studies, is essential for understanding and addressing obesity, the importance of public health communication cannot be overstated [9]. Effective health communication can help raise awareness, debunk myths, and promote healthy behaviors related to obesity [10,11]. While social media platforms can be sources of valuable health information, they often contain inaccurate or misleading information, which can be detrimental to public health [7]. The infodemic accompanying the pandemic exemplifies how misleading health information on social media platforms can contribute to fear, unfounded concerns, and poor health behaviors among the general population [9,12]. In addition to providing a platform for the dissemination of potentially harmful misinformation, social media can also perpetuate stereotypes and biases against individuals with obesity, further exacerbating the negative impacts of this condition [7].

Social media platforms (eg, Twitter; subsequently rebranded as X) offer a wealth of information that can be harnessed to gain insights into the dominant health behaviors, attitudes, and sentiments of the masses toward obesity. By analyzing tweets and other social media posts, researchers can examine the perspectives and opinions of the public, celebrities, and influential organizations on this critical health issue [13].

Advanced analytical techniques, such as sentiment analysis and topic modeling, can be used to explore the complex landscape of opinions and emotions surrounding obesity on social media platforms [14]. Sentiment analysis allows for the systematic identification and categorization of subjective information in textual data, providing insights into the prevailing emotions and opinions expressed in social media posts [15]. Topic modeling, on the other hand, is an unsupervised machine learning technique that can uncover hidden patterns and themes within large volumes of text data, revealing the most prominent topics and trends related to obesity in social media discussions [16]. By using these cutting-edge methodologies, this study aims to uncover the dominant narratives, attitudes, and emotions related to obesity on social media platforms. The findings from this analysis will not only provide a comprehensive understanding of public sentiment toward obesity but also inform the development of targeted public health interventions and communication strategies to address this global health challenge.

The research objectives of this study are (1) to examine the role of social media in perpetuating or challenging stereotypes, biases, and stigma against individuals with obesity, and assess the potential impact of these factors on their mental health and well-being; and (2) to identify the most prominent topics and trends in discussions surrounding obesity on social media, examining the content generated by the general public, celebrities, and influential organizations.

Methods

Overview

This study uses a combination of sentiment analysis and topic modeling techniques to analyze tweets related to obesity. This approach allowed us to examine both the sentiment and content of the discussions, providing a comprehensive understanding of the perceptions and concerns related to obesity.

Data Collection

Twitter data were collected using Tweepy, a Python-based open-source wrapper package for the Twitter application programming interface (API) [17]. Tweepy simplifies the communication process with the API and enables access to Twitter's historical records. Tweepy serves as a reliable, open-source Python package that streamlines communication with the Twitter API, enhancing the overall development experience. We did not apply any restrictions to the country the tweets originated from. However, only tweets in the English language were retrieved. We leveraged Tweepy's *paginator* function to manage pagination and retrieve all relevant data associated with a single request and its arguments [17]. Pagination refers to techniques for programmatically querying all pages to obtain the entire return dataset [17].

In total, we collected 53,414 tweets posted between April 2019 and December 2022 containing the terms "obese" and "obesity." We specifically analyzed original tweets to understand public, celebrity, and organizational sentiments about obesity during the COVID-19 pandemic. We chose original tweets as they often originate key discussions and reflect the genuine stance of the users, particularly influential figures whose opinions can significantly sway public discourse. This allowed us to distill the core narratives and sentiments surrounding obesity, unobscured by the potential dilution or bias that retweets and comments might introduce. These latter elements, while undeniably valuable for gauging the spread and engagement with certain opinions, can encompass a wide array of responses, from supportive echoes to direct rebuttals or even tangential discussions.

Data Preprocessing

Before conducting sentiment analysis and topic modeling, data preprocessing was performed to clean and normalize the collected tweets using *NeatText* package in Python [18]. This involved the following steps [18]:

1. Removing URLs, special characters, and numbers
2. Converting all text to lowercase

XSL•FO
RenderX

3. Tokenizing the tweets into individual words
4. Removing stop words (commonly used words that do not carry significant meaning)

Celebrity Influence Analysis

Our methodology for assessing the influence of celebrities and important organizations involved analyzing the engagement metrics of their tweets related to obesity. We identified tweets from these individuals that resulted in a marked increase in Twitter activity, visualized as spikes in tweets on the topic of obesity over time. The events and tweets corresponding to these spikes were identified by the researchers (AW and SSA) from the Twitter records. This analysis provided insight into how celebrity statements on obesity resonated with and mobilized their audience, serving as an indirect yet insightful indicator of a tweet's impact on public attitudes and, by extension, its potential public health implications.

Sentiment Analysis Using Transformer Language Model XLM-RoBERTa-Base-Twitter

Our aim was to classify tweets into 3 sentiment classes: positive, negative, and neutral. To achieve this, we used zero-shot text classification, which allows for the categorization of tweets without explicitly training the model. The pretrained transformer language model XLM-RoBERTa-base-Twitter (XLM-Twitter) by Cardiff NLP was used for this purpose [15]. This model has been fine-tuned on approximately 198 million tweets for various multilingual and monolingual apps, including sentiment analysis in 8 different languages [15]. The model processes the input tweets and generates an output with 3 labels (positive, neutral, and negative), 3 probability scores, and the maximum value from the scores as a label for the input tensor. The probability scores are calculated using the *softmax* function, which converts the raw output values (logits) into probabilities [15].

Topic Modeling Using BERTopic

In this study, we followed the original BERTopic framework, using BERT (Bidirectional Encoder Representations from Transformers) for document embeddings, UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction, and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) for clustering [16]. The process involves 4 main steps.

Mapping Words and Documents to Real-Value Vectors

BERT, a pretrained language model, is used to convert words and documents into meaningful real-value vectors. These embeddings capture semantic information, enabling the identification of similar words and documents [16].

Performing Dimensionality Reduction

UMAP is used to reduce the number of features of the embeddings while preserving the global structure of the vectors in a lower-dimensional space. UMAP is a manifold learning technique that uses concepts from differential geometry and algebraic topology. It has 2 main hyperparameters, the number of nearest neighbors (*n*) and the minimum distance (*$minDist$*), which control the balance between local and global structure preservation [16].

Clustering the Documents

HDBSCAN is used for clustering. This algorithm combines the advantages of hierarchical clustering methods with DBSCAN (Density-Based Spatial Clustering of Applications with Noise). It does not require previous knowledge of the ideal number of clusters and allows for the automatic recognition of noise data. HDBSCAN has 2 main hyperparameters, the minimum cluster size (*min_cluster_size*) and the minimum samples (*min_samples*), which control the density-based clustering process. HDBSCAN works by constructing a hierarchical tree of clusters and selecting the most stable clusters throughout the tree as the final result [16].

Retrieving the Most Representative Words for Each Cluster

A class-based variation of term frequency–inverse document frequency is used to identify the most representative words for each cluster [16]. Term frequency–inverse document frequency is a common weighting scheme that aims to provide a vector representation of each document, reflecting the importance of each word by calculating the normalized product of the term frequency and the inverse document frequency.

Spikes in Twitter Activity Over Time for Each Cluster

An analysis of Twitter response activity was performed with a focus, particularly on the spikes in public response toward different clusters. This involved examining the volume and velocity of engagement across the identified clusters, which helped us understand the public's reaction intensity to various obesity-related discussions. By mapping these spikes in response activities to specific clusters, we were able to further elucidate the public's concerns and interests regarding obesity during the COVID-19 pandemic, providing additional layers of context to our findings.

Finally, the 2 authors (AW and SSA) conducted a manual analysis for the top 20 clusters, for an in-depth understanding of the prevalent attitudes and perceptions regarding obesity among Twitter users. The top 20 clusters were assessed for homogeneity, and those representing similar broad themes were merged by the authors. Representative tweets were cited after anonymization, to corroborate authors' interpretations of the clusters.

Ethical Considerations

Our study exclusively analyzed publicly available tweets from Twitter. The data used in our analysis consist of information that was shared publicly by users on the platform, without any expectation of privacy for content posted in the public domain. Given the nature of our study, it falls under the category of research exempt from ethical review according to many institutional and national guidelines on research ethics, which typically do not require ethical approval for studies involving the analysis of publicly available, nonsensitive data where individuals cannot be identified.

XSL•FO
RenderX

Results

Sentiment Analysis for Tweets Regarding Obesity

The Twitter API was used to extract 53,414 tweets posted between April 2019 and December 2022 containing the terms "obese" and "obesity." Sentiment analysis revealed a substantially higher percentage of tweets (n=37,048, 69.36%) represented negative sentiments, followed by neutral (n=11,169, 20.91%) and positive (n=5196, 9.73%; Figure 1). A steadily increasing trend in negative sentiments was observed among Twitter users from April 2019 to July 2022, while the volume of neutral and positive tweets remained stagnant (Figure 2). Individual tweets classified according to their sentiments can be accessed (Multimedia Appendix 1).

Figure 1. Volume of tweets on obesity and their breakdown according to sentiments (April 2019 to December 2022).

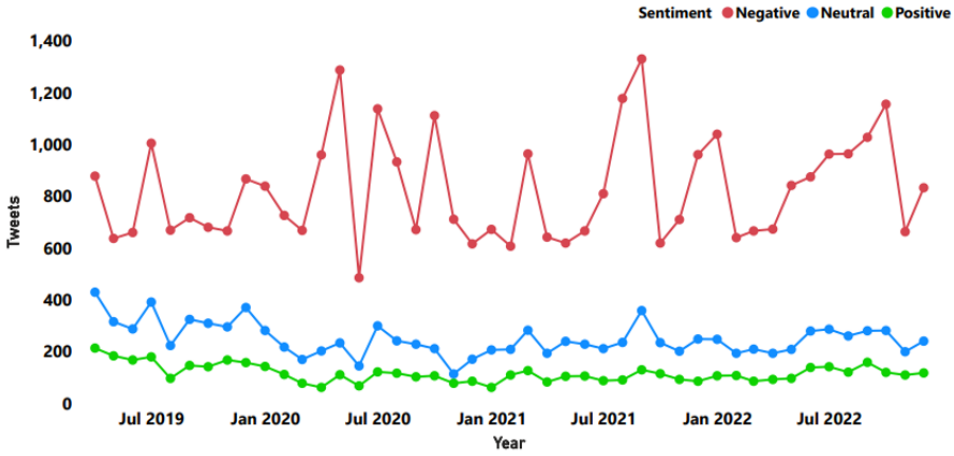

Figure 2. Sentiment-wise yearly trend in tweets regarding obesity (April 2019 to December 2022).

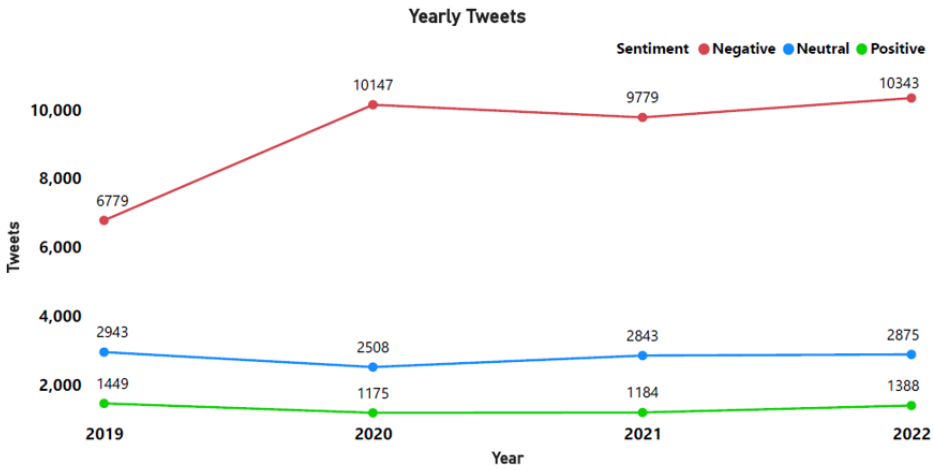

Spikes in Twitter activity (Figure 1) were associated with significant political and public health events. Increased Twitter activity was noted pertaining to the link between increased hospitalizations among people with obesity [19]. On March 4, 2021, research findings on increased risk of death by COVID-19 among people with obesity were reported. Similarly, on April 21, 2020, the US Centers for Disease Control and Prevention (CDC) reported higher frequency of COVID-19–related deaths among people with obesity. Other public health–related events attracting increased Twitter activity included an obesity campaign initiated by the UK government, which drew a lot of criticism from the public (July 27, 2020).

In a similar manner, several tweets from political personalities resulted in spiked Twitter activity. On August 23, 2021, a US-based political commentator's comments on the refusal to vaccinate people with obesity for COVID-19 sparked outrage. Increased Twitter activity was generated after a high-ranking official passed negative remarks on the US President's struggle with obesity (May 19, 2020). On October 2, 2020, the US President's diagnosis of COVID-19 also sparked negative sentiments toward obesity. Among celebrities associated with the entertainment industry, their struggle with obesity was also highlighted negatively (October 9, 2022).

Topic Modeling

Overview

Topic modeling revealed 243 parsimonious clusters (Multimedia Appendix 2). These represented different clusters of tweets on obesity. For brevity, the top 20 clusters of keywords comprising the highest proportion of tweets are presented in Figure 3. An intertopic distance map for obesity-related tweets generated using BERTopic is provided in Multimedia Appendix 3. This map presents the density of each cluster and the relationships and relative distances between clusters generated through topic

XSL•FO
RenderX

modeling. Among the 4 largest clusters (Figure 4), comments by the 52nd speaker of the United States House of Representatives on the US President's struggle with obesity generated large spikes followed by a decrease in tweets. Cluster 3 on obesity and racism generated increased activity in 2020 and cluster 0 on vaccination for COVID-19 generated increased activity after January 2021. A line plot showing trend of tweets from 20 largest clusters is provided in Multimedia Appendix 3. The dendrogram presenting clustering of similar and dissimilar clusters of tweets is presented in Figure 5. A qualitative analysis of the top 20 clusters of tweets was performed by the authors (Table 1).

Figure 3. Popular topics of tweets regarding obesity.

XSL•FO
RenderX

Figure 4. Frequency of tweets visualized over time for the 4 largest topic clusters.

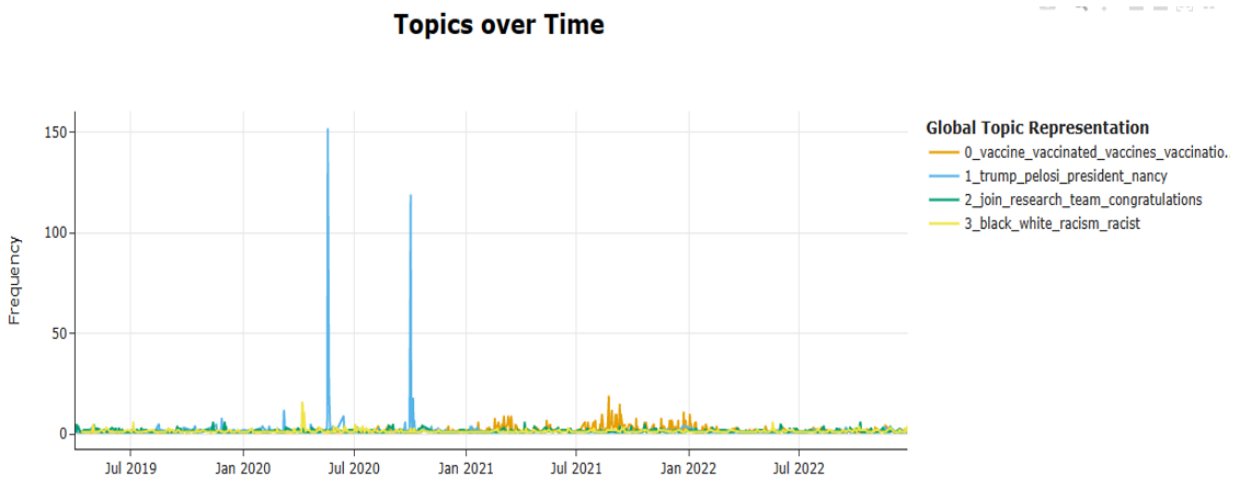

Figure 5. Dendrogram presenting clusters and subclusters of tweets regarding obesity during COVID-19 pandemic.

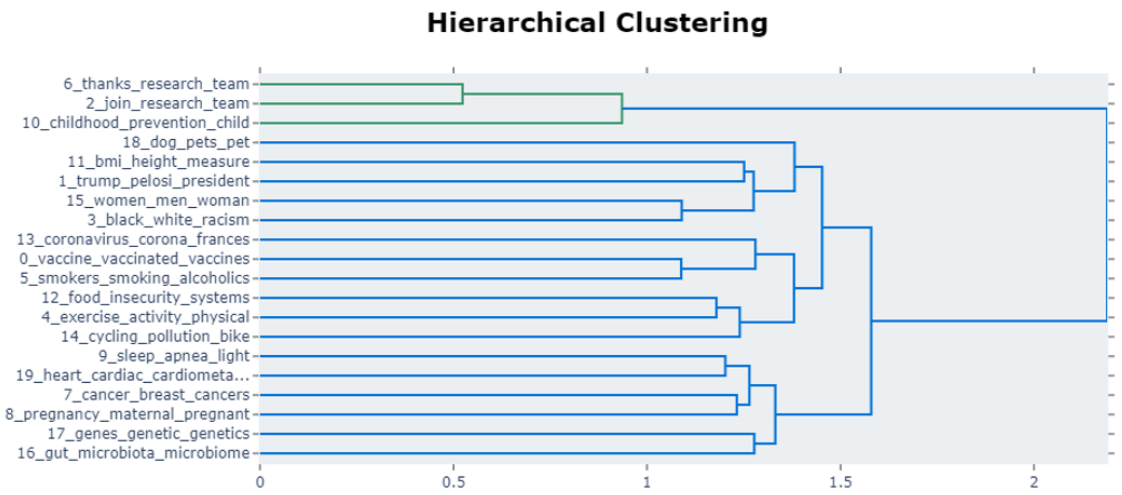

XSL•FO
RenderX

Table 1. Sample tweets pertaining to important clusters regarding obesity.

Cluster	Theme	Tweet
15	Body shaming	• This woman was extremely morbidly obese and a proponent of "body positivity" and "fat acceptance." She is now dead from obesity. She was literally killing herself on camera while being applauded. Fat acceptance is a murderous, suicidal ideology and should be called out as such. • Fat acceptance is a death cult. It's one of the most dangerous ideas in the world...It's probably time that we start thinking about the misinformation that tells people...that morbid, suicidal, self-destructive obesity is something to be celebrated. • Body positivity is not for obese people. • It's for burn victims. Stop trying to excuse yourself for unhealthy behavior.
1	US President's struggle with obesity	• Biden has Covid • Is resting, no fuss because he is in shape and fully vaccinated • When trump had covid was medivac to the hospital, had an army of medical people taking care of him because he is obese and ignored the vaccines. Being an idiot is costly to taxpayers
0 and 13	COVID-19 and obesity	• And yet, as usual, this latest study ignores that obesity is largely driving COVID hospitalizations and deaths, which is relevant if we're considering "all-cause health events." • @washingtonpost "New study"… so we're just going to ignore the fact that every study since the start of the pandemic showed that obesity was the main (and often only) factor in struggling with Covid. If you weren't really old or fat, you never had to worry. • This is dumb. By May 2020 we knew Covid was a disease of the elderly and obese. We should have locked down via BMI and age. Delivered healthy meals to the elderly and diet meals for obese. Could have done some good for the world. • 7. By waiting to get the vax, you're burdening our healthcare system if you get Covid. • The majority of ppl in hospitals w/ Covid have serious co-morbidities, which are the real strain on the system. Obesity kills almost as many ppl per yer as Covid has killed globally.
3	Racism	• Black grifter playbook: • Tell them their obesity, violent aggression, poor educational attainment & absent fathers are the result of factors outside of their control that they're not responsible for. • Tell them they're perfect as they are; blame everything on racism! • Obesity is the number one comorbidity when it comes to COVID. The black community, especially black women in America are disproportionately obese. • These are facts and this is why there is a higher death rate. But you push the racism angle.
4, 10, and 14	Physical activity, prevention, and interventions	• I never saw what people think the issue with going to the gym bc you don't want to be fat is, obesity is a massive issue around the world and can cause health issues that exercise and a good diet can help prevent lmao
2, 5, 7, 9, 14, 16, 17, and 19	Health research of obesity and its antecedents and consequences	• Hadassah-University Medical Center researchers said they have found the mechanism that contributes to the aggressive progression of breast cancer in overweight women. • It's amazing that a lot of Americans think they are healthy. • If you are overweight, obese, have high blood pressure, diabetes, sleep apnea, if you smoke, drink excessively, can't walk a flight of stairs w/o getting short of breath, etc... you aren't healthy. • Gut microbiota alterations in obesity remain a subject of debate. These findings report a decrease in gut microbes associated with inflammation (undesirable Bacteroides 2 enterotype) in individuals with obesity associated with the use of statin drugs. • In Brazil, 53% of men are obese. 45% of women are obese. High levels of cardiometabolic syndromes. They either fix their health or COVID will fix it for them. • It's not just about CO2. Biking is superior in every single way: congestion, road safety, public health (obesity, cardiovascular), air pollution, noise pollution (think animals, sleep quality), rents (think smaller roads, more housing) etc
10	Childhood obesity	• Join @UNICEF and @WHO on World Children's Day for a discussion on childhood obesity prevention and how to maintain momentum in the challenging new COVID-19 reality.
12	Food insecurity	• Interventions aimed at making food retail environments health promoting require targeting a combination of individual, intrapersonal and environmental factors, research by our @GLOBE_obesity centre has found. • UK #obesity plan will fail without action on unhealthy #food. Centre for Food Policy says efforts to lose weight are being thwarted by constant #advertising and exposure in retail.

XSL•FO
RenderX

Restricting Access to Vaccination for COVID-19 for People With Obesity

This domain was discussed in cluster 0 and 13, with a total of 1367 tweets. This cluster presented mainly negative perceptions of the public toward people with obesity. Themes surrounded restricting access of COVID-19 vaccination to people with obesity. This was mainly due to the public perception that obesity led to increased rates of COVID-19 hospitalization and deaths and that people with obesity were to be blamed for this.

US President's Struggle with Obesity

This theme was predominant in cluster 1 with 1003 tweets.

Health Research on Obesity and Its Antecedents and Consequences

This theme was discussed in clusters 2, 7, 9, 16, 17, and 19, comprising a corpus of 2910 tweets. This theme focuses on the cardiometabolic consequences of obesity, genetic risk factors, links with gut microbiota, and the association of different cancers with obesity. Although the dominant theme remained research discussions, however, a small corpus of tweets in cluster 5 discussed the syndemic association of obesity with alcohol and smoking or cited obesity as a risk factor for different disorders along with alcohol and smoking. This suggests that obesity is being perceived and treated similarly to alcohol consumption and smoking, in terms of social stigmatization, health concerns, and public policy. Research on sleep apnea and links with obesity was also highlighted.

Racism and High Obesity Rates Among Black American Population

This was a predominant theme in cluster 3 with 583 tweets. This corpus of tweets was predominated by prejudice toward Black American people. Race was associated with negative perceptions such as high prevalence of criminality, aggression, poor education, and obesity among Black American population.

Glorifying and Body Shaming

This corpus of tweets comprised 295 tweets and was a predominant theme in cluster 15. It explored the dichotomy in beauty standards that can be observed in society, particularly with regard to the perception of obesity and being overweight. It highlighted the complex nature of beauty standards in modern societies, where there is a growing awareness of body diversity and inclusivity. Several tweets celebrated and promoted larger body sizes as desirable and attractive, often as a response to the pressure and stigma faced by people with larger bodies. On the other hand, large groups of Twitter users promoted body shaming based on societal norms and expectations.

Childhood Obesity

Childhood obesity was a predominant topic in cluster 10 with 338 tweets. It highlighted the celebration of World Children's Day by the World Health Organization where UNICEF (United Nations International Children's Emergency Fund) and the World Health Organization discussed childhood obesity prevention, especially during the COVID-19 period.

Food Insecurity

This was a predominant topic in cluster 12 with 329 tweets. These discussed studies pertaining to healthy food retail policies in Australia, and policies and issues pertaining to healthier food choices and ultra-processed food.

Discussion

Principal Findings

By combining the sentiment analysis and topic modeling results, we were able to gain insights into the public discourse on obesity as reflected by the collected tweets.

Our analyses regarding the power of social media in shaping the public discourse on obesity are corroborated by similar studies published recently. White et al [20] examined the role of social messaging shared by persons in the public eye, based on 13 million tweets published during the COVID-19 pandemic. They demonstrated that COVID-19–related tweets published during the pandemic held a more negative tone. Authors also found that tweets shared by celebrities shape risk perceptions, political ideologies, and health-protective behaviors, and thus, influence public sentiment and discourse direction [20].

Our analysis showed that obesity was a frequently discussed topic on Twitter throughout the COVID-19 pandemic. The number of tweets concerning obesity matched those observed in studies for asthma [21] and bowel cancer [22], albeit falling short of the volume seen in discussions about breast cancer [23]. Through a comprehensive analysis, we found that conversations around obesity were not only widespread but also dominated by notably negative sentiments. As the pandemic unfolded, tweets from influential celebrities, politicians, and organizations on obesity generated a lot of debate among the users, who took to Twitter to share their thoughts, experiences, and concerns regarding obesity. When celebrities posted negative comments or opinions about obesity, their followers were more likely to engage in similar conversations, further perpetuating the negativity. These negative tweets could have various implications, such as reinforcing stereotypes and stigmatizing individuals struggling with obesity [24-26]. The ripple effect of these celebrity-driven conversations might lead to increased self-consciousness, social isolation, and negative self-image among those dealing with weight-related issues [27]. This demonstrates the vital role of celebrities in shaping public opinion and discourse around health topics, leading to a profound public health impact.

Exclusion of People With Obesity

Research has shown that obesity is associated with profound stigma, stereotypes, and implicit biases [27-29]. Our analyses revealed 2 important clusters of tweets demonstrating biases against people with obesity. These clusters of tweets revealed implicit health-related and racial biases toward people with obesity. This debate was generated after research insights were published on the role of obesity and ethnicity in worsening COVID-19–related outcomes [19]. The US CDC's communication [19] regarding the risk of severe outcomes related to the COVID-19 pandemic among Black people and people with obesity received significant attention on Twitter.

XSL•FO
RenderX

While CDC's communication was important from a public health perspective, it revealed biases against obesity among celebrities and public. These implicit biases against obesity led to unempathetic notions like denying COVID-19 vaccinations to people with obesity because they are responsible for their own condition. A large cluster of tweets also discussed obesity and COVID-19 outcomes from the lens of racism, leading to emotionally charged debates. Due to these issues in the United States, Black people faced a lot of stigma and prejudice.

Research and Intellectual Discourse on Obesity

In addition to its challenges, Twitter has proven to be an effective platform for intellectual discourse, fostering scientific communication and engagement. Scientists often published their research findings related to obesity, communicated with the public, and advertised their studies for recruitment. Twitter also allowed engagement with diverse audiences which resulted in real-time dissemination of scientific findings [30,31]. Moreover, such social media platforms also help in networking with other scientists, and foster collaboration within the scientific community [30,31]. Social media platforms like Twitter are also important media for exploring the public's sentiments regarding health policies. Twitter could provide an important avenue for discussing health policies and help shape intellectual discourse on obesity. Our analyses demonstrated several useful insights regarding the UK government's "Better Health" campaign [32].

Implications for Future Research and Practice

This research has several implications. It is essential for celebrities to be mindful of their impact on public opinion and the potential consequences of their statements. A recent international consensus statement published by multiple stakeholders highlighted the stigma and prejudice against people with obesity [6]. It emphasized that both the general public and medical professionals have been reluctant to accept obesity as a disease and widespread beliefs persist that obesity is entirely within an individual's control [33]. This leads to weight-based discrimination and unfair treatment at workplaces, health care settings, and educational institutes [6]. In health care environments, negative attitudes held by health care professionals regarding obesity, such as associating it with laziness, lack of willpower, and poor self-control, can negatively impact the quality of care provided to individuals with obesity [6]. Therefore, it is important to launch health information campaigns to promote knowledge regarding the complex nature of obesity and reduce stigma and prejudice against people with obesity [6,27]. Encouraging a more compassionate and supportive conversation about obesity can contribute to a healthier and more inclusive atmosphere, which is crucial for those dealing with the physical and mental struggles related to obesity.

Limitations

This study has some limitations. Our analyses excluded comments and retweets, which constrains our ability to fully delve into the dynamics of public discourse on Twitter. Nevertheless, it afforded us the clarity and manageability necessary for an effective sentiment analysis and topic modeling, presenting a focused snapshot of sentiments during a pivotal time. Acknowledging the limitations of this approach, we propose that subsequent research endeavors could broaden the scope to include retweets and comments. Such expansions would enrich our understanding of the engagement dynamics and the propagation of sentiments within the diverse ecosystem of Twitter, offering a more comprehensive view of the public discourse on obesity.

Given our study's aim to unravel public emotions and sentiments toward obesity during the COVID-19 pandemic, we opted for a more focused approach to ensure the relevance and coherence of our dataset. By limiting our analysis to "obese" and "obesity," we were able to maintain a clear and specific thematic focus, which is crucial for the accuracy of sentiment analysis and topic modeling. We acknowledge that this decision narrows the breadth of our investigation and may exclude certain relevant discussions. However, this specificity was necessary to achieve our research objectives and ensure the integrity of our findings. Future studies could expand on this work by exploring a broader array of terms and using more sophisticated filtering techniques to capture a wider spectrum of public discourse on body weight and health.

The scope of our research was intentionally focused on analyzing public sentiments and topics related to obesity as reflected in original tweets. This focus inherently limited our exploration into the dynamics of Twitter, such as the impact of retweets and comments, which could offer additional layers of understanding regarding public engagement and sentiment propagation. Furthermore, while the influence of celebrities and political personalities on public opinion was noted, a detailed quantitative analysis of their tweets and the direct correlation with increased Twitter activity was beyond the scope of our current study. Such an analysis would require an extension of our dataset to include a broader range of tweets, potentially encompassing various user categories and their interactions. While these limitations constrain the breadth of our analysis, they also highlight areas for future research. Future studies, with access to expanded datasets and resources, could explore these dimensions more comprehensively.

Finally, future studies could benefit from incorporating advanced bot detection techniques to further refine the analysis of Twitter data, ensuring that the insights derived are truly representative of human sentiment and discourse on the topic of obesity.

Conclusion

Twitter is an important source to gauge obesity-related sentiments and attitudes of the public, celebrities, and influential organizations. Sentiments regarding obesity were predominantly negative. Obesity was associated with racism, poorer life choices, and social evils such as illicit substance use and alcohol consumption. Influential politicians' negative portrayal of obesity among their colleagues may lead to poorer public sentiments. This has negative connotations for public health in general. The Conservative government's campaign for curbing the "epidemic" of obesity in Britain attracted criticism from the public. Obesity being a risk factor for severe COVID-19 also led to a negative portrayal on social media.

XSL•FO
RenderX

The study aims to contribute to a more comprehensive understanding of the role of social media in shaping public sentiment, attitudes, and health behaviors related to obesity. In addition, the findings will inform the development of evidence-based public health policies, prevention strategies, and treatment approaches that address the unique challenges posed by this global health issue in the context of social media and digital communication.

Data Availability

All data are available for download using the Twitter application programming interface (API). The datasets generated during and/or analyzed during this study are available in the X developer platform [34].

Authors' Contributions

JCC conceived the idea of the study. JCC, AW, and ZP designed the study. SSA downloaded the Twitter data and analyzed the dataset. SSA and AW interpreted the Results section. JCC, AW, HM, and SSA wrote the first draft of the paper. ZP supervised the whole project and critically reviewed and edited the paper. All authors approved the paper for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sentiment analysis for individual tweets.
[XLSX File (Microsoft Excel File), 11808 KB - jmir_v26i1e52142_app1.xlsx]

Multimedia Appendix 2
Results of cluster analysis.
[XLSX File (Microsoft Excel File), 4347 KB - jmir_v26i1e52142_app2.xlsx]

Multimedia Appendix 3
Intertopic distance map and topics over time.
[ZIP File (Zip Archive), 2323 KB - jmir_v26i1e52142_app3.zip]

References

1. Wong MCS, Huang J, Wang J, Chan PSF, Lok V, Chen X, et al. Global, regional and time-trend prevalence of central obesity: a systematic review and meta-analysis of 13.2 million subjects. Eur J Epidemiol 2020;35(7):673-683 [FREE Full text] [doi: 10.1007/s10654-020-00650-3] [Medline: 32448986]
2. Safaei M, Sundararajan EA, Driss M, Boulila W, Shapi'i A. Comput Biol Med 2021;136:104754 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104754] [Medline: 34426171]
3. Gunawardana S, Gunasinghe CB, Harshani MS, Seneviratne SN. Physical and psychosocial quality of life in children with overweight and obesity from Sri Lanka. BMC Public Health 2021;21(1):86 [FREE Full text] [doi: 10.1186/s12889-020-10104-w] [Medline: 33413215]
4. Endalifer ML, Diress G. Epidemiology, predisposing factors, biomarkers, and prevention mechanism of obesity: a systematic review. J Obes 2020;2020:6134362 [FREE Full text] [doi: 10.1155/2020/6134362] [Medline: 32566274]
5. Wu YK, Berry DC. Impact of weight stigma on physiological and psychological health outcomes for overweight and obese adults: a systematic review. J Adv Nurs 2018;74(5):1030-1042. [doi: 10.1111/jan.13511] [Medline: 29171076]
6. Rubino F, Puhl RM, Cummings DE, Eckel RH, Ryan DH, Mechanick JI, et al. Joint international consensus statement for ending stigma of obesity. Nat Med 2020;26(4):485-497 [FREE Full text] [doi: 10.1038/s41591-020-0803-x] [Medline: 32127716]
7. Pollack CC, Emond JA, O'Malley AJ, Byrd A, Green P, Miller K, et al. Characterizing the prevalence of obesity misinformation, factual content, stigma, and positivity on the social media platform reddit between 2011 and 2019: infodemiology study. J Med Internet Res 2022;24(12):e36729 [FREE Full text] [doi: 10.2196/36729] [Medline: 36583929]
8. Waqas A, Salminen J, Jung SG, Almerekhi H, Jansen BJ. Mapping online hate: a scientometric analysis on research trends and hotspots in research on online hate. PLoS One 2019;14(9):e0222194 [FREE Full text] [doi: 10.1371/journal.pone.0222194] [Medline: 31557227]
9. Zhang HT, Tham JS, Waheed M. The effects of receiving and expressing health information on social media during the COVID-19 infodemic: an online survey among malaysians. Int J Environ Res Public Health 2022;19(13):7991 [FREE Full text] [doi: 10.3390/ijerph19137991] [Medline: 35805646]

XSL•FO
RenderX

10. Shih SF, Liu CH, Liao LL, Osborne RH. Health literacy and the determinants of obesity: a population-based survey of sixth grade school children in Taiwan. BMC Public Health 2016;16:280 [FREE Full text] [doi: 10.1186/s12889-016-2879-2] [Medline: 27000035]
11. Lovrić B, Placento H, Farčić N, Lipič Baligač M, Mikšić Š, Mamić M, et al. Association between health literacy and prevalence of obesity, arterial hypertension, and diabetes mellitus. Int J Environ Res Public Health 2022;19(15):9002 [FREE Full text] [doi: 10.3390/ijerph19159002] [Medline: 35897372]
12. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
13. Hönings H, Knapp D, Nguy n BC, Richter D, Williams K, Dorsch I, et al. Health information diffusion on Twitter: the content and design of WHO tweets matter. Health Info Libr J 2022;39(1):22-35. [doi: 10.1111/hir.12361] [Medline: 33682996]
14. Lyu JC, Han EL, Luli GK. COVID-19 Vaccine-related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021;23(6):e24435 [FREE Full text] [doi: 10.2196/24435] [Medline: 34115608]
15. Barbieri F, Anke LE, Camacho-Collados J. XLM-T: Multilingual language models in twitterfor sentiment analysis and beyond. arXiv:2104.12250 2022.
16. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 2022.
17. Roesslein J. Tweepy: Twitter for Python!. GitHub. URL: https://github.com/tweepy/tweepy [accessed 2024-09-05]
18. Agbe E. NeatText:a simple NLP package for cleaning textual data and text preprocessing. GitHub. 2020. URL: https://github.com/Jcharis/neattext [accessed 2024-09-05]
19. Obesity, race/ethnicity, and COVID-19. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/obesity/data/obesity-and-covid-19.html [accessed 2024-09-05]
20. White BM, Melton C, Zareie P, Davis RL, Bednarczyk RA, Shaban-Nejad A. Exploring celebrity influence on public attitude towards the COVID-19 pandemic: social media shared sentiment analysis. BMJ Health Care Inform 2023;30(1):e100665 [FREE Full text] [doi: 10.1136/bmjhci-2022-100665] [Medline: 36810135]
21. Pratiwi H, Benkő R, Kusuma IY. Navigating the asthma network on Twitter: insights from social network and sentiment analysis. Digit Health 2024;10:20552076231224075 [FREE Full text] [doi: 10.1177/20552076231224075] [Medline: 38269370]
22. Pérez-Pérez M, Pérez-Rodríguez G, Fdez-Riverola F, Lourenço A. Using Twitter to understand the human bowel disease community: exploratory analysis of key topics. J Med Internet Res 2019;21(8):e12610 [FREE Full text] [doi: 10.2196/12610] [Medline: 31411142]
23. Modave F, Zhao Y, Krieger J, He Z, Guo Y, Huo J, et al. Understanding perceptions and attitudes in Breast Cancer discussions on Twitter. Stud Health Technol Inform 2019;264:1293-1297 [FREE Full text] [doi: 10.3233/SHTI190435] [Medline: 31438134]
24. Puhl RM, Heuer CA. The stigma of obesity: a review and update. Obesity (Silver Spring) 2009;17(5):941-964 [FREE Full text] [doi: 10.1038/oby.2008.636] [Medline: 19165161]
25. Puhl R, Suh Y. Health consequences of weight stigma: implications for obesity prevention and treatment. Curr Obes Rep 2015;4(2):182-190. [doi: 10.1007/s13679-015-0153-z] [Medline: 26627213]
26. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. Obes Rev 2015;16(4):319-326 [FREE Full text] [doi: 10.1111/obr.12266] [Medline: 25752756]
27. Stanford FC, Tauqeer Z, Kyle TK. Media and its influence on obesity. Curr Obes Rep 2018;7(2):186-192 [FREE Full text] [doi: 10.1007/s13679-018-0304-0] [Medline: 29637412]
28. Browne NT, Hodges EA, Small L, Snethen JA, Frenn M, Irving SY, et al. Childhood obesity within the lens of racism. Pediatr Obes 2022;17(5):e12878. [doi: 10.1111/ijpo.12878] [Medline: 34927392]
29. Shah H, Bohlen J. Implicit Bias. St. Petersburg, FL: StatPearls Publishing LLC; 2023.
30. Little JS, Romee R. Tweeting from the bench: Twitter and the physician-scientist benefits and challenges. Curr Hematol Malig Rep 2020;15(6):419-423 [FREE Full text] [doi: 10.1007/s11899-020-00601-5] [Medline: 33179209]
31. Klar S, Krupnikov Y, Ryan JB, Searles K, Shmargad Y. Using social media to promote academic research: identifying the benefits of twitter for sharing academic work. PLoS One 2020;15(4):e0229446 [FREE Full text] [doi: 10.1371/journal.pone.0229446] [Medline: 32251463]
32. New obesity strategy unveiled as country urged to lose weight to beat coronavirus (COVID-19) and protect the NHS [press release]. GOV.UK. 2020. URL: https://www.gov.uk/government/news/ [accessed 2024-09-05]
33. Kyle TK, Dhurandhar EJ, Allison DB. Regarding obesity as a disease: evolving policies and their implications. Endocrinol Metab Clin North Am 2016;45(3):511-520 [FREE Full text] [doi: 10.1016/j.ecl.2016.04.004] [Medline: 27519127]
34. Academic research. X Developer Platform. URL: https://developer.x.com/en/use-cases/do-research/academic-research [accessed 2024-09-27]

XSL•FO
RenderX

Abbreviations

API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
CDC: Centers for Disease Control and Prevention
DBSCAN: Density-Based Spatial Clustering of Applications with Noise
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
UMAP: Uniform Manifold Approximation and Projection
UNICEF: United Nations International Children's Emergency Fund

Edited by A Mavragani; submitted 24.08.23; peer-reviewed by H Sun, C Zheng, H Veldandi; comments to author 16.12.23; revised version received 06.02.24; accepted 27.06.24; published 11.10.24.

Please cite as:
Correia JC, Ahmad SS, Waqas A, Meraj H, Pataky Z
Exploring Public Emotions on Obesity During the COVID-19 Pandemic Using Sentiment Analysis and Topic Modeling: Cross-Sectional Study
J Med Internet Res 2024;26:e52142
URL: https://www.jmir.org/2024/1/e52142
doi:10.2196/52142
PMID:

©Jorge César Correia, Sarmad Shaharyar Ahmad, Ahmed Waqas, Hafsa Meraj, Zoltan Pataky. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Concurrent Media Exposure on Professional Identity: Cross-Sectional Study of 1087 Medical Students During Long COVID

Manli Wu[1], PhD; Jun Yan[1], PhD; Chongming Qiao[2], BA; Chu Yan[3], BA

[1]School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan, China
[2]School of Media and Communication, Shanghai Jiao Tong University, Shanghai, China
[3]School of Marxism, Huazhong University of Science and Technology, Wuhan, China

Corresponding Author:
Jun Yan, PhD
School of Journalism and Information Communication
Huazhong University of Science and Technology
School of Journalism and Information Communication, HUST
Number 1037, Luoyu Road
Wuhan, 430074
China
Phone: 86 18971629738
Email: junyan@hust.edu.cn

Abstract

Background: Long COVID has widened the health gap across society and highlighted the vulnerabilities and risks faced by health care systems. For instance, the global trend of medical workers resigning has become a prominent topic on social media. In response to this severe social problem in global public health within the digital society, it is urgent to investigate how the professional identity of medical students, who are digital natives and the future workforce of medical practitioners, is affected by the media environment.

Objective: This study aims to examine how media exposure relates to medical students' perceptions of informational and emotional support, and how these perceptions further influence the development of their professional identity.

Methods: Building on the Stimulus-Organism-Response (SOR) framework, this study develops a theoretical model to illustrate how media exposure affects medical students' professional identity through the mediation of social support. Specifically, media exposure was assessed through online news media and social media exposure; social support was evaluated in terms of informational and emotional support; and professional identity was measured through medical students' sense of belonging and professional commitment. A survey was conducted at a medical school in China, yielding 1087 valid responses that were analyzed using SmartPLS 4.0.

Results: Consistent with our expectations, online news media exposure was positively associated with both informational support (β=.163; P<.001) and emotional support (β=.084; P=.007). Similarly, social media exposure showed positive associations with informational support (β=.122; P<.001) and emotional support (β=.235; P<.001). Thereafter, informational support (β=.228; P<.001) and emotional support (β=.344; P<.001) were positively associated with students' sense of belonging. Meanwhile, both informational support (β=.245; P<.001) and emotional support (β=.412; P<.001) positively impacted medical students' professional commitment. In addition, a mediation test was conducted. The results confirmed that informational support and emotional support partially mediated the effect of online news media, while fully mediating the effect of social media on medical students' sense of belonging and professional commitment.

Conclusions: This study finds that exposure to online news media and social media can enhance medical students' sense of belonging and professional commitment through the formation of informational and emotional support. It expands the discussion on the role of media in providing social support and facilitating the development of medical students' professional identity. This is a valuable contribution to addressing complex public health crises through effective media governance in the network era.

(J Med Internet Res 2024;26:e50057) doi:10.2196/50057

XSL•FO
RenderX

KEYWORDS

COVID-19; media exposure; social support; professional identity; medical students; Stimulus-Organism-Response framework

Introduction

Background

Media factors in medical sociology have long been recognized as significant in influencing health management and medical outcomes. For example, media and media exposure have been shown to play a key role in heart disease prevention, vaccination, community care, and other health interventions [1,2]. The COVID-19 pandemic and long COVID in the digital age have caused unprecedented social disruptions due to the widespread use of social media and the circulation of mixed information. Therefore, it is essential to investigate how these new "media factors" impact the stability and development of society.

Furthermore, the prevalence of COVID-19 is accompanied by an infodemic [3]. The disaster stories about COVID-19 on social media serve not only as a significant force to unite the public in combating the pandemic but also as hidden triggers for the widespread resignation of health care workers. According to the World Health Organization (WHO), there will be a shortfall of 10 million health workers by 2030 [4]. Medical students are a vital reserve force for health care workers and play an essential role in public health events. For instance, during the COVID-19 pandemic, medical students' involvement in patient care and health consultation complemented the work of medical professionals [5,6]. Their professional identity is critical to their career perception and choices [7], which may further impact the stability of the medical workforce.

Professional identity refers to the self-conceptualization of one's work role [8] and is central to professional loyalty. In contrast to experienced medical professionals, medical students are young, vulnerable, and digital natives, with their professional identities still in formation and transformation [9]. Prior studies have typically applied descriptive analyses to explain the development of professional identity [10,11]. Although current studies are beginning to use cross-sectional methods to explore predictors of professional identity, they primarily discuss individual and educational environmental factors, such as clinical practice [12], experiences within the educational system [13], and working conditions [14]. Recently, there has been growing evidence suggesting that online media play an important role in influencing students' knowledge and attitudes regarding the medical profession [15]; however, the theoretical mechanisms behind this influence remain unclear. Given that the COVID-19 pandemic has highlighted the impact of media on people's daily lives [16], this research speculates that exposure to online media during the pandemic could also influence medical students' professional identity. As online news media and social media are the most commonly used types of media, this study proposes the following research question: *How does concurrent media exposure affect medical students' professional identity?*

In public health events such as the COVID-19 pandemic, the media play a crucial role in information distribution, pandemic warnings, gauging public emotions, facilitating social mobilization, and providing emotional support [17]. Specifically, online media not only deliver medical-related information [18] but also feature news reports that portray the professional and heroic images of health care workers [19]. Additionally, online media facilitate social interactions [20], allowing netizens to express their support for health care workers through understanding, encouragement, affirmation, sympathy, and care, among other forms [21]. As such, the social support gained through online media exposure may help medical students recognize the value of medical groups [22], develop their attitudes toward the medical profession, and understand public opinions about the field [23]. This study utilizes the Stimulus-Organism-Response (SOR) framework [24] to explain how media exposure relates to medical students' professional identity, emphasizing the mediating role of social support.

To address the proposed research question, a survey method is used, yielding a total of 1087 valid responses. The results reveal that media exposure during the COVID-19 pandemic affects medical students' professional identity, either directly or indirectly, through social support. Theoretically, this study provides a comprehensive understanding of the effects of media exposure on the development of medical students' professional identity by extending the SOR framework. Practically, this study underscores the importance of media and information management in fostering professional identity, offering a viable approach to addressing the shortage of the medical workforce, promoting social unity, and reducing health disparities.

Objectives

The objective of this study was to examine the effect of concurrent media exposure on medical students' professional identity, with a specific focus on the mediating role of social support. In particular, this study investigated how concurrent media exposure (ie, online news media and social media) influenced social support (ie, informational support and emotional support), which in turn affected the development of medical students' professional identity (ie, sense of belonging and professional commitment).

Theoretical Background

Professional Identity

As a crucial link between individuals and society, professional identity is an essential force in fostering social unity and mitigating social risks. Professional identity encompasses the attitudes, values, knowledge, beliefs, and skills shared among members of a professional group [8]. It not only influences individuals' career choices [25] but also impacts their job satisfaction [26] and turnover intentions [27]. Prior studies have typically focused on specific professions, such as teachers, lawyers, and health care workers [26,28,29]. Professional identity is especially significant in medical education, as it can influence the development of the medical workforce [25].

Although prior studies have generally examined professional identity as an integrated concept [26,30], it can also be understood as a complex structure [31]. Individuals typically

XSL•FO
RenderX

develop their identity based on their feelings and behaviors toward the profession [32]. To gain a nuanced understanding of professional identity, this study uses belonging and professional commitment as its key components. Belonging refers to the feeling of being an integral part of a professional group, encompassing perceptions of acceptance, security, and respect within the specific profession [33]. Professional commitment reflects both attitudes and behaviors toward the profession, indicating the relative strength of an individual's identification with their profession and their willingness to invest effort in and maintain membership within it [34]. Professional commitment reflects the alignment between personal beliefs and professional goals [35], which can lead individuals to invest effort and develop a sense of attachment to their profession. It comprises 3 dimensions: affective commitment, continuance commitment, and normative commitment. Affective commitment pertains to individuals' attitudes or feelings toward the profession, while continuance and normative commitment relate to individuals' behavioral intentions regarding their profession [36].

Belonging and professional commitment represent different aspects of professional identity, and it is important to recognize the distinctions between them. Belonging is a psychological concept that reflects the relationship between an individual and others within a particular group [33,37], while professional commitment measures the intensity of an individual's connection to their profession, characterized by the willingness to invest effort and the desire to remain in that profession [36]. In this context, belonging emphasizes individuals' psychological ownership of a group [37], while professional commitment serves as a stabilizing or obliging force that guides one's professional behavior [38].

SOR Framework

Overview

Derived from environmental psychology, the SOR framework posits that environmental factors can trigger cognitive and affective reactions in individuals, which subsequently influence their behaviors [24]. The SOR framework consists of 3 interconnected components: stimulus, organism, and response. Specifically, the stimulus refers to the environmental factors that impact individuals' psychological states [24]. Response refers to behavior in reaction to a certain stimulus [24,39]. The organism, consisting of cognitive and affective states [39], is often considered the mediator between stimulus and response [40,41]. Cognitive states involve the acquisition, processing, retention, and retrieval of information, while affective states include pleasure, arousal, dominance [24], and other emotion typologies relevant to specific contexts [39].

The SOR framework has been widely used to explain individuals' behaviors. For instance, previous research has used this framework to assess the impacts of social media overload [42] and technological features [43] on users' behaviors. In a recent study, the SOR framework was applied to examine the impact of the COVID-19 pandemic on behavioral changes among university students [44]. This study adopts the SOR framework for 2 reasons. First, the SOR framework delineates the stages of behavior formation by highlighting how external cues influence individuals' responses through changes in their internal states [24], thereby illuminating the development of medical students' professional identity. Second, the organism component encompasses both cognitive and affective reactions [39], allowing us to differentiate medical students' psychological responses resulting from media exposure.

Along with the rationale behind the SOR framework, this study treats media exposure as an external stimulus. Given that media exposure is a significant source of social support [45,46], this study conceptualizes social support as the organism component, with emotional support representing the affective aspect and informational support reflecting the cognitive aspect. Furthermore, professional identity is viewed as the response of medical students in reaction to media exposure.

Media Exposure as Stimulus

Media exposure refers to the frequency with which individuals are exposed to specific media [47]. It has been recognized as a critical predictor of various behaviors, including political participation [48], sexual behavior [49], and alcohol consumption [50]. During the COVID-19 pandemic, people increasingly relied on media to obtain real-time information for making health decisions [51]. Existing literature has validated the effects of media exposure on the public's preventive behaviors [1]. In general, media exposure can influence individuals' cognition and behaviors in the age of COVID-19. Through media exposure, medical students can learn about the responsibilities and challenges faced by medical workers [19], which may impact their professional identity.

Among all media categories, online news media and social media are the 2 most representative and frequently used channels in the new media environment [20]. Unlike traditional news media, such as television and newspapers, online news media rely on the internet to deliver news content. As a result of the news gatekeeping mechanisms, content on online news media typically originates from authoritative sources, reflecting official attitudes and the mainstream values of society [20]. Therefore, people often place their trust in online news media. By contrast, anyone can post and share information on social media, and quality control over content is limited [52]. As social media facilitate online interactions, they have become a prominent source of information during COVID-19 [53]. This study simultaneously considers online news media and social media as media stimuli, as they are the primary information sources and may influence medical students differently.

Social Support as an Organism

Through online media, social support can be delivered via information transmission and online interactions [53]. Social support refers to the tangible or intangible resources obtained from others through interpersonal ties within a social network [54]. Prior studies have confirmed the importance of social support in coping with stress and depression [55], improving psychological health [56], and enhancing life satisfaction [57]. Social support is available through online media in various forms, including companionship, solidarity, and information exchange [46,58]. Among all categories of social support, informational support and emotional support are the most frequently mentioned [45,59,60]. Informational support

XSL•FO
RenderX

emphasizes the provision of information through giving advice, offering appraisal support, sharing new information, or providing references [59,61]. Emotional support, by contrast, refers to the affective assistance characterized by caring, empathy, love, and trust [61]. Although both informational support and emotional support are psychological variables reflecting users' perceptions of support availability, informational support focuses on problem-solving, while emotional support pertains to affective assistance [62]. Therefore, these 2 dimensions may have differential impacts. Consequently, this study treats social support as the organism component, with emotional support representing the affective aspect and informational support representing the cognitive aspect.

Hypotheses Development

Study Rationale

This study examines how media exposure influences medical students' professional identity, with a particular focus on the underlying mechanisms. Based on the rationale of the SOR framework, exposure to online news media and social media can impact both informational and emotional support, which, in turn, affects medical students' professional identity.

The Effects of Online News Media Exposure on Social Support

Online news media play a crucial role in providing pandemic-related information. Given that people tend to have a high level of trust in information from authoritative sources [2,20], medical students are more likely to trust and adopt information presented in online news media. This exposure allows medical students to acquire professional knowledge to navigate the pandemic and to learn about the contributions and sacrifices of health care workers [19]. Therefore, exposure to online news media can offer valuable informational support to medical students.

In addition, news content on online news media typically embodies mainstream values and reflects official attitudes [20]. As a result of the dedication of health care workers during the pandemic, medical groups are often highly praised in news reports [19]. Exposure to online news media may help medical students recognize that health care workers are respected and valued by the public. Furthermore, by allowing readers to comment on specific news reports, online news media serve as channels for the public to express their respect for health care workers. Thus, online news media can also provide emotional support to medical students. Therefore, the following hypotheses are proposed:

- H1a: Online news media exposure is positively associated with informational support.
- H1b: Online news media exposure is positively associated with emotional support.

The Effects of Social Media Exposure on Social Support

With abundant user-generated content and multiple information sources, social media play an important role in the dissemination of medical information [63]. Through social media, exposure to content related to medical workers not only helps medical students protect their health but also enriches their professional knowledge. Furthermore, as mutual trust can be established through online interactions [20,22], medical students likely place significant value on information shared by individuals within their online social networks.

Moreover, exposure to social media can facilitate emotional communication through social network connections [64]. Users can share personal feelings with one another through online interactions. In doing so, emotional support—such as understanding, encouragement, affirmation, sympathy, or care—is conveyed [21]. During the pandemic, social media enabled medical students to stay in contact with their acquaintances and support one another. Therefore, the following hypotheses are proposed:

- H2a: Social media exposure is positively associated with informational support.
- H2b: Social media exposure is positively associated with emotional support.

The Effects of Informational Support on Professional Identity

During the pandemic, information provided by medical professionals not only helped the public but also benefited medical students. Individuals who receive help from others are more likely to develop a positive attitude toward those providing the assistance [14]. Belonging refers to the extent to which individuals integrate themselves into their recognized group [37]. Thus, medical students who receive informational support from medical professionals may come to appreciate the values of medical groups and foster long-term relationships with them. In this way, a sense of belonging is cultivated.

Information about the pandemic can also offer medical students guidelines for future practice and deepen their understanding of medical practitioners' responsibilities. These perceived messages help foster acceptance of professional norms and recognition of the obligations within the "psychological contract" established with the medical profession, both of which are crucial components of commitment [36,38]. In this context, informational support accelerates the development of medical students' professional commitment. Therefore, the following hypotheses are proposed:

- H3a: Informational support is positively associated with medical students' sense of belonging to the medical profession.
- H3b: Informational support is positively associated with medical students' professional commitment.

The Effects of Emotional Support on Professional Identity

Emotional support also plays a key role in shaping medical students' attitudes toward medical groups and the profession. Previous research has shown that emotional support from colleagues can strengthen an individual's attachment to the group [60]. Specifically, trust and respect toward medical workers can instill pride in the profession, further enhancing medical students' identification as part of the medical community [23]. In this way, emotional support can reinforce medical students' sense of belonging to the medical profession.

XSL•FO
RenderX

Emotional support can also influence professional commitment. For example, it has been shown to reduce burnout among nurses [65] and promote their commitment to the profession [66]. Similarly, emotional support may strengthen medical students' professional commitment by improving their psychological well-being. According to Ahmad et al [67], a supportive environment can increase student engagement and enhance their commitment to their profession. Emotional support from social networks can help create such an environment, further boosting medical students' professional commitment. The following hypotheses are proposed:

- H4a: Emotional support is positively associated with medical students' sense of belonging to the medical profession.
- H4b: Emotional support is positively associated with medical students' professional commitment.

The Mediation Effects of Social Support

Media exposure can influence the development of individuals' professional identity. Evidence indicates that exposure to media representations related to one's identity can affect their understanding of social groups [19,68] and shape their identification with a specific identity [69]. Notably, Geusens and Beullens [69] proposed that media might impact the perception of one's identity in an indirect manner. This study uses social support as a potential explanatory mechanism to clarify the relationship between media exposure and professional identity. Therefore, it hypothesizes that media exposure may indirectly influence medical students' professional identity through the mediation of social support. The following hypotheses are proposed:

- H5: Informational support mediates the effects of online news media on medical students' sense of belonging (a) and professional commitment (b).
- H6: Emotional support mediates the effects of online news media on medical students' sense of belonging (a) and professional commitment (b).
- H7: Informational support mediates the effects of social media on medical students' sense of belonging (a) and professional commitment (b).
- H8: Emotional support mediates the effects of social media on medical students' sense of belonging (a) and professional commitment (b).

This study also included gender, age, education, and early internship experience as control variables, as they may affect medical students' professional identity [8,12]. The research model is depicted in Figure 1.

Figure 1. Research model.

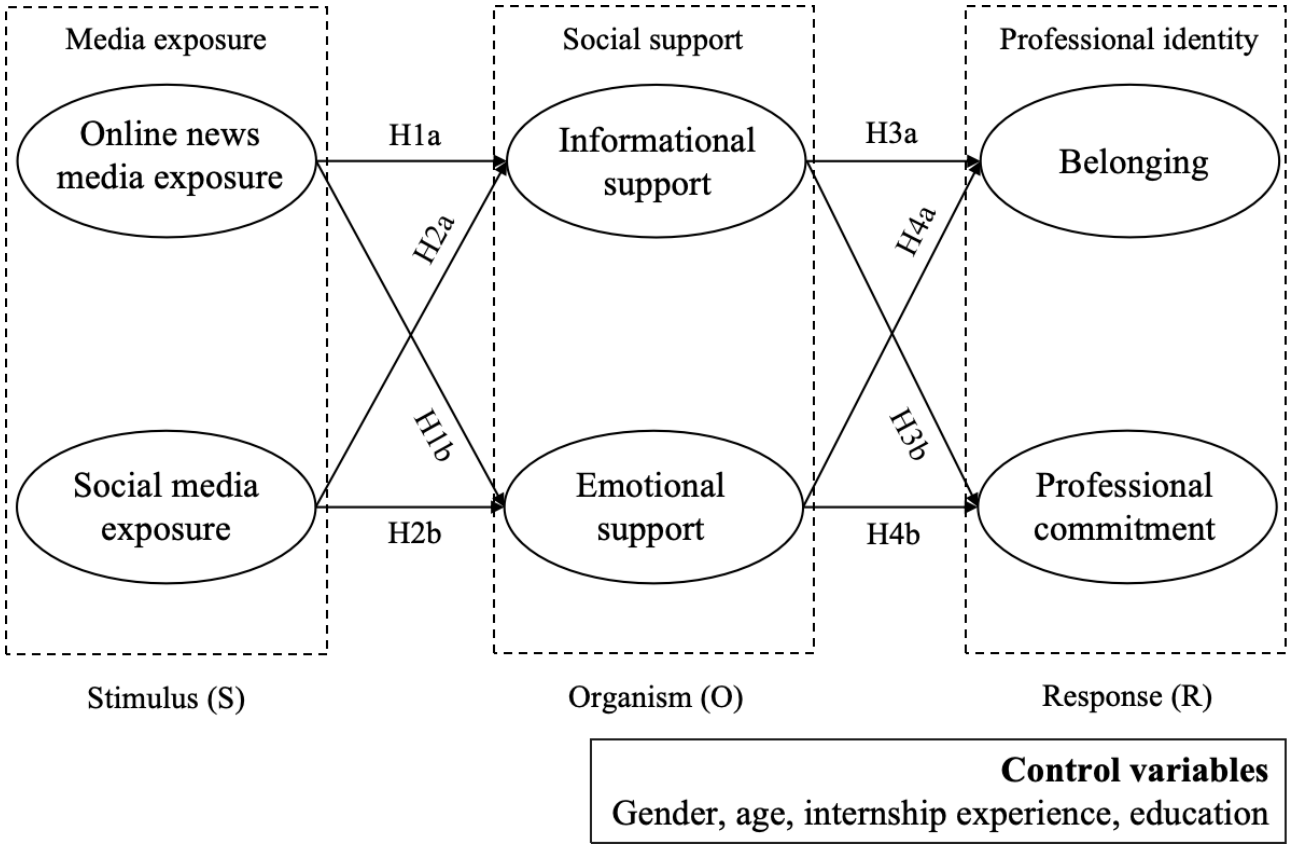

Methods

Survey Design

To empirically test the proposed research model, a cross-sectional survey was conducted to collect data. The survey design adhered to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [70] (see Multimedia Appendix 1 for the completed CHERRIES). A convenience sampling method was used, inviting medical students who were readily available to participate. Participants were recruited with the assistance of faculty members at a medical college. Specifically, the study team shared information about the survey

XSL•FO
RenderX

in several medical courses and invited students to participate voluntarily. They were informed that their anonymity would be ensured to encourage honest responses and reduce social desirability bias [71]. The survey was conducted in May 2022.

Overall, the main body of the questionnaire consisted of 4 sections: (1) media exposure; (2) social support; (3) professional identity; and (4) demographic characteristics. Additionally, an introductory page was included at the beginning of the questionnaire to inform participants about the research purpose, target population, length of the survey, and policies regarding data security and storage. Furthermore, all participants were required to provide informed consent.

Ethical Considerations

The questionnaire and methodology for this study were approved by the Ethics Committee of Tongji Medical College of Huazhong University of Science and Technology (approval number 2022S009). Informed consent was obtained from all participants. Measures were implemented to protect the privacy of all participants and ensure the confidentiality of the data. The questionnaire was anonymous, and the data file was stored on a secure drive accessible only to research team members who signed a confidentiality agreement. Each participant received a compensation of 10 RMB (approximately US $1.5) for their valid response.

Measurement Development

This study used several strategies to develop the measurement instruments. First, to ensure the content validity of the instrument, measures for the variables were primarily adapted from well-validated scales and modified to fit the Chinese context and the characteristics of medical students. The back-translation method was used to adapt the original English scale to Chinese and ensure consistency between the 2 language versions. Specifically, the questions were first framed in English, then translated into Chinese, and finally back-translated into English to confirm consistency.

Second, we conducted several rounds of pretests to refine the instruments. Three professors and a group of graduate students from a medical school were invited to participate in these pretests. They completed the questionnaires and took part in follow-up interviews to evaluate whether the instruments effectively measured the constructs under investigation. Participants were also asked to provide feedback on the wording of the questions, clarity of expression, and question sequence. Based on their feedback, we enhanced the face validity and content validity of the instruments.

Third, we conducted a pilot test to assess the reliability and validity of the instruments. The results of the data analysis (shown in Multimedia Appendix 2; also see [72-75]) indicated that all measurements demonstrated ideal reliability and validity. Additionally, the measurement model analysis in the formal test further confirmed the reliability and validity of the instruments. The measures for the constructs are provided as follows.

Measures

Social Media Exposure and Online News Media Exposure

The measures for media exposure were adapted from Gao et al [76]. Five items were used to assess medical students' media exposure on a 7-point Likert scale (1=never to 7=very often). Participants were asked to evaluate the frequency of their exposure to information about medical workers and the medical profession (eg, health information from medical professionals, news stories or comments about medical workers, or information regarding medical education) on online news media and social media during the COVID-19 pandemic.

Considering the Chinese media environment and the media usage habits of medical students, this study included common news media (eg, Xinhuanet and Tencent News) as well as medical news media (eg, Medsci and PubMed) as components of online news media. Social media included Sina Weibo (Weibo Corporation), WeChat (Tencent Holdings Limited), and QQ (Shenzhen Tencent Computer System Co., Ltd.), as these 3 platforms are representative applications among Chinese youth [77].

Informational Support and Emotional Support

Informational support and emotional support were measured using items adapted from Nick et al [59]. Three items were used to assess each type of support through a 7-point Likert scale (1=never to 7=very often). To measure informational support, participants were asked to evaluate the frequency with which they encountered the following situations online in the past year: (1) people provided me with helpful medical information online; (2) I received help from others online when I had difficulties with my medical learning; and (3) people helped me understand my profession better while I was online. To measure emotional support, participants were asked to evaluate the frequency with which they encountered the following situations on the internet in the past year: (1) people showed that they cared about medical workers online; (2) people supported the work of medical workers online; and (3) individuals on the internet liked the things that medical workers said or did.

Belonging and Professional Commitment

The existing literature often uses self-reported measures to evaluate professional identity. For instance, Heidari et al [78] used a cross-sectional survey to examine the effects of online social capital and social networking on the formation of students' professional identities. Similarly, Zhang et al [79] conducted an online cross-sectional survey to explore the relationship between psychological resilience and the sense of professional identity. Researchers typically regard professional identity as a psychological construct that emphasizes individuals' professional self-concept [6]. As professional identity is a complex structure [31] that encompasses individuals' feelings and behaviors toward their profession [32], this study uses belonging and professional commitment as indicators of professional identity and measures medical students' feelings and behavioral intentions toward the profession.

XSL•FO
RenderX

Belonging was measured using items adapted from Lin et al [80]. Participants were asked to indicate the extent to which they agree with the following statements, reflecting their social affiliation and emotional connection with the medical group: (1) I belong to the medical group; (2) I feel socially connected to medical workers; and (3) I will be part of the medical group in the future. All items were measured on a 7-point Likert scale (1=strongly disagree to 7=strongly agree).

As mentioned above, professional commitment consists of 3 dimensions. Despite their conceptual differences, these dimensions are simultaneously experienced by individuals [38], reflecting their attitudes and behavioral intentions. Therefore, it is appropriate to measure these 3 dimensions as components of professional commitment. The measurement instruments were adapted from Meyer et al [36] and modified to align with the medical profession. Participants were asked to rate their level of agreement with the following statements: (1) I feel proud to be a medical worker and have no regrets about pursuing the medical profession; (2) I do not regret entering the medical profession; (3) It would be costly for me to change my profession now; (4) There are no pressures preventing me from changing my profession; and (5) I feel a responsibility to continue in the medical profession. Notably, the first 2 items represent affective commitment, while the third and fourth items reflect normative commitment. The last item is designed to capture participants' continuance commitment. All items were measured on a 7-point Likert scale (1=strongly disagree to 7=strongly agree).

Data Collection and Samples

Medical students from 4 majors (clinical medicine, basic medicine, nursing, and public health) at a medical college in China were recruited to complete a set of paper questionnaires. We selected these 4 majors due to their critical roles in the frontlines of the COVID-19 pandemic. Specifically, doctors, nurses, and epidemiologists from clinical medicine, nursing, and public health actively participated in the prevention and control of the pandemic. Students majoring in basic medicine were also included, as they play a vital role in combating the new coronavirus. By choosing a diverse range of majors, we aimed to enhance the generalizability of this study.

In the formal data collection process, a screening question was included to allow only medical students with online media usage experience to participate in the survey. Notably, participants from the pilot test were not permitted to complete the questionnaire. The survey included questions regarding demographic information, professional identity, and factors related to media exposure and social support. Additionally, we incorporated attention-trap questions to ensure the quality of the responses. A total of 1200 questionnaires were collected. After removing 113 invalid responses (eg, those that failed the attention test, contained missing values, or had excessively duplicated answers), we finalized a sample of 1087 valid responses, with students from the 4 majors approximately evenly distributed.

Data Analysis

This study used structural equation modeling to test the research framework. We selected SmartPLS 4.0 (SmartPLS GmbH), based on partial least squares, as it allows for simultaneous testing of both the measurement model and the structural model [81]. In the current model, online news media exposure and social media exposure were incorporated as independent variables; informational support and emotional support served as mediators; and professional identity (ie, belonging and professional commitment) variables were included as outcomes. Following the approach suggested by Anderson and Gerbing [82], we first analyzed the measurement model to assess reliability and validity, and subsequently conducted structural model analysis to evaluate our research model. Furthermore, a mediation analysis was conducted to uncover the underlying mechanisms of media exposure's influence. To account for potential confounding effects, gender, age, education, and internship experience were included as control variables in the structural equation model.

Results

Demographic Information

As mentioned above, a final sample of 1087 valid responses was obtained. The demographic information of this sample is summarized in Table 1. More than half of the students were female and pursuing an undergraduate degree. Additionally, most of the students were in their 20s or younger and reported having internship experiences during their studies.

XSL•FO
RenderX

Table 1. Demographic information (N=1087).

Characteristic	Values, n (%)
Gender	
Male	461 (42.41)
Female	626 (57.59)
Age (years)	
<22	701 (64.49)
23-25	278 (25.57)
26-29	101 (9.29)
30-40	7 (0.64)
Internship (years)	
0	252 (23.18)
0.5	327 (30.08)
0.6-1.5	306 (28.15)
1.6-3	179 (16.47)
>3	23 (2.12)
Education	
Undergraduate	756 (69.55)
Master's	221 (20.33)
Doctor	110 (10.12)

Measurement Model Test

This study assesses the convergent and discriminant validity to evaluate the measurement model. The descriptive statistics of the variables in our research model and the results are presented in Table 2. Notably, Cronbach α for all constructs exceeded 0.7. Additionally, the composite reliability values for all constructs were greater than 0.8, and all average variance extracted (AVE) values were above 0.5, meeting the recommended threshold values of 0.7 and 0.5, respectively [72,73]. In addition, all item loadings, except for the one related to professional commitment, exceeded 0.7. As this factor loading was close to 0.7, we retained it to consider content validity. Thus, the model demonstrates acceptable reliability and convergent validity. We further assessed discriminant validity by comparing the interconstruct correlation coefficients with the square roots of AVE [72]. The results in Table 3 indicate that the square roots of AVE exceed all interconstruct correlation coefficients, confirming that discriminant validity is not an issue.

This study also examined multicollinearity. As a rule, multicollinearity is considered high if the variance inflation factor of a variable exceeds 10 [72]. In this study, the variance inflation factor values for all variables were below 3, indicating the absence of significant multicollinearity. Additionally, because all data were collected from a cross-sectional survey, common method bias (CMB) was assessed. First, we conducted the Harman single-factor test using SPSS Statistics 26.0 (IBM Corp.) and found that the most important factor explained only 23.33% of the total variance, which is lower than the reference value of 50% [74]. Second, CMB may lead to high correlations between constructs. Table 3 shows that the highest correlation between constructs was 0.675, which is below the recommended threshold of 0.90 [75]. Therefore, CMB was not a significant threat in this study.

XSL•FO
RenderX

Table 2. Reliability and convergent validity analysis.

Item	VIF[a]	Item loading	Mean (SD)	AVE[b]	CR[c]	Cronbach α
Online news media exposure				0.768	0.868	0.713
Online news media exposure 1	1.442	0.935	3.74 (1.82)			
Online news media exposure 2	1.442	0.814	3.34 (1.78)			
Social media exposure				0.647	0.845	0.726
Social media exposure 1	1.323	0.733	4.49 (2.04)			
Social media exposure 2	1.878	0.909	4.92 (1.68)			
Social media exposure 3	1.571	0.759	4.38 (1.81)			
Informational support				0.755	0.902	0.838
Informational support 1	2.107	0.892	5.12 (1.30)			
Informational support 2	2.017	0.857	5.02 (1.34)			
Informational support 3	1.837	0.858	5.18 (1.29)			
Emotional support				0.752	0.900	0.834
Emotional support 1	2.189	0.874	5.53 (1.31)			
Emotional support 2	2.785	0.918	5.58 (1.25)			
Emotional support 3	1.700	0.805	5.29 (1.40)			
Belonging				0.804	0.925	0.878
Belonging 1	2.343	0.888	5.05 (1.59)			
Belonging 2	2.618	0.919	5.52 (1.39)			
Belonging 3	2.333	0.883	5.29 (1.55)			
Professional commitment				0.581	0.873	0.819
Professional commitment 1	1.999	0.835	5.16 (1.49)			
Professional commitment 2	2.083	0.789	4.69 (1.60)			
Professional commitment 3	1.484	0.678	5.40 (1.56)			
Professional commitment 4	1.668	0.757	5.60 (1.47)			
Professional commitment 5	1.747	0.743	4.73 (1.57)			

[a]VIF: variance inflation factor.
[b]AVE: average variance extracted.
[c]CR: composite reliability.

Table 3. Discriminant validity analysis.

Analyzed variables	Online news media exposure	Social media exposure	Informational support	Emotional support	Belonging	Professional commitment
Online news media exposure	*0 .876* [a]	—[b]	—	—	—	—
Social media exposure	0.260	*0.804* [a]	—	—	—	—
Informational support	0.195	0.164	*0.869* [a]	—	—	—
Emotional support	0.145	0.256	0.593	*0.867* [a]	—	—
Belonging	0.160	0.188	0.437	0.473	*0.897* [a]	—
Professional commitment	0.165	0.189	0.494	0.552	0.675	*0.762* [a]

[a]The square roots of average variances extracted are in italics.
[b]Not applicable.

XSL•FO
RenderX

Structural Model Test

The results of the structural model test are shown in Figure 2. Consistent with our expectations, online news media exposure was positively associated with informational support (β=.163; *P*<.001) and emotional support (β=.084; *P*=.007). Additionally, social media exposure was positively associated with informational support (β=.122; *P*<.001) and emotional support (β=.235; *P*<.001). Therefore, H1a, H1b, H2a, and H2b are all supported.

Thereafter, this study analyzed the impact of social support on medical students' professional identity. The results showed that informational support (β=.228; *P*<.001) and emotional support (β=.344; *P*<.001) were positively associated with belonging. Meanwhile, we found that informational support (β=.245; *P*<.001) and emotional support (β=.412; *P*<.001) were positively associated with medical students' professional commitment. Overall, the model explained 28.1% of the variance in belonging and 35.6% of the variance in professional commitment. Therefore, H3a, H3b, H4a, and H4b are all supported.

This study also tested the impact of demographic variables on medical students' professional identity. The results showed that the impacts of gender on belonging (β=.064; *P*=.22) and professional commitment (β=–.008; *P*=.87) were insignificant. Meanwhile, the impacts of age on belonging (β=–.074; *P*=.08) and professional commitment (β=–.053; *P*=.17) were also insignificant. Furthermore, the impacts of education on belonging (β=–.061; *P*=.15) and professional commitment (β=–.022; *P*=.54) were likewise insignificant. Notably, internship experience had positive effects on both belonging (β=.114; *P*<.001) and professional commitment (β=.091; *P*<.001).

Figure 2. Data analysis results. [a]*P*<.01; [b]*P*<.001.

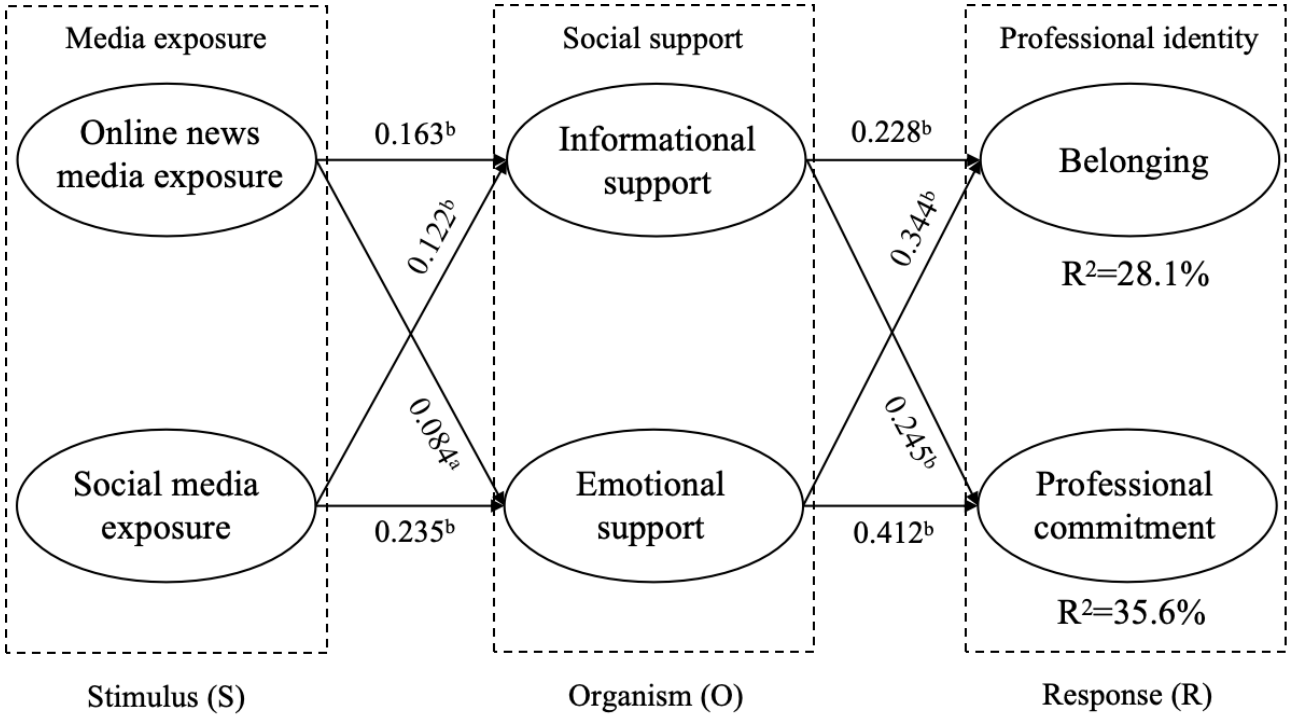

Mediation Analysis

The mediation effects of social support were further assessed following Hair et al [83] and Zhao et al [84]. Table 4 shows the results of our mediation analysis. The results indicated that specific indirect effects for both independent variables were significant, confirming the mediation effect of social support. Additionally, online news media exposure had significant direct effects on both belonging (β=.076; *P*=.01) and professional commitment (β=.059; *P*=.03). Therefore, informational support and emotional support partially mediated the relationship between online news media exposure and medical students' professional identity [81]. Thus, H5a, H5b, H6a, and H6b were supported. In addition, as the direct effects of social media on belonging (β=.060; *P*=.06) and professional commitment (β=.045; *P*=.13) were both insignificant, social support fully mediated the relationship between social media exposure and medical students' professional identity [81]. As such, H7a, H7b, H8a, and H8b were all supported.

XSL•FO
RenderX

Table 4. Results of the mediation analysis.

Independent variable and relationship	Specific indirect effect (*P* value)	Direct effect (*P* value)	Mediation effect
Online news media exposure			
Online news media exposure→Informational support→Belonging	0.035 (<.001)	0.076 (.01)	Partial mediation
Online news media exposure→Emotional support→Belonging	0.029 (.011)	0.076 (.01)	Partial mediation
Online news media exposure→Informational support→Professional commitment	0.039 (<.001)	0.059 (.03)	Partial mediation
Online news media exposure→Emotional support→Professional commitment	0.035 (.007)	0.059 (.03)	Partial mediation
Social media exposure			
Social media exposure→Informational support→Belonging	0.026 (.002)	0.060 (.06)	Full mediation
Social media exposure→Emotional support→Belonging	0.076 (<.001)	0.060 (.06)	Full mediation
Social media exposure→Informational support→Professional commitment	0.029 (<.001)	0.045 (.13)	Full mediation
Social media exposure→Emotional support→Professional commitment	0.092 (<.001)	0.045 (.13)	Full mediation

Discussion

Principal Findings

In light of the far-reaching crisis brought by the pandemic, this study aims to uncover how media factors can help mitigate social risks within the field of medical sociology. Utilizing the SOR framework, we propose a research model to examine the impacts of media exposure on medical students' professional identity. Our findings reveal several intriguing insights.

First and foremost, our findings indicate that both online news media exposure and social media exposure are positively related to perceived emotional and informational support. This suggests that medical students can derive both types of support from these media sources. Notably, the path coefficient between online news media and informational support was relatively higher than that for emotional support, whereas the opposite was true for social media exposure. The results can be explained by the inherent characteristics of the 2 types of media. Online news media are generally perceived as trustworthy due to their strict gatekeeping mechanisms [2], making them vital and reliable sources for informational support. By contrast, social media serves as an important platform for facilitating online interactions [46], thereby proving effective in providing emotional support.

Second, emotional support and informational support are positively associated with belonging and professional commitment, indicating that social support is a significant predictor of medical students' professional identity. Notably, the path coefficients for emotional support on both belonging and professional commitment are higher than those for informational support. The results suggest that emotional support may have a greater impact on fostering medical students' professional identity than informational support. This finding aligns with Wang et al [61], who emphasized the importance of emotional support in enhancing interpersonal relationships within online health support groups. Additionally, this can be explained by the nature of professional identity itself; both belonging and professional commitment relate to individuals' feelings toward their professional group and the profession [37,38], making them more likely to be driven by emotional factors.

Third, the impact of media exposure on professional identity was mediated by social support. Specifically, social support partially mediated the effect of online news media exposure and fully mediated the effect of social media exposure. These results indicate that social media primarily relies on the provision of social support to influence professional identity, whereas online news media exert their effects on professional identity both directly and indirectly through social support. One possible explanation for this finding lies in the characteristics of different media. During the pandemic, individuals primarily relied on social media to transmit and obtain social support. By contrast, by reporting news about medical workers, online news media can directly assist medical students in developing their understanding of professional identity [19]. These findings further clarify that social media and online news media might play different roles in shaping medical students' professional identity.

Fourth, the analysis of control variables shows that internship experience is positively associated with medical students' belonging and professional commitment. For medical students, the medical profession places great emphasis on practicality, making internships an essential avenue for gaining knowledge about the field. On the one hand, internships help students develop their self-efficacy by providing hands-on experience and exposure to real-world medical environments [85]; on the other hand, practical experience can enhance medical students' awareness of the profession [86]. Consequently, the training received during internships may better prepare students to become medical workers, which, in turn, inspires them to cultivate a sense of belonging and professional commitment to the field.

Implications

The internet has emerged as the predominant form of media globally, with social media serving as a crucial platform during the COVID-19 pandemic [3]. Given the media environment's

XSL•FO
RenderX

significant influence on shaping the cognition and behaviors of the younger generation [87], as well as the importance of medical students' professional identity for the stability of the medical community, this study explores the impact of media exposure on medical students' professional identity and seeks to uncover the mechanisms underlying this relationship. By comparing with prior work, this study offers several theoretical implications. First, it enriches the existing literature on professional identity in medical education. Previous studies typically focused on individual and educational environmental factors in the development of professional identity [12-14], often neglecting the role of the media environment. By integrating media exposure research with a social psychology perspective, this study empirically validates a new set of relationships and illustrates how media exposure influences medical students' perceived social support and professional identity. Unlike prior work, this study examines 2 types of media exposure simultaneously, developing a nuanced understanding of the effects of online news media and social media on medical education. The findings emphasize the significance of media in fostering medical students' professional identity, offering a new perspective for understanding the development of professional identity.

Second, this study extends the application of the SOR framework to explore the psychological processes underlying the development of professional identity, complementing the literature that examines the direct impacts of social factors on professional identity. It demonstrates that media exposure influences professional identity through both informational support and emotional support. Previous research has tended to focus on either cognitive or affective aspects of the organism component in the SOR framework [39]. This study further differentiates these 2 aspects by discussing the distinct impacts of informational support and emotional support on professional identity. Additionally, by illustrating how social support can be generated from different types of media exposure, this study contributes to the literature on the development of social support within various media environments.

This study also has practical implications for media platforms and medical educators. Above all, the findings confirm the importance of the media environment in fostering professional identity. These results are particularly relevant in the context of the Chinese media landscape. Since the outbreak of the COVID-19 pandemic, a substantial amount of information sharing praise and rewards for medical workers has circulated on Chinese media platforms. Exposure to such information may help students recognize the public's appreciation for their profession, thereby enhancing their perception of social support as well as their professional identity. Consequently, the proper use and management of media during public health crises can facilitate the development of medical students' professional identity. Considering the differences between online news media and social media, management departments and medical educators can leverage online news media to post reliable medical information, providing informational support, while encouraging favorable social interactions through social media to offer emotional support.

Additionally, our findings emphasize the importance of social support in enhancing professional identity. Sufficient social support should be provided to improve students' well-being during their learning experiences [88]. By alleviating mental health issues such as depression, stress, and anxiety, students' professional identity can be further promoted [89]. Furthermore, it is beneficial to present positive images of medical workers in online media so that medical students can develop a positive self-concept about the medical profession [77]. When reporting on medical workers, journalists should emphasize the value of medical professionalism rather than simply portraying them as angels or heroes [90,91]. In addition to focusing on providing social support through online media, medical educators and management departments can implement other measures to support students in coping with the challenges they encounter and improving their well-being, thereby reinforcing their professional identity.

Moreover, the significant effects of internships on belonging and professional commitment offer valuable implications for medical educators. Specifically, medical educators should provide opportunities for medical students to engage in practical learning. However, given that many medical students often have limited experience interacting with patients, leveraging media resources to provide professional knowledge can serve as an important complement to their internships. The practical experience gained from internships and the knowledge acquired through online media can be combined to motivate medical students to feel a sense of belonging and commitment to their profession.

Limitations and Future Research

There are several limitations that warrant future research. First, our study is designed based on the media environment in China. While the data illustrate the impacts of media exposure on the professional identity of Chinese medical students, future research could recruit respondents from other countries to enhance the generalizability of the findings. Second, this study focuses on online news media and social media without considering other media forms, such as short video platforms. Future research can consider the effects of other types of media. Third, this study collected self-reported data through a cross-sectional survey. While the relationships between variables can be tested, this approach limits the ability to validate causal relationships. Future studies could adopt longitudinal designs to collect data and examine the effects of media exposure. In addition, self-reported data may be less reliable in predicting actual behavior. Future research should consider using objective measures to assess professional identity. Fourth, while this study focuses on media exposure to uncover its impacts on the development of medical students' professional identity, it is important to acknowledge that professional identity can be influenced by various factors. Future studies should explore additional factors in different contexts, particularly those that may hinder the development of professional identity.

Conclusions

Whether faced with a plague or war, health workers have always been on the front lines of combating social risks. Their professional identity serves as both a foundation for their beliefs

XSL•FO
RenderX

and a source of strength. As members of the young generation in the age of social media, medical students' cognitions and behaviors are easily shaped by the media environment [87]. It is thus urgent to conduct in-depth research on the impacts of media exposure on medical students' professional identity, especially as the global wave of medical staff resignations intensifies. This study extends the SOR framework to examine the effects of media exposure related to long COVID on medical students' professional identity, aiming to provide insights into the critical issue of medical workforce shortages. Based on a survey of 1087 medical students, this study reveals that exposure to online news media and social media can enhance medical students' sense of belonging and professional commitment through emotional and informational support. It clarifies the role of media affordances in medical education, underscores the importance of media management and social support in shaping professional identity, and expands the understanding of the media's role in public health events. In the new media environment, these factors may become crucial components and governance tools in medical sociology, influencing health outcomes and serving as a viable approach to achieving the One Health goal in the long COVID era.

Acknowledgments

The authors acknowledge the financial support from the 2020 General Project of MOE (Ministry of Education) Foundation on Humanities and Social Sciences of China (20YJAZH117). The authors thank the JMIR editors and review team for constructive feedback and suggestions. No generative artificial intelligence was used in any portion of the manuscript writing.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

MW contributed to conceptualization, methodology, writing (original draft), review, editing, and project administration. JY was responsible for conceptualization, providing resources, writing (original draft), review, editing, and project administration. CQ participated in methodology, data analysis, writing (original draft), review, and editing. CY contributed to the writing (original draft).

Conflicts of Interest

None declared.

Multimedia Appendix 1
CHERRIES (Checklist for Reporting Results of Internet-Based e-Surveys) checklist.
[DOCX File , 18 KB - jmir_v26i1e50057_app1.docx]

Multimedia Appendix 2
Data analysis results of the pilot test.
[DOC File , 63 KB - jmir_v26i1e50057_app2.doc]

References

1. Ren W, Zhu X, Hu Y. Risk perception and preventive behaviors: a comparison of the multifaceted effects of social media and authoritative media during the outbreak of COVID-19. Chin J Journal Commun 2021 May 15;43(05):23-42. [doi: 10.13495/j.cnki.cjjc.2021.05.002]
2. Du Z, Luo X, Su L. Collective action under social expectation: the cognitive construction of media exposure on COVID-19 vaccination intention. Chin J Journal Commun 2022 Apr 15;44(4):53-77. [doi: 10.13495/j.cnki.cjjc.2022.04.008]
3. World Health Organization (WHO). Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation. WHO. URL: https://tinyurl.com/bdfcvuvx [accessed 2024-03-30]
4. World Health Organization (WHO). Global strategy on human resources for health: workforce 2030: reporting at Seventy-fifth World Health Assembly. WHO. URL: https://www.who.int/news/item/02-06-2022-global-strategy-on-human-resources-for-health--workforce-2030 [accessed 2024-03-30]
5. Passemard S, Faye A, Dubertret C, Peyre H, Vorms C, Boimare V, et al. Covid-19 crisis impact on the next generation of physicians: a survey of 800 medical students. BMC Med Educ 2021 Oct 13;21(1):529 [FREE Full text] [doi: 10.1186/s12909-021-02955-7] [Medline: 34645453]
6. Tempski P, Arantes-Costa FM, Kobayasi R, Siqueira MAM, Torsani MB, Amaro BQRC, et al. Medical students' perceptions and motivations during the COVID-19 pandemic. PLoS One 2021;16(3):e0248627 [FREE Full text] [doi: 10.1371/journal.pone.0248627] [Medline: 33730091]

XSL•FO
RenderX

7. Yan J, Wu M, Liao Y, Huang Y. Modelling the factors that affect medical students' occupational identity in long COVID: an integrated perspective of motivation, opportunity and ability. Humanit Soc Sci Commun 2024 Feb 12;11(1):1-9. [doi: 10.1057/s41599-024-02755-6]
8. Johnson M, Cowin LS, Wilson I, Young H. Professional identity and nursing: contemporary theoretical developments and future research challenges. Int Nurs Rev 2012 Dec;59(4):562-569. [doi: 10.1111/j.1466-7657.2012.01013.x] [Medline: 23134142]
9. Erikson H. Identity, Youth, and Crisis. New York, NY: Norton; 1968.
10. Shu M, Huang S, Chen J. Impact of medical practice environmental assessment on professional identity of clinical medical students. Med Soc 2019 Jun 10;32(6):131-134. [doi: 10.13723/j.yxysh.2019.06.035]
11. Sun L, Liang W, Mei Y, Lin L, Yu Y, Xia O. Investigation on the professional identity of master degree postgraduate of clinical medicine. Chin J Med Educ 2015 Jun 10;35(5):764-766. [doi: 10.3760/cma.j.issn.1673-677X.2015.05.035]
12. Helmich E, Derksen E, Prevoo M, Laan R, Bolhuis S, Koopmans R. Medical students' professional identity development in an early nursing attachment. Med Educ 2010 Jun 15;44(7):674-682. [doi: 10.1111/j.1365-2923.2010.03710.x]
13. Cavenagh P, Dewberry C, Jones P. Becoming professional: when and how does it start? A comparative study of first-year medical and law students in the UK. Med Educ 2000 Nov;34(11):897-902. [doi: 10.1046/j.1365-2923.2000.00680.x] [Medline: 11107013]
14. Duran S, Celik I, Ertugrul B, Ok S, Albayrak S. Factors affecting nurses' professional commitment during the COVID-19 pandemic: a cross-sectional study. J Nurs Manag 2021 Oct;29(7):1906-1915 [FREE Full text] [doi: 10.1111/jonm.13327] [Medline: 33794061]
15. Ruan B, Yilmaz Y, Lu D, Lee M, Chan TM. Defining the digital self: a qualitative study to explore the digital component of professional identity in the health professions. J Med Internet Res 2020 Sep 29;22(9):e21416 [FREE Full text] [doi: 10.2196/21416] [Medline: 32990636]
16. De Coninck D, d'Haenens L, Matthijs K. Forgotten key players in public health: news media as agents of information and persuasion during the COVID-19 pandemic. Public Health 2020 Jun;183:65-66 [FREE Full text] [doi: 10.1016/j.puhe.2020.05.011]
17. Perez‐Lugo M. Media uses in disaster situations: a new focus on the impact phase. Sociological Inquiry 2004 Apr;74(2):210-225. [doi: 10.1111/j.1475-682x.2004.00087.x]
18. Krawczyk K, Chelkowski T, Laydon DJ, Mishra S, Xifara D, Gibert B, et al. Quantifying online news media coverage of the COVID-19 pandemic: text mining study and resource. J Med Internet Res 2021 Jun 02;23(6):e28253 [FREE Full text] [doi: 10.2196/28253] [Medline: 33900934]
19. Bagnasco A, Catania G, Gallagher A, Morley G. Media representations of nurses in the pandemic: just doing our job? Nurs Ethics 2020 Jun;27(4):901-905. [doi: 10.1177/0969733020926352] [Medline: 32495717]
20. Lee J, Kim K, Park G, Cha N. The role of online news and social media in preventive action in times of infodemic from a social capital perspective: the case of the COVID-19 pandemic in South Korea. Telemat Inform 2021 Nov;64:101691 [FREE Full text] [doi: 10.1016/j.tele.2021.101691] [Medline: 36567818]
21. Bambina A. Online Social Support: The Interplay of Social Networks and Computer-Mediated Communication. Amherst, NY: Cambria Press; 2007.
22. Porter CE, Donthu N. Cultivating trust and harvesting value in virtual communities. Management Science 2008 Jan;54(1):113-128. [doi: 10.1287/mnsc.1070.0765]
23. Gazzola N, De Stefano J, Audet C, Theriault A. Professional identity among counselling psychology doctoral students: a qualitative investigation. Counselling Psychology Quarterly 2011 Dec;24(4):257-275. [doi: 10.1080/09515070.2011.630572]
24. Mehrabian A, Russell JA. An Approach to Environmental Psychology. Boston, MA: MIT Press; 1974.
25. Browne C, Wall P, Batt S, Bennett R. Understanding perceptions of nursing professional identity in students entering an Australian undergraduate nursing degree. Nurse Educ Pract 2018 Sep;32:90-96. [doi: 10.1016/j.nepr.2018.07.006] [Medline: 30098517]
26. Loi R, Hang-yue N, Foley S. The effect of professional identification on job attitudes: a study of lawyers in Hong Kong. Organ Anal 2004 Feb 1;12(2):109-128. [doi: 10.1108/eb028988]
27. Zhang W, Meng H, Yang S, Liu D. The influence of professional identity, job satisfaction, and work engagement on turnover intention among township health inspectors in China. Int J Environ Res Public Health 2018 May 14;15(5):988 [FREE Full text] [doi: 10.3390/ijerph15050988] [Medline: 29757985]
28. de Lasson L, Just E, Stegeager N, Malling B. Professional identity formation in the transition from medical school to working life: a qualitative study of group-coaching courses for junior doctors. BMC Med Educ 2016 Jun 24;16(1):165 [FREE Full text] [doi: 10.1186/s12909-016-0684-3] [Medline: 27342973]
29. Beijaard D, Meijer PC, Verloop N. Reconsidering research on teachers' professional identity. Teaching and Teacher Education 2004 Feb;20(2):107-128. [doi: 10.1016/j.tate.2003.07.001]
30. Brooks NG, Riemenschneider CK, Hardgrave BC, O'Leary-Kelly AM. IT professional identity: needs, perceptions, and belonging. European Journal of Information Systems 2017 Dec 19;20(1):87-102. [doi: 10.1057/ejis.2010.48]
31. Meijers F. The development of a career identity. Int J Adv Couns Spe 1998;20:191-207. [doi: 10.1023/A:1005399417256]

XSL•FO
RenderX

32. Ashforth BE, Schinoff BS. Identity under construction: how individuals come to define themselves in organizations. Annu Rev Organ Psychol Organ Behav 2016 Mar 21;3(1):111-137. [doi: 10.1146/annurev-orgpsych-041015-062322]
33. Patel SE, Chrisman M, Russell CL, Lasiter S, Bennett K, Pahls M. Cross-sectional study of the relationship between experiences of incivility from staff nurses and undergraduate nursing students' sense of belonging to the nursing profession. Nurse Educ Pract 2022 Jul;62:103320. [doi: 10.1016/j.nepr.2022.103320] [Medline: 35500414]
34. Aranya N, Ferris K. A reexamination of accountants' organizational-professional conflict. Account Rev Jan 1984;59(1):1-15 [FREE Full text]
35. Lachman R, Aranya N. Job attitudes and turnover intentions among professionals in different work settings. Organization Studies 1986 Jul 01;7(3):279-293. [doi: 10.1177/017084068600700305]
36. Meyer JP, Allen NJ, Smith CA. Commitment to organizations and occupations: extension and test of a three-component conceptualization. Journal of Applied Psychology 1993;78(4):538-551. [doi: 10.1037/0021-9010.78.4.538]
37. Hagerty BM, Lynch-Sauer J, Patusky KL, Bouwsema M, Collier P. Sense of belonging: a vital mental health concept. Archives of Psychiatric Nursing 1992 Jun;6(3):172-177. [doi: 10.1016/0883-9417(92)90028-H]
38. Meyer JP, Herscovitch L. Commitment in the workplace: toward a general model. Human Resource Management Review 2001 Sep;11(3):299-326. [doi: 10.1016/s1053-4822(00)00053-x]
39. Eroglu SA, Machleit KA, Davis LM. Atmospheric qualities of online retailing. Journal of Business Research 2001 Nov;54(2):177-184. [doi: 10.1016/s0148-2963(99)00087-9]
40. Luo P, Wang C, Guo F, Luo L. Factors affecting individual online rumor sharing behavior in the COVID-19 pandemic. Comput Human Behav 2021 Dec;125:106968 [FREE Full text] [doi: 10.1016/j.chb.2021.106968] [Medline: 34334932]
41. Hewei T, Youngsook L. Factors affecting continuous purchase intention of fashion products on social E-commerce: SOR model and the mediating effect. Entertainment Computing 2022 Mar;41:100474. [doi: 10.1016/j.entcom.2021.100474]
42. Fu S, Li H, Liu Y, Pirkkalainen H, Salo M. Social media overload, exhaustion, and use discontinuance: examining the effects of information overload, system feature overload, and social overload. Information Processing & Management 2020 Nov;57(6):102307. [doi: 10.1016/j.ipm.2020.102307]
43. Zhang H, Lu Y, Gupta S, Zhao L. What motivates customers to participate in social commerce? The impact of technological environments and virtual customer experiences. Information & Management 2014 Dec;51(8):1017-1030. [doi: 10.1016/j.im.2014.07.005]
44. Pandita S, Mishra HG, Chib S. Psychological impact of covid-19 crises on students through the lens of Stimulus-Organism-Response (SOR) model. Child Youth Serv Rev 2021 Jan;120:105783 [FREE Full text] [doi: 10.1016/j.childyouth.2020.105783] [Medline: 33518862]
45. Plantin L, Daneback K. Parenthood, information and support on the internet. A literature review of research on parents and professionals online. BMC Fam Pract 2009 May 18;10:34 [FREE Full text] [doi: 10.1186/1471-2296-10-34] [Medline: 19450251]
46. Selkie E, Adkins V, Masters E, Bajpai A, Shumer D. Transgender adolescents' uses of social media for social support. J Adolesc Health 2020 Mar;66(3):275-280. [doi: 10.1016/j.jadohealth.2019.08.011] [Medline: 31690534]
47. Drew D, Weaver D. Media attention, media exposure, and media effects. Journalism Quarterly 1990 Dec 01;67(4):740-748. [doi: 10.1177/107769909006700428]
48. Gil de Zúñiga H, Molyneux L, Zheng P. Social media, political expression, and political participation: panel analysis of lagged and concurrent relationships. J Commun 2014 Jun 19;64(4):612-634. [doi: 10.1111/jcom.12103]
49. Brown J, L'Engle K, Pardun C, Guo G, Kenneavy K, Jackson C. Sexy media matter: exposure to sexual content in music, movies, television, and magazines predicts black and white adolescents' sexual behavior. Pediatrics 2006 Apr;117(4):1018-1027. [doi: 10.1542/peds.2005-1406] [Medline: 16585295]
50. Anderson P, de Bruijn A, Angus K, Gordon R, Hastings G. Impact of alcohol advertising and media exposure on adolescent alcohol use: a systematic review of longitudinal studies. Alcohol Alcohol 2009;44(3):229-243. [doi: 10.1093/alcalc/agn115] [Medline: 19144976]
51. Chao M, Xue D, Liu T, Yang H, Hall BJ. Media use and acute psychological outcomes during COVID-19 outbreak in China. J Anxiety Disord 2020 Aug;74:102248 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102248] [Medline: 32505918]
52. Kim A, Dennis AR. Says who? The effects of presentation format and source rating on fake news in social media. MISQ 2019 Jan 1;43(3):1025-1039. [doi: 10.25300/misq/2019/15188]
53. Rosen AO, Holmes AL, Balluerka N, Hidalgo MD, Gorostiaga A, Gómez-Benito J, et al. Is social media a new type of social support? Social media use in Spain during the COVID-19 pandemic: a mixed methods study. Int J Environ Res Public Health 2022 Mar 26;19(7):3952 [FREE Full text] [doi: 10.3390/ijerph19073952] [Medline: 35409634]
54. Langford CPH, Bowsher J, Maloney JP, Lillis PP. Social support: a conceptual analysis. J Adv Nurs 1997 Jan;25(1):95-100. [doi: 10.1046/j.1365-2648.1997.1997025095.x] [Medline: 9004016]
55. Buschmann M, Hollinger L. Influence of social support and control on depression in the elderly. Clinical Gerontologist 1994 Jun 03;14(4):13-28. [doi: 10.1300/j018v14n04_03]
56. Oppong Asante K. Social support and the psychological wellbeing of people living with HIV/AIDS in Ghana. Afr J Psychiatry (Johannesbg) 2012 Sep;15(5):340-345. [doi: 10.4314/ajpsy.v15i5.42] [Medline: 23044888]

XSL•FO
RenderX

57. Castellá Sarriera J, Bedin L, Abs D, Casas F, Calza T. Relación entre soporte social, la satisfacción de vida y bienestar subjetivo en adolescentes brasileros. Univ Psychol 2015 Sep 07;14(2):459-474. [doi: 10.11144/javeriana.upsy14-2.rbss]
58. Lu W, Hampton KN. Beyond the power of networks: differentiating network structure from social media affordances for perceived social support. New Media & Society 2016 Jan 04;19(6):861-879. [doi: 10.1177/1461444815621514]
59. Nick EA, Cole DA, Cho S, Smith DK, Carter TG, Zelkowitz RL. The online social support scale: measure development and validation. Psychol Assess 2018 Sep;30(9):1127-1143 [FREE Full text] [doi: 10.1037/pas0000558] [Medline: 29781664]
60. Ruiller C, Van Der Heijden BI. Socio-emotional support in French hospitals: effects on French nurses' and nurse aides' affective commitment. Appl Nurs Res 2016 Feb;29:229-236. [doi: 10.1016/j.apnr.2015.06.006] [Medline: 26856519]
61. Wang Y, Kraut R, Levine J. To stay or leave? The relationship of emotionalinformational support to commitment in online health support groups. 2012 Feb 11 Presented at: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work; February 11-15, 2012; Seattle, WA p. 833-842 URL: https://dl.acm.org/doi/abs/10.1145/2145204.2145329 [doi: 10.1145/2145204.2145329]
62. Cronenwett LR. Parental network structure and perceived support after birth of first child. Nursing Research 1985;34(6):347-351. [doi: 10.1097/00006199-198511000-00007]
63. Sutherland S, Jalali A. Social media as an open-learning resource in medical education: current perspectives. AMEP 2017 Jun;Volume 8:369-375. [doi: 10.2147/amep.s112594]
64. Oh S, Lee SY, Han C. The effects of social media use on preventive behaviors during infectious disease outbreaks: the mediating role of self-relevant emotions and public risk perception. Health Commun 2021 Jul 16;36(8):972-981. [doi: 10.1080/10410236.2020.1724639] [Medline: 32064932]
65. Bellanti F, Lo Buglio A, Capuano E, Dobrakowski M, Kasperczyk A, Kasperczyk S, et al. Factors related to nurses' burnout during the first wave of coronavirus disease-19 in a university hospital in Italy. IJERPH 2021 May 11;18(10):5051. [doi: 10.3390/ijerph18105051] [Medline: 34064610]
66. Chang H, Shyu YL, Wong M, Chu T, Lo Y, Teng C. How does burnout impact the three components of nursing professional commitment? Scand J Caring Sci 2017 Dec 24;31(4):1003-1011. [doi: 10.1111/scs.12425] [Medline: 28439914]
67. Ahmad Z, Anantharaman RN, Ismail H. Students' motivation, perceived environment and professional commitment: an application of Astin's college impact model. Accounting Education 2012 Apr;21(2):187-208. [doi: 10.1080/09639284.2011.603472]
68. Kjærgaard A, Morsing M, Ravasi D. Mediating identity: a study of media influence on organizational identity construction in a celebrity firm. J Manage Stud 2011 Mar;48(3):514-543. [doi: 10.1111/j.1467-6486.2010.00954.x]
69. Geusens F, Beullens K. I see, therefore I am: exposure to alcohol references on social media, but not on traditional media, is related to alcohol consumption via drinking and non-drinking identity. Health Commun 2023 Feb 16;38(2):402-410. [doi: 10.1080/10410236.2021.1954301] [Medline: 34266317]
70. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
71. Burns AJ, Roberts TL, Posey C, Lowry PB, Fuller B. Going beyond deterrence: a middle-range theory of motives and controls for insider computer abuse. Information Systems Research 2023 Mar;34(1):342-362. [doi: 10.1287/isre.2022.1133]
72. Fornell C, Larcker DF. Structural equation models with unobservable variables and measurement error: algebra and statistics. Journal of Marketing Research 1981 Aug 01;18(3):382-388. [doi: 10.1177/002224378101800313]
73. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis: A Global Perspective (7th Edition). Upper Saddle River, NJ: Pearson Prentice Hall; 2010.
74. Harman HH. Modern Factor Analysis. Chicago, IL: University of Chicago Press; 1976.
75. Bagozzi RP, Yi Y, Phillips LW. Assessing construct validity in organizational research. Administrative Science Quarterly 1991 Sep;36(3):421. [doi: 10.2307/2393203]
76. Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S, et al. Mental health problems and social media exposure during COVID-19 outbreak. PLoS One 2020 Apr 16;15(4):e0231924 [FREE Full text] [doi: 10.1371/journal.pone.0231924] [Medline: 32298385]
77. Administration Center of China E-Government Network, Online Communication Research Institute of Nanjing University. A research on public awareness and information dissemination of novel coronavirus pneumonia was officially released. Administration Center of China E-Government Network. URL: http://www.sic.gov.cn/sic/83/79/0226/10410_pc.html [accessed 2022-04-22]
78. Heidari E, Salimi G, Mehrvarz M. The influence of online social networks and online social capital on constructing a new graduate students' professional identity. Interactive Learning Environments 2020 May 27;31(1):214-231. [doi: 10.1080/10494820.2020.1769682]
79. Zhang Z, Fu W, Tian C, Zhang F, Zhao B, Mao J, et al. Professional identity of Chinese nursing students during the COVID-19 pandemic outbreak: a nation-wide cross-sectional study. Nurse Education in Practice 2021 Mar;52:103040. [doi: 10.1016/j.nepr.2021.103040]
80. Lin H, Fan W, Chau PY. Determinants of users' continuance of social networking sites: a self-regulation perspective. Information & Management 2014 Jul;51(5):595-603. [doi: 10.1016/j.im.2014.03.010]

XSL•FO
RenderX

81. Urbach N, Ahlemann F. Structural equation modeling in information systems research using partial least squares. J Inform Technology Theory Appl 2010 Jun;11(2):5-40 [FREE Full text]
82. Anderson JC, Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach. Psychological Bulletin 1988 May;103(3):411-423. [doi: 10.1037/0033-2909.103.3.411]
83. Hair JF, Hult GTM, Ringle CM, Sarstedt MA. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM) (3rd Edition). Thousand Oaks, CA: Sage; 2021.
84. Zhao X, Lynch JG, Chen Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. J Consum Res 2010 Aug 01;37(2):197-206. [doi: 10.1086/651257]
85. Lang JC, Lee CH. Identity accumulation, others' acceptance, job-search self-efficacy, and stress. J Organiz Behav 2005 May;26(3):293-312. [doi: 10.1002/job.309]
86. Filstad C, Traavik LE, Gorli M. Belonging at work: the experiences, representations and meanings of belonging. JWL 2019 Apr 03;31(2):116-142. [doi: 10.1108/jwl-06-2018-0081]
87. Purba AK, Henery PM, Thomson RM, Pearce A, Henderson M, Katikireddi SV. Social media use and adolescent engagement in health risk behaviours: a systematic review and meta-analysis. The Lancet 2021 Nov;398:S93. [doi: 10.1016/s0140-6736(21)02636-2]
88. Guo K, Zhang X, Bai S, Minhat HS, Nazan AINM, Feng J, et al. Assessing social support impact on depression, anxiety, and stress among undergraduate students in Shaanxi province during the COVID-19 pandemic of China. PLoS One 2021 Jul 23;16(7):e0253891 [FREE Full text] [doi: 10.1371/journal.pone.0253891] [Medline: 34297731]
89. Hao C, Zhu L, Zhang S, Rong S, Zhang Y, Ye J, et al. Serial multiple mediation of professional identity, and psychological capital in the relationship between work-related stress and work-related well-being of ICU nurses in China: a cross-sectional questionnaire survey. Front Psychol 2020 Dec 22;11:535634 [FREE Full text] [doi: 10.3389/fpsyg.2020.535634] [Medline: 33414737]
90. Halberg N, Jensen PS, Larsen TS. We are not heroes-The flipside of the hero narrative amidst the COVID19-pandemic: a Danish hospital ethnography. J Adv Nurs 2021 May 22;77(5):2429-2436 [FREE Full text] [doi: 10.1111/jan.14811] [Medline: 33616210]
91. Stokes-Parish J, Elliott R, Rolls K, Massey D. Angels and heroes: the unintended consequence of the hero narrative. J Nurs Scholarsh 2020 Sep 27;52(5):462-466 [FREE Full text] [doi: 10.1111/jnu.12591] [Medline: 32856396]

Abbreviations

AVE: average variance extracted
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
CMB: common method bias
SOR: Stimulus-Organism-Response

Edited by T de Azevedo Cardoso; submitted 18.06.23; peer-reviewed by J Mira, A AL-Asadi; comments to author 06.02.24; revised version received 02.04.24; accepted 04.09.24; published 17.10.24.

Please cite as:
Wu M, Yan J, Qiao C, Yan C
Impact of Concurrent Media Exposure on Professional Identity: Cross-Sectional Study of 1087 Medical Students During Long COVID
J Med Internet Res 2024;26:e50057
URL: https://www.jmir.org/2024/1/e50057
doi:10.2196/50057
PMID:39418080

©Manli Wu, Jun Yan, Chongming Qiao, Chu Yan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association Between School-Related Google Trends Search Volume and Suicides Among Children and Adolescents in Japan During 2016-2020: Retrospective Observational Study With a Time-Series Analysis

Takahiro Arai[1,2,3], MSHM; Hiroe Tsubaki[4], PhD; Ayako Wakano[5], PhD; Yasuyuki Shimizu[3], BA

[1]School of Management and Information Sciences, Tama University, Tokyo, Japan
[2]Graduate School of Health Management, Keio University, Kanagawa, Japan
[3]Japan Suicide Countermeasures Promotion Center, Tokyo, Japan
[4]The Institute of Statistical Mathematics, Tokyo, Japan
[5]Department of Economics, School of Political Science and Economics, Tokai University, Kanagawa, Japan

Corresponding Author:
Takahiro Arai, MSHM
School of Management and Information Sciences
Tama University
4-1-1 Hijirigaoka
Tama City
Tokyo, 206-0022
Japan
Phone: 81 42 337 7149
Email: arai.t@keio.jp

Abstract

Background: Suicide is the leading cause of death among children and adolescents in Japan. Internet search volume may be useful in detecting suicide risk. However, few studies have shown an association between suicides attempted by children and adolescents and their internet search volume.

Objective: This study aimed to examine the relationship between the number of suicides and the volume of school-related internet searches to identify the search terms that could serve as the leading indicators of suicide prevention among children and adolescents.

Methods: We used data on weekly suicides attempted by elementary, middle, and high school students in Japan from 2016 to 2020, provided by the National Police Agency. Internet search volume was weekly data for 20 school-related terms obtained from Google Trends. Granger causality and cross-correlation analysis were performed to estimate the temporal back-and-forth and lag between suicide deaths and search volume for the related terms.

Results: The search queries “I do not want to go to school” and “study” showed Granger causality with suicide incidences. The cross-correlation analysis showed significant positive correlations in the range of –2 to 2 for “I do not want to go to school” (highest value at time lag 0, r=0.28), and –1 to 2 for “study” (highest value at time lag –1, r=0.18), indicating that the search volume increased as the number of suicides increased. Furthermore, during the COVID-19 pandemic period (January-December 2020), the search trend for “I do not want to go to school,” unlike “study,” was highly associated with suicide frequency.

Conclusions: Monitoring the volume of internet searches for “I do not want to go to school” could be useful for the early detection of suicide risk among children and adolescents and for optimizing web-based helpline displays.

(J Med Internet Res 2024;26:e51710) doi:10.2196/51710

KEYWORDS

adolescent; children; COVID-19; Google Trends; internet; Japan; monitoring; suicide; surveillance; time series analysis

XSL•FO
RenderX

Introduction

In Japan, suicide is the leading cause of death among people aged 10-19 years [1]. Suicide among children and adolescents has continued to increase since 2016, with data for 2020 showing the suicide rate (per 100,000 population) for primary, secondary, and high school students totaled 4.0, with 0.2 for primary school students, 4.5 for secondary school students, and 11.0 for high school students. One of the key reasons for suicide is school-related problems [2]. Therefore, suicide prevention among children and adolescents is primarily conducted in schools. However, school-based suicide prevention programs have exhibited limited effectiveness in reducing suicide attempts and rates [3,4]. In recent years, suicide prevention using the internet has received greater attention to complement the effectiveness of school-based suicide prevention programs directed toward the young population. As the internet has become a prevailing medium for young people, suicide prevention has been examined using internet spaces rather than mental health care in schools and gatekeeping using the phone [5]. Recent studies have analyzed the association between the volume of suicide-related terms derived from internet search queries and suicide [6-11].

The field of public health has explored the development of monitoring methodologies using internet searches. During the COVID-19 pandemic, there have been suggestions that Google Trends could be used to aid disease surveillance systems [12-14]; their complementary function has also been explored in the context of suicide prevention [15,16].

The Japanese government focused on the relationship between internet search queries and suicide and proposed a suicide prevention strategy using internet search mechanisms in 2017 [17]. However, the empirical findings were quite limited. Sueki [18] reported a significant positive correlation between the number of suicides and the search term “utsu” (depression). Hagihara et al [19] further reported a strong association between suicide rates and the search term “hydrogen sulfide,” a means of attempting suicide. Both studies covered the period before the COVID-19 pandemic, and search terms were related to mental illness and crisis interventions, such as the means of attempting suicide. Furthermore, during the COVID-19 pandemic, Taira et al [20] found that keywords such as “no money” and “divorce” could predict the number of suicides, discovering a new linguistic marker by targeting search terms related to social factors rather than words directly related to specific disease names or suicide.

Internet search terms may provide clues for identifying suicide crisis pathways and be useful for optimizing web-based suicide prevention [21]. However, the research investigating an association between suicide deaths among children and adolescents and internet search queries is scant. Therefore, new linguistic markers that have the potential to detect suicide risk among children and adolescents need to be explored.

Our study aims to identify the search terms that could serve as the leading indicators of suicide among children and adolescents. For this purpose, we investigated the relationship between the number of suicides and internet search volumes. We further examined how these trends changed in the social context during the COVID-19 pandemic. In this study, children and adolescents are defined based on their school attendance at elementary, junior high, and high schools rather than by age. This definition, as used in the school education system, is considered a prominent factor contributing to the occurrence of suicide among children and adolescents in Japan [2].

Several analytical challenges regarding the usefulness of search queries in suicide prevention remain unresolved. In other words, the coarseness of the monthly data on the number of suicides is problematic. To tackle this weakness, we propose that using weekly data can change the analysis results [22]. Our study has 2 unique features. First, we used weekly data on the number of suicides and internet search volumes because monthly analysis may not observe changes in highly fluid social conditions. Second, we used the number of suicides attempted by elementary, middle, and high school students rather than the total number of suicides. The Japanese suicide surveillance system includes occupational status.

Methods

Suicide Data

The National Police Agency provided data on the number of suicides. These were weekly data processed by the Japan Suicide Countermeasures Promotion Center under the supervision of the Ministry of Health, Labor, and Welfare after receiving confirmation from the National Police Agency. The data collection period was 5 years, from 2016 to 2020, and the analysis included elementary, middle, and high school students. The variables present in the National Police Agency data are demographic variables, gender (binary), age (integer), occupational status (51 categories including elementary, middle, and high school students), living together (3 levels—yes, no, and unknown), municipality of residence (character), suicide municipality (character), and other or not having a history of suicide attempts (3 levels—yes, no, and unknown). Selecting participants solely based on age would not adequately exclude the individuals who dropped out or did not attend high school after graduating from their junior high school. Therefore, this study used occupational status categories to accurately identify and include elementary, junior high, and high school students.

Internet Search Data

Search queries used weekly data (7-day interval from Sunday to Monday) from January 3, 2016, to December 26, 2020, using Google Trends. The Google Trends search volumes during this period were normalized between 0 and 100. According to Google Trends, the normalization method is described as “search results are normalized to the time and location of a query by the following process: Each data point is divided by the total searches for the geography and time range. The resulting numbers are then scaled on a scale of 0 to 100 based on a topic’s proportion to all searches on all topics” [23]. Therefore, it only shows a relative value rather than a pure search frequency. Relative search values (RSV) are calculated as follows.

XSL•FO
RenderX

For the search volume of a word "*i*" in a certain week "*t*," in a certain region "*j*," let $Qmax_{i,j}$ be the maximum search volume when it was the maximum in all periods.

$$Qmax_{i,j}=max\{search_{i,j,t}\}$$

(1)

The amount of retrieval of the search word "*i*" in a given week "*t*" is relative to the maximum value.

(2)

The search terms used in this study were selected based on a list of the causes of student suicide [2]. We excluded "suicide" and "I want to die," which have been used in many studies, because they are likely to be searched by many people unrelated to suicide and are not necessarily representative of suicide risks among children and adolescents [24]. In addition, we integrated words related to school resumption because the school resumption timing is considered to cause strong psychological stress among young people [25].

Based on the above information, 20 keywords were selected for this study, such as "school," "school reopening," "I do not want to go to school," "school hard," "school anxiety," "When is school?" "When does school start?" "school start," "school bullying," "school friends," "school grades," "bullying," "friendships," "study," "anxiety about studying," "I do not know how to study," "mid-term or term-end exam," "entrance examination," "future career," and "grades."

Statistical Analysis

We used the time-series analysis approach used in previous studies [26]. Referring to previous studies, we constructed a concept map to vividly illustrate time-series analysis (Figure 1).

First, a unit root test (an augmented Dickey-Fuller [ADF] test) was performed on the collected time-series data (n=260 weeks) using the R package (R Core Team) "tseries" [27] to examine whether the sequence was stationary or not. Stationarity is an important assumption in many statistical analyses, including Granger causality analysis. The null hypothesis of the ADF test (H0) assumes the existence of a unit root and indicates nonstationarity, while the alternative hypothesis (H1) indicates stationarity. With the significance level set at .05, a *P* value below it justifies the rejection of H0 and affirms the stationarity of the data. Next, the vector autoregressive (VAR) model was estimated using the R package "vars" [28]. It incorporates the optimal lag variables, whose order was determined by Akaike's information criterion (AIC), and the Granger causality was verified. Granger causality analysis is a statistical method used to determine if one time-series variable can predict another time-series variable. To assess the significance of the relationship, we performed the *F* test. If the *P* value was below the significance level ($P<.05$), we rejected the null hypothesis and concluded that there is evidence of Granger causality. However, if the *P* value was above the significance level, we failed to reject the null hypothesis, indicating that there was no evidence of Granger causality. Furthermore, to adjust for the influence of the school calendar, we incorporated dummy variables for weeks including April 8, September 1, and January 8, which are anticipated to have spikes in suicide risk as they mark the end of spring, summer, and winter breaks, respectively [25]. We corrected the number of suicides using these dummy variables in our Granger causality testing model (the school calendar model). We tested the Granger causality test using this adjusted number of suicides.

The Google Trends model for the Granger causality test is as follows.

(3)

(4)

Y_t is the number of suicides at time *t*. X_t is the causal variable (the number of searches for the keyword on Google) at time .

Adjustments to the number of suicides in the school calendar model are as follows:

The Poisson regression model takes the following form:

$$log(\lambda_t)=\gamma_0+\gamma_1 D_{spring,t}+\gamma_2 D_{summer,t}+\gamma_3 D_{winter,t}$$

(5)

λ_t is the expected number of suicides at week *t*. $D_{spring,t}$, $D_{summer,t}$, and $D_{winter,t}$ are dummy variables at the end of the spring, summer, and winter breaks, respectively. The actual number of suicides data is assumed to follow a Poisson distribution. By subtracting the predictions made by the Poisson regression model from the measured data, we generated suicide count data adjusted for the effects of the school calendar.

After performing the Granger causality test, as other past studies did, we also calculated the cross-correlation coefficients based on the same method as in previous studies [22]. Correlation coefficients with the time lag data were estimated, and a significance test was performed. We assumed that the variance was equivariant when the correlation coefficient was obtained. In this case, the square root transformation was used for the variance stabilization transformation owing to the data frequency. Finally, search terms and suicide frequencies that showed significant differences in the number of suicides were plotted against the number of weeks in 2020 to determine social conditions during the COVID-19 pandemic. All statistical analyses were performed using R (version 4.1.0; The R Foundation), and the significance level was set to 5%.

XSL•FO
RenderX

Figure 1. Concept map expressing the methodology of time series analysis. VAR: vector autoregressive.

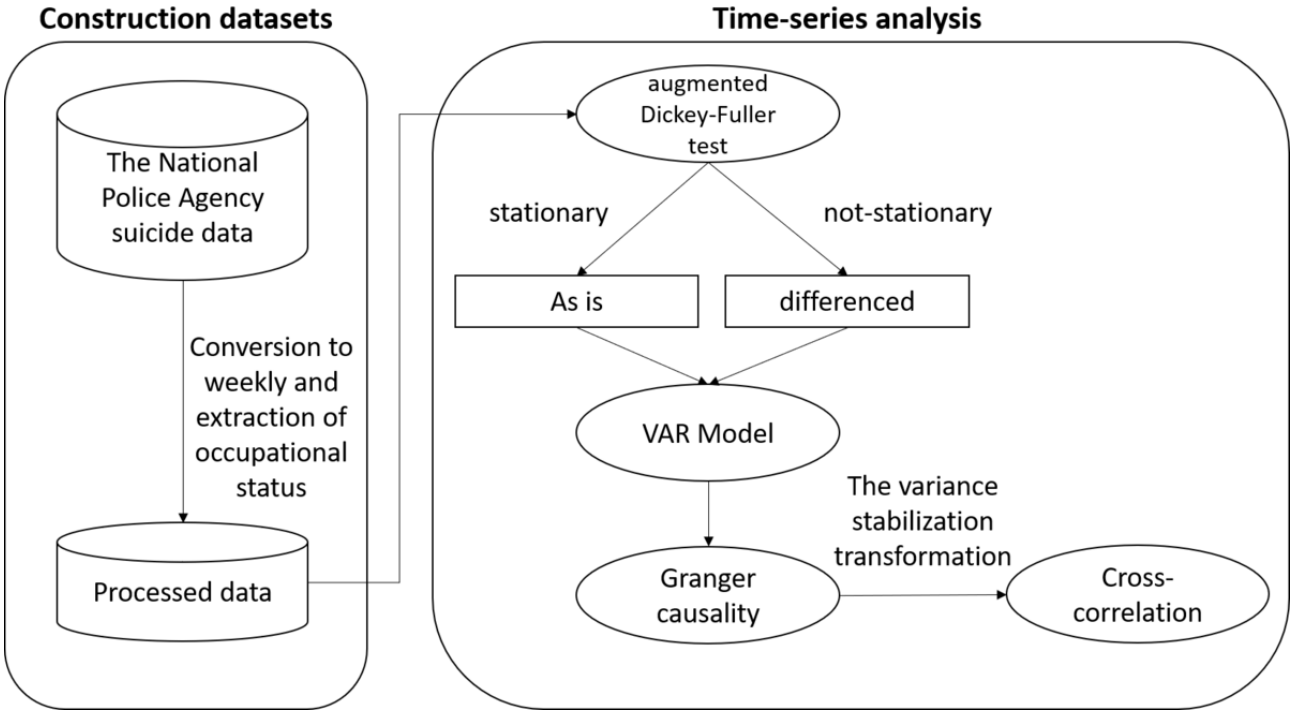

Ethical Considerations

Our study was approved by the Research Ethics Review Committee of Japan Suicide Countermeasures Promotion Center on September 16, 2021 (R03-01). Patient consent was waived because this study was a secondary analysis of existing data. The data were made available for research use by processing them on a weekly basis to prevent personal identification.

Results

Distribution of the Number of Suicides

The total number of suicides among elementary, middle, and high school students, from January 3, 2016, to December 26, 2020, was 1925 (mean 7.40, SD 3.42). The 52 weeks were divided into 5 periods. The summary statistics for each period were as follows: period 1 (January 3, 2016, to December 31, 2016) was 318 (mean 6.12, SD 2.94), period 2 (January 1, 2017, to December 30, 2017) was 352 (mean 6.77, SD 2.95), period 3 (December 31, 2017, to December 29, 2018) was 366 (mean 7.04, SD 2.81), period 4 (December 30, 2018, to December 28, 2019) was 394 (mean 7.58, SD 3.14), and period 5 (December 29, 2019, to December 26, 2019) was 495 (mean 9.52, SD 4.18). During our study period (260 total weeks), there were 0 weeks with no suicide deaths among elementary, middle, or high school students.

Evaluation Outcomes

The study variables were compiled and are presented in Table 1. They consist of 11 school-related terms and 9 terms integrated with school words, generating a total of 20 variables.

Unit root tests (ADF) were performed on the original series of suicide variables, all of which exhibited stationary behavior (Table 2). The test results are available upon request from the authors. Using the AIC, a VAR model was selected for Granger causality analysis. The results showed a statistically significant Granger causality between the search terms “I do not want to go to school” (P=.006) and “study” (P=.02) and the number of suicides (Table 3). The other search terms did not show Granger causality in either direction for the number of suicides.

XSL•FO
RenderX

Table 1. The Google Trends search terms selected as the subjects of analysis.

English equivalent search terms	Original Japanese search terms
Search terms combined with school	
school	学校 (gakkou)
school reopening	学校 再開 (gakkou saikai)
I do not want to go to school	学校 行きたくない (gakkou ikitakunai)
school hard	学校 辛い (gakkou tsurai)
school anxiety	学校 不安 (gakkou fuan)
When is school?	学校 いつ (gakkou itsu)
When does school start?	学校 いつから (gakkou itsukara)
school start	学校 始まる (gakkou hajimaru)
school bullying	学校 いじめ (gakkou ijime)
School-related search terms	
school friends	学校 友人 (gakkou yuujin)
school grades	学校 成績 (gakkou seiseki)
bullying	いじめ (ijime)
friendships	友人関係 (yuujinkankei)
study	勉強 (benkyou)
anxiety about studying	勉強 不安 (benkyou fuan)
I do not know how to study	勉強 わからない (benkyou wakaranai)
mid-term or term-end exam	試験 (shiken)
entrance examination	受験 (juken)
future career	進路 (shinro)
grades	成績 (seiseki)

XSL•FO
RenderX

Table 2. Results of augmented Dickey-Fuller test on time-series data.

Variables	ADF[a] test statistics	*P* value
the number of suicides	–4.61	<.001
school	–3.91	.01
school reopening	–3.91	.01
I do not want to go to school	–6.72	<.001
school hard	–5.96	<.001
school anxiety	–4.89	<.001
When is school?	–3.49	.04
When does school start?	–4.21	<.001
school start	–4.08	<.001
school bullying	–5.22	<.001
school friends	–5.38	<.001
school grades	–15.67	<.001
bullying	–4.94	<.001
friendship	–6.15	<.001
study	–5.77	<.001
anxiety about studying	–5.45	<.001
I do not know how to study	–6.37	<.001
mid-term or term-end exam	–5.84	<.001
entrance examination	–4.67	<.001
future career	–4.53	<.001
grades	–3.69	.03

[a]ADF: augmented Dickey-Fuller.

XSL•FO
RenderX

Table 3. Temporal pre-post relationship by Granger causality (Google Trends model).

Null hypothesis	Optimal VAR[a] lag length	*F* test (*df*)	*P* value
“school” does not Granger-cause the number of suicides	7	1.10 (7, 476)	.36
The number of suicides does not Granger-cause “school”	7	0.63 (7, 476)	.74
“school reopening” does not Granger-cause the number of suicides	9	0.79 (9, 464)	.63
The number of suicides does not Granger-cause “school reopening”	9	1.32 (9, 464)	.22
“I do not want to go to school” does not Granger-cause the number of suicides	1	7.55 (1, 512)	.006
The number of suicides does not Granger-cause “I do not want to go to school”	1	3.79 (1, 512)	.05
“school hard” does not Granger-cause the number of suicides	1	0.01 (1, 512)	.94
the number of suicides does not Granger-cause “school hard”	1	1.51 (1, 512)	.22
“school anxiety” does not Granger-cause the number of suicides	1	0.94 (1, 512)	.33
the number of suicides does not Granger-cause “school anxiety”	1	0.39 (1, 512)	.53
“When is school?” does not Granger-cause the number of suicides	3	0.19 (3, 500)	.90
the number of suicides does not Granger-cause “When is school?”	3	0.21 (3, 500)	.89
“When does school start?” does not Granger-cause the number of suicides	12	0.93 (12, 446)	.51
the number of suicides does not Granger-cause “When does school start?”	12	0.56 (12, 446)	.88
“school start” does not Granger-cause the number of suicides	3	0.69 (3, 500)	.56
the number of suicides does not Granger-cause “School start”	3	0.74 (3, 500)	.53
“school bullying” does not Granger-cause the number of suicides	1	2.80 (1, 512)	.09
the number of suicides does not Granger-cause “school bullying”	1	0.13 (1, 512)	.72
“school friends” does not Granger-cause the number of suicides	1	0.11 (1, 512)	.74
the number of suicides does not Granger-cause “school friends”	1	1.42 (1, 512)	.23
“school grades” does not Granger-cause the number of suicides	2	0.32 (2, 506)	.73
The number of suicides does not Granger-cause “school grades”	2	1.72 (2, 506)	.18
“bullying” does not Granger-cause the number of suicides	2	1.85 (2, 506)	.16
the number of suicides does not Granger-cause “bullying”	2	0.75 (2, 506)	.47
“friendship” does not Granger-cause the number of suicides	1	0.02 (1, 512)	.89
the number of suicides does not Granger-cause “friendship”	1	0.07 (1, 512)	.79
“study” does not Granger-cause the number of suicides	5	2.63 (5, 488)	.02
the number of suicides does not Granger-cause “study”	5	0.54 (5, 488)	.74
“anxiety about studying” does not Granger-cause the number of suicides	4	0.39 (4, 494)	.82
the number of suicides does not Granger-cause “anxiety about studying”	4	0.40 (4, 494)	.81
“I do not know how to study” does not Granger-cause the number of suicides	1	0.08 (1, 512)	.77
the number of suicides does not Granger-cause “I do not know how to study”	1	0.26 (1, 512)	.61
“mid-term or term-end exam” does not Granger-cause the number of suicides	3	0.27 (3, 500)	.85
the number of suicides does not Granger-cause “mid-term or term-end exam”	3	0.39 (3, 500)	.76
“entrance examination” does not Granger-cause the number of suicides	5	0.86 (5, 488)	.51
the number of suicides does not Granger-cause “entrance examination”	5	0.58 (5, 488)	.71
“future career” does not Granger-cause the number of suicides	4	0.74 (4, 494)	.56
the number of suicides does not Granger-cause “future career”	4	0.89 (4, 494)	.47
“grades” does not Granger-cause the number of suicides	2	2.11 (2, 506)	.12
the number of suicides does not Granger-cause “grades”	2	2.07 (2, 506)	.13

[a]VAR: vector autoregressive.

XSL•FO
RenderX

Robustness of the Result

For the search words that were found to be significant in the above analysis, we conducted the robustness check by generating a dataset that excludes the effect of the school calendar. The reason for this check is to test whether our search keywords remain statically significant after controlling for the effect of the school calendar. If it remains so, it infers that our search keywords have a powerful predicting power for the number of suicides among elementary, middle, and high school students.

The results of the school calendar model rejected the null hypotheses "I do not want to go to school" does not Granger-cause the number of suicides and "study" does not Granger-cause the number of suicides both as in the Google Trends model (Table 4).

Table 4. Temporal pre-post relationship by Granger causality (school calendar model in the search terms "I do not want to go to school" and "study").

Null hypothesis	Optimal VAR[a] lag length	*F* statistics (*df*)	*P* value
"I do not want to go to school" does not Granger-cause the number of suicides	1	10.91 (1, 512)	.001
The number of suicides does not Granger-cause "I do not want to go to school"	1	2.74 (1, 512)	.10
"study" does not Granger-cause the number of suicides	2	4.01 (2, 506)	.02
the number of suicides does not Granger-cause "study"	2	1.43 (2, 506)	.24

[a]VAR: vector autoregressive.

Correlation Analysis

Subsequently, variance stabilization transformations were applied to the phrases shown by Granger causality. A cross-correlation analysis, which revealed that the time-series of suicides lagged by 2 weeks concerning both "I do not want to go to school" and "study," was performed (Table 5).

In this study, the search queries "I do not want to go to school" and "study" showed Granger causality with suicide incidences. The results of the cross-correlation analysis showed significant positive correlations in the range of –2 to 2 for "I do not want to go to school" (highest value at time lag 0, r=0.28) and –1 to 2 for "study" (highest value at time lag –1, r=0.18), indicating that the search volume increased as the number of suicides increased.

Furthermore, in 2020, during the COVID-19 pandemic, the time-series trend for "I do not want to go to school" resembled that of the number of suicides. Figure 2 shows the social situation in 2020 and the time-series of the number of suicides and search terms. Both variables decreased from March to May, when schools were closed nationwide, and increased from May to June, when the state of emergency was lifted. This relationship persisted even after the summer vacation and increased from September to October. In addition, 2 more sharp increases were observed after November. Conversely, the search volume for "study" increased from March to May and remained high thereafter, showing no significant change compared with "I do not want to go to school."

Table 5. Cross-correlation with the number of suicides at lags –3 to +3 after variance stabilization.

Search terms	–3	–2	–1	0	1	2	3
"I do not want to go to school"	0.04	0.09[a]	0.21[a]	0.28[a]	0.26[a]	0.17[a]	0.06
"study"	0.05	0.04	0.18[a]	0.17[a]	0.13[a]	0.12[a]	0.09

[a]Statistically significant values at P<.05.

XSL•FO
RenderX

Figure 2. Number of suicides and search queries during the COVID-19 pandemic period. Highlights indicate social conditions. The top line plots weekly suicide, the middle line plots "I do not want to go to school," and the bottom line plots "study.".

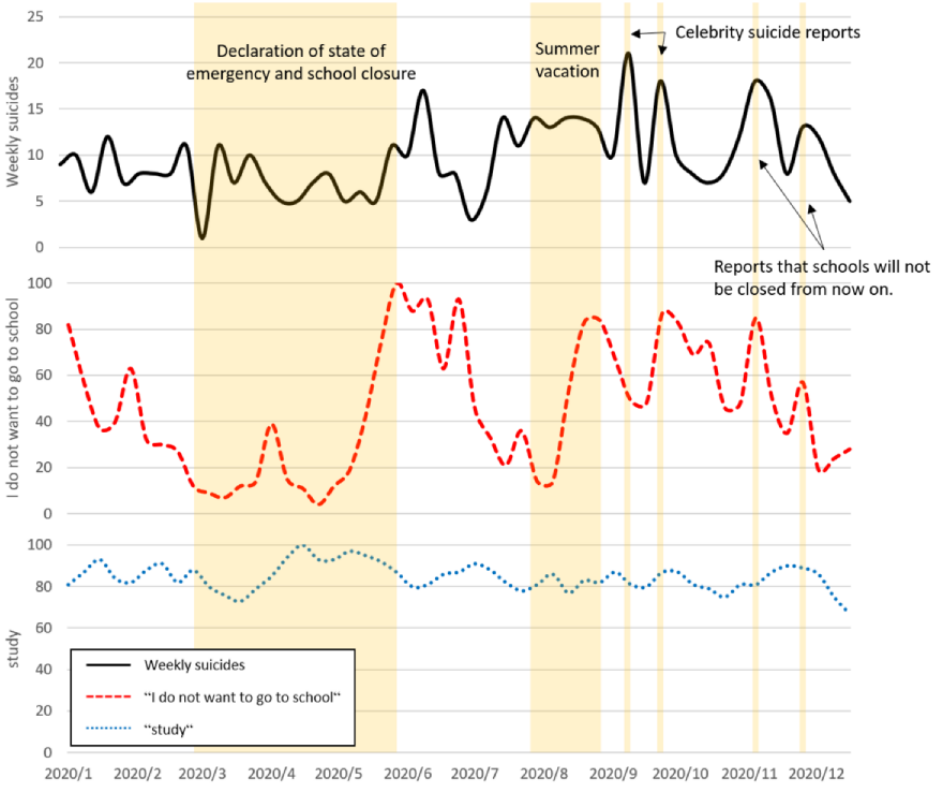

Discussion

Principal Findings

This study examines the association between the number of suicides among Japanese children and adolescents and the volume of school-related internet searches. We used Granger causality analysis to show that specific school-related internet search volumes may precede the number of suicides among Japanese elementary, middle, and high school students over time. In the Granger causality test, "I do not want to go to school" and "study" were statistically significant in both the Google Trends model and the school calendar model. Internet search terms that were significantly and positively correlated with the number of suicides were "I do not want to go to school" and "study." Regarding the time lag, a negative direction starting from 0 indicates the periods after the increase in search volumes, whereas a positive direction indicates the periods before the increase in search volumes. Therefore, to investigate whether the number of suicides increases after an increase in search volumes, the negative lag direction should be primarily considered. In this study, lags were calculated weekly, whereas previous studies used monthly aggregates, where a lag of approximately 2-3 months was observed [16,18,22]. However, the lag in this study was less than 2 weeks, indicating the high accuracy of this leading indicator. The results in Table 4 imply that a lag in either direction indicates a short-term association between the search terms and suicide. The positive correlation between the number of suicides and the search volumes suggests that search terms may be a leading indicator of suicide. Furthermore, the volume of searches for "I do not want to go to school" showed a time-series trend similar to the number of suicides among elementary, middle, and high school students, even in the social context of the 2020 COVID-19 pandemic. To the best of our knowledge, our study is the first to investigate the association between the weekly number of elementary, middle, and high school student suicides and the volume of internet searches for school-related terms in Japan. The discovery of new linguistic markers associated with suicide among elementary, middle, and high school students may contribute to research and practice related to suicide prevention. As our analysis is exploratory, aimed at generating hypotheses for future research, we did not apply multiple comparison corrections [29]. Future studies will need to apply more stringent statistical controls when testing these initial hypotheses to validate the findings.

In 2020, during the COVID-19 pandemic, similar time-series trends were revealed between the search volume "I do not want to go to school" and the number of suicides. In Japan, schools were closed nationwide on March 2, 2020, and were reopened when the state of emergency was lifted on May 25, 2020. During this period, a sharp increase was noted in both the number of suicides and search volumes, which could be attributed to the increasing tension over the school's reopening. This trend could also be observed during the summer vacation from August to September. The increase in the number of suicides in September cannot necessarily be explained solely by the psychological burden of the resumption of school, since reports of celebrity suicides occurred. However, the number of searches for "I do not want to go to school" gradually began to increase before school resumed in September. This phenomenon may be attributed to an increase in the number of searches considered as suicide risk owing to the timing of the end of the long

XSL•FO
RenderX

vacation. Therefore, this result may have reinforced the "broken promise effect" theory [30], which suggests that suicide rates tend to increase after a long break from school.

Other negative school-related terms searched during the same period did not yield significant Granger causality or correlation coefficients. This suggested the robustness of the "I do not want to go to school" term being the high-risk keyword for suicide among elementary, middle, and high school students.

The 2 peaks in search volumes in November 2020 might have been influenced by press conferences conducted by the Ministry of Education, Culture, Sports, Science, and Technology on November 10 and November 27, which stated that entrance examinations and the nationwide closure of all schools would not be canceled or requested [31,32]. This statement may have recalled the psychological stress children and adolescents feel about schools. Therefore, government statements related to schools should be made with caution.

This study's findings are consistent with those of previous studies. A meta-analysis and systematic review indicated that school absenteeism and dropout rates are associated with an increased risk of self-harm and suicide attempts [33,34], suggesting that educational factors may influence suicide. The search term "I do not want to go to school" in this study may be precisely a proxy variable for measuring educational factors such as expectations and the stress of attending school. The results imply that suicide is closely related to the educational system, especially the opening and closing of schools. In other words, it is possible that school opening and closing are positively and negatively correlated with suicide risk, respectively. However, caution must be exercised while interpreting these results. This is because it has been reported that school ties may be a powerful protective factor against youth suicide [35].

They should enable children who do not want to go to school to continue their learning at home or in other settings by providing them with extensive support. For example, the government could provide financial support to encourage participation in alternative education programs. This support could include subsidizing the cost of accessing online learning platforms or providing the necessary funding to attend free schools. Schools should also adopt more flexible scheduling to motivate students to learn. This allows students to progress at their own pace and helps them to balance their academic and personal interests and activities. In addition, mental health support in schools and communities needs to be strengthened to help children for whom school is a distressing experience. The findings of this study contribute to a better understanding of support strategies for children who do not want to go to school and to proposals for promoting comprehensive change in the education system.

In a study in South Korea on suicide-related internet search terms among children and adolescents, the search volume for "dropout" showed a significant positive correlation [36], indicating that school-related search terms were associated with the number of suicides, which is consistent with our findings. However, our study differs in that we used weekly data rather than daily data, which would be affected by each day of the week, Monday through Sunday. Furthermore, our study fills a gap in previous studies because it shows the potential to be used as a risk factor not only in normal times but also in emergencies such as school closures associated with the COVID-19 pandemic.

This study's strength is exhibited in the process of selecting search terms through an exploratory approach. The internet search terms selected in this study were chosen through an exploratory review of official sources on suicide in Japan. The selection process did not follow a rigorous and systematic methodology. However, relying solely on a systematic list of search terms based on previous research does not necessarily uncover search terms that are effective for suicide prevention. One study hypothesized that internet searches related to past orientation would be positively related to suicide rates because people with regret and depression are more likely to dwell on the past [37]. In another study, from a clinical psychology perspective, search terms associated with absolutist thinking, such as "completely" and "totally," were used [38]. These emerging linguistic markers may not have been identified through standardized and broad-ranging methodologies. Although we could have selected specific "words" related to suicide, it may have been difficult to select the sentence "I do not want to go to school."

Online gatekeeping reduces suicidal ideation [39]. However, "Suicide Prevention Results" are not currently implemented for the search term "I do not want to go to school" in the Japanese Google search system. It is believed that whether a helpline is displayed depends on the language and search terms used [40,41]. Our study has the potential to improve the suicide prevention results for children and adolescents and enable effective referrals to counseling services.

Suicide surveillance in Japan has been delayed for more than 6 months [42]. Monitoring high-risk suicide keywords will help improve such delays and aid in rapid administrative support and mental health planning at the national and municipal levels. Furthermore, monitoring and evaluating surveillance systems involves not only promptness but also cost-effectiveness factors [43]. The Google Trends data used in this study can be obtained in real time and free of charge and could complement existing surveillance systems. This study's results may provide useful information for many stakeholders, such as schools, internet service providers, governments, and municipalities.

Limitations

This study has several limitations. First, the Google Trends dataset may be susceptible to noise, as search volumes can be influenced by external factors such as social conditions and media coverage, independent of the searcher's characteristics. In addition, Google Trends has been criticized for its inability to capture user demographics such as age and gender [7]. Our study assumes that the use of school-related search terms may limit the demographics of searchers to younger age groups. Indeed, future experimentation is needed to determine the exact demographic distribution of individuals who perform school-related searches.

XSL•FO
RenderX

Second, an increase in search volumes for specific keywords does not always correspond to an increase in the number of suicides. Some search terms are accompanied by online helpline advertisements in the browser, which may reduce suicide risk by guiding individuals to seek professional help. Consequently, the correlation between internet queries and the number of suicides may be nullified.

Third, the data used in this study only observes search behavior. Recent forms of social media, such as Twitter (rebranded as X) and Instagram (Meta), offer the distinct feature of allowing individuals to express their thoughts and situations, which diverges from searching behavior [24]. Consequently, these new media forms are currently being investigated regarding adolescent suicides [44]. In the future, it will be necessary to not only continually observe search terms through Google Trends but also scrutinize new media that enable the expression of psychological states.

Finally, while this study is concentrated on suicide among elementary, junior high, and high school students in Japan, yielding findings specific to this demographic within the Japanese context, the global availability of Google Trends and suicide data presents an opportunity for broadening the research scope. Therefore, future studies are planned to explore the applicability of these findings across different regions and cultures, aiming to understand the universality and limitations of our conclusions.

Conclusions

This study identifies an association between the volume of internet searches for school-related terms such as “I do not want to go to school” and “study,” and the number of suicides among children and adolescents in Japan. By analyzing the search volume for school-related terms, which have not received much attention in previous studies, the study suggests the possibility of improving online helpline displays for children and adolescents. The volume of searches for school-related terms may also be useful in monitoring suicide risk among children and adolescents, not only during normal times but also during emergencies such as school closures. However, because internet search volume includes external factors unrelated to suicidal characteristics, it may not be easy to generalize the findings and incorporate them into national and municipal suicide prevention strategies. Further empirical studies will be needed in the future.

Acknowledgments

We thank all the members of the Japan Suicide Countermeasures Promotion Center for their warm guidance and encouragement throughout this research, as well as Professor Keita Yamauchi of Keio University for his important advice. This work was supported by JSPS KAKENHI (grants JP21H04403 and JP24K23761). The funding sources had no role in the study design, data collection, data analysis, data interpretation, or preparation of the manuscript.

Data Availability

Data supporting this study’s results require permission from the Japan Center for Suicide Prevention. Requests to access these datasets should be directed to TA, the corresponding author.

Authors' Contributions

TA contributed to the conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing original draft, writing-review and editing, visualization, supervision, and project administration. HT performed the formal analysis and writing-review and editing. AW assisted the interpretation of the findings, formal analysis, and writing-review and editing. YS performed the supervision.

Conflicts of Interest

None declared.

References

1. Ministry of Health, Labour and Welfare. Annual Report on Suicide Prevention. 2020. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/jisatsuhakusyo2022.html [accessed 2024-09-14]
2. Ministry of Education, Culture, Sports, Science and Technology. Conference of Research Collaborators on Suicide Prevention for Students. 2021. URL: https://www.mext.go.jp/content/20210216-mxt_jidou01-000012837_009.pdf [accessed 2024-09-14]
3. Cusimano MD, Sameem M. The effectiveness of middle and high school-based suicide prevention programmes for adolescents: a systematic review. Inj Prev 2011;17(1):43-49. [doi: 10.1136/ip.2009.025502] [Medline: 21059602]
4. Katz C, Bolton S, Katz LY, Isaak C, Tilston-Jones T, Sareen J, Swampy Cree Suicide Prevention Team. A systematic review of school-based suicide prevention programs. Depress Anxiety 2013;30(10):1030-1045. [doi: 10.1002/da.22114] [Medline: 23650186]
5. Sueki H, Ito J. Suicide prevention through online gatekeeping using search advertising techniques: a feasibility study. Crisis 2015;36(4):267-273. [doi: 10.1027/0227-5910/a000322] [Medline: 26440623]

XSL•FO
RenderX

6. Barros JM, Melia R, Francis K, Bogue J, O'Sullivan M, Young K, et al. The validity of google trends search volumes for behavioral forecasting of national suicide rates in Ireland. Int J Environ Res Public Health 2019;16(17):3201 [FREE Full text] [doi: 10.3390/ijerph16173201] [Medline: 31480718]
7. Capron DW, Andel R, Voracek M, Till B, Niederkrotenthaler T, Bauer BW, et al. Time-series analyses of firearm-related Google searches and U.S. suicide rates 2004-2016. Suicide Life Threat Behav 2021;51(3):554-563. [doi: 10.1111/sltb.12728] [Medline: 33426750]
8. Halford EA, Lake AM, Gould MS. Google searches for suicide and suicide risk factors in the early stages of the COVID-19 pandemic. PLoS One 2020;15(7):e0236777 [FREE Full text] [doi: 10.1371/journal.pone.0236777] [Medline: 32706835]
9. Jimenez A, Santed-Germán MA, Ramos V. Google searches and suicide rates in Spain, 2004-2013: correlation study. JMIR Public Health Surveill 2020;6(2):e10919 [FREE Full text] [doi: 10.2196/10919] [Medline: 32281540]
10. Lee JY. Search trends preceding increases in suicide: a cross-correlation study of monthly Google search volume and suicide rate using transfer function models. J Affect Disord 2020;262:155-164. [doi: 10.1016/j.jad.2019.11.014] [Medline: 31733460]
11. Lopez-Agudo LA. The association between internet searches and suicide in Spain. Psychiatry Res 2020;291:113215. [doi: 10.1016/j.psychres.2020.113215] [Medline: 32559673]
12. Adem Doganer ZJZ. An infodemiological analysis of google trends in COVID-19 outbreak: predict case numbers and attitudes of different societies. Journal of Database Management 2021;32(2):19. [doi: 10.4018/jdm.2021040101]
13. Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google trends time series. Sci Rep 2020;10(1):20693 [FREE Full text] [doi: 10.1038/s41598-020-77275-9] [Medline: 33244028]
14. Peng Y, Li C, Rong Y, Pang CP, Chen X, Chen H. Real-time prediction of the daily incidence of COVID-19 in 215 countries and territories using machine learning: model development and validation. J Med Internet Res 2021;23(6):e24285 [FREE Full text] [doi: 10.2196/24285] [Medline: 34081607]
15. McCarthy MJ. Internet monitoring of suicide risk in the population. J Affect Disord 2010;122(3):277-279 [FREE Full text] [doi: 10.1016/j.jad.2009.08.015] [Medline: 19748681]
16. Yang AC, Tsai S, Huang NE, Peng C. Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004-2009. J Affect Disord 2011;132(1-2):179-184. [doi: 10.1016/j.jad.2011.01.019] [Medline: 21371755]
17. Prime Minister's Office of Japan. The general principles of suicide prevention policy. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/taikou_h290725.html [accessed 2024-09-14]
18. Sueki H. Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: data from 2004 to 2009 in Japan. Psychiatry Clin Neurosci 2011;65(4):392-394 [FREE Full text] [doi: 10.1111/j.1440-1819.2011.02216.x] [Medline: 21569178]
19. Hagihara A, Miyazaki S, Abe T. Internet suicide searches and the incidence of suicide in young people in Japan. Eur Arch Psychiatry Clin Neurosci 2012;262(1):39-46. [doi: 10.1007/s00406-011-0212-8] [Medline: 21505949]
20. Taira K, Hosokawa R, Itatani T, Fujita S. Predicting the number of suicides in Japan using Internet search queries: vector autoregression time series model. JMIR Public Health Surveill 2021;7(12):e34016 [FREE Full text] [doi: 10.2196/34016] [Medline: 34823225]
21. Arendt F, Scherr S. Optimizing online suicide prevention: a search engine-based tailored approach. Health Commun 2017;32(11):1403-1408. [doi: 10.1080/10410236.2016.1224451] [Medline: 27739876]
22. Tran US, Andel R, Niederkrotenthaler T, Till B, Ajdacic-Gross V, Voracek M. Low validity of Google trends for behavioral forecasting of national suicide rates. PLoS One 2017;12(8):e0183149 [FREE Full text] [doi: 10.1371/journal.pone.0183149] [Medline: 28813490]
23. FAQ about Google trends data. Google. 2023. URL: https://support.google.com/trends/answer/4365533?hl=en [accessed 2024-09-14]
24. Fahey RA, Boo J, Ueda M. Covariance in diurnal patterns of suicide-related expressions on Twitter and recorded suicide deaths. Soc Sci Med 2020;253:112960 [FREE Full text] [doi: 10.1016/j.socscimed.2020.112960] [Medline: 32251933]
25. Matsubayashi T, Ueda M, Yoshikawa K. School and seasonality in youth suicide: evidence from Japan. J Epidemiol Community Health 2016;70(11):1122-1127. [doi: 10.1136/jech-2016-207583] [Medline: 27225682]
26. Sato K, Mano T, Iwata A, Toda T. Need of care in interpreting Google trends-based COVID-19 infodemiological study results: potential risk of false-positivity. BMC Med Res Methodol 2021;21(1):147 [FREE Full text] [doi: 10.1186/s12874-021-01338-2] [Medline: 34275447]
27. Trapletti. Time series analysis and computational finance. Tseries. 2019. URL: https://cran.r-project.org/web/packages/tseries/tseries.pdf [accessed 2024-09-14]
28. Pfaff B. VAR, SVAR and SVEC models: implementation within R package vars. Journal of Statistical Software 2008;27(4):1-32 [FREE Full text]
29. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. Malden, MA: Wiley-Blackwell; 2001.
30. Gabennesch H. When promises fail: a theory of temporal fluctuations in suicide. Social Forces 1988;67(1):129-145.
31. Minister of Education, Culture, Sports, Science and Technology Koichi Hagiuda Press Conference Transcript. 2020. URL: https://www.mext.go.jp/b_menu/daijin/detail/mext_00108.html [accessed 2023-07-31]

XSL•FO
RenderX

32. Minister of Education, Culture, Sports, Science and Technology Koichi Hagiuda Press Conference Transcript. 2020. URL: https://www.mext.go.jp/b_menu/daijin/detail/mext_00108.html [accessed 2023-07-31]
33. Castellví P, Miranda-Mendizábal A, Alayo I, Parés-Badell O, Almenara J, Alonso I, et al. Assessing the relationship between school failure and suicidal behavior in adolescents and young adults: a systematic review and meta-analysis of longitudinal studies. School Mental Health 2020;12(3):429-441. [doi: 10.1007/s12310-020-09363-0]
34. Epstein S, Roberts E, Sedgwick R, Polling C, Finning K, Ford T, et al. School absenteeism as a risk factor for self-harm and suicidal ideation in children and adolescents: a systematic review and meta-analysis. Eur Child Adolesc Psychiatry 2020;29(9):1175-1194 [FREE Full text] [doi: 10.1007/s00787-019-01327-3] [Medline: 30989389]
35. Tomek S, Burton S, Hooper LM, Bolland A, Bolland J. Suicidality in black American youth living in impoverished neighborhoods: is school connectedness a protective factor? School Mental Health 2018;10(1):1-11. [doi: 10.1007/s12310-017-9241-4]
36. Choi WS, Han J, Hong HJ. Association between Internet searches related to suicide/self-harm and adolescent suicide death in South Korea in 2016-2020: secondary data analysis. J Med Internet Res 2023;25:e46254 [FREE Full text] [doi: 10.2196/46254] [Medline: 37079349]
37. Lee D, Lee H, Choi M. Examining the relationship between past orientation and US suicide rates: an analysis using big data-driven Google search queries. J Med Internet Res 2016;18(2):e35 [FREE Full text] [doi: 10.2196/jmir.4981] [Medline: 26868917]
38. Adam-Troian J, Arciszewski T. Absolutist words from search volume data predict state-level suicide rates in the United States. Clinical Psychological Science 2020;8(4):788-793. [doi: 10.1177/2167702620916925]
39. Sueki H, Takahashi AT, Ito J. Changes in suicide ideation among users of online gatekeeping using search-based advertising. Arch Suicide Res 2023;27(4):1339-1350. [doi: 10.1080/13811118.2022.2131491] [Medline: 36237123]
40. Scherr S, Haim M, Arendt F. Equal access to online information? Google's suicide-prevention disparities may amplify a global digital divide. New Media & Society 2018;21(3):562-582. [doi: 10.1177/1461444818801010]
41. Arendt F, Haim M, Scherr S. Investigating Google's suicide-prevention efforts in celebrity suicides using agent-based testing: a cross-national study in four European countries. Soc Sci Med 2020;262:112692. [doi: 10.1016/j.socscimed.2019.112692] [Medline: 32057542]
42. Baran A, Gerstner R, Ueda M, Gmitrowicz A. Implementing real-time data suicide surveillance systems. Crisis 2021;42(5):321-327 [FREE Full text] [doi: 10.1027/0227-5910/a000829] [Medline: 34528451]
43. Live life: an implementation guide for suicide prevention in countries. WHO. 2021. URL: https://apps.who.int/iris/handle/10665/341726 [accessed 2023-07-31]
44. Niederkrotenthaler T, Stack S, Till B, Sinyor M, Pirkis J, Garcia D, et al. Association of increased youth suicides in the United states with the release of 13 reasons why. JAMA Psychiatry 2019;76(9):933-940 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0922] [Medline: 31141094]

Abbreviations

AIC: Akaike's information criterion
ADF test: augmented Dickey-Fuller test
RSV: relative search value
VAR: vector autoregressive

Edited by A Coristine; submitted 09.08.23; peer-reviewed by A Jimenez, H Kautz, A AL-Asadi; comments to author 24.02.24; revised version received 20.05.24; accepted 23.08.24; published 21.10.24.

Please cite as:
Arai T, Tsubaki H, Wakano A, Shimizu Y
Association Between School-Related Google Trends Search Volume and Suicides Among Children and Adolescents in Japan During 2016-2020: Retrospective Observational Study With a Time-Series Analysis
J Med Internet Res 2024;26:e51710
URL: https://www.jmir.org/2024/1/e51710
doi:10.2196/51710
PMID:

©Takahiro Arai, Hiroe Tsubaki, Ayako Wakano, Yasuyuki Shimizu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tracking the Spread of Pollen on Social Media Using Pollen-Related Messages From Twitter: Retrospective Analysis

Martín Pérez-Pérez[1,2,3], PhD; María Fernandez Gonzalez[4], PhD; Francisco Javier Rodriguez-Rajo[4], PhD; Florentino Fdez-Riverola[1,2,3], PhD

[1]CINBIO, Universidade de Vigo (University of Vigo), Vigo, Spain

[2]Department of Computer Science, School of Computer Engineering, Universidade de Vigo (University of Vigo), Ourense, Spain

[3]Next Generation Computer Systems Group, School of Computer Engineering, Galicia Sur Health Research Institute, Galician Health Service, SERGAS-UVIGO, Ourense, Spain

[4]Department of Plant Biology and Soil Sciences, Faculty of Sciences, Universidade de Vigo (University of Vigo), Ourense, Spain

Corresponding Author:
Martín Pérez-Pérez, PhD
Department of Computer Science
School of Computer Engineering
Universidade de Vigo (University of Vigo)
Polytechnic Building, University Campus As Lagoas s/n
Ourense, 32004
Spain
Phone: 34 988387039
Fax: 34 988387001
Email: martiperez@uvigo.gal

Abstract

Background: Allergy disorders caused by biological particles, such as the proteins in some airborne pollen grains, are currently considered one of the most common chronic diseases, and European Academy of Allergy and Clinical Immunology forecasts indicate that within 15 years 50% of Europeans will have some kind of allergy as a consequence of urbanization, industrialization, pollution, and climate change.

Objective: The aim of this study was to monitor and analyze the dissemination of information about pollen symptoms from December 2006 to January 2022. By conducting a comprehensive evaluation of public comments and trends on Twitter, the research sought to provide valuable insights into the impact of pollen on sensitive individuals, ultimately enhancing our understanding of how pollen-related information spreads and its implications for public health awareness.

Methods: Using a blend of large language models, dimensionality reduction, unsupervised clustering, and term frequency–inverse document frequency, alongside visual representations such as word clouds and semantic interaction graphs, our study analyzed Twitter data to uncover insights on respiratory allergies. This concise methodology enabled the extraction of significant themes and patterns, offering a deep dive into public knowledge and discussions surrounding respiratory allergies on Twitter.

Results: The months between March and August had the highest volume of messages. The percentage of patient tweets appeared to increase notably during the later years, and there was also a potential increase in the prevalence of symptoms, mainly in the morning hours, indicating a potential rise in pollen allergies and related discussions on social media. While pollen allergy is a global issue, specific sociocultural, political, and economic contexts mean that patients experience symptomatology at a localized level, needing appropriate localized responses.

Conclusions: The interpretation of tweet information represents a valuable tool to take preventive measures to mitigate the impact of pollen allergy on sensitive patients to achieve equity in living conditions and enhance access to health information and services.

(J Med Internet Res 2024;26:e58309) doi:10.2196/58309

KEYWORDS

pollen; respiratory allergies; Twitter; large language model; LLM; knowledge reconstruction; text mining

XSL•FO
RenderX

Introduction

Background

Allergies and associated diseases such as asthma, rhinosinusitis, atopic dermatitis, and food-oral syndrome affect at least 30% of the world population and close to 80% of families [1]. The World Health Organization estimates that 400 million people in the world experience allergic rhinitis and 300 million experience asthma [2]. In particular, allergy disorders caused by biological particles as proteins of some airborne pollen grains are currently considered one of the most common chronic diseases. European Academy of Allergy and Clinical Immunology forecasts indicate that, within 15 years, 50% of Europeans will have some kind of allergy as a consequence of urbanization, industrialization, pollution, and climate change [3-8]. Therefore, in less than half a century, allergy, originally perceived as a rare disease, has become a major public health threat.

According to recent data published by the European Academy of Allergy and Clinical Immunology, allergy constitutes a significant cause of morbidity worldwide, considered a pandemic with an impact on the quality of life and the health and medical systems of both resource-rich and resource-limited economies [9,10]. An important impact of pollen levels is observed in the increase in overconsultations in primary care for allergic rhinitis [11]. Moreover, it is estimated that, in Europe, asthma and rhinitis caused by allergies generates >100 million days of absenteeism at work and schools, in addition to the economic losses due to presenteeism, a circumstance in which people go to work but their productivity is extremely reduced [3]. Some studies estimate that the total cost of these allergic disease–induced situations to the United States was US $19.7 billion, and in the European Union, economic losses ranged from €55 to €151 billion (US $61 to US $167.5 billion) [4,12].

Preventive measures during high pollen seasons are medications; immunotherapy; and the adoption of strategies such as staying indoors during peak pollen times, using air purifiers, and keeping windows closed to reduce exposure [13]. The most common information available to prevent allergies for patients with pollinosis is the number of pollen grains or aeroallergens in the bioaerosol and their temporal distribution [10,14].

In addition, the timing of symptomatology of allergenic people represents suitable information to prevent allergies, but the difficulty of real-time patient symptom monitoring makes it very hard to obtain this kind of data. To solve this gap, epidemiological studies could be conducted with new technological tools such as the assessment of social media data to track the spread of pollen symptomatology–related messages [15,16]. During recent years, social networks have also become a significant tool for business and science dissemination and a reference to understand and study the trends, ideas, feelings, and thoughts of individual users and, therefore, are turning into a great opportunity to carry out new methodological studies in which individual people's accounts participate as coproducers of scientific knowledge [15,17,18]. Patient narratives on social networks can provide valuable insights into public awareness and seasonal patterns to take preventive measures to mitigate the impact of pollen on sensitive patients. A better understanding of exposure and symptom levels will help significantly improve new therapies [19].

Several key factors should be considered for an accurate social media data analysis. The accurate identification of platforms where users frequently discuss health-related topics, such as Twitter (subsequently rebranded X), Facebook, Instagram, and health forums, is essential to selecting relevant keywords related to pollen, allergies, or related symptoms [20,21]. Among all the currently existing social networks, Twitter has a leading role as a broadcasting communication medium for researchers, stakeholders, or organizations [15,16,22-25]. Twitter has been used to extract public opinions and attitudes in previous health studies related to COVID-19 [26] or Alzheimer disease [27], demonstrating how Twitter discourse analysis can provide insights into public opinions and attitudes on health and social issues. In addition, it is important to define the precise period for an accurate social media retrospective analysis considering seasonal variations in pollen levels. For accurate and current insights from global social media analysis, (1) irrelevant content such as retweets, spam, and duplicates must be meticulously removed during data cleaning [15,22]; (2) the frequency of pollen-related messages over time to identify peak periods should be determined; and (3) the geographical distribution of pollen-related discussions should be explored to identify regions with earlier prevalence, which serve as forecast for other regions, and trends in pollen-related discussions [17]. Comparing social media data with official pollen count data reveals valuable insights—online discussions correlate with actual pollen levels, pinpointing periods of peak pollen-related messages that suggest high pollen activity. This information can inform recommendations for health organizations, policy makers, and the public, tailored to address health trends related to different conditions [15,21,24].

Objectives

In this context, this study aimed to leverage advancements in data analysis and social networks using the semantic knowledge embedded in the latent space of large language models (LLMs), unsupervised clustering, term frequency–inverse document frequency (TF-IDF), and graph knowledge reconstruction techniques to track and scrutinize the dissemination of messages related to pollen symptoms from December 2006 to January 2022. This sophisticated approach facilitated a detailed examination of public discussions and trends on Twitter, offering insights into the impact of pollen on sensitive individuals.

Methods

Overview

This section summarizes the different steps and methods applied to obtain pollen-related knowledge on the Twitter social network. Figure 1 presents the workflow applied following the methods described in the following sections.

XSL•FO
RenderX

Figure 1. Applied analysis workflow. The analysis began with collecting pollen-related tweets and user data, filtered for English content and curated for problem domain relevancy. KNN: k-nearest neighbor; LLM: large language model; PCA: principal component analysis; TF-IDF: term frequency–inverse document frequency.

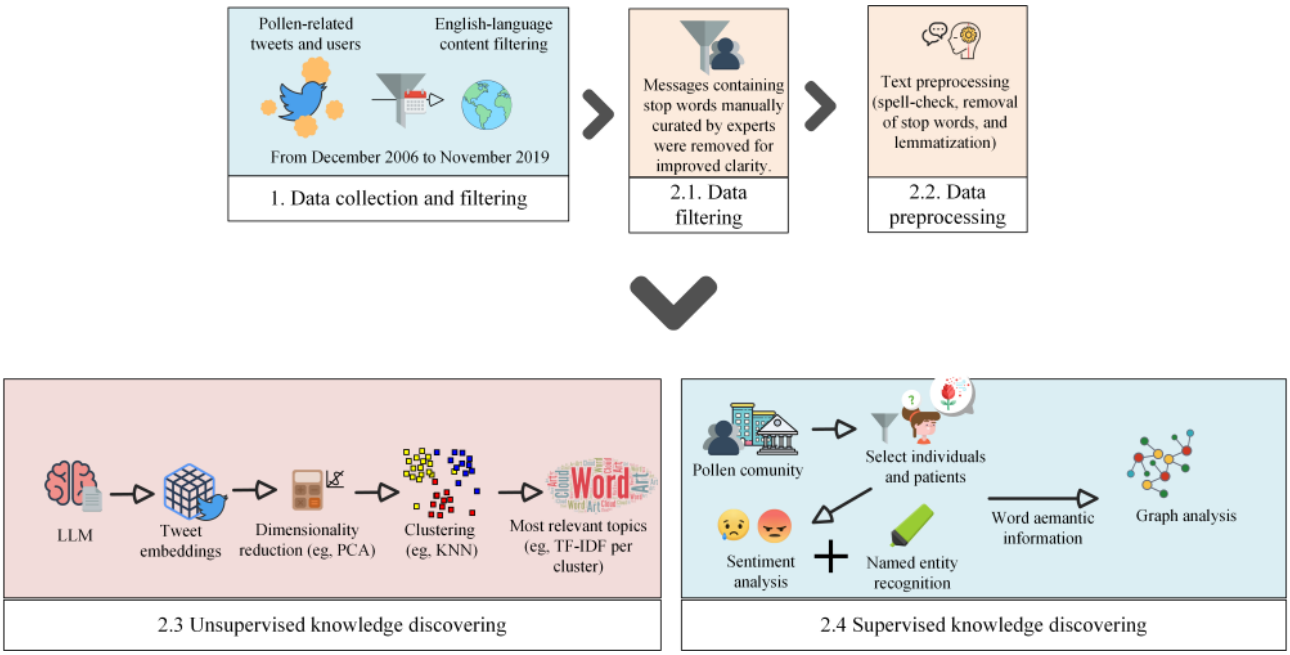

This analysis focused on pollen-related tweets and user data collected from December 2006 to January 2022. Tweets were filtered to retain only English-language entries and remove those irrelevant to the study domain using a handcrafted stop word list (eg, spam messages or nonrelevant topics). The refined data then underwent text preprocessing, including spell-checking, standard stop word removal, and lemmatization to normalize the word representation.

Upon completing the text preprocessing, both supervised and unsupervised methods were applied. In the unsupervised workflow, the messages were encoded using LLMs to map out semantic latent spaces as word embeddings. These embeddings were then dimensionally reduced via principal component analysis (PCA) to simplify the dataset complexity while preserving its essential structure. Subsequently, clustering algorithms such as k-nearest neighbor (KNN) were used to identify natural groupings within the data. To pinpoint the most pertinent topics within each cluster, further analysis applied the TF-IDF method, which highlights words that are more significant within a cluster relative to the entire dataset.

On the other hand, in the supervised workflow, usernames and descriptions were automatically analyzed to classify individual and nonindividual profiles (eg, public entities or news channels). Sentiment analysis was performed to assign emotional values to tweets and gauge public sentiment, whereas named entity recognition (NER) was applied to identify and categorize key entities relevant to the problem domain, such as symptoms or diseases. Finally, graph analysis was conducted to visualize the semantic interactions of the annotated words. This comprehensive approach provided a detailed understanding of the discourse and patterns related to pollen and its effects on the community throughout the analyzed period.

The following sections provide a more detailed exploration of the methods implemented in each component of the workflow described previously.

Data Collection and Filtering

To carry out the proposed analysis, a collection of 4.6 million English-language tweets that mentioned the word *pollen* from December 2006 to January 2022 were collected using the Twitter4J library, a Java tool for the Twitter application programming interface. To obtain relevant information in the topic domain, some repeated tweets and tweets mentioning noninteresting words were filtered out during the data collection process as well as retweets to prevent redundant information. However, it is important to note that not all collected messages contained relevant biomedical knowledge. The term *pollen* is used by the Twitter community to refer to a variety of topics, including video games, digital coins, insects, photography, and pollen level alerts. To ensure the relevance and specificity of the dataset, the tweets were meticulously filtered to exclude messages that were not relevant to the domain of the study. This semiautomatic filtering process involved retaining only English-language tweets and removing those with content irrelevant to the research focus. A carefully curated stop word list developed by a domain expert was used to exclude messages that were deemed unconnected to the health-related research topic. This list included words associated with unrelated subjects, such as *bees*, *weather channels*, and *pollen products for sale*, among others. This data curation approach helped refine the whole dataset, ensuring that the analysis was focused and relevant to the health domain.

Data Preprocessing

The content of the collected tweets could vary from useful and meaningful information to unconnected text. To extract high-quality information and features, certain preprocessing

XSL•FO
RenderX

techniques were required. Therefore, the following preprocessing steps were applied.

Different techniques were applied to normalize the user messages, namely, (1) URL identification and filtering, (2) word normalization (eg, *haaaaaaaaaaaaappy* to *happy*), (3) removal of extra white space, and (4) spelling error correction using the Hunspell dictionary with the normalized Levenshtein algorithm [28]. Finally, the mentions of Twitter users were identified and filtered to ensure compliance with privacy and ethical considerations.

Different techniques were applied to preprocess texts for topic discovery, clustering, and NER. The semantic normalization processing included tokenization to separate the text into meaningful elements such as words and phrases; removal of common and low-value words such as English stop words or domain-specific ones (eg, pollen); categorization each token using part-of-speech tagging; filtering of numbers and tokens with <2 characters; transformation of all tokens to lower case; and, finally, standardization of tokens to use only their root form through lemmatization techniques.

Extracting the Main Discussed Topics in Social Media Conversations

To uncover the main topics discussed by the community within the evaluated time frame (December 2006-January 2022), this study implemented (1) a topic modeling strategy using large language embeddings, specifically, Bidirectional Encoder Representations From Transformers (BERT), to explore the semantic latent space; (2) a vector dimension reduction (ie, PCA); (3) clustering (ie, KNN); and (4) a word relevance statistical analysis method (TF-IDF) to discover dense clusters (eg, group-discussed subjects) and easily describe and interpret the most relevant discussed topics. The following sections describe each processing technique.

Applying LLM Semantic Latent Space to Discover the Main Social Media Topics

Document embeddings are dense vector representations that capture the semantic latent space, enabling mathematical comparisons of text based on the vector space similarity. By representing documents as real-valued feature vectors, it becomes possible to perform topic modeling and other natural language processing (NLP) tasks, measuring the closeness of documents in vector space. BERT, one of the most widely used language representation models, has established itself as a benchmark for NLP tasks due to its ability to extract high-quality embeddings from text corpora [29,30]. The BERT model has been widely used for text classification tasks. To build its understanding of language, BERT pretrains a deep neural network on a large corpus of text data. This approach gives the model a comprehensive understanding of the semantic structure and context of language. BERT's bidirectional nature enables it to interpret the context of words by analyzing those that precede and follow each word in a sentence, leading to superior performance compared to traditional language models. BERT has been successfully applied to various text classification tasks, including sentiment analysis, topic classification, and fake news detection, often producing state-of-the-art results. BERT models can also be used for clustering tasks by leveraging their semantic latent space. BERT generates high-dimensional embeddings that capture the contextual and semantic information of the input text. In this sense, BERT embeddings were used to extract text features and group similar social media discussions together via clustering algorithms such as k-means clustering.

Dimensionality Reduction

The high-dimensional nature of LLMs such as the BERT model generates embeddings with hundreds or thousands of dimensions multiplied by the number of messages to analyze, which can be computationally expensive and memory intensive to work with. In addition, the large extension of the dataset generated from social media data makes it even more expensive to process the data.

In this context, dimensionality reduction techniques such as PCA are emerging as effective tools when combined with LLMs to decrease dataset dimensions. This integration simplifies data handling and enhances its visualization and interpretability [31]. PCA is a linear technique that identifies the directions in the data that capture the most variation and projects the data onto a lower-dimensional space along these directions. In this context, PCA was used to condense the semantic vector space matrix, thereby optimizing both performance and the representation of the evaluation steps.

Clustering Topic Analysis

Clustering methods such as KNN are fundamental techniques in the field of text mining. KNN is a straightforward and effective strategy for grouping similar data points based on their proximity to each other. This method facilitates the discovery of unsupervised knowledge by bringing together similar text documents, thereby revealing patterns and connections within the data that might not have been immediately clear. The integration of KNN with BERT latent space offers a powerful methodology for unsupervised text analysis. When texts are transformed into embeddings, the resulting vectors contain a vast amount of semantic information. This allows for the comparison of texts and the measurement of similarity based on these semantic features. In the context of clustering, these embeddings can serve as input for KNN, enabling it to identify and group texts based on their semantic similarities. In this way, a text can be assigned to the cluster of the k-nearest texts considering the similarities in their embeddings. This approach helps in categorizing text documents efficiently and effectively, illuminating hidden structures within large sets of data. When these embeddings are processed using KNN, the algorithm can accurately group tweets with similar content, thereby identifying prevailing topics within the Twitter discourse.

Word Relevance and Clustering Representation

TF-IDF is a common representation used in text analysis for the importance of a word in a document in relation to a corpus of documents. TF-IDF is calculated as the product of 2 measures: the term frequency and the inverse document frequency. The term frequency measures the frequency of a word in a document, whereas the inverse document frequency measures the rarity of a word in the dataset. In the context of clustering, the TF-IDF representation was used to identify the

XSL•FO
RenderX

keywords or phrases that are characteristic of a particular cluster and to provide a more interpretable and concise representation of how the social community expresses itself and associates every concept or eventual idea. TF-IDF is defined as follows (equation 1):

$$tf - idf(t, s, D) = tf(t, s) \times idf(t, D),$$

where t is the evaluated term, s stands for any given sentence of the dataset D, and $tf(t, s)$ expresses the ratio corresponding to the term t in a sentence s, described as follows (equation 2):

where n_t is the number of occurrences of the term t in a sentence s and n_k is the total number of terms in a sentence s. Moreover, in equation 1, $idf(t, D)$ stands for the logarithmic ratio of the term t in the dataset D and is computed as follows (equation 3):

Extracting Health-Related Patterns in Social Media Discussions

Overview

Analytical methods, including NER and coattention semantic networks, complement broader analytical efforts by extracting key insights from biomedical and health-related datasets. These techniques leverage labeled data and predefined categories to pinpoint crucial information effectively. By integrating the outcomes of this guided research with sentiment and semantic graph analyses, it is feasible to identify novel patterns and connections within the data that might otherwise remain elusive. Therefore, to discover relevant knowledge, the following techniques were applied.

NER Technique

NER enables a more accurate and comprehensive health-related analysis of medical topics discussed on social media compared to unsupervised methods that only rely on usual word frequency and patterns. By leveraging NER, relevant medical information can be extracted and analyzed, uncovering important insights and knowledge that might otherwise go unnoticed in the studied domain. Therefore, in terms of term recognition, an automatic processing workflow supported by different automatic state-of-the-art recognizer tools was established to identify relevant health topics in the problem domain.

For this process, the following 6 state-of-the-art NER taggers were used to assist the annotator and save efforts in the semiautomatic annotation workflow: TMCHEM [32] to identify chemical, drug brand, and trade names; LINNAEUS [33] to annotate species; disease name normalization (DNORM) [34] to recognize disease names; A Biomedical Named Entity Recognizer (ABNER) [35] to annotate genes and proteins; and Open Source Chemistry Analysis Routines (OSCAR4) [36] to recognize chemical names, reaction names, enzymes, chemical prefixes, and adjectives. In addition, to complement the annotation process and annotate the domain categories of *Disease*, *Food & Nutrition*, *Anatomy*, *Drug and Chemical compounds*, *Symptoms*, *Plants*, and *Diet*, an in-house ontology-based NER was established based on the following ontologies: the FoodOn ontology [37], the National Cancer Institute Thesaurus ontology [38], the Symptom Ontology [39], the Foundational Model of Anatomy ontology [40], the Medical Subject Headings ontology [41], the Chemical Entities of Biological Interest lexicon [42], the Disease Ontology [43], the Experimental Factor Ontology [44], the International Classification of Diseases [45], the Interlinking Ontology for Biological Concepts [46], the Human Phenotype Ontology [47], the DrugBank lexicon [48], the Kyoto Encyclopedia of Genes and Genomes lexicon [49], and a list of domain terms manually curated by experts.

This in-house ontology-based NER entailed a dictionary lookup as well as pattern and rule-based annotation to perform an inverted recognition strategy in which sentence words were used as patterns to be matched against the established lexicon from the ontologies and the manually curated list of domain terms [50]. In this way, the lexical knowledge provided by the different ontologies allowed for the recognition of perfect matches as well as lexical variations of the terms (ie, lemmatized entries and synonyms).

Coattention Semantic Graph Reconstruction

Coattention semantic graphs represent a valuable asset for extracting pertinent insights into the dynamics of how social media communities engage with different health-related subjects. This approach enables a more nuanced understanding of public sentiment, prevalent misconceptions, and emerging concerns, thereby offering a comprehensive view of the collective health discourse on social platforms. These graphs are built by analyzing the frequency and patterns of co-occurrence of terms in a corpus of text data, such as social media posts or online articles. The resulting graph represents a map of the semantic relationships between the different health-related concepts discussed, with nodes representing the word concepts (eg, *Itchy eye*, *Runny nose*, or *fatigue*), edges representing the existence of a discussed relationship in the community, and the weight of each edge representing the strength of the co-occurrence relationship between them (ie, the number of social media messages that discuss both terms in conjunction). Therefore, the concepts identified through NER were used to reconstruct co-occurrence semantic graphs of the Twitter messages that were evaluated. First, the relevant terms were identified using NER, which were then analyzed for their frequency of co-occurrence. The knowledge graph was then reconstructed by connecting the identified entities based on their semantic comention. This process enables the discovery of hidden relationships and patterns in the data, providing valuable insights into the health-related relationships between different biomedical entities identified in the text data. This technique serves as a valuable methodology to analyze the degree of association between various health topics (eg, disease, symptoms, or treatments) discussed by a specific social media community (eg, patients or users expressing themselves in the first person).

Sentiment Analysis

Different studies have used sentiment analysis to explore the emotional component of tweets. Using NLP techniques, these analyses assessed the words or symbols in tweets to quantify

XSL•FO
RenderX

the intensity of positive and negative sentiments or emotions [51,52]. Therefore, and to gain insights into the sentiment of the messages posted on the social network concerning the main topic, sentiment analysis was carried out using the pretrained model Twitter-Robustly Optimized BERT Pretraining Approach-base for Sentiment Analysis [53]. This model was trained on a vast dataset of 58 million tweets and fine-tuned for sentiment analysis using the TweetEval benchmark, making it an effective tool for analyzing sentiment in social media messages.

Graph Analysis

To visualize the results of semantic graphs, the Gephi tool [54] and the Circle Pack layout [55] were used. Gephi is a widely used open-source network analysis and visualization tool that allows for the exploration and interpretation of complex graphs. The Circle Pack layout is a hierarchical layout algorithm that arranges nodes in a circular packing order, with the most important nodes usually positioned in the center and the less important ones in the periphery. In this way, the Circle Pack layout can help identify the most central and significant entities in the graph, making it an effective way to visualize the results.

Finally, to gain a deeper insight into the structure and the embedded knowledge, various graph statistical metrics were computed to describe their characteristics. State-of-the-art metrics such as degree, betweenness centrality, closeness centrality, characteristic path length, clustering coefficient, and the average number of neighbors were used for this purpose [56-58]. Degree centrality quantifies the total number of direct links (eg, the number of relations) to other vertices, and higher degrees indicate greater centrality. Betweenness centrality measures a node's importance in facilitating communication between other nodes. A node with higher betweenness centrality plays a more critical role, as many paths between other nodes must pass through it. Finally, closeness centrality measures the ease of connectivity between vertices, with a high closeness centrality corresponding to nodes with short average shortest paths to other nodes (eg, the ease with which a term is mentioned by the community) [59,60].

Community Characterization

The user profiles in this study underwent data collection and processing to classify the accounts involved in the conversations. Consequently, to save manual processing efforts and improve the final results, an automatic user characterization workflow was established based on the previous works of the authors in the problem domain [61]. The purpose of this step was to discern between individual and nonindividual users (ie, organizations; stakeholders; and educational, informative, or commercial accounts without personal connections), determine their sex (male or female), and identify their association with the main topic (ie, accounts belonging to patients with mentioned pollen allergy). This user-centered analysis allowed for the stratification of the final outcomes in terms of health.

Ethical Considerations

All data were obtained from the public social media platform Twitter using their application programming interface. In addition, the information obtained was freely published by the different users of the platform and remains under their control, allowing them to delete it at any time. The results generated do not contain any user-identifiable data, ensuring the privacy and anonymity of the individuals involved. Our methodology emphasizes aggregate trends and patterns rather than individual user behaviors.

Results

Knowledge Base Description

As previously mentioned, all user messages were preprocessed to discern relevant and nonrelevant messages to the health topic. In this sense, a total of 4.6 million messages were processed to create a final evaluable knowledge base of relevant tweets. The final dataset consisted of 4,643,994 tweets after the preprocessing and filtering steps. Of these, 1,393,304 tweets (30%) were deemed relevant to health, while 3,188,607 (68.7%) tweets were considered irrelevant. The preprocessing filtering step ensures a clearer understanding and deeper insights into the topics of interest, improving the final results and making the data more useful for discovering knowledge and decision-making. To carry out this process, a subsample of tweets was revised by an expert to select the most common words that were not related to the problem domain. Then, a fully automatic process filtered all tweets, or user accounts, that contained these words. Textbox 1 represents an extract of the nonrelevant topics of the filtered messages, some of them related to weather accounts and bee, honey, or pollen seller messages.

XSL•FO
RenderX

Textbox 1. Nonrelevant words and bigrams of the filtered messages. The textbox illustrates an extract of the top 15 nonrelevant topics from the filtered messages scaled by their occurrence in the knowledge base.

Top 15 nonrelevant words

- Bee
- Health
- Honey
- Plant
- Grain
- Benefit
- Fenner
- Food
- Discover
- Try
- Use
- Superfood
- Purifier
- Profile
- Eat

Top 15 nonrelevant bigrams

- Bee-honey
- Bee-health
- Benefit- health
- Fever-hay
- Rain-wash
- Discover-health
- Profile-superfood
- Discover-profile
- Count-mold
- Grass-weed
- Contain-preservative
- Plant-release
- Health-increase
- Color-preservative
- Plant ragweed

To gain insights into the progression of the dataset over time, Figure 2 shows the evolution of discussions on social media related to the *pollen* topic. The figure shows the percentage of tweets mentioning "declared patients" affected by pollen allergies (y-axis) each year (x-axis). This percentage was calculated by dividing the number of tweets mentioning patients who were affected by pollen allergies by the total number of tweets in that year. The graph includes the overall trend in the total number of tweets and a polynomial trend line, indicating the general tendency over the years. This graphical approach not only provides insights into the frequency of conversations but also enables analysis of the volume of discussions over time. Therefore, as can be observed in Figure 2, the years in which the word *pollen* was used in the greatest volume worldwide (in the context of health) were 2012 to 2013 and 2020 to 2021.

Along the same line, and to visually analyze the trends in patient engagement with the topic of *pollen* over time and provide insights into the changing levels of interest and engagement among patients, Figure 2 also shows the trends in patient tweets as a proportion of the total number of tweets for each year from December 2006 to January 2022 (ie, patients who mentioned their allergic illness over time). The data show a clear upward trend, with a peak of 3.06% (4493/146,825) of patient tweets

XSL•FO
RenderX

in 2020. In particular, the percentage of patient tweets appeared to increase notably from 2018 to 2020, indicating a potential rise in pollen allergies and related discussions on social media. On the other hand, and with a deep granularity, the data suggest that there is a clear and expected seasonal pattern in the discussion of pollen on Twitter. The months between March and August had the highest volume of messages. April was the month in which the community discussed pollen the most. These data were in accordance with those of another health-related social media study conducted in the United States that pointed out that most tweets per month associated with telehealth for mental health or substance abuse were posted from March 2020 to June 2021, with a peak in May 2020 [62].

Accordingly, for the whole period under analysis, the days between May 7, 2018, and May 10, 2018, correspond to the period of the greatest popularity of the word *pollen* on Twitter (also in terms of favorites and retweets). In terms of year of popularity, 2021 was the year with the highest number of retweets and favorites concerning the topic, which demonstrates a growing interest in pollen and a growing information flow.

Figure 2. Evolution of "pollen" discussions on Twitter from December 2006 to January 2022. The figure shows the annual percentage of tweets mentioning “declared patients” affected by pollen allergies, with a polynomial trend line indicating overall tendencies.

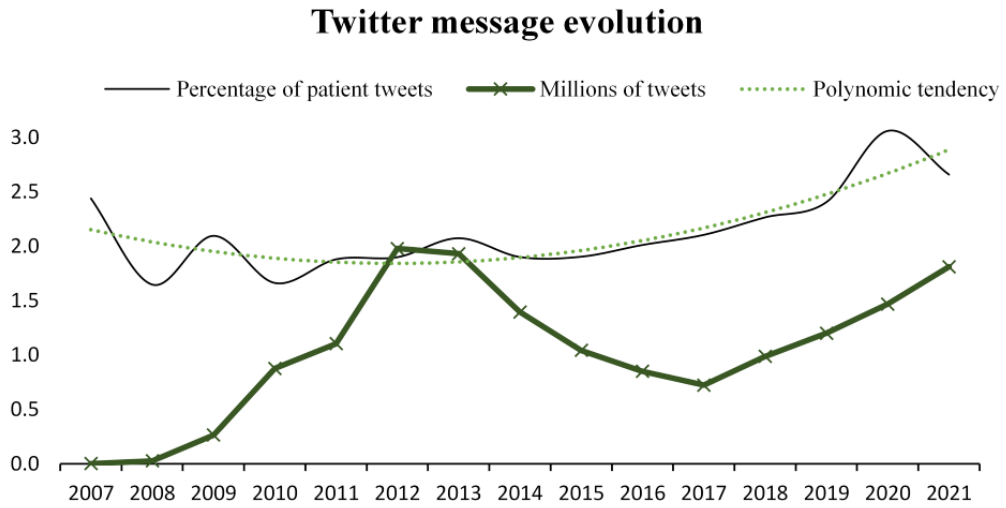

To identify the most discussed subjects in the community, an unsupervised topic analysis of the relevant messages (1.39 million) was carried out. Therefore, Table 1 represents the top 12 main topics discussed from December 2006 to January 2022. To obtain the most important words for each topic, the TF-IDF word relevance technique was used. The x-axis represents the TF-IDF word relevance of each word or term (ranging from 0 to 1) within the set of evaluated messages. The results obtained showed that the principal symptoms mentioned were related to ailments such as eye problems, allergic reactions, hay fever, and other symptoms (nose and head related). The data illustrate that conversations not only were centered on physical symptoms such as eye irritation, sneezing, and respiratory issues but also encompassed a strong emotional and social component.

The presence of strong language and expletives reflects elevated levels of frustration and discomfort among individuals affected by pollen allergies. Moreover, the association of pollen with weather-related terms such as *rain*, *wind*, and *mold* suggests that public awareness and concerns about pollen allergies are closely linked to environmental conditions. This could indicate that people are turning to social media to seek advice or share individual experiences related to weather changes affecting their allergy symptoms. One key finding was the significant mention of specific environments and activities, such as *street*, *night*, *morning*, and *work*, suggesting that discussions about pollen allergies are often contextualized within everyday scenarios and concerns. Furthermore, the prominence of terms related to emotional responses and coping strategies, such as *happy*, *hate*, *fight*, and *help*, underscores the emotional burden of pollen allergies and the community’s efforts to seek and offer support. Finally, the analysis revealed a nuanced conversation about weather and seasonal changes, indicated by the association with terms such as *weather*, *wind*, *rain*, and climatic change This suggests that public concerns and experiences with pollen allergies are closely tied to weather patterns and climate change phenomena. The same trend has been pointed out by other authors who measured the frequency of the words *allergy*, *asthma*, and *rhinitis* in tweets identifying the weather as the most relevant topic related to *asthma* and *rhinitis* [63].

In addition, a temporal analysis of the identified topic clusters was carried out to obtain insights into the dynamics of online discussions and the evolution of public opinion. Therefore, the temporal analysis of social media data can reveal historical trends and the relationships between different topics, providing a comprehensive view of the dynamics of online discussions and their evolution over time. Figure 3 illustrates the progression of the main topics identified over the assessed period. The graph shows the relevance of each cluster over time, illustrating the fluctuating interest and discussion trends within the Twitter community on pollen-related topics.

XSL•FO
RenderX

Table 1. Top 12 main topic clusters showing the different concepts discussed together from December 2006 to January 2022.

	TF-IDF[a]
Topic 0	
Street	0.22
Street social	0.21
Social	0.21
Thank	0.16
Great	0.13
Work	0.12
Topic 1	
Eye	0.38
Feel	0.29
Sneeze	0.26
Die	0.10
Face	0.09
Stop	0.09
Breathe	0.08
Topic 2	
Stupid	0.09
Head	0.07
Walk	0.05
Rid	0.04
Blame	0.04
Tired	0.03
Atlanta	0.03
Topic 3	
Fuck	0.36
Shit	0.31
Ass	0.28
Outside	0.25
Really	0.18
Kill	0.17
Man	0.14
Topic 4	
Bad	0.33
Well	0.23
Hope	0.14
Mask	0.14
Wear	0.10
Still	0.09
Year	0.09
Topic 5	
Low	0.36
Grass	0.35

XSL•FO
RenderX

	TF-IDF[a]
Mold	0.25
Weather	0.23
Moderate	0.22
Weed	0.15
Wind	0.14
Topic 6	
Plant	0.31
Love	0.22
Cat	0.11
Cover	0.11
Little	0.11
Male	0.10
Try	0.09
Topic 7	
Hate	0.28
Allergic	0.28
Nose	0.22
Throat	0.10
Sinus	0.09
Fucking	0.07
Itch	0.06
Topic 8	
Suffer	0.23
Hay fever	0.19
Help	0.19
Hayfever	0.19
Fever	0.19
Hay	0.18
Cause	0.15
Topic 9	
Rain	0.56
Away	0.33
Wash	0.32
Wash away	0.25
Rain wash	0.17
Snow	0.10
Morning	0.09

[a]TF-IDF: term frequency–inverse document frequency.

XSL•FO
RenderX

Figure 3. Evolution of the different topic clusters in the pollen-related discussions and their relevance from December 2006 to January 2022.

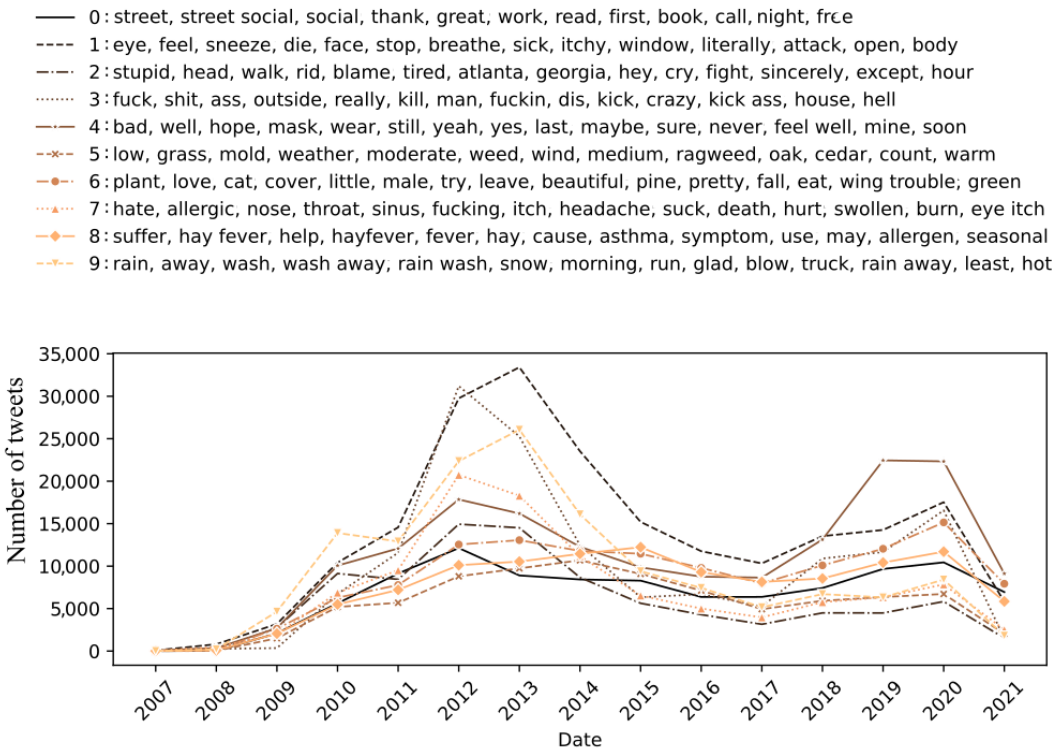

The consistent pattern observed across the data series presented in Figure 3 hints at a close connection between the different topics identified, suggesting that some years had a greater impact on the community (2011-2014). While most topics (topics 1, 2, 3, 7, and 8) primarily addressed medical concerns, expressions about health status, and symptomatology associated with pollen allergies, there were distinct topics that diverged to explore various other subjects. This variety in subject matter reflects the multifaceted nature of public discourse surrounding pollen allergies, encompassing not only health-related issues and community support but also broader environmental, social-cultural, and external events that have influenced the community. For example, topic 6 provides a broad perspective on pollen, addressing its presence in various contexts from environmental exposure to its role in natural beauty and even as a food supplement. The interplay of terms such as *plant*, *beautiful*, and *eat* suggests an acknowledgment of pollen's complexity, encompassing its allure and nutritional aspects. Since 2019, the discourse on mask use (topic 4) shifted significantly due to the COVID-19 pandemic, impacting discussions related to pollen allergies. This shift would be particularly visible in temporal data, where an uptick in mask-related discussions could coincide with the pandemic's onset, reflecting a broader societal adaptation to mask wearing.

Male Versus Female Topic Analysis

To better understand the primary conditions discussed in relation to accounts classified as male and female, the methodology described in the Extract Main Discussed Topics in Social Media Conversations section was applied to discover the main mentioned topics by sex. This approach specifically targeted messages from accounts identified by their sex, allowing for a detailed analysis of the health topics predominantly associated with each group. This sex classification was determined based on the demographic insights derived from the authors' previous research in the problem domain [61]. The messages attributed to each sex was segregated into 2 distinct clusters. Within these clusters, the top 7 topics demonstrating the highest impact, as measured using their TF-IDF scores, were determined. Table 2 represents the key topics discussed in relation to male- and female-classified accounts and measured by their term TF-IDF word relevance within the evaluated messages. Although no substantially differences were observed overall, it was found that asthma and sneezing were more significant discussed topics in the female messages.

XSL•FO
RenderX

Table 2. Most common pollen-related topics discussed by male and female individuals, and measured by their term frequency–inverse document frequency (TF-IDF).

	TF-IDF
Top male-discussed topics	
Topic 0	
Hayfever	0.03
Sinus	0.03
Eye	0.03
Rain	0.02
Nose	0.02
Grass	0.02
Kill	0.02
Topic 1	
Big snort	0.69
Snort sinus	0.68
Sinus big	0.63
Snort	0.39
Big	0.39
Sinus	0.25
Dammit sinus	0.05
Top female-discussed topics	
Topic 0	
Eye	0.03
Hayfever	0.03
Sinus	0.02
Sneeze	0.02
Asthma	0.02
Rain	0.02
Nose	0.02
Topic 1	
Big snort	0.88
Snort sinus	0.83
Sinus big	0.72
Big	0.43
Snort	0.40
Sinus	0.37
Pretty sure	0.06

Semantic Connections of Medical Topics

Different semantic knowledge networks were reconstructed to offer valuable insights into the behavior and semantics that the community uses to discuss. This approach has been used by other researchers as a method to uncover new knowledge about the topics being discussed. By analyzing the structure and connections within these networks, researchers can identify dominant themes and understand the contextual use of language within specific communities [64,65]. Figure 4 shows the relationships between the different hashtags created and represents the concepts that are deemed relevant by the community. The size and the color of the vertexes are based on the betweenness centrality of the hashtag (ie, the number of times a node acts as a bridge), whereas the edge size represents the number of users that used both connected hashtags in the same Twitter message. This graph provides valuable information about the public's interests proposed by the community itself,

XSL•FO
RenderX

assisting in the identification of the relationship between the topics that the community is most engaged with and their connection. Figure 4 shows the top comentioned hashtags raised by the community in the social media network. As can be observed, the high frequency of health-related hashtags in the semantic graph suggests that the Twitter community is primarily focused on providing support and information to individuals who may be experiencing symptoms or health issues. This highlights the importance of social media platforms to contribute to public health efforts, such as raising awareness about different health issues and encouraging healthy behaviors.

On the other hand, the hashtag comention analysis also revealed that the community was highly interested in topics such as *news*, *science*, *technology*, *did you know*, and *climate change*. These concepts had the highest betweenness centrality (ie, the number of times a vertex acts as a bridge) and a high degree (ie, a high-frequency mention) and were highly associated with health topics (eg, allergy, hay fever, or asthma). Therefore, this finding suggests that these concepts may play a key role in connecting and bridging different areas of interest and discussion within the community and may be particularly valuable for promoting greater engagement and interaction. In addition, these results indicate that the community was highly interested in staying informed about recent knowledge, scientific discoveries, and technological advancements, as well as advocating for environmental issues such as climate change. These concepts have also been used as keywords in other works; for example, *news* has been indicated as a theme to measure the frequency of tweets discussing *allergy* and *asthma* [63]. Other studies examining health communication during the Ebola outbreak analyzed the topic coverage and sentiment dynamics across 2 different media platforms: Twitter and traditional news publications. The findings suggest that Twitter and news media each presented distinct perspectives on the issue [66]. Another concept such as *technology* was used in a study that developed an app called Allergy Diary, in which the authors concluded that mobile technology is likely to become an important tool to better understand and manage allergenic rhinitis and asthma [67].

On the other hand, Figure 5 maps out the connections among distinct concepts identified through the NER process, highlighting their co-occurrence within the messages of the community. The vertex color denotes the biomedical domain category annotated (eg, disease), whereas the vertex size is dependent on its degree (ie, the number of annotated interactions). The edge color represents the evaluated sentiment (eg, positive or negative), whereas the width of the edge indicates the number of documents that evidenced the interaction. The semantic graph analysis provides a unique perspective on the health-related topics that are most frequently discussed in first-person messages, including those from patients, and the connections between them to better understand the community's interests, concerns, and behaviors. This approach offers a more comprehensive and nuanced understanding of the community's interests and priorities regarding health-related topics to better address the relationships and interactions between different relevant domain topics. These NER processes have been previously applied by other authors in biomedical tasks [68,69] or in the clinical domain [70]. The vertexes in the graph represent identified health-related concepts, whereas the edges denote semantic connections between them, in addition to indicating the predicted sentiment. To enhance the readability and specificity of the knowledge network, generalist terms within the domain, such as *pollen* and *allergy*, were filtered. Therefore, these terms did not contribute significantly to the comprehension of the relationships. By focusing on the most relevant and meaningful terms, the analysis was able to provide a more focused and insightful understanding of relevant health-related concepts.

Regarding the most relevant vertexes, it is evident that *cough*, *asthma*, and *hay fever* were the most prevalent terms, indicating their significance in clinical research. *Cough*, *hay fever*, *headache*, *asthma*, *sore throat*, *eye conditions*, and *sinus infections* were the common health concerns mentioned. Birch and ragweed were the plants or species with the biggest mentions in relation to the identified health-related topics. In terms of graph statistics, the clustering coefficient of the graph is relatively low, indicating that there are not many groups of nodes that are tightly connected to each other. This means that the community had a broad discussion about medical terms without focusing on a specific relationship. However, some nodes, such as *sore throat* and *sinus infection*, have a relatively high clustering coefficient, indicating that they are part of highly connected subnetworks (ie, most commonly mentioned symptoms concerning different health-related topics). On the other hand, some notable nodes, including *cough*, *hay fever*, and *headache*, have the lowest average shortest path lengths, meaning that they are connected to many other nodes and are highly central in the graph. This suggests that these symptoms were more likely to be mentioned in tweets related to pollen allergies and first-person experiences. Finally, a comparison between the hashtag and health-related mention graphs revealed that, while health-related terms associated with hashtags primarily centered on respiratory issues and eye-nose conditions (eg, *cough*, *asthma*, or *itchy eyes*), there was a notable lack of mentions related to symptoms such as *headaches*, *migraines*, or *fatigue* mentioned in the first-person term analysis graph. Such findings suggest a possible lack of community support or awareness concerning these conditions.

An analysis of the sentiment expressed in the knowledge graph revealed a predominance of negative sentiment. These data were aligned with those of another study focusing on tweets related to both *asthma* and *allergic rhinitis* that found that the average sentiment scores were predominantly negative [63].

The most frequently mentioned medical conditions were *watery eyes*, *runny nose*, *hay fever*, *headache*, and *migraine*, along with *cough* and *asthma*. However, there is also evidence of some positive sentiment relationships in the dataset, with a concentration of positive mentions around certain medications, such as *Flonase*, *Tylenol*, and *Cetirizine*, as well as improvements in symptoms related to hay *fever*, *migraine*, and *cough*.

XSL•FO
RenderX

Figure 4. Top comentioned hashtags in pollen-related discussions. Knowledge graph representing the comention hashtags from December 2006 to January 2022.

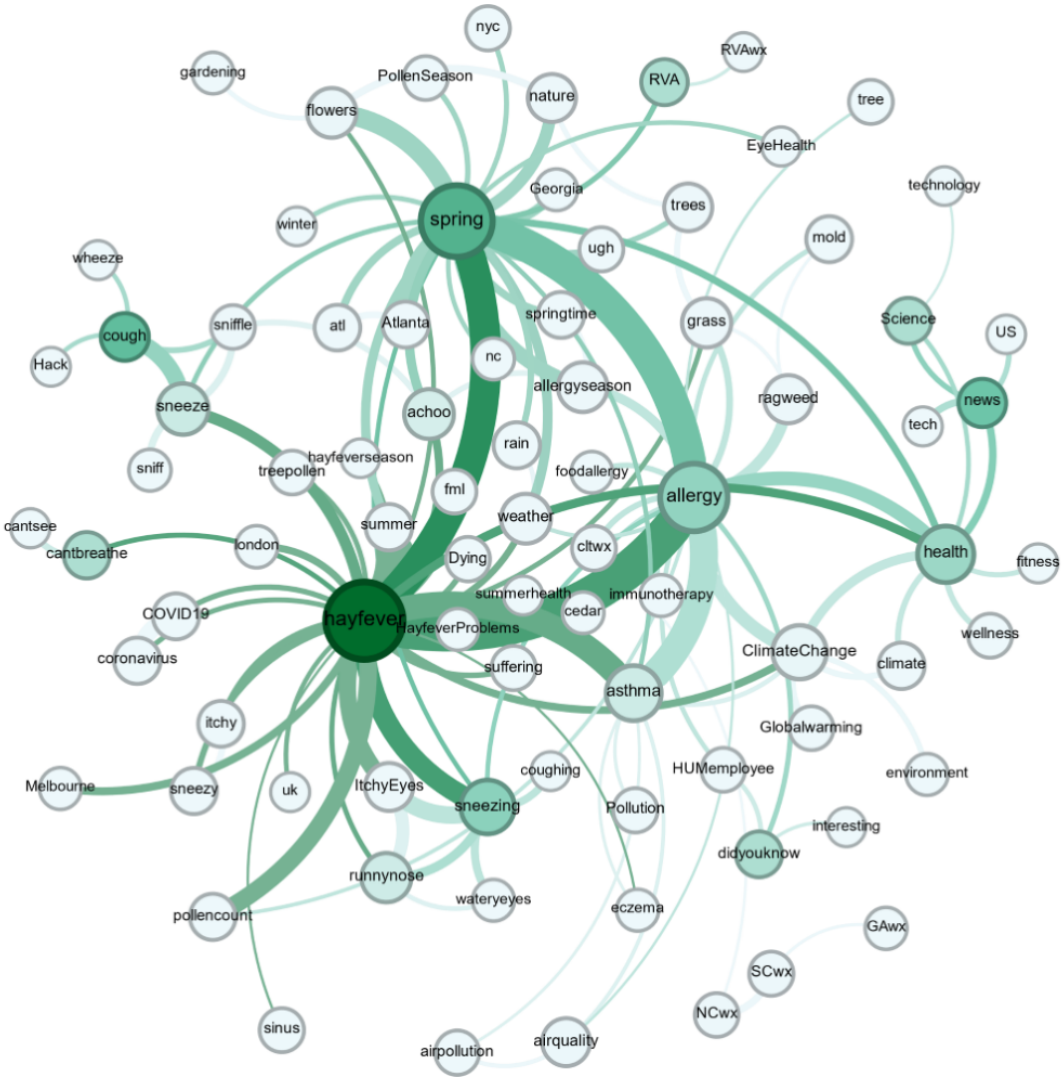

XSL•FO
RenderX

Figure 5. Top comentioned health-related topics identified in first-person messages. Knowledge graph representing the comentioned health-related topics identified from December 2006 to January 2022 in pollen-related discussions.

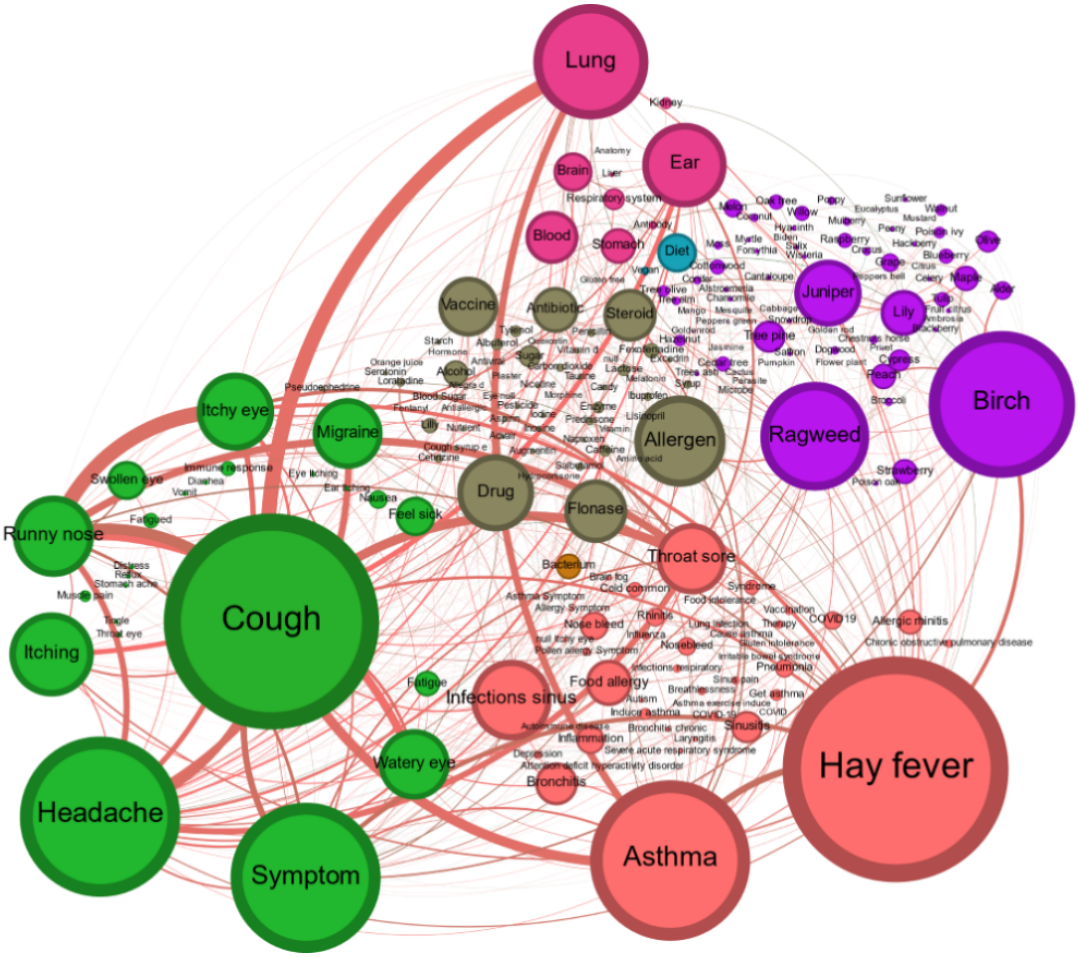

Health-Related Knowledge Analysis

To gain a deeper understanding of the medical knowledge contained in the messages and identify concepts and relationships that were not previously identified by the ontologies and NER techniques, an unsupervised analysis was conducted on all relevant messages that contained at least one medical term, identified through the proposed annotation workflow. The objective of this analysis was to identify any previously undiscovered correlations between the medical concepts mentioned in the messages, thus providing a more comprehensive understanding of the relationships between them. Table 3 represents the most discussed topics measured by their TF-IDF, and Figure 6 shows their evolution over time.

The unsupervised analysis of health-related messages reveals a wide-ranging discourse, from environmental appreciation and specific allergens (topic 0) to individual experiences with hay fever and associated symptoms (topics 2 and 3). Topic 1 revealed a comprehensive dialogue around distinct types of pollen and other allergens that can affect air quality and health. Topics such as topic 4 (allergic reactions and food) and topic 5 (asthma and respiratory issues) extended the conversation to broader allergy and respiratory conditions, suggesting a crossover between pollen allergies and other allergic reactions, including dietary concerns and asthma-related challenges. Discussions on treatment strategies (topic 6) highlighted a proactive community seeking both medical and natural remedies. Emotional responses to allergy struggles were vividly expressed (topic 7), whereas unique concerns such as pet and plant toxicity emerged (topic 8). Finally, the impact of weather, seasonal changes, and climatic change on allergies was acknowledged (topic 9), illustrating a comprehensive public understanding of the multifaceted nature of pollen allergies.

XSL•FO
RenderX

Table 3. Top 6 main topic clusters showing the different concepts discussed together in messages with identified health-related named entity recognition annotations from December 2006 to January 2022.

	TF-IDF[a]
Topic 0	
Plant	0.18
Love	0.12
Willow	0.12
Really	0.10
Willow wing	0.10
Wing trouble	0.10
Trouble	0.10
Topic 1	
Grass	0.34
Ragweed	0.22
Low	0.21
Oak	0.21
Juniper	0.19
Cedar	0.18
Birch	0.17
Topic 2	
Hay fever	0.46
Hayfever	0.45
Fever	0.42
Hay	0.42
Suffer	0.18
Hayfever suffer	0.07
Fever suffer	0.06
Topic 3	
Eye	0.43
Headache	0.25
Nose	0.22
Feel	0.19
Throat	0.14
Itch	0.13
Sinus	0.13
Topic 4	
Allergic	0.56
Food	0.09
Reaction	0.07
Never	0.06
Fruit	0.06
Body	0.05
Allergic reaction	0.04
Topic 5	

XSL•FO
RenderX

	TF-IDF[a]
Asthma	0.35
Cough	0.17
Mask	0.14
Cause	0.13
Well	0.11
Lung	0.11
Outside	0.10
Topic 6	
Help	0.18
Symptom	0.13
Allergen	0.11
Use	0.10
Work	0.10
Fever oil	0.09
Try	0.08
Topic 7	
Sneeze	0.21
Fuck	0.10
Kill	0.08
Thank	0.08
Stop	0.07
Away	0.07
Morning	0.06

[a]TF-IDF: term frequency–inverse document frequency.

XSL•FO
RenderX

Figure 6. Evolution of Twitter conversations related to "pollen" within the recognized topic, based on Twitter messages containing at least one medical term, from December 2006 to January 2022.

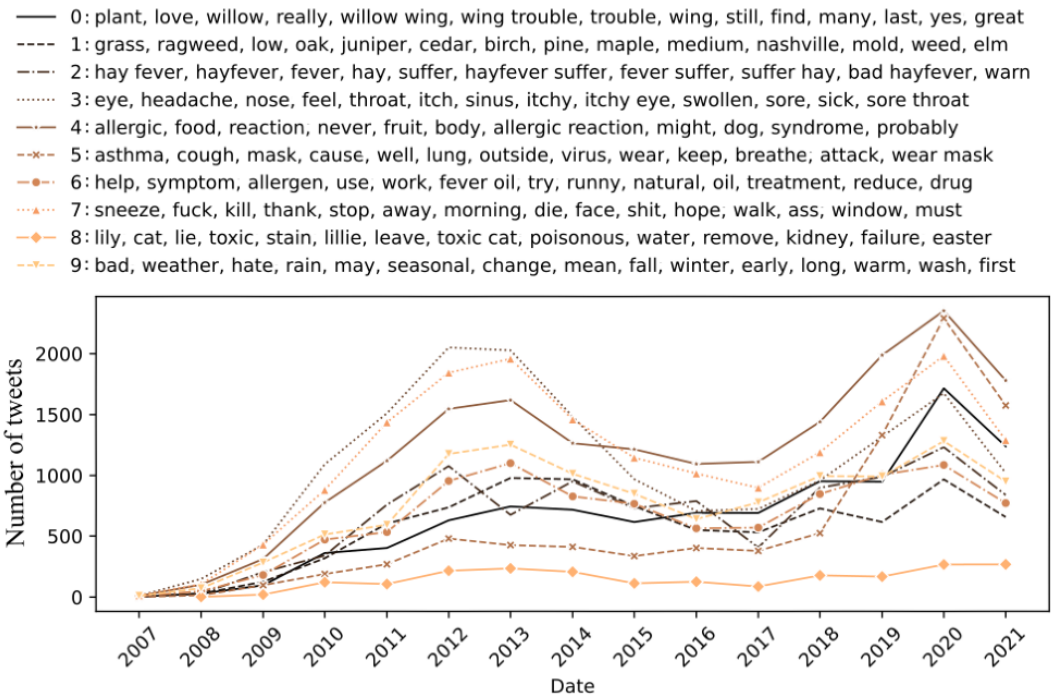

In-Depth Analysis of Previously Discovered Topics

Overview

In terms of health-related analysis, there were different key concepts discovered in previous sections (Figures 3, 4, and 5) that require further analysis to gain a deeper understanding of the correlation between these terms and the pollen domain. In this line, an expert in the field carefully selected these topics, revising the previous sections and focusing on areas that were not directly associated with respiratory issues or topics that demanded more detailed examination. For instance, pine pollen grains were traditionally considered a nonallergenic pollen, and their relevance in this work deserves further attention [71]. This expert-driven selection ensured that the analysis covered both expected and less obvious aspects related to pollen, providing a comprehensive view of the topic's impact on public health discussions. To accomplish this objective, a compilation of all relevant messages that refer to the specific terms outlined in the following sections was gathered and analyzed.

Headache and Migraine

While respiratory-related illnesses dominate health discussions, headaches and migraines were among the most commonly reported symptoms in the evaluated dataset. Therefore, given the significant impact that headaches and migraines can have on individuals' daily lives and overall well-being, further research in this area could provide new insights into strategies for prevention, management, and treatment, potentially enhancing the quality of life. Textbox 2 represents the most common unsupervised concepts related to this topic.

XSL•FO
RenderX

Textbox 2. Most common words and bigrams related to the headache and migraine discussions based on their term frequency–inverse document frequency.

Top 10 relevant words

- Giving
- Sinus
- Hate
- Bad
- Nose
- Worst
- Outside
- Sneezing
- Migraines
- Away

Top 10 relevant bigrams

- Eye-nose
- Migraine-sinus
- Sinus-wake
- Hate-sinus
- Rain-wash
- Nose-throat
- Sinus-thank
- Eye-sinus
- Morning-wake
- Cause-sinus

The comentioned words in the headache and migraine discussions reveal a clear comorbidity of the reported illness with sinus and sneezing, suggesting that these symptoms may be related to or exacerbate the patient's condition. These findings are comparable with those of another study that investigated the relationship between allergic rhinitis and migraines, implying a probable association [72]. Nevertheless, more recent studies suggest that allergens and headaches or migraines may not be related [73], emphasizing the need for greater research to fully understand any potential connections. In addition, there appeared to be a seasonal pattern in the reported symptoms, with a higher frequency of mentions referring to a specific time of the day. Notably, the data suggest that symptoms may be more concentrated in the morning, as evidenced by the prevalence of terms such as *morning*, *wake*, and *sleep*. This result aligns with the observations of a study conducted in the northeastern United States between May 2020 and June 2020, which identified that terms such as *morning* and *sleep* were predominantly used to categorize clusters of conversations using hierarchical clustering of subgraphs [73].

Figure 7 illustrates the changes in the frequency of mentions of migraines and headaches over time, with a focus on their stationarity. The top chart presents the number of messages mentioning headache and migraine obtained from the Twitter application programming interface. The bottom chart presents the number of messages mentioning both symptoms and pollen. The green line represents a polynomic tendency line of grade 2. In this sense, the data show a distinct pattern of stationarity between March and September, with the largest number of mentions (considering the total number of messages) recorded in 2012 and 2013. However, in terms of discussion volume, the last years showed that there was a growing tendency to discuss these conditions, indicating a possible increasing significance and impact on people's lives. Overall, these findings suggest the need for further research to better understand the comorbidity of headaches and migraines with sinus and sneezing symptoms, as well as the potential influence of time of day on the occurrence of these conditions.

XSL•FO
RenderX

Figure 7. Evolution of headache and migraine mentions on Twitter (top) and in pollen-related messages (bottom).

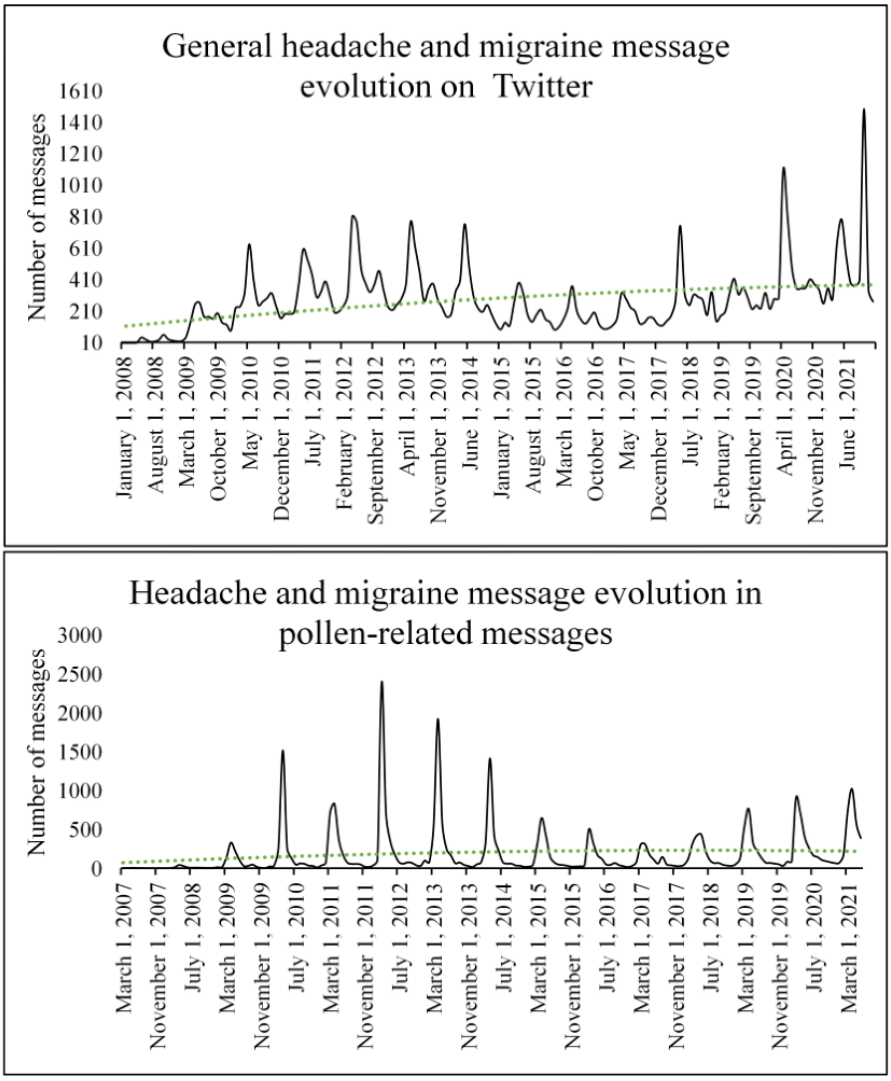

Pine

According to the findings, the plants that were mentioned most frequently in the community were ragweed, oak, cedar, pine, and birch. In this sense, one of the most mentioned plants was pine even though allergies to pine pollen are extremely rare [74]. To further explore the public discourse regarding pine, a topic analysis was conducted to analyze the public discussion along this line. Textbox 3 presents the most mentioned topics concerning pine pollen. Notably, a considerable number of mentions on this topic revolved around the belief that natural compounds contained in pollen have the potential to boost testosterone levels and, as a result, enhance physical performance. This concept sparked debates regarding the use of pine pollen as a natural supplement aimed at boosting testosterone-related attributes.

XSL•FO
RenderX

Textbox 3. Most common words and bigrams related to pine based on their term frequency–inverse document frequency.

Top 10 relevant words

- Hormones
- Testosterone
- Pine
- Natural
- Levels
- Bee
- Sex
- Boosting
- Allergies
- Balance
- Increase

Top 10 relevant bigrams

- Feel-hormone
- Hormone-mix
- Count-hormone
- Hormone-need
- Bee-hormone
- Cry-hormone
- Hormone-stress
- Hormone-woman
- Hormone-plant
- Headache-hormone
- Blame-hormone

Asthma

To enhance our comprehension of the primary topics connected to asthma and associated concerns, Textbox 4 illustrates the terms most frequently mentioned alongside this health condition. The analysis identified *pregnancy*, *mask*, and *climate change* as the most prevalent subjects in asthma-related discussions. This finding aligns closely with the observations presented in Figure 6 regarding female individuals. Specifically, the prominent frequency of discussions related to pregnancy and asthma within the community indicates a strong interest among pregnant individuals in obtaining information and support for managing asthma during pregnancy, with the goal of minimizing potential risks to their unborn children. This underlines the critical need for making reliable, easily accessible information available to those living with asthma and underscores the necessity for heightened awareness and education on managing asthma during pregnancy among health care providers and the wider community.

On the other hand, the discussion of climate change highlights the increasing concern about the impact of environmental factors on respiratory health. Climate change can exacerbate asthma by increasing air pollution, triggering more frequent and severe allergic reactions and leading to extreme weather events that can cause respiratory distress and, therefore, increased pollen levels [75]. This finding highlights the heightened sensitivity of the pollen community to the effects of environmental factors on respiratory health.

XSL•FO
RenderX

Textbox 4. Most common words and bigrams related to asthma based on their term frequency–inverse document frequency.

Top 10 relevant words

- Attack
- Triger
- Cause
- Suffer
- Breathe
- Grass
- Mask
- Help
- Feel
- Fever
- Weather

Top 10 relevant bigrams

- Fever-hay
- Change-climate
- Mask-wear
- Change-link
- Increase-risk
- Change-weather
- Kid-trigger
- Exposure-pregnancy
- Attack-trigger
- Child-risk
- Fever-hay

Tired, Fatigue, Kill, and Sleep

The analysis of the terms *tired*, *fatigue*, and *sleep* in relation to the main topic of pollen allergies revealed interesting insights (Textbox 5). These symptoms were highly associated with watery eyes and runny nose, indicating that this symptoms have a significant impact on sleep quality, resulting in tiredness and fatigue during the day. These results are in accordance with the findings of a study conducted in South Korea, in which the authors pointed out that individuals with allergic rhinitis had significantly greater sleep disorders, fatigue, and depressive symptoms than those without allergic rhinitis, so allergic rhinitis was an important factor influencing sleep disorders and fatigue [76]. Moreover, the analysis carried out highlight a possible association between pollen-related health conditions and work performance, suggesting that pollen allergies can negatively impact work productivity. However, a recent meta-study indicates that the relationship between allergic symptoms and sleep problems or daytime fatigue ranges from low to very low, suggesting caution when interpreting these results [77]. This implies that the findings of this study may be more reflective of patient perceptions than of confirmed interrelated symptoms. In addition, the analysis also indicates that the terms *need-rain* are highly associated with the main topic, suggesting that weather patterns can play a crucial role in mitigating pollen allergy symptoms.

XSL•FO
RenderX

Textbox 5. Most common words and bigrams related to tired, fatigue, and sleep based on their term frequency–inverse document frequency.

Top 10 relevant words

- Eye
- Feel
- Rain
- Fuck
- Hate
- Sneeze
- Weather
- Night
- Nose
- Sinus
- Work

Top 10 relevant bigrams

- Fever-hay
- Need-rain
- Rain-wash
- Open-window
- Eye-nose
- Eye-itch
- Mask-weather
- Burn-eye
- Cut-grass
- Love-weather
- Feel-head

Sinus Conditions

To gain a better understanding of the principal key topics related to sinus conditions, Textbox 6 represents the most comentioned terms related to this health condition. Regarding the mentioned words, headache and infection were some of the most mentioned cosymptomatology. Furthermore, the significance of discussions centered around mornings and the related health conditions, such as ear infections, should also be emphasized.

XSL•FO
RenderX

Textbox 6. Most common words and bigrams related to sinus based on their term frequency–inverse document frequency.

Top 10 relevant words

- Fuck
- Headache
- Infection
- Hate
- Feel
- Kill
- Hate
- Snort
- Eye
- Weather
- Rain

Top 10 relevant bigrams

- Fever-hay
- Rain-wash
- Mask-wear
- Eye-nose
- Change-weather
- Ass-kick
- Ear-infection
- Need-rain
- Feel-hope
- Feel-head
- Kill-try

Shared Resource Analysis

An analysis of the most shared resources (ie, URLs), grouped by topic using the unsupervised workflow, was carried out to reveal valuable insights into the most relevant and popular shared resources on the Twitter platform. Web-based data were scraped and grouped by topic considering their website descriptions, titles, and contents. It should be noted that not all websites could be obtained as some of the resources may no longer exist at present. This information can be used to better understand the needs and interests of the community and tailor communication and content strategies accordingly. In short, analyzing shared URLs is a powerful tool for gaining knowledge about the most important and interesting topics in the community that endure over time. In addition, by analyzing popular websites, it is possible to identify key influencers and thought leaders across various domains, facilitating targeted networking and outreach strategies. Insights derived from these patterns are invaluable for understanding community interests and trends, thereby enhancing strategic decision-making and communication initiatives.

Figure 8 shows the top shared resources by the community as well as unsupervised classified topics sorted by their share density (ie, the number of users mentioning the URL). Each bubble represents a domain, with colors indicating the topic and bubble sizes reflecting popularity based on user mentions. Major shared topics include news, research, pollen studies, symptoms, weather, and treatments, highlighting community interests in health and medical updates. As can be observed, the most valuable resources for the community were primarily related to news portals, research information, and new related pollen research projects. In addition, resources related to symptoms, weather, and treatments were highly valued by the community. This suggests that the community is highly interested in staying informed about the latest research and developments in the field of health and medicine. Furthermore, the significance of resources related to symptoms, weather, and treatments suggests that the community is striving to gain a better understanding of their health conditions and find effective ways to manage their symptoms, as well as be prepared for weather and climate conditions. Overall, these findings provide insights into the information needs of the community and can be used to inform the development of targeted resources and interventions to better support and empower individuals with health-related concerns.

XSL•FO
RenderX

Figure 8. Top shared resources by the community. This figure shows the most shared URLs on Twitter grouped by topic using website embeddings for clustering.

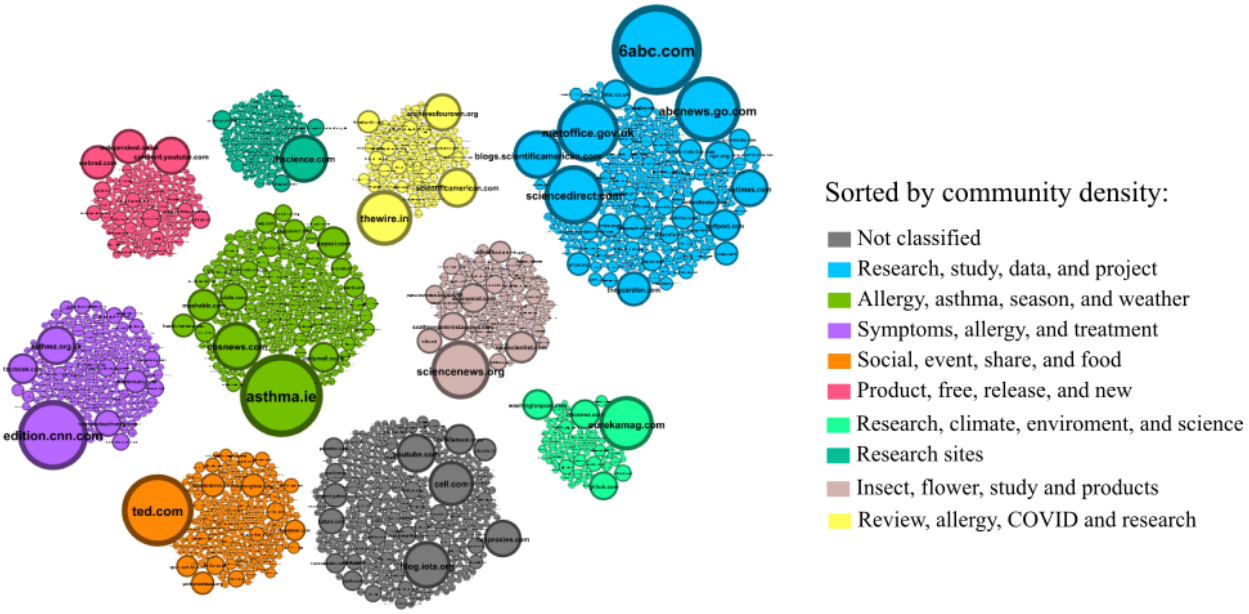

Discussion

Principal Findings

This study used different NLP and text mining techniques to analyze 4.6 million tweets (from December 2006 to January 2022) about pollen allergy symptomatology and its influence on patients with allergy. The study found 1.39 million relevant tweets, with most messages posted between March and August every year. In terms of message evolution, there was a significant rise in patient tweets concerning pollen-related health issues. Unsupervised analysis revealed a wide range of subjects, including environmental variables, personal hay fever experiences, allergic responses, asthma, and medicine, illustrating the complexities of pollen allergies. The presence of emphatic language in tweets reflects the frustration and discomfort experienced by those who have allergies. The association of pollen with weather-related terms such as *rain*, *wind*, and *mold* suggests that public awareness and concerns about pollen allergies are intricately linked to environmental conditions. This finding is consistent with previous studies that measured the frequency of weather-related words in tweets discussing *allergy*, *asthma*, and *rhinitis*. Furthermore, associations were found between medical concepts such as the comorbidity of headaches and migraines with nasal symptoms and sneezing, as well as a relevant debate concerning sleep problems and fatigue-related symptoms associated with pollen allergy symptoms. On the other hand, misinformation was found about the use of pine pollen as a stimulant for the generation of testosterone.

Strengths and Limitations

While this study contributes to the understanding of pollen-related discourse on social media platforms, several limitations and strengths should be noted. First, the data used in this analysis were sourced exclusively from a publicly available social media platform, which may not fully capture the breadth and depth of discussions on the topic. In addition, inherent biases within social media data, such as fluctuations in user engagement and platform popularity, may influence the findings. However, the use of publicly available data on Twitter allows for a large and diverse study, providing a broad overview of public sentiment and awareness. In this way, Twitter has become a highly valued source of knowledge in the scientific community [78]. A significant strength of this work lies in the use of public and distended Twitter data. The voluntary nature of the data used in this work also means that it reflects genuine and spontaneous patient experiences and concerns, making it a rich source of information.

On the other hand, the generalizability of the findings may be constrained by the language and geographical biases inherent in social media data and introduces the possibility of an impact on the interpretation of social media discourse on pollen. However, this inherent bias is a widespread challenge that highlights the multilayered complexities involved in interpreting social media data [79,80]. Conversely, the primary advantage of this work lies in the extensive reach of social media and data volume, allowing for the collection and analysis of a vast amount of social information from a diverse and widespread user base. This comprehensive dataset provides a more globalized perspective, enabling the identification of aggregate patterns and trends that might be overlooked in studies with smaller and regionalized sample sizes. This inclusivity enhances the reliability and relevance of the findings, making them more applicable to different demographic and geographic contexts.

Finally, the similarity between allergic symptoms and those of some viral infections introduces an inherent bias in analyzing social media discussions about pollen. Despite focusing the data query specifically on messages that contained the word *pollen*, reducing the likelihood of capturing discussions related to unrelated viral illnesses or common colds, it is not possible to eliminate this bias. The difficulty that individuals face in distinguishing between allergic and viral symptoms, which is well documented in the literature, must be acknowledged when

XSL•FO
RenderX

interpreting the findings of this study [81]. Nevertheless, a significant strength of this work lies in its ability to focus the results only on those individuals who offer information about their health or perceptions concerning the search query of this work (*pollen*), which may minimize the incorporation of discussions that deal exclusively with viral infections. In addition, the use of sophisticated analytical tools in this work, such as modern LLMs, enhances the reliability of the analysis and emphasizes the robustness of our methodology. These models enable precise differentiation of the discussed medical concepts based on their context. LLMs automatically distinguish between discussions of allergic and viral symptoms in contexts where these symptoms are explicitly mentioned.

Comparison With Prior Work

Health-related social network data have been used for large-scale analysis of issues related to *pollen* and other medical ailments, known as *infovigilance* [82]. Nonetheless, this work conducts a more globalized analysis of pollen and pollen allergy than other regional research, concentrating the analysis on more aggregated data rather than on targeted discussions [83-85]. This comprehensive analysis facilitates the identification of common themes and the discovery of patterns among patients from various regions worldwide. The findings of this study align with those of earlier studies showing the seasonality of social media activity related to health and the influence of environmental factors on respiratory disorders. Similar to earlier findings, the frequent correlation between tweets about pollen allergies and phrases such as *weather*, *wind*, *rain*, and *climate change* demonstrates that public concerns are related to climatic changes [83]. On the other hand, in 2020, the evaluated data revealed a peak in *pollen* tweets, which coincided with a surge in Google Trends related to media coverage of asthma [86]. In terms of seasonality, our data on pollen discussions fluctuated over the years, with a visible decrease between 2013 and 2018. This information does not fully align with reports showing a rising prevalence of allergies in recent years [87] and those indicating that an increase in carbon dioxide concentrations and atmospheric temperatures elevate pollen levels, causing high pollen counts to start earlier and last longer [88]. However, individual perceptions of allergic symptoms may be influenced by emotional and external factors, such as community attitudes and general perceptions. Therefore, some studies suggest that the impact of allergies, personal experiences, and perceptions can be significantly affected by emotional states [89,90].

Conclusions

The primary value of tweets is the insights that they provide into human behavior and thought processes; people share their genuine, spontaneous experiences related to pollen, providing information that would otherwise be difficult to obtain. In this sense, this study provides a detailed understanding from various perspectives of how pollen allergies are discussed on Twitter. In addition, it explores how incorporating modern LLMs can enhance the analysis and understanding of the public discussion. These advanced models can analyze vast amounts of text data, identifying patterns and nuances that traditional methods might overlook. The scientific understanding of pollen and allergies can greatly benefit from alternative data sources, including social media, to improve and validate related scientific knowledge. By analyzing the causal relationships among pollen counts, tweet numbers, and semantic patterns related to pollen-related health issues, valuable insights can be gained into how patients interact with pollen information. These insights can assist health stakeholders, organizations, and governments in monitoring public engagement, guiding decision-making, and designing targeted communication strategies. This approach may prove valuable in adopting preventive measures to mitigate the impact of pollen allergies on sensitive patients and improving access to health information. Leveraging social media data for public health surveillance and intervention has the potential to advance the development of new treatments and improve the overall management of pollen allergies. This study is expected to help public health officials, researchers, and medical professionals better understand the pollen-related community and use Twitter more efficiently for various objectives, such as tracking symptoms, disseminating information, and adopting preventive measures.

Acknowledgments

The Next Generation Computer Systems Group (SING) group thanks the Centro de Investigación, Transferencia e Innovación (Research, Transfer and Innovation Center-CITI) from the University of Vigo for hosting its IT infrastructure. This work was supported by the Associated Laboratory for Green Chemistry, financed by the Portuguese Foundation for Science and Technology (grant UIDB/50006/2020); the Consellería de Cultura, Educación e Universidade (Department of Culture, Education and University-Xunta de Galicia) under the scope of the strategic funding of Competitive Reference Groups (grants ED431C 2022/03-GRC and ED431C 2017/62 BV1); and the Centro Singular de Investigación de Galicia (Singular Research Center of Galicia) accreditation 2019-2022, funded by the European Regional Development Fund (grant ED431G2019/06). The authors also acknowledge the postdoctoral fellowship of MP-P, funded by the Xunta de Galicia (grant ED481D-2023-003). This research also received funding from the project control aerobiológico de Galicia (Aerobiological Control of Galicia 2021–2024) and CO-0034-2021 00VT of the Department of Health (Xunta de Galicia).

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Sánchez-Borges M, Martin BL, Muraro AM, Wood RA, Agache IO, Ansotegui IJ, et al. The importance of allergic disease in public health: an iCAALL statement. World Allergy Organ J 2018;11(1):8 [FREE Full text] [doi: 10.1186/s40413-018-0187-2] [Medline: 29743965]
2. Bousquet J, Dahl R, Khaltaev N. Global alliance against chronic respiratory diseases. Allergy 2007 Mar 08;62(3):216-223 [FREE Full text] [doi: 10.1111/j.1398-9995.2007.01307.x] [Medline: 17298337]
3. Calderon MA, Demoly P, Gerth van Wijk R, Bousquet J, Sheikh A, Frew A, et al. EAACI: a European declaration on immunotherapy. Designing the future of allergen specific immunotherapy. Clin Transl Allergy 2012 Oct 30;2(1):20 [FREE Full text] [doi: 10.1186/2045-7022-2-20] [Medline: 23110958]
4. Lake IR, Jones NR, Agnew M, Goodess CM, Giorgi F, Hamaoui-Laguel L, et al. Climate change and future pollen allergy in Europe. Environ Health Perspect 2017 Mar;125(3):385-391. [doi: 10.1289/ehp173]
5. Vara A, Fernández-González M, Aira M, Rodríguez-Rajo FJ. Oleaceae cross-reactions as potential pollinosis cause in urban areas. Sci Total Environ 2016 Jan 15;542(Pt A):435-440. [doi: 10.1016/j.scitotenv.2015.10.107] [Medline: 26520268]
6. Moreno-Grau S, Aira MJ, Elvira-Rendueles B, Fernández-González M, Fernández-González D, García-Sánchez A, et al. Assessment of the Olea pollen and its major allergen Ole e 1 concentrations in the bioearosol of two biogeographical areas. Atmos Environ 2016 Nov;145:264-271. [doi: 10.1016/j.atmosenv.2016.09.040]
7. Pawankar R. Allergic diseases and asthma: a global public health concern and a call to action. World Allergy Organ J 2014;7(1):12 [FREE Full text] [doi: 10.1186/1939-4551-7-12] [Medline: 24940476]
8. Platts-Mills TA. The allergy epidemics: 1870-2010. J Allergy Clin Immunol 2015 Jul;136(1):3-13 [FREE Full text] [doi: 10.1016/j.jaci.2015.03.048] [Medline: 26145982]
9. Prescott SL, Pawankar R, Allen KJ, Campbell DE, Sinn JK, Fiocchi A, et al. A global survey of changing patterns of food allergy burden in children. World Allergy Organ J 2013 Dec 04;6(1):21 [FREE Full text] [doi: 10.1186/1939-4551-6-21] [Medline: 24304599]
10. D'Amato G, Vitale C, Sanduzzi A, Molino A, Vatrella A, D'Amato M. Allergenic pollen and pollen allergy in Europe. In: Singh AB, editor. Allergy and Allergen Immunotherapy New Mechanisms and Strategies. New York, NY: Apple Academic Press; 2017:287-306.
11. Riveras CM, Luque SS, Verdes TQ, Lema ME, García LA. Impacto de los niveles de polen en la utilización de los servicios sanitarios en Santiago de Compostela. Rev Esp Salud Publica 2008;12(2):110-121 [FREE Full text]
12. Zuberbier T, Lötvall J, Simoens S, Subramanian SV, Church MK. Economic burden of inadequate management of allergic diseases in the European Union: a GA(2) LEN review. Allergy 2014 Oct;69(10):1275-1279. [doi: 10.1111/all.12470] [Medline: 24965386]
13. Bergmann KC, Berger M, Klimek L, Pfaar O, Werchan B, Werchan M, et al. Nonpharmacological measures to prevent allergic symptoms in pollen allergy: a critical review. Allergol Select 2021;5:349-360 [FREE Full text] [doi: 10.5414/ALX02294E] [Medline: 34870079]
14. Rodríguez-Rajo FJ, Jato V, González-Parrado Z, Elvira-Rendueles B, Moreno-Grau S, Vega-Maray A, et al. The combination of airborne pollen and allergen quantification to reliably assess the real pollinosis risk in different bioclimatic areas. Aerobiologia 2010 May 27;27(1):1-12. [doi: 10.1007/S10453-010-9170-2]
15. Martínez-Martínez F, Roldán-Álvarez D, Martín E, Hoppe HU. An analytics approach to health and healthcare in citizen science communications on Twitter. Digit Health 2023 Jan 05;9:205520762211453. [doi: 10.1177/20552076221145349]
16. Salzmann-Erikson M. Digital inclusion: a mixed-method study of user behavior and content on Twitter. Digit Health 2023 Nov 01;9:20552076231211277 [FREE Full text] [doi: 10.1177/20552076231211277] [Medline: 37928325]
17. Miyake E, Martin S. Long Covid: online patient narratives, public health communication and vaccine hesitancy. Digit Health 2021 Nov 29;7:20552076211059649 [FREE Full text] [doi: 10.1177/20552076211059649] [Medline: 34868622]
18. Ellison NB, Boyd DM. Sociality through social network sites. In: Dutton WH, editor. The Oxford Handbook of Internet Studies. Oxford, UK: Oxford University Press; 2013:1518-1572.
19. Pfaar O, Bastl K, Berger U, Buters J, Calderon MA, Clot B, et al. Defining pollen exposure times for clinical trials of allergen immunotherapy for pollen-induced rhinoconjunctivitis - an EAACI position paper. Allergy 2017 May 27;72(5):713-722. [doi: 10.1111/all.13092] [Medline: 27874202]
20. Hawn C. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. Health Aff (Millwood) 2009 Mar;28(2):361-368 [FREE Full text] [doi: 10.1377/hlthaff.28.2.361] [Medline: 19275991]
21. Mazumdar S, Thakker D. Citizen science on Twitter: using data analytics to understand conversations and networks. Future Internet 2020 Nov 26;12(12):210. [doi: 10.3390/fi12120210]
22. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control 2010 Apr;38(3):182-188 [FREE Full text] [doi: 10.1016/j.ajic.2009.11.004] [Medline: 20347636]
23. Shahi GK, Dirkson A, Majchrzak TA. An exploratory study of COVID-19 misinformation on Twitter. Online Soc Netw Media 2021 Mar;22:100104 [FREE Full text] [doi: 10.1016/j.osnem.2020.100104] [Medline: 33623836]
24. Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: Proceedings of the 43rd Hawaii International Conference on System Sciences. 2010 Presented at: HICSS '10; January 5-8, 2010; Honolulu, HI p. 1-10 URL: https://ieeexplore.ieee.org/document/5428313 [doi: 10.1109/hicss.2010.412]

XSL•FO
RenderX

25. Pershad Y, Hangge P, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018 May 28;7(6):121 [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
26. Suratnoaji C, Nurhadi, Arianto IW. Public opinion on lockdown (PSBB) policy in overcoming COVID-19 pandemic in Indonesia: analysis based on big data Twitter. Asian J Public Opin Res 2020;8(3):393-406 [FREE Full text] [doi: 10.15206/ajpor.2020.8.3.393]
27. Bacsu JD, Fraser S, Chasteen AL, Cammer A, Grewal KS, Bechard LE, et al. Using Twitter to examine stigma against people with dementia during COVID-19: infodemiology study. JMIR Aging 2022 Mar 31;5(1):e35677 [FREE Full text] [doi: 10.2196/35677] [Medline: 35290197]
28. Yujian L, Bo L. A normalized levenshtein distance metric. IEEE Trans Pattern Anal Machine Intell 2007 Jun;29(6):1091-1095. [doi: 10.1109/tpami.2007.1078]
29. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL HLT '19; June 6-7, 2019; Minneapolis, MN p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf
30. Giovanni MD, Brambilla M. Exploiting Twitter as source of large corpora of weakly similar pairs for semantic sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language. 2021 Presented at: EMNLP '21; November 7-11, 2021; Virtual Event p. 9902-9910 URL: https://aclanthology.org/2021.emnlp-main.780.pdf [doi: 10.18653/v1/2021.emnlp-main.780]
31. Ben Salem K, Ben Abdelaziz A. Principal component analysis (PCA). Tunis Med 2021;99(4):383-389 [FREE Full text] [Medline: 35244921]
32. Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform 2015;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S3 [FREE Full text] [doi: 10.1186/1758-2946-7-S1-S3] [Medline: 25810774]
33. Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics 2010 Feb 11;11(1):85 [FREE Full text] [doi: 10.1186/1471-2105-11-85] [Medline: 20149233]
34. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013 Nov 15;29(22):2909-2917 [FREE Full text] [doi: 10.1093/bioinformatics/btt474] [Medline: 23969135]
35. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005 Jul 15;21(14):3191-3192. [doi: 10.1093/bioinformatics/bti475] [Medline: 15860559]
36. Jessop DM, Adams SE, Willighagen EL, Hawizy L, Murray-Rust P. OSCAR4: a flexible architecture for chemical text-mining. J Cheminform 2011 Oct 14;3(1):41 [FREE Full text] [doi: 10.1186/1758-2946-3-41] [Medline: 21999457]
37. Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. NPJ Sci Food 2018 Dec 18;2(1):23 [FREE Full text] [doi: 10.1038/s41538-018-0032-6] [Medline: 31304272]
38. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thésaurus and Ontology. J Web Semant 2003 Dec;1(1):75-80. [doi: 10.1016/j.websem.2003.07.007]
39. Schriml LM. Symptom ontology. OBO Technical WG. 2018. URL: https://obofoundry.org/ontology/symp.html [accessed 2019-12-11]
40. Rosse C, Mejino Jr JL. The foundational model of anatomy ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy Ontologies for Bioinformatics: Principles and Practice. London, UK: Springer; 2008:59-117.
41. Welcome to medical subject headings. National Library of Medicine. URL: https://tinyurl.com/33fs5yaf [accessed 2023-02-08]
42. de Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, Haug K, et al. Chemical entities of biological interest: an update. Nucleic Acids Res 2010 Jan;38(Database issue):D249-D254 [FREE Full text] [doi: 10.1093/nar/gkp886] [Medline: 19854951]
43. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 2015 Jan;43(Database issue):D1071-D1078 [FREE Full text] [doi: 10.1093/nar/gku1011] [Medline: 25348409]
44. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an experimental factor ontology. Bioinformatics 2010 Apr 15;26(8):1112-1118 [FREE Full text] [doi: 10.1093/bioinformatics/btq099] [Medline: 20200009]
45. Almeida MS, Sousa Filho LF, Rabello PM, Santiago BM. International classification of diseases - 11th revision: from design to implementation. Rev Saude Publica 2020 Dec 14;54:104 [FREE Full text] [doi: 10.11606/s1518-8787.2020054002120] [Medline: 33175024]
46. Interlinking ontology for biological concepts. BioPortal. URL: https://tinyurl.com/yzksajv3 [accessed 2023-02-08]
47. Human phenotype ontology. OBO Foundry. URL: https://obofoundry.org/ontology/hp.html [accessed 2023-02-08]
48. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018 Jan 04;46(D1):D1074-D1082 [FREE Full text] [doi: 10.1093/nar/gkx1037] [Medline: 29126136]

XSL•FO
RenderX

49. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017 Jan 04;45(D1):D353-D361 [FREE Full text] [doi: 10.1093/nar/gkw1092] [Medline: 27899662]
50. Pérez-Pérez M, Pérez-Rodríguez G, Fdez-Riverola F, Lourenço A. Using Twitter to understand the human bowel disease community: exploratory analysis of key topics. J Med Internet Res 2019 Aug 15;21(8):e12610 [FREE Full text] [doi: 10.2196/12610] [Medline: 31411142]
51. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Weblogs Soc Media 2014 May 16;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]
52. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care Tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
53. Barbieri F, Camacho-Collados J, Neves L, Espinosa-Anke L. TweetEval: unified benchmark and comparative evaluation for tweet classification. arXiv Preprint posted online October 23, 2020 [FREE Full text] [doi: 10.18653/v1/2020.findings-emnlp.148]
54. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Proc Int AAAI Conf Weblogs Soc Media 2009;3(1):361-362 [FREE Full text] [doi: 10.1136/qshc.2004.010033]
55. Collins CR, Stephenson K. A circle packing algorithm. Comput Geom 2003 Jul;25(3):233-256. [doi: 10.1016/S0925-7721(02)00099-8]
56. Freeman LC, Roeder D, Mulholland RR. Centrality in social networks: ii. experimental results. Soc Netw 1979 Jan;2(2):119-141. [doi: 10.1016/0378-8733(79)90002-9]
57. Freeman LC. Centrality in social networks conceptual clarification. Soc Netw 1978 Jan;1(3):215-239. [doi: 10.1016/0378-8733(78)90021-7]
58. Freeman LC. A set of measures of centrality based on betweenness. Sociometry 1977 Mar;40(1):35. [doi: 10.2307/3033543]
59. Nieminen J. On the centrality in a graph. Scand J Psychol 1974 May;15(4):332-336. [doi: 10.1111/j.1467-9450.1974.tb00598.x] [Medline: 4453827]
60. Bolland JM. Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. Soc Netw 1988 Sep;10(3):233-253. [doi: 10.1016/0378-8733(88)90014-7]
61. Pérez-Pérez M, Igrejas G, Fdez-Riverola F, Lourenço A. A framework to extract biomedical knowledge from gluten-related tweets: the case of dietary concerns in digital era. Artif Intell Med 2021 Aug;118:102131 [FREE Full text] [doi: 10.1016/j.artmed.2021.102131] [Medline: 34412847]
62. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis. JAMIA Open 2022 Jul;5(2):ooac028 [FREE Full text] [doi: 10.1093/jamiaopen/ooac028] [Medline: 35495736]
63. Sousa-Pinto B, Jankin S, Vieira RJ, Marques-Cruz M, Fonseca JA, Bousquet J. English tweets on allergy: content analysis and association with surveillance data. Clin Exp Allergy 2024 Jul 03;54(7):500-508. [doi: 10.1111/cea.14479] [Medline: 38567657]
64. Abascal-Mena R, Lema R, Sèdes F. From tweet to graph: social network analysis for semantic information extraction. In: Proceedings of the 2014 IEEE Eighth International Conference on Research Challenges in Information Science. 2014 Presented at: RCIS '14; May 28-30, 2014; Marrakech, Morocco p. 1-10 URL: https://ieeexplore.ieee.org/document/6861047 [doi: 10.1109/rcis.2014.6861047]
65. Radicioni T, Saracco F, Pavan E, Squartini T. Analysing Twitter semantic networks: the case of 2018 Italian elections. Sci Rep 2021 Jun 24;11(1):13207 [FREE Full text] [doi: 10.1038/s41598-021-92337-2] [Medline: 34168169]
66. Kim EH, Jeong YK, Kim Y, Kang KY, Song M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. J Inf Sci 2016 Jul 11;42(6):763-781. [doi: 10.1177/0165551515608733]
67. Bousquet J, Devillier P, Anto JM, Bewick M, Haahtela T, Arnavielhe S, MACVIA working group. Daily allergic multimorbidity in rhinitis using mobile technology: a novel concept of the MASK study. Allergy 2018 Aug 24;73(8):1622-1631 [FREE Full text] [doi: 10.1111/all.13448] [Medline: 29569295]
68. Fries J, Wu S, Ratner A, Ré C. SwellShark: a generative model for biomedical named entity recognition without labeled data. arXiv Preprint posted online April 20, 2017 [FREE Full text]
69. Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inform 2013 Dec;46(6):1088-1098 [FREE Full text] [doi: 10.1016/j.jbi.2013.08.004] [Medline: 23954592]
70. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: challenges and opportunities. In: Proceedings of the 2016 IEEE International Conference on Big Data. 2016 Presented at: Big Data '16; December 5-8, 2016; Washington, DC p. 1937-1945 URL: https://ieeexplore.ieee.org/document/7840814 [doi: 10.1109/bigdata.2016.7840814]
71. Gastaminza G, Lombardero M, Bernaola G, Antepara I, Muñoz D, Gamboa PM, et al. Allergenicity and cross-reactivity of pine pollen. Clin Exp Allergy 2009 Sep 05;39(9):1438-1446. [doi: 10.1111/j.1365-2222.2009.03308.x] [Medline: 19573163]
72. Saberi A, Nemati S, Shakib RJ, Kazemnejad E, Maleki M. Association between allergic rhinitis and migraine. J Res Med Sci 2012 Jun;17(6):508-512 [FREE Full text] [Medline: 23626624]

XSL•FO
RenderX

73. Zhang P, Bhaskarabhatla S. How advocacy affects Twitter migraine conversations: a pilot cross-sectional survey of Northeast American "migraine" landscape on Twitter from May to June 2020. Cephalalgia Rep 2020 Nov 11;3:251581632097208. [doi: 10.1177/2515816320972085]
74. Behrendt H, Becker WM. Localization, release and bioavailability of pollen allergens: the influence of environmental factors. Curr Opin Immunol 2001 Dec;13(6):709-715. [doi: 10.1016/s0952-7915(01)00283-7] [Medline: 11677094]
75. D'Amato G, Chong-Neto HJ, Monge Ortega OP, Vitale C, Ansotegui I, Rosario N, et al. The effects of climate change on respiratory allergy and asthma induced by pollen and mold allergens. Allergy 2020 Sep 05;75(9):2219-2228. [doi: 10.1111/all.14476] [Medline: 32589303]
76. Kim O, Kim B, Jeong H, Lee J, Jung H. Sleep, fatigue, and depressive symptoms among female nurses with allergic rhinitis. Healthcare (Basel) 2021 Oct 05;9(10):1328 [FREE Full text] [doi: 10.3390/healthcare9101328] [Medline: 34683008]
77. Liu J, Zhang X, Zhao Y, Wang Y. The association between allergic rhinitis and sleep: a systematic review and meta-analysis of observational studies. PLoS One 2020 Feb 13;15(2):e0228533 [FREE Full text] [doi: 10.1371/journal.pone.0228533] [Medline: 32053609]
78. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
79. Zhao Y, He X, Feng Z, Bost S, Prosperi M, Wu Y, et al. Biases in using social media data for public health surveillance: a scoping review. Int J Med Inform 2022 Aug;164:104804. [doi: 10.1016/j.ijmedinf.2022.104804] [Medline: 35644051]
80. McCormick TH, Lee H, Cesare N, Shojaie A, Spiro ES. Using Twitter for demographic and social science research: tools for data collection and processing. Sociol Methods Res 2017 Aug 09;46(3):390-421 [FREE Full text] [doi: 10.1177/0049124115605339] [Medline: 29033471]
81. Nawaz Z, Pullen F, Rivera-Mariani FE, Rizvi SA, Sanchez-Gonzalez M, Smollar M, et al. Spring is here, now what? Know the difference between a cold, flu, coronavirus and allergy. Emerg Infect Dis Diag J 2020 [FREE Full text]
82. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
83. Rong J, Michalska S, Subramani S, Du J, Wang H. Deep learning for pollen allergy surveillance from twitter in Australia. BMC Med Inform Decis Mak 2019 Nov 08;19(1):208 [FREE Full text] [doi: 10.1186/s12911-019-0921-x] [Medline: 31699071]
84. Cowie S, Arthur R, Williams HT. @choo: tracking pollen and hayfever in the UK using social media. Sensors (Basel) 2018 Dec 14;18(12):4434 [FREE Full text] [doi: 10.3390/s18124434] [Medline: 30558222]
85. Sitaru S, Wecker H, Buters J, Biedermann T, Zink A. Social media to monitor prevalent diseases: hay fever and Twitter activity in Germany. Allergy 2023 Oct 14;78(10):2777-2780. [doi: 10.1111/all.15787] [Medline: 37317626]
86. Sousa-Pinto B, Heffler E, Antó A, Czarlewski W, Bedbrook A, Gemicioglu B, et al. Anomalous asthma and chronic obstructive pulmonary disease Google Trends patterns during the COVID-19 pandemic. Clin Transl Allergy 2020 Nov 02;10(1):47 [FREE Full text] [doi: 10.1186/s13601-020-00352-9] [Medline: 33292529]
87. Schmidt CW. Pollen overload: seasonal allergies in a changing climate. Environ Health Perspect 2016 Apr;124(4):A70. [doi: 10.1289/ehp.124-a70]
88. Oh JW. Pollen allergy in a changing planetary environment. Allergy Asthma Immunol Res 2022 Mar;14(2):168-181 [FREE Full text] [doi: 10.4168/aair.2022.14.2.168] [Medline: 35255535]
89. Rothman S, Walker SA. The problem of emotional factors in the allergies. Int Arch Allergy Immunol 2009 Jul 1;1(4):306-315. [doi: 10.1159/000227874]
90. De Peuter S, Van Diest I, Lemaigre V, Li W, Verleden G, Demedts M, et al. Can subjective asthma symptoms be learned? Psychosom Med 2005;67(3):454-461. [doi: 10.1097/01.psy.0000160470.43167.e2] [Medline: 15911910]

Abbreviations

ABNER: A Biomedical Named Entity Recognizer
BERT: Bidirectional Encoder Representations From Transformers
DNORM: disease name normalization
KNN: k-nearest neighbor
LLM: large language model
NER: named entity recognition
NLP: natural language processing
OSCAR4: Open Source Chemistry Analysis Routines
PCA: principal component analysis
TF-IDF: term frequency–inverse document frequency

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.03.24; peer-reviewed by B Audeh, S Tagliaferro; comments to author 25.04.24; revised version received 27.05.24; accepted 10.09.24; published 21.10.24.

Please cite as:
Pérez-Pérez M, Fernandez Gonzalez M, Rodriguez-Rajo FJ, Fdez-Riverola F
Tracking the Spread of Pollen on Social Media Using Pollen-Related Messages From Twitter: Retrospective Analysis
J Med Internet Res 2024;26:e58309
URL: https://www.jmir.org/2024/1/e58309
doi:10.2196/58309
PMID:39432897

©Martín Pérez-Pérez, María Fernandez Gonzalez, Francisco Javier Rodriguez-Rajo, Florentino Fdez-Riverola. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Discussion of Heated Tobacco Products on Twitter Following IQOS's Modified-Risk Tobacco Product Authorization and US Import Ban: Content Analysis

Minji Kim[1], PhD; Julia Vassey[2], PhD; Dongmei Li[3], PhD; Artur Galimov[2], PhD; Eileen Han[4], PhD; Matthew G Kirkpatrick[2,5], PhD; Cassandra A Stanton[6], PhD; Jenny E Ozga[6], PhD; Sarah Lee[3], BS; Jennifer B Unger[2], PhD

[1]Department of Health Promotion, Education, and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States

[2]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

[3]Department of Clinical and Translational Research, University of Rochester Medical Center, Rochester, NY, United States

[4]Center for Tobacco Control Research and Education, University of California, San Francisco, San Francisco, CA, United States

[5]Department of Psychology, University of Southern California, Los Angeles, CA, United States

[6]Behavioral Health and Health Policy Practice, Westat, Rockville, MD, United States

Corresponding Author:
Minji Kim, PhD
Department of Health Promotion, Education, and Behavior
Arnold School of Public Health
University of South Carolina
915 Greene Street, Suite 554
Columbia, SC, 29208
United States
Phone: 1 803 777 1904
Email: minjikim@sc.edu

Abstract

Background: Understanding public opinions about emerging tobacco products is important to inform future interventions and regulatory decisions. Heated tobacco products (HTPs) are an emerging tobacco product category promoted by the tobacco industry as a "better alternative" to combustible cigarettes. Philip Morris International's IQOS is leading the global HTP market and recently has been subject to important policy events, including the US Food and Drug Administration's (FDA) modified-risk tobacco product (MRTP) authorization (July 2020) and the US import ban (November 2021). Although limited in their legal implications outside the United States, these policy events have been quoted in global news outlets and Philip Morris International's promotional communications, showing how they may potentially impact global tobacco regulation. Given the impending return of IQOS to the US market, understanding how the policy events were received through social media discourse will provide valuable insights to inform global tobacco control policy.

Objective: This study aims to examine HTP-related social media discourse around important policy events.

Methods: We analyzed HTP-related posts on Twitter during the time period that included IQOS's MRTP authorization in the United States and the US import ban, examining personal testimonial, news/information, and direct marketing/retail tweets separately. We also examined how the tweets discussed health and policy. A total of 10,454 public English tweets (posted from June 2020 to December 2021) were collected using HTP-related keywords. We randomly sampled 2796 (26.7%) tweets and conducted a content analysis. We used pairwise co-occurrence analyses to evaluate connections across themes.

Results: Tweet volumes peaked around IQOS-related policy events. Among all tweets, personal testimonials were the most common (1613/2796, 57.7%), followed by news/information (862/2796, 30.8%) and direct marketing/retail (321/2796, 11%). Among personal testimonials, more tweets were positive (495/1613, 30.7%) than negative (372/1613, 23.1%), often comparing the health risks of HTPs with cigarettes (402/1613, 24.9%) or vaping products (252/1613, 15.6%). Approximately 10% (31/321) of the direct marketing/retail tweets promoted international delivery, suggesting cross-border promotion. More than a quarter of tweets (809/2796, 28.9%) discussed US and global policy, including misinterpretation about IQOS being a "safer" tobacco product after the US FDA's MRTP authorization. Neutral testimonials mentioning the IQOS brand (634/1613, 39.3%) and discussing policy (378/1613, 23.4%) showed the largest pairwise co-occurrence.

XSL•FO
RenderX

Conclusions: Results suggest the need for careful communication about the meaning of MRTP authorizations and relative risks of tobacco products. Many tweets expressed HTP-favorable opinions referring to reduced health risks, even though the US FDA has denied marketing of the HTP with reduced risk claims. The popularity of social media as an information source with global reach poses unique challenges in health communication and health policies. While many countries restrict tobacco marketing via the web, our results suggest that retailers may circumvent such regulations by operating overseas.

(J Med Internet Res 2024;26:e53938) doi:10.2196/53938

KEYWORDS

heated tobacco products; IQOS; social media; Twitter; tobacco control; modified-risk tobacco product authorization; MRTP authorization; tobacco regulatory science; import ban; observational study; public opinion; content analysis

Introduction

Heated tobacco products (HTPs) use electronic heating devices that heat processed tobacco sticks to create nicotine-containing aerosol [1-3]. Philip Morris International (PMI) introduced IQOS in Switzerland and Japan in 2014 and has been leading the global HTP market. Other popular products in the category include British American Tobacco's (BAT) Glo, Japan Tobacco International's Ploom, and KT&G's Lil [4].

IQOS launched in selected markets in the United States in 2019, after its premarket tobacco product authorization by the US Food and Drug Administration (FDA) in 2019 [5,6]. In July 2020, the FDA authorized IQOS's modified risk tobacco product (MRTP) claims that complete switching to IQOS from cigarettes would reduce exposure to harmful and potentially harmful chemicals ("reduced exposure" claims), while denying authorization marketing claims that using IQOS can reduce the risks of tobacco-related diseases [7]. Also in 2020, two major patent infringement lawsuits were filed against PMI by BAT (April 2020) and Healthier Choices Management Corp (November 2020), with hearings at the US International Trade Commission (ITC) starting in the end of January 2021 [8]. As a result, in May 2021, PMI lost the first round of BAT lawsuits. In September 2021, the US ITC banned the import of IQOS into the United States, which took effect in November 2021 [9]. As of March 2024, IQOS is still not available in the United States, although PMI recently reached an agreement with BAT to resolve all ongoing patent infringement litigation [10], which will result in resuming sales of IQOS in the United States.

Even before the US MRTP authorization, HTPs have been promoted as "cleaner," "high-tech," and "better" alternatives to combustible tobacco products in other countries such as Japan [11] and South Korea [3], using explicit and implicit claims, which were often interpreted as HTPs being "less harmful" than cigarettes [12,13]. The tobacco industry has argued that HTPs are different from other tobacco products, primarily by claiming that they pose a lower level of health risk, in an effort to avoid stricter regulation [14]. However, earlier research suggests that people who smoke cigarettes who adopted HTPs continued smoking [15], potentially leading to greater nicotine exposure and addiction by using multiple tobacco products.

The effects of promotional efforts, policy events, and media coverage on public opinion about emerging tobacco products, including HTPs, need further research. Although the MRTP authorization applies only to the United States, this decision was widely quoted by global news media, often including inaccurate and misleading information about IQOS being "safer" than cigarettes, and PMI has been observed using the MRTP decision to lobby for regulatory changes outside of the United States [16]. The use of IQOS's US market presence in global public relations by PMI also suggests the potential impact of the US import ban outside of the United States. Given the impending return of IQOS to the US market, which may again be used as a global public relations opportunity by PMI, a closer look at how these policy events were received on social media is required to better inform US and global tobacco control policy.

Social media plays a crucial role in analyzing tobacco industry marketing tactics and monitoring discussions on new trends and products, as social media data can provide insights into how general public, news sources, and industry stakeholders react to tobacco control policy changes [17-19]. Further, previous studies highlighted the effectiveness of analyzing Twitter (currently X) data for insights into public's discourse on tobacco control initiatives [19,20]. Twitter attracts a diverse audience, encompassing various age groups, geographic locations, and interests. Specifically, Twitter was used by 368.4 million people globally in 2022, the time of data collection [21]. In the United States, 23% of adults reported using Twitter, with 42% of its users logging in daily [22,23]. About a quarter of US adolescents between 13 and 17 years old also reported using Twitter [24]. This varied user demographic on Twitter can shed light on a wide array of viewpoints, beliefs, and opinions about HTPs (such as IQOS) after receiving MRTP authorization.

Existing social media studies on IQOS have found that people already saw IQOS as a less harmful alternative to other tobacco products before MRTP authorization [25], with the proportion of tweets with positive attitudes toward IQOS increasing during the month after the MRTP authorization [26]. However, their application in analyzing public reactions to the new policy events that may have a qualitatively different impact on tobacco market in the United States and globally, including the IQOS MRTP authorization in the United States and the US IQOS import ban, remains limited. Therefore, this study adds to the existing literature by collecting new data after new development in HTP-related policies, observing changes in social media discourse during months surrounding significant new policy events that may impact US as well as global tobacco control policies, hence providing further insights. The tobacco industry has been actively using social media for marketing, including promoting IQOS [27,28], warranting close examination of social

XSL•FO
RenderX

media content. To understand the potential detrimental impact on perceptions of IQOS and other HTPs, we analyzed HTP-related social media posts during the period surrounding the major policy events described above, analyzing different types of tweets separately with particular attention to user perceptions around health consequences of using HTPs and MRTP-related discussions. This study also adds to existing research by including general keywords related to HTPs, not just IQOS, to capture tweets only using general terms such as "HTP" or "heat-not-burn."

Methods

Data Collection and Procedures

Original Twitter posts (tweets) were collected between June 1, 2020, and December 31, 2021, from Twitter's Streaming Application Program Interface using IQOS- and HTP-related keywords and hashtags (eg, "iqos," "heatedtobacco," and "heatnotburn"; see Section S1 in Multimedia Appendix 1 for the full list). The initial time period for data collection originally was determined immediately before and 1 year after the US MRTP authorization was made in July 2020. During early data collection in fall 2021, a new regulatory event (ruling of IQOS's patent infringement against BAT's Glo and subsequent US import ban) emerged, and the research team decided to expand the data collection window to include when the ruling went into effect in November 2021 (see Figure S1 in Multimedia Appendix 1).

Similar to previous studies [25,29], we excluded all retweets to prevent overrepresenting repeated identical statements. Because of limitations in analytical capacity, we limited our analysis to English-language tweets. Geocodes (indicating where the account is located) were included in the metadata, but because people can opt out of including such information, only a small number of tweets included location information, and therefore, geocodes were excluded from analyses.

The initial search yielded 65,333 unique public tweets collected over 19 months; after removing undiscernible or ambiguous content (eg, non-English or emoji-only tweets), we had 10,423 English tweets. We selected a random subset (n=4138) of them (see Figure S1 in Multimedia Appendix 1 for a flow diagram of the tweet selection process) to facilitate manual content analysis.

We developed a codebook informed by prior research [25,26] and refined it during the initial coding and training process (see Tables 1 and 2 for codebook definitions and tweet distribution; see Figure S2 in Multimedia Appendix 1 for more details about the codebook structure and coding procedure). Previous studies coded emerging themes among tweets, including news, health claims, marketing, personal testimonial, policy, or other [25], and examined overall and across-time distributions of neutral-, positive-, or negative-sentiment tweets [26]. Major updates in this codebook include clarifying mutually exclusive categories of tweets ("personal testimonial," "news/information," or "direct marketing/retail") to better reflect the nature of individual tweets and allow separate analyses within each category. We aimed to better focus on personal testimonials that would more closely reflect population perceptions. In addition, 2 main overarching discussion topics, health and policy, were additionally coded irrespective of whether an individual tweet was categorized as personal testimonial, news/information, or direct marketing/retail.

XSL•FO
RenderX

Table 1. Distribution of tweets by 3 mutually exclusive categories in the analytic set (n=2796): direct marketing/retail, news/information, and personal testimonials.

Categories[a]	Tweets, n	Corpus (%)	Category (%)	Example
Personal testimonials: sharing one's opinions or direct/indirect experiences of HTPs[b], including perspective from users and nonusers	1613	57.7	100	Happy that TGA[c] rejected the selling of heated tobacco products in Australia due to public health and potential harm [URL]
Valence				
Positive	495	17.7	30.7	I recently bought an IQOS and since then, only smoked 2 cigarettes instead of 30-40. Good cut [URL]
Neutral	746	26.7	46.2	Watch me buy IQOS after [@] posted it. Always a trigger for me lol
Negative	372	13.3	23.1	Why are the people that do the IQOS kiosks in Racetrack super friendly? Like you don't expect someone selling you cancerous products to be as nice as almost all of them are. But they are.
Content				
Flavor or taste	30	1.1	1.9	[@] I should accept that I couldn't stop smoking and start vaping. I missed the tobacco taste. Then enters IQOS, now that's good.
Experiential remarks	270	9.7	16.7	IQOS smells like farts. How do you smoke that?!
Information seeking	307	11.0	19.0	[@] Hi, I bought IQOS yesterday and the blade broke. Can you help?
Comparison to combustible tobacco	401	14.3	24.9	The FDA[d] has approved #iQOS' [MRTP[e]] application to claim IQOS is safer than cigarettes, encouraging tobacco companies to make safer products. That's sound public policy.
Comparison to ENDS[f]	252	9.0	15.6	#IQOS is a joke. Get a vape or loose leaf vaporizer and avoid being subject to a single brand and its flavors
Tobacco marketing strategies	198	7.1	12.3	I declined an offer to try IQOS for free by an IQOS salesman. She turned away with a disgusted look #IQOS
News/information: news articles, statements, or summaries of information and facts about HTPs	862	30.8	100	New study demonstrates that heated tobacco products caused a notable reduction in Japan for combustible cigarettes sales. Read more: [URL]
Comparison to combustible tobacco	192	6.9	22.3	EU[g] countries request #electronic #cigarettes and HTPs to be #taxed like traditional tobacco products [@] #ecigarette #vaping [URL]
Comparison to ENDS	81	2.9	9.4	New study at Queen Mary University of London shows that IQOS was less effective for reducing cravings since it yielded less nicotine compared to JUUL. [URL] #VapeNews #Stoptober #vape
Tobacco marketing strategies	113	4.0	13.1	Following video with [@] discusses how the tobacco industry in Guatemala continues using old strategies to promote new products, such e-cigarettes and heated tobacco products. [URL]
Referring to scientific evidence	225	8.0	26.1	Only 10 of the 62 harmful compounds found in a new IQOS experiment were found in previous studies. Get help for free to quit at [URL] or [Phone number]
Direct marketing/retail: selling or promoting HTPs through direct marketing claims, including promoting point of sales	321	11.5	100	[@] New Marlboro IQOS Heat Sticks Black Menthol Flavor, available for international delivery [URL]
Direct marketing/retail: flavor	66	2.4	20.6	Promotion!!! Taste Amber, Yellow, Turquois, Purple [flavors]. You can pay via Alipay. #iqos #heets #amber #yellow #torquois #purple #heatnoburn #smoke [URL]

XSL•FO
RenderX

Categories[a]	Tweets, n	Corpus (%)	Category (%)	Example
Direct marketing/retail: product features	84	3.0	26.2	Easily eject heat sticks via a sliding button instead of cap removal. Learn more [URL] #london #uk #hnb [URL]

[a]Note: The 3 categories (direct marketing/retail, news/information, and personal testimonials) are mutually exclusive. Within personal testimonials, the 3 valence groups are also mutually exclusive. The remaining content themes within each category are not mutually exclusive. [@], [URL], and [Phone number] indicate redacted information.

[b]HTP: heated tobacco product.

[c]TGA: Therapeutic Goods Administration.

[d]FDA: Food and Drug Administration.

[e]MRTP: modified risk tobacco product.

[f]ENDS: electronic nicotine delivery systems.

[g]EU: European Union.

Table 2. Distribution of topics and relevant subthemes in the analytic set (n=2796)

Topics	Tweets, n	Corpus (%)	Topic (%)	Example[a]
Health	632	22.6	100	[@] warns that HTP[b] and traditional cigarettes have similar amounts of nicotine and secondhand aerosol despite being marketed as better alternative
Comparison to combustible tobacco	298	10.7	47.2	HTP health warnings should reflect their lower risks than cigarettes. Source: [URL] #VapeNews
Comparison to ENDS[c]	126	4.5	19.9	[@] Nicotine vapes are undeniably safer than smoking and HNB[d] products that deliver fewer toxins/carcinogens. Few teens vape. Why is this not seen as a way to quit?
Cessation	194	6.9	30.7	[@] I tried IQOS, very good for quitting cigarettes
Policy	809	28.9	100	Altria can't sell IQOS in US since Biden administration stays out of patent dispute – CNBC: After deadline passed, Altria and PMI unable to sell/import IQOS devices in US [URL] #patentnews
US-based policy: MRTP[e]	105	3.8	13.0	[@] After FDA[f] denied 'modified risk' order but PM marketed as one, FDA granted an 'Exposure Modified' order for #iQOS. No accountability? [@] response to [@] #IQOS order. [URL]
US-based policy: Other	450	16.1	55.6	If cigarettes with low nicotine get a PMTA[g], then all HTP or vaping HTP are bad joke.
Non-US policy	244	8.7	30.2	In Japan, e-cigs are illegal, HTPs are legal. Of course HTP lowered cigarette sales. Do not fear. WHO is trying to reverse that by restricting HTP use.
HTP brand mentioned	2796	100	100	
IQOS	1838	65.7	65.7	My JUUL and IQOS witness my daily pack of cedars blue cigarettes [URL]
Other brands	205	7.3	7.3	[URL] [@] will release new heated tobacco product Ploom X on Aug 17 at convenience stores in Japan and certain retail stores. Currently available for presale
General HTPs	835	29.9	29.9	Hong Kong Council on Smoking and Health advise the banning of HNB and e-cigarettes due to misleading lower health risks. These risks have been proven globally [URL]

[a]Note: [@], [URL], and [Phone number] indicate redacted information.

[b]HTP: heated tobacco product.

[c]ENDS: electronic nicotine delivery systems.

[d]HNB: heat-not-burn.

[e]MRTP: modified risk tobacco product.

[f]FDA: Food and Drug Administration.

[g]PMTA: premarket tobacco product authorization.

XSL•FO
RenderX

Tweets were categorized into four mutually exclusive categories: (1) personal testimonials, (2) news/information, (3) direct marketing/retail, or (4) noise (irrelevant or not understandable). We focused our analysis on the 2796 tweets that fit into categories 1-3.

The unit of analysis was the text of the tweet; images were not included in the analyses. When coding the tweets, we only considered information that was explicitly mentioned—that is, content that may have been included in the linked websites but not discussed in the tweets was not considered. Also, many tweets include a long list of hashtags at the end, with many hashtags only loosely related to the actual tweet content or adding any new information. Therefore, we only considered hashtags that were part of the actual tweet sentences (eg, hashtags considered: "Globally, #HTPs help millions of smokers and are a safer alternative"; not considered: "HTPs: next generation of smoke-free alternatives for teens #heatedtobacco #tobacco #prevention") [30].

The tweets within each category were coded into specific subthemes (not mutually exclusive), using similar approaches from previous research [25,26] as well as via iterative discussion during the coder meetings. Among personal testimonials, six subthemes emerged: (1-a) description of flavor or taste in the context of a personal recommendation, thoughts, or review; (1-b) personal experience using HTPs either favorably or unfavorably (eg, feeling of smoke and taste); (1-c) information seeking about HTPs (eg, troubleshooting and asking about product availability); (1-d) comparison to combustible tobacco; (1-e) comparison to electronic nicotine delivery systems (ENDS); or (1-f) HTP marketing strategies. Personal testimonial tweets were also coded for their valence: positive, neutral, or negative. News/information tweets were coded for four subthemes: (2-a) comparison to combustible tobacco, (2-b) comparison to ENDS, (2-c) tobacco marketing strategies, or (2-d) referring to scientific evidence. Direct marketing/retail tweets were coded for two subthemes: (3-a) flavors and (3-b) product features.

All tweets were further coded for 2 not mutually exclusive topics: health (mentioning health risks and/or benefits of using HTPs; subthemes included comparison to combustible tobacco, comparison to ENDS, and/or cessation) and/or policy (discussing HTP-related tobacco policy or regulation; subthemes included mentions of MRTP authorization, other US-based policy, and/or policy outside of the United States). Tweets were also coded for specific HTP brands (not mutually exclusive): whether they mentioned IQOS, other HTP brands (eg, Ploom and Glo), or referred to general product category.

Three independent coders coded the tweets using a binary system (yes/no) for each category, topic, and subtheme (see Figure S2 in Multimedia Appendix 1). To establish interrater reliability, 3 coders double-coded a subsample of tweets (total 300, approximately 10% of the analytic sample [19,30,31]) through an iterative procedure, where differences across coders in the initial practice set were discussed in real time and coding definitions were amended and refined. Coders applied the refined codebook to the next smaller set and repeated this process until consensus was reached and inter-rater agreement reached at least 90%. Because of a lack of variance for some codes (ie, too many 0's vs 1's), κ could not be reliability calculated; therefore, we used percent agreement instead to assess inter-rater reliability. The first author served as an arbitrator resolving discrepancies between the coders. Then the 3 coders each independently coded a third of the remaining tweets.

Data Analysis

After completing the coding process, we conducted a descriptive analysis of the tweets according to categories and themes. We also performed a pairwise co-occurrence analysis to assess connections across themes: for example, we examined the proportion of personal testimonials tweets that also discussed health-related content about HTPs. To visualize the co-occurrence matrix containing the counts of co-occurrences of themes, the *igraph* package version 1.2.6 in R (R Foundation for Statistical Computing) was used [25]. Examining the co-occurrence of themes, separate analyses were conducted within the mutually exclusive categories, and we present results from the personal testimonials and news/information tweets, excluding direct marketing/retail tweets that were the least frequent.

Ethical Considerations

As this study analyzed publicly accessible information only, no ethics board review has been conducted. All URLs and individual Twitter handles were redacted (eg, "[URL]" rather than reproducing what was included in the tweet) to protect privacy. To further protect privacy, posts used in this paper were paraphrased by one of the coders so that the tweets included cannot be easily traced back to original accounts.

Results

Distribution of Tweets Across Time

The distribution of tweets across the 19-month period is shown in Figure 1, along with several policy and business events. Compared with June 2020, a 50% increase in the number of tweets was observed in July 2020 after the IQOS MRTP authorization. Other peaks in the number of tweets were observed in February, May, and September 2021, related to events surrounding IQOS patent infringement lawsuits and the US import ban. This trend was mostly driven by news/information and policy-related tweets. Health-related tweets were most frequent in July 2020, immediately after the US FDA's MRTP authorization of IQOS reduced exposure claims. Personal testimonials were most frequent in February and September 2021, coinciding with legal disputes and the US import ban. Direct marketing/retail tweets were generally lower in number, with a peak observed in January 2021.

XSL•FO
RenderX

Figure 1. Distribution of tweets (n=2976) between June 2020 and December 2021 and relevant policy events. (A) Overall tweets (left axis) and 3 categories (right axis: direct marketing/retail, news/information, and personal testimonials). (B) Overall tweets (left axis) and 2 topical themes (right axis: health and policy). BAT: British American Tobacco; FDA: Food and Drug Administration; ITC: International Trade Commission; MRTP: modified-risk tobacco product; PMI: Philip Morris International.

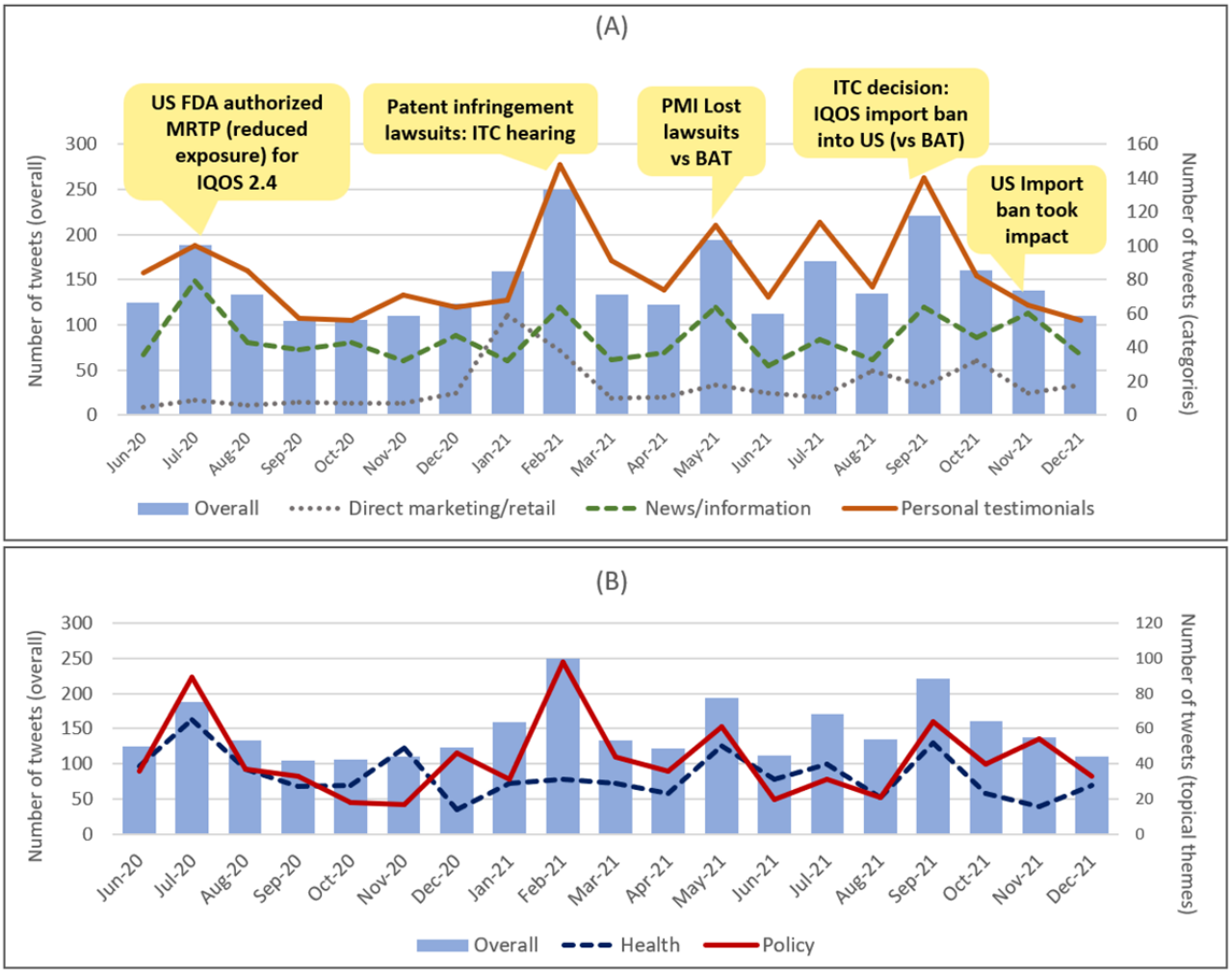

Categories and Topical Themes

Descriptive Characteristics

Tables 1 and 2 show the descriptive characteristics of tweets and paraphrased example tweets illustrating each category (Table 1: mutually exclusive) and topic (Table 2: not mutually exclusive) with relevant subthemes. Overall, the most prevalent category was personal testimonials (1613/2976, 57.7%), followed by news/information (862/2976, 30.8%), and direct marketing/retail (321/2976, 11.5%). In terms of topic, policy (809/2976, 28.9%) was more prevalent than health (632/2976, 22.6%).

Categories of Tweets: Personal Testimonials, News/Information, and Direct Retail/Marketing

Within personal testimonials, close to half of the posts (746/1613, 46.2%) were neutral, followed by positive (495/1613, 30.7%) and negative (372/1613, 23.1%) posts. Examples of positive tweets often included mentions of reduced harm (or its potential), for example, “For the past year, IQOS has been great for my lungs and throat” or discreet nature of using HTPs (eg, “I like how IQOS makes smoking discreet. I can smoke so much at home without anyone noticing.”). Some tweets simply expressed favorable attitudes toward IQOS and other HTPs, such as “Falling in love with my IQOS, more every day.” Examples of negative tweets included suspicion toward the tobacco industry (eg, “#PhillipMorris views menthol #cigarette ban as opportunity to market heated tobacco products and make money, not harm reduction. Stop regulatory loopholes. Heated tobacco is tobacco.”), or the perception that HTPs do not differ from smoking cigarettes (eg, “IQOS is same as smoking with many extra hassle, including charging and higher cost. Terrible.”).

Personal testimonial tweets comparing HTPs to combustible tobacco appeared more frequently (401/1613, 24.9%) than those comparing HTPs to ENDS (252/1613, 15.6%), with 8.1% (131/1613) mentioning both comparisons. In both comparisons, positive was the most prevalent sentiment (199/401, 49.6%, for combustible and 117/252, 46.4%, for ENDS comparisons), followed by neutral (119/401, 29.7%; 78/252, 31.0%) and then negative toward HTPs (83/401, 20.7%; 57/252, 22.6%). Positive comparisons of HTPs to combustible tobacco testimonials generally pointed to harm reduction potential and lack of smoke, whereas comparisons to ENDS were mixed. Some mentioned that HTPs and vaping are “on the same side” in terms of harm

XSL•FO
RenderX

reduction, whereas some mentioned differential preference between vaping and HTPs (eg, “ecigs just doesn't work for everyone. IQOS and snus are much more satisfying but heated tobacco is not everyone's cup of tea. Respect diverse tastes.”). Among negative tweets, comparisons to both combustible tobacco and ENDS often discussed that HTPs are similar to (or no better than) smoking (eg, “Researchers compared IQOS and cigarettes but weren't able to find enough data to show they are different on cardiovascular problems in testing over short time.”).

Nearly half (407/862, 47.2%) of news/information tweets discussed policy (see the “Topical Themes: Health and Policy” section below), including the IQOS MRTP authorization (eg, “FDA authorizes IQOS marketing with reduced exposure claims [URL]”). Tweets in this category often compared HTPs to combustible tobacco (192/862, 22.3%) or ENDS (81/862, 9.4%), most of which mentioned the potential health risks and benefits of HTPs (eg, “PMI claimed compared to cigarettes, IQOS aerosol includes 95% fewer toxins. However, this and other similar results all come from tobacco industry-funded research.”). About a quarter of news/information tweets (225/862, 26.1%) cited scientific evidence, such as peer-reviewed studies or other data sources (eg, “An #onlinefirst study showed that heated tobacco products, e-cigarettes, and cigarettes are correlated with biomarkers of stiff arteries.”).

Only 11.5% (321/2796) of all tweets belonged to this category, mostly referring to online retailers and accompanied by web links. Among these, 20.6% (66/321) mentioned flavors (eg, “Mellow tobacco blend and zesty aroma notes in Yellow Selection HEETS tobacco sticks for IQOS that are now available for purchase”). Although none of these tweets included geocodes, we noted that 45.2% (145/321) of the direct marketing/retail tweets included phone numbers, region names, or currency units, with the United Arab Emirates being the most frequent (92/321, 28.7%; eg, “IQOS HEETS Sienna available in Dubai/Abu Dhabi”), followed by the United Kingdom (22/321, 6.9%; eg, “Come visit our stores to learn about Ploom from in-store experts. We are always here to help you! #Ploom #PloomUK #HeatedTobacco #Westfield”). A total of 9.7% (31/321) of marketing/retail tweets mentioned “international delivery” as an available option.

Topical Themes: Health and Policy

Nearly a quarter (632/2796, 22.6%) of the tweets were health related. Similar to what we observed among personal testimonials, about half of health-related tweets (298/632, 47.2%) compared HTPs with combustible tobacco, whereas only 19.9% (126/632) compared HTPs with ENDS. More than a quarter (194/632, 30.7%) of health-related tweets discussed IQOS’s efficacy for cigarette cessation, approximately three-quarters of which included one’s own experience using IQOS either successfully (eg, “Try IQOS. I've quit traditional cigs for a year and it's awesome!”) or unsuccessfully (eg, “I was a smoker for 12 years. I tried to stop smoking by using IQOS, but I failed… now I need to quit both”). Only 9 direct marketing/retail tweets discussed health (eg, “If you want to switch from cigarettes, give Ploom S Kit a try. Order yours from the [Shop name]”).

More than a quarter (809/2796, 28.9%) of all tweets discussed policy: the majority of them (450/809, 55.6%) discussed US-based policy other than MRTP authorization, such as marketing authorization under premarket tobacco product application for IQOS 2.4 (April 2020) or IQOS 3 (December 2020), or IQOS import ban to the United States (September 2021). Discussion of non-US policy was also a common topic among policy tweets (244/809, 30.2%; eg, “New German tobacco law states tax increase of 3% on cigarettes, 26% on HTPs, and 1000% on e-cigs”). MRTP authorization (105/809, 13.0%) discussions were relatively uncommon.

Although the FDA made it clear that their MRTP marketing order for IQOS is limited to reduced exposure, not reduced risks, some tweets claimed that this decision proves that IQOS is safer than other tobacco products and referred to the authorization decision as FDA “approval” (eg, “Yes they are safe - FDA approved them because they see IQOS as appropriate for #publichealth protection.”). Some tweets discussed that this US policy will also impact global marketing of IQOS (eg, “…RJ Reynolds won the lawsuit against PMI, so IQOS cannot be sold in the US. But their FDA approval stays and will help PMI to promote IQOS globally.”).

Thematic Co-occurrences

Figures 2 and 3 display separate co-occurrence analyses within personal testimonial (Figure 2) and news/information (Figure 3) tweets. Among personal testimonials, neutral tweets mentioning IQOS brand (634/1613, 39.3%) showed the largest co-occurrence, followed by neutral tweets discussing policy (378/1613, 23.4%) and neutral tweets about general HTPs (268/1613, 16.6%). There were slightly more health-related positive testimonials (220/1613, 13.6%) than health-related negative testimonials (144/1613, 8.9%). News/information tweets had the highest co-occurrence with IQOS brand (447/862, 51.9%), policy (407/862, 47.2%), general HTPs (393/862, 45.6%), followed by scientific evidence (225/862, 26.1%), US-based policy (223/862, 25.9%), and health (178/862, 20.6%). The thickness of the lines represents the frequency of pairwise co-occurrences, and the size of the circles represents the frequency of theme occurrences.

XSL•FO
RenderX

Figure 2. Theme co-occurrences of personal testimonials (n=1613) with policy (n=399), health (n=445), and brand (IQOS: n=1173; other: n=82; general HTPs: n=424). HTP: heated tobacco product.

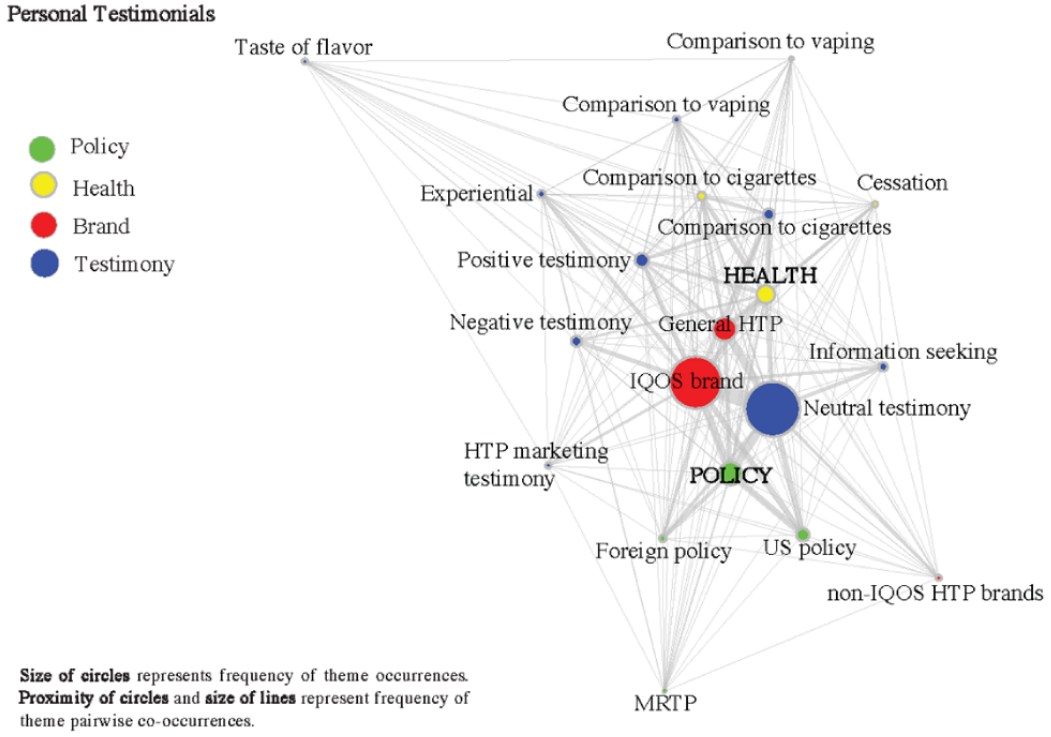

Figure 3. Theme co-occurrences of news/information (n=862) with policy (n=407), health (n=278), and brand (IQOS: n=447; other: n=30; general HTPs: n=393). HTP: heated tobacco product; MRTP: modified-risk tobacco product.

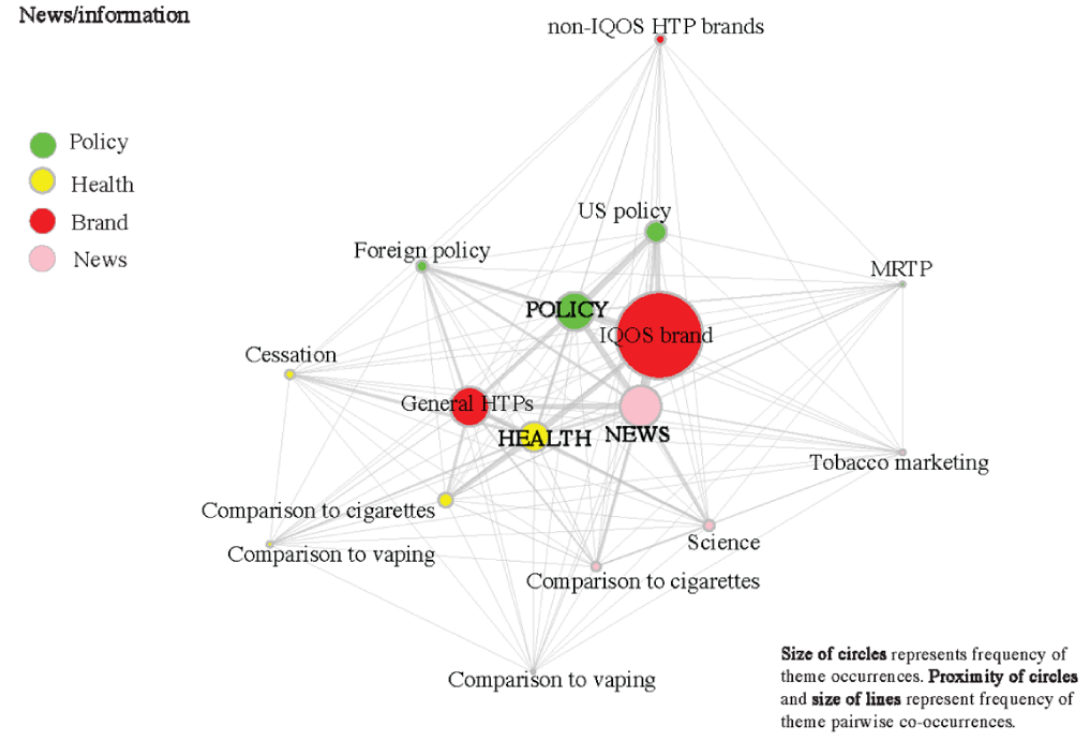

Discussion

Principal Findings

This study reviewed tweets related to HTPs between June 2020 (1 month before the US FDA's MRTP authorization for IQOS 2.4) and December 2021 (1 month after the IQOS US import ban took effect) to observe people's response to major policy decisions related to HTPs. Social media such as Twitter, Reddit, and TikTok have emerged as a major source of news and other information, especially among young people [32,33]. Tobacco industry has also been actively using social media as a promotion outlet for emerging tobacco products [27,28]. A recent study on Instagram suggests that despite the platform's attempts at regulating tobacco promotions, paid partnership

XSL•FO
RenderX

posts are abound promoting venues and events that feature tobacco use, suggesting limited compliance and/or enforcement [34]. Social media use is associated with greater exposure to tobacco marketing and promotions [35] and more favorable perceptions toward and use of nicotine and tobacco products [36,37]. Coupled with concerns related to the lack of regulations and enforcement due to their global nature and strong reliance on user-created content [38,39], a closer surveillance is warranted for social media discourse surrounding emerging tobacco products.

We found that more than half of the tweets were personal testimonials, expressing thoughts and opinions based on one's own experience and/or knowledge, which provide a valuable window into public opinion regarding HTPs. Close to half of the personal testimonials were neutral, either not expressing clear valence or mentioning both positive and negative perspectives. Among the non-neutral personal testimonials, consistent with a previous study [25], more tweets expressed a positive than negative valence toward HTPs. In positive-valence tweets, people described IQOS as less harmful, "better," or "safer" than combustible tobacco products, namely cigarettes, or discussed HTPs as being able to help people who smoke quit smoking. Another study examining IQOS-related tweets between November 2019 and August 2020 (2 months after MRTP authorization) found that overall there were more negative than positive tweets, but also found that positive tweets were more frequently observed around MRTP authorization [26]. This study suggests that this trend continued during the year after MRTP authorization. This is also consistent with PMI's harm reduction marketing strategies in and outside of the United States [3,40,41], including their global public relations efforts emphasizing the US FDA's MRTP authorization [16,42], and may reflect explicit and implicit marketing influence. However, it should be noted that the US FDA's MRTP authorization for IQOS was limited to reduced exposure to harmful chemicals and did not include reduced risks of tobacco-related disease, compared to cigarettes [7]. With insufficient evidence on whether or how much IQOS or other emerging tobacco products may reduce the health risks in the short and long term compared to cigarettes [43], such discussions should be closely monitored and, if necessary, corrected using effective communication campaigns.

Comparison With Prior Work

This study provides recent examination of social media discourse by collecting recent dataset from Twitter. This is also extremely timely given the impending return of IQOS to the US market after a legal settlement between PMI and BAT, which will resolve ongoing litigations [10]. Compared to an early study [25], tweets that directly market or promote HTPs comprised a substantially lower proportion of the corpus (Barker et al [25]: 32.3%; this study: 11.5%). The observed reduction in direct marketing may be a result of an increase in other categories. For example, personal testimonials may have increased as IQOS is known to more people after MRTP authorization and PMI's continued promotional effort globally. Or, this may reflect the tobacco industry's shift to more stealthy marketing where paid influencers post promotional content as personal opinion, as the tobacco industry has been historically doing [28]. In addition, our timeline involved multiple major policy events surrounding IQOS. There may have been more policy-related discussions in both the personal testimonial and news/information categories demonstrated by our co-occurrence analysis, resulting in a relative decrease in the proportion of direct marketing/retail tweets. This finding may also reflect relatively lower direct marketing/retail activities on Twitter than other social media platforms, such as Instagram and YouTube, by IQOS official accounts [27].

Direct marketing/retail tweets were observed most often during January 2021. This could be because the HTP retailers may target those who want to stop smoking combustible cigarettes as part of their new year's resolutions, suggesting switching to HTPs. Such cessation claims have been observed in early marketing and promotion of e-cigarettes [44-46]. However, there were no specific health or cessation claims observed in direct marketing/retail tweets in our analytic set. Most tweets surrounding cessation claims were personal testimonials, in which individual Twitter users discussed their own experience of smoking cessation using HTPs.

It should be noted that HTP marketing is regulated by many social media platforms [39]. Moreover, the US FDA's MRTP marketing order requires that all digital marketing for IQOS be allowed only with audience age verification and mandates continuous monitoring and reporting of all promotional activities [47]. However, content on social media and other online channels that originates overseas can easily reach an international audience, including people in the United States. Our search, conducted from the United States, included multiple direct marketing/retail tweets that could be seen without an age verification. Content of some direct marketing/retail tweets suggested that many originated from outside of the United States, such as from the United Arab Emirates or the United Kingdom, with some tweets offering "international delivery." Therefore, despite the US regulation that restricts shipping of tobacco products [48], it should be closely examined whether retailers operating from overseas may be able to circumvent national retail and promotional restrictions [49].

This study adds to existing information on social media discussions about other emerging tobacco products such as ENDS [29,30,50] by examining how an emerging product (ie, HTPs) is discussed in comparison to ENDS and cigarettes. Our findings are consistent with existing research that people who post on Twitter may misjudge the emerging products' relative risks to other tobacco products [50], as some posts we reviewed suggested confusion between MRTP authorization and approval, and included frequent mentions of reduced health risks compared to cigarettes. This study also adds to existing social media content analyses of HTPs [25,26] by including tweets from more recent time points that overlap with many meaningful policy events (eg, MRTP authorization and US import ban). Although these events' direct legal impact is limited to the United States, PMI has been using MRTP authorization in their global public relations efforts [42] and lobbying for international tobacco regulatory policy changes [16]. Moreover, a news content analysis in low- and middle-income countries found that half of the 50 identified news articles covering US MRTP authorization during the 6 months after the announcement

XSL•FO
RenderX

included misleading "reduced risk" language [51] even though the US FDA's decision only specifies "reduced exposure" to harmful and potentially harmful chemicals. This may lead to incorrect risk perceptions around IQOS and other HTPs in and outside of the United States and calls for the needs of close surveillance of public opinion and effective health communication campaigns to inform appropriate regulatory policies and interventions globally.

Limitations

This study also comes with limitations that are not uncommon in similar studies examining social media content. First, this study only considered posts from publicly accessible accounts and may not reflect the attitudes of people with private Twitter accounts. Additionally, because of the analytical capacity, we limited our corpus to English language only. This potentially misses much conversation happening in non–English speaking countries such as Japan, Germany, and Italy, where IQOS seems to have gained greater popularity [52], as well as conversations conducted in non-English languages commonly spoken in the United States (eg, Spanish). Previous research suggests that Twitter users' information-sharing networks largely cluster and segregate as a function of language and topic [53], and Spanish speakers in the United States may be more likely to receive and share online misinformation in general [54]. However, it is important to note that a recent analysis of Spanish tobacco-related tweets found similar themes to previous English-only studies [55].

Some of the personal testimonials may have come from influencers connected to and sponsored by the tobacco industry, without clear disclosure within the tweet [28,56]. We did not investigate the nature of the accounts posting tweets—if the text reflected a person's direct or indirect experience and/or opinion toward IQOS, we counted it as a personal testimonial. A similar approach was applied to news/information tweets, where tweets from tobacco industry public relations accounts may have inadvertently been included. A future study should explore the origins of such tweets and how the differences across account types may differentially impact viewers' perceptions.

The location of tweets (eg, US vs non-US) would have been helpful to better understand the impact of the US-based policy events on global tobacco control policy; however, because of the small number of tweets (approximately 2%) that include location information, such analyses could not be conducted. Future content analyses may provide further insights if they can verify information where the posts originated.

Moreover, our analyses focused on the textual content of tweets, not including external links or images. This may have led to some loss of information. For example, some tweets were vague about which authorization they referred to (eg, "Senate plenary to deliberate regulation of import, manufacturing, sales, distribution, and use of vapes and HTPs"), so it was unclear whether they should be coded as US or foreign policy. It is also possible that some additional themes and subthemes would have emerged if we included images in the coding process [57]. This study also collected data only from Twitter from a specific time period, and therefore the results may not generalize to other popular social media platforms such as Facebook, Instagram, or TikTok that have different format features, platform regulations, and user populations.

Conclusions

Our study suggests that social media content analysis can be a useful tool for public opinion surveillance related to emerging tobacco products, including the sentiment of social media users regarding emerging tobacco products. Such information should be actively used to determine what kind of interventions and communication campaigns should be prioritized. This study categorized and analyzed 3 different types of tweets discussing HTPs: personal testimonials, news/information, and direct sales/marketing. Our findings suggest that after the US FDA's IQOS MRTP authorization, many Twitter users held neutral or positive perspectives toward HTPs. They perceived IQOS as being less harmful or safer than cigarettes, even though the MRTP authorization was limited to "reduced exposure" claims, not "reduced risk." Given that many people turn to social media as a source of information, examining news/information tweets also provides insight into the nature of information about HTPs that people are exposed to. This includes some areas for concern, including misrepresentation of MRTP authorization as "FDA approval." This study also raises awareness about cross-border marketing and sales of HTPs and calls for stricter regulatory policy and effective enforcement on social media platforms globally. This kind of content analysis provides more detailed information driven by the themes of users' posts and a different perspective from traditional surveys, although continuous research is required to address its representativeness.

Acknowledgments

We thank Dr Jon-Patrick Allem, who assisted data retrieval using Twitter Application Program Interface, and the members of the Marketing Influence Special Interest Group sponsored by the Center for Coordination of Analytics, Science, Enhancement, and Logistics (CASEL) in Tobacco Regulatory Science, who provided feedback to our work in progress. This research was supported by the National Cancer Institute and the Food and Drug Administration (FDA) Center for Tobacco Products (CTP) under Award Numbers U54CA180905 and U54CA228110 and the CASEL in Tobacco Regulatory Science (U54-DA046060-01; National Institute of Drug Abuse and FDA CTP). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the FDA.

Data Availability

Analytic set and coding results may be shared with reasonable request.

XSL•FO
RenderX

Authors' Contributions

MK proposed and conceptualized the study, which further developed by feedback from JV, DL, CAS, JEO, and JBU. JBU provided the dataset and MK and JV cleaned the dataset. JV, DL, JEO, and SL coded the dataset. MK and JV led the analyses and writing, with added input from EH in health-related tweets and AG and MGK on policy-related tweets. SL created tables with paraphrased tweets. All authors reviewed and approved the manuscript and participated in the revision process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search keywords for relevant tweets, selection process of the analytic set, and coding procedure and category/content theme and subtheme structure.
[DOCX File , 109 KB - jmir_v26i1e53938_app1.docx]

References

1. Auer R, Concha-Lozano N, Jacot-Sadowski I, Cornuz J, Berthet A. Heat-not-burn tobacco cigarettes: smoke by any other name. JAMA Intern Med 2017;177(7):1050-1052 [FREE Full text] [doi: 10.1001/jamainternmed.2017.1419] [Medline: 28531246]
2. Caputi TL, Leas E, Dredze M, Cohen JE, Ayers JW. They're heating up: internet search query trends reveal significant public interest in heat-not-burn tobacco products. PLoS One 2017;12(10):e0185735 [FREE Full text] [doi: 10.1371/journal.pone.0185735] [Medline: 29020019]
3. Kim M. Philip Morris International introduces new heat-not-burn product, IQOS, in South Korea. Tob Control 2018;27(e1):e76-e78 [FREE Full text] [doi: 10.1136/tobaccocontrol-2017-053965] [Medline: 29170165]
4. Glantz SA. Heated tobacco products: the example of IQOS. Tob Control 2018;27(Suppl 1):s1-s6 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054601] [Medline: 30352841]
5. FDA permits sale of IQOS tobacco heating system through premarket tobacco product application pathway. U.S. Food and Drug Administration. 2019. URL: https://www.fda.gov/news-events/press-announcements/fda-permits-sale-iqos-tobacco-heating-system-through-premarket-tobacco-product-application-pathway [accessed 2019-04-30]
6. Churchill V, Weaver SR, Spears CA, Huang J, Massey ZB, Fairman RT, et al. IQOS debut in the USA: Philip Morris International's heated tobacco device introduced in Atlanta, Georgia. Tob Control 2020;29(e1):e152-e154 [FREE Full text] [doi: 10.1136/tobaccocontrol-2019-055488] [Medline: 32024772]
7. FDA authorizes marketing of IQOS tobacco heating system with 'reduced exposure' information. U.S. Food and Drug Administration. 2020. URL: https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-iqos-tobacco-heating-system-reduced-exposure-information [accessed 2020-07-07]
8. USITC attorney backs BAT in Philip Morris patent dispute. World Intellectual Property Review. 2021. URL: https://www.worldipreview.com/news/usitc-attorney-backs-bat-in-philip-morris-patent-dispute-20623 [accessed 2022-08-10]
9. Gretler C, Decker S. Philip Morris IQOS imports barred from U.S.; deadline passes. Bloomberg. 2021. URL: https://www.bloomberg.com/news/articles/2021-11-29/philip-morris-iqos-imports-barred-from-u-s-as-deadline-passes [accessed 2024-09-11]
10. PMI and BAT announce global patent settlement. Philip Morris International. 2024. URL: https://www.pmi.com/investor-relations/press-releases-and-events/press-releases-overview/press-release-details/?newsId=27101 [accessed 2024-03-10]
11. Hair EC, Bennett M, Sheen E, Cantrell J, Briggs J, Fenn Z, et al. Examining perceptions about IQOS heated tobacco product: consumer studies in Japan and Switzerland. Tob Control 2018;27(Suppl 1):s70-s73 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054322] [Medline: 29764957]
12. Kim M, Watkins SL, Koester KA, Mock J, Kim HC, Olson S, et al. Unboxed: US young adult tobacco users' responses to a new heated tobacco product. Int J Environ Res Public Health 2020;17(21):8108 [FREE Full text] [doi: 10.3390/ijerph17218108] [Medline: 33153143]
13. East KA, Tompkins CNE, McNeill A, Hitchman SC. 'I perceive it to be less harmful, I have no idea if it is or not': a qualitative exploration of the harm perceptions of IQOS among adult users. Harm Reduct J 2021;18(1):42 [FREE Full text] [doi: 10.1186/s12954-021-00490-8] [Medline: 33849549]
14. Watts C, Burton S, Freeman B. Creating a market for IQOS: analysis of Philip Morris' strategy to introduce heated tobacco products to the Australian consumer market. Tob Control 2022;31(3):458-463. [doi: 10.1136/tobaccocontrol-2020-056057] [Medline: 33191270]
15. Kim J, Yu H, Lee S, Paek YJ. Awareness, experience and prevalence of heated tobacco product, IQOS, among young Korean adults. Tob Control 2018;27(Suppl 1):s74-s77 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054390] [Medline: 30158210]

XSL•FO
RenderX

16. PMI promotion of IQOS using FDA MRTP order. Tobacco Tactics. 2021. URL: https://tobaccotactics.org/article/pmi-iqos-fda-mrtp-order/ [accessed 2023-12-10]
17. Harris JK, Moreland-Russell S, Choucair B, Mansour R, Staub M, Simmons K. Tweeting for and against public health policy: response to the Chicago Department of Public Health's electronic cigarette Twitter campaign. J Med Internet Res 2014;16(10):e238 [FREE Full text] [doi: 10.2196/jmir.3622] [Medline: 25320863]
18. Allem JP, Escobedo P, Chu KH, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on Twitter to e-cigarette education. Tob Control 2017;26(2):226-229 [FREE Full text] [doi: 10.1136/tobaccocontrol-2015-052757] [Medline: 26956467]
19. Kirkpatrick MG, Dormanesh A, Rivera V, Majmundar A, Soto DW, Chen-Sankey JC, et al. #FlavorsSaveLives: an analysis of Twitter posts opposing flavored e-cigarette bans. Nicotine Tob Res 2021;23(8):1431-1435 [FREE Full text] [doi: 10.1093/ntr/ntaa276] [Medline: 33394024]
20. Beard TA, Donaldson SI, Unger JB, Allem JP. Examining tobacco-related social media research in government policy documents: systematic review. Nicotine Tob Res 2024;26(4):421-426. [doi: 10.1093/ntr/ntad172] [Medline: 37712553]
21. Number of X (formerly Twitter) users worldwide from 2019 to 2024. Statista. 2024. URL: https://www.statista.com/statistics/303681/twitter-users-worldwide/ [accessed 2024-03-25]
22. Perrin A, Anderson M. Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center. 2019. URL: https://www.pewresearch.org/short-reads/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/ [accessed 2024-02-06]
23. Social media fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2024-02-06]
24. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2024-02-06]
25. Barker JO, Vassey J, Chen-Sankey JC, Allem JP, Cruz TB, Unger JB. Categorizing IQOS-related Twitter discussions. Int J Environ Res Public Health 2021;18(9):4836 [FREE Full text] [doi: 10.3390/ijerph18094836] [Medline: 33946546]
26. Zou C, Wang X, Xie Z, Li D. Perceptions of the IQOS heated tobacco product on Twitter in the United States. Front Commun 2021;6(163):728604. [doi: 10.3389/fcomm.2021.728604]
27. Abroms LC, Wysota CN, Tosakoon S, Khayat A, Duan Z, Bar-Zeev Y, et al. Industry marketing of tobacco products on social media: case study of Philip Morris International's IQOS. Tob Control 2024;33(4):518-524. [doi: 10.1136/tc-2022-057833] [Medline: 36927516]
28. Jackler RK, Ramamurthi D, Axelrod AK, Jung JK, Louis-Ferdinand NG, Reidel JE. Global marketing of IQOS: the Philip Morris campaign to popularize "Heat Not Burn" tobacco. Stanford Medicine Health Care. 2020. URL: https://stanfordhealthcare.org/publications/837/837223.html [accessed 2024-02-21]
29. Allem JP, Dharmapuri L, Unger JB, Cruz TB. Characterizing JUUL-related posts on Twitter. Drug Alcohol Depend 2018;190:1-5 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.05.018] [Medline: 29958115]
30. Galimov A, Vassey J, Galstyan E, Unger JB, Kirkpatrick MG, Allem JP. Ice flavor-related discussions on Twitter: content analysis. J Med Internet Res 2022;24(11):e41785 [FREE Full text] [doi: 10.2196/41785] [Medline: 36449326]
31. Galimov A, Kirkpatrick MG, Vassey J, Galstyan E, Smith A, Allem JP, et al. Oral nicotine gum discussions on Twitter: content analysis. Nicotine Tob Res 2024;26(4):503-507. [doi: 10.1093/ntr/ntad190] [Medline: 37791822]
32. Social nedia and news fact sheet. Pew Research Center. 2023. URL: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/ [accessed 2024-02-06]
33. News platform fact sheet. Pew Research Center. 2023. URL: https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/ [accessed 2024-01-15]
34. Laestadius L, van Hoorn K, Vassey J, Ozga J, Stanton C, Li D, et al. Tobacco, nicotine and counter-marketing promotions using Instagram's branded content tool. Tob Control 2023 [FREE Full text] [doi: 10.1136/tc-2023-058301] [Medline: 38160057]
35. Zheng X, Li W, Wong SW, Lin HC. Social media and e-cigarette use among US youth: longitudinal evidence on the role of online advertisement exposure and risk perception. Addict Behav 2021;119:106916. [doi: 10.1016/j.addbeh.2021.106916] [Medline: 33798917]
36. Pokhrel P, Fagan P, Herzog TA, Laestadius L, Buente W, Kawamoto CT, et al. Social media e-cigarette exposure and e-cigarette expectancies and use among young adults. Addict Behav 2018;78:51-58 [FREE Full text] [doi: 10.1016/j.addbeh.2017.10.017] [Medline: 29127784]
37. Lee S, Han DH, Chow A, Seo DC. A prospective longitudinal relation between elevated use of electronic devices and use of electronic nicotine delivery systems. Addict Behav 2019;98:106063. [doi: 10.1016/j.addbeh.2019.106063] [Medline: 31377448]
38. Albarracin D, Romer D, Jones C, Hall Jamieson K, Jamieson P. Misleading claims about tobacco products in YouTube videos: experimental effects of misinformation on unhealthy attitudes. J Med Internet Res 2018;20(6):e229 [FREE Full text] [doi: 10.2196/jmir.9959] [Medline: 29959113]
39. Kong G, Laestadius L, Vassey J, Majmundar A, Stroup AM, Meissner HI, et al. Tobacco promotion restriction policies on social media. Tob Control 2024;33(3):398-403 [FREE Full text] [doi: 10.1136/tc-2022-057348] [Medline: 36328589]

XSL•FO
RenderX

40. Berg CJ, Romm KF, Bar-Zeev Y, Abroms LC, Klinkhammer K, Wysota CN, et al. IQOS marketing strategies in the USA before and after US FDA modified risk tobacco product authorisation. Tob Control 2023;32(4):418-427 [FREE Full text] [doi: 10.1136/tobaccocontrol-2021-056819] [Medline: 34667105]
41. Khayat A, Berg CJ, Levine H, Rodnay M, Abroms L, Romm KF, et al. PMI's IQOS and cigarette ads in Israeli media: a content analysis across regulatory periods and target population subgroups. Tob Control 2024;33(e1):e54-e61 [FREE Full text] [doi: 10.1136/tc-2022-057671] [Medline: 36418166]
42. Gilmore AB, Braznell S. US regulator adds to confusion around heated tobacco products. BMJ 2020;370:m3528 [FREE Full text] [doi: 10.1136/bmj.m3528] [Medline: 32938588]
43. Tattan-Birch H, Hartmann-Boyce J, Kock L, Simonavicius E, Brose L, Jackson S, et al. Heated tobacco products for smoking cessation and reducing smoking prevalence. Cochrane Database Syst Rev 2022;1(1):CD013790 [FREE Full text] [doi: 10.1002/14651858.CD013790.pub2] [Medline: 34988969]
44. Kim M, Ling PM, Ramamurthi D, Halpern-Felsher B. Youth's perceptions of e-cigarette advertisements with cessation claims. Tob Regul Sci 2019;5(2):94-104 [FREE Full text] [doi: 10.18001/TRS.5.2.1] [Medline: 31840040]
45. Klein EG, Berman M, Hemmerich N, Carlson C, Htut S, Slater M. Online e-cigarette marketing claims: a systematic content and legal analysis. Tob Regul Sci 2016;2(3):252-262 [FREE Full text] [doi: 10.18001/TRS.2.3.5] [Medline: 27446984]
46. Grana RA, Ling PM. "Smoking revolution": a content analysis of electronic cigarette retail websites. Am J Prev Med 2014;46(4):395-403 [FREE Full text] [doi: 10.1016/j.amepre.2013.12.010] [Medline: 24650842]
47. Marketing granted order letter from FDA CTP to Philip Morris Products S.A. U.S. Food and Drug Administration. 2020. URL: https://www.fda.gov/media/144700/download [accessed 2023-07-01]
48. Prevent All Cigarette Trafficking (PACT) Act. U.S. Bureau of Alcohol, Tobacco, Firearms and Explosives. 2023. URL: https://www.atf.gov/alcohol-tobacco/prevent-all-cigarette-trafficking-pact-act [accessed 2023-07-01]
49. Freeman B, Watts C, Astuti PAS. Global tobacco advertising, promotion and sponsorship regulation: what's old, what's new and where to next? Tob Control 2022;31(2):216-221. [doi: 10.1136/tobaccocontrol-2021-056551] [Medline: 35241591]
50. Allem JP, Dormanesh A, Majmundar A, Rivera V, Chu M, Unger JB, et al. Leading topics in Twitter discourse on JUUL and puff bar products: content analysis. J Med Internet Res 2021;23(7):e26510 [FREE Full text] [doi: 10.2196/26510] [Medline: 34279236]
51. Robichaud MO, Puryear T, Cohen JE, Kennedy RD. How media stories in low- and middle-income countries discussed the U.S. Food and Drug Administration's modified risk tobacco product order for IQOS. Nicotine Tob Res 2023;25(10):1659-1666 [FREE Full text] [doi: 10.1093/ntr/ntad092] [Medline: 37310968]
52. Philip Morris International reports 2023 second-quarter and first-half results. Financial Post. 2023. URL: https://financialpost.com/pmn/business-wire-news-releases-pmn/philip-morris-international-reports-2023-second-quarter-and-first-half-results [accessed 2024-09-11]
53. Bastos MT, Puschmann C, Travitzki R. Tweeting across hashtags: overlapping users and the importance of language, topics, and politics. 2013 Presented at: Proceedings of the 24th ACM Conference on Hypertext and Social Media; May 1, 2013; Paris, France. [doi: 10.1145/2481492.2481510]
54. Inclusion, information, and intersection: the truth about connecting with U.S. Latinos. The Neilson Company. 2021. URL: https://www.nielsen.com/wp-content/uploads/sites/2/2021/09/nielsen-2021-hispanic-diverse-insights-report-210682-D9.pdf [accessed 2023-09-29]
55. Allem JP, Rodriguez V, Pattarroyo M, Ramirez CM, Beard TA, Soto D, et al. Spanish-language tobacco-related posts on Twitter: content analysis. Nicotine Tob Res 2024;26(6):759-763. [doi: 10.1093/ntr/ntad220] [Medline: 37942524]
56. Vassey J, Valente T, Barker J, Stanton C, Li D, Laestadius L, et al. E-cigarette brands and social media influencers on Instagram: a social network analysis. Tob Control 2023;32(e2):e184-e191 [FREE Full text] [doi: 10.1136/tobaccocontrol-2021-057053] [Medline: 35131947]
57. Kumar A, Garg G. Sentiment analysis of multimodal twitter data. Multimed Tools Appl 2019;78(17):24103-24119. [doi: 10.1007/s11042-019-7390-1]

Abbreviations

BAT: British American Tobacco
ENDS: electronic nicotine delivery system
FDA: Food and Drug Administration
HTP: heated tobacco product
ITC: International Trade Commission
MRTP: modified-risk tobacco product
PMI: Philip Morris International

XSL•FO
RenderX

Edited by T Leung; submitted 30.10.23; peer-reviewed by Z Duan, C Berg; comments to author 07.01.24; revised version received 01.04.24; accepted 14.08.24; published 24.10.24.

Please cite as:
Kim M, Vassey J, Li D, Galimov A, Han E, Kirkpatrick MG, Stanton CA, Ozga JE, Lee S, Unger JB
Discussion of Heated Tobacco Products on Twitter Following IQOS's Modified-Risk Tobacco Product Authorization and US Import Ban: Content Analysis
J Med Internet Res 2024;26:e53938
URL: https://www.jmir.org/2024/1/e53938
doi:10.2196/53938
PMID:39446431

©Minji Kim, Julia Vassey, Dongmei Li, Artur Galimov, Eileen Han, Matthew G Kirkpatrick, Cassandra A Stanton, Jenny E Ozga, Sarah Lee, Jennifer B Unger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining Quitting Experiences on Quit Vaping Subreddits From 2015 to 2021: Content Analysis

Elexis Kierstead[1], MPH; Nathan Silver[1], PhD; Michael Amato[1], PhD

Truth Initiative, Washington, DC, United States

Corresponding Author:
Elexis Kierstead, MPH
Truth Initiative
900 G St NW
Washington, DC, 20006
United States
Phone: 1 2024545746
Email: lkierstead@truthinitiative.org

Abstract

Background: Despite the prevalence of vaping nicotine, most nicotine cessation research remains focused on smoking cigarettes. However, the lived experience of quitting smoking is different from quitting vaping. As a result, research examining the unique experiences of those quitting vaping can better inform quitting resources and cessation programs specific to e-cigarette use. Examining Reddit forums (ie, subreddits) dedicated to the topics of quitting vaping nicotine can provide insight into the discussion around experiences on quitting vaping. Prior literature examining limited discussions around quitting vaping on Reddit has identified the sharing of barriers and facilitators for quitting, but more research is needed to investigate the content comprehensively across all subreddits.

Objective: The objective of this study is to examine content across quit vaping subreddits since their inception to better understand quitting vaping within the context of the expanding nicotine market.

Methods: All posts from January 2015 to October 2021 were scraped from all quit vaping subreddits: r/QuittingJuul, r/QuitVaping, r/quit_vaping, and r/stopvaping (N=7110). Rolling weekly average post volume was calculated. A codebook informed by a latent Dirichlet allocation topic model was developed to characterize themes in a subsample of 695 randomly selected posts. Frequencies and percentages of posts containing each coded theme were assessed along with the number of upvotes and comments.

Results: Post volume increased across all subreddits over time, spiking from August – September of 2019 when vaping lung injury emerged. Just over 52% of posts discussed seeking social support and 16.83% discussed providing social support. Posts providing support received the most positive engagements (i.e. upvotes) of all coded categories. Posts also discussed physical and psychological symptoms of withdrawal (30.65% and 18.85%, respectively), strategies for quitting including: quitting cold turkey (38.33%), using alternative nicotine products (17%), and tapering down nicotine content (10.50%). Most posts shared a personal narrative (92.37%) and some discussed quit motivation (28.20%) and relapse (14.99%).

Conclusions: This work identifies a desire for peer-to-peer support for quitting vaping, which reinforces existing literature and highlights characteristics of quitting vaping specific to a changing nicotine product environment. Given that posts providing social support were the most upvoted, this suggests that subreddit contributors are seeking support from their peers when discussing quitting vaping. Additionally, this analysis shows the sharing of barriers and facilitators for quitting, supporting findings from prior exploration of quit vaping subreddits. Finally, quitting vaping in an ever-growing nicotine market has led to the evolution of vaping-specific quit methods such as tapering down nicotine content. These findings have direct implications for quit vaping product implementation and development.

(J Med Internet Res 2024;26:e52129) doi:10.2196/52129

KEYWORDS

quitting vaping, social media, tobacco policy; cessation; e-cigarette; electronic cigarette; smoking; vaping; cessation programs; social support; peer support

XSL•FO
RenderX

Introduction

The nicotine product landscape, which used to be dominated almost exclusively by cigarettes, has expanded over the last decade to include new noncombustible products varying in nicotine delivery method and product characteristics [1,2]. Nicotine vapes, that is, e-cigarettes, are currently the most popular tobacco product used by US youth and young adults and have increased in popularity among adult users [1-3]. The drastic increase in vaping nicotine is accompanied by a corresponding increase in an interest to quit [4]. While nicotine vaping cessation shares many similarities with cigarette smoking cessation, it also differs in important ways, including variability of products, social acceptability, place-based restrictions of product use, and reasons for quitting [5,6]. As the e-cigarette and nicotine product markets continue to rapidly evolve, it is critical to continue monitoring the evolving nature of cessation from these products to ensure that interventions and treatment programs are effective at providing users with the support they need.

Reddit is a highly trafficked forum-based social media site used by just below 20% of US adults [7], wherein users can post on a discussion forum about a particular topic, that is, subreddit, and other users can comment, upvote, or downvote the post. Upvotes and downvotes indicate agreement and support or disagreement and dislike, respectively, and determine the likelihood that content will appear in a user's feed. Reddit users are more likely to be male, young adults, and White [8]. Reddit provides a unique source of information, granting insight into lived experiences without external researcher intervention and observation [9-11]. Examining content on subreddits—user-run forums organized around a topic or interest group—specific to quitting vaping nicotine provides insight into individual experiences within the context of the expanding tobacco product market. Additionally, prior literature has established that social media sites, such as Reddit, can provide social support to users [12-14], and social support on social media has even been tied to increases in self-efficacy to quit tobacco [15]. Reddit's utility as a discussion-based forum and social media site that individuals may turn to for social support emphasizes the importance of understanding if and how those trying to quit vaping are using this platform.

The changing nicotine product landscape has resulted in the availability of a wide variety of nicotine products for which quitting experiences may differ. The rise of JUUL products led to the widespread popularity of vape products, peaking in 2018 when the US Food and Drug Administration (FDA) characterized youth vape use as an "epidemic" [16]. Tobacco policy developments responding to this epidemic resulted in shifts in the nicotine product market. After the outbreak of lung injury cases initially connected to nicotine vape products [17], the FDA restricted the sale of flavored pod-based nicotine vapes [16]. As a result, disposable flavored nicotine vapes not restricted by this policy rose to popularity, leading to further proliferation of new vaping products and brands [3,18,19]. The FDA also required all tobacco products that emerged to submit a premarket tobacco product application (PMTA) by September 9, 2021 [20]. Through this process, the FDA reviews the scientific evidence for risks and benefits to public health for each product and determines if they are permitted to market their products in the United States. If denied authorization to market their products, said products could be removed from the US market. PMTA review is an important regulatory process implemented to address the presence of previously unregulated e-cigarette products that continue to be sold broadly on the US market [21]. Following the regulatory scrutiny surrounding e-cigarettes, recent years have seen the emergence of other noncombustible nicotine products—such as Zyn and On!—gaining popularity on the marketplace, especially among current tobacco users [22-24]. Further examination of the experiences of those quitting vaping nicotine within the context of this ever-changing market is needed.

The experiences of quitting vaping nicotine and quitting smoking are different [25,26]. Unlike quitting smoking, those quitting vaping must overcome the appeal of flavors, the inherent convenience and discreteness of the product, and a reported lack of self-awareness of how often one is using the product [6]. Moreover, those quitting vaping report a lack of trusted information on product safety and greater social acceptability of product use as unique challenges compared to quitting cigarettes [6]. Further investigation into the nuances of the quitting experience for vape product users is needed to better inform quitting interventions.

Reddit, as an online discussion-based social media platform, provides candid insight about the quitting process, as this public discourse is conducted outside of a research setting. Reddit has previously been used as a data source to examine quitting vaping [27-29]. One study examined posts from the r/QuitVaping subreddit over a 4-week period in 2020, qualitatively evaluating the barriers and facilitators to use. This work identified withdrawal symptoms and the intensity of addiction as the biggest barriers, and distracting oneself and maintaining a positive mindset as the most effective facilitators [29]. In light of useful insights gained from previous examinations of the quitting experience on Reddit, there is value in expanding the scope of inquiry to longitudinally examine the growth of multiple online quitting communities on the platform, as well as the most common topics of conversation.

We examine all posts from all subreddits related to quitting vaping, encompassing the entire lifetime of these forums, thus allowing us to contextualize our findings within a developing marketplace, a unique contribution to the literature. We conducted a content analysis to summarize themes present in the posts using an inductively developed codebook. Through this work, we asked (1) How had the volume of use of quit vaping subreddits changed across time? and (2) What were the most common topics of conversation on these forums?

Methods

Procedure

To comprehensively examine how those trying to quit vaping use Reddit, we analyzed all posts on subreddits dedicated to quitting vaping nicotine. Post content, number of comments, and upvotes were used to examine broad trends, among other

XSL•FO
RenderX

metadata. Moreover, we used the *Gensim* package, developed by RARE Technologies Ltd, for Python (version 3.10; Python Software Foundation) to conduct an unsupervised machine learning analysis (latent Dirichlet allocation [LDA] topic modeling) of all posts to develop a codebook for human-coded content analysis on a random sample of posts. Descriptive analysis of user engagement with Reddit posts, including comments and upvotes (a platform feature used to crowdsource evaluations of the relevance of a post from forum members), was aggregated for each coded feature.

Sample and Measures

Using the Pushshift application programming interface, all posts (N=7110) and comments (N=41,019) were scraped from 4 quit vaping subreddits: r/QuittingJuul, r/QuitVaping, r/quit_vaping, and r/stopvaping. The posts spanned from January 12, 2015, the first post in r/quit_vaping, to October 20, 2021, following the PMTA deadline in September of 2021 and other seismic shifts in the nicotine product marketplace. Post text, timestamp, subreddit of origin, post author, number of upvotes, and number of comments were collected for each post.

Codebook Development

First, rolling weekly averages of post volume over time within and across all 4 subreddits were calculated. Second, post and comment totals were aggregated at the user level. Next, we developed a codebook to characterize themes across all Reddit posts. Whereas previous research focused on a specific time interval or subreddit [27-29], we examine the full corpus of texts (7110 posts) on all quit-related subreddits from their inception. Given this volume of texts, we used an inductive, data-driven method—LDA topic modeling—to perform an initial "first pass" through the data to identify common topics. In contrast to theory-driven approaches to content analysis, this inductive approach can help identify common topics of discussion on quitting-related subreddits that may not be accounted for by existing theory [30,31]. LDA is a dimension reduction procedure analogous to confirmatory factor analysis wherein a corpus of texts is stripped of syntactical structure, including the removal of *stop words* like "and," "or," and "but"; *stemming* to remove suffixes (eg, "ed" and "ing"); *tokenizing* to make each unique word a variable; and *vectorization* to create a sparse matrix that essentially dummy codes the presence or absence of every possible word in each individual text [32]. The LDA algorithm then identifies latent topics based on words that both co-occur and help to differentiate between topics. We specify the number of topics a priori based on a perplexity graph, which visualizes model fit as a function of the optimal number of topics. In our case, a 12-topic model had the highest perplexity [33]. LDA then produces probabilistic factor loadings for each latent topic indicative of the likelihood that a given text shares a latent topic with other texts based on the weighted values of the words in that text. Consistent with convention [34], we set a coherence threshold of 0.70 to identify the most highly weighted words and example texts with high factor loadings for each latent topic that we used to interpret the topic model (Table 1).

XSL•FO
RenderX

Table 1. Latent Dirichlet allocation (LDA) topic model–determined themes to inform content analysis from a sample of 7110 posts on quit vaping subreddits from January 2015 to October 2021.

Topic number	Topic name	Relevant words	Example
1	Asking for help	need, hard, stop, trying	"Please help I can't quit: I've been vaping for 4 years now and I seriously wish I never started. I've tried to quit but I can't. I don't even get buzzed anymore. How did everyone else quit? I've been thinking of buying a CBD vape or something with a low percent nic to slowly get off, would this help?"
2	Initiation and progression	hard, stop, trying, gum, habit, addiction, hitting	"I never smoked cigarettes before but in the height of the Juul era several years ago all of my friends switched to vaping. I picked up the habit."
3	Sleeplessness and brain fog	ago, never, thought, felt, else, worse, sleep, wake, brain	"Like I feel like I'm not a person like I don't exist. Like I don't want to do anything like the only thing I did today was browse my computer. I just have been laying on my bed because I feel so useless and tired. I can't focus for the life of me."
4	Quit aids	using, juice, nic, gum, cigarettes, symptoms, CBD	"Mine: lollipops, fidget cube, Bubly water, those QuitGo fake cigarettes, Nicolette lozenges. Still looking for more, I'm on day 6 What are your coping mechanisms?"
5	Quit motivation	work, people, health, news, long, life	"I'm 17 and I quit vaping because my 10 month old niece is moving in with me on the 15th and I want to be better for her and I'd like to be completely nicotine free before she lives here."
6	Consumption method	salts, pod, used, around, juice, vs, mg/ml	"GUIDE: Easy weaning off JUUL w/ diminishing nicotine salt strengths (w/ links): If you JUUL you may have noticed how absolutely ass-kickingly strong the nicotine salt solution in JUUL is compared to analogs and non-nicotine salt vaping. You may have switched from JUUL 5% to the new 3% pods and found them still too strong. JUUL doesn't offer anything below 3%."
7	Anxiety, respiratory or cardiovascular	chest, pain, heart, breath, shortness, attack, panic, tightness	"It's day 17 and I am having 2-3 panic attacks a day and dealing with depression. All of these symptoms onset the day I quit. Is this normal? And how long until it starts to go away"
8	Study recruitment	study, link, survey, research, based, users, mobile	"If you use **e-cigarettes or vape** and are **16-20 years old**, you may qualify for a **remote research study** with the Medical University of South Carolina."
9	Relapse circumstances	relapsed, drunk, cigars, spliffs, scare	"I'm just going to go back to college where all of my friends have nicotine and there's no way I'm going to say no when they offer it to me, especially when drunk."
10	Weight gain	weight, gain, hungry	"I started vaping 3mg about 3 years ago because I learned that nicotine was an appetite suppressant, and that sounded perfect to my (at the time) anorexic brain."
11	Self-help or journey metaphor or guidance	allen, carr, book	"Any advice is appreciated and explaining what to expect in the coming days will be very helpful to my journey"
12	Relational problems while quitting	freebase, husband, refused, snappy, smashed	"My partner who I live with still vapes regularly. So when I decided to try to quit I knew this would be a challenge, as having it around, in sight all the time would be hard."

LDA also has significant limitations, leading us to prefer not to rely on topic modeling alone. First, perplexity is an imprecise measure with the identified topics indicative of patterns of shared words that may or may not be interpretable. As a result, some “topics” identified by the model may not be topics in reality. Most importantly, although LDA is an effective and efficient way to sift through large amounts of text data for broader trends, the accuracy of any given factor loading is limited. While we can confidently say that a topic identified by the model is prevalent across the broader corpus of text, a given text that loads high on a latent topic has a relatively low likelihood of being about that topic. However, the inductive generation of topics proved useful in developing a codebook to exhaustively represent the data.

The topic model identified 12 latent topics that were refined by the authors to be human coded on a random sample of 695 posts. Based on established conventions for media-based content analyses [35], we reserved 10% (n=69) of our sample for establishing reliability. Two trained research assistants coded the reserved sample, resolving discrepancies via discussion until intercoder reliability for each coded category exceeded a minimum Cohen κ threshold of 0.70 [35]. Coded categories with low reliability were dropped from further analysis, resulting in 6 major themes, 3 of which contained subthemes, for a total of 13 codes. Codes were not mutually exclusive, allowing for multiple codes in the same post. Table 2 provides a full accounting of our codebook, criteria, and reliability information.

XSL•FO
RenderX

Table 2. Expert-determined, topic model–informed codebook for examining themes among a random sample of 695 posts across all quit vaping subreddits from January 2015 to October 2021 with reliability between trained coders.

Code	Definition	Example	Cohen κ
Social support	Text mentions or contains users interacting with one another in a socially supportive way, seeking or providing peer support, or mentioning social support in their vaping cessation efforts	"I believe in you!"	0.91
Seeking (SS[a])	Text contains asking or searching for support or advice from other Reddit users	"throw me your strategies for not juuling at parties"	0.91
Providing (SS)	Text contains an offer for advice or explicitly gives social support to other Reddit users, either with a clear recipient or to the general userbase	"Here's a few helpful tips"	0.84
Withdrawal	Text mentions withdrawal from vaping products, either in apprehension or as lived experience	"2 weeks! The only cravings now are psychological- when I get bored, or am in situations where I used to Juul heavily (like playing video games)"	0.88
Physical (W[b])	Text mentions physical symptoms of withdrawal from vaping products, including cravings, appetite changes, weight gain or loss, respiratory symptoms, cardiovascular symptoms, flu-like symptoms (ie, aches and pains, runny nose), nausea, cosmetic appearance (ie, skin or acne or bags under eyes)	"Overall, definitely feeling much better physically. Way less dehydrated, happier, sleeping better and feeling more energetic in the am."	0.86
Psychological (W)	Text mentions emotional or mental symptoms of withdrawal from vaping products, including anxiety or depression, brain fog or confusion, anger, relationship struggles, confusion between emotional and physical symptoms	"I feel amazing! The first 3 days were awful - mostly mental symptoms; irritable, tired, increased appetite, headache, extreme cravings."	0.87
Quit strategies	Text mentions strategies used to quit using vaping products, which can include seeking or providing guidance or life experience	"For the craving and withdrawals all I did was drink ice cold water and play basketball outside in my backyard after work to keep myself busy ."	0.97
Cold turkey (QS[c])	Text mentions completely ceasing use of tobacco or nicotine products at one point in time as a strategy for quitting vaping	"I quit 25 mg nic salts cold turkey about 5 days and 2 hours ago."	1
Tapering (QS)	Text mentions the slow reduction of nicotine content in e-liquid in an effort to quit or the slow reduction in use in a vaping product over time. Can also include tapering regarding frequency of use of nicotine products, regardless of change in e-liquid content (eg, individual using juul 5 days a week, tapering down to 3 days a week, and so on)	"I had previously dropped from 50 mg to 25 mg as the first step of my quitting journey."	1
Alternative products (QS)	Text mentions the use of alternate tobacco or nicotine products as a quit strategy (eg, nicotine gum, nicotine pouches, cigarettes, cannabis products, non-nicotine vaping devices, etc.) This includes FDA[d]-approved nicotine replacement therapy cessation products.	"Im chewing gum now but if im honest its not helping as much as it should"	0.87
Personal narrative	Text contains a personal story on the forum often containing lived experiences to either ask for or give guidance and support	"So I haven't been using JUUL exclusively for the past year since I started using electronic cigs, vapes, etc."	0.88
Quit motivation	Text contains mention of sources of motivation to quit vaping or a personal assessment of how motivated they feel to quit. This sentiment can either be positive or negative, discussing why they are or aren't motivated to quit	"Those thoughts kept me motivated my first three days"	0.97
Relapse	Text mentions beginning to vape again after intending to quit using the product, or after a period of sustained abstinence from vaping, either as a lived experience or in apprehension	"Trying to quit but keep slipping up"	1

[a]SS: social support.

[b]W: withdrawal.

[c]QS: quitting strategy.

[d]FDA: Food and Drug Administration.

Statistical Analyses

Frequencies and percentages of posts containing each coded theme were tallied along with measures of central tendency for the number of upvotes and comments within each coded theme. The number of comments was counted for each post; however, these comments were not analyzed for content.

XSL•FO
RenderX

Ethical Considerations

Study data were collected from a publicly available data source, Reddit; therefore, no ethics approval was required. The data were anonymized for analysis and publication.

Results

Post Volume Over Time

Across all 4 subreddits, rolling weekly average posts showed that overall use of these subreddits had grown over time. After an initial peak in overall post volume from August to September 2019, posts trended upward over time (Figure 1). When examining post volume by subreddit, we identified that r/QuittingJuul was initially the most frequently used subreddit, until May to June 2020, when r/QuitVaping overtook r/QuittingJuul in post volume. The other 2 subreddits, r/quit_vaping and r/stopvaping, were infrequently used (Figure 2).

Figure 1. Rolling weekly average posts aggregated across 4 quit vaping subreddits from January 2015 to October 2021 (N=7110). FDA: Food and Drug Administration; PMTA: premarket tobacco product application.

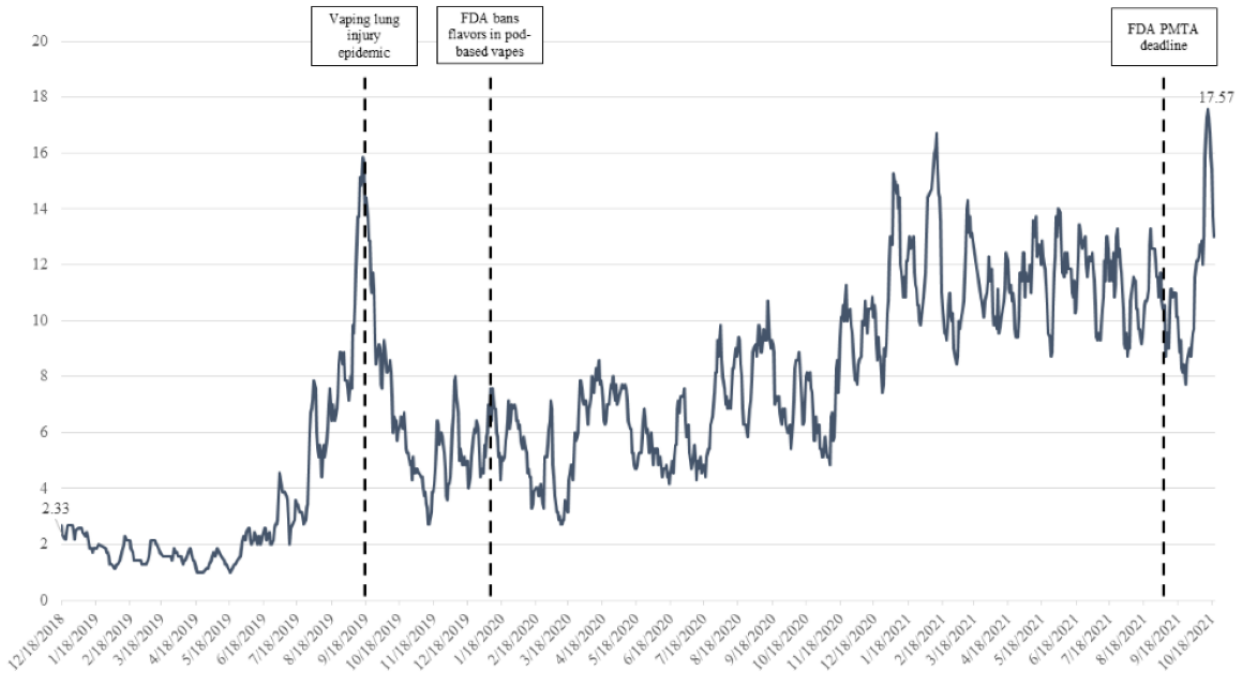

XSL•FO
RenderX

Figure 2. Post volume across quit vaping subreddits from January 2015 to October 2021 (N=7110). FDA: Food and Drug Administration; PMTA: premarket tobacco product application.

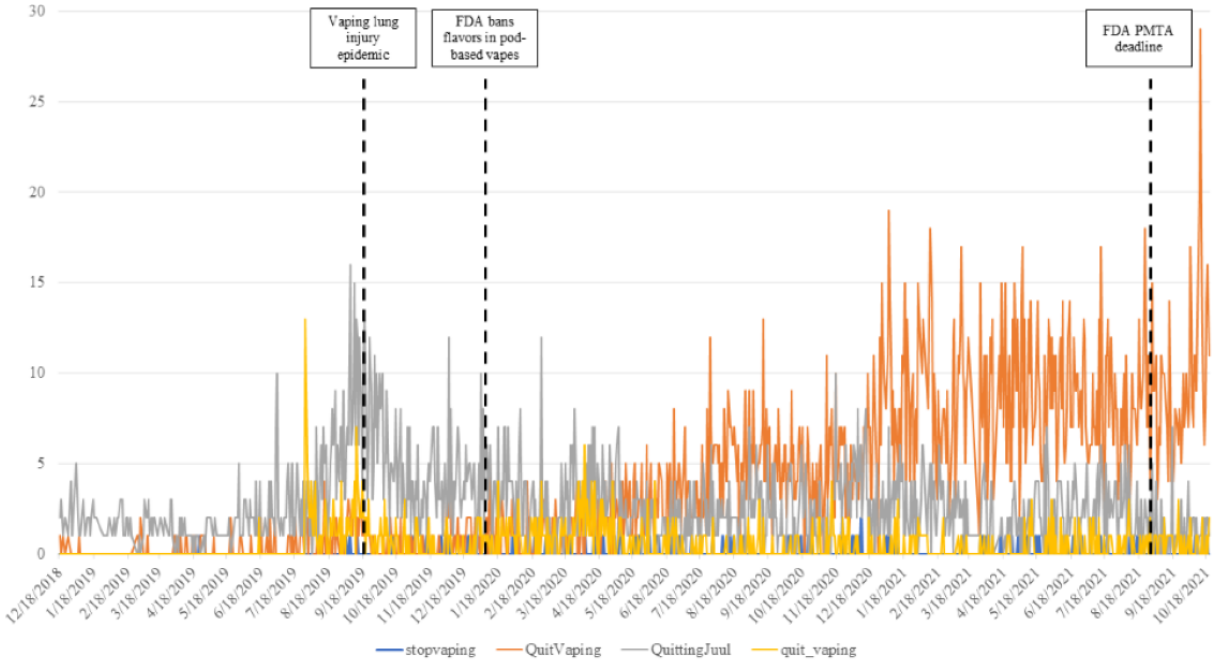

User Contribution Characteristics

The sample consisted of 7110 total posts by 3197 unique users. Notably, 2336 posts were not affiliated with an author name either because the associated account was deleted or the user requested that Pushshift delete their data. Most users posted infrequently (median 1, IQR 1-1), with only 23.8% (761/3197) of users posting more than once and 1.9% (60/3197) of users posting more than 5 times. The posts garnered 41,019 comments (median 4, IQR 2-6 comments per post). A total of 7873 unique users posted a median of 2 (IQR 1-4) comments, each with 23.9% (n=1882) of users posting 5 or more comments and 9.7% (n=764) posting 10 or more comments. However, 12 users each posted more than 100 comments. Also, 5396 comments were not associated with a username, suggesting there may have been more users commenting than those we documented.

Content Analysis

Table 3 summarizes all 13 codes, including the frequency at which they appeared in the sample, as well as median upvotes and median comments.

XSL•FO
RenderX

Table 3. Frequency and overall percent of coded sample of 695 quit vaping–related subreddit posts from January 2015 to October 2021 containing each theme from a topic-model derived, expert-determined codebook.

Code	Frequency, n (%)	Upvotes, median (IQR)	Comments, median (IQR)
Social support	467 (67.19)	5 (3-10)	4 (2-6)
Seeking (SS[a])	362 (52.09)	4 (2-7)	4 (2-7)
Providing (SS)	117 (16.83)	15 (8-22)	4 (2-6)
Withdrawal	291 (41.87)	6 (3-11)	4 (2-6)
Physical (W[b])	213 (30.65)	6 (3-12)	4 (2-6)
Psychological (W)	131 (18.85)	7 (3-13)	4 (2-6)
Quit strategies	389 (55.97)	6 (3-13)	4 (2-6)
Cold turkey (QS[c])	266 (38.33)	7 (3-13)	4 (2-6)
Tapering (QS)	73 (10.50)	6 (3-10)	4 (2-6)
Alternative products (QS)	118 (17)	5 (3-11)	4 (2-6)
Personal narrative	642 (92.37)	6 (3-13)	4 (2-6)
Quit motivation	196 (28.2)	7 (3-13)	4 (2-6)
Relapse	104 (14.99)	4 (3-9)	4 (2-5.5)

[a]SS: social support.

[b]W: withdrawal.

[c]QS: quitting strategy.

Social Support

Over two-thirds of posts contained mentions of *social support* (467/695, 67.19%). We found that 52.09% (362/695) of posts in the sample were *seeking social support*, making up 77.52% (362/467) of the posts coded as *social support*. Meanwhile, 16.83% (117/695) of the total coded posts and 25.05% (117/467) of the *social support* posts *provided social support*. Posts *seeking social support* received a median of 4 (IQR 2-7) upvotes and 4 (IQR 2-7) comments, while posts *providing social support* received a median of 15 (IQR 8-22) upvotes and 4 (IQR 2-6) comments.

Withdrawal

We found 41.87% (291/695) of posts discussed *withdrawal* from vaping nicotine. Of the total coded sample, 30.65% (213/695) of posts discussed *physical withdrawal* symptoms, making up 73.2% (213/291) of posts mentioning withdrawal. To a lesser extent, 18.85% (131/695) of the total sample and 45.02% (131/291) of *withdrawal* posts discussed *psychological withdrawal* symptoms. Examples of physical withdrawal symptoms include cravings, heart palpitations, mouth pain, nausea, or changes to appetite, while psychological symptoms of withdrawal include increased anxiety, changes to mood, or brain fog. Mentions of *physical withdrawal* received a median of 6 (IQR 3-12) upvotes and 4 (IQR 2-6) comments, while *psychological withdrawal* mentions received a median of 7 (IQR 3-13) upvotes and 4 (IQR 2-6) comments.

Quit Strategies

We found 55.97% (389/695) of posts discussed *strategies for quitting*. Of the total sample, 38.33% (266/695) of posts discussed *quitting cold turkey*, making up 68.38% (266/389) of the posts mentioning *quit strategies*. About 10.50% (73/695) of the full coded sample and 18.77% (73/389) of the *strategies for quitting* posts discussed *tapering down nicotine content* as a method of quitting. Approximately 17.00% (118/695) of the full sample and 30.33% (118/389) of the *strategies for quitting* posts mentioned *using alternative products* to quit. Notably, post authors reported use of both FDA-approved and unapproved quit aids ranging from nicotine replacement therapy (NRT) to recreational oral nicotine products (eg, Zyn). Posts discussing *quitting cold turkey* received a median of 7 (IQR 3-13) upvotes and 4 (IQR 2-6) comments, posts discussing *tapering down* received a median of 6 (IQR 3-10) upvotes and 4 (IQR 2-6) comments, and posts discussing *alternative products* received a median of 5 (IQR 3-11) upvotes and 4 (IQR 2-6) comments.

Personal Narratives, Quit Motivation, and Relapse

The majority of posts in the sample (642/695, 92.37%) contained *personal narratives* drawing on and sharing experiences of the post author in the text. These posts received a median of 6 (IQR 3-13) upvotes and 4 (IQR 2-6) comments.

We found 28.2% (196/695) of the sample contained discussion around *motivation to quit* vaping. Themes around health and wellness were commonly found as motivating reasons for quitting vaping. Posts discussing *quit motivation* received a median of 7 (IQR 3-13) upvotes and 4 (IQR 2-6) comments.

We found that 14.99% (104/695) of posts mentioned *relapsing* vaping after a period of cessation. These posts mentioning *relapse* garnered a median of 4 (IQR 3-9) upvotes and 4 (IQR 2-5.5) comments.

XSL•FO
RenderX

Discussion

Overview

Using data from all quit vaping subreddits, this study shows that post volume has been increasing over time. This finding emphasizes the increasing use of online forums for seeking quit vaping resources within the context of increasing population vape use. This work also identifies that people are using these forums to both request and provide social support from their peers. Broadly, these findings are further evidence that along with increasing vape product use, there is also an increasing desire to quit. Reddit forums are serving as a digital community where individuals gather to seek support to quit vaping and to provide quitting support to others. The presence of both support-seeking and support-providing behaviors suggests the forum encompasses variable types of engagement regarding the quitting process.

The increasing post volume on quit vaping subreddits reflects a growing public appetite for quitting resources as vaping continues to proliferate [1,2]. This increase in post volume also suggests a desire for the peer-to-peer social support offered by Reddit forums. We find this reflected in the post content, with 2 out of every 3 posts seeking or providing social support. It is also important to note that the posts *providing* social support garnered 3 times as many upvotes, that is, positive engagement, as the posts *seeking* social support, suggesting many Reddit users may be seeking out motivational support from their peers. Prior literature has established that peer-to-peer social support can be a valuable tool for tobacco cessation [14,25,36-38]. Furthermore, this type of peer support is replicated in the messaging design of successful text-to-quit vaping interventions [39] and is in line with prior literature around vaping cessation [25]. This analysis highlights the increasing desire for peer-to-peer social support in the quit vaping experience, outlining an important consideration for effective intervention implementation.

This study supports research suggesting that societal events shape individual quitting behaviors. While attribution of the observed patterns to discrete events is not possible with the current dataset, we offer the following speculative explanations based on temporally co-occurring events for investigation in future research. The sharp rise in post volume in the Fall of 2019 suggests that interest in quitting vaping spiked around the time when lung injury initially connected to vaping was making national headlines [40,41]. Similarly, we also note that, over time, the most trafficked subreddit changed from r/QuittingJuul to r/QuitVaping. This shift mirrors the market expansion that occurred after JUUL's sale was heavily restricted and product users began picking up disposable products that continue to be sold in a variety of flavors [18,19]. Finally, we note that post volume did increase from 2020 onward, after the dawn of the COVID-19 pandemic. Studies show that while some increased their use of tobacco during the pandemic, others felt motivated to quit during this time, likely turning to online accessible quit vaping resources [42]. Although few implications can be drawn from these overlapping phenomena, further research should investigate the extent to which information-seeking behavior is responsive to tobacco control policy events and media coverage. This could encourage interventionists to utilize future policy and media events to their advantage, seeking out potential quitters at these shared moments.

When examining the content of the posts, our work builds on prior literature, finding that battling withdrawal is a significant barrier to quitting while improving health is a frequently listed motivator [29]. Coping with withdrawal and other common topics, including quit strategies and relapse, align with evidence-based suggested topics for discussion with tobacco users from the US Public Health Service Clinical Practice Guidelines for Treating Tobacco Use and Dependence [43] as well as previous findings from other digital cessation communities [44].

In addition to confirming these trends, we also note that users shared specific strategies for quitting vaping that were contextualized inside of an expanding nicotine market. Specifically, participants mentioned tapering down their nicotine content, as others have previously observed [29]. Some participants went so far as to provide extensive guides to this stepping-down process. This methodology is specific to quitting vaping but bears similarity to clinical guidance provided for tapering nicotine with FDA-approved NRT [43]. Additionally, when considering strategies for quitting, participants mentioned using not only NRT designed for smoking cessation but also products not approved for cessation, such as newly popular nicotine pouch products. More research is needed to identify how those seeking to quit vaping are viewing these products within the context of quitting or cutting back on nicotine, as compared to FDA-approved NRT.

These findings have direct implications for the development and implementation of quit vaping resources. The characteristics of the Reddit forum itself could be indicative of what individuals might seek out from a quit vaping resource, for example, information sharing and social support. The content also indicates that quit vaping support messaging should be tailored to the way these products are used in an expanding nicotine market, such as tapering down nicotine content or product substitution. Additionally, beyond lessons learned from the content itself, the growing presence on these subreddits indicates a growing population of individuals who want to quit vaping and are turning to social media sites for help. These individuals have created their own quitting support, and many could likely benefit from quit programs. Future research should investigate ties between quit program use and tobacco policy events.

Limitations

This work has several limitations. Not all information was complete across post, comment, and user datasets as some posts or accounts had been removed by users. There is subjectivity in inductively creating and assigning codes for a dataset due to the inherent biases of those conducting the interpretation. The preliminary topic model was introduced to reduce the likelihood of these inherent biases affecting the interpretation of the data in the development of coded categories. Finally, we are attempting to understand a population based on a convenience sample; those participating in quit vaping subreddits may not be representative of all individuals wishing to quit vaping, nor

XSL•FO
RenderX

all individuals participating in other digital cessation communities, which may have different norms. Specifically, Reddit users are disproportionately White young adult men, limiting the generalizability of the results to the whole population, as is typical of a convenience sample.

Conclusions

We find an increasing desire for peer-to-peer support for quitting vaping over time on Reddit, indicating a greater need for quit vaping resources as vaping continues to gain popularity in the United States [1-3]. We also note that the content on the quit vaping subreddits affirms the sharing of barriers and facilitators to quitting identified by prior literature but also highlights important characteristics of quitting vaping specific to an expanding and changing nicotine product environment, providing critical guidance for the development of quit vaping tools.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

EK was involved in conceptualization, formal analysis, and writing—original draft preparation; NS was involved in conceptualization, data curation, formal analysis, visualization, and writing—review and editing; and MA was involved in supervision and writing—review and editing.

Conflicts of Interest

None declared.

References

1. Cornelius ME, Loretan CG, Jamal A, Davis Lynn BC, Mayer M, Alcantara IC, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023;72(18):475-483 [FREE Full text] [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
2. Gentzke AS, Wang TW, Cornelius M, Park-Lee E, Ren C, Sawdey MD, et al. Tobacco product use and associated factors among middle and high school students - national youth tobacco survey, United States, 2021. MMWR Surveill Summ 2022;71(5):1-29 [FREE Full text] [doi: 10.15585/mmwr.ss7105a1] [Medline: 35271557]
3. Boakye E, Dzaye O, Erhabor J, Osuji N, Obisesan O, Osei AD, et al. Impact of the food and drug administration enforcement policy on flavored e-cigarettes on the online popularity of disposable e-cigarettes: analyses of Google search query data. BMC Public Health 2022;22(1):1937 [FREE Full text] [doi: 10.1186/s12889-022-14367-3] [Medline: 36258175]
4. Cuccia AF, Patel M, Amato MS, Stephens DK, Yoon SN, Vallone DM. Quitting e-cigarettes: quit attempts and quit intentions among youth and young adults. Prev Med Rep 2021;21:101287 [FREE Full text] [doi: 10.1016/j.pmedr.2020.101287] [Medline: 33489721]
5. Amato MS, Bottcher MM, Cha S, Jacobs MA, Pearson JL, Graham AL. "It's really addictive and I'm trapped": a qualitative analysis of the reasons for quitting vaping among treatment-seeking young people. Addict Behav 2021;112:106599 [FREE Full text] [doi: 10.1016/j.addbeh.2020.106599] [Medline: 32950927]
6. Sanchez S, Kaufman P, Pelletier H, Baskerville B, Feng P, O'Connor S, et al. Is vaping cessation like smoking cessation? A qualitative study exploring the responses of youth and young adults who vape e-cigarettes. Addict Behav 2021;113:106687. [doi: 10.1016/j.addbeh.2020.106687] [Medline: 33045643]
7. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 7. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2023-08-01]
8. Barthel M, Stocking G, Holcomb J, Mitchell A. 1. Reddit news users more likely to be male, young and digital in their news preferences. Pew Research Center. 2016 Feb 25. URL: https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/ [accessed 2023-08-01]
9. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with reddit data. Soc Sci Comput Rev 2019;39(5):943-960. [doi: 10.1177/0894439319893305]
10. Jamnik M, Lane D. The use of reddit as an inexpensive source for high-quality data. Pract Assess Res Eval 2019;22(1):1-10 [FREE Full text]
11. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying reddit: a systematic overview of disciplines, approaches, methods, and ethics. Soc Media Soc 2021;7(2):1-14. [doi: 10.1177/20563051211019004]
12. Selkie E, Adkins V, Masters E, Bajpai A, Shumer D. Transgender adolescents' uses of social media for social support. J Adolesc Health 2020;66(3):275-280. [doi: 10.1016/j.jadohealth.2019.08.011] [Medline: 31690534]

XSL•FO
RenderX

13. Hayes RA, Carr CT, Wohn DY. It's the audience: differences in social support across social media. Soc Media Soc 2016;2(4):205630511667889. [doi: 10.1177/2056305116678894]
14. Cheung YTD, Chan CHH, Wang MP, Li HCW, Lam T. Online social support for the prevention of smoking relapse: a content analysis of the WhatsApp and Facebook social groups. Telemed J E Health 2017;23(6):507-516. [doi: 10.1089/tmj.2016.0176] [Medline: 27911654]
15. Phua J. Participating in health issue-specific social networking sites to quit smoking: how does online social interconnectedness influence smoking cessation self-efficacy? J Commun 2013;63(5):933-952. [doi: 10.1111/jcom.12054]
16. Statement from FDA Commissioner Scott Gottlieb, M.D., on new steps to address epidemic of youth e-cigarette use. U.S. Food and Drug Administration. 2018. URL: https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-new-steps-address-epidemic-youth-e-cigarette-use [accessed 2022-07-21]
17. Gartner C, Bonevski B, Hall W. Miscommunication about the causes of the US outbreak of lung diseases in vapers by public health authorities and the media. Drug Alcohol Rev 2020;39(1):3-6. [doi: 10.1111/dar.13024] [Medline: 31930631]
18. Ali FRM, Diaz MC, Vallone D, Tynan MA, Cordova J, Seaman EL, et al. e-Cigarette unit sales, by product and flavor type - United States, 2014-2020. MMWR Morb Mortal Wkly Rep 2020;69(37):1313-1318 [FREE Full text] [doi: 10.15585/mmwr.mm6937e2] [Medline: 32941416]
19. Wang TW, Gentzke AS, Neff LJ, Glidden EV, Jamal A, Park-Lee E, et al. Disposable e-cigarette use among U.S. youth - an emerging public health challenge. N Engl J Med 2021;384(16):1573-1576. [doi: 10.1056/NEJMc2033943] [Medline: 33725431]
20. Hammond H. FDA pushes back deadline to review PMTAs. CSP Daily News. URL: https://cspdailynews.com/tobacco/fda-pushes-back-deadline-review-pmtas [accessed 2023-08-01]
21. Products C for T. Premarket tobacco product applications. U.S. Food and Drug Administration. 2023. URL: https://www.fda.gov/tobacco-products/market-and-distribute-tobacco-product/premarket-tobacco-product-applications [accessed 2024-01-30]
22. Felicione NJ, Schneller LM, Goniewicz ML, Hyland AJ, Cummings KM, Bansal-Travers M, et al. Oral nicotine product awareness and use among people who smoke and vape in the U.S. Am J Prev Med 2022;63(4):611-618. [doi: 10.1016/j.amepre.2022.04.019] [Medline: 35667923]
23. Hrywna M, Gonsalves NJ, Delnevo CD, Wackowski OA. Nicotine pouch product awareness, interest and ever use among US adults who smoke, 2021. Tob Control 2023;32(6):782-785 [FREE Full text] [doi: 10.1136/tobaccocontrol-2021-057156] [Medline: 35217596]
24. Marynak KL, Wang X, Borowiecki M, Kim Y, Tynan MA, Emery S, et al. Nicotine pouch unit sales in the US, 2016-2020. J Am Med Assoc 2021;326(6):566-568 [FREE Full text] [doi: 10.1001/jama.2021.10366] [Medline: 34374729]
25. Berg CJ, Krishnan N, Graham AL, Abroms LC. A synthesis of the literature to inform vaping cessation interventions for young adults. Addict Behav 2021;119:106898 [FREE Full text] [doi: 10.1016/j.addbeh.2021.106898] [Medline: 33894483]
26. Villanti AC, West JC, Klemperer EM, Graham AL, Mays D, Mermelstein RJ, et al. Smoking-cessation interventions for U.S. young adults: updated systematic review. Am J Prev Med 2020;59(1):123-136 [FREE Full text] [doi: 10.1016/j.amepre.2020.01.021] [Medline: 32418800]
27. Pepper JK, Zitney LV, Preiss A, Bukowski T, Szczypka G, Kim AE. Can social media monitoring help identify the next EVALI? An examination of Reddit posts about vitamin E acetate and dank vapes. Drug Alcohol Depend 2022;230:109193 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.109193] [Medline: 34915270]
28. Sharma R, Wigginton B, Meurk C, Ford P, Gartner CE. Motivations and limitations associated with vaping among people with mental illness: a qualitative analysis of reddit discussions. Int J Environ Res Public Health 2016;14(1):7 [FREE Full text] [doi: 10.3390/ijerph14010007] [Medline: 28025516]
29. Struik L, Yang Y. e-Cigarette cessation: content analysis of a quit vaping community on reddit. J Med Internet Res 2021;23(10):e28303 [FREE Full text] [doi: 10.2196/28303] [Medline: 34694229]
30. Rodriguez MY, Storer H. A computational social science perspective on qualitative data exploration: using topic models for the descriptive analysis of social media data*. J Technol Hum Serv 2019;38(1):54-86. [doi: 10.1080/15228835.2019.1616350]
31. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. J Gen Intern Med 2020;35(7):2244-2247 [FREE Full text] [doi: 10.1007/s11606-020-05889-w] [Medline: 32399912]
32. Blei DM. Latent Dirichlet allocation. Geeks for Geeks. 2003. URL: https://www.geeksforgeeks.org/latent-dirichlet-allocation/ [accessed 2024-09-18]
33. Hagen L. Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models? Inf Process Manage 2018;54(6):1292-1307. [doi: 10.1016/j.ipm.2018.05.006]
34. Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. 2014 Presented at: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces; June 27, 2014; Baltimore, MD p. 63-70. [doi: 10.3115/v1/w14-3110]
35. Neuendorf KA. The Content Analysis Guidebook. London, United Kingdom: SAGE; 2017.

XSL•FO
RenderX

36. Ford P, Clifford A, Gussy K, Gartner C. A systematic review of peer-support programs for smoking cessation in disadvantaged groups. Int J Environ Res Public Health 2013;10(11):5507-5522 [FREE Full text] [doi: 10.3390/ijerph10115507] [Medline: 24169412]
37. Wagner J, Burg M, Sirois B. Social support and the transtheoretical model: Relationship of social support to smoking cessation stage, decisional balance, process use, and temptation. Addict Behav 2004;29(5):1039-1043. [doi: 10.1016/j.addbeh.2004.02.058] [Medline: 15219355]
38. Westmaas JL, Bontemps-Jones J, Bauer JE. Social support in smoking cessation: reconciling theory and evidence. Nicotine Tob Res 2010;12(7):695-707. [doi: 10.1093/ntr/ntq077] [Medline: 20513695]
39. Graham AL, Jacobs MA, Amato MS, Cha S, Bottcher MM, Papandonatos GD. Effectiveness of a quit vaping text message program in promoting abstinence among young adult e-cigarette users: protocol for a randomized controlled trial. JMIR Res Protoc 2020;9(5):e18327 [FREE Full text] [doi: 10.2196/18327] [Medline: 32356774]
40. Leas EC, Nobles AL, Caputi TL, Dredze M, Zhu S, Cohen JE, et al. News coverage of the e-cigarette, or vaping, product use associated lung injury (EVALI) outbreak and internet searches for vaping cessation. Tob Control 2021;30(5):578-582 [FREE Full text] [doi: 10.1136/tobaccocontrol-2020-055755] [Medline: 33051278]
41. The Lancet Respiratory Medicine. The EVALI outbreak and vaping in the COVID-19 era. Lancet Respir Med 2020;8(9):831 [FREE Full text] [doi: 10.1016/S2213-2600(20)30360-X] [Medline: 32805203]
42. Elling JM, Crutzen R, Talhout R, de Vries H. Tobacco smoking and smoking cessation in times of COVID-19. Tob Prev Cessat 2020;6:39 [FREE Full text] [doi: 10.18332/tpc/122753] [Medline: 33083672]
43. Treating tobacco use and dependence: 2008 update. US Department of Health and Human Services. URL: https://www.ahrq.gov/prevention/guidelines/tobacco/index.html [accessed 2023-07-17]
44. Myneni S, Fujimoto K, Cobb N, Cohen T. Content-driven analysis of an online community for smoking cessation: integration of qualitative techniques, automated text analysis, and affiliation networks. Am J Public Health 2015;105(6):1206-1212. [doi: 10.2105/AJPH.2014.302464] [Medline: 25880942]

Abbreviations

FDA: Food and Drug Administration
LDA: latent Dirichlet allocation
NRT: nicotine replacement therapy
PMTA: premarket tobacco product application

Edited by A Mavragani; submitted 23.08.23; peer-reviewed by A Palmer, M Metcalf, R Mermelstein; comments to author 11.01.24; revised version received 04.03.24; accepted 01.07.24; published 25.10.24.

Please cite as:
Kierstead E, Silver N, Amato M
Examining Quitting Experiences on Quit Vaping Subreddits From 2015 to 2021: Content Analysis
J Med Internet Res 2024;26:e52129
URL: https://www.jmir.org/2024/1/e52129
doi:10.2196/52129
PMID:

©Elexis Kierstead, Nathan Silver, Michael Amato. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Interest in Elf Bar in the United States: Google Health Trends Analysis

Akshaya Srikanth Bhagavathula[1*], PhD; Page D Dobbs[2*], PhD

[1]Department of Public Health, North Dakota State University, Fargo, ND, United States

[2]Department of Health, Human Performance and Recreation, University of Arkansas, Fayetteville, AR, United States

[*]all authors contributed equally

Corresponding Author:
Page D Dobbs, PhD
Department of Health, Human Performance and Recreation
University of Arkansas
Suite 317
346 West Ave
Fayetteville, AR, 72701
United States
Phone: 1 479 575 2858
Email: pdobbs@uark.edu

Abstract

Background: Despite the popularity of JUUL e-cigarettes, other brands (eg, Elf Bar) may be gaining digital attention.

Objective: This study compared Google searches for Elf Bar and JUUL from 2022 to 2023 using Google Health Trends Application Programming Interface data.

Methods: Using an infodemiology approach, we examined weekly trends in Google searches (per 10 million) for "Elf Bar" and "JUUL" at the US national and state levels from January 1, 2022, to December 31, 2023. Joinpoint regression was used to assess statistically significant trends in the search probabilities for "Elf Bar" and "JUUL" during the study period.

Results: Elf Bar had less online interest than JUUL at the beginning of 2022. When the US Food and Drug Administration denied JUUL marketing authority on June 23, 2022, JUUL searches peaked at 2609.3×10^7 and fell to 83.9×10^7 on September 3, 2023. Elf Bar searches surpassed JUUL on July 10, 2022, and steadily increased, reaching 523.2×10^7 on December 4, 2022. Overall, Elf Bar's weekly search probability increased by 1.6% (95% CI 1.5%-1.7%; P=.05) from January 2022 to December 2023, with the greatest increase between May 29 and June 19, 2022 (87.7%, 95% CI 35.9%-123.9%; P=.001). Elf Bar searches increased after JUUL's suspension in Pennsylvania (1010%), Minnesota (872.5%), Connecticut (803.5%), New York (738.1%), and New Jersey (702.9%).

Conclusions: Increasing trends in Google searches for Elf Bar indicate that there was a growing online interest in this brand in the United States in 2022.

(J Med Internet Res 2024;26:e50343) doi:10.2196/50343

KEYWORDS

e-cigarettes; Elf Bar; JUUL; tobacco; Google Trends; Google Health Trends

Introduction

Defined by 4 distinct "generations," the first e-cigarette products resembled traditional cigarettes and were referred to as "cig-a-likes" [1]. These devices evolved into vape pens (second generation) and modifiable tank-style systems (third generation). The current (fourth) generation of e-cigarette products now include cartridge-based and disposable devices. Cartridge-based devices include a prefilled "pod" that contains nicotine and flavors and can be discarded after use, and disposable devices are advertised to include a number of "puffs" (eg, 1500), after which they can be disposed [1]. Although youth and young adults' use of e-cigarettes was increasing from 2011 to 2020 [2], it peaked in 2019, when the fourth-generation flash drive–shaped product, JUUL, controlled the e-cigarette market [3-5]. Although the use of e-cigarettes has declined since 2019, they continue to be popular among young audiences, with 2.13

XSL•FO
RenderX

million middle and high school students reporting current use in the United States in 2023 [6].

In January 2020, the Food and Drug Administration (FDA) announced that a sales restriction of cartridge-based e-cigarette flavors (other than mint and tobacco) would go into effect later that month [7]. This policy was suggested to decrease the popularity of cartridge-based e-cigarettes such as JUUL. JUUL Labs Inc has now settled in court regarding cases where they were accused of marketing their products to those younger than 21 years, the legal sales age for tobacco products [8]. Although this policy was posited to reduce use among youth and young adults, researchers have speculated that the FDA's cartridge-based flavor restriction policy created a loophole for disposable devices, such as Puff Bar, that gained online attention after this policy was enacted [9,10].

On July 20, 2020, the FDA issued warning letters to 10 companies, including Puff Bar and other disposable e-cigarette brands, that they had not received the appropriate premarket authority to be marketed and sold in the United States [11]. After a short period of suspended sales, Puff Bar began selling their products again in March 2021, claiming that they used synthetic nicotine, which was not authorized for regulation by the FDA [12]; the FDA's authority over nicotine products is defined in the Tobacco Control Act as products that were made, contained, or derived from tobacco [13]. Thus, on March 11, 2022, President Joe Biden closed this regulatory gap by signing a Congress-approved amendment (Docket No. FDA-2022-N-3261) that expanded the tobacco product definition to "products that contain nicotine from any source," granting the FDA authority over synthetic nicotine products [12].

From 2021 to 2023, other brands of disposable e-cigarette devices, including Elf Bar, began to gain popularity in Great Britain and the United States, with 56.7% of US youths who reported current use of e-cigarettes indicating that they used Elf Bar most often in 2023 (compared with 15.5% for JUUL) [14-17]. Research indicates that discussion about Elf Bar and other emerging e-cigarette brands was increasing on Twitter in 2021, whereas discussion about JUUL was decreasing [18]. During this time, research suggests that disposable, nicotine salt–based e-cigarette products became the most popular e-cigarette devices on the market [15]. However, no research has explored the popularity of Elf Bar on a more widely available search engine, such as Google. Google Trends (GT) is a useful tool for measuring popularity (defined as online interest) of a wide range of topics including football players, disease prioritization, and the harms of drug use [19-21]. It has been used in prior tobacco-related research, to examine the online popularity of JUUL and Puff Bar [10], but it has not been used since the FDA's expansion of the tobacco product definition to determine the online popularity of other e-cigarette products, such as Elf Bar.

June 2022, the FDA announced the prohibition of JUUL sales in the US market [22], an order that was immediately appealed. Thus, although JUUL remained on the market, the popularity of this product continued to decline while newer disposable products that included flavors saw increases in use [9,10,23]. Although data suggest that there were shifts in e-cigarette product and brand use after the COVID-19 pandemic [15], and research has noted the popularity of Puff Bar [9,24,25], there remains a gap in literature about the online interest in other brands, particularly following the FDA's order to suspend Puff Bar sales. It is possible that regulatory announcements could have created changes in online interest of e-cigarette products. With growing regulations that focus on particular brand or product characterizations (eg, cartridge-based systems and synthetic nicotine-containing products), there could be growing online popularity for other e-cigarette brands, such as Elf Bar, a brand that has been noted for its recent popularity in the United Kingdom and United States [16,17]. The purpose of this study is to explore the online interest in Elf Bar in the United States, compared with JUUL from 2022 to 2023 using Google Health Trends Application Programming Interface (GHT-API) data.

Methods

GHT-API Data

GT is a popular data source that is widely used in public health research [26]. Given recent concerns about the data quality and accuracy of GT [27-29], some researchers have begun to use GHT-API, which provides raw probabilities of a short search session with no restrictions on the search volume index [30,31]. To collect and use the GHT-API data, researchers must have a valid API key supplied by Google. A detailed GHT-API technique and data features have already been published elsewhere [32,33]. Briefly, Google maintains a random sample of all Google search queries that can be obtained by researchers using the API. The GHT-API provides researchers with the relative share of all Google searches made by individuals within a specific geographic area and time range for given search terms. The API returns a scale proportion equal to the number of searches for specific search terms as a proportion of total searches multiplied by a constant of 10 million. Compared with traditional GT metrics, GHT-API facilitates more granular and unbiased analyses of temporal and spatial search trends pertinent to public health topics. Further research leveraging this promising data source may elucidate novel insights into online health information–seeking behaviors.

To identify Google search terms, we examined GT search queries to identify the most popular queries related to "Elf Bar" and "JUUL" from 2022 to 2023. Given JUUL's prior popularity [10], we included the term JUUL as a reference that would help interpret the popularity of Elf Bar as an emerging e-cigarette product. The study considered search queries labeled "Breakout" (indicates that the Google search term has grown by more than 5000%) [34]. Following these techniques and using a valid GHT-API key, we collected the data for the search terms "Elf Bar" and "JUUL" in the United States from January 1, 2022, through December 31, 2023, a time at which Elf Bar was noted for its popularity among US youth [17]. Like previous studies [30-33], we collected multiple samples (using the same dates) of weekly and monthly data searches for "Elf Bar" and "JUUL" within the United States and averaged them to obtain better estimates of their true values. We conducted weekly and monthly searches at the state level for all 50 states between

XSL•FO
RenderX

January 1 and October 16, 2022; however, 8 states (Alaska, Delaware, Montana, North Dakota, Rhode Island, South Dakota, Vermont, and Wyoming) and the District of Columbia were excluded because of insufficient data or suppression of the API by the GHT-API. These areas were not considered in the state-level analysis.

Statistical Analysis

To obtain GHT-API data, Python 3.11.0 (Python Software Foundation) was used. Excel (Microsoft Corp) was used to clean data, and the *geofacet* package for R version 4.2.1 (R Foundation for Statistical Computing) was used to create data visualizations. The coefficient of variance (COV) method was used to calculate the variability of data between samples. The COV can be calculated as the ratio of the SD to the mean:

where μ is the mean probability of searches and σ is the SD.

The COV measures how consistent the values of each sample are from the representative mean of the data samples. The lower the percentage COV value, the greater the homogeneity of the values in the data set. A total of 5 samples were collected with COV that were found to be less than 5% between the samples. Therefore, we used the mean of 5 samples for the analysis.

The probabilities of weekly Google searches for "Elf Bar" and "JUUL," from January 1, 2022, to December 31, 2023, were plotted to visualize the temporal trends. We calculated the percentage changes in Elf Bar after JUUL's suspension in July 2022 for each state using the following formula:

Weekly probabilities of the US public interest in Elf Bar and JUUL were calculated Using Joinpoint Regression Analysis Software (version 5.0.2; National Cancer Institute, May 2023). This joinpoint regression software uses regression modeling to assess trends while searching for temporal trend changes at time points known as joinpoints and estimating the regression function from previous joinpoints [35]. The number of joinpoints is calculated using a weighted Bayesian information criterion through the data-dependent selection method [36], and the analysis criteria were set to detect up to 5 joinpoints. This joinpoint regression analysis technique can capture nonlinear inflection points and trends in the data and provide a more nuanced understanding of the dynamics of information-seeking behavior on the internet. Weekly percentage changes between trend-change points were determined using 95% CIs.

Results

Overview

Figure 1 shows that the search interest in the Elf Bar was lower (highest weekly probability: 17.3×10^7) than interest in the "JUUL" (highest weekly probability: 198.4×10^7) at the beginning of 2022. Starting from week 24 (June 12, 2022), Elf Bar search interest (169.1×10^7) exceeded "JUUL" (165.2×10^7) and continued to steadily increase. During week 25, the search probability of JUUL peaked at 2609.3×10^7. During that week (June 23, 2022), the US FDA denied JUUL marketing authority in the United States (temporarily suspending JUUL products from the US market). After this time, online interest in JUUL decreased, reaching an all-time low of 83.9×10^7 on September 3, 2023 (week 88). Alternatively, Elf Bar search interest surpassed that of JUUL, again, on week 28 (July 10, 2022) and steadily increased over time, with the highest weekly probability of 523.2×10^7 reported on week 49 (December 4, 2022).

Figure 1. Weekly Google searches comparing "Elf Bar" and "JUUL" in 2022-2023 (105 weeks). FDA: Food and Drug Administration.

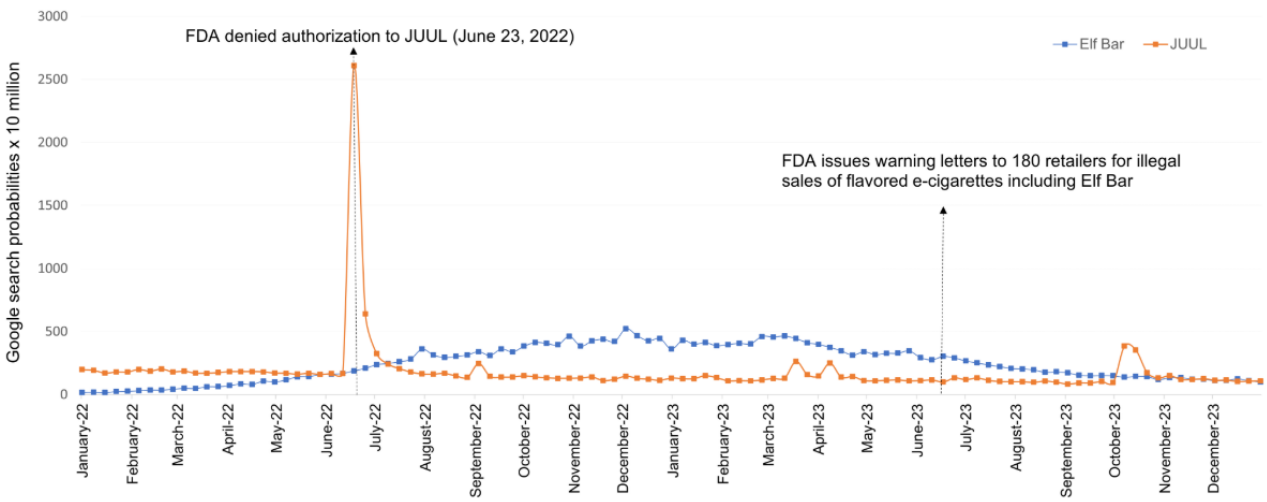

Elf Bar Searches at the State Level

Trends in monthly Elf Bar search volumes for 42 states from January 1, 2022, to December 31, 2023, are shown in Figure 2. Overall, there was an upward trend in Elf Bar–related search volumes across all states after the JUUL suspension. West Virginia (866.4×10^7), Louisiana (579.2×10^7), Ohio (542.8 ×

XSL•FO
RenderX

10^7), Mississippi (504.2×10^7), and Pennsylvania (480.6×10^7) had the highest mean search volume for Elf Bar, whereas Michigan (118.2×10^7), Massachusetts (144.5×10^7), New Hampshire (167.2×10^7), Texas (196.5×10^7), and Illinois (198.1×10^7) had the lowest. However, after the JUUL suspension in July 2022, changes in the percentages in Elf Bar searches were highest in Pennsylvania (1010%), Minnesota (872.5%), Connecticut (803.5%), New York (738.1%), and New Jersey (702.9%), whereas states such as Hawaii (–13.1%), Virginia (76.5%), Nebraska (76.6%), Louisiana (80.2%), and Idaho (95.2%) recorded the lowest search interest (see Figure 3).

Figure 2. "Elf Bar" weekly search volume from January 1, 2022, to December 31, 2023. The API suppressed data and excluded 8 states (Alaska, Delaware, Montana, North Dakota, Rhode Island, South Dakota, Vermont, and Wyoming) and the District of Columbia with small populations and data uncertainty. API: application programming interface.

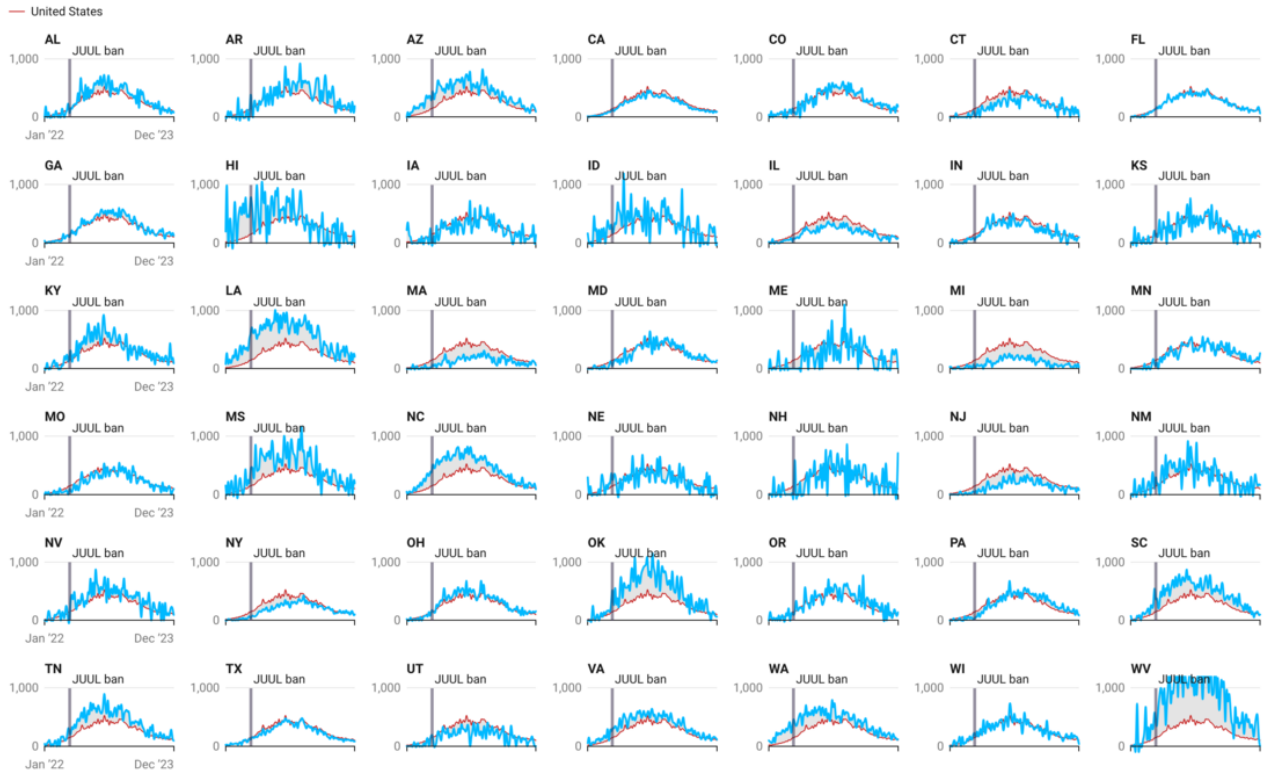

Figure 3. Percentage change in mean search volumes of Elf Bar after JUUL suspension in June 2022.

XSL•FO
RenderX

Trends in Elf Bar and JUUL Searches

From January 1, 2022, to December 31, 2023, three joinpoints for Elf Bar and JUUL were identified, indicating 3 distinct trends or changes in weekly searches that all demonstrated a significant search interest during the study period (Figure 4). Overall, there was an average increase of 1.6% (95% CI 1.5-1.7; *P*<.001) in Elf Bar's weekly search probability and an average decrease of –0.5% (95% CI –0.8 to –0.2; *P*<.001) in JUUL's weekly search probability (Table 1).

Figure 4. Joinpoint analysis of the weekly search trend of Elf Bar and JUUL from January 1, 2022, to December 31, 2023. *Significant at *P*<.05. Three joinpoints identified.

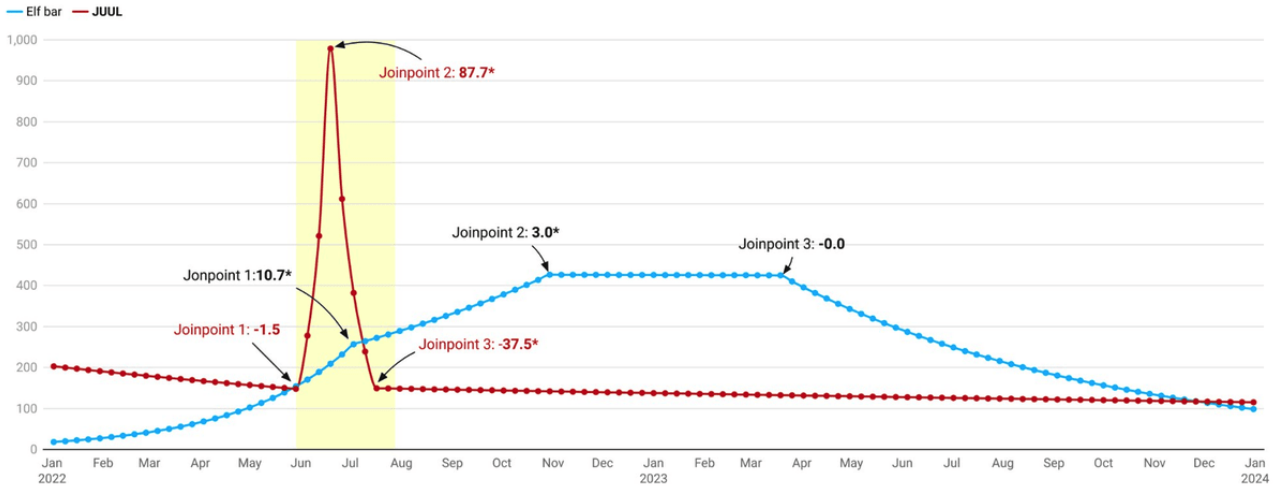

Table 1. Joinpoint analysis of the weekly search trend of Elf Bar and JUUL.

Search term and period[a]	Weekly percentage change (%; 95% CI)	*P* value
Elf Bar		
January 1 to July 3, 2022	10.7 (10.2 to 11.3)	<.001
July 3 to October 30, 2022	3.0 (2.1 to 4.6)	<.001
October 30 to March 19, 2023	–0.0 (–1.1 to 0.8)	.97
March 19 to December 31, 2022	–3.5 (–3.8 to –3.3)	.001
January 1, 2022, to December 31, 2023[b]	1.6 (1.5 to 1.7)	<.001
JUUL		
January 1 to May 29, 2022	–1.5 (–3.9 to 0.3)	.09
May 29 to June 19, 2022	87.7 (35.9 to 123.9)	.001
June 19 to July 17, 2022	–37.5 (–53.1 to –23.8)	.001
July 17 to December 2023	–0.3 (–0.6 to –0.02)	.03
January 1 to December 31, 2023[b]	–0.5 (–0.8 to –0.2)	<.001

[a]Three joinpoints.

[b]Average weekly percentage change during the study period.

Briefly, before JUUL's marketing denial order on June 23, 2022, the search interest for Elf Bar increased by 10.7% (95% CI 10.2%-11.3%; *P*<.001) per week from January to July 3, 2022. After July, the search probability for Elf Bar increased by 3.0% (95% CI 2.1%-4.6%; *P*<.001) each week until October 30, 2022, after which a negative trend was observed until December 31, 2023. In comparison, JUUL had 4 distinct time trends (1 increasing and 3 decreasing) in search interest during the study period. There was an exponential increase of 87.7% (95% CI 35.9%-123.9%; *P*=.001) in weekly search interest of JUUL during the time when the US FDA denied JUUL marketing authority in the United States, after which a significant continuous negative trend was observed.

Discussion

In this study, we investigated the trends in Elf Bar's and JUUL's online searches in the United States from January 2022 to December 2023. We found that Elf Bar searches increased both before and after the FDA denied JUUL marketing authority, until October 30, 2022, where online interest in Elf Bar began to steadily decrease. These findings demonstrate that there was a decreasing online interest in JUUL and an increasing online interest in Elf Bar through 2022, which is consistent with Twitter

XSL•FO
RenderX

trend data from 2021 [18]. We found there to be a peak in online interest about JUUL via GHT-API during the time that the FDA announced the suspension of JUUL products. Although this suspension was appealed by JUUL Labs the following day [37], such an announcement generated online discourse about JUUL. Including JUUL provides a comparison to understand the growing online popularity of Elf Bar, similar to prior comparisons between JUUL and Puff Bar [10]. This trend of increased interest in Elf Bar compared with JUUL continued until June 2023, around the time when the FDA announced that they had issued more than 180 warning letters to retailers for illegally selling Elf Bar products that had not received premarket approval.

Although the popularity of JUUL is well documented [5,38-42] and research suggests that the popularity of flavored disposable e-cigarette devices (ie, Puff Bar) emerged soon after the FDA's flavor restriction for cartridge-based e-cigarette systems went into effect [10], prior research has not examined the growing online interest in the emerging e-cigarette brand Elf Bar in the United States. Our findings show that in 2022, Elf Bar searches exceeded those of JUUL, a product that controlled the e-cigarette market from 2015 to 2019 [5]. Although Puff Bar was the first disposable device to replace JUUL's online popularity [10], emerging research suggests that Elf Bar was reported as the preferred product among youth in the United Kingdom in 2022 for reasons including better flavor and taste [17]. Further, findings from the US National Youth Tobacco Survey reported Elf Bar to be the most popular e-cigarette brand among middle school and high school students [16]. Although we cannot determine the reasons for using this search term, we can add to a growing body of literature that there was an online interest in Elf Bar during a time when it was popular among young audiences. Furthermore, research suggests that self-proclaimed small business owners are selling e-cigarettes illegally to youth on social media [43,44]. Thus, given a recent call for better regulation of tobacco e-commerce laws [45], our findings suggest the need to better understand how Google search terms were used to inform the online surveillance of e-cigarette products.

In 2019, eight states (Massachusetts, Michigan, Montana, New York, Oregon, Rhode Island, Utah, and Washington) temporarily restricted the sales of flavored e-cigarettes, in response to e-cigarette and vaping-related lung injury [46]. Since that time, 5 states (Massachusetts, New Jersey, New York, Rhode Island, and California) have prohibited the sale of all flavored e-cigarettes, and 2 states (Maryland and Utah) have restricted the sale of some flavored e-cigarettes. We found that several of these states had low weekly searches for Elf Bar after the announcement that JUUL was ordered to end sales; however, Massachusetts and New York saw a notable change in the percentage of searchers for Elf Bar after this announcement. Thus, the availability of these products may generate or curb online interest in these products. With a changing legal landscape of tobacco product (including e-cigarettes) regulation at the local, state, and federal levels [47-49], there remains a need to understand how policy changes may impact online interest in emerging tobacco products. As described in Twitter discussions, e-cigarette users often report confusion about the difference between leading products such as JUUL and Puff Bar, and some reported switching to other products rather than quitting e-cigarettes after the cartilage-based flavor restriction went into effect [23]. Thus, as state and federal tobacco regulatory policies change, GHT could help explore changes in online interest that may occur after these actions.

This study was subject to limitations. Given the limited searches provided via GT, search terms used may not capture the full landscape of interest in these brands. Further, although this study focused on the growing popularity of Elf Bar compared with JUUL, there may be other disposable devices that are also growing in popularity online. Although we were able to break down interest by location (state), using time-stamped data, we are unable to identify the demographic data of the persons looking for information about the search term. Thus, there is no way to know if the audience searching for Elf Bar was legally old enough to purchase the product (aged 21 years and older) or was younger than this age. Understanding online customers' age could help to understand policy enforcement of the local, state, and federal Tobacco 21 laws. We also used weekly GHT data rather than data captured monthly, which could provide a more detailed response to the FDA's JUUL suspension. Further, Google searches do not explain why the search was made; thus, we cannot confirm the reason for which people are conducting these searches. Finally, the data collected via GHT-API only capture searches for these terms online and do not capture data about use or attempts to purchase these products.

We found there to be an increased interest in Elf Bar during 2022. Although interest in JUUL increased during the time frame when the FDA announced that they were suspending the sale of JUUL products (only to be appealed the following day), the growing online interest in Elf Bar surpassed that of JUUL. Coupled with sales data and survey data about product preferences, understanding trends in online interest for particular products and brands can help inform marketing campaigns and prevention strategies that seek to reduce use among youths and young adults. Such trends data can help to describe online search history around time frames, such as announcements of tobacco regulatory responses.

Data Availability

Data will be made available upon request.

XSL•FO
RenderX

Authors' Contributions

ASB and PDD were both responsible for the study planning, conceptualization, and coordination. ASB drafted the original methods, results, and discussion sections, and PDD drafted the first introduction section. Both authors reviewed and edited the manuscript.

Conflicts of Interest

None declared.

References

1. Ozga JE, Felicione NJ, Douglas A, Childers M, Blank MD. Electronic cigarette terminology: where does one generation end and the next begin? Nicotine Tob Res 2022;24(3):421-424 [FREE Full text] [doi: 10.1093/ntr/ntab164] [Medline: 34379784]
2. Preventing tobacco use among youth and young adults: a report of the surgeon general. U.S. Department of Health and Human Services. 2012. URL: http://www.surgeongeneral.gov/library [accessed 2021-02-15]
3. Wang TW, Neff LJ, Park-Lee E, Ren C, Cullen KA, King BA. E-cigarette use among middle and high school students—United States, 2020. MMWR Morb Mortal Wkly Rep 2020;69(37):1310-1312 [FREE Full text] [doi: 10.15585/mmwr.mm6937e1] [Medline: 32941408]
4. Chu K, Colditz JB, Primack BA, Shensa A, Allem J, Miller E, et al. JUUL: spreading online and offline. J Adolesc Health 2018;63(5):582-586 [FREE Full text] [doi: 10.1016/j.jadohealth.2018.08.002] [Medline: 30348280]
5. Huang J, Duan Z, Kwok J, Binns S, Vera LE, Kim Y, et al. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. Tob Control 2019;28(2):146-151 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054382] [Medline: 29853561]
6. Birdsey J, Cornelius M, Jamal A, Park-Lee E, Cooper MR, Wang J, et al. Tobacco product use among U.S. middle and high school students—National Youth Tobacco Survey, 2023. MMWR Morb Mortal Wkly Rep 2023;72(44):1173-1182 [FREE Full text] [doi: 10.15585/mmwr.mm7244a1] [Medline: 37917558]
7. FDA finalizes enforcement policy on unauthorized flavored cartridge-based e-cigarettes that appeal to children, including fruitmint. U.S. Food and Drug Administration. 2020. URL: https://www.fda.gov/news-events/press-announcements/fda-finalizes-enforcement-policy-unauthorized-flavored-cartridge-based-e-cigarettes-appeal-children [accessed 2022-11-17]
8. The People of the State of New York vs. JUUL Labs Inc. Index No. 452168/19(2023). Office of the New York State Attorney General. URL: https://ag.ny.gov/sites/default/files/settlements-agreements/NY%20JUUL%20Final.pdf [accessed 2024-02-14]
9. Chu KH, Hershey TB, Hoffman BL, Wolynn R, Colditz JB, Sidani JE, et al. Puff bars, tobacco policy evasion, and nicotine dependence: content analysis of tweets. J Med Internet Res 2022;24(3):e27894 [FREE Full text] [doi: 10.2196/27894] [Medline: 35333188]
10. Dai H, Hao J. Online popularity of JUUL and Puff Bars in the USA: 2019-2020. Tob Control 2022;31(1):7-10. [doi: 10.1136/tobaccocontrol-2020-055727] [Medline: 33051277]
11. FDA notifies companies, including Puff Bar, to remove flavored disposable e-cigarettes and youth-appealing e-liquids from market for not having required authorization | FDA. U.S. Food and Drug Administration. 2020. URL: https://www.fda.gov/news-events/press-announcements/fda-notifies-companies-including-puff-bar-remove-flavored-disposable-e-cigarettes-and-youth [accessed 2020-08-12]
12. What is synthetic nicotine and what does it mean for the youth vaping epidemic? Truth Initative. 2022. URL: https://truthinitiative.org/research-resources/emerging-tobacco-products/what-synthetic-nicotine-and-what-does-it-mean-youth [accessed 2022-11-17]
13. Family smoking prevention and tobacco control act—an overview. U.S. Food and Drug Administration. URL: https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/family-smoking-prevention-and-tobacco-control-act-overview [accessed 2021-01-25]
14. Tattan-Birch H, Jackson S, Kock L, Dockrell M, Brown J. Rapid growth in disposable e-cigarette vaping among young adults in Great Britain from 2021 to 2022: a repeat cross-sectional survey. Addiction 2023;118(2):382-386 [FREE Full text] [doi: 10.1111/add.16044] [Medline: 36065820]
15. Hammond D, Reid JL. Trends in vaping and nicotine product use among youth in Canada, England and the USA between 2017 and 2022: evidence to inform policy. Tob Control 2023. [doi: 10.1136/tc-2023-058241] [Medline: 37940402]
16. Results from the annual National Youth Tobacco Survey. U.S. Food and Drug Administration. URL: https://www.fda.gov/tobacco-products/youth-and-tobacco/results-annual-national-youth-tobacco-survey [accessed 2024-06-13]
17. East K, Taylor EV, Simonavičius E, Reid JL, Burkhalter R, McNeill A, et al. Use of 'Elf Bar' among youth and young adults who currently vape in England: cross-sectional associations with demographics, dependence indicators and reasons for use. Addiction 2024 Mar 21. [doi: 10.1111/add.16463] [Medline: 38515247]
18. Tang Q, Zhou R, Xie Z, Li D. Monitoring and identifying emerging e-cigarette brands and flavors on Twitter: observational study. JMIR Form Res 2022;6(12):e42241 [FREE Full text] [doi: 10.2196/42241] [Medline: 36469415]

XSL•FO
RenderX

19. Bitirim Y. An evaluation on popularity of the prioritized diseases using Google Trends. J Med Imaging Health Inform 2019;9(4):682-687. [doi: 10.1166/jmihi.2019.2613]
20. Jankowski W, Hoffmann M. Can Google searches predict the popularity and harm of psychoactive agents? J Med Internet Res 2016;18(2):e38 [FREE Full text] [doi: 10.2196/jmir.4033] [Medline: 26916984]
21. Malagón-Selma P, Debón A, Domenech J. Measuring the popularity of football players with Google Trends. PLoS One 2023;18(8):e0289213 [FREE Full text] [doi: 10.1371/journal.pone.0289213] [Medline: 37585380]
22. Wise J. FDA orders JUUL to remove all its e-cigarettes from US market. BMJ 2022;377:o1568. [doi: 10.1136/bmj.o1568] [Medline: 35750332]
23. Allem J, Dormanesh A, Majmundar A, Rivera V, Chu M, Unger JB, et al. Leading topics in Twitter discourse on JUUL and Puff Bar products: content analysis. J Med Internet Res 2021;23(7):e26510 [FREE Full text] [doi: 10.2196/26510] [Medline: 34279236]
24. Tan ASL, Weinreich E. #PuffBar: how do top videos on TikTok portray Puff Bars? Tob Control 2021;30(6):712-713 [FREE Full text] [doi: 10.1136/tobaccocontrol-2020-055970] [Medline: 32934090]
25. Rosenthal H, Chow N, Mehta S, Pham D, Milanaik R. Puff Bars: a dangerous trend in adolescent disposable e-cigarette use. Curr Opin Pediatr 2022;34(3):288-294. [doi: 10.1097/MOP.0000000000001123] [Medline: 35152232]
26. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of Google Trends in health care research: a systematic review. PLoS One 2014;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
27. Cebrián E, Domenech J. Is Google Trends a quality data source? Applied Economics Letters 2022;30(6):811-815 [FREE Full text] [doi: 10.1080/13504851.2021.2023088]
28. Rovetta A. Reliability of Google Trends: analysis of the limits and potential of web infoveillance during COVID-19 pandemic and for future research. Front Res Metr Anal 2021;6:670226 [FREE Full text] [doi: 10.3389/frma.2021.670226] [Medline: 34113751]
29. Arora VS, McKee M, Stuckler D. Google Trends: opportunities and limitations in health and health policy research. Health Policy 2019;123(3):338-341. [doi: 10.1016/j.healthpol.2019.01.001] [Medline: 30660346]
30. Raubenheimer JE, Riordan BC, Merrill JE, Winter T, Ward RM, Scarf D, et al. Hey Google! will New Zealand vote to legalise cannabis? Using Google Trends data to predict the outcome of the 2020 New Zealand cannabis referendum. Int J Drug Policy 2021;90:103083. [doi: 10.1016/j.drugpo.2020.103083] [Medline: 33341700]
31. Riddell CA, Farkas K, Neumann K, Santaularia NJ, Ahern J, Mason SM. US shelter in place policies and child abuse Google search volume during the COVID-19 pandemic. Prev Med 2022;163:107215 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107215] [Medline: 35998763]
32. Raubenheimer JE. A practical algorithm for extracting multiple data samples from Google Trends extended for health. Am J Epidemiol 2022;191(9):1666-1669. [doi: 10.1093/aje/kwac088] [Medline: 35512308]
33. Neumann K, Mason SM, Farkas K, Santaularia NJ, Ahern J, Riddell CA. Harnessing Google health trends data for epidemiologic research. Am J Epidemiol 2023;192(3):430-437 [FREE Full text] [doi: 10.1093/aje/kwac171] [Medline: 36193858]
34. Google Trends: understanding the data. Google News Initiatives. URL: https://newsinitiative.withgoogle.com/resources/trainings/fundamentals/google-trends-understanding-the-data/ [accessed 2024-02-14]
35. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med 2000;19(3):335-351. [doi: 10.1002/sim.811]
36. Kim HJ, Chen HS, Midthune D, Wheeler B, Buckman DW, Green D, et al. Data-driven choice of a model selection method in joinpoint regression. J Appl Stat 2023;50(9):1992-2013 [FREE Full text] [doi: 10.1080/02664763.2022.2063265] [Medline: 37378270]
37. Houghtaling EQ. FDA Temporarily Suspends Ban on Juul E-Cigarettes. 2022. URL: https://www.npr.org/2022/07/06/1109962374/fda-temporarily-suspends-order-banning-juul-cigarettes [accessed 2024-09-21]
38. Katz SJ, Cohen EL, Kinzer HT. "Can I hit that?" Vaping knowledge, attitudes and practices of college students. J Am Coll Health 2022;70(6):1778-1787. [doi: 10.1080/07448481.2020.1820512] [Medline: 33048639]
39. Willett JG, Bennett M, Hair EC, Xiao H, Greenberg MS, Harvey E, et al. Recognition, use and perceptions of JUUL among youth and young adults. Tob Control 2019;28(1):115-116. [doi: 10.1136/tobaccocontrol-2018-054273] [Medline: 29669749]
40. Sidani JE, Colditz JB, Barrett EL, Chu K, James AE, Primack BA. JUUL on Twitter: analyzing tweets about use of a new nicotine delivery system. J Sch Health 2020 Feb;90(2):135-142 [FREE Full text] [doi: 10.1111/josh.12858] [Medline: 31828791]
41. Sidani JE, Colditz JB, Barrett EL, Shensa A, Chu K, James AE, et al. I wake up and hit the JUUL: analyzing Twitter for JUUL nicotine effects and dependence. Drug Alcohol Depend 2019;204:107500 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.06.005] [Medline: 31499242]
42. Case KR, Hinds JT, Creamer MR, Loukas A, Perry CL. Who is JUULing and why? An examination of young adult electronic nicotine delivery systems users. J Adolesc Health 2020;66(1):48-55 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.05.030] [Medline: 31481286]

XSL•FO
RenderX

43. Dobbs PD, Schisler ED, McCormick C. #Discreetshipping: selling e-cigarettes on TikTok. Nicotine Tob Res 2024:ntae081. [doi: 10.1093/ntr/ntae081] [Medline: 38778482]
44. Pearson G, Davidson DL, Schillo B, Kreslake JM. 'Discreet shipping' on TikTok enables selling of e-cigarettes to youth. Tob Control 2024. [doi: 10.1136/tc-2023-058315] [Medline: 38191236]
45. Leas EC. Need for improved regulation of tobacco e-commerce. Tob Control 2024 [FREE Full text] [doi: 10.1136/tc-2023-058515] [Medline: 38191235]
46. States and localities that have restricted the sale of flavored tobacco products. Campaign for Tobacco Free Kids. URL: https://assets.tobaccofreekids.org/factsheets/0398.pdf [accessed 2024-04-12]
47. Dobbs PD, Chadwick G, Dunlap CM, White KA, Cheney MK. Tobacco 21 policies in the U.S.: the importance of local control with federal policy. Am J Prev Med 2021;60(5):639-647. [doi: 10.1016/j.amepre.2020.12.009] [Medline: 33602533]
48. D Dobbs PD, Chadwick G, W Ungar KW, M Dunlap CM, White KA, Kelly MC, et al. Development of a Tobacco 21 policy assessment tool and state-level analysis in the USA, 2015-2019. Tob Control 2020;29(5):487-495 [FREE Full text] [doi: 10.1136/tobaccocontrol-2019-055102] [Medline: 31611425]
49. Dutra LM, Glantz SA, Arrazola RA, King BA. Impact of e-cigarette minimum legal sale age laws on current cigarette smoking. J Adolesc Health 2018;62(5):532-538 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.11.302] [Medline: 29422436]

Abbreviations

COV: coefficient of variance
FDA: Food and Drug Administration
GHT-API: Google Health Trends Application Programming Interface
GT: Google Trends

Edited by T Leung, T de Azevedo Cardoso; submitted 27.06.23; peer-reviewed by L Dutra, M Navarro; comments to author 02.02.24; revised version received 15.02.24; accepted 16.08.24; published 05.11.24.

Please cite as:
Bhagavathula AS, Dobbs PD
Online Interest in Elf Bar in the United States: Google Health Trends Analysis
J Med Internet Res 2024;26:e50343
URL: https://www.jmir.org/2024/1/e50343
doi:10.2196/50343
PMID:

©Akshaya Srikanth Bhagavathula, Page D Dobbs. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Social Media Marketing Strategies for Electronic Cigarettes: Content Analysis of Chinese Weibo Accounts

Xinyi Zhou[1,2*], MA; Xinyu Hao[3*], MA; Yuhang Chen[4], MA; Hui Deng[5,6], MA; Ling Fang[5,6], MA; Lingyun Zhang[5,6], MA; Xiaotao Yan[5,6], MA; Pinpin Zheng[5,6], PhD; Fan Wang[1], PhD

[1]Fudan Development Institute, Fudan University, Shanghai, China
[2]Asia Research Center, Fudan University, Shanghai, China
[3]School of Politics and International Relations, East China Normal University, Shanghai, China
[4]School of Journalism, Fudan University, Shanghai, China
[5]Department of Preventive Medicine and Health Education, School of Public Health, Fudan University, Shanghai, China
[6]Key Laboratory of Public Health Safety, Fudan University, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Fan Wang, PhD
Fudan Development Institute
Fudan University
Think Tank Building, Room 415
220 Handan Road
Shanghai, 200433
China
Phone: 86 21 55664081
Fax: 86 21 55670203
Email: wangfan512@126.com

Abstract

Background: E-cigarettes have gained popularity among teenagers due to extensive marketing strategies on social media platforms. This widespread promotion is a risk factor, as it fosters more positive attitudes toward e-cigarette use among teenagers and increases the perception that using e-cigarettes is normal. Therefore, the marketing of e-cigarettes on social media is a serious global health concern, and its strategies and impact should be clearly identified.

Objective: This study examined how e-cigarette companies popularize their products via Weibo and identified the specific strategies influencing the effectiveness of their marketing.

Methods: In phase 1, we conducted a search on Qcc.com and identified 32 e-cigarette brands with active Weibo accounts between October 1 and December 31, 2020, along with 863 Weibo posts. The data were investigated through content analysis. The codebook was developed into four categories: (1) product and features, (2) sales and promotions, (3) social contact and interaction, and (4) restrictions and warnings. To further understand the factors influencing e-cigarette brand marketing, we conducted a multiple linear regression analysis.

Results: Marketing tactics by e-cigarette companies on Chinese social media were documented, including emphasizing attractive product features, using trendy characters, implicit promotions, downplaying health concerns, and engaging with Weibo users in various ways. Out of 863 posts, 449 (52%) mentioned product characteristics. In 313 (36.3%) posts, visible figures were used to attract attention. Product promotion was absent in 762 (88.3%) posts, and purchase channels were not mentioned in 790 (98.3%) posts. Social interaction–related posts received attention (n=548, 63.5%), particularly those featuring hashtag content (n=538, 62.3%). Most posts did not include claims for restrictions on teenagers' purchases or use (n=687, 79.6%) or information on health warnings (n=839, 97.2%). Multiple linear regression analysis identified marketing strategies that effectively increase the exposure of e-cigarette posts on Weibo. Posts including engagement via posts encouraging reposts, comments, and likes ($P<.001$) and engagement topics related to e-cigarette brands were positively correlated with the number of reposts ($P=.009$). Posts highlighting nonmonetary incentives ($P=.004$), posts with age restriction statements ($P<.001$), engaging via stories and idea collection ($P<.001$), and engagement topics related to products ($P<.001$) and current affairs ($P=.002$) had a positive effect on the number of comments. Engagement topics related to brands ($P<.001$) or interactive sweepstakes ($P<.001$) had a positive effect on the number of likes.

XSL•FO
RenderX

Conclusions: E-cigarette posts on Weibo that focus on product features and social interaction attract public attention, especially from teenagers. Stricter regulations and monitoring should be adopted to restrict the social media marketing of e-cigarettes.

(J Med Internet Res 2024;26:e51594) doi:10.2196/51594

KEYWORDS

e-cigarette; marketing strategy; social media; teenagers; content analysis

Introduction

The prevalence of emerging nicotine and tobacco products, including electronic nicotine delivery systems (ENDS), has significantly increased worldwide [1], posing serious health concerns. E-liquids created by e-cigarettes contain toxic and potentially carcinogenic metals, increasing the likelihood of developing chronic health issues [2,3]. However, e-cigarettes have emerged as the most prevalent tobacco product among teenagers worldwide [4,5]. According to the US Centers for Disease Control, in 2022, over 2.55 million middle and high school students reported e-cigarette use within the past 30 days in the United States [6]. In China, the 2021 National Tobacco Prevalence Survey of High School and College Students reported that 86.6% of middle school students were aware of e-cigarettes, while 3.6% had used them. These figures represent a 9.2% and 0.8% increase, respectively, compared with 2019 [7]. In many countries, e-cigarette use rates among young people are higher than those among adults [8]. The e-cigarette trend prompted the World Health Organization (WHO) to strictly regulate ENDS to ensure the maximum protection of public health [1].

An increasing number of published studies have demonstrated that the popularity of e-cigarettes is largely attributed to their extensive propagation on social media platforms [9-11]. Notwithstanding this, e-cigarette marketing on social media is not closely monitored [12,13], likely contributing to a surge in the marketing and consumption of these products. For example, currently, in the United States, e-cigarette companies not bound by the Tobacco Master Settlement Agreement, an agreement that has committed to reducing smoking rates among Americans, can promote their products on social media platforms such as Twitter, Facebook, Instagram, and YouTube [14-16]. In the early years, a similar regulatory gap existed in China [17]. Since 2020, the Chinese government has recognized the potential health hazards of e-cigarettes, particularly for teenagers [18]. The Special Inspection Action Plan for the E-cigarette Market was implemented, which conducted a comprehensive clean-up of e-cigarette–related information on the internet [19]. Since 2021, the State Tobacco Monopoly Administration in China has the authority to supervise new tobacco products, including e-cigarettes [20], with stricter regulatory oversight.

Nevertheless, the implicit promotion of e-cigarettes on specific social media platforms has persisted in China for a considerable period. This promotion is driven by both e-cigarette companies and users, aiming to make the product more appealing through attractive flavors [12], eye-catching packaging design [13,21], and positioning it as a cessation tool [22]. Similar challenges are faced in Australia, the United States, the United Kingdom, and Indonesia [23-25]. Exposure to promotional content related to e-cigarettes on social media is identified as a risk factor contributing to heightened intentions to use e-cigarettes, more favorable attitudes toward e-cigarettes, and an increased perception of e-cigarette use as a normative behavior, particularly among teenagers [26,27].

Weibo is China's most popular social media platform, integrating the features of Twitter, YouTube, and Instagram. In 2020, 30% of Weibo active users were born after 2000, indicating that teenagers are an important user group [28]. Teenagers use Weibo for social interaction, self-presentation, and access to various types of information, including entertainment gossip, social news, brand marketing, and other types of information [29]. Given that most teenagers engage with numerous social media platforms several times per day [30], they are exposed to a considerable quantity of e-cigarette information on social media, are particularly vulnerable to such messaging, and have demonstrated considerable interest in vaping [14,31,32]. This expands the accessibility of e-cigarette information, influencing the likelihood of e-cigarette use and shaping social norms around e-cigarette use among young people [1].

Given the growing apprehension regarding the widespread appeal of e-cigarette–related content and the potential to cultivate a new generation of nicotine-dependent users [33], particularly among teenagers, examining Weibo posts on how e-cigarette companies have popularized their products is meaningful. The results can inform policy makers on how e-cigarettes are being promoted and marketed on social media, which is critical for intensifying regulations around e-cigarettes. To address these concerns, we aimed to (1) investigate the characteristics of marketing strategies of e-cigarette advertisements on Weibo and (2) examine the factors influencing the perceived attractiveness of e-cigarettes on Weibo.

Methods

Aim 1: Marketing Strategies of E-Cigarette Advertisements on Weibo

Data Collection

First, we conducted a search on the Chinese enterprise search website Qcc.com [34] using the keyword "e-cigarettes" with specific constraints such as "industry category=manufacturing," "search scope=company name/brand/product," and "registration status=current/existing." This search yielded a total of 3746 electronic cigarette companies. Second, we performed a data cleansing process in which we excluded 2403 non–e-cigarette companies that were unrelated to e-cigarettes or had names such as Electronic Commerce Centre, Experience Store, Authorized

XSL•FO
RenderX

Store, Studio, and similar contents. Next, we examined the remaining dataset and excluded 1239 companies that lacked a company website or had an inaccessible website. As a result, we obtained a final list of 104 e-cigarette companies. Third, using the names and relevant information of these 104 e-cigarette companies, we searched on Weibo and found 181 related Weibo accounts. Of these 181 accounts, 66 (36.5%) had updated Weibo posts in 2020 and 32 (17.7%) posted new content during our research period (October 1 to December 31, 2020). Finally, we selected 32 e-cigarette brands and their official Weibo accounts to include in our research.

Our analysis included all 863 posts published by the 32 accounts from October 1, 2020, to December 31, 2020. All the posts were scraped by researchers manually. The sample collection was completed in early January 2021.

Developing the Coding Framework

We reviewed the existing studies on e-cigarette social media marketing and browsed through basic account information and all 863 posts from 32 e-cigarette Weibo accounts to develop the coding framework. First, we reviewed previous relevant content analysis and summarized important strategies to test in our study [1] [35-46]. Second, we reviewed the 863 posts, combined the characteristics of Weibo, and created an initial codebook comprising four analysis categories: (1) product and features [22,36], (2) sales and promotion [23,37], (3) social contact and interaction, and (4) restrictions and warnings [47,48]. The coding framework was established through discussions and revisions among the researchers involved in this study. We entered the data into EpiData (EpiData Association) via the double-entered method and imported it into SPSS (version 29.0; IBM Corp) for data analysis. To examine the characteristics of the Weibo posts, we conducted a descriptive analysis by calculating the frequency and percentage distribution of each item in the codebase.

Aim 2: Examining Factors Influencing the Attractiveness of E-Cigarettes on Weibo

The codebook includes all independent variables for the analysis of influencing factors. To evaluate whether the items of the Weibo post characteristics in our codebook could influence the media effect of e-cigarette–related posts collected in our study, we used the Spearman bivariate correlation test to screen independent variables that were correlated with the numbers of reposts, comments, and likes. Then, we conducted 3 multiple linear regression analyses based on the number of reposts, comments, and likes. These analyses helped us identify the independent variables to be included in the final model. We used coefficient B (B) to indicate the amount of change in the dependent variables for one unit change in the number of reposts, comments, and likes. The SE measured the uncertainty of the regression coefficient estimates across the 3 models. We used standardized coefficient beta (β) to allow for a direct comparison of the relative influence of different variables in the regression analysis. T-value (T) tested the significance of the influence of independent variables on dependent variables. Variance inflation factor (VIF) detected multicollinearity between independent variables. A VIF of less than 10 is considered not to affect the model's stability. We conducted tests of statistical significance using $P<.05$ as a threshold to report statistical significance.

Ethical Considerations

This study observed the marketing strategies of e-cigarette companies and public behaviors on social media. This study did not involve human participants, as all data were acquired from public social media platforms. All the names of e-cigarette brands and users participating in the Weibo interaction are anonymous. Therefore, no additional ethical review was necessary.

Results

Sample of Posts

Figure 1 shows this study's screening process, which followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist. Multimedia Appendix 1 provides descriptive statistics for the characteristics of 863 posts from 32 selected Chinese e-cigarette Weibo accounts. Unlike traditional media, social media platforms facilitate individual interactions through comments and sharing. The number of reposts, comments, and likes reflected how e-cigarettes were perceived. Of the 863 Weibo posts, there were a total of 24,571 reposts, 8516 comments, and 22,272 likes. The average numbers of reposts, comments, and likes were 28.47 (SD 286.13), 9.87 (SD 41.27), and 25.81 (SD 237.63), respectively. Among these posts, the highest number of reposts, comments, and likes were 4903 (20%), 677 (7.9%), and 5029 (22.56%), respectively. The top 3 posts with the most reposts, comments, and likes all included interactive sweepstakes using e-cigarette and comparable products.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist that this study followed.

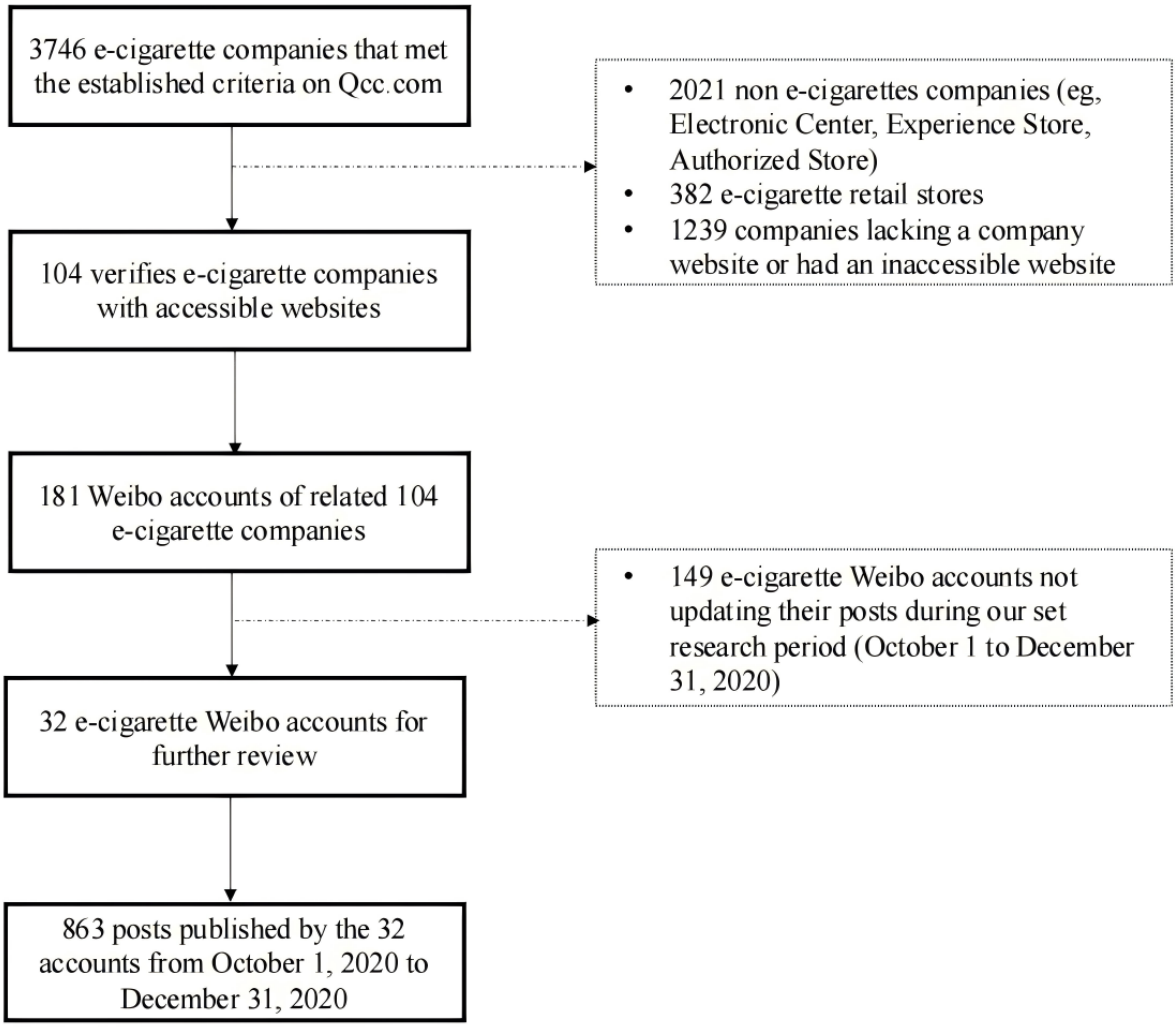

Product and Features

As Multimedia Appendix 1 shows, of the 863 Weibo posts, 449 (52%) posts described details of product features to attract others, including 351 (40.7%) posts relating to the shape of e-cigarettes and packaging attractiveness, 111 (12.9%) posts about flavors, and 48 (5.6%) posts highlighting the high quality and technique used in e-cigarettes. Out of 863 posts, 313 (36.3%) contained visible figures and 81 (9.4%) depicted fashion icons. Other visible characters with the highest percentage were everyday people (n=78, 9%), and the remaining were business people (n=26, 3%), sports people (n=16, 1.9%), celebrities (n=13, 1.5%), sales agents (n=12, 1.4%), and health professionals (n=2, 0.2%).

Sales and Promotions

A total of 101 (11.7%) posts had information on product promotions. Among them, 10 (1.2%) posts involved monetary promotional offers, including coupons, two-for-one offers, and discount sales. Moreover, 83 (9.6%) posts were related to nonmonetary promotions including giveaways and free samples, while 8 (0.9%) posts contained both incentives.

A total of 790 (91.5%) Weibo posts did not mention any information on purchase channels. The rest (n=73, 8.5%) still contained information on purchase links, phone numbers of offline purchases, and other purchase channels.

Social Contact and Interaction

Of the 863 posts, 548 (63.5%) interacted with fans, indicating that most posts tried to entice Weibo users to engage through various methods, including reposting, commenting, and clicking likes. A total of 538 (62.3%) posts featured hashtags; 41 (4.8%) encouraged fans to repost, comment, and like; and 29 (3.4%) interacted with fans by collecting stories and ideas.

E-cigarette–related posts generated interactions with Weibo users through a variety of topics. In 45.4% (n=392) of the posts, interaction topics were related to products. Current event topics accounted for 21.1% (n=182) of the posts. Topics related to interactive sweepstakes accounted for 8.7% (n=75). The post number of brand-related topics, “soul-soother” or emotional issues, and Weibo account promotion were 73 (8.5%), 15 (1.7%), and 1 (0.1%), respectively. Furthermore, joint promotion with other brands and promoting the Weibo accounts of brands by sending gifts were also the main themes of hashtag topics.

Restrictions and Warnings

Of the posts, 687 (79.6%) posts had no age restriction claims for minors to purchase or use e-cigarettes, 130 (15.1%) restricted minors’ use but did not limit their purchase, and 6 (0.7%)

XSL•FO
RenderX

restricted minors' purchases but did not limit their use. Only 40 (4.6%) posts restricted both purchases and use by minors. Meanwhile, only 24 (2.8%) posts provided health warnings.

We put all the variables in the codebook into the Spearman bivariate correlation test to screen out the independent variables that had a significant correlation with the dependent variable. These variables were included as independent variables in multiple linear regression. A multiple linear regression analysis was conducted to examine the effect of e-cigarette postmarketing strategy features on communication outcomes. The overall statistics of the 3 regression models are presented in Table 1. The results are presented in Table 2.

Table 1. Summary statistics of multiple linear regression models.

Models	Dependent variables	R^2	*F* test (*df*)	*P* value
Model 1	Number of reposts	0.183	9.224 (13)	<.001
Model 2	Number of comments	0.211	11.935 (12)	<.001
Model 3	Number of likes	0.204	10.512 (12)	<.001

XSL•FO
RenderX

Table 2. Results of multiple linear regression analysis for the influence of Weibo marketing strategies on the number of reposts, comments, and likes in China (N=863).

Model	Nonstandardized coeffificients		Standard coeffificients	T[a]	*P* value	VIF[b]
	B[c]	SE	B[d]			
Model 1						
Constant	–23.834	121.781	N/A[e]	–0.196	.85	N/A
Purchase channels mentioned	–160.161	54.865	–0.124	–2.919	.004	1.183
Nonmonetary incentives mentioned	145.616	110.899	0.139	1.313	.19	7.343
No monetary incentives mentioned	104.164	97.565	0.106	1.068	.47	1.202
Age restrictions mentioned	26.925	37.420	0.031	0.720	.47	1.202
Flavors mentioned	–66.201	45.943	–0.060	–1.441	.15	1.128
Function of helping reduce stress mentioned	12.966	119.122	0.005	0.109	.91	1.300
Function of improving social skills mentioned	–38.685	126.013	–0.019	–0.307	.76	2.440
Function of entertainment	–21.014	136.283	–0.009	–0.154	.88	2.199
Engagement via posts encouraging reposts, comments, and likes	426.267	73.226	0.314	5.821	<.001	1.897
Engagement topics related to products	49.217	38.181	0.062	1.289	.20	1.518
Engagement topics related to brands	120.660	46.003	0.115	2.623	.009	1.249
Engagement topics related to current affairs	62.660	36.701	0.083	1.707	.09	1.527
Engagement topics related to interactive sweepstakes	35.513	84.653	0.034	0.420	.68	4.327
Model 2						
Constant	–31.261	17.061	N/A	–1.832	.07	N/A
Purchase channels mentioned	–11.961	7.593	–0.068	–1.575	.12	1.248
Nonmonetary incentives mentioned	43.244	14.885	0.301	2.905	.004	7.283
No monetary incentives mentioned	19.737	13.098	0.147	1.507	.13	6.449
Age restrictions mentioned	18.580	5.056	0.155	3.675	<.001	1.208
Flavors mentioned	–9.652	6.260	–0.064	–1.542	.12	1.153
Engagement via stories and ideas collecting	37.903	8.805	0.173	4.305	<.001	1.094
Engagement topics related to products	18.462	5.183	0.170	3.562	<.001	1.540
Engagement topics related to brands	30.974	6.098	0.214	5.079	<.001	1.209
Engagement topics related to soul soother or emotional topics	1.487	12.691	0.005	0.117	.91	1.207
Engagement topics related to current affairs	15.813	4.965	0.152	3.185	.002	1.539
Engagement topics related to interactive sweepstakes	17.967	11.083	0.126	1.621	.11	4.084
Salesmen images appeared in the posts	0.002	0.005	0.019	0.461	.65	1.182
Model 3						

XSL•FO
RenderX

Model	Nonstandardized coeffificients		Standard coeffificients	T[a]	*P* value	VIF[b]
	B[c]	SE	B[d]			
Constant	579.228	130.244	N/A	4.447	<.001	N/A
Purchase channels mentioned	–17.749	44.922	–0.017	–0.395	.69	1.184
Nonmonetary incentives mentioned	–763.118	90.786	–0.880	–8.406	<.001	7.343
No incentives mentioned	–489.187	79.866	–0.603	–6.125	<.001	6.498
Age restrictions mentioned	5.712	31.559	0.008	0.181	.86	1.276
No health warnings mentioned	–68.884	75.751	–0.037	–0.909	.36	1.090
No product information mentioned	–74.531	32.205	–0.126	–2.314	.02	1.975
Function of entertainment	4.908	76.594	0.003	0.064	.95	1.036
Flavors mentioned	–64.256	39.541	–0.070	–1.625	.11	1.247
Engagement via stories and ideas collecting	90.489	51.911	0.068	1.743	.08	1.03
Engagement topics related to products	6.503	34.271	0.010	0.190	.85	1.824
Engagement topics related to brands	149.064	37.730	0.171	3.951	<.001	1.254
Engagement topics related to interactive sweepstakes	451.256	67.904	0.523	6.645	<.001	4.155
Engagement topics related to current affairs	22.219	30.368	0.035	0.732	.47	1.560

[a]T: T-value.

[b]VIF: variance inflation factor.

[c]B: coefficient B.

[d]β: standardized coefficient beta.

[e]N/A: not applicable.

Contents including engagement via posts encouraging reposts, comments, and likes (*P*<.001), and engagement topics related to e-cigarette brands were positively correlated with the number of reposts (*P*=.009). Posts highlighting nonmonetary incentives (*P*=.004), posts with age-restriction statements (*P*<.001), engaging via stories and idea collection (*P*<.001), and engagement topics related to products (*P*<.001) and current affairs (*P*=.002) had a positive effect on the number of comments. Engagement topics related to brands (*P*<.001) or interactive sweepstakes (*P*<.001) had a more significant positive effect on the number of likes than those without. Posts that mentioned purchase channels significantly negatively affected the number of reposts (*P*=.004) of e-cigarette posts. Posts that did not mention any incentives had a negative impact on receiving likes of posts (*P*<.001). Posts that had no product information had a negative correlation with the number of likes (*P*=.02).

Discussion

Principal Findings

We analyzed Weibo posts by e-cigarette brands and gained valuable insight into how e-cigarette promotion information spreads and proliferates via social media. Weibo posts by 32 official e-cigarette accounts collectively received many likes, comments, and reposts, gaining widespread attention. Our results show that e-cigarette content exposure on social media may attract Weibo users' attention and can successfully shape more positive attitudes toward e-cigarettes. Existing research has also shown that social media marketing of e-cigarettes has a particularly positive impact on teenagers' attitudes and risk behavior [49,50], which may pose a threat and challenge to youth e-cigarette use and global public health.

Marketing Tactics on Chinese Social Media

We observed several marketing tactics employed by e-cigarette companies on Chinese social media, documented as emphasizing attractive product features, using trendy characters, implicit promotions, downplaying health concerns, and engaging with Weibo users in various ways. These posts were visualized through pictures or videos, catering to teenagers' reading habits and the demand for more impactful content [51,52]. The persuasive power of images was crucial in making e-cigarette-related information more attractive [42]. Few posts mentioned health warnings and age restrictions.

Highlighting attractive product characteristics was one of the most significant e-cigarette marketing tactics on Weibo. Approximately half of the Weibo posts emphasized product characteristics, particularly showcasing unique shapes and attractive packaging, which may appeal to teenagers and lower

XSL•FO
RenderX

their perceptions of e-cigarette risks [38,39]. Even though only a small number of posts mentioned flavors, the promotion of flavors through beautifully designed posters and appealing flavor descriptions should be further monitored. This finding revealed that flavor features attract teenagers [22,40-42], as it is the most common reason for youth e-cigarette use [43]. In particular, some e-cigarette brands marketed flavors that represent Chinese culture, such as green tea. To the best of our knowledge, this is the first study to identify a marketing strategy that builds a responsible corporate image by promoting efforts to protect traditional Chinese culture. Among all the posts with images of people, fashion icons and ordinary people accounted for the largest proportion. The latest research suggests that attention should be paid to graphic characters in e-cigarette advertisements, as fashion models or ordinary people may glorify or normalize e-cigarette use among teenagers [53].

Even after online sales supervision was tightened in 2020 [19], a small number of e-cigarette companies continue to provide relevant promotional information on Weibo and continue to look for implicit promotion methods. The impact of nonmonetary incentives was uncertain in our findings. However, the existing published scoping reviews reveal that e-cigarette companies worldwide use monetary and nonmonetary incentives, such as promotional codes, discounts for new customers, coupons, and giveaways, in addition to e-cigarette advertising [44,54]. Double incentive strategies likely create excitement and reciprocity, making teenagers feel grateful toward the brand, thus promoting a positive perception of that e-cigarette brand [55].

The scarcity of age restrictions and health warnings in e-cigarette posts was common, potentially reducing Weibo users' sense of caution. Previous studies have also highlighted the lack of emphasis on health issues and warnings by e-cigarette companies on social media [1,5,21,56,57]. Previous studies have found nicotine warnings of different designs could influence young people's perception of harm and use intention to varying degrees [58]. In our study, we found relevant statements were in small print and vague, which might not draw enough attention or mislead teens. Therefore, warnings and statements should be carefully designed to alert or minimize appeal to teenagers. Additionally, such restrictions should be strictly regulated and monitored on social media.

One remarkable new finding regarding online engagement is that interactive sweepstakes on Weibo significantly expanded the influence of e-cigarette brands. In our study, the 3 posts with the highest reposts, comments, and likes all had interactive sweepstakes. Web-based contests and sweepstakes are also frequently utilized on social media platforms favored by young people in other countries [44,46]. A previous study emphasized the success of sweepstake-based marketing in enhancing user attention and engagement [59]. Reposts, comments, and likes are measures of user engagement [45,60] and indicators to increase the exposure of posts [61,62]. Web-based interactivity is a unique feature of social media that can create greater brand exposure and attention. Digital e-cigarette marketing involves various interactive methods to foster greater engagement. Our results show that various types of interaction positively impact the media effects of e-cigarette Weibo posts to varying degrees.

The Establishment of Community Culture on Weibo

Based on the content analysis, our study found that e-cigarette companies constructed a unique culture to appeal to both e-cigarette users and potential users on Weibo. Interactive strategies on social media help e-cigarette brands establish an online presence that posits e-cigarettes as fashionable trend and lifestyle choice, potentially creating a community and identity for vapers and fostering a sense of belonging and loyalty. This conclusion is consistent with the existing research [63].

Our results uncovered many posts with hashtags, similar to the results of other studies [4,21]. Many e-cigarette Weibo accounts post hashtags with their brand name to build a vaping community associated with a specific e-cigarette brand. Using brand-specific hashtags can further construct social bonds and common beliefs around e-cigarettes [1,64]. Posts including hashtags unrelated to tobacco directly are common [4], and existing studies revealed vaping posts contained health lifestyle hashtags [65]; thus, teenagers may view specific hashtags posted by e-cigarette companies without searching for e-cigarette-related content, which can greatly expand a post's reach. For example, one of the posts included hashtags such as #daretoimagine ("Gan Yu Xiang Xiang" in Chinese). Moreover, Chinese e-cigarette marketing on Weibo combines mainstream culture and subculture through hashtags related to current events. For example, some hashtags linked the new technology of e-cigarette products to China's overall technological power. Through this marketing strategy, e-cigarette Weibo posts achieve broader coverage.

Implications for Policy Making

China gradually strengthened regulations regarding e-cigarette marketing after this study was completed. The latest policy tightens restrictions on the manufacture and sale of e-cigarettes, specifically prohibiting the sale of flavored e-cigarettes, tobacco flavors, and private company atomized e-cigarettes [20]. After the ban, there has been a massive decline in the e-cigarette marketing strategies we observed in this study. However, more implicit posting, instead of direct promotion content on social media, may be a way to circumvent these restrictions. Until May 10, 2024, among all the 32 Weibo accounts of e-cigarette companies we investigated, 6 (18.9%) Weibo accounts continued updating posts on Weibo after the published ban in 2022 [20], and 1 (3.1%) e-cigarette company is still updating posts on Weibo as of April 2024. The content posted by these e-cigarette companies on Weibo did not directly promote e-cigarettes. Rather, most posts described e-cigarette flavors covertly via references to coffee, desserts, and other food-related items.

Our findings suggest that hidden online promotional content poses a significant threat and challenge to youth e-cigarette use and global public health. Thus, regulatory policies should be implemented to protect teenagers from marketing information regarding e-cigarettes and other tobacco products. Constant surveillance and nationwide legislation of internet marketing activities are essential to protect teenagers from ambiguous, inaccurate, or even deceptive information about e-cigarettes, as teenagers and other vulnerable populations are more susceptible to marketing claims [31,32,66]. There should be clear legal provisions that e-cigarette–related information must include

XSL•FO
RenderX

explicit health claims and age restriction statements in large print [67,68]. Furthermore, more relevant health education activities and campaigns for teenagers should be conducted, and health communication media should give more prominence to issues related to e-cigarette prevention and control.

In response to the rapid growth of e-cigarette marketing, China has been intensifying its efforts to regulate e-cigarette internet marketing, which could serve as a reference for enhancing electronic cigarette oversight globally.

Limitations

This study had several limitations that need to be addressed in future research. First, our study focused on the official Weibo accounts of e-cigarette brands, neglecting other stakeholders and social media platforms involved in e-cigarette social media marketing. Second, our data collection did not consider the possibility of data manipulation by e-cigarette companies. Third, although the existing research shows that e-cigarette marketing is mainly targeted toward teenagers, we could not confirm whether the e-cigarette Weibo marketing strategies are directly reflected in this age group due to the anonymity on social media. Direct evidence of the relationship between Weibo e-cigarette marketing and its impact on teenagers still needs to be teased out. Future research could gather data from more countries and platforms to conduct a more internationally comparative study.

Conclusion

The findings reveal marketing tactics used by e-cigarette companies on Chinese social media, and their influence on the numbers of reposts, comments, and likes on posts. Despite nationwide regulations, e-cigarette companies are finding ways to strengthen their marketing presence on Weibo and indirectly target adolescents, often without health warnings or age restriction disclaimers. E-cigarette marketing on social media requires global attention and coordinated efforts. Our study provides insights into how to strengthen the supervision of invisible marketing on social media of e-cigarette companies.

Acknowledgments

We would like to thank the study communities and undergraduate students WJD and JYJ for their involvement in the data collection and coding. This study was funded by the Shanghai Municipal Health Commission Health Promotion Young Talent Grant (JKKPYC-2022-01), the Bloomberg Initiative to Reduce Tobacco Use (CHINA-26-12), and the National Natural Science Foundation of China (82173635). The Tobacco Free Kids Action Fund provided technical support and played an important role in conducting this research and preparing this article.

Data Availability

Data are available on reasonable request.

Authors' Contributions

FW, PPZ, XYH, and HD discussed, conceptualized, and designed the study. LF and XYH collected the data. XYZ and XYH discussed and performed statistical analysis. XYZ, XYH, and YHC wrote the initial draft. XYZ, XYH, YHC, FW, PPZ, LYZ, and XTY redrafted the paper. XYZ completed the revision draft. FW is responsible for the overall content as the guarantor. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Codebook and descriptive statistics of marketing characteristics of 863 posts from 32 selected Chinese e-cigarette Weibo accounts (October 1 to December 31, 2020).
[DOCX File , 20 KB - jmir_v26i1e51594_app1.docx]

References

1. Czaplicki L, Kostygina G, Kim Y, Perks SN, Szczypka G, Emery SL, et al. Characterising JUUL-related posts on Instagram. Tob Control 2020 Nov;29(6). [doi: 10.1136/tobaccocontrol-2018-054824] [Medline: 31266903]
2. Shields PG, Berman M, Brasky TM, Freudenheim JL, Mathe E, McElroy JP, et al. A review of pulmonary toxicity of electronic cigarettes in the context of smoking: a focus on inflammation. Cancer Epidemiol Biomarkers Prev 2017 Aug;26(8):1175-1191 [FREE Full text] [doi: 10.1158/1055-9965.EPI-17-0358] [Medline: 28642230]
3. Muthumalage T, Prinz M, Ansah KO, Gerloff J, Sundar IK, Rahman I. Inflammatory and oxidative responses induced by exposure to commonly used e-cigarette flavoring chemicals and flavored e-liquids without nicotine. Front Physiol 2017;8:1130 [FREE Full text] [doi: 10.3389/fphys.2017.01130] [Medline: 29375399]
4. O'Brien EK, Hoffman L, Navarro MA, Ganz O. Social media use by leading US e-cigarette, cigarette, smokeless tobacco, cigar and hookah brands. Tob Control 2020 Dec;29(e1):e87-e97. [doi: 10.1136/tobaccocontrol-2019-055406] [Medline: 32217772]

XSL•FO
RenderX

5. Chen Y, Sun S, Zhao X, Zhou H, Wang F. Objective facts or misleading hype? Associations between features of e-cigarette marketing and sales on a Chinese e-commerce platform. IJERPH 2020 Sep 15;17(18):6711. [doi: 10.3390/ijerph17186711]
6. More than 2.5 million youth reported e-cigarette use in 2022. US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/media/releases/2022/p1007-e-cigarette-use.html [accessed 2023-08-23]
7. Tobacco epidemic surveillance results released for Chinese high school and university students in 2021. Chinese Center For Disease Control And Prevention. 2022 Mar 16. URL: https://www.chinacdc.cn/jkzt/sthd_3844/slhd_12885/ [accessed 2024-10-07]
8. Tobacco and nicotine industry tactics addict youth for life. World Health Organization. 2024 May 23. URL: https://www.who.int/news/item/23-05-2024-tobacco-and-nicotine-industry-tactics-addict-youth-for-life [accessed 2024-10-07]
9. Navarro MA, O'Brien EK, Ganz O, Hoffman L. Influencer prevalence and role on cigar brand Instagram pages. Tob Control 2021 Nov;30(e1):e33-e36. [doi: 10.1136/tobaccocontrol-2020-055994] [Medline: 33046583]
10. Vogel EA, Guillory J, Ling PM. Sponsorship disclosures and perceptions of e-cigarette Instagram posts. Tob Regul Sci 2020 Sep 01;6(5):355-368 [FREE Full text] [doi: 10.18001/trs.6.5.5] [Medline: 33778107]
11. Vassey J, Valente T, Barker J, Stanton C, Li D, Laestadius L, et al. E-cigarette brands and social media influencers on Instagram: a social network analysis. Tob Control 2022 Feb 07:184-191 [FREE Full text] [doi: 10.1136/tobaccocontrol-2021-057053] [Medline: 35131947]
12. Chu K, Unger JB, Cruz TB, Soto DW. Electronic cigarettes on Twitter - spreading the appeal of flavors. Tob Regul Sci 2015 Apr;1(1):36-41 [FREE Full text] [doi: 10.18001/TRS.1.1.4] [Medline: 27853734]
13. Kong G, Kuguru KE, Krishnan-Sarin S. Gender differences in U.S. adolescent e-cigarette use. Curr Addict Rep 2017 Dec;4(4):422-430 [FREE Full text] [doi: 10.1007/s40429-017-0176-5] [Medline: 29545987]
14. Dashtian H, Murthy D, Kong G. An exploration of e-cigarette-related search items on YouTube: network analysis. J Med Internet Res 2022 Jan 27;24(1):e30679 [FREE Full text] [doi: 10.2196/30679] [Medline: 35084353]
15. Donaldson SI, Dormanesh A, Perez C, Zaffer MO, Majmundar A, Unger JB, et al. Monitoring the official YouTube channels of e-cigarette companies: a thematic analysis. Health Educ Behav 2023 Jan 20;50(5):677-682. [doi: 10.1177/10901981221148964]
16. Galimov A, Vassey J, Galstyan E, Unger JB, Kirkpatrick MG, Allem J. Ice flavor-related discussions on Twitter: content analysis. J Med Internet Res 2022 Nov 30;24(11):e41785 [FREE Full text] [doi: 10.2196/41785] [Medline: 36449326]
17. Hu Y. Electronic cigarettes instead of traditional cigarettes, regulation can not be left blank. Guangming Online. 2018 Jun 29. URL: https://guancha.gmw.cn/2018-06/29/content_29547797.htm [accessed 2024-10-07]
18. Zhong S, Wang Y, Chen R, Zhou L. E-cigarette use among youth in China. Lancet Reg Health - West Pac 2022 Jul;24:100504. [doi: 10.1016/j.lanwpc.2022.100504]
19. The special inspection action plan for the e-cigarette market. State Tobacco Monopoly Administration. 2020 Jul 21. URL: https://www.gov.cn/zhengce/zhengceku/2020-07/15/content_5526939.htm [accessed 2024-10-06]
20. State Tobacco Monopoly Administration on strengthening the supervision of e-cigarettes related matters notice. State Tobacco Monopoly Administration. 2022 Mar 21. URL: https://www.gov.cn/gongbao/content/2022/content_5697988.htm [accessed 2024-10-07]
21. Alpert JM, Chen H, Riddell H, Chung YJ, Mu YA. Vaping and Instagram: a content analysis of e-cigarette posts using the Content Appealing to Youth (CAY) index. Subst Use Misuse 2021;56(6):879-887. [doi: 10.1080/10826084.2021.1899233] [Medline: 33749515]
22. McCausland K, Maycock B, Leaver T, Jancey J. The messages presented in electronic cigarette-related social media promotions and discussion: scoping review. J Med Internet Res 2019 Feb 05;21(2):e11953 [FREE Full text] [doi: 10.2196/11953] [Medline: 30720440]
23. McCausland K, Maycock B, Leaver T, Wolf K, Freeman B, Thomson K, et al. E-cigarette promotion on Twitter in Australia: content analysis of tweets. JMIR Public Health Surveill 2020 Nov 05;6(4):e15577 [FREE Full text] [doi: 10.2196/15577] [Medline: 33151159]
24. Wulan W, Kusuma D, Nurjanah N, Aprianti A, Ahsan A. Is exposure to social media advertising and promotion associated with e-cigarette use? Evidence from Indonesia. Asian Pac J Cancer Prev 2022 Apr 01;23(4):1257-1262. [doi: 10.31557/apjcp.2022.23.4.1257]
25. Cho YJ, Thrasher JF, Driezen P, Hitchman SC, Reid JL, Hammond D. Trends in exposure to and perceptions of e-cigarette marketing among youth in England, Canada and the United States between 2017 and 2019. Health Educ Res 2022 Jan 29;36(6):657-668 [FREE Full text] [doi: 10.1093/her/cyab039] [Medline: 34849859]
26. Vogel EA, Ramo DE, Rubinstein ML, Delucchi KL, Darrow S, Costello C, et al. Effects of Social Media on Adolescents' Willingness and Intention to Use E-Cigarettes: An Experimental Investigation. Nicotine Tob Res 2020 Jan 08:694-701. [doi: 10.1093/ntr/ntaa003] [Medline: 31912147]
27. Camenga D, Gutierrez KM, Kong G, Cavallo D, Simon P, Krishnan-Sarin S. E-cigarette advertising exposure in e-cigarette naïve adolescents and subsequent e-cigarette use: a longitudinal cohort study. Addict Behav 2018 Jun;81:78-83 [FREE Full text] [doi: 10.1016/j.addbeh.2018.02.008] [Medline: 29432916]
28. Weibo 2020 user development report. Sina Weibo Data Center. 2021 Mar 12. URL: https://data.weibo.com/report/reportDetail?id=456 [accessed 2024-10-07]

XSL•FO
RenderX

29. Bell K, Keane H. Nicotine control: E-cigarettes, smoking and addiction. Int J Drug Policy 2012 May;23(3):242-247. [doi: 10.1016/j.drugpo.2012.01.006] [Medline: 22365155]
30. Teens, social media and technology 2018. Pew Research Center. 2018 Mar 18. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2024-10-07]
31. Perikleous EP, Steiropoulos P, Paraskakis E, Constantinidis TC, Nena E. E-cigarette use among adolescents: an overview of the literature and future perspectives. Front Public Health 2018;6 [FREE Full text] [doi: 10.3389/fpubh.2018.00086] [Medline: 29632856]
32. Cho H, Li W, Shen L, Cannon J. Mechanisms of social media effects on attitudes toward e-cigarette use: motivations, mediators, and moderators in a national survey of adolescents. J Med Internet Res 2019 Jun 27;21(6):e14303 [FREE Full text] [doi: 10.2196/14303] [Medline: 31250830]
33. Koval R, Willett J, Briggs J. Potential benefits and risks of high-nicotine e-cigarettes. JAMA 2018 Oct 09;320(14):1429. [doi: 10.1001/jama.2018.12328]
34. Qi Cha Cha. 2021 Jan 21. URL: https://www.qcc.com/ [accessed 2021-01-31]
35. Pokhrel P, Fagan P, Herzog TA, Laestadius L, Buente W, Kawamoto CT, et al. Social media e-cigarette exposure and e-cigarette expectancies and use among young adults. Addict Behav 2018 Mar;78:51-58 [FREE Full text] [doi: 10.1016/j.addbeh.2017.10.017] [Medline: 29127784]
36. Sears CG, Walker KL, Hart JL, Lee AS, Siu A, Smith C. Clean, cheap, convenient: promotion of Electronic cigarettes on YouTube. Tob Prev Cessat 2017 Apr;3 [FREE Full text] [doi: 10.18332/tpc/69393] [Medline: 28725876]
37. O'Brien EK, Navarro MA, Hoffman L. Mobile website characteristics of leading tobacco product brands: cigarettes, smokeless tobacco, e-cigarettes, hookah and cigars. Tob Control 2018 Aug 30. [doi: 10.1136/tobaccocontrol-2018-054549] [Medline: 30166427]
38. Hardie L, McCool J, Freeman B. E-cigarette retailers' use of Instagram in New Zealand: a content analysis. IJERPH 2023 Jan 19;20(3):1897. [doi: 10.3390/ijerph20031897]
39. Stanton CA, Pasch KE, Pericot-Valverde I, Cruz-Cano R, Moran MB, Abadi MH, et al. Longitudinal associations between U.S. youth exposure to e-cigarette marketing and e-cigarette use harm perception and behavior change. Preventive Medicine 2022 Nov;164:107266. [doi: 10.1016/j.ypmed.2022.107266]
40. Soule EK, Sakuma KK, Palafox S, Pokhrel P, Herzog TA, Thompson N, et al. Content analysis of internet marketing strategies used to promote flavored electronic cigarettes. Addict Behav 2019 Apr;91:128-135. [doi: 10.1016/j.addbeh.2018.11.012] [Medline: 30606627]
41. Allem J, Majmundar A, Dharmapuri L, Cruz TB, Unger JB. E-liquid-related posts to Twitter in 2018: thematic analysis. Addict Behav Rep 2019 Dec;10:100196 [FREE Full text] [doi: 10.1016/j.abrep.2019.100196] [Medline: 31431917]
42. Lee J, Suttiratana SC, Sen I, Kong G. E-cigarette marketing on social media: a scoping review. Curr Addict Rep 2023 Jan 16;10(1):29-37. [doi: 10.1007/s40429-022-00463-2]
43. Kong G, Morean ME, Cavallo DA, Camenga DR, Krishnan-Sarin S. Reasons for electronic cigarette experimentation and discontinuation among adolescents and young adults. Nicotine Tob Res 2015 Jul;17(7):847-854 [FREE Full text] [doi: 10.1093/ntr/ntu257] [Medline: 25481917]
44. Alpert JM, Jaisle A, Chen H. A content analysis of the promotional strategies employed by e-cigarette brands on Twitter. Health Mark Q 2019;36(4):307-321. [doi: 10.1080/07359683.2019.1680121] [Medline: 31696789]
45. Chu K, Allem J, Cruz TB, Unger JB. Vaping on Instagram: cloud chasing, hand checks and product placement. Tob Control 2016 Sep;26(5) [FREE Full text] [doi: 10.1136/tobaccocontrol-2016-053052] [Medline: 27660111]
46. Moran MB, Heley K, Pierce JP, Niaura R, Strong D, Abrams D. Ethnic and socioeconomic disparities in recalled exposure to and self-reported impact of tobacco marketing and promotions. Health Commun 2019 Mar;34(3):280-289 [FREE Full text] [doi: 10.1080/10410236.2017.1407227] [Medline: 29236530]
47. Amalia B, Kapoor S, Sharma R, Fu M, Fernández E, Rana JS. Online sales compliance with the electronic cigarettes ban in India: a content analysis. Int J Public Health 2020 Nov;65(8):1497-1505. [doi: 10.1007/s00038-020-01480-6] [Medline: 33034737]
48. Laestadius LI, Wahl MM, Cho YI. #Vapelife: an exploratory study of electronic cigarette use and promotion on Instagram. Subst Use Misuse 2016 Dec 14;51(12):1669-1673. [doi: 10.1080/10826084.2016.1188958] [Medline: 27484191]
49. Booth P, Albery IP, Cox S, Frings D. Survey of the effect of viewing an online e-cigarette advertisement on attitudes towards cigarette and e-cigarette use in adults located in the UK and USA: a cross-sectional study. BMJ Open 2019 Jun 18;9(6):e027525. [doi: 10.1136/bmjopen-2018-027525]
50. Marynak K, Robichaud M, Puryear T, Kennedy R, Moran M. #Nicotineaddiction on TikTok: A quantitative content analysis of top-viewed posts. Tob Induc Dis 2022 Aug 5;20(August):69-69 [FREE Full text] [doi: 10.18332/tid/151868] [Medline: 35979474]
51. E-commerce News. Youth social media use report 2022: TikTok and YouTube are taking over young people. E-commerce News. 2022 Sep 28. URL: https://www.dsb.cn/196736.html [accessed 2024-10-07]
52. Struik LL, Dow-Fleisner S, Belliveau M, Thompson D, Janke R. Tactics for drawing youth to vaping: content analysis of electronic cigarette advertisements. J Med Internet Res 2020 Aug 14;22(8):e18943 [FREE Full text] [doi: 10.2196/18943] [Medline: 32663163]

XSL•FO
RenderX

53. Chen-Sankey J, Jeong M, Wackowski OA, Unger JB, Niederdeppe J, Bernat E, et al. Noticing people, discounts and non-tobacco flavours in e-cigarette ads may increase e-cigarette product appeal among non-tobacco-using young adults. Tob Control 2023 Dec 13;33(1):30-37. [doi: 10.1136/tobaccocontrol-2022-057269] [Medline: 35672144]
54. Collins L, Glasser AM, Abudayyeh H, Pearson JL, Villanti AC. E-cigarette marketing and communication: how e-cigarette companies market e-cigarettes and the public engages with e-cigarette information. Nicotine Tob Res 2019 Jan 01;21(1):14-24 [FREE Full text] [doi: 10.1093/ntr/ntx284] [Medline: 29315420]
55. Palmatier RW, Jarvis CB, Bechkoff JR, Kardes FR. The role of customer gratitude in relationship marketing. J Mark 2009 Sep 01;73(5):1-18. [doi: 10.1509/jmkg.73.5.1]
56. Laestadius LI, Wahl MM, Vassey J, Cho YI. Compliance with FDA nicotine warning statement provisions in e-liquid promotion posts on Instagram. Nicotine Tob Res 2020 Oct 08;22(10):1823-1830 [FREE Full text] [doi: 10.1093/ntr/ntaa092] [Medline: 32433737]
57. Romito LM, Hurwich RA, Eckert GJ. A snapshot of the depiction of electronic cigarettes in YouTube videos. Am J Health Behav 2015 Nov;39(6):823-831. [doi: 10.5993/AJHB.39.6.10] [Medline: 26450550]
58. Sontag J, Manderski MTB, Hammond D, Wackowski OA. US young adults' perceived effectiveness of draft pictorial e-cigarette warning labels. Tob Control 2019 Jun 05;28(e1):e49-e51. [doi: 10.1136/tobaccocontrol-2018-054802]
59. Jung W, Yang S, Kim H. Design of sweepstakes-based social media marketing for online customer engagement. Electron Commer Res 2019 Jan 4;20(1):119-146. [doi: 10.1007/s10660-018-09329-0]
60. Vassey J, Metayer C, Kennedy CJ, Whitehead TP. #Vape: measuring e-cigarette influence on Instagram with deep learning and text analysis. Front Commun (Lausanne) 2020;4 [FREE Full text] [doi: 10.3389/fcomm.2019.00075] [Medline: 35233388]
61. Sherman LE, Greenfield PM, Hernandez LM, Dapretto M. Peer influence via Instagram: effects on brain and behavior in adolescence and young adulthood. Child Dev 2018 Jan;89(1):37-47 [FREE Full text] [doi: 10.1111/cdev.12838] [Medline: 28612930]
62. Martinez-Pecino R, Garcia-Gavilán M. Likes and problematic Instagram use: the moderating role of self-esteem. Cyberpsychol Behav Soc Netw 2019 Jun;22(6):412-416. [doi: 10.1089/cyber.2018.0701] [Medline: 31107106]
63. Fournier S, Lee L. Getting brand communities right. Harvard Business Review. 2009 Apr 30. URL: https://hbr.org/2009/04/getting-brand-communities-right [accessed 2023-08-23]
64. Allem J, Ferrara E, Uppu SP, Cruz TB, Unger JB. E-cigarette surveillance with social data: social bots, emerging topics, and trends. JMIR Public Health Surveill 2017 Dec 20;3(4):e98 [FREE Full text] [doi: 10.2196/publichealth.8641] [Medline: 29263018]
65. Basáñez T, Majmundar A, Cruz TB, Unger JB. Vaping associated with healthy food words: a content analysis of Twitter. Addict Behav Rep 2018 Dec;8:147-153 [FREE Full text] [doi: 10.1016/j.abrep.2018.09.007] [Medline: 30320201]
66. Dai L, He Y, Tan Y, Yu Z, Zhu J. Online e-cigarette information exposure and its association with e-cigarette use among adolescents in Shanghai, China. IJERPH 2022 Mar 11;19(6):3329. [doi: 10.3390/ijerph19063329]
67. Guillory J, Kim AE, Fiacco L, Cress M, Pepper J, Nonnemaker J. An experimental study of nicotine warning statements in e-cigarette tweets. Nicotine Tob Res 2020 Apr 21;22(5):814-821 [FREE Full text] [doi: 10.1093/ntr/ntz029] [Medline: 30820571]
68. Lee H, Lin H, Seo D, Lohrmann DK. The effect of e-cigarette warning labels on college students' perception of e-cigarettes and intention to use e-cigarettes. Addict Behav 2018 Jan;76:106-112. [doi: 10.1016/j.addbeh.2017.07.033]

Abbreviations

ENDS: electronic nicotine delivery systems
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
T: T-value
VIF: variance inflation factor
WHO: World Health Organization

Edited by A Mavragani; submitted 06.08.23; peer-reviewed by E de Quincey, D Dinh; comments to author 09.01.24; revised version received 10.05.24; accepted 18.09.24; published 07.11.24.

Please cite as:
Zhou X, Hao X, Chen Y, Deng H, Fang L, Zhang L, Yan X, Zheng P, Wang F
Social Media Marketing Strategies for Electronic Cigarettes: Content Analysis of Chinese Weibo Accounts
J Med Internet Res 2024;26:e51594
URL: https://www.jmir.org/2024/1/e51594
doi:10.2196/51594
PMID:

XSL•FO
RenderX

©Xinyi Zhou, Xinyu Hao, Yuhang Chen, Hui Deng, Ling Fang, Lingyun Zhang, Xiaotao Yan, Pinpin Zheng, Fan Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tobacco and Alcohol Content in Top Vietnamese YouTube Music Videos: Content Analysis

Thi Phuong Thao Tran[1,2], PhD; Thu Trang Vu[2], MPhil; Yachao Li[3], PhD; Lucy Popova[1], PhD

[1]School of Public Health, Georgia State University, Atlanta, GA, United States

[2]Hanoi University of Public Health, Hanoi, Vietnam

[3]The College of New Jersey, Ewing Township, NJ, United States

Corresponding Author:
Thi Phuong Thao Tran, PhD
School of Public Health, Georgia State University
140 Decatur St. SE
Atlanta, GA, 30303
United States
Phone: 1 4709626086
Fax: 1 4044131140
Email: ttran217@gsu.edu

Abstract

Background: Seeing portrayals of tobacco and alcohol in music videos (MVs) may reduce perceived risks, increase susceptibility, and lead to the initiation of tobacco and alcohol use among adolescents and young adults. Previous studies have predominantly concentrated on assessing tobacco and alcohol contents in English-language MVs within Western countries. However, many other countries have not only been influenced by the English music market but have also produced music in their native languages, and this content remains underexamined.

Objective: This study aims to investigate the prevalence of tobacco- and alcohol-related content in top Vietnamese MVs on YouTube from 2013 to 2021, to describe how tobacco and alcohol are portrayed in these MVs, and to examine associations between these portrayals and MV characteristics.

Methods: A total of 410 Vietnamese MVs, including the top 40 or 50 most viewed released each year between 2013 and 2021, were analyzed. General information, such as the song name, its release date and ranking, age restriction, musical genre, and type of MV, was collected. We examined tobacco and alcohol content in the MVs, with specific details such as tobacco types, their brands, as well as the number, age, sex, and roles of individuals smoking or drinking.

Results: Among the 410 MVs, 36 (8.8%) contained tobacco-related content and 136 (33.2%) featured alcohol-related content. Additionally, 28 (6.8%) out of 410 MVs included both tobacco and alcohol content. The prevalence of videos with tobacco and alcohol content fluctuated over the years. In MVs with tobacco-related content, a higher proportion of hip-hop or rap songs contained tobacco-related content (n=6, 30%) compared to other music genres. In MVs with tobacco-related content, cigarettes were the most frequently shown product (n=28, 77.8%), and smoking scenes were often depicted at parties (n=13, 36.1%) and during dancing and singing scenes (n=12, 33.3%). Among the 31 MVs portraying actual tobacco use, tobacco use was typically depicted with 1 person, often a young adult male, while 38.7% (n=12) showed singer(s) smoking. For MVs with alcohol-related content, there was a high proportion showing alcohol images at parties, bars, or pubs (n=96, 70.6%). Among 87 MVs containing drinking scenes, 60.9% (n=53) involved groups of young adults of both sexes, and 64% (n=56) depicted singers drinking. Additionally, only 2 (5.6%) MVs included health warnings about tobacco harm, and 2 MVs (1.5%) included warnings about drinking restricted to individuals 18 years and above.

Conclusions: The notable prevalence of tobacco and alcohol content in leading Vietnamese YouTube MVs raises concerns, especially as most of this content is portrayed without any warnings. The study underscores a regulatory gap in addressing such content on the internet, emphasizing the urgent need for stricter regulations and age restrictions on platforms such as YouTube.

(J Med Internet Res 2024;26:e55555) doi:10.2196/55555

XSL•FO
RenderX

KEYWORDS

risk; risk factor; tobacco content; alcohol content; tobacco; alcohol; tobacco portrayal; alcohol portrayal; music video; Vietnam; Vietnamese; YouTube; social media; socials; youth; adolescent; teen; teenager; young adult

Introduction

The tobacco and alcohol industry has invested billions of dollars in a wide range of direct and indirect advertising, promotion, and sponsorship (APS) tactics to attract people worldwide to purchase its products [1,2]. These tactics encompass public entertainment, sponsoring sports events, advertising on radio, television, magazines, newspapers, point-of-sale, and other forms such as distributing free samples, loyalty schemes, and coupons. The negative impacts of APS tactics have been well documented, including smoking and drinking initiation [3-6], higher levels of tobacco and alcohol consumption [4,7-9], and the fostering of positive attitudes or misunderstandings about tobacco and alcohol use [10-12]. Therefore, prohibiting APS is a key element of tobacco control and one of the "best buys" for reducing the burden of harmful alcohol use [13,14].

However, advertising has extended well beyond traditional media in recent years thanks to the borderless growth of digital marketing [15,16]. With the rapid expansion of social media platforms, the tobacco and alcohol industries have increasingly leveraged them as powerful marketing tools, often through partnerships with celebrities and influencers to promote their products [17,18]. YouTube, the largest video-sharing platform and most popular social media among young people (used by 95% of teens in the United States), has served as an effective marketing medium through its music videos (MVs) [17,19]. For instance, in England, YouTube MVs delivered approximately 1006 million gross alcohol impressions (95% CI 748-1264) and 203 million tobacco impressions (95% CI 151- 255) to the British population in 2016 [20]. Alcohol- and tobacco-related imagery and lyrics have been blended into numerous MVs, indirectly and subtly promoting these unhealthy behaviors. A prior study found that some MVs overtly encouraged excessive drinking and drunkenness without showing any negative consequences to the drinker [21]. As a result, watching these MVs, particularly those with positive alcohol portrayals, is linked to adolescent drinking initiation and increased drinking [22].

Previous studies have primarily examined tobacco and alcohol content in English-language MVs within Western countries. However, many other countries have not only been influenced by the English music market but have also created music in their native languages, often accompanied by MVs. There is a lack of research on the prevalence of tobacco and alcohol content and the portrayal of related features in MVs in non-English languages [23].

In Vietnam, there are approximately 65 million social media users, with YouTube reigning supreme as the dominant platform for sharing MVs [24]. YouTube is also the second most widely used social media platform after Facebook, particularly among children, adolescents, and young adults who consider it their primary choice for music consumption [24]. Vietnam ranks among the top 10 countries worldwide in cigarette consumption, with approximately 15.6 million smokers [25]. It also ranks among the top 3 countries worldwide in terms of alcohol per capita consumption [26]. Given these factors, we anticipate that tobacco- and alcohol-related content is common in MVs in Vietnam. However, to date, no studies have examined the presence of such content in MVs. Therefore, our study aims to investigate the prevalence of such content in top Vietnamese MVs on YouTube over 9 years (2013 to 2021). Additionally, this study will describe other features related to the depiction of tobacco and alcohol in these MVs.

Methods

This study followed the guidelines for content analysis studies laid out by Neuendorf (2017) [27].

Study Sample

The study sample consisted of the top 40 or 50 most viewed Vietnamese MVs released each year between 2013 and 2021 (since 2017, the top 50 MVs were listed instead of the top 40, and 2021 was the most recently available full-year data at the time of the manuscript submission in 2023) [28]. These selections were determined based on viewership data derived from the domestic music chart available on YouTube. A total of 410 MVs were included in this study. The list of most viewed MVs from 2013 to 2017 was last updated in December 2021, while the list of those from 2018 to 2021 was last updated in January 2022.

Among the 410 MVs, 8 original MVs were removed from YouTube for unknown reasons. Because we focused on the MV content, we replaced these missing MVs with the same ones uploaded in later years. When a song had 2 distinct MVs, both were included as separate entries in our analysis since the MV content differed and each was in the top 40/50 as a unique entry. Furthermore, for 3 songs with the same MV listed in the top chart in different years (ie, were uploaded in different years by 2 different accounts), we retained all these MVs instead of eliminating the duplicates. This decision was driven by our objective to investigate the proportion of top MVs in Vietnam featuring tobacco and alcohol content over the years. This approach also acknowledges that audiences have the option to listen to and watch MVs, even if they are repeated from year to year.

Coding Scheme

Overview

We developed a codebook that contained the following sections: general information, lyrics, MV Tobacco content, and MV Alcohol content. The codebook for this study is shown in Multimedia Appendix 1.

Section 1: General Information

In this section, general information about the MVs was recorded. Each video was assigned a unique video ID number, and the

XSL•FO
RenderX

coding date was recorded. The Vietnamese song's name, release date, and ranking within its chart year were noted. Age restrictions, indicating who could access the MV, were categorized into no restriction, ≥18 years old, and ≥21 years old. The music genres (ie, mutually exclusive options, including pop, hip-hop/rap, R&B, electronic dance music, and other) and type of MV (ie, not mutually exclusive options, narrative-based MV, lyric, performance/dance, animated MV, and others) were noted. Finally, the presence of tobacco or alcohol references in the song name was recorded as yes or no.

Section 2: Lyrics

The song's lyrics were examined to indicate the presence (1) or absence (0) of mentions of tobacco or alcohol content in the lyrics.

Section 3: MV Tobacco Contents

This section coded the depiction of tobacco products in MVs. We coded for specific types of tobacco products shown (cigarettes, e-cigarettes, etc) and their brand/logo, whether tobacco use was shown (eg, lighting a cigarette, person smoking or vaping, etc). Among MVs containing smoking or vaping scenes, the number (ie, 1, 2, 3, or ≥ 4 persons), age (ie, young adult ages 18 to 29 years or adult over the age of 30 years), gender (ie, female, male, or both sexes), and roles of individuals smoking or vaping (ie, singers or others) were categorized. The portrayal of tobacco-related images was assessed, and the cumulative duration of these scenes was recorded (<10 or ≥10 seconds). The presence of warnings about tobacco product harms (yes or no) in MVs was noted.

Section 4: MV Alcohol Contents

Similar to the tobacco section, this section coded the depiction of alcohol contents in MVs. We coded images of alcohol containers (ie, beer bottles, cans, etc) and their brand/logos, images of glasses/cups, and scenes of drinking. The presence of scenes showing drunkenness (ie, a person appears to be inebriated) (yes or no) was recorded. Among MVs having drinking scenes, details about the number (ie, 1, 2, 3, or ≥ 4 persons), age (ie, young adult ages 18-29 or adult over the age of 30), gender (ie, female, male, or both sexes), and roles of individuals drinking (ie, singers or others) were documented. The portrayal of alcohol-related images was described, along with the cumulative duration of these scenes (<10 or ≥10 seconds). Next, the presence of warnings about the legal drinking age was assessed.

MV Coding

The coding team included 4 native Vietnamese speakers with backgrounds in biostatistics or public health. These coders were staff members working at the Hanoi University of Public Health, the University of Medicine and Pharmacy at Ho Chi Minh City, and Pham Ngoc Thach University of Medicine in Vietnam. Before the coding process, a comprehensive 2-hour training session was conducted for the coders, which included 5 parts: (1) introduction to this study (explaining the background and objectives), (2) coding scheme and explaining each variable, (3) working on coding examples from 3 MVs, (4) individual coding practices of 2 MVs, and (5) reconciliation of differences and summary. Each session involved an active discussion. Following the training, the intercoder reliability was established following Neuendorf's recommendations [29]. Specifically, 10% (41/410) of the overall sample was independently coded by 2 coders each, using a rotating system among 4 coders (A, B, C, and D). The reliability test was performed as follows. First, the 41 videos were coded by pairs of coders, with each coder responsible for 20 or 21 MVs (ie, Coder A: videos 1-20; Coder B: videos 11-30; Coder C: videos 21-41; and Coder D: videos 31-41 and 1-10). This design ensured that each video was coded by 2 different coders, creating overlapping segments for reliability assessment. Then, we calculated the Randolph free-marginal kappa for each variable [30]. The intercoder reliability met the high agreement criterion with α ranging from 0.77 to 1. Multimedia Appendix 2 shows the α for each variable. All discrepancies between coders on the first 41 MVs were resolved by the first author (TPTT). After establishing reliability, the remaining MVs were distributed equally among the 4 coders for the independent coding process (ie, Coder A: videos 42-133, Coder B: videos 134-225, Coder C: videos 226-317, and Coder D: videos 318-410).

The entire coding scheme was implemented within a REDCap (Research Electronic Data Capture) data entry form, enhancing systematic and robust data acquisition.

Statistical Analysis

Intercoder reliability among the independent coders was assessed using the Randolph free-marginal kappa [30], calculated using the Online Kappa Calculator [31]. A descriptive analysis was conducted to examine the prevalence of tobacco- and alcohol-related MVs across different years and the frequency of associated features. A descriptive analysis of tobacco and alcohol portrayals was also performed. χ^2 tests were used to assess statistically significant differences in characteristics, including release year, age restriction, and music genres between MVs with and without tobacco (or alcohol) content. The Fisher exact test was used instead of the χ^2 test when more than 20% of cells had expected frequencies <5. Simple logistic regression models were used to identify the factors associated with tobacco and alcohol contents in MVs. The predictor variables included release year, ranking, age restrictions, music genres, and the type of MV. All statistical analyses were performed with STATA software (version 17.0; StataCorp), and values of $P<.05$ were considered statistically significant.

Ethical Considerations

This study was reviewed by the Georgia State University Institutional Review Board (H24050) and was determined to be not human subjects research. No personally identifiable information has been collected or analyzed as part of this research.

Results

Table 1 shows the characteristics of the MVs. Most of the MVs analyzed had no age restriction to access them. In terms of music genres, pop music was the dominant category, accounting for 63.9% (n=262), followed by other types (ie, pop-rap) at 23.4% (n=96). Regarding the type of MVs, a significant proportion were narrative-based (n=281, 68.5%) and performance or

XSL•FO
RenderX

dance-oriented (n=287, 70%). The average duration of these MVs was 312 seconds (ie, 5 minutes and 12 seconds).

Of the 410 MVs examined, 36 (8.8%) contained tobacco-related content and 136 (33.2%) featured alcohol-related content. Additionally, 28 (6.8%) out of 410 MVs included both tobacco and alcohol content. Only 2 (5.6%) of 36 MVs with tobacco content and 3 (2.2%) of 136 with alcohol content had age restrictions for audiences aged 18 or older. Notably, hip-hop or rap songs contained more tobacco- and alcohol-related content compared to other music genres. Among narrative-based MVs, nearly half of them contained alcohol-related content. Only 3 (0.7%) and 2 (0.5%) of the MVs mentioned tobacco and alcohol content in the song titles, respectively.

Table 1. Characteristics of top Vietnamese YouTube music videos (MVs) between 2013 and 2021 (N=410).

Characteristics	Total, n (%)	MVs containing tobacco-related content			MVs containing alcohol-related content		
		Yes, n (%)	No, n (%)	*P* value[a]	Yes, n (%)	No, n (%)	*P* value
Total	410 (100)	36 (8.8)	374 (91.2)	—[b]	136 (33.2)	274 (66.8)	—
Release year				.14			.002[c]
2013	40 (9.8)	3 (7.5)	37 (92.5)		11 (27.5)	29 (72.5)	
2014	40 (9.8)	0 (0)	40 (100)		11 (27.5)	29 (72.5)	
2015	40 (9.8)	3 (7.5)	37 (92.5)		15 (37.5)	25 (62.5)	
2016	40 (9.8)	5 (12.5)	35 (87.5)		7 (17.5)	33 (82.5)	
2017	50 (12.2)	3 (6)	47 (94)		14 (28)	36 (72)	
2018	50 (12.2)	6 (12)	44 (88)		23 (46)	27 (54)	
2019	50 (12.2)	8 (16)	42 (84)		26 (52)	24 (48)	
2020	50 (12.2)	6 (12)	44 (88)		20 (40)	30 (60)	
2021	50 (12.2)	2 (4)	48 (96)		9 (18)	41 (82)	
Age restriction[d]				.04			.11
No restriction	406 (99)	34 (8.4)	372 (91.6)		133 (32.8)	273 (67.2)	
≥18 years old	4 (1)	2 (50)	2 (50)		3 (75)	1 (25)	
≥21 years old	0 (0)	0 (0)	0 (0)		0 (0)	0 (0)	
Music genre[d]				.01			.51
Rock	0 (0)	0 (0)	0 (0)		0 (0)	0 (0)	
Pop	262 (63.9)	18 (6.9)	244 (93.1)		88 (33.6)	174 (66.4)	
Hip-hop/rap	20 (4.9)	6 (30)	14 (70)		7 (35)	13 (65)	
R&B	6 (1.5)	0 (0)	6 (100)		1 (16.7)	5 (83.3)	
Electronic dance music	26 (6.3)	0 (0)	26 (100)		5 (19.2)	21 (80.8)	
Other	96 (23.4)	12 (12.5)	84 (87.5)		35 (36.5)	61 (63.5)	
Type of MV							
Narrative-based MV	281 (68.5)	33 (11.7)	248 (88.3)	.001	131 (46.6)	150 (53.4)	<.001
Lyric	68 (16.6)	2 (2.9)	66 (97.1)	.06	2 (2.9)	66 (97.1)	<.001
Performance/dance	287 (70)	29 (10.1)	258 (89.9)	.18	110 (38.3)	177 (61.7)	.001
Animated	13 (3.2)	0 (0)	13 (100)	.62	1 (7.7)	12 (92.3)	.07
Other (fan-made and background image)	17 (4.1)	0 (0)	17 (100)	.38	1 (5.9)	16 (94.1)	.02
MV length, mean (SD)	312.1 (101.9)	319 (122.7)	311.5 (99.9)		336.1 (117.7)	300.2 (91)	

[a]*P* values were calculated by the Fisher exact test (except for *P*=.002) to assess statistically significant differences in characteristics between MVs with and without tobacco (or alcohol) content.

[b]Not applicable.

[c]χ^2 test was used instead.

[d]Mutually exclusive response.

XSL•FO
RenderX

Table 2 presents the predictors of tobacco and alcohol contents in MVs. Compared to 2013, MVs released in 2019 had an increased likelihood of containing alcohol content. MVs with a higher ranking (ie, a lower number means a higher rank) were more likely to contain alcohol-related content. The hip-hop/rap genre was more likely to contain tobacco content (odds ratio [OR]=5.81, 95% CI=1.99-16.93) compared to pop. Regarding types of MVs, narrative-based MVs had higher odds of having both tobacco content (OR=5.59, 95% CI=1.68-18.58) and alcohol content (OR 21.66, 95% CI=8.59-54.58) compared to nonnarrative-based MVs. Additionally, performance/dance MVs had an increased likelihood of including alcohol content (OR 2.32, 95% CI=1.41-3.80) compared to MVs that were not focused on performance/dance.

Table 2. The predictors of tobacco and alcohol content in 410 Vietnamese YouTube music videos (MVs).

Characteristics	MVs containing tobacco-related content, OR[a] (95% CI)	MVs containing alcohol-related content, OR (95% CI)
Release year (versus 2013)		
2014	N/A[b]	1 (0.37-2.67)
2015	1 (0.19-5.28)	1.58 (0.62-4.07)
2016	1.76 (0.39-7.93)	0.56 (0.19-1.63)
2017	0.79 (0.15-4.13)	1.03 (0.40-2.60)
2018	1.68 (0.39-7.19)	2.25 (0.92-5.47)
2019	2.35 (0.58-9.51)	2.86* (1.17-6.94)
2020	1.68 (0.39-7.19)	1.76 (0.72-4.30)
2021	0.51 (0.08-3.24)	0.58 (0.21-1.57)
Ranking in the top chart	1.01 (0.99-1.04)	0.98* (0.97-1)
Age restriction (≥18 years old versus no restriction)	10.94* (1.49-80.13)	6.16 (0.63-59.76)
Music genres (versus pop)		
Hip-hop/rap	5.81** (1.99-16.93)	1.06 (0.41-2.76)
R&B	N/A	0.40 (0.05-3.44)
Electronic dance music	N/A	0.47 (0.17-1.29)
Other	1.94 (0.90-4.19)	1.13 (0.70-1.85)
Type of MVs		
Narrative-based MV (yes versus no)	5.59** (1.68-18.58)	21.66*** (8.59-54.58)
Lyric (yes versus no)	0.27 (0.06-1.17)	0.05*** (0.01-0.20)
Performance/dance (yes versus no)	1.86 (0.79-4.38)	2.32*** (1.41-3.80)
Animated MV (yes versus no)	N/A	0.16 (0.02-1.26)
Other (yes versus no)	N/A	0.12* (0.02-0.91)

[a]OR: odds ratio.

[b]N/A: not applicable.

*P<.05, ** P<.01, *** P<.001.

Multimedia Appendix 3 illustrates the proportion of tobacco- and alcohol-related content in the 410 MVs analyzed. While 7.6% (n=31) of MVs depicted the actual use of tobacco products, 3.7% (n=15) contained just images of tobacco products (eg, a cigarette pack lying on a table). In terms of alcohol-related content, 24.9% (n=102) of MVs featured images of alcohol containers, such as wine or beer bottles and cans. Additionally, 23.4% (n=96) included images of wine or beer glasses or cups, while 21.2% (n=87) depicted scenes of individuals consuming alcohol. Furthermore, 8% (n=33) of MVs portrayed someone in a state of inebriation.

The prevalence of MVs containing tobacco- and alcohol-related content had a fluctuating pattern over the years, as illustrated in Figure 1. Notably, for alcohol, the proportion of videos with alcohol content increased every year between 2016 and 2019, when more than half of the MVs contained alcohol-related imagery.

XSL•FO
RenderX

Figure 1. The prevalence of music videos (MVs) containing tobacco and alcohol-related content across the years.

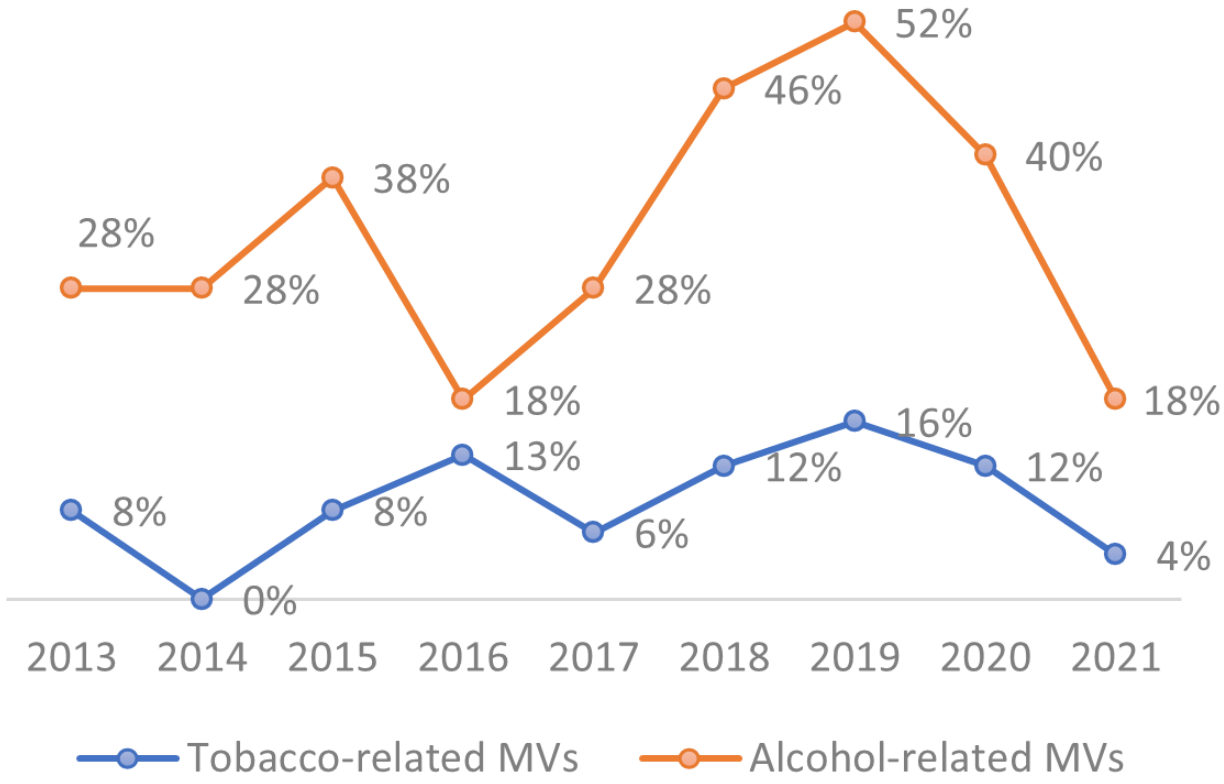

Table 3 describes the portrayal of tobacco in MVs. The following numbers depict the prevalence of MVs containing tobacco-related content (n=36, 8.8%). Cigarettes were the most frequently shown product (n=28, 77.8%), followed by bamboo waterpipe tobacco (n=3, 8.3%). Only 2.8% (n=1) of the MVs displayed tobacco brands, and 5.6% (n=2) included health warnings about tobacco harm. Tobacco imagery was commonly seen at parties, bars, and pubs (n=13, 36.1%), as well as during dancing and singing scenes (n=12, 33.3%). In 19.4% (n=7) of songs, tobacco content was mentioned in the lyrics. Regarding actual use of tobacco products, tobacco use was typically depicted with one person, often a young adult male, and 38.7% (n=12) showed singer(s) smoking. Only 5.6% (n=2) of MVs containing tobacco content had warnings about the harm of tobacco use.

XSL•FO
RenderX

Table 3. Portrayal of tobacco in 36 Vietnamese YouTube music videos (MVs) with tobacco content.

Characteristics	n (%)
Tobacco products	
Cigarette	28 (77.8)
Bamboo waterpipe tobacco	3 (8.3)
Hookah	1 (2.8)
Cigarillos/little cigar	1 (2.8)
Large cigar	3 (8.3)
Other (pipes, roll-your-own cigarettes)	2 (5.6)
Images of tobacco brand/logo[a]	
No	35 (97.2)
Yes	1 (2.8)
Tobacco portrayals	
Shown in sad scenes depicting someone crying, dying, etc	8 (22.2)
Shown when someone was stressed or anxious	7 (19.4)
Shown during dancing or singing scenes	12 (33.3)
Shown at parties, bars, or pubs	13 (36.1)
Duration of tobacco contents[a]	
<10 seconds	19 (52.8)
≥10 seconds	17 (47.2)
Warning or disclaimer about tobacco harmfulness[a]	
No	34 (94.4)
Yes	2 (5.6)
Tobacco mentioned in the lyrics of the song[a]	
No	29 (80.6)
Yes	7 (19.4)
Number of smoker(s)[a,b]	
1	18 (58.1)
2	7 (22.6)
3	2 (6.5)
≥4	4 (12.9)
Age of smoker(s)[b]	
Young adult (18 to 29 years)	24 (77.4)
Adult (30+ years)	10 (32.3)
Person(s) smoking[a]	
Singer(s)	12 (38.7)
Dancer or featured star (not singer)	23 (74.2)
Could not be determined	1 (3.2)
Biological sex of smoker(s)[a,b]	
Female	2 (6.5)
Male	26 (83.9)
Both sexes	3 (9.7)

XSL•FO
RenderX

[a]Mutually exclusive response.

[b]Among MVs containing actual tobacco use scenes (n=31).

Table 4 presents the portrayal of alcohol in the MVs. Images of alcoholic beverage brands or logos appeared in 26 (19.1%) MVs. There was a high proportion of MVs showing alcohol images in parties, bars, or pubs (n=96, 70.6%). Alcohol-related content was mentioned in lyrics in 45 (33.1%) MVs. Only 1.5% (n=2) of MVs included warnings about drinking being restricted to individuals 18 years old and above. In MVs depicting drinking, it frequently involved groups (ie, more than 4 people) (n=45, 51.7%), among young adults of both sexes (n=53, 60.9%), and 56 MVs (64%) showed singers drinking. Only 2 MVs (1.5%) containing alcohol content had warning about drinking being restricted to individuals 18 years old and above.

XSL•FO
RenderX

Table 4. Portrayal of alcohol in 136 Vietnamese YouTube music videos (MVs) with alcohol content.

Characteristics	n (%)
Images of wine/beer brand/logo[a]	
No	110 (80.9)
Yes	26 (19.1)
Alcohol portrayals	
Shown in sad scenes depicting someone crying, dying, etc	23 (16.9)
Shown when someone is stressed or anxious	14 (10.3)
Shown during dancing or singing scenes	32 (23.5)
Shown at parties, bars, or pubs	96 (70.6)
Duration of alcohol content[a]	
<10 seconds	28 (20.6)
≥10 seconds	108 (79.4)
Warning about drinking being restricted to individuals aged 18 years and above[a]	
No	134 (98.5)
Yes	2 (1.5)
Alcohol mentioned in the lyrics of the song[a]	
No	91 (66.9)
Yes	45 (33.1)
Number of drinker(s)[a,b]	
1	23 (26.4)
2	14 (16.1)
3	5 (5.7)
≥4	45 (51.7)
Age of drinker(s)[b]	
Young adult	82 (94.3)
Adult	21 (24.1)
Who the drinker(s) are[b]	
Singer(s)	56 (64.4)
Dancer or featured star (not singer)	75 (86.2)
Biological sex of the drinker(s)[a,b]	
Female	6 (6.9)
Male	28 (32.2)
Both sexes	53 (60.9)

[a]Mutually exclusive response.

[b]Among MVs containing drinking scenes (n=87).

Discussion

Principal Results

Tobacco and alcohol content were prevalent in top Vietnamese YouTube MVs. Our findings indicate that 8.8% (n=36) of MVs from 2013 to 2021 contained tobacco content, and 33.2% (n=136) featured alcohol content. This aligns with a study by Gruber et al [32], where alcohol was present in 34.5% of MVs and tobacco products were present in 10% of MVs broadcast on television channels in the United States in 2001. However, these rates were lower than recent studies in the United Kingdom [33], which showed alcohol and tobacco imagery in 45% and 22% of YouTube MVs, respectively, and in the United States, where nearly one-fourth of popular MVs from 2018 to 2020 contained tobacco imagery [34]. These differences across countries could be attributed to cultural attitudes toward tobacco

XSL•FO
RenderX

use, regulatory frameworks, and social norms. The higher prevalence of alcohol content compared to tobacco content in the MVs could be attributed to widespread direct endorsement by the artists who have financial associations with alcoholic products [33]. Additionally, alcohol is often portrayed as a social lifestyle and is more widely accepted in social settings such as parties, celebrations, and nightlife, which are common themes in MVs.

We observed an increase in the proportion of MVs containing tobacco and alcohol from 2013 to 2019, followed by a decrease in 2020 and 2021. Several factors might contribute to this increase, including evolving cultural norms, shifts in marketing strategies by the tobacco and alcohol industries, and changes in popular music trends. Conversely, the subsequent decline could be attributed to the impact of the COVID-19 pandemic, as suggested in a previous study [34]. For instance, the parody song "Ghen Cô Vy," which emphasized handwashing and personal hygiene to prevent COVID-19, went viral and topped the charts in 2020. Another factor could be the variation in the number of narrative-based MVs, being highest in 2019 and lowest in 2020 and 2021. Further monitoring is necessary to validate these trends.

Comparison With Prior Works

Our findings revealed that the hip-hop genre was associated with a higher rate of tobacco content in Vietnamese YouTube MVs. This was consistent with previous studies indicating that nearly half of tobacco-related incidences appeared in hip-hop songs among the Billboard top tracks in the United States from 2013 to 2017 [35]. Hip-hop artists often serve as role models for their fans owing to their prominence and the social, cultural, and political commentary in their music, which includes tobacco use [35]. While alcohol is likely to appear across various music genres, it is more common in hip-hop and pop MVs [22,32,33], consistent with our study.

The prevalence of smoking and drinking in society was reflected in the portrayals of tobacco and alcohol in MVs. Cigarettes were the most commonly depicted tobacco product in MVs, followed by bamboo waterpipe tobacco—a traditional tobacco product—representing the most commonly used types of tobacco in Vietnam (36.7% of men used cigarettes, and 13.7% used bamboo waterpipes) [25]. We did not find any MVs containing new tobacco products such as e-cigarettes and heated tobacco products, which are not popular in Vietnam, especially among adults, with a prevalence of less than 1% [25]. There is a notable disparity in smoking prevalence in society between men (45.3%) and women (1.1%) [25], as reflected in the MVs, where scenes of male smokers were very common, accounting for 83.9% (n=26) of all the MVs compared to female smokers at 6.5% (n=2). In terms of drinking behavior, there was also a disparity in alcohol consumption between men and women, with approximately 60% of men reporting past-week alcohol consumption compared to only 5% of women [36]. However, the open social norms surrounding women's participation in the modern "drinking culture" [37] could contribute to a high proportion of MVs depicting both male and female characters drinking (n=53, 60.9%). Additionally, smoking appeared to be solitary behavior, with more than half of the MVs depicting a scene with only 1 person smoking. In contrast, drinking was portrayed as a social behavior in Vietnam, with 51.7% (n=45) of the MVs showing a group of people (ie, 4 or more persons) drinking together.

One-third of MVs with tobacco content showcased singers engaging in smoking, while two-thirds of videos with alcohol content featured singers consuming alcohol. These depictions likely contribute to the formation of positive attitudes toward tobacco and alcohol use. It is noteworthy that the primary audience for these artists comprises adolescents and young adults, as they make up the predominant demographic using social media platforms like YouTube to follow music and swiftly adopt new social trends. Moreover, considering that a substantial proportion of individuals depicted as smokers or drinkers in MVs were young adults (n=24, 77.4% for tobacco and n=82, 94.3% for alcohol) in our study, the societal impact is considerable. Given that many of these singers boast millions of fans, including children, the potential influence on society is profound, effectively turning them into spokespeople for the industry [17]. Adolescents and young adults frequently look up to their idols, often seeking to emulate various aspects of their behavior. Consequently, the portrayal of tobacco and alcohol in MVs poses a significant public health risk to these age groups, potentially diminishing their awareness of the risks associated with tobacco and alcohol use, heightening their susceptibility to use tobacco, and even leading to initiation [4,22,38].

Implications for Regulations

Vietnam's regulations on tobacco and alcohol content in online platforms reveal a notable gap. Circular No 25/2018/TT-BVHTTDL, issued by the Ministry of Culture Sports and Tourism, restricts the depiction of actors using tobacco in theater and cinematic works. Additionally, Decree No 24/2020/ND-CP outlines regulations related to certain provisions of the law on preventing and combating the harms of alcoholic beverages. These regulations require approval from the authorized agency based on recommendations of the Evaluation Council for scenes depicting tobacco use or alcohol consumption in artistic works. The use of such images is permissible only when necessary for portraying historical characters, recreating a specific historical period, or critiquing and condemning smoking or drinking. Films or movies must be classified according to appropriate age groups and include health warnings regarding the harmful effects of tobacco and alcohol, such as "drinking alcohol may cause traffic accidents" and "persons under 18 years of age are forbidden to drink alcohol" either in text or images. However, it is unclear whether these regulations would apply to MVs, which are digital videos and are not released in the form of a tape or film. Our study found that most of the tobacco and alcohol content depicted in top Vietnamese MVs was not accompanied by any health warnings.

Managing activities on the internet, including MVs on platforms like YouTube, is challenging, as Vietnam is grappling with issues related to advertising and sponsorship on the internet. The widespread use of sophisticated online techniques, including social media, for promoting tobacco and alcohol products has led to challenges in regulation [16,39]. Artists, whether

XSL•FO
RenderX

intentionally or inadvertently, become effective and powerful marketing tools for these products by incorporating images of tobacco and alcohol in their MVs on social media platforms like YouTube [6,33,34]. Moreover, in Vietnam, certain exceptions exist for drinking, notably in the case of highlanders and specific ethnic groups. For instance, the traditional straw liquor of highlanders and other ethnic communities is regarded as a cultural expression cultivated over generations. Consequently, these images may provoke controversial opinions and pose challenges in adherence to regulations.

Although no laws specifically impose age restrictions for MVs, the Tobacco Control Law No 09/2012/QH13, issued by the National Assembly in 2012, prohibits the depiction of children under 18 years old smoking. Regulations related to alcohol use among children under 18 years were issued more recently. This might explain our findings, where the presence of tobacco content was significantly associated with MVs having an age restriction, while this was not seen for alcohol content. However, it is important to emphasize that the number of MVs containing tobacco and alcohol content with age restrictions remained very low.

YouTube does not allow content showing "harmful or dangerous acts involving minors," such as minors drinking alcohol or using tobacco [40]. While YouTube's Community Guidelines are universal policies that apply to users worldwide, it also has procedures to respond appropriately to valid legal requests that are based on relevant local laws [41]. For example, YouTube permits branding and informational advertising for alcoholic beverages in 68 countries, with the exception of 4. Vietnam only allows beverages with less than 5.5% alcohol by volume, Ecuador with less than 5%, and India and Indonesia with 0% alcohol [42]. Tobacco advertising on YouTube is prohibited. In addition, videos with tobacco content fall under the "limited or no ads" monetization state, meaning that content creators will not be able to receive revenues for advertisements placed in their videos. However, this information is not shown to the viewers and might not be applicable to MVs, as "artistic content such as MVs may contain elements such as inappropriate language, references to soft drug use, or nonexplicit sexual themes, and still be suitable for advertising" [43]. Content that depicts tobacco might be age-gated or restricted on YouTube. However, as our results showed, only 5.6% (n=2) of MVs with tobacco content had age restrictions. Thus, young people are not sufficiently protected from alcohol- and tobacco-related content in MVs on YouTube. Considering the billions of views that alcohol and tobacco content receive on YouTube and their potentially harmful effects, there is a pressing need to regulate such content. Given the other restrictions YouTube has put in place [44], it should expand these rules to include any portrayals of tobacco (and possibly alcohol) regardless of the type of video—at least for monetization purposes if an outright ban is not feasible. Governments should consider implementing the Framework Convention on Tobacco Control requirements on the internet to reduce protobacco content. Our study also revealed novel findings, showing that mentions of tobacco or alcohol in song lyrics were prevalent. This emphasizes the importance of regulating not only visual imagery but also lyrical content.

Limitations

This study has several limitations. First, some songs that are popular among adolescents and young adults were not included in this study because they did not make it onto the top charts. Second, although this study included only top MVs on YouTube, MVs can also be posted and viewed via other popular websites such as Spotify, Zing MP3, and social media platforms like Facebook. Third, this study was unable to identify the characteristics of individuals who viewed each video and their exposure to tobacco and alcohol content in the MVs. Therefore, further studies are needed to investigate how exposure to MVs containing tobacco and alcohol content impacts perceptions and behaviors related to smoking and drinking. Finally, the small sample size of tobacco-related MVs likely did not provide sufficient power to detect significant effects of the release year on tobacco content.

Conclusions

In conclusion, our analysis of top Vietnamese MVs from 2013 to 2021 revealed a concerning prevalence of tobacco and alcohol content. Approximately one-third of the MVs contained alcohol content, while nearly 1 in 10 featured tobacco content. Additionally, most of the tobacco and alcohol content was depicted in MVs without any warnings or health disclaimers. This study highlighted a regulatory gap in addressing such content online, emphasizing the need for more stringent regulations and age restrictions on platforms like YouTube. Furthermore, there is a critical need for additional research to understand how exposure to MVs influences the perceptions and behaviors of adolescents and young adults regarding smoking and drinking.

Acknowledgments

We would like to extend our gratitude to Thi My Linh Nguyen (Pham Ngoc Thach University of Medicine), Thi Thanh Nguyen (Ho Chi Minh City Medicine and Pharmacy University), and Ba Linh Tieu (Medical Genetics Institute) for their assistance in conducting MV coding in this study.

Funding provided by the Bloomberg Initiative to Reduce Tobacco Use through the CDC Foundation with a grant from Bloomberg Philanthropies.

XSL•FO
RenderX

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention Foundation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Updated codebook.
[DOCX File , 737 KB - jmir_v26i1e55555_app1.docx]

Multimedia Appendix 2
The α for each variable.
[XLSX File (Microsoft Excel File), 28 KB - jmir_v26i1e55555_app2.xlsx]

Multimedia Appendix 3
The proportion of tobacco- and alcohol-related content in the music videos (N=410).
[PNG File , 28 KB - jmir_v26i1e55555_app3.png]

References

1. Tactic: advertising and promotions. Alcohol Justice. URL: https://alcoholjustice.org/site/big-alcohol/industry-tactics/24-tactic-advertising-and-promotions.html [accessed 2021-02-08]
2. Tobacco industry marketing. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/tobacco_industry/marketing/index.htm [accessed 2021-02-09]
3. WHO Representative Office China. World Health Organization. URL: http://www.wpro.who.int/china/topics/tobacco/infographics/en/ [accessed 2021-02-09]
4. Smith LA, Foxcroft DR. The effect of alcohol advertising, marketing and portrayal on drinking behaviour in young people: systematic review of prospective cohort studies. BMC Public Health 2009 Feb 06;9(1):51 [FREE Full text] [doi: 10.1186/1471-2458-9-51] [Medline: 19200352]
5. Arora M, Gupta VK, Nazar GP, Stigler MH, Perry CL, Reddy KS. Impact of tobacco advertisements on tobacco use among urban adolescents in India: results from a longitudinal study. Tob Control 2012 May 29;21(3):318-324 [FREE Full text] [doi: 10.1136/tc.2010.040733] [Medline: 21803927]
6. Braverman MT, Aarø LE. Adolescent smoking and exposure to tobacco marketing under a tobacco advertising ban: findings from 2 Norwegian national samples. Am J Public Health 2004 Jul;94(7):1230-1238. [doi: 10.2105/ajph.94.7.1230] [Medline: 15226148]
7. Van den Bulck J, Beullens K, Mulder J. Television and music video exposure and adolescent 'alcopop' use. Int J Adolesc Med Health 2006;18(1):107-114. [doi: 10.1515/ijamh.2006.18.1.107] [Medline: 16639864]
8. Snyder LB, Milici FF, Slater M, Sun H, Strizhakova Y. Effects of alcohol advertising exposure on drinking among youth. Arch Pediatr Adolesc Med 2006 Jan 01;160(1):18-24. [doi: 10.1001/archpedi.160.1.18] [Medline: 16389206]
9. Emery S, Kim Y, Choi YK, Szczypka G, Wakefield M, Chaloupka FJ. The effects of smoking-related television advertising on smoking and intentions to quit among adults in the United States: 1999–2007. Am J Public Health 2012 Apr;102(4):751-757. [doi: 10.2105/ajph.2011.300443]
10. Lynch BS, Bonnie RJ, editors. Growing up Tobacco Free: Preventing Nicotine Addiction in Children and Youths. Washington, DC: National Academies Press; 1994.
11. The impact of tobacco advertising, promotion, and sponsorship. Campaign for Tobacco-Free Kids. URL: https://www.tobaccofreekids.org/assets/global/pdfs/en/APS_healthimpact.pdf [accessed 2021-02-09]
12. The seven key messages of the alcohol industry. European Centre for Monitoting Alcohol Marketing. 2021 Feb 09. URL: https://eucam.info/eucam-reports/the-7-key-messages-of-the-alcohol-industry/ [accessed 2021-02-09]
13. Key elements of tobacco control legislation. World Health Organization. URL: https://www.who.int/tobacco/research/legislation/key_elements/en/ [accessed 2021-02-08]
14. Tackling NCDs: 'best buys' and other recommended interventions for the prevention and control of noncommunicable diseases. World Health Organization. URL: https://iris.who.int/handle/10665/259232 [accessed 2021-02-08]

XSL•FO
RenderX

15. Room R, O'Brien P. Alcohol marketing and social media: a challenge for public health control. European Centre for Monitoting Alcohol Marketing. URL: https://eucam.info/2020/11/03/alcohol-marketing-and-social-media-a-challenge-for-public-health-control-2/ [accessed 2021-02-09]
16. O'Brien EK, Hoffman L, Navarro MA, Ganz O. Social media use by leading US e-cigarette, cigarette, smokeless tobacco, cigar and hookah brands. Tob Control 2020 Dec 26;29(e1):e87-e97. [doi: 10.1136/tobaccocontrol-2019-055406] [Medline: 32217772]
17. Industry influencer: how tobacco content is infiltrating social media. Truth Initiative. URL: https://truthinitiative.org/research-resources/tobacco-pop-culture/industry-influencer-how-tobacco-content-infiltrating-social [accessed 2024-08-15]
18. Barry A, Bates AM, Olusanya O, Vinal CE, Martin E, Peoples JE, et al. Alcohol marketing on Twitter and Instagram: evidence of directly advertising to youth/adolescents. Alcohol Alcohol 2016 Jul;51(4):487-492 [FREE Full text] [doi: 10.1093/alcalc/agv128] [Medline: 26597794]
19. Freeman B, Chapman S. Is "YouTube" telling or selling you something? Tobacco content on the YouTube video-sharing website. Tob Control 2007 Jun 12;16(3):207-210 [FREE Full text] [doi: 10.1136/tc.2007.020024] [Medline: 17565142]
20. Cranwell J, Opazo-Breton M, Britton J. Adult and adolescent exposure to tobacco and alcohol content in contemporary YouTube music videos in Great Britain: a population estimate. J Epidemiol Community Health 2016 May 14;70(5):488-492 [FREE Full text] [doi: 10.1136/jech-2015-206402] [Medline: 26767404]
21. Cranwell J, Britton J, Bains M. "F*ck it! Let's get to drinking-poison our livers!": a thematic analysis of alcohol content in contemporary YouTube music videos. Int J Behav Med 2017 Feb 6;24(1):66-76 [FREE Full text] [doi: 10.1007/s12529-016-9578-3] [Medline: 27601057]
22. Sloane K, Wilson N, Imlach Gunasekara F. A content analysis of the portrayal of alcohol in televised music videos in New Zealand: changes over time. Drug Alcohol Rev 2013 Jan 10;32(1):47-52. [doi: 10.1111/j.1465-3362.2012.00477.x] [Medline: 22679924]
23. Elkin L, Thomson G, Wilson N. Connecting world youth with tobacco brands: YouTube and the internet policy vacuum on Web 2.0. Tob Control 2010 Oct 25;19(5):361-366. [doi: 10.1136/tc.2010.035949] [Medline: 20739706]
24. Digital 2020: Vietnam. Datareportal. URL: https://datareportal.com/reports/digital-2020-vietnam [accessed 2021-02-17]
25. Global adult tobacco survey 2015. World Health Organization. URL: https://extranet.who.int/ncdsmicrodata/index.php/catalog/875 [accessed 2023-10-26]
26. Manthey J, Shield KD, Rylett M, Hasan OSM, Probst C, Rehm J. Global alcohol exposure between 1990 and 2017 and forecasts until 2030: a modelling study. Lancet 2019 Jun;393(10190):2493-2502. [doi: 10.1016/s0140-6736(18)32744-2]
27. Neuendorf KA. The Content Analysis Guidebook. Newbury Park, California: Sage Publications Inc; 2017.
28. World Music Charts. URL: https://www.youtube.com/@WorldMusicCharts [accessed 2024-10-28]
29. Neuendorf KA. Content analysis—a methodological primer for gender research. Sex Roles 2010 Nov 20;64(3-4):276-289. [doi: 10.1007/s11199-010-9893-0]
30. Randolph J. Free-marginal multirater kappa (multirater κfree): an alternative to Fleiss fixed-marginal multirater kappa. 2010 Presented at: Joensuu University Learning and Instruction Symposium 2005; October 14-15; Joensuu, Finland.
31. Randolph J. Online Kappa Calculator. 2008. URL: http://justusrandolph.net/kappa/ [accessed 2023-08-08]
32. Gruber EL, Thau HM, Hill DL, Fisher DA, Grube JW. Alcohol, tobacco and illicit substances in music videos: a content analysis of prevalence and genre. J Adolesc Health 2005 Jul;37(1):81-83. [doi: 10.1016/j.jadohealth.2004.02.034] [Medline: 15963915]
33. Cranwell J, Murray R, Lewis S, Leonardi-Bee J, Dockrell M, Britton J. Adolescents' exposure to tobacco and alcohol content in YouTube music videos. Addiction 2015 Apr 28;110(4):703-711 [FREE Full text] [doi: 10.1111/add.12835] [Medline: 25516167]
34. Rath JM, Dimaya B, O'Connor KM, Kreslake JM, Vallone DM, Hair EC. Content analysis of tobacco imagery in popular music Videos, 2018-2021. Prev Med Rep 2023 Jun;33:102188 [FREE Full text] [doi: 10.1016/j.pmedr.2023.102188] [Medline: 37223573]
35. Knutzen KE, Moran MB, Soneji S. Combustible and electronic tobacco and marijuana products in hip-hop music videos, 2013-2017. JAMA Intern Med 2018 Dec 01;178(12):1608-1615 [FREE Full text] [doi: 10.1001/jamainternmed.2018.4488] [Medline: 30326017]
36. Van Bui T, Blizzard CL, Luong KN, Van Truong NL, Tran BQ, Otahal P, et al. Alcohol consumption in Vietnam, and the use of 'standard drinks' to measure alcohol intake. Alcohol Alcohol 2016 Mar 08;51(2):186-195. [doi: 10.1093/alcalc/agv082] [Medline: 26884509]
37. Lincoln M. Alcohol and drinking cultures in Vietnam: A review. Drug Alcohol Depend 2016 Feb 01;159:1-8 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.10.030] [Medline: 26802499]
38. Russell CA, Régnier-Denois V, Chapoton B, Buhrau D. Impact of substance messages in music videos on youth: beware the influence of connectedness and its potential prevention-shielding effect. J Stud Alcohol Drugs 2017 Sep;78(5):674-683. [doi: 10.15288/jsad.2017.78.674] [Medline: 28930054]
39. Lack of regulation leaves door open to harmful digital marketing of alcohol. European Centre for Monitoting Alcohol Marketing. URL: https://eucam.info/2020/07/07/lack-of-regulation-leaves-door-open-to-harmful-digital-marketing-of-alcohol/ [accessed 2021-02-08]

XSL•FO
RenderX

40. Child safety policy. YouTube. URL: https://support.google.com/youtube/answer/2801999?sjid=6627684101682304310-EU [accessed 2024-08-13]
41. Rules and policies: legal removals. YouTube. URL: https://www.youtube.com/intl/ALL_uk/howyoutubeworks/policies/legal-removals/ [accessed 2024-08-18]
42. List of ad policies - alcohol. Google. URL: https://support.google.com/adspolicy/answer/6012382?hl=en#zippy=%2Ccountries-that-allow-such-ads%2Ctroubleshooter-alcohol-targeting [accessed 2024-08-15]
43. Advertiser-friendly content guidelines. YouTube. URL: https://support.google.com/youtube/answer/6162278?sjid=16551834878316171038-EU [accessed 2024-08-15]
44. Kong G, Laestadius L, Vassey J, Majmundar A, Stroup AM, Meissner HI, et al. Tobacco promotion restriction policies on social media. Tob Control 2024 Apr 19;33(3):398-403 [FREE Full text] [doi: 10.1136/tc-2022-057348] [Medline: 36328589]

Abbreviations

APS: advertising, promotion, and sponsorship
MV: music video
OR: odds ratio
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 15.12.23; peer-reviewed by J Vassey, T Tr n Bình; comments to author 07.08.24; revised version received 20.08.24; accepted 02.10.24; published 08.11.24.

Please cite as:
Tran TPT, Vu TT, Li Y, Popova L
Tobacco and Alcohol Content in Top Vietnamese YouTube Music Videos: Content Analysis
J Med Internet Res 2024;26:e55555
URL: https://www.jmir.org/2024/1/e55555
doi:10.2196/55555
PMID:

©Thi Phuong Thao Tran, Thu Trang Vu, Yachao Li, Lucy Popova. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predicting Prefecture-Level Well-Being Indicators in Japan Using Search Volumes in Internet Search Engines: Infodemiology Study

Myung Si Yang[1*], PHN, RN, MS; Kazuya Taira[2*], PHN, RN, PhD

[1]Course of Advanced Nursing Sciences, Human Health Sciences, Faculty of Medicine, Kyoto University, Kyoto, Japan

[2]Human Health Sciences, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[*]all authors contributed equally

Corresponding Author:
Myung Si Yang, PHN, RN, MS
Course of Advanced Nursing Sciences, Human Health Sciences, Faculty of Medicine, Kyoto University
53, Shogoin kawara-cho, Sakyo-ku
Kyoto, 606–8507
Japan
Phone: 81 0757513927
Email: yang.myungsi.k63@kyoto-u.jp

Abstract

Background: In recent years, the adoption of well-being indicators by national governments and international organizations has emerged as an important tool for evaluating state governance and societal progress. Traditionally, well-being has been gauged primarily through economic metrics such as gross domestic product, which fall short of capturing multifaceted well-being, including socioeconomic inequalities, life satisfaction, and health status. Current well-being indicators, including both subjective and objective measures, offer a broader evaluation but face challenges such as high survey costs and difficulties in evaluating at regional levels within countries. The emergence of web log data as an alternative source of well-being indicators offers the potential for more cost-effective, timely, and less biased assessments.

Objective: This study aimed to develop a model using internet search data to predict well-being indicators at the regional level in Japan, providing policy makers with a more accessible and cost-effective tool for assessing public well-being and making informed decisions.

Methods: This study used the Regional Well-Being Index (RWI) for Japan, which evaluates prefectural well-being across 47 prefectures for the years 2010, 2013, 2016, and 2019, as the outcome variable. The RWI includes a comprehensive approach integrating both subjective and objective indicators across 11 domains, including income, job, and life satisfaction. Predictor variables included *z* score–normalized relative search volume (RSV) data from Google Trends for words relevant to each domain. Unrelated words were excluded from the analysis to ensure relevance. The Elastic Net methodology was applied to predict RWI using RSVs, with α balancing ridge and lasso effects and λ regulating their strengths. The model was optimized by cross-validation, determining the best mix and strength of regularization parameters to minimize prediction error. Root mean square errors (RMSE) and coefficients of determination (R^2) were used to assess the model's predictive accuracy and fit.

Results: An analysis of Google Trends data yielded 275 words related to the RWI domains, and RSVs were collected for 211 words after filtering out irrelevant terms. The mean search frequencies for these words during 2010, 2013, 2016, and 2019 ranged from −1.587 to 3.902, with SDs between 3.025 and 0.053. The best Elastic Net model (α=0.1, λ=0.906, RMSE=1.290, and R^2=0.904) was built using 2010-2016 training data and 2-13 variables per domain. Applied to 2019 test data, it yielded an RMSE of 2.328 and R^2 of 0.665.

Conclusions: This study demonstrates the effectiveness of using internet search log data through the Elastic Net machine learning method to predict the RWI in Japanese prefectures with high accuracy, offering a rapid and cost-efficient alternative to traditional survey approaches. This study highlights the potential of this methodology to provide foundational data for evidence-based policy making aimed at enhancing well-being.

(J Med Internet Res 2024;26:e64555) doi:10.2196/64555

KEYWORDS

well-being; spatial indicator; infodemiology; search engine; public health; health policy; policy-making; Google; Japan

XSL•FO
RenderX

Introduction

Conceptualizing Well-Being

Well-being is a complex and multifaceted concept for which there is "no consensus around a single definition of well-being", as acknowledged by the US Centers for Disease Control and Prevention, but generally, well-being refers to "judging life positively," "feeling good," and the experience of good physical health [1]. It also includes dimensions such as physical health, emotional health, economic circumstances, life satisfaction, and engaging activities and works [2]. In the early 21st century, numerous scholars suggested that well-being should be quantified to create indicators that could guide national public policies. Consequently, many scales have been proposed [2-5]. However, the absence of a universally accepted definition of well-being has made it challenging to establish a single, comprehensive measure [6]. Current measures often fall into categories such as "life evaluation," "hedonic well-being," and "eudaimonic well-being," each capturing different aspects of the well-being spectrum [7]. Yet, because these dimensions are deeply interconnected, scales focused on specific aspects may not fully capture the overall well-being of individuals.

The Need for a Comprehensive Well-Being Indicator

Historically, well-being has been measured primarily by economic indicators such as gross domestic product (GDP). However, a prior study has suggested that economic indicators are "no longer a complete approximation of how well a nation is doing" [4]. GDP is only one aspect of well-being and does not capture socioeconomic inequalities, life satisfaction, or health status [8]. For the nation as a whole, previous research suggests that as the GDP or economic activity increases, the standard of living of the population improves [9]; however, in a well-developed society, there may be some citizens who sacrifice their well-being for economic efficiency. In fact, economic growth can lead to improved living conditions in some areas, but an examination of the United States revealed that despite a 3-fold GDP increase in the past 50 years, life satisfaction levels have not risen [3], accompanied by rising depression and anxiety rates [10]. Therefore, GDP growth is not equal to improved well-being, and more comprehensive well-being indicators that include multiple factors such as health, work, and social connections need to be used in policy evaluation. However, it is not proposed to completely replace traditional economic indicators, but to add new indicators of well-being and look at the relationship between both the economy and well-being [11].

Well-Being Indicators Currently Used for Policy Evaluation and Their Challenges

Several countries and international organizations use a variety of different well-being indicators, which are a mix of subjective and objective indicators. Organizations such as the Organization for Economic Co-operation and Development (OECD) and United Nations Development Program, as well as countries including New Zealand, the United Kingdom, France, and Italy, use objective well-being metrics encompassing health (eg, lifespan), job opportunities (eg, employment rate), environmental conditions (eg, greenhouse gas emissions), safety (eg, crime rate), and governance (eg, voter turnout) [12-18]. Subjective indicators involve self-evaluations of individuals of their lives, as exemplified by tools such as the Satisfaction With Life Scale [19] and the Subjective Happiness Scale [20]. In practical policy implementation, dashboards that visualize data for both objective and subjective indicators have been devised to holistically assess social well-being. In countries like the United Kingdom and New Zealand, unique well-being indices have been developed to gauge citizen welfare and guide policy and fiscal decisions [15-18].

In addition, some international organizations are involved in the development of well-being indicators and are publishing reports on international comparisons. Typical examples include the Better Life Index (BLI) of the OECD [12], the World Happiness Report of the United Nations [13], and the Human Development Index of the United Nations Development Program [14]. Moreover, while many well-being indicators target national units, there are initiatives to establish metrics for more localized regional assessments within countries [21], illustrating diverse methodologies to evaluate comprehensive well-being.

These preexisting well-being indicators have some limitations and challenges. Conducting large-scale epidemiological surveys, which are necessary for both objective and subjective indicators, incurs significant human, temporal, and financial costs. In several resource-limited nations, conducting surveys can be challenging, or individuals experiencing severe well-being deficiencies are often more prone to nonresponse, thereby hindering the assessment of their well-being. In addition, survey items rely on multiple government surveys, and survey years often vary and can only be evaluated once every few years. It is also difficult to standardize the items surveyed across all countries, making international comparisons difficult. Therefore, a reasonable indicator that is comparable to other regions can be obtained, and evaluated at the right time for policy evaluation is very useful for policymakers.

Web Log Data as a Policy Indicator

With the spread of the internet, methods for using web log data to predict statistics for policy evaluation have been reported. Statistics related to well-being indicators have also been associated with web log data, including health [22], job availability [23], environmental quality [24], safety [25], governance [26], and subjective elements, such as emotional well-being [27] and life satisfaction [28]. The web log data used in the previous study varied and included search volume logs from internet search engines, such as Google and Yahoo! Search, as well as log data from social networking services (SNSs), including X (Twitter), Facebook, and Instagram. This study used search volume log data from the search engine Google, which can be collected from Google Trends [29]. Google is one of the major search engines used in 195 countries around the world, making it easy to ensure reproducibility in other countries. Another advantage of social networking tools is that there is less bias in demographic information such as user age, gender, and race. While information posted on SNSs is often directed toward society and others and may contain only overly idealistic information, search behavior on search engines is an individual's internal process and might be able to reduce the

XSL•FO
RenderX

confirmation bias that amplifies the information that people find favorable.

Objective

This study aimed to develop a model that predicts comprehensive well-being indicators via search volume log data from internet search engines. This approach seeks to bypass the need for large-scale statistical surveys, thereby reducing budgetary and human resource requirements. In other words, it enables policymakers to assess the well-being of the public at low cost and at the right time, thereby facilitating more effective policy decisions.

Methods

Regional Well-Being Index as an Outcome

The Regional Well-Being Index (RWI) for Japan [30], structured based on the BLI methodology of the OECD, was used as the outcome variable. The RWI, like the BLI, consisted of 11 domains: “Income,” “Jobs,” “Housing,” “Health,” “Work-Life balance,” “Education,” “Community,” “Civic Engagement,” “Environment,” “Safety,” and “Life Satisfaction.” This index focuses on integrating both subjective and objective indicators to comprehensively evaluate well-being.

The “Regional Well-Being” of the OECD [21] provides detailed scores for well-being at subnational regions at smaller geographical scales than the national level. However, the administrative divisions that make policymaking do not coincide with these subnational regions in several countries. For instance, the OECD’s regional well-being for Japan is presented at a relatively macroscopic level, segmenting the nation into 10 regions, including Tohoku and Kansai, each encompassing several prefectures. This level of aggregation differs from the levels at which policy decisions are operationalized, typically at the prefectural and municipal levels. To use well-being indicators more efficiently, it is important to calculate them for each administrative level where policy evaluation and decision-making are conducted. Therefore, this study adopted the RWI, a comprehensive well-being indicator at the prefectural level in Japan, based on the BLI methodology, as its outcome.

Yang and Taira [30] provided domain-specific scores and an integrated RWI (IRWI) that aggregated all domains by prefecture. These RWIs assessed the well-being of all 47 Japanese prefectures for 2010, 2013, 2016, and 2019, affirming their reliability and validity relative to the BLI and the existing well-being index [30]. Due to the data unavailability of certain indicators constituting the RWI, the 2019 data remain the most recent. Subsequent updates have been delayed due to factors such as the COVID-19 pandemic and prolonged data aggregation processes.

Internet Search Log Data as Predictor Variables

The relative search volume (RSV) in the internet search engine Google was used as the predictor variable and was obtained from the Google Trends website. Google Trends enables the tracking of temporal variations in the popularity of specific search words on Google and YouTube while capturing regional search dynamics and related words. Google Trends was used to collect related words in addition to the RSVs for the main words. The “Related keywords” feature of Google Trends offers other words that are frequently searched in relation to a selected word. “Related keywords” include “Top searches” and “Rising searches,” where “Top searches” are the words most frequently searched within the same session as the selected word, and “Rising searches” are the words whose search frequency has increased the most significantly over a specific period. Google Trends collects up to 25 related words.

In Google Trends, RSVs are normalized on a 0-100 scale, where the word with the highest search volume in a specific period scores 100, and the frequencies of other words are adjusted accordingly. Moreover, Google Trends facilitates the comparison of search frequencies for words across different states or prefectures within a country, normalizing data to the area with the highest frequency of a word score of 100, whereas others are rated in comparison. For instance, a search frequency score of 50 for a word in one region suggests that its search volume is half of that of the highest-ranked region.

First, in the procedure for obtaining RSVs, following the methodology established in prior research that developed the RWI framework, representative words for each domain were selected based on their relevance to the specific domain, ensuring alignment with their Japanese translations (Figure 1). After selecting the main word for each domain, related search words were collected using Google Trends, which can gather up to 25 related words for any given word. The collection period spanned between January 1, 2010, and December 31, 2019, aligned with the RWI measurement years, and RSVs of related search words during this period were collected. In addition, we collected RSVs for web searches within Japan, without limiting them to any specific category but including all categories. Data collection was conducted without using quotation marks around the search words. Since searching multiple words simultaneously would result in normalization with the word having the highest search volume set to 100, we collected RSVs for each term individually.

XSL•FO
RenderX

Figure 1. Creation method for the word set. "Representative words" provide the English translations of the original Japanese words in parentheses.

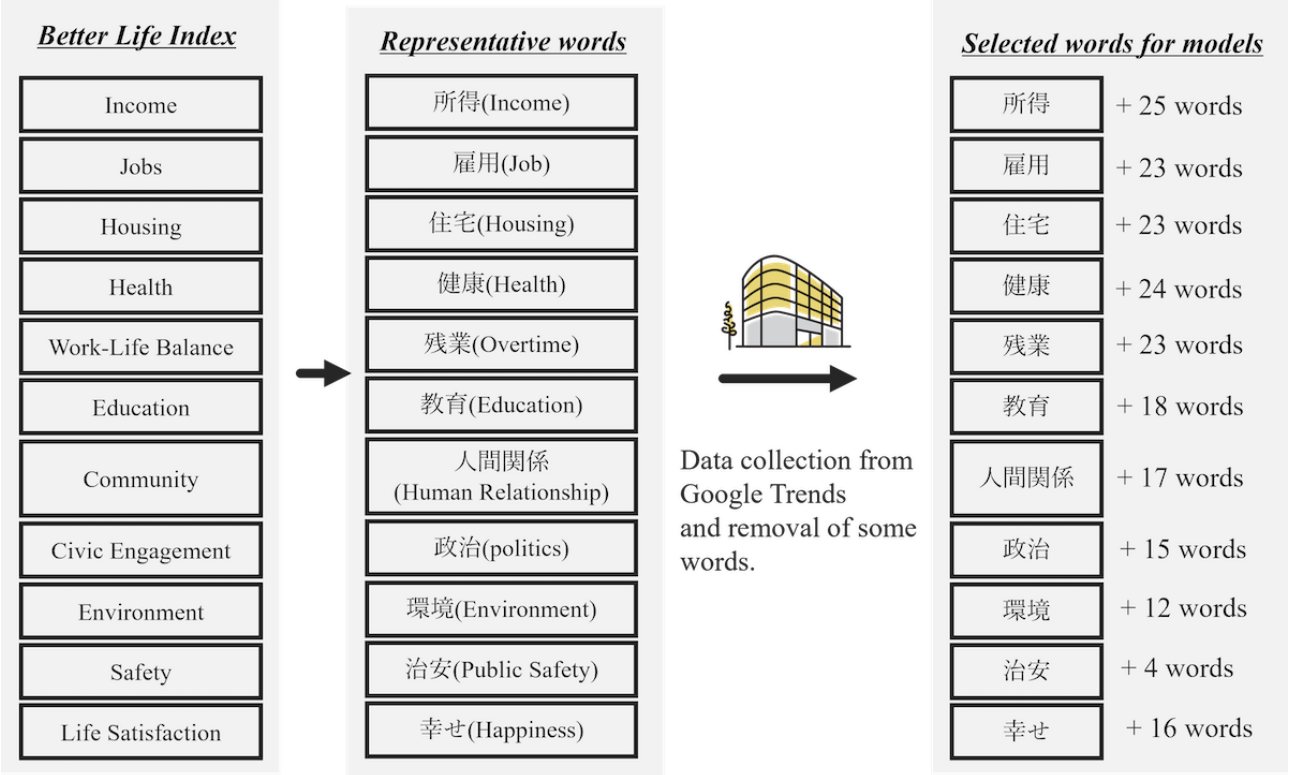

To extract the words for the predictive model, those that included specific regions, companies, or personal names were excluded. Additionally, 2 authors, YM and TK, double-checked and deleted words irrelevant to each domain. For example, words such as "福祉 住 環境" (welfare living environment) relating to the environment of an individual's residence were judged inappropriate and removed, since the original "environment" domain contains the words referring to the "natural environment." In the Community domain, words like "人間関係" (human relations) and "人間 関係," differing only in spacing, were identified to represent the same concept. Since the search trends were similar (Multimedia Appendix 1) and it was considered that searchers had the same intent, only "人間関係" was included.

Finally, we collected the RSVs for all main and related words by prefecture. Google Trends allows for the comparison of search frequencies of specific words across states or prefectures within a selected country using its "Interest by Subregion" function. The function provides a country map shaded according to the term's popularity. The color intensity represents the percentage of searches for the leading search term in a particular region. Search term popularity is relative to the total number of Google searches performed at a specific time, in a specific location. Data were collected for each keyword by setting specific time frames for the years 2010, 2013, 2016, and 2019 (eg, January 1, 2010, to December 31, 2010). In Google Trends, extremely low search frequencies are sometimes indicated as "less than 1" instead of "0," such cases were treated as "0" in this study. The normalization process for RSVs was conducted for the data of each domain within each year, standardizing the annual prefecture-specific data. This means, for example, that the prefecture-specific data in the "Income" domain for the year 2010 was standardized to ensure comparability across prefectures.

Statistical Analysis

A descriptive statistical analysis of the RWIs and RSVs was conducted for all words. We calculated the mean and SD of the RSVs for representative words of each domain for the years 2010, 2013, 2016, and 2019 by the prefecture to explore temporal variations in search behaviors across different regions. We also calculated Pearson partial correlation coefficients by adjusting the data year and population of prefectures [31] to evaluate the association between the RWI scores and the RSV of each keyword.

As a supplementary analysis, we conducted a spatial evaluation of IRWI scores to assess the geographical interrelationships of well-being across regions. Well-being may be influenced by regional cultural characteristics, leading to similarities in IRWI scores among neighboring areas. Understanding these spatial relationships can provide insights for considering regional collaboration in policy interventions.

We first applied Global Moran *I* [32] to assess the overall spatial autocorrelation of IRWI scores across Japan, identifying whether scores were clustered or dispersed on a national level. A significant positive Global Moran *I* indicates the clustering of similar values, while a negative value suggests the dispersion of dissimilar values. Additionally, Local Moran *I* [33] was used to analyze local similarities and differences between regions and their neighbors, allowing us to identify clusters of high or low scores. This analysis highlights regional disparities in well-being, which can inform targeted policy interventions. Prefectures without adjacent regions, such as Hokkaido and Okinawa, were excluded from this analysis.

XSL•FO
RenderX

To predict the RWI, we used the Elastic Net methodology, a machine-learning technique, that was designed to prevent the problems of multicollinearity and overfitting in a linear regression through regularization [34]. L1 regularization (lasso regression) applies a penalty to the absolute values of the coefficients, playing a role in excluding unnecessary variables from the model. L2 regularization (ridge regression) applies a penalty to the squared coefficients, thereby reducing the coefficients of highly correlated explanatory variables to overcome multicollinearity. The Elastic Net combines these 2 types of regularization, enabling the creation of a model that addresses the multicollinearity between explanatory variables and selects important features. To further prevent overfitting, we estimated model parameters using data from 2010, 2013, and 2016, and then assessed the prediction errors with 2019 data. This approach enabled us to evaluate the model's performance on out-of-sample predictions, ensuring that it mitigates overfitting.

In this model, RSVs and prefectures served as predictor variables (with prefectures coded as dummy variables), whereas the IRWI score was used as the outcome variable. The performance of Elastic Net varies with 2 parameters: α, the mix ratio of L1 and L2 regularization, and λ, the strength of regularization. The optimal values for α and λ were determined by fixing α and identifying the λ that minimized the mean squared error through 10-fold cross-validation. This procedure was repeated 11 times, incrementally adjusting α from 0 to 1 by 0.1, and the model with the lowest mean squared error was selected. The predictive accuracy of the model was assessed using the root mean square error (RMSE) and the coefficient of determination (R^2). All statistical analyses, including standard normalization, Pearson correlation coefficient calculation, and Elastic Net processing, were conducted using R software (version 4.1.3, The R Foundation).

Ethical Considerations

This study used publicly accessible data from Google Trends and open government statistics for secondary use to eliminate the need for ethical review.

Results

Summary of RWI Scores

The median IRWIs by prefecture in Japan were 0.67 (IQR −2.48 to 2.71), 0.00 (IQR −2.85 to 2.76), 0.13 (IQR −3.05 to 2.49), and 0.19 (IQR −2.75 to 3.06) for 2010, 2013, 2016, and 2019, respectively (Table 1) [30]. The RWIs by 11 domains were also shown in Multimedia Appendices 2-5.

From 2010 to 2019, the Global Moran *I* statistics for the spatial analysis of IRWI scores across Japan ranged from 0.297 to 0.526, reflecting significant spatial autocorrelation throughout the entire period (Figure 2). Corresponding *P* values varied from 4.955×10^{3} to 3.892×10 , indicating strong statistical significance. For Local Moran *I*, median values in 2010, 2013, 2016, and 2019 were 0.333 (IQR 0.036-0.858), 0.214 (IQR 0.006-0.746), 0.113 (IQR –0.020 to 0.825), and 0.146 (IQR 0.055 to 0.829), respectively. Figure 2 shows the spatial distribution of Local Moran *I* scores, with regions color-coded according to their values for each year. Additionally, each prefecture is marked with red or blue dots to indicate whether their IRWI scores were above or below the median. This visualization showed the spatial clustering of regions with high or low IRWI scores.

XSL•FO
RenderX

Table 1. Integrated scores of the Regional Well-Being Index. All scores were standardized and unitless.

Prefecture	Year			
	2010	2013	2016	2019
Hokkaido	−2.02	−2.66	−4.28	−4.46
Aomori	−4.13	−6.69	−2.48	−6.45
Iwate	−4.46	−3.82	−3.86	−7.35
Miyagi	−4.55	1.85	−2.77	−2.49
Akita	2.21	3.08	−0.74	0.19
Yamagata	0.91	1.90	1.22	2.04
Fukushima	−1.62	−1.68	−3.27	−1.14
Ibaraki	2.51	−0.43	0.55	0.27
Tochigi	0.67	−0.66	1.05	−1.91
Gunma	2.93	2.32	0.58	0.79
Saitama	−1.26	0.00	−0.22	0.53
Chiba	1.54	1.31	0.42	−1.80
Tokyo	2.30	1.74	5.82	4.80
Kanagawa	1.30	2.79	0.82	3.14
Niigata	1.76	1.62	2.54	2.98
Toyama	3.95	0.95	5.20	−0.20
Ishikawa	4.04	2.42	6.30	5.50
Fukui	4.22	6.52	9.67	4.31
Yamanashi	8.14	7.83	6.91	5.77
Nagano	5.36	4.92	1.85	2.60
Gifu	4.52	6.45	7.45	5.52
Shizuoka	6.62	5.34	4.24	3.81
Aichi	3.80	4.74	5.99	4.12
Mie	2.61	7.41	5.95	7.30
Shiga	4.31	2.93	4.13	5.81
Kyoto	−2.88	−3.08	−0.36	−0.73
Osaka	−8.97	−10.18	−7.35	−7.97
Hyogo	−3.27	−2.95	−0.03	2.30
Nara	2.81	5.67	1.04	4.57
Wakayama	−0.30	2.72	0.13	2.76
Tottori	2.00	−1.25	−0.43	−4.44
Shimane	6.28	3.56	3.35	−0.54
Okayama	−2.38	−1.39	−4.01	−4.63
Hiroshima	−2.32	−4.22	−2.48	−2.36
Yamaguchi	1.33	0.12	2.43	3.14
Tokushima	−2.58	−4.79	−4.00	−1.76
Kagawa	−3.98	−1.38	−2.01	−3.01
Ehime	−3.30	−3.19	−4.06	2.69
Kochi	−5.27	−7.26	−8.56	−3.58
Fukuoka	−6.88	−7.70	−10.52	−5.74
Saga	0.03	−1.85	−4.11	−3.36

XSL•FO
RenderX

Prefecture	Year			
	2010	2013	2016	2019
Nagasaki	−1.66	−2.12	−0.10	−0.83
Kumamoto	−1.34	−2.84	−4.43	−7.07
Oita	−2.17	−1.23	1.21	2.92
Miyazaki	−0.97	1.85	1.07	2.52
Kagoshima	1.06	−2.86	−2.82	−1.82
Okinawa	−10.91	−5.81	−7.05	−6.77
Overall, median (IQR)	0.67 (−2.48 to 2.71)	0.00 (−2.85 to 2.76)	0.13 (−3.05 to 2.49)	0.19 (−2.75 to 3.06)

Figure 2. Spatial distribution of IRWI scores across Japan (2010-2019). IRWI scores are categorized as "High" or "Low" based on whether they are above or below the median. IRWI: integrated Regional Well-Being Index.

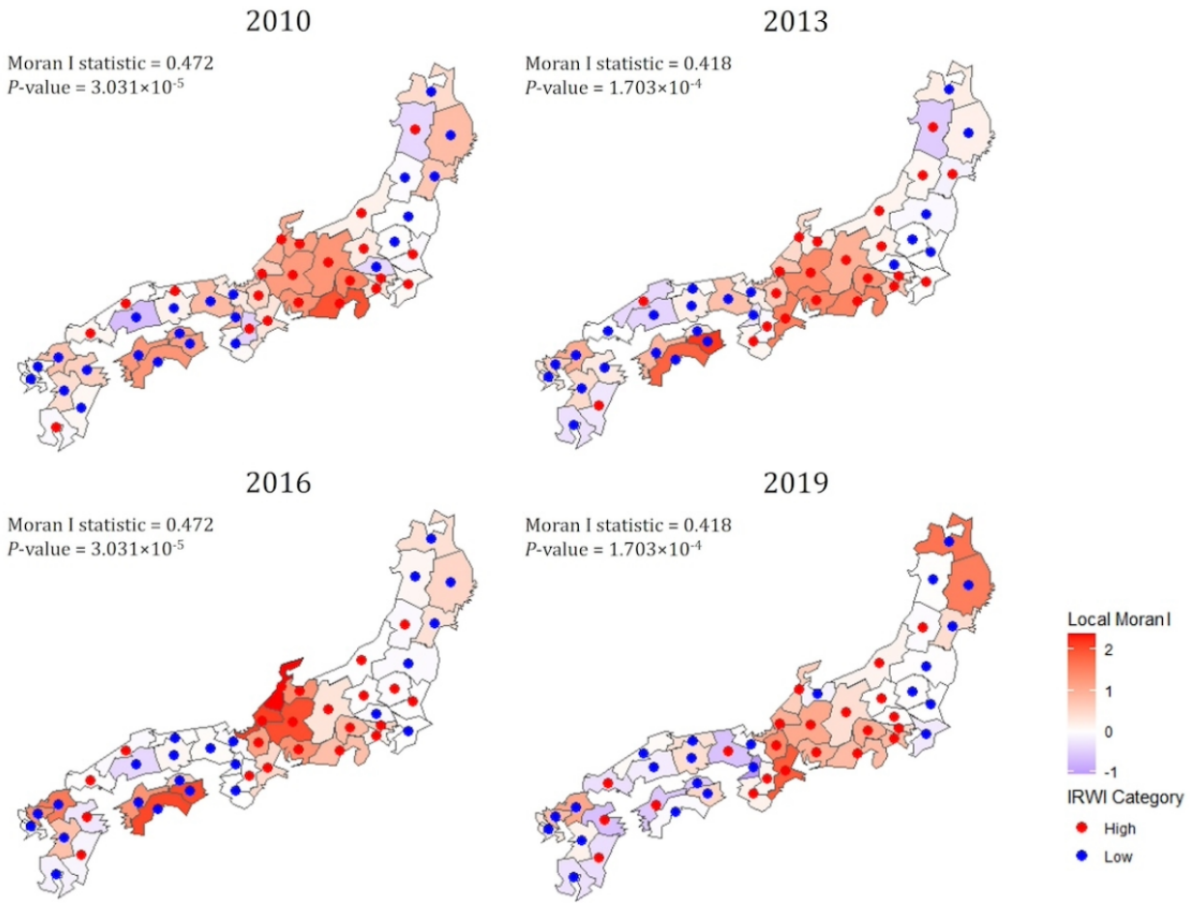

Characteristics of Words and Their RSVs

Eleven representative words were extracted for each domain, following a previous study [30], and 275 related words were associated with these representative words. Of these, the RSVs for 211 words were collected and those that met the exclusion criteria were excluded (Figure 1 and Multimedia Appendix 6). The mean search frequencies for the representative words of each domain during the data collection period (2010, 2013, 2016, and 2019) ranged from −1.587 to 3.902, with SDs ranging from 3.025 to 0.053 (Figure 3).

The partial correlation coefficients between each domain of the RWI and the RSV of each word varied in ranges, indicating the minimum and maximum extents of correlation. Specifically, the coefficients ranged as follows: for "Income," from −0.301 to 0.226; for "Jobs," from −0.315 to 0.133; for "Housing," from −0.604 to 0.225; for "Health," from −0.283 to 0.297; for "Work-Life Balance," from −0.285 to 0.350; for "Education," from −0.396 to 0.269; for "Community," from −0.216 to 0.063; for "Civic Engagement," from −0.233 to 0.269; for "Environment," from −0.070 to 0.261; for "Safety," from −0.183 to 0.036; and for "Life Satisfaction," from −0.112 to 0.219, as detailed in Multimedia Appendix 6. Additionally, the overall range for the partial correlation coefficient between the IRWI and RSV for each word was determined to be from −0.409 to 0.362, as also noted in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 3. Temporal variations in regional relative search volumes: Descriptive statistics of mean and SD. “avg” is the average relative search volumes for 2010, 2013, 2016, and 2019. “sd” is the SD of relative search volumes in 2010, 2013, 2016, and 2019.

		Income	Jobs	Housing	Health	Work-Life Balance	Education	Community	Civic Engagement	Environment	Safety	Life Satisfaction
Hokkaido	avg	-0.049	0.500	1.097	-0.160	0.202	1.074	-0.072	-0.038	-0.377	0.014	0.094
	sd	0.681	0.440	0.905	0.502	0.276	0.509	0.458	0.533	0.168	0.279	1.135
Aomori	avg	-0.210	0.442	-0.405	-0.925	-0.906	0.085	-0.158	-0.795	-0.246	-0.821	-0.153
	sd	0.104	0.340	0.368	0.406	0.963	0.543	0.577	0.542	0.645	0.515	0.736
Iwate	avg	0.888	0.949	1.322	-0.661	0.110	0.218	0.428	-0.475	0.074	-0.752	0.227
	sd	0.497	0.607	1.405	0.845	0.468	0.669	1.127	0.645	0.473	0.494	1.057
Miyagi	avg	-0.346	0.391	1.217	-0.709	0.262	0.758	0.085	0.236	0.012	-0.151	0.060
	sd	0.449	0.921	1.017	0.773	0.481	0.387	0.679	0.075	0.286	0.432	0.880
Akita	avg	0.575	0.544	1.074	-0.250	-0.933	0.185	1.192	0.762	0.261	-1.066	-0.324
	sd	0.192	0.483	0.447	0.575	1.173	0.125	1.343	0.564	0.297	0.432	0.845
Yamagata	avg	0.028	0.725	-0.384	-0.300	-0.348	-0.742	-0.130	-0.417	-0.021	-0.763	-0.194
	sd	1.186	0.401	0.600	0.690	0.917	0.680	1.066	1.127	0.694	0.854	0.853
Fukushima	avg	0.076	0.948	0.776	0.131	0.014	-0.210	0.043	-0.388	0.161	-0.948	-0.234
	sd	1.100	0.514	0.823	0.645	1.595	0.510	0.574	0.398	0.530	0.454	0.522
Ibaraki	avg	-1.064	-0.966	-0.096	-0.713	0.150	1.132	-0.424	-0.039	0.857	-0.030	-1.023
	sd	0.606	0.433	0.971	0.372	0.228	1.204	0.724	0.564	0.677	0.378	0.534
Tochigi	avg	-1.216	-1.211	-0.053	-0.534	0.279	-1.329	0.469	-0.532	-0.943	-0.167	-0.004
	sd	0.827	0.734	1.459	1.106	0.875	0.717	0.574	0.643	0.918	0.552	1.302
Gunma	avg	-0.464	-0.986	0.339	0.664	0.102	-0.935	-0.341	-0.283	-0.761	-0.653	-0.792
	sd	0.396	0.624	0.556	0.745	0.253	0.258	0.886	0.596	0.356	0.594	0.544
Saitama	avg	-1.250	-1.118	-0.118	-0.290	0.080	-1.195	-0.318	0.236	-0.624	0.187	-0.023
	sd	0.698	0.704	0.449	0.994	0.545	0.305	0.129	0.259	0.593	0.166	0.641
Chiba	avg	-1.330	-1.146	-0.366	-0.362	0.380	-1.142	-0.315	0.098	-0.998	0.574	-0.559
	sd	0.721	0.537	0.280	0.630	0.589	0.433	0.115	0.215	0.324	0.240	0.772
Tokyo	avg	-0.045	0.489	-0.581	0.210	2.655	-1.037	-0.239	0.697	0.576	1.502	-0.353
	sd	1.959	2.323	1.350	2.358	2.533	1.173	0.659	2.064	2.036	1.143	1.603
Kanagawa	avg	-0.818	-0.991	-0.616	0.170	0.875	-1.154	-0.257	0.676	0.226	1.369	-0.300
	sd	0.502	0.270	0.381	0.801	0.629	0.453	0.383	0.907	0.687	0.336	0.564
Niigata	avg	-0.033	0.103	-0.493	0.477	0.130	0.443	-0.205	-0.254	-0.368	-0.424	0.447
	sd	0.594	0.141	0.626	0.285	0.256	0.264	0.828	0.420	0.551	0.458	0.681
Toyama	avg	0.059	0.382	-0.021	0.957	0.523	-0.384	0.054	0.252	0.729	-0.454	0.008
	sd	0.772	0.625	0.573	0.677	1.083	0.402	0.267	0.727	0.530	0.940	0.943
Ishikawa	avg	0.124	-0.022	-0.627	-0.153	0.366	-0.673	0.371	0.358	0.084	-0.244	-0.172
	sd	0.635	0.409	0.656	0.612	0.808	0.270	1.219	0.470	0.556	0.112	0.676
Fukui	avg	0.073	0.059	-0.869	1.328	0.479	-0.267	-0.104	-0.397	0.666	-0.295	-0.947
	sd	0.906	0.574	0.796	1.097	0.762	0.225	1.074	0.565	0.587	0.388	1.386
Yamanashi	avg	0.176	-0.416	-0.409	2.263	-0.148	0.250	-0.401	1.032	0.375	0.217	0.160
	sd	0.838	0.669	0.679	0.248	0.908	0.816	1.160	0.677	0.230	1.670	1.722
Nagano	avg	-0.173	-0.443	-0.051	-0.151	0.139	-0.356	0.001	-0.369	-0.324	-0.630	-0.719
	sd	0.527	0.388	0.590	0.161	0.664	0.339	0.324	0.927	0.429	0.344	0.743
Gifu	avg	-0.170	0.071	-0.262	-0.104	-0.712	-0.600	-0.138	-0.513	-0.234	-0.577	-0.158
	sd	0.236	0.934	1.009	0.674	0.316	0.547	0.590	0.597	0.204	0.434	0.788
Shizuoka	avg	-1.100	-1.029	-0.055	-0.515	0.247	-1.366	-0.163	-0.574	-1.038	0.012	-0.985
	sd	0.496	0.631	0.565	0.909	0.278	0.137	0.658	0.383	0.582	0.323	0.839
Aichi	avg	-0.686	-0.751	-0.045	0.229	0.424	-0.855	-0.073	-0.917	-0.341	0.165	-0.554
	sd	1.104	0.435	0.271	0.860	0.431	0.396	0.285	0.197	0.649	0.070	0.412
Mie	avg	-0.774	-0.749	-1.587	-1.168	0.386	-0.789	0.073	-0.666	-0.739	-0.132	-0.351
	sd	0.279	0.464	0.621	0.373	0.689	0.463	0.990	0.673	0.053	0.284	0.780
Shiga	avg	-0.760	-0.844	-0.978	-0.458	-0.093	-0.166	0.655	0.356	-0.030	1.027	-0.095
	sd	0.500	0.553	0.745	0.551	0.336	0.428	0.782	0.314	0.480	0.307	0.633
Kyoto	avg	-0.547	-0.575	-1.139	0.010	0.219	0.629	-0.408	1.971	0.495	2.805	-0.304
	sd	0.157	0.274	0.656	0.813	0.311	0.597	0.709	0.942	0.617	0.680	0.418
Osaka	avg	-0.719	-0.661	0.346	-0.023	0.551	-0.718	-0.291	-0.496	-0.699	3.096	1.308
	sd	1.082	1.003	1.461	1.263	1.135	0.826	0.765	0.304	1.047	0.514	1.278
Hyogo	avg	-0.036	-0.987	-0.187	0.398	0.253	0.098	-0.393	-0.099	-0.183	1.702	0.319
	sd	0.348	0.132	0.236	0.904	0.234	0.586	0.485	0.280	0.612	0.459	0.687
Nara	avg	-1.036	-1.121	-0.898	0.517	-0.692	1.323	0.252	-0.128	-1.296	1.455	0.792
	sd	0.292	0.588	0.693	1.534	0.756	1.216	0.667	0.510	0.448	0.408	0.699
Wakayama	avg	0.107	-0.601	-0.662	-1.078	-0.402	0.256	0.124	-0.837	-1.151	-0.191	0.641
	sd	1.030	1.084	0.723	0.749	1.129	0.441	0.840	0.498	0.290	0.149	0.547
Tottori	avg	1.524	1.547	0.870	-0.048	-0.226	1.455	1.010	-0.950	3.902	0.039	1.082
	sd	0.669	1.093	0.445	0.699	0.617	0.258	1.824	1.190	0.977	0.819	1.397
Shimane	avg	1.811	1.386	0.777	0.386	-0.460	1.682	-0.863	-0.867	1.461	-0.899	0.252
	sd	0.557	0.447	0.496	0.741	0.328	0.597	2.796	0.943	0.369	1.145	1.567
Okayama	avg	0.143	0.254	-0.361	0.661	0.140	0.177	-0.436	-0.112	0.117	0.237	0.198
	sd	0.274	0.293	0.451	0.623	0.326	0.562	0.684	0.326	0.369	0.561	0.305
Hiroshima	avg	-0.420	-0.499	-0.547	-0.522	0.304	-0.191	0.411	-0.252	-0.319	-0.002	-0.142
	sd	0.256	0.394	0.429	0.340	0.467	0.348	0.181	0.367	0.292	0.243	1.003
Yamaguchi	avg	0.641	-0.274	-0.139	0.942	-1.022	-0.224	0.427	0.239	0.079	-0.681	1.406
	sd	0.684	0.429	1.001	0.515	0.660	0.549	0.783	0.804	0.511	0.458	1.553
Tokushima	avg	0.219	0.083	-0.736	0.820	0.577	2.653	1.144	-0.193	1.081	-0.012	0.094
	sd	0.865	0.187	0.253	0.081	0.687	0.385	2.113	0.400	0.482	1.139	0.859
Kagawa	avg	0.741	-0.083	-0.113	0.238	-0.146	0.251	-0.221	0.537	-0.001	0.722	0.433
	sd	0.847	0.249	0.899	0.452	1.129	0.349	1.194	0.156	0.447	0.301	0.984
Ehime	avg	-0.041	-0.022	-0.066	-0.767	-0.769	-0.248	0.848	0.088	-0.328	-0.740	0.411
	sd	0.056	0.707	0.425	0.667	0.301	0.377	0.623	0.346	0.380	1.239	0.462
Kochi	avg	1.265	1.328	1.145	-0.171	-1.042	1.563	0.350	1.454	0.009	-0.116	-0.444
	sd	0.992	0.827	0.591	0.708	1.370	0.502	1.239	1.291	0.143	0.372	0.769
Fukuoka	avg	-0.241	-0.517	0.184	-0.420	0.208	-0.922	-0.302	-0.483	-0.531	0.668	-0.091
	sd	0.478	0.554	0.583	0.522	0.324	0.367	0.533	0.724	0.714	0.390	0.635
Saga	avg	0.901	0.914	0.137	0.285	-0.660	0.446	0.559	1.376	0.507	-0.692	0.760
	sd	1.438	0.799	0.588	0.847	1.599	0.830	0.719	3.025	0.488	0.573	0.988
Nagasaki	avg	0.250	0.667	0.145	-0.461	0.203	-0.376	-0.144	0.006	0.381	-0.650	0.455
	sd	0.993	1.114	0.985	1.699	1.111	0.448	1.190	1.157	1.548	0.295	0.381
Kumamoto	avg	0.359	0.660	0.810	0.245	-0.261	-0.029	-0.924	-0.283	-0.176	-0.444	0.385
	sd	1.064	0.551	1.419	0.816	0.914	0.556	0.958	0.705	0.605	0.256	1.238
Oita	avg	0.713	0.990	-0.259	0.491	-0.585	0.995	0.007	0.405	0.071	-0.159	-0.658
	sd	0.759	0.241	0.515	0.705	0.212	0.466	0.697	1.040	0.391	1.252	0.529
Miyazaki	avg	0.826	0.523	2.242	-0.484	-0.740	-0.068	1.125	0.385	-0.073	-0.519	-0.043
	sd	0.837	0.470	0.683	0.762	0.925	0.450	1.060	0.835	0.151	0.424	0.421
Kagoshima	avg	0.644	0.404	0.140	0.636	-1.005	0.244	-1.330	-0.499	-0.218	-0.878	0.760
	sd	0.691	1.130	0.657	1.505	0.437	0.900	0.592	1.369	1.281	0.287	1.179
Okinawa	avg	0.382	0.651	0.501	-0.629	-0.107	0.361	-0.973	0.693	-0.103	-0.400	-0.670
	sd	0.429	0.644	0.667	0.858	1.356	0.459	0.697	1.105	0.308	0.656	1.373

Prediction Model Using Search Log Data and Elastic Net

The best Elastic Net model was constructed using training data from 2010 to 2016, incorporating 2 to 13 variables per domain (α=0.1, λ=0.906, RMSE=1.290, and R^2=0.940). This model was then used to predict outcomes for the 2019 test data, yielding an RMSE of 2.328 and an R^2 of 0.665 (Tables 2 and 3). The model included 2-13 variables per domain as features. The standardized partial regression coefficients for words ranged from −0.386 to 0.489, and for the selected prefectures as features, they ranged from −0.704 to 0.439 (Tables 2 and 3).

XSL•FO
RenderX

Table 2. Elastic Net regression analysis metrics.

Index	Words selected for the model (standardized partial regression coefficient)[a]
Intercept	4.078×10^{-15}
Income	給与 (−0.063), 所得 申告 (−0.006), 所得 証明 (−0.008), 所得 証明 書 (−0.021), 確定 申告 (0.489), 扶養 所得 (0.030), 譲渡 所得 (0.117), 退職 所得 (0.028), 住民 税 (0.013)
Jobs	雇用 契約 (0.226), 雇用 保険 証 (0.020), 雇用 保険 者 証 (−0.125), 雇用 保険 被 保険 者 証 (0.043), 助成 金 雇用 (−0.104), 雇用 保険 料率 (−0.011), 失業 保険 (−0.386)
Housing	住宅 (−0.116), 住宅 控除 (0.087), 賃貸 住宅 (−0.143), 市営 住宅 (−0.107), 住宅 ローン 控除 (−0.046), 注文 住宅 (0.033), マンション (−0.264), 住宅 展示 場 (−0.127), 住宅 情報 (−0.381), 高齢 者 住宅 (−0.025), リフォーム (0.059), 分譲 住宅 (0.299), エコ ポイント 住宅 (0.171)
Health	健康 保険 組合 (0.360), 健康 センター (0.035), 健康 ランド (0.320), 健康 管理 (−0.048), 健康 保険 協会 (−0.026), 保険 証 (−0.069), 全国 健康 保険 協会 (−0.275), 健康 保険 とは (−0.135), 健康 管理 センター (0.075)
Work-life balance	残業 手当 (−0.018), 労働 時間 (−0.164), 労働 基準 法 (0.001), 転職 (0.130), 月 残業 時間 (−0.159)
Education	教育 (−0.118), 教育 大学 (−0.068), 教育 委員 会 (−0.338), 特別 教育 (−0.145), 教育 指導 (−0.247), 教育 ローン (−0.179), 教育 研究 所 (0.055), 義務 教育 (0.351), 教育 問題 (0.113)
Community	疲れ た (−0.019), 人間関係 悩み (−0.053), 人間 関係 苦手 (0.141)
Civic engagement	政治 (−0.048), 政治 ブログ (−0.307), 選挙 (0.250), 政治 問題 (−0.160)
Environment	環境 基準(−0.330), 自然 環境 (0.259)
Safety	治安 (−0.134), 日本 治安 (−0.016)
Life satisfaction	幸せ (−0.063), 幸せ の 時間 (−0.209), 幸せ 画像 (−0.007), 幸せ に なりたい (−0.133), 幸せ に なろう (−0.154), 幸せ に なるために (−0.034), 小さな 幸せ (−0.094)
Prefecture	Aomori (−0.260), Iwate (−0.246), Akita (0.010), Fukushima (−0.081), Saitama (−0.031), Tokyo (0.167), Toyama (0.033), Ishikawa (0.126), Fukui (0.275), Yamanashi (0.319), Nagano (0.089), Gifu (0.439), Shizuoka (0.323), Aichi (0.415), Mie (0.334), Shiga (0.243), Osaka (−0.704), Shimane (0.269), Okayama (−0.171), Hiroshima (−0.159), Yamaguchi (0.050), Tokushima (−0.188), Kagawa (−0.245), Kochi (−0.232), Fukuoka (−0.505), Kumamoto (−0.069), Miyazaki (0.044), Okinawa (−0.497)

[a]Please refer to Multimedia Appendix 6 for English translations of Japanese words.

Table 3. Model accuracy statistics.

	Training data (2010-2016)	Test data (2019)
α[a]	0.1	—[b]
λ[c]	0.906	—
RMSE[d]	1.290	2.328
R^{2e}	0.904	0.665

[a]α represents the proportion of L1 to L2 regularization in Elastic Net. This value is calculated using the training data and is part of the model used to predict the test data but is not recalculated for the test data.

[b]Not applicable.

[c]λ represents the intensity of regularization. This value is calculated using the training data and is part of the model used to predict the test data but is not recalculated for the test data.

[d]RMSE: root mean square error; the square root of the mean squared error.

[e]R^2 (coefficient of determination) is the determination coefficient's value.

Discussion

Principal Findings

The primary aim of this study was to predict the RWI for each prefecture in Japan using internet search log data. The best model in this study achieved an out-of-sample R^2 value of 0.665 (in-sample R^2 of 0.904; Tables 2 and 3), which is relatively high compared with the R^2 values ranging from 0.005 to 0.830 reported in previous regional-level well-being studies using web data (Table 4) [27,28,35-39]. Most previous studies have focused on predicting subjective well-being or single-objective indicators, representing only one aspect of well-being. Unlike earlier research, this study's comprehensive approach to predicting the RWI underscores the efficacy of internet search data in evaluating overall well-being.

XSL•FO
RenderX

Table 4. Previous studies predicting well-being-related indicators using web data.

Levels and reference	Outcomes	Big Data measure	Big Data source	R^{2a}
Individuals				
Kosinski et al [35]	Life satisfaction	Type of Facebook pages liked	Facebook	0.003
Schwartz et al [36]	Life satisfaction	Posts	Facebook	0.090
Panicheva et al [37]	Life satisfaction	Message, Posts, App Usage, App Time, Number of friends	VKontakte[b]	0.116
Panicheva et al [37]	Psychological well-being	Message, App Usage, App Time, Number of friends, Posts	VKontakte	0.119
National				
Algan et al [28]	Life evaluation today	Word searches	Google Trends	0.940
Algan et al [28]	Life evaluation in 5 years	Word searches	Google Trends	0.840
Algan et al [28]	Happiness	Word searches	Google Trends	0.590
Carpi et al [38]	Human Development Index	Tweets	Twitter	0.640
State				
Durahim and Coskun [39]	Happiness	Tweets	Twitter	0.005
Algan et al [28]	Life evaluation today	Word searches	Google Trends	0.720
Algan et al [28]	Life evaluation in 5 years	Word searches	Google Trends	0.820
Algan et al [28]	Happiness	Word searches	Google Trends	0.830
Jaidka et al [27]	Life satisfaction	Tweets	Twitter	0.384
Jaidka et al [27]	Happiness	Tweets	Twitter	0.260

[a]R^2 (coefficient of determination) is the determination coefficient's value.
[b]"VKontakte" is Russia's social networking platform.

The primary advantage of this study is its potential to reduce the time and economic resources required for conventional well-being assessments significantly. The developed model allows for frequent and quick assessment of comprehensive well-being at the prefectural level in Japan using upcoming search log data. This approach could be invaluable for policymaking aimed at enhancing well-being. Moreover, this methodology can be adapted beyond Japan, allowing countries where the BLI is measured to calculate their RWI based on the Japanese approach and to predict regional well-being using web data based on this study's methods. If the accuracy of the model is ensured, it will allow immediate and repeated assessments of comprehensive well-being. Additionally, the use of Google, a platform extensively used globally, in this study suggests its potential applicability even in many resource-limited nations where large-scale surveys pose challenges. The methodology proposed in this study could be tested in various countries, potentially enabling the assessment of well-being and the realization of evidence-based policy making (EBPM) for well-being improvement.

The analysis revealed marked disparities in mean search frequencies and their SDs when segmented by prefecture (Tables 2 and 3). These differences indicate variations in interest in well-being-related topics across regions. Notably, the tendency for higher search frequencies related to work-life balance in major urban areas, such as Tokyo and Osaka, might suggest an increased awareness of the working environment in these regions. Furthermore, the magnitude of the SDs reflects the degree of variability in search behavior during the observation period, suggesting the potential to assess changes in the interests and states of well-being among residents of different regions.

In the analysis of the relationship between each word and regional well-being indicators (RWI) as well as IRWI in this study, correlation coefficients showed both positive and negative values, with many remaining within the range of ±0.3. Notably, in the "Housing" domain, there were relatively high positive correlations, such as a correlation coefficient of 0.6 for "マンション." A positive correlation coefficient suggests that as the search frequency for a word increases, so does the RWI domain or IRWI score. Although these correlations do not establish causality, they indicate a potential relationship between search behaviors related to specific words and the well-being of people in that area or with more comprehensive well-being. This implies that internet search data could be a viable means of understanding well-being.

As a supplementary analysis, Global Moran *I* results indicate spatial autocorrelation in Japan's IRWI scores, suggesting that well-being in Japan is not randomly distributed but exhibits spatial patterns. Additionally, Local Moran *I* indicated these spatial patterns specifically form. High-score clusters were predominantly found in the Chubu and Tokai regions, while low-score clusters were identified in parts of the Tohoku and Kyushu regions, as well as in Shikoku. This suggests that regional differences in well-being may not be solely due to isolated factors but could also be influenced by interrelationships with adjacent areas, possibly reflecting common factors across

XSL•FO
RenderX

multiple regions. This finding could provide important implications for policy interventions aimed at improving well-being.

Limitations

While focusing on predicting a comprehensive well-being index (IRWI), this study did not consider the causal relationship between fluctuations in search log data and changes in each well-being indicator of the RWI. Although the out-of-sample R^2 and RMSE indicate a high degree of accuracy, the model might exhibit slight overfitting. This may be due to the limited data available, as the evaluation is restricted to a single year (2019). The RWIs used as outcomes in this study are based on somewhat outdated data, with 2019 being the most recent year available due to the unavailability of certain indicators constituting the RWI. Subsequent updates have been delayed by factors such as the COVID-19 pandemic and extended data aggregation processes. However, the year-to-year variations in RWIs are typically not substantial, suggesting that insights derived from data up to 2019 remain valuable and pertinent. Furthermore, including data from 2020 onward would introduce significant confounding effects due to the pandemic's impact. Thus, the examination of these postpandemic trends and their implications remains an important issue for future research. Additionally, since this was an ecological study that used data at the prefecture-level rather than at the individual level, we could not evaluate associations at the individual level. It also did not address the correlations with specific RWI domains or the relative importance of each domain. The well-being indicators, BLI and RWI, are designed to allow users to assess the significance of each domain in a flexible manner. Therefore, the results of this study, which treated each domain uniformly, may differ from the interpretation of the RWI when used in a prefecture. Consequently, implementation in the field should be a topic for future research. Another limitation is the lack of consideration for the emotional valence of search queries and the reason why certain specific words significantly correlated with well-being. Although RSV captures public interest—which implicitly includes emotional aspects—the technical challenges of applying natural language processing techniques to individual search terms limited their use in this study. Therefore, future research is warranted to better understand regional characteristics and to derive more accurate interpretations of searchers' intentions from search terms associated with well-being scores. Addressing these challenges could provide deeper insights into the relationship between web-based behavior and well-being.

In addition to these limitations, Google Trends data was limited by its relative nature, which might not have fully captured actual search activity. While it is possible that infrequently searched words would be overestimated, we believe that the impact is limited in this study because niche words were excluded during the process of extracting representative words. Additionally, the algorithms behind Google's search functions introduced uncertainty in interpreting search intent [40]. It was also noted that the granularity of Google Trends data could have led to limitations in predictive accuracy [41].

Comparison With Prior Work

This study successfully improved the predictive accuracy of comprehensive well-being indices at the regional level through the use of web data, achieving relatively higher accuracy compared with the outcomes of previous studies, as shown in Table 4. Kosinski et al [35] and Schwartz et al [36], used Facebook data and predicted individual-level life satisfaction but achieved relatively low R^2 values of 0.003 and 0.090, respectively. Panicheva et al [37] integrated various types of web data, including messages and attributes from VKontakte, a SNS popular in Russia and its neighboring countries, to predict individual-level life satisfaction and mental well-being; however, the improvement in R^2 was limited.

In contrast, predictions at broader national and state levels yielded more accurate results. A notable example is Algan et al [28], who achieved high predictive accuracy with R^2 values of 0.940 at the national level and 0.720 at the state level for Life Satisfaction. Our study showed similar levels of accuracy in forecasting a more comprehensive well-being index, effectively capturing trends in regional well-being.

Furthermore, the predictive model in this study demonstrated a level of accuracy comparable to that reported by Carpi et al [38], who predicted the Human Development Index, a comprehensive composite index analogous to the RWI used in this research. This result suggests that our methodology can efficiently and accurately predict comprehensive well-being indicators using web data, suggesting that it can quickly identify broader well-being trends while saving time and money.

Implications and Actions Needed

Using the RWI calculation methodology of this study, OECD countries have the opportunity to calculate the RWI for their regions and apply these in regional policymaking. Moreover, the creation of high-accuracy models using internet search data facilitates the timely and continuous assessment of well-being. Governments and international organizations are shifting their focus from merely economic development and life expectancy to improving well-being; however, integrating well-being indicators into national and regional policy goals is traditionally time-consuming and costly. This study demonstrates the efficacy of our approach in predicting comprehensive well-being indicators through web data, indicating its potential to rapidly assess broader trends in well-being with cost and time efficiency. Furthermore, these internet search data are likely to be less susceptible to biases common in traditional survey methods, such as recall and social desirability biases, possibly unveiling aspects of well-being that conventional approaches overlook. Comparing this approach with traditional methodologies may yield insights into societal well-being and provide foundational data for policymaking and evaluations aimed at improving well-being. If this approach is effective, it could extend the reach of well-being assessments to additional countries and regions, thereby accelerating the adoption of policies designed to improve societal well-being.

XSL•FO
RenderX

Conclusion

This study predicted RWIs for Japanese prefectures with high accuracy using RSVs from internet search engines and an Elastic Net machine learning method. This approach provides an immediate and cost-effective alternative to traditional survey methods for comprehensive well-being assessments. It enables ongoing and quick assessments and serves as foundational data for EBPM focused on enhancing well-being. Adding the well-being indicators predicted by the method proposed in this study to conventional policy indicators will enable the agile detection of changes in the population and provide basic data for the discovery of new health issues and policy formulation. Moreover, this methodology suggests a potential solution for assessing well-being in resource-limited nations and regions, where large-scale epidemiological surveys are impractical. However, this study reflects ecological trends rather than individual behaviors, and further research is warranted to identify causal relationships between individual search terms and well-being.

Acknowledgments

We thank Dr Takahiro Itaya of Kyoto University for the insightful comments and constructive feedback during the initial stages of this study. We also wish to express our profound thanks to Dr Shiomi Misa for her persistent guidance and advice, which were invaluable over the entire course of our research. During the preparation of this work, the author used DeepL [42] and ChatGPT (OpenAI) [43] in order to improve the English language. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of search frequencies for "人間関係" (human relationship) and "人間 関係" on Google Trends.
[DOCX File , 26 KB - jmir_v26i1e64555_app1.docx]

Multimedia Appendix 2
Scores of the Regional Well-Being Index for the year 2010.
[DOCX File , 38 KB - jmir_v26i1e64555_app2.docx]

Multimedia Appendix 3
Scores of the Regional Well-Being Index for the year 2013.
[DOCX File , 38 KB - jmir_v26i1e64555_app3.docx]

Multimedia Appendix 4
Scores of the Regional Well-Being Index for the year 2016.
[DOCX File , 38 KB - jmir_v26i1e64555_app4.docx]

Multimedia Appendix 5
Scores of the Regional Well-Being Index for the year 2019.
[DOCX File , 38 KB - jmir_v26i1e64555_app5.docx]

Multimedia Appendix 6
Representative words in the Better Life Index domains and related words for each domain.
[DOCX File , 74 KB - jmir_v26i1e64555_app6.docx]

References

1. Health-related quality of life (HRQOL): well-being concepts. Centers for Disease Control and Prevention. 2024. URL: https://archive.cdc.gov/www_cdc_gov/hrqol/wellbeing.htm [accessed 2024-08-29]
2. Lee MT, Kubzansky LD, VanderWeele TJ. Measuring Well-Being: Interdisciplinary Perspectives from the Social Sciences and the Humanities. Oxford, United Kingdom: Oxford University Press; 2021.
3. Diener E, Seligman MEP. Beyond money: toward an economy of well-being. Psychol Sci Public Interest 2004;5(1):1-31. [doi: 10.1111/j.0963-7214.2004.00501001.x] [Medline: 26158992]
4. Diener E, Seligman MEP. Measure for measure: the case for a national well-being index. Sci Spirit 2006;17(2):36-37. [doi: 10.3200/sspt.17.2.36-37]

XSL•FO
RenderX

5. Kahneman D, Krueger AB, Schkade D, Schwarz N, Stone A. Toward national well-being accounts. Am Econ Rev 2004;94(2):429-434. [doi: 10.1257/0002828041301713]
6. Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ Open 2016;6(7):e010641. [doi: 10.1136/bmjopen-2015-010641] [Medline: 27388349]
7. Steptoe A, Deaton A, Stone AA. Subjective wellbeing, health, and ageing. Lancet 2015;385(9968):640-648. [doi: 10.1016/S0140-6736(13)61489-0] [Medline: 25468152]
8. Stiglitz J, Sen A, Fitoussi JP. Report by the commission on the measurement of economic performance and social progress. INSEE 2009. [doi: 10.2139/ssrn.1714428]
9. Kakwani N. Performance in living standards. J Dev Econ 1993;41(2):307-336. [doi: 10.1016/0304-3878(93)90061-q]
10. Twenge JM. The age of anxiety? birth cohort change in anxiety and neuroticism, 1952-1993. J Pers Soc Psychol 2000;79(6):1007-10021. [doi: 10.1037//0022-3514.79.6.1007] [Medline: 11138751]
11. Diener E, Tov W. In: Land KC, Michalos AC, Sirgy MJ, editors. Handbook of Social Indicators and Quality of Life Research. Amsterdam, the Netherlands: Springer; 2012:137-157.
12. How's life?: measuring well-being. Organization for Economic Co-operation and Development. 2011. URL: https://www.oecd-ilibrary.org/economics/how-s-life_9789264121164-en [accessed 2023-12-09]
13. Helliwell JF, Layard R, Sachs JD, Aknin LB, de Neve NE, Wang S. World Happiness Report.: Sustainable Development Solutions Network URL: https://worldhappiness.report/ed/2023/ [accessed 2023-12-09]
14. Human development index. United Nations Development Programme. 2023. URL: https://hdr.undp.org/data?center/human?development?index#/indicies/HDI [accessed 2023-12-09]
15. New Zealand Treasury. Living Standards Framework Dashboard. 2023. URL: https://lsfdashboard.treasury.govt.nz/wellbeing/ [accessed 2023-12-09]
16. UK measures of national well-being dashboard. Office for National Statistics. 2023. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/ukmeasuresofnationalwellbeing/dashboard [accessed 2023-12-09]
17. National wealth indicators. French National Institute for Statistics Economic Studies. 2024. URL: https://www.insee.fr/fr/statistiques/3281635?sommaire=3281778#documentation-sommaire [accessed 2024-08-26]
18. Measuring-well-being. National Institute of Statistics. 2024. URL: https://www.istat.it/statistiche-per-temi/focus/benessere-e-sostenibilita/la-misurazione-del-benessere-bes/ [accessed 2024-08-26]
19. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
20. Lyubomirsky S, Lepper HS. A measure of subjective happiness: preliminary reliability and construct validation. Soc Indic Res 1999;46(2):137-155. [doi: 10.1023/A:1006824100041]
21. OECD regional well-being. Organization for Economic Co-operation and Development. URL: https://www.oecdregionalwellbeing.org/ [accessed 2023-12-09]
22. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015;26(2):159-169. [doi: 10.1177/0956797614557867] [Medline: 25605707]
23. Llorente A, Garcia-Herranz M, Cebrian M, Moro E. Social media fingerprints of unemployment. PLoS One 2015;10(5):e0128692. [doi: 10.1371/journal.pone.0128692] [Medline: 26020628]
24. Mendoza M, Poblete B, Valderrama I. Nowcasting earthquake damages with Twitter. EPJ Data Sci 2019;8(1):3. [doi: 10.1140/epjds/s13688-019-0181-0]
25. Kadar C, Rosés BR, Pletikosa I. Measuring ambient population from location-based social networks to describe urban crime. 2017 Presented at: International conference on social informatics; September 3, 2017; Glasgow, United Kingdom p. 521-535. [doi: 10.1007/978-3-319-67217-5_31]
26. Xie Z, Liu G, Wu J, Wang L, Liu C. Wisdom of fusion: prediction of 2016 Taiwan election with heterogeneous big data. 2016 Presented at: 13th International Conference on Service Systems and Service Management (ICSSSM); June 24-26, 2016; Kunming, China p. 1-6. [doi: 10.1109/icsssm.2016.7538625]
27. Jaidka K, Giorgi S, Schwartz HA, Kern ML, Ungar LH, Eichstaedt JC. Estimating geographic subjective well-being from Twitter: a comparison of dictionary and data-driven language methods. Proc Natl Acad Sci U S A 2020;117(19):10165-10171. [doi: 10.1073/pnas.1906364117] [Medline: 32341156]
28. Algan Y, Murtin F, Beasley E, Higa K, Senik C. Well-being through the lens of the internet. PLoS One 2019;14(1):e0209562 [FREE Full text] [doi: 10.1371/journal.pone.0209562] [Medline: 30633759]
29. Google Trends. URL: https://trends.google.co.jp/trends/?hl=en?US [accessed 2024-10-10]
30. Yang MS, Taira K. Calculating a prefecture-level well-being index in Japan: applying the framework of the OECD's better life index. Nihon Koshu Eisei Zasshi 2024;71(9):455-465 [FREE Full text] [doi: 10.11236/jph.24-002] [Medline: 38866533]
31. Ministry of Internal Affairs and Communications. Population Census. 2020. URL: https://www.e-stat.go.jp/stat-search/file-download?statInfId=000013168605&fileKind=4 [accessed 2023-12-09]
32. Moran PAP. Notes on continuous stochastic phenomena. Biometrika 1950;37(1-2):17-23. [doi: 10.1093/biomet/37.1-2.17]

XSL•FO
RenderX

33. Anselin L. Local indicators of spatial association—LISA. Geogr Anal 2010;27(2):93-115. [doi: 10.1111/j.1538-4632.1995.tb00338.x]
34. Hui Z, Trevor H. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005;67(2):301-320 [FREE Full text] [doi: 10.1111/j.1467-9868.2005.00503.x]
35. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci U S A 2013;110(15):5802-5805 [FREE Full text] [doi: 10.1073/pnas.1218772110] [Medline: 23479631]
36. Schwartz HA, Sap M, Kern ML, Eichstaedt JC, Kapelner A, Agrawal M, et al. Predicting individual welk-being through the language of social media. Pac Symp Biocomput 2016;21:516-527. [Medline: 26776214]
37. Panicheva P, Mararitsa L, Sorokin S, Koltsova O, Rosso P. Predicting subjective well-being in a high-risk sample of russian mental health app users. EPJ Data Sci 2022;11(1):21 [FREE Full text] [doi: 10.1140/epjds/s13688-022-00333-x] [Medline: 35402139]
38. Carpi T, Hino A, Iacus S, Porro G. A Japanese subjective well-being indicator based on Twitter data. Soc Sci Jpn J 2022;25(2):273-296 [FREE Full text] [doi: 10.1093/ssjj/jyac002]
39. Durahim AO, Coşkun M. #iamhappybecause: gross national happiness through Twitter analysis and big data. Technol Forecast Soc Change 2015;99:92-105. [doi: 10.1016/j.techfore.2015.06.035]
40. Bilić P. Search algorithms, hidden labour and information control. Big Data Soc 2016;3(1):205395171665215. [doi: 10.1177/2053951716652159]
41. Jun SP, Yoo HS, Choi S. Ten years of research change using google trends: from the perspective of big data utilizations and applications. Technol Forecast Soc Change 2018;130:69-87. [doi: 10.1016/j.techfore.2017.11.009]
42. DeepL. URL: https://www.deepl.com/ [accessed 2024-10-28]
43. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2024-10-28]

Abbreviations

BLI: Better Life Index
EBPM: evidence-based policy making
GDP: gross domestic product
IRWI: integrated Regional Well-Being Index
OECD: Organization for Economic Co-operation and Development
RWI: Regional Well-Being Index
RSV: relative search volume
RMSE: root mean square error
SNS: social networking service

Edited by A Mavragani; submitted 23.07.24; peer-reviewed by Y Matsuda, R Guo, I Boumahdi; comments to author 13.08.24; revised version received 31.08.24; accepted 08.10.24; published 11.11.24.

Please cite as:
Yang MS, Taira K
Predicting Prefecture-Level Well-Being Indicators in Japan Using Search Volumes in Internet Search Engines: Infodemiology Study
J Med Internet Res 2024;26:e64555
URL: https://www.jmir.org/2024/1/e64555
doi:10.2196/64555
PMID:

©Myung Si Yang, Kazuya Taira. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Concurrent Mentions of Vaping and Alcohol on Twitter: Latent Dirichlet Analysis

Lynsie R Ranker[1], MPH, PhD; David Assefa Tofu[2], MA; Manyuan Lu[2], MS; Jiaxi Wu[3], PhD; Aruni Bhatnagar[4], PhD; Rose Marie Robertson[5,6], MD; Derry Wijaya[2,7], PhD; Traci Hong[8], PhD; Jessica L Fetterman[9], PhD; Ziming Xuan[1], MA, MSc, SCD

[1]Community Health Sciences, Boston University School of Public Health, Boston, MA, United States
[2]Department of Computer Science, Boston University, Boston, MA, United States
[3]Annenberg School of Communication, University of Pennsylvania, Philadelphia, PA, United States
[4]Department of Medicine, University of Louisville, Louisville, KY, United States
[5]American Heart Association Tobacco Regulation and Addiction Center, Dallas, TX, United States
[6]Department of Medicine, Vanderbilt University, Nashville, TN, United States
[7]Monash University, Jakarta, Indonesia
[8]College of Communication, Boston University, Boston, MA, United States
[9]Evans Department of Medicine and Whitaker Cardiovascular Institute, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

Corresponding Author:
Lynsie R Ranker, MPH, PhD
Community Health Sciences
Boston University School of Public Health
801 Massachusetts Avenue, Crosstown Center, 4th Floor
Boston, MA, 02215
United States
Phone: 1 617 358 2310
Email: lranker@bu.edu

Abstract

Background: Co-use of alcohol and e-cigarettes (often called vaping) has been linked with long-term health outcomes, including increased risk for substance use disorder. Co-use may have been exacerbated by the COVID-19 pandemic. Social networking sites may offer insights into current perspectives on polysubstance use.

Objective: The aims of this study were to investigate concurrent mentions of vaping and alcohol on Twitter (subsequently rebranded X) during a time of changing vaping regulations in the United States and the emergence of the COVID-19 pandemic.

Methods: Tweets including both vape- and alcohol-related terms posted between October 2019 and September 2020 were analyzed using latent Dirichlet allocation modeling. Distinct topics were identified and described.

Results: Three topics were identified across 6437 tweets: (1) flavors and flavor ban (n=3334, 51.8% of tweets), (2) co-use discourse (n=1119, 17.4%), and (3) availability and access regulation (n=1984, 30.8%). Co-use discussions often portrayed co-use as positive and prosocial. Tweets focused on regulation often used alcohol regulations for comparison. Some focused on the perceived overregulation of vaping (compared to alcohol), while others supported limiting youth access but not at the expense of adult access (eg, stronger age verification over product bans). Across topics, vaping was typically portrayed as less harmful than alcohol use. The benefits of flavors for adult smoking cessation were also discussed. The distribution of topics across time varied across both pre– and post–regulatory change and pre– and post–COVID-19 pandemic declaration periods, suggesting shifts in topic focus salience across time.

Conclusions: Co-use discussions on social media during this time of regulatory change and social upheaval typically portrayed both vaping and alcohol use in a positive light. It also included debates surrounding the differences in regulation of the 2 substances—particularly as it related to limiting youth access. Emergent themes from the analysis suggest that alcohol was perceived as more harmful but less regulated and more accessible to underage youth than vaping products. Frequent discussions and comparisons of the 2 substances as it relates to their regulation emphasize the still-evolving vaping policy landscape. Social

XSL•FO
RenderX

media content analyses during times of change may help regulators and policy makers to better understand and respond to common concerns and potential misconceptions surrounding drug-related policies and accessibility.

(J Med Internet Res 2024;26:e51870) doi:10.2196/51870

KEYWORDS

e-cigarettes; alcohol; social media; vape; tweet; vaping; alcohol use; co-use; substance use disorder; social networking site; insight; regulation; youth; vaping policy

Introduction

Since 2014, nicotine-containing products such as e-cigarettes (often referred to broadly as vaping) have been the most used tobacco product among youth in the United States [1]. An estimated 14% of US high schoolers currently vape [2], while nearly one-third report consuming alcohol in the past 30 days [1]. Alcohol and vaping can often co-occur. A recent meta-analysis found that vaping was associated with a 6-fold increased odds of alcohol consumption as well as binge drinking or drunkenness [3]. Nicotine, the addictive chemical found in e-cigarettes, activates pathways in the brain that may reinforce addictive behavior and has been found to enhance the pleasurable effects of alcohol consumption as well as increase cravings [4,5]. A recent study using ecological momentary assessment found that youth who reported co-use of e-cigarettes and alcohol were more likely to report high-risk behaviors, including binge drinking, and that co-use was more common in social contexts [6].

Polysubstance use is concerning to clinicians and public health practitioners, as it may increase the risk of adverse long-term health and social outcomes including increased substance use disorder and reduced educational attainment [7-11]. Substance use and addiction concerns have been further exacerbated by the COVID-19 pandemic, which led to increased mental health strain and alcohol use [12]. While research on the long-term health effects of e-cigarettes is still emerging, addiction to alcohol and tobacco has been found to have health consequences including increased cancer risk and exacerbation of mental health disorders [13,14]. For young people, exposure to nicotine is linked to detrimental effects on learning, memory, and attention [15,16]. Further, addiction to nicotine and alcohol can stress relationships, increase feelings of isolation, and increase the risk for injuries and mortality [13,17].

Youth today report being on the web “near constantly” and spend the majority of that time on social networking sites such as YouTube, Twitter (subsequently rebranded X), and Instagram [18]. On these platforms, they are increasingly exposed to substance use–related content generated by peers, influencers, or businesses [19-24]. In one recent survey, nearly three-quarters of middle and high school youth who used social media reported seeing e-cigarette–related content on these platforms [25]. Social learning theory posits that individuals learn vicariously through observing and modeling the behaviors of others [26]. Thus, social media portrayals, which commonly place substances in a positive light, could create or reinforce positive use expectancies by directing attention to youth-appealing content, encouraging modeling of this behavior on their own accounts, and increasing motivation for such behaviors by portraying them as part of a social norm. Indeed, exposure to substance use content on social media is linked with positive use expectancies, norms, and initiation [27-29], and prior work has found associations between exposure to alcohol content on social media and binge drinking among college students [30,31].

One commonly used social networking platform is Twitter. It is a microblogging platform that, as of early 2022, had an estimated 229 million daily active users [32]. Individuals aged 18-49 years are most likely to engage in vaping, and they also represent an estimated 76% of Twitter’s active users [33]. The text-based, timely nature of tweets offers an opportunity to explore social media discourse, and analyses of posts using methods such as content analysis may provide valuable insight into individual’s perceptions and experiences.

Content analyses of tweets have been used to understand a variety of substance use–related topics, including e-cigarette perceptions [34], e-cigarette policy reactions [35], addiction concerns during COVID-19 [36], and e-cigarette cessation campaign responses [37]. In addition, such approaches have been applied to the analysis of other social media forums [38,39]. Findings from this body of literature have helped inform health communication and intervention efforts by assessing the current understanding and viewpoints of potential audiences [35,40]. Analyses have also been conducted to monitor brands that are increasingly using social media as a marketing tool—potentially in violation of rules surrounding marketing to underage youth who use these platforms frequently [38,41,42]. While content analyses have historically relied upon human coding, the application of machine learning techniques to identify emergent topics related to substance use is becoming more common in order to analyze large amount of social media data concurrently [43-45]. Yet, few studies have applied these tools to examine polysubstance use discussions on social media [46,47]. Furthermore, none have examined such social media conversations in times where mental health, lifestyles, and substance use behaviors may be in flux.

The emergence of the COVID-19 pandemic in the United States was a time period characterized by social upheaval as well as key federal tobacco regulatory changes. Specifically, the federal Tobacco 21 law prohibiting sales of tobacco products to those younger than 21 years of age was implemented on December 20, 2019 [48], and the federal ban prohibiting the sale of characterizing flavors (excluding tobacco and menthol) in cartridge-based e-cigarettes was implemented on January 2, 2020 [49]. While several states (eg, Hawaii, California, and Oregon) had already increased the purchase age from 18 to 21 years or placed restrictions on flavored products, both of these federal US-wide policies were major updates to the regulation of tobacco in the United States.

XSL•FO
RenderX

The goal of this study was to apply a computational content analysis method to characterize the common topics—or themes—of a sample of tweets that include both vaping-related and alcohol-related terms. We specifically focus on a time period of upheaval in the United States by using a dataset, which includes tweets during both (1) the emergence and initial peaks of COVID-19 in the United States and (2) the implementation of federal restrictions on tobacco products, including the Tobacco 21 law and the federal flavor ban of cartridge-based e-cigarettes. A deeper understanding of the key terms and emergent themes surrounding co-use may help inform communication campaigns surrounding harm reduction and access to treatment.

Methods

Data Collection and Sample

Tweets were collected using a Twitter firehose application programming interface through Brandwatch, a subscription-based social analytics software. Each tweet was selected for the dataset if (1) it was posted between October 2019 and September 2020, (2) it used at least 1 vaping-related term such as e-cig, vape, and e-liquid (see Table S1 in Multimedia Appendix 1 for full term list), (3) it originated from a handle (username) not identified as an organization or business, and (4) it originated from a geocode within the United States. This led to an initial sample of 63,008 tweets. The selected time period reflects a period of change in the United States, including both the emergence and initial peaks of COVID-19 and new federal restrictions on tobacco products (December 2019: Tobacco 21 and February 2020: federal flavor ban of cartridge-based e-cigarettes).

Similar to prior work, retweets (tweets originally composed by a different Twitter user and reshared by another user) and replies (responses to tweets that may or may not include the original tweet) were retained, as they were assumed to reflect an endorsement of the original post [50].

All tweets containing 1 or more alcohol terms were selected from the overall vaping tweet sample using the *grepl()* function in R (R Foundation for Statistical Computing) to form the final analytic sample of tweets mentioning both vaping and alcohol (N=6437 tweets). Alcohol-related terms used in prior topic modeling studies were reviewed by the study team, and a final set was selected for the current analysis based on a review of the content analysis literature, team consensus, and team review of initial test searches of the dataset. The final list of terms used was selected to ensure a broad search strategy reflected by tweets, which included use-related (tipsy and drunk), event-related (bar crawl and Thirsty Thursday), and product-related (beer and vodka) terms (see Table S1 in Multimedia Appendix 1 for full list) [46,47,51,52]. A subsample of tweets included in the final dataset was checked by the study team to confirm that they contained at least 1 vaping and 1 alcohol-related search term and that the content of the tweet was relevant to vaping or alcohol.

Latent Dirichlet Allocation Modeling and Topic Description

We used latent Dirichlet allocation (LDA) to evaluate the major topics—or themes—among the sampled tweets. LDA is a machine learning topic modeling approach to identify naturally occurring topics. Model selection was based on the coherence score [53-55]. Specifically, coherence scores were evaluated for each topic count, and the highest score was selected to maximize model fit [56]. Tweets within each topic were then reviewed, and the labels for each topic were created by the research team, reflecting qualitatively emerging themes.

Tweet-level descriptive statistics were generated overall and by topic group. The top 15 most salient words overall are reported. The salience of a word is how informative it was for identifying a distinct topic within the model. More salient terms help differentiate across topics. For each topic, we report the top 15 most relevant terms (relevancy metric λ was set to 0.6 to assist topic interpretation as recommended by Sievert and Shirley [57]), sorted by frequency, along with mock tweets reflecting the topic theme. To protect user anonymity, no tweets were quoted verbatim. Instead, we created mock tweets reflecting the shared perspectives and common themes within each topic group.

Metadata and Imputed Covariates

Each tweet contains metadata for the tweet itself and the Twitter handle (Twitter's term for a username). These data are used as descriptive information and for demographic prediction.

We used deep learning techniques such as M3 demographic inference [58] and Bidirectional Encoder Representations From Transformers (BERT) [59] models to predict whether a Twitter handle was likely an individual or an organization, as well as the account user's gender and age. Using machine learning approaches to predict demographics provides additional contextual data to tweets often missing from user accounts [45]. Given that vaping and drinking vary by age and gender, it is critical to understand potential variations in post activity and content by these key demographics [25,60,61]. A handle was required to have ≥50 tweets for gender and organizational prediction via M3 inference using information from the user's Twitter bio, handle, username, and profile image (when available) [58]. M3 is a previously validated deep learning system to predict demographic information with an accuracy of 91.8% and 89.8% for gender and organizational status, respectively [58]. For the current dataset, gender was additionally validated against the gender prediction available in Brandwatch, and M3 inference was found to have a 93% agreement rate. Following prior applications of M3 inference, individuals with a probability score ≥0.5 for being a man were categorized as men, while those with scores <0.5 were categorized as women. Users with ≥0.5 probability score for being an organization were categorized as organizations, while those with scores <0.5 were categorized as individuals. Those categorized as organizations were not included in the analytic dataset, as previously noted in the *Data Collection and Sample* section.

XSL•FO
RenderX

For age prediction (<21 and ≥21 years), we built and trained a BERT [59] model using a dataset of 3000 prelabeled tweets and their last 100 tweets. The training model yielded high accuracy in age prediction (81%). We then used the trained age-BERT model to predict the age (<21 and ≥21 years) of the users using their last 100 tweets (see Multimedia Appendix 1 for additional details on model development).

Tweet sentiment was calculated using Valence Aware Dictionary for Sentiment Reasoning, a lexicon designed for quantitatively analyzing the sentiment of social media text, which has been previously validated and applied to Twitter data [62]. Valence Aware Dictionary for Sentiment Reasoning reviews text and automatically calculates a compound score ranging from –1 (most extreme negative) and +1 (most extreme positive). Scores were used to summarize average tweet sentiment overall and by topic group.

Time Period Indicators and Analyses

Two federal regulations related to tobacco products were announced during the study period. The federal Tobacco 21 law (December 20, 2019) raised the minimum purchasing age in the United States from 18 to 21 years [48]. On January 2, 2020, a federal ban on flavored cartridge-based e-cigarettes was instituted (with exceptions for tobacco- and menthol-flavored products) [49]. We use the second regulatory change as the regulatory change indicator, given the close proximity of their implementation. Dates from January 2, 2020, and later were considered to have occurred in the post–regulatory change period, while dates prior to this date were coded as the pre–regulatory change period. Variation in the prevalence of each topic across regulatory time periods was tested via chi-square tests, and the resulting *P* value was reported.

We also examined variations in topic prevalence by whether they occurred before or after the declaration of the COVID-19 pandemic. Specifically, the World Health Organization declared COVID-19 a global pandemic on March 11, 2020. For this study, dates before versus on and after the pandemic declaration were considered pre– versus post–COVID-19 pandemic declaration periods, respectively. Variation in the prevalence of each topic by pandemic time periods was tested via chi-square tests, and the resulting *P* value was reported.

Ethical Considerations

This study used publicly available Twitter data. Only study staff had access to the full tweets and associated metadata. All data were stored on password-protected devices. No tweets are quoted verbatim to ensure anonymity. There is no path from this manuscript or any supporting material to individual users, usernames, or tweets. The Boston University Institutional Review Board determined that studies using such data did not meet the definition of human participants and was thus exempt.

Results

Demographics

Roughly 10% of tweets that mentioned e-cigarettes or related terms included alcohol-related terms (6437/63,008, 10.2% after the exclusion of organizational accounts). Among the vaping-alcohol tweets where gender and age could be predicted, 61.9% (n=3448) of tweets were designated as likely originating from men, and 78.9% (n=3708) were likely posted by individuals ≥21 years of age (Table 1). The median follower count was 384 (IQR 141-1146), and the accounts followed were 505 (IQR 218-1170). Average tweet sentiment was largely neutral (80.1%). The majority of the tweets (n=5188, 80.6%) were posted in January 2020 or later, after the Tobacco 21 law and cartridge-based e-cigarette bans went into effect. This is slightly higher than would be expected if tweets were evenly distributed across months (75% of tweets would be from January 2020 or later). A little over half of tweets (n=3773, 58.6%) were posted in March 2020 or later, after the COVID-19 pandemic emerged in the United States. This is roughly what would be expected if tweets were evenly distributed across the time periods.

XSL•FO
RenderX

Table 1. Account characteristics among vaping-alcohol tweets overall and by topic.

Account characteristics	Overall (N=6437)	Topic 1: flavors and flavor ban (n=3334)	Topic 2: co-use discourse (n=1119)	Topic 3: availability and access regulation (n=1984)
Gender[a], n (%)				
Man	3448 (61.9)	1942 (65.9)	465 (48.7)	1041 (62.5)
Woman	2119 (38.1)	1005 (34.1)	490 (51.3)	624 (37.5)
Age (>21 years), n (%)[b]	3703 (78.9)	1933 (75.8)	610 (82.4)	1160 (82.2)
Followers, median (IQR)	384.0 (141.0-1146.0)	396.0 (128.0-1351.0)	350.0 (168.0-658.0)	398.5 (150.0-1384.5)
Following, median (IQR)	505.0 (218.0-1170)	529.0 (202.0-1325.0)	375.0 (221.0-694.0)	592.0 (241.0-1459.0)
Tweet sentiment[c] (%)				
Negative	11.3	10	18.3	9.6
Neutral	80.1	81.5	71.6	82.6
Positive	8.6	8.5	10	7.8

[a]Among those where gender could be imputed, overall: n=5567, topic 1: n=2947, topic 2: n=955, and topic 3: n=1665.

[b]Among those where age could be imputed, overall: n=4700, topic 1: n=2549, topic 2: n=740, and topic 3: n=1411).

[c]Percentages based on average compound sentiment score across tweets.

LDA Model Results

A 3-topic model was the best fit for the data (coherence=0.39). Including additional groups resulted in less distinct topics, without improving model fit. Vape was the most common among the top 15 salient terms, followed by alcohol, drink, care, smoke, flavor, and store (Figure S1 in Multimedia Appendix 1). While COVID-19–specific terms (ie, pandemic and COVID-19) were mentioned, no specific COVID-19–related themes or terms were frequent or salient across topics. After qualitative examination of relevant terms and full tweets, the following topic themes were identified: (1) flavors and flavor ban (n=3334, 51.8% of tweets; Figure 1), (2) co-use discourse (n=1119, 17.4%), and (3) availability and access regulation (n=1984, 30.8%).

Figure 1. Top 15 most relevant terms ranked by term frequency among the 3 identified topics (N=6437 tweets): (A) topic 1: flavors and flavor ban (n=3334, 51.8%); (B) topic 2: co-use discourse (n=1119, 17.4%); and (C) topic 3: availability and access regulation (n=1984, 30.8%). * Indicates a term that was in the search term list for tweet selection (see Table S1 in Multimedia Appendix 1 for the full list).

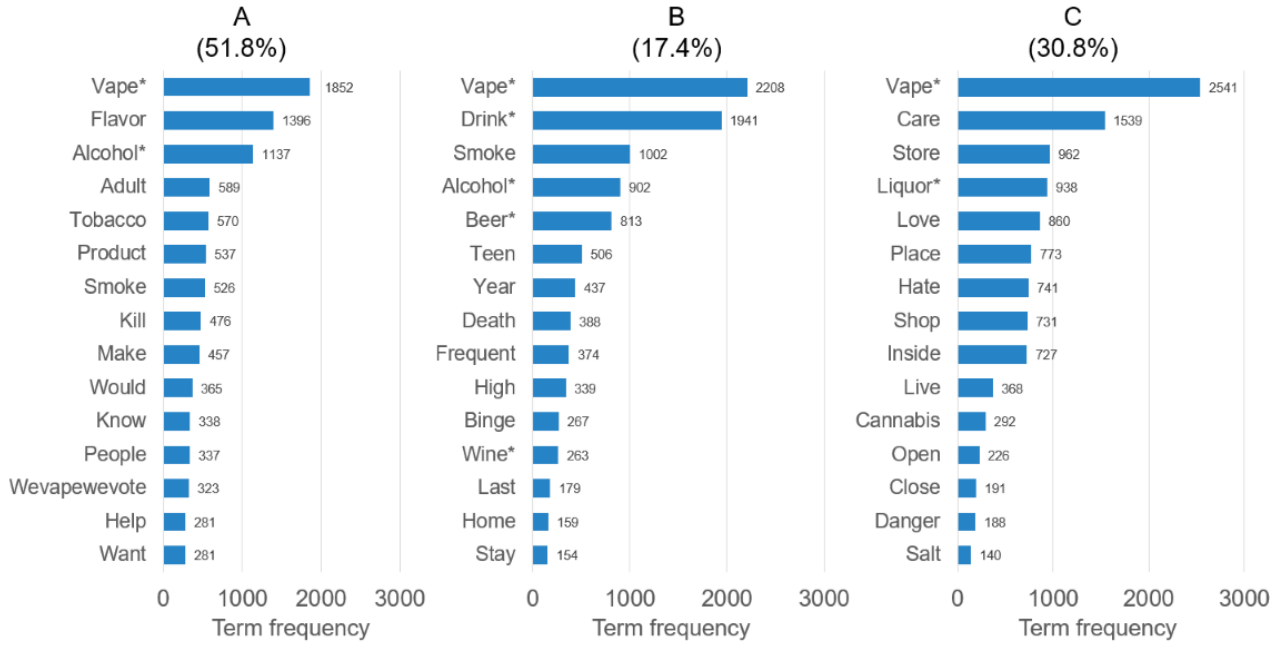

Within the flavors and flavor ban topic (topic 1), the most relevant and frequent terms were vape, flavor, alcohol, adult, tobacco, and product (Figure 1). Tweets tended to include arguments focused on how banning vaping flavors was incongruent with alcohol regulation where flavors are allowed and common. Tweets also often discussed the benefits of flavors for adults who made the switch from combustible tobacco products or espoused the opinion that e-cigarette use was less

XSL•FO
RenderX

detrimental to youth's health and safety than alcohol. This topic had the lowest proportion of tweets from individuals ≥21 years (76% vs 82% in other groups). Mock tweets representing common themes and tone within this group include:

> *We choose which flavor of liquor or candy to buy, why can't adults have options when it comes to vaping flavors?*
>
> *Banning vape flavors because kids use them makes no sense. Especially when alcohol is still available in flavors like bubble gum and vanilla.*
>
> *Alcohol kills more youth a year than vaping ever has or will.*

For the co-use topic (topic 2; n=1119, 17.4%), the most frequent words were vape, drink, smoke, alcohol, beer, and teen. Tweets often discussed plans or experiences related to using e-cigarettes and alcohol together in a positive light. This topic included tweets mentioning regulation, but these were less common compared with the other groups. Tweets within this group had a more even gender distribution (51.3% women vs approximately 35% in the other groups; Table 1). They were also more likely (compared to topic 1) to include individuals predicted to be older than 21 years of age (n=610, 82.4%). This topic also had the highest prevalence of both negative (18.3% vs roughly 10% in the other groups) and positive (10% vs roughly 8% in the other groups) sentiment, suggesting more polarization within this topic. Mock tweet examples include:

> *Listening to the Lumineers with my vape in one hand and a glass of wine in the other is my happy place*
>
> *Just saw someone vaping in a non-smoking area, while throwing back a beer. What a flex!*
>
> *I love vaping when I am drunk*

The availability and access regulations topic (topic 3) represented roughly a third of tweets (n=1984, 30.8%). The most relevant and frequent terms were vape, care, store, liquor, love, and place. Tweets focused on regulation (similar to topic 1) but tended to include discussions of whether and how regulation restricts access. Some tweets supported regulations restricting youth access (ie, Tobacco 21) but expressed concern or anger that stricter regulations would limit access among legal-age adults. Tweets often mentioned age verification or limiting locations of sale as strategies rather than stricter regulations seen as moving toward "prohibition." Alcohol regulations were held up as a comparison—either as an example of how complete bans of certain vape products may not work or to acknowledge the need to strengthen or enforce age verification processes and limit sale locations similar to alcohol. Statistics surrounding alcohol use and harm were also used for comparison, as were regulations and use prevalence of other substances, including cannabis.

> *Banning vapor products will not work, just like alcohol prohibition did not work.*
>
> *Vapes should never have been available at gas stations. Allow products to be sold in age verified stores so adults have access while restricting access to kids—just like alcohol.*

Topic Prevalence Before and After Federal Tobacco Policy Changes

We next examined the distribution of topics by when in the regulatory change period they were posted. While the flavor ban topic (topic 1) represented over half of the tweets across the study period (n=3334, 51.8%, Table 2), flavor-related tweets were more common in the pre–regulatory change period prior to the changes to federal policy at the start of 2020 (Tobacco 21 and ban on flavored cartridge-based e-cigarettes) than in the post–regulatory change period (860/1249, 68.8% of pre–regulatory change period tweets compared to 2474/5188, 47.6% of post–regulatory change period tweets). Co-use discussions were more common in the post–regulatory change period (n=11, <1% of tweets prior to policy change vs n=1108, 21.4% of all tweets in the post–regulatory change period), while availability and access discussions (topic 3) were similar in prevalence across the 2 time periods (roughly 30%). Variations in topic prevalence across time were statistically significant across the 3 topics (P<.001).

XSL•FO
RenderX

Table 2. Variation in the prevalence of vaping-alcohol tweets during policy shifts and pandemic declaration overall and by topic.

Time period	Topic 1: flavors and flavor ban (n=3334), n (%)	Topic 2: co-use discourse (n=1119), n (%)	Topic 3: availability and access regulation (n=1984), n (%)	*P* value[a]
Overall (N=6437)	3334 (51.8)	1119 (17.4)	1984 (30.8)	__[b]
Federal policy change status[c]				<.001
Pre–regulatory change period	860 (68.8)	11 (0.9)	378 (30.3)	
Post–regulatory change period	2474 (47.6)	1108 (21.4)	1606 (31)	
COVID-19 pandemic declaration[d]				<.001
Pre–COVID-19 pandemic declaration period	1628 (61.1)	273 (10.3)	763 (28.6)	
Post–COVID-19 pandemic declaration period	1706 (45.2)	846 (22.4)	1221 (32.4)	

[a]Resulting from chi-square test for difference in proportions by time period across topic groups.

[b]Not applicable.

[c]Tweet was posted on January 1, 2020, or after.

[d]Tweet was posted on March 1, 2020, or after (World Health Organization pandemic declaration was on March 13, 2020).

Topic Prevalence Before and After the Pandemic Declaration

Finally, variations in topic prevalence before and after the pandemic declaration were examined. Topic prevalence varied significantly in the pre– and post–COVID-19 pandemic declaration periods (*P*<.001). The availability and access topic (topic 3) was slightly more common in the post–COVID-19 pandemic period compared to the pre-period (n=1221, 32.4% vs n=763, 28.6%, respectively) as was the co-use topic (topic 2; n=846, 22.4% vs n=273, 10.3%, respectively).

Discussion

Principal Findings

In a sample of tweets mentioning e-cigarettes and alcohol, we identified 3 topics. Most tweets fell into 2 topics focused on vaping regulation while drawing comparisons to alcohol regulations. One additional topic including discussions of both alcohol and vaping (co-use) was also identified. Our findings demonstrated that while there were co-use discussions on Twitter, a large subset of comentions during this time focused on the regulation of these substances both before and after the implementation of federal policies related to tobacco. We also found that topic prevalence varied before and after key regulatory events as well as in the postpandemic period. Specifically, co-use discussions were most common after the pandemic declaration, which may suggest increased stress and substance use during a public health emergency. In addition, while flavor and flavor ban–related discussions were common across the study, they were more prevalent in the time period leading up to Tobacco 21 and the ban on flavored cartridge-based e-cigarettes. This may have reflected increased discussion prior to implementation as a way to express support or concerns surrounding the regulations.

Comparison to Prior Work

These analyses add to our understanding of the social media discourse surrounding vaping regulation. Prior work examining Twitter discussion in response to the 2016 Deeming Rule by the Food and Drug Administration (FDA) on e-cigarettes found a rise in negative reactions after the announcement of the new regulations [35]. They also found that tweets often debated whether the rule would harm health or focused on how it would impact the new, emerging e-cigarette market [35]. Another more recent analysis that applied LDA topic modeling before and after the FDA's flavor enforcement policy found similarly negative sentiment after implementation and discussions surrounding purchasing and accessing products after regulatory implementation [63]. Similar to these prior studies, tweets within the current analyses, which were collected during a period of changing policies, often debated the value of regulating e-cigarettes. In particular, how regulations implemented to protect youth would limit adult access to flavors and make purchasing products more challenging. Public health officials and policy makers should be aware of regulatory discourse in order to understand current public knowledge and sentiment as well as to contribute accurate and timely information to address concerns surrounding regulatory shifts. This may be particularly important in situations where posts are generated or elevated by the industry either directly or indirectly (ie, astroturfing) [64].

A novel finding was the use of alcohol as a regulatory comparator when reflecting on vaping regulation. While some tweets acknowledged alcohol regulation was beyond the FDA's purview, the opinion that vaped nicotine products are more regulated than alcohol was consistently used to argue against vaping regulations. Additionally, use statistics among youth were commonly shared as evidentiary support. Specifically, some reported alcohol as a more common, harmful substance used by youth compared with e-cigarettes.

The prominence of access and regulatory commentary during this time period may have been in response to the new, evolving nature of regulations. One study analyzing tweets responding to an antivaping campaign found frequent objections to regulation as well as tweets touting e-cigarette health benefits [37]. In another study focused on marijuana use mentions—a

XSL•FO
RenderX

substance experiencing regulatory change—calls for legalization were common [65]. Recent regulatory changes may have been more salient after the pandemic declaration due to challenges with access from business-related closures during this time period.

The identified co-use topic demonstrates an opportunity for those on these platforms (including youth audiences) to be exposed to content, which may normalize joint use. Common terms defining this subtopic included broad alcohol terms such as “alcohol” and “beer” but also terms that may be associated with heavy drinking such as “frequent” and “binge.” Our search strategy of including both general alcohol terms as well as common excessive drinking terms in the United States such as “blacked out” and “trashed” may have allowed us to more readily identify co-use tweets that would have not been captured with more standard drinking terms. These findings emphasize the importance of understanding cultural terms and specificity when collecting and analyzing social media data.

Counteradvertising efforts may be particularly useful during times of crisis and change as pathways for sharing accurate information in a timely fashion [44]. The unique time period of data collection may have also affected co-use discussions. Interestingly, the co-use topic was almost entirely restricted to the postpolicy period and was most common during the pandemic period. While regulatory shifts may have brought substance use discussions to the forefront, the early stages of the pandemic likely influenced co-use discussions. Over three-quarters of tweets within this topic were more specifically from the period after the COVID-19 pandemic was declared in March 2020. A prior study found alcohol use mentions related to coping during COVID-19 and addiction concerns during this time period [36]. With individuals spending time at home and living their social lives in virtual spaces, individuals during this time period may have increased their likelihood of sharing experiences related to substance use and co-use in particular. As LDA is a data-driven process, it is possible that if we had run the LDA separately across time periods for the regulatory change period or the COVID-19 declaration period, different themes may have emerged. Future research could adopt this approach to explore additional themes that may be missed within a larger dataset.

Strengths and Limitations

This study has important limitations. First, analyses were limited to Twitter and to the text of tweets. Findings may not be generalizable to other social networks or more visual content (eg, videos and images), which are an important element of social media discourse. Second, although we included retweets, we did not explore key influencers that may drive content and were unable to examine the demographics of tweet viewers (ie, whether tweets reached underage youth). In addition, we did not distinguish between human and bot accounts. Furthermore, it is unclear whether users expressing frustration over regulations were industry funded or endorsed accounts. An analytic limitation is that LDA uses a “bag of words” method, which ignores word order and context. In addition, our focus on tweets mentioning both vape- and alcohol-related terms and the resulting smaller sample size likely limited the identification of additional topics. Finally, while we reviewed a subsample of tweets to ensure they included the relevant vape and alcohol search terms, we did not examine tweets for whether they were germane to vaping or alcohol use. It is therefore possible that although a term was used, the tweet was not focused on these topics.

This study also has several significant strengths. First, we used a data-driven approach to identify tweets mentioning both vaping and alcohol. We made no a priori assumptions about the number of topics or the thematic groupings and were able to quantify salient and frequent terms mentioned within identified themes. Additionally, we used state-of-the-art machine learning algorithms to predict account user demographics. Finally, we analyzed tweets from a time period, which included many changes in the United States, shedding light on how policy change and major national emergencies may shift social media discourse surrounding substance use.

Conclusions

In conclusion, our study identified 3 distinctive topics across tweets including both vape- and alcohol-related terms. Specifically, a discussion of flavors and pending flavor bans was the largest topic subgroup, followed by a discussion of potential restrictions to e-cigarette availability and adult access resulting from regulatory changes, and finally, a portion of tweets included co-use of alcohol and vape products. While there were co-use mentions, tweets more often mentioned alcohol to make comparisons regarding how the substances were differentially regulated. Our findings demonstrate how the still-evolving landscape of electronic nicotine product policies influences social media discussion and debate. It also signals an opportunity to better understand and prepare for common concerns surrounding proposed regulatory changes—particularly as it relates to limiting youth access. A deeper understanding of social media discussions may inform health communication campaigns and policies focused on restricting underage youth’s access to substances and advancing youth’s health.

Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the Food and Drug Administration, or the American Heart Association. LRR, JW, JLF, AB, and RMR were supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (award U54HL120163). JLF is also supported by the National Heart, Lung, and Blood Institute (K01 HL143142). This work was partially supported by the Evans Center for Interdisciplinary Biomedical Research Affinity Research Collaborative on Tobacco Regulatory Science at Boston University.

XSL•FO
RenderX

Data Availability

Analyses use publicly available posts from Twitter. The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LRR conceptualized the study, drafted the initial manuscript, critically reviewed and revised the manuscript, and prepared the final manuscript draft. DAT and ML conducted the data analyses. JW, AB, and RMR assisted with the interpretation of results and critically reviewed and revised the manuscript. DW, TH, and JLF contributed to the design of the study, oversaw data collection, assisted with the interpretation of results, and critically reviewed and revised the manuscript. ZX contributed to the design of the study, assisted with the interpretation of results, and critically reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methods details, alcohol and vaping terms list, and list of the top 15 most salient terms across the data set.
[DOCX File , 79 KB - jmir_v26i1e51870_app1.docx]

References

1. Creamer MR, Everett Jones S, Gentzke AS, Jamal A, King BA. Tobacco product use among high school students—Youth Risk Behavior Survey, United States, 2019. MMWR Suppl 2020 Aug 21;69(1):56-63 [FREE Full text] [doi: 10.15585/mmwr.su6901a7] [Medline: 32817607]
2. Park-Lee E, Ren C, Cooper M, Cornelius M, Jamal A, Cullen KA. Tobacco product use among middle and high school students—United States, 2022. MMWR Morb Mortal Wkly Rep 2022 Nov 11;71(45):1429-1435 [FREE Full text] [doi: 10.15585/mmwr.mm7145a1] [Medline: 36355596]
3. Rothrock AN, Andris H, Swetland SB, Chavez V, Isaak S, Pagane M, et al. Association of e-cigarettes with adolescent alcohol use and binge drinking-drunkenness: a systematic review and meta-analysis. Am J Drug Alcohol Abuse 2020 Nov 01;46(6):684-698. [doi: 10.1080/00952990.2020.1771723] [Medline: 32795246]
4. Thrul J, Gubner NR, Tice CL, Lisha NE, Ling PM. Young adults report increased pleasure from using e-cigarettes and smoking tobacco cigarettes when drinking alcohol. Addict Behav 2019 Jun;93:135-140 [FREE Full text] [doi: 10.1016/j.addbeh.2019.01.011] [Medline: 30710807]
5. Sagheddu C, Scherma M, Congiu M, Fadda P, Carta G, Banni S, et al. Inhibition of N-acylethanolamine acid amidase reduces nicotine-induced dopamine activation and reward. Neuropharmacology 2019 Jan;144:327-336. [doi: 10.1016/j.neuropharm.2018.11.013] [Medline: 30439418]
6. Yang JJ, Lin HC, Ou TS, Tong Z, Li R, Piper ME, et al. The situational contexts and subjective effects of co-use of electronic cigarettes and alcohol among college students: an ecological momentary assessment (EMA) study. Drug Alcohol Depend 2022 Oct 01;239:109594 [FREE Full text] [doi: 10.1016/j.drugalcdep.2022.109594] [Medline: 35988530]
7. Connor JP, Gullo MJ, White A, Kelly AB. Polysubstance use: diagnostic challenges, patterns of use and health. Curr Opin Psychiatry 2014 Jul;27(4):269-275. [doi: 10.1097/YCO.0000000000000069] [Medline: 24852056]
8. Zuckermann AME, Williams GC, Battista K, Jiang Y, de Groh M, Leatherdale ST. Prevalence and correlates of youth poly-substance use in the COMPASS study. Addict Behav 2020 Aug;107:106400. [doi: 10.1016/j.addbeh.2020.106400] [Medline: 32222564]
9. Kelly AB, Evans-Whipp TJ, Smith R, Chan GCK, Toumbourou JW, Patton GC, et al. A longitudinal study of the association of adolescent polydrug use, alcohol use and high school non-completion. Addiction 2015 Apr;110(4):627-635 [FREE Full text] [doi: 10.1111/add.12829] [Medline: 25510264]
10. Silveira ML, Green VR, Iannaccone R, Kimmel HL, Conway KP. Patterns and correlates of polysubstance use among US youth aged 15-17 years: wave 1 of the Population Assessment of Tobacco and Health (PATH) Study. Addiction 2019 May;114(5):907-916 [FREE Full text] [doi: 10.1111/add.14547] [Medline: 30614093]
11. Richter L, Pugh BS, Smith PH, Ball SA. The co-occurrence of nicotine and other substance use and addiction among youth and adults in the United States: implications for research, practice, and policy. Am J Drug Alcohol Abuse 2017 Mar;43(2):132-145. [doi: 10.1080/00952990.2016.1193511] [Medline: 27494436]
12. Czeisler ME, Board A, Thierry JM, Czeisler CA, Rajaratnam SMW, Howard ME, et al. Mental health and substance use among adults with disabilities during the COVID-19 pandemic—United States, February-March 2021. MMWR Morb Mortal Wkly Rep 2021 Aug 27;70(34):1142-1149 [FREE Full text] [doi: 10.15585/mmwr.mm7034a3] [Medline: 34437518]
13. Addiction and health. National Institute on Drug Abuse. URL: https://nida.nih.gov/publications/drugs-brains-behavior-science-addiction/addiction-health [accessed 2024-05-08]

XSL•FO
RenderX

14. Lai HMX, Cleary M, Sitharthan T, Hunt GE. Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990-2014: a systematic review and meta-analysis. Drug Alcohol Depend 2015 Sep 01;154:1-13. [doi: 10.1016/j.drugalcdep.2015.05.031] [Medline: 26072219]
15. e-Cigarette use among youth and young adults: a report of the Surgeon General. US Department of Health and Human Services. 2016. URL: https://www.cdc.gov/tobacco/sgr/e-cigarettes/pdfs/2016_sgr_entire_report_508.pdf [accessed 2024-09-11]
16. Surgeon General's Advisory on e-cigarette use among youth: the e-cigarette epidemic among youth. US Department of Health and Human Services. Office of the Surgeon General. 2018. URL: https://health.gov/healthypeople/tools-action/browse-evidence-based-resources/surgeon-generals-advisory-e-cigarette-use-among-youth [accessed 2024-09-11]
17. Cleary M, Thomas SP. Addiction and mental health across the lifespan: an overview of some contemporary issues. Issues Ment Health Nurs 2017 Jan;38(1):2-8. [doi: 10.1080/01612840.2016.1259336] [Medline: 27936330]
18. Anderson M, Jiang J. Teens, social media and technology. Pew Research Center. 2018. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2022-05-10]
19. Clendennen SL, Loukas A, Vandewater EA, Perry CL, Wilkinson AV. Exposure and engagement with tobacco-related social media and associations with subsequent tobacco use among young adults: a longitudinal analysis. Drug Alcohol Depend 2020 Aug 01;213:108072 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108072] [Medline: 32563115]
20. Soneji S, Yang J, Knutzen KE, Moran MB, Tan ASL, Sargent J, et al. Online tobacco marketing and subsequent tobacco Use. Pediatrics 2018;141(2):e20172927 [FREE Full text] [doi: 10.1542/peds.2017-2927] [Medline: 29295893]
21. Vogel EA, Guillory J, Ling PM. Sponsorship disclosures and perceptions of e-cigarette Instagram posts. Tob Regul Sci 2020 Sep;6(5):355-368 [FREE Full text] [doi: 10.18001/trs.6.5.5] [Medline: 33778107]
22. Cabrera-Nguyen EP, Cavazos-Rehg P, Krauss M, Bierut LJ, Moreno MA. Young adults' exposure to alcohol- and marijuana-related content on Twitter. J Stud Alcohol Drugs 2016 Mar;77(2):349-353 [FREE Full text] [doi: 10.15288/jsad.2016.77.349] [Medline: 26997194]
23. Barry AE, Bates AM, Olusanya O, Vinal CE, Martin E, Peoples JE, et al. Alcohol marketing on Twitter and Instagram: evidence of directly advertising to youth/adolescents. Alcohol Alcohol 2016 Jul;51(4):487-492 [FREE Full text] [doi: 10.1093/alcalc/agv128] [Medline: 26597794]
24. Primack BA, Colditz JB, Rosen EB, Giles LM, Jackson KM, Kraemer KL. Portrayal of alcohol brands popular among underage youth on YouTube: a content analysis. J Stud Alcohol Drugs 2017 Sep;78(5):654-664. [doi: 10.15288/jsad.2017.78.654] [Medline: 28930052]
25. Gentzke AS, Wang TW, Cornelius M, Park-Lee E, Ren C, Sawdey MD, et al. Tobacco product use and associated factors among middle and high school students—National Youth Tobacco Survey, United States, 2021. MMWR Surveill Summ 2022 Mar 11;71(5):1-29 [FREE Full text] [doi: 10.15585/mmwr.ss7105a1] [Medline: 35271557]
26. Bandura A. Social Learning Theory. Englewood Cliffs, NJ: Prentice Hall; 1977.
27. Davis JP, Pedersen ER, Tucker JS, Dunbar MS, Seelam R, Shih R, et al. Long-term associations between substance use-related media exposure, descriptive norms, and alcohol use from adolescence to young adulthood. J Youth Adolesc 2019 Jul;48(7):1311-1326 [FREE Full text] [doi: 10.1007/s10964-019-01024-z] [Medline: 31025156]
28. Elmore KC, Scull TM, Kupersmidt JB. Media as a "Super Peer": how adolescents interpret media messages predicts their perception of alcohol and tobacco use norms. J Youth Adolesc 2017 Feb;46(2):376-387 [FREE Full text] [doi: 10.1007/s10964-016-0609-9] [Medline: 27837371]
29. Ranker LR, Wu J, Hong T, Wijaya D, Benjamin EJ, Bhatnagar A, et al. Social media use, brand engagement, and tobacco product initiation among youth: evidence from a prospective cohort study. Addict Behav 2024 Jul;154:108000. [doi: 10.1016/j.addbeh.2024.108000] [Medline: 38461745]
30. Moreno MA, Cox ED, Young HN, Haaland W. Underage college students' alcohol displays on Facebook and real-time alcohol behaviors. J Adolesc Health 2015 Jun;56(6):646-651 [FREE Full text] [doi: 10.1016/j.jadohealth.2015.02.020] [Medline: 26003580]
31. Yang B, Zhao X. TV, social media, and college students' binge drinking intentions: moderated mediation models. J Health Commun 2018;23(1):61-71. [doi: 10.1080/10810730.2017.1411995] [Medline: 29265924]
32. Twitter Inc. Twitter announces first quarter 2022 results. Recent Financial Releases. 2022. URL: https://www.prnewswire.com/news-releases/twitter-announces-first-quarter-2022-results-301535206.html [accessed 2022-05-10]
33. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2021-04-07]
34. Wu J, Wang Y, Xu YA, Fetterman JL, Hong T. Morally driven and emotionally fueled: the interactive effects of values and emotions in the social transmission of information endorsing e-cigarettes. Int J Commun 2023;17(2023):1190-1210 [FREE Full text]
35. Lazard AJ, Wilcox GB, Tuttle HM, Glowacki EM, Pikowski J. Public reactions to e-cigarette regulations on Twitter: a text mining analysis. Tob Control 2017 Dec;26(e2):e112-e116. [doi: 10.1136/tobaccocontrol-2016-053295] [Medline: 28341768]
36. Glowacki EM, Wilcox GB, Glowacki JB. Identifying #addiction concerns on Twitter during the COVID-19 pandemic: a text mining analysis. Subst Abus 2021;42(1):39-46. [doi: 10.1080/08897077.2020.1822489] [Medline: 32970973]

XSL•FO
RenderX

37. Allem JP, Escobedo P, Chu KH, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on Twitter to e-cigarette education. Tob Control 2017 Mar;26(2):226-229 [FREE Full text] [doi: 10.1136/tobaccocontrol-2015-052757] [Medline: 26956467]
38. Cornacchione Ross J, Lazard AJ, Hedrick McKenzie A, Reffner Collins MK, Sutfin EL. What cigarillo companies are putting on Instagram: a content analysis of Swisher Sweets' marketing from 2013 to 2020. Nicotine Tob Res 2023 Mar 22;25(4):755-762. [doi: 10.1093/ntr/ntac255] [Medline: 36326790]
39. Irizar P, Puddephatt JA, Warren JG, Field M, Jones A, Rose AK, et al. "Drinkers Like Me": a thematic analysis of comments responding to an online article about moderating alcohol consumption. Front Psychol 2022;13:780677 [FREE Full text] [doi: 10.3389/fpsyg.2022.780677] [Medline: 35360610]
40. Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, et al. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res 2015 Oct 27;17(10):e243 [FREE Full text] [doi: 10.2196/jmir.4969] [Medline: 26508089]
41. Ezike NC, Ames Boykin A, Dobbs PD, Mai H, Primack BA. Exploring factors that predict marketing of e-cigarette products on twitter: infodemiology approach using time series. JMIR Infodemiology 2022;2(2):e37412 [FREE Full text] [doi: 10.2196/37412] [Medline: 37113447]
42. Liang Y, Zheng X, Zeng DD, Zhou X, Leischow SJ, Chung W. Exploring how the tobacco industry presents and promotes itself in social media. J Med Internet Res 2015 Jan 21;17(1):e24 [FREE Full text] [doi: 10.2196/jmir.3665] [Medline: 25608524]
43. Vassey J, Metayer C, Kennedy CJ, Whitehead TP. #Vape: measuring e-cigarette influence on Instagram with deep learning and text analysis. Front Commun (Lausanne) 2020;4:75 [FREE Full text] [doi: 10.3389/fcomm.2019.00075] [Medline: 35233388]
44. Jordan SE, Hovet SE, Fung ICH, Liang H, Fu KW, Tse ZTH. Using Twitter for public health surveillance from monitoring and prediction to public response. Data 2019;4(1):6. [doi: 10.3390/data4010006]
45. Fu R, Kundu A, Mitsakakis N, Elton-Marshall T, Wang W, Hill S, et al. Machine learning applications in tobacco research: a scoping review. Tob Control 2023 Jan;32(1):99-109. [doi: 10.1136/tobaccocontrol-2020-056438] [Medline: 34452986]
46. Russell AM, Colditz JB, Barry AE, Davis RE, Shields S, Ortega JM, et al. Analyzing Twitter chatter about tobacco use within intoxication-related contexts of alcohol use: "Can someone tell me why nicotine is so fire when you're drunk?". Nicotine Tob Res 2022 Jul 13;24(8):1193-1200 [FREE Full text] [doi: 10.1093/ntr/ntab195] [Medline: 34562100]
47. Krauss MJ, Grucza RA, Bierut LJ, Cavazos-Rehg PA. "Get drunk. Smoke weed. Have fun.": A content analysis of tweets about marijuana and alcohol. Am J Health Promot 2017 May;31(3):200-208 [FREE Full text] [doi: 10.4278/ajhp.150205-QUAL-708] [Medline: 26559715]
48. Newly signed legislation raises federal minimum age of sale of tobacco products to 21. Food and Drug Administration. 2020. URL: https://public4.pagefreezer.com/content/FDA/23-11-2021T07:28/https:/www.fda.gov/tobacco-products/ctp-newsroom/newly-signed-legislation-raises-federal-minimum-age-sale-tobacco-products-21 [accessed 2023-02-14]
49. FDA finalizes enforcement policy on unauthorized flavored cartridge-based e-cigarettes that appeal to children, including fruit and mint. FDA. 2020. URL: https://www.fda.gov/news-events/press-announcements/fda-finalizes-enforcement-policy-unauthorized-flavored-cartridge-based-e-cigarettes-appeal-children [accessed 2023-02-14]
50. McCausland K, Maycock B, Leaver T, Wolf K, Freeman B, Jancey J. e-Cigarette advocates on Twitter: content analysis of vaping-related tweets. JMIR Public Health Surveill 2020 Oct 14;6(4):e17543 [FREE Full text] [doi: 10.2196/17543] [Medline: 33052130]
51. Cavazos-Rehg PA, Krauss MJ, Sowles SJ, Bierut LJ. "Hey Everyone, I'm Drunk." An evaluation of drinking-related Twitter chatter. J Stud Alcohol Drugs 2015 Jul;76(4):635-643 [FREE Full text] [doi: 10.15288/jsad.2015.76.635] [Medline: 26098041]
52. West JH, Hall PC, Hanson CL, Prier K, Giraud-Carrier C, Neeley ES, et al. Temporal variability of problem drinking on Twitter. OJPM 2012;02(01):43-48. [doi: 10.4236/ojpm.2012.21007]
53. Hoffman MD, Blei DM, Bach F. Online Learning for Latent Dirichlet Allocation. 2010 Dec 06 Presented at: Proceedings of the 23rd International Conference on Neural Information Processing Systems—Volume 1; 2010; Vancouver, British Columbia, Canada p. 856-864. [doi: 10.5555/2997189.2997285]
54. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text]
55. LDA model. Gensim. 2022. URL: https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html#sphx-glr-auto-examples-tutorials-run-lda-py [accessed 2022-10-14]
56. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. 2015 Presented at: WSDM 2015 Proceedings of the 8th ACM International Conference on Web Search and Data Mining Association for Computing Machinery; February 2, 2018; New York, NY, United States p. 399-408. [doi: 10.1145/2684822.2685324]
57. Sievert C, Shirley KE. LDAvis: a method for visualizing and interpreting topics. 2014 Presented at: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces; September 3, 2024; Baltimore, MD, United States p. 63-70. [doi: 10.3115/v1/w14-3110]
58. Wang Z, Hale SA, Adelani D, Grabowicz PA, Hartmann T, Flöck F, et al. Demographic inference and representative population estimates from multilingual social media data. 2019 Presented at: The Web Conference 2019—Proceedings of

XSL•FO
RenderX

the World Wide Web Conference, WWW 2019 Association for Computing Machinery, Inc; May 15, 2019; New York, NY, United States p. 2056-2067. [doi: 10.1145/3308558.3313684]
59. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Jun Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019; Minneapolis, MN, United States p. 4171-4186 URL: https://doi.org/10.48550/arXiv.1810.04805 [doi: 10.18653/v1/N19-1423]
60. Slade T, Chapman C, Swift W, Keyes K, Tonks Z, Teesson M. Birth cohort trends in the global epidemiology of alcohol use and alcohol-related harms in men and women: systematic review and metaregression. BMJ Open 2016 Oct 24;6(10):e011827 [FREE Full text] [doi: 10.1136/bmjopen-2016-011827] [Medline: 27797998]
61. Keyes KM, Miech R. Age, period, and cohort effects in heavy episodic drinking in the US from 1985 to 2009. Drug Alcohol Depend 2013 Sep 01;132(1-2):140-148 [FREE Full text] [doi: 10.1016/j.drugalcdep.2013.01.019] [Medline: 23433898]
62. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; June 1-4, 2014; Ann Arbor, MI, United States p. 216-225 URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 [doi: 10.1609/icwsm.v8i1.14550]
63. Lu X, Sun L, Xie Z, Li D. Perception of the Food and Drug Administration electronic cigarette flavor enforcement policy on Twitter: observational study. JMIR Public Health Surveill 2022 Mar 29;8(3):e25697 [FREE Full text] [doi: 10.2196/25697] [Medline: 35348461]
64. University of Bath. Astroturfing. Tobacco Tactics. 2022. URL: https://tobaccotactics.org/wiki/astroturfing/ [accessed 2022-05-31]
65. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza RA, Bierut LJ. Twitter chatter about marijuana. J Adolesc Health 2015 Feb;56(2):139-145 [FREE Full text] [doi: 10.1016/j.jadohealth.2014.10.270] [Medline: 25620299]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
FDA: Food and Drug Administration
LDA: latent Dirichlet allocation

Edited by A Mavragani; submitted 15.08.23; peer-reviewed by F Geusens, G Humphreys; comments to author 24.02.24; revised version received 30.05.24; accepted 26.06.24; published 12.11.24.

Please cite as:
Ranker LR, Tofu DA, Lu M, Wu J, Bhatnagar A, Robertson RM, Wijaya D, Hong T, Fetterman JL, Xuan Z
Concurrent Mentions of Vaping and Alcohol on Twitter: Latent Dirichlet Analysis
J Med Internet Res 2024;26:e51870
URL: https://www.jmir.org/2024/1/e51870
doi:10.2196/51870
PMID:

©Lynsie R Ranker, David Assefa Tofu, Manyuan Lu, Jiaxi Wu, Aruni Bhatnagar, Rose Marie Robertson, Derry Wijaya, Traci Hong, Jessica L Fetterman, Ziming Xuan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Wellness Influencer Responses to COVID-19 Vaccines on Social Media: A Longitudinal Observational Study

Gabrielle O'Brien[1], PhD; Ronith Ganjigunta[2]; Paramveer S Dhillon[1], PhD

[1]School of Information, University of Michigan, Ann Arbor, MI, United States

[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Gabrielle O'Brien, PhD
School of Information
University of Michigan
105 S. State Street
Ann Arbor, MI, 48103
United States
Phone: 1 (734) 764 1555
Email: elleobri@umich.edu

Abstract

Background: Online wellness influencers (individuals dispensing unregulated health and wellness advice over social media) may have incentives to oppose traditional medical authorities. Their messaging may decrease the overall effectiveness of public health campaigns during global health crises like the COVID-19 pandemic.

Objective: This study aimed to probe how wellness influencers respond to a public health campaign; we examined how a sample of wellness influencers on Twitter (rebranded as X in 2023) identified before the COVID-19 pandemic on Twitter took stances on the COVID-19 vaccine during 2020-2022. We evaluated the prevalence of provaccination messaging among wellness influencers compared with a control group, as well as the rhetorical strategies these influencers used when supporting or opposing vaccination.

Methods: Following a longitudinal design, wellness influencer accounts were identified on Twitter from a random sample of tweets posted in 2019. Accounts were identified using a combination of topic modeling and hand-annotation for adherence to influencer criteria. Their tweets from 2020-2022 containing vaccine keywords were collected and labeled as pro- or antivaccination stances using a language model. We compared their stances to a control group of noninfluencer accounts that discussed similar health topics before the pandemic using a generalized linear model with mixed effects and a nearest-neighbors classifier. We also used topic modeling to locate key themes in influencer's pro- and antivaccine messages.

Results: Wellness influencers (n=161) had lower rates of provaccination stances in their on-topic tweets (20%, 614/3045) compared with controls (n=242 accounts, with 42% or 3201/7584 provaccination tweets). Using a generalized linear model of tweet stance with mixed effects to model tweets from the same account, the main effect of the group was significant (β_1=–2.2668, SE=0.2940; P<.001). Covariate analysis suggests an association between antivaccination tweets and accounts representing individuals (β=–0.9591, SE=0.2917; P=.001) but not social network position. A complementary modeling exercise of stance within user accounts showed a significant difference in the proportion of antivaccination users by group (χ^2_1[N=321]=36.1, P<.001). While nearly half of the influencer accounts were labeled by a K-nearest neighbor classifier as predominantly antivaccination (48%, 58/120), only 16% of control accounts were labeled this way (33/201). Topic modeling of influencer tweets showed that the most prevalent antivaccination themes were protecting children, guarding against government overreach, and the corruption of the pharmaceutical industry. Provaccination messaging tended to encourage followers to take action or emphasize the efficacy of the vaccine.

Conclusions: Wellness influencers showed higher rates of vaccine opposition compared with other accounts that participated in health discourse before the pandemic. This pattern supports the theory that unregulated wellness influencers have incentives to resist messaging from establishment authorities such as public health agencies.

(J Med Internet Res 2024;26:e56651) doi:10.2196/56651

XSL•FO
RenderX

KEYWORDS

social media, COVID-19, vaccination; personal brands; public health; wellness; global health; pandemic; Twitter; tweets; vaccine; longitudinal design; wellness influencers; hand-annotation; anti-vaccination; infodemiology

Introduction

Social media has profoundly transformed the landscape of health information sharing and dissemination. Platforms such as Facebook, Twitter (rebranded as X in 2023), and Reddit have become crucial venues for health discussions [1-6], where individuals navigating medical decision-making may find knowledge and community support [7-14]. At the same time, media platforms may provide opportunities for unverified and potentially dangerous medical misinformation to reach wider audiences [15-22].

Social media platforms democratize who gets to command an online audience, which presents both opportunities and liabilities for public understanding of science and medicine [23]. Scholars have noted the rise of a "wellness influencer" class, online personalities who dispense health and lifestyle advice as part of their personal brand [24-28]. For these influencers, a lack of professional credentials (such as medical or scientific training) is often a source of authority: by eschewing ties to traditional institutions like universities or public health agencies, influencers can position themselves as relatable, authentic, and uncorrupted. This rhetorical strategy may be especially effective when trust in longstanding institutions is low [24].

Indeed, there is converging evidence that in some countries (such as the United States), trust in science has diverged from trust in scientific institutions. Surveys indicate that while confidence in "science" remains strong, confidence in the professional and governmental organizations that traditionally represent scientific and medical authority has waned [29-31]. This has prompted sociological examinations of who gets to "speak for science" [27].

If social media influencers rush to fill the void left by skeptical attitudes toward conventional institutions, this may have important ramifications for public science communication strategies. While considerable focus has been directed toward online misinformation research in recent years, there are still important gaps in the literature. For example, while there are many case studies of wellness influencers contradicting the guidelines of medical and public health authorities [26,32], it is more difficult to measure the prevalence and persuasiveness of this content [33,34]. On the other hand, while there are numerous studies leveraging computational social science techniques to detect patterns in health information sharing in large volumes of social media data [3,35-41], it is still challenging to connect this work to sociological theories of who claims to speak authoritatively on health and why.

While it has been reported that a handful of influential Twitter accounts (some that could be reasonably described as wellness influencers) shared a disproportionate volume of antivaccine messages [42] during the pandemic, this type of measure does not directly establish the prevalence of those attitudes among wellness influencers generally. It remains possible that most wellness influencers did not oppose vaccination, and their salience in misinformation research is due to a small but active minority. If this is true, we will need to refine theories of how the authority of science is negotiated online.

In this study, we aim to directly test whether being a wellness influencer online before a major public health crisis is associated with sharing anti-establishment opinions toward a public health intervention later [43]. Specifically, we use the COVID-19 pandemic and the ensuing vaccine rollout as a probe, as there is a sizable foundation of research tools available for analyzing this type of social media content—particularly on Twitter, where we locate our study [44-47]. In addition to the researcher tools available for studying antivaccine attitudes on social media, we leverage a variety of methods from natural language processing, a family of techniques for analyzing unstructured text, that has been previously applied to understanding online social interactions around health matters ranging from mental illness [6,8,12,37,48,49] to nutrition [5,43,50].

If being a wellness influencer truly requires opposing traditional health authorities as a core part of brand building, then we should expect influencers who were established before the pandemic to voice antagonistic opinions toward vaccines (which are necessarily created, tested, and distributed by scientific and medical institutions). Our study uses a longitudinal design to test the hypothesis that wellness influencers active before the pandemic were more likely to express antivaccination attitudes during the pandemic compared with a control group of accounts. We use a previously developed language model [51], trained on a high-quality dataset of labeled tweets, to classify opinions expressed toward or against the COVID-19 vaccine in thousands of tweets collected throughout the pandemic.

Notably, we hand-annotate user accounts to establish a representative "wellness influencer" sample, operationalized here as individuals who provide health and wellness advice without formal regulatory constraints. In addition to investigating the prevalence of anti-public health attitudes among our sample of influencers, we also analyze the key themes and rhetorical tactics used by wellness influencers. This combination of quantitative and qualitative analyses provides a new estimate for the prevalence of anti-establishment health beliefs among a clearly defined segment of users, as well as insight into their messaging strategies.

Methods

Overview

In our study, we first created a cohort of wellness influencers on Twitter (note that Twitter has been rebranded as X, but throughout the manuscript, we will refer to the platform as Twitter, as this was the name at the time of data collection). The cohort of influencers was selected according to definitions established by Baker and Rojek [25]. Specifically, these individuals are characterized by their online behavior providing health and wellness guidance directly to their followers.

XSL•FO
RenderX

Subsequently, we tracked these influencers' posts and perspectives on COVID-19 vaccination from the onset of the pandemic in 2020 until the end of 2022. This strategy is depicted in Figure 1. The source of our data was the Twitter Decahose, a 10% random sample of all tweets. Our analysis was restricted to English-language tweets only.

Figure 1. Diagram of the study design.

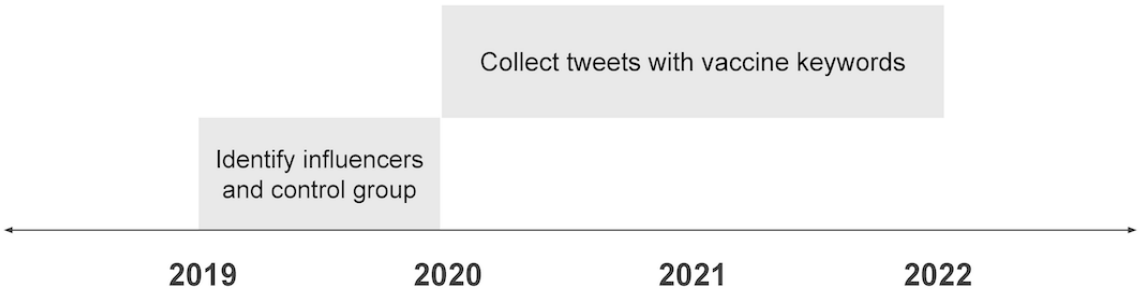

Identifying Influencer Accounts Before the Pandemic

Our process for identifying wellness influencers began by analyzing Twitter activity in 2019 (a year before the COVID-19 outbreak), focusing on accounts that significantly contributed to health and wellness conversations, as evidenced by high retweet counts.

Estimating the 2019 Retweet Network for Health Discourse

To construct a retweet network on the topics of health and wellness, we began with the task of identifying a keyword list. Using all available 2019 tweets in the Decahose, we trained a word embedding model [52] using the Python Word2Vec library [53]. We then identified the nearest 400 tokens in the embedding space to the seed "#wellness." We also identified the nearest 400 tokens to the nonhashtag form of the seed, "wellness." This gave us 400 hashtags and 400 nonhashtag keywords. Using our keyword list, we looked for accounts that had used keywords at least 50 times in the 2019 tweet corpus. Note that the threshold of 50 is somewhat arbitrary and conservative; by choosing a relatively high threshold, we are selecting for accounts with a markedly high volume of tweets with health keywords, but this may exclude accounts with clear health opinions that post less frequently. This stage yielded a sample of 2414 accounts.

Clustering and Annotating Influencer Accounts

To simplify labeling the 2414 accounts as influencers or not, we first used an unsupervised learning approach to cluster the accounts based on their 2019 tweets. After identifying relevant clusters, we hand-annotated the users in those clusters.

Clustering was done with the BERTopic library [54], treating each user's collected tweets as a single document. BERTopic uses a general-purpose, pretrained sentence embedding model [55] to map documents into a 768-dimensional space. Then, BERTopic applies a dimensionality reduction technique (Uniform Manifold Approximation and Projection [UMAP] [56]) and clusters the reduced embeddings using a hierarchical density-based clustering technique, hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [57]. HDBSCAN automatically selects a parsimonious number of clusters (in this case, 36). All users were assigned to their closest cluster. We use BERTopic's default models and hyperparameters for document embedding, dimensionality reduction, and clustering.

After documents have been clustered, BERTopic optionally allows users to fine-tune the keywords that are used to represent each cluster. This does not change the documents assigned to each cluster but may be useful to help researchers interpret the clusters. We used the KeyBert topic representation sub-model, as this approach has been suggested by the BERTopic authors to produce more interpretable topic clusters [54].

Following the manual inspection of the user clusters, exemplar tweets, and cluster keywords identified by BERTopic, we determined that 9 clusters dealt with wellness topics (other notable clusters represented news, politics, medical practice, and technology). There were more than 900 accounts in the selected clusters. We manually inspected these accounts for adherence to two criteria:

1. Individual accounts: The account needed to be that of an individual, such as a personal trainer, nutrition coach, book author, or entertainer, rather than an organization or company.
2. Wellness advice provision: The account was required to actively offer health and wellness tips to its followers, such as practical lifestyle suggestions or personal health strategies.

To judge if an account represented an individual, we supplemented the analysis of collected tweets with a manual scrape of Twitter bios and a collection of Twitter bios we had pulled from the API (application performing interface) before it became inaccessible (Twitter removed its free API, along with support for researcher API access, in February 2023; the Twitter API is required to programmatically access user bios, locations, follower and following counts, and full tweet history). Through this manual annotation process, we pinpointed 186 accounts that satisfied both prerequisites. A random sample of 10 such accounts, with example wellness advice, is presented in Table 1.

XSL•FO
RenderX

Table 1. Anonymized sample of 10 randomly selected accounts in the influencer group. Their Twitter bios have been paraphrased, and advice tweets have been lightly reworded to prevent reidentification.

Self-description from bio	Example advice
Spiritual guide	Add clay to smoothies to eliminate toxins from your body.
Tech executive	Drink 3 liters of water daily for a month for clearer skin.
Food blogger	Eat coconut flour for healthy digestion.
Lifestyle medicine consultant	Intermittent fasting will lower your cholesterol.
Alchemist and herbalist	For optimal health, have alkaline foods and high pH water.
Metabolic health coach	Be mindful about cutting carbohydrates from your diet for improved health.
Psychic	Improve brain health with nutritional supplement drinks.
Motivational coach	To relieve your anxiety, meditate.
Writer	Avoid junk food, as it cannot be digested by your body.
Pop culture fan	The best thing for your body and mind is a simple lifestyle.

Snowball Sampling

To increase our sample size, we also conducted a "snowball" sampling stage: using our curated list of wellness influencers, we identified any accounts they retweeted more than once in 2019 that also contained wellness keywords. We then manually annotated those accounts to ensure adherence to our influencer criteria, which expanded our influencer count to 264.

While snowball sampling is a commonplace strategy to expand a sample, it may exacerbate sampling biases in some ways. For example, wellness influencers may be especially likely to follow other influencers who share their demographic characteristics, topical interests, and beliefs, a well-known phenomenon called homophily [58,59]. Therefore, we caution that this second stage of sampling may increase sample size without necessarily improving our sample's representativeness of the broader population of wellness influencer accounts.

Considering Twitter Bots

It is well-established that Twitter contains "bots," automated accounts that may serve functions such as aggregating content, tweeting advertisements, or providing followers to paying clients [60]. Defining and measuring the prevalence of bot accounts is complex, as estimates can vary based on the measurement technique. However, one estimate from 2017 put the percentage between 9% and 15% [61], and another using Twitter data from 2022 estimated between 8% and 20%, depending on how inactive and deleted accounts were handled in the calculation [62].

Ideally, we would have liked to use a research tool like Botometer [63], a model for labeling probable bot accounts. However, due to the shuttering of the Twitter developer API to researchers, we were unable to obtain the API access required to use Botometer or similar services. Still, we can make some inferences about the role of bots in our sample. As our influencer selection process involved hand-labeling accounts that represent themselves as individuals, which relied on visual inspection of the accounts' bios (when available) and tweets, this group does not contain any accounts that appeared to be obviously bots, such as content aggregators or accounts that only post e-commerce links.

It is possible—and in the case of high-profile public figures, perhaps likely—that some influencer accounts use social media management software to schedule tweets with links to their blogs and websites. However, we consider this sort of messaging to be in the "voice" of the individual's personal brand and do not regard it as inherently problematic for identifying the individual's vaccination stance.

Analyzing Pandemic-Era Tweets From Influencers

Collecting Vaccine-Related Tweets

We extracted all tweets between January 1, 2020, and December 31, 2022, from the Decahose for our list of influencers. Retweets made by influencers were retained, as these provide potentially important information about an account holder's stance toward vaccination. We then filtered the collected tweets to only those containing a list of vaccine-related keywords and phrases. Our vaccine keyword list was created by combining lists from previously published studies of vaccine-related tweets [45,51]. Tweets were deduplicated before analysis, as some accounts reshared content on multiple days. The distribution of these tweets across the timeline is depicted in Figure 2.

XSL•FO
RenderX

Figure 2. The number of tweets containing a vaccine-related keyword per day for each group over the investigation period.

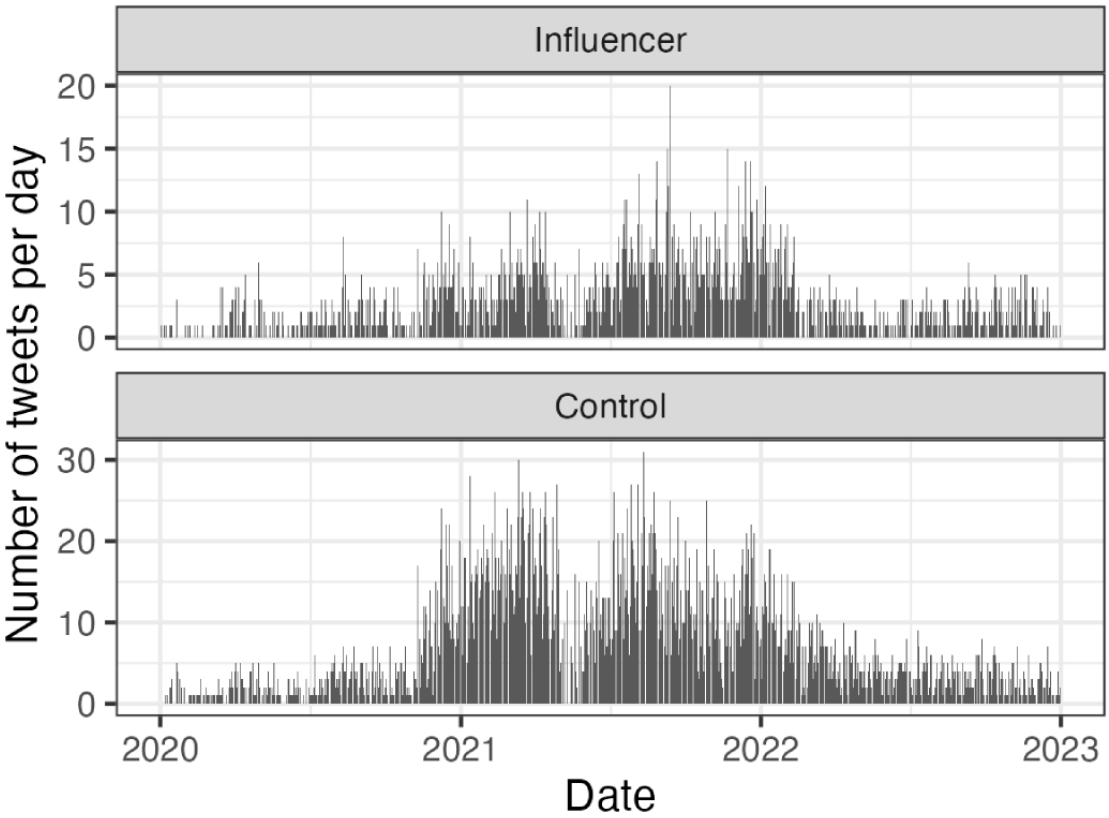

Tweet Stance Labeling With the VaxxHesitancy Model

To evaluate the stance expressed in these tweets regarding the COVID-19 vaccination, we used a pretrained transformer-based model. This model was fine-tuned on the VaxxHesitancy dataset—a curated, annotated collection of 3101 English-language tweets about COVID-19 vaccines, gathered from November 2020 to April 2022 [51]. The tweets in this dataset were categorized by human annotators into one of four categories reflecting their stance:

1. Provaccination: Posts supportive of COVID-19 vaccination.
2. Antivaccination: Posts opposing COVID-19 vaccination and seeking to convince others to do the same.
3. Hesitant: Posts expressing uncertainty or a wish to delay or refuse vaccination.
4. Irrelevant: Posts not explicitly stating a stance on COVID-19 vaccination.

The published VaxxHesitancy dataset does not include the text content of the collected tweets, as it was published with the expectation that the Twitter API would remain freely available to researchers to reconstruct the tweet text from a list of IDs. After the Twitter API changes, it was no longer possible for us to reconstruct the tweet text from the ID list. In response to this challenge, the authors of the VaxxHesitancy dataset graciously shared with us a sequence classification model binary trained on a test set of 2670 tweets and evaluated on a test set of the remaining 431 (the test set consisted of only tweets that were at least double annotated with interrater agreement, ensuring high confidence in their stance labels). The stance classification model is based on the VaxxBert transformer model, with fine-tuning for the stance labeling task. The VaxxHesitancy team benchmarked the model with an accuracy of 74.5% and an F_1-score of 70.5 on the test set for the 4-label classification task.

A preliminary analysis of our dataset revealed a notable pattern: tweets categorized as “hesitant” were usually from influencers who also produced “antivaccination” content. Despite these stances representing distinct categories at the tweet level, their frequent overlap within the same accounts indicated a shared behavioral pattern. Given that our original hypothesis did not distinguish between hesitant and antivaccination stances, and considering their common co-occurrence within the same users, we merged these categories for our analysis.

Tweet Topic Modeling

In our exploratory analysis of themes in influencers’ tweets, we again used the BERTopic library [54]. We separated tweets into two corpora: provaccination tweets by influencers and opposed tweets (comprised of hesitant and antivaccination tweets). For each corpus, we fit a BERTopic model (again, with the KeyBert topic representation submodel). Each tweet was treated as its own document for clustering.

Constructing A Control Group

Understanding the influence of wellness accounts on vaccination stances necessitates a benchmark for comparative analysis. Thus, we established a control group consisting of Twitter accounts that fall within similar topic domains as the influencer cohort but do not fulfill the criteria to be classified as influencers—those accounts either do not personify individuals

XSL•FO
RenderX

or do not dispense wellness advice in the tweets we analyzed. Examples of such accounts include medical professionals and scientists who refrain from giving health advice on Twitter, public health campaign initiatives, various media organizations that feature wellness segments, advocates for mental health awareness, and commentators on public health. This control group will help us delineate the specific impact of wellness influencers as compared to the broader wellness discourse on the platform.

Ethical Considerations for Studying Twitter Users

While some accounts in our sample correspond to public figures, others are unlikely to be widely known offline (for example, an account with followers numbering in the low thousands). Because we are studying individuals who take stances on a highly contentious public health issue, we expect these account holders to be at risk of harassment or other negative outcomes if they are identified in our reporting. Therefore, following the guidelines for social media researchers put forth by a committee at the Economic and Social Research Council at the University of Aberdeen [64], we present only anonymized quotes and account descriptions throughout our results.

Results

Influencer Stances Toward Vaccination

We began our analysis by examining overall patterns in vaccine-related tweets collected during the pandemic period (Table 2). Within the influencer group, we recovered 3045 relevant tweets from 161 accounts. Using the vaccine stance detection model, roughly 40% of these tweets were classified as containing no stance, roughly 20% were provaccination, and the remaining (~40%) tweets were labeled as hesitant or antivaccination. Meanwhile, for the control group, 7584 relevant tweets were collected from 242 accounts. The majority were labeled as not containing a stance (~50%) or provaccination (~40%), with only ~10% of tweets labeled hesitant or antivaccination. Overall, the proportion of stance-taking tweets was significantly higher in the influencer group (2-sample test for equality of proportions (χ^2_1[N=5616]=26.9, P<.001).

Table 2. Descriptive statistics of collected pandemic-era tweets.

	Influencers (n=161)	Control (n=242)
Tweets with vaccine keywords, n	3045	7584
Stance detected, n		
No stance	1315	3698
Provaccination	614	3201
Hesitant or antivaccination	1116	685

Since our study focuses on evaluating attitudes toward the vaccination public health initiative, specifically measuring the percentage of provaccination tweets among those that expressed a clear stance, thus, we excluded tweets categorized as "No stance." To account for the possibility of users contributing multiple tweets, we used a mixed-effects modeling approach to test the hypothesis that tweets from influencer accounts displayed more negative stances compared with other stance-expressive tweets.

We estimated a model, as outlined in Equation 1, with the stance of the tweet (coded as 1 for provaccination and 0 for antivaccination or hesitant) serving as the dependent variable. Our model incorporates random effects for users (α_j) since a user could potentially contribute multiple tweets and an intercept term (β_0). The coefficient of interest is β_1, which measures the impact of a user group (influencer vs. control) on tweet stance. Finally, given that the outcome variable (stance) was binary, our analysis was performed using a binomial generalized linear model to provide the most accurate representation of the data.

$$s_{ij}=\beta_0+\alpha_j+\beta_1\text{Group}_j+\varepsilon_{ij} \quad \textbf{(1)}$$

where i is a tweet index, j is a user index, and tweet stance s_{ij} is defined as

1, if tweet stance is labeled provaccination, or

0, if tweet stance is labeled negative or hesitant.

The model estimation results are presented in Table 3, where the control group is dummy-coded as the reference or baseline condition. The main effect of the group was significant (β_1=–2.2668, SE=0.2940, P<.001). Based on the direction of the effect, tweets from the influencer group that express a stance on vaccination are, on average, more negative than tweets from the control group. We, therefore, reject the null hypothesis that stance-taking tweets are equally provaccination across groups.

To unpack the factors that may predispose an account to support or oppose vaccination, we also investigated the potential contributions of two other variables: "whether a Twitter account represents an individual and the network centrality of the account before the pandemic." Accounts representing individuals may present opinions that are less moderated than accounts representing organizations or collectives, which could increase the propensity toward antivaccination messages. We did not have a strong a priori hypothesis about how an account's position in the social graph would relate to vaccine stance; thus, this analysis should be considered post hoc.

XSL•FO
RenderX

Table 3. Mixed-effects generalized linear model of tweet stance.

	Dependent variable (Tweet stance, s_{ij})
Group	–2.2668 (0.2940)
Constant	2.1706 (0.1567)
Observations	5616
Log Likelihood	–1955.119
Akaike Information Criterion	3920.237
Bayesian Information Criterion	3953.404

Are Twitter Accounts Representing Individuals More Likely to Oppose Vaccination?

By definition, every influencer account must represent an individual. However, in the control group, 79 accounts were labeled as individuals (there were 29 accounts in the control group that we could not confidently label as individuals or not; these accounts are excluded from the following analysis). To test whether individual status adds explanatory value to a model of tweet stance, we added an additional term representing individual status to our baseline model:

$$s_{ij}=\beta_0+\alpha_j+\beta_1\text{Group}_j+\beta_2\text{Individual}_j+\varepsilon_{ij} \quad \textbf{(2)}$$

where the indicator variable Individual is 1 if the account corresponds to an individual and 0 if not. Once again, as in Equation 2, the model fits coefficients αj (a random effect of the user), β0 (an intercept), and β1 (the effect of the user group). The new coefficient β2 refers to the main effect of individual status.

As can be seen from Table 4 (“Model 1”), the main effect of the group was still significant (β_1=–1.3032, SE=0.3612; *P*<.001), and so was the main effect of individual status (β=–0.9591, SE=0.2917; *P*=.001). Based on the sign of the individual status main effect, we can see that tweets from accounts representing individuals were more likely to take antivaccination stances. Thus, we can see that the “individual” status helps explain additional variation in Tweet stance beyond what is explained by influencer status alone.

Table 4. Mixed-effects generalized linear models of tweet stance with covariates.

	Model 1	Model 2
Group	–1.3032 (0.3612)	–2.035 (0.282)
Individual	–0.9591 (0.2917)	__[a]
Log-centrality	—	–0.076 (0.095)
Constant	2.5420 (0.1965)	1.436 (1.002)
Observations	5206	5206
Log Likelihood	–1797.0530	–1803.645
Akaike Information Criterion	3606.1060	3615.289
Bayesian Information Criterion	3656.4510	3651.520

[a]Not available.

Does Social Network Position Affect Vaccine Stance?

Engagement levels on Twitter vary significantly from one account to another. Some enjoy high retweet rates, signaling frequent engagement, while others remain largely overlooked. Despite our selection of accounts based on their high retweet count, this does not preclude significant variability within our sample. To quantify engagement within health-related conversations on Twitter, we analyzed the in-degree centrality of each account using the retweet network (before COVID-19 in 2019).

In-degree centrality gauges an account’s influence by the volume of retweets it receives from distinct users. Upon comparison, we found no significant difference in the centrality of influencer and control group accounts (Wilcoxon rank-sum test; *W*=8401.5, *P*=.27). This indicates that being an influencer does not necessarily correlate with higher retweet centrality within our study’s context. We also tested a mixed-effects generalized linear model as before, except with each user’s centrality measure included as a main effect. Centrality was log-transformed before modeling, as it is nonnormally distributed with a long tail. The model results are presented in Table 4 (“Model 2”).

$$s_{ij}=\beta_0+\alpha_j+\beta_1\text{Group}_j+\beta_2\text{log-centrality}_j+\varepsilon_{ij} \quad \textbf{(3)}$$

The main effect of log-centrality was not significant (β_2=–0.076, SE=0.09472; *P*=.42). Hence, we did not find evidence that the connectedness of Twitter accounts prepandemic had an

XSL•FO
RenderX

association with their stances on vaccinations during the pandemic.

Cluster Analysis of Pro- and Antivaccination Accounts

So far, we have used a supervised learning-based modeling strategy to explore the influence of Twitter account properties on Tweet stance. An alternative analysis strategy is to model users' stances in an unsupervised fashion, which yields more readily interpretable estimates of how many accounts in the influencer and control groups supported vaccination, respectively. To this end, we clustered users into two broad categories: vaccination supporters and opponents.

For all accounts that had at least one tweet labeled as containing a stance (120 influencer accounts and 201 control accounts), we calculated a stance vector as follows:

(4)

Where *i* is an index for each of the three stance labels produced by the VaxxHesitancy model ("positive," "hesitant," or "opposed"), n indexes every tweet with label *i* for a given account, and the "score" is the stance model's confidence score for that label (a number between 0 and 1). The stance vector is normalized by dividing by the total number of tweets per account so that its values sum to 1.

For example, an account with two tweets labeled "opposed," each with a confidence score of 1, would have a stance vector of [0, 0, 1]. If the account had two tweets, one labeled "opposed" and one "hesitant" (again with a confidence score of 1), the resulting vector would be [0, 0.5, 0.5]. Confidence scores are incorporated so that Tweets that are not confidently labeled by the model will contribute less to the stance vector.

Next, we used the *k*-means clustering algorithm to assign every account's stance vector to *k*=2 clusters. Accounts from both groups (influencer and control) were pooled together for clustering. We did not attempt to correct the class imbalance (ie, that there were more control accounts than influencer accounts).

Inspecting the resulting clusters (Figure 3), 91 accounts were assigned to a predominantly antivaccination cluster, whereas 230 were assigned to a predominantly provaccination cluster. Within the influencer group, 58 out of 120 (48%) accounts were assigned to the antivaccination cluster. Within the control group, 33 out of 201 accounts (16%) were assigned to the antivaccination cluster. The difference in proportions was statistically significant (2-sample test for equality of proportions; χ^2_1[N=321]=36.1, P<.001). This modeling exercise suggests that while antivaccination stances were a minority opinion in both groups, they were relatively more common among influencer accounts.

Figure 3. Result of k-means clustering of user stance vectors split by group. The x-axis is the provaccination component of the stance vector (defined in Equation 4). The y-axis is the sum of the antivaccination and hesitant components of the stance vector. Light jittering (+/- 0.1) has been added to reduce overplotting.

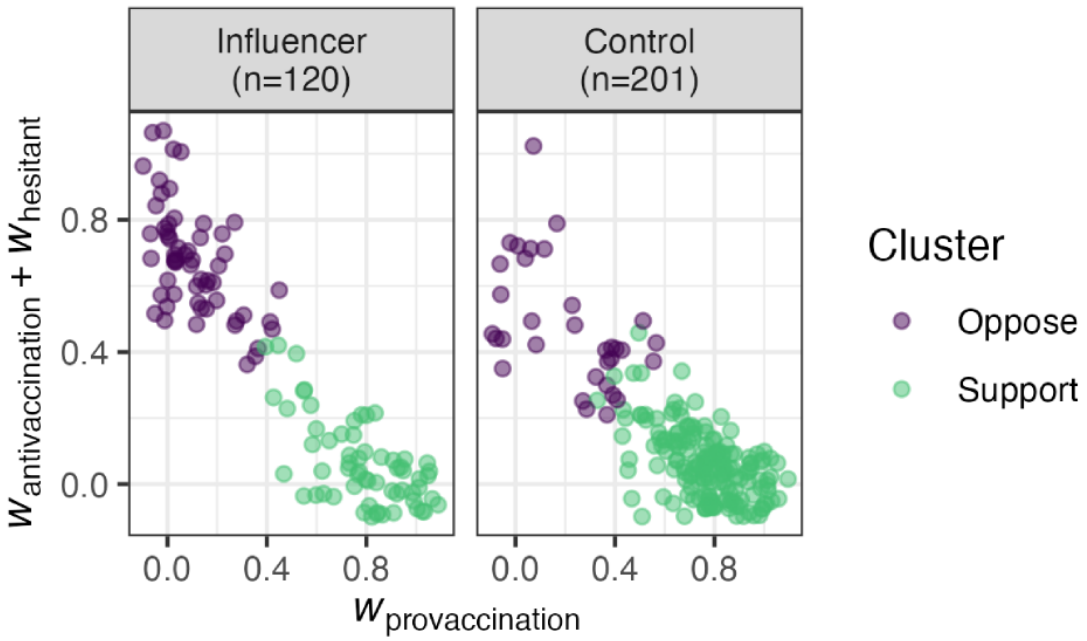

Topical Analysis of Argument Strategies for and Against Vaccination

Overview

To understand the rhetorical strategies used to promote or oppose vaccination, we once again used the BERTopic modeling library to cluster influencers' tweets (more details in the Methods section). Tables 5 and 6 show the top 5 themes that emerged from topic modeling for both anti- and provaccination tweets by wellness influencers. Among tweets opposed to vaccination, the most prevalent theme was criticism of recommending COVID-19 vaccination for children. Other major themes included government overreach through vaccine mandates, corruption of the pharmaceutical industry ("big pharma"), opposition to recurring booster shots, and rejection of vaccine passports. Taken together, these topics reveal that

XSL•FO
RenderX

wellness influencers who opposed COVID-19 vaccination often did so by sowing broader suspicions toward government and scientific institutions.

Among provaccination tweets, the largest cluster appears to represent encouragement and calls to action. Other major themes include broadcasting information about vaccine efficacy and safety, criticizing "antivaxxers," and discussing immunity. In addition, 3 of the clusters (vaccine efficacy, vaccine safety, and immunity) appear to involve sharing scientific (or at least scientific-sounding) information. While it is beyond the scope of this study to fact-check all scientific appeals made in the tweets, it is notable that adopting a scientific framing is a popular rhetorical strategy.

Table 5. Antivaccination themes in wellness influencer tweets.

Cluster theme	Number of tweets, n	Example
Protecting children	56	"White house says it is time for the vulnerable 5-year-olds to roll up their sleeves and take the COVID vax."
Government overreach	55	"Vaccines were supposed to be the way to freedom, they're clearly the way to more authoritarianism."
Big pharma	55	"CDC admits it: Big Pharma's injured 387,087; seriously injured 31,240."
Recurring booster shots	45	"If you want a picture of the future, imagine a booster injecting into a human arm, forever."
Vaccine passports	45	"Just say no to vaccine passports."

Table 6. Provaccination themes in wellness influencer tweets.

Cluster theme	Number of tweets, n	Example
Encouraging vaccination	108	"We all have to do our part and be vaccinated for the greater good."
Vaccine efficacy	98	"The Covid Vaccine is 80% effective after 8 months of development, when the flu vaccine is 40% effective after 70 years."
Vaccine safety	55	"Vaccines are absolutely safe. There is nothing to worry about. There is MUCH MORE risk from a Covid illness than from a vaccine."
Criticizing vaccine skeptics	41	"The Wisconsin vaccine saboteur was a microchip guy and a flat Earther."
Immunity	32	"We need 80-90% of us vaccinated to reach herd immunity and put covid in the rear view mirror."

Temporal Dynamics in Vaccine Support

Finally, we investigated temporal dynamics in vaccine support: was support relatively consistent across time? Were there different temporal dynamics for the influencer and control groups? Note that these questions were developed post hoc, and as such, we had no hypotheses in mind about what we would observe. These findings should, therefore, be considered exploratory.

Trends in daily support (as a percentage of stance-taking tweets) for the two groups are shown in Figure 4. We see that both groups followed a relatively parallel trajectory, with the key difference being the baseline level of support (ie, the trend curves are similar, except influencers are shifted lower on the y-axis compared to controls). The proportion of supportive messaging peaked in the first half of 2021 for both groups. It is challenging to directly align this peak with a clear indicator of vaccine access, as we are unable to verify the location of most users in our sample, and vaccine rollout schedules differ by country.

Despite this ambiguity, we can note a few milestones in global vaccine rollout: the first authorized COVID-19 vaccines after a large clinical trial became available in December 2020, starting in the United Kingdom and quickly followed by other nations [66] (though note that Russia and China had already begun distributing candidate vaccines based on intermediate clinical trial results). The number of COVID-19 vaccines administered per day around the world increased from December 2020 until it peaked around June 2021, with a smaller peak in daily vaccination counts around December 2021 [67].

If a large share of provaccination messaging from influencers related to encouraging people to get vaccinated (more details in Table 6), then it seems reasonable that this messaging would be concentrated while vaccine distribution was reaching its highest velocity—during the first half of 2021. Although trendlines for both groups appear to rise again at the end of 2022, we urge caution interpreting this—there are simply fewer on-topic tweets per day at the beginning and end of the observation period (more details in Figure 2), so estimates of the prevalence of provaccination stances will necessarily be noisier. This is visible in the wider standard error ranges in early 2020 and late 2022.

XSL•FO
RenderX

Figure 4. The proportion of provaccination tweets per day, out of all stance-taking tweets about vaccines, during the observation period. Loess smoothing has been applied for ease of interpretation, and shaded ranges indicate one standard error. Standard errors are wider at the start and end of the observation period when there were relatively fewer on-topic tweets per day.

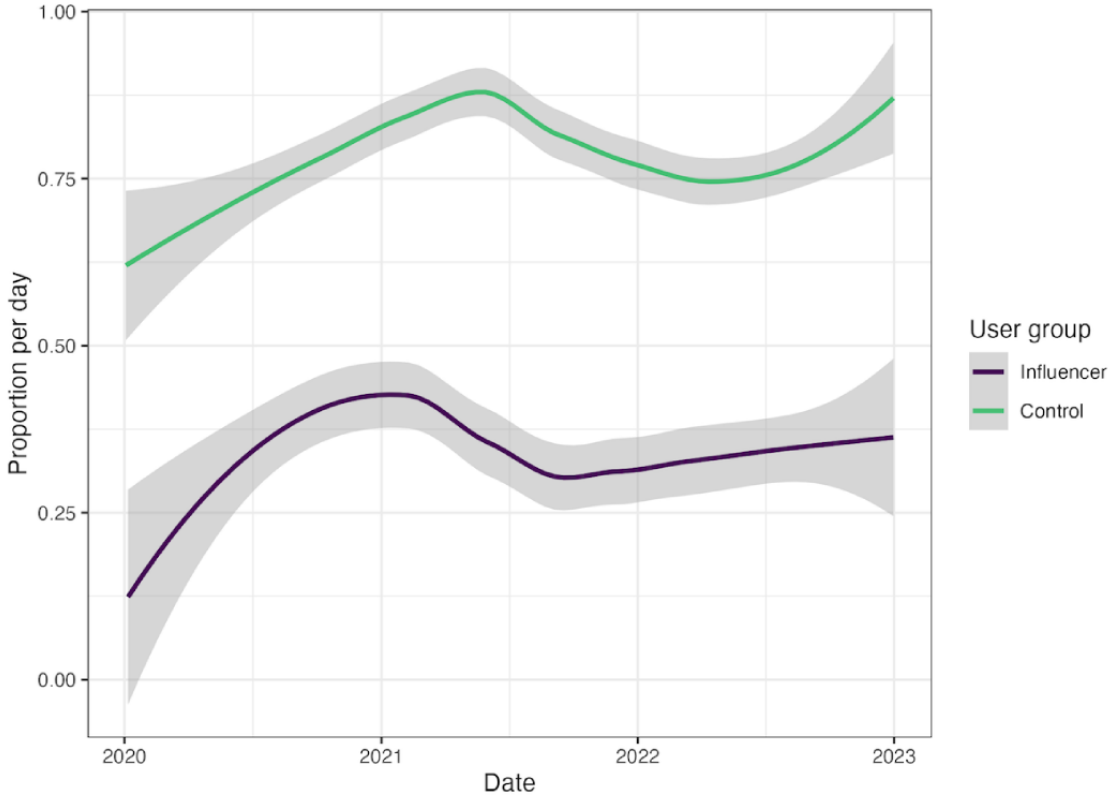

While it is possible that temporal dynamics in supportive messaging are related to accounts changing stances, it seems more likely that the burst of provaccination messaging reflects activity by a group of accounts that were consistently provaccine. This prediction is based largely on our user clustering analysis, in which the influencers group appears to have distinct stance clusters rather than a uniform continuum (see Figure 3). If many users were switching their opinions, we would expect a more continuous distribution, as averaging a mixture of pro- and antivaccination stances should lead to an intermediate valence.

However, further study would be required to definitively estimate the prevalence of stance changes. Detecting changes reliably would require statistical power beyond what we are able to achieve through a 10% random sample of tweets alone—at an absolute minimum, there would need to be at least two stance-taking tweets present per account at sufficiently separate time points (and in practice, the number will be even higher because the stance-detection model is not perfectly reliable). This question could ideally be answered using a more complete archive of social media activity.

Discussion

Summary of Findings

This study provides evidence that wellness influencers on Twitter were more likely to oppose COVID-19 vaccination compared with other Twitter users participating in health discussions before the pandemic. In our analysis, we identified a cohort of wellness influencer accounts during 2019, before the onset of the global coronavirus pandemic. Compared to a control group of accounts that posted on similar topics, wellness influencers were more likely to tweet messages expressing antivaccination stances during the rollout of the COVID-19 vaccine. Among wellness influencer accounts for which we could estimate vaccine stance, roughly half (48%) were identified as opposing vaccines, compared with 16% of accounts in the control group. This overall finding was robust to incorporating covariates into our statistical models, such as the social network centrality of an account or whether an account represents an individual or not.

In addition, we conducted an exploratory analysis of themes in pro- and antivaccination tweets by influencers. Our topic modeling approach provides further evidence that anti-establishment messaging comprises a core part of many wellness influencers' rhetoric. These accounts invoked themes like parental rights, government overreach, and distrust of corrupt pharmaceutical companies when opposing vaccination. Provaccination influencers, on the other hand, encouraged followers to get vaccinated and shared scientifically framed information about vaccine safety, efficacy, and its relationship with immunity. Their use of scientific framing suggests these influencers recognize the cultural authority science holds.

Implications

Our findings are congruent with the hypothesis raised by Baker and Rojek [24,25] that the rise of wellness influencers is a direct response to waning public trust in traditional authorities like medical professionals or public health agencies. Notably, the dominant antivaccine messaging themes among wellness

XSL•FO
RenderX

influencers we observed were related to parental rights, government overreach, and corruption. While their tweets often contained scientific-sounding language, the rhetoric was markedly political.

If these influencers are indeed popular because of preexisting low trust in public institutions that make health knowledge and policies, then attempts to better regulate or counteract wellness content online may have limited impact. Indeed, a 2021 meta-analysis of COVID-19 misinformation mitigation strategies, such as offering corrections by experts or peers, found overall no statistically significant effect for these interventions online or offline [21]. This does not prove counter-messaging against vaccine misinformation never works, but suggests useful interventions are likely to be context-specific (ie, tailored to a particular community) with modest effect sizes.

Based on our results and the current literature, we would recommend a more direct—though ultimately far more challenging—approach: proactively investing in efforts to restore trust in public institutions to stem the tide of interest in wellness influencers before major public health crises occur.

Limitations

Sources of Measurement Error

Our work has several limitations. The findings here are limited by reliance on Twitter data, particularly with a 10% random sample of tweets. Because we cannot query any user's entire tweet history and the stance detection model is not perfectly accurate, our estimates of the prevalence of pro- and antivaccination messaging are all probabilistic. In addition, while some users produced a high volume of stance-taking tweets in our dataset, for others, we could only recover a single stance-taking tweet. Our group-level analyses should, therefore, be treated with higher confidence than our individual-level analyses, which are surely noisier.

Sample Size and Generalizability

Second, our sample size is relatively small for a social media study with only a few hundred Twitter accounts. Our criteria for labeling influencers are conservative, meaning they likely have a high false negative rate and a low false positive rate. In other words, while we are confident that the accounts we have identified as influencers meet our criteria, we have likely missed similar accounts that simply did not have as many tweets present in the Twitter Decahose random sample to be detected. Because we are attempting to make an argument about how a distinct subset of Twitter users behave, we prefer this conservative approach, which prioritizes the validity of our sample over a larger dataset. However, this design choice implies that our results should not be taken as a comprehensive view of support for vaccination on Twitter broadly.

Consequences of The Twitter API Change

When we first designed this study and carried out our initial exploratory work in the space, researchers had broad access to Twitter's free API, which could be used to query information about accounts like their bio, locations, full tweet history, and followers. In February 2023, however, Twitter ended free access to its API, including for researchers. This blockage required us to rethink several aspects of our study, as it effectively limited us to using only what data could be accessed through the Twitter Decahose. The Decahose, while an extraordinary resource for researchers, stores only information about tweets rather than user accounts. Consequently, we can only learn about users through the tweets attached to their account handle in the random 10% sample.

The API closure required us to make several major changes to our research strategy. While we had initially intended to identify influencers through a network analysis using Twitter lists of wellness influencers as the seed, the API shutdown meant we could no longer easily assemble a complete follower network (or collection of lists). Instead, we reconstructed the social network through retweets in the Decahose dataset during 2019 and used keywords to ensure topical relevance.

In addition, we planned to use researcher tools such as Botometer [63] (to detect bots in our sample) and M3-Inference [68] (to label accounts corresponding to individuals and infer demographic variables) to programmatically identify candidate influencer accounts at scale. However, both these tools require Twitter API access, and as such, we were unable to incorporate them. This led us to use stricter criteria for selecting candidate influencer accounts from keyword usage to produce a smaller volume of candidate accounts amenable to manual review.

Furthermore, we had intended to pull the entire available tweet history for users identified as influencers or members of the control group through the API, which would have given us a better statistical estimate for vaccine stance per individual. Although we were still able to address our core research questions, better estimates of stance at the individual level would have allowed us to look at interactions with demographic variables, topical interests, and time (to measure the prevalence of stance-changing).

While we hope our work in its current form is still useful to the broader research community, our experience also testifies to the serious consequences of the API shutdown on social media research—both for data access and for the tremendous ecosystem of open-source software tools for researchers built around Twitter.

Further Research Directions

While our study provides an initial estimate of the prevalence of antipublic health messaging in a cohort of wellness influencers, further research will be needed to fully understand the reach and impact of this messaging. The measures in our study focus on the content of tweets rather than their effect on readers. We do not know how many people saw the vaccine-related tweets in our sample or the demographics of the audience for the tweets.

Perhaps more importantly, we cannot decisively say whether the rhetoric used by wellness influencers was persuasive to followers. There has been recent attention on the "wellness to conspiracy" pipeline [26,69,70], which proposes that participation in online wellness communities can lead to exposure to—and even adoption of—conspiracy beliefs (for example, the American far-right conspiracy Q-Anon). While it is clear from the literature that some individuals have followed

XSL•FO
RenderX

such a path, it remains possible that the pipeline is narrow: just as science denial is characterized by picking and choosing select aspects of scientific consensus to reject [71-73], people who follow wellness influencer accounts for diet suggestions might discount influencer's opinions on vaccines.

One way to directly assess the downstream effects of following wellness influencers or participating in wellness communities online would be to follow users over time, an approach that has been used to study pathways toward and away from conspiracy beliefs on Reddit [74,75]. Considering that the researcher API for Twitter has disappeared indefinitely, further studies would likely have to be conducted on other social media platforms. Transitioning to another social media platform for research would also provide an opportunity to check that our main findings replicate in other online spaces.

An additional direction for research would be on protective factors observed in wellness influencers who go on to support public health initiatives. We observed that roughly half of the wellness influencers in our sample who expressed a stance on the COVID-19 vaccines voiced support. These users need to craft messaging that preserves their own authority as health experts of some sort while also expressing a position that broadly supports establishment sources of knowledge. Identifying their messaging strategies and potential motivations (both commercial and ideological) could illuminate productive avenues for public health leaders and social media personalities to work together.

Acknowledgments

We thank the Gates NLP team, especially Yida Mu, for supplying us with the vaccine stance detection model and documentation.

Conflicts of Interest

None declared.

References

1. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015;22(1):243-256 [FREE Full text] [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
2. Cooper CP, Gelb CA, Rim SH, Hawkins NA, Rodriguez JL, Polonec L. Physicians who use social media and other internet-based communication technologies. J Am Med Inform Assoc 2012;19(6):960-964 [FREE Full text] [doi: 10.1136/amiajnl-2011-000628] [Medline: 22634078]
3. Moukarzel S, Rehm M, Del Fresno M, Daly AJ. Diffusing science through social networks: the case of breastfeeding communication on twitter. PLoS One 2020;15(8):e0237471 [FREE Full text] [doi: 10.1371/journal.pone.0237471] [Medline: 32790712]
4. Ghenai A, Mejova Y. Fake cures: user-centric modeling of health misinformation in social media. 2018 Presented at: Proc ACM Hum Comput Interact Association for Computing Machinery; 2018 March 10-15; Singapore p. 1-20. [doi: 10.1145/3274327]
5. Abbar S, Mejova Y, Weber I. You tweet what you eat: studying food consumption through Twitter. 2015 Presented at: ACM CHI Conference on Human Factors in Computing Systems (CHI); 2024 May 11-16; Yokohama, Japan p. 3197-3206. [doi: 10.1145/2702123.2702153]
6. Stupinski AM, Alshaabi T, Arnold MV, Adams JL, Minot JR, Price M, et al. Quantifying changes in the language used around mental health on twitter over 10 years: observational study. JMIR Ment Health 2022;9(3):e33685 [FREE Full text] [doi: 10.2196/33685] [Medline: 35353049]
7. Pagoto S, Schneider KL, Evans M, Waring ME, Appelhans B, Busch AM, et al. Tweeting it off: characteristics of adults who tweet about a weight loss attempt. J Am Med Inform Assoc 2014;21(6):1032-1037 [FREE Full text] [doi: 10.1136/amiajnl-2014-002652] [Medline: 24928175]
8. McClellan C, Ali MM, Mutter R, Kroutil L, Landwehr J. Using social media to monitor mental health discussions - evidence from twitter. J Am Med Inform Assoc 2017;24(3):496-502 [FREE Full text] [doi: 10.1093/jamia/ocw133] [Medline: 27707822]
9. Zhang H, Wheldon C, Dunn AG, Tao C, Huo J, Zhang R, et al. Mining twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. J Am Med Inform Assoc 2020;27(2):225-235 [FREE Full text] [doi: 10.1093/jamia/ocz191] [Medline: 31711186]
10. Vydiswaran VGV, Romero DM, Zhao X, Yu D, Gomez-Lopez I, Lu JX, et al. Uncovering the relationship between food-related discussion on twitter and neighborhood characteristics. J Am Med Inform Assoc 2020;27(2):254-264 [FREE Full text] [doi: 10.1093/jamia/ocz181] [Medline: 31633756]
11. Fox S. The social life of health informatio. California Healthcare Foundation. 2011. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf [accessed 2024-10-16]

XSL•FO
RenderX

12. De Choudhury MD, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. 2014 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; 2024 June 3-6; Buffalo, New York, USA p. 71-80. [doi: 10.1609/icwsm.v8i1.14526]
13. De Choudhury M, Kıcıman E. The language of social support in social media and its effect on suicidal ideation risk. Proc Int AAAI Conf Weblogs Soc Media 2017;2017:32-41 [FREE Full text] [Medline: 28840079]
14. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. J Gen Intern Med 2020;35(7):2244-2247 [FREE Full text] [doi: 10.1007/s11606-020-05889-w] [Medline: 32399912]
15. Czerniak K, Pillai R, Parmar A, Ramnath K, Krocker J, Myneni S. A scoping review of digital health interventions for combating COVID-19 misinformation and disinformation. J Am Med Inform Assoc 2023;30(4):752-760 [FREE Full text] [doi: 10.1093/jamia/ocad005] [Medline: 36707998]
16. Vijaykumar S, Rogerson DT, Jin Y, de Oliveira Costa MS. Dynamics of social corrections to peers sharing COVID-19 misinformation on whatsApp in Brazil. J Am Med Inform Assoc 2021;29(1):33-42 [FREE Full text] [doi: 10.1093/jamia/ocab219] [Medline: 34672323]
17. Hua Y, Jiang H, Lin S, Yang J, Plasek JM, Bates DW, et al. Using twitter data to understand public perceptions of approved versus off-label use for COVID-19-related medications. J Am Med Inform Assoc 2022;29(10):1668-1678 [FREE Full text] [doi: 10.1093/jamia/ocac114] [Medline: 35775946]
18. Grimes DR. Health disinformation & social media: The crucial role of information hygiene in mitigating conspiracy theory and infodemics. EMBO Rep 2020;21(11):e51819 [FREE Full text] [doi: 10.15252/embr.202051819] [Medline: 33155436]
19. Grimes DR. Medical disinformation and the unviable nature of COVID-19 conspiracy theories. PLoS One 2021;16(3):e0245900 [FREE Full text] [doi: 10.1371/journal.pone.0245900] [Medline: 33711025]
20. Himelein-Wachowiak M, Giorgi S, Devoto A, Rahman M, Ungar L, Schwartz HA, et al. Bots and misinformation spread on social media: implications for COVID-19. J Med Internet Res 2021;23(5):e26933 [FREE Full text] [doi: 10.2196/26933] [Medline: 33882014]
21. Janmohamed K, Walter N, Nyhan K, Khoshnood K, Tucker JD, Sangngam N, et al. Interventions to mitigate COVID-19 misinformation: a systematic review and meta-analysis. J Health Commun 2021;26(12):846-857. [doi: 10.1080/10810730.2021.2021460] [Medline: 35001841]
22. Jiang S, Fang W. Misinformation and disinformation in science: examining the social diffusion of rumours about GMOs. Cultures of Science 2019;2(4):327-340. [doi: 10.1177/209660831900200407]
23. Solomon M. Trust: the need for public understanding of how science works. Hastings Cent Rep 2021;51 Suppl 1:S36-S39. [doi: 10.1002/hast.1227] [Medline: 33630343]
24. Rojek C, Baker SA. Lifestyle Gurus: Constructing Authority and Influence Online. Hoboken, New Jersey: John Wiley & Sons; 2020.
25. Baker SA, Rojek C. The belle gibson scandal: the rise of lifestyle gurus as micro-celebrities in low-trust societies. Journal of Sociology 2019;56(3):388-404. [doi: 10.1177/1440783319846188]
26. Baker SA. Alt. health influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the COVID-19 pandemic. European Journal of Cultural Studies 2022;25(1):3-24. [doi: 10.1177/13675494211062623]
27. Allchin D. Who Speaks for Science? Sci Educ (Dordr) 2022;31(6):1475-1492 [FREE Full text] [doi: 10.1007/s11191-021-00257-4] [Medline: 34690433]
28. Nguyen A, Catalan-Matamoros D. Digital Mis/Disinformation and public engagment with health and science controversies: fresh perspectives from Covid-19. MaC 2020;8(2):323-328. [doi: 10.17645/mac.v8i2.3352]
29. Gauchat G. The cultural authority of science: public trust and acceptance of organized science. Public Underst Sci 2011;20(6):751-770. [doi: 10.1177/0963662510365246] [Medline: 22397083]
30. Millstone E, van Zwanenberg P. A crisis of trust: for science, scientists or for institutions? Nat Med 2000;6(12):1307-1308. [doi: 10.1038/82102] [Medline: 11100103]
31. Mann M, Schleifer C. Love the Science, hate the scientists: conservative identity protects belief in science and undermines trust in scientists. Soc Forces Oxford Academic 2020;99(1):305-332. [doi: 10.1093/sf/soz156]
32. Baker SA, Walsh MJ. 'A mother's intuition: it's real and we have to believe in it': how the maternal is used to promote vaccine refusal on instagram. Inf Commun Soc 2022;26(8):1675-1692. [doi: 10.1080/1369118x.2021.2021269]
33. Mu Y, Jiang Y, Heppell F, Singh IC, Bontcheva K, Song X. A large-scale comparative study of accurate COVID-19 information versus misinformation. arXiv:2304.04811 2023. [doi: 10.48550/arXiv.2304.04811]
34. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
35. Teodoro R, Naaman M. Fitter with twitter: understanding personal health and fitness activity in social media. ICWSM 2021;7(1):611-620. [doi: 10.1609/icwsm.v7i1.14417]
36. Lee JL, DeCamp M, Dredze M, Chisolm MS, Berger ZD. What are health-related users tweeting? A qualitative content analysis of health-related users and their messages on twitter. J Med Internet Res 2014;16(10):e237 [FREE Full text] [doi: 10.2196/jmir.3765] [Medline: 25591063]

XSL•FO
RenderX

37. Makita M, Mas-Bleda A, Morris S, Thelwall M. Mental health discourses on twitter during mental health awareness week. Issues Ment Health Nurs 2021;42(5):437-450. [doi: 10.1080/01612840.2020.1814914] [Medline: 32926796]
38. Ghenai A, Mejova Y. Catching zika fever: application of crowdsourcing and machine learning for tracking health misinformation on twitter. : IEEE; 2017 Presented at: 2017 IEEE International Conference on Healthcare Informatics (ICHI); 2017 August 23-26; Park City, UT, USA. [doi: 10.1109/ichi.2017.58]
39. Strudwicke IJ, Grant WJ. #JunkScience: investigating pseudoscience disinformation in the Russian internet research agency tweets. Public Underst Sci 2020;29(5):459-472. [doi: 10.1177/0963662520935071] [Medline: 32597365]
40. Samory M, Mitra T. The government spies using our webcams': the language of conspiracy theories in online discussions. 2018 Presented at: Proc ACM Hum Comput Interact Association for Computing Machinery; 2018 November 1; United States p. 1-24. [doi: 10.1145/3274421]
41. Lavorgna A, Carr L. Tweets and quacks: network and content analyses of providers of non-science-based anticancer treatments and their supporters on twitter. Sage Open 2021;11(1). [doi: 10.1177/21582440211003084]
42. Project TV, Cryst E, DiResta R, Meyersohn L. Memes, magnets and microchips: Narrative dynamics around COVID-19 vaccines. URL: https://purl.stanford.edu/mx395xj8490 [accessed 2024-10-29]
43. Mejova Y. Information sources and needs in the obesity and diabetes twitter discourse. 2018 Presented at: Proceedings of the 2018 International Conference on Digital Health; 2018 April 23 - 26; Lyon France p. 21-29. [doi: 10.1145/3194658.3194664]
44. Burki T. Vaccine misinformation and social media. The Lancet Digital Health 2019;1(6):e258-e259. [doi: 10.1016/S2589-7500(19)30136-0]
45. Jiang B, Sheth P, Li B, Liu H. CoVaxNet: an online-offline data repository for COVID-19 vaccine research. arXiv:2207.01505 2023. [doi: 10.36190/2023.07]
46. Wang H, Li Y, Hutch MR, Kline AS, Otero S, Mithal LB, et al. Patterns of diverse and changing sentiments towards COVID-19 vaccines: a sentiment analysis study integrating 11 million tweets and surveillance data across over 180 countries. J Am Med Inform Assoc 2023;30(5):923-931 [FREE Full text] [doi: 10.1093/jamia/ocad029] [Medline: 36821435]
47. Jamison A, Broniatowski DA, Smith MC, Parikh KS, Malik A, Dredze M, et al. Adapting and extending a typology to identify vaccine misinformation on twitter. Am J Public Health 2020;110(S3):S331-S339. [doi: 10.2105/AJPH.2020.305940] [Medline: 33001737]
48. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
49. Pavlova A, Berkers P. Mental health discourse and social media: which mechanisms of cultural power drive discourse on twitter. Soc Sci Med 2020;263:113250 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113250] [Medline: 32862081]
50. Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. Int J Inf Manag 2018;38(1):1-6. [doi: 10.1016/j.ijinfomgt.2017.08.002]
51. Mu Y, Jin M, Grimshaw C, Scarton C, Bontcheva K, Song X. VaxxHesitancy: a dataset for studying hesitancy towards COVID-19 vaccination on Twitter. ICWSM 2023;17:1052-1062. [doi: 10.1609/icwsm.v17i1.22213]
52. Mikolov T, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst. 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf [accessed 2024-04-14]
53. Rodriguez D. word2vec. PyPI. URL: https://pypi.org/project/word2vec/ [accessed 2024-10-16]
54. Haque A, Ginsparg P. BERTopic: neural topic modeling with a class-based TF-IDF procedure. J. Am. Soc. Inf. Sci 2009;60(11):2203-2218 [FREE Full text] [doi: 10.1002/asi.21166]
55. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. 2019 Presented at: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2024 October 24; Hong Kong, China p. 3982-3992. [doi: 10.18653/v1/d19-1410]
56. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. JOSS 2018;3(29):861. [doi: 10.21105/joss.00861]
57. McInnes L, Healy J, Astels S. hdbscan: hierarchical density based clustering. JOSS 2017;2(11):205. [doi: 10.21105/joss.00205]
58. Kang J, Lerman K. Using lists to measure homophily on Twitter. Workshops at the twenty-sixth AAAI conference on artificial intelligence. 2012. URL: https://cdn.aaai.org/ocs/ws/ws0934/5425-22771-1-PB.pdf [accessed 2024-04-14]
59. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. Annu Rev Sociol 2001;27(1):415-444. [doi: 10.1146/annurev.soc.27.1.415]
60. Yang KC, Varol O, Nwala AC, Sayyadiharikandeh M, Ferrara E, Flammini A, et al. Social Bots: Detection and Challenges. 2023. URL: https://arxiv.org/abs/2312.17423v1 [accessed 2024-04-11]
61. Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: detection, estimation, and characterization. ICWSM 2017;11(1):280-289. [doi: 10.1609/icwsm.v11i1.14871]
62. Varol O. Should we agree to disagree about twitter's bot problem? OSNEM 2023;37-38:100263. [doi: 10.1016/j.osnem.2023.100263]

XSL•FO
RenderX

63. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. BotOrNot: a system to evaluate social bots. 2016 Presented at: 25th International Conference on World Wide Web Association for Computing Machinery; 2016 April 11 - 15; Montréal Québec Canada p. 273-274. [doi: 10.1145/2872518.2889302]
64. Townsend L, Wallace C, Harte D. Social Media Research: A Guide to Ethics. University of Aberdeen. 2016. URL: https://www.gla.ac.uk/media/Media_487729_smxx.pdf [accessed 2024-04-14]
66. Ledford H, Cyranoski D, Van Noorden R. The UK has approved a COVID vaccine - here's what scientists now want to know. Nature 2020;588(7837):205-206. [doi: 10.1038/d41586-020-03441-8] [Medline: 33288887]
67. Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, et al. A global database of COVID-19 vaccinations. Nat Hum Behav 2021;5(7):947-953. [doi: 10.1038/s41562-021-01122-8] [Medline: 33972767]
68. Wang Z, Hale S, Adelani D, Grabowicz P, Hartmann T, Flöck F, et al. Demographic inference and representative population estimates from multilingual social media data. 2019 Presented at: The World Wide Web Conference; 2019 May 13; Singapore p. 2056-2067. [doi: 10.1145/3308558.3313684]
69. Mclaughlin M. The neoliberal wellness journey down the rabbit hole. California State University. San Bernardino; 2021. URL: https://scholarworks.lib.csusb.edu/etd/1277 [accessed 2024-04-21]
70. Sandlin JA, Gómez AE. Toward new critical pedagogies of conspirituality consumption: exploring and combatting the COVID‐19 new‐age grifters. New Dir Adult Contin Educ 2023;2023(178):41-57. [doi: 10.1002/ace.20486]
71. Hornsey MJ. Why facts are not enough: understanding and managing the motivated rejection of science. Curr Dir Psychol Sci 2020;29(6):583-591. [doi: 10.1177/0963721420969364]
72. Lewandowsky S, Oberauer K. Motivated rejection of science. Curr Dir Psychol Sci 2016;25(4):217-222. [doi: 10.1177/0963721416654436]
73. Pasek J. It's not my consensus: motivated reasoning and the sources of scientific illiteracy. Public Underst Sci 2018;27(7):787-806. [doi: 10.1177/0963662517733681] [Medline: 28942728]
74. Engel K, Phadke S, Mitra T. Learning from the ex-believers: individuals' journeys in and out of conspiracy theories online. 2023 Presented at: Proceedings of the ACM on Human-Computer Interaction; 2016 May 11-16; New York, USA p. 1-37. [doi: 10.1145/3610076]
75. Phadke S, Samory M, Mitra T. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. 2022 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; 2024 June 3-6; Buffalo, New York, USA p. 770-781. [doi: 10.1609/icwsm.v16i1.19333]

Abbreviations

API: application programming interface
HDBSCAN: hierarchical density-based spatial clustering of applications with noise
UMAP: Uniform Manifold Approximation and Projection

Edited by A Mavragani; submitted 22.01.24; peer-reviewed by M Elbattah, C Ruiz-Nunez, S Wang; comments to author 21.03.24; revised version received 22.04.24; accepted 17.10.24; published 27.11.24.

Please cite as:
O'Brien G, Ganjigunta R, Dhillon PS
Wellness Influencer Responses to COVID-19 Vaccines on Social Media: A Longitudinal Observational Study
J Med Internet Res 2024;26:e56651
URL: https://www.jmir.org/2024/1/e56651
doi:10.2196/56651
PMID:

©Gabrielle O'Brien, Ronith Ganjigunta, Paramveer S Dhillon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Long COVID Discourse in Canada, the United States, and Europe: Topic Modeling and Sentiment Analysis of Twitter Data

Ahmed Ghassan Tawfiq AbuRaed[1], PhD; Emil Azuma Prikryl[2], MD; Giuseppe Carenini[1], Prof Dr; Naveed Zafar Janjua[3], DrPH, Prof Dr Med

[1]Department of Computer Science, The University of British Colombia, Vancouver, BC, Canada

[2]NOSM University, Thunder Bay, BC, Canada

[3]BC Centre for Disease Control, Vancouver, BC, Canada

Corresponding Author:
Ahmed Ghassan Tawfiq AbuRaed, PhD
Department of Computer Science
The University of British Colombia
2366 Main Mall
ICICS Building
Vancouver, BC, V6T 1Z4
Canada
Phone: 1 7789294413
Email: ahmed.aburaed@ubc.ca

Abstract

Background: Social media serves as a vast repository of data, offering insights into public perceptions and emotions surrounding significant societal issues. Amid the COVID-19 pandemic, long COVID (formally known as post–COVID-19 condition) has emerged as a chronic health condition, profoundly impacting numerous lives and livelihoods. Given the dynamic nature of long COVID and our evolving understanding of it, effectively capturing people's sentiments and perceptions through social media becomes increasingly crucial. By harnessing the wealth of data available on social platforms, we can better track the evolving narrative surrounding long COVID and the collective efforts to address this pressing issue.

Objective: This study aimed to investigate people's perceptions and sentiments around long COVID in Canada, the United States, and Europe, by analyzing English-language tweets from these regions using advanced topic modeling and sentiment analysis techniques. Understanding regional differences in public discourse can inform tailored public health strategies.

Methods: We analyzed long COVID–related tweets from 2021. Contextualized topic modeling was used to capture word meanings in context, providing coherent and semantically meaningful topics. Sentiment analysis was conducted in a zero-shot manner using Llama 2, a large language model, to classify tweets into positive, negative, or neutral sentiments. The results were interpreted in collaboration with public health experts, comparing the timelines of topics discussed across the 3 regions. This dual approach enabled a comprehensive understanding of the public discourse surrounding long COVID. We used metrics such as *normalized pointwise mutual information* for coherence and *topic diversity* for diversity to ensure robust topic modeling results.

Results: Topic modeling identified five main topics: (1) long COVID in people including children in the context of vaccination, (2) duration and suffering associated with long COVID, (3) persistent symptoms of long COVID, (4) the need for research on long COVID treatment, and (5) measuring long COVID symptoms. Significant concern was noted across all regions about the duration and suffering associated with long COVID, along with consistent discussions on persistent symptoms and calls for more research and better treatments. In particular, the topic of persistent symptoms was highly prevalent, reflecting ongoing challenges faced by individuals with long COVID. Sentiment analysis showed a mix of positive and negative sentiments, fluctuating with significant events and news related to long COVID.

Conclusions: Our study combines natural language processing techniques, including contextualized topic modeling and sentiment analysis, along with domain expert input, to provide detailed insights into public health monitoring and intervention. These findings highlight the importance of tracking public discourse on long COVID to inform public health strategies, address misinformation, and provide support to affected individuals. The use of social media analysis in understanding public health issues is underscored, emphasizing the role of emerging technologies in enhancing public health responses.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59425) doi:10.2196/59425

KEYWORDS

long COVID; topic modeling; sentiment analysis; Twitter; public perception; social media analysis; public health

Introduction

As of December 2023, there have been more than 700 million cases of COVID-19 globally, leading to nearly 7 million deaths [1]. These figures likely constitute an underestimation, given reduced reporting requirements for COVID-19 in most countries.

Beyond the acute effects of infection, a significant portion of survivors of COVID-19 experience a broad spectrum of ongoing symptoms several months after infection, which are generally captured under terms such as long COVID and post–COVID-19 condition. The estimated proportion of survivors with long COVID varies extremely widely from below 10% to around 60% across different studies [2], but population-based studies have reported that about 20% of people developed long COVID [3].

While the definition of long COVID continues to evolve, it has been defined as new, returning, or ongoing health problems that persist or occur 4 or more weeks after SARS-CoV-2 infection, with wide-ranging symptoms that can include fatigue, postexertional malaise, shortness of breath, palpitations, trouble sleeping, cognitive deficits, anxiety, and depression [4].

The characteristics and subtypes of long COVID, which likely represents several overlapping entities, continue to be an active area of research using both traditional observational study designs [5] and other innovative data-driven approaches [6].

During the COVID-19 pandemic, people have used social media such as Twitter (rebranded as “X” in 2023) to share information, opinions, and sentiments about COVID-19 and long COVID. This kind of information can help inform health care organizations and public health organizations and assist them in developing approaches and interventions that are sensitive to the concerns and perceptions of the public. This is of relevance to long COVID because it is a diverse condition that is experienced and perceived by the public in many different ways, making the identification of key issues and concerns associated with the condition of interest to health organizations to inform their communications, interventions, and research objectives around long COVID.

The use of topic modeling and sentiment analysis is widespread for identifying issues and public opinions in the field of public health. These approaches are also being used to gain insights into COVID-19–related matters. Analyses have been performed to unveil patterns in health communications across a variety of data sources, communities, and geographic locations.

While certain studies explored news articles [7] or research papers [8,9], the majority of research concentrated on social media platforms such as Reddit posts [10,11] and tweets [10,12-31]. Various approaches have been used to address the tasks of topic modeling, its visualization, and sentiment analysis.

Latent Dirichlet allocation (LDA) has been widely regarded as the best approach for topic modeling until recent years, due to its probabilistic foundation and effectiveness in uncovering latent themes within a corpus of text [32]. While most of the studies have used LDA to achieve topic modeling [8,12,14,16,20,24,33,34], Ridhwan and Hargreaves [13] have used Gibbs Sampling Dirichlet Multinomial Mixture [35], while Zheng et al [36] used conditional random field for a named entity recognition task aiming to extract the most frequent terms and applying the Jaccard similarity coefficient. Furthermore, Yan et al [37] applied biterm topic modeling, and Sussman et al [10] have used Cision’s Brandwatch software to perform both topic modeling and sentiment analysis.

On the other hand, sentiment analysis has been tackled by various rule-based and machine learning methods. Valence Aware Dictionary for Sentiment Reasoning [38] is a famous rule-based system that has been used by many studies [12-14]. TextBlob is a library for processing text that also implements a rule-based sentiment analysis system used by several studies [17,20,22]. Marcec and Likic [18] have used the AFINN lexicon [39], a list of English terms manually rated for valence.

Moreover, other studies have used machine learning methods to improve sentiment analysis. These include classical machine learning methods such as logistic regression, AdaBoost [39], XGBoost [40], and advanced techniques such as long short-term memory networks [41], which have been used by Guo et al [21]. Multilayer perceptron [42], naïve Bayes [43], and support vector machine [44] algorithms have been used by Masood et al [23]. In addition, Kumar et al [19] used bidirectional encoder representations from transformers [45], a deep neural network based on transformer architecture, demonstrating the efficacy of these advanced models in capturing the nuances of sentiment in social media text.

By integrating these approaches, researchers have been able to develop more sophisticated models for sentiment analysis, contributing to a deeper understanding of public opinion and emotional responses on social media platforms. Moreover, some studies have also explored sentiments in the context of specific aspects selected by domain experts [46]. However, such methods have not been used extensively to investigate issues related to long COVID, except for a few select papers (Table 1). While the research involved variations of topic modeling or sentiment analysis techniques, they generally did not combine these approaches.

XSL•FO
RenderX

Table 1. Related work on topic modeling and sentiment analysis on long COVID–related data[a].

Authors	Source	Posters	Time	Location	Language	Topic modeling	Sentiment analysis
Déguilhem et al [47]	Twitter, Reddit, Doctissimo, Facebook, and other forums	Public	January 1, 2020, to August 10, 2021	France	French	Yes, biterm topic modeling [37]	No
Bhattacharyya et al [48]	Twitter	Public	August 28, 2022, to September 6, 2022	Not specified	English	No	Yes, National Research Council Emotion Lexicon [49]
Southwick et al [33]	Reddit	Public	Not specified	Not specified	English	Yes, LDA[b] [32]	Yes, Affective Norms for English Words lexicon [50]
Miyake and Martin [51]	Twitter, Facebook, blogs, news posts on social media, Reddit, forums, and other platforms	Public	January 1, 2020, to January 1, 2021	United Kingdom	English	No	Yes, but only of hashtags and emojis using IBM Watson emotional lexicon
Fu [34]	Twitter	Public	March 26, 2022, to April 26, 2022	United States	English	Yes, LDA [32]	Yes, VADER[c] [38]
Ramakrishnan et al [52]	Twitter	Public	May 1 to Sep 30, 2021	Not specified	English	No	Yes, IBM Watson Tone Analyzer and 6 classical ML[d] algorithms

[a]This table summarizes previous studies on topic modeling and sentiment analysis across various platforms and time frames. It includes information on the source of data, the population being studied, the time period covered, the geographic location, the language used, and the specific methodologies applied for topic modeling and sentiment analysis.

[b]LDA: latent Dirichlet allocation.

[c]VADER: Valence Aware Dictionary for Sentiment Reasoning.

[d]ML: Machine learning.

We have used new methods for our tasks that, according to our knowledge, have never been used before in the context of long COVID analysis. First, we used contextualized topic modeling (CTM) [53], which takes advantage of the entire context of the text rather than LDA, which, by applying neural variational techniques, uses a bag-of-words representation of the text, meaning that it ignores the order of words and considers only the frequency of words. Also, we use Llama 2 [54] to represent the text and classify its sentiment rather than rule-based methods or classical machine learning methods.

In summary, this paper addressed 3 key knowledge gaps related to long COVID and methods for topic modeling and sentiment analysis. First, it investigated topics and sentiments related to long COVID across 3 regions (Canada, the United States, and Europe), providing an opportunity for comparative analysis of issues and public perceptions in different geographical contexts. Second, the study introduced new techniques, such as CTM and large language model (LLM) sentiment analysis, to obtain nuanced insights related to long COVID. These advanced methods surpass traditional approaches by capturing the context and subtleties of the discourse more effectively. Third, the study integrated expertise from public health and computer sciences, combining methodological rigor with domain-specific knowledge to interpret the results in a meaningful way. This interdisciplinary approach ensures that the analytical methods are aligned with public health objectives and implications. By addressing these gaps, the study contributes valuable insights into public health monitoring and intervention strategies, demonstrating the use of social media analysis in understanding complex public health issues such as long COVID.

The innovation in this study lies in the integration of CTM and an LLM (Llama 2) for sentiment analysis. CTM goes beyond traditional LDA by capturing word meanings in context, leading to more coherent and semantically meaningful topics. Using Llama 2 for sentiment analysis provides a nuanced understanding of sentiments expressed in tweets, leveraging the model's robust language comprehension capabilities.

Methods

Data and Data Processing

We used Twitter data for this analysis. In collaboration with public health experts working on long COVID, we identified relevant hashtags and keywords to extract Twitter data related to long COVID. Table 2 provides the list of hashtags and keywords that were used to collect the long COVID–relevant tweets.

XSL•FO
RenderX

Table 2. Terms used to collect long COVID–relevant tweets provided by health care professionals[a].

Term	Description
#Long COVID	This hashtag refers to the long-term effects and symptoms experienced by individuals after recovering from COVID-19.
#mecfs	Abbreviation for "Myalgic Encephalomyelitis/Chronic Fatigue Syndrome," a complex and debilitating condition characterized by extreme fatigue that is not alleviated by rest.
#MyalgicEncephalomyelitis	Also known as ME[b], this is a chronic condition characterized by profound fatigue, pain, sleep disturbances, and other symptoms.
#Fibromyalgia	A chronic disorder characterized by widespread musculoskeletal pain, fatigue, and tenderness in localized areas.
#PostViralSyndrome	A condition that occurs after a viral infection, characterized by persistent symptoms such as fatigue, joint pain, and cognitive issues.
#Dysautonomia	A disorder of the autonomic nervous system, which controls involuntary bodily functions such as heart rate, blood pressure, and digestion. Symptoms can include lightheadedness, fainting, and difficulty regulating body temperature.
#ChronicFatigueSyndrome	Also known as CFS[c], this is a complex disorder characterized by extreme fatigue that does not improve with rest and may be worsened by physical or mental activity.
#PwME	Abbreviation for "People with Myalgic Encephalomyelitis," a term used to refer to individuals living with ME/CFS.
#MyalgicE	An abbreviation for ME, a chronic condition characterized by extreme fatigue and other symptoms.
#PostCovidSyndrome	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.
#postcovid	A term used to describe symptoms or conditions that persist after recovering from COVID-19.
#PostViralFatigueSyndrome	A condition characterized by persistent fatigue and other symptoms following a viral infection.
#postviralillness	A term used to describe the lingering effects of a viral illness, which can include fatigue, muscle pain, and cognitive difficulties.
"Long Covid," "long COVID," and "long COVID syndrome"	A term used to describe the long-term effects of COVID-19, including persistent symptoms such as fatigue, shortness of breath, and cognitive issues.
"Post acute COVID"	A term used to describe the period of time after a COVID-19 infection where symptoms continue to persist.
"Post-COVID syndrome"	A term used to describe the syndrome of lingering symptoms and health issues experienced by individuals after recovering from COVID-19.
"Post-acute sequelae of SARS-CoV-2"	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.
"Long-term COVID"	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.
"Long haulers"	A colloquial term used to describe individuals who continue to experience symptoms and health issues after recovering from COVID-19.
"Chronic COVID syndrome"	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.

[a]This table lists the hashtags and keywords identified by public health experts to extract tweets related to long COVID from Twitter. Each term includes a description to explain its relevance to long COVID.

[b]ME: myalgic encephalomyelitis.

[c]CFS: chronic fatigue syndrome.

We focused on English-language tweets due to the availability and accessibility of large volumes of Twitter data in English, allowing for more accurate and reliable analysis. Moreover, this focus enables a comparative study across regions where English is the predominant language, reducing the complexity of multilingual analysis and ensuring the consistency of sentiment and topic modeling methods.

The hashtags in Table 2 were selected based on their relevance to long COVID as identified by public health experts. These experts provided terms frequently used in discussions related to long COVID, ensuring comprehensive data collection. However, we acknowledge that some terms such as #Fibromyalgia, #Dysautonomia, and #ChronicFatigueSyndrome may not exclusively refer to long COVID. To mitigate this, we applied additional filtering and manual inspection to confirm

XSL•FO
RenderX

that the majority of tweets were indeed related to long COVID. Despite these efforts, we recognize that some tweets may not be entirely relevant, and this limitation is addressed in the *Discussion* section.

For this study, we collected tweets alongside their metadata for the year 2021. We used the Twitter application programming interface [55] to extract the data, and afterward, we applied necessary preprocessing of the tweets. For preprocessing, we used the *tweet-preprocessor* toolkit, which applies cleaning, tokenizing, and parsing of URLs, hashtags, mentions, reserved words ("RT," which stands for "Retweet" and "FAV," which stands for "Favorite"), emojis, and smileys.

Among the 814,951 tweets extracted in total (Canada: n=98,796, United States: n=289,856, and Europe: n=426,299), we included only those tweets written in English using tweet metadata and the *spacy-langdetect* toolkit [56]. This process resulted in 782,089 tweets in total: 95,743 for Canada, 278,392 for the United States, and 407,954 for Europe including the United Kingdom. To remove tweet-specific keywords and URLs, we used the *tweet-preprocessor* toolkit [57]. We did not remove hashtags and mentions because they can be informative for our study. We lowercased and tokenized using the *Spacy* toolkit [58]. Since the methods we used in this paper are all unsupervised, we did not split the data for training and testing (we will share our data according to Twitter policy once this paper has been accepted).

Topic Modeling

In order to analyze the public perception of long COVID, we applied CTM. CTM uses pretrained representations of language (eg, bidirectional encoder representations from transformers) to support topic modeling based on the context. To identify the best number of topics that can be extracted from the tweets with the best coherence and diversity between the topics, we applied the following automatic evaluation metrics with different settings: for coherence, we used normalized pointwise mutual information [59], Cv [60], and word embedding [61], while for diversity, we used topic diversity and inverted rank-biased overlap [62]. These metrics evaluate the semantic consistency and distinctiveness of topics, ensuring meaningful results.

First, we applied the measures on 5, 10, 15, and 20 topics. However, the best topic coherence and diversity metrics results were obtained by extracting 5 topics. We tried the same metrics over 4 and 6 topics to see whether they would have better results, but they did not. Moreover, to confirm the automatic evaluation we have shown the 5 extracted topics to the health care professionals.

To assess changes in topics of discussion over time, we compared timelines of topic distributions across the 3 regions and examined how this reflected change in public perception of long COVID. To discover topics and track the topic change over time, we constructed topic models on our Twitter data using zero-shot CTM implementation.

We used the *pyLDAvis* tool [63] to visualize the 5 topics generated by the CTM model. *PyLDAvis* visualizes the relationship between topics using multidimensional scaling. It helps understand topic overlap and distinctiveness, aiding in the interpretation of topic coherence and diversity metrics. This methodological detail enhances the reliability of our topic modeling results. Two public health experts examined the visualization and one of them labeled each topic to make it easier to understand the context of each topic. More specifically, we performed a basic analysis based on an examination of the estimates of the vector, a document-to-topic distribution, produced by the model. We first divided tweets into weekly buckets using Coordinated Universal Time–12 time stamps (eg, January 21-26, January 27 to February 2, and February 3-9, 2021). We then computed a mean vector for tweets in each bucket as done by Griffiths and Steyvers [64].

Sentiment Analysis

Sentiment analysis was conducted to gauge the emotional tone expressed in tweets related to long COVID. The objective was to understand how individuals on Twitter perceived and communicated their sentiments regarding long COVID throughout the year 2021. To analyze the sentiments expressed in tweets related to long COVID, we used sentiment analysis at the tweet level. For tweet sentiment classification, we used Llama 2, which have been recognized as one of the most widely used LLMs recently. The robustness of Llama 2 for sentiment analysis is supported by recent studies demonstrating its effectiveness across various domains [65-68].

The accuracy of sentiment analysis using Llama 2 was validated through manual inspection of a random sample of classified tweets. Misclassification rates were low, demonstrating the model's robustness. Prompts were designed to capture the overall sentiment of each tweet accurately, considering context and phrasing nuances.

The sentiment analysis pipeline provided by Hugging Face's Transformers library is designed to analyze the sentiment of a given piece of text. Pipelines are high-level objects that allow users to easily apply pretrained models to various natural language processing (NLP) tasks. This specific pipeline uses the Llama 2 model for sentiment analysis, facilitating efficient processing of text data. Through tokenization, encoding, and model inference, the pipeline predicts sentiment, categorizing it as positive, negative, or neutral. This process offers valuable insights into the emotional tone of tweets related to long COVID.

Ethical Considerations

This study was conducted using publicly available Twitter data, which are anonymized and deidentified. The data collection and analysis were carried out in accordance with Twitter's terms of service. No personal or sensitive information was accessed, ensuring the privacy and confidentiality of Twitter users.

As the research involved publicly available, anonymized data from Twitter, it did not require formal ethics review approval. This approach is consistent with TCPS 2 (Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans), article 2.2, regarding secondary analysis of public data where no direct interaction with human subjects occurs.

Given that only publicly accessible data were analyzed, specific informed consent from Twitter users was not required. The

XSL•FO
RenderX

original consent provided by Twitter users for public sharing of their data under Twitter's terms of service allows for secondary analysis without additional consent.

All data analyzed in this study were anonymized and deidentified. No identifying information was collected, and data were handled in ways that fully comply with ethical guidelines on protecting participant confidentiality.

No compensation was provided to Twitter users, as the study used only publicly available data and did not involve direct participation or interaction with users.

Results

Topic Modeling

The 5 topics identified through topic modeling and a domain expert effort were as follows:

1. Long COVID in people including children in the context of vaccination (T1)
2. Duration and suffering associated with long COVID (T2)
3. Persistent symptoms of long COVID (T3)
4. The need for research on long COVID treatment (T4)
5. Measuring long COVID symptoms (T5)

A set of sample tweets associated with each topic was then examined to qualitatively validate the topic labels. The sample tweets provided contextual information that could not be inferred from the salient terms alone. Table 3 highlights the topics and their labels.

Table 3. Labels for top 5 prevalent topics in Canada, the United States, and Europe relating to long COVID[a].

Topic number	Salient terms (top 10 in italics)	Label based on salient terms	Label based on sample tweets
T1	"*people*," "*children*," "*kids*," "*covid*," "*long*," "*get*," "*many*," "*risk*," "*vaccinated*," "*deaths*," "know," "cases," "vaccine," "schools," "death," "even," "also," "still," "die," "amp," "term," "vaccines," "young," "school," "delta," "health," "protect," "masks," "spread," and "immunity"	Long COVID in people, including children, in the context of vaccination	More objective perspectives: research, monitoring, experience, and news items highlighting significant impact of symptoms of long COVID in adults and children, including potential benefits of vaccination in reducing long COVID risk
T2	"*day*," "*feel*," "*like*," "*back*," "*days*," "*year*," "*got*," "*time*," "*pain*," "*last*," "week," "feeling," "months," "since," "one," "better," "friend," "ago," "still," "felt," "work," "went," "hope," "years," "body," "go," "march," "going," "first," and "today"	Duration and suffering associated with long COVID	More subjective perspectives: suffering and frustration associated with long COVID
T3	"*symptoms*," "*19*," "*covid*," "*long*," "*study*," "*post*," "*patients*," "*term*," "*infection*," "*haulers*," "new," "covid19," "months," "syndrome," "coronavirus," "fatigue," "acute," "effects," "common," "studies," "viral," "via," "persistent," "reported," "sars," "found," "weeks," "experiencing," "brain," and "lingering"	Persistent symptoms of long COVID	Persistent symptoms of long COVID, with greater focus on formal and informal advocacy and awareness-raising around long COVID
T4	"*longcovid*," "*mecfs*," "*research*," "*please*," "*cfs*," "*support*," "*amp*," "*pwme*," "*help*," "*thank*," "patients," "community," "need," "share," "us," "longcovidkids," "funding," "sign," "patient," "awareness," "join," "millionsmissing," "myalgicencephalomyelitis," "treatments," "illness," "treatment," "guidelines," "nicecomms," "needs," and "thanks"	Need for research on long COVID treatment	Calls to address the plight of people with long COVID, either by researchers or by governments, or frustration at having been neglected, and frustration at having ongoing symptoms.
T5	"*fewer*," "*ontario*," "*vast*," "*matters*," "*measure*," "*amongst*," "*likelihood*," "*possibility*," "*polio*," "*contract*," "ha," "tracking," "victims," "de," "clinically," "eg," "vs," "admitted," "telegraph," "exposure," "estimate," "fda," "surge," "impacted," "reducing," "70," "american," "ages," "yo," and "largely"	Measuring long COVID symptoms	The large number of people affected by, or projected to be affected, by long COVID symptoms, and associated societal impacts.

[a]Derived from topic modeling of Twitter content from 2021, with labels for each topic based on human interpretation of keywords and associated sample tweets. This table shows the topics identified, the salient terms associated with each topic, and the final labels given to each topic based on the interpretation of sample tweets.

Through expert's evaluation, it became clear that the accurate and meaningful labels could not be consistently developed through salient terms alone, devoid of the context of the sample tweets from which they were derived. In particular, we observed that topic T1 tended to be associated with more objective perspectives on long COVID such as research, monitoring, objective descriptions of experiences, and news items, whereas T2 tended to be associated with more subjective perspectives such as suffering and frustration associated with having long COVID. In addition, the sample tweets suggested that T3 and T4 had advocacy for more to be done to address long COVID as a prominent element. Finally, T5 had societal impacts as a prominent element within the sample tweets.

Examination of sample tweets by the health care professionals also revealed overlap in the content between some of the topics. This is also reflected to some extent in the intertopic distances that were computed and visualized in Figure 1, which

XSL•FO
RenderX

demonstrates some intertopic overlap among topics T1-T4 but not for T5. Since *pyLDAvis* is an interactive visualization of the topics, we asked the health care professionals to explore the adjacent metric settings. By default, *pyLDAvis* is set for λ=1, which sorts words just by their frequency within the specific topic (by their red bars). By contrast, setting λ=0 words sorts words by their "lift." This means that words whose red bars are nearly as long as their blue bars will be sorted at the top.

Figure 1. Intertopic distance map via multidimensional scaling. This figure visualizes the distance between the 5 topics identified from long COVID–related tweets in 2021, showing intertopic overlap among topics T1-T4 but not for T5. The study includes tweets from Canada, the United States, and Europe. T1: Long COVID in people including children in the context of vaccination; T2: duration and suffering associated with long COVID; T3: persistent symptoms of long COVID; T4: need for research on long COVID treatment; and T5: measuring long COVID symptoms. In the figure, T1 is selected.

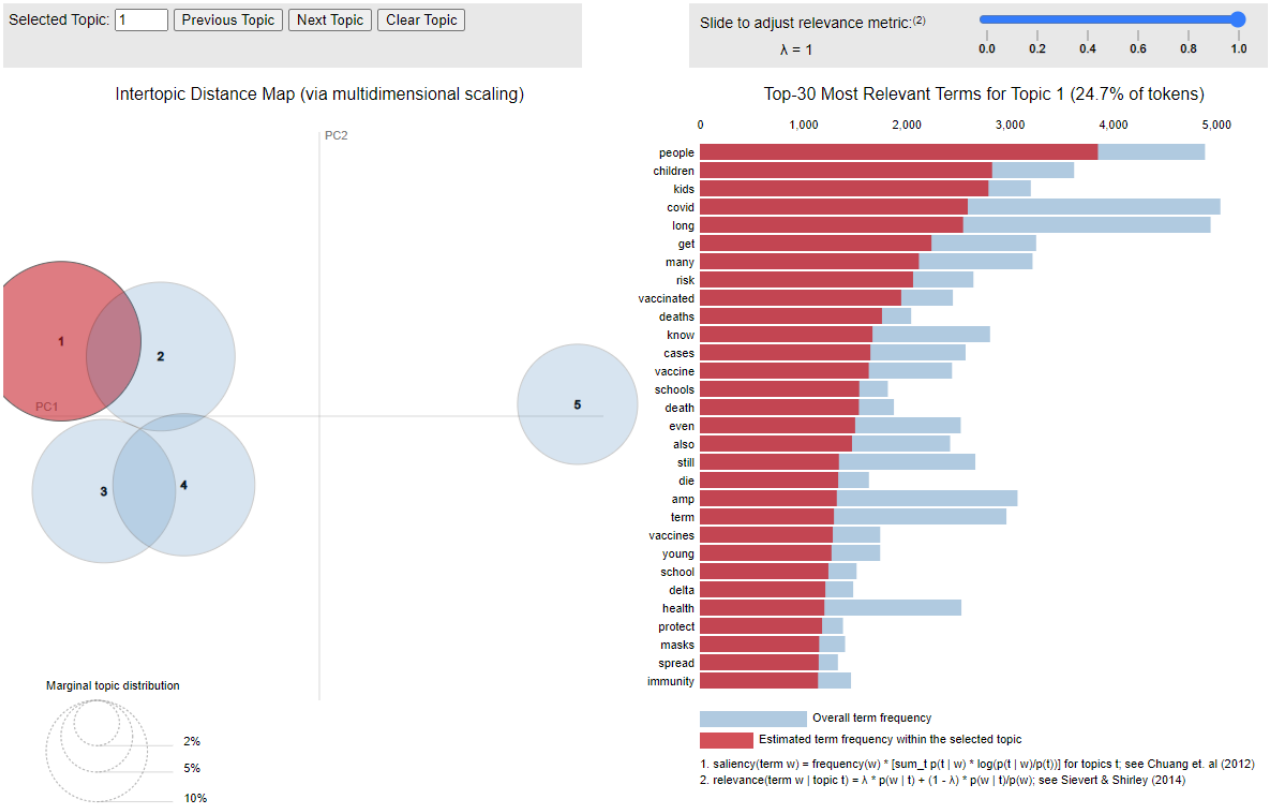

The types of discrepancies found between the salient terms–based topic labels and some of the sample tweets were due to low-label specificity, misrepresentation, or topic overlap. Some tweets were also found to be irrelevant to long COVID altogether; in particular, a portion of the sample tweets was irrelevant because of misclassification of long COVID–related tweets with tweets on COVID-19. Examples of these discrepancies are shown in Multimedia Appendix 1.

Based on the mean vector for each bucket, we drew graphs of long COVID topics over time as shown in Figures 2-7. We observed similar patterns between tweets in Canada and the United States. For example, the topic about subjective experiences around the duration and suffering associated with long COVID (T2) was less prominent early in the year but increased in prominence over time, peaking in late July-early August, and remained higher for the rest of the year, while there was no discernible trend in Europe. On the other hand, the topic on more objective perspectives on long COVID (T1) was relatively low in prominence in Europe compared with Canada and the United States. In contrast, the topic on persistent symptoms of long COVID (T3) was prominent throughout the year in all 3 regions.

XSL•FO
RenderX

Figure 2. Analyzing topic trends: visualizing the prominence of different topics over time through a bar chart for the Canada region. This figure shows the weekly distribution of the 5 identified topics in long COVID–related tweets in Canada during 2021.

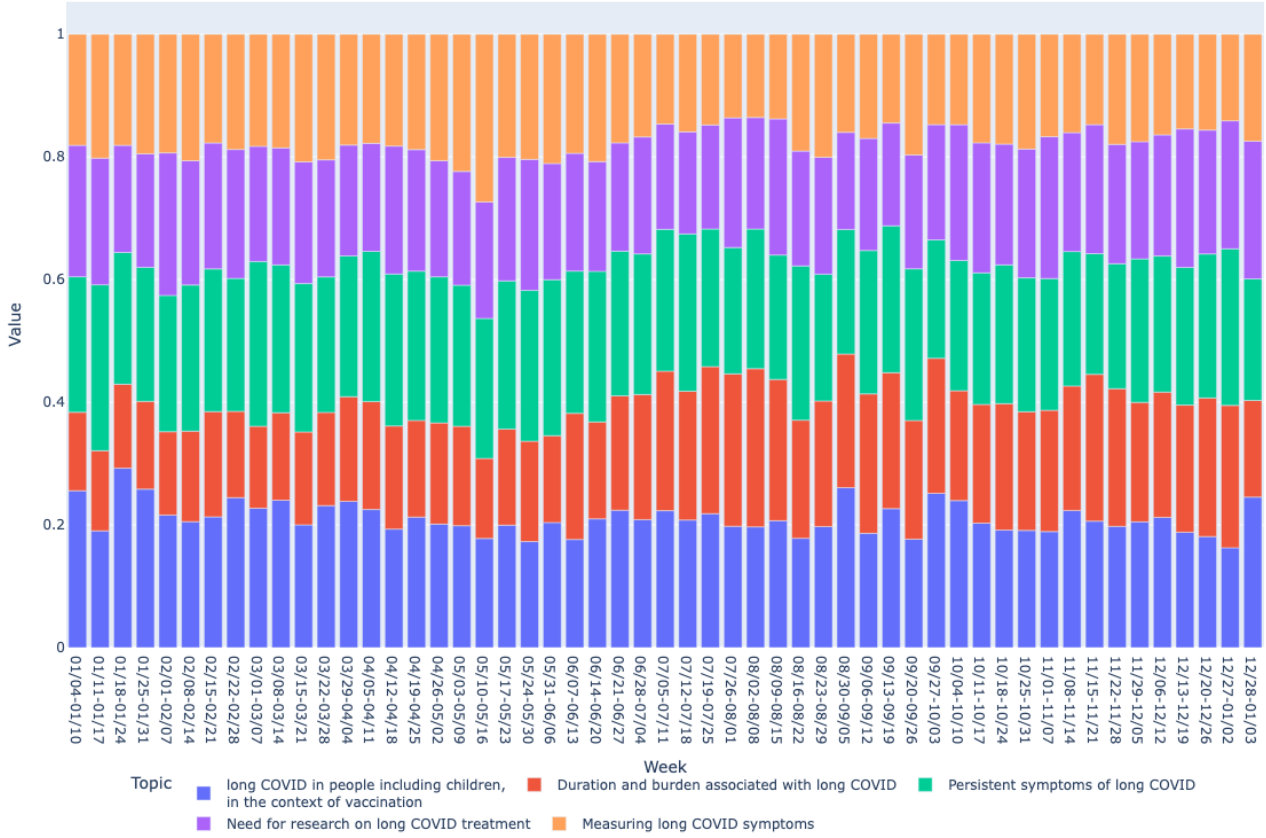

Figure 3. Analyzing topic trends: visualizing the prominence of different topics over time through a bar chart for the US region. This figure displays the weekly distribution of the 5 identified topics in long COVID–related tweets in the United States during 2021.

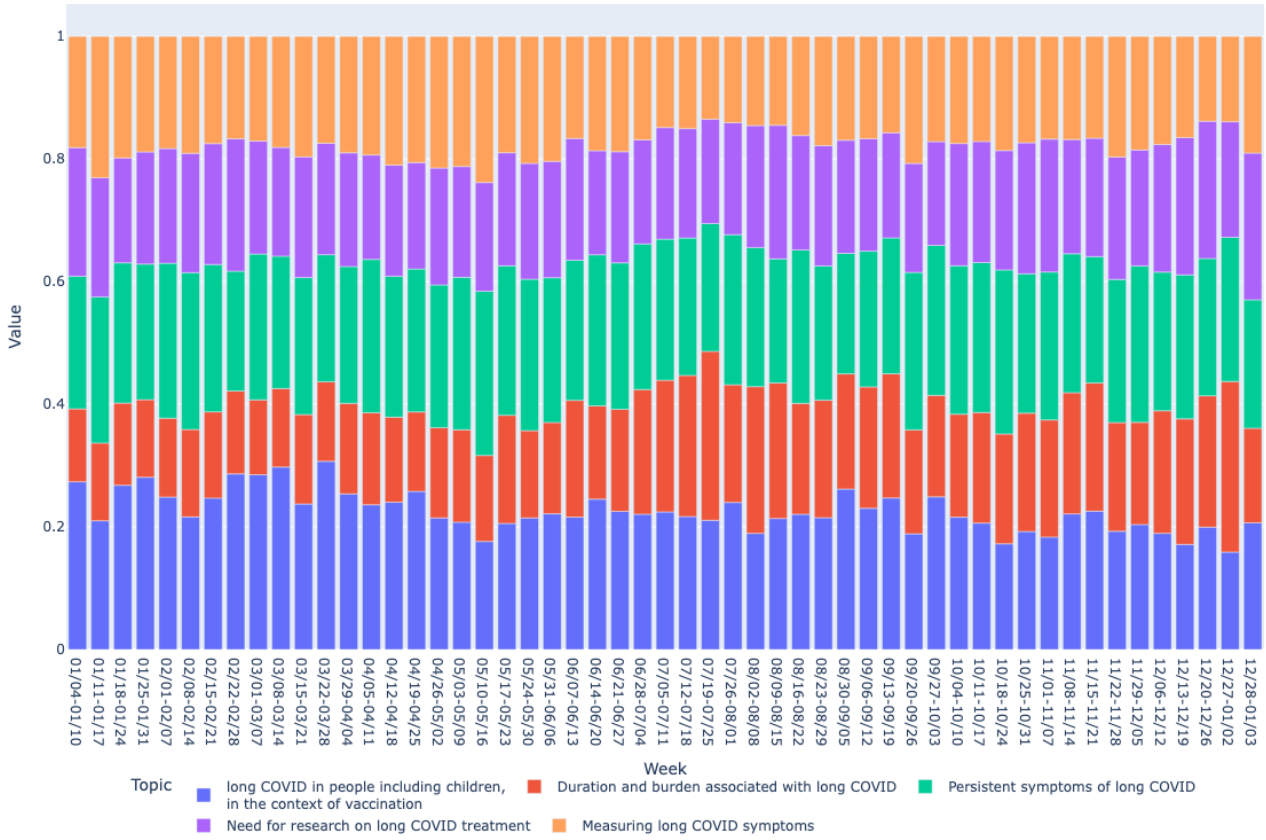

XSL•FO
RenderX

Figure 4. Analyzing topic trends: visualizing the prominence of different topics over time through a bar chart for Europe region. This figure depicts the weekly distribution of the 5 identified topics in long COVID–related tweets in Europe during 2021.

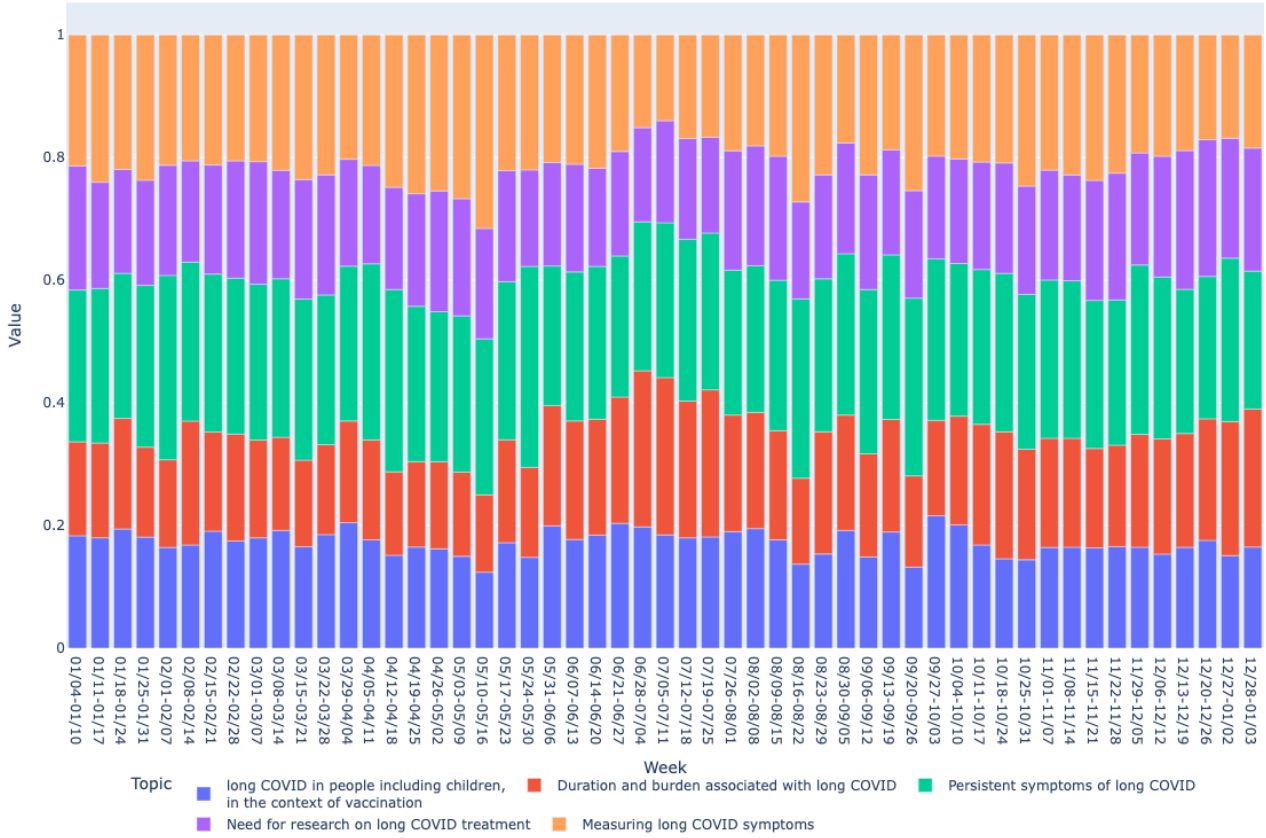

Figure 5. Analyzing topic trends: visualizing the prominence of different topics over time through a line chart for the Canada region. This figure shows the trends of the 5 identified topics in long COVID–related tweets in Canada over the year 2021.

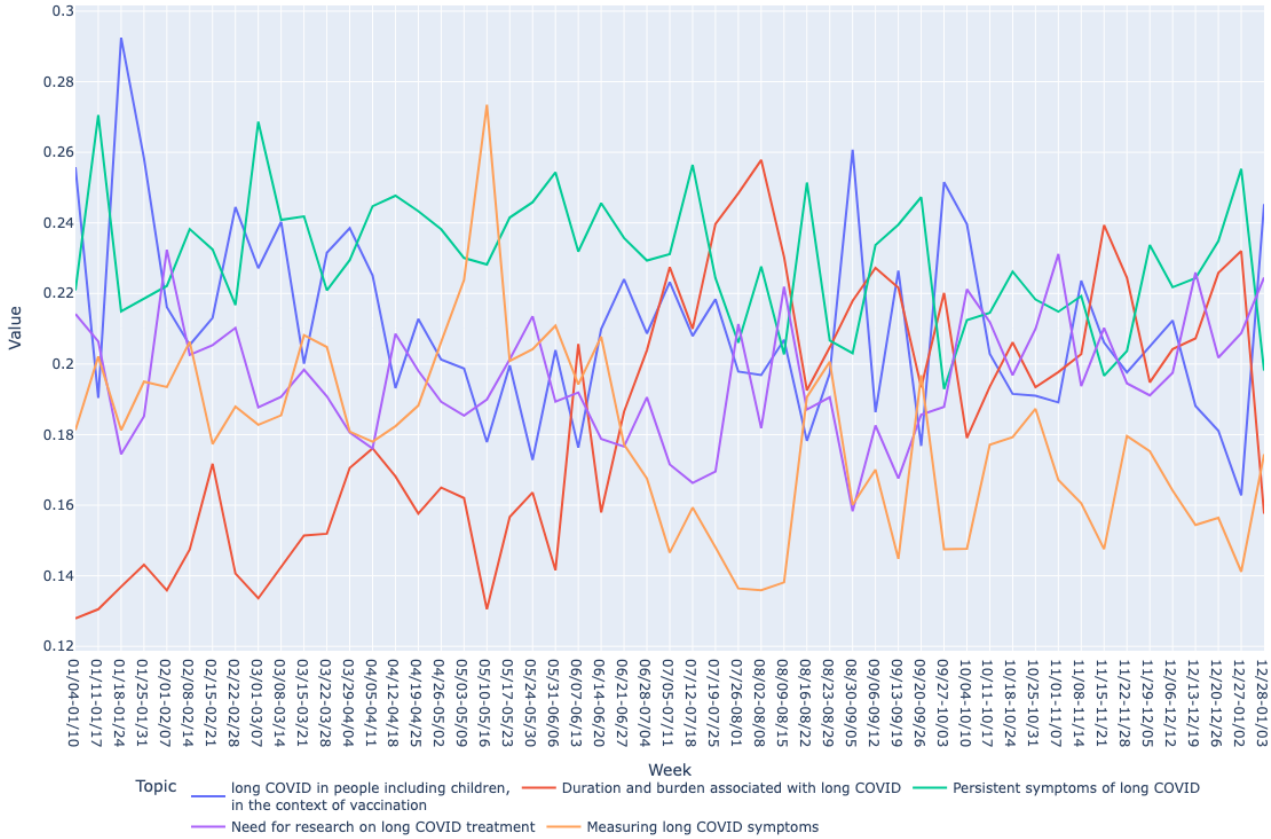

XSL•FO
RenderX

Figure 6. Analyzing topic trends: visualizing the prominence of different topics over time through a line chart for the US region. This figure illustrates the trends of the 5 identified topics in long COVID–related tweets in the United States over the year 2021.

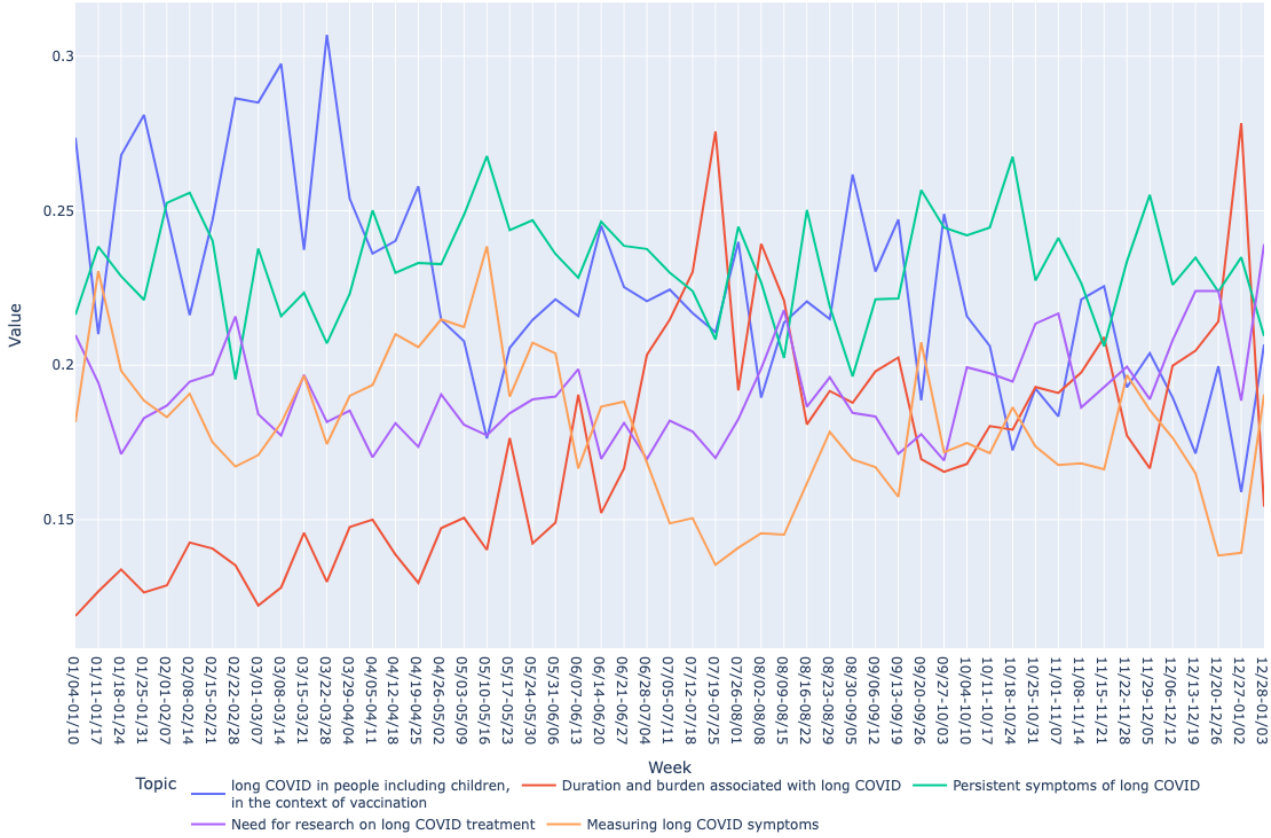

Figure 7. Analyzing topic trends: visualizing the prominence of different topics over time through a line chart for Europe region. This figure shows the trends of the 5 identified topics in long COVID–related tweets in Europe over the year 2021.

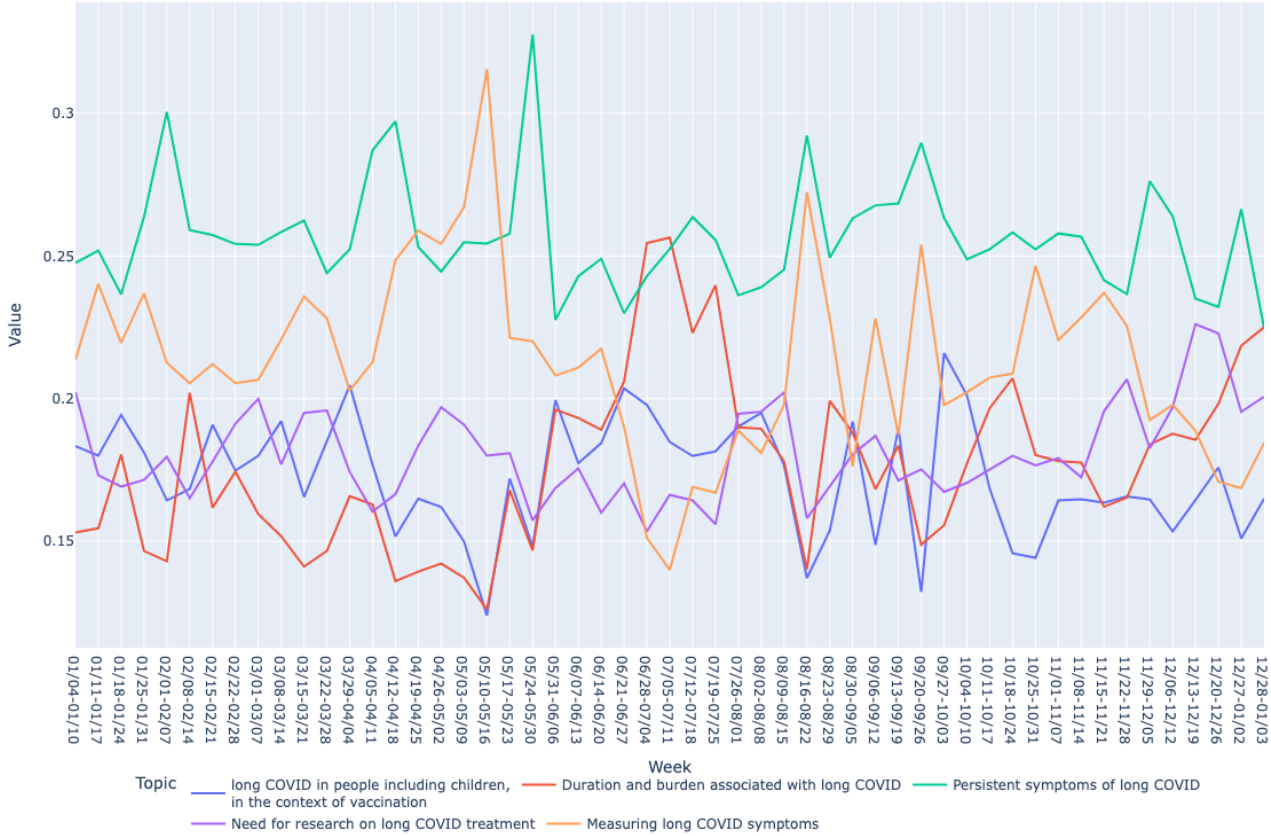

XSL•FO
RenderX

Both bar and line graphs were used to provide different perspectives on the data. Bar graphs offer a snapshot of weekly topic distribution, while line graphs illustrate trends over time, providing a comprehensive understanding of topic dynamics.

Sentiment Analysis

The sentiment distribution across long COVID–related tweets shows that there was no specific pattern. However, there was an increase in negative sentiment in July and August, followed by an increase in positive sentiment in September and again an increase in negative sentiment in December. In the United States, the trend was different, with a large amount of negative sentiment from January to March, a spike in July, and continued negative sentiment during August and September. April was notable for more positive sentiment and low negative sentiment. In particular, July had a lot of activity with some positive sentiment in addition to a large number of negative-sentiment tweets. The pattern in Europe was also different, with a large number of positive-sentiment tweets and substantial negative-sentiment tweets throughout the year.

There are many plausible explanations for these swings in positive and negative sentiments. For example, in July 2021, US President Joe Biden said that the long-term effects of COVID-19 can be considered a disability under federal civil rights laws, which may have spurred some positive sentiments [69]. In July, the National Institutes of Health also announced US $40 million for the study of long COVID and multisystem inflammatory syndrome in children [70]. In addition, multiple studies presented both positive and negative news about long COVID. All of these may have spurred both negative and positive discourse related to long COVID.

While we cannot establish a direct causal relationship, potential influences on sentiment swings can be hypothesized. Significant events, such as US President Joe Biden's statement in July 2021 that the long-term effects of COVID-19 could be considered a disability under federal civil rights laws, and the National Institutes of Health's announcement of US $40 million for the study of long COVID and multisystem inflammatory syndrome in children, likely contributed to these fluctuations. In addition, the presentation of both positive and negative news about long COVID in various studies may have further influenced public sentiment. These factors illustrate the complexity of inferring causality from observational data and highlight the multifaceted nature of public discourse.

Figures 8-10 show the count distributions of positive, negative, and neutral sentiments for the Canada, US, and European regions, respectively, providing a detailed view of the monthly sentiment counts. These figures help identify specific periods of increased positive or negative sentiments, which can be correlated with significant events.

Figures 11-13 show the percentage distributions of positive, negative, and neutral sentiments for the same regions, providing additional insights into the proportionate sentiment trends over the months. These figures complement the count-based visualizations by showing the relative proportions of each sentiment category, offering a more nuanced understanding of the sentiment landscape.

Figure 8. Sentiment analysis: visualizing the monthly sentiment counts produced by Llama 2 through a bar chart for the Canada region. This figure displays the count distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Canada during each month of 2021.

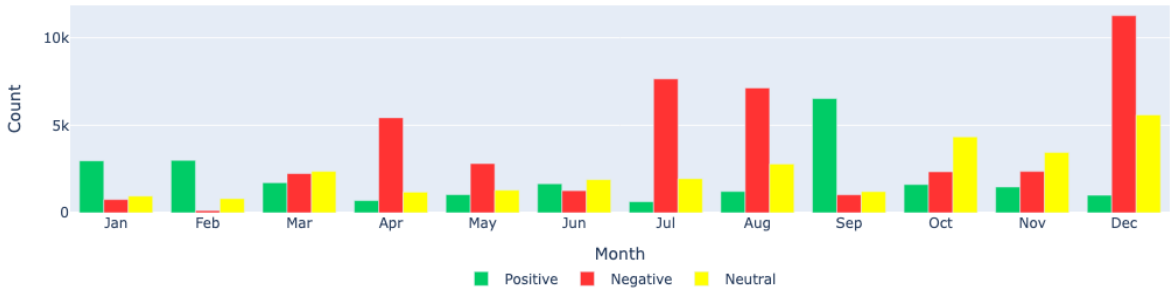

Figure 9. Sentiment analysis: visualizing the monthly sentiment counts produced by Llama 2 through a bar chart for the US region. This figure shows the count distribution of positive, negative, and neutral sentiments in long COVID–related tweets in the United States during each month of 2021.

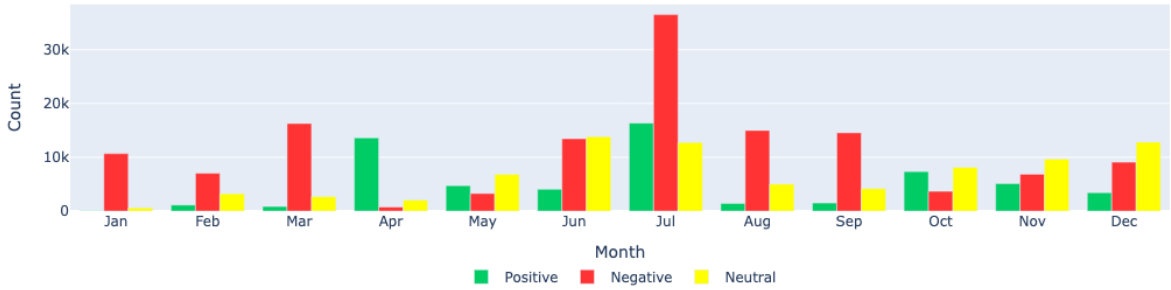

XSL•FO
RenderX

Figure 10. Sentiment analysis: visualizing the monthly sentiment counts produced by Llama 2 through a bar chart for Europe region. This figure illustrates the count distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Europe during each month of 2021.

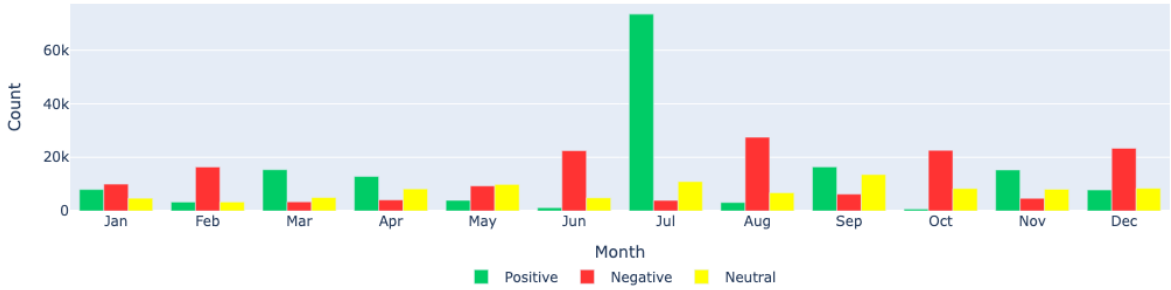

Figure 11. Sentiment analysis: visualizing the monthly sentiment percentages produced by Llama 2 through a bar chart for the Canada region. This figure displays the percentage distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Canada during each month of 2021.

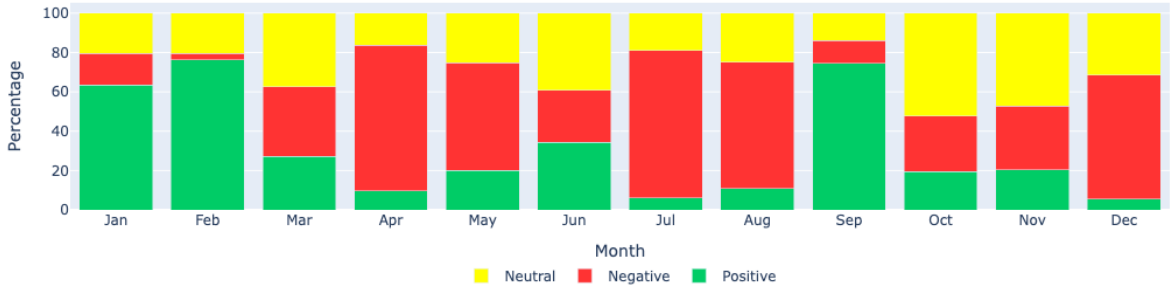

Figure 12. Sentiment analysis: visualizing the monthly sentiment percentages produced by Llama 2 through a bar chart for the US region. This figure shows the percentage distribution of positive, negative, and neutral sentiments in long COVID–related tweets in the United States during each month of 2021.

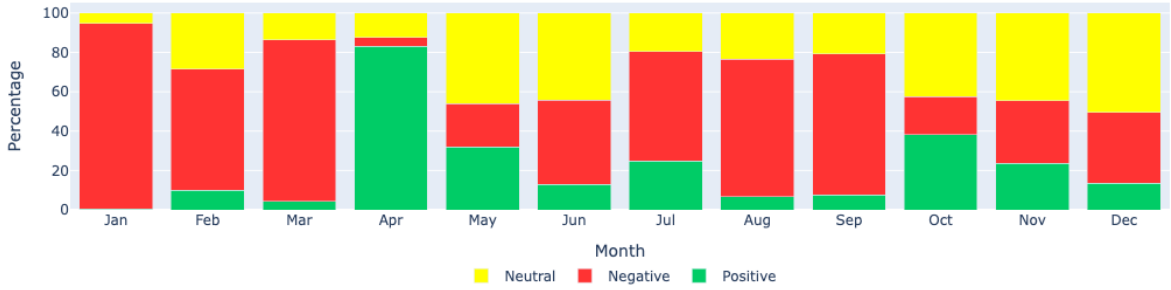

Figure 13. Sentiment analysis: visualizing the monthly sentiment percentages produced by Llama 2 through a bar chart for Europe region. This figure illustrates the percentage distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Europe during each month of 2021.

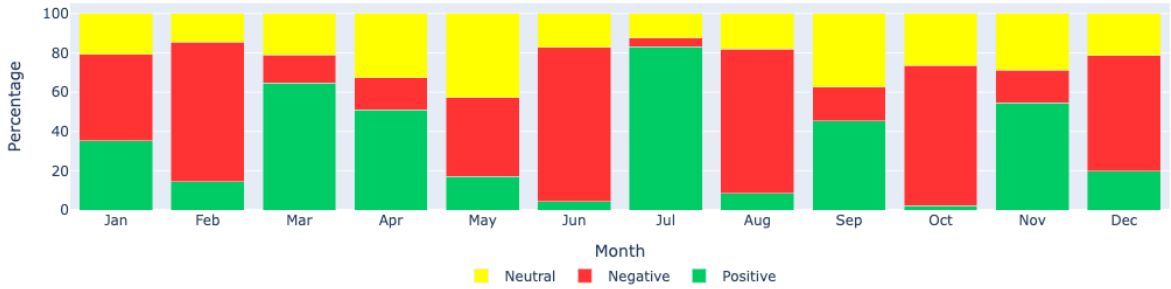

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we used topic modeling and sentiment analysis on Twitter data from Canada, the United States, and Europe in 2021. We identified 5 key topics: objective and subjective descriptions of the experience of people living with long COVID; persistent symptoms associated with long COVID; the need for more attention to be given to long COVID by researchers and governments; and measuring the impact of long COVID. The overlapping nature of these topics is likely due to the complex and ill-defined nature of long COVID, which shares symptoms with other conditions. Sentiment analysis showed a mix of positive and negative sentiments over time, with notable variations across different countries. The fluctuations in sentiment may be influenced by various events and news during the study period. For instance, misinformation such as claims that long COVID is merely psychological and not a real medical condition could lead to negative sentiments and confusion among the public. These insights into discussions and sentiments related to long COVID can help understand patient concerns and combat misinformation. The code used for data analysis is available in the study by AbuRaed [71].

The combined use of topic modeling and sentiment analysis in this study provided a richer, multidimensional view of the public discourse on long COVID. Topic modeling identified key themes such as persistent symptoms (T3) and the need for more research on treatments (T4), while sentiment analysis added an emotional layer to these discussions. For example, negative sentiments were consistently associated with topics about suffering and the call for more research, reflecting public frustration with the ongoing impact of long COVID. In contrast, topics such as vaccination (T1) often displayed more balanced or positive sentiments, indicating public optimism during vaccine rollout periods. By integrating these 2 approaches, we captured not only what people were discussing but also how they felt about these issues, offering a more comprehensive understanding of public sentiment and its fluctuations over time. This insight is crucial for public health communication and intervention strategies, as it highlights the emotional responses tied to different aspects of the long COVID discourse.

Most previous studies focused on either topic modeling or sentiment analysis related to long COVID but not both. In addition, many of these studies used LDA for topic modeling, which operates under the bag-of-words assumption and ignores word order. We used CTM, which captures word meanings in context, allowing for more coherent and semantically meaningful topics. For sentiment analysis, we used Llama 2, instead of rule-based or hybrid systems, demonstrating its robustness in sentiment classification. Our findings showed both consistencies and discrepancies compared with existing studies, underscoring the robustness of our identified topics and sentiment trends while also highlighting the nuanced capabilities of our methodology.

CTM addresses the limitations of traditional methods by capturing word meanings in context, enhancing topic coherence and relevance. This approach is particularly suited for complex and nuanced topics such as long COVID.

The topic labels, based on salient terms, were found to be inadequate upon examination of sample tweets. This discrepancy is expected due to the nonrepresentative nature of sample tweets and the overlapping representation among topics T1-T4, as indicated by the intertopic distances. Misinterpretation due to word-sense disambiguation challenges also contributed to misrepresentative labeling. Salient terms such as “like” or “long” are difficult to interpret without context, leading to irrelevance in the labels concerning long COVID.

The evaluation of topic labels involved expert review of sample tweets, ensuring that labels accurately reflected tweet content. This iterative process improves label accuracy and relevance. The methodology and results of this evaluation are detailed in the “Methods” and “Results” sections.

When examining topic prominence over time, it was challenging to identify clear patterns due to the overlapping nature of topics. However, there was an overall increasing trend in discussions around the duration and suffering associated with long COVID, consistent with the growing awareness and burden of long COVID as the pandemic progressed.

Public Health Implications

The results of this study have significant public health implications. Monitoring discussions and perceptions of people with long COVID can provide valuable insights into their needs and changing issues over time. Although geographical analysis is limited due to data constraints, this system can assess the impact of various actions and measures across jurisdictions. Furthermore, this approach can be applied to other emerging public health issues to monitor discourse, address concerns, and correct misinformation.

By identifying key topics and sentiment trends, our study can inform public health messaging and interventions to address misinformation. For example, if we observe a spike in negative sentiment and misinformation regarding the efficacy of long COVID treatments, public health officials can launch targeted information campaigns to provide accurate information and resources about available treatments and their effectiveness. Understanding public concerns and sentiment shifts allows for these targeted information campaigns to correct false narratives and provide timely support to affected individuals.

Limitations

There were several limitations to this study related to the data source, data collection and processing, the NLP methods used, and the approach to interpretation of the outputs.

With regard to the data source, a portion of the tweets were cut off, either because Twitter’s character limit per post forced the user to post multiple tweets to express their thoughts or because of the extraction process itself. In addition, some tweets are repeated multiple times (sometimes more than 80 times), because of retweets. Moreover, because of the design of the Twitter platform, such as character limits, and the intentions of Twitter users (eg, to grab people’s attention), the ideas expressed are abbreviated and unclear. We extracted only those tweets

XSL•FO
RenderX

that were in the English language and further limited tweets to those originating from Canada, the United States, and Europe. Thus, the data are unlikely to be representative of discussions and perceptions on long COVID elsewhere in the world. Finally, Twitter users constitute a small subset of the world population, further limiting the representativeness of these data.

With regard to data collection and data processing, the collection of Twitter data related to long COVID using long COVID–related keywords was subject to misclassification, since tweets relating to long COVID had significant overlap with tweets unrelated to long COVID. While this misclassification cannot be quantified without examining and categorizing each tweet one by one, an examination of a random sample of 164 tweets identified that approximately 13% (21/164) of the tweets were unrelated to long COVID. This misclassification may be in part because of the still poorly defined nature of long COVID as a condition, and in part a result of the long COVID–related keywords that were used to extract the tweets. For example, fibromyalgia, a chronic condition with similar symptomatology to long COVID, was discussed in some tweets that were irrelevant to long COVID.

With respect to NLP methods used to analyze the Twitter data, there was significant overlap between many of the topics in the topic model, as discussed above. In addition, challenges with word-sense disambiguation limit the interpretability of results.

With respect to the approach to interpretation of the outputs, having 1 domain expert to label the topics based on the sample tweets and a second domain expert to label the topics based on the salient terms would have helped enhance objectiveness in the labeling process. In this study, since the same domain expert did the labeling using these 2 approaches, there was likely some bias in how the labels were determined. Moreover, while no formal method was applied in this study to account for how public health measures impacted public perceptions on long COVID, the change in public health measures over time and region over the course of 2021 does constitute a potential source of bias and confounding in our analysis. For example, more stringent measures may prompt a more favorable perception toward government among those concerned about long COVID, since stringent measures may be perceived as part of an effort to reduce the risk to individuals not only of COVID-19 but also of long COVID. Conversely, less stringent measures may be perceived less favorably for similar reasons.

Conclusions

Our key findings include the identification of 5 main topics related to long COVID, significant regional variations in public discourse, and the fluctuating nature of sentiments influenced by major events. These findings provide a nuanced understanding of public concerns and the potential to inform public health strategies.

This study shows exploratory results from topic modeling and sentiment analysis on long COVID–related tweets in Canada, the United States, and Europe. By comparing results across these regions, we demonstrated changes in topic prominence over time. Public health domain experts played a crucial role in interpreting the results, emphasizing the importance of a human-in-the-loop approach in sentiment analysis. Although we identified some regional and temporal differences in topics, the main interpretation is the evolving public discourse and increasing awareness of long COVID throughout 2021.

The insights gained from this study can help public health officials and policy makers design more effective communication strategies and interventions tailored to the needs and concerns of those experiencing long COVID. By understanding the evolving public sentiment and key issues discussed on social media, public health responses can be better aligned with the population's needs. This research contributes to the broader understanding of how social media data can be leveraged for public health surveillance and intervention.

Future Directions

Future studies should explore more refined methods to enhance the specificity and accuracy of topic modeling and sentiment analysis. In addition, integrating more diverse social media platforms and languages could provide a more comprehensive understanding of global long COVID discourse. Finally, continuous monitoring and real-time analysis of social media data can further aid in understanding public perceptions and developing timely public health interventions.

Acknowledgments

The authors acknowledge the Canadian Institutes of Health Research (CIHR) and Michael Smith for Health System Impact Fellowship. The author thank Gaetano Cimino for providing feedback on the paper. This manuscript was prepared with the assistance of an artificial intelligence–based tool for language processing and editing, specifically ChatGPT by OpenAI [72], which was used to improve language clarity and coherence. All content and conclusions are the responsibility of the authors.

Conflicts of Interest

NZJ participated in advisory boards and has spoken for AbbVie and Gilead, not related to this work.

Multimedia Appendix 1
Sample tweets.
[DOCX File , 15 KB - jmir_v26i1e59425_app1.docx]

References

XSL•FO
RenderX

1. Coronavirus (COVID-19) dashboard. World Health Organization. URL: https://covid19.who.int [accessed 2024-06-30]
2. Altmann DM, Whettlock EM, Liu S, Arachchillage DJ, Boyton RJ. The immunology of long COVID. Nat Rev Immunol 2023;23(10):618-634. [doi: 10.1038/s41577-023-00904-7] [Medline: 37433988]
3. Binka M, Klaver B, Cua G, Wong AW, Fibke C, Velásquez García HA, et al. An elastic net regression model for identifying long COVID patients using health administrative data: a population-based study. Open Forum Infect Dis 2022;9(12):ofac640. [doi: 10.1093/ofid/ofac640] [Medline: 36570972]
4. Long-term effects of COVID-19. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html#:~:text=People%20with%20Long%20COVID%20can,can%20sometimes%20result%20in%20disability [accessed 2024-06-30]
5. Natarajan A, Shetty A, Delanerolle G, Zeng Y, Zhang Y, Raymont V, et al. A systematic review and meta-analysis of long COVID symptoms. Syst Rev 2023;12(1):88 [FREE Full text] [doi: 10.1186/s13643-023-02250-0] [Medline: 37245047]
6. Zhang H, Zang C, Xu Z, Zhang Y, Xu J, Bian J, et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. Nat Med 2023;29(1):226-235 [FREE Full text] [doi: 10.1038/s41591-022-02116-3] [Medline: 36456834]
7. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. J Med Internet Res 2020;22(4):e19118. [doi: 10.2196/19118] [Medline: 32302966]
8. Liu J, Nie H, Li S, Chen X, Cao H, Ren J, et al. Tracing the pace of COVID-19 research: topic modeling and evolution. Big Data Res 2021;25:100236. [doi: 10.1016/j.bdr.2021.100236]
9. Dong M, Cao X, Liang M, Li L, Liu G, Liang HY. Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling. medRXiv Preprint posted online April 10. [doi: 10.1101/2020.03.26.20044164]
10. Sussman KL, Bouchacourt L, Bright LF, Wilcox GB, Mackert M, Norwood AS, et al. COVID-19 topics and emotional frames in vaccine hesitation: a social media text and sentiment analysis. Digit Health 2023;9:20552076231158308 [FREE Full text] [doi: 10.1177/20552076231158308] [Medline: 36896330]
11. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. J Gen Intern Med 2020;35(7):2244-2247. [doi: 10.1007/s11606-020-05889-w] [Medline: 32399912]
12. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with twitter data in the United States: spatiotemporal perspective. J Med Internet Res 2021;23(9):e30854 [FREE Full text] [doi: 10.2196/30854] [Medline: 34346888]
13. Ridhwan MK, Hargreaves CA. Leveraging twitter data to understand public sentiment for the COVID‐19 outbreak in Singapore. Int J Inf Manag Data Insights 2021;1(2):100021. [doi: 10.1016/j.jjimei.2021.100021]
14. Huangfu L, Mo Y, Zhang P, Zeng DD, He S. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling. J Med Internet Res 2022;24(2):e31726 [FREE Full text] [doi: 10.2196/31726] [Medline: 34783665]
15. Beliga S, Martinčić-Ipšić S, Matešić M, Petrijevčanin Vuksanović I, Meštrović A. Infoveillance of the Croatian online media during the COVID-19 pandemic: one-year longitudinal study using natural language processing. JMIR Public Health Surveill 2021;7(12):e31540 [FREE Full text] [doi: 10.2196/31540] [Medline: 34739388]
16. Jafarzadeh H, Pauleen DJ, Abedin E, Weerasinghe K, Taskin N, Coskun M. Making sense of COVID-19 over time in New Zealand: assessing the public conversation using Twitter. PLoS One 2021;16(12):e0259882 [FREE Full text] [doi: 10.1371/journal.pone.0259882] [Medline: 34910732]
17. Maulana FI, Adi PDP, Lestari D, Purnomo A, Prihatin SY. Twitter data sentiment analysis of COVID-19 vaccination using machine learning. 2022 Presented at: 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI); December 8-9, 2017; Yogyakarta, Indonesia p. 582-587. [doi: 10.1109/ISRITI56927.2022.10053035]
18. Marcec R, Likic R. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgrad Med J 2022;98(1161):544-550. [doi: 10.1136/postgradmedj-2021-140685] [Medline: 34373343]
19. Kumar AHS, Shausan A, Demartini G, Rahimi A. Automatic identification of 5C vaccine behaviour on social media. 2022 Presented at: Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022); October 10, 2024; Gyeongju, Republic of Korea p. 136-146.
20. Akpatsa SK, Addo PC, Lei H, Li X, Dorgbefu M, Fiawoo DD, et al. Sentiment analysis and topic modeling of Twitter data: a text mining approach to the US-Afghan war crisis. SSRN Journal Preprint posted online on March 23, 2022. [doi: 10.2139/ssrn.4064560]
21. Guo Y, Zhu J, Huang Y, He L, He C, Li C, et al. Public opinions toward COVID-19 vaccine mandates: a machine learning-based analysis of U.S. tweets. AMIA Annu Symp Proc 2022;2022:502-511 [FREE Full text] [Medline: 37128441]
22. Maulana FI, Heryadi Y, Suparta W, Arifin Y. Social media analysis using sentiment analysis on COVID-19 from twitter. 2022 Presented at: 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE); December 13-14, 2022; Yogyakarta, Indonesia p. 286-290. [doi: 10.1109/icitisee57756.2022.10057932]

XSL•FO
RenderX

23. Masood A, Iqbal MM, Nayab A, Farooq M, Saeed MS. Impact of COVID-19 on human health using social media sentiment analysis. J Comput Biomed Inf 2023;5(01):41-51.
24. Upadhyay K. Topic modelling and sentiment analysis architecture for social networks in the COVID pandemic. Int J Comput Artif Intell 2022;3(2):56-59. [doi: 10.33545/27076571.2022.v3.i2a.55]
25. Sha H, Al HM, Mohler G, Brantingham P. Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. arXiv Preprint posted online on April 19, 2020 [FREE Full text]
26. Hosseini P, Hosseini P, Broniatowski DA. Content analysis of Persian/Farsi tweets during the COVID-19 pandemic in Iran using NLP. arXiv Preprint posted online on May 17, 2020 [FREE Full text] [doi: 10.18653/v1/2020.nlpcovid19-2.26]
27. Sharma K, Seo S, Meng C, Rambhatla S, Liu Y. COVID-19 on social media: analyzing misinformation in Twitter conversations. arXiv Preprint posted online on March 26, 2020 [FREE Full text]
28. Odlum M, Cho H, Broadwell P, Davis N, Patrao M, Schauer D, et al. Application of topic modeling to tweets as the foundation for health disparity research for COVID-19. Stud Health Technol Inform 2020 Jun 26;272:24-27 [FREE Full text] [doi: 10.3233/SHTI200484] [Medline: 32604591]
29. Wang X, Zou C, Xie Z, Li D. Public opinions towards COVID-19 in California and New York on Twitter. medRxiv Preprint posted online on July 14, 2020. [doi: 10.1101/2020.07.12.20151936] [Medline: 32699856]
30. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 2020;22(4):e19016. [doi: 10.2196/19016] [Medline: 32287039]
31. Ordun C, Purushotham S, Raff E. Exploratory analysis of Covid-19 tweets using topic modeling, UMAP, and DiGraphs. arXiv Preprint posted online on May 6, 2020 [FREE Full text]
32. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res 2003 Jan;3:993-1022. [doi: 10.7551/mitpress/1120.003.0082]
33. Southwick L, Guntuku SC, Klinger EV, Pelullo A, McCalpin H, Merchant RM. The role of digital health technologies in COVID-19 surveillance and recovery: a specific case of long haulers. Int Rev Psychiatry 2021;33(4):412-423. [doi: 10.1080/09540261.2020.1854195] [Medline: 33860736]
34. Fu YB. Investigating public perceptions regarding the long COVID on Twitter using sentiment analysis and topic modeling. Med Data Min 2022;5(4):24 [FREE Full text] [doi: 10.53388/mdm20220520024]
35. Yin J, Wang J. A Dirichlet multinomial mixture model-based approach for short text clustering. 2014 Presented at: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14); August 24-27, 2014; New York, NY p. 233-242. [doi: 10.1145/2623330.2623715]
36. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, et al. Conditional random fields as recurrent neural networks. 2015 Presented at: Proceedings of the IEEE International Conference on Computer Vision; June 20-23, 1995; Cambridge, MA p. 1529-1537. [doi: 10.1109/iccv.2015.179]
37. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. 2013 Presented at: Proceedings of the 22nd International Conference on World Wide Web; 2013 May 13-17; Rio de Janeiro, Brazil p. 1445-1456. [doi: 10.1145/2488388.2488514]
38. Hutto CJ, Gilbert EE. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: Eighth International Conference on Weblogs and Social Media (ICWSM-14); May 16, 2014; Ann Arbor, MI p. 216-225. [doi: 10.1609/icwsm.v8i1.14550]
39. Nielsen F. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. 2011 Presented at: Proceedings of the ESWC2011 Workshop on "Making Sense of Microposts": big things come in small packages; May 30, 2011; Crete, Greece p. 93-98 URL: http://arxiv.org/abs/1103.2903
40. An TK, Kim MH. A new diverse AdaBoost classifier. 2010 Presented at: Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence—Volume 01; October 23-24, 2010; Sanya, China p. 359-363. [doi: 10.1109/AICI.2010.82]
41. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
42. Haykin S. Neural Networks: A Comprehensive Foundation. Hoboken, NJ: Prentice Hall PTR; 1994.
43. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. In: Encyclopedia of Machine Learning. Boston, MA: Springer; 2010:713-714.
44. Suthaharan S, Suthaharan S. Support vector machine. In: Machine Learning Models and Algorithms for Big Data Classification: Thinking With Examples for Effective Learning. Boston, MA: Springer; 2016:207-235.
45. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018 [FREE Full text]
46. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-Based sentiment analysis. J Med Internet Res 2021;23(2):e25431. [doi: 10.2196/25431] [Medline: 33497352]
47. Déguilhem A, Malaab J, Talmatkadi M, Renner S, Foulquié P, Fagherazzi G, et al. Identifying profiles and symptoms of patients with long COVID in France: data mining infodemiology study based on social media. JMIR Infodemiology 2022;2(2):e39849. [doi: 10.2196/39849] [Medline: 36447795]

XSL•FO
RenderX

48. Bhattacharyya A, Seth A, Rai S. The effects of long COVID-19, its severity, and the need for immediate attention: analysis of clinical trials and Twitter data. Front Big Data 2022;5:1051386. [doi: 10.3389/fdata.2022.1051386] [Medline: 36588926]
49. Mohammad SM, Turney PD. NRC Emotion Lexicon. Ottawa, ON: National Research Council; 2013:234.
50. Bradley MM, Lang PJ. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Gainesville, FL: Technical Report C-1, the Center for Research in Psychophysiology, University of Florida; 1999:25-36.
51. Miyake E, Martin S. Long Covid: online patient narratives, public health communication and vaccine hesitancy. Digit Health 2021;7:20552076211059649 [FREE Full text] [doi: 10.1177/20552076211059649] [Medline: 34868622]
52. Ramakrishnan K, Balakrishnan V, Han GJ, Seong NK. Long COVID emotion analyzer: using machine learning approach. 2023 Presented at: Proceedings of the 9th International Conference on Computational Science and Technology; August 27-28, 2022; Johor Bahru, Malaysia p. 529-541. [doi: 10.1007/978-981-19-8406-8_42]
53. Bianchi F, Terragni S, Hovy D, Nozza D, Fersini E. Cross-lingual contextualized topic models with zero-shot learning. arXiv Preprint posted online on April 16, 2020 [FREE Full text]
54. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv Preprint posted online on February 27, 2023 [FREE Full text]
55. Twitter Developer Platform. URL: https://developer.twitter.com [accessed 2024-06-30]
56. spacy-langdetect 0.1.2. The Python Package Index. 2020. URL: https://pypi.org/project/spacy-langdetect/ [accessed 2020-06-15]
57. tweet-preprocessor 0.6.0. The Python Package Index. 2020. URL: https://pypi.org/project/tweet-preprocessor/ [accessed 2020-06-15]
58. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Scientometrics Research. URL: https://sentometrics-research.com/publication/72/ [accessed 2024-10-10]
59. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. 2015 Presented at: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; February 2-6, 2015; Shanghai, China p. 399-408. [doi: 10.1145/2684822.2685324]
60. Syed S, Spruit M. Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. 2017 Presented at: IEEE International Conference on Data Science and Advanced Analytics (DSAA); October 19-21, 2017; Tokyo, Japan p. 165-174. [doi: 10.1109/DSAA.2017.61]
61. Fang A, Macdonald C, Ounis I, Habel P. Using word embedding to evaluate the coherence of topics from twitter data. 2016 Presented at: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval; July 17-21, 2016; Pisa, Italy p. 1057-1060. [doi: 10.1145/2911451.2914729]
62. Bianchi F, Terragni S, Hovy D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. 2020 Presented at: 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; August 1-6, 2021; Online p. 759-766. [doi: 10.18653/v1/2021.acl-short.96]
63. Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. 2014 Presented at: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces; June 27, 2014; Baltimore, MD p. 63-70. [doi: 10.3115/v1/W14-3110]
64. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A 2004;101 Suppl 1(Suppl 1):5228-5235. [doi: 10.1073/pnas.0307752101] [Medline: 14872004]
65. Jeong C. Fine-tuning and utilization methods of domain-specific LLMs. arXiv Preprint posted online on January 1, 2024 [FREE Full text]
66. Pavlyshenko BM. Analysis of disinformation and fake news detection using fine-tuned large language model. arXiv Preprint posted online on September 9, 2023 [FREE Full text]
67. Yang H, Zhang Y, Xu J, Lu H, Heng PA, Lam W. Unveiling the generalization power of fine-tuned large language models. arXiv Preprint posted online on April 14, 2024 [FREE Full text]
68. Pavlyshenko BM. Financial news analytics using fine-tuned Llama 2 GPT model. arXiv Preprint posted online on August 24, 2023 [FREE Full text]
69. Biden administration recognizes long COVID as a disability under federal civil rights laws. The Washington Post. 2021. URL: https://www.washingtonpost.com/politics/biden-ada-long-covid-disability/2021/07/26/972f2a04-ee20-11eb-a452-4da5fe48582d_story.html [accessed 2024-06-30]
70. Children's National Hospital and NIAID launch study on long-term impacts of COVID-19 in kids. Children's National Hospital. 2021. URL: https://www.childrensnational.org/about-us/newsroom/2021/cnh-and-niaid-launch-large-study-on-long-term-impacts-of-covid-19-and-mis-c-on-kids [accessed 2024-06-30]
71. AbuRaed A. COVID-19 data and analysis. GitHub repository. URL: https://github.com/AhmedAbuRaed/covid19 [accessed 2024-06-30]
72. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2024-12-03]

XSL•FO
RenderX

Abbreviations

CTM: contextualized topic modeling
LDA: latent Dirichlet allocation
LLM: large language model
NLP: natural language processing

Edited by A Mavragani; submitted 12.04.24; peer-reviewed by A Yagahara, N Yiannakoulias; comments to author 22.06.24; revised version received 28.07.24; accepted 24.09.24; published 09.12.24.

Please cite as:
AbuRaed AGT, Prikryl EA, Carenini G, Janjua NZ
Long COVID Discourse in Canada, the United States, and Europe: Topic Modeling and Sentiment Analysis of Twitter Data
J Med Internet Res 2024;26:e59425
URL: https://www.jmir.org/2024/1/e59425
doi:10.2196/59425
PMID:

©Ahmed Ghassan Tawfiq AbuRaed, Emil Azuma Prikryl, Giuseppe Carenini, Naveed Zafar Janjua. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Toxicity on Social Media During the 2022 Mpox Public Health Emergency: Quantitative Study of Topical and Network Dynamics

Lizhou Fan[1], BS, PhD; Lingyao Li[1], BS, MS, PhD; Libby Hemphill[1], BA, PhD

School of Information, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Lizhou Fan, BS, PhD
School of Information
University of Michigan
105 S State St
Ann Arbor, MI, 48109
United States
Phone: 1 734 763 2285
Email: lizhouf@umich.edu

Abstract

Background: Toxicity on social media, encompassing behaviors such as harassment, bullying, hate speech, and the dissemination of misinformation, has become a pressing social concern in the digital age. Its prevalence intensifies during periods of social crises and unrest, eroding a sense of safety and community. Such toxic environments can adversely impact the mental well-being of those exposed and further deepen societal divisions and polarization. The 2022 mpox outbreak, initially called “monkeypox” but later renamed to reduce stigma and address societal concerns, provides a relevant context for this issue.

Objective: In this study, we conducted a comprehensive analysis of the toxic online discourse surrounding the 2022 mpox outbreak. We aimed to dissect its origins, characterize its nature and content, trace its dissemination patterns, and assess its broader societal implications, with the goal of providing insights that can inform strategies to mitigate such toxicity in future crises.

Methods: We collected >1.6 million unique tweets and analyzed them with 5 dimensions: context, extent, content, speaker, and intent. Using topic modeling based on bidirectional encoder representations from transformers and social network community clustering, we delineated the toxic dynamics on Twitter.

Results: By categorizing topics, we identified 5 high-level categories in the toxic online discourse on Twitter, including disease (20,281/43,521, 46.6%), health policy and health care (8400/43,521, 19.3%), homophobia (10,402/43,521, 23.9%), politics (2611/43,521, 6%), and racism (1784/43,521, 4.1%). Across these categories, users displayed negativity or controversial views on the mpox outbreak, highlighting the escalating political tensions and the weaponization of stigma during this infodemic. Through the toxicity diffusion networks of mentions (17,437 vertices with 3628 clusters), retweets (59,749 vertices with 3015 clusters), and the top users with the highest in-degree centrality, we found that retweets of toxic content were widespread, while influential users rarely engaged with or countered this toxicity through retweets.

Conclusions: Our study introduces a comprehensive workflow that combines topical and network analyses to decode emerging social issues during crises. By tracking topical dynamics, we can track the changing popularity of toxic content on the internet, providing a better understanding of societal challenges. Network dynamics highlight key social media influencers and their intentions, suggesting that engaging with these central figures in toxic discourse can improve crisis communication and guide policy making.

(J Med Internet Res 2024;26:e52997) doi:10.2196/52997

KEYWORDS

social media; network analysis; pandemic risk; health care analytics; infodemiology; infoveillance; health communication; mpox

XSL•FO
RenderX

Introduction

Background

The 2022 mpox outbreak was first reported in several countries in Europe and quickly became a global health crisis [1]. Mpox is a viral illness that can be transmitted from animals to humans or between humans through contact with blood or bodily fluids [2]. This disease was originally named "monkeypox" and was later renamed to reduce stigma and other issues during the 2022 outbreak [3]. This public health emergency was characterized by a large number of cases and a high rate of transmission, which posed a significant threat to public health globally. Health authorities and public health organizations were quick to respond to the outbreak, implementing measures to control the spread of the disease, providing health care and support to those affected, and making vaccination efforts to protect those who were not yet infected [4-6].

During the mpox public health emergency, social media platforms were used for public health communication and information sharing about the disease, its spread, and people's feelings about it. This led to both positive and negative consequences, with accurate information being shared alongside misinformation and toxic comments [7]. Toxicity on social media is prevalent during health crises, with many individuals spreading misinformation, fear, and hate [8,9]. This can undermine public health communication efforts and create confusion and fear among the public. In addition, toxicity on social media disproportionately impacts communities considered historically marginalized, exacerbating existing health disparities and making it difficult for these communities to access accurate and trustworthy information during a crisis. Homophobia and racism were common in mpox discussions [7].

Understanding the online toxic discourse during the 2022 mpox public health emergency is crucial for several reasons. First and foremost, studying toxicity in mpox discussions helps identify the factors contributing to the spread of misinformation, fear, and panic, which can exacerbate the public's response to the outbreak. Second, by analyzing the motives and patterns behind such toxic behavior, public health officials and researchers can develop effective communication strategies to counteract negativity and promote accurate information. Finally, understanding the prevalence and impact of toxic discourse allows us to explore the broader implications of online behavior on societal discourse and public opinion formation during health crises. To achieve this understanding, we chronicled original tweets, the online posts on the Twitter (Twitter, Inc) social media platform, from May to October 2022. We then used the Perspective application programming interface (API) [9,10] to identify *toxic tweets*, which are the rude, disrespectful, or unreasonable comments on Twitter that are likely to encourage individuals to leave conversations. We analyzed the toxic tweets to answer the following research questions:

- Aboutness—What are the toxic tweets about and how do they change through time?
- Diffusion—How do toxic tweets spread through online social networks?

The *aboutness* of toxic tweets highlights the connection between the evolution of toxic discourse topics and the motivations of the people posting these tweets. This analysis helps us understand the thoughts and concerns of the ongoing public health emergency for a significant portion of the public, including 330 million monthly users who are active on Twitter's social network [11]. The *diffusion* of *toxic tweets* is summarized as the retweets network and mentions network. The retweets network can inform who initiated or distributed toxic comments, while the mentions network shows who was frequently mentioned by other users and thus should be aware of toxicity dissemination. In particular, we followed the analytical framework of 5 dimensions, namely, *context*, *extent*, *content*, *speaker*, and *intent*. The 5 dimensions in characterizing toxic discourse, as adapted from the Rabat Plan of Action [12], are crucial for understanding the relationship between toxicity on social media and public health policy.

Our study aimed to unravel the complex dynamics of toxicity on social media during the 2022 mpox public health emergency using advanced computational techniques. Specifically, we aimed to identify the thematic structures, the *aboutness*, and the network behaviors, the *diffusion*, that perpetuate toxic discourse to understand how such narratives spread and the role of influential network actors in this process. Ultimately, our study sought to offer actionable insights that can help design effective interventions to mitigate toxicity on social media in future public health emergencies and other crisis situations.

Toxicity on Social Media

Toxicity on social media refers to rude, aggressive, and degrading attitudes and behaviors, which are exhibited in various forms, including harassment, bullying, or even the spread of hate speech and misinformation [13,14]. One of the primary causes of toxicity on social media is the anonymity and physical disconnect provided by the cyberspaces of online platforms. The online disinhibition effect magnifies the toxicity and facilitates the implementation of toxic ideas in daily life [15], which makes toxicity on social media a useful tool to anticipate extremes in public opinions and social dynamics. At the same time, toxicity on social media is easily contagious and can propagate quickly through social networks [16], where algorithms deployed by online social media platforms can contribute to the spread of toxic content by amplifying it and showing it to a larger audience [17].

Consequently, toxicity on social media can negatively impact the mental health and well-being of individuals, contributing to anxiety, depression, and low self-esteem, among other issues [18,19]. The aggravating toxicity on social media can exacerbate the culture of hate and division, causing harm to communities considered marginalized and making it difficult for racial and sexual minority individuals to engage in meaningful and productive online discourses and face-to-face activities [20].

Recent efforts in academia and industry aim to improve understanding of online toxicity and implement detection and moderation using socio-technical approaches. Researchers have gained insights into toxicity on social media from different perspectives. Guberman et al [21] quantified toxicity and verbal violence through crowdsourcing, which can be useful for the

XSL•FO
RenderX

moderation of toxic contents. Almerekhi et al [22] and Lwin et al [23] identified triggers of toxicity on social media and provided insights into the causes of toxic discussions by analyzing topical and sentiment shifts in interactions. Wijesirivardene et al [19] found that meaningful context of online conversations can help highlight or exonerate purported toxicity.

Benefiting from increasingly less expensive cloud storage and computing, social media platforms have also started developing and deploying toxicity on social media moderation applications. Perspective API developed by the Counter Abuse Technology team of Jigsaw and Google provides free access to toxic content detectors that aim at enabling healthy conversations and reducing toxicity and abusive behavior [9,10]. Similarly, the OpenAI moderation end point is a tool for checking content's compliance with OpenAI's content policy, including the prohibition of the generation of hateful, harassing, or violent content [24].

In addition to analysis and moderation of toxic content, other research calls for public engagement to tackle toxicity on social media, including encouraging individual responsibility and positive behaviors, as well as raising awareness and education [25,26]. Despite efforts to study online toxicity, it remains challenging to fully eliminate the impact of existing toxicity and prevent the spread of new toxic content. It is still imminent and meaningful to keep track, enhance comprehension, and intensify awareness of toxicity on social media.

Health Crisis Communications on Social Media

Health crisis communications on social media refer to the use of social media platforms to disseminate information and communicate during public health emergencies and crises [27]. Social media has become a vital tool for public health organizations and governments to communicate during crises and emergencies, as it can quickly and effectively reach a large, diverse audience, facilitate public engagement, and gather feedback [28,29]. Individual users of social media can also widely read others' opinions about a health crisis, freely express their feelings, and receive timely feedback [30].

However, there are also several challenges associated with using social media for health crisis communications. One of the main challenges is health-related toxicity on social media, which can lead to confusion and fear among the public. For instance, during the COVID-19 pandemic, there was a surge in misinformation, conspiracy theories, and discriminatory remarks, which not only hindered effective public health responses but also fueled stigma, discrimination, and even violence against certain groups [31,32]. In addition, the volume of information and the number of sources on social media can be overwhelming for the public, making it difficult for them to discern what information is reliable and relevant [33]. Without proper planning and intervention, information on social media can negatively influence health crisis communications and even result in infodemics [34,35]. Thus, public health organizations and governments should develop clear and consistent communication strategies for monitoring and responding to social media and provide accurate and trustworthy information to the public.

There are several guidelines that are helpful for public health organizations and governments to refer to. The Center for Risk Communication suggests six best practices in public health risk and crisis communication: (1) accepting and involving stakeholders as legitimate partners; (2) listening to people; (3) being truthful, honest, frank, and open; (4) coordinating, collaborating, and partnering with other credible sources; (5) meeting the needs of the media; and (6) communicating clearly and with compassion [36]. Recent research on the COVID-19 pandemic also suggests providing relevant, accurate, and sensitive information to key public groups to minimize communication noise and guide desirable coordinated actions [33]. While these principles are carefully written, it remains challenging to implement them in practice, especially due to the complexity of social media's role in health crisis communications.

Methods

Overview

We retrieved a large corpus of toxic online discourse on Twitter and applied computational methods for analysis. In this section, we describe the data and methods used in this paper (Figure 1). We first introduce the data retrieval process with *extent* as a content relevance filter, followed by a preliminary analysis of *context*. We then provide the details of methods for characterizing topical and network dynamics, supporting the comprehensive analysis of the *content*, *speaker*, and *intent* (of social media users).

XSL•FO
RenderX

Figure 1. Workflow overview. We first used external application programming interfaces (APIs) to retrieve and identify toxic Twitter data and mpox context data. We then applied topic modeling and network analysis methods to categorize temporal topical dynamics and cluster network communities. We also mapped the results in each step to the analytical framework. BERTopic: bidirectional encoder representations from transformers–based topic modeling technique; CNM: Clauset, Newman, and Moore; RQ: research question.

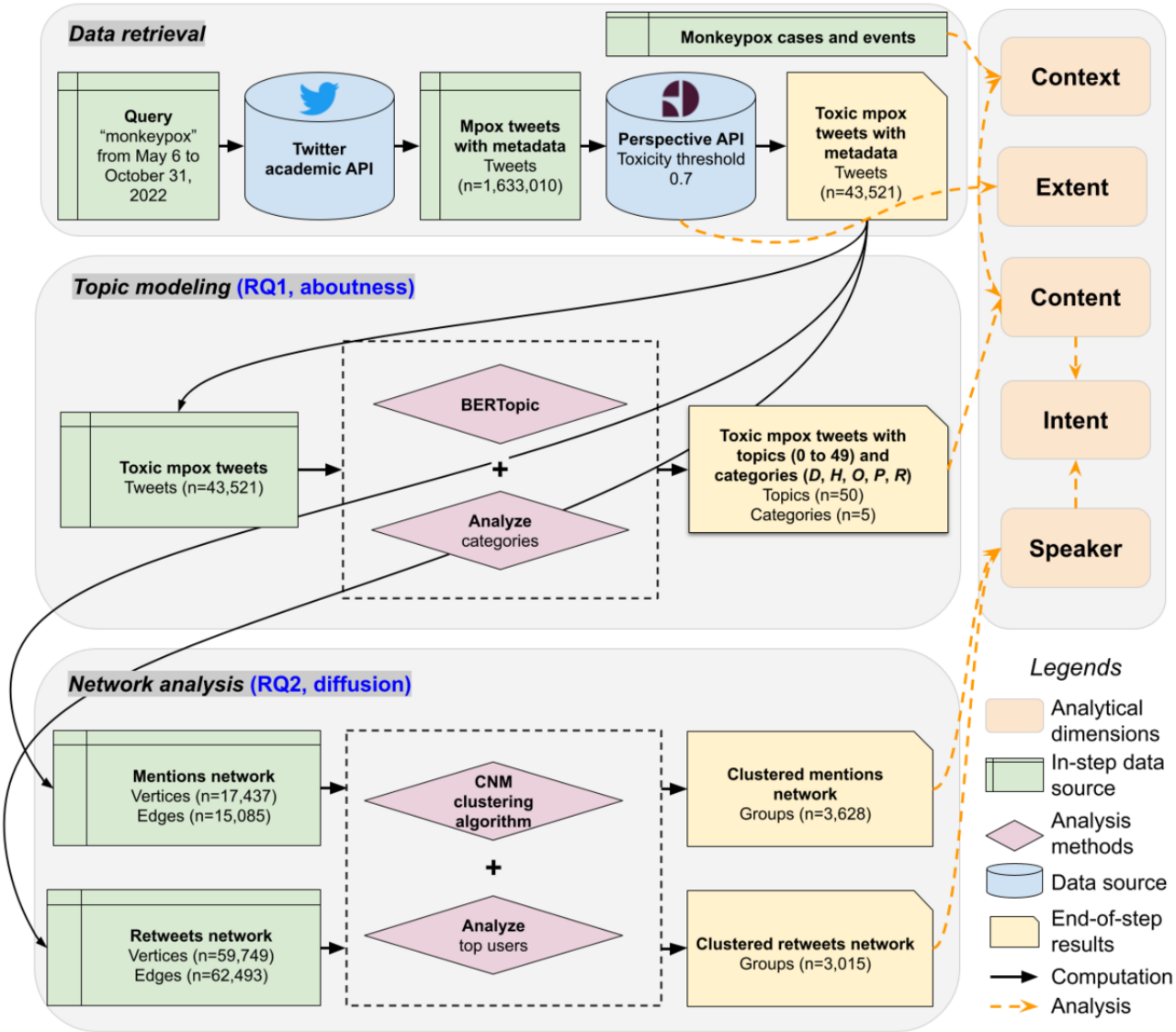

Ethical Considerations

This study uses publicly available Twitter data, strictly adhering to ethical guidelines and platform policies. All data were collected in compliance with Twitter's terms of service, and efforts were made to anonymize any identifiable information to protect user privacy. While Twitter users consent to the platform's terms, we recognize that this does not imply informed consent for research purposes; thus, the analysis of individual content was conducted with careful consideration to focus on public figures and avoid harm or misrepresentation. The research focuses on aggregate trends, ensuring that individual tweets are not decontextualized or used in a manner that could stigmatize individuals or groups. Finally, the findings are presented transparently, with an emphasis on societal benefit and respect for the original context of the data.

Data

In this section, we demonstrate the data collection process and the context of volume peaks.

Retrieval of Toxic Tweets

To chronicle online discourse on the 2022 mpox public health emergency, we used the Twitter Academic API [37] to query tweets with the keyword "monkeypox." Some relevant but less frequently used words, for example, "monkey pox" or "mpox," were not included for query simplicity and API efficiency. As such, the name change did not influence our data collection. We archived a collection of 1,633,010 unique English-language tweets started on May 6, 2022, when the initial cluster of cases was found in the United Kingdom [38] and ended on October 31, 2022. This study solely analyzed publicly available and anonymized social media data. No identifiable information about individual users was collected, stored, or analyzed. All data processing and analysis strictly adhered to Twitter's developer agreement and policy, ensuring compliance with ethical standards for privacy and confidentiality.

We then applied the Perspective API developed by the Counter Abuse Technology team of Jigsaw and Google to identify toxic tweets [9]. They define *toxicity* as a rude, disrespectful, or

XSL•FO
RenderX

unreasonable comment that is likely to disengage others' participation [9], especially those who are targeted by such toxicity. We adopted their definition of toxicity.

The Perspective API assigns each text submitted a probability score that corresponds to the proportion of people who would consider the text toxic. While choosing an appropriate threshold depends on the specific use case, the Perspective API team suggests that researchers experiment with a threshold between 0.7 and 0.9 to classify toxicity [9]. On the basis of our dataset, we observed that a tweet with a score ≥0.7 generally implied toxicity, and therefore, we chose 0.7 as the threshold to identify toxicity. When we limited our dataset to tweets that received a score of ≥0.7 from Perspective API, 43,521 toxic tweets remained.

Temporal Context and Hashtags

As shown in Figure 2 [1], we observed 2 substantial peaks in the volume of toxic public discourse on Twitter related to the mpox public health emergency. The first peak occurred from mid to late May, with a daily high of >1200 toxic tweets. This spike in toxic discourse coincided with the reporting of the initial cases of the 2022 mpox outbreak [39]. Notably, this peak preceded the World Health Organization (WHO) public acknowledgment of their tracking efforts of the global disease development of mpox by approximately 1 month.

Figure 2. Development of the 2022 mpox outbreak [1] and the volume of toxic tweets. CDC: Centers for Disease Control and Prevention; WHO: World Health Organization.

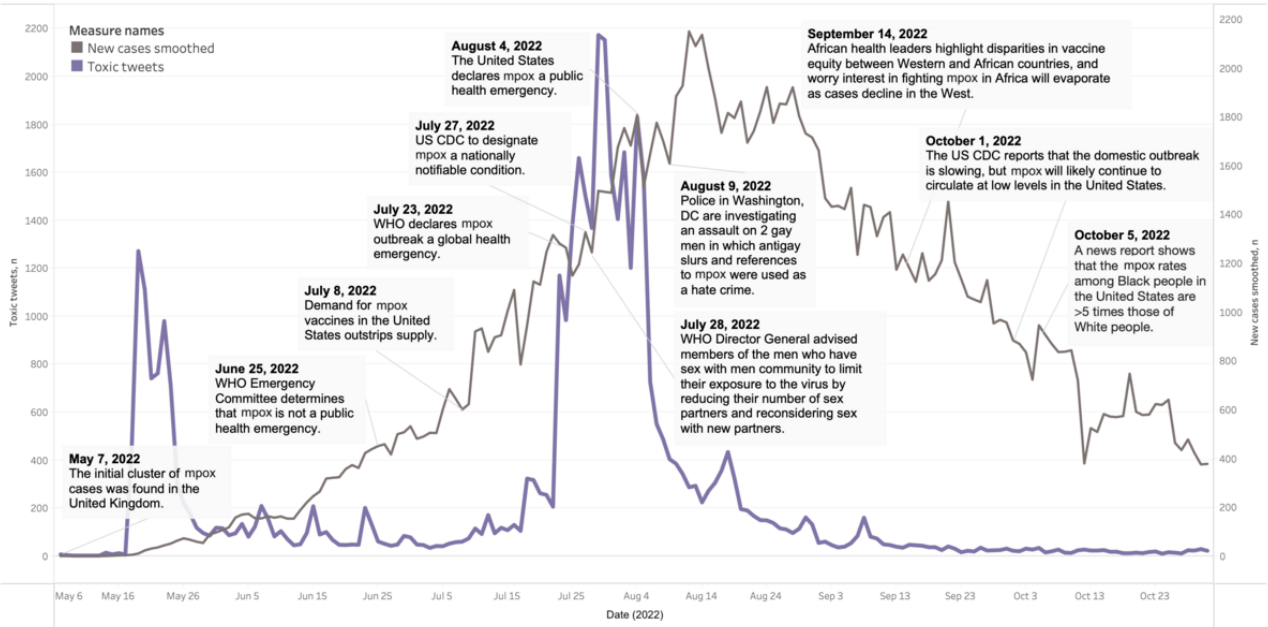

The second peak in toxic discourse was observed from late July to early August. This surge, which saw a daily high of approximately 2200 toxic tweets, nearly doubled the volume of the first peak. This increase followed 2 significant events: the WHO declared the mpox outbreak a global health emergency on July 23, 2022, and the Centers for Disease Control and Prevention designated mpox as a nationally notifiable condition on July 27, 2022 [1,2].

It is important to note that numerous mpox-related reports, particularly those with social and political implications, were released around and following the peak of new cases and the 2 peaks in toxic discourse volume. For example, on August 9, 2022, local police in Washington, DC, investigated an assault on 2 gay men as a hate crime; antigay rhetoric and references to mpox were used during the assault [40]. Then, on September 14 and October 5, 2022, there were reports on regional and racial disparities in health care, in particular, mpox vaccine equity, and health outcomes, that is, mpox rates [41,42]. These incidents underscore the profound real-world consequences of misinformation and toxic discourse, emphasizing the need for accurate and responsible communication on mpox and related issues.

Table 1 presents the summary of the top 20 hashtags in the toxic tweets, which highlights the frequent topics and provides a simplified overview of toxic online discourse on mpox. There is a diverse use of hashtags on different aspects of the mpox outbreak, including health emergency (eg, "pandemic" and "covid19"), health care (eg, "vaccine"), and population (eg, "lgbtq" and "gay"). These hashtags demonstrate the variety of topics in toxic tweets, which occur in different aspects of the online discourse.

XSL•FO
RenderX

Table 1. Top 20 hashtags used in toxic tweets (N=3742).

Hashtag	Count, n (%)
monkeypox	2889 (77.47)
monkeypoxvirus	106 (2.83)
covid19	102 (2.73)
covid	96 (2.57)
texasschoolmassacre	77 (2.06)
gay	76 (2.03)
who	60 (1.6)
lgbtq	50 (1.34)
aids	33 (0.88)
biden	30 (0.8)
cdc	30 (0.8)
billgatesbioterrorist	27 (0.72)
lgbt	23 (0.61)
trump	22 (0.59)
pandemic	22 (0.59)
vaccine	21 (0.56)
monkeypoxalypse	21 (0.56)
foxnews	20 (0.53)
pride	19 (0.51)
idiots	18 (0.48)

We also observed different types of entities in the hashtags relating to health administration organizations (eg, “who” and “cdc”), impactful individuals (eg, “biden” and “trump”), and news agencies (eg, “foxnews”). These hashtags indicate that further identification of users in this Twitter collection can help understand different stakeholders and participants in health and crisis information diffusion. Meanwhile, except for the query word “monkeypox,” all other hashtags were used <107 times. The variety of hashtags used and their low individual frequencies suggest that hashtags should not be the only source of identifying topics for this corpus. Alternative methods, such as topic modeling directly from the tweets, can be useful.

Analytical Methods

Overview

In this section, we provide the details of the analytical framework for toxicity on social media in our analysis. We also illustrate the characterization of topical and network patterns in the toxic tweets on mpox. We first demonstrate the technical details of topic modeling. We then discuss social network analysis methods for analyzing toxic information diffusion.

Adaptation of an Analytical Framework of Hate Speech Analysis for Toxicity on Social Media

We leveraged the hate speech analysis by the Rabat Plan of Action [12,43] and adapted it to the context of toxicity on social media. Our adapted analytical framework includes the following 5 dimensions:

- Context—the social and public health landscape behind toxic online discourse, including what
  events
  co-occur during the discourse and how
  public health metrics
  change relative to the volume and content trends
- Extent—the severity to which the message can be considered
  abusive or harmful
  to the targeted group, which can be assessed with a score from 0 to 1
- Content—the semantic summary of toxic online discourse, which reveals
  attributes of the targeted group
  (eg, vulnerability, political representation, and social construct) and the discourse’s
  co-occurrence with other narratives
  that are dominant in toxic discourse (ie, major semantic clusters in the corpus)
- Speaker—the status of the social media user who posts toxic content, which can influence the
  dissemination quantity
  (indicated by network metrics) and
  quality
  (depending on the user’s credibility, influence, and capacity)
- Intent—the assumed high-level summary of
  objectives
  and
  intended audience

XSL•FO
RenderX

for creating and spreading toxic content

These 5 dimensions are fundamental for our analysis and can benefit different stakeholders. Analyzing the context of toxic discourse can help policy makers identify key events that may be contributing to the proliferation of harmful messages, enabling them to address misinformation and foster a more supportive public health environment. Assessing the extent of toxic messages can help public health officials allocate resources and target interventions to counter the most severe cases of online abuse. Evaluating the content of toxic discourse reveals the attributes of targeted groups and dominant narratives, which can inform the development of tailored public health campaigns and interventions. Examining the speaker dimension provides insights into the dissemination of toxic content, allowing officials to monitor influential sources and mitigate their impact. Finally, understanding the intent behind toxic content can help public health policy makers craft strategies to engage with diverse audiences and counteract the harmful consequences of such discourse. By examining these 5 dimensions collectively, public health officials can gain a comprehensive understanding of the online toxic landscape during health crises, allowing them to devise timely and effective policy interventions. Table 2 further compares these 5 dimensions for analyzing toxic information diffusion with the original hate speech analysis framework by the Rabat Plan of Action [12].

Table 2. A comparison of analytical frameworks.

Dimension	Original definition	Definition for toxicity on social media in health crises	Analysis of adaptation
Context	Context denotes the social, cultural, and political landscape where the target of the hate speech is vulnerable.	It is defined as the social and public health landscape behind toxicity on social media, including what *events* co-occur during the discourse and how *public health metrics* change relative to the volume and content trends.	We limited the scope to social and public health and refocused the context from only the targeted group to the health crisis–related sociality.
Extent	Extent denotes the magnitude of the dissemination efforts or the extent of the hate speech act.	It is defined as the severity to which the message can be considered abusive or harmful to the targeted group, which can be assessed with a score from 0 to 1.	We measured the semantic intensity instead of the diffusion extent. We used Perspective API[a] [9] to score toxicity. Diffusion is analyzed through the concept of "Speaker" in our definition.
Content	Content and form include the provocative degree or aggressiveness of the message, the form taken by the expression, directness, call to action degree, correlation with other dominant hate narratives, and legal status.	It is defined as the semantic summary of toxic online discourse, which reveals attributes of the targeted group (eg, vulnerability, political representation, and social construct) and the discourse' co-occurrence with other narratives that are dominant in toxic discourse (ie, major semantic clusters in the corpus).	We kept 2 relevant subdimensions from the original definition.
Speaker	This dimension denotes the influence the speaker has on the audience to whom the SMS text message has been presented, including status, capacity, credibility, and influence on the targeted group.	It is defined as the status of the social media user who posts toxic content, which can influence the dissemination quantity (indicated by network metrics) and quality (depending on the user's credibility, influence, and capacity).	We summarized the original definition into 2 aspects, namely, quantity and quality, which are about the speaker's influence.
Intent	The intent of the speaker is estimated from past actions, reactions after promoting the hate message, probable objectives, and the intended audience.	It is defined as the assumed high-level summary of objectives and intended audience for creating and spreading toxic content.	We directly used the original definition in our context.
Likelihood of immediate actions	This dimension denotes the likelihood of the speech act generating a situation that represents a clear and immediate danger to the targeted social group, which is useful to evaluate as being sufficiently extreme to require a criminal investigation of censorship from state institutions.	__[b]	Our purpose was not to intervene in toxicity on social media as independent researchers but rather to identify and analyze them. Thus, the abovementioned 5 dimensions were sufficient.

[a]API: application programming interface.

[b]Not applicable.

Modeling Topics in Toxic Online Discourse

Topic modeling is a method for detecting and analyzing latent semantic topics from large volumes of unstructured text data. It assumes that each text document (eg, a tweet) is a combination of multiple latent topics, where each topic is represented by a probability distribution of words while representing topics by grouping together words that have similar meanings based on their probability distributions [44,45]. Topic modeling identifies groups of words or vectors that appear together, and those

XSL•FO
RenderX

groups are referred to as "topics." They are not necessarily topics in the colloquial sense of a "subject" or "theme." Identifying the content themes within and across topics requires manual inspection of the topics produced by the model. We refer to this step as "categorizing" and manually identified 5 themes ("categories") that capture all 50 topics. Topic modeling and other semantic presentation methods have been used as big data analysis tools in a variety of fields of research, including social media studies, health informatics, and crisis informatics [46-48].

With the development of deep learning techniques, such as transformers [49] and bidirectional encoder representations from transformers (BERT) [50], recent topic modeling methods take advantage of the embedding-based approach that better represents semantic relationships among words. These algorithms approach topic modeling as a clustering task and provide flexible language representation and text mining options [51,52]. As shown in Figure 3, our study follows this state-of-the-art development in topic modeling. We implemented a human-computer hybrid methodology sequence including both computational steps (in purple frames) and a human step (in a gray frame) for modeling topics in toxic online discourse during the mpox public health emergency. We followed the default steps and setting in BERT-based topic modelling technique (BERTopic) [52], which is a neural topic modeling method with a class-based term frequency-inverse document frequency (c-TF-IDF) procedure. We also extended the standard procedure by adding the preprocessing and the categorizing steps.

Figure 3. An extended sequence of steps with bidirectional encoder representations from transformers–based topic modeling technique (BERTopic). c-TF-IDF: class-based term frequency-inverse document frequency; SBERT: Sentence–bidirectional encoder representations from transformers; UMAP: uniform manifold approximation and projection.

The first part of the BERTopic methodology sequence is modeling topics and clustering tweets. We started with preprocessing the toxic mpox tweets by removing Twitter-related characters, including "@" and "RT" marks for social media networks and links starting with "http" for external web information. We then used Sentence-BERT, a transformer-based pretrained natural language processing (NLP) model, to derive semantically meaningful sentence embeddings for each of the cleaned tweets [53]. In particular, we used the Sentence-BERT Python package and the pretrained model "all-MiniLM-L6-v2" [54], which enables clustering and semantic searching by mapping search tweets to a 384-dimensional vector space, making it effective for semantic similarity tasks. To better handle the high-dimensional tweet vectors for clustering, we implemented a dimensionality reduction technique (uniform manifold approximation and projection [UMAP]) [55]. UMAP helps cluster models handle dimensionality [56] while maintaining a dataset's local and global structure. This feature of UMAP is important for constructing topic models, which depend on word vectors' structural similarities.

We then used the scikit-learn implementation of the k-means clustering algorithm by Lloyd to group similar sentences' embedding vectors into topics [57]. We used k-means clustering because it ensures that every vector is clustered into a topic and allows to select the number of clusters through experimentation. It first initiates centroids in the 384-dimensional vector space and randomly assigns a centroid to each of the 43,521 vectors. It then uses Euclidean distance [58] to update the centroid assignments recursively, which stops when the centroid assignments no longer update. We experimented with 3 different numbers of clusters: 30, 50, and 100. A total of 30 clusters produced clusters with a mix of topics and prevented us from disambiguating them. A total of 100 clusters were produced, resulting in sparse clusters that required manual combination of many topics. We settled on 50 clusters as a compromise. While the sentence embeddings remained the same, the clustering results could slightly vary with different random seeds initiated in the background. As there were only trivial differences among the clusters of sentences, we picked one of the results for our analysis. We shared the model on Hugging Face (Hugging Face, Inc) [59].

The second part of the sequence is representing topics and categories. We first tokenized topics using the count vectorizer in the *scikit-learn* Python package, which performs cluster-level (topic-level) bag-of-words representation that calculates and vectorizes the frequency of each word in each cluster [60]. In this study, we obtained word frequency vectors for each topic for representation purposes (ie, extracting keywords in each topic). This is fundamentally different from bag-of-words topic modeling, which uses corpus-level frequencies for creating topics. We then used c-TF-IDF to extract the difference of topical keywords, which helped distinguish among the clusters. After converting each cluster (topic) into a single document, we extracted the frequency of word x in class c [52]. In c-TF-IDF, we then had the importance score per word in each class:

(1)

where $tf_{X,C}$ is the frequency of word in class c, f_x is the frequency of word x across all classes, and A is the average number of words per class. In this way, we were able to represent topics with unique and frequent words as the keywords.

We then characterized the 50 topics based on the keywords and original tweets. After reviewing related literature, iterative refining categories, and labeling samples to study each category,

XSL•FO
RenderX

we annotated each topic with the following 5 categories: disease, health policy and health care, homophobia, politics, and racism.

Measuring User Influence in the Toxic Tweet Networks

Understanding how toxic information spreads on social media during public health crises is critical. Relationships, or user interactions, in social networks are often used to facilitate understanding of information diffusion in infodemiology [61]. We focused on 2 types of relationships on Twitter: mentions and retweets. A mention (ie, @username) is a tweet that quotes another user's name in the text. The user who is mentioned will receive a notification from Twitter. A retweet is a reposting of a tweet that starts with "RT @username" [62]. We calculated 3 measures of influence in the network, including in-degree centrality, outdegree centrality, and betweenness centrality, with close-degree centrality as a representative metric. Degree centrality refers to the number of edges a vertex has to other vertices, and it defines 3 types of centrality [63]. In our study, we particularly focus on (1) in-degree centrality, which measures the number of incoming connections a node has, indicating its popularity or influence within the network; and (2) betweenness centrality, which assesses the extent to which a node acts as a bridge along the shortest path between other nodes. Given a network $G=(V,E)$ with V vertices and E edges (defined as $deg(v)$), in-degree can be computed as [63] $deg_{1(v)} = N_V$ **(2)**

where N_V denotes the total number of all the incoming edges into vertex v. The betweenness of vertex v in a network is the fraction of all shortest paths between every pair of other vertices (s, t) that pass through vertex v. This is computed in three steps: (1) for each pair of vertices (s, t), shortest path between them is computed; (2) for each pair of vertices (s, t), the fraction of the shortest paths that pass through the vertex v is determined; and (3) the fraction over all pairs of vertices (s, t) is summed. Then, the betweenness of the vertex is calculated as [63]:

(3)

where σ_{st} represents the total number of shortest paths from vertex s to vertex t, $\sigma_{st}(v)$ is the number of those shortest paths that pass through vertex v. Further details of network analysis and metric calculation are provided in Multimedia Appendix 1.

Next, we applied a tool called NodeXL (Social Media Research Foundation) to generate the social network [64]. NodeXL is a visualization tool for social network analysis that is implemented as an add-on in Excel (Microsoft Corp) [64]. We further applied the Clauset, Newman, and Moore (CNM) algorithm [65] to investigate the social network in communicating toxicity on Twitter. The CNM algorithm infers the community structure from network topology that works by optimizing the modularity. It also provides insights into how vertices in social networks function and affect each other. One issue addressed by the CNM algorithm was to understand opinion leaders (eg, opinion leaders had high in-degree centrality) and distributors (eg, distributors had high betweenness centrality) in disseminating information.

As such, we generated social networks using mentions and retweets and investigated the following metrics in each network. The statistics for social networks M and R are summarized in Table 3 with the following attributes:

- Vertices—Twitter users in the social network
- Edges—relationships between 2 Twitter users (ie, retweet and mention)
- Duplicated edges—mention or retweet multiple times between 2 same Twitter users
- Self-loops—users mention or retweet their own tweets that form self-loops
- Connected components—a set of users in the network who are linked to each other by edges (ie, clusters in the social network)
- Geodesic distance—the length of the number of edges of the shortest path between 2 Twitter users (ie, 2 vertices in the network)

Table 3. Social network statistics for mentions and retweets networks.

Network metric	Mentions network	Retweets network
Network type	Directed	Directed
Vertices, n	17,437	59,749
Total edges, n	15,085	62,493
Duplicated edges, n	750	1663
Unique edges, n	14,335	60,830
Connected components, n	3628	3015
Self-loops, n	5	375
Maximum geodesic distance, n	26	21
Average geodesic distance, n	8.041	5.336

XSL•FO
RenderX

Results

Topic Modeling and Categorization

In this section, we report the topic modeling and categorization results, including the overall composition, temporal patterns, and representative tweets in each category. We summarize the results of the 50 topics into 5 *toxicity categories*, that is, the toxicity about 5 topical discourses, including disease (20,281/43,521, 46.6%), health policy and health care (8400/43,521, 19.3%), homophobia (10,402/43,521, 23.9%), politics (2611/43,521, 6%), and racism (1784/43,521, 4.1%).

Figure 4 shows the daily volume trend and weekly composition trend as well as the composition overview of categories in toxic tweets during the mpox outbreak (from May 6 to October 31, 2022). The overall trend indicates that a wide range of topics exist in the semantic dispersion of tweets: except for the first week when only 2 categories of discourse occurred, we observed no discontinuity in any category, which shows strong topical diversity in toxic discourse.

Figure 4. Overview of toxicity categories and change over time.

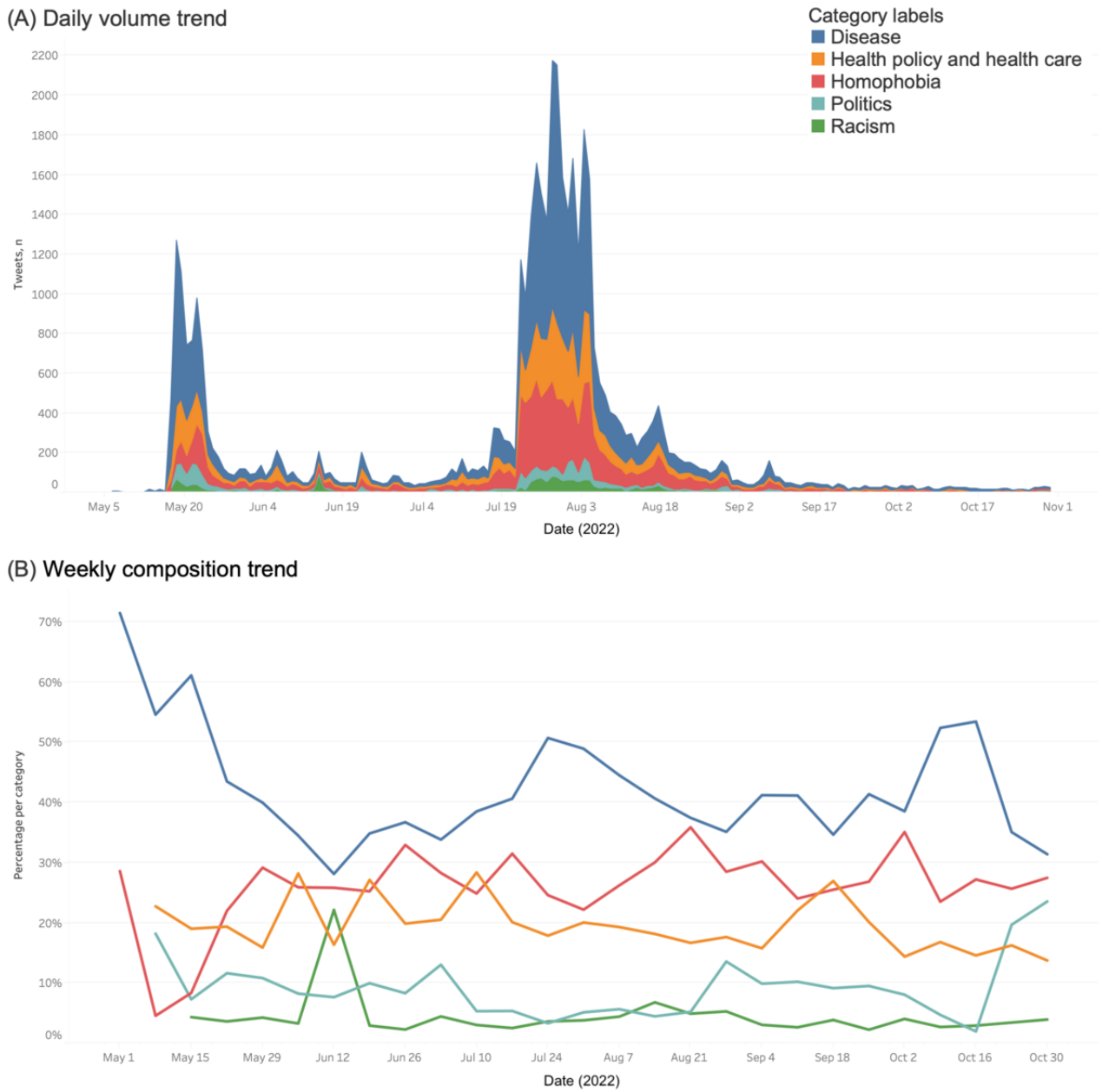

Disease, health policy and health care, and homophobia were the dominant categories that had higher and comparatively stable compositions (often >10% of the overall discourse). At the same time, the politics and racism categories had lower daily volumes (often <100 tweets) and weekly compositions (often <10% of the discourse). These 2 categories also had more fluctuations in volume: there were continuous large discourses in the politics category between May 20 and June 4, 2022, and in August as well as a composition increase toward the end of October. There was also a large discourse in the racism category around May 20, 2022, and June 12, 2022, and between July 19 and August

XSL•FO
RenderX

23, 2022, with the cluster around June 12, 2022, comprising >20% of the discourse of that week.

Figure 5 demonstrates the interdocument and intercategory distances based on the UMAP mapping of tweets to the semantic spaces of topics in a 2D visualization [55], where the vertical and horizontal dashed lines are the axes of the 2 dimensions. The semantic space of toxic tweets is divided into several clusters, primarily based on the summative categories of Twitter topics. The overall dispersion of the topics shows that the related topics belonging to the same category are close to each other, which indicates the comprehensiveness of the topic modeling and the categorization process. The collocation of topics further indicates internal relevance among toxic tweets of different topical focuses.

Figure 5. A 2D visualization of toxic tweets with categories.

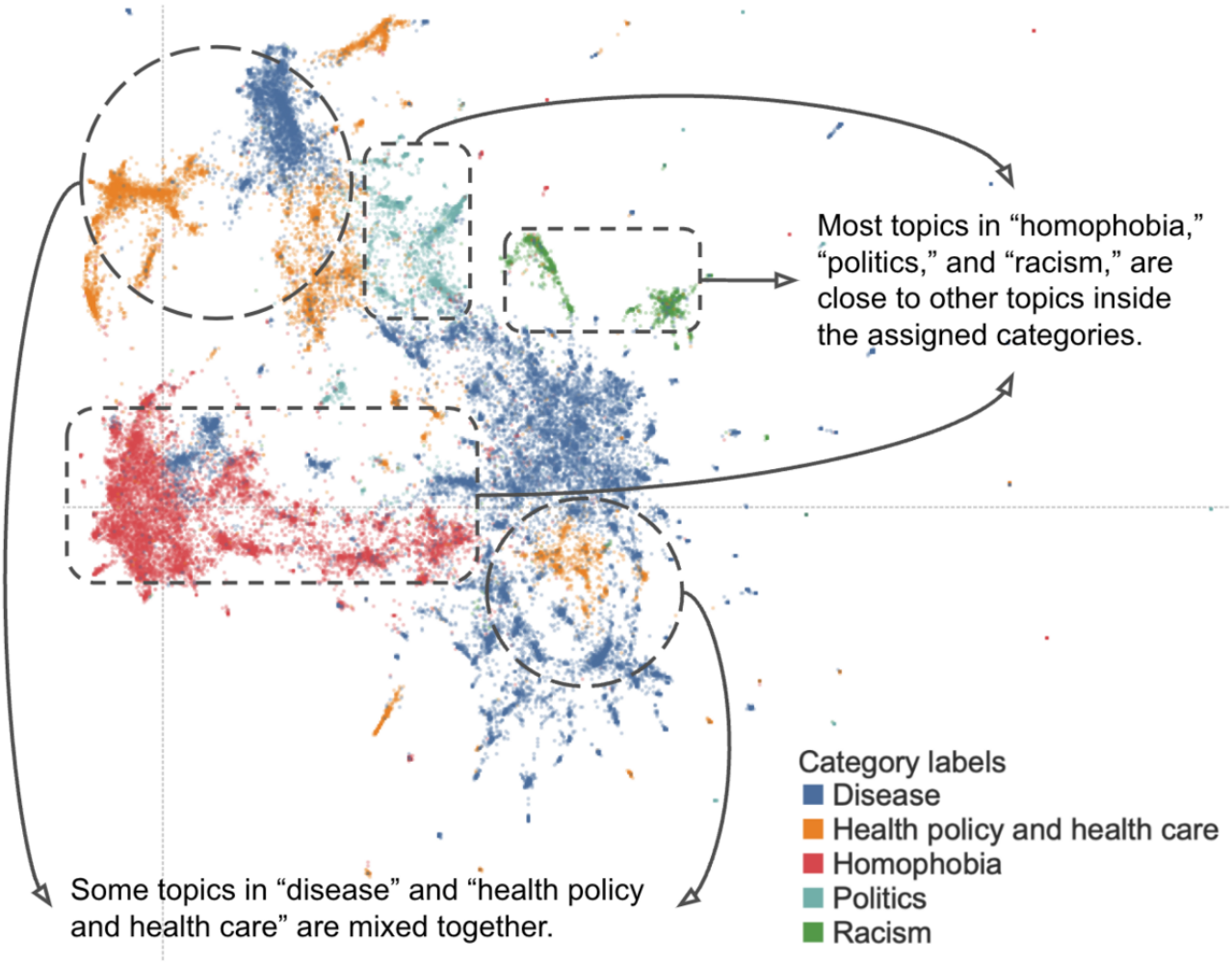

The disease category and the health policy and health care category are mixed together in multiple positions due to their topical relevance (highlighted in dashed circular frames in Figure 5). Practically, negative emotions about the disease could either be amplified or mitigated depending on the effectiveness of health policies and services, which is the reason why the topics in these 2 categories are visually close. It is often difficult to talk about health care without talking about disease and illness because they are inherently related. Homophobia, politics, and racism categories are enclosed in connected areas, which demonstrate the different sociopolitical focuses in their individual discourse (highlighted in a dashed rectangular frame in Figure 5). To better demonstrate the categorization criteria and contents in each category, we also provide example tweets in Table 4 and more examples with notes in Tables S1-S5 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 4. Example toxicity categories, topics, keywords, and tweets.

Category	Topics, n (%)	Keywords	Example tweet
Disease	6 (12)	scary_scared_shit_scaring	"If monkeypox was a person lol I swear that face kills me but no more mate, he's scary as f*ck[a] lol"
Health policy and health care	12 (24)	health_emergency_outbreak_cdc	"Hey CDC, F*ck You and your #monkeypox"
Homophobia	7 (14)	sex_anal_transmitted_spread	"Monkeypox is very serious, as serious as HIV for gay men having anal sex. The rest of us are Ok Follow Health Guidelines... avoid anal sex with gay men. Listen to the science! Nuff said? #Canada"
Politics	11 (22)	biden_ukraine_gates_f*ck	"You bet your ass they will.. School shooting, monkey-pox. Magically the story has changed away from Biden and his sh*tty gas prices, baby formula shortages, massive inflation, etc"
Racism	18 (36)	n*gg*s_n*gg*_finna_yall	"Laughing at someone catching monkeypox. You n*gg*s are lame frfr"

[a]The vowels in inappropriate words are masked; emojis and some special characters are removed; user names are removed; and some capital and lower-case letters, spaces, and punctuations are adjusted.

Information Diffusion Network

As mentioned in the Analytical Methods section, we focused on 2 relationships on Twitter: mentions and retweets. The mentions network aimed to reveal which accounts were frequently mentioned and so were encouraged to respond, while the retweet network aimed to reveal which accounts diffused toxicity. The research objective was to locate users who perpetuated or spread toxicity in the network. The mentions network (Figure 6) included 17,437 vertices, 15,085 edges, and 3,628 connected components (ie, clusters in the network). The average geodesic distance between 2 vertices was 8.041. The retweets network (Figure 7) included 59,749 vertices, 62,493 edges, and 3015 connected components. The average geodesic distance was 5.336. Overall, the mentions network had fewer vertices, edges, and connected components but a greater geodesic distance than the retweets network. This observation implies fewer interactions between users in the mentions network, possibly because those mentioned users did not respond to such toxic comments.

Figure 6. Network of Twitter users based on mentions. Each group number starts with a "G".

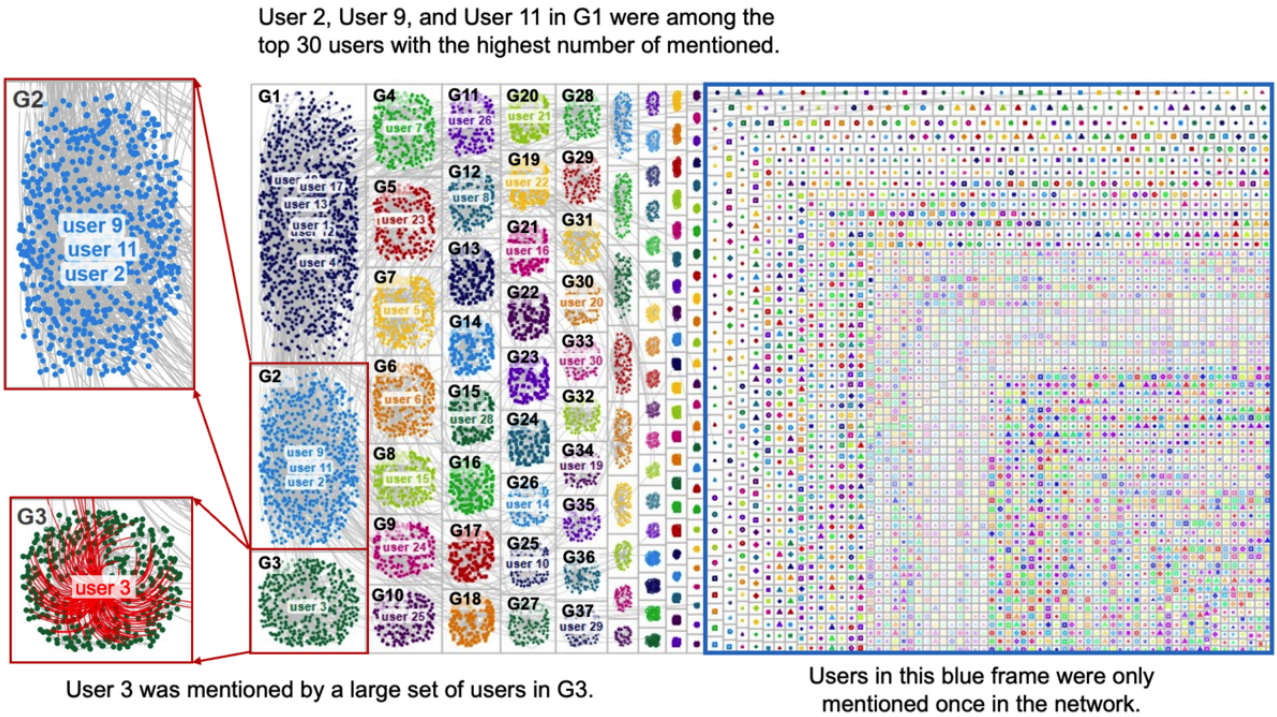

Regarding the users, we observed that a few users dominate the network, as centered in the cluster and surrounded by a large set of users in Figures 6 and 7. These "centered" users were frequently mentioned or retweeted by other users in the community and thus had the highest in-degree centrality. For some clusters, 1 user dominated the entire cluster (eg, user 3 in cluster G3 in Figure 6). For other clusters, several users colocated in the same cluster (eg, user 2, user 9, and user 11 in

XSL•FO
RenderX

cluster 2 in Figure 6), implying that their tweets shared similar outreaches and responses. We also observed that the users in the mentions network were more dispersed than those in the retweets network, as illustrated by a greater geodesic distance and a higher ratio of connected components divided by vertices. This reflects that a few users' tweets are repeatedly retweeted in the community, but users' mentions are arbitrary. However, many users were only mentioned or retweeted once according to both networks.

Figure 7. Network of Twitter users based on retweets. Each group number starts with a "G".

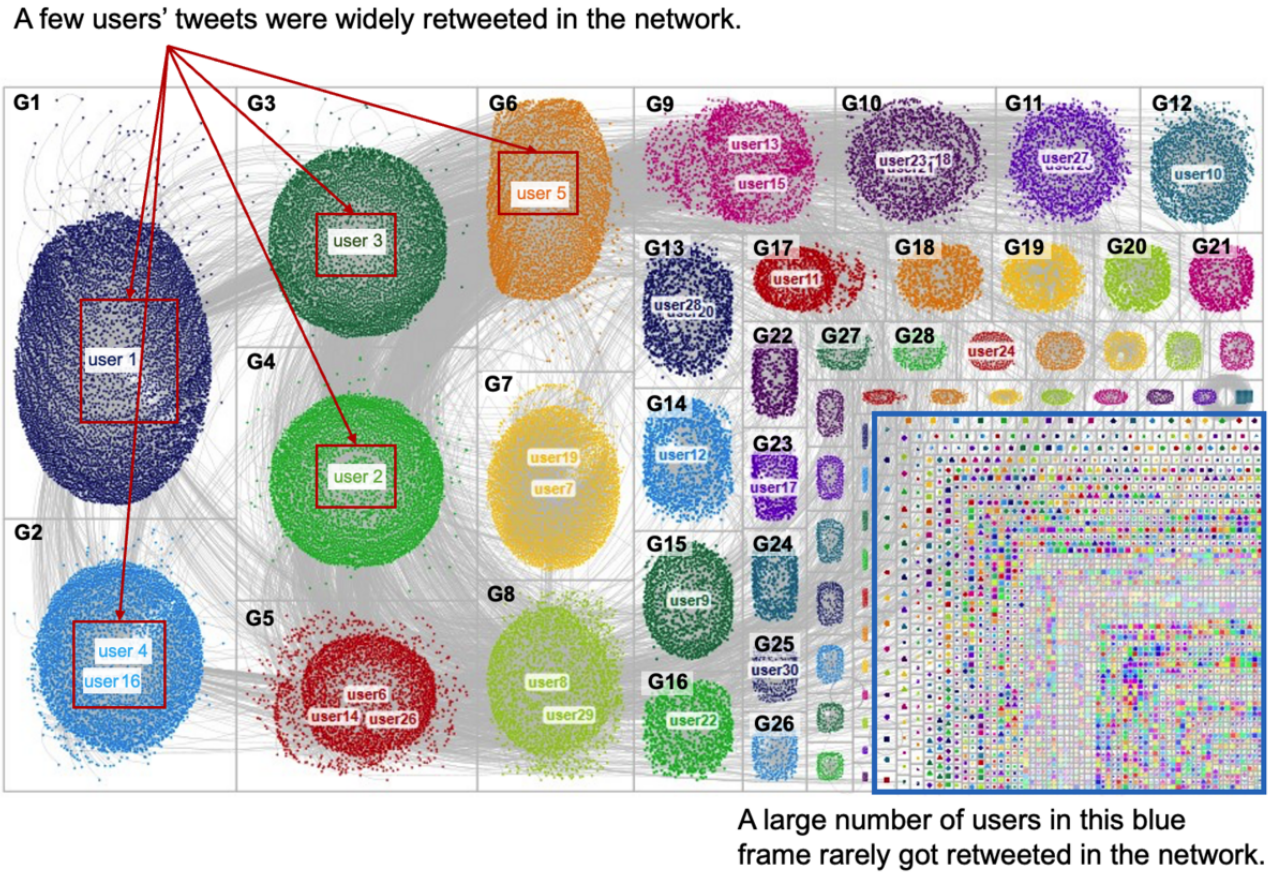

We further listed the top 30 users with the highest in-degree centrality with their account types in Tables 5 and 6. We used 2 attributes to describe a Twitter account. One attribute was the verification. A verified account may be an account of public interest, such as government agencies, politics, journalism, media, and influential public figures. For the verified accounts, we provided their usernames, while the usernames for nonverified accounts were masked. The other attribute was their account type. We manually interpreted each top user's account description and classified it into one of the following categories, as listed below:

- Organization—news media (org_media) and government agencies (org_government)
- Individual users—politician (ind_politician), journalists (ind_journalist), high-impact users (ind_impact), and other users (ind_other)

We differentiated an individual account as a "high-impact user" or "other user" account based on its number of followers; an account with >50,000 followers is identified as a "high-impact user" account.

XSL•FO
RenderX

Table 5. Top 30 users based on in-degree centrality in the mentions network.

Username	In-degree centrality, n	Cluster	Verified status	Account type	Occurrence of each category, n
User 1 (POTUS)	229	G1	True	org_government	*D*[a] *(68), H*[b] *(47), O*[c] *(67), P*[d] *(46), and R*[e] *(2)*
User 2 (WHO)	201	G2	True	org_government	*D (57), H (39), O (46), P (34), and R (35)*
User 3 (RepMTG)	155	G3	True	ind_politician	*D (49), H (9), O (67), P (31), and R (1)*
User 4 (CDCgov)	153	G1	True	org_government	*D (44), H (50), O (49), P (14), and R (1)*
User 5 (TimRunsHisMouth)	115	G7	True	ind_impact	*D (36), H (12), O (66), P (8), and R (5)*
User 6 (FoxNews)	111	G6	True	org_media	*D (37), H (14), O (33), P (18), R and (13)*
User 7	80	G4	False	org_media	*D (26), H (12), O (31), P (10), and R (1)*
User 8 (nypost)	68	G12	True	org_media	*D (10), H (10), O (42), P (9), and R (4)*
User 9 (CNN)	67	G2	True	org_media	*D (19), H (5), O (34), P (9), and R (5)*
User 10	63	G25	False	ind_impact	*D (17), H (4), O (37), and P (6)*
User 11 (DrTedros)	61	G2	True	ind_politician	*D (22), H (12), O (16), P (7), and R (6)*
User 12 (JoeBiden)	59	G1	True	ind_politician	*D (24), H (12), O (9), P (12), and R (1)*
User 13 (CDCDirector)	57	G1	True	org_government	*D (16), H (21), O (14), P (7), and R (1)*
User 14 (SkyNews)	50	G26	True	org_media	*D (21), H (2), O (13), P (5), and R (12)*
User 15 (MrAndyNgo)	46	G8	True	ind_journalist	*D (16), H (4), O (26), P (2), and R (2)*
User 16 (JackPosobiec)	44	G21	True	ind_politician	*D (13), H (4), O (24), P (3), and R (1)*
User 17 (DrEricDing)	41	G1	True	ind_impact	*D (20), H (8), O (15), and P (3)*
User 18 (nytimes)	40	G1	True	org_media	*D (11), H (6), O (18), P (7), and R (1)*
User 19 (thehill)	40	G34	True	org_media	*D (13), H (5), O (20), P (4), and R (2)*
User 20	39	G30	False	org_media	*D (12), H (5), O (16), P (5), and R (1)*
User 21 (Reuters)	38	G20	True	org_media	*D (12), H (13), O (13), P (3), and R (12)*
User 22 (Timcast)	38	G19	True	ind_journalist	*D (18), H (4), O (13), P (1), and R (2)*
User 23 (newsmax)	37	G5	True	org_media	*D (18), H (4), O (11), P (2), and R (1)*
User 24 (washingtonpost)	36	G9	True	org_media	*D (7), H (3), O (24), P (1), and R (1)*
User 25 (Scott_Wiener)	35	G10	True	ind_politician	*D (4), H (5), O (22), P (3), and R (1)*
User 26 (ZubyMusic)	35	G11	True	ind_impact	*D (12), H (4), O (19), and P (3)*

XSL•FO
RenderX

Username	In-degree centrality, n	Cluster	Verified status	Account type	Occurrence of each category, n
User 27 (BetoORourke)	35	G5	True	ind_politician	*D (14), H (1), O (11), and P (11)*
User 28 (bethanyshondark)	34	G15	True	ind_impact	*D (15), H (5), and O (15)*
User 29 (unusual_whales)	33	G37	True	org_media	*D (10), H (7), and O (16)*
User 30 (MattWalshBlog)	33	G33	True	ind_impact	*D (5), H (5), O (25), and R (1)*

[a]D: disease.

[b]H: health policy and health care.

[c]O: homophobia.

[d]P: politics.

[e]R: racism.

XSL•FO
RenderX

Table 6. Top 30 users based on in-degree centrality in the retweets network.

Username	In-degree centrality, n	Cluster	Verified status	Account type	Occurrence of each category, n
User 1	7107	G1	False	ind_other	D[a] (1)
User 2 (TimRunsHisMouth)	5038	G4	True	ind_impact	D (1), H[b] (1), and O[c] (1)
User 3	4325	G3	False	ind_impact	D (1)
User 4	4072	G2	False	ind_impact	D (1)
User 5	3222	G6	False	ind_impact	D (1)
User 6	2258	G5	False	ind_impact	H (1)
User 7	1947	G7	False	ind_impact	D (1)
User 8	1000	G8	False	ind_impact	H (1) and R[d] (1)
User 9	897	G15	False	ind_impact	D (1) and O (1)
User 10 (bahjarodriguez)	804	G12	True	ind_impact	D (1)
User 11	599	G17	False	ind_impact	D (1)
User 12 (jennawadsworth)	571	G14	True	ind_politician	D (1)
User 13	535	G9	False	ind_impact	D (1)
User 14	391	G5	False	ind_impact	D (1)
User 15 (DrEricDing)	345	G9	True	ind_impact	H (1)
User 16 (AngryBlackLady)	336	G2	True	ind_impact	D (1)
User 17	332	G23	False	ind_other	H (1) and O (1)
User 18	291	G10	False	ind_impact	H (1)
User 19	285	G7	False	ind_impact	D (1)
User 20	284	G13	False	ind_other	H (1)
User 21	282	G10	False	ind_impact	H (2)
User 22	272	G16	False	ind_politician	D (1), H (1), and P[e] (1)
User 23	260	G10	False	ind_other	H (1)
User 24	259	G29	False	ind_impact	D (1)
User 25 (johncardillo)	258	G11	True	ind_impact	D (1)
User 26	256	G5	False	ind_impact	D (2)
User 27	254	G11	False	ind_impact	*P* (1)
User 28	237	G13	False	ind_impact	R (1)
User 29	232	G8	False	ind_impact	D (1)
User 30	227	G25	False	ind_other	O (1)

[a]D: disease.

[b]H: health policy and health care.

[c]O: homophobia.

[d]R: racism.

[e]P: politics.

The sum of mentions in each category may be larger than the in-degree centrality in Table 5 because the in-degree centrality is computed based on the unique edges between 2 users.

We noted a couple of observations regarding the top users in the networks. For the mentions network, 90% (27/30) of the top-mentioned users had verified accounts, primarily news agencies, government portals, politicians, and independent high-impact users. In particular, the most frequently mentioned accounts were from government portals or politicians. By contrast, for the retweets network, most top users had nonverified accounts (24/30, 80%), and they were independent influencers (23/30, 77%) with large followings (ie, >50,000 followers). There is also a clear distinction regarding the organization accounts between mentions and retweets networks. More than half of the most frequently mentioned accounts (16/30, 55%) were from organization accounts, but none were among the top users in the retweets network.

XSL•FO
RenderX

Regarding the categories, these top users were frequently mentioned in tweets relative to disease, health policy and health care, and homophobia but comparatively less mentioned in the categories of politics and racism. On the basis of the retweets network, those tweets discussing disease and health policy and health care were most likely to obtain attention from the online community, while the other 3 categories did not. We observed that the categories of homophobia, politics, and racism were more likely to be shorter lived and more locally lived than the categories of disease and health policy and health care toxicity because these topics rarely appeared among the top retweeted users. Toxic tweets in the categories of homophobia and racism did not receive much attention, given that categories of "O" and "R" were rarely mentioned in Table 6.

Discussion

Principal Findings

Toxicity on social media is widespread during health crises, with many individuals spreading misinformation, fear, and hatred [8,9]. This can undermine public health communication efforts and lead to confusion and anxiety among the public. The discussion of toxic narratives during the 2022 mpox outbreak is an example of controversy in public communication during health crises. Building on prior work that leverages either topic modeling or network analysis techniques [66-68], our study further demonstrates the value of combining topical and network analyses to understand emerging social issues and crises. By examining the topical dynamics, we were able to uncover the prevalent themes in the toxic discourse during the 2022 mpox outbreak and observe their temporal shifts. Network dynamics revealed the key users and their roles in propagating toxicity, suggesting that addressing these high-impact users and their narratives could be crucial for effective crisis communication and policy decision-making. Our findings highlight the importance of monitoring and addressing toxicity on social media to foster a more inclusive and constructive public dialogue during health emergencies.

We also adapted the Rabat Plan of Action analytical framework for hate speech analysis to study toxicity on social media during the mpox outbreak. This framework takes into account the context, extent, content, speaker, and intent. By examining the context of the discussions, including the events that led to the discourse, the extent of toxic comments and hate speech, the content and themes, the speakers involved, and the intent behind the messages, the adapted analytical framework provided a comprehensive understanding of the toxicity landscape in the mpox scenario.

Toxicity Aboutness Reveals an Infodemic: Negative Feelings, Political Unrest, and Weaponized Stigma

Topical dynamics summarize temporal content popularity and provide an extended context of social issues in the mpox health crisis. In this sense, the understanding of context and content of toxicity represents the extremes of public opinion, respectively, from event and topical trend perspectives, which are mutually beneficial in profiling the problems in health communications during the 2022 mpox public health emergency. By examining groups of related topics, we can grasp an overarching view of the main subjects being discussed. In other words, we can identify the primary categories that highlight the most commonly mentioned topics. By analyzing temporal topical swifts, we further understand when the topical discourses, especially their peaks, occur in each of the categories. Such a summary of contents can reveal what categories of topics are discussed together, which quantitatively demonstrates public opinions around context and facilitates a multifaceted understanding of toxic contents in this health crisis. In particular, there are 3 outstanding problems associated with toxicity on social media: negative feelings, political unrest, and weaponized stigma.

Negative feelings during the 2022 mpox public health emergency often came out of emotions toward the disease or how the disease was dealt with. There were negative feelings, such as fear and anger, due to the actual or imagined physical symptoms caused by mpox or negativity because of mental anxiety. Specific to the 2022 mpox outbreak, there were also negative feelings carried on from the COVID-19 pandemic because some initial public health guidance were similar (eg, vaccination, wearing masks, and self-quarantine). Thus, similar to the COVID-19 pandemic, people could also be unhappy about how the health emergency was dealt with by the health authorities and health care providers. These negative feelings are not produced in isolation: Twitter users read and watch news from different media platforms and share their ideas on the platform. When some personal beliefs, which might not be scientifically mature, are put together with the practical inconvenience, the public uses toxicity to express their negativity toward the disease and the health services they receive.

Political unrest during the 2022 mpox outbreak was fueled by divisive reactions from politicians and the public, leading to the spread of conspiracy theories and misinformation. Disagreements over health policies, allocation of resources, and the overall handling of the outbreak often manifested in toxic discourse, further polarizing society. For example, some Twitter users propagated unfounded claims that the mpox outbreak was a result of a laboratory leak or a government conspiracy, which led to increased distrust in the authorities and health care providers. This toxic environment can be understood through the lens of psychological factors, such as fear, uncertainty, and a tendency toward confirmation bias, where individuals are more likely to believe and spread information that aligns with their pre-existing beliefs and fears. Societal factors, such as political polarization and a general erosion of trust in public institutions, also played a critical role in amplifying toxic behavior on the internet, as people sought out and shared content that validated their anxieties and skepticism.

Weaponized stigma became another significant issue during the outbreak, as incidents of attacks toward minority groups based on (perceived) sexuality, gender identity, and race were reported. This stigmatization was often rooted in misinformation and fear, with people associating certain groups with the spread of the disease or accusing them of not following public health guidelines. Toxicity on social media facilitated the perpetuation of these stigmatizing narratives, further marginalizing these communities and exacerbating existing social divisions.

XSL•FO
RenderX

Psychological factors such as xenophobia, scapegoating, and the need to find a tangible source of blame during a crisis contributed to the spread of these harmful narratives. Societal factors, including systemic discrimination and historical prejudices, were also at play, as these pre-existing biases were amplified in the digital space, leading to more virulent expressions of hate and intolerance. Understanding these dynamics is crucial for developing strategies to mitigate toxicity on social media and support affected communities.

Toxicity Diffusion Suggests Improvements in Health Communication: Influential Users Should Respond to and Counter Toxicity

Network dynamics reveal frequent speakers and their intentions, highlighting key priorities for public health communication and health policy. On the basis of the analytical framework, we used the social network theory to calculate the in-degree and betweenness centrality. Through this approach, we were able to identify influential speakers and mentioned users as well as gain insights into what they said and the likelihood of responses in online communities. Our analysis reveals several key observations regarding speakers and their intentions that are worth discussing.

In our analysis of speakers, we discovered that a few nonverified yet influential users dominated the retweets network. Their tweets garnered broad attention from the online community and resonated with many others who shared the same opinion regarding disease- and health-related negativity. By contrast, users who were mentioned in these toxic messages appeared to be scattered and chosen randomly, and they seldom responded to the negativity. Our findings suggest that toxic information, regardless of its intent, typically does not elicit responses from those who are mentioned.

For those most frequently mentioned users, our analysis revealed that verified government channels, news agencies, and politicians dominated the top-mentioned list. This finding highlights the importance of government and health agents being aware of toxic information and taking appropriate action. For example, some users expressed concerns on Twitter about the transmission routes of a disease, albeit in a toxic manner. To help limit such toxicity, we suggest that public health organizations such as the WHO and Centers for Disease Control and Prevention should inform the public about the transmission routes of the disease and the severity of the health crisis. This underscores the need for timely and effective communication from official sources in response to public concerns.

Upon analyzing intents in the retweets network, we observed that attributions of diseases to homophobia and racism were not frequently mentioned by the top speakers. While 1 interpretation could be that most online users view such attributions as malicious during public health crises, another perspective is that these top speakers, who are already prominent and attract attention regardless of their content, might not engage in such rhetoric. If these influential speakers had used this type of rhetoric, it is possible that it would have still received significant engagement. In contrast, tweets related to disease- and health care policy–related negativity were more generalized among a broader set of users and were retweeted for a more extended period. Moreover, our findings highlight the importance of focusing on accurate and relevant information during public health crises, as misinformation or toxic narratives can be quickly dismissed by online users. We suggest that public health agencies prioritize accurate information dissemination, which can help combat the spread of harmful narratives and promote healthy dialogue and public understanding of health-related issues.

After analyzing intents in the mentions network, we found that verified users were primarily mentioned in topics related to disease-related negativity, health policy–related negativity, and homophobia. Our findings reveal widespread dissatisfaction regarding health policies for mpox and concerns about the severity of the disease outbreak. This highlights the urgent need for government or health channels to release transparent and reliable information to address public concerns. Our findings also imply that some users might misunderstand the disease transmission or use mpox to stigmatize homosexuality. This indicates a need for relevant agencies to take immediate action to interrupt the dissemination of toxic information. By doing so, we can mitigate the influence of toxicity and reduce the harm to groups considered historically marginalized, such as the lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ+) community.

Comparison to Prior Work

Previous research has demonstrated that NLP techniques, such as topic modeling [66] and text classification [69], could be useful for uncovering toxicity on social media and hate speech during health crises. In the broader context of health communication, our study highlights several significant implications. One key implication involves the use of NLP and computational techniques to analyze social media narratives. Our research further shows that combining topic modeling and network analysis can provide a nuanced understanding of online toxic narratives and their dissemination.

Next, our topic modeling results suggest common underlying causes of toxicity and hate speech on social media, including emotional, political, and stigmatizing responses. Toxic narratives by negative feelings have been widely reported in previous studies [70,71], often linked to policies such as lockdowns and mask mandates during the COVID-19 pandemic [72]. In addition, political unrest and misinformation exacerbate these emotions, with conspiracy theories and distrust in authorities intensifying toxic narratives. This pattern, observed during the mpox outbreak, was also prevalent in past pandemics, such as the COVID-19 pandemic [67]. Moreover, the stigmatization of certain groups, such as the gay community during the mpox outbreak, mirrors the scapegoating and xenophobia seen in previous crises. For instance, previous studies reported widespread anti-Asian sentiment on social media during the COVID-19 pandemic [73,74]. Understanding these causes in terms of misinformation, distrust of government or health agencies, and societal biases on certain groups is crucial for developing strategies to mitigate the harmful impact of toxicity on social media during health crises.

XSL•FO
RenderX

Our network analysis highlights the crucial need to engage with influential users and address key narratives to mitigate the spread of toxicity on social media during health crises. Consistent with previous research [32], our findings suggest that official sources, such as government agencies and health care organizations, must prioritize timely, transparent, and accessible communication across major social media platforms, as these entities are frequent targets of toxic discourse. While earlier studies have found that right-wing sources are often linked to higher levels of toxicity and scientific sources to lower levels [67], our research indicates that general users, rather than verified accounts, are more likely to lead and propagate toxic narratives. This underscores the importance of not only monitoring influential users but also addressing the broader network of general users who contribute to the dissemination of harmful content.

Limitations and Outlook

One limitation could result from the use of Perspective API. For example, Perspective API might incorrectly identify the toxicity if a tweet's toxic words or patterns do not appear similar to its training samples. In particular, social media language is informal, and Perspective API might not be able to identify internet slang or abbreviations correctly. In addition, the threshold selection may have impacted the toxicity results. A smaller threshold increases the likelihood of identifying a toxic tweet. However, it simultaneously increases the number of false positives (ie, a nontoxic tweet is identified as toxic) [10].

Another limitation relates to the types of toxicity used. Toxicity on social media may include both emphasis of emotions (eg, fear and anxiety) toward the crisis and attacks of associated minority groups (eg, Asian Americans during the COVID-19 pandemic and LGBTQ+ people during the 2020 mpox outbreak). There is no clear boundary between the users who spread these 2 types of toxicity or between the language they use. For example, while a significant portion of the toxicity targets LGBTQ+ or African and African American communities, it is difficult to determine if some tweets are using toxic language to highlight inequity or discrimination and condemn identity-based attacks. Future work could include an analysis of nontoxic discourse to serve as a comparison to facilitate the overall understanding of toxic tweets about mpox. By examining the overlap between the discourse in toxic versus nontoxic discourse, especially the influential users behind it, we can better understand the extent to which these users are involved in combating misinformation and toxic speech. This comparative analysis would provide a more comprehensive depiction of user discourse on social media, with a focus on showcasing the difference between emphasis and toxicity.

In addition, algorithmic bias is an important consideration in our study, particularly concerning the algorithms used for data analysis, such as BERT and UMAP. BERT, a transformer-based model, is pretrained on large text corpora, which may contain inherent biases reflecting societal stereotypes and prejudices. These biases can influence the model's understanding and representation of language, potentially skewing the identification and clustering of topics. Similarly, UMAP, a dimensionality reduction technique, might introduce biases through its assumptions about data structure and the preservation of local and global relationships within the dataset. These algorithmic biases can impact the study's findings by potentially misrepresenting the semantic relationships and topic distributions within the toxic discourse, leading to conclusions that may not fully or accurately reflect the underlying data. Acknowledging these biases is crucial, and future work should focus on using bias mitigation strategies, such as algorithmic auditing and using debiased training datasets, to enhance the fairness and accuracy of the analysis.

At the same time, toxicity can look different in different parts of the world, among different cultural groups, and in different languages. While our analysis focused on English-speaking countries, it is important to recognize that many affected populations are not represented in the tweets we collected. Our approach may inherently overlook the perspectives and nuances present in non-English tweets, which may result in a cultural bias. We chose English content primarily due to the availability of robust language processing tools and resources for English, which facilitates more reliable analysis. We recognize that this focus might limit the generalizability of our findings to non–English-speaking audiences, and we have listed this as a limitation in the Discussion. The discourse around public health crises, such as the mpox outbreak, is multifaceted and culturally dependent. Non-English content may reveal different concerns, misinformation patterns, and public reactions that our study does not capture. As such, we could investigate toxicity with more granularity in the future and characterize it with regard to attack versus emphasis as well as demographic factors, including language, country, and culture.

Conclusions

Toxic online discourse can have detrimental impacts on public health crises. In this study, we collected tweet data during the 2022 mpox outbreak and analyzed toxicity from multiple dimensions, including context, extent, content, speaker, and intent. To better understand toxic dynamics on Twitter, we used BERTopic and social network community clustering techniques.

The temporal discourse analysis revealed that toxic tweets during the outbreak covered a diverse range of topical categories. The predominant topics were toxicity on disease, health policy and health care, and homophobia. While toxicity related to politics and racism had lower daily volumes, they reached respective peaks when triggering events happened. On the other hand, verified government channels, news agencies, and politicians were among the top-mentioned users in the social network and were primarily associated with the categories of disease, health policy and health care, and homophobia. Meanwhile, a few nonverified but influential users posted tweets that received high volumes of retweets, and tweets related to homophobia, politics, and racism were more likely to be shorter lived and have a local impact.

As such, to mitigate the toxicity on social media of mpox or similar infodemics, public health authorities should leverage advanced NLP tools, such as sentiment analysis and toxicity detection algorithms, to identify and address harmful content in real time. In addition, digital literacy campaigns can educate the public about the dangers of misinformation and the

XSL•FO
RenderX

importance of respectful online discourse. Establishing rapid response teams comprising public health experts, communication specialists, and community leaders can help counteract false narratives and provide accurate information swiftly. Finally, fostering partnerships with influential social media figures and organizations can amplify positive messages and mitigate the spread of toxic content.

To summarize, the topical dynamics revealed that Twitter users were expressing negativity and making controversial remarks about the mpox public health emergency, indicating a worsening of political unrest and the increased weaponization of stigma during the corresponding infodemic. The network dynamics highlight the need for government and health agencies to release transparent and reliable information to address public concerns. Overall, our study demonstrates a workflow that combines topical and network analyses to understand emerging social issues and crises. Our findings emphasize the importance of proactive measures needed from government and health agencies to combat harmful narratives and promote accurate information during public health crises.

Acknowledgments

This study is supported by the Carnegie Endowment for Faculty Development at the University of Michigan School of Information. Third-party grammar check applications were used (Grammarly; Grammarly, Inc).

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to Twitter's developer agreement and policy but are available from the corresponding author on reasonable request. Raw data, consisting of tweets and user information, cannot be shared directly due to privacy and data protection regulations. Note that the Twitter Academic application programming interface service was discontinued as of mid-2023, while access to Twitter data are still available at a cost [37].

Authors' Contributions

LF, LL, and LH conceptualized the study, and LF curated the data. LF and LL were involved in the methodology, formal analysis, and writing the original draft. LH was involved in project administration, resource collection, and supervision. LF, LL, and LH were involved in reviewing and editing the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Method notes on network analysis.
[DOCX File , 147 KB - jmir_v26i1e52997_app1.docx]

Multimedia Appendix 2
Notes on topic modeling.
[DOCX File , 17 KB - jmir_v26i1e52997_app2.docx]

References

1. Posner L, Turilli I. Monkeypox timeline: a frequently updated tracker of emerging developments from the beginning of the 2022 monkeypox outbreak. Think Global Health. URL: https://www.thinkglobalhealth.org/article/monkeypox-timeline [accessed 2023-09-07]
2. Mpox (monkeypox). World Health Organization (WHO). URL: https://www.who.int/news-room/fact-sheets/detail/monkeypox [accessed 2023-09-06]
3. WHO recommends new name for monkeypox disease. World Health Organization (WHO). URL: https://www.who.int/news/item/28-11-2022-who-recommends-new-name-for-monkeypox-disease [accessed 2023-09-06]
4. Factsheet for health professionals on mpox (monkeypox). European Centre for Disease Prevention and Control. URL: https://www.ecdc.europa.eu/en/all-topics-z/monkeypox/factsheet-health-professionals [accessed 2023-09-06]
5. FDA Mpox response. U.S. Department of Food and Agriculture. URL: https://www.fda.gov/emergency-preparedness-and-response/mcm-issues/fda-mpox-response [accessed 2023-09-06]
6. HHS response to the Mpox outbreak. Assistant Secretary for Public Affairs (ASPA), U.S. Department of Health and Human Services. URL: https://www.hhs.gov/mpox/index.html [accessed 2023-09-06]
7. Edinger A, Valdez D, Walsh-Buhi E, Trueblood JS, Lorenzo-Luaces L, Rutter LA, et al. Misinformation and public health messaging in the early stages of the Mpox outbreak: mapping the Twitter narrative with deep learning. J Med Internet Res 2023 Jun 06;25:e43841 [FREE Full text] [doi: 10.2196/43841] [Medline: 37163694]

XSL•FO
RenderX

8. Kolhatkar V, Wu H, Cavasso L, Francis E, Shukla K, Taboada M. The SFU opinion and comments corpus: a corpus for the analysis of online news comments. Corpus Pragmat 2020 Nov 02;4(2):155-190 [FREE Full text] [doi: 10.1007/s41701-019-00065-w] [Medline: 32685909]
9. Using machine learning to reduce toxicity online. Perspective. URL: https://perspectiveapi.com [accessed 2023-09-06]
10. Rieder B, Skop Y. The fabrics of machine moderation: studying the technical, normative, and organizational structure of Perspective API. Big Data Soc 2021 Oct 16;8(2):21. [doi: 10.1177/20539517211046181]
11. Turner A. How many users does X (formerly Twitter) have? Bankmycell. 2023. URL: https://www.bankmycell.com/blog/how-many-users-does-twitter-have [accessed 2023-09-21]
12. The Rabat plan of action. United Nations Human Rights. URL: https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action [accessed 2023-09-06]
13. Sheth A, Shalin VL, Kursuncu U. Defining and detecting toxicity on social media: context and knowledge are key. Neurocomputing 2022 Jun;490:312-318. [doi: 10.1016/j.neucom.2021.11.095]
14. TDL brief: online toxicity. The Decision Lab. URL: https://thedecisionlab.com/insights/health/tdl-brief-online-toxicity [accessed 2023-02-13]
15. Suler J. The online disinhibition effect. Cyberpsychol Behav 2004 Jun;7(3):321-326. [doi: 10.1089/1094931041291295] [Medline: 15257832]
16. Shen C, Sun Q, Kim T, Wolff G, Ratan R, Williams D. Viral vitriol: predictors and contagion of online toxicity in world of tanks. Comput Human Behav 2020 Jul;108:106343. [doi: 10.1016/j.chb.2020.106343]
17. Kim JW, Guess A, Nyhan B, Reifler J. The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. J Commun 2021;71(6):922-946 [FREE Full text] [doi: 10.1093/joc/jqab034]
18. Parent MC, Gobble TD, Rochlen A. Social media behavior, toxic masculinity, and depression. Psychol Men Masc 2019 Jul;20(3):277-287 [FREE Full text] [doi: 10.1037/men0000156] [Medline: 38250140]
19. Wijesiriwardene T, Inan H, Kursuncu U, Gaur M, Shalin VL, Thirunarayan K, et al. ALONE: a dataset for toxic behavior among adolescents on Twitter. In: Proceedings of the 12th International Conference on Social Informatics. 2020 Presented at: SocInfo '20; October 6-9, 2020; Pisa, Italy p. 427-439 URL: https://link.springer.com/chapter/10.1007/978-3-030-60975-7_31 [doi: 10.1007/978-3-030-60975-7_31]
20. Suarez Estrada M, Juarez Y, Piña-García CA. Toxic social media: affective polarization after feminist protests. Soc Media Soc 2022 May 17;8(2):1-12. [doi: 10.1177/20563051221098343]
21. Guberman J, Schmitz C, Hemphill L. Quantifying toxicity and verbal violence on Twitter. In: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion. 2016 Presented at: CSCW '16; February 26-March 2, 2016; San Francisco, CA p. 277-280 URL: https://dl.acm.org/doi/10.1145/2818052.2869107 [doi: 10.1145/2818052.2869107]
22. Almerekhi H, Kwak H, Jansen BJ, Salminen J. Detecting toxicity triggers in online discussions. In: Proceedings of the 30th ACM Conference on Hypertext and Social Media. 2019 Presented at: HT '19; September 17-20, 2019; Hof, Germany p. 291-292 URL: https://dl.acm.org/doi/10.1145/3342220.3344933 [doi: 10.1145/3342220.3344933]
23. Lwin MO, Lu J, Sheldenkar A, Schulz PJ, Shin W, Gupta R, et al. Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. JMIR Public Health Surveill 2020 May 22;6(2):e19447 [FREE Full text] [doi: 10.2196/19447] [Medline: 32412418]
24. Moderation. OpenAI. URL: https://platform.openai.com/docs/guides/moderation/overview [accessed 2023-02-15]
25. Record I, Miller B. People, posts, and platforms: reducing the spread of online toxicity by contextualizing content and setting norms. Asian J Philos 2022 Aug 19;1(2):41 [FREE Full text] [doi: 10.1007/s44204-022-00042-2] [Medline: 38013952]
26. Salminen J, Sengün S, Corporan J, Jung SG, Jansen BJ. Topic-driven toxicity: exploring the relationship between online toxicity and news topics. PLoS One 2020 Feb 21;15(2):e0228723 [FREE Full text] [doi: 10.1371/journal.pone.0228723] [Medline: 32084164]
27. Denecke K, Atique S. Social media and health crisis communication during epidemics. In: Syed-Abdul S, Gabarron E, Lau AY, editors. Participatory Health Through Social Media. New York, NY: Academic Press; 2016:427-466.
28. Sellnow TL, Sellnow DD, Helsel EM, Martin JM, Parker JS. Risk and crisis communication narratives in response to rapidly emerging diseases. J Risk Res 2018 Jan 19;22(7):897-908. [doi: 10.1080/13669877.2017.1422787]
29. Gesser-Edelsburg A, Shir-Raz Y, Walter N, Mordini E, Dimitriou D, James JJ, et al. The public sphere in emerging infectious disease communication: recipient or active and vocal partner? Disaster Med Public Health Prep 2015 Aug 17;9(4):447-458. [doi: 10.1017/dmp.2015.31] [Medline: 26186579]
30. van Velsen L, van Gemert-Pijnen JE, Beaujean DJ, Wentzel J, van Steenbergen JE. Should health organizations use web 2.0 media in times of an infectious disease crisis? An in-depth qualitative study of citizens' information behavior during an EHEC outbreak. J Med Internet Res 2012 Dec 20;14(6):e181 [FREE Full text] [doi: 10.2196/jmir.2123] [Medline: 23257066]
31. Majó-Vázquez S, Nielsen RK, Verdú J, Rao N, De Domenico M, Papaspiliopoulos O. Volume and patterns of toxicity in social media conversations during the COVID-19 pandemic. Reuters Institute for the Study of Journalism. URL: https://reutersinstitute.politics.ox.ac.uk/volume-and-patterns-toxicity-social-media-conversations-during-covid-19-pandemic [accessed 2024-04-29]

XSL•FO
RenderX

32. Pascual-Ferrá P, Alperstein N, Barnett DJ, Rimal RN. Toxicity and verbal aggression on social media: polarized discourse on wearing face masks during the COVID-19 pandemic. Big Data Soc 2021 Jun 10;8(1):1-17. [doi: 10.1177/20539517211023533]
33. Kim DK, Kreps GL. An analysis of government communication in the United States during the COVID-19 pandemic: recommendations for effective government health risk communication. World Med Health Policy 2020 Dec 27;12(4):398-412 [FREE Full text] [doi: 10.1002/wmh3.363] [Medline: 32904935]
34. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676 [FREE Full text] [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
35. Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. J Med Internet Res 2020 Jun 29;22(6):e21820 [FREE Full text] [doi: 10.2196/21820] [Medline: 32589589]
36. Covello VT. Best practices in public health risk and crisis communication. J Health Commun 2003;8 Suppl 1:5-8. [doi: 10.1080/713851971] [Medline: 14692565]
37. Twitter academic API. Twitter. URL: https://developer.twitter.com/en/products/twitter-api/academic-research [accessed 2023-02-05]
38. Bryant M. Patient treated for monkeypox in isolation at London hospital. The Guardian. URL: https://www.theguardian.com/uk-news/2022/may/07/patient-treated-for-monkeypox-in-isolation-at-london-hospital [accessed 2023-09-06]
39. Mathieu E, Spooner F, Dattani S, Ritchie H, Roser M. Mpox (monkeypox). Our World in Data. 2022. URL: https://ourworldindata.org/monkeypox [accessed 2024-04-29]
40. Hermann P. Police: assailants used slur and referenced monkeypox in Shaw attack. The Washington Post. 2022 Aug 09. URL: https://www.washingtonpost.com/dc-md-va/2022/08/09/shaw-attack-monkeypox-slur-hate/ [accessed 2023-09-07]
41. Kates J, Artiga S, Dawson L. National data show continuing disparities in Monkeypox (MPX) cases and vaccinations among black and Hispanic people. Kaiser Family Foundation (KFF). URL: https://www.kff.org/racial-equity-and-health-policy/issue-brief/national-data-show-continuing-disparities-in-mpx-monkeypox-cases-and-vaccinations-among-black-and-hispanic-people/ [accessed 2023-09-07]
42. As Monkeypox eases in the West, "nothing has changed" for Africa, where doctors remain helpless. CBS News. URL: https://www.cbsnews.com/news/monkeypox-vaccine-shortage-africa-as-virus-eases-in-us-europe/ [accessed 2023-09-07]
43. Toolkit for analysing a case of hate speech. European Court of Human Rights, The European Union. URL: https://rm.coe.int/advanced-guide-toolkit-how-to-analyse-hate-speech/1680a217cd [accessed 2024-04-29]
44. Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat 2007 Jun 1;1(1):17-35. [doi: 10.1214/07-aoas114]
45. Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer TK, McNamara DS, Dennis S, Kintsch W, editors. Handbook of Latent Semantic Analysis. New York, NY: Psychology Press; 2007.
46. Manna S, Phongpanangam O. Exploring topic models on short texts: a case study with crisis data. In: Proceedings of the 2018 Second IEEE International Conference on Robotic Computing. 2018 Presented at: IRC '18; January 31-February 2, 2018; Laguna Hills, CA p. 377-382 URL: https://ieeexplore.ieee.org/document/8329942 [doi: 10.1109/irc.2018.00078]
47. Yu H, Fan L, Gilliland AJ. Disparities and resilience: analyzing online health information provision, behaviors and needs of LBGTQ + elders during COVID-19. BMC Public Health 2022 Dec 13;22(1):2338 [FREE Full text] [doi: 10.1186/s12889-022-14783-5] [Medline: 36514032]
48. Li L, Ma Z, Fan L, Lee S, Yu H, Hemphill L. ChatGPT in education: a discourse analysis of worries and concerns on social media. Educ Inf Technol 2023 Oct 11;29(9):10729-10762 [FREE Full text] [doi: 10.1007/S10639-023-12256-9]
49. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online August 2, 2023 2023 [FREE Full text]
50. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online May 24, 2019 2019 [FREE Full text]
51. Sia S, Dalmia A, Mielke SJ. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too!. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual Event p. 1728-1736 URL: https://aclanthology.org/2020.emnlp-main.135.pdf [doi: 10.18653/v1/2020.emnlp-main.135]
52. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online March 11, 2022 2022 [FREE Full text]
53. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv. Preprint posted online August 27, 2019 2019 [FREE Full text] [doi: 10.18653/v1/d19-1410]
54. Cuenca NR. all-MiniLM-L6-v2. Hugging Face. URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 [accessed 2023-09-21]
55. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. J Open Source Softw 2018 Sep;3(29):861. [doi: 10.21105/joss.00861]
56. Indyk P, Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the 13th annual ACM symposium on Theory of computing. 1998 Presented at: STOC '98; May 24-26, 1998; Dallas, TX p. 604-613 URL: https://dl.acm.org/doi/10.1145/276698.276876 [doi: 10.1145/276698.276876]

XSL•FO
RenderX

57. sklearn.cluster.KMeans. Scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html [accessed 2023-09-23]
58. Danielsson PE. Euclidean distance mapping. Comput Graph Image Process 1980 Nov;14(3):227-248. [doi: 10.1016/0146-664x(80)90054-4]
59. Fan L. lizhouf/mpox_50_topics_modeling. Hugging Face. URL: https://huggingface.co/lizhouf/mpox_50_topics_modeling [accessed 2023-09-21]
60. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 2010 Aug 28;1(1-4):43-52. [doi: 10.1007/S13042-010-0001-0]
61. Chu AM, Chong AC, Lai NH, Tiwari A, So MK. Enhancing the predictive power of Google trends data through network analysis: infodemiology study of COVID-19. JMIR Public Health Surveill 2023 Sep 07;9:e42446 [FREE Full text] [doi: 10.2196/42446] [Medline: 37676701]
62. About different types of posts. X (Formerly Twitter). URL: https://help.twitter.com/en/using-twitter/types-of-tweets [accessed 2023-01-03]
63. Knoke D, Yang S. Social Network Analysis. Thousand Oaks, CA: Sage Publications; 2008.
64. Smith M, Milic-Frayling N, Shneiderman B, Mendes RE, Leskovec J, Dunne C. NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010. ScienceOpen. 2019. URL: https://www.scienceopen.com/document?vid=0b85110c-3853-4589-95cb-def966c20c49 [accessed 2024-04-29]
65. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys 2004 Dec;70(6 Pt 2):066111. [doi: 10.1103/PhysRevE.70.066111] [Medline: 15697438]
66. Obadimu A, Khaund T, Mead E, Marcoux T, Agarwal N. Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. Inf Process Manag 2021 Sep;58(5):102660 [FREE Full text] [doi: 10.1016/j.ipm.2021.102660] [Medline: 36567973]
67. Chipidza W. The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: an affiliation network approach. Int J Inf Manage 2021 Dec;61:102397 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2021.102397] [Medline: 34545262]
68. Alshalan R, Al-Khalifa H, Alsaeed D, Al-Baity H, Alshalan S. Detection of hate speech in COVID-19-related tweets in the Arab region: deep learning and topic modeling approach. J Med Internet Res 2020 Dec 08;22(12):e22609 [FREE Full text] [doi: 10.2196/22609] [Medline: 33207310]
69. Wani MA, ELAffendi M, Shakil KA, Abuhaimed IM, Nayyar A, Hussain A, et al. Toxic fake news detection and classification for combating COVID-19 misinformation. IEEE Trans Comput Soc Syst 2024 Aug;11(4):5101-5118. [doi: 10.1109/tcss.2023.3276764]
70. Khaskheli A, Jiang Y, Raza SA, Qamar Yousufi S. Social isolation and toxic behavior of students in e-learning: evidence during the time of the COVID-19 pandemic. Interact Learn Environ 2022 Oct 16:1-20. [doi: 10.1080/10494820.2022.2133145]
71. Awal MR, Cao R, Mitrovic S, Lee RK. On analyzing antisocial behaviors amid COVID-19 pandemic. arXiv. Preprint posted online July 21, 2020 2020 [FREE Full text]
72. DiCicco K, Noor N, Yousefi N, Maleki M, Spann B, Agarwal N. Toxicity and networks of COVID-19 discourse communities: a tale of two social media platforms. In: Proceedings of the 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2023: the 45th European Conference on Information Retrieval. 2023 Presented at: ROMCIR '23; April 2-6, 2023; Dublin, Ireland p. 1-13 URL: https://ceur-ws.org/Vol-3406/paper3_jot.pdf [doi: 10.4135/9781529604184]
73. Gover AR, Harper SB, Langton L. Anti-Asian hate crime during the COVID-19 pandemic: exploring the reproduction of inequality. Am J Crim Justice 2020 Jul 07;45(4):647-667 [FREE Full text] [doi: 10.1007/s12103-020-09545-1] [Medline: 32837171]
74. Fan L, Yu H, Yin Z. Stigmatization in social media: documenting and analyzing hate speech for COVID-19 on Twitter. Proc Assoc Inf Sci Technol 2020 Oct 22;57(1):e313 [FREE Full text] [doi: 10.1002/pra2.313] [Medline: 33173820]

Abbreviations

API: application programming interface
BERT: bidirectional encoder representations from transformers
BERTopic: bidirectional encoder representations from transformers–based topic modeling technique
CNM: Clauset, Newman, and Moore
c-TF-IDF: class-based term frequency-inverse document frequency
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and questioning
NLP: natural language processing
UMAP: Uniform Manifold Approximation and Projection
WHO: World Health Organization

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 21.09.23; peer-reviewed by M Haupt, A Regmi; comments to author 02.07.24; revised version received 22.07.24; accepted 28.10.24; published 12.12.24.

Please cite as:
Fan L, Li L, Hemphill L
Toxicity on Social Media During the 2022 Mpox Public Health Emergency: Quantitative Study of Topical and Network Dynamics
J Med Internet Res 2024;26:e52997
URL: https://www.jmir.org/2024/1/e52997
doi:10.2196/52997
PMID:

©Lizhou Fan, Lingyao Li, Libby Hemphill. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Discovering Time-Varying Public Interest for COVID-19 Case Prediction in South Korea Using Search Engine Queries: Infodemiology Study

Seong-Ho Ahn[1*]; Kwangil Yim[2*], MD, PhD; Hyun-Sik Won[1], MS; Kang-Min Kim[1,3*], PhD; Dong-Hwa Jeong[1,4*], PhD

[1]Department of Artificial Intelligence, The Catholic University of Korea, Bucheon-Si, Republic of Korea
[2]Department of Hospital Pathology, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
[3]Department of Data Science, The Catholic University of Korea, Bucheon-Si, Republic of Korea
[4]Department of Healthcare and Aritifical Intelligence, The Catholic University of Korea, Bucheon-Si, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Kang-Min Kim, PhD
Department of Artificial Intelligence
The Catholic University of Korea
Jibong-Ro 43 3-1
Bucheon-Si
Republic of Korea
Phone: 82 2 2164 5564
Email: kangmin89@catholic.ac.kr

Abstract

Background: The number of confirmed COVID-19 cases is a crucial indicator of policies and lifestyles. Previous studies have attempted to forecast cases using machine learning techniques that use a previous number of case counts and search engine queries predetermined by experts. However, they have limitations in reflecting temporal variations in queries associated with pandemic dynamics.

Objective: This study aims to propose a novel framework to extract keywords highly associated with COVID-19, considering their temporal occurrence. We aim to extract relevant keywords based on pandemic variations using query expansion. Additionally, we examine time-delayed web-based search behavior related to public interest in COVID-19 and adjust for better prediction performance.

Methods: To capture temporal semantics regarding COVID-19, word embedding models were trained on a news corpus, and the top 100 words related to “Corona” were extracted over 4-month windows. Time-lagged cross-correlation was applied to select optimal time lags correlated to confirmed cases from the expanded queries. Subsequently, ElasticNet regression models were trained after reducing the feature dimensions using principal component analysis of the time-lagged features to predict future daily case counts.

Results: Our approach successfully extracted relevant keywords depending on the pandemic phase, encompassing keywords directly related to COVID-19, such as its symptoms, and its societal impact. Specifically, during the first outbreak, keywords directly linked to COVID-19 and past infectious disease outbreaks similar to those of COVID-19 exhibited a high positive correlation. In the second phase of the pandemic, as community infections emerged, keywords related to the government’s pandemic control policies were frequently observed with a high positive correlation. In the third phase of the pandemic, during the delta variant outbreak, keywords such as “economic crisis” and “anxiety” appeared, reflecting public fatigue. Consequently, prediction models trained by the extracted queries over 4-month windows outperformed previous methods for most predictions 1-14 days ahead. Notably, our approach showed significantly higher Pearson correlation coefficients than models based solely on the number of past cases for predictions 9-11 days ahead (P=.02, P<.01, and P<.01), in contrast to heuristic- and symptom-based query sets.

Conclusions: This study proposes a novel COVID-19 case-prediction model that automatically extracts relevant queries over time using word embedding. The model outperformed previous methods that relied on static symptom-based or heuristic queries,

XSL•FO
RenderX

even without prior expert knowledge. The results demonstrate the capability of our approach to track temporal shifts in public interest regarding changes in the pandemic.

(J Med Internet Res 2024;26:e63476) doi:10.2196/63476

KEYWORDS

COVID-19; confirmed case prediction; search engine queries; query expansion; word embedding; public health; case prediction; South Korea; search engine; infodemiology; infodemiology study; policy; lifestyle; machine learning; machine learning techniques; utilization; temporal variation; novel framework; temporal; web-based search; temporal semantics; prediction model; model

Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus has highly affected the clinical system as well as social systems worldwide. COVID-19 is highly contagious and can be lethal to some individuals, especially older adults, those with underlying medical conditions, and immunocompromised individuals. Owing to the rapid global spread of COVID-19 in March 2020, 3 months after the first case was reported in December 2019 in Wuhan, China, the World Health Organization declared the outbreak a public health emergency of international concern. Most countries use the daily number of confirmed COVID-19 cases as a key indicator for determining quarantine policies to prevent its spread. These quarantine policies impact people's daily lives and economic activities, such as increasing delivery orders for food, allowing telemedicine, and enabling teleworking. Therefore, the daily number of COVID-19 confirmed cases has become a significant barometer for guiding policies and lifestyle changes during the post–COVID-19 era.

Numerous attempts have been made across several countries to forecast the number of confirmed COVID-19 cases using machine learning techniques [1-12]. Most studies have used a previous number of cases to capture the temporal dynamics of time-series data using autoregressive [2,3], conventional machine learning [4-6], and deep learning [7,8] models. However, using only the previous number of cases cannot reflect temporal variation due to external factors such as outbreaks with social events, new variants, and quarantine policy. To mitigate this limitation, some studies used additional features including traffic [9,10], mobile roaming [11], and search engine query [12] data. These approaches have demonstrated improvements in forecasting performance by reflecting human behavior and interest during the COVID-19 pandemic.

Web-based behaviors such as searching for and communicating on social networking services are significantly associated with the spread of infectious diseases [12-14]. In particular, the frequency of search engine queries has proven to be one of the most important complementary datasets for predicting infectious diseases such as COVID-19 [12,15-18]. Specifically, previous studies [15-18] demonstrated the feasibility of accurately predicting the number of confirmed COVID-19 cases by using search engine query data. A previous study [15] demonstrated that models incorporating search engine queries related to COVID-19 symptoms exhibited improved forecasting performance compared with those that did not include such queries. In these studies, query selection was important for extracting useful search engine queries that correlated with the confirmed cases. Previous studies [15-18] were selected based on correlation studies on medical expertise and COVID-19–related symptoms. However, because some studies [15-18] selected queries based on heuristic human knowledge (ie, selection by expertise knowledge), choosing relevant keywords demands a high level of expertise and continuous updating as time progresses. Consequently, previous studies may exhibit limitations in accurately reflecting real-world situations, as they did not consider the temporal fluctuations of search terms associated with COVID-19. To address this issue, it is essential to automatically extract relevant queries by considering the search behavior across different phases of the pandemic. We applied query expansion using word embedding to develop an automatic query extractor. Word embedding enables the understanding of word meanings by analyzing contextual word occurrences and empowering the automatic extraction of strong relationships among COVID-19–related terms in a corpus.

In this study, we propose a novel framework to extract terms strongly associated with COVID-19 considering their temporal occurrence. We used query expansion using word embedding to overcome the limitations of previous heuristic studies. From 8 major media outlets in South Korea, we collected news articles containing the word "COVID" by period. Then we trained word embedding models with the collected articles and selected the top 100 words with a high word similarity with "COVID" by period. We used these words to obtain expanded search engine query data. Additionally, we propose a time-lagged cross-correlation (TLCC) based feature selection method that considers the time delay between search engine queries and the number of confirmed cases. Subsequently, these terms were used as extended queries with an optimal time lag within the search engine data to predict the number of confirmed COVID-19 cases 1-14 days ahead.

In summary, our contributions are three-fold: (1) we developed a novel framework for daily COVID-19 case prediction by extracting efficient queries for query expansion by considering the temporal importance of queries, (2) the proposed query expansion can discover efficient queries for daily COVID-19 case predictions by capturing the transition of query trends over time, and (3) our framework can predict important keywords that change over time in real-world use by significantly outperforming the previous method of using search engine query data 1-14 days ahead of prediction.

XSL•FO
RenderX

Methods

Ethical Considerations

No ethics approval was required for this study, as it was not considered human participant research. All data were collected from publicly available sources. Data on confirmed COVID-19 cases were obtained from Our World in Data [19]. Search engine queries were collected from NAVER Data Lab (NDL) [20], one of the largest search engines in Korea.

Our Framework

Overview

Our framework consists of 3 components: query expansion, TLCC-based feature selection, and daily number of confirmed case predictions. Through query expansion, we performed word embedding of a COVID-19–related corpus collected beforehand, segmented by time period, followed by the extraction of word vectors. We then calculated the cosine similarity between these word vectors and the COVID-19–related terms to extract COVID-19–related queries. The extracted queries were used to gather search engine query data. By using the TLCC-based feature selection, we reflected on the time delay between search engine queries and the number of confirmed cases to determine the optimal time lag. Using the obtained optimal time lag, we conducted daily confirmed case predictions using expanded search engine queries. An overview of the proposed framework is presented in Figure 1.

Figure 1. Overview of our framework. The framework comprises 3 modules: query expansion, feature selection, and prediction model. PCA: principal component analysis; TLCC: time-lagged cross-correlation.

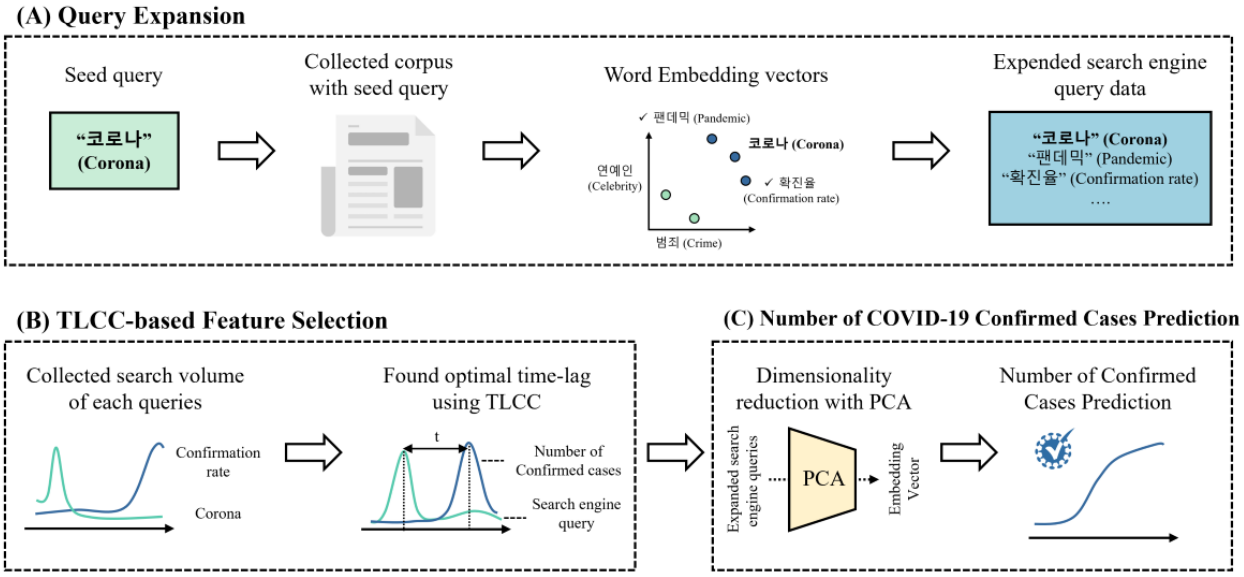

Query Expansion

Query expansion is a method aimed at enhancing search efficiency by identifying multiple words related to the seed query. Various approaches apply query expansion methods that consider numerous words similar to the seed queries. The first automatic query expansion method applied using a statistical approach was discussed in a previous study [21]. However, this approach has limitations, because it inadequately considers the semantic features between words. Word embedding, a technique for learning vector representations of word semantics, has been extensively studied in natural language processing. Word2Vec [22] is one of the most widely used word-embedding methods. It vectorizes the meaning of a word by considering the contextual meaning of the sentence and uses a shallow neural network to match the words appearing around it.

Previous studies [23,24] have demonstrated the applicability of word-embedding features to query expansion for generating more relevant keywords. A previous study [23] applied the Word2Vec model to word embedding for query expansion. Another study [24] proposed a method that considers topic-specific words in query expansion using Word2Vec. While previous methods [23] have focused on the global meaning of words, this approach [24] considers the local meaning of words based on domain knowledge. For example, in the previous method [23], terms closely associated with "gasoline tax," such as "cutting," "squeeze," and "reduce," are generated, capturing the global context of the topic. In contrast, the approach proposed in [24] generates terms like "tax," "deficit," and "vote," emphasizing a more localized semantic interpretation of domain knowledge.

Therefore, we applied this study's concept to COVID-19–specific words. This approach aids in extracting specific keywords related to COVID-19 and automatically extracts queries from valuable search engine query data to predict daily COVID-19 cases. We trained Word2Vec on a collected COVID-19–specific corpus, extracting similar words between "코로나" (translated as "Corona") and others using cosine similarity on the word vectors.

However, search engine query data are primarily driven by people's interests, which can change over time, leading to shifts in the perceived importance of keywords. Previous methods

XSL•FO
RenderX

have struggled to account for this temporal variation in semantics. To address this, we used a simple approach to transform data, considering the temporal variation in semantics over a window of 4 months based on the article publication date. We applied word embedding to each article with a striding 1-month interval to evaluate the model's performance throughout the entire period. This method not only accommodates temporal changes in semantics but also conserves storage space, requiring relatively small amounts of training data in real-world use environments and minimizing the training time. We excluded single-letter words and extracted the top 100 words with the highest cosine similarity because single-letter words are likely to be meaningless. We extracted search engine query data using queries obtained by query expansion including "코로나 (Corona)."

TLCC-Based Feature Selection

In previous studies [25-27], search engine queries were shown to exhibit a time lag, occurring earlier than the confirmation of cases. Because query expansion involves the extraction of various keywords, each feature has a critical time lag for prediction. Consequently, we propose a feature selection method based on the TLCC. TLCC is a method that considers time lag, facilitating the measurement of similarity between 2 signals. This method enables the extraction of features with the maximum Pearson correlation coefficient (PCC) between search engine queries and the number of confirmed cases at the optimal time lag. We calculated the TLCC up to 14 days prior to the present day for search engine queries extracted through query expansion and the number of confirmed cases. The equations for TLCC-based feature selection are denoted as Equations 1, 2, and 3:

where q is a set of search engine queries; i is an index of each search engine query; and l is the range of the time lag, which is a hyperparameter for the TLCC and is the search engine query selected by the TLCC-based feature selection. We used the range of l to $-14 \leq l \leq 0$.

Prediction of the Number of Confirmed COVID-19 Cases

After extracting 101 words related to "코로나 (Corona)" through query expansion, we obtained search engine query data. Using the NDL application programming interface (API), we collected data for both the extracted words and "코로나 (Corona)" over a time series spanning 4 months. Within this period, we designated the last 30 days as the test set, and the preceding days as the training set. Using feature selection based on the TLCC, we applied it to each of the 100 keywords extracted through query expansion to extract features with an optimal time lag in the training set.

Not all 100 features automatically extracted through query expansion contribute effectively to the prediction; we used principal component analysis [28] to filter and retain only pertinent information for prediction. We reduced the dimensionality of the features using principal component analysis, with an explained variance of 0.90. Additionally, following the approach of previous studies [17], we used the cumulative number of confirmed cases in the past 3 days as an additional feature.

For data normalization, we applied a min-max scaler. To predict daily COVID-19 cases using search engine query data, we used ElasticNet [29], a linear regression model with L1 and L2 regularizations. ElasticNet is commonly used for the time-series prediction of high-dimensional data. It has been applied to predict daily COVID-19 cases in a previous study [18]. We selected the L1 and L2 regularization hyperparameters through 5-fold cross-validation using the prequential approach, considering the characteristics of the time series [30].

Dataset

Daily Confirmed COVID-19 Case Data and Search Engine Query Data in South Korea

Daily COVID-19 case data represent the number of patients diagnosed with COVID-19 per day. This data was sourced from Our World in Data [19]. We used this dataset, covering the period from February 1, 2020, to November 30, 2021, for South Korea. The search engine query data represent the frequency of a search term on a search engine portal site per day. This study uses NDL API and Google Trends (GT), using queries extracted through the query expansion method, to predict confirmed COVID-19 cases in South Korea. We primarily conducted correlation analysis using NDL data, while the GT data were used for a comparative analysis of prediction performance against the NDL data. Data were collected for each of the 4 months, spanning from February 1, 2020, to November 30, 2021, considering various options such as age, sex, and device type. These values were normalized and ranged from 0 to 100, representing the actual search frequency within the user-defined period.

COVID-19–Specific Corpus Collection

Our focus was to build a corpus of texts highly relevant to COVID-19 to extract queries from search engine query data to predict daily COVID-19 cases. News articles from major media outlets tend to be more reliable than community posts and provide valuable information related to COVID-19; therefore, we collected news articles for query expansion. A COVID-19–specific corpus was amassed from news articles obtained from 8 news agency websites [31-38].

To mitigate the potential impact of political bias on keyword extraction, we balanced the selection of media outlets based on political learning. A total of 3 conservative, 3 liberal, and 2 centrist outlets were considered. To ensure the retrieval of exclusively COVID-19–related news articles, we crawled articles identified in search results using keywords such as "코로나19 (Corona19)," "코로나 (Corona)," "코로나 바이러스 (Coronavirus)," "신종 코로나바이러스 (Novel Coronavirus)," "COVID-19," and "코비드19 (COVID-19)" from February 2020 to December 2021, leveraging BIG KINDS [39], a news big data analysis tool. Basic information regarding the collected text data is summarized in Multimedia Appendix 1. To extract meaningful Korean words, we removed all non-Korean languages and special characters. To extract nouns from the corpus for training word embeddings, we used an Open Korean Text tokenizer [40], a Korean language tokenizer. Open Korean

XSL•FO
RenderX

Text tokenizer is trained on Twitter data and exhibits robustness to out-of-vocabulary issues, allowing it to distinguish COVID-19–related words, including neologisms such as "트윈데믹 (Twindemic)."

Model Evaluation

In the evaluation, we used the mean absolute percentage error (MAPE), root-mean-squared percentage error (RMSPE), and PCC to measure the performance of the time-series prediction. Previous studies [41,42] used the mean absolute error (MAE), root-mean-squared error (RMSE), and PCC. Our aim was for the model to perform consistently well across all 15 periods. However, owing to the dependence of the MAE and RMSE on the scale of the ground truth, it is challenging to consistently measure the model's performance across all periods in a consistent manner. Hence, we used MAPE and RMSPE as the percentage versions of MAE and RMSE, respectively. The MAPE and RMSPE quantify the disparity between the actual and predicted values as percentage values, with lower values being preferable. In contrast, PCC gauges the statistical resemblance between the actual and predicted values, with higher values being preferable. We compared the performance of our method with that of previous methods by averaging the evaluation metrics across all 15 periods. The equation of these evaluation metrics is denoted as Equations 4, 5, and 6.

Where y^t is the ground truth, is the mean of the ground truth, y is the predicted value, and is the mean of the predicted value.

In addition, we performed the Kruskal-Wallis H-test to compare our method with previous methods for multiple comparisons. We performed the Mann-Whitney U test for paired comparisons with False Discovery Rate Correction. This statistical analysis demonstrated that our method significantly outperformed previous methods in a statistically meaningful manner.

Results

Qualitative Analysis of Search Engine Query Data

We used extracted search engine query data based on the cosine similarity of word vectors correlated to "코로나" (translated as "Corona") which is a Korean word on subword of "코로나바이러스-19" (translated as "Coronavirus-19"). Therefore, it was necessary to collect a sufficient number of news articles to extract useful word vectors. Figure 2 demonstrates the number of confirmed COVID-19 cases along with the number of news articles when searching articles with the keywords "코로나19 (Corona 19)," "코로나 (Corona)," "코로나바이러스 (Coronavirus)," "신종 코로나바이러스 (Novel Coronavirus)," "COVID-19," and "코비드19 (Covid19)," which are paraphrased Korean word of "COVID-19." After the first case occurred on January 20, the number of articles changed over the following months. The highest number of news articles reported group infections caused by Shincheonji in March 2020. As the number of confirmed cases decreased, the number of news articles also declined gradually. In December 2020, with the resurgence of COVID-19, and in July 2021, with the increase in confirmed cases due to the delta variant, the number of news articles increased. However, following the emergence of the omicron variant in December 2021, despite a rapid increase in confirmed cases, the number of news articles decreased. In November 2021, the Korean government announced "recovery to a new normal," easing quarantine policy as vaccination rates increased and public fatigue with social distancing grew [43]. News articles were collected from February 2020 to November 2021, and queries were extracted within this period.

Figure 2. Number of articles and number of COVID-19 confirmed cases. The bar plot denotes the number of news articles related to COVID-19, while the blue line indicates the number of COVID-19 confirmed cases. The table at the top shows the social distancing system and levels announced by the Korean government.

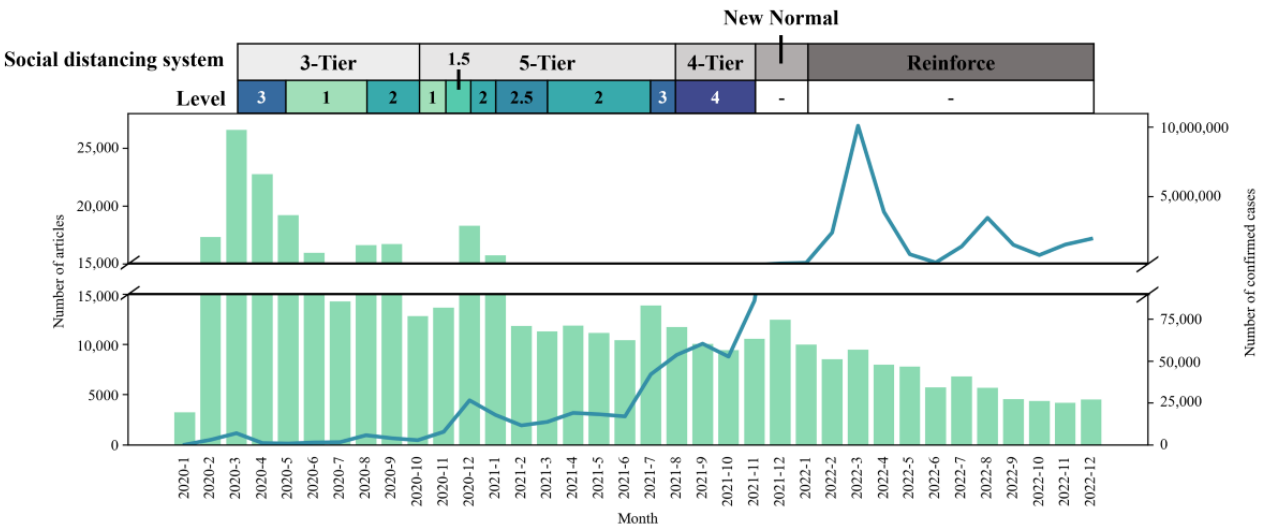

Next, we extracted word-embedding vectors from news articles over a 4-month time window, sliding this window by 1 month at a time. Subsequently, after removing single-character words, we selected 100-word vectors with the highest cosine similarity

XSL•FO
RenderX

with vectors corresponding to "코로나 (Corona)" as keyword queries for further analysis. As a result, 19 sets of 100-word embedding vectors were established between February 2020 and November 2021. For each 4-month interval, we collected search engine query data corresponding to the extracted keywords using the NDL API. We examined the association between the collected search engine queries and the actual number of COVID-19-confirmed cases by calculating the PCC for each period (Figure 3). All extracted keywords and PCC values are presented in Multimedia Appendix 2. Among the expanded queries, words directly related to COVID-19, such as "confirmation rate," "pandemic," and "influenza," exhibited a high PCC in the early stages of the pandemic, but demonstrated a decreasing trend over time (Figure 3A). Past outbreaks of infectious disease, "Middle East respiratory syndrome (MERS)," demonstrated a similar tendency in keywords directly related to COVID-19. Words related to government policy and public interest, such as "shutdown," "daily life," "tourism," "international travel," and "economic crisis," partially emerged in response to the decline or resurgence of COVID-19 (Figure 3B). Particularly, the term "economic crisis" exhibited a high positive correlation during periods, including December 2020, when there was a notable increase in COVID-19 cases. Conversely, "international travel" exhibited a negative correlation with the number of confirmed COVID-19 cases during these periods. In Figure 3C, words related to variants or transient events, such as "Delta variant," "twindemic," "the beginning of the semester," "canceled classes," and "face-to-face" are infrequently observed but show a high correlation when the corresponding events occur.

XSL•FO
RenderX

Figure 3. Correlation analysis about extracted queries by query expansion based on the pandemic phase. After categorizing each keyword using expert knowledge, PCC between the number of confirmed cases and (A) keywords directly related to COVID-19, (B) keywords related to public interest, and (C) keywords related to transient events were compared. If a keyword does not appear during the middle period, it is represented by a dotted line. MERS: Middle East respiratory syndrome; PCC: Pearson correlation coefficient.

Given the evidence from previous studies indicating time delay patterns between the number of confirmed COVID-19 cases and search engine queries, we conducted an analysis using TLCC to account for the possibility that search engine queries may precede or follow confirmed cases [25-27]. After shifting the search engine queries within the range of 14 days before to 14 days after, we calculated the PCC of each query based on the number of confirmed cases. The TLCC results are presented in Multimedia Appendix 3. Figure 4 illustrates the TLCC results of distinct keywords during the first outbreak (March 2020 to June 2020), the resurgence of COVID-19 resulting from community infections (December 2020 to March 2021), and the emergence of the delta variant (June 2021 to September 2021). During the initial outbreak, words directly related to

XSL•FO
RenderX

"Corona," such as "MERS," and "Influenza" exhibited a high correlation with the number of confirmed cases closely aligned with a time lag of zero, consistent with "Corona" (Figure 4A). Conversely, the term "economic activity" exhibited a negative correlation with the number of confirmed cases, reaching its highest absolute value with a time lag of 1, indicating 1 day after. During the resurgence of COVID-19, "corona," "pandemic," and "government," associated with cluster infections and subsequent policies during this period, exhibited a high positive correlation at negative time lags, particularly 9 days before the onset (Figure 4B). Conversely, the PCCs of these terms decreased at positive time lags. Similarly, the term "international travel" displayed a high negative correlation at negative time lags, especially 7 days before the onset. During the period when the delta variants emerged, high positive PCCs of words related to public and social responses, such as "economic crisis" and "anxiety," were observed at positive time lags, indicating that these search queries preceded the confirmation rate (Figure 4C). The term "Delta variant" showed a high negative correlation at positive time lags. However, the term "corona" displayed low PCCs during this period, differing from the higher relevance observed in previous periods of the first outbreak and community infections.

Figure 4. Time-lagged cross-correlation analysis depends on different pandemic phases. Keywords exhibiting high positive or negative PCCs at certain time lag were selected for each phase: (A) the first outbreak, (B) the community infection, and (C) the delta variants. Negative time lags indicate that a query preceded changes in COVID-19–confirmed cases, while positive time lags indicate that a query followed changes in case numbers. PCC: Pearson correlation coefficient.

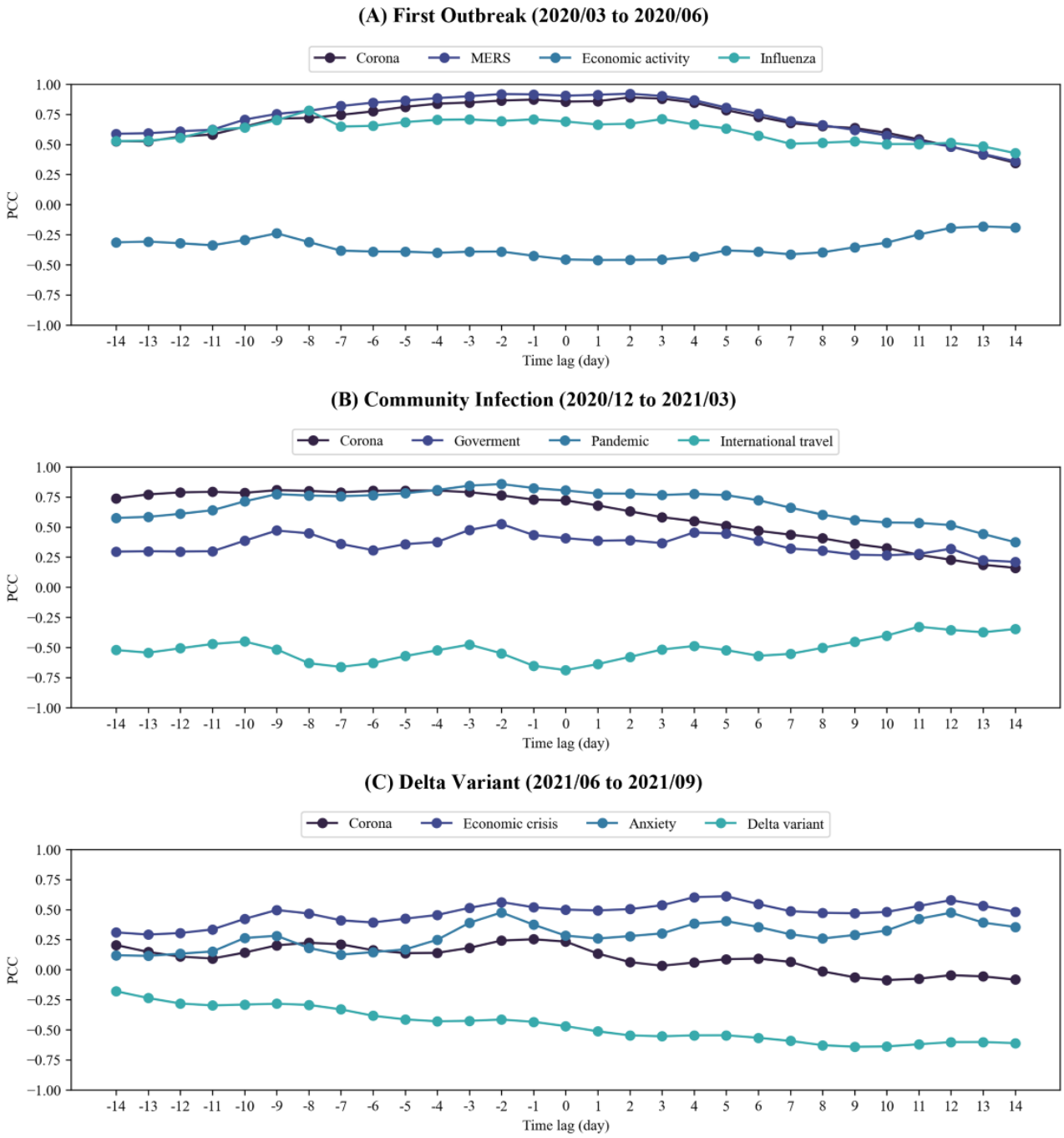

XSL•FO
RenderX

Prediction Performance for Daily Confirmed COVID-19 Cases

To evaluate the capability of the query expansion technique to dynamically extract informative keywords as an infectious disease progresses, we proposed a machine learning framework to predict the number of daily confirmed COVID-19 cases using the extracted queries. For each of the 19 four-month period, the first 3 months were used as the training set, whereas the last month was used as the testing set.

To demonstrate that query expansion-based keyword extraction is helpful in predicting daily COVID-19 cases by reflecting the variation in public interest in COVID-19, we compared the prediction performance of symptom-based keywords [18], heuristic-based keywords [17], and past daily confirmed COVID-19 cases only (case-only) [44], which were used in previous studies. Symptom-based keywords for the first few hundred [45] surveys identified 19 symptoms related to daily COVID-19 cases and their frequency of occurrence. Owing to the direct association between COVID-19 and its symptoms, it was used as the basis for symptom-based keywords for COVID-19 in a previous study [18]. To use these keywords, we translated 19 symptoms from the first few hundred into Korean to extract search engine query data from search engines in South Korea. Heuristic-based keywords are COVID-19–related keywords useful for predicting daily COVID-19 cases in South Korea, as selected in a previous correlation study [46]. These keywords also included terms indirectly related to COVID-19 (eg, social distancing and masks). We used the same keywords as in previous studies (Multimedia Appendix 4).

In Figure 5 [17,18], we averaged the results of 19 predictions over the entire time period and compared them to previous methods [17,18,44]; the results for each time period are shown in Tables S1, S2, and S3 in Multimedia Appendix 5. We predicted 1-14 days ahead to see how search engine query data extracted by query expansion affects the n-day-ahead prediction. In Figures 5A and 5B [17,18], our method has a lower MAPE than the other methods at 2 and 14 days ahead and a lower RMSPE than the other methods at 8, 10, and 14 days ahead. At 5, 7, and 14 days ahead, the search engine query data for the heuristic method showed a significant increase in MAPE and RMSPE compared with case-only. The MAPE and RMSPE were higher when the search engine query data on the symptom method were 5 days ahead and 11-14 days ahead than case-only. We can see that our method does not change much in MAPE as N-days-ahead increases, has less performance drop than the other methods, and has a lower MAPE and RMSPE than case-only as N-days-ahead increases. Regarding PCC, Figure 5C [17,18] shows that our method outperforms the other methods, except for 6, 7, and 13 days ahead. In particular, at 2, 9, 10, and 11 days ahead, PCC showed significant differences in multiple comparisons using the Kruskal-Wallis H-test (P=.04, P=.01, P<.01, and P<.01, respectively). In addition, the PCC showed significant differences between the case-only method and our method at 9, 10, and 11 days ahead of the post-hoc analysis using the Benjamini-Hochberg procedure (P=.02, P<.01, and P<.01). Overall, for most n-day-ahead evaluations, our method showed through MAPE, RMSPE, and PCC that the difference between the predicted and actual values was smaller than that of the other methods, while the predicted values were simultaneously highly correlated with the actual values.

XSL•FO
RenderX

Figure 5. Comparison of prediction performance between our approach, models without queries, and 2 previous query-based methods using experts' heuristics [17] and symptoms [18]. *P<.05, **P<.01, and ns=not significant. MAPE: mean absolute percentage error; PCC: Pearson correlation coefficient; RMSPE: root-mean-squared percentage error.

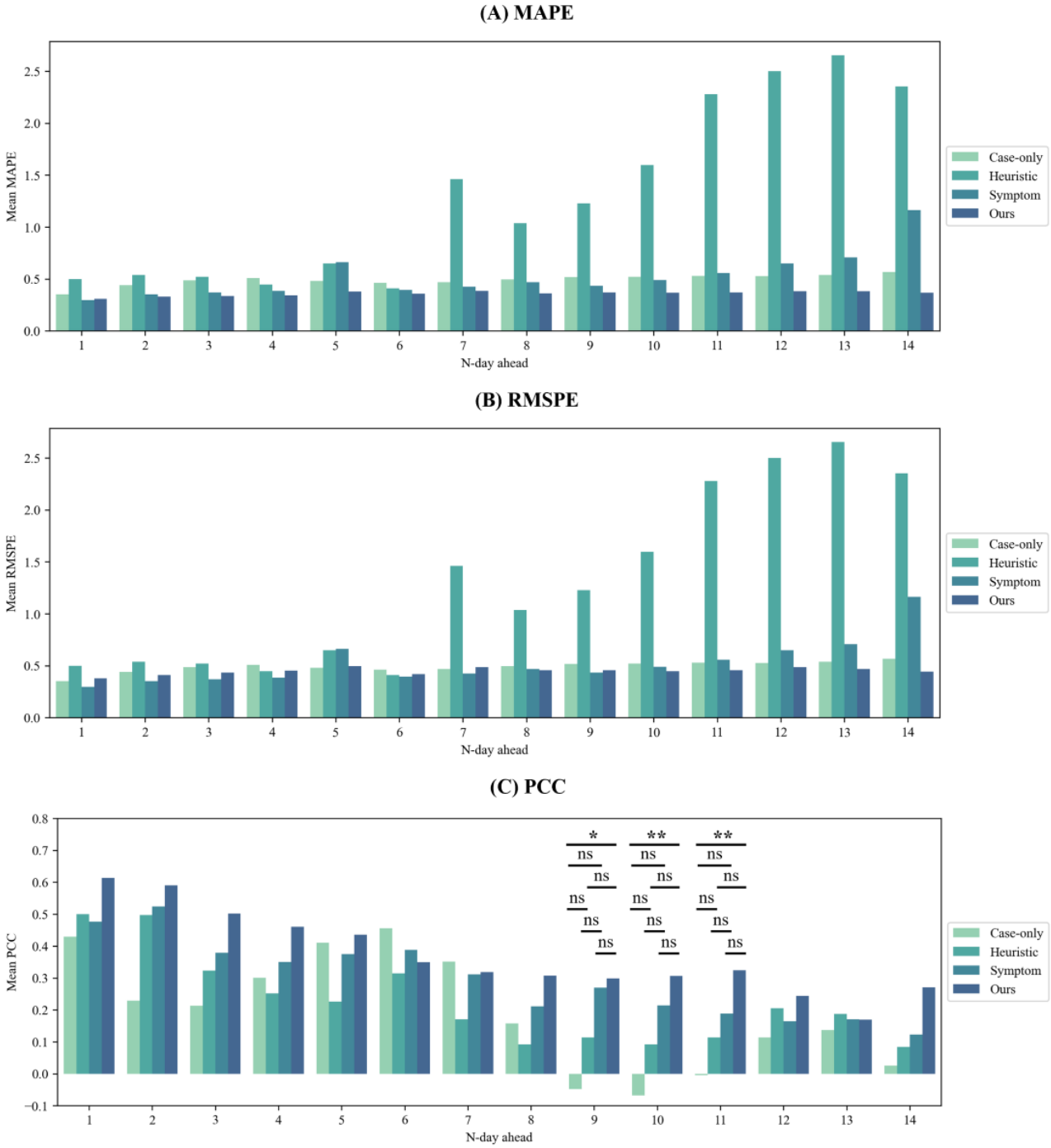

Figure 6 presents the regression plots comparing the ground truth values and the predicted values from the case-only method and our method. For the 1-day, 7-day, and 10-day ahead predictions, our method demonstrates a time series pattern that is more closely aligned with the ground truth compared with the case-only method. Notably, in the 10-day ahead prediction, where the PCC demonstrates statistical significance, both the case-only method and our method struggle to accurately predict sharp increases in the number of confirmed cases. However, our method more effectively captures the overall pattern of increases and decreases than the case-only method.

XSL•FO
RenderX

Figure 6. Regression plots on the ground truth values and the predicted values of the case-only method and our method in 1-, 7-, and 10-day ahead prediction.

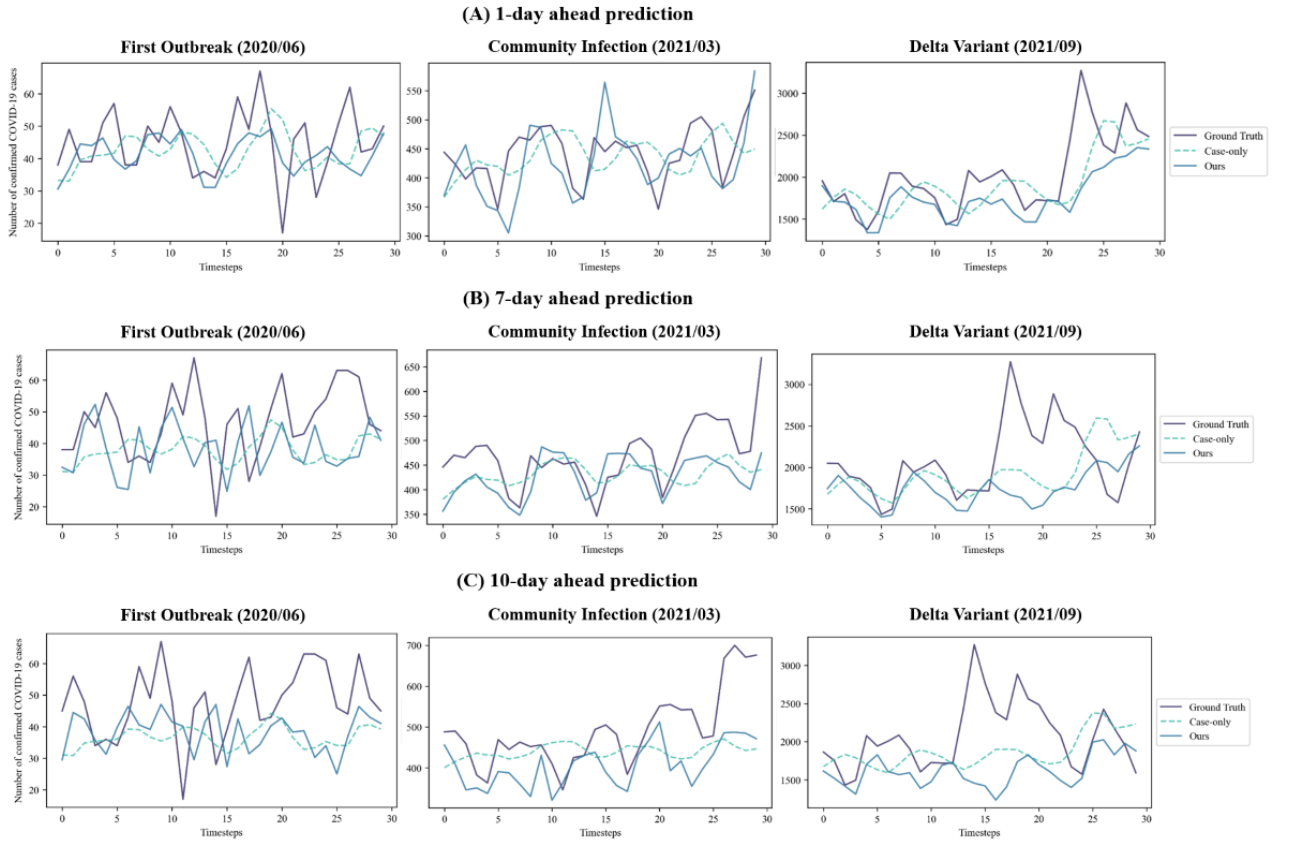

To evaluate the effectiveness between the GT data and NDL data, we compared MAPE, RMSPE, and PCC for the 10-day ahead predictions. Additionally, we conducted a post hoc analysis using the Mann-Whitney *U* test with the Benjamini-Hochberg procedure to verify whether the case-only method and other methods based on search engine query data show significant differences in performance. In Table 1, our method using NDL data achieved average MAPE, RMSPE, and PCC values of 0.3654, 0.4462, and 0.3069, respectively, outperforming our method using GT data, which recorded 0.4229, 0.5199, and 0.1918. The post-hoc analysis confirmed that our method using NDL data is the only approach that showed a statistically significant improvement over the case-only method (*P*=.02).

Table 1. Prediction performance comparison on Google Trends (GT) data and NAVER Data Lab (NDL) data in 10-day ahead prediction.

Method	Mean absolute percentage error, mean (SD)	Root mean squared percentage error, mean (SD)	Pearson correlation coefficient, mean (SD)	*P* value
Case-only	0.5214 (0.7251)	0.6177 (0.8590)	–0.0678 (0.2990)	__[a]
Heuristic (GT)	0.4824 (0.4581)	0.5738 (0.5711)	0.0664 (0.2891)	.44
Heuristic (NDL)	1.5956 (4.8904)	2.2247 (7.0910)	0.0921 (0.3270)	.44
Symptom (GT)	0.4542 (0.3421)	0.5333 (0.3982)	0.0587 (0.2816)	.44
Symptom (NDL)	0.4907 (0.4518)	0.6868 (0.7460)	0.2142 (0.3362)	.11
Ours (GT)	0.4229 (0.4143)	0.5199 (0.5567)	0.1918 (0.3276)	.15
Ours (NDL)	0.3654 (0.2983)	0.4462 (0.3958)	0.3069 (0.3074)	.02

[a]Not applicable.

Discussion

Principal Results

In this study, we proposed a novel approach that uses query expansion techniques to forecast the number of confirmed COVID-19 cases. Specifically, our approach involves automatically extracting search engine queries that are highly relevant to COVID-19 from internet documents. We crawled words that appeared alongside “corona” from news articles and embedded them as word vectors based on their similarity to “corona.” Consequently, the proposed system could effectively adjust to temporal variations in keyword prominence across the various phases of infectious diseases. Internet search behavior

XSL•FO
RenderX

patterns vary depending on the prevalence of keywords during specific periods. For example, queries related to symptoms, government policies, public concerns, and social events can emerge at various periods. To reflect these dynamics, queries from particular time intervals that demonstrated the highest TLCC were selected to predict the number of confirmed cases. To the best of our knowledge, this is the first attempt to forecast the number of COVID-19 cases by extracting textual data from web-based documentation using the word embedding method.

In addition, our model demonstrated superior performance compared with those from previous studies [17,18,44]. Furthermore, in predicting confirmed cases 9-11 days after keyword appearance, we exhibited statistically significant superior performance than using only the number of confirmed cases, in contrast to previous studies (P=.02, P<.01, and P<.01; Figure 5C [17,18]). From a medical perspective, the number of COVID-19 cases in the previous period serves as a powerful predictor of future cases [47]. Mathematical models of infectious diseases, such as susceptible-infected-recovered frameworks, have also used a number of confirmed cases as an essential predictor [47]. Moreover, in the initial stage, the number of infected populations was the most important factor, because changes in other factors were minimal. Therefore, we were able to statistically demonstrate that our model, unlike previous models, does not rely solely on the number of confirmed cases for prediction.

Predicting the daily number of confirmed COVID-19 cases can be useful for making decisions related to medical resource allocation and formulating epidemic/pandemic prevention policies [48]. Therefore, it can aid medical and governmental decision-making to prevent the spread of infectious diseases. Conventional studies that forecast the number of patients with infectious diseases have relied on time-series analyses that consider the temporal dynamics of past patient numbers, such as the autoregressive model and LSTM [2,3,8,44]. However, our results demonstrate that the prediction performance of models based solely on past confirmed cases deteriorates for future confirmed cases beyond 7 days (Figure 5C [17,18]). This could arise from external factors such as the emergence of variants, government policies, or social events that may intervene and cause fluctuations in daily COVID-19 cases owing to the virus' high contagiousness [49].

To address this issue, several studies incorporating search engine queries selected keywords based on researchers' experiences [17,46] or by using trending COVID-19–related keywords obtained from GTs [12,15,18,41,42], Baidu [16], and NDL [17]. Previous studies have mainly focused on keywords directly related to COVID-19 [17,41,42], keywords associated with COVID-19 symptoms [15,18], and individual preventive materials [17] (Multimedia Appendix 4). These heuristic approaches often functioned effectively, particularly in collaborations with experts. However, this process is time-consuming and resource-intensive, and the results can vary significantly depending on the expertise of the executor. We evaluated the predictive performance of our approach using 2 previously established query sets: symptom-based queries [15,18] and heuristic queries [17]. We identified that heuristic queries deteriorate the prediction performance in most cases compared with models that do not use queries (Figure 5 [17,18]), especially for those beyond 6 days. The symptom-based query resulted in higher PCCs compared with models without a query in all cases, except for 5, 6, and 7 days ahead; however, no statistical difference was found. Additionally, for predictions 11-14 days ahead, the MAPE and RMSPE showed considerable increases, particularly for longer prediction horizons.

We demonstrated that our approach outperforms previous methods in most scenarios. Our method exhibited the lowest MAPE in all N-day-ahead predictions, except for prediction 1-day ahead, where it ranked second following the model with a symptom-based query. Similarly, we also proved the robustness of our approach by observing the highest PCCs in all cases, except for predictions 6 and 7 days ahead, where it placed second, following models without queries. Across all 19 periods, the PCCs between the predicted values and actual outputs were positive, unlike in the other methods, which exhibited negative PCCs in several periods (Figure S3 in Multimedia Appendix 5). In particular, our framework showed the capability to predict confirmed cases more than 1 week in advance, unlike other approaches.

The superior prediction performance of our approach suggests that our framework, which incorporates query expansion techniques and TLCC, adeptly identifies keywords temporally associated with COVID-19 by tracking shifts in public interest regarding the pandemic [26]. Specifically, our approach demonstrated that, beyond keywords directly related to COVID-19 and keywords associated with symptoms or personal protective equipment, a variety of other keywords, including past outbreak infectious diseases such as "MERS" and societal impacts of COVID-19 such as "economic crisis," "the start of the semester," and "international travel," could also be effectively used to predict COVID-19 cases (Figure 3). These keywords suggest that using not only articles directly related to COVID-19 but also non–COVID-related or indirectly related articles can aid in extracting valuable keywords. Interestingly, MERS, a past infectious disease outbreak similar to COVID-19, exhibited a comparable trend of a high positive correlation during this stage. This result stemmed from numerous media articles that drew parallels and distinctions between COVID-19 and MERS.

To evaluate the generalization of our method in other infectious diseases, we applied the query expansion approach to MERS-CoV-2 (MERS) outbreak, which occurred in South Korea in 2015. After crawling news articles using the keyword "MERS" from the 2-month outbreak period, spanning May 20 (ie, the date of the first confirmed case) to July 19 (ie, the release date of the Intensive Care Hospital) in 2015, we analyzed the correlation between MERS-related queries and the number of cases. To evaluate the effectiveness of query expansion, we compared the correlation coefficients of our query set with those from previous studies [50,51], as presented in Multimedia Appendix 6. When comparing the top 3 keywords in terms of PCC, our method extracted "확진" ("confirmation," PCC=0.6533), "환자" ("patient," PCC=0.6472), and "감염" ("infection," PCC=0.5784). In contrast, the heuristic method extracted "메르스" ("MERS," PCC=0.5708), "MERS" (PCC=0.5478), and "메르스병원" ("MERS hospital,"

XSL•FO
RenderX

PCC=0.5176), while the symptom-based method extracted "기침" ("cough," PCC=0.3956), "설사" ("diarrhea," PCC=0.2687), and "두통" ("headache," PCC=0.2663). The results indicate that our method successfully extracted keywords with stronger correlations, demonstrating its potential for application across various infectious disease outbreaks to obtain search engine query data significantly correlated with confirmed case trends, thereby enhancing predictive performance. However, the MERS outbreak was much shorter than the COVID-19 pandemic, lasting approximately 2 months, with a peak in reported cases in early June. Developing predictive models for this epidemic is challenging due to the limited time-series data available, particularly for the period preceding the June outbreak. A primary limitation of our query expansion framework is its dependence on a substantial amount of textual data. The amount of text data from May 20 to early June may be insufficient to perform effective word embedding. The objective of correlation analysis was to determine whether the keywords generated through query expansion significantly correlated with the actual case counts, thereby assessing the effectiveness of our method in extracting relevant search queries under limited data conditions.

In addition, we discovered that keywords exhibiting a negative TLCC could also be effectively used to predict COVID-19 cases (Figures 3 and 4). Keywords such as "economic activity" and "international travel" may reflect this negative correlation due to a shift in public focus toward daily life as incidences of COVID-19 decline. The keyword "Delta variant" also showed a negative correlation, which could be attributed to the extensive attention it received prior to its spread in South Korea. However, once present in South Korea, the epidemiological characteristics did not significantly differ from those of previous variants [52,53], which may explain the observed negative correlation. For the MERS outbreak, we discovered keywords related to the MERS, quarantine, and infected facilities using TLCC analysis. The MERS-related keywords included "메르스" ("MERS"), "감염" ("Infection"), "전염" ("Contagion"), "환자" ("Patient"), "확진" ("Confirmation"), and "사망자" ("Dead"). Keywords related to quarantine were "격리" ("Isolation"), "보건복지부" ("Ministry of Health and Welfare"), and "정부" ("Government"). For infected facilities, "병원" ("Hospital") and "평택" ("Pyeongtaek") were extracted. Among the MERS-related keywords, "메르스" ("MERS"), "감염" ("Infection"), and "전염" ("Contagion"), as well as those related to quarantine and infected facilities, showed the highest correlation within a time lag range of -3 to -5 days, suggesting that these keywords reflect public interest prior to the confirmation of cases. In contrast, "환자" ("Patient"), "확진" ("Confirmation"), and "사망자" ("Dead") exhibited the highest correlations at time lags of 1, 2, and 4 days, respectively, indicating that these keywords capture increased public interest following case confirmations. Detailed TLCC results for MERS are provided in Multimedia Appendix 7.

We discovered changes in public interest toward the pandemic over time. During the first outbreak, there was a high positive correlation observed among keywords directly linked to COVID-19, such as "Corona," "Influenza," and "Confirmation rate" (Figures 3 and 4). This trend was likely a reflection of intensified public attention toward COVID-19 as a result of frequent search queries related to confirmed case counts. The PCCs of these keywords exhibited a declining trend over time, suggesting a potential decrease in public interest in the number of confirmed COVID-19 cases. In addition, by analyzing the TLCC during the first outbreak, most keywords were aligned with no time lag. During this period, the government accurately tracked all confirmed cases, and people's interest in COVID-19 was exceptionally high. Consequently, it can be hypothesized that there is a correlation between the frequency of keywords appearing in news articles and the number of COVID-19 cases, which may be closely aligned in real time.

In the community infection period, because the government started to fail to track all routes of confirmed cases, keywords regarding the government's pandemic control policies, such as "Shutdown" and "Daily life," were frequently observed with high positive correlations, alongside keywords directly associated with COVID-19 (Figure 3B). This observation suggests that increased public interest in government pandemic control policies could be a contributing factor. Interestingly, the appearance of COVID-19-related keywords tends to align well with the confirmed case numbers 7-14 days after their appearance, as indicated by the peaks in the negative time lags (Figure 4). It was hypothesized that an increase in COVID-19 cases impacts daily life, leading to an increase in keywords in news articles before individuals receive a confirmed diagnosis. Thus, during this period, an internet search–based artificial intelligence model could significantly aid in predicting COVID-19 cases. Consequently, our approach exhibited a superior prediction performance during these periods, particularly with the highest number of positive PCCs (Figure S3 in Multimedia Appendix 5).

During the delta variant period, there was a general decrease in the PCC between keyword frequency in news articles and the number of confirmed cases. Specifically, keywords directly related to COVID-19, including "Corona," showed relatively low TLCC in all time lags. Instead, keywords related to the societal impacts of COVID-19, such as "economic crisis" and "anxiety," exhibited high TLCCs in positive time lags. This indicates a tendency for these keywords to appear in news articles following an increase in confirmed cases. The patterns during this period suggest that public interest in COVID-19 diminished overall, and even in cases of symptoms or confirmed infections, the issue was not treated with urgency or importance. Because the prediction of confirmed cases should be performed based on previous search engine queries (ie, TLCCs at negative time lags), our approach demonstrated relatively low PCCs during these periods, despite showing comparably low MAPE and RMSPE values (Figure S3 in Multimedia Appendix 5).

Limitations

Despite the remarkable predictive performance of our approach, particularly for predictions beyond 1 week ahead, our study has several limitations. First, we collected and used a corpus of news articles to ensure data reliability. During the initial stages, when there was significant interest in the pandemic, the number of news searches and confirmed cases were relatively well correlated. However, as time progressed and interest in the

XSL•FO
RenderX

disease declined, the predictive capacity decreased. Further research incorporating a corpus that can effectively predict confirmed cases even when interest in news searches declines may hold substantial potential for improvement.

Second, owing to the reliance on news articles for data collection, the total amount of data yielded was relatively small. Consequently, we were limited to training Word2Vec on 4 months' worth of data for COVID-19 outbreak. This limitation could potentially be addressed by collecting additional, validated textual data on a large scale, or by devising methodologies that can enable word embedding or keyword extraction even with minimal data. From a methodological perspective on word embedding, FastText was found to be less suitable compared with Word2Vec-CBOW, as it tends to focus more on morphological features rather than the semantic aspects of words. Additionally, Word2Vec-Skipgram, which requires a larger number of training iterations and more data than Word2Vec-CBOW, demonstrated a tendency to extract less relevant keywords when trained under the same parameter settings and dataset, making it comparatively less effective. Overcoming this limitation may allow the extraction of keywords within a narrower temporal range, thereby enabling a more sensitive capture of shifts in public interest toward infectious diseases.

Conclusions

In this study, a COVID-19 case prediction model was developed using web-based search data and the Word2Vec algorithm with adaptability to temporal variations. This was achieved through direct text extraction from internet-based documents. The model outperformed previous models, even with automated keyword selection and without expert experience. Additionally, analysis of the keywords used provided insights into shifts in the public interest, correlating with changes in case numbers. This study highlighted the potential utility of using machine learning and search data to track and predict disease progression.

Acknowledgments

This work was supported by the National Research Foundation of Korea grant funded by the Ministry of Science and ICT (NRF-2021R1G1A1094236, 40%; NRF-2022R1C1C1010317, 30%). This work was also supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD; IITP-2024-RS-2024-00438207, 30%).

Authors' Contributions

SHA, DHJ, and KMK conceptualized the study. SHA and HSW collected the databases, refined the data, and developed the system. SHA and KY wrote the original manuscript. SHA analyzed and visualized the data. KMK and DHJ supervised this study and edited the manuscript. KMK and DHJ acquired the project funding. All the authors reviewed and approved the final version of the manuscript. SHA and KY equally contributed as first authors. DHJ and KMK equally contributed as corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Basic information of the collected text data from news articles.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v26i1e63476_app1.xlsx]

Multimedia Appendix 2
Extracted query lists using query expansion per period.
[XLSX File (Microsoft Excel File), 50 KB - jmir_v26i1e63476_app2.xlsx]

Multimedia Appendix 3
Time-lagged cross-correlation (TLCC) values and optimal time lag on all periods and queries.
[XLSX File (Microsoft Excel File), 61 KB - jmir_v26i1e63476_app3.xlsx]

Multimedia Appendix 4
Query lists about heuristics and symptoms in previous studies.
[XLSX File (Microsoft Excel File), 19 KB - jmir_v26i1e63476_app4.xlsx]

Multimedia Appendix 5
Performance tables and figures about averaged n-day ahead and averaged periods.
[DOCX File , 233 KB - jmir_v26i1e63476_app5.docx]

Multimedia Appendix 6

XSL•FO
RenderX

Pearson correlation coefficients of Middle East respiratory syndrome (MERS)–related queries.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v26i1e63476_app6.xlsx]

Multimedia Appendix 7
Time-lagged cross-correlation (TLCC) analysis of Middle East respiratory syndrome (MERS)–related queries.
[XLSX File (Microsoft Excel File), 229 KB - jmir_v26i1e63476_app7.xlsx]

References

1. Fard SG, Rahimi HM, Motie P, Minabi MAS, Taheri M, Nateghinia S. Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review. Heliyon 2021;7(10):e08143 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e08143] [Medline: 34660935]
2. Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health 2020;13(7):914-919 [FREE Full text] [doi: 10.1016/j.jiph.2020.06.001] [Medline: 32546438]
3. Chaurasia V, Pal S. COVID-19 pandemic: ARIMA and regression model-based worldwide death cases predictions. SN Comput Sci 2020;1(5):288 [FREE Full text] [doi: 10.1007/s42979-020-00298-6] [Medline: 33063056]
4. Kasilingam D, Sathiya Prabhakaran SP, Rajendran DK, Rajagopal V, Santhosh Kumar T, Soundararaj A. Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning. Transbound Emerg Dis 2021;68(3):1001-1018 [FREE Full text] [doi: 10.1111/tbed.13764] [Medline: 32749759]
5. Quintero Y, Ardila D, Camargo E, Rivas F, Aguilar J. Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables. Comput Biol Med 2021;134:104500 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104500] [Medline: 34052570]
6. Ribeiro MHDM, da Silva RG, Mariani VC, Coelho LDS. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Fractals 2020;135:109853 [FREE Full text] [doi: 10.1016/j.chaos.2020.109853] [Medline: 32501370]
7. Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V. Modeling the spread of COVID-19 infection using a multilayer perceptron. Comput Math Methods Med 2020;2020:5714714 [FREE Full text] [doi: 10.1155/2020/5714714] [Medline: 32565882]
8. Zeroual A, Harrou F, Dairi A, Sun Y. Deep learning methods for forecasting COVID-19 time-Series data: a comparative study. Chaos Solitons Fractals 2020;140:110121 [FREE Full text] [doi: 10.1016/j.chaos.2020.110121] [Medline: 32834633]
9. Zeng C, Zhang J, Li Z, Sun X, Olatosi B, Weissman S, et al. Spatial-temporal relationship between population mobility and COVID-19 outbreaks in South Carolina: time series forecasting analysis. J Med Internet Res 2021;23(4):e27045 [FREE Full text] [doi: 10.2196/27045] [Medline: 33784239]
10. Schwabe A, Persson J, Feuerriegel S. Predicting COVID-19 spread from large-scale mobility data. 2021 Presented at: KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2021 August 14; United States. [doi: 10.1145/3447548.3467157]
11. Choi SB, Ahn I. Forecasting imported COVID-19 cases in South Korea using mobile roaming data. PLoS One 2020;15(11):e0241466 [FREE Full text] [doi: 10.1371/journal.pone.0241466] [Medline: 33147252]
12. Ma S, Sun Y, Yang S. Using internet search data to forecast COVID-19 trends: a systematic review. Analytics 2022;1(2):210-227. [doi: 10.3390/analytics1020014]
13. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
14. Chew AWZ, Pan Y, Wang Y, Zhang L. Hybrid deep learning of social media big data for predicting the evolution of COVID-19 transmission. Knowl Based Syst 2021;233:107417 [FREE Full text] [doi: 10.1016/j.knosys.2021.107417] [Medline: 34690447]
15. Rabiolo A, Alladio E, Morales E, McNaught AI, Bandello F, Afifi AA, et al. Forecasting the COVID-19 epidemic by integrating symptom search behavior into predictive models: infoveillance study. J Med Internet Res 2021;23(8):e28876 [FREE Full text] [doi: 10.2196/28876] [Medline: 34156966]
16. Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, Davis J, et al. Real-time forecasting of the COVID-19 outbreak in Chinese provinces: machine learning approach using novel digital data and estimates from mechanistic models. J Med Internet Res 2020;22(8):e20285 [FREE Full text] [doi: 10.2196/20285] [Medline: 32730217]
17. Husnayain A, Shim E, Fuad A, Su ECY. Predicting new daily COVID-19 cases and deaths using search engine query data in South Korea from 2020 to 2021: infodemiology study. J Med Internet Res 2021;23(12):e34178 [FREE Full text] [doi: 10.2196/34178] [Medline: 34762064]
18. Lampos V, Majumder MS, Yom-Tov E, Edelstein M, Moura S, Hamada Y, et al. Tracking COVID-19 using online search. NPJ Digit Med 2021;4(1):17 [FREE Full text] [doi: 10.1038/s41746-021-00384-w] [Medline: 33558607]

XSL•FO
RenderX

19. Edouard M, Hannah R, Lucas R, Cameron A, Charlie G, Joe H, et al. COVID-19 pandemic. Our World in Data. URL: https://ourworldindata.org/coronavirus [accessed 2024-06-20]
20. Search terms trends. Naver Data Lab. 2024. URL: https://datalab.naver.com/keyword/trendSearch.naver [accessed 2024-06-20]
21. Maron ME, Kuhns JL. On relevance, probabilistic indexing and information retrieval. J ACM 1960;7(3):216-244. [doi: 10.1145/321033.321035]
22. Mikolov T, Chen K, Corrado G, Dean J. editors. Efficient estimation of word representations in vector space. 2013 Presented at: International Conference on Learning Representations, ICLR; 2013 May 4; USA.
23. Kuzi S, Shtok A, Kurland O. Query expansion using word embeddings. 2016 Presented at: CIKM'16: ACM Conference on Information and Knowledge Management; 2016 October 28; United States. [doi: 10.1145/2983323.2983876]
24. Diaz F, Mitra B, Craswell N. Query expansion with locally-trained word embeddings. In: 54th Annual Meeting of the Association for Computational Linguistics, ACL. 2016 Presented at: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024 November 15; Germany. [doi: 10.18653/v1/p16-1035]
25. Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with internet search volumes: a google trends analysis. Int J Infect Dis 2020;95:192-197 [FREE Full text] [doi: 10.1016/j.ijid.2020.04.033] [Medline: 32305520]
26. Higgins TS, Wu AW, Sharma D, Illing EA, Rubel K, Ting JY. Correlations of online search engine trends with coronavirus disease (COVID-19) incidence: infodemiology study. JMIR Public Health Surveill 2020;6(2):e19702 [FREE Full text] [doi: 10.2196/19702] [Medline: 32401211]
27. Cinarka H, Uysal MA, Cifter A, Niksarlioglu EY, Çarkoğlu A. The relationship between google search interest for pulmonary symptoms and COVID-19 cases using dynamic conditional correlation analysis. Sci Rep 2021;11(1):14387 [FREE Full text] [doi: 10.1038/s41598-021-93836-y] [Medline: 34257381]
28. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016;374(2065):20150202 [FREE Full text] [doi: 10.1098/rsta.2015.0202] [Medline: 26953178]
29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc 2005;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
30. Cerqueira V, Torgo L, Mozetič I. Evaluating time series forecasting models: an empirical study on performance estimation methods. Mach Learn 2020;109(11):1997-2028. [doi: 10.1007/s10994-020-05910-7]
31. Chosun Ilbo. 2024. URL: https://www.chosun.com/ [accessed 2011-07-26]
32. Joongang Ilbo. 2024. URL: https://www.joongang.co.kr/ [accessed 2024-11-15]
33. Dong-A Ilbo. 2024. URL: https://www.donga.com/ [accessed 2024-06-20]
34. Kyunghyang Shinmun. 2024. URL: https://www.khan.co.kr/ [accessed 2024-06-20]
35. The Hankyoreh. 2024. URL: https://www.hani.co.kr/ [accessed 2024-06-20]
36. Korean Broadcasting System. KBS News. 2024. URL: https://news.kbs.co.kr/news/pc/main/main.html [accessed 2024-06-20]
37. Munhwa Broadcasting Corporation. MBC News. 2024. URL: https://imnews.imbc.com/pc_main.html [accessed 2024-06-20]
38. Seoul Broadcasting System. SBS News. 2024. URL: https://news.sbs.co.kr/news/newsMain.do [accessed 2024-06-20]
39. Big Kinds. URL: https://www.bigkinds.or.kr/ [accessed 2024-06-20]
40. Open-Korean-Text. Github. 2018. URL: https://github.com/open-korean-text/open-korean-text [accessed 2024-06-20]
41. Ma S, Yang S. COVID-19 forecasts using internet search information in the United States. Sci Rep 2022;12(1):11539 [FREE Full text] [doi: 10.1038/s41598-022-15478-y] [Medline: 35798774]
42. Wang T, Ma S, Baek S, Yang S. COVID-19 hospitalizations forecasts using internet search data. Sci Rep 2022;12(1):9661 [FREE Full text] [doi: 10.1038/s41598-022-13162-9] [Medline: 35690619]
43. Chun JY, Jeong H, Kim Y. Identifying susceptibility of children and adolescents to the omicron variant (B.1.1.529). BMC Med 2022;20(1):451 [FREE Full text] [doi: 10.1186/s12916-022-02655-z] [Medline: 36419108]
44. Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 2020;135:109864 [FREE Full text] [doi: 10.1016/j.chaos.2020.109864] [Medline: 32390691]
45. Boddington NL, Charlett A, Elgohari S, Byers C, Coughlan L, Vilaplana TG, et al. Epidemiological and clinical characteristics of early COVID-19 cases, United Kingdom of Great Britain and Northern Ireland. Bull World Health Organ 2021;99(3):178-189 [FREE Full text] [doi: 10.2471/BLT.20.265603] [Medline: 33716340]
46. Husnayain A, Shim E, Fuad A, Su ECY. Understanding the community risk perceptions of the COVID-19 outbreak in South Korea: infodemiology study. J Med Internet Res 2020;22(9):e19788 [FREE Full text] [doi: 10.2196/19788] [Medline: 32931446]
47. Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. Nature 2020;584(7821):420-424. [doi: 10.1038/s41586-020-2554-8] [Medline: 32674112]
48. Xie L, Yang H, Zheng X, Wu Y, Lin X, Shen Z. Medical resources and coronavirus disease (COVID-19) mortality rate: evidence and implications from hubei province in China. PLoS One 2021;16(1):e0244867 [FREE Full text] [doi: 10.1371/journal.pone.0244867] [Medline: 33449940]

XSL•FO
RenderX

49. Lopez VK, Cramer EY, Pagano R, Drake JM, O'Dea EB, Adee M, et al. Challenges of COVID-19 case forecasting in the US, 2020-2021. PLoS Comput Biol 2024;20(5):e1011200 [FREE Full text] [doi: 10.1371/journal.pcbi.1011200] [Medline: 38709852]
50. Shin SY, Seo DW, An J, Kwak H, Kim SH, Gwack J, et al. High correlation of middle east respiratory syndrome spread with google search and twitter trends in Korea. Sci Rep 2016;6:32920 [FREE Full text] [doi: 10.1038/srep32920] [Medline: 27595921]
51. Alnuqaydan AM, Almutary AG, Sukamaran A, Yang BTW, Lee XT, Lim WX, et al. Middle east respiratory syndrome (MERS) virus-pathophysiological axis and the current treatment strategies. AAPS PharmSciTech 2021;22(5):173 [FREE Full text] [doi: 10.1208/s12249-021-02062-2] [Medline: 34105037]
52. Ryu BH, Hong SI, Lim SJ, Cho Y, Hong KW, Bae IG, et al. Features of COVID-19 among children and adolescents without risk factors before and after the delta variant outbreak in South Korea. Pediatr Infect Dis J 2022;41(1):e34-e35 [FREE Full text] [doi: 10.1097/INF.0000000000003394] [Medline: 34773397]
53. Choe YJ, Choi EH, Choi JW, Eun BW, Eun LY, Kim YJ, et al. Change in severity and clinical manifestation of MIS-C Over SARS-CoV-2 variant outbreaks in Korea. J Korean Med Sci 2023;38(30):e225 [FREE Full text] [doi: 10.3346/jkms.2023.38.e225] [Medline: 37527908]

Abbreviations

API: application programming interface
GT: Google Trends
MAE: mean absolute error
MAPE: mean absolute percentage error
MERS: Middle East respiratory syndrome
NDL: NAVER Data Lab
PCC: Pearson correlation coefficient
RMSE: root-mean-squared error
RMSPE: root-mean-squared percentage error
TLCC: time-lagged cross-correlation

Edited by A Mavragani; submitted 21.06.24; peer-reviewed by JY Chun, S Yang; comments to author 20.09.24; revised version received 11.10.24; accepted 29.10.24; published 16.12.24.

Please cite as:
Ahn SH, Yim K, Won HS, Kim KM, Jeong DH
Discovering Time-Varying Public Interest for COVID-19 Case Prediction in South Korea Using Search Engine Queries: Infodemiology Study
J Med Internet Res 2024;26:e63476
URL: https://www.jmir.org/2024/1/e63476
doi:10.2196/63476
PMID:

©Seong-Ho Ahn, Kwangil Yim, Hyun-Sik Won, Kang-Min Kim, Dong-Hwa Jeong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Introducing and Evaluating the Effectiveness of Online Cognitive Behavior Therapy for Gambling Disorder in Routine Addiction Care: Comparative Cohort Study

Olof Molander[1], PhD; Anne H Berman[2], PhD; Miriam Jakobson[3], MSc; Mikael Gajecki[3], PhD; Hanna Hällström[3], MSc; Jonas Ramnerö[1†], PhD; Johan Bjureberg[1], PhD; Per Carlbring[4], PhD; Philip Lindner[1,3], PhD

[1]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm Health Care Services, Region Stockholm, Stockholm, Sweden

[2]Department of Psychology, Uppsala University, Uppsala, Sweden

[3]Stockholm Centre for Dependency Disorders, Stockholm Health Care Services, Region Stockholm, Stockholm, Sweden

[4]Department of Psychology, Stockholm University, Stockholm, Sweden

[†]deceased

Corresponding Author:
Olof Molander, PhD
Centre for Psychiatry Research, Department of Clinical Neuroscience
Karolinska Institutet
Stockholm Health Care Services, Region Stockholm
Stationsgatan 69, 7th floor
Stockholm, SE-113 64
Sweden
Phone: 46 700011241
Email: olof.molander@ki.se

Abstract

Background: Several treatment-related challenges exist for gambling disorder, in particular at-scale dissemination in health care settings.

Objective: This study describes the introduction of a newly developed internet-delivered cognitive behavioral therapy (iCBT) program for gambling disorder (GD), provided with therapist support in routine addiction care, in a nationally recruited sample in Sweden. The study details the introduction of the iCBT program, evaluates its effectiveness and acceptability, and compares registry outcomes among iCBT patients with other patients with GD at the clinic who received face-to-face psychological treatment as usual.

Methods: The study site was the Stockholm Addiction eClinic, which offers digital interventions for addictive disorders in routine care. The iCBT program was introduced nationally for treatment-seeking patients through the Swedish eHealth platform. After approximately 2 years of routine treatment provision, we conducted a registry study, including ordinary patients in routine digital care (n=218), and a reference sample receiving face-to-face psychological treatment for GD (n=216).

Results: A statistically significant reduction in the Gambling Symptom Assessment Scale scores during the treatment was observed (B=–1.33, SE=0.17, $P<.001$), corresponding to a large within-group Cohen d effect size of $d=1.39$. The iCBT program was rated high for satisfaction. A registry-based survival analysis, controlling for psychiatric comorbidity, showed that patients receiving iCBT exhibited posttreatment outcomes (re-engagement in outpatient addiction care, receiving new psychiatric prescriptions, enrollment in psychiatric inpatient care, and care events indicative of contact with social services) similar to comparable patients who underwent face-to-face treatment-as-usual.

Conclusions: A lack of randomized allocation notwithstanding, the iCBT program for GD evaluated in this study was well-received by patients in routine addiction care, was associated with the expected symptom decrease during treatment, and appears to result in posttreatment registry outcomes similar to face-to-face treatment. Future studies on treatment mechanisms and moderators are warranted.

International Registered Report Identifier (IRRID): RR2-10.1186/s40814-020-00647-5

XSL•FO
RenderX

(J Med Internet Res 2024;26:e54754) doi:10.2196/54754

KEYWORDS

gambling disorder; internet-delivered cognitive behavioral therapy; routine addiction care; registry study; gambling; addiction; health care setting; iCBT; Sweden; feasibility; clinic; hospital; psychological treatment; digital intervention; addictive disorder; eHealth; digital care; survival analysis; comorbidity

Introduction

Previous-year population prevalence of problem gambling varies between 0.3% and 5.3% across countries [1] and in Sweden has been estimated to 2.1% [2]. Negative consequences related to problem gambling may occur in important life domains such as finances, relationships, or health, including a high rate of suicide ideation and attempts [3-5]. Problem gambling is a broad public health-based term, generally associated with poor mental and physical well-being. From a health care perspective, problem gambling behavior is subsumed under the gambling disorder (GD) diagnosis, formerly defined as an impulsive disorder, but included among addictive disorders at the introduction of the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition (*DSM-5*) [6]. The *DSM-5* specifies 9 diagnostic criteria for GD, observed within the past year. If at least 4 criteria are endorsed, the disorder is graded as mild (4-5 criteria), moderate (6-7), or severe (8-9) [6].

In Sweden, national prevalence rates, in combination with early studies indicating a higher problem gambling prevalence in social services and primary care samples [7,8], have led to political action. Since 2019, Sweden has introduced a semiopen gambling market, regulated by the Swedish Gambling Authority. Currently, there are 45 providers with commercial betting licenses and 56 with nonexclusive internet-based gambling licenses. These forms of gambling are most common among Swedish individuals with GD [9]. Licensed providers are required by law to provide responsible gambling tools. The primary responsible gambling tool is the Spelpaus.se voluntary self-exclusion registry, which allows gamblers to exclude themselves from all licensed land-based and internet-based gambling for various durations [10]. A recent survey from the Swedish Gambling Authority [11] indicates that 93% of daily gamblers are aware of Spelpaus.se. However, up to 30% of daily gamblers intentionally gamble with nonlicensed providers, which do not have the same duty of care obligations, including mandatory checks with the Spelpaus.se registry [11]. Furthermore, research shows that almost 40% of gamblers with an active Spelpaus.se self-exclusion continue to gamble with unlicensed providers [9].

Significant new legislation was also enacted in Sweden in 2018, which stipulated obligatory provision of GD treatment within health care and social services [12]. In recent years, evidence has accumulated for the effectiveness of psychological treatments, in particular cognitive behavioral therapy (CBT) [13]. CBT has been found effective in reducing gambling behavior and related problems in face-to-face individual and group therapies [14,15]. A recent meta-analysis also indicates lower, but promising effects for digitally delivered GD treatment [13]. In internet-delivered cognitive behavioral therapy (iCBT), treatment content is made accessible to the patient through a digital platform. The iCBT has been shown to be effective for a wide range of psychiatric as well as somatic disorders, where effects have shown equivalence to face-to-face delivered CBT [16]. Advantages specific to the internet format include cost-effectiveness, wide accessibility, scalability, and reducing perceived barriers to seeking treatment due to, for example, shame and stigma [17,18].

From a treatment dissemination and provision perspective, challenges remain. While introduction into regular health care is an obvious final goal for internet-based psychological treatment forms in general and is emphasized in treatment development models [19] and national eHealth strategies, few treatments that have been scientifically developed and evaluated are later introduced in routine health care settings and disseminated among regular patients (a study by Titov et al [20] shows an analysis of implementation of iCBT for anxiety and depression in routine care). Specific challenges exist regarding internet-based treatment for GD. First, most clinical trials evaluating iCBT for GD have been conducted in research settings, with participants recruited from the general population [17]. Second, although CBT and iCBT can be considered evidence-based treatments, the GD research field has not yet reached the same level of maturity as for other disorders [21]. Current treatments for GD often consist of a broad mix of general CBT interventions found effective for other psychiatric and addictive disorders (a study by Gooding and Tarrier [22] shows a review of GD treatment components) but lack a solid theoretical base in relation to gambling behavior [21]. It has not been understood how GD-related behavior is maintained in terms of clinical presentation, and researchers have recommended aligning treatment approaches with basic behavioral research [23,24]. However, current treatment manuals typically do not take contextual factors into consideration, including clinically pertinent factors such as exposure to gambling advertisements, low barriers to engage internet gambling, use of self-exclusion registries, and similar strategies. Third, GD is associated with high levels of psychiatric comorbidity [25]. In a meta-analysis, Dowling et al [26], found that 75% of treatment-seeking individuals with GD fulfilled an additional current *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (*DSM-IV*) Axis 1 diagnosis, such as mood disorders, alcohol use disorders, anxiety disorders, and substance use disorders. Addressing this high prevalence of psychiatric comorbidity is a pressing treatment-related GD challenge.

This study describes the introduction and subsequent effectiveness evaluation of a newly developed iCBT GD program provided with therapist support through asynchronous secure messages, within routine addiction care. The aims of the study were to (1) estimate the effectiveness and acceptability of the iCBT in a registry sample and (2) compare registry outcomes among iCBT patients with other patients with GD at

XSL•FO
RenderX

the clinic who received face-to-face psychological treatment-as-usual (TAU).

Methods

Setting, Study Design, and Ethics

This study was conducted at the Stockholm Addiction eClinic, a specialized unit within the Stockholm Centre for Dependency Disorders offering internet-delivered psychological treatments for addictive disorders to patients with alcohol [27] or cannabis use disorders [28]. In 2019, the treatment repertoire was expanded to include GD. When the GD treatment program was introduced as part of routine care, feasibility data were initially collected in a small sample. After clinical quality assurance revealed satisfactory outcomes, a decision was made to continue to offer iCBT within the context of routine care. After approximately 2 years of routine treatment provision, a registry study was conducted on routinely collected data. A parallel cohort design was used, comparing patients in the iCBT GD treatment to patients who had received TAU at regular face-to-face outpatient clinics at the Stockholm Centre for Dependency Disorders.

At the Stockholm Centre for Dependency Disorders, patients first seek treatment for GD, either through clinical or self-referral, are clinically assessed, and choose either face-to-face treatment (for a description on the clinical treatment seeking process, see [29]) or iCBT. Both iCBT and face-to-face treatments are intended to be standalone treatments.

Interventions

A novel iCBT program, grounded in learning theory, experimental behavioral research, and findings from interviews with patient representatives [21,24], was developed as part of a larger project [30] which aimed to conceptualize a clinical framework for the development and maintenance of GD. Briefly, the iCBT program emphasized 4 psychological processes relevant to GD. First, “access to money” is a key trigger for gambling behavior. Second, individuals with GD are likely to intensify their gambling behavior due to an exceedingly increased "reward expectancy". Third, “chasing behaviors” can be performed to chase wins and losses, and autopilot-chasing renders gambling behavior insensitive to actual aversive consequences. Fourth, “the gambling zone,” a psychological space where everything outside the gambling experience becomes irrelevant to gamblers as they become completely absorbed by the game [31], reinforces gambling behavior in itself, as well as an inability to terminate gambling once engaged.

The iCBT program consisted of 1 assessment module and 9 subsequent treatment modules, developed to encompass a simple, delimited set of core treatment components presumed to be of greatest importance for achieving control over problematic gambling behavior. Briefly, the program included psychoeducation on the 4 above-mentioned psychological processes. Patients were first instructed to register individual gambling-related situations. Thereafter, a treatment rationale was presented, emphasizing increased control over gambling behavior by completion of various behavioral exercises in gambling-related situations, in which patients were encouraged to act differently or in ways opposite than before to widen their behavioral repertoire and challenge presumptions. Patients then rated various behavioral exercises in a difficulty-rating task. After this, they performed repeated behavioral exercises with increasing difficulty, under the supervision of their assigned therapist (a licensed clinical psychologist) through weekly asynchronous secure platform messages. In these continuous messages, the iCBT therapists monitored treatment engagement and progress, provided personalized feedback on tasks, and were available to answer questions. If need be, the iCBT therapists contacted the patients, who in turn could also contact their iCBT therapist by telephone. In the beginning of the iCBT program, a standard information text instructed the patients to complete the treatment in 16-18 weeks, but time extensions were allowed if deemed clinically appropriate by the iCBT therapists.

TAU consisted of systematic psychological treatments delivered face-to-face at any of the outpatient clinics at the Stockholm Center for Dependency Disorders (nonsystematic interventions such as support conversations were excluded). Current GD treatments often lack a solid theoretical base in relation to gambling behavior [21], as suggested by the 4 central GD processes in the iCBT program. The face-to-face treatments delivered in routine addiction care therefore included a range of disparate methodological approaches and interventions, such as CBT, cognitive therapy, motivational interviewing, relapse prevention, and psychodynamic therapy, with the following national treatment provision codes: DU008, DU009, DU010, DU011, DU013, DU020, DU043, DU113, DU118, DU119, and DU120.

Recruitment Criteria

Standard clinical eligibility criteria were used to maximize external validity. All treatment-seeking gamblers with problem gambling or GD during the past year, who were deemed clinically suitable, were offered iCBT; no prospective patient was excluded exclusively on the grounds of psychiatric comorbidities. Since patients were recruited from a clinical setting, it was not possible to accurately log how many were exposed to the offer of iCBT but turned it down.

Treatment Introduction and Feasibility Study

The iCBT program was made available using the Swedish national eHealth platform (Stöd och behandling, Treatment and Support [TAS] platform). The TAS platform is integrated into the national health care guide 1177.se, allowing prospective patients to conveniently access information, patient services, and internet-based treatments within the same platform. Access required a bank-issued digital ID, after which notifications by SMS text message and email could be enabled by each individual user. During this first phase of the study, a feasibility study was conducted among 23 patients (“Feasibility study” in Multimedia Appendix 1 [32,33]). When the feasibility study was terminated in July 2020, clinical quality assurance work and preliminary effects indicated promising results, and the provision of the iCBT continued within the framework of regular specialist addiction care.

XSL•FO
RenderX

Registry Study

After approximately 2 years of routine treatment provision, data from iCBT patients starting between July 2020 and August 2022 were compiled using a stepwise procedure to create a registry cohort. First, data were exported from the TAS platform that included all iCBT patients finalized at least 3 months before data export, in order to provide a minimal follow-up period for the registry outcomes. This resulted in 218 participants remaining, after removal of duplicates and 1 mislabeled test patient. Since duplicates could only result from a given patient starting and ending the program multiple times, we opted to keep only one enrollment per patient in order to approximate the intention-to-treat principle; the first treatment enrollment was selected if the TAS log file contained any activity (3 cases out of 4). Next, a health care informatician identified up to 3 reference patients for each included iCBT patient (randomly selected in case of >3 or more available) based on age (±5 years) and gender. Only iCBT patients residing in Stockholm County were matched with reference patients, to avoid introducing possible bias clustered on region of residence, that is, all patients had the same theoretical opportunity to engage in new outpatient addiction care in Region Stockholm during the registry outcome period. Raw registry data were compiled for both arms as per the description below and were then shared with the researchers. While the raw reference sample included 488 patients, 193 were subsequently excluded due to a parallel, documented care commitment at the Addiction eClinic at any time during the examined time frame, leaving 295 reference patients. Next, only reference patients who had a care event marked as constituting a systematic psychological treatment (provision codes DU008, DU009, DU010, DU011, DU013, DU020, DU043, DU113, DU118, DU119, and DU120) were retained, leaving 225 reference patients. An additional 6 reference patients were then excluded due to having care events only before 2019. This exclusion criterion was applied to better correspond with the same point in time during which the iCBT patients were undergoing treatment; this included new gambling market legislation coming into effect in 2019 in Sweden, including introduction of a national self-exclusion registry used by licensed gambling providers. Finally, 3 patients in the reference group were excluded for being <18 years, the age required for licensed gambling in Sweden, and the age limit applied at the iCBT clinic. Demographic characteristics are given in Table 1.

XSL•FO
RenderX

Table 1. Participant characteristics.

	iCBT[a]		TAU[b]	
	Total (n=218)	Stockholm-residing (n=123)	Total (n=216)	Matched (n=123)
Age (years), mean (SD)	36.3 (10.5)	36.1 (10.1)	37.8 (10.6)	37.3 (10.7)
Gender, n (%)				
Men	167 (77)	96 (78)	175 (81)	102 (83)
Women	51 (23)	27 (22)	41 (19)	21 (17)
Diagnostic codes[c], n (%)				
Gambling disorder (F630)	213 (98)	119 (97)	203 (94)	117 (95)
Other *ICD*[d] diagnoses	29 (13)	22 (18)	84 (39)	21 (17)
Mental and behavioral disorders due to use of psychoactive substances	19 (9)	15 (12)	51 (24)	17 (14)
Mood/affective disorders Neurotic, stress-related and somatoform disorders	5 (2)	4 (3)	29 (13)	3 (2)
	5 (2)	4 (3)	29 (13)	4 (3)
Disorders of personality and behavior in adult persons	3 (1)	2 ()	6 (3)	1 (1)
Previous treatments[e], n (%)				
Psychological treatment in outpatient addiction care[f]	12 (6)	11 (9)	20 (9)	7 (6)
Prescriptions of psychiatric medication in addiction care	11 (5)	11 (9)	44 (20)	10 (8)
Psychiatric inpatient enrollment Involvement of social services	2 (1)	2 (2)	13 (6)	2 (2)
	1 (0)	1 (1)	14 (6)	1 (1)

[a]iCBT: internet-based cognitive behavioral treatment at the Stockholm Addiction eClinic. Non-Stockholm patients were excluded in the Stockholm sample.

[b]TAU: treatment-as-usual, that is, psychological treatment delivered face-to-face at any of the outpatient clinics at the Stockholm Center for Dependency Disorders. Patients at the eAddiction Clinic and non-Stockholm patients were excluded.

[c]Diagnoses linked to care-events; 2 years before treatment start to 6 months follow-up.

[d]ICD: International Classification of Diseases.

[e]2 years before treatment start to treatment start.

[f]The systematic psychological treatments included the following national treatment provision codes: DU008, DU009, DU010, DU011, DU013, DU020, DU043, DU113, DU118, DU119, and DU120.

Outcomes

Severity of Gambling Symptoms

GD symptoms were measured using the Gambling Symptom Assessment Scale (GSAS) [34], which served as the primary outcome according to the study protocol [32]. The GSAS is a 12-item self-report measure to assess the severity of gambling symptoms over the past week, aiding clinicians and researchers in tracking symptom progression during treatment. The GSAS has been found to be sensitive to treatment-related change [34]. The GSAS was administered with each iCBT module during treatment. No equivalent measure was available from the TAU arm of the cohort sample.

Gambling Activity

Gambling during the iCBT treatment was measured during each module using the Timeline Followback method [35]. Due to the low prevalence of gambling observed during treatment, along with high prevalence of reported use of stimuli control techniques, we opted to analyze gambling habits as a binary variable (any or none). No equivalent measure was available from the TAU arm of the cohort sample.

Negative Effects and Treatment Satisfaction

Any negative effects of the iCBT program were measured in conjunction with the last module, using the Negative Effects Questionnaire [36]. Relative occurrence of each reported negative effects is reported, along with distribution of attribution (treatment itself or other causes). Treatment satisfaction was measured at the same time using the Client Satisfaction Questionnaire (CSQ)–8 [37]. Furthermore, a tailored evaluation form was included with which patients rated which components of the treatment program they found helpful. No equivalent measures were available from the TAU arm of the cohort sample.

XSL•FO
RenderX

Registry Outcomes

From raw, timestamped data on care events, care commitments, and medication prescriptions, a set of registry-based outcomes were created using a stepwise procedure. First, an index date for treatment onset was created using the timestamp of relevant treatment provision codes. To avoid bias stemming from presumed differences in treatment duration between iCBT and TAU arms, patients in both arms were assigned a second index date corresponding to 6 months after treatment onset. Next, registry outcomes suggesting continued need for psychiatric treatment were computed using timestamped event data: new treatment onset within addiction care in Stockholm (GD-specific or not), new prescription of psychiatric medication, new enrollment in psychiatric inpatient care (GD-specific or not), and a registry entry indicative of social service involvement (in Sweden, health care and social services have a shared responsibility for providing treatment for GD). Examining both GD-specific and other treatment occurrences served to both control for detection bias (eg, systematic differences in tagging care event with the relevant diagnosis code), and to widen the scope of the outcome to cover, for example, psychiatric comorbidity (detailed outcome definitions in Table S2 in Multimedia Appendix 1 [32,33]).

Statistical Analyses

To examine within-group change in the iCBT cohort during participants' (varying) treatment windows, we modeled GSAS scores and gambling activity as a function of assessment occasion, the latter captured by a numeric variable (0-10) corresponding to the module. Since program composition changed during the cohort data collection period (10 modules being collapsed into 9), normalization of the time variable was required across program versions. While we initially considered making full use of the linear time variable by including noninteger timepoints, we opted instead to include all available data with the original sequence-based time integers but also to cap the primary end point at module 9. This decision was in part guided by the finding that only 1 patient undergoing the 10-module version reported 10 measurements. Changes during the treatment window were modeled using mixed-effects models (lmer and lmerTest) [38], either linear (GSAS) or logistic (any gambling), including random intercepts and slopes. GSAS within-group effect sizes were calculated using the standard Cohen *d* equation by extrapolating means from the mixed models (ensuring compliance with intention-to-treat and equal n) and observed SDs at the corresponding end point.

Registry outcomes were (retrospectively) gathered continuously (ie, not at fixed timepoints), and were therefore analyzed using survival analysis [39]. To adjust for the increased presence of psychiatric severity in the age- and sex-matched comparison patients (Table 1), survival analyses were run using both the complete TAU sample, as well as a matched subsample. This subsample was created using a propensity score matching procedure, using the 1:1 nearest neighbor method (MatchIt R package) [40], with past diagnoses and care as separate predictors.

Ethical Considerations

All stages of the study were approved by the Swedish Ethical Review Authority (#2019-05479 and #2022-04987-02). All feasibility participants provided informed consent to data sharing for research purposes.

Results

Treatment Engagement

In the total iCBT cohort, patients sent an average of 5.87 (SD 6.10, median 4, range 0-28) messages to their therapists and began on average 4.16 (SD 3.10) modules, defined as completing at least 1 exercise per module (Figure S1 in Multimedia Appendix 1 for histogram [32,33]). The mean number of iCBT weeks during the treatment window was 11 (SD 7, range 0-26) weeks. A total of 6 therapists provided the iCBT treatment. The therapists were not randomized to patients, but assigned patients based on availability, that is, quasi-random. For the total TAU cohort, the mean treatment duration during the treatment window was 15 (SD 7, range 0-26) weeks. On average, patients received 6.34 (SD 4.5) treatment sessions face-to-face.

Change in Symptoms and Gambling Activity

Outcome measures were available for 188 (out of 218) patients who began the iCBT program. A statistically significant reduction in GSAS scores during the treatment window was observed (B=–1.33, SE=0.17, $P<.001$), corresponding to a large within-group Cohen *d* effect size, $d=1.39$. Visualization of change in symptoms are shown in Figure 1. No change in gambling activity was observed ($P=.68$). However, the average occurrence of any gambling during the treatment window was low, reported on only 19.4% of assessments, consistent with high rates (86.4%) of used stimuli control techniques (eg, the Spelpaus.se self-exclusion registry) during treatment.

XSL•FO
RenderX

Figure 1. Change in gambling disorder (F630) symptoms during the internet-based cognitive behavioral therapy (n=188). GSAS: Gambling Symptom Assessment Scale.

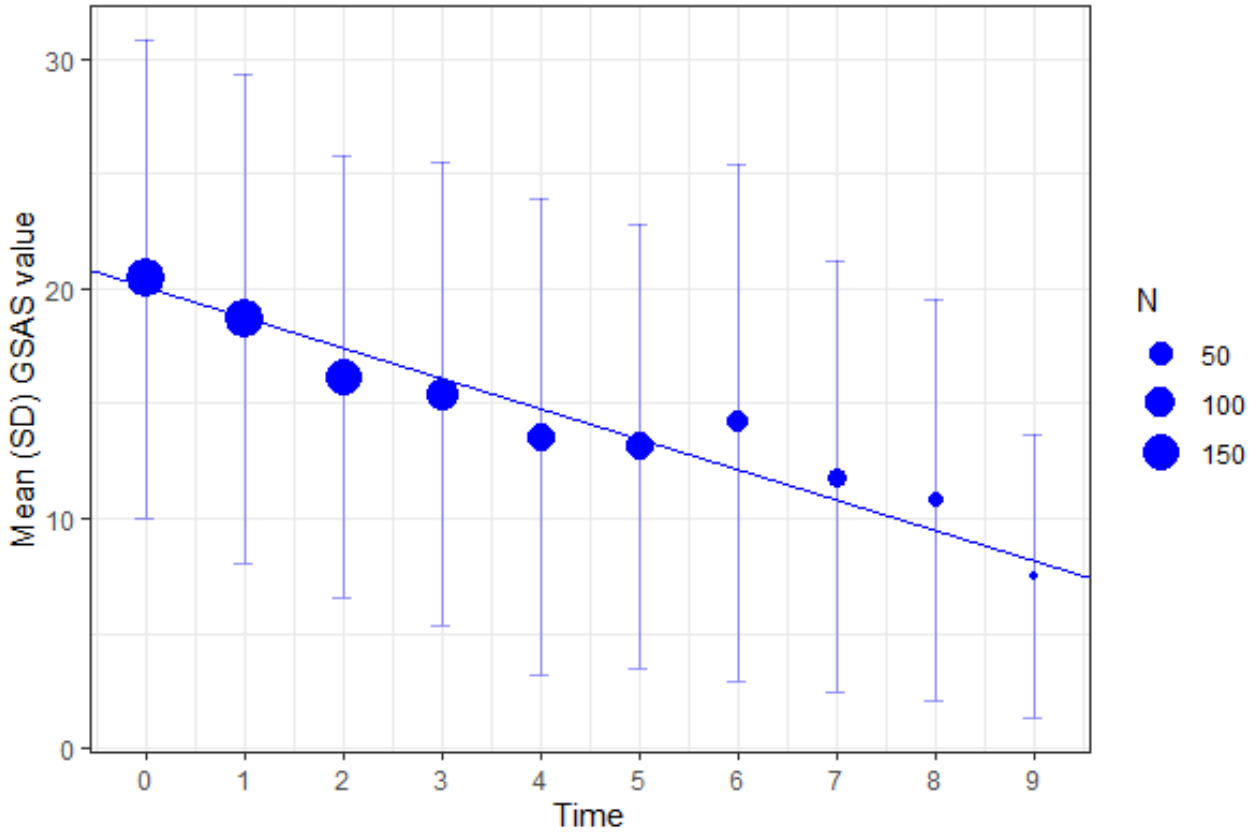

A post hoc, exploratory analysis examined whether having any care event-marked comorbid psychiatric diagnosis moderated the change in GSAS score during treatment. A significant time×subgroup interaction effect (B=1.33, SE=0.67, *P*=.048) was found, revealing that not only did the comorbidity subgroup have a different slope, but also no significant change over time at all, despite similar intercepts (B=–0.42, SE=2.67, *P*=.88). Subsequent analyses revealed that the comorbidity subgroup also engaged in significantly less iCBT modules (ΔM=–1.63, *P*=.01)

Negative Effects and Treatment Satisfaction

The assessment of negative effects was completed by 30 patients at the end of the iCBT program, and 20 patients reported at least 1 negative effect, with an average of 3.05 (SD 2.28) negative effects reported per patient. The most commonly reported negative effects included increased adverse emotional experiences such as unpleasant memories, stress, anxiety, or negative emotions. Patients attributed an average of 72% of the reported negative effects to the iCBT program and 28% to other circumstances, but this differed substantially between specific negative effects (100%-0% attribution). In addition, 2 patients reported suicidal thoughts, both unrelated to the iCBT (Table 2). Regarding treatment satisfaction, patients rated the iCBT program with a posttreatment mean rating of 27.9 (SD 2.94; n=32) out of 32 on the CSQ-8. In a separate evaluation form, patients (n=25) reported which specific useful treatment components they found useful, the most prevalent being reading about gambling-related loss of control and common reactions, reading about why people get stuck in gambling problems, and doing behavioral exercises when thinking about chasing losses (Table S3 in Multimedia Appendix 1 [32,33]).

XSL•FO
RenderX

Table 2. Self-reported negative effects of treatment in cohort sample.

Type of negative effect (NEQ[a])	Reported by proportion of patients (total n=20), %	Attributed to iCBT[b], %	Other circumstances, %
Unpleasant memories resurfaced	80	94	6
I did not always understand my treatment	40	88	12
I experienced more anxiety	35	57	43
I experienced more hopelessness	35	86	14
I became afraid that other people would find out about my treatment	25	80	20
I felt like I was under more stress	20	50	50
I felt more worried	20	50	50
I started feeling ashamed in front of other people because I was having treatment	15	67	33
I got thoughts that it would be better if I did not exist anymore and that I should take my own life	10	0	100
I had more problems with my sleep	5	0	100
I experienced more hopelessness	5	0	100
I stopped thinking that things could get better	5	0	100
I did not always understand my therapist	5	100	0
I felt that the treatment was not motivating	5	0	100

[a]NEQ: Negative Effects Questionnaire [36].

[b]iCBT: internet-delivered cognitive behavioral therapy.

Registry Outcomes

Survival analyses revealed that patients in the iCBT arm, compared with the full TAU arm, were significantly less likely to re-engage in outpatient addiction care, GD-specific (P=.01) or not (P<.001), less likely to receive new psychiatric prescriptions (P<.001), less likely to be enrolled in psychiatric inpatient care (P=.01), and to show care events indicative of contact with social services (P=.04). When the iCBT arm was compared with the matched TAU subsample, none of the differences in outcomes remained significant, except non–GD-specific re-engagement in outpatient addiction care (P=.01) (Figure 2 and Table 3). The figure shows survival analyses of re-engagement in outpatient treatments and prescriptions of psychiatric medications in routine addiction care in 2 GD treatment samples; a total sample (n=339) that received either iCBT (n=123) or TAU (n=216) and a subsample (n=246) matched on comorbidity that received either iCBT (n=123) or TAU (n=213).

XSL•FO
RenderX

Figure 2. Survival analyses: new treatments and prescriptions. GD: gambling disorder (F630); iCBT: internet-delivered cognitive behavioral therapy at the Stockholm Addiction eClinic; TAU: treatment-as-usual, that is, systematic psychological treatment delivered face-to-face at any of the outpatient clinics at the Stockholm Center for Dependency Disorders.

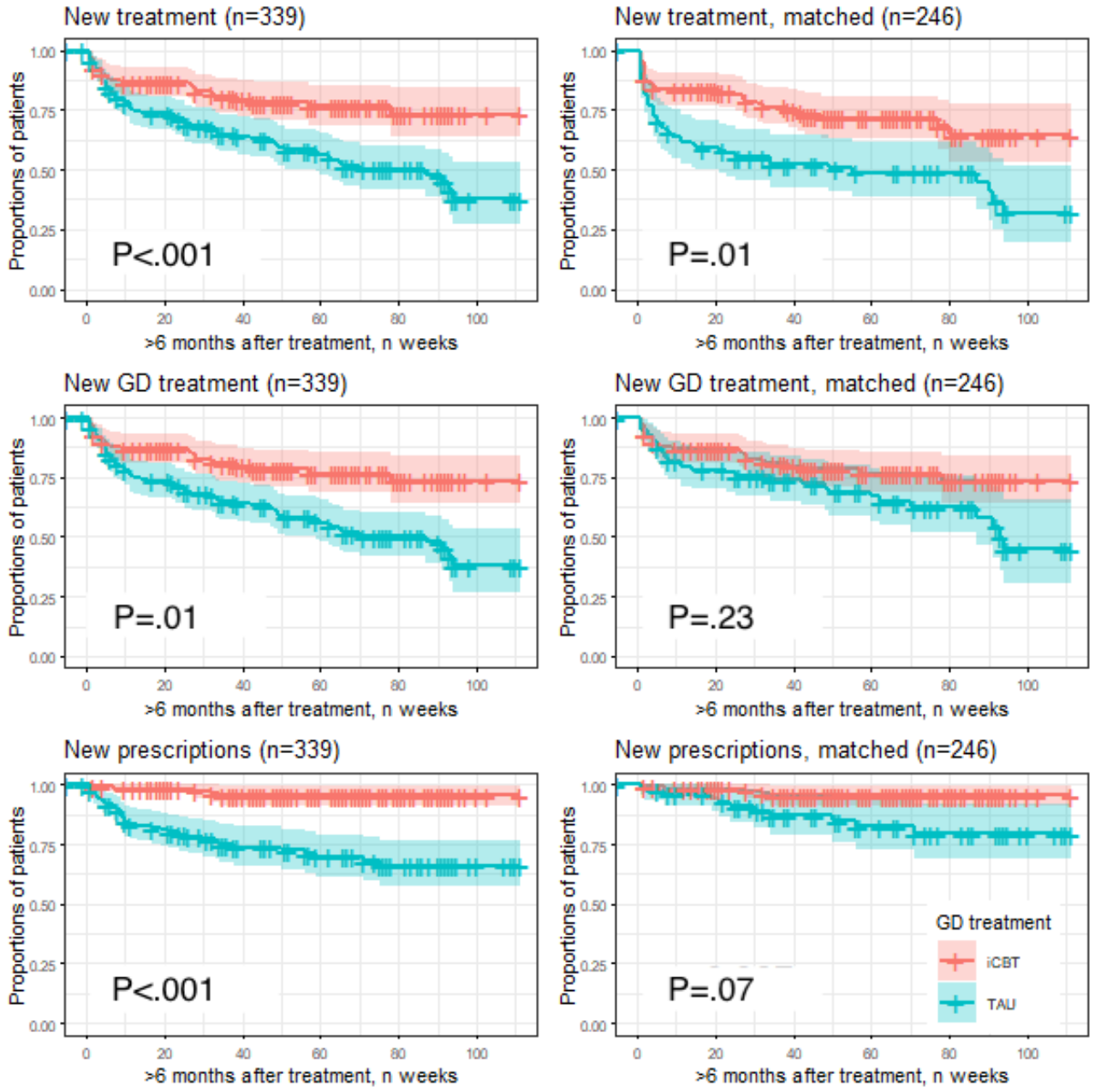

XSL•FO
RenderX

Table 3. Registry outcomes 6 months after treatment start.

Outcomes	iCBT[a] (n=123), n	TAU[b] (n=216), n	Matched TAU (n=123), n	Chi-square[c] (*df*)	*P* value
Psychiatric inpatient enrollment					
Any diagnosis	0	14	5	__[d]	—
iCBT vs TAU	—	—	—	6.76 (2)	.01
iCBT vs matched TAU[e]	—	—	—	3.27 (2)	.07
Gambling Disorder	0	0	0	—	—
iCBT vs TAU	—	—	—	—	—
iCBT vs matched TAU	—	—	—	—	—
Involvement of social services	1	13	2	—	—
iCBT vs TAU	—	—	—	4.13 (2)	.04
iCBT vs matched TAU	—	—	—	0 (2)	≥.99

[a]iCBT: internet-delivered cognitive behavioral treatment at the Stockholm Addiction eClinic.

[b]TAU = treatment-as-usual, that is, systematic psychological treatment delivered face-to-face at any of the outpatient clinics at the Stockholm Center for Dependency Disorders.

[c]Pearson chi-square test with Yates continuity correction.

[d]Not available.

[e]Matched TAU: treatment-as-usual, sample matched on comorbidity.

Discussion

Principal Findings

This paper describes the introduction of a newly developed iCBT program for GD provided within routine addiction care.

Patients who began the program were found to significantly decrease GD symptoms during the treatment window, with large within-group effect sizes in both the feasibility and registry samples. These effects were roughly equivalent to those reported in trials including participants with problem gambling from the general population [15] and iCBT with weekly telephone support [41]. Patient engagement in the iCBT modules was roughly comparable to other iCBT gambling trials in the general population, for example, studies by Dowling et al [42] and Magnusson et al [43]. The iCBT program was rated high in terms of treatment satisfaction; however, it should be noted that this score may not be fully representative. Regarding registry-based outcomes, iCBT patients showed overall similar outcomes once psychiatric severity was controlled for. Since our cohort sample lacked randomized allocation to the 2 arms, the findings of this study should not be interpreted to suggest causality. Although the final TAU sample only included patients who had undergone a systematic form of psychotherapy, it should be emphasized that the psychotherapeutic content differed between conditions, and that some face-to-face TAU treatments (eg, psychodynamic therapy) are not recommended for GD in national guidelines [12]. These study design limitations notwithstanding, the totality of our findings does show that the iCBT program did not merely serve to funnel patients into regular care, but appears to be effective in reducing GD symptoms in its own right. Of importance, while measures of gambling activity and GD symptoms were only collected during treatment (entailing missing data among those who never began treatment), the registry outcomes included all Stockholm-residing patients that enrolled in the iCBT, meaning that the included survival analyses were in full compliance with the intention-to-treat principle and thus provide an unbiased estimate of effectiveness. Ratings of treatment satisfaction were congruently high, although this measurement bears the same limitation of representativity as above. Furthermore, using a tailored instrument like the Internet Evaluation and Utility Questionnaire [44] would likely have been a more suitable choice than the general CSQ, as it includes evaluation items specific to internet-delivered treatment format.

Some patients reported negative treatment effects such as unpleasant memories, stress, anxiety, or negative emotions. This was an expected result, since the iCBT program actively encouraged patients with GD to confront gambling-related situations and manage them differently; similar negative effects have been reported for CBT in the context of anxiety [36]. It should be noted that negative effects were only reported by a small group of patients who completed large parts of the program, and the overall low occurrence of negative effects may thus not generalize to the entire treatment population. Nonetheless, no patient reported that the iCBT failed to produce positive results, or that gambling-related deterioration occurred. Furthermore, no GD patient reported any negative effects related to the iCBT exercises. This finding is noteworthy, as the behavioral exercises emphasize increased control in gambling-related situations, in some cases even involving gambling for small amounts of money under controlled circumstances (including asynchronous therapist support). In addition, 2 patients reported suicidal thoughts, although these were not attributed to the iCBT program itself. A high prevalence of suicidal thoughts and behaviors among patients with GD has been reported in several previous studies, for example, studies by Black et al [3] and Håkansson and Karlsson

XSL•FO
RenderX

[45]. This study contributes to this knowledge base, suggesting the importance of continuous clinical monitoring and assessment of suicidal thoughts and behaviors among treatment-seeking patients with GD. To reduce future attrition of the reporting of negative effects, assessments of negative effects could be administered earlier in treatment.

From a clinical perspective, the field of GD research is still partly in its infancy, and a prioritized challenge related to treatment involves the inclusion of disorder-specific interventions grounded in basic behavioral research. In this study, we have reported the introduction and evaluation of a newly developed iCBT program, which emphasizes 4 key psychological processes relevant to GD, that are access to money, reward expectation, chasing behaviors, and the gambling zone. The iCBT behavioral exercises targeting these GD processes appear to have been well-received, as indicated by a survey distributed toward the end of the treatment. However, it is important to note that this finding may not generalize to the entire sample due to the high attrition rate, and it is reasonable to hypothesize that patients who appreciated these exercises were more likely to complete the program in its entirety. As with mapping of negative effects, future iterations of the program should aim to include this survey in an earlier module to reduce missing data. Exploring how early experiences with the program influence subsequent treatment, engagement should be considered an important topic for future research, as well as the development of high-accuracy prediction models to identify who is likely to benefit from treatment. Although beyond the scope of this study, it is worth emphasizing that the iCBT program was designed to build on a simple, delimited set of theoretically derived interventions. This design enables future research on specific mediators and treatment mechanisms, a recommended feature for clinical GD trials [46].

A final GD treatment-related challenge concerns psychiatric comorbidity. Throughout the clinical procedure used in this study, caution was taken not to exclude patients based on comorbidity. However, we noted that the prevalence of psychiatric comorbidity seemed lower in the iCBT sample, compared with the TAU sample, as well as in comparison to prevalence estimates among treatment-seeking patients with GD reported elsewhere (Dowling et al [26] shows a meta-analysis). This is likely due to a study design detail, with comorbidity diagnoses being extracted from linked care events (to ensure up-to-date and still applicable diagnoses) at an Addiction eClinic that does not typically offer treatment for comorbidity. However, since moderating effects were indeed observed, even with a smaller subgroup sample, there is no reason to believe that this post hoc analysis was underpowered. Findings from this study indicate that patients with GD with psychiatric comorbidities can benefit from additional targeted treatment interventions, although it should also be noted that face-to-face treatments for GD in general have been found effective, even when additional disorders were present [47]. Whether the potential impact of comorbidity is inherently linked to the internet format, which places greater demands on the patients' abilities to independently plan and engage with the treatment material, remains unknown. It should also be stressed that the subgroup with comorbidity that was identified, likely represents those with more severe comorbidity; to what degree moderation can be observed also at lower severity levels, also remains unknown. In sum, given the high rates of comorbidity among patients with GD, elucidating the moderating and mediating effects of this on GD outcomes should be considered a research priority.

The iCBT program was introduced and made nationally available to treatment-seeking patients, through the TAS platform which is integrated into the national health care guide at 1177.se. Treatment dissemination is a crucial challenge, and we conclude that the TAS platform is a viable technical infrastructure to distribute internet-delivered psychological treatments among patients in the health care system. Importantly, the fact that measurements were collected from the same platform used for treatment, entails that there is little reason for patients who drop out of treatment to continue reporting outcomes, making it difficult to conduct true intention-to-treat analyses using this data source. Another issue that became obvious, unrelated to the TAS platform but with obvious implications for design of future studies, was the low rates of any gambling activity, and the concurrent high rates of used stimuli control techniques. Given that the national gambling self-exclusion registry Spelpaus.se, at time of writing, has over 100,000 individuals signed up and is explicitly referred to as a "First Aid" for GD, this is not unexpected. It also illustrates the necessity of informed study design considerations when attempting to estimate effectiveness of GD treatments.

Strengths and Limitations

Strengths of this study include a comprehensive cohort design with side-by-side reporting of findings from 2 different samples, the use of a matched TAU comparison sample, and registry-based outcomes covering a follow-up period of up to 2 years, with no (theoretical) missing data during this period. Since the study was conducted within the very setting and under the same conditions that continue to apply, we are convinced that our findings have high generalizability. The study also has several limitations, both in terms of design and individual measures. Regarding outcomes collected during the iCBT treatment window, these were only available (including missing data estimation) from those who actually began treatment and completed at least 1 measure (188/218); generalizability is thus limited to patients who initiated treatment, not all patients allocated to treatment. Regarding registry outcomes, the ethical approval did not cover registry linkage with the Swedish address registry, entailing that changes in region of residence during the follow-up period could not be taken into account, constituting a possible source of bias under the assumption that the 2 arms are differentially likely to change region of residence. Only having Stockholm-residing iCBT patients be matched to TAU patients should mitigate this issue, at least in part. However, patients in Sweden are free to seek treatment in other regions, including the iCBT program evaluated here. It should also be noted that while registry data does not bear missing data in the typical sense, since all health care contact by law needs to be recorded, the registry data examined in this study does not provide a complete view since it does not cover visits to nonregional health care interventions, such as contact with social services or gambling support lines, or participation in

XSL•FO
RenderX

peer-support groups. Another important limitation is that the registry follow-up data did not include measures of actual gambling, only treatment-seeking (including for gambling). Such data could have been collected by, for instance, sending out invitations to patients to respond to a survey; alas, resources for this were not available. In theory, it would also be possible to extend the registry follow-up to also include enrollment in the Spelpaus.se self-exclusion register; such linkage should be attempted in future research. A final limitation is that propensity score matching could only be performed on registry variables, which did not include potentially influential factors such as education. Furthermore, the treatment provided to the TAU cohort was not necessarily matched to the iCBT cohort in terms of treatment structure or content. As in any nonrandomized observational study, causal interpretations of difference should hence be done with great caution, yet we argue that findings of outcomes in the TAU cohort nonetheless constitute a meaningful clinical benchmark with which to compare newer treatments.

Conclusions

A newly developed iCBT program for GD, introduced into routine addiction care, was feasible, well-received by patients, and was associated with the expected decrease in symptoms during the treatment window. The study highlights the need for GD treatment development in tandem with ongoing research and monitoring of treatment effects, especially in the context of psychiatric comorbidity among patients with GD. The findings also underscore the importance of careful study design and technical considerations when introducing internet-based treatments into the routine care setting. Future randomized controlled trials and studies of iCBT treatment mechanisms, and the moderating effects of psychiatry comorbidity, are warranted.

Acknowledgments

This study was made possible by funding from the Independent Research Council of Svenska Spel (PL and MJ), the Centre for Psychiatry Research (PL and OM), and the Swedish Research Council for Health, Working Life and Welfare (FORTE) program grant “Responding to and Reducing Gambling Problems—Studies in Help-seeking, Measurement, Comorbidity and Policy Impacts” (REGAPS; FORTE grant 2016-07091; OM and AHB).

Data Availability

The data sets generated and analyzed during this study are not publicly available due to patient confidentiality and cannot be shared without permission from the health care provider, but are available from the corresponding author on reasonable request.

Authors' Contributions

PL, OM, and AHB conceived the study. OM developed the iCBT in collaboration with JR, JB, AHB, and PL. OM and PL introduced the iCBT in collaboration with MJ, MG, and HH, who also assessed and treated patients in the study. OM and PL performed statistical analyses and wrote the first paper draft. All authors contributed to the final version of the paper.

Conflicts of Interest

Authors PL and PC report past and ongoing industry-academia collaborations with several licensed gambling providers on independent scientific evaluations of responsible gambling tools, including project-specific financing, but have no personal financial or other ties to the gambling industry. Authors OM, PC, and PL report having received funding from the Independent Research Council of Svenska Spel; author AHB sits on the Council. JB reports no conflicts of interest.

Multimedia Appendix 1
Feasibility Study.
[DOCX File , 37 KB - jmir_v26i1e54754_app1.docx]

References

1. Dąbrowska K, Moskalewicz J, Wieczorek Ł. Barriers in access to the treatment for people with gambling disorders. Are they different from those experienced by people with alcohol and/or drug dependence? J Gambl Stud 2017;33(2):487-503 [FREE Full text] [doi: 10.1007/s10899-016-9655-1] [Medline: 27832520]
2. Abbott M, Romild U, Volberg R. The prevalence, incidence, and gender and age-specific incidence of problem gambling: results of the Swedish longitudinal gambling study (Swelogs). Addiction 2018;113(4):699-707. [doi: 10.1111/add.14083] [Medline: 29105942]
3. Black DW, Coryell W, Crowe R, McCormick B, Shaw M, Allen J. Suicide ideations, suicide attempts, and completed suicide in persons with pathological gambling and their first-degree relatives. Suicide Life Threat Behav 2015;45(6):700-709 [FREE Full text] [doi: 10.1111/sltb.12162] [Medline: 25845522]
4. Newman SC, Thompson AH. A population-based study of the association between pathological gambling and attempted suicide. Suicide Life Threat Behav 2003;33(1):80-87. [doi: 10.1521/suli.33.1.80.22785] [Medline: 12710543]

XSL•FO
RenderX

5. Rychtarik RG, McGillicuddy NB. Preliminary evaluation of a coping skills training program for those with a pathological-gambling partner. J Gambl Stud 2006;22(2):165-178. [doi: 10.1007/s10899-006-9008-6] [Medline: 16841184]
6. Diagnostic and Statistical Manual of Mental Disorders. Washington, DC: American Psychiatric Association; 2013.
7. Dahlberg M, Anderberg M. Detecting problem gambling - a pilot project at a social service office. Socialmedicinsk Tidskr 2015;92(4):505-518.
8. Nehlin C, Nyberg F, Jess K. Brief intervention within primary care for at-risk gambling: a pilot study. J Gambl Stud 2016;32(4):1327-1335. [doi: 10.1007/s10899-016-9610-1] [Medline: 27038815]
9. Håkansson A, Widinghoff C. Gambling despite nationwide self-exclusion-a survey in online gamblers in Sweden. Front Psychiatry 2020;11:599967 [FREE Full text] [doi: 10.3389/fpsyt.2020.599967] [Medline: 33343428]
10. Lindner P, Forsström D, Jonsson J, Berman A, Carlbring P. Transitioning between online gambling modalities and decrease in total gambling activity, but no indication of increase in problematic online gambling intensity during the first phase of the COVID-19 outbreak in Sweden: a time series forecast study. Front Public Health 2020;8:554542 [FREE Full text] [doi: 10.3389/fpubh.2020.554542] [Medline: 33117770]
11. The Swedish gambling authority. Gaming Habits Online. 2023. URL: https://www.government.se/government-agencies/gaming-board-for-sweden-lotteriinspektionen/ [accessed 2024-08-13]
12. The Swedish National Board of Health and Welfare. Treatment for gambling disorder. National Guidelines for Health Care and Social Services. 2017. URL: https://www.socialstyrelsen.se/publikationer2017/2017-12-20 [accessed 2024-08-13]
13. Eriksen JW, Fiskaali A, Zachariae R, Wellnitz KB, Oernboel E, Stenbro AW, et al. Psychological intervention for gambling disorder: a systematic review and meta-analysis. J Behav Addict 2023;12(3):613-630 [FREE Full text] [doi: 10.1556/2006.2023.00034] [Medline: 37450372]
14. Cowlishaw S, Merkouris S, Chapman A, Radermacher H. Pathological and problem gambling in substance use treatment: a systematic review and meta-analysis. J Subst Abuse Treat 2014 Feb;46(2):98-105. [doi: 10.1016/j.jsat.2013.08.019] [Medline: 24074847]
15. Pallesen S, Mitsem M, Kvale G, Johnsen BH, Molde H. Outcome of psychological treatments of pathological gambling: a review and meta-analysis. Addiction 2005;100(10):1412-1422. [doi: 10.1111/j.1360-0443.2005.01204.x] [Medline: 16185203]
16. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023;22(2):305-314 [FREE Full text] [doi: 10.1002/wps.21088] [Medline: 37159350]
17. van der Maas M, Shi J, Elton-Marshall T, Hodgins DC, Sanchez S, Lobo DS, et al. Internet-based interventions for problem gambling: scoping review. JMIR Ment Health 2019;6(1):e65 [FREE Full text] [doi: 10.2196/mental.9419] [Medline: 30617046]
18. Andersson G, Carlbring P. Internet Interventions in Clinical Psychology. 2022. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1654543 [accessed 2023-11-20]
19. Clark DM. Developing new treatments: on the interplay between theories, experimental science and clinical innovation. Behav Res Ther 2004;42(9):1089-1104. [doi: 10.1016/j.brat.2004.05.002] [Medline: 15325903]
20. Titov N, Dear B, Nielssen O, Staples L, Hadjistavropoulos H, Nugent M, et al. ICBT in routine care: a descriptive analysis of successful clinics in five countries. Internet Interv 2018;13:108-115 [FREE Full text] [doi: 10.1016/j.invent.2018.07.006] [Medline: 30206525]
21. Molander O, Ramnerö J, Bjureberg J, Berman AH. What to target in cognitive behavioral treatment for gambling disorder-a qualitative study of clinically relevant behaviors. BMC Psychiatry 2022;22(1):510 [FREE Full text] [doi: 10.1186/s12888-022-04152-2] [Medline: 35902829]
22. Gooding P, Tarrier N. A systematic review and meta-analysis of cognitive-behavioural interventions to reduce problem gambling: hedging our bets? Behav Res Ther 2009;47(7):592-607. [doi: 10.1016/j.brat.2009.04.002] [Medline: 19446287]
23. Dixon M. Why behavior analysts should study gambling behavior. Anal Gambl Behav 2007;1(1):1 [FREE Full text]
24. Ramnerö J, Molander O, Lindner P, Carlbring P. What can be learned about gambling from a learning perspective? A narrative review. Nordic Psychology 2019;71(4):303-322. [doi: 10.1080/19012276.2019.1616320]
25. Håkansson A, Mårdhed E, Zaar M. Who seeks treatment when medicine opens the door to pathological gambling patients-psychiatric comorbidity and heavy predominance of online gambling. Front Psychiatry 2017;8:255 [FREE Full text] [doi: 10.3389/fpsyt.2017.00255] [Medline: 29238309]
26. Dowling NA, Cowlishaw S, Jackson AC, Merkouris SS, Francis KL, Christensen DR. Prevalence of psychiatric co-morbidity in treatment-seeking problem gamblers: a systematic review and meta-analysis. Aust N Z J Psychiatry 2015;49(6):519-539 [FREE Full text] [doi: 10.1177/0004867415575774] [Medline: 25735959]
27. Johansson M, Sinadinovic K, Gajecki M, Lindner P, Berman AH, Hermansson U, et al. Internet-based therapy versus face-to-face therapy for alcohol use disorder, a randomized controlled non-inferiority trial. Addiction 2021;116(5):1088-1100 [FREE Full text] [doi: 10.1111/add.15270] [Medline: 32969541]

XSL•FO
RenderX

28. Sinadinovic K, Johansson M, Johansson A, Lundqvist T, Lindner P, Hermansson U. Guided web-based treatment program for reducing cannabis use: a randomized controlled trial. Addict Sci Clin Pract 2020;15(1):9 [FREE Full text] [doi: 10.1186/s13722-020-00185-8] [Medline: 32070417]
29. Månsson V, Molander O, Carlbring P, Rosendahl I, Berman AH. Emotion regulation-enhanced group treatment for gambling disorder: a non-randomized pilot trial. BMC Psychiatry 2022 Jan 06;22(1):16 [FREE Full text] [doi: 10.1186/s12888-021-03630-3] [Medline: 34991511]
30. Molander O. Developing novel measures and treatments for gambling disorder [Doctoral Dissertation, Karolinska Institutet]. Dept of Clinical Neuroscience. 2022. URL: https://www.psykiatriforskning.se/globalassets/verksamheter/forskning-och-utveckling/centrum-for-psykiatriforskning/forskning/spelberoende/thesis_olof_molander.pdf [accessed 2022-04-10]
31. Schüll ND. Addiction by Design: Machine Gambling in Las Vegas. Princeton, New Jersey: Princeton University Press; 2012.
32. Molander O, Lindner P, Ramnerö J, Bjureberg J, Carlbring P, Berman AH. Internet-based cognitive behavior therapy for problem gambling in routine care: protocol for a non-randomized pilot and feasibility trial. Pilot Feasibility Stud 2020;6(1):106 [FREE Full text] [doi: 10.1186/s40814-020-00647-5] [Medline: 32699645]
33. Ferris J, Wynne H. The Canadian Problem Gambling Index: Final Report. Ottawa: Canadian Centre on Substance Abuse; 2001.
34. Kim SW, Grant JE, Potenza MN, Blanco C, Hollander E. The gambling symptom assessment scale (G-SAS): a reliability and validity study. Psychiatry Res 2009;166(1):76-84 [FREE Full text] [doi: 10.1016/j.psychres.2007.11.008] [Medline: 19200607]
35. Weinstock J, Whelan JP, Meyers AW. Behavioral assessment of gambling: an application of the timeline followback method. Psychol Assess 2004;16(1):72-80. [doi: 10.1037/1040-3590.16.1.72] [Medline: 15023094]
36. Rozental A, Kottorp A, Forsström D, Månsson K, Boettcher J, Andersson G, et al. The Negative Effects Questionnaire: psychometric properties of an instrument for assessing negative effects in psychological treatments. Behav Cogn Psychother 2019 Sep;47(5):559-572 [FREE Full text] [doi: 10.1017/S1352465819000018] [Medline: 30871650]
37. Attkisson CC, Greenfield TK. In: Maruish ME, editor. Client Satisfaction Questionnaire-8 and Service Satisfaction Scale-30. The Use of Psychological Testing for Treatment Planning and Outcome Assessment. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc; 1994:402-420.
38. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Softw 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
39. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer 2003;89(2):232-238 [FREE Full text] [doi: 10.1038/sj.bjc.6601118] [Medline: 12865907]
40. Stuart E, King G, Imai K, Ho D. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011:1-28 [FREE Full text] [doi: 10.32614/cran.package.matchit]
41. Carlbring P, Degerman N, Jonsson J, Andersson G. Internet-based treatment of pathological gambling with a three-year follow-up. Cogn Behav Ther 2012;41(4):321-334. [doi: 10.1080/16506073.2012.689323]
42. Dowling NA, Merkouris SS, Rodda SN, Smith D, Aarsman S, Lavis T, et al. GamblingLess: a randomised trial comparing guided and unguided internet-based gambling interventions. JCM 2021;10(11):2224. [doi: 10.3390/jcm10112224]
43. Magnusson K, Nilsson A, Andersson G, Hellner C, Carlbring P. Internet-delivered cognitive-behavioral therapy for significant others of treatment-refusing problem gamblers: a randomized wait-list controlled trial. J Consult Clin Psychol 2019;87(9):802-814. [doi: 10.1037/ccp0000425] [Medline: 31403816]
44. Ritterband LM, Ardalan K, Thorndike FP, Magee JC, Saylor DK, Cox DJ, et al. Real world use of an Internet intervention for pediatric encopresis. J Med Internet Res 2008;10(2):e16 [FREE Full text] [doi: 10.2196/jmir.1081] [Medline: 18653440]
45. Håkansson A, Karlsson A. Suicide attempt in patients with gambling disorder-associations with comorbidity including substance use disorders. Front Psychiatry 2020;11:593533 [FREE Full text] [doi: 10.3389/fpsyt.2020.593533] [Medline: 33304287]
46. Walker M, Toneatto T, Potenza MN, Petry N, Ladouceur R, Hodgins DC, et al. A framework for reporting outcomes in problem gambling treatment research: the banff, alberta Consensus. Addiction 2006;101(4):504-511. [doi: 10.1111/j.1360-0443.2005.01341.x] [Medline: 16548930]
47. Dowling NA, Merkouris SS, Lorains FK. Interventions for comorbid problem gambling and psychiatric disorders: advancing a developing field of research. Addict Behav 2016;58:21-30. [doi: 10.1016/j.addbeh.2016.02.012] [Medline: 26900888]

Abbreviations

CBT: cognitive behavioral therapy
CSQ: Client Satisfaction Questionnaire
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, fifth edition
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, fourth edition
GD: gambling disorder

XSL•FO
RenderX

GSAS: Gambling Symptom Assessment Scale
iCBT: internet-delivered cognitive behavioral therapy
TAS: Treatment and Support
TAU: treatment-as-usual

Edited by T McCall; submitted 20.11.23; peer-reviewed by S Rodda, N Titov; comments to author 26.01.24; revised version received 21.03.24; accepted 24.05.24; published 18.09.24.

Please cite as:
Molander O, Berman AH, Jakobson M, Gajecki M, Hällström H, Ramnerö J, Bjureberg J, Carlbring P, Lindner P
Introducing and Evaluating the Effectiveness of Online Cognitive Behavior Therapy for Gambling Disorder in Routine Addiction Care: Comparative Cohort Study
J Med Internet Res 2024;26:e54754
URL: https://www.jmir.org/2024/1/e54754
doi:10.2196/54754
PMID:

©Olof Molander, Anne H Berman, Miriam Jakobson, Mikael Gajecki, Hanna Hällström, Jonas Ramnerö, Johan Bjureberg, Per Carlbring, Philip Lindner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Optimizing Telehealth for Diabetes Management in the Deep South of the United States: Qualitative Study of Barriers and Facilitators on the Patient and Clinician Journey

Alessandra N Bazzano[1], MPH, PhD; Tejal Patel[1], MS; Elizabeth Nauman[2], PhD; Dana Cernigliaro[3], PhD; Lizheng Shi[4], PhD

[1]Department of Social, Behavioral, and Population Sciences, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, United States

[2]Louisiana Public Health Institute, New Orleans, LA, United States

[3]Public Health Innovation and Action, New York, NY, United States

[4]Department of Health Policy and Management, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, United States

Corresponding Author:
Alessandra N Bazzano, MPH, PhD
Department of Social, Behavioral, and Population Sciences
Tulane University School of Public Health and Tropical Medicine
1440 Canal St
New Orleans, LA, 70112
United States
Phone: 1 5049882338
Email: abazzano@tulane.edu

Abstract

Background: The Deep South of the United States, and Louisiana in particular, bears a greater burden of obesity, diabetes, and heart disease compared with other regions in the United States. Throughout the COVID-19 pandemic, there has been a substantial increase in telehealth visits for diabetes management to protect the safety of patients. Although there have been significant advancements in telehealth and chronic disease management, little is known about patient and provider perspectives on the challenges and benefits of telehealth visits among people living with diabetes and providers who care for patients with diabetes in Louisiana.

Objective: This study aimed to explore barriers, facilitators, challenges, and benefits to telehealth for patients with diabetes and health care providers as they transitioned from in-person to remote care during the early COVID-19 pandemic to understand potential optimization.

Methods: A total of 24 semistructured qualitative interviews were conducted with 18 patients living with diabetes and 6 clinicians who served patients with diabetes to explore their experiences and perceptions of telehealth services for diabetes care. Approximately half of the participants identified as Black or African American, half as White, and 75% as female. Interviews were recorded, transcribed, and coded by experienced qualitative researchers using inductive and deductive techniques. A narrative, descriptive approach to the patient and clinician journey framed the study, including the development of internal journey maps, and reflexive thematic analysis was applied to the transcripts, with special attention to barriers and facilitators.

Results: In total, 5 themes illustrated barriers and facilitators for participants: convenience, safety, and comfort are the benefits of telehealth for patients and clinicians; yet telehealth and in-person visits are valued differently; the convenience of telehealth may have a downside; technology acts as a double-edged sword; and managing expectations and efficiency of the visit experience was an important factor. Individual experiences varied in relation to several factors, including comfort level and access to technology, health system protocols for providing telemedicine, and level of diabetes control among patients.

Conclusions: Recommendations for optimization include providing support to help guide and inform patients about what to expect and how to prepare for telehealth visits as well as allowing clinicians to schedule telehealth and in-person visits during discrete blocks of time to improve efficiency. Further research should address how hybrid models of telehealth and in-person care may differentially impact health outcomes for patients with diabetes, particularly for people with multiple chronic conditions in settings where access to technology and connectivity is not optimal.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e43583) doi:10.2196/43583

KEYWORDS

telemedicine; diabetes mellitus; chronic disease; patient participation; delivery of health care; COVID-19; mobile phone

Introduction

Background

In the United States, telehealth services have significantly advanced during the last 20 years, including in the Deep South, where high rates of cardiometabolic disorders, such as obesity [1], diabetes [2], and heart disease [3], are prevalent. The growth of new technologies and the high rates of internet and electronic device use in the United States have enabled telehealth to flourish—between 2010 and 2017, the percentage of US hospitals using technology to connect with patients increased from 35% to 76% [4]. With the COVID-19 pandemic, the role of telehealth has become even more important, providing care to those who need it without in-person visits that could put patients at increased risk of exposure. During the initial peak of the COVID-19 pandemic, telehealth visits between 2019 and 2020 increased from <1% to as high as 80% in places with high COVID-19 prevalence [5]. Although the increase in telehealth visits was substantial, disparities among patient subgroups remained prevalent, with video telehealth rates lowest among populations without a high school diploma; among adults aged ≥65 years; and among Asian, Black or African American, and Latine individuals [5].

This significant increase and expansion of telehealth services has been especially important for patients with chronic diseases, such as diabetes, which requires consistent management by health care providers to prevent serious complications and mortality [6]. Diabetes affects an estimated 34.2 million people or 10.5% of the US population. For adults, there are significant disparities, with the highest prevalence of diabetes among minoritized populations, those without high school education, and populations that live below the federal poverty level [7]. Diabetes is one of the leading causes of death in Louisiana, with approximately 14.2% of the population being diagnosed with diabetes [7]. People living with diabetes, particularly those who also have other health conditions, are at an increased risk for serious complications due to COVID-19 and therefore may benefit from remote care [8]. Telehealth visits likely reduce the risk of exposure to COVID-19 infection, as patients do not have to interact with others in waiting rooms, examination rooms, and health care facilities, while also avoiding the potential risk of exposure during transportation.

The potential benefits of telehealth services can be seen most acutely in states that bear the greatest burden of COVID-19 prevalence. Early in the pandemic, Louisiana experienced the fastest growth rate of COVID-19 infections in the world, with 2 parishes comprising the New Orleans metro area exhibiting one of the highest per capita death rates among metropolitan cities in the United States [9]. There were significant racial disparities in death rates in Louisiana at a magnitude greater than other epicenters of the pandemic in the United States [10], particularly among the Black or African American population, where the rate of death due to COVID-19 was significantly higher than that of other races [10]. In March 2020, the Louisiana Department of Health directed health care providers to postpone care for 30 days and encouraged providers to use telehealth services. At the same time, the Louisiana Department of Health eased Medicaid billing restrictions on the provision of telehealth visits, expanding the reach for patients who need care [11,12]. This paved the way for the expansion of and increase in telehealth services in Louisiana.

With the significant increase and continuation of telehealth visits during the COVID-19 pandemic and evidence that telehealth programs can be an effective way of managing diabetes compared with face-to-face care [13], a better understanding of both patient and provider perspectives is required to identify opportunities to understand and improve the role of telehealth in patient care for this condition. A recent qualitative study of technology use for diabetes care assessed patient perspectives, indicating a good potential for benefit, with some hesitation on the part of patients about the use of technology [14]. With the rapid transition to telehealth early in the COVID-19 pandemic, remote services have been shown to be well received and result in positive encounters despite technical challenges [15,16]. From the health care provider's perspective, the advantages of telehealth included increased usage of services [17], improved provider well-being [17,18], accommodation of patients who face challenges for in-person care [18,19], and fostering a sense of family-centered care [18]. The disadvantages from the clinician perspective include issues with technology and disparities in patient access to technology [17-20], less personal connection [19], inability to complete in-person diagnostic tests [20], reimbursement complications and out-of-state licensure restrictions [17], and a preference for in-person visits for patient populations that would benefit more from in-person care [17]. Studies exploring patient perspectives have reported that benefits, such as reduced travel time, shorter wait time, and cost savings [15,21], as well as convenience and safety [15,22], were seen as advantages, whereas internet issues, technical barriers [21,22], lack of connection with their provider, and unfamiliarity with the telehealth process [22] were seen as barriers. One study of patient perspectives on cardiology telehealth found that although both in-person and telehealth visits were viewed favorably, patient satisfaction was rated slightly higher for in-person visits. They found that the domain of clinical competence had the only lower mean score for telehealth [21]. A recent study highlighted racial disparities in telehealth use in the South with the transition to telehealth during the COVID-19 pandemic, finding an increase in the proportion of female and Black and Hispanic patients. However, discrepancies were observed in the likelihood of using an audio-video telehealth service, with older, Black, urban, and Medicaid or Medicare insurance carriers less likely to use audio-video telehealth services [23].

Although there have been significant advancements in telehealth and chronic disease management, they have not yet met their

XSL•FO
RenderX

potential for improving and addressing the needs of all patients equally. The demographic context of patients with chronic diseases, such as diabetes, may be misaligned with access to effective telehealth modalities, such as video visits [23]; thus, to investigate broader issues around the value of telehealth, we applied qualitative methods to an exemplar population of people living with diabetes.

Objectives

This study was nested within a larger natural experiment study [24-28] and comprised the qualitative component of a mixed method approach that aimed to explore barriers, facilitators, challenges, and benefits to telehealth for patients with diabetes and health care providers in Louisiana as they transitioned from in-person to telehealth care following the COVID-19 pandemic. Data from this study complement the larger study's quantitative examination of facilitators and barriers to the uptake, adoption, and implementation of telehealth services among Medicare patients with diabetes from the perspectives of health systems, health care providers, and patients. The larger study provided quantitative comparison of diabetes control and continuity of care between patients with and without telehealth use during COVID-19, providing a mixed method approach alongside the qualitative data presented here [28]. This study also aimed to provide insights into the broader context of telehealth value for patients with chronic diseases in lower socioeconomic contexts.

Methods

Overview

This study is a part of the larger Louisiana Experiment Assessing Diabetes Outcomes study, which aimed to assess the reach, effectiveness, adoption, implementation, and maintenance of chronic care management services. The details of the larger study can be found elsewhere [28]. This qualitative study included individual interviews with 18 patients with diabetes and 6 clinicians who provided diabetes care, who participated in semistructured interviews to explore their experiences, including barriers and facilitators to engaging in telehealth services for diabetes care. To guide the preparation of this study, the Standards for Reporting Qualitative Research Reporting Guidelines [29] were used at each stage of the process.

Sampling and Recruitment

Purposive sampling was used to identify participants who were either patients who received a telehealth visit for diabetes care after March 1, 2020, or health system staff who had conducted or coordinated such visits. Recruitment of patients occurred through patient portal messages and secure emails, with phone call follow-up, and participants were purposively selected to represent the broader population of patients with diabetes in Louisiana who depend on Medicare and a mix of insurance providers. Chain referral sampling was also used, and interviewees from the previous year of the study assisted in the identification of potential participants. In total, 60 patients were invited and 18 of them agreed to participate, and among 10 clinicians who were invited, 6 agreed to participate, fulfilling the expectation of a sample sufficient for qualitative research using the principles of code saturation in data analysis [30,31]. Health care provider participants were defined as clinicians and included primary care physicians, endocrinologists, and advanced practice nurses, allowing for a variety of practice experiences. Clinicians were invited via email or phone to participate in interviews and given information about the study. The study steering committee, including patient partners and researchers, reviewed and facilitated participant recruitment strategies.

Data Collection Methods

Interview guides were used to support the semistructured interviews and were developed in partnership with study steering committee members. To strengthen the validity, the guide was pretested with patient partners who had diabetes and experienced telehealth services during the pandemic. Interviews were conducted and recorded on receiving participant consent through videoconferencing for participants who stated that the technology was available to them and via telephone for those who preferred not to use videoconferencing. No patients had to travel for the interview, and all phone or videoconference interviews were conducted in private locations. Patient and provider participants were interviewed between October 2020 and April 2021 by researchers with advanced degrees in public health and experience in qualitative research. To acknowledge their contributions, participants received gift cards. A study description explaining the purpose of the research, goals for the interview, and participants' rights was provided before the interviews, at which only the researchers and participants were present, and informed consent was obtained.

Data Analysis

All data were securely stored on encrypted computers and accessed only by the researchers who completed the analysis. A narrative descriptive approach provided a guiding orientation to the analysis, with special attention paid to the participant journey and attendant barriers and facilitators. The study researchers have used reflexive thematic analyses [32]. A total of 2 experienced qualitative researchers collaborated in coding, and through inductive and deductive strategies, they proceeded to creation of categories and candidate themes from the interview transcripts. Open coding was used to describe candidate themes, whereas deductive coding was used to check themes against the data. Journey maps were also developed with human-centered design tools applied to interview data to explore barriers and facilitators, paint points, and high points on participants' journeys with telehealth. During the analysis process, peer debriefing and reflexive journaling through memos were conducted. NVivo qualitative analysis software (QSR International) was used to manage data. Preliminary and final themes and analysis progress were shared with the study leaders and researchers, partner organizations, and patient stakeholders to receive feedback on the findings. The researchers discussed preliminary data during the review process and consulted with patient partners for validity.

Ethics Approval

Institutional review board approval was provided for this research by Tulane University under reference number 906810.

XSL•FO
RenderX

Results

Participant Characteristics

A total of 18 patients with diabetes and 6 health care providers were interviewed for this study. Patient participants were predominantly female (14/18, 78%) and married (8/18, 44%). The average age of participants was 60 (SE 3.0; range: 35-78) years, and, on average, patients had been living with diabetes for 13 to 14 (SE 13.5; range 1-48) years at the time of study. Half of the patient participants described themselves as Black or African American (9/18, 50%) and half described themselves as White (9/18, 50%). Table 1 presents an overview of the participant demographic data. For participating health care providers, the average age was 50 (SE 5.1; range 38-75) years, and they had been practicing in their profession for an average of 20 (SE 5.8; range 7-46) years. The participants included primary care physicians, endocrinologists, and advanced practice nurses, allowing for a variety of practice experiences. A total of 4 provider participants identified themselves as female and 2 identified themselves as male. In addition, 2 providers identified as Black or African American, 1 identified as South Asian, 2 identified as White, whereas 1 provider participant did not disclose this information.

Most patient participants had Medicaid or Medicare health insurance coverage. Almost all participants lived within a 45-minute drive of their health care provider or hospital, except 1 rural residing participant who lived 4 hours away from their provider. The participants described managing several other health conditions in addition to diabetes, including asthma, overweight or obesity, heart conditions, thyroid disease, breast cancer, vision issues, sleep apnea, kidney disease, genetic disorders, and most commonly high blood pressure. Most participants identified as independent and self-sufficient, although many had friends and family members who checked in with them and ensured that they were managing their health and helped with transportation, technology, or advice on diet and lifestyle.

In terms of previous experience with telehealth, patient participants had an average of 2 to 3 previous telehealth visits (range 1-5 visits). Some participants reported using the phone for telephonic visits (8/18, 44%), whereas others reported using the telephone and video (10/18, 56%) for telehealth visits. Of 18 patients, 14 (78%) had returned to mainly in-person care at the time of the study while continuing to have some telehealth visits. Table 2 illustrates the participants' telehealth usage patterns.

Table 1. Participant characteristics.

Characteristics[a]	Patient participants (n=18)	Provider participants (n=6)
Age (y), mean (SE; range)	60.39 (3.03; 35-78)	50.16 (5.08; 38-72)
Years living with diabetes (patients only), mean (SE; range)	13.53 (3.56; 1-48)	N/A[b]
Years in clinical practice (providers only), mean (SE; range)	N/A	20.08 (5.84; 7-46)
Sex, n (%)		
Female	14 (78)	4 (67)
Male	4 (22)	2 (33)
Race, n (%)		
Black or African American	9 (50)	2 (33)
South Asian	0 (0)	1 (17)
White	9 (50)	2 (33)
Marital status, n (%)		
Married	8 (44)	—[c]
Divorced	4 (22)	—
Single	4 (22)	—
Widowed	2 (11)	—

[a]Characteristics, including race, refer to categorizations used by the United States Census Bureau and other entities; these are not an indicator of biological difference but are presented to provide context about socially constructed experiences.

[b]N/A: not applicable.

[c]Not available.

XSL•FO
RenderX

Table 2. Patient telehealth use characteristics (n=18).

Visit details and characteristics	Values
Telehealth visits, mean (range)	2.35 (1-5)
Mode of visit, n (%)	
Telephone only	8 (44)
Telephone and video	10 (56)
Returned to in-patient care, n (%)	
No	4 (22)
Yes, with some telehealth	14 (78)

Overview of the Telehealth Journey for Patients and Clinicians

Both the patient and provider participants were asked to describe their user journey during the telehealth process. For patients, the activities for scheduling and accessing telehealth visits were generally similar across the different health systems they used. Participants typically described receiving a call or notification at least 1 day before their scheduled visit and receipt of a link to log into the visit. If the patient had trouble navigating the log-in process, some participants described that a member of the health care staff would reach out to help guide them. On average, telehealth visits lasted between 10 and 20 minutes. However, the reported duration of the full range of visits was between 1 minute for a patient who did not have results to discuss to a full hour for a participant who had a more complex health issue to discuss. Patients used phones and laptops to connect to visits, and video or phone visits were often conducted at home (locations within the home included kitchens, offices, living rooms, bedrooms, and bathrooms), but some took place in a car or outside the home. Where patients took telehealth visits at home, they also reported that family members might be at home or in the same room as the patient during the telehealth visit.

From the health care provider perspective, the telehealth process varied based on the standard operating procedures of the individual health system. Health systems had different scheduling approaches for telehealth; some allocated full days or half days for scheduling strictly telehealth visits, with in-person appointments scheduled during a different block of time; others had both in-person and telehealth visits interspersed throughout the day on their schedule. The latter format was considered the least efficient by clinicians due to longer wait times when switching between in-person and telehealth visits, as well as difficulties making connections with patients. Health care providers reported variable visit durations for telehealth visits between 10 and 20 minutes on average, and these were taken from different locations. Certain facilities had physician workrooms, conference rooms, or offices that the provider could use. A couple of providers reported taking calls from home; however, challenges with this approach were mentioned, including distractions. Clinicians described using smartphones, laptops, and tablets for telehealth and also mentioned apps, such as Doximity, which conceals the phone number of the caller, as being used in telehealth care provision [27].

Facilitators and Barriers to Telehealth Care for Diabetes

Overview

The 5 themes drawn from the qualitative analysis are presented in subsequent sections. Overall, participants described positive experiences with telehealth during the pandemic, although perspectives ranged from a strong preference for telehealth to a strong preference for in-person care. The participants described valuing the convenience, safety, and comfort of telehealth visits. However, patients seemed to value telehealth visits differently than in-person visits, indicating the importance of in-person interactions with their provider and physical check-ups. Furthermore, health care providers stated that patients had different expectations for a telehealth visit and that sometimes the flexibility and convenience of the telehealth visit translated to patients not being prepared for visits, having unrealistic expectations about wait times in telehealth, and taking calls from locations that were inappropriate for health care (eg, while walking through a retail store).

The role of technology as both a barrier and a facilitator was a key theme for both patients and clinicians. Clinicians who experienced it as most beneficial were typically more comfortable navigating technology, had more experience with telehealth, or were part of a health system that provided support around technology. Patients with diabetes who were more comfortable with technology used with telehealth (eg, those who could easily navigate devices, log into portals for video telehealth visits, and receive visit summaries from the patient portal) described the most positive experience with telehealth visits.

Clinicians described valuing the convenience and safety of telehealth for patients, particularly the ability of patients to check in with them without traveling to a health care facility, the efficiency of the system for those who were tech savvy, and improved safety for patients with diabetes who were at high risk for severe COVID-19. Similar to patient participants, clinicians described challenges and delays in care because of issues with technology. For providers, telehealth was perceived to be the best option for patients with controlled diabetes. Both clinicians and patients endorsed a preference for a hybrid model consisting of both in-person and telehealth care.

XSL•FO
RenderX

Convenience, Safety, and Comfort of Telehealth for Patients and Providers

Patients described many facilitators for using telehealth, with the most common being convenience, particularly for those with mobility issues, those with busy schedules, or those who lived at greater geographical distances from their provider. Patients described valuing the convenience and comfort of not having to drive, park, check in, and wait for their appointment as well as the ease of talking to their provider over the phone:

> *Yes, I love [telehealth visits] I don't have to get dressed, I don't have to drive, I don't have to do nothing, but answer my phone...That's a big help, you don't have to get dressed to go out and you don't have to spend no money on no gas. It's just worth [it] all around. You can definitely be safe because you are not going into no crowd.* [Black or African American female patient aged 77 years, living with diabetes for 11 years]

> *If I'm too busy, I'd rather do telehealth than trying to finagle my schedule. "Oh, I got to go to the doctor. I got to do this. I got to do that." Telehealth would be much easier. I probably could sit in the car, going wherever I'm going, and do my telehealth visit and still be at the place wherever I need to be.* [Black or African American male patient aged 46 years, living with diabetes for 2 years]

> *I feel like you still get the same healthcare as if you were standing there in front your doctor, unless they have to look at something, like your wound or something ...i think it is very accessible and very easy to use.* [White female patient aged 43 years, living with diabetes for 8 years]

One participant's view was that she felt more honest and open sitting at home, speaking with her provider during a telehealth visit:

> *I think I'm more honest when I'm sitting in my chair at home. It's harder for them to judge me virtually, in my mind, at least. I think I'm just so much more comfortable here, that I'm more open.* [White female patient aged 60 years, living with diabetes for 20 years]

Participants also discussed valuing their safety. Patients who did not have a strong preference for in-person or telehealth visits mentioned that they would do whatever was the safest and recommended during the pandemic. Although an in-person visit was discussed as important if they had a physical issue they felt their physician should assess, patients acknowledged that traveling and being in a hospital could put them more at risk for COVID-19 infection:

> *Either way it's fine with me, but for me, the virtual is the best for everybody's sake. I believe that's the best way to do it. I'm fine with it. A lot of other people are not.* [Black or African American male patient aged 69 years, living with diabetes for 18 years]

> *No. I think [I don't have a preference for] either one, I mean, I like the convenience of not having to go into the office, especially right now having COVID and then hearing that I could get it again, the safety part of it.* [White female patient aged 43 years, living with diabetes for 8 years]

> *One thing I did like, I was at my house. I didn't have to go through all the shenanigans at the hospital with checking in and all this kind of stuff and social distancing stuff, worrying about blood pressure. Some people social distance, some people don't. I didn't have to be bothered with that.* [Black or African American male patient aged 46 years, living with diabetes for 2 years]

> *I liked it because my doctors know I'm taking my medicine and I like him because he has a pleasant spirit. I don't like driving there and finding parking. I didn't feel rushed. I would do a virtual visit, I'm nervous of the virus so I don't want to go in the office.* [Black or African American female patient aged 66 years, living with diabetes for 18 years]

Health care providers also mentioned safety from COVID-19 as a factor in telehealth visits. This was especially important for providers who worked in locations where it was difficult to physically maintain the distance between patients or in crowded waiting rooms. In such cases, the risk was perceived as especially high, particularly for patients with multiple conditions or those who were not vaccinated:

> *I don't think it's really fair or safe, [for] the patients who aren't vaccinated to come in and sit in a small lobby early with patients who have diabetes and the like. I just don't think it's good public health. You choose not to be vaccinated then you're probably going to get a telehealth or televideo call...I'm excited about that.* [Health care provider aged 75 years, practicing for 46 years]

Other facilitators for telehealth included patients feeling that their telehealth visits provided sufficient time with their provider and that their provider was attentive and took an appropriate amount of time with the telehealth visits:

> *I liked that I didn't have to go out there, and still I felt like I was right there talking to him. He is tall and big and younger than me and just like a sweet teddy bear not really overweight just like a teddy bear. He was a doctor that I could sit and talk to him and he would listen. It took me years to find him.* [White female patient aged 75 years, living with diabetes for about 10-15 years]

> *It works fine for me. I feel like my doctor is very attentive and I can ask any questions, so it works fine either way for me.* [Black or African American female patient aged 51 years, living with diabetes for 7 years]

> *I didn't feel [rushed]. Everything was answered. Everything was fine as if I was in the office because I got my instructions on what I needed to do with my blood work. It was just fine.* [Black or African American female patient aged 48 years, living with diabetes for 2 years]

XSL•FO
RenderX

Participants also mentioned the comfort and ease of telehealth visits. One participant mentioned that given the ease of a telehealth visit, she was more inclined to call the physician, knowing she would not have the hassle of driving in:

> *Because it's so much easier for me to be able to do one of these visits just over my phone from home, you know that sometimes, there's something that goes wrong, you think, "Oh, I should probably go to the doctor, but I don't want to mess with it"? I think I would be more inclined to go ahead and contact the doctor knowing that I wouldn't have to go in.* [White female patient aged 60 years, living with diabetes for 20 years]

A patient reported that their telehealth visit allowed them to be prepared and enabled a thorough assessment, especially as it was the first visit:

> *It allows me to go ahead and have my questions prepared to raise the questions that I need to raise. They go through reports...with the first virtual visit that I had [with my doctor], I almost died. It was 45 minutes long. [That's] the thing my son said was good about it is that he was here, so he heard everything.* [White female aged 78 years, living with diabetes for about 45 years]

Valuing Telehealth Differently From In-Person Visits

Although the patient participants described appreciating the convenience, safety, and comfort of telehealth visits, a model of care that included both in-patient and telehealth visits was considered optimal. One caveat is the need for laboratory tests, such as hemoglobin A_{1c} (HbA_{1c}), for which many patients underwent laboratory work at outside facilities before their telehealth visits. Patients who had difficulty uploading previously obtained laboratory results and navigating technology preferred in-person care:

> *[The telehealth visit] was good, but I told him I couldn't wait to see him. I would rather go into the office. It was alright, but I like when I go to see him. He's a very patient doctor the [telehealth] conversation didn't feel rushed, I just think in person was better.* [Black or African American female patient aged 67 years, living with diabetes for 7 years]

> *I would use [telehealth] in the future unless I was having problems, like if I had other issues like high blood pressure or something like that. I wouldn't use it all the time, but to be convenient like, if I just needed a prescription refill, the telehealth visit would be fine for me.* [Black or African American female patient aged 48 years, living with diabetes for 2 years]

In addition to endorsing the importance of a physical check, participants wondered about the billing associated with telehealth visits. One patient described this as follows:

> *Well, virtual—I didn't get any information, no blood tests and there's no A1C. It was a pointless endeavor. It was just [hospital] saying, "We can't figure this out. We're just going to do a pretend appointment so we can bill the insurance company." It was a phone call, so I had the phone in my hand, I'm in my room by myself. They asked me how I'm doing, I said fine and that was the end of it. Probably less than a minute.* [White male aged 70 years, living with diabetes for 3 years]

For clinicians, telehealth visits were generally described as a good asset for diabetes patient care. Provider sentiment was that telehealth worked best as an additional element for patients but could never fully replace in-person care. Similar to patients, clinicians discussed the convenience of telehealth visits, especially for patients who lived in rural areas, had issues with transportation, or were older adult patients who have mobility challenges:

> *I think probably 80% of patients are fine. We could do this over the video or no video or phone, not in person, let's say. Unless you have open wounds or unless there's changes if we had seen you before, and you're saying, "Well, Doc, everything's similar."...I think probably majority is fine. Not to say we should convert all of their visits to tele-visits, but possibly two out of three, or every other, or some, whatever is convenient.* [Female health provider aged 38 years, practicing for 7 years]

In certain circumstances, providers could not substitute a telehealth visit for an in-person visit, such as with a new patient, patients with multiple chronic conditions, where a patient had major changes in their health status, or where a patient had symptoms of uncontrolled diabetes. One provider noted difficulty in seeing patients who were not English speaking for telehealth visits, as it was challenging to get an interpreter for a telehealth call. A physician described the need to see new patients in person as follows:

> *I'm personally seeing all my new people in person. If they were seen by another endocrine provider, or if they were a new referral, I see them the first time in person. I just feel more comfortable that way, knowing that I've examined them thoroughly. We have their vitals, we've gone through the meds. After that, I usually do follow up tele-visits.* [Female health provider aged 40 years, practicing for about 5 years]

Clinicians discussed telehealth as generally best for patients whose diabetes was well controlled as a check-in to ensure that patients were continuing to do well. Overall, however, providers would prefer to see patients in person if there were any issues or complications:

> *For the patients who are doing very well, who are largely on auto pilot, for whom there's not very much to discuss...It's usually just, "Oh the labs look very good, just keep doing what you're doing." It works well, it works quite well, but for the patients who have more complexity, more comorbidities for some, there's a fair amount that needs to be done. You end the meeting feeling that there's still more that needs to be done. In some cases, you actually have to tell the patient that they do need to set up an appointment to be seen in person. It all depends on how much current*

XSL•FO
RenderX

> *comorbidity and how well controlled the patient is.* [Male health care provider aged 53 years, practicing for 28 years]

Given the benefits and minor drawbacks of telehealth visits for diabetes care, provider participants described telehealth as being a permanent part of patient care. One physician noted:

> *I really do think there's definitely place for telemedicine for these patients even past pandemic. Maybe they come into the clinic once in six months or even once a year depending on how well their control is for diabetes. Of course, if they truly need to be seen, they need to be seen.* [Female health provider aged 38 years, practicing for 7 years]

Convenience of Telehealth May Have a Downside

A downside to the flexibility and convenience of telehealth visits may be that it also affects how well prepared patients are for a visit and lessens focus and attention on the visit content itself. As telehealth visits can be taken from a wide variety of locations, patients sometimes reported having their visits in places that were not private or where significant distractions prevented them from focusing on the visit. This was especially noted during telephonic visits.

One participant described taking a call from the car from the side of the road, whereas another reported taking a call while out shopping. Patients who were multitasking and engaging in driving or shopping could be less likely to have laboratory or blood glucose results available for the visit:

> *With me, it was fine the first time because after we talked on text, I didn't have to leave work and go over there...I pulled over on the side of the road and we talked and after...I went on to my destination.* [Black or African American male patient aged 69 years, living with diabetes for 18 years]

> *I like the fact that even if I am at work or home, or even in the car, it's just the flexibility, like, "Hey, if I can't make it to the—If I couldn't get to the doctor's office, I could do it from wherever."* [Black or African American male patient aged 47 years, living with diabetes for 1 year]

Although participants valued the comfort and convenience of telehealth visits, health care providers noted that they may not be as well prepared for such visits. One provider mentioned that some patients do not see the telehealth visit as a "real" physician's visit, so they can be less prepared for the visit or in a location that is distracting:

> *I find patients are going about their day and then realizing their appointment is live and [saying] "I don't have my data because I'm at Target." They didn't really think of it as a doctor's visit and so they didn't prepare as well as they normally would...so it's a little bit distracting. They don't have their pills with them or they are not at home to say, "Oh, I'm taking this much of that or this medicine or that medicine."* [Female health provider aged 38 years, practicing for 7 years]

A clinic coordinator also noted that patients sometimes did not follow through laboratory tests in the context of telehealth visits, impacting their ability to follow-up:

> *The doctor does the telehealth visit, he put in his plan, he wants the patient to follow up in 3 months with labs. Then I mail the patient a reminder appointment, advise them, and get the labs done before their appointment. Once a patient clicks on for the visit, if they don't do the lab—usually I prep the schedule before an appointment and be like "You didn't do your lab?" That kind of thing. The patient being compliant is a big factor with a telehealth visit.* [Male health provider aged 53 years, practicing for 23 years]

When patients are on the go, the technology can also be less reliable, as noted by another clinician:

> *The negative side of [telehealth] is the internet connection...If the patient is not at home or a home setting, like if they are in a car, I find those patients won't be able to get good reception.* [Female health provider aged 38 years, practicing for 7 years]

Clinicians also noted often having to wait on the line for patients to locate blood glucose logs or medicines. This situation was contrasted with in-person visits, for which patients know they are expected to arrive with their information. Where patients needed to find laboratory results or other information during a visit or call, it could significantly delay visit length having a knock-on effect for subsequent telehealth patients:

> *With telehealth, usually when they come into the clinic, their vitals are done and the nurse reconciles their meds for them...Most people do have the equipment at home to check. It's not really been an issue, but sometimes you have to wait for them to do it while you're on the phone with them just to make sure that their blood pressure is okay. The same thing with medication, sometimes they don't remember or they have to go check. You're waiting a little bit for that to happen because you don't have support of staff, and you're doing it yourself, those things take longer sometimes.* [Female health provider aged 40 years, practicing for about 5 years]

Technology as a Double-Edged Sword

Comfort level and proficiency with technology were variable among both patients and health care providers, as were the different modes and platforms used by health systems to provide telehealth visits.

Some patients were only comfortable using their phones for a visit, whereas others had no trouble navigating video telehealth visits accessed through a patient portal. Those who felt most comfortable with technology were more amenable to telehealth as part of their care and seemed to have a better and more efficient experience. One patient described this as follows:

> *No, [no difficulties with the technology]. I was fortunate enough, we had been using that type of communication at work, so I was a little familiar with the process, how it works. No, I didn't have any issues*

XSL•FO
RenderX

with it at all. [White male patient aged 57 years, living with diabetes for 1 year]

However, navigating technology for many patient participants was an issue. Depending on their technological capabilities and schedules, patients preferred either phone or video visits. Patients who were less technologically savvy described frustrations navigating the system as well as getting used to the telehealth visit experience and were more likely to value in-person care. Some participants mentioned having help from family members or having someone from the hospital call before their first visit to ensure that they could log on; however, others did not have support with the technology. One participant mentioned the following:

I just think that when it's your first time on there, if they actually had help for someone, an elderly person to do telehealth. For the first time, I think somebody might need to coach that person or assist that person logging on. Just make it friendlier for the elderly or the up in age people. [Black or African American female patient aged 46 years, living with diabetes for 2 years]

I was a little concerned for the very first time, I was like, "Oh, I hope I don't have a problem with the phone, what if my video doesn't work?" I had no problems from day one, everything worked out perfectly. [Black or African American female patient aged 51 years, living with diabetes for 7 years]

The first time it was aggravating because I couldn't make the [technology] connection, so they just got me later. Generally speaking, I prefer to go in person, but depending on which doctor and what reason [for it], telehealth is just fine. [White female patient aged 59 years, living with diabetes for about 15 years]

From the clinician perspective, the specific technology, for example, the internet connection and how comfortable the patient was in navigating the connection, was mentioned as being very challenging. This was likely to be more of an issue for health systems primarily serving low-income patients. A clinician within a federally qualified health center described it as follows:

It was challenging, even to get them registered between the back and forth, we would have my medical assistant, myself and an IT person and the nurse, so you would have 4 people trying to launch one visit. It would take 45 min to do a back and forth...just trying to troubleshoot and talk a patient through on a landline, to help launch something on an iPad. [Female health provider aged 51 years, practicing for 22 years]

One provider described how some lower-income patients had "Katrina phones," which were provided just after Hurricane Katrina in 2005 and do not have video, voicemail, or text options, making it difficult to call back or get in touch with the patients. Other issues described were patients not answering a call from a phone number that they did not recognize or not being savvy with text or voicemail, which hampered scheduling or contacting patients:

Some patients have phones that they call Katrina phones. It was government- sponsored phones that don't have voicemail, that have very few minutes and so…they either pick up or they don't pick up. The other issues is when we call from the hospital, the phone [number] may be private...most of our patients…don't answer any of those...Then some patients have a fancy phone…and they don't really know how to use it...I think, definitely, a good portion of my population is not really tech-savvy. Again, I think training them to use this more and more it will work, it's fine. It's just we couldn't make it all happen in a month. [Provider 2]

There were technical issues going on that I think made [the telehealth process], particularly at the beginning, it wasn't always very smooth and patients weren't looking at times and everything. The elderly were having difficulty linking in. [Male health provider aged 75 years, practicing for 46 years]

Telehealth visits were most beneficial and efficient when patients could navigate the system and upload their health information, such as blood pressure, glucose levels, and weight. When patients did not have this information available, telehealth visits were not as comprehensive:

Some patients don't have access to take their own vital signs, like their blood pressure, pulse, they don't [have] a pressure kit. When they log into the app, the link is going to ask them their weight, blood pressure, and if they have that available, they're supposed to input it...some patients input their vitals some don't. [Female health provider aged 44 years, practicing for 12 years]

However, it was noted that as patients had more telehealth visits, the process improved. One clinician stated as follows:

We still have occasional glitches where you have a patient who is scheduled and signed up for a telehealth visit, and there's trouble with them logging in. It's a lot less now because people are becoming more accustomed. If they've done it once, they've done it twice, now it's a matter of habit. They're familiar with it now. [Female health provider aged 51 years, practicing for 22 years]

Managing Expectations and Efficiency of the Visit Experience

Expectations about how efficient telehealth visits should be, including scheduling, logging in, and undertaking a visit, were frequently noted. On the one hand, patient expectations for timely and efficient care were not met, whereas on the other hand, care providers were limited by their ability to meet expectations based on how their schedules were set up and how their health system operated telehealth visits.

Patients sometimes described long wait times for health care providers to log into their telehealth visits when they had an appointment scheduled for a specific time. As telehealth calls can be taken from anywhere and had less of a wait time to schedule, there was a perception that there would not be a wait

XSL•FO
RenderX

time for a visit. The participants were frustrated when these expectations were not met:

> *I did have a problem when it was time for my appointment because on the app, you have like a grey screen, and it stays like that for a while because you're waiting on the doctor to come in, so I called him twice because the appointment was going over time. My appointment was for 10:00. It was 10:30 and nobody had came on the line. I thought I was doing something wrong or I thought something had happened or I missed the appointment.* [Black or African American male patient aged 46 years, living with diabetes for 2 years]

Providers noted that they could not meet patients' expectations of efficiency because of scheduling and other issues that were outside their control. This contrasted with the expectation that patients had with in-person visits. One clinician described this as follows:

> *I think when a patient is in your office and they're sitting for 20 minutes, it's expected my doctor is running late...Unfortunately, that's the reality. It shouldn't be that way, but that's how it is. On the phone, when you say I'm going to call around 11, they really think it's going to be exactly 11. You may be 11:30 just like you would be in clinic or even harder because you can't get in touch with them themselves let alone the rest of your clinic flow.* [Female health provider aged 38 years, practicing for 7 years]

One patient participant described being in his car, not expecting to have to wait for the provider to log on. Participants mentioned that once they had 1 or 2 visits, the process became easier, and they knew what to expect:

> *I just assumed that I could [log in to telehealth] if I was out. I assumed that I can still have a telehealth visit where I was at, but I remember saying I can go sit in my car. I think it was going to be too long to sit in the car. It had to be rescheduled.* [Black or African American female patient aged 48 years, living with diabetes for 2 years]

Another participant waited 15 minutes to reschedule because of technical issues from the provider. They stated the following:

> *For me, it wasn't hard, but I know on the other end, I don't know if it's because just they were just starting using it because there were a couple of times where I've called...I logged in for the telehealth visit and nobody was there and I waited 15 minutes and I said, "Either they are running over or they forgot, or something happened." I log out, then I get a message saying, "Hey sorry, we were having technical difficulties. Can we reschedule for such and such a time?"* [Black or African American male patient aged 47 years, living with diabetes for 1 year]

Clinicians lamented the challenges and need to temper patient expectations of timeliness in telehealth scheduling:

> *If you had patients scheduled on half an hour visit, but it took you 45 minutes just to launch it, and then another 15 to 20 minutes on the visit, we would be behind. We're trying to call the patient saying, "We're sorry we're late. Just give us a minute." One patient could take upwards of of an our to do a 15 minute visit or a 20 minute visit.* [Female health provider aged 44 years, practicing for 12 years]

> *Yes, definitely, patients have expressed their frustration with, "Well, I thought they were going to call at 11 and it's 11:30, or 40, or 20, whatever it may be." I think they expect a little more easier contact just because it's by the phone, and to us, it's harder, I think.* [Female health provider aged 38 years, practicing for 7 years]

Scheduling issues and wait times were commonly discussed by health care providers as some of the biggest barriers to efficient care:

> *There are patients who they may get in 30 minutes before the appointment and then leave out 10 minutes before...and then we have to track them down and say, "You didn't wait long enough...your appointment was at 2:00, you went in at 1:30 and left at 1:45." Then they've lost the link...particularly at the beginning it wasn't always very smooth and patients weren't looking at times and everything.* [Male health provider aged 75 years, practicing for 46 years]

Clinician participants described challenges with scheduling in-person versus telehealth visits, which affected patient wait times and depended on how their health system organized the clinician workflow:

> *When we initially started, we had them interjected into the day and that's when we realized the inefficiencies with getting the system to launch. We quickly transitioned to one or two, and each provider had their preference. I had a telehealth day. I had one day that was designated for telehealth and that's it. Then some providers they had a certain time frame designated for telehealth with maybe their mornings, they saw patients in the office, and then in the afternoon they would have a few telehealth scheduled.* [Female health provider aged 51 years, practicing for 22 years]

> *I try to group them all together. We're going to do virtual in the morning and in-person evening. I don't like to have a virtual and then in-person, because...you're going to run late behind if you have an in-person patient and then you got a televisit. Because somewhere you might have went over with the in-person visit, and the patient is on to the televisit. Those kinds of situations make a doctor run late too.* [Female health provider aged 44 years, practicing for 12 years]

Discussion

Principal Findings

The study found that both patients and providers appreciated the flexibility and convenience of telehealth visits. However,

XSL•FO
RenderX

telehealth visits were considered by clinicians to be best for those whose diabetes is well controlled or less complicated for patients who were more tech savvy and have internet connectivity and for those with transportation challenges. A caveat of telehealth visits was that patients seemed not to approach the visit with the same focus and preparation as an in-person visit—patients could be distracted, lack important information on their condition, or hold different expectations about the visit, such as short wait times. In addition, some patients felt that telehealth visits were not as comprehensive. A mixed model of both in-person and telehealth visits was considered optimal by most patients and providers of diabetes care.

A recent study in Missouri found that women, older patients, and those with Medicare, Medicaid, and self-pay statuses used telehealth more during the beginning of the COVID-19 pandemic, although older patients; Black patients; urban patients; and those with Medicare, Medicaid, and self-pay status were more likely to use phone only compared with audio-video telehealth services [23]. Our study sheds light on some of the nuances of the experiences of a similar population in the South, as our patient population identified as predominantly female, older, and using Medicaid or Medicare and half identified as Black.

Our results showed that overall, for both patients and health care providers, telehealth was viewed as an important part of diabetes management and care; however, a hybrid model incorporating both in-person and telehealth visits was considered optimal, as has been noted in other studies [15]. Despite issues with connectivity and scheduling complications, both patients and providers appreciated the flexibility and convenience of telehealth visits, as has been described in other studies [19,21,22,33]. However, from both client and clinician perspectives, telehealth visits were considered best for those whose diabetes was well controlled, for patients without complications, and for those who were tech savvy or have good internet or cellular connectivity, which echoes the findings from prior studies on telehealth [5,17,19]. The benefits of telehealth appear to be especially important for patients with diabetes who have difficulty traveling to see their physician, including older adult patients, with the caveat that they must also be comfortable navigating technology [33,34], and in particular, telehealth has been shown to be beneficial for patients in rural communities [35].

For patients with chronic diseases, such as diabetes, health management through telemedicine can be crucial for continuity of care and avoiding exposure to COVID-19 [6,36]. However, the convenience and flexibility of telehealth visits must be weighed against the perception that virtual care may be less comprehensive or effective. Health care providers in the study discussed how patients sometimes would not approach a telehealth visit with the same focus and preparation as an in-person visit. Motivation to continue engaging in diabetes care is another important consideration, as one study noted the importance of combining “eHealth” with regular face-to-face consultations to avoid reducing patient motivation for engagement [37]. There may be downsides to the flexibility and convenience of telehealth, as patients and providers described visits taking place in distracting environments, difficulty in managing expectations regarding wait times for visits, or lack of diagnostic information, such as available blood glucose results.

Comparison With Prior Studies

Studies have indicated that most patients feel that telehealth visits can be as effective as in-person visits [16,38]. However, one study identified that patient respondents with higher HbA_{1c} levels who completed a telehealth visit were more likely to report that video care did not save them time, money, or stress, although a majority of those respondents still felt that video care was effective or beneficial in some way. In the same study, respondents with higher HbA_{1c} levels who had *not yet had* a video appointment were also more likely to report difficulty connecting remotely and were more concerned about the quality of care using video [15].

Technology was discussed by participants in this study as both a barrier and a facilitator for telehealth. Both patients and providers described a learning curve for using telehealth services. Older adults who were not as comfortable with technology seemed to prefer telephonic visits compared with video visits because of the ease of cell phone use, which has been noted in other studies [17,23,39,40]. Our study highlights some barriers to audio-video telehealth visits for this population, such as outdated technology (Katrina phones), inconsistent internet service, and lack of comfort with technology. Furthermore, age, sex, median household income, insurance status, and marital status have been found to be associated with patient participation in telehealth [41], providing evidence of structural inequality affecting patient comfort and the ability to engage in and benefit from telehealth. One study found that communities at greater risk of needing support after a disaster were significantly more likely to experience more barriers to telehealth, including access to reliable internet, low uptake of use of web-based medical portals, not feeling comfortable with technology, and more likely to report language barriers as a concern [42]. The Gulf South and Louisiana in particular are at high risk for similar events.

For patients, having the support to navigate early visits appeared to be important for improving the experience, and with more exposure to telehealth, the process became easier. Patients in this study had a broad range of experiences with telehealth (from 1 to 5 visits), which has been noted to impact perceptions of telehealth, particularly related to comfort with technology [35]. Issues and challenges with technology and computer literacy have been found to be significant barriers to the adoption of telehealth [43]. Similar hindrances have also been noted with other forms of non–face-to-face care in this setting [24,26], indicating that in all such remote clinical encounters, patients will benefit from support along the journey of chronic care management. Given the potential for telemedical interventions to be clinically effective in improving diabetes control overall and significantly improving HbA_{1c} concentrations [44], it is vital that these strategies be optimized to address potential challenges in advance, ensuring quality of care and value in health.

XSL•FO
RenderX

Understanding the experiences of patients with diabetes and clinicians who care for populations with diabetes is important, especially given the significant risks that patients with diabetes face due to infectious diseases such as COVID-19 [8]. The ability to avoid facility waiting rooms, offices, and exposure to large numbers of people makes telehealth an important tool for people with diabetes, and this factor coupled with the dramatic increase in the use of telehealth services in recent years [11] makes it crucial to investigate lived experiences with telehealth. Some of this study's findings may be representative of an abrupt change to telehealth; for example, the facilitator of increased safety and comfort of telehealth during the pandemic and technological challenges. However, other themes are likely to be more persistent, such as the downside of convenience and mismatched expectations of telehealth visit efficiency and wait times. Additional studies should focus on whether barriers to telehealth seen in the immediate period after the declaration of the COVID-19 pandemic are persistent for people with diabetes and their health care providers or whether these fade for patients and clinicians.

Strengths and Limitations

This study used a rigorous qualitative approach to understand the telehealth experiences of patients living with diabetes during COVID-19 as well as the experiences of clinicians who treat patients with diabetes in Louisiana. This study had some limitations. First, the participants were mostly older adults, which may not be fully representative of all populations with diabetes who use telehealth, but were representative of key populations of interest in the region with the largest burden of chronic disease. Clinicians from a range of ages and years of practice were included. Different characteristics and experiences did provide a range of perspectives on telehealth barriers and facilitators. Second, some challenges were encountered in recruitment, including pandemic shutdowns and a hurricane in the region, which may have affected which participants were ultimately able to participate in the study. Finally, participant reports of experiences may have been influenced by recall bias or social desirability bias. Interviews took place at different lengths of time from the participants' last experiences with telehealth. Certain limitations were outside our control (pandemic and hurricane); however, efforts to ensure reliability and trustworthiness were used throughout the study. As with all studies, there is a risk that participants may selectively present positive and socially desirable responses to experiences.

Conclusions

Telehealth plays an important role in the management of diabetes and may be especially important in areas with high prevalence, such as Louisiana. Issues of internet connectivity and proficiency using technology must be addressed to ensure equitable access across patient populations. In addition, support to help guide and inform patients on what to expect and how to prepare for telehealth visits is recommended to improve the experience of both health systems and patients. Preparing patients with diabetes for what to expect during telehealth visits, including having important health information available during the visit and managing expectations over wait times, is an important strategy. Similarly, allowing clinicians to schedule telehealth and in-person visits during discrete blocks of time could encourage efficiency. Further research should address how hybrid models of telehealth and in-person care may differentially impact health outcomes for patients with diabetes, particularly for people with multiple chronic conditions and in settings where access to technology and connectivity is not optimal.

Acknowledgments

The Louisiana Experiment Assessing Diabetes study would like to acknowledge the contributions of our partners, without whom this research would not be possible. These partners include Ochsner Health System, Tulane Medical Center, University Medical Center New Orleans, Research Action for Health Network (REACHnet, a PCORnet Clinical Research Network) and their multistakeholder Diabetes Advisory Groups, Pennington Biomedical Research Center, and Blue Cross and Blue Shield of Louisiana. Patients and community partners were also actively engaged in the research and contributed to this study. Patient Engagement Statement: patients and other relevant stakeholders as research partners were engaged in the planning of the study as well as the refining and finalizing of data collection forms and protocols and the results and discussion of this study. The authors collaborated with the program-wide steering committee, including patient partners, to ensure that the final study and analytical protocol were consistent with the relevant US Centers for Disease Control and Prevention, National Institute of Diabetes and Digestive and Kidney Diseases, and Patient-Centered Outcomes Research Institute (PCORI) methodological standards. Cathy Glover and Patricia Dominick are gratefully acknowledged for their contributions and specifically for the review of this work. This work was supported by a PCORI Award (NEN 1508-32257) through the NEXT-D2 Program, jointly sponsored by the US Centers for Disease Control and Prevention, the National Institute of Diabetes and Digestive and Kidney Diseases, and PCORI. All statements in this paper, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of PCORI, its Board of Governors, or Methodology Committee.

Data Availability

All data generated or analyzed during this study are included in this published paper.

XSL•FO
RenderX

Authors' Contributions

ANB and LS contributed to conceptualization. ANB, TP, and DC contributed to data curation and formal analysis. LS, ANB, and EN contributed to funding acquisition. ANB and TP contributed to investigation. ANB contributed to methodology. ANB, TP, and DC contributed to project administration. ANB, TP, and DC contributed to writing the original draft. ANB and DC contributed to reviewing and editing the paper. All authors reviewed and approved the final version of this manuscript.

Conflicts of Interest

None declared.

References

1. Adult obesity maps. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/obesity/data/prevalence-maps.html [accessed 2023-11-21]
2. National and state diabetes trends. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/diabetes/data/statistics-report/index.html [accessed 2023-11-29]
3. Heart disease mortality by state. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/nchs/pressroom/sosmap/heart_disease_mortality/heart_disease.htm [accessed 2023-11-29]
4. Fact sheet: telehealth. American Heart Association. 2019. URL: https://www.aha.org/system/files/2019-02/fact-sheet-telehealth-2-4-19.pdf [accessed 2023-11-21]
5. Karimi M, Lee E, Couture S, Gonzales A, Grigorescu V, Smith S, et al. National survey trends in telehealth use in 2021: disparities in utilization and audio vs video services. Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. 2022. URL: https://collections.nlm.nih.gov/catalog/nlm:nlmuid-9918590087506676-pdf [accessed 2023-11-29]
6. Diabetes management: team-based care for patients with type 2 diabetes. The Community Guide. 2016. URL: https://www.thecommunityguide.org/findings/diabetes-management-team-based-care-patients-type-2-diabetes [accessed 2023-11-29]
7. National diabetes statistics report website. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/diabetes/data/statistics-report/index.html [accessed 2023-11-29]
8. How COVID-19 impacts people with diabetes. American Diabetes Association. 2022. URL: https://tinyurl.com/4bfnke7m [accessed 2023-11-21]
9. Adhikari S, Pantaleo NP, Feldman JM, Ogedegbe O, Thorpe L, Troxel AB. Assessment of community-level disparities in coronavirus disease 2019 (COVID-19) infections and deaths in large US metropolitan areas. JAMA Netw Open 2020 Jul 01;3(7):e2016938 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.16938] [Medline: 32721027]
10. Price-Haywood EG, Burton J, Fort D, Seoane L. Hospitalization and mortality among black patients and white patients with COVID-19. N Engl J Med 2020 Jun 25;382(26):2534-2543 [FREE Full text] [doi: 10.1056/NEJMsa2011686] [Medline: 32459916]
11. Poeran J, Cho LD, Wilson L, Zhong H, Mazumdar M, Liu J, et al. Pre-existing disparities and potential implications for the rapid expansion of telemedicine in response to the Coronavirus Disease 2019 pandemic. Med Care 2021 Aug 01;59(8):694-698 [FREE Full text] [doi: 10.1097/MLR.0000000000001585] [Medline: 34054024]
12. Dave D, Friedson AI, Matsuzawa K, Sabia JJ. When do shelter-in-place orders fight COVID-19 best? Policy heterogeneity across states and adoption time. Econ Inq 2021 Jan;59(1):29-52 [FREE Full text] [doi: 10.1111/ecin.12944] [Medline: 32836519]
13. Santos DS, Batistelli CR, Lara MM, Ferreira ED, Moreira TR, Cotta RM. The effectiveness of the use of telehealth programs in the care of individuals with hypertension and, or diabetes mellitus: systematic review and meta-analysis. Diabetol Metab Syndr 2022 May 28;14(1):76 [FREE Full text] [doi: 10.1186/s13098-022-00846-5] [Medline: 35643495]
14. Lauffenburger JC, Barlev RA, Sears ES, Keller PA, McDonnell ME, Yom-Tov E, et al. Preferences for mHealth technology and text messaging communication in patients with type 2 diabetes: qualitative interview study. J Med Internet Res 2021 Jun 11;23(6):e25958 [FREE Full text] [doi: 10.2196/25958] [Medline: 34114964]
15. Crossen SS, Romero CC, Loomba LA, Glaser NS. Patient perspectives on use of video telemedicine for type 1 diabetes care in the United States during the COVID-19 pandemic. Endocrines 2021 Dec 01;2(4):449-456 [FREE Full text] [doi: 10.3390/endocrines2040040] [Medline: 35373189]
16. The rapid transition to telemedicine:insights and early trends. Press Ganey. 2020. URL: https://www.matrc.org/wp-content/uploads/2020/04/Press-Ganey-Telemedicine_-1.pdf [accessed 2023-11-29]
17. Lipschitz JM, Connolly SL, Van Boxtel R, Potter JR, Nixon N, Bidargaddi N. Provider perspectives on telemental health implementation: lessons learned during the COVID-19 pandemic and paths forward. Psychol Serv (Forthcoming) 2022 Mar 24 [FREE Full text] [doi: 10.1037/ser0000625] [Medline: 35201809]
18. DePuccio MJ, Gaughan AA, Shiu-Yee K, McAlearney AS. Doctoring from home: physicians' perspectives on the advantages of remote care delivery during the COVID-19 pandemic. PLoS One 2022 Jun 2;17(6):e0269264 [FREE Full text] [doi: 10.1371/journal.pone.0269264] [Medline: 35653337]

XSL•FO
RenderX

19. Rao L, Comfort AB, Dojiri SS, Goodman S, Yarger J, Shah N, et al. Telehealth for contraceptive services during the COVID-19 pandemic: provider perspectives. Womens Health Issues 2022 Sep;32(5):477-483 [FREE Full text] [doi: 10.1016/j.whi.2022.05.001] [Medline: 35691762]
20. Lee M, Luna P, Lynch S, Nagpal S, Dominguez YC, Ahmed Z, et al. Telehealth provider perspectives during COVID-19: insights from an academic cardiology practice. J Am Coll Cardiol 2021 May;77(18):3212. [doi: 10.1016/s0735-1097(21)04567-8]
21. Singh A, Mountjoy N, McElroy D, Mittal S, Al Hemyari B, Coffey N, et al. Patient perspectives with telehealth visits in cardiology during COVID-19: online patient survey study. JMIR Cardio 2021 Jan 22;5(1):e25074 [FREE Full text] [doi: 10.2196/25074] [Medline: 33385234]
22. Auchus IC, Jaradeh K, Tang A, Marzan J, Boslett B. Transitioning to telehealth during the COVID-19 pandemic: patient perspectives and attendance at an HIV clinic in San Francisco. AIDS Patient Care STDS 2021 Jul;35(7):249-254 [FREE Full text] [doi: 10.1089/apc.2021.0075] [Medline: 34242090]
23. Pierce RP, Stevermer JJ. Disparities in the use of telehealth at the onset of the COVID-19 public health emergency. J Telemed Telecare 2023 Jan;29(1):3-9 [FREE Full text] [doi: 10.1177/1357633X20963893] [Medline: 33081595]
24. Wharton MK, Shi L, Eragoda S, Monnette AM, Nauman E, Price-Haywood EG, et al. Qualitative analysis of health systems utilizing non-face-to-face chronic care management for medicare-insured patients with diabetes. J Ambul Care Manage 2020 Oct;43(4):326-334. [doi: 10.1097/JAC.0000000000000342] [Medline: 32858736]
25. Bazzano AN, Monnette AM, Wharton MK, Price-Haywood EG, Nauman E, Dominick P, et al. Older patients' preferences and views related to non-face-to-face diabetes chronic care management: a qualitative study from southeast Louisiana. Patient Prefer Adherence 2019 May;13:901-911 [FREE Full text] [doi: 10.2147/PPA.S201072] [Medline: 31213782]
26. Bazzano AN, Wharton MK, Monnette A, Nauman E, Price-Haywood E, Glover C, et al. Barriers and facilitators in implementing non-face-to-face chronic care management in an elderly population with diabetes: a qualitative study of physician and health system perspectives. J Clin Med 2018 Nov 20;7(11):451 [FREE Full text] [doi: 10.3390/jcm7110451] [Medline: 30463310]
27. Shao Y, Stoecker C, Hong D, Nauman E, Fonseca V, Hu G, et al. The impact of reimbursement for non-face-to-face chronic care management on comprehensive metabolic biomarkers among multimorbid patients with type 2 diabetes. Med Care 2023 Mar 01;61(3):157-164. [doi: 10.1097/MLR.0000000000001816] [Medline: 36728398]
28. Walker B, Stoecker C, Shao Y, Nauman E, Fort D, Shi L. Telehealth and medicare type 2 diabetes care outcomes: evidence from Louisiana. Med Care 2023 Apr 01;61(Suppl 1):S77-S82 [FREE Full text] [doi: 10.1097/MLR.0000000000001724] [Medline: 36893422]
29. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
30. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
31. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar 26;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
33. Goldberg EM, Lin MP, Burke LG, Jiménez FN, Davoodi NM, Merchant RC. Perspectives on telehealth for older adults during the COVID-19 pandemic using the quadruple aim: interviews with 48 physicians. BMC Geriatr 2022 Mar 08;22(1):188 [FREE Full text] [doi: 10.1186/s12877-022-02860-8] [Medline: 35260091]
34. Kichloo A, Albosta M, Dettloff K, Wani F, El-Amir Z, Singh J, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. Fam Med Community Health 2020 Aug;8(3):e000530 [FREE Full text] [doi: 10.1136/fmch-2020-000530] [Medline: 32816942]
35. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221 [FREE Full text] [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
36. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
37. Lie SS, Karlsen B, Oord ER, Graue M, Oftedal B. Dropout from an eHealth intervention for adults with type 2 diabetes: a qualitative study. J Med Internet Res 2017 May 30;19(5):e187 [FREE Full text] [doi: 10.2196/jmir.7479] [Medline: 28559223]
38. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 Dec;15(3):225-229 [FREE Full text] [doi: 10.1370/afm.2095] [Medline: 28483887]
39. Yeager VA, Wharton MK, Monnette A, Price-Haywood EG, Nauman E, Angove RS, et al. Non-face-to-face chronic care management: a qualitative study assessing the implementation of a new CMS reimbursement strategy. Popul Health Manag 2018 Dec;21(6):454-461. [doi: 10.1089/pop.2017.0196] [Medline: 29658847]

XSL•FO
RenderX

40. Friedman EE, Devlin SA, Gilson SF, Ridgway JP. Age and racial disparities in telehealth use among people with HIV during the COVID-19 pandemic. AIDS Behav 2022 Aug;26(8):2686-2691. [doi: 10.1007/s10461-022-03607-7] [Medline: 35133528]
41. Darrat I, Tam S, Boulis M, Williams AM. Socioeconomic disparities in patient use of telehealth during the coronavirus disease 2019 surge. JAMA Otolaryngol Head Neck Surg 2021 Mar 01;147(3):287-295 [FREE Full text] [doi: 10.1001/jamaoto.2020.5161] [Medline: 33443539]
42. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post-COVID era. Milbank Q 2021 Jun;99(2):340-368 [FREE Full text] [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
43. Scott KC, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
44. Eberle C, Stichling S. Clinical improvements by telemedicine interventions managing type 1 and type 2 diabetes: systematic meta-review. J Med Internet Res 2021 Feb 19;23(2):e23244 [FREE Full text] [doi: 10.2196/23244] [Medline: 33605889]

Abbreviations

HbA1c: hemoglobin A1c

Edited by G Eysenbach, T Leung; submitted 24.10.22; peer-reviewed by D McElroy, N Haff; comments to author 24.12.22; revised version received 01.03.23; accepted 17.11.23; published 01.01.24.

Please cite as:
Bazzano AN, Patel T, Nauman E, Cernigliaro D, Shi L
Optimizing Telehealth for Diabetes Management in the Deep South of the United States: Qualitative Study of Barriers and Facilitators on the Patient and Clinician Journey
J Med Internet Res 2024;26:e43583
URL: https://www.jmir.org/2024/1/e43583
doi:10.2196/43583
PMID:37976468

©Alessandra N Bazzano, Tejal Patel, Elizabeth Nauman, Dana Cernigliaro, Lizheng Shi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Key Factors Influencing the Operationalization and Effectiveness of Telemedicine Services in Henan Province, China: Cross-Sectional Analysis

Xianying He[1*], MM; Fangfang Cui[1*], MM; Minzhao Lyu[2], PhD; Dongxu Sun[1], MSc; Xu Zhang[1], MSc; Jinming Shi[1], BSc; Yinglan Zhang[3], BM; Shuai Jiang[4], PhD; Jie Zhao[1], PhD

[1]National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

[2]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

[3]College of Public Health, Zhengzhou University, Zhengzhou, China

[4]Finance Department, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

[*]these authors contributed equally

Corresponding Author:
Jie Zhao, PhD
National Engineering Laboratory for Internet Medical Systems and Applications
The First Affiliated Hospital of Zhengzhou University
1 Jianshe Road
Zhengzhou, 450052
China
Phone: 86 037167966286
Fax: 86 037167966286
Email: zhaojie@zzu.edu.cn

Abstract

Background: Telemedicine has demonstrated its potential in alleviating the unbalanced distribution of medical resources across different regions. Henan, a province in China with a population of approximately 100 million, is especially affected by a health care divide. The province has taken a proactive step by establishing a regional collaborative platform for telemedicine services provided by top-tier provincial hospitals.

Objective: We aim to identify the key factors that influence the current operationalization and effectiveness of telemedicine services in Henan province. The insights gained from this study will serve as valuable references for enhancing the efficient operation of telemedicine platforms in low- and middle-income regions.

Methods: We analyzed service reports from the performance management system of telemedicine services in Henan province throughout 2020. Using descriptive statistics and graphical methods, we examined key influencing factors, such as management competency; device configuration; and hospital capability, capacity, and service efficacy, across hospitals at 2 different tiers. In addition, we used generalized linear models and multiple linear regression models to identify key operational factors that significantly affect the service volume and efficacy of 2 major telemedicine services, namely teleconsultation and tele-education.

Results: Among the 89 tier 3 hospitals and 97 tier 2 hospitals connected to the collaborative telemedicine platform, 65 (73%) and 55 (57%), respectively, have established standardized management procedures for telemedicine services. As the primary delivery method for telemedicine services, 90% (80/89) of the tier 3 hospitals and 94% (91/97) of the tier 2 hospitals host videoconferencing consultations through professional hardware terminals rather than generic computers. Teleconsultation is the dominant service type, with an average annual service volume per institution of 173 (IQR 37-372) and 60 (IQR 14-271) teleconsultations for tier 3 and tier 2 hospitals, respectively. Key factors influencing the service volume at each hospital include available funding, management competency, the number of connected upper tiers, and the number of professional staff. After receiving teleconsultations from tier 3 (65/89, 73%) and tier 2 (61/97, 63%) hospitals, patients reported significant improvements in their medical conditions. In addition, we observed that service efficacy is positively influenced by management competency, financial incentives, the number of connected upper or lower tiers, and the involvement of participating medical professionals.

Conclusions: Telemedicine has become increasingly popular in Henan province, with a notable focus on teleconsultation and tele-education services. Despite its popularity, many medical institutions, especially tier 2 hospitals, face challenges related to

XSL•FO
RenderX

management competency. In addition to enhancing the effectiveness of existing telemedicine services, health care decision-makers in Henan province and other low- and middle-income regions should consider expanding the service categories, such as including remote emergency care and telesurgery, which have promise in addressing crucial health care needs in these regions.

(J Med Internet Res 2024;26:e45020) doi:10.2196/45020

KEYWORDS

telemedicine; service statistics; efficiency; quality management

Introduction

Background

To alleviate the unbalanced distribution of medical resources between metropolitan cities and rural or remote regions, telemedicine has emerged as a viable approach to connect experts at top-tier hospitals with underprivileged medical institutions and remote patients [1,2]. In the early stages, telemedicine services were primarily delivered via telephone and email for simple consultations [3], which proved ineffective in dealing with complex patient conditions. With the use of modern IT and telecommunication technologies, telemedicine services can now be delivered through high-speed networks, enabling high-resolution video, audio, and medical record sharing in real time for better efficiency [4,5]. Recent technological advances such as wearable technologies, telemedicine appliances, and 5G networks have further expanded the possibilities of new service genres in telemedicine. These include telemonitoring, remote ward rounds, telediagnosis, telepathology, remote emergency care, and tele-education [6-8].

China currently categorizes its medical institutions into 3 tiers. In general, tier 3 hospitals possess the most medical resources and are often located in metropolitan cities; tier 2 hospitals have fewer medical capabilities than tier 3 hospitals; and the tier 1 medical centers are only equipped for basic consultations, such as general practitioner services. Telemedicine significantly benefits patients at tier 2 and tier 1 hospitals from underprivileged regions by connecting them with experts at tier 3 hospitals. In addition, medical practitioners at lower-tier hospitals can receive training from experts at the higher-tier hospitals through regional telemedicine platforms; for instance, the provincial telemedicine platform in Henan province, as studied in this paper, linked 97.1% (99/102) of tier 3 hospitals and 89.6% (326/364) of tier 2 hospitals in the province. We note that 89 (87.3%) of the 102 tier 3 hospitals and 97 (26.6%) of the 364 tier 2 hospitals were covered in this study because sufficient data were available only for these hospitals. Such interhospital collaboration through telemedicine services plays a vital role not only in easing the unbalanced distribution of medical resources but also in making professional treatment for patients in remote regions more affordable [9,10].

Given the immense potential and benefits of telemedicine, many countries, particularly high-income ones, have embraced its practice. Telemedicine services in the United States have seen rapid development since the 1960s when NASA used a telemonitoring system to track the physiological indexes of astronauts on missions [11]. This pioneering effort paved the way for the emergence of telediagnosis and teleconsultation services in the country [12,13]. By 2013, 40% to 50% of hospitals in the United States provided telemedicine services. Likewise, Europe and countries such as Australia, Canada, and the United Kingdom have made significant strides to promote telehealth or telemedicine, primarily to serve their citizens in rural and remote areas [14-16]. Telemedicine now encompasses a broad range of services, including teleconsultation, telepathology, telediagnosis, remote emergency care, remote surgery, and tele-education, spanning across various medical disciplines [17,18].

The development of telemedicine in China commenced relatively later, following a similar pattern to many other low- and middle-income countries. Remote consultations in China commenced in the 1990s. In 1997, Shanghai Zhongshan Hospital and Anhui Fuyang People's Hospital pioneered the launch of remote consultations [19]. Since the late 20th century, regional collaborative telemedicine platforms have been established in many provinces of China, connecting premium medical resources from top-tier hospitals to underprivileged ones [20-22]. As of 2021, in China, >29 provinces have set up regional platforms, covering >24,000 medical institutions to provide telemedicine services [23-26]. As a specific example, the collaborative telemedicine platform studied in this paper currently serves 425 tier 3 and tier 2 hospitals across Henan province. Its aim is to address the shortage of high-quality medical resources in underprivileged regions.

Objectives

To the best of our knowledge, prior research articles have examined the effectiveness of telemedicine services delivered through regional collaborative platforms with regard to their service delivery processes [27,28]. In addition, some studies have focused on specific service types such as telepathology [29] and teleconsultation [23,26]. However, limited information is available concerning the operational factors crucial to ensuring the effective delivery of telemedicine services. Moreover, Henan province serves as a typical representative of low- and middle-income regions that actively promote telemedicine. Therefore, this paper presents a representative case study of the collaborative telemedicine platform in Henan province. The objective is to understand how various operational factors, including management competency, device configurations, institution capability, and the status of connected hospitals, affect the effectiveness of telemedicine services. The insights gained from this study offer valuable guidance to health care decision makers for the effective operation of regional collaborative telemedicine platforms not only in China but also in other low- and middle-income regions striving to bridge their health care divide through telemedicine.

XSL•FO
RenderX

Methods

Study Design

To understand the influence of different operational factors on the effectiveness of telemedicine services in low- and middle-income regions, this study centered on the collaborative telemedicine platform in Henan province for a representative case study. The platform, initiated by the provincial health department, is responsible for overseeing the operation of telemedicine services in the province. It directly connects hospitals in 18 cities of Henan province through web-based private networks.

To achieve our study objective, we designed a questionnaire to be completed by administrative and managerial personnel at all connected tier 3 and tier 2 hospitals. The questionnaire aimed to gather information on the operationalization and effectiveness of each telemedicine service from management personnel at the studied hospitals. The questions underwent a comprehensive review by technical engineers and medical professionals from 10 (2.4%) of the 425 connected hospitals and were revised based on the feedback provided before distribution. The questionnaire comprises 47 questions categorized into 5 general aspects, including management competency, equipment configuration, service capability, service efficacy, and development potentials. Each aspect is further delineated by a set of detailed factors with specific measures (Multimedia Appendix 1).

The questionnaire also underwent reliability testing, resulting in a Cronbach α coefficient of .875. The validity testing revealed a Kaiser-Meyer-Olkin coefficient of 0.766 and a significant Bartlett sphericity test ($P<.001$). Overall, the designed questionnaire demonstrated decent reliability and validity.

Data Collection

Our survey questions were integrated into the collaborative telemedicine platform as a performance management system, which collected feedback (ie, answers to our designed questions) from the corresponding administrative and managerial personnel at the hosting hospitals. Two example screenshots of the performance management system are shown in Figure 1, including the log-in page and a snippet of questions soliciting feedback. The validity of the feedback was assessed by the hosting hospitals before it was uploaded to the performance management system. Professional staff operating the system also gave regular callbacks to randomly selected respondents to verify the reliability of the feedback. The performance management system further aggregated the collected feedback per hospital, region, and year and generated evaluation reports for the respective management personnel to review. This paper covers the feedback collected by the system during the whole of 2020 (ie, from January 1, 2020, to December 31, 2020).

Figure 1. Screenshots of the performance management system of telemedicine services.

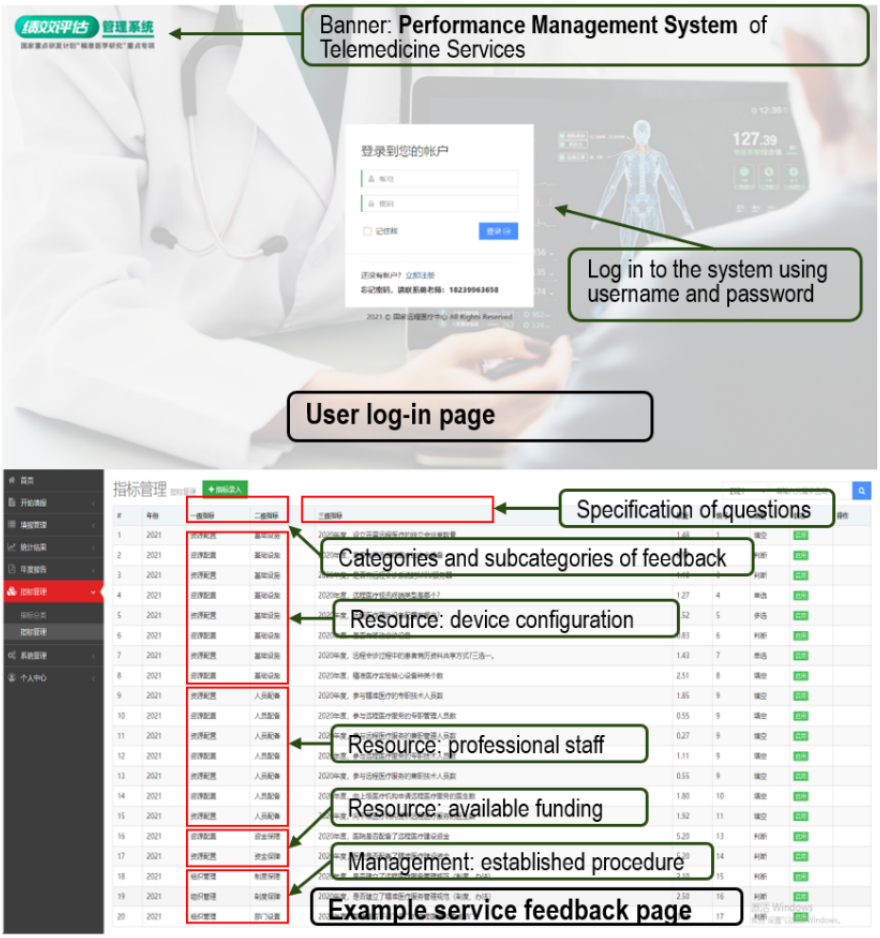

XSL•FO
RenderX

Ethical Considerations

We obtained ethics approval for this study from the host institution, the First Affiliated Hospital of Zhengzhou University (2021-KY-1092-001). The ethics approval allows the researchers to collect operational information of telemedicine services through questionnaires administered to administrative and management personnel at the studied hospitals connected to the Henan provincial telemedicine platform. All participant personnel were formally informed about the use of the collected questionnaires and the objective of this study. Ethics approval was granted on the condition that the collected data would not reveal respondents' identities and that only aggregated information (such as at hospital level) would be discussed in the research manuscript.

Statistical Analysis

Completed questionnaires were collected from 189 participant hospitals in Henan province; after data preprocessing, which involved data cleansing and the removal of missing and invalid data, we obtained valid survey responses from 186 (98.4%) hospitals that provided telemedicine services during 2020. Of these 186 hospitals, 89 (47.8%) were tier 3 hospitals, and 97 (52.2%) were tier 2 hospitals. In our analysis, we used median values to represent statistical distributions and absolute counts with percentage values to describe counting data. Bar charts were plotted for the volume of different service types. We performed a chi-square test and a rank sum test using SPSS (version 23.0; IBM Corp) to compare the management competency, device configuration, and service capability of tier 2 and tier 3 hospitals. Furthermore, we used a generalized linear model that took the aforementioned operational factors (eg, management competency and device configuration) as independent variables to analyze their impacts on the service volume of various types (eg, teleconsultation and tele-education). A multiple linear regression model was used to analyze the influence of the identified operational factors on service efficacy with a significance level of .05 (ie, α=.05). We note that wrong and incomplete data were treated as missing values.

Results

Management Competency

Among the 186 studied hospitals that provided telemedicine services in 2020 via the collaborative platform in Henan province, 58% (52/89) of the tier 3 hospitals and 45% (44/97) of the tier 2 hospitals have dedicated departments managing telemedicine services. Moreover, in terms of professional staff, on average, tier 3 and tier 2 hospitals have 2.8 and 2.1 persons, respectively. In addition, 73% (65/89) of the tier 3 hospitals and 57% (55/97) of the tier 2 hospitals have established a standard management procedure for telemedicine services to be followed, meaning that the difference in this regard between tier 3 hospitals and tier 2 hospitals is statistically significant (χ^2_1=5.4; *P*=.02). To encourage medical practitioners to actively participate in telemedicine services as part of their daily duty, 54% (48/89) of the tier 3 hospitals offer financial incentives (eg, an additional bonus), whereas the corresponding figure for tier 2 hospitals is much lower (30/97, 31%; χ^2_1=5.4; *P*<.001). A telemedicine service is often requested by a lower-tier hospital to access the medical resources (eg, experts) available at an upper-tier hospital; therefore, a reasonable amount of revenue sharing between the 2 parties also contributes to the successful collaboration in a telemedicine service. However, we found that only 16% (14/89) of the tier 3 hospitals and 11% (11/97) of the tier 2 hospitals had revenue-sharing arrangements with other participating institutions, whereas the difference between tier 3 and tier 2 hospitals in this regard was not statistically significant (Table 1).

Table 1. Management competency regarding telemedicine services at tier 3 and tier 2 hospitals in Henan province in 2020.

Studied factors	Tier 3 hospitals (n=89), n (%)	Tier 2 hospitals (n=97), n (%)	Chi-square (*df*)	*P* value
Has dedicated department	52 (58)	44 (45)	3.2 (1)	.08
Has established management procedure	65 (73)	55 (57)	5.4 (1)	.02
Has financial incentives for participating medical practitioners	48 (54)	30 (31)	10.1 (1)	.001
Has revenue-sharing arrangements with participating hospitals	14 (16)	11 (11)	0.8 (1)	.38

Configuration of Telemedicine Devices

The configuration of telemedicine devices is of vital importance in the effective delivery of service. The multipoint control unit (MCU) server is 1 of the critical devices that manage videoconferencing processes among participants in a telemedicine service. As revealed in our analysis, 48 (54%) of the 89 tier 3 hospitals have configured their MCU servers, approximately twice the number of tier 2 hospitals (24/97, 25%), with strong statistical significance (χ^2_1=16.7; *P*<.001). In addition to the MCU server, videoconferencing terminals that directly record and stream video and audio from participants are also deterministic influencing factors. Professional hardware terminals for videoconferencing are the optimal choice because they provide well-optimized performance and robustness. We observed that 90% (80/89) of the tier 3 hospitals and 94% (91/97) of the tier 2 hospitals had configured professional hardware terminals for telemedicine services, whereas the rest (tier 3 hospitals: 9/89, 10%; tier 2 hospitals: 6/97, 6%) used generic conferencing software (eg, Zoom and Tencent Meeting) on computers. Although the higher deployment rate of hardware terminals at tier 2 hospitals seems counterintuitive, after interviewing some administrative officers at these hospitals, we believe that tier 2 hospitals have more motivation and willingness with regard to purchasing such high-quality terminals because they are less complex to configure than MCUs

XSL•FO
RenderX

and can directly benefit their patients who seek medical advice from experts at tier 3 hospitals (as well as create a better impression). As for the configuration of other professional telemedicine devices, 58% (52/89) of the tier 3 hospitals and 45% (44/97) of the tier 2 hospitals were equipped with digital pathology slide scanners, and 35% (31/89) of the tier 3 hospitals and 25% (24/97) of the tier 2 hospitals had professional recording equipment for tele-education. Fairly similar insights (ie, the device configuration status at tier 3 hospitals is better than that at tier 2 hospitals) can be obtained for other types of devices as listed in Table 2, all with statistical significance (P<.05).

Table 2. Configuration of telemedicine devices at tier 3 and tier 2 hospitals in Henan province in 2020.

Type of device	Tier 3 hospitals (n=89), n (%)	Tier 2 hospitals (n=97), n (%)	Chi-square (*df*)	*P* value
MCU[a] server	48 (54)	24 (25)	16.7 (1)	<.001
Professional videoconferencing hardware	80 (90)	91 (94)	0.1 (1)	.33
Videoconferencing software on generic computer	9 (10)	6 (6)	0.1 (1)	.33
Digital pathology slide scanner	52 (58)	44 (45)	3.2 (1)	.08
Tele-ECG[b] monitoring systems	57 (64)	40 (41)	9.7 (1)	.002
Multisource surgical information collection terminal	15 (17)	2 (2)	12.2 (1)	.001
Recording equipment	31 (35)	24 (25)	2.3 (1)	.13
Remote mobile ward-round vehicle	16 (18)	3 (3)	11.2 (1)	.001
Professional teleconsultation equipment	31 (35)	10 (10)	16.2 (1)	.001
Network security appliances	64 (72)	54 (56)	5.3 (1)	.02

[a]MCU: multipoint control unit.

[b]Tele-ECG: tele-electrocardiogram.

Capability and Capacity of Hospitals

Now, we look at the capability of hospitals with regard to offering each type of telemedicine service. Among the tier 3 hospitals covered in this study, 94% (84/89), 75% (68/89), and 75% (68/89) offered teleconsultation, tele-education, and tele-electrocardiogram (tele-ECG), respectively. Among the tier 2 hospitals, 75% (73/97), 63% (61/97), and 53% (51/97) offered teleconsultation, tele-education, and telepathology, respectively. The capabilities of tier 2 and tier 3 hospitals with regard to other service types are depicted in Multimedia Appendix 2.

As shown in Table 3, with regard to the number of connected medical institutions for tier 3 and tier 2 hospitals, the median values for the connected upper-tier hospitals and lower-tier hospitals are 3 (IQR 2-9) and 16 (IQR 5-27), respectively, per tier 3 hospital and 2 (IQR 1-4) and 4 (IQR 0-15), respectively, per tier 2 hospital. Tier 3 hospitals have more connected low tier hospitals than tier 2 hospitals (H=−4.80; P<.001). As for the average service volume per hospital, that is, the number of services that have been delivered on average by the hospital, tier 3 hospitals have delivered 173 teleconsultation services, 4281 tele-ECG services, and 542 telediagnosis services, whereas tier 2 hospitals have delivered 60 teleconsultation services, 1065 tele-ECG services, and 51 telediagnosis services. Overall, the volume of the aforementioned 3 service types at tier 3 hospitals is significantly higher than that at tier 2 hospitals (P<.05). The average service volumes of telepathology at tier 3 and tier 2 hospitals are 120 and 114, respectively, with no statistical difference. Tele-education and remote emergency care have low volumes in both tiers (Table 4).

We highlight that the service volumes for each type of telemedicine service, which serve as the dependent variables, exhibit strong correlations with several influencing factors (ie, independent variables), including management competency, device configuration, and engagement of staff or medical practitioners. In this study, we further investigated how these factors affect the 2 most popular types of telemedicine services (ie, teleconsultation and tele-education). To be more specific, we took the service volumes of teleconsultation and tele-education as 2 dependent variables that are influenced by independent variables, including the tier to which a hospital belongs, available funding, and dedicated management department. A full list of independent variables is shown in Table 4. We constructed a generalized linear model using the variables defined herein. Our model was evaluated as a good fit by the omnibus test (P<.001).

Now, we highlight the factors that significantly affect the service volumes of teleconsultation and tele-education (Tables 4 and 5). First, having dedicated management departments and professional staff positively affects the service volume of teleconsultation (β=.70; P=.04 and β=.19; P=.002, respectively). Second, both the number of connected upper-tier hospitals and configured MCU servers positively influence the service volumes of both teleconsultation and tele-education (β=.04; P=.04 and β=.96; P=.005, respectively). Third, as shown in Table 5, the available funding, management procedure, and the number of participating medical practitioners are key influencing factors for the service volume of tele-education.

XSL•FO
RenderX

Table 3. Number of connected upper-tier and lower-tier hospitals and service volumes at tier 3 and tier 2 hospitals in Henan province in 2020.

Studied factors	Tier 3 hospitals, median (IQR)	Tier 2 hospitals, median (IQR)	*Z*	*P* value
Upper-tier hospitals, n	3 (2-9)	2 (1-4)	−1.88	.06
Lower-tier hospitals, n	16 (5-27)	4 (0-15)	−4.80	<.001
Service volume of teleconsultation, n	173 (37-372)	60 (14-271)	−2.70	.007
Service volume of tele-ECG[a], n	4281 (294-24,812)	1065 (50-5103)	−2.16	.03
Service volume of telediagnosis, n	542 (62-9245)	51 (19-957)	−2.66	.008
Service volume of telepathology, n	120 (58-252)	114 (49-250)	0.43	.67
Service volume of tele-education, n	28 (10-60)	29 (9-50)	0.53	.59
Service volume of remote emergency care, n	5 (1-10)	18 (2-23)	975.00	.35

[a]Tele-ECG: tele-electrocardiogram.

Table 4. Analysis of key influencing factors for the service volume of teleconsultation using a generalized linear model.

Variable	β	SE	95% CI	Wald test	*P* value
Constant	5.00	.49	4.03 to 5.97	101.98	.001
Tier to which hospital belongs: tier 3 hospital (reference=tier 2 hospital)	−.15	.28	−.70 to .39	0.31	.58
Available funding (reference=no)	−1.12	.28	−1.68 to .57	15.62	<.001
Dedicated department (reference=no)	.70	.35	.02 to 1.39	4.12	.04
Management procedure (reference=no)	−.30	.37	−1.02 to .42	0.67	.41
Financial incentive (reference=no)	.22	.28	−.34 to .77	0.59	.44
Revenue-sharing arrangements (reference=no)	−.50	.36	−1.20 to .19	2.00	.16
Connected upper-tier hospitals	.04	.02	.01 to .07	4.05	.04
Connected lower-tier hospitals	.001	.000	.000 to .003	2.16	.14
Professional staff	.19	.06	.07 to .31	9.63	.002
MCU[a] server (reference=no)	.96	.34	.29 to 1.64	7.82	.005
Type of videoconferencing terminals: professional hardware (reference=generic PC)	.18	.47	−.74 to 1.11	0.15	.70

[a]MCU: multipoint control unit.

Table 5. Analysis of key influencing factors for the service volume of tele-education using a generalized linear model.

Variable	β	SE	95% CI	Wald test	*P* value
Constant	3.15	.24	2.68 to 3.62	172.29	<.001
Tier to which hospital belongs: tier 3 hospital (reference=tier 2 hospital)	−.20	.23	−.64 to 2.42	0.79	.37
Available funding (reference=no)	−.81	.22	−1.25 to −.38	13.44	<.001
Dedicated department (reference=no)	.20	.24	−.26 to .66	0.72	.40
Management procedure (reference=no)	1.26	.23	.80 to 1.72	28.85	<.001
Financial incentive (reference=no)	.009	.21	−.390 to .410	0.00	.97
Revenue-sharing arrangements (reference=no)	.31	.30	−.28 to .91	1.08	.30
Connected upper-tier hospitals	−.02	.02	−.06 to .01	1.71	.19
Connected lower-tier hospitals	.000	.00	.000 to .001	0.42	.52
Professional staff	.07	.05	−.02 to .17	2.30	.13
Participating medical practitioners	.004	.00	.001 to .008	6.98	.008
Recording device (reference=no)	−.28	.21	−.69 to .14	1.70	.19

XSL•FO
RenderX

Efficacy of Telemedicine Services

Administrative and managerial personnel at the hosting hospitals were encouraged to take surveys on the efficacy of each telemedicine service they provide. To measure service efficacy, we took into account the number of patients whose medical conditions improved after receiving telemedicine services. Our results showed that the service efficacy at tier 3 hospitals was 73%, whereas that at tier 2 hospitals was 63%. We constructed a multiple linear regression model to identify the operational factors (ie, independent variables as listed in the first column of Table 6) that could have significant impacts on service efficacy (ie, the dependent variable). The model was evaluated as a good fit by the omnibus test (*P*=.006).

We now discuss the factors that have a significant impact on service efficacy. As shown in the first column of Table 6, proper management procedures, established financial incentives, the number of connected upper-tier and lower-tier hospitals, and the number of participating medical practitioners are important influencing factors for the efficacy of teleconsultation services. With a proper management procedure (β=.15; *P*=.04) and financial incentives for participating staff (β=.14; *P*=.03), the efficacy could be significantly improved. However, having a higher number of participating medical practitioners may not always lead to better teleconsultation efficacy (β=−.001; *P*=.03), possibly owing to potential conflicts related to the treatment plans proposed by different participating physicians.

Table 6. Analysis of key influencing factors for the service efficacy of teleconsultation using a multiple linear regression model.

Variable	β	SE	95% CI	Wald test	*P* value
Constant	.46	.07	.34 to .59	49.512	<.001
Tier to which hospital belongs: tier 3 hospital (reference=tier 2 hospital)	.09	.06	−.03 to .21	2.363	.12
Available funding (reference=no)	.01	.07	−.12 to .15	0.037	.85
Dedicated department (reference=no)	−.10	.08	−.24 to .05	1.625	.20
Management procedure (reference=no)	.15	.07	.01 to .28	4.405	.04
Financial incentive (reference=no)	.14	.06	.02 to .26	4.825	.03
Revenue-sharing arrangements (reference=no)	−.11	.08	−.27 to .06	1.653	.20
Connected upper-tier hospitals	.010	.01	.004 to .024	6.952	.008
Connected lower-tier hospitals	−.000	.00	−.001 to .000	4.770	.03
Professional staff	−.03	−.01	.06 to .00	3.677	.06
Participating medical practitioners	−.001	.00	−.002 to .000	5.005	.03
Use of emerging IT technologies (reference=no)	.05	.05	−.09 to .19	0.514	.47

Future Development

To estimate the future development of telemedicine services in Henan province, we studied the willingness of development and the level of financial investment at each hospital. We found that 58% (52/89) of the tier 3 hospitals and 49% (48/97) of the tier 2 hospitals had sufficient funding dedicated for the future development of telemedicine services; 48% (43/89) of the tier 3 hospitals and 38% (37/97) of the tier 2 hospitals have a detailed strategic plan; and 30% (27/89) of the tier 3 hospitals and 20% (19/97) of the tier 2 hospitals have started their exploration journey with emerging technologies such as 5G, virtual reality, and wearable devices in telemedicine services.

Discussion

Principal Findings

Given the important role of telemedicine in easing the unbalanced distribution of medical resources, it has been widely deployed in many low- and middle-income countries and regions, including China [28,30,31]. Henan province, as a representative of low- and middle-income regions in China, stands among the earliest provinces to establish regional collaborative telemedicine platforms, effectively connecting premium hospitals in metropolitan cities with underprivileged ones in remote and rural regions. In this study, we focus on analyzing service feedback from the provincial collaborative telemedicine platform. Through this analysis, we aim to identify key operational factors that significantly influence the effectiveness of telemedicine services provided by tier 3 and tier 2 hospitals. Our results serve as valuable references for future development and service improvement, particularly in low- and middle-income regions.

Inadequate Managerial Capacity

Management incompetency is a prominent problem at many hospitals covered in this study: 42% (37/89) of the tier 3 hospitals and 54% (52/97) of the tier 2 hospitals did not have a dedicated management department for telemedicine services. Furthermore, the absence of proper management procedures is also a common problem. The average professional staff per hospital is <3 persons, which is particularly true at tier 2 hospitals. Our obtained insights highlight that although telemedicine has reached a high speed of deployment in China, many hospitals do not have a dedicated management and professional team for telemedicine services [21]. In addition, at lower-tier hospitals, technical staff with insufficient training may not be able to promptly handle emergency issues with their telemedicine devices and networks, negatively affecting service

XSL•FO
RenderX

effectiveness [32]. Our results indicate that an established management procedure and the number of professional staff dedicated to telemedicine services at a hospital are critical factors that influence service volume and efficacy. Therefore, in low- and middle-income regions such as Henan province, seeking improvements in management competency and the training of professional staff are of high priority for increasing service effectiveness.

Absence of Revenue-Sharing Arrangements

Revenue from the telemedicine sector is an important financial motivation for hospitals to deliver high-quality services to patients in remote and rural regions [33]. In the United States, 1 virtual clinical encounter could cost approximately US $120 [34]. In China, the cost of 1 virtual clinical encounter usually ranges from CNY 200 (US $27.9) to CNY 1000 (US $139.7) [22]. The revenue of each service is expected to be fairly shared by the hospitals that host patients and those that provide remote services. However, we found that a proper revenue-sharing arrangement has not been set up by most of the studied hospitals, which inevitably hinders interhospital collaborations. Therefore, establishing fair revenue-sharing arrangements among participating institutions is of vital importance in the operationalization of telemedicine services.

Incapable Telemedicine Devices

Professional hardware videoconferencing terminals served by MCU servers are capable of providing real-time video and audio for telemedicine services with high resolution, security, reliability, and robustness, especially when serving multiple participating hospitals [35]. However, the high cost of such hardware combinations slows down their adoption at many hospitals [36-38]. A study conducted in 2018 showed that 61.3% of the tier 3 hospitals in China were equipped with hardware videoconferencing terminals for telemedicine services [21]. In this study, we found that 90% (80/89) of the tier 3 hospitals and 94% (91/97) of the tier 2 hospitals in Henan province have such hardware terminals. In addition, to enhance network security during the provision of telemedicine services, 72% (64/89) of the tier 3 hospitals and 56% (54/97) of the tier 2 hospitals have configured security appliances such as firewalls and intrusion detection systems. We suggest that hospitals with sufficient funding should consider making professional hardware videoconferencing terminals and network security appliances mandatory to guarantee the quality and reliability of their telemedicine services.

Insufficient Coverage of Service Types

Our study shows that the popular telemedicine service types are teleconsultation, tele-education, tele-ECG, telepathology, and telediagnosis. Among them, tele-ECG has the most service volume because many lower-tier hospitals have the screening capability but not the capability to provide a diagnosis. Therefore, they prefer to request a tele-ECG service with upper-tier hospitals for a better diagnosis of patient conditions. Similar reasons are applied to telediagnosis and telepathology [29]. Other service types such as remote emergency care and remote surgery have low volumes currently, and their supporting devices have a low deployment rate at the respective hospitals. Given the fact that telemedicine services in remote surgery, remote ward round, and intensive care units are becoming mature [8,39,40], it is suggested that hospitals with sufficient funding could start trials with the emerging service types.

Limitations

First, this paper focused on the critical factors for the operationalization of telemedicine services and their impacts on service effectiveness only at tier 3 (n=89) and tier 2 (n=97) hospitals in China's Henan province; tier 1 medical institutions, which mostly serve rural and remote regions, have not been covered in this study owing to their currently low involvement in telemedicine. Second, owing to the limitation of a survey-based study, we are only able to cover the general aspect of management competency without examining detailed breakdowns, which may require rigorous interviews with the respective personnel and thus is left for future research. Third, our results primarily concern the major service types (eg, teleconsultation, tele-ECG, and tele-education), whereas the others (eg, remote surgery and remote emergency care) require sufficient data to be able to draw a trustable conclusion. In addition, this study does not cover how various factors affect patient satisfaction, which could be a valuable direction for future research.

Conclusions

As a representative of low- and middle-income regions in China, Henan province has taken steps to develop its provincial collaborative telemedicine platform to encourage the sharing of medical resources among hospitals through various types of telemedicine services. Our study focused on the data collected from the platform throughout 2020, and it revealed that a majority of tier 2 and tier 3 hospitals offering telemedicine services are now equipped with professional devices for teleconsultation, telediagnosis, tele-ECG, and tele-education. Among all service types, tele-ECG stands out as the dominant service in terms of volume. However, our findings also highlighted some prominent issues. Among the main concerns are management incompetency, characterized by the absence of a dedicated management department; the lack of financial incentives for medical practitioners; and improper revenue sharing among participating hospitals. In addition, the current service capability is relatively limited, with only a small set of service types available, including teleconsultation, tele-ECG, telediagnosis, and tele-education. Hence, an important future direction is to expand the service categories at most hospitals, incorporating services such as remote surgery guidance and remote emergency care.

Acknowledgments

This research was supported by grants from the National Key R&D Program of China (2022ZD0160704), the Key Science and Technology Program in Henan province (201400210400), the Science and Technology Research Program in Henan province

XSL•FO
RenderX

(232102311057), the Key Research of Higher Education in Henan province (22A330004 and 23B330004), and the Joint Co-construction Program of Medical Science and Technology of Henan province (LHGJ20210360 and LHGJ20200349). The funding bodies played no role in the design, data collection, or interpretation of this work. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the funding bodies.

Data Availability

The data used in this research are available upon individual request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of information collected in this study.
[DOCX File , 17 KB - jmir_v26i1e45020_app1.docx]

Multimedia Appendix 2
Types of telemedicine services delivered by tier 3 and tier 2 hospitals in Henan province and their volume in percentage per hospital tier in 2020.
[PNG File , 10 KB - jmir_v26i1e45020_app2.png]

References

1. Waller M, Stotler C. Telemedicine: a primer. Curr Allergy Asthma Rep 2018 Aug 25;18(10):54. [doi: 10.1007/s11882-018-0808-4] [Medline: 30145709]
2. Combi C, Pozzani G, Pozzi G. Telemedicine for developing countries. a survey and some design issues. Appl Clin Inform 2016 Nov 02;7(4):1025-1050 [FREE Full text] [doi: 10.4338/ACI-2016-06-R-0089] [Medline: 27803948]
3. Thrall JH, Boland G. Telemedicine in practice. Semin Nucl Med 1998 Apr;28(2):145-157. [doi: 10.1016/s0001-2998(98)80004-4] [Medline: 9579416]
4. Zheng J, Wang Y, Zhang J, Guo W, Yang X, Luo L, et al. 5G ultra-remote robot-assisted laparoscopic surgery in China. Surg Endosc 2020 Jul 22;34(11):5172-5180. [doi: 10.1007/s00464-020-07823-x]
5. Mat Kiah ML, Al-Bakri SH, Zaidan AA, Zaidan BB, Hussain M. Design and develop a video conferencing framework for real-time telemedicine applications using secure group-based communication architecture. J Med Syst 2014 Oct;38(10):133. [doi: 10.1007/s10916-014-0133-y] [Medline: 25199651]
6. Tian W. [Exploration and prospect of 5G application in telemedicine]. Zhonghua Wai Ke Za Zhi 2020 Jan 01;58(1):1-4. [doi: 10.3760/cma.j.issn.0529-5815.2020.01.001] [Medline: 31902160]
7. Lin TW, Hsu CL, Le T, Lu CF, Huang BY. A smartcard-based user-controlled single sign-on for privacy preservation in 5G-IoT telemedicine systems. Sensors (Basel) 2021 Apr 20;21(8):2880 [FREE Full text] [doi: 10.3390/s21082880] [Medline: 33924024]
8. Majumder S, Mondal T, Deen MJ. Wearable sensors for remote health monitoring. Sensors (Basel) 2017 Jan 12;17(1):130 [FREE Full text] [doi: 10.3390/s17010130] [Medline: 28085085]
9. Speyer R, Denman D, Wilkes-Gillan S, Chen Y, Bogaardt H, Kim J, et al. Effects of telehealth by allied health professionals and nurses in rural and remote areas: a systematic review and meta-analysis. J Rehabil Med 2018 Feb 28;50(3):225-235 [FREE Full text] [doi: 10.2340/16501977-2297] [Medline: 29257195]
10. Akiyama M, Yoo B. A systematic review of the economic evaluation of telemedicine in Japan. J Prev Med Public Health 2016 Jul;49(4):183-196 [FREE Full text] [doi: 10.3961/jpmph.16.043] [Medline: 27499161]
11. Fuchs M. Provider attitudes toward STARPAHC: a telemedicine project on the Papago reservation. Med Care 1979 Jan;17(1):59-68. [doi: 10.1097/00005650-197901000-00005] [Medline: 759743]
12. Chaet D, Clearfield R, Sabin JE, Skimming K, Council on Ethical and Judicial Affairs American Medical Association. Ethical practice in telehealth and telemedicine. J Gen Intern Med 2017 Oct;32(10):1136-1140 [FREE Full text] [doi: 10.1007/s11606-017-4082-2] [Medline: 28653233]
13. Hatcher-Martin JM, Adams JL, Anderson ER, Bove R, Burrus TM, Chehrenama M, et al. Telemedicine in neurology: telemedicine work group of the American academy of neurology update. Neurology 2020 Jan 07;94(1):30-38. [doi: 10.1212/WNL.0000000000008708] [Medline: 31801829]
14. Almufleh A, Lee C, Tsang MY, Gin K, Tsang TS, Nair P. The need for telemedicine integration into adult cardiology training curricula in Canada. Can J Cardiol 2021 Jun;37(6):929-932. [doi: 10.1016/j.cjca.2021.03.001] [Medline: 33992489]
15. Agarwal P, Kithulegoda N, Umpierre R, Pawlovich J, Pfeil JN, D'Avila OP, et al. Telemedicine in the driver's seat: new role for primary care access in Brazil and Canada: The Besrour Papers: a series on the state of family medicine in Canada and Brazil. Can Fam Physician 2020 Feb;66(2):104-111 [FREE Full text] [Medline: 32060190]

XSL•FO
RenderX

16. Owens B. Telemedicine on the rise but lagging in Canada. CMAJ 2018 Sep 24;190(38):E1149-E1150 [FREE Full text] [doi: 10.1503/cmaj.109-5634] [Medline: 30249766]
17. Mussetti A, Peric Z, Figueroa C. COVID19 in hematological patients and telemedicine: lessons learned across Europe and the US. Curr Opin Infect Dis 2022 Aug 01;35(4):295-301. [doi: 10.1097/QCO.0000000000000843] [Medline: 35849519]
18. Kristensen MB, Høiberg L, Nøhr C. Updated mapping of telemedicine projects in Denmark. Stud Health Technol Inform 2019;257:223-228. [Medline: 30741200]
19. Zheng Y, Lin Y, Cui Y. Teledermatology in China: history, current status, and the next step. J Investig Dermatol Symp Proc 2018 Dec;19(2):S71-S73 [FREE Full text] [doi: 10.1016/j.jisp.2018.09.003] [Medline: 30471754]
20. Wang Z, Gu H. A review of telemedicine in China. J Telemed Telecare 2009;15(1):23-27. [doi: 10.1258/jtt.2008.080508] [Medline: 19139216]
21. Cui F, Ma Q, He X, Zhai Y, Zhao J, Chen B, et al. Implementation and application of telemedicine in China: cross-sectional study. JMIR Mhealth Uhealth 2020 Oct 23;8(10):e18426 [FREE Full text] [doi: 10.2196/18426] [Medline: 33095175]
22. Zhang W, He D, Wang G, Zhu C, Evans R. Analyzing national telemedicine policies in China from the perspective of policy instrument (1997-2020). Int J Med Inform 2022 Oct;166:104854. [doi: 10.1016/j.ijmedinf.2022.104854] [Medline: 35981479]
23. Cai H, Wang H, Guo T, Bao G. Application of telemedicine in Gansu province of China. PLoS ONE 2016 Jun 22;11(6):e0158026. [doi: 10.1371/journal.pone.0158026]
24. Wang Y, Yang J, Ma H, Dong X, Xie G, Ye S, et al. Application of telemedicine in the COVID-19 epidemic: an analysis of Gansu province in China. PLoS One 2021;16(8):e0249872 [FREE Full text] [doi: 10.1371/journal.pone.0249872] [Medline: 34347779]
25. Wang TT, Li JM, Zhu CR, Hong Z, An DM, Yang HY, et al. Assessment of utilization and cost-effectiveness of telemedicine program in western regions of China: a 12-year study of 249 hospitals across 112 cities. Telemed J E Health 2016 Nov;22(11):909-920. [doi: 10.1089/tmj.2015.0213] [Medline: 27314300]
26. Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine during the COVID-19 pandemic: experiences from western China. J Med Internet Res 2020 May 08;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]
27. Zhai Y, Gao J, Chen B, Shi J, Wang L, He X, et al. Design and application of a telemedicine system jointly driven by videoconferencing and data exchange: practical experience from Henan province, China. Telemed J E Health 2020 Jan 01;26(1):89-100. [doi: 10.1089/tmj.2018.0240] [Medline: 31050599]
28. Cui F, He X, Zhai Y, Lyu M, Shi J, Sun D, et al. Application of telemedicine services based on a regional telemedicine platform in China from 2014 to 2020: longitudinal trend analysis. J Med Internet Res 2021 Jul 12;23(7):e28009 [FREE Full text] [doi: 10.2196/28009] [Medline: 34255686]
29. He X, Wang L, Wang L, Gao J, Cui F, Ma Q, et al. Effectiveness of a cloud-based telepathology system in China: large-sample observational study. J Med Internet Res 2021 Jul 29;23(7):e23799 [FREE Full text] [doi: 10.2196/23799] [Medline: 34326037]
30. Khanal S, Burgon J, Leonard S, Griffiths M, Eddowes LA. Recommendations for the improved effectiveness and reporting of telemedicine programs in developing countries: results of a systematic literature review. Telemed J E Health 2015 Nov;21(11):903-915 [FREE Full text] [doi: 10.1089/tmj.2014.0194] [Medline: 26103465]
31. Kamdar NV, Huverserian A, Jalilian L, Thi W, Duval V, Beck L, et al. Development, implementation, and evaluation of a telemedicine preoperative evaluation initiative at a major academic medical center. Anesth Analg 2020 Dec;131(6):1647-1656 [FREE Full text] [doi: 10.1213/ANE.0000000000005208] [Medline: 32841990]
32. Ma Q, Sun D, Tan Z, Li C, He X, Zhai Y, et al. Usage and perceptions of telemedicine among health care professionals in China. Int J Med Inform 2022 Oct;166:104856 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104856] [Medline: 36037593]
33. Lu W, Wang X, Zhao J, Zhai Y. Research on teleconsultation service quality based on multi-granularity linguistic information: the perspective of regional doctors. BMC Med Inform Decis Mak 2020 Jun 18;20(1):113 [FREE Full text] [doi: 10.1186/s12911-020-01155-5] [Medline: 32552734]
34. Lin JC, Kavousi Y, Sullivan B, Stevens C. Analysis of outpatient telemedicine reimbursement in an integrated healthcare system. Ann Vasc Surg 2020 May;65:100-106. [doi: 10.1016/j.avsg.2019.10.069] [Medline: 31678131]
35. Fatehi F, Armfield NR, Dimitrijevic M, Gray LC. Technical aspects of clinical videoconferencing: a large scale review of the literature. J Telemed Telecare 2015 Feb 19;21(3):160-166. [doi: 10.1177/1357633x15571999]
36. Burke BL, Hall RW, Section on Telehealth Care. Telemedicine: pediatric applications. Pediatrics 2015 Jul;136(1):e293-e308 [FREE Full text] [doi: 10.1542/peds.2015-1517] [Medline: 26122813]
37. Xu T, Pujara S, Sutton S, Rhee M. Telemedicine in the management of type 1 diabetes. Prev Chronic Dis 2018 Jan 25;15:E13 [FREE Full text] [doi: 10.5888/pcd15.170168] [Medline: 29369757]
38. Jumreornvong O, Yang E, Race J, Appel J. Telemedicine and medical education in the age of COVID-19. Acad Med 2020 Dec;95(12):1838-1843 [FREE Full text] [doi: 10.1097/ACM.0000000000003711] [Medline: 32889946]
39. Li J, Yang X, Chu G, Feng W, Ding X, Yin X, et al. Application of improved robot-assisted laparoscopic telesurgery with 5G technology in urology. Eur Urol 2023 Jan;83(1):41-44. [doi: 10.1016/j.eururo.2022.06.018] [Medline: 35817641]
40. Yager PH. Remote parent participation in intensive care unit rounds. Pediatr Clin North Am 2020 Aug;67(4):655-659. [doi: 10.1016/j.pcl.2020.04.008] [Medline: 32650863]

XSL•FO
RenderX

Abbreviations

MCU: multipoint control unit
tele-ECG: tele-electrocardiogram

Edited by T Leung, G Eysenbach; submitted 13.12.22; peer-reviewed by H Ayatollahi, L Wang, J Yang, V Brahmbhatt, M Sousa; comments to author 13.04.23; revised version received 26.04.23; accepted 17.11.23; published 05.01.24.

Please cite as:
He X, Cui F, Lyu M, Sun D, Zhang X, Shi J, Zhang Y, Jiang S, Zhao J
Key Factors Influencing the Operationalization and Effectiveness of Telemedicine Services in Henan Province, China: Cross-Sectional Analysis
J Med Internet Res 2024;26:e45020
URL: https://www.jmir.org/2024/1/e45020
doi:10.2196/45020
PMID:38180795

©Xianying He, Fangfang Cui, Minzhao Lyu, Dongxu Sun, Xu Zhang, Jinming Shi, Yinglan Zhang, Shuai Jiang, Jie Zhao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Costs and Cardiovascular Benefits of a Fourth-Generation Synchronous Telehealth Program on Mortality and Cardiovascular Outcomes for Patients With Atrial Fibrillation: Retrospective Cohort Study

Hao-Yun Chang[1,2,3], MD; Hui-Wen Wu[4,5], MSc, RN; Chi-Sheng Hung[1,4,5,6], MD, PhD; Ying-Hsien Chen[1,4,5,6], MD; Ching-Chang Huang[1,4,5,6], MD; Li-Tan Yang[1,4,5,6], MD; Shin-Tsyr Hwang[5,7], MSc, RN; Jiun-Yu Yu[5,8], PhD; Jen-Kuang Lee[1,4,5,6,9], MD, PhD; Yi-Lwun Ho[1,4,5,6], MD, PhD

[1]Division of Cardiology, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
[2]Division of Cardiology, Department of Internal Medicine, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu, Taiwan
[3]Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan
[4]Cardiovascular Center, National Taiwan University Hospital, Taipei, Taiwan
[5]Telehealth Center, National Taiwan University Hospital, Taipei, Taiwan
[6]Department of Internal Medicine, National Taiwan University College of Medicine, Taipei, Taiwan
[7]Department of Nursing, National Taiwan University Hospital, Taipei, Taiwan
[8]Department of Business Administration, College of Management, National Taiwan University, Taipei, Taiwan
[9]Department of Laboratory Medicine, National Taiwan University College of Medicine, Taipei, Taiwan

Corresponding Author:
Jen-Kuang Lee, MD, PhD
Division of Cardiology, Department of Internal Medicine, National Taiwan University Hospital
No. 7, Chung-Shan South Road
Taipei, 100
Taiwan
Phone: 886 2 23123456 ext 62152
Fax: 886 227044688
Email: b85401104@gmail.com

Abstract

Background: The prevalence of atrial fibrillation (AF) continues to increase in modern aging society. Patients with AF are at high risk for multiple adverse cardiovascular events, including heart failure, stroke, and mortality. Improved medical care is needed for patients with AF to enhance their quality of life and limit their medical resource utilization. With advances in the internet and technology, telehealth programs are now widely used in medical care. A fourth-generation telehealth program offers synchronous and continuous medical attention in response to physiological parameters measured at home. Although we have previously shown the benefits of this telehealth program for some patients with a high risk of cardiovascular disease, its benefits for patients with AF remains uncertain.

Objective: This study aims to investigate the benefits of participating in a fourth-generation telehealth program for patients with AF in relation to cardiovascular outcomes.

Methods: This was a retrospective cohort study. We retrospectively searched the medical records database of a tertiary medical center in Northern Taiwan between January 2007 and December 2017. We screened 5062 patients with cardiovascular disease and enrolled 537 patients with AF, of which 279 participated in the telehealth program and 258 did not. Bias was reduced using the inverse probability of treatment weighting adjustment based on the propensity score. Outcomes were collected and analyzed, including all-cause readmission, admission for heart failure, acute coronary syndrome, ischemic stroke, systemic embolism, bleeding events, all-cause mortality, and cardiovascular death within the follow-up period. Total medical expenses and medical costs in different departments were also compared. Subgroup analyses were conducted on ischemic stroke stratified by several subgroup variables.

XSL•FO
RenderX

Results: The mean follow-up period was 3.0 (SD 1.7) years for the telehealth group and 3.4 (SD 1.9) years for the control group. After inverse probability of treatment weighting adjustment, the patients in the telehealth program had significantly fewer ischemic strokes (2.0 vs 4.5 events per 100 person-years; subdistribution hazard ratio [SHR] 0.45, 95% CI 0.22-0.92) and cardiovascular deaths (2.5 vs 5.9 events per 100 person-years; SHR 0.43, 95% CI 0.18-0.99) at the follow-up. The telehealth program particularly benefited patients comorbid with vascular disease (SHR 0.11, 95% CI 0.02-0.53 vs SHR 1.16, 95% CI 0.44-3.09; *P*=.01 for interaction). The total medical expenses during follow-up were similar in the telehealth and control groups.

Conclusions: This study demonstrated the benefits of participating in the fourth-generation telehealth program for patients with AF by significantly reducing their ischemic stroke risk while spending the same amount on medical expenses.

(J Med Internet Res 2024;26:e48748) doi:10.2196/48748

KEYWORDS

atrial fibrillation; cardiovascular death; fourth-generation synchronous program; ischemic stroke; telehealth

Introduction

In the past 3 decades, atrial fibrillation (AF) has become one of the most frequent cardiac arrhythmias worldwide. Its high disease burden has made it a significant public health issue. The worldwide incidence of AF in 2017 was estimated at 3046 million new cases, a 31% increase compared to 1997 [1]. The incidence of AF varies with age. AF affects <0.2% of individuals aged 49 years and younger, but 10%-17% of individuals aged 80 years or older [2]. Given the aging trend in societies worldwide, AF prevalence is expected to increase continuously [3]. Taiwan is also not exempt from this worldwide phenomenon. The prevalence of AF in Taiwan in 2011 stood at 1.07%, and it was projected to reach an estimated 4.01% by the year 2050 [4]. While multiple treatment modalities and strategies have been developed for AF, patients with AF are still associated with multiple adverse cardiovascular outcomes, including ischemic stroke, myocardial infarction, heart failure, cognitive decline, and even mortality [5]. An innovative modality to provide better outcomes for patients with AF is needed.

With advancements in technology, we have entered the internet era. Telemedicine has been well established as an essential part of contemporary medical practice. A fourth-generation telehealth program provides internet-based, synchronous, round-the-clock disease management and immediate response. Studies have shown that patients receiving telemedicine have better control over numerous cardiovascular risk factors, including diabetes mellitus, hypertension, and hyperlipidemia [6]. It even reduced all-cause mortality in heart failure patients [7,8]. Nevertheless, Wootton [9] conducted a literature review and argued that the evidence base for the value of telemedicine in managing chronic diseases was somewhat contradictory. Therefore, the benefits of telemedicine for a particular disease need to be evaluated. According to our annual report, the characteristics of the population that tends to use telehealth are male (66%), aged 65 years or older (53%), people with hypertension (55%), and people with coronary artery disease (53%). Our group previously showed the benefits of participating in a fourth-generation telehealth program for patients with peripheral artery disease [10]. However, the benefits to patients with AF remains to be determined.

This study aimed to investigate the benefits of participating in a fourth-generation telehealth program for patients with AF in relation to cardiovascular outcomes, including cardiovascular events, readmission, mortality, and financial costs.

Methods

Data Source and Patients

This was a retrospective cohort study. All patients aged 20 years or older diagnosed with cardiovascular disease between January 2007 and December 2017 were screened and had their charts reviewed. Cardiovascular disease included AF, coronary artery disease, myocardial infarction, heart failure, peripheral artery disease, stroke, and hypertension. Only patients with AF were selected for this study. Patients with AF who also participated in the fourth-generation synchronous telehealth program at the National Taiwan University Hospital (NTUH; Taipei, Taiwan) telehealth center were enrolled as the study group. Patients with AF who did not participate in the telehealth program receiving usual standard treatment were enrolled as the control group. The decision to participate in the telehealth program depended primarily on the patients and their caregivers. The cohort entry date was defined as the day of inviting the patients to participate in the telehealth program.

Telehealth Care Program

The fourth-generation synchronous telehealth program at NTUH was an internet-based, multiaspects-integrated, remote management health care program for chronic disease. The Graduate Institute of Biomedical Electronics and Bioinformatics at the National Taiwan University developed the integrated platform. We previously reported the details of this program [11]. Briefly, this advanced telehealth program collects participants' biometric data, including heart rate, blood pressure, oximetry data, and single-lead electrocardiography. These data are transmitted to our telehealth center daily and on demand. Nurses manage each participant by making telephone calls periodically and on demand, providing communication and health promotion. On-call cardiologists are available 24/7 to provide immediate responses and recommendations in any emergency. Following acute events, long-term medication and health care management are discussed with the patient's primary care physician and provided to the patient. This telehealth program emphasizes education, medication adherence, prevention, and early condition deterioration recognition. It strengthens the bridge between acute events and long-term home care.

XSL•FO
RenderX

Usual Standard Care

The patients in the control group did not receive the telehealth program but the usual standard treatment provided at our cardiovascular center outpatient clinics. Their primary care physician provided the usual standard care according to, but not limited to, the updated guidelines for managing chronic AF, guidelines for managing stable ischemic heart disease, American Diabetes Association guidelines for diabetes management, and the American Heart Association's guidelines for lifestyle modification and primary prevention to reduce cardiovascular risks. There was no contact between the telehealth center, the primary care physician, and patients receiving the usual standard care.

Covariates and Outcomes

The covariates were demographics (age, sex, and BMI), comorbidities (hypertension, diabetes, hyperlipidemia, coronary artery disease, chronic kidney disease, myocardial infarction, stroke history, heart failure, cancer, and peripheral arterial occlusive disease), "congestive heart failure, hypertension, age, diabetes mellitus, previous stroke or transient ischemic accident or thromboembolism, vascular disease, age, and sex category" (CHA_2DS_2-VASc) score, Charlson Comorbidity Index score, vital signs (mean arterial pressure and heart rate), laboratory data (serum creatinine and low-density lipoprotein cholesterol), left ventricular systolic function, and use of 10 medication types at enrollment. The outcomes were all-cause readmission, admission for heart failure, acute coronary syndrome, ischemic stroke, systemic embolism, bleeding events (major bleeding, gastrointestinal bleeding, and intracranial hemorrhage), all-cause mortality, and death due to cardiovascular disease during follow-up. The definition of all-cause readmission and admission for heart failure encompassed any admission to NTUH, and admission to NTUH specifically under the diagnosis of heart failure, subsequent to the cohort entry date, respectively. The date and cause of death were linked by the Taiwan Death Registry database. Each patient was followed from their cohort entry date until the day of event occurrence, the end of the database (December 31, 2017), or the day of death, whichever came first. The total number of admissions and medical expenses during follow-up were also analyzed. The medical expenses were retrieved from the NTUH electronic billing system, which encompassed comprehensive records of costs associated with outpatient services, emergency department visits, and hospitalizations. The total cost used was defined as medical expenses.

Statistical Analysis

This study reduced selection bias using inverse probability of treatment weighting (IPTW) based on a propensity score. The propensity score was the predicted probability of being in the telehealth group given certain values of the covariates using a multivariable logistic regression model. The covariates used in the propensity score calculation are listed in Table 1, where the follow-up duration is replaced by the cohort entry date. To mitigate the impact of extreme weights, we truncated the outlier weights at the 99th percentile of the weights [12]. The balance of covariates between the telehealth and nontelehealth groups was evaluated using the absolute value of the standardized difference between groups, where a value of <0.1 was considered negligible. In addition, there were some missing data (eg, BMI, vital signs, and lipid profile), which were imputed using the single expectation-maximization algorithm. The assumption of missing completely at random was evaluated using the Little test. The IPTW was conducted on the imputed data without missing values.

XSL•FO
RenderX

Table 1. Baseline characteristics of patients with atrial fibrillation participating or not in the telehealth program. The number of patients has been inflated after inverse probability treatment weighting (IPTW) adjustment; therefore, n values are not provided with the percentages.

	Before IPTW			After IPTW			Available number
	Telehealth (n=279)	Nontelehealth (n=258)	STD[a]	Telehealth	Nontelehealth	STD	
Age (years), mean (SD)	68.0 (13.1)	69.2 (13.2)	–0.09	68.3 (13.0)	68.4 (13.2)	–0.01	537
Male, n (%)	156 (55.9)	157 (60.9)	–0.10	58.3	57.1	0.02	537
BMI (kg/m^2), mean (SD)	24.4 (4.4)	24.4 (3.7)	–0.01	24.4 (4.3)	24.4 (3.6)	0.02	522
Comorbid condition, n (%)							537
Hypertension	148 (53)	146 (56.6)	–0.07	55	54	0.02	
Diabetes mellitus	78 (28)	63 (24.4)	0.08	27.7	27	0.02	
Hyperlipidemia	95 (34.1)	84 (32.6)	0.03	32.9	31.9	0.02	
Coronary artery disease	113 (40.5)	126 (48.8)	–0.17	44.4	44.6	<0.01	
Chronic kidney disease	36 (12.9)	34 (13.2)	–0.01	13.3	13.3	<0.01	
Old myocardial infarction	12 (4.3)	19 (7.4)	–0.13	7	6.4	0.03	
Ischemic stroke	61 (21.9)	57 (22.1)	–0.01	21.4	21.1	0.01	
Hemorrhagic stroke	2 (0.7)	3 (1.2)	–0.05	1.4	1.1	0.03	
Transient ischemic attack	10 (3.6)	9 (3.5)	0.01	3.6	3.9	–0.01	
Any stroke	62 (22.2)	59 (22.9)	–0.02	21.9	21.9	<0.01	
Heart failure hospitalization	90 (32.3)	102 (39.5)	–0.15	35.5	36.2	–0.02	
Cancer	29 (10.4)	17 (6.6)	0.14	8.8	7.7	0.04	
Peripheral arterial occlusive disease	17 (6.1)	29 (11.2)	–0.18	8.7	8.8	<0.01	
CHA_2DS_2-VASc[b] score, mean (SD)	3.4 (2.2)	3.6 (2.0)	–0.11	3.4 (2.1)	3.5 (2.0)	–0.03	
Charlson Comorbidity Index score, mean (SD)	1.78 (1.78)	1.83 (1.78)	–0.03	1.9 (1.9)	1.8 (1.8)	0.02	
Vital sign, mean (SD)							507
Mean arterial pressure (mmHg)	88.3 (11.4)	89.5 (14.3)	–0.09	89.0 (10.8)	88.6 (14.4)	0.03	
Heart rate (beat/min)	73.5 (14.0)	73.7 (16.0)	–0.01	73.5 (13.4)	73.6 (15.3)	–0.01	
Laboratory data, mean (SD)							
Creatinine (mg/dL)	1.3 (1.3)	1.5 (1.5)	–0.12	1.38 (1.47)	1.37 (1.40)	0.01	536
eGFR[c] (mL/min/1.73 m^2)	70.2 (27.6)	67.9 (28.8)	0.08	67.7 (27.6)	70.1 (29.0)	–0.08	536
LDL-C[d] (mg/dL)	97.2 (34.1)	96.2 (30.8)	0.03	96.5 (26.4)	96.2 (25.6)	0.01	360
LVEF[e] (%), mean (SD)							
Teich	61.7 (13.7)	60.9 (13.6)	0.06	61.4 (13.4)	61.3 (13.6)	0.01	407
MOD-sp4[f]	44.0 (12.4)	43.8 (12.2)	0.02	43.4 (11.9)	44.2 (11.9)	–0.07	122
sp4-el[g]	43.7 (12.3)	43.4 (13.1)	0.02	43.2 (12.0)	43.4 (12.9)	–0.02	103
Medication, n (%)							537
ACEI[h] or ARB[i]	140 (50.2)	140 (54.3)	–0.08	53	52.3	0.01	
β-blocker	160 (57.3)	134 (51.9)	0.11	55	55.2	<0.01	
dCCB[j]	102 (36.6)	105 (40.7)	–0.09	38.9	38	0.02	
Diuretics	152 (54.5)	144 (55.8)	–0.03	54.5	54.9	–0.01	
Aspirin	97 (34.8)	106 (41.1)	–0.13	37.1	37.8	–0.02	

XSL•FO
RenderX

	Before IPTW			After IPTW			Available number
	Telehealth (n=279)	Nontelehealth (n=258)	STD[a]	Telehealth	Nontelehealth	STD	
Clopidogrel or cilostazol	43 (15.4)	55 (21.3)	–0.15	18.2	19.2	–0.03	
Oral anticoagulants	152 (54.5)	116 (45)	0.19	50.3	49.6	0.01	
Oral hypoglycemic agents	52 (18.6)	37 (14.3)	0.12	16.8	16.3	0.01	
Insulin	12 (4.3)	7 (2.7)	0.09	3.8	4.2	–0.02	
Statin	76 (27.2)	73 (28.3)	–0.02	28.7	28.1	0.01	
Follow-up year, mean (SD)	3.0 (1.7)	3.4 (1.9)	–0.25	3.2 (1.8)	3.2 (1.7)	–0.01	

[a]STD: standardized difference.

[b]CHA_2DS_2-VASc: congestive heart failure, hypertension, age, diabetes mellitus, previous stroke or transient ischemic accident or thromboembolism, vascular disease, age, and sex category.

[c]eGFR: estimated glomerular filtration rate.

[d]LDL-C: low-density lipoprotein cholesterol.

[e]LVEF: left ventricular ejection fraction.

[f]MOD-sp4: method of disc-single plane 4 chamber view.

[g]sp4-el: single plane 4 chamber view-ellipsoid model.

[h]ACEi: angiotensin-converting enzyme inhibitor.

[i]ARB: angiotensin receptor blocker.

[j]dCCB: dihydropyridine calcium channel blocker.

The incidence of outcomes during follow-up was calculated using the incidence density, expressed as the number of events per 100 person-years. The absolute risk difference between the groups was also calculated in the IPTW-adjusted cohort. Furthermore, the risks of outcomes during follow-up were compared between groups using the Cox proportional hazards model for fatal outcomes (eg, mortality) or the Fine-Gray subdistribution hazard model for nonfatal outcomes (eg, ischemic stroke). The outcomes were assessed during the follow-up period. The average number of all-cause readmissions during follow-up between groups was compared using a Poisson model, which treated the logarithm of follow-up duration as an offset variable. The accumulated medical expenditures during follow-up were compared between groups using a linear regression model. The study group was the only explanatory factor in these regression models.

A subgroup analysis was conducted on ischemic stroke outcome stratified by several subgroup variables, including CHA_2DS_2-VASc score (<4 vs ≥4), sex, age (<75 years vs ≥75 years), hypertension, heart failure, diabetes, previous ischemic stroke, vascular disease, and use of oral anticoagulants. A 2-sided *P* value of <.05 was considered statistically significant. However, the clinical significance of subgroup analyses was loosened to a *P* value of <.10 because the interaction test was known to be conservative [13,14]. All statistical analyses were performed using SAS (version 9.4; SAS Institute).

Ethical Considerations

This study was approved by the institutional review board of NTUH (201804072RINA). This study was conducted under the relevant regulation and protocols.

Results

Patients' Demographics and Clinical Characteristics

This study screened 5062 patients, of whom 537 had AF and were selected for analysis. Among them, a total of 279 and 258 patients were in the telehealth and nontelehealth groups, respectively. Their detailed demographics and clinical characteristics are listed in Table 1. The mean age was 68.0 (SD 13.1) years in the telehealth group and 69.2 (SD 13.2) years in the nontelehealth group. Before IPTW, patients in the telehealth group were less likely to be male; had lower prevalence of coronary artery disease, myocardial infarction, heart failure hospitalization, and peripheral arterial occlusive disease; had lower CHA_2DS_2-VASc scores; had lower serum creatinine levels; were more likely to take β-blockers, oral anticoagulants, and oral hypoglycemic agents; and were less likely to take aspirin and clopidogrel or cilostazol. Patients in the telehealth group had a higher prevalence of cancer. However, after IPTW adjustment, the characteristics were well-balanced between the groups, with all the absolute standardized difference values being <0.1. The mean follow-up time was 3.0 (SD 1.7) years for the telehealth group and 3.4 (SD 1.9) years for the nontelehealth group. In addition, the assumption of missing completely at random was held given the insignificant Little test.

Prognosis of Patients With AF Participating in the Telehealth Program

After IPTW adjustment, the all-cause admission rate during the follow-up period was significantly higher in the telehealth group (50.4 vs 39.7 events per 100 person-years; subdistribution hazard ratio [SHR] 1.38, 95% CI 1.15-1.65). Notably, the incidence of

XSL•FO
RenderX

ischemic stroke was significantly lower in the telehealth group (2.0 vs 4.5 events per 100 person-years; SHR 0.45, 95% CI 0.22-0.92). In addition, cardiovascular death was also significantly lower in the telehealth group (2.5 vs 5.9 events per 100 person-years; SHR 0.43, 95% CI 0.18-0.99; Table 2). The cumulative incidence of ischemic stroke was consistently lower in the telehealth group, deviating further from the nontelehealth group over time (Figure 1A). Separately considering each follow-up year, the cumulative incidence of ischemic stroke was consistently lower in the telehealth group in both the first and second years of follow-up (Figure 1B). The absolute risk difference for each outcome is shown in Figure 2.

Figure 1. The cumulative incidence of ischemic stroke in patients with atrial fibrillation participating (red line) and not participating (blue line) in the telehealth program in the inverse probability treatment weighting–adjusted cohort over (A) the follow-up period and (B) in each follow-up year.

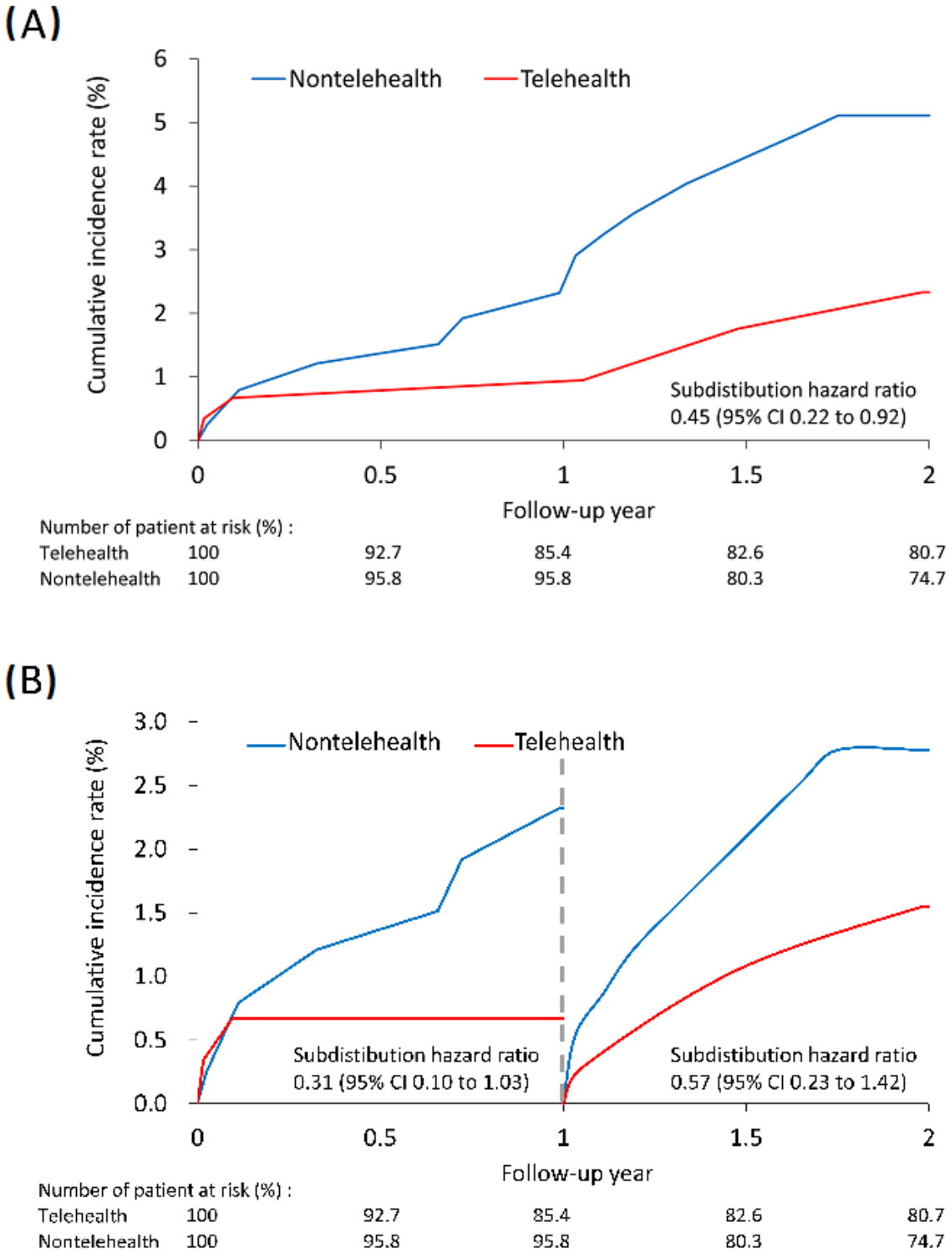

XSL•FO
RenderX

Figure 2. The absolute risk difference of each outcome during 2-year follow-up in patients with atrial fibrillation participating and not participating in the telehealth program in the inverse probability treatment weighting–adjusted cohort. ARD: absolute risk difference.

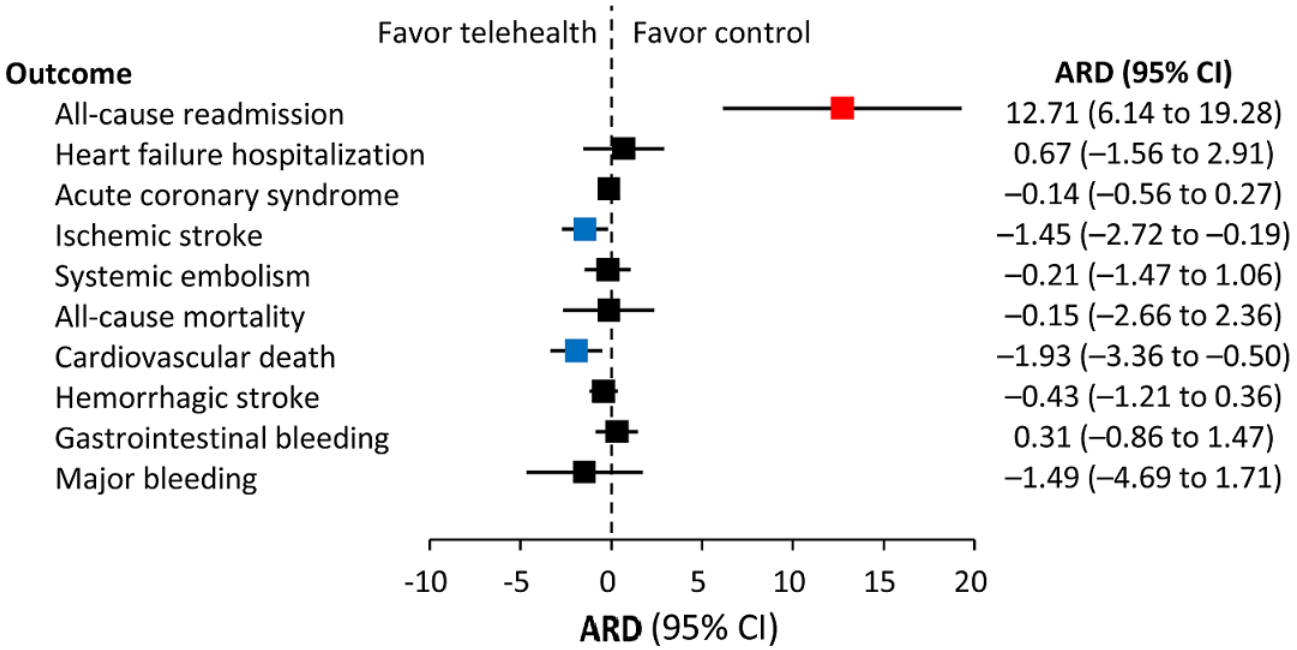

XSL•FO
RenderX

Table 2. Clinical outcomes of patients with atrial fibrillation at the follow-up. The number of patients has been inflated after inverse probability treatment weighting adjustment; therefore, n values are not provided with the percentages.

Outcome	Telehealth		Nontelehealth		ARD[a] (95% CI)	Ratio of telehealth, HR[b] or SHR[c] (95% CI)	*P* value
	%	ID[d] (95% CI)[e]	%	ID (95% CI)[e]			
All-cause readmission	50.4	43.1 (38.0 to 48.2)	39.7	30.4 (26.3 to 34.5)	12.71 (6.14 to 19.28)	1.38[f] (1.15 to 1.65)	.001
Heart failure hospitalization	9.6	6.0 (4.4 to 7.6)	8.8	5.3 (3.8 to 6.9)	0.67 (–1.56 to 2.91)	1.12[f] (0.76 to 1.67)	.57
Acute coronary syndrome	0.2	0.13 (–0.10 to 0.37)	0.5	0.28 (–0.06 to 0.62)	–0.14 (–0.56 to 0.27)	0.48[f] (0.06 to 4.13)	.50
Ischemic stroke	2	1.2 (0.5 to 1.9)	4.5	2.6 (1.6 to 3.7)	–1.45 (–2.72 to –0.19)	0.45[f] (0.22 to 0.92)	.03
Systemic embolism	3	1.8 (0.9 to 2.6)	3.4	2.0 (1.1 to 2.9)	–0.21 (–1.47 to 1.06)	0.90[f] (0.46 to 1.77)	.77
Major bleeding	5.4	5.8 (3.7 to 7.9)	6.7	7.3 (4.9 to 9.7)	–1.49 (–4.69 to 1.71)	0.72[f] (0.45 to 1.17)	.19
Gastrointestinal bleeding	2.9	1.8 (0.9 to 2.6)	2.5	1.5 (0.7 to 2.2)	0.31 (–0.86 to 1.47)	1.23[f] (0.59 to 2.53)	.58
Intracranial hemorrhage	0.9	0.53 (0.06 to 1.00)	1.7	0.95 (0.32 to 1.58)	–0.43 (–1.21 to 0.36)	0.56[f] (0.18 to 1.69)	.30
All-cause mortality	12.7	7.5 (5.7 to 9.2)	13.3	7.6 (5.9 to 9.4)	–0.15 (–2.66 to 2.36)	0.97[g] (0.58 to 1.65)	.92
Cardiovascular death	2.5	1.5 (0.7 to 2.3)	5.9	3.4 (2.2 to 4.6)	–1.93 (–3.36 to –0.50)	0.43[g] (0.18 to 0.99)	.049

[a]ARD: absolute risk difference.

[b]HR: hazard ratio.

[c]SHR: subdistribution hazard ratio.

[d]ID: incidence density.

[e]100 person-years.

[f]Subdistribution hazard ratio.

[g]Hazard ratio.

Subgroup Analysis of Ischemic Stroke

The subgroup analysis showed that the observed favorable effect in the telehealth group was more apparent in patients with CHA_2DS_2-VASc scores ≥4 than <4 (SHR 0.13, 95% CI 0.03-0.66 vs SHR 0.82, 95% CI 0.34-1.99; *P*=.05 for interaction) but not statistically significant. However, the observed favorable effect in the telehealth group was more pronounced in patients with than without vascular disease (SHR 0.11, 95% CI 0.02-0.53 vs SHR 1.16, 95% CI 0.44-3.09; *P*=.01 for interaction). The remaining subgroup variables had no significant interaction effects (Figure 3).

XSL•FO
RenderX

Figure 3. The different subgroup analysis of ischemic stroke risk during the 2-year follow-up period. The subdistribution hazard ratio (SHR) and *P* value for variable interaction are shown in the right columns. CHA2DS2-VASc: congestive heart failure, hypertension, age, diabetes mellitus, previous stroke or transient ischemic accident or thromboembolism, vascular disease, age, and sex category; NA: not applicable.

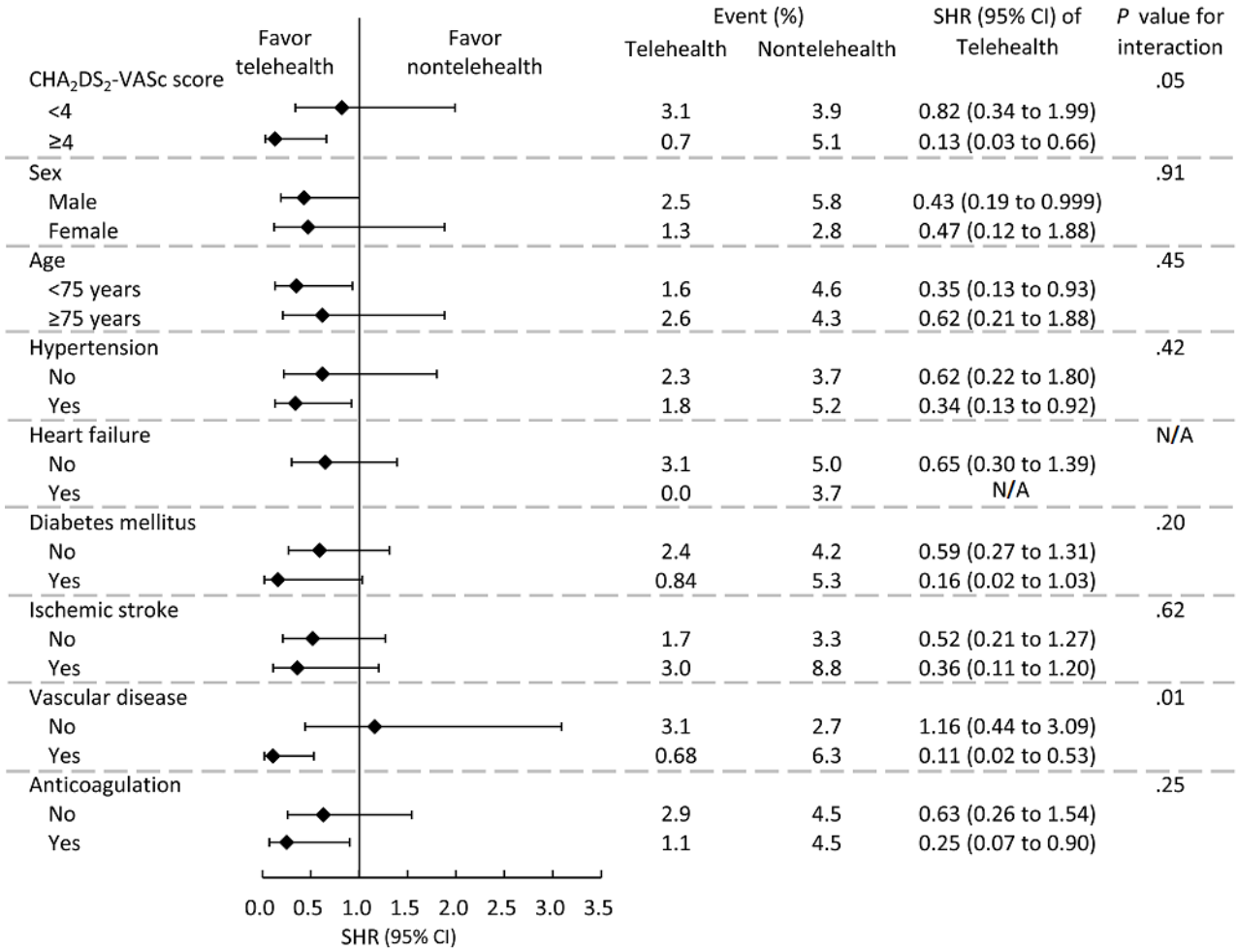

Medical Use and Costs for Patients Participating in the Telehealth Program

Readmissions were significantly higher in the telehealth group (1.6, SD 2.3 times) than in the nontelehealth group (1.2, SD 1.7 times; rate ratio 1.35, 95% CI 1.22-1.50). Nevertheless, medical expenditure did not differ significantly between groups during the entire follow-up period, including outpatient clinic, hospitalization, emergency department, and total expenses (Table 3).

Table 3. Medical use and costs for patients with atrial fibrillation participating or not in the telehealth program.

Outcome	Telehealth (n=537.2), mean (SD)	Nontelehealth (n=532.9), mean (SD)	Ratio of telehealth, RR[a] or B[b] (95% CI)	*P* value
Number of readmissions	1.6 (2.3)	1.2 (1.7)	1.35[c] (1.22 to 1.50)	<.001
Medical expenditure (US$ × 10^3)				
Outpatient	11.0 (10.5)	9.5 (10.1)	1.50[d] (–0.25 to 3.24)	.09
Hospitalization	19.4 (18.8)	19.8 (19.5)	–0.37[d] (–3.60 to 2.87)	.83
Emergency department	1.2 (1.5)	1.2 (1.6)	–0.02[d] (–0.29 to 0.24)	.87
Total	37.9 (58.6)	33.3 (35).8	4.6[d] (–3.6 to 12.8)	.27

[a]RR: rate ratio.

[b]B: regression coefficient.

[c]Rate ratio.

[d]Regression coefficient.

XSL•FO
RenderX

Discussion

Principal Results

This study marked a pioneering investigation into the efficacy of implementing the fourth-generation synchronous telehealth program for patients with AF in terms of clinical outcomes. The results demonstrated a favorable effect of the telehealth program, effectively reducing the risk of stroke in patients diagnosed with AF. It also shows that the patients in the telehealth program had significantly fewer cardiovascular deaths at the follow-up. Besides, the total medical expenses during follow-up were similar in the telehealth and control groups.

Overviews

Among all possible complications of AF, ischemic stroke is the most feared and highly related. The average annual risk of ischemic stroke in patients with AF was 4.4% [15], and AF increased the relative risk of stroke 4-5-fold [16]. Since it was frequently encountered and had severe consequences leading to disability and mortality, any method to reduce this risk has great clinical benefit and impact. This study demonstrated that participating in the telehealth program significantly decreased the stroke rate. The difference was evident as early as the first year and continuously had a favorable effect on ischemic stroke events in the second year of follow-up. By the end of the 2-year follow-up, the risk was reduced by half in the patients receiving standard care plus a telehealth program compared to those receiving standard in-person clinic follow-up only (SHR 0.45, 95% CI 0.22-0.92). The results showed that the difference between the 2 groups increased over time. The greater the adherence to the telehealth program, the greater the benefit to the patient. Similarly, Guo et al [17,18] showed that using mobile health technology for AF patient management could improve quality of life, increase medication adherence, improve patient education, and reduce the risk of rehospitalization and adverse events.

Besides ischemic stroke, the absolute risk difference for cardiovascular death also showed a favorable outcome in the telehealth group. There were significantly fewer cardiovascular deaths in the telehealth group. While the telehealth program appeared to have a negative effect in this study, causing more readmissions, it translated into better clinical outcomes by reducing ischemic stroke and cardiovascular death. There were many possible reasons for these results. The fourth-generation synchronous telehealth program provided real-time feedback on physiological measurements with periodic web-based check-ins by the nurse managers and timely interpretation by the cardiologists. Early recognition of disease deterioration, symptom development, or condition change triggered a hospital visit and possible subsequent hospitalization but ultimately contributed to better clinical outcomes. In addition, this program also facilitated timely management and increased drug adherence.

Despite many studies showing favorable telemedicine outcomes, others failed to show better clinical outcomes in patients participating in telehealth programs [9,19]. Differences in patient population enrollment, telehealth programs, and telemonitor devices across studies could explain these contradictory results. Therefore, the effectiveness of each telehealth program for a specific disease must be evaluated individually. Our group has previously reported the favorable outcomes of participating in the fourth-generation synchronous telehealth program for patients with chronic cardiovascular and peripheral artery disease [10,11]. This study demonstrated and confirmed that applying this telehealth program to patients with AF was also beneficial.

Subgroup analysis provided further insight into which population benefits from this telehealth program the most in relation to reduced ischemic stroke risk. Pre-existing vascular disease showed the strongest interaction regarding fewer stroke events. Peripheral artery disease, coronary artery disease, and cerebral vascular disease shared common risk factors with ischemic stroke and were all associated with a higher stroke rate [20-22]. It is well known that patients with higher CHA_2DS_2-VASc scores have an increased risk for ischemic stroke. Notably, there is a trend for clinical significance that a protective effect was observed within the telehealth group with a CHA_2DS_2-VASc ≥4 points, but the *P* value of .05 for interaction did not attain statistical significance. The limited sample size may also conceal the true significance. We previously demonstrated that patients with higher CHA_2DS_2-VASc scores had higher risks of cardiovascular admissions, but the fourth-generation telehealth program diminished the outcome difference [23]. In clinical practice, it has also been observed that patients at high risk of stroke derive the greatest benefit from stroke prevention measures [24]. Our results concurred with those findings and indicated that those with high ischemic stroke risk (comorbid with vascular disease or a high CHA_2DS_2-VASc score) were the most valuable target population for this telehealth program.

One of the most frequently mentioned barriers to applying telemedicine to patient care was the increasing financial cost and lack of reimbursement [25]. Some studies performed a cost-effectiveness analysis to support the extra expenditure and advocate telemedicine [26,27]. This study even showed that the fourth-generation telehealth program used for patients with AF was cost-efficient and did not increase total medical expenses. Therefore, the benefits of this telehealth program did not come at the expense of financial harm to the patients.

Limitations and Strengths

This study had several strengths and limitations. Most published studies discussing telemedicine used a follow-up duration of only several months. The mean follow-up duration in this study was >3 years in both the control and telehealth groups. For most chronic disease complications, prevalence positively correlated with time. Some have argued that the beneficial evidence was not robust due to the short follow-up period. This issue was somewhat overcome in this study, which provided evidence regarding relative long-term clinical outcomes. However, this study was retrospective and nonrandomized, resulting in heterogeneity in the patient population and disease severity. The 2 groups were well-matched after IPTW adjustment, minimizing the possible confounding effect of clinical factors. Second, the clinical outcomes were retrieved from our hospital's medical records and electronic billing system. If the patient received extra medical care outside of our hospital, it was not

XSL•FO
RenderX

recorded. Potential resources used but not charged for were also neglected and overlooked when calculating expenditures.

Conclusions

In summary, this study showed the benefits to patients with AF of participating in the fourth-generation synchronous telehealth program. Their stroke risk was significantly decreased, especially for those with pre-existing vascular disease and high CHA_2DS_2-VASc scores. There was no increase in total medical expenditure when participating in this telehealth program. Future prospective, randomized trials are needed to further confirm our findings and fully integrate them into the optimal multimodality management of patients with AF.

Acknowledgments

YLH and JKL correspond equally to this study. We thank the staff of the Department of Medical Research at National Taiwan University Hospital (NTUH) for the integrated medical database. We also thank Alfred Hsing-Fen Lin for his assistance with the statistical analysis. This research was supported by the Ministry of Science and Technology of Taiwan (MOST 108-2221-E-002-163- and MOST 109-2221-E-002-083-) and NTUH (107-EDN11, 108-N4406, 108-EDN02, 109-O20, 109-S4579, and 109-EDN11).

Conflicts of Interest

None declared.

References

1. Lippi G, Sanchis-Gomar F, Cervellin G. Global epidemiology of atrial fibrillation: an increasing epidemic and public health challenge. Int J Stroke 2021;16(2):217-221. [doi: 10.1177/1747493019897870] [Medline: 31955707]
2. Zoni-Berisso M, Lercari F, Carazza T, Domenicucci S. Epidemiology of atrial fibrillation: European perspective. Clin Epidemiol 2014;6:213-220 [FREE Full text] [doi: 10.2147/CLEP.S47385] [Medline: 24966695]
3. Krijthe BP, Kunst A, Benjamin EJ, Lip GYH, Franco OH, Hofman A, et al. Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060. Eur Heart J 2013;34(35):2746-2751 [FREE Full text] [doi: 10.1093/eurheartj/eht280] [Medline: 23900699]
4. Chao TF, Liu CJ, Tuan TC, Chen TJ, Hsieh MH, Lip GYH, et al. Lifetime risks, projected numbers, and adverse outcomes in Asian patients with atrial fibrillation: a report from the Taiwan nationwide AF cohort study. Chest 2018;153(2):453-466. [doi: 10.1016/j.chest.2017.10.001] [Medline: 29017957]
5. Rahman F, Kwan GF, Benjamin EJ. Global epidemiology of atrial fibrillation. Nat Rev Cardiol 2014;11(11):639-654. [doi: 10.1038/nrcardio.2014.118] [Medline: 25113750]
6. Vernooij JWP, Kaasjager HAH, van der Graaf Y, Wierdsma J, Grandjean HMH, Hovens MMC, et al. Internet based vascular risk factor management for patients with clinically manifest vascular disease: randomised controlled trial. BMJ 2012;344:e3750 [FREE Full text] [doi: 10.1136/bmj.e3750] [Medline: 22692651]
7. Kitsiou S, Paré G, Jaana M. Effects of home telemonitoring interventions on patients with chronic heart failure: an overview of systematic reviews. J Med Internet Res 2015;17(3):e63 [FREE Full text] [doi: 10.2196/jmir.4174] [Medline: 25768664]
8. Inglis SC, Clark RA, McAlister FA, Stewart S, Cleland JGF. Which components of heart failure programmes are effective? A systematic review and meta-analysis of the outcomes of structured telephone support or telemonitoring as the primary component of chronic heart failure management in 8323 patients: abridged Cochrane review. Eur J Heart Fail 2011;13(9):1028-1040 [FREE Full text] [doi: 10.1093/eurjhf/hfr039] [Medline: 21733889]
9. Wootton R. Twenty years of telemedicine in chronic disease management--an evidence synthesis. J Telemed Telecare 2012;18(4):211-220 [FREE Full text] [doi: 10.1258/jtt.2012.120219] [Medline: 22674020]
10. Lee JK, Hung CS, Huang CC, Chen YH, Wu HW, Chuang PY, et al. The costs and cardiovascular benefits in patients with peripheral artery disease from a fourth-generation synchronous telehealth program: retrospective cohort study. J Med Internet Res 2021;23(5):e24346 [FREE Full text] [doi: 10.2196/24346] [Medline: 34003132]
11. Ho YL, Yu JY, Lin YH, Chen YH, Huang CC, Hsu TP, et al. Assessment of the cost-effectiveness and clinical outcomes of a fourth-generation synchronous telehealth program for the management of chronic cardiovascular disease. J Med Internet Res 2014;16(6):e145 [FREE Full text] [doi: 10.2196/jmir.3346] [Medline: 24915187]
12. Austin PC, Stuart EA. Moving towards best practice when using Inverse Probability of Treatment Weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 2015;34(28):3661-3679 [FREE Full text] [doi: 10.1002/sim.6607] [Medline: 26238958]
13. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11(5):561-570 [FREE Full text] [doi: 10.1097/00001648-200009000-00012] [Medline: 10955409]
14. Normand ST, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol 2001;54(4):387-398. [doi: 10.1016/s0895-4356(00)00321-8] [Medline: 11297888]

XSL•FO
RenderX

15. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA 2001;285(22):2864-2870 [FREE Full text] [doi: 10.1001/jama.285.22.2864] [Medline: 11401607]
16. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham study. Stroke 1991;22(8):983-988 [FREE Full text] [doi: 10.1161/01.str.22.8.983] [Medline: 1866765]
17. Guo Y, Chen Y, Lane DA, Liu L, Wang Y, Lip GYH. Mobile health technology for atrial fibrillation management integrating decision support, education, and patient involvement: mAF app trial. Am J Med 2017;130(12):1388-1396.e6 [FREE Full text] [doi: 10.1016/j.amjmed.2017.07.003] [Medline: 28847546]
18. Guo Y, Lane DA, Wang L, Zhang H, Wang H, Zhang W, et al. Mobile health technology to improve care for patients with atrial fibrillation. J Am Coll Cardiol 2020;75(13):1523-1534 [FREE Full text] [doi: 10.1016/j.jacc.2020.01.052] [Medline: 32241367]
19. Takahashi PY, Pecina JL, Upatising B, Chaudhry R, Shah ND, Van Houten H, et al. A randomized controlled trial of telemonitoring in older adults with multiple health issues to prevent hospitalizations and emergency department visits. Arch Intern Med 2012;172(10):773-779 [FREE Full text] [doi: 10.1001/archinternmed.2012.256] [Medline: 22507696]
20. Kolls BJ, Sapp S, Rockhold FW, Jordan JD, Dombrowski KE, Fowkes FGR, et al. Stroke in patients with peripheral artery disease. Stroke 2019;50(6):1356-1363 [FREE Full text] [doi: 10.1161/strokeaha.118.023534]
21. Bhatia R, Sharma G, Patel C, Garg A, Roy A, Bali P, et al. Coronary artery disease in patients with ischemic stroke and TIA. J Stroke Cerebrovasc Dis 2019;28(12):104400. [doi: 10.1016/j.jstrokecerebrovasdis.2019.104400] [Medline: 31606321]
22. Hu X, De Silva TM, Chen J, Faraci FM. Cerebral vascular disease and neurovascular injury in ischemic stroke. Circ Res 2017;120(3):449-471 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.308427] [Medline: 28154097]
23. Lee JK, Hung CS, Huang CC, Chen YH, Chuang PY, Yu JY, et al. Use of the CHA2DS2-VASc score for risk stratification of hospital admissions among patients with cardiovascular diseases receiving a fourth-generation synchronous telehealth program: retrospective cohort study. J Med Internet Res 2019;21(1):e12790 [FREE Full text] [doi: 10.2196/12790] [Medline: 30702437]
24. Gao X, Passman R. Stroke prevention in atrial fibrillation. Curr Cardiol Rep 2022;24(11):1765-1774. [doi: 10.1007/s11886-022-01787-1] [Medline: 36136225]
25. Al Jamal N, Abi-Saleh B, Isma'eel H. Advances in telemedicine for the management of the elderly cardiac patient. J Geriatr Cardiol 2021;18(9):759-767 [FREE Full text] [doi: 10.11909/j.issn.1671-5411.2021.09.004] [Medline: 34659382]
26. Grustam AS, Severens JL, De Massari D, Buyukkaramikli N, Koymans R, Vrijhoef HJM. Cost-effectiveness analysis in telehealth: a comparison between home telemonitoring, nurse telephone support, and usual care in chronic heart failure management. Value Health 2018;21(7):772-782 [FREE Full text] [doi: 10.1016/j.jval.2017.11.011] [Medline: 30005749]
27. Hwang R, Morris NR, Mandrusiak A, Bruning J, Peters R, Korczyk D, et al. Cost-utility analysis of home-based telerehabilitation compared with centre-based rehabilitation in patients with heart failure. Heart Lung Circ 2019;28(12):1795-1803. [doi: 10.1016/j.hlc.2018.11.010] [Medline: 30528811]

Abbreviations

AF: atrial fibrillation
CHA2DS2-VASc: congestive heart failure, hypertension, age, diabetes mellitus, previous stroke or transient ischemic accident or thromboembolism, vascular disease, age, and sex category
IPTW: inverse probability treatment weighting
NTUH: National Taiwan University Hospital
SHR: subdistribution hazard ratio

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 11.05.23; peer-reviewed by S Han; comments to author 11.06.23; revised version received 20.07.23; accepted 11.10.23; published 08.01.24.

Please cite as:
Chang HY, Wu HW, Hung CS, Chen YH, Huang CC, Yang LT, Hwang ST, Yu JY, Lee JK, Ho YL
Costs and Cardiovascular Benefits of a Fourth-Generation Synchronous Telehealth Program on Mortality and Cardiovascular Outcomes for Patients With Atrial Fibrillation: Retrospective Cohort Study
J Med Internet Res 2024;26:e48748
URL: https://www.jmir.org/2024/1/e48748
doi:10.2196/48748
PMID:38190237

XSL•FO
RenderX

©Hao-Yun Chang, Hui-Wen Wu, Chi-Sheng Hung, Ying-Hsien Chen, Ching-Chang Huang, Li-Tan Yang, Shin-Tsyr Hwang, Jiun-Yu Yu, Jen-Kuang Lee, Yi-Lwun Ho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reliability of Telepsychiatry Assessments Using the Attention-Deficit/Hyperactivity Disorder Rating Scale-IV for Children With Neurodevelopmental Disorders and Their Caregivers: Randomized Feasibility Study

Shunya Kurokawa[1], MD, PhD; Kensuke Nomura[1,2], MD, PhD; Nana Hosogane[3], MD; Takashi Nagasawa[4], MD; Yuko Kawade[2,5], MD; Yu Matsumoto[2], MD; Shuichi Morinaga[1,6], MD; Yuriko Kaise[1], MA; Ayana Higuchi[1], MA; Akiko Goto[5], MD, PhD; Naoko Inada[7], MA, PhD; Masaki Kodaira[3], MD, PhD; Taishiro Kishimoto[8], MD, PhD

[1]Department of Neuropsychiatry, School of Medicine, Keio University, Tokyo, Japan
[2]Department of Child Psychiatry, Shimada Ryoiku Medical Center for Challenged Children, Tokyo, Japan
[3]Department of Child and Adolescent Mental Health, Aiiku Clinic, Tokyo, Japan
[4]Department of Child and Adolescent Psychiatry, Tokyo Metropolitan Children's Medical Center, Tokyo, Japan
[5]Tsurugaoka Garden Hospital, Tokyo, Japan
[6]Hiratsuka City Hospital, Kanagawa, Japan
[7]Department of Clinical Psychology, Taisho University, Tokyo, Japan
[8]Hills Joint Research Laboratory for Future Preventive Medicine and Wellness, Keio University School of Medicine, Tokyo, Japan

Corresponding Author:
Taishiro Kishimoto, MD, PhD
Hills Joint Research Laboratory for Future Preventive Medicine and Wellness
Keio University School of Medicine
Mori JP Tower F7
1-3-1, Azabudai, Minato-ku
Tokyo, 106-0041
Japan
Phone: 81 3 5363 3219
Email: tkishimoto@keio.jp

Abstract

Background: Given the global shortage of child psychiatrists and barriers to specialized care, remote assessment is a promising alternative for diagnosing and managing attention-deficit/hyperactivity disorder (ADHD). However, only a few studies have validated the accuracy and acceptability of these remote methods.

Objective: This study aimed to test the agreement between remote and face-to-face assessments.

Methods: Patients aged between 6 and 17 years with confirmed *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* diagnoses of ADHD or autism spectrum disorder (ASD) were recruited from multiple institutions. In a randomized order, participants underwent 2 evaluations, face-to-face and remotely, with distinct evaluators administering the ADHD Rating Scale-IV (ADHD-RS-IV). Intraclass correlation coefficient (ICC) was used to assess the reliability of face-to-face and remote assessments.

Results: The participants included 74 Japanese children aged between 6 and 16 years who were primarily diagnosed with ADHD (43/74, 58%) or ASD (31/74, 42%). A total of 22 (30%) children were diagnosed with both conditions. The ADHD-RS-IV ICCs between face-to-face and remote assessments showed “substantial” agreement in the total ADHD-RS-IV score (ICC=0.769, 95% CI 0.654-0.849; $P<.001$) according to the Landis and Koch criteria. The ICC in patients with ADHD showed “almost perfect” agreement (ICC=0.816, 95% CI 0.683-0.897; $P<.001$), whereas in patients with ASD, it showed “substantial” agreement (ICC=0.674, 95% CI 0.420-0.831; $P<.001$), indicating the high reliability of both methods across both conditions.

Conclusions: Our study validated the feasibility and reliability of remote ADHD testing, which has potential benefits such as reduced hospital visits and time-saving effects. Our results highlight the potential of telemedicine in resource-limited areas, clinical trials, and treatment evaluations, necessitating further studies to explore its broader application.

Trial Registration: UMIN Clinical Trials Registry UMIN000039860; http://tinyurl.com/yp34x6kh

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51749) doi:10.2196/51749

KEYWORDS

acceptability; ADHD; application; attention-deficit/hyperactivity disorder; autism spectrum disorders; autism; child; children; diagnosis; management; neurodevelopmental disorder; neurodevelopmental; psychiatrists; reliability; telepsychitatry

Introduction

Attention-deficit/hyperactivity disorder (ADHD) and autism spectrum disorder (ASD) are caused by genetic and environmental factors. Children with ADHD or ASD are recognized to be at risk for emotional and behavioral difficulties during adolescence, young adulthood, and beyond, owing to social adjustment difficulties. Hence, early detection and intervention are crucial to prevent the development of additional comorbidities [1-3].

The most recent US epidemiological study reported a prevalence of 9.5% for ADHD and 2.5% for ASD [4], making ADHD the most common neurodevelopmental disorder. In Japan, although epidemiological studies on ADHD in children are lacking, the prevalence rate in adults has been reported to be 1.65% [5]. However, children with ADHD or ASD face many obstacles in accessing specialized medical care. One of the biggest challenges is the global shortage of child psychiatrists, leading to long waiting times for diagnosis. For example, in Canada and the United States, the average wait times are 7 and 13 months, respectively [6,7]. This is particularly concerning in rural or isolated areas without access to specialized medical facilities.

In addition, these children may have difficulties leaving their homes and engaging in social interactions (ie, social withdrawal). Even when they have access to specialized medical facilities, some may find it extremely difficult to see a doctor in person. Moreover, it has been reported that these children have trouble with the time management associated with ADHD, resulting in high dropout rates during treatment [8,9].

Moreover, accurate diagnosis requires close observation by medical staff and a thorough interview with caregivers regarding the child's developmental history [10,11]. However, a lack of trained evaluators can lead to difficulty in accurate assessment. Children's evaluations are susceptible to being influenced by the halo effect (the distortion of evaluations about other characteristics due to the salient features they possess) and the contrast effect (a greater perception of difference than the actual difference) [12,13].

One possible approach to solving these problems is remote diagnosis and evaluation. In recent years, because a large part of psychiatric treatment and assessment has consisted of conversations with patients, remote evaluation and treatment using remote digital tools can be effective, particularly in clinical trials.

Therefore, the US Food and Drug Administration and the European Medicines Agency are promoting remote central evaluation using videoconferencing systems in psychiatric trials [14] and physician-led clinical trials [15], which are becoming increasingly common. To remotely perform a severity diagnosis in clinical trials, it is necessary to verify the degree of agreement with the usual face-to-face assessment.

Only a few studies have validated the agreement of remote assessment tests for developmental disorders with the usual face-to-face assessment [16-18]. A study reported that the Autism Diagnostic Interview–Revised Edition score from face-to-face interviews with caregivers of 20 children with ASD was equivalent to the Autism Diagnostic Interview–Revised Edition score from telephone interviews (r=0.73-0.90) [16]. The Autism Diagnostic Observation Schedule–2, considered the gold standard for ASD observation ratings, was administered to 23 adults with ASD and showed an intraclass correlation coefficient (ICC) of 0.92 between face-to-face and remote assessments [17], with the limitation that the ratings were made by the same rater. One study used a mobile app to video record scenes in which parents were concerned about their child's development at home [18]. Among the 40 children with ASD and 11 children with typical development, 88.2% of diagnoses agreed with the face-to-face diagnosis using the Autism Diagnostic Observation Schedule and other tools (κ=0.75; sensitivity=84.9%; specificity=94.4%).

However, research on the remote assessment of ADHD ratings is limited. The ADHD Rating Scale-IV (ADHD-RS-IV) [19,20] is the gold standard tool used to support the diagnosis and severity assessment of ADHD, and it has been used in practice in many clinical trials [21,22]. It has been reported that the sensitivity and specificity of ADHD-RS-IV in Japan above the 90th percentile were 89.13% and 94.07%, respectively [23]. This study aimed to test whether the remote method is equivalent to the face-to-face method for children with ADHD and ASD and their caregivers by administering ADHD-RS-IV and determining the ICC and to verify the validity and feasibility of remote assessment.

Methods

Participants

Patients were recruited at Keio University Hospital and 4 collaborating institutions (Shimada Ryoiku Medical Center for Challenged Children, Aiiku Clinic, Tokyo Metropolitan Children's Medical Center, and Tsurugaoka Garden Hospital). The inclusion criteria were as follows: (1) confirmation of *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) [24] diagnoses of ADHD or ASD; (2) aged between 6 and 17 years at the time of obtaining consent; and (3) if receiving pharmacotherapy, the treatment was stable for at least 3 months before obtaining consent.

The exclusion criteria were as follows: (1) either the child or caregiver had a hearing or visual impairment that made it difficult to use remote tools, even with corrective devices, such as glasses or hearing aids; (2) there were no caregivers with information related to the participants' early childhood; (3)

XSL•FO
RenderX

individuals had comorbid symptoms, such as hallucinations and delusions, which made it challenging to engage in the study, as determined by the clinician; and (4) there were plans to initiate new treatments, such as pharmacotherapy or psychotherapy, during the observation period.

Baseline Assessments

Background information and data, such as age, sex, diagnosis, medication, duration of illness, and intelligence test results, were collected from medical records. For the baseline evaluation, the Autism-Spectrum Quotient–Japanese version for children, the Conners 3 Japanese version to evaluate ADHD symptoms by caregivers, the Strengths and Difficulties Questionnaire to assess social adaptation, and the Short Sensory Profile to assess sensory characteristics were administered.

Study Procedure

The ADHD-RS-IV and Childhood Autism Rating Scale–2 scores were used as primary outcomes to compare face-to-face and remote developmental evaluations.

After the baseline assessments, the participants underwent evaluations twice, either face-to-face or remotely. Recall bias could have occurred if the same evaluator had conducted both assessments. Therefore, the order of the face-to-face and remote evaluations was randomized for each participant, and different evaluators conducted the evaluations. To reduce the burden on the participants, minimize the practice effect, and avoid temporal changes, the interval between face-to-face and remote evaluations was set at least 2 weeks and no more than 3 months apart.

After completing the 2 assessments, the participants were given a questionnaire to report their satisfaction with the remote assessment. Additionally, given that previous studies showed long waiting periods for accessing child psychiatry specialists, we gathered background information on the waiting period at a specialist's initial meeting. We also considered cases where specialized hospitals and clinics were not within one's living area and collected data on the average travel time for outpatient visits and waiting time for their usual visit.

The 3 evaluators were licensed psychologists in developmental testing who had undergone sufficient training and had confirmed their agreement rates before conducting the evaluations. Evaluators rotated as needed for both in-person and remote assessments, rather than having specific evaluators for each. None of them was involved in the participants' assessments or treatment during their usual visits.

As preliminary training before starting the research, 3 examiners administered the ADHD-RS-IV to 5 patients.

The remote smartphone assessment tool "Curon" by MICIN Co Ltd was used for the telemedicine evaluation. The participants were seated in front of a smartphone in their homes and were introduced to a remote evaluator. Remote evaluators administered their evaluations in a room at the Keio University Hospital using a PC. Although it was assumed that a download and upload environment of 50 Mbps or higher would be perfect for remote video applications, stable communication was achieved at approximately 5 Mbps, which, in most cases, fulfilled the Japanese standard 4G network. The assessments began after confirming that there were no interruptions in the video or audio environment.

Statistical Analysis

Test-retest reliability was assessed using the ICC for continuous variables, with the following standard performance parameters: almost perfect (0.81-1.00); substantial (0.61-0.80); moderate (0.41-0.60); fair (0.21-0.40); and slight (0.0-0.20) [25]. The ICC within the same rater when testing the same patient multiple times was assumed to be 0.90. Assuming that the ICC between the remote and face-to-face tests was 0.8, and to achieve a 95% CI width of 0.3 with a probability of >80%, a sample size of 31 was required. It was also assumed that the analysis would be conducted by dividing the participants into 2 groups: ASD and ADHD. Therefore, the required sample size was set at 62 [26]. Normally distributed data are described as mean (SD). Categorical variables are presented as numbers and percentages. All variables were inspected using histograms, q-q plots, and Kolmogorov-Smirnov tests before statistical analyses were conducted to test the distribution. All analyses were 2-sided with an α value of .05. Statistical analyses were conducted using the SPSS software (version 25.0; SPSS Inc).

Ethical Considerations

The study protocol was approved by the ethics committee of Keio University School of Medicine (20190301). The study was registered in the UMIN Clinical Trial Registry (UMIN000039860). Informed consent was obtained from all caregivers. For the children, an informed assent form was used, tailored to their age and understanding, and written confirmation of their consent was obtained after careful explanation. The personal information collected in this study was securely stored in a locked cabinet located within a lockable room belonging to the corresponding author. The information obtained through this study was anonymized by removing personal identifiers and assigning research numbers at each institution and then was stringently managed and preserved under the supervision of the corresponding author. In the event of an exacerbation of psychiatric symptoms as a result of this study, appropriate treatment was to be administered. No financial compensation was provided.

Results

Participants

The demographic characteristics of the participants are shown in Table 1. A total of 75 patients consented to participate in the study. Overall, 2 participants were from Keio University Hospital, 44 were from Shimada Ryoiku Medical Center for Challenged Children, 13 were from Aiiku Clinic, 10 were from Tokyo Metropolitan Children's Medical Center, and 6 were from Tsurugaoka Garden Hospital. Of these, 1 participant from Shimada Ryoiku Medical Center for Challenged Children dropped out owing to time limitation. Among the 74 remaining participants, 17 (23%) were girls and 57 (77%) were boys. All patients were Asian (Japanese). Their ages ranged from 6 to 16 years, with an average of 10.4 (SD 2.5) years. The primary clinical diagnoses were as follows: 31 (42%) individuals were

XSL•FO
RenderX

diagnosed with ASD, 43 (58%) were diagnosed with ADHD, and 22 (30%) had comorbid ASD and ADHD. Of the 61 caregivers, 47 (77%) had previous experience with remote video calls, whereas 14 (23%) did not.

Table 1. Demographic characteristics.

Characteristics	All patients (n=74)	Patients with ADHD[a] (n=43)	Patients with ASD[b] (n=31)
Sex (female), n (%)	17 (16)	12 (28)	5 (16)
Age (years), mean (SD)	10.4 (2.5)	10.5 (2.4)	10.1 (2.5)
SDQ[c], mean (SD)	21.94 (5.31)	21.09 (5.35)	22.93 (5.16)
AQ[d], mean (SD)	24.68 (7.38)	22.86 (6.72)	27.23 (7.62)
Conners 3, mean (SD)	112.84 (38.22)	108.34 (39.94)	118.53 (35.77)
SSP[e], mean (SD)	70.78 (20.26)	67.44 (20.38)	75.19 (19.55)
CARS-2[f] (face-to-face), mean (SD)	35.97 (4.90)	34.19 (4.49)	38.37 (4.43)
ADHD-RS-IV[g] total (face-to-face), mean (SD)	27.25 (9.88)	26.88 (10.38)	27.77 (9.29)
ADHD-RS-IV hyperactivity/impulsiveness (face-to-face), mean (SD)	8.08 (6.08)	7.95 (6.60)	8.27 (5.39)
ADHD-RS-IV total inattention (face-to-face), mean (SD)	19.17 (5.29)	18.93 (5.28)	19.50 (5.37)

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]ASD: autism spectrum disorder.

[c]SDQ: Strength and Difficulties Questionnaire.

[d]AQ: Autism-Spectrum Quotient.

[e]SSP: Short Sensory Profile.

[f]CARS-2: Childhood Autism Rating Scale-2.

[g]ADHD-RS-IV, Attention-deficit/Hyperactivity Disorder Rating Scale-IV.

Participants were asked how long the waiting period was until the first visit for a child psychiatrist appointment, and the average was found to be 79.0 (SD 57.1; range 1-200) days. Regarding their usual visits, it took 36.8 (SD 20.7; range 5-100) minutes from leaving home to arriving at the hospital. The average waiting time for a usual appointment was 24.0 (SD 14.6; range 5-60) minutes. The potential time that could be saved using remote visits (by adding visiting time × 2 and waiting time) was 97.7 (SD 42.5; range 40-260) minutes.

ICCs of ADHD-RS-IV Between Face-to-Face and Remote Assessments

The ICCs obtained after preliminary training were as follows: 0.987 (95% CI 0.934-0.999; $P<.001$) for the total score; 0.991 (95% CI 0.957-0.999; $P<.001$) for the hyperactivity/impulsivity subscore; and 0.807 (95% CI 0.022-0.978; $P=.02$) for the inattention subscore.

The ICCs are shown in Figures 1-3. The ICC of the total ADHD-RS-IV was 0.769 (95% CI 0.654-0.849; $P<.001$), which was “substantial,” according to the Landis and Koch criteria [25]. The ICC of the ADHD-RS-IV hyperactivity/impulsiveness and inattention subscores were 0.779 (95% CI 0.669-0.856; $P<.001$) and 0.667 (95% CI 0.515-0.778; $P<.001$), respectively, indicating “substantial” agreement. In patients with ADHD as their primary diagnosis, the ICC for total score was 0.816 (95% CI 0.683-0.897; $P<.001$), indicating “almost perfect” agreement; the ICC for hyperactivity/impulsiveness was 0.861 (95% CI 0.756-0.923; $P<.001$), indicating “almost perfect” agreement; and the ICC for inattention score was 0.642 (95% CI 0.423-0.790; $P<.001$), indicating “substantial” agreement. In patients with ASD as their primary diagnosis, the ICC for total score was 0.674 (95% CI 0.420-0.831; $P<.001$), indicating “substantial” agreement; the ICC for hyperactivity/impulsiveness was 0.591 (95% CI 0.299-0.782; $P<.001$), indicating “moderate” agreement; and ICC for inattention score was 0.733 (95% CI 0.511-0.863; $P<.001$), indicating “substantial” agreement.

XSL•FO
RenderX

Figure 1. Intraclass correlations of total Attention-Deficit/Hyperactivity Disorder Rating Scale-IV (ADHD-RS-IV) score between face-to-face and remote assessments in all participants (autism spectrum disorder [ASD] and attention-deficit/hyperactivity disorder [ADHD]).

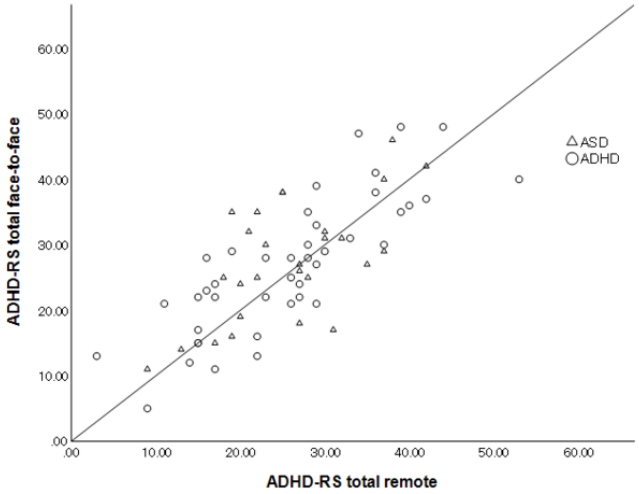

Figure 2. Intraclass correlations of Attention-Deficit/Hyperactivity Disorder Rating Scale-IV (ADHD-RS-IV) hyperactivity/impulsiveness subscore between face-to-face and remote assessments in all participants (autism spectrum disorder [ASD] and attention-deficit/hyperactivity disorder [ADHD]).

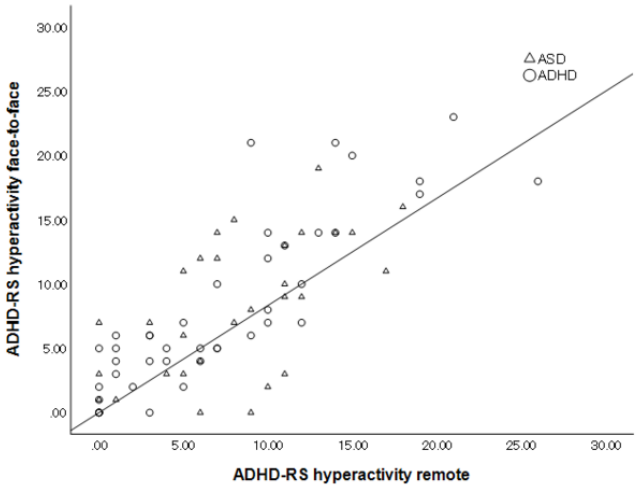

Figure 3. Intraclass correlations of Attention-Deficit/Hyperactivity Disorder Rating Scale-IV (ADHD-RS-IV) inattention subscore between face-to-face and remote assessments in all participants (autism spectrum disorder [ASD] and attention-deficit/hyperactivity disorder [ADHD]).

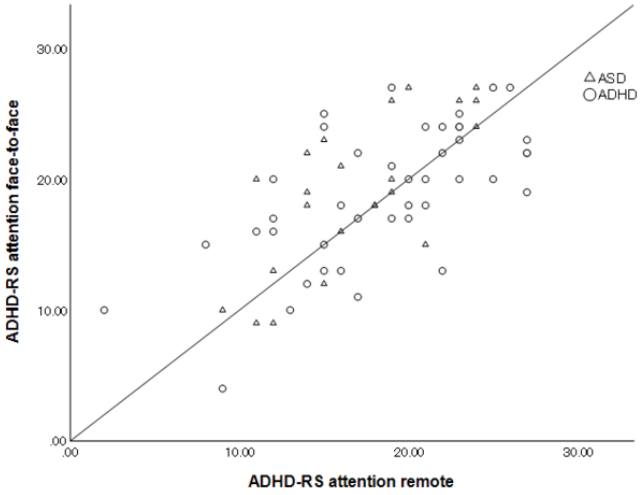

There were no clear sex differences in the ICCs. In regard to age, comparing the older age group (≥11 years; n=37) and the younger age group (≤10 years; n=37), the ICC for total score was 0.675 (95% CI 0.464-0.814; *P*<.001) in the younger age

XSL•FO
RenderX

group and 0.757 (95% CI 0.558-0.873; *P*<.001) in the older age group. The ICC for hyperactivity/impulsiveness was 0.706 (95% CI 0.509-0.833; *P*<.001) in the younger age group and 0.730 (95% CI 0.515-0.858; *P*<.001) in the older age group. The ICC for inattention score was 0.599 (95% CI 0.357-0.766; *P*<.001) in the younger age group and 0.659 (95% CI 0.408-0.818; *P*<.001) in the older age group.

Discussion

Principal Findings

This is the largest study to date on the validation and feasibility of remote ADHD testing. We found good agreement between the remote- and face-to-face–administered ADHD-RS-IV. This study also showed significant potential benefits for children and their caregivers in terms of reducing hospital visits and waiting times.

Additionally, as many as 14 (23%) out of the 61 caregivers who participated in the study had no previous experience with remote video calls; however, this did not pose a significant problem. This can be partly attributed to the staff providing detailed instructions. However, it is also possible that children who had attended internet-based classes at school because of the COVID-19 pandemic were able to provide guidance on the operation of the technology.

The results showed that the agreement rate between the face-to-face and remote assessments was lower in patients with ASD than in those with ADHD, according to the ADHD-RS-IV. One possible reason for the difference in results between patients with ADHD and ASD is the variability in interpretation among caregivers and evaluators. For example, when asked, “Does it seem like the child is not listening when spoken to?” caregivers of children with ASD, who may already have limited social interaction, may readily answer “yes” or “always.” In contrast, they may consider it a symptom of autism rather than inattentiveness related to ADHD and answer “no.” The wide range of interpretations provided by caregivers and evaluators may be a contributing factor [27].

The ADHD-RS-IV results indicated that the inattention ICC scores were numerically lower than the hyperactivity scores. This may be due to the remote assessment environment. At the time of enrollment, a quiet environment was recommended for remote assessments. However, because of the space and density of patients’ homes, situations often arose where there were numerous stimuli, such as the presence of toys or the assessment interview being conducted while other siblings were nearby or in the same room. These circumstances, compared with a controlled hospital consultation room, may have influenced the raters’ impressions of the participants in remote assessments. Moreover, when assessments take place at home, caregivers may feel uncomfortable explaining the severe or negative condition of their children in the presence of their child or other family members. It would be desirable to confirm in advance whether a space can be secured or a time when no other siblings are present can be arranged for remote assessment.

Our results have shown higher agreement rates in the older age group (≥11 years) compared to the younger age group (≤10 years). Symptoms such as hyperactivity and inattention in younger children may be more apparent in their home environment than in controlled settings due to familiarity and less stimulation control, resulting in a slight decrease in agreement with the assessment. Although there may be a benefit to internet-based assessment in regard to observing how the child typically behaves in relaxed homes, older children may be better suited for remote assessment in terms of agreement with face-to-face evaluation.

The potential time-saving effect, considering the time for visits to the hospital and waiting time at the hospital, was found to be 97.7 (SD 42.5; range 40-260) minutes. Given the increase in dual-income households in Japan, which has resulted in less time being spent with children, it would be highly meaningful if we could save this amount of time using remote assessments without reducing quality.

In addition, this research was conducted during the COVID-19 pandemic (the state of emergency was first declared on March 13, 2020, and was removed from the special measures law on May 8, 2023), which may have influenced the results. Although telemedicine was introduced in Japan, it was not widely adopted because of reluctance from the perspective of medical fees. Therefore, in many cases, people had to go to the hospital, facing the risk and anxiety of infection. The importance of having such tools ready to prepare for future outbreaks cannot be overlooked. While the demand for and evidence of telemedicine are expanding, there are also challenges. Issues such as the digital divide, which refers to the disparity that arises between people who can use the internet and computers and those who cannot, and in Japan specifically, the difficulty in widespread adoption due to regulations preventing billing for such services, are notable.

Based on the results of our study, telemedicine may be used under the following conditions: (1) in areas where there is a shortage of medical resources, such as public health and developmental support centers, by collaborating with child psychiatrists, clinical psychologists, and other medical professionals; (2) in central evaluations in clinical trials; and (3) to evaluate treatment effectiveness by combining and complementing face-to-face visits. Although further studies are needed, it may be used to screen children and their parents who are unsure whether to see a specialist and to provide assessment support for clinics without developmental testing capabilities.

Nevertheless, a careful balance must be maintained, as direct in-person assessment provides advantages such as observing nonverbal cues and behaviors, which may be important for a comprehensive understanding of the child’s condition.

Limitations

This study had several limitations. First, it was limited to children who had already been diagnosed and had received medical care and treatment. Therefore, these results do not apply to undiagnosed neurotypes. This study aimed to explore whether remote assessment tools could be helpful when children with developmental issues and their caregivers seek medical assistance. Therefore, the study design did not encompass typically developing children. Nonetheless, ICC, which usually

XSL•FO
RenderX

tends to exhibit higher values when there is a diverse range of patient scores, manifested a relatively high degree of agreement, specifically in the group of children affected in this study. This can be interpreted as endorsing the validity of remote assessment procedures in this context. Second, although our study effectively indicated the potential for high-accuracy remote assessments, this does not necessarily guarantee a remote diagnosis. However, considering the frequent use of the ADHD-RS-IV in the diagnostic process for ADHD, the fact that we have demonstrated its robustness in the context of remote assessment may suggest its utility for future diagnoses. Even if remote ADHD-RS-IV assessments do not replace diagnosis during the first visit, they can be used to identify individuals who should be prioritized for early assessment by conducting severity evaluations and triages, thus expediting their initial consultation. To compensate for the weaknesses of this study, future studies should focus on examining the congruence between severity assessments and diagnoses conducted remotely in comparison with in-person evaluations, in addition to evaluating the efficacy of remote methods across diverse subpopulations.

Comparison With Previous Work

There have been few studies comparing ADHD assessment scales in face-to-face and remote settings. In previous research in the field of neurodevelopmental disorders, there is a report on the usability and reliability of the Autism Diagnostic Observation Schedule conducted face-to-face and remotely with 23 adults with ASD [16]. The ICC between face-to-face and remote was high at 0.92. In their report, technicians were present in the room during the remote assessment to provide assistance with technical operations, whereas in this study, we made it possible for participants to operate the equipment themselves at home without the help of technicians. Additionally, in that study, the same examiner conducted the tests in both face-to-face and remote settings for all cases. In contrast, in our study, different examiners conducted the tests in face-to-face and remote settings. Although it is difficult to compare because the disorder and assessment tools are different, considering these factors, the ICC value of our results is comparable, indicating a new finding in this field.

Conclusions

The results of this study showed that developmental assessments can be conducted with the same level of accuracy using remote tools as compared to face-to-face assessments. This means that even medical institutions where specialized assessments are not available, as well as health care centers, can benefit from these assessments, thereby improving the convenience for children who require early detection and intervention. Future research is needed to investigate the consistency of remote assessments and diagnoses compared with the initial face-to-face examination as well as the effectiveness of remote examinations in various subpopulations.

Acknowledgments

This study was supported in part by the Japan Science and Technology Agency Program on Open Innovation Platform With Enterprises, Research Institute, and Academia (JST-OPERA; grant JPMJOP1842; Principal Investigator Hiroaki Miyata) and by MICIN Inc, Tokyo, Japan.

Conflicts of Interest

KN has received speaker's honoraria from Takeda, Janssen, Eli Lilly, Eisai, MSD, Otsuka, and Shionogi. MK has received honoraria from Takeda and Shionogi. TK has received consultant fees from Dainippon Sumitomo, Novartis, and Otsuka; speaker's honoraria from Banyu, Eli Lilly, Dainippon Sumitomo, Janssen, MSD, Novartis, Otsuka, and Pfizer; and grant support from Takeda, Dainippon-Sumitomo, and Otsuka. The remaining authors have no conflicts of interest to declare.

References

1. Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, et al. Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics 2010;125(1):e17-e23 [FREE Full text] [doi: 10.1542/peds.2009-0958] [Medline: 19948568]
2. Kamio Y, Inada N, Koyama T. A nationwide survey on quality of life and associated factors of adults with high-functioning autism spectrum disorders. Autism 2013;17(1):15-26. [doi: 10.1177/1362361312436848] [Medline: 22399449]
3. Halperin JM, Marks DJ. Practitioner review: assessment and treatment of preschool children with attention-deficit/hyperactivity disorder. J Child Psychol Psychiatry 2019;60(9):930-943. [doi: 10.1111/jcpp.13014] [Medline: 30690737]
4. Zablotsky B, Black LI, Maenner MJ, Schieve LA, Danielson ML, Bitsko RH, et al. Prevalence and trends of developmental disabilities among children in the United States: 2009-2017. Pediatrics 2019;144(4):e20190811 [FREE Full text] [doi: 10.1542/peds.2019-0811] [Medline: 31558576]
5. Nakamura S, Ohnishi M, Uchiyama S. Epidemiological survey of adult attention deficit hyperactivity disorder (ADHD) in Japan. Jpn J Psychiatr Treat 2013;28:155-162.
6. Penner M, Anagnostou E, Ungar WJ. Practice patterns and determinants of wait time for autism spectrum disorder diagnosis in Canada. Mol Autism 2018;9:16 [FREE Full text] [doi: 10.1186/s13229-018-0201-0] [Medline: 29541438]

XSL•FO
RenderX

7. McKenzie K, Forsyth K, O'Hare A, McClure I, Rutherford M, Murray A, et al. Factors influencing waiting times for diagnosis of autism spectrum disorder in children and adults. Res Dev Disabil 2015;45-46:300-306. [doi: 10.1016/j.ridd.2015.07.033] [Medline: 26280693]
8. McCarthy S, Asherson P, Coghill D, Hollis C, Murray M, Potts L, et al. Attention-deficit hyperactivity disorder: treatment discontinuation in adolescents and young adults. Br J Psychiatry 2009;194(3):273-277 [FREE Full text] [doi: 10.1192/bjp.bp.107.045245] [Medline: 19252159]
9. Nomura K, Tarumi R, Yoshida K, Sado M, Suzuki T, Mimura M, et al. Cancellation of outpatient appointments in patients with attention-deficit/hyperactivity disorder. PLoS One 2021;16(11):e0260431 [FREE Full text] [doi: 10.1371/journal.pone.0260431] [Medline: 34797891]
10. National Institute for Health and Care Excellence (UK). Attention Deficit Hyperactivity Disorder: Diagnosis and Management. London: National Institute for Health and Care Excellence (NICE); 2019.
11. Pliszka S, AACAP Work Group on Quality Issues. Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 2007;46(7):894-921 [FREE Full text] [doi: 10.1097/chi.0b013e318054e724] [Medline: 17581453]
12. Stevens J, Quittner AL, Abikoff H. Factors influencing elementary school teachers' ratings of ADHD and ODD behaviors. J Clin Child Psychol 1998;27(4):406-414. [doi: 10.1207/s15374424jccp2704_4] [Medline: 9866077]
13. Simonoff E, Pickles A, Hervas A, Silberg JL, Rutter M, Eaves L. Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction. Psychol Med 1998;28(4):825-837. [doi: 10.1017/s0033291798006886] [Medline: 9723139]
14. Kobak KA, Leuchter A, DeBrota D, Engelhardt N, Williams JBW, Cook IA, et al. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. J Clin Psychopharmacol 2010;30(2):193-197. [doi: 10.1097/JCP.0b013e3181d20912] [Medline: 20520295]
15. Kane JM, Robinson DG, Schooler NR, Mueser KT, Penn DL, Rosenheck RA, et al. Comprehensive versus usual community care for first-episode psychosis: 2-year outcomes from the NIMH RAISE Early Treatment Program. Am J Psychiatry 2016;173(4):362-372 [FREE Full text] [doi: 10.1176/appi.ajp.2015.15050632] [Medline: 26481174]
16. Ward-King J, Cohen IL, Penning H, Holden JJA. Brief report: telephone administration of the autism diagnostic interview—revised: reliability and suitability for use in research. J Autism Dev Disord 2010;40(10):1285-1290. [doi: 10.1007/s10803-010-0987-x] [Medline: 20195733]
17. Schutte JL, McCue MP, Parmanto B, McGonigle J, Handen B, Lewis A, et al. Usability and reliability of a remotely administered adult autism assessment, the Autism Diagnostic Observation Schedule (ADOS) Module 4. Telemed J E Health 2015;21(3):176-184 [FREE Full text] [doi: 10.1089/tmj.2014.0011] [Medline: 25569603]
18. Smith CJ, Rozga A, Matthews N, Oberleitner R, Nazneen N, Abowd G. Investigating the accuracy of a novel telehealth diagnostic approach for autism spectrum disorder. Psychol Assess 2017;29(3):245-252 [FREE Full text] [doi: 10.1037/pas0000317] [Medline: 27196689]
19. DuPaul G, Power TJ, Anastopoulos AD, Reid R. ADHD Rating Scale-IV : Checklists, Norms, and Clinical Interpretation. New York: Guilford Press; 1998.
20. Tani I, Okada R, Ohnishi M, Nakajima S, Tsujii M. Japanese version of home form of the ADHD-RS: an evaluation of its reliability and validity. Res Dev Disabil 2010;31(6):1426-1433. [doi: 10.1016/j.ridd.2010.06.016] [Medline: 20638822]
21. McGough JJ, Sturm A, Cowen J, Tung K, Salgari GC, Leuchter AF, et al. Double-blind, sham-controlled, pilot study of trigeminal nerve stimulation for attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 2019;58(4):403-411.e3 [FREE Full text] [doi: 10.1016/j.jaac.2018.11.013] [Medline: 30768393]
22. Riggs PD, Winhusen T, Davies RD, Leimberger JD, Mikulich-Gilbertson S, Klein C, et al. Randomized controlled trial of osmotic-release methylphenidate with cognitive-behavioral therapy in adolescents with attention-deficit/hyperactivity disorder and substance use disorders. J Am Acad Child Adolesc Psychiatry 2011;50(9):903-914 [FREE Full text] [doi: 10.1016/j.jaac.2011.06.010] [Medline: 21871372]
23. Takayanagi N, Yoshida S, Yasuda S, Adachi M, Kaneda-Osato A, Tanaka M, et al. Psychometric properties of the Japanese ADHD-RS in preschool children. Res Dev Disabil 2016;55:268-278. [doi: 10.1016/j.ridd.2016.05.002] [Medline: 27164481]
24. American Psychiatric Association, American Psychiatric Association DSM-5 Task Force. Diagnostic and statistical manual of mental disorders DSM-5, 5th Edition. Washington, DC: American Psychiatric Association; 2013.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]
26. Doros G, Lew R. Design based on intra-class correlation coefficients. Curr J Biostat 2010;1(1):1-8. [doi: 10.3844/amjbsp.2010.1.8]
27. Grzadzinski R, Dick C, Lord C, Bishop S. Parent-reported and clinician-observed autism spectrum disorder (ASD) symptoms in children with attention deficit/hyperactivity disorder (ADHD): implications for practice under DSM-5. Mol Autism 2016;7:7 [FREE Full text] [doi: 10.1186/s13229-016-0072-1] [Medline: 26788284]

XSL•FO
RenderX

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ADHD-RS-IV: Attention-deficit/Hyperactivity Disorder Rating Scale-IV
ASD: autism spectrum disorder
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ICC: intraclass correlation coefficient

Edited by T de Azevedo Cardoso, Y Hong; submitted 10.08.23; peer-reviewed by T Ewais, J Hamid, E Ikefuama; comments to author 30.11.23; revised version received 13.12.23; accepted 14.12.23; published 19.02.24.

Please cite as:
Kurokawa S, Nomura K, Hosogane N, Nagasawa T, Kawade Y, Matsumoto Y, Morinaga S, Kaise Y, Higuchi A, Goto A, Inada N, Kodaira M, Kishimoto T
Reliability of Telepsychiatry Assessments Using the Attention-Deficit/Hyperactivity Disorder Rating Scale-IV for Children With Neurodevelopmental Disorders and Their Caregivers: Randomized Feasibility Study
J Med Internet Res 2024;26:e51749
URL: https://www.jmir.org/2024/1/e51749
doi:10.2196/51749
PMID:38373022

©Shunya Kurokawa, Kensuke Nomura, Nana Hosogane, Takashi Nagasawa, Yuko Kawade, Yu Matsumoto, Shuichi Morinaga, Yuriko Kaise, Ayana Higuchi, Akiko Goto, Naoko Inada, Masaki Kodaira, Taishiro Kishimoto. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Wait Times for Scheduling Appointments for Prevention of Macrovascular and Microvascular Complications of Diabetes: Cross-Sectional Descriptive Study

Corey H Basch[1], EdD, MPH; Grace C Hillyer[2], EdD, MPH; Charles E Basch[3], PhD

[1]Department of Public Health, William Paterson University, Wayne, NJ, United States

[2]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, United States

[3]Department of Health Studies and Applied Educational Psychology, Teachers College Columbia University, New York, NY, United States

Corresponding Author:
Corey H Basch, EdD, MPH
Department of Public Health
William Paterson University
University Hall
Wayne, NJ, 07470
United States
Phone: 1 973 720 2603
Email: baschc@wpunj.edu

Abstract

Background: Diabetes is a chronic disease that requires lifelong management and care, affecting around 422 million people worldwide and roughly 37 million in the United States. Patients newly diagnosed with diabetes must work with health care providers to formulate a management plan, including lifestyle modifications and regular office visits, to improve metabolic control, prevent or delay complications, optimize quality of life, and promote well-being.

Objective: Our aim is to investigate one component of system-wide access to timely health care for people with diabetes in New York City (NYC), namely the length of time for someone with newly diagnosed diabetes to obtain an appointment with 3 diabetes care specialists: a cardiologist, an endocrinologist, and an ophthalmologist, respectively.

Methods: We contacted the offices of 3 different kinds of specialists: cardiologists, endocrinologists, and ophthalmologists, by telephone, for this descriptive cross-sectional study, to determine the number of days required to schedule an appointment for a new patient with diabetes. The sampling frame included all specialists affiliated with any private or public hospital in NYC. The number of days to obtain an appointment with each specialist was documented, along with “time on hold” when attempting to schedule an appointment and the presence of online booking capabilities.

Results: Of the 1639 unique physicians affiliated with (private and public) hospitals in the 3 subspecialties, 1032 (cardiologists, endocrinologists, and ophthalmologists) were in active practice and did not require a referral. The mean wait time for scheduling an appointment was 36 (SD 36.4; IQR 12-51.5) days for cardiologists; 82 (SD 47; IQR 56-101) days for endocrinologists; and 50.4 (SD 56; IQR 10-72) days for ophthalmologists. The median wait time was 27 days for cardiologists, 72 days for endocrinologists, and 30 days for ophthalmologists. The mean time on hold while attempting to schedule an appointment with these specialists was 2.6 (SD 5.5) minutes for cardiologists, 5.4 (SD 4.3) minutes for endocrinologists, and 3.2 (SD 4.8) minutes for ophthalmologists, respectively. Over 46% (158/341) of cardiologists enabled patients to schedule an appointment on the web, and over 55% (128/228) of endocrinologists enabled patients to schedule an appointment on the web. In contrast, only approximately 25% (117/463) of ophthalmologists offered web-based appointment scheduling options.

Conclusions: The results indicate considerable variation in wait times between and within the 3 specialties examined for a new patient in NYC. Given the paucity of research on wait times for newly diagnosed people with diabetes to obtain an appointment with different specialists, this study provides preliminary estimates that can serve as an initial reference. Additional research is needed to document the extent to which wait times are associated with complications and the demographic and socio-economic characteristics of people served by different providers.

(J Med Internet Res 2024;26:e55351) doi:10.2196/55351

XSL•FO
RenderX

KEYWORDS

appointment scheduling; cardiologists; chronic disease; cross-sectional study; diabetes; endocrinologists; healthcare utilization; lifestyle modification; management plan; metabolic control; online booking; ophthalmologists; prevention; quality of life; scheduling appointments; scheduling; specialist; timeliness; United States; wait time; well-being

Introduction

Diabetes is a chronic disease that requires lifelong management and care, affecting around 422 million people worldwide [1] and roughly 37 million in the United States (1 in 10 Americans) [2]. The 3 main kinds of diabetes are type 1, an autoimmune disease in which the body stops producing insulin; type 2, where the body stops using insulin well and can no longer regulate normal glucose levels; and gestational diabetes, which occurs during pregnancy. In the United States, approximately 90%-95% of people affected have type 2 diabetes [2], which is the most expensive chronic condition [3], costing an estimated US $327 billion in 2017.

Patients newly diagnosed with diabetes must work with health care providers to formulate a management plan, including lifestyle modifications and regular office visits, to improve metabolic control and associated long-term disease outcomes, prevent or delay complications, optimize quality of life, and promote well-being. In addition to attaining stable metabolic control through lifestyle changes, diabetes self-care clinical management also includes weight management and cardiovascular and retinopathy risk monitoring [4]. According to the Standards of Care in Diabetes 2023, treatment decisions must be timely, include social support, and be formulated collaboratively between patients and their health care providers, with consideration of the individual's preferences, prognoses, comorbidities, and financial factors [5]. Defined as "the system's capacity to provide care quickly after a need is recognized [6], timeliness in health care is associated with insurance type [7,8] and a lack thereof has been shown to delay treatment initiation [7,9,10].

Therapeutic inertia, or "lack of timely adjustment to therapy when a patient's treatment goals are not met" [11], can be dire, yet research demonstrates it remains prevalent in Western countries [12]. Delays in treatment are known to cause more severe complications [13-16], including hypertension and cardiovascular disease, nonperipheral and peripheral retinopathy, nephropathy, and peripheral vascular diseases [10,17]. Unintentional delays in diagnosis further complicate screening and treatment and are medically damaging [18].

Timely care begins with initial appointments at the time of diagnosis. There are gaps between high-quality health care and what patients actually receive, and more timely care is one of several key focal points for improvement [4]. Overcoming population-wide delays in assessment, monitoring, and treatment is complex and involves interactions among patients, health care providers, and the health care system [19,20]. Effective interventions to decrease hyperglycemia and improve glycemic control remain a major challenge [21-23]. Therefore, researchers suggest that monitoring systems that provide a holistic view of the quality of care are necessary to combat this issue [20].

The purpose of this study was to investigate one component of system-wide access to timely health care for people with diabetes in New York City (NYC), namely the length of time for someone with newly diagnosed diabetes to obtain an appointment with 3 diabetes care specialists: a cardiologist, an endocrinologist, and an ophthalmologist, respectively. As type 1 and type 2 diabetes require similar specialist management, we do not differentiate between the types but refer to diabetes as a single entity for the purposes of this study. The sampling frame included all specialists affiliated with any private or public hospital in NYC in 2022. We documented the number of days to obtain an appointment with each specialist. Secondary objectives were to describe "time on hold" when attempting to schedule an appointment and the presence of web-based booking capabilities.

Methods

Design

The design of this study was cross-sectional, and the study was limited to a descriptive analysis.

Study Setting

The offices of three different kinds of specialists—cardiologists, endocrinologists, and ophthalmologists—in NYC were contacted by telephone to determine the number of days required to schedule an appointment for a new patient with diabetes.

Sampling, Eligibility Criteria, and Data Collection

The sample frame was cardiologists, endocrinologists, and ophthalmologists, listed on the websites of all accredited private or public hospitals in NYC in the fall of 2022.

Each website was reviewed, and contact information for each provider was excerpted into a separate file. Each office telephone number listed on the website was called to determine the following information: whether the physician was accepting new patients, type of insurance accepted (private, public, and none), and number of days for a person with newly diagnosed diabetes to obtain an appointment.

Exclusion criteria included physicians who were not accepting new patients, no longer worked at the hospital, those who only saw specific types of patients (eg, pediatrics), and those who did not have a working phone number. Furthermore, when a doctor was listed at multiple locations, a random number generator was used to randomly select only one office to include. Time on hold while scheduling appointments and availability of web-based booking ("ZocDoc" or the hospital website) were noted, as well as the type of insurance accepted by providers (Medicaid, commercial, or none).

Statistical Analysis

Of the 1639 unique physicians affiliated with (private and public) hospitals in the three subspecialties, a total of 1032 (341

XSL•FO
RenderX

of 670 cardiologists, 228 of 315 endocrinologists, and 463 of 654 ophthalmologists) were in active practice and did not require a referral (Figure 1). For each subspecialty, we determined the mean (SD), IQR, and total range for the continuous variables: the number of days until the first available appointment and the time to reach someone at the office on the telephone to make an appointment. Categorical variables included types of insurance accepted and whether it was possible to schedule an appointment through ZocDoc (a NYC-based web-based appointment booking service) or on the web at the hospital website. All analyses were performed using SPSS (version 28; IBM Corp) [21].

Figure 1. Diabetes specialists affiliated with New York City (NYC) hospitals.

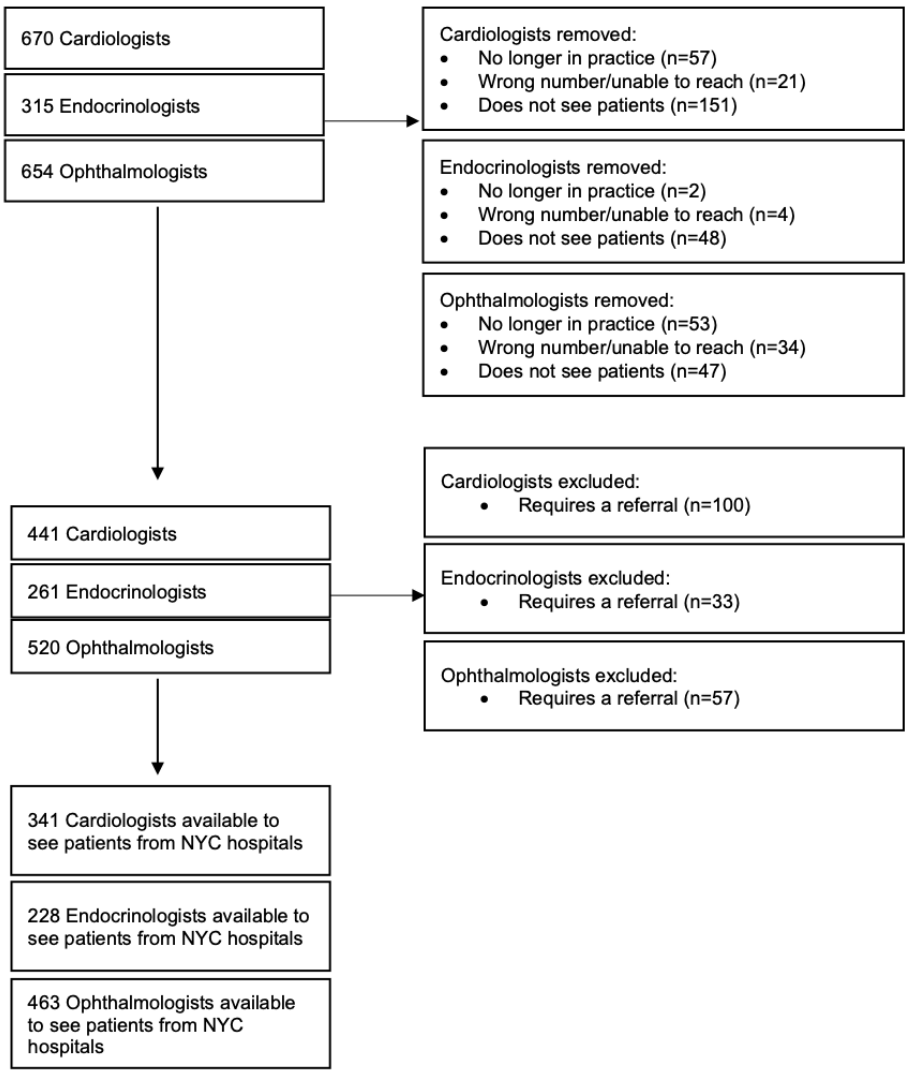

Ethical Considerations

This study was approved by the institutional review board at Teachers College, Columbia University (protocol #23-029). This study did not meet the definition of human participant research for the institutional review board at William Paterson University.

Results

The mean wait time for scheduling an appointment was 36 (SD 36 days; IQR 12.0-51.5) days for cardiologists; 82 (SD 47; IQR 56.0-101.0) days for endocrinologists; and 50.4 (SD 56; IQR 10.0-72.0) days for ophthalmologists (Table 1). The median wait time was 27.0 days for cardiologists, 72.0 days for endocrinologists, and 30.0 days for ophthalmologists. The mean time on hold while attempting to schedule an appointment with these specialists was 2.6 (SD 5.5) minutes for cardiologists, 5.4 (SD 4.3) minutes for endocrinologists, and 3.2 (SD 4.8) minutes for ophthalmologists, respectively (Table 1). Over 46% (158/341) of cardiologists enabled patients to schedule an appointment on the web (38/341, 11.1% on ZocDoc and 120/341, 35.2% on their website), and over 56% (128/228) of endocrinologists enabled patients to schedule an appointment on the web (52/228, 22.8% on ZocDoc and 76/228, 33.3% on their website). In contrast, only approximately 25% (117/463) of ophthalmologists offered web-based appointment scheduling options (43/463, 9.3% on ZocDoc and 74/463, 16.0% on their website).

XSL•FO
RenderX

Table 1. Characteristics of timeliness of care among New York City specialists (N=1032).

	Cardiologists (n=341)	Endocrinologists (n=228)	Ophthalmologists (n=463)
Time on hold (minutes)			
Mean (SD)	2.6 (5.5)	5.4 (4.3)	3.2 (4.8)
Range	<1-48	<1-24.2	<1-66
Time until appointment (days)			
Mean (SD)	36 (36)	82 (47)	50.4 (56)
Median (IQR)	27.0 (12.0-51.5)	72.0 (56.0-101.0)	30.0 (10.0-72.0)
Insurance accepted, n (%)			
Private only	9 (2.6)	14 (6.1)	5 (1.1)
Private and Medicare	44 (12.9)	79 (34.6)	108 (23.3)
Private, Medicare, and Medicaid	285 (83.6)	131 (57.5)	349 (75.4)
None	3 (0.9)	4 (1.8)	1 (0.2)
Appointments available on ZocDoc, n (%)	38 (11.1)	52 (22.8)	43 (9.3)
Appointments available on website, n (%)	120 (35.2)	76 (33.3)	74 (16.0)

Discussion

Principal Results

The importance of early detection and treatment of macrovascular and microvascular complications caused by diabetes results in considerable personal and societal costs [22,23]. Living with diabetes requires considerable self-management on a daily basis, as well as scheduling multiple clinical appointments for screening and checkups. Because of the increased risks of macrovascular and microvascular complications, diabetes is the most costly chronic condition in the United States [24], with some estimates stating that 25% of all US health care costs are attributed to the disease [25]. Timely access to health care and education at an early stage in the natural history of the disease can help people with diabetes learn self-management knowledge and skills and delay the incidence of complications [5,26]. Surprisingly, we did not identify any published studies in peer-reviewed biomedical journals that documented the time required for a new patient to obtain an appointment with each of the 3 main clinical specialists who care for people with diabetes: cardiologists, endocrinologists, and ophthalmologists.

The results from this study provide an initial estimate of the wait times required for a new patient to see each of the 3 specialists in NYC. The results indicate considerable variation in wait times both between and within the 3 specialties examined. For example, compared with the mean wait time for a new patient to see a cardiologist, the mean wait time to see an endocrinologist was more than twice as long (36 days vs 82 days), with the mean wait time for seeing an ophthalmologist falling in between (approximately 50 days). The 25% of cardiologists with the shortest wait times were 12 or fewer days, while the 25% with the longest wait times were 51.5 days or longer. For endocrinologists, the 25% with the shortest wait times were 56 or fewer days, while the 25% with the longest wait times were 101 days or longer. The largest IQR in wait times was observed for ophthalmologists; the 25% with the shortest wait times was 10 or fewer days, while the 25% with the longest wait times was 72 days or longer.

While the concept of surveillance may have a negative connotation from socio-political perspectives, surveillance is one of the main public health strategies used for disease prevention and control. By identifying subgroups within the population at high risk, scarce resources can be directed in an appropriate and efficient way. Disease surveillance, therefore, plays a prominent role in national disease prevention and health promotion efforts, including documentation of causes of death, comorbidity, disability, the use and cost of health care, and a wide range of preventive health behaviors [27,28]. Nevertheless, we did not identify any published studies in peer-reviewed biomedical literature regarding wait times for medical appointments, a gap in current knowledge this study aimed to address.

Limitations

As this was a cross-sectional study, the findings cannot be generalized over time. There were some inaccuracies on the web pages used to construct the sampling frame (including physicians listed who did not provide clinical care or were no longer employed at the respective clinical setting). This in itself is an important finding because it shows the frustration a person with diabetes may face in attempting to schedule an appointment with a provider who is not available. Another limitation is that the data were based on reports from office staff, which may not have been accurate. Nevertheless, this simulates the kinds of information that would have been available to prospective patients calling the respective offices. The doctors included in this study were all affiliated with a hospital in NYC, which provides an incomplete sampling frame of all practices.

Conclusions

This study was focused on health care providers addressing cardiology (macrovascular complications), endocrinology (metabolic control), and ophthalmology (microvascular complications causing diabetic retinopathy) [29]. These 3

XSL•FO
RenderX

aspects of care influence the natural history of diabetes and its major complications. Endocrine-level control of metabolism is paramount for those with diabetes, as it mediates the progression of disease [29]. Poor metabolic control over time causes a wide range of negative clinical consequences [29], including increased risk for heart disease and heart failure [30], which is the leading cause of death in the United States, as well as increased risk, progression, and severity of diabetic retinopathy, the leading cause of nontraumatic blindness among working age adults [30,31]. Both cardiovascular disease and diabetic retinopathy can be treated effectively in the early stages; however, once clinical signs and symptoms are apparent, treatment is less effective [29,30].

Given the paucity of research on wait times for newly diagnosed people with diabetes to obtain an appointment with different specialists, this study provides preliminary estimates that can serve as an initial reference. The geographic scope of this study was limited to NYC, which has a far greater concentration of clinical specialists than many other geographic areas [32]. It is reasonable to expect that wait times may be considerably longer in areas with fewer providers. However, this requires confirmation since providers in some areas may offer remote accessibility. Those with the greatest needs and at the highest risk of experiencing complications from diabetes tend to have the least access to timely health care services [33-35]. While time spent waiting for an appointment is relevant for everyone in the population, it is especially important for people who require multiple annual appointments [36,37]. Additional research is needed to document the extent to which wait times are associated with complications and the demographic and socio-economic characteristics of people served by different providers.

Acknowledgments

The authors would like to thank Erin Jacques, Allison Kincaid, and Helen Yousaf. This research was funded in part by the William Paterson Center for Research Grant Program.

Data Availability

The data sets generated during and analyzed during this study are available upon reasonable request.

Conflicts of Interest

CHB serves as an Editorial Board Member for JMIR; she did not have a role in the review or editorial process for this article. All other authors declare no conflicts of interest.

References

1. Diabetes. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/diabetes [accessed 2024-02-16]
2. What is diabetes? Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/diabetes/basics/diabetes.html [accessed 2024-02-16]
3. Power of prevention: health and economic benefits of diabetes interventions. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/chronicdisease/programs-impact/pop/diabetes.htm [accessed 2024-02-16]
4. American Diabetes Association. Standards of medical care in diabetes-2022 abridged for primary care providers. Clin Diabetes 2022;40(1):10-38 [FREE Full text] [doi: 10.2337/cd22-as01] [Medline: 35221470]
5. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 1. Improving care and promoting health in populations: standards of care in diabetes-2023. Diabetes Care 2023;46(Supple 1):S10-S18 [FREE Full text] [doi: 10.2337/dc23-S001] [Medline: 36507639]
6. Chartbook on access to health care: elements of access to health care: timeliness. Agency for Healthcare Research and Quality. 2018. URL: https://tinyurl.com/2z5ddtmv [accessed 2024-02-16]
7. Oostrom T, Einav L, Finkelstein A. Outpatient office wait times and quality of care for medicaid patients. Health Aff (Millwood) 2017;36(5):826-832 [FREE Full text] [doi: 10.1377/hlthaff.2016.1478] [Medline: 28461348]
8. Ibrahim H, Harhara T, Athar S, Nair SC, Kamour AM. Multi-Disciplinary discharge coordination team to overcome discharge barriers and address the risk of delayed discharges. Risk Manag Healthc Policy 2022;15:141-149 [FREE Full text] [doi: 10.2147/RMHP.S347693] [Medline: 35140535]
9. Halpern MT, Holden DJ. Disparities in timeliness of care for U.S. medicare patients diagnosed with cancer. Curr Oncol 2012;19(6):e404-e413 [FREE Full text] [doi: 10.3747/co.19.1073] [Medline: 23300364]
10. American Diabetes Association Professional Practice Committee. 10. Cardiovascular disease and risk management: standards of medical care in diabetes-2022. Diabetes Care 2022;45(Suppl 1):S144-S174 [FREE Full text] [doi: 10.2337/dc22-S010] [Medline: 34964815]
11. Overcoming therapeutic inertia. American Diabetes Association. 2023. URL: https://www.therapeuticinertia.diabetes.org/ [accessed 2024-02-16]

XSL•FO
RenderX

12. Khunti K, Gomes MB, Pocock S, Shestakova MV, Pintat S, Fenici P, et al. Therapeutic inertia in the treatment of hyperglycaemia in patients with type 2 diabetes: a systematic review. Diabetes Obes Metab 2018;20(2):427-437 [FREE Full text] [doi: 10.1111/dom.13088] [Medline: 28834075]
13. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HAW. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med 2008;359(15):1577-1589 [FREE Full text] [doi: 10.1056/NEJMoa0806470] [Medline: 18784090]
14. Shah BR, Hux JE, Laupacis A, Zinman B, van Walraven C. Clinical inertia in response to inadequate glycemic control: do specialists differ from primary care physicians? Diabetes Care 2005;28(3):600-606 [FREE Full text] [doi: 10.2337/diacare.28.3.600] [Medline: 15735195]
15. Fu AZ, Qiu Y, Davies MJ, Radican L, Engel SS. Treatment intensification in patients with type 2 diabetes who failed metformin monotherapy. Diabetes Obes Metab 2011;13(8):765-769 [FREE Full text] [doi: 10.1111/j.1463-1326.2011.01405.x] [Medline: 21457427]
16. Khunti K, Damci T, Meneghini L, Pan CY, Yale JF. Study of once daily levemir (SOLVE): insights into the timing of insulin initiation in people with poorly controlled type 2 diabetes in routine clinical practice. Diabetes Obes Metab 2012;14(7):654-661. [doi: 10.1111/j.1463-1326.2012.01602.x] [Medline: 22443213]
17. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000;321(7258):405-412 [FREE Full text] [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
18. Institute FHI. Improvement stories: across the chasm aim #4: health care should be timely. Institute for Healthcare Improvement. 2023. URL: https://tinyurl.com/yjw7fpur [accessed 2024-02-16]
19. Allen JD, Curtiss FR, Fairman KA. Nonadherence, clinical inertia, or therapeutic inertia? J Manag Care Pharm 2009;15(8):690-695 [FREE Full text] [doi: 10.18553/jmcp.2009.15.8.690] [Medline: 19803559]
20. Andreozzi F, Candido R, Corrao S, Fornengo R, Giancaterini A, Ponzani P, et al. Clinical inertia is the enemy of therapeutic success in the management of diabetes and its complications: a narrative literature review. Diabetol Metab Syndr 2020;12:52 [FREE Full text] [doi: 10.1186/s13098-020-00559-7] [Medline: 32565924]
21. IBM SPSS statistics for windows, version 28.0. IBM Corp. Armonk, NY: IBM Corp URL: https://tinyurl.com/4pu9yrsv [accessed 2024-02-16]
22. Diabetes complications are costly in older adults. Centers for Disease Control and Prevention. 2022. URL: https://tinyurl.com/2cp6sne7 [accessed 2024-02-16]
23. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
24. Dieleman JL, Baral R, Birger M, Bui AL, Bulchis A, Chapin A, et al. US spending on personal health care and public health, 1996-2013. JAMA 2016;316(24):2627-2646 [FREE Full text] [doi: 10.1001/jama.2016.16885] [Medline: 28027366]
25. New American Diabetes Association report finds annual costs of diabetes to be $412.9 billion. American Diabetes Association. 2023. URL: https://tinyurl.com/ynfz5h7v [accessed 2024-02-16]
26. Mendez I, Lundeen EA, Saunders M, Williams A, Saaddine J, Albright A. Diabetes self-management education and association with diabetes self-care and clinical preventive care practices. Sci Diabetes Self Manag Care 2022;48(1):23-34 [FREE Full text] [doi: 10.1177/26350106211065378] [Medline: 35023406]
27. About the national health and nutrition examination survey. Centers for Disease Control and Prevention, National Center for Health Statistics. 2023. URL: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm [accessed 2024-02-16]
28. About behavioral risk factor surveillance system. Centers for Disease Control Prevention. 2014. URL: https://www.cdc.gov/brfss/about/index.htm [accessed 2024-02-16]
29. Fowler MJ. Microvascular and macrovascular complications of diabetes. Clin Diabetes 2008;26(2):77-82 [FREE Full text] [doi: 10.2337/diaclin.26.2.77]
30. Cardiovascular disease and diabetes. American Heart Association. 2022. URL: https://tinyurl.com/bdfdb8n6 [accessed 2024-02-16]
31. Diabetes and vision loss. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/diabetes/managing/diabetes-vision-loss.html [accessed 2024-02-16]
32. Number of active physicians in New York in 2023, by specialty area. Statista. 2023. URL: https://tinyurl.com/3u5hbwkb [accessed 2024-02-16]
33. Hsu CC, Lee CH, Wahlqvist ML, Huang HL, Chang HY, Chen L, et al. Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. Diabetes Care 2012;35(11):2286-2292 [FREE Full text] [doi: 10.2337/dc11-2052] [Medline: 22912425]
34. Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, et al. Social determinants of health and diabetes: a scientific review. Diabetes Care 2020;44(1):258-279 [FREE Full text] [doi: 10.2337/dci20-0053] [Medline: 33139407]
35. Zhang X, Bullard KM, Gregg EW, Beckles GL, Williams DE, Barker LE, et al. Access to health care and control of ABCs of diabetes. Diabetes Care 2012;35(7):1566-1571 [FREE Full text] [doi: 10.2337/dc12-0081] [Medline: 22522664]

XSL•FO
RenderX

36. Pizer SD, Prentice JC. What are the consequences of waiting for health care in the veteran population? J Gen Intern Med 2011;26(Suppl 2):676-682 [FREE Full text] [doi: 10.1007/s11606-011-1819-1] [Medline: 21989621]
37. Prentice JC, Pizer SD. Delayed access to health care and mortality. Health Serv Res 2007;42(2):644-662 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00626.x] [Medline: 17362211]

Abbreviations

NYC: New York City

Edited by A Mavragani; submitted 10.12.23; peer-reviewed by M Amrollahi-Sharifabadi, E Popa; comments to author 24.01.24; revised version received 30.01.24; accepted 13.02.24; published 26.03.24.

Please cite as:
Basch CH, Hillyer GC, Basch CE
Wait Times for Scheduling Appointments for Prevention of Macrovascular and Microvascular Complications of Diabetes: Cross-Sectional Descriptive Study
J Med Internet Res 2024;26:e55351
URL: https://www.jmir.org/2024/1/e55351
doi:10.2196/55351
PMID:38530352

©Corey H Basch, Grace C Hillyer, Charles E Basch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of Telehealth Services that are Clinically Appropriate for Reimbursement in the US Medicaid Population: Mixed Methods Study

Sanjeev Saravanakumar[1], BA; Andrey Ostrovsky[2], MD

[1]GW School of Medicine & Health Sciences, Washington, DC, United States
[2]Social Innovation Ventures, Washington, DC, United States

Corresponding Author:
Sanjeev Saravanakumar, BA
GW School of Medicine & Health Sciences
2300 I St NW
Washington, DC, 20052
United States
Phone: 1 6307303002
Email: sanjeev@gwu.edu

Abstract

Background: When the US Department of Health and Human Services instituted a State of Public Health Emergency (PHE) during the COVID-19 pandemic, many telehealth flexibilities were fast-tracked to allow state Medicaid agencies to reimburse new specialty services, sites of care, and mediums such as FaceTime to communicate with patients.. This resulted in expanded access to care for financially vulnerable Medicaid patients, as evidenced by an uptick in telehealth use. Research has mostly focused on telehealth reimbursement for limited use cases such as rural primary care, without broader consideration for how telehealth can be appropriately mainstreamed and maintained.

Objective: This study sought to (1) evaluate the continuation of flexible telehealth reimbursement broadly, beyond the COVID-19 pandemic; (2) analyze the clinical effectiveness of the new telehealth services; and (3) offer code-by-code reimbursement guidance to state Medicaid leaders.

Methods: We surveyed 10 state Medicaid medical directors (MMDs) who are responsible for the scientific and clinical appropriateness of Medicaid policies in their respective states. Participants were asked to complete an internet-based survey with a list of medical billing codes, grouped by service type, and asked if they believed they should be reimbursed by Medicaid on a permanent basis. Additional questions covered more detailed recommendations, such as reimbursing video with audio versus audio-only, guardrails for certain specialty services, and motivations behind responses.

Results: The MMDs felt that the majority of services should be reimbursed via some modality of telehealth after the PHE, with the most support for video combined with audio compared to audio-only. There were exceptions on both ends of the spectrum, where services such as pulmonary diagnostics were not recommended to be reimbursed in any form and services such as psychotherapy for mental health had the most support for audio-only. The vast majority of MMDs were supportive of reimbursement for remote monitoring services, but some preferred to have some reimbursement guardrails. We found that 90% (n=9) of MMDs were supportive of reimbursement for telehealth interprofessional services, while half (n=5) of the respondents felt that there should be continued guardrails for reimbursement. Motivations for continuing reimbursement flexibility were largely attributed to improving access to care, improving outcomes, and improving equity among the Medicaid patient population.

Conclusions: There is a strong clinical endorsement to continue the telehealth flexibility enabled by the PHE, primarily for video combined with audio telehealth, with caution against audio-only telehealth in situations where hands-on intervention is necessary for diagnosis or treatment. There is also support for reimbursing remote monitoring services and telehealth interprofessional services, albeit with guardrails. These results are primarily from a perspective of improving access, outcomes, and equity; other state-specific factors such as fiscal impact and technical implementation may need to be taken into account when considering reimbursement decisions on telehealth.

(J Med Internet Res 2024;26:e46412) doi:10.2196/46412

XSL•FO
RenderX

KEYWORDS

mobile phone; telehealth; Medicaid reimbursement; health equity; Center for Medicare & Medicaid Services; telemedicine; reimbursement; digital health; Medicaid; Public Health Emergency; access; equity; health insurance; coverage; reimburse; equitable; health policy; telehealth expansion

Introduction

The COVID-19 pandemic created unprecedented disruption to health care systems, particularly to systems serving the most vulnerable populations, such as beneficiaries of the US Medicaid services. Stay-at-home and social distancing orders forced rapid uptake of new technologies for remote patient care, most notably telehealth. Under the state of public health emergency (PHE), state Medicaid programs allowed for additional flexibility for providers to receive reimbursement for telehealth, including allowing the home as an originating site for patients to use telehealth and expanding allowable technological devices to include mainstream smartphone apps such as FaceTime, WhatsApp, and Skype [1,2]. Many of these technological advances were developed before the pandemic but saw a surge in popularity as the public needed creative solutions to work around the pandemic [1]. We now see that telemedicine adoption and use has expanded, which provides new evidence on the impact of telemedicine on cost, quality, and access for those unsure about implementing telehealth in their own practices.

The US Medicaid program was a major beneficiary of the surge in telehealth popularity, being one of the largest insurers with over 85 million people covered in the United States. Medicaid beneficiaries primarily include people whose income is below the federal poverty level, children, pregnant women, older individuals, and individuals with disabilities [3]. While Medicaid is an entitlement program by the federal government, it is largely state-run, with the Social Security Act (SSA) and associated waivers (section 1115, section 1915b, and section 1915c) providing states the flexibility on Medicaid eligibility, covered benefits, and provider payment rates [3]. The flexibility of Medicaid administration has therefore created state-by-state differences in Medicaid's day-to-day operations, including the implementation and oversight of telemedicine services [4].

The National Medicaid Medical Directors Network (MMDN) is a nonpartisan, nonprofit collective that helps convene Medicaid leaders across all 50 states to discuss differences in state implementation, share best practices, and solve pressing needs among the states [5]. State Medicaid medical directors (MMDs) are typically responsible for ensuring the scientific and clinical appropriateness of state Medicaid policies. When the PHE provided flexibility for telehealth reimbursement during the COVID-19 pandemic, each state was guided by its respective medical director to responsibly ramp up its telehealth capabilities. Historically, these benefits were limited to rural areas or special circumstances for those with intellectual and developmental disabilities [6]. During the COVID-19 pandemic, the expansion of these benefits to all Medicaid beneficiaries marked a period of unprecedented telehealth access.

Before the pandemic, telehealth use represented a negligible amount (<1%) of total outpatient visits, exploded to 13% during the March 2020 to August 2020 period, and has since stabilized to a new level of 8% from March 2021 to August 2021 [1]. There was high variability in telemedicine uptake based on specialty ranging from endocrinologists (67.2%), gastroenterologists (57%), neurologists (56.3%), pain management physicians (50.6%), psychiatrists (50.2%), orthopedic surgeons (20.7%), and ophthalmologists (9.3%) using any telemedicine during the pandemic [7]. While telehealth offers numerous benefits to patients and the Medicaid program, including improved access to care, the impacts on clinical quality and financial maintainability remain unclear [2]. State Medicaid programs are facing budget constraints, forcing conversations about the cost-effectiveness of telehealth services going forward [5]. While numerous questions remain, the growing potential of telehealth has led to several legislative efforts to consider extending telehealth flexibility beyond the PHE and making it permanent [8].

Given the implications of these changes, state Medicaid leaders have expressed needing code-by-code guidance on the specific services that are appropriate to reimburse when delivered via telehealth. Because of the expansion flexibility paired with state-to-state differences in implementation, an existing universal precedent has yet to be established. The 2021 MMDN report and Government Accountability Office (GAO) report on considerations for telehealth reimbursement continuation delve into the topic but do not provide strong prescriptive recommendations.

This paper aims to aggregate the expertise of leadership who oversaw state telehealth expansion during the COVID-19 pandemic, identify code-by-code reimbursement guidance, and provide recommendations at-large for state Medicaid directors. Stakeholders such as providers, payers, and patients are important in shaping this conversation and should be considered with the perspective of the MMDs provided in this study. It is also historically evident that there are challenges for telehealth expansion, which range across cost, technology, and organizational domains, making these data all the more important for the exchange of lessons learned from state to state and to provide guidance for leaders both in the United States and abroad who are uncertain about risk and financial reimbursement policies [9].

We hypothesized that MMDs would advocate to continue reimbursing video and audio for most specialty service codes but would be wary of audio-only telehealth for certain conditions. We summarize and synthesize the findings from a survey of the MMDN which offers evidence-backed input through a lens of clinical appropriateness and quality.

Methods

Overview

Study participants were selected from the MMDN, a collaborative platform for senior clinical leaders within state

XSL•FO
RenderX

Medicaid agencies, which meets periodically to exchange best practices aimed at enhancing the health outcomes of Medicaid beneficiaries [5]. Considering the study's objective to discern the clinical justifiability for telehealth reimbursement, engaging MMDs as respondents for this survey was ideal.

The survey was designed to capture MMD's viewpoints on Medicaid's reimbursement and benefit coverage policies for telehealth services extending beyond the PHE phase (Multimedia Appendix 1). The formulation of survey questions was informed by an examination of 10 state Medicaid program bulletins—namely, California, Massachusetts, New York, Florida, Pennsylvania, Michigan, North Carolina, Georgia, Ohio, and Illinois—concerning coding and coverage specifications for telehealth services amid the PHE toward the end of 2020 (Multimedia Appendix 2). Prior to distribution, the survey underwent a pilot test conducted by a medical director specializing in virtual care at a major academic medical center, leading to revisions that enhanced the readability and precision of question content. The survey, disseminated via email, was self-administered and required approximately 10-12 minutes for completion. Participants were presented with a compilation of Current Procedural Terminology (CPT) codes and Healthcare Common Procedure Coding System (HCPCS) codes, organized according to service type, and were instructed to indicate which services they deemed warranting permanent Medicaid reimbursement after the PHE. Further inquiries were made regarding the preferred modes of service delivery, encompassing both video and audio, video with audio only, audio-only, or an option for respondents unfamiliar with the code. Regarding a set of services traditionally offered over the telephone, respondents were queried on the appropriateness of categorizing and reimbursing these as telehealth services by Medicaid beyond the PHE, with the home identified as the originating site, and whether such coverage should be extended with or without specific guardrails. The concluding segment of the survey sought to understand the rationale underpinning the respondents' selections in the preceding sections. For a full examination of the survey, refer to Multimedia Appendix 1.

In collaboration with Academy Health, we distributed the survey to all MMDs (N=50) between January 5, 2021, and February 5, 2021. Qualtrics was the designated survey software and AcademyHealth directly sent the survey via email to all the MMDs. To increase the response rate, a single follow-up email was sent a week later. There was no compensation for completing the survey.

To understand the implications of the survey responses, we used a mixed methods analytical approach. Quantitative data from the survey were analyzed using descriptive statistics to identify trends and patterns in the responses. This included the percentage of respondents who supported the continuation of telehealth reimbursement for specific services, as well as the preferred modality of telehealth (video with audio vs audio-only). We also examined the distribution of responses across different service types and specialties. Qualitative data from the open-ended questions were analyzed thematically to gain insights into the motivations behind the responses and to identify potential guardrails for certain specialty services. This analytical approach allowed us to appreciate the complexities of telehealth reimbursement to provide nuanced recommendations for state Medicaid leaders.

The analysis in our study was primarily descriptive, aimed at discerning the perspectives of MMDs regarding the continued reimbursement for services rendered via telehealth post the PHE.

Given the constraint of small sample sizes, we refrained from performing formal statistical tests that might not yield robust conclusions under such circumstances. To maintain the integrity and relevance of the data, we filtered out responses where less than 50% of the survey questions were completed, ensuring that the analysis was grounded on substantial and meaningful respondent input.

Ethical Considerations

We did not pursue an institutional review board (IRB) review because our study survey was administered to a nonvulnerable population, entailing no greater than usual risk, with a stringent anonymization process applied to all data collected. This study aligns with the exemption criteria delineated by institutional and regional guidelines, such as those specified by the US Department of Health & Human Services (45 CFR 46.101(b)). Particularly, our methodology resonates with scenarios for which IRB approval may be exempt, including research involving educational tests, surveys, interviews, or observation of public behavior, provided the information obtained is recorded in such a manner that subjects cannot be identified, directly or through identifiers linked to them. This alignment with established guidelines underscores the justification for forgoing IRB approval in our study, thereby adhering to the ethical framework governing human subject research while facilitating timely and resource-efficient data collection. Using diligent data management and confidentiality measures was paramount in our study, particularly as an IRB assessment was not obtained for the survey administered. All survey responses were anonymized at the point of collection, with no personally identifiable information recorded, adhering to the principles of confidentiality and privacy. A robust data management system was used to securely store, process, and analyze the collected data, ensuring restricted access only to authorized members of the research team.

Results

A total of 12 (24%) MMDs completed the survey, and we included 10 respondents in the analysis (20% net response rate), after removing 2 respondents with incomplete surveys. Figure 1 displays the respondents' preferences for reimbursement for groups of codes that can be received via telehealth or telemedicine. See Multimedia Appendix 2 for specific CPT codes. Respondents felt that the majority of services should be reimbursed via some modality of telehealth after the PHE, though respondents were generally more comfortable with video combined with audio compared to audio-only. The main exception was psychotherapy for mental health or substance use disorder treatment, which had the most support of any services for audio-only with 60% (n=6) of respondents supporting it. There was only one category of codes, pulmonary

XSL•FO
RenderX

diagnostics, that a majority of respondents felt should not be reimbursed in any telehealth form; other categories that did not have favorable reimbursement support with 40% (n=4) of respondents including personal care services, adult day services, and physical performance tests.

Figure 1. Reimbursement of services that Medicaid beneficiaries can receive via telehealth or telemedicine. CHC: community health center; E&M: evaluation and management; ESRD: end-stage renal disease; FHQC: federally qualified health center; HRA: health risk assessment.

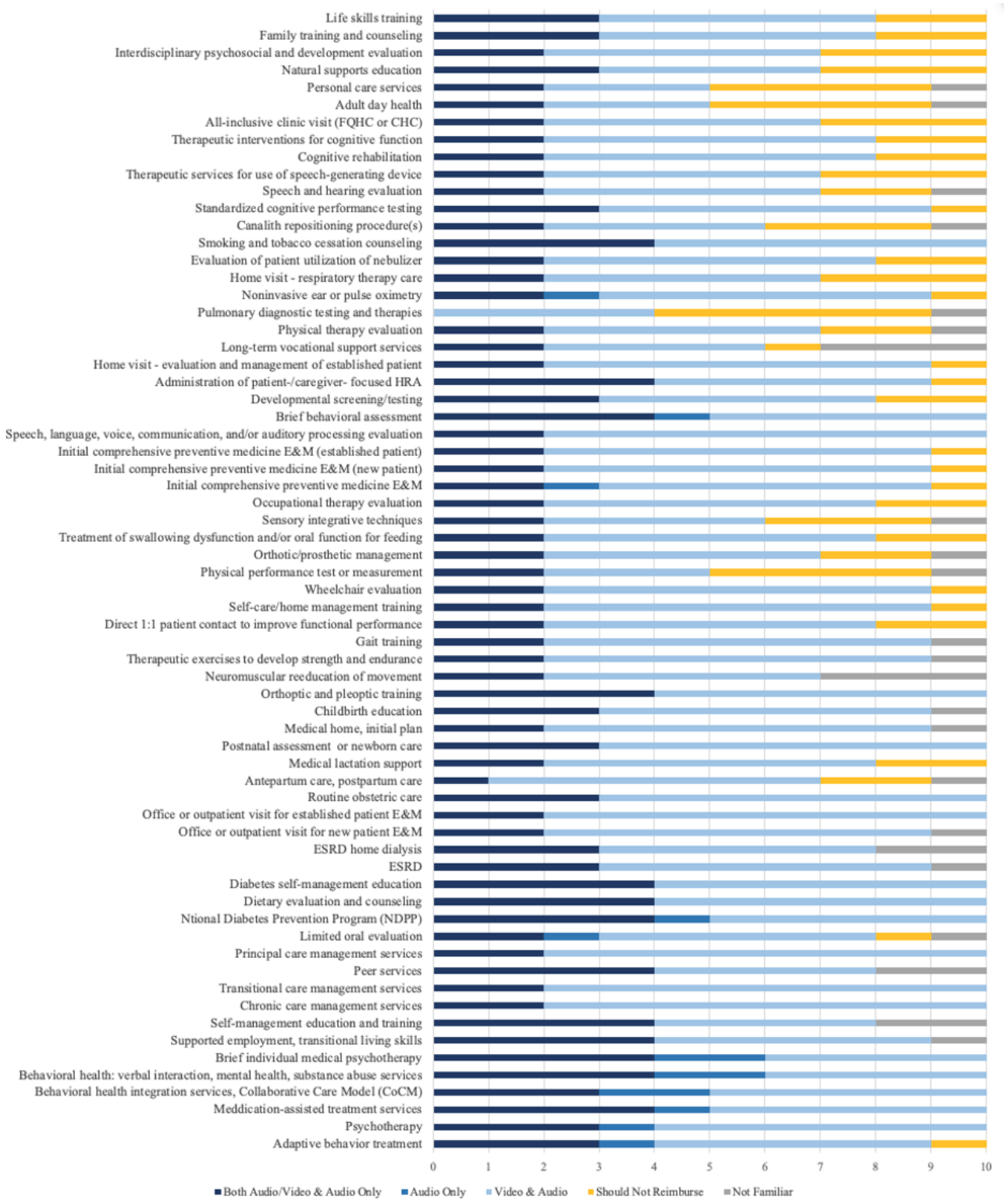

Figure 2 displays responses about continued reimbursement after the PHE for remote monitoring services and services generally delivered via telephone. Respondents supported continued reimbursement with 70% (n=7) to 90% (n=9) in favor of reimbursing remote monitoring services and 50% (n=5) to 70% (n=7) in favor of reimbursing services that are generally delivered by telephone. However, more than half of supportive respondents felt there should be continued guardrails for reimbursement. The proposed guardrails were inconsistent or not specified across these responses. Responses showed a similar

XSL•FO
RenderX

pattern for reimbursement of interprofessional services in Table 1.

When asked about motivations for reimbursement decisions in the previous questions, respondents reported a variety of motivations in Table 2. The most common responses were the service's likelihood to improve access to care (n=10, 100%), the service's likelihood to improve outcomes (n=10, 100%), and the service's likelihood to improve equity (n=8, 80%), while the least common response was the level of evidence for the service and the service's contribution to the financial sustainability of the Medicaid program (n=3, 30%).

Figure 2. Reimbursement for remote monitoring services and services that Medicaid beneficiaries generally receive by telephone. E&M: evaluation and management.

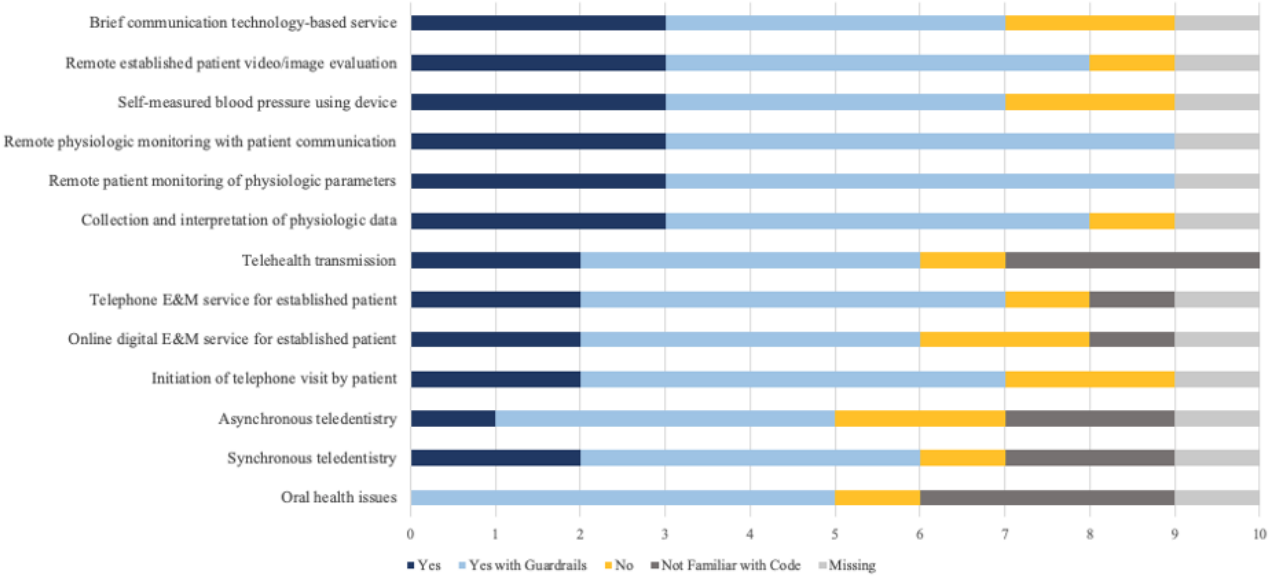

Table 1. Telehealth reimbursement of interprofessional services (n=10)[a].

Responses	Number of MMDs[b], n
Yes	4
Yes, with guard rails	5
No	1
Not familiar with code	0
Missing response	0

[a]Services include interprofessional telephone or internet or electronic health records assessment and management services provided by a consultative physician, including a verbal and written report to the patient's treating or requesting physician or other qualified health care professional.

[b]MMD: Medicaid medical director.

Table 2. Motivation for telehealth reimbursement recommendation (n=10).

Motivation characteristic	Number of MMDs[a], n (%)
Service's likelihood to improve access to care	10 (100)
Service's likelihood to improve outcomes	10 (100)
Service's likelihood to improve equity	8 (80)
Service's likelihood to improve beneficiary experience	7 (70)
Service's likelihood to reduce provider burden	7 (70)
Service's cost-effectiveness	7 (70)
Service's contribution to the financial sustainability of Medicaid program	3 (30)
Level of evidence for the service	3 (30)

[a]MMD: Medicaid medical director.

XSL•FO
RenderX

Discussion

Principal Findings

The expansion of telehealth capabilities during the COVID-19 pandemic offered a unique opportunity for clinical leaders to evaluate the strengths and weaknesses of virtually delivered care. The survey administered to MMDs across the United States provides us with insight into the specialty billing codes recommended for reimbursement specifically via telehealth. Several states across the United States were involved in the data aggregation process and the results validated our hypothesis that there would be support for permanent reimbursement for video with audio telehealth after the pandemic for a majority of services with special caution for audio-only telehealth.

Support for continued audio or video telehealth use was prominent; however, there were mixed opinions on continued support for audio-only telehealth from the MMDs. From a clinical perspective, MMDs were hesitant to support audio-only telehealth services where hands-on evaluation or intervention was traditionally necessary. MMDs were concerned about the inability to identify nonverbal cues such as signs of oversedation, anxious fidgeting, or environmental risks such as unsafe housing conditions, which could pose unidentified risks from using telehealth. However, MMDs mostly agree on the usefulness of audio-only telehealth for behavioral health services, aligning with the Centers for Medicare & Medicaid Services (CMS) Physician Fee Schedule 2022 policy approving reimbursement of audio-only telebehavioral health under certain conditions [10].

The benefits of audio-only telehealth are notable, especially in improving access for individuals lacking reliable internet, addressing health equity concerns. Most surveyed MMDs preferred having some regulations if such services are to be reimbursed, to avoid impacting disadvantaged populations adversely. Audio-only services have notably reduced missed appointments and enhanced patient engagement, without compromising care quality in areas like maternal care [11,12].

Furthermore, the results showed support for reimbursing remote monitoring services and telephone-delivered services, though coding and reimbursement challenges have impeded access for Medicaid and Medicare beneficiaries [13]. Ambiguity in coding has historically hindered adoption and patient access to remote physiologic monitoring codes but the creation of new coding options for remote therapeutic monitoring aims to mitigate this issue.

Finally, to address the longstanding issue of specialist availability, MMDs support reimbursing virtual interprofessional services, aiming to attract more specialists to treat Medicaid beneficiaries. In alignment with these findings, the Center for Medicaid and the Children's Health Insurance Program (CHIP) Services (CMCS) recently issued subregulatory guidance encouraging states to cover and reimburse virtual interprofessional services [14]. This is a critical and necessary step to ensure specialist access for patients from all geographic and socioeconomic backgrounds. The endorsement for virtual interprofessional services reimbursement by this study and by CMCS, combined with the dearth of e-consultation platforms, presents a potential market opportunity for providers targeting Medicaid beneficiaries.

Comparison to Prior Work

Recent reports by the GAO on Medicare and Medicaid detailed that the telehealth flexibility during the PHE period was enabled with the goal of lessening barriers affecting beneficiary care, augmenting provider availability, and ensuring adequate reimbursement to cover the additional costs of virtual care. Prior to the PHE, telehealth use had a strictly defined scope for Medicare and Medicaid reimbursement with gradual movement toward additional use based on emergent use cases.

The precedent for telehealth reimbursement is now being revisited due to the growing number of use cases and its potential to transform health care access, given its vital role during the COVID-19 pandemic. However, the barriers to universal adoption involve a perceived lack of research proving that telehealth is safe and effective on a broader scale.

Further reports by the MMDN have emphasized the importance of evidence as a key criterion for deciding reimbursement. They also noted that the impact of COVID-19 pandemic-related telehealth waivers on the quality of care remains unknown, with current broad evidence being inconclusive.

In the setting of these reports and prior precedent, our survey for MMDs explored clinical factors for continuing or discontinuing telehealth reimbursement which provides additional evidence to the discussion. The results showed that the primary factors influencing the responses were whether the service code could enhance access, outcomes, and equity of care. Conversely, the least common considerations were the level of evidence and financial maintainability of the service code. Our data highlight a discrepancy between prioritizing services that enhance access and outcomes, and those backed by a solid level of evidence.

Continuing the existing telehealth flexibility offers an opportunity to maintain equitable access and simultaneously expand the evidence base through additional measurement and analyses. With this study, we have identified, on a code-by-code basis, the services that MMDs find clinically appropriate for reimbursement while further evidence is gathered and evaluated.

Cost and Technology Implementation

While some MMDs have expressed concerns about cost and implementation challenges associated with telehealth, safety-net providers are generally accepting of permanent telehealth adoption as long as there is sufficient reimbursement [9]. The evidence that Medicaid beneficiaries are eagerly willing to use telehealth services paired with appropriate Medicaid reimbursement can make the venture cost-effective for health centers in the long run. Studies evaluating telehealth cost-effectiveness have already proven to be efficient in certain areas, such as the management of type 2 diabetes and chronic heart failure when compared to usual methods of in-person care [15].

One approach to ease the implementation and cost burden for administrators and health centers is to default to full

XSL•FO
RenderX

reimbursement for services in Figures 1 and 2 that have strong clinical support from MMDs. For services that did not have MMD support for reimbursement, state Medicaid programs can develop low-burden use management approaches with rapid cycle health services research to balance maintaining provisional and targeted coverage while being good stewards of state budgets.

Policy Recommendations

A key workflow barrier that the COVID-19 PHE flexibility resolved was the originating site, or the physical location of the patient when using telehealth devices. Medicare and many Medicaid programs expanded the types of originating sites a patient could be at to include the home and other locations [9]. Waiver of the requirement for the home to be the originating site per section 1834(m) of the SSA is a requisite and foundational policy to enable the benefits of telehealth to be realized from an access, outcomes, and equity perspective. This PHE flexibility is due to expire in December 2024, which creates a legislative cliff that will imperil the most vulnerable Americans and wipe out the access gains achieved during the pandemic. To maintain the equitable access and outcomes gained during the pandemic the US Congress must create a legislative fix to permanently strike the SSA section 1834(m) restriction on the home as the originating site. If Congress cannot create a legislative fix before the end of 2024, then CMS needs to extend this flexibility to at least the end of 2025 to allow Congress to address the telehealth cliff. Additionally, CMCS could offer states waivers to the 1834(m) restriction on the home as the originating site through 1115 demonstrations or 1332 waivers.

In addition to addressing the 1834(m) restriction, CMCS should actualize the learnings from this study by issuing subregulatory guidance through a CMCS informational bulletin or state health official letter advising State Medicaid Directors to undertake the following steps: (1) fully reimburse telehealth services that are video and audio; (2) fully reimburse audio-only services that do not entirely require hands-on evaluation or intervention with the patient, such as telebehavioral health; (3) develop use management processes with a low-provider burden to consider full reimbursement of audio-only services that predominantly require hands-on evaluation or intervention with the patient while balancing access with safety and cost-effectiveness; and (4) promote the adoption of the newly created CPT 93 modifier for audio-only telehealth to ensure more research can be done on audio-only telehealth [16].

CMCS Process Recommendations

To support the successful implementation of these policies, the CMCS approval processes for state plan amendments, waivers, and demonstrations should err on the side of enabling provider discretion and should be viewed through the lens of access, outcomes, and equity. The approval processes should also deeply involve the CMCS chief medical officer as well as the state-specific MMDs. Additionally, there should be standard quality measures that can be used across virtual and nonvirtual interventions.

Limitations

First, our study had a primary limitation related to the size of our participant sample. Although our surveys were sent to all MMDs in the United States and our response rate was a reasonable 20% (n=10), the total sampling frame was small to begin with. We appreciate the support from AcademyHealth for distributing surveys and attempting follow-up with all the MMDs to optimize the response rate.

Second, within the survey itself, we acknowledge room for improvement. Specifically, question 2 could have been more explicit, addressing remote patient monitoring more directly. Additionally, we did not explicitly inquire about the preferred reimbursement levels for each service, so we can only draw limited conclusions about the precision of reimbursement.

Third, we encountered a technical issue during the initial survey distribution that hindered 2 respondents from selecting multiple options. Fortunately, we rectified this issue promptly but this technical issue may have led to incomplete responses by 2 respondents leading to their responses not being included in the analysis.

Finally, in terms of illness considerations within the survey, we intentionally excluded the assessment of illness severity as it was deemed beyond the scope of our study with respect to telehealth management.

Conclusions

While telehealth existed before the COVID-19 outbreak, flexibility in telehealth policy has allowed for a more complete array of capabilities to be used across the United States. When considering the continuation of these policies, we validated our hypothesis that MMDs would be supportive of continued reimbursement of audio or video telehealth but would be wary of audio-only telehealth. The data show that hesitation for supporting audio-only telehealth was driven by clinical situations where visualization or hands-on intervention with the patient was necessary for diagnosis or treatment. There is also support for reimbursing remote monitoring services and telehealth interprofessional services, albeit with guardrails. These results are primarily from a perspective of improving access, outcomes, and equity; other state-specific factors such as fiscal impact and technical implementation may need to be taken into account when considering reimbursement decisions on telehealth.

These results on the clinical appropriateness of maintaining telehealth coverage should be taken in the context of broader research from the National Association of Medicaid Directors (NAMD) as well as the MMDN. The NAMD is conducting a parallel exercise with state Medicaid directors which offers a complementary perspective on the fiscal appropriateness and policy alignment of telehealth reimbursement. One way to strike a balance between equitably improving access and outcomes with fiscal stewardship is through accelerating value-based payment models in Medicaid. While state-specific decisions are being made on telehealth reimbursement, Congress must act to stave off the telehealth cliff.

XSL•FO
RenderX

Acknowledgments

We are grateful to the Medicaid medical directors that responded to the survey. We thank Sandy Varatharajah, Alina Schnake-Mahl, and Oren Gordon for their help with the study design and analysis. We also thank Anna Kalkanis-Ellis for her assistance with creating custom coding for the survey. Finally, we appreciate the input on survey methodology and results interpretation from Helen Hughes, Brian Hasselfeld, the Medicaid Medical Directors Network, Academy Health, and the National Association of Medicaid Directors. No external sources of funding were collected for this study.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AO contributed to the conceptualization, data curation, methodology, and project administration. SS shared responsibility in conceptualization and project administration with AO, and additionally took charge of formal analysis, validation, and writing. Both authors reviewed and approved the final version of the study.

Conflicts of Interest

AO is an investor in Social Innovation Ventures. Their role did not interfere with this research and its findings. SS declares no conflicts of interest.

Multimedia Appendix 1
Survey.
[DOCX File , 6771 KB - jmir_v26i1e46412_app1.docx]

Multimedia Appendix 2
Current Procedural Terminology coding overview.
[DOCX File , 31 KB - jmir_v26i1e46412_app2.docx]

References

1. Lo J, Rae M, Amin K, Cox C. Outpatient telehealth use soared early in the COVID-19 pandemic but has since receded. Peterson-KFF Health System Tracker. 2022. URL: https://tinyurl.com/tebh6p7f [accessed 2022-10-06]
2. Farb J, Yocom C. GAO-21-575T, medicare and medicaid: COVID-19 program flexibilities and considerations for their continuation. United States Government Accountability Office. 2021. URL: https://www.gao.gov/assets/gao-21-575t.pdf [accessed 2022-08-16]
3. Mitchell A, Baumrucker EP, Colello KJ, Napili A, Binder C, Braun SK. Medicaid: an overview. Congressional Research Service. 2023. URL: https://crsreports.congress.gov/product/pdf/R/R43357 [accessed 2022-08-16]
4. Levinson DR. Access to care: provider availability in medicaid managed care: report (OEI-02-13-00670) 11-08-2014. U.S. Department of Health and Human Services, Office of Inspector General. 2014. URL: https://www.oig.hhs.gov/oei/reports/oei-02-13-00670.asp [accessed 2022-08-16]
5. Medicaid medical directors offer key perspectives on telehealth. AcademyHealth. 2021. URL: https://academyhealth.org/publications/2021-03/medicaid-medical-directors-offer-key-perspectives-telehealth [accessed 2022-08-16]
6. Friedman C. Telehealth service delivery in medicaid home- and community-based services for people with intellectual and developmental disabilities. Int J Telerehabil 2022;14(1):e6478 [FREE Full text] [doi: 10.5195/ijt.2022.6478] [Medline: 35845163]
7. Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Variation in telemedicine use and outpatient care during the COVID-19 pandemic in the United States. Health Aff (Millwood) 2021;40(2):349-358 [FREE Full text] [doi: 10.1377/hlthaff.2020.01786] [Medline: 33523745]
8. HHS fact sheet: telehealth flexibilities and resources and the COVID-19 public health emergency. US Department of Health and Human Services. 2023. URL: https://tinyurl.com/yc5j2f6u [accessed 2024-02-23]
9. Lori UP, Bouskill KE, Jessica S, Mimi S, Shira HF. Experiences of medicaid programs and health centers in implementing telehealth. Rand Health Q 2020;8(4):RR-2564-ASPE [FREE Full text] [Medline: 32582468]
10. Fact sheet: calendar year (CY) 2022 medicare physician fee schedule final rule. Centers for Medicare and Medicaid Services. 2021. URL: https://www.cms.gov/newsroom/fact-sheets/calendar-year-cy-2022-medicare-physician-fee-schedule-final-rule [accessed 2022-08-16]
11. Muppavarapu K, Saeed SA, Jones K, Hurd O, Haley V. Study of impact of telehealth use on clinic "no show" rates at an academic practice. Psychiatr Q 2022;93(2):689-699 [FREE Full text] [doi: 10.1007/s11126-022-09983-6] [Medline: 35412100]

XSL•FO
RenderX

12. Duryea EL, Adhikari EH, Ambia A, Spong C, McIntire D, Nelson DB. Comparison between in-person and audio-only virtual prenatal visits and perinatal outcomes. JAMA Netw Open 2021;4(4):e215854 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.5854] [Medline: 33852002]
13. Jarrin R, Barrett MA, Kaye L, Sayiner S, von Leer A, Johns J, et al. Need for clarifying remote physiologic monitoring reimbursement during the COVID-19 pandemic: a respiratory disease case study. NPJ Digit Med 2021;4(1):50 [FREE Full text] [doi: 10.1038/s41746-021-00421-8] [Medline: 33712676]
14. Coverage and payment of interprofessional consultation in Medicaid and the Children's Health Insurance Program (CHIP). Pennsylvania Office of Rural Health. 2023. URL: https://tinyurl.com/5j428ht3 [accessed 2024-02-23]
15. Grustam AS, Severens JL, De Massari D, Buyukkaramikli N, Koymans R, Vrijhoef HJM. Cost-effectiveness analysis in telehealth: a comparison between home telemonitoring, nurse telephone support, and usual care in chronic heart failure management. Value Health 2018;21(7):772-782 [FREE Full text] [doi: 10.1016/j.jval.2017.11.011] [Medline: 30005749]
16. Beschloss A, Van Ramshorst R, Bachireddy C, Chen CP, Ostrovsky A. New coding modifier offers opportunity to investigate audio-only telehealth. Health Affairs Forefront. 2022. URL: https://tinyurl.com/43kcf6dw [accessed 2024-02-23]

Abbreviations

CHIP: Children's Health Insurance Program
CMCS: Center for Medicaid and CHIP Services
CMS: Centers for Medicare & Medicaid Services
CPT: Current Procedural Terminology
GAO: Government Accountability Office
HCPCS: Healthcare Common Procedure Coding System
IRB: institutional review board
MMD: Medicaid medical director
MMDN: Medical Directors Network
NAMD: National Association of Medicaid Directors
PHE: public health emergency
SSA: Social Security Act

Edited by G Eysenbach, T Leung; submitted 10.02.23; peer-reviewed by S Gordon, D Kohen, J Zhu; comments to author 06.04.23; revised version received 31.08.23; accepted 17.02.24; published 28.03.24.

Please cite as:
Saravanakumar S, Ostrovsky A
Evaluation of Telehealth Services that are Clinically Appropriate for Reimbursement in the US Medicaid Population: Mixed Methods Study
J Med Internet Res 2024;26:e46412
URL: https://www.jmir.org/2024/1/e46412
doi:10.2196/46412
PMID:38546706

©Sanjeev Saravanakumar, Andrey Ostrovsky. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telehealth Care Through Internet Hospitals in China: Qualitative Interview Study of Physicians' Views on Access, Expectations, and Communication

Yuqiong Zhong[1,2], Mphil; Jessica Hahne[3], MA, MPH; Xiaomin Wang[4,5], PhD; Xuxi Wang[1], Bphil; Ying Wu[1], MPhil; Xin Zhang[2,6*], MD, PhD; Xing Liu[6,7*], PhD

[1]School of Humanities, Central South University, Changsha, China
[2]Xiangya Hospital, Central South University, Changsha, China
[3]Department of Psychological & Brain Sciences, Washington University in St Louis, St Louis, MO, United States
[4]Center for Clinical Pharmacology, The Third Xiangya Hospital of Central South University, Changsha, China
[5]Center for Medical Ethics, Central South University, Changsha, China
[6]Medical Humanities Research Center, Central South University, Changsha, China
[7]Office of International Cooperation and Exchanges, Xiangya Hospital, Central South University, Changsha, China
[*]these authors contributed equally

Corresponding Author:
Xing Liu, PhD
Office of International Cooperation and Exchanges
Xiangya Hospital, Central South University
No 87 Xiangya Road, Kaifu District
Changsha, 410008
China
Phone: 86 18229765509
Email: xingliu0706@csu.edu.cn

Abstract

Background: Internet hospitals in China are an emerging medical service model similar to other telehealth models used worldwide. Internet hospitals are currently in a stage of rapid development, giving rise to a series of new opportunities and challenges for patient care. Little research has examined the views of chronic disease physicians regarding internet hospitals in China.

Objective: We aimed to explore the experience and views of chronic disease physicians at 3 tertiary hospitals in Changsha, China, regarding opportunities and challenges in internet hospital care.

Methods: We conducted semistructured qualitative interviews with physicians (n=26) who had experience working in internet hospitals affiliated with chronic disease departments in 3 tertiary hospitals in Changsha, Hunan province, south central China. Interviews were transcribed verbatim and analyzed by content analysis using NVivo software (version 11; Lumivero).

Results: Physicians emphasized that internet hospitals expand opportunities to conduct follow-up care and health education for patients with chronic illnesses. However, physicians described disparities in access for particular groups of patients, such as patients who are older, patients with lower education levels, patients with limited internet or technology access, and rural patients. Physicians also perceived a gap between patients' expectations and the reality of limitations regarding both physicians' availability and the scope of services offered by internet hospitals, which raised challenges for doctor-patient boundaries and trust. Physicians noted challenges in doctor-patient communication related to comprehension and informed consent in internet hospital care.

Conclusions: This study explored the experience and views of physicians in 3 tertiary hospitals in Changsha, China, regarding access to care, patients' expectations versus the reality of services, and doctor-patient communication in internet hospital care. Findings from this study highlight the need for physician training in telehealth communication skills, legislation regulating informed consent in telehealth care, public education clarifying the scope of internet hospital services, and design of internet hospitals that is informed by the needs of patient groups with barriers to access, such as older adults.

(J Med Internet Res 2024;26:e47523) doi:10.2196/47523

XSL•FO
RenderX

KEYWORDS

China; internet hospital; health care access; telehealth; doctor-patient relationship; mobile phone

Introduction

As information technology develops rapidly in the current era, telehealth is growing exponentially in use [1-3]. Particularly in the wake of the COVID-19 pandemic, the use of telehealth for both primary and specialist care has accelerated around the globe [4]. In particular, telehealth is being implemented at an increasing scale in various countries with aging populations to improve health care access and quality for growing numbers of patients with chronic diseases [5-9].

The internet hospital is 1 major emerging telehealth model that is distinct to China, a country with a particularly large aging population and a high chronic disease burden [10-12]. Designed to make health care services more available, convenient, affordable, and efficient, internet hospitals are a type of online platform through which certain health care services can be conducted remotely. There are 3 main types of internet hospitals—those initiated by physical hospitals, those jointly established by physical hospitals and business enterprises, and those initiated by business enterprises relying on physical medical institutions. Research suggests that internet hospitals initiated solely by physical hospitals are the most widespread type [13]. In terms of the target patient population, internet hospitals primarily aim to facilitate services for patients with common illnesses requiring relatively simple treatment [14], patients with chronic diseases (diabetes, hypertension, and cancer) [15], and patients in remote and rural areas [16].

However, the scope of internet hospitals goes beyond telemedicine services for patients. Services provided by internet hospitals can be classified into three categories, (1) “core medical services,” which mainly include follow-up care for in-person medical services, telemedicine consultations, guidance on chronic disease management, and guidance on medication use; (2) “non-core medical services,” which mainly include medical consultations between health care providers and remote education for health care providers; and (3) “convenience services,” which mainly include health care appointment scheduling, mobile payment for health services, remote examination of medical test results, and dispensation and distribution of some medications [10,13,17,18]. Thus, the internet hospital model has the potential to increase access to health care for patients and training for providers, and to decrease costs across the health care system.

A number of recent policies by the Chinese national government have promoted rapid development and uptake of the internet hospital model. In 2015 [19] and 2018 [20], the State Council issued guidelines promoting “‘Internet +’ Healthcare,” which emphasized the development of internet hospitals as part of the “Health China” strategy for health care reform. Concurrently in 2018, the National Health Commission formulated specific regulations on internet hospital management, which officially authorized internet hospitals to facilitate a range of telehealth services and marked the start of their standardized development [21-23]. In 2020, the National Health Commission issued the “Notice on Strengthening Informatization to Support the Prevention and Control of the Novel Coronavirus Pneumonia Epidemic,” emphasizing the advantages of internet hospitals in controlling the spread of the COVID-19 pandemic [24]. In the wake of these policies, by June 2023, the number of internet hospitals had reached more than 3000, and 364 million of China’s 1.079 billion internet users were using online medical services [25,26]. However, research suggests that most internet hospitals are not yet fully developed or providing the full scope of services intended to achieve these goals [27-29].

At this early stage of the model’s development, little research to date has evaluated the views of Chinese medical professionals and patients regarding internet hospitals. However, research on telehealth in other countries reveals that telehealth raises many new concerns and challenges alongside the aforementioned opportunities [30-32]. One of the most common concerns raised by patients is the potential for misdiagnosis due to the inability to conduct physical examinations through telehealth [33,34]. Particular groups such as lower-income older adults also commonly report barriers to the use of telehealth such as lack of familiarity with technology or limited access to technological devices or internet connections [35]. Smartphone data or internet connection problems can also lower patient satisfaction and limit access among rural patient populations [36,37]. Various groups of patients also commonly report feeling concerned about patient privacy and the security and protection of medical data when using telehealth [38].

In order to guide the direction of internet hospital development in China, further research is needed to examine the emerging challenges and opportunities to patient care presented by this country-specific telehealth platform. The aim of this study was to explore the experience and views of chronic disease physicians at 3 tertiary hospitals in Changsha, China, regarding opportunities and challenges presented by internet hospital care.

Methods

Overview

The methodology whereby this study was designed and conducted is reported following the items in the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist [39]. See Multimedia Appendix 1 [39] for more information.

Setting, Participant Recruitment, and Eligibility Criteria

We conducted in-depth, semistructured interviews with physicians at 3 tertiary hospitals in Changsha, Hunan Province, south central China. Inclusion criteria for participants were 18 years of age or older, experience working in internet hospitals, and employment in a chronic disease department at one of the study hospitals. Our rationale for these inclusion criteria was to select doctors who had work experience relevant to the research questions. Enrollment occurred over a 2-month period from April to May 2022. Using a purposive sampling approach, we obtained a list of doctors who had previous experience

XSL•FO
RenderX

working in internet hospitals. We then messaged or called the doctors on the list to briefly explain the primary and secondary objectives of the study, invited them to share their perspectives related to the study, and asked them to be available for interviews. Out of the 28 doctors contacted, 26 agreed to participate, while 2 physicians declined due to a lack of time. Those who responded positively to the invitation were subsequently contacted by the author, YZ, either via WeChat (Tencent Holdings Limited) or telephone to schedule an interview.

We recruited participants until reaching data saturation, at which point no new information about the meaning of codes or themes and the relationship between them continued to appear [40].

In order to allow spontaneous answers and mitigate bias, participants were given minimum information in advance about specific interview topics.

Ethical Considerations

In April 2022, the research protocol was approved by the institutional review board of Xiangya Hospital, Central South University (#202204092). No prior relationships existed between study participants and members of the research team. Verbal informed consent was recorded via an audio recorder for each participant before participation. Participants were informed in advance that their interviews would be recorded, with the assurance that these recordings would be subject to encryption for security purposes, and they provided their verbal consent accordingly. All participants received a compensation of 200 RMB (1 CNY=US $0.15 on May 2022) for their participation, which was disbursed through a WeChat transfer. To protect the information of the interviewees, the interview data were deidentified in the process of transcription from audio recordings.

Data Collection

The interview guide was collaboratively developed and then subjected to pilot-testing by the research team. Throughout the concurrent phases of data collection and analysis, the interview guide underwent iterative refinement in response to emerging insights and participant responses. This adaptive approach is considered vital to the robustness of qualitative research [41]. Revisions were implemented subsequent to discussions involving YZ, JH, and Xiaomin Wang, aiming to clarify, define, and critically examine emerging content from interviews as relevant to the research questions. All questions from the finalized interview guide are listed in Textbox 1.

Textbox 1. Semistructured interview guide.

Q1: What are your views on internet hospitals?

Q2: Could you tell me about your experience working in internet hospitals?

Q3: What do you think are the biggest advantages of internet hospitals? Can you give some examples?

Q4: What do you think are the most troubling or difficult aspects of internet hospitals? Can you give some examples?

Q5: How do you inform the patients who come to the internet hospitals before treatment?

Q6: Do you think there are any differences between doctor-patient communication in internet hospitals and physical hospitals?

Q7: What impact do you think internet hospitals have on doctor-patient communication? Can you give some examples?

Q8: What training have you participated in regarding internet hospitals? What do you think of this training?

Q9: What do you think about the current status of the development of internet hospital laws and regulations in China? How could they be improved? What other aspects can promote the further development of internet hospitals?

Q10: Is there anything else you would like to add about doctor-patient communication in internet hospitals?

Data for this study were collected from April to May 2022. We carried out interviews through a combination of online and offline modalities, depending on each participant's preference and availability. Online interviews were conducted remotely by video call, via the mobile app WeChat. Offline interviews were conducted in private rooms at the study hospitals. The interviews were conducted in Mandarin Chinese by the authors, YZ (a postgraduate student) and Xiaomin Wang (an associate professor). Both interviewers have received professional training in qualitative interviewing and had extensive experience conducting qualitative research prior to this study.

The research team discussed possible probes and follow-up questions before beginning interviews, and interviewers used them when necessary to draw out more information relevant to the main research question. Concurrently, a second interviewer assumed the role of an observer to ensure the standardization of interview methods and to mitigate potential biases.

Interviews were audio recorded, transcribed, and uploaded into qualitative data management NVivo software (version 11; Lumivero) on password-protected computers to facilitate the analysis. Field notes made by interviewers during the interview process were also stored on password-protected computers, to be used for reference by the research team during analysis. Interviews ranged from 20 to 50 minutes long. Transcripts were sent to participants upon request, but no corrections, comments, or notes were made to transcripts.

Data Analysis

Analysis of the data was performed through conventional content analysis, using guidelines described by Hsieh and Shannon [42]. An advantage of conventional content analysis is that it avoids using preconceived categories, to generate codes inductively from the data. This modality is considered appropriate when current knowledge of the phenomenon being researched is limited [42].

XSL•FO
RenderX

Authors YZ and Xuxi Wang transcribed all interviews verbatim and reviewed transcripts several times to acquire a thorough understanding of the whole data set. They then read transcripts line-by-line and highlighted keywords and sentences from a set of initial transcripts, to generate primary codes that captured key concepts. Primary codes were repeatedly reviewed and revised through discussion among the authors and comparison across the transcripts. A finalized codebook including 17 codes and 61 subcodes was used to code all interviews, using NVivo software (version 11). Data saturation was reached after 26 interviews, once the research team determined by the consensus that we had interviewed a sufficiently varied sample of physicians from the 3 study hospitals, while also having obtained sufficiently content-rich data.

Following the coding of all transcripts, all coded segments of the interview data were translated into English by authors YZ and Xuxi Wang, native Mandarin speakers, and double-checked for accuracy by author JH, a native English speaker. Codes and subcodes were repeatedly reviewed and were grouped into clusters according to similarities and differences. Clusters of codes were then treated as subcategories and aggregated into the main categories that were representative of the key findings. These categories were repeatedly reviewed until fully developed, through a process of identifying and comparing exemplary excerpts for each code, category, and subcategory.

This process of analysis culminated in three main categories describing the experiences and views of chronic disease physicians regarding the opportunities and challenges presented by internet hospital care, (1) advancements and shortcomings in care access due to internet hospitals, (2) patients' expectations versus limitations on doctors' availability and the scope of services—implications for doctor-patient boundaries and trust, and (3) advantages and downsides of online communication for comprehension and informed consent. These main categories are shown in Textbox 2, alongside the subcategories from which they were aggregated, and further explained in the results narrative below.

Textbox 2. Main study findings.

Advancements and shortcomings in care access due to internet hospitals

- Advancements
  - Enhanced ability to conduct follow-up care for patients with chronic illness
  - More efficient channels for health education
- Shortcomings
  - Disparities in access (ie, for older adults, patients with lower education levels, patients with limited internet or technology access, rural patients)

Patients' expectations versus limitations on doctors' availability and the scope of services

- Patients' expectations of doctors' availability create unclear professional boundaries.
- Patients' expectations of the service scope of internet hospitals affect doctor-patient trust.

Advantages and downsides of online communication for comprehension and informed consent:

- Advantages
  - Doctors value having extra time to think carefully about replies to patients' messages, compared to in-person communication.
- Downsides
  - Internet hospitals' restrictions on consultation times, procedures, and arbitrary rules or schedules can hinder effective patient communication.
  - Doctors have concerns about the quality of online diagnoses and advice, as well as patient accuracy and comprehension, due to the limitations of online care.
  - Doctors have concerns about the completeness and uniformity of clinical informed consent in internet hospitals.

Results

Description of Study Participants

The 26 participants came from 3 different affiliated hospitals with 10, 5, and 11 participants interviewed from each hospital, respectively. Participants ranged in age from 29 to 49 years, and all 26 participants had PhD degrees. Only 5 participants stated that they had received specific training for working in internet hospitals, and 1 of them stated that training included discussion of clinical ethics in internet hospital care. We interviewed doctors from several departments involved in care for patients with chronic disease—oncology, cardiovasology, hematology, endocrinology, gastroenterology, nephrology, and infection departments. Aggregated participant characteristics are presented in Table 1.

XSL•FO
RenderX

Table 1. Participant characteristics (N=26).

Characteristics	Participants, n (%)
Gender	
Male	10 (38)
Female	16 (62)
Department[a]	
Oncology	6 (23)
Cardiovasology	6 (23)
Hematology	1 (4)
Endocrinology	2 (8)
Gastroenterology	3 (12)
Nephrology	5 (19)
Infection departments	3 (12)
Job status	
Resident physician	1 (4)
Attending physician	20 (77)
Deputy chief physician	4 (15)
Chief physician	1 (4)
Work experience (years)	
<5	2 (8)
5-10	13 (50)
>10	11 (42)
Internet hospital work experience (years)	
<1	7 (27)
1-3	12 (46)
≥4	7 (27)
Received training specific to internet hospital care	
Yes	5 (19)
No	21 (81)
Received clinical ethics training in internet hospital care	
Yes	1 (4)
No	25 (96)

[a]Some percentages may not add up to 100 due to rounding.

Advancements and Shortcomings in Care Access Due to Internet Hospitals: Follow-Up Care, Health Education, and Disparities

Most doctors stated that internet hospitals affiliated with their physical hospitals of employment were still in the early stages of development and that their internet hospital work experience mainly took place in enterprise-initiated internet hospitals. Doctors stated that internet hospitals initiated by physical hospitals were “not fully operational yet,” (Dr B) and “the consultation volume of patients is relatively small” (Dr C). They also suggested that there was currently a “lack of incentives” (Dr D) to work in internet hospitals initiated by physical hospitals, whereas enterprise-initiated internet hospitals offered “higher income” (Dr B) and a setting where “doctors set their own prices” (Dr C).

In both internet hospitals initiated by physical hospitals and enterprise-initiated internet hospitals, doctors stated that the majority of their work consisted of online consultation for common or easily diagnosable diseases, and follow-up services for patients with chronic diseases who had previously received care at physical hospitals, such as adjusting medications and ordering medical tests to be scheduled in person. Most doctors were motivated to work for internet hospitals particularly because of the opportunity to be part of expanding follow-up care for patients with chronic diseases.

XSL•FO
RenderX

> *Most of the patients who come to the internet hospitals are chronic disease patients, with conditions such as hypertension, diabetes, coronary heart disease, etc. These patients have been clearly diagnosed in our hospital, and some of them need to be guided or communicated with about what needs to be paid attention to in the process of home-based management. For example, patients' blood pressure might fluctuate, or they can consult online if they have any uncomfortable symptoms, which is quite common.* [Dr X]

Some doctors also believed based on experience that internet hospitals could serve as a more efficient channel to provide health education for patients, particularly for the management of chronic diseases.

> *We also feel that doctors in tertiary hospitals do not have much time to do health education with patients, but through the internet hospitals platforms, we can explain to patients the concept of health or a healthy way of life. For example, for a patient with heart failure, he isn't expected to come back to the hospital again and again, because I have instilled him with an understanding of healthy lifestyle and diet, and the workload of the doctor will be reduced in the long run.* [Dr C]

Despite the ways in which doctors felt internet hospitals expanded access to care and services, they had also observed disparities in access to internet hospitals across several groups, including older adults and patients with lower levels of education or technological literacy: "Older patients may not use smartphones or might need assistance to do so from family members" (Dr E).

> *Relatively speaking, if the patients come to the internet hospital for consultation, the education level of these patients will be higher, otherwise they will not be able to fully communicate with their doctor.* [Dr O]

Most doctors mentioned that internet hospitals are especially suitable for patients with chronic diseases. Doctors also stated that while older adults are one of the most common groups of patients with chronic diseases, many older adults have difficulties in using or accessing internet hospitals (Dr D and Dr X). Some doctors mentioned that internet hospitals currently have limited connections and overlap with local health services in rural areas. They believed moving toward more connection with local services was an important goal—particularly because the demand at tertiary hospitals frequently outstrips resources (Dr P and Dr Q), and because many rural patients travel long distances to receive care at tertiary hospitals.

> *Even for follow-up visits for chronic diseases or common diseases, many patients will still go to tertiary hospitals. Instead, the patient can go to a qualified local hospital and send us the results of the test, and then [through internet hospital care,] we can advise the patient or tell him how to adjust the medication, or refer the patient to a tertiary hospital for testing. But at present, internet hospitals have not played a big enough role in these aspects.* [Dr E]

Patients' Expectations Versus Limitations on Doctors' Availability and the Scope of Services: Implications for Doctor-Patient Boundaries and Trust

Patients' Expectations of Doctors' Availability Create Unclear Professional Boundaries

When asked about new challenges in patient care posed by internet hospitals, only 1 doctor mentioned risks related to patient privacy and data security.

> *The internet hospital platform where I am located requires patients to provide their name, gender, age, ID number, and other information, which can be seen on the doctor's portal, but as a doctor I will definitely not disclose the patient's private information but just give him diagnostic advice according to the necessary information provided by the patient.* [Dr G]

By contrast, many doctors expressed concerns about their own privacy. Some doctors shared stories from their work in physical hospitals of willingly sharing their personal WeChat with patients in case patients had questions after discharge (Dr D, Dr M, and Dr R). While some doctors did not seem to mind-bending this boundary with patients, others remembered negative experiences when patients had sent messages making demands of doctors' time at all hours (Dr E, Dr M, Dr P, and Dr Z). They also recalled times when patients obtained the doctor's personal contact information through their own means and contacted them after leaving the hospital without the doctor's consent (Dr V and Dr Z). As a result, some doctors had positive views of internet hospitals because they can serve as a means for online communication with patients without requiring the doctor to disclose their own personal contact information.

> *I prefer to use the official platform to communicate with patients, rather than through private WeChat or phone, because I really don't want to receive phone calls or text messages from patients after I work. But if the phone does ring, I will take into account that he is an old patient of mine, and I will still answer it, because I am not sure if there is any emergency. But for patients on such online hospital platforms, I rarely give them my phone number and personal WeChat.* [Dr P]

Several doctors were also uncomfortable that they were required to post information about themselves when working on internet hospital platforms, such as their name, location, and credentials (Dr P and Dr X). Because the audience of patients in internet hospitals is wider, they worried that patients who were dissatisfied with care may have the ability to post negative information about them in public forums online, citing their personal information. As a result, doctors stated that they would be more cautious in diagnosis and giving advice when dealing with patients in internet hospitals with whom they were less familiar (Dr J, Dr O, and Dr V).

XSL•FO
RenderX

Patients' Expectations of the Service Scope of Internet Hospitals Affect Doctor-Patient Trust

Another concern expressed by many doctors was that patients held unrealistic expectations of the scope of services that internet hospital doctors provide. Some doctors mentioned that some patients feel that just because they spend money in an internet hospital, they should be able to get all their problems solved at once or get immediate treatment (Dr E and Dr G)—when in reality in many cases, the doctor might need to conduct tests over multiple online consultations or might recommend that patients seek further medical services in offline, physical settings. Doctors were concerned that patients' dissatisfaction with unmet expectations might generate distrust toward the doctor.

> *What patients don't know is that in fact, most of the time, seeing a doctor is a step-by-step process, and it is necessary to do examinations step-by-step to exclude diseases or diagnose diseases. They often have high expectations for the effect of consultation in internet hospitals, and they think that doctors should be able to diagnose their diseases at one time; and not only to diagnose them, but also to propose a treatment plan.* [Dr E]

Some doctors suggested that this gap between expectations and reality could be especially strong for new patients whom the doctor had not seen previously for in-person care. They described that they often recommend for new internet hospital patients to go to physical hospitals to be examined before receiving further internet hospital care or advice, and that patients who are expecting immediate solutions can find this disappointing (Dr U and Dr X). One doctor suggested the need to educate patients and the general public on the scope of internet hospital care, in light of this mismatch in patients' and doctors' expectations (Dr Y).

However, some doctors raised concerns about issues related to doctor-patient trust that had less to do with adjusting patients' expectations and more to do with the format of online communication itself.

> *Doctor-patient trust will be better in physical hospitals. Because the doctor-patient relationship is a very special relationship, offline communication can observe the patient's expression, speed of speech, action, etc, and is more suitable for empathy with patients. When it comes to the doctor-patient relationship and trust, I think face-to-face consultation is still necessary.* [Dr J]

> *Face-to-face communication in physical hospitals may be more detailed, because if it is through text messages or phone calls, I may be able to talk to the patient in a few words, but if the patient is communicating in our hospital, it may take me half an hour. Because there is unequal information in medicine itself, the patient himself is not very clear about medicine, and without adequate communication, there is no trust between doctors and patients.* [Dr W]

Advantages and Downsides of Online Communication for Comprehension and Informed Consent

Doctors working in internet hospitals mainly used pictures and texts, and rarely video calls, to communicate with patients. Some doctors valued the extra time gained by this format to think carefully about their replies to patients' messages (Dr E, Dr U, and Dr X). However, most doctors pointed out how the lack of nonverbal communication could increase miscommunication and misunderstandings.

> *Online communication is through typing, and some doctors can't see the facial expressions of patients, which is very inconvenient. The communication between doctors and patients may need body language, facial expressions and other aspects.... I want the patient to really understand me in terms of attitude or tone or feeling or whatever.* [Dr G]

Doctors also expressed dissatisfaction with limitations on the time and procedures for consultation through various internet hospitals, and how sometimes arbitrary rules or schedules hindered communication with patients.

> *The doctors in our department need to be on duty every month in the internet hospitals. When it is my turn to be on duty, a patient will send his questions to me through the platform, but I think this mode is not good. For example, the patient might leave a message for me, but I am busy and don't reply to him in time, and he may not see my reply in time when I reply. If I go back and forth with him several times, this problem will not be solved until I come back on duty next month, and then the patient's problem will not be solved at all.* [Dr C]

> *The internet hospital at our hospital stipulates that patients can ask five questions at a time.... Sometimes doctors are not able to inquire in detail in order to understand the condition.* [Dr G]

Doctors had concerns about the quality of diagnoses and advice that they provided online, due to the inability to do direct physical examination. These concerns were intensified by their perception that many patients could not describe their symptoms clearly and accurately.

> *Currently, a lot of people still lack of basic medical knowledge, which will lead to ineffective or inefficient consultation on the internet, because they cannot describe their own symptoms, or cannot collect their own data and then summarize it. Patients cannot provide information about their condition sufficiently and accurately, which will seriously affect the efficiency of consultation.* [Dr E]

Two doctors also specifically mentioned the difficulty in internet hospital care of not being able to use the "four-diagnosis method"—a method used by doctors in traditional Chinese medicine for diagnosing illness, including diagnosis through observation, diagnosis through auscultation and olfaction, diagnosis through inquiry, and diagnosis through pulse feeling [43]. Although doctors in this study practiced mainly "Western" medicine, they described integrating certain traditional practices

XSL•FO
RenderX

such as this method into their care at physical hospitals (Dr D and Dr H).

In light of concerns about the potential for miscommunication with patients, a few doctors also expressed uncertainty about the completeness and uniformity of clinical informed consent as it is currently practiced in internet hospitals. While they believed that a standardized process of online informed consent for medical advice and treatment was needed, they did not know of any relevant laws or procedures.

> *Because we have so little time to communicate online, and such a narrow scope of care services, we don't usually obtain informed consent online. I might listen to the patient explain his symptoms. I might tell him what tests he needs before I give my advice, or if I'm dealing with a familiar patient, I might just prescribe his medication, so there's no need for informed consent. However, I think how to issue online informed consent, whether online informed consent is legally effective, and how to sign online informed consent all need to be considered. This is also for the protection of medical staff.* [Dr U]

Discussion

Principal Findings

This study sheds light on previously underresearched aspects of internet hospitals in China, as both the first interview study to examine physicians' perceptions of internet hospitals and one of the few studies on internet hospitals conducted in China outside of its most major cities. Our research revealed that physicians see enhanced opportunities in internet hospitals to conduct follow-up care for patients with chronic illnesses and to provide health education. However, physicians noted disparities in access for different groups, such as older adults, patients with lower education levels, patients with limited internet or technology access, and rural patients. One particularly novel finding was the conflict between patients' expectations and the reality of limitations on doctors' availability and the scope of services available through internet hospitals. Physicians perceived that this gap affected both boundaries and trust in the doctor-patient relationship. Physicians also discussed opportunities and challenges in doctor-patient communication, including issues of comprehension and informed consent. Considering that the development of internet hospitals involves multiple industries, including medical institutions, national policymaking departments, and technology providers, we raise several suggestions below on physician training, patient education, regulations, and design, as well as directions for future research.

Training for Doctors

Internet hospital care involves real-time online sharing of medical data. Information about both doctors and patients is centralized and easily accessible to authorized users on the internet hospital platform. Some doctors in our study were uncomfortable when required to publicly post their names, basic personal information, and credentials on internet hospital platforms because it might make them more vulnerable to public criticism. Doctors' reasons for being concerned about this were in line with previous research showing a high degree of conflict in the doctor-patient relationship in China [44,45]. This underscores the importance of current efforts both locally and internationally that aim to rebuild trust in the doctor-patient relationship [46-48]. Considering doctors' concerns about patients requesting for them to disclose their WeChat in both physical and internet hospital work, communication skills training for doctors should prepare doctors for how to interact with patients with empathy and care, while also maintaining their preferred professional boundaries.

It was also notable that only 1 doctor who was interviewed discussed concerns related to patient privacy and data security when asked about challenges presented by internet hospitals. By contrast, Li et al's [49] study on the determinants of patients' use of internet hospitals in China showed that while patients generally desire to use internet hospitals, they are apprehensive about the associated risk of their personal information being leaked. Due to the heightened potential for data leaks and breaches of patient health information associated with the use of internet hospitals, it is imperative that health care professionals undergo training to raise their awareness of data security precautions. For instance, physicians should be trained on proactive measures that they can take to guarantee that the internet hospital services they are affiliated with implement adequate security protocols around patient information. Furthermore, physicians should be trained to communicate with internet hospital patients or their legal proxies about potential risks related to data security and to apprise them of protective measures enacted to safeguard information. Future research should also evaluate the frequency with which data leaks and breaches in internet hospitals actually occur.

Findings from our study also suggest that internet hospitals have led to changes in doctor-patient communication. Doctors in our study considered it to be an advantage of internet hospital care that they generally had more time to communicate with patients compared to in-person care. However, a previous study on internet hospitals suggested that while doctors can obtain key information from patients within a few minutes through in-person communication and examination, information received in the same amount of time online tends to be more limited [43]. Research conducted by Deng et al [50] also highlights that engaging in online consultation work while simultaneously engaging in a main career providing in-person medical consultation may place excessive demands on doctors' time and energy. This phenomenon of work overload could potentially impede the widespread adoption of internet hospitals and introduce added risk to medical practice.

Relatedly, doctors in our study mentioned that when working in internet hospitals, they could only communicate with patients in the form of text, pictures, or video-based consultations, and that they had to rely largely on patient self-report. Both of these factors caused doctors to worry about the accuracy of their diagnoses. This aligns with recent research showing that about 70% of surveyed health care providers believe communication difficulties between patients and health care providers result in online consultations being insufficient [51], and about 70% of providers report feeling apprehensive about the possibility of

XSL•FO
RenderX

misdiagnosis when providing care through internet hospitals [51]. Recent research has also found that patients who use internet hospitals have more negative views on the doctor-patient relationship than nonusers—including both interpersonal factors such as the degree to which patients trust doctors and practical factors such as the degree to which patients agree with their doctors' medical opinions. Studies from other countries have similarly shown that telehealth can present new challenges or deficiencies in communication [52-55].

To address such challenges, telehealth communication competencies need to become a core component of both future research and physician training for internet hospitals in China—just as similar competencies are emerging as a priority for telehealth enhancement around the world [56]. Physicians providing care through internet hospitals should undergo standardized training for web-based communication skills, as research from other countries suggests such training can adapt interpersonal skills to the telehealth environment [57] and enhance empathic expression. More training for physicians on this skill set might reduce their apprehension about communicating through internet hospitals and assist them in communicating in a manner that improves outcomes for patients. Considering that doctors in our study expressed concern about patient comprehension and diagnostic accuracy, further research is also needed to evaluate and establish methods for measuring patient satisfaction, patient comprehension of information communicated by doctors, and diagnostic accuracy in internet hospital care. Future research should also examine the feasibility of integrating traditional Chinese medical practices such as the four-diagnosis method into telehealth care in China.

Education for Patients or the Public

Findings from our study highlight new challenges in the doctor-patient relationship posed by internet hospital care. One especially novel finding in our study was doctors' perception that patients subconsciously expected them to be online 24 hours a day, while doctors actually had limited hours working in the internet hospital and could not meet this expectation. Particularly when patients still needed to ask questions after the end of the physician's available time for consultation, doctors described the risk of conflicts with patients. These findings suggest many patients may be unaware when message-based interactions with physicians in internet hospitals are discontinuous or asynchronous. Therefore, public information about internet hospitals should specify the boundaries of physicians' availability for internet hospital consultations. While the scope of services may expand as internet hospitals continue to develop, information disseminated to the public should make it clear that internet hospital care is currently only intended for either follow-up care for previously diagnosed patients with chronic diseases, or for new patients with common and more easily diagnosable conditions. Finally, public education should equip patients or their proxies for distinct ways in which they might self-advocate for optimal care in the context of internet hospitals compared to in-person care. This may involve the development of interventions such as question-prompt lists that are specific to equipping patients for internet hospital consultations.

Regulations and Laws

Doctors in our study believed that difficulties with nonverbal communication in internet hospitals often led to miscommunication and misunderstanding, and many raised concerns that there were no specific laws regulating online doctor-patient communication. As a result, most doctors in our study expressed that they felt they were walking on eggshells concerning possible conflicts with patients. This builds on findings from the "2022 China E-hospital development research report" [51], in which one of the most common suggestions made by health care providers for the further development of internet hospitals was to standardize legal protection for doctors practicing in internet hospitals. Gaps in relevant laws and regulations may reduce the willingness of risk-conscious clinicians to provide medical services through internet hospitals.

Existing internet hospital laws and regulations in China are still mainly in the trial stage [51,58], and are being outpaced by evolving challenges in internet hospital care. The doctors in our study believed that internet hospitals may increase the difficulty of diagnosis and treatment, increase medical safety risks related to miscommunication, and increase the risk of medical malpractice liability. Research by Zhi et al describes how the inability of doctors to perform physical examinations or certain laboratory or imaging examinations through internet hospitals may compromise the accuracy of doctors' judgments [59]. However, the legal responsibilities of physical medical institutions, internet hospitals, and doctors regarding issues such as these have not been fully clarified. We suggest that further refinement and clarification of these and other aspects of internet hospital law will help doctors feel more protected in their work and increase the motivation of doctors to work in internet hospitals.

Doctors in our study mentioned that China also lacks detailed legal provisions on the implementation of online informed consent. Internationally recognized ethics standards highlight 4 core elements of informed consent—capacity to consent, information disclosure, comprehension, and voluntary authorization [60]. Informed consent issues involved in telehealth in other countries are similar to those described by doctors working in internet hospitals in this study, namely, the degree of discernment required from providers to ensure that patients are sufficiently informed to provide consent increases dramatically in telehealth [61]. In the United States, different states have different regulations on remote informed consent, and no federal policy has been formed at present. Some states require patients to fill out and sign written consent forms, while others do not [62]. In China, the Administrative Measures for internet hospitals stipulate that "internet hospitals must warn patients of risks and obtain informed consent from patients" [22]. However, current laws in China do not provide clear rules regarding the validity of electronic signatures for informed consent in internet hospital care. Informed consent in internet hospital care also involves unique information security issues due to the use of electronic health records, but there is currently no specific legal guidance for internet hospital platform developers or doctors concerning data security protection of informed consent in internet hospital care.

XSL•FO
RenderX

Tackling Disparities

Our research revealed that while older adults are at higher risk for chronic illness and are the main target population for internet hospitals, they are also reported by doctors to experience a number of barriers to internet hospital use. This finding aligns with research from various countries showing that older adults are less likely than younger patients to express positive attitudes toward using telehealth [63,64]. Health care providers in China and other countries have observed that older adults may be apprehensive about telehealth due to difficulty in operating computers or smartphones [65], may need the help of care partners to log into telehealth accounts successfully, and may need more time on average to download and set up applications [43,66]. Previous surveys have also shown that medical personnel believe that the low efficiency of online communication between doctors and patients and the low internet use rate of some patient groups (such as older adults) are the main factors hindering the development of internet hospitals [51].

Previous research in various countries has shown that the ease of use and perceived usefulness of telehealth systems have a positive impact on the acceptance of telehealth in patients who are older [67,68]. However, to date, China has not established an effective quality control system for internet hospitals [27,69], and the aforementioned ways in which internet hospitals currently pose increased risks for patient safety may affect general patients', let alone older adults' willingness to use them [49]. We recommend that the needs of older adults be considered in the design and development of internet hospital platforms and that older adults participate in the system design process [70,71]. Community health workers may be a workforce that could be mobilized to support telehealth training efforts among patients who are older, assist individuals with limited telehealth literacy in attending online appointments, and provide culturally and linguistically appropriate information about telehealth to rural patients and communities [64,72,73]. In general, health care organizations should invest in developing internet hospitals that are functional and easy to use. Drawing from research on telehealth design improvement in other countries, internet hospitals could be designed with features in mind to help physicians communicate more clearly with patients, such as providing notifications to physicians when patients read messages [74]. In addition, it may be beneficial for platforms to provide training materials to patients when patients register and log into internet hospitals for the first time [75]. Considering that a major goal of internet hospital development is to expand health care access, it will be crucial to address disparities in internet hospital use through these and other educational and design considerations.

Limitations

This study should be interpreted in the light of certain limitations. As most participants interviewed were attending physicians, findings may not be generalizable to the perspective of other health care workers or patients. The generalizability of our study findings from a single region and time point may also be limited, as there may be variations in internet hospital features and practice in other regions in China, and over time as internet hospitals continue to develop rapidly.

Conclusions

This study explored the experience and views of physicians in 3 tertiary hospitals in Changsha, China regarding access to care, patients' expectations versus the reality of services, and doctor-patient communication in internet hospital care. Findings from this study indicate that there is a need to train physicians in telehealth-specific communication skills. National policymaking departments should also further refine laws and regulations concerning internet hospitals, particularly those related to online informed consent. Technology developers should take the needs of older adults into particular account in the design of internet hospital platforms.

Acknowledgments

This study was supported by Major Program of National Social Science Fund of China, Research on Moral Issues in the Field of Contemporary Science and Technology (22&ZD044), the Xiangya Medical Humanities Series: Principles of Biomedical Ethics (monograph Award) Fund (KTZZPT019), Hunan Provincial Innovation Foundation for Postgraduate (CX20220133), the Fundamental Research Funds for the Central Universities of Central South University (2022ZZTS0041), the China Scholarship Council (CSC,202206370069), and the Natural Science of Changsha City (kq2202362).

Data Availability

The data generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YZ and XL conceptualized this study and designed the methodology. YZ and Xiaomin Wang conducted the interviews for data collection. Xuxi Wang and YZ transcribed the interviews. YZ, Xuxi Wang, YW, and XZ conducted and provided resources for preliminary analysis of the data. YZ and JH wrote and edited the paper. JH, XL, and XZ oversaw the implementation of all study activities. All authors read and approved the final paper.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 598 KB - jmir_v26i1e47523_app1.pdf]

References

1. Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Public Health 2021 Apr 01;42:463-481 [FREE Full text] [doi: 10.1146/annurev-publhealth-090519-093711] [Medline: 33798406]
2. Kirkland EB, DuBose-Morris R, Duckett A. Telehealth for the internal medicine resident: a 3-year longitudinal curriculum. J Telemed Telecare 2021 Oct;27(9):599-605. [doi: 10.1177/1357633X19896683] [Medline: 31888396]
3. Losorelli SD, Vendra V, Hildrew DM, Woodson EA, Brenner MJ, Sirjani DB. The future of telemedicine: revolutionizing health care or flash in the pan? Otolaryngol Head Neck Surg 2021 Aug;165(2):239-243. [doi: 10.1177/0194599820983330] [Medline: 33399500]
4. Hafner M, Yerushalmi E, Dufresne E, Gkousis E. The potential socio-economic impact of telemedicine in Canada. Rand Health Q 2022 Jun;9(3):6 [FREE Full text] [Medline: 35837517]
5. Alhamam NM, Buhalim RA, Almakhayitah IH, AlBahr AW, AlYaeesh IA. Telemedicine for musculoskeletal care during the COVID-19 pandemic: evaluating readiness of Saudi citizens. Cureus 2021 Feb 16;13(2):e13380 [FREE Full text] [doi: 10.7759/cureus.13380] [Medline: 33614363]
6. Bhaskar S, Bradley S, Chattu VK, Adisesh A, Nurtazina A, Kyrykbayeva S, et al. Telemedicine across the globe-position paper from the COVID-19 pandemic health system resilience PROGRAM (REPROGRAM) international consortium (part 1). Front Public Health 2020;8:556720 [FREE Full text] [doi: 10.3389/fpubh.2020.556720] [Medline: 33178656]
7. Cioti AM, Stanescu AMA, Grajdeanu IV, Serban B, Popescu E, Bratu OG, et al. Telemedicine in Europe—current status and future perspectives. Rev Med Mod 2019;26(4):165-168 [FREE Full text] [doi: 10.31689/rmm.2019.26.4.165]
8. Raja M, Bjerkan J, Kymre IG, Galvin KT, Uhrenfeldt L. Telehealth and digital developments in society that persons 75 years and older in European countries have been part of: a scoping review. BMC Health Serv Res 2021 Oct 26;21(1):1157 [FREE Full text] [doi: 10.1186/s12913-021-07154-0] [Medline: 34696789]
9. Brucksch S. Japan and its rapid ageing society: does e-health technology provide a solution? J Aging Sci 2018;6(2):1-4 [FREE Full text] [doi: 10.4172/2329-8847.1000192]
10. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020 Jul 29;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
11. Scott R, Mars M. Telehealth in the developing world: current status and future prospects. SHTT 2015 Feb;2015(3):25-37 [FREE Full text] [doi: 10.2147/SHTT.S75184]
12. Yip W, Fu H, Chen AT, Zhai T, Jian W, Xu R, et al. 10 years of health-care reform in China: progress and gaps in Universal Health Coverage. Lancet 2019 Sep 28;394(10204):1192-1204. [doi: 10.1016/S0140-6736(19)32136-1] [Medline: 31571602]
13. 2021 China E-hospital development report. National Telemedicine and Connected Health Center and CN-Healthcare. 2021. URL: https://zk.cn-healthcare.com/doc-show-53644.html [accessed 2023-09-17]
14. Chen X, Wu X, Zhang Q, Jing R, Cheng W, Tian J, et al. The construction and operational models of internet hospitals in China: a hospital-based survey study. BMC Health Serv Res 2023 Jun 21;23(1):669 [FREE Full text] [doi: 10.1186/s12913-023-09675-2] [Medline: 37344831]
15. Houston TK, Allison JJ. Users of Internet health information: differences by health status. J Med Internet Res 2002;4(2):E7 [FREE Full text] [doi: 10.2196/jmir.4.2.e7] [Medline: 12554554]
16. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021 Jul 20;23(7):e22330 [FREE Full text] [doi: 10.2196/22330] [Medline: 34283025]
17. Yu T, Jin C, Wu X, Yue D. Implementation of shared decision-making within internet hospitals in China based on patients' needs: feasibility study and content analysis. JMIR Form Res 2023 Jan 06;7:e39965 [FREE Full text] [doi: 10.2196/39965] [Medline: 36607710]
18. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015 Aug;3(8):e445-e446 [FREE Full text] [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
19. Guiding opinions of the State Council on actively promoting the "Internet +" action. The State Council of the People's Republic of China. 2015. URL: https://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm [accessed 2023-09-16]
20. Opinions on promoting the development of "Internet + medical health". General Office of the State Council. 2018. URL: https://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm [accessed 2023-09-16]
21. Internet diagnosis and treatment management measures (trial implementation). National Health Commission of the People's Republic of China. 2018. URL: https://www.gov.cn/gongbao/content/2019/content_5358684.htm [accessed 2023-09-16]
22. Internet hospital management measures (trial implementation). National Health Commission of the People's Republic of China. 2018. URL: https://www.gov.cn/gongbao/content/2019/content_5358684.htm [accessed 2023-09-16]
23. Telemedicine service management standards (trial implementation). National Health Commission of the People's Republic of China. 2018. URL: https://www.gov.cn/gongbao/content/2019/content_5358684.htm [accessed 2023-09-16]

XSL•FO
RenderX

24. Notice of the general office of the NHC on strengthening informatization to support the prevention and control of the novel coronavirus pneumonia epidemic. National Health Commission of China. 2020. URL: https://www.gov.cn/zhengce/zhengceku/2020-02/05/content_5474692.htm [accessed 2023-09-16]
25. The 52nd statistical report on China's internet development. China Internet Network Information Center. 2023. URL: https://www.cnnic.cn/NMediaFile/2023/0908/MAIN1694151810549M3LV0UWOAV.pdf [accessed 2024-02-28]
26. There are more than 3,000 internet hospitals, and it is necessary to explore a better development path. Health Times. 2023. URL: http://www.jksb.com.cn/index.php?m=content&c=index&a=show&catid=788&id=200853 [accessed 2024-02-28]
27. Qiu Y, Liu Y, Ren W, Qiu Y, Ren J. Internet-based and mobile-based general practice: cross-sectional survey. J Med Internet Res 2018 Sep 25;20(9):e266 [FREE Full text] [doi: 10.2196/jmir.8378] [Medline: 30257819]
28. Xie X, Zhou W, Lin L, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017 Jul 04;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]
29. Yuan J, He F. Status quo and problems of the internet hospitals in China. Chin Rural Health Serv Adm 2023;43(02):126-128+140. [doi: 10.19955/j.cnki.1005-5916.2023.02.009]
30. Pogorzelska K, Chlabicz S. Patient satisfaction with telemedicine during the COVID-19 pandemic-a systematic review. Int J Environ Res Public Health 2022;19(10):6113 [FREE Full text] [doi: 10.3390/ijerph19106113] [Medline: 35627650]
31. Sagaro GG, Battineni G, Amenta F. Barriers to sustainable telemedicine implementation in Ethiopia: a systematic review. Telemed Rep 2020;1(1):8-15 [FREE Full text] [doi: 10.1089/tmr.2020.0002] [Medline: 35722252]
32. Şahin E, Yavuz Veizi BG, Naharci MI. Telemedicine interventions for older adults: a systematic review. J Telemed Telecare 2024 Feb;30(2):305-319. [doi: 10.1177/1357633X211058340] [Medline: 34825609]
33. Imlach F, McKinlay E, Middleton L, Kennedy J, Pledger M, Russell L, et al. Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. BMC Fam Pract 2020 Dec 13;21(1):269 [FREE Full text] [doi: 10.1186/s12875-020-01336-1] [Medline: 33308161]
34. Kumar S, Kumar A, Kumar M, Kumar A, Arora R, Sehrawat R. Feasibility of telemedicine in maintaining follow-up of orthopaedic patients and their satisfaction: a preliminary study. J Clin Orthop Trauma 2020 Oct;11(Suppl 5):S704-S710 [FREE Full text] [doi: 10.1016/j.jcot.2020.07.026] [Medline: 32837105]
35. Latulipe C, Gatto A, Nguyen HT, Miller DP, Quandt SA, Bertoni AG, et al. Design considerations for patient portal adoption by low-income, older adults. Proc SIGCHI Conf Hum Factor Comput Syst 2015 Apr;2015:3859-3868 [FREE Full text] [doi: 10.1145/2702123.2702392] [Medline: 27077140]
36. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in access to and use of electronic personal health information between rural and urban residents in the United States. J Rural Health 2018;34(Suppl 1):s30-s38 [FREE Full text] [doi: 10.1111/jrh.12228] [Medline: 28075508]
37. Johnson BA, Lindgren BR, Blaes AH, Parsons HM, LaRocca CJ, Farah R, et al. The new normal? Patient satisfaction and usability of telemedicine in breast cancer care. Ann Surg Oncol 2021;28(10):5668-5676 [FREE Full text] [doi: 10.1245/s10434-021-10448-6] [Medline: 34275045]
38. Solimini R, Busardò FP, Gibelli F, Sirignano A, Ricci G. Ethical and legal challenges of telemedicine in the era of the COVID-19 pandemic. Medicina (Kaunas) 2021;57(12):1314 [FREE Full text] [doi: 10.3390/medicina57121314] [Medline: 34946259]
39. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
40. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
41. Klem NR, Bunzli S, Smith A, Shields N. Demystifying qualitative research for musculoskeletal practitioners part 5: rigor in qualitative research. J Orthop Sports Phys Ther 2022;52(2):60-62. [doi: 10.2519/jospt.2022.10487] [Medline: 35100817]
42. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
43. Li Y, Hu H, Rozanova L, Fabre G. COVID-19 and internet hospital development in China. Epidemiologia (Basel) 2022;3(2):269-284 [FREE Full text] [doi: 10.3390/epidemiologia3020021] [Medline: 36417257]
44. Jiang MM, Wu ZY, Tu AX. Research on the cooperative governance path of multiple stakeholders in doctor-patient disputes under the environment of information asymmetry. Int J Environ Res Public Health 2023;20(2):1597 [FREE Full text] [doi: 10.3390/ijerph20021597] [Medline: 36674353]
45. Zhao D, Zhang Z. Changes in public trust in physicians: empirical evidence from China. Front Med 2019;13(4):504-510. [doi: 10.1007/s11684-018-0666-4] [Medline: 30194619]
46. Du L, Xu J, Chen X, Zhu X, Zhang Y, Wu R, et al. Rebuild doctor-patient trust in medical service delivery in China. Sci Rep 2020;10(1):21956 [FREE Full text] [doi: 10.1038/s41598-020-78921-y] [Medline: 33319841]
47. Fu Y, Tang T, Long J, Lin B, Li J, Quan G, et al. Factors associated with using the internet for medical information based on the doctor-patient trust model: a cross-sectional study. BMC Health Serv Res 2021;21(1):1268 [FREE Full text] [doi: 10.1186/s12913-021-07283-6] [Medline: 34819044]

XSL•FO
RenderX

48. Tucker JD, Wong B, Nie JB, Kleinman A, Patient-Physician Trust Team. Rebuilding patient-physician trust in China. Lancet 2016;388(10046):755 [FREE Full text] [doi: 10.1016/S0140-6736(16)31362-9] [Medline: 27560268]
49. Li D, Hu Y, Pfaff H, Wang L, Deng L, Lu C, et al. Determinants of patients' intention to use the online inquiry services provided by internet hospitals: empirical evidence from China. J Med Internet Res 2020;22(10):e22716 [FREE Full text] [doi: 10.2196/22716] [Medline: 33006941]
50. Deng W, Yang T, Deng J, Liu R, Sun X, Li G, et al. Investigating factors influencing medical practitioners' resistance to and adoption of internet hospitals in China: mixed methods study. J Med Internet Res 2023;25:e46621 [FREE Full text] [doi: 10.2196/46621] [Medline: 37523226]
51. 2022 China E-hospital development research report. Shanghai Jiao Tong University. 2022. URL: https://www.100ec.cn/detail--6620464.html [accessed 2023-01-08]
52. Alpert JM, Dyer KE, Lafata JE. Patient-centered communication in digital medical encounters. Patient Educ Couns 2017;100(10):1852-1858 [FREE Full text] [doi: 10.1016/j.pec.2017.04.019] [Medline: 28522229]
53. Alpert JM, Krist AH, Aycock RA, Kreps GL. Applying multiple methods to comprehensively evaluate a patient portal's effectiveness to convey information to patients. J Med Internet Res 2016 May 17;18(5):e112 [FREE Full text] [doi: 10.2196/jmir.5451] [Medline: 27188953]
54. Gerber DE, Beg MS, Duncan T, Gill M, Craddock Lee SJ. Oncology nursing perceptions of patient electronic portal use: a qualitative analysis. Oncol Nurs Forum 2017 Mar 01;44(2):165-170 [FREE Full text] [doi: 10.1188/17.onf.165-170] [Medline: 28222081]
55. Sieck CJ, Hefner JL, Schnierle J, Florian H, Agarwal A, Rundell K, et al. The rules of engagement: perspectives on secure messaging from experienced ambulatory patient portal users. JMIR Med Inform 2017;5(3):e13 [FREE Full text] [doi: 10.2196/medinform.7516] [Medline: 28676467]
56. van Galen LS, Wang CJ, Nanayakkara PWB, Paranjape K, Kramer MHH, Car J. Telehealth requires expansion of physicians' communication competencies training. Med Teach 2019;41(6):714-715. [doi: 10.1080/0142159X.2018.1481284] [Medline: 29944031]
57. Wright HH, O'Shea MC, Sekula J, Mitchell LJ. Assessment of communication skills using telehealth: considerations for educators. Front Med (Lausanne) 2022;9:841309 [FREE Full text] [doi: 10.3389/fmed.2022.841309] [Medline: 35979204]
58. Huang M, Wang J, Nicholas S, Maitland E, Guo Z. Development, status quo, and challenges to China's health informatization during COVID-19: evaluation and recommendations. J Med Internet Res 2021;23(6):e27345 [FREE Full text] [doi: 10.2196/27345] [Medline: 34061761]
59. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021;23(9):e18307 [FREE Full text] [doi: 10.2196/18307] [Medline: 34342267]
60. World Medical Association. World Medical Association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194 [FREE Full text] [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
61. Blake JH, Schwemmer MK, Sade RM. The patient-surgeon relationship in the cyber era: communication and information. Thorac Surg Clin 2012;22(4):531-538 [FREE Full text] [doi: 10.1016/j.thorsurg.2012.07.002] [Medline: 23084616]
62. Fields BG. Regulatory, legal, and ethical considerations of telemedicine. Sleep Med Clin 2020;15(3):409-416 [FREE Full text] [doi: 10.1016/j.jsmc.2020.06.004] [Medline: 32762973]
63. Chiang KF, Wang HH, Chien IK, Liou JK, Hung CL, Huang CM, et al. Healthcare providers' perceptions of barriers in implementing of home telecare in Taiwan: a qualitative study. Int J Med Inform 2015;84(4):277-287. [doi: 10.1016/j.ijmedinf.2015.01.007] [Medline: 25649842]
64. Gray DM, Joseph JJ, Olayiwola JN. Strategies for digital care of vulnerable patients in a COVID-19 world-keeping in touch. JAMA Health Forum 2020;1(6):e200734 [FREE Full text] [doi: 10.1001/jamahealthforum.2020.0734] [Medline: 36218529]
65. Miller DP, Latulipe C, Melius KA, Quandt SA, Arcury TA. Primary care providers' views of patient portals: interview study of perceived benefits and consequences. J Med Internet Res 2016;18(1):e8 [FREE Full text] [doi: 10.2196/jmir.4953] [Medline: 26772771]
66. Varsi C, Gammon D, Wibe T, Ruland CM. Patients' reported reasons for non-use of an internet-based patient-provider communication service: qualitative interview study. J Med Internet Res 2013;15(11):e246 [FREE Full text] [doi: 10.2196/jmir.2683] [Medline: 24220233]
67. Cimperman M, Brenčič MM, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an extended UTAUT model. Int J Med Inform 2016;90:22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
68. Zhou M, Zhao L, Kong N, Campy KS, Qu S, Wang S. Factors influencing behavior intentions to telehealth by Chinese elderly: an extended TAM model. Int J Med Inform 2019;126:118-127. [doi: 10.1016/j.ijmedinf.2019.04.001] [Medline: 31029253]
69. 2020 research report on the development of internet hospitals in China. Institute of Healthcare. 2020. URL: http://zk.cn-healthcare.com/doc-show-39773.html [accessed 2023-09-17]
70. Bhattarai P, Phillips JL. The role of digital health technologies in management of pain in older people: an integrative review. Arch Gerontol Geriatr 2017;68:14-24. [doi: 10.1016/j.archger.2016.08.008] [Medline: 27584871]

XSL•FO
RenderX

71. Cosco TD, Fortuna K, Wister A, Riadi I, Wagner K, Sixsmith A. COVID-19, social isolation, and mental health among older adults: a digital catch-22. J Med Internet Res 2021;23(5):e21864 [FREE Full text] [doi: 10.2196/21864] [Medline: 33891557]
72. Huang W, Long H, Li J, Tao S, Zheng P, Tang S, et al. Delivery of public health services by Community Health Workers (CHWs) in primary health care settings in China: a systematic review (1996-2016). Glob Health Res Policy 2018;3:18 [FREE Full text] [doi: 10.1186/s41256-018-0072-0] [Medline: 29992191]
73. Lohr AM, Ingram M, Nuñez AV, Reinschmidt KM, Carvajal SC. Community-clinical linkages with community health workers in the United States: a scoping review. Health Promot Pract 2018;19(3):349-360 [FREE Full text] [doi: 10.1177/1524839918754868] [Medline: 29363334]
74. Laukka E, Huhtakangas M, Heponiemi T, Kujala S, Kaihlanen AM, Gluschkoff K, et al. Health care professionals' experiences of patient-professional communication over patient portals: systematic review of qualitative studies. J Med Internet Res 2020;22(12):e21623 [FREE Full text] [doi: 10.2196/21623] [Medline: 33289674]
75. Hefner JL, Sieck CJ, Walker DM. Patient and physician perspectives on training to improve communication through secure messaging: clarifying the rules of engagement. Health Care Manage Rev 2022;47(1):3-11. [doi: 10.1097/HMR.0000000000000279] [Medline: 32379081]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by T de Azevedo Cardoso; submitted 22.03.23; peer-reviewed by Y Cao, N Mungoli, A Gangadhara Rao; comments to author 06.09.23; revised version received 27.09.23; accepted 26.02.24; published 29.03.24.

Please cite as:
Zhong Y, Hahne J, Wang X, Wang X, Wu Y, Zhang X, Liu X
Telehealth Care Through Internet Hospitals in China: Qualitative Interview Study of Physicians' Views on Access, Expectations, and Communication
J Med Internet Res 2024;26:e47523
URL: https://www.jmir.org/2024/1/e47523
doi:10.2196/47523
PMID:38551618

©Yuqiong Zhong, Jessica Hahne, Xiaomin Wang, Xuxi Wang, Ying Wu, Xin Zhang, Xing Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impacts of an Acute Care Telenursing Program on Discharge, Patient Experience, and Nursing Experience: Retrospective Cohort Comparison Study

Courtenay R Bruce[1], MA, JD; Steve Klahn[1], RN, MBA; Lindsay Randle[1], MBA; Xin Li[1], BS; Kelkar Sayali[1], BS; Barbara Johnson[1], BSN, MBA, DNP; Melissa Gomez[1], MBA; Meagan Howard[1], MHA; Roberta Schwartz[1], PhD; Farzan Sasangohar[1], PhD

Houston Methodist, Houston, TX, United States

Corresponding Author:
Courtenay R Bruce, MA, JD
Houston Methodist
8100 Greenbriar Drive
Houston, TX, 77030
United States
Phone: 1 281 620 9040
Email: crbruce@houstonmethodist.org

Abstract

Background: Despite widespread growth of televisits and telemedicine, it is unclear how telenursing could be applied to augment nurse labor and support nursing.

Objective: This study evaluated a large-scale acute care telenurse (ACTN) program to support web-based admission and discharge processes for hospitalized patients.

Methods: A retrospective, observational cohort comparison was performed in a large academic hospital system (approximately 2100 beds) in Houston, Texas, comparing patients in our pilot units for the ACTN program (telenursing cohort) between June 15, 2022, and December 31, 2022, with patients who did not participate (nontelenursing cohort) in the same units and timeframe. We used a case mix index analysis to confirm comparable patient cases between groups. The outcomes investigated were patient experience, measured using the Hospital Consumer Assessment of Health Care Providers and Systems (HCAHCPS) survey; nursing experience, measured by a web-based questionnaire with quantitative multiple-choice and qualitative open-ended questions; time of discharge during the day (from electronic health record data); and duration of discharge education processes.

Results: Case mix index analysis found no significant case differences between cohorts (P=.75). For the first 4 units that rolled out in phase 1, all units experienced improvement in at least 4 and up to 7 HCAHCPS domains. Scores for “communication with doctors” and “would recommend hospital” were improved significantly (P=.03 and P=.04, respectively) in 1 unit in phase 1. The impact of telenursing in phases 2 and 3 was mixed. However, “communication with doctors” was significantly improved in 2 units (P=.049 and P=.002), and the overall rating of the hospital and the ”would recommend hospital” scores were significantly improved in 1 unit (P=.02 and P=04, respectively). Of 289 nurses who were invited to participate in the survey, 106 completed the nursing experience survey (response rate 106/289, 36.7%). Of the 106 nurses, 101 (95.3%) indicated that the ACTN program was very helpful or somewhat helpful to them as bedside nurses. The only noticeable difference between the telenursing and nontelenursing cohorts for the time of day discharge was a shift in the volume of patients discharged before 2 PM compared to those discharged after 2 PM at a hospital-wide level. The ACTN admissions averaged 12 minutes and 6 seconds (SD 7 min and 29 s), and the discharges averaged 14 minutes and 51 seconds (SD 8 min and 10 s). The average duration for ACTN calls was 13 minutes and 17 seconds (SD 7 min and 52 s). Traditional cohort standard practice (nontelenursing cohort) of a bedside nurse engaging in discharge and admission processes was 45 minutes, consistent with our preimplementation time study.

Conclusions: This study shows that ACTN programs are feasible and associated with improved outcomes for patient and nursing experience and reducing time allocated to admission and discharge education.

(J Med Internet Res 2024;26:e54330) doi:10.2196/54330

XSL•FO
RenderX

KEYWORDS

telenursing; telemedicine; patient discharge; health personnel; surveys and questionnaires; patient outcome assessment

Introduction

Telemedicine, particularly video televisits, has greatly expanded in the wake of the COVID-19 pandemic [1,2]. Televisits have shown promise as a robust, practical, efficacious, and scalable alternative to in-person office visits that could ameliorate labor supply shortages [3,4]. The published evidence suggests a generally positive attitude toward televisit appointments for chronic care, focused on addressing financial and transportation barriers and improving patients' access to care [5-7]. Despite the promise shown by televisits, limited attention has been paid to applying this method in the acute care setting and, in particular, on how this promising technology can be leveraged to support nurses.

Estimates suggest that approximately 200,000 open nursing positions will become available each year between 2021 and 2031 [8]. Telenursing can augment nursing labor supply, decrease nursing workload, maintain patient and nurse safety, and positively impact nursing and patient experiences [9]. However, the impact of telenursing on outcomes in acute care settings remains a research gap.

To address this gap, this study aimed to evaluate the outcomes associated with a large-scale acute care telenurse (ACTN) program to support web-based admission and discharge processes for hospitalized patients compared to patients who did not undergo the ACTN program intervention. Admission and discharge are 2 substantive and time-consuming acute care nursing tasks that involve tedious documentation in the electronic health record (EHR) and extensive interaction with patients and families to gather history and provide patient education [10,11]. We aimed to develop an ACTN program to augment nursing care by conducting admission and discharge processes through telenursing in a large health system. Subsequently, we discuss the impacts on 4 end points: patient experience, nursing experience, time of discharge during the day, and length of time for discharge education processes. We hypothesized that the ACTN program would be associated with higher patient experience scores and improved nursing experience compared to standard admission and discharge practices.

Methods

Overview

This study was conducted in a large academic hospital system (approximately 2100 beds) in Houston, Texas. The preimplementation methods are reported more extensively in the studies by Hehman et al [12] and Schwartz et al [13]. Program implementation was first informed by nursing time and workload surveys and pilot implementation in 4 comparatively understaffed units. The chief innovation officer, along with nursing leaders and ACTN program administrators, met with the bedside nursing staff of these 4 understaffed units to solicit their input on where and how ACTN would add value to their workflow. Bedside nursing staff provided critical input on admission processes that could be delegated to individuals working remotely with no perceived negative impact on patient experience. We conducted participatory workflow design sessions with bedside nursing staff on the ACTN program to cocreate workflow integration points where the remote team could assist [13].

Pilot Implementation and Procedures

Before implementation, the ACTN administrators trained bedside nurses in pilot units by demonstrating the use of technology during shift huddles. Then, the trainers presented slides on contact information and available support and provided a role demarcation process map, showing what the remote telenurse staff would be doing compared to what the bedside nurses needed to do to launch and conduct discharge education. Furthermore, the trainers invited the nursing staff to observe several discharges to learn how to conduct them. A software with Health Insurance Portability and Accountability Act compliance was uploaded to iPads (Apple Inc) and stored on each unit. Handheld iPads were available, and roaming iPads were made available for patients who could not hold an iPad.

The pilot implementation was staggered in a phased rollout, consisting of 3 sequenced phases, as shown in Figure 1. Upon admission, the acute care bedside nurse contextualized the ACTN program with patients and families by handing the patient an iPad with a preloaded remote program app (Caregility) and then pressing a soft key to allow the ACTN to enter the patient's room via the iPad screen. The ACTN introduced themselves, completed the nursing admission profile in the EHR, placed a request for a consultation, and notified the bedside nurse that the admission was completed using secure SMS text messaging [13]. A similar process was followed for discharge workflow processes, where the ACTN completed patient education on discharge instructions, confirmed the patient's pharmacy details, confirmed discharge transportation, and arranged for departure.

XSL•FO
RenderX

Figure 1. Overview of the 3-phase pilot implementation in 12 units. CV: cardiovascular; CVIMU: cardiovascular intermediate unit; GI: gastrointestinal.

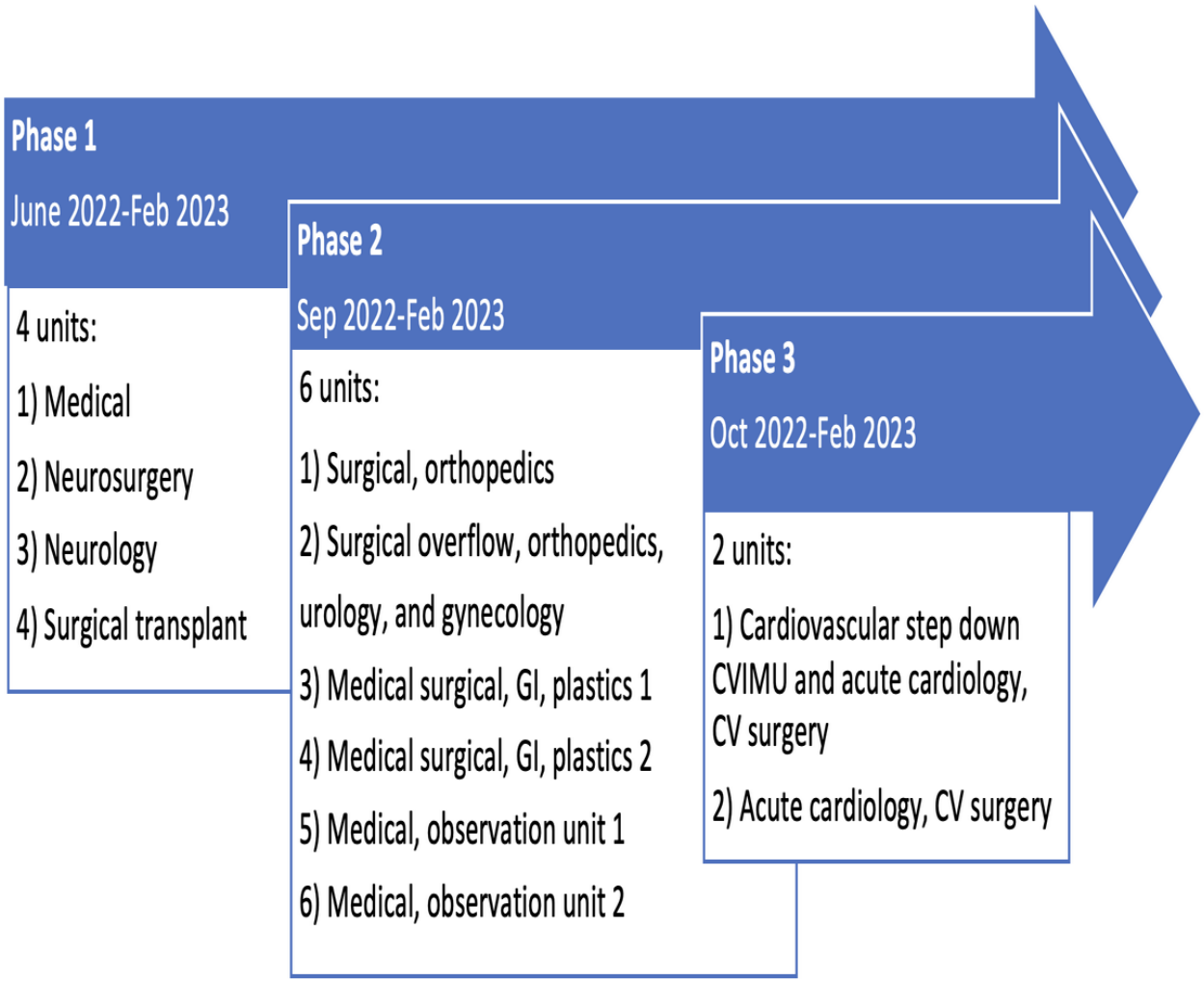

Bedside nurses used their discretion regarding which patients would be appropriate for the ACTN program. They based this determination principally on whether documentation was needed and whether the patient could benefit from the undivided attention the ACTN program could afford. Furthermore, they excluded patients from the ACTN program if the patients expressed discomfort using an iPad. After the initial rollout, patients' input was sought on their experience with the ACTN program to identify where and how improvements could be made, and this feedback was incorporated into iterative revisions in subsequent rollouts.

Pilot Outcomes Monitoring

A retrospective, observational cohort comparison was performed, in which all patients in our pilot units for the ACTN program (telenursing cohort) between June 15, 2022, and December 31, 2022, were compared with all patients who did not participate (nontelenursing cohort) in the same units in the same timeframe.

Our primary outcomes were patient experience and nursing experience. Patient experience scope was any process observable by patients [14]. We compared patient experiences in the telenursing and nontelenursing cohorts by evaluating patients' responses to the widely used Hospital Consumer Assessment of Health Care Providers and Systems (HCAHPS) survey [15], which represented 8 aspects (called dimensions) of patient satisfaction. Each dimension was measured using a continuous variable (0 to 100 points).

For the telenursing cohort, we analyzed bedside nurses' collective responses using a Forms (Microsoft Corp) survey conducted in April 2023. The survey consisted of 5 questions, asking them to indicate whether the ACTN program was helpful using a Likert scale with 5 items (very helpful to very unhelpful). Nurses were asked to provide open-ended comments to explain the reasons for their evaluation. At the end of the survey, we included 2 open-ended fields for nurses to describe opportunities for improvement in future rollouts and provide any additional comments.

Furthermore, we explored the time at which discharge occurred using the EHR admission, discharge, and transfer date and time. We compared the hour of the day the patient was discharged in the telenursing cohort with the hour of the day the patient was discharged in the nontelenursing cohort, hypothesizing a priori that patients might be discharged earlier in the day in the telenursing cohort. Finally, we analyzed the duration of discharge education for both cohorts, measured in minutes.

XSL•FO
RenderX

Data Analysis

The patient demographic data were available for all patients. To confirm that the telenursing cohort had similar patient demographics as the nontelenursing cohort (and therefore to confirm that nurse biases in patient selection for the ACTN program were unlikely), we conducted a case mix index (CMI) evaluation. We first isolated the population of both cohorts into adults (aged ≥18 y). We compared only those patients who were discharged home and excluded those who were on extracorporeal membrane oxygenation or those who underwent a tracheostomy. The remaining population was evaluated to determine whether there was a difference in patient acuity and severity. After confirming that patient acuity and severity were of no significant difference, we included the inpatient and observation populations to evaluate the intervention results.

For the patient experience data, independent sample *t* tests (2-tailed) were used to compare the telenursing and nontelenursing cohorts across different HCAHPS dimensions and units. Analysis was conducted using R software (R Foundation for Statistical Computing). For the nursing experience survey data, we used Excel (Microsoft Corp) to analyze the responses to multiple-choice, discrete questions and thematic analysis to evaluate the open-text fields. Thematic analysis allows for eliciting key themes that emerge based on recurring statements [16]. The analysis followed an inductive approach. This approach uses open-ended questions, allowing themes to emerge with a few previously articulated assumptions on responses. Given the limited content, CRB served as the primary coder. Coding labels were used for data contextualizing, allowing for new themes to emerge throughout the coding process, using a codebook [16,17]. We stored emergent patterns and themes in an electronic format.

Ethical Considerations

The hospital's review board determined that the ACTN pilot would not be considered regulated human subjects research. All data reported in this study were aggregated and deidentified.

Results

Overview

The demographics of the telenursing and nontelenursing cohorts were relatively similar. Both cohorts had an average age of 60 years with an SD of 16.91; had a similar distribution in race and ethnicity (approximately 92/2319, 3.96% Asian; 525/2319, 22.64% Black; 425/2319, 18.33% Hispanic; 70/2319, 3.02% Native American, declined to identify, or other categories; and 1202/2319, 51.83% White); and had a similar distribution in female participants versus male participants (1249/2319, 53.86% vs 1070/2319, 46.14%). To further understand the population, the CMI analysis for acuity and severity showed that the CMI was slightly higher in the telenursing cohort than in the nontelenursing cohort, but the difference was not statistically significant (*P*=.75).

Patient Experience

Among the first 4 units that rolled out in phase 1, all units experienced improvement in at least 4 and up to 7 HCAHPS domains (Table S1 in Multimedia Appendix 1). On average, 6 out of 8 HCAHPS domains were improved for patients in the telenursing cohort. All 4 units experienced improvements in the "overall rating" domain, and 3 of the 4 units experienced improvements in "likelihood to recommend" domain for patients in the telenursing cohort compared to those in the nontelenursing cohort within the same units. The improvement scores ranged from 1.4% for the neurosurgery unit (36 beds) to 11.6% for the medical unit (37 beds). Furthermore, all 4 units in the first phase of roll out experienced improved scores in the "responsiveness" domain by >4 points (ranging from 5% to 10.1%). A total of 2 out of the 4 units also experienced improvements in the "communication with nurses" (ranging from 1.7% to 3%) and "communication about medicines" (ranging from 3.3% to 11.7%) domains. The 2 units that did not experience improvement in the communication domains were the combined medical and surgery neurology and neurosurgical units (36 beds). Only the neurosurgical unit showed statistically significant improvements in 2 dimensions: "communication with doctors" (*P*=.03) and "would recommend hospital" (*P*=.04).

For the 7 units that rolled out during phase 2, only 1 orthopedic surgery unit (28 beds) experienced improvements in every domain (ranging from 0.9% to 12.5%). Medical observation unit 1 also improved in 5 areas. However, only improvements in "communication with doctors" (*P*=.002), "overall rating of hospital" (*P*=.02), and "would recommend hospital" (*P*=.04) were statistically significant. The remaining units experienced improvements in some domains for the telenursing cohort compared to the nontelenursing cohort, with no improvement in other domains. However, the scores for "communication with nurses" and "communication with doctors" domains were improved for most of the units that rolled out in phase 2 (Table S2 in Multimedia Appendix 1).

For the 2 units that rolled out in phase 3, both of which were surgical cardiac units with 36 beds, 1 unit experienced improvement in every domain except "responsiveness" (ranging from 1% to 12%). The other unit only experienced improvement in the "communication with doctors" (4.9%) and "care transitions" domains (1.1%). However, none of these improvements were statistically significant (Table S3 in Multimedia Appendix 1).

Nursing Experience

Of the 289 nurses who were invited to participate in the survey, 106 completed the survey (36.7% response rate). Of the 106 nurses, 101 (95.3%) indicated that the ACTN program was "very helpful" or "somewhat helpful" to them as bedside nurses.

Quantitative Findings

The main reasons nurses gave for the program's helpfulness included that it saved them time (94/106, 88.7%), allowed them to focus on more urgent clinical needs (90/106, 84.9%), allowed them to focus on activities they felt were more in line with their skill level (55/106, 51.9%), and allowed patients to have undivided attention for their discharge education (52/106, 49.1%). Among the 5 nurses who indicated that the ACTN program was somewhat unhelpful or very unhelpful, 3 (60%) indicated that workflows were not clear or needed further

XSL•FO
RenderX

refinement or clarification. Furthermore, the nurse respondents shared several barriers and provided opportunities for improvement, with 91 (85.8%) out of 106 nurses offering suggestions.

Qualitative Findings

Overview

For the free-text explanation fields, all but 3 nurses (103/106, 97.2%) provided additional comments on the ACTN program helpfulness. Three themes emerged from the qualitative analysis of the free-text comments: (1) most of the nurses' comments reflected that telenurses help bedside nurses save time, (2) respondents indicated that extra hands provided emotional and physical support in providing patient care, and (3) respondents perceived an improvement in patient safety by having a telenurse who could "catch missed" issues.

Time Saving

One of the perceived benefits of the telenursing program was saving time. One nurse said the following:

> *... Just putting in home medications alone takes up so much time. This new telenurse service helps* [save time]

Several nurses highlighted that admission and discharge processes are so complex and time-consuming that shifting this work to the ACTN program freed nurses to perform other activities, as reflected by this nurse:

> *The tele RN is able to spend as much time possible sufficiently educating an admission or discharge while allowing me time to respond to the needs of my other patients saving me time on one patient especially charting.*

Emotional and Physical Support

For the second theme, several responses focused less on time management and perceived efficiencies and instead centered more on the emotional appeal and support in having an extra hand, as one nurse mentioned:

> *Being in such a fast-paced unit, it can be a bit stressful with so many discharges and admissions. Having a helpful hand is beneficial.*

Improved Patient Safety

Finally, the third theme was perceived improvement in patient safety by having a telenurse who could "catch missed" issues (eg, an incorrectly identified pharmacy details), simultaneously allowing the primary bedside nurse to focus more intensely on other needs, essentially creating a 2-fold safety promotion. Some nurses noted that they could begin carrying out orders while the telenurses began completing the admission, facilitating quicker treatment and resolution of care needs, thereby improving the safety and quality of care. One nurse mentioned the following:

> *Allows [telenurses] to take on thorough and accurate admissions, while also preventing any rushing the patient might experience from the primary RN.*

When asked for areas of improvement, the most recurring theme was having 24 hours of support during the weekend and during the week. The second theme for improvement was the reduced time to connect to a telenurse. The third theme was the availability of iPads. Nurses mentioned that iPads could sometimes be unavailable in patients' rooms or they may not be fully charged.

Time of Discharge

The time of day distribution is presented in Figure 2. The only noticeable difference between the telenursing and nontelenursing cohorts was a shift in the volume of patients discharged before 2 PM compared with those discharged after 2 PM at a hospital-wide level (Table 1). At an individual unit level, these results were not consistent and could be further explored by patient population and their needs to discharge. The variation was further illustrated when reviewing the length of stay of patients in the telenursing and nontelenursing cohorts. Only 5 out of the 12 units showed a decrease in the average inpatient length of stay.

XSL•FO
RenderX

Figure 2. Comparison of percentage distribution of discharge by time of the day.

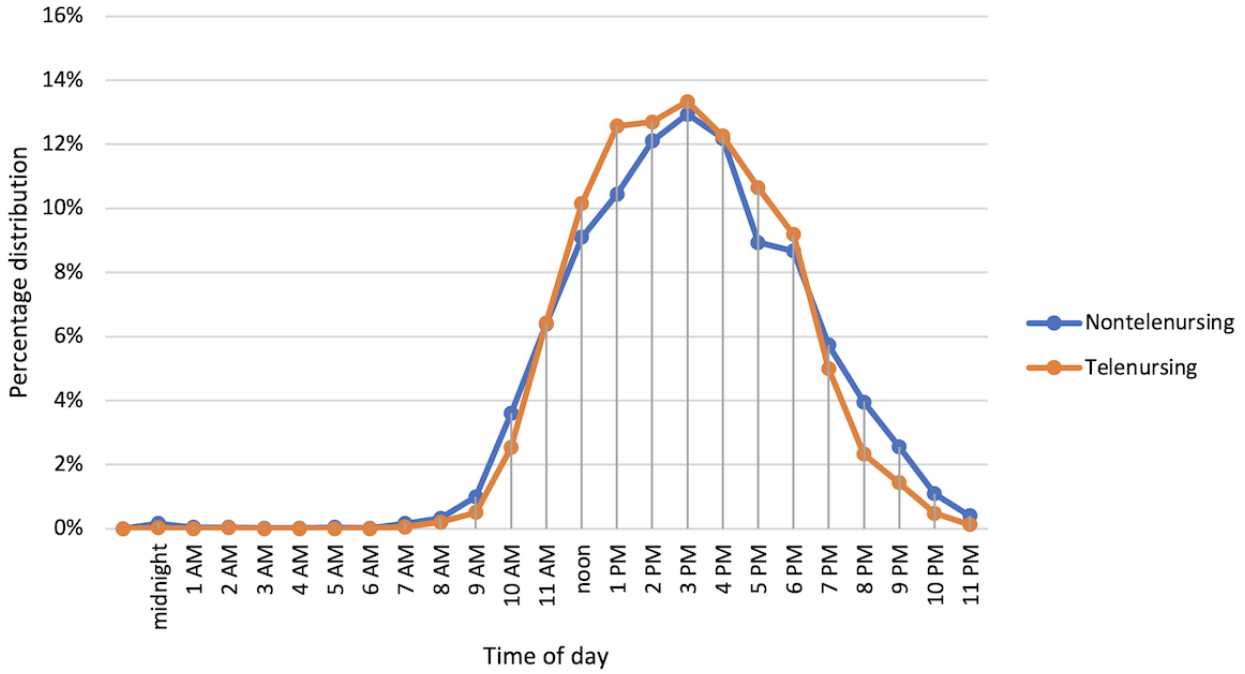

Table 1. Time of the day distributions for the nontelenursing cohort compared to the telenursing cohort.

Hour of day	Nontelenursing (n=4220), n (%)	Telenursing (n=3907), n (%)
Before and up to 2 PM	1837 (43.53)	1766 (45.2)
After 2 PM	2383 (56.47)	2141 (54.8)

Discharge Length

The ACTN admissions averaged 12 minutes and 6 seconds (SD 7 min and 29 s), and the discharges averaged 14 minutes and 51 seconds (SD 8 min and 10 s). The average duration for ACTN calls was 13 minutes and 17 seconds (SD 7 min and 52 s). Traditional cohort standard practice of a bedside nurse engaging in discharge and admission processes was 45 minutes, consistent with our preimplementation nursing time study.

Discussion

Principal Findings

Our results suggest that the ACTN program was associated with positive nursing experiences because it saved time. Furthermore, the ACTN program was associated with higher HCAHPS scores in several domains but only in the first series of units that piloted the intervention. In phase 1, the improvement in “communication with doctors” and “would recommend hospital” scores in 1 unit was statistically significant. In phase 2, the improvement in “communication with doctors” score was significant in 2 units and that in “overall rating of hospital” and “would recommend hospital” scores were significant in 1 unit. The time of day discharge was nearly the same in both the telenursing and nontelenursing cohorts. The duration for discharge processes was less than half in the ACTN cohort compared to the nonintervention cohort.

At the time of writing this paper, the United States was experiencing a critical nursing shortage that will likely reach an epidemic level in the next few decades [8]. Despite the promise shown by telenursing, to our knowledge, only 1 existing paper documents the impact of ACTN programs on HCAHPS-measured patient satisfaction using a small cohort of patients in a single, time-limited pre- and posttelenursing analysis [18]. A study by Schuelke et al [18] revealed a 6.2% increase in “communication with meds” and 12.7% increase in “communication with nursing” domain scores; other HCAHPS domains were not evaluated. This research builds upon the promising work of Schuelke et al [18], evaluating the impact of an ACTN program on several units with a much larger cohort of patients using a staggered rollout and comparing all HCAHPS domains between telenursing and nontelenursing cohorts within the same time frame and in the same units.

By conducting granular HCAHPS analyses, we identified what we believed to be a time sequence variability in that units that rolled out in phase 1 performed considerably stronger in HCAHPS impacts than units that rolled out in later phases. An explanation for this sequence effect might be that some later adopters had less potential for high effect size, given that the first 4 units of the rollout were specifically chosen for their staffing problems compared to later units. ACTN support might have augmented the staffing support to such a degree that allowed the impacts of the program to be more salient. An alternative explanation is that the early adopters and promoters tend to have greater diffusion uptake, greater saturation and

XSL•FO
RenderX

adoptability, and greater impacts compared to late adopters or those resistant to adoption [19,20]. Our anecdotal evidence suggests that early adopters might have *wanted* the telenursing program to succeed; therefore, they applied consistent implementation practices to ensure success. Adopters in later stages were more aware of barriers and potential downsides and might have been more ambivalent about telenursing and, therefore, less likely to modify their behaviors to promote the telenursing program's success.

Another interesting finding was that the ACTN program seemed to be effective for both medical and surgical units of all specialties. Phase 1 was a mix of medical and surgical units; however, all units experienced increases in scores. Phases 2 and 3 experienced mixed results, without a clear lead for one specialty over the other. This may suggest that ACTN programs are broadly applicable across acute settings and that success depends most crucially on the need and desire of unit leaders.

Our time of day discharge findings showed only a few quantitative positive efficiencies. However, our discharge duration analysis and nursing experience survey results showed that ACTN has major time-saving benefits for nurses, suggesting a discrepancy between perceived and actual time savings versus time-of-day discharge savings. One explanation for this discrepancy may be that many factors beyond nursing impact the time of the day a patient is discharged; therefore, while the bedside nurses' time is saved, the remaining discharge processes beyond nurses remain unaffected. Specifically, there are 3 segments of time during discharge processes: (1) the time for the discharge order and medication reconciliation [21] to the time the after-visit summary (AVS) is populated and printed [22]; (2) the time the AVS is completed and printed to the time the discharge instructions are provided; and (3) the time from providing the discharge instructions to the actual discharge (Figure 3). Notably, telenurses' involvement is currently limited to only the second segment of time. Specifically, telenurses' involvement is not initiated until the AVS is printed by the nurse, which means that telenurses cannot positively impact any discharge activity that occurs between the time the discharge order is written and the time the AVS is printed. However, there are inefficiencies and bottlenecks in discharge processes that occur well before the AVS is printed [23,24]. For instance, the discharging physician may write a conditional discharge order early in the morning, listing conditions that cannot be fulfilled within a few hours or it may take bedside nursing longer than anticipated time to print the AVS.

Figure 3. Overview of the discharge process in our health system. AVS: after-visit summary.

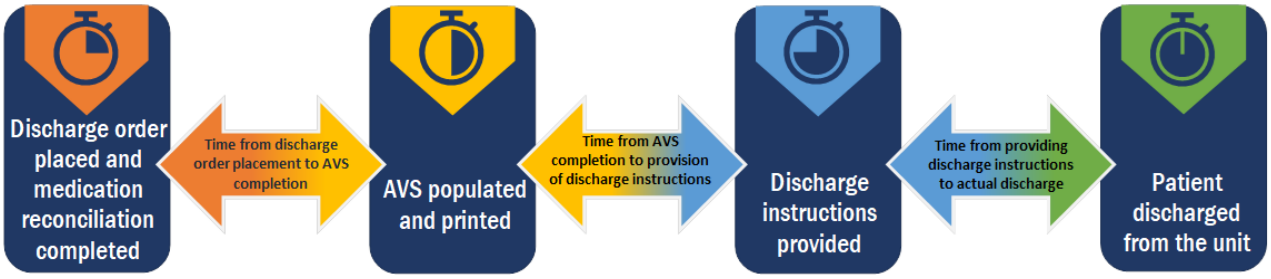

To create a wider cascade effect for positively impacting the discharge processes for all segments of time, we are currently trying to obtain greater transparency through EHR reporting in what occurs for segments 1 and 3. For instance, at present, we know that at least 2 hospitals in our 8-hospital system have high incidence rates of conditional discharge orders that should be reduced. One hospital anecdotally reports that the discharging physician identifies incorrect pharmacies, which requires a nurse to send the scripts back to the discharging pharmacist to reconcile before discharge education can occur [25]; however, the prevalence and location of these issues remain speculative. Segment 3 is a black box of time [26]—the time it takes for hospital transport or an ambulance to arrive and move the patient to their destination and the time it takes for the family to pick up the patient. All these factors impact the discharge processes and need to be fully elucidated, explored, and streamlined. Furthermore, we hope to facilitate processes that enable telenurses to print the AVS, to remove the dependency on bedside nurses to begin the discharge education process.

Limitations

This study has several noteworthy limitations. First, the study was conducted in 1 health system and the results may not be generalizable to other settings with different patient populations, processes, and implementation strategies [27]. Second, in this study, we did not control for other factors that could impact patient and provider satisfaction as well as discharge times; telenursing can only improve upon one component in a complex set of factors limiting discharge efficiency and satisfaction outcomes. Finally, participating nurses were aware of the ongoing study, and this knowledge might have affected their behavior [28].

Future Directions

After the completion of this pilot study, the ACTN admission and discharge program has been rolled out to pilot medical units and all surgical and observation units. Our rationale for expansion rested on the premise that nursing experience is important to maintain and strengthen, particularly at a time when turnover is high in the health care industry in general. It is important to reduce staff inefficiencies in workload as a means of preserving or strengthening organizational morale and cost saving. Because our nursing experience findings for the ACTN program heavily supported the program, this served as the primary motivation for expansion. The nursing experience findings, coupled with the findings related to time-savings in discharge education and modest improvement, though not negative, in the HCAHPS findings for the ACTN program compared to the nontelenursing cohort, further supported expansion.

XSL•FO
RenderX

The initial scope for expansion included a complete system-wide implementation for all admissions and discharges. Furthermore, we are planning to expand the ACTN program beyond admissions and discharges. Responsive to qualitative feedback reported earlier, the next phase of the ACTN program will add safeguards on high-risk medications by having telenurses conduct double-checks, skin assessments, hourly rounding assistance, and auditing of safety functions and educational activities. These activities were chosen because they are time-intensive for nursing staff on the patient floors. Additional support in these areas would be a staff morale booster in addition to improved efficiencies for bedside nursing. Conducting hourly rounding using the ACTN program will require more time and resources; however, conducting high-quality, uninterrupted hourly rounds is known to be effective at improving patient safety and patient experience outcomes [29]. Therefore, we suspect that the ACTN program will have some positive impacts if rounds are consistently conducted, even if conducted virtually.

In addition, the ACTNs have been motivating other specialties to adopt or consider a similar program as the ACTN program to support stretched staffing. These specialties include respiratory care, in which virtual support can quickly identify patients in need of intensive on-site support; pharmacy, in which direct communication with staff on medications and patient training can happen through virtual means; infection control, in which room environments can be reviewed through virtual audits, moving quickly from floor to floor; and guest relations and spiritual care, in which patients can be visited virtually upon patient request. Furthermore, physicians who wish to either virtually enter inpatient rooms during their clinic days or from home can quickly drop in to see patients using the virtual program. For these groups to further develop advanced inpatient telemedicine programs, additional technology will be required, including cameras that can zoom into various portions of the room and advanced sound capabilities. Future work could expand programs similar to ACTN to specialties such as respiratory therapy, pharmacy, infection prevention, and spiritual care.

Conclusions

This study provides preliminary evidence suggesting that telenursing may effectively address nursing shortages in acute care settings and positively impact patient and provider satisfaction as well as admission and discharge times. More work is needed to validate the findings in other settings, use other satisfaction metrics, and investigate the impact of telenursing on the quality of care and cost.

Acknowledgments

The authors would like to thank Jacob M Kolman, MA, ISMPP CMPP, senior scientific writer, Houston Methodist Academic Institute, for the critical review and for providing formatting feedback on this manuscript. The authors would also like to thank Amir Hossein Javid for his help with statistical analysis.

Data Availability

Data sharing is not applicable as no data sets were generated during this study.

Authors' Contributions

All authors were involved in the conceptualization, review and approval, and writing of the manuscript. LR, BJ, MG, RS, SK, and MH were extensively involved in the implementation of the project. BJ, MH, SK, and MG conducted the training. SK and XL conducted the analyses. CRB wrote and edited the manuscript, inserted and refined the citations, and provided critical feedback during implementation and analyses. CRB and FS were involved in all stages of writing and publication. All authors meaningfully contributed to the drafting, writing, brainstorming, executing, finalizing, and approving of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional outcome information for Hospital Consumer Assessment of Health Care Providers and Systems, time of day discharges, and discharge education processes.
[DOCX File , 34 KB - jmir_v26i1e54330_app1.docx]

References

1. Greiwe J. Telemedicine lessons learned during the COVID-19 pandemic. Curr Allergy Asthma Rep 2022 Jan 21;22(1):1-5 [FREE Full text] [doi: 10.1007/s11882-022-01026-1] [Medline: 35061150]
2. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
3. Ekeland AG, Bowes A, Flottorp S. Effectiveness of telemedicine: a systematic review of reviews. Int J Med Inform 2010 Nov;79(11):736-771. [doi: 10.1016/j.ijmedinf.2010.08.006] [Medline: 20884286]

XSL•FO
RenderX

4. Nanda M, Sharma R. A review of patient satisfaction and experience with telemedicine: a virtual solution during and beyond COVID-19 pandemic. Telemed J E Health 2021 Dec 01;27(12):1325-1331. [doi: 10.1089/tmj.2020.0570] [Medline: 33719577]
5. Yu X, Bayram A. Managing capacity for virtual and office appointments in chronic care. Health Care Manag Sci 2021 Dec;24(4):742-767 [FREE Full text] [doi: 10.1007/s10729-021-09546-4] [Medline: 33759065]
6. Laskowski ER, Johnson SE, Shelerud RA, Lee JA, Rabatin AE, Driscoll SW, et al. The telemedicine musculoskeletal examination. Mayo Clin Proc 2020 Aug;95(8):1715-1731 [FREE Full text] [doi: 10.1016/j.mayocp.2020.05.026] [Medline: 32753146]
7. Hughes L, Petrella A, Phillips N, Taylor RM. Virtual care and the impact of COVID-19 on nursing: a single centre evaluation. J Adv Nurs 2022 Feb;78(2):498-509 [FREE Full text] [doi: 10.1111/jan.15050] [Medline: 34590738]
8. Pulse on the nation's nurses survey series: COVID-19 two-year impact assessment survey. American Nurses Foundation. 2022 Mar 1. URL: https://www.nursingworld.org/~4a2260/contentassets/872ebb13c63f44f6b11a1bd0c74907c9/covid-19-two-year-impact-assessment-written-report-final.pdf [accessed 2023-04-14]
9. Abraham C, Jensen C, Rossiter L, Dittman Hale D. Telenursing and remote patient monitoring in cardiovascular health. Telemed J E Health 2023 Sep 08 (forthcoming). [doi: 10.1089/tmj.2023.0187] [Medline: 37682280]
10. Rouleau G, Gagnon M, Côté J, Payne-Gagnon J, Hudson E, Dubois CA. Impact of information and communication technologies on nursing care: results of an overview of systematic reviews. J Med Internet Res 2017 Apr 25;19(4):e122 [FREE Full text] [doi: 10.2196/jmir.6686] [Medline: 28442454]
11. Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. J Am Med Inform Assoc 2021 Apr 23;28(5):998-1008 [FREE Full text] [doi: 10.1093/jamia/ocaa325] [Medline: 33434273]
12. Hehman MC, Fontenot NM, Drake GK, Musgrove RS. Leveraging digital technology in nursing. Health Emerg Disaster Nurs 2023;10:41-45. [doi: 10.24298/hedn.2022-0014]
13. Schwartz RL, Hamlin SK, Vozzella GM, Randle LN, Klahn S, Maris GJ, et al. Utilizing telenursing to supplement acute care nursing in an era of workforce shortages: a feasibility pilot. Comput Inform Nurs 2024 Feb 01;42(2):151-157. [doi: 10.1097/CIN.0000000000001097] [Medline: 38252545]
14. Cleary PD. Evolving concepts of patient-centered care and the assessment of patient care experiences: optimism and opposition. J Health Polit Policy Law 2016 Aug 28;41(4):675-696. [doi: 10.1215/03616878-3620881] [Medline: 27127265]
15. CAHPS hospital survey. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/cahps/surveys-guidance/hospital/index.html [accessed 2023-10-12]
16. Strauss A, Corbin JM. Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Thousand Oaks, CA: Sage Publications; 1990.
17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
18. Schuelke S, Aurit S, Connot N, Denney S. Virtual nursing: the new reality in quality care. Nurs Adm Q 2019;43(4):322-328. [doi: 10.1097/NAQ.0000000000000376] [Medline: 31479052]
19. Rogers E. Diffusion of Innovations. New York, NY: Free Press of Glenco; 1962.
20. Mohammadi MM, Poursaberi R, Salahshoor MR. Evaluating the adoption of evidence-based practice using Rogers's diffusion of innovation theory: a model testing study. Health Promot Perspect 2018;8(1):25-32 [FREE Full text] [doi: 10.15171/hpp.2018.03] [Medline: 29423359]
21. Fernandes BD, Almeida PH, Foppa AA, Sousa CT, Ayres LR, Chemello C. Pharmacist-led medication reconciliation at patient discharge: a scoping review. Res Social Adm Pharm 2020 May;16(5):605-613. [doi: 10.1016/j.sapharm.2019.08.001] [Medline: 31395445]
22. Federman A, Sarzynski E, Brach C, Francaviglia P, Jacques J, Jandorf L, et al. Challenges optimizing the after visit summary. Int J Med Inform 2018 Dec;120:14-19 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.09.009] [Medline: 30409339]
23. Bristol AA, Elmore CE, Weiss ME, Barry LA, Iacob E, Johnson EP, et al. Mixed-methods study examining family carers' perceptions of the relationship between intrahospital transitions and patient readiness for discharge. BMJ Qual Saf 2023 Aug;32(8):447-456 [FREE Full text] [doi: 10.1136/bmjqs-2022-015120] [Medline: 36100445]
24. Piniella NR, Fuller TE, Smith L, Salmasian H, Yoon CS, Lipsitz SR, et al. Early expected discharge date accuracy during hospitalization: a multivariable analysis. J Med Syst 2023 May 12;47(1):63 [FREE Full text] [doi: 10.1007/s10916-023-01952-1] [Medline: 37171484]
25. Schnipper JL, Reyes Nieva H, Yoon C, Mallouk M, Mixon AS, Rennke S, et al. What works in medication reconciliation: an on-treatment and site analysis of the MARQUIS2 study. BMJ Qual Saf 2023 Aug;32(8):457-469. [doi: 10.1136/bmjqs-2022-014806] [Medline: 36948542]
26. Abu HO, Anatchkova MD, Erskine NA, Lewis J, McManus DD, Kiefe CI, et al. Are we "missing the big picture" in transitions of care? Perspectives of healthcare providers managing patients with unplanned hospitalization. Appl Nurs Res 2018 Dec;44:60-66 [FREE Full text] [doi: 10.1016/j.apnr.2018.09.006] [Medline: 30389062]
27. Leung L. Validity, reliability, and generalizability in qualitative research. J Family Med Prim Care 2015;4(3):324-327 [FREE Full text] [doi: 10.4103/2249-4863.161306] [Medline: 26288766]

XSL•FO
RenderX

28. Merrett F. Reflections on the Hawthorne effect. Educ Psychol 2006 Jan;26(1):143-146. [doi: 10.1080/01443410500341080]
29. Mitchell M, Lavenberg JG, Trotta RL, Umscheid CA. Hourly rounding to improve nursing responsiveness: a systematic review. J Nurs Adm 2014 Sep;44(9):462-472 [FREE Full text] [doi: 10.1097/NNA.0000000000000101] [Medline: 25148400]

Abbreviations

ACTN: acute care telenurse
AVS: after-visit summary
CMI: case mix index
EHR: electronic health record
HCAHPS: Hospital Consumer Assessment of Health Care Providers and Systems

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 06.11.23; peer-reviewed by C Jensen; comments to author 08.12.23; revised version received 16.01.24; accepted 17.02.24; published 04.04.24.

Please cite as:
Bruce CR, Klahn S, Randle L, Li X, Sayali K, Johnson B, Gomez M, Howard M, Schwartz R, Sasangohar F
Impacts of an Acute Care Telenursing Program on Discharge, Patient Experience, and Nursing Experience: Retrospective Cohort Comparison Study
J Med Internet Res 2024;26:e54330
URL: https://www.jmir.org/2024/1/e54330
doi:10.2196/54330
PMID:38573753

©Courtenay R Bruce, Steve Klahn, Lindsay Randle, Xin Li, Kelkar Sayali, Barbara Johnson, Melissa Gomez, Meagan Howard, Roberta Schwartz, Farzan Sasangohar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Monitoring of Clinician-Patient Video Visits During the Peak of the COVID-19 Pandemic: Adoption and Sustained Challenges in an Integrated Health Care Delivery System

Jessica A Palakshappa[1,2], MD, MS; Erica R Hale[1,2], MS; Joshua D Brown[1], PhD; Carol A Kittel[2], MA; Emily Dressler[2], PhD; Gary E Rosenthal[1,2], MD; Sarah L Cutrona[3,4], MD, MPH; Kristie L Foley[2], MS, PhD; Emily R Haines[2], PhD; Thomas K Houston II[1,2], MD, MPH

[1]Atrium Health Wake Forest Baptist, Winston Salem, NC, United States

[2]Wake Forest University School of Medicine, Winston Salem, NC, United States

[3]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[4]Center for Healthcare Organization and Implementation Research, Veterans Affairs Bedford Healthcare System, Bedford, MA, United States

Corresponding Author:
Jessica A Palakshappa, MD, MS
Wake Forest University School of Medicine
1 Medical Center Blvd
Winston Salem, NC, 27157
United States
Phone: 1 336 716 8465
Email: jpalaksh@wakehealth.edu

Abstract

Background: Numerous prior opinion papers, administrative electronic health record data studies, and cross-sectional surveys of telehealth during the pandemic have been published, but none have combined assessments of video visit success monitoring with longitudinal assessments of perceived challenges to the rapid adoption of video visits during the pandemic.

Objective: This study aims to quantify (1) the use of video visits (compared with in-person and telephone visits) over time during the pandemic, (2) video visit successful connection rates, and (3) changes in perceived video visit challenges.

Methods: A web-based survey was developed for the dual purpose of monitoring and improving video visit implementation in our health care system during the COVID-19 pandemic. The survey included questions regarding rates of in-person, telephone, and video visits for clinician-patient encounters; the rate of successful connection for video visits; and perceived challenges to video visits (eg, software, hardware, bandwidth, and technology literacy). The survey was distributed via email to physicians, advanced practice professionals, and clinicians in May 2020. The survey was repeated in March 2021. Differences between the 2020 and 2021 responses were adjusted for within-respondent correlation across surveys and tested using generalized estimating equations.

Results: A total of 1126 surveys were completed (511 surveys in 2020 and 615 surveys in 2021). In 2020, only 21.7% (73/336) of clinicians reported no difficulty connecting with patients during video visits and 28.6% (93/325) of clinicians reported no difficulty in 2021. The distribution of the percentage of successfully connected video visits ("Over the past two weeks of scheduled visits, what percentage did you successfully connect with patients by video?") was not significantly different between 2020 and 2021 (P=.74). Challenges in conducting video visits persisted over time. Poor connectivity was the most common challenge reported by clinicians. This response increased over time, with 30.5% (156/511) selecting it as a challenge in 2020 and 37.1% (228/615) in 2021 (P=.01). Patients not having access to their electronic health record portals was also a commonly reported challenge (109/511, 21.3% in 2020 and 137/615, 22.3% in 2021, P=.73).

Conclusions: During the pandemic, our health care delivery system rapidly adopted synchronous patient-clinician communication using video visits. As experience with video visits increased, the reported failure rate did not significantly decline, and clinicians continued to report challenges related to general network connectivity and patient access to technology.

(J Med Internet Res 2024;26:e54008) doi:10.2196/54008

XSL•FO
RenderX

KEYWORDS

telehealth; telemedicine; e-health; eHealth; video visits; video; ICT; information and communication technology; survey; surveys; adoption; usability; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; COVID-19

Introduction

Interest in telehealth from policy makers, health care providers, patients, and families continues to grow [1], including newer modalities such as video visits [2]. Video visits refer to clinician-patient communication that includes real time video and audio assessment of the patient when the clinician is in a different location. Video visits have the potential to improve efficiency for clinicians and to improve access for patients, particularly those who reside in rural areas or with transportation barriers [3]. The technology to support video visits has existed for decades; however, only a small minority of clinicians used this form of telehealth in their practices [4] prior to the COVID-19 pandemic. There are multiple reasons why telehealth adoption was slow despite its potential benefits including changing cost and reimbursement policies, federal and state licensing laws, incompatible electronic health records, and gaps in internet access in certain areas.

The COVID-19 pandemic and subsequent public health emergency led to fundamental shifts in how health care was delivered in the United States, including the rapid adoption of telehealth services. Before the public health emergency, approximately 13,000 fee-for-service Medicare beneficiaries received telehealth services in a week and that number increased to nearly 1.7 million beneficiaries by the last week of April 2020 [5]. While the need to avoid in-person contact fueled the initial rapid rise, regulations and restrictions were temporarily lifted during this time facilitating its use. Clinicians were also paid for telehealth services at the same rate as in-person medical services. Several studies have reported on the rapid uptake of telehealth, including video visits, in this context [6-8]. However, few reports have explored rates of success and failure of video visits over time. The challenges clinicians face in conducting video visits have also not been explored. Understanding these challenges will be important for improving and expanding the reach of telehealth services after the pandemic has ended.

In the context of the rapidly increasing use of telehealth to conduct video visits, and consistent with the sociotechnical model's [9] emphasis on monitoring the implementation of health information technology in complex adaptive health care systems, our health care delivery system initiated a series of brief assessments of video visit adoption. The research objective of this report is to summarize the findings of the video visit monitoring including (1) the use of video visits (compared with in-person and telephone visits) over time during the pandemic, (2) video visit successful connection rates, and (3) changes in perceived video visit challenges. With patient and clinician skills and experience with video visits increasing over time, our primary hypothesis was that clinicians' perceived challenges to completing video visits (eg, software, hardware, bandwidth, and technology literacy) would decline over time.

Methods

Study Design

The design was a longitudinal series of 2 cross-sectional assessments (2020 and 2021). In summary, for the dual purpose of monitoring and improving telehealth implementation in our health care delivery system, institutional leaders developed a brief web-based survey regarding the use of video visits and challenges. The survey was initially distributed in 2020. Given the ongoing public health emergency and the need to re-evaluate telehealth use, the survey was repeated in 2021. Institutional leaders encouraged clinicians to complete the survey, communicating encouragement via emails and announcements at in-person faculty and departmental meetings.

Ethical Considerations

As the brief assessments were distributed by the clinical system as part of ongoing quality improvement, the project was approved as an research protocol as exempt from human participants approval by the Wake Forest University School of Medicine Institutional Review Board (IRB00077473). The survey did not collect identifying information from the participants.

Survey Development

Published in 2010 by Sittig and Singh [9], the sociotechnical model of health information technology was the first to fully emphasize the importance of system monitoring in implementation frameworks. Key aspects of monitoring, including measuring how the technology is being used by clinicians and whether implementation outcomes are being achieved, were considered when developing the survey. The survey was developed with a literature search, expert review, and iterative pilot-testing (see Multimedia Appendix 1). The final survey included 12 questions related to rates of in-person, telephone, and video for clinician-patient encounters; the rate of successful connection for video visits; and perceived challenges to video visits (eg, software, hardware, bandwidth, and technology literacy).

Study Population

The study population included all outpatient clinicians practicing across the health care delivery system; we excluded clinicians without direct patient care responsibilities. The system includes 5 hospitals and over 350 primary care and specialty clinics that provide care to over 2 million persons annually. The brief assessment was distributed to clinicians in 2020 and 2021 (1937 clinicians and 2843 clinicians, respectively).

Survey Distribution and Data Collection

As we are an integrated health care delivery system, we had access to the contact details of all providers. Our group practice clinical operations executive committee facilitated the survey distribution by requesting that each department chair and clinical

XSL•FO
RenderX

service line director send an email to their team of providers to notify them of the survey and encourage completion. Surveys were collected and managed using REDCap (Research Electronic Data Capture), a secure, web-based app designed to support data capture [10,11]. A unique survey link was distributed via email to each clinician in May 2020 and March 2021. The system sent up to 2 reminder emails if the recipient had not yet completed the survey.

Statistical Analysis

To take full advantage of the data collected, we first analyzed the results as 2 cross-sectional surveys. In this primary analysis, we included all respondents in each year. We recognize that a subset of clinicians also responded in both years. Thus, as a secondary analysis, we analyzed the data limited to the longitudinal cohort who participated in both years. First, summary statistics are presented as count (frequency) for categorical variables and mean (SD) or median (IQR) for continuous variables as appropriate. Generalized estimating equations were then used to model frequency distributions of in-person, telephone, and video visits, and patient and clinician challenges. These logit models were adjusted for within-respondent correlation across surveys via an exchangeable correlation structure. *P* values of .05 were considered statistically significant. *P* values for multiple comparisons in frequency distributions of in-person, telephone, and video visits between physicians, advanced practice professionals (APPs), and other clinicians were adjusted via the Tukey-Kramer method to control for type I errors with a corrected *P* value <.05 deemed statistically significant [12]. All statistical analyses were performed with R (version 4.2.1; R Core Team) [13].

We recognize that a subset of clinicians responded in both years. Thus, as a secondary analysis, we analyzed the data limited to the longitudinal cohort who participated in both surveys. For the secondary analysis, matched pairs analyses were performed as were performed in the entire sample with only those responses from clinicians that completed both surveys.

Results

Surveillance Participation and Participant Characteristics

In 2020, 1937 surveys were sent and 511 responses were received (response rate 26.4%). In 2021, 2843 surveys were sent and 615 responses were received (response rate 21.6%). In both years, over 55% of the respondents were physicians from a wide range of clinical specialties. About half of the clinicians who completed the survey in 2020 also completed it in 2021 (Table 1).

Table 1. Characteristics of health care clinicians.

	Response rate in 2020, n (%)	Response rate in 2021, n (%)
Clinician type[a]		
Physician	300 (58.7)	353 (57.4)
Physician assistant or nurse practitioner	195 (38.2)	216 (35.1)
Other allied health professionals	16 (3.1)	46 (7.5)
Physician specialties[b]		
Internal medical subspecialty	84 (28)	68 (19.3)
General internal medicine and geriatrics	46 (15.3)	43 (12.2)
Family medicine	34 (11.3)	42 (11.9)
Pediatrics (general)	27 (9)	45 (12.7)
Surgery subspecialties	26 (8.7)	34 (9.6)
Pain medicine, anesthesia, and rehabilitation	25 (8.3)	18 (5.1)
Pediatric subspecialties	20 (6.7)	16 (4.5)
Neurology	19 (6.3)	14 (4)
Psychiatry and psychology	9 (3)	42 (11.9)
Obstetrics and gynecology	7 (2.3)	22 (6.2)
Others	3 (1)	9 (2.5)

[a]n=511 responses in 2020 and n=615 responses in 2021.

[b]n=300 responses in 2020 and n=353 responses in 2021.

Health Care Delivery by In-Person and Telephone

To place the volume of telehealth in context, we first asked about the number of in-person encounters completed over the past 2 weeks (Table 2). The distribution of responses differed between 2020 and 2021 (*P*<.001). Modeled probabilities show fewer respondents reported zero (22.2% vs 4.9%) or 1 to 10 (33.5% vs 11.5%) in-person visits in 2021 as compared with 2020. The volume of in-person visits increased over time (Table 2). Further, the majority of respondents (399/509, 78.4%)

XSL•FO
RenderX

reported at least 1 telephone visit in 2020 and 65.7% (369/562) in 2021 although the distribution of responses differed from 2020 to 2021 (*P*<.001).

Table 2. Health care delivery (through in-person visits, telephone visits, and video visits) reported by providers in 2020 and 2021.

Over the past 2 weeks, thinking of your outpatient clinics, please approximate how many of your patient encounters were scheduled as follows	Health care delivery in 2020 (N=511), n (%)	Health care delivery in 2021 (N=615), n (%)	*P* value[a]
In-person visits			<.001
0	113 (22.2)	28 (4.9)	
1 to 10	170 (33.5)	65 (11.5)	
11 to 25	105 (20.7)	132 (23.3)	
26 to 50	69 (13.6)	139 (24.6)	
Over 50	51 (10)	202 (35.7)	
Telephone visits			<.001
0	110 (21.6)	193 (34.3)	
1 to 10	230 (45.2)	316 (56.2)	
11 to 25	110 (21.6)	44 (8)	
26 to 50	37 (7.3)	6 (1.1)	
Over 50	22 (4.3)	2 (0.4)	
Video visits			<.001
0	174 (34.1)	239 (42.4)	
1 to 10	199 (39)	260 (46.1)	
11 to 25	95 (18.6)	38 (6.7)	
26 to 50	30 (5.9)	16 (2.8)	
Over 50	12 (2.4)	11 (2)	

[a]Differences in frequency distributions between 2020 and 2021 tested via generalized estimating equation modeling; *P* value adjusted using Tukey-Kramer method to control for type I errors.

Health Care Delivery by Video Visits

Many health care providers were engaged in virtual care, with 65.9% (336/510) health care providers reporting video visit encounters in 2020 and 57.6% (325/564) health care providers reporting video visit encounters in 2021 (Table 2) although the distribution of responses again changed from 2020 to 2021 (*P*<.001). Compared with 2020, fewer 2021 respondents reported 11-25 (18.6% vs 6.7%), 26-50 (5.9% vs 2.8%), or over 50 (2.4% vs 2%) visits.

Secondary analyses were robust to missing data and showed that the results (distributions of in-person, phone, and video visits) did not change when limiting the data to only respondents who participated in both surveys.

Comparing Health Care Delivery by Physicians, APPs, and Others

We also compared health care delivery modality by type of clinician (physicians, APPs, or others). Patterns of health care delivery reported in the overall sample were similar in the physician, APP, and other subgroups. There were no significant differences between physicians and APPs in the number of patient encounters that were completed as in-person, telephone visits, or video visits in 2020 or 2021.

Perceived Challenges to Patient-Clinician Connection Using Video Visits

The use of video visits came with challenges. In 2020, only 21.7% (73/336) of clinicians reported no difficulty connecting with patients during video visits and 28.6% (93/325) of clinicians reported no difficulty in 2021 (Figure 1). The distribution of the percentage of successfully connected video visits ("Over the past two weeks of scheduled visits, what percentage did you successfully connect with patients by video?") was not significantly different between 2020 and 2021 (*P*=.74, Figure 1). There was also no significant difference between physicians and APPs in the rate of successful video connection with patients in either year.

XSL•FO
RenderX

Figure 1. Percentage of successful provider and patient connection in the scheduled video visits, by comparing 2020 and 2021 results. Survey Question: "Over the past two weeks, of the scheduled video visits, what percentage did you successfully connect with the patients by video?" Comparing 2020 versus 2021, no significant difference (*P*=.74); the *P* value was adjusted using the Tukey-Kramer method to control for type I errors.

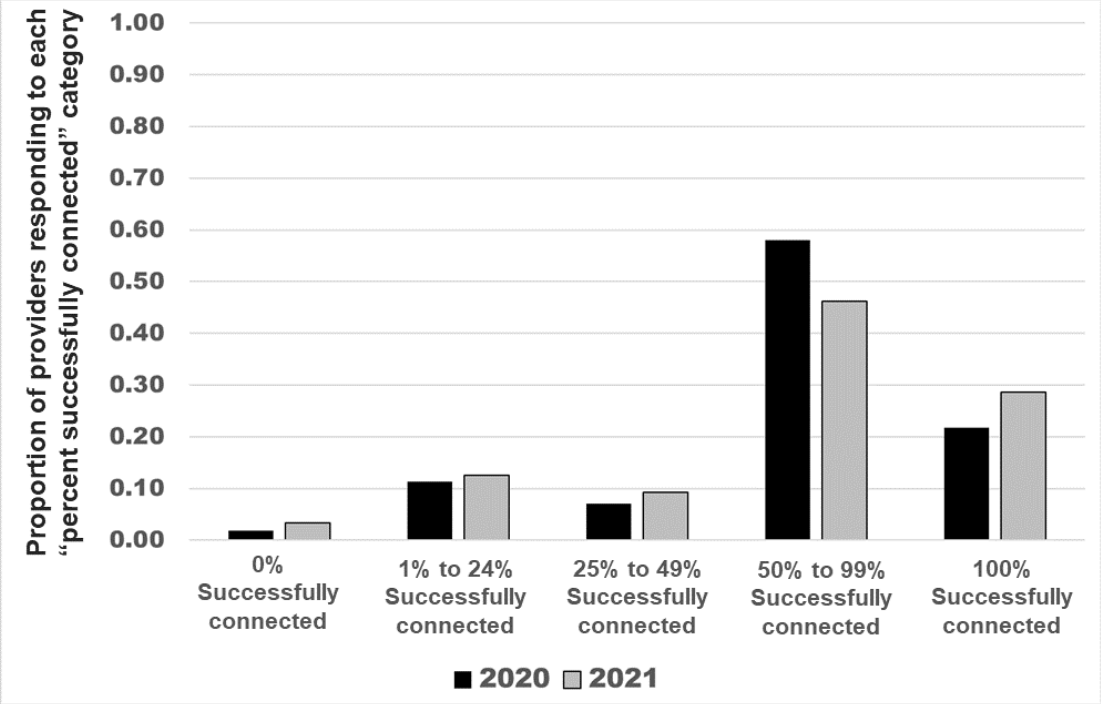

Clinicians were asked about the challenges in successfully completing video visits (see Table 3). The most commonly reported challenge was poor connectivity. This response increased over time with 30.5% (156/511) selecting it as a challenge in 2020 and 37.1% (228/615) selecting it as a challenge in 2021 (*P*=.01). Patients not having access to their electronic health record portals was also a commonly reported challenge (109/511, 21.3% in 2020 and 137/615, 22.3% in 2021; *P*=.73).

Table 3. Clinician-reported challenges encountered on scheduled clinical video telehealth during the COVID-19 pandemic.

If you had challenges with successfully completing video visits with patients, what were the reasons for the challenges (check all that apply)?	Challenges encountered in 2020 (N=511), n (%)	Challenges encountered in 2021 (N=615), n (%)	*P* value[a]
Patient challenges			
The patient did not have access to a smartphone or computer with camera.	135 (26.4)	179 (29.1)	.28
The patient did not have patient portal access.[b]	109 (21.3)	137 (22.3)	.73
The patient did not have access to the internet.	73 (14.3)	112 (18.2)	.08
The patient did not have the needed software.	69 (13.5)	74 (12)	.45
Clinician challenges			
There was a problem with network connectivity (poor connectivity).	156 (30.5)	228 (37.1)	.008
There was a problem with the clinician software.	54 (10.6)	55 (8.9)	.34
There was a problem with clinician hardware.	28 (5.5)	30 (4.9)	.63

[a]*P* value adjusted using the Tukey-Kramer method to control for type I errors.

[b]For some clinical video visits, a prerequisite was that patients needed to have registered with the patient portal.

Discussion

Principal Findings

Although telehealth technology was available in our health care delivery system prior to the COVID-19 pandemic, it saw only limited use for providing synchronous care to patients prior to the pandemic. Inconsistent reimbursement for services, restrictions on the physical location of patients and clinicians during telehealth, and rules about types of visits that were acceptable for telehealth services all contributed to its limited use [14,15]. With the pandemic, and consistent with other reports, our health care delivery system rapidly expanded the provision of clinical care by way of video visits. Later in the pandemic, in-person visits did increase, but the use of video visits remained well above prepandemic levels.

XSL•FO
RenderX

Overall, there has been a shift toward a more positive sentiment about telehealth and telemedicine since the start of the pandemic. A scoping review by Doraiswamy et al [16] reported 543 telehealth-related papers (mostly opinions, commentaries, and perspectives; 61%) published across 331 different journals from January to June 2020. Most of these new reports had a "celebratory" or favorable sentiment about the use of telehealth. The scope of the increase in telehealth during the public health emergency likely contributed to this sentiment though concerns about patient and clinician connection, the lack of physical examinations, and cost-effectiveness were still noted by some. Although our providers reported benefits for clinical video telehealth beyond audio-only calls for patient-provider visits, failure to connect using clinical video visits was common.

While reports have documented challenges with clinical video telehealth [17], few have monitored these challenges over time. During the pandemic, as our health care delivery system's experience with video visits grew, clinicians did not report a meaningful reduction in connection failure rate. The most frequently reported challenges were general network connectivity and those related to the digital divide (eg, patient lack of internet access, needed software, or internet-connected cameras). Gaps in access to technology and the internet for telehealth may impact some patient groups more than others. For example, older age, rural residence, dual Medicare and Medicaid enrollment, and non-Hispanic Black or Hispanic race or ethnicity have been shown to be associated with a lower probability of technology ownership, access to the internet, and use of the internet for communication in cancer survivors [18]. Further, over 10% of clinicians also reported that they experienced software or hardware challenges (eg, limited availability of internet cameras at a clinical location). Expanding telehealth will require ongoing investments in technology for clinicians. New workflows to support successful connection during video visits and follow-up processes may also be needed.

Our video visit monitoring results were shared with clinical operations leadership. In response to the sustained challenges noted, we initiated a new video visit program to provide patient support prior to scheduled video visits. Our technology navigators are a specially trained, centralized team and are directed to reach out to vulnerable patients and families to facilitate video visit access. A new electronic health record dashboard identified patients with (1) a scheduled video visit and (2) 1 or more risk factors (eg, lack of a prior successful video visit and lack of patient portal access). We further prioritized patients older than 65 years and those living in rural areas. Technology navigators reached out by telephone to contact these at-risk patients to assess their technology access (eg, internet, software, webcam, or smartphone), technology literacy and perceived competence, and availability of at-home support from family and friends. The technology navigators then troubleshoot any challenges noted by the patients and offer to conduct a "practice" video visit. Evaluation of this program is ongoing. In 2022-2023, the navigators contacted 1266 patients at high risk for video visit failure. Among those contacted, 515 requested and were provided assistance. With previsit support from the navigators, the patient-provider scheduled video visit completion rate was 84% as compared with a 60% completion rate among those patients who did not receive support.

Limitations of our video visit surveillance analysis include that the data were collected across 1 health care delivery system with an integrated electronic health care record system. The perceptions and challenges may be different in a smaller health care system and in those with different health care record systems. The survey measured only clinician-reported telehealth use and success rates which may be limited by recall. As with all surveys, our results may be biased as only about one-quarter of the sample responded. It is possible that respondents experienced more challenges conducting video visits than those who did not respond. Further, not all clinicians longitudinally completed both the 2020 and 2021 surveys—due both to response rates and providers leaving and entering the health care system. Thus, changes over time may represent differences in the underlying sample. For example, new clinicians may have been more or less familiar with conducting video visits.

Conclusions

Recent reviews have noted the need for more evidence related to telehealth's implementation, effectiveness, and health equity in telehealth access [16,19,20]. Although internet and smartphone access has increased over the last decade (with older adults being one of the fastest-growing subgroups of new adoption), our longitudinal video visit surveillance reveals that the digital divide is still a significant barrier to video visit access.

Although Healthy People 2030 (a set of national objectives to improve health and well-being) includes developmental and research objectives related to patient portals and increasing the use of telehealth to improve access to health services [21], some social determinants of health taxonomies do not include technology access. If telehealth is increasingly an important component of health care access, then technology access (eg, internet, smartphone, patient portal, and connected hardware, such as internet-connected video) should be considered a social determinant of health [22]. A comprehensive solution to overcoming the digital divide has not yet been achieved. However, some partial solutions include directly providing technology to patients, providing detailed instructions, and support services (eg, our technology navigator program), and engaging trusted caregivers (family and friends) who may be able to assist patients [23-27].

Acknowledgments

The project was supported through the National Cancer Institute Cancer Moonshot initiative—iDAPT: Implementation and Informatics–Developing Adaptable Processes and Technologies for Cancer Control (P50 CA244693); the Wake Forest Clinical and Translational Science Award (5UL1TR001420); and also a career development award to JAP (1K23AG073529).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Brief Provider Telehealth Assessment.
[PDF File (Adobe PDF File), 40 KB - jmir_v26i1e54008_app1.pdf]

References

1. World Health Organization. Telemedicine: Opportunities and Developments in Member States: Report on the Second Global Survey on eHealth 2009. Geneva, Switzerland: World Health Organization; 2010.
2. Moulaei K, Shanbehzadeh M, Bahaadinbeigy K, Kazemi-Arpanahi H. Survey of the patients' perspectives and preferences in adopting telepharmacy versus in-person visits to the pharmacy: a feasibility study during the COVID-19 pandemic. BMC Med Inform Decis Mak 2022;22(1):99 [FREE Full text] [doi: 10.1186/s12911-022-01834-5] [Medline: 35418072]
3. Nesbitt TS. The evolution of telehealth: where have we been and where are we going? In: Lustig TA, editor. The Role of Telehealth in an Evolving Health Care Environment: Workshop Summary. Washington, DC: National Academies Press; 2012:11-16.
4. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018;37(12):1923-1930 [FREE Full text] [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
5. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
6. Weiner JP, Bandeian S, Hatef E, Lans D, Liu A, Lemke KW. In-person and telehealth ambulatory contacts and costs in a large US insured cohort before and during the COVID-19 pandemic. JAMA Netw Open 2021;4(3):e212618 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2618] [Medline: 33755167]
7. Gilson SF, Umscheid CA, Laiteerapong N, Ossey G, Nunes KJ, Shah SD. Growth of ambulatory virtual visits and differential use by patient sociodemographics at one urban academic medical center during the COVID-19 pandemic: retrospective analysis. JMIR Med Inform 2020;8(12):e24544 [FREE Full text] [doi: 10.2196/24544] [Medline: 33191247]
8. Hatef E, Wilson RF, Hannum SM, Zhang A, Kharrazi H, Weiner JP, et al. Use of Telehealth During the COVID-19 Era. Rockville, MD: Agency for Healthcare Research and Quality; 2023.
9. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010;19(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
10. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
11. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
12. Midway S, Robertson M, Flinn S, Kaller M. Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. PeerJ 2020;8:e10387 [FREE Full text] [doi: 10.7717/peerj.10387] [Medline: 33335808]
13. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
14. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
15. Hyder MA, Razzak J. Telemedicine in the United States: an introduction for students and residents. J Med Internet Res 2020;22(11):e20839 [FREE Full text] [doi: 10.2196/20839] [Medline: 33215999]
16. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
17. Garfan S, Alamoodi AH, Zaidan BB, Al-Zobbi M, Hamid RA, Alwan JK, et al. Telehealth utilization during the Covid-19 pandemic: a systematic review. Comput Biol Med 2021;138:104878 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104878] [Medline: 34592585]
18. Lama Y, Davidoff AJ, Vanderpool RC, Jensen RE. Telehealth availability and use of related technologies among medicare-enrolled cancer survivors: cross-sectional findings from the onset of the COVID-19 pandemic. J Med Internet Res 2022;24(1):e34616 [FREE Full text] [doi: 10.2196/34616] [Medline: 34978531]
19. Moulaei K, Moulaei R, Bahaadinbeigy K. Barriers and facilitators of using health information technologies by women: a scoping review. BMC Med Inform Decis Mak 2023;23(1):176 [FREE Full text] [doi: 10.1186/s12911-023-02280-7] [Medline: 37670281]
20. Rodriguez JA, Shachar C, Bates DW. Digital inclusion as health care: supporting health care equity with digital-infrastructure initiatives. N Engl J Med 2022;386(12):1101-1103 [FREE Full text] [doi: 10.1056/NEJMp2115646] [Medline: 35302722]

XSL•FO
RenderX

21. Healthy people 2030: health IT. U.S. Department of Health and Human Services. URL: https://health.gov/healthypeople/objectives-and-data/browse-objectives/health-it [accessed 2024-03-14]
22. Benda NC, Veinot TC, Sieck CJ, Ancker JS. Broadband internet access is a social determinant of health!. Am J Public Health 2020;110(8):1123-1125. [doi: 10.2105/AJPH.2020.305784] [Medline: 32639914]
23. Francis J, Kadylak T, Makki TW, Rikard RV, Cotten SR. Catalyst to connection: when technical difficulties lead to social support for older adults. Am Behav Sci 2018;62(9):1167-1185. [doi: 10.1177/0002764218773829]
24. Cotten SR, Yost EA, Berkowsky RW, Winstead V, Anderson WA. Designing Technology Training for Older Adults in Continuing Care Retirement Communities, 1st Edition. Boca Raton, FL: CRC Press; 2016.
25. Sadasivam RS, Kinney RL, Lemon SC, Shimada SL, Allison JJ, Houston TK. Internet health information seeking is a team sport: analysis of the Pew internet survey. Int J Med Inform 2013;82(3):193-200. [doi: 10.1016/j.ijmedinf.2012.09.008] [Medline: 23149121]
26. Luger TM, Hogan TP, Richardson LM, Cioffari-Bailiff L, Harvey K, Houston TK. Older veteran digital disparities: examining the potential for solutions within social networks. J Med Internet Res 2016;18(11):e296 [FREE Full text] [doi: 10.2196/jmir.6385] [Medline: 27881361]
27. Houston TK, Richardson LM, Cotten SR. Patient-directed digital health technologies: is implementation outpacing evidence? Med Care 2019;57(2):95-97. [doi: 10.1097/MLR.0000000000001068] [Medline: 30601341]

Abbreviations

APP: advanced practice professional
REDCap: Research Electronic Data Capture

Edited by G Eysenbach, T Leung; submitted 26.10.23; peer-reviewed by K Moulaei; comments to author 19.12.23; revised version received 24.01.24; accepted 09.03.24; published 08.04.24.

Please cite as:
Palakshappa JA, Hale ER, Brown JD, Kittel CA, Dressler E, Rosenthal GE, Cutrona SL, Foley KL, Haines ER, Houston II TK Longitudinal Monitoring of Clinician-Patient Video Visits During the Peak of the COVID-19 Pandemic: Adoption and Sustained Challenges in an Integrated Health Care Delivery System
J Med Internet Res 2024;26:e54008
URL: https://www.jmir.org/2024/1/e54008
doi:10.2196/54008
PMID:38587889

©Jessica A Palakshappa, Erica R Hale, Joshua D Brown, Carol A Kittel, Emily Dressler, Gary E Rosenthal, Sarah L Cutrona, Kristie L Foley, Emily R Haines, Thomas K Houston II. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis

Jian Huo[1*], MSc; Yan Yu[2*], MMS; Wei Lin[3], MMS; Anmin Hu[2,3,4], MMS; Chaoran Wu[2], MD, PhD

[1]Boston Intelligent Medical Research Center, Shenzhen United Scheme Technology Company Limited, Boston, MA, United States

[2]Department of Anesthesia, Shenzhen People's Hospital, The First Affiliated Hospital of Southern University of Science and Technology, Shenzhen Key Medical Discipline, Shenzhen, China

[3]Shenzhen United Scheme Technology Company Limited, Shenzhen, China

[4]The Second Clinical Medical College, Jinan University, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Chaoran Wu, MD, PhD
Department of Anesthesia
Shenzhen People's Hospital, The First Affiliated Hospital of Southern University of Science and Technology
Shenzhen Key Medical Discipline
No 1017, Dongmen North Road
Shenzhen, 518020
China
Phone: 86 18100282848
Email: wu.chaoran@szhospital.com

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e59628

Abstract

Background: The continuous monitoring and recording of patients' pain status is a major problem in current research on postoperative pain management. In the large number of original or review articles focusing on different approaches for pain assessment, many researchers have investigated how computer vision (CV) can help by capturing facial expressions. However, there is a lack of proper comparison of results between studies to identify current research gaps.

Objective: The purpose of this systematic review and meta-analysis was to investigate the diagnostic performance of artificial intelligence models for multilevel pain assessment from facial images.

Methods: The PubMed, Embase, IEEE, Web of Science, and Cochrane Library databases were searched for related publications before September 30, 2023. Studies that used facial images alone to estimate multiple pain values were included in the systematic review. A study quality assessment was conducted using the Quality Assessment of Diagnostic Accuracy Studies, 2nd edition tool. The performance of these studies was assessed by metrics including sensitivity, specificity, log diagnostic odds ratio (LDOR), and area under the curve (AUC). The intermodal variability was assessed and presented by forest plots.

Results: A total of 45 reports were included in the systematic review. The reported test accuracies ranged from 0.27-0.99, and the other metrics, including the mean standard error (MSE), mean absolute error (MAE), intraclass correlation coefficient (ICC), and Pearson correlation coefficient (PCC), ranged from 0.31-4.61, 0.24-2.8, 0.19-0.83, and 0.48-0.92, respectively. In total, 6 studies were included in the meta-analysis. Their combined sensitivity was 98% (95% CI 96%-99%), specificity was 98% (95% CI 97%-99%), LDOR was 7.99 (95% CI 6.73-9.31), and AUC was 0.99 (95% CI 0.99-1). The subgroup analysis showed that the diagnostic performance was acceptable, although imbalanced data were still emphasized as a major problem. All studies had at least one domain with a high risk of bias, and for 20% (9/45) of studies, there were no applicability concerns.

Conclusions: This review summarizes recent evidence in automatic multilevel pain estimation from facial expressions and compared the test accuracy of results in a meta-analysis. Promising performance for pain estimation from facial images was established by current CV algorithms. Weaknesses in current studies were also identified, suggesting that larger databases and metrics evaluating multiclass classification performance could improve future studies.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD42023418181; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=418181

(J Med Internet Res 2024;26:e51250) doi:10.2196/51250

KEYWORDS

computer vision; facial image; monitoring; multilevel pain assessment; pain; postoperative; status

Introduction

The definition of pain was revised to "an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" in 2020 [1]. Acute postoperative pain management is important, as pain intensity and duration are critical influencing factors for the transition of acute pain to chronic postsurgical pain [2]. To avoid the development of chronic pain, guidelines were promoted and discussed to ensure safe and adequate pain relief for patients, and clinicians were recommended to use a validated pain assessment tool to track patients' responses [3]. However, these tools, to some extent, depend on communication between physicians and patients, and continuous data cannot be provided [4]. The continuous assessment and recording of patient pain intensity will not only reduce caregiver burden but also provide data for chronic pain research. Therefore, automatic and accurate pain measurements are necessary.

Researchers have proposed different approaches to measuring pain intensity. Physiological signals, for example, electroencephalography and electromyography, have been used to estimate pain [5-7]. However, it was reported that current pain assessment from physiological signals has difficulties isolating stress and pain with machine learning techniques, as they share conceptual and physiological similarities [8]. Recent studies have also investigated pain assessment tools for certain patient subgroups. For example, people with deafness or an intellectual disability may not be able to communicate well with nurses, and an objective pain evaluation would be a better option [9,10]. Measuring pain intensity from patient behaviors, such as facial expressions, is also promising for most patients [4]. As the most comfortable and convenient method, computer vision techniques require no attachments to patients and can monitor multiple participants using 1 device [4]. However, pain intensity, which is important for pain research, is often not reported.

With the growing trend of assessing pain intensity using artificial intelligence (AI), it is necessary to summarize current publications to determine the strengths and gaps of current studies. Existing research has reviewed machine learning applications for acute postoperative pain prediction, continuous pain detection, and pain intensity estimation [10-14]. Input modalities, including facial recordings and physiological signals such as electroencephalography and electromyography, were also reviewed [5,8]. There have also been studies focusing on deep learning approaches [11]. AI was applied in children and infant pain evaluation as well [15,16]. However, no study has focused on pain intensity measurement, and no comparison of test accuracy results has been made.

Current AI applications in pain research can be categorized into 3 types: pain assessment, pain prediction and decision support, and pain self-management [14]. We consider accurate and automatic pain assessment to be the most important area and the foundation of future pain research. In this study, we performed a systematic review and meta-analysis to assess the diagnostic performance of current publications for multilevel pain evaluation.

Methods

This study was registered with PROSPERO (International Prospective Register of Systematic Reviews; CRD42023418181) and carried out strictly following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [17].

Study Eligibility

Studies that reported AI techniques for multiclass pain intensity classification were eligible. Records including nonhuman or infant participants or 2-class pain detection were excluded. Only studies using facial images of the test participants were accepted. Clinically used pain assessment tools, such as the visual analog scale (VAS) and numerical rating scale (NRS), and other pain intensity indicators, were rejected in the meta-analysis. Textbox 1 presents the eligibility criteria.

XSL•FO
RenderX

Textbox 1. Study eligibility criteria.

Study characteristics and inclusion criteria

- Participants: children and adults aged 12 months or older
- Setting: no restrictions
- Index test: artificial intelligence models that measure pain intensity from facial images
- Reference standard: no restrictions for systematic review; Prkachin and Solomon pain intensity score for meta-analysis
- Study design: no need to specify

Study characteristics and exclusion criteria

- Participants: infants aged 12 months or younger and animal subjects
- Setting: no need to specify
- Index test: studies that use other information such as physiological signals
- Reference standard: other pain evaluation tools, e.g., NRS, VAS, were excluded from meta-analysis
- Study design: reviews

Report characteristics and inclusion criteria

- Year: published between January 1, 2012, and September 30, 2023
- Language: English only
- Publication status: published
- Test accuracy metrics: no restrictions for systematic reviews; studies that reported contingency tables were included for meta-analysis

Report characteristics and exclusion criteria

- Year: no need to specify
- Language: no need to specify
- Publication status: preprints not accepted
- Test accuracy metrics: studies that reported insufficient metrics were excluded from meta-analysis

Search Strategy

In this systematic review, databases including PubMed, Embase, IEEE, Web of Science, and the Cochrane Library were searched until December 2022, and no restrictions were applied. Keywords were "artificial intelligence" AND "pain recognition." Multimedia Appendix 1 shows the detailed search strategy.

Data Extraction

A total of 2 viewers screened titles and abstracts and selected eligible records independently to assess eligibility, and disagreements were solved by discussion with a third collaborator. A consentient data extraction sheet was prespecified and used to summarize study characteristics independently. Table S5 in Multimedia Appendix 1 shows the detailed items and explanations for data extraction. Diagnostic accuracy data were extracted into contingency tables, including true positives, false positives, false negatives, and true negatives. The data were used to calculate the pooled diagnostic performance of the different models. Some studies included multiple models, and these models were considered independent of each other.

Study Quality Assessment

All included studies were independently assessed by 2 viewers using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool [18]. QUADAS-2 assesses bias risk across 4 domains, which are patient selection, index test, reference standard, and flow and timing. The first 3 domains are also assessed for applicability concerns. In the systematic review, a specific extension of QUADAS-2, namely, QUADAS-AI, was used to specify the signaling questions [19].

Meta-Analysis

Meta-analyses were conducted between different AI models. Models with different algorithms or training data were considered different. To evaluate the performance differences between models, the contingency tables during model validation were extracted. Studies that did not report enough diagnostic accuracy data were excluded from meta-analysis.

Hierarchical summary receiver operating characteristic (SROC) curves were fitted to evaluate the diagnostic performance of AI models. These curves were plotted with 95% CIs and prediction regions around averaged sensitivity, specificity, and area under the curve estimates. Heterogeneity was assessed visually by forest plots. A funnel plot was constructed to evaluate the risk of bias.

Subgroup meta-analyses were conducted to evaluate the performance differences at both the model level and task level,

XSL•FO
RenderX

and subgroups were created based on different tasks and the proportion of positive and negative samples.

All statistical analyses and plots were produced using RStudio (version 4.2.2; R Core Team) and the R package *meta4diag* (version 2.1.1; Guo J and Riebler A) [20].

Results

Study Selection and Included Study Characteristics

A flow diagram representing the study selection process is shown in (Figure 1). After removing 1039 duplicates, the titles and abstracts of a total of 5653 papers were screened, and the percentage agreement of title or abstract screening was 97%. After screening, 51 full-text reports were assessed for eligibility, among which 45 reports were included in the systematic review [21-65]. The percentage agreement of the full-text review was 87%. In 40 of the included studies, contingency tables could not be made. Meta-analyses were conducted based on 8 AI models extracted from 6 studies. Individual study characteristics included in the systematic review are provided in Tables 1 and 2. The facial feature extraction method can be categorized into 2 classes: geometrical features (GFs) and deep features (DFs). One typical method of extracting GFs is to calculate the distance between facial landmarks. DFs are usually extracted by convolution operations. A total of 20 studies included temporal information, but most of them (18) extracted temporal information through the 3D convolution of video sequences. Feature transformation was also commonly applied to reduce the time for training or fuse features extracted by different methods before inputting them into the classifier. For classifiers, support vector machines (SVMs) and convolutional neural networks (CNNs) were mostly used. Table 1 presents the model designs of the included studies.

Figure 1. Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) flowchart of study selection.

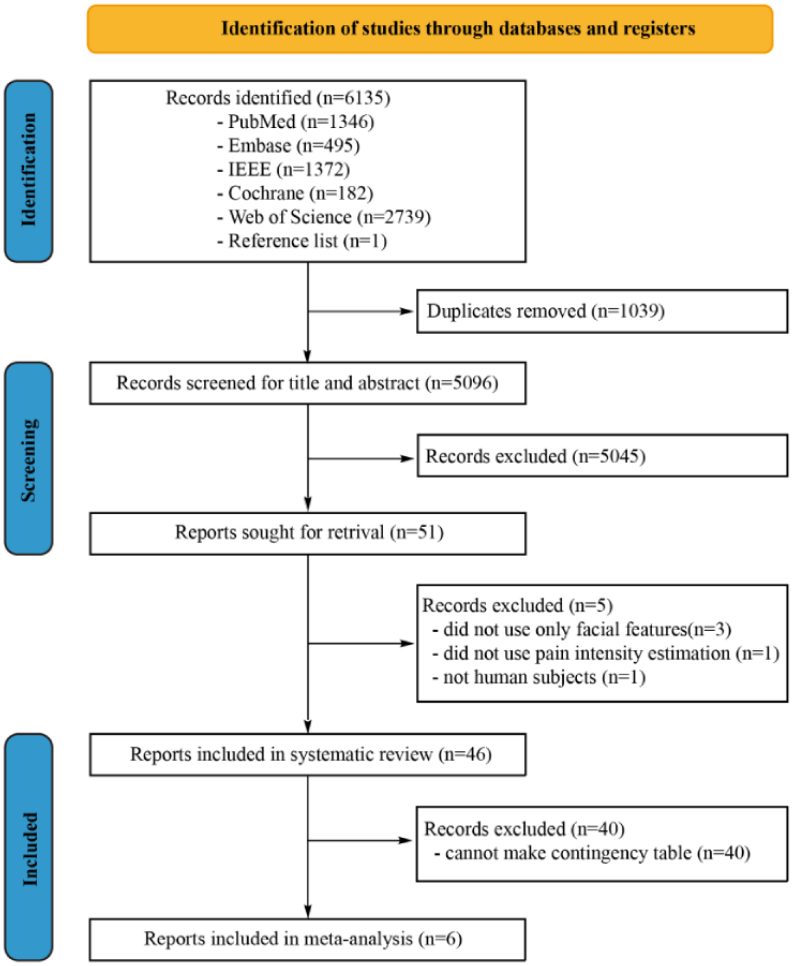

XSL•FO
RenderX

Table 1. Model designs of studies included in the systematic review.

Author and year	Facial feature descriptor	Temporal features[a]	Feature transformation	Classification method
Hammal and Cohn (2012) [21]	CAPP	–	Log-normal filters	SVM[b]
Adibuzzaman et al (2015) [22]	PCA	–	None	Euclidean distance; angular distance; SVM
Majumder et al (2015) [23]	GF[c]; DDF	+	None	GMM[d]; SVM
Rathee and Ganotra (2015) [24]	TPS[e]	+	DML[f]	SVM
Sikka et al (2015) [25]	CERT	++	None	Linear regression
Rathee and Ganotra (2016) [26]	Gabor; LBP; HOG	–	MDML[g]	SVM
Zhou et al (2016) [27]	AAM[h]	++	Flattening	RCNN
Egede et al (2017) [28]	GF; HOG; CNN;	++	RVR[i]	RVR
Martinez et al (2017) [29]	PSPI[j]; I-FES[k];	++	None	LSTM[l]-RNN; RNN-HCRF[m]
Bourou et al (2018) [30]	GF; Color	–	Statistical metrics	GLMM[n]
Haque et al (2018) [31]	CNN-RGB	++	Fine-tuned VGGFace	CNN
Semwal et al (2018) [32]	2D-Conv	–	Maxpooling	CNN
Tavakolian and Hadid (2018) [33]	Pretrained CASIA	++	VLAD[o]	BE
Tavakolian and Hadid (2018) [34]	3D-convolution	++	Average pooling	CNN
Wang and Sun (2018) [35]	3D-convolution; HOG; DFGS	++	SVR[p]	SVR
Bargshady et al (2019) [36]	Fine-tuned VGGFace	–	None	RNN
Casti et al (2019) [37]	LBP	–	MDS[q]	CNN
Lee and Wang (2019) [38]	CNN-RGB	–	None	ELM[r]
Saha et al (2019) [39]	PCA	–	None	NR
Tavakolian and Hadid (2019) [40]	3D-convolution	++	None	CNN
Bargshady et al (2020) [41]	VGGFace	–	PCA;DNN	EDLM[s]
Bargshady et al (2020) [42]	VGGFace	–	PCA	EDLM[s]
Dragomir et al (2020) [43]	ResNet	–	None	ResNet
Huang et al (2020) [44]	AAM	++	RNN-GRU	SVM
Mallol-Ragolta et al (2020) [45]	GF; HOG; HOG; OpenFace; VGGFace; ResNet-50	–	None	LSTM-RNN
Peng et al (2020) [46]	DCNN[t]	–	Probabilistic combination	Multiscale deep fusion network
Tavakolian et al (2020) [47]	GSM[u]	++	Aggregation	SNN
Xu and de Sa (2020) [48]	Handcrafted	++	None	NN
Pikulkaew et al (2021) [49]	None	–	None	DCNN
Rezaei et al (2021) [50]	CNN	++	Flattening	NN
Semwal and Londhe (2021) [51]	CNN	–	None	CNN
Semwal and Londhe (2021) [52]	VGGNet; MobileNet; GoogLeNet	–	None	CNN
Szczapa et al (2021) [53]	Landmark trajectory	++	None	SVR
Ting et al (2021) [54]	None	++	DOML[v]	NN

XSL•FO
RenderX

Author and year	Facial feature descriptor	Temporal features[a]	Feature transformation	Classification method
Xin et al (2021) [55]	CNN	++	None	LIAN[w]
Alghamdi and Alaghband (2022) [56]	OpenCV	–	Flattening	Shallow CNN
Barua et al (2022) [57]	P-DarkNet19	–	INCA	k-NN
Fontaine et al (2022) [58]	OpenCV	–	None	CNN; SVM
Hosseini et al 2022) [59]	Convolution	–	None	DCNN
Huang et al (2022) [60]	3D-CNN (S3D-G)	++	None	CNN
Islamadina et al (2022) [61]	CNN	–	None	CNN
Swetha et al (2022) [62]	None	–	None	CNN
Wu et al (2022) [63]	CNN	+	Siamese network; BiLSTM[x]	NN
Ismail and Waseem (2023) [64]	CNN	–	None	CNN
Vu and Beurton-Aimar (2023) [65]	CNN	–	Average pooling	LSTM network

[a]No temporal features are shown by – symbol, time information extracted from 2 images at different time by +, and deep temporal features extracted through the convolution of video sequences by ++.
[b]SVM: support vector machine.
[c]GF: geometric feature.
[d]GMM: gaussian mixture model.
[e]TPS: thin plate spline.
[f]DML: distance metric learning.
[g]MDML: multiview distance metric learning.
[h]AAM: active appearance model.
[i]RVR: relevance vector regressor.
[j]PSPI: Prkachin and Solomon pain intensity.
[k]I-FES: individual facial expressiveness score.
[l]LSTM: long short-term memory.
[m]HCRF: hidden conditional random field.
[n]GLMM: generalized linear mixed model.
[o]VLAD: vector of locally aggregated descriptor.
[p]SVR: support vector regression.
[q]MDS: multidimensional scaling.
[r]ELM: extreme learning machine.
[s]Labeled to distinguish different architectures of ensembled deep learning models.
[t]DCNN: deep convolutional neural network.
[u]GSM: gaussian scale mixture.
[v]DOML: distance ordering metric learning.
[w]LIAN: locality and identity aware network.
[x]BiLSTM: bidirectional long short-term memory.

XSL•FO
RenderX

Table 2. Characteristics of model training and validation.

Author and year	Database	Objects	Output levels	Validation method	External validation	Evaluation metrics
Hammal and Cohn (2012) [21]	UNBC[a]	Frame	4	5-fold; LOSO[b]	No	ICC[c] 0.85, 0.55; F_1-score 0.96, 0.67
Adibuzzaman et al (2015) [22]	Self-prepared	Image	3	10-fold	Yes	Sensitivity 0.53; Specificity 0.7
Majumder et al (2015) [23]	UNBC	Frame	16	5-fold	No	Accuracy 87.43
Rathee and Ganotra (2015) [24]	UNBC	Frame	16	LOO; 10-fold	No	Accuracy 0.96; CT[d]
Sikka et al (2015) [25]	Self-prepared	Sequence	11	LOSO	No	AUC[e] 0.94; Cohen κ0.61
Rathee and Ganotra (2016) [26]	UNBC	Frame	4	5-fold	No	Accuracy 0.75
Zhou et al (2016) [27]	UNBC	Frame	16	LOSO	No	MSE[f] 1.54; PCC[g] 0.65
Egede et al (2017) [28]	UNBC	Frame	16	LOSO	No	RMSE[h]<1; PCC 0.67
Martinez et al (2017) [29]	UNBC	Sequence	11	Split	No	MAE[i] 2.8; ICC[j] 0.19
Bourou et al (2018) [30]	BioVid	Frame	5	10-fold	No	Accuracy 0.27; RCI 0.03
Haque et al (2018) [31]	MIntPain	Frame	5	5-fold	No	CT
Semwal and Londhe (2018) [32]	UNBC	Frame	3	Split	No	CT; Accuracy 0.93
Tavakolian and Hadid (2018) [33]	UNBC	Frame	16	LOSO	No	MSE 0.69; PCC 0.81
Tavakolian and Hadid (2018) [34]	UNBC	Frame	16	LOSO	No	MSE 0.53; ICC 0.75; PCC 0.84
Wang and Sun (2018) [35]	UNBC	Frame	16	LOSO	No	RMSE 0.94; PCC 0.68
Bargshady et al (2019) [36]	UNBC	Frame	4	LOSO	No	Accuracy 0.75; AUC 0.83; MSE 0.95
Casti et al (2019) [37]	UNBC	Frame	11	Split	No	Recall 0.92; Precision 0.82
Lee and Wang (2019) [38]	UNBC	Frame	16	5-fold	No	MSE 1.22; PCC 0.5
Saha et al (2019) [39]	Self-prepared	Image	3	10-fold	No	Accuracy 0.71; CT
Tavakolian and Hadid (2019) [40]	UNBC	Frame	5; 16	LOSO	No	MSE 0.32; PCC 0.92; AUC 0.86
Bargshady et al (2020) [41]	MIntPain; UNBC	Frame	5	10-fold	No	Accuracy 0.89; AUC 0.93
Bargshady et al (2020) [42]	UNBC	Frame	4	10-fold	No	Accuracy 0.91; AUC 0.98
Dragomir et al (2020) [43]	BioVid	Frame	5	CV	No	Accuracy 36.6
Huang et al (2020) [44]	UNBC	Frame	6	Split	No	PCC 0.89; ICC 0.72; MSE 0.21; MAE 0.24
Mallol-Ragolta et al (2020) [45]	EmoPain	Frame	11	Split	No	CCC[k] 0.174
Peng et al (2020) [46]	UNBC	Frame	5	NR	No	Accuracy 0.80; PCC 0.6; MAE 0.57; MSE 0.82
Tavakolian et al (2020) [47]	BioVid; UNBC	Frame	5; 16	LOSO	Yes	MSE 1.03, 0.92; AUC 0.69, 0.71
Xu and de Sa (2020) [48]	UNBC	Sequence	6; 11; 16; 16	5-fold	No	MSE 4.61; MAE 1.73; ICC 0.61; PCC 0.67
Pikulkaew et al (2021) [49]	UNBC	Frame	3	NR	No	Accuracy 0.93
Rezaei et al (2021) [50]	UofR; UNBC	Frame	16	5-fold	Yes	PCC 0.48-0.7; ICC 0.31-0.59[l]
Semwal and Londhe (2021) [51]	Self-prepared	Frame	4	5-fold	No	CT; Accuracy 0.97
Semwal and Londhe (2021) [52]	UNBC	Frame	5	10-fold	No	CT; F_1-score 0.91

XSL•FO
RenderX

Author and year	Database	Objects	Output levels	Validation method	External validation	Evaluation metrics
Szczapa et al (2021) [53]	UNBC	Sequence	11	5-fold; LOO; LOSO	No	MAE 2.44; RMSE 3.15
Ting et al (2021) [54]	UNBC	Sequence	11	5-fold; LOSO	No	MAE 1.62; MSE 4.39; ICC 0.66
Xin et al (2021) [55]	UNBC	Frame	4	LOSO	No	Accuracy 0.89; ICC 0.61; PCC 0.81; MAE 0.45; MSE 0.66
Alghamdi and Alaghband (2022) [56]	UNBC	Frame	4	Split	No	Accuracy 0.99
Barua et al (2022) [57]	DISFA; UNBC	Frame	4	10-fold	No	CT; Accuracy 0.95
Fontaine et al (2022) [58]	Self-prepared	Frame	4	Split	No	Sensitivity 0.90
Hosseini et al (2022) [59]	UNBC	Frame	7	NR	No	Accuracy 0.85; AUC 0.88; PCC 0.83
Huang et al (2022) [60]	UNBC	Frame	16	LOSO	No	MAE 0.4; MSE 0.76; PCC 0.82
Islamadina et al (2022) [61]	MIntPian	Frame	5	CV	No	CT; Accuracy 1.0
Swetha et al (2022) [62]	Self-prepared	Frame	4	NR	No	Accuracy 0.75
Wu et al (2022) [63]	Self-prepared	Frame; sequence	3	Split	No	Accuracy 0.81
Ismail and Waseem 2023 [64]	UNBC	Frame	16	5-fold	No	MAE 0.36; MSE 1.73; Accuracy 0.82
Vu and Beurton-Aimar 2023 [65]	DISFA; UNBC	Frame	16	LOSO	No	MSE 0.57; MAE 0.35; ICC 0.83; PCC 0.81

[a]UNBC: University of Northern British Columbia-McMaster shoulder pain expression archive database.

[b]LOSO: leave one subject out cross-validation.

[c]ICC: intraclass correlation coefficient.

[d]CT: contingency table.

[e]AUC: area under the curve.

[f]MSE: mean standard error.

[g]PCC: Pearson correlation coefficient.

[h]RMSE: root mean standard error.

[i]MAE: mean absolute error.

[j]ICC: intraclass coefficient.

[k]CCC: concordance correlation coefficient.

[l]Reported both external and internal validation results and summarized as intervals.

Table 2 summarizes the characteristics of model training and validation. Most studies used publicly available databases, for example, the University of Northern British Columbia-McMaster shoulder pain expression archive database [57]. Table S4 in Multimedia Appendix 1 summarizes the public databases. A total of 7 studies used self-prepared databases. Frames from video sequences were the most used test objects, as 37 studies output frame-level pain intensity, while few measure pain intensity from video sequences or photos. It was common that a study redefined pain levels to have fewer classes than ground-truth labels. For model validation, cross-validation and leave-one-subject-out validation were commonly used. Only 3 studies performed external validation. For reporting test accuracies, different evaluation metrics were used, including sensitivity, specificity, mean absolute error (MAE), mean standard error (MSE), Pearson correlation coefficient (PCC), and intraclass coefficient (ICC).

Methodological Quality of Included Studies

Table S2 in Multimedia Appendix 1 presents the study quality summary, as assessed by QUADAS-2. There was a risk of bias in all studies, specifically in terms of patient selection, caused by 2 issues. First, the training data are highly imbalanced, and any method to adjust the data distribution may introduce bias. Next, the QUADAS-AI correspondence letter [19] specifies that preprocessing of images that changes the image size or resolution may introduce bias. However, the applicability concern is low, as the images properly represent the feeling of pain. Studies that used cross-fold validation or leave-one-out cross-validation were considered to have a low risk of bias. Although the Prkachin and Solomon pain intensity (PSPI) score was used by most of the studies, its ability to represent individual pain levels was not clinically validated; as such, the risk of bias and applicability concerns were considered high

XSL•FO
RenderX

when the PSPI score was used as the index test. As an advantage of computer vision techniques, the time interval between the index tests was short and was assessed as having a low risk of bias. Risk proportions are shown in Figure 2. For all 315 entries, 39% (124) were assessed as high-risk. In total, 5 studies had the lowest risk of bias, with 6 domains assessed as low risk [26,27,31,32,59].

Figure 2. Risk of bias and applicability proportions. QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2.

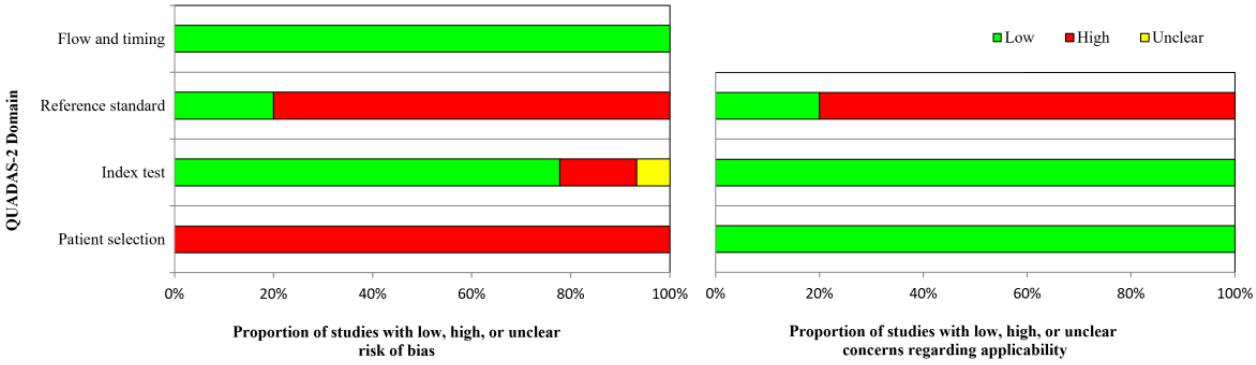

Pooled Performance of Included Models

In 6 studies included in the meta-analysis, there were 8 different models. The characteristics of these models are summarized in Table S1 in Multimedia Appendix 2 [23,24,26,32,41,57]. Classification of PSPI scores greater than 0, 2, 3, 6, and 9 was selected and considered as different tasks to create contingency tables. The test performance is shown in Figure 3 as hierarchical SROC curves; 27 contingency tables were extracted from 8 models. The sensitivity, specificity, and LDOR were calculated, and the combined sensitivity was 98% (95% CI 96%-99%), the specificity was 98% (95% CI 97%-99%), the LDOR was 7.99 (95% CI 6.73-9.31) and the AUC was 0.99 (95% CI 0.99-1).

Figure 3. Summary receiver operating characteristic (SROC) curve plots of the summarized results.

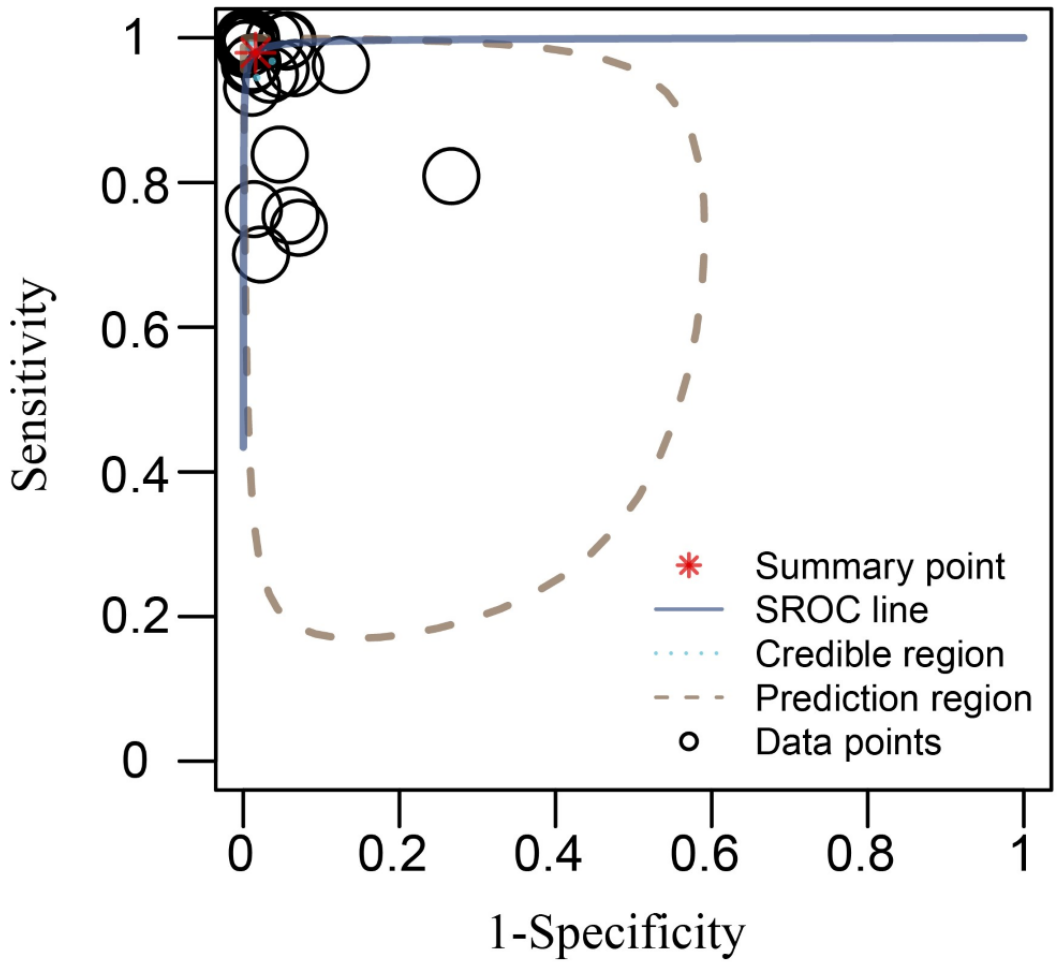

XSL•FO
RenderX

Subgroup Analysis

In this study, subgroup analysis was conducted to investigate the performance differences within models. A total of 8 models were separated and summarized as a forest plot in Multimedia Appendix 3 [23,24,26,32,41,57]. For model 1, the pooled sensitivity, specificity, and LDOR were 95% (95% CI 86%-99%), 99% (95% CI 98%-100%), and 8.38 (95% CI 6.09-11.19), respectively. For model 2, the pooled sensitivity, specificity, and LDOR were 94% (95% CI 84%-99%), 95% (95% CI 88%-99%), and 6.23 (95% CI 3.52-9.04), respectively. For model 3, the pooled sensitivity, specificity, and LDOR were 100% (95% CI 99%-100%), 100% (95% CI 99%-100%), and 11.55% (95% CI 8.82-14.43), respectively. For model 4, the pooled sensitivity, specificity, and LDOR were 83% (95% CI 43%-99%), 94% (95% CI 79%-99%), and 5.14 (95% CI 0.93-9.31), respectively. For model 5, the pooled sensitivity, specificity, and LDOR were 92% (95% CI 68%-99%), 94% (95% CI 78%-99%), and 6.12 (95% CI 1.82-10.16), respectively. For model 6, the pooled sensitivity, specificity, and LDOR were 94% (95% CI 74%-100%), 94% (95% CI 78%-99%), and 6.59 (95% CI 2.21-11.13), respectively. For model 7, the pooled sensitivity, specificity, and LDOR were 98% (95% CI 90%-100%), 97% (95% CI 87%-100%), and 8.31 (95% CI 4.3-12.29), respectively. For model 8, the pooled sensitivity, specificity, and LDOR were 98% (95% CI 93%-100%), 97% (95% CI 88%-100%), and 8.65 (95% CI 4.84-12.67), respectively.

Heterogeneity Analysis

The meta-analysis results indicated that AI models are applicable for estimating pain intensity from facial images. However, extreme heterogeneity existed within the models except for models 3 and 5, which were proposed by Rathee and Ganotra [24] and Semwal and Londhe [32]. A funnel plot is presented in Figure 4. A high risk of bias was observed.

Figure 4. In the funnel plot of the test results, significant heterogeneity was observed. DOR: diagnostic odds ratio; LDOR: log diagnostic odds ratio.

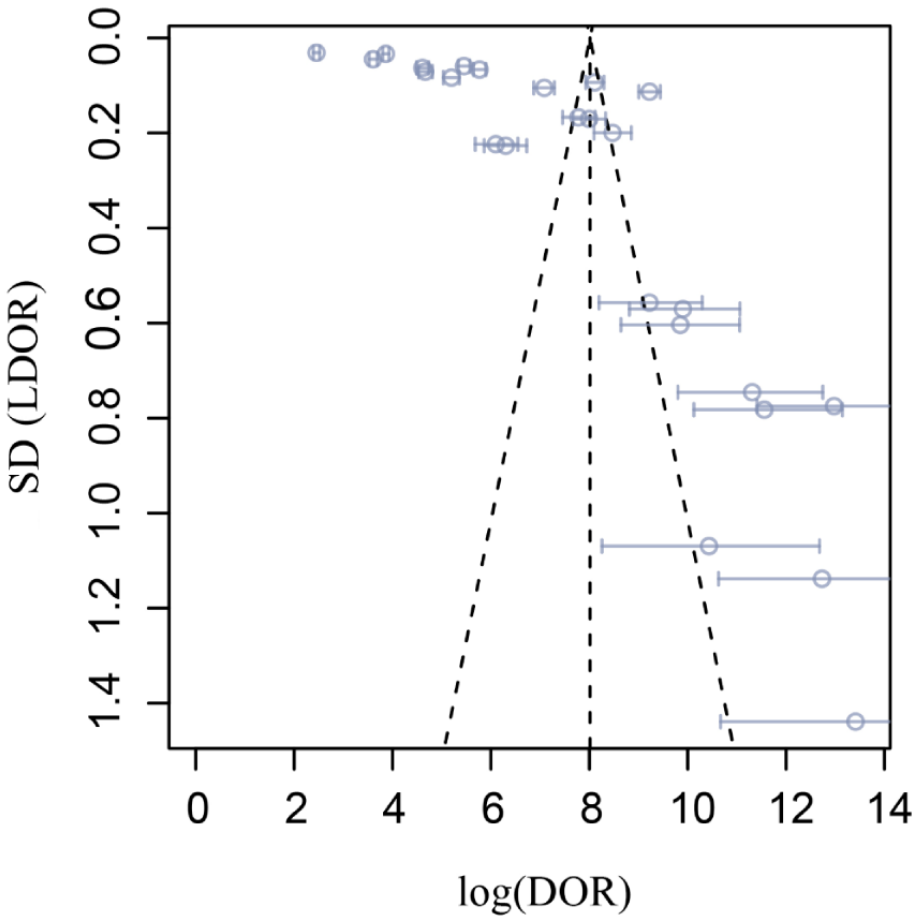

Discussion

Pain management has long been a critical problem in clinical practice, and the use of AI may be a solution. For acute pain management, automatic measurement of pain can reduce the burden on caregivers and provide timely warnings. For chronic pain management, as specified by Glare et al [2], further research is needed, and measurements of pain presence, intensity, and quality are one of the issues to be solved for chronic pain studies. Computer vision could improve pain monitoring through real-time detection for clinical use and data recording for prospective pain studies. To our knowledge, this is the first meta-analysis dedicated to AI performance in multilevel pain level classification.

In this study, one model's performance at specific pain levels was described by stacking multiple classes into one to make each task a binary classification problem. After careful selection in both the medical and engineering databases, we observed

XSL•FO
RenderX

promising results of AI in evaluating multilevel pain intensity through facial images, with high sensitivity (98%), specificity (98%), LDOR (7.99), and AUC (0.99). It is reasonable to believe that AI can accurately evaluate pain intensity from facial images. Moreover, the study quality and risk of bias were evaluated using an adapted QUADAS-2 assessment tool, which is a strength of this study.

To investigate the source of heterogeneity, it was assumed that a well-designed model should have familiar size effects regarding different levels, and a subgroup meta-analysis was conducted. The funnel and forest plots exhibited extreme heterogeneity. The model's performance at specific pain levels was described and summarized by a forest plot. Within-model heterogeneity was observed in Multimedia Appendix 3 [23,24,26,32,41,57] except for 2 models. Models 3 and 5 were different in many aspects, including their algorithms and validation methods, but were both trained with a relatively small data set, and the proportion of positive and negative classes was relatively close to 1. Because training with imbalanced data is a critical problem in computer vision studies [66], for example, in the University of Northern British Columbia-McMaster pain data set, fewer than 10 frames out of 48,398 had a PSPI score greater than 13. Here, we emphasized that imbalanced data sets are one major cause of heterogeneity, resulting in the poorer performance of AI algorithms.

We tentatively propose a method to minimize the effect of training with imbalanced data by stacking multiple classes into one class, which is already presented in studies included in the systematic review [26,32,42,57]. Common methods to minimize bias include resampling and data augmentation [66]. This proposed method is used in the meta-analysis to compare the test results of different studies as well. The stacking method is available when classes are only different in intensity. A disadvantage of combined classes is that the model would be insufficient in clinical practice when the number of classes is low. Commonly used pain evaluation tools, such as VAS, have 10 discrete levels. It is recommended that future studies set the number of pain levels to be at least 10 for model training.

This study is limited for several reasons. First, insufficient data were included because different performance metrics (mean standard error and mean average error) were used in most studies, which could not be summarized into a contingency table. To create a contingency table that can be included in a meta-analysis, the study should report the following: the number of objects used in each pain class for model validation, and the accuracy, sensitivity, specificity, and F_1-score for each pain class. This table cannot be created if a study reports the MAE, PCC, and other commonly used metrics in AI development. Second, a small study effect was observed in the funnel plot, and the heterogeneity could not be minimized. Another limitation is that the PSPI score is not clinically validated and is not the only tool that assesses pain from facial expressions. There are other clinically validated pain intensity assessment methods, such as the Faces Pain Scale-revised, Wong-Baker Faces Pain Rating Scale, and Oucher Scale [3]. More databases could be created based on the above-mentioned tools. Finally, AI-assisted pain assessments were supposed to cover larger populations, including incommunicable patients, for example, patients with dementia or patients with masked faces. However, only 1 study considered patients with dementia, which was also caused by limited databases [50].

AI is a promising tool that can help in pain research in the future. In this systematic review and meta-analysis, one approach using computer vision was investigated to measure pain intensity from facial images. Despite some risk of bias and applicability concerns, CV models can achieve excellent test accuracy. Finally, more CV studies in pain estimation, reporting accuracy in contingency tables, and more pain databases are encouraged for future studies. Specifically, the creation of a balanced public database that contains not only healthy but also nonhealthy participants should be prioritized. The recording process would be better in a clinical environment. Then, it is recommended that researchers report the validation results in terms of accuracy, sensitivity, specificity, or contingency tables, as well as the number of objects for each pain class, for the inclusion of a meta-analysis.

Acknowledgments

WL, AH, and CW contributed to the literature search and data extraction. JH and YY wrote the first draft of the manuscript. All authors contributed to the conception and design of the study, the risk of bias evaluation, data analysis and interpretation, and contributed to and approved the final version of the manuscript.

Data Availability

The data sets generated during and analyzed during this study are available in the Figshare repository [67].

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist, risk of bias summary, search strategy, database summary and reported items and explanations.
[DOC File , 154 KB - jmir_v26i1e51250_app1.doc]

XSL•FO
RenderX

Multimedia Appendix 2
Study performance summary.
[XLS File (Microsoft Excel File), 35 KB - jmir_v26i1e51250_app2.xls]

Multimedia Appendix 3
Forest plot presenting pooled performance of subgroups in meta-analysis.
[PNG File , 257 KB - jmir_v26i1e51250_app3.png]

References

1. Raja SN, Carr DB, Cohen M, Finnerup NB, Flor H, Gibson S, et al. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. Pain 2020;161(9):1976-1982 [FREE Full text] [doi: 10.1097/j.pain.0000000000001939] [Medline: 32694387]
2. Glare P, Aubrey KR, Myles PS. Transition from acute to chronic pain after surgery. Lancet 2019;393(10180):1537-1546. [doi: 10.1016/S0140-6736(19)30352-6] [Medline: 30983589]
3. Chou R, Gordon DB, de Leon-Casasola OA, Rosenberg JM, Bickler S, Brennan T, et al. Management of postoperative pain: a clinical practice guideline from the American Pain Society, the American Society of Regional Anesthesia and Pain Medicine, and the American Society of Anesthesiologists' Committee on Regional Anesthesia, Executive Committee, and Administrative Council. J Pain 2016;17(2):131-157 [FREE Full text] [doi: 10.1016/j.jpain.2015.12.008] [Medline: 26827847]
4. Hassan T, Seus D, Wollenberg J, Weitz K, Kunz M, Lautenbacher S, et al. Automatic detection of pain from facial expressions: a survey. IEEE Trans Pattern Anal Mach Intell 2021;43(6):1815-1831. [doi: 10.1109/TPAMI.2019.2958341] [Medline: 31825861]
5. Mussigmann T, Bardel B, Lefaucheur JP. Resting-State Electroencephalography (EEG) biomarkers of chronic neuropathic pain. A systematic review. Neuroimage 2022;258:119351 [FREE Full text] [doi: 10.1016/j.neuroimage.2022.119351] [Medline: 35659993]
6. Moscato S, Cortelli P, Chiari L. Physiological responses to pain in cancer patients: a systematic review. Comput Methods Programs Biomed 2022;217:106682 [FREE Full text] [doi: 10.1016/j.cmpb.2022.106682] [Medline: 35172252]
7. Thiam P, Hihn H, Braun DA, Kestler HA, Schwenker F. Multi-modal pain intensity assessment based on physiological signals: a deep learning perspective. Front Physiol 2021;12:720464 [FREE Full text] [doi: 10.3389/fphys.2021.720464] [Medline: 34539444]
8. Rojas RF, Brown N, Waddington G, Goecke R. A systematic review of neurophysiological sensing for the assessment of acute pain. NPJ Digit Med 2023;6(1):76 [FREE Full text] [doi: 10.1038/s41746-023-00810-1] [Medline: 37100924]
9. Mansutti I, Tomé-Pires C, Chiappinotto S, Palese A. Facilitating pain assessment and communication in people with deafness: a systematic review. BMC Public Health 2023;23(1):1594 [FREE Full text] [doi: 10.1186/s12889-023-16535-5] [Medline: 37608263]
10. El-Tallawy SN, Ahmed RS, Nagiub MS. Pain management in the most vulnerable intellectual disability: a review. Pain Ther 2023;12(4):939-961 [FREE Full text] [doi: 10.1007/s40122-023-00526-w] [Medline: 37284926]
11. Gkikas S, Tsiknakis M. Automatic assessment of pain based on deep learning methods: a systematic review. Comput Methods Programs Biomed 2023;231:107365 [FREE Full text] [doi: 10.1016/j.cmpb.2023.107365] [Medline: 36764062]
12. Borna S, Haider CR, Maita KC, Torres RA, Avila FR, Garcia JP, et al. A review of voice-based pain detection in adults using artificial intelligence. Bioengineering (Basel) 2023;10(4):500 [FREE Full text] [doi: 10.3390/bioengineering10040500] [Medline: 37106687]
13. De Sario GD, Haider CR, Maita KC, Torres-Guzman RA, Emam OS, Avila FR, et al. Using AI to detect pain through facial expressions: a review. Bioengineering (Basel) 2023;10(5):548 [FREE Full text] [doi: 10.3390/bioengineering10050548] [Medline: 37237618]
14. Zhang M, Zhu L, Lin SY, Herr K, Chi CL, Demir I, et al. Using artificial intelligence to improve pain assessment and pain management: a scoping review. J Am Med Inform Assoc 2023;30(3):570-587 [FREE Full text] [doi: 10.1093/jamia/ocac231] [Medline: 36458955]
15. Hughes JD, Chivers P, Hoti K. The clinical suitability of an artificial intelligence-enabled pain assessment tool for use in infants: feasibility and usability evaluation study. J Med Internet Res 2023;25:e41992 [FREE Full text] [doi: 10.2196/41992] [Medline: 36780223]
16. Fang J, Wu W, Liu J, Zhang S. Deep learning-guided postoperative pain assessment in children. Pain 2023;164(9):2029-2035 [FREE Full text] [doi: 10.1097/j.pain.0000000000002900] [Medline: 37146182]
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

XSL•FO
RenderX

19. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021;27(10):1663-1665 [FREE Full text] [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
20. Guo J, Riebler A. meta4diag: Bayesian bivariate meta-analysis of diagnostic test studies for routine practice. J Stat Soft 2018;83(1):1-31. [doi: 10.18637/jss.v083.i01]
21. Hammal Z, Cohn JF. Automatic detection of pain intensity. Proc ACM Int Conf Multimodal Interact 2012;2012:47-52 [FREE Full text] [doi: 10.1145/2388676.2388688] [Medline: 32724903]
22. Adibuzzaman M, Ostberg C, Ahamed S, Povinelli R, Sindhu B, Love R, et al. Assessment of pain using facial pictures taken with a smartphone. 2015 Presented at: 2015 IEEE 39th Annual Computer Software and Applications Conference; July 01-05, 2015; Taichung, Taiwan p. 726-731. [doi: 10.1109/compsac.2015.150]
23. Majumder A, Dutta S, Behera L, Subramanian VK. Shoulder pain intensity recognition using Gaussian mixture models. 2015 Presented at: 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE); December 19-20, 2015; Dhaka, Bangladesh p. 130-134. [doi: 10.1109/wiecon-ece.2015.7444016]
24. Rathee N, Ganotra D. A novel approach for pain intensity detection based on facial feature deformations. J Vis Commun Image Represent 2015;33:247-254. [doi: 10.1016/j.jvcir.2015.09.007]
25. Sikka K, Ahmed AA, Diaz D, Goodwin MS, Craig KD, Bartlett MS, et al. Automated assessment of children's postoperative pain using computer vision. Pediatrics 2015;136(1):e124-e131 [FREE Full text] [doi: 10.1542/peds.2015-0029] [Medline: 26034245]
26. Rathee N, Ganotra D. Multiview distance metric learning on facial feature descriptors for automatic pain intensity detection. Comput Vis Image Und 2016;147:77-86. [doi: 10.1016/j.cviu.2015.12.004]
27. Zhou J, Hong X, Su F, Zhao G. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 26-July 01, 2016; Las Vegas, NV. [doi: 10.1109/cvprw.2016.191]
28. Egede J, Valstar M, Martinez B. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. 2017 Presented at: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017); May 30-June 03, 2017; Washington, DC p. 689-696. [doi: 10.1109/fg.2017.87]
29. Martinez DL, Rudovic O, Picard R. Personalized automatic estimation of self-reported pain intensity from facial expressions. 2017 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); July 21-26, 2017; Honolulu, HI p. 2318-2327. [doi: 10.1109/cvprw.2017.286]
30. Bourou D, Pampouchidou A, Tsiknakis M, Marias K, Simos P. Video-based pain level assessment: feature selection and inter-subject variability modeling. 2018 Presented at: 2018 41st International Conference on Telecommunications and Signal Processing (TSP); July 04-06, 2018; Athens, Greece p. 1-6. [doi: 10.1109/tsp.2018.8441252]
31. Haque MA, Bautista RB, Noroozi F, Kulkarni K, Laursen C, Irani R. Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities. 2018 Presented at: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); May 15-19, 2018; Xi'an, China p. 250-257. [doi: 10.1109/fg.2018.00044]
32. Semwal A, Londhe ND. Automated pain severity detection using convolutional neural network. 2018 Presented at: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS); December 21-22, 2018; Belgaum, India p. 66-70. [doi: 10.1109/ctems.2018.8769123]
33. Tavakolian M, Hadid A. Deep binary representation of facial expressions: a novel framework for automatic pain intensity recognition. 2018 Presented at: 2018 25th IEEE International Conference on Image Processing (ICIP); October 07-10, 2018; Athens, Greece p. 1952-1956. [doi: 10.1109/icip.2018.8451681]
34. Tavakolian M, Hadid A. Deep spatiotemporal representation of the face for automatic pain intensity estimation. 2018 Presented at: 2018 24th International Conference on Pattern Recognition (ICPR); August 20-24, 2018; Beijing, China p. 350-354. [doi: 10.1109/icpr.2018.8545324]
35. Wang J, Sun H. Pain intensity estimation using deep spatiotemporal and handcrafted features. IEICE Trans Inf & Syst 2018;E101.D(6):1572-1580. [doi: 10.1587/transinf.2017edp7318]
36. Bargshady G, Soar J, Zhou X, Deo RC, Whittaker F, Wang H. A joint deep neural network model for pain recognition from face. 2019 Presented at: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS); February 23-25, 2019; Singapore p. 52-56. [doi: 10.1109/ccoms.2019.8821779]
37. Casti P, Mencattini A, Comes MC, Callari G, Di Giuseppe D, Natoli S, et al. Calibration of vision-based measurement of pain intensity with multiple expert observers. IEEE Trans Instrum Meas 2019;68(7):2442-2450. [doi: 10.1109/tim.2019.2909603]
38. Lee JS, Wang CW. Facial pain intensity estimation for ICU patient with partial occlusion coming from treatment. 2019 Presented at: BIBE 2019; The Third International Conference on Biological Information and Biomedical Engineering; June 20-22, 2019; Hangzhou, China p. 1-4.
39. Saha AK, Ahsan GMT, Gani MO, Ahamed SI. Personalized pain study platform using evidence-based continuous learning tool. 2019 Presented at: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC); July 15-19, 2019; Milwaukee, WI p. 490-495. [doi: 10.1109/compsac.2019.10254]

XSL•FO
RenderX

40. Tavakolian M, Hadid A. A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics. Int J Comput Vis 2019;127(10):1413-1425 [FREE Full text] [doi: 10.1007/s11263-019-01191-3]
41. Bargshady G, Zhou X, Deo RC, Soar J, Whittaker F, Wang H. Ensemble neural network approach detecting pain intensity from facial expressions. Artif Intell Med 2020;109:101954. [doi: 10.1016/j.artmed.2020.101954] [Medline: 34756219]
42. Bargshady G, Zhou X, Deo RC, Soar J, Whittaker F, Wang H. Enhanced deep learning algorithm development to detect pain intensity from facial expression images. Expert Syst Appl 2020;149:113305. [doi: 10.1016/j.eswa.2020.113305]
43. Dragomir MC, Florea C, Pupezescu V. Automatic subject independent pain intensity estimation using a deep learning approach. 2020 Presented at: 2020 International Conference on e-Health and Bioengineering (EHB); October 29-30, 2020; Iasi, Romania p. 1-4. [doi: 10.1109/ehb50910.2020.9280190]
44. Huang D, Xia Z, Mwesigye J, Feng X. Pain-attentive network: a deep spatio-temporal attention model for pain estimation. Multimed Tools Appl 2020;79(37-38):28329-28354. [doi: 10.1007/s11042-020-09397-1]
45. Mallol-Ragolta A, Liu S, Cummins N, Schuller B. A curriculum learning approach for pain intensity recognition from facial expressions. 2020 Presented at: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020); November 16-20, 2020; Buenos Aires, Argentina p. 829-833. [doi: 10.1109/fg47880.2020.00083]
46. Peng X, Huang D, Zhang H. Pain intensity recognition via multi‐scale deep network. IET Image Process 2020;14(8):1645-1652 [FREE Full text] [doi: 10.1049/iet-ipr.2019.1448]
47. Tavakolian M, Lopez MB, Liu L. Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. Pattern Recognit Lett 2020;140:26-33. [doi: 10.1016/j.patrec.2020.09.012]
48. Xu X, de Sa VR. Exploring multidimensional measurements for pain evaluation using facial action units. 2020 Presented at: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020); November 16-20, 2020; Buenos Aires, Argentina p. 786-792. [doi: 10.1109/fg47880.2020.00087]
49. Pikulkaew K, Boonchieng W, Boonchieng E, Chouvatut V. 2D facial expression and movement of motion for pain identification with deep learning methods. IEEE Access 2021;9:109903-109914. [doi: 10.1109/access.2021.3101396]
50. Rezaei S, Moturu A, Zhao S, Prkachin KM, Hadjistavropoulos T, Taati B. Unobtrusive pain monitoring in older adults with dementia using pairwise and contrastive training. IEEE J Biomed Health Inform 2021;25(5):1450-1462. [doi: 10.1109/JBHI.2020.3045743] [Medline: 33338024]
51. Semwal A, Londhe ND. S-PANET: a shallow convolutional neural network for pain severity assessment in uncontrolled environment. 2021 Presented at: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC); January 27-30, 2021; Las Vegas, NV p. 0800-0806. [doi: 10.1109/ccwc51732.2021.9376052]
52. Semwal A, Londhe ND. ECCNet: an ensemble of compact convolution neural network for pain severity assessment from face images. 2021 Presented at: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence); January 28-29, 2021; Noida, India p. 761-766. [doi: 10.1109/confluence51648.2021.9377197]
53. Szczapa B, Daoudi M, Berretti S, Pala P, Del Bimbo A, Hammal Z. Automatic estimation of self-reported pain by interpretable representations of motion dynamics. 2021 Presented at: 2020 25th International Conference on Pattern Recognition (ICPR); January 10-15, 2021; Milan, Italy p. 2544-2550. [doi: 10.1109/icpr48806.2021.9412292]
54. Ting J, Yang YC, Fu LC, Tsai CL, Huang CH. Distance ordering: a deep supervised metric learning for pain intensity estimation. 2021 Presented at: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA); December 13-16, 2021; Pasadena, CA p. 1083-1088. [doi: 10.1109/icmla52953.2021.00177]
55. Xin X, Li X, Yang S, Lin X, Zheng X. Pain expression assessment based on a locality and identity aware network. IET Image Process 2021;15(12):2948-2958 [FREE Full text] [doi: 10.1049/ipr2.12282]
56. Alghamdi T, Alaghband G. Facial expressions based automatic pain assessment system. Appl Sci 2022;12(13):6423 [FREE Full text] [doi: 10.3390/app12136423]
57. Barua PD, Baygin N, Dogan S, Baygin M, Arunkumar N, Fujita H, et al. Automated detection of pain levels using deep feature extraction from shutter blinds-based dynamic-sized horizontal patches with facial images. Sci Rep 2022;12(1):17297 [FREE Full text] [doi: 10.1038/s41598-022-21380-4] [Medline: 36241674]
58. Fontaine D, Vielzeuf V, Genestier P, Limeux P, Santucci-Sivilotto S, Mory E, et al. Artificial intelligence to evaluate postoperative pain based on facial expression recognition. Eur J Pain 2022;26(6):1282-1291. [doi: 10.1002/ejp.1948] [Medline: 35352426]
59. Hosseini E, Fang R, Zhang R, Chuah CN, Orooji M, Rafatirad S, et al. Convolution neural network for pain intensity assessment from facial expression. 2022 Presented at: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); July 11-15, 2022; Glasgow, Scotland p. 2697-2702. [doi: 10.1109/embc48229.2022.9871770]
60. Huang Y, Qing L, Xu S, Wang L, Peng Y. HybNet: a hybrid network structure for pain intensity estimation. Vis Comput 2021;38(3):871-882. [doi: 10.1007/s00371-021-02056-y]
61. Islamadina R, Saddami K, Oktiana M, Abidin TF, Muharar R, Arnia F. Performance of deep learning benchmark models on thermal imagery of pain through facial expressions. 2022 Presented at: 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT); November 03-05, 2022; Solo, Indonesia p. 374-379. [doi: 10.1109/comnetsat56033.2022.9994546]

XSL•FO
RenderX

62. Swetha L, Praiscia A, Juliet S. Pain assessment model using facial recognition. 2022 Presented at: 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS); May 25-27, 2022; Madurai, India p. 1-5. [doi: 10.1109/iciccs53718.2022.9788265]
63. Wu CL, Liu SF, Yu TL, Shih SJ, Chang CH, Mao SFY, et al. Deep learning-based pain classifier based on the facial expression in critically ill patients. Front Med (Lausanne) 2022;9:851690 [FREE Full text] [doi: 10.3389/fmed.2022.851690] [Medline: 35372435]
64. Ismail L, Waseem MD. Towards a deep learning pain-level detection deployment at UAE for patient-centric-pain management and diagnosis support: framework and performance evaluation. Procedia Comput Sci 2023;220:339-347 [FREE Full text] [doi: 10.1016/j.procs.2023.03.044] [Medline: 37089761]
65. Vu MT, Beurton-Aimar M. Learning to focus on region-of-interests for pain intensity estimation. 2023 Presented at: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG); January 05-08, 2023; Waikoloa Beach, HI p. 1-6. [doi: 10.1109/fg57933.2023.10042583]
66. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv 2019;52(4):1-36. [doi: 10.1145/3343440]
67. Data for meta-analysis of pain assessment from facial images. Figshare. 2023. URL: https://figshare.com/articles/dataset/Data_for_Meta-Analysis_of_Pain_Assessment_from_Facial_Images/24531466/1 [accessed 2024-03-22]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
CV: computer vision
DF: deep feature
GF: geometrical feature
ICC: intraclass correlation coefficient
LDOR: log diagnostic odds ratio
MAE: mean absolute error
NRS: numerical rating scale
PCC: Pearson correlation coefficient
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis
PROSPERO: International Prospective Register of Systematic Reviews
PSPI: Prkachin and Solomon pain intensity
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
SROC: summary receiver operating characteristic
SVM: support vector machine
VAS: visual analog scale

Edited by A Mavragani; submitted 26.07.23; peer-reviewed by M Arab-Zozani, M Zhang; comments to author 18.09.23; revised version received 08.10.23; accepted 28.02.24; published 12.04.24.

Please cite as:
Huo J, Yu Y, Lin W, Hu A, Wu C
Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e51250
URL: https://www.jmir.org/2024/1/e51250
doi:10.2196/51250
PMID:38607660

©Jian Huo, Yan Yu, Wei Lin, Anmin Hu, Chaoran Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patients' Representations of Perceived Distance and Proximity to Telehealth in France: Qualitative Study

Amélie Loriot[1*], MSc; Fabrice Larceneux[1*], PhD; Valérie Guillard[1*], PhD; Jean-Philippe Bertocchio[2,3*], MD, PhD

[1]Paris Dauphine–PSL (Paris Sciences & Lettres) University, Paris, France

[2]Service Thyroïde – Tumeurs Endocrines, Hôpital de la Pitié-Salpêtrière, Assistance Publique – Hôpitaux de Paris, Paris, France

[3]SKEZI, Annecy, France

[*]all authors contributed equally

Corresponding Author:
Amélie Loriot, MSc
Paris Dauphine–PSL (Paris Sciences & Lettres) University
Place du Maréchal de Lattre de Tassigny
Paris, 75116
France
Phone: 33 144054405
Email: amelie.loriot@dauphine.psl.eu

Abstract

Background: In the last 2 decades, new technologies have emerged in health care. The COVID-19 pandemic further accelerated the adoption of technology by both health care professionals and patients. These technologies create remote care practices that bring several benefits to the health care system: easier access to care, improved communication with physicians, and greater continuity of care. However, disparities in the acceptance and use of telehealth tools still exist among patients. These tools also disrupt conventional medical practices and prompt a new reassessment of the perceptions of distance and proximity as physical (ie, time and space dimensions) and nonphysical (ie, behavioral dimensions) concepts. The reasons why patients do or do not adopt telehealth tools for their care and therefore their perspectives on telehealth remain unanswered questions.

Objective: We explored the barriers as well as the motivations for patients to adopt telehealth tools. We specifically focused on the social representations of telehealth to establish a comprehensive conceptual framework to get a better understanding of how telehealth is perceived by patients.

Methods: This study uses a qualitative design through in-depth individual interviews. Participants were recruited using a convenience sampling method with balanced consideration of gender, age, location (urban/rural), and socioeconomic background. After collecting informed consent, interviews were transcribed and analyzed using the thematic analysis methodology.

Results: We conducted 14 interviews, with which data saturation was reached. The 2 main opposed dimensions, perceived proximity and distance, emerged as an essential structure for understanding the social representations of telehealth. A logic of engagement versus hostility emerged as the main tension in adopting telehealth, almost ideological. Interestingly, practical issues emerged regarding the adoption of telehealth: A logic of integration was opposed to a logic of constraints. Altogether, those dimensions enabled us to conceptualize a semiotic square, providing 4 categories with a coherent body of social representations. Due to the dynamic nature of these representations, we proposed 2 "paths" through which adherence to telehealth may improve.

Conclusions: Our semiotic square illustrating patients' adherence to telehealth differentiates socially beneficial versus socially dangerous considerations and pragmatic from ideological postures. It shows how crucial it is to consider perceived distance and proximity to better understand barriers and motivations to adopting telehealth. These representations can also be considered as leverage that could be modified to encourage the step-by-step adhesion process. Even if reducing the perceived temporal distance to in-person meeting and enhancing the perceived proximity of access to care may be seen as efficient ways to adopt telehealth tools, telehealth can also be perceived as a care practice that threatens the patient-physician relationship. The patient-oriented perceived value turns out to be critical in the future development of and adherence to telehealth tools.

(J Med Internet Res 2024;26:e45702) doi:10.2196/45702

KEYWORDS

telehealth; teleconsultation; social representations; perceived proximity; semiotic square

XSL•FO
RenderX

Introduction

Telehealth, a Subset of eHealth Still Ongoing

Many technologies have been developed in eHealth in recent years. Defined as the "use of information and communication technologies in support of health and health-related fields, including health care services, health monitoring, health literature, and health education, knowledge and research" [1], eHealth covers a wide range of practices. First, mobile apps and connected devices are referred to as mobile health (mHealth). Second, telehealth (ie, the practice of medicine using information and communication technologies) covers 5 practices: teleconsultation, teleexpertise, medical regulation, remote medical monitoring, and remote medical assistance.

Recent research focusing on remote care has indicated some confusion regarding the wording used to refer to health-related technologies [2]. For instance, the terms "telemedicine" and "telehealth" are often used interchangeably [3]. However, some researchers highlight a difference between these 2 concepts. Whereas telemedicine is limited to remote clinical services, telehealth is broader and refers to remote clinical services as well as remote nonclinical services, like administrative meetings [4]. Thus, telehealth has been defined as "the use of electronic information and telecommunication technologies to support and promote long-distance clinical health care, patient and professional health-related education, public health and health administration" [5].

eHealth is expected to lead to significant changes in the delivery of care and medical practices [6]. Because of (1) disparities in access to health care, (2) the aging population, and (3) budget constraints limiting public policies, the development of eHealth devices can be seen as a solution to the future challenges faced by the health system in many high-income countries [7]. Before the COVID-19 crisis, there were significant disparities in the use of eHealth tools between European countries [8]. A global shift occurred during the pandemic: The use of many eHealth tools became necessary, democratizing their use in terms of communication, monitoring, or care delivery, and the use of technology to provide health services has accelerated [9,10]. Telehealth may now concern everyone.

Benefits and Barriers of Telehealth

The practice of telehealth presents many benefits for patients, including (1) better access to health care services, in particular in isolated regions like rural areas; (2) improved continuity of care; (3) increased availability of health information [11]; and (4) empowerment of patients [12]. As such, telehealth is supposed to increase efficiency and quality of care [10] and favors patient-centered care by enabling better communication between patients and health care professionals [13].

However, researchers have pointed out that many barriers exist that limit health equity for all patients. Significant disparities remain regarding the access to, adherence to, and use of telehealth tools [14,15]. In particular, little is known about the role of digital health literacy [13,15,16] (ie, "the ability to search for, find, understand and evaluate health information from electronic sources and to apply knowledge acquired to solve a health problem" [17]). Among individuals in rural areas, low levels of education are associated with lower use of digital health tools [18]. Some scholars argue that online interactions are impersonal and dangerous because of the lack of a physical examination [10] and that telehealth may threaten the quality of the relationship between physicians and patients [19].

Studies among health care professionals have also shown a reluctance to adopt these technologies because of a fear of "dehumanization" by virtualizing patients and care [20]. This feeling of dehumanization of care could explain negative attitudes toward telehealth [21].

Patients tend to attribute significant importance to health care professionals' physical and emotional presence [22] and direct interactions with them [7]. However, the digitalization of health is transforming these relationships [23]: Telehealth disrupts medical practices and reduces physical interactions between patients and physicians. However, it leads to reconsidering notions of distance and proximity [23], including physical and nonphysical dimensions (ie, cognitive or relational aspects that are perceived by individuals) [24-26]. Physical proximity and perceived proximity are not necessarily aligned. Indeed, individuals can feel themselves close to an element that is physically far but also to perceive it far when it is physically close [26]. Perceived proximity has a cognitive dimension that refers to "a mental assessment of how distant someone else seems" and an affective component, since these representations are subject to emotions rather than rational thought [26,27]. In health care, Talbot et al [28] investigated the perceptions French physicians may have about telehealth using the conceptual framework of proximity of Boschma [25] that includes the following 5 dimensions of proximity: cognitive, organizational, social, institutional, and geographical. However, how patients react to these changes in care delivery and the representations of these practices remain unanswered questions. Therefore, exploring patient's representations of telehealth is important to better understand psychological mechanisms underlying the adherence to telehealth. The theory of social representations is fruitful in overcoming this limitation.

The Social Representations Theory

The theoretical background of social representations provides a framework for understanding how new concepts become common knowledge. Defined as a collective elaboration "of a social object by the community for the purpose of behaving and communicating" [29], social representations consist of a system of values, ideas, and practices that enable individuals to orient themselves in their material and social world as well as to master it and provide a code for social exchange [30]. Therefore, social representations provide people with a common frame of communication that is built in everyday interactions. More precisely, a social representation corresponds to thoughts and feelings being expressed in verbal and overt behavior of actors that constitutes an object for a social group [31].

Although social representations are commonly shared, some may be more polemical, reflecting oppositions between social groups in society [32]. In addition, social representations have a dynamic nature across and within social groups of people, and societal practices, communication, and the process of

XSL•FO
RenderX

knowledge are strongly connected, particularly in the health field, which has been one of the leading research areas for this theory [33-35].

Interestingly, social representations constitute a structure explaining behaviors that result not only from an individual cognitive process but also from social and cultural representations and that are shared collectively [36,37]. Social representations have been shown to be a significant indicator of attitudes [38,39]. However, social representations of patients have never been studied in the context of telehealth specifically. A qualitative study is well suited to understand these representations. The objective of our qualitative research was to establish a comprehensive conceptual framework to gain a better understanding of how telehealth influences perceived proximity or distance for patients and therefore, to better apprehend their barriers as well as their motivations to adopting telehealth tools.

Methods

Study Design

A qualitative study was conducted with an interpretative approach to explore patients' representations of telehealth and their perception of proximity toward it. We adopted an inductive, constructivist perspective, assuming that people construct their life-worlds through their representations and interpretations of telehealth as a social fact to which they attribute specific terms and meanings.

Setting and Sample

Qualitative in-depth individual interviews were set up using a semistructured thematic interview guide. Convenience sampling was used to recruit participants. Variation sampling was sought [40] with consideration of gender, age, location (urban/rural), and socioeconomic background (Multimedia Appendix 1). We used the saturation criterion to stop recruitment. This criterion is the point at which gathering more data about a theoretical construct reveals no new properties nor yields any further theoretical insights [41]. This saturation point is usually reached with 9 to 17 interviews [42].

Data Collection

After obtaining informed consent, patients were contacted, and an appointment for an interview was set. Interviews lasted from 45 minutes to 75 minutes and were performed directly inside the family home or conducted through the digital platform Microsoft Teams because of the geographical distance between the researcher and the participant. The study took place in May 2022. A total of 14 interviews were gathered: 8 participants were female, 6 were male, and their mean age was 52 (range 23-83) years. Of the interviews, 11 interviews were run face to face, and 4 were online.

The interview guide explored various aspects of how health and telehealth are perceived; including defining what constitutes perceived good health; understanding respondents' relationship with their own health; examining how they seek health-related information; discussing challenges in accessing care as related to geographical, temporal, and perceived distances; and evaluating respondents' overall and specific relationships with technology within the context of health care. This comprehensive approach aimed to gain insights into how individuals perceive telehealth and their level of engagement with it.

During each interview, we wrote down our impressions that could possibly impact the interpretation of results. Interviews were digitally audio-recorded with permission, and verbatims were transcribed.

Ethical Considerations

At the beginning of each interview, potential participants were given comprehensive information about the context, objectives, and methods of the study. The interviewees were informed that they could withdraw from this study at any time. After allowing enough time for any questions or clarification they may have required, all the participants gave their informed consent. The study design was reviewed and approved by the Research Ethics Committee of Paris Dauphine–PSL (Paris Sciences & Lettres) University (20231128/01). Additionally, following national legislation, data were pseudonymized during the transcription process in a way that no participant could be directly identified: A number was assigned to each participant with no record of any directly identifying data. Participants received no compensation for participating in this research.

Data Analysis

First, we conducted a vertical analysis and read the transcripts to get an impression of the whole data set. Second, transcribed data were analyzed using a horizontal thematic analysis to develop a narrative of the findings through a categorical approach using qualitative software (NVivo Version 12). We followed the grounded theory approach to code verbatim [43]: Each transcript was coded inductively by manually marking central key words that could represent a code. The codes were then grouped under themes that emerged through the analysis process. Finally, we categorized the data by collapsing codes that conveyed similar meanings. Multimedia Appendix 2 presents an example of our analysis process.

After the first step of the analysis of social representations, which was to record all the dimensions that emerged from the participants, we used the semiotic square method to map semantic categories highlighting opposing and complementary concepts [44]. This structure enables the understanding of the tension among symbolic meanings and the elements by which meaning is being expressed [45]. The semiotic square has been often applied in consumer research [46] and specifically to explore consumers' relationships with technology ideology [47].

Results

Overview

First, a specific definition of telehealth emerged from the patient perspective. If researchers define telehealth broadly, the interview analysis revealed that telehealth is associated with teleconsultation for a large majority of patients and rarely with other practices. It concerns mainly remote care and is associated

XSL•FO
RenderX

with questions about the quality of interactions with the physician.

Second, the content analysis revealed 4 main types of social representations of telehealth: the expected opposition between engagement and hostility and a more subtle distinction between integration and constraint.

Representations of Proximity: the Logic of Engagement

Our analysis of interviews revealed the first category of very positive social representations related to telehealth that led to a logic of engagement and adherence to this practice. This commitment is based on the idea of optimization of health services. The strong proximity with its practice is explained by a feeling of comfort and a perception of convenience. Telehealth is considered an easy, practical tool. Participant 4 (P4) mentioned:

> *I found it practical and comfortable.*

Perceived practicality and convenience underline the actual benefits of adopting telehealth. Indeed, this practice enables a reduction of the perceived temporal distance to the consultation, leading to representations of efficiency and effectiveness (P13) on one hand and allowing reinforcement of access to care, which creates a feeling of personal usefulness (P8), on the other hand. Participant 13 (P13) mentioned:

> *Now that everything is overbooked in their appointments, (...), we are at about 15 days/3 weeks for getting any new appointment, both by phone or by Doctolib, in video, it is a little faster.*

In addition, participant 8 (P8) said:

> *It is so quick, it makes everyday life easier!*

From this perspective, the main issue behind social representations of proximity is related to an improvement of the functional proximity to telehealth.

Representations of Distance: the Logic of Hostility

At the opposite end to that of the first category, the second category of social representations follows a logic of hostility toward telehealth. It reveals a strong rejection of its development. Although adherence follows a view of functional proximity, rejection is explained by a lack of perceived relational proximity caused by telehealth. These representations of perceived distance reveal a profound fear of the dehumanization of medicine. Telehealth is seen as a dehumanizing practice that is destructive of human interaction by virtualizing both patients and care, as Kaplan [20] mentioned. This was confirmed by participant 6 (P6), who stated:

> *It kills the human contact, which is really important to me. I definitely prefer having the secretary over the phone to tell me there is an appointment in three weeks.*

The major component of this category is the perceived deterioration of the relationship with the physician. Great importance is given to the human dimension in care. However, the interviews revealed these representations are based on a feeling of detachment from the caregiver caused by telehealth. This emphasizes the impersonal nature of the relationship. Participant 2 (P2) stated:

> *We dematerialize everything. It brings detachment from the caregiver.*

From this perspective, the development of a relationship with perceived proximity and trust seems incompatible with distant and remote care. The virtual nature of this link is intrinsically considered as the opposite of human interaction. Participant 1 (P1) stated:

> *I do not like it. I like to see the person right in front of me.*

Here, social representations of telehealth found an increase of perceived distance between the patient and physician. The perception of actual proximity to the physician tends to disappear with telehealth, which reinforces emotional and affective distance [48]. These representations finally highlight the fact that telehealth cannot replace an in-person consultation. For instance, participant 5 (P5) stated:

> *I would not make [a remote physician] my referring physician. There need to be a close relationship with him. I must be able to give him my trust. I am not sure that I will always have the same doctor when using teleconsultation.*

Altogether, these depictions of distance nurture the perception that telehealth has a detrimental or potentially harmful impact on society, as it undermines the interpersonal nature of care.

In addition to these 2 opposite categories of representations, proximity versus distance, more nuanced types of social representations also emerged within the verbatims. We labelled them “nondistance” and “nonproximity” representations.

Representations of Nondistance: the Logic of Integration

The third category of social representations reflects “nondistance” to telehealth, as these representations are related neither to total adherence nor to rejection but rather follow a logic of integration: Participants highlighted the actual possibility to choose to use (or not) telehealth tools. Representations do not reflect a full engagement with this practice but rather a nonrejection of telehealth.

First, these representations of nondistance highlight the functional aspects of this practice. In this context, developing a relational proximity with the physician was not judged as necessary. For instance, participant 4 (P4) stated:

> *I felt more like I was with a teleoperator than a physician. It felt like there was a script behind it, but why not, that is not necessarily a bad thing.*

This situation is not seen as a problem; the efficient and nonrelational aspect of the consultation is valued here. Thus, this representation shows a greater emphasis on the functional proximity rather than on the relational proximity [24].

The importance given to the functional aspects of telehealth was also revealed through the way specific health practices are elicited. For instance, telehealth was mainly seen as a backup

XSL•FO
RenderX

or emergency solution, leading to occasional use according to the situation. Participant 5 (P5) stated:

> *It can be a first step to detect an emergency. For example, if you cannot get a doctor during the weekend, we have remote visits (...) So, to me, it is an emergency solution.*

Because it is convenient, patients do not reject telehealth, especially when there is no need to be seen in person, for example for a prescription renewal, as suggested by participant 4 (P4):

> *It depends on what you are looking for in the consultation. If it is for a medication renewal, yes, I would recommend it.*

Thus, these representations of nondistance do not refer to hostility nor engagement toward telehealth but rather to tolerance. The practice is adopted but not entirely accepted. Indeed, the use of telehealth should remain occasional. Participant 11 (P11) said:

> *If I were starting using a teleconsultation system, I would say to myself 'no more than three times in a row.' The fourth time, you still have to go, once every two years for a check-up, I would tend to say that.*

Tolerance comes also with some reluctance about the reliability of this practice. Telehealth was perceived as less reliable than a physical consultation because there is no physical contact and no auscultation, which seems to lead to mistrust, as suggested by participants 10 (P10) and 14 (P14).

> *Auscultation is one of the first things you learn in medicine, like touching the patient. Try to get an auscultation from a machine, to put its hands on the belly.* [P10]

> *When I had my operation, I had a consultation with the anesthesiologist by teleconsultation. It was silly, he told me to pull my tongue out (...) No, for me this is ridiculous!* [P14]

Overall, social representations related to nondistance reveal nonrejection of telehealth under conditions of efficiency and reliability. The choice of using telehealth tools is made under specific circumstances and leads to occasional use, based on high value placed on simplicity and functional aspects.

Representations of Nonproximity: the Logic of Constraint

Within the fourth category, social representations are related to "nonproximity," a label that reflects a logic of constraint. Whereas representations of proximity highlight engagement and active behavior toward telehealth, representations of nonproximity depict situations of the use of telehealth when there is no other choice, as participant 10 (P10) mentioned:

> *Is telehealth a good thing? Like everyone else, I use it because I am left with no alternative option.*

Patients come to telehealth whenever they have no or few alternatives, considering telehealth as a last option, such as during a lockdown for example, as explained by participant 3 (P3):

> *If I had to use it, it would really be out of obligation, like during a lockdown, and because I do not have the possibility to move around.*

In this perspective, telehealth tools are not really accepted and should remain a second option to physical in-person consultations, mainly because telehealth requires digital literacy. Participant 7 (P7) explained:

> *For the elderly, it is a problem! I have to schedule their appointments from my own mobile phone because they do not have access to the internet.*

Thus, like the representations of perceived distance, the representations of nonproximity are also mostly negative. However, they do not reflect a total rejection of the practice of telehealth but rather a nonadherence as patients come to it when they have no other option.

Finally, our qualitative analysis allowed us to structure a semiotic square (Figure 1) with 2 main categories of social representations of telehealth (ie, perceived proximity and perceived distance) as well as subsequent tensions in the discourse. The negation of these 2 terms forms 2 other categories illustrating 4 distinctive classes of meanings highlighting nuanced representations of perceived proximity and distance to telehealth and the opposite and complementary relationships [49]. The main components of the 4 categories are summarized in Figure 1. Interestingly, 2 additional analyses of the semiotic square improve our vision of social representations of telehealth, one based on a vertical reading and the other based on a horizontal reading.

XSL•FO
RenderX

Figure 1. Overview of the 4 categories of social representations of telehealth, including the corresponding logic and main drivers, that resulted from 14 qualitative interviews run with patients in May 2022, with a major horizontal dimension (proximity versus distance) and its opposites (vertical).

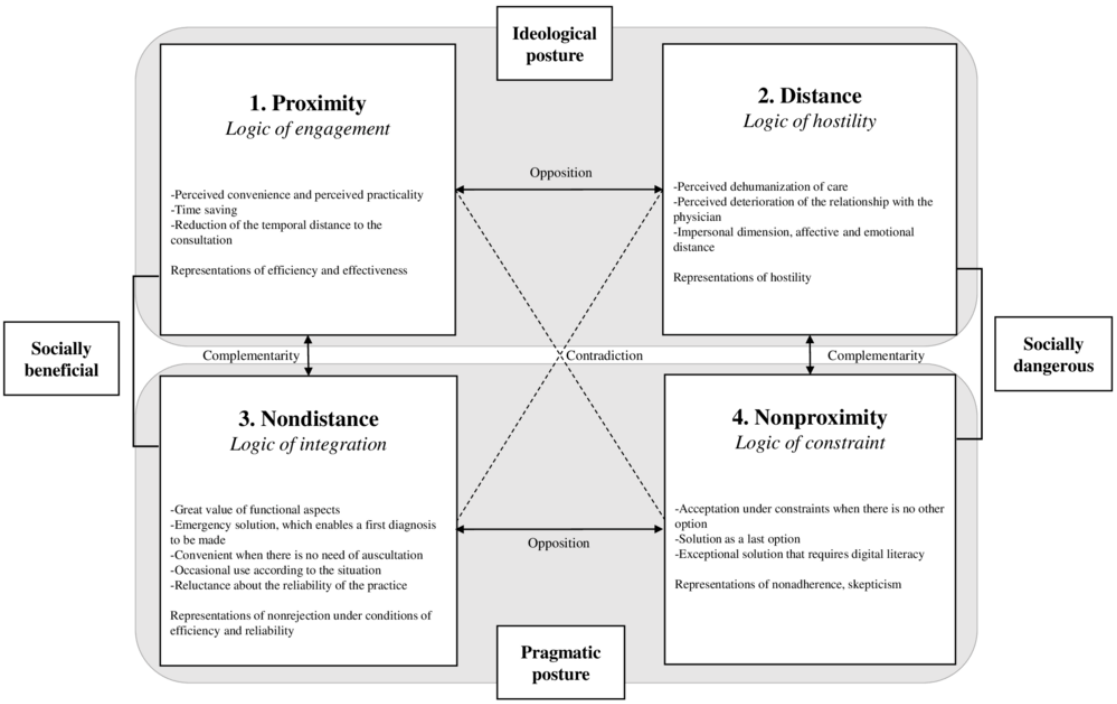

Telehealth: Socially Beneficial Versus Socially Dangerous

From a vertical reading of the relationships between the categories, there are 2 structuring representations of telehealth. In the left part of the semiotic square (Figure 1), the complementary relationship, linking proximity and nondistance, refers to favorable representations as well as to discourse encouraging the development of telehealth. These tools are perceived as socially beneficial for all stakeholders, but there is room for improvement in generalizing their use.

Within these favorable representations of telehealth, adherence and nonrejection are based on 2 main drivers. First, trust in the physician is crucial as he or she is considered a legitimate expert, as suggested by participant 4 (P4):

> *I feel that doctors are experts (...), I trust them entirely because to me they seem to be experts.*

Consequently, positive representations of telehealth seem to be linked to the perceived relational proximity with the health care professional. Second, these representations stem from familiarity with the tool. Being familiar with the term "telehealth" and knowing what it means generate a feeling of closeness toward it. Participant 5 (P5) stated:

> *I heard [about telehealth] because in my profession—I work with pharmacies—we talk about it.*

In the right part of the square is the complementary relationship combining perceived distance and nonproximity. This underlines representations of hostility and skepticism toward telehealth, which are considered socially harmful or even dangerous for society. Rejection and nonadherence seem to be explained mainly by insufficient digital literacy as well as difficulties accessing the Internet and telehealth tools, as suggested by participant 9 (P9):

> *No, I do not use the Internet at all! (...) There are surely many things to do but I do not know how to do them...*

This revealed a substantial cognitive distance to telehealth and ultimately making care practices feel more complex. The ancestral role of auscultation in medical consultation and the importance given to touching patients are noted, showing that the lack of perceived physical proximity between the patient and physician tends to reinforce the psychological distance toward telehealth and ultimately the rejection of its practice. Participant 4 (P4) said:

> *The ability itself of performing an actual auscultation by touching people and listening to them using a stethoscope is being lost at the expense of the care to improve the development of technology.*

Telehealth: Ideological Versus Pragmatic Postures

A horizontal reading highlights the similarity of the logics of engagement and hostility, both based on ideological postures: pros and cons of the practice of telehealth depending on whether it seems to belong to the "good" versus "bad" for the society. More efficiencies appear to be pros, and less of a human relationship appears to be a con. Conversely, the logics of integration and constrain reflect pragmatic postures: how to deal with the tool and on what occasion. Sometimes, it appears to be accepted because it is convenient and adapted to specific situations, sometimes because there is no other choice. Interestingly, ideological postures tend to separate opposite groups, while the pragmatic views tend to rebuild a link between the nondistance and nonproximity groups. These nuanced, more balanced perceptions invite us to think about the practical

XSL•FO
RenderX

implications, elaborating "paths" of social representations to drive patients toward less rejection of and more adherence to telehealth.

Discussion

Main Findings

Using qualitative methods, our findings suggest a new conceptual framework to apprehend telehealth from patients' perspectives based on 4 categories of social representations. First, perceived proximity was associated with social representations reflecting the idea that telehealth is intrinsically an efficient, practical, and effective solution. This logic of engagement is in line with a strong belief in progress and technological tools to face the challenges of the health care system, namely the issue of access to care. On the opposite side, social representations were more related to a feeling of distance from telehealth, enforcing an unfavorable attitude and leading to a rejection of these tools. This logic of hostility is mainly anchored in a fear of dehumanization of society. Telehealth is blamed for compromising the quality of the relations and for accelerating the loss of human contact between patients and physicians. This perceived distance from telehealth highlights a situation of exclusion, especially for patients who do not have access to digital technology or who do not have sufficient digital literacy. Aside from these 2 categories, 2 more nuanced types of representations emerged. First, from a logic of integration, social representations revealed that telehealth is appealing but showed worries and fears about its reliability. This practice can be conditionally accepted according to a situational approach. Second, a logic of constraint reflected social representations based on skepticism but leading to acceptance when there are no alternatives.

From a theoretical point of view, our results, based on a semiotic square, bring new elements to the literature of perceived proximity. We have shown that telehealth leads to reconsidering proximity through several dimensions. Although not diminishing the geographical or physical gap between the patient and the health care provider, technological tools, such as a teleconsultation from home, can enhance accessibility to health care. The relational dimension of proximity, already identified by Boschma [25], seems to also be impacted by telehealth. Indeed, many social representations have shown that this perception of proximity with the caregiver is reduced by telehealth and revealed a fear of dehumanization in the relationship. In addition, we showed that perceived functional proximity to telehealth leads to increased adherence and a favorable attitude to its development, which should encourage policymakers to strengthen this aspect in communication strategies for telehealth. These findings also constitute a societal contribution. In addition, this research has revealed 2 major oppositions embedded in the social representations. The first one consists of "good," or a socially beneficial position, versus "bad," or a socially dangerous position. The second one highlights the posture, rather "ideological" or "pragmatic," leading to contributions to public policy aiming to foster adherence to eHealth tools.

Building a semiotic square also revealed potential changes in people's representations of telehealth and thus the potential to contribute to change attitudes toward these tools. They may be adapted to patients' concerns and aspects that patients value in the practice of consultation. Our qualitative material brings insight to how these representations can be obstacles to the adoption of telehealth, as well as elements that can foster adherence. We propose considering paths through which patients' representations could evolve. Mobilizing the social representations along these paths could first alleviate the perception of distance to the health care professional then enable the perceived proximity to telehealth. Our analysis emphasized some risks in how telehealth is implemented. If telehealth is developed without considering representations expressing reluctancy, individuals who are subjected to the use of telehealth may remain hostile to its development, may gradually feel a distance to it, and may finally totally reject this practice (coming from nonproximity to distance). To avoid such a vicious circle, 2 paths (Figure 2) may create an increased feeling of proximity to telehealth.

Figure 2. Illustration of 2 paths to create proximity with telehealth valuing either functional (path 1) or relational (path 2) aspects.

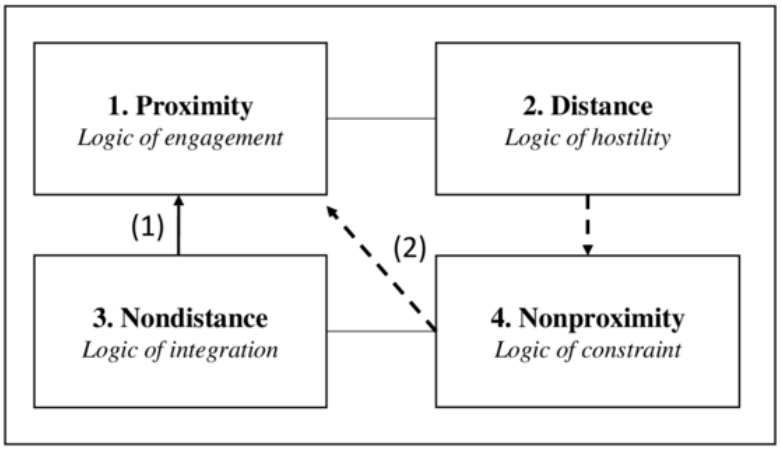

The first path consists of transforming representations related to a perceived nondistance into a perceived proximity to telehealth. This pathway adopts a functional approach to consultation. The challenge is to dispel fears about the technological feasibility of using digital health tools to eliminate skepticism and reinforce favorable representations. It would

XSL•FO
RenderX

then be necessary to reassure patients about the importance of any human contact during medical consultations. Highlighting the regular and immediate exchanges with physicians that telehealth allows would be perceived as helpful. Developing remote auscultation solutions and increasing communication about them by highlighting the accuracy and reliability of these technologies would help to alleviate these concerns and encourage adherence to these tools. Finally, reinforcing the benefits in terms of efficiency, time optimization, and practicality would contribute to (1) reducing the perceived temporal distance of access to care and (2) increasing the perceived functional proximity to telehealth.

The second path consists of (1) transforming representations related to a perceived distance into a feeling of nonproximity and subsequently (2) fostering the perceived proximity to telehealth. This path is mainly aimed at individuals who attach great importance to the relational and human dimension of care. The first challenge would be to strengthen trust in the health care system because representations and attitudes toward telehealth are intrinsically linked to the relationship patients develop with the health care system and physicians. It is also necessary to improve access to digital technology to reduce the cognitive distance and to increase their perception of proximity. Finally, highlighting and communicating about the strengthening of relational and affective proximity, allowed by telehealth when facilitating contact between patients and physicians, could lead to favorable representations and attitudes. Therefore, conceiving a system of medical support with a health care professional in telehealth booths could be an effective solution.

Limitations and Research Avenues

This study has some limitations. First, our sample did not include patients with a broad range of diseases: Very few of them had chronic diseases. Due to the sample size, we could not cover all medical specialties: For instance, ophthalmology and the need for emergency surgery may bring specific representations of telehealth for patients. It could also be interesting to interview people from other rural areas known as “medical deserts” (ie, regions with inadequate access to health care). In addition, we interviewed patients who do not practice as health care professionals. To broaden our research findings, we could incorporate additional insights by examining the perceptions of telehealth among other groups, particularly caregivers.

Conclusion

The development of telehealth tools leads to new challenges in medical practice. The social representations telehealth brings go beyond the perception of proximity and distance, are multifaceted, and include postures and attitudes. The social representations revealed by the semiotic square on perceived proximity to telehealth underscore the importance of designing health care strategies based on a patient-centric approach in the implementation of digital health tools.

Acknowledgments

The authors would like to thank all the participants in this study.

Data Availability

The data sets generated and analyzed during this study are available upon reasonable request from the corresponding author.

Authors' Contributions

This study was carried out by all authors working collaboratively. AL initiated the proposal. AL and FL conceived the study, collected the data, and performed the qualitative analysis. AL wrote the first draft of the manuscript. FL, VG, and JPB participated in data interpretation and revised the manuscript. All authors approved the final version of the manuscript.

Conflicts of Interest

JPB is a physician and works at SKEZI, a company that develops digital tools in health. The remaining authors have no conflicts of interest to declare.

Multimedia Appendix 1
Characteristics of patients.
[DOCX File , 13 KB - jmir_v26i1e45702_app1.docx]

Multimedia Appendix 2
Example of the qualitative analysis process used in this research.
[DOCX File , 14 KB - jmir_v26i1e45702_app2.docx]

References

1. 58th World Health assembly report. World Health Organization. 2005. URL: https://apps.who.int/gb/ebwha/pdf_files/WHA58-REC1/english/A58_2005_REC1-en.pdf [accessed 2024-04-12]

XSL•FO
RenderX

2. Queirós A, Alvarelhão J, Cerqueira M, Silva A, Santos M, Pacheco Rocha N. Remote care technology: a systematic review of reviews and meta-analyses. Technologies 2018 Feb 10;6(1):22. [doi: 10.3390/technologies6010022]
3. Fatehi F, Wootton R. Telemedicine, telehealth or e-health? A bibliometric analysis of the trends in the use of these terms. J Telemed Telecare 2012 Dec 01;18(8):460-464. [doi: 10.1258/jtt.2012.gth108] [Medline: 23209265]
4. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: A rapid systematic review. Digit Health 2021 Apr 19;7:20552076211009396 [FREE Full text] [doi: 10.1177/20552076211009396] [Medline: 33959378]
5. Telemedicine and Telehealth. The Office of the National Coordinator for Health Information Technology (ONC). URL: https://www.healthit.gov/topic/health-it-health-care-settings/telemedicine-and-telehealth [accessed 2024-04-12]
6. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017 Sep;3:38-38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
7. Lindberg J, Bhatt R, Ferm A. Older people and rural eHealth: perceptions of caring relations and their effects on engagement in digital primary health care. Scand J Caring Sci 2021 Dec 14;35(4):1322-1331 [FREE Full text] [doi: 10.1111/scs.12953] [Medline: 33448031]
8. From innovation to implementation: eHealth in the WHO European Region. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/326317 [accessed 2024-04-12]
9. Wang W, Sun L, Liu T, Lai T. The use of E-health during the COVID-19 pandemic: a case study in China's Hubei province. Health Sociol Rev 2022 Nov 23;31(3):215-231 [FREE Full text] [doi: 10.1080/14461242.2021.1941184] [Medline: 34161186]
10. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221 [FREE Full text] [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
11. Gagnon M, Duplantie J, Fortin J, Landry R. Implementing telehealth to support medical practice in rural/remote regions: what are the conditions for success? Implement Sci 2006 Aug 24;1(1):18 [FREE Full text] [doi: 10.1186/1748-5908-1-18] [Medline: 16930484]
12. Clemensen J. Patient empowerment and new citizen roles through telehealth technologies - The early stage. Proceedings of The Third International Conference on eHealth, Telemedicine, and Social Medicine 2011:114-119 [FREE Full text]
13. Goodridge D, Marciniuk D. Rural and remote care: Overcoming the challenges of distance. Chron Respir Dis 2016 May 21;13(2):192-203 [FREE Full text] [doi: 10.1177/1479972316633414] [Medline: 26902542]
14. Smith B, Magnani JW. New technologies, new disparities: The intersection of electronic health and digital health literacy. Int J Cardiol 2019 Oct 01;292:280-282 [FREE Full text] [doi: 10.1016/j.ijcard.2019.05.066] [Medline: 31171391]
15. Reiners F, Sturm J, Bouw LJ, Wouters EJ. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019 Feb 21;16(4):645 [FREE Full text] [doi: 10.3390/ijerph16040645] [Medline: 30795623]
16. Dunn P, Hazzard E. Technology approaches to digital health literacy. Int J Cardiol 2019 Oct 15;293:294-296. [doi: 10.1016/j.ijcard.2019.06.039] [Medline: 31350037]
17. First meeting of WHO GCM/NCD Working Group on health education and health literacy for noncommunicable diseases. Geneva Global Health Hub. 2017 Feb 17. URL: https://g2h2.org/posts/event/first-meeting-of-who-gcmncd-working-group-on-health-education-and-health-literacy-for-noncommunicable-diseases/ [accessed 2024-04-12]
18. Salemink K, Strijker D, Bosworth G. Rural development in the digital age: A systematic literature review on unequal ICT availability, adoption, and use in rural areas. Journal of Rural Studies 2017 Aug;54:360-371. [doi: 10.1016/j.jrurstud.2015.09.001]
19. Hjelm NM. Benefits and drawbacks of telemedicine. J Telemed Telecare 2005 Jun 24;11(2):60-70. [doi: 10.1258/1357633053499886] [Medline: 15829049]
20. Kaplan B. Revisiting health information technology ethical, legal, and social issues and evaluation: telehealth/telemedicine and COVID-19. Int J Med Inform 2020 Nov;143:104239 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104239] [Medline: 33152653]
21. Huryk L. Factors influencing nurses' attitudes towards healthcare information technology. J Nurs Manag 2010 Jul;18(5):606-612. [doi: 10.1111/j.1365-2834.2010.01084.x] [Medline: 20636510]
22. Rørtveit K, Sætre Hansen B, Leiknes I, Joa I, Testad I, Severinsson E. Patients' experiences of trust in the patient-nurse relationship—a systematic review of qualitative studies. Open Journal of Nursing 2015;05(03):195-209 [FREE Full text] [doi: 10.4236/ojn.2015.53024]
23. Oudshoorn N. Physical and digital proximity: emerging ways of health care in face-to-face and telemonitoring of heart-failure patients. Sociol Health Illn 2009 Apr;31(3):390-405 [FREE Full text] [doi: 10.1111/j.1467-9566.2008.01141.x] [Medline: 19144084]
24. Torre A, Rallet A. Proximity and localization. Regional Studies 2005 Feb;39(1):47-59. [doi: 10.1080/0034340052000320842]
25. Boschma R. Proximity and innovation: a critical assessment. Regional Studies 2005 Feb;39(1):61-74. [doi: 10.1080/0034340052000320887]

XSL•FO
RenderX

26. Wilson JM, Boyer O'Leary M, Metiu A, Jett QR. Perceived proximity in virtual work: explaining the paradox of far-but-close. Organization Studies 2008 May 20;29(7):979-1002. [doi: 10.1177/0170840607083105]
27. van Zoonen W, Sivunen A, Rice RE, Treem JW. Organizational information and communication technologies and their influence on communication visibility and perceived proximity. International Journal of Business Communication 2021 Oct 06;60(4):1267-1289. [doi: 10.1177/23294884211050068]
28. Talbot D, Charreire Petit S, Pokrovsky A. La proximité comme perception de la distance. Le cas de la télémédecine. Rev. Fr. Gest 2020 Nov 11;46(289):51-74. [doi: 10.3166/rfg.2020.00439]
29. Moscovici S. Attitudes and opinions. Annu Rev Psychol 1963 Jan;14(1):231-260. [doi: 10.1146/annurev.ps.14.020163.001311] [Medline: 13936119]
30. Moscovici S. Foreword. In: Herzlich C, editor. Health and illness A social psychological analysis. London, England: Academic Press; 1973:IX-XIV.
31. Wagner W, Duveen G, Farr R, Jovchelovitch S, Lorenzi‐Cioldi F, Marková I, et al. Theory and method of social representations. Asian J of Social Psycho 2002 Dec 18;2(1):95-125. [doi: 10.1111/1467-839x.00028]
32. Manuti A, Mininni G. Social representations of medically assisted fecundation a study on the discursive construction of “media texts”. Psychology 2010;01(05):337-348 [FREE Full text] [doi: 10.4236/psych.2010.15043]
33. de Rosa AS, Wagner W. The social representations of mental illness in children and adults. In: Doise W, Moscovici S, editors. Current issues in European Social Psychology. Cambridge, United Kingdom: Cambridge University Press; 1987:47-138.
34. de Rosa AS, Bocci E. Resisting cognitive polyphasia in social representations of madness. In: de Rosa AS, editor. Social Representations in the 'Social Arena'. Abingdon, United Kingdom: Routledge; 2012:310.
35. de Rosa AS, Mannarini T. The “Invisible Other”: social representations of COVID-19 pandemic in media and institutional discourse. Papers on Social Representations 2020;29(2):1-35 [FREE Full text]
36. Moliner P, Tafani E. Attitudes and social representations: a theoretical and experimental approach. European Journal of Social Psychology 1998 Dec 04;27(6):687-702. [doi: 10.1002/(sici)1099-0992(199711/12)27:6<687::aid-ejsp839>3.0.co;2-7]
37. Bidjari AF. Attitude and social representation. Procedia - Social and Behavioral Sciences 2011;30:1593-1597. [doi: 10.1016/j.sbspro.2011.10.309]
38. Monterrubio JC, Andriotis K. Social representations and community attitudes towards spring breakers. Tourism Geographies 2014 Mar 03;16(2):288-302. [doi: 10.1080/14616688.2014.889208]
39. Schultes M, Kollmayer M, Mejeh M, Spiel C. Attitudes toward evaluation: An exploratory study of students' and stakeholders' social representations. Eval Program Plann 2018 Oct;70:44-50. [doi: 10.1016/j.evalprogplan.2018.06.002] [Medline: 29936398]
40. Patton MQ. Qualitative Research & Evaluation Methods. Thousand Oaks, CA: SAGE Publications, Inc; 2002.
41. Bryant A, Charmaz K. The SAGE Handbook of Grounded Theory. Thousand Oaks, CA: SAGE Publications Ltd; 2007.
42. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
43. Strauss A, Corbin JM. Grounded Theory in Practice. Thousand Oaks, CA: SAGE Publications, Inc; 1997.
44. Greimas AJ. Structural semantics: An attempt at a method. Lincoln, NE: University of Nebraska Press; 1983.
45. Signori P, Flint DJ. Revealing the unique blend of meanings in corporate identity: An application of the semiotic square. Journal of Marketing Theory and Practice 2019 Nov 20;28(1):26-42. [doi: 10.1080/10696679.2019.1662312]
46. Floch J. The contribution of structural semiotics to the design of a hypermarket. International Journal of Research in Marketing 1988 Jan;4(3):233-252. [doi: 10.1016/s0167-8116(88)80007-7]
47. Kozinets RV. Technology/ideology: how ideological fields influence consumers' technology narratives. Journal of Consumer Research 2008;34(6):865-881. [doi: 10.1086/523289]
48. Fiedler K. Construal level theory as an integrative framework for behavioral decision‐making research and consumer psychology. J Consum Psychol 2008 Jan 22;17(2):101-106. [doi: 10.1016/s1057-7408(07)70015-3]
49. Greimas AJ, Rastier F. The interaction of semiotic constraints. Yale French Studies 1968(41):86. [doi: 10.2307/2929667]

Abbreviations

mHealth: mobile health
PSL: Paris Sciences & Lettres

XSL•FO
RenderX

Edited by A Mavragani; submitted 13.01.23; peer-reviewed by S Tubeuf, H Yu; comments to author 13.03.23; revised version received 08.05.23; accepted 19.12.23; published 26.04.24.

Please cite as:
Loriot A, Larceneux F, Guillard V, Bertocchio JP
Patients' Representations of Perceived Distance and Proximity to Telehealth in France: Qualitative Study
J Med Internet Res 2024;26:e45702
URL: https://www.jmir.org/2024/1/e45702
doi:10.2196/45702
PMID:38669676

©Amélie Loriot, Fabrice Larceneux, Valérie Guillard, Jean-Philippe Bertocchio. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

IT-Related Barriers and Facilitators to the Implementation of a New European eHealth Solution, the Digital Survivorship Passport (SurPass Version 2.0): Semistructured Digital Survey

Ismay A E de Beijer[1], MSc; Selina R van den Oever[1], MSc; Eliana Charalambous[2,3], MD; Giorgio Cangioli[2], PhD; Julia Balaguer[4], MD, PhD; Edit Bardi[5,6], MD, PhD; Marie Alfes[7], PhD; Adela Cañete Nieto[4], MD, PhD; Marisa Correcher[8], MSc; Tiago Pinto da Costa[7], MSc; Alexander Degelsegger-Márquez[9], PhD; Vanessa Düster[10], MSc; Anna-Liesa Filbert[11], MSc; Desiree Grabow[11], PhD; Gerald Gredinger[9], PhD; Hannah Gsell[7], MSc; Riccardo Haupt[12], MD; Maria van Helvoirt[13], MD; Ruth Ladenstein[5,10,14], MBA, MD, PhD; Thorsten Langer[15], MD, PhD; Anja Laschkolnig[9], MPH; Monica Muraca[12], MD; Saskia M F Pluijm[1], PhD; Jelena Rascon[16,17], MD, PhD; Günter Schreier[18], PhD; Zuzana Tomášikova[7], MSc; Florian Trauner[9], MSc, MD; Justas Trinkūnas[16], PhD; Kathrin Trunner[9], DI, MSc; Anne Uyttebroeck[13], MD, PhD; Leontien C M Kremer[1,19,20], MD, PhD; Helena J H van der Pal[1*], MD, PhD; Catherine Chronaki[2*], MSc; PanCareSurPass Consortium[21]

[1]Princess Máxima Center for Pediatric Oncology, Utrecht, Netherlands
[2]Health Level Seven Europe, Brussels, Belgium
[3]Venizeleio General Hospital of Heraklion, Heraklion, Greece
[4]Hospital Universitario y Politécnico La Fe, Valencia, Spain
[5]St Anna Children's Hospital, Vienna, Austria
[6]Department of Paediatrics and Adolescent Medicine, Kepler University Hospital, Johannes Kepler University Linz, Linz, Austria
[7]Childhood Cancer International Europe, Vienna, Austria
[8]Instituto Investigación Sanitaria La Fe, Valencia, Spain
[9]Gesundheit Österreich GmbH, Vienna, Austria
[10]Department of Studies and Statistics for Integrated Research and Projects, St Anna Children's Cancer Research Institute, Vienna, Austria
[11]Division of Childhood Cancer Epidemiology, German Childhood Cancer Registry, Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany
[12]Diagnosi, Osservazione, Prevenzione dopo trattamento Oncologico Clinic, Department of Hematology/Oncology, Istituto di Ricovero e Cura a Carattere Scientifico Istituto Giannina Gaslini, Genova, Italy
[13]University Hospitals Leuven, Katholieke Universiteit Leuven, Leuven, Belgium
[14]Department of Paediatrics, Medical University of Vienna, Vienna, Austria
[15]Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[16]Vilnius University Hospital Santaros Klinikos, Vilnius, Lithuania
[17]Clinics for Children's Diseases, Institute of Clinical Medicine, Faculty of Medicine, Vilnius University, Vilnius, Lithuania
[18]Center for Health and Bioresources, Austrian Institute of Technology, Graz, Austria
[19]Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, Netherlands
[20]Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands
[21]See Acknowledgments
[*]these authors contributed equally

Corresponding Author:
Ismay A E de Beijer, MSc
Princess Máxima Center for Pediatric Oncology
Heidelberglaan 25
Utrecht, 3484 CS
Netherlands
Phone: 31 638960162
Email: i.a.e.debeijer-3@prinsesmaximacentrum.nl

XSL•FO
RenderX

Abstract

Background: To overcome knowledge gaps and optimize long-term follow-up (LTFU) care for childhood cancer survivors, the concept of the Survivorship Passport (SurPass) has been invented. Within the European PanCareSurPass project, the semiautomated and interoperable SurPass (version 2.0) will be optimized, implemented, and evaluated at 6 LTFU care centers representing 6 European countries and 3 distinct health system scenarios: (1) national electronic health information systems (EHISs) in Austria and Lithuania, (2) regional or local EHISs in Italy and Spain, and (3) cancer registries or hospital-based EHISs in Belgium and Germany.

Objective: We aimed to identify and describe barriers and facilitators for SurPass (version 2.0) implementation concerning semiautomation of data input, interoperability, data protection, privacy, and cybersecurity.

Methods: IT specialists from the 6 LTFU care centers participated in a semistructured digital survey focusing on IT-related barriers and facilitators to SurPass (version 2.0) implementation. We used the fit-viability model to assess the compatibility and feasibility of integrating SurPass into existing EHISs.

Results: In total, 13/20 (65%) invited IT specialists participated. The main barriers and facilitators in all 3 health system scenarios related to semiautomated data input and interoperability included unaligned EHIS infrastructure and the use of interoperability frameworks and international coding systems. The main barriers and facilitators related to data protection or privacy and cybersecurity included pseudonymization of personal health data and data retention. According to the fit-viability model, the first health system scenario provides the best fit for SurPass implementation, followed by the second and third scenarios.

Conclusions: This study provides essential insights into the information and IT-related influencing factors that need to be considered when implementing the SurPass (version 2.0) in clinical practice. We recommend the adoption of Health Level Seven Fast Healthcare Interoperability Resources and data security measures such as encryption, pseudonymization, and multifactor authentication to protect personal health data where applicable. In sum, this study offers practical insights into integrating digital health solutions into existing EHISs.

(J Med Internet Res 2024;26:e49910) doi:10.2196/49910

KEYWORDS

pediatric oncology; long-term follow up care; survivorship; cancer survivors; Survivorship Passport; SurPass, eHealth; information and technology

Introduction

More and more children and adolescents successfully survive cancer into adulthood due to improvements in childhood cancer treatment [1,2]. There are currently around 500,000 childhood cancer survivors in Europe [1-3], with around 8000-10,000 new survivors each year [4]. However, despite the increasing survival rates, childhood cancer survivors remain at risk of impaired quality of life and extensive morbidity and mortality due to disease relapse or late health complications caused by cancer treatments (late effects) [5-7]. Late effects can include physical as well as psychological and social conditions, ranging from subsequent neoplasms and cardiotoxicity to chronic pain and poor psychological well-being [4-16]. To improve or preserve the quality of life of childhood cancer survivors, long-term follow-up (LTFU) care focusing on late-effects surveillance and timely intervention is essential [17]. However, comprehensive LTFU programs are lacking in many pediatric cancer centers [18]. In particular, the coordination of LTFU care between health care providers (HCPs), care managers, and childhood cancer survivors, as well as the available knowledge about late effects and the transition from pediatric to adult health care services, often call for improvement [18-20].

Previous studies have highlighted the need for a treatment summary and care plan for childhood cancer survivors as part of successful LTFU care [21-23]. To increase knowledge among HCPs and childhood cancer survivors and optimize long-term survivorship care, the Survivorship Passport (SurPass) was developed by the Pan-European Network for Care of Survivors after Childhood and Adolescent Cancer (PanCare), a multidisciplinary and international association of professionals, childhood cancer survivors, and their families, aiming to reduce the impact of late health effects for childhood cancer survivors [24-26]. The SurPass summarizes cancer and treatment-related data of childhood cancer survivors and, thanks to built-in algorithms, suggests personalized follow-up recommendations based on evidence-based surveillance guidelines developed by the International Late Effects of Childhood Cancer Guideline Harmonization Group and consensus-based recommendations formulated within several PanCare projects (PanCareSurFup and PanCareFollowUp) [17,26]. In addition, the SurPass provides plain language information on late effects and self-care. All in all, the SurPass supports personalized follow-up care and can improve understanding of late effects among HCPs and childhood cancer survivors. Over the years, multiple versions of SurPass have been developed. At the Istituto Giannina Gaslini, Italy, the SurPass (version 1.2) was found to have an overall positive impact on survivors and their families [24]. Ultimately, the SurPass has the potential to be used throughout Europe and beyond to improve LTFU care and empower childhood cancer survivors to take charge of their own health [24].

Previously developed versions of the SurPass require manual entry of individual treatment data to be entered manually into

XSL•FO
RenderX

the SurPass database, making its use in daily clinical practice rather time-consuming [24]. As a result, SurPass is currently being upgraded to a semiautomated and interoperable version (SurPass version 2.0) as part of the European Horizon 2020-funded PanCareSurPass (PCSP) project [27]. Like previous versions of SurPass (version 2.0) will generate a survivor-specific treatment summary and survivorship care plan (SCP) using algorithms that link treatment data with available follow-up recommendations. Unlike previous versions, SurPass (version 2.0; hereafter referred to as SurPass) will facilitate semiautomated data entry from electronic health information systems (EHISs) and integration of SurPass into national or regional electronic health records (EHRs). A high level of interoperability and data protection or security are essential to achieve semiautomated data transfer. The development and harmonization of interfaces is required to enable data exchange between systems and the storage of information in different systems (eg, hospital systems, clinical trials, and cancer registries). Interoperability between EHISs and EHRs and the SurPass platform is facilitated by the Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) interoperability standard [28]. HL7 FHIR supports the exchange of data between health care software systems and combines the hallmarks of the established HL7 (version 2 and 3) and Clinical Document Architecture (CDA) standards while leveraging current web standards such as XML, JSON, HTTP, and the OAuth (Open Authorization Standard) [29]. In addition, HL7's international patient summary (IPS) standard specifies an EHR extract containing essential health information intended for use in cross-border care scenarios. IPS models support continuity of care for patients and coordination of care across health systems [30]. Specifically, the treatment summary variables in SurPass are linked according to the IPS. Finally, SurPass has been certified as a Class 1 medical device and must guarantee data security in terms of availability, confidentiality, and data integrity by the General Data Protection Regulation (GDPR) [31] and national data protection and privacy requirements.

Upon successful implementation of SurPass throughout 3 European health system scenarios (national EHISs, regional EHISs, and cancer registries or hospital-based EHISs), the PCSP project must focus on the three main challenges described above: (1) semiautomation of data input, (2) interoperability, and (3) data protection or privacy and cybersecurity. To support the most appropriate implementation strategy for SurPass throughout the 3 health system scenarios and overcome the 3 main challenges, a digital survey study was designed. The results of the first part of the survey, which addressed barriers and facilitators related to the care process and ethical, legal, social, and economic aspects of implementation, are described elsewhere [32]. This report describes the results of the second part of the survey, with which we aimed to identify IT-related barriers and facilitators to SurPass implementation. Subsequently, we aimed to derive insights that could be broadly applied to countries with similar types of health systems interested in implementing SurPass.

Methods

Study Design and Participants

The PCSP project representatives were asked to provide the email addresses of all the IT specialists working in their center who are responsible for the management of IT systems used to document the treatment of patients with cancer and the future implementation of SurPass (N=20). Specifically, the IT specialists invited were based in 6 LTFU care centers (hereafter referred to by their country name): 1 IT specialist from Austria (Children's Cancer Research Institute St Anna Kinderkrebsforschung), 1 from Belgium (Katholieke Universiteit Leuven and University Hospitals Leuven), 5 from Germany (University Medical Center Mainz [German Childhood Cancer Registry] and Universität zu Lübeck), 7 from Italy (Istituto Giannina Gaslini), 2 from Lithuania (Viesoji Istaiga Vilniaus Universiteto Ligonine Santaros Klinikos), and 4 from Spain (Fundación para la Investigación del Hospital Universitario la Fe de la Comunidad Valenciana). The participating centers represented the 3 European health system scenarios, including (1) nationally based EHISs or EHRs (Austria and Lithuania), (2) institutional or regional EHISs or EHRs (Italy and Spain), and (3) national cancer registries and hospital-based EHISs or EHRs (Germany and Belgium).

Survey Development

The survey was designed using the input from 6 earlier semistructured interviews with IT specialists from each of the participating centers, conducted by researchers from HL7 Europe [33]. The interviews were conducted to build up a picture of the relevant issues to be explored in the survey. In turn, the survey aimed to collect detailed data on the health system scenarios represented by the centers. First, we inquired about individual respondent characteristics, such as country of residence and organization of the IT department, followed by questions about which systems could be accessed by survivorship care staff; whether the information could be downloaded, entered, or updated; whether the information was integrated transparently, via common identifiers, or by other means; and whether the systems could exchange information using application programming interfaces (APIs). Second, respondents were asked to indicate which health data exchange (HDE) standards or interoperability frameworks were used or currently implemented in their institution. Third, the availability of childhood cancer survivor information (eg, medical history, diagnostic imaging, or pathology laboratories) and its format (eg, hardcopy, PDF, or HL7 CDA) and accessibility were examined. Similar questions about availability and format were asked about cancer diagnosis and treatment (eg, histology-cytology reports, cumulative doses of chemotherapy or radiotherapy, and types of immunotherapy) and noncancer medical information (eg, comorbidities, surgical procedures, and hereditary syndromes). Lastly, respondents were asked about data protection, data storage, and available resources related to the implementation of SurPass. The survey concluded with open-ended questions about barriers and facilitators to the implementation of the interoperable SurPass. The full survey is presented in Multimedia Appendix 1.

XSL•FO
RenderX

Ethical Considerations

Ethical review board or institutional approvals were required and obtained in Belgium (Ethics Committee Research UZ Leuven; S65576), Italy (Comitato Etico Regionale della Liguria; 385/2021, databank ID 11633), Germany (Ethik-Komission Universität zu Lübeck; 21-257), and Spain (Comité de Ética de la Investigación con medicamentos; 2019-170-1). Ethical approval was not required in Austria and Lithuania. Institutional approval for pseudonymized data collection, analysis, and storage was granted by the Princess Máxima Center, the Netherlands (Clinical Research Committee Princess Máxima Center). This study was performed in line with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants included in this study. All responses to the survey were pseudonymized. Participants were informed that by participating in this study, they consented to the publication of their pseudonymized responses.

Data Collection

Survey distribution and data collection were conducted using the cloud-based central data management platform Castor electronic data capture, a GDPR-compliant digital survey tool. The survey was conducted in English. On-site language assistance was provided to participants with limited English proficiency. The survey was launched on August 16, 2021, and closed on October 4, 2021. The results were sent to the centers for validation, which took place from December 2021 to March 2022.

Data Analysis

Pseudonymized survey data were exported from Castor electronic data capture and processed and analyzed using SPSS (version 25.0, IBM Corp). The survey results were categorized and analyzed per health system scenario and according to the three main challenges: (1) semiautomation of data input, (2) interoperability, and (3) data protection or privacy and cybersecurity. Additionally, the availability of sufficient resources (ie, staff, funding, and knowledge) to implement SurPass was considered separately for each center.

Predefined barriers to semiautomated data input and interoperability included: unaligned EHIS infrastructures without the availability of a personalized EHR or structured centralized clinical trial database from which (clinical) data could be transferred; the availability or accessibility of less than half of the requested health data information sources and patient data; the nonuse of HL7 FHIR; and the use of proprietary (noninteroperable) coding systems. Barriers to data protection or privacy and cybersecurity included not being able to privacy-protect SurPass record linkage (in line with GDPR), less than half of the IT specialists (per center) being able to guarantee data protection, and less than half of the IT specialists (per center) being familiar with data retention regulations. Finally, the unavailability of sufficient resources as indicated by at least half of the IT specialists (per center), was also considered a barrier.

Predefined facilitators for semiautomation of data input and interoperability included having a standardized and structured HDE process; the availability of a personalized EHR or structured centralized clinical trial database from which (clinical) data could be transferred in the case of nonaligned EHISs; the availability or accessibility of at least half of the requested health data information sources and patient data, ideally with the possibility to download, enter, update, and integrate the data using APIs; the use of HL7 FHIR; and the use of international coding systems such as the Anatomical Therapeutic Chemical (ATC) code system, the *International Classification of Diseases for Oncology, 3rd Edition* (*ICD-O-3*), and the *International Classification for Childhood Cancer, 3rd Edition* (*ICCC-3*) as used in SurPass. Facilitators for data protection or privacy and cybersecurity included pseudonymization of personal health data per the European Regulation 2016/679 of the European Parliament and of the Council of April 27, 2016, on the protection of natural persons per the processing of personal data and the free movement of such data, at least half of the IT specialists (per center) being able to ensure data protection, and at least half of the IT specialists (per center) being familiar with data retention regulations. Similarly, the availability of sufficient resources, as indicated by at least half of the IT specialists (per center), was considered to facilitate the implementation of SurPass.

Furthermore, we used the fit-viability model (FVM) as described by Liang et al [34] to assess the successful implementation of SurPass across the 3 unique health system scenarios. Specifically, the FVM evaluates the technological factors and organizational readiness essential for implementing information technology. Liang et al [34] defined “fit” as the match between task requirements and technology capabilities and “viability” as the economic feasibility, IT infrastructure maturity, and organizational support. In our study, we used the FVM to assess how well SurPass fits into the 3 health system scenarios and the extent to which SurPass is feasible, sustainable, and likely to be successfully integrated into the existing EHIS in the different health system scenarios, taking into account the system-specific barriers and facilitators.

Results

Survey Participants

A total of 13 IT specialists (1 from Austria and Belgium, 2 from Lithuania and Spain, 3 from Germany, and 4 from Italy) responded to the survey invitation. Further, 3 IT specialists represented the first scenario (national EHISs), 5 IT specialists represented the second scenario (regional or local EHISs), and 4 IT specialists represented the third scenario (cancer registries and hospital-based EHISs). The overall response rate was 65%. The response rates per country were 100% for Austria (1/1), Belgium (1/1), and Lithuania (2/2); 60% (3/5) for Germany; 57% (4/7) for Italy, and 50% (2/4) for Spain.

Barriers and Facilitators per Health System Scenario

Overview of Available Data Systems and Resources

Inherent to the 3 health system scenarios, there are barriers and facilitators relevant to the implementation of SurPass. The characteristics of these scenarios, categorized according to the 3 main implementation challenges, are described below. The availability of data systems and resources in each center is

XSL•FO
RenderX

discussed accordingly (see also Tables 1-4). The center-specific HDE standards, interoperability frameworks, and coding systems for each center are described separately by Chronaki et al [33]. A general overview of all the identified barriers and facilitators to the implementation of the SurPass is provided in Table 5.

XSL•FO
RenderX

Table 1. Information system sources in each health system scenario and center.

Accessible data systems	Scenario 1		Scenario 2		Scenario 3	
	Lithuania	Austria	Italy	Spain	Belgium	Germany
National EHR[a]	Yes	Yes	Yes	Yes	Yes	No
Possibility to download data	Yes	Yes	Yes	Yes	Yes	No
Possibility to enter data	Yes	*m*[b]	Yes	No	Yes	No
Possibility to update data	Yes	Yes	Yes	No	No	No
Integrated transparently	No	*m*	Yes	No	No	No
APIs[c] available to integrate	Yes	Yes	Yes	No	Yes	No
Regional EHR	?[d]	*m*	~[e]	Yes	No	No
Possibility to download data	?	*m*	Yes	No	No	No
Possibility to enter data	?	*m*	Yes	Yes	No	No
Possibility to update data	?	*m*	Yes	Yes	No	No
Integrated transparently	?	*m*	?	~	No	No
APIs available to integrate	?	*m*	Yes	~	No	No
Cancer registry	Yes	Yes	Yes	Yes	Yes	Yes
Possibility to download data	No	Yes	Yes	?	Yes	No
Possibility to enter data	No	Yes	Yes	Yes	Yes	Yes
Possibility to update data	No	Yes	Yes	Yes	No	Yes
Integrated transparently	?	*m*	?	Yes	No	?
APIs available to integrate	No	Yes	?	No	Yes	No
Hospital EMR[f]	Yes	*m*	Yes	Yes	Yes	Yes
Possibility to download data	Yes	*m*	Yes	Yes	Yes	No
Possibility to enter data	Yes	*m*	Yes	Yes	Yes	?
Possibility to update data	Yes	*m*	Yes	Yes	Yes	?
Integrated transparently	Yes	*m*	Yes	Yes	Yes	?
APIs available to integrate	Yes	*m*	Yes	Yes	No	?
Patient records at LTFU[g] **care center**	Yes	*m*	Yes	Yes	Yes	?
Possibility to download data	Yes	*m*	Yes	Yes	Yes	?
Possibility to enter data	Yes	*m*	Yes	Yes	Yes	?
Possibility to update data	Yes	*m*	Yes	Yes	Yes	?
Integrated transparently	*m*	*m*	Yes	No	Yes	?
APIs available to integrate	*m*	*m*	Yes	No	No	?
Patient records at other outpatient center	Yes	*m*	No	No	Yes	?
Possibility to download data	Yes	*m*	No	No	Yes	?
Possibility to enter data	Yes	*m*	No	No	No	?
Possibility to update data	Yes	*m*	No	No	No	?
Integrated transparently	Yes	*m*	No	No	Yes	?
APIs available to integrate	Yes	*m*	No	No	No	?
Appointment scheduling system	Yes	Yes	Yes	Yes	Yes	Yes
Possibility to download data	Yes	*m*	Yes	Yes	Yes	?
Possibility to enter data	Yes	Yes	Yes	Yes	Yes	?
Possibility to update data	Yes	Yes	Yes	Yes	Yes	?

XSL•FO
RenderX

Accessible data systems	Scenario 1		Scenario 2		Scenario 3	
	Lithuania	Austria	Italy	Spain	Belgium	Germany
Integrated transparently	?	*m*	Yes	Yes	Yes	?
APIs available to integrate	Yes	*m*	Yes	Yes	No	?
Pharmacy	Yes	*m*	Yes	Yes	Yes	Yes
Possibility to download data	?	*m*	Yes	Yes	Yes	?
Possibility to enter data	?	*m*	Yes	Yes	Yes	?
Possibility to update data	?	*m*	Yes	Yes	Yes	?
Integrated transparently	Yes	*m*	Yes	Yes	Yes	?
APIs available to integrate	?	*m*	Yes	Yes	Yes	?
Laboratories	Yes	Yes	Yes	Yes	Yes	Yes
Possibility to download data	Yes	Yes	Yes	Yes	Yes	?
Possibility to enter data	No	No	Yes	Yes	Yes	?
Possibility to update data	No	No	Yes	Yes	Yes	?
Integrated transparently	?	*m*	Yes	Yes	Yes	?
APIs available to integrate	?	*m*	Yes	Yes	Yes	?
Radiology	Yes	Yes	Yes	Yes	Yes	?
Possibility to download data	Yes	Yes	Yes	Yes	Yes	?
Possibility to enter data	Yes	No	Yes	Yes	Yes	?
Possibility to update data	Yes	No	Yes	Yes	Yes	?
Integrated transparently	?	*m*	Yes	Yes	Yes	?
APIs available to integrate	?	*m*	Yes	Yes	Yes	?
Clinical trial systems	Yes	*m*	No	No	Yes	Yes
Possibility to download data	Yes	*m*	No	No	No	?
Possibility to enter data	Yes	*m*	No	No	No	?
Possibility to update data	Yes	*m*	No	No	No	?
Integrated transparently	Yes	*m*	No	No	Yes	?
APIs available to integrate	Yes	*m*	No	No	No	?
Primary health care information system	Yes	Yes	No	Yes	No	Yes
Possibility to download data	*m*	Yes	No	Yes	No	No
Possibility to enter data	*m*	Yes	No	Yes	No	?
Possibility to update data	*m*	Yes	No	Yes	No	?
Integrated transparently	*m*	*m*	No	Yes	No	?
APIs available to integrate	*m*	*m*	No	Yes	No	?

[a]EHR: electronic health record.
[b]*m*: Answers missing.
[c]API: application programming interface.
[d]?: Respondents did not know.
[e]~: Unclear, for example, 1 × “yes” and 1 × “no.”
[f]EMR: electronic medical record.
[g]LTFU: long-term follow-up.

XSL•FO
RenderX

Table 2. Patient data that is electronically available in each health system scenario and center[a].

Characteristics	Scenario 1	Scenario 2		Scenario 3	
	Lithuania	Italy	Spain	Belgium	Germany
Patient data type					
Patient summary or medical history	Yes	Yes	Yes	Yes	Yes
Hospital admissions	Yes	Yes	Yes	Yes	Yes
Diagnostic imaging (images)	Yes	Yes	Yes	Yes	Yes
Diagnostic imaging (reports)	Yes	Yes	Yes	Yes	Yes
Biochemical laboratories	Yes	Yes	Yes	Yes	Yes
Pathology laboratories	Yes	Yes	Yes	Yes	Yes
CC[b] treatment summary	Yes	Yes	Yes	?[c]	Yes
LTFU[d] care visit report	?	Yes	No	?	?
SurPass[e]	No	Yes	No	No	No
LTFU center appointments	Yes	?	?	?	?
Other medical databases	?	?	No	Yes	Yes

[a]An earlier version of the Survivorship Passport has been implemented and evaluated previously in Italy. Austria was intentionally left out of the table because all answers were missing.

[b]CC: childhood cancer.

[c]Respondents did not know.

[d]LTFU: long-term follow-up.

[e]SurPass: Survivorship Passport.

XSL•FO
RenderX

Table 3. Overview of accessible patient data formats in each health system scenario and center. Austria was intentionally left out of the table because all answers were missing.

Patient data type	Scenario 1	Scenario 2		Scenario 3	
	Lithuania	Italy	Spain	Belgium	Germany
General medical history					
Comorbidities	****[a]	****	?[b]	?	***[c]
Allergies	****	****	****	***	****
Medication (noncancer related)	****	****	****	***	**[d]
Surgical procedures (noncancer related)	****	****	****	***	**
Admissions	***	****	***	**	***
Trauma	****	****	**	**	No
Hereditary syndromes	***	No	**	***	No
Cancer diagnosis					
Cancer diagnosis	****	?	***	****	**
Cancer diagnosis date	****	*[e]	***	[f]	****
Histology-cytology report	***	***	****	**	***
Imaging reports	**	***	****	**	**
Laboratory reports	****	****	***	**	**
Cancer treatment					
Surgical interventions	**	****	****	**	**
Stem cell or bone marrow transplantations	***	**	?	**	No
Chemotherapy start or end date	?	**	***	***	**
Chemotherapy type	?	***	***	****	**
Chemotherapy cumulative dose	**	***	***	?	?
Chemotherapy treatment complications	?	**	***	***	?
Immunotherapy start or end date	?	?	?	**	?
Immunotherapy type	?	?	?	**	?
Immunotherapy cumulative dose	?	?	?	?	?
Immunotherapy treatment complications	?	?	?	?	?
Hormonal therapy start or end date	?	?	?	**	?
Hormonal therapy type	?	?	?	**	?
Hormonal therapy cumulative dose	?	?	?	?	?
Hormonal therapy treatment complications	?	?	?	?	?
Radiotherapy start or end date	**	*	***	**	?
Radiotherapy type	?	*	***	**	?
Radiotherapy cumulative dose	?	*	***	?	?
Radiotherapy site	?	*	***	?	?
Radiotherapy treatment complications	?	*	***	?	?

[a]Electronically available, coded, and interconnected.
[b]Respondents did not know.
[c]Yes, electronically available and coded.
[d]Yes, electronically available.
[e]Yes, hardcopy available (awarded when at least one respondent mentioned “yes”).
[f]Electronically available, free text, and interconnected.

XSL•FO
RenderX

Table 4. Scarcity of resources in each health system scenario and center. The table shows the number of IT specialists per country that indicated either having (yes) or not having (no) sufficient staff, time, funds, and knowledge.

Resource type lacking, n	Scenario 1		Scenario 2		Scenario 3	
	Lithuania	Austria	Italy	Spain	Belgium	Germany
Staff						
Yes	0	1	1	1	0	2
No	2	0	3	1	1	
Time						
Yes	0	1	2	1	0	2
No	2	0	2	1	1	1
Funds						
Yes	0	1	2	1	0	1
No	2	0	2	1	1	2
Knowledge						
Yes	0	1	2	0	0	1
No	2	0	2	2	1	2

XSL•FO
RenderX

Table 5. Summary of the identified barriers and facilitators to the implementation of the SurPass[a] (version 2.0) throughout all 3 health care system scenarios.

Characteristics	Barriers	Facilitators
Scenario 1: national EHISs[b]		
Austria[c]	• Uncertainty about data retention • Lack of resources	• Standardized and structured data exchange • Straightforward GDPR[d] compliance using pseudonymization • Data protection can be guaranteed • Use of ATC[e] code system for medication • Use of ICD-O-3[f] for histology
Lithuania	• Uncertainty about data retention	• Standardized and structured data exchange • Majority of the data systems is accessible • Majority of patient data is available • Use of HL7 FHIR[g] • Data protection can be guaranteed • Sufficiency of resources • Use of ICD-O-3 for histology
Scenario 2: regional EHISs		
Italy	• EHIS infrastructure discrepancies • Uncertainty about data protection • Uncertainty about data retention • Uncertainty about the sufficiency of resources	• Availability of personalized EHR[h] (Fascicolo Sanitario Elettronico) • Majority of the data systems is accessible • Majority of patient data is available • Use of HL7 FHIR • Use of ATC code system for medication
Spain	• EHIS infrastructure discrepancies • Unavailability of HDE[i] standards • Uncertainty about data retention • Uncertainty about the sufficiency of resources	• Availability of personalized EHR (Historia de Salud Electrónica) • Majority of the data systems is accessible • Majority of patient data is available • Data protection can be guaranteed
Scenario 3 cancer registries and hospital-based EHISs		
Belgium	• Data origination from different sources • EHIS infrastructure discrepancies • Uncertainty about data protection • Uncertainty about data retention • Uncertainty about the sufficiency of resources	• Majority of the data systems is accessible • Majority of patient data is available • Use of HL7 FHIR • Use of ATC code system for medication
Germany	• Data origination from different sources • EHIS infrastructure discrepancies • Majority of the data systems is accessible but not downloadable • Uncertainty about data protection • Uncertainty about data retention • Lack of resources	• Availability of structured centralized clinical trial databases • Majority of patient data is available • Use of HL7 FHIR • Use of ICD-O-3 and ICCC-3[j] for cancer diagnosis and histology

[a]SurPass: Survivorship Passport.
[b]EHIS: electronic health information system.
[c]Accessibility of data systems and patient data availability is unknown due to missing answers from the information technology specialist from Austria.
[d]GDPR: General Data Protection Regulation.
[e]ATC: Anatomical Therapeutic Chemical.
[f]ICD-O-3: International Classification of Diseases for Oncology, 3rd Edition.
[g]HL7 FHIR: Health Level 7 Fast Healthcare Interoperability Resources.
[h]EHR: electronic health record.
[i]HDE: health data exchange.
[j]ICCC-3: International Classification of Childhood Cancer, Third edition.

National EHISs (Austria and Lithuania)

In this scenario, the data required for the patient-specific oncological treatment summary will be provided by HCPs through a national EHISs, such as the ELGA (Austrian National Electronic Health Record Systems) or the Lithuanian Electronic Health Services and Infrastructure Cooperation System (ESPBI IS). The patient-specific oncological treatment summary, which

XSL•FO
RenderX

takes into account the cumulative treatment burden and extends to the end of treatment of the first pediatric cancer, will be systematically generated in CDA format. This scenario thus provides a standardized and structured way of exchanging data. In addition, the accessibility of digital data is fundamental to enable semiautomated data input from EHISs and to facilitate interoperability using HL7 FHIR. In Lithuania, these conditions are met as all data systems and the majority of patient data are accessible (Tables 1-3) and HL7 FHIR is used. In addition, the IT specialists from Lithuania indicated that they have sufficient resources (staff, funds, and knowledge) available to implement SurPass (version 2.0) in their center (Table 4). On the other hand, the IT specialist from Austria indicated that he did not have sufficient resources at hand. In addition, no information was available on the accessibility of the Austrian data system. In terms of interoperability, Austria currently uses HL7 CDA in ELGA and is working toward the adoption of HL7 FHIR in the future. For medication, Austria uses the ATC code system. For cancer diagnosis, Lithuania uses the *ICCC-3*. For histology, both Lithuania and Austria use the *ICD-O-3*. Other coding systems varied substantially between the 2 centers.

Per privacy and security, the structured data in this scenario will be pseudonymized where necessary, for example by using the European Patient Identity Service (EUPID; pseudonymization and privacy-preserving record linkage tool to facilitate secondary use of data sets in biomedical research and health care [35,36]). Next, the data will be transferred to the SurPass platform to generate the SurPass. EUPID will certainly be used in Austria, but there is no consensus yet in Lithuania. In the future, the use of EUPID for the HDE process will ensure GDPR compliance. In this study, survey results indicated that IT professionals in Austria and Lithuania have existing solutions that ensure SurPass data protection. Once the pseudonymized SurPass is generated, either the entire SurPass (in the case of Lithuania's Electronic Health Services and Infrastructure Cooperation System) or the SCP (in the case of Austria's ELGA) will be returned to the HCP and the childhood cancer survivor's local care team, where the SCP within the SurPass can be tailored to the childhood cancer survivor's individual needs and national care pathways in the local health IT system. The SurPass will be accessible to both childhood cancer survivors and HCPs through the national EHIS or EHR. However, the IT specialists from Austria and Lithuania were uncertain about SurPass data retention. In the Austrian ELGA, data retention is currently limited to 10 years, which is likely to be too short for the SurPass concept.

In terms of the FVM, this scenario shows a high fit for SurPass, particularly in Lithuania given their current use of HL7 FHIR and the accessibility of data systems. In Austria, where a transition to HL7 FHIR is underway, SurPass compatibility is growing. Viability is strong with evident economic feasibility for Lithuania and effective data protection strategies present in both countries. Thus, prospects for SurPass adoption in this scenario are promising, especially in Lithuania, where ample resources contribute to a favorable implementation environment. Provided that uncertainties about data retention are resolved and Austria allocates sufficient resources, the implementation of SurPass is likely to be successful.

Regional EHISs (Italy and Spain)

In the second scenario, the main pediatric cancer institutions in each region have local registries or treatment protocol databases with demographic information and treatment data. Here, EHIS infrastructure is not fully aligned between regions or local hospitals in the same region, limiting semiautomated data exchange and interoperability. Similarly, EHISs may have been adopted at different times and therefore differ in terms of completeness or level of sophistication of the system, limiting interoperability. However, SurPass implementation can be aided by the availability of personalized EHRs (Fascicolo Sanitario Elettronico in Italy and Historia de Salud Electrónica in Spain), which contains medical information that is accessible to all HCPs in the region or country and which also allows patients to access their data. However, it is worth noting that the FSE in Italy is currently undergoing a significant reorganization, transitioning from a regional to a national homogeneous system that will use HL7 FHIR. This transition may lead to changes in the way medical information is accessed and exchanged between HCPs, but it will allow SurPass to be integrated into the Italian EHR. Furthermore, the survey results showed that in both Italy and Spain, the majority of data systems and patient data are accessible to survivorship care staff (Tables 1-3). Apart from the cancer registry, all data systems in Italy are also transparently integrated, including APIs, supporting interoperability between these systems together with the use of HL7 FHIR. Spain, on the other hand, has not yet adopted HL7 FHIR. For medication, Italy uses the ATC system. Other coding systems varied substantially between the 2 centers.

In this scenario, the SurPass (version 2.0) is intended to be integrated into the EHIS of the treating institution for follow-up care, as well as into the EHR for use in possible emergency admissions to other hospitals and during the transition from pediatric to adult LTFU care. Similar to the first scenario, the cross-border data transfer could be pseudonymized by using a GDPR-compliant double pseudonymization system such as EUPID. Personalization of the SurPass data would only take place within the treating institution itself, or in other local databases with permission to hold identifying personal data. The results of the survey showed that, in contrast to the Italian respondents, the majority of IT specialists from Spain indicated that they were able to guarantee the protection of SurPass data. Neither center's IT specialists knew how long SurPass records had to be stored. Besides, IT specialists in both centers were unsure about the availability of sufficient resources to implement SurPass (Table 4).

In the context of the FVM, this scenario shows a moderate fit for SurPass implementation within regional EHIS structures. Viability is impacted by EHIS infrastructures that are not fully aligned between regions and local hospitals, hindering seamless semiautomated data exchange and interoperability. Despite these barriers, the availability of personalized EHRs offers potential support for SurPass integration. Furthermore, viability is dependent on data protection concerns and GDPR compliance. Provided these challenges are addressed and sufficient resources can be secured, the adoption of SurPass in centers representing this scenario appears promising.

XSL•FO
RenderX

Cancer Registries and Hospital-Based EHISs (Belgium and Germany)

The third scenario is the least favorable in terms of semiautomated data input, interoperability, and data protection. Unlike the scenarios with nationally or regionally organized health data, this scenario requires the integration of data from a multitude of sources to generate a SurPass. Epidemiological data will need to be obtained from cancer or bone marrow transplant registries such as the German Childhood Cancer Registry, while disease and treatment-specific clinical data will need to be retrieved from hospital databases. In addition, although it is possible to access and transfer clinical data from central clinical trial databases in Germany, this process adds complexity and is not as streamlined as data collection in the other 2 scenarios. Both Belgium and Germany store SurPass data at CINECA, the largest data center in Italy [37]. The resulting SurPass will be returned digitally to the centers in PDF format. Currently, in both Belgium and Germany, most data systems and patient data are accessible, although in Germany without the possibility of downloading or further exploiting the data (Tables 1-3). Yet, if the data systems are already accessible, the possibilities to download, enter, update, and integrate the data can be explored shortly as applicable during the PCSP project. Moreover, both centers in Belgium and Germany are using HL7 FHIR, which improves interoperability between local health data systems. For cancer diagnosis and histology, Germany uses the cancer-specific coding systems ICD-O-3 and ICCC-3, and Belgium uses the ATC code system for medication. Other coding systems varied substantially between the centers. Lastly, at the time of completing the questionnaire, IT specialists from Belgium and Germany were uncertain about guaranteeing data protection, data retention, and the availability of sufficient resources in their centers (Table 4).

In this scenario, the FVM underscores significant barriers to SurPass implementation. SurPass fit is challenged by the need to integrate data from diverse sources. However, viability is reasonable, with accessible data systems and HL7 FHIR adoption improving interoperability. SurPass adoption or implementation supported by accessible data systems is feasible but will depend on overcoming challenges related to data transfer, data protection, and resource availability.

Discussion

Principal Findings

Using a digital survey, we assessed IT-related barriers and facilitators to the implementation of the digital SurPass throughout 3 health system scenarios (national EHISs—Austria and Lithuania; regional or local EHISs—Italy and Spain; and cancer registries or hospital-based EHISs—Belgium and Germany). The survey involved 13 IT specialists from the participating centers. The application of the FVM to the 3 distinct health system scenarios provides valuable insights into the feasibility of implementing SurPass in pediatric cancer survivorship care. Scenario 1 (national EHISs) shows a positive outlook for a successful SurPass implementation, especially in Lithuania, where current HL7 FHIR usage and accessible data systems contribute to a high fit and good viability. In Austria, ongoing efforts to transition to HL7 FHIR indicate growing compatibility. Viability is strong in both countries, setting the stage for promising SurPass adoption, provided that uncertainties around data retention in Austria are addressed and sufficient resources are allocated. Scenario 2 (regional or local EHISs) presents a more nuanced picture with a moderate fit. Viability is hampered by misaligned EHIS infrastructures, impacting semiautomated data exchange and interoperability. Despite these barriers, personalized EHRs offer support for SurPass, provided that data protection concerns are dealt with. However, uncertainties among IT professionals regarding data protection, storage, and resource availability suggest that careful consideration is needed in the adoption process. Finally, scenario 3 (cancer registries or hospital-based EHISs) presents significant barriers to good SurPass fit due to the complexity of integrating data from multiple sources. Nevertheless, a reasonable viability, characterized by accessible data systems and HL7 FHIR adoption, suggests successful SurPass implementation in the future, provided that challenges related to data transfer, data protection, and resource availability are effectively handled. These findings highlight the importance of context-specific barriers and facilitators to the implementation of SurPass in the 3 different health care scenarios.

SurPass implementation requires the exchange of medical data between various data sources, especially in the second and third health system scenarios. National EHISs, as in Austria and Lithuania, are best suited for semiautomated data transfer from the EHIS of the treating institution to the SurPass platform, followed by the regional or local EHIS infrastructures, as in the second and third scenarios. Ultimately, semiautomated data input is paramount to making clinical care processes more efficient. Besides the EHIS infrastructure, the use of HL7 FHIR is of major importance for connecting SurPass to EHISs, registries, and national or regional EHRs. Due to its international recognition and adoption following the open API paradigm, FHIR is an ideal choice for promoting interoperability and the standardized exchange of health care data. In addition to the IPS, FHIR's flexibility, extensibility, and wide adoption within the health care industry support our decision to recommend FHIR. All centers except Spain and Austria are already using HL7 FHIR. Austria is currently using HL7 CDA in ELGA but is also working toward the adoption of HL7 FHIR. Relatedly, a systematic literature analysis on barriers and facilitators for the implementation of eHealth services found that having well-established HDE standards is an important success factor for eHealth services [38]. Furthermore, the lack of time for HCPs to create and update the SurPass is an important barrier in the survivorship care process [32]. Since the time needed to prepare and update SurPass depends on the level of interoperability that can be achieved between existing EHISs and SurPass, we recommend the LTFU care center in Spain and any other institution with the ambition to implement SurPass to adopt HL7 FHIR in all EHISs.

Complications with data accessibility and data privacy or security are listed in the top 10 barriers mentioned in 38 papers on the implementation of eHealth solutions [38]. On the one hand, this study illustrated that the majority of information sources and patient data types is accessible and downloadable

XSL•FO
RenderX

in all health system scenarios, except in Germany (scenario 3), where the data are accessible but not downloadable. On the other hand, only in the first scenario (Austria and Lithuania) and in Spain (scenario 2) did IT specialists report being able to ensure SurPass data protection. Per data retention, IT specialists exhibited uncertainty and inconsistency in all scenarios. It is important to consider the context of the survey when interpreting these results. For example, in Germany, IT specialists were responsible for software development but were not involved in detailed data protection discussions, which were handled by the official data protection officer. Each of the IT specialists involved is aware of the necessary prerequisites to ensure data protection. They may need to establish a specific framework to effectively integrate SurPass data and comply with data protection standards. It remains imperative for all centers to have a data protection solution in place, as compliance with GDPR and national requirements for data protection and privacy is an essential prerequisite for the successful implementation of SurPass in all health system scenarios. Data encryption, pseudonymization, and multifactor authentication—a security process in which a user provides two or more authentication factors to verify their identity before accessing a protected resource or system—could form appropriate data protection measures for all health system scenarios.

The insights gained from SurPass implementation have broader applicability, offering valuable lessons for the introduction of similar eHealth solutions in various health care settings. SurPass serves as a practical example illustrating the integration of digital health tools into existing EHISs, and the application of the FVM framework provides essential technological considerations for implementing such solutions beyond the SurPass context. In addition, our study is of strategic importance in light of the upcoming European Health Data Space (EHDS) regulation, which refines the GDPR and requires electronic health systems to be able to exchange data in the European Electronic Health Record Exchange Format [39]. SurPass has many common elements with the IPS, which is one of the priority data categories in the EHDS regulation. In view of this, our study, despite its specific focus, sheds light on some of the hurdles of digital transformation in Europe as it explores the health data economy in realistic care pathways. As such, SurPass lays the groundwork for the pragmatic adoption of key regulations and upcoming directives such as the EHDS, the Medical Device Directive, and the Artificial Intelligence Act [39-41].

Limitations

The limitations of this study include missing data from Austria regarding accessibility of data systems and availability of patient data, and inconsistent or uncertain responses within individual centers. This may be because SurPass is a newly developed tool and many IT specialists were unfamiliar with SurPass and its requirements. In addition, there were many unknown responses from German centers regarding information system sources, which may be attributed to the complexity of the IT systems in place and the lack of involvement of German IT specialists in detailed discussions on data protection. We, therefore, acknowledge that the overview of computing infrastructures and IT landscapes across the 6 centers may be incomplete. Furthermore, the number of IT specialists participating in this study may appear relatively small. The selection process was based on their direct involvement in the project to ensure a relevant perspective on the SurPass implementation. This study's design including 1 center per country was based on resource allocation and practical considerations, as outlined in the PCSP project agreement. Despite initially inviting IT specialists from each of the 6 participating centers (N=20), only 13 provided complete responses. This limited participation reflects the challenges of understaffed and less engaged hospital IT departments across Europe. We acknowledge the potential for selection bias, as their individual expertise was not systematically assessed. However, it is important to note that after analyzing the survey results, we conducted a thorough verification of the IT specialists' responses with all participating centers. While we recognize this study's limitations, we emphasize that our findings are of substantial value in shaping an effective implementation strategy for SurPass in all 3 health system scenarios. To facilitate the wider adoption of SurPass, we plan to publish an implementation toolkit at the end of the PCSP project.

Conclusions

This paper described the findings from a digital survey study that assessed IT-related barriers and facilitators to SurPass implementation in 3 European health system scenarios, that is nationally based EHISs or EHRs (Austria and Lithuania); institutional or regional EHISs or EHRs (Italy and Spain); and national cancer registries and hospital-based EHISs or EHRs (Germany and Belgium). In all scenarios, barriers and facilitators were related to 3 main challenges including semiautomated data input, interoperability, and data protection or privacy and cybersecurity. For all scenarios, we recommend the adoption of HL7 FHIR to support interoperability. In addition, we emphasize the importance of GDPR compliance in all scenarios, such as using encryption, EUPID pseudonymization, and multifactor authentication where applicable. Our results will support the SurPass implementation strategies for all 3 health system scenarios. Ultimately, SurPass implementation provides lessons for introducing similar eHealth solutions across diverse health care settings, offering practical insights into integrating digital health solutions into existing and future EHISs and paving the way for the pragmatic adoption of key regulations such as the EHDS, the Medical Device Directive, and the Artificial Intelligence Act.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation program (grant 899999). The funding source was not involved in this study's design, data collection, data analysis, data interpretation, the writing of this paper, or the decision to submit this paper for publication.

XSL•FO
RenderX

The members of the *PanCareSurPass Consortium* are as follows:

Universitaetsmedizin der Johannes Gutenberg-Universitaet Mainz (UMC-Mainz): Desiree Grabow, Anna-Liesa Filbert, Dorothea Niehoff, Diana Walz, Claudia Spix, Friederike Erdmann (until Dec 2022), Cecile Ronkers (after Dec 2022).

Istituto Giannina Gaslini (IGG): Riccardo Haupt, Monica Muraca, Simone Lightwood, Francesca Bagnasco, Giacomo Cavalca, Sara Oberti, Ramona Tallone (after Dec 2022), Andrea Beccaria (after Dec 2022).

St Anna Children's Cancer Research Institute (CCRI)/St Anna Children's Hospital (SAK): Ruth Ladenstein, Edit Bardi, Vanessa Düster (until Dec 2022), Serena Mancini (after Dec 2022), Leo Kager (after Dec 2022).

Katholieke Universiteit Leuven (KU Leuven): Anne Uyttebroeck, Maria van Helvoirt, Jurgen Lemiere, Marleen Renard, An Michiels, Laura Tuts.

Universitaet zu Luebeck (UzL): Thorsten Langer, Ann-Kristin Kock-Schoppenhauer, Lea Hildebrand, Anke Neumann, Anne Katrin Jahnke (until Aug 2022), Sarah Rieken (after Dec 2022).

Viesoji Istaiga Vilniaus Universiteto Ligonine Santaros Klinikos (VULSK): Jelena Rascon, Justas Trinkūnas, Audronė Ciesiūnienė, Paulius Ragauskas.

Instituto de Investigación Sanitaria La Fe Hospital Universitario y Politécnico La Fe (IISLaFe—HUyPLaFe): Adela Cañete Nieto, Julia Balaguer Guill, Maria Teresa Tormo Alcañiz, Antonio Orduña Galan, Marisa Correcher Palau, Lucas Cervero Beltrán.

Cineca Consorzio interuniversitario (CINECA): Davide Saraceno, Alessandra Berti, Carlo Contino, Nikos Thomopulos, Giulia Stabile (until Dec 2022).

Azienda Ligure Sanitaria della Regione Liguria (A.Li.Sa): Maria Franca Tomassi, Igor Zamberlan, Barbara Nichel.

AIT Austrian Institute of Technology GmbH (AIT): Günter Schreier, Dieter Hayn, Karl Kreiner, Stefan Beyer.

Health Level Seven (HL7 Europe): Catherine Chronaki, Giorgio Cangioli, Eliana Charalambous.

Gesundheit Österreich GmbH (GÖG): Alexander Degelsegger Márquez, Gerald Gredinger, Kathrin Trunner, Florian Trauner, Anja Laschkolnig.

Prinses Máxima Centrum voor Kinderoncologie (PMC): Leontien Kremer, Heleen van der Pal, Saskia Pluijm, Selina van den Oever, Ismay de Beijer, Jessica Trollip (until Feb 2022), Emma Hardijzer (until Aug 2022).

PanCare: Heleen van der Pal, Jaap den Hartogh, Jeroen te Dorsthorst.

SIOP Europe ASBL (SIOPE): Samira Essiaf, William Sciberras (until Aug 2022), Giorgia Manuzi (after Aug 2022).

Childhood Cancer International (CCI): Anita Kienesberger, Hannah Gsell, Carina Schneider, Zuzana Tomasikova.

Pintail LTD (PT) Not applicable.

Data Availability

Study participants did not consent to data sharing outside the PanCareSurPass project. Access to participant data is therefore limited to national and international supervisory authorities. This study's protocol can be made available upon request to the corresponding author.

Authors' Contributions

All authors contributed to the conception and design of this study and the design of the surveys. Ethical approvals were obtained and data were collected by authors affiliated with the 6 participating institutions. Data were analyzed by IAEdB, SRvdO, LCMK, HJHvdP, SMFP, CC, and EC, and results were interpreted by all authors. This paper was drafted by IAEdB and critically revised by all authors. All authors approved the final version of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full survey.
[PDF File (Adobe PDF File), 224 KB - jmir_v26i1e49910_app1.pdf]

References

XSL•FO
RenderX

1. Steliarova-Foucher E, Fidler MM, Colombet M, Lacour B, Kaatsch P, Piñeros M, et al. Changing geographical patterns and trends in cancer incidence in children and adolescents in Europe, 1991-2010 (automated childhood cancer information system): a population-based study. Lancet Oncol 2018;19(9):1159-1169 [FREE Full text] [doi: 10.1016/S1470-2045(18)30423-6] [Medline: 30098952]
2. The SIOPE strategic plan: a European cancer plan for children and adolescents. SIOP Europe: The European Society for Paediatric Oncology. 2015. URL: https://siope.eu/media/wp-content/uploads/2018/02/the-siope-strategic-plan.pdf [accessed 2024-03-29]
3. van Deuren S, Penson A, van Dulmen-den Broeder E, Grootenhuis MA, van der Heiden-van der Loo M, Bronkhorst E, et al. Prevalence and risk factors of cancer-related fatigue in childhood cancer survivors: a DCCSS LATER study. Cancer 2022;128(5):1110-1121 [FREE Full text] [doi: 10.1002/cncr.33993] [Medline: 34724201]
4. Gatta G, Botta L, Rossi S, Aareleid T, Bielska-Lasota M, Clavel J, et al. Childhood cancer survival in Europe 1999-2007: results of EUROCARE-5—a population-based study. Lancet Oncol 2014;15(1):35-47. [doi: 10.1016/S1470-2045(13)70548-5] [Medline: 24314616]
5. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. Nat Rev Cancer 2014;14(1):61-70 [FREE Full text] [doi: 10.1038/nrc3634] [Medline: 24304873]
6. Bhakta N, Liu Q, Ness KK, Baassiri M, Eissa H, Yeo F, et al. The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). Lancet 2017;390(10112):2569-2582 [FREE Full text] [doi: 10.1016/S0140-6736(17)31610-0] [Medline: 28890157]
7. Hudson MM, Ness KK, Gurney JG, Mulrooney DA, Chemaitilly W, Krull KR, et al. Clinical ascertainment of health outcomes among adults treated for childhood cancer. JAMA 2013;309(22):2371-2381 [FREE Full text] [doi: 10.1001/jama.2013.6296] [Medline: 23757085]
8. Leerink JM, de Baat EC, Feijen EAM, Bellersen L, van Dalen EC, Grotenhuis HB, et al. Cardiac disease in childhood cancer survivors: risk prediction, prevention, and surveillance: JACC CardioOncology state-of-the-art review. JACC CardioOncol 2020;2(3):363-378 [FREE Full text] [doi: 10.1016/j.jaccao.2020.08.006] [Medline: 34396245]
9. Christen S, Roser K, Mulder RL, Ilic A, Lie HC, Loonen JJ, et al. Recommendations for the surveillance of cancer-related fatigue in childhood, adolescent, and young adult cancer survivors: a report from the International Late Effects of Childhood Cancer Guideline Harmonization Group. J Cancer Surviv 2020;14(6):923-938 [FREE Full text] [doi: 10.1007/s11764-020-00904-9] [Medline: 32839902]
10. Mulder RL, Font-Gonzalez A, Hudson MM, van Santen HM, Loeffen EAH, Burns KC, et al. Fertility preservation for female patients with childhood, adolescent, and young adult cancer: recommendations from the PanCareLIFE Consortium and the International Late Effects of Childhood Cancer Guideline Harmonization Group. Lancet Oncol 2021;22(2):e45-e56. [doi: 10.1016/S1470-2045(20)30594-5] [Medline: 33539753]
11. Mulder RL, Font-Gonzalez A, Green DM, Loeffen EAH, Hudson MM, Loonen J, et al. Fertility preservation for male patients with childhood, adolescent, and young adult cancer: recommendations from the PanCareLIFE Consortium and the International Late Effects of Childhood Cancer Guideline Harmonization Group. Lancet Oncol 2021;22(2):e57-e67. [doi: 10.1016/S1470-2045(20)30582-9] [Medline: 33539754]
12. Norsker FN, Rechnitzer C, Andersen EW, Linnet KM, Kenborg L, Holmqvist AS, et al. Neurologic disorders in long-term survivors of neuroblastoma—a population-based cohort study within the Adult Life after Childhood Cancer in Scandinavia (ALiCCS) research program. Acta Oncol 2020;59(2):134-140 [FREE Full text] [doi: 10.1080/0284186X.2019.1672892] [Medline: 31591921]
13. Armenian SH, Hudson MM, Mulder RL, Chen MH, Constine LS, Dwyer M, et al. Recommendations for cardiomyopathy surveillance for survivors of childhood cancer: a report from the International Late Effects of Childhood Cancer Guideline Harmonization Group. Lancet Oncol 2015;16(3):e123-e136 [FREE Full text] [doi: 10.1016/S1470-2045(14)70409-7] [Medline: 25752563]
14. Bardi E, Mulder RL, van Dalen EC, Bhatt NS, Ruble KA, Burgis J, et al. Late hepatic toxicity surveillance for survivors of childhood, adolescent and young adult cancer: recommendations from the International Late Effects of Childhood Cancer Guideline Harmonization Group. Cancer Treat Rev 2021;100:102296. [doi: 10.1016/j.ctrv.2021.102296] [Medline: 34571378]
15. Teepen JC, van Leeuwen FE, Tissing WJ, van Dulmen-den Broeder E, van den Heuvel-Eibrink MM, van der Pal HJ, et al. Long-term risk of subsequent malignant neoplasms after treatment of childhood cancer in the DCOG LATER Study Cohort: role of chemotherapy. J Clin Oncol 2017;35(20):2288-2298 [FREE Full text] [doi: 10.1200/JCO.2016.71.6902] [Medline: 28530852]
16. Friend AJ, Feltbower RG, Hughes EJ, Dye KP, Glaser AW. Mental health of long-term survivors of childhood and young adult cancer: a systematic review. Int J Cancer 2018;143(6):1279-1286 [FREE Full text] [doi: 10.1002/ijc.31337] [Medline: 29468674]
17. Michel G, Mulder RL, van der Pal HJH, Skinner R, Bárdi E, Brown MC, et al. Evidence-based recommendations for the organization of long-term follow-up care for childhood and adolescent cancer survivors: a report from the PanCareSurFup Guidelines Working Group. J Cancer Surviv 2019;13(5):759-772. [doi: 10.1007/s11764-019-00795-5] [Medline: 31396878]

XSL•FO
RenderX

18. Essig S, Skinner R, von der Weid NX, Kuehni CE, Michel G. Follow-up programs for childhood cancer survivors in Europe: a questionnaire survey. PLoS One 2012;7(12):e53201 [FREE Full text] [doi: 10.1371/journal.pone.0053201] [Medline: 23300889]
19. Salz T, McCabe MS, Onstad EE, Baxi SS, Deming RL, Franco RA, et al. Survivorship care plans: is there buy-in from community oncology providers? Cancer 2014;120(5):722-730 [FREE Full text] [doi: 10.1002/cncr.28472] [Medline: 24327371]
20. Svedberg P, Einberg EL, Wärnestål P, Stigmar J, Castor A, Enskär K, et al. Support from healthcare services during transition to adulthood—experiences of young adult survivors of pediatric cancer. Eur J Oncol Nurs 2016;21:105-112. [doi: 10.1016/j.ejon.2016.02.008] [Medline: 26952685]
21. Brennan ME, Gormally JF, Butow P, Boyle FM, Spillane AJ. Survivorship care plans in cancer: a systematic review of care plan outcomes. Br J Cancer 2014;111(10):1899-1908 [FREE Full text] [doi: 10.1038/bjc.2014.505] [Medline: 25314068]
22. Ford JS, Tonorezos ES, Mertens AC, Hudson MM, Casillas J, Foster BM, et al. Barriers and facilitators of risk-based health care for adult survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. Cancer 2020;126(3):619-627 [FREE Full text] [doi: 10.1002/cncr.32568] [Medline: 31626337]
23. Linge HM, Follin C. Mixed methods assessment of impact on health awareness in adult childhood cancer survivors after viewing their personalized digital treatment summary and follow-up recommendations. BMC Cancer 2021;21(1):347 [FREE Full text] [doi: 10.1186/s12885-021-08051-9] [Medline: 33794815]
24. Haupt R, Essiaf S, Dellacasa C, Ronckers CM, Caruso S, Sugden E, et al. The 'Survivorship Passport' for childhood cancer survivors. Eur J Cancer 2018;102:69-81 [FREE Full text] [doi: 10.1016/j.ejca.2018.07.006] [Medline: 30138773]
25. Jankovic M, Haupt R, Spinetta JJ, Beck JD, Byrne J, Calaminus G, et al. Long-term survivors of childhood cancer: cure and care-the Erice Statement (2006) revised after 10 years (2016). J Cancer Surviv 2018;12(5):647-650. [doi: 10.1007/s11764-018-0701-0] [Medline: 29946794]
26. van Kalsbeek RJ, van der Pal HJH, Hjorth L, Winther JF, Michel G, Haupt R, et al. The European multistakeholder PanCareFollowUp project: novel, person-centred survivorship care to improve care quality, effectiveness, cost-effectiveness and accessibility for cancer survivors and caregivers. Eur J Cancer 2021;153:74-85. [doi: 10.1016/j.ejca.2021.05.030] [Medline: 34153717]
27. PanCareSurPass. 2021. URL: http://www.pancaresurpass.eu [accessed 2024-03-29]
28. Chronaki C, Cangioli G, Dellacasa C, Haupt R. SurPass-IPS: delivering on the social value of health data for childhood cancer survivors. HL7 Newsletter. 2019. URL: http://www.hl7.eu/download/eun-09-2019.pdf [accessed 2024-03-29]
29. Kasthurirathne SN, Mamlin B, Kumara H, Grieve G, Biondich P. Enabling better interoperability for HealthCare: lessons in developing a standards based application programing interface for electronic medical record systems. J Med Syst 2015;39(11):182. [doi: 10.1007/s10916-015-0356-6] [Medline: 26446013]
30. The International Patient Summary. 2023. URL: https://international-patient-summary.net [accessed 2024-03-29]
31. General Data Protection Regulation. 2022. URL: https://gdpr.eu [accessed 2024-03-29]
32. van den Oever SR, de Beijer IAE, Kremer LCM, Alfes M, Balaguer J, Bardi E, PanCareSurPass consortium. Barriers and facilitators to implementation of the interoperable Survivorship Passport (SurPass) v2.0 in 6 European countries: a PanCareSurPass online survey study. J Cancer Surviv 2023 Feb 20. [doi: 10.1007/s11764-023-01335-y] [Medline: 36808389]
33. Chronaki C, Charalambous E, Cangioli G, Schreier G, van den Oever S, van der Pal H, et al. Factors influencing implementation of the survivorship passport: the IT perspective. Stud Health Technol Inform 2022;293:161-168. [doi: 10.3233/SHTI220363] [Medline: 35592976]
34. Liang TP, Huang CW, Yeh YH, Lin B. Adoption of mobile technology in business: a fit‐viability model. Ind Manag Data Syst 2007;107(8):1154-1169. [doi: 10.1108/02635570710822796]
35. Nitzlnader M, Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. Stud Health Technol Inform 2014;198:211-218. [Medline: 24825705]
36. European Patient Identity Services. EUPID. URL: https://services.eupid.eu [accessed 2024-04-03]
37. CINECA. URL: https://www.cineca.it/ [accessed 2024-04-03]
38. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
39. Commission recommendation of 6.2.2019 on a European Electronic Health Record exchange format. European Commission. 2019. URL: https://digital-strategy.ec.europa.eu/en/policies/electronic-health-records [accessed 2024-03-29]
40. French-Mowat E, Burnett J. How are medical devices regulated in the European Union? J R Soc Med 2012 Apr;105 Suppl 1(Suppl 1):S22-S28 [FREE Full text] [doi: 10.1258/jrsm.2012.120036] [Medline: 22508970]
41. EU AI Act: first regulation on artificial intelligence. European Commission. 2023. URL: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2024-03-29]

Abbreviations

ATC: Anatomical Therapeutic Chemical

XSL•FO
RenderX

CDA: clinical document architecture
EHDS: European Health Data Space
EHIS: electronic health information system
EHR: electronic health record
ELGA: Austrian National Electronic Health Record Systems
EUPID: European Patient Identity Service
FHIR: Fast Healthcare Interoperability Resources
FVM: fit-viability model
GDPR: General Data Protection Regulation
HCP: health care provider
HDE: health data exchange
HL7: Health Level Seven
ICCC-3: International Classification for Childhood Cancer, 3rd Edition
ICD-O-3: International Classification of Diseases for Oncology, 3rd Edition
IPS: international patient summary
LTFU: long-term follow-up
OAuth: Open Authorization Standard
PCSP: PanCareSurPass
SCP: survivorship care plan
SurPass: Survivorship Passport

Edited by A Mavragani; submitted 13.06.23; peer-reviewed by W van Harten, B Schreiweis; comments to author 05.10.23; revised version received 24.10.23; accepted 28.02.24; published 02.05.24.

Please cite as:
de Beijer IAE, van den Oever SR, Charalambous E, Cangioli G, Balaguer J, Bardi E, Alfes M, Cañete Nieto A, Correcher M, Pinto da Costa T, Degelsegger-Márquez A, Düster V, Filbert AL, Grabow D, Gredinger G, Gsell H, Haupt R, van Helvoirt M, Ladenstein R, Langer T, Laschkolnig A, Muraca M, Pluijm SMF, Rascon J, Schreier G, Tomášikova Z, Trauner F, Trinkūnas J, Trunner K, Uyttebroeck A, Kremer LCM, van der Pal HJH, Chronaki C, PanCareSurPass Consortium
IT-Related Barriers and Facilitators to the Implementation of a New European eHealth Solution, the Digital Survivorship Passport (SurPass Version 2.0): Semistructured Digital Survey
J Med Internet Res 2024;26:e49910
URL: https://www.jmir.org/2024/1/e49910
doi:10.2196/49910
PMID:38696248

©Ismay A E de Beijer, Selina R van den Oever, Eliana Charalambous, Giorgio Cangioli, Julia Balaguer, Edit Bardi, Marie Alfes, Adela Cañete Nieto, Marisa Correcher, Tiago Pinto da Costa, Alexander Degelsegger-Márquez, Vanessa Düster, Anna-Liesa Filbert, Desiree Grabow, Gerald Gredinger, Hannah Gsell, Riccardo Haupt, Maria van Helvoirt, Ruth Ladenstein, Thorsten Langer, Anja Laschkolnig, Monica Muraca, Saskia M F Pluijm, Jelena Rascon, Günter Schreier, Zuzana Tomášikova, Florian Trauner, Justas Trinkūnas, Kathrin Trunner, Anne Uyttebroeck, Leontien C M Kremer, Helena J H van der Pal, Catherine Chronaki, PanCareSurPass Consortium. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Prevalence and Disparities in Telehealth Use Among US Adults Following the COVID-19 Pandemic: National Cross-Sectional Survey

Erin M Spaulding[1,2,3], RN, PhD; Michael Fang[3,4], MHS, PhD; Yvonne Commodore-Mensah[1,3,5], RN, MHS, PhD; Cheryl R Himmelfarb[1,4,5], RN, PhD; Seth S Martin[2,3,5,6], MHS, MD; Josef Coresh[3,4,7,8], MD, PhD

[1]Johns Hopkins University School of Nursing, Baltimore, MD, United States

[2]Digital Health Innovation Laboratory, Ciccarone Center for the Prevention of Cardiovascular Disease, Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

[3]Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[4]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[5]Center for Health Equity, Johns Hopkins University, Baltimore, MD, United States

[6]Johns Hopkins University Whiting School of Engineering, Baltimore, MD, United States

[7]Optimal Aging Institute, New York University Grossman School of Medicine, New York, NY, United States

[8]Division of Epidemiology, Department of Population Health, New York University Grossman School of Medicine, New York, NY, United States

Corresponding Author:
Erin M Spaulding, RN, PhD
Johns Hopkins University School of Nursing
525 North Wolfe Street
Baltimore, MD, 21205
United States
Phone: 1 4109554766
Email: espauld2@jhu.edu

Abstract

Background: Telemedicine expanded during the COVID-19 pandemic, though use differed by age, sex, race or ethnicity, educational attainment, income, and location. It is unclear if high telehealth use or inequities persisted late into the pandemic.

Objective: This study aims to evaluate the prevalence of, inequities in, and primary reasons for telehealth visits a year after telemedicine expansion.

Methods: We used cross-sectional data from the 2022 Health Information National Trends Survey (HINTS 6), the first cycle with data on telemedicine. In total, 4830 English- and Spanish-speaking US adults (aged ≥18 years) were included in this study. The primary outcomes were telehealth visit attendance in the 12 months before March 7, 2022, to November 8, 2022, and the primary reason for the most recent telehealth visit. We evaluated sociodemographic and clinical predictors of telehealth visit attendance and the primary reason for the most recent telehealth visit through Poisson regression. Analyses were weighted according to HINTS 6 standards.

Results: We included 4830 participants (mean age 48.3, SD 17.5 years; 50.28% women; 65.21% White). Among US adults, 38.78% reported having a telehealth visit in the previous year. Telehealth visit attendance rates were similar across age, race or ethnicity, income, and urban versus rural location. However, individuals with a telehealth visit were less likely to live in the Midwest (adjusted prevalence ratio [aPR] 0.65, 95% CI 0.54-0.77), and more likely to be women (aPR 1.21, 95% CI 1.06-1.38), college graduates or postgraduates (aPR 1.24, 95% CI 1.05-1.46), covered by health insurance (aPR 1.56, 95% CI 1.08-2.26), and married or cohabitating (aPR 1.17, 95% CI 1.03-1.32), adjusting for sociodemographic characteristics, frequency of health care visits, and comorbidities. Among participants with a telehealth visit in the past year, the primary reasons for their most recent visit were minor or acute illness (32.15%), chronic disease management (21%), mental health or substance abuse (16.94%), and an annual exam (16.22%). Older adults were more likely to report that the primary reason for their most recent telehealth visit was for chronic disease management (aPR 2.08, 95% CI 1.33-3.23), but less likely to report that it was for a mental health or substance abuse issue (aPR 0.19, 95% CI 0.10-0.35), adjusting for sociodemographic characteristics and frequency of health care visits.

XSL•FO
RenderX

Conclusions: Among US adults, telehealth visit attendance was high more than a year after telemedicine expansion and did not differ by age, race or ethnicity, income, or urban versus rural location. Telehealth could continue to be leveraged following COVID-19 to improve access to care and health equity.

(J Med Internet Res 2024;26:e52124) doi:10.2196/52124

KEYWORDS

telehealth; telemedicine; delivery of health care; health care disparities; COVID-19

Introduction

Access to health care, or the timely use of personal health services, is critical to achieving the best possible health outcomes [1,2]. However, many individuals face barriers to obtaining care, such as a lack of health insurance, poor access to transportation, and appointment times only being offered during typical work hours [2,3]. Telehealth has the potential to improve access to care in the United States. The Centers for Medicare and Medicaid Services define telehealth as real-time communication between a patient and a physician or practitioner through telecommunications equipment, including synchronous (both audio-video and audio-only) visits [4]. Previously, telemedicine coverage was largely limited to individuals living in rural areas and had to be conducted by video in a health care facility [5]. In response to COVID-19, on March 17, 2020, the Centers for Medicare and Medicaid Services temporarily expanded telemedicine, enabling beneficiaries to receive services outside a health care facility [6].

Studies using claims and electronic health record data demonstrated that telehealth use increased significantly from the prepandemic through the pandemic period [7-11]. While telemedicine expansion can improve access to health care, it also has the potential to exacerbate existing disparities among individuals with limited digital access and literacy. As such, digital access and literacy are increasingly being recognized as overarching social determinants of health [12]. Data from the 2018 American Community Survey (ACS) demonstrated that 26% of Medicare beneficiaries lacked digital access at home (ie, were without a computer or internet or a smartphone or data plan) and were likely unprepared for the dramatic shift in care toward telemedicine [13]. This was especially true among older adults, non-Hispanic Black and Hispanic adults, widowed individuals, adults with a high school education or less, adults with an income 200% below the federal poverty level, adults enrolled in Medicaid, and those with a disability [13].

Cross-sectional data between March 19, 2020, and March 24, 2020, from the Pew Research Center's American Trends Panel of US adults, immediately following the temporary expansion of telemedicine, demonstrated that approximately 17% of individuals reported using telehealth, with use being higher among racial and ethnic minorities [14]. The 2021 National Health Interview Survey (NHIS) data demonstrated that 37% of adults used telemedicine in the 12 months before January 2021-December 2021, during the early stages of COVID-19 [15]. Significant disparities in telehealth use were found, with older adults, women, non-Hispanic White and non-Hispanic American Indian or Alaska Native adults, adults with higher income and education, and adults living in urban areas and in the Northeast and West of the United States being more likely to have used telehealth in the past year [15]. Studies using claims and electronic health record data also found disparities in telehealth use during the early stages of the pandemic [8-11,16-18]. Understanding the prevalence and predictors of telehealth visit attendance nationally, beyond the first year of the COVID-19 pandemic, could aid in decision-making around the continued use of telehealth to improve access to care.

We used 2022 Health Information National Trends Survey (HINTS 6) data to evaluate the prevalence of and sociodemographic and clinical predictors of telehealth visit attendance in the 12 months before March 2022-November 2022 and the primary reasons for the most recent telehealth visits.

Methods

Data Source and Study Population

This study used publicly available cross-sectional data from the 2022 HINTS 6 data set administered by the National Cancer Institute. HINTS 6 is a nationally representative survey of English- and Spanish-speaking civilian, noninstitutionalized adults living in the United States. Participants could respond to the survey on paper or the web [19]. HINTS 6, conducted between March 7, 2022, and November 8, 2022, is the first survey cycle to ask questions about telehealth visits [19]. The sampling strategy consisted of a two-stage design: (1) a stratified sample of addresses was selected from a database of residential addresses in the United States, and (2) one adult was selected within each sampled household [19].

The sampling frame of addresses was grouped into four sampling strata: (1) high minority urban, (2) low minority urban, (3) high minority rural, and (4) low minority rural [19]. The high minority stratum was defined to include addresses in census tracts from the 2015-2019 ACS with at least a 34% population proportion of African Americans or Hispanics [19,20]. The rural or urban strata were based on the latest 2013 US Department of Agriculture, Economic Research Service, rural-urban continuum codes [19,21]. The overall weighted response rate, calculated using the Response Rate 4 formula of the American Association of Public Opinion Research [22], was 28% and differed by strata (high minority urban: 21%, low minority urban: 31%, high minority rural: 22%, and low minority rural: 29%) [19]. The high minority strata were oversampled, and within the low minority strata, the rural stratum was sampled at a higher rate than the urban stratum [19]. Survey weights were demographically calibrated using 2021 ACS estimates (age, sex, educational attainment, marital status, race, ethnicity, and census region) and HINTS 6–reported insurance and cancer

XSL•FO
RenderX

status [19]. We included participants in this analysis with data on telehealth use, sociodemographics, frequency of health care visits, and comorbidities.

Outcome Measures

All participants were asked "In the past 12 months, did you receive care from a doctor or health professional using telehealth?" to assess telehealth visit attendance from March 2022 to November 2022 [23]. A telehealth visit was defined as a "telephone or video appointment with a doctor or health professional" [23].

Participants with a telehealth visit in the past 12 months were asked "What was the primary reason for your most recent telehealth visit?" [23]. Response options include annual visit; minor illness or acute care (eg, fever or sinus infection); managing chronic condition or disease (eg, high blood pressure, diabetes, heart disease, obesity, or cancer); medical emergency; mental health, behavioral, or substance abuse issue (eg, depression, anxiety, or drug or alcohol abuse); or other [23].

Covariates

Covariates included age, sex, race or ethnicity, education, annual household income, health insurance, marital status, location, census region, number of times in the past 12 months the participant received care from a health professional (excluding emergency room visits), and comorbidities (diabetes; hypertension; heart condition [heart attack, angina, or heart failure]; chronic lung disease, asthma, emphysema, or chronic bronchitis; and depression or anxiety disorder). Education level was categorized as high school graduate or less, some college or vocational or technical school, or college graduate or postgraduate. Annual household income was categorized as ≤US $34,999, US $35,000-US $74,999, or ≥US $75,000. Health insurance was recorded as being covered or not covered by any kind of health insurance plan. Marital status was categorized as married or cohabitating or divorced, widowed, separated, or single as a proxy for social support. Location was categorized as urban (metropolitan) or rural (nonmetropolitan). The census region was categorized as Northeast, Midwest, South, or West. The number of times the participant received care from a health professional in the past 12 months was dichotomized into less than 5 times and 5 times or more.

Statistical Analysis

We estimated the proportion of US adults with a telehealth visit in the past 12 months and the primary reason for their most recent telehealth visit. We used recommended methods for HINTS 6, including survey weights and the delete one jackknife replication method, to account for the complex survey design and generate nationally representative estimates [19,24]. This jackknife variance estimation technique deleted one primary sampling unit at a time from the full sample to create a set of 50 replicate weights [19,24]. We examined prevalence ratios of sociodemographic and clinical predictors of telehealth visit attendance in the past 12 months and their primary reason for their most recent telehealth visit using generalized linear models with a Poisson distribution and logarithmic link. Prevalence ratios are the preferred statistical estimate for cross-sectional studies when the outcome or outcomes are not rare [25]. The reported percentages are weighted and the sample sizes are unweighted. Analyses were conducted using Stata (version 18.0; StataCorp).

Ethical Considerations

HINTS 6 (2022) was given a nonhuman subjects research designation from the National Institutes of Health Office of Human Subjects Research Protections on August 16, 2021. It was reviewed and approved by the Westat Institutional Review Board on May 10, 2021, with a subsequent amendment reviewed and approved on November 24, 2021 (6632.03.51). Thus, informed consent was not required for the primary data collection. This analysis using HINTS 6 data also met criteria for nonhuman subject research by the Johns Hopkins University School of Medicine institutional review board. Thus, ethical review and approval were not needed for this study because it involves secondary data analysis of existing, deidentified, publicly available data.

Results

This study included 4830 participants, of which the mean age was 48.3 (SD 17.5) years, 50.28% (2886/4830) were women, 65.21% (2895/4830) were non-Hispanic White adults, 39.51% (1367/4830) had some college or vocational or technical schooling, 45.88% (1991/4830) had an annual household income of at least US $75,000, 89.59% (4437/4830) had health insurance, 56.88% (2573/4830) were married or cohabitating, 88.23% (4225/4830) lived in an urban location, 38.4% (2176/4830) lived in the South, and 78.25% (3589/4830) had less than 5 visits with a health professional in the previous 12 months. Survey-weighted characteristics are provided in Table 1.

XSL•FO
RenderX

Table 1. Predictors of telehealth visits in 12 months before Health Information National Trends Survey (HINTS 6; March 2022-November 2022) among weighted US adults (N=4830). Weights are calibrated using data from the 2021 American Community Survey (age, sex, educational attainment, marital status, race, ethnicity, and census region) conducted by the US Census Bureau.

Sociodemographic characteristics	US adults weighted % (95% CI)[a]	Telehealth visit in past 12 months, weighted % (95% CI)[b]	Unadjusted PR[c] (95% CI)	Model 1 adjusted PR (95% CI)[d]	Model 2 adjusted PR (95% CI)[e]
Overall	N/A[f]	38.78 (36.68-40.91), n=1986	N/A	N/A	N/A
Age (years)					
18-44	42.70 (40.91-44.51)	37.9 (34.03-41.93)	1 (reference)	1 (reference)	1 (reference)
45-64	37.47 (35.69-39.29)	40.75 (37.21-44.38)	1.08 (0.94-1.23)	1.04 (0.91-1.19)	1.02 (0.89-1.18)
≥65	19.83 (19.06-20.62)	36.93 (33.71-40.27)	0.97 (0.83-1.14)	0.90 (0.76-1.06)	0.87 (0.73-1.03)
Sex					
Male	49.72 (48.44-51)	33.47 (30.42-36.67)	1 (Reference)	1 (Reference)	1 (Reference)
Female	50.28 (49-51.56)	44.02 (40.83-47.26)	1.32 (1.16-1.49)[g]	1.29 (1.14-1.47)[g]	1.21 (1.06-1.38)[h]
Race or ethnicity					
Non-Hispanic White	65.21 (64.36-66.05)	39.68 (36.79-42.65)	1 (Reference)	1 (Reference)	1 (Reference)
Non-Hispanic Black	11.48 (10.9-12.08)	33.66 (28.18-39.62)	0.85 (0.70-1.03)	0.89 (0.73-1.07)	0.98 (0.82-1.17)
Hispanic	17.24 (16.47-18.04)	38.84 (33.55-44.41)	0.98 (0.83-1.16)	1 (0.83-1.22)	1.11 (0.92-1.34)
Non-Hispanic Asian	6.08 (5.53-6.68)	38.52 (26.41-52.26)	0.97 (0.68-1.38)	0.89 (0.65-1.23)	1.05 (0.77-1.42)
Education					
High school graduate or less	26.87 (25.13-28.68)	33.79 (29.68-38.17)	1 (reference)	1 (reference)	1 (reference)
Some college or vocational or technical school	39.51 (37.93-41.12)	37.22 (33.58-41.01)	1.10 (0.94-1.30)	1.09 (0.92-1.29)	1.09 (0.91-1.30)
College graduate or postgraduate	33.62 (32.79-34.46)	44.59 (41.45-47.77)	1.32 (1.15-1.52)[g]	1.22 (1.03-1.44)[h]	1.24 (1.05-1.46)[h]
Annual household income (US $)					
≤34,999	23.75 (21.69-25.94)	36.13 (31.24-41.32)	1 (reference)	1 (reference)	1 (reference)
35,000-74,999	30.37 (28.26-32.56)	34.79 (30.84-38.96)	0.96 (0.80-1.16)	0.92 (0.76-1.13)	1.01 (0.84-1.21)
≥75,000	45.88 (43.51-48.27)	42.78 (38.85-46.8)	1.18 (0.99-1.42)	1 (0.81-1.23)	1.13 (0.92-1.38)
Health insurance					
No	10.41 (9.57-11.32)	22.15 (15.21-31.1)	1 (reference)	1 (reference)	1 (reference)
Yes	89.59 (88.68-90.43)	40.71 (38.31-43.15)	1.84 (1.25-2.70)[h]	1.73 (1.16-2.59)[h]	1.56 (1.08-2.26)[h]
Marital status					
Divorced, widowed, separated, or single	43.12 (41.82-44.44)	35.39 (31.67-39.29)	1 (reference)	1 (reference)	1 (reference)
Married or cohabitating	56.88 (55.56-58.18)	41.34 (38.8-43.94)	1.17 (1.03-1.33)[h]	1.12 (0.98-1.27)	1.17 (1.03-1.32)[h]
Location					
Urban	88.23 (87.34-89.07)	39.21 (37.05-41.42)	1 (reference)	1 (reference)	1 (reference)
Rural	11.77 (10.93-12.66)	35.49 (30.09-41.29)	0.91 (0.77-1.06)	0.97 (0.81-1.15)	0.93 (0.80-1.10)
Census region					
Northeast	17.55 (16.69-18.44)	45.16 (39.7-50.73)	1 (reference)	1 (reference)	1 (reference)
Midwest	21.8 (20.87-22.75)	27.9 (24.16-31.98)	0.62 (0.51-0.74)[g]	0.63 (0.52-0.76)[g]	0.65 (0.54-0.77)[g]
South	38.4 (37.37-39.43)	37.63 (34.11-41.29)	0.83 (0.71-0.98)[h]	0.88 (0.76-1.03)	0.87 (0.74-1.01)

XSL•FO
RenderX

Sociodemographic characteristics	US adults weighted % (95% CI)[a]	Telehealth visit in past 12 months, weighted % (95% CI)[b]	Unadjusted PR[c] (95% CI)	Model 1 adjusted PR (95% CI)[d]	Model 2 adjusted PR (95% CI)[e]
West	22.26 (21.33-23.21)	46.36 (41.76-51.02)	1.03 (0.89-1.19)	1.06 (0.91-1.23)	1.05 (0.89-1.23)
Frequency of visits with health professional					
<5	78.25 (76.18-80.2)	33.56 (30.99-36.23)	1 (reference)	N/A	1 (reference)
≥5	21.75 (19.8-23.82)	57.55 (53.59-61.42)	1.72 (1.54-1.91)[g]	N/A	1.47 (1.32-1.63)[g]
Diabetes					
No	83.63 (82.01-85.13)	37.36 (35.21-39.56)	1 (reference)	N/A	1 (reference)
Yes	16.37 (14.87-17.99)	45.99 (40.18-51.91)	1.23 (1.07-1.41)[h]	N/A	1.09 (0.96-1.25)
Hypertension					
No	63.51 (61.34-65.63)	36.57 (33.53-39.73)	1 (reference)	N/A	1 (reference)
Yes	36.49 (34.37-38.66)	42.61 (38.77-46.55)	1.17 (1.01-1.34)[h]	N/A	1.12 (0.97-1.31)
Heart condition					
No	92.94 (91.84-93.9)	37.87 (35.67-40.11)	1 (reference)	N/A	1 (reference)
Yes	7.06 (6.1-8.16)	50.73 (42.97-58.45)	1.34 (1.14-1.58)[h]	N/A	1.20 (1.02-1.42)[h]
Lung disease					
No	88.51 (87.17-89.74)	37.41 (34.98-39.92)	1 (reference)	N/A	1 (reference)
Yes	11.49 (10.26-12.83)	49.27 (44.13-54.43)	1.32 (1.15-1.51)[g]	N/A	1.06 (0.91-1.23)
Depression or anxiety					
No	71.92 (69.55-74.18)	31.75 (29.48-34.1)	1 (reference)	N/A	1 (reference)
Yes	28.08 (25.82-30.45)	56.77 (52.16-61.27)	1.79 (1.60-2)[g]	N/A	1.65 (1.47-1.86)[g]

[a]US adults' characteristics are weighted column percentages.

[b]Telehealth visit in past 12 months by sociodemographic characteristics are weighted row percentages.

[c]PR: prevalence ratio.

[d]Model 1: adjusting for age, sex, race or ethnicity, education, income, health insurance, marital status, location, and census region.

[e]Model 2: adjusting for age, sex, race or ethnicity, education, income, health insurance, marital status, location, census region, frequency of visits with a health professional, and comorbidities.

[f]Not applicable.

[g]$P<.001$.

[h]$P<.05$.

The weighted prevalence of telehealth visit attendance in the past 12 months was 38.78% (1986/4830, 95% CI 36.68%-40.91%). Telehealth visit attendance in the past 12 months was similar by age (aged between 18 and 44 years: 634/1456, 37.9%; aged between 45 and 64 years: 729/1738, 40.75%; ≥65 years: 623/1636, 36.93%), race or ethnicity (non-Hispanic White: 1198/2895, 39.68%; non-Hispanic Black: 298/787, 33.7%; Hispanic: 393/885, 38.8%; non-Hispanic Asian: 97/263, 38.5%), annual household income (≤US $34,999: 517/1367, 36.13%; US $35,000-US $74,999: 570/1472, 34.79%; ≥US $75,000: 899/1991, 42.78%), and location (urban: 1782/4225, 39.21%; rural: 204/605, 35.5%). However, individuals who reported having a telehealth visit in the past 12 months were less likely to live in the Midwest (266/834, 22.9% vs 312/715, 45.2% Northeast; adjusted prevalence ratio [aPR] 0.65, 95% CI 0.54-0.77), and more likely to be women (1276/2886, 44.02% vs 710/1944, 33.47% men; aPR 1.21, 95% CI 1.06-1.38), college graduates or postgraduates (1081/2362, 44.59% vs 362/1101, 33.79% high school graduate or less; aPR 1.24, 95% CI 1.05-1.46), covered by health insurance (1872/4437, 41% vs 114/393, 22.2% not covered; aPR 1.56, 95% CI 1.08-2.26), married or cohabitating (1096/2573, 41.34% vs 890/2257, 35.4% divorced, widowed, separated, or single; aPR 1.17, 95% CI 1.03-1.32), have 5 or more visits in the previous 12 months with a health professional (700/1241, 57.55% vs 1286/3589, 33.56% with less than 5 visits; aPR 1.47, 95% CI 1.32-1.63), have a heart condition (224/451, 50.7% vs 1762/4379, 37.87%; aPR 1.20, 95% CI 1.02-1.42), and have depression or an anxiety disorder (753/1317, 56.77% vs 1233/3513, 31.75%; aPR 1.65, 95% CI 1.47-1.86), after adjusting for sociodemographic characteristics, frequency of visits with a health professional, and comorbidities (Table 1). While individuals with diabetes, hypertension, and lung disease were more likely to have reported a telehealth visit in the past 12 months in unadjusted models, this did not hold true after adjusting for other sociodemographic characteristics, frequency

XSL•FO
RenderX

of visits with a health professional, and comorbidities. Similarly, adults living in the South were less likely to have reported a telehealth visit in the past 12 months in the unadjusted model but not in the adjusted models.

The primary reasons for participants' most recent telehealth visit in the past 12 months (Figure 1; N=1878) were for a minor illness or acute care (510/1878, 32.15%, 95% CI 28.1%-36.49%), managing a chronic disease (462/1878, 21%, 95% CI 18.27%-24.03%), mental health, behavioral or substance abuse issues (288/1878, 16.94%, 95% CI 14.15%-20.16%), annual visits (352/1878, 16.22%, 95% CI 13.62%-19.2%), and other reasons (266/1878, 13.68%, 95% CI 11.26%-16.53%). Individuals who reported that the primary reason for their most recent telehealth visit in the past year was for an annual visit were more likely to be aged 65 years and older (158/560, 28.3% vs 93/622, 14.5% aged between 18 and 44 years; aPR 2.16, 95% CI 1.52-3.08) and non-Hispanic Black adults (83/271, 28.2% vs 189/1145, 15.39% White; aPR 1.96, 95% CI 1.28-3), and less likely to be women (195/1206, 13.6% vs 157/672, 19.7% men; aPR 0.68, 95% CI 0.50-0.93) and have had 5 or more visits with a health professional in the previous 12 months (85/673, 9.2% vs 267/1205, 19.6% less than 5 visits; aPR 0.44, 95% CI 0.32-0.59). No differences were found by education, annual household income, health insurance, marital status, location, or census region in the adjusted model. Individuals who reported that the primary reason for their most recent telehealth visit in the past year was for a minor illness or acute care were more likely to be married or cohabitating (322/1053, 36.4% vs 188/825, 25.4% divorced, widowed, separated, or single; aPR 1.30, 95% CI 1-1.68), and less likely to be non-Hispanic Black adults (51/271, 20.9% vs 321/1145, 33.2% White; aPR 0.62, 95% CI 0.41-0.95) and living in a rural location (50/193, 22.1% vs 460/1685, 33.37% urban; aPR 0.64, 95% CI 0.42-0.97). No differences were found by age, sex, education, annual household income, health insurance, census region, or frequency of visits with a health professional in the previous 12 months in the adjusted model.

Figure 1. Prevalence of the primary reasons for telehealth visits in the 12 months before the Health Information National Trends Survey (HINTS 6; March 2022-November 2022) among weighted US adults (n=1878). Estimates are based on HINTS 6 data fielded from March through November 2022. Question text: "What was the primary reason for your most recent telehealth visit?" Response options (select 1): annual visit, minor illness or acute care (eg, fever or sinus infection), managing my chronic health condition or disease (eg, high blood pressure, diabetes, heart disease, obesity, or cancer), mental health, behavioral or substance abuse issues (eg, depression, anxiety, or drug or alcohol abuse), medical emergency, or other. This question was asked of participants who reported having a telehealth visit in the past 12 months.

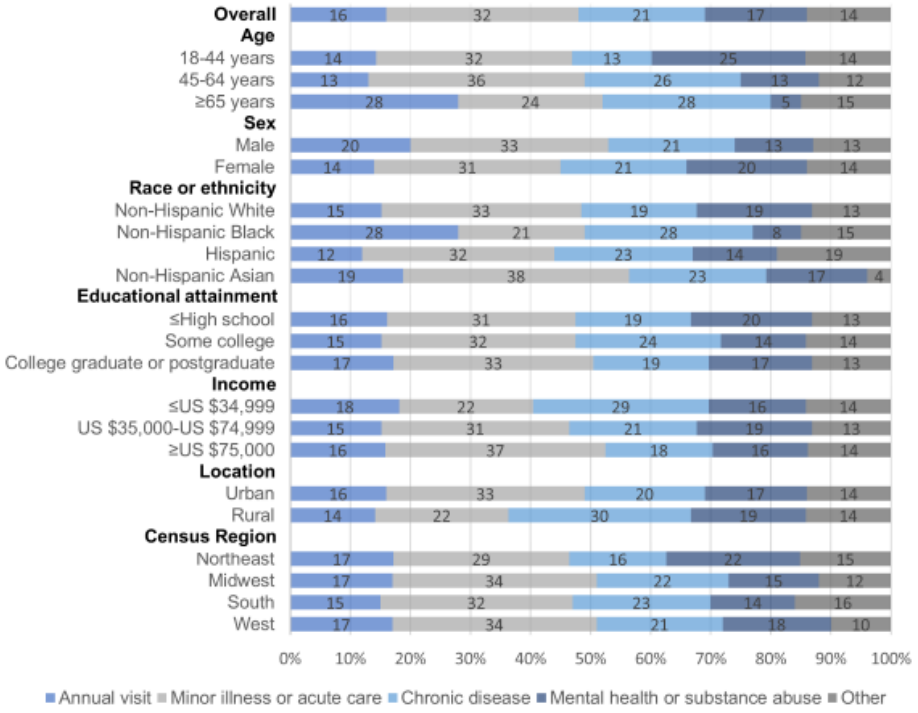

Individuals who reported that the primary reason for their most recent telehealth visit was for chronic disease management were more likely to be aged between 45 and 64 years and 65 years and older (192/696, 26.1% and 184/560, 28.4% vs 86/622, 13.3% aged between 18 and 44 years; aPR 1.93, 95% CI 1.27-2.93 and aPR 2.08, 95% CI 1.33-3.23) and to have had 5 or more visits with a health professional in the previous 12 months (229/673, 29.9% vs 233/1205, 16.72% less than 5 visits; aPR 1.67, 95% CI 1.23-2.26), and less likely to have an annual household income of ≥US $75,000 (188/875, 17.6% vs 141/466, 29.5% ≤US $34,999; aPR 0.60, 95% CI 0.37-0.98). No differences were found by sex, race or ethnicity, education, health insurance, marital status, location, or census region in the adjusted model. Individuals who reported that the primary reason for their most recent telehealth visit was for mental health, behavioral, or substance abuse issues were less likely to be aged between 45 and 64 years and 65 years and older (110/696, 13.2% and 29/560, 4.7% vs 149/622, 25.5% aged between 18 and 44 years; aPR 0.61, 95% CI 0.45-0.82 and aPR 0.19, 95% CI 0.10-0.35), non-Hispanic Black and Hispanic (24/271, 7.9% and 47/367, 13.5% vs 208/1145, 19.09% White; aPR 0.42, 95% CI 0.24-0.73 and aPR 0.61, 95% CI 0.37-0.99),

XSL•FO
RenderX

and married or cohabitating (131/1053, 11.33% vs 157/825, 25.9% divorced, widowed, separated, or single; aPR 0.44, 95% CI 0.28-0.67). No differences were found by sex, education, annual household income, health insurance, location, census region, or frequency of visits with a health professional in the previous 12 months in the adjusted model. Multimedia Appendix 1 shows full details on the predictors of primary reasons for telehealth visits in the 12 months before HINTS 6.

Discussion

Summary of Findings

In this nationally representative sample, 38.78% (1986/4830) of US adults had a telehealth visit in the 12 months before March 2022-November 2022. Similarly, 2021 NHIS data demonstrated that 37% of adults had a telehealth visit in the 12 months before January 2021-December 2021 [15]. Thus, telehealth use has remained steady over the past year following telemedicine expansion. While 2021 NHIS data demonstrated significant disparities in telehealth visits by age, race or ethnicity, income, and location [15], this analysis did not.

The 2021 NHIS data demonstrated that young adults were less likely to have a telemedicine visit in the past year [15]. While telehealth visit attendance was similar by age in this analysis, older adults were more likely to report that the primary reason for their most recent telehealth visit was for an annual visit or for chronic disease management. Meanwhile, young adults were more likely to report that their primary reason was mental health, behavioral, or substance abuse issue. The high prevalence of telehealth use for mental health, behavioral, and substance abuse issues among young adults during COVID-19 may explain the narrowing of telehealth disparities by age since 2021. Claims data from January 2020 to June 2020 demonstrated that 57% of psychiatry visits were conducted through telemedicine [26].

The 2021 NHIS data demonstrated that individuals with a lower annual income, living in more rural areas, and non-Hispanic Black, non-Hispanic Asian, and Hispanic adults were less likely to have a telehealth visit in the past year [15]. In this analysis, there were no significant differences in telehealth visit attendance by income, location, or race or ethnicity. COVID-19 restrictions may have forced people of all socioeconomic statuses to move toward telehealth. Additionally, while telehealth was originally intended for individuals living in rural areas, it may also be an important modality for increasing access to care among individuals living in urban settings. While disparities in telehealth use may be narrowing, the impact of telehealth use on patient outcomes remains unclear [27]. Other studies have evaluated national telehealth use [28,29]; however, we primarily compared our findings to 2021 NHIS data due to its similar methodology (ie, it is a nationally representative survey asking about telehealth use in the previous 12 months).

Our findings largely align with those of other studies. For example, the US Census Bureau's Household Pulse Survey data, from April 14, 2021, through August 8, 2022, found that 23% of adult respondents (N=1,180,248) reported having a telehealth visit within the last 4 weeks [29]. In this study, we observed a higher prevalence of participants having a telehealth visit. It is important to note that HINTS 6 evaluated telehealth use over 12 months, whereas the Household Pulse Survey assessed it over a shorter period of 4 weeks [28,29]. The Pulse Survey found that those who self-reported as Hispanic or Latino, Black, and multiracial or other; aged less than 65 years old; making less than US $100,000 per year; having a bachelor's degree or higher; with Medicare and Medicaid; and with other insurance types were more likely to have had a telehealth visit [29]. However, individuals who were uninsured and from the Northeast, Midwest, and South regions of the United States were less likely to have had a telehealth visit [29]. In our analysis, we did not find any disparities in telehealth use by age, race or ethnicity, or income. The Pulse Survey actually found that telehealth use may be higher in individuals that have a lower annual household income or are from underrepresented racial and ethnic groups [29]. However, unlike our findings, the Pulse Survey demonstrated that individuals aged 65 years and older were less likely to have reported having a telehealth visit over the previous 4 weeks [29].

Similar to our findings, the Pulse Survey also found that individuals living in the Midwest were less likely to have had a telehealth visit, and individuals with health insurance were more likely to have had a telehealth visit than adults who were uninsured [29]. The Pulse Survey found that approximately 9% of uninsured adults had a telehealth visit in the previous 4 weeks [29] while in HINTS 6, 22% of uninsured adults had a telehealth visit in the past 12 months. This difference may be a reflection of the length of the assessment periods between the 2 studies. Almost a quarter of the uninsured participants in this analysis had a telehealth visit in the past 12 months. While lower than insured individuals, the number of uninsured adults with telehealth visits was still fairly high. Telehealth may present a cheaper option for uninsured individuals than in-person care, saving them money on travel and loss of wages.

The Pulse Survey also evaluated predictors of video versus audio-only telehealth visits from July 21, 2021, to August 8, 2022 [28,29]. Video telehealth usage was lower among Hispanic or Latino, Black, and Asian adults; men; older adults (aged ≥65 years); and adults with less than a bachelor's degree and an annual household income of less than US $100,000 [29]. While we did not evaluate predictors of video versus audio-only telehealth visits in this analysis, we will explore them in future work. There may be some instances when audio-only telehealth visits, depending on the reason for the telehealth visit, are not sufficient and result in lower quality care than an in-person visit or a video telehealth visit. Patient outcomes should be compared by health problem across in-person, video telehealth, and audio-only telehealth visits in the future.

Having a telehealth visit in the past 12 months for a minor illness or acute care (510/1878, 32.15%) was the most common primary reason for participants' most recent telehealth visit, followed by managing a chronic disease (462/1878, 21%), mental health, behavioral, or substance abuse issues (288/1878, 16.94%), and an annual visit (352/1878, 16.22%). A review of large-scale claims studies published during the first year of the pandemic found that telehealth use varied widely across specialties and diagnoses, but the highest levels were seen for behavioral health [30]. Meanwhile, a 2020 analysis of Doximity data demonstrated

XSL•FO
RenderX

that the top 10 specialties using telemedicine largely treated chronic illnesses [31]. In addition to supporting chronic disease management and mental or behavioral health, telemedicine may play an important role in addressing minor or acute illnesses (eg, fever or sinus infection) in the future. In this study, we also found that individuals, especially those who cited that the primary reason for their most recent telehealth visit was for chronic disease management, with a higher frequency of visits with a health professional, may be more likely to engage in telehealth.

In this analysis, middle-aged and older adults were less likely to report that the primary reason for their most recent telehealth visit was for a mental health, behavioral, or substance abuse issue (110/696, 13.2% aged between 44 and 64 years and 29/560, 4.7% aged ≥65 years, vs 149/622, 25.5% aged between 18 and 44 years). However, according to 2019 NHIS data, the prevalence of experiencing any level of depressive symptoms is similar across age groups (approximately 18%) [32]. While mental health issues are prevalent in older adults, this population may be more likely to schedule telehealth visits for chronic disease management. In this analysis, non-Hispanic Black and Hispanic adults were also less likely than non-Hispanic White adults to report that the primary reason for their most recent telehealth visit was due to a mental health, behavioral, or substance abuse issue. However, 2019 NHIS data also demonstrated that the prevalence of experiencing any level of depressive or anxiety symptoms was similar across these 3 racial or ethnic groups [32]. Additional research is needed to examine the reason for these disparities in telehealth use for mental health, behavioral, and substance abuse issues.

Limitations and Strengths

Limitations include the inability to determine causation; the potential for selection bias due to the HINTS 6 low response rate (28% [7], consistent with declining rates across many national surveys [33]) and recall bias, the combining of audio-only and audio-video telehealth visits, and not having data on health insurance plan type. However, the survey weighting was calibrated to adjust for demographic variables by response rate. We will explore predictors of audio-only and audio-video telehealth visits separately in future work. Despite these limitations, HINTS 6 is the first national study to ask about telehealth use over 12 months during the latter stages of COVID-19. Furthermore, this study also looked beyond the prevalence of telehealth visits to the prevalence and predictors of participants' primary reasons for their most recent telehealth visit.

Conclusions

Telehealth visit attendance, over a year following telemedicine expansion in response to COVID-19, remained high (1986/4830, 38.78%). Disparities in telehealth visit attendance by age, race or ethnicity, income, and location may be less pronounced than at the beginning of COVID-19. There may be continued opportunities to leverage telehealth following the pandemic to increase access to care and promote health equity.

Acknowledgments

SSM reports support from the American Heart Association (20SFRN35380046, 20SFRN35490003, #878924, #882415, and #946222), the Patient-Centered Outcomes Research Institute (ME-2019C1-15 328 and IHS-2021C3-24147), the National Institutes of Health (P01HL108800 and R01AG071032), the David and June Trone Family Foundation, the Pollin Digital Innovation Fund, Sandra and Larry Small, CASCADE FH, Google, Amgen, and Merck. EMS reports support from the American Heart Association (20SFRN35380046 and #878924) and the National Institutes of Health (U01HL096812). MF was supported by NIH/NIDDK grant K01-DK138273.

Data Availability

The HINTS 6 data set analyzed in this study are made publicly available by the National Cancer Institute [34].

Authors' Contributions

EMS had full access to all of the data in the study, took responsibility for the integrity of the data and the accuracy of the data analysis, and was also responsible for statistical analysis. EMS, MF, SSM, and JC were responsible for concept and design, acquisition, analysis, and interpretation of the data. All the authors were responsible for drafting the manuscript and critical revision of the manuscript for important intellectual content.

Conflicts of Interest

Under a license agreement between Corrie Health and Johns Hopkins University, the university owns equity in Corrie Health. The university and SSM are entitled to royalty distributions related to Corrie Health. Additionally, SSM is a cofounder of and holds equity in Corrie Health. This arrangement has been reviewed and approved by Johns Hopkins University in accordance with its conflict-of-interest policies. SSM has also received research and material support from Apple and iHealth. Furthermore, SSM is on the Advisory Board for Care Access and reports personal consulting fees from Amgen, AstraZeneca, BMS, Chroma, Kaneka, Merck, NewAmsterdam, Novartis, Novo Nordisk, Premier, Sanofi, and 89bio. EMS reports personal consulting fees from Corrie Health. All other authors declare no conflicts of interest.

XSL•FO
RenderX

Multimedia Appendix 1
Predictors of primary reasons for telehealth visits in 12 months before Health Information National Trends Survey 6 (March 2022-November 2022) among weighted US adults (n=1878).
[PDF File (Adobe PDF File), 115 KB - jmir_v26i1e52124_app1.pdf]

References

1. Institute of Medicine (US). Committee on monitoring access to personal health care services. In: Millman M, editor. Access to Health Care in America. Washington (DC): National Academies Press (US); 1993.
2. Access to health services. Healthy People 2030. URL: https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/access-health-services#cit1 [accessed 2024-02-02]
3. Institute of Medicine (US). Committee on understanding and eliminating racial and ethnic disparities in health care. In: Smedley BD, Stith AY, Nelson AR, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington (DC): National Academies Press (US); 2003.
4. Telehealth resource center: definitions. American Medical Association. Philadelphia, PA; 2023. URL: https://www.ama-assn.org/practice-management/digital/telehealth-resource-center-definitions [accessed 2023-10-17]
5. Mehrotra A, Huskamp HA, Souza J, Uscher-Pines L, Rose S, Landon BE, et al. Rapid growth in mental health telemedicine use among rural medicare beneficiaries, wide variation across states. Health Aff (Millwood) 2017;36(5):909-917 [FREE Full text] [doi: 10.1377/hlthaff.2016.1461] [Medline: 28461359]
6. CDC museum COVID-19 timeline. Centers for Medicare and Medicaid Services. URL: https://www.cdc.gov/museum/timeline/covid19.html [accessed 2023-07-26]
7. Cantor JH, McBain RK, Pera MF, Bravata DM, Whaley CM. Who is (and is not) receiving telemedicine care during the COVID-19 pandemic. Am J Prev Med 2021;61(3):434-438 [FREE Full text] [doi: 10.1016/j.amepre.2021.01.030] [Medline: 33781622]
8. Friedman AB, Gervasi S, Song H, Bond AM, Chen AT, Bergman A, et al. Telemedicine catches on: changes in the utilization of telemedicine services during the COVID-19 pandemic. Am J Manag Care 2022;28(1):e1-e6 [FREE Full text] [doi: 10.37765/ajmc.2022.88771] [Medline: 35049260]
9. Xu P, Hudnall M, Zhao S, Raja U, Parton J, Lewis D. Pandemic-triggered adoption of telehealth in underserved communities: descriptive study of pre- and postshutdown trends. J Med Internet Res 2022;24(7):e38602 [FREE Full text] [doi: 10.2196/38602] [Medline: 35786564]
10. Pagán VM, McClung KS, Peden CJ. An observational study of disparities in telemedicine utilization in primary care patients before and during the COVID-19 pandemic. Telemed J E Health 2022;28(8):1117-1125. [doi: 10.1089/tmj.2021.0412] [Medline: 34935517]
11. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
12. Benda NC, Veinot TC, Sieck CJ, Ancker JS. Broadband internet access is a social determinant of health!. Am J Public Health 2020;110(8):1123-1125. [doi: 10.2105/AJPH.2020.305784] [Medline: 32639914]
13. Roberts ET, Mehrotra A. Assessment of disparities in digital access among medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020;180(10):1386-1389 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]
14. Campos-Castillo C, Anthony D. Racial and ethnic differences in self-reported telehealth use during the COVID-19 pandemic: a secondary analysis of a US survey of internet users from late March. J Am Med Inform Assoc 2021;28(1):119-125 [FREE Full text] [doi: 10.1093/jamia/ocaa221] [Medline: 32894772]
15. Lucas JW, Villarroel MA. Telemedicine use among adults: United States, 2021. NCHS Data Brief 2022(445):1-8 [FREE Full text] [Medline: 36255940]
16. Chunara R, Zhao Y, Chen J, Lawrence K, Testa PA, Nov O, et al. Telemedicine and healthcare disparities: a cohort study in a large healthcare system in New York City during COVID-19. J Am Med Inform Assoc 2021;28(1):33-41 [FREE Full text] [doi: 10.1093/jamia/ocaa217] [Medline: 32866264]
17. Qian AS, Schiaffino MK, Nalawade V, Aziz L, Pacheco FV, Nguyen B, et al. Disparities in telemedicine during COVID-19. Cancer Med 2022;11(4):1192-1201 [FREE Full text] [doi: 10.1002/cam4.4518] [Medline: 34989148]
18. Adepoju OE, Chae M, Ojinnaka CO, Shetty S, Angelocci T. Utilization gaps during the COVID-19 pandemic: racial and ethnic disparities in telemedicine uptake in federally qualified health center clinics. J Gen Intern Med 2022;37(5):1191-1197 [FREE Full text] [doi: 10.1007/s11606-021-07304-4] [Medline: 35112280]
19. Health Information National Trends Survey 6 (HINTS 6) methodology report. Westat. 2023. URL: https://hints.cancer.gov/docs/methodologyreports/HINTS_6_MethodologyReport.pdf [accessed 2023-07-26]
20. 2015-2019 ACS 5-year estimates. United States Census Bureau. URL: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2019/5-year.html [accessed 2023-10-17]
21. Rural-urban continuum codes. Economic Research Service, United Stated Department of Agriculture. URL: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx [accessed 2023-10-17]

XSL•FO
RenderX

22. Standard definitions: final dispositions of case codes and outcome rates for surveys. The American Association for Public Opinion Research. 2023. URL: https://aapor.org/wp-content/uploads/2023/05/Standards-Definitions-10th-edition.pdf [accessed 2023-10-20]
23. Health information national trends survey. National Institutes of Health. URL: https://hints.cancer.gov/docs/Instruments/HINTS6-AnnotatedEnglishInstrument.pdf [accessed 2023-07-26]
24. Rust K. Variance estimation for complex estimators in sample surveys. J Off Stat 1985;1(4):381-397 [FREE Full text]
25. Petersen MR, Deddens JA. A comparison of two methods for estimating prevalence ratios. BMC Med Res Methodol 2008;8:9 [FREE Full text] [doi: 10.1186/1471-2288-8-9] [Medline: 18307814]
26. Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Variation in telemedicine use and outpatient care during the COVID-19 pandemic in the United States. Health Aff (Millwood) 2021;40(2):349-358 [FREE Full text] [doi: 10.1377/hlthaff.2020.01786] [Medline: 33523745]
27. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
28. Karimi M, Lee EC, Couture SJ, Gonzales A, Grigorescu V, Smith SR, et al. National survey trends in telehealth use in 2021: disparities in utilization and audio vs. video services. US Department of Health and Human Services. 2022. URL: https://aspe.hhs.gov/sites/default/files/documents/4e1853c0b4885112b2994680a58af9ed/telehealth-hps-ib.pdf [accessed 2023-11-28]
29. Adams EC, Adams N, Adams N, Adams SR, Samson LW, Adams AB, et al. Updated national survey trends in telehealth utilization and modality (2021-2022). U. S. Department of Health and Human Services. 2023. URL: https://connectwithcare.org/wp-content/uploads/2023/04/household-pulse-survey-telehealth-covid-ib.pdf [accessed 2023-11-20]
30. Harju A, Neufeld J. Telehealth utilization during the COVID-19 pandemic: a preliminary selective review. Telemed Rep 2022;3(1):38-47 [FREE Full text] [doi: 10.1089/tmr.2021.0040] [Medline: 35720447]
31. 2020 state of telemedicine report: examining patient perspectives and physician adoption of telemedicine since the COVID-19 pandemic. Doximity. URL: https://c8y.doxcdn.com/image/upload/Press%20Blog/Research%20Reports/2020-state-telemedicine-report.pdf [accessed 2023-07-26]
32. Moron LP, Irimata KE, Parker JD. Comparison of mental health estimates by sociodemographic characteristics in the research and development survey 3 and the 2019 national health interview survey. Natl Health Stat Report 2023(188):1-11 [FREE Full text] [Medline: 37440240]
33. Fakhouri T, Chen TC, Afful J, Parker J. Non-response bias in the national health and nutrition examination surveys: a comparison of cooperative participants versus late, inaccessible, or reluctant participants. National Center for Health Statistics. URL: https://www.cdc.gov/nchs/data/bsc/bscpres_fakhouri_january_2018.pdf [accessed 2023-07-26]
34. HINTS data. National Cancer Institute. URL: https://hints.cancer.gov/data/Default.aspx [accessed 2023-12-22]

Abbreviations

ACS: American Community Survey
aPR: adjusted prevalence ratio
HINTS 6: Health Information National Trends Survey
NHIS: National Health Interview Survey

Edited by G Eysenbach, A Mavragani; submitted 23.08.23; peer-reviewed by S Kujala; comments to author 22.09.23; revised version received 31.10.23; accepted 28.02.24; published 10.05.24.

Please cite as:
Spaulding EM, Fang M, Commodore-Mensah Y, Himmelfarb CR, Martin SS, Coresh J
Prevalence and Disparities in Telehealth Use Among US Adults Following the COVID-19 Pandemic: National Cross-Sectional Survey
J Med Internet Res 2024;26:e52124
URL: https://www.jmir.org/2024/1/e52124
doi:10.2196/52124
PMID:38728070

©Erin M Spaulding, Michael Fang, Yvonne Commodore-Mensah, Cheryl R Himmelfarb, Seth S Martin, Josef Coresh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Travel Distance Between Participants in US Telemedicine Sessions With Estimates of Emissions Savings: Observational Study

Mollie R Cummins[1,2,3], MSN, PhD; Sukrut Shishupal[2], MSc; Bob Wong[1], PhD; Neng Wan[4], MS, PhD; Jiuying Han[4], MS; Jace D Johnny[1], DNP; Amy Mhatre-Owens[1], MS; Ramkiran Gouripeddi[2], MS, MBBS; Julia Ivanova[3], PhD; Triton Ong[3], PhD; Hiral Soni[3], MS, MBA, PhD; Janelle Barrera[3,5], MPH; Hattie Wilczewski[3], BS; Brandon M Welch[3,6], MS, PhD; Brian E Bunnell[3,5], PhD

[1]College of Nursing, University of Utah, Salt Lake City, UT, United States
[2]Spencer Fox Eccles School of Medicine, Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States
[3]Doxy.me Inc, Charleston, SC, United States
[4]Department of Geography, University of Utah, Salt Lake City, UT, United States
[5]Department of Psychiatry and Behavioral Neurosciences, Morsani College of Medicine, University of South Florida, Salt Lake City, UT, United States
[6]Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Mollie R Cummins, MSN, PhD
College of Nursing
University of Utah
10 South 2000 East
Salt Lake City, UT, 84112-5880
United States
Phone: 1 8015859740
Email: mollie.cummins@utah.edu

Abstract

Background: Digital health and telemedicine are potentially important strategies to decrease health care's environmental impact and contribution to climate change by reducing transportation-related air pollution and greenhouse gas emissions. However, we currently lack robust national estimates of emissions savings attributable to telemedicine.

Objective: This study aimed to (1) determine the travel distance between participants in US telemedicine sessions and (2) estimate the net reduction in carbon dioxide (CO_2) emissions attributable to telemedicine in the United States, based on national observational data describing the geographical characteristics of telemedicine session participants.

Methods: We conducted a retrospective observational study of telemedicine sessions in the United States between January 1, 2022, and February 21, 2023, on the doxy.me platform. Using Google Distance Matrix, we determined the median travel distance between participating providers and patients for a proportional sample of sessions. Further, based on the best available public data, we estimated the total annual emissions costs and savings attributable to telemedicine in the United States.

Results: The median round trip travel distance between patients and providers was 49 (IQR 21-145) miles. The median CO_2 emissions savings per telemedicine session was 20 (IQR 8-59) kg CO_2). Accounting for the energy costs of telemedicine and US transportation patterns, among other factors, we estimate that the use of telemedicine in the United States during the years 2021-2022 resulted in approximate annual CO_2 emissions savings of 1,443,800 metric tons.

Conclusions: These estimates of travel distance and telemedicine-associated CO_2 emissions costs and savings, based on national data, indicate that telemedicine may be an important strategy in reducing the health care sector's carbon footprint.

(J Med Internet Res 2024;26:e53437) doi:10.2196/53437

KEYWORDS

air pollution; environmental health; telemedicine; greenhouse gases; clinical research informatics; informatics; data science; telehealth; eHealth; travel; air quality; pollutant; pollution; polluted; environment; environmental; greenhouse gas; emissions;

XSL•FO
RenderX

retrospective; observational; United States; USA; North America; North American; cost; costs; economic; economics; saving; savings; finance; financial; finances; CO2; carbon dioxide; carbon footprint

Introduction

Diminished air quality, marked by high levels of pollutants and particulate matter, has been associated with varied cardiovascular, respiratory, and other health issues and premature mortality [1-12]. Fossil fuel–based transportation directly causes air pollution by releasing particulate matter, nitrogen oxides, and volatile organic compounds. Additionally, the emission of greenhouse gases (GHGs), including carbon dioxide (CO_2) from fossil fuel consumption (including transportation-related consumption), contributes profoundly to climate change [13]. The heat-trapping properties of CO_2 and other GHGs are considered primary drivers of global warming [14]. With climate change, we anticipate that air quality and human health will be further diminished by increased levels of ground-level ozone, particulate matter due to wildfires, and airborne allergens [13]. Beyond its effects on air quality, climate change is associated with varied additional health problems resulting in morbidity and mortality, ranging from heat-related illness to infectious disease transmission and food insecurity [15]. During the first year of the COVID-19 pandemic, we observed sharp decreases in atmospheric CO_2 worldwide, attributed to decreases in transportation and other CO_2-emitting human activities in compliance with public health orders [16-18]. However, atmospheric CO_2 has since increased and is forecasted to rise with fossil fuel consumption in the coming years, and there is a critical, global need to reduce emissions and mitigate impacts on human health [18,19].

Ironically, hospitals and health care organizations are major contributors to CO_2 emissions. These emissions relate to both direct and indirect energy expenditures of health care facilities, many of which operate 24 hours per day, 7 days per week. Emissions also result from the transportation of health care providers and other employees who provide services, and the transportation of patients and other visitors traveling to facilities. Most health care emissions result from the supply chain, which is dependent upon fossil fuels and includes the production, transportation, and operation of equipment, devices, and material supplies [20]. In a 2016 EIO-LCA (economic input-output life cycle assessment) modeling analysis, Eckelman and Sherman [21] estimated that 10% of US CO_2 emissions were attributable to the health care sector in 2013, prompting a re-examination of health care delivery practices in light of the environmental impact. In a 2020 update to the original analysis, the authors analyzed more recent data and incorporated additional state-level analyses of access and quality [22]. They determined that the US national health care GHG emissions increased by 6% from 2010 to 2018 and that electricity is the largest contributor to GHG emissions from the health care sector [22]. Health care pollution is a growing global concern. It is estimated that the health care sector is responsible for 4.4% of global GHG emissions and the United States appears particularly culpable [20]. In an international comparison of health carbon footprints, defined as global supply chain CO_2 emissions related to health care expenditures and investments, the United States was the second largest emitter of health care–related CO_2 emissions [23].

Telemedicine is a potentially important strategy for mitigating health care related GHG emissions, and numerous studies have examined the GHG emissions attributable to telemedicine. A 2022 systematic review of 31 studies conducted between 2000 and 2021 by Donald and Irukulla [24] found that telemedicine is associated with substantial CO_2 emission savings. However, previous US studies that examined telemedicine-associated CO_2 emissions savings were conducted entirely in local or regional settings [24]. For example, a 2021 study by Jiang et al [25] examined travel-related emissions savings in a convenience sample of 100 veterans receiving teleoncology care at a single site. They estimated that there was a savings of 35.5 metric tons of CO_2 with 560 sessions. In a recent larger study conducted at Stanford Health Care in California, researchers estimated a savings of 17,000 metric tons in 2021 due to telemedicine use [26].

However, specific estimates of CO_2 emissions savings from prior research vary widely, as they are based upon varied assumptions, regions, and medical specialties. Among US studies, estimates of the CO_2 emissions savings per telemedicine session range from 11.2 (vascular surgery, Michigan) to 893 kg CO_2 (otorhinolaryngology head and neck surgery, New Mexico) [24,27,28]. Additionally, most estimates are based on fairly small regional samples with under 500 participants. In the Donald and Irukulla [24] review, only 2 US studies, 1 conducted in California and 1 conducted in Utah, had more than 1000 participants [29,30]. Consequently, the overall extent of telemedicine's contribution to CO_2 emissions savings in the United States is unknown.

Health care pollution is a critical concern, and we must consider environmental impacts including GHG emissions when designing health care programs, services, and facilities to avoid unintended adverse consequences on human health. Digital health and telemedicine are potentially important strategies to decrease the health care sector's adverse environmental health impacts. However, we currently lack robust national estimates of emissions savings attributable to telemedicine that would enable the quantification of telemedicine's emissions savings in modeling efforts. This study aimed to (1) determine the travel distance between participants in US telemedicine sessions and (2) estimate the net reduction in CO_2 emissions attributable to telemedicine in the United States, based on national observational data describing the geographical characteristics of telemedicine session participants.

Methods

Study Design

We conducted a retrospective observational study of telemedicine sessions in the United States between January 1,

XSL•FO
RenderX

2022, and February 21, 2023, on the doxy.me platform. We calculated the approximate travel distance between providers and patients for a proportional sample of sessions, including estimated emissions savings and expenditures. Further, based on the best available public data, we calculated the emissions savings attributable to telemedicine in the United States.

Ethical Considerations

The data used in this study were determined to be deidentified, as per the Health Insurance Portability and Accountability Act (HIPAA) of 1996, using the expert determination method. All study procedures were reviewed by the University of Utah Institutional Review Board and determined to be nonhuman subjects research.

Doxy.me Session Data

Doxy.me is a commercial telemedicine platform commonly used by individual providers, clinics, and health care organizations [31]. Based on third-party survey research and other estimates, doxy.me is used in 8%-30% of telemedicine sessions in the United States daily [32-34]. A HIPAA-compliant platform, it does not store patient names, addresses, or medical information. For operational purposes, doxy.me stores the IP addresses of session participants and limited account information about providers. For this study, we defined a telemedicine session as any session between 1 patient and 1 provider, with a duration between 5 and 120 minutes. Doxy.me assigned approximate locations (geospatial coordinates) to the telemedicine sessions based on IP addresses using a free IP geolocation service, ipstack (iPstack API).

Sampling and Time Frame

We analyzed a proportional sample of 79,904 sessions, drawn from a randomized sample of 8,000,000 doxy.me telemedicine sessions that occurred in the United States, between 2 participants (1 patient and 1 provider), between January 1, 2022, and February 21, 2023, meeting inclusion criteria and less than 400 miles geodesic distance. The strata used for proportions were call region (Northeast, Southeast, West, Midwest, and Southwest); session length (5-30 minutes, or greater than 30 minutes); day of the week (Monday through Friday or weekend); the hour of the day (1 PM through 11 PM Coordinated Universal Time, or all other hours). Then, we calculated travel distance and time using the Google Distance Matrix Application Programming Interface for car travel [35].

Inclusion Criteria

Telemedicine sessions occurring between dyads (1 provider and 1 patient), using any device, between January 1, 2022, and February 21, 2023.

Exclusion Criteria

We excluded calls of less than 5 minutes, as calls of this duration generally do not represent actual clinical encounters. Rather, they are attributable to changes in equipment or devices or a need to adjust settings. We also excluded group sessions and sessions longer than 2 hours, which likely represent the use of doxy.me for remote monitoring. We excluded sessions with >400 miles geodesic distance to mitigate bias related to extreme distances and based on the conservative assumption that most patients receiving care from a highly distant health care provider, should they need to see that provider in person, would seek alternative local care or not seek care at all.

Analysis

For the proportional sample of 79,904 sessions, we first geocoded the location of health care seekers and providers based on the coordinates derived from IP addresses, then calculated car travel time and distance using the Google Distance Matrix Application Programming Interface [35]. We calculated descriptive statistics and box plots to describe session characteristics, including duration, specialty, patient travel time, and travel distance (round trip). Then we adjusted the travel distance for a more realistic approximation of travel savings due to telemedicine. We multiplied by 0.848, based on the assumption that approximately 84.8% of travel was car travel, consistent with American Community Survey data on adult commuting patterns in the United States [36].

CO_2 Emission Savings

We calculated the tailpipe CO_2 emission savings associated with the telemedicine sessions by multiplying the total adjusted travel distance in miles by 404 grams, which is the 2022 US Environmental Protection Agency estimate of CO_2 emissions per mile for an average passenger vehicle, assuming an average fuel economy of 22.0 miles per gallon [37].

CO_2 Expenditure

To approximate the CO_2 expenditure associated with each telemedicine session, we first calculated the energy expenditure associated with videoconferencing using the estimates and methods described by Blenkinsop and colleagues [38] and Mytton [39], corresponding to the use of 720p high-definition video by 2 participants. Following that approach, we calculated the total amount of data transferred during each session using a rate of 0.036 GB per minute [38]. We assumed an electricity use of 0.015 kWh/GB, based on fixed-line energy transmission estimates [40]. Using the US Environmental Protection Agency conversion rate of 4.33×10^{-4} metric tons CO_2/kWh, we estimated the CO_2 emissions expenditure in metric tons. Then, we subtracted the CO_2 expenditure from CO_2 savings to calculate the net CO_2 savings per session.

No publicly available data describes the total number of telemedicine sessions that took place in the United States during the dates of this study. To obtain an approximate estimate of annual US telemedicine sessions, we used publicly reported data on the number of telemedicine sessions reimbursed via the US Medicare and Medicaid programs. Centers for Medicare & Medicaid Services (CMS) spending accounted for 38% of National Health Expenditures in 2021. CMS reimbursed for 27,691,878 telemedicine sessions during 1 year, from March 1, 2020, to February 28, 2021 [41]. While expenditures do not translate directly to the volume of health care delivery or the number of telemedicine sessions that take place, we used it as a best-available approximate of the overall volume of health care delivered using telemedicine. Based upon these figures, we estimate that CMS-reimbursed telemedicine sessions represented 38% of a total of 72,873,363 telemedicine sessions

XSL•FO
RenderX

that occurred in the United States in 2021, and we base our estimate of CO_2 emissions savings on this number of sessions and the postpandemic policy environment of 2021 and 2022.

Representativeness of Doxy.me Session Data

We used multiple data sources describing national telemedicine use patterns to assess the representativeness of doxy.me data for national patterns of use [41-44]. We calculated descriptive statistics for 2 points of comparability, the regional distribution of sessions and the specialty of providers, and compared the observed distributions with that reported nationally for Medicare claims. See Multimedia Appendix 1 for a detailed comparison.

Geographic Distribution of CO_2 Emissions Savings

Using ArcGIS, we created 2 maps that visualize CO_2 emissions savings at the zip code level. The first map shows the overall pattern of CO_2 savings. The second map shows the CO_2 savings per session.

Results

Overview

A random sample of 8,000,000 telemedicine sessions, hosted by doxy.me, between January 1, 2022, and February 21, 2023, were examined. Of the sample, 6,231,614 sessions met our inclusion and exclusion criteria. After proportional sampling, the analytic data set consisted of 79,904 sessions.

Session Characteristics

The duration of sessions was bimodal in distribution. Of the 79,904 telemedicine sessions, 35,204 (44.06%) were 5-30 minutes in length, and 44,700 (55.94%) were >30-120 minutes in length. The Northeast region had the most sessions 27,311 (34.18%), followed by the Southeast 16,318 (20.42%), West 15,179 (18.99%), Midwest 13,788 (17.26%), and finally Southwest 7308 (9.15%). Tuesday, Wednesday, and Thursday were most frequent: 18,134 (22.69%), 17,751 (22.22%), and 17,052 (21.34%), respectively. From 2 PM-5 PM Eastern (11 AM-2 PM Pacific) were the most common hours, each at approximately 8150 sessions (10.19%). The geographic distribution of the sessions by region is described in Table 1 and depicted in Figures 1 and 2. Comparisons of the doxy.me data to reference data sources, by state and by specialty are provided in Multimedia Appendix 1.

Table 1. Session characteristics include duration and geographic distribution.

Characteristics	Full data set (N=6,231,614), n (%)	Proportional sample (n=79,904), n (%)
Duration (min)		
5-30	2,747,610 (44.06)	35,204 (44.06)
30-120	3,484,004 (56.94)	44,700 (55.94)
Geographical location (by region)		
Northeast	2,125,693 (34.18)	27,311 (3.18)
Southeast	1,272,704 (20.42)	16,318 (20.42)
West	1,183,814 (18,99)	15,179 (18.99)
Midwest	1,076,509 (17.26)	13,788 (17.26)
Southwest	572,884 (9.15)	7308 (9.15)
Missing	10 (0)	0 (0)

XSL•FO
RenderX

Figure 1. Geographical distribution of telemedicine sessions and total number of sessions by county.

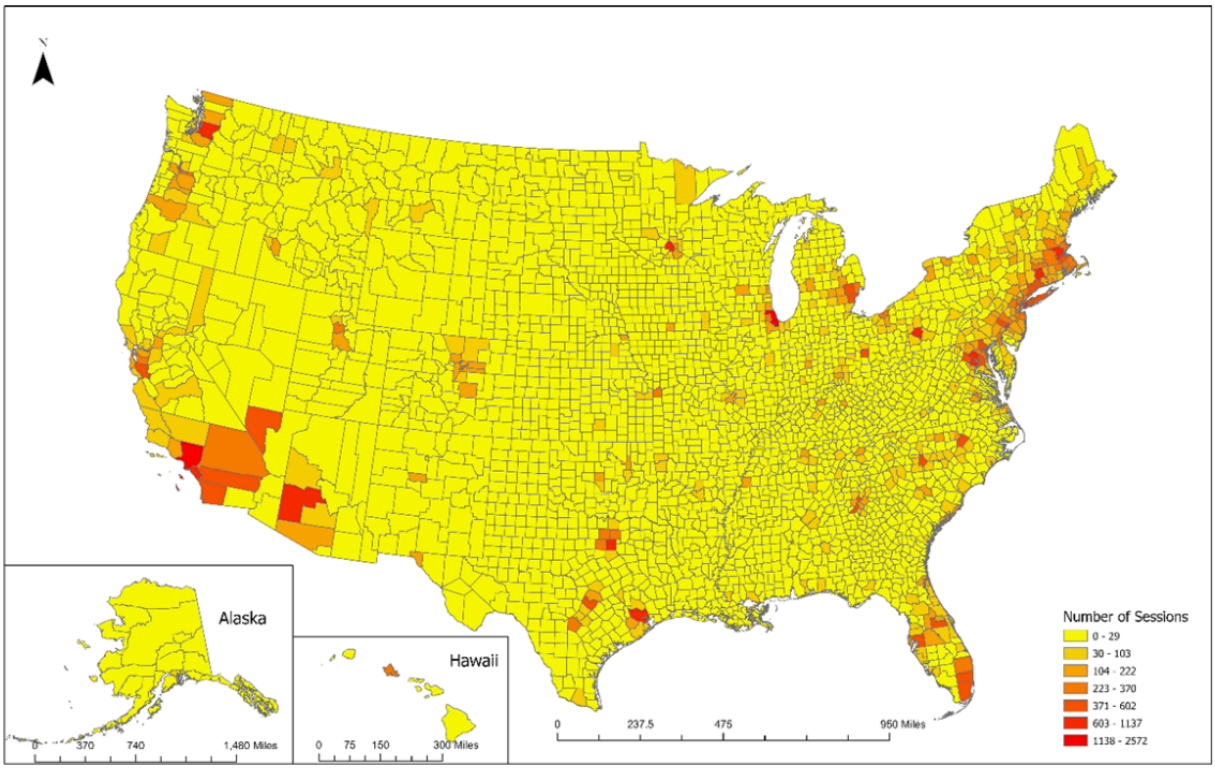

Figure 2. Geographical distribution of telemedicine sessions and number of sessions per 100,000 population by county.

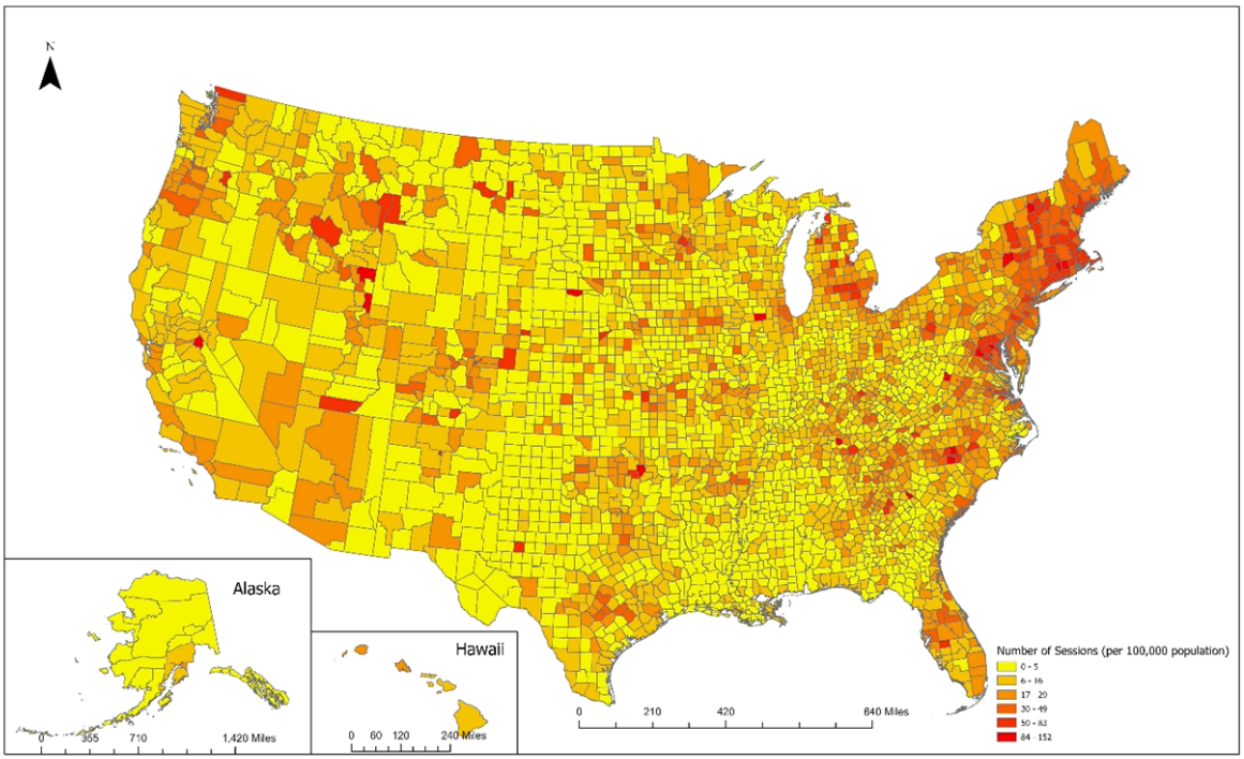

XSL•FO
RenderX

Travel Distance

Travel distance was not normally distributed. Overall, the median round trip distance was 49 (IQR 21-15) miles; the mean round trip travel distance was 134 (SD 196) miles. Descriptive statistics for round trip travel distance (in miles) grouped by day of week, region, and telehealth session length are presented in Table 2. Weekends show the highest round trip distance (mean 150, SD 210, and median 52, IQR 23-182). The Southeast region showed the highest distance of travel (mean 167, SD 214, and median 63, IQR 25-234). Telemedicine sessions of 5-30 minutes duration showed the highest travel distance (mean 144, SD 204, and median 54, IQR 23-166). Figure 3 shows box plots of round trip travel distance, by region. The box plots illustrate regional variation, the skewed distribution of distance across all regions, and the presence of high-distance outliers.

Table 2. Round trip travel distance (miles) for proportional sample (N=79,904).

Characteristics	Sample, n	Mean (SD)	Median (IQR)	Minimum to maximum
Day of week				
Monday	13,777	131 (194)	49 (20-141)	0-1228
Tuesday	18,134	132 (195)	49 (20-144)	0-1191
Wednesday	17,751	135 (198)	49 (20-147)	0-1130
Thursday	17,052	133 (195)	49 (21-143)	0-1098
Friday	10,711	136 (198)	50 (22-149)	0-1065
Saturday and Sunday	2479	150 (210)	52 (23-182)	0-1065
Region				
Midwest	13,788	143 (209)	48 (10-159)	0-1127
Northeast	27,311	106 (157)	46 (20-109)	0-1065
Southeast	16,318	167 (214)	63 (25-234)	0-1104
Southwest	7308	145 (205)	49 (22-177)	0-1040
West	15,179	136 (217)	44 (18-128)	0-1228
Session length				
30-120 min	44,700	126 (190)	46 (19-130)	0-1228
5-30 min	35,204	144 (204)	54 (22-166)	0-1130

Figure 3. Boxplots of round trip travel distance (miles) by US region for proportional sample (N=79,904).

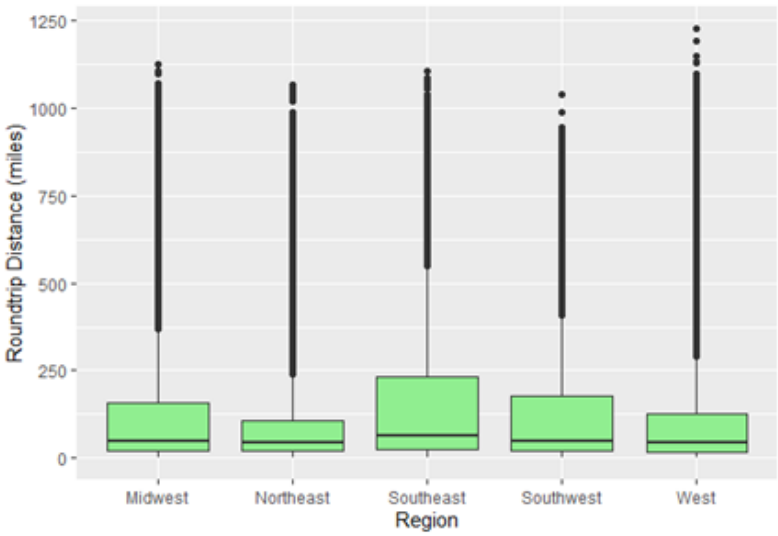

Emissions Savings and Expenditures

The emissions savings and expenditures are detailed in Table 3. The mean CO_2 savings per telemedicine session was 54,077 (SD 79,362) g and the median was 19,812 (IQR 8322-58,772) g. The mean CO_2 savings per minute of telemedicine care was 2896 (SD 6258) g, and the median was 723 (IQR 239-2597) g. For the 6,231,614 doxy.me telemedicine sessions, we estimate a savings of 123,461 metric tons of CO_2 (calculated with median). Further, we estimate that nationally, the use of telemedicine in the United States during the years 2021-2022 resulted in approximate annual CO_2 emissions savings of

XSL•FO
RenderX

1,443,800 metric tons (calculated with median). Multimedia Appendix 2 contains a map showing the geographical distribution of the CO_2 emissions savings per telemedicine session.

Table 3. Travel and emissions calculations per session for proportional sample (N=79,904).

Characteristics	Mean (SD)	Median (IQR)	Minimum to maximum
Travel time, round trip, automobile[a] (h)	2.38 (3.01)	1.15 (0.64-2.66)	0.00 to 22.94
Travel distance, round trip, automobile[a] (miles)	133.87 (196.44)	49.05 (20.62-145.49)	0.00 to 1227.55
Transportation-related CO_2 emissions savings (g)	54,085.06 (79,362.04)	19,818.13 (8331.32-58,776.59)	0.00 to 495,928.64
Duration (min)	34.53 (20.03)	36.73 (14.48-51.73)	5.00 to 119.83
CO_2 emissions equivalent of telemedicine energy use (g)	8.07 (4.68)	8.59 (3.39-12.09)	1.17 to 28.02
Net emissions savings per telemedicine session (g)	54,076.99 (79,362.26)	19,812.45 (8322.11-58,772.37)	–27.89 to 495,921.13
Net emissions savings per minute of session length (g)	2896.17 (6258.47)	722.99 (238.65-2596.56)	–0.23 to 79,230.29

[a]Calculated using Google Distance Matrix Application Programming Interface.

Discussion

Principal Findings

We used national, observational data to determine the travel distance between patients and providers participating in telemedicine sessions. Further, we used our findings to generate national estimates of CO_2 emissions costs and savings. Under the relatively stable relevant policy conditions of the years 2021-2022, we estimate the annual emissions savings associated with US telemedicine as 1,462,932 metric tons or nearly 1.5 million metric tons. This is equivalent to the CO_2 emissions from approximately 3.4 million barrels of oil or 165 million gallons of gasoline [45]. This estimate is unique in that it was based upon a large national set of observational data, calculation of travel distance using Google Distance Matrix, and a set of conservative assumptions related to telemedicine connections and behavior. Further, we estimated emissions savings per telemedicine session and per minute of telemedicine care delivery, potentially useful estimates for modeling the environmental impact of health care programs and services. Previously published estimates of emissions savings were based on regional or local telemedicine use and are not directly comparable. Here, we generated estimates based on a large national sample of US telemedicine sessions conducted in varied geographic areas by health care providers of varied specialties.

We used round trip travel distance determined using Google Distance Matrix rather than geodesic distance as the basis for emissions savings estimates. With geodesic distance ≤400, travel distance by car still exceeded 1000 miles in some rare cases, particularly in the West, but was typically much lower. We found that travel distance varied according to region and day of the week. On weekends, there were fewer telemedicine sessions, but the travel distance between providers and patients was greater. This may reflect weekend telemedicine staffing patterns that use more distant health care providers, or the use of telemedicine services based in more distant geographical areas when local, in-person health care services are closed over a weekend. Additionally, we observed some regional variation in the distribution of travel distance. We observed the highest median travel distance between patients and providers in the southeastern United States, and the lowest median distance in the West. Additionally, shorter sessions were characterized by greater distances, which could reflect decisions to manage brief encounters such as follow-up visits by telemedicine when patients are particularly distant from the health care provider.

We found a median round trip travel distance by automobile of 49 (IRQ 21-145) miles, based on the approximate locations of patients and providers participating in telemedicine sessions. This estimate is far lower than the travel distance savings found in larger prepandemic research studies [24]. For example, Thota et al [29] found a mean round trip distance savings of 332 miles per encounter for telemedicine cancer care in rural Utah. A prepandemic study of telemedicine consultations in the UC Davis system found a mean round trip distance savings of 278 miles [30]. A larger and more recent pandemic-era study of telemedicine in 5 UC health systems showed a substantially lower round trip travel distance savings of 17.6 miles, likely reflecting widespread pandemic-era telemedicine adoption and increased use of telemedicine in urban areas of the United States [46]. Our results align most closely with those of Sharma and colleagues [46], consistent with the similar timeframe and pandemic-era telemedicine use patterns. Our results likely differ because we used an entirely different approach to calculating travel distance; we used Google Distance Matrix to calculate automobile travel distance based on observational data describing the approximate actual location of patients and providers during telemedicine sessions, rather than geodesic (point-to-point) distance and historical street address data as in most previous studies. Further, our analysis was based on national data that reflects greater geographical diversity and diversity in patterns of telemedicine use.

To slow global warming and climate change, we must reduce CO_2 emissions in the health care sector. Though we estimate a substantial net savings of CO_2 emissions through telemedicine, these savings are equivalent to merely 0.02% of the total US CO_2 emissions in 2021 (6340.2 million metric tons) [47]. Telemedicine is clearly important in reducing GHG but must be combined with other innovative programs, services, and changes in behavior within and beyond the health care sector. An increasing number of health care organizations in the United

XSL•FO
RenderX

States have committed to reducing their carbon footprint. For example, Kaiser Permanente achieved carbon neutrality in 2020, and the University of California San Francisco is working toward carbon neutrality by 2025 [48]. Given our findings, the environmental impact of telemedicine should be duly considered in designing programs and services, as part of the effort to reduce health care's carbon footprint.

Limitations and Considerations

The estimates in this study depend upon a series of assumptions related to energy-consuming human behaviors. However, many of the assumptions are conservative and, if erroneous, would underestimate rather than overestimate CO_2 emissions savings. For example, we used adult commuting patterns to determine the percentage of travel occurring by car. However, patients who use alternative transportation to work (eg, public transportation, biking, and walking) may still travel by car to health care appointments. However, the use of cars for travel to health care appointments may differ for children, the retired, old adults, or disabled persons.

Similarly, we assumed patients would not travel to in-person appointments in place of telemedicine when the geodesic distance exceeds 400 miles and excluded sessions with a geodesic distance >400 from the analysis. Given this distance, if telemedicine was not an option, we reason that patients would have sought alternative local care or missed care. In reality, a certain proportion of these patients may travel unusually long distances, especially those seeking specialty care (eg, care at a cancer center) and patients in frontier areas. However, we decided to exclude high-distance telemedicine sessions to avoid inflated estimates.

Another critical assumption is that doxy.me data are nationally representative of patients and providers participating in telemedicine. We compared the geographic distribution of doxy.me session data to that reported for Medicare fee-for-service payments in Grace [44] by state. We found them to be highly similar (Multimedia Appendix 1). States with higher dense populations, such as California, New York, and Texas, are consistent in higher telemedicine session use in both sources. In contrast, more rural states like Wyoming have consistently low use (Multimedia Appendix 1). The overall market share of the doxy.me platform has been estimated by third parties as between 7% and 30% of US telemedicine [32-34,49-51].

For doxy.me specialty data, we were limited to the data available through deterministic linkage of doxy.me provider accounts to NPPES (National Plan and Provider Enumeration System), which was available for only 43.55% of the sessions. To gain insight into whether the doxy.me data are nationally representative in terms of specialty, we compared the distribution of specialty in the doxy.me data to that of the Medicare fee-for-payment data described by Grace [44]. There were marked differences in the distribution of specialties, which may or may not be attributable to quality issues in the linked specialty data (Multimedia Appendix 1). However, mental health services, psychiatry, and social work were consistent and frequent specialties in both doxy.me and the Medicare data. The types of telemedicine supported by the doxy.me telemedicine platform may differ somewhat from those of other platform providers. Doxy.me has a firm market share among mental health providers, offering integrated digital health functionality for mental health care. Thus, the sample that we analyzed may overrepresent mental health services and underrepresent other health care specialties (Adhere.ly; Adhere.ly, LLC). Given the large amount of missing data for specialty and substantial differences compared to Medicare claims data, we did not estimate telemedicine emissions savings by provider specialty.

The total number of telemedicine sessions that occur in the United States during a given year is unknown. We based our estimate on actual CMS reimbursement for telemedicine during 2021 and the proportion of US health care expenditures funded by CMS. Both 2021 and 2022 were years during the federal health emergency declaration related to the COVID-19 pandemic. Federal and state policies related to reimbursement and licensing can substantially affect telemedicine use patterns. However, policy was relatively stable during the years 2021-2022.

In future work, we plan to update and refine these estimates based on observational data describing real-world telemedicine connection characteristics and the type of devices used to connect, which affect the energy costs associated with telemedicine. We also plan to incorporate energy costs associated with standard, in-person care delivery including built infrastructure and personnel transportation. More robust specialty data would enable additional insights into potential variations in CO_2 emissions costs and savings across different types of health care. Additionally, future efforts could refine model assumptions based on observed health care use behavior patterns and regional differences in health care service delivery, as the body of evidence describing telemedicine use grows.

Conclusions

US health care is under increasing scrutiny for contributing to air pollution, climate change, and related adverse health outcomes. To duly consider environmental impacts in decision-making related to health care facilities, programs, and services, we must be able to quantify and compare the environmental effects of varied approaches to health care delivery. Here, we determined the travel distance between participants in US telemedicine sessions and, on that basis, estimated the associated annual CO_2 emissions savings for US telemedicine in 2021-2022 as nearly 1.5 million tons. Further, we estimated emissions savings per telemedicine session, and per minute of telemedicine care delivery based on a large national sample of telemedicine sessions. In doing so, we adopted conservative assumptions related to patient behavior and transportation and considered both energy savings and expenditures of telemedicine. The results indicate that telemedicine is an important strategy for reducing the health care sector's carbon footprint. Additionally, the findings enable quantification of telemedicine-associated CO_2 emissions savings.

XSL•FO
RenderX

Acknowledgments

The support and resources from the Center for High Performance Computing and the Center of Excellence in Exposure Health Informatics at the University of Utah are gratefully acknowledged. We also gratefully acknowledge the contributions of the Doxy.me Inc business intelligence team.

Conflicts of Interest

BMW is the founder, chief executive officer, and shareholder of Doxy.me Inc, a commercial telemedicine company. MRC, JI, TO, HS, JB, HW, and BEB are employees of the same company. The authors declare no other conflicts of interest.

Multimedia Appendix 1
Comparison of doxy.me session characteristics to other national data (state, specialty).
[DOCX File , 22 KB - jmir_v26i1e53437_app1.docx]

Multimedia Appendix 2
Geographical distribution of CON2 emissions savings per session (N=79,904).
[DOCX File , 5184 KB - jmir_v26i1e53437_app2.docx]

References

1. Timonen KL, Pekkanen J. Air pollution and respiratory health among children with asthmatic or cough symptoms. Am J Respir Crit Care Med 1997;156(2 Pt 1):546-552 [FREE Full text] [doi: 10.1164/ajrccm.156.2.9608044] [Medline: 9279238]
2. Thurston GD, Balmes JR, Garcia E, Gilliland FD, Rice MB, Schikowski T, et al. Outdoor air pollution and new-onset airway disease. An official American Thoracic Society workshop report. Ann Am Thorac Soc 2020;17(4):387-398 [FREE Full text] [doi: 10.1513/AnnalsATS.202001-046ST] [Medline: 32233861]
3. Szyszkowicz M, Kousha T, Castner J, Dales R. Air pollution and emergency department visits for respiratory diseases: a multi-city case crossover study. Environ Res 2018;163:263-269 [FREE Full text] [doi: 10.1016/j.envres.2018.01.043] [Medline: 29459308]
4. Zhou M, He G, Liu Y, Yin P, Li Y, Kan H, et al. The associations between ambient air pollution and adult respiratory mortality in 32 major Chinese cities, 2006-2010. Environ Res 2015;137:278-286. [doi: 10.1016/j.envres.2014.12.016] [Medline: 25601729]
5. Meo SA, Suraya F. Effect of environmental air pollution on cardiovascular diseases. Eur Rev Med Pharmacol Sci 2015;19(24):4890-4897 [FREE Full text] [Medline: 26744881]
6. Pozzer A, Dominici F, Haines A, Witt C, Münzel T, Lelieveld J. Regional and global contributions of air pollution to risk of death from COVID-19. Cardiovasc Res 2020;116(14):2247-2253 [FREE Full text] [doi: 10.1093/cvr/cvaa288] [Medline: 33236040]
7. Pascal M, de Crouy Chanel P, Wagner V, Corso M, Tillier C, Bentayeb M, et al. The mortality impacts of fine particles in France. Sci Total Environ 2016;571:416-425. [doi: 10.1016/j.scitotenv.2016.06.213] [Medline: 27453142]
8. Hanigan IC, Rolfe MI, Knibbs LD, Salimi F, Cowie CT, Heyworth J, et al. All-cause mortality and long-term exposure to low level air pollution in the '45 and up study' cohort, Sydney, Australia, 2006-2015. Environ Int 2019;126:762-770 [FREE Full text] [doi: 10.1016/j.envint.2019.02.044] [Medline: 30878871]
9. Wei Y, Wang Y, Wu X, Di Q, Shi L, Koutrakis P, et al. Causal effects of air pollution on mortality rate in Massachusetts. Am J Epidemiol 2020;189(11):1316-1323 [FREE Full text] [doi: 10.1093/aje/kwaa098] [Medline: 32558888]
10. Strak M, Weinmayr G, Rodopoulou S, Chen J, de Hoogh K, Andersen ZJ, et al. Long term exposure to low level air pollution and mortality in eight European cohorts within the ELAPSE project: pooled analysis. BMJ 2021;374:n1904 [FREE Full text] [doi: 10.1136/bmj.n1904] [Medline: 34470785]
11. Kim H, Kim J, Kim S, Kang SH, Kim HJ, Kim H, et al. Cardiovascular effects of long-term exposure to air pollution: a population-based study with 900 845 person-years of follow-up. J Am Heart Assoc 2017;6(11):e007170 [FREE Full text] [doi: 10.1161/JAHA.117.007170] [Medline: 29118034]
12. Turner MC, Krewski D, Diver WR, Pope CA, Burnett RT, Jerrett M, et al. Ambient air pollution and cancer mortality in the cancer prevention study II. Environ Health Perspect 2017;125(8):087013 [FREE Full text] [doi: 10.1289/EHP1249] [Medline: 28886601]
13. Nolte CA, Dolwick PD, Fann N, Horowitz LW, Naik V, Pinder RW, et al. Ch. 13: Air quality. Impacts, risks, and adaptation in the United States: fourth national climate assessment, volume II.: U.S. Global Change Research Program; 2018. URL: https://nca2018.globalchange.gov/chapter/13/ [accessed 2023-12-12]
14. Causes of climate change. European Commission. URL: https://climate.ec.europa.eu/climate-change/causes-climate-change_en [accessed 2023-06-12]

XSL•FO
RenderX

15. Romanello M, Di Napoli C, Drummond P, Green C, Kennard H, Lampard P, et al. The 2022 report of the Lancet countdown on health and climate change: health at the mercy of fossil fuels. Lancet 2022;400(10363):1619-1654 [FREE Full text] [doi: 10.1016/S0140-6736(22)01540-9] [Medline: 36306815]
16. Nguyen XP, Hoang AT, Ölçer AI, Huynh TT. Record decline in global CO2 emissions prompted by COVID-19 pandemic and its implications on future climate change policies. Energ Source Part A 2021:1-4 [FREE Full text] [doi: 10.1080/15567036.2021.1879969]
17. Kumar A, Singh P, Raizada P, Hussain CM. Impact of COVID-19 on greenhouse gases emissions: a critical review. Sci Total Environ 2022;806(Pt 1):150349 [FREE Full text] [doi: 10.1016/j.scitotenv.2021.150349] [Medline: 34555610]
18. Smith LV, Tarui N, Yamagata T. Assessing the impact of COVID-19 on global fossil fuel consumption and CO2 emissions. Energy Econ 2021;97:105170 [FREE Full text] [doi: 10.1016/j.eneco.2021.105170] [Medline: 33612887]
19. United Nations Environment Programme. Emissions Gap Report 2022: The Closing Window—Climate Crisis Calls for Rapid Transformation Of Societies. Nairobi: United Nations; 2022.
20. How the Health Sector Contributes to the Global Climate Crisis and Opportunities for Action. Health Care's Climate Footprint. URL: https://noharm-global.org/sites/default/files/documents-files/5961/HealthCaresClimateFootprint_092319.pdf
21. Eckelman MJ, Sherman J. Environmental impacts of the U.S. health care system and effects on public health. PLoS One 2016;11(6):e0157014 [FREE Full text] [doi: 10.1371/journal.pone.0157014] [Medline: 27280706]
22. Eckelman MJ, Huang K, Lagasse R, Senay E, Dubrow R, Sherman JD. Health care pollution and public health damage in the United States: an update. Health Aff (Millwood) 2020;39(12):2071-2079 [FREE Full text] [doi: 10.1377/hlthaff.2020.01247] [Medline: 33284703]
23. Pichler PP, Jaccard IS, Weisz U, Weisz H. International comparison of health care carbon footprints. Environ Res Lett 2019;14(6):064004 [FREE Full text] [doi: 10.1088/1748-9326/ab19e1]
24. Donald N, Irukulla S. Greenhouse gas emission savings in relation to telemedicine and associated patient benefits: a systematic review. Telemed e-Health 2022;28(11):1555-1563. [doi: 10.1089/tmj.2022.0047] [Medline: 35446668]
25. Jiang CY, Strohbehn GW, Dedinsky RM, Raupp SM, Pannecouk BM, Yentz SE, et al. Teleoncology for veterans: high patient satisfaction coupled with positive financial and environmental impacts. JCO Oncol Pract 2021;17(9):e1362-e1374 [FREE Full text] [doi: 10.1200/OP.21.00317] [Medline: 34406817]
26. Thiel CL, Mehta N, Sejo CS, Qureshi L, Moyer M, Valentino V, et al. Telemedicine and the environment: life cycle environmental emissions from in-person and virtual clinic visits. NPJ Digit Med 2023;6(1):87 [FREE Full text] [doi: 10.1038/s41746-023-00818-7] [Medline: 37160996]
27. Beswick DM, Vashi A, Song Y, Pham R, Holsinger FC, Rayl JD, et al. Consultation via telemedicine and access to operative care for patients with head and neck cancer in a Veterans Health Administration population. Head Neck 2016;38(6):925-929. [doi: 10.1002/hed.24386] [Medline: 26899939]
28. Paquette S, Lin JC. Outpatient telemedicine program in vascular surgery reduces patient travel time, cost, and environmental pollutant emissions. Ann Vasc Surg 2019;59:167-172. [doi: 10.1016/j.avsg.2019.01.021] [Medline: 31077768]
29. Thota R, Gill DM, Brant JL, Yeatman TJ, Haslem DS. Telehealth is a sustainable population health strategy to lower costs and increase quality of health care in rural Utah. JCO Oncol Pract 2020;16(7):e557-e562 [FREE Full text] [doi: 10.1200/JOP.19.00764] [Medline: 32463765]
30. Dullet NW, Geraghty EM, Kaufman T, Kissee JL, King J, Dharmar M, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017;20(4):542-546 [FREE Full text] [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
31. National Plan and Provider Enumeration System. URL: https://nppes.cms.hhs.gov/#/ [accessed 2023-05-09]
32. Connor J, Zheng Y, Houle K, Cox L. Adopting telehealth during the COVID-19 era: the urologist's perspective. Urology 2021;156:289-295 [FREE Full text] [doi: 10.1016/j.urology.2021.03.051] [Medline: 34052257]
33. Fox-Fuller JT, Rizer S, Andersen SL, Sunderaraman P. Survey findings about the experiences, challenges, and practical advice/solutions regarding teleneuropsychological assessment in adults. Arch Clin Neuropsychol 2022;37(2):274-291 [FREE Full text] [doi: 10.1093/arclin/acab076] [Medline: 34564721]
34. Vendor performance in response to the COVID-19 crisis. KLAS Research. 2020. URL: https://drive.google.com/file/d/1Ll_eea2V68uFaodQOGJcGVOK3nCQDAYl/view [accessed 2023-06-06]
35. Distance matrix API overview. Google Developers. URL: https://developers.google.com/maps/documentation/distance-matrix/overview [accessed 2023-05-09]
36. Michael B, Charlynn B, Brian M. Commuting by public transportation in the United States: 2019. U.S. Census Bureau. 2021. URL: https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs-48.pdf [accessed 2024-04-02]
37. Office of Transportation and Air Quality. Greenhouse gas emissions from a typical passenger vehicle. U.S. Environmental Protection Agency. 2018. URL: https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100U8YT.pdf [accessed 2024-04-02]
38. Blenkinsop S, Foley A, Schneider N, Willis J, Fowler HJ, Sisodiya SM. Carbon emission savings and short-term health care impacts from telemedicine: an evaluation in epilepsy. Epilepsia 2021;62(11):2732-2740 [FREE Full text] [doi: 10.1111/epi.17046] [Medline: 34462915]

XSL•FO
RenderX

39. Mytton D. Zoom, video conferencing, energy, and emissions. David Mytton Blog. URL: https://davidmytton.blog/zoom-video-conferencing-energy-and-emissions/ [accessed 2023-12-20]
40. Aslan J, Mayers K, Koomey JG, France C. Electricity intensity of internet data transmission: untangling the estimates. J Ind Ecol 2018;22(4):785-798 [FREE Full text] [doi: 10.1111/jiec.12630]
41. Medicare telemedicine snapshot. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/medicare-telemedicine-snapshot [accessed 2023-12-06]
42. Lucas JW, Villarroel MA. Telemedicine use among adults: United States, 2021. NCHS Data Brief 2022(445):1-8 [FREE Full text] [Medline: 36255940]
43. Telemedicine use. National Center for Health Statistics. 2022. URL: https://www.cdc.gov/nchs/covid19/pulse/telemedicine-use.htm [accessed 2023-06-06]
44. Grace S. Telehealth expansion in medicare: policy changes, recent trends in adoption, and future impact. CareJourney. URL: https://carejourney.com/telehealth-expansion-in-medicare-policy-changes-recent-trends-in-adoption-and-future-impact/ [accessed 2023-06-06]
45. Greenhouse Gas Equivalencies calculator. United States Environmental Protection Agency. 2015. URL: https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator [accessed 2023-06-22]
46. Sharma S, Yellowlees PM, Gotthardt CJ, Luce MS, Avdalovic MV, Marcin JP. Environmental impact of ambulatory telehealth use by a statewide university health system during COVID-19. Telemed e-Health 2023;29(8):1134-1142. [doi: 10.1089/tmj.2022.0396] [Medline: 36576982]
47. Inventory of U.S. greenhouse gas emissions and sinks. United States Environmental Protection Agency. 2017. URL: https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks [accessed 2023-06-22]
48. Decarbonization. UCSF Office of Sustainability Campus Life Services. URL: https://campuslifeserviceshome.ucsf.edu/sustainability/carbon-neutral-2025 [accessed 2023-06-22]
49. Top telehealth companies by market share. Definitive Healthcare. URL: https://www.definitivehc.com/resources/healthcare-insights/top-telehealth-companies-market-share [accessed 2023-06-09]
50. Adams K. The 13 telehealth platforms physicians use the most. Becker's Health IT. URL: https://www.beckershospitalreview.com/telehealth/the-13-telehealth-platforms-physicians-use-the-most.html [accessed 2023-06-09]
51. Ritchie JB, Allen CG, Morrison H, Nichols M, Lauzon SD, Schiffman JD, et al. Utilization of health information technology among cancer genetic counselors. Mol Genet Genomic Med 2020 Aug;8(8):e1315 [FREE Full text] [doi: 10.1002/mgg3.1315] [Medline: 32468681]

Abbreviations

CMS: Centers for Medicare & Medicaid Services
CO_2: carbon dioxide
EIO-LCA: economic input-output lifecycle assessment
GHG: greenhouse gas
HIPAA: Health Insurance Portability and Accountability Act
NPPES: National Plan and Provider Enumeration System

Edited by G Eysenbach; submitted 06.10.23; peer-reviewed by P Yellowlees; comments to author 17.11.23; revised version received 22.11.23; accepted 26.01.24; published 15.05.24.

Please cite as:
Cummins MR, Shishupal S, Wong B, Wan N, Han J, Johnny JD, Mhatre-Owens A, Gouripeddi R, Ivanova J, Ong T, Soni H, Barrera J, Wilczewski H, Welch BM, Bunnell BE
Travel Distance Between Participants in US Telemedicine Sessions With Estimates of Emissions Savings: Observational Study
J Med Internet Res 2024;26:e53437
URL: https://www.jmir.org/2024/1/e53437
doi:10.2196/53437
PMID:38536065

©Mollie R Cummins, Sukrut Shishupal, Bob Wong, Neng Wan, Jiuying Han, Jace D Johnny, Amy Mhatre-Owens, Ramkiran Gouripeddi, Julia Ivanova, Triton Ong, Hiral Soni, Janelle Barrera, Hattie Wilczewski, Brandon M Welch, Brian E Bunnell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public and Research Interest in Telemedicine From 2017 to 2022: Infodemiology Study of Google Trends Data and Bibliometric Analysis of Scientific Literature

Andrea Maugeri[1], PhD; Martina Barchitta[1], PhD; Guido Basile[2], MD; Antonella Agodi[1], PhD

[1]Department of Medical and Surgical Sciences and Advanced Technologies "GF Ingrassia", University of Catania, Catania, Italy

[2]Department of General Surgery and Medical-Surgical Specialties, University of Catania, Catania, Italy

Corresponding Author:
Antonella Agodi, PhD
Department of Medical and Surgical Sciences and Advanced Technologies "GF Ingrassia"
University of Catania
Via Santa Sofia 87
Catania, 95123
Italy
Phone: 39 0953792183
Email: agodia@unict.it

Abstract

Background: Telemedicine offers a multitude of potential advantages, such as enhanced health care accessibility, cost reduction, and improved patient outcomes. The significance of telemedicine has been underscored by the COVID-19 pandemic, as it plays a crucial role in maintaining uninterrupted care while minimizing the risk of viral exposure. However, the adoption and implementation of telemedicine have been relatively sluggish in certain areas. Assessing the level of interest in telemedicine can provide valuable insights into areas that require enhancement.

Objective: The aim of this study is to provide a comprehensive analysis of the level of public and research interest in telemedicine from 2017 to 2022 and also consider any potential impact of the COVID-19 pandemic.

Methods: Google Trends data were retrieved using the search topics "telemedicine" or "e-health" to assess public interest, geographic distribution, and trends through a joinpoint regression analysis. Bibliographic data from Scopus were used to chart publications referencing the terms "telemedicine" or "eHealth" (in the title, abstract, and keywords) in terms of scientific production, key countries, and prominent keywords, as well as collaboration and co-occurrence networks.

Results: Worldwide, telemedicine generated higher mean public interest (relative search volume=26.3%) compared to eHealth (relative search volume=17.6%). Interest in telemedicine remained stable until January 2020, experienced a sudden surge (monthly percent change=95.7%) peaking in April 2020, followed by a decline (monthly percent change=–22.7%) until August 2020, and then returned to stability. A similar trend was noted in the public interest regarding eHealth. Chile, Australia, Canada, and the United States had the greatest public interest in telemedicine. In these countries, moderate to strong correlations were evident between Google Trends and COVID-19 data (ie, new cases, new deaths, and hospitalized patients). Examining 19,539 original medical articles in the Scopus database unveiled a substantial rise in telemedicine-related publications, showing a total increase of 201.5% from 2017 to 2022 and an average annual growth rate of 24.7%. The most significant surge occurred between 2019 and 2020. Notably, the majority of the publications originated from a single country, with 20.8% involving international coauthorships. As the most productive country, the United States led a cluster that included Canada and Australia as well. European, Asian, and Latin American countries made up the remaining 3 clusters. The co-occurrence network categorized prevalent keywords into 2 clusters, the first cluster primarily focused on applying eHealth, mobile health (mHealth), or digital health to noncommunicable or chronic diseases; the second cluster was centered around the application of telemedicine and telehealth within the context of the COVID-19 pandemic.

Conclusions: Our analysis of search and bibliographic data over time and across regions allows us to gauge the interest in this topic, offer evidence regarding potential applications, and pinpoint areas for additional research and awareness-raising initiatives.

(J Med Internet Res 2024;26:e50088) doi:10.2196/50088

XSL•FO
RenderX

KEYWORDS

telemedicine; eHealth; digital medicine; COVID-19; Google Trends; bibliometric; Google; data spanning; COVID-19; accessibility; cost reduction; cost; noncommunicable disease; eHealth; mobile health; awareness; policy decision

Introduction

The concept of telemedicine—defined as the use of technology to provide health care services remotely—has been around since the early 20th century, but it has gained significant attention globally only in recent years [1]. The development of technology such as videoconferencing, remote monitoring devices, and mobile health apps has made it possible to deliver health care services remotely, paving the way for the widespread adoption of telemedicine [2-4]. It is worth noting that the term telemedicine also covers health data analysis and the application of big data and artificial intelligence methods for epidemiological research and diagnosis support [3,5-7].

The use of technology allows patients to access care from anywhere, at any time and reduces the need for in-person visits, which can be particularly beneficial for individuals with mobility issues, those living in rural or remote areas, and individuals with chronic diseases [8,9]. The potential benefits of telemedicine are numerous, including increased access to health care services, reduced health care costs, and improved patient outcomes [8-11]. Telemedicine can also help to address workforce shortages in health care, particularly in rural and remote areas, by enabling health care providers to deliver care to patients in those regions without the need for travel [10,12].

The COVID-19 pandemic has underscored the significance of telemedicine in maintaining uninterrupted health care delivery while mitigating the risk of virus transmission [13-18]. As an illustration, in light of the imperative to reduce COVID-19 exposure among patients and health care providers, many elective surgical procedures were rescheduled, prompting surgeons to adopt telemedicine as an alternative for preoperative, follow-up, and urgent surgical care consultations [15,19]. The pandemic has therefore forced health care providers to adapt quickly to the new reality of delivering care remotely, leading to a significant increase in the adoption and use of telemedicine globally [3,20]. As demand grew exponentially during the pandemic period, even the telehealth market is expected to grow to US $218.5 billion by 2025 [3]. Nevertheless, even after the COVID-19 pandemic recedes, it is highly improbable that this mode of health care delivery will be disregarded.

Despite the potential benefits of telemedicine, its adoption and implementation have been slow in some regions. Barriers to adoption include regulatory challenges, technological limitations, and resistance from health care providers and patients. People living in poorer regions, women, the elderly, and those living in rural or remote areas are far less likely to be online than those in wealthier regions [1,11]. Of those connected, nearly 90% use mobile devices to access the internet, which might not be appropriate for delivering digital health services [3]. A digital divide also exists in terms of digital literacy or low skills, which is a concern for the poorest, elderly, and others with limited access to technology [3]. For these reasons, understanding the level of interest in telemedicine among the public and research community can help identify areas for improvement. Google Trends is a valuable tool for investigating the level of interest of the general public in a specific topic. As such, it has been used in previous research to analyze the public's interest in various health and health care–related subjects [21-32]. The same applies to a topic such as telemedicine, for which there is little evidence. From a research standpoint, we have recently observed a significant increase in the number of publications related to telemedicine, indicating a growing interest in this field among researchers. An interest that needs to be mapped in terms of scientists and groups of researchers, countries that are contributing most to the research, and areas of applications.

Overall, this study aims to provide a comprehensive analysis of the level of public and research interest in telemedicine, also consider any potential impact of the COVID-19 pandemic. We therefore limit our analysis to the period between 2017 and 2022, a period of 6 years preceding and following the pandemic. We use 2 different approaches to analyze the level of interest in telemedicine. The first approach is an analysis of Google Trends data, which allows us to analyze the level of interest in telemedicine among the general public. The second approach is a bibliometric data analysis, which involves examining publications related to telemedicine over the past 6 years. Our analysis of search and bibliographic data over time and across regions helps us understand the level of interest in this topic, identify areas for further research and awareness-raising efforts, and inform policy decisions regarding telemedicine adoption and implementation.

Methods

Data Collection

Google Trends Data

Google Trends provides open access to time-series data related to Google searches for specific terms and topics [33]. According to the framework proposed by Mavragani and Ochoa [34], we retrieved Google Trends data separately by using the search topics "telemedicine" or "e-health," encompassing all search categories. We queried Google Trends on March 9, 2023, and Google Trends data were obtained at the global level, as well as by country, for the period between January 1, 2017, and December 31, 2022. It is worth mentioning that search topics are a group of terms that share the same concept across languages, covering an array of variations, typos, and related searches [34,35]. This precludes the need to enter a set of individual keywords, while maintaining the consistency of search queries across all regions and timeframes [34,35]. Furthermore, using specific search topics without any search category restrictions proves beneficial for capturing the general interest of diverse populations [34,35]. We also obtained data for the "top related topics" that are most frequently searched with the topics under investigation (ie, telemedicine" or "e-health").

XSL•FO
RenderX

In general, Google Trends data are provided as a normalized measure (ie, relative search volume [RSV]), obtained by dividing the search volume for a given term or topic by the total number of searches. This normalization process resulted in a percentage scale, with 100% corresponding to the peak in search volume in any given time frame and location. The value 0% does not necessarily indicate no searches, but rather a very low search volume for a given term or topic [34].

Bibliometric Data

Before describing the collection of bibliometric data, it is necessary to distinguish bibliometric analysis from reviews and systematic reviews of scientific literature. The first primarily uses a mechanistic method to track the global research trends in a certain field based on the outputs of scientific literature databases. Reviews and systematic reviews are instead characterized by methodical and replicable methodologies to find, select, and synthesize all available evidence on a particular topic or clinical question.

In our study, we searched the Scopus database for all articles mentioning the terms “telemedicine” or “eHealth” in the title, abstract, and keywords. We have chosen Scopus because it is recognized as the largest scientific literature database of peer-reviewed articles covering a wide range of subjects [36]. The literature search was conducted on 9th March 2023 and was limited to original articles published in English and in the subject area of medicine from 2017 to 2022. The query string used for the search was: TITLE-ABS-KEY (“telemedicine” OR “eHealth”) AND PUBYEAR > 2016 AND PUBYEAR < 2023 AND (LIMIT-TO (DOCTYPE , “ar”)) AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (SUBJAREA , “MEDI”)). The Scopus search result was exported in the format of a CSV file with all data elements, including information on citation, bibliography, abstract, and keywords.

Data Analysis

Google Trends Data Analysis

We first conducted a univariate analysis and compared the public interest on the topics of telemedicine and eHealth at the global level. Next, a joinpoint regression analysis was carried out to identify possible time points at which public interest trends changed. This analysis was conducted on log-transformed RSV, with 5000 permutations and assuming uncorrelated errors. The grid search method was chosen to determine where to locate joinpoints on the timescale [37]. Results were reported as the monthly percent change (MPC), calculated as the average percentage change per month between different joinpoints. Joinpoint regression analysis was performed using the Joinpoint Regression Program (version 4.3.1.0; Statistical Research and Applications Branch, National Cancer Institute) provided by the Surveillance, Epidemiology, and End Results Program (National Cancer Institute) on the website. We next mapped the public interest in telemedicine and eHealth by country, selecting the top 5 countries with the greatest interest. In these countries, Google Trends data were correlated with the number of new confirmed COVID-19 cases, deaths, hospitalizations, and patients in the intensive care unit per million residents. These data were obtained from the Our World in Data website [38]. Results were reported as the Spearman rank correlation coefficient (ρ). All statistical analyses were 2-tailed and performed with a significance level of 0.05.

Descriptive Analysis of Bibliometric Data

Descriptive analysis of bibliometric data was performed using Bibliometrix (K-Synth), an open-source R-tool for automating the stages of data analysis and data visualization [39]. After loading and converting bibliometric data in R (R Core Team), the main descriptive results were summarized as the number of documents, authors, sources, keywords, timespan, and average number of citations. Accordingly, tables and visualizations were obtained for the annual scientific production, top articles per number of citations, most productive authors, most productive countries, total citations per country, most relevant journals, and most relevant keywords.

Network Analysis of Bibliometric Data

Next, the VOSviewer software (version 1.6.16; Centre for Science and Technology Studies, Leiden University) was used to construct networks projecting authors’ and countries’ collaborations, as well as trending research topics through the analysis of keywords. This mapping method is generally used to estimate the association strength between different bibliometric items (ie, the nodes of the network), which may for example be publications, authors, or keywords [40]. The relation between 2 items is represented by a link (ie, the edges of the network), which may be a bibliographic coupling link between publications, a coauthorship link between authors, or a co-occurrence link between keywords. Each link has a strength, indicated by a positive number with higher values associated with stronger links [40]. The association strength may for example be indicated as the number of cited references 2 publications have in common for the bibliographic coupling link; as the number of publications 2 authors have coauthored for the coauthorship links; or as the number of publications in which 2 keywords occur together for the co-occurrence link [40]. Accordingly, each item receives different attributes, such as the weight and score attributes. Weight is a nonnegative numerical attribute indicating the importance of the item in the network. From a graphical perspective, items with higher weights are shown more prominently than those with lower weights [40]. Score attributes instead indicate additional numerical properties of the item, which can be only visualized in the overlay visualization of a map. There are also 2 standard weight attributes that can be used for descriptive purposes and computed for each item or the entire network, the links attribute and the total link strength attribute [40]. By creating the network map, items can be grouped into clusters, which correspond to linked items, labeled with colors and numbers.

We first performed network analyses of coauthorship at the author and country level, using the fractional method to reduce the influence of documents with many authors. With this approach, the strength of a coauthorship between 2 authors is determined considering the number of documents coauthored normalized for the total number of authors of each coauthored document. At the author level, we included authors who have published at least 20 articles on the topic, with no restrictions on the number of citations. At the country level, we included

XSL•FO
RenderX

the first 50 countries with the greatest number of articles, with no restrictions on the number of citations. To identify research areas of greatest interest and their connections, we also applied a co-occurrence analysis of author keywords occurring more than 50 times.

Results

Public Interest Over Time

Globally, the mean public interest in telemedicine and eHealth—expressed as RSV—was 26.3% (SD 18.1%) and 17.6% (SD 8.8%), respectively. The greater interest in telemedicine rather than in eHealth is evident from Figure 1, which shows RSVs from January 2017 to December 2022. According to the above figure, public interest in both topics was stable before the COVID-19 pandemic and then increased rapidly for telemedicine and more gradually for eHealth. Based on the joinpoint regression analysis (Figure S1 in Multimedia Appendix 1), public interest in telemedicine was stable until January 2020 (MPC=0.36%), which corresponded to the first joinpoint (37th month of the time series; 95% CI 36-38). Then, public interest suddenly increased (MPC=95.7%) to the highest peak reached in April 2020, which was the second joinpoint detected in the 40th month of the time series (95% CI 39-41). From that point on, public interest decreased (MPC=–22.7%) until August 2020 (ie, the third joinpoint on the 44th month; 95% CI 42-48) and then returned to being stable (MPC=–0.14%). Similarly, public interest in eHealth was stable until January 2020 (MPC=0.34%), which again was the first joinpoint of the time series (37th month; 95% CI 32-43). From that point on, public interest gradually increased (MPC=5.5%) to the highest peak reached in January 2022, which corresponded to the second joinpoint detected in the 61st month of the time series (95% CI 46-63). Then, public interest rapidly decreased (MPC=–20.1%) until May 2022 (ie, the third joinpoint on the 65th month; 95% CI 58-67) and then returned to being almost stable (MPC=1.8%).

Figure 1. Public interest in telemedicine and eHealth from January 2017 to December 2020, was assessed through the Google Trends analysis. This graph shows the RSVs for the topics of telemedicine (blue line) and eHealth (orange line) over time. The dotted lines represent the average values. RSV: relative search volume.

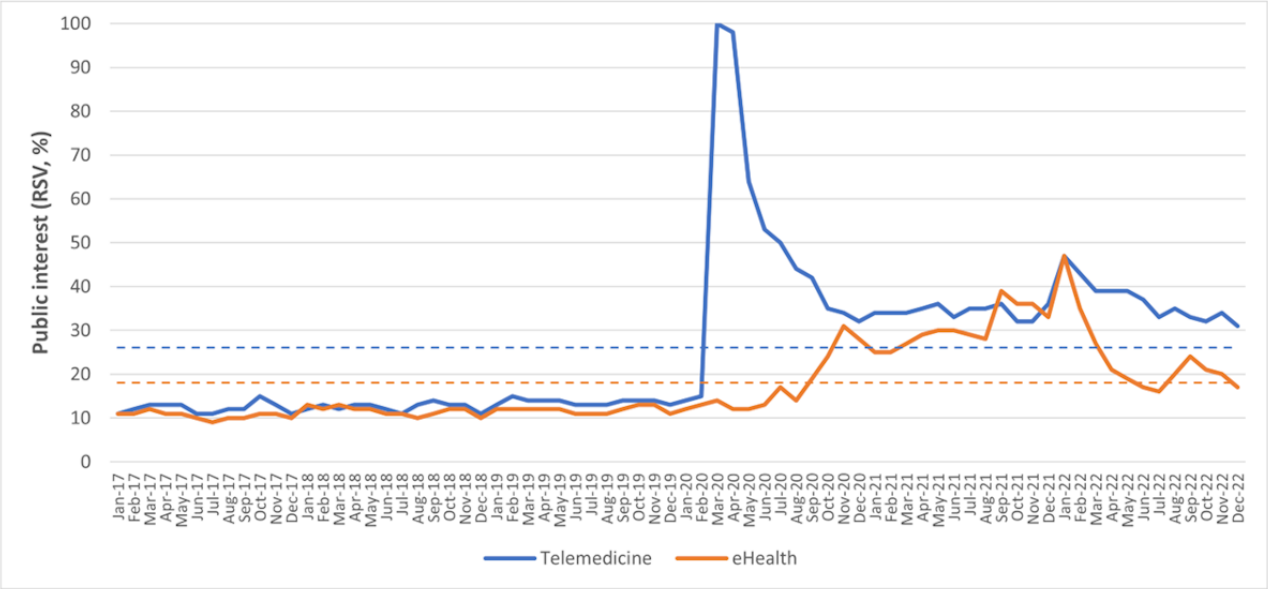

Geographic Distribution of Public Interest and Correlations With COVID-19 Data

The map in Figure 2 shows the geographic distribution of public interest in telemedicine from 2017 to 2022. The top 5 countries were Chile, Australia, Canada, the United States, and Puerto Rico (Figure S2 in Multimedia Appendix 1). Public interest in eHealth was less widespread (Figure S3 in Multimedia Appendix 1), therefore we have not considered this topic for further analyses. For those countries included in the top 5, we evaluated correlations between the RSV for telemedicine and COVID-19 data per million residents. Public interest in Chile was strongly correlated with the number of new cases (ρ=0.718; P<.001), and weakly with the number of new deaths (ρ=0.321; P=.006). In Australia, public interest moderately correlated with the number of new cases (ρ=0.537; P<.001), new deaths (ρ=0.505; P<.001), hospitalized cases (ρ=0.559; P<.001), and patients in the intensive care unit (ρ=0.483; P<.001). Public interest in Canada was weakly correlated with the number of new cases (ρ=0.265; P=.001) and deaths (ρ=0.333; P<.001). In the United States, public interest was weakly correlated with the number of new deaths (ρ=0.295; P<.001). No correlations were evident for the public interest in Puerto Rico.

XSL•FO
RenderX

Figure 2. Geographic distribution of the public interest in telemedicine from January 2017 to December 2022, was assessed through the Google Trends analysis. This choropleth map illustrates the RSVs for the topic of telemedicine at the country level. RSV: relative search volume.

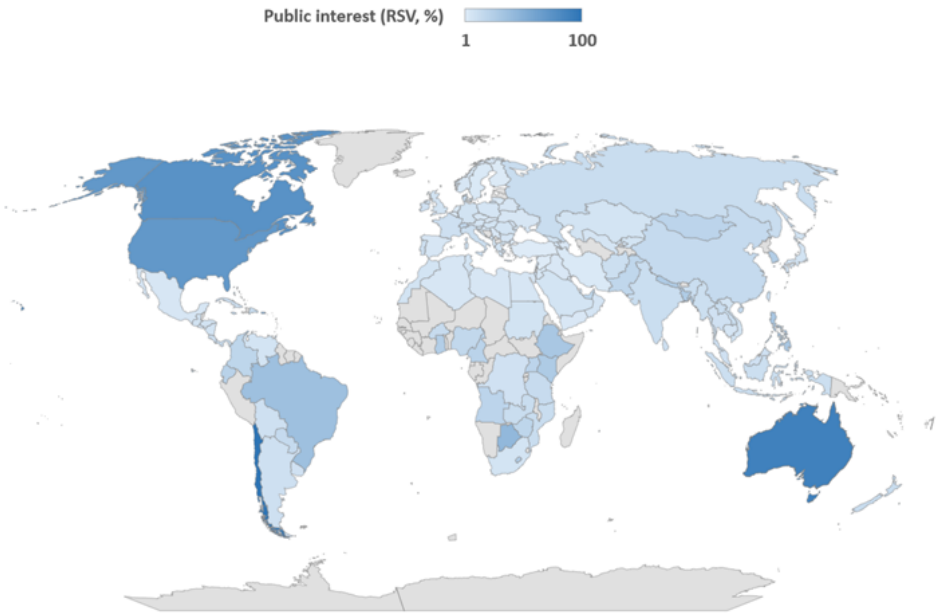

Common Topics Related to Telemedicine

Despite the considerable impact of the pandemic on the public interest in telemedicine, there were few COVID-19-related topics among those that were commonly searched with telemedicine (Figure S4 in Multimedia Appendix 1). In particular, Google users interested in telemedicine mainly searched for general topics including health, physician, health care, medicine, and therapy. Moreover, there were some common topics related to specific telemedicine providers (eg, "Telehealth Ontario," "Medicare," "Santa Catarina," "Cigna," and "CVS Pharmacy").

Description of Bibliometric Data

We identified 19,539 original medical articles published in English and indexed in the Scopus database from 2017 to 2022. These articles were published by 2824 different sources (eg, journals and books), with a document average age of 2.8 years. Notably, the first 2 sources (JMIR and Telemedicine and eHealth) published a number of articles that were nearly 40% of the sum of the first 10 sources (Table S1 in Multimedia Appendix 1). The overall average number of citations per document was 11.2, while the average number of citations per document and per year was 2.6. The top 10 articles per citation are reported in Table S2 in Multimedia Appendix 1, their total citations ranged from 506 to 999, while their total citations per year ranged from 121 to 221. A total of 553,913 references were found in the bibliometric analysis of all the articles. The top 10 cited references are reported in Table S3 in Multimedia Appendix 1, with total citations ranging from 70 to 161.

Research Interest Over Time

Our analysis of the bibliometric data also revealed a significant increase in the number of publications related to telemedicine over the past 6 years (Figure 3A). The number of publications per year increased from 1615 in 2017 to 4870 in 2022, showing an overall increase of 201.5% and an annual average growth rate of 24.7%. The greatest increase has been reported between 2019 and 2020, with an annual growth rate of 89.0%. Regarding citations, the average total citations per year decreased from 2017 to 2022 (Figure 3B), probably as a consequence of different document ages. By contrast, the average article citations per year were stable from 2017 to 2019, followed by a peak in 2020, and then decreased until 2022 (Figure 3C).

XSL•FO
RenderX

Figure 3. Research interest in telemedicine from 2017 to 2022, assessed through the analysis of bibliometric data. (A) Number of publications included in the bibliometric analysis from 2017 to 2022. (B) Average total citations per year from 2017 to 2022. (C) Average article citations per year from 2017 to 2022.

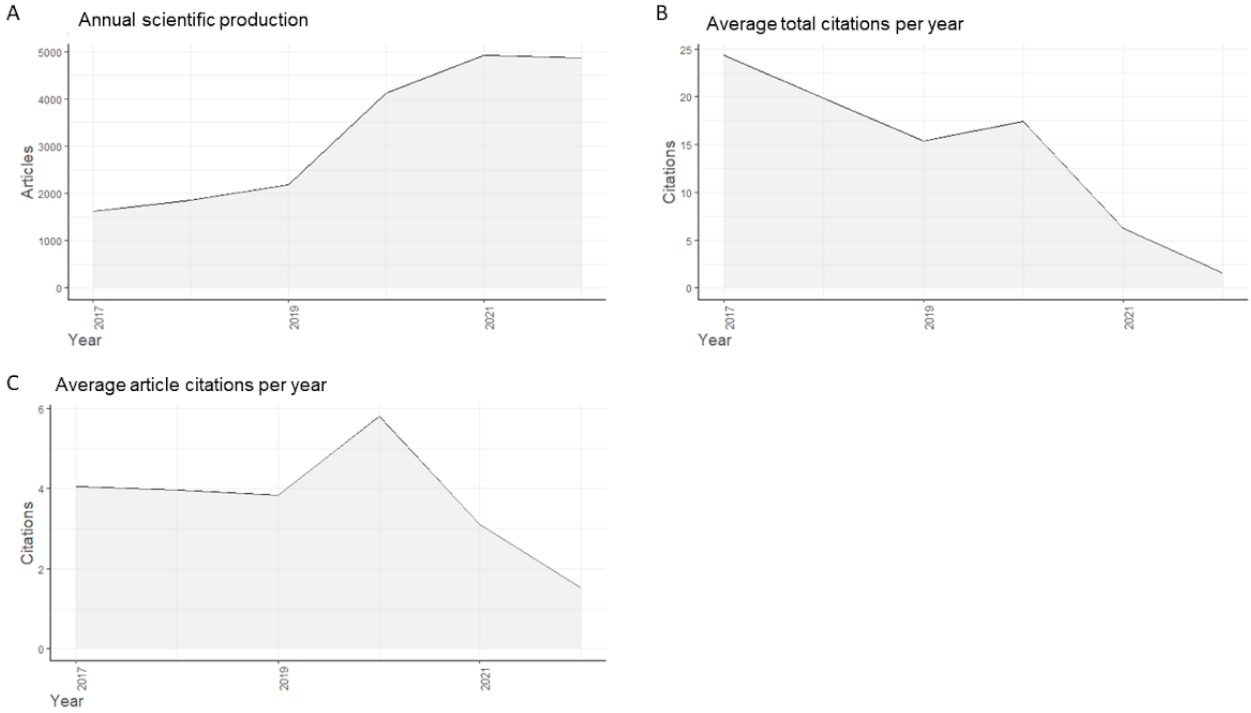

Most Productive Authors and Coauthorship Network

Among the included articles, 93,394 authors were appearing 136,956 times overall. In particular, there were 690 single-authored documents written by 642 independent authors. Accordingly, the number of documents per author was 0.2, while the number of coauthors per document was 7.0. Table S4 in Multimedia Appendix 1 shows the list of the 10 most productive authors in terms of total and fractionalized articles. Their contribution to the field ranged from 48 to 80 articles (7.2 to 10.0 fractionalized articles). In Figure 4 it is shown that 71 of the authors who published at least 20 documents (n=80) were well connected. There were 488 links between these authors—which were grouped into 7 clusters—with a total link strength of 988.

XSL•FO
RenderX

Figure 4. Coauthorship network of the most productive authors. The graphs show coauthorships between 71 authors with at least 20 publications on telemedicine, weighted by number of links. (A) Network visualization representing 7 clusters of collaborations. (B) Overlay visualization weighted by the number of links and scored by the average number of citations. The figures were prepared using VOSviewer.

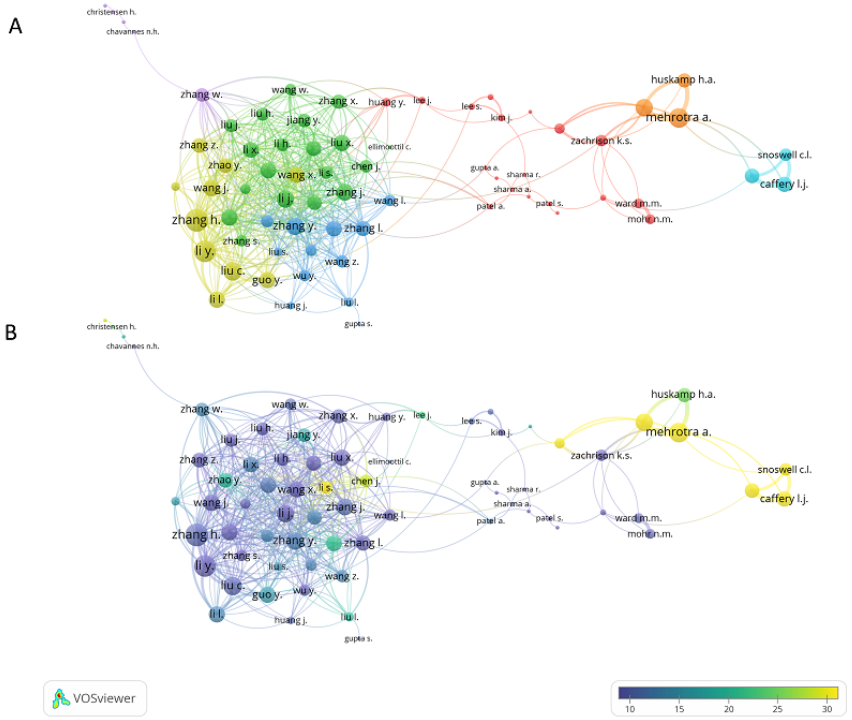

Most Contributing Countries and International Collaboration Network

The majority of publications came from a single country, while 20.8% featured international coauthorships. Table 1 shows the top 10 corresponding author's countries per document, also considering differences between single-country and multiple-country publications. One thing to note was that the United States alone published almost the same number of articles as the other 9 (6610 vs 6653), representing 33.8% of all articles analyzed. Countries with the highest proportion of multiple-country publications were Germany (31.5%), the United Kingdom (27.9%), and the Netherlands (26.8%); those with the lowest proportion were the United States (11.4%), India (20.3%), and Italy (22.5%). Similar results were evident in terms of citations, with average citations per document ranging from 8.9 for Spain to 15.2 for India (Table S5 in Multimedia Appendix 1).

Table 1. Top 10 corresponding authors' countries per document from 2017 to 2022.

Countries	Single country publications, n	Multiple country publications, n
United States	5858	752
Australia	792	281
United Kingdom	745	288
Canada	629	219
China	532	180
Germany	475	218
Netherlands	489	179
Italy	504	146
Spain	405	123
India	357	91

The coauthorship network based on the top 50 most contributing countries is illustrated in Figure 5. Overall, there were 1051 links with a total link strength of 14,751. Accordingly, countries were divided into 4 clusters. Cluster 1 consisted exclusively of European countries (n=21) with the United Kingdom, Germany, and Italy being the most interconnected; cluster 2 consisted of Asian countries (n=18) led by India and China; cluster 3 consisted of 6 countries, with the United States, Canada, and Australia being the most interconnected; and cluster 4 consisted of 5 countries of Latin America (ie, Brazil, Argentina, Chile, Colombia, and Mexico).

XSL•FO
RenderX

Figure 5. Collaboration network of the most contributing countries. The graphs show collaborations between the top 50 most contributing countries, weighted by the number of documents. (A) Network visualization representing 4 clusters of countries. (B) Overlay visualization weighted by the number of links and scored by the average number of citations. The figures were prepared using VOSviewer.

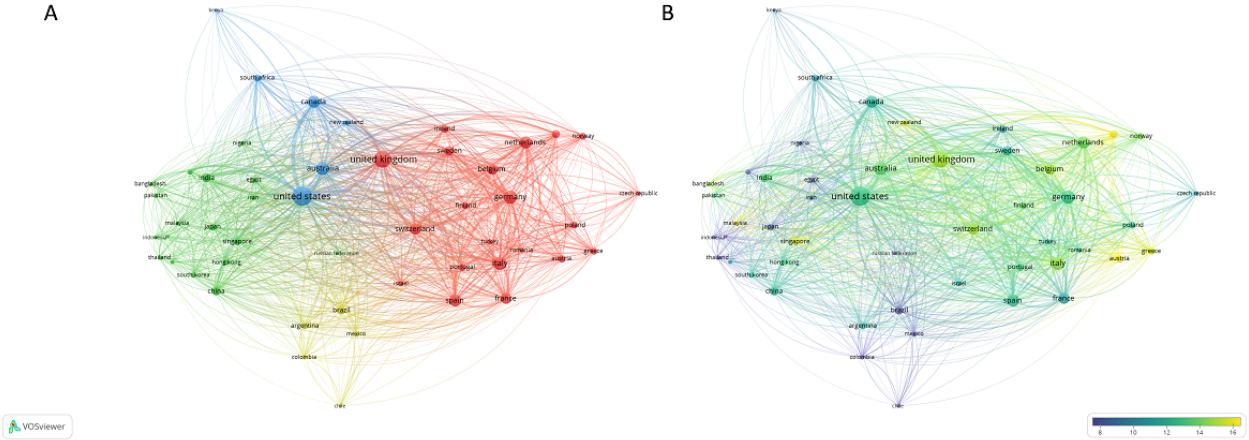

Most Relevant Keywords and Co-Occurrence Network

Overall, 22,798 author's keywords were found among the included studies. Based on the analysis of the most common keywords, COVID-19-related terms appear in the top 10 (Table S6 in Multimedia Appendix 1). In particular, this ranking put "COVID-19" and "pandemic," respectively, in the 2nd and 8th place. In addition to common terms related to telemedicine (eg, telehealth, eHealth, and mobile health [mHealth]), the analysis also revealed mental health as an area of relevant interest.

The co-occurrence network based on the top 50 most common keywords is illustrated in Figure 6. Overall, there were 1030 links with a total link strength of 21,275. Accordingly, keywords were divided into 2 clusters. Cluster 1 seemed mostly related to the application of eHealth, mHealth, or digital health to noncommunicable or chronic diseases (eg, diabetes, hypertension, depression, and anxiety) and cluster 2 was instead related to the application of telemedicine and telehealth in the context of the COVID-19 pandemic.

Figure 6. Co-occurrence network of the most common keywords. The graphs show the co-occurrence of the top 50 most common keywords, weighted by the number of links. (A) Network visualization representing 2 clusters of common keywords. (B) Overlay visualization weighted by the number of links and scored by the average number of citations. The figures were prepared using VOSviewer.

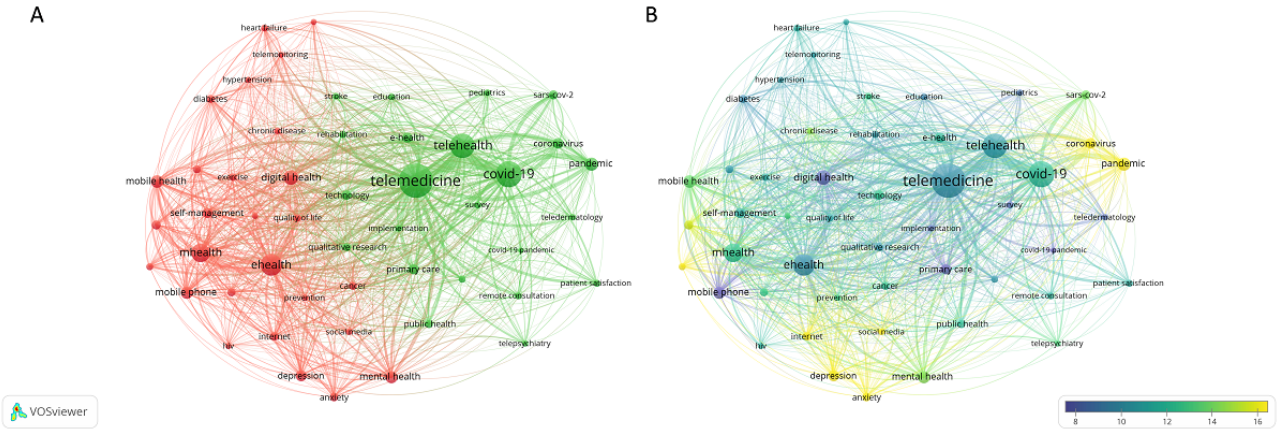

Discussion

Principal Results and Comparison With Prior Work

Our research highlights the increasing interest in telemedicine among the general public and researchers, particularly in response to the COVID-19 pandemic. Even prior to the outbreak, telemedicine had already demonstrated significant potential across various health care sectors. It offered a practical solution for delivering remote consultations, conducting preoperative evaluations, and facilitating postoperative follow-ups, particularly for patients residing in remote or underserved areas [9,19]. Although telemedicine comes with numerous advantages, it also presents certain drawbacks. For example, it may prove unsuitable for emergency situations requiring immediate hands-on medical attention, necessitating traditional, in-person emergency care [41]. Additionally, specific diagnostic procedures, like imaging or laboratory tests, may demand specialized equipment unavailable in a patient's home, potentially impacting diagnostic accuracy [42]. Furthermore, not all individuals have access to the requisite technology for telemedicine consultations, such as a reliable internet connection, a suitable device, or the technical skills for virtual appointments [43]. This discrepancy in access can result in health care disparities. Finally, telemedicine involves transmitting sensitive health information over digital networks, posing challenges in ensuring the privacy and security of patient

XSL•FO
RenderX

data and carrying a risk of data breaches or unauthorized access [43].

Despite these advantages and disadvantages, the pandemic served as a catalyst for the widespread adoption and acceptance of telemedicine [14,15,18]. Notably, telemedicine has garnered increased attention, especially in the aftermath of the COVID-19 pandemic, where its significance became evident due to the implementation of social distancing measures. However, limited studies have explored the population-level interest in telemedicine, as measured by tools such as Google Trends. Some of these studies are summarized in Table S7 in Multimedia Appendix 1 [44-51]. To the best of our knowledge, our study provides the most recent and comprehensive analysis in this field, shedding light on the level of interest among both the general public and researchers. An example of the use of infodemiological methods to explore global interest in telemedicine during the COVID-19 pandemic can be seen in the study conducted by Leochico et al [51] in 2020. Their study revealed a significant surge in online searches for telemedicine and related terms after the outbreak of the pandemic [51]. Extending the analysis period until the end of 2022, our study showed that while the level of public interest in telemedicine experienced a significant increase from January to April 2020, it later declined until August 2020 and eventually stabilized. Nevertheless, public interest has remained slightly higher compared to the pre-pandemic period. This finding is consistent with the results of a Google Trends analysis conducted by Wong et al [47] on the 50 countries most affected by the COVID-19 pandemic until July 2020. According to our findings, Chile, Australia, Canada, and the United States demonstrated the highest levels of public interest in telemedicine. In these countries, we observed moderate to strong correlations between Google Trends and COVID-19 data, including new cases and deaths, as well as hospitalizations. This finding is consistent with the results reported by Arshad Ali et al [45], who found a significant global correlation between the increase in COVID-19 cases and deaths and the interest in telemedicine.

With regard to research interest, there are examples of bibliometric analyses that existed even before the COVID-19 pandemic. For instance, Armfield et al [52] analyzed nearly 18,000 publication records, published between 2009 and 2013, to investigate the themes in telemedicine and telehealth literature. They found that the majority of studies focused on the clinical effectiveness of telemedicine. Other research questions include the adoption and implementation of telemedicine and eHealth technologies in health care systems [52]. Edirippulige et al [53] conducted a bibliometric analysis of telemedicine-related literature published until 2018 in highly ranked clinical journals and revealed that the acceptance of telemedicine research by these journals indicated a maturing of the telemedicine field. However, the pandemic has led to a surge of research interest in telemedicine and related fields. Our bibliometric analysis, in fact, revealed a considerable increase in the number of publications, particularly from 2020 onwards. Most of these publications came from a single country, with only 1 in 5 featuring international collaborations. Despite the importance of fostering collaborations among various stakeholders, including academics, health administrators, practitioners, policymakers, and communities, which involve reciprocal knowledge translation, such partnerships are often lacking [54]. The United States led the way as the most productive country, with Canada and Australia following in a cluster. Meanwhile, European, Asian, and Latin American countries comprised the other 3 clusters. Previous studies already demonstrated the predominant role of the United States, with Lan et al [55] providing a general overview, and Kumar et al [56] analyzing the trends in orthopedics and trauma-related telemedicine during the COVID-19 pandemic. Another aspect of our analysis focused on whether the different terms, such as telemedicine, telehealth, eHealth, mHealth, and so forth, can be used interchangeably or if each of them refers to a specific area of research. Fatehi and Wootton [57] conducted a bibliometric analysis in 2012 to examine the trends in the use of terms such as telemedicine, telehealth, and eHealth. They discovered that these terms were frequently used interchangeably, with a growing prevalence of the term “eHealth” in more recent years. On the contrary, our analysis revealed a specific focus on the use of eHealth, mHealth, or digital health for noncommunicable and chronic diseases, while telemedicine and telehealth were predominantly used in the context of the COVID-19 pandemic. The analysis by Lan et al [55], limited to the application of telemedicine to COVID-19, showed similar results. In particular, telemedicine was mainly used to provide mental health services, health care services delivery, and to control cross-infection. In contrast, the term “mobile apps” was closely associated with chronic illness entities such as diabetes, heart failure, and asthma, as well as health service entities such as patient education and self-care [55].

Limitations

While Google Trends provides a valuable tool for analyzing public interest, there are several limitations to this approach. First, Google Trends only provides information on internet searches and does not account for offline discussions, media coverage, or other forms of engagement with the topic. This means that our results may not be representative of the entire population or capture the full extent of public interest. Second, the data provided by Google Trends are aggregated and anonymous, making it difficult to determine the specific demographics, motivations, or intentions behind the searches. This can limit the ability to draw meaningful conclusions or make accurate predictions about public behavior or attitudes toward telemedicine. Third, Google Trends data may be subject to various biases, such as the effect of media coverage or search engine optimization strategies. Additionally, the results may be influenced by factors such as seasonality, news events, or changes in search algorithms, making it challenging to compare trends over time or across different regions. Some of these limitations also apply to bibliometric analysis. First, bibliometric data may not reflect the complete picture of research interest as not all research is published and indexed in databases. Some research may be unpublished or published in non-indexed sources. Second, bibliometric analysis may not capture changes in research interest in real time. It can take some time for research to be published, indexed, and reflected in bibliometric data, meaning that the data may not reflect the most current

XSL•FO
RenderX

state of research interest. Third, bibliometric analysis may not capture the full range of research interest as it is limited to the keywords used in publications. For instance, to offer a more comprehensive overview, one might consider examining interest in other trending subjects, like generative artificial intelligence and large language models. Finally, bibliometric analysis does not provide insights into the reasons behind the trends observed. It is limited to providing quantitative data on the number and frequency of publications and citations, and cannot provide qualitative insights into the motivations or drivers behind the research interest. For all these reasons, while the analysis of Google Trends and bibliometric data can provide a useful starting point for understanding public and research interest in telemedicine, it is essential to supplement this analysis with other sources of data and to interpret the results with caution.

Conclusions

Our study offers a comprehensive picture of the evolving landscape of telemedicine and its growing importance in health care delivery. By analyzing search and bibliographic data across regions and over time, our study provides valuable insights into the level of interest in telemedicine. This information serves to pinpoint potential application fields, identify gaps in this research, and emphasize areas that warrant additional attention and efforts in raising awareness.

Data Availability

The data sets analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AM, MB, and AA conceived and designed the study. AM and MB contributed to data collection. AM and MB contributed to the data analysis. AM, MB, GB, and AA contributed to the interpretation of results. AM, MB, GB, and AA contributed to writing the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional outputs from the examination of Google Trends and bibliometric data.
[DOCX File , 249 KB - jmir_v26i1e50088_app1.docx]

References

1. A health telematics policy in support of WHO's health-for-all strategy for global health development: report of the WHO group consultation on health telematics. WHO. 1997. URL: https://pesquisa.bvsalud.org/portal/resource/pt/who-63857 [accessed 2024-02-03]
2. Craig J, Patterson V. Introduction to the practice of telemedicine. J Telemed Telecare 2005;11(1):3-9. [doi: 10.1177/1357633X0501100102] [Medline: 15829036]
3. The rise of digital health technologies during the pandemic. European Parliamentary Research Service. 2021. URL: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)690548 [accessed 2024-02-03]
4. Barchitta M, Maugeri A, Favara G, Lio RMS, Riela PM, Guarnera L, et al. Development of a web-app for the ecological momentary assessment of dietary habits among college students: the HEALTHY-UNICT project. Nutrients 2022;14(2):330 [FREE Full text] [doi: 10.3390/nu14020330] [Medline: 35057511]
5. Barchitta M, Maugeri A, Favara G, Riela PM, Gallo G, Mura I, et al. A machine learning approach to predict healthcare-associated infections at intensive care unit admission: findings from the SPIN-UTI project. J Hosp Infect 2021;112:77-86 [FREE Full text] [doi: 10.1016/j.jhin.2021.02.025] [Medline: 33676936]
6. Barchitta M, Maugeri A, Favara G, Riela PM, La Mastra C, La Rosa MC, et al. Cluster analysis identifies patients at risk of catheter-associated urinary tract infections in intensive care units: findings from the SPIN-UTI network. J Hosp Infect 2021;107:57-63 [FREE Full text] [doi: 10.1016/j.jhin.2020.09.030] [Medline: 33017617]
7. Barchitta M, Maugeri A, Favara G, Riela PM, Gallo G, Mura I, et al. Early prediction of seven-day mortality in intensive care unit using a machine learning model: results from the SPIN-UTI project. J Clin Med 2021;10(5):992 [FREE Full text] [doi: 10.3390/jcm10050992] [Medline: 33801207]
8. Flodgren G, Rachas A, Farmer AJ, Inzitari M, Shepperd S. Interactive telemedicine: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2015;2015(9):CD002098 [FREE Full text] [doi: 10.1002/14651858.CD002098.pub2] [Medline: 26343551]
9. Gonçalves-Bradley DC, Maria J, Ricci-Cabello I, Villanueva G, Fønhus MS, Glenton C, et al. Mobile technologies to support healthcare provider to healthcare provider communication and management of care. Cochrane Database Syst Rev 2020;8(8):CD012927 [FREE Full text] [doi: 10.1002/14651858.CD012927.pub2] [Medline: 32813281]

XSL•FO
RenderX

10. Wootton R. Telemedicine support for the developing world. J Telemed Telecare 2008;14(3):109-114. [doi: 10.1258/jtt.2008.003001] [Medline: 18430271]
11. Jennett PA, Hall LA, Hailey D, Ohinmaa A, Anderson C, Thomas R, et al. The socio-economic impact of telehealth: a systematic review. J Telemed Telecare 2003;9(6):311-320. [doi: 10.1258/135763303771005207] [Medline: 14680514]
12. Rao B, Lombardi A. Telemedicine: current status in developed and developing countries. J Drugs Dermatol 2009;8(4):371-375. [Medline: 19363855]
13. Tilahun B, Gashu KD, Mekonnen ZA, Endehabtu BF, Angaw DA. Mapping the role of digital health technologies in prevention and control of COVID-19 pandemic: review of the literature. Yearb Med Inform 2021;30(1):26-37 [FREE Full text] [doi: 10.1055/s-0041-1726505] [Medline: 34479378]
14. Blue R, Yang AI, Zhou C, De Ravin E, Teng CW, Arguelles GR, et al. Telemedicine in the era of coronavirus disease 2019 (COVID-19): a neurosurgical perspective. World Neurosurg 2020;139:549-557 [FREE Full text] [doi: 10.1016/j.wneu.2020.05.066] [Medline: 32426065]
15. Gachabayov M, Latifi LA, Parsikia A, Latifi R. The role of telemedicine in surgical specialties during the COVID-19 pandemic: a scoping review. World J Surg 2022;46(1):10-18 [FREE Full text] [doi: 10.1007/s00268-021-06348-1] [Medline: 34743242]
16. Curcic IB, Berkovic MC, Kizivat T, Varzic SC, Smolic R, Smolic M. Effect of COVID-19 on management of type 1 diabetes: pushing the boundaries of telemedical healthcare. World J Diabetes 2021;12(6):780-785 [FREE Full text] [doi: 10.4239/wjd.v12.i6.780] [Medline: 34168727]
17. Umano GR, Di Sessa A, Guarino S, Gaudino G, Marzuillo P, Del Giudice EM. Telemedicine in the COVID-19 era: taking care of children with obesity and diabetes mellitus. World J Diabetes 2021;12(5):651-657 [FREE Full text] [doi: 10.4239/wjd.v12.i5.651] [Medline: 33995852]
18. Kronenfeld JP, Penedo FJ. Novel coronavirus (COVID-19): telemedicine and remote care delivery in a time of medical crisis, implementation, and challenges. Transl Behav Med 2021;11(2):659-663 [FREE Full text] [doi: 10.1093/tbm/ibaa105] [Medline: 33098426]
19. Ghomrawi HMK, Holl JL, Abdullah F. Telemedicine in surgery-beyond a pandemic adaptation. JAMA Surg 2021 Oct 01;156(10):901-902 [FREE Full text] [doi: 10.1001/jamasurg.2021.2052] [Medline: 34259819]
20. Bouabida K, Lebouché B, Pomey M. Telehealth and COVID-19 pandemic: an overview of the telehealth use, advantages, challenges, and opportunities during COVID-19 pandemic. Healthcare (Basel) 2022;10(11):2293 [FREE Full text] [doi: 10.3390/healthcare10112293] [Medline: 36421617]
21. Bragazzi NL, Alicino C, Trucchi C, Paganino C, Barberis I, Martini M, et al. Global reaction to the recent outbreaks of Zika virus: insights from a big data analysis. PLoS One 2017;12(9):e0185263 [FREE Full text] [doi: 10.1371/journal.pone.0185263] [Medline: 28934352]
22. Galido A, Ecleo JJ, Husnayain A, Su ECY. Exploring online search behavior for COVID-19 preventive measures: the Philippine case. PLoS One 2021;16(4):e0249810 [FREE Full text] [doi: 10.1371/journal.pone.0249810] [Medline: 33831076]
23. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012-1014 [FREE Full text] [doi: 10.1038/nature07634] [Medline: 19020500]
24. Greiner B, Ottwell R, Vassar M, Hartwell M. Public interest in preventive measures of coronavirus disease 2019 associated with timely issuance of statewide stay-at-home orders. Disaster Med Public Health Prep 2020;14(6):765-768 [FREE Full text] [doi: 10.1017/dmp.2020.189] [Medline: 32498752]
25. Hamulka J, Jeruszka-Bielak M, Górnicka M, Drywień ME, Zielinska-Pukos MA. Dietary supplements during COVID-19 outbreak. Results of Google Trends analysis supported by PLifeCOVID-19 online studies. Nutrients 2020;13(1):54 [FREE Full text] [doi: 10.3390/nu13010054] [Medline: 33375422]
26. Keitoku K, Nishimura Y, Hagiya H, Koyama T, Otsuka F. Impact of the world antimicrobial awareness week on public interest between 2015 and 2020: a Google Trends analysis. Int J Infect Dis 2021;111:12-20 [FREE Full text] [doi: 10.1016/j.ijid.2021.08.018] [Medline: 34391908]
27. Kurian SJ, Bhatti AUR, Alvi MA, Ting HH, Storlie C, Wilson PM, et al. Correlations between COVID-19 cases and Google Trends data in the United States: a state-by-state analysis. Mayo Clin Proc 2020;95(11):2370-2381 [FREE Full text] [doi: 10.1016/j.mayocp.2020.08.022] [Medline: 33164756]
28. Maugeri A, Barchitta M, Basile G, Agodi A. How COVID-19 has influenced public interest in antimicrobials, antimicrobial resistance and related preventive measures: a Google Trends analysis of Italian data. Antibiotics (Basel) 2022;11(3):379 [FREE Full text] [doi: 10.3390/antibiotics11030379] [Medline: 35326842]
29. Maugeri A, Barchitta M, Agodi A. Using Google Trends to predict COVID-19 vaccinations and monitor search behaviours about vaccines: a retrospective analysis of Italian data. Vaccines (Basel) 2022;10(1):119 [FREE Full text] [doi: 10.3390/vaccines10010119] [Medline: 35062780]
30. Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google Trends time series. Sci Rep 2020;10(1):20693 [FREE Full text] [doi: 10.1038/s41598-020-77275-9] [Medline: 33244028]
31. Mayasari NR, Ho DKN, Lundy DJ, Skalny AV, Tinkov AA, Teng IC, et al. Impacts of the COVID-19 pandemic on food security and diet-related lifestyle behaviors: an analytical study of Google trends-based query volumes. Nutrients 2020;12(10):3103 [FREE Full text] [doi: 10.3390/nu12103103] [Medline: 33053656]

XSL•FO
RenderX

32. Maugeri A, Barchitta M, Perticone V, Agodi A. How COVID-19 pandemic has influenced public interest in foods: a Google Trends analysis of Italian data. Int J Environ Res Public Health 2023;20(3):1976 [FREE Full text] [doi: 10.3390/ijerph20031976] [Medline: 36767342]
33. Arora VS, McKee M, Stuckler D. Google Trends: opportunities and limitations in health and health policy research. Health Policy 2019;123(3):338-341. [doi: 10.1016/j.healthpol.2019.01.001] [Medline: 30660346]
34. Mavragani A, Ochoa G. Google Trends in infodemiology and infoveillance: methodology framework. JMIR Public Health Surveill 2019;5(2):e13439 [FREE Full text] [doi: 10.2196/13439] [Medline: 31144671]
35. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical approaches in Google Trends research: systematic review. J Med Internet Res 2018;20(11):e270 [FREE Full text] [doi: 10.2196/jmir.9366] [Medline: 30401664]
36. Khudzari JM, Kurian J, Tartakovsky B, Raghavan GSV. Bibliometric analysis of global research trends on microbial fuel cells using scopus database. Biochem Eng J 2018;136:51-60 [FREE Full text] [doi: 10.1016/j.bej.2018.05.002]
37. Lerman PM. Fitting segmented regression models by grid search. J R Stat Soc 1980;29(1):77-84. [doi: 10.2307/2346413]
38. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Gavrilov D, Giattino C, et al. Coronavirus pandemic (COVID-19). Our World in Data. URL: https://ourworldindata.org/coronavirus [accessed 2024-02-03]
39. Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr 2017:959-975 [FREE Full text] [doi: 10.1016/j.joi.2017.08.007]
40. van Eck NJ, Waltman L. VOSviewer Manual. Manual for VOSviewer version 1 .6.19. 2023. URL: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.19.pdf [accessed 2024-02-03]
41. Reed M, Huang J, Somers M, Hsueh L, Graetz I, Millman A, et al. Telemedicine versus in-person primary care: treatment and follow-up visits. Ann Intern Med 2023;176(10):1349-1357. [doi: 10.7326/M23-1335] [Medline: 37844311]
42. Calton BA, Nouri S, Davila C, Kotwal A, Zapata C, Bischoff KE. Strategies to make telemedicine a friend, not a foe, in the provision of accessible and equitable cancer care. Cancers (Basel) 2023;15(21):5121 [FREE Full text] [doi: 10.3390/cancers15215121] [Medline: 37958296]
43. Shaw RJ. Access to technology and digital literacy as determinants of health and health care. Creat Nurs 2023;29(3):258-263. [doi: 10.1177/10784535231211682] [Medline: 37909069]
44. Hong YR, Lawrence J, Williams D, Mainous I. Population-level interest and telehealth capacity of US hospitals in response to COVID-19: cross-sectional analysis of Google search and national hospital survey data. JMIR Public Health Surveill 2020;6(2):e18961 [FREE Full text] [doi: 10.2196/18961] [Medline: 32250963]
45. Arshad Ali S, Bin Arif T, Maab H, Baloch M, Manazir S, Jawed F, et al. Global interest in telehealth during COVID-19 pandemic: an analysis of Google Trends. Cureus 2020 Sep 16;12(9):e10487 [FREE Full text] [doi: 10.7759/cureus.10487] [Medline: 33083187]
46. Jimenez AJ, Estevez-Reboredo RM, Santed MA, Ramos V. COVID-19 symptom-related Google searches and local COVID-19 incidence in Spain: correlational study. J Med Internet Res 2020;22(12):e23518 [FREE Full text] [doi: 10.2196/23518] [Medline: 33156803]
47. Wong MYZ, Gunasekeran DV, Nusinovici S, Sabanayagam C, Yeo KK, Cheng CY, et al. Telehealth demand trends during the COVID-19 pandemic in the top 50 most affected countries: infodemiological evaluation. JMIR Public Health Surveill 2021;7(2):e24445 [FREE Full text] [doi: 10.2196/24445] [Medline: 33605883]
48. Alonto AHD, Jamora RDG, Leochico CFD, Espiritu AI. Low online search interest in teleneurology before and during COVID-19 pandemic: an infodemiological study. Neurol Sci 2022;43(5):2929-2934 [FREE Full text] [doi: 10.1007/s10072-022-05902-6] [Medline: 35075573]
49. Kinoshita T, Matsumoto T, Taura N, Usui T, Matsuya N, Nishiguchi M, et al. Public interest and accessibility of telehealth in Japan: retrospective analysis using Google Trends and national surveillance. JMIR Form Res 2022;6(9):e36525 [FREE Full text] [doi: 10.2196/36525] [Medline: 36103221]
50. van Kessel R, Kyriopoulos I, Wong BLH, Mossialos E. The effect of the COVID-19 pandemic on digital health-seeking behavior: big data interrupted time-series analysis of Google Trends. J Med Internet Res 2023;25:e42401 [FREE Full text] [doi: 10.2196/42401] [Medline: 36603152]
51. Leochico CFD, Austria EMV, Espiritu AI. Global online interest in telehealth, telemedicine, telerehabilitation, and related search terms amid the COVID-19 pandemic: an infodemiological study. Acta Med Philipp 2021;56(11):66-75 [FREE Full text] [doi: 10.47895/amp.vi0.3037]
52. Armfield NR, Edirippulige S, Caffery LJ, Bradford NK, Grey JW, Smith AC. Telemedicine--a bibliometric and content analysis of 17,932 publication records. Int J Med Inform 2014;83(10):715-725. [doi: 10.1016/j.ijmedinf.2014.07.001] [Medline: 25066950]
53. Edirippulige S, Senanayake B, Fatehi F, Hansen J, Bambling M, Smith AC, et al. Telemedicine: niche or mainstream? A bibliometric analysis and review of the output of highly ranked clinical journals. J Telemed Telecare 2024;30(1):53-63. [doi: 10.1177/1357633X211043376] [Medline: 34581621]
54. Gerber T, Olazabal V, Brown K, Pablos-Mendez A. An agenda for action on global e-Health. Health Aff (Millwood) 2010;29(2):233-236. [doi: 10.1377/hlthaff.2009.0934] [Medline: 20348066]
55. Lan X, Yu H, Cui L. Application of telemedicine in COVID-19: a bibliometric analysis. Front Public Health 2022;10:908756 [FREE Full text] [doi: 10.3389/fpubh.2022.908756] [Medline: 35719666]

XSL•FO
RenderX

56. Kumar A, Sinha S, Jameel J, Kumar S. Telemedicine trends in orthopaedics and trauma during the COVID-19 pandemic: a bibliometric analysis and review. J Taibah Univ Med Sci 2022;17(2):203-213 [FREE Full text] [doi: 10.1016/j.jtumed.2021.09.003] [Medline: 34690642]
57. Fatehi F, Wootton R. Telemedicine, telehealth or e-health? A bibliometric analysis of the trends in the use of these terms. J Telemed Telecare 2012 Dec;18(8):460-464. [doi: 10.1258/jtt.2012.gth108] [Medline: 23209265]

Abbreviations

mHealth: mobile health
MPC: monthly percent change
RSV: relative search volume

Edited by A Mavragani; submitted 19.06.23; peer-reviewed by D Dinh, G Gallo, D Gunasekeran; comments to author 30.11.23; revised version received 01.12.23; accepted 03.01.24; published 16.05.24.

Please cite as:
Maugeri A, Barchitta M, Basile G, Agodi A
Public and Research Interest in Telemedicine From 2017 to 2022: Infodemiology Study of Google Trends Data and Bibliometric Analysis of Scientific Literature
J Med Internet Res 2024;26:e50088
URL: https://www.jmir.org/2024/1/e50088
doi:10.2196/50088
PMID:38753427

©Andrea Maugeri, Martina Barchitta, Guido Basile, Antonella Agodi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Findings From a National Survey of Older US Adults on Patient Willingness to Use Telehealth Services: Cross-Sectional Survey

Olufeyisayo O Odebunmi[1,2*], BPharm, MPH; Tamera D Hughes[3*], PharmD, PhD; Austin R Waters[1,2*], MSPH; Benjamin Y Urick[3,4*], PharmD, PhD; Caroline Herron[5*], PharmD; Mary Wangen[6*], MPH; Catherine Rohweder[6*], MPH, DrPH; Renée M Ferrari[2,7*], PhD; Macary W Marciniak[3*], BCACP, BCPS, PharmD; Stephanie B Wheeler[1,2,6*], PhD; Alison T Brenner[2,6,8*], PhD; Parth D Shah[9*], PharmD, PhD

[1]Department of Health Policy and Management, University of North Carolina, Chapel Hill, NC, United States
[2]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, United States
[3]Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, United States
[4]Prime Therapeutics, Eagan, MN, United States
[5]Syracuse Veterans Affairs Medical Center, Syracuse, NY, United States
[6]Center for Health Promotion and Disease Prevention, University of North Carolina, Chapel Hill, NC, United States
[7]Department of Maternal and Child Health, University of North Carolina, Chapel Hill, NC, United States
[8]Division of General Medicine and Clinical Epidemiology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, NC, United States
[9]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, United States
[*]all authors contributed equally

Corresponding Author:
Parth D Shah, PharmD, PhD
Public Health Sciences Division
Fred Hutchinson Cancer Center
1100 Fairview Avenue
PO Box 19024
Seattle, WA, 98109
United States
Phone: 1 2066676120
Email: pshah@fredhutch.org

Abstract

Background: Telehealth (telemedicine and telepharmacy) services increase access to patient services and ensure continuity of care. However, few studies have assessed factors that influence patients' willingness to use telehealth services, and we sought to investigate this.

Objective: This study aims to examine respondents' (aged between 45 and 75 years) willingness to use telehealth services (telepharmacy and telemedicine) and the correlates of the willingness to use telehealth services.

Methods: We administered a cross-sectional national survey of 1045 noninstitutionalized US adults aged between 45 and 75 years in March and April 2021. Multiple logistic regression analyses were used to identify demographic and health service use correlates of self-reported willingness to use telehealth services.

Results: Overall willingness to use telemedicine was high (674/1045, 64.5%). Adults aged 55 years and older were less willing to use telemedicine (aged between 55 and 64 years: odds ratio [OR] 0.61, 95% CI 0.42-0.86; aged 65 years or older: OR 0.33, 95% CI 0.22-0.49) than those younger than 55 years. Those with a regular provider (OR 1.01, 95% CI 1-1.02) and long travel times (OR 1.75, 95% CI 1.03-2.98) were more willing to use telemedicine compared to those without a regular provider and had shorter travel times, respectively. Willingness to use telemedicine services increased from 64.5% (674/1045) to 83% (867/1045) if the service was low-cost or insurance-covered, was with their existing health care provider, or was easy-to-use. Overall willingness to use telepharmacy was 76.7% (801/1045). Adults aged older than 55 years were less willing to use telepharmacy (aged between 55 and 64 years: OR 0.57, 95% CI 0.38-0.86; aged 65 years or older: OR 0.24, 95% CI 0.15-0.37) than those younger than 55 years. Those who rated pharmacy service quality higher were more willing to use telepharmacy (OR 1.06, 95% CI 1.03-1.09) than those who did not.

XSL•FO
RenderX

Conclusions: Respondents were generally willing to use telehealth (telemedicine and telepharmacy) services, but the likelihood of their being willing to use telehealth decreased as they were older. For those initially unwilling (aged 55 years or older) to use telemedicine services, inexpensive or insurance-covered services were acceptable.

(J Med Internet Res 2024;26:e50205) doi:10.2196/50205

KEYWORDS

community pharmacy; telehealth; telemedicine; telepharmacy; pharmacy service quality; patient willingness; willingness; cross-sectional; national survey; telehealth cost

Introduction

Telehealth (eg, telemedicine and telepharmacy) has increased access to and sustainment of inpatient and outpatient services, specialty services, and pharmacy dispensing and nondispensing services (eg, vaccination counseling) [1,2]. Telehealth is defined as the provision of clinical care, patient education, and public health remotely through secure telecommunication technologies (eg, Zoom; Zoom Video Communications) [3], to enhance access to high-quality care [1,4-6]. Telehealth is a broad term that encompasses remote health care services provided by health care professionals, including physicians, nurses, and pharmacists [7,8]. Telepharmacy specifically refers to the delivery of pharmacy services, including remote dispensing, medication therapy management, and remote consultation [2,9,10]. Telemedicine is defined as the provision of clinic and diagnostic services remotely by physicians through telecommunication platforms [7]. Hence, telehealth includes telemedicine and telepharmacy.

Past studies have reported the success of remote clinic or doctor's appointments remotely providing mental health services, medication dispensing, patient education, and counseling [5], which can improve medication adherence and outcomes in chronic disease management [6,8,11,12]. Precautions during the COVID-19 pandemic, such as social distancing, necessitated telehealth services in health care settings [10]. Through this shift, the perceived value of telehealth services has grown, as it means a lower risk of infection for people who are immunocompromised, lower travel expenses, convenience for patients, increased access to health services, and better continuity of care [13,14]. Additionally, the Coronavirus Aid Relief, Economic Security Act (CARES Act) of 2020 and the Centers for Medicare and Medicaid Services eased restrictions on Medicare reimbursement for telehealth services, further increasing access to health care services delivered through telehealth [15,16].

Telehealth services (including telemedicine and telepharmacy) show great promise to improve health care access equity, but few studies examine factors associated with patients' willingness to use telehealth among older US adults. The Andersen health care utilization model posits that predisposing and enabling factors influence what type of care and how much care a person uses [17]. Predisposing factors for using telehealth include distance to the nearest physical provider or pharmacy, patient age, insurance status, or income. Telehealth (telepharmacy and telemedicine) could be considered enabling factors to increase access to patient care services. We aimed to evaluate the correlates of patients' willingness to use telehealth (referred to as telemedicine and telepharmacy in this study) services in clinical and pharmacy settings among US adults aged between 45 and 75 years.

Methods

Participants and Procedures

We conducted a cross-sectional survey, the National PharmFIT Patient Survey, on noninstitutionalized US adults aged between 45 and 75 years who were eligible for colorectal cancer screening. More details about our participants and survey sampling procedures are published elsewhere [18] and briefly described here. Our national survey included members of market research panels managed by Qualtrics, a web-based survey research company [19]. We applied sampling quotas for participants to match the 2010 US census data for race, ethnicity, and sex and oversampled rural residents to comprise a third of our sample. Respondents provided informed consent and completed the survey between March and April 2021. This study sample includes 1045 panelists who were eligible and completed the survey. Using the American Association for Public Opinion Research response rate 4 [20], we accounted for ineligibility, data quality issues, and overquotas, yielding a response rate of 62% (1274/2054). Respondents represented all 50 states; Washington, DC; and Puerto Rico.

Measures

Overview

The National PharmFIT Patient Survey developed by our multidisciplinary research team comprised 9 different domains: colorectal cancer screening experience, health care use, pharmacy type, PharmFIT program design, diffusion of innovation, willingness to use PharmFIT, follow-up care, telehealth (telemedicine and telepharmacy), and demographics. In this study, we evaluated telehealth items embedded in the National PharmFIT Patient Survey [18].

Telemedicine and Telepharmacy

In this study, we define telemedicine and telepharmacy as the delivery of health services remotely by a health care provider or pharmacist [7,8], respectively. Hereinafter, we refer to telehealth as both telemedicine and telepharmacy. A total of 3 survey items assessed respondents' previous use of and willingness to use telehealth services. The first item assessed respondents' willingness to use telemedicine services with a health care provider. The item used a 5-point response scale ranging from "strongly disagree" (1) to "strongly agree" (5). Those who strongly disagreed, disagreed, or neither agreed nor disagreed were categorized as "not willing to use telemedicine"

XSL•FO
RenderX

(coded 0), while those who agreed or strongly agreed were categorized as "willing to use telemedicine" (coded 1). Participants who were "not willing to use telemedicine" received a follow-up multiple-choice question (second item) assessing under which conditions they would be willing to use telemedicine services with a health care provider. These conditions included: (1) getting an appointment sooner than an in-person appointment, (2) getting an appointment with the respondent's usual health care provider, (3) whether the system was easy to use, (4) whether the system was private, (5) whether the appointments were available during extended hours, (6) whether insurance covered visits, (7) other reasons, or (8) none of these. The third item was a multiple-choice question assessing which health care services the participant would be willing to access through telepharmacy. We recoded this item as a dichotomous variable: 0 for "not willing to use telepharmacy" (for those who selected "none of these services" or gave a negative response in the open-ended response option) or 1 for "willing to use telepharmacy" (for those who selected any pharmacy service or stated other pharmacy services in the open-ended response section).

Demographic and Health Use Characteristics

The survey assessed respondents' sociodemographic characteristics, their usual place of receiving care, whether they had a regular provider, travel time to their regular provider, the type of pharmacy they typically use for health care needs, travel time to their typical pharmacy, and their perceptions of pharmacy service quality. Demographic characteristics included age (categorized in 10-year increments), gender, race, ethnicity, insurance type, educational attainment, household income (US $20,000 increments), and rurality. Rurality was grouped into "rural" and "nonrural" by mapping respondents' zip codes to rural-urban continuum codes [21]. Pharmacy service quality rating was developed as a composite score from items assessing how sympathetic, trustworthy, and responsive the pharmacy team was and how safe and familiar respondents felt with the pharmacy staff. These items used a 5-point response scale ranging from "strongly disagree" (1) to "strongly agree" (5) and were summarized as a continuous score, ranging from 0 to 30. High scores indicated better pharmacy service quality. The full survey instrument can be accessed online [22].

Statistical Analyses

We performed summary statistics to describe the demographic and health use factors (age, gender, race, ethnicity, insurance type, rurality, education, income, health status, place of care, regular source of care, pharmacy service quality rating, and pharmacy type) of our sample. Next, using willingness to use telemedicine and telepharmacy as our outcomes and the demographic and health usage factors listed above, which we determined a priori from Andersen's model and past literature as the predictors [17,23-25], we conducted multiple logistic regressions to assess the correlates of respondents' willingness to use telemedicine and telepharmacy. Finally, we described different conditions under which respondents indicated willingness to use telemedicine if they were initially unwilling. All analyses used a critical α=.05 and were conducted using Stata (version 16.1; StataCorp).

Ethical Considerations

The University of North Carolina institutional review board approved this study (18-1337). All participants provided informed consent online before accessing and completing the survey.

Results

Respondents' Characteristics and Willingness to Use Telemedicine and Telepharmacy Services

The majority of respondents were older than 55 years, non-Hispanic White, nonrural, had private insurance, and had household incomes below US $80,000. At least two-thirds of respondents were willing to use telemedicine (674/1045, 64.5%), and telepharmacy (801/1045, 76.7%; Table 1).

XSL•FO
RenderX

Table 1. Characteristics of respondents (N=1045).

Characteristics	Value
Sex, n (%)	
Male	523 (50)
Female	522 (50)
Age (years), n (%)	
45-54	348 (33)
55-64	348 (33)
≥65	349 (33)
Race, n (%)	
White	770 (74)
Black or African American	134 (13)
Asian	68 (7)
Multiracial or other	73 (7)
Ethnicity, n (%)	
Not Hispanic or Latinx	919 (88)
Hispanic or Latinx	126 (12)
Insurance type, n (%)	
Private	569 (56)
Public	384 (38)
Uninsured	60 (6)
Rurality, n (%)	
Not rural	731 (70)
Rural	314 (31)
Education, n (%)	
High school or less	217 (21)
Some college	408 (39)
College degree	259 (25)
Graduate education or higher	161 (15)
Household income (US $), n (%)	
<20,000	156 (15)
20,000-39,000	247 (24)
40,000-59,000	219 (21)
60,000-79,000	151 (14)
80,000-99,000	89 (9)
>100,000	183 (18)
Health status, n (%)	
Excellent	77 (7)
Very good	301 (29)
Good	443 (42)
Fair	190 (18)
Poor	34 (3)
Place to receive health care, n (%)	
A doctor's office or health care center	895 (85)

XSL•FO
RenderX

Characteristics	Value
Urgent care center or clinic in a drug store or grocery store	73 (7)
Emergency room	32 (3)
A VA[a] medical center or VA outpatient clinic	39 (4)
Some other place	6 (1)
Have provider as a regular source of care, n (%)	
No	112 (10)
Yes	933 (89)
Willingness to use telemedicine, n (%)	
No	371 (36)
Yes	674 (64)
Travel time to provider (minutes), mean (SD)	18.73 (14.19)
Pharmacy type, n (%)	
A retail chain pharmacy (like CVS[b] or Walgreens)	515 (49)
A pharmacy in a grocery store (like Kroger or Albertsons or Safeway)	164 (16)
A pharmacy in a department store or wholesaler (like Walmart or Costco)	166 (16)
A pharmacy in a clinic or hospital where medical care is received	90 (9)
A local, independently owned pharmacy	110 (11)
Willingness to use telepharmacy, n (%)	
No	244 (23)
Yes	801 (77)
Travel time to nearest pharmacy (minutes), mean (SD)	18.79 (14.51)
Pharmacy service quality rating, mean (SD)	17.51 (4.91)

[a]VA: Veteran's Affairs.

[b]CVS: A retail pharmacy chain in the United States.

Correlates of Respondents' Willingness to Use Telemedicine Services

In the multiple logistic regression model, compared to participants aged between 45 and 54 years, respondents aged between 55 and 64 years (odds ratio [OR] 0.61, 95% CI 0.42-0.86) and respondents aged 65 years or older (OR 0.33, 95% CI 0.22-0.49) were less likely to be willing to use telemedicine. Respondents who had a regular provider (OR 1.01, 95% CI 1-1.02) and had long travel times to their provider (OR 1.75, 95% CI 1.03-2.98) were more likely to be willing to use telemedicine (Table 2).

XSL•FO
RenderX

Table 2. Correlates of respondents' willingness to use telehealth services (N=1045).

Characteristics	Willing to use telemedicine			Willing to use telepharmacy		
	Value	OR[a] (95% CI)	*P* value	Value	OR (95% CI)	*P* value
Sex, n/N (%)						
Male	333/523 (63.7)	Reference	__[b]	400/523 (76.5)	Reference	—
Female	341/522 (65.3)	1.12 (0.84-1.48)	.44	401/522 (76.8)	1.10 (0.80-1.5)	.57
Age (years), n/N (%)						
45-54	260/348 (74.7)	Reference	—	296/348 (85.1)	Reference	—
55-64	231/348 (66.4)	0.61 (0.42-0.86)	.01	275/348 (79)	0.57 (0.38-0.86)	.01
≥65	183/349 (52.4)	0.33 (0.22-0.49)	<.001	230/349 (65.9)	0.24 (0.15-0.37)	<.001
Race, n/N (%)						
White	484/770 (62.9)	Reference	—	580/770 (75.3)	Reference	—
Black or African American	98/134 (73.1)	1.51 (0.97-2.37)	.07	109/134 (81.3)	1.37 (0.82-2.29)	.23
Asian	44/68 (65)	0.87 (0.49-1.54)	.64	53/68 (78)	1.17 (0.61-2.25)	.63
Multiracial or other	48/73 (66)	1.04 (0.59- 1.82)	.88	59/73 (81)	1.43 (0.73-2.79)	.30
Ethnicity, n/N (%)						
Not Hispanic or Latinx	589/919 (64.1)	Reference	—	701/919 (76.3)	Reference	—
Hispanic or Latinx	85/126 (67.5)	1.02 (0.65-1.59)	.94	100/126 (79.4)	1.13 (0.67-1.90)	.64
Insurance type, n/N (%)						
Private	272/384 (70.8)	0.63 (0.32-1.27)	.20	306/384 (79.7)	0.81 (0.39-1.67)	.57
Public	338/569 (59.4)	0.61 (0.31-1.22)	.16	421/569 (74)	1.14 (0.56-2.33)	.71
Uninsured	43/60 (72)	Reference	—	47/60 (78)	Reference	—
Rurality, n/N (%)						
Not rural	488/731 (66.8)	Reference	—	570/731 (78)	Reference	—
Rural	186/314 (59.2)	0.80 (0.58-1.1)	.17	231/314 (73.6)	0.75 (0.52-1.08)	.12
Education, n/N (%)						
High school or less	129/217 (59.4)	Reference	—	161/217 (74.2)	Reference	—
Some college	255/408 (62.5)	1.02 (0.71-1.48)	.90	304/408 (74.5)	0.89 (0.59-1.35)	.60
College degree	180/259 (69.5)	1.37 (0.89-2.12)	.15	211/259 (81.5)	1.35 (0.83-2.22)	.23
Graduate education or higher	110/161 (68.3)	1.26 (0.76-2.08)	.37	125/161 (77.6)	1.16 (0.66-2.02)	.61
Household income (US $), n/N (%)						
<20,000	97/156 (62.2)	Reference	—	117/156 (75)	Reference	—
20,000-39,000	156/247 (63.2)	1.11 (0.71-1.73)	.65	181/247 (73.3)	1.08 (0.65-1.78)	.77
40,000-59,000	123/219 (56.2)	0.86 (0.53-1.38)	.53	167/219 (76.3)	1.52 (0.88-2.64)	.13
60,000-79,000	109/151 (72.2)	1.59 (0.93-2.72)	.09	120/151 (79.5)	1.42 (0.77-2.6)	.26
80,000-99,000	59/89 (66)	1.07 (0.57-1.99)	.83	72/89 (81)	1.63 (0.78-3.39)	.19
>100,000	130/183 (71)	1.34 (0.78-2.32)	.29	144/183 (78.7)	1.34 (0.72-2.48)	.35
Health status, n/N (%)						
Excellent	57/77 (74)	1.70 (0.68-4.26)	.26	63/77 (82)	0.85 (0.28-2.59)	.77
Very good	189/301 (62.8)	1.09 (0.50-2.37)	.83	232/301 (77.1)	0.79 (0.30-2.1)	.63
Good	287/443 (64.8)	1.28 (0.60-2.75)	.52	333/443 (75.2)	0.72 (0.28-1.88)	.50
Fair	121/190 (63.7)	1.18 (0.54-2.62)	.68	145/190 (76.3)	0.74 (0.27-2)	.55
Poor	20/34 (59)	Reference	—	28/34 (82)	Reference	—

XSL•FO
RenderX

Characteristics	Willing to use telemedicine			Willing to use telepharmacy		
	Value	OR[a] (95% CI)	*P* value	Value	OR (95% CI)	*P* value
Place to receive health care[c], n/N (%)						
A doctor's office or health care center	578/895 (64.6)	1.54 (0.66-3.58)	.32	—	—	—
Urgent care center or clinic in a drug store or grocery store	49/73 (67)	1.79 (0.71-5.53)	.22	—	—	—
Emergency room	18/32 (56)	Reference		—	—	—
A VA[d] medical center or VA outpatient clinic	26/39 (67)	1.81 (0.61-5.35)	.29	—	—	—
Some other place	3/6 (50)	1.38 (0.18-10.75)	.76	—	—	—
Have provider as a regular source of care[c], n/N (%)						
No	65/112 (58)	Reference		—	—	—
Yes	609/933 (65)	1.75 (1.03-2.98)	.04	—	—	—
Travel time to provider[c], mean (SD)	18.73 (14.19)	1.01 (1-1.02)	.05	—	—	—
Pharmacy type[e], n/N (%)						
A retail chain pharmacy (like CVS[f] or Walgreens)	—	—	—	395/515 (76.7)	Reference	—
A pharmacy in a grocery store (like Kroger or Albertsons or Safeway)	—	—	—	125/164 (76.2)	0.99 (0.64-1.53)	.95
A pharmacy in a department store or wholesaler (like Walmart or Costco)	—	—	—	124/166 (74.7)	1.03 (0.67-1.60)	.89
A pharmacy in a clinic or hospital where you receive medical care	—	—	—	75/90 (83)	1.87 (0.98-3.54)	.06
A local, independently owned pharmacy	—	—	—	82/110 (74.5)	0.98 (0.58-1.66)	.94
Travel time to nearest pharmacy[e], mean (SD)	—	—		18.79 (14.51)	1 (0.99-1.01)	.73
Pharmacy service quality rating[e], mean (SD)	—	—		17.51 (4.91)	1.06 (1.03-1.09)	<.001

[a]OR: odds ratio.
[b]Not available.
[c]Items specific to telemedicine.
[d]VA: Veteran's Affairs.
[e]Items specific to telepharmacy.
[f]A retail pharmacy chain in the United States.

Conditions That Make Respondents Willing to Use Telemedicine

A minority of respondents reported ambivalence toward or disagreed with the use of telemedicine (371/1045, 35.5%). About a third of these respondents indicated they would be willing to use telemedicine if appointments were low-cost or insurance-covered (135/371, 36.4%) or if the appointment was with their usual health care provider (125/371, 33.7%). About a quarter of these respondents were willing to use telemedicine if the platform was easy to use (98/371, 26.4%). To a lesser extent, respondents indicated the other conditions as important criteria to meet before they were willing to use telemedicine (Table 3). After accounting for the top 3 conditions for using telemedicine (low-cost or insurance-covered appointments, having the appointment with a health care provider, and a user-friendly platform), respondents' willingness to use telemedicine increased from 64.5% (674/1045) to 82.9% (867/1045).

XSL•FO
RenderX

Table 3. Conditional use of telemedicine services with a health care provider (n=371).

"I would be willing to use telemedicine to communicate with a health care provider about my health only if..."	Respondents, n (%)
My insurance would pay for it, or it would be low-cost for me	135 (37)
My appointment was with my health care provider	125 (34)
The system was easy to use	98 (26)
The system was private	91 (25)
I could get an appointment sooner than an in-person appointment	66 (18)
The appointments were available during extended hours, such as evening	48 (13)

Willingness to Use Patient Care Services Through Telepharmacy

Respondents were willing to use telepharmacy services for medication counseling (727/1045, 69.6%), vaccination consultation (511/1045, 48.9%), urgent care consultation (293/1045, 28%), chronic disease management (283/1045, 27.1%), travel medicine consultation (252/1045, 24.1%), genetic testing counseling (159/1045, 15.2%), and smoking cessation counseling (85/1045, 8.1%; Table 4).

Table 4. Proportion of respondents willing to use telepharmacy services (N=1045).

Patient care services	Respondents, n (%)
Medication counseling	727 (70)
Vaccination consultation	511 (49)
Urgent care consultation	252 (28)
Chronic disease management and education	283 (27)
Travel medicine consultation	252 (24)
Genetic testing counseling	159 (15)
Smoking or tobacco cessation counseling	85 (8)
Other	15 (1)

Correlates of Respondents' Willingness to Use Telepharmacy Services

In the multiple logistic regression model, compared to participants aged between 45 and 54 years, respondents aged between 55 and 64 years, (OR 0.57, 95% CI 0.38-0.86) and respondents aged 65 years or older (OR 0.24, 95% CI 0.15-0.37) were less likely to be willing to use telepharmacy services. Respondents who reported a higher pharmacy service quality rating were also more likely to be willing to use telepharmacy (OR 1.06, 95% CI 1.03-1.09; Table 2).

Discussion

Principal Results and Comparison With Previous Work

The capacity of telehealth (ie, telemedicine and telepharmacy) to expand access to health services, especially in remote areas and to marginalized populations, has been established [1,4,5]. This study adds to this literature by revealing that a substantial proportion of our national sample of older adults were willing to use telehealth (which refers to both telemedicine and telepharmacy in this study) services. Further, we found that age, having a regular provider, longer travel time to the provider, and perceived higher pharmacy service quality were factors associated with willingness to use telehealth services.

We found that older adults (aged 55 years and older) were less willing to access telehealth services than those aged between 45 and 54 years. Previous studies have reported similar results and noted that older patients could be more hesitant to use telehealth services in general due to their inexperience with modern technology and the perception that their health care needs cannot be effectively addressed in remote appointments [24-26]. Another study that assessed the readiness of adults to use telehealth reported that among people aged 65 years or older in the US population, over 30% were not ready to access medical services through telehealth, with 20% of this group being resistant due to physical disabilities such as difficulty seeing and hearing [24]. As such, to increase the use of telehealth, health care settings should develop educational materials or training, as well as more user-friendly telehealth models for diverse adult populations that account for disabilities like visual and hearing impairment [27-29].

We also found that participants who had a regular health care provider and reported higher appraisals of pharmacy service quality were more willing to use telemedicine and telepharmacy, respectively. Strong rapport in the form of familiarity, trustworthiness, and responsiveness are established characteristics of good patient-provider relationships [30,31]. While patients tend to have established care with primary care and other specialty providers, they may be less likely to have an "established" pharmacy and, by extension, usual pharmacists [32,33],from whom they receive care. Our findings suggest that

XSL•FO
RenderX

the use of telepharmacy services among older adults may increase if community pharmacists are able to establish better interpersonal connections with their patients. This could be achieved by assuming a patient-centered approach in the delivery of pharmacy services or constant patient engagement through regular check-ins to address patients' health-related questions.

Unsurprisingly, we found that increased travel time to a health care provider was associated with a higher willingness to use telemedicine. Barriers to attending in-person clinic appointments are well established, such as access to reliable transportation, taking time off work, or childcare services. Similar to our finding, a systematic review showed that patients were more satisfied with a telemedicine option because of lower financial costs and resource commitments [34]. Multiple studies have shown that with telemedicine, patients were able to save hours on the road, lower transportation expenses, shorten clinic wait times, and avoid missing work [14,35,36]. These external factors can be nontrivial, especially for older adults on fixed incomes or who have difficulty traveling and should be highlighted as benefits of telehealth services in both medical and pharmacy settings.

Past studies have demonstrated that low uptake of telehealth may be due to unfamiliarity with the telehealth technology or the cost of the telehealth service [24-26].This study revealed similar reasons why people may not want to participate in telemedicine, such as high costs or no insurance coverage, not having an appointment with their provider, and complicated telemedicine platforms. As clinics and pharmacies are adopting telehealth services, they should consider the characteristics of their patient populations and select simpler and more user-friendly platforms. From a policy standpoint, accounting for the rollback of compensation for telehealth services by the federal government, insurance companies and regulators should consider ways to create and maintain reimbursement models and incentives that could increase the use of telehealth services, as these services may help contain costs and be more patient-centered and convenient to some [37,38].

Strengths and Limitations

Strengths of this study include oversampling of medically underserved groups, as evidenced by our sample having a higher proportion of rural residents and those with low socioeconomic status. Although a convenience sample, our quotas matched the US census. Another strength is that we evaluated the correlates of willingness to use telemedicine and telepharmacy. Previous studies have either focused on telemedicine or telepharmacy only and may not have evaluated factors that lead to telemedicine service use. Additionally, our survey was administered during the pandemic, so familiarity with and use of telehealth likely increased during the time of this study, increasing the timeliness and relevance of our results.

A limitation to consider is that this study was cross-sectional, so we are unable to establish temporality and whether the uptake of telehealth services among this population increased over time. Also, this was an internet-based survey, and all participants had internet access, so we could not assess telehealth needs or preferences among those that may have poorer access to these technologies. Another limitation was that we sampled specific adults that met the criteria for our larger PharmFIT Patient Survey, which may not be wholly representative of the older adult population. Additional studies might need to be conducted to assess the overall generalizability of our findings to the older adult population. Also, we could not assess whether a participant's health plan covered telehealth services. However, due to the COVID-19 pandemic, telehealth services have become more widespread, more types of patient care services are provided remotely, and insurance plans more readily cover the costs of these services [39].

Conclusion

Our results revealed that respondents were generally willing to use telehealth (telemedicine and telepharmacy) services, but the likelihood of them being willing to use these services decreased as respondents were older. We also found that for those unwilling to use telemedicine services, if certain stipulations (eg, inexpensive services) were met, then their willingness to use these services increased. Telehealth services have increased during the COVID-19 pandemic, showing the relevance of this approach for health care providers to adopt across the country in closing the gap in access to health services [10]. There were uncertainties about whether telehealth services would continue to be reimbursed after the pandemic, but evidence suggests that federal and private insurers would continue to reimburse these services [40,41]. It is critical that telehealth services continue to be available, as they improve equitable access and have a tendency to reduce health-related costs [34]. Hence, to increase adoption, more considerations, such as providing low-cost or easy-to-use telehealth options, are needed to design telehealth platforms that can accommodate patients with certain characteristics (eg, older adults). Similarly, providers in clinics and pharmacies need to be aware of situations that increase the willingness to use telehealth services, such as establishing a good rapport with patients, and decrease the willingness, such as a lack of insurance coverage.

Acknowledgments

This study was supported by the Centers for Disease Control and Prevention of the US Department of Health and Human Services (cooperative agreement #U48 DP006400; multiple principal investigators Jennifer Leeman DrPH, MPH, ATB, and SBW) and a donation by the Safeway Foundation to Fred Hutch (principal investigator PDS). ARW is supported by the National Cancer Institute's National Research Service Award sponsored by the Lineberger Comprehensive Cancer Center at the University of North Carolina (T32 CA116339). The manuscript contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by Safeway, Fred Hutch, University of North Carolina, National Cancer Institute, Centers for Disease Control and Prevention of the US Department of Health and Human Services, or the US government.

XSL•FO
RenderX

Data Availability

The data used in this study are available on request from the corresponding author.

Authors' Contributions

All the authors were responsible for conceptualization, methodology, investigation, validation, project administration, and writing (review and editing). OOO was responsible for data curation and visualization. BYU, CH, OOO, and PDS were responsible for formal analyses. PDS was responsible for supervision. OOO and PDS were responsible for writing the original draft. ATB, PDS, and SBW were responsible for funding acquisition and contributed as cosenior authors.

Conflicts of Interest

SBW receives grant funding paid to her institution from Pfizer and AstraZeneca.

References

1. Kovačević M, Ćulafić M, Vezmar Kovačević S, Borjanić S, Keleč B, Miljković B, et al. Telepharmacy service experience during the COVID-19 pandemic in the Republic of Srpska, Bosnia and Herzegovina. Health Soc Care Community 2022 Sep;30(5):e1639-e1650 [FREE Full text] [doi: 10.1111/hsc.13590] [Medline: 34582591]
2. Le T, Toscani M, Colaizzi J. Telepharmacy: a new paradigm for our profession. J Pharm Pract 2020 Apr;33(2):176-182. [doi: 10.1177/0897190018791060] [Medline: 30060679]
3. Office for the advancement of telehealth. Health Resources & Services Administration. 2024. URL: https://www.hrsa.gov/telehealth [accessed 2024-01-10]
4. Pathak S, Haynes M, Qato DM, Urick BY. Telepharmacy and quality of medication use in rural areas, 2013-2019. Prev Chronic Dis 2020 Sep 03;17:E101 [FREE Full text] [doi: 10.5888/pcd17.200012] [Medline: 32886060]
5. Inch J, Notman F, Watson M, Green D, Baird R, Ferguson J, Telepharmacy Research Team. Tele-pharmacy in rural Scotland: a proof of concept study. Int J Pharm Pract 2017 Jun;25(3):210-219 [FREE Full text] [doi: 10.1111/ijpp.12376] [Medline: 28466547]
6. Baldoni S, Amenta F, Ricci G. Telepharmacy services: present status and future perspectives: a review. Medicina (Kaunas) 2019 Jul 01;55(7):327 [FREE Full text] [doi: 10.3390/medicina55070327] [Medline: 31266263]
7. Telehealth, telemedicine, and telecare: what's what? Federal Communications Commission. 2024. URL: https://www.fcc.gov/general/telehealth-telemedicine-and-telecare-whats-what [accessed 2024-01-10]
8. Perisetti A, Goyal H. Successful distancing: telemedicine in gastroenterology and hepatology during the COVID-19 pandemic. Dig Dis Sci 2021 Apr;66(4):945-953 [FREE Full text] [doi: 10.1007/s10620-021-06874-x] [Medline: 33655456]
9. Regulation of telepharmacy practice. Federal Register. 2021. URL: https://www.federalregister.gov/documents/2021/11/17/2021-24948/regulation-of-telepharmacy-practice [accessed 2024-01-10]
10. Unni EJ, Patel K, Beazer IR, Hung M. Telepharmacy during COVID-19: a scoping review. Pharmacy (Basel) 2021 Nov 11;9(4):183 [FREE Full text] [doi: 10.3390/pharmacy9040183] [Medline: 34842823]
11. Maxwell LG, McFarland MS, Baker JW, Cassidy RF. Evaluation of the impact of a pharmacist-Led telehealth clinic on diabetes-related goals of therapy in a veteran population. Pharmacotherapy 2016 Mar;36(3):348-356. [doi: 10.1002/phar.1719] [Medline: 26877253]
12. Varker T, Brand RM, Ward J, Terhaag S, Phelps A. Efficacy of synchronous telepsychology interventions for people with anxiety, depression, posttraumatic stress disorder, and adjustment disorder: a rapid evidence assessment. Psychol Serv 2019 Nov;16(4):621-635. [doi: 10.1037/ser0000239] [Medline: 29809025]
13. Ramsey A, Yang L, Vadamalai K, Mustafa SS. Appointment characteristics in an allergy/immunology practice in the immediate aftermath of COVID-19 restrictions. J Allergy Clin Immunol Pract 2020 Sep;8(8):2771-2773 [FREE Full text] [doi: 10.1016/j.jaip.2020.05.017] [Medline: 32470444]
14. Nguyen M, Waller M, Pandya A, Portnoy J. A review of patient and provider satisfaction with telemedicine. Curr Allergy Asthma Rep 2020 Sep 22;20(11):72 [FREE Full text] [doi: 10.1007/s11882-020-00969-7] [Medline: 32959158]
15. Cubanski J, Kates J, Tolbert J, Guth M, Pollitz K, Freed M. What happens when COVID-19 emergency declarations end? Implications for coverage, costs, and access. KFF. 2023. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/what-happens-when-covid-19-emergency-declarations-end-implications-for-coverage-costs-and-access [accessed 2024-04-10]
16. CARES act: AMA COVID-19 pandemic telehealth fact sheet. American Medical Association. 2023. URL: https://www.ama-assn.org/delivering-care/public-health/cares-act-ama-covid-19-pandemic-telehealth-fact-sheet [accessed 2024-04-10]
17. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 1995 Mar;36(1):1-10. [Medline: 7738325]
18. Brenner AT, Waters AR, Wangen M, Rohweder C, Odebunmi O, Marciniak MW, et al. Patient preferences for the design of a pharmacy-based colorectal cancer screening program. Cancer Causes Control 2023 Dec;34(Suppl 1):99-112 [FREE Full text] [doi: 10.1007/s10552-023-01687-x] [Medline: 37072526]
19. The leading experience management software. Qualtrics XM. 2023. URL: https://www.qualtrics.com/ [accessed 2024-04-10]

XSL•FO
RenderX

20. Standard definitions: final dispositions of case codes and outcome rates for surveys. 9th Edition. The American Association for Public Opinion Research. 2016. URL: https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf [accessed 2024-04-10]
21. Rural-Urban Continuum Codes. U.S. Department of Agriculture: Economic Research Service. 2023. URL: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx [accessed 2024-04-10]
22. PharmFIT - the 4CNC Core Project (2019-2024). Dataverse. 2023. URL: https://dataverse.unc.edu/dataverse/cpcrn-4cnc-pharmfit [accessed 2024-06-10]
23. Mackwood MB, Tosteson TD, Alford-Teaster JA, Curtis KM, Lowry ML, Snide JA, et al. Factors influencing telemedicine use at a northern new England cancer center during the COVID-19 pandemic. JCO Oncol Pract 2022 Jul;18(7):e1141-e1153 [FREE Full text] [doi: 10.1200/OP.21.00750] [Medline: 35446680]
24. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 01;180(10):1389-1391 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
25. Kong SS, Otalora Rojas LA, Ashour A, Robinson M, Hosterman T, Bhanusali N. Ability and willingness to utilize telemedicine among rheumatology patients-a cross-sectional survey. Clin Rheumatol 2021 Dec;40(12):5087-5093 [FREE Full text] [doi: 10.1007/s10067-021-05759-8] [Medline: 34219187]
26. Omboni S, Tenti M. Telepharmacy for the management of cardiovascular patients in the community. Trends Cardiovasc Med 2019 Feb;29(2):109-117. [doi: 10.1016/j.tcm.2018.07.002] [Medline: 30037524]
27. Improving access to telehealth. Telehealth.HHS.gov. 2023. URL: https://telehealth.hhs.gov/providers/health-equity-in-telehealth/improving-access-to-telehealth [accessed 2024-04-10]
28. Telehealth. ADA.gov: U.S. Department of Justice Civil Rights Division. 2024. URL: https://www.ada.gov/topics/telehealth/ [accessed 2024-04-10]
29. Valdez RS, Rogers CC, Claypool H, Trieshmann L, Frye O, Wellbeloved-Stone C, et al. Ensuring full participation of people with disabilities in an era of telehealth. J Am Med Inform Assoc 2021 Feb 15;28(2):389-392 [FREE Full text] [doi: 10.1093/jamia/ocaa297] [Medline: 33325524]
30. Worley MM. Testing a pharmacist-patient relationship quality model among older persons with diabetes. Res Social Adm Pharm 2006 Mar;2(1):1-21. [doi: 10.1016/j.sapharm.2005.12.006] [Medline: 17138498]
31. Esmalipour R, Salary P, Shojaei A. Trust-building in the pharmacist-patient relationship: a qualitative study. Iran J Pharm Res 2021;20(3):20-30 [FREE Full text] [doi: 10.22037/ijpr.2020.114113.14675] [Medline: 34903966]
32. Mc Namara KP, Dunbar JA, Philpot B, Marriott JL, Reddy P, Janus ED. Potential of pharmacists to help reduce the burden of poorly managed cardiovascular risk. Aust J Rural Health 2012 Apr;20(2):67-73. [doi: 10.1111/j.1440-1584.2012.01259.x] [Medline: 22435766]
33. Smith KB, Humphreys JS, Lenard Y, Jones JA, Prince V, Han GS. Still the doctor - by a country mile! Preferences for health services in two country towns in north-west New South Wales. Med J Aust 2004;181(2):91-95 [FREE Full text] [doi: 10.5694/j.1326-5377.2004.tb06182.x] [Medline: 15257646]
34. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: A systematic review and narrative analysis. BMJ Open 2017 Aug 03;7(8):e016242 [FREE Full text] [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
35. Mustafa SS, Yang L, Mortezavi M, Vadamalai K, Ramsey A. Patient satisfaction with telemedicine encounters in an allergy and immunology practice during the coronavirus disease 2019 pandemic. Ann Allergy Asthma Immunol 2020 Oct;125(4):478-479 [FREE Full text] [doi: 10.1016/j.anai.2020.06.027] [Medline: 32585178]
36. Taylor L, Capling H, Portnoy JM. Administering a telemedicine program. Curr Allergy Asthma Rep 2018 Sep 15;18(11):57. [doi: 10.1007/s11882-018-0812-8] [Medline: 30220060]
37. Niu B, Mukhtarova N, Alagoz O, Hoppe K. Cost-effectiveness of telehealth with remote patient monitoring for postpartum hypertension. J Matern Fetal Neonatal Med 2022 Dec;35(25):7555-7561. [doi: 10.1080/14767058.2021.1956456] [Medline: 34470135]
38. Agha Z, Schapira RM, Maker AH. Cost effectiveness of telemedicine for the delivery of outpatient pulmonary care to a rural population. Telemed J E Health 2002;8(3):281-291. [doi: 10.1089/15305620260353171] [Medline: 12419022]
39. Demeke HB, Pao LZ, Clark H, Romero L, Neri A, Shah R, et al. Telehealth practice among health centers during the COVID-19 pandemic - United States, July 11-17, 2020. MMWR Morb Mortal Wkly Rep 2020 Dec 18;69(50):1902-1905 [FREE Full text] [doi: 10.15585/mmwr.mm6950a4] [Medline: 33332297]
40. Salmanizadeh F, Ameri A, Bahaadinbeigy K. Methods of reimbursement for telemedicine services: a scoping review. Med J Islam Repub Iran 2022;36:68 [FREE Full text] [doi: 10.47176/mjiri.36.68] [Medline: 36128316]
41. MLN90175 - Telehealth Services. Centers for Medicare and Medicaid Services. Medicare Learning Network. 2024. URL: https://www.cms.gov/files/document/mln901705-telehealth-services.pdf [accessed 2024-01-10]

Abbreviations

CARES Act: Coronavirus Aid Relief, Economic Security Act

XSL•FO
RenderX

OR: odds ratio

Edited by T de Azevedo Cardoso; submitted 23.06.23; peer-reviewed by G Ricci, B Wolfe-Piggott, D Dinh, K Drude; comments to author 12.01.24; revised version received 16.02.24; accepted 20.03.24; published 23.05.24.

Please cite as:
Odebunmi OO, Hughes TD, Waters AR, Urick BY, Herron C, Wangen M, Rohweder C, Ferrari RM, Marciniak MW, Wheeler SB, Brenner AT, Shah PD
Findings From a National Survey of Older US Adults on Patient Willingness to Use Telehealth Services: Cross-Sectional Survey
J Med Internet Res 2024;26:e50205
URL: https://www.jmir.org/2024/1/e50205
doi:10.2196/50205
PMID:38780994

©Olufeyisayo O Odebunmi, Tamera D Hughes, Austin R Waters, Benjamin Y Urick, Caroline Herron, Mary Wangen, Catherine Rohweder, Renée M Ferrari, Macary W Marciniak, Stephanie B Wheeler, Alison T Brenner, Parth D Shah. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Structured Multilevel Telehealth Service on Hospital Admissions and Mortality During COVID-19 in a Resource-Limited Region in Brazil: Retrospective Cohort Study

Clara Rodrigues Alves Oliveira[1,2], MSc, MD, PhD; Magda Carvalho Pires[3], MD, PhD; Karina Cardoso Meira[4], MD, PhD; Jordana Cristina de Jesus[4], MD, PhD; Isabela Nascimento Borges[1,2], MSc, MD, PhD; Maria Cristina Paixão[2], BSc; Mayara Santos Mendes[2], MSc, MD; Leonardo Bonisson Ribeiro[2], BSc; Milena Soriano Marcolino[1,2,5], MSc, MD, PhD; Maria Beatriz Moreira Alkmim[2], MSc, MD; Antonio Luiz Pinho Ribeiro[1,2,5], MSc, MD, PhD

[1]Department of Internal Medicine, Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2]Telehealth Center, University Hospital and Telehealth Network of Minas Gerais, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[3]Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[4]Health School, Universidade Federal do Rio Grande do Norte, Natal, Brazil

[5]Institute for Health Technology Assessment, Porto Alegre, Brazil

Corresponding Author:
Clara Rodrigues Alves Oliveira, MSc, MD, PhD
Department of Internal Medicine
Medical School
Universidade Federal de Minas Gerais
110 Avenida Professor Alfredo Balena, Room 107
Ala Sul Santa Efigênia
Belo Horizonte, 30130-100
Brazil
Phone: 55 3134099669
Email: claralves@gmail.com

Abstract

Background: The COVID-19 pandemic represented a great stimulus for the adoption of telehealth and many initiatives in this field have emerged worldwide. However, despite this massive growth, data addressing the effectiveness of telehealth with respect to clinical outcomes remain scarce.

Objective: The aim of this study was to evaluate the impact of the adoption of a structured multilevel telehealth service on hospital admissions during the acute illness course and the mortality of adult patients with flu syndrome in the context of the COVID-19 pandemic.

Methods: A retrospective cohort study was performed in two Brazilian cities where a public COVID-19 telehealth service (TeleCOVID-MG) was deployed. TeleCOVID-MG was a structured multilevel telehealth service, including (1) first response and risk stratification through a chatbot software or phone call center, (2) teleconsultations with nurses and medical doctors, and (3) a telemonitoring system. For this analysis, we included data of adult patients registered in the Flu Syndrome notification databases who were diagnosed with flu syndrome between June 1, 2020, and May 31, 2021. The exposed group comprised patients with flu syndrome who used TeleCOVID-MG at least once during the illness course and the control group comprised patients who did not use this telehealth service during the respiratory illness course. Sociodemographic characteristics, comorbidities, and clinical outcomes data were extracted from the Brazilian official databases for flu syndrome, Severe Acute Respiratory Syndrome (due to any respiratory virus), and mortality. Models for the clinical outcomes were estimated by logistic regression.

Results: The final study population comprised 82,182 adult patients with a valid registry in the Flu Syndrome notification system. When compared to patients who did not use the service (n=67,689, 82.4%), patients supported by TeleCOVID-MG (n=14,493, 17.6%) had a lower chance of hospitalization during the acute respiratory illness course, even after adjusting for sociodemographic characteristics and underlying medical conditions (odds ratio [OR] 0.82, 95% CI 0.71-0.94; *P*=.005). No difference in mortality was observed between groups (OR 0.99, 95% CI 0.86-1.12; *P*=.83).

XSL•FO
RenderX

Conclusions: A telehealth service applied on a large scale in a limited-resource region to tackle COVID-19 was related to reduced hospitalizations without increasing the mortality rate. Quality health care using inexpensive and readily available telehealth and digital health tools may be delivered in areas with limited resources and should be considered as a potential and valuable health care strategy. The success of a telehealth initiative relies on a partnership between the involved stakeholders to define the roles and responsibilities; set an alignment between the different modalities and levels of health care; and address the usual drawbacks related to the implementation process, such as infrastructure and accessibility issues.

(J Med Internet Res 2024;26:e48464) doi:10.2196/48464

KEYWORDS

COVID-19; telehealth; health care; clinical outcomes; hospital admission; mortality; adoption; effectiveness; digital health tool; flu; teleconsultation; digital health; digital literacy; telemonitoring

Introduction

The COVID-19 pandemic has accelerated the adoption of telehealth in outpatient settings [1]. The demands generated by the pandemic, such as the need for social distancing and home isolation in an effort to reduce virus spread and infection rates, to decrease the pressure on overwhelmed health systems, and to increase health care providers' safety, determined the scaled-up use of telehealth tools worldwide [2].

Taking these new challenges into account, the Telehealth Network of Minas Gerais (TNMG), one of the largest public telehealth services in Latin America, developed a synchronous teleconsultation and telemonitoring service to assist patients with suspected or confirmed COVID-19 [3,4]. This service, named TeleCOVID-MG, was first deployed in two Brazilian medium-sized cities and was then expanded to assist the entire community of Universidade Federal de Minas Gerais (UFMG), a large public Brazilian university [3].

Although many telehealth services were developed and implemented during the pandemic, data addressing the clinical effectiveness of the use of a telehealth service for the health assistance of patients with respiratory symptoms are scarce [5,6]. Most of the available studies have focused on different outcomes such as user satisfaction, system usability, and emotional comfort [3,5]. However, there is a lack of studies analyzing the impact of telehealth services on more objective outcomes such as hospitalization and mortality rates compared to standard health care support, which is usually delivered in an in-person manner. Moreover, the studies that do exist in this regard often include small sample sizes and report conflicting results [6].

Casariego-Vales et al [7] conducted a study to assess the effectiveness of a proactive telemonitoring approach for patients with COVID-19. The study was conducted in the northwest region of Spain during the third wave of the pandemic. The results demonstrated that the patients who were enrolled in a systematic telemonitoring program, which included different intensities of approaches tailored to the patient's risk for adverse outcomes, presented lower rates of emergency department visits and hospitalization as compared to those of the patients who did not participate in this systematic telemonitoring program. Additionally, these systematically telemonitored patients presented shorter hospital stays and a lower mortality rate during their first hospitalization [7].

Another study examined the impact of a telehealth intervention in Michigan, United States, which was a nurse-led, telephone-based active management protocol for individuals with COVID-19 who were in home isolation. Although the intervention group showed a lower rate of hospitalization within 30 days compared to that of the control group, this result was not statistically significant [8].

To obtain additional information on the effects of telehealth adoption with respect to clinical outcomes, the main goal of this study was to evaluate the impact of the adoption of a structured multilevel telehealth service, including teleconsultation and telemonitoring services and digital health tools, on relevant clinical outcomes such as hospital admission and mortality. Toward this end, we compared the outcomes presented by the patients with flu syndrome who did and did not use the telehealth service during the respiratory illness course in a limited-resource region in Brazil where TeleCOVID-MG was deployed.

Methods

The TeleCOVID-MG Service

TeleCOVID-MG was developed by the TNMG, which represents a telehealth network created in 2005 through a partnership among seven public universities of Minas Gerais State in the southeast region of Brazil. The TNMG is a large telehealth service in Brazil and develops different activities in the areas of clinical support, research, and tele-education [9].

Soon after the first case of COVID-19 was confirmed in Brazil, the multidisciplinary team of the TNMG, including physicians with expertise in telemedicine, infectious diseases specialists, nurses, managers, and information technology specialists, started to develop the TeleCOVID-MG system, according to the issued World Health Organization and Brazilian Ministry of Health Guidelines for COVID-19. Based on these guidelines and other available scientific evidence about COVID-19, instructional material was developed for training the TeleCOVID-MG team. As new scientific evidence emerged, the system and the instructional material were updated. TeleCOVID-MG was designed as a structured multilevel telehealth service, including teleconsultation and telemonitoring services and digital health tools (ie, a chatbot) [10]. TeleCOVID-MG software runs on a web environment, which allows the full recording of all activities taking place on the platform [3,4]. The

XSL•FO
RenderX

TeleCOVID-MG structured database and the data collected in the software were used for this study.

The service was first implemented in two Brazilian cities, Divinópolis and Teófilo Otoni, in May 2020. The service was then expanded in November 2020 to assist the faculty, public servants, and students of the UFMG. The service provided health support for the populations of Divinópolis and Teófilo Otoni until December 2021 and for the UFMG community until March 2023. TeleCOVID-MG was offered to the entire populations of these cities and the users were not charged for using it.

TeleCOVID-MG was built to work through the intersection of four different levels, as described elsewhere [3]. In brief, level 1 represents the user's gateway to the service, which is achieved via a chatbot developed by the TNMG team or through a special telephone call center [10]; level 2 represents the nursing staff; level 3 represents the medical staff; and level 4 represents the telemonitoring service, conducted by medical students under supervision. Every health professional and medical student in the program was extensively trained to use the TeleCOVID-MG system and to provide health assistance to the patients, as previously described [3,4].

At level 1, the screening and risk stratification of the patient with respiratory complaints was conducted by the chatbot software or by professionals of the call center who were trained for the task, which was performed through a list of questions that were established by drawing upon the best available evidence [10]. According to the severity of the symptoms and the comorbidities, the patient was sent to teleconsultation with the nursing staff, at level 2, or with the medical staff, at level 3. At the end of this teleconsultation, the health professional could advise the patient to adhere to home isolation or to seek onsite evaluation at a primary care center or an emergency unit. In addition to carrying out teleconsultations, the TeleCOVID-MG system allowed the health professionals to issue prescriptions, reports, and orders for diagnostic COVID-19 tests. All of these documents could be easily downloaded by the users. The software also enabled the generation of the compulsory report of suspected or confirmed COVID-19 cases, in compliance with requirements of the Brazilian Health Ministry [3,4].

All patients who were assessed by the nursing or medical teams at levels 2 and 3 were included in the telemonitoring program, which provided support for at least 10 days after the onset of respiratory symptoms. This telemonitoring service was delivered through a phone call; the patients with no alarm signs or decompensated comorbidities were monitored every 48 hours, while the patients with alarm signs and/or decompensated comorbidities were monitored every 24 hours. Undergraduate medical students from the local universities working under the supervision of a physician or a nurse composed the telemonitoring team [3,4].

There was also a nurse or a physician available to supervise and offer support to the level 2 and level 3 professionals, respectively. All of the supervisors were trained for the task and were available for clinical discussions and doubts clarification throughout the teleconsultation and telemonitoring duties. It is important to emphasize that, according to the clinical course of illness presented by the patient, a professional of any level, at any time, might request an evaluation by the staff of another level.

Finally, throughout the period that TeleCOVID-MG was in operation, a periodic and open dialogue was maintained with the local health managers to define the roles and responsibilities of the stakeholders and to set an alignment between the different modalities and levels of health care.

Study Design and Procedure

A retrospective cohort study was performed with adult patients (≥18 years old) who were registered in the Divinópolis and Teófilo Otoni Flu Syndrome compulsory notification databases between June 1, 2020, and May 31, 2021. The Divinópolis and Teófilo Otoni Flu Syndrome databases include patients who live and/or have received health support during the illness course in one of these two cities.

The exposed group (TeleCOVID-MG group) was composed of the registered patients with flu syndrome who used TeleCOVID-MG at least once during their illness course and the unexposed group (control group) was composed of patients who did not use this telehealth service during the respiratory illness course. The main outcomes investigated were hospital admissions during the acute respiratory illness course and mortality.

Divinópolis and Teófilo Otoni are both medium-sized cities. However, Divinópolis has a higher demographic density and Human Development Index, and better socioeconomic, educational, and health indicators compared to those of Teófilo Otoni (Table 1) [11,12].

XSL•FO
RenderX

Table 1. Sociodemographic, economic, education, and health indicators of Divinópolis and Teófilo Otoni compared to Brazil's overall indicators [11,12].

Indicators	Divinópolis	Teófilo Otoni	Brazil
Sociodemographic indicators			
Population in 2022, n	231,091	137,418	203,062,512
Demographic density (inhabitants/km^2) in 2022	326.35	42.38	23.86
Human Development Index in 2010	0.764	0.701	0.760
Economic indicators			
Gross Domestic Product (Brazilian Real[a]) in 2020	29,331.04	19,873.45	35,935.74
Employed population rate in 2020, %	27.4	22.1	28.4
Average monthly salary of formal workers in minimum-wage jobs (Brazilian Real[a]) for 2021	2.1	1.8	2.3
Education indicators			
Schooling rate from 6 to 14 years old in 2010, %	98.6	96.6	97.9
Basic Education Development Index for the early years of elementary school in 2021	6.5	5.5	5.5
Health indicators			
Infant mortality rate (per 1000 live births) for 2020	9.13	13.66	11.20
Number of physicians per 1000 inhabitants for 2022	3.16	1.97	2.38
Households with sanitary sewage for 2010, %	90.1	77.1	63.2

[a]US $1=4.93 Real for June 2020.

Data Acquisition

The data used in the analysis were extracted from four different databases: (1) the TeleCOVID-MG database, (2) Flu Syndrome compulsory notification database, (3) Severe Acute Respiratory Syndrome (SARS) compulsory notification database, and (4) Mortality Information System (MIS) database. Although the latter three are public databases, access to the identified data of the Flu Syndrome and SARS databases was obtained upon request to the municipal governments. In contrast, access to the identified data of the MIS database was solicited from the federal government.

The Flu Syndrome database includes data for the compulsory notification of patients with flu syndrome, the SARS database is composed of the compulsory notification of patients who were admitted to the hospital with SARS, and the MIS database is composed of the compulsory registration of the deaths of Brazilian citizens. Flu syndrome is defined by the Brazilian Health Ministry as the occurrence of an acute respiratory illness characterized by the presence of at least two of the following symptoms: fever, chills, sore throat, headache, cough, runny nose, and smell or taste disturbances. SARS is defined by the Brazilian Health Ministry as the occurrence of flu syndrome and at least one of the following symptoms: oxygen saturation below 95% (room air), dyspnea, persistent pressure or pain in the chest, and cyanosis of the lips or face [13].

It is important to note that the notification of flu syndrome and SARS can be made by any health professional at any level of health care. The patient's data necessary for completing all required notification documents are standardized, but not all the fields are mandatory.

Data regarding hospital admission due to acute respiratory illness were obtained from the SARS database. To obtain more accurate mortality data, this register was extracted from all three public databases. Identifying the patient's death in at least one of these databases was sufficient to characterize the final clinical outcome as "death." For the mortality analysis, all the deaths that occurred throughout the study period were considered, regardless of their primary causes.

Potential cofounders used as adjustment variables included city of flu syndrome notification; age; sex; and underlying medical conditions such as chronic respiratory diseases, chronic cardiovascular diseases, diabetes, chronic kidney disease, immunosuppression, high-risk pregnancy, and chromosomal diseases. These data were extracted from the Flu Syndrome database. Finally, the linkage between the four databases was made using the patient's name and their individual taxpayer register number.

This manuscript was written according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for reporting observational studies [14].

Statistical Analysis

Statistical analyses were performed in three steps: (1) descriptive analysis, (2) bivariate analysis (ie, evaluation of the association of the outcome with each variable of interest), and (3) multivariate analysis.

XSL•FO
RenderX

Descriptive analyses were run to summarize all variables, stratified into exposed and unexposed groups. Categorical variables (sex, city of flu syndrome notification, city of residence, underlying medical conditions, and clinical outcomes) are summarized by absolute and relative frequencies. The Shapiro-Wilk normality test was performed to determine whether the continuous variable (age) was normally distributed. This variable was found to have a nonnormal distribution and is therefore summarized using the median and IQR. There were no missing variables included in this analysis.

In the bivariate analysis, sociodemographic characteristics, city of flu syndrome notification, underlying medical conditions, and clinical outcomes were assessed using the χ^2 test to compare proportions. The Wilcoxon rank sum test was used to compare medians of continuous variables.

Models for the primary outcomes were estimated by logistic regression; the shaping process of the prediction models divided variables into four blocks by adopting a forward approach, mutually inserted in regression models 1-4. As the primary goal of the analysis was to identify the association of the TeleCOVID-MG service with the clinical outcomes (mortality and hospital admissions during the acute respiratory illness course), this variable was tested in all four models. Model 1 included use of the TeleCOVID-MG service only, model 2 added the city of the flu syndrome notification, model 3 added sex and age, and model 4 added the underlying medical conditions. For the regression models, odds ratios (ORs) and their respective 95% CIs were estimated.

Concerning the choice of variables, we considered the recommendations of Steyerberg [15] and Harrell [16]. According to these authors, it is better to use subject matter knowledge than statistical methods for variable selection. In this strategy, significance testing of adjustment variables is not necessary, especially if subject-specific knowledge supports the estimated effects [14,15]. Thus, we grouped variables parsimoniously in blocks, aggregating variables more distal to those that are more proximal: geography (cities), biological features (age and sex), and clinical predisposing features (underlying medical conditions).

All analyses were performed in R software (version 4.0.2) with the tidyverse, lubridate, stringi, rlang, jsonlite, Rcurl, writexl, openxlsx, readxl, and lmtest packages. A P value <.05 was considered statistically significant.

Ethical Considerations

Ethical approval for this study was obtained from the UFMG Research Ethics Committee (CAAE: 35953620.9.0000.5149). Informed consent was provided by all the patients who were supported by the TeleCOVID-MG service. The researchers signed a confidentiality term to access the identified data of the public databases. After the linkage of the databases, data were deidentified for the analysis. No compensation was provided to the research participants.

Results

Throughout the study period, 116,488 registers were found in the Flu Syndrome notification systems of Teófilo Otoni and Divinópolis. Among these, 4272 registers were canceled in the database, 16,780 patients had more than one notification in the register, and 13,254 patients were under 18 years old. Therefore, the final study population was composed of 82,182 patients (Figure 1). For the analysis, only the first register of the patient in the Flu Syndrome database was used and none of the included variables contained missing data.

XSL•FO
RenderX

Figure 1. Flowchart of patient inclusion.

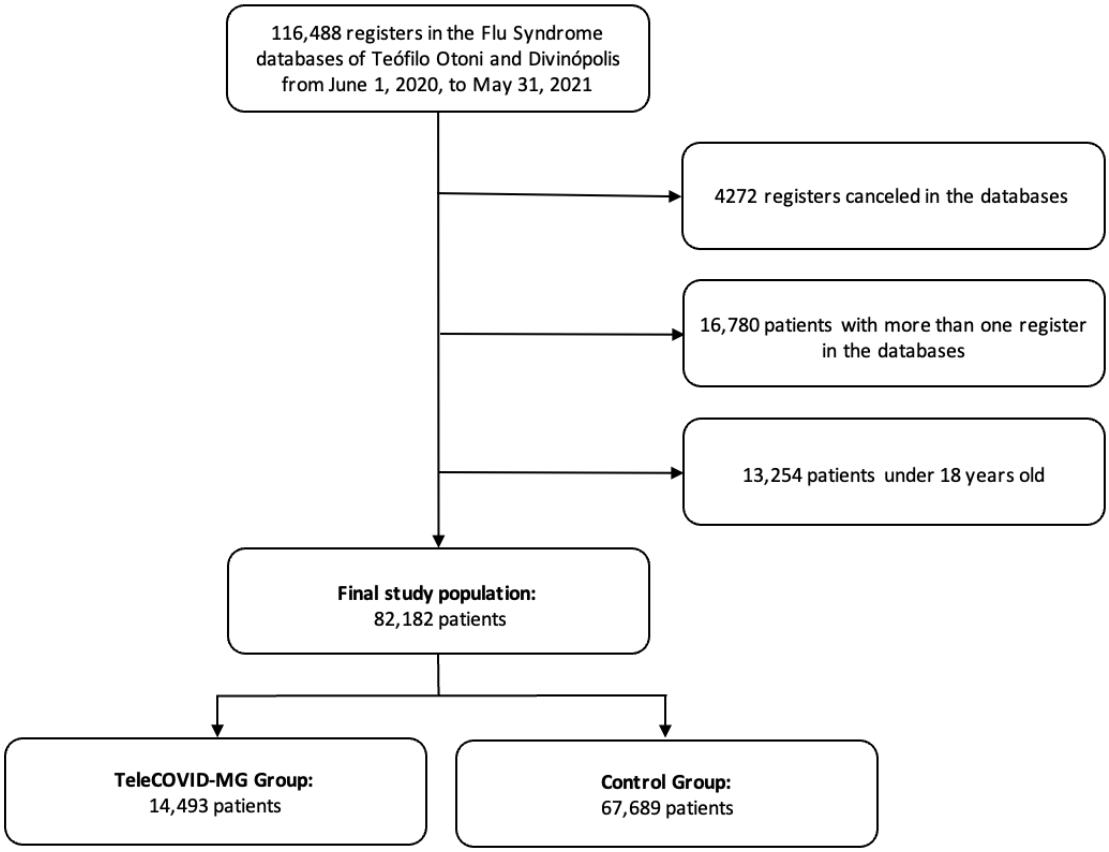

Among the 82,182 included patients, 14,493 (17.6%) used the TeleCOVID-MG service at least one time during the acute illness course (TeleCOVID-MG group) and the other 67,689 (82.4%) patients did not receive any kind of telehealth support by the TeleCOVID-MG service (control group). The sociodemographic and clinical characteristics of the patients are summarized in Table 2.

The TeleCOVID-MG group had a median age 2 years lower than that of the control group; a higher frequency of women; and slightly but significantly higher frequencies of patients with diabetes, chronic respiratory and cardiac diseases, and high-risk pregnancy (Table 2).

With regard to outcomes, the TeleCOVID-MG group had lower frequencies of both hospital admission ($P<.001$) and mortality ($P=.003$) compared to those of the control group (Table 2).

Among the 14,493 included patients who were accessed through the TeleCOVID-MG system during the illness course, 290 (2.0%) were advised to seek an onsite evaluation at a primary care center and 594 (4.1%) at an emergency department. The other 13,609 (93.9%) patients were advised to maintain domiciliary isolation and were monitored by the TeleCOVID-MG team.

In the regression analysis for the outcome hospital admission during the acute illness course, the telehealth intervention remained associated with a lower chance of hospitalization after adjusting for sociodemographic variables and underlying medical conditions (Table 3). In the fully adjusted model (model 4), there was a lower chance of hospitalization in patients who used the TeleCOVID-MG service, and a greater chance of this outcome was detected among the patients for whom the flu syndrome notification was reported in Teófilo Otoni, male patients, and patients with the following diseases: chronic cardiovascular disease, diabetes, chronic kidney disease, immunosuppression, and high-risk pregnancy (Table 3).

In the regression analysis for mortality (Table 4), the telehealth support was associated with a reduction in the chance of death in the model with the telehealth intervention alone (model 1) and in model 2 that included the city of flu syndrome notification. However, in models 3 and 4, which included sex, age, and underlying medical conditions, the OR values were close to 1.0. After the inclusion of all variables (model 4), age; Teófilo Otoni as the city of flu syndrome notification; male sex; and having chronic cardiovascular diseases, diabetes, immunosuppression, and chromosomal diseases remained associated with a greater chance of death due to flu syndrome (Table 4).

XSL•FO
RenderX

Table 2. Sociodemographic and clinical characteristics of the patients.

Characteristics	Overall (N=82,182)	TeleCOVID-MG group (n=14,493)	Control group (n=67,689)	*P* value[a]
Age (years), median (IQR)	38 (28-51)	37 (27-49)	39 (29-51)	<.001
Women, n (%)	44,847 (54.6)	8854 (61.1)	35,993 (53.2)	<.001
City of flu syndrome notification, n (%)				<.001
Divinópolis	46,313 (56.4)	3662 (25.3)	42,651 (63.0)	
Teófilo Otoni	25,276 (30.8)	6269 (43.3)	19,007 (28.1)	
Other	10,593 (12.9)	4562 (31.5)	6031 (8.9)	
City of residence, n (%)				<.001
Divinópolis	47,507 (57.8)	6091 (42.0)	41,416 (61.2)	
Teofilo Otoni	27,232 (33.1)	8205 (56.6)	19,027 (28.1)	
Other	7443 (9.1)	197 (1.4)	7246 (10.7)	
Underlying medical conditions, n (%)				
Chronic respiratory diseases	1272 (1.5)	253 (1.7)	1019 (1.5)	.03
Chronic cardiovascular diseases	2794 (3.4)	581 (4.0)	2213 (3.3)	<.001
Diabetes	2107 (2.6)	448 (3.1)	1659 (2.5)	<.001
Chronic kidney disease	203 (0.2)	20 (0.1)	183 (0.3)	.004
Immunosuppression	431 (0.5)	82 (0.6)	349 (0.5)	.45
High-risk pregnancy	407 (0.5)	103 (0.7)	304 (0.4)	<.001
Chromosomal diseases	139 (0.2)	28 (0.2)	111 (0.2)	.44
Clinical outcomes, n (%)				
Hospital admission	1980 (2.4)	269 (1.9)	1711 (2.5)	<.001
Death	1978 (2.4)	299 (2.1)	1679 (2.5)	.003

[a]*P* values are based on the Pearson χ^2 test for categorical variables or the Wilcoxon rank sum test for the continuous variable (age).

XSL•FO
RenderX

Table 3. Predictors of hospital admission according to logistic regression models.

Model	OR[a] (95% CI)	*P* value
Model 1		
TeleCOVID-MG group	0.73 (0.64-0.83)	<.001
Model 2		
TeleCOVID-MG group	0.64 (0.56-0.73)	<.001
City of flu syndrome notification (Teófilo Otoni)	1.58 (1.44-1.73)	<.001
Model 3		
TeleCOVID-MG group	0.85 (0.74-0.97)	.02
City of flu syndrome notification (Teófilo Otoni)	1.42 (1.29-1.56)	<.001
Male sex	1.75 (1.59-1.92)	<.001
Age	1.07 (1.07-1.08)	<.001
Model 4		
TeleCOVID-MG group	0.82 (0.71-0.94)	.005
City of flu syndrome notification (Teófilo Otoni)	1.50 (1.36-1.65)	<.001
Male sex	1.77 (1.61-1.95)	<.001
Age	1.07 (1.07-1.07)	<.001
Chronic respiratory diseases	1.14 (0.85-1.52)	.37
Chronic cardiovascular diseases	1.35 (1.15-1.58)	<.001
Diabetes	1.83 (1.54-2.17)	<.001
Chronic kidney disease	2.18 (1.34-3.43)	.001
Immunosuppression	1.87 (1.25-2.73)	.002
High-risk pregnancy	3.32 (1.29-6.96)	.005
Chromosomal diseases	1.33 (0.63-2.54)	.42

[a]OR: odds ratio.

XSL•FO
RenderX

Table 4. Predictors of mortality according to logistic regression models.

Model	OR[a] (95% CI)	*P* value
Model 1		
TeleCOVID-MG group	0.83 (0.73-0.94)	.003
Model 2		
TeleCOVID-MG group	0.79 (0.69-0.89)	<.001
City of flu syndrome notification (Teófilo Otoni)	1.2 (1.09-1.32)	<.001
Model 3		
TeleCOVID-MG group	1.01 (0.89-1.15)	.85
City of flu syndrome notification (Teófilo Otoni)	1.07 (0.97-1.18)	.16
Male sex	1.5 (1.37-1.65)	<.001
Age	1.07 (1.06-1.07)	<.001
Model 4		
TeleCOVID-MG group	0.99 (0.86-1.12)	.83
City of flu syndrome notification (Teófilo Otoni)	1.12 (1.02-1.24)	.02
Male sex	1.52 (1.38-1.66)	<.001
Age	1.06 (1.06-1.07)	<.001
Chronic respiratory diseases	1.28 (0.96-1.67)	.08
Chronic cardiovascular diseases	1.34 (1.14-1.57)	<.001
Diabetes	1.51 (1.26-1.8)	<.001
Chronic kidney disease	2.28 (1.41-3.54)	<.001
Immunosuppression	2.37 (1.63-3.34)	<.001
High-risk pregnancy	2.02 (0.71-4.45)	.13
Chromosomal diseases	2.37 (1.28-4.1)	.004

[a]OR: odds ratio.

Discussion

Principal Results

In an effort to avoid virus spread, preserve home isolation, and protect health care professionals from potential SARS-CoV-2 contamination, multiple telehealth initiatives emerged worldwide [1,17,18]. Along with the large adoption of telehealth tools in the pandemic context, concerns about usability, safety, costs, and efficacy related to this innovative practice have emerged. In this retrospective cohort study, we investigated the impact of the adoption of a structured multilevel public COVID-19 telehealth service on hospital admissions and mortality compared to the outcomes presented by patients who did not use the studied telehealth service. The results demonstrated a lower chance of hospitalization during the acute illness course in the TeleCOVID-MG group, even after adjusting for sociodemographic characteristics and underlying medical conditions, whereas there was no difference in mortality between the two groups.

The COVID-19 pandemic brought about an extra challenge for providing quality health assistance, which was particularly important for countries such as Brazil characterized by a large land area and significant health inequities [9,19]. In such regions, telehealth solutions offer even more advantages; thus, establishing that remote health services have the potential to provide high-quality health assistance will provide valuable information for health managers to better plan health interventions.

In addition to the clinical outcomes attained using the structured multilevel public COVID-19 telehealth service, it is reasonable to consider that this telehealth initiative contributed to achieving other important outcomes that were not measured in this study. In providing remote health support to the patients who did not need to seek an on-site evaluation in most situations, the telehealth service likely contributed to a decrease in contamination rates and thus an increase in health professionals' safety, which are also important outcomes that should be considered in future research in this field.

There are various issues regarding the deployment of a telehealth service that must be carefully examined, including costs, professional and patient acceptance and satisfaction, technology availability, and data safety. Regarding costs, there is a large body of evidence demonstrating that telehealth initiatives can be a cost-saving option in different scenarios. Specifically in the context of COVID-19, our group has already demonstrated that the TeleCOVID-MG service increased the access to health care and was an economically attractive strategy [20-24].

XSL•FO
RenderX

Regarding the acceptance and availability of the technology, it is important to remark that teleconsultation was not regulated in Brazil before March 2020 and TeleCOVID-MG was a pioneer in providing this type of service to the population [25]. Approximately 17.6% of the adult population with respiratory symptoms in the two studied cities used the service during the analyzed period; in contrast to our initial concerns about patient acceptance, we consider this use rate to represent a promising kickoff of this initiative. Compared to other telehealth services described in the literature that reported higher use rates [26], this COVID-19 teleconsultation service was offered in a medium-income country for a population with a low digital literacy level and no culture of using telehealth services. Therefore, the results of this study provide an important demonstration that, in specific scenarios, it is possible to offer high-quality health care using inexpensive and readily available technology in areas with limited resources and low digital literacy. However, it is important to note that the success of a telehealth initiative in achieving good adoption and positive clinical results also relies on the agreements made with the local health managers and on the integration of this modality of care with existing in-person health care at different levels. Maintaining a partnership is essential to solve the usual drawbacks associated with implementation of a telehealth program, such as infrastructure and accessibility issues, and to renegotiate the roles and responsibilities of the stakeholders whenever necessary.

Finally, with the increasing use of technology worldwide, the use of telehealth tools is becoming more feasible, and legislations in this regard are being reviewed or created along with growing evidence about professional and user acceptance and usability [26-29]. It seems that the use of telehealth tools is occurring in a one-way direction, without return. Therefore, scientific efforts are needed to clarify their broader potential and limitations.

Comparison With Prior Work

A large amount of evidence in different fields of telehealth has been produced; however, data regarding the clinical efficacy of this practice in the context of COVID-19 remain scarce [5,6].

A systematic review of 64 studies focusing on telehealth-based services for COVID-19 [5] revealed that only two studies measured mortality as an outcome and neither of these studies included control groups in their designs [30,31]. Another systematic review prepared by the Johns Hopkins University Center for Evidence-Based Practice evaluated the use of telehealth during the pandemic in different settings, but no comparative studies on the effectiveness of COVID-19 telehealth programs in reducing mortality were available [6].

Only two studies are available regarding the comparison of hospitalization rates between patients who were assisted by a specific COVID-19 telehealth program during the illness course and patients who did not receive telehealth support, with conflicting results [7,8]. Considering the sample sizes of both studies and the follow-up duration, the conclusion of the systematic review was that among the patients who received care for COVID-19, those who received an initial telehealth visit might have higher hospitalization rates compared with those of patients who received only in-person care (strength of evidence: low). This statement is open for discussion, as the primary study responsible for this conclusion in the systematic review presented additional information and results [7]. Although the primary study indicated that a high-intensity and protocolized telemonitoring program is related to higher hospitalization rates, this result was predictable given that the intensity of the telemonitoring actions was tailored to the risk of the patients in presenting adverse outcomes. Conversely, when considering all patients included in the study, the comparison between the groups showed that patients who were involved in a systematic telemonitoring program, which was tailored to the age and clinical characteristics of the patients, presented lower rates of hospital admissions than those of the patients who did not undergo the systematic telemonitoring program for COVID-19. Our results are consistent with the findings of this previous study, since we also demonstrated that engaging in systematic telehealth support provided through widely available technology can lead to positive clinical outcomes. Thus, our results further underscore the potential of the use systematic telehealth support as part of a comprehensive health care strategy.

It is worth emphasizing that both studies included in the systematic review [7,8] were conducted in high-income countries and involved relatively short follow-up (68 days and 30 days). In contrast, we evaluated the clinical effectiveness of a structured multilevel public COVID-19 telehealth service in a limited-resource region and included 82,182 patients who were followed up for 1 year. In this pioneering analysis, which spans an extended duration and involves a significant number of patients, the findings highlighted that embracing a public COVID-19 telehealth service yielded a positive impact with respect to hospital admissions without increasing mortality.

Limitations

The most important limitation of this study is related to the data quality of the public databases used in the analysis, mainly regarding the comorbidities, socioeconomic status, and ethnicity of the included patients.

Regarding the comorbidities, although these data were not missing, there is no orientation about the proper definition of each comorbidity. As the notification fields can be filled in by any health professional, the understanding about each comorbidity may vary. The last Brazilian survey showed a prevalence of self-reported hypertension of 23.9% in the adult population, while the estimated prevalence of self-reported diabetes is 7.7% [32]. Our data indicated a lower prevalence of comorbidities, with a registered prevalence of 3.4% and 2.6% for chronic cardiovascular diseases and diabetes, respectively, in the studied population. Despite this limitation, data misfiling likely occurred in both groups in the same fashion.

Regarding the socioeconomic status and ethnicity of the patients, although this information could offer additional context for interpreting the results, as these data were missing, they were not included in the analysis. However, it is important to note that since TeleCOVID-MG was a public and free service, most patients supported by this telehealth program did not have private health insurance.

XSL•FO
RenderX

Conclusions

Despite the massive growth of telehealth initiatives during the pandemic, data regarding the clinical effectiveness of this approach targeting flu syndrome are still lacking [5,6]. This was a retrospective cohort study conducted in a limited-resource region to evaluate the impact of the adoption of a public, structured, multilevel COVID-19 telehealth service for relevant clinical outcomes among adult patients with flu syndrome in comparison with the standard health assistance currently provided, which is primarily delivered in an in-person manner. This study demonstrates that a structured multilevel COVID-19 telehealth service adopted over a large scale can contribute to decreasing the rate of hospital admissions without increasing the mortality and may be considered as a potential and valuable health care strategy, even in regions with limited resources. The success of telehealth initiatives relies on a partnership between the involved stakeholders to coordinate actions within an established health care plan that addresses the specificities of the different target populations.

Acknowledgments

The authors would like to thank the entire TeleCOVID-MG team, including physicians, nurses, medical students, managers, and all other health professionals. This work was supported by Assembleia Legislativa de Minas Gerais (SESU/MEC; SEI-23072.211119/2020-10) and Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES; grant number 88887.507149/2020-00). ALPR is supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; grants 310679/2016-8 and 465518/2014-1). MS Marcolino is supported in part by CNPq (grant 10561/2021-3).

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to restrictions related to the confidentiality and privacy of the research participants.

Authors' Contributions

CRAO and ALPR coordinated the research. CRAO, MC Pires, KCM, JCdJ, MC Paixão, MS Mendes, LBR, MBMA, and ALPR obtained and organized the data. CRAO, MC Pires, KCM, JCdJ, INB, LBR, MS Marcolino, and ALPR analyzed the data. CRAO, KCM, JCdJ, INB, and MS Marcolino drafted the manuscript. All authors reviewed and edited the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

1. Dhaliwal JK, Hall TD, LaRue JL, Maynard SE, Pierre PE, Bransby KA. Expansion of telehealth in primary care during the COVID-19 pandemic: benefits and barriers. J Am Assoc Nurse Pract 2021 Jun 07;34(2):224-229. [doi: 10.1097/JXX.0000000000000626] [Medline: 34107501]
2. Wijesooriya NR, Mishra V, Brand PLP, Rubin BK. COVID-19 and telehealth, education, and research adaptations. Paediatr Respir Rev 2020 Sep;35:38-42 [FREE Full text] [doi: 10.1016/j.prrv.2020.06.009] [Medline: 32653468]
3. Marcolino MS, Diniz CS, Chagas BA, Mendes MS, Prates R, Pagano A, et al. Synchronous teleconsultation and monitoring service targeting COVID-19: leveraging insights for postpandemic health care. JMIR Med Inform 2022 Dec 22;10(12):e37591 [FREE Full text] [doi: 10.2196/37591] [Medline: 36191175]
4. Alkmim MBM, Marcolino MS, de Oliveira CRA, Borges IN, Cardoso CS, Rocha GM, et al. TeleCOVID-19: A multifaceted strategy from a public Brazilian telehealth service during the COVID-19 pandemic. In: Maeder AJ, Higa C, Van Den Berg MEL, Gough C, editors. Telehealth Innovations in Remote Healthcare Services Delivery: Global Telehealth 2020. Amsterdam: IOS Press; 2020:1-10.
5. Khoshrounejad F, Hamednia M, Mehrjerd A, Pichaghsaz S, Jamalirad H, Sargolzaei M, et al. Telehealth-based services during the COVID-19 pandemic: a systematic review of features and challenges. Front Public Health 2021;9:711762 [FREE Full text] [doi: 10.3389/fpubh.2021.711762] [Medline: 34350154]
6. Hatef E, Wilson R, Hannum S, Zhang A, Kharrazi H, Weiner J, et al. Use of Telehealth During the COVID-19 Era. Report No. 23-EHC005. Rockville, MD: Agency for Healthcare Research and Quality; 2023.
7. Casariego-Vales E, Blanco-López R, Rosón-Calvo B, Suárez-Gil R, Santos-Guerra F, Dobao-Feijoo MJ, On Behalf Of The Telea-Covid Lugo Comanagement Team. Efficacy of telemedicine and telemonitoring in at-home monitoring of patients with COVID-19. J Clin Med 2021 Jun 29;10(13):2893 [FREE Full text] [doi: 10.3390/jcm10132893] [Medline: 34209725]
8. Korycinski S, Metcalf D, Keteyian C. Effectiveness of a telephone-based nursing intervention to reduce hospital utilization by COVID-19 patients. Public Health Nurs 2022 Sep 25;39(5):940-948 [FREE Full text] [doi: 10.1111/phn.13074] [Medline: 35334128]

XSL•FO
RenderX

9. Soriano Marcolino M, Minelli Figueira R, Pereira Afonso Dos Santos J, Silva Cardoso C, Luiz Ribeiro A, Alkmim MB. The experience of a sustainable large scale Brazilian telehealth network. Telemed J E Health 2016 Nov;22(11):899-908. [doi: 10.1089/tmj.2015.0234] [Medline: 27167901]
10. Chagas BA, Pagano AS, Prates RO, Praes EC, Ferreguetti K, Vaz H, et al. Evaluating user experience with a chatbot designed as a public health response to the COVID-19 pandemic in Brazil: mixed methods study. JMIR Hum Factors 2023 Apr 03;10:e43135 [FREE Full text] [doi: 10.2196/43135] [Medline: 36634267]
11. Cities and states: Teófilo Otoni. Brazilian Institute of Geography and Statistics. 2010. URL: https://www.ibge.gov.br/cidades-e-estados/mg/teofilo-otoni.html [accessed 2023-04-04]
12. Cities and states: Divinópolis. Brazilian Institute of Geography and Statistics. URL: https://www.ibge.gov.br/cidades-e-estados/mg/divinopolis.html [accessed 2023-04-04]
13. Guia de Vigilância Epidemiológica Covid-19: Emergência da Saúde Pública de Importância Nacional pela Doença pelo Coronavírus 2019 - Covid-19. Ministério da Saúde, Brasil. URL: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/coronavirus/guia-de-vigilancia-epidemiologica-covid-19_2021.pdf/view [accessed 2023-04-04]
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007 Oct 16;4(10):e296 [FREE Full text] [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
15. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Cham: Springer; 2020.
16. Harrell Jr FE. Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer; 2016.
17. Garfan S, Alamoodi A, Zaidan B, Al-Zobbi M, Hamid RA, Alwan JK, et al. Telehealth utilization during the Covid-19 pandemic: a systematic review. Comput Biol Med 2021 Nov;138:104878 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104878] [Medline: 34592585]
18. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
19. Alkmim MB, Minelli Figueira R, Soriano Marcolino M, Silva Cardoso C, Pena de Abreu M, Rodrigues Cunha L, et al. Improving patient access to specialized health care: the Telehealth Network of Minas Gerais, Brazil. Bull World Health Org 2012 May 01;90(5):373-378. [doi: 10.2471/blt.11.099408]
20. Souza GFD, Figueira RM, Alkmim MB, Sousa LAPD, Bonisson L, Ribeiro ALP, et al. Teleophthalmology screening for diabetic retinopathy in Brazil: applicability and economic assessment. Telemed J E Health 2020 Mar 01;26(3):341-346. [doi: 10.1089/tmj.2018.0241] [Medline: 30994411]
21. Etges APBDS, Zanotto BS, Ruschel KB, da Silva RS, Oliveira M, de Campos Moreira T, et al. Telemedicine versus face-to-face care in ophthalmology: costs and utility measures in a real-world setting. Value Health Reg Issues 2022 Mar;28:46-53 [FREE Full text] [doi: 10.1016/j.vhri.2021.06.011] [Medline: 34800831]
22. Rodrigues DLG, Belber GS, Padilha FVDQ, Spinel LF, Moreira FR, Maeyama MA, et al. Impact of teleconsultation on patients with type 2 diabetes in the Brazilian public health system: protocol for a randomized controlled trial (TELEconsulta Diabetes Trial). JMIR Res Protoc 2021 Jan 21;10(1):e23679 [FREE Full text] [doi: 10.2196/23679] [Medline: 33475516]
23. Zanotto BS, Etges APBDS, Siqueira AC, Silva RSD, Bastos C, Araujo ALD, et al. Economic evaluation of a telemedicine service to expand primary health care in Rio Grande do Sul: TeleOftalmo's microcosting analysis. Cien Saude Colet 2020 Mar;25(4):1349-1360 [FREE Full text] [doi: 10.1590/1413-81232020254.28992019] [Medline: 32267437]
24. de Oliveira CRA, da Silva Etges APB, Marcolino MS, Paixão MC, Mendes MS, Ribeiro LB, et al. COVID-19 telehealth service can increase access to the health care system and become a cost-saving strategy. Telemed J E Health 2023 Jul 23;29(7):1043-1050. [doi: 10.1089/tmj.2022.0240] [Medline: 36445772]
25. Portaria nº 467, de 20 de Março de 2020. Diário Oficial da União. Ministério da Saúde. Gabinete do Ministro, Brasil. 2020 Mar 20. URL: https://www.in.gov.br/en/web/dou/-/portaria-n-467-de-20-de-marco-de-2020-249312996 [accessed 2023-04-04]
26. Omboni S, Padwal RS, Alessa T, Benczúr B, Green BB, Hubbard I, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 04;1:7-35 [FREE Full text] [doi: 10.20517/ch.2021.03] [Medline: 35233563]
27. Gonçalves RL, Pagano AS, Reis ZSN, Brackstone K, Lopes TCP, Cordeiro SA, et al. Usability of telehealth systems for noncommunicable diseases in primary care from the COVID-19 pandemic onward: systematic review. J Med Internet Res 2023 Mar 16;25:e44209 [FREE Full text] [doi: 10.2196/44209] [Medline: 36787223]
28. Isautier JM, Copp T, Ayre J, Cvejic E, Meyerowitz-Katz G, Batcup C, et al. People's experiences and satisfaction with telehealth during the COVID-19 pandemic in Australia: cross-sectional survey study. J Med Internet Res 2020 Dec 10;22(12):e24531 [FREE Full text] [doi: 10.2196/24531] [Medline: 33156806]
29. Nanda M, Sharma R. A review of patient satisfaction and experience with telemedicine: a virtual solution during and beyond COVID-19 pandemic. Telemed J E Health 2021 Dec 01;27(12):1325-1331. [doi: 10.1089/tmj.2020.0570] [Medline: 33719577]

XSL•FO
RenderX

30. Harris DA, Archbald-Pannone L, Kaur J, Cattell-Gordon D, Rheuban KS, Ombres RL, et al. Rapid telehealth-centered response to COVID-19 outbreaks in postacute and long-term care facilities. Telemed J E Health 2021 Jan;27(1):102-106 [FREE Full text] [doi: 10.1089/tmj.2020.0236] [Medline: 32644899]
31. Sitammagari K, Murphy S, Kowalkowski M, Chou S, Sullivan M, Taylor S, et al. Insights from rapid deployment of a "virtual hospital" as standard care during the COVID-19 pandemic. Ann Intern Med 2021 Feb;174(2):192-199 [FREE Full text] [doi: 10.7326/M20-4076] [Medline: 33175567]
32. Pesquisa Nacional de Saúde: 2019: percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal: Brasil e grandes regiões. IBGE, Coordenação de Trabalho e Rendimento, Brasil. 2020. URL: https://www.pns.icict.fiocruz.br/wp-content/uploads/2021/02/liv101764.pdf [accessed 2023-04-04]

Abbreviations

MIS: Mortality Information System
OR: odds ratio
SARS: Severe Acute Respiratory Syndrome
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TNMG: Telehealth Network of Minas Gerais
UFMG: Universidade Federal de Minas Gerais

Edited by T de Azevedo Cardoso; submitted 25.04.23; peer-reviewed by O Cohen, X Zhang; comments to author 09.08.23; revised version received 02.10.23; accepted 23.04.24; published 10.06.24.

Please cite as:
Oliveira CRA, Pires MC, Meira KC, de Jesus JC, Borges IN, Paixão MC, Mendes MS, Ribeiro LB, Marcolino MS, Alkmim MBM, Ribeiro ALP
Effect of a Structured Multilevel Telehealth Service on Hospital Admissions and Mortality During COVID-19 in a Resource-Limited Region in Brazil: Retrospective Cohort Study
J Med Internet Res 2024;26:e48464
URL: https://www.jmir.org/2024/1/e48464
doi:10.2196/48464
PMID:38857068

©Clara Rodrigues Alves Oliveira, Magda Carvalho Pires, Karina Cardoso Meira, Jordana Cristina de Jesus, Isabela Nascimento Borges, Maria Cristina Paixão, Mayara Santos Mendes, Leonardo Bonisson Ribeiro, Milena Soriano Marcolino, Maria Beatriz Moreira Alkmim, Antonio Luiz Pinho Ribeiro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Care Integration for Hepatitis C Virus Treatment Through Facilitated Telemedicine Within Opioid Treatment Programs: Qualitative Study

Suzanne S Dickerson[1], PhD; Saliyah J George[2], MPH; Ana Ventuneac[3], PhD; Arpan Dharia[4], MD; Andrew H Talal[4], MPH, MD

[1]Division of Tenured and Tenure-track Faculty Development, School of Nursing, University at Buffalo, State University of New York, Buffalo, NY, United States

[2]Division of Liver Diseases, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[3]START Treatment & Recovery Centers, Brooklyn, NY, United States

[4]Division of Gastroenterology, Hepatology and Nutrition, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY, United States

Corresponding Author:
Andrew H Talal, MPH, MD
Division of Gastroenterology, Hepatology and Nutrition
Jacobs School of Medicine and Biomedical Sciences
University at Buffalo, State University of New York
UB-CTRC, Suite 6090
875 Ellicot Street
Buffalo, NY, 14203
United States
Phone: 1 716 829 5354
Fax: 1 716 854 1397
Email: ahtalal@buffalo.edu

Abstract

Background: Telemedicine has the potential to remove geographic and temporal obstacles to health care access. Whether and how telemedicine can increase health care access for underserved populations remains an open question. To address this issue, we integrated facilitated telemedicine encounters for the management of hepatitis C virus (HCV), a highly prevalent condition among people with opioid use disorder (OUD), into opioid treatment programs (OTPs). In New York State, OTPs are methadone-dispensing centers that provide patient-centered, evidence-based treatment for OUD. We investigated the integration and impact of facilitated telemedicine into OTP workflows in these settings.

Objective: This study aims to understand OTP staff experiences with integrating facilitated telemedicine for HCV treatment into OTPs, including best practices and lessons learned.

Methods: We conducted semistructured interviews with 45 OTP staff members (13 clinical, 12 administrative, 6 physicians, and 14 support staff members) at least one year after the implementation of facilitated telemedicine for HCV management. We used hermeneutic phenomenological analysis to understand OTP staff experiences.

Results: We identified 4 overarching themes illustrating the successful integration of facilitated telemedicine for HCV care into OTPs. First, integration requires an understanding of the challenges, goals, and values of the OTP. As OTP staff learned about new, highly effective HCV therapies, they valued an HCV cure as a “win” for their patients and were excited about the potential to eliminate a highly prevalent infectious disease. Second, the integration of facilitated telemedicine into OTPs fosters social support and reinforces relationships between patients and OTP staff. OTP staff appreciated the ability to have “eyes on” patients during telemedicine encounters to assess body language, a necessary component of OUD management. Third, participants described high levels of interprofessional collaboration as a care team that included the blurring of lines between disciplines working toward a common goal of improving patient care. Study case managers were integrated into OTP workflows and established communication channels to improve patient outcomes. Fourth, administrators endorsed the sustained and future expansion of facilitated telemedicine to address comorbidities.

XSL•FO
RenderX

Conclusions: OTP staff were highly enthusiastic about facilitated telemedicine for an underserved population. They described high levels of collaboration and integration comparable to relevant integrative frameworks. When situated within OTPs, facilitated telemedicine is a high-value application of telemedicine that provides support for underserved populations necessary for high-quality health care. These experiences support sustaining and scaling facilitated telemedicine in comparable settings and evaluating its ability to address other comorbidities.

Trial Registration: ClinicalTrials.gov NCT02933970; https://clinicaltrials.gov/study/NCT02933970

(J Med Internet Res 2024;26:e53049) doi:10.2196/53049

KEYWORDS

hepatitis C virus; integrated treatment; facilitated telemedicine; substance users; people with opioid use disorder; opioid; opioids; telemedicine; telehealth; eHealth; e-health; ICT; substance use; substance abuse; HCV; hepatitis; liver; interview; interviews; qualitative; hermeneutic; phenomenological; implementation; integration; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; addict; addiction; addictions; addicts; hepatic

Introduction

Care integration adds value to existing health care venues by providing new point-of-care options, especially for underserved populations. Fragmented health care is a prevailing problem, exemplified by the separation of medical and behavioral health care [1]. The artificial separation can be particularly problematic for stigmatized populations, such as people with opioid use disorder (OUD), who frequently face difficulties navigating and accessing traditional health care settings [2]. Additionally, people with OUD have high hepatitis C virus (HCV) incidence and prevalence due to injection drug use and limited access to HCV therapy [3]. To achieve HCV elimination by 2030, as proposed in several jurisdictions [4-6], HCV treatment access must increase.

We sought to address restricted HCV care access for people with OUD through a randomized controlled trial of care integration using technology that is at the intersection of the internet or videoconferencing, and health care. Specifically, we used a facilitated telemedicine model integrated into 12 opioid treatment programs (OTPs) throughout New York State [7,8]. Facilitated telemedicine is an approach to health care delivery that combines social interactions between patients and staff (ie, case managers and clinical staff) with bidirectional videoconferencing between patients and geographically separated providers to promote trust. Facilitated telemedicine can integrate HCV treatment into sites, such as OTPs, that are convenient and familiar to patients with OUD but that lack HCV services and infrastructure. Methadone dispensing in the United States is highly structured and requires attendance at a licensed OTP, a system that was created through specific regulations [9]. A physician or advanced practice provider initially evaluates patients, and nurses administer methadone under direct observation.

OTPs in New York State assess, titrate, and dispense methadone in combination with patient-centered counseling to manage people with OUD, including symptoms of opioid withdrawal [2,10,11]. These approaches, combined with onsite medical services such as phlebotomy and HCV medication dispensing in an OTP setting, promote the integration of HCV treatment into OUD care. While OTPs can be stigmatizing, in our experience, people with OUD generally consider OTPs destigmatizing; they can be viewed as “safe spaces” and potential “medical homes” for integrating expanded health care delivery [12,13]. We integrated facilitated telemedicine into OTPs to connect patients with HCV specialists.

Since knowledge is limited concerning staff considerations that affect facilitated telemedicine integration into OTP settings, we investigated OTP staff’s experiences of facilitated telemedicine integration for HCV care into OTPs. We used the sociotechnical system framework to identify telemedicine delivery issues and potential sociotechnical concerns for people with OUD [14]. We sought to understand the meaning of facilitated telemedicine integration for staff in everyday clinical practice and the role that staff played in the facilitation. Our goal is to provide insights and potential considerations for future integration. Investigating staff’s common meanings and shared practices promotes understanding of integration successes, challenges, and lessons learned. We compared our results to the conceptual care integration framework developed by the Center for Integrated Health Solutions [15].

Methods

Parent Study Summary

We recently completed a stepped-wedge, randomized controlled trial comparing HCV treatment through facilitated telemedicine integrated into 12 OTPs throughout New York State compared to offsite referral (usual care) [16]. In New York State, OTPs are methadone-dispensing centers that provide patient-centered harm reduction approaches using an evidence-based medical treatment for OUD [10]. Although methadone as an OUD treatment is supported by New York State Medicaid, OTPs do not routinely offer primary care. Of the 12 study sites, 4 were community organizations, 4 were university-affiliated, and 4 were affiliated with health systems. In total, 10 of the 12 sites had their own institutional review boards; the board from the parent study site covered the other 2 sites. At the initiation of the study (ie, 2017), 10 of 12 sites offered onsite phlebotomy. By the end of the study (ie, 2022), all sites offered onsite phlebotomy. We commenced participant recruitment on March 1, 2017, and concluded on February 29, 2020.

We recruited a total of 602 participants. Both in-person and telemedicine encounters were managed by hepatitis specialists, who forwarded notes of the HCV consultation to the OTP for inclusion in the medical record. The entire facilitated

XSL•FO
RenderX

telemedicine encounter was integrated into the OTP with bidirectional videoconferencing between the patient-participant and the hepatitis provider. Third-party coverage of clinical care and direct-acting antivirals were study inclusion criteria. A study-supported case manager was the research team representative at each study site and was credentialed in human subjects training. The case manager facilitated the telemedicine encounter and was also an HCV educator, patient navigator, and advocate. Direct-acting antivirals for HCV treatment were dispensed in the OTP alongside methadone for OUD. Take-home doses were available on a case-by-case basis. Subsequently, we followed cured participants for 2 years posttreatment cessation to assess for reinfection.

Methodology

We used hermeneutic phenomenology to reveal understandings of human situations as experienced within a context of time, place, and situational influences [17,18]. In this approach, we gain an understanding of the experience of a technology from the perspective of health care as a human experience interacting with one another [19] versus a replacement of social aspects of care. In a postpositivist or interpretive framework, hermeneutic phenomenology uses an a priori (ie, no preconceived theory) approach to formulate open-ended questions, interpret the interview text, and explicate common meanings and shared practices. Themes are revealed through the interpretation of the language of experiences, as portrayed in staff narrative interviews. For details concerning hermeneutic phenomenology, please see [20,21] and Multimedia Appendices 1 and 2 [20,22-27].

Setting and Recruitment

We used purposive sampling to recruit OTP staff and administrators involved in HCV care integration at least one year after the initiation of facilitated telemedicine. We interviewed approximately 4 participants per site, representing 11 sites. We interviewed 45 participants, including 16 clinical, 15 administrative, and 14 support staff members (Table 1).

XSL•FO
RenderX

Table 1. Demographics of the 45 participants interviewed.

Characteristic	Participants, n (%)
Age (years)	
21-40	9 (20)
41-60	26 (58)
61-80	9 (20)
Not disclosed	1 (2)
Sex	
Male	17 (38)
Female	28 (62)
Race	
Black	18 (40)
White	22 (49)
Other (Latino, Puerto Rican, or mixed)	3 (7)
Not disclosed	2 (4)
Ethnicity	
Hispanic	3 (7)
Non-Hispanic	38 (84)
Other (this included Jamaican and Jewish)	2 (4)
Not disclosed	2 (4)
Highest educational attainment	
Less than bachelor's	4 (9)
Bachelor's	13 (29)
More than bachelor's	11 (24)
Professional certification-board certified, LMSW[a], PA[b], or CASAC[c]	17 (38)
Staff type	
Administrator (directors or program administrators)	12 (27)
Clinical (registered nurse, nurse practitioner, or PA)	13 (29)
Physicians	6 (13)
Support (counselors, patient account rep, or others)	14 (31)

[a]LMSW: licensed medical social worker.

[b]PA: physician assistant.

[c]CASAC: certified alcohol and substance abuse counselor.

Interview Conduct

SSD interviewed all participants from January to July 2021. We developed separate interview guides for staff and administrators (Multimedia Appendix 2). The initial open-ended question inquired about what the participants viewed as the most important aspect of integrated care through facilitated telemedicine. If not discussed in the open-ended inquiry, we subsequently used probes for further elaboration on participants' experiences of facilitated telemedicine and integration frameworks [14,21]. Interviews were recorded by Zoom (Zoom Video Communications), transcribed, deidentified, and verified by staff by comparing transcriptions against the recording for accuracy. One participant opted for an audio-only recording. The interviews lasted 30-60 minutes, and no participant prematurely discontinued the interview. The interviewer confirmed the final transcript version before the analysis.

Analysis

The multidisciplinary analysis team included a hermeneutic phenomenology expert (SSD), a social psychologist (AV), and a case manager (SJG), who performed the initial thematic coding. Subsequently, the study principal investigator (AHT) and study director (AD) joined the analysis team for the final coalescence of themes. SSD guided the analysis using de Witt and Ploeg's [22] framework for rigor while simultaneously maintaining balanced integration between the participants' voices and the researchers' interpretations. The analysis team

XSL•FO
RenderX

interpreted the transcripts and field notes in a reflective process that followed iterative steps [20,23]. The team met weekly to identify emergent themes with supportive quotes, which were discussed and refined in an iterative fashion. The team achieved openness by exposing the team's preunderstandings (biases) through a continuous, systematic process of auditing the team's interpretations to support theme development consistent with the participants' experiences. Any thematic disagreements among team members resulted in a return to the original interview transcript for verification. By team consensus, new themes were added as relevant, and the process continued until themes were refined to explicate the phenomenon. The final themes were examined by team members for coherence and comprehensiveness, providing a warranted interpretation.

Ethical Considerations

The study was approved by the institutional review board at the University of Buffalo (parent institution; STUDY00000724) and by the institutional review boards at each of the sites where the study was conducted. All participants signed informed consent and provided verbal consent reconfirmation before the interview. No participants opted out of the study. All data transcripts were deidentified. Participants received US $25 for their participation, if permissible. Included in this article is an aggregate summary of the data generated during the study. Confidentiality concerns prohibit us from sharing individual data transcripts.

Results

Participant interviews revealed 4 themes related to the integration of facilitated telemedicine into a behavioral health setting (Figure 1). OTP staff and administrators shared their experiences of HCV care through facilitated telemedicine integrated into OTPs.

Figure 1. Staff experiences of hepatitis C virus (HCV) care integration are illustrated as themes (green) and subthemes (blue). OTP: opioid treatment program.

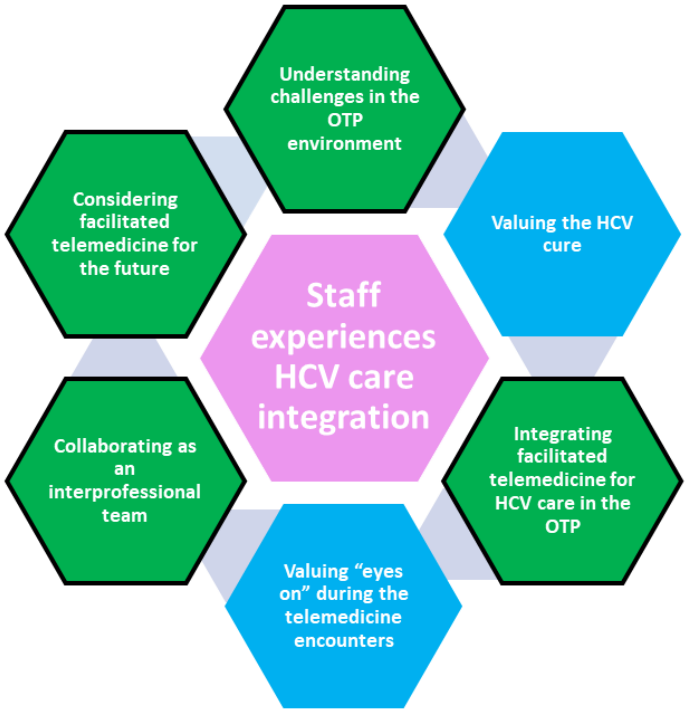

Theme 1: Understanding Challenges in the OTP Environment as a Point of HCV Care for People With OUD

Overview

In the context of working in the OTP, participants reflected on their common experiences managing physical withdrawal symptoms and the psychological effects of the OUD. Staff frequently addressed barriers to care, such as the patients' competing priorities as well as the difficulties engaging people with OUD who have a "disease of the brain." Staff narratives additionally explained situations when patients previously attempted to access health care outside of the OTP that were frequently accompanied by unjustified scorn. One staff member explained:

> *If [patients] go to their primary doctor and say, "I'm addicted to opiates, I'm on methadone," some doctors won't even treat them, they will shun them.*

Participants alternatively felt that people with OUD considered the OTP environment a safe, destigmatizing space in which to receive health care. Thus, the OTP was an appropriate place in which to provide point-of-care treatment for HCV. As one staff member reflected:

> *Care integration makes so much sense, we should have done it long ago.... Here is a subset of a population you are not going to reach unless you're willing to do something like this.*

OTP staff create nonjudgmental environments where patients feel comfortable and accepted to build patient trust. As one administrator noted, "[Patients] understand they're here for

XSL•FO
RenderX

treatment, they're surrounded by people who care about them, who want to listen." Another administrator reflected that OTP staff adopted a problem-solving, patient-centered approach to OUD treatment by, "trying to meet, treat patients in an individualized way." Another participant reflected on the care that OTP staff seek to provide: "We're not here to judge them, we're here to take care of them." Staff also face challenges in hectic OTP environments that have large patient loads, as one administrator reflected:

> *The staff overall do a remarkable job under challenging circumstances with inadequate resources, difficult environments [including] challenging [external] communities that don't want these programs [due to] stigma.*

Nonetheless, participants understood that the OTP's culture fosters their recognition as "safe spaces" for people with OUD. Participants related that successful integration required an understanding of the OTP's challenges and values.

Valuing the HCV Cure

Participants were very excited to learn about offering the benefit of a newer treatment that provides an HCV cure to their patients. Early in the study, however, participants were uncertain if patients would accept the newer treatment option for HCV care. A staff member related, "You may not get patient buy - in [for HCV treatment] right at the beginning.... They're in crisis, they could be in withdrawal." To promote buy-in for HCV treatment, study personnel provided HCV education to promote program endorsement for both staff and patients, as both groups were more familiar with older, less-effective therapeutic regimens. As a staff participant explained, "I was really excited hearing that there were [HCV] medications.... Highly effective."

As time went on, staff-participants also appreciated the opportunity for onsite HCV treatment to benefit their patients. Staff described how the patients who completed treatment shared their endorsement with other peers and became a peer pipeline for the facilitated telemedicine approach, as one staff member reported:

> *A number of my patients have been cured and their lives have improved.... When folks were introduced to the possibility of being in the study, they didn't run from it, they actually ran to it.*

Participants appreciated the accessibility and convenience of integrating HCV care into venues already frequented by people with OUD. The enthusiasm for the program grew over time as OTP staff became involved in HCV care integration in the OTP. As one administrator suggested:

> *Everyone benefited, mostly the patients. Not only did we help save lives, but [we] injected back into the program opportunities for learning and feeling that we're participating in advancing science.*

Participants experienced the merits of an HCV cure through patient success stories of accessing treatment that provided hope for the future. As one staff member reflected, "There's no need to languish, you can get treatment and there is a better tomorrow." An administrator commented that patients are "absolutely thrilled [with the] opportunity to have a successful medical intervention, it's life changing." In addition, the clinical staff reflected on the observed outcomes of treatment. One said, "Their countenance changed, [it was] brighter. Their level of confidence [increased]." Another clinical staff person commented on the outcomes:

> *The responses were great, in fact their methadone doses decreased, their mental status got better, their overall symptoms got better. Commitment to the [HCV] treatment gave them another responsibility that helped in their recovery process.*

Theme 2: Integrating Facilitated Telemedicine for HCV Care Into the OTP

Overview

Participants were overwhelmingly positive toward HCV treatment through facilitated telemedicine integrated into the OTP that fosters social support and relationship building between staff and patients, which is foundational for effective health care delivery. One staff member reflected:

> *They build trust here because they come here every day. They're used to our doctors [and] nurses.... They don't want to go anywhere else to get treated.*

Therefore, a quality program that endorses patient-centered approaches is supported within the OTP even when it includes the technology of telemedicine as long as a concurrent focus is on the social aspects of care, which are foundational to the OTP. One administrator explained that telemedicine "allows increased access to treatment that improves quality and monitoring." Staff described the integration of behavioral and medical health with infectious disease treatment as creating a "one-stop shop of healthcare delivery."

Facilitated telemedicine is built on a foundation of preexisting trusting relationships between patients and staff [2,21]. Case managers coordinated the telemedicine encounters, and familiar clinical staff were present during the initial encounter. OTP staff's presence during the initial telemedicine encounter provided the necessary sociotechnical connections that were adapted to include the facilitated component that occurred within the OTP. A staff member described the essential components of the facilitated telemedicine encounter: "What makes it positive is that they have a familiar person onsite who they've already seen. That's the seamless aspect." Staff were pleased with the patients' quick and easy adaptation to facilitated telemedicine when the provider on the screen became familiar to the patients, which resulted in the transfer of trust across the screen. As one staff member explained:

> *The technology was new, and it took one or two visits and then they became very comfortable with [telemedicine] and talked [as] if the person was in the room.*

The presence of case managers during the telemedicine encounters created a bridge between the patient and the HCV specialist, rephrasing questions and addressing patient concerns. One staff person explained how the OTP clinical staff also facilitated the interaction for the specialist by providing essential

XSL•FO
RenderX

information and performing clinical assessments: "The PA [was] in the room [during the initial telemedicine encounter].... Someone who knew [the patient] and about [his or her] situation." Furthermore, physician assistants or nurse practitioners used their clinical skills, adding benefit to the facilitated telemedicine encounter as one clinical staff member explained, "For the initial visits, I did their physical [exam]." The benefit of facilitated telemedicine is bidirectional; the telemedicine provider collaborated with the onsite clinician to perform physical examinations as indicated, and the onsite clinician achieved enhanced job satisfaction through the application of new and enhanced clinical assessments to gauge liver health. Participants further labeled the facilitation as a "warm handoff" and "seamless" transition to the telemedicine provider. Over time, the interactions became more efficient, as one staff participant explained: "[Patients] asked appropriate questions [and] felt comfortable, that really helped with the compliance.... [They] got more involved in their own treatment." Furthermore, for the remote specialist, the support of those who facilitate the encounter provides the social aspect of the sociotechnical interaction. As one staff participant reported, "If someone is there, who is the specialist's eyes and ears, then it [telemedicine] works."

Valuing "Eyes On" During Telemedicine Encounters

Staff members who participated in the facilitated telemedicine encounters deliberated regarding the need for having an "eyes on" perspective of patients during the encounters to observe nonverbal cues and OUD-related behavioral manifestations. One clinical staff member compared telemedicine to face-to-face encounters to explain the value of being present in the same room, as necessary, for appropriate assessments:

> *Often when you're talking to them [patients], you notice things that look a little odd and that will trigger a question. "You have a black eye, did someone hit you?"*

The "eyes on" perspective was described by a clinical staff person as an important skill when assessing individuals' health:

> *You really want to lay eyes on the problematic patients that are more likely to be impaired.... When they're seeing the doctor, [patients] can get their act together.*

Counselors agreed on the importance of direct visualization to assess for manifestations of OUD: "Sometimes in the world of addiction psychiatry, body language gives a lot of information." Some staff also commented on the necessity of "eyes on" for other conditions such as rashes, swelling, and inflammation that could benefit from an "eyes on" perspective.

Theme 3: Collaborating as an Interprofessional Team

The OTP and study staff collaborated as an interprofessional team to integrate facilitated telemedicine into the OTP (Figure 2). This collaboration was initiated and supported by a sentiment of hope and excitement for the ability to offer needed HCV treatment for their patients. Staff spoke of its dedication to supporting an underserved population with a treatment that actually offered a cure for HCV, which was "a big deal." Staff were excited and open to this opportunity, which some called an innovation. One staff participant concluded:

> *We've learned how to collaborate from the staff and especially the counselors...[who] were excited that their [HCV] patients will be treated. The collaboration was at the highest level possible.*

The staff developed daily workflow processes to promote integration within each OTP according to the constellation of staff at each setting. An administrator explained how it worked at one OTP: "The counselors would notify the patients [of] an appointment, the account rep [representative] would remind patients, and the nurse practitioner helped to round up patients." Another site staff member elaborated, "Between the nurses and the counselors, we all work [together], encouraging the clients to participate, making them feel welcome."

Participants enthusiastically reported how the study case managers integrated within the OTP staff to focus on the necessary treatment processes. Case managers provided the necessary structure for the staff to participate and integrate HCV treatment through facilitated telemedicine within their usual workflow, positively influencing the OTP staff. Thus, OTP leadership viewed case managers as essential to the OTP workflow as one administrator related: "It really felt like the [case managers] were team members, not outside folks coming in to do a project." Case managers' collaborations made the HCV treatment seamless. An administrator reported:

> *Having the case manager was very helpful because there were so many pieces. If that [HCV treatment] had fallen on our staff, that would have really made this challenging...The case manager knew how to navigate our system.*

The navigation included identifying positive HCV cases, understanding specific treatment parameters, obtaining insurance approval for HCV medications, being present during all telemedicine encounters, following the patient through the HCV treatment course, and obtaining outcome measurements. One administrator reflected:

> *The [case manager role] is important because it allows the counselors to focus on the substance use disorder and the social workers to deal with the mental health issues.*

The case managers were also essential for patient enrollment support, as staff described: "The case manager was very involved with all these patients; she was able to relate to them within their world."

Participants worked collaboratively as a "care team" consisting of clinicians, nurses, counselors, and case managers. Clinicians answered patients' questions, provided support, and encouraged adherence. Nurses coordinated medication storage, dispensing, and take-home doses. One staff participant explained:

> *The nurses would get things set up. The client was dosed at the window, and they have their take-out [methadone] bottles and their hep C medicine in the same bag, and off they would go, it was seamless.*

XSL•FO
RenderX

Counselors, as part of their routine employment duties, communicated with patients regularly, which enabled case managers to effectively educate, recruit, engage, and retain participants. An administrator reported:

> *[Facilitated telemedicine] has helped the counselors work with patients towards the treatment plan throughout the course of the treatment, [counselors are] able to provide information, communicate with the patient, and help move things along.*

Figure 2. Collaborating as an interprofessional team. Study-supported case managers are integrated into the opioid treatment program (OTP) as critical to supporting OTP staff workflows. Clinical staff, including medical doctors (MDs), nurse practitioners (NPs), and physician assistants (PAs), conduct physical examinations during telemedicine encounters. Nurses coordinate medication storage, dispensing, and take-home doses. Support staff, such as counselors, incorporate hepatitis C virus (HCV) care into the treatment plan through engagement, education, recruitment, and retention of care. The HCV specialist orders medications and labs for HCV management through telemedicine encounters.

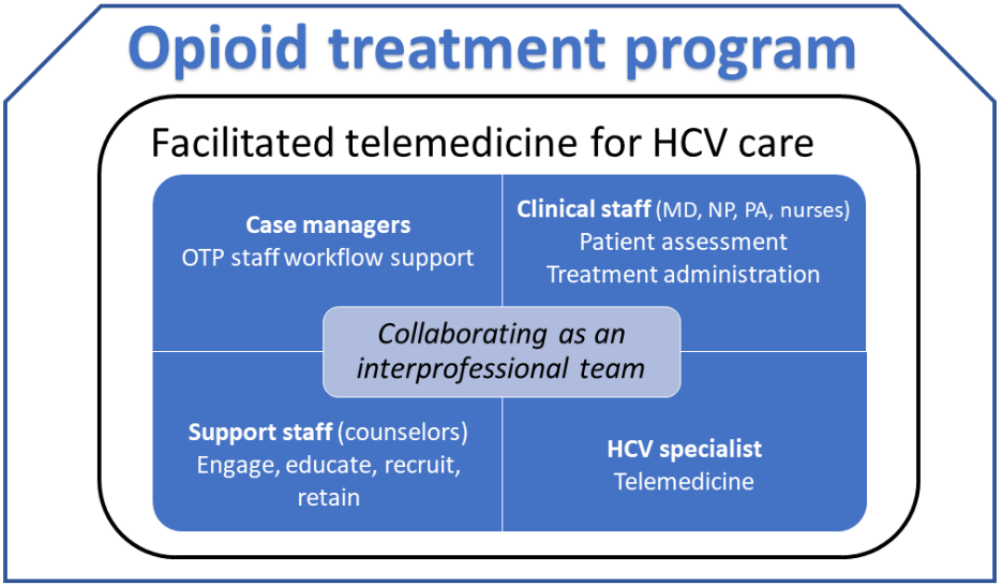

Theme 4: Considering Facilitated Telemedicine for the Future

OTP administrators endorsed sustaining and growing the facilitated telemedicine model. Participants provided meanings and suggestions for facilitated telemedicine integration (Table 2), and they advocated that the OTP environment provides an important foundation. Administrators agreed that facilitated telemedicine improved care access: “It makes so much sense [and] has made me think about the integration of other services, not just co-location, but real integration.” Participants commonly experienced their patients’ lack of follow-through with offsite health care referrals and suggested that facilitated telemedicine be expanded to address comorbidities. Additionally, participants felt the facilitated telemedicine model could be translated to other OTP settings, as one administrator explained:

> *[Patients] don’t like the idea of shopping around to different places to get care, so the idea is that they can come here to get their methadone, counseling, medical, primary, HIV [care], and diabetic* [care]

Staff suggested that additional potential expansion opportunities, especially through telemedicine, include “HIV, AIDS” and “internal medicine and chronic disease management including diabetes and hypertension.” Administrators indicated that their future includes leveraging OTP infrastructure to address comorbid health care issues.

Participants reflected on the positive outcomes of facilitated telemedicine for their patients, endorsing hope for their future health. One staff participant said:

> *I think having that little success, even just completing hep C treatment, is a win for them, it makes them more willing to treat other medical issues.*

However, using telemedicine solely without the facilitated portion could lead to health care that lacks the social component that is so necessary to build relationships and quality care. If health care becomes a standardized protocol devoid of human social interaction, there is a risk of dehumanizing health care. As one staff member related, now after the study, there is a change in HCV care as it is protocol-driven. Thus, if HCV is managed solely with technology, treatment effectiveness and satisfaction may be diminished, especially among underserved populations, such as people with OUD. One staff member said, “The hepatitis treatment is protocol driven. It’s very stepwise in process, a standardized workup...You can get a good workflow down” but not at the risk of eliminating human interaction, trust, and support.

XSL•FO
RenderX

Table 2. Participant insights of facilitated telemedicine.

Themes and meanings	Advice with supporting quotes
Theme 1: Understanding challenges in the OTP[a] environment as a point of care for people with OUD[b]	
Challenges of the OTP environment	Be flexible...to work with the patients, be ready for initial resistance. Get ready for a lot of no shows in the beginning. Persistence. [Staff]
	[What] would have helped [is being] able to do the blood draws onsite. We had to send our patients to the hospital to get their blood work done and even two blocks [distance] are a hurdle. [Administrator]
Valuing telemedicine in a familiar environment	Make sure that the client is aware of what telemedicine is, so that they are not scared...This population can be standoffish...So explain "it's to help you and your health conditions"...Communicate that it's beneficial and convenient." [Staff]
	I would recommend [telemedicine] remain in the program...Convenience and accessibility [are key]." [Staff]
	One patient was private [and] initially apprehensive. When he realized that he didn't have to go somewhere else meeting strangers [and he could remain] within the environment [and people] that he's accustomed to, he willingly participated." [Staff]
Understanding the HCV[c] cure	People who were completely cured would never have been cured...They would never go through the whole series of visits. [Staff]
	Clients speak to [other] clients about treatment and their experiences. [Staff]
Theme 2: Integrating facilitated telemedicine for HCV care into the OTP	
Understanding facilitated telemedicine	By seeing a telemedicine interaction, you see what is required and what is accomplished. You sing a different tune because you can see that you can solve problems. [Staff]
	Education is the most important part...Let them know that [telemedicine] is HIPAA-compliant. [Staff]
	Nothing's going to be shared, it's confidential, you have to establish patient trust. [Staff]
	I think telemedicine has been very helpful because of the ease of treatment...I've been astonished by the degree of compliance with treatment. The way we were able to successfully collaborate and integrate the HEP C treatment into the medication assisted treatment, to improve that compliance and to improve the quality and outcome of care, it's been amazing. [Administrator]
Onsite accessibility	You don't have to worry about the weather. Patients get nervous; "Am I going to get a ride to the clinic? What if I miss my appointment with my counselor? Are they going to let my probation officer know that I was noncompliant?" This way [telemedicine] takes so much anxiety away for the patient. [Administrator]
Information technology infrastructure needs	An onsite IT person to assist when things went wrong would be very helpful. [Administrator]
	If the Wi-Fi went down, we'd be really stuck. [It] would have helped to have technical backup in place. [Staff]
	Make sure you have good solid equipment because sometimes if the audio was bad, you always have poor connection, and it takes so long to set up. [Staff]
	[I would] like an integrated [HCV medication] dispensing record.... Our electronic system didn't support it. [Staff]
Theme 3: Collaborating as an interprofessional team-clinical support	[The case manager] has the ability to...be inviting to patients...He was enabling people. Choosing the [right] staff is important. [Staff]
	We provided "directly observed therapy," where we would give them their medicines when they come to the window. Some of these patients had to take their meds home because they don't come every day. I had to call patients on a number of occasions [to remind them]. Give a patient a beeper or a smartphone…We would set it [smartphone] up to alarm so that they'll remember to take the medicine. [Staff]
	Having a nurse available to do vitals, room a patient, and update medications, would have been nice. [Staff]
Theme 4: Considering facilitated telemedicine for the future-future considerations	"We're able to reach this subset of population that we would not have otherwise been able to reach This is a very viable approach.... I think it shows that there's a role for this" [Staff].
	"People who would benefit the most are freestanding clinics and small clinic networks… These independent clinics really stand to gain a lot. You're providing more services for your people, and you can really market this as a reason to go [to the clinic] as opposed to somewhere else" [Staff].
	If you have clients that have received the treatment, clients speaking to clients about treatment that they have already received and their experiences, that is the most impressionable to other people. [Staff]

[a]OTP: opioid treatment program.

[b]OUD: opioid use disorder.

[c]HCV: hepatitis C virus.

XSL•FO
RenderX

Discussion

Principal Results

OTP staff and administrators shared their experiences of HCV care through facilitated telemedicine integrated into OTPs. Participants understood that the OTPs' culture fosters their recognition as "safe spaces" for people with OUD. Participants related that successful integration required an understanding of the OTP's challenges and values (theme 1). Participants appreciated the accessibility and convenience of integrating HCV care into venues already frequented by people with OUD. The OTP environment, coupled with the trust and familiarity of individuals conducting telemedicine encounters, potentiates positive provider-patient relationships across the screen [12,13]. Similarly, the literature has identified these factors as essential components of the provider-patient relationship [28]. Participants valued an HCV cure as a "win" for their patients and as a motivator to address life challenges. As a result, participants became committed to HCV cure within their OTPs. The opportunities to interact with hepatitis specialists through facilitated telemedicine improved the provision of specialty care offered to OTP patients (theme 2). Counselors, nurses, and clinicians supported facilitated telemedicine for HCV care, which supplemented internal workflows and enriched job satisfaction. Counselors included HCV management in their patients' treatment plans, while nurses and clinicians monitored and educated patients. Case managers were responsible for HCV care delivery and were fully integrated into the OTP workflow (theme 3). Similarly, others have recently illustrated how integrated health care offers opportunities to promote interprofessional collaboration [1,29]. Displaying empathy, especially during telemedicine encounters, promotes high patient satisfaction [11,30] and is an essential component of the sociotechnical approach.

Comparisons With Previous Work and Standardized Frameworks

We compared participants' experiences to the Center for Integrated Health Solutions framework categories of collaboration, clinical delivery, and patient experience [15]. Participants indicated that they were mission-driven and patient-centered. Related to *collaboration*, participants integrated study-supported case managers into the OTP, established seamless workflows, and blended roles and responsibilities. The COVID-19 pandemic-associated implementation of telemedicine encounters induced similar changes in staff roles, workflows, and patient interactions [31]. Furthermore, amelioration of cultural differences between health care disciplines (ie, differences between medical and behavioral providers, such as interprofessional hierarchies, differences in treatment vocabularies, and variations in staff supervision) has been shown to occur most effectively by developing, supporting, sustaining, and growing integrated care teams [1]. A recent study that used video-observed therapy for HCV was permissive for support service offerings with the involvement of multidisciplinary teams [32]. A previous meta-analysis illustrated improved HCV therapeutic outcomes with the involvement of a multidisciplinary team [33]. Related to clinical delivery, participants described HCV treatment as protocolized, and they successfully adapted site-level OTP workflows to promote facilitated telemedicine integration with social support existent in the OTP [7]. Related to patient experience, facilitated telemedicine expanded access to patient-centered and equitable HCV care. Consistent with the sociotechnical system framework, participants reported that a dedicated, familiar individual (ie, case manager) is critical for successful telemedicine integration [14]. Thus, participants perceived that facilitated telemedicine increased health care quality and value [34,35]. The investigation described here begins to address knowledge gaps, including understanding of telemedicine's role as a component of behavioral-medical care integration [36].

Participant Meanings and Practical Advice

Participants recommended shared practices of facilitated telemedicine in the OTP, such as having "eyes on," that is, direct patient visualization, to assess OUD manifestations [37]. These observations underscore the potential ineffectiveness of audio-only solutions when evaluating people with OUD. Participants indicated that codispensing HCV medications and methadone increases HCV medication adherence and adds value by potentiating improvements in OUD recovery [38]. New York State regulations and OTP administrators permitted the codispensing of HCV medications and methadone. Additionally, managing HCV and OUD has been shown to improve treatment outcomes, adherence, and retention in care [39-41]. Patients' commitment to achieving an HCV cure improved their confidence and sense of self-worth, including an improved interest in their whole health and overall well-being [2]. Our observations, consistent with others, illustrate that telemedicine interventions for HCV treatment promote retention in OUD care. Treloar et al [42] found that an HCV and OUD colocated treatment model was overwhelmingly successful, largely because of the familiarity between patients and staff in the OTP setting. Other recent data illustrated improved HCV treatment uptake with integrated care [43]. Participants also offered suggestions for future facilitated telemedicine use (theme 4), suggestions supported by previous data illustrating behavioral health integration into different practice settings [1,29] and successful management of primary care conditions through telemedicine (Table 2) [36,44,45].

Strengths and Limitations

Due to the novelty of facilitated telemedicine targeted at people with OUD, we used hermeneutic phenomenology to investigate OTP staff's experiences. Examining the technological implications in a protocol-driven world illuminates the value of facilitated telemedicine, which emphasizes the importance of a trusting relationship between patients and providers. This therapeutic alliance creates a bridge to effective use of technology that does not abandon person-centeredness or widen the digital divide. Another study of Indigenous Australian individuals illustrated that telehealth enabled specialist consultations to occur in the safe environment of the Aboriginal Community Controlled Health Services [46]. In that model, an Indigenous health worker facilitated telehealth interactions and communication between community members and specialists. Subsequent investigations should evaluate facilitated

XSL•FO
RenderX

telemedicine integration into different settings through implementation frameworks.

In our case, we interviewed participants at least one year after telemedicine initiation and during COVID-19–associated lockdowns, thereby permitting evaluation of longer-term effects, including assessment of site-level, pandemic-associated changes in health care delivery. Our analysis team, which consisted of experts with diverse backgrounds in qualitative analysis, clinical research, and OUD, maintained data credibility [20,22,23]. Our telemedicine encounters were conducted as clinical research, which may limit their transferability to clinical care. Additionally, the OTP sample reflects the environment, policies, and regulations of New York State, which may limit transferability to other states that have less generous substance use treatment policies. In contrast to other states without governmental support for methadone, New York State Medicaid supports methadone as an OUD-indicated therapy. While we acknowledge that OTPs can have substantial limitations in terms of staff and resources, they are generally less stigmatizing than traditional health care settings, according to people with OUD [2,12,13]. Methadone program staff, however, can be viewed as highly mistrustful, which may interfere with the successful integration of HCV therapy [47]. To build trust between patients and OTP staff, promotion of success stories through a peer pipeline is recommended [47,48]. Some OTP regulations may actually hamper access to methadone and may promote patient stigmatization [49]. As a potential approach to decreasing stigma and improving health outcomes, annual staff training on OUD and OUD-related infectious diseases is recommended. In fact, New York State has annual continuing education requirements for OTP staff and providers. While we did not specifically evaluate reimbursement for services delivered through facilitated telemedicine, administrator interviews provided important insights into the financial aspects of facilitated telemedicine [50]. Administrators also commented on how the expansion of telemedicine during the COVID-19 pandemic enhanced reimbursement for telemedicine services performed at OTPs.

Conclusions

Participants were highly enthusiastic about facilitated telemedicine for HCV care integrated into OTPs. They identified several practices to promote integrated care, including seamless workflows, high-level collaboration, and blurring of lines between disciplines that portended successful HCV care integration. Participants noted the importance of facilitated telemedicine by familiar staff and identified practices to improve its quality and value. Thus, facilitated telemedicine is a high-value application of telemedicine that bridges treatment protocols to human interactions and provides supports that are essential for quality care [51]. The staff participants perceived profound HCV cure-related changes that motivated subsequent improvements in overall health and well-being. Thus, facilitated telemedicine could be promoted to expand access, improve quality, and add value to health care delivery to underserved populations, including people with OUD.

Acknowledgments

We thank the staff of the participating opioid treatment programs for sharing their views. This work was supported by a Patient-Centered Outcomes Research Institute Award (IHS-1507-31640) and partially supported by the Troup Fund of the Kaleida Health Foundation (AHT). The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of PCORI, its board of governors, or the Methodology Committee. Study funders were not involved in data collection, analysis, or manuscript preparation. This work was presented at the American Telemedicine Association 2023 Annual Conference and Expo in San Antonio, Texas on March 4-6, 2023.

Data Availability

The data transcripts generated and analyzed during the study are not publicly available due to confidentiality concerns but are available from the corresponding author on reasonable request.

Authors' Contributions

SSD was responsible for conceptualization, data collection, methodology, resources, software, supervision, formal analysis, investigation, visualization, writing-original draft, writing-review, and editing. SJG was responsible for data curation, formal analysis, writing–original draft, writing–review, and editing. AV was responsible for conceptualization, methodology, software, formal analysis, investigation, visualization, writing–original draft, writing-review, and editing. AD was responsible for data curation, formal analysis, visualization, writing–original draft, writing–review, and editing. AHT was responsible for conceptualization, data curation, funding acquisition, investigation, project administration, resources, supervision, validation, visualization, writing-original draft, writing-review, and editing.

Conflicts of Interest

AHT has received research funds for his institution from Gilead Sciences, Merck, and Abbvie Inc. He has also received honoraria from Gilead Sciences and Abbvie Inc. None of the other authors declare any conflicts of interest.

Multimedia Appendix 1
Consolidated Criteria for Reporting Qualitative Studies (COREQ checklist).

XSL•FO
RenderX

[PDF File (Adobe PDF File), 227 KB - jmir_v26i1e53049_app1.pdf]

Multimedia Appendix 2
Qualitative methodology supplement and interview guides for staff and administrators.
[PDF File (Adobe PDF File), 209 KB - jmir_v26i1e53049_app2.pdf]

References

1. Bornstein S. The challenges of behavioral health integration: the persistence of the mind-body problem. Ann Intern Med 2020;173(2):151-152. [doi: 10.7326/M20-2887] [Medline: 32479171]
2. Talal AH, Jaanimägi U, Dharia A, Dickerson SS. Facilitated telemedicine for hepatitis C virus: Addressing challenges for improving health and life for people with opioid use disorder. Health Expect 2023 Dec;26(6):2594-2607 [FREE Full text] [doi: 10.1111/hex.13854] [Medline: 37641398]
3. Ghany MG, Morgan TR, AASLD-IDSA Hepatitis C Guidance Panel. Hepatitis C guidance 2019 update: American Association for the Study of Liver Diseases-infectious diseases society of America recommendations for testing, managing, and treating Hepatitis C virus infection. Hepatology 2020;71(2):686-721 [FREE Full text] [doi: 10.1002/hep.31060] [Medline: 31816111]
4. Fleurence RL, Collins FS. A national Hepatitis C elimination program in the United States: a historic opportunity. JAMA 2023;329(15):1251-1252. [doi: 10.1001/jama.2023.3692] [Medline: 36892976]
5. Combating hepatitis B and C to reach elimination by 2030: advocacy brief. World Health Organization. URL: https://apps.who.int/iris/handle/10665/206453 [accessed 2021-07-03]
6. Viral hepatitis national strategic plan for the United States: a roadmap to elimination (2021-2025). U.S. Department of Health and Human Services. URL: https://www.hhs.gov/sites/default/files/Viral-Hepatitis-National-Strategic-Plan-2021-2025.pdf [accessed 2021-01-19]
7. Talal AH, Markatou M, Sofikitou EM, Brown LS, Perumalswami P, Dinani A, et al. Patient-centered HCV care via telemedicine for individuals on medication for opioid use disorder: Telemedicine for Evaluation, Adherence and Medication for Hepatitis C (TEAM-C). Contemp Clin Trials 2022;112:106632 [FREE Full text] [doi: 10.1016/j.cct.2021.106632] [Medline: 34813962]
8. Comparing ways to provide hepatitis C treatment for people who take methadone. Patient-Centered Outcomes Research Institute. URL: https://tinyurl.com/bdhesdjw [accessed 2022-11-22]
9. Simon C, Vincent L, Coulter A, Salazar Z, Voyles N, Roberts L, et al. The Methadone Manifesto: treatment experiences and policy recommendations From methadone patient activists. Am J Public Health 2022 Apr;112(S2):S117-S122. [doi: 10.2105/AJPH.2021.306665] [Medline: 35349312]
10. New York State Office of Addiction Services and Supports. Person-centered care guidance for OASAS certified programs. URL: https://oasas.ny.gov/system/files/documents/2020/01/oasasperson-centeredcareguidance.pdf [accessed 2021-09-09]
11. Talal AH, Sofikitou EM, Wang K, Dickerson S, Jaanimägi U, Markatou M. High satisfaction with patient-centered telemedicine for hepatitis C virus delivered to substance users: a mixed-methods study. Telemed J E Health 2023;29(3):395-407 [FREE Full text] [doi: 10.1089/tmj.2022.0189] [Medline: 35925809]
12. Islam MM. Missed opportunities for hepatitis C testing and other opportunistic health care. Am J Public Health 2013;103(12):e6 [FREE Full text] [doi: 10.2105/AJPH.2013.301611] [Medline: 24134357]
13. Earnshaw V, Smith L, Copenhaver M. Drug addiction stigma in the context of methadone maintenance therapy: an investigation into understudied sources of stigma. Int J Ment Health Addict 2013;11(1):110-122 [FREE Full text] [doi: 10.1007/s11469-012-9402-5] [Medline: 23956702]
14. Talal AH, Sofikitou EM, Jaanimägi U, Zeremski M, Tobin JN, Markatou M. A framework for patient-centered telemedicine: application and lessons learned from vulnerable populations. J Biomed Inform 2020;112:103622 [FREE Full text] [doi: 10.1016/j.jbi.2020.103622] [Medline: 33186707]
15. Heath B, Wise Romero P, Reynolds KA. CIHS standard framework for levels of integrated healthcare. SAMHSA-HRSA Center for Integrated Health Solutions. URL: https://www.thenationalcouncil.org/wp-content/uploads/2020/01/CIHS_Framework_Final_charts.pdf?daf=375ateTbd56 [accessed 2021-01-03]
16. Talal AH, Markatou M, Liu A, Perumalswami PV, Dinani AM, Tobin JN, et al. Integrated hepatitis C-opioid use disorder care through facilitated telemedicine: a randomized trial. JAMA 2024 Apr 23;331(16):1369-1378. [doi: 10.1001/jama.2024.2452] [Medline: 38568601]
17. Heidegger M. Being and Time. New York, NY: Robinson J. Harper and Row; 1962.
18. Smythe EA, Ironside PM, Sims SL, Swenson MM, Spence DG. Doing heideggerian hermeneutic research: a discussion paper. Int J Nurs Stud 2008;45(9):1389-1397. [doi: 10.1016/j.ijnurstu.2007.09.005] [Medline: 17950738]
19. Heidegger M. The Question Concerning Technology. New York, NY: Harper & Collins; 1982.
20. Dibley L, Dickerson S, Duffy M, Vandermause R. Doing Hermeneutic Phenomenological Research: A Practical Guide. London, UK: SAGE Publications Ltd; 2020.

XSL•FO
RenderX

21. Talal AH, Jaanimägi U, Davis K, Bailey J, Bauer BM, Dharia A, et al. Facilitating engagement of persons with opioid use disorder in treatment for hepatitis C virus infection via telemedicine: stories of onsite case managers. J Subst Abuse Treat 2021;127:108421 [FREE Full text] [doi: 10.1016/j.jsat.2021.108421] [Medline: 34134875]
22. de Witt L, Ploeg J. Critical appraisal of rigour in interpretive phenomenological nursing research. J Adv Nurs 2006;55(2):215-229. [doi: 10.1111/j.1365-2648.2006.03898.x] [Medline: 16866813]
23. Ironside PM. Hermeneutics. In: Fitzpatrick JJ, Kazer MW, editors. Encyclopedia of Nursing Research, 4th Edition. New York, NY: Springer Publishing Company; 2017.
24. Smith J. Beyond the divide between cognition and discourse: Using interpretative phenomenological analysis in health psychology. Psychology & Health 1996 Feb;11(2):261-271. [doi: 10.1080/08870449608400256]
25. Brocki J, Wearden A. A critical evaluation of the use of interpretative phenomenological analysis (IPA) in health psychology. Psychology & Health 2007 Feb 01;21(1):87-108. [doi: 10.1080/14768320500230185]
26. Smith JA. Reflecting on the development of interpretative phenomenological analysis and its contribution to qualitative research in psychology. Qual Res Psychol 2004;1:39-54.
27. Charmaz K. Constructing Grounded Theory. London: SAGE; 2014.
28. Eveleigh RM, Muskens E, van Ravesteijn H, van Dijk I, van Rijswijk E, Lucassen P. An overview of 19 instruments assessing the doctor-patient relationship: different models or concepts are used. J Clin Epidemiol 2012;65(1):10-15 [FREE Full text] [doi: 10.1016/j.jclinepi.2011.05.011] [Medline: 22118265]
29. Malâtre-Lansac A, Engel CC, Xenakis L, Carlasare L, Blake K, Vargo C, et al. Factors influencing physician practices' adoption of behavioral health integration in the United States: a qualitative study. Ann Intern Med 2020;173(2):92-99. [doi: 10.7326/M20-0132] [Medline: 32479169]
30. Rose S, Hurwitz HM, Mercer MB, Hizlan S, Gali K, Yu PC, et al. Patient experience in virtual visits hinges on technology and the patient-clinician relationship: a large survey study with open-ended questions. J Med Internet Res 2021;23(6):e18488 [FREE Full text] [doi: 10.2196/18488] [Medline: 34152276]
31. Srinivasan M, Asch S, Vilendrer S, Thomas SC, Bajra R, Barman L, et al. Qualitative assessment of rapid system transformation to primary care video visits at an academic medical center. Ann Intern Med 2020;173(7):527-535 [FREE Full text] [doi: 10.7326/M20-1814] [Medline: 32628536]
32. Karasz A, Merchant K, Arnsten J, Feinberg J, Kim AY, Lum PJ, et al. A video-observed treatment strategy to improve adherence to treatment among persons who inject drugs infected with hepatitis C virus: qualitative study of stakeholder perceptions and experiences. J Med Internet Res 2023;25:e38176 [FREE Full text] [doi: 10.2196/38176] [Medline: 37266986]
33. Dimova RB, Zeremski M, Jacobson IM, Hagan H, Des Jarlais DC, Talal AH. Determinants of hepatitis C virus treatment completion and efficacy in drug users assessed by meta-analysis. Clin Infect Dis 2013;56(6):806-816 [FREE Full text] [doi: 10.1093/cid/cis1007] [Medline: 23223596]
34. Herzer KR, Pronovost PJ. Ensuring quality in the era of virtual care. JAMA 2021;325(5):429-430. [doi: 10.1001/jama.2020.24955] [Medline: 33528544]
35. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press; 2001.
36. Albritton J, Ortiz A, Wines R, Booth G, DiBello M, Brown S, et al. Video teleconferencing for disease prevention, diagnosis, and treatment : a rapid review. Ann Intern Med 2022;175(2):256-266. [doi: 10.7326/M21-3511] [Medline: 34871056]
37. Rodriguez NJ, Okwara NC, Shen L, Jajoo K, Chan WW. Impact of telemedicine modalities on equitable access to ambulatory gastroenterology care. Gastroenterology 2021;161(3):742-747.e3 [FREE Full text] [doi: 10.1053/j.gastro.2021.05.042] [Medline: 34051240]
38. Losikoff P, Bosse JD, Martin SA, Wilson A, Chiodo LM. Integrated hepatitis C treatment is associated with improved retention and success in outpatient treatment for opioid use disorder at a private clinic. Front Psychiatry 2022;13:932306 [FREE Full text] [doi: 10.3389/fpsyt.2022.932306] [Medline: 36186876]
39. Springer SA, Del Rio C. Co-located opioid use disorder and hepatitis C virus treatment is not only right, but it is also the smart thing to do as it improves outcomes!. Clin Infect Dis 2020;71(7):1723-1725 [FREE Full text] [doi: 10.1093/cid/ciaa111] [Medline: 32011653]
40. Rosenthal ES, Silk R, Mathur P, Gross C, Eyasu R, Nussdorf L, et al. Concurrent initiation of hepatitis C and opioid use disorder treatment in people who inject drugs. Clin Infect Dis 2020;71(7):1715-1722 [FREE Full text] [doi: 10.1093/cid/ciaa105] [Medline: 32009165]
41. Artenie AA, Cunningham EB, Dore GJ, Conway B, Dalgard O, Powis J, et al. Patterns of drug and alcohol use and injection equipment sharing among people with recent injecting drug use or receiving opioid agonist treatment during and following hepatitis C virus treatment with direct-acting antiviral therapies: an international study. Clin Infect Dis 2020;70(11):2369-2376 [FREE Full text] [doi: 10.1093/cid/ciz633] [Medline: 31300820]
42. Treloar C, Rance J, Grebely J, Dore GJ. Client and staff experiences of a co-located service for hepatitis C care in opioid substitution treatment settings in New South Wales, Australia. Drug Alcohol Depend 2013;133(2):529-534. [doi: 10.1016/j.drugalcdep.2013.07.023] [Medline: 23932843]

XSL•FO
RenderX

43. Cunningham EB, Wheeler A, Hajarizadeh B, French CE, Roche R, Marshall AD, et al. Interventions to enhance testing, linkage to care, and treatment initiation for hepatitis C virus infection: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022;7(5):426-445. [doi: 10.1016/S2468-1253(21)00471-4] [Medline: 35303490]
44. Totten AM, Hansen RN, Wagner J, Stillman L, Ivlev I, Davis-O'Reilly C, et al. Telehealth for Acute and Chronic Care Consultations. Rockville, MD: Agency for Healthcare Research and Quality (US); 2019.
45. Totten AM, Womack DM, Eden KB, McDonagh MS, Griffin JC, Grusing S, et al. Telehealth: Mapping the Evidence for Patient Outcomes from Systematic Reviews. Rockville, MD: Agency for Healthcare Research and Quality (US); 2016.
46. Caffery LJ, Bradford NK, Smith AC, Langbecker D. How telehealth facilitates the provision of culturally appropriate healthcare for Indigenous Australians. J Telemed Telecare 2018;24(10):676-682 [FREE Full text] [doi: 10.1177/1357633X18795764] [Medline: 30343658]
47. Treloar C, Rance J, ETHOS Study Group. How to build trustworthy hepatitis C services in an opioid treatment clinic? A qualitative study of clients and health workers in a co-located setting. Int J Drug Policy 2014;25(5):865-870 [FREE Full text] [doi: 10.1016/j.drugpo.2014.01.011] [Medline: 24559604]
48. Talal AH, George SJ, Talal LA, Dharia A, Ventuneac A, Baciewicz G, et al. Engaging people who use drugs in clinical research: integrating facilitated telemedicine for HCV into substance use treatment. Res Involv Engagem 2023;9(1):63 [FREE Full text] [doi: 10.1186/s40900-023-00474-x] [Medline: 37533127]
49. Overview of opioid treatment program regulations by state. Pew Charitable Trust. URL: https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2022/09/overview-of-opioid-treatment-program-regulations-by-state [accessed 2024-01-21]
50. Ventuneac A, Dickerson SS, Dharia A, George SJ, Talal AH. Scaling and sustaining facilitated telemedicine to expand treatment access among underserved populations: a qualitative study. Telemed J E Health 2023;29(12):1862-1869. [doi: 10.1089/tmj.2022.0534] [Medline: 37252770]
51. Mehrotra A, Bhatia RS, Snoswell CL. Paying for telemedicine after the pandemic. JAMA 2021;325(5):431-432 [FREE Full text] [doi: 10.1001/jama.2020.25706] [Medline: 33528545]

Abbreviations

HCV: hepatitis C virus
OTP: opioid treatment program
OUD: opioid use disorder

Edited by T de Azevedo Cardoso; submitted 15.10.23; peer-reviewed by A Fronczek, L Taylor; comments to author 09.01.24; revised version received 28.01.24; accepted 29.04.24; published 12.06.24.

Please cite as:
Dickerson SS, George SJ, Ventuneac A, Dharia A, Talal AH
Care Integration for Hepatitis C Virus Treatment Through Facilitated Telemedicine Within Opioid Treatment Programs: Qualitative Study
J Med Internet Res 2024;26:e53049
URL: https://www.jmir.org/2024/1/e53049
doi:10.2196/53049
PMID:38865703

©Suzanne S Dickerson, Saliyah J George, Ana Ventuneac, Arpan Dharia, Andrew H Talal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Antibiotic Prescribing by Digital Health Care Providers as Compared to Traditional Primary Health Care Providers: Cohort Study Using Register Data

Andy Wallman[1], PharmD, PhD; Kurt Svärdsudd[2], MD, PhD; Kent Bobits[3], RN; Thorne Wallman[2,4], MD, PhD

[1]Department of Medical and Translational Biology, Umeå University, Umeå, Sweden

[2]Family Medicine and Preventive Medicine Section, Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden

[3]Region Sörmland County Council, Eskilstuna, Sweden

[4]Centre for Clinical Research Sörmland, Uppsala University, Eskilstuna, Sweden

Corresponding Author:
Andy Wallman, PharmD, PhD
Department of Medical and Translational Biology
Umeå University
Biology Building (House H)
Linnaeus Väg 9
Umeå, 901 87
Sweden
Phone: 46 0705500971
Email: andy.wallman@umu.se

Abstract

Background: "Direct-to-consumer (DTC) telemedicine" is increasing worldwide and changing the map of primary health care (PHC). Virtual care has increased in the last decade and with the ongoing COVID-19 pandemic, patients' use of online care has increased even further. In Sweden, online consultations are a part of government-supported health care today, and there are several digital care providers on the Swedish market, which makes it possible to get in touch with a doctor within a few minutes. The fast expansion of this market has raised questions about the quality of primary care provided only in an online setting without any physical appointments. Antibiotic prescribing is a common treatment in PHC.

Objective: This study aimed to compare antibiotic prescribing between digital PHC providers (internet-PHC) and traditional physical PHC providers (physical-PHC) and to determine whether prescriptions for specific diagnoses differed between internet-PHC and physical-PHC appointments, adjusted for the effects of attained age at the time of appointment, gender, and time relative to the COVID-19 pandemic.

Methods: Antibiotic prescribing data based on Anatomical Therapeutic Chemical (ATC) codes were obtained for Region Sörmland residents from January 2020 until March 2021 from the Regional Administrative Office. In total, 160,238 appointments for 68,332 Sörmland residents were included (124,398 physical-PHC and 35,840 internet-PHC appointments). Prescriptions issued by internet-PHC or physical-PHC physicians were considered. Information on the appointment date, staff category serving the patient, *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) diagnosis codes, ATC codes of prescribed medicines, and patient-attained age and gender were used.

Results: A total of 160,238 health care appointments were registered, of which 18,433 led to an infection diagnosis. There were large differences in gender and attained age distributions among physical-PHC and internet-PHC appointments. Physical-PHC appointments peaked among patients aged 60-80 years while internet-PHC appointments peaked at 20-30 years of age for both genders. Antibiotics with the ATC codes J01A-J01X were prescribed in 9.3% (11,609/124,398) of physical-PHC appointments as compared with 6.1% (2201/35,840) of internet-PHC appointments. In addition, 61.3% (6412/10,454) of physical-PHC infection appointments resulted in antibiotic prescriptions, as compared with only 25.8% (2057/7979) of internet-PHC appointments. Analyses of the prescribed antibiotics showed that internet-PHC followed regional recommendations for all diagnoses. Physical-PHC also followed the recommendations but used a wider spectrum of antibiotics. The odds ratio of receiving an antibiotic prescription (after adjustments for attained age at the time of appointment, patient gender, and whether the prescription was issued before or during the COVID-19 pandemic) during an internet-PHC appointment was 0.23-0.39 as compared with a physical-PHC appointment.

XSL•FO
RenderX

Conclusions: Internet-PHC appointments resulted in a significantly lower number of antibiotics prescriptions than physical-PHC appointments, adjusted for the large differences in the characteristics of patients who consult internet-PHC and physical-PHC. Internet-PHC prescribers showed appropriate prescribing according to guidelines.

(J Med Internet Res 2024;26:e55228) doi:10.2196/55228

KEYWORDS

telehealth prescribing; physical-primary health care; internet-primary health care; antibiotics; prescription; infectious disease; antibiotic; prescriptions; prescribing; telehealth; health care; traditional; digital; telemedicine; virtual care; Swedish; Sweden; primary care; quality of care; online setting; ePrescription; ePrescriptions; ePrescribing; eHealth; compare; comparison; online consultation; digital care; patient record; patient records; mobile phone

Introduction

"Digital care" is a broad term covering interactions in health care that do not require the care provider and patient to be in the same place at the same time [1]. The concept includes terms such as "telehealth," "direct-to-consumer telemedicine," "patient-initiated, on-demand health care with health care personnel at a distance," and "virtual visits." [2]

Technologies such as telehealth and online video appointments increase patients' access to health care, for example, when transport problems, schedules, or physical disabilities make office appointments difficult. These options may save time at work or home, as well as travel time, and missed and rescheduled assignments [3]. Less travel time and waiting time at primary health care (PHC) centers mean less absence from work, which benefits both patients and their employers, but also increases health care costs for society [4].

Availability, flexibility, simplicity, and convenience attract many patients to online health care [5]. Use of the telehealth market and services also increased globally during the COVID-19 pandemic [6].

Digital care meetings are a growing part of Swedish PHC, and there are several private and public digital care providers in the Swedish market [7]. Digital care can also potentially improve access to care for residents living in the rural areas of Sweden [4,5].

Despite these positive results, prescribing antibiotics through remote consultations might be difficult because of the inability to perform physical examinations and carry out necessary testing. Given the global concern regarding antimicrobial resistance, prescribing patterns for digital care providers compared with traditional PHC are important to investigate [8].

The Strategy Group for Rational Antibiotic Use and Reduced Antibiotic Resistance (Strama) aims to counteract antibiotic resistance in Sweden by providing recommendations for the treatment of common infections [9] and has accordingly suggested quality indicators for internet-PHC prescribing for common infection diagnoses [9].

Studies have suggested that physicians among private digital care providers (internet-PHC) in Sweden are less thorough in terms of following Strama's guidelines and recommendations than physicians practicing in traditional physical PHC (physical-PHC), and therefore prescribe antibiotics more often [10,11]. However, the results of a previous review paper showed insufficient evidence of the impact on antibiotic prescribing in primary care [8]. Studies revealed reduced antibiotic prescriptions for some infectious diagnoses by internet-PHC physicians [12,13], with others showing no differences [12-15], and specific age groups exhibiting higher antibiotic prescription rates [16]. A previous study of Swedish internet-based primary health care (PHC) showed a lower rate of prescriptions for sore throat and respiratory symptoms, with no differences observed for dysuria symptoms [12]. Another study showed that adult patients presenting with sinusitis received a higher rate of guideline-concordant diagnosis and a lower prescription of antibiotics at digital appointments as compared with in-office primary care appointments [13]. However, other studies demonstrated no difference [14,15,17,18], for example, in areas of acute respiratory infection [14] and management of urinary tract infections for most average-risk female patients, confirming that the national guidelines were followed in the majority of cases [17].

An American study showed no difference in prescribing or laboratory and imaging tests in most of the 20 studied diagnoses [18]. A randomized controlled trial demonstrated no difference and similar clinical outcomes for patients with diabetes mellitus having video compared with in-clinic consultations, as well as greater satisfaction among patients and primary care providers who used telemedicine consultations [15]. However, antibiotic prescribing was higher among children aged 0-17 years using digital appointments compared with in-office urgent care and primary care, and concordance with antibiotic prescribing guidelines was lower for digital appointments [16]. This study thus aimed to examine possible differences in antibiotic prescription patterns between internet-PHC and physical-PHC physicians in a sufficiently large study population to provide high statistical power, taking account of the effects of patient-attained age, gender, and time relative to the COVID-19 pandemic.

Methods

Setting

All Swedish residents have a personal identification number, which includes their date of birth and gender, and this is used for identification in all official registers in the country. Moreover, all residents are covered by the National Health Insurance Act, which allows them to see any physician they choose at heavily reduced fees [19].

XSL•FO
RenderX

Drugs may be prescribed only by licensed physicians and by midwives and nurses who meet specific requirements [20]. All medical care in Sweden is organized within regions, geographically corresponding to counties. PHC (public or private), including physical-PHC or internet-PHC, is usually the first point of call for patients because of its accessibility [21].

Internet-PHCs were established in Sweden in 2016, including services allowing patients to access health care using a smartphone or computer [4]. The 4 largest digital health care companies in Region Sörmland have approximately 80% of the collective digital market share in the region. The Swedish digital care system is designed to work as part of the Swedish welfare system, which means that the region reimburses the main costs for health care, prescribed drugs, and also for internet-PHC.

Ethical Considerations

The study was carried out in accordance with the tenets of the Declaration of Helsinki and the Swedish Ethical Review Authority (diary number 2020-04098) approved the study protocol. The previous local ethics committees have been closed down and since the 1990s have been replaced by the Ethical Review Authority, a government agent checking that ethics applications follow the law of research ethics. Data were collected in standard procedures in the Region Sörmland Administrative Office for financial reimbursement to PHC caregivers and quality assurance of PHC provided within the region. The approved study protocol allowed for analyses of these collected data.

Data Sources

Data for all appointments with physical-PHC and the 4 largest internet-PHCs in Region Sörmland from January 1, 2020, until March 31, 2021, irrespective of residency in Sweden, were obtained from the Region Sörmland Administrative Office. Data were collected in standard reports from the caregivers operating in the region. The data set included 3,064,680 appointment records for 1,238,183 patients, including 1,482,931 appointment records for 785,583 Sörmland and non-Sörmland residents consulting physical-PHC or the 4 largest internet-PHCs. Out of this material, the main study population for this report was 160,238 appointments for 68,332 Sörmland residents consulting regional physical-PHC or the 4 largest internet-PHCs. Information on the size of the Region Sörmland population by age and gender was obtained from the Swedish government agency Statistics Sweden. Data were analyzed using the Statistical Analysis System (SAS; version 9.3) [22].

Variables obtained for each appointment included the patient's personal identification number, which provided information on gender and attained age, residential area code, care provider (physical-PHC or internet-PHC), appointment date, diagnosis according to the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) [23], prescriptions according to Anatomical Therapeutic Chemical code [24], and date with respect to the COVID-19 pandemic. The data set was complete with no missing data.

Statistical Analysis

Simple differences between continuous variables were tested with Student *t* test and differences between categorical variables with chi-square tests. The effects of care provider (physical-PHC or internet-PHC) on antibiotic treatments for 6 diagnoses according to Strama [25], adjusted for the effects of attained age at the time of appointment, gender, and pandemic period, were analyzed by descending logistic regression analysis (probability modeled as the presence of antibiotic treatment). The time in relation to the pandemic period (before or during) was defined as 0 (before March 1, 2020) and 1 (from March 1, 2020, onwards).

Statistical power was computed using the SAS power procedure. The statistical power of the study was >99.9%. The SAS logistic regression procedure provided information on estimates and their 95% CIs, Wald chi-square (test parameter), odds ratios and their 95% CIs, the concordance index C (a measure of explanation similar to the receiver operating characteristic curve), and *P* values. A large number of statistical tests were performed, and Bonferroni adjustment was made to avoid mass significance issues. $P<.005$ was therefore used to indicate statistical significance.

Results

Table 1 shows the entire data set population for Sörmland and non-Sörmland residents according to age and gender. Non-Sörmland patients accounted for the vast majority of internet-PHC appointments.

Table 2 shows the total and mean numbers of appointments among Sörmland residents according to age and gender. Women accounted for more appointments than men in physical-PHC as well as internet-PHC appointments, and overall, there were more physical-PHC than internet-PHC appointments. The average number of appointments per patient was slightly higher among physical-PHC consultations.

Figure 1 shows the gender and attained age distributions among physical-PHC and internet-PHC appointments. The levels of physical-PHC appointments peaked among patients aged 60-80 years for both genders, while internet-PHC appointments peaked at 20-30 years of age for both genders.

Antibiotics with the Anatomical Therapeutic Chemical codes J01A-J01X were prescribed for all purposes during 11,609 of 124,398 (9.3%) physical-PHC appointments (Table 3) as compared with 2201 of 35,840 (6.14%) internet-PHC appointments. The most common class of antibiotics prescribed was J01C for both men and women. Patients were prescribed antibiotics in J01 Groups D, E, and F less frequently during internet-PHC appointments, while patients with physical-PHC appointments were exposed to a broader prescription pattern. Appointments leading to antibiotic prescriptions were more frequent before the COVID-19 pandemic than during the pandemic.

XSL•FO
RenderX

Table 1. Characteristics of the study population (age and gender distribution) in the data set examined over a period of 15 months (from January 1, 2020, to March 31, 2021). Data for all physical-primary health care providers in the Region of Sörmland and all visits to the 4 largest internet-primary health care providers in Sweden.

Characteristics	Sörmland residents			Non-Sörmland residents	
	Sörmland population	Physical-PHC[a]	Internet-PHC[b]	Physical-PHC	Internet-PHC
Age of women (years), n (%)					
0-9	17,123 (11.5)	1384 (5)	1621 (15.8)	67 (6.1)	57,447 (13.4)
10-19	17,536 (11.8)	1654 (5.9)	1383 (13.5)	106 (9.7)	51,282 (12.0)
20-29	15,524 (10.4)	2090 (7.5)	2555 (24.9)	193 (17.7)	114,504 (26.8)
30-39	17,672 (11.9)	2905 (10.4)	1808 (17.6)	160 (14.7)	81,873 (19.2)
40-49	17,640 (11.8)	3365 (12.6)	1319 (12.8)	115 (10.5)	57,027 (13.4)
50-59	18,799 (12.6)	4142 (14.8)	938 (9.1)	155 (14.2)	39,209 (9.2)
60-69	17,245 (11.6)	4298 (15.4)	444 (4.3)	143 (13.1)	17,906 (4.2)
70-79	17,343 (11.6)	5137 (18.4)	185 (1.8)	104 (9.5)	6903 (1.6)
80+	10,226 (6.9)	2932 (10.5)	20 (0.2)	49 (4.5)	1128 (0.3)
Total, n (%)	149,108 (100)	27,907 (55)	10,273 (58.4)	1092 (53.6)	427,279 (59.7)
Age of men (years), n (%)					
0-9	18,226 (12.1)	1498 (6.6)	1679 (22.9)	55 (5.8)	58,049 (20.2)
10-19	18,823 (12.5)	1191 (5.2)	856 (11.7)	69 (7.3)	31,649 (11)
20-29	17,070 (11.4)	1426 (6.3)	1422 (19.4)	133 (14)	59,267 (20.6)
30-39	18,103 (12.9)	1895 (8.3)	1199 (16.4)	138 (14.6)	51,902 (18)
40-49	17,894 (11.9)	2326 (10.2)	930 (12.7)	127 (13.4)	39,050 (13.6)
50-59	19,350 (12.9)	3391 (14.9)	679 (9.3)	124 (13.1)	27,395 (9.5)
60-69	16,916 (11.3)	4146 (18.2)	364 (5)	142 (15)	13,920 (4.8)
70-79	16,421 (10.9)	4705 (20.6)	167 (2.3)	128 (13.5)	5,755 (2)
80+	7418 (5)	2242 (9.8)	36 (0.5)	31 (3.3)	946 (0.3)
Total, n (%)	150,221 (100)	22,820 (45)	7332 (41.6)	947 (46.4)	287,933 (40.3)

[a]Physical-PHC: number of Sörmland and non-Sörmland residents using Sörmland-based PHC.

[b]Internet-PHC: number of Sörmland and non-Sörmland residents using-based PHC.

Table 2. Number of appointments to primary health care among Sörmland residents.

Number of appointments	Physical-PHC[a]	Internet-PHC[b]	*P* value
Women, n (%)	69,176 (75.7)	22,270 (24.4)	<.001
Average, n	2.72	2.53	<.001
IQR	1-3	1-3	__[c]
Men, n (%)	55,222 (80.3)	13,570 (19.7)	<.001
Average, n	2.64	2.21	<.001
IQR	1-3	1-2	—

[a]Physical-PHC: Sörmland-based primary health care.

[b]Internet-PHC: internet-based primary health care.

[c]Not applicable.

XSL•FO
RenderX

Figure 1. Patient age and gender distributions among physical-primary health care and internet-primary health care appointments. PHC: primary health care.

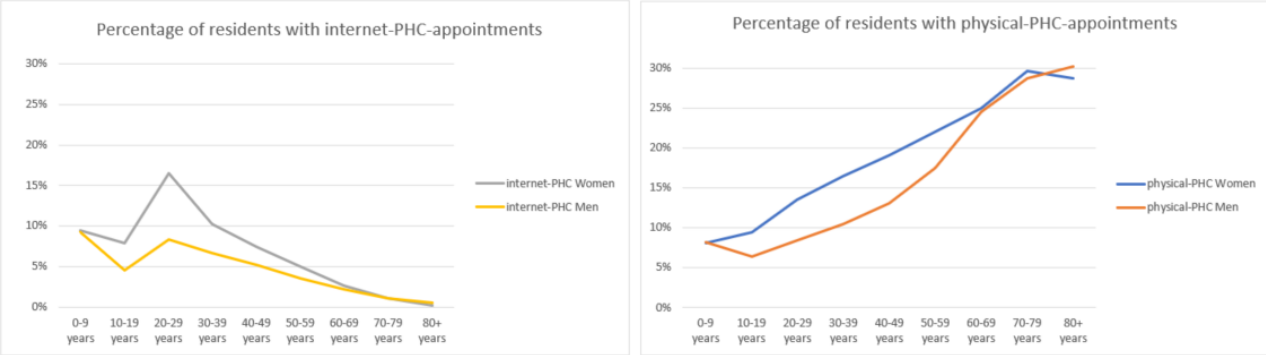

Table 3. Antibiotic prescriptions among Sörmland residents. All gender-specific differences between physical-primary health care and internet-primary health care, as well as those in the period between before and during the COVID-19 pandemic, were significant (P<.001).

Prescriptions	Physical-PHC[a], n (%)		Internet-PHC[b], n (%)	
	Women	Men	Women	Men
Antibiotics (ATC[c])				
J01A[d]	671 (1)	467 (0.9)	215 (1)	94 (0.7)
J01C[e]	4804 (6.9)	3211 (5.8)	1008 (4.5)	258 (1.9)
J01D[f]	43 (0.06)	29 (0.05)	0 (0)	3 (0.02)
J01E[g]	130 (0.2)	39 (0.07)	5 (0.02)	0 (0)
J01F[h]	339 (0.05)	248 (0.5)	30 (0.1)	20 (0.2)
J01M[i]	183 (0.3)	300 (0.5)	4 (0.02)	3 (0.02)
J01X[j]	1080 (1.6)	213 (0.4)	562 (2.5)	6 (0.04)
COVID-19 pandemic				
Before the pandemic	1338 (11.5)	912 (9.7)	132 (8.4)	22 (1.7)
During the pandemic	5793 (10.1)	3566 (7.8)	1685 (6.8)	362 (3)
Total	7131 (10.3)	4478 (8.1)	1817 (8.2)	384 (2.8)

[a]Physical-PHC: Sörmland primary health care.
[b]Internet-PHC: internet-based primary health care.
[c]ATC: Anatomical Therapeutic Chemical.
[d]J01A: tetracyclines.
[e]J01C: beta-lactamases.
[f]J01D: other beta-lactamases.
[g]J01E: sulfonamides.
[h]J01F: macrolides.
[i]J01M: quinolones.
[j]J01X: other antibacterial agents.

The total number of appointments with Strama-defined infection diagnoses was 18,433/160,238 (11.5%; Table 4). Physical-PHC appointments led to an infection diagnosis in 8.4% of cases compared with 22.3% of internet-PHC appointments. Appointments resulting in infection diagnoses were more frequent for internet-PHC (7979/35,840, 22.3%) compared with physical-PHC (10,454/124,398, 8.4%), and more frequent before the COVID-19 pandemic for both providers.

XSL•FO
RenderX

Table 4. Number and proportions of appointments for Sörmland residents with infectious diseases according to Strama antibiotic guidelines in Sweden. All gender-specific differences between physical-primary health care and internet-primary health as well as the period between before and during the COVID-19 pandemic were significant (*P*<.001).

Infections	Physical-PHC[a], n (%)			Internet-PHC[b], n (%)		
	Women	Men	Total	Women	Men	Total
Acne	305 (0.44)	170 (0.31)	475 (0.38)	839 (3.77)	342 (2.52)	1181 (3.3)
Acute bronchitis	291 (0.42)	207 (0.37)	498 (0.4)	56 (0.25)	31 (0.23)	87 (0.24)
Acute cystitis	1869 (2.7)	419 (0.76)	2288 (1.84)	1451 (6.52)	21 (0.15)	1472 (4.11)
Acute otitis media	402 (0.58)	402 (0.73)	804 (0.65)	13 (0.06)	14 (0.1)	27 (0.08)
Acute rhinosinusitis	299 (0.43)	126 (0.23)	425 (0.34)	212 (0.95)	82 (0.6)	294 (0.82)
Lyme borreliosis	235 (0.34)	180 (0.35)	415 (0.33)	116 (0.52)	110 (0.81)	226 (0.63)
Erysipelas	139 (0.2)	170 (0.31)	309 (0.25)	12 (0.05)	12 (0.09)	24 (0.07)
Pharyngotonsillitis	432 (0.62)	264 (0.48)	696 (0.56)	469 (2.11)	216 (1.59)	685 (1.91)
Cough	727 (1.05)	617 (1.12)	1344 (1.08)	456 (2.05)	366 (2.7)	822 (2.29)
Impetigo	86 (0.12)	73 (0.13)	159 (0.13)	177 (0.79)	150 (1.11)	327 (0.91)
Carbuncle, furuncle, etc	128 (0.19)	142 (0.26)	270 (0.22)	35 (0.16)	32 (0.24)	67 (0.19)
Chlamydia infection[c]	3 (0)	7 (0.01)	10 (0.01)	26 (0.12)	26 (0.19)	52 (0.15)
Genital mycoplasma[c]	2 (0)	1 (0)	3 (0)	2 (0.01)	8 (0.06)	10 (0.03)
Ingrown nail infection	67 (0.1)	77 (0.14)	144 (0.12)	39 (0.18)	45 (0.33)	84 (0.23)
Unspecified skin infection	399 (0.58)	373 (0.68)	772 (0.62)	212 (0.95)	146 (1.08)	358 (1)
Pneumonia	152 (0.22)	116 (0.21)	268 (0.22)	2 (0.01)	3 (0.02)	5 (0.01)
Upper respiratory tract infection	931 (1.35)	643 (1.16)	1574 (1.27)	1419 (6.37)	969 (7.14)	2388 (6.66)
COVID-19 pandemic						
Before COVID-19 pandemic	1565 (13.48)	1064 (11.32)	2629 (12.51)	558 (25.82)	290 (22.41)	848 (25.54)
During COVID-19 pandemic	4902 (8.51)	2923 (6.38)	7825 (7.57)	4893 (24.33)	2238 (18.23)	7131 (22.02)
Total	6467 (9.35)	3987 (7.22)	10,454 (8.4)	5451 (24.48)	2,528 (18.63)	7979 (22.26)

[a]Physical-PHC: Sörmland primary health care.

[b]Internet-PHC: internet-based primary health care.

[c]Patients not treated in PHC in Region Sörmland.

A total of 6412 (61.3%/10,454) physical-PHC appointments for patients with Strama-defined infection diagnoses resulted in antibiotic prescriptions, as compared with 2057 (of 7979, 25.8%) for internet-PHC appointments (Table 5). For all diagnoses, antibiotic prescriptions were lower for internet-PHC appointments. Internet-PHC prescribing was approximately at Strama-recommended levels (<5%) for bronchitis, cough, and upper respiratory tract infections (no Strama goals were available for cystitis, chlamydia, and mycoplasma infections), but higher than recommended for other diagnoses (8.3%-37% of diagnoses).

Analyses of the prescribed antibiotics showed that internet-PHC followed regional recommendations for all diagnoses (Multimedia Appendix 1), while physical-PHC also followed the recommendations, but used a wider spectrum of antibiotics.

Logistic regression analyses (Table S1 in Multimedia Appendix 2) of medication in relation to disease showed that prescriptions of antibiotics for internet-PHC were still lower (OR 0.23-0.39) than those for physical-PHC for the 6 main diagnoses leading to antibiotic prescriptions, even after adjustments for attained age at the time of appointment, patient sex, and whether the prescription was issued before or during the COVID-19 pandemic. J01C prescribing for acute cystitis showed the lowest OR of 0.23, and the highest OR was shown for J01A prescribing for acne diagnosis.

XSL•FO
RenderX

Table 5. Antibiotic prescriptions in Anatomical Therapeutic Chemical groups J01A-J01X for infectious diseases in relation to the number of patients with the disease among Sörmland residents. All differences between physical-primary health care and internet-primary health care were significant (*P*<.001).

Infections	Treatment with any of J01A-J01X	
	Regional-based PHC[a]	Internet-based PHC
Acne	185 (39)	212 (18)
Acute bronchitis	146 (29.3)	2 (2.3)
Acute cystitis	2235 (97.7)	1209 (82.1)
Acute otitis media	727 (90.4)	10 (37)
Acute rhinosinusitis	274 (64.5)	29 (9.9)
Lyme borreliosis	408 (98.3)	194 (85.8)
Erysipelas	289 (93.5)	11 (45.8)
Pharyngotonsillitis	573 (82.3)	214 (31.2)
Cough	76 (5.7)	0 (0)
Impetigo	96 (60.4)	63 (19.3)
Carbuncle, furuncle, etc	222 (82.2)	9 (13.4)
Chlamydia infection[b]	10 (100)	15 (28.9)
Genital mycoplasma[b]	3 (100)	8 (80)
Ingrown nail infection	120 (83.3)	7 (8.3)
Unspecified skin infection	653 (84.6)	59 (16.5)
Pneumonia	239 (89.2)	1 (20)
Upper respiratory tract infection	156 (9.9)	14 (0.6)
Total	6412 (61.3)	2057 (25.8)

[a]PHC: primary health care.

[b]Patients not treated in PHC in Region Sörmland.

Discussion

Main Findings

The results of this study showed that the patients differ in age and gender between internet-PHC and physical-PHC. Infection diagnoses were more common in internet-PHC (22.3%, 7979/35,840 of appointments) than physical-PHC (8.4%, 10,454/124,398) but antibiotics were prescribed in less degree for infection diagnoses appointments, internet-PHC, 25.8% (2057/7979) versus 61.3% (6412/10,454) for physical-PHC appointments. On average, internet-PHC providers prescribed antibiotics for 6 infection diagnoses 70% less often than physical-PHC providers (OR 0.23-0.39), after taking account of the effects of differences in gender and age distributions between providers, and the period in relation to the COVID-19 pandemic. The results were similar for all 6 diagnoses tested in multivariate analysis.

Study Strengths and Limitations

The strengths of this study included the use of data that were representative of a defined geographical area, thus minimizing the risk of selection bias regarding the included population. Moreover, the study population was large enough to avoid problems associated with low statistical power. Because the outcome variables were dichotomous, logistic regression was used to obtain the main results. The study limitations included the lack of access to medical record data. However, there was no reason to believe that the registration of diagnoses and treatments would be biased to an extent that would affect the results.

Prescribing of antibiotics in Sweden decreased by more than a quarter over the past 10 years [26], and strategies and goals to decrease antibiotic use are also set globally to decrease antimicrobial resistance.

These results indicated that infection diagnoses were much more frequent in internet-PHC appointments (22.3% [7979/35,840] of all appointments) compared with physical-PHC (8.4% [10,454/124,398]). This might be because these diagnoses are more common among younger people. Most patients chose internet-PHC for acne, impetigo, acute cystitis, and upper respiratory diseases. This might be an age effect, but these are also diagnoses that can be made by medical history or video observations. The high number of upper respiratory disease diagnoses may also be related to the COVID-19 pandemic.

Considering the effect of the pandemic, the degree of infection diagnoses by physical-PHC dropped from 12.7% to 7.6%, compared with 25.5% to 22% for internet-PHC. Hygiene recommendations probably decreased the transmission of common infections and physical appointments [27], with

XSL•FO
RenderX

patients with an infection diagnosis shifting from physical-PHC before the pandemic (2629 physical- vs 848 internet-based) to internet-PHC during the pandemic (7825 physical vs 7131 internet appointments).

This study demonstrated that internet-PHC providers prescribed fewer antibiotics than physical-PHC providers during the study period, in accordance with earlier, smaller studies in Sweden [12], but in contrast to reports in the Swedish daily press [10]. The higher and broader pattern of antibiotic prescriptions in physical-PHC is probably related to the broader range of diagnoses and possibly to laboratory test results and more complicated or multiple diseases [25].

Antibiotic prescribing levels for defined infectious diseases were significantly lower for internet-PHC for all 17 Strama [25] defined infections studied. Various respiratory infections were rarely treated with antibiotics by internet-PHC, following Strama's recommendations (<5%), despite the fact that the antibiotic prescription limit is not always possible or likely to be met [25].

Infection diagnoses often need to be confirmed by laboratory tests before prescribing, leading to less prescribing in internet-PHC. Diagnoses for which antibiotics were frequently prescribed included acute cystitis (82.1% of diagnosed) and Lyme borreliosis (85.5%), which was in line with Strama recommendations [25] for internet-PHCs. Patients with these diagnoses were prescribed antibiotics more frequently in physical-PHC.

Uncomplicated acute cystitis can be diagnosed only with an online appointment and treated with the first-line antibiotics nitrofurantoin and pivmecillinam, without laboratory testing, using internet-PHC. This appeared to be the case for other infections as well as with a higher degree of prescribing for visually diagnosable infections, such as skin infections and acne. Patients may also initiate contact with internet-PHC providers to change the administration formula after earlier contact with a physical-PHC provider. Further studies are therefore needed to establish the timing for those who make the first diagnosis and when antibiotics are prescribed to determine if the recommended Strama <5% level is adequate.

Internet appointments have several advantages, of which the most important is quick access. Internet appointments save time for the patients, are comfortable, suitable, and seem to be sufficient for making simpler diagnoses and issuing prescriptions. Internet-PHC appointments increased during the COVID-19 pandemic to avoid exposure to other patients in waiting rooms or in-person consultations. In Sweden, prescribing is almost totally digitalized, making internet-PHC prescribing easier, with patients receiving their prescriptions online instantly.

In addition to the above advantages of internet-PHC, it may have disadvantages, including leading to more appointments for symptoms that would not lead to a physical-PHC appointment. Younger patients and women tend to use internet-PHC more often due to a higher degree of digital literacy, whereas the technical threshold may exclude elderly and disabled patients. Further studies should determine if an increase in internet-PHC use releases physical-PHC resources for these groups. Internet-based PHC may fill a gap in the health care system and physical-PHC and internet-PHC may complement each other.

Conclusions

This study demonstrated substantial differences in the characteristics of patients who consult internet-PHC and physical-PHC providers in Sweden. The proportion of patients who chose internet-PHC increased during the COVID-19 pandemic, but the rate of infection diagnoses decreased in internet-PHC. The probability of a patient with an infection diagnosis receiving antibiotics after an internet-PHC visit was reduced compared with physical-PHC, after adjusting for attained age at the time of appointment, gender, and the pandemic. Internet physicians thus prescribe fewer antibiotics than physicians in physical consultations and show appropriate prescribing according to the recommendations.

Acknowledgments

This study was supported by the Centre for Clinical Research Sörmland, Uppsala University and Umeå University. We thank Susan Furness, PhD, from Edanz's website for editing a draft of this paper.

Data Availability

The data sets generated during and/or analyzed during this study are not publicly available due to ethical approval restraints regarding patients' personal health data but are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the writing–review and editing. AW and TW wrote the original draft and contributed to the conceptualization. AW contributed to the methodology. KB and KS performed formal analysis and data curation. KB and TW contributed to the resources. AW performed the visualization.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Prescriptions for infectious diseases in relation to the number of patients with the disease among Sörmland residents.
[DOCX File , 25 KB - jmir_v26i1e55228_app1.docx]

Multimedia Appendix 2
Logistic regression analyses of antibiotics for the most common diagnoses in relation to disease, regional- versus internet-based primary health care, age at the time of appointment, and prescription before or during the COVID-19 pandemic.
[DOCX File , 19 KB - jmir_v26i1e55228_app2.docx]

References

1. Becevic M, Sheets LR, Wallach E, McEowen A, Bass A, Mutrux ER, et al. Telehealth and telemedicine in Missouri. Mo Med 2020;117(3):228-234 [FREE Full text] [Medline: 32636555]
2. McGrail KM, Ahuja MA, Leaver CA. Virtual visits and patient-centered care: results of a patient survey and observational study. J Med Internet Res 2017;19(5):e177 [FREE Full text] [doi: 10.2196/jmir.7374] [Medline: 28550006]
3. Donelan K, Barreto EA, Sossong S, Michael C, Estrada JJ, Cohen AB, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019;25(1):40-44 [FREE Full text] [Medline: 30667610]
4. Blix M, Jeansson J. Telemedicine and the Welfare State: The Swedish Experience. Sweden: Research Institute of Industrial Economics; 2018. URL: https://www.ifn.se/media/3adc1ks3/wp1238.pdf [accessed 2024-05-28]
5. Blix M, Jeansson J. Digital transformation and public services: societal impacts in Sweden and beyond. In: Larsson A, Teigland R, editors. Telemedicine and the Welfare State: The Swedish Experience. New York, NY: Routledge; 2019:15-32.
6. Hoffman DA. Increasing access to care: telehealth during COVID-19. J Law Biosci 2020;7(1):lsaa043 [FREE Full text] [doi: 10.1093/jlb/lsaa043] [Medline: 32843985]
7. Ekman B. Cost analysis of a digital health care model in Sweden. Pharmacoecon Open 2018;2(3):347-354 [FREE Full text] [doi: 10.1007/s41669-017-0059-7] [Medline: 29623633]
8. Han SM, Greenfield G, Majeed A, Hayhoe B. Impact of remote consultations on antibiotic prescribing in primary health care: systematic review. J Med Internet Res 2020;22(11):e23482 [FREE Full text] [doi: 10.2196/23482] [Medline: 33031045]
9. Samverkan mot antibiotikaresisten (the Swedish strategic programme against antibiotic resistance). URL: https://strama.se/; [accessed 2023-05-11]
10. Cederberg J. Nätläkare förskriver antibiotika på »dåliga grunder« (Netdoctors prescribing antibiotics on poor foundations).: Läkartidningen; 2017 May 17. URL: https://tinyurl.com/mrxce979 [accessed 2024-06-05]
11. Persson E. Granskning: Nätläkare ställer felaktiga diagnoser. 2017. URL: https://www.dagensarena.se/innehall/granskning-natlakare-staller-felaktiga-diagnoser/ [accessed 2023-05-15]
12. Entezarjou A, Calling S, Bhattacharyya T, Milos Nymberg V, Vigren L, Labaf A, et al. Antibiotic prescription rates after eVisits versus office visits in primary care: observational study. JMIR Med Inform 2021;9(3):e25473 [FREE Full text] [doi: 10.2196/25473] [Medline: 33720032]
13. Johnson KM, Dumkow LE, Burns KW, Yee MA, Egwuatu NE. Comparison of diagnosis and prescribing practices between virtual visits and office visits for adults diagnosed with sinusitis within a primary care network. Open Forum Infect Dis 2019;6(9):ofz393 [FREE Full text] [doi: 10.1093/ofid/ofz393] [Medline: 31660415]
14. Yao P, Clark S, Gogia K, Hafeez B, Hsu H, Greenwald P. Antibiotic prescribing practices: is there a difference between patients seen by telemedicine versus those seen in-person? Telemed J E Health 2020;26(1):107-109. [doi: 10.1089/tmj.2018.0250] [Medline: 30762493]
15. Sood A, Watts SA, Johnson JK, Hirth S, Aron DC. Telemedicine consultation for patients with diabetes mellitus: a cluster randomised controlled trial. J Telemed Telecare 2018;24(6):385-391. [doi: 10.1177/1357633X17704346] [Medline: 28406066]
16. Yao P, Clark S, Gogia K, Hafeez B, Hsu H, Greenwald P. Antibiotic prescribing practices: is there a difference between patients seen by telemedicine versus those seen in-person? Telemed J E Health 2020;26(1):107-109. [doi: 10.1089/tmj.2018.0250] [Medline: 30762493]
17. Rastogi R, Martinez KA, Gupta N, Rood M, Rothberg MB. Management of urinary tract infections in direct to consumer telemedicine. J Gen Intern Med 2020;35(3):643-648 [FREE Full text] [doi: 10.1007/s11606-019-05415-7] [Medline: 31667749]
18. Norden JG, Wang JX, Desai SA, Cheung L. Utilizing a novel unified healthcare model to compare practice patterns between telemedicine and in-person visits. Digit Health 2020;6:2055207620958528 [FREE Full text] [doi: 10.1177/2055207620958528] [Medline: 32995039]
19. Health Care Act. 2017. URL: http://rkrattsbaser.gov.se/sfst?bet=2017:30; [accessed 2022-05-15]
20. Läkemedelsverkets föreskrifter (HSLF-FS 2021:75) om förordnande och utlämnande av läkemedel och teknisk sprit. 2021. URL: https://www.lakemedelsverket.se/sv/lagar-och-regler/foreskrifter/2021-75-konsoliderad [accessed 2022-05-15]
21. Socialstyrelsen, About the Swedish healthcare system. 2020. URL: https://www.socialstyrelsen.se/en/about-us/healthcare-for-visitors-to-sweden/about-the-swedish-healthcare-system/ [accessed 2024-06-10]

XSL•FO
RenderX

22. Statistical Analysis System. Durham, NC: SAS Institute; 2024. URL: https://www.sas.com/en_in/software/stat.html [accessed 2024-05-29]
23. ICD-10 Version:2019.: World Health Organization URL: https://www.who.int/publications/m/item/international-statistical-classification-of-diseases-and-related-health-problems---volume-2 [accessed 2022-05-15]
24. Anatomical Therapeutic Chemical (ATC) Classification.: World Health Organization; 2023. URL: https://www.who.int/tools/atc-ddd-toolkit/atc-classification [accessed 2024-06-10]
25. Strama. Rekommendationer för kvalitetsindikatorer vid digitala vårdmöten. 2022. URL: https://strama.se/wp-content/uploads/2022/04/Kvalitetsindikatorer-for-digitala-vardmoten-2022.pdf [accessed 2022-05-15]
26. Swedish Strategy to Combat Antibiotic Resistance 2020-2023. 2020. URL: https://strama.se/wp-content/uploads/2023/06/amr_strategi_eng_web.pdf [accessed 2024-06-10]
27. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]

Abbreviations

ATC: Anatomical Therapeutic Chemical
DTC: direct-to-consumer
ICD-10: International Statistical Classification of Diseases, Tenth Revision
PHC: primary health care
SAS: Statistical Analysis System
Strama: Strategy Group for Rational Antibiotic Use and Reduced Antibiotic Resistance

Edited by A Mavragani; submitted 07.12.23; peer-reviewed by H De Loof, S Pesälä, E van der Velde; comments to author 28.02.24; revised version received 08.03.24; accepted 09.05.24; published 26.06.24.

Please cite as:
Wallman A, Svärdsudd K, Bobits K, Wallman T
Antibiotic Prescribing by Digital Health Care Providers as Compared to Traditional Primary Health Care Providers: Cohort Study Using Register Data
J Med Internet Res 2024;26:e55228
URL: https://www.jmir.org/2024/1/e55228
doi:10.2196/55228
PMID:

©Andy Wallman, Kurt Svärdsudd, Kent Bobits, Thorne Wallman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Evaluation of Telemedicine Consultations Using Health Outcomes and User Attitudes and Experiences: Scoping Review

Pranavsingh Dhunnoo[1,2], MD, MSc; Bridie Kemp[3], BSc, PhD; Karen McGuigan[3], PhD; Bertalan Meskó[2], MD, PhD; Vicky O'Rourke[4], EdD; Michael McCann[1], PhD

[1]Department of Computing, Atlantic Technological University, Letterkenny, Ireland
[2]The Medical Futurist Institute, Budapest, Hungary
[3]School of Nursing and Midwifery, Queen's University Belfast, Belfast, United Kingdom
[4]Faculty of Business, Atlantic Technological University, Letterkenny, Ireland

Corresponding Author:
Pranavsingh Dhunnoo, MD, MSc
Department of Computing
Atlantic Technological University
CoLab Building, ATU
Port Road
Letterkenny, F92 FC93
Ireland
Phone: 353 74 918 6000
Email: pranavsingh.dhunnoo@research.atu.ie

Abstract

Background: Despite a recent rise in adoption, telemedicine consultations retention remains challenging, and aspects around the associated experiences and outcomes remain unclear. The need to further investigate these aspects was a motivating factor for conducting this scoping review.

Objective: With a focus on synchronous telemedicine consultations between patients with nonmalignant chronic illnesses and health care professionals (HCPs), this scoping review aimed to gain insights into (1) the available evidence on telemedicine consultations to improve health outcomes for patients, (2) the associated behaviors and attitudes of patients and HCPs, and (3) how supplemental technology can assist in remote consultations.

Methods: PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guided the scoping review process. Inclusion criteria were (1) involving adults with nonmalignant, noncommunicable chronic conditions as the study population; (2) focusing on health outcomes and experiences of and attitudes toward synchronous telemedicine consultations between patients and HCPs; and (3) conducting empirical research. A search strategy was applied to PubMed (including MEDLINE), CINAHL Complete, APA PsycNet, Web of Science, IEEE, and ACM Digital. Screening of articles and data extraction from included articles were performed in parallel and independently by 2 researchers, who corroborated their findings and resolved any conflicts.

Results: Overall, 4167 unique articles were identified from the databases searched. Following multilayer filtration, 19 (0.46%) studies fulfilled the inclusion criteria for data extraction. They investigated 6 nonmalignant chronic conditions, namely chronic obstructive pulmonary disease, diabetes, chronic kidney disease, ulcerative colitis, hypertension, and congestive heart failure, and the telemedicine consultation modality varied in each case. Most observed positive health outcomes for patients with chronic conditions using telemedicine consultations. Patients generally favored the modality's convenience, but concerns were highlighted around cost, practical logistics, and thoroughness of clinical examinations. The majority of HCPs were also in favor of the technology, but a minority experienced reduced job satisfaction. Supplemental technological assistance was identified in relation to technical considerations, improved remote workflow, and training in remote care use.

Conclusions: For patients with noncommunicable chronic conditions, telemedicine consultations are generally associated with positive health outcomes that are either directly or indirectly related to their ailment, but sustained improvements remain unclear. These modalities also indicate the potential to empower such patients to better manage their condition. HCPs and patients tend to be satisfied with remote care experience, and most are receptive to the modality as an option. Assistance from supplemental technologies mostly resides in addressing technical issues, and additional modules could be integrated to address challenges

XSL•FO
RenderX

relevant to patients and HCPs. However, positive outcomes and attitudes toward the modality might not apply to all cases, indicating that telemedicine consultations are more appropriate as options rather than replacements of in-person visits.

(J Med Internet Res 2024;26:e53266) doi:10.2196/53266

KEYWORDS

telemedicine; internet-based consultation; chronic illnesses; cyberpsychology; digital health; scoping review

Introduction

Background

While the underpinning technology has been available for several decades [1], telemedicine consultations experienced a surge in adoption following the COVID-19 pandemic [2]. In particular, for patients with chronic conditions, these modes of care lessen travel-associated pain as well as facilitate the management and treatment of their disease [3,4]. Such technology-mediated consultations are also favored by physicians treating patients with chronic conditions, as they can improve productivity, patient's health, and management [4]. Continued interest in telemedicine consultation from researchers, investors, and policy makers indicates that such modalities are being actively considered as viable options within health care systems beyond the pandemic [5-9].

However, persisting issues have been raised regarding aspects such as privacy, reliability, safety, and accessibility when using remote consultation means [10-13]. Due to difficulties in identifying nonverbal cues through such means, Kilvert et al [14] recommend that the first meeting between the health care professional (HCP) and the patient be conducted face to face rather than through internet-based means, suggesting that the latter approach is not totally apt to fully replace physical meetings. The scoping review by Leone et al [15] also identified discrepancies in telehealth consultation guidelines, which indicate that the modality is still nascent and evolving.

Recent studies also reflect a need to better understand telemedicine consultation adoption by patients with chronic conditions [16,17]. Notably, a 2021 analysis by McKinsey & Company identified a stabilized use of telehealth following an initial peak early in the pandemic as telehealth was borne out of necessity [18]. Even if the use of the modality is higher than prepandemic levels, only approximately 40% of responders in the United States would be inclined to use telehealth after the pandemic subsides. However, 40% to 60% would be interested in internet-based health care with more options, such as a *digital front door* or a lower-cost virtual-first health plan. While this analysis might be more representative of the US market, it nevertheless indicates the potential for supplemental technologies to telehealth for addressing flailing retention rates and barriers to adoption.

Such challenges to maintaining and encouraging remote care use highlight the importance of gaining a deeper understanding of the technical and psychological elements, that is, cyberpsychological elements, involved in telemedicine consultations, as well as indicate potentials for exploring how novel technological approaches can effectively tackle them [4]. Incorporating supplementary technologies presents an opportunity for innovation and for transforming the adoption of web-based modalities for patients with chronic conditions. For instance, the study by Yuan et al [19] on the remote care outcomes of patients with heart failure during the pandemic found that telephone consultations were associated with increased mortality. By supplementing this modality with other technologies, there is a possibility of achieving better outcomes. Researchers further consider the metaverse and enabling technologies such as extended reality as potential approaches to supplement remote care [20]. These approaches represent potential novel avenues to consider in telemedicine consultation research.

While recent reviews have analyzed the state of telemedicine consultations from the perspective of either those living with chronic conditions [21-23] or HCPs [24], they did not investigate extensively, within the same review, perspectives of both HCPs and patients with chronic conditions and their associated behaviors or potentials for technological improvements during synchronous telemedicine consultations. Furthermore, a preliminary search for existing scoping reviews and systematic reviews performed in April 2023 on Google Scholar and Open Science Framework did not identify such investigations. Researchers have recently accentuated the need to better understand the preferences and concerns of remote health users [25]. They have also recommended further exploration of technological assistance to reduce limitations experienced during telemedicine consultations [4]. These reflect the conceptual framework developed by Hensel et al [26] that highlights the importance of investigating aspects of cyberpsychological obtrusiveness and the adoption of telehealth [26]. However, such aspects have not been widely investigated in recent reviews and indicate the need for new research in this area.

Aims and Objectives of This Scoping Review

In addition to the potentials of telemedicine consultations, the limited availability of a synthesis of the literature in this field was a motivating factor for this scoping review, which is the first step in an on-going PhD project in Ireland at the Atlantic Technological University [27].

The project aims to develop a novel, evidence-based artificial intelligence assistive tool to supplement real-time, synchronous telemedicine consultations between patients with nonmalignant chronic illnesses and their HCPs. Considering the higher prevalence of certain chronic conditions in Ireland such as diabetes mellitus, cardiovascular conditions, and chronic obstructive pulmonary disease (COPD) [28], particular emphasis is paid on these nonmalignant conditions. Therefore, the nascent aspect of telemedicine consultations [15] and the exploratory nature of this research were deemed appropriate for the adoption of a scoping review design [29,30] that aims to investigate

XSL•FO
RenderX

synchronous telemedicine consultations for nonmalignant chronic illnesses through the following research questions: (1) What is the available evidence on synchronous telemedicine consultations to improve health outcomes for patients with chronic conditions? (2) What are the associated behaviors and attitudes of patients and HCPs during synchronous telemedicine consultations? (3) How can supplemental technology assist in remote consultations?

Methods

This scoping review was conducted by following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines and the methodology developed by the Joanna Briggs Institute [30,31]. The protocol for this review was also registered on Open Science Framework on April 28, 2023 [32].

Rationale for Inclusion Criteria

The inclusion criteria were based on the Population, Concept, and Context framework, which is recommended to design relevant objectives and eligibility criteria for scoping reviews, where each component of the framework (population, concept, and context) guides the identification of the respective areas of interest of the review [33]. This framework also assisted in designing the search terms to ensure relevance to the review's aims. Textbox 1 summarizes the inclusion and exclusion criteria of this study based on the Population, Concept, and Context framework.

Textbox 1. Inclusion and exclusion criteria based on Population, Concept, and Context framework for study selection.

Inclusion criteria

- Population
  - Adults with nonmalignant, noncommunicable chronic conditions
- Concept
  - Health outcomes, experiences and attitudes toward synchronous internet-based consultations between patients and health care professionals
  - Aspects around attitudes, experiences, engagement, behaviors, and intentions during such interactions
- Context
  - Clinical settings and beyond
  - Research centers

Exclusion criteria

- Population
  - Patients with no or any other condition
  - <18 years of age
  - Malignant or communicable conditions
- Concept
  - Traditional or face-to-face or non–internet-based or undefined means of consultations
  - Interactions with non–health care professionals
  - Asynchronous internet-based care
- Context
  - Nonempirical research (eg, reviews and editorials)

Upon agreement among the authors, and with the input from a specialist librarian, the following databases and repositories were selected given their relevance to provide published as well as gray literature around the research questions: PubMed (which includes MEDLINE), CINAHL Complete, APA PsycNet, Web of Science, IEEE, and ACM Digital.

Adopted Search Strategy

A search strategy was devised by the research team to identify sources of evidence relevant to the research aims. Terms such as "metaverse" and "mixed reality" were included given their potentials in remote care [20] and to maximize the potential to identify publications that involve novel telemedicine consultation approaches, considering the exploratory nature of this review. The search terms also included specific chronic conditions such as "diabetes mellitus" and "COPD" given their prevalence in Ireland, where subsequent stages of a project involving this scoping review will take place, and to thus provide more specific results [28].

The selected databases were searched from their inception till March 6, 2023, with no language or article type filters applied;

XSL•FO
RenderX

however, for CINAHL Complete via EBSCOhost, the “Apply Equivalent Subjects” and “Suggest Subject Terms” filters were selected. The following combination of search terms was used: (“virtual consultation” OR “remote consultation” OR “telemedicine” OR “telehealth” OR “metaverse” OR “virtual reality” OR “augmented reality” OR “mixed reality” OR “extended reality”) AND (“chronic conditions” OR “chronic illnesses” OR “chronic disease” OR “diabetes mellitus” OR “chronic respiratory illnesses” OR “cardiovascular conditions” OR “obesity” OR “COPD”) AND (“attitudes” OR “experiences” OR “engagement” OR “behaviours” OR “intentions” OR “motivations” OR “psychology” OR “barriers”). Medical Subject Headings were used in CINAHL Complete, which involved the database’s recommended terms. Multimedia Appendix 1 details the latter search.

Multilayer Filtering of Search Results

The search results were screened independently by 2 researchers (PD and BK) via the web-based referencing software Rayyan (Rayyan Systems Inc) against the selection criteria (Textbox 1). Articles were excluded if (1) their full texts were not in English, (2) they did not involve real-time consultation with an HCP [34], (3) the (internet-based) consultation modality was unclear, (4) no consultation was involved, and (5) they were not empirical research; thus, articles such as reviews and editorials were excluded. Citation tracking was performed on publications identified from the databases whose full texts were eligible for data extraction. More specifically, backward search of their respective reference lists and forward search of articles that cite back the individual study were performed, as these are recommended practices for added value to health-related reviews [35]. After their respective screenings, the reviewers discussed and resolved any conflicts over their selection. This was done by reverting to the selection criteria and consulting with KM. The results of the screening process are summarized in Figure 1 [36], which is based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram. HCP: health care professional [36].

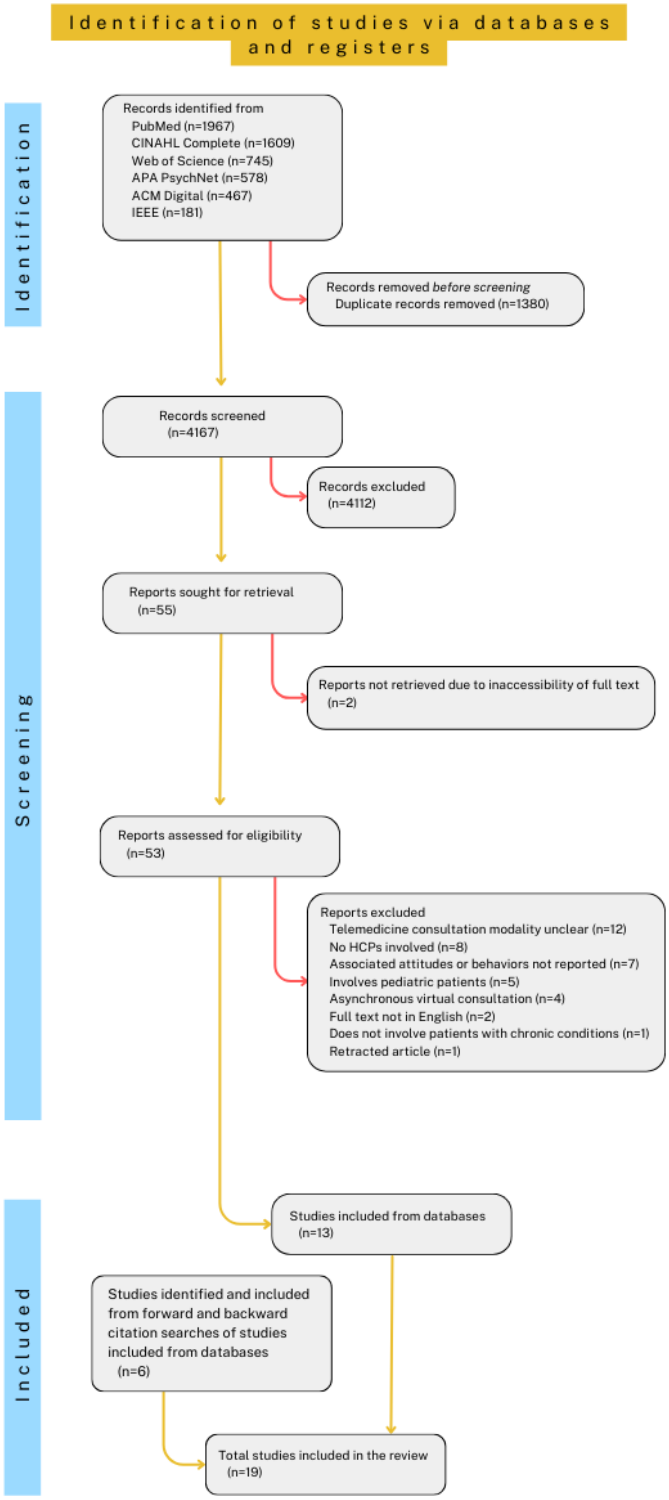

Data relevant to this scoping review's aims were independently extracted from the final list of included sources of evidence and summarized in a data-charting form by PD and KM, who resolved any conflicts after comparing their findings. This was done by reverting to the individual studies. The data-charting form was designed by the research team with items relevant to this review. Data extracted pertained to the study characteristics, including authors, year of publication, and country of origin, study design, research aims, chronic conditions involved, internet-based consultation modality used, health outcome reported by the studies, related attitudes and behaviors toward the internet-based consultation modality, and potentials for technological improvement identified in the articles.

For the synthesis of results, the included studies were grouped based on the chronic condition involved. Thereafter, data on the outcomes, type of internet-based consultation used, associated behaviors, and potentials for technological improvements were summarized. As a scoping review design

XSL•FO
RenderX

was adopted for this investigation to provide an overview of current evidence relevant to the research questions, no critical appraisal of the included sources was performed, which is typical for such studies [31,37].

Results

Overview

Searching the databases yielded a total of 5547 articles, out of which 4167 (75.12%) unique articles were identified after the removal of duplicates. Screening of texts and abstracts against the inclusion criteria resulted in 55 (1.32%) texts for full-text screening. Of these 55 articles, 13 (24%) were found to fulfill the inclusion criteria and were selected for further data extraction and analysis. To maximize the identification of publications relevant to this review, we performed further backward and forward citation searching from these 13 articles. Such citation tracking after initial study selection is recommended practice to enhance the identification of potentially eligible reports [35]. This led to the identification of 27 additional articles for full-text screening against the inclusion criteria. This yielded 6 additional articles for inclusion in data extraction. Thus, in total, 19 studies were included for data extraction after screening 53 full texts from the database searches (of which 13 were included) and 27 full texts from backward and forward searches (of which 6 were included). The data extracted from the 19 included articles are presented in summarized form in Table 1, with more details provided in Multimedia Appendix 2 [38-56].

XSL•FO
RenderX

Table 1. Characteristics of sources of evidence.

Study, year, and country	Study design and aims	Chronic condition and telemedicine consultation modality	Reported health outcomes	Related attitudes and behaviors	Potentials for technological improvement
Mair et al [38], 1999, United Kingdom	Mixed method feasibility pilot of technology assistance for COPD[a] exacerbations at home	COPD; video telemedicine consultation	Not reported	• Users required getting accustomed to the setup	Need to improve image quality
Tudiver et al [39], 2007, United States	Phone survey interview to evaluate HCPs'[b] perceptions of remote diabetes care delivery	Diabetes; video telemedicine consultation	Increased control, confidence, compliance, and motivation in managing diabetes among patients	• Positivity toward telemedicine consultation	Improve communication between HCPs and remote workflow
Whitten and Mickus [40], 2007, United States	Mixed method study to evaluate the use of home telehealth for patients with COPD and CHF[c]	COPD and CHF; video telemedicine consultation	No significant effects	• Satisfied with care delivery and technology	Telehealth may provide added value if the associated costs are reduced, even if the outcomes are similar to those of traditional care
Trief et al [41], 2008, United States	Qualitative study to describe the lived experiences of older adults with diabetes involved in a telemedicine management intervention	Diabetes; video telemedicine consultation	Positive physical health changes and some emotional benefit	• HCPs' encouragement can positively influence participation. • Participation motivated by desire to feel and be healthier	Some frustrations with operating the technology
Nilsson et al [42], 2009, Sweden	Quantitative evaluation of the feasibility and quality of uncomplicated hypertension care in rural areas, with comparison between telemedicine and face-to-face consultations	Primary hypertension; video telemedicine consultation	Patients treated via telemedicine consultation had a higher probability of improving their blood pressure	• Patients: most found telemedicine consultation to be as good as in-person GP[d] meetings. • HCPs: appreciated detailed remote patient evaluation; telemedicine consultation allowed them to work more independently	Equipment not found to be useful for remote lung examination
Cook et al [43], 2010, United States	Quantitative evaluation of remote audio nurse counseling to address cognitive and emotional barriers to medication adherence in ulcerative colitis	Ulcerative colitis; audio telemedicine consultation	Treatment adherence up to 6 months was higher than the expected rate.	• Treatment discontinuation due to negative beliefs about treatment and breakthrough symptoms	Concerns around cost and treatment logistics
Sorknæs et al [44], 2010, Denmark	Mixed method interventional study to investigate the effect of telemedicine video consultations (telemedicine consultation) on early readmissions for patients with COPD in their homes after a hospital discharge	COPD; video telemedicine consultation	10% to 14% reduction in readmission risk for patients using telemedicine	• High satisfaction rate with patients and nurses • Nurses trusted the measurement collected from consultations.	Minimize audiovisual delay

XSL•FO
RenderX

Study, year, and country	Study design and aims	Chronic condition and telemedicine consultation modality	Reported health outcomes	Related attitudes and behaviors	Potentials for technological improvement
Campbell et al [45], 2011, Canada	Mixed method feasibility study of remote care provision to rural communities and to assess the level of satisfaction among patients and HCPs	CKD[e]; video telemedicine consultation	Patients experienced less stress during remote visits than during in-person visits	• High level of satisfaction • The ease of access to care and time and cost savings were appreciated.	Improve the flow of paperwork and access to patient information
Berkhof et al [46], 2014, the Netherlands	Quantitative pilot study to determine the effects of telemedicine on the health care use and health status of patients with COPD	COPD; audio telemedicine consultation	No improvement of health status among the telemedicine consultation group and worsened score for the telemedicine consultation group	• Use of audio telephone consultations led to little loss to follow-up	Include educational or pulmonary rehabilitation or training component to drive a more successful telemedicine model
Mathar et al [47], 2015, Denmark	Qualitative interviews investigating the telemedicine consultation experiences and preferences of patients with COPD	COPD; video telemedicine consultation	No special value attributed	• Patients are receptive of the technology and to adopting at-home monitoring equipment	Interference with daily routines, cost considerations, and unwanted obligations
Raymond et al [48], 2016, United States	Pilot study to investigate the feasibility and acceptability of a telemedicine intervention	Diabetes; video telemedicine consultation	Potentials to improve mental health, diabetes education, and condition management	• Satisfactory experience, given improved convenience, ease of access to HCPs' insights, and time efficiency	None reported
Reid et al [49], 2018, United States	Quantitative cohort pilot study to assess feasibility, acceptability, care retention and follow-up rates, patient satisfaction, and adherence to guidelines in a telemedicine intervention	Diabetes; video telemedicine consultation	No significant change in HbA_{1c}[f] values in either the telemedicine consultation or control group	• Higher engagement, adherence, and frequency of clinical visits in the telemedicine consultation group • Higher satisfaction rates in the telemedicine consultation group regarding care experiences and technology	Barrier to access encountered due to institutional internet firewall settings
Lambooy et al [50], 2021, Australia	Quantitative case-controlled longitudinal observational cohort study to assess the feasibility, sustainability, and clinical outcomes of telehealth videoconferencing (telemedicine consultation) for patients with CKD	CKD; video telemedicine consultation	No between- group differences during follow-up for 2 years	• Very high patient satisfaction with the care provided • Telemedicine consultation was comparable to standard care. • Telemedicine consultation uptake reduced over time.	Reduce concerns with technology and remind that telemedicine consultation is an option
Magliah et al [51], 2021, Saudi Arabia	Cross-sectional quantitative analysis of web-based surveys to evaluate patient perception of phone clinics during the COVID-19 pandemic	Diabetes; audio telemedicine consultation	Improved diabetes control and reduced risk of contracting COVID-19	• Patients: satisfied with the modality; most expressed interest in remote audio follow-ups • Physicians: good understanding of patients' condition	Improve internet access

XSL•FO
RenderX

Study, year, and country	Study design and aims	Chronic condition and telemedicine consultation modality	Reported health outcomes	Related attitudes and behaviors	Potentials for technological improvement
Singh et al [52], 2021, United States	Web-based surveys with quantitative analyses to determine the limitations of telehealth accessibility, patient satisfaction with telehealth relative to in-person visits, and the perceived advantages and disadvantages to telehealth	Cardiac condition; any telemedicine consultation modality	None reported	• High level of satisfaction, with time and cost savings seen as advantageous • Clinical examination not perceived as thorough by some patients	Improve internet connectivity, improve the coordination of telehealth appointments, and incorporate remote patient monitoring solutions
Balut et al [53], 2022, United States	Mixed methods study to understand telemedicine use before and after the COVID-19 pandemic and identify relevant barriers and facilitators to its implementation	Cardiology; audio or video visits	None reported	• Patients: racial and ethnic minority groups less likely to use telemedicine; preference of telephone over video visits • Physicians: preferred video to phone visits; willingness to keep using telemedicine consultation after the pandemic	Improve access to video consultations with equipment, video bandwidth, and user-friendliness of the videoconferencing platform
Heyck Lee et al [54], 2022, Canada	Quantitative cross-sectional observational survey to evaluate patients' and physicians' perspectives on the key advantages and disadvantages of telephone consultations in a nephrology outpatient clinic setting during the COVID-19 pandemic	CKD; audio telemedicine consultation	Reduced anxiety related to contracting COVID-19 by traveling to clinic	• Patients: Comfortable with the modality • Low preference for video telemedicine consultation • HCPs: Reduced job satisfaction and sense of connection with patients	Improve translation and auditory feedback to increase patient access to audio consultations
Hitchcock and Heath [55], 2022, United Kingdom	Web-based questionnaire and telephone interviews to understand patients' perspectives on telemedicine consultation and the accessibility of the service and to inform its development	Diabetes; audio or video telemedicine consultation	None reported	• Appreciated at-home convenience	Increase flexibility and access to the service and team
Löhnberg et al [56], 2022, United States	Quantitative program evaluation with phone surveys to report on the transition to a remote modality for a weight management program during the COVID-19 pandemic	Obesity; audio telemedicine consultation	None reported	• About half found telephone visits comparable to in-person visits. • Half indicated preference for remote visits after the pandemic. • Time and cost efficiency of remote care compared to in-person care favored	Video meetings for interdisciplinary team cumbersome due to hardware and software requirements; require adequate training

[a]COPD: chronic obstructive pulmonary disease.

[b]HCP: health care professional.

XSL•FO
RenderX

[c]CHF: congestive heart failure.

[d]GP: general practitioner.

[e]CKD: chronic kidney disease.

[f]HbA_{1c}: hemoglobin A_{1c}.

Characteristics of Sources of Evidence

Details about the 19 included studies' authors, origin, year of publication, aims, design, and participants can be found in Table 1. They were published between 1999 and 2022 [38-56] and conducted in 8 different countries, with 9 (47%) from the United States [39-41,43,48,49,52,53,56], 2 (11%) from the United Kingdom [38,55], 2 (11%) from Denmark [44,47], 2 (11%) from Canada [45,54], 1 (5%) from Sweden [42], 1 (5%) from the Netherlands [46], 1 (5%) from Australia [50], and 1 (5%) from Saudi Arabia [51].

Of the 19 included studies, 9 (47%) adopted a quantitative design [42,43,46,49-52,54,56], 7 (37%) adopted a mixed method design [38-40,44,45,48,53], and 3 (16%) adopted a qualitative design [41,47,55]. As each study had differing aims and scopes, the number of participants varied in each case and averaged at 401 (SD 1211.19).

Among the studies included in this review, 6 nonmalignant chronic conditions were investigated altogether, namely COPD, diabetes, chronic kidney disease, ulcerative colitis, hypertension, and congestive heart failure. However, some were more frequently investigated than others, and some studies involved more than 1 chronic condition. Of the 19 included studies, 6 (32%) involved patients with diabetes [39,41,48,49,51,55], 5 (26%) involved patients with COPD [38,40,44,46,47], 4 (21%) involved patients with chronic cardiovascular conditions (CCCs) [40,42,52,53], 3 (16%) involved patients with chronic kidney disease [45,50,54], 1 (5%) involved patients with ulcerative colitis [43], and 1 (5%) involved patients with obesity [56]. Among the 4 studies that involved CCCs, 1 (25%) involved both patients with COPD and patients with congestive heart failure [40], 1 (25%) involved patients with hypertension [42], and 2 (50%) did not specify the conditions but involved patients with unspecified CCCs [52,53].

The remote consultation modality varied among the 19 included studies, with 12 (63%) using only videoconferencing [38-42,44,45,47-50,55] and 5 (26%) using only telephone visits [43,46,51,54,56]. Furthermore, 1 (5%) study considered all telehealth modalities [52], and 1 (5%) study involved both telephone and video visits [53]. Among the 12 included studies that were published before the COVID-19 pandemic [38-49], the vast majority (10/12, 83%) used video consultations [38-42,44,45,47-49], while only 2 (17%) used only telephone visits [43,46]. In comparison, among the 7 included studies that were published after the pandemic [50-56], the telemedicine consultation modality was more diverse, with 3 (43%) studies involving telephone consultations only [51,54,56], 2 (29%) involving video consultations only [50,55], 1 (14%) involving both video and telephone means [53], and 1 (14%) considering any telemedicine consultation modality [52].

Reported Health Outcomes

The majority of the included studies (14/19, 74%) reported the health outcomes of the patients undertaking internet-based consultations, while 5 (26%) did not provide such data due to different scopes of their research aims [38,52,53,55,56]. Of the reviewed articles that reported on these, most (n=9, 64%) observed positive health outcomes for patients with chronic conditions using telemedicine consultation modalities [39,41-45,48,51,54]. These modes of care delivery were associated with better condition management [39,51], improved physiological parameters [41,42], higher treatment adherence [43], and reduced risk of readmission [44]. Remote visits also positively impacted patients' mental health, with reported reduction in anxiety [45,48,54]. Some articles found no notable improvements or particular contribution to addressing health-related issues when using telemedicine consultations [40,46,47,49,50]. Of the 19 included studies, 1 (5%) even found deteriorating symptoms among patients in the telemedicine group based on the adopted Clinical COPD Questionnaire scale [46].

Diabetes and COPD were the most commonly investigated conditions in the included studies with reported health outcomes. Patients with diabetes generally experienced positive health outcomes relating to their physical and mental health as well as the management of their condition [39,41,48,50], while a minority experienced no changes compared to in-person consultation [49]. In the case of patients with COPD, most of the included studies with reported health outcomes indicated no additional health benefits via telemedicine consultations [40,46,47], while only a minority associated telemedicine consultations with positive health outcomes [44]. However, despite technological improvements over the timeline of the included studies, the reported health outcomes in either diabetes or COPD cases do not seem to be strictly related to the telemedicine consultation technology.

Associated Behaviors and Attitudes of Patients and HCPs

All the included studies reported on the associated behaviors and attitudes of patients and HCPs during their participation in telemedicine consultations. Patients were generally comfortable, receptive, and satisfied with the approach [39,41,49-51], even if it required the adoption of supplemental at-home monitoring equipment [47]. They appreciated the convenience [45,48,55], time [45,52,56], and cost efficiency [45,52,56] that remote care offered, as well as the ease of access to HCPs' insights [47,48]. Some also perceived the technology as a welcome source of information [39,47]. In some studies, patients found remote consultations to be as good as in-person visits [40,42,50,56], with some indicating an interest in or preference for remote visits for follow-up appointments [45,51,53,56]. However, some patients raised concerns about cost, treatment, and practical logistics and that clinical examinations were not thorough during

XSL•FO
RenderX

internet-based visits [38,43,52,55]. HCPs viewed certain aspects of remote care delivery as helpful, such as having access to patient data, specialist's insights, and user-friendliness [39,42]. Despite expressing a need to get better acquainted with the remote modality, HCPs appreciated that they could gain comprehensive insights into their patient's conditions remotely [38,42,44,51]. However, a minority of HCPs reported a reduction in job satisfaction and a diminished sense of connection with their patients [54].

Telemedicine consultations for patients with diabetes in the included studies were met with positive attitudes. Patients with diabetes were more motivated during telemedicine consultations [39,41,49], favored the convenience provided by such means [48,55], and expressed a preference to have follow-up appointments remotely [51]. HCPs appreciated the assistive aspect of telemedicine consultations with patients with diabetes and the insights they could derive remotely [39,51]. While most studies (4/6, 67%) involving patients with diabetes used only video telemedicine consultations [39,41,48,49], 1 (17%) study involved only audio visits [51], and 1 (17%) study included both video and audio visits [55]; the satisfactory experience and positive attitude were shared in all remote modalities among this patient group. Among the included studies that involved COPD cases, patients and HCPs were satisfied with the technology [40,44,47], while a fraction required some familiarization with the setup [38], which is likely reflective of the time of publication (1999). Patients with COPD were receptive to telemedicine consultations [46] and were willing to adopt supplemental technologies for remote monitoring [47]. HCPs further trusted the remote readings [44]. The majority (4/5, 80%) used video consultations [38,40,44,47], and only 1 (20%) study involved audio-only telemedicine consultations [46], but the satisfaction rate was similar in either mode for patients with COPD.

Potentials for Supplemental Technology Assistance

Among the included studies, potentials for supplemental technology assistance during internet-based consultations were inferred in most studies based on the barriers and concerns reported. Most concerns were related to technical considerations such as malfunctioning computers, audiovisual delay, internet firewall settings, video bandwidth, internet access, and connectivity [41,44,49,51-53]. The need to improve communication and coordination between HCPs for a more efficient workflow during remote consultations was also identified [39,52,56]. Minimizing associated paperwork, data load, resource use, and associated expenses could improve adoption among HCPs [39,40,43,45,47]. Improvements were identified to be associated with the assessment and monitoring of patients remotely, which could involve better image quality or supplemental equipment for specific physical examinations [38,42,52]. Some suggestions included incorporating an educational component for patients into the telemedicine modality, improving the user-friendliness of the platform, and improving translation and auditory feedback [46,53,54].

In particular, for patients with diabetes, improvements for video telemedicine consultations pertained to the web-based workflow [39] and functioning of the setup [41,49], while audio telemedicine consultations could be improved with increased internet connectivity and flexibility of the remote care team [51,55]. For patients with COPD, video consultations would benefit from improved audiovisual quality [38,44], reduced associated costs, and less disruptions to daily routines [40,47]. The identified improvements for audio telemedicine consultations for patients with COPD could involve educational and training components for both patients and HCPs to better use the modality [46].

Critical Appraisal

No critical appraisal of the included studies was conducted, as it is generally not conducted in scoping reviews [31].

Discussion

Principal Findings

This scoping review set out to investigate, in relation to synchronous telemedicine consultations for nonmalignant chronic illnesses, (1) the available evidence on telemedicine consultations to improve health outcomes for patients, (2) the associated behaviors and attitudes of patients and HCPs, and (3) how supplemental technology can assist in remote consultations.

Study Characteristics

We identified 19 studies that fulfilled the selection criteria, and these were published between 1999 and 2022. Among these, 12 (63%) were conducted before the COVID-19 pandemic in only 2 continents (Europe and North America), while the remaining 7 (36%) publications included studies conducted in Australia and Saudi Arabia. This potentially indicates that the expansion of interest in remote care-related research to more locations might have been precipitated by the shift to adopting such consultation modalities amid the pandemic [57].

The higher number of identified studies relating to diabetes, cardiovascular conditions, and COPD likely results from these specific terms being included in the search strategy. However, recent systematic reviews investigating remote consultations also identified more studies that relate to these conditions [58,59], which could indicate a higher likelihood of remote care being adopted for consulting patients with such conditions, considering their higher prevalence [60-63].

As the majority of the included studies in this scoping review used video consultations, this modality appears to be the most popular one. This is contrary to certain cases such as in the United Kingdom, where remote audio consultations appear to be preferred [64]. Nevertheless, based on the reviewed studies, there appear to be high satisfaction rate among adopters of video consultations [39,40,44,45,48-50,52] and even a preference for video consultations by some HCPs for the nonverbal cues and visual insights that this modality enables [42,53]. While most of the included studies published before the pandemic used video visits [38-42,44,45,47-49], studies that were conducted after the COVID-19 pandemic adopted more heterogeneous means [50-56], with slightly more studies investigating audio-only telemedicine consultations [51,54,56]. The higher diversity in telemedicine consultation means after the pandemic

XSL•FO
RenderX

could be attributed to the necessity of using remote care options during the health crisis [18,65]. While significant technological improvements have been achieved since 1999 [66], which is the publication year of this review's earliest included study, it appears that there has not been much change in terms of the remote modalities adopted, but their means of access has evolved. More novel means of telemedicine consultation such as extended reality and the metaverse, which researchers have considered as potential supplements to telemedicine consultations [20], have not been identified in the included studies. This potentially indicates their nascent aspect, which has not been extensively investigated and could represent a future research topic.

Reported Health Outcomes for Patients With Chronic Conditions Using Telemedicine Consultations

From the 19 included studies, 14 (74%) reported on the health outcomes of patients using internet-based consultations. The majority experienced positive health outcomes, with improvements directly related to their chronic condition [41,42,44] as well as their mental health [45,48,54]. Improvements were also noticeable regarding treatment control, adherence, and condition management [39,43,51]. Considering these findings from this scoping review, telemedicine consultations indicate the potential to be beneficial not only for patients' physical health but also for aspects that are, to some extent, tangential to their chronic conditions, improvements to which have been associated with better health outcomes [67,68]. This was highlighted by the recent study by Salah et al [69] investigating the psychological impact of the pandemic on patients with chronic conditions among the Egyptian population. While their findings are limited to the Egyptian population, the researchers found increased rates of anxiety, depression, and stress among this demographic, partly associated with the fear of contracting the SARS-CoV-2 virus, owing to which they subsequently did not attend follow-up consultations. Similar effects were also identified in the review by Dubey et al [70] on the psychosocial impact of COVID-19.

Furthermore, studies identified in this scoping review indicated an improvement in mental health as well as adherence; this indicates the potential for remote care models to address no-shows and anxiety related to in-person visits. This corroborates with the review by Kendzerska et al [71], which identified benefits of telemedicine for chronic disease management to include improved treatment compliance and follow-up rates. Thus, remote care could potentially be associated with empowering patients with chronic conditions to better manage their condition and treatment regime while being beneficial for their physical and mental health.

However, these findings might not be applicable to every case, as a minority of studies did not find telemedicine consultations to add any significant contribution and found that they were comparable to traditional in-person modalities in regard to health outcomes, even if participants were receptive or satisfied with the internet-based care experience [40,46,47,49,50]. Of the 19 included studies, 1 (5%) [46] even identified deteriorating symptoms among patients using telemedicine based on the Clinical COPD Questionnaire scale, which is a validated and reliable measure [72-74]. Such findings are in accord with previous research that highlighted that internet-based care does not offer a "one size fits all" approach [75,76]. Even if a minority of the included studies did not find added health value or even identified adverse health outcomes with telemedicine consultations, such modalities could be better considered as potential options tailored for individual patients' needs rather than a replacement of traditional face-to-face visits in every case.

Considering the most investigated conditions in the included studies, diabetes and COPD, the need for a tailored telemedicine consultation approach is further highlighted. For instance, patients with diabetes using telemedicine consultations experienced positive health outcomes in most cases [39,41,48,50]. In comparison, positive health outcomes were only identified in a minority of patients with COPD [44], while most experienced no particular health benefits [40,44,46]. These could be attributed to the varying needs of patients with diabetics and patients with COPD, and telemedicine consultations might provide more health benefits for the former group of patients. However, this represents potentials to enhance telemedicine consultations for patients with COPD using supplemental technology aimed at improving outcomes for this specific group.

Associated Behaviors and Attitudes of Patients and HCPs During Telemedicine Consultations

Each of the 19 included studies provided data relating to the behaviors and attitudes of patients and HCPs during telemedicine consultations. On the basis of these studies, it is notable that even before the pandemic, most patients and HCPs were receptive to and satisfied by technology-mediated consultations [39,41,42,44,45,47-49]. As for the identified studies that were conducted after the pandemic was declared, patients' and HCPs' acceptability toward or satisfaction with remote care was high in every case [50-56]. This is in line with previous studies that found a positive attitude toward telemedicine use among patients and HCPs during the pandemic [77,78]. However, such attitudes might have been influenced by the need to adopt remote care technologies imposed by the pandemic [65], rather than these being an optional alternative. Given that the prepandemic studies identified in this scoping review still reported a general receptiveness to and satisfaction with the modality, telemedicine consultations might be considered as a viable option.

From patients' perspectives, the reported benefits of telemedicine consultations are related to their convenience, time and cost efficiency, ease of access to HCPs' insights, and informational aspects [39,45,47,48,52,55,56]. Barriers to these aspects have been identified as limitations to health care access for patients with chronic conditions [78-81]. The findings of this scoping review indicate that internet-based modalities can potentially address some of those pertinent barriers, especially considering that some patients viewed internet-based visits as comparable to in-person visits [40,42,50,56]. Considering telemedicine consultations for patients with diabetes, the approach, whether audio or video based, was positively received in each case [39,41,48,49,51,55], and acceptance was also high among patients with COPD [40,44,47], although a minority

XSL•FO
RenderX

needed to get accustomed to the video telemedicine consultation setup [38]. This need could be attributed to the outdated equipment used in the study conducted in 1999, and video telemedicine consultation setup has likely improved in recent years.

Nevertheless, such a positive outlook toward telemedicine consultations might not be applicable to every patient, as a minority of the studies included in this review reported that patients were less receptive toward remote care. They expressed negative attitudes toward the modality's cost, meeting scheduling and logistics, and the thoroughness of clinical examinations [38,43,52,55]. One of the included studies found that racial and ethnic minority groups were less likely to use telemedicine [53]. In addition to reinforcing the notion that internet-based care might not be appropriate in every case [75,76], it reflects the observation of recent studies that highlighted limited adoption by and accessibility of telemedicine for minority groups [82,83]. However, these behaviors toward telemedicine consultations were identified in only a minority of the included studies and could indicate avenues for further investigations.

Regarding the included studies that reported on HCPs' relevant attitudes and behaviors, the majority indicated positive attitudes and behaviors toward telemedicine consultations. They appreciated the access to patient data and specialists' insights provided by these modalities and the user-friendliness of the technology [39,42]. They were also satisfied with and trusted the interpretation of patients' conditions remotely [42,44,51], even if some were unfamiliar with the modality [38,51]. In one of the included studies, HCPs further expressed interest in the continued use of telemedicine after the pandemic [53]. In the case of audio or video telemedicine consultations with patients with diabetes, HCPs were receptive toward the technology [39,41,48,49,51,55]. A similar finding was observed among HCPs involved in telemedicine consultations with patients with COPD [38,40,44,46,47]. Recent studies in different countries have also identified positive attitude from HCPs toward remote care [84-86].

However, those same recent studies have identified limitations to telemedicine from the perspectives of HCPs. For instance, the nationwide survey conducted by Ma et al [85] in China found HCPs' concerns with the modality to relate to infrastructure, service process, cost, and popularity of such services. The quantitative study conducted by Andronic et al [84] on HCPs' perception of telemedicine in Romania found limitations to internet-based care associated with the lack of human interaction, dependency on technology, and legislative concerns. The exploratory survey conducted by Ncube et al [86] among HCPs in Botswana identified privacy, associated cost, and required infrastructure and human resource among inhibitors to telemedicine use. While in minority, this scoping review also identified similar behaviors from HCPs [39,54]. They reported mixed perceptions of telemedicine efficiency in regard to the number and length of visits, diminished job satisfaction, and a lowered sense of connection with their patients during telemedicine consultations. This could potentially indicate that some HCPs might have some reservations in regard to remote care use, which could be investigated in future studies to determine factors that could increase satisfaction with telemedicine consultations.

Regarding studies that reported on the HCPs' attitudes toward different remote care modalities, most were inclined toward video, rather than audio-only, consultation for the audiovisual feedback and nonverbal cues such remote care modalities provide [42,53]. In contrast, a minority showed a more positive attitude toward telephone consultations considering its accessibility for older adults or those with physical impairments [54]. While the low number of included studies that reported on this aspect limits the ability to draw a consensus, this finding might indicate that video consultations might be a preferred option, while audio-only visits could be a consideration based on individual needs and preferences [64].

Potentials Identified for Supplementing Telemedicine Consultations With Additional Technological Support

On the basis of the included studies' reported barriers and concerns regarding telemedicine consultation use for patients with noncommunicable chronic conditions, potentials for supplemental technology assistance were inferred. The majority of the included studies reported the need to improve certain technical aspects related to telemedicine consultations. These include malfunctioning computers, audiovisual delay, internet firewall settings, video bandwidth, and internet access and connectivity [41,44,49,51-53]. Considering telemedicine consultations specifically for the most common illnesses investigated, patients with diabetes experienced challenges related to the telemedicine consultation equipment in both video and audio formats [41,49], while audiovisual quality was the major technical issue for patients with COPD using video telemedicine consultation [38,44]. It is somewhat surprising that relatively basic and infrastructural issues were encountered during recent studies when in parallel researchers are encouraging the adoption of more advanced technologies such as the Internet of Things and blockchain in the health care setting [87,88]. This potentially indicates challenges to effectively implementing remote care technologies due to inadequate resources and infrastructure. Such challenges have also been shown to be tied to design aspects, where remote consultation areas in hospitals require a different setting and design from traditional and existing infrastructure [89]. In relation to supplemental technologies providing assistance to address such issues, supplemental technologies providing assistance could be extrapolated from the review by Baker and Stanley [90] of technical aspects of telemedicine technology. The authors suggest that planning for successful telemedicine programs should consider technical aspects and include features such as high-speed and secure internet access, a "telemedicine hub" for conducting remote interactions, patient access software, and access to IT professionals who can address malfunctioning issues.

Some studies included in this review identified the need for better communication and coordination between HCPs [39,52,56]; this could potentially indicate the need to better train and support HCPs in effectively using telehealth. Both patients with diabetes and COPD, the 2 most commonly investigated conditions in the included studies of this review, indicated

XSL•FO
RenderX

challenges with the remote workflow [39,47,55], which potentially indicate a need for better training to effectively use remote care. Researchers have also identified a lack of standard recommendations for telemedicine curricula, as well as a limitation in telehealth education and training among HCPs [91-93]. A potential technological assistance could be including within telemedicine software an HCP training module based on a competency-based, outcome-oriented framework, as recommended by Stovel et al [91]. This could be integrated in tandem with an educational component for patients according to their specific condition as well as for telemedicine use, given that this was a suggestion in one of the included studies in this review [46].

In regard to other suggestions for improvement from the included studies, user-friendliness and improving translation and auditory feedback were highlighted [53,54]. While involving a limited number of studies from those included in this review, these observations might indicate the need for a participatory design approach in designing telemedicine services [94]. The review by Fouquet and Miranda [95] further concluded that telemedicine design flaws could be circumvented by involving stakeholders, from HCPs to patients, in the design and implementation of such services. As such, improvements to factors such as user-friendliness and auditory feedback could be undertaken by an iterative development approach such as that highlighted in the Integrate, Design, Assess, and Share framework [96]. During development cycles, supplemental technologies could be integrated to address the identified needs such as translation [54], which is a real-time feature of the internet-based health care company CirrusMD [97]. Other needs identified in the included studies pertained to reducing associated paperwork and data load, resource use, and expenses [39,40,43,45,47]. These could, to some extent, be addressed by integrating speech recognition software for automated medical reporting, which has been identified to reduce administrative burden and lower associated costs [98].

Supplemental technologies can also extend beyond software to include hardware. This could address the need for better assessment and monitoring of patients remotely, as identified in some of the included studies [38,42,52]. Telemedicine consultations could integrate the use of wearable devices appropriate for patients' conditions with the remote capture of patient data that can be shared with HCPs on the web [99], which the review by de Farias et al [100] found to improve patient care and treatment effectiveness.

To better help visualize the identified potentials for supplemental technology assistance during telemedicine consultations, we summarized and illustrated this scoping review's findings in this regard in Figure 2.

Figure 2. Potentials for technological improvements during internet-based consultations with patients with nonmalignant chronic conditions. (Individual images were generated with the assistance of the generative artificial intelligence tool Dall-E [Open AI].) HCP: health care professional.

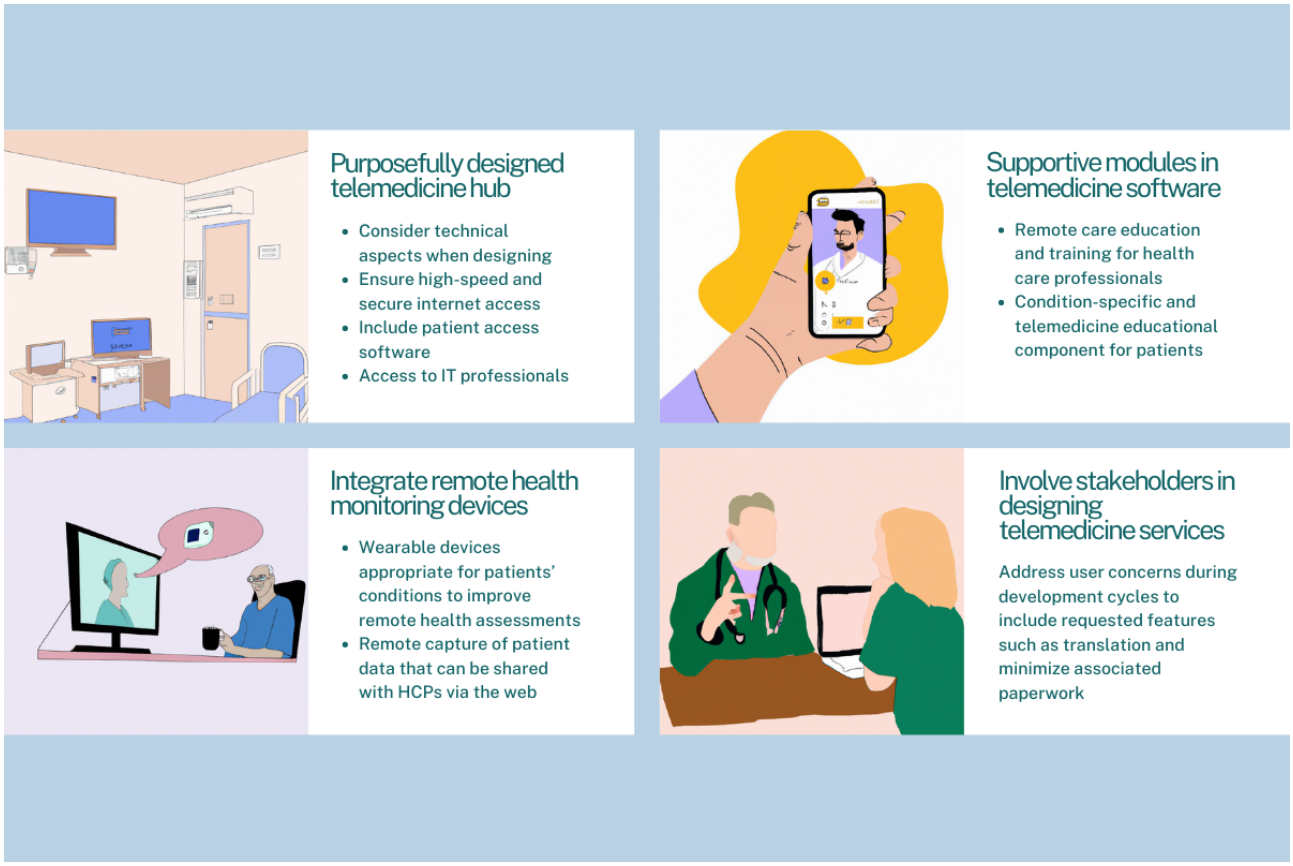

Limitations

There are some limitations to the included studies as well as to the scoping review process that require a measure of caution when generalizing the findings. Regarding limitations of the included sources of evidence, while the average participant size was 401, this number was disproportionately affected by 1 study that included >5000 participants [53], which can influence the

XSL•FO
RenderX

research findings [101]. Some measure of discretion is advised regarding the generalizability of the findings of health outcomes, as only a minority of the reviewed studies included patients through a follow-up period [39,41,43,46] or involved control groups [40,42,44,46,49,50]. Only 1 of the reviewed studies included both follow-ups and control groups [46]. Despite the difference in study design and scope, including follow-ups and control groups are recommended practices in telemedicine research to gain more comprehensive insights into outcomes [102,103]. As a minority of included studies did not specify the conditions involved [52,53], the results need to be interpreted with a layer of caution.

While we followed recommended guidelines for conducting this scoping review, some limitations persist. Most prominently, the number of included studies can limit the generalizability of the findings. This is due to the filtration process, which, for example, excluded studies that did not involve synchronous consultations or studies that did not involve interactions with HCPs, and insights from excluded studies were not considered. However, the inclusion criteria were designed by the research team to adhere to the research aims. In addition, this review considered extracted data and synthesized insights from different chronic conditions and did not consider individual conditions in relation to the review's objectives. Therefore, caution is advised when extrapolating the findings to specific conditions.

Conclusions

This scoping review provided important insights into the health outcomes, attitudes, and potentials for technological improvements in relation to using telemedicine consultations for patients with nonmalignant chronic illnesses. On the basis of the reviewed studies, patients with noncommunicable chronic conditions who use such remote care modalities are mostly patients with COPD or diabetes, although remote means is also used for providing care for people with other chronic conditions. According to our findings, patients with chronic conditions generally experience positive health outcomes that are either directly or indirectly related to their ailment. However, sustained improvements cannot be confidently established due to limited follow-up data, and positive health outcomes might depend on the condition, with patients with diabetes mostly experiencing positive outcomes, while patients with COPD generally did not find added health benefits. Nevertheless, our findings also identified that telemedicine consultations could be considered as an empowering tool for patients with chronic conditions to assist them in better managing their condition and treatment regime.

This review indicates a general receptiveness toward telemedicine consultations based on the positive related attitudes and behaviors when such consultations are an option. Most patients favor the convenience, increased access to care, expert insights, and health education that remote care options provide. HCPs tend to be satisfied with the interpretation of patients' conditions remotely and have an apparent preference for video-based modalities considering the audiovisual feedback and nonverbal cues that they provide.

Potentials for supplemental technological assistance in internet-based consultations have also been identified from the included studies. These mostly relate to technical challenges due to inadequate resources and infrastructure and could be addressed during the planning stages of telemedicine programs. Telemedicine software could also integrate telehealth education and training modules for HCPs and patients to address the respective challenges they experience. It could further be recommended that the design of telemedicine services adopts a participatory aspect with stakeholders.

However, this review highlights that telemedicine consultations might not be the most appropriate option for a minority of cases. In relation to remote care, some patients with nonmalignant chronic conditions might not experience positive health outcomes, might not find the modality accessible, or might not be satisfied by the modality. A portion of HCPs might hold reservations to using telehealth due to a diminished sense of job satisfaction and a reduced sense of connection with their patients during telemedicine consultations. This reinforces the notion that telemedicine consultations might not represent a "one size fits all" approach and should be better considered as an option that can be tailored based on specific needs rather than a complete replacement of in-person visits.

Nevertheless, considering the general receptiveness and positive health outcomes associated with such modalities, telemedicine consultations might be considered as viable options to be recommended to patients with noncommunicable chronic conditions and their HCPs, with the potential to improve their perception of the modality with supplemental technology.

Acknowledgments

The authors would like to acknowledge the assistance of the artificial intelligence tool Dall-E (Open AI) in creating the individual images featured in Figure 2 of this scoping review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search term combination used in CINAHL Complete (via EBSCOhost).
[DOCX File , 7 KB - jmir_v26i1e53266_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Extracted data from included studies.
[DOCX File , 16 KB - jmir_v26i1e53266_app2.docx]

Multimedia Appendix 3
PRISMA-ScR Checklist.
[PDF File (Adobe PDF File), 119 KB - jmir_v26i1e53266_app3.pdf]

References

1. Craig J, Petterson V. Introduction to the practice of telemedicine. J Telemed Telecare 2016 Jun 24;11(1):3-9. [doi: 10.1177/1357633x0501100102]
2. Jiménez-Rodríguez D, Santillán García A, Montoro Robles J, Rodríguez Salvador MD, Muñoz Ronda FJ, Arrogante O. Increase in video consultations during the COVID-19 pandemic: healthcare professionals' perceptions about their implementation and adequate management. Int J Environ Res Public Health 2020 Jul 15;17(14):5112 [FREE Full text] [doi: 10.3390/ijerph17145112] [Medline: 32679848]
3. Li J, Chen G, De Ridder H, Cesar P. Designing a social VR clinic for medical consultations. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI EA '20; April 25-30, 2020; Honolulu, HI p. 1-9 URL: https://dl.acm.org/doi/10.1145/3334480.3382836 [doi: 10.1145/3334480.3382836]
4. Rego N, Pereira HS, Crispim J. Perceptions of patients and physicians on teleconsultation at home for diabetes mellitus: survey study. JMIR Hum Factors 2021 Nov 23;8(4):e27873 [FREE Full text] [doi: 10.2196/27873] [Medline: 34817394]
5. Alsaif SS, Kelly JL, Little S, Pinnock H, Morrell MJ, Polkey MI, et al. Virtual consultations for patients with obstructive sleep apnoea: a systematic review and meta-analysis. Eur Respir Rev 2022 Dec 31;31(166):220180 [FREE Full text] [doi: 10.1183/16000617.0180-2022] [Medline: 36517048]
6. Remote and video consultations: guidance for community pharmacy team. National Health Service. URL: https://www.england.nhs.uk/wp-content/uploads/2022/04/B1400-remote-and-video-consultations-guidance-for-community-pharmacy-teams.pdf [accessed 2024-04-29]
7. Richardson E, Aissat D, Williams GA, Fahy N. Keeping what works: remote consultations during the COVID-19 pandemic. Eurohealth 2020;26(2):73-76 [FREE Full text]
8. Government encouraging the use of eHealth (telehealth). Ministerie van Algemene Zaken. URL: https://www.government.nl/topics/ehealth/government-encouraging-use-of-ehealth [accessed 2024-04-29]
9. Ellery S. New report shows remote mental health consultations make care more accessible but are not the right solution for all patients. Health Innovation Network. 2021. URL: https://healthinnovationnetwork.com/news/new-report-shows-remote-mental-health-consultations-make-care-more-accessible-but-are-not-the-right-solution-for-all-patients/ [accessed 2024-04-29]
10. Baker DC, Bufka LF. Preparing for the telehealth world: navigating legal, regulatory, reimbursement, and ethical issues in an electronic age. Prof Psychol Res Pr 2011 Dec;42(6):405-411. [doi: 10.1037/a0025037]
11. Caffery LJ, Smith AC. Investigating the quality of video consultations performed using fourth generation (4G) mobile telecommunications. J Telemed Telecare 2015 Mar 11;21(6):348-354. [doi: 10.1177/1357633x15577311]
12. van Eck CF. Web-based follow-up after total joint arthroplasty proves to be cost-effective, but is it safe? J Bone Joint Surg 2014;96(22):e192. [doi: 10.2106/jbjs.n.00829]
13. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440 [FREE Full text] [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
14. Kilvert A, Wilmot EG, Davies M, Fox C. Virtual consultations: are we missing anything? Pract Diabetes 2020 Aug 15;37(4):143-146. [doi: 10.1002/pdi.2288]
15. Leone E, Eddison N, Healy A, Royse C, Chockalingam N. Exploration of implementation, financial and technical considerations within allied health professional (AHP) telehealth consultation guidance: a scoping review including UK AHP professional bodies' guidance. BMJ Open 2021 Dec 27;11(12):e055823 [FREE Full text] [doi: 10.1136/bmjopen-2021-055823] [Medline: 34969656]
16. Khan N, Jones D, Grice A, Alderson S, Bradley S, Carder P, et al. A brave new world: the new normal for general practice after the COVID-19 pandemic. BJGP Open 2020 Aug;4(3):bjgpopen20X101103 [FREE Full text] [doi: 10.3399/bjgpopen20X101103] [Medline: 32487520]
17. Verma P, Kerrison R. Patients' and physicians' experiences with remote consultations in primary care during the COVID-19 pandemic: a multi-method rapid review of the literature. BJGP Open 2022 Jun;6(2) [FREE Full text] [doi: 10.3399/BJGPO.2021.0192] [Medline: 35031558]
18. Bestsennyy O, Gilbert G, Harris A, Rost J. Telehealth: a quarter-trilliondollar post-COVID-19 reality? McKinsey & Company. URL: https://www.mckinsey.com/~/media/McKinsey/Industries/Healthcare%20Systems%20and%20Services/Our%20Insights/Telehealth%20A%20quarter%20trillion%20dollar%20post%20COVID%2019%20reality/Telehealth-A-quarter-trilliondollar-post-COVID-19-reality.pdf [accessed 2024-04-29]

XSL•FO
RenderX

19. Yuan N, Botting PG, Elad Y, Miller SJ, Cheng S, Ebinger JE, et al. Practice patterns and patient outcomes after widespread adoption of remote heart failure care. Circ Heart Fail 2021 Oct;14(10):e008573 [FREE Full text] [doi: 10.1161/CIRCHEARTFAILURE.121.008573] [Medline: 34587763]
20. Chengoden R, Victor N, Huynh-The T, Yenduri G, Jhaveri RH, Alazab M, et al. Metaverse for healthcare: a survey on potential applications, challenges and future directions. IEEE Access 2023;11:12765-12795. [doi: 10.1109/access.2023.3241628]
21. van den Bergh R, Bloem B, Meinders MJ, Evers LJ. The state of telemedicine for persons with Parkinson's disease. Curr Opin Neurol 2021 Aug 01;34(4):589-597 [FREE Full text] [doi: 10.1097/WCO.0000000000000953] [Medline: 33990100]
22. Sasseville M, LeBlanc A, Boucher M, Dugas M, Mbemba G, Tchuente J, et al. Digital health interventions for the management of mental health in people with chronic diseases: a rapid review. BMJ Open 2021 Apr 05;11(4):e044437 [FREE Full text] [doi: 10.1136/bmjopen-2020-044437] [Medline: 33820786]
23. Alvarez P, Sianis A, Brown J, Ali A, Briasoulis A. Chronic disease management in heart failure: focus on telemedicine and remote monitoring. Rev Cardiovasc Med 2021 Jun 30;22(2):403-413 [FREE Full text] [doi: 10.31083/j.rcm2202046] [Medline: 34258907]
24. Jonasdottir SK, Thordardottir I, Jonsdottir T. Health professionals' perspective towards challenges and opportunities of telehealth service provision: a scoping review. Int J Med Inform 2022 Nov;167:104862. [doi: 10.1016/j.ijmedinf.2022.104862] [Medline: 36179598]
25. Moulaei K, Sheikhtaheri A, Fatehi F, Shanbehzadeh M, Bahaadinbeigy K. Patients' perspectives and preferences toward telemedicine versus in-person visits: a mixed-methods study on 1226 patients. BMC Med Inform Decis Mak 2023 Nov 15;23(1):261 [FREE Full text] [doi: 10.1186/s12911-023-02348-4] [Medline: 37968639]
26. Hensel BK, Demiris G, Courtney KL. Defining obtrusiveness in home telehealth technologies: a conceptual framework. J Am Med Inform Assoc 2006 Jul 01;13(4):428-431. [doi: 10.1197/jamia.m2026]
27. Dhunnoo P, McGuigan K, O'Rourke V, McCann M. Artificial intelligence enhancement for remote virtual consultations in healthcare provision for patients with chronic conditions. In: Proceedings of the 25th International Conference on Human-Computer Interaction. 2023 Presented at: HCII '23; July 23–28, 2023; Copenhagen, Denmark p. 45-50 URL: https://link.springer.com/chapter/10.1007/978-3-031-35992-7_7 [doi: 10.1007/978-3-031-35992-7_7]
28. Evidence for policies to prevent chronic conditions. University College Cork. URL: https://www.ucc.ie/en/esprit/research/epicc/ [accessed 2024-04-29]
29. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
30. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
32. Dhunnoo P, McCann M, O'Rourke V, McGuigan K, Kemp BJ, Meskó B. Protocol for scoping review: synchronous virtual consultations for chronic illnesses - perceptions of patients and healthcare professionals, related behaviours, and technological assistance. Open Science Framework. URL: https://osf.io/c2564 [accessed 2024-04-29]
33. Pollock D, Peters MD, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 01;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
34. Transforming and scaling up health professionals' education and training. World Health Organization. 2013. URL: https://www.who.int/publications/i/item/transforming-and-scaling-up-health-professionals%E2%80%99-education-and-training [accessed 2024-04-29]
35. Hirt J, Nordhausen T, Appenzeller-Herzog C, Ewald H. Citation tracking for systematic literature searching: a scoping review. Res Synth Methods 2023 May 12;14(3):563-579. [doi: 10.1002/jrsm.1635] [Medline: 37042216]
36. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
37. Peters M, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla K, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2017:1-24.
38. Mair FS, Wilkinson M, Bonnar SA, Wootton R, Angus RM. The role of telecare in the management of exacerbations of chronic obstructive pulmonary disease in the home. J Telemed Telecare 1999 Dec 02;5 Suppl 1(1_suppl):S66-S67. [doi: 10.1258/1357633991932603] [Medline: 10534847]
39. Tudiver F, Wolff LT, Morin PC, Teresi J, Palmas W, Starren J, et al. Primary care providers' perceptions of home diabetes telemedicine care in the IDEATel project. J Rural Health 2007 Jan 29;23(1):55-61. [doi: 10.1111/j.1748-0361.2006.00068.x] [Medline: 17300479]

XSL•FO
RenderX

40. Whitten P, Mickus M. Home telecare for COPD/CHF patients: outcomes and perceptions. J Telemed Telecare 2007 Jun 24;13(2):69-73. [doi: 10.1258/135763307780096249] [Medline: 17359569]
41. Trief PM, Sandberg J, Izquierdo R, Morin PC, Shea S, Brittain R, et al. Diabetes management assisted by telemedicine: patient perspectives. Telemed J E Health 2008 Sep;14(7):647-655. [doi: 10.1089/tmj.2007.0107] [Medline: 18817493]
42. Nilsson M, Rasmark U, Nordgren H, Hallberg P, Skönevik J, Westman G, et al. The physician at a distance: the use of videoconferencing in the treatment of patients with hypertension. J Telemed Telecare 2009 Nov 30;15(8):397-403. [doi: 10.1258/jtt.2009.090509] [Medline: 19948706]
43. Cook PF, Emiliozzi S, El-Hajj D, McCabe MM. Telephone nurse counseling for medication adherence in ulcerative colitis: a preliminary study. Patient Educ Couns 2010 Nov;81(2):182-186. [doi: 10.1016/j.pec.2009.12.010] [Medline: 20079598]
44. Sorknaes AD, Madsen H, Hallas J, Jest P, Hansen-Nord M. Nurse tele-consultations with discharged COPD patients reduce early readmissions--an interventional study. Clin Respir J 2011 Jan;5(1):26-34. [doi: 10.1111/j.1752-699X.2010.00187.x] [Medline: 21159138]
45. Campbell M, Akbari A, Amos S, Keyes C. Feasibility of providing nephrology services to remote communities with videoconferencing. J Telemed Telecare 2012 Jan 03;18(1):13-16. [doi: 10.1258/jtt.2011.110321] [Medline: 22052964]
46. Berkhof FF, van den Berg JW, Uil SM, Kerstjens HA. Telemedicine, the effect of nurse-initiated telephone follow up, on health status and health-care utilization in COPD patients: a randomized trial. Respirology 2015 Feb 17;20(2):279-285. [doi: 10.1111/resp.12437] [Medline: 25400242]
47. Mathar H, Fastholm P, Sandholm N. A qualitative study of televideo consultations for COPD patients. Br J Nurs 2015 Feb 26;24(4):205-209. [doi: 10.12968/bjon.2015.24.4.205] [Medline: 25723262]
48. Raymond JK, Berget CL, Driscoll KA, Ketchum K, Cain C, Fred Thomas JF. CoYoT1 clinic: innovative telemedicine care model for young adults with type 1 diabetes. Diabetes Technol Ther 2016 Jun;18(6):385-390 [FREE Full text] [doi: 10.1089/dia.2015.0425] [Medline: 27196443]
49. Reid MW, Krishnan S, Berget C, Cain C, Thomas JF, Klingensmith GJ, et al. CoYoT1 clinic: home telemedicine increases young adult engagement in diabetes care. Diabetes Technol Ther 2018 May;20(5):370-379. [doi: 10.1089/dia.2017.0450] [Medline: 29672162]
50. Lambooy S, Krishnasamy R, Pollock A, Hilder G, Gray NA. Telemedicine for outpatient care of kidney transplant and CKD patients. Kidney Int Rep 2021 May;6(5):1265-1272 [FREE Full text] [doi: 10.1016/j.ekir.2021.02.016] [Medline: 34013104]
51. Magliah SF, Alsabban AS, Turkistani HA, Abulaban BA, Alsharif AH, Alsharif SS, et al. Perception of virtual clinics among Saudi adults with type 1 diabetes during the COVID-19 pandemic. Diabetes Metab Syndr 2021 Jul;15(4):102175 [FREE Full text] [doi: 10.1016/j.dsx.2021.06.012] [Medline: 34186374]
52. Singh A, Mountjoy N, McElroy D, Mittal S, Al Hemyari B, Coffey N, et al. Patient perspectives with telehealth visits in cardiology during COVID-19: online patient survey study. JMIR Cardio 2021 Jan 22;5(1):e25074 [FREE Full text] [doi: 10.2196/25074] [Medline: 33385234]
53. Balut MD, Wyte-Lake T, Steers WN, Chu K, Dobalian A, Ziaeian B, et al. Expansion of telemedicine during COVID-19 at a VA specialty clinic. Healthc (Amst) 2022 Mar;10(1):100599 [FREE Full text] [doi: 10.1016/j.hjdsi.2021.100599] [Medline: 34999492]
54. Heyck Lee S, Ramondino S, Gallo K, Moist LM. A quantitative and qualitative study on patient and physician perceptions of nephrology telephone consultation during COVID-19. Can J Kidney Health Dis 2022 Jan 05;9:20543581211066720 [FREE Full text] [doi: 10.1177/20543581211066720] [Medline: 35024152]
55. Hitchcock M, Heath J. Understanding a young adult diabetes service's accessibility: have virtual appointments helped? Pract Diabetes 2022 Aug 02;39(4):32-37. [doi: 10.1002/pdi.2408]
56. Lohnberg JA, Salcido L, Frayne S, Mahtani N, Bates C, Hauser ME, et al. Rapid conversion to virtual obesity care in COVID-19: impact on patient care, interdisciplinary collaboration, and training. Obes Sci Pract 2022 Feb 28;8(1):131-136 [FREE Full text] [doi: 10.1002/osp4.550] [Medline: 34540265]
57. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
58. Peyroteo M, Ferreira IA, Elvas LB, Ferreira JC, Lapão LV. Remote monitoring systems for patients with chronic diseases in primary health care: systematic review. JMIR Mhealth Uhealth 2021 Dec 21;9(12):e28285 [FREE Full text] [doi: 10.2196/28285] [Medline: 34932000]
59. Walker RC, Tong A, Howard K, Palmer SC. Patient expectations and experiences of remote monitoring for chronic diseases: systematic review and thematic synthesis of qualitative studies. Int J Med Inform 2019 Apr;124:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.01.013] [Medline: 30784430]
60. Harrison C, Britt H, Miller G, Henderson J. Prevalence of chronic conditions in Australia. PLoS One 2013 Jul 23;8(7):e67494 [FREE Full text] [doi: 10.1371/journal.pone.0067494] [Medline: 23935834]
61. Boersma P, Black LI, Ward BW. Prevalence of multiple chronic conditions among US adults, 2018. Prev Chronic Dis 2020 Sep 17;17:E106 [FREE Full text] [doi: 10.5888/pcd17.200130] [Medline: 32945769]

XSL•FO
RenderX

62. Jürisson M, Pisarev H, Uusküla A, Lang K, Oona M, Kalda R. Prevalence of chronic conditions and multimorbidity in Estonia: a population-based cross-sectional study. BMJ Open 2021 Oct 05;11(10):e049045 [FREE Full text] [doi: 10.1136/bmjopen-2021-049045] [Medline: 34610934]
63. Singh K, Patel S, Biswas S, Shivashankar R, Kondal D, Ajay VS, et al. Multimorbidity in South Asian adults: prevalence, risk factors and mortality. J Public Health (Oxf) 2019 Mar 01;41(1):80-89 [FREE Full text] [doi: 10.1093/pubmed/fdy017] [Medline: 29425313]
64. Greenhalgh T, Shaw SE, Alvarez Nishio A, Byng R, Clarke A, Dakin F, et al. Remote care in UK general practice: baseline data on 11 case studies. NIHR Open Res 2022 Nov 29;2:47 [FREE Full text] [doi: 10.3310/nihropenres.13290.2] [Medline: 36814638]
65. Shiferaw KB, Mengiste SA, Gullslett MK, Zeleke AA, Tilahun B, Tebeje T, et al. Healthcare providers' acceptance of telemedicine and preference of modalities during COVID-19 pandemics in a low-resource setting: an extended UTAUT model. PLoS One 2021 Apr 22;16(4):e0250220 [FREE Full text] [doi: 10.1371/journal.pone.0250220] [Medline: 33886625]
66. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017 Sep;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
67. Khan R, Socha-Dietrich K. Investing in medication adherence improves health outcomes and health system efficiency: adherence to medicines for diabetes, hypertension, and hyperlipidaemia. The Organization for Economic Cooperation and Development (OECD). URL: https://www.oecd-ilibrary.org/content/paper/8178962c-en [accessed 2024-04-29]
68. Harvey PW, Petkov JN, Misan G, Fuller J, Battersby MW, Cayetano TN, et al. Self-management support and training for patients with chronic and complex conditions improves health-related behaviour and health outcomes. Aust Health Rev 2008 May;32(2):330-338. [doi: 10.1071/ah080330] [Medline: 18447824]
69. Salah H, Ibrahim Rabie AS, Said AS, AlAhmad MM, Shaaban AH, Khalil DM, et al. COVID-19's psychological impact on chronic disease patients seeking medical care. Healthcare (Basel) 2023 Mar 19;11(6):888 [FREE Full text] [doi: 10.3390/healthcare11060888] [Medline: 36981545]
70. Dubey S, Biswas P, Ghosh R, Chatterjee S, Dubey MJ, Chatterjee S, et al. Psychosocial impact of COVID-19. Diabetes Metab Syndr 2020 Sep;14(5):779-788 [FREE Full text] [doi: 10.1016/j.dsx.2020.05.035] [Medline: 32526627]
71. Kendzerska T, Zhu DT, Gershon AS, Edwards JD, Peixoto C, Robillard R, et al. The effects of the health system response to the COVID-19 pandemic on chronic disease management: a narrative review. Risk Manag Healthc Policy 2021;14:575-584 [FREE Full text] [doi: 10.2147/RMHP.S293471] [Medline: 33623448]
72. Reda AA, Kotz D, Kocks JW, Wesseling G, van Schayck CP. Reliability and validity of the clinical COPD questionniare and chronic respiratory questionnaire. Respir Med 2010 Nov;104(11):1675-1682 [FREE Full text] [doi: 10.1016/j.rmed.2010.04.023] [Medline: 20538445]
73. Tsiligianni IG, van der Molen T, Moraitaki D, Lopez I, Kocks JW, Karagiannis K, et al. Assessing health status in COPD. A head-to-head comparison between the COPD assessment test (CAT) and the clinical COPD questionnaire (CCQ). BMC Pulm Med 2012 May 20;12(1):20 [FREE Full text] [doi: 10.1186/1471-2466-12-20] [Medline: 22607459]
74. Zhou Z, Zhou A, Zhao Y, Chen P. Evaluating the clinical COPD questionnaire: a systematic review. Respirology 2017 Feb 19;22(2):251-262. [doi: 10.1111/resp.12970] [Medline: 28102972]
75. Haig-Ferguson A, Loades M, Whittle C, Read R, Higson-Sweeney N, Beasant L, et al. "It's not one size fits all"; the use of videoconferencing for delivering therapy in a specialist paediatric chronic fatigue service. Internet Interv 2019 Mar;15:43-51 [FREE Full text] [doi: 10.1016/j.invent.2018.12.003] [Medline: 30619719]
76. Ayabakan S, Bardhan IR, Zheng Z. Impact of telehealth and process virtualization on healthcare utilization. Inf Syst Res 2024 Mar;35(1):45-65. [doi: 10.1287/isre.2023.1220]
77. Raj M, Rai P, Narasimha GV, Onkar A, Angral S, Varshney S. Feasibility and acceptability of teleconsultation during COVID-19: a cross-sectional study. Cureus 2022 Oct;14(10):e30937 [FREE Full text] [doi: 10.7759/cureus.30937] [Medline: 36465724]
78. Mohammed R, Elmajid EA, Amine H, Khadija C. Acceptance factors of telemedicine technology during COVID-19 pandemic among health professionals: a qualitative study. Healthc Technol Lett 2023 Mar 10;10(1-2):23-33 [FREE Full text] [doi: 10.1049/htl2.12042] [Medline: 37077882]
79. Schwarz T, Schmidt AE, Bobek J, Ladurner J. Barriers to accessing health care for people with chronic conditions: a qualitative interview study. BMC Health Serv Res 2022 Aug 14;22(1):1037 [FREE Full text] [doi: 10.1186/s12913-022-08426-z] [Medline: 35964086]
80. Lazar M, Davenport L. Barriers to health care access for low income families: a review of literature. J Community Health Nurs 2018 Jan 11;35(1):28-37. [doi: 10.1080/07370016.2018.1404832] [Medline: 29323941]
81. Brundisini F, Giacomini M, DeJean D, Vanstone M, Winsor S, Smith A. Chronic disease patients' experiences with accessing health care in rural and remote areas: a systematic review and qualitative meta-synthesis. Ont Health Technol Assess Ser 2013;13(15):1-33 [FREE Full text] [Medline: 24228078]
82. Adepoju OE, Chae M, Ojinnaka CO, Shetty S, Angelocci T. Utilization gaps during the COVID-19 pandemic: racial and ethnic disparities in telemedicine uptake in federally qualified health center clinics. J Gen Intern Med 2022 Apr 02;37(5):1191-1197 [FREE Full text] [doi: 10.1007/s11606-021-07304-4] [Medline: 35112280]

XSL•FO
RenderX

83. White-Williams C, Liu X, Shang D, Santiago J. Use of telehealth among racial and ethnic minority groups in the united states before and during the COVID-19 pandemic. Public Health Rep 2023 Sep 16;138(1):149-156 [FREE Full text] [doi: 10.1177/00333549221123575] [Medline: 36113138]
84. Andronic O, Petrescu GE, Artamonov AR, Bolocan A, Rădăvoi D, Bran M, et al. Healthcare professionals' specialists' perception of telemedicine in Romania-a quantitative study of beliefs, practices, and expectations. Healthcare (Basel) 2023 May 25;11(11):1552 [FREE Full text] [doi: 10.3390/healthcare11111552] [Medline: 37297692]
85. Ma Q, Sun D, Tan Z, Li C, He X, Zhai Y, et al. Usage and perceptions of telemedicine among health care professionals in China. Int J Med Inform 2022 Oct;166:104856 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104856] [Medline: 36037593]
86. Ncube B, Mars M, Scott RE. Perceptions and attitudes of patients and healthcare workers towards the use of telemedicine in Botswana: an exploratory study. PLoS One 2023 Feb 16;18(2):e0281754 [FREE Full text] [doi: 10.1371/journal.pone.0281754] [Medline: 36795740]
87. Zeadally S, Bello O. Harnessing the power of internet of things based connectivity to improve healthcare. Internet Things 2021 Jun;14:100074. [doi: 10.1016/j.iot.2019.100074]
88. Srivastava G, Parizi RM, Dehghantanha A. The future of blockchain technology in healthcare internet of things security. In: Choo KK, Dehghantanha A, Parizi RM, editors. Blockchain Cybersecurity, Trust and Privacy. Cham, Switzerland: Springer; 2020:161-184.
89. Smits M, van Dalen D, Popping D, Bleeker R, Stommel MW, van Goor H. Designing a video consultation area for hybrid care delivery: the Garden Room with a view. Front Digit Health 2023 Jul 25;5:1198565 [FREE Full text] [doi: 10.3389/fdgth.2023.1198565] [Medline: 37564883]
90. Baker J, Stanley A. Telemedicine technology: a review of services, equipment, and other aspects. Curr Allergy Asthma Rep 2018 Sep 26;18(11):60. [doi: 10.1007/s11882-018-0814-6] [Medline: 30259201]
91. Stovel RG, Gabarin N, Cavalcanti RB, Abrams H. Curricular needs for training telemedicine physicians: a scoping review. Med Teach 2020 Aug 05;42(11):1234-1242. [doi: 10.1080/0142159x.2020.1799959]
92. Robbins E, Johal K, Keswani A. The use and teaching of telemedicine in allergy/immunology training programs. Curr Allergy Asthma Rep 2023 Aug 07;23(8):463-470. [doi: 10.1007/s11882-023-01096-9] [Medline: 37418070]
93. Edirippulige S, Armfield NR. Education and training to support the use of clinical telehealth: a review of the literature. J Telemed Telecare 2017 Feb 08;23(2):273-282. [doi: 10.1177/1357633X16632968] [Medline: 26892005]
94. Danbjorg DB, Clemensen J, Rothmann MJ. Participatory design methods in telemedicine. Int J Integr Care 2016 Nov 09;16(5):11 [FREE Full text] [doi: 10.5334/ijic.2559]
95. Fouquet SD, Miranda AT. Asking the right questions-human factors considerations for telemedicine design. Curr Allergy Asthma Rep 2020 Aug 29;20(11):66 [FREE Full text] [doi: 10.1007/s11882-020-00965-x] [Medline: 32862299]
96. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
97. Real-time language translation for virtual healthcare. CirrusMD. 2023. URL: https://www.cirrusmd.com/real-time-language-translation-for-virtual-healthcare [accessed 2024-04-29]
98. Kwint E, Zoet A, Labunets K, Brinkkemper S. How different elements of audio affect the word error rate of transcripts in automated medical reporting. In: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies. 2023 Presented at: BIOSTEC '23; February 16-18, 2023; Lisbon, Portugal p. 179-187 URL: https://dspace.library.uu.nl/handle/1874/437593 [doi: 10.5220/0011794100003414]
99. Vegesna A, Tran M, Angelaccio M, Arcona S. Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed J E Health 2017 Jan;23(1):3-17 [FREE Full text] [doi: 10.1089/tmj.2016.0051] [Medline: 27116181]
100. de Farias FA, Dagostini CM, Bicca YD, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020 May 01;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
101. Faber J, Fonseca LM. How sample size influences research outcomes. Dental Press J Orthod 2014 Aug;19(4):27-29 [FREE Full text] [doi: 10.1590/2176-9451.19.4.027-029.ebo] [Medline: 25279518]
102. Timpel P, Harst L. Research implications for future telemedicine studies and innovations in diabetes and hypertension-a mixed methods study. Nutrients 2020 May 08;12(5):1340 [FREE Full text] [doi: 10.3390/nu12051340] [Medline: 32397096]
103. Goharinejad S, Hajesmaeel-Gohari S, Jannati N, Goharinejad S, Bahaadinbeigy K. Review of systematic reviews in the field of telemedicine. Med J Islam Repub Iran 2021 Oct 30;35:184 [FREE Full text] [doi: 10.47176/mjiri.35.184] [Medline: 36042824]

Abbreviations

CCC: chronic cardiovascular condition
COPD: chronic obstructive pulmonary disease
HCP: health care professional
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by G Eysenbach; submitted 01.10.23; peer-reviewed by KM Surapaneni, R Rosenbacke; comments to author 02.01.24; revised version received 27.01.24; accepted 19.03.24; published 09.07.24.

Please cite as:
Dhunnoo P, Kemp B, McGuigan K, Meskó B, O'Rourke V, McCann M
Evaluation of Telemedicine Consultations Using Health Outcomes and User Attitudes and Experiences: Scoping Review
J Med Internet Res 2024;26:e53266
URL: https://www.jmir.org/2024/1/e53266
doi:10.2196/53266
PMID:38980704

©Pranavsingh Dhunnoo, Bridie Kemp, Karen McGuigan, Bertalan Meskó, Vicky O'Rourke, Michael McCann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perception and Acceptance of Telemedicine Use in Health Care Among the General Public in China: Web-Based Cross-Sectional Survey

Yulan Lin[1*], MD, PhD; Xiaonan Xu[1*], MSc; Yiyang Liu[1*], MSc; Haridah Alias[2], MSc; Zhijian Hu[1], PhD; Li Ping Wong[1,2], PhD

[1]Fujian Key Laboratory of Environmental Factors and Cancer, Department of Epidemiology and Health Statistics, School of Public Health, Fujian Medical University, Fuzhou, China

[2]Centre of Population Heath (CePH), Department of Social and Preventive Medicine, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, Malaysia

[*]these authors contributed equally

Corresponding Author:
Zhijian Hu, PhD
Fujian Key Laboratory of Environmental Factors and Cancer
Department of Epidemiology and Health Statistics, School of Public Health
Fujian Medical University
1 Xue Yuan Road
Fuzhou, 350122
China
Phone: 86 5912286 ext 2573
Email: huzhijian@fjmu.edu.cn

Abstract

Background: The COVID-19 pandemic is bringing about substantial changes in health care systems, leading to a significant shift toward telemedicine for the delivery of health care services.

Objective: This study aims to examine the relationship between perceived usefulness and ease of use of telemedicine services and their association with the behavioral intention to use telemedicine.

Methods: An anonymous cross-sectional survey was conducted in China. Partial least squares structural equation modeling was used to determine significant predictors of intention to use telemedicine consultation. Types of illnesses that favored seeking telemedicine consultation, as well as the most preferred platform for conducting telemedicine consultations, were also investigated.

Results: In total, 1006 participants completed the survey. A total of 44.3% (n=446) reported being very likely and 49.3% (n=496) reported being likely to seek telemedicine consultation. Overall, the majority of participants expressed strong agreement or agreement regarding the perceived usefulness of telemedicine. Likewise, the majority indicated strong agreement or agreement when it came to their perception of the ease of using telemedicine. In the partial least squares structural equation modeling, perceived usefulness (β=0.322; P<.001) and perceived ease of use (β=0.118; P=.01) were significantly associated with a higher likelihood of seeking telemedicine consultation. A considerable number of participants expressed willingness to use telemedicine services for various medical conditions, particularly respiratory (n=340, 33.8%), skin (n=316, 31.4%), and musculoskeletal issues (n=316, 31.4%) while showing less interest in seeking telemedicine consultations for reproductive health (n=44, 4.4%) and cancer (n=64, 6.4%). The majority preferred video chat (n=443, 44%) and text chat (n=317, 31.5%) as their most preferred platforms for telemedicine consultation, while a smaller proportion preferred telephone (n=193, 19.2%) and email (n=53, 5.3%).

Conclusions: Telemedicine has the potential to play a larger role in China's health care system. The preferences for certain platforms over others may influence service design and implementation.

(J Med Internet Res 2024;26:e53497) doi:10.2196/53497

KEYWORDS

telemedicine; acceptance; China; general public; COVID-19; pandemic; health care; public health; health care delivery; health care services; survey; cross-sectional survey; consultation; teleconsultation; health care system

XSL•FO
RenderX

Introduction

Telemedicine in China had its beginnings earlier, but the COVID-19 pandemic accelerated its widespread adoption and practice [1]. In China, before the pandemic, telemedicine was an emerging health care delivery approach that was undergoing rapid development. Telemedicine has been instrumental in enhancing the distribution of top-tier health care resources from urban centers to remote rural areas. Prior to the pandemic, it played a pivotal role in ensuring equitable health care services between urban and underserved regions while also advancing the development of China's tiered medical system [2]. The telemedicine network in China was activated immediately following the initial COVID-19 outbreak in January 2020 [3]. The country leveraged telemedicine and tele-education effectively throughout the pandemic, ensuring the safety of both patients and health care workers while delivering timely health care services. Telemedicine proved its feasibility, acceptability, and effectiveness, leading to substantial enhancements in health care outcomes during the pandemic in China [3].

Telemedicine has indeed proven to be useful during both public health emergencies and nonemergency times, offering cost savings and maintaining quality medical care through remote consultations [4]. As the world transitions back to prepandemic normalcy, uncertainties persist regarding whether the general public's acceptance and use of telemedicine have evolved. It remains essential for the public to sustain their backing for telemedicine services, even as the pandemic subsides. Research is warranted to assess this acceptance, as it holds significant importance, particularly in identifying areas where telemedicine can be improved, such as addressing concerns around privacy, security, and quality of care. This helps ensure that telemedicine services align with patient needs, improve health outcomes, and optimize resource use.

Perceived ease of use was reinforced as a key driver of using telemedicine [5]. Telemedicine enhances health care access by allowing practitioners to offer services beyond geographical boundaries, reducing noncritical hospital visits, and alleviating congestion in health care facilities [5]. Telemedicine also effectively bridges the gap between providers and patients in remote locations, ensuring easy access to health care services for patients in these areas [6]. Both perceived ease of use and perceived usefulness in telemedicine are important indicators for assessing the successful adoption and integration of this innovative health care approach into our modern health care systems. These 2 factors play a pivotal role in influencing individuals' decisions to embrace telemedicine, ultimately shaping its impact on health care accessibility and efficiency.

The first aim of this study was to examine the relationship between perceived usefulness and ease of use of telemedicine services and their association with the general public's behavioral intention to use telemedicine. Using the technology acceptance model (TAM) as a theoretical framework, we postulate that actual behaviors are predicted by behavioral intention, and behavioral intention is underpinned by perceived usefulness and perceived ease of telemedicine use. The TAM is widely used to investigate the acceptance and use of technology by individuals. It is one of the most popular frameworks for understanding technology adoption at the individual level [7]. Second, this study also explored the types of illnesses that people would prefer to seek telemedicine consultations as well as their preferred platforms for conducting these consultations. The study's findings aim to contribute to a deeper understanding of the factors driving telemedicine adoption in China while also providing valuable insights for policy makers and health care providers looking to improve the accessibility and acceptance of telemedicine services in the aftermath of the COVID-19 pandemic.

Methods

Study Participants and Survey Design

A nationwide cross-sectional open survey using a web-based questionnaire was conducted from August 30, 2021, to May 1, 2022. A convenient sampling method was used. The study used Wenjuanxing (Changsha Jingwei Cloud Computing Technology Co, Ltd), a widely accepted web-based questionnaire survey platform in China for data collection. The survey platform prevents multiple responses by the same individual by disallowing users with the same IP address from accessing the survey more than once. The survey was disseminated mainly using WeChat (Tencent Holdings Ltd), China's largest social media platform. Responses were automatically captured by the survey platform and transferred to SPSS (IBM Corp) for analysis. The inclusion criteria were that the participants were Chinese citizens aged 18 years or older. Data collection took place across 6 main regions in mainland China: North, Northeast, East, Southcentral, Southwest, and Northwest. First, the research team advertised and disseminated the survey with colleagues, students, and their network contacts. Second, the snowball sampling technique was used to enlist more participants across all provinces in mainland China, where individuals who received the survey link were encouraged to pass on the invitations to their own network members.

The calculated sample size was 385, based on a normal approximation of binomial distribution with a finite population correction applied, assuming an observer proportion of participants selecting a specific response option of 50%, a 95% confidence level, and a margin of error of 5%. The sample size was multiplied by the predicted design effect of 2 to account for the use of convenience sampling and a web-based survey [8]. Therefore, the minimum sample size for this study was 770 (385×2) participants. Survey participants were encouraged to share the survey link with everyone in their contact list.

Measures

Overview

The demographic characteristics of study participants were first determined. Subsequently, the survey questionnaire (Multimedia Appendix 1) was divided into 3 sections that assessed participants' perceptions of telemedicine services in terms of usefulness, ease of use, and behavioral intention to use them if it is available.

XSL•FO
RenderX

Demographic Characteristics

In this study, demographic information such as participants' age, sex, highest educational attainment, annual family income, and birthplace were collected.

Perceived Usefulness of Telemedicine

The questions on perceived usefulness were self-developed to evaluate participants' perceptions of telemedicine's use and consist of 4 items. These questions assess the usefulness of telemedicine during public health emergencies and nonemergency times, its cost-effectiveness compared to conventional medical care-seeking methods, and its capacity to deliver quality health care services on par with traditional face-to-face consultations. The response choices were presented using a 4-point Likert scale: 4=strongly agree, 3=agree, 2=disagree, and 1=strongly disagree. The items have a Cronbach α value of 0.631, composite reliability (CR) of 0.760, and average variance extracted (AVE) of 0.391.

Perceived Ease of Telemedicine Use

Likewise, the perceived ease of use questionnaire was also self-developed. The questionnaire comprises 5 questions aimed at evaluating the ease with which participants perceive telemedicine as an effective way to enhance health care access, provide services across geographical boundaries, deliver care to remote patients, reduce noncritical hospital visits, and alleviate hospital congestion. Likewise, the response choices were presented using a 4-point Likert scale: 4=strongly agree, 3=agree, 2=disagree, and 1=strongly disagree. The Cronbach α value for perceived ease of telemedicine use is 0.513, with a CR of 0.731, and AVE of 0.406.

Behavioral Intention

The question about behavioral intention asked participants how inclined they were to seek telemedicine consultation if it were accessible to them. Option answers were "very likely," "likely," "unlikely," and "strongly unlikely." They were also asked about the types of illnesses they would like to seek telemedicine consultation as well as their most preferred platform for conducting telemedicine consultations.

All questions in the questionnaire require a response; participants must fill them out before proceeding to the next page. As a result, all responses received through the survey platform are complete. Before the survey commenced, the questionnaire underwent content validation by a panel of experts in the field and was subsequently pilot-tested with the general public before its official dissemination.

Statistical Analyses

Partial least squares structural equation modeling (PLS-SEM) was used to predict factors influencing the likelihood of seeking telemedicine consultation if it is available. This technique assesses the reliability of the data set, the statistical significance of the coefficients, and the error of the estimated path coefficients [9]. The bootstrapped significance calculation was performed in SmartPLS software (version 3.2.8; SmartPLS GmbH). Prior to running the path model, the construct validity (convergent and discriminant) was tested. In the PLS-SEM analysis, perceived usefulness and perceived ease of use were considered as reflective constructs, and all other independent variables and the outcome—likelihood to seek telemedicine consultation if it is available—were single-item constructs. As PLS-SEM is a very useful technique for evaluating complex theoretical relationships between multiple variables, the PLS-SEM analysis is also used to explain the effect of demographic variables on the association between seeking telemedicine consultation and the perceived usefulness of telemedicine or perceived ease of telemedicine. Results of the measurement model indicated that all indicators had an acceptable outer loading (range 0.541-0.721) with a Cronbach α value of 0.513 for perceived usefulness and 0.631 for perceived ease of use, a CR value of 0.731 for perceived usefulness and 0.760 for perceived ease of use, and an AVE value of 0.406 for perceived usefulness and 0.391 for perceived ease of use. The variance inflation factors for all indicators were below 2.5 (ranging from 1.108 to 1.278), which revealed that all indicators belonging to these 2 constructs were adequately independent. Discriminant validity assessment through the heterotrait-monotrait ratio of correlations method also indicated that all heterotrait-monotrait values were lower than the most restrictive threshold (0.85; range 0.028-0.558).

Ethical Considerations

This study was implemented according to the principles of the Declaration of Helsinki and granted ethical approval by the Fujian Medical University Research Ethics Committee, China (approval FJMU 2022 No 64). Consent was acquired by presenting the consent statement at the survey's outset and mandating participants to click an agreement button, signifying their consent to participate, before proceeding to answer the survey questions. Informed consent involves providing participants with information sheets, informing them about the survey's timeframe and the estimated time required to complete it, and assuring them of the anonymity of their responses. The participant's information sheet also contains details about the investigators and the institution conducting the survey. The participants were informed that their participation was voluntary, and no incentives, whether monetary or nonmonetary, were provided. The web-based consent used in the web-based survey received approval from the ethical committee. The survey did not collect any identifying information from participants. Data are securely stored on a cloud platform accessible only to the principal investigator.

Results

User Statistics

A total of 1006 complete responses were received in the survey. Table 1 shows the demographics of the study participants. The age of the participants ranged from 18 to 64 (mean 28.2, SD 7.1) years. A substantial portion of the participants fell within the 18-29 years age bracket, accounting for 64.9% (n=653) of the total. Sex distribution was nearly equal, with 50.7% (n=510) male participants and 49.3% (n=496) female participants. In terms of their highest educational attainment, 16% (n=161) had completed high school, while the majority (n=793, 78.8%) were university graduates. When considering income categories, the majority of participants reported family annual incomes in the

XSL•FO
RenderX

range of CN ¥150,001 to CN ¥200,000 (approximately US $20,655 to US $27,540), comprising 53.2% (n=416) of the sample. A substantial proportion of participants (n=801, 79.6%) were from urban areas. The majority of responses were from Northern (45.8%, n=461) and Eastern (n=302, 30%) regions (Figure 1).

Table 1. Participants' characteristics (N=1006).

Sociodemographic characteristics	Values, n (%)
Age group (years)	
18-29	653 (64.9)
30-39	275 (27.3)
40-64	78 (7.8)
Sex	
Male	510 (50.7)
Female	496 (49.3)
Highest educational level	
Secondary school and below	52 (5.2)
High school or technical school	161 (16)
Tertiary	793 (78.8)
Family annual income (CN ¥; US $)	
<50,000 (<6885)	83 (8.3)
50,000-100,000 (6885-13,770)	222 (22.1)
100,001-150,000 (13,770-20,655)	285 (28.3)
150,001-200,000 (20,655-27,540)	250 (24.9)
250,001-300,000 (27,540-41,310)	130 (12.9)
>300,000 (>41,310)	36 (3.6)
Birthplace	
Urban	801 (79.6)
Rural	205 (20.4)

XSL•FO
RenderX

Figure 1. The proportion of responses by region and population distribution.

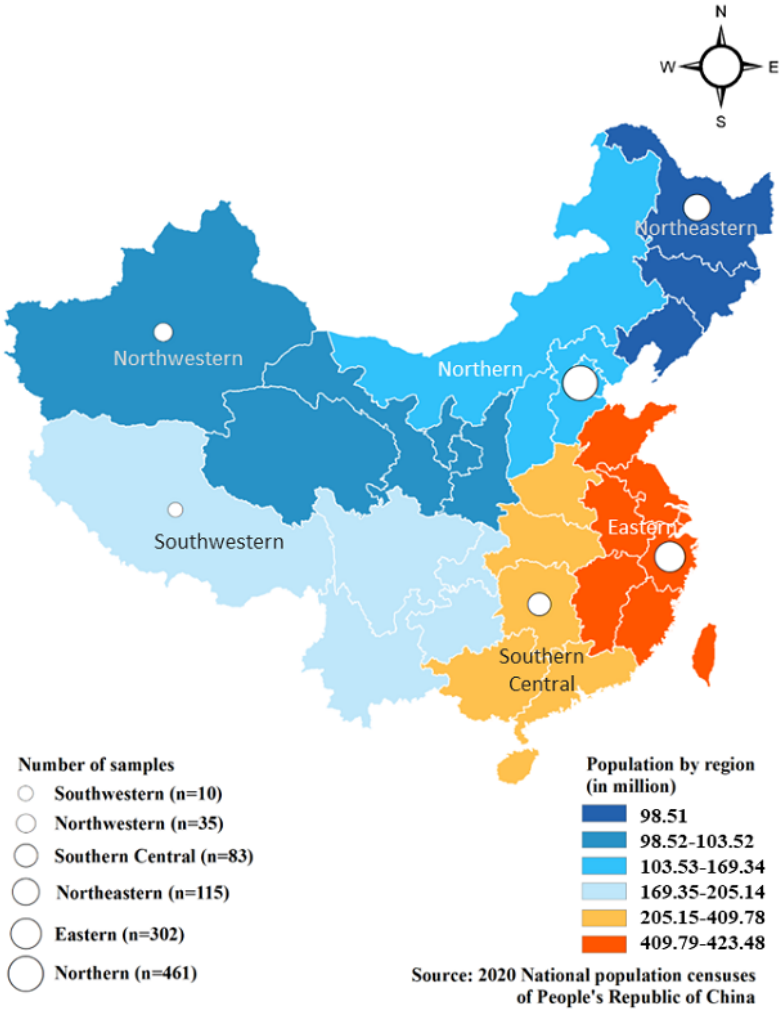

Evaluation Outcomes

Table 2 shows the proportion of responses on individual items of perceived usefulness and ease of use. The highest percentage (n=482, 47.9%) strongly agreed that telemedicine is essential in nonpublic emergency situations. In contrast, a relatively lower proportion (n=213, 21.2%) strongly agreed that telemedicine is useful during public health emergencies. However, there was a nearly equal split among participants who strongly agreed that telemedicine visits are cost-saving compared to conventional medical care–seeking methods (n=345, 34.3%) as well as those who strongly agreed that telemedicine provides medical services of equal quality to traditional face-to-face consultations (n=317, 31.5%). Overall, the majority of participants expressed strong agreement or agreement regarding the perceived usefulness of telemedicine.

Likewise, the majority of participants indicated strong agreement or agreement when it came to their perception of the ease of using telemedicine. Among all the items, the highest proportion of strong agreement was observed in the perception that telemedicine is an effective means to improve access to health services (n=427, 42.4%), followed by the belief that telemedicine can enable health care providers to deliver services to patients in remote locations, with 37.7% (n=379) strongly agreeing. A relatively lower proportion of strong agreement was noted for the statement that telemedicine can replace hospital visits for noncritical illnesses, with 31.4% (n=316) strongly agreeing.

Overall, a total of 44.3% (n=446) reported being very likely to seek telemedicine consultation, and 49.3% (n=496) reported being likely to seek telemedicine consultation. Only 5.2% (n=52) reported being unlikely, and 1.2% (n=12) reported being strongly unlikely. Acceptance of telemedicine consultation by region is shown in Figure 2. The Northeastern region exhibits the highest likelihood of seeking telemedicine, with 54.8% (n=153) of individuals reported very likely to seek telemedicine consultation. The PLS-SEM in Figure 3 shows the associations of all the factors associated with the likelihood of seeking telemedicine consultation. As depicted in the figure, perceived usefulness (β=0.322; P<.001) showed the greatest effect in influencing a higher likelihood to seek telemedicine consultation. Perceived ease of use (β=0.118; P=.01) is also associated with a higher likelihood to seek telemedicine consultation. Results for adjusted R^2 indicated that this model explained 17.1% of the total variance in the likelihood of seeking telemedicine consultation.

Table 3 shows the responses of participants' perspectives on the types of illnesses that they would like to seek telemedicine consultation. Only 16.4% (n=165) indicated a likelihood of

XSL•FO
RenderX

seeking telemedicine for all illnesses. A high proportion of participants reported that they were likely to seek telemedicine consultation for respiratory (n=340, 33.8%), skin diseases, and musculoskeletal (n=316, 31.4%)-related diseases. However, a low proportion reported the likelihood of seeking telemedicine for reproductive health (n=44, 4.4%) and cancer (n=64, 6.4%). The findings on the most preferred platform for telemedicine consultation indicate that the majority preferred video chat (n=443, 44%), followed by text chat (n=317, 31.5%). A small proportion reported preferring telephone (n=193, 19.2%) and via email (n=53, 5.3%).

Table 2. Responses regarding perceived usefulness and ease of use of telemedicine services (N=1006).

Items	Strongly agree, n (%)	Agree, n (%)	Disagree, n (%)	Strongly disagree, n (%)
PU[a]				
PU1: Telemedicine is essential for use during public health emergencies (eg, pandemics, infectious disease outbreaks, and flooding)	213 (21.2)	681 (67.7)	101 (10)	11 (1.1)
PU2: Telemedicine is essential for use during the time of no public health emergencies	482 (47.9)	401 (39.9)	103 (10.2)	20 (2)
PU3: Telemedicine visits are cost savings compared to conventional way of seeking medical care	345 (34.3)	529 (52.6)	108 (10.7)	24 (2.4)
PU4: Telemedicine can equally provide quality medical service as the traditional face-to-face consultation	317 (31.5)	523 (52)	134 (13.3)	32 (3.2)
PEU[b]				
PEU1: Telemedicine is an effective way to improve access to health services	427 (42.4)	453 (45)	102 (10.1)	24 (2.4)
PEU2: Telehealth practitioners can provide medical services across geographic borders	360 (25.8)	506 (50.3)	116 (11.5)	24 (2.4)
PEU3: Telemedicine can enable providers deliver health services to patients at remote locations	379 (37.7)	519 (51.6)	77 (7.7)	31 (3.1)
PEU4: Telemedicine can circumvent hospital visits of noncritical illnesses	316 (31.4)	499 (49.6)	132 (13.1)	59 (5.9)
PEU5: Telemedicine can help to alleviate hospital congestion	311 (30.9)	537 (53.4)	119 (11.8)	39 (3.9)

[a]PU: perceived usefulness.

[b]PEU: perceived ease of use.

XSL•FO
RenderX

Figure 2. Acceptance of telemedicine consultations by region.

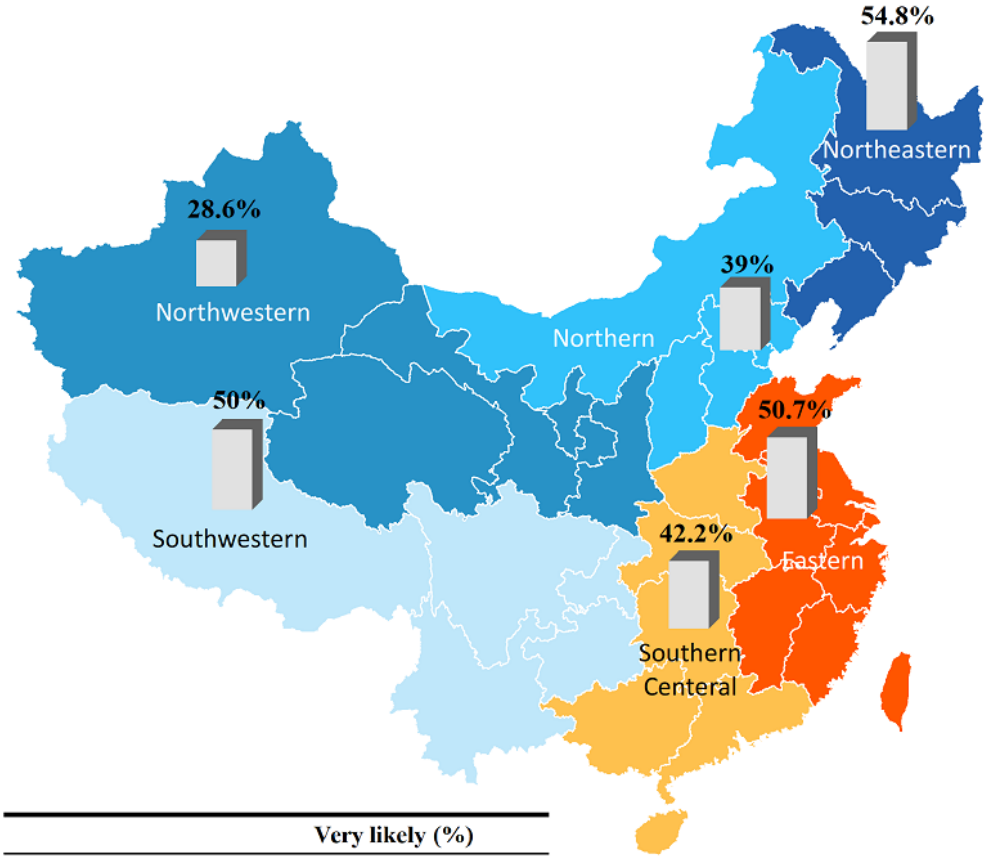

	Very likely (%)
Northern	39
Northeastern	54.8
Eastern	50.7
Southern Central	42.2
Southwestern	50
Northwestern	28.6

Figure 3. Partial least squares structural equation modeling of factors influencing the likelihood of seeking telemedicine services. PEU: perceived ease of use; PU: perceived usefulness.

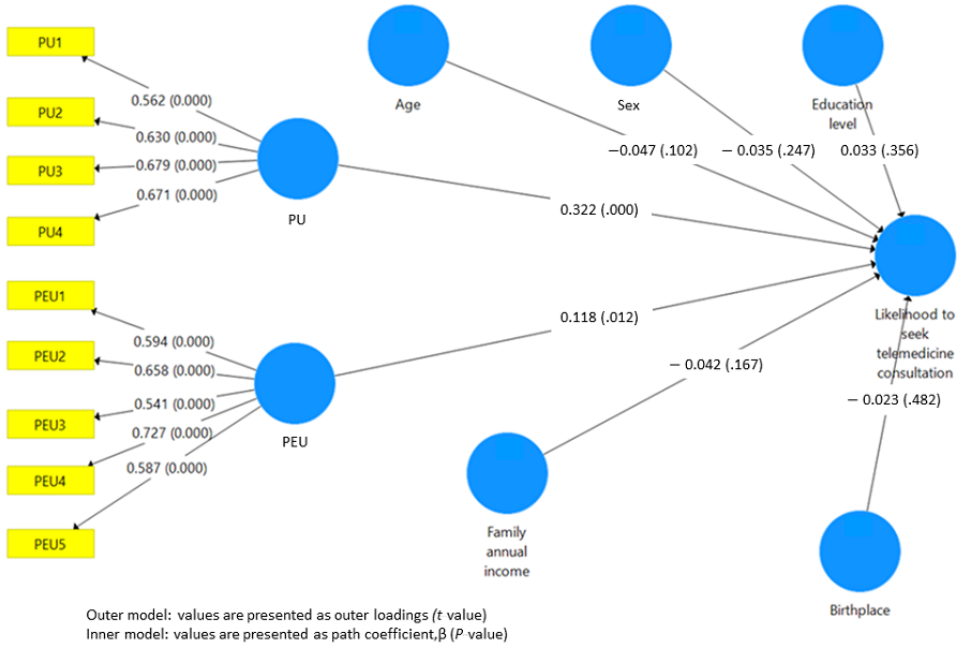

XSL•FO
RenderX

Table 3. Telemedicine consultation preferences by type of illness (N=1006).

Type of illness	Values, n (%)
Respiratory (eg, asthma, bronchitis, and tuberculosis)	340 (33.8)
Musculoskeletal (eg, arthritis, rheumatism, and backache)	316 (31.4)
Skin diseases	316 (31.4)
Endocrine (eg, thyroid disease, diabetes, and lupus	274 (27.2)
Circulatory (hypertension)	272 (27)
Cardiovascular diseases (heart problem and stroke)	260 (25.8)
Neurologic diseases (migraine, sclerosis, and epilepsy)	259 (25.7)
Psychological (depression, anxiety, and other emotional issue)	243 (24.2)
Digestive diseases (trouble with mouth, gums, ulcers, indigestion, diarrhea, and constipation)	243 (24.2)
All illnesses	165 (16.4)
Infectious diseases	98 (9.7)
Cancer	64 (6.4)
Reproductive health and related diseases	44 (4.4)

Discussion

Principal Findings

This study shed light on the perceptions of both the perceived usefulness and ease of use of telemedicine among the study participants. First, in regard to findings on the perceived usefulness of telemedicine, it is noteworthy that the highest percentage of participants strongly agreed that telemedicine is essential in nonpublic emergency situations. This finding reflects a broad acknowledgment of telemedicine's importance, even as the pandemic subsides, indicating continued public support for its use in daily health care. A considerably high proportion of participants strongly agreed that telemedicine visits are cost-saving compared to conventional medical care–seeking methods, and those who strongly agreed that telemedicine provides medical services of equal quality to traditional face-to-face consultations align with the core objectives of telemedicine implementation [10-12].

Second, with regard to the findings on the perceived ease of use of telemedicine, the majority of participants indicated strong agreement or agreement across all items related to ease of use. The highest proportion of strong agreement was observed in the perception that telemedicine is an effective means to improve access to health services. This highlights the belief that telemedicine can bridge geographical barriers and enhance health care accessibility, especially for those in remote or underserved areas. Moreover, a high proportion of participants strongly agreed that telemedicine can enable health care providers to deliver services to patients in remote locations, reaffirming the perception that telemedicine has the potential to expand the reach of health care services, even to geographically isolated communities. The existing health inequity and scarcity of medical resources in rural China provide a unique opportunity for telemedicine to play a crucial role in bridging this gap [13]. The positive perceptions of telemedicine suggest that, overall, the public is open to adopting new technologies to address their health care needs.

Finally, while a relatively lower proportion of participants strongly agreed that telemedicine can replace hospital visits for noncritical illnesses, it still signifies a significant level of acceptance. This suggests that individuals may be open to using telemedicine for a range of health care needs, albeit with some reservations about replacing in-person visits for more serious conditions. The public should be made aware that there is a growing body of evidence that supports the role of telemedicine in the management of serious health conditions. Telemedicine has been shown to be effective in managing chronic diseases, specifically, congestive heart failure, stroke, and chronic obstructive pulmonary disease [14]. While telemedicine should not replace in-person care entirely, it can supplement it and provide greater flexibility and convenience for patients and health care providers. Resistance to telemedicine for serious illnesses can be overcome by implementing a hybrid model that includes remote consultations, specialist consultations, remote monitoring, follow-up appointments, access to timely and appropriate care, and continuity of care [6]. This approach can increase patient trust and confidence in telemedicine, which can lead to higher adoption rates, resulting in better health outcomes and increased patient satisfaction.

This study found that the overall acceptance of telemedicine is quite substantial, with a substantial proportion of participants expressing a willingness to seek telemedicine consultations. However, it is also essential to note that a small minority of participants expressed that they were unlikely. These findings suggest that while the majority of individuals are receptive to telemedicine, there is still a segment of the population that may have reservations, concerns, or barriers to using telemedicine services. Identifying and addressing these concerns is critical to achieving broader telemedicine adoption.

Moreover, this study highlights notable regional discrepancies in the acceptance of telemedicine, with the Northeastern and Eastern regions exhibiting the greatest propensity for seeking remote medical consultations. This phenomenon can be attributed, in part, to the established correlation between regional

XSL•FO
RenderX

economic prosperity and health care disparities within China. In China, high-quality medical resources are concentrated in the economically advanced Eastern region, while the Central and Western regions have a relative lack of health resources [15]. Furthermore, individuals residing in the affluent Eastern region tend to possess superior health outcomes [16]. Consequently, factors, such as health care infrastructure, public familiarity, and educational initiatives surrounding telemedicine, as well as local attitudes toward web-based health care delivery, might all play a role in shaping these regional variances. By recognizing and addressing these disparities, policy makers can develop tailored measures to enhance telemedicine adoption and ensure uniform access to quality health care across China.

Our finding also highlights the importance of addressing the regional disparities in telemedicine acceptance, particularly in poorer regions such as the Western area. These regions may potentially benefit the most from telemedicine due to their limited access to quality health care services. Enhancing acceptance of telemedicine in these areas can be achieved through various means, including increasing awareness and education about the benefits of telemedicine, providing accessible and user-friendly technology, and offering culturally sensitive telemedicine services that cater to the unique needs of diverse patient populations. Additionally, implementing policies that encourage the use of telemedicine, such as reimbursement models that support remote consultations, can help overcome financial barriers and promote greater adoption in resource-poor settings.

The findings of PLS-SEM on the relationships between various factors and the likelihood of seeking telemedicine consultations provide valuable insights into the key determinants influencing individuals' intentions to use telemedicine. The analysis highlights that perceived usefulness and perceived ease of use play a significant and robust role in influencing the likelihood of seeking telemedicine consultations. It is worth noting that similar findings have been reported in other studies [17,18], underscoring the consistency of these factors as critical drivers of telemedicine adoption across various health care contexts. These consistent findings emphasize the importance of addressing perceived usefulness and ease of use when designing and promoting telemedicine services to ensure their acceptance and effectiveness among a diverse patient population.

This study also provides important insights into the varying degrees of acceptance among participants to use telemedicine for different types of illnesses. Notably, only a small proportion, representing 16.4% (n=165) of the participants, indicated a likelihood of seeking telemedicine for all types of illnesses. This finding suggests that while telemedicine is gaining acceptance as a viable option for medical consultation, there remains a significant portion of the population who may not be fully comfortable or confident in using telemedicine for a comprehensive range of health concerns. Respiratory issues emerged as a primary area of interest among participants, with a significant number indicating their likelihood to use telemedicine for such conditions. This heightened interest in telemedicine for respiratory issues could be attributed to the fact that respiratory problems, such as common colds and seasonal allergies, are highly prevalent and often less threatening compared to other serious health conditions. Patients may feel more comfortable seeking remote consultations for these common ailments, which they are more likely to encounter in their daily lives.

However, the results also reveal areas where telemedicine adoption may face challenges. For instance, a relatively low proportion of participants reported their likelihood to seek telemedicine for reproductive health issues. This suggests that individuals may have reservations or concerns related to the sensitive nature of reproductive health concerns and the preference for in-person consultations in this context. Likewise, a relatively small percentage expressed a willingness to use telemedicine for cancer-related matters. Cancer is a complex and serious health condition, and individuals may feel more comfortable receiving such diagnoses and treatments in traditional health care settings, where physical examinations and emotional support are readily available.

The results regarding participants' preferences for the platform used in telemedicine consultations provide valuable insights into the communication methods they find most suitable and comfortable. First, it is noteworthy that many expressed a preference for video chat as their most favored platform for telemedicine consultations. Video chat offers the advantage of face-to-face interaction, allowing health care providers to visually assess patients and create a more personal connection. This preference may reflect the desire for a more comprehensive and visual consultation experience, which can be particularly important for conditions that require visual examination. Second, text chat emerged as the second most preferred platform. Text chat offers a written communication channel, which can be beneficial for discussing medical concerns, sharing information, and seeking advice without the need for real-time interaction. Text chat offers the advantage of enabling patients to furnish written descriptions of their medical concerns, thereby ensuring important details are accurately conveyed to health care providers.

Limitations

This study shares the same constraints commonly associated with web-based surveys. Aside from socially desirable responses, this survey particularly tends to draw participants with computer literacy and a notable interest or predisposition toward the subject matter. Second, this study is constrained by an uneven distribution of samples, as it lacked adequate representation from the Southwestern and Northwestern regions. China's Southwestern and Northwestern regions have relatively lower income compared to the higher-income Eastern and Coastal regions. Therefore, careful interpretation is necessary, and the findings may not be applicable to the general population. In particular, further research is required to explore and evaluate TAMs in resource-constrained environments.

In general, Cronbach $\alpha>0.7$, CR>0.7, and AVE≥0.5 are considered acceptable benchmarks for validity and reliability [19-21]. In this study, the items concerning perceived usefulness and ease of use of telemedicine exhibited Cronbach α values lower than 0.7 and AVE values lower than 0.5, although the CR value is of acceptable value. While the general benchmark for Cronbach α is often set at 0.7 or higher for internal

XSL•FO
RenderX

consistency, values slightly below this threshold can still be acceptable. Particularly in psychological research, a Cronbach α value of 0.5 was deemed acceptable or reliable [22-24]. Regarding the AVE values, both item values ranged between 39% and 41%, which are below the recommended level of 0.5. Nevertheless, it has been rationalized that the AVE may be a more conservative estimate of the validity of the measurement model [25], and based on CR alone, the researcher may conclude that the convergent validity of the construct is adequately reliable [26]. The CR for both perceived usefulness and ease of use of telemedicine ranges from 0.73 to 0.76, surpassing the recommended threshold of 0.60 [10,25]. Hence, the internal reliability of both measurement items could be deemed adequately acceptable.

Conclusions

This study's results highlight a generally positive attitude toward telemedicine, with a substantial proportion of individuals expressing a willingness to seek telemedicine consultations. However, it also underscores the need to address the concerns of those who remain hesitant or unlikely to use telemedicine. The regional differences in acceptance emphasize the importance of tailoring telemedicine implementation strategies to specific geographic areas to ensure widespread access to this valuable health care resource. While there is strong interest in using telemedicine for respiratory, skin, and musculoskeletal issues, health care providers and policy makers should be mindful of the lower willingness to adopt telemedicine for reproductive health and cancer-related concerns. Addressing patient concerns, ensuring privacy and security, and offering appropriate support and guidance in these areas could be essential for promoting broader telemedicine adoption across diverse health care needs. The findings from this research can provide valuable insights for government agencies, telemedicine implementers, and policy makers. They can use this information to comprehend the critical factors influencing the behavior of prospective telemedicine users and craft precise strategies and policies to facilitate its widespread adoption effectively.

Acknowledgments

The authors would like to express their thanks to all the participants whose participation made this study possible.

Authors' Contributions

The conceptualization of the study was carried out by Y Lin, LPW, and ZH. LPW, Y Lin, XX, and Y Liu contributed to the development of the methodology. LPW, XX, and HA were responsible for the validation process. Formal analysis was conducted by LPW and HA. Data curation was handled by XX and Y Lin. The original draft of the manuscript was prepared by LPW and Y Lin. All authors reviewed and provided edits for the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File , 23 KB - jmir_v26i1e53497_app1.docx]

Multimedia Appendix 2
CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist.
[PDF File (Adobe PDF File), 140 KB - jmir_v26i1e53497_app2.pdf]

References

1. Gao J, Fan C, Chen B, Fan Z, Li L, Wang L, et al. Telemedicine is becoming an increasingly popular way to resolve the unequal distribution of healthcare resources: evidence from China. Front Public Health 2022;10:916303 [FREE Full text] [doi: 10.3389/fpubh.2022.916303] [Medline: 35874991]
2. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015;3(8):e445-e446 [FREE Full text] [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
3. Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine during the COVID-19 pandemic: experiences from Western China. J Med Internet Res 2020;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]
4. Snoswell CL, Taylor ML, Comans TA, Smith AC, Gray LC, Caffery LJ. Determining if telehealth can reduce health system costs: scoping review. J Med Internet Res 2020;22(10):e17298 [FREE Full text] [doi: 10.2196/17298] [Medline: 33074157]
5. Holtz B, Mitchell K, Hirko K, Ford S. Using the technology acceptance model to characterize barriers and opportunities of telemedicine in rural populations: survey and interview study. JMIR Form Res 2022;6(4):e35130 [FREE Full text] [doi: 10.2196/35130] [Medline: 35436207]
6. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]

XSL•FO
RenderX

7. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018;9(3):604-634 [FREE Full text] [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
8. Wejnert C, Pham H, Krishna N, Le B, DiNenno E. Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. AIDS Behav 2012;16(4):797-806 [FREE Full text] [doi: 10.1007/s10461-012-0147-8] [Medline: 22350828]
9. Chin WW. The partial least squares approach to structural equation modeling. In: Marcoulides G, editor. Modern Methods for Business Research. New York: Psychology Press; 1998:295-336.
10. Ringle C, Wende S, Becker JM. "SmartPLS 3." Boenningstedt: SmartPLS GmbH. 2015. URL: http://www.smartpls.com [accessed 2023-10-08]
11. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022;39(1):168-182 [FREE Full text] [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
12. Sorwar G, Rahamn MM, Uddin R, Hoque MR. Cost and time effectiveness analysis of a telemedicine service in Bangladesh. Stud Health Technol Inform 2016;231:127-134. [Medline: 27782024]
13. Langabeer JR, Champagne-Langabeer T, Alqusairi D, Kim J, Jackson A, Persse D, et al. Cost-benefit analysis of telehealth in pre-hospital care. J Telemed Telecare 2017;23(8):747-751. [doi: 10.1177/1357633X16680541] [Medline: 27913657]
14. Du Y, Zhou Q, Cheng W, Zhang Z, Hoelzer S, Liang Y, et al. Factors influencing adoption and use of telemedicine services in rural areas of China: mixed methods study. JMIR Public Health Surveill 2022;8(12):e40771 [FREE Full text] [doi: 10.2196/40771] [Medline: 36563026]
15. Bashshur RL, Shannon GW, Smith BR, Alverson DC, Antoniotti N, Barsan WG, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014;20(9):769-800 [FREE Full text] [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
16. Xie X, Liu P, Zheng Y, Zhou W, Zou J, Wang X, et al. Equity of health resource distribution in China during 2009–15: an analysis of cross-sectional nationwide data. Lancet 2017;390(Suppl 4):S6. [doi: 10.1016/s0140-6736(17)33144-6]
17. Liu J, Zhang Y. Health status and health disparity in China: a demographic and socioeconomic perspective. China Popul Dev Stud 2018;2(3):301-322 [FREE Full text] [doi: 10.1007/s42379-018-0015-y]
18. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
19. Qi M, Cui J, Li X, Han Y. Perceived factors influencing the public intention to use e-consultation: analysis of web-based survey data. J Med Internet Res 2021;23(1):e21834 [FREE Full text] [doi: 10.2196/21834] [Medline: 33470934]
20. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate Data Analysis. 7th Edition. Upper Saddle River, NJ: Pearson Prentice Hall; 2010.
21. Hair JF, Ringle CM, Sarstedt M. PLS-SEM: indeed a silver bullet. J Mark Theory Pract 2011;19(2):139-152. [doi: 10.2753/MTP1069-6679190202]
22. Byrne BM. Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming. 3rd Edition. London: Routledge; 2016.
23. Dall'Oglio AM, Rossiello B, Coletti MF, Caselli MC, Ravà L, Di Ciommo V, et al. Developmental evaluation at age 4: validity of an Italian parental questionnaire. J Paediatr Child Health 2010;46(7-8):419-426. [doi: 10.1111/j.1440-1754.2010.01748.x] [Medline: 20546104]
24. Dall'Oglio I, Nicolò R, Di Ciommo V, Bianchi N, Ciliento G, Gawronski O, et al. A systematic review of hospital foodservice patient satisfaction studies. J Acad Nutr Diet 2015;115(4):567-584. [doi: 10.1016/j.jand.2014.11.013] [Medline: 25634093]
25. Nguyen MC, Gabbe SG, Kemper KJ, Mahan JD, Cheavens JS, Moffatt-Bruce SD. Training on mind-body skills: feasibility and effects on physician mindfulness, compassion, and associated effects on stress, burnout, and clinical outcomes. J Posit Psychol 2019;15(2):194-207. [doi: 10.1080/17439760.2019.1578892]
26. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 2018;18(1):39-50. [doi: 10.1177/002224378101800104]

Abbreviations

AVE: average variance extracted
CR: composite reliability
PLS-SEM: partial least squares structural equation modeling
TAM: technology acceptance model

XSL•FO
RenderX

Edited by S Ma; submitted 09.10.23; peer-reviewed by D Dinh, A Gangadhara Rao; comments to author 28.02.24; revised version received 06.03.24; accepted 03.04.24; published 16.07.24.

Please cite as:
Lin Y, Xu X, Liu Y, Alias H, Hu Z, Wong LP
Perception and Acceptance of Telemedicine Use in Health Care Among the General Public in China: Web-Based Cross-Sectional Survey
J Med Internet Res 2024;26:e53497
URL: https://www.jmir.org/2024/1/e53497
doi:10.2196/53497
PMID:39012687

©Yulan Lin, Xiaonan Xu, Yiyang Liu, Haridah Alias, Zhijian Hu, Li Ping Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Smartwatch-Detected Arrhythmias in Patients After Transcatheter Aortic Valve Replacement (TAVR): Analysis of the SMART TAVR Trial

Jiaqi Fan[1], MD; Hanyi Dai[1], MD; Yuchao Guo[1], MD; Jianguo Xu[1], MD; Lihan Wang[1], MD; Jubo Jiang[1], MD; Xinping Lin[1], MD; Cheng Li[1], MD; Dao Zhou[1], MD; Huajun Li[1], MD; Xianbao Liu[1], MD; Jian'an Wang[1], MD

department of Cardiology, Zhejiang University, Hangzhou, China

Corresponding Author:
Jian'an Wang, MD
department of Cardiology
Zhejiang University
Jiefang Road 88
Hangzhou, 310000
China
Phone: 86 15267029492
Email: wangjianan111@zju.edu.cn

Abstract

Background: There are limited data available on the development of arrhythmias in patients at risk of high-degree atrioventricular block (HAVB) or complete heart block (CHB) following transcatheter aortic valve replacement (TAVR).

Objective: This study aimed to explore the incidence and evolution of arrhythmias by monitoring patients at risk of HAVB or CHB after TAVR using smartwatches.

Methods: We analyzed 188 consecutive patients in the prospective SMART TAVR (smartwatch-facilitated early discharge in patients undergoing TAVR) trial. Patients were divided into 2 groups according to the risk of HAVB or CHB. Patients were required to trigger a single-lead electrocardiogram (ECG) recording and send it to the Heart Health App via their smartphone. Physicians in the central ECG core lab would then analyze the ECG. The incidence and timing of arrhythmias and pacemaker implantation within a 30-day follow-up were compared. All arrhythmic events were adjudicated in a central ECG core lab.

Results: The mean age of the patients was 73.1 (SD 7.3) years, of whom 105 (55.9%) were men. The mean discharge day after TAVR was 2.0 (SD 1.8) days. There were no statistically significant changes in the evolution of atrial fibrillation or atrial flutter, Mobitz I, Mobitz II, and third-degree atrial ventricular block over time in the first month after TAVR. The incidence of the left bundle branch block (LBBB) increased in the first week and decreased in the subsequent 3 weeks significantly (P<.001). Patients at higher risk of HAVB or CHB received more pacemaker implantation after discharge (n=8, 9.6% vs n=2, 1.9%; P=.04). The incidence of LBBB was higher in the group with higher HAVB or CHB risk (n=47, 56.6% vs n=34, 32.4%; P=.001). The independent predictors for pacemaker implantation were age, baseline atrial fibrillation, baseline right bundle branch block, Mobitz II, and third-degree atrioventricular block detected by the smartwatch.

Conclusions: Except for LBBB, no change in arrhythmias was observed over time in the first month after TAVR. A higher incidence of pacemaker implantation after discharge was observed in patients at risk of HAVB or CHB. However, Mobitz II and third-degree atrioventricular block detected by the smartwatch during follow-ups were more valuable indicators to predict pacemaker implantation after discharge from the index TAVR.

Trial Registration: ClinicalTrials.gov NCT04454177; https://clinicaltrials.gov/study/NCT04454177

(J Med Internet Res 2024;26:e41843) doi:10.2196/41843

KEYWORDS

arrhythmias; transcatheter aortic valve replacement; smartwatch; ambulatory electrocardiography; smartphone; mobile phone

XSL•FO
RenderX

Introduction

Transcatheter aortic valve replacement (TAVR) has been established as a preferred treatment for patients with symptomatic severe aortic stenosis with overall risk profiles. However, there are some drawbacks to this procedure, such as the new-onset rhythm and conduction disturbance (CD). The management of arrhythmias and CDs is still challenging even with tailored postoperative management, according to the recent scientific expert panel [1,2]. The scientific expert panel proposed an algorithm based on baseline and post-TAVR electrocardiography within the index hospitalization [1]. Patients were categorized into groups suitable for early discharge (within 48 hours after TAVR), groups at high risk of high-degree atrioventricular block (HAVB) or complete heart block (CHB), and groups in need of pacemaker implantation before discharge [1]. However, this tailored prespecified algorithm falls short in accurately identifying patients at high risk of HAVB or CHB [2]. Therefore, the adoption of ambulatory electrocardiogram (ECG) monitoring in a specific subset of these patients will enhance patient-centered management after discharge. The 2019 algorithm recommends using 30-day ambulatory ECG monitoring in TAVR recipients [1]. Recently, 2 studies using the patch and implantable loop recorders observed a considerable rate of delayed HAVB or CHB after TAVR [3,4]. A smartwatch, a type of ambulatory ECG monitoring, enables single-lead ECG checks after activation by patients, showing enough efficacy and safety in patients who had a TAVR [5]. To date, no study has evaluated the incidence and evolution of arrhythmias and CDs after TAVR using a noninvasive and convenient smartwatch.

In this study, we aim to explore the incidence and evolution of arrhythmias and CDs detected by smartwatches over time in patients at high risk of HAVB or CHB following TAVR within a 30-day follow-up.

Methods

This was a single-center, prospective trial, called the SMART TAVR (smartwatch-facilitated early discharge in patients undergoing TAVR) trial, registered with the ClinicalTrials.gov registry (NCT04454177).

Ethical Considerations

Informed consent was provided by all patients to participate in the study before the procedures. The protocol was approved by the medical ethics committee of the Second Affiliated Hospital of Zhejiang University (IR2020001223). This study was conducted according to the Declaration of Helsinki and the Good Clinical Practice guidelines. The data were anonymous and deidentified for every participant.

Patient Population

Patients with symptomatic severe aortic stenosis who had received the elective transfemoral TAVR in our hospital were consecutively included in the study. Severe aortic stenosis was defined as a median transvalvular gradient of more than 40 mmHg, a peak jet velocity greater than 4 meters per second, or an initial aortic valve area less than 1 cm^2. Exclusion criteria included severe complications of TAVR (eg, death, conversion to surgical aortic valve replacement, or unplanned cardiopulmonary resuscitation) during the procedure, life expectancy less than 12 months, severe dementia (ie, unable to sign for informed consent, take care of themselves, or communicate with the research team during a study visit). In addition, patients with pacemaker implantation at baseline and predischarge were also excluded.

Participants were classified into groups of high and low risk of HAVB or CHB based on our conduction management protocol in ECG. The incidence and time of arrhythmias and pacemaker implantation within a 30-day follow-up were compared.

Algorithm for Patients at High Risk of HAVB or CHB

Twelve-lead ECGs were recorded and analyzed at baseline, immediate postoperative period, 4 hours and 24 hours after the procedure, on a daily basis, and thereafter, during the index hospitalization, if needed. Higher risk of HAVB or CHB was defined as the occurrence of at least one of the following: (1) increased PR or QRS interval ≥20 milliseconds within the last two 12-lead ECGs before discharge; (2) PR ≥240 millisecond or QRS ≥150 milliseconds (QRS ≥140 ms for patients with atrial fibrillation) in the last 12-lead ECG before discharge; and (3) transient or persisted HAVB or CHB occurring within the 12-lead ECGs mentioned above. Our CD management protocol was described in detail in the supplementary methods in Multimedia Appendix 1. This protocol was modified based on the previously published Expert Consensus Algorithm for the management of CDs in 2019 [1].

Smartwatch Equipment and Monitoring Frequency

The enrolled patients were provided with a smartwatch within 24 hours before the scheduled TAVR procedure and were required not to share their smartwatches with anyone else. HUAWEI Watch GT series could continuously monitor and record multiple biometric parameters, including heart rate, step counts, sleep cycles, RR interval, and pulse oxygen saturation. Moreover, these devices could record the single-lead ECG upon activation and analyze QRS complexes and P waves. Patients were required to activate the watch to record ECG twice per day in the week following TAVR discharge and at least 2 days a week for the rest of the month, as described in a previous study [6]. The duration of the recording was at least 30 seconds, and the recordings were applied in the morning and afternoon each day. Moreover, patients were required to upload smartwatch readings at the onset of any cardiovascular symptoms, including dyspnea, chest pain, palpitations, dizziness, or presyncope. The data recorded by HUAWEI Watch were transmitted to the HUAWEI phone app. The Heart Health App (developed by Second Affiliated Hospital of Zhejiang University) would then receive, process, and store the data and transfer it to the remote database after securing patients' approval. A designated heart team member would access and download the data via a cloud database. All arrhythmic events were adjudicated in a central ECG core lab.

XSL•FO
RenderX

Data Collection and Definitions

Baseline clinical, electrocardiographic, imaging, and periprocedural characteristics data were collected. The single-lead ECG recording arrhythmias were detected by the smartwatch and evaluated by the central ECG core lab. A persistent left bundle branch block (LBBB) was defined as an LBBB that persisted during the whole 30-day follow-up period. All living patients were assessed face to face in the 30-day follow-up. All TAVR-related variables and outcomes were defined according to the Valve Academic Research Consortium-2 criteria [7]. Patients were categorized into a group at higher risk of HAVB or CHB and a group at lower risk of HAVB or CHB who were eligible for next-day discharge. Variables and outcomes were compared between the two groups. All data were stored in the TORCH registry and SMART TAVR registry databases and were trackable.

Statistical Analysis

Continuous variables were presented as median (SD) values. Categorical data were shown as counts (percentages). Student *t* test (2-tailed), Mann-Whitney *U* test, or Kruskal-Wallis test were used for continuous variables, while chi-square or Fisher exact test were used for categorical data. Predictors of pacemaker implantation were assessed by Cox regression models. Variables with $P<.1$ in the univariate analysis and commonly reported predictor factors in previous studies were included in the multivariate stepwise regression model. Kaplan-Meier estimates were calculated to describe the probability of pacemaker implantation with the date of TAVR as a starting point according to the risk of HAVB or CHB classification. *P* values less than .05 were considered statistically significant.

Results

After excluding 7 patients who had the pacemaker implantation at baseline and 16 at predischarge, a total of 188 consecutive patients who underwent elective transfemoral TAVR between July 16, 2020, and December 31, 2021, were enrolled. No patients withdrew from the study within the 30-day follow-up. The mean age of the patients was 73.1 (SD 7.3) years, of whom 105 (55.9%) were men, and the mean discharge days after TAVR was 2.0 (SD 1.8) days. The baseline prevalence of atrial fibrillation or flutter seemed to be higher in the group at higher risk of HAVB or CHB compared to the group at lower risk of HAVB or CHB, without statistical significance (n=15, 18.1% vs n=9, 8.6%; $P=.05$). The baseline and procedural characteristics of the study population are presented in Table 1.

Two patients died within the 30-day follow-up in the group with higher HAVB or CHB risk. The patients at higher risk of HAVB or CHB received more pacemaker implantation after discharge within the 30-day follow-up (n=8, 9.6% vs n=2, 1.9%; $P=.04$). The Kaplan-Meier analysis of pacemaker implantation after discharge between the two groups is shown in Figure 1. Up to 80% of patients received the pacemaker implantation within 14 days after TAVR, and the mean time to pacemaker implantation after TAVR was similar between these two groups (mean 10.8, SD 7.2 days vs mean 9.0, SD 7.1 days; $P=.71$). For the arrhythmias detected by the smartwatch, the incidence of LBBB was higher in the group at higher HAVB or CHB risk (n=47, 56.6% vs n=34, 32.4%; $P=.001$), and more persistent LBBB was also observed in the same group (n=24, 28.9% vs n=10, 9.5%; $P=.001$). The incidence of Mobitz I, Mobitz II, and third-degree atrioventricular block (AVB) in the group with a higher risk of HAVB or CHB seemed to be higher in numerical terms compared to the group with lower HAVB or CHB risk, without statistical significance (Mobitz I: n=4, 5% vs n=1, 1%; $P=.26$; Mobitz II: n=3, 3.8% vs n=1, 1%; $P=.44$; third-degree AVB: n=4, 5% vs n=1, 1%; $P=.23$). Other arrhythmias detected by the smartwatch were comparable between these two groups. No differences were observed between these two groups in echocardiographic data during the 30-day follow-up. All outcomes during the 30-day follow-up are presented in Table 2.

There were no statistically significant changes in the evolution of atrial fibrillation or atrial flutter, Mobitz I, Mobitz II, and third-degree AVB during the first month after TAVR. The incidence of LBBB increased in the first week and decreased in the subsequent 3 weeks significantly ($P<.001$). Consistently, the QRS duration increased in the first week and decreased significantly afterward ($P<.001$). Moreover, the incidence of LBBB was higher in the higher-risk HAVB or CHB group when compared with the lower-risk group at any given time. The detailed information about different arrhythmias over time in the two groups is shown in Table S1 in Multimedia Appendix 1. The evolution of arrhythmias during the first 7 days and the first 4 weeks after TAVR is presented in Figure S1 in Multimedia Appendix 1, and the incidences of arrhythmias in total, in the higher-risk and lower-risk HAVB or CHB groups are presented in Figure S2 in Multimedia Appendix 1.

The predictors for pacemaker implantation after discharge were baseline right bundle branch block (RBBB), higher risk of HAVB or CHB, Mobitz II, and third-degree AVB detected by smartwatch in the univariate analysis. The independent predictors after discharge for pacemaker implantation were age, baseline atrial fibrillation, baseline RBBB, Mobitz II, and third-degree AVB detected by the smartwatch. The Cox regression model for pacemaker implantation after discharge within the 30-day follow-up is presented in Table 3.

XSL•FO
RenderX

Table 1. Baseline, electrocardiographic, imaging, and procedural characteristics of the study population, according to the high-degree atrioventricular block (HAVB) or complete heart block (CHB) risk.

Characteristics	Overall (N=188)	Higher risk of HAVB or CHB[a] (n=83)	Lower risk of HAVB or CHB (n=105)	*P* value
Age (years), mean (SD)	73.1 (7.3)	73.9 (8.1)	72.4 (6.5)	.17
Gender (male), n (%)	105 (55.9)	46 (55.4)	59 (56.2)	.92
BMI (kg/m^2), mean (SD)	23.40 (3.57)	23.63 (3.62)	23.21 (3.53)	.43[b]
STS[c] (%), mean (SD)	3.78 (3.07)	4.30 (3.62)	3.40 (2.52)	.13[b]
Hypertension, n (%)	95 (51.6)	45 (54.2)	50 (49.5)	.52
Diabetes mellitus, n (%)	38 (20.4)	20 (24.1)	18 (17.5)	.27
Prior stroke, n (%)	7 (3.7)	5 (6)	2 (1.9)	.27
Atrial fibrillation or flutter, n (%)	24 (12.8)	15 (18.1)	9 (8.6)	.05
Echocardiographic and computed tomographic variables , mean (SD)				
LVEF[d] (%)	58.8 (11.2)	58.0 (12.2)	59.4 (10.4)	.51[b]
Max velocity (ms)	4.80 (0.74)	4.86 (0.81)	4.76 (0.67)	.50[b]
Mean gradient (mmHg)	54.8 (18.5)	55.9 (20.1)	53.9 (17.2)	.66[b]
AVA[e] (cm^2)	0.66 (0.23)	0.62 (0.24)	0.70 (0.22)	.01[b]
Perimeter derived diameter (mm)	25 (4)	25.1 (5.4)	24.9 (2.4)	.65
Electrocardiographic variables, n (%)				
Preexisting atrial fibrillation	23 (12.6)	12 (15)	11 (10.7)	.38
Preexisting RBBB[f]	13 (7.1)	7 (8.6)	6 (5.8)	.46
Preexisting LBBB[g]	9 (4.9)	7 (8.6)	2 (1.9)	.08
Increased PR or QRS interval ≥20 ms	112 (62.9)	56 (73.7)	56 (54.6)	.01
PR ≥240 ms or QRS ≥150 ms[h]	69 (36.7)	42 (50.6)	27 (25.7)	<.001
Periprocedural characteristics				
Predilatation, n (%)	171 (93.4)	74 (91.4)	97 (95.1)	.31
Postdilatation, n (%)	120 (65.9)	52 (65)	68 (66.7)	.81
Self-expanding valve, n (%)	160 (88.4)	71 (88.8)	89 (88.1)	.90
Prosthetic valve size >26mm, n (%)	134 (74)	56 (70)	78 (77.2)	.27
Oversizing ratio, mean (SD)	1.06 (0.09)	1.06 (0.11)	1.06 (0.07)	.69
Days discharged after TAVR[i] (days), mean (SD)	2 (1.8)	1 (0)	3.3 (2.1)	<.001[b]

[a]Higher risk of HAVB or CHB was defined by our conduction management protocol in electrocardiogram (supplementary methods in Multimedia Appendix 1).

[b]Mann-Whitney *U* test was used.

[c]STS: Society of Thoracic Surgeons.

[d]LVEF: left ventricular ejection fraction.

[e]AVA: aortic valve area.

[f]RBBB: right bundle branch block.

[g]LBBB: left bundle branch block.

[h]QRS ≥140 ms for patients with atrial fibrillation.

[i]TAVR: transcatheter aortic valve replacement.

XSL•FO
RenderX

Figure 1. Kaplan-Meier analysis of permanent pacemaker implantation (PPMI) after discharge between groups with higher and lower risk of high-degree atrioventricular block (HAVB) or complete heart block (CHB). HR: heart rate; TAVR: transcatheter aortic valve replacement.

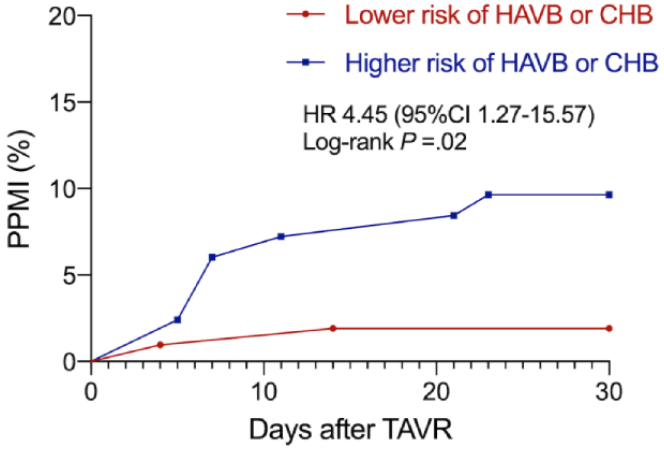

XSL•FO
RenderX

Table 2. thirty days outcomes of the study population, according to the high-degree atrioventricular block (HAVB) or complete heart block (CHB) risk group.

Characteristics	Overall (N=188)	Higher risk of HAVB or CHB[a] (n=83)	Lower risk of HAVB or CHB (n=105)	*P* value
Mortality, n (%)	2 (1.1)	2 (2.4)	0 (0)	.19
PPMI[b] after discharge, n (%)	10 (5.3)	8 (9.6)	2 (1.9)	.04
Time to PPMI since TAVR[c], mean (SD)	10.4 (6.8)	10.8 (7.2)	9.0 (7.1)	.60[d]
PPMI within 14 days since TAVR, n (%)	8 (80)	6 (75)	2 (100)	>.99
Arrhythmias detected by smartwatch, n (%)				
Bradyarrhythmia	84 (44.7)	48 (57.8)	36 (34.3)	.002
LBBB[e]	81 (43.1)	47 (56.6)	34 (32.4)	.001
Persistent LBBB	34 (18.1)	24 (28.9)	10 (9.5)	.001
Mobitz I	5 (2.7)	4 (5)	1 (1)	.23
Mobitz II	4 (2.2)	3 (3.8)	1 (1)	.44
Third AVB[f]	5 (2.7)	4 (5)	1 (1)	.23
Tachyarrhythmia	33 (17.9)	17 (21.3)	16 (15.4)	.30
Atrial fibrillation or atrial flutter	33 (17.9)	17 (21.3)	16 (15.4)	.30
ventricular tachycardia	0 (0)	0 (0)	0 (0)	—[g]
PVST[h]	0 (0)	0 (0)	0 (0)	—
Significant pause (>5 s)	7 (4.1)	4 (5.4)	3 (3.1)	.72
Echocardiographic data, mean (SD)				
LVEF[i] (%)	60.5 (9.4)	58.9 (11.8)	61.5 (7.3)	.57[d]
Max velocity (ms)	2.31 (0.51)	2.30 (0.50)	2.31 (0.52)	.81[d]
Mean gradient (mmHg)	11.3 (5.1)	11.3 (4.6)	11.3 (5.4)	.83[d]
AVA[j] (cm^2)	1.67 (0.43)	1.66 (0.45)	1.68 (0.42)	.42[d]

[a]Higher risk of HAVB/CHB was defined by our conduction management protocol in electrocardiogram (supplementary methods in Multimedia Appendix 1).

[b]PPMI: pacemaker implantation.

[c]TAVR: transcatheter aortic valve replacement.

[d]Mann-Whitney *U* test was used.

[e]LBBB: left bundle branch block.

[f]AVB: atrioventricular block.

[g]Not applicable.

[h]PVST: paroxysmal supraventricular tachycardia.

[i]LVEF: left ventricular ejection fraction.

[j]AVA: aortic valve area.

XSL•FO
RenderX

Table 3. Cox regression models for pacemaker implantation after discharge within 30-day follow-up.

Variables before discharge	Univariate regression		Multivariate regression	
	β (SE)	*P* value	β (SE)	*P* value
Age	0.023 (0.044)	.61	–0.106 (0.047)	.03
Gender (male)	0.621 (0.690)	.37	__[a]	—
Baseline atrial flutter or atrial fibrillation	1.128 (0.690)	.10	2.923 (0.999)	.003
Baseline RBBB[b]	1.821 (0.691)	.01	3.395 (1.079)	.002
Baseline LBBB[c]	0.832 (1.054)	.43	—	—
Predilatation	0.500 (1.054)	.64	—	—
Self-expanding valve	0.663 (0.791)	.40	—	—
Oversizing ratio	1.421 (2.913)	.63	—	—
Postdilatation	1.102 (0.646)	.09	—	—
Higher risk of HAVB[d] or CHB[e]	1.648 (0.791)	.04	—	—
Atrial fibrillation or atrial flutter[f]	1.168 (0.646)	.07	—	—
LBBB[f]	0.272 (0.632)	.67	—	—
Mobitz I[f]	1.445 (1.054)	.17	—	—
Mobitz II[f]	3.155 (0.697)	<.001	4.523 (1.160)	<.001
Third AVB[f,g]	2.849 (0.693)	<.001	2.898 (0.971)	.003

[a]Not applicable.

[b]RBBB: right bundle branch block.

[c]LBBB: left bundle branch block.

[d]HAVB: high-degree atrioventricular block.

[e]CHB: complete heart block. Higher risk of HAVB or CHB was defined by our conduction management protocol in electrocardiogram supplementary methods in Multimedia Appendix 1).

[f]Arrhythmias after discharge detected by the smartwatch.

[g]AVB: atrioventricular block.

Discussion

Principal Findings

The main findings of this study are summarized as follows: (1) during the 30-day follow-up, patients at higher risk of HAVB or CHB received more pacemaker implantation after discharge compared to patients at lower risk of HAVB or CHB who were eligible for next-day discharge; (2) the incidence of LBBB was higher in patients at higher risk of HAVB or CHB while the incidence of other arrhythmias and CDs seemed similar in these two groups; (3) the age, baseline atrial fibrillation, baseline RBBB, Mobitz II, and third-degree AVB detected by the smartwatch were independent predictors for pacemaker implantations after discharge from the index TAVR hospitalization.

Data from the Healthcare Cost and Utilization Project Nationwide Readmission Database reported an increase from 5.7% to 13% in the proportion of pacemaker implantation in patients receiving TAVR after the promotion of minimalist TAVR and early discharge strategies in recent years [8]. The increasing trend of pacemaker implantation was likely to be driven by the reduced length of stay after TAVR, and hence, a risk stratification and monitoring protocol was urgently needed. However, the data using the ambulatory ECG monitoring to observe the proportion of pacemaker implantation were scarce. In their study, Tian et al [9] performed a remote monitoring system—BodyGuardian (Preventice Technologies)—in patients who received the TAVR procedure and found within 30 days after discharge, 8.7% of patients received pacemaker implantation, of which 7.1% were for symptomatic second or third-degree AVB and 1.6% were for symptomatic sinus node dysfunction after discharge within the 30-day follow-up [9]. A recent study by Muntane-Carol et al [3] also performed the systematic 2-week ambulatory ECG monitoring (CardioSTAT [Icentia] or Zio AT Patch [iRhythm Technologies]) following minimalist TAVR and detected an incidence of 2.2% for pacemaker implantation in the group with no ECG changes, an incidence of 6.6% in the new ECG CD group, and an incidence of 13.2% in the baseline RBBB group [3]. Similarly, Okoh et al [10] observed an incidence of 0%, 4%, and 8.5% for pacemaker implantation in group I (normal pre-TAVR, periprocedural, and discharge ECGs), group II (normal pre-TAVR and abnormal subsequent ECGs), and group III (abnormal baseline and abnormal subsequent ECGs), respectively, with a real-time home continuous ECG system

XSL•FO
RenderX

(Zio AT Patch [iRhythm Technologies]). The overall incidence of pacemaker implantation in our study was 5.3%, while the incidence in the higher-risk HAVB or CHB group was up to 9.6%. This was similar to previous studies with ambulatory ECG monitoring after TAVR. Our data also showed that 80% of pacemakers were implanted within 14 days after TAVR, similar to the results reported by Mazzella et al [8]. The rate of pacemaker implantation after discharge was similar in these studies, whether or not ambulatory ECGs were used. However, the rate increased in the specific group of patients with a higher risk of HAVB or CHB according to the modified protocol in our study (similar to the new ECG CDs group or baseline RBBB group in the study by Muntane-Carol et al [3]. Therefore, we can draw the inference that ambulatory ECG use does not increase the rate of pacemaker implantation after discharge, but it can monitor the arrhythmias or CDs in higher-risk patients who may be more likely to receive the pacemaker implantation on the discharging day. However, this conclusion should be interpreted cautiously, as our study did not compare the incidence of arrhythmic events monitored with or without smartwatches.

In a recent study by Reiter et al [4], 59 consecutive patients who received implantable loop recorders were enrolled; an LBBB incidence of 28.8% on the first day after TAVR, 32.2% on the second day, and 33.9% on the third day were observed, while the incidence decreased to 25.9% at the first-month follow-up. The incidence of LBBB in the above-mentioned study was higher than that in our study, which might be due to the higher oversizing ratio of the implanted prosthetic valve. Consistent with our previous pilot research data, the prevalence of LBBB increased within the first week after TAVR and decreased in the subsequent weeks following the TAVR, which was in line with the resolution of edema and inflammation in the area of the His bundle and the left bundle branch [11]. The prevalence of LBBB was up to 39.4% in the whole cohort, while the prevalence was 51.8% in patients at a higher risk of HAVB or CHB. Urena et al [12] and Muntane-Carol et al [13] reported that persistent LBBB may result in more HAVB or CHB episodes requiring pacemaker implantation. In our study, only 2 of the 10 patients requiring pacemaker implantation developed new-onset persistent LBBB, while 3 had transient LBBB (with 2 patients developing the LBBB on the first day after TAVR and 1 developing it on the third day). Moreover, in our study, we observed a higher prevalence of LBBB in patients at risk of HAVB or CHB. Nearly half a percentage of LBBB was persistent LBBB in this group, while only one-third of LBBB was persistent in patients eligible for next-day discharge.

Of the 44.1% of patients classified as being at higher risk of HAVB or CHB, 9.6% ultimately received the pacemaker implantation after discharge within the 30-day follow-up. Most of the pacemakers were implanted within 14 days after the TAVR procedure. However, 2 patients in the group with a higher risk of HAVB or CHB received the pacemaker implantation on the 21st and 23rd days post TAVR for symptomatic significant pause (atrial flutter and LBBB in the previous single ECG) and delayed HAVB, respectively. Our study indicated that the intensive remote monitoring of arrhythmias and CDs within the first 14 days after TAVR was very necessary, and the monitor continued to play a critical role in the subsequent days of the first months. It was reasonable that the Mobitz II and third-degree AVB detected by the smartwatch were independent predictors of pacemaker implantation. Additionally, age, baseline atrial fibrillation, and baseline RBBB were independent predictors for pacemaker implantation. However, the study did not observe baseline LBBB, self-expanding valve, and prosthesis oversizing ratio as independent predictors of pacemaker implantation after discharge as most commonly reported [14]. Although clinical judgment evaluating patients at a higher risk of HAVB or CHB was recommended by expert consensus, it was still challenging to identify this population [2]. In our previous case, 1 patient who was eligible for next-day discharge received the pacemaker implantation after discharge for symptomatic delayed HAVB detected by the smartwatch [15]. Therefore, additional strategies for refining the risk assessment are warranted [2]. Our strategy of arrhythmias monitored by a smartwatch during follow-ups may provide more information, proving to be a better predictor for pacemaker implantation after discharge.

Although a significant portion of the technology is still in its nascent stages, encompassing numerous technical challenges, such as data handling, analysis, storage, security, and privacy, it is imperative to foster collaboration between the health and technical domains, involving a diverse range of experts with varied skills and domain knowledge [6]. Engaging health care professionals in smart technology research will not only ensure the relevance of these tools but also enhance their successful implementation.

Study Limitations

First, this was a single-center, nonrandomized cohort study. There may be some bias since nearly all enrolled patients were in good medical compliance, while 1 patient with asymptomatic HAVB declined the pacemaker implantation. The SMART TAVR study involved patients who had undergone a TAVR procedure, which likely contributed to higher compliance levels, driven by their postprocedural monitoring and psychological care requirements [5]. Moreover, the small number of TAVR recipients with balloon-expandable prosthetic valves (11.6%) may interfere with the results. Additionally, numerous ongoing studies are currently investigating the use of smart wearable devices for monitoring atrial fibrillation. There is a noticeable gap in research regarding the detection of bradyarrhythmia. Finally, the limited number of pacemaker implantation cases may influence the independent predictor evaluation.

Conclusions

Except for LBBB, no change in arrhythmias was observed over time in the first month after TAVR. Patients at a high risk of HAVB or CHB received more pacemaker implantation after discharge. More prevalence of LBBB was observed in the HAVB or CHB group compared with the patients who were eligible for next-day discharge while the other arrhythmias were similar between these two groups. Smartwatch-monitored arrhythmias during follow-up were valuable factors in predicting pacemaker implantation after discharge from the index TAVR.

XSL•FO
RenderX

Acknowledgments

The authors sincerely thank the Huawei Heart Health Research Team for the development and optimization of the ECG algorithm, headed by Mr Xiaoxiang He. Team members include Jiabing Yan, Wenjuan Chen, Lian Wu, and Jianhua Guo.

Data Availability

The data sets generated during and/or analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Authors' Contributions

JF, X Liu, and J Wang contributed to the study concept and design. JF and HD did the statistical analyses. JF wrote the manuscript. Data collection was conducted by JF, HD, YG, JX, L Wang, JJ, X Lin, CL, and DZ. Critical revision of the manuscript was done by JF, X Liu, and J Wang. JF accepts full responsibility for the conduct of the study and has access to the data. All authors controlled the decision to publish.

Conflicts of Interest

J Wang and X Liu have been consultants at Venus MedTech. All other authors report no conflicts of interest.

Multimedia Appendix 1
Additional information and statistics.
[DOCX File , 3520 KB - jmir_v26i1e41843_app1.docx]

References

1. Puri R, Rodés-Cabau J. Transcatheter aortic valve replacement: a revolution in evolution. JACC Cardiovasc Interv 2016 Feb 22;9(4):364-366 [FREE Full text] [doi: 10.1016/j.jcin.2015.12.019] [Medline: 26803421]
2. Malebranche D, Bartkowiak J, Ryffel C, Bernhard B, Elsmaan M, Nozica N, et al. Validation of the 2019 expert consensus algorithm for the management of conduction disturbances after TAVR. JACC Cardiovasc Interv 2021 May 10;14(9):981-991 [FREE Full text] [doi: 10.1016/j.jcin.2021.03.010] [Medline: 33958172]
3. Muntané-Carol G, Okoh AK, Chen C, Nault I, Kassotis J, Mohammadi S, et al. Ambulatory electrocardiographic monitoring following minimalist transcatheter aortic valve replacement. JACC Cardiovasc Interv 2021 Dec 27;14(24):2711-2722 [FREE Full text] [doi: 10.1016/j.jcin.2021.08.039] [Medline: 34949396]
4. Reiter C, Lambert T, Kellermair J, Blessberger H, Fellner A, Nahler A, et al. Delayed total atrioventricular block after transcatheter aortic valve replacement assessed by implantable loop recorders. JACC Cardiovasc Interv 2021 Dec 27;14(24):2723-2732 [FREE Full text] [doi: 10.1016/j.jcin.2021.09.003] [Medline: 34949397]
5. Liu X, Fan J, Guo Y, Dai H, Xu J, Wang L, et al. JAHA 2022 Apr 05;11(7):e023219. [doi: 10.1161/jaha.121.023219]
6. Morita PP, Sahu KS, Oetomo A. Health monitoring using smart home technologies: scoping review. JMIR Mhealth Uhealth 2023 Apr 13;11:e37347. [doi: 10.2196/37347]
7. Kappetein AP, Head SJ, Généreux P, Piazza N, van Mieghem NM, Blackstone EH, Valve Academic Research Consortium (VARC)-2. Updated standardized endpoint definitions for transcatheter aortic valve implantation: the Valve Academic Research Consortium-2 consensus document (VARC-2). Eur J Cardiothorac Surg 2012 Nov 01;42(5):S45-S60 [FREE Full text] [doi: 10.1093/ejcts/ezs533] [Medline: 23026738]
8. Mazzella AJ, Hendrickson MJ, Arora S, Sanders M, Li Q, Vavalle JP, et al. Shifting trends in timing of pacemaker implantation after transcatheter aortic valve replacement. JACC Cardiovasc Interv 2021 Jan 25;14(2):232-234 [FREE Full text] [doi: 10.1016/j.jcin.2020.09.034] [Medline: 33183993]
9. Tian Y, Padmanabhan D, McLeod CJ, Zhang P, Xiao P, Sandhu GS, et al. Utility of 30-day continuous ambulatory monitoring to identify patients with delayed occurrence of atrioventricular block after transcatheter aortic valve replacement. Circ: Cardiovascular Interventions 2019 Dec;12(12):e007635. [doi: 10.1161/circinterventions.118.007635]
10. Okoh AK, Chen C, Kassotis J, Soliman F, Patel K, Sohitliv A, et al. Utilizing ambulatory electrocardiogram monitoring to reduce conduction related death after transcatheter aortic valve replacement. Catheter Cardiovasc Interv 2022 Mar;99(4):1243-1250. [doi: 10.1002/ccd.30027] [Medline: 34851550]
11. Liu X, Fan J, Guo Y, Chen Y, Li C, Jiang J, et al. Smart watch measured conduction disturbance after transcatheter aortic valve replacement. J Am Coll Cardiol 2021 May;77(18):1136. [doi: 10.1016/s0735-1097(21)02495-5]
12. Urena M, Mok M, Serra V, Dumont E, Nombela-Franco L, DeLarochellière R, et al. Predictive factors and long-term clinical consequences of persistent left bundle branch block following transcatheter aortic valve implantation with a balloon-expandable valve. J Am Coll Cardiol 2012 Oct 30;60(18):1743-1752 [FREE Full text] [doi: 10.1016/j.jacc.2012.07.035] [Medline: 23040577]

XSL•FO
RenderX

13. Muntané-Carol G, Nombela-Franco L, Serra V, Urena M, Amat-Santos I, Vilalta V, et al. Late arrhythmias in patients with new-onset persistent left bundle branch block after transcatheter aortic valve replacement using a balloon-expandable valve. Heart Rhythm 2021 Oct;18(10):1733-1740. [doi: 10.1016/j.hrthm.2021.05.031] [Medline: 34082083]
14. Sammour Y, Krishnaswamy A, Kumar A, Puri R, Tarakji KG, Bazarbashi N, et al. Incidence, predictors, and implications of permanent pacemaker requirement after transcatheter aortic valve replacement for patients undergoing transcatheter aortic valve replacement. JACC Cardiovasc Interv 2021 Jan 25;14(2):115-134 [FREE Full text] [doi: 10.1016/j.jcin.2020.09.063] [Medline: 33478630]
15. Fan J, Liu X, Xu Y, Xu J, Wang J. Smartwatch detected high-degree atrioventricular block requiring pacemaker implantation after transcatheter aortic valve replacement during COVID-19. Eur Heart J Digit Health 2021 Sep;2(3):545-546 [FREE Full text] [doi: 10.1093/ehjdh/ztab051] [Medline: 36713599]

Abbreviations

AVB: atrioventricular block
CD: conduction disturbance
CHB: complete heart block
ECG: electrocardiogram
HAVB: high-degree atrioventricular block
LBBB: left bundle branch block
RBBB: right bundle branch block
SMART TAVR: smartwatch-facilitated early discharge in patients undergoing TAVR
TAVR: transcatheter aortic valve replacement

Edited by A Mavragani; submitted 11.08.22; peer-reviewed by D Filos, M Alasnag, A Cobo; comments to author 20.07.23; revised version received 28.09.23; accepted 02.05.24; published 19.07.24.

Please cite as:
Fan J, Dai H, Guo Y, Xu J, Wang L, Jiang J, Lin X, Li C, Zhou D, Li H, Liu X, Wang J
Smartwatch-Detected Arrhythmias in Patients After Transcatheter Aortic Valve Replacement (TAVR): Analysis of the SMART TAVR Trial
J Med Internet Res 2024;26:e41843
URL: https://www.jmir.org/2024/1/e41843
doi:10.2196/41843
PMID:

©Jiaqi Fan, Hanyi Dai, Yuchao Guo, Jianguo Xu, Lihan Wang, Jubo Jiang, Xinping Lin, Cheng Li, Dao Zhou, Huajun Li, Xianbao Liu, Jian'an Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Use of Online Consultation Systems and Patient Experience of Primary Care: Cross-Sectional Analysis Using the General Practice Patient Survey

Xiaochen Ge[1,2], MSc; Paul Chappell[1,2], PhD; Jean Ledger[2], PhD; Minal Bakhai[2], MBBS; Geraldine M Clarke[1], PhD

[1]Improvement Analytics Unit, The Health Foundation, London, United Kingdom
[2]NHS England, London, United Kingdom

Corresponding Author:
Geraldine M Clarke, PhD
Improvement Analytics Unit
The Health Foundation
8 Salisbury Square
London, EC4Y 8AP
United Kingdom
Phone: 44 207257 ext 8000
Email: geraldine.clarke@health.org.uk

Abstract

Background: NHS England encourages the use of online consultation (OC) systems alongside traditional access methods for patients to contact their general practice online and for practices to manage workflow. Access is a key driver of patients' primary care experience. The use of online technology and patient experience vary by sociodemographic characteristics.

Objective: This study aims to assess the association between OC system use and patient experience of primary care in English general practice and how that varies by OC system model and practice sociodemographic characteristics (rurality, deprivation, age, and ethnicity).

Methods: We categorized practices as "low-use" or "high-use" according to the volume of patient-initiated contacts made via the OC system. We considered practices using one of 2 OC systems with distinct designs and implementation models—shorter "free text" input with an embedded single workflow OC system (FT practices) and longer "mixed text" input with variation in implemented workflow OC system (MT practices). We used 2022 General Practice Patient Survey data to capture 4 dimensions of patient experience—overall experience, experience of making an appointment, continuity of care, and use of self-care before making an appointment. We used logistic regression at the practice level to explore the association between OC system use and patient experience, including interaction terms to assess sociodemographic variation.

Results: We included 287,194 responses from 2423 MT and 170 FT practices. The proportions of patients reporting positive experiences at MT and FT practices were similar or better than practices nationally, except at high-use MT practices. At high-use MT practices, patients were 19.8% (odds ratio [OR] 0.802, 95% CI 0.782-0.823) less likely to report a good overall experience; 24.5% (OR 0.755, 95% CI 0.738-0.773) less likely to report a good experience of making an appointment; and 18.9% (OR 0.811, 95% CI 0.792-0.83) less likely to see their preferred general practitioner; but 27.8% (OR 1.278, 95% CI 1.249-1.308) more likely to use self-care, compared with low-use MT practices. Opposite trends were seen at FT practices. Sociodemographic inequalities in patient experience were generally lower at high-use than low-use practices; for example, gaps in overall experience between practices with the most and fewest White patients decreased by 2.7 percentage points at MT practices and 6.4 percentage points at FT practices. Trends suggested greater improvements in experience for traditionally underserved groups—patients from urban and deprived areas, younger patients, and non-White patients.

Conclusions: An OC system with shorter free text input and an integrated single workflow can enhance patient experience and reduce sociodemographic inequalities. Variation in patient experience between practices with different sociodemographic characteristics and OC systems underscores the importance of tailored design and implementation. Generalizing results across different OC systems is difficult due to variations in how they are integrated into practice workflows and communicated to patients.

(J Med Internet Res 2024;26:e51272) doi:10.2196/51272

XSL•FO
RenderX

KEYWORDS

general practice; online consultation; patient experience; modern general practice; sociodemographic health inequalities; General Practice Patient Survey; cross-sectional study

Introduction

Improving patients' experience of care has been an important policy focus for the National Health Service (NHS) in England [1]. Patient experience is a key outcome of health care [2,3] and a key component of the quality of care [4,5]. Patient experience is also an important factor in effective treatment outcomes, as satisfied patients are more likely to accept, adhere to, and continue with medication and treatment [6].

Over the past few years, general practices have started to adopt online consultation (OC) systems to support more inclusive and flexible access and delivery of primary care services. OC systems allow patients to use digital channels to contact their general practice, ask health-related questions, report symptoms, submit an administrative query, or receive health advice or help. The system gathers relevant information upfront, either directly from the patient or their caregiver or via care navigators or other practice staff who complete the form on behalf of patients over the phone. Practices can then use this information to understand the patient's need and respond in the right way by navigating and triaging the requests to the right service or professional [7].

A main driver of change has been the health policy in the English NHS. The 2019 *NHS Long Term Plan* committed to improving patient access to primary care by 2023 or 2024 [8] with a focus on the use of digital routes, particularly through the use of OC systems. The COVID-19 pandemic accelerated the adoption and use of digital access channels and contributed to a rapid growth in the use of OC systems to minimize the risk of exposure to COVID-19 infection for patients and staff [9]. Many practices in England now use OC systems combined with an embedded single workflow where all requests are managed equitably regardless of the route of contact. This model of implementation can support consistent filtering and navigation of requests, enabling fairer and safer allocation of the right type of appointment, correctly matched to the patients' needs on the first attempt. In 2023, in the context of rising demand and pressures on frontline staff, NHS England set out a commitment to implementing this "modern general practice" approach in the Delivery Plan for Recovering Access to Primary Care [10]. The modern general practice approach is a way of organizing work in general practice to provide more inclusive access and improve the understanding of demand. Using digital tools, practices can better align existing capacity with need through structured information gathering at the point of contact and consistent use of care navigation across all access channels to get patients to the right person or service at the right time. The approach aims to ensure that care is provided safely and equitably (including the continuity of care), optimizing the use of a multiprofessional team and improving the efficiency of practice processes.

There are numerous, constantly evolving OC systems now available for general practices in England with different designs and models of implementation. In particular, differences in request input formats and length, functionality (use of algorithms), workflow, and integration with other software have been found to have different impacts on patient experience [11].

Advancing equality and minimizing health inequalities is a key ambition of the *NHS Long Term Plan*. There is well-documented variation in patient experience of primary care in relation to sociodemographic factors—ethnic minority patients [12,13], younger patients [14], and patients living in the most deprived [15] and urban areas [16] tend to report the least positive experiences. However, there is little evidence on how the shift toward the use of OC systems in general practice will affect sociodemographic inequalities in patients' experience of primary care.

Given the rapid scale-up and diversity of OC systems, we sought to investigate how patients' experience of general practice varied between practices with high or low use of an OC system. We examined 2 OC systems with distinct designs and implementation models that are likely to result in different patient experiences. Key differences in the design relate to how information is gathered in a request (input)—either as shorter "free text" structure questions or as a longer "mixed input" of free text and logic-based, multiple-choice questions. Key differences in intended implementation relate to whether the OC system is consistently embedded with a single workflow for managing all patient requests regardless of the access route or not. We also explored changes in gaps in experience between practices with the most and least positive experience in relation to sociodemographic characteristics, including practice location (rural or urban) and deprivation, patient age, and ethnicity profiles. The aim of this research is to inform the design and implementation of OC systems and how they might be used to improve equity in patients' experience of primary care.

Methods

Data and Data Sources

This was a retrospective, cross-sectional study using aggregate practice-level data from the 2022 General Practice Patient Survey (GPPS), a major national survey that captures multiple aspects of patient experiences in local practices in England [17]. Questionnaires are sent to a random sample of patients who are 16 years of age or older and have been continuously registered with their general practice for at least 6 months. Data are weighted to account for some people being more likely to respond than others. This adjusts the data to account for any demographic differences between all eligible patients in the practice and those who respond to the survey. Additional details regarding the sampling, weighting, questionnaire design, and data collection are published elsewhere [18]. Aggregate practice-level data from the 2019 GPPS was used to control for prior patient experience.

We obtained volumes of patient-initiated requests made via an OC system between March 2021 and February 2022 at practices

XSL•FO
RenderX

using 1 of the 2 OC systems with key differences in their design and implementation model:

- Free text input with an embedded single workflow OC system (FT OC system): Patients enter their requests using short, structured free text responses [19]. Tick-box questions allow patients to indicate whether they are a patient, parent, or caregiver; their request type as a new medical problem, an existing problem, or any other question; and whether they would like to see a preferred clinician. Practices are encouraged to capture all incoming requests, whether initiated online, in person, or by telephone (and completed on behalf of patients by care navigators or other practice staff), in the OC system to support a model of modern general practice, where there is parity of access and prioritization regardless of the access route [20].
- Mixed text input with variation in implemented workflow OC system (MT OC system): Patients enter their requests using a mix of long free text and logic-based, multiple-choice questions. This questionnaire is longer than the one used in the FT OC system. An algorithm triages and flags response according to the clinical urgency. Patients may then be automatically redirected to urgent or emergency care, for example, to call NHS 111 or go directly to an emergency department. There is variation in implementation models, with some practices managing all requests via a single workflow and others via multiple workflows. For example, requests initiated in person or by telephone may be handled separately from those incoming via the OC system.

Both systems provide patients with clear self-care instructions, the ability to nominate a preferred clinician, and aim to respond to requests within a fixed timeframe. Practices may respond via online or text message, telephone call, or a face-to-face consultation. GPPS and patient-initiated request data were linked at the practice level to publicly available information capturing practice characteristics. These included sociodemographic characteristics of registered patients and practice workforce data from NHS Digital [21], area deprivation from the English Index of Multiple Deprivation [22], rural or urban location from the Office of National Statistics [23], and Care Quality Commission ratings from NHS Digital [24].

Study Population and Use of the OC System

We included all practices using the MT or FT OC system between March 2021 and February 2022 (hereafter referred to as MT or FT practices, respectively). We determined the number of requests that were initiated online via the OC system per 1000 registered patients per practice per month (usage rate). Practices were categorized as high-use or low-use according to the usage rate over the study period (Figure S1 in Multimedia Appendix 1). This was assumed to be a proxy for whether a practice was fully (high-use) or only partially (low-use) using the OC system for patient-initiated demand. Due to the much larger number of MT practices, we also included a medium-use category for MT practices. Multimedia Appendix 1 presents the practice inclusion criteria and use categorization.

GPPS Experience Dimensions

We selected 4 dimensions of patient-reported experience of primary care—overall experience, experience of making an appointment, ability to see or speak to a preferred general practitioner (GP; hereafter referred to as the continuity of care), and trying to get information or advice before trying to get an appointment (hereafter referred to as self-care). These dimensions were chosen based on their likelihood of both being impacted by the design of the OC system and where past research has demonstrated their importance in driving overall experience of primary care [25]. Full question wording and details are presented in Table S1 in Multimedia Appendix 1.

Study Covariates

For each practice, we determined list size; number of full-time equivalent GPs; Care Quality Commission rating; proportion of registered patients by age, gender, ethnicity, and education level; rural or urban practice location; practice deprivation; whether the practice starting using the OC system before or after the COVID-19 pandemic began in March 2020; and prior patient experience scores for each of the patient experience dimensions in 2019. Practice deprivation was categorized according to deprivation score quintile, with quintile 1 representing a practice located in 20% of most deprived areas in England [26]. Practice rural or urban location was determined by Lower Layer Super Output Area Statistics [23].

Statistical Methods

Multiple-choice responses to the GPPS questions (Table S1 in Multimedia Appendix 1) were summarized as follows. For 3 of the 4 questions with a 5-point ordinal scale response, binomial variables were created by combining the 2 most positive responses together versus the other responses. For example, the overall experience score was measured as the proportion of responders at a given practice reporting either a “very good” or “fairly good” overall experience. For the question relating to the use of self-care before seeking an appointment, binomial variables were created by combining all positive responses indicating that some action had been taken versus the response of “no action taken.”

We used logistic regression [27] to explore the relationship between the use of the OC system and each dimension of patient experience with adjustment for covariates. We standardized continuous variables to ensure they were on the same scale [28]. To explore whether gaps in experience between practices with the most and least positive experience in relation to practice sociodemographic characteristics varied between low-use and high-use MT and FT practices, we tested for an interaction between each sociodemographic characteristic and use level in turn. For interaction analyses, continuous age and ethnicity variables were converted into categorical variables representing different proportions of registered patients older than 65 years of age and of White ethnicity, respectively. Predicted experience scores were calculated by averaging practice characteristics in each interaction group [29] and were used to trace changes in patient experience in low-use and high-use practices by their sociodemographic characteristics.

XSL•FO
RenderX

All analyses were carried out on a secure analysis server at the Health Foundation using R (version 4.0.3; R Core Team) [30]. Data were provided to the research team under a data sharing agreement with NHS England and are not publicly available.

Sensitivity Analysis

We conducted sensitivity analyses to assess the reliability of our findings. One sensitivity analysis involved weighting negative responses (proportions of patients reporting a "poor" or a "very poor" experience) when rating patient experience. Another sensitivity analysis applied a different set of usage rate thresholds to address the robustness of the current approach used to categorize use.

Ethical Considerations

The authors declare that this research meets ethical guidelines. The GPPS is designed to give patients the opportunity to give feedback about their experiences of their general practice. The original data collection is approved by the Central Office for Research Ethics Committee (COREC) and carried out by Ipsos MORI for NHS England. Ipsos MORI is a registered and independent survey organization that strictly adheres to the Market Research Society's ethical code of conduct. The patient-initiated request counts from OC systems were anonymized and not identifiable to the research team carrying out the research. This analysis did not receive an exemption from an institutional review board as its review was not required for this research, which is limited to the secondary use of information previously collected in the course of normal care and where the patients are not identifiable to the research team carrying out the research.

Results

Practice Characteristics

A total of 287,164 responses from the 2022 GPPS survey from 2423 MT and 170 FT practices were included in the analysis. The overall response rates at MT and FT practices were 28.3% and 32.1%, respectively. Sociodemographic characteristics of MT and FT practices are broadly comparable to practices in the rest of England (Table 1) with a few notable exceptions—high-use MT practices were less likely to be in the top 20% of most deprived areas or a rural area and more likely to have more registered patients compared with practices nationally; high-use FT practices were also less likely to be in the top 20% of most deprived areas but more likely to be in a rural area and have a greater proportion of patients older than 65 years of age and of White ethnicity compared with practices nationally.

XSL•FO
RenderX

Table 1. Patient sociodemographic characteristics of general practices using MT[a] and FT[b] OC[c] systems and English practices nationally (2022).

Characteristic	MT practices			FT practices		National practices (n=6518)
	Low use (n=932)	Medium use (n=825)	High use (n=666)	Low use (n=91)	High use (n=79)	
Deprivation, n (%)						
1 (most deprived)	271 (29.3)	218 (26.8)	146 (22.1)	24 (26.7)	16 (20.5)	1806 (27.78)
2	217 (23.4)	170 (20.9)	156 (23.6)	24 (26.7)	23 (29.5)	1501 (23.09)
3	174 (18.8)	163 (20.0)	138 (20.9)	14 (15.6)	18 (23.1)	1248 (19.2)
4	143 (15.4)	145 (17.8)	109 (16.5)	17 (18.9)	12 (15.4)	1017 (15.64)
5 (least deprived)	121 (13.1)	117 (14.4)	112 (16.9)	11 (12.2)	9 (11.5)	929 (14.29)
Location, n (%)						
Rural	140 (15.2)	92 (11.4)	56 (8.5)	19 (21.1)	20 (25.6)	954 (14.74)
Small town	271 (29.4)	325 (40.1)	242 (36.7)	18 (20.0)	22 (28.2)	2516 (38.89)
Urban	510 (55.4)	393 (48.5)	361 (54.8)	53 (58.9)	36 (46.2)	3000 (46.37)
CQC[d] rating, n (%)						
1	4(0.4)	1(0.1)	3(0.5)	0(0.0)	0(0.0)	29 (0.46)
2	48 (5.3)	36 (4.5)	27 (4.1)	3(3.3)	3(4.0)	281 (4.46)
3	829 (90.7)	730 (90.9)	603 (92.2)	82 (90.1)	70 (92.1)	5,681 (90.23)
4	33 (3.6)	36 (4.5)	21 (3.2)	6 (6.6)	3 (4.0)	305 (4.84)
Practice list size, median (IQR)	6465 (4294-9174)	8685 (5892-12,126)	9645 (6612-13,666)	8402 (5258-10,906)	8348 (5764-11,899)	8028 (5229-11,710)
Number of full-time equivalent doctors per 1000 registered patients, median (IQR)	0.49 (0.37-0.66)	0.54 (0.40-0.72)	0.56 (0.42-0.71)	0.54 (0.42-0.71)	0.58 (0.45-0.76)	0.53 (0.40-0.71)
Registered patients aged >65 years (%), median (IQR)	17.07 (12.61-21.93)	17.98 (13.72-22.17)	17.13 (13.03-21.62)	19.69 (17.44-23.73)	19.89 (17.81-22.78)	18.18 (13.77-22.34)
Registered patients by ethnicity group (%), median (IQR)						
Asian	4.17 (1.14-13.02)	3.22 (1.24-10.35)	4.33 (1.47-12.86)	1.76 (0.92-5.43)	1.28 (0.87-3.88)	3.52 (1.2-11.21)
Black	1.46 (0.29-7.09)	1.11 (0.3-5.39)	1.51 (0.34-7.06)	0.61 (0.31-1.18)	0.53 (0.32-0.93)	1.03 (0.32-4.85)
Mixed	1.90 (0.94-3.96)	1.76 (1.00-3.66)	2.02 (1.10-3.93)	1.05 (0.76-1.95)	1.02 (0.77-1.78)	1.72 (0.98-3.47)
White	90.85 (64.8-97.27)	93.22 (74.55-97.16)	90.56 (68.42-96.74)	96.31 (90.68-97.86)	96.89 (92.77-97.86)	92.66 (75.14-97.25)
Other	0.55 (0.18-2.43)	0.50 (0.18-1.91)	0.58 (0.21-2.31)	0.27 (0.14-0.72)	0.23 (0.15-0.39)	0.42 (0.17-1.49)
Registered male patients (%), median (IQR)	49.36 (48.66-50.30)	49.15 (48.50-50.14)	49.21 (48.61-50.13)	49.31 (48.75-49.71)	49.09 (48.64-50.01)	49.24 (48.65-50.02)
Registered patients with higher education (%), median (IQR)	11.33 (10.03-12.24)	11.78 (10.67-12.61)	11.81 (10.62-12.67)	12.11 (11.68-12.67)	12.03 (11.35-12.90)	11.70 (10.67-12.52)

[a]MT: mixed text input with variation in implemented workflow online consultation system.

[b]FT: free text input with an embedded single workflow online consultation system.

[c]OC: online consultation.

[d]CQC: Care Quality Commission.

XSL•FO
RenderX

Association Between Use of the OC System and Patient Experience

The proportions of patients reporting positive experiences at MT and FT practices were similar or better than practices nationally, except at high-use MT practices (Table 2). High-use MT practices had lower proportions of patients reporting a positive experience than both low-use MT practices and practices nationally across all experience dimensions except self-care. Conversely, high-use FT practices had higher proportions of patients reporting a positive experience than both low-use FT practices and practices nationally across all experience dimensions except self-care. For example, 68.8% and 78.2% of respondents at high-use MT and FT practices, respectively, reported a good overall experience, compared with 72.4% of practices nationally. For self-care, 66% and 61.6% of respondents at high-use MT and FT practices, respectively, reported use of self-care before making an appointment, compared with 61.5% of practices nationally.

These trends are confirmed after adjusting for underlying differences between the practices (Figure 1 and Table S2 in Multimedia Appendix 1) in regression modeling. At MT practices, high use of the OC system is associated with poorer overall experience (odds ratio [OR] 0.8, 95% CI 0.78-0.82; *P*<.001), experience of making an appointment (OR 0.75, 95% CI 0.74-0.77; *P*<.001), and continuity of care (OR 0.81, 95% CI 0.79-0.83; *P*<.001) but better use of self-care (OR 1.28, 95% CI 1.25-1.31; *P*<.001), compared with low use. At FT practices, high use of the OC system is associated with better overall experience (OR 1.23, 95% CI 1.13-1.32; *P*<.001), experience of making an appointment (OR 1.31, 95% CI 1.23-1.4; *P*<.001), and continuity of care (OR 1.28, 95% CI 1.20-1.37; *P*<.001) but not with use of self-care (OR 0.99, 95% CI 0.93-1.06; *P*=.87), compared with low use.

Table 2. Numbers of included responses and weighted national patient experience in England at general practices using MT[a] and FT[b] OC[c] systems and English practices nationally (2022).

Dimension	MT practices, n (%)			FT practices, n (%)		National practices, n (%)
	Low use	Medium use	High use	Low use	High use	
Good overall experience	81,184 (75.5)	96,289 (72.8)	86,837 (68.8)	9406 (72.7)	9410 (78.2)	710,421 (72.4)
Good overall experience of making an appointment	76,210 (61.9)	90,215 (56.1)	81,315 (49.9)	8788 (57.1)	8807 (66.2)	666,553 (56.2)
Continuity of care	34,225 (44.4)	37,433 (37.6)	31,125 (33.8)	3567 (39.6)	3422 (44.7)	274,407 (38.2)
Self-care	75,787 (58.9)	89,645 (62.2)	80,964 (66)	8756 (62.3)	8732 (61.6)	662,857 (61.5)

[a]MT: mixed text input with variation in implemented workflow online consultation system.

[b]FT: free text input with an embedded single workflow online consultation system.

[c]OC: online consultation.

XSL•FO
RenderX

Figure 1. Association between patient experience and use of an OC system at (A) MT practices and (B) FT practices. FT: free text input with an embedded single workflow online consultation system; MT: mixed text input with variation in implemented workflow online consultation system; OC: online consultation.

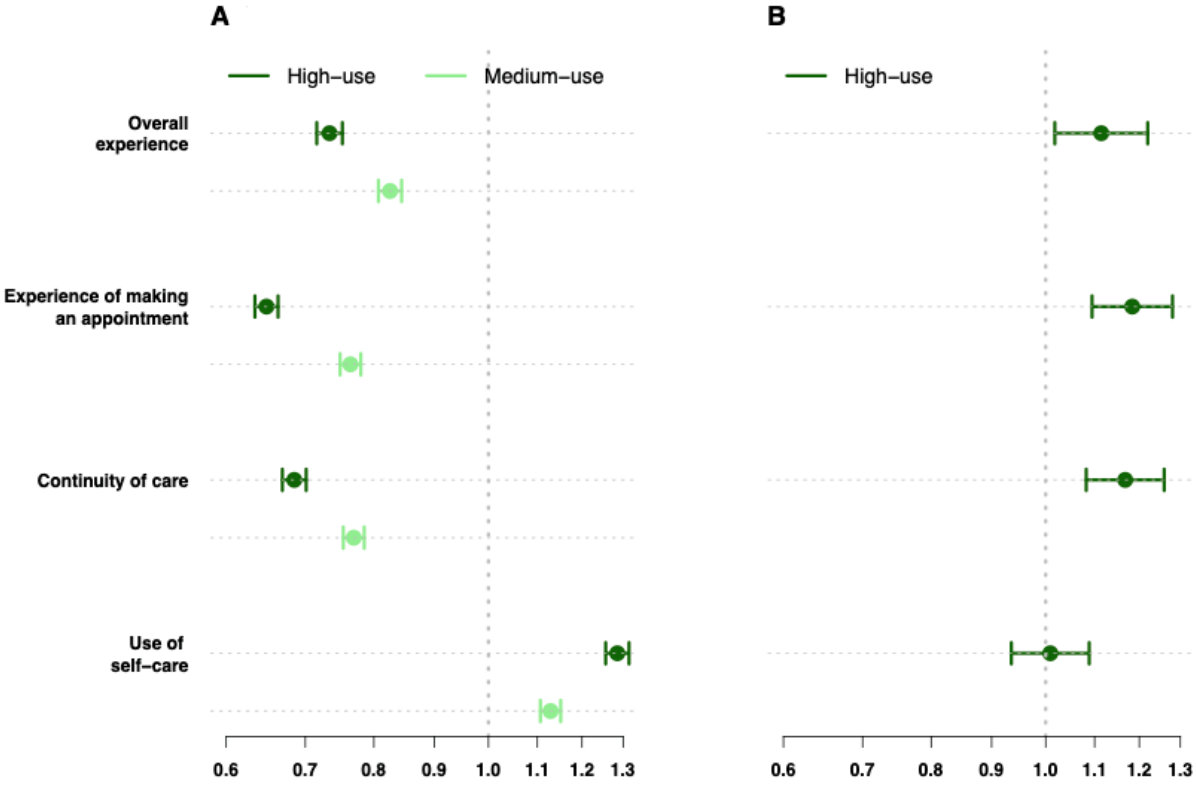

Sociodemographic Inequalities

Including an interaction in our multivariable model allowed us to predict average reported experience scores at low-use and high-use MT and FT practices in different sociodemographic groups (Figure 2, Table 3, and Figures S2-S4 in Multimedia Appendix 1). We have denoted in blue and red the sociodemographic groups that traditionally experience the worst and best experience of primary care, respectively.

For all patient experience dimensions other than self-care, (1) predicted patient experience at low-use MT and FT practices was poorest for the practices in the traditionally most challenged sociodemographic groups—namely, those in urban and most deprived areas and with the greatest proportion of patients aged younger than 65 years and of non-White ethnicity; (2) declines in predicted patient experience between low-use and high-use MT practices were more pronounced for practices in the least challenged sociodemographic groups—namely, those in rural and least deprived areas and with the greatest proportion of patients aged older than 65 years and of White ethnicity; and (3) improvements in average patient experience between low-use and high-use FT practices were chiefly driven by practices in the most challenged sociodemographic groups.

Improvements in the use of self-care between low-use and high-use MT practices tended to be most pronounced at practices in the least challenged sociodemographic groups. There was no significant sociodemographic variation in the use of self-care between low-use and high-use FT practices.

Table 3 summarizes how the difference in average experience between the most and least positive practice sociodemographic groups changed between low-use and high-use practices. A surfeit of significant negative estimates overwhelmingly suggests that sociodemographic inequalities are narrower at high-use than at low-use practices. For example, the difference in overall experience between rural and urban practices is 4.3 (P<.001) and 10.6 (P<.001) percentage points smaller at high-use compared with low-use MT and FT practices, respectively. There were some notable nonnegligible exceptions to this trend in relation to age—differences in experience between practices with the greatest and least proportions of patients older than 65 years of age were wider at high-use compared with low-use practices for continuity of care (3.1 percentage points, P<.001) at MT practices and for overall experience (7.0 percentage points, P<.001) and experience of making an appointment (5.0 percentage points, P<.001) at FT practices. Results were robust to sensitivity analyses.

XSL•FO
RenderX

Figure 2. Trends in inequalities in overall patient experience in (A) MT practices and (B) FT practices. The y-axis presents adjusted predictions of the proportion of patients reporting a good response for the overall experience in primary care. Average estimates with 95% CIs are presented for each practice sociodemographic group. The absolute change between low-use and high-use practices can be seen visually in the figure and corresponds to the values presented in Table 3 column 1. Groups represented in red or blue are those that traditionally report the least or most challenges in their experience of primary care. FT: free text input with an embedded single workflow online consultation system; MT: mixed text input with variation in implemented workflow online consultation system.

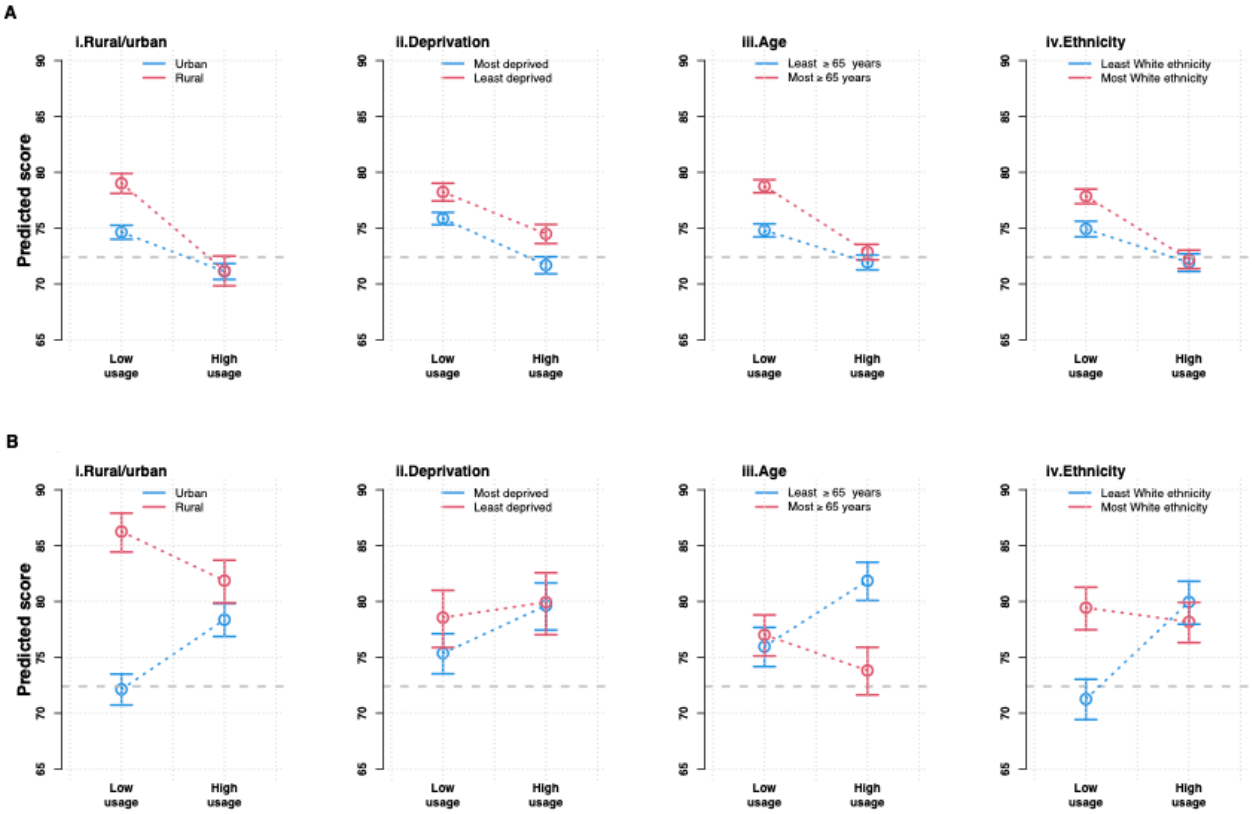

XSL•FO
RenderX

Table 3. Changes in inequalities in patient experience between high-use and low-use practices using MT[a] and FT[b] OC[c] systems[d].

Practices and inequalities	Overall experience	Experience of making an appointment	Continuity of care	Use of self-care
MT practices				
Location (rural or urban)				
Change, n	–4.31	–3.01[e]	–2.54[e]	–1.26[e]
Interaction *P* value	<.001[f]	<.001[f]	<.001[f]	<.001[f]
Deprivation				
Change, n	0.43	–0.65	–0.06	0.02
Interaction *P* value	<.001[f]	<.001[f]	<.001[f]	<.001[f]
Age				
Change, n	–3.03	–2.55	3.12	–2.21
Interaction *P* value	<.001[f]	<.001[f]	<.001[f]	<.001[f]
Ethnicity				
Change, n	–2.64	–2.01[e]	–4.57	–2.2
Interaction *P* value	.003[f]	<.001[f]	<.001[f]	.28
FT practices				
Location (rural or urban)				
Change, n	–10.63	–13.03	3.93[e]	–2.44
Interaction *P* value	<.001[f]	<.001[f]	<.001[f]	.06
Deprivation				
Change, n	–2.89	–1.74[e]	0.04[e]	2.23[e]
Interaction *P* value	.03[f]	.05	.01[f]	.1
Age				
Change, n	6.99[1]	5.04[e]	–2.34	0.03[e]
Interaction *P* value	<.001[f]	<.001[f]	<.001[f]	.3
Ethnicity				
Change, n	–6.39[e]	–5.51[e]	5.76[e]	2.59
Interaction *P* value	<.001[f]	<.001[f]	<.001[f]	.42

[a]MT: mixed text input with variation in implemented workflow online consultation system.

[b]FT: free text input with an embedded single workflow online consultation system.

[c]OC: online consultation.

[d]The interaction *P* values presented in this table can be interpreted as a measure of the strength of evidence that inequalities in patient experience changed between low-use and high-use practices for each sociodemographic characteristic and for each experience outcome. The summary measure (change) is the difference in predicted experience between the absolute difference in predicted experience between the most and least positive practice sociodemographic groups at low-use practices and the equivalent measure at high-use practices. Negative values indicate that inequalities between the most and least positive practice groups have narrowed or improved; the converse is applicable for positive values. For all dimensions other than self-care, low-use practices experiencing the most challenge are generally those in urban and most deprived areas, with the greatest proportion of patients aged younger than 65 years and from ethnic minority groups. For self-care, the opposite is true, with low-use practices in rural and least deprived areas, and with the greatest proportion of patients aged 65 years or older and of White ethnicity experiencing the most challenge at low-use practices.

[e]The inequality has flipped with the most challenged sociodemographic group at low-use practices now the least challenged sociodemographic group at high-use practices.

[f]Significant results ($P<.05$).

XSL•FO
RenderX

Discussion

Patient Experience and Use of an OC system

Practice use of an OC system was associated with substantial differences in the patient-reported experience of primary care across dimensions of the overall experience, the experience of making an appointment, continuity of care, and self-care at the practice level. These results varied by the type of OC system and by practice sociodemographic factors providing important new insights into optimal OC system design and implementation.

Opposing effects on patient experience between low-use and high-use MT practices compared with FT practices suggests that differences in the design and implementation model of each OC system are critical.

Design

The MT OC system uses multiple-choice questions for inputting requests, which has been shown to increase patient workload and decrease patient satisfaction. Problems include the length of time taken to complete [31-33], confusing navigation [34], or being redirected to services that do not address the patient's needs [35]. In the latter case, patients may amend their responses to achieve a different outcome [31]. Conversely, the FT OC system uses shorter free text for inputting requests, enabling patients to express their requests in their own words and potentially reduce the time for completion. These differences in how requests are input might help explain the contrasting effects on patient overall experience and experience of making an appointment between low-use and high-use MT and FT practices. Use of self-care is better at high-use than at low-use MT practices but is not associated with use at FT practices. Self-care options may be more apparent or readily accessible on MT OC systems, or alternatively, algorithms used by MT OC systems may be more likely to guide patients toward self-care alternatives. Although careful consideration needs to be given to the design of self-care options in an OC system to assess their appropriateness for each patient, these findings suggest that OC systems could be effectively used to encourage self-care.

Implementation

Both MT and FT OC systems allow patients to nominate a preferred clinician, which is expected to improve continuity of care [36]. However, only patients at FT practices reported enhanced continuity at high-use compared with low-use practices. FT OC systems offers a model of digitally supported information gathering and care navigation, where all incoming requests, whether by telephone or online, are processed through a single workflow allowing for consistent understanding of needs, navigation, and triage. Greater variability in workflow models at MT practices may offer some insights into why reported continuity is worse at high-use than at low-use MT practices, but more research would be needed to confirm this theory [36].

Inequalities in Patient Experience

The association between patient use of an OC system and patient experience of primary care varied by practice-level sociodemographic factors. Practices in urban and most deprived areas, as well as those with the greatest proportion of patients aged younger than 65 years and from ethnic minority groups, tended to benefit the most from high use of the OC system. Notably, these are the groups that typically report the worst experience of primary care [12,14,16]. Here, we found that these groups reported greater improvements (FT practices) or less pronounced declines (MT practices) in experience in all dimensions except self-care at high-use compared with low-use practices than their counterparts—namely, those in rural and least deprived areas and with the greatest proportion of patients aged older than 65 years and of White ethnicity.

These findings suggest that the FT OC system design and indicated implementation model is more likely that the MT OC system design and indicated implementation model to address the challenges faced by those patients who traditionally report the worst experience of primary care. Factors that may explain differences in experience across different sociodemographic groups are explained in the following sections.

Work or Caring Commitments

OC systems provide patients with flexible ways to contact their local practices, allowing them to avoid the inconvenience and cost of traveling to their practice or waiting on the telephone [37]. The improved access may be of greater benefit to people from the most deprived communities who are more likely to find it difficult to take time away from work or caring commitments [38].

Language Barriers

Research shows that shorter question sets and free text entry can be easier and faster for patients to answer and align best with their speaking and communication styles [11]. This may be particularly important for patients for whom English is not their first language and who may feel more confident or at ease sending an online request, where they can take more time to express themselves or can receive help via online translation systems or from relatives or friends [39].

Population Health and the Demand-Capacity Gap Within a Practice

People from ethnic minority backgrounds or those residing in disadvantaged areas shoulder a higher burden of health problems compared with people of White ethnicity or those living in wealthier areas [40]. Practices in deprived areas tend to be overrepresented by patients from ethnic minority groups [40] and are underfunded and underresourced relative to need [41,42]. Hence, a greater demand-capacity gap might make getting an appointment or having sufficient time during the appointment more difficult at practices in deprived areas or with a high proportion of patients from ethnic minority backgrounds. If an OC system helps practices to enhance their care navigation and better align capacity with needs more efficiently and effectively, it has the potential to have a greater impact on these challenged practices than on well-funded and well-resourced practices.

XSL•FO
RenderX

Digital Inclusion and Adoption

People in rural areas often face greater challenges accessing fast and reliable digital technologies [43] and have a greater proportion of older patients compared with people living in urban areas. Older patients also have greater health care needs [44] and may be slower to adopt new technologies [45] or find it more difficult or frustrating to use digital systems compared with younger patients [46]. The greater need for care combined with less familiar access routes to care may lead to a poorer experience at high than at low-use practices for older patients and those living in rural areas. Research also shows that both nondigital and digital users can benefit from modern general practice, with older patients and those with greater clinical needs being prioritized as well as faster access being provided for older patients, young children, those with new presentations, and those with more complex care needs [47]. Hence, improving support for the implementation and use of digital tools as part of modern general practice has the potential to improve patient experience for older patients over time [48].

Patient Expectations

Patient expectations and perceptions prior to experiencing care are theorized to be a mediator of patient satisfaction [49]. Individuals from ethnic minority backgrounds often have lower expectations of general practice compared with those of White ethnicity [50]. High-use practices with large proportions of patients from ethnic minority backgrounds generally report better experiences (except for self-care) compared to high-use practices with mostly White patients. This might suggest that the expectations of patients from ethnic minority backgrounds are easier to meet, rather than indicating a truly better experience.

The finding that the continuity of care was better at high-use FT practices compared with both low-use FT practices and practices nationally may offer further evidence that the approach to navigation and triage indicated by the FT OC system can enable greater continuity of care. Similarly, self-care was better at high-use MT practices compared with both low-use MT practices and practices nationally for all sociodemographic groups. This may suggest that the ability to incorporate options for self-care in an OC system can be useful in addressing some of the barriers or challenges to self-management known to be greater in some populations than others, for example, people experiencing socioeconomic deprivation [51].

Narrowing of Inequalities

Our findings show that differences in reported patient experience between the most and least challenged practice groups tend to be smaller at high-use than at low-use practices. This may support leveling up of access across sociodemographic groups and so contribute to more inclusive access. It may also suggest that practices that have fully implemented an OC system may have invested more time and resources in raising awareness and supporting its use. More work is required to focus on design and implementation across the 2 systems and to understand in greater depth the variation between different patient cohorts and why benefits are not realized for practices using MT OC systems.

Limitations

There were 4 main limitations of our study. First, we only captured total counts of patients using an OC system at practice, not who or how patients were using it, nor whether all patients responding to the survey had actually used them. We also could not account for variations in how they are integrated into practice workflows, how they are communicated to patients, and how their data are collected. These factors may have reduced the power to detect effects and to disentangle which features of design and implementation are responsible for the effects we see. Second, the analysis was conducted at the practice rather than the patient level, so it was not possible to make conclusions about the impact of OC systems on individual patients or to adjust at the patient level for patient comorbidities, which may be correlated with patient experience and the sociodemographic characteristics considered.

Third, as with most patient experience surveys [52], some practices included in this analysis have a relatively low survey response rate, which may introduce systematic bias into the data. However, the GPPS data are weighted to account for differences between all patients at a practice and the subset that takes part in the survey. This mitigates the issue of a low response rate and other studies have reported little evidence to support concerns related to nonresponse bias in analyses using the GPPS survey [53,54]. Finally, responses could not restrict all patients' interaction with practices within a given timeframe. Hence, recall bias because of systematic differences between the ability of respondents to accurately recall details of their most recent GP experience may be introduced.

Conclusions

The OC systems have the potential to increase overall experience of general practice, experience of making an appointment, continuity of care, and use of self-care and to reduce sociodemographic disparities in access to care. However, stark differences in reported patient experience at practices using different types of OC systems suggest that careful consideration as to how the systems are designed and implemented, particularly in relation to local population needs and demographics, is required to achieve this. Generalizing and evaluating impacts across different OC systems is difficult due to variations in how they are integrated into practice workflows, how they are communicated to patients, and their constant evolution

Acknowledgments

The authors would like to thank representatives of the online consultation (OC) systems studied in this evaluation for sharing counts of patient-initiated OC requests and for their advice in understanding their OC systems. The authors would like to thank

XSL•FO
RenderX

Richard Brine and Liz Crellin for data management and procurement. They thank Arne Wolters and Kaat de Corte for reviewing the paper prior to submission. They also thank Jenny Newbould (RAND Europe) for her advice and insights.

Data Availability

Data on patient use of online consultation systems were provided to the research team under a data sharing agreement with NHS England and are not publicly available due to commercial sensitivities. The General Practice Patient Survey (GPPS) data sets analyzed during this study are freely available [17].

Disclaimer

This research should not be considered to constitute or imply any form of recommendation to buyers in relation to current or future procurement activities. It provides a "point of time" view of systems which are out in the General Practice Market, therefore, it does not attempt to be comprehensive or represent the future market; it should not be considered to be predictive or anticipatory in any way in relation to ongoing market refresh in this area, which occurs periodically.

Authors' Contributions

XG, MB, JL, and GMC conceived the design of the study. XG conducted all statistical analyses. All authors interpreted the results. XG, MB, and GMC wrote the paper. All authors reviewed the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online consultation system use, General Practice Patient Survey 2022 questions and available answers for selected domains, trends in inequalities in overall experience of making an appointment, trends in inequalities in the ability to see or speak to a preferred general practitioner (continuity of care), and trends in inequalities in seeking any kind of information or advice before trying to make an appointment (self-care utilization).
[DOCX File , 553 KB - jmir_v26i1e51272_app1.docx]

References

1. Improving experience of care: a shared commitment for those working in health and care systems. NHS England. URL: https://tinyurl.com/mszk85rn [accessed 2023-02-23]
2. NHS outcomes framework (NHS OF). NHS Digital. URL: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-outcomes-framework [accessed 2023-06-15]
3. Patient experience improvement framework. NHS England. URL: https://www.england.nhs.uk/publication/patient-experience-improvement-framework/ [accessed 2023-06-15]
4. Donabedian A. The quality of care. How can it be assessed? JAMA 1988;260(12):1743-1748. [doi: 10.1001/jama.260.12.1743] [Medline: 3045356]
5. Anhang Price R, Elliott MN, Zaslavsky AM, Hays RD, Lehrman WG, Rybowski L, et al. Examining the role of patient experience surveys in measuring health care quality. Med Care Res Rev 2014;71(5):522-554 [FREE Full text] [doi: 10.1177/1077558714541480] [Medline: 25027409]
6. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013;3(1):e001570 [FREE Full text] [doi: 10.1136/bmjopen-2012-001570] [Medline: 23293244]
7. Bakhai M, Croney L, Waller O, Henshall N, Felstead C. Using online consultations in primary care: implementation toolkit. NHS England. URL: https://tinyurl.com/rj9ks4ph [accessed 2023-07-10]
8. NHS Long Term Plan. NHS. URL: https://www.longtermplan.nhs.uk [accessed 2023-03-16]
9. Advice on how to establish a remote 'total triage' model in general practice using online consultations. Coronavirus. URL: https://www.england.nhs.uk/coronavirus/documents/advice-on-how-to-establish-a-remote-total-triage-model-in-general-practice-using-online-consultations/ [accessed 2023-05-22]
10. Delivery plan for recovering access to primary care. NHS England. URL: https://www.england.nhs.uk/publication/delivery-plan-for-recovering-access-to-primary-care/ [accessed 2023-07-14]
11. Darley S, Coulson T, Peek N, Moschogianis S, van der Veer SN, Wong DC, et al. Understanding how the design and implementation of online consultations affect primary care quality: systematic review of evidence with recommendations for designers, providers, and researchers. J Med Internet Res 2022;24(10):e37436 [FREE Full text] [doi: 10.2196/37436] [Medline: 36279172]

XSL•FO
RenderX

12. Lyratzopoulos G, Elliott M, Barbiere JM, Henderson A, Staetsky L, Paddison C, et al. Understanding ethnic and other socio-demographic differences in patient experience of primary care: evidence from the English general practice patient survey. BMJ Qual Saf 2012;21(1):21-29 [FREE Full text] [doi: 10.1136/bmjqs-2011-000088] [Medline: 21900695]
13. Magadi JP, Magadi MA. Ethnic inequalities in patient satisfaction with primary health care in England: evidence from recent general practitioner patient surveys (GPPS). PLoS One 2022;17(12):e0270775 [FREE Full text] [doi: 10.1371/journal.pone.0270775] [Medline: 36542601]
14. Kontopantelis E, Roland M, Reeves D. Patient experience of access to primary care: identification of predictors in a national patient survey. BMC Fam Pract 2010;11:61 [FREE Full text] [doi: 10.1186/1471-2296-11-61] [Medline: 20799981]
15. Levene LS, Baker R, Walker N, Williams C, Wilson A, Bankart J. Predicting declines in perceived relationship continuity using practice deprivation scores: a longitudinal study in primary care. Br J Gen Pract 2018;68(671):e420-e426 [FREE Full text] [doi: 10.3399/bjgp18X696209] [Medline: 29739778]
16. Iqbal I, Thompson L, Wilson P. Patient satisfaction with general practice in urban and rural areas of Scotland. Rural Remote Health 2021;21(4):6634 [FREE Full text] [doi: 10.22605/RRH6634] [Medline: 34706547]
17. GP Patient Survey (GPPS). NHS England. URL: https://gp-patient.co.uk/downloads/2023/GPPS_2023_National_infographic_PUBLIC.pdf [accessed 2024-04-04]
18. GP patient survey 2022. Technical Annex.: Ipsos Mori URL: https://www.gp-patient.co.uk/Downloads/2022/GPPS_2022_Technical_Annex_PUBLIC.pdf [accessed 2024-01-23]
19. Eccles A, Hopper M, Turk A, Atherton H. Patient use of an online triage platform: a mixed-methods retrospective exploration in UK primary care. Br J Gen Pract 2019;69(682):e336-e344 [FREE Full text] [doi: 10.3399/bjgp19X702197] [Medline: 30910874]
20. Advice on how to establish a remote 'total triage' model in general practice using online consultations. NHS England and NHS Improvement. URL: https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/03/C0098-total-triage-blueprint-september-2020-v3.pdf [accessed 2024-06-20]
21. General Practice Workforce. NHS Digital. URL: https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services [accessed 2023-06-19]
22. English indices of deprivation 2019. HM Government. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2024-06-20]
23. The 2011 rural-urban classification for small area geographies: a user guide and frequently asked questions (v1.0). URL: https://assets.publishing.service.gov.uk [accessed 2024-06-20]
24. Quality and Outcomes Framework (QOF) - 2016-17 - NHS Digital. NHS England. URL: https://digital.nhs.uk [accessed 2024-06-20]
25. Paddison CAM, Abel GA, Roland MO, Elliott MN, Lyratzopoulos G, Campbell JL. Drivers of overall satisfaction with primary care: evidence from the English general practice patient survey. Health Expect 2015;18(5):1081-1092 [FREE Full text] [doi: 10.1111/hex.12081] [Medline: 23721257]
26. The English indices of deprivation 2019. In: Frequently Asked Questions (FAQs). London, UK: Ministry of Housing, Communities & Local Government; 2019.
27. Dunn PK, Smyth GK. Generalized linear models with examples in R. In: Springer Texts in Statistics. Berlin: Springer; 2018.
28. Schielzeth H. Simple means to improve the interpretability of regression coefficients. Methods Ecol Evol 2010;1(2):103-113. [doi: 10.1111/j.2041-210x.2010.00012.x]
29. Lüdecke D. ggeffects: tidy data frames of marginal effects from regression models. JOSS 2018;3(26):772 [FREE Full text]
30. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
31. Jones RB, Tredinnick-Rowe J, Baines R, Maramba IDC, Chatterjee A. Use and usability of GP online services: a mixed-methods sequential study, before and during the Covid-19 pandemic, based on qualitative interviews, analysis of routine eConsult usage and feedback data, and assessment of GP websites in Devon and Cornwall, England. BMJ Open 2022;12(3):e058247 [FREE Full text] [doi: 10.1136/bmjopen-2021-058247] [Medline: 35256447]
32. Cowie J, Calveley E, Bowers G, Bowers J. Evaluation of a digital consultation and self-care advice tool in primary care: a multi-methods study. Int J Environ Res Public Health 2018;15(5):896 [FREE Full text] [doi: 10.3390/ijerph15050896] [Medline: 29724040]
33. Moschogianis SF, Darley S, Coulson T, Peek N, Cheraghi-Sohi S, Brown B. Patient experiences of an online consultation system: a qualitative study in English primary care post-Covid-19. Br J Gen Pract 2024:BJGP.2023.0076. [doi: 10.3399/BJGP.2023.0076] [Medline: 38164550]
34. Greater Manchester Digital First Primary Care. Health Innovation Manchester. URL: https://healthinnovationmanchester.com/our-work/gm-digital-first-primary-care/ [accessed 2023-03-16]
35. Nilsson E, Sverker A, Bendtsen P, Eldh AC. A human, organization, and technology perspective on patients' experiences of a chat-based and automated medical history-taking service in primary health care: interview study among primary care patients. J Med Internet Res 2021;23(10):e29868 [FREE Full text] [doi: 10.2196/29868] [Medline: 34661544]

XSL•FO
RenderX

36. Turner A, Morris R, Rakhra D, Stevenson F, McDonagh L, Hamilton F, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2022;72(715):e128-e137 [FREE Full text] [doi: 10.3399/BJGP.2021.0426] [Medline: 34903520]
37. Atherton H, Brant H, Ziebland S, Bikker A, Campbell J, Gibson A, et al. The potential of alternatives to face-to-face consultation in general practice, and the impact on different patient groups: a mixed-methods case study. Health Serv Deliv Res 2018;6(20):1-200. [doi: 10.3310/hsdr06200] [Medline: 29889485]
38. Farr M, Banks J, Edwards HB, Northstone K, Bernard E, Salisbury C, et al. Implementing online consultations in primary care: a mixed-method evaluation extending normalisation process theory through service co-production. BMJ Open 2018;8(3):e019966 [FREE Full text] [doi: 10.1136/bmjopen-2017-019966] [Medline: 29555817]
39. Using online consultations in primary care: implementation toolkit. NHS England. URL: https://www.england.nhs.uk [accessed 2024-06-20]
40. Office for National Statistics. People living in deprived neighbourhoods. In: Ethnicity Facts and Figures. UK: GOV.UK; 2020.
41. Nussbaum C, Massou E, Fisher R, Morciano M, Harmer R, Ford J. Inequalities in the distribution of the general practice workforce in England: a practice-level longitudinal analysis. BJGP Open 2021;5(5):BJGPO.2021.0066 [FREE Full text] [doi: 10.3399/BJGPO.2021.0066] [Medline: 34404634]
42. Fisher R, Dunn P, Asaria M, Thorlby R. Briefing: Level or Not?. England, UK: The Health Foundation; 2020.
43. Hollman AK, Obermier TR, Burger PR. Rural measures: a quantitative study of the rural digital divide. J Inf Policy 2021;11:176-201.
44. Vestesson EM, de Corte KLA, Crellin E, Ledger J, Bakhai M, Clarke GM. Consultation rate and mode by deprivation in English general practice from 2018 to 2022: population-based study. JMIR Public Health Surveill 2023;9:e44944 [FREE Full text] [doi: 10.2196/44944] [Medline: 37129943]
45. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (create). Psychol Aging 2006;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
46. Zanaboni P, Fagerlund AJ. Patients' use and experiences with e-consultation and other digital health services with their general practitioner in Norway: results from an online survey. BMJ Open 2020;10(6):e034773 [FREE Full text] [doi: 10.1136/bmjopen-2019-034773] [Medline: 32554721]
47. Chappell P, Dias A, Bakhai M, Ledger J, Clarke GM. How is primary care access changing? a retrospective, repeated cross-sectional study of patient-initiated demand at general practices in England using a modern access model, 2019-2022. BMJ Open 2023;13(8):e072944 [FREE Full text] [doi: 10.1136/bmjopen-2023-072944] [Medline: 37591638]
48. Leach B, Parkinson S, Gkousis E, Abel G, Atherton H, Campbell J, et al. Digital facilitation to support patient access to web-based primary care services: scoping literature review. J Med Internet Res 2022;24(7):e33911 [FREE Full text] [doi: 10.2196/33911] [Medline: 35834301]
49. Linder-Pelz S. Toward a theory of patient satisfaction. Soc Sci Med 1982;16(5):577-582. [doi: 10.1016/0277-9536(82)90311-2] [Medline: 7100990]
50. Ogden J, Jain A. Patients' experiences and expectations of general practice: a questionnaire study of differences by ethnic group. Br J Gen Pract 2005;55(514):351-356 [FREE Full text] [Medline: 15904553]
51. Woodward A, Davies N, Walters K, Nimmons D, Stevenson F, Protheroe J, et al. Self-management of multiple long-term conditions: a systematic review of the barriers and facilitators amongst people experiencing socioeconomic deprivation. PLoS One 2023;18(2):e0282036 [FREE Full text] [doi: 10.1371/journal.pone.0282036] [Medline: 36809286]
52. Anhang Price R, Quigley DD, Hargraves JL, Sorra J, Becerra-Ornelas AU, Hays RD, et al. A systematic review of strategies to enhance response rates and representativeness of patient experience surveys. Med Care 2022;60(12):910-918 [FREE Full text] [doi: 10.1097/MLR.0000000000001784] [Medline: 36260705]
53. Roland M, Elliott M, Lyratzopoulos G, Barbiere J, Parker RA, Smith P, et al. Reliability of patient responses in pay for performance schemes: analysis of national general practitioner patient survey data in England. BMJ 2009;339:b3851 [FREE Full text] [doi: 10.1136/bmj.b3851] [Medline: 19808811]
54. Saunders CL, Elliott MN, Lyratzopoulos G, Abel GA. Do differential response rates to patient surveys between organizations lead to unfair performance comparisons? evidence from the English cancer patient experience survey. Med Care 2016;54(1):45-54 [FREE Full text] [doi: 10.1097/MLR.0000000000000457] [Medline: 26595223]

Abbreviations

COREC: Central Office for Research Ethics Committee
FP practices: practices using the free text input with an embedded single workflow online consultation system
FT OC system: free text input with an embedded single workflow online consultation system
GP: general practitioner
GPPS: General Practice Patient Survey

XSL•FO
RenderX

MT practices: practices using the mixed text input with variation in implemented workflow online consultation system
MT OC system: mixed text input with variation in implemented workflow online consultation system
NHS: National Health Service
OC: online consultation
OR: odds ratio

Edited by A Mavragani; submitted 28.07.23; peer-reviewed by E Nilsson, M Blanker; comments to author 27.10.23; revised version received 22.12.23; accepted 17.04.24; published 26.07.24.

Please cite as:
Ge X, Chappell P, Ledger J, Bakhai M, Clarke GM
The Use of Online Consultation Systems and Patient Experience of Primary Care: Cross-Sectional Analysis Using the General Practice Patient Survey
J Med Internet Res 2024;26:e51272
URL: https://www.jmir.org/2024/1/e51272
doi:10.2196/51272
PMID:

©Xiaochen Ge, Paul Chappell, Jean Ledger, Minal Bakhai, Geraldine M Clarke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Virtual Clinic Telehealth Abortion Services in the United States One Year After Dobbs: Landscape Review

Leah R Koenig[1,2,3], MSPH; Jennifer Ko[1,2], MLIS; Ushma D Upadhyay[1,2], MPH, PhD

[1]Advancing New Standards in Reproductive Health, Department of Obstetrics, Gynecology & Reproductive Sciences, University of California, San Francisco, Oakland, CA, United States

[2]Center for Gender and Health Justice, Global Health Institute, University of California, Oakland, CA, United States

[3]Department of Epidemiology and Biostatistics, Univeristy of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Ushma D Upadhyay, MPH, PhD
Advancing New Standards in Reproductive Health
Department of Obstetrics, Gynecology & Reproductive Sciences
University of California, San Francisco
1330 Broadway
Suite 1100
Oakland, CA, 94612
United States
Phone: 1 415 353 4626
Email: ushma.upadhyay@ucsf.edu

Abstract

Background: Telehealth abortion has taken on a vital role in maintaining abortion access since the *Dobbs v. Jackson Women's Health Organization* Supreme Court decision. However, little remains known about the landscape of new telehealth-only virtual clinic abortion providers that have expanded since telehealth abortion first became widely available in the United States in 2021.

Objective: This study aimed to (1) document the landscape of telehealth-only virtual clinic abortion care in the United States, (2) describe changes in the presence of virtual clinic abortion services between September 2022, following the *Dobbs* decision, and June 2023, and (3) identify structural factors that may perpetuate inequities in access to virtual clinic abortion care.

Methods: We conducted a repeated cross-sectional study by reviewing web search results and abortion directories to identify virtual abortion clinics in September 2022 and June 2023 and described changes in the presence of virtual clinics between these 2 periods. In June 2023, we also described each virtual clinic's policies, including states served, costs, patient age limits, insurance acceptance, financial assistance available, and gestational limits.

Results: We documented 11 virtual clinics providing telehealth abortion care in 26 states and Washington DC in September 2022. By June 2023, 20 virtual clinics were providing services in 27 states and Washington DC. Most (n=16) offered care to minors, 8 provided care until 10 weeks of pregnancy, and median costs were US $259. In addition, 2 accepted private insurance and 1 accepted Medicaid, within a limited number of states. Most (n=16) had some form of financial assistance available.

Conclusions: Virtual clinic abortion providers have proliferated since the *Dobbs* decision. We documented inequities in the availability of telehealth abortion care from virtual clinics, including age restrictions that exclude minors, gestational limits for care, and limited insurance and Medicaid acceptance. Notably, virtual clinic abortion care was not permitted in 11 states where in-person abortion is available.

(J Med Internet Res 2024;26:e50749) doi:10.2196/50749

KEYWORDS

medication abortion; telehealth; virtual clinics; abortion; access; policy; health equity

Introduction

Following the 2022 *Dobbs v. Jackson Women's Health Organization* US Supreme Court decision, demand for abortion surged in states where abortion care remains legally accessible [1]. Telehealth abortion has emerged as an important model of abortion provision in the US in recent years. In 2021, the US Food and Drug Administration (FDA) lifted a restriction that required mifepristone, the first drug in the medication abortion

XSL•FO
RenderX

regimen, to be dispensed only inside medical facilities, thus allowing direct-to-patient telehealth abortion care to expand. As of December 2023, telehealth constituted 19% of abortions in the US health care system [1].

Telehealth reduces geographic barriers to abortion and decreases wait times to care [2-4]. However, nearly all Southern and Midwest states ban abortion entirely or permit abortion but prohibit telehealth for abortion, limiting telehealth's potential to help maintain abortion access. Such restrictions on telehealth include in-person counseling, ultrasound, or other testing requirements and prohibitions of telehealth for abortion [5].

Direct-to-patient telehealth abortion is safe and effective [6-9]. Patients typically complete synchronous (over videoconferencing or a phone call) or asynchronous (using secure messaging) screening with a clinician to assess for medical eligibility. Once deemed eligible, patients are mailed medications, typically from a mail-order pharmacy. They then take the medications, pass the pregnancy, and complete follow-up interactions with their provider, from home or another place they choose [10].

Virtual abortion clinics—telehealth abortion providers without brick-and-mortar facilities in the state where they are providing abortion care—have also proliferated. However, because virtual abortion clinics are so new, little is known about their availability, reach, and policies. We aimed to document the landscape of virtual clinic abortion care in the US, to describe changes in the availability of virtual abortion clinics over time between September 2022, just after the *Dobbs* decision, and June 2023, and to identify structural factors that may perpetuate inequities in access to virtual clinic abortion care.

Methods

This repeated cross-sectional landscape review involved web searches (search terms listed in Textbox 1) and 3 abortion directory websites (Abortionfinder, Ineedana, and Plan C) that document abortion service availability. We synthesized all available information from the 3 abortion directories and clinic websites.

Virtual clinics were eligible if they provided telehealth abortion care within the US health care system during the search period. We included virtual clinics that provided telehealth abortion services in states where they did not have a brick-and-mortar location.

LRK and JK conducted the first and second searches in September 2022 and June 2023, respectively. They conducted independent searches to identify eligible virtual clinics and document their policies. Discrepancies were then resolved iteratively through a third collaborative search.

Our primary measures of interest were the number of unique virtual clinics operational at the time of each search and the number of states served. These were assessed in both September 2022 and June 2023. Our secondary measures of interest were the service policies of each virtual clinic, including ages served, costs, insurance and Medicaid acceptance, gestational limits, languages offered, whether the service offered synchronous care (involving video or phone interactions) and/or asynchronous care (entirely over secure messaging), and whether the virtual clinic provided medication abortion through advanced provision. These were assessed only in June 2023. We described each outcome using descriptive statistics including frequencies, percentages, medians, and modes. We adhered to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.

Textbox 1. Search terms.

"telehealth"

"telemedicine"

"online"

"abortion pills"

"medication abortion"

"mifepristone"

"misoprostol"

Results

In September 2022, 11 virtual clinics provided telehealth abortion care in 26 states and Washington DC (Table S1 in Multimedia Appendix 1). By June 2023, 20 virtual clinics provided telehealth abortion care in 27 states and Washington DC (Figure 1). In June 2023, 23 states had no virtual clinic providers, among which 14 had banned abortion entirely and 9 allowed abortion but had restrictions on telehealth for abortion.

In June 2023, gestational limits ranged from 10 weeks, 0 days to 13 weeks, 0 days, although the modal gestational limit was 10 weeks, 0 days (n=8; Table 1). Among the 20 virtual clinics, 4 did not serve patients younger than 18 years. Overall, 8 virtual clinics provided only synchronous care, 6 provided asynchronous care, and 6 offered patients a choice between the two. In addition, 6 offered medication abortion by advanced provision.

Costs for virtual clinic telehealth abortion care ranged from US $90 to US $600, and median costs were US $259. Few (n=2) accepted private insurance and 2 accepted Medicaid in certain states. Most virtual clinics (n=15) offered some form of financial assistance to patients. About half offered services in languages

XSL•FO
RenderX

other than English (n=11). However, while some of the virtual clinic providers were multilingual, many platforms were only available in English, and some charged additional costs for translations.

Figure 1. Map of virtual clinic telehealth abortion care availability in the United States, September 2022 (A) and June 2023 (B). The figure was created with mapchart.net.

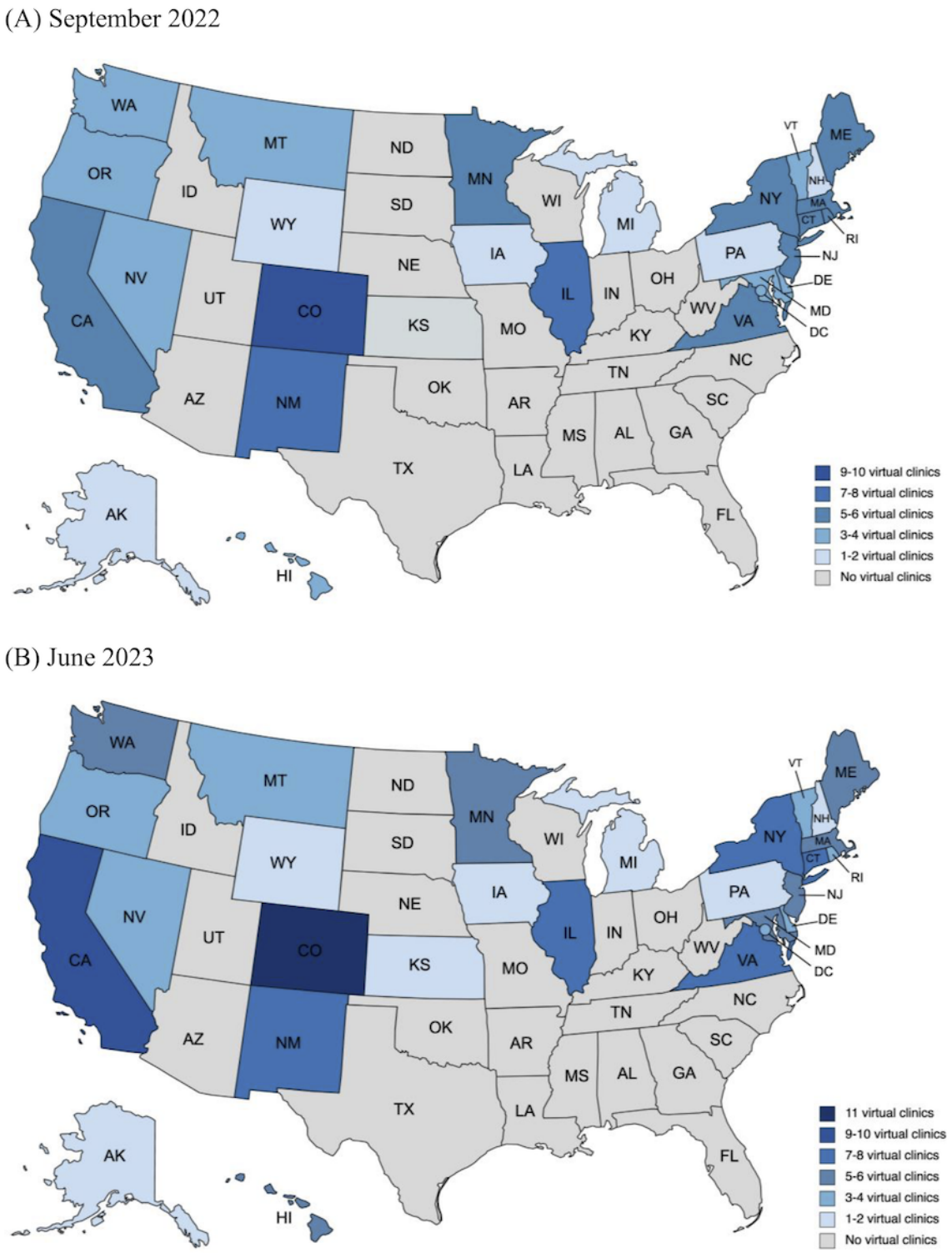

XSL•FO
RenderX

Table 1. Availability and policies of virtual abortion clinics in the United States, June 2023.

Clinic name	Gestational limit	Telehealth care model	Age limit	Languages offered	Advanced provision	Cost (in US $)	Accepted private insurance	Accepted Medicaid	Financial assistance available	States served
145 Abortion Telemedicine	13 weeks 0 days	Synchronous	None	Over 500 languages with translational service (US $100)	No	145	No	No	No	18 (CA, CO, CT, DE, DC, HI, IL, ME, MA, MT, NJ, NM, NY, OR, RI, VT, VA, and WA)
Aid Access	12 weeks 0 days	Asynchronous	13+	English	Yes	150	No	No	Yes	20 (AK, CA, CO, DC, HI, IL, ME, MD, MA, MI, MN, NV, NJ, NM, NY, OR, RI, VT, VA, and WA)
Abortion on Demand	10 weeks 0 days	Synchronous	18+	English	No	239-389	No	No	Yes	24 (CA, CO, CT, DE, DC[a], HI, IL, KS, ME, MD, MA, MN, MT, NV, NH, NJ, NM, NY, OR, PA[a], RI, VT, VA, and WA)
carafem	11 weeks 0 days	Synchronous	None	Spanish translator available by phone	No	249	No	IL only	Yes	15 (CO, CT, DE, DC, IL, IA, ME, MA, MN, NV, NJ, NM, RI, VT, and VA)
Choices Rising	10 weeks 0 days	Asynchronous	15+	English, Spanish	No	297	No	CA only	No	2 (CA and HI)
Choix	10 weeks 5 days	Asynchronous	15+ (certain states)	English, Spanish	Yes	399	No	No	Yes	6 (CA, CO, IL, ME, NM, and VA)
Forward Midwifery	12 weeks 3 days	Synchronous or asynchronous	None	English, Spanish, Translation Services	No	150	No	No	Yes	6 (CA, CO, MD, MA, NM, and OR)
Hey Jane	10 weeks 0 days	Synchronous or asynchronous	18+	English	No	199-409	Yes	No	Yes	9 (CA, CO, CT, IL, NJ, NM, NY, VA, and WA)
Jennifer Boyd	10 weeks 0 days	Synchronous	None	English	No	550	No	No	Yes	2 (CT and NY)
Juniper Midwifery	11 weeks 0 days	Synchronous or asynchronous	None	English, Spanish	Yes	200	No	No	Yes	6 (CA, CO, CT, NJ, NM, and NY)
Just the Pill	10 weeks 0 days	Synchronous	None	English, Spanish	No	350	No	No	Yes	4 (CO, MN, MT, and WY)
Lilith Care	12 weeks 0 days	Asynchronous	18+	English	No	140	No	No	No	3 (HI, MA, and RI)
Luna Flow Health	10 weeks 6 days	Synchronous	13+	English	Yes	195-395	No	No	Yes	1 (CA)

XSL•FO
RenderX

Clinic name	Gestational limit	Telehealth care model	Age limit	Languages offered	Advanced provision	Cost (in US $)	Accepted private insurance	Accepted Medicaid	Financial assistance available	States served
Maitri Wellness	12 weeks 0 days	Synchronous or asynchronous	18+	English, Hindi, Spanish	Yes	150-500	No	No	No	3 (MI, NJ, and NY)
Metro Area	12 weeks 6 days	Synchronous or asynchronous	None	English	Yes	165	Yes	No	Yes	3 (DC, MD, and VA)
Joan Fleischman	10 weeks 0 days	Synchronous	None	English	No	600	No	No	Yes	3 (CT, NJ, and NY)
Pills by post	12 weeks 0 days	Asynchronous	14+	English, Spanish	No	150	No	No	Yes	4 (CO, IL, MN, and NY)
The Satanic Temple	11 weeks, 0 days	Synchronous	17+	English	No	90	No	No	Yes	1 (NM)
Sunny	10 weeks 0 days	Asynchronous	16+	English, Spanish, Translation Services	No	269	No	No	No	2 (CA and NY)
Wisp	10 weeks 0 days	Synchronous or asynchronous	18+	English	No	200	No	No	Yes	8 (CO, CT, IL, ME, MD, NM, NY, and WA)

[a]Telehealth abortion care available with in-person pickup in a neighboring state for certain states.

Discussion

Principal Findings

Virtual clinics increased dramatically in the United States between 2022 and 2023, the year after the *Dobbs* decision; they doubled in number and increased service provision to reach 27 states and Washington, DC. Our results show that telehealth abortion is an important and rapidly expanding model of abortion care.

We documented several structural barriers in the landscape of virtual clinic abortion that could be addressed through changes in virtual clinic policies. Most virtual clinics we examined offered medication abortion care up to 10 weeks of pregnancy, although medication abortion is commonly provided off-label in the United States until 11 weeks and is recommended for use by the World Health Organization up to 12 weeks [11,12]. As abortion bans created additional logistical hurdles that delay abortion care, virtual clinics should expand gestational limits to serve patients throughout the first trimester. Future research should assess the safety of telehealth abortion provided at gestations beyond 11 weeks.

Some virtual clinics had minimum age requirements that are not legally mandated in the states they serve. These restrictions limit abortion access for adolescents, who face even greater barriers to abortion and stand to benefit from the privacy and ability to avoid travel that telehealth offers [13]. In addition, few virtual clinics accepted private insurance or Medicaid, highlighting an important accessibility gap. Most virtual clinics we identified offered only asynchronous or synchronous care. Research has demonstrated that both are safe and effective and that each offers unique benefits to patients [4,7,9]. It is also critical that virtual abortion clinics offer multilingual services to reach immigrant and undocumented populations—groups for whom travel for abortion care may be especially difficult. Future qualitative research can highlight changes needed to enable virtual clinics to remove these barriers to care.

We also identified state policy changes that could help reduce inequities in telehealth abortion care. First, states can take action to improve equitable access to telehealth medical care more broadly through actions such as increasing coverage for asynchronous telehealth care and promoting equitable access to reliable internet connection. Second, they can promote access to telehealth abortion by improving Medicaid and insurance reimbursement [14]. Third, in-person counseling, ultrasound, and other requirements are not based on medical evidence, and states should remove them to allow telehealth abortion in all states where abortion is legal.

This analysis had several limitations. First, we did not document virtual clinic policies during the first search in September 2022, and therefore we could not examine changes in these policies over time. Second, we may have missed virtual abortion clinics not documented on the websites we examined or in our searches. Third, in some cases, the policies listed on virtual clinic websites or abortion directories may differ from actual practice. For example, advance provision may be available from providers who did not publicize those services. Fourth, this review does not encompass telehealth services provided by brick-and-mortar clinics. However, these results provide a novel picture of the landscape of telehealth abortion care from virtual clinics in the United States in the year following the *Dobbs* decision. By conducting a landscape review of the information available from web searches and abortion directories, we provide a sense of the information available to prospective abortion patients across the United States.

XSL•FO
RenderX

Virtual clinics are emerging as key providers in the US abortion landscape in the face of dire restrictions on abortion access. In 2022 and 2023, several states that protect abortion passed "shield laws," which create legal protections for clinicians providing telehealth abortion to patients in states with bans [15]. These laws are further expanding the role of telehealth by allowing patients residing in states where abortion is banned to access telehealth abortion care within the US health care system. Since our review in June 2023, several virtual clinics have begun to offer telehealth abortion care to residents of states with abortion bans under these shield laws [16,17]. However, as long as patients experience legal risks using such services, access will never be equitable. To ensure health equity, all people should have access to abortion care offered by virtual clinics regardless of the state they live in.

Conclusion

Virtual abortion clinics have grown in prominence in the US since 2022. As of 2023, many virtual clinics retained policies that may limit equitable access to abortion care, such as minimum age requirements and gestational limits not required by law and lack of insurance acceptance. Given the increasingly restricted US abortion landscape, it is critical to address barriers to telehealth abortion care to ensure as many people as possible can access abortion care.

Acknowledgments

We appreciate the "I Need an A," "Abortion Finder," and "Plan C" teams for collecting and collating vital data on abortion care availability. This research was supported by the BaSe Family Fund, Erik E. and Edith H. Bergstrom Foundation, Isabel Allende Foundation, Jess Jacobs, Kahle/Austin Foundation, Lisa and Douglas Goldman Fund, and Preston-Werner Ventures. LRK was funded in part by training grants from the National Institute on Minority Health and Health Disparities of the National Institutes of Health (award T32MD015070) and the National Institute of Child Health and Human Development (award 1F31HD111277) during the study period. The funders had no role in the study design, data collection, and analysis, the writing of the report, or the decision to submit the paper for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Availability of virtual abortion clinics in the United States, September 2022.
[DOCX File , 14 KB - jmir_v26i1e50749_app1.docx]

References

1. #WeCount May 2024 report. #WeCount May 2024 report.: Society of Family Planning; 2024 May 14. URL: https://societyfp.org/wp-content/uploads/2024/05/WeCount-report-6-May-2024-Dec-2023-data_Final.pdf [accessed 2024-06-22]
2. Koenig LR, Becker A, Ko J, Upadhyay UD. The role of telehealth in promoting equitable abortion access in the United States: spatial analysis. JMIR Public Health Surveill 2023;9:e45671 [FREE Full text] [doi: 10.2196/45671] [Medline: 37934583]
3. Aiken AR, Upadhyay UD. The future of medication abortion in a post-Roe world. BMJ 2022 Jun 02;377:o1393. [doi: 10.1136/bmj.o1393] [Medline: 35654599]
4. Koenig LR, Ko J, Valladares ES, Coeytaux FM, Wells E, Lyles CR, et al. Patient acceptability of telehealth medication abortion care in the United States, 2021 2022: a cohort study. Am J Public Health 2024;114(2):241-250. [doi: 10.2105/AJPH.2023.307437] [Medline: 38237103]
5. The Availability and Use of Medication Abortion. Kaiser Family Foundation. 2024 Jun 13. URL: https://www.kff.org/womens-health-policy/fact-sheet/the-availability-and-use-of-medication-abortion/ [accessed 2024-06-22]
6. Upadhyay UD, Raymond EG, Koenig LR, Coplon L, Gold M, Kaneshiro B, et al. Outcomes and safety of history-based screening for medication abortion: a retrospective multicenter cohort study. JAMA Intern Med 2022;182(5):482-491 [FREE Full text] [doi: 10.1001/jamainternmed.2022.0217] [Medline: 35311911]
7. Upadhyay UD, Koenig LR, Meckstroth KR. Safety and efficacy of telehealth medication abortions in the US during the COVID-19 pandemic. JAMA Netw Open 2021 Aug 02;4(8):e2122320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.22320] [Medline: 34427682]
8. Anger HA, Raymond EG, Grant M, Haskell S, Boraas C, Tocce K, et al. Clinical and service delivery implications of omitting ultrasound before medication abortion provided via direct-to-patient telemedicine and mail in the U.S. Contraception 2021 Dec;104(6):659-665. [doi: 10.1016/j.contraception.2021.07.108] [Medline: 34329607]
9. Upadhyay UD, Koenig LR, Meckstroth K, Ko J, Valladares ES, Biggs MA. Effectiveness and safety of telehealth medication abortion in the USA. Nat Med 2024;30(4):1191-1198. [doi: 10.1038/s41591-024-02834-w] [Medline: 38361123]
10. Raymond EG, Grossman D, Mark A, Upadhyay UD, Dean G, Creinin MD, et al. Commentary: No-test medication abortion: A sample protocol for increasing access during a pandemic and beyond. Contraception 2020 Jun;101(6):361-366 [FREE Full text] [doi: 10.1016/j.contraception.2020.04.005] [Medline: 32305289]

XSL•FO
RenderX

11. NAF 2022 Clinical Policy Guidelines for Abortion Care. National Abortion Federation. URL: https://prochoice.org/wp-content/uploads/2022-CPGs.pdf [accessed 2024-06-22]
12. Abortion care guideline. Human Reproduction Programme. Geneva: World Health Organization; 2022. URL: https://www.who.int/publications/i/item/9789240039483
13. Myers CK, Ladd D. Did parental involvement laws grow teeth? The effects of state restrictions on minors' access to abortion. J Health Econ 2020 May;71:102302 [FREE Full text] [doi: 10.1016/j.jhealeco.2020.102302] [Medline: 32135395]
14. Upadhyay UD, Schroeder R, Kaller S, Stewart C, Berglas NF. Pricing of medication abortion in the United States, 2021–2023. Perspect Sex Reprod Health 2024 Jul 02. [doi: 10.1111/psrh.12280]
15. Cohen DS, Donley G, Rebouché R. Abortion Shield Laws. NEJM Evid 2023 Apr;2(4):EVIDra2200280. [doi: 10.1056/EVIDra2200280] [Medline: 38320012]
16. Belluck P, Bazelon E. New York Passes Bill to Shield Abortion Providers Sending Pills Into States With Bans. URL: https://www.nytimes.com/2023/06/20/health/abortion-shield-law-new-york.html [accessed 2023-08-08]
17. Kitchener C. Blue-state doctors launch abortion pill pipeline into states with bans. Washington Post. 2023. URL: https://www.washingtonpost.com/politics/2023/07/19/doctors-northeast-launch-abortion-pill-pipeline-into-states-with-bans/ [accessed 2024-06-21]

Abbreviations

FDA: US Food and Drug Administration
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by T de Azevedo Cardoso; submitted 13.07.23; peer-reviewed by A Swartzendruber, J Seymour, D Constant, L Dodge; comments to author 22.11.23; revised version received 04.01.24; accepted 06.06.24; published 05.08.24.

Please cite as:
Koenig LR, Ko J, Upadhyay UD
Virtual Clinic Telehealth Abortion Services in the United States One Year After Dobbs: Landscape Review
J Med Internet Res 2024;26:e50749
URL: https://www.jmir.org/2024/1/e50749
doi:10.2196/50749
PMID:

©Leah R Koenig, Jennifer Ko, Ushma D Upadhyay. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Potential, Pitfalls, and Future Directions for Remote Monitoring of Chronic Respiratory Diseases: Multicenter Mixed Methods Study in Routine Cystic Fibrosis Care

Martinus C Oppelaar[1], MD; Yvette Emond[2], MSc, PhD; Michiel A G E Bannier[3], MD, PhD; Monique H E Reijers[4], MD, PhD; Hester van der Vaart[5], MD, PhD; Renske van der Meer[6], MD, PhD; Josje Altenburg[7], MD, PhD; Lennart Conemans[8,9], MD; Bart L Rottier[10,11], MD, PhD; Marianne Nuijsink[12], MD, PhD; Lara S van den Wijngaart[1], MD, PhD; Peter J F M Merkus[1], MD, PhD; Maud Heinen[2*], RN, PhD; Jolt Roukema[1*], MD, PhD

[1]Department of Pediatric Pulmonology, Amalia Children's Hospital, Radboud University Medical Center, Nijmegen, Netherlands

[2]IQ Health Science Department, Radboud University Medical Center, Nijmegen, Netherlands

[3]Department of Paediatric Pulmonology, MosaKids Children's Hospital, Maastricht University Medical Centre+, Maastricht, Netherlands

[4]Department of Pulmonology, Radboud University Medical Center, Nijmegen, Netherlands

[5]Department of Pulmonary Diseases, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[6]Department of Pulmonology, Haga Teaching Hospital, The Hague, Netherlands

[7]Department of Respiratory Medicine, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, Netherlands

[8]Department of Respiratory Medicine, Maastricht University Medical Centre+, Maastricht, Netherlands

[9]Division of Respiratory & Age-related Health, Department of Respiratory Medicine, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht, Netherlands

[10]Department of Pediatric Pulmonology and Pediatric Allergology, University Medical Center Groningen, Beatrix Children's Hospital, University of Groningen, Groningen, Netherlands

[11]Groningen Research Institute for Asthma and COPD, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[12]Haga Teaching Hospital, Juliana Children's Hospital, The Hague, Netherlands

[*]these authors contributed equally

Corresponding Author:
Martinus C Oppelaar, MD
Department of Pediatric Pulmonology
Amalia Children's Hospital
Radboud University Medical Center
Geert Grooteplein 10
Nijmegen, 6500 HB
Netherlands
Phone: 31 24 361 44 30
Email: marc.oppelaar@radboudumc.nl

Abstract

Background: The current literature inadequately addresses the extent to which remote monitoring should be integrated into care models for chronic respiratory diseases (CRDs).

Objective: This study examined a remote monitoring program (RMP) in cystic fibrosis (CF) by exploring experiences, future perspectives, and use behavior over 3 years, with the aim of developing future directions for remote monitoring in CRDs.

Methods: This was a mixed methods, multicenter, observational study in 5 Dutch CF centers following a sequential explanatory design. Self-designed questionnaires using the technology acceptance model were sent out to people with CF who had a minimum of 12 months of experience with the RMP and local health care professionals (HCPs). Questionnaire outcomes were used to inform semistructured interviews with HCPs and people with CF. Qualitative findings were reported following the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist. Anonymous data on use frequency of all people with CF were analyzed.

Results: Between the second quarter of 2020 and the end of 2022, a total of 608 people with CF were enrolled in the program, and a total of 9418 lung function tests and 2631 symptom surveys were conducted. In total, 65% (24/37) of HCPs and 89% (72/81)

XSL•FO
RenderX

of people with CF responded to the questionnaire, and 7 HCPs and 12 people with CF participated in semistructured interviews. Both people with CF and HCPs were positive about remote monitoring in CF care and found the RMP a good addition to daily care (people with CF: 44/72, 61%; HCPs: 21/24, 88%). Benefits ranged from supporting individual patients to reducing health care consumption. The most valued monitoring tool was home spirometry by both people with CF (66/72, 92%) and HCPs (22/24, 92%). Downsides included the potential to lose sight of patients and negative psychosocial effects, as 17% (12/72) of people with CF experienced some form of stress due to the RMP. A large majority of people with CF (59/72, 82%) and HCPs (22/24, 92%) wanted to keep using the RMP in future, with 79% (19/24) of HCPs and 75% (54/72) of people with CF looking forward to more replacement of in-person care with digital care during periods of well-being. Future perspectives for the RMP were centered on creating hybrid care models, personalizing remote care, and balancing individual benefits with monitoring burden.

Conclusions: Remote monitoring has considerable potential in supporting people with CF and HCPs within the CF care model. We identified 4 practice-based future directions for remote monitoring in CF and CRD care. The strategies, ranging from patient driven to prediction driven, can help clinicians, researchers, and policy makers navigate the rapidly changing digital health field, integrate remote monitoring into local care models, and align remote care with patient and clinician needs.

(J Med Internet Res 2024;26:e54942) doi:10.2196/54942

KEYWORDS

telemonitoring; digital health; chronic respiratory diseases; telespirometry; interviews; mixed methods; qualitative study; remote monitoring; evaluation; cystic fibrosis; pediatrics; mixed method; observational study; health care professionals; semistructured interview; psychosocial; clinicians; researchers; policy makers; telehealth

Introduction

Background

Chronic respiratory disease (CRD) management needs to undergo rapid changes to respond to growing challenges in the coming decades. CRDs are one of the largest contributors to the global noncommunicable disease burden, and their prevalence is increasing worldwide [1]. At the same time, health systems are under pressure due to rising health care costs and health worker shortages. Consequently, there are increasing calls for innovative solutions such as remote monitoring to relieve some of this burden and ensure health care continuity for patients with asthma, chronic obstructive pulmonary disease, and other CRDs [2]. The COVID-19 pandemic has shown that remote monitoring in CRD management can be feasible and effective, but there are still uncertainties regarding to what extent remote monitoring should be integrated into existing care models [3-5].

The field of cystic fibrosis (CF) provides a perfect case study to analyze this problem. CF is an autosomal recessive hereditary disease caused by a defect in the gene coding for the CF transmembrane conductance regulator (CFTR) protein [6]. CF is a multisystem disease, but progressive pulmonary deterioration due to chronic inflammation and recurrent pulmonary infections is the most frequent cause of morbidity and mortality for people with CF [6]. Up until recently, the median age of mortality was between 30 and 40 years but is now rapidly increasing due to new CFTR protein–modulating drugs [6]. These drugs specifically target the cellular defect underlying CF and lead to significant clinical improvements that will result in a larger adult CF population in the near future [7,8]. Moreover, CF will more closely resemble other CRDs as pulmonary deterioration is slowed down significantly [9]. Care models need to quickly adjust to these changes, potentially with digital solutions [9]. Therefore, the changes we witness in digital health in CF today could provide valuable guidance for the implementation of digital health in the wider field of CRD management.

Until now, a prominent focus of the literature on remote monitoring in CF has been the early identification of pulmonary exacerbations using remote monitoring. Some studies have shown that this may allow for earlier identification of pulmonary exacerbations at a population level, but there is no evidence that remote monitoring also improves clinical outcomes or quality of life for individual patients [10]. Moreover, although remote symptom and spirometry monitoring appears highly feasible, adequate uptake of remote lung function monitoring is rarely sustained over time [10,11]. As a result, the current literature now calls for a different direction, with increasing attention to evidence for improvements in other domains such as health system strengthening, patient empowerment, and workload of health care professionals (HCPs). At the same time, there is a pressing need for more guidance on the integration of remote monitoring into existing CF care models [5,8].

Objectives

This study is part of the Airlift-CF project, which provides remote care for patients with asthma and CF in the Netherlands [12-14]. Over 40% of all Dutch people with CF in 5 Dutch CF centers are using this program in their regular care to monitor symptoms and lung function at home. The aims of this study were to determine the role of remote monitoring in CF care models by exploring experiences, future perspectives, and actual remote monitoring program (RMP) use during the 3 years after the onset of the COVID-19 pandemic and develop future directions for remote monitoring in CF and other CRDs. The findings of this study will help researchers align their efforts with clinical needs and will empower patients, HCPs, and policy makers to make informed decisions about the use of remote monitoring in routine care within different contexts.

XSL•FO
RenderX

Methods

Study Design

This was a mixed methods, multicenter, observational study that followed a sequential explanatory design (ie, quantitative analyses were followed by qualitative analyses to provide in-depth explanations of the findings) guided by the Mixed Methods Appraisal Tool 2018 [15]. The Strengthening the Reporting of Observational Studies in Epidemiology checklist was used to design and report this study (Multimedia Appendix 1). The study was conducted in 5 of the 7 Dutch CF centers (Radboud University Medical Center; University Medical Center Groningen; Maastricht University Medical Center+; Academic Medical Center, Amsterdam; and Haga Hospital, The Hague).

eHealth Program

The RMP for CF was introduced in March 2020 to provide health care continuity for people with CF in response to the COVID-19 pandemic and was based on our preexisting RMP for pediatric asthma. No formal implementation project could be developed during this turbulent period. Instead, implementation was guided by our experience in implementation of remote monitoring for pediatric asthma [16]. The RMP is used to (1) monitor disease symptoms using a 7-item modified Fuchs questionnaire [17], (2) monitor lung function using a Bluetooth-connected portable spirometer (Spirobank Smart; Medical International Research), and (3) facilitate easy and secure patient-HCP contact. Further details are described in Multimedia Appendix 2 [12-14].

Participants

Both people with CF and HCPs were asked to participate in this study. People with CF were eligible for inclusion if they were aged ≥6 years and had used the RMP for at least 12 months. The criterion for experience was chosen because little to no evidence exists on the use of remote monitoring for CF for >12 months. Therefore, this study aimed to focus explicitly on the experiences and perspectives of long-term users as our previous experiences show that adaption and habituation to remote monitoring occurs over longer periods. Parents of younger children who were unable to participate in this study themselves were asked to provide their experiences instead. People with CF or their parents were invited to participate by their HCPs through the RMP when they met the inclusion criteria. HCPs (medical doctors and nurses) working with the RMP were also invited to participate. No sample size calculation was performed as this was not possible for this exploratory study. Recruitment of participants started in March 2022 and was scheduled to last 6 months.

Questionnaires

Participants received self-designed questionnaires based on the themes of perceived usefulness, perceived ease of use, intention to use, and use behavior from the technology acceptance model (TAM) [18,19]. Most questions were closed ended (ie, 5-item Likert scale). Some questions were open ended and focused on experienced advantages or disadvantages, suggestions for future use of the RMP, and incentives or disincentives (Multimedia Appendix 3). The content, relevance, and language of the questions were scrutinized by patient representatives of the Dutch CF Foundation and a small group of people with CF and HCPs. We used a modified 10-item System Usability Scale tailored to the study population to quantify perceived ease of use (Multimedia Appendix 3) [20,21]. Questionnaire invitations were sent via email, with automatic reminders after 1, 2, and 4 weeks.

Interviews

Semistructured interviews were performed to further explore and substantiate questionnaire results. The coauthors (MCO, YE, MH, PJFMM, and JR) discussed questionnaire results to reach a consensus on which questionnaire results warranted further investigation during the qualitative part of this study. These factors included (1) the relevance of the findings for clinical practice, (2) the unexplained ambiguity of the results (eg, differences between HCPs and people with CF or varying opinions within subgroups), (3) the potential impact of the findings on people with CF and HCPs, and (4) the frequency of themes that emerged in open-ended questions. After the identification of subjects was completed, interview guides were designed using the TAM framework and complementary domains from the work by Flottorp et al [18,19,22] (the full translated interview guides are available in Multimedia Appendix 4).

For each age group (6-12 years, 12-16 years, and >16 years), people with CF who had indicated interest in interviews in the questionnaire were invited randomly via email. Selection by age groups was chosen because age is an important factor in the focus and organization of CF care—care for younger children tends to be more focused on development, whereas care for adult people with CF is more focused on pulmonary health. Therefore, perspectives and experiences will likely differ between these groups. Moreover, by including different age groups, we also allowed for the inclusion of parents of people with CF, who likely have their own unique experiences and perspectives, hence providing a more valid representation of the target population. The coauthors recruited HCPs for interviews from their teams through purposive sampling based on experience and availability. Recruitment continued until data saturation was reached within subgroups. Interviews were held by video in Dutch for up to 60 minutes; moderated by the first author (MCO), who has a medical background; and audio recorded. Verbatim transcription of the recordings was carried out by a professional service.

Data on RMP Use

Anonymous data on frequency of use of all RMP users were extracted from January 2020 to December 2022. We described the monthly enrollment rate of participants over time, the monthly lung function rates, and the monthly symptom survey rates.

Data Analysis

Quantitative analyses were performed using SPSS Statistics (version 27; IBM Corp), and qualitative analysis of the interviews was performed using ATLAS.ti (version 22.0.11; ATLAS.ti Scientific Software Development GmbH). Questionnaire results were summarized for HCPs and people

XSL•FO
RenderX

with CF separately. Results of equivalent questionnaire questions were compared between these subgroups using the Fisher exact test. Open-ended questionnaire results were analyzed using open coding (MCO) and categorized into subthemes that emerged (MCO, PJFMM, and JR).

Transcripts were analyzed independently using open coding by MCO and YE. Themes that emerged were categorized into predefined themes from the TAM or into new themes. The codes and analyses were discussed until a consensus was reached (MCO, YE, and MH). Findings were reported following the COREQ (Consolidated Criteria for Reporting Qualitative Research; Multimedia Appendix 5 [23]) checklist.

Ethical Considerations

Local ethical committees waived formal approval considering the negligible burden of participation and absence of imposed risks (file number for local ethical committee Arnhem-Nijmegen region: 2021-13214). All eligible people with CF and their parents provided informed e-consent on the RMP before participation in the study in accordance with the Dutch Central Committee on Research Involving Human Subjects guidelines. All data were pseudonymized using an encrypted code that was only accessible to HCPs directly involved in the treatment of people with CF. Participation was voluntary and no compensation was provided to the participants.

Results

Demographics

A total of 81 people with CF gave informed consent to participate in this study. Of all participants, 89% (72/81) of people with CF and 65% (24/37) of HCPs responded to the questionnaires. The full questionnaire results are presented in Multimedia Appendix 3. We randomly invited 21 people with CF from the questionnaire respondents to participate in semistructured interviews. Due to the recruitment procedure of HCPs for interviews, the total number of invited HCPs was unknown. A subgroup of 12 people with CF and 7 HCPs participated in semistructured interviews until data saturation was reached. The demographics of the participants are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographics of participants.

	Questionnaire respondents	Interviewees
People with CF[a] (n=81), n (%)		
Overall	72 (89)	12 (15)
Sex (male), n (%)	34 (47)	5 (42)
Age (y), median (IQR)	31 (15.5-43.5)	15 (10.75-37.25)
Age distribution (y), n (%)		
6-12	8 (11)	3 (25)
12-18	12 (17)	4 (33)
18-30	14 (19)	1 (8)
≥30	38 (53)	4 (33)
Age (y), range	8-61	8-53
Hospital, n (%)		
Radboudumc	14 (19)	4 (33)
HagaZiekenhuis	13 (18)	2 (17)
Maastricht UMC+[b]	21 (29)	2 (17)
UMC Groningen	11 (15)	2 (17)
Amsterdam UMC	13 (18)	2 (17)
CFTR[c] genotype, n (%)		
F508del homozygous	46 (64)	9 (75)
F508del heterozygous	22 (31)	3 (25)
Other	4 (6)	0 (0)
Elexacaftor/tezacaftor/ivacaftor in 2022, n (%)	61 (85)	9 (75)
Pancreas enzyme use, n (%)	61 (85)	11 (92)
Months of using the program at recruitment, median (IQR)	21 (19-22)	21 (19.5-22)
Health care professionals (n=37), n (%)		
Overall	24 (65)	7 (19)
Specialty, n (%)		
Pediatric pulmonology	12 (50)	3 (43)
Pulmonology	10 (42)	3 (43)
Both	2 (8)	1 (14)
Profession, n (%)		
Medical doctor	12 (50)	2 (29)
Specialist nurse or nurse practitioner	12 (50)	5 (71)
Sex (male), n (%)	3 (12)	0 (0)
Age group (y), n (%)		
<30	2 (8)	0 (0)
30-35	2 (8)	0 (0)
35-40	8 (33)	2 (29)
40-45	4 (17)	3 (43)
45-50	0 (0)	0 (0)
>50	8 (33)	2 (29)
Hospital, n (%)		

XSL•FO
RenderX

	Questionnaire respondents	Interviewees
Radboudumc	6 (25)	3 (43)
HagaZiekenhuis	3 (13)	1 (14)
Maastricht UMC+	5 (21)	2 (29)
UMC Groningen	6 (25)	1 (14)
Amsterdam UMC	4 (17)	0 (0)
Years of work experience, median (IQR)	9 (4-17.5)	10 (1.5-18)

[a]CF: cystic fibrosis.

[b]UMC: University Medical Center.

[c]CFTR: cystic fibrosis transmembrane conductance regulator.

Qualitative Analyses

Open-ended questionnaire analysis resulted in 28 subthemes. Most questionnaire findings were included in the qualitative part of this study. Findings that were not included either had high agreement rates (eg, no people with CF or parents felt too closely watched by their CF team) or required no further qualitative explanations (eg, the devices that people wanted to be able to access the RMP on). Qualitative analyses of interviews resulted in 380 specific codes divided into 7 themes and 12 subthemes. The themes and subthemes are presented in Multimedia Appendix 6.

Use Behavior

Between the second quarter of 2020 and the end of 2022, a total of 608 people with CF were enrolled on the program, and a total of 9418 lung function tests and 2631 symptom surveys were conducted. Questionnaire respondents reported that they used the program weekly (7/72, 10%), monthly (19/72, 26%), during symptoms (28/72, 39%), or rarely (18/72, 25%). Figure 1 shows the monthly enrollment rate, lung function rate, and symptom survey rate for all users enrolled in the program between January 2020 and December 2022. In interviews, HCPs reported a reduction in use frequency over time and especially after the introduction of a new modulator therapy (elexacaftor/tezacaftor/ivacaftor [ETI]). HCPs generally categorized people with CF into 4 different user groups: those who had never started using the program, those who started but discontinued using the program due to difficulties (technical, psychosocial, or otherwise), those who used the program only on indication, or those who used the program regularly. Consequently, HCPs doubted whether the program was suitable for everyone (Textbox 1, quote 1).

XSL•FO
RenderX

Figure 1. (A) Waterfall plot of the number of new users by month. (B) The number of lung function tests by month. (C) The number of symptom surveys by month. The colored bars represent Dutch COVID-19 lockdown periods. The dashed line represents the introduction of the new modulator drug elexacaftor/tezacaftor/ivacaftor in the Netherlands. FEV1: forced expiratory volume in one second.

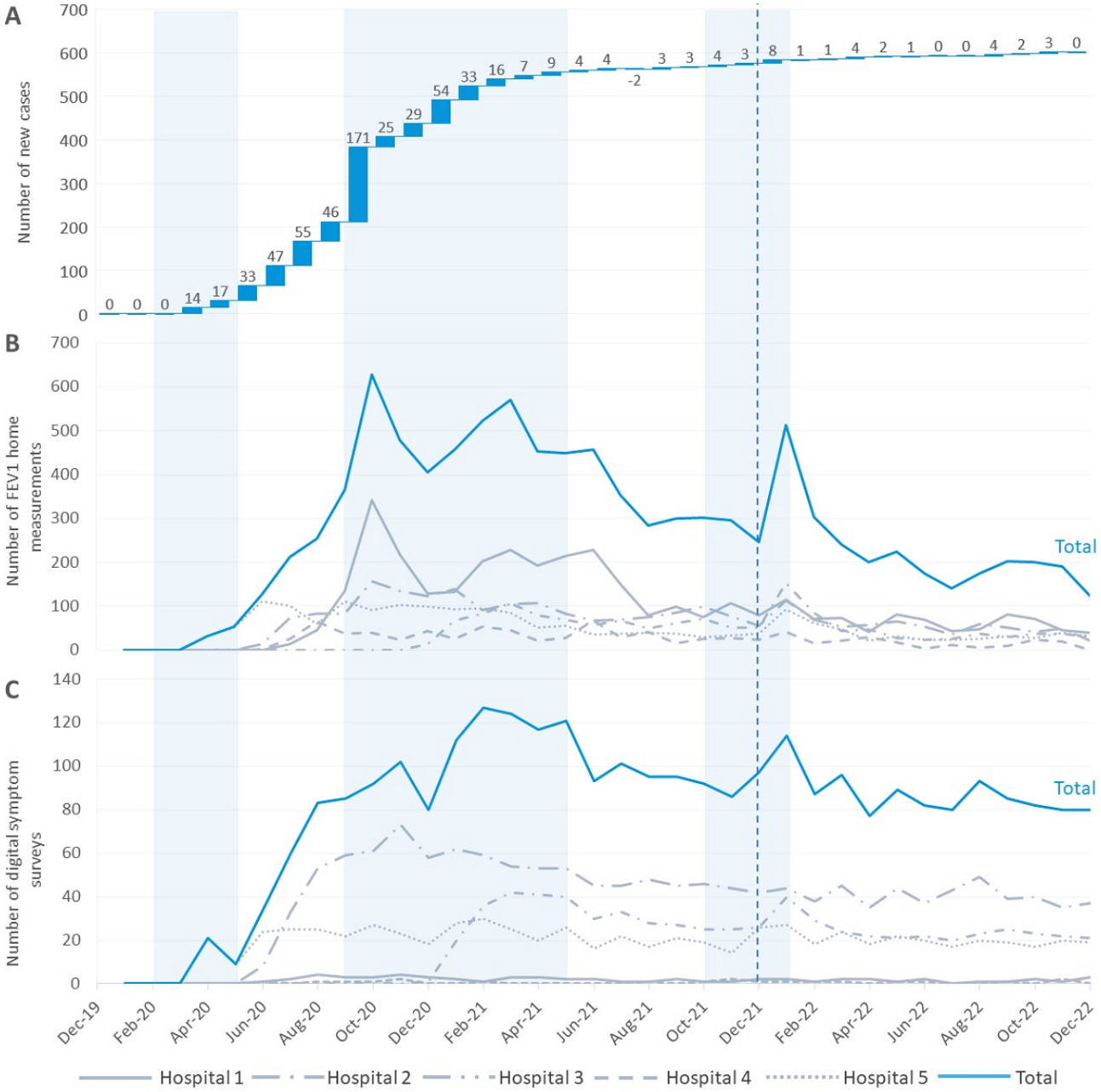

XSL•FO
RenderX

Textbox 1. Selection of supporting quotes from interviews translated from Dutch.

Quote 1

- "At first, we gave everyone a spirometer. For some, it quickly ended up in a box somewhere. Others started using it but couldn't get it to work, so it still ended up in a box somewhere. Quite some patients do actually use it. Some somewhat more diligently at first. I set a reminder for a symptom survey every 14 days and ask patients to also measure their lung function at these moments. Some were really adherent at first, but then couldn't sustain this pattern but take it up again every now and then. And others remain very adherent. For myself, I let go of the idea that I should set up everyone with these devices [...]. Now, I ask patients more directly whether they actually want to use these devices before I give them one." (Interviewee 1; health care professional [HCP]).

Quote 2

- "The measurement method is different than in the hospital. Different support, different mouthpiece, different technique. And that also means that I don't always agree with the outcome for 100%." (Interviewee 2; person with cystic fibrosis [CF]).

Quote 3

- "[Low results] don't really encourage them to keep measuring at home because it is disappointing. We try to motivate patients by explaining that this is a recurring issue, and that it takes time to get used to measuring lung functions at home, and that it will resolve itself [...]. So, we try to make it clear beforehand that there are differences and that the differences will become smaller when they keep practicing [...]. Sometimes this helps, but sometimes it doesn't" (Interviewee 3; HCP).

Quote 4

- "I think that mostly patients themselves really appreciate it [...]. There are different types of patients of course. Some really want to stay in control and want to be able to see how they are doing, and the program is really easy to use for these ends. These patients usually use it well and frequently. There are also patients who still aren't doing so well and who dislike feeling like a patient and having to go the hospital for a full day. They would prefer being at home to work or something. I think that this program can be used very well as a tool to have check-ups at home. I also think, for myself at least, that there is a third group, those people which I have very little control over and don't show up to outpatient visits. I hope this program can be an intermediate solution to stay in touch with them" (Interviewee 4; HCP).

Quote 5

- "[The benefit], right now, is being able to keep track of my stability. Really being able to see whether the trend is going up or down [...]. Just a type of certainty, which also helps to reduce hospital visits. Now, I can just skip a few and then go again after four months instead of much more often [...]. That is a very big advantage for me, also because I live so far away from the hospital" (Interviewee 5; person with CF).

Quote 6

- "We noticed that before we used the program, we felt tense before outpatient visits because of how [my child's] lung function would be. And we actually had no insight in it. Since we use the program, outpatient visits feel so much more relaxed and pleasant, because we know that everything is fine." (Interviewee 6; parent of child with CF).

Quote 7

- Interviewee: "For some children, but that is a minority, it also gives stress. We withdrew a few patients from the program, because the home measurements gave too much stress. Obsessive amounts of measurements and compulsive parents. The benefits just didn't outweigh the burden anymore. Maybe we can restart when [their children] are older, but this can also be an effect."
- Interviewer: "How does this present itself, and where does this stress originate from?"
- Interviewee: "A child blows very variably, because he apparently isn't skilled enough yet and parents really struggle with this. They think: 'well this can't be right.' They might want their child to blow too often, and at a certain moment the child will think: 'let it go, I don't want to anymore.' Of course the child really needs to have a good technique, so then the interaction between [parents and the child] just isn't right and then it's smart to just pause it for a while" (Interviewee 7; HCP).

Quote 8

- "I asked [my son]: what do you think about the device yourself? [...] He answered 'I don't really like it.' Then I asked him: 'And what if it could help you to go to the hospital less?' Then it was absolutely fine" (Interviewee 8; parent of child with CF).

Quote 9

- "How often do I use it? Well, not so often. Once a month, once a quarter. Mostly on moments when I feel like checking myself quickly. And that has two reasons. First, Trikafta [...] immediately led to a drastic decrease of typical CF symptoms like coughing and lots of sputum. As a result, my lung function is not really an issue anymore. So I don't feel the need to measure my lung function every week. Like, 'am I doing well or not?' No, I feel well and I know my lung function is relatively stable. I was ill during last Christmas and during these moments it's great to be able to check my lung function; like does it have an impact or not? But I use it sporadically. On the one hand, because I feel fine, and on the other hand because I sometimes doubt the reliability" (Interviewee 2; person with CF).

Quote 10

XSL•FO
RenderX

- "If I am stable, than I would say I only need to visit the hospital physically every half year. In between, depending on how I am doing, I could have short communications every two to three months [...]. But then you only need to go to the hospital physically every six months [...]. During these visits you can hand in sputum, blood, get a physical examination like blood pressure, that kind of stuff. I imagine that when you are extremely stable that maybe you could even go to the hospital every nine months. And have extra check-ups when you don't trust how it is going, like, having contact by phone. Then you can still always go to the hospital to hand in sputum or something. I think, both for myself and the health system, this is a very good addition" (Interviewee 10; person with CF).

Quote 11

- Interviewer: "How do you feel about [frequent monitoring to predict exacerbations]?"
- Interviewee: "I think [daily measurements] are too much [...] I think it depends on your situation. When you are doing worse than others, then maybe it could give you an advantage, when you can truly detect exacerbations early. But in the case of people who are stable, I don't think that it really has an added value."
- Interviewer: "That's clear. Would it have been different in the period before Trikafta?"
- Interviewee: "Then I would have done it, I think. Yes. If it could predict: 'now you need to pay attention and possibly take action.' Yes, then I would have definitely used it" (Interviewee 5; person with CF).

Quote 12

- "The group of patients for whom I would find it more difficult [to replace outpatient visits with remote monitoring] are those with a certain rejection of their child's condition. Those parents who prefer not to talk about their child's CF, or who would prefer to have as few therapies as possible [...]. The danger is that you lose sight of these parents and patients [...]. The group of parents for whom it would be possible are those who have fully accepted their child's condition. Those who know that you sometimes have to visit the hospital or to be alert during symptoms, and who also stay alert. That is of course our biggest population, and they know the potential consequences for the future of their children when they deny them care. But you really have to know your patients before you can make this assessment together. And you really don't want to miscalculate and find out in hindsight that you have lost grip on your patient" (Interviewee 11; HCP).

Quote 13

- "Eventually, the gain will be in using online monitoring [to reduce regular outpatient visits]. For some patients this will be really suitable. For others, seeing them in the consultation room adds to their therapy adherence and disease knowledge. So, for some this will definitely be possible. However, this will require a certain 'transition-management' for some doctors who have to accept that this is how it is going to be in the future. That our hospital visits are not some sacred towers of strength when we talk about CF management. That really requires a change in mentality" (Interviewee 11; HCP).

Perceived Ease of Use

Overall, usability was well rated. Most users found the RMP easy to use (HCPs: 16/24, 67%; people with CF: 55/72, 76%; *P*=.42), and almost all respondents thought that most users would be able to quickly learn to use the RMP (HCPs: 23/24, 96%; people with CF: 60/72, 83%; *P*=.17). The most reported negative aspect of the RMP's usability were technical errors related to the portable spirometer. In questionnaires, people with CF reported that they experienced technical errors sometimes (14/72, 19%), often (7/72, 10%), or very frequently (2/72, 3%). Moreover, 42% (10/24) of HCPs and 29% (21/72) of people with CF had doubts about the reliability of the home spirometry measurements, and 25% (6/24) of HCPs compared to 74% (53/72) of people with CF thought that people with CF used a good lung function technique at home.

In interviews, both people with CF and HCPs reported that home measurements were generally lower than hospital measurements or more variable. The mouthpiece of the home spirometer was often found difficult to use, especially by children. Some people with CF reported missing the support of a pulmonary function technician at home (Textbox 1, quote 2). Others found the home spirometer easy and quick to use and found that good education and experience helped improve technique and interpretation. HCPs reported that technical and reliability issues had caused some people with CF to lose motivation to use the program (Textbox 1, quote 3). For the HCPs themselves, the reproducibility of measurements, the flow-volume loop, and context (eg, increased symptoms, deviation from baseline, or own assessment of technique) were important to interpret results.

People with CF suggested adding percentage predicted values for home spirometry outcomes, and HCPs suggested better integration of the program into existing electronic health records and better education and communication about the program and new developments.

Perceived Usefulness

Questionnaire results showed that most people with CF and HCPs found the program a good addition to daily CF care (people with CF: 44/72, 61%; HCPs: 21/24, 88%; *P*=.02) or their job (HCPs: 21/24, 88%). Less than half of HCPs felt that the program improved work performance (5/24, 21%), productivity (8/24, 33%), or effectiveness (10/24, 42%). On the other hand, HCPs found that the RMP helped recognize early deterioration of people with CF (16/24, 67%) and gave more insights into people with CF's disease courses (18/24, 75%). In interviews, HCPs reported that most benefits were present at the patient level rather than at a work performance level (Textbox 1, quote 4). However, HCPs especially appreciated being able to remotely monitor acute deterioration, which facilitated remote triage and adjustment and follow-up of treatments.

XSL•FO
RenderX

A total of 49% (35/72) of people with CF (people with CF: 22/54, 41%; parents of people with CF: 13/18, 72%; *P*=.03) believed that they were able to recognize deterioration sooner than without the RMP, 49% (35/72) believed that they needed less in-person outpatient visits when they were doing well, 36% (26/72) felt more in control, and 35% (25/72) felt more motivated for their treatment. From qualitative analyses, the most important benefits for people with CF included improved understanding of their condition, improved control over their condition, better symptom perception, and reductions in unscheduled and routine outpatient visits (Textbox 1, quote 5). Some parents of people with CF mentioned that the program improved communication with their children about their condition. Benefits varied across people with CF.

Psychosocial Effects

According to questionnaire results, 17% (12/72) of people with CF experienced stress due to the RMP. Qualitative analyses revealed positive psychosocial effects such as reassurance about health status and less anxiety before outpatient visits (Textbox 1, quote 6). However, for others, the program caused negative psychosocial effects such as increased stress and worries, confrontations between parents and children who disliked home measurements, frustrations with technical errors or bad outcomes, fears of deterioration, increased disease burden, and confrontation with health identities. HCPs were mostly aware of these effects, which sometimes required intervention (Textbox 1, quote 7). Most interviewed people with CF believed that the benefits outweighed the negative effects, whereas others had not experienced any psychosocial effects.

Intention to Use

In both the questionnaire and interviews, the presence of symptoms was found to be an important incentive to use the RMP regularly. People with CF reported that a future reduction in outpatient visits would be a strong incentive for self-monitoring also during periods of well-being (Textbox 1, quote 8). Other incentives included agreements with and encouragement of HCPs regarding measurement frequency, creating a habit or setting reminders, living far away from the hospital, intrinsic motivation and a perceived necessity for self-monitoring, the fact that measurements are easy and quick to perform, and individually experienced benefits. Ultimately, people with CF wanted an adequate balance between monitoring burden and the benefits of use.

Stable periods were accompanied by little motivation for frequent self-monitoring. For many, these stable periods were more prevalent after initiation of ETI. Other disincentives for regular self-monitoring included having a good symptom perception, unclear implementation and goals of remote monitoring, the existence of other communication media (eg, direct phone lines or email), little intrinsic motivation, other health priorities, technical errors or doubts about reliability of outcomes, remote monitoring being mostly supplemental to regular care without reductions in therapy burden elsewhere, and the aforementioned negative psychosocial effects (Textbox 1, quote 9).

Future Use

A total of 79% (19/24) of HCPs and 75% (54/72) of people with CF wanted more care remotely in the future rather than physically when people with CF were feeling well (ie, hybrid care). Although some people with CF wanted to keep using the RMP as supplemental to routine care, many people with CF and HCPs also supported hybrid care in interviews, especially in adult CF care. The most important driver was improvement in people with CF's condition due to new modulator drugs. Hybrid care was thought to help people with CF reclaim agency over their lives and reduce the burden of hospital visits. However, there was a consensus among interviewed participants that in-person care could not be replaced fully by remote care and that there was a strong need for individualized approaches (Textbox 1, quote 10). In pediatrics, it was mentioned that evaluating the physical and psychosocial development of children was also important during outpatient follow-up and that this might not be fully replaceable by digital care.

Table 2 shows additional remote monitoring functions identified as useful for CF care in questionnaires. Additional remote microbiology assessments was the only repeated suggestion for future additions in interviews. One pulmonologist mentioned that, when additional measurements are required to assess patients remotely, it would be better to see the patient in person. Interviewees showed little support for pulmonary exacerbation prediction models requiring high self-monitoring frequencies since the introduction of ETI, but some acknowledged that they would have used these models before they started ETI (Textbox 1, quote 11).

XSL•FO
RenderX

Table 2. Number of respondents who found the following options useful for a remote monitoring program for cystic fibrosis (CF) care[a].

	HCPs[b] (n=24), n (%)	People with CF (n=72), n (%)	*P* value	CF group, n (%)		*P* value
				People with CF (n=54)	Parents (n=18)	
Monitoring of lung function	22 (92)	66 (92)	>.99	51 (94)	15 (83)	.16
Monitoring of pulmonary symptoms	17 (71)	26 (36)	.004[c]	19 (35)	7 (39)	.78
A function that alerts people with CF when they are not doing well	15 (63)	20 (28)	.003[c]	14 (26)	6 (33)	.56
A function that alerts HCPs when people with CF are not doing well	13 (54)	32 (44)	.48	21 (39)	11 (61)	.11
Monitoring gastrointestinal symptoms	6 (25)	15 (21)	.78	10 (19)	5 (28)	.50
Tracking dietary requirements from the dietician or pancreatic enzyme use	6 (25)	7 (10)	.08	6 (11)	1 (6)	.67
Tracking physiotherapy exercises or activity (eg, number of steps)	5 (21)	9 (13)	.33	9 (17)	0 (0)	.10
Tracking medication use or side effects	13 (54)	24 (33)	.09	17 (31)	7 (39)	.58
Tracking how people with CF feel with regard to their CF (eg, fear or confidence)	11 (46)	13 (18)	.01[c]	10 (19)	3 (17)	>.99
As little functions as possible	2 (8)	4 (6)	.64	3 (6)	1 (6)	>.99

[a]Proportions of respondents who selected the functions were compared to each other using the Fisher exact test (health care professionals vs people with CF and people with CF vs parents of people with CF).

[b]HCP: health care professional.

[c]Significant difference.

Prerequisites

Only 54% (13/24) of HCPs reported that it was easy to predict beforehand which people with CF would benefit from these programs. HCPs identified several requirements for remote monitoring in interviews: regular home measurements to maintain baseline values, a reliable lung function technique, motivation of people with CF, adequate technology savviness or (digital) health literacy, and adequate psychosocial coping strategies. For hybrid care, requirements for patients were found to be stricter and included a stable condition, adequate symptom recognition, alertness and responsibility during deterioration, and adequate therapy adherence. Therefore, an important prerequisite to be able to assess eligibility was being well acquainted with patients and maintaining this relationship remotely (Textbox 1, quote 12).

To optimize remote monitoring, requirements for HCPs included a more coaching role, adopting shared decision-making in deciding remote monitoring goals and methods, a change in mindset, including remote monitoring in regular follow-up conversations, clear department policies with a division of tasks and adequate safety nets, better education of (new) employees on remote monitoring, and improved communication between HCPs and developers. In most centers, the availability of motivated specialist nurses was essential as they performed most remote monitoring tasks (Textbox 1, quote 13).

Discussion

Principal Findings

This study identified experiences, future perspectives, and use behavior of remote monitoring in CF care. Our results show broad support among both HCPs and people with CF for remote monitoring. Benefits ranged from improved symptom perception to reduced routine outpatient care. Lung function home measurements were most appreciated. Pitfalls included negative psychosocial effects and the potential to lose sight of patients. Multiple incentives or disincentives to use the program were identified, such as the severity of physical symptoms. Future perspectives were centered on hybrid care models, personalized remote monitoring strategies, and balancing experienced benefits and monitoring burden.

A recent study identified similar patient benefits of remote monitoring, such as improved control and symptom recognition [24]. In previous work, we also found these benefits in pediatric asthma [12]. Unfortunately, these benefits have received little attention in the literature, potentially because they have not been systematically identified, lack good objective end points, or are hard to express in financial savings. Nevertheless, these benefits can be supportive to individual patients or parents. Future research should distinguish the varieties of individual patient benefits as well as identify objective end points. One possible

XSL•FO
RenderX

example of end points could be personalized electronic patient-reported outcome measures [25].

The literature suggests that between a third and a half of people with CF are eager to replace regular care with remote care [11,26]. In our study, >75% of HCPs and people with CF wanted more remote care and less in-person care during periods of well-being. Presently, a minimum of 4 outpatient visits annually is a quality criterion for CF outpatient care in the Netherlands and according to the European CF Society best practice guideline [27]. Most interviewees considered a minimum of 2 annual outpatient visits feasible and desirable when supplemented with remote monitoring. These findings are in line with those of Hendra et al [28]. Our findings may be explained by the improved physical condition of our respondents as 85% used ETI and by the overrepresentation of adults in our study. Parents and HCPs in pediatrics also encouraged hybrid care models but were more reluctant to reduce follow-up visits because of safety and the specialty's broader developmental focus. Nevertheless, a reduction of 50% in regular outpatient visits with remote monitoring has already been proven safe and efficient for pediatric asthma [13,14].

An important benefit not previously identified was the ability of people with CF to objectify symptoms at home, which helped guide HCPs and people with CF through periods of deterioration and recovery. Although the reliability of lung function outcomes was frequently questioned, especially by HCPs, adequate training, education, and regular practice could mostly overcome this issue. Consequently, remote monitoring has potential to reduce unscheduled health care consumption (eg, outpatient visits and emergency department visits), but this needs to be studied further.

Up to 20% of questionnaire respondents experienced increased stress from using the program. For some, the psychological burden of home monitoring required intervention. These results are more discouraging than previously reported acceptance rates [5,11,24,29]. This might be due to selection bias or a lack of good psychosocial screening methods. These findings emphasize that HCPs need to make time for education, regular evaluation, and support. This also requires motivation and new digital health skills for HCPs and especially for CF nurses, who played a crucial role in the remote monitoring of people with CF. Moreover, it illustrates that not everyone might benefit from remote monitoring. Patients should be responsible and motivated and have sufficient (digital) health literacy and adequate coping strategies. Importantly, these interventions should not be forced on patients but should be evaluated on a patient-by-patient basis through open communication and shared decision-making. This requires an adequate relationship between patients and HCPs not just at the initiation of remote monitoring but also throughout the complete patient journey, where a long-term relationship is required to evaluate the success of remote monitoring and make adaptions to the individual strategy accordingly. Maintaining this relationship will pose a new challenge in the era of digital health as the nexus between patients and HCPs will be increasingly distant in both time and space.

Up to 40% of respondents only used the program during periods of increased symptoms. People with CF likely prioritize reducing their disease burden as they improve physically with new modulator drugs [8,30,31]. Therefore, people with CF might be disincentivized to regularly self-monitor when it is mostly supplemental to regular care. In addition, similarly to the findings by Simpson et al [32], people with CF identified less monitoring functions as useful than HCPs. These findings were reinforced by the objective observation that home measurement rates declined after COVID-19 lockdowns and after the introduction of ETI. These disincentives offer an explanation for the low monitoring adherence in many studies [5,11,24,29,33]. It also corresponds to the negative change in attitude of people with CF toward frequent monitoring to predict deterioration after starting ETI. Nonetheless, many people with CF have no access to modulator therapies or have disappointing responses, and they might still benefit from future prediction models. Generally, a balance must be struck between individual benefits and the burden of remote monitoring.

Strengths and Limitations

Our mixed methods design enabled the integration of multiple sources of data from a large group of people with CF. To prevent both interpretation and confirmation bias, 2 experienced qualitative researchers (YE and MH) with different backgrounds (psychology [YE] and nursing sciences [MH]) were involved in the collection and analysis of qualitative data. These researchers were not involved in the RMP. To our knowledge this is the most comprehensive appraisal of remote monitoring in CF within long-term, multicenter care. Recently, there have been increasing calls for studies such as this one [5]. Moreover, the overlap in findings with those of other studies in pediatric and adult asthma suggests that our results translate well to other pulmonary diseases [12,14].

The COVID-19 pandemic was the direct cause for the initiation of the RMP. Hence, the time of implementation was turbulent and not representative of regular CF care before the pandemic. This should be considered in interpreting the results. For example, few HCPs found that the RMP improved their work performance and productivity, but this might also be explained by a preoccupation with COVID-19–related care and managing the overburdened health system. Consequently, the full potential of the RMP might not have been reached as there was little time to create clear remote monitoring strategies and policies even though they were found to be important prerequisites for successful implementation. We initially aimed to include more medical doctors in interviews because we aimed for equal representation of different HCPs involved in the remote monitoring of people with CF. However, data saturation was reached after 7 interviews. Possibly, CF nurses were more likely to respond to interview invitations as their role was crucial in the implementation of remote monitoring in routine care. Finally, because we recruited participants through informed e-consent to our program, we were unable to reach people with CF who stopped using the program. HCPs were added to provide some insights into this population.

Future Directions

The results of this study can be used to synthesize 4 future directions of remote monitoring in CF care, which are summarized in Textbox 2 (sensible strategies for remote

XSL•FO
RenderX

monitoring). These strategies can be used simultaneously and alternately depending on active needs, capacities, and available resources of people with CF, HCPs, and the overarching health system. In more practical terms, this means that the different sensible strategies should be implemented based on what resources are available to support remote monitoring within the local context and what local priorities for digital support of the health system are in place. At a patient level, the sensible strategies should translate to an iterative process of shared decision-making on remote monitoring goals, methods, and evaluation and adaption that evolves alongside the changing and dynamic needs of patients. This requires clear remote monitoring policies and tasks and should foster the continuing relationship between patients and HCPs.

Textbox 2. Sensible strategies for remote monitoring in chronic respiratory diseases.

- Patient-driven strategies aim to support individual patients. These strategies include improving symptom perception, self-management, and therapy adherence or tracking disease trajectories. Importantly, patients themselves have most control over what, why, and when they monitor. Patient-driven strategies require the least effort from health care professionals (HCPs) and are mostly supplemental to regular care, but they require patients to be motivated and skilled. In addition, these strategies can be used to gain experience with remote monitoring in a small group of users or to provide niche monitoring functions for a minority of users. A challenge for these strategies will be reimbursement. Future research should focus on identifying the variety of possible goals for patient-driven strategies as well as how these strategies can be evaluated and valued.
- Symptom-driven strategies are used to guide patients and HCPs through periods of acute and subacute deterioration. They encompass remote symptom objectification, triage, and the initiation and follow-up of treatments. Symptom-driven strategies are mostly supplemental to regular care and require little system-wide changes but can provide immediate remote point-of-care benefits for both patients and HCPs. Symptom-driven strategies require a minimal monitoring frequency to track baseline values and maintain skills during stable periods. Research should focus on reducing irregular health care consumption, such as emergency department visits and inpatient stays, and on cost-effectiveness.
- Care-driven strategies are used to replace regular care through remote monitoring, whereas symptom-driven strategies have more potential for irregular care. Hybrid care models can potentially improve patient emancipation; reduce health care costs and workload; and, therefore, strengthen health systems. However, care-driven strategies require significant effort from both patients and HCPs, system-wide changes, well-designed policies, and adequate safety nets. Consequently, there is a need for more research into the requirements, cost-effectiveness, safety, and implementation of these strategies.
- Prediction-driven strategies are not yet practice based but have received significant attention in the literature. These strategies distinguish themselves from symptom-driven strategies through frequent, multiparametric monitoring and algorithms designed to predict exacerbations or other clinical outcomes. Prediction-driven strategies might be suitable for patients with an unstable condition and frequent acute and subacute deterioration, but there is a need for the validation of these models as well as of their clinical application and efficacy. Importantly, these strategies need to strike a balance between benefits and burden.

The sensible strategies provide a first goal-based frame to conceptualize remote monitoring in clinical practice while recognizing the need for flexibility in applying these principles to local contexts and individual patients. Consequently, the sensible strategies can also be applied in resource-scarce settings where conventional communication media (eg, telephone or SMS text messages) are more accessible than digital programs. As at least patient-driven and care-driven approaches have been identified in previous work in pediatric asthma, it is probable that the sensible strategies can also be used for other CRDs [12,14].

We recognize that the sensible strategies might not be complete, but it is the first goal-based classification of remote monitoring strategies that originated within routine, long-term CRD care. These strategies have the potential to be a first step in resolving the disconnect between current research efforts and clinical applicability as the strategies allow researchers to align their efforts with a clinically relevant strategy that is accompanied by clinical considerations. Moreover, the strategies will empower patients, clinicians, and policy makers to make informed decisions about the use of remote monitoring within their own contexts. Nevertheless, future efforts are needed to make the sensible strategies more actionable by expanding them, validating them, and defining potential end points while acknowledging the inequities in available resources for digital health worldwide.

Conclusions

Remote monitoring can offer a range of benefits for HCPs and people with CF at both the individual and collective levels. It is essential to integrate remote monitoring strategies into care models according to local capacities and needs to maximize benefits while ensuring feasibility. Many previous studies have focused on predicting deterioration. These interventions are complex and expensive and require significant long-term effort from patients. Therefore, they will not be beneficial for everyone or always feasible in daily practice. The sensible strategies for remote monitoring in chronic respiratory diseases in Textbox 2 can help facilitate the integration of remote monitoring into the care models of CF and other CRDs as it aims to help clinicians, researchers, and policy makers align remote monitoring with local demands and capacities. Future research should pay more attention to the other sensible strategies as they better correspond to value-based health care and could provide immediate support for patients, HCPs, and the overarching health systems.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the Dutch Cystic Fibrosis Foundation for donating the portable spirometers and for providing valuable insights and guidance in the development of the questionnaires. They would like to thank the Dutch Cystic Fibrosis Foundation and the Steering Group for supplying the Dutch Cystic Fibrosis Registry data used in this study. They would also like to thank all people with cystic fibrosis and health care professionals who participated in the interviews. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Authors MCO, LSvdW, JR, and PJFMM are involved in the development of the software used in this study. The software is property of the Radboud University Medical Center and, hence, does not generate any profits for the authors. No generative artificial intelligence was used at any stage of this study.

Data Availability

Anonymized data from this study will be made available upon reasonable request.

Authors' Contributions

All authors were involved in the conceptualization of this study and recruitment of participants. MCO, YE, MH, JR, and PJFMM were responsible for the study design. MCO managed the project administration, visualization of results, writing of the manuscript, and quantitative analyses supervised by PJFMM and JR. MCO and YE conducted the formal analysis of the qualitative results supervised by MH, JR, and PJFMM. All authors were involved in reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[DOCX File , 35 KB - jmir_v26i1e54942_app1.docx]

Multimedia Appendix 2
Details of the remote monitoring program.
[DOCX File , 1329 KB - jmir_v26i1e54942_app2.docx]

Multimedia Appendix 3
Questionnaires including answers (translated).
[DOCX File , 28 KB - jmir_v26i1e54942_app3.docx]

Multimedia Appendix 4
Interview guides (translated into English).
[DOCX File , 25 KB - jmir_v26i1e54942_app4.docx]

Multimedia Appendix 5
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 416 KB - jmir_v26i1e54942_app5.pdf]

Multimedia Appendix 6
Themes and subthemes.
[DOCX File , 16 KB - jmir_v26i1e54942_app6.docx]

References

1. GBD 2019 Chronic Respiratory Diseases Collaborators. Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019. EClinicalMedicine 2023 May;59:101936 [FREE Full text] [doi: 10.1016/j.eclinm.2023.101936] [Medline: 37229504]
2. Dunn J, Coravos A, Fanarjian M, Ginsburg GS, Steinhubl SR. Remote digital health technologies for improving the care of people with respiratory disorders. Lancet Digit Health 2024 Apr;6(4):e291-e298 [FREE Full text] [doi: 10.1016/S2589-7500(23)00248-0] [Medline: 38402128]

XSL•FO
RenderX

3. 2023 Global Initiative for Asthma main report. Global Initiative for Asthma. 2023 Jul 10. URL: https://ginasthma.org/2023-gina-main-report/#:~:text=The%202023%20update%20of%20the,on%20the%20GINA%20Science%20Committee [accessed 2023-11-03]
4. Global strategy for prevention, diagnosis and management of COPD: 2023 report. Global Initiative for Chronic Obstructive Lung Disease. 2023. URL: https://goldcopd.org/2023-gold-report-2/ [accessed 2023-11-03]
5. Vagg T, Shanthikumar S, Ibrahim H, O'Regan P, Chapman WW, Kirwan L, et al. Telehealth in cystic fibrosis. A systematic review incorporating a novel scoring system and expert weighting to identify a 'top 10 manuscripts' to inform future best practices implementation. J Cyst Fibros 2023 Jul;22(4):598-606 [FREE Full text] [doi: 10.1016/j.jcf.2023.05.012] [Medline: 37230808]
6. Ratjen F, Bell SC, Rowe SM, Goss CH, Quittner AL, Bush A. Cystic fibrosis. Nat Rev Dis Primers 2015 May 14;1:15010 [FREE Full text] [doi: 10.1038/nrdp.2015.10] [Medline: 27189798]
7. Keogh RH, Tanner K, Simmonds NJ, Bilton D. The changing demography of the cystic fibrosis population: forecasting future numbers of adults in the UK. Sci Rep 2020 Jun 30;10(1):10660 [FREE Full text] [doi: 10.1038/s41598-020-67353-3] [Medline: 32606329]
8. Prickett MH, Flume PA, Sabadosa KA, Tran QT, Marshall BC. Telehealth and CFTR modulators: accelerating innovative models of cystic fibrosis care. J Cyst Fibros 2023 Jan;22(1):9-16 [FREE Full text] [doi: 10.1016/j.jcf.2022.07.002] [Medline: 35879227]
9. Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, et al. The future of cystic fibrosis care: a global perspective. Lancet Respir Med 2020 Jan;8(1):65-124 [FREE Full text] [doi: 10.1016/S2213-2600(19)30337-6] [Medline: 31570318]
10. Wong CH, Smith S, Kansra S. Digital technology for early identification of exacerbations in people with cystic fibrosis. Cochrane Database Syst Rev 2023 Apr 14;4(4):CD014606. [doi: 10.1002/14651858.CD014606.pub2] [Medline: 37057835]
11. Edmondson C, Westrupp N, Seddon P, Olden C, Wallis C, Dawson C, et al. The feasibility of home monitoring of young people with cystic fibrosis: results from CLIMB-CF. J Cyst Fibros 2022 Jan;21(1):70-77 [FREE Full text] [doi: 10.1016/j.jcf.2021.09.018] [Medline: 34635459]
12. Oppelaar MC, van den Wijngaart LS, Merkus PJ, Croonen EA, Hugen CA, Brouwer ML, et al. It is not just the FEV1 that matters, but the personal goals we reach along the way: qualitative, multicenter, prospective, observational study. J Med Internet Res 2021 Oct 20;23(10):e29218 [FREE Full text] [doi: 10.2196/29218] [Medline: 34668868]
13. van den Wijngaart LS, Kievit W, Roukema J, Boehmer AL, Brouwer ML, Hugen CA, et al. Online asthma management for children is cost-effective. Eur Respir J 2017 Oct;50(4):1701413 [FREE Full text] [doi: 10.1183/13993003.01413-2017] [Medline: 28982768]
14. van den Wijngaart LS, Roukema J, Boehmer AL, Brouwer ML, Hugen CA, Niers LE, et al. A virtual asthma clinic for children: fewer routine outpatient visits, same asthma control. Eur Respir J 2017 Oct;50(4):1700471 [FREE Full text] [doi: 10.1183/13993003.00471-2017] [Medline: 28982775]
15. Hong Q, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, et al. Mixed Methods Appraisal Tool (MMAT) version 2018: user guide. McGill University. 2018. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf [accessed 2023-11-03]
16. van den Wijngaart LS, Geense WW, Boehmer AL, Brouwer ML, Hugen CA, van Ewijk BE, et al. Barriers and facilitators when implementing web-based disease monitoring and management as a substitution for regular outpatient care in pediatric asthma: qualitative survey study. J Med Internet Res 2018 Oct 30;20(10):e284 [FREE Full text] [doi: 10.2196/jmir.9245] [Medline: 30377147]
17. Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash ML, Ramsey BW, et al. Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. N Engl J Med 1994 Sep 08;331(10):637-642. [doi: 10.1056/NEJM199409083311003] [Medline: 7503821]
18. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
19. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
20. Sauro J. Measuring usability with the System Usability Scale (SUS). MeasuringU. 2011 Feb 3. URL: https://measuringu.com/sus/ [accessed 2022-07-07]
21. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. London, UK: CRC Press; 1996.
22. Flottorp SA, Oxman AD, Krause J, Musila NR, Wensing M, Godycki-Cwirko M, et al. A checklist for identifying determinants of practice: a systematic review and synthesis of frameworks and taxonomies of factors that prevent or enable improvements in healthcare professional practice. Implement Sci 2013 Mar 23;8:35 [FREE Full text] [doi: 10.1186/1748-5908-8-35] [Medline: 23522377]
23. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]

XSL•FO
RenderX

24. Nash EF, Choyce J, Carrolan V, Justice E, Shaw KL, Sitch A, et al. A prospective randomised controlled mixed-methods pilot study of home monitoring in adults with cystic fibrosis. Ther Adv Respir Dis 2022;16:17534666211070133 [FREE Full text] [doi: 10.1177/17534666211070133] [Medline: 35274585]
25. Muilwijk D, van Paridon TJ, van der Heijden DC, Faber-Bisschop BM, Zomer-van Ommen DD, Heijerman HG, et al. Development and validation of a novel personalized electronic patient-reported outcome measure to assess quality of life (Q-LIFE): a prospective observational study in people with cystic fibrosis. EClinicalMedicine 2023 Aug;62:102116 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102116] [Medline: 37554124]
26. Solomon GM, Bailey J, Lawlor J, Scalia P, Sawicki GS, Dowd C, et al. Patient and family experience of telehealth care delivery as part of the CF chronic care model early in the COVID-19 pandemic. J Cyst Fibros 2021 Dec;20 Suppl 3:41-46 [FREE Full text] [doi: 10.1016/j.jcf.2021.09.005] [Medline: 34930542]
27. Castellani C, Duff AJ, Bell SC, Heijerman HG, Munck A, Ratjen F, et al. ECFS best practice guidelines: the 2018 revision. J Cyst Fibros 2018 Mar;17(2):153-178 [FREE Full text] [doi: 10.1016/j.jcf.2018.02.006] [Medline: 29506920]
28. Hendra K, Neemuchwala F, Chan M, Ly NP, Gibb ER. Patient and provider experience with cystic fibrosis telemedicine clinic. Front Pediatr 2021 Nov 24;9:784692 [FREE Full text] [doi: 10.3389/fped.2021.784692] [Medline: 34900879]
29. Lechtzin N, Mayer-Hamblett N, West NE, Allgood S, Wilhelm E, Khan U, et al. Home monitoring of patients with cystic fibrosis to identify and treat acute pulmonary exacerbations. eICE study results. Am J Respir Crit Care Med 2017 Nov 01;196(9):1144-1151 [FREE Full text] [doi: 10.1164/rccm.201610-2172OC] [Medline: 28608719]
30. Calthorpe RJ, Smith SJ, Rowbotham NJ, Leighton PA, Davies G, Daniels T, et al. What effective ways of motivation, support and technologies help people with cystic fibrosis improve and sustain adherence to treatment? BMJ Open Respir Res 2020 Aug;7(1):e000601 [FREE Full text] [doi: 10.1136/bmjresp-2020-000601] [Medline: 32816834]
31. Sawicki GS, Sellers DE, Robinson WM. High treatment burden in adults with cystic fibrosis: challenges to disease self-management. J Cyst Fibros 2009 Mar;8(2):91-96 [FREE Full text] [doi: 10.1016/j.jcf.2008.09.007] [Medline: 18952504]
32. Simpson AJ, Honkoop PJ, Kennington E, Snoeck-Stroband JB, Smith I, East J, et al. Perspectives of patients and healthcare professionals on mHealth for asthma self-management. Eur Respir J 2017 May;49(5):1601966 [FREE Full text] [doi: 10.1183/13993003.01966-2016] [Medline: 28461291]
33. Beaufils F, Enaud R, Gallode F, Boucher G, Macey J, Berger P, et al. Adherence, reliability, and variability of home spirometry telemonitoring in cystic fibrosis. Front Pediatr 2023;11:1111088 [FREE Full text] [doi: 10.3389/fped.2023.1111088] [Medline: 36911035]

Abbreviations

CF: cystic fibrosis
CFTR: cystic fibrosis transmembrane conductance regulator
COREQ: Consolidated Criteria for Reporting Qualitative Research
CRD: chronic respiratory disease
ETI: elexacaftor/tezacaftor/ivacaftor
HCP: health care professional
RMP: remote monitoring program
TAM: technology acceptance model

Edited by S Ma, T Leung; submitted 28.11.23; peer-reviewed by T Sirari, H Oh; comments to author 18.03.24; revised version received 09.04.24; accepted 27.05.24; published 06.08.24.

Please cite as:
Oppelaar MC, Emond Y, Bannier MAGE, Reijers MHE, van der Vaart H, van der Meer R, Altenburg J, Conemans L, Rottier BL, Nuijsink M, van den Wijngaart LS, Merkus PJFM, Heinen M, Roukema J
Potential, Pitfalls, and Future Directions for Remote Monitoring of Chronic Respiratory Diseases: Multicenter Mixed Methods Study in Routine Cystic Fibrosis Care
J Med Internet Res 2024;26:e54942
URL: https://www.jmir.org/2024/1/e54942
doi:10.2196/54942
PMID:39106098

©Martinus C Oppelaar, Yvette Emond, Michiel A G E Bannier, Monique H E Reijers, Hester van der Vaart, Renske van der Meer, Josje Altenburg, Lennart Conemans, Bart L Rottier, Marianne Nuijsink, Lara S van den Wijngaart, Peter J F M Merkus, Maud Heinen, Jolt Roukema. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient and Provider Experiences With Compassionate Care in Virtual Physiatry: Qualitative Study

Marina B Wasilewski[1], PhD; Abirami Vijayakumar[1], MSc; Zara Szigeti[1], MSc; Amanda Mayo[1], MD, MHSc; Laura Desveaux[2], PhD; James Shaw[3], PhD; Sander L Hitzig[1], PhD; Robert Simpson[3], MBChB, PhD

[1]Sunnybrook Research Institute, St. John's Rehab, Toronto, ON, Canada
[2]Institute for Better Health, Mississauga, ON, Canada
[3]University of Toronto, Toronto, ON, Canada

Corresponding Author:
Marina B Wasilewski, PhD
Sunnybrook Research Institute
St. John's Rehab
285 Cummer Avenue, Room B114
Toronto, ON, M2M 2G1
Canada
Phone: 1 416 226 6780 ext 57957
Email: marina.wasilewski@sunnybrook.ca

Abstract

Background: Telemedicine in the realm of rehabilitation includes the remote delivery of rehabilitation services using communication technologies (eg, telephone, emails, and video). The widespread application of virtual care grants a suitable time to explore the intersection of compassion and telemedicine, especially due to the impact of COVID-19 and how it greatly influenced the delivery of health care universally.

Objective: The purpose of this study was to explore how compassionate care is understood and experienced by physiatrists and patients engaged in telemedicine.

Methods: We used a qualitative descriptive approach to conduct interviews with patients and physiatrists between June 2021 and March 2022. Patients were recruited across Canada from social media and from a single hospital network in Toronto, Ontario. Physiatrists were recruited across Canada through social media and the Canadian Association for Physical Medicine and Rehabilitation (CAPM&R) email listserve. Interviews were recorded and transcribed. Data were analyzed thematically.

Results: A total of 19 participants were interviewed—8 physiatrists and 11 patients. Two themes capturing physiatrists' and patients' experiences with delivering and receiving compassionate care, especially in the context of virtual care were identified: (1) compassionate care is inherently rooted in health care providers' inner intentions and are, therefore, expressed as caring behaviors and (2) virtual elements impact the delivery and receipt of compassionate care.

Conclusions: Compassionate care stemmed from physiatrists' caring attitudes which then manifest as caring behaviors. In turn, these caring attitudes and behaviors enable individualized care and the establishment of a safe space for patients. Moreover, the virtual care modality both positively and negatively influenced how compassion is enacted by physiatrists and received by patients. Notably, there was large ambiguity around the norms and etiquette surrounding virtual care. Nonetheless, the flexibility and person-centeredness of virtual care cause it to be useful in health care settings.

(J Med Internet Res 2024;26:e51878) doi:10.2196/51878

KEYWORDS

compassionate care; rehabilitation; physiatry; telemedicine; burnout; care; qualitative study; patient experience; compassion; compassionate; patient centered; virtual care; communication; health care delivery; patient engagement; physiatrist; physiatrists; Canada; social media; physical medicine; technology; communication technology; ICT; experience

XSL•FO
RenderX

Introduction

The COVID-19 pandemic has profoundly impacted health care delivery around the world. Like many health disciplines [1,2], physical medicine and rehabilitation (PM&R) (also known as "physiatry" or "rehabilitation medicine") pivoted to "virtual care"—or "telemedicine"—to mitigate the spread of COVID-19 and to ensure that patients had continued access to vital rehabilitation services throughout the pandemic [3,4]. "Telemedicine" in the context of rehabilitation involves the remote provision of rehabilitation services using information and communication technologies either synchronously (eg, phone calls and videoconferences) or asynchronously (eg, email) to improve patient health [5,6]. Telemedicine models enabled physiatrists (PM&R physicians) to continue providing care to patients, to promote functional recovery and improve quality of life [7]. An abundance of evidence across typical PM&R patient populations (eg, stroke, chronic musculoskeletal conditions, and chronic obstructive pulmonary disorder) has demonstrated telemedicine to be as effective as in-person care at improving several physical (eg, motor function) and social (eg, quality of life) health outcomes [8-12]. Despite these clinical and practical advantages, the implementation of telemedicine was limited prior to the pandemic [13,14], when it was driven forward by necessity at a pace that left little room to optimize its delivery.

Telemedicine introduces the potential to disrupt the relational nature of the traditional doctor-patient relationship, including complex, relational processes such as empathic connection, rapport building, and compassionate responding, which form the basis of compassionate care [15]. Compassion itself is a multifaceted concept that can be broadly defined as the emotions that arise when noticing someone suffer, and feeling motivated to help reduce such suffering through relational care and action [16]. Viewed widely as a core aspect of high-quality health care, compassionate patient care entails human-to-human "contact," personal interaction, and bidirectional communication [17]. Many of these implicit processes that take place during in-person encounters may not be the same in virtual care contexts. For example, health care workers have expressed that many long hours of providing care virtually leads to exhaustion, due to a heightened need to concentrate on the screen and overcompensate for the challenges associated with processing nonverbal cues (eg, body language, facial expressions) [18]. This potentially puts providers at higher risk for burnout and compassion fatigue [18], especially concerning given that physiatrists are the third most burnt out medical specialty in North America [19].

The widespread implementation of virtual care presents an opportune time to explore the emerging intersection of compassion and telemedicine. A recent scoping review on the conceptualization, use, and outcomes associated with compassion in PM&R found that no studies had explored the concept of compassion in a virtual care setting [20]. What compassionate care means to health care providers and patients in PM&R settings and how it is conveyed and experienced virtually represents a key knowledge gap. The aim of this study was to explore how compassionate care is understood and experienced by physiatrists and patients engaged in telemedicine. Although there are a range of theoretical approaches to answer research questions, many designs may not be optimal for studies that do not necessitate a highly theoretical framework and instead focus closely on capturing and understanding participants' experiences [21]. Hence, a qualitative descriptive approach may be used to provide experiences and perceptions of study participants [22] to better understand the conceptualization of compassionate care among patients and health care providers.

Methods

Research Design

We used a qualitative descriptive approach, which entails a concise and descriptively rich analysis that remains true to participants' own words. Thus, it produces a data-near report that is representative of participants' views, making it meaningful for key stakeholders and relevant for justifying actionable change [22,23].

Participants

Participant Profile

Eight Physiatrists and 11 patients across Canada were recruited for this study.

Inclusion and Exclusion Criteria

Participants were eligible to participate if they were English speaking adults and judged able to provide informed consent. Patient participants had to have received care virtually, at least once, from a physiatrist in the 90 days preceding recruitment. Physiatrist had to have provided virtual care at least once in the previous 90 days. Virtual care included video, telephone, and email consultation or follow-up visits (or both).

Recruitment Procedure

Between June 2021 and March 2022, we recruited physiatrists across Canada via social media and the Canadian Association for Physical Medicine and Rehabilitation (CAPM&R) email listserve. Patients were recruited across Canada from social media and [REHABILITATION HOSPITAL—DE-IDENTIFIED] outpatient clinic via referral by a member of their circle of care.

Ethical Considerations

This study was approved by the Research Ethics Board at Sunnybrook Hospital (ID: 4960). Informed consent was obtained prior to data collection.

Data Collection

Once informed consent was collected for participants, one-on-one interviews were conducted by two trained qualitative researchers (AV and ZS) via telephone or Zoom (Zoom Video Communications) between June 2021 and March 2022. Interview questions were focused on participants' experiences with providing and receiving virtual care and the ways that compassion featured in these interactions (see Multimedia Appendices 1 and 2 for interview questions). Data were collected until saturation of ideas was achieved (ie, no new

XSL•FO
RenderX

information emerged from subsequent interviews) [24]. The participants had no prior relationship with the interviewers and understood that the study goals were to explore experiences with virtual care. Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Interviews ranged from 25 to 75 minutes, were audio-recorded, and transcribed verbatim. All identifying information was removed from the transcripts, which were uploaded to NVivo [25] for organization and analysis. Sociodemographic information was collected from both patients and physiatrists—for the latter we also collected professional practice information (eg, years of practice, practice setting).

Data Analysis

We used an inductive thematic approach following the steps outlined by Braun and Clarke (2006), whereby data were deconstructed into isolated fragments, followed by reconstruction into overarching themes which describe the higher-level messaging in the data [26]. Two independent researchers (AV and ZS) completed the coding process and 2 additional researchers (RS and MBW) participated in the thematic analysis.

Rigor

Analytic rigor was enhanced by triangulating between multiple individuals throughout analysis, having regular team meetings, and practicing reflexivity (discussing and journaling the study team's own biases and experiences that may influence data interpretation) [24]. We also adhered to the consolidated criteria for reporting qualitative research (COREQ) guidelines (See Appendix A). The COREQ checklist is used to encourage explicit and clear documentation of interview methodologies, ensuring important details are included in this document. Refer to Domains 2 and 3 of the COREQ to see how study design, analysis and findings were outlined. Furthermore, a generative AI was not used in any portion of the manuscript writing.

Results

We interviewed n=19 participants (n=11 patients and n=8 physiatrists; see Table 1 for participant characteristics). We identified 2 key themes that captured patients' and physiatrists' experiences with receiving and providing compassionate care, broadly, and their experiences within the virtual context, specifically.

XSL•FO
RenderX

Table 1. Demographic information for patient and health care professional participants (n=19).

Characteristics	Patients (n=11)	Health care professionals (n=8)
Age (years), mean (SD)	62.2 (9.5)	42.8 (8.5)
Assigned sex at birth, n (%)		
Male	4 (36)	4 (50)
Female	7 (64)	4 (50)
Gender identity, n (%)		
Male	4 (36)	4 (50)
Female	7 (64)	4 (50)
Ethnicity, n (%)		
White-European	7 (64)	3 (37.5)
White-North American	3 (27)	3 (37.5)
Asian-East	1 (9)	2 (25)
Reason for psychiatric care, n (%)		
Amputation	7 (64)	__[a]
Stroke	3 (27)	—
Bypass surgery	1 (9)	—
Received virtual care prior to the COVID-19 pandemic?, n (%)		
Yes	8 (73)	—
No	3 (27)	—
Primary areas of practice, n (%)		
Traumatic brain injury	—	1 (12.5)
Spinal cord injury	—	3 (37.5)
Stroke/neurology	—	5 (62.5)
Musculoskeletal	—	3 (37.5)
Neuromuscular/electromyography	—	2 (25)
Geriatric	—	1 (12.5)
Burns	—	1 (12.5)
Years practicing in physiatry, mean (SD)	—	8.1 (6.8)
Primary practice settings, n (%)		
Acute care facility	—	2 (25)
Specialist inpatient rehabilitation hospital	—	8 (100)
Long-term care or other congregate care setting	—	1 (12.5)
Other: outreach clinics	—	2 (25)

[a]Not applicable.

Theme #1: Compassionate Care Is Rooted in Providers' Inner Caring Intentions that Are Expressed as Outward Caring Behaviors

Our findings suggest that compassionate care (or lack thereof) transcended health care discipline and was understood as stemming from caring attitudes which then manifested as caring behaviors. Caring attitudes and behaviors were enacted to create a safe environment, where patients felt comfortable in discussing their health concerns. Physiatrists were, thus, able to appreciate patients' perspectives, which facilitated individualization of care.

Caring Attitudes Manifest as Caring Behaviors

Compassionate behaviors demonstrated by health care professionals stemmed from their compassionate attitudes and intentions. When patients talked about caring attitudes, they explained that their physiatrists were being "nice" and "pleasant" (Patient 10), "kind" (health care provider [HCP] 08), as well as exercising "patience" (Patient 09). Additionally, physiatrists commonly reported being "sympathetic and empathetic"

XSL•FO
RenderX

(HCP05, HCP08) to better understand the needs of their patients and respond in a "respectful and humane" (HCP01) manner.

While having kindhearted intent was highlighted as a foundation for physiatrists to exercise compassion, caring attitudes were not sufficient in isolation. Rather, participants noted that compassionate intentions needed to be expressed in the form of caring behaviors to "show me [my physiatrist] cares" (Patient 03). For example, patients noted that caring attitudes, such as "empathy" (Patient 01, Patient 03, Patient 05, Patient 09, HCP02, HCP03, HCP05, HCP08), were reflected in attentive behaviors like "actively listening about the situation" (Patient 01). In one instance, a patient described how a physiatrist "respond[ed] to me within ten minutes of my email and getting me an appointment quickly", reinforcing that the patient's concerns were going to be "nipped in the bud" (Patient 03). Patients also felt empathy from their physiatrists when they "spent time with you and talking to you [...], that I think there's an element of understanding, some compassion" (Patient 05).

Physiatrists' caring behaviors were often focused on cultivating interpersonal connections and expressing interest in and respect for patients' concerns and values—understanding that "there's a person there [...] it's not just the body, it's the mind, and how that will affect [the patient] in the future" (Patient 02). Caring behaviors were rooted in "recognizing that this is a person and the person is experiencing suffering" (HCP04); hence, a compassionate approach involved alleviating this suffering rather than just "making a diagnosis and giving somebody a medication" (HCP04). Patients additionally described caring behaviors as humanizing, such as when physiatrists referred to "patients as people and using their name" (Patient 08) and when there was "the caring feeling that you get [from] somebody when you talk to them, you know they're really interested or they really care. It's almost like they become your friend" (Patient 05). Patients often described the importance of feeling "that interconnectedness," which was facilitated by physiatrists "asking how I'm feeling" (Patient 09) or "call[ing] back to things about [patients'] lives that they've mentioned … because [the patient] knows that I've been listening" (HCP03). Interconnectedness was described as being facilitated by "being present and listening and taking no notes" (HCP01), "offer[ing] an assuring pat on the arm" and "reassuring them that they have my undivided attention. I'll say a statement like, I'm here for you. How can I help you?" (HCP07).

At times, caring behaviors were clinically oriented like exercising therapeutic touch or "going out of their way" (Patient 06) to provide care, such as in the case of Patient 06's physiatrist who was "teaching other doctors how to save limbs. And, he's also going out to the Indigenous community and doing the services for them. So, when you hear somebody doing that, then you know that they care."

Having Caring Attitudes and Behaviors Facilitates the Creation of a Safe Space for Individualized Care

Through a combination of caring attitudes and behaviors, patients felt that physiatrists created a safe and comfortable zone where psychological stressors were reduced, and space was created to ask questions and express concerns and fears while their physiatrist remained "non-judgmental" (Patient 08). A safe space was shaped by "giving the time for me" (Patient 02, Patient 03, Patient 10); time for physiatrists "to listen to me, hear me out […] compassionate care to me is just accepting me as I am […] not making me feel judged or less than […] that I can be myself" (Patient 03). Participants noted that a safe space was created over a prolonged period characterized by their physiatrist "know[ing] me" (Patient 04) because physiatrists "really develop[ed] rapport over a longer period of time" (HCP05) with their patients, making them feel safe and "comfortable" (Patient 04, Patient 08) enough to "start rattling off question after question […] and it's like, perfect. She knows me, she knows that I want to try to answer questions and that she immediately feels comfortable" (HCP05). Some patients also spoke about the importance of feeling "wanted" (Patient 08), which stemmed from physiatrists being attentive and allowing them to "ramble on and communicate because I was coming to terms with certain thoughts [...] she allowed for [some self-discovery]" (Patient 01). Physiatrists described the skills needed to create a safe environment, where "listening is a big part of providing an atmosphere that feels open, supportive, so that patients feel like they can open up a fair bit and say what is actually on their mind" (HCP05).

When patients were able to express their questions, concerns, and needs in a safe, nonjudgmental space, physiatrists were better able to individualize care. Patients felt that care was individualized when physiatrists remembered "who I am and that you know me… you don't have to flip through your notes to find out who the heck I am" (Patient 03). Patients noted that their physiatrist knowing who they were, individualizing their care, and "understand[ing] what's best for the individual" (Patient 01) occurred when physiatrist were able to "identify that not everyone's the same and use their professional experience to alter that message to be more specific to what they hear from a condition standpoint. I feel that's compassion" (Patient 01). Conversely, patients felt that the antithesis to individualized care was a "cookie-cutter approach" (Patient 01) that left patients feeling that their physiatrists were "putting [them] in one of the buckets" (Patient 01) without acknowledging their individual histories, limitations, and goals. Physiatrists described individualizing care by incorporating patients' interests and meaningful activities into care plans, which was referred to as "personalizing the experience" (HCP08). For example, "let's say, somewhere in the interview, they mention that they love to walk. And they can't walk because of their foot drop, I will try to work that into conversation later" (HCP01). Physiatrists emphasized that "putting myself in [the patient's] shoes" (HCP06) allowed them to better understand and acknowledge patients' unique interests and circumstances (eg, availability of family support).

Compassionate Care Is Absent When Caring Attitudes and Behaviors Are Lacking

Patients referenced encounters with health care providers more broadly to describe that uncaring attitudes manifested as providers "being rude" (Patient 02, Patient 05, Patient 07), "not car[ing] about [losing my leg]" (Patient 10) and being unsympathetic toward patient limitations, which hindered the provision of compassionate care. For some, descriptions of uncompassionate HCPs included individuals who were

XSL•FO
RenderX

"excellent doctors but crappy human beings [that] had no bedside manners" (Patient 03). For others, this meant "a gruff attitude…and the feeling that they don't give a damn" (Patient 05). Patients described an absence of compassionate care when their providers exhibited uncaring behaviors such as "rushing" (Patient 02, Patient 06) them and having "one-sided discussions" (Patient 01) where patients were not given the opportunity to discuss their concerns. In some cases, the absence of caring attitudes was attributable to a lack of time. For instance, one patient said "I have found [physiatrist] to be quite compassionate when she's not too busy. If she's really busy and time is short, I would say she's very matter of fact as opposed to being compassionate" (Patient 05).

Theme #2: Virtual Elements Influence the Provision and Receipt of Compassionate Care

Participants highlighted many unique advantages to virtual care including that "you can pretty much do it anywhere" (HCP02) and provide it to "anyone" (HCP01), thereby making care "more efficient, time-wise" (HCP04). Despite the fact that virtual care "expanded a lot of opportunity" (HCP02) to access care, participants also highlighted challenges associated with virtual care, such as technical glitches or the fact that not everyone "knew how to do [virtual care]" (HCP03).

The Flexibility of Virtual Care Modalities Allows for Individualized Compassionate Care

The ability to "[give] patients a choice" (HCP01, HCP06, HCP07) of virtual modality (eg, phone, videoconferencing, and email) and platform (eg, Zoom and Doxy.me) was viewed across participants as being compassionate as it promoted a sense of "flexibility" (Patient 03, Patient 10, HCP01), "ease" (Patient 08), "comfort" (HCP01, HCP04, HCP05, Patient 06, Patient 08, Patient 09), and autonomy. When options were provided, patients felt like their physiatrists were being "more personal, more private and more one on one" (Patient 03). Physiatrists explained that "when the [patient] actually felt like [the virtual visit] was more convenient for them" it resulted in a "really meaningful visit" and that "giving [patients] that choice" is what they "enjoy the most" (HCP02, HCP07). Providing virtual care options, specifically those with visual features (eg, Zoom and Doxy.me), was also described as advantageous for individualizing care since physiatrists could "get a sneak peek into their [patient's] home life" (HCP02) and see elements of a patient that perhaps could not be seen otherwise. For example, physiatrists mentioned how they have "had patients show me them dancing with their grandchildren in the backyard on their cell phone […] or having them get up and show me something from their house that wouldn't necessarily be available if they were just in a clinic, [that] personalis[es] the experience [...] And that really is helpful on virtual care because you get those intimate moments, even though you're distanced" (HCP08). Furthermore, the unique features and flexibility of virtual modalities allowed for better family engagement in care appointments because "[patients] can have a family member with them, which often times in the pandemic or even before the pandemic, they couldn't if they had come into clinic physically" (HCP05). Ultimately, this facilitated care individualization because families could offer supplemental information that physiatrists could then use to better-tailor care to the patient.

The Ambiguous Nature of Virtual Care Norms Altered How Caring Attitudes and Behaviors Were Enacted

Ambiguity Concerning Safety

Physiatrists emphasized that the highest priorities of medicine were to avoid patient harm and ensure patient safety. This norm was still noted in a virtual world, where physiatrists described safety concerns that prevented them from "examin[ing] [the patient]" because they were in an inappropriate or unsafe environment, like "sitting in a truck at the side of the road on a construction site" (HCP03). Physiatrists noted how taking appointments from inappropriate locations was "a bit of a challenge because some patients don't treat it like a medical appointment … they're distracted" (HCP01). Patients agreed that they did not always take their medical appointments seriously and that they were "a little bit too comfortable because […] I didn't have to get dressed up, I didn't have to make any kind of an impression" (Patient 08), illustrating uncertainty as to what was and was not appropriate etiquette for virtual appointments.

Physiatrists also said that, because of the types of populations that they worked with (eg, people with complex disabilities, impaired mobility, high falls risks), they "have to be cognisant of certain safety aspects of what's happening with the patient. If they're not steady, you can't get them up to stand because they might fall" (HCP02). In addition, the ambiguity around ensuring patient safety in a virtual setting was described in relation to concerns around "consent, confidentiality, and documentation…It was like…we need to do virtual care yesterday. Here's what you [need] to do [to] make sure you've covered your bases" (HCP07). Overall, physiatrists' comments highlighted that physical safety and privacy were of paramount importance. This meant first ensuring the protection of patients' health information and offering virtual care options that were used "in a safe way" (HCP03)—a way that considered both confidentiality and patient choice because "as long as patients provide the level of consent, you can use any virtual platform that you wish" (HCP03).

Ambiguity Concerning Presence

Participants' perspectives varied when it came to norms around facilitating a sense of presence and feeling as if their physiatrist was in the moment with them. Patients noted how certain behaviors, like a physiatrist "tak[ing] notes while they were typing into a computer," left patients feeling like their physiatrist was not really present and "not really engaged with me" (Patient 09). Some patients highlighted how feeling connected to their physiatrist was particularly difficult in an online setting because of uncaring behaviors like when "there's one eye always on the clock to make sure that you can make the next meeting. I just found that you get that impression in any video meeting, but I don't get that as much in the in-person meetings" (Patient 07). Physiatrists agreed that "it's difficult to engage in the same way" (HCP02) because interpersonal caring behaviors that help facilitate a person-to-person connection are often intangible and can be challenging to enact in an online setting. Regardless, it

XSL•FO
RenderX

was universally understood that enacting interpersonal caring behaviors and facilitating a sense of engagement were easier "when you already have a relationship with the patient" (HCP07) because "virtually it's really difficult to create that connection […] that interconnection is just not being established as well. That rapport is just not happening" (Patient 09).

There were also varied comments around promptness and timing in a virtual setting. In some cases, physiatrists expressed being "more comfortable" if "a phone [appointment] ends up being 15 or 20 minutes late, because [the patient] is on the phone … they're not sitting in clinic, doing nothing but wait[ing] for me" (HCP01). Conversely, other physiatrists felt worried that if they were late to a virtual appointment, patients may "sign off or get cut off" (HCP02). The prevailing sentiment among physiatrists, however, was that regardless of the modality, being "on time is very good" (Patient 07) because a lack of punctuality was viewed as cold, disrespectful of patient's time, and inherently "uncaring."

Compassionate Behaviors and Individualized Care Are Challenged in Virtual Physiatry

Virtual care challenged physiatry practice both in terms of exercising core clinical tasks that required physical presence, and in exercising compassion. Physiatrists strongly emphasized that being physically present with patients was particularly important for physiatry because a "physical examination is very much a standard part of a physiatry assessment. And it's important for multiple reasons in terms of assessing function […] thinking about things like dexterity, things like mobility […] things that [have] to be assessed hands on" (HCP04). Patients echoed similar sentiments explaining that "if [PHYSIATRIST] wanted to look at my leg [to help with the prosthetic] […] well you can talk about it over the phone, but you can't fix a prosthetic unless you go in" (Patient 07). The inability to perform clinical tasks and physically assess a patient further challenged physiatrists in getting "a sense of who the person is, what their needs are" (HCP01), which made it difficult for them to individualize and tailor care. Physiatrists also noted that other clinical behaviors like "handing [the patient] a tissue if they start to cry" (HCP02) or "putting your hand on their shoulder. […] even maybe hugging a patient" (HCP05) were tested when there was physical separation. The inability to enact many caring behaviors in the absence of physical presence left physiatrists feeling like they did not always "know the patient" or "know what [they] can do to support them" (HCP05). Since the virtual environment precludes many typical caring behaviors, patients stated that their physiatrists needed to "compensate for the fact you're not physically there. The fact that you can't tap me on the arm or you can't rub my arm and say you're going to be fine if I start to cry, I think you need something else to make me feel that you're listening" (Patient 09).

Discussion

Principal Findings

This qualitative descriptive study explored the experiences of patients and physiatrists engaged in virtual care with the goal of understanding how compassion was expressed and received in the virtual modality. First, findings underscored that compassionate care stemmed from providers' caring attitudes, which manifested as caring behaviors. Participants felt compassionate care was characterized by the creation of a psychologically safe space for patients where care could be individualized. Second, findings suggest that the virtual modality both positively and negatively impacts how compassion is enacted by physiatrists and received by patients. There was also ambiguity around the norms and etiquette for virtual care.

The way that compassionate care was described in our study aligns with the growing body of literature on this topic. Participants emphasized that compassion began with providers' internal motivations and benevolent intentions but that those intentions necessarily had to manifest as tangible behaviors and expressions of care. This echoes other research that has demonstrated that while "virtuous intent" is a main driver of compassion, compassionate care is cultivated through "contemplative practices" (eg, actions, behaviors, self-reflection) [27]. In essence, providers' internal feelings, values, and morals are what drive actions that aim to alleviate pain and suffering in others, which is the heart of compassionate care [28]. These findings suggest that providers' intentions and motivations positively impact the care received by patients. Many participants described compassionate care in terms of the psychologically safe environment that it enabled. This included statements about feeling "safe" to ask questions and express concerns, knowing that their physiatrist knew them "as a person" (eg, recalling details about their life and family without checking notes), and feeling that their physiatrist gave them "time" (ie, not rushing through appointments). In accordance to previous studies, creating a "safe space" and making patients feeling valued, important, and "known" are all important outcomes of compassionate acts by providers [29,30]. Hence, these results support the idea that compassionate care often is received by patients in a psychologically safe location.

Conversely, participants in our study felt that compassion was lacking when providers' behaviors and acts did not reflect concern or care for the patient (eg, provider being rude, rushing through an appointment, or not listening to the patient). This highlights that even though these providers may have had good intentions, patients will not experience compassionate care unless those intentions translate into caring acts and behaviors. Even though individual providers' "baseline compassion aptitude" varies, skills in compassionate responding can be taught, for example, learning to enhance affective and relational components through role playing [31]. Indeed, while arguments have been made in the past that compassionate is innate and that you either "have it or you do not," mounting research supports the notion that compassion can be cultivated through learned behaviors, principles, and strategies [32-34]. These findings raise intriguing questions regarding the nature of compassion and how it is learned and expressed by providers.

The virtual modality was viewed as impacting *expressions* of compassion by providers and *receipt* of compassion by patients. A notable theme here was that virtual care's naissance underpinned many ambiguities about the appropriate norms and etiquette within this setting. In the most extreme cases, physiatrists expressed concerns about their patients engaging

XSL•FO
RenderX

in virtual appointments from inappropriate, potentially unsafe locations, which undermined compassionate care in as far as physiatrists were primarily concerned with ensuring their patient's safety. This issue has been noted in other literature and points to the challenge of providers having limited control over a patient's virtual care environment and etiquette [35]. Some recommendations to address this issue include providers communicating expectations to patients around how to safely and optimally engage in virtual care appointments [35] and providing education that helps patients value virtual care in the same way as in-person care [36].

On the provider side, the idea of "webside" manner has gained a lot of traction with the rapid pivot to virtual care throughout the pandemic and refers to "virtual care etiquette" to be exercised by providers [37]. Good "webside" manner entails noting patient needs, taking time to clarify any concerns, and following up with their care as a way of exercising kindness and compassion in the virtual setting [37]. In our study, a lack of "webside manner" was experienced by some patients who felt that an interpersonal connection with their provider was absent and that their provider was not always "present" and "in the moment" (eg, taking notes on screen, looking off at a clock). One element that was ubiquitously mentioned by participants as mitigating this challenge was having an established patient-physiatrist relationship that could then translate to the virtual setting. This aligns with other studies of compassionate care that have emphasized the importance of mutual trust and continuity of care between patients and providers [38]. In the absence of a pre-existing relationship, providers must be additionally intentional in building rapport with patients and working to make appointments feel less rushed and more conversational [38].

Our focus on physiatry elucidated that virtual care can be especially challenging for the field of PM&R due to physiatrists' scope of practice and the core clinical tasks in which they routinely engage. In this context, physical examination is a central part of standard practice (eg, physical manipulation of limbs, assessing prosthetics, assessing function and dexterity, etc). Although telemedicine can be optimally used for symptom monitoring, prescribing, medical advice, and psychological support (ie, tasks that typically do not require any physical presence [39]), other rehabilitation interventions (eg, physical assessments and examinations) do not necessarily have good validity when conducted virtually compared to in person [40,41]. This emphasizes that in the field of PM&R especially, there may be many opportunities to leverage the advantages of virtual care, but that in-person assessment and care remain necessary under certain conditions.

Although the virtual setting posed several challenges to compassionate care, it is important to note that the flexibility afforded by the modality was unanimously considered as an enabler of compassionate and person-centered care by patients and physiatrists. Person-centered care is defined as care that respects an individual's preferences, needs, and values, and is provided in an empathic and compassionate way [42]. In alignment with this definition, participants in our study felt that the flexibility afforded by virtual care led to care being more individualized, considerate of unique needs and circumstances. Moving forward—during and beyond the COVID-19 pandemic—it is important to build on the many advantages of virtual care that have been identified, particularly in terms of enhancing accessibility to diverse patient populations, while also mitigating less favorable aspects of the virtual environment which otherwise might serve to undermine the provision of compassionate patient care.

Strengths and Limitations

To the best of our knowledge, this is the first study to focus on compassionate care in the field of physiatry and in the context of virtual care. We were able to achieve a more complete understanding of compassionate care by including both physiatrist and patient perspectives. This allowed us to elucidate how compassion is both *enacted* and *experienced* in physiatry practice broadly, and in the virtual care context more specifically. One limitation is that nearly all participants were Caucasian and lived or practiced in major metropolitan areas, making our findings less transferable to patients and providers of other ethnicities and those who live or practice in rural settings. Our study also took place during a period where compensation mechanisms for virtual care where enhanced and streamlined due to the COVID-19 pandemic, which eased provision of this type of care and allowed for exploration of compassion in this context. Thus, our findings may not be transferable to circumstances when these mechanisms are no longer in place (due to reduction or elimination of virtual care compensation codes).

Conclusions

Although virtual care has the potential to optimize patients' access to therapy and treatment, great caution must be taken to maintain the compassionate elements of care; especially in the field of PM&R. Compassion is rooted in providers' inner caring attitudes which manifest as caring behaviors and enable individualized care and the creation of a safe space for patients. Notably, the physical nature of core clinical tasks in physiatry was challenged by the virtual setting. Moving forward, the flexibility and person-centeredness of virtual care render it a useful modality in a routine sense, as long as in-person care is provided too, when necessary. Greater patient education, clinical guidance, and competency training for physiatrists in this area may help optimize safe and compassionate virtual care in PM&R settings.

Acknowledgments

This work was supported by an AMS Small Grant in Compassion and Artificial Intelligence. No generative AI was used for this study.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient Interview Guide.
[DOCX File , 17 KB - jmir_v26i1e51878_app1.docx]

Multimedia Appendix 2
Physiatrist Interview Guide.
[DOCX File , 17 KB - jmir_v26i1e51878_app2.docx]

References

1. Reddy A, Arthur J, Dalal S, Hui D, Subbiah I, Wu J, et al. Rapid transition to virtual care during the COVID-19 epidemic: experience of a supportive care clinic at a tertiary care cancer center. J Palliat Med 2021;24(10):1467-1473. [doi: 10.1089/jpm.2020.0737] [Medline: 33535019]
2. Malliaras P, Merolli M, Williams CM, Caneiro JP, Haines T, Barton C. 'It's not hands-on therapy, so it's very limited': telehealth use and views among allied health clinicians during the coronavirus pandemic. Musculoskelet Sci Pract 2021;52:102340 [FREE Full text] [doi: 10.1016/j.msksp.2021.102340] [Medline: 33571900]
3. Tenforde AS, Hefner JE, Kodish-Wachs JE, Iaccarino MA, Paganoni S. Telehealth in physical medicine and rehabilitation: a narrative review. PMR 2017;9(5S):S51-S58. [doi: 10.1016/j.pmrj.2017.02.013] [Medline: 28527504]
4. Leochico CFD, Mojica JAP, Rey-Matias RR, Supnet IE, Ignacio SD. Role of telerehabilitation in the rehabilitation medicine training program of a COVID-19 referral center in a developing country. Am J Phys Med Rehabil 2021;100(6):526-532. [doi: 10.1097/PHM.0000000000001755] [Medline: 33998606]
5. Prvu Bettger J, Resnik LJ. Telerehabilitation in the age of COVID-19: an opportunity for learning health system research. Phys Ther 2020;100(11):1913-1916 [FREE Full text] [doi: 10.1093/ptj/pzaa151] [Medline: 32814976]
6. Howard IM, Kaufman MS. Telehealth applications for outpatients with neuromuscular or musculoskeletal disorders. Muscle Nerve 2018;58(4):475-485. [doi: 10.1002/mus.26115] [Medline: 29510449]
7. American Academy of Physical Medicine and Rehabilitation. URL: https://www.aapmr.org/about-physiatry/about-physical-medicine-rehabilitation [accessed 2024-06-28]
8. Appleby E, Gill ST, Hayes LK, Walker TL, Walsh M, Kumar S. Effectiveness of telerehabilitation in the management of adults with stroke: a systematic review. PLoS One 2019;14(11):e0225150 [FREE Full text] [doi: 10.1371/journal.pone.0225150] [Medline: 31714924]
9. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020;1(1):CD010255 [FREE Full text] [doi: 10.1002/14651858.CD010255.pub3] [Medline: 32002991]
10. Chen J, Jin W, Zhang XX, Xu W, Liu XN, Ren CC. Telerehabilitation approaches for stroke patients: systematic review and meta-analysis of randomized controlled trials. J Stroke Cerebrovasc Dis 2015;24(12):2660-2668. [doi: 10.1016/j.jstrokecerebrovasdis.2015.09.014] [Medline: 26483155]
11. Cottrell MA, Galea OA, O'Leary SP, Hill AJ, Russell TG. Real-time telerehabilitation for the treatment of musculoskeletal conditions is effective and comparable to standard practice: a systematic review and meta-analysis. Clin Rehabil 2017;31(5):625-638. [doi: 10.1177/0269215516645148] [Medline: 27141087]
12. Tabak M, Vollenbroek-Hutten MM, van der Valk PD, van der Palen J, Hermens HJ. A telerehabilitation intervention for patients with chronic obstructive pulmonary disease: a randomized controlled pilot trial. Clin Rehabil 2014;28(6):582-591. [doi: 10.1177/0269215513512495] [Medline: 24293120]
13. Bezuidenhout L, Joseph C, Thurston C, Rhoda A, English C, Conradsson DM. Telerehabilitation during the COVID-19 pandemic in Sweden: a survey of use and perceptions among physiotherapists treating people with neurological diseases or older adults. BMC Health Serv Res 2022;22(1):555 [FREE Full text] [doi: 10.1186/s12913-022-07968-6] [Medline: 35473602]
14. Fang BK, Jiang JJ, Loh JKS, Ismail SAB. Telerehabilitation acceptance among patients during circuit breaker period: a retrospective study. Dialogues Health 2022;1:100049 [FREE Full text] [doi: 10.1016/j.dialog.2022.100049] [Medline: 38515891]
15. Gomez T, Anaya YB, Shih KJ, Tarn DM. A qualitative study of primary care physicians' experiences with telemedicine during COVID-19. J Am Board Fam Med 2021;34(Suppl):S61-S70 [FREE Full text] [doi: 10.3122/jabfm.2021.S1.200517] [Medline: 33622820]
16. Goetz JL, Keltner D, Simon-Thomas E. Compassion: an evolutionary analysis and empirical review. Psychol Bull 2010;136(3):351-374 [FREE Full text] [doi: 10.1037/a0018807] [Medline: 20438142]
17. Calabrò RS. Teleneurorehabilitation in the COVID-19 era: what are we doing now and what will we do next? Med Sci (Basel) 2021;9(1):15 [FREE Full text] [doi: 10.3390/medsci9010015] [Medline: 33668321]

XSL•FO
RenderX

18. Shachak A, Alkureishi MA. Virtual care: a 'Zoombie' apocalypse? J Am Med Inform Assoc 2020;27(11):1813-1815 [FREE Full text] [doi: 10.1093/jamia/ocaa185] [Medline: 32940711]
19. Bateman EA, Viana R. Burnout among specialists and trainees in physical medicine and rehabilitation: a systematic review. J Rehabil Med 2019;51(11):869-874 [FREE Full text] [doi: 10.2340/16501977-2614] [Medline: 31608964]
20. Posa S, Wasilewski MB, Mercer SW, Simpson S, Robinson LR, Simpson R. Conceptualization, use, and outcomes associated with empathy and compassion in physical medicine and rehabilitation: a scoping review. Int J Rehabil Res 2022;45(4):291-301 [FREE Full text] [doi: 10.1097/MRR.0000000000000542] [Medline: 35837691]
21. Doyle L, McCabe C, Keogh B, Brady A, McCann M. An overview of the qualitative descriptive design within nursing research. J Res Nurs 2020;25(5):443-455 [FREE Full text] [doi: 10.1177/1744987119880234] [Medline: 34394658]
22. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
23. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
24. Thorogood N, Green J. Qualitative Methods for Health Research. London, England: SAGE Publications Ltd; 2009.
25. NVivo Qualitative Data Analysis Software.: QSR International Pty Ltd; 2010. URL: https://yosigo.ugr.es/en/recurso/nvivo-qualitative-data-analysis-software-qsr-international/ [accessed 2024-06-28]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Sinclair S, Hack TF, McClement S, Raffin-Bouchal S, Chochinov HM, Hagen NA. Healthcare providers perspectives on compassion training: a grounded theory study. BMC Med Educ 2020;20(1):249 [FREE Full text] [doi: 10.1186/s12909-020-02164-8] [Medline: 32758216]
28. Perez-Bret E, Altisent R, Rocafort J. Definition of compassion in healthcare: a systematic literature review. Int J Palliat Nurs 2016;22(12):599-606. [doi: 10.12968/ijpn.2016.22.12.599] [Medline: 27992278]
29. Sharp S, McAllister M, Broadbent M. The vital blend of clinical competence and compassion: how patients experience person-centred care. Contemp Nurse 2016;52(2-3):300-312. [doi: 10.1080/10376178.2015.1020981] [Medline: 26077823]
30. Sinclair S, Norris JM, McConnell SJ, Chochinov HM, Hack TF, Hagen NA, et al. Compassion: a scoping review of the healthcare literature. BMC Palliat Care 2016;15:6 [FREE Full text] [doi: 10.1186/s12904-016-0080-0] [Medline: 26786417]
31. Sinclair S, Kondejewski J, Jaggi P, Dennett L, Roze des Ordons AL, Hack TF. What is the state of compassion education? A systematic review of compassion training in health care. Acad Med 2021;96(7):1057-1070 [FREE Full text] [doi: 10.1097/ACM.0000000000004114] [Medline: 33830949]
32. Shield RR, Tong I, Tomas M, Besdine RW. Teaching communication and compassionate care skills: an innovative curriculum for pre-clerkship medical students. Med Teach 2011;33(8):e408-e416. [doi: 10.3109/0142159X.2011.586748] [Medline: 21774636]
33. Adamson E, Dewar B. Compassionate care: student nurses' learning through reflection and the use of story. Nurse Educ Pract 2015;15(3):155-161. [doi: 10.1016/j.nepr.2014.08.002] [Medline: 25754833]
34. Betcher DK. Elephant in the room project: improving caring efficacy through effective and compassionate communication with palliative care patients. Medsurg Nurs 2010;19(2):101-105. [Medline: 20476520]
35. Tewksbury C, Deleener ME, Dumon KR, Williams NN. Practical considerations of developing and conducting a successful telehealth practice in response to COVID-19. Nutr Clin Pract 2021;36(4):769-774 [FREE Full text] [doi: 10.1002/ncp.10742] [Medline: 34156725]
36. McCrady E, Strychowsky JE, Woolfson JP. Experience of pediatricians and pediatric surgeons with virtual care during the COVID-19 pandemic: descriptive study. JMIR Pediatr Parent 2022;5(2):e34115 [FREE Full text] [doi: 10.2196/34115] [Medline: 35666938]
37. Wong A, Bhyat R, Srivastava S, Boissé Lomax L, Appireddy R. Patient care during the COVID-19 pandemic: use of virtual care. J Med Internet Res 2021;23(1):e20621 [FREE Full text] [doi: 10.2196/20621] [Medline: 33326410]
38. Imlach F, McKinlay E, Middleton L, Kennedy J, Pledger M, Russell L, et al. Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. BMC Fam Pract 2020;21(1):269 [FREE Full text] [doi: 10.1186/s12875-020-01336-1] [Medline: 33308161]
39. Snoswell CL, Chelberg G, De Guzman KR, Haydon HM, Thomas EE, Caffery LJ, et al. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. J Telemed Telecare 2023;29(9):669-684. [doi: 10.1177/1357633X211022907] [Medline: 34184580]
40. Mani S, Sharma S, Omar B, Paungmali A, Joseph L. Validity and reliability of internet-based physiotherapy assessment for musculoskeletal disorders: a systematic review. J Telemed Telecare 2017;23(3):379-391. [doi: 10.1177/1357633X16642369] [Medline: 27036879]
41. Russell T, Truter P, Blumke R, Richardson B. The diagnostic accuracy of telerehabilitation for nonarticular lower-limb musculoskeletal disorders. Telemed J E Health 2010;16(5):585-594. [doi: 10.1089/tmj.2009.0163] [Medline: 20575726]
42. Boyd K. The curriculum of caring: fostering compassionate, person-centered health care. AMA J Ethics 2016;18(4):384-392 [FREE Full text] [doi: 10.1001/journalofethics.2016.18.4.medu1-1604] [Medline: 27099187]

XSL•FO
RenderX

Abbreviations

CAPM&R: Canadian Association for Physical Medicine and Rehabilitation
COREQ: consolidated criteria for reporting qualitative research
HCP: health care provider
PM&R: physical medicine and rehabilitation

Edited by T de Azevedo Cardoso; submitted 22.08.23; peer-reviewed by H Pilabré, S Kaur; comments to author 05.12.23; revised version received 20.03.24; accepted 30.05.24; published 06.08.24.

Please cite as:
Wasilewski MB, Vijayakumar A, Szigeti Z, Mayo A, Desveaux L, Shaw J, Hitzig SL, Simpson R
Patient and Provider Experiences With Compassionate Care in Virtual Physiatry: Qualitative Study
J Med Internet Res 2024;26:e51878
URL: https://www.jmir.org/2024/1/e51878
doi:10.2196/51878
PMID:39106094

©Marina B Wasilewski, Abirami Vijayakumar, Zara Szigeti, Amanda Mayo, Laura Desveaux, James Shaw, Sander L Hitzig, Robert Simpson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identification of Motivational Determinants for Telemedicine Use Among Patients With Rheumatoid Arthritis in Germany: Secondary Analysis of Data From a Nationwide Cross-Sectional Survey Study

Felix Muehlensiepen[1,2*], BA, MPH, Dr rer medic; Pascal Petit[2*], PhD; Johannes Knitza[2,3,4], MHBA, MD, PD, PhD; Martin Welcker[5], MD; Nicolas Vuillerme[2,6,7], PhD

[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Rüdersdorf bei Berlin, Germany

[2]AGEIS, Université Grenoble Alpes, Grenoble, France

[3]Department of Internal Medicine 3, Friedrich-Alexander-University Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

[4]Institute for Digital Medicine, University Hospital Giessen-Marburg, Philipps University Marburg, Marburg, Germany

[5]Medizinisches Versorgungszentrum für Rheumatologie Dr M Welcker GmbH, Planegg, Germany

[6]Institut Universitaire de France, Paris, France

[7]LabCom Telecom4Health, Orange Labs & Université Grenoble Alpes, CNRS, Inria, Grenoble, France

[*]these authors contributed equally

Corresponding Author:
Felix Muehlensiepen, BA, MPH, Dr rer medic
Center for Health Services Research
Faculty of Health Sciences Brandenburg
Brandenburg Medical School Theodor Fontane
Seebad 82/83
Rüdersdorf bei Berlin, 16816
Germany
Phone: 49 15119126024
Email: Felix.Muehlensiepen@mhb-fontane.de

Abstract

Background: Previous studies have demonstrated telemedicine to be an effective tool to complement rheumatology care and address workforce shortage. With the COVID-19 outbreak, telemedicine experienced a massive upswing. An earlier analysis revealed that the motivation of patients with rheumatic and musculoskeletal diseases to use telemedicine is closely connected to their disease. It remains unclear which factors are associated with patients' motivation to use telemedicine in certain rheumatic and musculoskeletal diseases groups, such as rheumatoid arthritis (RA).

Objective: This study aims to identify factors that determine the willingness to try telemedicine among patients diagnosed with RA.

Methods: We conducted a secondary analysis of data from a German nationwide cross-sectional survey among patients with RA. Bayesian univariate logistic regression analysis was applied to the data to determine which factors were associated with willingness to try telemedicine. Predictor variables (covariates) studied individually included sociodemographic factors (eg, age, sex) and health characteristics (eg, health status). All the variables positively and negatively associated with willingness to try telemedicine in the univariate analyses were then considered for Bayesian model averaging analysis after a selection based on the variance inflation factor (≤ 2.5) to identify determinants of willingness to try telemedicine.

Results: Among 438 surveyed patients in the initial study, 210 were diagnosed with RA (47.9%). Among them, 146 (69.5%) answered either yes or no regarding willingness to try telemedicine and were included in the analysis. A total of 22 variables (22/55, 40%) were associated with willingness to try telemedicine (region of practical equivalence %≤5). A total of 9 determinant factors were identified using Bayesian model averaging analysis. Positive determinants included desiring telemedicine services provided by a rheumatologist (odds ratio [OR] 13.7, 95% CI 5.55-38.3), having prior knowledge of telemedicine (OR 2.91, 95% CI 1.46-6.28), residing in a town (OR 2.91, 95% CI 1.21-7.79) or city (OR 0.56, 95% CI 0.23-1.27), and perceiving one's health status as moderate (OR 1.87, 95% CI 0.94-3.63). Negative determinants included the lack of an electronic device (OR 0.1, 95% CI 0.01-0.62), absence of home internet access (OR 0.1, 95% CI 0.02-0.39), self-assessment of health status as bad (OR 0.44,

XSL•FO
RenderX

95% CI 0.21-0.89) or very bad (OR 0.47, 95% CI 0.06-2.06), and being aged between 60 and 69 years (OR 0.48, 95% CI 0.22-1.04) or older than 70 years (OR 0.38, 95% CI 0.16-0.85).

Conclusions: The results suggest that some patients with RA will not have access to telemedicine without further support. Older patients, those not living in towns, those without adequate internet access, reporting a bad health status, and those not owning electronic devices might be excluded from the digital transformation in rheumatology and might not have access to adequate RA care. These patient groups certainly require support for the use of digital rheumatology care.

(J Med Internet Res 2024;26:e47733) doi:10.2196/47733

KEYWORDS

telemedicine; rheumatoid arthritis; rheumatology; primary care; health services research; eHealth; data analysis; survey; Germany; tool; care; willingness; sociodemographic; age; telehealth; digital transition

Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory joint disease, which can cause cartilage and bone damage as well as disability [1]. RA is the most common rheumatic disease with a global age-standardized point prevalence and annual incidence rates of 246.6 and 14.9 per 100,000 population in 2017 [2]. RA is associated with high medical costs [3] and contributes to a significant deterioration in quality of life [4]. Patients in rural areas usually have limited access to rheumatology care and therefore, accept longer diagnosis times [5]. The increasing shortage of rheumatologists alongside rising demand makes RA a major global public health challenge [2].

Telemedicine offers a promising opportunity to support RA care [6,7]. It has the potential to address workforce shortage [8] as well as disparity in rheumatology care for rural patients [4]. The effective integration of telemedicine into standard rheumatology care, depends on the willingness and ability of end users to engage with telemedicine [9,10].

A previous study revealed that patients' motivation to use telemedicine is closely connected to their specific disease [11]. However, it is still unclear which factors determine the motivation of patients with RA for using telemedicine. To gain a better understanding of these factors, we performed a secondary analysis using data from a German nationwide cross-sectional survey conducted earlier [10]. The aim of this study was to identify the factors associated with RA patients' willingness to try telemedicine.

Methods

Study Design

This work reports on results from a secondary analysis of data collected as part of a cross-sectional, self-completed, and paper-based survey of German patients with rheumatic and musculoskeletal diseases (RMDs) in collaboration with the patient organization German League against Rheumatism (Deutsche Rheuma-Liga, Landesvertretung Brandenburg) and outpatient rheumatologists. The eligibility criteria to participate in the survey were being (1) a patient in rheumatology care; (2) aged 18 years and older; and (3) in Germany. The survey was embedded in a more than 2-year mixed methods study, investigating acceptance, opportunities, and obstacles to the implementation of telemedicine [10]. The actual survey was conducted between September 1, 2019, and December 30, 2019. The exact methodology applied for the survey has been described previously [10].

Data Selection or Population Considered

From the aforementioned German nationwide survey, a data set included 438 patients in total, only 210 (47.9%) patients were diagnosed with RA and were considered in this study. These patients answered a total of 26 questions (Table 1) [11]. Individuals with missing answer regarding willingness to try telemedicine or that answered, "do not know" (Q11: "Would you like to try telemedicine?") were excluded from this study. As a result, a total of 146 (146/210, 69.5%) patients were considered for analysis.

To check if this sample size was enough to answer the goal of our study, we conducted a sample size estimation based on a 1-sample proportion test aiming at determining whether the proportion of individuals who wanted to try telemedicine was significantly different from participants who did not want to try telemedicine. Due to the lack of similar studies regarding this topic, the sample size estimation was based on conservative measures. To that end, we considered a significant level of .05, a statistical power of 0.80, and a small effect size (h=0.25) [12], resulting to an estimated sample size of 126. The choice of this size effect value was motivated by a previous study [11]. In this previous study, there was a response rate of 71.7% regarding the willingness to try telemedicine among patients, with 116 of 282 patients (41.1%) who were willing to try telemedicine. Assuming that a similar proportion of patients is willing to try telemedicine and a similar response rate among the subsample of patients diagnosed with RA, the corresponding Cohen *h* (effect size) would be 2asin(sqrt(0.41))–2asin(sqrt(0.410.717))=0.244≈0.25, which corresponds to a medium effect size [12-14]. Using the exact Cohen *h* value would have yielded an estimated sample size of 132.

To address the selection bias introduced by excluding individuals with missing answer regarding willingness to try telemedicine or that answered, "do not know," a comparison of the characteristics of included and excluded participants was performed (Multimedia Appendix 1). To that end, a Wilcoxon-Mann-Whitney rank sum test was performed for continuous variables, while the chi-square test for count data was used for categorical variables. The Wilcoxon effect size and its 95% CI were computed to quantify the difference

XSL•FO
RenderX

between both groups using the Vargha and Delaney *A* statistic [15]. To measure the strength of the association between categorical variables, the Cramér *V* was computed, with a value of 1 corresponding to complete association and of 0 corresponding to no association between the variables.

Table 1. Regression analysis—variables considered (n=146).

	Modality	Response rate, n (%)
Dependent variable		
Q11: "Would you like to try telemedicine?"	2 categories: yes or no	146 (100)
Independent variables		
Q1: "How far do you drive to your rheumatology doctor's office?"	6 categories: up to 10 km, 10-20 km, 20-30 km, 30-40 km, >40 km, not answered	145 (99.3)
Q2: "How far do you drive to your general practitioner's office?"	5 categories: up to 5 km, 5-10 km, 10-15 km, >15 km, not answered	146 (100)
Q3: "Have you ever contacted your doctor's office using an electronic means?"	3 categories: yes, no, not answered	146 (100)
Q4: "Do you own an electronic device?"	3 categories: yes, no, not answered	145 (99.3)
Q5: "Do you have internet access at home?"	3 categories: yes, no, not answered	146 (100)
Q8: "Prior to this survey, had you ever heard the term 'telemedicine'?"	4 categories: yes, no, do not know, not answered	145 (99.3)
Q14: "Would you like your rheumatologist offer you telemedicine services?"	4 categories: yes, no, do not know, not answered	139 (95.2)
Q16: "Do you document your health status?"	4 categories: yes on paper, yes digitally, no, not answered	139 (95.2)
Q17: Age	4 categories: <60 years, 60-69 years, >70 years, not answered	144 (98.6)
Q18: Biological sex	3 categories: female, male, not answered	144 (98.6)
Q20: "How do you rate your health status?"	6 categories: very good, good, moderate, bad, very bad, not answered	141 (96.6)
Q21: "Are you in rheumatology treatment?"	4 categories: yes, no I am a new patient, do not know, not answered	146 (100)
Q23: "My place of residence is…"	5 categories: city (>100,000 inhabitants), town (20,000-100,000 inhabitants), provincial town (5000-20,000 inhabitants), rural area (<5000 inhabitants), not answered	144 (98.6)

Choice of Variables

Questions related to other RMDs were not considered in this study, leaving a total of 14 questions (Table 1). "Would you like to try telemedicine?" (Q11) was used as dependent variable while age (Q17), biological sex (Q18), distance to the rheumatologist's office (Q1), distance to the GP's office (Q2), previous electronic contact with physician (Q3), electronic device ownership (Q4), internet access (Q5), self-reported knowledge regarding telemedicine (Q8), self-reported willingness that rheumatologist offer telemedicine services (Q16), self-reported health status (Q20), means of health status documentation (Q16), rheumatology treatment (Q21), and place of residence (Q23) were used as independent variables. For questions other than willingness to try telemedicine (Q11: "Would you like to try telemedicine?"), missing values (no answer) were considered as a new category for the regression analysis [16]. For instance, Q4 ("Do you own an electronic device?") had previously 2 categories and was considered with 3 (yes, no, and not answered). As a result, 55 answers from 13 independent variables were considered (Table 1). For statistical analysis, all the categorical variables having more than 2 modalities (eg, "yes," "no," "do not know") were transformed into dummy or binary variables. For instance, Q4 ("Do you own an electronic device?") was transformed into 3 dummy variables. Age was considered as both continuous and categorical variables.

Regression Analysis

Bayesian univariate logistic regression analyses was applied to the data in order to determine which factors were associated with willingness to try telemedicine (Q11: "Would you like to try telemedicine?").

For each model, odds ratios (ORs) with 95% CI are presented. All the individual variables associated (positively or negatively) with willingness to try telemedicine in the Bayesian univariate analysis were considered for analysis in later Bayesian multivariate analysis after variable selection. This variable selection was based on the ROPE (region of practical equivalence) percentage (ROPE%≤ 5) [17-19], and a subsequent selection based on the variance inflation factor (VIF) [20]. Collinear covariates, with a VIF>2.5, were excluded in the

XSL•FO
RenderX

multivariate models [21]. Finally, determinants of willingness to try telemedicine were identified through Bayesian model averaging (BMA) [22].

All statistical analyses were performed using R software (version 4.1.2; R Foundation for Statistical Computing) for Windows 10. The sample size estimation was performed using the *pwr* package version 1.3-0 [13]. Bayesian estimation was performed using the *rstanarm* package version 2.21.1 [23,24]. Weakly informative priors (default prior in *rstanarm*) were used. The default priors in *rstanarm* package are designed to be weakly informative. The Bayesian model adds priors (independent by default) on the coefficients of the generalized linear model. The Bayesian estimation was performed via Markov chain Monte Carlo Bernoulli model, with 4 randomly initialized Markov chains, each for 2000 iterations (including a warm-up period of 1000 iterations that is discarded). The BMA was undertaken with the *BMA* package version 3.18.15 [25]. Regarding priors for BMA, we assumed that all candidate models were equally likely a priori (same prior weight).

Ethical Considerations

Primary data collection was conducted in compliance with the current data protection regulations of the General Data Protection Regulation [26] and the Helsinki Declaration. All the study participants were informed about the research project. Sending the questionnaire back to the study center was considered as consent. No personal data was collected in the survey; therefore, no written informed consent was required in accordance with the General Data Protection Regulation [26]. Participants did not receive any compensation for participating in this study. The institutional review board of the Theodor Fontane Medical School in Brandenburg issued a waiver for the secondary analyses of the survey data (164112023-ANF).

Results

Population Characteristics

The response rate from the 146 patients with RA included in this study varied from 95.2% (139/146) for Q14 (“Would you like your rheumatologist offer you telemedicine services?”) and Q16 (“Do you document your health status?”) to 100% (146) for Q2 (“How far do you drive to your general practitioner’s office?”), Q3 (“Have you ever contacted your doctor’s office using an electronic means?”), Q5 (“Do you have internet access at home?”), and Q21 (“Are you in rheumatology treatment?”), respectively (Table 1). Most patients with RA were female (111/146, 76.0%), with an average age of 62 (SD 13.2) years and lived in provincial town or rural area (86/146, 58.9%; Table 2). Half of the patients with RA considered they had bad or very bad health status (70/146, 47.9%), while 41.1% (60/146) considered their health status as moderate. Most patients with RA had internet access at home (120/146, 82.2%) and owned an electronic device (131/146, 89.7%) while half of the patients had prior telemedicine knowledge (79/146, 54.1%).

To address potential selection bias, we compared the characteristics of included (146/210, 69.5%) and excluded (64/210, 30.5%) participants (Multimedia Appendix 1). The mean age was similar between excluded (60, SD 15) years and included participants (62, SD 13) years, with no significant difference (P=.46). Similar proportions of included (79/146, 54.1%) and excluded (34/64, 53%) participants had prior knowledge of telemedicine, with no significant difference (P=.98). There were no significant differences in the distance to the rheumatologist’s or general practitioner’s office between included and excluded participants (P=.28 and P=.69, respectively). Similarly, no significant differences were found in terms of electronic contact with physicians (P=.48), health status (P=.60), rheumatology treatment (P=.20), or place of residence (P=.36). A slightly higher proportion of included participants had internet access at home (120/146, 82.1% vs 47/63, 73%), but this difference was not significant (P=.21). There were no significant differences in health status documentation methods (P=.27) or sex distribution (P=.31) between the groups. Most participants in both groups owned an electronic device, with no significant difference (P=.94). Yet, a higher proportion of included participants answered yes or no regarding their wish that telemedicine services were offered by their rheumatologist (116/146, 79.4%) than excluded participants (30/63, 48%), with a higher proportion of included participants who wished for telemedicine services offered by their rheumatologist (33/146, 22.6% vs 8/63, 12%) than excluded participants (P<.001). Because we excluded participants who answered “do not know” regarding willingness to try telemedicine (Q11), there was, by construction, a significant difference between excluded and included participants regarding Q11.

XSL•FO
RenderX

Table 2. Characteristics of the study population[a].

Question	All participants (n=146)	Participants who answered yes to willingness to try telemedicine (Q11)	Participants who answered no to willingness to try telemedicine (Q11)	Chi-square (*df*)	*P* value	Cramer *V*
Want to try TM (Q11), n (%)				140 (1)	1.2×10^{-32}	0.98
Yes, n (%)	51 (34.9)	51 (100)	0 (0)			
No, n (%)	95 (65.1)	0 (0)	95 (100)			
Age, mean (SD)	62 (13)	65 (12)	57 (15)	3183[b]	7.0×10^{-4}	0.67 (0.58-0.75)[c]
Prior knowledge of TM (Q8), n (%)				8.4 (2)	.02	0.24
Yes	79 (54.1)	36 (71)	43 (45)			
No	63 (43.1)	14 (27)	49 (52)			
Do not know	3 (2.1)	1 (2)	2 (2)			
Missing answer or not answered	1 (0.68)	0 (0)	1 (1)			
Distance to rheumatologist's office (Q1), n (%)				3.6 (6)	.74	0.16
Up to 10 km	52 (35.6)	19 (3)	33 (35)			
10-20 km	27 (18.5)	8 (16)	19 (20)			
20-30 km	23 (15.7)	7 (14)	16 (17)			
30-40 km	18 (12.3)	9 (18)	9 (10)			
40-50 km	14 (9.6)	3 (6)	11 (12)			
50-60 km	3 (2.1)	1 (2)	2 (2)			
More than 60 km	8 (5.5)	3 (6)	5 (5)			
Missing answer or not answered	1 (0.68)	1 (2)	0 (0)			
General practitioner's office distance (Q2), n (%)				6.4 (5)	.26	0.21
Up to 5 km	103 (70.5)	33 (65)	70 (74)			
5-10 km	22 (15.1)	9 (18)	13 (14)			
10-15 km	11 (7.5)	4 (8)	7 (7)			
15-20 km	7 (4.8)	2 (4)	5 (5)			
20-25 km	1 (0.68)	1 (2)	0 (0)			
25-30 km	2 (1.4)	2 (4)	0 (0)			
>30 km	0 (0)	0 (0)	0 (0)			
Electronic contact with physician (Q3), n (%)				1.9 (1)	.16	0.11
No	34 (23.2)	8 (16)	26 (27)			
Yes	112 (76.7)	43 (84)	69 (73)			
Missing answer or not answered	0 (0)	0 (0)	0 (0)			
Health status (Q20), n (%)				7.5 (4)	.11	0.23
Very good	3 (2.1)	2 (4)	1 (1.1)			
Good	8 (5.5)	4 (8)	4 (4)			
Moderate	60 (41.1)	26 (51)	34 (36)			
Bad	61 (41.7)	15 (29)	46 (48)			
Very bad	9 (6.2)	2 (4)	7 (7)			

XSL•FO
RenderX

Question	All participants (n=146)	Participants who answered yes to willingness to try telemedicine (Q11)	Participants who answered no to willingness to try telemedicine (Q11)	Chi-square (*df*)	*P* value	Cramer *V*
Missing answer or not answered	5 (3.4)	2 (4)	3 (3)			
Rheumatology treatment (Q21)						
Yes	139 (95.2)	47 (92)	92 (97)	3.9 (2)	.14	0.16
No, I am a new patient	5 (3.4)	2 (4)	3 (3)			
Do not know	2 (1.4)	2 (4)	0 (0)			
Place of residence (Q23), n (%)				6.1 (3)	.11	0.21
City (>100,000 inhabitants)	35 (23.9)	9 (18)	26 (27)			
Town (20,000-100,000 inhabitants)	23 (15.7)	13 (25)	10 (11)			
Provincial town (5000-20,000 inhabitants)	46 (31.5)	16 (31)	30 (32)			
Rural area (< 5000 inhabitants)	40 (27.4)	13 (25)	27 (28)			
Missing answer or not answered	2 (1.4)	0 (0)	2 (2)			
Internet access at home (Q5), n (%)				8.9 (1)	2.8×10^{-3}	0.25
Yes	120 (82.2)	49 (96)	71 (75)			
No	26 (17.8)	2 (4)	24 (25)			
Wish of telemedicine services offered by rheumatologist (Q14), n (%)				56 (2)	8.8×10^{-13}	0.63
Yes	33 (22.6)	26 (51)	7 (7)			
No	83 (56.8)	8 (16)	75 (79)			
Do not know	23 (15.8)	12 (24)	11 (12)			
Missing answer or not answered	7 (4.8)	5 (10)	2 (2)			
Health status documentation (Q16), n (%)				0.88 (2)	.64	0.08
Yes, on paper	35 (23.9)	13 (25)	22 (23)			
Yes digitally	19 (13.0)	8 (16)	11 (12)			
No	85 (58.2)	27 (53)	58 (61)			
Missing answer or not answered	7 (4.8)	3 (6)	4 (4)			
Sex (Q18), n (%)				0 (1)	>.99	0
Male	33 (22.6)	12 (24)	21 (22)			
Female	111 (76.0)	39 (76)	72 (76)			
Missing answer or not answered	2 (1.4)	0 (0)	2 (2)			
Possession of an electronic device (Q4), n (%)				4.1 (1)	.04	0.17
Yes	131 (89.7)	50 (98)	81 (85)			
No	14 (9.6)	1 (2)	13 (14)			

XSL•FO
RenderX

Question	All participants (n=146)	Participants who answered yes to willingness to try telemedicine (Q11)	Participants who answered no to willingness to try telemedicine (Q11)	Chi-square (*df*)	*P* value	Cramer *V*
Missing answer or not answered	1 (0.68)	0 (0)	1 (1)			

[a]All variables are reported as number (percentage), except for age (mean, SD). Wilcoxon-Mann-Whitney rank sum test for age and chi-square test for count data was used to compare participants' characteristics between those who answered yes and no.

[b]*W* statistic.

[c]Vargha and Delaney *A* statistic.

Bayesian Univariate Logistic Regression Analysis

A total of 22 (40%) out of 55 variables or factors (answers to the 13 questions) were found to be positively or negatively associated (ROPE%≤5) with willingness to try telemedicine (Figure 1). After removing collinear variables (VIF>2.5), a total of 9 (40%) out of 22 variables were considered for the BMA analysis.

Figure 1. Bayesian univariate logistic regression. The percentage indicates ROPE, that is, the probability that the considered credible factor values are not negligible. The dashed lines indicate the 95% CI of the ROPE. ROPE: region of practical equivalence.

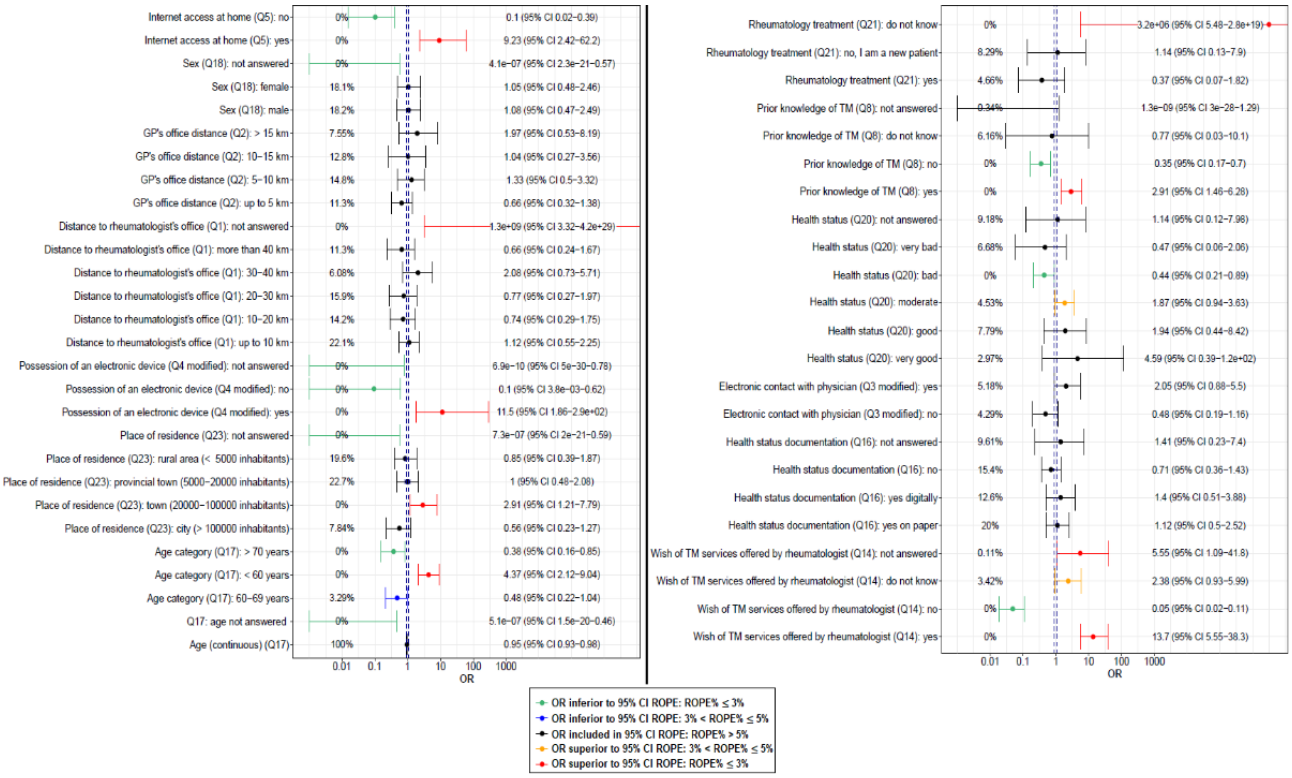

BMA Analysis

Figure 2 presents the factors identified through BMA. The value in each cell corresponds to the posterior probability that the considered variable is nonzero (in percent). The darker the color, the higher the posterior probability percentage. Cells with color from light yellow to red and the "+" sign refer to factors positively associated with willingness to try telemedicine. By contrast, cells with colors from light green to dark blue and the "–" sign refer to factors negatively associated with willingness to try telemedicine. A total of 8 determinant factors were identified. Wishing that telemedicine services were offered by a rheumatologist, possessing prior knowledge of telemedicine, living in a town and perceiving one's health status as moderate were the 4 factors positively associated with willingness to try telemedicine. By contrast, not owning an electronic device, not having home internet access, perceiving one's health status as bad and being over 60 years were the 4 factors negatively associated with willingness to try telemedicine.

XSL•FO
RenderX

Figure 2. Determinants of willingness to try telemedicine identified through Bayesian model averaging analysis.

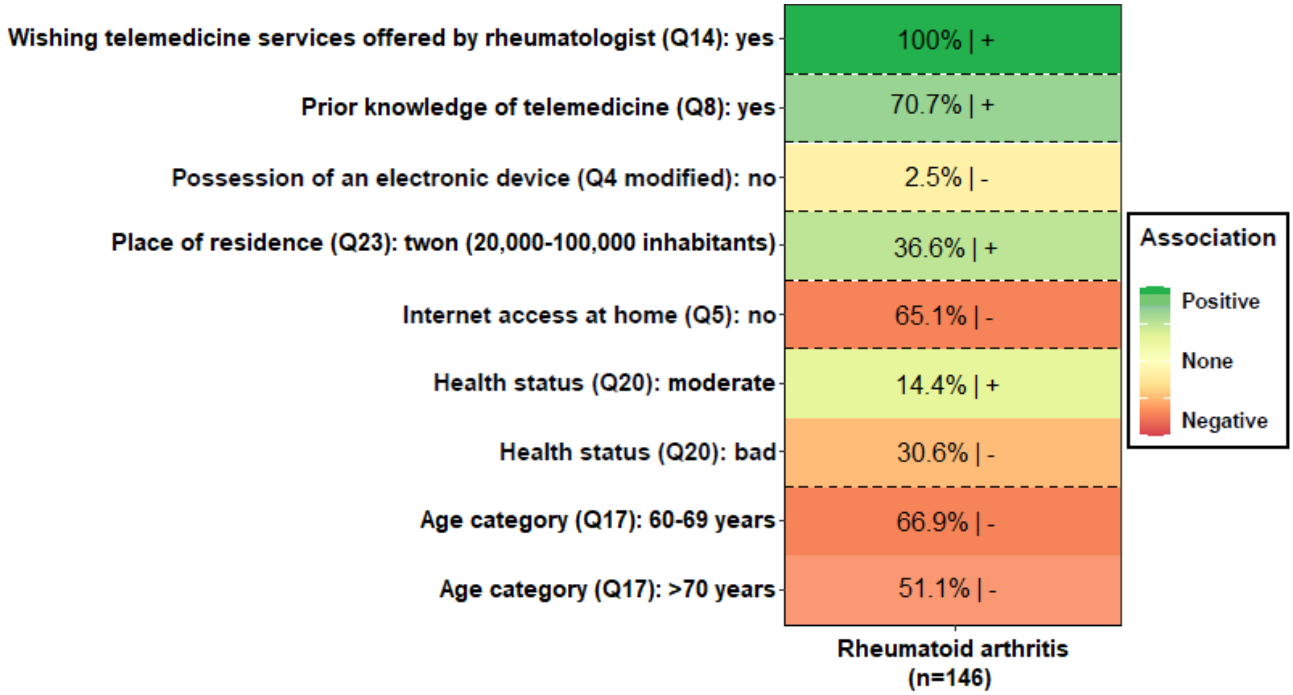

Figure 3 presents the detailed results of the BMA analysis, which tests multiple plausible or candidate models to explain willingness to try telemedicine based on the observed data. A total of 57 models were tested (x-axis), with the independent variables selected (y-axis) for each model depicted with colored cells. For a given model, green cells refer to independent variables that were positively associated with willingness to try telemedicine, while red cells indicate independent variables that are negatively associated with willingness to try telemedicine.

Figure 3. Bayesian model averaging analysis—factors selected for each candidate model tested. Green cells refer to variables that were positively associated with willingness to try telemedicine while red cells indicate variables that are negatively associated with willingness to try telemedicine.

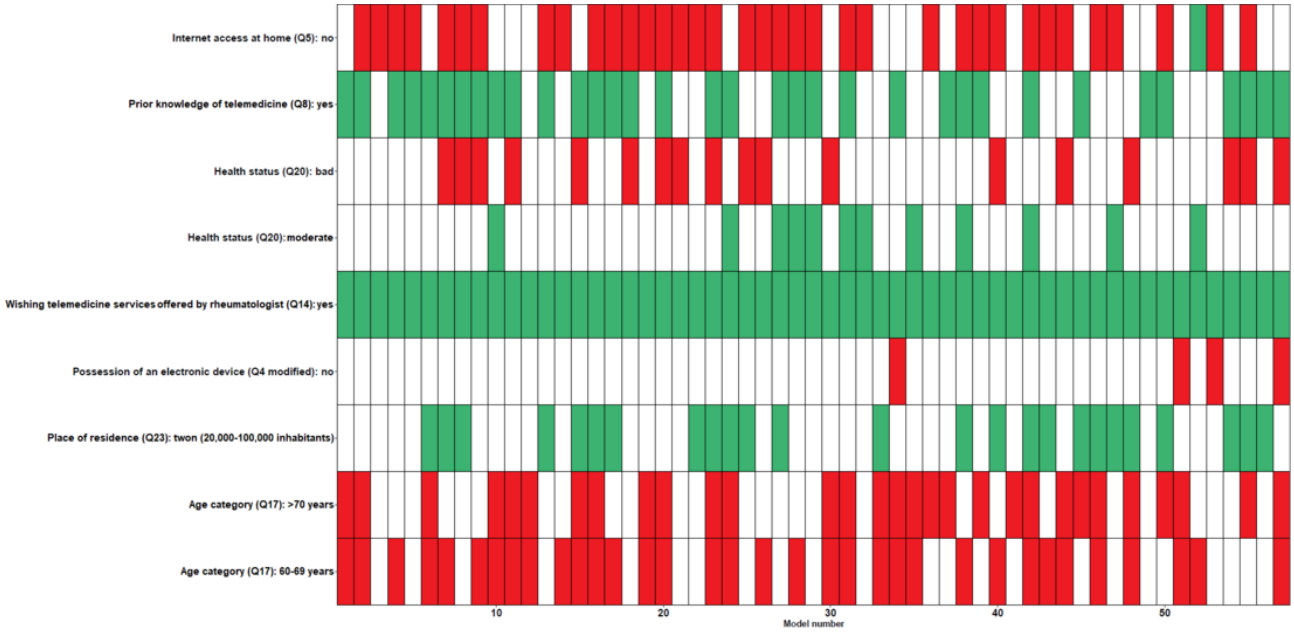

Patients' Profile

The BMA analysis delineated 2 distinct profiles of RA patients based on their motivation toward telemedicine (Figure 4). Figure 4 shows the profiles identified with patients who are motivated to try telemedicine highlighted in green and patients who do not want to try telemedicine highlighted in red.

XSL•FO
RenderX

Figure 4. Profile of patients with RA motivated to try telemedicine versus patients with RA not motivated to try telemedicine. The percentages refer to the percentage of individuals with the answer specified for each question. RA: rheumatoid arthritis.

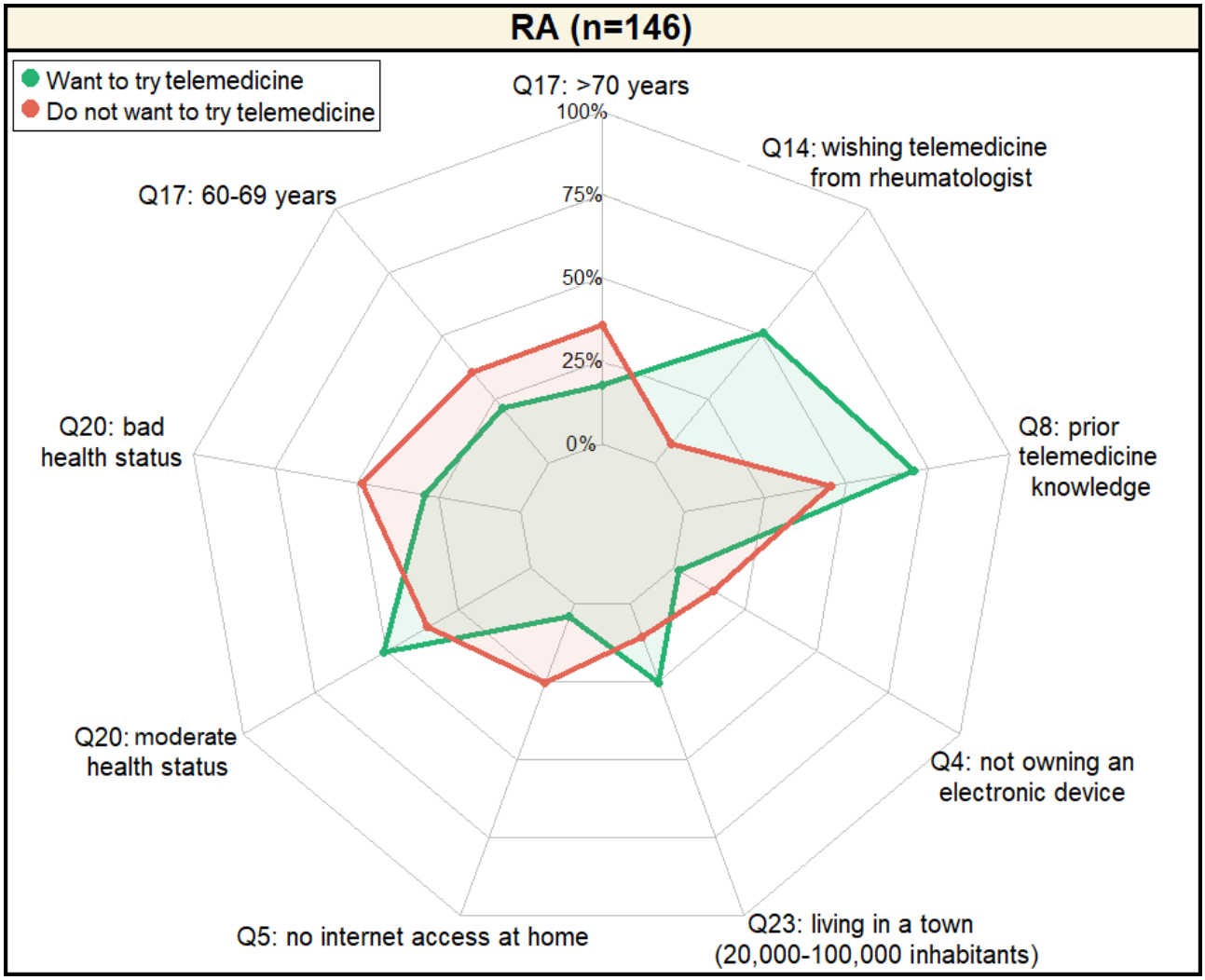

Older patients with RA, particularly those aged over 60 years, demonstrated less inclination toward telemedicine, with 72% (68/95) showing reluctance, in contrast to 39% (20/51) of the more motivated younger cohort. This motivation disparity extends to health perceptions, where 48% (46/95) of the telemedicine-reluctant group reported poor health status, compared to 29% (15/51) among those keen on telemedicine.

A closer examination of technological engagement reveals that patients with a positive attitude toward telemedicine possess a stronger foundation in digital literacy, evidenced by 70% (36/51) having prior telemedicine knowledge and 96.1% (49/51) having internet access at home. This group also boasts a higher rate of electronic device ownership (50/51, 98%) and a marked preference for living in urban areas (13/51, 25%). Interestingly, these patients perceive their health status more moderately than their counterparts.

The data underscores a strong demand among telemedicine-enthusiastic patients for the integration of telemedicine services by their rheumatologists, with 51% (26/51) expressing a wish for such offerings, a stark contrast to the 7% (7/95) among the less motivated group. This segmentation of patients with RA by telemedicine motivation offers insightful parameters for tailoring telehealth services to meet diverse patient needs and preferences effectively.

The percentage indicates the ROPE percentage, that is, the probability that the considered credible factor values are not negligible. The dashed lines indicate the 95% CI of the ROPE.

The percentages refer to the percentage of patients with RA with the answer specified for each question according to the patient motivation to try telemedicine or not (ie, patients with RA who want to try telemedicine and patients with RA who do not want to try telemedicine). For instance, 100% (the outer circular line, the farthest from the radar center) means that all patients with RA have answered the considered question with the specified answer (eg, being 60-69 years). By contrast, 0% (the inner circular line the closest to the center) means that no patient with RA chose the specified answer for the considered question (eg, being greater than 70 years). The points indicate for each question the percentage of patients with RA that chose the specified answer. Green points and lines refer to patients that wanted to try telemedicine. Red points and lines correspond to patients not wanting to try telemedicine. For each question, there were 3 possible situations. First, when the green and red points overlapped (were similar), it meant that there was no difference between patients with RA whether they were

XSL•FO
RenderX

motivated or not to use telemedicine, the proportion of similar answers was close. Second, when the green point was higher (higher percentage) than the red point, it indicates that patients with RA motivated to use telemedicine chose the specified answer more often than patients with RA not motivated to use telemedicine, which means that this factor (answer to the question) had a positive impact on willingness to try telemedicine. Finally, when the green point was lower (lower percentage) than the red point, it indicates that patients with RA motivated to use telemedicine chose the specified answer less often than patients with RA not motivated to use telemedicine, which means that this factor (answer to the question) had a negative impact on willingness to try telemedicine.

Discussion

Overview

The effective use of telemedicine in RA care offers the highest potential for improving health outcomes and self-management of patients along with relieving rheumatology workforce and reducing health care costs [27-29]. Therefore, we conducted a secondary analysis using data from a German nationwide cross-sectional survey among patients with RMD [11] to identify the factors associated with the willingness to try telemedicine, aiming to enable more effective telemedicine strategies.

Principal Results

Our results revealed that factors determining the motivation of patients with RA toward telemedicine use were multidimensional.

On one hand, patients with RA who wanted to try telemedicine more often owned an electronic device and had internet access at home. They were less likely to be younger than 60 years, perceived their health condition as moderate, had prior knowledge of traditional medicine, and more often lived in a city or town.

On the other hand, patients with RA not wanting to try telemedicine, had less access to internet at home, no prior knowledge of telemedicine and considered to have a bad health status.

Comparison With Prior Work

Only limited research findings on the effectiveness and acceptability of telemedicine approaches in rheumatology have been published prior to the beginning of the COVID-19 pandemic [7,30,31]. However, our results are consistent with previous studies: Müskens et al [27] reported that the Dutch users of a telemedicine platform were younger, more highly educated, and had better health outcomes than the total population with RA. Concurrently, our data confirms a demographic divide in the use of eHealth and mobile health: patients with RA aged less than 60 years were willing to try telemedicine, whereas those aged 60 to 80 years did not want their rheumatologists to offer them telemedicine. Furthermore, similar to the corresponding findings among physicians [32], the knowledge of telemedicine was an important determinant of telemedicine motivation among patients with RA. Our results are consistent with previous studies. Indeed, Hansen et al [33] revealed a digital divide in terms of education level in the use of eHealth platforms (apps, search engines, video services, and social media sites) among people with diabetes in Norway. Tennant et al [34] showed that being younger, using more electronic devices, and possessing a higher level of education positively influences eHealth literacy among residents of the US state of Florida. The perceived health status is also a key determinant in terms of telemedicine motivation among patients with RA. Patients, who considered their health status to be bad rarely wanted to use telemedicine, whereas patients with a moderate health status were interested in telemedicine. We hypothesize here that the worse the disease state, the higher the desire to talk face-to-face with a rheumatologist. In contrast, Ferrucci et al [35] reported that video telemedicine was more likely to be used by Alaskan patients with RA with higher disease activity. In conclusion, the relationship between disease state and telemedicine use needs to be further explored.

Complementing the demographic and health status divide, we also found signs of an economic divide in telemedicine motivation. Our results emphasize that the individual possession of technical, telemedicine-ready, equipment determines the motivation for usage. Those patients with RA who do not own technical devices will not use telemedicine. These findings are in line with previous studies for other diseases pointing to the digital divide and the danger of socioeconomic inequalities in the use of eHealth. Based on data of National Cancer Institute's 2012 Health Information National Trends Survey, Kontos et al [36] identified that that lower socioeconomic status, older, and male US adult internet-users were less likely to engage in a number of eHealth activities, such as using the internet to look for a health care provider, using email or the internet to communicate with a doctor, using web-based tracking of their personal health information, using a website to help track diet, weight, and physical activity, or downloading health information to a smartphone device. Latulippe et al [37] recently performed a literature review with the result that ethnicity and low income are the most commonly used characteristics to identify people at risk of social health inequality, which might be reduced via eHealth by aiming for universal access to the tool of eHealth, considering of users' literacy level, creating eHealth tools that respect the cultural attributes of future users, and encouraging the participation of people at risk of social health inequalities. Since educational level, socioeconomic status, ethnicity, and income were not assessed in the original survey, there is an urgent need for further research to gain a better understanding of access to telemedicine in rheumatology.

Implications

In the realm of rheumatology, the implementation of telemedicine presents both opportunities and challenges. Our findings underscore the necessity of ensuring that all patients with RA can access telemedicine, highlighting the importance of addressing barriers related to age, geographical location, internet connectivity, and the availability of electronic devices. To circumvent these obstacles, targeted support and educational initiatives are imperative. Programs aimed at patient education, bolstered by the active involvement of patient associations, have been identified as crucial factors in facilitating the adoption of telehealth among patients with RA [38]. Such measures not

XSL•FO
RenderX

only promote digital literacy but also ensure that patients are not excluded from receiving comprehensive RA care due to technological disparities.

The clinical practice landscape is similarly affected, with physicians identified as key facilitators in the adoption of telemedicine [39]. However, the uptake of telemedicine by health care professional is hindered by challenges mirroring those faced by patients, including a lack of familiarity with telemedicine, demographic factors such as age, and the geographical context of their practice [40]. Overcoming these barriers is essential to broaden the use of telemedicine in clinical settings. Proposals such as the integration of assisted telemedicine services in pharmacies or health kiosks, particularly in rural and underserved regions, are promising strategies that align with broader digitalization efforts in Germany [41]. These initiatives can play a pivotal role in enhancing access to specialized rheumatology care, ensuring that digital health resources extend their reach effectively.

Moreover, the evolution of medical teaching to incorporate telemedicine knowledge is critical. The education of health care professionals must prioritize digital literacy, ensuring that future clinicians are proficient in using digital tools for patient care. This includes a comprehensive understanding of how to navigate the challenges and opportunities presented by telemedicine. Furthermore, as telemedicine becomes increasingly integrated into standard health care delivery, it is paramount to maintain the option of high-quality, traditional, and analogue health care. This approach necessitates the implementation of a triage mechanism [42] to match patients with the most suitable form of care, whether digital or traditional, tailored to their preferences and circumstances.

Limitations

The primary data on which this analysis is based was collected until December 30, 2019, that is, shortly prior to SARS-CoV-2 outbreak in Germany (January 27, 2020). Due to the need to reduce physical contact and thus minimize the risk of infection, usage of telemedicine initially received a major uptake in global health care delivery [43]. Hence, more patients with RMDs and likely other subgroups will have tried telemedicine by now [44]. A replication of the initial survey is essential to identify whether and how the identified factors have changed. Apart from this, the limitations of the primary data still apply [10]. These are primarily the high potential of self-selection and nonresponse bias. These sources of bias are due to the fact that the primary data collection was a questionnaire survey in which participation was voluntary. These biases may have contributed to a disproportionate representation of technically proficient and interested patients in the survey population, potentially leading to a more favorable attitude toward telemedicine. Additionally, the study presents a selection bias because participants who either did not answer Q11 (“Would you like to try telemedicine?”) or answered “do not know” were not included in our analysis, possibly excluding those who are not familiar with the term “telemedicine.” However, this selection bias should be limited because the characteristics of the excluded individuals were similar to those of the included participants (Multimedia Appendix 1). To mitigate this limited selection bias, a multinomial regression analysis, considering all participants, regardless of their answer, could have been performed. Furthermore, it is important to emphasize that there is a considerable difference between the inclination expressed in a survey and the actual use of telemedicine by patients with RMD. Our data must be validated against real-world data from clinical practice in RA care.

Regarding statistical analyses, we used a Bayesian approach to conduct the secondary analysis of the earlier survey. A practical limitation of the Bayesian approach is that it requires the specification of prior distributions both on parameters of each model and on the distribution of models themselves. As we had no a priori assumption, we used weakly informative priors. Choosing another prior distribution may have substantial influence on the outcome [45,46]. Regarding variable selection, a widespread approach consisting of including significant variables from univariate analysis in a multivariate analysis was carried out [47,48]. To be more accurate, all the individual variables associated (positively or negatively) with willingness to try telemedicine in the Bayesian univariate analysis were selected based on the ROPE percentage (ROPE %≤5). A ROPE-only decision rule was used as suggested in other works [16-18]. Choosing another ROPE percentage threshold may have yielded different results. Then, we performed a conservative selection based on the VIF (≤2.5) to deal with potential variable multicollinearity. Finally, we used the remaining variables with BMA for model selection and identification of determinants. BMA was chosen in particular because it reduces overconfidence and is relatively robust against model misspecification [46,49-51]. The Markov chain Monte Carlo model was used to deal with the intractable computational challenge of BMA that comes from the candidate model enumeration [52].

Conclusions

Our results suggest that health status, telemedicine knowledge, age, and access to technical equipment and infrastructure influence the motivation of patients with RA to use telehealth. In particular, older patients with RA with bad health status, are currently not motivated to use telemedicine and, thus, might be left out from digital transition in rheumatology care.

Acknowledgments

The authors would like to thank the survey participants, their teams, and all other supporters of TeleRheumaBB. They also owe special gratitude to KV Consult- und Managementgesellschaft mbH, which initiated the study in the first place.

This study is part of the PhD thesis of FM (AGEIS, Université Grenoble Alpes, Grenoble, France). This work forms a part of a broader transnational and interdisciplinary collaboration between Université Grenoble Alpes (France), Philipps-University Marburg (Germany), and Brandenburg Medical School (Germany). This work was supported by the French National Research Agency

XSL•FO
RenderX

(France) under the framework of the Investissements d'avenir program (ANR-10-AIRT-05 and ANR-15-IDEX-02) and by MIAI@Grenoble Alpes (ANR-19-P3IA-0003). The sponsors were not involved in the review or approval of the manuscript for publication.

During the preparation of this work, the authors used ChatGPT (March 2023 version) to improve language and readability. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in drafting the article and critically revising it for important intellectual content, and all authors approved the final version to be submitted for publication. FM had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

FM contributed to the study conception and design, acquired data, and was involved in the analysis and interpretation of data. PP contributed to the study conception and design, as well as the analysis and interpretation of data. JK participated in the study conception and design and contributed to the analysis and interpretation of data. MW assisted in the acquisition of data. NV was involved in the study conception and design, data acquisition, and analysis and interpretation of data. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparison of the characteristics (continuous variable) of the included and excluded participants.

[DOCX File , 23 KB - jmir_v26i1e47733_app1.docx]

References

1. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet 2016;388(10055):2023-2038 [FREE Full text] [doi: 10.1016/S0140-6736(16)30173-8] [Medline: 27156434]
2. Safiri S, Kolahi AA, Hoy D, Smith E, Bettampadi D, Mansournia MA, et al. Global, regional and national burden of rheumatoid arthritis 1990-2017: a systematic analysis of the Global Burden of Disease study 2017. Ann Rheum Dis 2019;78(11):1463-1471. [doi: 10.1136/annrheumdis-2019-215920] [Medline: 31511227]
3. Hresko A, Lin T, Solomon DH. Medical care costs associated with rheumatoid arthritis in the US: a systematic literature review and meta-analysis. Arthritis Care Res 2018;70(10):1431-1438 [FREE Full text] [doi: 10.1002/acr.23512] [Medline: 29316377]
4. Malm K, Bergman S, Andersson ML, Bremander A, Larsson I. Quality of life in patients with established rheumatoid arthritis: a phenomenographic study. SAGE Open Med 2017;5:2050312117713647 [FREE Full text] [doi: 10.1177/2050312117713647] [Medline: 28611920]
5. Lennep DS, Crout T, Majithia V. Rural health issues in rheumatology: a review. Curr Opin Rheumatol 2020;32(2):119-125. [doi: 10.1097/BOR.0000000000000694] [Medline: 31913162]
6. Piga M, Cangemi I, Mathieu A, Cauli A. Telemedicine for patients with rheumatic diseases: systematic review and proposal for research agenda. Semin Arthritis Rheum 2017;47(1):121-128. [doi: 10.1016/j.semarthrit.2017.03.014] [Medline: 28420491]
7. McDougall JA, Ferucci ED, Glover J, Fraenkel L. Telerheumatology: a systematic review. Arthritis Care Res 2017;69(10):1546-1557 [FREE Full text] [doi: 10.1002/acr.23153] [Medline: 27863164]
8. Miloslavsky EM, Bolster MB. Addressing the rheumatology workforce shortage: a multifaceted approach. Semin Arthritis Rheum 2020;50(4):791-796 [FREE Full text] [doi: 10.1016/j.semarthrit.2020.05.009] [Medline: 32540672]
9. Muehlensiepen F, Knitza J, Marquardt W, Engler J, Hueber A, Welcker M. Acceptance of telerheumatology by rheumatologists and general practitioners in Germany: nationwide cross-sectional survey study. J Med Internet Res 2021;23(3):e23742 [FREE Full text] [doi: 10.2196/23742] [Medline: 33690147]
10. Muehlensiepen F, Knitza J, Marquardt W, May S, Krusche M, Hueber A, et al. Opportunities and barriers of telemedicine in rheumatology: a participatory, mixed-methods study. Int J Environ Res Public Health 2021;18(24):13127 [FREE Full text] [doi: 10.3390/ijerph182413127] [Medline: 34948737]
11. Muehlensiepen F, Petit P, Knitza J, Welcker M, Vuillerme N. Factors associated with telemedicine use among patients with rheumatic and musculoskeletal disease: secondary analysis of data from a German nationwide survey. J Med Internet Res 2023;25:e40912 [FREE Full text] [doi: 10.2196/40912] [Medline: 36705950]

XSL•FO
RenderX

12. Cohen J. A power primer. Psychol Bull 1992;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
13. Champely S. pwr: basic functions for power analysis. The R Project for Statistical Computing. 2020. URL: https://CRAN.R-project.org/package=pwr [accessed 2024-06-25]
14. Dey R, Mulekar M. Effect size as a measure of difference between two populations. In: Alhajj R, Rokne J, editors. Encyclopedia of Social Network Analysis and Mining. New York, NY: Springer; 2018.
15. Vargha A, Delaney HD. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. J Educ Behav Stat 2016;25(2):101-132. [doi: 10.3102/10769986025002101]
16. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. J Clin Epidemiol 2020;125:183-187. [doi: 10.1016/j.jclinepi.2020.03.028] [Medline: 32540389]
17. Kruschke J. Rejecting or accepting parameter values in Bayesian estimation. Adv Methods Pract Psychol Sci 2018;1(2):270-280. [doi: 10.1177/2515245918771304]
18. Makowski D, Ben-Shachar MS, Chen SHA, Lüdecke D. Indices of effect existence and significance in the Bayesian framework. Front Psychol 2019;10:2767 [FREE Full text] [doi: 10.3389/fpsyg.2019.02767] [Medline: 31920819]
19. Schwaferts P, Augustin T. Bayesian decisions using regions of practical equivalence (ROPE): foundations. University of Munich. URL: https://epub.ub.uni-muenchen.de/74222/1/TR%20ROPE%2BHDI%20SUBMITTED.pdf [accessed 2022-03-30]
20. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. J Interdisc Math 2010;13(3):253-267. [doi: 10.1080/09720502.2010.10700699]
21. Zuur AF, Iono EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 2010;1(1):3-14.
22. Depaoli S, Lai K, Yang Y. Bayesian model averaging as an alternative to model selection for multilevel models. Multivariate Behav Res 2021;56(6):920-940. [doi: 10.1080/00273171.2020.1778439] [Medline: 32619364]
23. Brilleman S, Crowther M, Moreno-Betancur M, Buros NJ, Wolfe R. StanCon—joint longitudinal and time-to-event models via Stan. GitHub. URL: https://github.com/stan-dev/stancon_talks/ [accessed 2023-03-30]
24. Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: bayesian applied regression modeling via stan. The R Project for Statistical Computing. URL: https://mc-stan.org/rstanarm [accessed 2023-03-30]
25. Raftery A, Hoeting J, Volinsky C, Painter I, Yeung K. BMA: bayesian model averaging. Package for Bayesian model averaging and variable selection for linear models, generalized linear models and survival models (cox regression). URL: https://github.com/hanase/BMA [accessed 2023-03-30]
26. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04 [accessed 2023-03-30]
27. Müskens WD, Rongen-van Dartel SAA, Vogel C, Huis A, Adang E, van Riel PLCM. Telemedicine in the management of rheumatoid arthritis: maintaining disease control with less health-care utilization. Rheumatol Adv Pract 2021;5(1):rkaa079 [FREE Full text] [doi: 10.1093/rap/rkaa079] [Medline: 33688619]
28. Ma Y, Zhao C, Zhao Y, Lu J, Jiang H, Cao Y, et al. Telemedicine application in patients with chronic disease: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022;22(1):105 [FREE Full text] [doi: 10.1186/s12911-022-01845-2] [Medline: 35440082]
29. Bernard L, Valsecchi V, Mura T, Aouinti S, Padern G, Ferreira R, et al. Management of patients with rheumatoid arthritis by telemedicine: connected monitoring. A randomized controlled trial. Joint Bone Spine 2022;89(5):105368. [doi: 10.1016/j.jbspin.2022.105368] [Medline: 35248737]
30. Taylor PC. The opportunities and challenges of telemedicine in the management of people with rheumatoid arthritis. J Rheumatol 2022;49(11):1188-1190. [doi: 10.3899/jrheum.220655] [Medline: 36318992]
31. Jackson LE, Edgil TA, Hill B, Owensby JK, Smith CH, Singh JA, et al. Telemedicine in rheumatology care: a systematic review. Semin Arthritis Rheum 2022;56:152045. [doi: 10.1016/j.semarthrit.2022.152045] [Medline: 35843158]
32. Muehlensiepen F, Petit P, Knitza J, Welcker M, Vuillerme N. Factors associated with telemedicine use among German general practitioners and rheumatologists: secondary analysis of data from a nationwide survey. J Med Internet Res 2022;24(11):e40304 [FREE Full text] [doi: 10.2196/40304] [Medline: 36449333]
33. Hansen AH, Bradway M, Broz J, Claudi T, Henriksen Ø, Wangberg SC, et al. Inequalities in the use of eHealth between socioeconomic groups among patients with type 1 and type 2 diabetes: cross-sectional study. J Med Internet Res 2019;21(5):e13615 [FREE Full text] [doi: 10.2196/13615] [Medline: 31144669]
34. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
35. Ferucci ED, Holck P, Day GM, Choromanski TL, Freeman SL. Factors associated with use of telemedicine for follow-up of rheumatoid arthritis. Arthritis Care Res 2020;72(10):1404-1409 [FREE Full text] [doi: 10.1002/acr.24049] [Medline: 31421020]

XSL•FO
RenderX

36. Kontos E, Blake KD, Chou WS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
37. Latulippe K, Hamel C, Giroux D. Social health inequalities and eHealth: a literature review with qualitative synthesis of theoretical and empirical studies. J Med Internet Res 2017;19(4):e136 [FREE Full text] [doi: 10.2196/jmir.6731] [Medline: 28450271]
38. Mühlensiepen F, Kurkowski S, Krusche M, Mucke J, Prill R, Heinze M, et al. Digital health transition in rheumatology: a qualitative study. Int J Environ Res Public Health 2021;18(5):2636 [FREE Full text] [doi: 10.3390/ijerph18052636] [Medline: 33807952]
39. Magnol M, Eleonore B, Claire R, Castagne B, Pugibet M, Lukas C, et al. Use of eHealth by patients with rheumatoid arthritis: observational, cross-sectional, multicenter study. J Med Internet Res 2021;23(1):e19998 [FREE Full text] [doi: 10.2196/19998] [Medline: 33512320]
40. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. There's an app for that, but nobody's using it: insights on improving patient access and adherence to digital therapeutics in Germany. Digit Health 2022;8:20552076221104672 [FREE Full text] [doi: 10.1177/20552076221104672] [Medline: 35811758]
41. [Digitalization strategy for healthcare and nursing. Digital together]. Federal Ministry of Health. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/D/Digitalisierungsstrategie/BMG_Broschuere_Digitalisierungsstrategie_bf.pdf [accessed 2023-03-30]
42. Kulcsar Z, Albert D, Ercolano E, Mecchella JN. Telerheumatology: a technology appropriate for virtually all. Semin Arthritis Rheum 2016;46(3):380-385. [doi: 10.1016/j.semarthrit.2016.05.013] [Medline: 27395561]
43. Omboni S, Padwal RS, Alessa T, Benczúr B, Green BB, Hubbard I, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022;1:7-35 [FREE Full text] [doi: 10.20517/ch.2021.03] [Medline: 35233563]
44. Peine A, Paffenholz P, Martin L, Dohmen S, Marx G, Loosen SH. Telemedicine in Germany during the COVID-19 pandemic: multi-professional national survey. J Med Internet Res 2020;22(8):e19745 [FREE Full text] [doi: 10.2196/19745] [Medline: 32568724]
45. Wagenmakers E, Marsman M, Jamil T, Ly A, Verhagen J, Love J, et al. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. Psychon Bull Rev 2018;25(1):35-57 [FREE Full text] [doi: 10.3758/s13423-017-1343-3] [Medline: 28779455]
46. Hinne M, Gronau QF, van den Bergh D, Wagenmakers E. A conceptual introduction to Bayesian model averaging. Adv Methods Pract Psychol Sci 2020;3(2):200-215. [doi: 10.1177/2515245919898657]
47. Alpoim PN, Godoi LC, Pinheiro MDB, Freitas LG, Carvalho MDG, Dusse LM. The unexpected beneficial role of smoking in preeclampsia. Clin Chim Acta 2016;459:105-108. [doi: 10.1016/j.cca.2016.05.030] [Medline: 27259465]
48. Dabi Y, El Mrini M, Duquesnes I, Delongchamps NB, Sibony M, Zerbib M, et al. Impact of body mass index on the oncological outcomes of patients treated with radical nephroureterectomy for upper tract urothelial carcinoma. World J Urol 2018;36(1):65-71. [doi: 10.1007/s00345-017-2095-4] [Medline: 29032451]
49. Genell A, Nemes S, Steineck G, Dickman PW. Model selection in medical research: a simulation study comparing Bayesian model averaging and stepwise regression. BMC Med Res Methodol 2010;10:108 [FREE Full text] [doi: 10.1186/1471-2288-10-108] [Medline: 21134252]
50. Mu Y, See I, Edwards JR. Bayesian model averaging: improved variable selection for matched case-control studies. Epidemiol Biostat Public Health 2019;16(2):e13048 [FREE Full text] [doi: 10.2427/13048] [Medline: 31772926]
51. Wang D, Zhang W, Bakhai A. Comparison of bayesian model averaging and stepwise methods for model selection in logistic regression. Stat Med 2004;23(22):3451-3467. [doi: 10.1002/sim.1930] [Medline: 15505893]
52. Lu Z, Lou W. Bayesian approaches to variable selection: a comparative study from practical perspectives. Int J Biostat 2021;18(1):83-108. [doi: 10.1515/ijb-2020-0130] [Medline: 33761580]

Abbreviations

BMA: Bayesian model averaging
OR: odds ratio
RA: rheumatoid arthritis
RMD: rheumatic and musculoskeletal disease
ROPE: region of practical equivalence
VIF: variance inflation factor

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 30.03.23; peer-reviewed by R Knudsen, L Gossec; comments to author 30.01.24; revised version received 16.02.24; accepted 26.06.24; published 19.08.24.

Please cite as:
Muehlensiepen F, Petit P, Knitza J, Welcker M, Vuillerme N
Identification of Motivational Determinants for Telemedicine Use Among Patients With Rheumatoid Arthritis in Germany: Secondary Analysis of Data From a Nationwide Cross-Sectional Survey Study
J Med Internet Res 2024;26:e47733
URL: https://www.jmir.org/2024/1/e47733
doi:10.2196/47733
PMID:

©Felix Muehlensiepen, Pascal Petit, Johannes Knitza, Martin Welcker, Nicolas Vuillerme. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Health Care Usage During the COVID-19 Pandemic and the Adoption of Telemedicine: Retrospective Study of Chronic Disease Cohorts

Margrét Vilborg Bjarnadóttir[1*], PhD; David Anderson[2*], PhD; Kelley M Anderson[3], PhD, FNP; Omar Aljwfi[4], PhD; Alina Peluso[5], PhD; Adam Ghannoum[6]; Gayle Balba[7], MD; Nawar Shara[4,8], PhD

[1]Decisions, Operations and Information Technology, University of Maryland, College Park, College Park, MD, United States
[2]Villanova School of Business, Villanova, PA, United States
[3]School of Nursing, Georgetown University, Washington, DC, United States
[4]MedStar Health Research Institute, Hyattsville, MD, United States
[5]Oak Ridge National Laboratory, Oak Ridge, TN, United States
[6]University of Maryland, College Park, College Park, MD, United States
[7]Department of Infectious Diseases, MedStar Georgetown University Hospital, Washington, DC, United States
[8]Department of Endocrinology, MedStar Georgetown University Hospital, Washington, DC, United States
[*]these authors contributed equally

Corresponding Author:
Margrét Vilborg Bjarnadóttir, PhD
Decisions, Operations and Information Technology
University of Maryland, College Park
4353 Van Munching Hall
College Park, MD, 20742
United States
Phone: 1 301 405 2061
Email: mbjarnad@umd.edu

Abstract

Background: The COVID-19 pandemic accelerated telehealth adoption across disease cohorts of patients. For many patients, routine medical care was no longer an option, and others chose not to visit medical offices in order to minimize COVID-19 exposure. In this study, we take a comprehensive multidisease approach in studying the impact of the COVID-19 pandemic on health care usage and the adoption of telemedicine through the first 12 months of the COVID-19 pandemic.

Objective: We studied the impact of the COVID-19 pandemic on in-person health care usage and telehealth adoption across chronic diseases to understand differences in telehealth adoption across disease cohorts and patient demographics (such as the Social Vulnerability Index [SVI]).

Methods: We conducted a retrospective cohort study of 6 different disease cohorts (anxiety: n=67,578; depression: n=45,570; diabetes: n=81,885; kidney failure: n=29,284; heart failure: n=21,152; and cancer: n=35,460). We used summary statistics to characterize changes in usage and regression analysis to study how patient characteristics relate to in-person health care and telehealth adoption and usage during the first 12 months of the pandemic.

Results: We observed a reduction in in-person health care usage across disease cohorts (ranging from 10% to 24%). For most diseases we study, telehealth appointments offset the reduction in in-person visits. Furthermore, for anxiety and depression, the increase in telehealth usage exceeds the reduction in in-person visits (by up to 5%). We observed that younger patients and men have higher telehealth usage after accounting for other covariates. Patients from higher SVI areas are less likely to use telehealth; however, if they do, they have a higher number of telehealth visits, after accounting for other covariates.

Conclusions: The COVID-19 pandemic affected health care usage across diseases, and the role of telehealth in replacing in-person visits varies by disease cohort. Understanding these differences can inform current practices and provides opportunities to further guide modalities of in-person and telehealth visits. Critically, further study is needed to understand barriers to telehealth service usage for patients in higher SVI areas. A better understanding of the role of social determinants of health may lead to more support for patients and help individual health care providers improve access to care for patients with chronic conditions.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e54991) doi:10.2196/54991

KEYWORDS

telehealth utilization; health care utilization; demographic differences; cohort study; telehealth; COVID-19; telehealth adaption

Introduction

The COVID-19 pandemic exerted both direct and indirect impacts on individual health and well-being. Prior to March 2020, routine clinic visits were performed as preventative measures to identify conditions that may result in severe diseases and as a component of chronic condition management. When COVID-19 lockdowns were instituted across the United States, many clinical offices and laboratories closed. Routine medical care was no longer an option for many patients, and in other cases patients chose not to visit medical offices in order to minimize COVID-19 exposure.

Across studies, the median reduction in health care usage during the pandemic (across visits, admissions, diagnostics, and therapeutics) was 37%, with healthier individuals being more likely to reduce their use of health care services [1]. Many patients with chronic disease experienced detrimental effects: the literature documents a significant decrease in the treatment and detection of chronic disease [2,3], as well as adverse effects on chronic disease management as patients reduced their usage and experienced barriers in access to care. For example, New York City witnessed a 139% increase in deaths related to heart disease during the pandemic [4]. Experts attribute this increase in mortality to a decrease in diagnostic procedures as patients were unable to access facilities and as those facilities experienced an overwhelming influx of patients with COVID-19. Cancer rates also increased during the pandemic. For instance, due to COVID-19 lockdowns, routine breast cancer screening was suspended throughout 2020, and it is anticipated that this delay in detection will lead to an estimated 7.9%-9.6% short-term increase in deaths due to breast cancer [5].

Beyond the numerous studies that have evaluated health care outcomes, there is a growing body of studies about health care resource usage during and post the pandemic [6-8] and about the widespread adoption of telehealth. Telehealth, that is, the use of audio and video consultations, is time-efficient, socially distanced, and low cost [9]. Therefore, unsurprisingly, the use of telehealth surged during the pandemic, as many of the barriers to its adoption were removed to support clinical care [10]. Telehealth was used by patients who were immunocompromised and by others wary of possible COVID-19 infection or affected by lockdowns [9].

While telemedicine consultations are not suitable for all patients, the pandemic has permanently changed telemedicine usage patterns, and the use of telemedicine may well become a new standard of care, although patient and health care providers' satisfaction and views are mixed [11-16]. Therefore, disparities in the adoption of telehealth are an important consideration. Such disparities have been documented in a number of studies, which show that overall, there is less adoption among older patients [17,18], patients of lower socioeconomic status [17-21], minorities [18,20,21] and men [19,20]. These disparities are driven by lack of access to the necessary technology, particularly for patients of lower socioeconomic status and older patients [19].

In this study, we take a comprehensive approach to study health care usage and the adoption of telemedicine through the first 12 months of the COVID-19 pandemic. We selected conditions that are common with high health care usage in the categories of mental health (depression and anxiety), cardiovascular disease (heart failure), cancer, renal disease (kidney failure), and endocrine disorder (diabetes) [22-24]. We also evaluated the impact of specific social determinants of health indicators on health care usage in general and telehealth adoption in particular. Importantly, this study allows us to compare the usage and adoption of telehealth across chronic disease cohorts in the context of health care disparities. We aim to contribute to our understanding of how various patient populations used telemedicine during the pandemic, which can inform improved access to care for vulnerable groups.

Methods

Cohort Definition

Information about diagnosis and usage was extracted from the electronic health records of a regional health care system serving Virginia, Maryland, and the District of Columbia. For each of our study diseases (anxiety, depression, diabetes, kidney failure, heart failure, and cancer), we defined a disease cohort of patients to include any patient diagnosed with the disease before October 2019. The corresponding *ICD-9* (*International Classification of Diseases, Ninth Revision*) and *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes are provided in Section A of Multimedia Appendix 1. Since a single patient experiencing multiple chronic diseases may be in more than 1 disease cohort, we conducted a robustness check that independently repeated the analysis only on those members who are in a single disease cohort. The results are materially the same and are included in Section B of Multimedia Appendix 1.

We then evaluated members' health care usage in 6 months prior to the pandemic lockdowns, from October 2019 through March 2020. We next compared usage in this period with usage during a 12-month pandemic period, from June 2020 through May 2021. Note that we excluded April and May 2020 from the study, as health care services in the system under study were significantly disrupted during that period.

Demographic, Socioeconomic, and Clinical Variables

We extracted each patient's age, gender, race, and county from the electronic health records data. For gender, due to low numbers in the nonbinary and not reported categories, we grouped all patients in those categories together. Similarly, and also due to low numbers, we grouped together all patients not identified as Black or White. We extracted the Federal

XSL•FO
RenderX

Information Processing Standards county-level Social Vulnerability Index (SVI) from the Centers for Disease Control and Prevention [25] and matched it with the patient's zip code. Finally, we calculated each member's Charlson Comorbidity Index, based on the patient's visits from January 2018 to the end of the prepandemic period. As we are explicitly using age in our models, the age-free version was calculated [26].

Usage

Before the pandemic telehealth services were available, primarily through an eVisit app, telehealth services were available across various ambulatory sites, offering options such as video visits, audio-only phone visits, and eConsults. Despite the availability, usage remained modest. With the pandemic onset there was a significant surge in telehealth demand, peaking at more than 400 visits daily. In response, strategic adaptations were made, including temporarily waiving patient fees until June 30, 2021. In this study, we define a telehealth visit as both video and audio visits. We defined in-person usage as the number of inpatient, outpatient, and emergency department visits in a given period for any reason (not disease-specific). We then defined overall usage as the sum of in-person usage and the number of telehealth visits.

Statistical Analysis

We conducted a longitudinal cohort study comparing usage using summary statistics and regression analysis. First, for each disease we compared the in-person and overall usage in the prepandemic period with usage during the pandemic time period (broken into two 6-month subperiods to highlight temporal patterns).

Then, for each disease, we constructed a regression model that explains the number of pandemic period in-person visits as a function of the usage in the prepandemic period and patients' demographic, socioeconomic, and clinical characteristics. Because the distribution of in-person visits is left-skewed, we transformed the dependent variable to ensure statistical fit. Specifically, we fit a regression model with the natural logarithm of the pandemic period in-person visits as the dependent variable and in-person visits in the prepandemic period, Charlson Index score, race, age, gender, and county average SVI as the independent variables.

Finally, we ran a regression model on the number of telehealth visits. Since only one-third of patients have a telehealth visit in their data, we fit a zero-inflated negative binomial regression model [27]. The zero-inflated model is a 2-step regression model, and in the first step a logistic regression model is fit to explain which patients have no telehealth appointment. In the second step, for the patients who have a telehealth appointment, a negative binomial regression model is fit (to best fit the distribution in the number of telehealth visits). We use the same independent variables: in-person visits in the prepandemic period, Charlson Index score, race, age, gender, and county average SVI. We finally highlight the adoption patterns of telehealth, broken down by race and SVI.

Ethical Considerations

The original data collection and the study protocol were reviewed by the Institutional Review Board of Georgetown-Howard Universities Center for Clinical and Translational Science (approval ID: STUDY00003813). The analysis followed the approved protocol and the authors therefore had the permission to use the data for this study. The data were deidentified and the study was an except retrospective study; therefore, consent was not sought and no compensation was provided.

Results

Cohort Characteristics

Table 1 summarizes the cohort characteristics for the different disease groups. The average age is lowest for the anxiety and depression cohorts and highest for those experiencing kidney failure and heart failure. Interestingly, while men constitute almost half of the kidney failure and heart failure cohorts, they constitute close to 45% (15,922/35,460 and 36,930/81,885, respectively) of patients with cancer and diabetes and close to 30% (19,733/67,578 and 13,808/45,570, respectively) of those with anxiety and depression. We also observe variation by race: White patients are the majority in the anxiety, depression, and cancer cohorts, while Black patients are the majority in the diabetes, kidney failure, and heart failure cohorts. These differences reflect both expected variation by disease (ie, typically older patients have kidney and heart failure) and the demographics served by the health care system we study. Across disease cohorts, between 22% (14,597/67,578 for anxiety) and 33% (27,104/81,885 for diabetes) of patients live in counties with high SVI. The highest percentage of patients living in high-SVI counties are patients in the diabetes, kidney failure, and heart failure cohorts. Between 24% (6940/29,284 for kidney failure and 5055/21,152 for heart failure) and 32% (11,117/35,460 for cancer) of patients live in counties with low SVI.

XSL•FO
RenderX

Table 1. Summary statistics for the 6 disease cohorts.

	Anxiety	Depression	Diabetes	Kidney failure	Heart failure	Cancer
Cohort size, n	67,578	45,570	81,885	29,284	21,152	35,460
Age (years), mean	51.9	54.4	65.5	70.9	71.3	68.2
Charlson Index[a]	1.00	1.28	2.37	2.63	3.54	3.48
Sex, %						
Female	70.7	69.7	54.9	48.9	50.3	55.1
Male	29.2	30.3	45.1	51.1	49.7	44.9
Other	0.01	0.01	0.004	—[b]	—	—
Race, %						
Black	30.3	35.0	53.4	58.2	54.3	35.3
Other	8.1	7.3	9.0	6.8	4.9	9.2
White	61.6	57.6	37.6	35.1	40.8	55.5
SVI[c], %						
High	21.6	24.4	33.1	32.2	32.6	22.1
Medium	46.2	47.0	41.0	43.4	42.7	45.8
Low	31.3	27.9	25.2	23.7	23.9	31.5
Unknown	0.8	0.7	0.7	0.7	0.8	0.7
Number of visits[d]						
Pre–COVID-19	4.3	4.7	4.5	5.9	6.2	5.4
During COVID-19	3.8	4.0	4.0	4.8	4.8	4.3
Any telehealth (%)[e]	35.0	33.4	30.3	31.1	27.7	32.1

[a]Without age.

[b]Not applicable.

[c]SVI: Social Vulnerability Index.

[d]Average number of visits in a 6-month period.

[e]In the pandemic period.

Overall Usage

Figure 1A summarizes the in-person health care usage for the different disease cohorts. In order to capture temporal patterns, we break the 12-month pandemic period into two 6-month subperiods (with June 2020 through November 2020 as pandemic subperiod 1 and December 2020 through May 2021 as pandemic subperiod 2). We note that in-person usage decreased from the prepandemic period for every disease cohort. The reduction was the greatest for cancer (–18% and –22% in pandemic subperiods 1 and 2, respectively), heart failure (–20% and –24%), and kidney failure (–18% and –19%). Smaller decreases are seen for anxiety, depression, and diabetes. We then compare Figure 1A with Figure 1B, which shows the overall usage. We observe that during the pandemic period, the number of telehealth visits exceeds the reduction in in-person visits for both depression and anxiety; we see an increase in overall usage for these disease cohorts during this period. We also note that for the diabetes cohort, the number of telehealth visits almost equals the decrease in in-person visits. However, for the cancer, heart failure, and kidney failure cohorts, there remains a significant reduction in the total number of visits.

XSL•FO
RenderX

Figure 1. The percent difference in in-person visits (A) and overall utilization (B) during the 12-month pandemic period from June 2020 through November 2020 (subperiod 1) and December 2021 to May 2021 (subperiod 2) for the 6 disease cohorts.

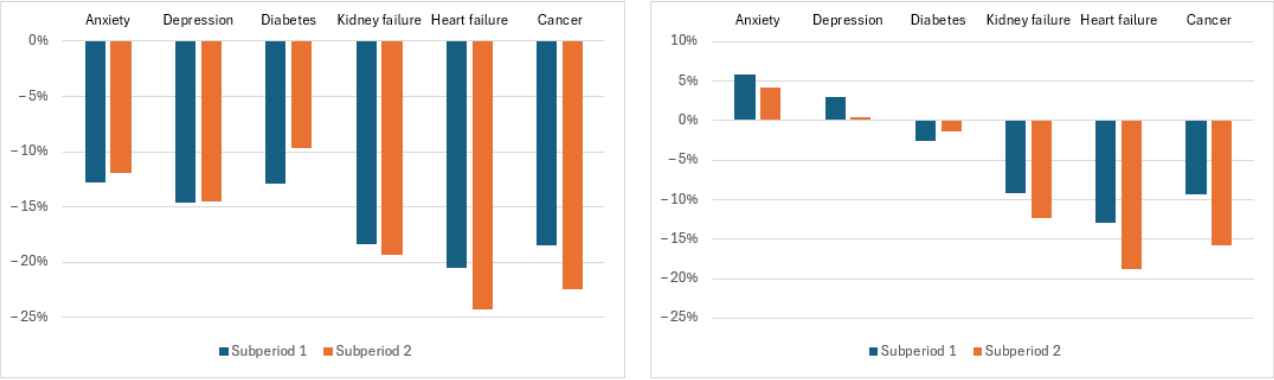

In-Person Visits

We run a linear regression model for each disease group for the number of visits in the pandemic period. Recall that the dependent variable was the natural logarithm of the number of in-person visits; therefore, we exponentiate the regression coefficients and interpret the coefficients as multiplicative.

The regression results are summarized in Table 2. We note that across disease groups, the higher the number of preperiod visits, the higher the expected number of pandemic period visits. We then observe that both White patients and, to a larger extent, those of other races have on average fewer in-person encounters than Black patients, after accounting for all other variables in the model. (Compared with similar Black patients, White patients have 1.8%-9.8% fewer visits and those of other races had 14.2%-19.7% fewer visits). For all diseases except heart failure, the Charlson Index is negatively associated with number of visits, indicating that, all else being equal, sicker patients have fewer visits. While the impact of age is statistically significant across diseases, the directionality of the impact differs by disease. Age is negatively associated with the number of visits for patients with both heart and kidney failure, has a small association with the number of visits for patients with diabetes and those with cancer, and is positively associated with the number of visits for anxiety and depression. Across diseases, men have on average fewer visits than women (holding everything else constant, estimates range from –5.6% for heart failure to –13.9% and –15.8% for depression and anxiety, respectively). Finally, across most diseases and holding everything else constant, patients from high and medium SVI counties have more visits on average than patients from counties with the lowest SVI (The exception are patients from high-SVI counties in the anxiety cohort, where the association is not statistically significant.)

XSL•FO
RenderX

Table 2. In-person regression results by disease cohort; for each disease cohort the dependent variable is the natural logarithm of the number of in-person visits during the 12-month pandemic period, June 2020 to May 2021 (*P* values are reported within parentheses).

	Anxiety, *r* (*P* value)	Depression, *r* (*P* value)	Diabetes, *r* (*P* value)	Kidney failure, *r* (*P* value)	Heart failure, *r* (*P* value)	Cancer, *r* (*P* value)
Intercept	1.12 (<.001)	1.14 (<.001)	1.37 (<.001)	1.78 (<.001)	1.84 (<.001)	1.38 (<.001)
Pre–COVID-19 visits	0.08 (<.001)	0.08 (<.001)	0.08 (<.001)	0.06 (<.001)	0.07 (<.001)	0.07 (<.001)
Charlson Index	–0.01 (<.001)	–0.02 (<.001)	–0.02 (<.001)	–0.03 (<.001)	–0.03 (<.001)	–0.03 (<.001)
Age (years)	0.005 (<.001)	0.004 (<.001)	0.0004 (.14)	–0.004 (<.001)	–0.01 (<.001)	0.001 (.002)
Sex						
Male	–0.16 (<.001)	–0.14 (<.001)	–0.10 (<.001)	–0.08 (<.001)	–0.06 (<.001)	–0.10 (<.001)
Unknown	0.27 (.47)	0.05 (.89)	–0.34 (.52)	—[a]	—	—
Race						
Other	–0.15 (<.001)	–0.17 (<.001)	–0.16 (<.001)	–0.16 (<.001)	–0.17 (<.001)	–0.22 (<.001)
White	–0.07 (<.001)	–0.07 (<.001)	–0.03 (<.001)	–0.05 (<.001)	–0.02 (.30)	–0.1 (<.001)
SVI[b]						
High	–0.002 (.82)	0.02 (.09)	0.03 (.004)	0.02 (.16)	0.01 (.49)	0.07 (<.001)
Medium	0.05 (<.001)	0.07 (<.001)	0.1 (<.001)	0.09 (<.001)	0.1 (<.001)	0.13 (<.001)
Missing	–0.01 (.88)	–0.05 (.31)	–0.09 (.03)	–0.05 (.52)	–0.01 (.95)	–0.02 (.77)

[a]Not applicable.

[b]SVI: Social Vulnerability Index.

Use of Telehealth

We studied 2 aspects of telehealth usage: first the percentage of patients using telehealth, and second the average number of telehealth visits for those patients who have at least 1 telehealth visit. In Figure 2, we break down the usage by race and SVI.

The highest rate of telehealth adoption (any usage) is by the anxiety and depression cohorts. We note that patients of other races are most likely to have at least 1 telehealth visit for both conditions (Figure 2A); however, among patients who use telehealth, Black patients have a higher average number of visits (Figure 2B). For kidney failure, heart failure, and cancer, while the rate of any telehealth usage by White patients is moderately lower than for Black patients, the average number of visits for those patients who do use telehealth are similar.

When telehealth usage for anxiety and depression is broken down by SVI, we note a more nuanced pattern. SVI does not appear to be associated with whether or not members in the Anxiety cohort use telehealth (Figure 2C), but among those who do, those with the lowest SVI have the lowest number of visits (Figure 2D). For depression we also note that patients from low-SVI areas have significantly fewer telehealth visits on average. Across all disease cohorts except anxiety, patients from low-SVI areas are more likely to use telehealth, followed by patients from medium and then high-SVI areas (Figure 2C). However, except for anxiety and depression the differences in number of telehealth visits for patients who use telehealth are generally less.

To quantify these relationships, and account for the many other factors that may affect usage, we next regress the number of telehealth visits in the pandemic period on the independent variables. Table 3 summarizes the results by disease group. For each disease group there are 2 regression coefficients. The first corresponds to the logistic regression model that explains which members do not have a telehealth visit. The second coefficient is associated with the negative binomial regression model that explains the number of visits.

We note that the more in-person pre–COVID-19 visits patients had, the less likely they are to have a telehealth visit (holding everything else constant). On average, for each additional in-person visit in the pre–COVID-19 period, the odds of not having a telehealth visit increases by 31.0% and 29.7% for anxiety and depression, respectively. We further note that across diseases, for those using telehealth, a higher number of pre–COVID-19 in-person visits is associated with fewer telehealth visits during the pandemic period.

Similarly, we note that the higher the Charlson Index, the less likely the patient is to use telehealth during COVID-19 period. However, while for kidney failure and heart failure, a higher Charlson Index is associated with a higher number of telehealth visits, for other disease groups, it is negatively associated with the number of visits.

Across disease groups, older patients are more likely to use telehealth, and for patients who do use telehealth, older patients typically have a higher number of visits (holding everything else constant). Male patients are similarly more likely to use telehealth and have a higher number of telehealth than female patients (holding everything else constant).

XSL•FO
RenderX

For anxiety, Black patients are less likely to use telehealth than White patients, and for diabetes, kidney failure, and heart failure this is reversed: Black patients in these cohorts are more likely to have at least 1 telehealth appointment than White patients (holding everything else constant). Across diseases, among those who use telehealth, White patients have a higher number of visits on average, holding everything else constant.

Finally, there is a diverging relationship between SVI group and telehealth usage. Patients from high-SVI counties are less likely to use telehealth across all disease cohorts (except for kidney failure). However, among patients who do use telehealth, patients from high-SVI counties have a higher number of telehealth visits on average than patients from low-SVI counties (everything else being the same).

Figure 2. The percentage of patients who have at least 1 telehealth visit (A) and the average number of telehealth visits for patients with at least 1 visit (B), during the 12-month pandemic period broken down by race; the percentage of patients who have at least 1 telehealth visit (C), and the average number of telehealth visits for patients with at least 1 visit (D) during the 12-month pandemic period, broken down by Social Vulnerability Index group.

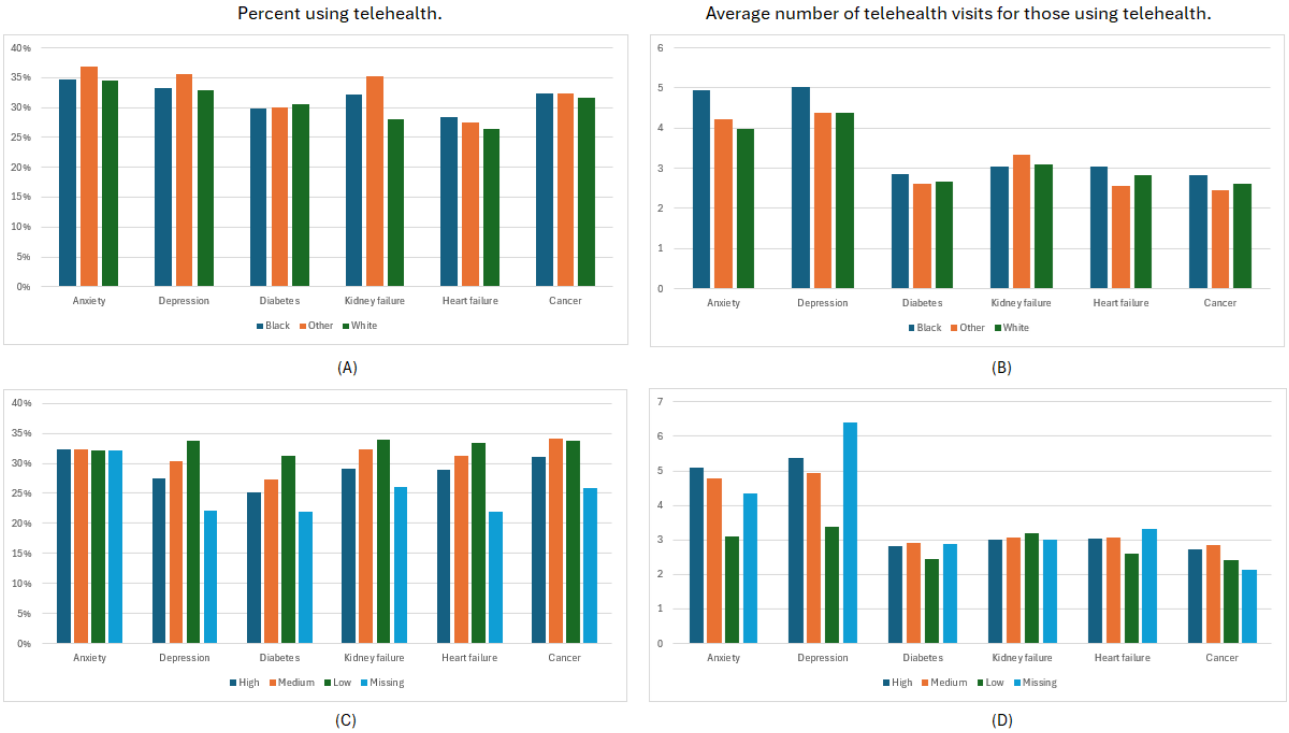

XSL•FO
RenderX

Table 3. Telehealth regression results for the 6 disease cohorts: for each disease group a zero-inflated negative binomial regression model is run; the first column corresponds to a logistic regression model that explains which members do not have a telehealth visit in the 12-month pandemic period; and the second column reflects the results from a negative binomial regression model that explains the number of visits in the same period (*P* values are reported within parentheses below each regression coefficient).

	Anxiety, *r* (*P* value)		Depression, *r* (*P* value)		Diabetes, *r* (*P* value)		Kidney failure, *r* (*P* value)		Heart failure, *r* (*P* value)		Cancer, *r* (*P* value)	
Variable	No visits	How many	No visits	How many	No visits	How many	No visits	How many	No visits	No visits	No visits	How many
Intercept	1.9 (<.001)	0.7 (<.001)	1.96 (<.001)	0.79 (<.001)	1.51 (<–.001)	0.15 (<.001)	1.84 (<.001)	–0.48 (<.001)	1.34 (<.001)	–0.25 (<.001)	0.85 (<.001)	–0.29 (<.001)
Pre–COVID-19 visits	0.27 (<.001)	–2.28 (<.001)	0.26 (<.001)	–2.33 (<.001)	0.17 (<.001)	–2.43 (<.001)	0.14 (<.001)	–2.28 (<.001)	0.15 (<.001)	–2.67 (<.001)	0.15 (<.001)	–1.73 (<.001)
Charlson Index	0.03 (<.001)	–0.02 (<.001)	0.0003 (<.001)	–0.01 (<.001)	0.09 (<.001)	–0.001 (<.001)	0.06 (<.001)	0.05 (<.001)	0.07 (<.001)	0.02 (<.001)	0.11 (<.001)	–0.01 (<.001)
Age (years)	–0.01 (<.001)	0.0004 (<.001)	–0.01 (<.001)	–0.001 (<.001)	–0.01 (<.001)	0.01 (<.001)	–0.01 (<.001)	0.02 (<.001)	–0.01 (<.001)	0.02 (<.001)	–0.01 (<.001)	0.01 (<.001)
Sex												
Male	–0.002 (<.001)	0.26 (<.001)	0.001 (<.001)	0.3 (<.001)	–0.2 (<.001)	0.12 (<.001)	–0.11 (<.001)	0.1 (<.001)	–0.16 (<.001)	0.1 (<.001)	–0.07 (<.001)	0.09 (<.001)
Unknown	–1.48 (<.001)	–1.86 (<.001)	–1.27 (<.001)	–5.12 (<.001)	–0.72 (<.001)	–0.17 (<.001)	—[a]	—	—	—	—	—
Race												
Other	–0.1 (<.001)	–0.09 (<.001)	–0.09 (<.001)	–0.09 (<.001)	–0.06 (<.001)	0.03 (<.001)	0.12 (<.001)	0.01 (<.001)	–0.14 (<.001)	0.01 (<.001)	–0.16 (<.001)	0.01 (<.001)
White	–0.09 (<.001)	0.03 (<.001)	0.01 (<.001)	0.05 (<.001)	0.03 (<.001)	0.02 (<.001)	0.12 (<.001)	(<.001)	0.07 (<.001)	0.13 (<.001)	–0.003 (<.001)	0.02 (<.001)
SVI[b]												
Low	–0.47 (<.001)	–0.29 (<.001)	–0.47 (<.001)	–0.23 (<.001)	–0.14 (<.001)	–0.41 (<.001)	0.07 (<.001)	–0.3 (<.001)	–0.12 (<.001)	–0.5 (<.001)	–0.08 (<.001)	–0.07 (<.001)
Medium	–0.004 (<.001)	–0.14 (<.001)	–0.06 (<.001)	–0.17 (<.001)	0.05 (<.001)	–0.16 (<.001)	0.05 (<.001)	–0.19 (<.001)	0.03 (<.001)	–0.19 (<.001)	0.11 (<.001)	–0.27 (<.001)
Missing	–0.1 (<.001)	0.04 (<.001)	0.22 (<.001)	0.23 (<.001)	0.07 (<.001)	0.29 (<.001)	–0.002 (<.001)	0.33 (<.001)	0.2 (<.001)	0.14 (<.001)	–0.31 (<.001)	–0.01 (<.001)

[a]Not applicable.

[b]SVI: Social Vulnerability Index.

Discussion

Principal Findings

This study compares the impact of COVID-19 pandemic on health care usage across multiple disease cohorts. We found that, as reported in previous studies, in-person visits decreased across all disease cohorts. The largest decreases in in-person visits were observed among patients with cancer, heart failure, and kidney failure. There are a number of possible reasons for this, including fear of COVID-19 infection, as all of these diseases are risk factors for COVID-19 complications, and age, as the average age was also the highest in these cohorts [28]. Across diseases, we found that the higher the pre–COVID-19 health care usage, the higher the in-person usage in the pandemic period, and the lower the odds of using telehealth. Interestingly, conditioned on a person using telehealth, the higher the pre–COVID-19 usage, the lower the number of pandemic period telehealth appointments, perhaps indicating that sicker patients were being routed to in-person visits.

Telehealth usage for some disease cohorts was equivalent to the reduction in in-person visits. In the 2 mental health cohorts we studied, the number of telehealth visits exceeded the reduction of in-person visits, leading to an increase in overall health care usage for both anxiety and depression. In contrast to the other diseases we studied, depression and anxiety may be treated and managed without physical examination [29,30], and therefore these findings are not surprising. Our finding that telehealth adoption is concurrent with the decrease in in-person mental health appointments is congruent with other recent studies [31,32]. There are additional possible explanations for the increase in usage among the anxiety and depression cohorts, including the impact of the COVID-19 pandemic on mental health.

During the pandemic, telehealth provided access to care and may have supplemented in-person visits to varying degrees. Therefore, it is critical that this mode of health care delivery does not introduce new disparities or exacerbate preexisting ones. When studying the impact of patients' demographic

XSL•FO
RenderX

characteristics association with telehealth usage, we find that women (except for anxiety and depression) and younger patients were less likely to use telehealth, consistent with previous publications [16-19]. In contrast to earlier works, we do not find consistent differences between telehealth adoption of Black and White patients after accounting for all other covariates. However, conditioned on that patients use telehealth, and accounting for all other factors, White patients have more visits on average than Black patients. This is in contrast with the average numbers presented in Figure 2, highlighting the importance of accounting for the many factors influencing telehealth adoption with regression modeling.

The observed relationship between SVI and telehealth usage patterns offers nuanced understanding of telehealth adoption across SVI areas. Across disease groups, patients from low-SVI counties are more likely to use telehealth but for fewer visits on average. On the other hand, patients from high-SVI counties are less likely to use telehealth, but if they do, they use telehealth more on average. Understanding barriers to telehealth adoption, including inadequate connectivity, lack of privacy, and other hidden barriers, is therefore especially important as telehealth visits may become a standard of care.

Telemedicine can facilitate effective communication between patients and health care workers. However, some patients find it difficult to adapt to telehealth communication. In addition, a number of older patients who are not as familiar with telehealth communication reported a relative lack of an emotional connection with the health care worker compared with in-person visits [33]. Telehealth consultations have been found to be more effective in primary care appointments and mental health consultations than they are for patients with high-risk conditions, based on both patient feedback and staff evaluations [8]. Our results reflect previous observations in which telehealth visits did not compensate for decreased in-person visits for patients with cancer, heart failure, or kidney failure, who experienced the greatest pandemic-related decrease in visits. As patients with these diagnoses require frequent visits to ensure stabilization of these chronic conditions, further research is needed to understand these differences. Our study also indicates that sicker patients (as captured by the Charlson index, and accounting for all other covariates) are less likely to use telehealth, and if they do they have lower number of telehealth visits on average, except for patients in the heart failure and kidney failure groups. Further study is needed to understand the nuances behind these usage patterns.

Limitations

This study has several limitations. In only rare cases did we observe death of patients in our study population, so we cannot confirm whether some of the reduction in visits is due to mortality in our disease cohorts. We note that cancer, heart failure, and kidney failure were the 3 disease cohorts with the largest reduction in visits in the second pandemic subperiod, and this may reflect higher mortality rates in these cohorts. We also note that we do not observe patient health care encounters outside the regional health system. Therefore, the extent to which patients moved away or sought services elsewhere is not captured. Finally, since our cohort has limited geographical scope, the results may not generalize beyond the regional service area.

Conclusions

The COVID-19 pandemic has affected health care usage across multiple disease conditions. The impacts of the pandemic on health care usage and the role of telehealth in replacing in-person visits vary among different disease cohorts. Understanding these differences can inform current practices and provides opportunities to further guide modalities of in-person and telehealth visits. Critically, further study is needed to ensure that all patients have equal access to telehealth services. A better understanding of the role of social determinants of health may lead to more support for patients and help individual health care providers improve access to care for patients with chronic conditions.

In summary, this study offers a unique comparison across disease cohorts from the same health system. In doing so, we highlight that many of the previously observed demographic differences consistently hold across disease cohorts. The study further takes a detailed analytical approach to study telehealth adoption offering new insights. For example, the study highlights that while telehealth adoption is lower for high SVI areas, if patients from high SVI areas use telehealth, then their usage exceeds that of other patients from lower SVI areas.

Acknowledgments

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (award UL1TR001409). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data sets generated for and analyzed during this study are not publicly available due the approved study protocol restrictions and patient privacy considerations. The data can be made available following a study protocol review and an appropriate institutional review board approval.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Health care codes used to identify our disease cohorts, and sensitivity analysis complementing the main results of the paper.
[PDF File (Adobe PDF File), 196 KB - jmir_v26i1e54991_app1.pdf]

References

1. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open 2021;11(3):e045343 [FREE Full text] [doi: 10.1136/bmjopen-2020-045343] [Medline: 33727273]
2. Roy C, Bollman EB, Carson LM, Northrop AJ, Jackson EF, Moresky RT. Assessing the indirect effects of COVID-19 on healthcare delivery, utilization and health outcomes: a scoping review. Eur J Public Health 2021;31(3):634-640 [FREE Full text] [doi: 10.1093/eurpub/ckab047] [Medline: 33755130]
3. Shinan-Altman S, Levkovich I, Tavori G. Healthcare utilization among breast cancer patients during the COVID-19 outbreak. Palliat Support Care 2020;18(4):385-391. [doi: 10.1017/s1478951520000516] [Medline: 32594966]
4. American College of Cardiology. COVID-19 pandemic indirectly disrupted heart disease care. 2021. URL: https://www.acc.org/about-acc/press-releases/2021/01/11/16/40/covid19-pandemic-indirectly-disrupted-heart-disease-care [accessed 2024-08-20]
5. Maringe C, Spicer J, Morris M, Purushotham A, Nolte E, Sullivan R, et al. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. Lancet Oncol 2020;21(8):1023-1034. [doi: 10.1016/s1470-2045(20)30388-0] [Medline: 32702310]
6. Nouri S, Lyles CR, Sherwin EB, Kuznia M, Rubinsky AD, Kemper KE, et al. Visit and between-visit interaction frequency before and after COVID-19 telehealth implementation. JAMA Netw Open 2023;6(9):e2333944. [doi: 10.1001/jamanetworkopen.2023.33944] [Medline: 37713198]
7. Villas-Boas SB, Kaplan S, White JS, Hsia RY. Adolescent total and mental health-related emergency department visits during the COVID-19 pandemic. JAMA Netw Open 2023;6(10):e2336463 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36463] [Medline: 37796500]
8. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
9. Albornoz SCD, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022;39(1):168-182 [FREE Full text] [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
10. Betancourt JA, Rosenberg MA, Zevallos A, Brown JR, Mileski M. The impact of COVID-19 on telemedicine utilization across multiple service lines in the United States. Healthcare (Basel) 2020;8(4):380 [FREE Full text] [doi: 10.3390/healthcare8040380] [Medline: 33019667]
11. Costa M, Reis G, Pavlo A, Bellamy C, Ponte K, Davidson L. Tele-mental health utilization among people with mental illness to access care during the COVID-19 pandemic. Community Ment Health J 2021;57(4):720-726 [FREE Full text] [doi: 10.1007/s10597-021-00789-7] [Medline: 33566269]
12. Miner H, Fatehi A, Ring D, Reichenberg JS. Clinician telemedicine perceptions during the COVID-19 pandemic. Telemed J E Health 2021;27(5):508-512. [doi: 10.1089/tmj.2020.0295] [Medline: 32946364]
13. Nitiema P. Telehealth before and during the COVID-19 pandemic: analysis of health care workers' opinions. J Med Internet Res 2022;24(2):e29519 [FREE Full text] [doi: 10.2196/29519] [Medline: 34978532]
14. Harsono D, Deng Y, Chung S, Barakat LA, Friedland G, Meyer JP, et al. Experiences with telemedicine for HIV care during the COVID-19 pandemic: a mixed-methods study. AIDS Behav 2022;26(6):2099-2111 [FREE Full text] [doi: 10.1007/s10461-021-03556-7] [Medline: 35064390]
15. Labisi T, Regan N, Davis P, Fadul N. HIV care meets telehealth: a review of successes, disparities, and unresolved challenges. Curr HIV/AIDS Rep 2022;19(5):446-453 [FREE Full text] [doi: 10.1007/s11904-022-00623-z] [Medline: 36063284]
16. Ladin K, Porteny T, Perugini JM, Gonzales KM, Aufort KE, Levine SK, et al. Perceptions of telehealth vs in-person visits among older adults with advanced kidney disease, care partners, and clinicians. JAMA Netw Open 2021;4(12):e2137193 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.37193] [Medline: 34870680]
17. Lattimore CM, Kane WJ, Fleming MA, Martin AN, Mehaffey JH, Smolkin ME, et al. Disparities in telemedicine utilization among surgical patients during COVID-19. PLoS One 2021;16(10):e0258452 [FREE Full text] [doi: 10.1371/journal.pone.0258452] [Medline: 34624059]
18. Luo J, Tong L, Crotty BH, Somai M, Taylor B, Osinski K, et al. Telemedicine adoption during the COVID-19 pandemic: gaps and inequalities. Appl Clin Inform 2021;12(4):836-844 [FREE Full text] [doi: 10.1055/s-0041-1733848] [Medline: 34496419]
19. Luo J, Tong L, Crotty BH, Somai M, Taylor B, Osinski K, et al. Telemedicine adoption during the COVID-19 pandemic: gaps and inequalities. Appl Clin Inform 2021;12(4):836-844 [FREE Full text] [doi: 10.1055/s-0041-1733848] [Medline: 34496419]
20. Ruberto R, Schweppe EA, Ahmed R, Swindell HW, Cordero CA, Lanham NS, et al. Disparities in telemedicine utilization during COVID-19 pandemic: analysis of demographic data from a large academic orthopaedic practice. JB JS Open Access 2022;7(2):e21.00116 [FREE Full text] [doi: 10.2106/JBJS.OA.21.00116] [Medline: 35425872]

XSL•FO
RenderX

21. Zhu J, Lois AW, Gitonga B, Chen-Meekin JY, Williams EJ, Khandelwal S, et al. The impact of socioeconomic status on telemedicine utilization during the COVID-19 pandemic among surgical clinics at an academic tertiary care center. Surg Endosc 2022;36(12):9304-9312 [FREE Full text] [doi: 10.1007/s00464-022-09186-x] [Medline: 35332387]
22. de Boer AG, Wijker W, de Haes HC. Predictors of health care utilization in the chronically ill: a review of the literature. Health Policy 1997;42(2):101-115. [doi: 10.1016/s0168-8510(97)00062-6] [Medline: 10175619]
23. van Oostrom SH, Picavet HSJ, de Bruin SR, Stirbu I, Korevaar JC, Schellevis FG, et al. Multimorbidity of chronic diseases and health care utilization in general practice. BMC Fam Pract 2014;15:61 [FREE Full text] [doi: 10.1186/1471-2296-15-61] [Medline: 24708798]
24. Sporinova B, Manns B, Tonelli M, Hemmelgarn B, MacMaster F, Mitchell N, et al. Association of mental health disorders with health care utilization and costs among adults with chronic disease. JAMA Netw Open 2019;2(8):e199910 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9910] [Medline: 31441939]
25. CDC/ATSDR SVI data and documentation dowload. Agency for Toxic Substances and Disease Registry. URL: https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html [accessed 2024-08-20]
26. Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 update and ICD-10 translation. Am Health Drug Benefits 2019;12(4):188-197 [FREE Full text] [Medline: 31428236]
27. Cameron AC, Trivedi PK. Microeconometrics Using Stata. College Station, TX: Stata Press; 2010.
28. Esai Selvan M. Risk factors for death from COVID-19. Nat Rev Immunol 2020;20(7):407 [FREE Full text] [doi: 10.1038/s41577-020-0351-0] [Medline: 32461673]
29. Scott AM, Clark J, Greenwood H, Krzyzaniak N, Cardona M, Peiris R, et al. Telehealth face-to-face provision of care to patients with depression: a systematic review and meta-analysis. Psychol Med 2022;52(14):2852-2860 [FREE Full text] [doi: 10.1017/S0033291722002331] [Medline: 35959559]
30. Krzyzaniak N, Greenwood H, Scott AM, Peiris R, Cardona M, Clark J, et al. The effectiveness of telehealth versus face-to face interventions for anxiety disorders: a systematic review and meta-analysis. J Telemed Telecare 2024;30(2):250-261. [doi: 10.1177/1357633X211053738] [Medline: 34860613]
31. McBain RK, Cantor J, Pera MF, Breslau J, Bravata DM, Whaley CM. Mental health service utilization rates among commercially insured adults in the US during the first year of the COVID-19 pandemic. JAMA Health Forum 2023;4(1):e224936 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.4936] [Medline: 36607697]
32. Kalmin MM, Cantor JH, Bravata DM, Ho P, Whaley C, McBain RK. Utilization and spending on mental health services among children and youths with commercial insurance. JAMA Netw Open 2023;6(10):e2336979 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36979] [Medline: 37787996]
33. Ghaderi H, Stowell JR, Akhter JR, Norquist M, Pugsley P, Subbian V. Impact of COVID-19 pandemic on emergency department visits: a regional case study of informatics challenges and opportunities. AMIA Annu Symp Proc 2021;2021:496-505 [FREE Full text] [Medline: 35308996]

Abbreviations

ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Statistical Classification of Diseases, Tenth Revision
SVI: Social Vulnerability Index

Edited by A Mavragani; submitted 29.11.23; peer-reviewed by A Gangadhara Rao, O Nguyen; comments to author 29.02.24; revised version received 14.05.24; accepted 04.07.24; published 03.10.24.

Please cite as:
Bjarnadóttir MV, Anderson D, Anderson KM, Aljwfi O, Peluso A, Ghannoum A, Balba G, Shara N
Health Care Usage During the COVID-19 Pandemic and the Adoption of Telemedicine: Retrospective Study of Chronic Disease Cohorts
J Med Internet Res 2024;26:e54991
URL: https://www.jmir.org/2024/1/e54991
doi:10.2196/54991
PMID:39361360

©Margrét Vilborg Bjarnadóttir, David Anderson, Kelley M Anderson, Omar Aljwfi, Alina Peluso, Adam Ghannoum, Gayle Balba, Nawar Shara. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Statistical Analysis of Telehealth Use and Pre- and Postpandemic Insurance Coverage in Selected Health Care Specialties in a Large Health Care System in Arkansas: Comparative Cross-Sectional Study

Aysenur Betul Cengil[1], PhD; Sandra Eksioglu[1], PhD; Burak Eksioglu[1], PhD; Hari Eswaran[2], PhD; Corey J Hayes[2], MPH, PharmD, PhD; Cari A Bogulski[2], PhD

[1]Industrial Engineering Department, College of Engineering, University of Arkansas, Fayetteville, AR, United States

[2]Department of Obstetrics and Gynecology, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, United States

Corresponding Author:
Sandra Eksioglu, PhD
Industrial Engineering Department
College of Engineering
University of Arkansas
4183 Bell Engineering Center
Fayetteville, AR, 72701
United States
Phone: 1 4795756042
Email: sandra@uark.edu

Abstract

Background: The COVID-19 pandemic triggered policy changes in 2020 that allowed insurance companies to reimburse telehealth services, leading to increased telehealth use, especially in rural and underserved areas. However, with many emergency rules ending in 2022, patients and health care providers face potential challenges in accessing these services.

Objective: This study analyzed telehealth use across specialties in Arkansas before and after the pandemic (2017-2022) using data from electronic medical records from the University of Arkansas for Medical Sciences Medical Center. We explored trends in insurance coverage for telehealth visits and developed metrics to compare the performance of telehealth versus in-person visits across various specialties. The results inform insurance coverage decisions for telehealth services.

Methods: We used pre- and postpandemic data to determine the impacts of the COVID-19 pandemic and changes in reimbursement policies on telehealth visits. We proposed a framework to calculate 3 appointment metrics: indirect waiting time, direct waiting time, and appointment length. Statistical analysis tools were used to compare the performance of telehealth and in-person visits across the following specialties: *obstetrics and gynecology*, *psychiatry*, *family medicine*, *gerontology*, *internal medicine*, *neurology*, and *neurosurgery*. We used data from approximately 4 million in-person visits and 300,000 telehealth visits collected from 2017 to 2022.

Results: Our analysis revealed a statistically significant increase in telehealth visits across all specialties ($P<.001$), showing an 89% increase from 51,589 visits in 2019 to 97,461 visits in 2020, followed by a 21% increase to 117,730 visits in 2021. Around 92.57% (134,221/145,001) of telehealth patients from 2020 to 2022 were covered by *Medicare*, *Blue Cross* and *Blue Shield*, *commercial and managed care*, *Medicaid*, and *Medicare Managed Care*. In-person visits covered by *Medicare* and *Medicaid* decreased by 15%, from 313,196 in 2019 to 264,696 in 2022. During 2020 to 2022, about 22.84% (33,123/145,001) of total telehealth visits during this period were covered by *Medicare* and 53.58% (86,317/161,092) were in *psychiatry*, *obstetrics and gynecology*, and *family medicine*. We noticed a statistically significant decrease ($P<.001$) in the average indirect waiting time for telehealth visits, from 48.4 to 27.7 days, and a statistically significant reduction in appointment length, from 93.2 minutes in 2020 to 39.59 minutes in 2022. The indirect waiting time for *psychiatry* telehealth visits was almost 50% shorter than that for in-person visits. These findings highlight the potential benefits of telehealth in providing access to health care, particularly for patients needing psychiatric care.

XSL•FO
RenderX

Conclusions: Reverting to prepandemic regulations could negatively affect Arkansas, where many live in underserved areas. Our analysis shows that telehealth use remained stable beyond 2020, with *psychiatry* visits continuing to grow. These findings may guide insurance and policy decisions in Arkansas and other regions facing similar access challenges.

(J Med Internet Res 2024;26:e49190) doi:10.2196/49190

KEYWORDS

appointment scheduling metrics; insurance coverage; statistical hypothesis testing; telehealth

Introduction

Background

Since 2020, there has been a significant increase in the use of telehealth services in the United States due to the COVID-19 pandemic, which was declared a public health emergency [1-4]. Telehealth visits have become an essential means of providing patients with access to care while reducing the risk of exposure for both patients and health care providers [5,6]. In response, the Health and Human Services Office for Civil Rights relaxed certain regulations, making it easier to provide health services via remote communication technology [7,8]. The Affordable Care Act also allowed Medicare and Medicaid to cover and reimburse the use of certain telehealth services, which further increased their use. In addition, the rapid growth of telecommunication technologies in recent years has facilitated the delivery of care via telehealth [9].

However, with the end of many emergency rules in 2022, federal government offices and private insurance companies are changing their reimbursement policies [10-12]. Thus, patients and health care providers are concerned about losing access to telehealth [13,14]. The rollback of COVID-19 waivers, coverage and payment policies, lack of insurer coverage of telehealth visits, and low or no reimbursement are among the top barriers to using telehealth, according to a recent survey of 1545 physicians by the American Medical Association [15]. Nevertheless, studies have shown that patients and health care providers are willing to use telehealth visits after the COVID-19 pandemic [16-19]. To evaluate the trends in telehealth use in Arkansas from 2017 to 2022, we conducted a study using data from the electronic medical records of the University of Arkansas for Medical Sciences Medical Center (UAMS Health), a network of health care providers across the state. Our analysis revealed that there continues to be a strong demand for telehealth services for certain specialties, such as psychiatry, obstetrics and gynecology (OB/GYN), and family medicine.

Objectives

Our study aims to analyze trends in telehealth use across specialties and insurance coverage before and after the pandemic. The study also uses performance metrics, such as indirect waiting time, direct waiting time, and appointment length, to compare telehealth and in-person visits. We aim to identify the most popular specialties among telehealth users and highlight observations about the performance of telehealth visits within these specialties. By doing so, we aim to provide data-driven evidence of telehealth use in Arkansas that can be used by insurance companies and policy makers to reevaluate their telehealth reimbursement policies.

Methods

Study Design

We used the patient data collected from the Arkansas Clinical Data Repository, which is maintained by the Department of Biomedical Informatics in the College of Medicine at the University of Arkansas for Medical Sciences (UAMS). We conducted a time series comparative cross-sectional study to investigate insurance coverage trends before and after the pandemic for different specialties. We compared the performance of telehealth and in-person visits and evaluated telehealth performance over time for these specialties.

Setting

The Arkansas Clinical Data Repository collects demographic, inpatient and outpatient health care use, and pharmacy prescription–related data from electronic medical records. Figure 1 illustrates the number of patients and visits for all health care specialties from 2017 to 2022, including both telehealth and in-person appointments. The data analyzed were collected from patients residing in Arkansas and neighboring states who scheduled appointments at UAMS Health. The data include insurance coverage information for both in-person and telehealth visits and appointment time stamps, including appointment creation day, appointment day, check-in and checkout times, and roomed and visit end times.

XSL•FO
RenderX

Figure 1. Number of patients and visits (2017-2022).

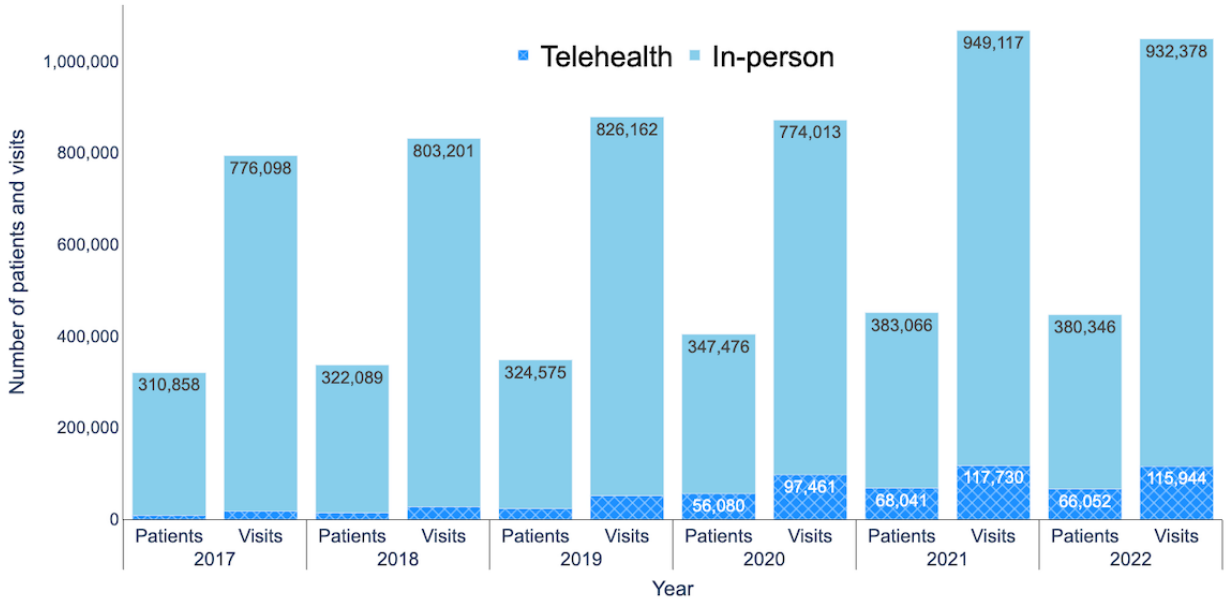

Data Preprocessing

Overview

We defined participants by indicating the inclusion and exclusion criteria for insurance coverage and appointment scheduling metrics. We have provided a detailed explanation of how the study size was determined. In addition, we have highlighted the outcomes and exposures for the analyses (insurance coverage and appointment scheduling metrics).

Insurance Coverage

In our analyses, we excluded visits that do not have insurance provider information. Table 1 summarizes the number and the corresponding percentage of visits without insurance provider information for in-person and telehealth visits. For in-person visits without insurance information, *radiology* accounts for 55.56% (347,386/625,295) visits of the observations, *family medicine* accounts for 12.91% (80,722/625,295), *radiation oncology* accounts for 11.94% (74,643/625,295), and 15.43% (96,477/625,295) do not have any specialty information. Among telehealth visits without insurance information, *call center* [20] accounts for 67.68% (184,989/273,309) of the observations, *OB/GYN* accounts for 19.5% (53,292/273,309), and *HealthNow* [21,22] accounts for 9.64% (26,339/273,309). It should be noted that patients are not billed for *call center* services.

Table 1. Number and percentage of in-person and telehealth visits without insurance provider information.

Year	In-person visits (n=625,295), n (%)	Telehealth visits (n=273,309), n (%)
2017	65,985 (8.5)	13,754 (79.45)
2018	83,074 (10.34)	24,746 (89.03)
2019	96,490 (11.68)	48,675 (94.35)
2020	79,594 (10.28)	57,944 (59.45)
2021	108,964 (11.48)	62,198 (52.83)
2022	191,188 (20.51)	65,992 (56.92)

Figures 2 and 3 present a breakdown of insurance coverage for telehealth and in-person visits from 2017 to 2022. Our analysis excluded the following insurance providers: *international*, *prepay non-Arkansas resident*, and *worker's compensation*, as these insurance providers cover only 0.36% (564/154,523) and 1.06% (46,894/4,435,674) of total telehealth and in-person visits, respectively. After preprocessing the data, we evaluated the insurance coverage trends (our primary outcome) of telehealth visits over time. We identified top insurance providers for certain specialties (our secondary outcome) based on the number of telehealth visits over time (our exposures).

XSL•FO
RenderX

Figure 2. Insurance coverage breakdown (2017-2022) for telehealth visits.

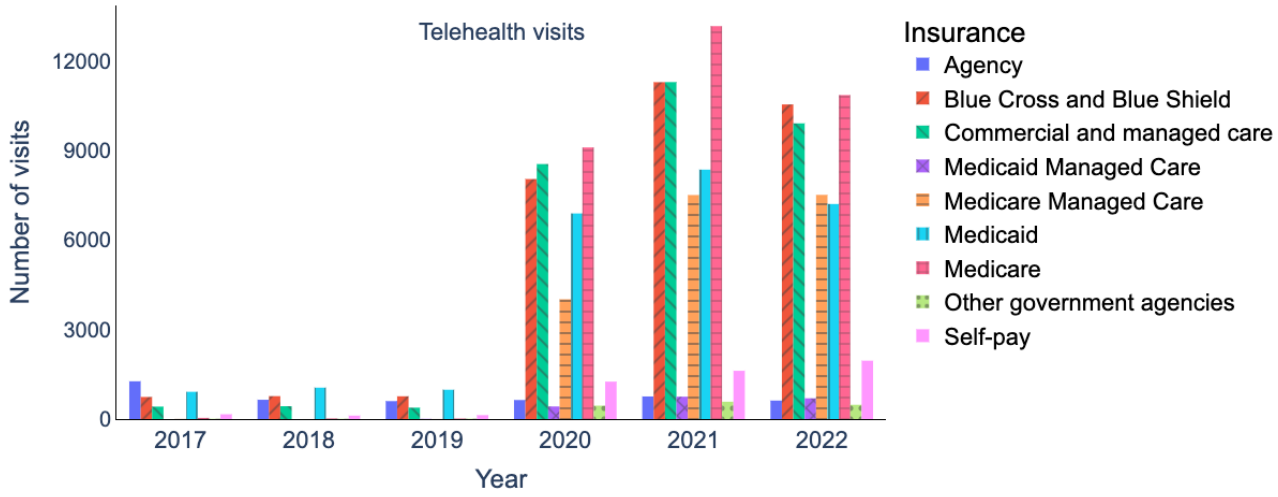

Appointment Scheduling

The workflow diagram of appointment scheduling at UAMS Health, whether for telehealth or in-person visits, is presented in Figure 4. To evaluate the scheduling efficiency, we used 3 metrics: *indirect waiting time*, *direct waiting time*, and *appointment length*. Indirect waiting time refers to the number of days between the *appointment creation* and the *appointment day*. Direct waiting time represents the difference between *check-in* and *roomed time*, while appointment length is the duration between *roomed time* and *visit end time*, both measured in minutes. It is important to note that some visits may not have all the necessary time stamps to calculate these metrics.

Figure 4. Workflow for both telehealth and in-person appointments.

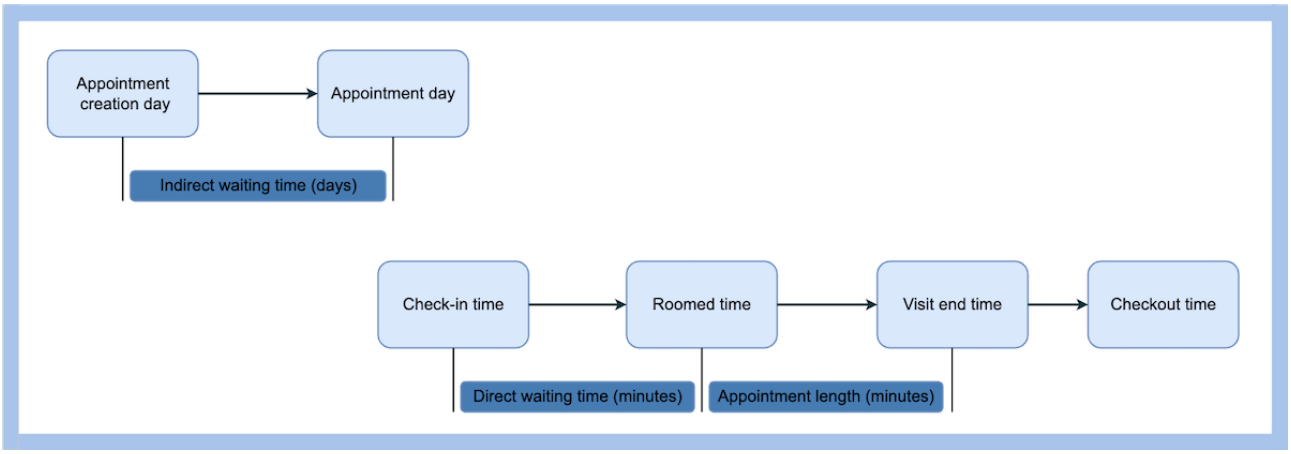

The process used to extract the data, build the dataset, and calculate the scheduling metrics for our dataset is shown in Figure 5. We established rules to identify errors and outliers and removed visits with incomplete time stamps, errors, and outliers. The first column of Figure 5 provides an overview of the process for calculating the indirect waiting time metric. For this metric, we excluded visits without an appointment creation time and visits where the creation day is after the appointment day. In cases where a patient created multiple appointments on the same day, we only considered the earliest one. We also identified appointments with an indirect waiting time >365 days as outliers. Similar procedures were applied to calculate the direct waiting time and appointment length metrics. The total number of telehealth and in-person visits considered in each step of this process are presented in Figure 5. Using the subset of visits for each scheduling metric, we compared the performance of telehealth and in-person visits (our primary outcome) via different scheduling metrics (our exposures).

XSL•FO
RenderX

Figure 5. Flow diagrams of data extraction and analysis for each scheduling metric. n_i: number of in-person visits; n_t: number of teleheath visits.

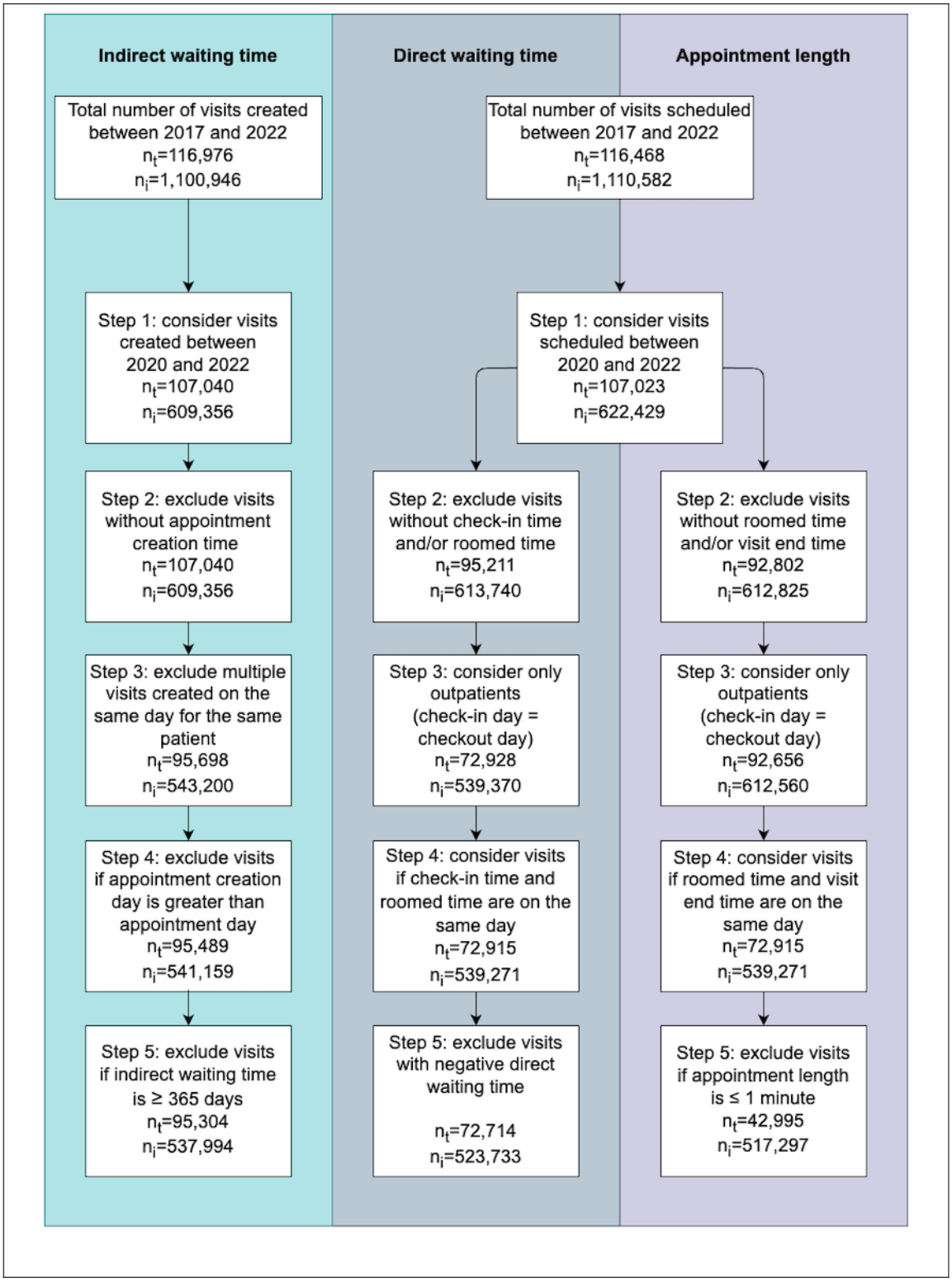

Data Analysis: Statistical Analysis for Insurance Coverage

To determine whether there is a statistically significant change in insurance coverage, we conducted additional hypothesis tests. An example hypothesis we tested is the following:

- H_0: There is no difference between the mean number of insurance-covered telehealth visits in 2019 and 2020.
- H_a: There is a difference between the mean number of insurance-covered telehealth visits in 2019 and 2020.

We conducted similar hypothesis testing to compare insurance coverage for telehealth visits during 2020 and 2021 as well as 2021 and 2022. The same set of hypothesis tests was conducted for in-person visits.

We conducted tests to determine if the data are normally distributed. We used the paired 2-tailed *t* test for normally

XSL•FO
RenderX

distributed data and the Wilcoxon signed rank test for nonnormally distributed data. The significance threshold (α) is .05 [23,24]. First, we used a 2-tailed test to evaluate if the population means are significantly different. For the case where the null hypothesis was rejected, we conducted additional 1-tailed tests to determine whether the difference in means is >0.

Statistical Analysis for Appointment Scheduling Metrics

The objective of this analysis was to investigate whether there is a statistically significant difference in performance metrics between in-person and telehealth services from 2020 to 2022 and between consecutive years for telehealth (or in-person) services. As telehealth use differs among specialties, we focused on the specialties with the most telehealth visits during 2020 to 2022. We conducted several hypothesis testing using the processed data. Here is an example of one of the hypotheses tested for *OB/GYN* visits:

- H_0: There is no difference between the means of direct waiting time for in-person and telehealth OB/GYN visits in 2020.
- H_a: There is a difference between the means of direct waiting time for in-person and telehealth OB/GYN visits in 2020.

We performed similar hypothesis testing for indirect waiting time, direct waiting time, and appointment length for the years 2021 and 2022. In addition, we compared the appointment scheduling metrics for telehealth (in-person) visits in consecutive years and extended the hypothesis testing to other specialties. Because our sample size is large, we used a 2-tailed *z* test with a significance threshold (α) of .05 [23,24].

For those cases where the null hypothesis was rejected, we conducted additional tests. We performed 1-tailed *z* tests to determine whether the difference between the means is >0.

Ethical Considerations

Data collection was approved by the Institutional Review Board of the UAMS (IRB 275271). The study was determined to not be human participant research. The data were deidentified, and there was no interaction with individuals.

Results

Trends in Telehealth and In-Person Visits

Telehealth Visits

In Figure 6, we present data on the number of telehealth visits at UAMS Health from 2017 to 2022, covering 7 different specialties (*psychiatry*, OB/GYN, *family medicine*, *gerontology*, *internal medicine*, *neurology*, and *neurosurgery*) as well as the *call center* and *HealthNow*. Before 2020, telehealth services were only used by *OB/GYN* patients due to the Antenatal and Neonatal Guidelines, Education and Learning System (ANGELS) program, which began in 2003 to provide telemedicine consultations to pregnant women considered high risk at the UAMS [25]. The number of *OB/GYN* visits increased from 17,250 in 2017 to 26,381 in 2018. This corresponds to a 53% increase. However, this number decreased significantly in 2019. Between 2020 and 2022, these 7 specialties, along with the *call center* and *HealthNow*, accounted for almost 87.52% (289,829/331,135) of all telehealth visits. Interestingly, the corresponding number of in-person visits for these 7 specialties only represented about 21.95% (582,865/2,655,508) of the total in-person visits.

In 2020, telehealth visits experienced a substantial increase in response to the COVID-19 pandemic. As shown in Figure 6, in 2021, the number of telehealth visits for *family medicine* went from 3381 to 11,035, which corresponds to a 226.4% increase. The number of *neurology* visits went from 3175 to 6062, a 90.9% increase, and the number of *internal medicine* visits went from 1299 to 2407, an 85.3% increase. However, in 2022, telehealth visits decreased for all specialties, except *neurology*, with an accompanying rise in the use of the *call center*. This trend is consistent with the national decrease in telehealth use, possibly due to the decreasing number and severity of COVID-19 cases [26].

In Figure 7, we provide a summary of the number of visits for different appointment types for various telehealth visits. Notably, except for *psychiatry* and *OB/GYN*, about 69.74% (39,270/56,307) of telehealth visits across all specialties were follow-up visits and phone conversations. The convenience of telehealth for follow-up visits is evident, as it saves patients the time and effort required to travel to health care facilities [27]. In addition, several studies have suggested that phone visits are more convenient for patients with inadequate broadband access or technological illiteracy [28,29].

Regarding *psychiatry* visits, approximately 61.95% (20,692/33,399) of visits focused on psychotherapy. This finding is consistent with other research that suggests telehealth visits as an alternative to in-person *psychiatry* visits, particularly for less common mental and physical health conditions [30,31].

XSL•FO
RenderX

Figure 6. Number of visits for the selected health care specialties, call center and HealthNow (2017-2022) for telehealth visits. OB/GYN: obstetrician and gynecologist.

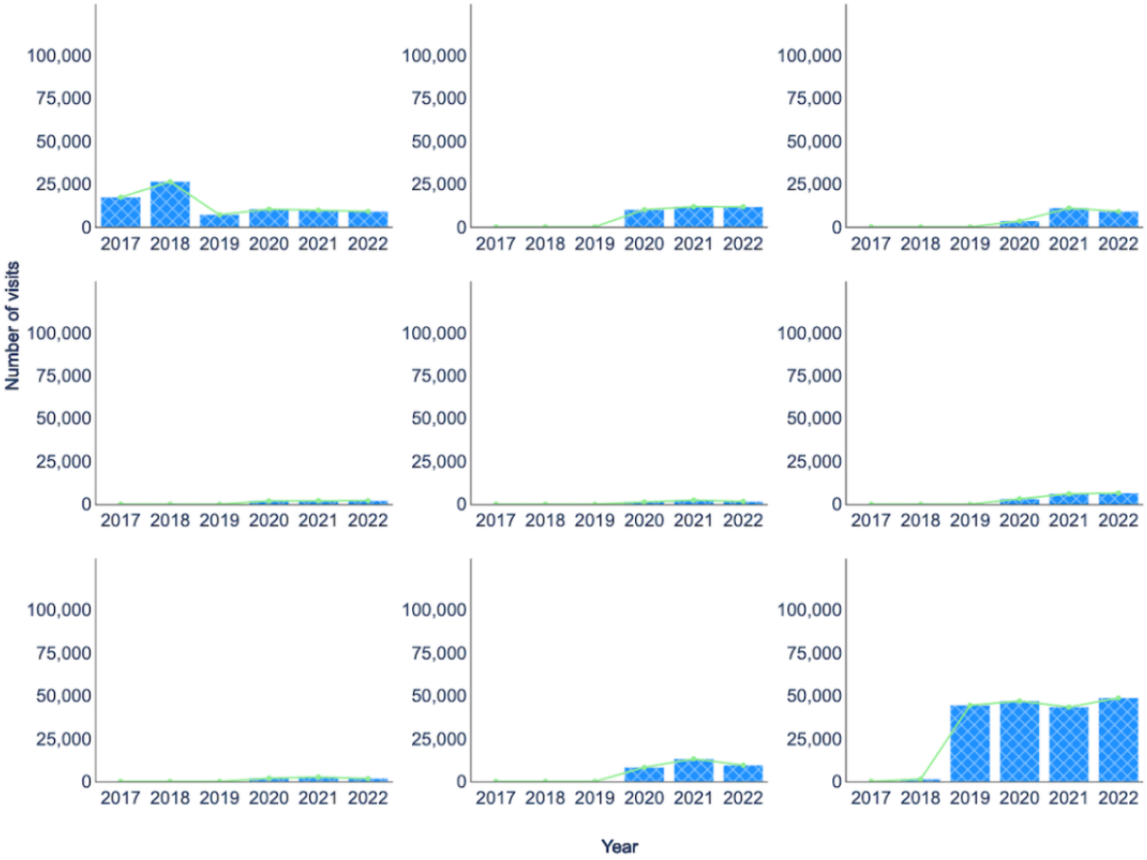

Figure 7. Appointment types of telehealth visits (2020-2022). IDHI: Institute for Digital Health & Innovation; OB/GYN: obstetrician and gynecologist.

XSL•FO
RenderX

In-Person Visits

Figure 8 provides the number of in-person visits at UAMS Health for 7 different specialties between 2017 and 2022. In 2020, the number of in-person visits decreased for all services except for *family medicine*. The low use of telehealth in *family medicine* may be due to physicians' lack of training [32,33]. The number of in-person visits for *neurosurgery* went from 5605 in 2019 to 2184 in 2020. This corresponds to a 61% decrease. During the same period, the number of *psychiatry* visits went from 18,366 to 10,719, a 41.6% decrease, and the number of *gerontology* visits went from 18,393 to 11,140, a 39.4% decrease. The decrease in *neurosurgery* in-person visits resulted from the suspension of elective surgeries during the COVID-19 pandemic [34,35].

Figure 8. Number of visits for the selected health care specialties, call center and HealthNow (2017-2022) for in-person visits. OB/GYN: obstetrician and gynecologist.

Because *psychiatry* visits do not require a physical examination, the decrease in the number of in-person visits observed is expected [34,36,37]. Telehealth visits are equivalent to in-person visits in terms of diagnostic accuracy, treatment effectiveness, quality of care, patient satisfaction, privacy, and confidentiality, as per the American Psychiatric Association (APA) [38]. Furthermore, the decrease in in-person *gerontology* services is likely due to the high vulnerability of older adults to COVID-19 infection [39].

In 2021, there was an increase in the number of in-person visits for all specialties except *neurosurgery*. However, recent studies suggest that most patients and health care providers preferred telehealth *neurosurgery* visits and were satisfied with the experience [34,40]. Overall, in 2022, we noticed minimal changes in the number of in-person visits, which indicates that demand for services is stabilizing. However, there were fewer total visits in 2022 compared to 2021, which included both in-person and telehealth visits. In summary, our observations suggest that patient visits shifted from in-person to telehealth during 2020 and 2021 compared with 2017 to 2019, with the exception of *OB/GYN*. When comparing 2022 to 2021, we observed only slight changes in the number of in-person and telehealth visits.

Insurance Coverage

Most patients who used telehealth services during the period of 2020 to 2022 had their expenses covered by several insurance providers, including *Medicare*, *Blue Cross* and *Blue Shield*, *commercial and managed care*, *Medicaid*, and *Medicare Managed Care*. These insurance providers accounted for 92.9% (134,221/144,437) of all telehealth visits during this time, as shown in Figure 2. It is worth noting that most of the in-person visits were also covered by these same insurance providers, as depicted in Figure 3. Specifically, these insurance providers

XSL•FO
RenderX

covered 91.8% (2,067,072/2,251,199) of all in-person visits that occurred between 2020 and 2022.

The mean values, SDs, and sample sizes (presented as n) of insurance-covered visits are indicated in Table 2. In addition, the result of the hypothesis tests and the corresponding *P* values are presented in Table 3. According to Table 3, the increase in insurance coverage for telehealth visits in 2020 compared to 2019 and in 2021 compared to 2020 was statistically significant. There were no statistically significant differences in insurance coverage for telehealth visits during 2021 and 2022. There were no statistically significant differences in insurance coverage for in-person visits during 2019 and 2020. However, the increase in insurance coverage for in-person visits in 2021 versus 2020 and the decrease in insurance coverage for in-person visits in 2022 versus 2021 were statistically significant.

Figure 3. Insurance coverage breakdown (2017-2022) for in-person visits.

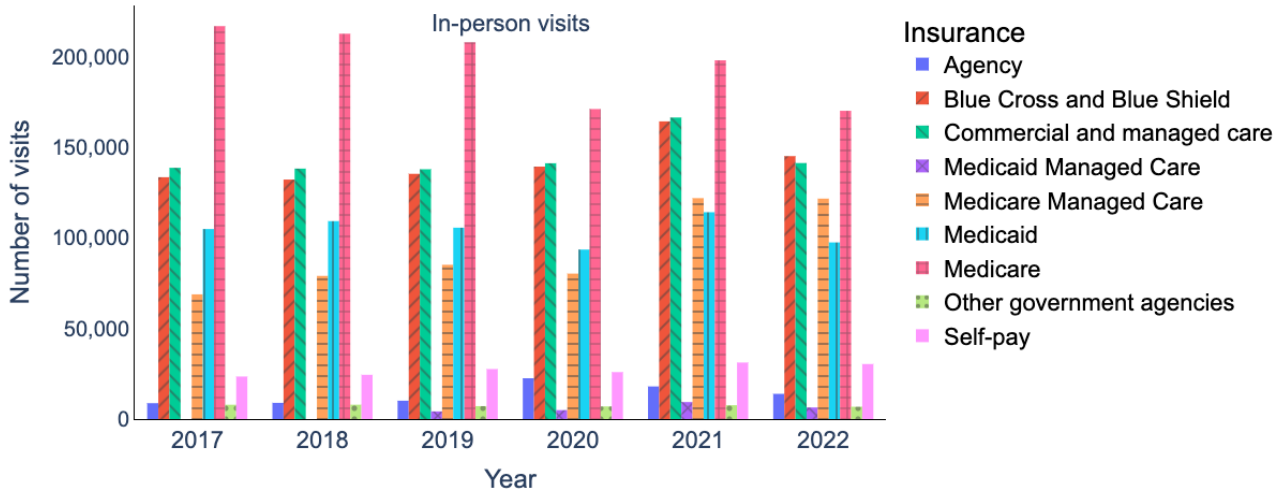

Table 2. Descriptive statistics for insurance coverage (sample size: n=9 insurance providers).

Visit type	Visits (n) conducted in 2019 across insurance providers, mean (SD)			
	2019	2020	2021	2022
Telehealth	323 (376)	4372 (3781)	6145 (5227)	5530 (4544)
In-person	80,153 (72,629)	76,239 (64,081)	92,387 (76,280)	81,506 (66,888)

Table 3. Hypothesis testing results for insurance coverage for the average number of visits conducted across insurance providers in 2019, 2020, 2021, and 2022 (μ^{2019}, μ^{2020}, μ^{2021}, and μ^{2022}, respectively).

Visit type	H_a (*P* value)					
	$\mu^{2019} \neq \mu^{2020}$	$\mu^{2019} < \mu^{2020}$	$\mu^{2020} \neq \mu^{2021}$	$\mu^{2020} < \mu^{2021}$	$\mu^{2021} \neq \mu^{2022}$	$\mu^{2021} > \mu^{2022}$
Telehealth	.004	.002	.004	.002	.06	—[e]
In-person	.43	—	.010	.007	.02	.001

[a]Not applicable.

During 2020 to 2022, telehealth visits covered by these insurance providers surged (Figure 2). *Medicare* telehealth visits increased by 19.3%, going from 9101 to 10,856. Telehealth visits covered by the *Blue Cross* and *Blue Shield* increased by 31.1%, going from 8042 to 10,542. Telehealth visits covered by *commercial and managed care* increased by 16%, growing from 8544 to 9908. Medicaid telehealth visits experienced a slight increase from 6889 to 7206, amounting to a 4.6% increase. *MedicareManaged Care* telehealth visits nearly doubled, increasing from 4013 to 7512, a notable 87.6% increase.

In contrast, in-person visits covered by these 5 insurance providers experienced different trends (Figure 3). During 2020 to 2022, in-person visits covered by *Medicare* decreased by 0.6%, going from 171,200 to 170,224. In-person visits covered by the *Blue Cross* and *Blue Shield* increased by 4.2%, going from 139,406 to 145,20. Similarly, in-person visits covered by *Medicaid* increased by 4.2%, going from 93,600 to 97,540. In-person visits covered by *commercial and managed care* increased by 0.1%, going from 141,257 to 141,368. *Medicare Managed Care* in-person visits saw a significant increase of 51.6%, going from 80,323 to 121,734.

Figure 9 provides a summary of the number of telehealth visits categorized by the insurance provider and specialty from 2020 to 2022. Specifically in Arkansas, the *Medicare* plan covers

XSL•FO
RenderX

adults aged >65 years and younger adults with a disability [41]. In *gerontology*, *neurology*, and *neurosurgery*, the highest number of visits was covered by *Medicare*, accounting for 60.81% (3754/6173), 28.14% (4409/15,667), and 24% (1484/6181), respectively. It is worth noting that a significant proportion of patients visiting these specialties are retirees and individuals with disabilities.

Figure 9. Insurance coverage of telehealth visits for the selected health care specialty, HealthNow (2020-2022). OB/GYN: obstetrician and gynecologist.

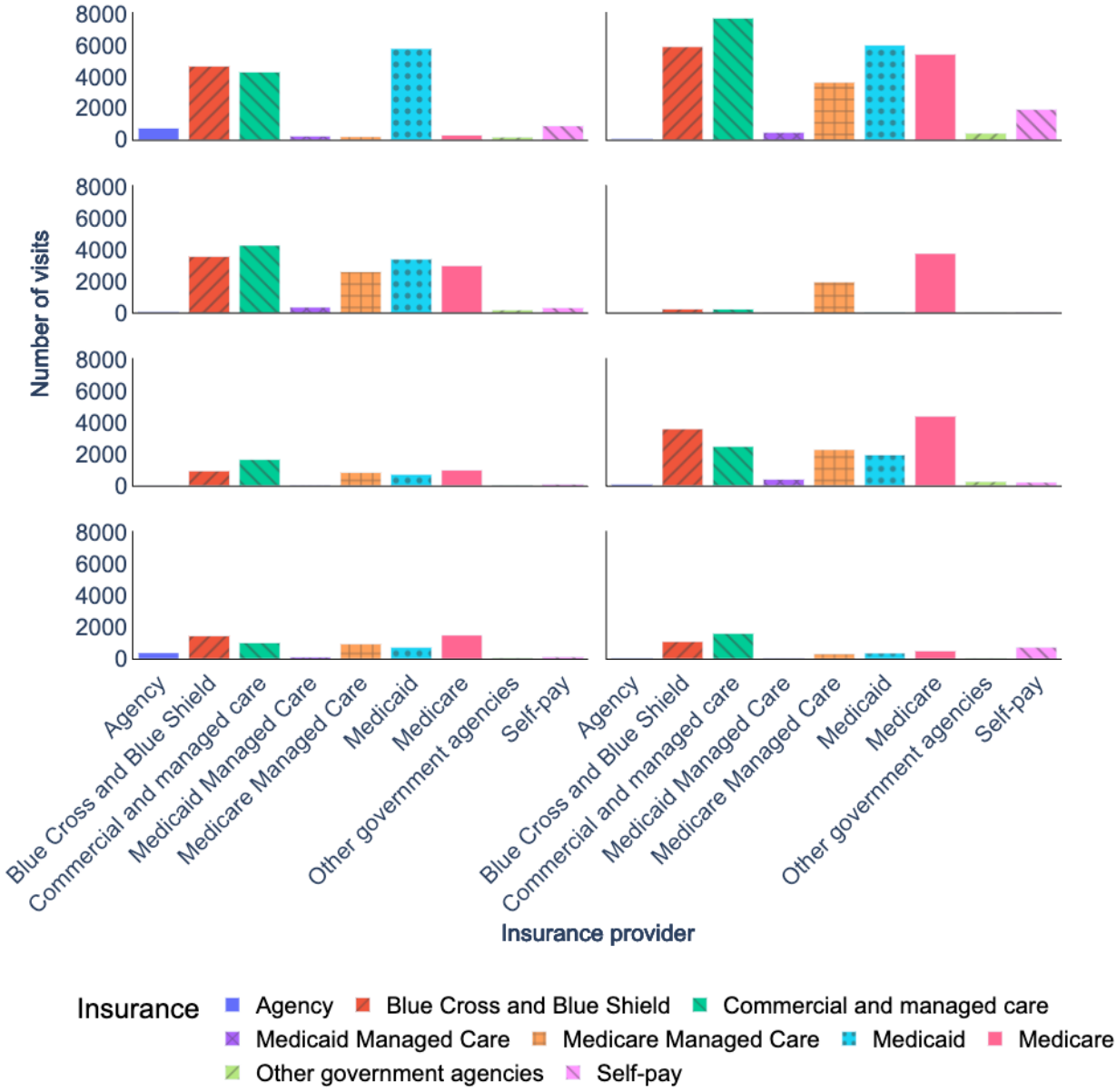

Moreover, in Arkansas, the *Medicaid* program covers pregnant women and children from low-income families [42]. Approximately 27% of Arkansas' population is covered by *Medicaid* or *CHIP*, and approximately 35.67% (5312/14,892) of pregnant women who use *OB/GYN* services are unemployed [43]. This explains why *Medicaid* covers a large percentage (5797/17,120, 33.86%) of *OB/GYN* visits.

Our analysis revealed that approximately 15.58% (718/4607) of *HealthNow* visits, 6.07% (1913/31,516) of *psychiatry* visits, and 5% (858/17,130) of *OB/GYN* visits were self-pay. *HealthNow* services have proven attractive to patients who are willing to self-pay, as they offer accessible health care services 24/7 [22]. Patients also choose to self-pay for psychotherapy services due to limited coverage by insurance companies for certain psychiatric treatments [44]. In addition, UAMS Health offers uninsured patients a 60% discount, which may explain why patients opt to self-pay to receive essential prenatal care visits necessary for a healthy pregnancy [45,46].

Appointment Scheduling

Figures 10-12 depict the average indirect waiting time, direct waiting time, and appointment length for each specialty for in-person and telehealth visits. In these figures, the bars with solid colors represent performance measures for in-person visits and the bars with patterns represent performance measures for

XSL•FO
RenderX

telehealth visits. The mean values, SDs, and sample sizes (presented as *n*) of these appointment scheduling metrics are indicated in Tables 4-12. These metrics were obtained through an analysis of the processed data, as illustrated in Figure 5. The result of the hypothesis tests and the corresponding *P* values are also presented in Tables 4-12.

Table 4 compares indirect waiting times for in-person and telehealth visits by specialty for the years 2020, 2021, and 2022. Tables 7 and 10 present similar information but for direct waiting time and appointment length, respectively.

Table 4 examines whether there were statistically significant differences in indirect waiting times of in-person visits between 2021 and 2020 and between 2022 and 2021. Tables 8 and 11 are similar to Table 5 but for direct waiting times and appointment length, respectively. Tables 6, 9, and 12 evaluate whether there were statistically significant differences between 2021 and 2020 and between 2022 and 2021 in indirect waiting times, direct waiting times, and appointment length of telehealth visits, respectively.

UAMS Health experienced changes in average waiting times and appointment length for both telehealth and in-person visits between 2020 and 2022.

In 2020, the average indirect waiting time for telehealth visits for these specialties was 48.36 days, while the corresponding indirect waiting time for in-person visits was 37.99 days. In 2021, the average indirect waiting time for telehealth visits decreased by 5.6 days, and in 2022, it decreased even further by 15.2 days. However, in 2021, the average indirect waiting time for in-person visits increased by 3.78 days, and in 2022, it decreased by 7.05 days. Overall, the average indirect waiting time at UAMS Health in 2022 decreased (compared to 2020) by 20.8 days for telehealth visits and by 3.27 days for in-person visits.

Similarly, in 2020, the average direct waiting time for telehealth visits was 14.49 minutes, and the corresponding direct waiting time for in-person visits was 12.82 minutes. In 2021, the average direct waiting time for telehealth visits increased by 3.38 minutes, and in 2022, it increased further by 3.6 minutes. The average direct waiting time for in-person visits increased by 2.75 minutes in 2021 but decreased by 1 minute in 2022. As a result, the overall average direct waiting time at UAMS Health in 2022 increased by 6.99 minutes for telehealth visits and by 1.75 minutes for in-person visits.

Finally, in 2020, the average appointment length for telehealth visits was 93.24 minutes, and for in-person visits, it was 75.39 minutes. In 2021, the average appointment length for telehealth visits decreased by 28.3 minutes, and in 2022, it decreased by an additional 25.35 minutes. By contrast, the average appointment length for in-person visits decreased by 0.38 minutes in 2021 and further decreased by 1.64 minutes in 2022. Overall, the average appointment length at UAMS Health in 2022 decreased by 53.65 minutes for telehealth visits and by 2.02 minutes for in-person visits.

Figure 10. Indirect waiting time (days) for in-person and telehealth visits (2020-2022), OB/GYN: obstetrician and gynecologist.

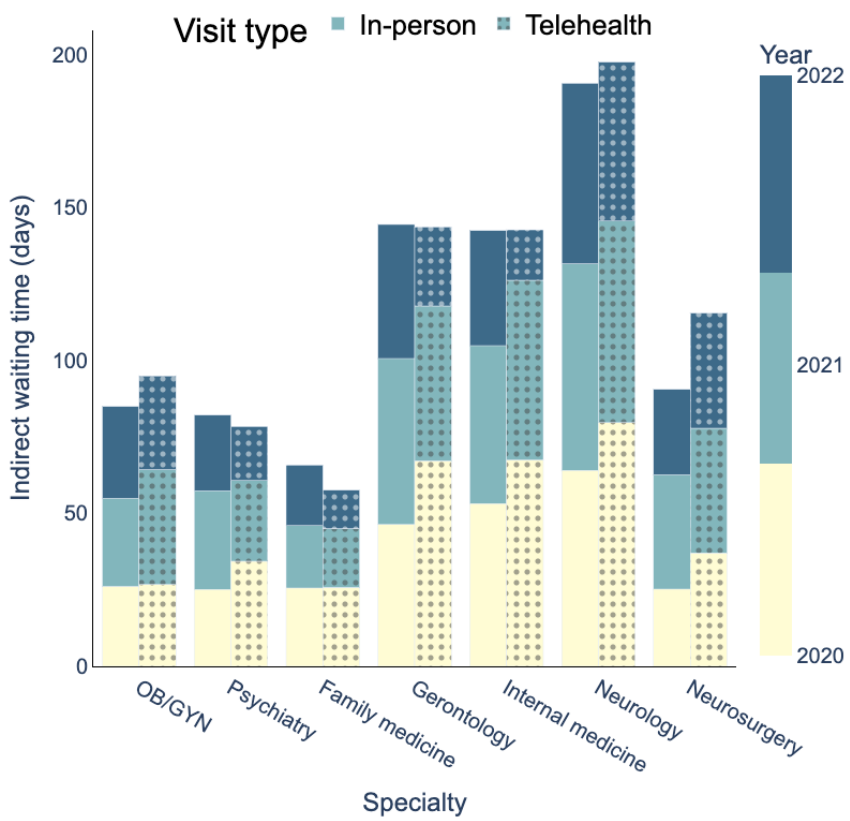

XSL•FO
RenderX

Figure 11. Direct waiting time (minutes) for in-person and telehealth visits (2020-2022), OB/GYN: obstetrician and gynecologist.

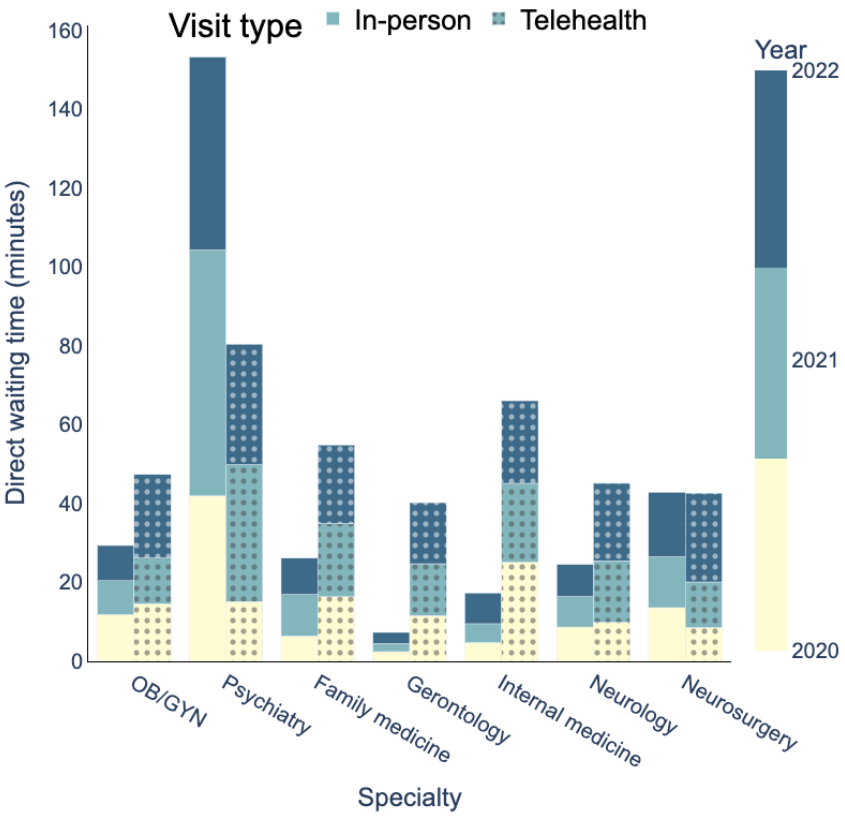

Figure 12. Appointment length (minutes) for in-person and telehealth visits (2020-2022), OB/GYN: obstetrician and gynecologist.

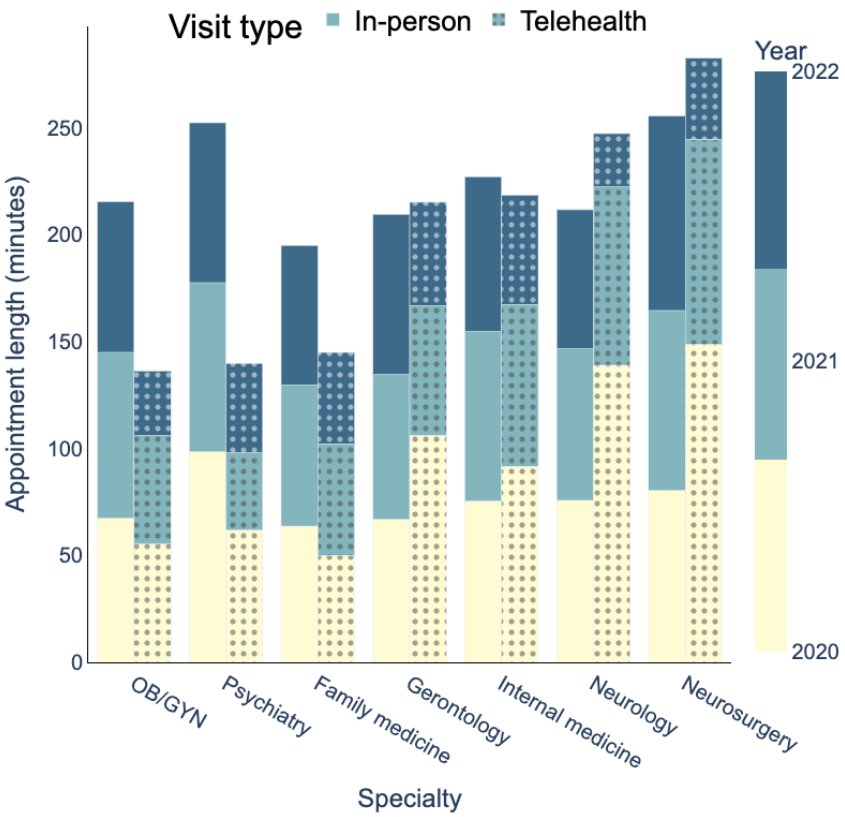

XSL•FO
RenderX

Table 4. Descriptive statistics and hypothesis testing results for mean indirect waiting time for in-person visits ($\mu^{i,ind}$) vs telehealth visits ($\mu^{t,ind}$)

Year and specialty	$\mu^{i,ind}$		$\mu^{t,ind}$		H_a (*P* value)	
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{i,ind} \neq \mu^{t,ind}$	$\mu^{i,ind} > \mu^{t,ind}$
2020						
Obstetrics and gynecology	26.1 (38.2)	33,580	26.8 (29.6)	6436	.21	—[a]
Psychiatry	25.2 (33.0)	7481	34.4 (36.8)	7459	<.001	.99
Family medicine	25.6 (42.5)	37,548	25.8 (41.2)	3539	.79	—
Gerontology	46.4 (64.8)	10,280	67.2 (67.9)	2144	<.001	.99
Internal medicine	53.3 (68.3)	13,198	67.5 (72.7)	1477	<.001	.99
Neurology	64.1 (60.0)	10,321	79.7 (65.0)	4100	<.001	.99
Neurosurgery	25.3 (46.7)	1750	37.1 (49.1)	2147	<.001	.99
2021						
Obstetrics and gynecology	28.8 (39.7)	36,857	37.6 (36.8)	4918	<.001	.99
Psychiatry	32.3 (37.1)	9883	26.4 (32.2)	8595	<.001	<.001
Family medicine	20.4 (35.7)	120,737	19.2 (32.9)	11,386	.001	<.001
Gerontology	54.3 (70.0)	14,706	50.6 (67.0)	2163	.02	.01
Internal medicine	51.6 (70.9)	19,585	58.8 (73.0)	2195	<.001	.99
Neurology	67.6 (59.2)	16,361	66.0 (58.4)	6076	.07	—
Neurosurgery	37.4 (64.7)	2098	40.8 (41.9)	2525	.04	.98
2022						
Obstetrics and gynecology	30.0 (31.5)	33,815	30.7 (27.5)	4043	.23	—
Psychiatry	24.7 (30.4)	11,793	17.7 (23.4)	8303	<.001	<.001
Family medicine	19.8 (33.1)	110,954	12.6 (23.8)	8143	<.001	<.001
Gerontology	43.9 (56.5)	11,221	25.9 (42.4)	1709	<.001	<.001
Internal medicine	37.7 (55.6)	18,705	16.5 (33.5)	1404	<.001	<.001
Neurology	59.0 (53.4)	14,775	52.0 (53.6)	5054	<.001	<.001
Neurosurgery	27.9 (36.4)	2383	37.7 (39.1)	1479	<.001	.99

[a]Not available.

XSL•FO
RenderX

Table 5. Descriptive statistics and hypothesis testing results for indirect waiting time for in-person visits for 2020 ($\mu^{i,ind,20}$), 2021 ($\mu^{i,ind,21}$), and 2022 ($\mu^{i,ind,22}$).

Specialty	$\mu^{i,ind,20}$		$\mu^{i,ind,21}$		$\mu^{i,ind,22}$		H_a (*P* value)			
	Days, mean (SD)	Visits, n	Days, mean (SD)	Visits, n	Days, mean (SD)	Visits, n	$\mu^{i,ind,20} \neq \mu^{i,ind,21}$	$\mu^{i,ind,20} > \mu^{i,ind,21}$	$\mu^{i,ind,21} \neq \mu^{i,ind,22}$	$\mu^{i,ind,21} > \mu^{i,ind,22}$
Obstetrics and gynecology	26.1 (38.2)	33,580	28.8 (39.7)	36,857	30.0 (31.5)	33,815	<.001	.99	<.001	.99
Psychiatry	25.2 (33.0)	7481	32.3 (37.1)	9883	24.7 (30.4)	11,793	<.001	.99	<.001	<.001
Family medicine	25.6 (42.5)	37,548	20.4 (35.7)	120,737	19.8 (33.1)	110,954	<.001	<.001	<.001	<.001
Gerontology	46.4 (64.8)	10,280	54.3 (70.0)	14,706	43.9 (56.5)	11,221	<.001	.99	<.001	<.001
Internal medicine	53.3 (68.3)	13,198	51.6 (70.9)	19,585	37.7 (55.6)	18,705	<.001	.02	<.001	<.001
Neurology	64.1 (60.0)	10,321	67.6 (59.2)	16,361	59.0 (53.4)	14,775	<.001	.99	<.001	<.001
Neurosurgery	25.3 (46.7)	1750	37.4 (64.7)	2098	27.9 (36.4)	2383	.90	—[a]	<.001	<.001

[a]Not applicable.

Table 6. Descriptive statistics and hypothesis testing results for indirect waiting time for telehealth visits for 2020 ($\mu^{t,ind,20}$), 2021 ($\mu^{t,ind,21}$), and 2022 ($\mu^{t,ind,22}$).

Specialty	$\mu^{t,ind,20}$		$\mu^{t,ind,21}$		$\mu^{t,ind,22}$		H_a (*P* value)			
	Days, mean (SD)	Visits, n	Days, mean (SD)	Visits, n	Days, mean (SD)	Visits, n	$\mu^{t,ind,20} \neq \mu^{t,ind,21}$	$\mu^{t,ind,20} > \mu^{t,ind,21}$	$\mu^{t,ind,21} \neq \mu^{t,ind,22}$	$\mu^{t,ind,21} > \mu^{t,ind,22}$
Obstetrics and gynecology	26.8 (29.6)	6436	37.6 (36.8)	4918	30.7 (27.5)	4043	<.001	.99	<.001	<.001
Psychiatry	34.4 (36.8)	7459	26.4 (32.2)	8595	17.7 (23.4)	8303	<.001	<.001	<.001	<.001
Family medicine	25.8 (41.2)	3539	19.2 (32.9)	11,386	12.6 (23.8)	8143	<.001	<.001	<.001	<.001
Gerontology	67.2 (67.9)	2144	50.6 (67.0)	2163	25.9 (42.4)	1709	<.001	<.001	<.001	<.001
Internal medicine	67.5 (72.7)	1477	58.8 (73.0)	2195	16.5 (33.5)	1404	<.001	<.001	<.001	<.001
Neurology	79.7 (65.0)	4100	66.0 (58.4)	6076	52.0 (53.6)	5054	<.001	<.001	<.001	<.001
Neurosurgery	37.1 (49.1)	2147	40.8 (41.9)	2525	37.7 (39.1)	1479	.08	—[a]	.02	.01

[a]Not applicable.

XSL•FO
RenderX

Table 7. Descriptive statistics and hypothesis testing results for direct waiting time in-person ($\mu^{i,dir}$) and for telehealth ($\mu^{t,dir}$).

Year and specialty	$\mu^{i,dir}$		$\mu^{t,dir}$		H_a (*P* value)	
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{i,dir} \neq \mu^{t,dir}$	$\mu^{i,dir} > \mu^{t,dir}$
2020						
Obstetrics and gynecology	11.9 (19.7)	40,859	14.6 (52.3)	4485	<.001	.99
Psychiatry	42.0 (66.9)	2261	15.2 (47.8)	380	<.001	<.001
Family medicine	6.4 (13.3)	33,312	16.4 (29.4)	1827	<.001	.99
Gerontology	2.4 (10.4)	11,039	11.6 (38.7)	1530	<.001	.99
Internal medicine	4.8 (19.2)	13,296	25.2 (47)	927	<.001	.99
Neurology	8.7 (12.9)	10,027	9.9 (30)	2768	.002	.99
Neurosurgery	13.6 (21.3)	2130	8.6 (36.1)	1683	<.001	<.001
2021						
Obstetrics and gynecology	8.6 (21.5)	47,900	11.6 (32)	4004	<.001	.99
Psychiatry	62.4 (81)	6219	34.7 (51.1)	5156	<.001	<.001
Family medicine	10.5 (25.4)	101,967	18.5 (36.7)	8119	<.001	.99
Gerontology	2.1 (8.6)	15,681	13.1 (27.8)	1782	<.001	.99
Internal medicine	4.8 (16.5)	16,404	20.0 (38.2)	1738	<.001	.99
Neurology	7.7 (11.3)	15,292	15.6 (30.3)	5250	<.001	.99
Neurosurgery	12.9 (20.5)	1933	11.5 (33.1)	2328	.11	__[a]
2022						
Obstetrics and gynecology	8.9 (20.9)	33,741	21.2 (23.6)	2335	<.001	.99
Psychiatry	48.9 (68.33)	5744	30.6 (30.7)	10,990	<.001	<.001
Family medicine	9.2 (17.5)	112,516	19.9 (36.4)	7552	<.001	.99
Gerontology	2.8 (11)	13,932	15.5 (26.2)	1692	<.001	.99
Internal medicine	7.8 (17.8)	20,248	21.0 (40)	1031	<.001	.99
Neurology	8.2 (11.3)	16,725	19.7 (24.1)	5671	<.001	.99
Neurosurgery	16.3 (27.3)	2507	22.6 (43.7)	1466	<.001	.99

[a]Not applicable.

XSL•FO
RenderX

Table 8. Descriptive statistics and hypothesis testing results for direct waiting time for in-person visits for 2020 ($\mu^{i,dir,20}$), 2021 ($\mu^{i,dir,21}$), and 2022 ($\mu^{i,dir,22}$).

Specialty	$\mu^{i,dir,20}$		$\mu^{i,dir,21}$		$\mu^{i,dir,22}$		H_a (*P* value)			
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{i,dir,20} \neq \mu^{i,dir,21}$	$\mu^{i,dir,20} > \mu^{i,dir,21}$	$\mu^{i,dir,21} \neq \mu^{i,dir,22}$	$\mu^{i,ind,21} > \mu^{i,ind,22}$
Obstetrics and gynecology	11.9 (19.7)	40,859	8.6 (21.5)	47,900	8.9 (20.9)	33,741	<.001	<.001	.02	.99
Psychiatry	42.0 (66.9)	2261	62.4 (81)	6219	48.9 (68.33)	5744	<.001	.99	<.001	<.001
Family medicine	6.4 (13.3)	33,312	10.5 (25.4)	101,967	9.2 (17.5)	112,516	<.001	.99	<.001	<.001
Gerontology	2.4 (10.4)	11,039	2.1 (8.6)	15,681	2.8 (11)	13,932	.006	<.001	<.001	.99
Internal medicine	4.8 (19.2)	13,296	4.8 (16.5)	16,404	7.8 (17.8)	20,248	.83	—[a]	<.001	.99
Neurology	8.7 (12.9)	10,027	7.7 (11.3)	15,292	8.2 (11.3)	16,725	<.001	<.001	.001	.99
Neurosurgery	13.6 (21.3)	2130	12.9 (20.5)	1933	16.3 (27.3)	2507	.28	—	<.001	.99

[a]Not applicable.

Table 9. Descriptive statistics and hypothesis testing results for direct waiting time for telehealth visits for 2020 ($\mu^{t,dir,20}$), 2021 ($\mu^{t,dir,21}$), and 2022 ($\mu^{t,dir,22}$).

Specialty	$\mu^{t,dir,20}$		$\mu^{t,dir,21}$		$\mu^{t,dir,22}$		H_a (*P* value)			
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{t,dir,20} \neq \mu^{t,dir,21}$	$\mu^{t,dir,20} > \mu^{t,dir,21}$	$\mu^{t,dir,21} \neq \mu^{t,dir,22}$	$\mu^{t,dir,21} > \mu^{t,dir,22}$
Obstetrics and gynecology	14.6 (52.3)	4485	11.6 (32.0)	4004	21.2 (23.6)	2335	.002	<.001	<.001	.99
Psychiatry	15.2 (47.8)	380	34.7 (51.1)	5156	30.6 (30.7)	10,990	<.001	.99	<.001	<.001
Family medicine	16.4 (29.4)	1827	18.5 (36.7)	8119	19.9 (36.4)	7552	.02	.99	.02	.99
Gerontology	11.6 (38.7)	1530	13.1 (27.8)	1782	15.5 (26.2)	1692	.19	—[a]	.01	.99
Internal medicine	25.2 (47.0)	927	20.0 (38.2)	1738	21.0 (40)	1031	.002	<.001	.53	—
Neurology	9.9 (30)	2768	15.6 (30.3)	5250	19.7 (24.1)	5671	<.001	.99	<.001	.99
Neurosurgery	8.6 (36.1)	1683	11.5 (33.1)	2328	22.6 (43.7)	1466	.007	.99	<.001	.99

[a]Not applicable.

XSL•FO
RenderX

Table 10. Descriptive statistics and hypothesis testing results for appointment length in-person ($\mu^{i,app}$) and for telehealth ($\mu^{t,app}$).

Year and specialty	$\mu^{i,app}$		$\mu^{t,app}$		H_a (*P* value)	
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{i,app} \neq \mu^{t,app}$	$\mu^{i,app} > \mu^{t,app}$
2020						
Obstetrics and gynecology	67.4 (45.6)	40,246	55.3 (88.4)	99	.09	.004
Psychiatry	98.4 (87.7)	1769	61.8 (74.1)	63	.001	.001
Family medicine	63.7 (34.3)	33,423	50.0 (78.1)	960	<.001	<.001
Gerontology	66.8 (35.2)	9825	106.2 (100.4)	290	<.001	.99
Internal medicine	75.4 (43.3)	13,224	91.7 (123.1)	449	<.001	.99
Neurology	75.8 (48.8)	9911	139.0 (96.8)	2524	<.001	.99
Neurosurgery	80.3 (52.9)	2080	148.7 (90.1)	1523	<.001	.99
2021						
Obstetrics and gynecology	77.6 (51.6)	47,879	50.7 (91.2)	1051	<.001	<.001
Psychiatry	79.1 (85.6)	5105	36.2 (39.3)	2152	<.001	<.001
Family medicine	66.1 (39.2)	106,370	52.2 (76.2)	5535	<.001	<.001
Gerontology	67.8 (34)	13,741	60.6 (67.1)	1081	<.001	<.001
Internal medicine	79.3 (43.2)	16,466	75.7 (103)	1195	.02	.008
Neurology	70.9 (43.4)	15,247	83.4 (81.2)	4388	<.001	.99
Neurosurgery	84.2 (51.8)	1878	95.8 (74.3)	1682	<.001	.99
2022						
Obstetrics and gynecology	70.4 (38.1)	33,291	30.2 (71.7)	878	<.001	<.001
Psychiatry	74.8 (77.6)	5279	41.8 (38.1)	5907	<.001	<.001
Family medicine	65.2 (37.6)	110,065	42.7 (58.8)	6246	<.001	<.001
Gerontology	74.9 (38.6)	12,603	48.4 (51.4)	1517	<.001	<.001
Internal medicine	72.4 (38.6)	19,824	51.1 (66.6)	838	<.001	<.001
Neurology	65.0 (38.1)	16,635	24.9 (40.1)	3835	<.001	<.001
Neurosurgery	90.9 (53.7)	2436	38.0 (42.2)	782	<.001	<.001

Table 11. Descriptive statistics and hypothesis testing results for mean appointment length for in-person visits for 2020 ($\mu^{i,app,20}$), 2021 ($\mu^{i,app,21}$), and 2022 ($\mu^{i,app,22}$).

Specialty	$\mu^{i,app,20}$		$\mu^{i,app,21}$		$\mu^{i,app,22}$		H_a (*P* value)			
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{i,app,20} \neq \mu^{i,app,21}$	$\mu^{i,app,20} > \mu^{i,app,21}$	$\mu^{i,app,21} \neq \mu^{i,app,22}$	$\mu^{i,app,21} > \mu^{i,app,22}$
Obstetrics and gynecology	67.4 (45.6)	40,246	77.6 (51.6)	47,879	70.4 (38.1)	33,291	<.001	1.0	<.001	<.001
Psychiatry	98.4 (87.7)	1769	79.1 (85.6)	5105	74.8 (77.6)	5279	<.001	<.001	.007	<.001
Family medicine	63.7 (34.3)	33,423	66.1 (39.2)	106,370	65.2 (37.6)	110,065	<.001	.99	<.001	<.001
Gerontology	66.8 (35.2)	9825	67.8 (34)	13,741	74.9 (38.6)	12,603	.02	.99	<.001	.99
Internal medicine	75.4 (43.3)	13,224	79.3 (43.2)	16,466	72.4 (38.6)	19,824	<.001	.99	<.001	<.001
Neurology	75.8 (48.8)	9911	70.9 (43.4)	15,247	65.0 (38.1)	16,635	<.001	<.001	<.001	<.001
Neurosurgery	80.3 (52.9)	2080	84.2 (51.8)	1878	90.9 (53.7)	2436	.02	.99	<.001	.99

XSL•FO
RenderX

Table 12. Descriptive statistics and hypothesis testing results for appointment length for telehealth visits for 2020 ($\mu^{t,app,20}$), 2021 ($\mu^{t,app,21}$), and 2022 ($\mu^{t,app,22}$).

Specialty	$\mu^{t,app,20}$		$\mu^{t,app,21}$		$\mu^{t,app,22}$		H_a (*P* value)			
	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	Minutes, mean (SD)	Visits, n	$\mu^{t,app,20} \neq \mu^{t,app,21}$	$\mu^{t,app,20} > \mu^{t,app,21}$	$\mu^{t,app,21} \neq \mu^{t,app,22}$	$\mu^{t,app,21} > \mu^{t,app,22}$
Obstetrics and gynecology	55.3 (88.4)	99	50.7 (91.2)	1051	30.2 (71.7)	878	.63	—[a]	<.001	<.001
Psychiatry	61.8 (74.1)	63	36.2 (39.3)	2152	41.8 (38.1)	5907	<.001	<.001	<.001	.99
Family medicine	50.0 (78.1)	960	52.2 (76.2)	5535	42.7 (58.8)	6246	.41	—	<.001	<.001
Gerontology	106.2 (100.4)	290	60.6 (67.1)	1081	48.4 (51.4)	1517	<.001	<.001	<.001	<.001
Internal medicine	91.7 (123.1)	449	75.7 (103)	1195	51.1 (66.6)	838	.008	<.001	<.001	<.001
Neurology	139.0 (96.8)	2524	83.4 (81.2)	4388	24.9 (40.1)	3835	<.001	<.001	<.001	<.001
Neurosurgery	148.7 (90.1)	1523	95.8 (74.3)	1682	38.0 (42.2)	782	<.001	<.001	<.001	<.001

[a]Not applicable.

Discussion

Principal Findings

This study explores trends in telehealth use across specialties and insurance coverage in Arkansas by analyzing historical electronic medical records. We observed a significant increase in telehealth use in 2020 due to the COVID-19 pandemic. This increase is not uniform across different specialties. We focused our analysis on those specialties that have the highest number of telehealth visits. Most telehealth services during 2020 to 2022 were covered by *Medicare*, *Blue Cross* and *Blue Shield*, *commercial and managed care*, *Medicaid*, and *Medicare Managed Care*. We used 3 scheduling metrics (ie, indirect waiting time, direct waiting time, and appointment length) to highlight differences between in-person and telehealth services. Our analysis points to the potential benefits of telehealth in providing access to health care, particularly for patients who need psychiatric care.

The relaxation of regulations on remote communication technology for health care services by the Health and Human Services Office for Civil Rights due to the COVID-19 pandemic was one of the main factors that led to the increased use of telehealth during this period. This change in policy led to *Medicare* and *Medicaid* programs, as well as private insurance companies, expanding coverage for telehealth visits. Within the UAMS Health system, during 2020 to 2022, 92.9% (134,221/144,437) of telehealth patients were covered by *Medicare*, *Blue Cross* and *Blue Shield*, *commercial and managed care*, *Medicare*, and *Medicare Managed Care*, which also covered 91.8% (2,067,072/2,251,199) of in-person visits during this period. As a result, telehealth was more commonly used for specialties, such as *OB/GYN*, *psychiatry*, *family medicine*, *gerontology*, *internal medicine*, *neurology*, and *neurosurgery*. Before 2020, *OB/GYN* patients used telehealth visits covered by the ANGELS program, which started in 2003 and provided telemedicine consultations to pregnant patients considered high risk at UAMS.

Since 2020, UAMS Health has been providing both in-person and telehealth visits. Initially, the transition to telehealth caught health care facilities unprepared and employees were not trained, resulting in longer average indirect waiting times and appointment lengths for telehealth visits. In 2020, telehealth visits had an average indirect waiting time that was 10.37 days longer than in-person visits, an average direct waiting time that was 1.66 minutes longer, and an appointment length that was 17.85 minutes longer.

However, in 2021 and 2022, UAMS Health made significant improvements to its telehealth service. The average indirect waiting time decreased by 20.8 days for telehealth visits and 3.27 days for in-person visits compared to 2020, and the average appointment length decreased by 53.65 minutes for telehealth visits and 2.02 minutes for in-person visits. These improvements could be attributed to the flexibility that telehealth visits provide in using valuable resources, such as staff and physician time. These improvements could as well be attributed to the heightened adaptability of patients to technology and telehealth. In addition, *HealthNow* was newly established in January 2020 and may have improved in many ways as program staff and health care providers gained additional experience.

Although the decrease in appointment length might be due to improved usability and familiarity with the service, shorter appointment time may be a cause for concern because patients have less time with their doctors. This could negatively impact the quality of care provided. In addition, the average direct waiting time increased by 6.99 minutes for telehealth visits and 1.75 minutes for in-person visits in 2022.

Overall, UAMS Health has made significant strides in improving its telehealth service since its initial implementation in 2020. The decrease in waiting times and appointment length is a positive development, although the increase in direct waiting times and the potential impact on the quality of care provided require further monitoring.

XSL•FO
RenderX

From 2020 to 2022, *psychiatry* had the highest number of telehealth visits among the different specialties that offer this service. According to the APA, telehealth *psychiatry* visits are just as accurate as in-person visits in terms of diagnosis, treatment effectiveness, quality of care, patient satisfaction, privacy, and confidentiality [38]. This is why the APA and the American Telemedicine Association collaborated to develop best practices guidelines for telemental health delivery [47].

Our study demonstrates that the appointment scheduling metrics for *psychiatry* telehealth visits are shorter than those for in-person visits. This suggests that patients would benefit from continued access to and reimbursement for telehealth *psychiatry* visits. Overall, the high number of telehealth *psychiatry* visits and their comparable diagnostic accuracy and treatment effectiveness to in-person visits highlight the value of telehealth in this field. The APA's best practices guidelines and the shorter appointment scheduling metrics for telehealth *psychiatry* visits provide further support for the continued use and reimbursement of this service.

In 2020, there was a significant decrease in the number of in-person visits for patients covered by *Medicare*. However, in 2021, we observed a trend of increased in-person visits for all patients, including those covered by *Medicare*. Despite this, the number of in-person visits covered by *Medicare* remained lower than that in 2019, which was not the case for visits covered by other insurance providers. In 2022, there was a decrease in the number of in-person visits covered by *Medicare*, *Medicaid*, and *other government agencies* compared to 2019.

These findings indicate that telehealth visits have improved health care accessibility in Arkansas. Many patients continued to use telehealth visits in 2022, even after the end of the COVID-19 public health emergency. Therefore, we suggest that insurance companies and policy makers take note of these observations and continue to cover these services. Telehealth services are vital for patients in Arkansas, especially because 59 out of the total 75 counties are classified as medically underserved and another 15 are classified as partially underserved [48]. These classifications are due to a shortage of primary care health service providers in Arkansas, making it challenging for patients to visit health care providers. This shortage is particularly pressing for patients living in rural areas, people on *Medicaid*, and children in ARKids [49].

Providing access to health care through telehealth reduces these vulnerabilities, making it a crucial aspect of health care in Arkansas. Thus, moving forward, insurance companies will need to (1) rethink their reimbursement policies to ensure payment equity between telehealth and in-person visits [50,51] and (2) change policies, such as lift restrictions on new patients.

Our study used a large amount of data to evaluate essential appointment scheduling metrics for both telehealth and in-person visits, including direct and indirect waiting times and appointment length. In 2022, we found that the average indirect waiting time for telehealth visits was 7.17 days shorter than for in-person visits, while the average direct waiting time was 6.89 minutes longer. In addition, the appointment length was 33.78 minutes shorter for telehealth visits than that for in-person visits. This could be attributed to telehealth visits not requiring a physical room, making them easier to schedule and resulting in shorter indirect waiting times.

Although longer direct waiting times may not be as concerning for telehealth visits, the shorter appointment length may be an area of concern. Further research is needed to identify the reasons behind this difference.

While many recent studies have analyzed telehealth use [2,3,52-57], only a few have focused on evaluating appointment scheduling metrics [58-62]. Previous studies have primarily focused on appointment adherence rather than appointment duration and waiting time. Our study provides crucial insights into comparing the efficiency of telehealth and in-person visits and can motivate changes in current practices to improve patient experience.

Limitations

Although our study provides valuable insights into telehealth visits for UAMS Health, it has a few limitations that should be acknowledged. First, the results of our analysis may not be generalizable to other health care providers in Arkansas, as their patient populations and care settings may differ. Nonetheless, our research framework can be applied to other health care providers to assess their telehealth visits. Second, our study used data from multiple health care providers at UAMS Health, with some offering only telehealth services and others providing a mix of in-person and telehealth visits. Therefore, our results reflect an overall evaluation of telehealth and in-person visits for all health care providers at UAMS Health and may not be representative of individual health care providers. Finally, we excluded records with incomplete or inaccurate time stamps from our analysis, which may have limited our sample size and introduced bias in our metrics calculations. Future research could address these limitations by using more comprehensive and accurate datasets and by comparing the performance of telehealth visits across different health care providers and patient populations.

Conclusions

In 2020, telehealth services in the United States saw an uptick due to policy changes allowing insurance companies to reimburse for these services. Most telehealth services during 2020 to 2022 were covered by *Medicare*, *Blue Cross* and *Blue Shield*, *commercial and managed care*, *Medicaid*, and *Medicare Managed Care*. Our analysis of electronic medical records from 2017 to 2022 collected by UAMS Health shows that telehealth use increased in Arkansas. This increase in telehealth services provided better health care access to individuals living in medically underserved and rural areas.

Our observations indicate that there was a shift in patient visits from in-person to telehealth visits during 2020 and 2021 for *psychiatry*, *family medicine*, *gerontology*, *internal medicine*, *neurology*, and *neurosurgery* visits. However, we did not observe this shift for *OB/GYN patients*. The reason for this is because pregnant women had access to telehealth before 2020 through the ANGELS program, which was established in 2003 and offered telemedicine consultations to UAMS patients in high-risk pregnancies.

XSL•FO
RenderX

When comparing the data from 2022 to 2021, we noticed only minor differences in the number of telehealth and in-person visits. Our findings indicate that *psychiatry* had the highest number of telehealth visits, followed by *OB/GYN* and *family medicine*. Notably, *psychiatry* visits do not require a physical examination and offer comparable diagnostic accuracy, treatment effectiveness, quality of care, patient satisfaction, privacy, and confidentiality as in-person visits.

To evaluate the efficiency of current telehealth appointment scheduling processes at UAMS Health, we compared the performance of telehealth and in-person visits using direct and indirect waiting times and appointment length. Our analysis of these metrics can inform improvements to current practices and motivate changes to optimize the scheduling of telehealth visits.

Future Directions

The scope of our research is impacted by the data available. We were not able to delve deeper in our analysis of telehealth use within each specialty because we did not have access to cultural, demographic, and socioeconomic data about patients. Researchers who have access to such data could use the proposed framework to determine what factors influence patients' decisions to use telehealth.

Our research is limited to the state of Arkansas. The scope of this research can be extended by investigating telehealth use in other states. Such an analysis would determine how the distribution of population between urban and rural areas, access to broadband networks, access to health care, and health care resources available do impact the use of telehealth.

Our study uses 3 metrics to evaluate the performance of telehealth and in-person visits, which are direct and indirect waiting time and appointment duration. Researchers who have access to data about resources available in a health care facility (the number of physicians or nurses and their schedules, the number of rooms, etc) could calculate other relevant metrics, such as resource use rate for in-person and telehealth visits. This information is important to design strategies (develop schedules) for the optimal use of resources and to provide high-quality service to in-person and telehealth patients.

Acknowledgments

The project described was partially supported by the Translational Research Institute (UL1 TR003107) through the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the use of medical data from a health care organization.

Authors' Contributions

ABC, SE, and BE conceptualized the study, determined the methodology used, and conducted the formal analysis of the data. ABC and CJH acquired the datasets. ABC curated and processed the data for analysis. ABC, SE, and BE wrote the original manuscript draft, and all authors reviewed and edited the final manuscript. SE, BE, and HE provided study supervision.

Conflicts of Interest

None declared.

References

1. Declarations of a public health emergency. US Department of Health & Human Services. 2022. URL: https://aspr.hhs.gov/legal/PHE/Pages/default.aspx [accessed 2024-04-29]
2. Koonin LM, Hoots B, Tsang CA, Leroy Z, Farris K, Jolly B, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic - United States, January-March 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 30;69(43):1595-1599 [FREE Full text] [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]
3. Demeke HB, Pao LZ, Clark H, Romero L, Neri A, Shah R, et al. Telehealth practice among health centers during the COVID-19 pandemic - United States, July 11-17, 2020. MMWR Morb Mortal Wkly Rep 2020 Dec 18;69(50):1902-1905 [FREE Full text] [doi: 10.15585/mmwr.mm6950a4] [Medline: 33332297]
4. Baum A, Kaboli PJ, Schwartz MD. Reduced in-person and increased telehealth outpatient visits during the COVID-19 pandemic. Ann Intern Med 2021 Jan;174(1):129-131. [doi: 10.7326/m20-3026]
5. Wosik J, Fudim M, Cameron B, Gellad ZF, Cho A, Phinney D, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 01;27(6):957-962 [FREE Full text] [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
6. Maugeri A, Barchitta M, Basile G, Agodi A. Public and research interest in telemedicine from 2017 to 2022: infodemiology study of Google trends data and bibliometric analysis of scientific literature. J Med Internet Res 2024 May 16;26:e50088 [FREE Full text] [doi: 10.2196/50088] [Medline: 38753427]

XSL•FO
RenderX

7. Medicare telemedicine health care provider fact sheet. Centers for Medicare & Medicaid Services and others. 2020. URL: https://www.cms.gov/newsroom/fact-sheets/medicare-telemedicine-health-care-provider-fact-sheet [accessed 2024-04-29]
8. Telehealth: delivering care safely during COVID-19. US Department of Health and Human Services. URL: https://telehealth.hhs.gov/providers/policy-changes-during-the-covid-19-public-health-emergency/ [accessed 2024-04-29]
9. Adler-Milstein J, Kvedar J, Bates DW. Telehealth among US hospitals: several factors, including state reimbursement and licensure policies, influence adoption. Health Aff (Millwood) 2014 Feb;33(2):207-215. [doi: 10.1377/hlthaff.2013.1054] [Medline: 24493762]
10. Telehealth policy changes after the COVID-19 public health emergency. US Department of Health and Human Services. 2020. URL: https://telehealth.hhs.gov/providers/telehealth-policy/policy-changes-after-the-covid-19-public-health-emergency [accessed 2023-11-15]
11. COVID-19 public health emergency (PHE). US Department of Health and Human Services. URL: https://www.hhs.gov/coronavirus/covid-19-public-health-emergency/index.html [accessed 2023-11-15]
12. Cubanski J, Kates J, Tolbert J. What happens when COVID-19 emergency declarations end? Implications for coverage, costs, and access. KFF. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/what-happens-when-covid-19-emergency-declarations-end-implications-for-coverage-costs-and-access/ [accessed 2023-11-15]
13. What happens to telemedicine after COVID-19? Association of American Medical Colleges. 2021. URL: https://www.aamc.org/news-insights/what-happens-telemedicine-after-covid-19 [accessed 2024-04-29]
14. Mallow J, Davis S. Health insurers are starting to roll back coverage for telehealth – even though demand is way up due to COVID-19. PBS News. URL: https://www.pbs.org/newshour/health/health-insurers-are-starting-to-roll-back-coverage-for-telehealth-even-though-demand-is-way-up-due-to-covid-19 [accessed 2023-11-14]
15. Adams K. 10 biggest barriers to telehealth, as told by physicians. Becker's Healthcare. URL: https://www.beckershospitalreview.com/telehealth/10-biggest-barriers-to-telehealth-as-told-by-physicians.html [accessed 2022-11-30]
16. Andrews E, Berghofer K, Long J, Prescott A, Caboral-Stevens M. Satisfaction with the use of telehealth during COVID-19: an integrative review. Int J Nurs Stud Adv 2020 Nov;2:100008 [FREE Full text] [doi: 10.1016/j.ijnsa.2020.100008] [Medline: 33083791]
17. Odebunmi OO, Hughes TD, Waters AR, Urick BY, Herron C, Wangen M, et al. Findings from a national survey of older US adults on patient willingness to use telehealth services: cross-sectional survey. J Med Internet Res 2024 May 23;26:e50205 [FREE Full text] [doi: 10.2196/50205] [Medline: 38780994]
18. Schriger SH, Klein MR, Last BS, Fernandez-Marcote S, Dallard N, Jones B, et al. Community mental health clinicians' perspectives on telehealth during the COVID-19 pandemic: mixed methods study. JMIR Pediatr Parent 2022 Mar 03;5(1):e29250 [FREE Full text] [doi: 10.2196/29250] [Medline: 35023839]
19. Phenicie R, Acosta Wright R, Holzberg J. Patient satisfaction with telehealth during COVID-19: experience in a rural county on the United States-Mexico border. Telemed J E Health 2021 Aug 01;27(8):859-865. [doi: 10.1089/tmj.2021.0111] [Medline: 34096790]
20. UAMS physician's call center. UAMS Institute for Digital Health & Innovation. 2022. URL: https://idhi.uams.edu/physicians-call-center/ [accessed 2024-04-28]
21. Boulden B. Now offers 24/7 digital health, live video access to convenient care. UAMS News. URL: https://news.uams.edu/2020/01/21/new-uams-healthnow-offers-24-7-digital-health-live-video-access-to-convenient-care/ [accessed 2024-04-28]
22. Digital health and innovation. UAMSHealth. URL: https://uamshealth.com/expertise/digital-health-innovation/ [accessed 2024-04-28]
23. Bain BZ, Engelhardt M. Introduction to Probability and Mathematical Statistics. Volume 4. Pacific Grove, CA: Duxbury Press; 1992.
24. Hailemeskel Abebe TH. The derivation and choice of appropriate test statistic (Z, t, F and Chi-Square test) in research methodology. Math Lett 2019;5(3):33. [doi: 10.11648/j.ml.20190503.11]
25. Lowery C, Bronstein J, McGhee J, Ott R, Reece EA, Mays GP. ANGELS and University of Arkansas for Medical Sciences paradigm for distant obstetrical care delivery. Am J Obstet Gynecol 2007 Jun;196(6):534.e1-534.e9. [doi: 10.1016/j.ajog.2007.01.027] [Medline: 17547884]
26. Melchionna M. New data shows telehealth usage drops by 4% nationally. TechTarget. URL: https://tinyurl.com/3s2meydz [accessed 2023-03-28]
27. Donelan K, Barreto EA, Sossong S, Michael C, Estrada JJ, Cohen AB, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44 [FREE Full text] [Medline: 30667610]
28. Verma S. Early impact of CMS expansion of Medicare telehealth during COVID-19. Health Affairs Forefront. 2020. URL: https://www.healthaffairs.org/content/forefront/early-impact-cms-expansion-medicare-telehealth-during-covid-19 [accessed 2024-04-29]
29. Jaklevic MC. Telephone visits surge during the pandemic, but will they last? JAMA 2020 Oct 27;324(16):1593-1595. [doi: 10.1001/jama.2020.17201] [Medline: 33027503]

XSL•FO
RenderX

30. Greenwood H, Krzyzaniak N, Peiris R, Clark J, Scott AM, Cardona M, et al. Telehealth versus face-to-face psychotherapy for less common mental health conditions: systematic review and meta-analysis of randomized controlled trials. JMIR Ment Health 2022 Mar 11;9(3):e31780 [FREE Full text] [doi: 10.2196/31780] [Medline: 35275081]
31. Krzyzaniak N, Greenwood H, Scott AM, Peiris R, Cardona M, Clark J, et al. The effectiveness of telehealth versus face-to face interventions for anxiety disorders: a systematic review and meta-analysis. J Telemed Telecare 2021 Dec 03;30(2):250-261 [FREE Full text] [doi: 10.1177/1357633x211053738]
32. Moore MA, Coffman M, Jetty A, Petterson S, Bazemore A. Only 15% of FPs report using telehealth; training and lack of reimbursement are top barriers. Am Fam Physician 2016 Jan 15;93(2):101 [FREE Full text] [Medline: 26926405]
33. Rangachari P, Mushiana SS, Herbert K. A narrative review of factors historically influencing telehealth use across six medical specialties in the United States. Int J Environ Res Public Health 2021 May 08;18(9):4995 [FREE Full text] [doi: 10.3390/ijerph18094995] [Medline: 34066829]
34. De Biase G, Freeman WD, Bydon M, Smith N, Jerreld D, Pascual J, et al. Telemedicine utilization in neurosurgery during the COVID-19 pandemic: a glimpse into the future? Mayo Clin Proc Innov Qual Outcomes 2020 Dec;4(6):736-744 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2020.07.013] [Medline: 33324948]
35. COVID-19: guidance for triage of non-emergent surgical procedures. American College of Surgeons. 2020. URL: https://www.facs.org/for-medical-professionals/covid-19/clinical-guidance/triage/ [accessed 2023-02-11]
36. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
37. Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Variation in telemedicine use and outpatient care during the COVID-19 pandemic in the United States. Health Aff (Millwood) 2021 Feb 01;40(2):349-358 [FREE Full text] [doi: 10.1377/hlthaff.2020.01786] [Medline: 33523745]
38. What is telepsychiatry? American Psychiatric Association. 2020. URL: https://www.psychiatry.org/patients-families/telepsychiatry [accessed 2023-02-11]
39. Doraiswamy S, Jithesh A, Mamtani R, Abraham A, Cheema S. Telehealth use in geriatrics care during the COVID-19 pandemic-a scoping review and evidence synthesis. Int J Environ Res Public Health 2021 Feb 11;18(4):1755 [FREE Full text] [doi: 10.3390/ijerph18041755] [Medline: 33670270]
40. Nie JZ, Karras CL, Texakalidis P, Trybula SJ, Dahdaleh NS. A systematic review of outpatient telemedicine use in neurosurgery since the start of coronavirus disease 2019. World Neurosurg 2022 Nov;167:e1090-e1102. [doi: 10.1016/j.wneu.2022.08.145] [Medline: 36115568]
41. Norris L. Medicare in Arkansas. Health Insurence. 2022. URL: https://www.healthinsurance.org/medicare/arkansas/ [accessed 2024-04-29]
42. Health care programs. Arkansas Department of Human Services. 2022. URL: https://humanservices.arkansas.gov/divisions-shared-services/county-operations/health-care-programs/ [accessed 2024-04-29]
43. Medicaid state fact sheets. Kaiser Family Foundation. URL: https://www.kff.org/interactive/medicaid-state-fact-sheets/ [accessed 2024-04-29]
44. Benjenk I, Chen J. Trends in self-payment for outpatient psychiatrist visits. JAMA Psychiatry 2020 Dec 01;77(12):1305-1307 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.2072] [Medline: 32667644]
45. UAMS financial assistance. UAMSHealth. 2023. URL: https://uamshealth.com/patients-and-guests/uams-financial-assistance/ [accessed 2023-02-13]
46. What is prenatal care and why is it important? National Institutes of Child Health and Human Development. 2017. URL: https://www.nichd.nih.gov/health/topics/pregnancy/conditioninfo/prenatal-care [accessed 2023-02-28]
47. Shore JH, Yellowlees P, Caudill R, Johnston B, Turvey C, Mishkind M, et al. Best practices in videoconferencing-based telemental health April 2018. Telemed J E Health 2018 Nov;24(11):827-832. [doi: 10.1089/tmj.2018.0237] [Medline: 30358514]
48. What is shortage designation? HRSA Health Workforce. 2023. URL: https://bhw.hrsa.gov/workforce-shortage-areas/shortage-designation#mups [accessed 2024-04-29]
49. Norris L. Addressing Arkansas's health services shortages by empowering nurse practitioners. Arkansas Center for Research in Economics UoCA. 2021. URL: https://uca.edu/acre/files/2021/02/ACRE_Health_Services_Shortages.pdf [accessed 2023-02-28]
50. Shachar C, Engel J, Elwyn G. Implications for telehealth in a postpandemic future: regulatory and privacy issues. JAMA 2020 Jun 16;323(23):2375-2376. [doi: 10.1001/jama.2020.7943] [Medline: 32421170]
51. Fiscal considerations for the future of telehealth. Committee for a Responsible Federal Budget. 2022. URL: https://www.crfb.org/papers/fiscal-considerations-future-telehealth [accessed 2023-03-28]
52. Howie F, Kreofsky BL, Ravi A, Lokken T, Hoff MD, Fang JL. Rapid rise of pediatric telehealth during COVID-19 in a large multispecialty health system. Telemed J E Health 2022 Jan 01;28(1):3-10. [doi: 10.1089/tmj.2020.0562] [Medline: 33999718]
53. Mishra S, Dhuna N, Lancki N, Yeh C, Larson DN. Telehealth utilization and patient satisfaction in an ambulatory movement disorders center during the COVID-19 pandemic. J Telemed Telecare 2022 Dec 27:1357633X2211468. [doi: 10.1177/1357633x221146819]

XSL•FO
RenderX

54. Haider Z, Aweid B, Subramanian P, Iranpour F. Telemedicine in orthopaedics during COVID-19 and beyond: a systematic review. J Telemed Telecare 2020 Aug 06;28(6):391-403. [doi: 10.1177/1357633x20938241]
55. Schulz T, Long K, Kanhutu K, Bayrak I, Johnson D, Fazio T. Telehealth during the coronavirus disease 2019 pandemic: rapid expansion of telehealth outpatient use during a pandemic is possible if the programme is previously established. J Telemed Telecare 2020 Jul 19;28(6):445-451. [doi: 10.1177/1357633x20942045]
56. Siddiqui S, Farr E, Dusto N, Chen L, Kocherginsky M, Skelton F, et al. Telemedicine use among physiatrists during the early phase of the COVID-19 pandemic and potential for future use. Telemed J E Health 2023 Feb 01;29(2):242-252 [FREE Full text] [doi: 10.1089/tmj.2022.0030] [Medline: 35833791]
57. Fu R, Sutradhar R, Li Q, Eskander A. Virtual and in-person visits by Ontario physicians in the COVID-19 era. J Telemed Telecare 2022 Mar 16;30(4):706-714 [FREE Full text] [doi: 10.1177/1357633x221086447]
58. Jeganathan S, Prasannan L, Blitz MJ, Vohra N, Rochelson B, Meirowitz N. Adherence and acceptability of telehealth appointments for high-risk obstetrical patients during the coronavirus disease 2019 pandemic. Am J Obstet Gynecol MFM 2020 Nov;2(4):100233 [FREE Full text] [doi: 10.1016/j.ajogmf.2020.100233] [Medline: 32984803]
59. Childs AW, Bacon SM, Klingensmith K, Li L, Unger A, Wing AM, et al. Showing up is half the battle: the impact of telehealth on psychiatric appointment attendance for hospital-based intensive outpatient services during COVID-19. Telemed J E Health 2021 Aug 01;27(8):835-842. [doi: 10.1089/tmj.2021.0028] [Medline: 33999738]
60. Zoran S, Turcott C, Whitehead A, Hrabik L, Harris A, Scott Schwoerer J. Rapid transition to telemedicine during the COVID-19 pandemic: medical genetics experience. WMJ 2021 Oct;120(3):218-221 [FREE Full text] [Medline: 34710304]
61. Snoswell CL, Comans TA. Does the choice between a telehealth and an in-person appointment change patient attendance? Telemed J E Health 2021 Jul 01;27(7):733-738. [doi: 10.1089/tmj.2020.0176] [Medline: 32831007]
62. Adams AM, Wu H, Zhang FR, Wajsberg JR, Bruney TL. Postpartum care in the time of COVID-19: the use of telemedicine for postpartum care. Telemed J E Health 2023 Feb;29(2):235-241. [doi: 10.1089/tmj.2022.0065] [Medline: 35727135]

Abbreviations

ANGELS: Antenatal and Neonatal Guidelines, Education and Learning System
APA: American Psychiatric Association
OB/GYN: obstetrics and gynecology
UAMS: University of Arkansas for Medical Sciences
UAMS Health: University of Arkansas for Medical Sciences Medical Center

Edited by T de Azevedo Cardoso; submitted 21.05.23; peer-reviewed by P Codyre, K Fultz Hollis, M Erkoc, S Matathil, R Talayeh; comments to author 08.11.23; revised version received 14.12.23; accepted 16.08.24; published 18.10.24.

Please cite as:
Cengil AB, Eksioglu S, Eksioglu B, Eswaran H, Hayes CJ, Bogulski CA
Statistical Analysis of Telehealth Use and Pre- and Postpandemic Insurance Coverage in Selected Health Care Specialties in a Large Health Care System in Arkansas: Comparative Cross-Sectional Study
J Med Internet Res 2024;26:e49190
URL: https://www.jmir.org/2024/1/e49190
doi:10.2196/49190
PMID:39423000

©Aysenur Betul Cengil, Sandra Eksioglu, Burak Eksioglu, Hari Eswaran, Corey J Hayes, Cari A Bogulski. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient-, Provider-, and Facility-Level Contributors to the Use of Cardiology Telehealth Care in the Veterans Health Administration: Retrospective Cohort Study

Rebecca Lauren Tisdale[1,2], MPA, MD; Jacqueline M Ferguson[1], PhD; James Van Campen[1], MS; Liberty Greene[1,2], MEd, MS; Charlie M Wray[3,4], MS, DO; Donna M Zulman[1,2], MS, MD

[1]Center for Innovation to Implementation, Veterans Affairs Palo Alto Health Care System, Menlo Park, CA, United States

[2]Division of Primary Care and Population Health, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

[3]Veterans Affairs San Francisco, San Francisco, CA, United States

[4]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Rebecca Lauren Tisdale, MPA, MD
Center for Innovation to Implementation
Veterans Affairs Palo Alto Health Care System
Building 324
795 Willow Road
Menlo Park, CA, 94025
United States
Phone: 1 650 269 5193
Email: rebecca.tisdale2@va.gov

Abstract

Background: Telehealth (care delivered by phone or video) comprises a substantial proportion of cardiology care delivered in the Veterans Health Administration (VHA). Little is known about how factors specific to patients, clinicians, and facilities contribute to variation in cardiology telehealth use.

Objective: The aim of this study is to estimate the relative extent to which patient-, clinician-, and facility-level factors affect cardiology telehealth use in VHA.

Methods: This was a retrospective, nation-wide cohort study of veterans' use of VHA cardiology telehealth care during the first 2 years of the COVID-19 pandemic (March 11, 2020, to March 10, 2022). We constructed multilevel, multivariable, logistic regression models of patient-level cardiology telehealth use (telephone or video-based care). Models included random effects for the patient, the patient's main cardiology provider, and the patient's primary facility (ie, VHA medical center) for specialty care and fixed effects for patient sociodemographic and clinical characteristics.

Results: Our analytic cohort comprised 223,809 veterans with 989,271 encounters among 2235 unique clinicians. The veterans' average age was 70.2 years, and 3.4% (n=7616) were women. Of the 989,271 encounters, 4.2% (n=41,480) were video based and 34.3% (n=338,834) were phone based. Adjusted odds of telehealth use were slightly higher for women versus men (adjusted odds ratio [AOR] 1.08, 95% CI 1.05-1.10), individuals identifying as Hispanic or Latino versus not Hispanic or Latino (AOR 1.46, 95% CI 1.43-1.49), and those with medium and long drive times versus short drive time (AOR 1.11, 95% CI 1.10-1.12 and AOR 1.09, 95% CI 1.07-1.10, respectively). Further, 40.5% of the variation in a veteran's likelihood of using cardiology telehealth care was found at the patient level, 30.8% at the clinician level, and 7% at the facility level.

Conclusions: The largest share of the attributable variability in VHA cardiology telehealth use in this cohort was explained by the patient, followed closely by the clinician. Little variability was attributed to the primary facility through which the veteran received their cardiology care. These results suggest that policy solutions intended to improve equity of cardiology telehealth care use in VHA may be most impactful when directed at patients and clinicians.

(J Med Internet Res 2024;26:e53298) doi:10.2196/53298

KEYWORDS

telehealth; mixed models; veteran; patient; virtual cardiology; cardiology care; cohort study

XSL•FO
RenderX

Introduction

The coronavirus (COVID-19) pandemic has been a catalyst for the expansion of telehealth (medical care delivered by phone or video), in the Veterans Health Administration (VHA) system and in health care settings worldwide [1,2]. While in-person care has largely been restored as the pandemic has subsided, telehealth continues to play a crucial role in improving access to medical care. However, as telehealth shifts from a requirement to an option [3], this role is evolving; beyond replacing an in-person encounter, telehealth can be used to address staffing gaps at a given location. For example, VHA has implemented a hub-and-spoke model across primary, mental health, and specialty care to compensate for staffing shortages, wherein clinicians at the "hub" site see patients at a "spoke" sites via telehealth [4].

For optimal and equitable use, such novel applications of telehealth depend on a thorough understanding of the drivers of telehealth use. Unsurprisingly, these drivers may differ by medical specialty and clinical conditions. Acknowledging this, professional societies have called for more research into determinants of telehealth provision for cardiovascular disease specifically [5]. Use of cardiology telehealth care varies depending on patients' sociodemographic characteristics, such as age, rurality, socioeconomic status, and number of comorbid conditions [6,7]. At the clinician level, significant variability exists regarding comfort level with and support for telehealth adoption among cardiologists [8]. Additionally, facility-level characteristics have also been shown to influence telehealth adoption [9]. Given that all 3 levels are likely influencing telehealth use, understanding the *relative* contribution of each is useful for informing decisions about where to direct policy change for the highest impact [10]. While other studies have examined this question across a range of specialties [11], none has specifically focused on cardiology telehealth care.

In this study, we estimate the variability in cardiology telehealth care use attributable to patients, clinicians, and health care facilities. Understanding the relative contribution of factors at these levels can inform policy initiatives and interventions to promote the optimal and equitable use of telehealth in specialty settings.

Methods

Data

This retrospective cohort study focused on active users of VHA cardiology care. Veterans aged 18 and older were included in the cohort if they had at least 1 outpatient cardiology visit for a cardiology diagnosis (the full list of diagnoses is provided in Table S1 in Multimedia Appendix 1) in the approximately 15-month period leading up the COVID-19 pandemic's onset (calendar year 2019 and the beginning of 2020; January 1, 2019, to March 10, 2020) and had at least 2 outpatient cardiology visits in the first 2 years of the pandemic (March 11, 2020, to March 10, 2022). The first day of the pandemic was considered to be March 11, 2020, consistent with World Health Organization directives [12].

We sourced all data from VHA's Corporate Data Warehouse, a repository for veteran's electronic health care records (172VA10P2: VHA Corporate Data Warehouse - VA 79 FR 4377). In addition to veterans' dates of birth and, when applicable, dates of death, the dataset included all cardiology outpatient encounters in VHA, as defined by VHA-specific "stop codes," where the visits were associated with an independent licensed practitioner with a unique National Provider Identifier (NPI). These 3-digit codes, available through VHA's Managerial Cost Accounting system, characterize all VHA outpatient encounters and associated clinical work units (Table S2 in Multimedia Appendix 1). Stop codes were also used to define whether the visit was conducted in person or via phone or video, with telehealth visits defined as those taking place by phone or video. Visits for cardiac rehabilitation were not included.

We captured veteran sociodemographic and clinical characteristics in a manner consistent with existing VHA telehealth literature [1,2,6]. Age at the beginning of the pandemic was categorized into 4 groups, roughly corresponding to quartiles for the study population: <50, 50-64, 65-74, and ≥75 years. Race was categorized as American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, unknown, or White, and ethnicity was categorized as Hispanic or Latino, not Hispanic or Latino, or unknown, both based on the most frequently recorded race or ethnicity identification in the electronic health record. Separate categories were created for missing race or ethnicity classification. Rurality, defined as highly rural (population density of fewer than 7 people per square mile), rural, or urban, was based on US Census Bureau criteria [13] and derived from VHA's Planning Systems Support Group (PSSG) Geocoded Enrollee Files in the Corporate Data Warehouse. Drive time to secondary care, which includes cardiology care, was also sourced from PSSG files and was categorized as short (≤30 min), medium (31-60 min), or long (>60 min). We included VHA enrollment priority category as a measure of social and medical risk; this system groups veterans based on military service–connected disability, recent military service, income, and other factors [14]. As in prior work [1,6], the 8 enrollment priority categories were condensed into 4: high disability (>50% service-connected disability or VHA catastrophically disabled), low/moderate disability (10%-40% service-connected disability or military exposure), low income (annual income below area-adjusted mean), or without special VHA enrollment priority. Noncardiology care was captured as primary care visits in the year prior to the study period and categorized into tertiles thereof (0-4, 5-8, or ≥9). Use of mental health care and emergency department or urgent care visits in the year prior to the analysis period were included as binary variables. Veteran chronic conditions were calculated out of a predefined group of 47 possible *International Classification of Disease, Tenth Revision* (*ICD-10*) diagnosis groups, constructed in prior studies in VHA's population [15-18] (Table S3 in Multimedia Appendix 1). Housing instability was based on a combination of outpatient stop codes denoting use of Veterans Affairs housing services and diagnosis codes (Table S2 in Multimedia Appendix 1). Calendar year (2020, 2021, and 2022) was also included.

XSL•FO
RenderX

While in practice patients may see multiple clinicians (including, eg, a nurse practitioner and a physician, or a trainee and an attending physician), sometimes across different facilities, for the purposes of this analysis, we assigned patients a primary cardiology clinician and both patients and clinicians a primary facility. If patients had cardiology encounters with multiple independent licensed practitioners, we defined their main cardiology provider as the provider they had seen most often, or in the case of ties, most recently. In this cohort, 25.1% (56,176/223,809) of patients saw a single cardiology provider; 27.9% (62,443/223,809/N) saw 2, and 47% (105,190/223,809) saw 3 or more cardiology providers. Likewise, for patients seen at multiple VHA medical centers (ie, facilities) for cardiology care, the patient's home site was the site at which the majority of their cardiology encounters took place. In total, 92.6% (207,247/223,809) of patients received care from a single VHA medical center, and 99.6% (222,914/223,809) received care from 2 or fewer.

Statistical Analysis

In addition to descriptive statistics for patient characteristics, we constructed multilevel logistic regression models of our primary outcome, a patient's odds of receiving cardiology telehealth care (ie, care delivered by phone or video). These models included random effects for the patient, the patient's main cardiology provider, and the patient's home VHA facility for specialty cardiology care. Models were adjusted for the patient sociodemographic and clinical characteristics delineated above and calendar year as fixed effects. Statistical analyses were conducted in Stata 17 (StataCorp, LLC).

Ethical Considerations

This analysis was carried out as part of the Virtual Access Quality Enhancement Research Initiative, which is designated as nonresearch quality improvement by VHA program office partners in the VHA Office of Rural Health. The institutional review board at the Stanford Research Compliance Office determined this evaluation does not meet the requirement of research or clinical investigation per Federal Regulations 45CFR 46.104 (Subsection 4) [19] and VA 38CFR 16.104 (Subsection 4) [20]. To protect the privacy and confidentiality of human subjects, study data were anonymous.

Results

Our analytic cohort comprised 989,271 encounters for 223,809 veteran patients, among 2235 clinicians and 138 facilities (VHA medical centers). Figure 1 [21] illustrates the structure of a basic 3-level multilevel statistical model [21] and the numbers included at each level.

Figure 1 depicts examples of how varying numbers of patients and clinicians may be grouped under a given facility (in this case, a Veterans Affairs medical center). Numbers in parentheses denote the level numbers in the model.

Overall, among the 989,271 encounters, 4.2% (n=41,480) of encounters were video based and 34.3% (n=338,834) were phone based. Of these, 385,707 (39%) were virtual (telehealth) over the 2-year period, although this dropped from 52.5% (278,053/529,401) of encounters being conducted virtually in the first year of the pandemic to 23.4% (107,654/459,870) in the second year. Of these veterans, 161,424 had at least 1 telehealth (video or phone) visit (Table 1). The majority of these individuals (n=137,004, 84.9%) used only the phone option for telehealth, as there were only 24,420 video care users.

The average age for veterans in the overall cohort and the subset of telehealth users was the same: 70.2 years (Table 1). Users of video care were a bit younger, with an average age of 68.7 years. Women comprised 3.4% (7616/223,809 and 5528/161,424, respectively) of the overall and telehealth groups. Most sociodemographic and clinical characteristics were similar between the general cohort and the subset of telehealth users, including proportions of each racial or ethnic group, rurality, drive time to care, enrollment priority, use of prior care, and numbers of chronic conditions.

Veterans and providers varied considerably in their encounters per person (Figure S1 in Multimedia Appendix 1). For patients, the mean number of encounters was 4 (SD 3), and for providers, the mean number of encounters among this patient cohort was 443 (SD 648). The mean patients per facility was 1749 (SD 1139) and the mean cardiology providers per facility was 17 (SD 14).

In adjusted multilevel models with random intercepts for the patient, patient's main clinician, and patient's home site, most differences by patient characteristics were small in magnitude if present (Figure 2; Table S4 in Multimedia Appendix 1).

Figure 1. Structure of multilevel model and the number of contributors at each level (based on the model from Leyland and Groenewegen [21]). VA: Veterans Affairs.

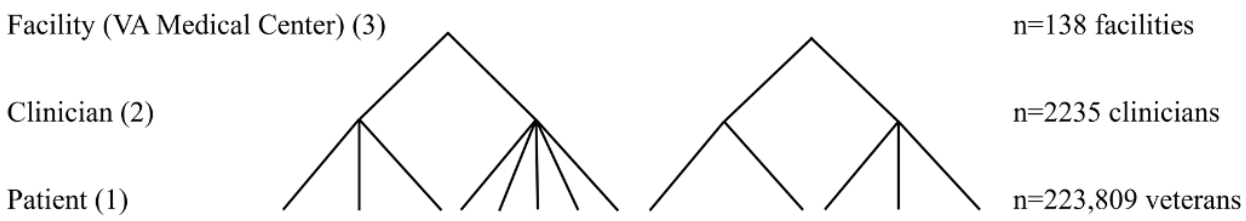

XSL•FO
RenderX

Table 1. Characteristics of the study population (N=223,809) and telehealth users.

Characteristics	All patients[a] (N=223,809)	Telehealth[b] users (n=161,424)	Telehealth nonusers (n=62,385)
Age (years), mean (SD)	70.2 (9.6)	70.2 (9.6)	70.1 (9.4)
Age group (years), n (%)			
18-49	6646 (3)	4838 (3)	1808 (2.9)
50-64	44,218 (19.8)	31,957 (19.8)	12,261 (19.7)
65-74	110,813 (49.5)	79,447 (49.2)	31,366 (50.3)
75+	62,132 (27.8)	45,182 (28)	16,950 (27.2)
Race, n (%)			
American Indian or Alaska Native	1865 (0.8)	1304 (0.8)	561 (0.9)
Asian	1310 (0.6)	981 (0.6)	329 (0.5)
Black or African American	36,180 (16.2)	26,477 (16.4)	9703 (15.6)
Native Hawaiian or other Pacific Islander	1937 (0.9)	1420 (0.9)	517 (0.8)
Unknown	9572 (4.3)	7033 (4.4)	2539 (4.1)
White	172,945 (77.3)	124,209 (76.9)	48,736 (78.1)
Ethnicity, n (%)			
Not Hispanic or Latino	206,846 (92.4)	148,404 (91.9)	58,442 (93.7)
Hispanic or Latino	10,574 (4.7)	8225 (5.1)	2349 (3.8)
Unknown	6389 (2.9)	4795 (3)	1594 (2.6)
Gender, n (%)			
Women	7616 (3.4)	5528 (3.4)	2088 (3.3)
Men	216,193 (96.6)	155,896 (96.6)	60,297 (96.7)
Rurality, n (%)			
Urban	144,217 (64.4)	105,038 (65.1)	39,179 (62.8)
Rural	70,350 (31.4)	49,891 (30.9)	20,459 (32.8)
Highly rural	9242 (4.1)	6495 (4)	2747 (4.4)
Drive time to secondary care (min), mean (SD)	43.6 (34.0)	43.2 (33.3)	44.9 (35.6)
Drive time to secondary care, n (%)			
Short	98,295 (43.9)	70,897 (43.9)	27,398 (43.9)
Medium	71,178 (31.8)	52,369 (32.4)	18,809 (30.1)
Long	52,708 (23.6)	37,025 (22.9)	15,683 (25.1)
Missing	1628 (0.7)	1133 (0.7)	495 (0.8)
Enrollment priority, n (%)			
No service disability	27,791 (12.4)	19,912 (12.3)	7879 (12.6)
Low/moderate disability	42,037 (18.8)	29,791 (18.5)	12,246 (19.6)
High disability	100,122 (44.7)	72,309 (44.8)	27,813 (44.6)
Low income	53,859 (24.1)	39,412 (24.4)	14,447 (23.2)
History of housing instability, n (%)			
No	215,367 (96.2)	155,211(96.2)	60,156 (96.4)
Yes	8442 (3.8)	6213 (3.9)	2229 (3.6)
Primary care visits in year prior to analysis period, mean (SD)	7.2 (6.6)	7.3 (6.7)	6.9 (6.2)

XSL•FO
RenderX

Characteristics	All patients[a] (N=223,809)	Telehealth[b] users (n=161,424)	Telehealth nonusers (n=62,385)
Primary care visits in year prior to analysis period, n (%)			
0-4	93,879 (41.9)	66,452 (41.2)	27,427 (44)
5-8	68,713 (30.7)	49,582 (30.7)	19,131 (30.7)
9+	61,217 (27.4)	45,390 (28.1)	15,827 (25.4)
Number of chronic medical conditions, mean (SD)	7.9 (3.5)	8.0 (3.5)	7.6 (3.4)
Number of chronic medical conditions, n (%)			
0-3	17,837 (8)	12,472 (7.7)	5365 (8.6)
4-7	95,444 (42.6)	67,335 (41.7)	28,109 (45.1)
8-11	76,940 (34.4)	56,135 (34.8)	20,805 (33.3)
12+	33,588 (15)	25,482 (15.8)	8106 (13)
At least 1emergency visit in the year prior to analysis period, n (%)			
No	120,972 (54.1)	85,801 (53.2)	35,171 (56.4)
Yes	102,837 (45.9)	75,623 (46.8)	27,214 (43.6)
At least 1 mental health visit in the year prior to analysis period, n (%)			
No	151,467 (67.7)	108,520 (67.2)	42,947 (68.8)
Yes	72,342 (32.3)	52,904 (32.8)	19,438 (31.2)

[a]Figures in parentheses represent standard deviation for continuous variables and percentages for categorical variables.

[b]Telehealth comprises phone and video care.

Figure 2. Adjusted odds of using cardiology telehealth care among established cardiology patients, selected characteristics (N=223,809). Full model results are included in Table S4 in Multimedia Appendix 1. AOR: adjusted odds ratio.

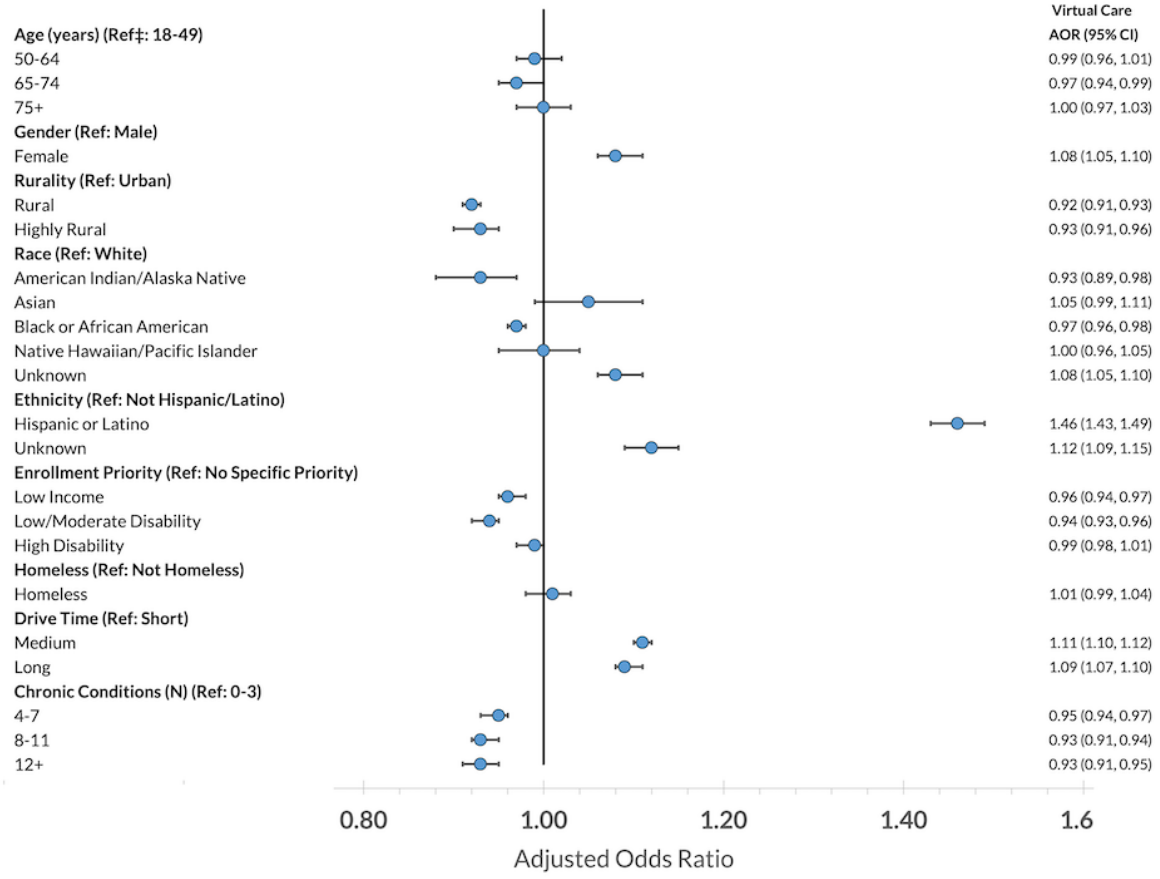

XSL•FO
RenderX

However, there were several groups of veterans with higher adjusted odds of using cardiology telehealth care: most notably, women (adjusted odds ratio [AOR] 1.08, 95% CI 1.05-1.1); Hispanic or Latino veterans (AOR 1.46, 95% CI 1.43-1.49); and those with medium or long drive time (AOR 1.11, 95% CI 1.10-1.12 and AOR 1.09, 95% CI 1.07-1.10, respectively). Adjusted odds of using cardiology telehealth care differed minimally by age (for age 65-74 years compared to the reference category of age 18-49 years, AOR 0.97, 95% CI 0.94-0.99; adjusted odds for the other age categories were not statistically different from 1). Adjusted odds of using telehealth were slightly lower for rural and highly rural veterans compared to those living in urban settings (AOR 0.92, 95% CI 0.91-0.93 and AOR 0.93, 95% CI 0.91-0.96, respectively). Among veterans of different race and ethnicity groups, adjusted odds were lower for those of American Indian or Alaska Native or Black or African American race compared to White veterans (AOR 0.93, 95% CI 0.89-0.98 and AOR 0.97, 95% CI 0.96-0.98, respectively) and higher for those of unknown (to VHA) race (AOR 1.08, 95% CI 1.05-1.10). As above, we did see a larger-magnitude increase for veterans identifying as Hispanic or Latino, and to a lesser extent, unknown ethnicity (AOR 1.12, 95% CI 1.09-1.15).

Compared to those enrolled in VHA without special considerations, patients with low income had a slightly lower AOR for using cardiology telehealth care (AOR 0.96, 95% CI 0.94-0.97), as did those with low or moderate levels of disability (AOR 0.94, 95% CI 0.93-0.96); there was no difference from the reference group for those with high levels of disability. We also saw no differences in AOR for telehealth use among patients with a history of housing instability compared to those without such a history. Having a higher number of chronic health conditions was associated with slightly lower odds of using cardiology telehealth care compared to our reference of having 0-3 chronic conditions (4-7 conditions, AOR 0.95, 95% CI 0.94-0.97; 8-11 conditions, AOR 0.93, 95% CI 0.91-0.94; 12 or more conditions, AOR 0.93, 95% CI 0.91-0.95).

Examining relative proportions of variation in a patient's likelihood of cardiology telehealth care use (ie, the intraclass correlation coefficient), 40.5% (95% CI 35.8%-45.3%) of variation was found at the patient level, 30.8% (95% CI 25.8%-36.2%) at the clinician level, and 7.0% (95% CI 6.3%-7.7%) at the facility level. In total, 21.7% of variation remained unexplained.

Discussion

Principal Findings

This analysis examines predictors of cardiology telehealth use among active users of cardiology care in this nation-wide integrated health care system. We found that patient-level characteristics explained the largest share of the attributable variability in VHA cardiology telehealth use. Clinician-level characteristics explained a more modest share, while facility-level factors contributed little to the variability seen.

Our findings may reflect a high level of standardization in telehealth-related policies (eg, around reimbursement) across VHA facilities. These results suggest that policy solutions intended to improve access and equity in use of cardiology telehealth care should focus on the patient and clinician levels. Such policy levers could include device or technical support for both patients and clinicians to increase the uptake of telehealth. For example, VHA has implemented a Digital Divide consult and tablet distribution program [22] to offer video-enabled devices to veterans with access barriers. Other strategies might include additional support staff to "room" a patient virtually and enter vital signs or chief complaint information; travel reimbursement for in-person appointments; and differential reimbursement for different visit modalities for clinicians, among many others. It should be noted that the desirability of increasing equity of telehealth use depends on the extent to which similar outcomes result from in-person and telehealth care. If the 2 modes are unequal in quality for a given use case, greater use of telehealth in a given subpopulation could represent worse access to quality care. Ongoing research comparing quality of telehealth and in-person care will help to unravel this issue.

The highest-yield drivers may be different in other environments. For example, Tzeng et al [11] found that facility-level factors accounted for more than clinician-level factors in total variance in outpatient virtual clinic use across multiple specialties in Taiwan. Rodriguez et al [23] found contributions to variance in the opposite order from our study (38% attributable to practices, 26% to clinicians, and 9% to patients) in video use (rather than telehealth use more generally) in primary and specialty care practices in the Mass General Brigham system. This variability underscores that drivers of telehealth use are likely to depend heavily on context: the level of the facility (department, practice, or medical center), whether the analysis focuses only on video or on all telehealth, and which specialties are included.

An important caveat when considering patient- versus clinician-level characteristics or policies is that it is challenging to completely disentangle factors at the level of patient versus clinician, as some subset of "patient" contribution may actually be clinician response to a given patient-level characteristic (eg, preferentially seeing older patients in person rather than via telehealth). To the extent that identified clinician-level variability reflects true clinician-level variation in practice, this raises the question of what the optimal breakdown of modalities ought to be for a given clinician, and how much variability in that breakdown across individuals is reasonable. Given that clinicians have differing preferences for telehealth use [24], are all clinicians obliged to offer some telehealth? Or is it acceptable—for quality and patient satisfaction—for some "virtualists" to offer mostly or exclusively telehealth [25] and others, primarily in-person care? These questions require further attention in the postpandemic period.

The reimbursement landscape for telehealth modalities will continue to evolve and affect the type of care offered. Reimbursing clinical time differentially depending on how care is delivered influences clinicians' preferences around telehealth—even in integrated health care systems [8], where the clinician's income is not directly impacted. If the Centers of Medicare and Medicaid (CMS) and health insurers choose

XSL•FO
RenderX

to reimburse differently for video, telephone, and in-person visits, the drivers of telehealth use will undoubtedly change.

With regard to the association between telehealth use and other patient-level characteristics, consistent with other studies of VHA cardiology [6], Hispanic or Latino veterans had higher adjusted odds of cardiology telehealth use compared to not Hispanic or Latino individuals, and rural-dwelling veterans had lower odds of telehealth use compared to urban-dwellers. Earlier in the pandemic, we found that men were more likely than women to use cardiology telehealth care, an effect reversed in this study (potentially because of evolving patterns of use over time). Previously, we also found no association between telehealth use and long drive time, whereas in this study, those with a longer drive time were more likely to use telehealth. As in prior work examining disparities in cardiology telehealth care in VHA and across other health care systems, there were small or no differences by race [6,7,26]; whether or not differences in telehealth use appear by age has varied across systems [6,7,26].

Limitations

Our study has limitations common to observational data and work in VHA's system. First, generalizability beyond VHA—a national, integrated health care system with long-standing, well-established processes in place to standardize practices across facilities—or to populations with less engaged users of care may be limited. Second, because we constructed our facility-level variable at the VHA medical center level, we cannot assess whether more granular facility levels (eg, individual clinics) could explain additional variation in telehealth use. Third, while all patients included had at least 1 cardiology diagnosis, we lack the specific diagnoses that were being addressed at these encounters. Fourth, we have analyzed telehealth visits with phone and video visits combined; in this population, video visits were rare enough that 3-level multilevel models were not feasible, and thus we cannot specifically report on the differential contributions to variability in video care use or whether those differ from the pattern among all telehealth visits. Fifth, our results represent a pooled estimate across the study period of 2020-2022, while use patterns may have varied across the pandemic [2,27]. Sixth, we do not include facility-level or provider-level characteristics due to computational limitations and data availability. Finally, we attributed patients to a single clinician and site to allow our statistical models to run; as a majority of veterans did see more than 1 cardiology provider during the study period, this required a significant simplification of our complex real-world data.

Conclusions

In sum, within VHA, a nationally integrated health care system, there are marked differences in the degree to which patient, clinician, and facility factors influence use of cardiology telehealth care. Given that most variability occurs at the patient and clinician levels, it suggests that these levels might be optimal targets for interventions intended to alter the mix of telehealth versus in-person care use.

Acknowledgments

The authors thank the Virtual Access QUERI team for supporting this evaluation, including Cindie Slightam, MPH, and Camila Chaudhary, MPH, from Veterans Affairs Palo Alto Health Care System for providing project management support. Preliminary results of this work were presented at the national meeting for the Society of General Internal Medicine, May 10-13, 2023. Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award UL1TR003142. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by a US Department of Veterans Affairs Quality Enhancement Research Initiative PEI 18-205 (Principal Investigator DMZ) and by the Veterans Administration Office of Academic Affairs Advanced Fellowship in Health Services Research (RLT). Funding provided by the US Department of Veterans Affairs (Office of Rural Health. Views expressed are those of the authors and the contents do not represent the views of the US Department of Veterans Affairs or the United States Government. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

Due to US Department of Veterans Affairs (VA) regulations and our ethics agreements, the analytic datasets used for this study are not permitted to leave the VA firewall without a Data Use Agreement. This limitation is consistent with other studies based on VA data. However, VA data are made freely available to researchers with an approved VA study protocol.

Authors' Contributions

RLT and JMF participated in the conceptualization and design of the study and performed the formal analysis, investigation, and validation. RLT wrote the original manuscript text. JVC and LG curated the data. DMZ supervised all work. All authors participated in the review and revision of the final manuscript draft for submission.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Cardiac Diagnosis Codes; Stop Codes; Chronic Condition Groups; Full Results of Multilevel Multivariable Logistic Regression Model; Variability in Utilization and Volume across Patients, Providers, and Facilities.
[DOCX File , 436 KB - jmir_v26i1e53298_app1.docx]

References

1. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the veterans health administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc 2021;28(3):453-462 [FREE Full text] [doi: 10.1093/jamia/ocaa284] [Medline: 33125032]
2. Ferguson JM, Wray CM, Jacobs J, Greene L, Wagner TH, Odden MC, et al. Variation in initial and continued use of primary, mental health, and specialty video care among veterans. Health Serv Res 2023;58(2):402-414 [FREE Full text] [doi: 10.1111/1475-6773.14098] [Medline: 36345235]
3. Goettsche K, DellaCava E. Telehealth as an option, not a requirement-who may benefit, and how do we decide? The American Journal of Geriatric Psychiatry 2023;31(3):S3. [doi: 10.1016/j.jagp.2022.12.269]
4. Burnett K, Stockdale SE, Yoon J, Ragan A, Rogers M, Rubenstein LV, et al. The clinical resource hub initiative: first-year implementation of the veterans health administration regional telehealth contingency staffing program. J Ambul Care Manage 2023;46(3):228-239. [doi: 10.1097/JAC.0000000000000468] [Medline: 37079357]
5. Takahashi EA, Schwamm LH, Adeoye OM, Alabi O, Jahangir E, Misra S, et al. An overview of telehealth in the management of cardiovascular disease: a scientific statement from the American heart association. Circulation 2022;146(25):e558-e568. [doi: 10.1161/CIR.0000000000001107] [Medline: 36373541]
6. Tisdale RL, Ferguson J, van Campen J, Greene L, Sandhu AT, Heidenreich PA, et al. Disparities in virtual cardiology visits among veterans health administration patients during the COVID-19 pandemic. JAMIA Open 2022;5(4):ooac103 [FREE Full text] [doi: 10.1093/jamiaopen/ooac103] [Medline: 36531138]
7. Osmanlliu E, Kalwani NM, Parameswaran V, Qureshi L, Dash R, Scheinker D, et al. Sociodemographic disparities in the use of cardiovascular ambulatory care and telemedicine during the COVID-19 pandemic. Am Heart J 2023;263:169-176 [FREE Full text] [doi: 10.1016/j.ahj.2023.06.011] [Medline: 37369269]
8. Gray C, Wray C, Tisdale R, Chaudary C, Slightam C, Zulman D. Factors influencing how providers assess the appropriateness of video visits: interview study with primary and specialty health care providers. J Med Internet Res 2022;24(8):e38826 [FREE Full text] [doi: 10.2196/38826] [Medline: 36001364]
9. Jacobs J, Ferguson JM, van Campen J, Yefimova M, Greene L, Heyworth L, et al. Organizational and external factors associated with video telehealth use in the veterans health administration before and during the COVID-19 pandemic. Telemed J E Health 2022;28(2):199-211. [doi: 10.1089/tmj.2020.0530] [Medline: 33887166]
10. Connolly SL, Charness ME, Miller CJ. To increase patient use of video telehealth, look to clinicians. Health Serv Res 2023;58(1):5-8 [FREE Full text] [doi: 10.1111/1475-6773.14041] [Medline: 35904201]
11. Tzeng YH, Yin WH, Lin KC, Wei J, Liou HR, Sung HJ, et al. Factors associated with the utilization of outpatient virtual clinics: retrospective observational study using multilevel analysis. J Med Internet Res 2022;24(8):e40288 [FREE Full text] [doi: 10.2196/40288] [Medline: 35917486]
12. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. World Health Organization. 2020 Mar 11. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [accessed 2021-04-20]
13. Urban and rural. US Census Bureau. 2020. URL: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html [accessed 2021-02-26]
14. Wang ZJ, Dhanireddy P, Prince C, Larsen M, Schimpf M, Pearman G. 2019 Survey of veteran enrollees' health and use of health care. Department of Veterans Affairs. 2020 Sep 24. URL: https://www.va.gov/VHASTRATEGY/SOE2021/2021_Enrollee_Data_Findings_Report-508_Compliant.pdf [accessed 2024-10-11]
15. Yu W, Ravelo A, Wagner TH, Phibbs CS, Bhandari A, Chen S, et al. Prevalence and costs of chronic conditions in the VA health care system. Med Care Res Rev 2003;60(3 Suppl):146S-167S. [doi: 10.1177/1077558703257000] [Medline: 15095551]
16. Fishman P, Von Korff M, Lozano P, Hecht J. Chronic care costs in managed care. Health Aff (Millwood) 1997;16(3):239-247. [doi: 10.1377/hlthaff.16.3.239] [Medline: 9141341]
17. Ray GT, Collin F, Lieu T, Fireman B, Colby CJ, Quesenberry CP, et al. The cost of health conditions in a health maintenance organization. Med Care Res Rev 2000;57(1):92-109. [doi: 10.1177/107755870005700106] [Medline: 10705704]
18. Yoon J, Chee CP, Su P, Almenoff P, Zulman DM, Wagner TH. Persistence of high health care costs among VA patients. Health Serv Res 2018;53(5):3898-3916 [FREE Full text] [doi: 10.1111/1475-6773.12989] [Medline: 29862504]
19. Title 45 § 46.104 Exempt research. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104 [accessed 2024-10-23]
20. Title 38 § 16.104 Exempt research. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-38/chapter-I/part-16/section-16.104 [accessed 2024-10-23]
21. Leyland AH, Groenewegen PP. Multilevel data structures. In: Multilevel Modelling for Public Health and Health Services Research. Cham, Switzerland: Springer; 2020:49-68.

XSL•FO
RenderX

22. Zulman DM, Wong EP, Slightam C, Gregory A, Jacobs JC, Kimerling R, et al. Making connections: nationwide implementation of video telehealth tablets to address access barriers in veterans. JAMIA Open 2019;2(3):323-329 [FREE Full text] [doi: 10.1093/jamiaopen/ooz024] [Medline: 32766533]
23. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021;27(1):21-26 [FREE Full text] [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
24. Connolly SL, Miller CJ, Gifford AL, Charness ME. Perceptions and use of telehealth among mental health, primary, and specialty care clinicians during the COVID-19 pandemic. JAMA Netw Open 2022;5(6):e2216401 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.16401] [Medline: 35671053]
25. Nochomovitz M, Sharma R. Is it time for a new medical specialty?: The medical virtualist. JAMA 2018;319(5):437-438. [doi: 10.1001/jama.2017.17094] [Medline: 29181539]
26. Eberly LA, Khatana SAM, Nathan AS, Snider C, Julien HM, Deleener ME, et al. Telemedicine outpatient cardiovascular care during the COVID-19 pandemic: bridging or opening the digital divide? Circulation 2020;142(5):510-512 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.048185] [Medline: 32510987]
27. Ferguson JM, Wray CM, van Campen J, Zulman DM. A new equilibrium for telemedicine: prevalence of in-person, video-based, and telephone-based care in the veterans health administration, 2019-2023. Ann Intern Med 2024;177(2):262-264. [doi: 10.7326/M23-2644] [Medline: 38252945]

Abbreviations

CMS: Centers for Medicare and Medicaid Services
ICD-10: International Classification of Disease, Tenth Revision
NPI: National Provider Identifier
PSSG: Planning Systems Support Group
VHA: Veterans Health Administration

Edited by N Cahill; submitted 02.10.23; peer-reviewed by C Turvey, K Kaphingst; comments to author 24.02.24; revised version received 11.03.24; accepted 14.08.24; published 25.10.24.

Please cite as:
Tisdale RL, Ferguson JM, Van Campen J, Greene L, Wray CM, Zulman DM
Patient-, Provider-, and Facility-Level Contributors to the Use of Cardiology Telehealth Care in the Veterans Health Administration: Retrospective Cohort Study
J Med Internet Res 2024;26:e53298
URL: https://www.jmir.org/2024/1/e53298
doi:10.2196/53298
PMID:

©Rebecca Lauren Tisdale, Jacqueline M Ferguson, James Van Campen, Liberty Greene, Charlie M Wray, Donna M Zulman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Importance of Race and Gender Concordance Between Patients and Physical Therapists in Digital Rehabilitation for Musculoskeletal Conditions: Observational, Longitudinal Study

Anabela C Areias[1], PhD; Dora Janela[1], PT; Maria Molinos[1], PhD; Virgílio Bento[1], PhD; Carolina Moreira[1,2], MD; Vijay Yanamadala[1,3,4], MD; Steven P Cohen[5,6], MD, PhD; Fernando Dias Correia[1,7], MD, PhD; Fabíola Costa[1], PhD

[1]Sword Health, Inc, Draper, UT, United States
[2]Instituto de Ciências Biomédicas Abel Salazar, Porto, Portugal
[3]Department of Surgery, Quinnipiac University Frank H Netter School of Medicine, Hamden, CT, United States
[4]Department of Neurosurgery, Hartford Healthcare Medical Group, Westport, CT, United States
[5]Northwestern Feinberg School of Medicine, Chicago, IL, United States
[6]Walter Reed National Military Medical Center, Uniformed Services University of the Health Sciences, Bethesda, MD, United States
[7]Neurology Department, Centro Hospitalar e Universitário do Porto, Porto, Portugal

Corresponding Author:
Fabíola Costa, PhD
Sword Health, Inc
13937 Sprague Lane, Ste 100
Draper, UT, 84020
United States
Phone: 1 385 308 8034
Fax: 1 801 206 3433
Email: f.costa@swordhealth.com

Abstract

Background: Race/ethnicity and gender concordance between patients and providers is a potential strategy to improve health care interventions. In digital health, where human interactions occur both synchronously and asynchronously, the effect of concordance between patients and providers is unknown.

Objective: This study aimed to evaluate the impact of race/ethnicity or gender concordance between patients and physical therapists (PTs) in engagement and the clinical outcomes following a digital care program (DCP) in patients with musculoskeletal (MSK) conditions.

Methods: This secondary analysis of 2 prospective longitudinal studies (originally focused on assessing the acceptance, engagement, and clinical outcomes after a remote DCP) examined the impact of both race/ethnicity concordance and gender concordance between patients and PTs on outcomes for a digital intervention for MSK conditions. Outcomes included engagement (measured by the completion rate and communication, assessed by text interactions), satisfaction, and clinical outcomes (response rate, ie, percentage of patients achieving at least a minimal clinically important change in pain, measured by the Numerical Pain Rating Scale [NPRS]; anxiety, measured by the Generalized Anxiety Disorder 7-item scale [GAD-7]; depression, measured by the Patient Health Questionnaire 9-item [PHQ-9]; and daily activity impairment, measured by the Work Productivity and Activity Impairment [WPAI] questionnaire).

Results: Of 71,201 patients, 63.9% (n=45,507) were matched with their PT in terms of race/ethnicity, while 61.2% (n=43,560) were matched for gender. Concordant dyads showed a higher completion rate among White (adjusted odds ratio [aOR] 1.11, 95% CI 1.05-1.19, $P<.001$) and Hispanic (aOR 1.27, 95% CI 1.08-1.54, $P=.009$) groups, as well as women (aOR 1.10, 95% CI 1.06-1.18, $P<.001$), when compared to discordant dyads. High and similar levels of interaction between patients and PTs were observed across race/ethnicity and gender dyads, except for Asian concordant dyads (adjusted β coefficient 5.32, 95% CI 3.28-7.36, $P<.001$). Concordance did not affect satisfaction, with high values (>8.52, 95% CI 8.27-8.77) reported across all dyads. Response rates for pain, anxiety, and daily activity impairment were unaffected by race/ethnicity concordance. An exception was observed for depression, with White patients reporting a higher response rate when matched with PTs from other races/ethnicities (aOR 1.20, 95% CI 1.02-1.39, $P=.02$). In terms of gender, men had a slightly higher pain response rate in discordant dyads (aOR 1.08, 95% CI 1.01-1.15, $P=.03$) and a higher depression response rate in concordant dyads (aOR 1.23, 95% CI 1.05-1.47, $P=.01$).

XSL•FO
RenderX

Conclusions: Race/ethnicity and gender concordance between patients and PTs does not translate into higher satisfaction or improvement for most clinical outcomes, aside from a positive effect on treatment completion. These results highlight the importance of other PT characteristics, in addition to race/ethnicity or gender concordance, suggesting the potential benefit of experience, languages spoken, and cultural safety training as ways to optimize care.

Trial Registration: ClinicalTrials.gov NCT04092946, NCT05417685; https://clinicaltrials.gov/study/NCT05417685, https://clinicaltrials.gov/study/NCT04092946

(J Med Internet Res 2024;26:e65354) doi:10.2196/65354

KEYWORDS

musculoskeletal pain; physical therapy; telerehabilitation; eHealth; racial/ethnic concordance; patient–provider concordance; physical therapy; digital rehabilitation; musculoskeletal conditions

Introduction

Individual characteristics, including race and ethnicity, are recognized as major determinants of health according to the World Health Organization. Health disparities among racial and ethnic minorities are well documented, revealing decreased health care access, poorer health status, and prognosis among these populations [1-3].

Although future projections forecast an increase in population and diversity in the United States [4], diversity within the medical community is not growing at the same pace [5]. This imbalance has raised concerns about the ability to provide culturally competent care and address the specific needs of these patients, when the majority of the workforce is White [6]. As such, the concept of concordance, when a patient and a clinician have the same race/ethnicity or gender, has been explored to understand its potential impact on engagement and clinical outcomes.

Previous studies, predominantly conducted in primary care and psychology settings, have focused on understanding patients' perspectives regarding clinicians' demographics and their impact on therapeutic alliance, communication quality, and overall satisfaction. Findings from these studies are varied: some suggest that patient-clinician racial/ethnic or gender concordance offers benefits [7-11], while others do not [12-14]. Moreover, a recent systematic review studying the association between racial, gender, and multifactorial concordance reported inconclusive results on patient outcomes, advocating for further research [15]. This heterogeneity might be explained by singularities in the population, conditions, and settings under study.

In digital health, where care can be delivered both synchronously and asynchronously, the impact of concordance between patients and providers remains largely unexplored, particularly in the field of physical therapy. Digital rehabilitation has arisen as a solution to address geographical and time barriers, as well as clinician shortages. Previous research supports its effectiveness among several musculoskeletal (MSK) conditions [16-18] regardless of race/ethnicity [19] and socioeconomic backgrounds [20]. Given the rising adoption of digital rehabilitation programs, and the unique dynamics they introduce to the patient-therapist relationship (eg, duration and frequency of contacts, communication channels used), it is important to assess the association between concordance, engagement, and patient outcomes.

This study aimed to evaluate the impact of race/ethnicity or gender concordance between patients and physical therapists (PT) in engagement and the clinical outcomes following an asynchronous digital care program (DCP) in patients with MSK conditions in a real-world setting. Additional insights into this topic can enable providers and constituents to develop strategies that consider these aspects.

Methods

Study Design

This was a secondary analysis of 2 prospective, longitudinal, single-arm investigations focusing on the acceptability, feasibility, engagement, and clinical outcomes of a remote DCP for a population with acute and chronic MSK conditions [20-24]. The overarching aim of the original protocol also included the exploration of associations between the outcomes and patients' demographic and clinical profiles. Patients participated in this study between January 1, 2022, and December 15, 2023. This study was reported in accordance with STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Table S1 in Multimedia Appendix 1).

Study Population

Patients, who were beneficiaries of employer health plans, reporting MSK pain in any of the following areas were included in accordance with eligibility criteria in the original investigations: ankle, elbow, hip, knee, low back, neck, shoulder, or wrist/hand. Patients self-reported their race/ethnicity in accordance with 2020 U.S. Census Bureau categories by the U.S. Office of Management and Budget (OMB) [25] as non-Hispanic Asian (hereafter referred to as Asians), non-Hispanic Black or African Americans (hereafter referred to as Black), Hispanic or Latino (hereafter referred to as Hispanic), non-Hispanic White (hereafter referred to as White), American Indian or Alaska Native, and Native Hawaiian or Other Pacific Islander.

Exclusion criteria defined in the original investigations were (1) health conditions incompatible with >20 minutes of light-to-moderate exercise; (2) ongoing cancer treatment; (3) the presence of red flags, defined as signs or symptoms suggestive of serious pathology (eg, rapid progressive motor weakness or sensory alterations, or bowel or bladder

XSL•FO
RenderX

dysfunction), if uncleared by a physician; and (4) missing patient or PT gender or race/ethnicity. Additional exclusion criteria specific for this secondary analysis were patient race/ethnicity from American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander, given the small sample size of these subgroups that precluded statistical analyses. Patients who were compliant with the intervention but missed a given reassessment survey were included.

Intervention

This completely remote DCP, developed according to current guidelines [26-28], consisted of home-based exercise, education, and cognitive behavioral therapy (CBT) for up to 12 weeks (at least 3 exercise sessions per week were recommended), as described elsewhere [19,20].

Briefly, after enrollment, patients completed a baseline form on a dedicated website collecting demographic and clinical information. At the end of this form, patients could choose from a pool of available PTs presented with a photo and a mini biosketch (Figure 1). PTs were diverse in gender, race/ethnicity (Figure S2 in Multimedia Appendix 1), and spoken languages (English, Spanish, French, Portuguese), having a minimum experience of 3 years. All PTs completed cultural safety training, encompassing implicit bias, cultural awareness and humility, motivational interview training and strategies to build rapport with patients from diverse backgrounds. During onboarding, patients had a synchronous video call with their PT for an average duration of 30-60 minutes. During this call, the PT collected further medical history, conducted a clinical assessment, established goals through shared decision-making, and defined a tailored program following a patient-centered approach. Subsequent interactions between the patient and the PT were mostly asynchronous, where the therapeutic relationship was fostered through bidirectional communication via a secure chat within a smartphone app (Figure 1). Nonetheless, additional video calls were scheduled, whenever deemed appropriate by either patients or PTs.

Figure 1. Screenshots of the patient app showcasing the content offered upon onboarding PT selection, during video calls, and during text messaging along the DCP (no real patients are portrayed in this figure). DCP: digital care program; PT: physical therapist.

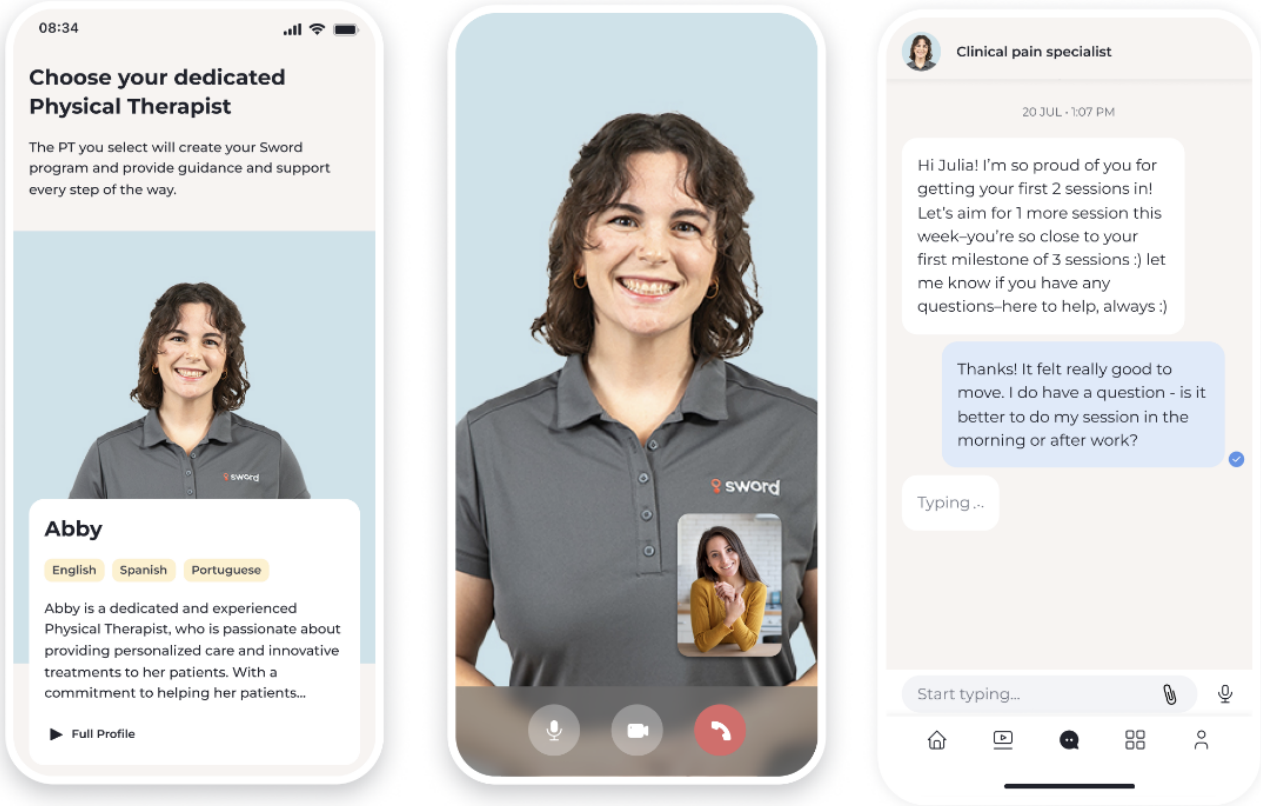

Patients received a Food and Drug Administration (FDA)–listed class II medical device composed of a dedicated tablet with a mobile app, motion tracking, and a cloud-based portal. The DCP was conducted autonomously and asynchronously at the patient's convenience, with exercise sessions being displayed on the tablet through videos with real-time audio and visual biofeedback on performance. The PT monitored and adjusted the program asynchronously through the cloud-based portal. Education covering pathophysiology, pain reconceptualization, active coping skills, the role of exercise, and fear-avoidance behaviors [26-28] was provided through short articles or videos via a smartphone app. The CBT component, delivered through interactive and audio modules via email, focused on mindfulness, acceptance and commitment therapy, and empathy-focused therapy [29].

Outcome Measures

Patient engagement was evaluated by the completion rate, defined as the percentage of patients who completed the DCP, and by communication, estimated through the number of text interactions between patients and PTs.

XSL•FO
RenderX

Program satisfaction was assessed through the question "On a scale from 0 to 10, how likely is it that you would recommend this intervention to a friend or neighbor?", recorded at the last reassessment. Clinical outcome data were collected at baseline and at the 9th, 18th, and 27th sessions as long as participants reached that milestone.

Clinical outcomes were analyzed for participants who reached the program end (ie, completers), where the last-available assessment data were used to determine the minimum clinically important change (MCIC), as described in Table 1.

Table 1. Clinical outcomes and corresponding patient-reported outcome measures with MCIC[a] thresholds.

Response rate	Patient-reported outcome measure	MCIC threshold
Pain	The NPRS[b] specific to the symptomatic body region: "Please rate your average pain over the last 7 days, from 0 (no pain at all) to 10 (worst pain imaginable)."	Improvement in pain by at least 30% according to IMMPACT (Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials) guidelines [30]
Daily activity impairment	WPAI[c] questionnaire: General Health v2.0 (range 0-100) [31]	Reduction of ≥9.1 [23]
Anxiety	GAD-7[d] (range 0-21) [32,33]	Reduction of ≥3.8 [34]
Depression	PHQ-9[e] (range 0-27) [33,35]	Reduction of ≥5 [36]

[a]MCIC: minimum clinically important change.

[b]NPRS: Numerical Pain Rating Scale.

[c]WPAI: Work Productivity and Activity Impairment.

[d]GAD-7: Generalized Anxiety Disorder 7-item scale.

[e]PHQ-9: Patient Health Questionnaire 9-item.

Statistical Analysis

Descriptive statistics were reported using means and proportions. Comparisons between concordant and discordant groups were performed using independent-sample *t* tests or Mann-Whitney *U* tests for continuous variables (with Bonferroni correction) or chi-squared tests for categorical variables. Completion and clinical outcomes were binary variables, while satisfaction was a continuous variable. Two predictive variables describing concordance on race/ethnicity or on gender were created. Concordance was defined as an exact match on race/ethnicity between the patient and the PT, or a match on gender, depending on the variable of interest. Response rates were calculated using the last patient reassessment. In the case of missing reassessment at the discharge time point, the last available reassessment was carried forward.

Due to variability in the distribution of members allocated to each PT, a consequence of some PTs having been in the study longer, using PTs as random effects was not feasible. Generalized linear or logistic regression models were used to assess the effects of concordance on both race/ethnicity and gender on the described outcomes. Models of race/ethnicity concordance included an interaction term between concordance and patient race/ethnicity to explore the effect of concordance on outcomes stratified by race/ethnicity. Similarly, models of gender concordance included an interaction term between concordance and patient gender to explore the effect of concordance on outcomes stratified by gender. Given the baseline differences observed between groups in variables previously reported as predictors of recovery in patients with MSK conditions after physical therapy [37-39], all models were adjusted for the BMI (body mass index); age; education levels (dichotomized as low [no college degree] or high [at least a bachelor's degree]); employment status (dichotomized as employed either full- or part-time or not currently employed); rurality; acuity (acute: pain<3 months; chronic: pain≥3 months) [40]); and baseline levels of pain, anxiety, and depression. Models assessing clinical outcomes were also adjusted for the discharged time point. A false discovery rate was used to adjust for multiple comparisons, and statistical significance was set at $P<.05$. Descriptive statistics were performed using commercially available software (IBM SPSS v22), and statistical analyses were performed using R Studio version 2023.09.1+494 (R Foundation for Statistical Computing).

Ethical Considerations

All research was performed in accordance with relevant guidelines and regulations set by the Declaration of Helsinki. The original studies were prospectively approved by the New England (120190313) and Advarra (Pro00063337) independent Institutional Review Boards (IRBs) and registered on ClinicalTrials (NCT04092946, NCT05417685) on September 17, 2019, and June 14, 2022, respectively. All patients provided informed consent. Both IRB approvals and respective original consent covered secondary analysis without need for additional consent.

Prior to data analysis, all collected data underwent a rigorous anonymization process to safeguard the privacy of participants involved in the research. All individuals whose images are included in this publication (Figure 1) provided explicit written consent for the display of identifiable features in this manuscript, in accordance with journal guidelines and ethical standards. All data collection and analysis methods complied with relevant guidelines and regulations. Participants were not provided with any form of compensation.

XSL•FO
RenderX

Results

Participant Details

Of 84,326 eligible participants, 71,201 (84.4%) started the program, of which 49,877 (70.1%) completed the intervention (Figure 2). Of those who started the program, 63.9% (n=45,507) matched their PT's race/ethnicity, while 61.2% (n=43,560) shared the same gender. Patients' demographic and clinical characteristics stratified by race/ethnicity and gender concordance are reported in Table S2 in Multimedia Appendix 1. Race/ethnicity and gender distributions of the PTs (Figure S1 in Multimedia Appendix 1) were similar to the membership reported by the American Physical Therapy Association (APTA). The cohort was predominantly composed of middle-aged (mean 48.7 years, SD 11.8) women (n=42,464, 59.6%), who identified as White (n=49,920, 70.1%) and had a BMI above 30 (n=28,337, 39.9%). These individuals tended to have a high level of education (n=45,483, 63.8%), lived in urban locations (n=62,405, 87.6%), and reported full-time employment (n=61,298, 86.1%).

Figure 2. Study flowchart.

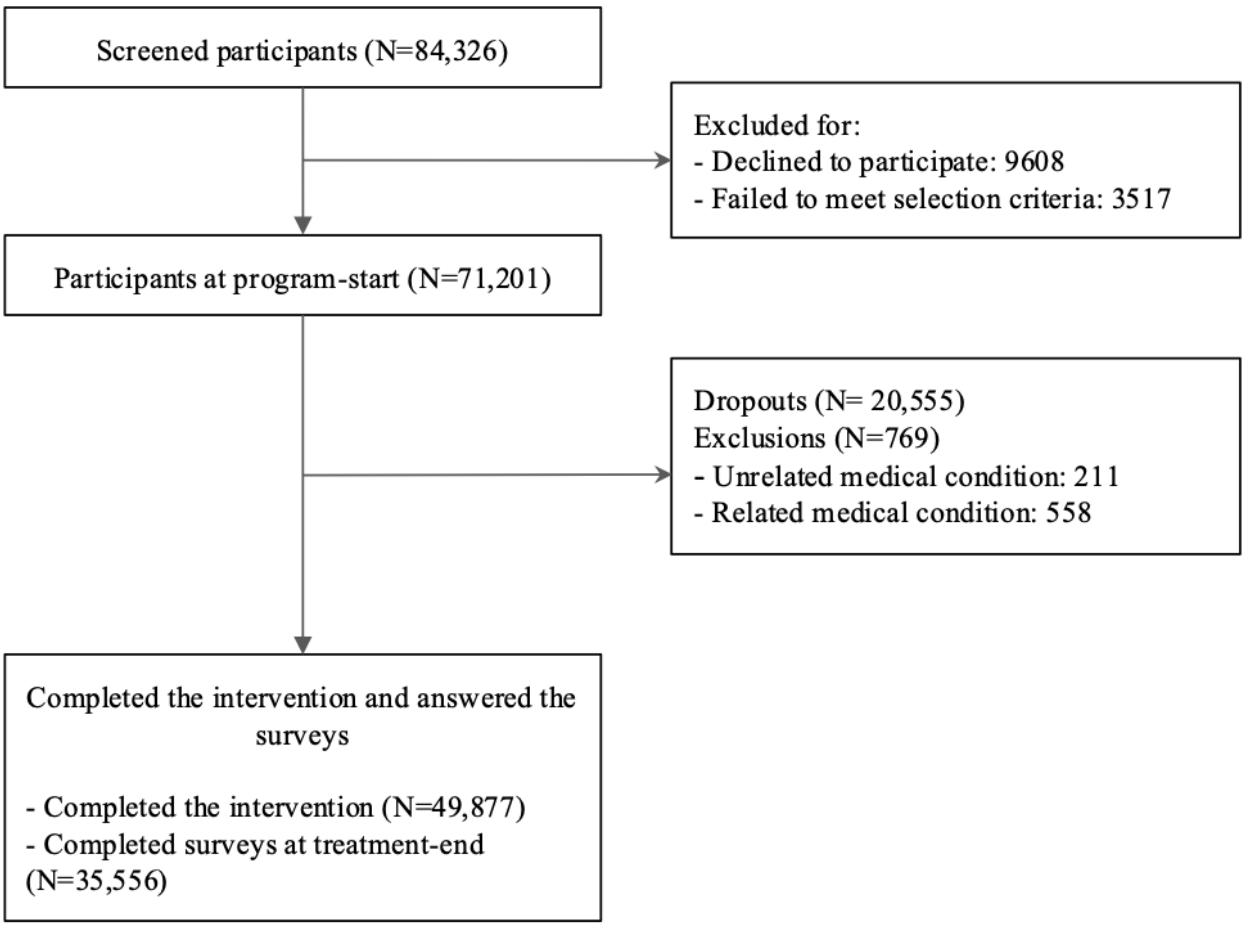

Regarding race/ethnicity, concordant and discordant groups were different in some demographic characteristics, such as age (mean 46.9, SD 11.5, vs mean 49.7, SD 11.8; *P*<.001), employment status (employed: n=22,898, 89.1%, vs n=38,400, 84.4%; *P*<.001), and rurality (n=1761, 6.9%, vs n=6708, 14.7%; *P*<.001). Differences were also observed among baseline clinical data, namely acuity (chronic: n=19,461, 75.9%, vs n=36,013, 79.3%; *P*<.001) and pain (mean 4.9, SD 2.0, vs mean 4.7, SD 1.9; *P*<.001).

Regarding gender, the concordant group had a larger proportion of women (n=33,854, 77.7%, vs n=8610, 31.1%; *P*<.001), patients working part-time (n=2215, 5.1%, vs n=770, 2.8%; *P*<.001), and patients with depression (mean 9.6, SD 4.3, vs mean 9.4, SD 4.2; *P*=.004) and daily activity impairment (mean 37.2, SD 22.8, vs mean 35.8, SD 22.4; *P*<.001).

Engagement Outcomes

The effect of race/ethnicity and gender concordance/discordance on completion rates and PT-patient interactions is presented in Figure 3 and Table 2, respectively. Both race/ethnicity and gender concordance were positively associated with the completion rate when compared to discordant dyads, particularly for those who identified as White (adjusted odds ratio [aOR] 1.11, 95% CI 1.05-1.19, *P*<.001), Hispanic (aOR 1.27, 95% CI 1.08-1.54, *P*=.009), and women (aOR 1.10, 95% CI 1.06-1.18, *P*<.001), as shown in Figure 3. PT-patient interactions were consistently high across the different race and gender dyads. Although there was a statistically significant difference between concordant and discordant dyads, this may not represent a clinically significant difference, as this variation ranged from 1 to 3 text interactions. An exception was observed among Asian

XSL•FO
RenderX

patients, where a wider difference was observed, with discordant dyads reporting significantly higher interactions (adjusted β coefficient 5.32, 95% CI 3.28-7.36) compared to concordant dyads.

Figure 3. The aORs for concordance stratified by patient race/ethnicity and gender (reference: PT-patient concordance) for the following outcomes: (A) completion rate, (B) pain response rate, (C) daily activity response rate, (D) depression response rate; and (E) anxiety response rate. Statistically significant aORs are depicted in red. aOR: adjusted odds ratio.

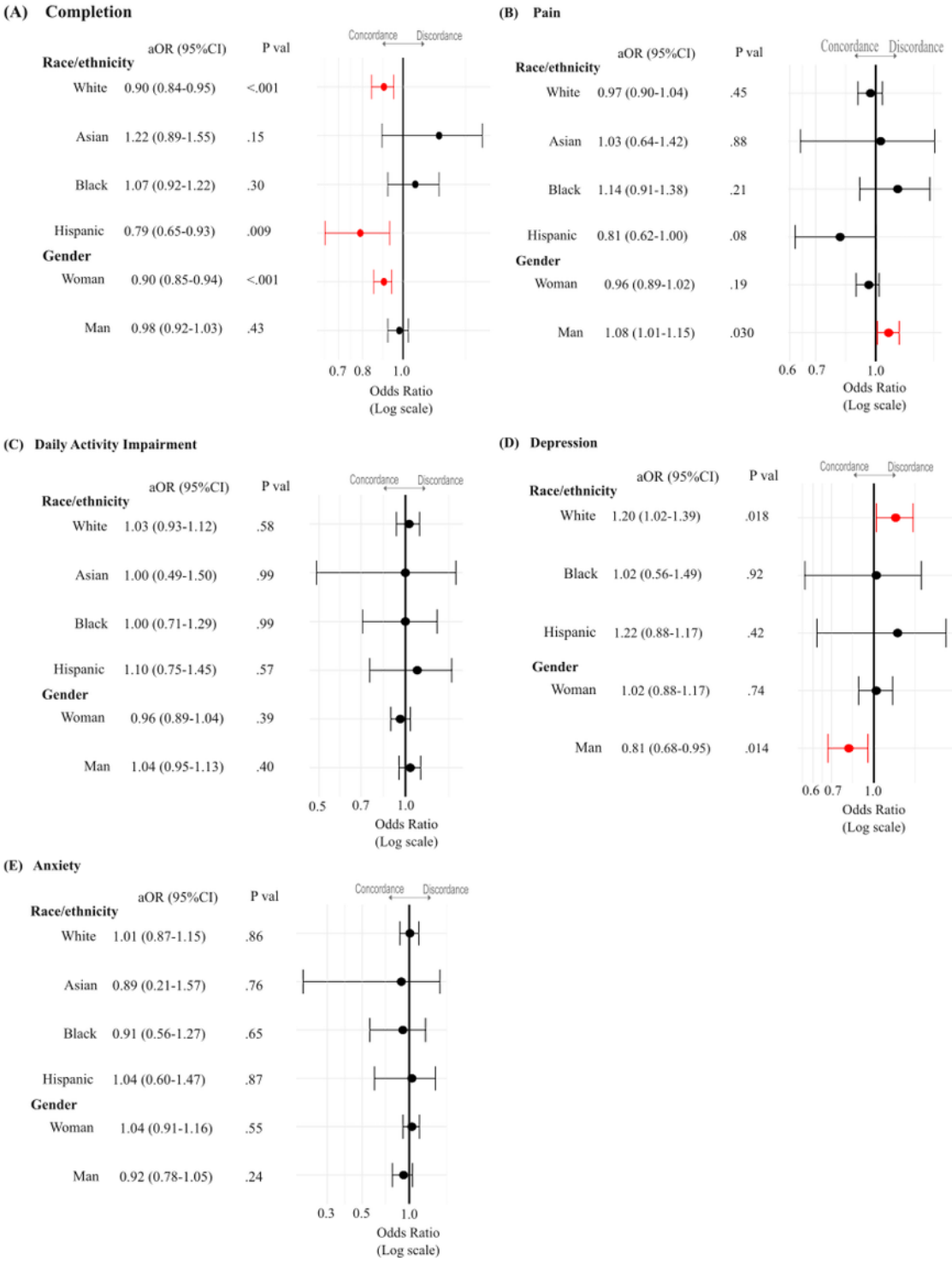

XSL•FO
RenderX

Table 2. Generalized linear model estimates of text interactions between patients and PTs[a] stratified by patient race/ethnicity and gender (reference: PT-patient concordance).

Variable	Concordance, mean (95% CI); n (%)	Discordance, mean (95% CI); n (%)	Adjusted β coefficient (95% CI)	*P*-value[b]
Race/ethnicity				
White	22.76 (22.13 to 23.39); 43,523 (87.2)	20.10 (19.39 to 20.80); 6397 (12.8)	–2.66 (–3.03 to –2.30)	<.001
Asian	16.76 (14.65 to 18.87); 254 (3.6)	22.08 (21.37 to 22.78); 6773 (96.4)	5.32 (3.28 to 7.36)	<.001
Black	20.17 (19.12 to 21.23); 1061 (14.4)	21.53 (20.82 to 22.24); 6284 (85.6)	1.36 (0.44 to 2.28)	.004
Hispanic	19.48 (18.31 to 20.65); 669 (9.7)	21.20 (20.50 to 21.91); 6240 (90.3)	1.72 (0.67 to 2.77)	.001
Gender				
Women	23.07 (22.44 to 23.70); 33,854 (79.7)	21.57 (20.89 to 22.25); 8610 (20.3)	–1.50 (–1.83 to –1.17)	<.001
Men	20.60 (19.93 to 21.27); 9706 (66.2)	21.44 (20.80 to 22.09); 19,031 (33.8)	0.84 (0.51 to 1.17)	<.001

[a]PT: physical therapist.

[b]All *P* values were significant.

Satisfaction

Program satisfaction was high across different races/ethnicities (mean 8.52, 95% CI 8.27-8.77, to mean 9.37, 95% CI 9.18-9.55) regardless of concordance status (all *P*>.05; Table 3), except for Asian patients who were statistically more likely to report higher satisfaction when matched in discordant dyads (adjusted β coefficient 0.36, 95% CI 0.12-0.59, *P*=.003; reference: concordance). Men showed a statistically higher likelihood to report satisfaction when matched in discordant dyads (adjusted β coefficient 0.15, 95% CI 0.10-0.20, *P*<.001; reference: concordance). In both cases (Asian patients and men), differences were small and therefore might not be clinically significant.

Table 3. Generalized linear model estimates of satisfaction level by patient race/ethnicity and gender (reference: PT[a]-patient concordance).

Variable	Concordance, mean (95% CI); n (%)	Discordance, mean (95% CI); n (%)	Adjusted β coefficient (95% CI)	*P*-value
Race/ethnicity				
White	8.93 (8.84 to 9.02); 43,523 (87.2)	8.88 (8.78 to 8.98); 6397 (12.8)	–0.04 (–0.09 to 0.01)	.08
Asian	8.52 (8.27 to 8.77); 254 (3.6)	8.88 (8.78 to 8.98); 6773 (96.4)	0.36 (0.12 to 0.59)	.003[b]
Black	9.17 (9.02 to 9.31); 1061 (14.4)	9.19 (9.09 to 9.29); 6284 (85.6)	0.02 (–0.11 to 0.15)	.76
Hispanic	9.37 (9.18 to 9.55); 669 (9.7)	9.24 (9.14 to 9.34); 6240 (90.3)	–0.13 (–0.30 to 0.04)	.13
Gender				
Women	9.06 (8.97 to 9.15); 33,854 (79.7)	9.07 (8.98 to 9.17); 8610 (20.3)	0.01 (–0.03 to 0.06)	.61
Men	8.68 (8.58 to 8.77); 9706 (33.8)	8.82 (8.73 to 8.92); 19,031 (66.2)	0.15 (0.10 to 0.20)	<.001[b]

[a]PT: physical therapist.

[b]Significant *P* values.

Clinical Outcomes

The impact of race/ethnicity and gender on clinical outcomes is presented in Figure 3 and Table S3 in Multimedia Appendix 1.

Significant and similar pain response rates (adjusted: 62.16%, 95% CI 52.45-70.99, to 68.88%, 95% CI 63.10-74.12) were observed across different races/ethnicities. Regarding gender, men reported a slightly higher response rate when supported by women PTs (discordant dyads) in relation to concordant dyads (aOR 1.08, 95% CI 1.01-1.15, *P*=.03). Alongside pain recovery, daily activity impairment followed a similar trend, with high and significant response rates across all races/ethnicities and genders (69.03%, 95% CI 56.94-78.98 to 74.46%, 95% CI 70.81-77.81) regardless of concordance (all *P*>.05; Figure 3). Anxiety was similarly not impacted by concordance. In contrast, White patients had a higher depression response rate when supported by PTs from other races/ethnicities (aOR 1.20, 95% CI 1.02-1.39; reference: concordance), while men supported by men PTs were more likely to report significant depression recovery than discordant dyads (aOR 1.23, 95% CI 1.05-1.47, *P*=.01; reference: discordance).

Effect of Covariates on Outcomes

The effect of covariates on outcomes among concordant/discordant dyads is depicted in Figure S2 in Multimedia Appendix 1 for race/ethnicity and Figure S3 in Multimedia Appendix 1 for gender.

Several covariates had an effect on the completion rate in both analyses (race/ethnicity concordance and gender concordance). A higher BMI, pain, depression and anxiety levels, and lower education levels were associated with a lower completion rate.

XSL•FO
RenderX

In contrast, older age and not being employed (unemployed or retired) were associated with a higher completion rate. A lower pain response rate was observed across those with a higher BMI, lower education, and higher mental health scores. Chronic conditions and not being employed were also associated with smaller improvements in clinical outcomes.

Discussion

Principal Findings

The findings of this study demonstrated that being concordant on race/ethnicity and gender did not translate into higher satisfaction or better clinical outcomes in asynchronous digital care, despite the positive effect on treatment completion. Concordant dyads were associated with higher completion rates when compared to their counterparts, particularly for White participants (aOR 1.11, 95% CI 1.05-1.19, *P*<.001), Hispanic participants (aOR 1.27, 95% CI 1.08-1.54, *P*=.009), and women (aOR 1.10, 95% CI 1.06-1.18, *P*<.001). Race and gender dyads had frequent interactions, ranging from 16.8 to 23.01, with a minimal difference between groups. An exception was observed among Asian discordant dyads, which exchanged an adjusted mean of 5.32 more texts (95% CI 3.28-7.36, *P*<.001) than concordant dyads.

Program satisfaction was high regardless of concordance status (>8.52, 95% CI 8.27-8.77). Race/ethnicity concordance did not impact clinical outcomes for pain, anxiety, and daily activity impairment, as all improved similarly (eg, pain response rates ranged from 62.16%, 95% CI 52.45-70.99, to 68.88%, 95% CI 63.10-74.12, across different race/ethnicity dyads). An exception was observed for depression, in which White patients experienced a higher success rate when matched with PTs from other races/ethnicities (aOR 1.20, 95% CI 1.02-1.39; reference: concordance). In terms of gender concordance, men had a slightly higher pain response rate in discordant dyads (aOR 1.08, 95% CI 1.01-1.15, *P*=.03; reference: concordance) and a higher depression response rate in concordant dyads (aOR 1.23, 95% CI 1.05-1.47, *P*=.01; reference: discordance).

Comparison With the Literature

Interactions in a digital context have distinct characteristics as they are primarily facilitated through video calls and text messaging. These forms of communication have been shown to reduce fears of embarrassment and foster a sense of support [41,42], as providers are readily accessible. However, it remains unclear whether the demographic characteristics of providers, and their concordance with patients, significantly influence intervention outcomes. Studies evaluating the effect of race/ethnicity or gender concordance on outcomes have been mainly conducted in primary/outpatient care [7-9,43,44] and behavioral health [45-47], making this study the first of its kind in the context of remote digital health. Herein, we explored this association in a large cohort whose demographic characteristics matched the US general population [48], particularly those reporting MSK pain [49]. Analyses were adjusted for key baseline characteristics that differed statistically between concordant and discordant groups and were previously reported to influence outcomes [37-39]. This approach reduces bias and permits a clearer understanding of the effect of racial/ethnic and gender concordance on outcomes. Our findings showed that concordance does not have a uniformly significant impact across all race/ethnicity, gender, and outcomes, as previously reported [15,50]. Importantly, the differences noted in some variables did not form a sustained pattern that can be operationalized in clinical practice.

Herein, concordance seemed to support higher completion rates, particularly among White and Hispanic groups and among women. Albeit in a different setting, a previous study reported that race concordance is important for retention in psychology interventions in adolescents [47]. Points of similarity may help enhance therapeutic relationships, resulting in a more trusting patient, willing to commit to the therapeutic plan developed collaboratively [51,52]. For example, women patients may find it easier to trust women PTs, who in turn may be more likely to empathize with them, given their common backgrounds [53].

High interaction levels were observed across all race/ethnicity and gender dyads, which are suggestive of therapeutic alliance establishment regardless of concordance. Further investigation will allow us to fully understand which factors of asynchronous communication may contribute more significantly to rapport building and therapeutic alliance establishment.

Similarly, high satisfaction was reported across dyads but was not generally impacted by concordance. Asian patients also reported slightly higher satisfaction when paired in discordant dyads, although this difference was small (adjusted β coefficient 0.36, 95% CI 0.12-0.59, *P*=.003), suggesting that it may lack clinical significance. The reason underlying these findings is unclear but may be related to unmeasured confounders. One can speculate that the diversity within the U.S. Census "Asian" category, encompassing groups ranging from Koreans to Asian Indians, may not reflect the true cultural or linguistic heterogeneity of the group [50,54]. The literature evaluating the effect of racial concordance on satisfaction has also shown inconclusive and contradictory findings [15].

Race/ethnicity concordance was not associated with significant improvements in most clinical outcomes (pain, daily activity impairment, and anxiety). An exception was depression, in which PTs' demographic characteristics may have helped foster better outcomes. White discordant dyads reported better depression response rates, suggesting that although shared identity may facilitate a trustworthy relationship and engagement with the treatment, it is not the prime determining factor in sustaining a significant clinical outcome [15]. Additional factors, in addition to provider background, can carry significant weight in determining outcomes. In studies examining outcome predictors across nearly all pain conditions and therapies, the risk factors for treatment failure almost entirely comprise patient-related variables, including psychiatric morbidities, socioeconomic factors, or other clinical characteristics (eg, secondary gain, obesity, and smoking) [55-57]. Although some of these variables are modifiable, many are not, and even for those which are, addressing them may be time-consuming, prohibitively expensive, and fraught with other roadblocks.

Herein, the clinical team consisted of experienced clinicians, many with subspecialty training, all trained in equity, cultural humility, and unconscious bias, which may have reduced any

XSL•FO
RenderX

subtle effects of concordance to the point of being undetectable with our study design [58,59]. Following Centers for Disease Control and Prevention (CDC) recommendations, training in cultural safety and humility is vital for empowering clinicians to adapt to diverse cultural groups, mitigating implicit racial bias and imbalances in patient-provider relationships, and promoting patient-centered care [44,60,61]. This is especially important in the light of the scarcity of diversity among providers compared to the general population, which is reflected in the physical therapy workforce [62,63].

Regarding the effect of gender concordance on clinical outcomes, only depression was influenced by concordance, with men showing a greater recovery trajectory when matched with a man PT. The reasons behind this finding are unknown, and the literature evaluating the effect of gender concordance on depression and other psychiatric outcomes is scarce. Depression is often deeply rooted in emotional beliefs and experiences [64], where societal and cultural aspects might play a significant role and influence patient-provider interactions [52]. Men might have felt more open and less stigmatized to share and discuss mental health issues when treated by a PT sharing the same gender identity. Interestingly, in the context of psychotherapy, men seem to exhibit no preference for therapist gender [45]. This evidence suggests that preferences observed in one clinical field may not necessarily apply to others and that caution is needed when generalizing findings across different outcomes, settings, or specialties, as the effects of concordance may be unique to each outcome and specific context.

Strengths and Limitations

This study contains several important strengths: (1) the topic itself—the effect of race/ethnicity and gender concordance on engagement and clinical outcomes in the context of digital health using a fully remote DCP, laying the foundation for future research on this issue in the context of digital health; (2) the large sample size and proportional representation of patients from minority backgrounds, addressing a limitation identified previously [7,65]; and (3) the inclusion of validated outcome measures, previously lacking in comparable literature [50].

This study also displays limitations that warrant discussion. Considering that this DCP was mostly asynchronous, the results may not be generalizable for typical in-person or synchronous telerehabilitation contexts, where the face-to-face interaction between PTs and members is expected to be higher. Although the analysis included a wide range of patient demographic and clinical characteristics, it did not account for other variables that influence a patient's experience with the provider (eg, years of experience, subspecialty training, and personality traits and compatibility). Additional subjective measures, such as empathy and cultural competency, which are difficult to capture, may have also impacted outcomes. Another limitation is the heterogeneity within race/ethnicity categories, particularly among Asian patients (which includes widely diverse groups). Although we considered adjusting for the random effects of the PT, the high variability of patients allocated to each PT precluded this analysis and should be addressed in future studies. Additionally, although potential bias from missing data cannot be ruled out, using the last observation carried forward provides the most conservative results.

Future research should focus on the impact of communication channels, cadence, and duration on the therapeutic alliance, particularly when comparing different delivery models (eg, in-person physical therapy). Additional research should address the impact of training (cultural, biased, or otherwise) and other PT characteristics as potential predictors of recovery outcomes in physical therapy. Another area for investigation is the concept of intersectionality in concordance (eg, multiracial and lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, and other [LGBTQIA+] identity). Understanding these aspects is crucial, given the current demographic trends in the US and world populations, and could lead to more personalized and effective care strategies.

Conclusion

Race/ethnicity and gender concordance between patients and PTs translated into higher treatment completion but did not impact satisfaction. For some clinical outcomes, discordant dyads reported greater improvements, suggesting that addressing provider diversity shortages requires more than just matching race/ethnicity or gender. Yet, the findings of this study did not identify a consistent pattern, suggesting that PTs' characteristics, namely cultural safety training, experience, subspecialty training, and linguistic ability, may significantly impact engagement and outcomes in diverse populations. Future interventions should focus on these aspects to better understand their potential in optimizing patient experience and outcomes for MSK conditions.

Acknowledgments

The authors extend their gratitude to the team of physical therapists for their diligent participant management. The authors also acknowledge the contributions of João Tiago Silva, Margarida Morais, and Guilherme Freches (Sword Health, Inc) to data validation and Ashley Bass (former Sword Health, Inc, employee) for diversity, equity, and inclusion insights. This work was supported by Sword Health, Inc, and also developed within the scope of project number 62 ("Responsible AI"), financed by European Union funds, namely the Recovery and Resilience Plan—"Componente 5: Agendas Mobilizadoras para a Inovação Empresarial," included in the NextGenerationEU funding program.

Data Availability

The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to privacy restrictions.

XSL•FO
RenderX

Authors' Contributions

Conceptualization was managed by FC; data curation by ACA and MM; formal analysis by ACA; writing—original draft preparation by ACA, DD, and FC; funding by VB; and interpretation of data and writing—reviewing and editing by all authors. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

ACA, DJ, MM, CM, VY, FDC, and FC were employed by the company Sword Health, Inc, the sponsor of this study. VB, VY, and FDC also hold equity in Sword Health, Inc, and VB is the chief executive officer of the same company. SPC is an independent scientific and clinical consultant who received adviser honorarium from Sword Health, Inc. The study sponsor, Sword Health, Inc, was involved in the study design, data collection and interpretation, and writing of the manuscript.

Multimedia Appendix 1

Table S1: STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting checklist. Supplementary. Table S2: Baseline characteristics for the overall cohort and stratified by race/ethnicity concordance. Figure S1: Comparison of distribution of physical therapists between the study clinical team and the American Physical Therapy Association (APTA) in terms of (A) race/ethnicity and (B) gender. The study clinical team demographics were obtained from employer records. Table S3: Generalized linear model analysis of clinical outcome estimates by patient race and gender (reference: physical therapist-patient concordance). Figure S2: Effect of covariates on the outcome within the race concordance/discordance analysis. Figure S3. Effect of covariates on the outcome within the gender concordance/discordance analysis.

[PDF File (Adobe PDF File), 604 KB - jmir_v26i1e65354_app1.pdf]

References

1. Lane E, Barnes C, Fritz JM. Differences in pain experience among different racial and ethnic groups. Phys Ther 2024 Jan 05:pzae001. [doi: 10.1093/ptj/pzae001] [Medline: 38181396]
2. Odonkor C, Esparza R, Flores L, Verduzco-Gutierrez M, Escalon M, Solinsky R, et al. Disparities in Health Care for Black Patients in Physical Medicine and Rehabilitation in the United States: A Narrative Review. PM&R 2020 Oct 08;13(2):180-203. [doi: 10.1002/pmrj.12509] [Medline: 33090686]
3. Richter RR, Chrusciel T, Salsich G, Austin T, Scherrer JF. Disparities exist in physical therapy utilization and time to utilization between black and white patients with musculoskeletal pain. Phys Ther 2022 Oct 06;102(10):pzac095. [doi: 10.1093/ptj/pzac095] [Medline: 35871435]
4. Vespa J, Medina L, Armstrong D. Demographic turning points for the United States: population projections for 2020 to 2060. U.S. Government Publishing Office. 2018. URL: https://www.govinfo.gov/app/details/GOVPUB-C3-PURL-gpo93743 [accessed 2024-07-16]
5. Salsberg E, Richwine C, Westergaard S, Portela Martinez M, Oyeyemi T, Vichare A, et al. Estimation and Comparison of Current and Future Racial/Ethnic Representation in the US Health Care Workforce. JAMA Netw Open 2021 Mar 01;4(3):e213789 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.3789] [Medline: 33787910]
6. Xierali IM, Nivet MA. The racial and ethnic composition and distribution of primary care physicians. J Health Care Poor Underserved 2018 Jan 01;29(1):556-570 [FREE Full text] [doi: 10.1353/hpu.2018.0036] [Medline: 29503317]
7. Takeshita J, Wang S, Loren AW, Mitra N, Shults J, Shin DB, et al. Association of Racial/Ethnic and Gender Concordance Between Patients and Physicians With Patient Experience Ratings. JAMA Netw Open 2020 Nov 02;3(11):e2024583 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.24583] [Medline: 33165609]
8. Rogo-Gupta LJ, Haunschild C, Altamirano J, Maldonado YA, Fassiotto M. Physician gender is associated with Press Ganey patient satisfaction scores in outpatient gynecology. Womens Health Issues 2018 May;28(3):281-285. [doi: 10.1016/j.whi.2018.01.001] [Medline: 29429946]
9. Ku L, Vichare A. The association of racial and ethnic concordance in primary care with patient satisfaction and experience of care. J Gen Intern Med 2022 Jun 10;38(3):727-732 [FREE Full text] [doi: 10.1007/s11606-022-07695-y] [Medline: 35688996]
10. Alexander J, Hearld L, Mittler JN. Patient–physician role relationships and patient activation: the moderating effects of race and ethnicity. Med Care Res Rev 2014 Jul 15;71(5):472-495. [doi: 10.1177/1077558714541967] [Medline: 25027408]
11. Bhati KS. Effect of client-therapist gender match on the therapeutic relationship: an exploratory analysis. Psychol Rep 2014 Oct 01;115(2):565-583. [doi: 10.2466/21.02.PR0.115c23z1] [Medline: 25243363]
12. Jerant A, Bertakis KD, Fenton JJ, Tancredi DJ, Franks P. Patient-provider sex and race/ethnicity concordance: a national study of healthcare and outcomes. Med Care 2011 Nov;49(11):1012-1020. [doi: 10.1097/MLR.0b013e31823688ee] [Medline: 22002644]
13. Behn A, Davanzo A, Errázuriz P. Client and therapist match on gender, age, and income: does match within the therapeutic dyad predict early growth in the therapeutic alliance? J Clin Psychol 2018 Mar 24;74(9):1403-1421. [doi: 10.1002/jclp.22616] [Medline: 29573351]

XSL•FO
RenderX

14. Sweeney CF, Zinner D, Rust G, Fryer GE. Race/ethnicity and health care communication: does patient-provider concordance matter. Med Care 2016 Nov;54(11):1005-1009. [doi: 10.1097/MLR.0000000000000578] [Medline: 27213546]
15. Otte SV. Improved patient experience and outcomes: is patient-provider concordance the key? J Patient Exp 2022 May 29;9:23743735221103033 [FREE Full text] [doi: 10.1177/23743735221103033] [Medline: 35664933]
16. Cui D, Janela D, Costa F, Molinos M, Areias AC, Moulder RG, et al. Randomized-controlled trial assessing a digital care program versus conventional physiotherapy for chronic low back pain. NPJ Digit Med 2023 Jul 07;6(1):121 [FREE Full text] [doi: 10.1038/s41746-023-00870-3] [Medline: 37420107]
17. Areias AC, Costa F, Janela D, Molinos M, Moulder RG, Lains J, et al. Long-term clinical outcomes of a remote digital musculoskeletal program: an ad hoc analysis from a longitudinal study with a non-participant comparison group. Healthcare 2022 Nov 23;10(12):2349 [FREE Full text] [doi: 10.3390/healthcare10122349] [Medline: 36553873]
18. Janela D, Costa F, Weiss B, Areias AC, Molinos M, Scheer JK, et al. Effectiveness of biofeedback-assisted asynchronous telerehabilitation in musculoskeletal care: a systematic review. Digit Health 2023 Jun 09;9:20552076231176696 [FREE Full text] [doi: 10.1177/20552076231176696] [Medline: 37325077]
19. Scheer J, Costa F, Molinos M, Areias A, Janela D, Moulder RG, et al. Racial and ethnic differences in outcomes of a 12-week digital rehabilitation program for musculoskeletal pain: prospective longitudinal cohort study. J Med Internet Res 2022 Oct 31;24(10):e41306 [FREE Full text] [doi: 10.2196/41306] [Medline: 36189963]
20. Areias AC, Molinos M, Moulder RG, Janela D, Scheer JK, Bento V, et al. The potential of a multimodal digital care program in addressing healthcare inequities in musculoskeletal pain management. npj Digit Med 2023 Oct 10;6(1):188 [FREE Full text] [doi: 10.1038/s41746-023-00936-2] [Medline: 37816899]
21. Costa F, Janela D, Molinos M, Lains J, Francisco GE, Bento V, et al. Telerehabilitation of acute musculoskeletal multi-disorders: prospective, single-arm, interventional study. BMC Musculoskelet Disord 2022 Jan 04;23(1):29 [FREE Full text] [doi: 10.1186/s12891-021-04891-5] [Medline: 34983488]
22. Janela D, Costa F, Areias A, Molinos M, Moulder R, Lains J, et al. Digital care programs for chronic hip pain: a prospective longitudinal cohort study. Healthcare 2022 Aug 22;10(8):1595. [doi: 10.3390/healthcare10081595] [Medline: 36011251]
23. Areias AC, Costa F, Janela D, Molinos M, Moulder RG, Lains J, et al. Impact on productivity impairment of a digital care program for chronic low back pain: a prospective longitudinal cohort study. Musculoskelet Sci Pract 2023 Feb;63:102709 [FREE Full text] [doi: 10.1016/j.msksp.2022.102709] [Medline: 36543719]
24. Areias AC, Janela D, Moulder RG, Molinos M, Bento V, Moreira C, et al. Applying AI to Safely and Effectively Scale Care to Address Chronic MSK Conditions. J Clin Med 2024 Jul 26;13(15):4366 [FREE Full text] [doi: 10.3390/jcm13154366] [Medline: 39124635]
25. Measuring racial and ethnic diversity for the 2020 census. United States Census Bureau. 2021. URL: https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html [accessed 2024-07-16]
26. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American College of Rheumatology/Arthritis Foundation guideline for the management of osteoarthritis of the hand, hip, and knee. Arthritis Care Res 2020 Jan 06;72(2):149-162. [doi: 10.1002/acr.24131] [Medline: 31908149]
27. National Guideline Centre (UK). Low back pain and sciatica in over 16s: assessment and management. National Institute for Health and Care Excellence (NICE). London, United Kingdom; 2016. URL: https://www.nice.org.uk/guidance/ng59/resources/low-back-pain-and-sciatica-in-over-16s-assessment-and-management-pdf-1837521693637 [accessed 2023-11-27]
28. Qaseem A, Wilt TJ, McLean RM, Forciea MA. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2017 Feb 14;166(7):514-530. [doi: 10.7326/M16-2367] [Medline: 28192789]
29. Williams ACDC, Fisher E, Hearn L, Eccleston C. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2020 Aug 12;8:CD007407. [doi: 10.1002/14651858.CD007407.pub4] [Medline: 32794606]
30. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008 Feb;9(2):105-121. [doi: 10.1016/j.jpain.2007.09.005] [Medline: 18055266]
31. Ospina MB, Dennett L, Waye A, Jacobs P, Thompson AH. A systematic review of measurement properties of instruments assessing presenteeism. Am J Manag Care 2015 Feb 01;21(2):e171-e185 [FREE Full text] [Medline: 25880491]
32. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
33. Bijker L, Sleijser-Koehorst MLS, Coppieters MW, Cuijpers P, Scholten-Peeters GGM. Preferred self-administered questionnaires to assess depression, anxiety and somatization in people with musculoskeletal pain - a modified Delphi study. J Pain 2020 Mar;21(3-4):409-417 [FREE Full text] [doi: 10.1016/j.jpain.2019.08.006] [Medline: 31487562]
34. Toussaint A, Hüsing P, Gumz A, Wingenfeld K, Härter M, Schramm E, et al. Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder questionnaire (GAD-7). J Affect Disord 2020 Mar 15;265:395-401. [doi: 10.1016/j.jad.2020.01.032] [Medline: 32090765]
35. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

36. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care 2004 Dec;42(12):1194-1201. [doi: 10.1097/00005650-200412000-00006] [Medline: 15550799]
37. Artus M, Campbell P, Mallen CD, Dunn KM, van der Windt DAW. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. BMJ Open 2017 Jan 17;7(1):e012901 [FREE Full text] [doi: 10.1136/bmjopen-2016-012901] [Medline: 28096253]
38. Verkerk K, Luijsterburg PAJ, Heymans MW, Ronchetti I, Pool - Goudzwaard A, Miedema HS, et al. Prognosis and course of pain in patients with chronic non - specific low back pain: a 1 - year follow - up cohort study. Eur J Pain 2015 Jan 06;19(8):1101-1110. [doi: 10.1002/ejp.633] [Medline: 25565501]
39. Nieminen LK, Pyysalo LM, Kankaanpää MJ. Prognostic factors for pain chronicity in low back pain: a systematic review. Pain Rep 2021 Jan 01;6(1):e919 [FREE Full text] [doi: 10.1097/PR9.0000000000000919] [Medline: 33981936]
40. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. A classification of chronic pain for ICD-11. Pain 2015 Jun;156(6):1003-1007 [FREE Full text] [doi: 10.1097/j.pain.0000000000000160] [Medline: 25844555]
41. Lewis E, Hassmén P, Pumpa KL. Participant perspectives of a telehealth trial investigating the use of telephone and text message support in obesity management: a qualitative evaluation. BMC Health Serv Res 2021 Jul 09;21(1):675 [FREE Full text] [doi: 10.1186/s12913-021-06689-6] [Medline: 34243772]
42. Huxley CJ, Atherton H, Watkins JA, Griffiths F. Digital communication between clinician and patient and the impact on marginalised groups: a realist review in general practice. Br J Gen Pract 2015 Nov 29;65(641):e813-e821. [doi: 10.3399/bjgp15X687853] [Medline: 26622034]
43. Walker RJ, Dawson AZ, Thorgerson A, Campbell JA, Engel S, Kastner M, et al. Relationship between satisfaction scores and racial/ethnic and sex concordance in primary care. Healthcare 2023 Aug 12;11(16):2276 [FREE Full text] [doi: 10.3390/healthcare11162276] [Medline: 37628474]
44. Lin OM, Reid HW, Fabbro RL, Johnson KS, Batch BC, Olsen MK, et al. Association of provider perspectives on race and racial health care disparities with patient perceptions of care and health outcomes. Health Equity 2021 Jul 01;5(1):466-475 [FREE Full text] [doi: 10.1089/heq.2021.0018] [Medline: 34316530]
45. Ilagan GS, Heatherington L. Advancing the understanding of factors that influence client preferences for race and gender matching in psychotherapy. Couns Psychol Q 2021 Jul 29;35(3):694-717. [doi: 10.1080/09515070.2021.1960274]
46. Cheng AW, Nakash O, Cruz-Gonzalez M, Fillbrunn MK, Alegría M. The association between patient–provider racial/ethnic concordance, working alliance, and length of treatment in behavioral health settings. Psychol Serv 2023;20(Suppl 1):145-156. [doi: 10.1037/ser0000582] [Medline: 34472952]
47. Wintersteen MB, Mensinger JL, Diamond GS. Do gender and racial differences between patient and therapist affect therapeutic alliance and treatment retention in adolescents? Prof Psychol Res Pract 2005 Aug;36(4):400-408. [doi: 10.1037/0735-7028.36.4.400]
48. S0201 selected population profile in the United States. United States Census Bureau. 2022. URL: https://data.census.gov/table?t=-00:-01:Race%20and%20Ethnicity [accessed 2024-07-31]
49. Burden of musculoskeletal diseases in the United States (BMUS). United States Bone Joint Initiative. 3rd Edition. 2015. URL: https://www.usbji.org/index.php?option=com_content&view=article&id=50:burden-of-musculoskeletal-diseases-in-the-united-states&catid=20:site-content&Itemid=148
50. Miller AN, Duvuuri VNS, Vishanagra K, Damarla A, Hsiao D, Todd A, et al. The relationship of race/ethnicity concordance to physician-patient communication: a mixed-methods systematic review. Health Commun 2023 Jun 20;39(8):1543-1557. [doi: 10.1080/10410236.2023.2223402] [Medline: 37338139]
51. Cruwys T, Lee G, Robertson A, Haslam C, Sterling N, Platow M, et al. Therapists who foster social identification build stronger therapeutic working alliance and have better client outcomes. Compr Psychiatry 2023 Jul 01;124:152394 [FREE Full text] [doi: 10.1016/j.comppsych.2023.152394] [Medline: 37216806]
52. Eken HN, Dee EC, Powers AR, Jordan A. Racial and ethnic differences in perception of provider cultural competence among patients with depression and anxiety symptoms: a retrospective, population-based, cross-sectional analysis. Lancet Psychiatry 2021 Nov;8(11):957-968 [FREE Full text] [doi: 10.1016/S2215-0366(21)00285-6] [Medline: 34563316]
53. Hall J, Roter D. Do patients talk differently to male and female physicians? A meta-analytic review. Patient Educ Couns 2002 Dec 01;48(3):217-224. [doi: 10.1016/s0738-3991(02)00174-x] [Medline: 12477606]
54. Gordon NP, Lin TY, Rau J, Lo JC. Aggregation of Asian-American subgroups masks meaningful differences in health and health risks among Asian ethnicities: an electronic health record based cohort study. BMC Public Health 2019 Nov 25;19(1):1551 [FREE Full text] [doi: 10.1186/s12889-019-7683-3] [Medline: 31760942]
55. Cohen SP, Doshi TL, Kurihara C, Reece D, Dolomisiewicz E, Phillips CR, et al. Multicenter study evaluating factors associated with treatment outcome for low back pain injections. Reg Anesth Pain Med 2021 Dec 08;47(2):89-99 [FREE Full text] [doi: 10.1136/rapm-2021-103247] [Medline: 34880117]
56. Wilkens P, Scheel IB, Grundnes O, Hellum C, Storheim K. Prognostic factors of prolonged disability in patients with chronic low back pain and lumbar degeneration in primary care: a cohort study. Spine 2013 Jan 01;38(1):65-74. [doi: 10.1097/BRS.0b013e318263bb7b] [Medline: 22718223]

XSL•FO
RenderX

57. Linton SJ, Shaw WS. Impact of psychological factors in the experience of pain. Phys Ther 2011 May;91(5):700-711. [doi: 10.2522/ptj.20100330] [Medline: 21451097]
58. Dickson T, Wilkinson T, Hegedus E, Gisselman A. Transforming the patient experience: moving beyond cultural competence to cultural safety. Phys Ther 2022 Jun 03;102(6):pzac034. [doi: 10.1093/ptj/pzac034] [Medline: 35358317]
59. Lekas H, Pahl K, Fuller Lewis C. Rethinking cultural competence: shifting to cultural humility. Health Serv Insights 2020 Dec 20;13:1178632920970580 [FREE Full text] [doi: 10.1177/1178632920970580] [Medline: 33424230]
60. Cultural competence in health and human services. Centers for Disease Control and Prevention (CDC). 2024. URL: https://npin.cdc.gov/pages/cultural-competence-health-and-human-services#what [accessed 2024-07-30]
61. Cultural competence. American Physical Therapy Association. 2020. URL: https://www.apta.org/siteassets/pdfs/policies/cultural-competence.pdf [accessed 2024-07-30]
62. Physical therapy worforce analysis: a report from the American Physical Therapy Association. American Physical Therapy Association. 2020 Dec. URL: https://www.apta.org/contentassets/5997bfa5c8504df789fe4f1c01a717eb/apta-workforce-analysis-2020.pdf [accessed 2024-07-30]
63. Yeowell G. 'Isn't it all whites?' Ethnic diversity and the physiotherapy profession. Physiotherapy 2013 Dec;99(4):341-346. [doi: 10.1016/j.physio.2013.01.004] [Medline: 23537883]
64. Remes O, Mendes JF, Templeton P. Biological, psychological, and social determinants of depression: a review of recent literature. Brain Sci 2021 Dec 10;11(12):1633 [FREE Full text] [doi: 10.3390/brainsci11121633] [Medline: 34942936]
65. Schoenthaler A, Ravenell J. Understanding the Patient Experience Through the Lenses of Racial/Ethnic and Gender Patient-Physician Concordance. JAMA Netw Open 2020 Nov 02;3(11):e2025349 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.25349] [Medline: 33165607]

Abbreviations

aOR: adjusted odds ratio
BMI: body mass index
CBT: cognitive behavioral therapy
DCP: digital care program
GAD-7: Generalized Anxiety Disorder 7-item scale
MCIC: minimum clinically important change
MSK: musculoskeletal
NPRS: Numerical Pain Rating Scale
PHQ-9: Patient Health Questionnaire 9-item
PT: physical therapist
WPAI: Work Productivity and Activity Impairment

Edited by G Eysenbach, A Schwartz; submitted 13.08.24; peer-reviewed by A Bove, S Rousset; comments to author 19.09.24; revised version received 25.09.24; accepted 08.10.24; published 29.10.24.

Please cite as:
Areias AC, Janela D, Molinos M, Bento V, Moreira C, Yanamadala V, Cohen SP, Correia FD, Costa F
Exploring the Importance of Race and Gender Concordance Between Patients and Physical Therapists in Digital Rehabilitation for Musculoskeletal Conditions: Observational, Longitudinal Study
J Med Internet Res 2024;26:e65354
URL: https://www.jmir.org/2024/1/e65354
doi:10.2196/65354
PMID:39470695

©Anabela C Areias, Dora Janela, Maria Molinos, Virgílio Bento, Carolina Moreira, Vijay Yanamadala, Steven P Cohen, Fernando Dias Correia, Fabíola Costa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characterization of Telecare Conversations on Lifestyle Management and Their Relation to Health Care Utilization for Patients with Heart Failure: Mixed Methods Study

Mojisola Erdt[1], PhD; Sakinah Binte Yusof[1], BA; Liquan Chai[2], MSc; Siti Umairah Md Salleh[1], BA; Zhengyuan Liu[1], MSc; Halimah Binte Sarim[3]; Geok Choo Lim[3]; Hazel Lim[1], BSocSci; Nur Farah Ain Suhaimi[1], BA; Lin Yulong[1], BA, MEng; Yang Guo[1], MSc; Angela Ng[3], BSN; Sharon Ong[3], BSN; Bryan Peide Choo[3], MPH; Sheldon Lee[3], MBBS, MMed; Huang Weiliang[3], MBBS, MRCP (UK); Hong Choon Oh[3], PhD; Maria Klara Wolters[2], PhD; Nancy F Chen[1], PhD; Pavitra Krishnaswamy[1], PhD

[1]Institute for Infocomm Research (I[2]R), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

[2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

[3]Changi General Hospital, Singapore, Singapore

Corresponding Author:
Pavitra Krishnaswamy, PhD
Institute for Infocomm Research (I²R)
Agency for Science, Technology and Research (A*STAR)
1 Fusionopolis Way, #21-01
Connexis South Tower
Singapore, 138632
Singapore
Phone: 65 6408 2450
Email: pavitrak@i2r.a-star.edu.sg

Abstract

Background: Telehealth interventions where providers offer support and coaching to patients with chronic conditions such as heart failure (HF) and type 2 diabetes mellitus (T2DM) are effective in improving health outcomes. However, the understanding of the content and structure of these interactions and how they relate to health care utilization remains incomplete.

Objective: This study aimed to characterize the content and structure of telecare conversations on lifestyle management for patients with HF and investigate how these conversations relate to health care utilization.

Methods: We leveraged real-world data from 50 patients with HF enrolled in a postdischarge telehealth program, with the primary intervention comprising a series of telephone calls from nurse telecarers over a 12-month period. For the full cohort, we transcribed 729 English-language calls and annotated conversation topics. For a subcohort (25 patients with both HF and T2DM), we annotated lifestyle management content with fine-grained dialogue acts describing typical conversational structures. For each patient, we identified calls with unusually high ratios of utterances on lifestyle management as lifestyle-focused calls. We further extracted structured data for inpatient admissions from 6 months before to 6 months after the intervention period. First, to understand conversational structures and content of lifestyle-focused calls, we compared the number of utterances, dialogue acts, and symptom attributes in lifestyle-focused calls to those in calls containing but not focused on lifestyle management. Second, to understand the perspectives of nurse telecarers on these calls, we conducted an expert evaluation where 2 nurse telecarers judged levels of concern and follow-up actions for lifestyle-focused and other calls (not focused on lifestyle management content). Finally, we assessed how the number of lifestyle-focused calls relates to the number of admissions, and to the average length of stay per admission.

Results: In comparative analyses, lifestyle-focused calls had significantly fewer utterances (P=.01) and more dialogue acts (P_{adj}=.005) than calls containing but not focused on lifestyle management. Lifestyle-focused calls did not contain deeper discussions on clinical symptoms. These findings indicate that lifestyle-focused calls entail short, intense discussions with greater emphasis on understanding patient experience and coaching than on clinical content. In the expert evaluation, nurse telecarers identified 24.2% (29/120) of calls assessed as concerning enough for follow-up. For these 29 calls, nurse telecarers were more attuned to concerns about symptoms and vitals (19/29, 65.5%) than lifestyle management concerns (4/29, 13.8%). The number of

XSL•FO
RenderX

lifestyle-focused calls a patient had was modestly (but not significantly) associated with a lower average length of stay for inpatient admissions (Spearman ρ=-0.30; P_{adj}=.06), but not with the number of admissions (Spearman ρ=-0.03; P_{adj}=.84).

Conclusions: Our approach and findings offer novel perspectives on the content, structure, and clinical associations of telehealth conversations on lifestyle management for patients with HF. Hence, our study could inform ways to enhance telehealth programs for self-care management in chronic conditions.

(J Med Internet Res 2024;26:e46983) doi:10.2196/46983

KEYWORDS

telehealth; telecare; heart failure; chronic disease; self-management; lifestyle management; behavior; health care utilization; conversation; dialogue; medical informatics

Introduction

Lifestyle and behavior modifications are widely recognized as cornerstones of effective management of chronic cardiometabolic conditions such as heart failure (HF) or type 2 diabetes mellitus (T2DM). For instance, HF guidelines encourage patients to commit to substantial lifestyle changes involving fluid and dietary restrictions, daily weighing, physical exercise, and adhering to complex medication regimens with regular self-monitoring [1]. Given these elaborate requirements, patients find it difficult to cope with the burden of the disease, its comorbidities, and self-care management challenges. As such, lifestyle management issues persist as unresolved last-mile challenges in cardiometabolic disease management [2-5], which, in turn, lead to potentially avoidable hospitalizations. This is especially evident among patients with HF, whose inpatient costs constitute the largest component of their overall economic burden [6].

In response, numerous education, coaching, and self-management interventions have emerged to support patients in self-care management [7,8]. Increasingly, to enhance access and convenience, these interventions are delivered through telehealth [9-11]. Telehealth interventions for chronic disease management have been evaluated in randomized controlled trials [12,13] and shown to have positive effects on health outcomes and health care utilization [10,14-16].

Commonly, telehealth interventions include nurse-led telecare programs, where trained nurses (telecarers) make regular telephone calls to patients, empowering them with knowledge and skills to independently manage their chronic conditions at home. These telecare conversations serve a central role in the program and involve personalized coaching and self-management support. Such nurse-led telecare programs have proven effective in improving the quality of life and self-care ability of patients with chronic disease in the community [17,18]. However, the communicative content and conversational structure of lifestyle-focused telecare conversations and their relation to health and clinical indicators are not well understood.

Previous works have examined the content of health care conversations using quantitative or qualitative approaches: for example, using visual and correlational analysis to analyze communicative behavior in medical interactions [19,20], applying machine learning to model conversational topic content [21], predicting the patient perception of communication quality [22], and examining narrative arcs over time [23]. Qualitative studies have used thematic coding, ethnography, and conversation analysis to understand clinical conversations [24-26]. However, these studies largely focused on physical encounters and have not assessed the content and structure of lifestyle management exchanges within telecare conversations. Furthermore, related work explores how patient-clinician communication relates to objective or subjective health and clinical indicators [27-29]. For example, increased patient participation in medical decision-making has been linked to improvements in glycated hemoglobin, functional status, and quality of life [30], while physician communication skills have been linked to patient adherence [28]. However, these studies typically considered single, in-person encounters. There has been limited work on telehealth encounters, particularly in chronic disease, where the conversations evolve over time (with the patient's condition). Addressing these gaps could distill effective structures of communicative behavior that may enhance the design and delivery of telehealth interventions.

Real-world data from routine telehealth delivery offer a substrate to elucidate the content and structure of self-care management conversations and to uncover relations to health and clinical indicators. However, multimodal datasets comprising telehealth conversations and linked medical records are not readily available. Furthermore, addressing such questions requires interfacing quantitative and qualitative analysis of conversational content, communicative behavior, and clinical data, but there has been limited work on such integrative analyses. More multipronged approaches encompassing a mixture of methods could uncover nonapparent patterns of communicative behavior in telehealth settings and their interconnections with health and clinical indicators.

This study aims to characterize the content and structure of telecare conversations on lifestyle management in HF and study how these conversations relate to health care utilization. To this end, we curated a multimodal dataset comprising transcripts of telecare conversations alongside electronic health records for patients with HF enrolled in a postdischarge telehealth program in Singapore. We analyzed this dataset using a mixed methods approach. First, we quantitatively analyzed the communicative content and conversational structures underlying telecarer-patient interactions in lifestyle management calls. Second, we qualitatively assessed telecarer perspectives on how lifestyle management conversations relate to key issues and levels of concern for follow-on clinical actions. Finally, we examined the association between the number of

XSL•FO
RenderX

lifestyle-focused calls a patient received and the frequency and length of their inpatient admissions.

Methods

HF Telesupport Program

The 12-month HF telesupport program delivers postdischarge support to patients with HF. The intervention was a series of telecarer-patient telephone calls (scheduled and ad hoc) focused on providing personalized education and support. These calls were conducted in English, Mandarin, Malay, or Tamil, or a combination thereof. Each patient received 11 to 15 scheduled calls (based on the patient's hospital-assessed risk score at enrolment), vital signs telemonitoring, and surveys on symptoms and lifestyle habits. Nurses reviewed patients' vitals, symptoms, and medical records to determine who needed more urgent support and ad hoc calls. There was a 12-month postintervention follow-up period.

Study Design

Figure 1 illustrates the steps in our study. First, we annotated calls with topics and dialogue acts (Figure 1A). Second, we identified calls with heightened lifestyle management content as lifestyle-focused calls (Figure 1B). Third, we processed structured electronic health records and aligned these with call data (Figure 1C). Subsequently, we investigated the calls from 3 angles (Figure 1D): (1) analysis of the content and conversational structures of lifestyle-focused calls; (2) an expert evaluation of concern levels with respect to lifestyle-focused calls; and (3) the association of lifestyle-focused calls with frequency and length of inpatient admissions. These steps are detailed in the following subsections.

Figure 1. Overview of steps in the study, depicting (A-C) data preparation and (D) analysis of lifestyle-focused calls. HF: heart failure; LOS: length of stay.

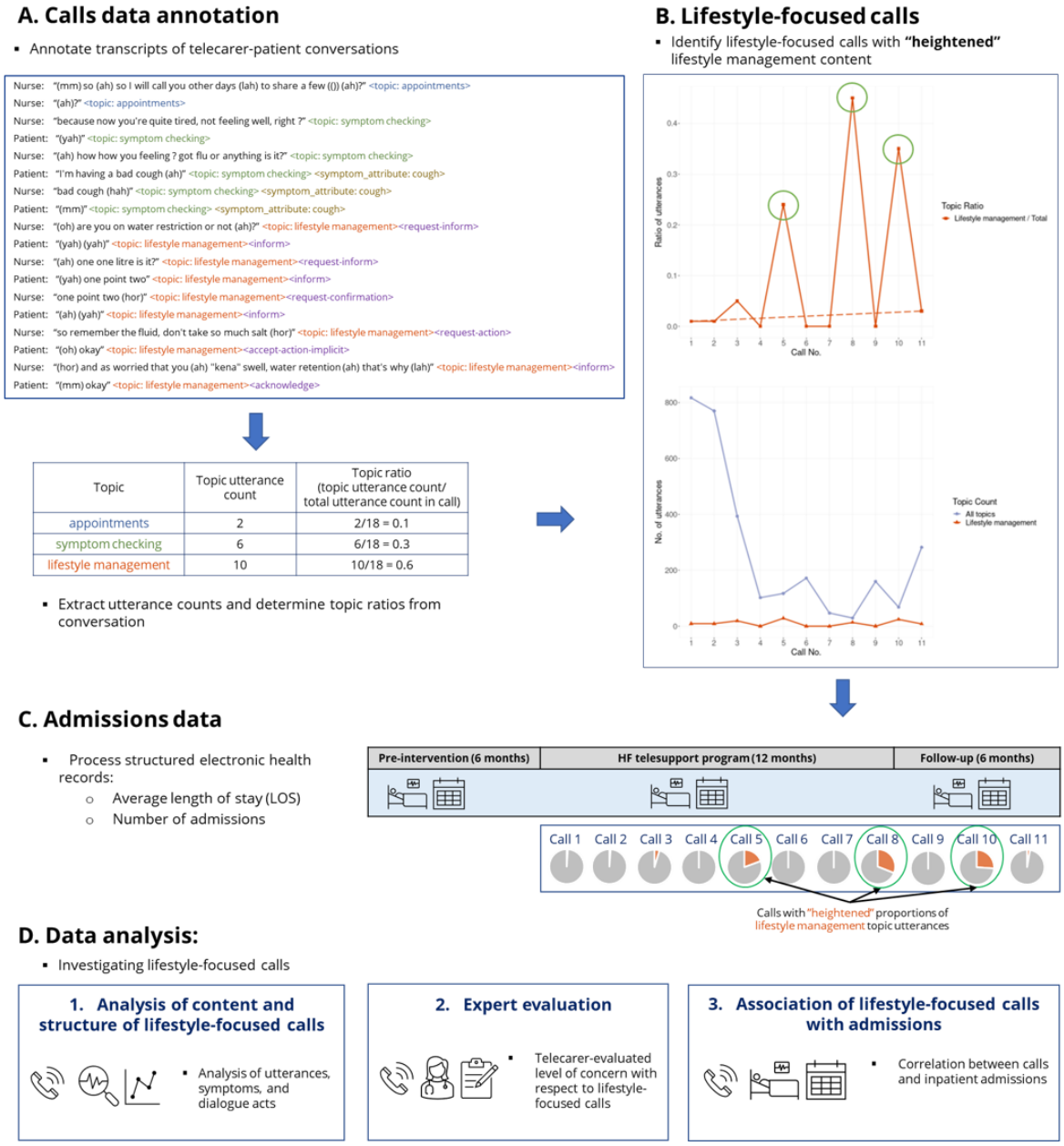

XSL•FO
RenderX

Data Preparation

Overview of Dataset

We extracted inpatient admissions data, recordings of (scheduled and ad hoc) calls (accompanied by telecarers' task notes), demographics, and medical profiles for this study. For each patient, admissions covered dates from 6 months before and 6 months after the intervention, and calls covered dates from the beginning of the intervention to 6 months after. This timeframe was patient specific. The 6-month postintervention period allowed for admissions data to accrue and to track calls afterward due to varying intervention durations across patients.

In an iterative process, we applied progressive and purposive sampling to select 50 patients with HF (full cohort) with English-language calls from the 150-patient cohort in the original nonrandomized controlled study [14]. Within the specified timeframe, the 50 patients with HF for our retrospective study had 729 English-language, content-driven calls, comprising 117 hours of audio recordings (if multiple calls were made on the same day, eg, due to a faulty connection or separate calls with the caregiver and patient, we counted these as 1 call). We transcribed, deidentified, and manually annotated these 729 calls with topics, symptoms, and attributes. Of these 50 patients, 25 patients had references to T2DM management in their calls (cohort of patients with both HF and T2DM [HFDM]). For these 25 patients (379 calls), we annotated lifestyle management content with dialogue structure information (described below). Finally, from the 25-patient HFDM cohort, we purposively sampled 11 patients who had at least 1 call with lifestyle management content and at least 1 lifestyle-focused call to facilitate a qualitative evaluation (expert evaluation cohort).

Calls Data Annotation

We used the Transcriber software (version 1.5.1; developed by Claude Barras and Edouard Geoffrois at DGA in Paris, in collaboration with LDC [University of Pennsylvania]) [31] for transcription and annotation. As the transcribers (SUMS, HL, and NFAS) listened to each audio conversation between telecarers and patients or caregivers, the speakers were manually identified and labeled. Each dialogue was split into utterances. Following convention in computational linguistics and discourse analysis [32,33], we defined utterances as meaningful units of speech delimited by clear pauses. Mostly, utterances correspond to sentences, but they can also correspond to parts of a sentence or a feedback sound such as "mm-hmm." We preserved the informal and spontaneous styles of spoken interactions, including interlocutor interruption, backchanneling, hesitation, false starts, and repetition [34]. We manually annotated all utterances with the topic of discussion (Table 1).

Table 1. Overview of topics, symptoms, attributes, speakers, and dialogue acts. Attributes may fall under any topic. Symptom attributes correspond to respective heart failure symptoms.

Variable	Annotations
Topics	Introduction, identification, appointments, telemonitoring, general education, customized coaching, related medical experience, symptom-checking, vitals, medication management, lifestyle management, social-chatting, others
Symptoms	Breathlessness, swelling, cough, dizziness, chest pain, heartbeat and palpitations, bleeding, headache
Symptom attributes	Location, frequency, extent, time, activity
Lifestyle attributes	Fluid and salt intake, smoking, and alcohol
Vitals attributes	Blood pressure, weight
Speakers	Nurse telecarer, patient, caregiver, others
Dialogue acts	
Exchanging information	Request-inform, inform
Understanding information	Acknowledge, request-clarification, request-confirmation
Performing action	Request-action, accept-action-implicit, accept-action-explicit, reject-action-implicit, reject-action-explicit
Evaluation of health condition	Evaluate, evaluate-positive, evaluate-negative
Social-emotional	Socio-emotional
Incomplete dialogue act	Back-channel, fragment, stall
Others	Others

Our study focused on conversation content related to the lifestyle management topic, comprising discussion on adherence to salt and fluid intake restrictions, tobacco and alcohol consumption, and compliance to recommended diet and exercise regimens (Multimedia Appendix 1). Unlike other topics, the lifestyle management topic uniquely focuses on understanding patient experience in relation to lifestyle management.

We annotated lifestyle management utterances (for the HFDM cohort) with dialogue acts to describe dialogue structure. Dialogue acts reflect high-level communication actions that a

XSL•FO
RenderX

speaker makes through the utterance, such as exchanging information, understanding information, performing action, evaluation of health condition, or social-emotional utterances [35]. We conducted an interrater reliability test of the dialogue acts annotated by 3 annotators (LC, SUMS, and HL) using Fleiss kappa.

Lifestyle-Focused Calls

We identified lifestyle-focused calls to investigate conversation content and communicative behavior relating to the lifestyle management topic. We consider changes in the prevalence of this topic over multiple conversations as suggestive of changes in conversational behavior for each patient, which could possibly indicate corresponding changes in their health and lifestyle management. To identify these changes, we quantified how conversations and behavior evolve over time, using a previously published approach [36,37] that leverages ratios of count and time to quantify behavioral indicators.

Specifically, for each transcribed call (for the full cohort) with lifestyle management content, we calculated utterance ratios for the lifestyle management content within the calls (Figure 1B). We defined an utterance ratio as the utterance count for a topic in a call divided by the total number of utterances in the same call across all topics. This gives a proportion of the utterances made for a specific topic in the call. We defined deviations from the norm as topic-focused calls with heightened content focusing far more on a given topic than the typical call. For each patient, we designated topic-focused calls as those calls fulfilling the criteria: (1) the topic ratio for a call is above the median regression line fitted for each patient's individual call trajectory, and (2) either the topic ratio for a call is substantially heightened in relation to that of the previous call, or the topic ratio for a call remained heightened following a prior increase from the previous call. Furthermore, 2 authors (ME and SBY) carried out a manual inspection to resolve any ambiguity in distinguishing topic-focused calls from other calls. Notably, to avoid arbitrary thresholds across patients, we highlight that the above process for the selection of topic-focused calls is patient-specific, thereby keeping the quantification of patients' conversational behavior personalized and within the context of each patient's calls.

We term calls with heightened lifestyle management content as lifestyle-focused calls and other calls with content relating to the lifestyle management topic, but not to such a heightened extent as calls with lifestyle management content (not focused).

Admissions Data

We considered admissions for the full cohort with any primary diagnosis (all-cause inpatient admissions). Elective or planned admissions, such as day surgery, were excluded. We considered the number of admissions and length of stay (LOS, or bed days) in hospital per admission as the 2 indicators for our analyses. These indicators can be considered proxies for the frequency and degree of worsening of a patient's health. For each patient, we calculated the average LOS per admission by dividing the cumulative LOS over all his or her admissions by his or her total number of admissions within the study timeframe. We could not consider HF-related admissions as the number of such admissions within our dataset was limited.

Data Analysis

Analysis of Content and Structure of Lifestyle-Focused Calls

We comparatively analyzed the number of utterances and dialogue acts in lifestyle-focused calls against those in other calls with lifestyle management content (not focused). As some calls in the 2 groups may come from the same patient, we accounted for this intrapatient correlation using linear regression with standard errors clustered by patient. We corrected for multiple comparisons using Holm-Bonferroni correction. We analyzed symptom attributes, comparing the proportion of lifestyle-focused calls with a mention of symptom attributes to that of calls with lifestyle management content (not focused). We used R software (version 4.3.3; R Foundation for Statistical Computing), including the *fixest* (version 0.12.0) and *stats* (version 4.3.3) packages. We set significance to $P<.05$.

Expert Evaluation

To understand how telecarers perceive lifestyle-focused calls, we conducted an expert evaluation study based on the Delphi method [38]. We sought to understand what call contents telecarers find concerning enough to follow up on after a lifestyle-focused call in comparison to other calls (not focused on lifestyle management content). For patients in the expert evaluation cohort, we requested 2 experienced telecarers (HBS and GCL) to review call transcripts in chronological order alongside the patient's demographic and medical profile and provide their experience-based judgment on which calls they find concerning enough to warrant follow-up actions. We asked them to select their level of concern from predefined follow-up categories already operationalized in the program: business as usual, nurse follow-up: case review, nurse follow-up: call patient, or advise the patient to go to a GP (general practitioner) or polyclinic. Notably, these follow-ups are related to health care utilization, whether actualized or otherwise.

We conducted a pilot study in which 3 patients were studied by the telecarers. Thereafter, the remaining 8 patients were divided equally between the 2 telecarers. We initially designed the study for telecarers to categorize concerns under lifestyle management and medication management. However, during the pilot, the telecarers requested to express additional concerns. Hence, we allowed them to indicate additional concerns as free text remarks. We collated and anonymized the telecarers' responses. To understand areas where they indicated additional concerns, we analyzed their remarks and other concerns. These were viewed holistically. The resulting categories and assignments of concerns to these categories were based on consensus amongst 2 independent readers (ME and PK) after several rounds of categorization and discussion. We only included concerns resulting clearly from or related to potential issues explicitly discussed in the call. Remarks were only considered when they gave context to telecarers' concerns and were aligned with the conversation content. We perused the transcripts to ensure no confounds regarding lifestyle management content.

XSL•FO
RenderX

Association of Lifestyle-Focused Calls With Admissions

To assess the degree of associations between lifestyle-focused calls and the frequency and length of inpatient admissions, we conducted correlation tests. We computed correlations between the number of lifestyle-focused calls per patient and the number and length (average LOS) of all-cause inpatient admissions. We applied the nonparametric Spearman rank correlation test, as the monotonicity assumption was fulfilled, but normality assumptions could not be guaranteed due to limited dataset size. We corrected for multiple comparisons using Holm-Bonferroni correction. We set significance to $P<.05$. We descriptively compared the number and length of all-cause inpatient admissions between groups of patients with at least 1 lifestyle-focused call and patients with no lifestyle-focused calls. We measured effect size using the Cohen *d* estimate. We used R software, including the *stats* (version 4.3.3) and *effsize* packages (version 0.81).

Ethical Considerations

This study is set within a Heart Failure Telesupport Program run at the Changi General Hospital (CGH), a 1000-bed acute care tertiary hospital in Singapore. Our retrospective study entailed a secondary analysis of data collected in an earlier nonrandomized controlled study on a 150-patient cohort from January 2014 to October 2017 [14]. The 150-patient cohort in the original study was recruited by trained research coordinators who approached patients eligible for enrolment and obtained written informed consent for their participation in the Heart Failure Telesupport Program and the use of their data for research purposes [14]. As the original nonrandomized controlled study [14] was conducted as part of the follow-up services offered to patients, it was approved by the SingHealth Centralised Institutional Review Board (CIRB) with waiver of ethics review. Our retrospective study was approved by the SingHealth CIRB (Protocol 2018/2761) with waiver of full informed consent. All data used in our retrospective study were deidentified before it was analyzed by the study team.

Results

Data Preparation

Overview of Dataset

Calls in our dataset involved 50 patients with HF and 8 trained telecarers from the Health Management Unit (HMU) at CGH. The 50 patients had these demographics: 13 (26%) female and 37 (74%) male participants; mean age of 59 (SD 14.3) years; and 29 (58%) Chinese, 8 (16%) Malay, 7 (15%) Indian, and 6 (12%) other participants. The 8 telecarers had these demographics: 1 (12) male and 7 (88%) female participants; mean age of 50.6 (SD 11.1) years; 1 (12) Malay and 7 (88%) Chinese participants; and 2-14 (mean 7.4, SD 4.3) years of experience at CGH [14].

The duration of calls ranged from 33 seconds to 64 minutes 15 seconds (median 7 minutes 30 seconds). Utterances per transcript ranged from 11 to 1289 (median 178). Typically, calls had no or very little lifestyle management content (3584/160,251, 2.24%; Multimedia Appendix 2). The interrater reliability test for the dialogue act annotations resulted in Fleiss $\kappa=0.70$. Examples of dialogue acts annotated in the dataset are in Multimedia Appendix 3. Distributions of the frequency and length of inpatient admissions are in Multimedia Appendix 4.

Cohort Characteristics

The 3 patient cohorts generally have comparable demographics, with patients being predominantly male, of Chinese ethnicity, and aged between 50 and 70 years (Table 2). Patients had, on average, 15 to 16 calls, with very few (2 to 3) lifestyle-focused calls. The expert evaluation cohort, however, due to its small size and having been purposively sampled to facilitate a qualitative evaluation, had very few (3/11, 27%) female patients and slightly more calls (mean 16.3, SD 4.7) and lifestyle-focused calls (mean 3.4, SD 1.3). On average, patients had 3 to 4 all-cause inpatient admissions, with an average LOS of about 6 days; this was comparable across cohorts. The cohorts were also comparable regarding the Charlson Comorbidity Index (CCI), the New York Heart Association (NYHA) classes, and medication classes (Multimedia Appendix 5).

XSL•FO
RenderX

Table 2. Characteristics of the full cohort (50 patients with HF[a]), HFDM[b] cohort (25 patients with both HF and T2DM[c]), and expert evaluation cohort (11 patients with both HF and T2DM, a subset of the HFDM cohort).

	Full cohort (N=50)[d]	HFDM cohort (n=25)[d]	Expert evaluation cohort (n=11)[d]
Age in years, mean (SD)	59 (14.3)	58.3 (12.8)	61.1 (12)
Gender, n (%)			
Male	37 (74)	21 (84)	8 (73)
Female	13 (26)	4 (16)	3 (27)
Ethnicity, n (%)			
Chinese	29 (58)	13 (52)	5 (45)
Malay	8 (16)	5 (20)	2 (18)
Indian	7 (14)	3 (12)	3 (27)
Others	6 (12)	4 (16)	1 (9)
Number of inpatient admissions, mean (SD)			
All-cause	3.4 (2.8)	3.8 (2.4)	3.3 (2.6)
CVD[e]-related	2.0 (2.1)	1.8 (1.1)	1.4 (0.5)
HF-related	1.8 (1.3)	1.8 (1.4)	1.4 (0.8)
Average LOS [f], mean (SD)			
All-cause	6.5 (2.4)	6.1 (2.6)	5.7 (2.6)
CVD-related	6.5 (2.3)	6.0 (1.9)	5.9 (1.9)
HF-related	6.6 (2.6)	5.9 (2.4)	5.7 (2.8)
Calls, mean (SD)			
All calls	14.6 (7.1)	15.2 (7)	16.3 (4.7)
Lifestyle-focused calls	2.0 (1.9)	2.5 (1.9)	3.4 (1.3)
Calls with lifestyle management content (not focused)	2.8 (3.6)	3.7 (4.3)	3.5 (2.7)
Calls without lifestyle management content	9.8 (8.1)	9.0 (7.9)	9.4 (6.3)

[a]HF: heart failure.
[b]HFDM: both heart failure and type 2 diabetes mellitus.
[c]T2DM: type 2 diabetes mellitus.
[d]For some indicators, data may not be available for all patients.
[e]CVD: cardiovascular disease.
[f]LOS: length of stay.

Data Analysis

Analysis of Content and Structure of Lifestyle-Focused Calls

To give an impression of the content and conversation structures in lifestyle-focused calls, we present 2 excerpts of conversation with dialogue act annotations. In the first example (Multimedia Appendix 6), the patient is feeling unwell, and the telecarer checks on his general condition and discusses different aspects of HF lifestyle management. Request-inform and inform are used to check for compliance with fluid restriction and to educate the patient. Request-action is used to reinforce correct adherence behavior, although this takes up a small part of the conversation. The second example (Multimedia Appendix 7) demonstrates how socio-emotional dialogue acts are used for relationship-building and expressing emotions. Evaluate-negative indicates where the telecarer highlights an issue with the patient's condition, while inform provides background regarding the patient's condition and approach. This call shows the difficulty in establishing the patient's actual actions relating to lifestyle management and highlights the shared burden of care between the caregiver and telecarer in helping to manage the patient's condition.

We now present the results of the analyses. First, for the full cohort, we find that lifestyle-focused calls had significantly fewer utterances (mean 203.1, SD 147.3) than calls with lifestyle management content (not focused) (mean 310.5, SD 217.1; P=.01). Lifestyle-focused calls are therefore shorter than calls with lifestyle content (not focused) in terms of number of utterances. Furthermore, the proportion of lifestyle-focused calls with a mention of symptom attributes is comparable with that

XSL•FO
RenderX

of calls with lifestyle management content (not focused) (Multimedia Appendix 8).

Next, we see significantly more dialogue acts on average in lifestyle-focused calls (mean 17.5, SD 12.4) than in calls with lifestyle management content (not focused; mean 10.1, SD 7.7; P_{adj}=.005; Table 3). We observe similar trends across categories of dialogue acts: exchanging information (mean 10.8, SD 7.8 versus mean 7.0, SD 5.2; P_{adj}=.01), understanding information (mean 3.0, SD 2.8 versus mean 2.0, SD 2.0; P_{adj}=.04), and incomplete dialogue acts (mean 1.4, SD 2.4 versus mean 0.4, SD 0.8; P_{adj}=.01). As dialogue act annotations were only available for the 25-patient HFDM cohort, we focused these analyses on them.

Table 3. Linear regression results comparing dialogue acts, grouped by category, for lifestyle-focused calls (n=49) and calls with lifestyle management content (not focused; n=80). Multiple dialogue acts may be annotated within a single utterance. SEs clustered by patient, with P<.05, and Holm-Bonferroni correction for multiple comparisons.

	Lifestyle-focused calls (n=49)			Calls with lifestyle management content (not focused) (n=80)			Linear regression with clustered SEs[a]			
	Dialogue acts, n	Mean (SD)	Median (IQR)	Dialogue acts, n	Mean (SD)	Median (IQR)	Estimate	95% CI	P value	P_{adj} value[b]
Overall dialogue acts	857	17.5 (12.4)	15.0 (9.0-24.0)	807	10.1 (7.7)	8.0 (5.0-13.0)	7.4	3.8 to 11.0	<.001	.005
Exchanging information	531	10.8 (7.8)	10.0 (5.0-16.0)	557	7.0 (5.2)	6.0 (3.0-8.0)	3.9	1.7 to 6.0	<.001	.01
Understanding information	146	3.0 (2.8)	2.0 (1.0-4.0)	154	2.0 (2.0)	1.0 (0.0-3.0)	1.1	0.3 to 1.8	<.001	.04
Performing action	39	0.8 (1.4)	0.0 (1.0-2.0)	15	0.2 (0.7)	0.0 (0.0-0.0)	0.6	0.1 to 1.2	.03	.12
Evaluation of health condition	21	0.4 (1.0)	0.0 (0.0-0.0)	16	0.2 (0.7)	0.0 (0.0-0.0)	0.2	–0.2 to 0.6	.24	.40
Social-emotional	29	0.6 (1.7)	0.0 (0.0-0.0)	15	0.2 (0.6)	0.0 (0.0-0.0)	0.4	–0.2 to 1.0	.17	.40
Incomplete dialogue act	68	1.4 (2.4)	1.0 (0.0-2.0)	30	0.4 (0.8)	0.0 (0.0-0.3)	1.0	0.4 to 1.6	<.001	.01
Others	25	0.5 (0.8)	0.0 (0.0-1.0)	21	0.3 (0.9)	0.0 (0.0-0.0)	0.2	–0.1 to 0.6	.13	.40

[a]Lifestyle-focused call (versus call with lifestyle management content [not focused]).

[b]Holm-Bonferroni correction.

Overall, while lifestyle-focused calls had fewer utterances than other calls with lifestyle management content (not focused), the discussions in lifestyle-focused calls were more intense with more dialogue acts. In particular, the higher numbers of exchanging and understanding information dialogue acts suggest that lifestyle-focused calls have more probing and coaching exchanges. Conversely, discussions on symptoms were comparable across both types of calls, suggesting that clinical topics did not have greater attention in lifestyle-focused calls.

Expert Evaluation

The telecarers did not flag many calls as concerning (Table 4). Only 24.2% (29/120) of the calls were found concerning enough to follow up on. Of the 30 lifestyle-focused calls among the 120 calls, 30% (9/30) were flagged with at least 1 concern. We found that the telecarers had concerns in 8 areas: 1-HF symptoms (eg, swelling, bloated stomach, cough, chest pain, and shortness of breath); 2-Influenza symptoms (eg, fever, flu, viral infection, and runny nose); 3-Gastrointestinal symptoms (eg, stomach discomfort, abdominal pain, constipation, and diarrhea); 4-Vitals (eg, weight gain, blood pressure, and weight loss); 5-Comorbidities (eg, cancer and diabetes); 6-Lifestyle management (eg, fluid and salt intake); 7-Medication management (eg, medication adjustment, self-administration, and adherence); and 8-Others (eg, dental issues leading to weight loss).

In terms of areas of concern, telecarers were more attuned to concerns related to 1-HF symptoms (12/120, 10%) and 4-Vitals (12/120, 10%), with 5 calls related to both, rather than to concerns related to 6-Lifestyle management (4/120, 3.3%). Even within lifestyle-focused calls, only 2 of the calls were flagged with concerns related to 6-Lifestyle management.

XSL•FO
RenderX

Table 4. Areas of concern identified by the telecarers, their level of concern, and the proportion of calls with concerns across all calls (n=120) and across lifestyle-focused calls (n=30).

	Areas of concern identified by telecarers:								Calls[a] with at least 1 concern	Calls[a] with no concern
	1-HF[b] symptoms	2-Influenza symptoms	3-Gastrointestinal symptoms	4-Vitals	5-Comorbidities	6-Lifestyle management	7-Medication management	8-Others		
All calls										
Business as usual, n	1	0	0	2	0	0	1	1	—[c]	—
Nurse follow-up: case review, n	6	4	2	4	4	4	2	0	—	—
Nurse follow-up: call patient, n	1	0	0	3	3	0	0	0	—	—
Advise patient to go to GP[d] or polyclinic, n	4	2	1	3	0	0	0	0	—	—
Total number of calls[a], n (%)	12 (10.0)	6 (5.0)	3 (2.5)	12 (10.0)	7 (5.8)	4 (3.3)	3 (2.5)	1 (0.8)	29 (24.2)	91 (75.8)
Lifestyle-focused calls										
Business as usual, n	0	0	0	0	0	0	1	0		
Nurse follow-up: case review, n	1	1	1	1	1	2	0	0		
Nurse follow-up: call patient, n	1	0	0	1	0	0	0	0		
Advise patient to go to GP or polyclinic, n	3	0	0	1	0	0	0	0		
Total number of calls[a], n (%)	5 (16.7)	1 (3.3)	1 (3.3)	3 (10.0)	1 (3.3)	2 (6.7)	1 (3.3)	0	9 (30)	21 (70)

[a]Calls may have overlapping concerns.

[b]HF: heart failure.

[c]Not applicable.

[d]GP: general practitioner.

The 4 calls flagged with concerns related to 6-Lifestyle management (including 2 lifestyle-focused calls) were all accompanied by concerns relating to 4-Vitals; 1 of the lifestyle-focused calls additionally had a concern relating to 1-HF symptoms. The telecarers assigned the level of concern "Nurse follow-up: case review" to all 4 of these calls.

In summary, the expert evaluation shows that telecarers do not often flag calls as concerning enough to follow up on beyond the call. Concerns, when flagged, primarily focus on HF symptoms and vitals (19/29, 65.5%). The few calls flagged with lifestyle management concerns (4/29, 13.8%) were consistently accompanied by concerns relating to other clinical issues and followed up via case review. As such, telecarers do not typically find lifestyle-focused calls any more concerning than other calls (not focused on lifestyle management content).

Association of Lifestyle-Focused Calls with Admissions

We present results of the association between the number of lifestyle-focused calls and (1) average LOS for all-cause inpatient admissions and (2) number of all-cause inpatient admissions in Figure 2. Each data point corresponds to 1 patient. Black dots represent the average LOS and number of admissions per patient, respectively. Trends in average LOS and number of admissions are represented with smoothed lines fitting locally weighted regression curves (LOWESS; N=50). The number of lifestyle-focused calls has a modest negative correlation with average LOS for all-cause inpatient admissions (ρ=–0.30; P=.03; P_{adj}=.06) (Figure 2A), but a negligible negative correlation with the number of all-cause inpatient admissions (ρ=-0.18; P=.84; P_{adj}=.84; Figure 2B). Both findings are not statistically significant after accounting for multiple comparisons.

XSL•FO
RenderX

Figure 2. Correlation between the number of lifestyle-focused calls and (A) average length of stay (LOS) for all-cause inpatient admissions and (B) number of all-cause inpatient admissions.

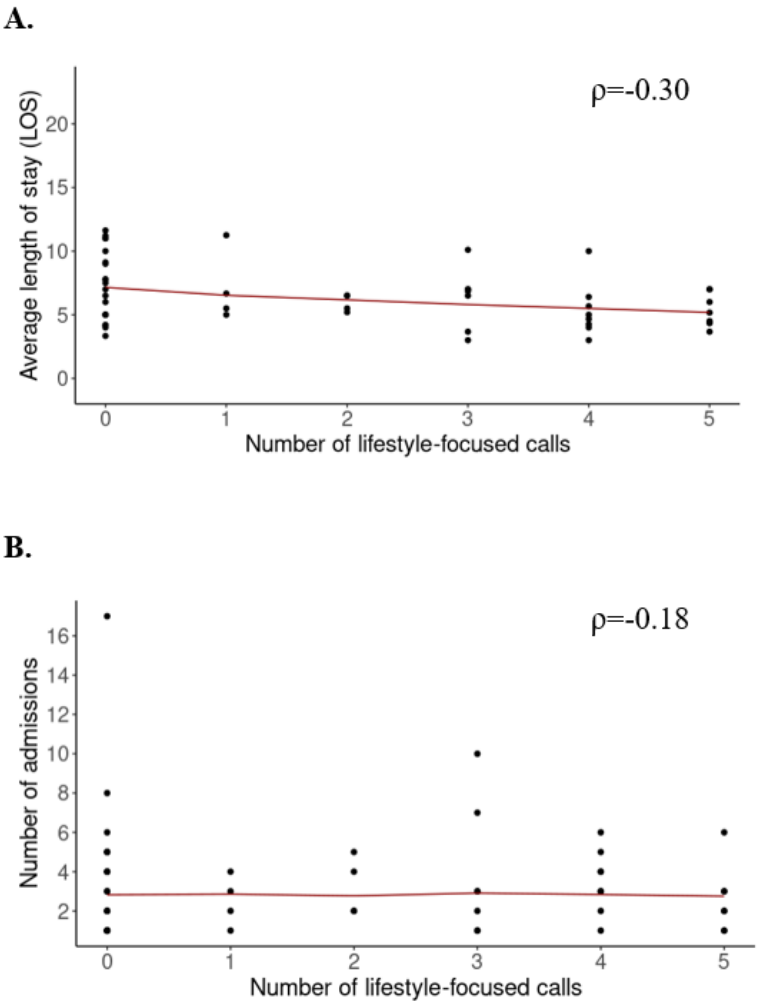

We further compared the frequency and length of admissions for patients with no lifestyle-focused calls (20/50, 40%) versus for patients with at least 1 lifestyle-focused call (30/50, 60%; Multimedia Appendix 9). Patients with at least 1 lifestyle-focused call have lower average LOS on average (mean 3.2, SD 2.1) compared to patients with no lifestyle-focused calls (mean 4.0, SD 1.4), with a medium effect size (Cohen *d*=–0.62). Similarly, patients with at least 1 lifestyle-focused call have slightly fewer average admissions (mean 3.2, SD 2.1) compared to patients with no lifestyle-focused calls (mean 4.0, SD 1.4), with a small effect size (Cohen *d*=–0.21).

We considered whether the association of lifestyle-focused calls to average LOS could be attributed to a higher number of symptoms mentioned in the calls (Multimedia Appendix 8). Lifestyle-focused calls have fewer to similar proportions of symptom attributes as compared to calls with lifestyle management content (not focused).

Overall, these findings suggest that having more calls focused on discussing lifestyle management may be related to shorter (but not fewer) inpatient stays.

Discussion

Principal Findings

Nurse-led telehealth interventions improve the self-care behaviors of patients with chronic cardiometabolic conditions, but the content, structure, and clinical associations of lifestyle-focused telecarer-patient conversations are not well understood. We leveraged a real-world multimodal dataset with a unique mixed methods approach to elucidate novel insights on telecarer-patient conversations focused on lifestyle management. First, we found that compared to typical calls with lifestyle management content, calls with a greater focus on lifestyle management were shorter and more intense but did not emphasize discussion of symptoms. Furthermore, we found that experienced telecarers are more attuned to symptoms and vitals as causes for concern than to lifestyle management. Finally, we found that a greater focus on lifestyle management in telecare conversations is associated with modest but nonsignificant drops in the extent of inpatient care required.

Implications

Our quantitative analyses showed that lifestyle-focused calls entail conversational structures reflective of coaching activities

XSL•FO
RenderX

informed by an understanding of patient experiences on behavioral or lifestyle changes [39]. Such coaching activities are beneficial to patient participation and lifestyle management for chronic disease [12,17,20]. Notably, while previous studies have studied communication in telehealth programs as a whole [26,39,40], we zoomed in on conversations where telecarers prioritized lifestyle management (perhaps in response to changes in patient experience or behavior) and conducted rigorous linguistic analyses (spanning utterances, dialogue acts, and attributes) on the underlying conversational structures. Hence, our work provides targeted insights into content and structures of conversations promoting lifestyle management in telehealth settings and could inform enhancements to telecare programs.

That said, our qualitative evaluations revealed the inherent challenges of addressing lifestyle management issues in telehealth settings. From the telecarers' perspective, lifestyle management issues were rarely concerning enough to follow up on, partly because they assessed that any lifestyle management issues that arose during calls were (to the extent possible) addressed and reinforced within the calls themselves. Indeed, with time constraints and increasing numbers of patients, telecarers may often grapple with the contention of addressing acute needs arising from symptom checking and vitals monitoring versus investing in long-term gains from lifestyle management. Hence, supplementing calls with integrated care processes or digital solutions could enable more resource-efficient means to detect and follow up on lifestyle management issues.

Our association analyses showed that a greater focus on intensive lifestyle management support in telecarer-patient conversations had a discernable (but nonsignificant) association with the average length of inpatient stays but not with the number of admissions. While it is known that nurse-led telecoaching improves patients' HF-related self-care behaviors [10,41], our study is notable for suggesting possible links between lifestyle-focused telecare conversations and health care utilization. Intuitively, one might expect that patients with more lifestyle-focused calls would be more symptomatic and, hence, have longer hospital stays. However, our results point to a different possibility: that having more calls focusing on lifestyle management issues may be linked to the increased likelihood of such issues being promptly addressed; that is, these calls could correspond to timely interventions that reinforce positive behaviors and mitigate future progression to severe admissions. The potential effectiveness of these lifestyle-focused interventions possibly also explains the lack of associations with the frequency of inpatient admissions [14,17,18]. That said, the associations are modest, possibly as the length of inpatient stay is influenced by a wide range of complex factors [27,42,43], and telehealth conversation contents are only indirect indicators of clinical state [44].

Collectively, our findings motivate the design of processes and solutions to enable telecarers to effectively address longer-term lifestyle management issues alongside acute clinical priorities for HF and other cardiometabolic conditions. For example, digital tools could regularly bring telecarers' attention to lifestyle management issues, thereby enabling timely and focused coaching and reinforcement in calls. Furthermore, the inclusion of systematic follow-up actions that telecarers could suggest to patients experiencing lifestyle management issues beyond typical follow-ups for symptoms and clinical issues may be beneficial.

Strengths and Limitations

A notable strength of our study is the multimodal dataset linking unstructured telehealth conversations with structured electronic health records. Another strength is the analysis approach leveraging behavioral science–driven strategies [36,37] to assess changes in conversational behavior over time, as this allowed us to pinpoint calls where telecarers accorded heightened priority to lifestyle management. Further strengths include the rigorous linguistic analyses of these calls using fine-grained dialogue acts, the expert evaluation, and the associations of calls with admissions, as these together provide a holistic picture of interfacing behavioral, conversational, and clinical considerations for new insights.

Resource constraints for data curation led to some limitations. Our dataset was limited to English conversations for only 50 patients, and potential patient selection bias should be noted. Singapore is a multi-ethnic society in Southeast Asia, with many residents being conversant in at least 2 languages (English, Mandarin, Malay, or Tamil). However, fluency in English could possibly correlate to higher education and socioeconomic status. Furthermore, for efficient annotation, only a single topic was annotated per utterance, but this was mitigated by assigning the predominant topic to each utterance. We also could not perform qualitative coding and analysis of the call transcripts, but future efforts in this direction could provide more nuanced findings. Furthermore, the paucity and irregularity of admissions data across patients introduced challenges for association analyses. While we mitigated these by adopting a patient-specific timeframe, limitations in dataset size curtailed granular analyses (eg, of HF-related admissions) and statistical power. Future efforts with larger cohorts could be valuable.

Conclusion

This study provides new perspectives on the content and structure of telecare conversations and highlights that a focus on lifestyle management could play an important role in improving the care of patients with HF. As such, our findings offer the potential to inform ways to enhance fast-proliferating telehealth programs for improved self-care in chronic conditions.

Acknowledgments

This research is supported by funding and infrastructure from A*STAR, Singapore (grants IAF H19/01/a0/023 on Diabetes Clinic of the Future, SSF A1818g0044 on Digital Health, and SSF A1718g0045 on Deep Learning 2.0). This work was conducted using resources and infrastructure provided by the Aural Intelligence Department at I^2R. Telehealth data acquisition was funded by the

XSL•FO
RenderX

Economic Development Board (EDB), Singapore Living Lab Fund, and Philips Electronics Hospital to Home Pilot Project (EDB grant S14-1035-RF-LLF H and W). We acknowledge Royal Philips, Singapore, for support in the design and setup of the telesupport program.

The authors would like to acknowledge valuable inputs from I^2R colleagues: Ai Ti for infrastructure support, Savitha Ramaswamy for data cross-linking, Priscilla Ong for testing the expert study, and Nurensah Beevi for compliance and governance support. We acknowledge valuable inputs from Michael MacDonald, Ying Zi Oh, and Zhenwei Teo (CGH Department of Cardiology), data extraction assistance from Shao Chuen Tong (CGH Health Services Research department) and Chui Yan Ang (CGH Health Management Unit), inputs on conceptualization of the overall study from Wai Leng Chow (CGH Health Services Research department), statistical inputs from Koh Xuan Han (CGH Health Services Research department) and support from Yian Chin Lee (CGH Care Coordination).

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to the need to protect patient privacy but may be made available on reasonable request to Dr. Oh Hong Choon (oh.hong.choon@singhealth.com.sg), under restrictions subject to obtaining ethics approval from the institutional review board(s) and an appropriate data-use and/or research agreement with Changi General Hospital.

Authors' Contributions

ME, SBY, LC, SL, HCO, MKW, NFC, and PK conceptualized the study and designed the methodology. ME, SBY, LC, ZL, LY, and YG developed software for data curation and analysis. ME, SBY, LC, SUMS, HL, HBS, GCL, AN, SO, HCO, MKW, and PK conducted formal analysis. ME, SBY, LC, HL, NFAS, SUMS, ZL, LY, YG, AN, SO, BPC, SL, HW, HCO, MKW, NFC, and PK contributed to data acquisition and curation. AN, SL, HCO, MKW, NFC, and PK supervised the project. SL, HCO, NFC, and PK contributed to project administration and funding acquisition. ME, SBY, LC, SUMS, MKW, and PK contributed to writing the original draft. ME, SBY, SUMS, ZL, AN, SO, BPC, SL, HW, HCO, MKW, NFC, and PK contributed to reviewing and editing the manuscript. All authors read the final manuscript and approved of its submission. HCO and PK are the corresponding authors of this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of topics, symptoms, attributes, speakers, and dialogue acts annotated in the dataset. Symptoms, lifestyle attributes, and vitals attributes may fall under any of the topics listed. Symptom attributes correspond to the respective heart failure symptoms.
[DOCX File , 28 KB - jmir_v26i1e46983_app1.docx]

Multimedia Appendix 2
Proportions of utterances annotated in the dataset (n=160,251) across topics and speakers. Proportions of symptoms (n=35,114), symptom attributes (n=1036), lifestyle attributes (n=4992), and vitals attributes (n=45,474) annotated in the dataset. Symptom attributes (location, frequency, extent, time, and activity) correspond to the respective heart failure symptoms.
[DOCX File , 27 KB - jmir_v26i1e46983_app2.docx]

Multimedia Appendix 3
Examples of dialogue acts annotated in the dataset.
[DOCX File , 26 KB - jmir_v26i1e46983_app3.docx]

Multimedia Appendix 4
Distributions (histograms) of (A) the number of inpatient admissions and (B) average length of stay (LOS) per patient, for all-cause inpatient admissions (50 patients). Vertical dotted lines indicate mean values for number of inpatient admissions and average LOS.
[PNG File , 96 KB - jmir_v26i1e46983_app4.png]

Multimedia Appendix 5
Characteristics of the full cohort (50 patients with HF), HFDM cohort (25 patients with both HF and T2DM), and expert evaluation cohort (11 patients with both HF and T2DM, a subset of the HFDM cohort).
[DOCX File , 24 KB - jmir_v26i1e46983_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Excerpt of conversation on lifestyle management between patient P05 and nurse telecarer N02.
[DOCX File , 26 KB - jmir_v26i1e46983_app6.docx]

Multimedia Appendix 7
Excerpt of conversation on lifestyle management between patient P03's spouse and nurse telecarer N06.
[DOCX File , 26 KB - jmir_v26i1e46983_app7.docx]

Multimedia Appendix 8
Comparison of proportions of lifestyle-focused calls (n=100) and calls with lifestyle management content (not focused; n=138) that include mentions of symptom attributes.
[DOCX File , 24 KB - jmir_v26i1e46983_app8.docx]

Multimedia Appendix 9
Comparative analysis of the average length of stay (LOS) for all-cause inpatient admissions and number of all-cause inpatient admissions between patients with at least 1 lifestyle-focused call (n=30) and patients with no lifestyle-focused calls (n=20).
[DOCX File , 24 KB - jmir_v26i1e46983_app9.docx]

References

1. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2022;79(17):e263-e421. [doi: 10.1016/j.jacc.2021.12.012] [Medline: 35379503]
2. Carter J, Donelan K, Thorndike AN. Patient perspectives on home-based care and remote monitoring in heart failure: a qualitative study. J Prim Care Community Health 2022;13:21501319221133672 [FREE Full text] [doi: 10.1177/21501319221133672] [Medline: 36305386]
3. Sokalski T, Hayden KA, Raffin Bouchal S, Singh P, King-Shier K. Motivational interviewing and self-care practices in adult patients with heart failure: a systematic review and narrative synthesis. J Cardiovasc Nurs 2020;35(2):107-115. [doi: 10.1097/JCN.0000000000000627] [Medline: 31851149]
4. Santesmases-Masana R, González-de Paz L, Hernández-Martínez-Esparza E, Kostov B, Navarro-Rubio MD. Self-care practices of primary health care patients diagnosed with chronic heart failure: a cross-sectional survey. Int J Environ Res Public Health 2019;16(9):1625 [FREE Full text] [doi: 10.3390/ijerph16091625] [Medline: 31075932]
5. van der Wal MHL, Jaarsma T, van Veldhuisen DJ. Non-compliance in patients with heart failure; how can we manage it? Eur J Heart Fail 2005;7(1):5-17 [FREE Full text] [doi: 10.1016/j.ejheart.2004.04.007] [Medline: 15642526]
6. Hessel FP. Overview of the socio-economic consequences of heart failure. Cardiovasc Diagn Ther 2021;11(1):254-262 [FREE Full text] [doi: 10.21037/cdt-20-291] [Medline: 33708497]
7. Keller R, Hartmann S, Teepe GW, Lohse K, Alattas A, Tudor Car L, et al. Digital behavior change interventions for the prevention and management of type 2 diabetes: systematic market analysis. J Med Internet Res 2022;24(1):e33348 [FREE Full text] [doi: 10.2196/33348] [Medline: 34994693]
8. Snoswell CL, Chelberg G, de Guzman KR, Haydon HH, Thomas EE, Caffery LJ, et al. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. J Telemed Telecare 2023;29(9):669-684. [doi: 10.1177/1357633X211022907] [Medline: 34184580]
9. de Groot J, Wu D, Flynn D, Robertson D, Grant G, Sun J. Efficacy of telemedicine on glycaemic control in patients with type 2 diabetes: a meta-analysis. World J Diabetes 2021;12(2):170-197 [FREE Full text] [doi: 10.4239/wjd.v12.i2.170] [Medline: 33594336]
10. Imanuel Tonapa S, Inayati A, Sithichoksakulchai S, Daryanti Saragih I, Efendi F, Chou F. Outcomes of nurse-led telecoaching intervention for patients with heart failure: a systematic review and meta-analysis of randomised controlled trials. J Clin Nurs 2022;31(9-10):1125-1135. [doi: 10.1111/jocn.16025] [Medline: 34535943]
11. Kebapci A, Ozkaynak M, Lareau SC. Effects of eHealth-based interventions on adherence to components of cardiac rehabilitation: a systematic review. J Cardiovasc Nurs 2020;35(1):74-85. [doi: 10.1097/JCN.0000000000000619] [Medline: 31738217]
12. Timpel P, Oswald S, Schwarz PEH, Harst L. Mapping the evidence on the effectiveness of telemedicine interventions in diabetes, dyslipidemia, and hypertension: an umbrella review of systematic reviews and meta-analyses. J Med Internet Res 2020;22(3):e16791 [FREE Full text] [doi: 10.2196/16791] [Medline: 32186516]
13. Ma Y, Zhao C, Zhao Y, Lu J, Jiang H, Cao Y, et al. Telemedicine application in patients with chronic disease: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022;22(1):105 [FREE Full text] [doi: 10.1186/s12911-022-01845-2] [Medline: 35440082]

XSL•FO
RenderX

14. Leng Chow W, Aung CYK, Tong SC, Goh GS, Lee S, MacDonald MR, et al. Effectiveness of telemonitoring-enhanced support over structured telephone support in reducing heart failure-related healthcare utilization in a multi-ethnic Asian setting. J Telemed Telecare 2020;26(6):332-340. [doi: 10.1177/1357633X18825164] [Medline: 30782070]
15. Wu C, Wu Z, Yang L, Zhu W, Zhang M, Zhu Q, et al. Evaluation of the clinical outcomes of telehealth for managing diabetes: a PRISMA-compliant meta-analysis. Medicine (Baltimore) 2018;97(43):e12962 [FREE Full text] [doi: 10.1097/MD.0000000000012962] [Medline: 30412116]
16. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
17. Lee AYL, Wong AKC, Hung TTM, Yan J, Yang S. Nurse-led telehealth intervention for rehabilitation (Telerehabilitation) among community-dwelling patients with chronic diseases: systematic review and meta-analysis. J Med Internet Res 2022;24(11):e40364 [FREE Full text] [doi: 10.2196/40364] [Medline: 36322107]
18. Wong AKC, Bayuo J, Wong FKY, Yuen WS, Lee AYL, Chang PK, et al. Effects of a nurse-led telehealth self-care promotion program on the quality of life of community-dwelling older adults: systematic review and meta-analysis. J Med Internet Res 2022;24(3):e31912 [FREE Full text] [doi: 10.2196/31912] [Medline: 35311680]
19. Angus D, Watson B, Smith A, Gallois C, Wiles J. Visualising conversation structure across time: insights into effective doctor-patient consultations. PLoS One 2012;7(6):e38014 [FREE Full text] [doi: 10.1371/journal.pone.0038014] [Medline: 22693629]
20. Street RL, Millay B. Analyzing patient participation in medical encounters. Health Commun 2001;13(1):61-73. [doi: 10.1207/S15327027HC1301_06] [Medline: 11370924]
21. Idris MY, Alema-Mensah E, Olorundare E, Mohammad M, Brown M, Ofili E, et al. Exploring the discursive emphasis on patients and coaches who participated in technology-assisted diabetes self-management education: clinical implementation study of Health360x. J Med Internet Res 2022;24(3):e23535 [FREE Full text] [doi: 10.2196/23535] [Medline: 35302506]
22. Mayfield E, Laws MB, Wilson IB, Penstein Rosé C. Automating annotation of information-giving for analysis of clinical conversation. J Am Med Inform Assoc 2014;21(e1):e122-e128 [FREE Full text] [doi: 10.1136/amiajnl-2013-001898] [Medline: 24029598]
23. Ross L, Danforth CM, Eppstein MJ, Clarfeld LA, Durieux BN, Gramling CJ, et al. Story arcs in serious illness: natural language processing features of palliative care conversations. Patient Educ Couns 2020;103(4):826-832. [doi: 10.1016/j.pec.2019.11.021] [Medline: 31831305]
24. Dahm MR, Raine SE, Slade D, Chien LJ, Kennard A, Walters G, et al. Older patients and dialysis shared decision-making. Insights from an ethnographic discourse analysis of interviews and clinical interactions. Patient Educ Couns 2024;122:108124 [FREE Full text] [doi: 10.1016/j.pec.2023.108124] [Medline: 38232671]
25. Dowell A, Stubbe M, Macdonald L, Tester R, Gray L, Vernall S, et al. A longitudinal study of interactions between health professionals and people with newly diagnosed diabetes. Ann Fam Med 2018;16(1):37-44 [FREE Full text] [doi: 10.1370/afm.2144] [Medline: 29311173]
26. Seuren LM, Ilomäki S, Dalmaijer E, Shaw SE, Stommel WJP. Communication in telehealth: a state-of-the-art literature review of conversation-analytic research. Res Lang Soc Interact 2024;57(1):73-90 [FREE Full text] [doi: 10.1080/08351813.2024.2305045] [Medline: 38741749]
27. Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2014;9(4):e94207 [FREE Full text] [doi: 10.1371/journal.pone.0094207] [Medline: 24718585]
28. Zolnierek KBH, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009;47(8):826-834 [FREE Full text] [doi: 10.1097/MLR.0b013e31819a5acc] [Medline: 19584762]
29. Griffin SJ, Kinmonth A, Veltman MWM, Gillard S, Grant J, Stewart M. Effect on health-related outcomes of interventions to alter the interaction between patients and practitioners: a systematic review of trials. Ann Fam Med 2004;2(6):595-608 [FREE Full text] [doi: 10.1370/afm.142] [Medline: 15576546]
30. Greenfield S, Kaplan SH, Ware JE, Yano EM, Frank HJL. Patients' participation in medical care: effects on blood sugar control and quality of life in diabetes. J Gen Intern Med 1988;3(5):448-457. [doi: 10.1007/BF02595921] [Medline: 3049968]
31. Barras C, Geoffrois E, Wu Z, Liberman M. Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication 2001;33(1-2):5-22. [doi: 10.1016/s0167-6393(00)00067-4]
32. Asher N, Hunter J, Morey M, Benamara F, Afantenos S. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. 2016 Presented at: Proceedings of the 10th International Conference on Language Resources and Evaluation; May 23-28, 2016; Portoroz, Slovenia p. 23-28 URL: https://aclanthology.org/L16-1432
33. Liu Z, Chen NF. Improving multi-party dialogue discourse parsing via domain integration. 2021 Presented at: Proceedings of the 2nd Workshop on Computational Approaches to Discourse; November 10-11, 2021; Punta Cana, Dominican Republic p. 122-127. [doi: 10.18653/v1/2021.codi-main.11]
34. Liu Z, Ng A, Lee S, Aw A, Chen N. Topic-aware pointer-generator networks for summarizing spoken conversations. 2019 Presented at: Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); December 14-18, 2019; Singapore, Singapore p. 814-821. [doi: 10.1109/asru46091.2019.9003764]

XSL•FO
RenderX

35. Asher N, Lascarides A. Logics of Conversation. Cambridge, United Kingdom: Cambridge University Press; 2003.
36. Johnston JM, Pennypacker HS, Green G. Strategies and Tactics of Behavioral Research and Practice. Oxfordshire, England: Routledge; 2020.
37. Johnston JM, Hodge CW. Describing behavior with ratios of count and time. Behav Anal 1989;12(2):177-185 [FREE Full text] [doi: 10.1007/BF03392494] [Medline: 22478031]
38. Brady SR. Utilizing and adapting the delphi method for use in qualitative research. Int J Qual Methods 2015;14(5):160940691562138. [doi: 10.1177/1609406915621381]
39. Gambling T, Long A. Tailoring advice and optimizing response: a case study of a telephone-based support for patients with type 2 diabetes. Fam Pract 2010;27(2):179-185. [doi: 10.1093/fampra/cmp097] [Medline: 20032169]
40. Mattisson M, Börjeson S, Årestedt K, Lindberg M. Interaction between telenurses and callers - a deductive analysis of content and timing in telephone nursing calls. Patient Educ Couns 2024;123:108178 [FREE Full text] [doi: 10.1016/j.pec.2024.108178] [Medline: 38387390]
41. Nick JM, Roberts LR, Petersen AB. Effectiveness of telemonitoring on self-care behaviors among community-dwelling adults with heart failure: a quantitative systematic review. JBI Evid Synth 2021;19(10):2659-2694 [FREE Full text] [doi: 10.11124/JBIES-20-00329] [Medline: 33896907]
42. Cleland JG, Louis AA, Rigby AS, Janssens U, Balk AH, TEN-HMS Investigators. Noninvasive home telemonitoring for patients with heart failure at high risk of recurrent admission and death: the Trans-European Network-Home-Care Management System (TEN-HMS) study. J Am Coll Cardiol 2005 May 17;45(10):1654-1664 [FREE Full text] [doi: 10.1016/j.jacc.2005.01.050] [Medline: 15893183]
43. Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N. A systematic review of the prediction of hospital length of stay: towards a unified framework. PLOS Digit Health 2022;1(4):e0000017 [FREE Full text] [doi: 10.1371/journal.pdig.0000017] [Medline: 36812502]
44. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]

Abbreviations

CCI: Charlson Comorbidity Index
CGH: Changi General Hospital
CIRB: Centralised Institutional Review Board
GP: general practitioner
HF: heart failure
HFDM: heart failure and type 2 diabetes mellitus
HMU: Health Management Unit
LOS: length of stay
NYHA: New York Heart Association
T2DM: type 2 diabetes mellitus

Edited by A Mavragani; submitted 12.01.24; peer-reviewed by B Tyl; comments to author 09.02.24; revised version received 10.06.24; accepted 20.08.24; published 30.10.24.

Please cite as:
Erdt M, Yusof SB, Chai L, Md Salleh SU, Liu Z, Sarim HB, Lim GC, Lim H, Suhaimi NFA, Yulong L, Guo Y, Ng A, Ong S, Choo BP, Lee S, Weiliang H, Oh HC, Wolters MK, Chen NF, Krishnaswamy P
Characterization of Telecare Conversations on Lifestyle Management and Their Relation to Health Care Utilization for Patients with Heart Failure: Mixed Methods Study
J Med Internet Res 2024;26:e46983
URL: https://www.jmir.org/2024/1/e46983
doi:10.2196/46983
PMID:

©Mojisola Erdt, Sakinah Binte Yusof, Liquan Chai, Siti Umairah Md Salleh, Zhengyuan Liu, Halimah Binte Sarim, Geok Choo Lim, Hazel Lim, Nur Farah Ain Suhaimi, Lin Yulong, Yang Guo, Angela Ng, Sharon Ong, Bryan Peide Choo, Sheldon Lee, Huang Weiliang, Hong Choon Oh, Maria Klara Wolters, Nancy F Chen, Pavitra Krishnaswamy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey

Yuka Sugawara[1], MD, PhD; Yosuke Hirakawa[1], MD, PhD; Masao Iwagami[2], MSc, MD, PhD; Haruo Kuroki[3], MD, PhD; Shuhei Mitani[1], MD; Ataru Inagaki[4], MD, PhD; Hiroki Ohashi[5], MD; Mitsuru Kubota[6], MD, PhD; Soichi Koike[7], MD, PhD; Rie Wakimizu[8], PhD; Masaomi Nangaku[1], MD, PhD

[1]Division of Nephrology and Endocrinology, The University of Tokyo, Tokyo, Japan
[2]Department of Health Services Research, Institute of Medicine, University of Tsukuba, Ibaraki, Japan
[3]Clinic Pauroom, Tokyo, Japan
[4]Department of Education, College of Education, Psychology and Human Studies, Aoyama Gakuin University, Tokyo, Japan
[5]TAMA Family Clinic, Tokyo, Japan
[6]Department of General Pediatrics and Interdisciplinary Medicine, National Center for Child Health and Development, Tokyo, Japan
[7]Division of Health Policy and Management, Center for Community Medicine, Jichi Medical University, Tochigi, Japan
[8]Department of Child Health and Development Nursing, Institute of Medicine, University of Tsukuba, Ibaraki, Japan

Corresponding Author:
Masaomi Nangaku, MD, PhD
Division of Nephrology and Endocrinology
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 3 3815 5411
Email: mnangaku@m.u-tokyo.ac.jp

Related Article:

This is a corrected version. See correction statement: http://mhealth.jmir.org/2024/1/e68439/

Abstract

Background: Telemedicine, or online medical care, has gained considerable attention worldwide. However, it has not been widely adopted in Japan, and the detailed status of received and provided online medical care and the reasons for its lack of popularity remain unknown.

Objective: This study aims to investigate the current status of online medical care in Japan and the factors limiting its adoption from the perspective of both patients receiving and medical professionals providing online medical care.

Methods: In total, 2 nationwide questionnaire surveys were conducted. The first survey, targeting both patients and healthy individuals, screened approximately 40,000 participants among 13 million people. The participants were selected to match the age distribution of the Japanese population based on government data, and their online medical care experience and medical visit status were recorded. To further investigate online medical care use and satisfaction, a web-based survey was conducted with 15% (6000/40,000) of the screened participants. The second survey, targeting medical professionals, was administered to a physician, a nurse, and a member of the administrative staff in each of 4900 randomly selected medical facilities to inquire about their online medical care practices and impressions. In addition, both surveys investigated the factors limiting online medical care expansion in Japan.

Results: The response rates among patients and healthy individuals targeted for the screening and main surveys were 92.5% (36,998/40,000) and 80% (1312/1478, 88.77%; 1281/1522, 84.17%; 404/478, 84.5%; and 2226/2522, 88.26% in 4 survey groups), respectively. The survey of medical professionals yielded 1552 responses (n=618, 39.82% physicians; n=428, 27.58% nurses; n=506, 32.6% administrative staff). Although the facility-level response rate was low (794/4900, 16.2%), some facility categories had relatively high response rates. Only 5.29% (1956/36,998) of the patients and healthy individuals had online medical care experience. When there were more hospitals nearby and they felt it was more work to see a physician in person, they were more

XSL•FO
RenderX

likely to use online medical care (more nearby hospitals: adjusted odds ratio [aOR] 1.33, 95% CI 1.18-1.50; more work: aOR 1.48, 95% CI 1.35-1.63 per survey response point in the patient group). Similarly, these factors were substantially associated with satisfaction (more nearby hospitals: aOR 1.40, 95% CI 1.14-1.73; more work: aOR 1.50, 95% CI 1.27-1.76 per survey response point in the patient group). In both surveys, the most frequently selected factor preventing the widespread use of online medical care was patients' need to switch to face-to-face medical care for mandatory tests and procedures. Inadequate awareness of and education on online medical care were also frequently selected.

Conclusions: Our nationwide surveys provided insights into the current status of online medical care in Japan and simultaneously identified several problems and issues related to it, which will be useful in promoting its wider adoption.

(J Med Internet Res 2024;26:e64159) doi:10.2196/64159

KEYWORDS

telemedicine; online medical care; telehealth; eHealth; mobile phone

Introduction

Background

"Telemedicine" involves the delivery of all types of medical, diagnostic, and treatment-related services using telecommunication technologies [1]. This includes various medical processes such as conducting diagnostic tests, closely monitoring patient progress after treatment or therapy, and facilitating access to specialists who may not be geographically close to the patient. Similar terms for telemedicine include "telehealth" and "eHealth," which are broader than telemedicine and often encompass digital app to support self-management. In Japan, the term "online medical care" is used, which is defined by the government more specifically than the term "telemedicine" as follows: the act of examining and diagnosing a patient, communicating the results of the diagnosis, and prescribing medical treatment in real time between a physician and a patient via information and telecommunications technology [2]. This definition emphasizes real-time communication between physicians and patients.

Telemedicine, including online medical care, has steadily evolved over time; however, the COVID-19 pandemic substantially accelerated its adoption and implementation worldwide. An analysis using Google Trends data revealed a sharp increase in public interest in telemedicine in April 2020 after the onset of the COVID-19 pandemic [3]. The number of telemedicine practices in the United States has been increasing rapidly since March 2020 with the expansion of public health insurance coverage [4]. The US Food and Drug Administration launched the Digital Health Center of Excellence in September 2020, which focuses on the development of various technologies such as mobile health devices, medical device software, and wearable technology. The United Kingdom and other European countries have been actively promoting telemedicine services through their governments, and telemedicine has become more widespread since the COVID-19 pandemic in many countries [5-8].

However, despite this global trend, Japan has not widely adopted telemedicine. Previously, telemedicine was highly restricted and not reimbursable in Japan; however, since 2018, online medical care medication expenses for specific diseases have been reimbursed. Despite deregulation and reimbursement revisions in Japan during the COVID-19 pandemic to promote online medical care [7], its implementation remains limited at many facilities and for many patients.

While online medical care is assumed to have the advantage of increasing convenience for patients, there are many unknowns, including whether it is convenient for medical institutions, whether accurate medical care can be provided, and to what extent patients are willing to use it. In clarifying these issues, it is also desirable to understand what types of facilities have offered online medical care the most, what types of patients have received online medical care, and in what types of situations patients were most satisfied with online medical care.

Objectives

To understand the current situation following the COVID-19 pandemic and identify issues related to online medical care, we conducted a nationwide questionnaire survey of the 2 largest relevant populations: patients and healthy individuals and health care professionals. The primary objective of this study was to identify the key factors that limit the adoption of online medical care.

Methods

Study Design

In total, 2 nationwide questionnaire surveys were conducted at approximately the same time to analyze the actual status of online medical care and the potential barriers to online medical care expansion in Japan from the perspective of both patients and healthy individuals who receive online medical care and medical professionals who provide it. This study was registered with University hospital Medical Information Network- Clinical Trials Registry before initiation (registration ID: UMIN000051219).

Development of Questionnaires

To develop the questionnaires, interviews were conducted at 2 sites with patients, health care professionals (physicians and nurses), and administrative staff members regarding their experiences with telemedicine, its advantages, and possible reasons why it is not currently desired or promoted. The interviews with health care professionals or administrative staff members were conducted in person or web-based, whereas patients were exclusively interviewed in person. No personal information was obtained from the participants.

XSL•FO
RenderX

The primary objective of this study was to determine the actual status of online medical care in Japan. In the patient group, we investigated whether they had experienced online medical care (a binary variable) and their level of satisfaction (on a 4-point scale) to determine the experience rate and the level of satisfaction of those who had experienced online medical care. In addition, factors related to the experience of online medical care or the satisfaction with online medical care (as outcome) were analyzed, including sex, age (surveyed as an integer and used as a binary variable for whether the patient was aged ≥60 years), number of nearby medical facilities (3 levels), and perceived effort involved in seeing a physician (4 levels; as exposure factors). In the group of health care professionals, the rate of online medical care implementation was examined overall, by type of facility, and by use of electronic medical records (EMRs).

The second objective of this study was to identify issues of online medical care. To this end, we obtained responses regarding factors that prevent the further spread of online medical care. In addition, as data to reinforce this, responses regarding the various patient burdens associated with medical visits (online medical care vs face-to-face medical care on a 5-point scale) and the ease of performing various medical procedures (online medical care vs face-to-face medical care on a 5-point scale) in the medical professional group were also surveyed.

On the basis of the summary of these interviews and previous findings, 2 questionnaires were developed and refined for patients and healthy individuals and for medical professionals. The final version of each questionnaire (in Japanese and English) is provided in Multimedia Appendices 1-4. The survey of patients and healthy individuals consisted of six parts: (1) background of respondents, (2) online medical care use and satisfaction among those with online medical care experience, (3) online medical care use prediction among those without online medical care experience, (4) various patient burdens, (5) comparison of face-to-face medical care and online medical care, and (6) factors impeding the adoption of online medical care. The survey of medical professionals consisted of seven parts: (1) background of respondent facilities and respondents, (2) status of online medical care use and its issues, (3) various patient burdens, (4) ease of performing various clinical procedures, (5) medical facilities' financial burdens and medical fees with online medical care, (6) appropriate patients for online medical care, and (7) factors hindering the adoption of online medical care. Usability and technical functionality were tested by the multiple research team members before fielding the questionnaire.

Survey of Patients and Healthy Individuals

To be eligible for the screening survey, respondents must be members of the Freeasy panel (iBRIDGE) [9], an internet-based survey panel of 13 million people. It consists of people who have registered with the site or application for the purpose of accumulating points and exchanging them for cash or e-money. In addition, to participate in the subsequent main survey, the eligibility criteria were that the respondent had to have provided sufficient information in the screening survey and have experience with online medical care or have been selected through random sampling described later if they had no experience with online medical care.

First, 40,000 individuals were selected from the Freeasy panel [9] to match the age and sex distribution of the Japanese population [10] following statistics published by the Japanese government (ie, the respondents were selected in order of earliest response [nonprobabilistic sampling], but the maximum number of respondents in each category was set based on the aforementioned distribution). They were asked to complete a screening questionnaire about their regular medical visits (currently visiting a medical facility on a regular basis; that is, more than once every 3 months) and experience with online medical care. Note that the survey panel used in this study did not include those who had difficulty accessing the internet as web-based operations were required for participation. The respondents to the screening survey were classified into the following four categories: (1) with regular visits and with online medical care experience, (2) with regular visits and without online medical care experience, (3) without regular visits and with online medical care experience, and (4) without regular visits and without online medical care experience. Second, the main survey was conducted on 3000 individuals with regular visits (ie, patients) and 3000 individuals without regular visits (ie, healthy individuals). As the number of individuals with online medical care experience was small in both groups, all the individuals were sampled. Alternatively, there were a large number of individuals who had no experience with online medical care; thus, stratified random sampling by age and sex was used to select target individuals in each group without regular visits (probabilistic sampling). All members of group 1 and part of group 2 were selected so that the total number of respondents in these groups was 3000. Similarly, all members of group 3 and part of group 4 were selected so that these groups would have a total of 3000 respondents. This approach ensured that each combined set of groups (groups 1 and 2 and groups 3 and 4) had a similar age and sex distribution. The participants were then asked to complete the main questionnaire. The survey request was sent out via the Freeasy system to the email addresses registered with Freeasy or posted on Freeasy's affiliated website, and responses were completed on the web. The response period for the screening survey was September 25, 2023, to October 8, 2023, and that for the main survey was October 19, 2023, to November 2, 2023. The Checklist for Reporting Results of Internet E-Surveys [11] (Multimedia Appendix 5) provides further details on this survey.

Survey of Medical Professionals

The eligibility criterion for this survey was that the respondent must be a physician, nurse, or administrative staff member of a medical institution covered by insurance in Japan as of June 2023.

Stratified random sampling by region and hospital category (probabilistic sampling) was used to select from 96,269 insured medical facilities in Japan as of June 2023. The regions were divided into 7 categories, facilities were divided into 2 categories (hospitals and clinics), and 350 facilities per category (350 × 2 × 7 = 4900 facilities in total) were selected. The sampling rate

XSL•FO
RenderX

was 5% in total; however, it was set higher for the following facility categories that were anticipated to be underrepresented (ie, to include only a very small number of participants) through simple random sampling: specific functioning hospitals (sampling rate of 100%), including university hospitals and hospitals for advanced medical care; clinics for medical services in remote areas; and core hospitals for medical services in remote areas (sampling rates of 50%). The study descriptions and requests for survey responses were mailed to the hospital administrator of each facility, who was instructed to distribute survey completion requests to 3 respondents at their facility (1 physician, 1 nurse, and 1 member of the administrative staff). The respondents were asked to complete a web-based form using a shortened URL or QR code. In cases of duplicate responses, only the first response was considered valid. The response period was October 30, 2023, to December 8, 2023. The Checklist for Reporting Results of Internet E-Surveys (Multimedia Appendix 5) [11] provides further details on this survey.

Data Analysis

The questionnaire responses were summarized for descriptive purposes and then compared between groups of respondents with or without online medical care experience and background factors. The chi-square and Mann-Whitney *U* tests were used for binary or categorical and ordinal variables, respectively, and the McNemar test was used when comparing two groups with correspondence. *P* values of <.05 were considered statistically significant. A multivariate logistic regression analysis was also performed of the outcome of experience or satisfaction with online medical care, with age, sex, number of nearby medical facilities (3-point Likert scale), and perceived effort of visiting the hospital (4-point Likert scale) included as covariates. All analyses were performed using Stata (version 17.0; StataCorp).

Ethical Considerations

This entire study (including the interviews, the patient and healthy individual survey, and the medical professional survey) was approved by the Research Ethics Committee of the Faculty of Medicine at the University of Tokyo (2023024NI) and was conducted in accordance with the tenets of the Declaration of Helsinki. For the interviews, which were conducted to develop the subsequent questionnaires for patients and healthy individuals and medical professionals, written informed consent was obtained from all the participants, and the interview responses did not contain any personal information. In the questionnaire survey of patients and healthy individuals, the recruitment of patients was performed through the iBRIDGE company (Freeasy system), informed consent was obtained on the web at the time of each response on Freeasy-affiliated websites, and responses were provided by the company in a completely anonymized form. For each question answered, respondents received points worth ¥1 (US $0.007) as a reward (according to the survey company's rules and regulations). Points can be exchanged for gift certificates or discount coupons through Freeasy-affiliated websites. In the questionnaire survey of medical professionals, a consent form was not required because the submission of responses was considered as consent to participate. To exclude duplicate responses, respondents were asked to provide only their facility name, title, and initials. In other words, no explicit personal information was included in the survey content, but potential identifiers were included. The responses to the medical professional survey, which included potential identifiers, were stored in a password-locked and encrypted hard disk in a secure room. No compensation was provided for the medical professional participants.

Results

Patient and Healthy Individual Characteristics

The proportion of women aged 70 years in the panel and the allocation to this group were slightly lower than those in the population distribution. In total, 92.5% (36,998/40,000) of individuals responded to the screening questionnaire and were classified into four categories: (1) regular visits and with online medical care experience (1478/36,998, 3.99%), (2) regular visits and without online medical care experience (16,269/36,998, 43.97%), (3) without regular visits and with online medical care experience (478/36,998, 1.29%), and (4) without regular visits and without online medical care experience (18,773/36,998, 50.74%). Among the respondents to the screening questionnaire, only 5.29% (1956/36,998) had online medical care experience. Second, all participants in group 1 (1478/1478, 100%) and some in group 2 (1522/16,269, 9.36%) were extracted (Figures S1A and S1B in Multimedia Appendix 6). Similarly, all participants in group 3 (478/478, 100%) and some in group 4 (2522/18,773, 13.43%) were extracted (Figures S1C and S1D in Multimedia Appendix 6). The main questionnaire was then administered to the selected individuals. The response rate in each group was >80% (1312/1478, 88.77%; 1281/1522, 84.17%; 404/478, 84.5%; and 2226/2522, 88.26% in groups 1, 2, 3, and 4, respectively). More than 80% of the respondents (4438/5223, 84.97%) were aged <60 years.

Regarding the number of nearby medical facilities, a major proportion of respondents with online medical care experience reported that there were many medical facilities nearby (Figure S2A in Multimedia Appendix 6). A greater percentage of respondents with online medical care experience reported that it took considerable effort to go to the hospital (Figure S2B in Multimedia Appendix 6). Among the respondents with regular visits, those with online medical care experience visited the clinic more frequently than those without (Figure S3A in Multimedia Appendix 6). One-way trips to the clinic were shorter for the group with no online medical care experience (Figure S3B in Multimedia Appendix 6).

Characteristics of Medical Professionals

Of the 1580 responses, 1552 (98.23%) were valid after excluding duplicate responses and responses from out-of-scope facilities. When categorized by job, 39.82% (618/1552) of the responses were from physicians, 27.58% (428/1552) were from nurses, and 32.6% (506/1552) were from administrative staff members. When divided by online medical care practice, 78.87% (1224/1552) did not practice online medical care at their facility, 9.15% (142/1552) practiced online medical care at their facility but with no direct involvement, and the remaining 11.98% (186/1552) were directly involved in online medical care.

XSL•FO
RenderX

At the facility level, responses were received from 16.2% (794/4900) of the sampled facilities. The response rates from specific functioning hospitals, core hospital for medical services in remote areas, and clinic for medical services in remote areas were higher than those from other hospitals (specific functioning hospitals: 38/88, 43%; clinic for medical services in remote areas: 130/504, 25.8%; core hospitals for medical services in remote areas: 35/170, 20.6%; other hospitals: 297/2196, 13.5%; other clinics: 298/1946, 15.31%. Note that 4 facilities were categorized as specific functioning hospitals and at the same time as core hospitals for medical services in remote areas).

Patients' and Healthy Individuals' Perspectives on Current Online Medical Care Use

In the group with online medical care experience, >50% of the patients selected internal medicine as the department in which they received online medical care (705/1312, 53.73% and 231/404, 57.2% of respondents with and without regular visits); followed by dermatology and psychiatry or psychosomatic medicine: 18.14% (238/1312), 16.6% (67/404), 18.06% (237/1312), 8.4% (34/404) for respondents with and without regular visits, respectively. In the group with no experience with online medical care, when asked to indicate which departments they would be willing to consider receiving online medical care in, internal medicine and psychiatry or psychosomatic medicine were the most common (587/1281, 45.82%; 786/2226, 35.31%; and 397/1281, 31%; 518/2226, 23.27% of respondents with and without regular visits, respectively; Table S1 in Multimedia Appendix 6).

Table 1 shows the results of the multivariate analysis of the online medical care experience with the variables of age of >60 years (binary variable), sex, number of nearby medical facilities (3-point Likert scale), and perceived effort of visiting the hospital (4-point Likert scale). The more nearby medical facilities, the more likely the respondent was to have experienced online medical care (adjusted odds ratio [aOR] 1.33, 95% CI 1.18-1.50 per point in the patient group; aOR 1.43, 95% CI 1.22-1.68 per point in the healthy individual group). The more effort they felt it was to visit the hospital, the more likely the respondent was to have experienced online medical care (aOR 1.48, 95% CI 1.35-1.63 per point in the patient group; aOR 1.57, 95% CI 1.38-1.78 per point in the healthy individual group).

Table 1. Multivariate analysis of factors related to experience and satisfaction with online medical care.

Factors	Experience, adjusted odds ratio (95% CI)		Satisfaction among respondents with experience, adjusted odds ratio (95% CI)	
	Patients (n=2593)	Healthy individuals (n=2630)	Patients (n=1312)	Healthy individuals (n=404)
Age of >60 y	1.09 (0.89-1.34)	1.07 (0.77-1.49)	1.04 (0.74-1.46)	0.54 (0.27-1.07)
Male sex	1.01 (0.86-1.18)	1.07 (0.86-1.32)	0.88 (0.67-1.16)	0.67 (0.40-1.13)
Medical facilities nearby[a]	*1.33 (1.18-1.50)* [b]	*1.43 (1.22-1.68)*	*1.40 (1.14-1.73)*	*1.77 (1.23-2.54)*
Effort for hospital visits[c]	*1.48 (1.35-1.63)*	*1.57 (1.38-1.78)*	*1.50 (1.27-1.76)*	*1.77 (1.31-2.40)*

[a]Ordinal variable (1=few, 2=reasonable, and 3=many).

[b]Italicization indicate statistical significance (the 95% CI of the adjusted odds ratio does not exceed 1).

[c]Ordinal variable (1=no, 2=negligible, 3=minimal, and 4=substantial).

The participants were asked to rate their level of satisfaction with online medical care on a 4-point scale (*very satisfied*, *generally satisfied*, *not very satisfied*, and *not satisfied at all*). Approximately 80% of the respondents in both groups selected the top 2 choices, "very satisfied" or "generally satisfied" (1023/1312, 77.97% and 322/404, 79.7% of respondents with and without regular visits, respectively; Table S2 in Multimedia Appendix 6). When stratified, respondents without regular visits aged 60 years or who had more medical facilities nearby were substantially more satisfied with online medical care (Table S2 in Multimedia Appendix 6). The groups that felt that going to the hospital required more work and that felt that online medical care was less burdensome than face-to-face medical care were more satisfied with online medical care (Table S2 in Multimedia Appendix 6).

As with the online medical care experience, a multivariate logistic analysis of satisfaction with online medical care (selecting the top 2 levels on a 4-point satisfaction scale was defined as being satisfied) showed that the greater the number of nearby medical facilities, the more satisfied the participants were (aOR 1.40, 95% CI 1.14-1.73 per point in the patient group; aOR 1.77, 95% CI 1.23-2.54 per point in the healthy individual group; Table 1). The more effort they felt it was to visit the hospital, the more satisfied they were (aOR 1.50, 95% CI 1.27-1.76 per point in the patient group; aOR 1.77, 95% CI 1.31-2.40 per point in the healthy individual group).

In the group with online medical care experience, the respondents answered that they wanted to use online medical care and searched for a medical facility that could provide it (870/1312, 66.31% and 217/404, 53.7% of the respondents with and without regular visits) rather than their physicians or medical facilities recommending the use of online medical care (442/1312, 33.69% and 152/404, 37.6% of those with and without regular visits). When the respondents without online medical care experience were asked about situations in which they would be willing to use online medical care, a relatively high number of respondents chose "when offered the option of online medical care by a physician or medical facility where they had already received face-to-face medical care" (696/1281, 54.33% and 932/2226, 41.87% of those with and without regular visits).

XSL•FO
RenderX

Medical Professionals' Perspectives on Current Online Medical Care Practice

Among the health care facilities that responded to the survey, the implementation rate of online medical care was 20.4% (162/794). Of the responding facilities, only 71.4% (567/794) reported using EMRs, and facilities that used EMR systems had significantly higher online medical care practice rates (142/567, 25% vs 20/227, 8.8%; $P<.001$). The implementation rate of online medical care did not differ between hospitals and clinics (75/366, 20.5% for hospitals and 87/428, 20.3% for clinics); however, among hospitals, the rate was higher for specific functioning hospitals than for core hospital for medical services in remote areas or other hospitals (16/38, 42.1% for specific functioning hospitals; 7/35, 20% for core hospital for medical services in remote areas; and 53/297, 17.8% for other hospitals).

The most frequently selected scenarios in which online medical care was provided were prescriptions of the same medication as usual (106/186, 57%), routine medical checkups (98/186, 52.7%), and difficulty attending hospital consultations owing to transportation issues and living in a remote area (68/186, 36.6%; Table 2). The most common problem with the system when implementing online medical care was that it could not be implemented on the same computer as the EMR system (Table 2).

Table 2. Online medical care situations and problems with its system as answered by health care professionals who are directly involved in online medical care (results of the medical professional survey; n=186).

Situations and problems	Participants, n (%)
Situations in which online medical care is provided	
Prescription of the same medication as usual	106 (57)
Routine medical checkups	98 (52.7)
Difficulty attending hospital consultations (eg, transportation issues and living in a remote area)	68 (36.6)
Outpatient visits for fever and other situations in which some type of infectious disease is anticipated	62 (33.3)
Explanation of test results	61 (32.8)
Visit for sudden onset of illness (acute illness) other than infectious diseases	19 (10.2)
Second opinion	13 (7)
Consultations when there is no specialist physician nearby	12 (6.5)
Consultation on whether to see a physician	12 (6.5)
Nutritional guidance	7 (3.8)
Consultation services related to childcare (eg, developmental counseling and childcare support)	4 (2.2)
Conferences for home medical care	4 (2.2)
Details not known by the respondents	3 (1.6)
Perinatal counseling services (eg, motherhood classes and genetic, infertility, and pregnancy complication counseling)	2 (1.1)
Other	12 (6.5)
Problems identified with the online medical care system	
Cannot be implemented on the same terminal as the electronic medical record system	95 (51.1)
Inability to screen share documents such as test results (must be captured on a camera)	82 (44.1)
Inability to draw illustrations on the screen (must be written on a piece of paper by hand and captured on a camera)	72 (38.7)
Communication environment problems causing issues	64 (34.4)
Difficulty operating on the patient side	59 (31.7)
Difficulty operating on the medical institution side	19 (10.2)
Can be performed on the same terminal as the electronic medical record system, but the software is not integrated and is difficult to use	13 (7)
None that apply	26 (14)

Most respondents (128/186, 68.8%) indicated that the hospital tasks required for online medical care were "obviously/somewhat more complicated" than those for face-to-face medical care (Table S3 in Multimedia Appendix 6). A major proportion of respondents among the administrative staff (33/43, 77%) chose "obviously/somewhat more complicated," and none chose "obviously easier."

Only 14% (26/186) of the respondents reported that they could see more patients per unit time with online medical care than with face-to-face medical care, whereas 44.1% (82/186) reported that they could see fewer patients with online medical care (Table S3 in Multimedia Appendix 6).

XSL•FO
RenderX

Patient Burdens Associated With Medical Visits

Both patients and healthy individuals and medical professionals were asked how the patient burden associated with medical visits differed between online medical care and face-to-face medical care. Regarding time and physical burden, most respondents indicated that online medical care was less burdensome, with comparable results for patients and healthy individuals and medical professionals (1352/2226, 60.74% to 307/404, 76% and 1285/2226, 57.73% to 116/142, 81.7%, respectively, depending on the respondent category; Table 3). Most patients and healthy individuals with online medical care experience indicated that online medical care reduces mental and financial burden (mental burden: 735/1312, 56.02% and 252/404, 62.4% for those with and without regular visits, respectively; financial burden: 671/1312, 51.14% and 172/404, 42.6% for those with and without regular visits, respectively).

Table 3. Percentage of respondents reporting that online medical care (OMC) is less burdensome than face-to-face medical care (results of the patient and healthy individual and medical professional surveys)[a].

Respondents	Time burden, n (%)	Physical burden, n (%)	Mental burden, n (%)	Financial burden, n (%)
Patients and healthy individuals				
With regular visits (OMC experience)				
Experienced (n=1312)	986 (75.15)	891 (67.91)	735 (56.02)	671 (51.14)
Not experienced (n=1281)	929 (72.52)	859 (67.06)	502 (39.19)	463 (36.14)
Without regular visits (OMC experience)				
Experienced (n=404)	307 (75.99)	310 (76.73)	252 (62.38)	172 (42.57)
Not experienced (n=2226)	1352 (60.74)	1285 (57.73)	792 (35.58)	753 (33.83)
Medical personnel (OMC practice)				
Not provided (n=1224)	897 (73.28)	933 (76.23)	438 (35.78)	549 (44.85)
Provided but not involved[b] (n=142)	103 (72.54)	116 (81.69)	70 (49.3)	65 (45.77)
Involved (n=186)	137 (73.66)	140 (75.27)	84 (45.16)	77 (41.4)

[a]In this survey, the financial burden included medical fees, transportation costs, line charges, and various other expenses related to medical visits.

[b]The facility provides OMC practice but the respondent himself or herself is not involved in it.

The acceptability of medical visits was also surveyed according to the frequency and modality of medical visits. The acceptability rate did not differ significantly between face-to-face medical care and online medical care in patients with online medical care experience (Table 4). When stratified by age, those aged ≥60 years with no online medical care experience tended to prefer face-to-face medical care over online medical care regardless of the frequency of medical visits.

XSL•FO
RenderX

Table 4. Acceptability of online medical care (OMC) and face-to-face medical care (FMC) in relation to the frequency of visits (results of the patient and healthy individual survey).

Patients and healthy individuals	Once a week			Once every 2 weeks			Once a month		
	FMC, n (%)	OMC, n (%)	*P* value	FMC, n (%)	OMC, n (%)	*P* value	FMC, n (%)	OMC, n (%)	*P* value
With regular visits (OMC experience)									
Experienced: all (n=1312)	797 (60.74)	833 (63.49)	.09	849 (64.71)	886 (67.53)	.07	1037 (79.04)	1028 (78.35)	.62
Experienced: aged <60 y (n=1074)	672 (62.57)	696 (64.8)	.21	702 (65.36)	741 (69)	*.03* [a]	832 (77.47)	845 (78.68)	.43
Experienced: aged ≥60 y (n=238)	125 (52.52)	137 (57.56)	.17	147 (61.76)	145 (60.92)	.81	205 (86.13)	183 (76.89)	*.0 05*
Not experienced: all (n=1281)	529 (41.3)	558 (43.56)	.17	663 (51.76)	617 (48.17)	*.03*	1075 (83.92)	778 (60.73)	*<.001*
Not experienced: aged <60 y (n=1039)	426 (41)	463 (44.56)	*.05*	536 (51.59)	515 (49.57)	.25	862 (82.96)	644 (61.98)	*<.001*
Not experienced: aged ≥60 y (n=242)	103 (42.56)	95 (39.26)	.41	127 (52.48)	102 (42.15)	*.01*	213 (88.02)	134 (55.37)	*<.001*
Without regular visits (OMC experience)									
Experienced: all (n=404)	205 (50.74)	258 (63.86)	*<.001*	246 (60.89)	280 (69.31)	*.001*	319 (78.96)	320 (79.21)	.90
Experienced: aged <60 y (n=354)	183 (51.69)	231 (65.25)	*<.001*	218 (61.58)	251 (70.9)	*.001*	279 (78.81)	286 (80.79)	.35
Experienced: aged ≥60 y (n=50)	22 (44)	27 (54)	.20	28 (56)	29 (58)	.78	40 (80)	34 (68)	.08
Not experienced: all (n=2226)	902 (40.52)	912 (40.97)	.69	1114 (50.04)	996 (44.74)	*<.001*	1531 (68.78)	1198 (53.82)	*<.001*
Not experienced: aged <60 y (n=1971)	763 (38.71)	796 (40.39)	.15	950 (48.2)	874 (44.34)	*.001*	1318 (66.87)	1063 (53.93)	*<.001*
Not experienced: aged ≥60 y (n=255)	139 (54.51)	116 (45.49)	*.02*	164 (64.31)	122 (47.84)	*<.001*	213 (83.53)	135 (52.94)	*<.001*

[a]Italicization indicates statistical significance.

Comparison of Face-to-Face Medical Care and Online Medical Care for Various Medical Procedures

The overwhelming majority of medical professionals answered that general impressions and physical examination findings were easier (chose “somewhat easier” or “clearly easier”) to obtain in face-to-face medical care than in online medical care (1467/1552, 94.52% and 1500/1552, 96.65%, respectively; Table 5). Most respondents also answered that verbal communication, interviews with family members, and building trust between patients and physicians were easier in face-to-face medical care (1059/1552, 68.23%; 772/1552, 49.74%; and 1207/1552, 77.77%, respectively). In contrast, a higher proportion of respondents indicated that visual understanding of the patient’s medical condition and behavior at home and visual understanding of the home and other environments were relatively easier in online medical care (642/1552, 41.37% and 1057/1552, 68.11%, respectively).

XSL•FO
RenderX

Table 5. Ease of performing various medical procedures (results of the medical personnel survey).

Procedures	Clearly easier in OMC[a], n (%)	Somewhat easier in OMC, n (%)	Almost the same, n (%)	Somewhat easier in FMC[b], n (%)	Clearly easier in FMC, n (%)
General impression (n=1552)	7 (0.45)	15 (0.97)	63 (4.06)	364 (23.45)	1103 (71.07)
Physical examination findings (n=1552)	8 (0.52)	7 (0.45)	37 (2.38)	310 (19.97)	1190 (76.68)
Verbal communication (n=1552)	7 (0.45)	33 (2.13)	453 (29.19)	476 (30.67)	583 (37.56)
Interviews with family members (n=1552)	35 (2.26)	174 (11.21)	571 (36.79)	348 (22.42)	424 (27.32)
Visual understanding of the patient's medical condition and behavior at home (n=1552)	181 (11.66)	461 (29.7)	297 (19.14)	241 (15.53)	372 (23.97)
Visual understanding of the home and other environments (n=1552)	507 (32.67)	550 (35.44)	222 (14.3)	108 (6.96)	165 (10.63)
Building trust between patients and physicians (n=1552)	2 (0.13)	14 (0.9)	329 (21.2)	519 (33.44)	688 (44.33)

[a]OMC: online medical care.

[b]FMC: face-to-face medical care.

Patients Appropriate for Online Medical Care

When asked what percentage of patients attending their hospital would benefit from online medical care, 72.62% (1127/1552) of medical professionals answered "10%" to "30%," with "10%" being the most common answer (636/1552, 40.98%; Figure S4A in Multimedia Appendix 6). A few respondents (203/1552, 13.08%) answered "0%." These results showed no significant difference in opinions between respondents who provided online medical care and those who did not (Figure S4B in Multimedia Appendix 6).

Regarding the maximum age at which a patient can receive online medical care alone, the most common response from medical professionals was "60 years old" (369/1552, 23.78%), and a similar proportion of respondents (344/1552, 22.16%) answered that age was not a relevant factor (Figure S4C in Multimedia Appendix 6). No significant differences were observed in the distribution of responses between respondents who provided online medical care or not (Figure S4D in Multimedia Appendix 6).

Factors Limiting Online Medical Care Expansion

In the survey of patients and healthy individuals, 16 options were presented as factors preventing the expansion or adoption of online medical care, and each respondent was allowed to select 3 options. The most selected factor was "online medical care requires patients to visit a hospital when tests or procedures are needed." In addition, "concerns about the adequacy of the clinical examination conducted during online medical care," "despite being aware of online medical care, there is uncertainty or lack of clarity regarding when online medical care is appropriate or desired," and "face-to-face communication is easier than web-bsed interactions" were commonly chosen (Table 6). The top 3 choices slightly differed depending on whether the patient had regular visits and whether the patient had previous experience with online medical care.

XSL•FO
RenderX

Table 6. Factors hindering online medical care (OMC) adoption and expansion.

Factors	Patients and healthy individuals, n (%)				Medical professionals, n (%)			
	Without regular visits and with OMC experience (n=404)	Without regular visits and without OMC experience (n=2226)	With regular visits and with OMC experience (n=1312)	With regular visits and without OMC experience (n=1281)	All (n=1552)	OMC practice		
						Not provided (n=1224)	Provided but not involved[a] (n=142)	Involved (n=186)
OMC increases administrative procedures on the part of medical institutions	43 (10.64)	250 (11.23)	195 (14.86)	131 (10.2)	598 (38.53)	466 (38.07)	51 (35.92)	*81 (43.55)* [b]
OMC requires more time and effort on the part of physicians	__[c]	—	—	—	397 (25.58)	322 (26.31)	29 (20.42)	46 (24.73)
OMC requires more time and effort on the part of the patient	40 (9.9)	269 (12.08)	181 (13.8)	132 (10.3)	380 (24.48)	303 (24.75)	29 (20.42)	48 (25.81)
OMC places a heavy financial burden on the medical institution	37 (9.16)	191 (8.58)	191 (14.56)	81 (6.3)	489 (31.51)	404 (33.01)	31 (21.83)	54 (29.03)
OMC imposes a heavy financial burden on patients	47 (11.63)	221 (9.93)	148 (11.28)	116 (9.1)	140 (9.02)	104 (8.5)	16 (11.27)	20 (10.75)
Difficult for the medical institution to construct a system and communication environment for OMC	62 (15.35)	307 (13.79)	215 (16.39)	172 (13.4)	*655 (42.2)*	*561 (45.83)*	36 (25.35)	58 (31.18)
Difficult for patients to download applications and build a communication environment for OMC	94 (23.27)	515 (23.14)	303 (23.09)	284 (22.2)	*880 (56.7)*	*701 (57.27)*	*82 (57.75)*	*97 (52.15)*
Concerns regarding the adequacy of the clinical examination during OMC	*128 (31.68)*	*705 (31.67)*	298 (22.71)	*420 (32.8)*	565 (36.4)	476 (38.89)	34 (23.94)	55 (29.57)
Face-to-face communication is easier than web-based interaction	112 (27.72)	592 (26.59)	273 (20.81)	*395 (30.84)*	620 (39.95)	507 (41.42)	48 (33.8)	65 (34.95)
OMC requires patients to visit a hospital to undergo tests or procedures	*126 (31.19)*	*718 (32.26)*	*400 (30.49)*	*509 (39.73)*	*906 (58.38)*	*736 (60.13)*	*80 (56.34)*	*90 (48.39)*
OMC may not be feasible for a larger proportion of patients	59 (14.6)	257 (11.55)	157 (11.97)	132 (10.3)	585 (37.69)	489 (39.95)	*53 (37.32)*	43 (23.12)
Concerns regarding breach of personal information through OMC	50 (12.38)	293 (13.16)	187 (14.25)	153 (11.94)	375 (24.16)	323 (26.39)	26 (18.31)	26 (13.98)
FMC[d] is preferable to OMC for the education of medical students and young physicians	—	—	—	—	249 (16.04)	207 (16.91)	22 (15.49)	20 (10.75)
Lack of awareness regarding OMC	79 (19.55)	440 (19.77)	262 (19.97)	222 (17.33)	341 (21.97)	260 (21.24)	38 (26.76)	43 (23.12)
Despite awareness, there is a lack of clarity on when OMC is appropriate or desired	*114 (28.22)*	*610 (27.4)*	*354 (26.98)*	357 (27.87)	386 (24.87)	305 (24.92)	35 (24.65)	46 (24.73)

XSL•FO
RenderX

Factors	Patients and healthy individuals, n (%)				Medical professionals, n (%)			
	Without regular visits and with OMC experience (n=404)	Without regular visits and without OMC experience (n=2226)	With regular visits and with OMC experience (n=1312)	With regular visits and without OMC experience (n=1281)	All (n=1552)	OMC practice		
						Not provided (n=1224)	Provided but not involved[a] (n=142)	Involved (n=186)
Lack of information on OMC providers	94 (23.27)	581 (26.1)	*326 (24.85)*	290 (22.64)	326 (21.01)	255 (20.83)	38 (26.76)	33 (17.74)
The level of satisfaction with FMC is high and many people do not need OMC	71 (17.57)	371 (16.67)	233 (17.76)	227 (17.72)	501 (32.28)	401 (32.76)	43 (30.28)	57 (30.65)
Regardless of the level of satisfaction with FMC, people tend to maintain the status quo	56 (13.86)	358 (16.08)	213 (16.23)	222 (17.33)	321 (20.68)	249 (20.34)	26 (18.31)	46 (24.73)

[a]The facility provides OMC practice but the respondent himself or herself is not involved in it.

[b]Italicization indicates the top 3 in each category.

[c]These items were not included as options in the patient and healthy individual survey.

[d]FMC: face-to-face medical care.

In the survey of medical professionals, 18 options were presented, and each respondent was allowed to select all of those that applied. When categorized by job type and practice status of online medical care, the top 2 choices for all groups were "online medical care requires patients to visit hospitals for mandatory tests and procedures" and "difficult for patients to download applications and build a communication environment for online medical care" (Table 6). When stratified by facility type, the results suggest that the complexity of administrative procedures was a disincentive for specific functioning hospitals, and 54% (43/80) of respondents selected "online medical care increases administrative procedures on the part of medical institutions" (Table S4 in Multimedia Appendix 6).

Discussion

Principal Findings

By conducting 2 nationwide surveys—one for patients and healthy individuals and the other for medical professionals—we acquired a comprehensive understanding of the current status of online medical care in Japan and identified critical concerns that must be addressed to facilitate its further adoption and expansion in the future. Although only 5.29% (1956/36,998) of the respondents had received online medical care and only 20.4% (162/794) of the institutions offered online medical care, satisfaction was high among approximately 80% (1023/1312, 77.97% and 322/404, 79.7%) of the patients and healthy persons who had received online medical care. If they had a large number of nearby medical facilities and felt that a lot of work was associated with medical visits, they had significantly more experience with online medical care and were more satisfied with it. The survey of medical professionals found that they felt that various medical procedures were more difficult to perform in online medical care compared to face-to-face medical care, whereas online medical care was perceived to reduce various patient burdens compared to face-to-face medical care. The number 1 factor currently preventing the adoption of online medical care in Japan, selected the most in both the patient and medical professional surveys, was that patients who receive online medical care are forced to switch to face-to-face medical care when tests or procedures become necessary. In addition, the lack of sufficient awareness of and education on online medical care both among patients and healthy individuals and among medical professionals may present challenges in performing certain medical procedures compared to face-to-face medical care.

In addition, Japan's unique background needs to be mentioned. Japan has a large number of medical institutions, ranking third after Colombia and South Korea in the number of medical institutions per population in 2020 according to an Organisation for Economic Co-operation and Development report [12]. The fact that medical institutions are typically located nearby and face-to-face medical care is relatively readily available may impact online medical care diffusion. In addition, laboratory centers that perform blood and urine tests and x-rays for online medical care are only experimental and not widely used in Japan.

Current Status of Online Medical Care in Japan

In the screening survey of approximately 40,000 patients and healthy individuals, only 5.29% (1956/36,998) of respondents had online medical care experience. Almost all the participants in the panel had smartphones, PCs, or other terminals and were presumed to have a high affinity for telemedicine as this respondent group had to respond via the internet; therefore, the actual experience rate of online medical care in Japan may be much lower.

XSL•FO
RenderX

In the survey of medical professionals, only 20.4% (162/794) of the facilities that responded had implemented online medical care. This result is similar to or slightly higher than the 15.6% telemedicine implementation rate in Japanese clinics in 2022 [13]. The implementation rate did not differ between hospitals and clinics; however, it was particularly high among specific functioning hospitals compared with other hospitals. This may be because specific functioning hospitals are involved in the treatment of or second opinions on rare or intractable diseases.

Although basic tools for telemedicine are available in Japan, many facilities are still operating in a traditional paper-based manner, with 28.6% (227/794) of the responding facilities not using EMRs. In the United States, data show that 88.2% of office-based physicians use EMRs [14]. In Japan, the prevalence of EMR use is 91.2% for large hospitals with ≥400 beds, which is almost the same as that in the United States. However, it is lower for general hospitals and general clinics at 57.2% and 49.9%, respectively [15]. The implementation rate of online medical care is high among facilities using EMR systems, suggesting that further dissemination of these systems may be necessary before online medical care becomes widespread.

Online medical care is most frequently used in internal medicine and dermatology. In addition to these departments, patients and healthy individuals with no online medical care experience preferred online medical care implementation in psychiatry and psychosomatic medicine. Although obtaining the opinions of psychiatry and psychosomatic medicine specialists regarding the feasibility of online medical care is necessary, many patients and healthy individuals are demanding online medical care in these departments. Consequently, telepsychiatry has become widespread in many countries, and in 15 of 17 regions in the world, telepsychiatry was reimbursed at the same rate as or a higher rate than in-person consultations during the COVID-19 pandemic [7].

Facilities engaging in medical services in remote areas, such as core hospital for medical services in remote areas and clinic for medical services in remote areas, also had a higher implementation rate of online medical care than others, which reflects the attempt to overcome geographical disadvantages through online medical care. In fact, in the survey of medical professionals, “difficulty in attending hospital consultations owing to transportation issues” and “living in a remote area” were often selected as situations in which online medical care was provided. However, in the survey of patients and healthy individuals, respondents with fewer nearby medical facilities expressed lower satisfaction with online medical care than those with more nearby medical facilities, although the reason for this contrasting finding was not clear. In previous reports, it has been noted that access to telemedicine in rural areas remains an important issue [16,17].

In the survey of patients and healthy individuals, satisfaction with online medical care was high. Online medical care was more satisfactory to patients and healthy individuals, especially those who felt that hospital visits required considerable effort and those who felt that online medical care reduced the burden of medical visits (Table 1). Furthermore, despite the availability of medical facilities nearby, those with online medical care experience perceived that online medical care reduced the efforts required to physically visit a hospital. The efforts required to make a hospital visit were determined by several other factors (eg, childcare, caregiving, work, busyness, and physical mobility difficulties) rather than distance from a medical facility. This suggests that online medical care may become even more widespread not only in underpopulated areas where online medical services are considered necessary but also in urban areas where there are many medical facilities. There is also a report from the United States that telemedicine has become more prevalent in urban areas than in rural areas after the COVID-19 pandemic [16].

Our survey revealed that medical professionals also recognize that online medical care reduces the burden on patients, improving the access of medical visits. In a survey of 31 countries conducted by the Organisation for Economic Co-operation and Development on the status of telemedicine implementation and other issues, national experts agreed that telemedicine services can positively impact several aspects of health system performance (ie, equity, efficiency, access, cost-effectiveness, effectiveness, safety, and quality) [17]. Online medical care is an important means of connecting patients who have interrupted their visits due to the burden of hospital visits but who still require medical care. However, this is only the case if they are able to use information and communications technology. In promoting the spread of online medical care, it is important to consider those who are not skilled in information and communications technology or else the so-called digital divide may be exacerbated.

In contrast, medical professionals implied that online medical care may not be as convenient to implement as face-to-face medical care as it is more difficult to perform various medical procedures necessary for diagnosis and determination of severity and there are still some problems with the system for performing online medical care.

Our survey results indicate that, while online medical care offers overwhelming advantages in terms of patient convenience and access to medical care, it is difficult for health care providers to offer a variety of medical services and set up a system for online medical care. As the magnitude of benefits for patients seems to outweigh the concerns on the part of health care providers, it seems that online medical care ought to be more widespread.

Factors Limiting Online Medical Care Expansion and Possible Solutions

The major limiting factor for online medical care adoption was that “online medical care requires patients to make a hospital visit when tests or procedures are needed.” The high selection rate for this option indicates that the lack of access to tests and physical examinations in online medical care is a drawback. This is consistent with the fact that many health care providers responded that performing various medical procedures was more difficult in online medical care than in face-to-face medical care (Table 5). To overcome this, it is necessary to develop separate access to tests or procedures specifically tailored for online medical care or alternative means of obtaining this diagnostic information. In particular, patients desire technologies

XSL•FO
RenderX

or tools that make it easier to obtain information for accurate condition assessment, such as video calls with higher resolution and 3D effects, as well as alternative methods for assessing physical findings. Furthermore, for some types of telemedicine, it may be beneficial to have a nurse by the patient's side while communicating with the physician online. The nurse's presence on the patient's side eliminates operational concerns and problems and provides information that is difficult to obtain in online medical care, such as physical findings. A previous systematic review summarizing the barriers to adapting to telemedicine reported that the top barriers included technically challenged staff, resistance to change, cost, reimbursement, and patient age and educational level; however, it did not address the need to visit the hospital for physical examinations and tests [18]. In a previous questionnaire survey with medical professionals and another interview survey with patients, concerns related to the lack of physical examinations were cited as barriers to the use of online medical care [19,20]. Such factors can be identified through questionnaires and interviews.

The following responses were also common: "difficult for patients to download applications and build a communication environment for online medical care," "despite being aware, people do not know when online medical care is appropriate or desired," and "concerns regarding the adequacy of the clinical examination in online medical care." Thus, applications for online medical care that can be easily understood from the patients' perspective should be developed. In the free-response section of the questionnaire, some respondents suggested that booths for online medical care should be set up at city halls, convenience stores, and other public places and that a unified standardization of applications is needed to avoid confusion (data not shown). In addition, offering educational activities regarding online medical care and releasing a list of facilities that implement online medical care in an easy-to-read format for the general public should be considered. Attitudes toward online medical care may change after individuals gain firsthand experience with it.

Promoting telemedicine among patients who are suitable for online medical care or patients who have a disease or condition that is suitable for online medical care could be a realistic strategy. For example, medical professionals answered that it seems to be easier to "visually grasp the patient at home" and "visually grasp the environment of the patient's daily life at home" with online medical care than with face-to-face medical care (Table 5). The conditions in which these medical procedures are highly important include caring for children and older patients whose usual condition is difficult to understand during hospital visits. The fact that those aged 60 years exhibited higher satisfaction rates with online medical care than those aged 60 years does not preclude its use by older adults; however, it may be a good idea to appeal to younger people, including those caring for older individuals. Health care providers who are younger are more likely to adopt telemedicine [21]. Given that many respondents expressed their willingness to switch to online medical care if offered by their trusted physicians and medical institutions that currently provide face-to-face medical care, approaches from the medical provider side should be considered for further promotion of online medical care. Notably, patients prefer continuity of care and receiving online medical care from a physician they already know to receiving face-to-face medical care from unfamiliar providers [22].

Limitations

Despite these advantages, our study has certain limitations that should be acknowledged. First, because this was a voluntary survey, there may be a selection bias in that those who are satisfied with or interested in online medical care may respond more positively to the survey. Furthermore, it is estimated that the participation of those with low digital literacy was low due to the web-based nature of the survey. Second, compared to the high response rate for the patient and healthy individual survey, the overall response rate for the medical professionals' survey was low (794/4900, 16.2%), and it is possible that the results do not reflect the whole picture because of nonresponse bias. Specifically, it is possible that responses were not received from medical professionals who were too busy to respond. In addition, although our survey did not allow us to investigate the reasons for nonresponse, if those opposed to online medical care tended to be nonrespondents, the results of this survey could be skewed from the overall trend. Third, although part of our results can be extrapolated to other countries, they reflect the actual situation in Japan as of 2023. Furthermore, because this was a questionnaire survey, recall bias and erroneous answers may have been included in the results.

Conclusions

In conclusion, our nationwide surveys revealed the current status of online medical care in Japan and simultaneously identified several problems and issues related to online medical care, which will be useful in considering its expansion or adoption. On the basis of the identified factors preventing the spread of online medical care, it is important to develop technology and improve the system of medical care so that tests and procedures can be performed seamlessly with online medica care, as well as informing and educating the general population, including patients and the medical community.

Acknowledgments

The authors thank all those who participated in the survey. They would also like to thank Editage for the English-language editing. This study was supported by Ministry of Health, Labour and Welfare (23IA2001).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

YS, YH, and MN contributed to study design and conception. Formulation of questionnaires was conducted by HK, SM, AI, HO, MK, SK, and RW. YS, YH, and MI analyzed and interpreted the data. YS, YH, and MI wrote the original draft. All authors equally contributed to the work, met the criteria for authorship stated in the Uniform Requirements for Manuscripts Submitted to Biomedical Journals, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Japanese questionnaire for patients and healthy individuals.
[DOC File , 164 KB - jmir_v26i1e64159_app1.doc]

Multimedia Appendix 2
English translated questionnaire for patients and healthy individuals.
[DOC File , 120 KB - jmir_v26i1e64159_app2.doc]

Multimedia Appendix 3
Japanese questionnaire for medical professionals.
[PDF File (Adobe PDF File), 408 KB - jmir_v26i1e64159_app3.pdf]

Multimedia Appendix 4
English translated questionnaire for medical professionals.
[DOC File , 142 KB - jmir_v26i1e64159_app4.doc]

Multimedia Appendix 5
Checklist for Reporting Results of Internet E-Surveys.
[XLSX File (Microsoft Excel File), 26 KB - jmir_v26i1e64159_app5.xlsx]

Multimedia Appendix 6
Tables S1-S4 and Figures S1-S4.
[PPTX File , 148 KB - jmir_v26i1e64159_app6.pptx]

References

1. Telehealth, telemedicine, and telecare: what's what? Federal Communications Commission. 2024. URL: https://www.fcc.gov/general/telehealth-telemedicine-and-telecare-whats-what [accessed 2024-05-27]
2. Guidelines for the appropriate implementation of online medical care. Ministry of Health, Labour and Welfare. 2022 Feb 28. URL: https://www.mhlw.go.jp/content/12601000/000901835.pdf [accessed 2024-05-27]
3. Maugeri A, Barchitta M, Basile G, Agodi A. Public and research interest in telemedicine from 2017 to 2022: Infodemiology study of Google trends data and bibliometric analysis of scientific literature. J Med Internet Res 2024 May 16;26:e50088 [FREE Full text] [doi: 10.2196/50088] [Medline: 38753427]
4. Koonin LM, Hoots B, Tsang CA, Leroy Z, Farris K, Jolly B, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic - United States, January-March 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 30;69(43):1595-1599 [FREE Full text] [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]
5. von der Groeben S, Czaplicki A, Hegerl U, Reich H. Telemedicine during the COVID-19 pandemic in Germany: results from three nationally representative surveys on use, attitudes and barriers among adults affected by depression. Internet Interv 2023 Apr;32:100622 [FREE Full text] [doi: 10.1016/j.invent.2023.100622] [Medline: 37091132]
6. Walley D, McCombe G, Broughan J, O'Shea C, Crowley D, Quinlan D, et al. Use of telemedicine in general practice in Europe since the COVID-19 pandemic: a scoping review of patient and practitioner perspectives. PLOS Digit Health 2024 Feb;3(2):e0000427 [FREE Full text] [doi: 10.1371/journal.pdig.0000427] [Medline: 38354211]
7. Kinoshita S, Cortright K, Crawford A, Mizuno Y, Yoshida K, Hilty D, et al. Changes in telepsychiatry regulations during the COVID-19 pandemic: 17 countries and regions' approaches to an evolving healthcare landscape. Psychol Med 2022 Oct;52(13):2606-2613 [FREE Full text] [doi: 10.1017/S0033291720004584] [Medline: 33243311]
8. Habib S, Alsulaim KB, Mobeirek OA, Alsaeed AM, Albawardi FA, Alqahtani YK, et al. Barriers and facilitators of telemedicine among physicians at a university hospital. Cureus 2023 Sep;15(9):e45078 [FREE Full text] [doi: 10.7759/cureus.45078] [Medline: 37842354]
9. Free and easy self-survey tool. Freeasy. URL: https://freeasy24.research-plus.net/ [accessed 2024-08-14]

XSL•FO
RenderX

10. Population statistics as of October 1, 2020. e-Stat. URL: https://www.e-stat.go.jp/dbview?sid=0003448228 [accessed 2024-08-14]
11. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
12. OECD data explorer - archive. Organisation for Economic Co-operation and Development. URL: https://data-explorer.oecd.org/vis?tenant=archive&df[ds]=DisseminateArchiveDMZ&df[id]=DF_HEALTH_REAC&df[ag]=OECD&dq=HOSPTHOS..&lom=LASTNPERIODS&lo=5&to[TIME_PERIOD]=false&ly[cl]=TIME_PERIOD&ly[rs]=COU&ly[rw]=UNIT&vw=tb [accessed 2024-08-15]
13. Kuwayama T, Kotani K. Relationship between clinics offering telemedicine and population density in Japan: an ecological study. Telemed Rep 2024;5(1):99-104 [FREE Full text] [doi: 10.1089/tmr.2023.0054] [Medline: 38595726]
14. Electronic medical records/electronic health records (EMRs/EHRs). US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm [accessed 2024-08-15]
15. Survey of medical facilities: trends in the diffusion of electronic medical record systems. Ministry of Health, Labour, and Welfare. URL: https://www.mhlw.go.jp/content/10800000/000938782.pdf [accessed 2024-08-15]
16. Leung LB, Yoo C, Chu K, O'Shea A, Jackson NJ, Heyworth L, et al. Rates of primary care and integrated mental health telemedicine visits between rural and urban Veterans Affairs beneficiaries before and after the onset of the COVID-19 pandemic. JAMA Netw Open 2023 Mar 01;6(3):e231864 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.1864] [Medline: 36881410]
17. The future of telemedicine after COVID-19: OECD policy responses to coronavirus (COVID-19). Organisation for Economic Co-operation and Development. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/the-future-of-telemedicine-after-covid-19_d46e9a02-en [accessed 2024-08-15]
18. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: A systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
19. Poljo A, Tynes DM, Timper K, Süsstrunk J, Kraljević M, Peterli R, et al. Breaking down barriers to bariatric care: a qualitative study on how telemedicine could transform patient experiences in a Swiss monocentric setting. BMJ Open 2024 May 20;14(5):e083142 [FREE Full text] [doi: 10.1136/bmjopen-2023-083142] [Medline: 38772586]
20. Thomassen EE, Berg IJ, Kristianslund EK, Tveter AT, Østerås N. Willingness, perceived facilitators and barriers to use remote care among healthcare professionals - a cross-sectional study. BMC Health Serv Res 2023 Nov 27;23(1):1307 [FREE Full text] [doi: 10.1186/s12913-023-10301-4] [Medline: 38012633]
21. Phillips G, Millhollon R, Elenwo C, Ford AI, Bray N, Hartwell M. Associations of clinical personnel characteristics and telemedicine practices. J Osteopath Med 2024 Oct 01;124(10):455-462 [FREE Full text] [doi: 10.1515/jom-2024-0023] [Medline: 38855816]
22. Pogorzelska K, Marcinowicz L, Chlabicz S. Understanding satisfaction and dissatisfaction of patients with telemedicine during the COVID-19 pandemic: An exploratory qualitative study in primary care. PLoS One 2023;18(10):e0293089 [FREE Full text] [doi: 10.1371/journal.pone.0293089] [Medline: 37847684]

Abbreviations

aOR: adjusted odds ratio
EMR: electronic medical record

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 11.07.24; peer-reviewed by W Kayo; comments to author 08.08.24; revised version received 25.08.24; accepted 11.10.24; published 01.11.24.

Please cite as:
Sugawara Y, Hirakawa Y, Iwagami M, Kuroki H, Mitani S, Inagaki A, Ohashi H, Kubota M, Koike S, Wakimizu R, Nangaku M
Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey
J Med Internet Res 2024;26:e64159
URL: https://www.jmir.org/2024/1/e64159
doi:10.2196/64159
PMID:

©Yuka Sugawara, Yosuke Hirakawa, Masao Iwagami, Haruo Kuroki, Shuhei Mitani, Ataru Inagaki, Hiroki Ohashi, Mitsuru Kubota, Soichi Koike, Rie Wakimizu, Masaomi Nangaku. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Internet Hospital Consultations on Outpatient Visits and Expenses: Quasi-Experimental Study

Yayuan Liu[1], BA; Haofeng Jin[1], PhD; Zhuoyuan Yu[2], BMed; Yu Tong[1,3], PhD

[1]School of Management, Zhejiang University, Hangzhou, China

[2]Department of Medical Administration, The Affiliated Dongyang Hospital, Wenzhou Medical University, Dongyang, China

[3]Center for Research on Zhejiang Digital Development and Governance, Hangzhou, China

Corresponding Author:
Yu Tong, PhD
School of Management
Zhejiang University
Hangzhou, 310058
China
Phone: 86 57188206827
Email: tong_yu@zju.edu.cn

Abstract

Background: Internet hospital consultations are emerging in China as a new channel for patients to access health care services. Unlike third-party health care platforms such as Haodf, Teladoc Health, and MDLive, internet hospitals seamlessly integrate patients' offline medical records with online consultations, offering a cohesive online and offline health care experience. However, its impact on outpatient visits remains ambiguous. While it may encourage outpatient visits due to better continuity of care, it could also reduce face-to-face visits because of the convenience of online consultations. Given that patients in China have the autonomy to freely choose their health care providers, it is critical for hospitals to understand the effect of this telehealth technology on outpatient visits.

Objective: This study aimed to analyze the impact of patients' adoption of internet hospital consultations on their outpatient frequency and expenses, and whether these impacts vary between urban and rural patients.

Methods: The data used in this study were collected from a public tertiary hospital situated in a southeastern county of China, covering internet hospital consultations from January 2021 to October 2022, and offline outpatient records from January 2020 to October 2022. The dataset also includes patient demographic information. To estimate the causal effect, we used a quasi-experimental design, combining the difference-in-differences (DiD) analysis with the propensity score matching (PSM). After performing PSM, 2065 pairs of patients (4130 patients) were obtained for data analysis.

Results: Our findings highlight 3 key results. First, patients' adoption of internet hospital consultations increases their frequency of outpatient visits by 2.4% per month ($P<.001$), and the associated expenses by 15.5% per month ($P<.001$). Second, such positive effects are more pronounced for patients residing in rural areas. Specifically, for every 1% increase in the distance between patients' residences and the county government (an urban center), the positive effect on monthly outpatient visits increases by 0.3% ($P=.06$), and the positive effect on monthly outpatient expenses increases by 2.4% ($P=.03$). Third, our post hoc analysis shows that rural patients living in areas with higher local health care quality experience a mitigated positive effect of internet hospital consultations, compared with those in areas with lower health care quality.

Conclusions: This study extends the research scope of telehealth technologies by investigating internet hospitals, which are characterized by the integration of online and offline services. Our findings suggest that patients' adoption of internet hospital consultations is associated with an increase in both the frequency and expenses of outpatient visits. In addition, these effects vary based on patients' urban-rural status and local health care quality. These insights offer valuable guidance for policy makers and health care providers in promoting and optimizing the development and operation of internet hospitals.

(J Med Internet Res 2024;26:e57609) doi:10.2196/57609

KEYWORDS

internet hospital; online consultation; telehealth; outpatient visits; outpatient expenses; urban-rural healthcare disparity

XSL•FO
RenderX

Introduction

Over the past decade, telehealth technologies are increasingly embraced by worldwide governments and hospitals to improve medical accessibility and efficiency. One promising innovation in China that has gained widespread attention is the internet hospitals [1]. Internet hospitals are online medical platforms managed by offline medical institutions, primarily by large public hospitals [2]. Through internet hospital, offline medical records and resources can be used to offer online health care services for remote patients [3,4].

Compared with third-party health care platforms operated by private enterprises such as Teladoc Health, MDLive, and Haodf [5-11], internet hospitals have advantages in terms of service scope and continuity. Without access to hospitals' internal records, third-party health care platforms offer limited services such as outpatient appointment booking and online medical consultations. In contrast, internet hospitals can integrate patients' offline medical records when providing online health care services [4]. This integration enables more comprehensive services that can be streamlined with patients' offline care, such as follow-up consultations, electronic prescriptions and chronic disease management, thereby having the potential for continuous and well-informed medical care [1,12].

Despite of the touted opportunities, whether the adoption of internet hospital consultation services (termed as internet hospital consultations thereafter) can benefit a hospital's offline business remains unclear. On one hand, internet hospital consultations may help a hospital reach and attract more outpatient visits [13]. In China, public hospitals operate in a competitive market where patients can freely choose which hospital to visit [14]. Internet hospital consultations help physical hospitals overcome geographical limitations and potentially engage with a larger number of patients. By facilitating online interactions and offering continuous medical services, internet hospitals may strengthen the bond between the hospital and patients and attract more patients. On the other hand, having internet hospital consultations might also reduce a hospital's offline outpatient visits by providing an alternative communication channel for doctors and patients [9], leading to a decrease in offline outpatient revenue. This issue is particularly salient for lower-tier hospitals in China, which operate below capacity and struggle with patient attrition [15,16].

Previous literature on the impact of online consultation services on offline visits is primarily based on third-party health care platforms. For instance, several studies have used secondary data from third-party consultation platforms, indicating that doctors offering online consultation services tended to receive more offline visits [8,10,13,17]. These findings highlight the role of online consultation services in demonstrating doctors' service quality and enhancing their reputation [13,18]. However, it is important to note that an increase in visits at the doctor's level does not necessarily imply a corresponding rise in visits for individual patient. The increase in offline visits may result from attracting a broader range of patients rather than increasing the frequency of visits by individual patients. [19]. In the context of internet hospitals, online and offline integration can make online consultations more efficient and accurate [12]. It remains unclear whether this enhanced capability will substitute a patient's offline outpatient visits or increase them.

With the objective of assessing the value of internet hospital consultations, it is imperative to understand the impact of internet hospital consultations at the patient level. Few previous related studies on this direction rely on the survey method, which is often limited by sample size and self-reporting bias [5,20-22]. In this paper, we use large scale secondary data obtained from internet hospitals to empirically evaluate the effectiveness of internet hospital consultation services on changing patients' offline visits. This understanding will help hospitals determine if they should invest additional resources in operating internet hospital consultations. Hence, we propose the first research question: What is the impact of patients' adoption of internet hospital consultations on their outpatient visits in the same hospital?

In addition to assessing the overall effect of internet hospital consultations, we are also interested in further exploring its heterogeneity between urban and rural patients. Online consultations hold promise in reducing physical constraints for patients to access medical services and alleviating the urban and rural health care disparities [14]. Previous research has demonstrated that health information technology can facilitate the flow of health care knowledge and social support from urban to rural areas [23-25]. For instance, Hwang et al [24] used exponential random graph model and presented evidence indicating that online consultations connect doctors in resourceful regions and patients in underserved areas. However, these studies often focus on medical knowledge disparities while overlooking the effort needed for in-person visits. Given that online medical knowledge is not adequate in addressing all medical issues, it is often necessary for patients to visit hospitals offline [26]. Rural patients typically encounter higher costs associated with seeking medical care offline, such as time and travel expenses [14]. These increased costs are likely to limit the effectiveness of internet hospital consultations in promoting offline visits for rural patients, thus challenging its ability to alleviate the urban-rural health care disparity. In order to gain a better understanding of the societal impact and contextual limitations of internet hospital consultations, we propose the second research question: Does the impact of internet hospitals consultations on offline outpatient visits differ between urban and rural patients?

In summary, this study aims to examine the impact of patients' adoption of internet hospital consultations on their outpatient visits (including outpatient frequency and outpatient expenses), and whether such impact varies between urban and rural patients. We leverage a dataset obtained from a public hospital situated in a southeastern county of China, and use a quasi-experimental design that combines the propensity score matching (PSM) technique with the difference-in-differences (DiD) analysis.

XSL•FO
RenderX

Methods

Data Background

The data for this study are obtained from a public tertiary hospital located in a southeastern county of China. With over 70 departments and nearly 3000 staff members, the hospital offers comprehensive health care services to the local residents of the county as well as patients from surrounding counties.

The focal hospital launched its internet hospital on January 1, 2021, in response to the government's policy on promoting "Internet Plus" health care [27]. Through the internet hospital consultation service, patients are able to communicate with doctors of the hospital by sending text and graphic messages, regarding matters such as medication, symptoms diagnosis, and treatment. Note that the consultation service is integrated with the offline electronic medical record system of the hospital. This allows doctors to access patients' outpatient history during the online consultation process, while patients could make outpatient appointment with doctors.

To promote the use of internet hospital consultations, the hospital offers this service free of charge to all patients. Meanwhile, to encourage doctors to participate in this service, the hospital provides doctors with a subsidy of 10 CNY (approximately US $1.4) for each completed consultation. In addition, for patients who are unable to use online consultation services themselves, such as the elderly or children, the hospital permits next-of-kin to consult on their behalf.

Data Collection and Processing

To investigate the impact of internet hospital consultations on outpatient visits, we focus specifically on patients with chronic diseases, as these patients typically require long-term management and higher levels of health care resources [28,29]. We select patients with representative chronic diseases from 4 typical chronic disease departments: neurology, rheumatology, cardiology, and endocrinology, resulting in a total sample size of 56,749 patients. Patients are not excluded based on age, gender, or other demographic factors. Details of selection process are provided in Multimedia Appendix 1. We collect 3 datasets of these patients: internet hospital consultation records, offline outpatient records, and demographic data of patients. To protect patient privacy, all records have been anonymized and carefully stripped of any personally identifiable information. Figure 1 illustrates the overall process of sample processing.

XSL•FO
RenderX

Figure 1. Flowchart of sample processing.

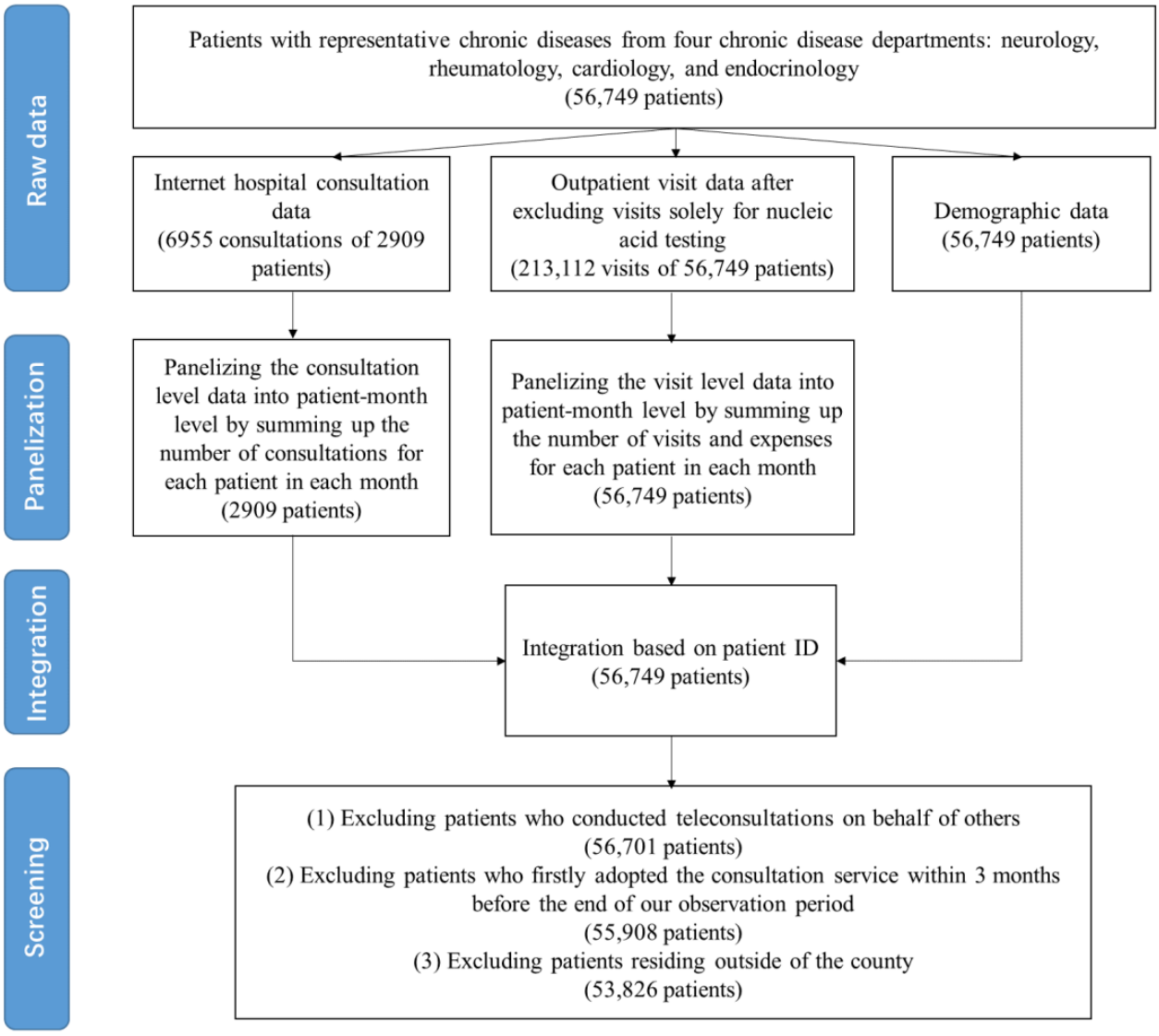

Internet Hospital Consultation Data

The internet hospital consultation data contains all consultations records of these patients from January 1, 2021 (the date of launch), to October 31, 2022. Each consultation record includes the date of consultation and the chief complaint of the patient. Throughout our sample period, a total of 6955 online consultations were conducted, involving 52 doctors and 2909 patients. We panelize the consultation level data into the patient-month level by summing up the number of consultations for each patient in each month. Accordingly, we obtain a panel data containing 98,906 observations for 2909 patients in 34 months.

Outpatient Data

We collect outpatient records for these patients between January 1, 2020, and October 31, 2022. This dataset captures information including the date of each visit and the corresponding medication expenses. To mitigate the potential impact of COVID-19 pandemic on our results, we excluded outpatient visits that were solely for nucleic acid testing, resulting in a total of 213,112 visits made by 56,749 patients. Note that the observation window for outpatient records was 12 months earlier than that for internet hospital consultations, which enables us to measure the impact of internet hospital consultations by comparing a patient's outpatient visits before and after adopting internet hospital consultations.

Consistent with previous literature [28,30], we panelize the outpatient visit level data into the patient-month level by summing up the number of visits and expenses for each patient in each month. In cases where a patient did not go to the hospital within a given month, the values of outpatient frequency and expenses are assigned to zero. Accordingly, we obtain a panel data containing 1,929,499 observations for 56,749 patients in 34 months.

Demographic Data

We gather demographic data of patients, including age, gender, and home address. The home address information will be used to capture the residential area of patients.

XSL•FO
RenderX

Data Integration

Finally, we integrate these 3 datasets based on the unique ID of patients, resulting in a panel dataset containing each patient's demographic information, consulting behavior, and outpatient behavior in the month level. To eliminate potential confounding factors that may interfere with our measurement of the impact of internet hospital consultations on outpatient visits, we impose 3 restrictions on the sample. We exclude patients who have conducted teleconsultations on behalf of others. To minimize the censored data problem, we exclude patients who used the consultation service for the first time within 3 months before the end of our observation period. To eliminate outliers, we exclude patients residing outside of the county where the hospital is located. These restrictions leave us with a final sample of 53,826 patients, and 2068 of them have used the internet hospital consultation service at least once.

Variables and Data Statistics

We construct 2 variables to capture the residential areas of patients. The first variable Area is a categorical variable, with patients divided into 3 categories based on the administrative region of their home address: patients residing in urban areas (Area=0), patients residing in rural areas bordering on urban areas (Area=1), and patients residing in rural areas not bordering on urban areas (Area=2). In addition, to capture the residential areas with more precision, we establish a continuous variable Distance by calculating the distance between patients' residence and the county government (based on latitude and longitude coordinates). The underlying rationale is that patients who reside farther away from the county government are more likely to be rural patients. Further elaboration regarding these 2 variables is presented in Figure 2.

Figure 2. Visualization of area and distance construction.

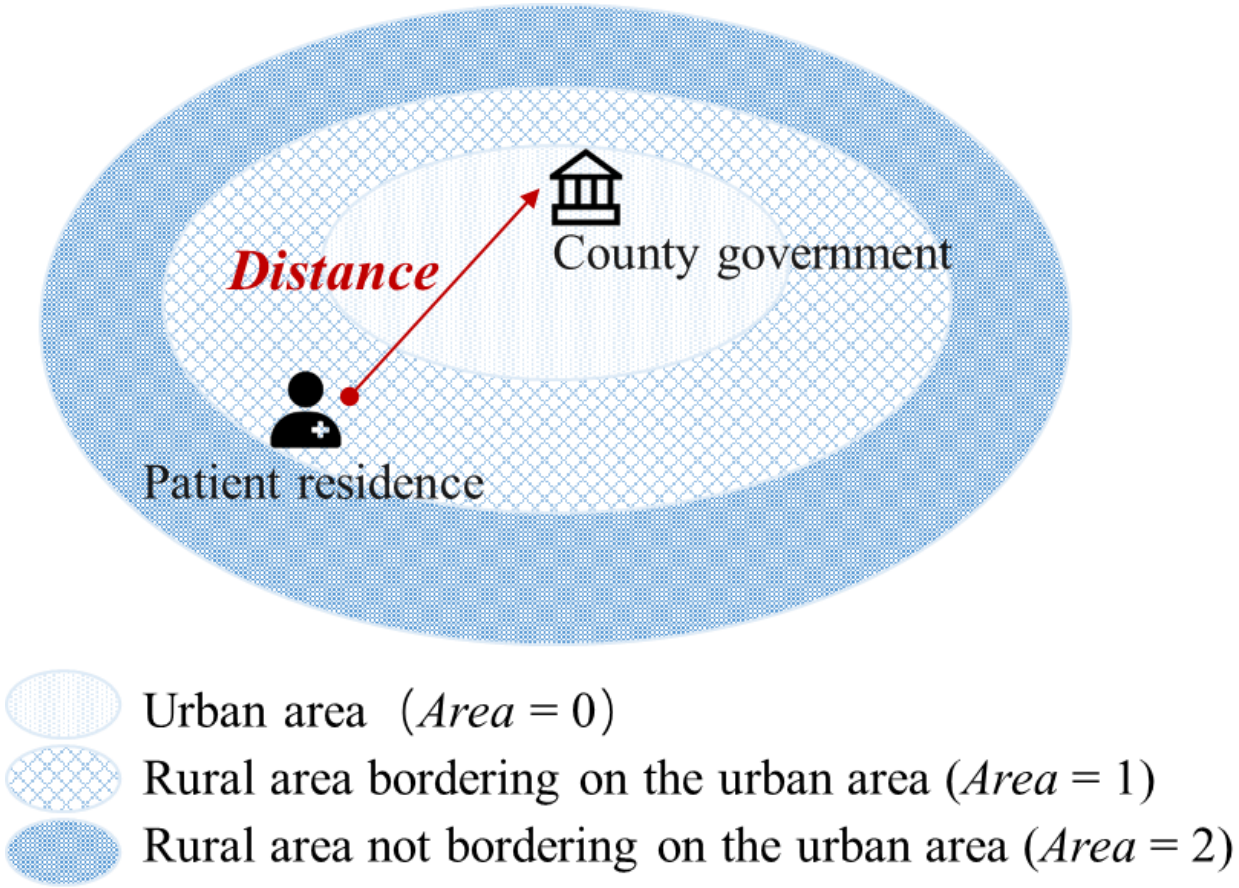

Table 1 reports descriptive statistics for all patients, adopters of internet hospital consultations, and nonadopters of internet hospital consultations. We use the term "adopters" for patients who have used the internet hospital service at least once during our observation window. Consistent with the traditional technology adoption patterns, we find that adopters are on average, about 16 years younger than nonadopters (age 45 years vs 61 years in 2022). Furthermore, adopters are more likely to be female (1096 vs 972), and living in urban areas (1338 vs 522). In terms of outpatient visits and expenses, adopters visited the hospital an average of 0.136 times per month with an expense of 17.87 CNY (approximately US $2.5), which is higher than nonadopters, who visited 0.109 times per month with an expense of 15.868 CNY (approximately US $2.3).

XSL•FO
RenderX

Table 1. Summary statistics of the patients.

Variables	Total sample (N=53,826)	Adopters (n=2068)	Nonadopters (n=51,758)
Categorical variables, n (%)			
Gender			
Men	27,667 (51.4)	972 (47)	26,695 (51.6)
Women	26,159 (48.6)	1096 (53)	25,063 (48.4)
Area			
0 (Patients living in the urban area)	31,847 (59.2)	1338 (64.7)	30,524 (58.9)
1 (Patients living in the rural area bordering on the urban area)	15,318 (28.4)	522 (25.2)	14,803 (28.6)
2 (Patients living in the rural area not bordering on the urban area)	6661 (12.4)	208 (10.1)	6458 (12.5)
Continuous Variables, mean (SD)			
Age (year)	60.645 (15.330)	45.439 (12.685)	61.244 (15.114)
Distance (km)	11.159 (10.245)	9.859 (9.450)	11.210 (10.272)
Monthly outpatient visits (times)	0.110 (0.154)	0.136 (0.158)	0.109 (0.154)
Monthly outpatient expenses (CNY, 1 CNY=US $0.14)	15.943 (15.944)	17.870 (43.674)	15.868 (44.221)

Empirical Analysis: Propensity Score Matching

We use a quasi-experimental design that combines the PSM technique and the DiD analysis to examine the impact of internet hospital consultations [31,32]. The purpose of PSM is to mitigate the self-selection bias, as the treatment, whether a patient has adopted the internet hospital consultations, was not randomly assigned but selected by patients themselves. Its basic idea is to find untreated patients who share highly similar characteristics with the treated patients before the launch of the internet hospital. Then the differences in outcomes between treated and untreated groups can be attributed to the treatment assignment [33].

Specifically, we conduct a logit regression that regresses whether a patient has adopted internet hospital consultations on a set of covariates, including patient age, gender, distance from home to the county government, number of outpatient visits (Outpatient_n), number of examinations (Exam_n), department most frequently visited (Department), and whether has visited a doctor who will offer online consultation services in the future (Doctor_online). The variable Doctor_online aims to address the issue of self-selection among doctors. By ensuring that the treated and control patients are had either visited or not visited doctors providing internet hospital consultations, we mitigated the influence of doctors' self-selection on patients' adoption of internet hospital consultations. All of these covariates are calculated based on the data window before the launch of the internet hospital [34]. The regression analysis results in a propensity score for each patient, indicating their likelihood of adopting internet hospital consultations.

Consequently, we match each treated patient with a control patient based on their propensity scores using one-to-one nearest neighboring matching, note that the covariates Department and Doctor_online are conducted exact matching. We successfully obtain 2065 pairs of observations (with 3 samples failing to match) to be used for the DiD analysis in the next subsection. Table 2 demonstrates that our matching effectively achieves a well-balanced distribution between treated and control groups, as indicated by mean differences between the 2 groups being insignificant for each of the covariates after matching [33]. The summary statistics for the 2065 paired patients across the entire observation window are provided in Table 3.

XSL•FO
RenderX

Table 2. Balance test.

Variable	Mean		*t* test	
	Treated	Control	*t* test (*df*)	*P* value
Age				
Unmatched	45.510	61.250	–46.707 (53,824)	<.001
Matched	45.539	45.499	0.099 (4128)	.92
Gender				
Unmatched	0.470	0.516	–4.082 (53,824)	<.001
Matched	0.471	0.492	–1.370 (4128)	.17
Distance				
Unmatched	9.904	11.209	–5.683 (53,824)	<.001
Matched	9.907	9.619	0.982 (4128)	.33
Outpatient_n				
Unmatched	0.853	1.210	–7.028 (53,824)	<.001
Matched	0.852	0.797	1.008 (4128)	.31
Exam_n				
Unmatched	0.998	1.025	–0.434 (53,824)	.66
Matched	0.998	0.842	1.481 (4128)	.14
Doctor_online				
Unmatched	0.284	0.407	–11.181 (53,824)	<.001
Matched	0.284	0.284	0 (4128)	>.99
Department				
Unmatched	0.681	1.038	–13.238 (53,824)	<.001
Matched	0.679	0.679	0 (4128)	>.99

Table 3. Summary statistics after matching.

Variables	Adopters (n=2065)	Nonadopters (n=2065)
Categorical Variables, n (%)		
Gender		
Men	972 (47.1)	1016 (49.2)
Women	1093 (52.9)	1049 (50.8)
Area		
0 (patients living in the urban area)	1335 (64.6)	1362 (66)
1 (patients living in the rural area bordering on the urban area)	522 (25.3)	499 (24.1)
2 (patients living in the rural area not bordering on the urban area)	208 (10.1)	204 (9.9)
Continuous Variables, mean (SD)		
Age (year)	45.539 (12.719)	45.499 (13.136)
Distance (km)	9.907 (9.490)	9.619 (9.379)
Monthly outpatient visits (times)	0.136 (0.158)	0.083 (0.105)
Monthly outpatient expenses (CNY, 1 CNY= US $0.14)	17.900 (43.516)	9.229 (25.686)

XSL•FO
RenderX

Difference-in-Differences Model

The PSM procedure allows us to simulate a quasi-experiment, where we match each treated patient with a comparable untreated patient based on observable covariates. To address concerns regarding selection bias on unobservable factors, we conduct a generalized DiD (2-way fixed effects) analysis with the following specification:

$$\ln(DV_{it}) = \beta_1 Adopter_{it} + \gamma_i + \eta_t + \varepsilon_{it} \quad (1)$$

where i denotes a patient index and t denotes the year-month index. The dependent variables (DVs) are the monthly frequency of outpatient visits, denoted as $Visits_{it}$, and the monthly expense of outpatient visits, denoted as $Expense_{it}$, respectively. We take the natural logarithm transformation to DVs (+1) to control their skewness. The binary variable $Adopter_{it}$ equals 1 if patient i has adopted the internet hospital consultation service on or before year-month t. For nonadopters, the value of $Adopter_{it}$ remains 0 for all months. We introduce patient fixed effects γ_i to control time-invariant patient characteristics, as well as year-month fixed effects to control for time trends, denoted by η_t. The error term is denoted by ε_{it}. The coefficient β_1 is of our interest, which measures the effect of patient adoption of internet hospital consultations on the frequency or expenses of outpatient visits.

To investigate the moderating effect of patients' residence on the relationship between internet hospital consultation adoption and outpatient visits, we add an interaction term between $Adopter_{it}$ and $Area_i$ to the baseline model, as shown in Equation (2). We do not include the term of $Area_i$ in the model because it has been absorbed by the patient fixed effects γ_i. Similarly, we add the interaction term between $Adopter_{it}$ and $Distance_i$ as shown in Equation (3).

Our primary focus is on the coefficient β_2, where a positive value would indicate that the effect of internet hospital consultation adoption on the frequency and expenses of outpatient visits is amplified for patients residing in rural areas.

$$\ln(DV_{it}) = \beta_1 Adopter_{it} + \beta_2 Adopter_{it} \times Area_i + \gamma_i + \eta_t + \varepsilon_{it} \quad (2)$$

$$\ln(DV_{it}) = \beta_1 Adopter_{it} + \beta_2 Adopter_{it} \times \ln(Distance_i) + \gamma_i + \eta_t + \varepsilon_{it} \quad (3)$$

Ethical Considerations

This study has been granted ethical approval by the medical ethics committee of Dongyang Hospital (NO.2024-YX-019). The data used in this study were anonymous and did not contain any personally identifiable information. Furthermore, reidentification from this dataset was not feasible. This study did not involve any experimental manipulation of human subjects, and did not pose any physical, psychological, or financial harm or risks to the patients. The online consultations and offline outpatient visits involved in this research were truly initiated by patients and were not subjected to any artificial intervention or control.

Results

Main Results

The estimated coefficients are presented in Table 4, with P values provided in parentheses. The first 2 columns represent estimation results of Equation (1). The significantly positive coefficient of Adopter indicates that the adoption of internet hospital consultations could increase a patient's monthly frequency of outpatient visits by 2.4% (P<.001), and could lead to a 15.5% increase in a patient's monthly expenses (P<.001).

XSL•FO
RenderX

Table 4. Effects of internet hospital consultation adoption on outpatient visits.

Variables	Estimation Results of Equation (1)		Estimation Results of Equation (2)		Estimation Results of Equation (3)	
	DV[a]=Visit	DV[a]=Expense	DV[a]=Visit	DV[a]=Expense	DV[a]=Visit	DV[a]=Expense
Adopter, coefficient (*P* value)	0.024 (<.001)	0.155 (<.001)	0.019 (<.001)	0.126 (<.001)	0.012 (<.001)	0.115 (<.001)
Adopter×1.Area (Rural area bordering on urban areas), coefficient (*P* value)	__[b]	__[b]	0.003 (.76)	0.051 (.34)	__[b]	__[b]
Adopter×2.Area (Rural area not bordering on urban areas), coefficient (*P* value)	__[b]	__[b]	0.028 (.04)	0.161 (.07)	__[b]	__[b]
Adopter×ln (Distance), coefficient (*P* value)	__[b]	__[b]	__[b]	__[b]	0.003 (.06)	0.024 (.03)
Constant, coefficient (*P* value)	0.036 (<.001)	0.193 (<.001)	0.036 (<.001)	0.141 (<.001)	0.036 (<.001)	0.141 (<.001)
Patient FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations, n	138,382	138,382	138,382	138,382	138,382	138,382
Patients, n	4130	4130	4130	4130	4130	4130
R^2	0.012	0.011	0.012	0.011	0.012	0.011

[a]DV: dependent variable.

[b]—: not applicable.

The last 4 columns in Table 4 present the results of the moderating effects of patients' residence. Coefficients of Adopter×2.Area turn out to be significantly positive (*P*=.04, *P*=.07). This suggests that compared with the baseline, that is, patients living in urban areas, patients living in rural areas not bordering on urban areas experience a higher increase in the frequency of outpatient visits (or the expenses of outpatient visits) after adopting internet hospital consultations. However, patients living in rural areas bordering on urban towns (Area=1) do not show a significant difference with the baseline, as indicated by the insignificant coefficients of Adopter×1.Area. Consistently, the significantly positive coefficients of Adopter×ln(Distance) imply that patients residing in more remote areas would experience a greater increase in outpatients visits and outpatient expenses after the adoption (*P*=.06 and *P*=.03). In addition, we conduct subgroup analysis by the department that the patient most frequently visited (Department) and the region where the patient resides (Area). The results are provided in Multimedia Appendix 2.

Post Hoc Analysis: The Moderating Effect of Local Health Care Quality

Based on the above analysis, we find that the internet hospital consultation has a positive impact on both the frequency and expenses of outpatient visits, and this impact is more pronounced among rural patients. These findings can be attributed to the educational function facilitated by internet hospital consultations, which effectively enhances access to medical knowledge, providing patients with a greater understanding of their health care needs and encouraging them to seek offline medical treatment [35]. In particular, for rural patients with limited medical knowledge and inadequate local medical resources, internet hospital consultations would play a more pivotal role in promoting and facilitating their outpatient visits.

To verify this possible mechanism, we focus on patients living in rural areas (Area=1 and Area=2), and investigate how the local health care quality moderates the positive impact of internet hospital consultations on outpatient visits. We expect that for patients living in rural towns with relatively better health care quality, the positive effect would be diminished, as patients are more likely to choose nearby hospitals for treatment after receiving health knowledge and advice from internet hospital consultations.

We construct the variable CenHospStaff to capture the health care quality of each rural town, which represents the number of medical staff at the central hospital of the town. In China, the central hospital is typically the leading medical institution within the primary health care system, and serves as a convenient option for comprehensive treatment locally. We posit that towns with better-staffed central hospital generally boast more extensive departments and better-equipped medical facilities, and exhibit a higher health care quality [36,37]. The average value of CenHospStaff is 99.25 among 12 rural towns that are included in our sample. The SD of 113.04 indicates significant variability in health care quality among these towns.

To investigate how the relationship between internet hospital consultation adoption and outpatient visits varies based on the local health care quality, we add the interaction terms of Adopter×ln(CenHospStaff) to Equation (1). The estimated coefficients are presented in Table 5, with *P* values provided in parentheses. The coefficient of Adopter×ln(CenHospStaff) are both significantly negative when DV is visit frequency and expenses (*P*=.08 and *P*=.10), indicating that after adopting the

XSL•FO
RenderX

internet hospital consultations, patients residing in rural towns with better health care quality experience a mitigated increased effect on outpatient visits and expenses.

Table 5. Results of the moderating effect of the local health care quality.

Variables	DV[a]=Visit	DV[a]=Expense
Adopter, coefficient (*P* value)	0.067 (.01)	0.382 (.01)
Adopter×ln (CenHospStaff), coefficient (*P* value)	–0.010 (.08)	–0.049 (.10)
Constant, coefficient (*P* value)	0.035 (<.001)	0.183 (<.001)
Patient FE	YES	YES
Year-month FE	YES	YES
Observations, n	48,923	48,923
Patients, n	1460	1460
R^2	0.012	0.011

[a]DV: dependent variable.

In summary, the results generally align with our expectations. That is, patients residing in rural areas with relatively worse health care quality would experience an amplified positive impact of consultations on both frequency and expenses of outpatient visits. These findings suggest that internet hospital consultations hold the potential to alleviate health care disparities by providing rural patients with advanced medical knowledge, and encouraging them to seek offline medical treatment.

Discussion

Principal Findings

Using a unique dataset from a county-level hospital and using DiD analysis with PSM, we find that the adoption of internet hospital consultation triggers additional outpatient visits and expenses by patients. Furthermore, we find that the positive effect of internet hospital consultations on outpatient visits and expenses is amplified for patients residing in rural areas. Our post hoc analysis further reveals that online consultations would have a larger impact among patients residing in areas with inadequate health care resources. All of these results suggest that internet hospital consultations have the potential to alleviate health care disparity between rural and urban areas by providing advanced medical knowledge, and encouraging more outpatient visits.

Limitations

Our analysis has the following 3 limitations that may motivate future work. First, a potential limitation is the generalizability of our findings. Given the limited availability of data, our sample only includes patients with representative chronic diseases in 4 specific departments. However, different diseases may exhibit distinct characteristics and treatment patterns, necessitating caution when extrapolating the study findings to other chronic diseases. In addition, the empirical analysis is based on a single internet hospital established by a county-level hospital, which primarily serves local residents. However, for some internet hospitals established by top-ranking hospitals in China, they may serve patients from the entire province or even the whole country [14]. Future studies could focus on the impact of internet hospitals established by hospitals at different levels, enabling a more comprehensive understanding of the role of internet hospitals in the health care system.

A second limitation is that we have only considered the impact on the frequency and expenses of patients' outpatient visits, but inpatient visits have not been considered due to the data limitation. Despite the growing relevance of internet hospitals, the literature remains limited in this context. Future analyses can adopt the analytical framework and methodologies used in this paper, while considering more comprehensive outcomes, such as inpatient visits and patient satisfaction.

Third, while we are able to design a quasi-experimental method combining PSM with DiD analysis to alleviate the potential endogeneity issues, this method is limited in its ability to fully establish causality. An ideal study would leverage randomized controlled trial to better examine the causal relationship between internet hospital consultations and patients' health care behaviors.

Theoretical Contributions

This paper makes 3 key contributions to existing literature. First, it extends the research scope of telehealth technologies by being one of the first attempts to investigate the impact of online consultation services on outpatient visits in the context of internet hospitals [4]. Current literature on online consultations primarily concentrates on third-party online health care platforms, which often lack patients' offline medical records and resources necessary for providing physical medical services [8,9,17,21,22]. Few previous related studies on this direction rely on the survey method, which could be limited by sample size and self-reporting bias [5,20-22]. Our focus adds to the literature by providing solid evidence on the positive effect of hospital-initiated platform service, that is, internet hospital consultations, which can be seamlessly integrated with the hospital's offline medical care services. Our data analysis results reveal that internet hospital consultation services increase both the frequency of patients' offline outpatient visits and their associated expenses. This suggests that internet hospitals can, to a certain extent, help hospitals attract more patients and enhance their competitive edge.

XSL•FO
RenderX

Second, this study expands our understanding of patients' reaction toward online consultation services. Previous literature primarily focused on doctors as the unit of analysis, revealing that those offering online medical consultations tended to see an increase in offline patient visits [10,17]. These studies highlight the role of online consultation services in showcasing doctors' service quality and enhancing their reputation [13,18]. However, it remains unclear whether the increase in outpatient visits is attributed to an expanding patient population or to individual patients making more frequent visits [19]. This study focuses on patients as the unit of analysis and hence can obtain a nuanced understanding on how internet hospital consultations affect offline visits of patients with representative chronic conditions. Our findings indicate that internet hospital consultations promote patient adherence to ongoing diagnosis and treatment, thereby facilitating continuous medical care. This expands our comprehension of how online medical consultations can effectively support patient management strategies.

Third, to the best of our knowledge, this study is among the first to empirically examine the role of patients' urban-rural status in moderating the impact of online consultations on offline visits [38]. Previous research has suggested that online consultation is a solution to alleviate health care disparities by providing useful health information to rural patients [14,23-25], but it remains unclear whether the information acquired will actually impact the outpatient behavior of rural patients, particularly in terms of their outpatient visits. This study goes 1 step further to demonstrate that such health information will encourage rural patients to increase their outpatient visits and prescriptions at the hospital. These findings contribute to a deeper understanding of the social value of internet hospital consultations.

Conclusions

In recent years, internet hospital consultations have emerged as a novel approach to providing health care services and education in China. Our research suggests that the adoption of this service by patients potentially increases their frequency of offline outpatient visits, particularly among individuals residing in remote and rural areas. These findings provide critical insights for multiple stakeholders. For hospitals and doctors, the results indicate that internet hospital consultations may serve as a complementary tool to attract and retain patients, particularly those from underserved regions. The seamless integration of online and offline services can enhance patient engagement by offering continuity of care and strengthening the bond between patients and the hospital. For policymakers, the findings suggest that internet hospital consultations could play a role in improving health care access, especially in rural areas where health care services are often limited. Furthermore, by highlighting the potential for internet hospitals to increase hospital revenue through more frequent outpatient visits and higher expenses, these findings may provide an incentive for more hospitals to adopt and expand internet hospital services.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (grant 72394371, 72091312, and 71821002) and Zhejiang University Qizhen Program (grant 226-2024-00232).

Data Availability

The aggregated data of this study (eg, summary statistics) are available from the corresponding author on reasonable request. The individual-level data are not publicly available because of the request of the hospital and data privacy laws and regulations in China.

Authors' Contributions

YL contributed to the conceptualization, investigation, data curation, formal analysis, validation, writing the original draft, and review and editing. HJ contributed to the investigation, methodology, validation, and review and editing. ZY contributed to the resources and methodology. YT contributed to the conceptualization, funding acquisition, supervision, and review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screening process for patients with chronic disease.
[DOCX File , 68 KB - jmir_v26i1e57609_app1.docx]

Multimedia Appendix 2
Split sample analysis by department and patient residence.
[DOCX File , 38 KB - jmir_v26i1e57609_app2.docx]

References

XSL•FO
RenderX

1. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015;3(8):e445-e446 [FREE Full text] [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
2. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
3. Li L, Liu G, Xu W, Zhang Y, He M. Effects of Internet Hospital Consultations on Psychological Burdens and Disease Knowledge During the Early Outbreak of COVID-19 in China: Cross-Sectional Survey Study. J Med Internet Res 2020 Aug 04;22(8):e19551 [FREE Full text] [doi: 10.2196/19551] [Medline: 32687061]
4. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in china: documentary analysis and qualitative interview study. J Med Internet Res 2021;23(7):e22330 [FREE Full text] [doi: 10.2196/22330] [Medline: 34283025]
5. Qi M, Cui J, Li X, Han Y. Influence of E-consultation on the intention of first-visit patients to select medical services: results of a scenario survey. J Med Internet Res 2023;25:e40993 [FREE Full text] [doi: 10.2196/40993] [Medline: 37115615]
6. Lu W, Wu H. How online reviews and services affect physician outpatient visits: content analysis of evidence from two online health care communities. JMIR Med Inform 2019;7(4):e16185 [FREE Full text] [doi: 10.2196/16185] [Medline: 31789597]
7. Wu H, Deng Z. Do physicians' online activities impact outpatient visits? an examination of online health communities. PACIS 2019 Proceedings 2019:207 [FREE Full text]
8. Wu H, Lu N. Online written consultation, telephone consultation and offline appointment: an examination of the channel effect in online health communities. Int J Med Inform 2017;107:107-119. [doi: 10.1016/j.ijmedinf.2017.08.009] [Medline: 29029686]
9. Huang N, Yan Z, Yin H. Effects of Online–Offline Service Integration on e‐Healthcare Providers: A Quasi‐Natural Experiment. Prod Oper Manag 2021 Aug 01;30(8):2359-2378. [doi: 10.1111/poms.13381]
10. Wang L, Yan L, Zhou T, Guo X, Heim GR. Understanding physicians' online-offline behavior dynamics: an empirical study. Inf Syst Res 2020;31(2):537-555. [doi: 10.1287/isre.2019.0901]
11. Graham K, Wilson RRA, Ewing A, Smith J, Dalkin BH, Woods BC, et al. Comparison of patient satisfaction between face-to-face and telehealth modalities in nephrolithiasis nutritional counseling. Urol Pract 2024;11(2):271-275. [doi: 10.1097/UPJ.0000000000000498] [Medline: 38215012]
12. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021;23(9):e18307 [FREE Full text] [doi: 10.2196/18307] [Medline: 34342267]
13. Liu X, Guo X, Wu H, Wu T. The impact of individual and organizational reputation on physicians' appointments online. Int J Electron Commer 2016;20(4):551-577. [doi: 10.1080/10864415.2016.1171977]
14. Tong Y, Tan CH, Sia CL, Shi Y, Teo HH. Rural-urban healthcare access inequality challenge: transformative roles of information technology. MIS Q 2022;31(2):159-169.
15. 2020 China health statistics yearbook.: National Health Commission of China; 2021. URL: http://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202112/dcd39654d66c4e6abf4d7b1389becd01.shtml [accessed 2024-10-11]
16. Li Z. Government subsidy schemes for coordinating healthcare referral in hierarchical service systems. J Ind Eng Eng Manag 2023;37(3):103-117.
17. Fan W, Zhou Q, Qiu L, Kumar S. Should Doctors Open Online Consultation Services? An Empirical Investigation of Their Impact on Offline Appointments. Inf Syst Res 2023 Jun;34(2):629-651. [doi: 10.1287/isre.2022.1145]
18. Chiu YL, Wang JN, Yu H, Hsu YT. Consultation pricing of the online health care service in China: hierarchical linear regression approach. J Med Internet Res 2021;23(7):e29170 [FREE Full text] [doi: 10.2196/29170] [Medline: 34259643]
19. Zhou M, Li X, Burtch G. Healthcare Across Boundaries: Urban-Rural Differences in the Consequences of Telehealth Adoption. Inf Syst Res 2024 Sep;35(3):1092-1113 [FREE Full text] [doi: 10.1287/isre.2021.0380]
20. Gleason N, Prasad PA, Ackerman S, Ho C, Monacelli J, Wang M, et al. Adoption and impact of an eConsult system in a fee-for-service setting. Healthc (Amst) 2017;5(1-2):40-45. [doi: 10.1016/j.hjdsi.2016.05.005] [Medline: 27469441]
21. López Seguí F, Vidal-Alaball J, Sagarra Castro M, García-Altés A, García Cuyàs F. General practitioners' perceptions of whether teleconsultations reduce the number of face-to-face visits in the catalan public primary care system: retrospective cross-sectional study. J Med Internet Res 2020;22(3):e14478 [FREE Full text] [doi: 10.2196/14478] [Medline: 32175914]
22. Xing W, Hsu PY, Chang YW, Shiau WL. How does online doctor–patient interaction affect online consultation and offline medical treatment? Ind Manag Data Syst 2020;120(1):196-214. [doi: 10.1108/imds-05-2019-0261]
23. Muñoz RF. Using evidence-based internet interventions to reduce health disparities worldwide. J Med Internet Res 2010;12(5):e60 [FREE Full text] [doi: 10.2196/jmir.1463] [Medline: 21169162]
24. Hwang EH, Guo X, Tan Y, Dang Y. Delivering healthcare through teleconsultations: implications for offline healthcare disparity. Inf Syst Res 2022;33(2):515-539. [doi: 10.1287/isre.2021.1055]
25. Goh JM, Gao GG, Agarwal R. The creation of social value: can an online health community reduce rural-urban health disparities? MIS Q 2016;40(1):247-263.
26. Gromisch ES, Turner AP, Haselkorn JK, Lo AC, Agresta T. Mobile health (mHealth) usage, barriers, and technological considerations in persons with multiple sclerosis: a literature review. JAMIA Open 2021;4(3):ooaa067 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa067] [Medline: 34514349]

XSL•FO
RenderX

27. The general office of the state council opinions on promoting the development of "internet + medical health".: General Office of the State Council; 2018. URL: https://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm [accessed 2024-10-11]
28. Ayabakan S, Bardhan IR, Zheng Z. Impact of telehealth and process virtualization on healthcare utilization. Inf Syst Res 2023;35(1):45-65. [doi: 10.1287/isre.2023.1220]
29. Blackwell DL, Martinez ME, Gentleman JF, Sanmartin C, Berthelot J. Socioeconomic status and utilization of health care services in Canada and the United States: findings from a binational health survey. Med Care 2009;47(11):1136-1146. [doi: 10.1097/MLR.0b013e3181adcbe9] [Medline: 19786920]
30. Bavafa H, Hitt LM, Terwiesch C. The impact of E-Visits on visit frequencies and patient health: evidence from primary care. Manage Sci 2018;64(12):5461-5480 [FREE Full text] [doi: 10.1287/mnsc.2017.2900] [Medline: 33033417]
31. Truong M, Yeganeh L, Cook O, Crawford K, Wong P, Allen J. Using telehealth consultations for healthcare provision to patients from non-Indigenous racial/ethnic minorities: a systematic review. J Am Med Inform Assoc 2022;29(5):970-982 [FREE Full text] [doi: 10.1093/jamia/ocac015] [Medline: 35150266]
32. Bo Y, Liu QB, Tong Y. The effects of adopting mobile health and fitness apps on hospital visits: quasi-experimental study. J Med Internet Res 2023;25:e45681 [FREE Full text] [doi: 10.2196/45681] [Medline: 37505809]
33. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. J Econ Surv 2008;22(1):31-72. [doi: 10.1111/j.1467-6419.2007.00527.x]
34. Qiao D, Lee SY, Whinston AB, Wei Q. Financial incentives dampen altruism in online prosocial contributions: a study of online reviews. Inf Syst Res 2020;31(4):1030-1492. [doi: 10.1287/isre.2020.0949]
35. Venkatesh V, Sykes TA, Zhang X. ICT for development in rural India: a longitudinal study of women's health outcomes. MIS Q 2020;44(2):605-629. [doi: 10.25300/misq/2020/12342]
36. McFarland DC, Ornstein KA, Holcombe RF. Demographic factors and hospital size predict patient satisfaction variance--implications for hospital value-based purchasing. J Hosp Med 2015;10(8):503-509 [FREE Full text] [doi: 10.1002/jhm.2371] [Medline: 25940305]
37. Person SD, Allison JJ, Kiefe CI, Weaver MT, Williams OD, Centor RM, et al. Nurse staffing and mortality for Medicare patients with acute myocardial infarction. Med Care 2004;42(1):4-12. [doi: 10.1097/01.mlr.0000102369.67404.b0] [Medline: 14713734]
38. Savira F, Orellana L, Hensher M, Gao L, Sanigorski A, Mc Namara K, et al. Use of General practitioner telehealth services during the COVID-19 pandemic in regional victoria, Australia: retrospective analysis. J Med Internet Res 2023;25:e39384 [FREE Full text] [doi: 10.2196/39384] [Medline: 36649230]

Abbreviations

DiD: difference-in-differences
DV: dependent variable
PSM: propensity score matching

Edited by A Coristine; submitted 26.02.24; peer-reviewed by Z Sijia, L Song; comments to author 06.05.24; revised version received 29.06.24; accepted 17.09.24; published 11.11.24.

Please cite as:
Liu Y, Jin H, Yu Z, Tong Y
Impact of Internet Hospital Consultations on Outpatient Visits and Expenses: Quasi-Experimental Study
J Med Internet Res 2024;26:e57609
URL: https://www.jmir.org/2024/1/e57609
doi:10.2196/57609
PMID:

©Yayuan Liu, Haofeng Jin, Zhuoyuan Yu, Yu Tong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Barriers and Facilitators Associated With Remote Concussion Physical Assessments From the Perspectives of Clinicians and People Living With Workplace Concussions: Focus Group Study

Keely Barnes[1,2,3], MHK; Heidi Sveistrup[1,2,4,5], PhD; Motahareh Karimijashni[1,3], MSc; Mark Bayley[6,7], MD; Mary Egan[1,2], PhD; Martin Bilodeau[1,2,4], PhD; Michel Rathbone[8], MD, PhD; Monica Taljaard[3,9], PhD; Shawn Marshall[2,3,10], MSc, MD

[1]School of Rehabilitation Sciences, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada
[2]Bruyère Research Institute, Ottawa, ON, Canada
[3]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[4]School of Human Kinetics, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada
[5]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
[6]KITE Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada
[7]Division of Physical Medicine and Rehabilitation, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[8]Department of Medicine, Division of Neurology, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[9]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[10]Department of Medicine, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:
Keely Barnes, MHK
School of Rehabilitation Sciences
Faculty of Health Sciences
University of Ottawa
200 Lees Ave.
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6136126127
Email: kbarn076@uottawa.ca

Abstract

Background: Evaluating the clinical status of concussions using virtual platforms has become increasingly common. While virtual approaches to care are useful, there is limited information regarding the barriers and facilitators associated with a virtual concussion assessment.

Objective: This study aims to identify the barriers and facilitators associated with engaging in virtual concussion assessments from the perspective of people living with workplace concussions; identify the barriers and facilitators to completing virtual concussion assessments from the perspectives of clinicians; and identify the clinical measures related to 4 clinical domains that would be most appropriate in virtual practice: general neurological examination and vestibular, oculomotor, and cervical spine assessment. We also evaluated effort.

Methods: Separate online focus groups were conducted with expert concussion clinicians and people living with workplace concussions. A moderator led the focus groups using a semistructured interview guide that targeted a discussion of participants' experiences with virtual assessments. The discussions were recorded, transcribed, and analyzed by 2 reviewers using content analysis. Barriers and facilitators associated with completing the physical concussion examination were categorized based on the domain of the concussion examination and more general barriers and facilitators. Clinician-selected measures believed to work best in a virtual practice were described using frequency counts.

Results: A total of 4 focus groups with 15 people living with workplace concussions and 3 focus groups with 14 clinicians were completed using Microsoft Teams. Barriers were identified, such as triggering of symptoms associated with completing an assessment over video (mentioned 13/162 (8%) and 9/201 (4%) of the time for patient and clinician participants, respectively); challenges with location and setup (mentioned 16/162 (10%) of the time for patient participants); communication (mentioned

XSL•FO
RenderX

34/162 (21%) and 9/201 (4%) of the time for patient and clinician participants, respectively); and safety concerns (mentioned 11/162 (7%) of the time for patient and 15/201 (7%) for clinician participants). Facilitators were identified, such as having access to support (mentioned 42/154 (27%) and 21/151 (14%) of the time for patient and clinician participants, respectively); implementing symptom management strategies throughout the assessment (mentioned 11/154 (7%) of the time for patient participants); and having access to resources (mentioned 25/151 (17%) of the time for clinician participants). From the perspective of the clinician participants included in this study, the clinical measures recommended most for a virtual practice were finger to nose testing; balance testing; the Vestibular/Ocular Motor Screening tool; saccades; and cervical spine range of motion within their respective domains (ie, neurological examination, vestibular, oculomotor, and cervical spine assessment).

Conclusions: Virtual assessments appear to be useful for both people living with workplace concussions and clinicians. While barriers were identified, such as challenges associated with exposure to screens, virtual assessments have benefits such as improved access to care. The clinician-selected measures that were considered best in a virtual practice will be investigated in an upcoming evaluative study.

International Registered Report Identifier (IRRID): RR2-10.2196/40446

(J Med Internet Res 2024;26:e56158) doi:10.2196/56158

KEYWORDS

remote care; mild traumatic brain injury; telehealth; assessment; workplace injury; concussion; telemedicine; brain injury

Introduction

Background

Concussions result from indirect or direct impacts to the head and usually lead to neurological symptoms [1]. While most individuals who sustain a concussion recover within days of the injury, an estimated 10% to 20% of individuals experience persistent symptoms after the injury requiring intervention [2]. Following a concussion sustained in the workplace, there is an increased likelihood of reporting persistent symptoms [3,4]. This may be attributed to several factors, including access to worker's compensation, fear of returning to the site of the injury [3], and the demands of work that can exacerbate symptoms. Timely access to care can be challenging for workers in remote and rural communities due to poor access to specialized clinicians [5-8]. The COVID-19 pandemic further presented challenges to accessing in-person care for workplace injuries [9,10].

Virtual care, the remote delivery of health care services that may address some of these challenges, often uses visual and audio communication or videoconferencing technology [11,12]. For patients with a concussion, virtual care could improve access to care [7,13-16] and has been shown to be beneficial from the perspective of clinicians, although not without limitations [17]. A recent study by van Ierssel et al [17] explored the perspectives of 25 clinicians on assessing and managing concussions through virtual care approaches. They reported that clinicians believed that virtual care could be beneficial for monitoring the functional status at follow-up, for screening patients, and for identifying those who may require further in-person assessment. Moreover, participants did not believe care to be jeopardized by virtual follow-up assessments. In general, the literature supports the view that clinicians recognize advantages to using virtual care approaches for concussion assessment and management, but there is a need to further explore the effectiveness of virtual assessment in the context of concussion care.

There is limited information on the barriers and facilitators associated with virtual care from the perspectives of people with a concussion. One study [18] explored virtual care experiences from the perspectives of patients, caregivers, and health care providers in the geriatric medicine setting. Uncertainty about accuracy (ie, the ability to validly identify deficits) was reported as a concern with patients and caregivers, who questioned whether a comprehensive understanding of patient status was captured [18]. However, patient and caregiver participants in this study highlighted the benefits, such as improved access to care and engagement with caregivers in appointments, emphasizing that a hybrid approach to care may be superior to virtual-only approaches [18]. However, with regard to concussion care, experiences may differ from those in the geriatric medicine setting. People with concussion may experience greater fatigue when compared to other populations due to screen exposure [19]. Furthermore, most concussions occur in younger people (<30 years) when compared to the geriatric population [20], who are potentially more comfortable and less intimidated by the technology involved [21]. Both the clinician and patient perspectives are important for a comprehensive understanding of the potential usefulness of the technology [11], and therefore, a detailed exploration of experiences associated with virtual assessments from the perspectives of people living with workplace concussion is needed.

There is a critical need to identify the barriers and facilitators associated with the main domains of the physical examination (ie, neurological examination, vestibular, oculomotor, and cervical spine), when completing a hands-on comprehensive assessment is not possible [22-24]. There is an additional need to evaluate the effort used during an assessment, which is particularly the case in the context of injured workers, as the presence of and access to compensation following a work-related injury could influence performance (if low effort is used) and subsequent findings on an examination [25]. While the evaluation of effort is not a clinical domain, it is considered a component of the examination used to interpret the clinical results and is therefore presented as a separate "domain" throughout this paper.

XSL•FO
RenderX

Objectives

This study aimed to (1) explore the barriers and facilitators related to engaging in virtual concussion assessments from the perspective of people living with workplace concussions; (2) document the barriers and facilitators associated with completing a virtual concussion examination (ie, neurological examination, vestibular, oculomotor, and cervical spine domains and effort) as identified by concussion expert clinicians with experience in completing virtual assessments; and (3) document concussion expert clinicians' opinions regarding which clinical measures from the aforementioned domains work best in a virtual practice. This study focused on the assessment of physical domains due to substantial adaptations required to administer these measures virtually and due to individual interpretations of how best to administer these when compared with measurements of emotional or cognitive health.

Methods

Overview

In this paper, we report on a part of a mixed methods study. The full protocol [26] and initial findings from surveys and a working group to identify targeted clinical measures [27] have been recently published. This paper reports on focus group discussions related to the previously identified clinical measures. This paper follows the guidelines outlined by the Standards for Reporting Qualitative Research (Multimedia Appendix 1).

Ethical Considerations

Research ethics approval was obtained from the Ottawa Health Sciences Network Research Ethics Board (20210575-01H), the University of Ottawa Board of Ethics (H-02-22-7611), and the Bruyère Research Ethics Board (M16-22-006). Verbal consent over the telephone was obtained from all participants before participation in the focus groups. Privacy and confidentiality were maintained by using deidentified data. Participants were provided with a CAD $30 (US $22) gift card for their participation in the focus group.

Participants and Recruitment

Patient participants were recruited through the Ontario Workers Network clinic at the Ottawa Hospital. Eligible patient participants were adults (aged >18 years) who had attended at least one initial or follow-up virtual assessment following a workplace concussion injury through a videoconferencing platform. Purposive sampling was used to identify these eligible patient participants [28]. The sampling strategy was designed to ensure adequate distribution of specific characteristics such as patient age and sex. Eligible participants were approached by telephone and were provided with a description of the aims of the study before consenting. We attempted to recruit at least 5 participants per group [29,30]; however, some participants did not attend the scheduled group or had to cancel, so the groups ranged from 2 to 5 participants. We attempted to include a diverse group of participants in terms of age, date since injury, and sex.

Eligible clinician participants were Canadian-practicing clinicians (ie, neurologists, physiatrists, family medicine physicians, sports medicine physicians, physiotherapists, athletic therapists, and kinesiologists) with expertise in treating concussions (not specifically workplace concussion) and with experience in completing virtual concussion assessments. Eligible clinician participants were recruited through targeted emails and outreach to rehabilitation clinics across Canada. Purposive sampling was used [26] to ensure the distribution of certain characteristics, such as clinician profession. Similar to the patient participant groups, we attempted to recruit 5 to 7 clinicians per group; however, due to scheduling challenges, the groups ranged from 4 to 5 participants. We attempted to include a diverse group of clinicians in terms of clinical profession.

Data Collection

In a previous paper [24], we surveyed clinicians concerning their use of clinical measures in their in-person practice. The clinicians rank-ordered the measures based on relevant measures to their in-person practice within the following domains: neurological examination, vestibular, oculomotor, cervical spine, and effort assessment. We then documented the perceived feasibility of administering these measures virtually, based on expert clinician opinion [27]. The measures identified in this work were further discussed in this focus group study.

Focus groups were completed over a videoconferencing platform (Microsoft Teams) from August 2022 to January 2023. Consenting participants attended one focus group (either a clinician group or a patient group). Semistructured guides (refer to the study by Barnes et al [26]) were used to direct the discussion for both clinician and patient focus groups. The guides were reviewed and revised by the study team before the commencement of the focus groups to confirm the appropriate wording of the questions and ensure the questions captured the aspects that the team was aiming to explore. The patient interview guide consisted of 9 questions regarding practical and technical barriers and facilitators to undergoing a virtual assessment and sought feedback and recommendations for virtual assessments. The clinician interview guide included 12 questions related to (1) practical and technical barriers and facilitators to virtual assessments and (2) domain-specific questions related to the neurological examination, vestibular, oculomotor, and cervical spine domains and effort assessment, as well as those prompting the participants to discuss and select the outcome measures [24,25] within each clinical domain that they felt would work best in a virtual practice.

Focus groups were moderated by KB, a female registered kinesiologist. KB held a master's degree in human kinetics and was a PhD candidate in rehabilitation sciences at the time of the study. KB worked as a kinesiologist for 2 years both in a complex care clinic where Workplace Safety and Insurance Board concussion-injured workers were assessed and treated as well as in home-based kinesiology services. A notetaker (MK) was present to prevent loss of data, record nonverbal expressions and key discussion points to add further context to the recording, and aid with technical issues. The focus groups were audio-video recorded and transcribed verbatim by KB.

All participants completed a demographic form following the completion of the focus group. Questions related to age, sex,

XSL•FO
RenderX

work status, virtual assessments completed after the concussion, and symptoms were completed for patient participants. Questions related to age, sex, clinical profession, clinical experience, volume of practice of patients with a concussion, virtual assessments completed in the last year, and self-perceived competency associated with completing the in-person and virtual assessment were completed for clinician participants.

Data Analysis

Clinician and Patient Participant Identification of Barriers and Facilitators

Two reviewers (KB and MK) used inductive qualitative content analysis [31] in NVivo qualitative data analysis software (version 12; QSR International Inc). Analysis started simultaneously with data collection so that data saturation (the point at which no new themes were identified) could be monitored. The analytic process is presented in Figure 1 [29,32-34].

To ensure that inferences being derived from the data were valid, the coders analyzed the data using the same coding scheme (consisting of codes to be used by the reviewers based on an initial review of the transcripts) [29,32]. They met weekly to peer review data analyses and discuss interpretations until reaching consensus. Frequency counts were generated for each theme by tabulating the number of times each theme was identified in the transcripts, and the percentage of times each theme was reported by the participants was presented.

Figure 1. Content analysis process flow diagram reflecting the 7 steps that led to the development of results.

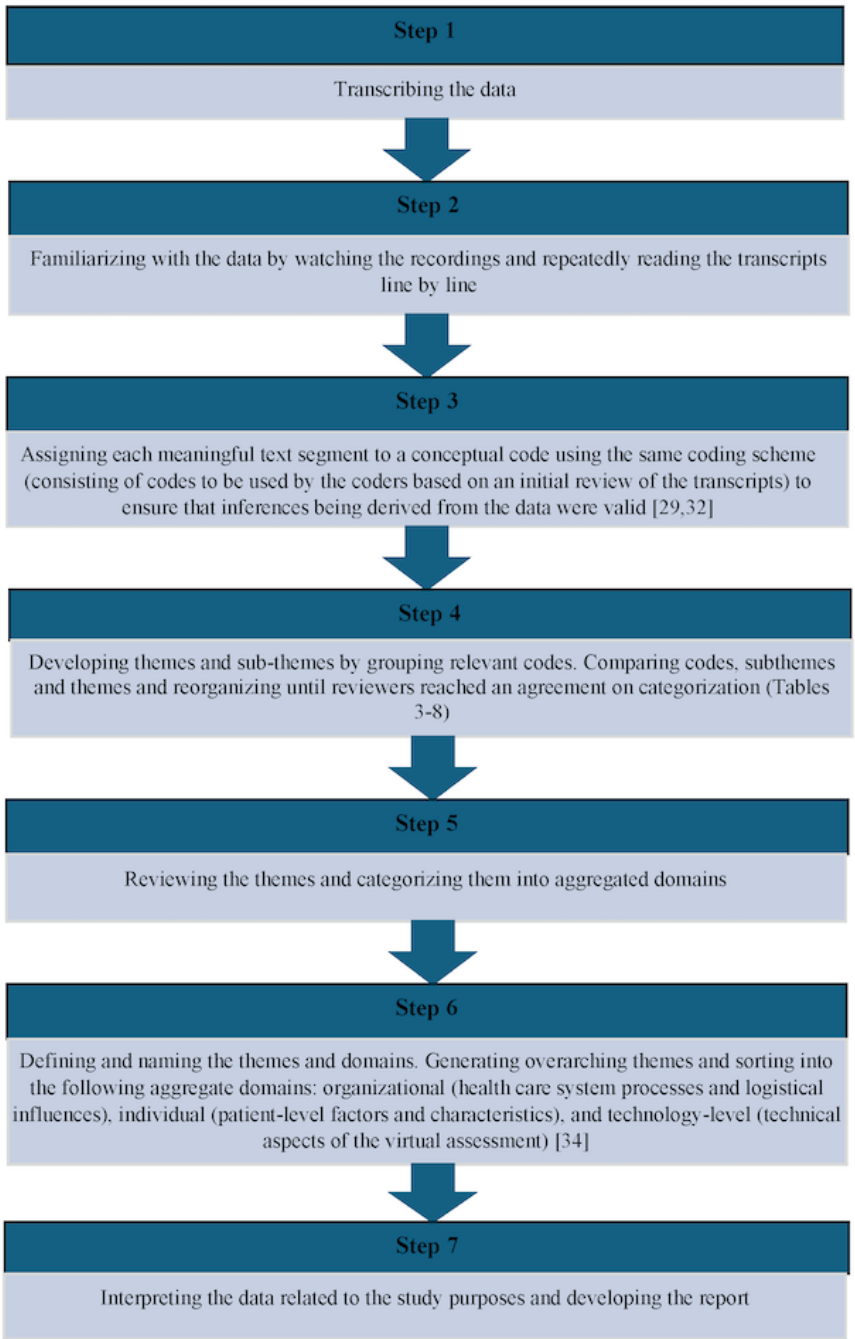

XSL•FO
RenderX

The clinician subthemes were further categorized into the predefined categories (domain-specific and general categories). Domain specific was defined as any barrier or facilitator that was discussed in relation to specific measures under the following domains: neurological examination, vestibular, oculomotor, cervical spine, and effort. General was defined as any barrier or facilitator that relates to the virtual assessment not specific to the domains. The patient participant subthemes were all categorized as general barriers and facilitators. NVivo was used to calculate Cohen κ coefficients to determine the agreement between the 2 coders for all the codes.

Clinician Participant Selection of Clinical Measures

Prereading materials were provided, to the clinician participants only, before the focus group. These materials contained descriptions of the clinical measures identified by clinicians in an earlier study [27] to be discussed in the focus groups, YouTube links with video demonstrations of the measures, and a summary of select psychometric properties of each measure. Details of these materials are described elsewhere [27]. From this list, the clinicians selected the ones that they perceived would provide the most information when administered virtually. The 2 reviewers independently documented the frequency of the clinician selection of each clinical measure.

Results

Focus Group Structure

Overall, 7 focus groups (4 with people living with workplace concussions and 3 with clinicians) were completed. Saturation was reached after 3 groups for both the patient and clinician groups; however, as we experienced challenges recruiting male participants, we conducted one additional male-only focus group. This allowed us to have 2 mixed-sex focus groups, a female-only focus group, and a male-only focus group. Focus groups were made up of 2 to 5 participants. A total of 15 people with workplace concussions and 14 clinicians participated in the focus groups. The focus group discussion lasted between 30 and 90 minutes each as some groups contained only 2 participants and some groups contained up to 5. Cohen κ was calculated as 0.80 across all groups (considered substantial agreement) for all codes for both the clinician and patient participant focus groups.

Participants

Patient Participants

Table 1 presents the demographic information of the patient participants. The age range of the patient participants was between 20 and 59 (mean 40.9, SD 13.1) years. Most participants were female (11/15, 73%). All participants were experiencing persistent symptoms after workplace concussion injury at the time of their virtual assessment. Injury dates upon participation in the focus group ranged from 9 months to approximately 4 years after a concussion. All participants reported physical symptoms as most common (15/15, 100%), followed by cognitive (6/15, 40%) and emotional symptoms (2/15, 13%).

XSL•FO
RenderX

Table 1. Patient participant demographic characteristics (n=15).

Characteristics	Participants, n (%)
Age range (y)	
20-29	3 (20)
30-39	4 (27)
40-49	5 (33)
50-59	3 (20)
>60	0 (0)
Sex	
Female	11 (73)
Male	4 (27)
Current work status	
Off work	2 (13)
Full return to work	5 (33)
Modified return to work, same occupation	6 (40)
Modified return to work, different occupation	2 (13)
Virtual assessments completed after the concussion	
1-5	7 (47)
6-10	4 (27)
>10	4 (27)
Currently experienced symptoms	
Somatic symptoms (headaches, tinnitus, vision blurriness, fatigue, nausea, migraines, and dizziness)	15 (100)
Emotional symptoms (overstimulation, anxiety, anger, and frustration)	2 (13)
Cognitive symptoms (slow thinking and expression of thoughts, memory, issues processing information, and inability to concentrate)	6 (40)

Clinician Participants

Table 2 presents the demographic information for the clinician participants. Each focus group consisted of a mix of clinicians. A total of 8 physicians (ie, sports medicine physicians, physiatrists, and a family medicine physician) and 6 other clinicians (ie, athletic therapists and physiotherapists) participated. The clinicians' virtual assessment experience ranged from ≤10 virtual assessments completed in the last year to over 20 completed virtual assessments completed in the last year. Most clinicians self-reported their competency of completing the in-person and virtual assessment as "strongly competent" and "competent," respectively.

XSL•FO
RenderX

Table 2. Clinician participant demographic characteristics (n=14).

Characteristics	Participants, n (%)
Age range (y)	
20-29	1 (7)
30-39	5 (36)
40-49	2 (14)
50-59	3 (21)
>60	3 (21)
Sex	
Female	6 (43)
Male	8 (57)
Clinical profession	
Physiotherapist	4 (29)
Physiatrist	5 (36)
Athletic therapist	2 (14)
Sports medicine physician	2 (14)
Family medicine physician	1 (7)
Clinical experience (y)	
1-5	3 (21)
5-10	4 (29)
>10	7 (50)
Current volume of practice of patients with a concussion	
0-25	5 (36)
25-50	2 (14)
>50	6 (43)
Did not respond	1 (7)
Virtual assessments completed in the last year	
0-10	3 (21)
10-20	1 (7)
>20	7 (50)
Unsure	3 (21)
Competency of in-person neurological assessments	
Strongly incompetent	0 (0)
Incompetent	0 (0)
Neutral	1 (7)
Competent	4 (29)
Strongly competent	9 (64)
Competency of virtual neurological assessments	
Strongly incompetent	0 (0)
Incompetent	1 (7)
Neutral	3 (21)
Competent	8 (57)
Strongly competent	2 (14)

XSL•FO
RenderX

Themes

Barriers

Overview

Tables 3 and 4 present the barriers for both the patient and clinician participants. Quotes are represented as numbers in the text within each aggregate category. For example, "P12-Q1" refers to quote number 1 from Patient participant 12 or "C2-Q7" refers to quote number 7 from Clinician participant 2. For the clinicians only, categorization of the subthemes (domain specific or general) are presented. Percentage of times each theme was reported by the participants is presented. This was calculated by dividing the frequency of times a specific theme was mentioned by participants by the total number of times all themes were mentioned by participants. For example, location and setup at home was identified as a barrier by patient participants (ie, due to limited space, noise at home, being in a different province, etc) 9.9% (16/162) of the time. Clinical process–related factors were identified as a barrier by clinician participants 10.5% (21/201) of the time. Rankings associated with the barriers reported by the patient and clinician participants are presented in Multimedia Appendix 2.

XSL•FO
RenderX

Table 3. Patient and clinician participant barriers associated with virtual concussion assessments.

Aggregated domains, overarching themes, and subthemes	Patients		Clinicians		
	Identified by the patient participant	Patients reported (out of a total of 162 themes mentioned), n (%)	Identified by the clinician participant	Clinicians reported (out of a total of 201 themes mentioned), n (%)	Clinician categories
Organizational level					
Environmental setup					
Location and setup at home	✓	16 (9.9)	__[a]	—	—
Clinical process–related factors		—		21 (10.5)	
Lack administration support	—	—	✓		General
Time	—	—	✓		Domain specific: cervical; neurological
Individual level					
Communication and engagement		34 (21)		9 (4.5)	
Challenges with communication	✓		✓		Domain specific: neurological
Challenges building rapport and connecting with clinicians	✓		✓		General
Triggering physical and emotional symptoms		13 (8)		9 (4.5)	
Screens trigger physical symptoms	✓		✓		General
Emotional	✓		—	—	—
Safety		11 (6.8)		15 (7.5)	
Unease working to limit ability	✓		—	—	—
Feelings of tests being safer in person	✓		—	—	—
Safety	—	—	✓		Domain specific: vestibular; neurological
Comfort				9 (4.5)	
Patient comfort and preference	—	—	✓		Domain specific: effort
Psychological impact	—	—	✓		General
Environmental setup					
Environmental and patient setup	—	—	✓	16 (8)	Domain specific: vestibular; oculomotor; neurological; effort
Technology level					
Technology and internet		32 (19.7)		41 (20.4)	
Health care system barriers	✓		—	—	—
Issues or lack comfort with technology and the internet	✓		—	—	—

XSL•FO
RenderX

Aggregated domains, overarching themes, and subthemes	Patients		Clinicians		
	Identified by the patient participant	Patients reported (out of a total of 162 themes mentioned), n (%)	Identified by the clinician participant	Clinicians reported (out of a total of 201 themes mentioned), n (%)	Clinician categories
Format of required documents	✓		—	—	—
Technical and internet issues	—	—	✓		Domain specific: oculomotor; neurological
Camera and device limitations	—	—	✓		Domain specific: vestibular; oculomotor; neurological
Accuracy and completeness of physical examination		56 (34.6)		53 (26.4)	
Incomplete evaluation or lack accuracy	✓		—	—	—
Lacking physical contact	✓		—	—	—
Unable to complete assessments in entirety	—	—	✓		Domain specific: vestibular; oculomotor; cervical; neurological
Lack accuracy	—	—	✓		Domain specific: vestibular; oculomotor; cervical; neurological
Unable to assess certain conditions and subtleties	—	—	✓		Domain specific: vestibular; oculomotor; cervical; neurological
Visibility					
Observability	—	—	✓	28 (13.9)	Domain specific: vestibular; oculomotor; cervical; neurological

[a]Not identified by participants.

XSL•FO
RenderX

Table 4. Patient and clinician participant barriers quotes associated with virtual concussion assessments.

Aggregated domains, overarching themes, and subthemes	Quotes
Organizational level	
Environmental setup	
Location and setup at home	• "A lot of the times they couldn't do my assessment virtually because I lived in a different province" [P12[a]-Q1[b]]
Clinical process–related factors	• "But think about the time that it takes to explain to somebody, try to see what it is that you're doing. You know it's, it takes far more time than seeing them in person." [C2[c]-Q7]b
Individual level	
Communication and engagement	• "Harder to explain um what was going on. I also have speech issues so trying to express sometimes over virtual is difficult." [P6-Q4] • "I found the virtual assessment created a bit of a barrier on creating relationships, though, and you weren't able to as easily connect or especially if it is your first time seeing them, athlete or patient." [C3-Q11]
Triggering physical and emotional symptoms	• "My big, bothersome side effects was ah like technology. Just the audio like the kind of the sound that comes out of the audio really irritated me. And with the visual like the screen that would really drive my eyes and my head crazy, so it wasn't really convenient for, to be in front of a screen trying to recover from a concussion." [P14-Q3] • "Depending on how symptomatic they are, screen time isn't great for them. So, um you know, decreased tolerance to screen time is one of the things and you might not be able to make it through a full assessment." [C6-Q10]
Safety	• "Once COVID hit, we were all like on video with the clinic and it was a little bit harder because I've fallen a lot. Umm. I'm in a walker now. A walker. Um. And sometimes there aren't people here at the house, so if I do fall I, it's a little scarier in a way." [P1-Q2] • "If they had really poor balance, you're not there to guard them as a physio. We're always kind of told to guard when you're doing that kind of stuff and they have nobody else there to help guard then that's a problem." [C9-Q8]
Comfort	• "When you think about effort, it's like, well, part of it is that the individual has to be comfortable, right, in the virtual assessment... I wouldn't draw any conclusions around effort from a virtual assessment." [C4-Q12]
Environmental setup	• "A barrier for gait and tandem gait is kind of having enough space where you can see them walking on the screen too. So probably back to that camera umm set up and everything and then making sure there's enough space in their house so they can actually do a long enough walk that you can see any abnormalities too." [C6-Q9]
Technology level	
Technology and internet	• "The Internet stuff because I have a hard time doing computers right now and due to vision loss and stuff and I just forget how to do it. So, um that's something that's, if I didn't have somebody here, I wouldn't have been, I wouldn't have been able to do it." [P3-Q6] • "I find the eye tests are the most challenging and mainly it's you know, it's whose finger to follow is 1 but also the resolution especially if depending on the quality of their camera it's the hardest and the closer they get to the screen of course the more blurry it becomes to them also, so it's really hard." [C10-Q15]
Accuracy and completeness of physical examination	• "I have only had one in person um assessment with both OT and the doctor. And it was like very hands on compared to any of the other assessments before. The other assessments like that were online kind of just felt like we were going through a checklist to make sure like you know, nothing's like terrible condition, but the one in person, for sure, that felt longer and more in depth." [P8-Q5] • "If you're gonna do this, you're gonna have them walk by the camera and you sort of say, ok, now aim the camera down at your feet as you're walking or that, go back far enough there just really isn't enough space for that, camera is not wide enough and whatnot or goes low enough to really get a good assessment. You can get a generally gestalt of it, but I don't think it's an accurate assessment." [C1-Q13]
Visibility	• "It's the visibility of, you know, trying to count errors on camera." [C2-Q14]

XSL•FO
RenderX

[a]P: patient participant.

[b]Q: quote.

[c]C: clinician participant.

Patient Participant Barriers

Organizational Level

Patient participants reported some issues with being in a province different from that of the assessing clinician. For example, some professional associations require clinicians to be licensed to practice in the province in which the patient is located for the virtual assessment (the patient may be located in Quebec and the clinician in Ontario). Participants further voiced concerns regarding privacy associated with using a videoconferencing platform such as the fear of people overhearing conversations and concerns regarding the security of the videoconferencing platform (Q1).

Individual Level

Patient participants expressed safety concerns (Q2). Participants believed that certain components of the assessment would feel safer when done in person, specifically the balance tests. They believed this would have an impact on the results; participants felt that they could "perform better" and "push out of their comfort zone" with the in-person assessments when compared to the virtual assessments.

Patient participants felt that virtual assessments could activate physical and emotional symptoms. The use of screens was commonly reported to trigger physical concussion symptoms (headaches, dizziness, vision issues, etc; Q3). Feelings of isolation and depression could be increased during virtual assessment: "you do feel very isolated" and "you are all alone and it gets very depressing."

Patient participants also reported challenges communicating issues related to their condition and challenges building rapport with their clinicians through videoconferencing platforms (Q4).

Technology Level

A barrier identified included the perspective that less accurate information is obtained with the virtual assessment and the perspective that the virtual assessment is incomplete. The participants perceived that clinicians were unable to see subtle issues associated with their condition, such as "eye twitches" or "tremors." Participants perceived the in-person examination to feel more in-depth and hands-on, whereas the virtual assessment felt unfinished (Q5). Furthermore, participants expressed a lack of physical contact with the virtual assessment, which seemed to be particularly noticeable for the cervical spine examination.

Participants reported issues with the internet connection and with the videoconferencing platforms used for the virtual assessment. Different clinicians using different platforms, videoconferencing platforms that are difficult to use, and the format of the required documents, such as needing to complete web-based questionnaires, were identified as challenges to completing the virtual assessment (Q6).

Clinician Participant Barriers

Organization Level

Lack of time (due to the additional time required to set the patient up correctly on the screen in a virtual appointment) and lengthy clinical measures were identified as domain-specific barriers (Q7). Challenges due to the lack of administrative support to aid with the virtual assessment process was identified as a barrier, which also relates to the processes in the clinical setting.

Individual Level

Safety concerns were expressed with all balance measures and components of the neurological examination (eg, pronator drift due to a need for the patient to close their eyes during the test and functional squat) by clinician participants (Q8).

Environmental and patient setup was identified as a domain-specific barrier. Specifically, clinicians reported challenges with the patients' environment, such as lacking space, and challenges getting patients in appropriate positions to complete clinical measures (Q9).

Clinicians reported that computer screens can trigger physical concussion symptoms, which makes it challenging to complete a comprehensive assessment virtually, representing a general barrier. Clinicians expressed that patients have minimal tolerance to screens, which leads to a need to take frequent breaks (Q10).

Clinicians identified challenges with communicating with patients through screens, including challenges with explaining how to complete a measure virtually and challenges associated with building rapport and forming connections with patients, reflecting a general barrier (Q11).

A lack of patient comfort with the virtual assessment and clinician stress or frustration associated with completing the virtual assessment due to technical challenges were identified as barriers. The importance of patient comfort was particularly relevant to the discussion regarding the evaluation of effort (observing if the patient is performing at optimal capacity) in a virtual assessment. Clinicians reported that due to the lack of comfort and confounding factors (ie, fatigue, symptom aggravation associated with the screen, and unclear picture), it would be very challenging to draw conclusions regarding the effort used in a virtual assessment (Q12).

Technology Level

A domain-specific barrier identified by clinician participants included an inability to complete the physical examination in its entirety. For example, clinicians reported an inability to complete a full cranial nerve examination (a part of the neurological examination) virtually and an inability to touch a patient's neck when completing cervical spine palpation. Clinicians also questioned accuracy for certain measures (such as the modified Balance Error Scoring System and oculomotor tests) when administered virtually (Q13). Furthermore, clinicians

XSL•FO
RenderX

expressed challenges associated with identifying subtle deficits experienced by patients and an inability to assess comorbid conditions. Most of these were limitations identified due to the quality of current technology (clarity of the image on the screen).

Clinicians identified an inability to get an understanding of the full picture, such as missing body language cues due to only being able to see a patient's upper body. Camera limitations were discussed (Q14).

Camera, device, and internet issues were identified as general barriers to completion of the virtual assessment. Unstable internet, the use of different devices (particularly mobile phones), and positioning of the camera were all barriers identified that could influence the quality of the virtual assessment (Q15).

Facilitators

Tables 5 and 6 present the facilitators for both the patient and clinician participants. Rankings associated with the facilitators reported by the patient and clinician participants are presented in Multimedia Appendix 3. A figure presentation of the overarching themes, subthemes, and codes is presented in Multimedia Appendix 4 (patient participants) and Multimedia Appendix 5 (clinician participants). Definitions of the subthemes associated with Tables 3 and 4 are presented in Multimedia Appendices 6 and 7, respectively.

Tables 7 and 8 present the clinician- and patient participant–identified benefits associated with the virtual concussion assessment. While not necessarily facilitators to the use of virtual assessments, the participants identified some benefits to this approach. The information is presented in the same format as has been for the barriers and facilitators, as described earlier.

XSL•FO
RenderX

Table 5. Patient- and clinician-participant facilitators associated with virtual concussion assessment.

Aggregated domains, overarching themes, subthemes	Patients		Clinicians		
	Identified by patient participants	Patients reported (out of a total of 154 themes mentioned), n (%)	Identified by clinician participants	Clinicians reported (out of a total of 151 themes mentioned), n (%)	Clinician categories
Organizational level					
Format of delivered care and material		23 (14.9)		29 (19.2)	
Use of functions available on the virtual platform	✓		__[a]	—	—
Introduction to technology and resources	✓		✓		Domain specific: vestibular; oculomotor; cervical; neurological
Integration and patient selection of virtual	✓		✓		General
Support		42 (27.3)		21 (13.9)	
Opportunity to involve more care team members	✓		—	—	—
Easy contact with clinician or technical support	✓		—	—	—
Support	—	—	✓		Domain specific: vestibular; oculomotor; neurological
Use of resources					
Use of resources to complete assessment	—	—	✓	25 (16.6)	Domain specific: vestibular; oculomotor; cervical; neurological
Individual level					
Symptom management					
Offering symptom management strategies	✓	11 (7.1)	—	—	—
Support		42 (27.3)		21 (13.9)	
Home support	✓		—	—	—
Support	—	—	✓		Domain specific: vestibular; oculomotor; neurological
Environmental setup					
Environmental and patient setup	—	—	✓	16 (10.6)	Domain specific: vestibular; oculomotor; neurological; effort
Technology level					
Use of measure to identify gross deficits					
Use of measure as global screen and rely on subjective information	—	—	✓	24 (15.9)	Domain specific: vestibular; oculomotor; cervical; neurological

[a]Not identified by participants.

XSL•FO
RenderX

Table 6. Patient- and clinician-participant facilitators quotes associated with virtual concussion assessment.

Aggregated domains, overarching themes, subthemes	Quotes
Organizational level	
Format of delivered care and material	• One thing I would suggest is, for things like the neurology appointment, I know I think it would be good if someone could just go um be comfortable just for your appointment, be comfortable. We might need to do this, this, because then, for me, it sounds like for you know, we wanna appreciate the person's time and also like I'm setting up and making sure things are like, you know, whatever, like, please if I need to do this so having knowing kind of what your spatial requirements are gonna be might be a good idea. Just so that it can help us." [P2[a]-Q17[b]] • "I find that a lot of times the patient gets on to the appointments virtually and they have this sense of anxiety because they don't know what they're going to go through in the assessment. They don't know how long it's going to last. They don't know how taxing is going to be for those who are severely concussed it can be incredibly taxing. So, I find just having a basic outline that we'll be doing some history taking some brief history taking some measurements range of motion and that sort of thing, Balance testing, whatever it is we're going to be testing just to give them an advance." [C5[c]-Q22]
Support	• "But I do find that when you are having technical difficulties, that calling them and they'll send you a new link or whatever, or they'll even call you on the phone umm to do your appointments just by voice, which helps me a lot as well." [P3-Q16] • "I've done this where an occupational therapist helped me with the eye exam. It was so much. I felt so much more confident. In what we were testing." [C11-Q24]
Use of resources	• "The only way you could do this would be with some sort of app, that sort of, uh, can detect that that, that, that pupil or whatnot and then track it for you on the screen as you're, they're doing it." [C1-Q23]
Individual level	
Symptom management	• "During our appointments, we were taught about the blue light ah and certain glasses we can wear during virtual, and then we made-up a rule cause I was always very stubborn at first with my injury and I really like to make eye contact through virtual. I thought it made it look more professional and how it should be. And they told me you don't need to do that because you are a patient. And so, we did 20 seconds looking at the, looking at the person speaking to you and then nobody will be offended if you look 2 minutes at your wall. So, 20 seconds, 2 minutes. Whenever you start feeling symptoms." [P7-Q19]
Support	• "I still have to do the balancing stuff, but I always make sure that somebody is here with me um that my partner is here with me so he can, if something happens." [P3-Q20] • "For some patients, if they have like a, you know, a partner or family member or somebody that you know, you can just ask to come for the physical exam, right? You ask them to just, you know, bring them along. And that helps with some of the safety aspects of it." [C4-Q26]
Environmental setup	• "I think sometimes actually just having like a shoe box and a laptop is pretty helpful cause the shoebox lets the laptop be up at more eye height." [C5-Q27]
Technology level	
Use of measure to identify gross deficits	
Use of measure as global screen and rely on subjective information	• I would say I would say that like eye movements was is really like in my practice a screen to make sure that they have eye movement, didn't have some kind of odd, you know like palsy, right. But beyond that, it just really you can't really interpret much more. Virtual." [C4-Q29]

[a]P: patient participant.

[b]Q: quote.

[c]C: clinician participant.

XSL•FO
RenderX

Table 7. Patient and clinician participant benefits associated with virtual concussion assessments.

Aggregated domains, overarching themes, and subthemes	Patients		Clinicians		
	Identified by patient participants	Patients reported (out of a total of 80 themes mentioned), n (%)	Identified by clinician participants	Clinicians reported (out of a total of 22 themes mentioned), n (%)	Clinician categories
Organizational level					
Access		36 (45)		15 (68)	
Improved access	—	—	✓		General
Cost saving	✓		—	—	—
Travel	✓		—	—	—
Technology level					
Convenience, comfort, and ease					
Sense of it being easier and more convenient at home	✓	44 (55)	✓	7 (31.8)	General

Table 8. Patient and clinician participant benefits quotes associated with virtual concussion assessments.

Aggregated domains, overarching themes, and subthemes	Quotes
Organizational level	
Access	• "I also had difficult to, traveling the distance because I live a far ways away. So it was, you know, it was nicer to have the virtual just because I didn't have to travel." [P6[a]-Q9[b]] • "I found that virtual assessments also allowed for patients who couldn't necessarily drive due to symptoms it facilitates having quicker access as well as, you know, being at least able to check in on them." [C3[c]-Q25]
Convenience, comfort, and ease	
Sense of it being easier and more convenient at home	• "it's convenient. It was very convenient to take it in the house and not have to go anywhere and they like again, it was a negative partly, but also a positive was the comfort I felt in my own home. I did feel better, my symptoms were I'd say better than they would have been in the hospital. And. Yeah. So, I think that it was like I, yeah, yeah, I felt I feel safer and more comfortable in my home." [P11-Q21] • "It could be easier to reach out to them a little bit earlier in the process because there aren't those barriers are of, you know, having to book an in-person appointment, them having to find transportation or having to find support to be brought in. So, they're in the comfort of their own homes and it's better for them in terms of their recovery because for a lot of these patients coming in is like a 1/2 day or a full day process." [C11-Q28]

[a]P: patient participant.

[b]Q: quote.

[c]C: clinician participant.

Patient-Participant Facilitators and Benefits

Organizational Level

A facilitator identified by the participants included having another health care professional in person with them (if in clinic), such as an occupational therapist, while the physician completed the virtual examination. Participants expressed that having easy access to technical support, when needed, would be helpful in troubleshooting technical difficulties and easing feelings of anxiety (Q16).

Having a choice of device used in the virtual assessment, being prepared in advance of the assessment, and using the features that videoconferencing offers (such as the record function) were facilitators identified by the patient participants that all relate to the process and format in which the assessment is delivered. Specifically, participants reported that having the ability to choose the device and screen size used for the virtual assessment would encourage use of the virtual assessment. Participants felt that having at least one in-person assessment, reflecting a hybrid approach, was helpful to enhance their feelings of comfort with the virtual assessment. Advanced preparation was another key facilitator identified. Participants felt that being contacted before

XSL•FO
RenderX

the assessment and being provided with information regarding what to expect would be helpful (Q17).

Having easy access to care was a benefit of virtual assessment use. Participants valued having access to care particularly when they were unable to attend in-person assessments due to various reasons (COVID-19 pandemic restrictions, comorbidities, unable to drive, etc; Q18).

Individual Level

Implementing symptom management strategies throughout the virtual assessment was identified as a facilitator. When clinicians offered screen breaks or recommended symptom management strategies to use during the assessment, such as using blue light glasses, the virtual assessments were more manageable for participants (Q19).

Having equipment such as a chair or another person (a partner or family member) at home were facilitators related to both the individual- and organization-level aggregate domains identified by patient participants. Participants reported feeling safe completing their assessments at home with these safety measures in place (Q20).

Technology Level

A sense of ease, comfort, and control with virtual assessments was expressed as a benefit. Specifically, feeling that care was not jeopardized when assessed virtually and a sense of virtual assessments being more convenient, comfortable, and requiring less energy were facilitators identified (Q21). Furthermore, the participants felt that they could be more independent with the virtual approach (eg, not rely on others to drive to be able to attend an assessment) and could control their environments better at home when compared to a hospital or clinic setting.

Clinician-Participant Facilitators and Benefits

Organization Level

Adequate preparation and integration of the virtual assessment with the in-person assessment (hybrid approach to care where the virtual option is available, if needed) were identified domain-specific facilitators. Clinicians expressed that sending information in advance of the assessment and informing patients of what to expect improved ease of completing a virtual assessment (Q22). In addition, having a backup in case the primary connection for the virtual assessment failed would help with the flow of the virtual assessment. These facilitators relate to the format in which the assessment and material needed for the assessment are delivered.

Using resources to support administration of certain measures virtually, such as mobile apps that can measure cervical spine angles or eye movements or systems that could analyze vestibular deficits, were identified as domain-specific facilitators by clinicians (Q23). For example, there are apps (such as HeadCheck) that clinicians reported using in their virtual practice that provide objective information on eye movements.

A domain-specific facilitator identified by the clinician participants included having access to clinical support in person. Clinicians reported that this would be particularly helpful due to safety concerns and due to challenges seeing subtleties over camera (Q24).

Clinicians reported the benefit of improved access to care that virtual care provides. Clinicians perceived that virtual assessment approaches could eliminate the need for patients to drive to assessment centers (or rely on a family member to drive) and perceived that this could facilitate easier and quicker access to care for the patient (Q25).

Individual Level

Similar to patient participants, clinicians expressed that holding onto a physical support (eg, chair, wall, and another person) facilitated the administration of certain measures virtually and was expressed as useful for most domains of the physical examination. Furthermore, access to emotional and technical support (a family member or partner) at home could help with the virtual assessment process. This also allows clinicians to gain the family member perspective, which could aid in getting collateral information and obtaining a more comprehensive understanding of clinical status (Q26).

Having an appropriate space to complete the virtual assessment at home and appropriately setting up the patient on the screen were identified as domain-specific facilitators. Clinician participants identified strategies, such as placing the computer on a shoebox and asking the patient to stand in a corner to complete certain measures virtually such as balance tests, would facilitate completion of the physical examination (Q27).

Technology Level

Clinicians reported a sense of ease, comfort, and convenience with virtual assessments. Particularly, they expressed that the virtual assessment allows patients to be in the comfort of their homes and eliminates some of the obstacles that need to be overcome to get to assessment centers (time, driving, and costs), which relates to a general benefit (Q28).

Clinician participants reported that using most of the measures to identify gross deficits (rather than specific deficits) or that using some of the measures for subjective information rather than objective information was a facilitator to its use in a virtual context (Q29). For example, when using the modified Balance Error Scoring System, clinicians reported that counting errors is a challenge over camera, so the measure is administered to identify gross balance deficits rather than identify subtle issues. Similarly, when administering the Vestibular/Ocular Motor Screening (VOMS) tool, clinicians rely on subjective reports of symptom changes and do not use this measure to identify specific ocular or vestibular issues.

Clinician Selection of Measures That Would Work Best in Virtual Assessments

The frequency of selection of each clinical measure is presented in Table 9. Within their respective domains, clinicians identified the finger to nose test, balance testing, the VOMS tool, saccades, and cervical spine range of motion as the measures that would work best and give the most relevant information in their virtual practice out of the list of measures provided.

XSL•FO
RenderX

Table 9. Clinician selection of measures that would work best in a virtual practice within the respective domains.

Domain and measures	Description of measures	Selection, n (%)
Neurological examination		
Coordination: finger to nose; rapid, alternating movements	• Finger to nose: patient extends arm and touches examiner's outstretched finger, and then touches patient's nose • Rapid alternating movements: patient alternately taps the back and palm surface of one hand onto the other hand; toe tapping: patient repeatedly taps foot on floor [35]	9 (64)
Motor (pronator drift and functional squat)	• Pronator drift: stretch both arms forward, with palms facing upward, and eyes closed [35] • Functional squat: observe and perform 10 body-weight squats [36]	3 (21)
Cranial nerve	• I: Smell object • II: Read the Snellen chart (a chart with varying-sized letters; cover one eye). Ask the patient to look at your eyes, place hands on both sides of the patient's head with the index fingers extended, ask the patient to indicate which finger you are moving (left or right) • Fundoscopy examination: clinician views the eye using ophthalmoscope • III, IV, and VI: eye movements; pursuit (follow moving target), convergence (eyes move at the same time, such as observing an object nearing the patient's nose), and saccades (alternate gaze between right and left and up and down targets) • V: facial sensation; light touch and pinprick both sides; clench teeth, and open jaw against resistance • VII: facial movements (show teeth, whistle, and close eyes tightly) • VIII: hearing test (tuning fork), test balance, and gait test for nystagmus (beating or twitching of eyes) • IX and X: voice (say "ah" with a tongue depressor), swallowing, cough, and gag reflex • XI: shrug shoulders and turning the head left and right (against resistance) • XII: open the mouth and stick out the tongue [35]	1 (7)
Vestibular		
Balance (feet together, single leg stance, and tandem stance)	• Feet together: standing with feet together; time-based tests • Single leg stance: standing on one leg and eyes open and eyes closed; time-based tests • Tandem stance: stand in heel to toe stance and eyes closed; time-based test [37]	9 (64)
VOMS[a]	• Smooth pursuits: the patient maintains focus on the target as the examiner moves the target smoothly in the horizontal and vertical direction • Saccades: the examiner holds 2 single points horizontally. Instruct the patient to move their eyes as quickly as possible from point to point. Repeat the test with 2 points held vertically • Convergence: the patient focuses on a target and slowly brings it toward the tip of their nose • Vestibular ocular reflex: the patient is asked to rotate their head horizontally while maintaining focus on the target. The test is repeated vertically • Visual motion sensitivity: the patient holds an arm outstretched and focuses on their thumbs. The patient rotates, together as a unit, their head, eyes, and trunk to the right and left [38]	9 (64)
Romberg	• Stand with feet together on a flat, hard surface. The examinee is asked to stand still and keep their eyes open for approximately 30 seconds. Next, the examinee is asked to close their eyes and stand for 30 seconds; time-based test [35]	3 (21)
VOR[b] test	• Move the head quickly side to side; repeat up and down [35]	3 (21)
Gait or tandem gait	• Observe the patient walk. Observe the heel to toe walk [35]	1 (7)
mBESS[c]	• Single leg stance (stand on one leg), tandem (stand in the heel to toe position), and double leg stance (stand with feet together) on flat ground and foam surfaces, with eyes closed. The examiner records errors during each condition (open eyes, lift hands off hips, etc) [39]	0 (0)
Oculomotor		
Saccades	• Rapid movement of eyes between fixed points. Instruct the patient to look from one target to the other as quickly as possible—vertical and horizontal [40]	7 (50)
Smooth pursuits	• Pursuits are performed in an H pattern. Participants track the object [41]	5 (36)

XSL•FO
RenderX

Domain and measures	Description of measures	Selection, n (%)
Convergence	• Use a pencil, placed just above the nose between the eyes. Move the target toward the patient at a rate of approximately 1-2 cm/s, encouraging them to keep the target single. Measure the patient's reported subjective break (target becomes double) in centimeters. Measure in centimeters [42]	2 (14)
Cervical		
Range of motion	• Active range of motion: head turn left, right, up, down, and side flexion left and right [37]	13 (93)
Self-palpation	• Patient feels with fingers and hands during physical examination of the neck [37]	2 (14)

[a]VOMS: Vestibular/Ocular Motor Screening.

[b]VOR: vestibular ocular reflex.

[c]mBESS: modified Balance Error Scoring System.

Discussion

Principal Findings

This study used focus groups to identify barriers and facilitators associated with virtual concussion assessments. To our knowledge, this work is the first to explore the barriers and facilitators associated with virtual concussion assessments from the perspectives of people living with workplace concussions. The findings of this study indicate that a hybrid approach (first assessment completed in person and follow-ups completed virtually) to concussion assessment is preferred by both people living with workplace concussions and clinicians. It appears that the virtual approach to concussion assessment is of value, highlighting a need to further explore and consolidate information on the effectiveness of the clinical measures used in concussion assessments when administered virtually. People living with workplace concussions seem to prefer having the initial assessment with the care team in person, followed by virtual assessments as needed. Virtual assessments were perceived as most appropriate for follow-ups and for a global screen of deficits rather than for identifying specific deficits.

In this study, overlapping themes were identified for both clinician and patient participants. Interestingly, both clinicians and patients commented on difficulties associated with completing the required documents (questionnaires) virtually. Both expressed a need to verbally complete the required documents with the patients over the phone or have support in place so that they have help to complete them before the virtual assessment. This is in line with Beaton et al [43], who documented that the preferred method of information delivery for people living with a concussion is in auditory format.

Both clinicians and people living with workplace concussions highlighted that inaccurate values may be obtained with the virtual assessment, which was the most common barrier identified (identified by patient and clinician participants approximately 35% and 26% of the time, respectively). Objectivity in assessments conducted using virtual care is a challenge, particularly with the lack of clarity when observing certain movements through videoconferencing, which in turn influences clinicians' abilities to adequately assess patients [24]. Virtual care has been used for the management of concussions; however, an identified limitation is the challenge associated with the remote physical examination [24]. Subtle abnormalities such as those observed during objective oculomotor assessments could be difficult to capture through videoconferencing [7]. Furthermore, a neurological examination in its entirety is not currently feasible through videoconferencing [44,45]. Particularly, a complete evaluation of the cranial nerves (fundoscopy) cannot be completed along with motor and sensory function, cervical spine palpation, strength testing, muscle reflexes, and neuro-otology maneuvers such as those for vertigo [7,15,36,46]. Both clinicians and patients expressed challenges associated with using technology, including the virtual approach to assessment that can be symptom-triggering, safety concerns with the virtual assessment in relation to potential for falling, challenges with the setup at home, and barriers associated with communicating and establishing a clinician-patient rapport through a screen.

This study identified technical, individual, and organizational-level barriers that are in line with some of the literature on virtual care in people living with neurological conditions. Ownsworth et al [47] reported limitations associated with videoconferencing specifically, which include issues with technology and incompatibility for the patient (inability to use technology). Technical issues such as poor video or audio quality due to either the varying quality of devices used or Wi-Fi signal strength and difficulty placing the camera are identified barriers specific to videoconferencing [11]. Logistical challenges when a caregiver is not present at home to aid with the virtual session contributes to the nonfeasibility of virtual care approaches [13]. Concerns of lack of security and concerns regarding privacy in terms of family members overhearing sessions are described barriers [11], which was a concern expressed by patient participants in this study. Consistent with the findings of this study, the literature documented that the lack of resources including the lack of access to the administrative or skilled personnel support needed and lack of time, as virtual assessments are perceived to take more time to appropriately set up the patient, influence the feasibility and use of videoconferencing approaches [48]. A limitation associated with virtual assessments identified by the participants in this study includes lack of physical presence [47]. Lack of physical presence has been documented to be associated with challenges in terms of building rapport when not face to face as well as an inability to gain a full understanding of the

XSL•FO
RenderX

patient's status [47]. The patient-clinician relationship may be jeopardized if technologies are used due to the difficulties associated with building trust through screens and the challenges of clear communication [15,44,46].

From a facilitation standpoint, both the clinicians and people living with workplace concussions in this study highlighted the convenience and improved access that virtual assessments provide. Both groups emphasized the need to have appropriate supports and preparation in place; however, only the patient participants identified this as the most common facilitator (identified approximately 27% of the time). The patient participants in this study identified the importance of having easy access to technical support. This is in line with Hale-Gallardo et al [13] who reported that virtual care coordinators and technical clinicians are viewed as resources that could facilitate the adoption of technology-based care in clinical practice. Having assistants help set up the system is a necessary facilitator to videoconferencing [16]. Family support is additionally viewed as a facilitator due to the help they provided with the technical and clinical aspects of videoconferencing [11]. The increased likelihood of family members being present during videoconferencing is an advantage, which could aid with gathering information [49]. This is also helpful to aid with remembering recommendations and next steps of care. Clinician participants identified the format of the assessment and materials (adequate preparation and offering virtual assessments as a complement to in-person assessments) as the most common facilitator (identified approximately 19% of the time). Development of clear instructions and advanced planning to prepare for issues are all necessary facilitators to videoconferencing [16]. The importance of having an initial in-person touchpoint and preference for the hybrid approach was highlighted in the literature. This approach enhances the comfort of the patient with the clinician [11]. Clinician and patient understanding of the benefits of virtual approaches to care, which includes cost savings, autonomy for the user, and reduced travel and wait times act as facilitators for the use of technology-based approaches to assessment [11,16,46].

Several unique findings of this study are specific to a virtual assessment after a concussion. Specifically, the challenges with completing the virtual assessment due to screen intolerance and resulting symptom aggravation were a notable barrier to engaging in a virtual assessment. Depending on the time point in recovery, exposure to screens could be a very challenging barrier to overcome to complete a safe and effective assessment virtually. One strategy to address this issue, which was identified as a facilitator by participants, was implementing symptom management strategies (screen breaks and blue light glasses) so that the virtual assessment could be more tolerable. In addition, due to subtle deficits that are commonly experienced by people living with workplace concussions, virtual care approaches may not be currently adequate to capture these subtleties, and therefore, in-person care may be needed to identify issues that go beyond gross deficits. Furthermore, there are potential contradictions with the focus group data. For example, balance and the VOMS tool were selected as the vestibular measures that would work best in a virtual practice by clinicians. However, several issues were identified with these measures such as the challenges with visualizing subtle deficits associated with eye movements and balance when assessing virtually. Even with these challenges, the clinicians still believed that out of the measures identified in the surveys and discussed in the focus groups, these measures would be most appropriate for virtual assessments.

This work expands on some of the previous work on the clinician perspectives of virtual concussion care to outline barriers and facilitators to completing components of the physical examination [17]. Finally, this work reports on clinician-identified measures that will work best in a virtual context and identifies the barriers and facilitators associated with using these measures virtually so that the most appropriate measure for virtual administration can be selected. This work supports the identification of clinical measures to include in a virtual version of a concussion assessment, which will be further explored in a future planned evaluation study with the aim of documenting the psychometric properties of the measures. These measures will be included in a virtual assessment toolkit, which will be tested in a future study.

Limitations

There are some important limitations to acknowledge in this work. The context of this work is within workplace injury, and all the patient participants who participated in the focus groups were experiencing persistent symptoms after injury at the time of their virtual assessment. Their experiences may be unique to those experiencing injuries occurring in the workplace, and therefore, the results may not be generalizable to people with concussions from other causes. Furthermore, only 27% of the patient participants were male, which indicates that the results of this study may be more reflective of female experiences. However, Merritt et al [50] reported that the prevalence of concussions is higher in female individuals, so the distribution of male individuals and female individuals in this study may be a close reflection of concussion occurrence in the working population. In addition, while attempts were made to include clinicians from various clinical fields (while focusing on professions that typically complete the physical examination), certain professions were not represented in the focus groups, such as neurology and occupational therapy. Finally, the selection of measures as most appropriate to use in a virtual assessment may be biased based on clinician experiences and practices. Clinician level of comfort with each clinical measure may have influenced their decisions.

Conclusions

It is important to note that patients and clinicians identified limitations and challenges associated with virtual care. A hybrid approach to concussion assessment seems to be the most accepted model to assessment from the perspectives of both clinicians and people living with workplace concussions. This study demonstrates the acceptance and perceived value of virtual approaches in concussion assessment and therefore provides a rationale for further exploration of the psychometric properties of the virtual concussion assessment. There is an identified need to establish reliability and validity properties of the virtual concussion examination.

XSL•FO
RenderX

Acknowledgments

KB is supported by an admission scholarship from the University of Ottawa, an Early Professionals, Inspired Careers in AgeTech (EPIC-AT) Health Research Training Program fellowship through AGE-WELL, and the Ontario Graduate Scholarship. MK is supported by an admission scholarship, international doctoral scholarship from the University of Ottawa, Hans K Uhthoff, MD Fellow of the Royal College of Surgeons of Canada Graduate fellowship, and BMO Financial Group Graduate Scholarship. This project was funded by a grant provided by the Workplace Safety and Insurance Board (WSIB; Ontario). The provision of grant support by WSIB does not in any way infer or imply endorsement of the content by the WSIB.

Data Availability

All data generated or analyzed during this study are included in this published manuscript and its multimedia appendices.

Authors' Contributions

KB, HS, and SM contributed to the development of the protocol, with supporting guidance from M Bayley, ME, M Bolideau, MR, and MT. KB, HS, SM, M Bayley, ME, M Bolideau, MR, and MT reviewed and approved the protocol. KB moderated the focus group. MK participated as a notetaker in the focus groups. KB transcribed the focus groups. KB and MK coded the transcripts and generated the themes. KB drafted the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standards for Reporting Qualitative Research checklist.
[PDF File (Adobe PDF File), 1157 KB - jmir_v26i1e56158_app1.pdf]

Multimedia Appendix 2
Barrier rankings.
[DOCX File , 15 KB - jmir_v26i1e56158_app2.docx]

Multimedia Appendix 3
Facilitator rankings.
[DOCX File , 15 KB - jmir_v26i1e56158_app3.docx]

Multimedia Appendix 4
Patient participant overarching themes, subthemes, and codes.
[DOCX File , 328 KB - jmir_v26i1e56158_app4.docx]

Multimedia Appendix 5
Clinician participant overarching themes, subthemes, and codes.
[DOCX File , 468 KB - jmir_v26i1e56158_app5.docx]

Multimedia Appendix 6
Patient subtheme definitions.
[DOCX File , 15 KB - jmir_v26i1e56158_app6.docx]

Multimedia Appendix 7
Clinician subtheme definitions.
[DOCX File , 14 KB - jmir_v26i1e56158_app7.docx]

References

1. Ventura RE, Balcer LJ, Galetta SL. The neuro-ophthalmology of head trauma. Lancet Neurol 2014 Oct;13(10):1006-1016. [doi: 10.1016/S1474-4422(14)70111-5] [Medline: 25231523]

XSL•FO
RenderX

2. Greenberg MS, Wood NE, Spring JD, Gurvits TV, Nagurney JT, Zafonte RD, et al. Pilot study of neurological soft signs and depressive and postconcussive symptoms during recovery from mild traumatic brain injury (mTBI). J Neuropsychiatry Clin Neurosci 2015;27(3):199-205. [doi: 10.1176/appi.neuropsych.14050111] [Medline: 26222967]
3. Lilley R, Davie G, Langley J, Ameratunga S, Derrett S. Do outcomes differ between work and non-work-related injury in a universal injury compensation system? Findings from the New Zealand Prospective Outcomes of Injury Study. BMC Public Health 2013 Oct 22;13:995 [FREE Full text] [doi: 10.1186/1471-2458-13-995] [Medline: 24148609]
4. Colantonio A, Comper P. Post-injury symptoms after work related traumatic brain injury in Canadian population. Work 2012;43(2):195-201. [doi: 10.3233/WOR-2012-1377] [Medline: 22927617]
5. Caze Ii T, Knell GP, Abt J, Burkhart SO. Management and treatment of concussions via tele-concussion in a pediatric setting: methodological approach and descriptive analysis. JMIR Pediatr Parent 2020 Aug 13;3(2):e19924 [FREE Full text] [doi: 10.2196/19924] [Medline: 32634106]
6. Vargas BB, Channer DD, Dodick DW, Demaerschalk BM. Teleconcussion: an innovative approach to screening, diagnosis, and management of mild traumatic brain injury. Telemed J E Health 2012 Dec;18(10):803-806. [doi: 10.1089/tmj.2012.0118] [Medline: 23101482]
7. Ellis MJ, Boles S, Derksen V, Dawyduk B, Amadu A, Stelmack K, et al. Evaluation of a pilot paediatric concussion telemedicine programme for northern communities in Manitoba. Int J Circumpolar Health 2019 Dec;78(1):1573163 [FREE Full text] [doi: 10.1080/22423982.2019.1573163] [Medline: 30714513]
8. Bishop SA, Bath B, Wiechnik C, Mendez I, Johnson R, Lovo S. Use of virtual care strategies to join multidisciplinary teams evaluating work-related injuries in rural residents. Telemed J E Health 2023 Jan;29(1):116-126. [doi: 10.1089/tmj.2021.0548] [Medline: 35584260]
9. Rush KL, Seaton CL, Corman K, Hawe N, Li EP, Dow-Fleisner SJ, et al. Virtual care prior to and during COVID-19: cross-sectional survey of rural and urban adults. JMIR Form Res 2022 Aug 22;6(8):e37059 [FREE Full text] [doi: 10.2196/37059] [Medline: 35849794]
10. Gross DP, Asante A, Pawluk J, Niemeläinen R. A descriptive study of the implementation of remote occupational rehabilitation services due to the COVID-19 pandemic within a workers' compensation context. J Occup Rehabil 2021 Jun;31(2):444-453 [FREE Full text] [doi: 10.1007/s10926-020-09934-7] [Medline: 33118130]
11. Almathami HK, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407 [FREE Full text] [doi: 10.2196/16407] [Medline: 32130131]
12. Venkataraman K, Morgan M, Amis KA, Landerman LR, Koh GC, Caves K, et al. Tele-assessment of the berg balance scale: effects of transmission characteristics. Arch Phys Med Rehabil 2017 Apr;98(4):659-64.e1 [FREE Full text] [doi: 10.1016/j.apmr.2016.10.019] [Medline: 27894732]
13. Hale-Gallardo JL, Kreider CM, Jia H, Castaneda G, Freytes IM, Cowper Ripley DC, et al. Telerehabilitation for rural veterans: a qualitative assessment of barriers and facilitators to implementation. J Multidiscip Healthc 2020;13:559-570 [FREE Full text] [doi: 10.2147/JMDH.S247267] [Medline: 32669850]
14. Hoaas H, Andreassen HK, Lien LA, Hjalmarsen A, Zanaboni P. Adherence and factors affecting satisfaction in long-term telerehabilitation for patients with chronic obstructive pulmonary disease: a mixed methods study. BMC Med Inform Decis Mak 2016 Feb 25;16:26 [FREE Full text] [doi: 10.1186/s12911-016-0264-9] [Medline: 26911326]
15. Tenforde AS, Hefner JE, Kodish-Wachs JE, Iaccarino MA, Paganoni S. Telehealth in physical medicine and rehabilitation: a narrative review. PM R 2017 May;9(5S):S51-S58. [doi: 10.1016/j.pmrj.2017.02.013] [Medline: 28527504]
16. Turkstra LS, Quinn-Padron M, Johnson JE, Workinger MS, Antoniotti N. In-person versus telehealth assessment of discourse ability in adults with traumatic brain injury. J Head Trauma Rehabil 2012;27(6):424-432 [FREE Full text] [doi: 10.1097/HTR.0b013e31823346fc] [Medline: 22190010]
17. van Ierssel J, O'Neil J, King J, Zemek R, Sveistrup H. Clinician perspectives on providing concussion assessment and management via telehealth: a mixed-methods study. J Head Trauma Rehabil 2023;38(3):E233-E243. [doi: 10.1097/HTR.0000000000000827] [Medline: 36731011]
18. Watt JA, Fahim C, Straus SE, Goodarzi Z. Barriers and facilitators to virtual care in a geriatric medicine clinic: a semi-structured interview study of patient, caregiver and healthcare provider perspectives. Age Ageing 2022 Jan 06;51(1):afab218. [doi: 10.1093/ageing/afab218] [Medline: 34794177]
19. Toresdahl BG, Young WK, Quijano B, Scott DA. A systematic review of telehealth and sport-related concussion: baseline testing, diagnosis, and management. HSS J 2021 Feb;17(1):18-24 [FREE Full text] [doi: 10.1177/1556331620975856] [Medline: 33967637]
20. Langer L, Levy C, Bayley M. Increasing incidence of concussion: true epidemic or better recognition? J Head Trauma Rehabil 2020;35(1):E60-E66. [doi: 10.1097/HTR.0000000000000503] [Medline: 31246881]
21. Chun YJ, Patterson PE. A usability gap between older adults and younger adults on interface design of an internet-based telemedicine system. Work 2012;41 Suppl 1:349-352. [doi: 10.3233/WOR-2012-0180-349] [Medline: 22316747]
22. Ansary AM, Martinez JN, Scott JD. The virtual physical exam in the 21st century. J Telemed Telecare 2021 Jul;27(6):382-392. [doi: 10.1177/1357633X19878330] [Medline: 31690169]

XSL•FO
RenderX

23. McIntyre M, Robinson LR, Mayo A. Practical considerations for implementing virtual care in physical medicine and rehabilitation: for the pandemic and beyond. Am J Phys Med Rehabil 2020 Jun;99(6):464-467 [FREE Full text] [doi: 10.1097/PHM.0000000000001453] [Medline: 32324617]
24. Hill S, Barr C, Killington M, McLoughlin J, Daniels R, van den Berg M. The design and development of MOVE-IT: a system for remote vestibular and oculomotor assessment in people with concussion. In: Maeder AJ, Higa C, van den Berg ME, Gough C, editors. Studies in Health Technology and Informatics. New York, NY: IOS Press; 2021:27-36.
25. Bigler ED. Symptom validity testing, effort, and neuropsychological assessment. J Int Neuropsychol Soc 2012 Jul;18(4):632-640. [doi: 10.1017/s1355617712000252] [Medline: 23057080]
26. Barnes K, Sveistrup H, Bayley M, Egan M, Rathbone M, Taljaard M, et al. Identification of clinical measures to use in a virtual concussion assessment: protocol for a mixed methods study. JMIR Res Protoc 2022 Dec 22;11(12):e40446 [FREE Full text] [doi: 10.2196/40446] [Medline: 36548031]
27. Barnes K, Sveistrup H, Bayley M, Egan M, Bilodeau M, Rathbone M, et al. Clinician-prioritized measures to use in a remote concussion assessment: delphi study. JMIR Form Res 2024 Sep 02;8:e47246 [FREE Full text] [doi: 10.2196/47246] [Medline: 39222352]
28. Etikan I, Musa SA, Alkassim RS. Comparison of convenience sampling and purposive sampling. Am J Theor Appl Stat 2016;5(1):1 [FREE Full text] [doi: 10.11648/j.ajtas.20160501.11]
29. Hennink MM, Kaiser BN, Weber MB. What influences saturation? Estimating sample sizes in focus group research. Qual Health Res 2019 Aug;29(10):1483-1496 [FREE Full text] [doi: 10.1177/1049732318821692] [Medline: 30628545]
30. Guest G, Namey E, McKenna K. How many focus groups are enough? Building an evidence base for nonprobability sample sizes. Field Methods 2016 Jul 24;29(1):3-22. [doi: 10.1177/1525822X16639015]
31. Moretti F, van Vliet L, Bensing J, Deledda G, Mazzi M, Rimondini M, et al. A standardized approach to qualitative content analysis of focus group discussions from different countries. Patient Educ Couns 2011 Mar;82(3):420-428. [doi: 10.1016/j.pec.2011.01.005] [Medline: 21292424]
32. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
33. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
34. Weber RP. Basic Content Analysis. Thousand Oaks, CA: Sage Publications; 1990.
35. Fuller G. Neurological Examination Made Easy. 6th edition. New Brunswick, QC: Elsevier Health Sciences; 2019.
36. Al Hussona M, Maher M, Chan D, Micieli JA, Jain JD, Khosravani H, et al. The virtual neurologic exam: instructional videos and guidance for the COVID-19 era. Can J Neurol Sci 2020 Sep;47(5):598-603 [FREE Full text] [doi: 10.1017/cjn.2020.96] [Medline: 32434626]
37. Johnston S, Leddy J, Reed N, Cortel-LeBlanc A, Hafizi R, Laing S, et al. Virtual concussion exam (VCE) manual. Foundation for Advancing Family Medicine and CMA Foundation. 2022. URL: https://pedsconcussion.com/wp-content/uploads/VirtualConcussionExamManual.pdf [accessed 2024-04-29]
38. Mucha A, Collins MW, Elbin RJ, Furman JM, Troutman-Enseki C, DeWolf RM, et al. A brief vestibular/ocular motor screening (VOMS) assessment to evaluate concussions: preliminary findings. Am J Sports Med 2014 Oct;42(10):2479-2486 [FREE Full text] [doi: 10.1177/0363546514543775] [Medline: 25106780]
39. Azad AM, Al Juma S, Bhatti JA, Delaney JS. Modified Balance Error Scoring System (M-BESS) test scores in athletes wearing protective equipment and cleats. BMJ Open Sport Exerc Med 2016;2(1):e000117 [FREE Full text] [doi: 10.1136/bmjsem-2016-000117] [Medline: 27900181]
40. Termsarasab P, Thammongkolchai T, Rucker JC, Frucht SJ. The diagnostic value of saccades in movement disorder patients: a practical guide and review. J Clin Mov Disord 2015;2:14 [FREE Full text] [doi: 10.1186/s40734-015-0025-4] [Medline: 26788350]
41. Fife TD, Tusa RJ, Furman JM, Zee DS, Frohman E, Baloh RW, et al. Assessment: vestibular testing techniques in adults and children: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. Neurology 2000 Nov 28;55(10):1431-1441. [doi: 10.1212/wnl.55.10.1431] [Medline: 11094095]
42. Scheiman M, Gallaway M, Frantz KA, Peters RJ, Hatch S, Cuff M, et al. Nearpoint of convergence: test procedure, target selection, and normative data. Optom Vis Sci 2003 Mar;80(3):214-225. [doi: 10.1097/00006324-200303000-00011] [Medline: 12637833]
43. Beaton MD, Hadly G, Babul S. Stakeholder recommendations to increase the accessibility of online health information for adults experiencing concussion symptoms. Front Public Health 2020;8:557814 [FREE Full text] [doi: 10.3389/fpubh.2020.557814] [Medline: 33505948]
44. Ellis MJ, Russell K. The potential of telemedicine to improve pediatric concussion care in rural and remote communities in Canada. Front Neurol 2019;10:840 [FREE Full text] [doi: 10.3389/fneur.2019.00840] [Medline: 31428043]
45. Wechsler LR, Tsao JW, Levine SR, Swain-Eng RJ, Adams RJ, Demaerschalk BM, American Academy of Neurology Telemedicine Work Group. Teleneurology applications: report of the telemedicine work group of the american academy of neurology. Neurology 2013 Mar 12;80(7):670-676 [FREE Full text] [doi: 10.1212/WNL.0b013e3182823361] [Medline: 23400317]

XSL•FO
RenderX

46. Dorsey ER, Glidden AM, Holloway MR, Birbeck GL, Schwamm LH. Teleneurology and mobile technologies: the future of neurological care. Nat Rev Neurol 2018 May;14(5):285-297. [doi: 10.1038/nrneurol.2018.31] [Medline: 29623949]
47. Ownsworth T, Theodoros D, Cahill L, Vaezipour A, Quinn R, Kendall M, et al. Perceived usability and acceptability of videoconferencing for delivering community-based rehabilitation to individuals with acquired brain injury: a qualitative investigation. J Int Neuropsychol Soc 2020 Jan;26(1):47-57. [doi: 10.1017/S135561771900078X] [Medline: 31983367]
48. Tran V, Lam MK, Amon KL, Brunner M, Hines M, Penman M, et al. Interdisciplinary eHealth for the care of people living with traumatic brain injury: a systematic review. Brain Inj 2017;31(13-14):1701-1710. [doi: 10.1080/02699052.2017.1387932] [Medline: 29064300]
49. Larner AJ. Teleneurology: an overview of current status. Pract Neurol 2011 Oct;11(5):283-288. [doi: 10.1136/practneurol-2011-000090] [Medline: 21921003]
50. Merritt VC, Padgett CR, Jak AJ. A systematic review of sex differences in concussion outcome: what do we know? Clin Neuropsychol 2019 Aug;33(6):1016-1043. [doi: 10.1080/13854046.2018.1508616] [Medline: 30618335]

Abbreviations

VOMS: Vestibular/Ocular Motor Screening

Edited by A Mavragani; submitted 09.01.24; peer-reviewed by M Fenner, S Ren, G Srinivasan, A Svingos; comments to author 01.07.24; revised version received 22.08.24; accepted 25.09.24; published 13.11.24.

Please cite as:
Barnes K, Sveistrup H, Karimijashni M, Bayley M, Egan M, Bilodeau M, Rathbone M, Taljaard M, Marshall S
Barriers and Facilitators Associated With Remote Concussion Physical Assessments From the Perspectives of Clinicians and People Living With Workplace Concussions: Focus Group Study
J Med Internet Res 2024;26:e56158
URL: https://www.jmir.org/2024/1/e56158
doi:10.2196/56158
PMID:

©Keely Barnes, Heidi Sveistrup, Motahareh Karimijashni, Mark Bayley, Mary Egan, Martin Bilodeau, Michel Rathbone, Monica Taljaard, Shawn Marshall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing the Quality of Direct-to-Consumer Telemedicine Dominated and Delivered by Public and Private Sector Platforms in China: Standardized Patient Study

Faying Song[1*], BS; Xue Gong[2*], MS; Yuting Yang[3], MS; Rui Guo[1], PhD

[1]School of Public Health, Capital Medical University, Beijing, China

[2]Beijing Luhe Hospital, Capital Medical University, Beijing, China

[3]Hospital Management Research Institute, Peking University Third Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Rui Guo, PhD
School of Public Health
Capital Medical University
No 10 Youanmenwai, Xitoutiao
Beijing,
China
Phone: 86 010 83911055
Email: guorui@ccmu.edu.cn

Abstract

Background: Telemedicine is expanding rapidly, with public direct-to-consumer (DTC) telemedicine representing 70% of the market. A key priority is establishing clear quality distinctions between the public and private sectors. No studies have directly compared the quality of DTC telemedicine in the public and private sectors using objective evaluation methods.

Objective: Using a standardized patient (SP) approach, this study aimed to compare the quality of DTC telemedicine provided by China's public and private sectors.

Methods: We recruited 10 SPs presenting fixed cases (urticaria and childhood diarrhea), with 594 interactions between them and physicians. The SPs evaluated various aspects of the quality of care, effectiveness, safety, patient-centeredness (PCC), efficiency, and timeliness using the Institute of Medicine (IOM) quality framework. Ordinary least-squares (OLS) regression models with fixed effects were used for continuous variables, while logistic regression models with fixed effects were used for categorical variables.

Results: Significant quality differences were observed between public and private DTC telemedicine. Physicians from private platforms were significantly more likely to adhere to clinical checklists (adjusted β 15.22, $P<.001$); provide an accurate diagnosis (adjusted odds ratio [OR] 3.85, $P<.001$), an appropriate prescription (adjusted OR 3.87, $P<.001$), and lifestyle modification advice (adjusted OR 6.82, $P<.001$); ensure more PCC (adjusted β 3.34, $P<.001$); and spend more time with SPs (adjusted β 839.70, $P<.001$), with more responses (adjusted β 1.33, $P=.001$) and more words (adjusted β 50.93, $P=.009$). However, SPs on private platforms waited longer for the first response (adjusted β 505.87, $P=.001$) and each response (adjusted β 168.33, $P=.04$) and paid more for the average visit (adjusted β 40.03, $P<.001$).

Conclusions: There is significant quality inequality in different DTC telemedicine platforms. Private physicians might provide a higher quality of service regarding effectiveness and safety, PCC, and response times and words. However, private platforms have longer wait times for their first response, as well as higher costs. Refining online reviews, establishing standardized norms and pricing, enhancing the performance evaluation mechanism for public DTC telemedicine, and imposing stricter limitations on the first response time for private physicians should be considered practical approaches to optimizing the management of DTC telemedicine.

(J Med Internet Res 2024;26:e55400) doi:10.2196/55400

XSL•FO
RenderX

KEYWORDS

telemedicine; direct-to-consumer telemedicine; standardized patient; China; public; private; platform; objective evaluation; quality of care; effectiveness; safety; timeliness; regression model; management

Introduction

Telehealth (telecare or telemedicine) has emerged as a sustainable alternative to traditional in-person visits, including clinician-to-clinician, clinician-to-patient, and patient-to-mobile health technology [1,2]. A virtual visit delivered by internet hospitals in China is equivalent to direct-to-consumer (DTC) telemedicine/telehealth [3], which directly provides medical services (appointments, consultations, e-prescriptions, drug delivery, health care management, etc) to patients by using information technology to extend medical resources from hospitals to the internet [4,5]. The COVID-19 pandemic has boosted telemedicine in medical care visits, and increasingly more people are beginning to use telemedicine [6,7]. According to the statistics of the China Internet Network Information Center, as of June 2023, the number of internet medical users in China reached 364 million, marking an increase of 1.62 million compared to December 2022 [8]. The demand for and the supply of DTC telemedicine in China have attained a novel level.

There are primarily public brick-and-mortar hospital-sponsored and private enterprise–sponsored internet hospitals in China's DTC telemedicine market [3]. Public DTC telemedicine is considered a nonprofit medical service, and its workforce comprises practicing physicians in public hospitals [9]. Its charges must be implemented in accordance with the government's price policy, and the government pays the input costs. Public DTC telemedicine assigns patients to doctors and mainly provides basic medical services. In 2014, the Second Provincial Hospital of Guangdong established the first internet hospital in China [10]. In 2016, the first public tertiary hospital campus–sponsored internet hospital (the First Affiliated Hospital of Zhejiang University) was launched [11]. In 2018, the State Council of China issued the "Internet Plus Healthcare of 2018" guideline [12], and the National Health Commission accordingly issued specific regulation schemes for online medical diagnosis and treatment in 2018 [13], which have brought about the rapid development of DTC telemedicine [6], especially public DTC telemedicine. According to statistics, approximately 70% of internet hospitals in China have been initiated by public hospitals (public DTC telemedicine) [14].

Private DTC telemedicine operates as a self-contained for-profit institution, with services provided by platform physicians or doctors with multiple practices. The pricing can be carried out according to the market law [9]. Around 2010, private enterprises, such as WeDoctor and Haodf, took the lead in launching online health care, providing physician comments/ratings, appointments and referrals, and postconsultation management, which is the prototype of China's DTC telemedicine. These enterprises use physicians' brands to attract patients, which can better address patient demands and offer a broader spectrum of services. The integration of advanced technologies, such as artificial intelligence and big data analytics, has further enhanced the capabilities of these platforms, enabling more personalized and precise health care services.

Theoretically, opponents of the private sector argue that profit-driven private health care providers can manipulate services to maximize profits rather than quality [15,16]. In contrast, a competing theory claims that private health care providers can use more up-to-date and flexible management to improve quality [17]. With the rapid development of DTC telemedicine in China, it is a priority to clarify the difference in health care quality between public and private health care providers, which can provide some information and inspiration for patients when choosing services. However, the existing quality evaluations of private and public medical services mainly focus on face-to-face services, and the evidence is insufficient and conflicting [18-21]. There is no nationally representative sample, as previous studies have compared the quality of private and public primary health care within specific Chinese cities. Internationally, 2 earlier reviews [22,23] of 80 and 102 studies, respectively, came to opposite conclusions regarding the relative quality of the 2 sectors. One concluded that the private sector is more responsive, puts in more effort, and can prioritize clients' needs [22], while the other failed to support the claim that the private sector is generally more efficient, responsible, or medically effective than the public sector [23]. Another 2 reviews (including the latest 2017 one) [24,25] added no certainty to the debate. Regarding DTC telemedicine, some scholars have evaluated its subjective quality based on patient perception and recall-based surveys rather than actual clinical ability; other scholars have explored its objective quality based on patient data and the standardized patient (SP) approach [26,27]. For instance, Resneck et al [28] contended that although telemedicine may enhance the availability of high-value health care, its accuracy of diagnoses and quality of prescriptions and treatment recommendations have not met expectations because of fragmentation. Gong et al [29] assessed the quality of DTC telemedicine services using an SP approach. They discovered that doctors with better response times and more words are significantly more likely to adhere to clinical checklists and provide accurate diagnoses, appropriate prescriptions, and lifestyle modification advice. Crucially, there is yet no evidence from large samples of quality differences between public and private telemedicine.

Studies have suggested that comparing services provided by the public and private sectors needs to address 3 significant barriers: a clear definition of quality, an accurate measurement of quality, and a sufficient adjustment for the case mix [30]. For this study, we referred to the framework developed by the Institute of Medicine (IOM) [31] and the specific measurement indicators proposed by Xu et al [30] based on this model to define quality. Furthermore, we suggested using SPs to measure quality and control the case mix accurately. This method has gained rapid acceptance for quality measurement in countries worldwide [32-34]. The SP approach enables researchers to compare the care that is provided against predetermined national

XSL•FO
RenderX

standards of care; avoids Hawthorne effects, whereby doctors change their behavior because they know they are being observed; and allows researchers to control for potential confounding arising from differences in the case mix and patient recall bias [35].

Text-based consultation is among the most frequently used DTC telemedicine services in China. Therefore, we used SPs to collect and compare the quality between public and private internet text-based consultations. This study is the first large sample of evidence to evaluate and compare the quality of public and private DTC telemedicine services using an SP approach and the IOM framework in the context of China. In anticipation, our findings will inform policy makers in China and similar countries that provide public/private DTC telemedicine.

Methods

Case Selection

Each SP presented 1 of 2 medical cases, urticaria or childhood diarrhea of an absent child. Some scholars have proposed 10 questions to consider when assessing suitability as an SP case [36]. The specific reasons for choosing the 2 cases are presented in Multimedia Appendix 1. In general, these diseases (1) have a high incidence in China, (2) are easy to be portrayed online through pictures and texts, (3) can be diagnosed preliminarily without an offline invasive examination, and (4) have national guidelines and clear clinical treatment pathways (medication, patient education, etc).

Study Design and Setting

We considered creating a nationally representative probability sample of China's DTC telemedicine providers in tertiary hospitals. The provinces have been divided into 3 levels based on a comprehensive ranking that considers both economic development (judging from the real gross domestic product [GDP] per capita) and health status (judging from life expectancy at the provincial level), with each factor weighted equally (Figure 1), which also corresponds to the regional distribution (eastern, central. and western) [37]. Ultimately, for feasibility reasons, we chose 6 provinces (Guangdong, Zhejiang, Hunan, Anhui, Guangxi, and Xinjiang) to represent China. Multimedia Appendix 2 presents the process of dividing the provinces into 3 levels.

Figure 1. Six selected sample provinces on the map of China.

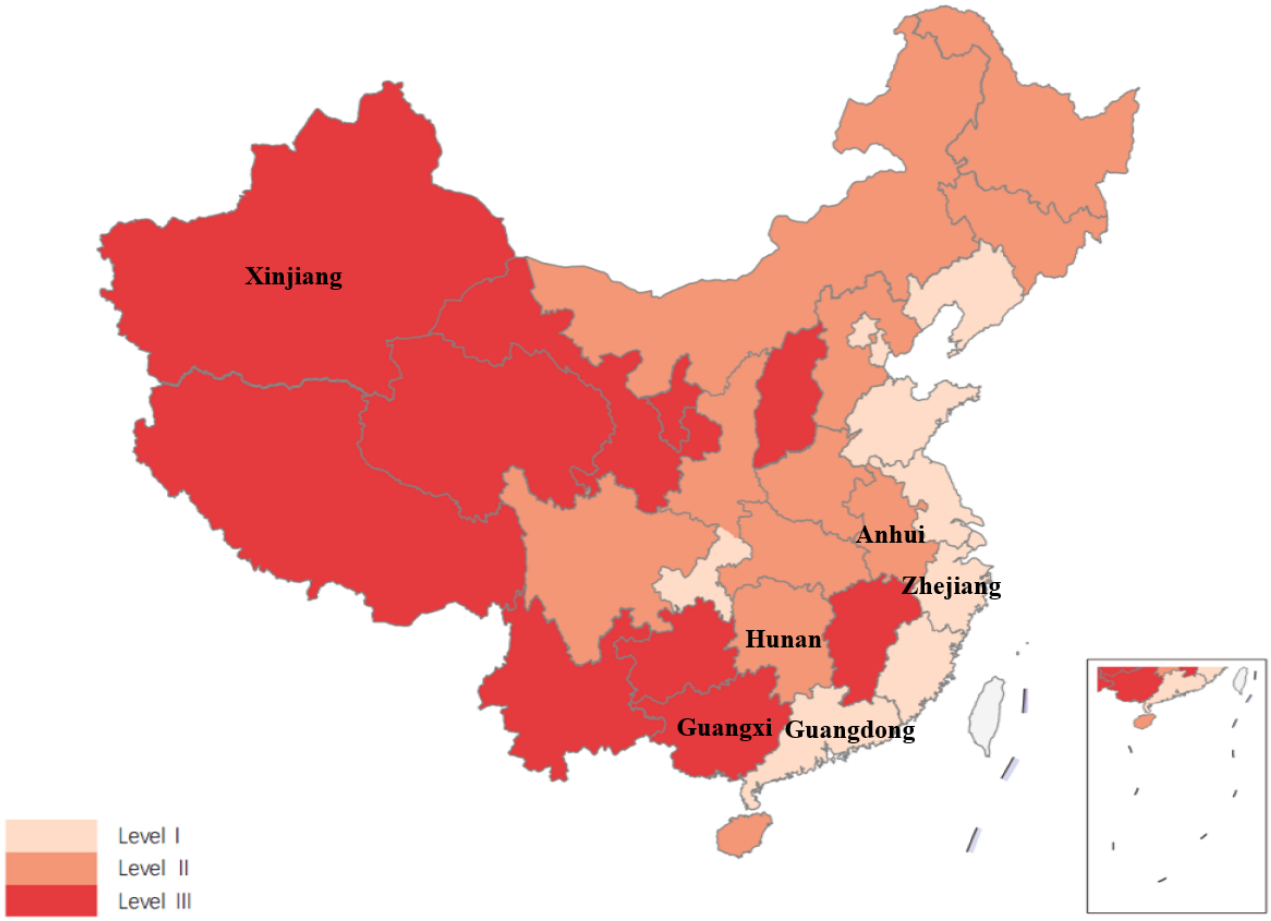

In each provincial capital city, we selected medical institutions that met all our following criteria:

- Public sector platforms: (1) tertiary public hospitals with (2) independent platforms based on offline physical hospitals established to provide DTC telemedicine, (3) dermatology/pediatrics DTC telemedicine services, and (4) asynchronous graphic consultations
- Private sector platforms: (1) DTC telemedicine services on Haodf (one of China's earliest and most popular private online medical platforms) while practicing at physical tertiary public hospitals, (2) dermatology/pediatrics DTC

XSL•FO
RenderX

telemedicine services, and (3) asynchronous graphic consultations

Dermatologists and pediatricians from eligible institutions constituted the sampling frame.

The sample size was calculated according to the following formula:

where σ is the SD from the population, δ is the difference between the 2 groups that is clinically significant, and N is the sample size for each setting. Assuming a 5% type I error (α) and 95% power (1 – β), according to previous research [38], δ=0.15 and σ=0.31, the sample size for 1 setting should reach 110, that is, the visits should be more than 110 times for each setting. Given that some doctors may fail to visit (unable to conduct a second visit, the SP is denied or identified on the spot, etc) or the quality of the records may be low, and other reasons, the sample size was expanded by 35%, so the sample size of this study was 150 times. Therefore, 600 visits for 600 doctors were implemented in this study.

Proportionate sampling was used to determine the weight of each provincial capital city stratum. The weight of the stratum was the proportion of the population contained in that stratum, and the population comprised the permanent residents of the Seventh National Census in each city. The definition of telemedicine areas was based on the location of the hospital where the doctors practice offline. We calculated that 45, 28, 24, 22, 21, and 10 physicians should be included in the provincial capital city in each setting. However, due to the limitation of the total number of pediatricians in the capital city of Guangxi Province on Haodf, only 15 physicians were included in the setting of pediatrics on private sector platforms. A total of 594 physicians were eventually included as respondents in this study. Figure 2 summarizes the sampling procedure.

Figure 2. Sampling procedure. SP: standardized patient.

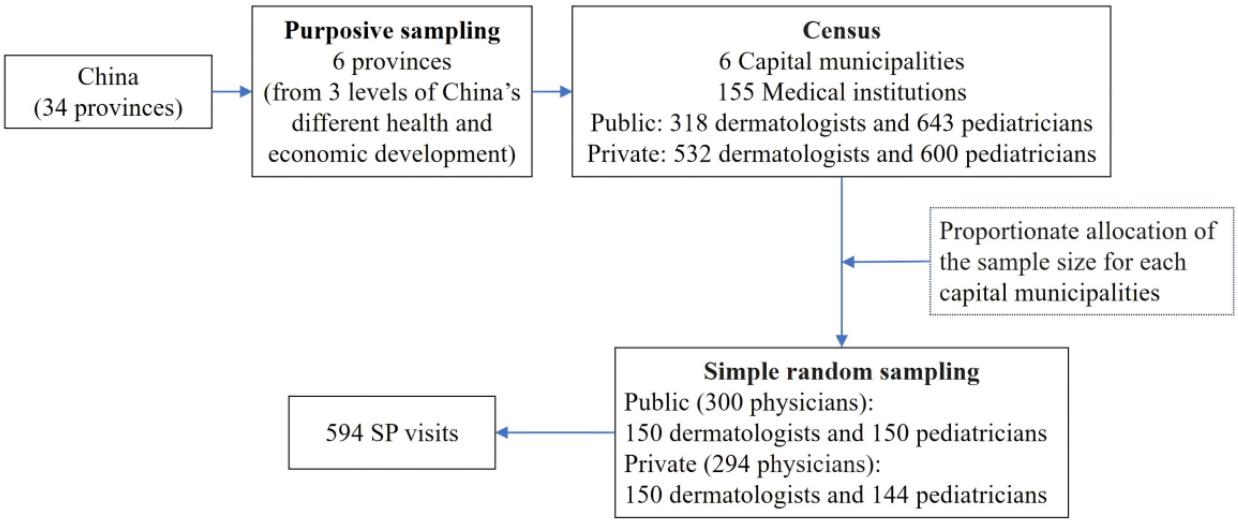

Standardized Patient Procedure

Script and Checklist

We developed a detailed script that consulted doctors and professors with extensive experience for the SPs to use in their visits with the doctors. In both cases, urticaria and childhood diarrhea, preliminary scripts based on established disease guidelines were distributed to 3 pediatricians and 3 dermatologists to improve professionalism and accuracy. Two professors were assigned to each script for fluency and colloquial language. The final scripts covered all possible questions a physician may ask and the SPs' answers during the consultation. Each script included (1) a detailed background story, (2) an opening statement, (3) an illness history presented in question-and-answer format, (4) the language to insist on diagnosis/prescription/lifestyle advice if not given, and (5) a peroration. The checklists of specific cases provide recommended questions that a physician should ask or perform during the interactions (the condensed scripts and checklists are provided in Multimedia Appendix 3).

Recruitment and Training of Standardized Patients

Recruitment and training protocols ensured that the SPs conformed closely to the doctors' regular patient populations. All SPs were recruited from a college following these essential criteria: (1) SPs must be in good physical condition without confounding symptoms, and (2) students of medicine were not recruited as SPs because their medical knowledge and behaviors might affect the diagnosis and treatment of the counterpart physician. In total, 10 SPs were recruited.

A cross-disciplinary team of professors, medical experts, and investigators trained the SPs for 1 week to consistently perform in each encounter according to case scenarios and scripts. This extensive training led to low detection rates of SPs by the sampled doctors. No doctor voiced any suspicions, so the detection rate was, thus, considered 0.

Standardized Patient Visit

The 10 SPs were randomly allocated and visited each physician independently during clinic or nonclinic hours from July 4 to August 5, 2022 (n=5, 50%, SPs describing urticaria and n=5,

XSL•FO
RenderX

50%, SPs describing childhood diarrhea). Each encounter was performed according to the scripts developed. The SPs saved mobile phone screenshots to record the interactions with physicians, allowing us to accurately evaluate interactions without recall bias.

Standardized Patient Exit Survey

SPs completed an exit survey with a structured questionnaire immediately after the visit based on the screenshots to extract information about what each doctor had done, which included (1) a technical quality checklist (adherence to the checklist, defined as the proportion of recommended questions that were asked during the SP consultation, the accuracy of diagnosis, the appropriateness of the prescription, and the lifestyle modification advice), (2) a doctor-patient communication scale (patient-centeredness [PCC]), and (3) a list that recorded the total cost and timeliness of services (time waiting for the first response, time waiting for each response, time for consultation, total number of the doctor's responses, and total words in all of the doctor's responses). The case developers checked the screen recordings to ensure the quality of the SPs' performance and the accuracy of filling out the exit questionnaire at each visit.

Evaluating the Quality of DTC Telemedicine

We measured the quality of DTC telemedicine based on various aspects of the IOM framework (Tables 1 and 2): (1) effectiveness and safety (adherence to the checklist, accuracy of diagnosis, appropriateness of the prescription, and lifestyle modification advice), (2) PCC (following the Patient Perception of Patient-Centeredness [PPPC] rating scale) [39,40], (3) timeliness (time waiting for the first response, time waiting for each response, time for consultation, total number of the doctor's responses, and total words in all of the doctor's responses), and (4) efficiency.

Table 1. Outcome variables used in this study and their definitions.

Variables	Definition	Type	Coding	Source
Effectiveness and safety				
Adherence to the checklist	Adherence to the case-specific checklist in China's national clinical guidelines [41,42]	Continuous	0-1	SP[a] checklist
Accurate diagnosis	Urticaria, not allergy, erythema annulare, popular urticaria, or other diseases (viral diarrhea, infective diarrhea)	Dichotomous	0: no; 1: yes	SP checklist
Appropriate prescription	A definite second-line medication prescribed, with clearly defined course duration and dosage	Dichotomous	0: no; 1: yes	SP checklist
Lifestyle modification advice	Lifestyle modification advice provided	Dichotomous	0: no; 1: yes	SP checklist
PCC[b]				
PCC1: exploring illness experiences	Table S4 in Multimedia Appendix 4	Continuous	1-5	PPPC[c] rating scale
PCC2: understanding the whole person	Table S4 in Multimedia Appendix 4	Continuous	0-3	PPPC rating scale
PCC3: finding the common ground	Table S4 in Multimedia Appendix 4	Continuous	3-17	PPPC rating scale
Timeliness				
Time waiting for the first response	Time interval between the patient's first question and the doctor's first response	Continuous	Minutes	Time record
Time waiting for each response	Average time interval between each patient's question and the doctor's response	Continuous	Minutes	Time record
Time for consultation	Time interval between the patient's first question and the doctor's last response	Continuous	Minutes	Time record
Total number of the doctor's responses	Total number of times of doctor responds in 1 encounter	Continuous	Times	Records
Total words in all of the doctor's responses	Total number of words spoken by the doctor in 1 encounter	Continuous	Words	Records
Efficiency				
Total cost	Charge of each visit	Continuous	RMB[d]	Cost record

[a]SP: standardized patient.

[b]PCC: patient-centeredness.

[c]PPPC: Patient Perception of Patient-Centeredness.

[d]RMB: renminbi.

XSL•FO
RenderX

Table 2. Other variables used in this study and their definitions.

Variables	Definition	Type	Coding	Source
Exposure				
Ownership type	Private and public	Dichotomous	1: public; 2: private	Records
Possible clinician factors influencing quality				
Region	Doctor's practice place	Ordinal	1: level Ⅰ; 2: level II; 3: level III	Records
Gender	Doctor's gender	Dichotomous	0: male; 1: female	Records
Institutional type	Type of institution where the doctor works	Dichotomous	0: traditional Chinese medicine hospital; 1: general hospital	Records
Title	Doctor's title	Ordinal	1: associate chief physician; 2: chief physician; 3: resident physician; 4: attending physician	Records
Possible patient factors influencing quality				
Timing of the visit	Whether SPs[a] visit the doctor during clinic hours	Dichotomous	1: clinic hours; 2: nonclinic hours	Records

[a]SP: standardized patient.

Statistical Analysis

Our analysis unit was the interaction between physicians and SPs. Descriptive statistics identified the sociodemographic characteristics of the final sample. We used the median and interquartile distances to describe continuous variables and used the number and percentage to describe categorical variables. The Wilcoxon rank-sum test was performed to assess the differences in service quality. Ordinary least-squares (OLS) regression models with fixed effects were used for the continuous variables and logistic regression models with fixed effects for the categorical variables, reporting coefficients (β) with the accompanying clustering robust standard error in OLS regressions and the odds ratio (OR) with the accompanying clustering robust standard error in logistic regressions. All regressions were adjusted for critical covariates (region, sex, institution, and title). We performed statistical analyses using Stata version 15 and considered $P<.05$ statistically significant (2-tailed).

Sensitivity Analysis

We recompared the quality of public and private DTC telemedicine services in 2 cases: the first included all interactions contributed by SPs acting as patients with urticaria, and the second included all interactions contributed by SPs acting as patients with childhood diarrhea. The quality of services in public and private DTC telemedicine was assessed for all cases.

Ethical Considerations

In this study, there was no personal information of a single patient because we used SPs to present both urticaria and childhood diarrhea. In addition, our study did not involve the personal information of a single doctor but focused on the overall distribution of the sociodemographic characteristics of a sample of doctors. Informed consent was obtained from all SPs. However, to avoid the Hawthorne effect, the informed consent of physicians was not required for this study. Ethical approval for this study was obtained from the Ethics Committee of Capital Medical University (2021SY074).

Results

Basic Characteristics

Consultations were categorized as effective or ineffective depending on whether the doctor provided a diagnosis, a prescription, or a referral recommendation. In total, 321 effective consultations were included. Of them, interactions of public and private sectors accounted for 54.5% (175/321) and 45.5% (146/321), respectively. In addition, 173 (53.9%) cases of urticaria and 148 (46.1%) cases of childhood diarrhea were analyzed. Furthermore 220 (68.5%) physicians were from general hospitals, most physicians were female (n=207, 64.5%), were located in a level Ⅰ city (n=161, 50.2%), and had the title of attending physician (n=124, 38.6%) or associate chief physician (n=109, 34%). In 169 (52.6%) interactions, SPs visited the physician during clinic hours, and in the other 152 (47.4%) interactions, they visited the physician during nonclinic hours. Interactions varied by case, type of institution, and physician gender between public and private sectors; however, there were no differences in other variables. More details can be found in Table 3.

XSL•FO
RenderX

Table 3. Characteristics of interactions between physicians and SPs[a].

Characteristics	Total (N=321), n (%)	Public sector (n=175), n (%)	Private sector (n=146), n (%)
Case (χ^2_1=4.387, *P*=.04)			
Urticaria	173 (53.9)	85 (48.6)	88 (60.3)
Childhood diarrhea	148 (46.1)	90 (51.4)	58 (39.7)
Region (χ^2_2=1.535, *P*=.46)			
Level Ⅰ	161 (50.2)	93 (53.1)	68 (46.6)
Level Ⅱ	103 (32.1)	54 (30.9)	49 (33.6)
Level Ⅲ	57 (17.8)	28 (16.0)	29 (19.9)
Type of institution (χ^2_2=13.192, *P*=.001)			
General hospital	220 (68.5)	105 (60.0)	115 (78.8)
Specialized hospital	36 (11.2)	26 (14.9)	10 (6.8)
Traditional Chinese medicine hospital	65 (20.2)	44 (25.1)	21 (14.4)
Physician gender (χ^2_2=7.393, *P*=.03)			
Female	207 (64.5)	124 (70.9)	83 (56.8)
Male	95 (29.6)	44 (25.1)	51 (34.9)
Missing	19 (5.9)	7 (4.0)	12 (8.2)
Physician title (χ^2_3=3.493, *P*=.32)			
Chief physician	47 (14.6)	23 (13.1)	24 (16.4)
Associate chief physician	109 (34.0)	61 (34.9)	48 (32.9)
Attending physician	124 (38.6)	73 (41.7)	51 (34.9)
Resident physician	41 (12.8)	18 (10.3)	23 (15.8)
Timing of the visit (χ^2_1=0.414, *P*=.52)			
Clinic hours	169 (52.6)	95 (54.3)	74 (50.7)
Nonclinic hours	152 (47.4)	80 (45.7)	72 (49.3)

[a]SP: standardized patient.

Quality of DTC Telemedicine in Private and Public Sectors

Figures 3-5 and Table S5 in Multimedia Appendix 5 describe the quality of DTC telemedicine provided by the public and private sectors. The median percentage (29.17%) of items completed for checklists in the private sector was higher than that (16.67%) in the public sector (*P*<.001). Furthermore, the effectiveness and safety of consultations provided by private sector platforms were consistently higher than those provided by public hospitals (accurate diagnosis: 91.8% vs 75.4%; appropriate prescription: 33.6% vs 12.0%; lifestyle modification advice: 92.5% vs 73.1%). The median PCC scores were 15 (IQR 11-18) and 18 (IQR 16-20) for the public and the private sector, respectively. The median total number of doctor's responses from the private sector was 5 (IQR 4-6), which was more in than that in the public sector (median 4, IQR 3-5; *P*<.001). The median total number of words in the doctor's responses in the private sector was 160 (IQR 114.00-232.25), which was also more than that in the public sector (median 109, IQR 66-198; *P*<.001). In contrast, public sector physicians had shorter waiting times for the first response than private sector physicians (median 68, IQR 9-222, vs median 118, IQR 22.25-672.75). The median cost of each interaction was lower in the public sector (median 10, IQR 0-30) than in the private sector (median 42, IQR 27-72; *P*<.001).

XSL•FO
RenderX

Figure 3. Effectiveness and safety of DTC telemedicine provided by public and private sectors. DTC: direct to consumer.

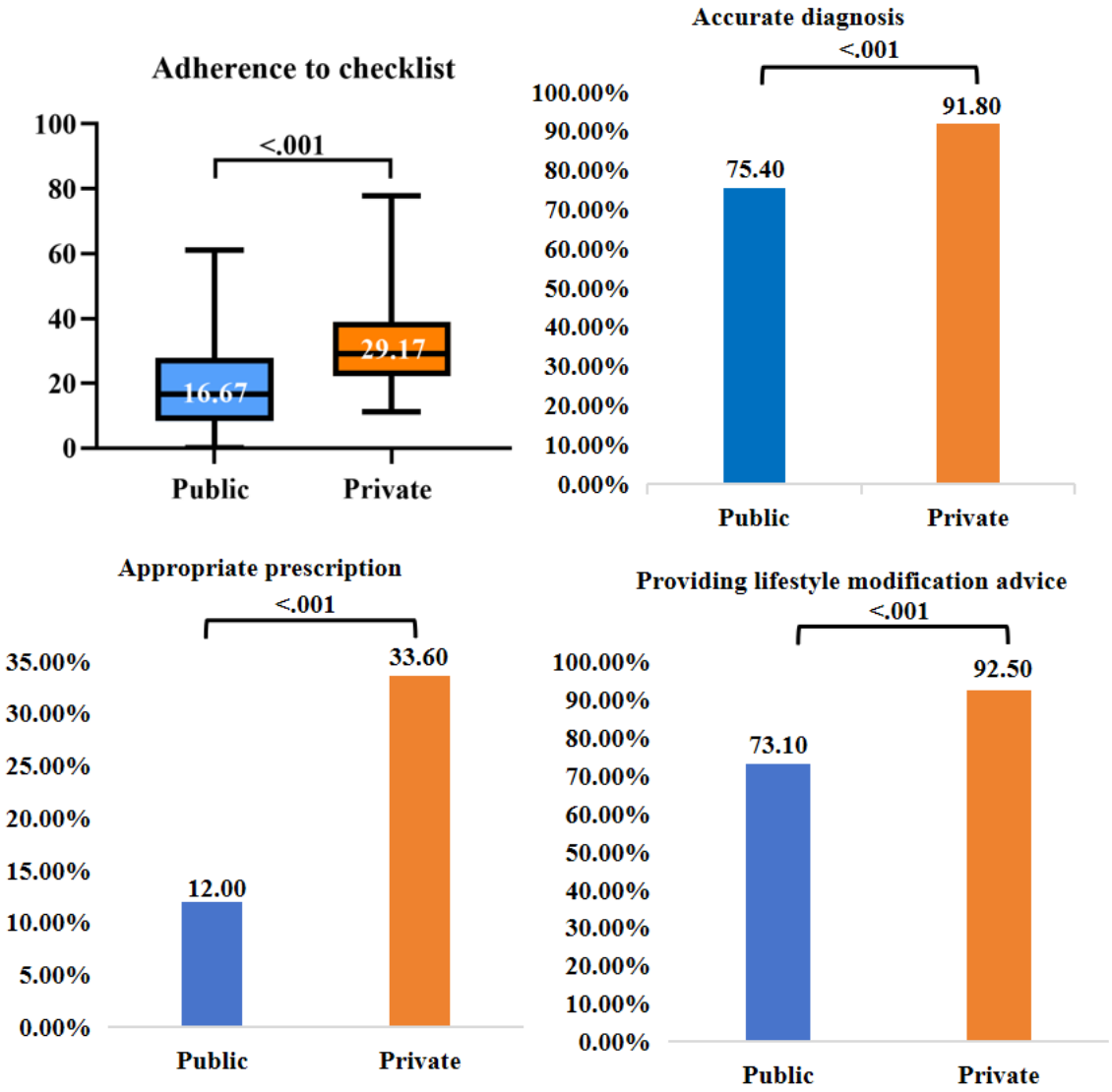

XSL•FO
RenderX

Figure 4. PCC of DTC telemedicine provided by public and private sectors. DTC: direct to consumer; PCC: patient-centeredness.

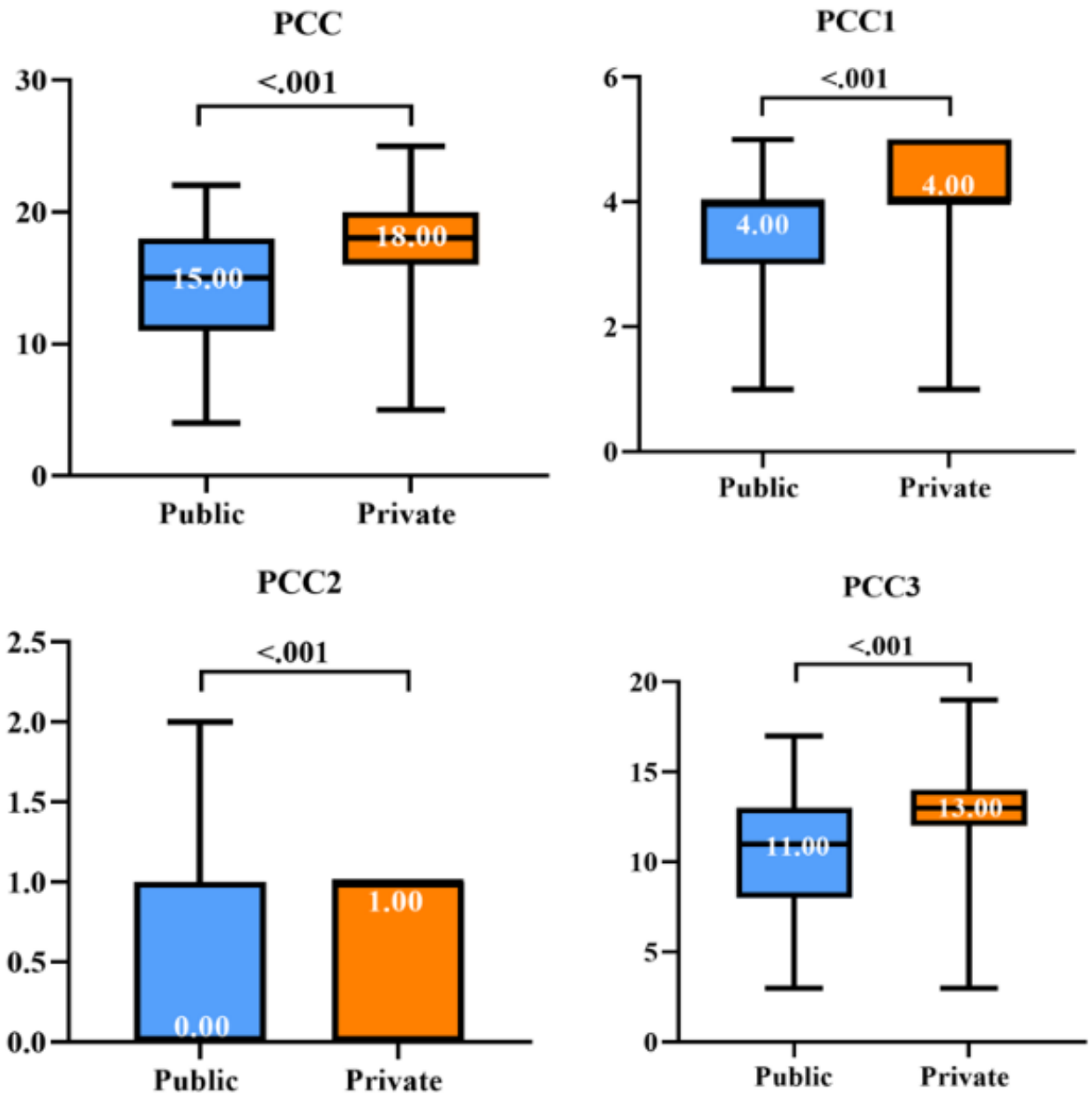

XSL•FO
RenderX

Figure 5. Timeliness and efficiency of DTC telemedicine provided by public and private sectors. DTC: direct to consumer.

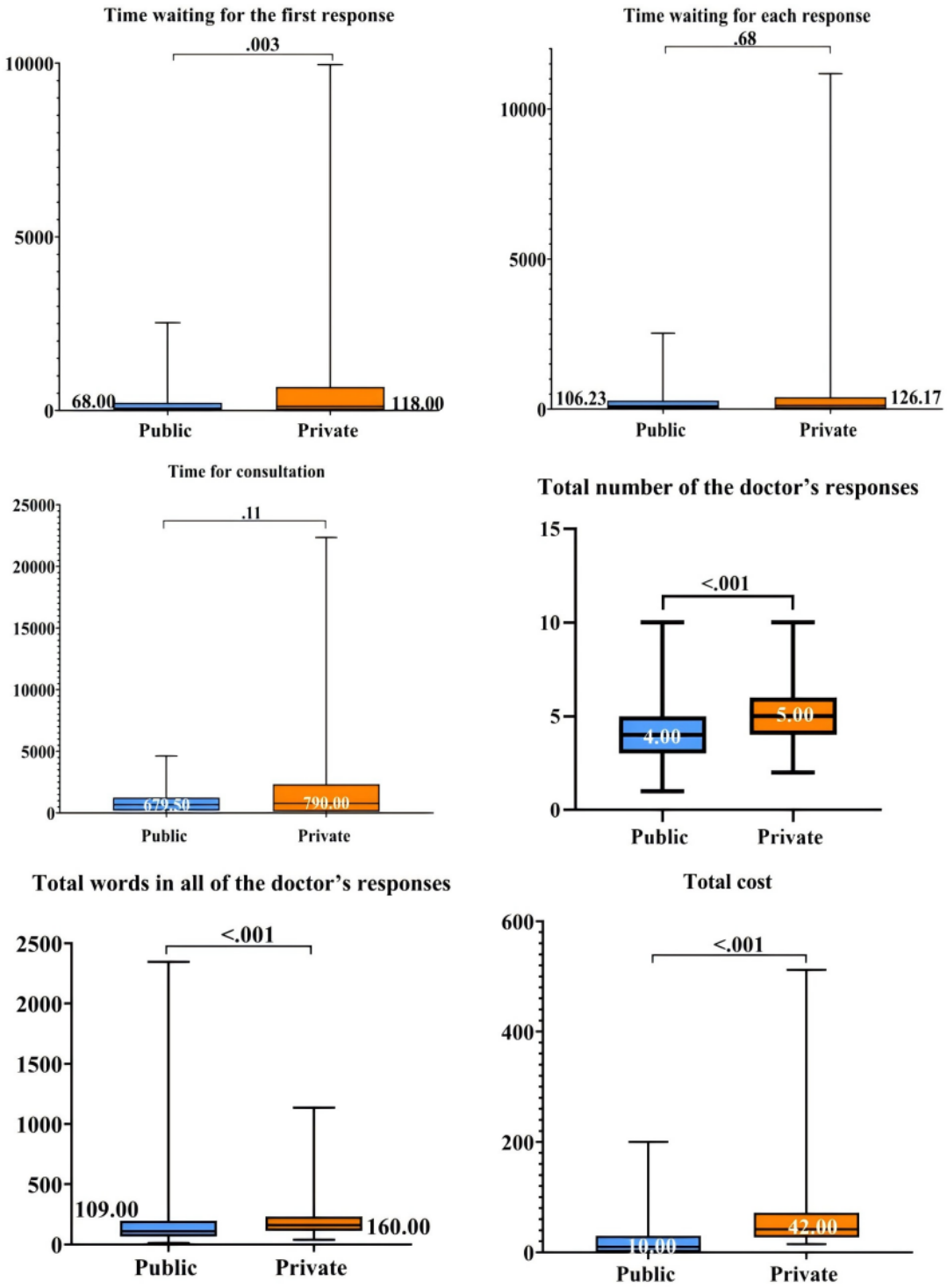

Association Between the Models and the Quality of DTC Telemedicine

Table 4 shows the association between the models and the quality of DTC telemedicine. After adjustments were made for case, region, type of institution, physician gender, physician title, and timing of the visit, significant quality differences were observed between the public and private sectors. The results showed that physicians from private platforms were significantly more likely to adhere to clinical checklists (adjusted β 15.22, *P*<.001); provide an accurate diagnosis (adjusted OR 3.85, *P*<.001), an appropriate prescription (adjusted OR 3.87, *P*<.001), and lifestyle modification advice (adjusted OR 6.82, *P*<.001); ensure more PCC (adjusted β 3.34, *P*<.001); and spend more time with SPs (adjusted β 839.70, *P*<.001), with more responses (adjusted β 1.33, *P*=.001) and more words (adjusted β 50.93, *P*=.009). However, SPs on private platforms waited longer for the first response (adjusted β 505.87, *P*=.001) and each response

XSL•FO
RenderX

(adjusted β 168.33, *P*=.04) and paid more for the average visit (adjusted β 40.03, *P*<.001).

Table 4. Effects of models on the quality of DTC[a] telemedicine[b].

Characteristics	Public sector adjusted β/OR[c] (SE)	Private sector adjusted β/OR (SE)	*P* value
Effectiveness and safety			
Adherence to the checklist	Reference	15.22 (1.63)	<.001
Accurate diagnosis	Reference	3.85 (1.46)	<.001
Appropriate prescription	Reference	3.87 (1.41)	<.001
Lifestyle modification advice	Reference	6.82 (2.57)	<.001
PCC[d]			
PCC	Reference	3.34 (0.34)	<.001
PCC1	Reference	0.60 (0.08)	<.001
PCC2	Reference	0.33 (0.14)	.04
PCC3	Reference	2.41 (0.34)	<.001
Timeliness			
Time waiting for the first response	Reference	505.87 (95.85)	.001
Time waiting for each response	Reference	168.33 (68.89)	.04
Time for consultation	Reference	839.70 (151.31)	<.001
Total number of the doctor's responses	Reference	1.33 (0.26)	.001
Total words in all of the doctor's responses	Reference	50.93 (15.30)	.009
Efficiency			
Total cost	Reference	40.03 (4.12)	<.001

[a]DTC: direct to consumer.

[b]The variables were controlled as follows: case, region, type of institution, physician gender, physician title, and timing of the visit, and SE was the clustering robust standard error.

[c]OR: odds ratio.

[d]PCC: patient-centeredness.

Sensitivity Analysis Outcomes

Table S6 in Multimedia Appendix 6 shows the sensitivity analysis results for urticaria. The results of the regressions were close to our initial analysis results, except for the time waiting for each response (*P*=.10) and the total words in all of the doctor's responses (*P*=.10), reporting no statistical effect.

Table S7 in Multimedia Appendix 6 shows the sensitivity analysis results for childhood diarrhea. The results of the regressions differed from the entire sample in prescriptions (*P*=.10), time waiting for the first response (*P*=.07), time waiting for each response (*P*=.44), and time for consultation (*P*=.07), presenting no statistical effect.

Discussion

Principal Findings

This study is the first large sample of evidence to evaluate and compare the quality of DTC telemedicine between public and private health care providers using an SP approach. Previous studies have suggested that quality studies have rarely been based on large random samples, sufficient numbers of tracer conditions, and gold-standard assessment tools [32]. Our study conducted a more thorough and valid comparison between private and public health care providers, with an adequately powered and representative sample of doctors. We found the private sector fares better regarding technical quality indicators, such as adherence to clinical checklists, PCC, and total numbers of and words in doctors' responses. However, the public sector performed better in the time waiting for the first response and the total cost. The results could provide a strategic reference for improving the quality of public and private DTC telemedicine in China.

Advantages of Private DTC Telemedicine Quality

Compared to public DTC telemedicine, private doctors were more likely to adhere to clinical checklists, give a higher proportion of accurate diagnoses and appropriate prescriptions, and provide lifestyle modification advice. This aligns with a prior study comparing the quality of tuberculosis care in offline public and private health care facilities in Mumbai, India, where private doctors asked more medical history questions and were more likely to complete more clinical checklists and history collection programs [35]. However, it contradicts the findings of 2 studies comparing primary health care services sectors in China [43]. The differences in study conclusions could be attributed to discrepancies in study settings, diseases, and

XSL•FO
RenderX

physician populations. Reliability and ease of use are critical for physicians in telemedicine [44]. DTC telemedicine in the public sector in China started later than in the private sector, increasing mainly during the COVID-19 pandemic. The incomplete construction and mechanism of DTC telemedicine platforms in the public sector may affect the doctors' experience, ultimately hindering their efforts in consultations. Doctors who exert more effort perform more clinical checklist recommendations and diagnose more precisely and can also avoid unnecessary care [45]. However, more empirical studies are required to test specific reasons.

Furthermore, private DTC telemedicine providers exhibited greater PCC than those in the public sector, contradicting a study on community health centers [43]. A possible interpretation is that compensation mechanisms, education levels, and relevant rules and regulations are all factors that affect doctors' behavior [46,47]. Reward and punishment systems, the professional environment, and reactions to the market competition also impact doctors' behavior. DTC telemedicine in the private sector amplifies the role of a doctor's brand in attracting patients. The private sector has established a sound and standardized postvisit evaluation mechanism, which displays a doctor's comprehensive score and online reviews on the doctor's homepage. Online reviews create an easily accessible digital word of mouth for patient-perceived care quality [48]. A recent national survey in the United States showed that 59% of participants referred to online reviews when choosing a physician [49]. Patients' ratings of doctors depend on whether they have received tangible help from those doctors and have had an excellent medical experience. Therefore, doctors who explore more patients' disease experiences, family and social backgrounds, and doctor-patient consensus can obtain high scores and build a good brand and reputation to attract more patients.

In addition, responses from private sector doctors were more detailed and frequent, meaning they provided more numbers of and words in responses. A large proportion of the difference may be due to financial incentives tied to service prices. Internet medical service items and prices are important factors affecting the patient acceptance and participation enthusiasm of doctors and medical institutions [50]. The lower cost of DTC telemedicine in public sectors leads to lower motivation. The cost of a single consultation on the Haodf platform belongs to the doctor in full [51]. It allows pricing according to the quality of service, which promotes the doctor's active service provision behavior. Doctors provide better online services in response to patients' financial efforts [52].

The public sector should learn some practices from the private sector. First, a standardized postvisit evaluation mechanism should be established to show patients' disease experience and encourage doctors to be more patient centered. Second, corresponding norms and price standards for DTC telemedicine should be formulated, and the performance incentive evaluation mechanism should be improved so that doctors with different titles can fully realize their labor value. One hospital in Shenzhen pioneered a successful strategy of holding weekly meetings, while using DTC telemedicine fees as a performance reward to increase doctors' enthusiasm [53].

Advantages of Public DTC Telemedicine Quality

The time waiting for the first response differed significantly between the 2 sectors, with the public sector performing better. One possible explanation is that nearly 60% of public sector consultations require doctors to receive consultations within 24 hours and about 22% require consultations within 48 hours, while in the private sector, doctors can respond to graphic consultations initiated by patients after 1 week. This results in a shorter wait time for the public sector doctor's first response (ie, a timelier response). Establishing regulations to govern the response time for initial DTC telemedicine consultations is crucial in ensuring a prompt reply. For instance, health care professionals must respond within 48 hours.

The private sector had a higher total cost than public DTC telemedicine. This may be due to the different pricing strategies of the 2 models. The public sector primarily operates as a nonprofit and adheres to government-guided fee schedules, with some institutions even providing free consultation services. In this study, free consultation accounted for 28.6% (50/175) of effective consultations in the public sector. The private sector mainly operates for profit, implementing market-regulated prices, where the cost of a single consultation can reach up to Chinese yuan (CNY) 512 (~US $72).

Disadvantages of Public and Private DTC Telemedicine Quality

We further analyzed several poor aspects of DTC telemedicine services. DTC telemedicine had poor rates of appropriate prescribing and adherence to clinical checklists, at 21.8% and 25%, respectively, with a low PCC2 score. Previous studies have found that although most physicians acknowledge the positive effects of guidelines, understanding and adhering to the guidelines still need to be improved. Factors related to guidelines, individuals, patients, organizations, and the environment conspire to contribute to noncompliance [54]. The results of this study reflect that most prescriptions do not have the necessary 3 elements (definite second-line medication, clearly defined course duration, and dosage). One interpretation is that DTC telemedicine may exacerbate uncertainty in doctors' understanding of patients' illnesses. To avoid risk, doctors may make vague prescribing recommendations [55]. Regarding PCC2, doctors are busy, pay more attention to diseases when facing patients, and make judgments on time. The separation of the biological and social nature of patients causes doctors to pay little attention to the social background of patients [56].

These findings highlight several strategies that may help improve the quality of DTC telemedicine. On the one hand, it is essential that organizations, especially public DTC telemedicine institutions, add advisory templates with clinical guidelines, such as significant symptoms and dosage, on the platforms. These can also help save time because doctors complete e-consults outside their commonly designated work [57]. Doctors argue that using the current system for documentation is complex and inefficient and creates a time burden. On the other hand, public and private sectors can reestablish scientific and reasonable doctor scheduling and collect information on the time distribution when patients initiate consultations. For instance, a hospital in western China

XSL•FO
RenderX

discovered online prescription searches had 2 distinct peaks at 9:00 AM and 8:00 PM [58].

Limitations

This study should be interpreted in the context of several potential limitations. Due to the study area, the conclusion may need more generalizability. First, we only carried out an SP survey on childhood diarrhea and urticaria. Both diseases do not require invasive examinations. Considering that the diagnosis and treatment of each disease are quite different, the study's conclusion may not be extrapolated to other diseases. Second, even though we surveyed 6 cities in the context of DTC telemedicine, we still need to be cautious about generalizing the findings to the overall quality of DTC telemedicine. Moreover, our research was conducted only on 1 of the earliest and most popular private platforms in China. We will experiment with a broader sample of health care systems to address some of these challenges.

Conclusion

This study is the first large sample of evidence to evaluate and compare the quality of DTC telemedicine between public and private sector doctors using the SP approach in China. The study added evidence suggesting that comparisons of the quality of DTC telemedicine services provided by the public and private sectors need to consider different quality indicators. Compared with public DTC telemedicine, private DTC telemedicine providers are more likely to adhere to clinical checklists; providing a higher proportion of accurate diagnoses, appropriate prescriptions, and lifestyle modification advice; and being more patient centered. However, DTC telemedicine in the public sector has lower costs and faster responses. Bearing in mind the findings, this study has highlighted several strategies that might be helpful to improve the quality of DTC telemedicine. First, DTC telemedicine in the public sector should make efforts to enhance the patient experience by developing online reviews. Formulating corresponding norms and price standards for DTC telemedicine and improving the performance incentive evaluation mechanism is essential. Second, private DTC telemedicine should more strictly limit the first response time of health service providers. Finally, organizations should add advisory templates with clinical guidelines, such as significant symptoms and dosage, on online platforms in public and private DTC telemedicine. In addition, they should reestablish scientific and reasonable doctor scheduling and collect information on the time distribution when patients initiate consultations.

Acknowledgments

We would like to thank all the participants for their support in developing this paper. This study was supported by the National Natural Science Foundation of China (grant number 72174131) and the R&D Program of the Beijing Municipal Education Commission (grant number SZ202210025010).

Authors' Contributions

FYS, XG, and RG designed and conducted the research. FYS and XG performed the analysis and interpretation of data and drafted the initial manuscript. YTY participated in the standardized patient survey and provided support for submission. RG critically reviewed and revised the final manuscript. All authors have approved the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Selection conditions and characteristics of SP cases. SP: standardized patient.
[DOCX File , 21 KB - jmir_v26i1e55400_app1.docx]

Multimedia Appendix 2
Level and ranking of provinces.
[DOCX File , 15 KB - jmir_v26i1e55400_app2.docx]

Multimedia Appendix 3
Condensed scripts and checklists.
[DOCX File , 24 KB - jmir_v26i1e55400_app3.docx]

Multimedia Appendix 4
Scoring method of patient-centeredness (PCC).
[DOCX File , 13 KB - jmir_v26i1e55400_app4.docx]

Multimedia Appendix 5
Quality of DTC telemedicine provided by public and private sectors. DTC: direct to consumer.

XSL•FO
RenderX

[DOCX File , 16 KB - jmir_v26i1e55400_app5.docx]

Multimedia Appendix 6
Effect of models on the quality of DTC telemedicine for different case. DTC: direct to consumer.
[DOCX File , 23 KB - jmir_v26i1e55400_app6.docx]

References

1. Tuckson RV, Edmunds M, Hodgkins ML. Telehealth. N Engl J Med 2017 Oct 19;377(16):1585-1592. [doi: 10.1056/NEJMsr1503323] [Medline: 29045204]
2. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: Capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
3. Liu L, Shi L. Chinese Patients' Intention to Use Different Types of Internet Hospitals: Cross-sectional Study on Virtual Visits. J Med Internet Res 2021 Aug 13;23(8):e25978 [FREE Full text] [doi: 10.2196/25978] [Medline: 34397388]
4. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020 Jul 29;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
5. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of Internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021 Jul 20;23(7):e22330 [FREE Full text] [doi: 10.2196/22330] [Medline: 34283025]
6. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 01;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
7. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
8. Digital China development report 2022. Cyberspace Administration of China. 2023. URL: https://www.cac.gov.cn/2023-05/22/c_1686402318492248.htm [accessed 2024-10-28]
9. Miao W, Yue H, Kunkun Z, Da H, Li Y, Guangjun Y. Comparison on operation model of internet hospitals in China. Chin Health Resour 2020;23(02):110-113. [doi: 10.13688/j.cnki.chr.2020.20240]
10. Wanmin Lian: how did the country's first internet hospital become. China Hospital Association Information Committee. 2019. URL: https://www.chima.org.cn/Html/News/Articles/4848.html [accessed 2024-10-28]
11. Internet hospital of the first affiliated hospital of Zhejiang university opens today, the first public hospital sponsored internet hospital in China. People's Daily Online. URL: http://health.people.com.cn/n1/2016/0217/c398004-28131507.html [accessed 2024-10-28]
12. State Council issues guideline on internet plus healthcare. State Council of China. URL: https://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm?trs=1 [accessed 2024-10-28]
13. Measures for the administration of Internet diagnosis and treatment (for trial implementation). National Health Commission of the People's Republic of China. URL: https://www.gov.cn/gongbao/content/2019/content_5358684.htm [accessed 2024-10-28]
14. China internet hospital development report. National Center for Telemedicine and Internet Medicine. 2022. URL: https://zk.cn-healthcare.com/label-show-346.html [accessed 2024-10-28]
15. Hansmann HB. The role of nonprofit enterprise. Yale Law J 1980 Apr;89(5):835. [doi: 10.2307/796089]
16. Patouillard E, Goodman CA, Hanson KG, Mills AJ. Can working with the private for-profit sector improve utilization of quality health services by the poor? A systematic review of the literature. Int J Equity Health 2007 Nov 07;6:17 [FREE Full text] [doi: 10.1186/1475-9276-6-17] [Medline: 17988396]
17. Gaynor M, Town RJ. Competition in health care markets. In: Pauly MV, Mcguire TG, Barros PP, editors. Handbook of Health Economics Volume 2. Amsterdam, the Netherlands: Elsevier; 2011:499-637.
18. Eggleston K, Lu M, Li C, Wang J, Yang Z, Zhang J, et al. Comparing public and private hospitals in China: evidence from Guangdong. BMC Health Serv Res 2010 Mar 23;10:76 [FREE Full text] [doi: 10.1186/1472-6963-10-76] [Medline: 20331886]
19. Eggleston K, Shen Y, Lau J, Schmid CH, Chan J. Hospital ownership and quality of care: what explains the different results in the literature? Health Econ 2008 Dec;17(12):1345-1362. [doi: 10.1002/hec.1333] [Medline: 18186547]
20. Chou S. Asymmetric information, ownership and quality of care: an empirical analysis of nursing homes. J Health Econ 2002 Mar;21(2):293-311. [doi: 10.1016/s0167-6296(01)00123-0] [Medline: 11939243]
21. Wei X, Yin J, Wong SYS, Griffiths SM, Zou G, Shi L. Private ownership of primary care providers associated with patient perceived quality of care: a comparative cross-sectional survey in three big Chinese cities. Medicine (Baltimore) 2017 Jan;96(1):e5755 [FREE Full text] [doi: 10.1097/MD.0000000000005755] [Medline: 28072718]

XSL•FO
RenderX

22. Basu S, Andrews J, Kishore S, Panjabi R, Stuckler D. Comparative performance of private and public healthcare systems in low- and middle-income countries: a systematic review. PLoS Med 2012;9(6):e1001244 [FREE Full text] [doi: 10.1371/journal.pmed.1001244] [Medline: 22723748]
23. Berendes S, Heywood P, Oliver S, Garner P. Quality of private and public ambulatory health care in low and middle income countries: systematic review of comparative studies. PLoS Med 2011 Apr;8(4):e1000433 [FREE Full text] [doi: 10.1371/journal.pmed.1000433] [Medline: 21532746]
24. Coarasa J, Das J, Gummerson E, Bitton A. A systematic tale of two differing reviews: evaluating the evidence on public and private sector quality of primary care in low and middle income countries. Glob Health 2017 Apr 12;13(1):24 [FREE Full text] [doi: 10.1186/s12992-017-0246-4] [Medline: 28403871]
25. Kruse FM, Stadhouders NW, Adang EM, Groenewoud S, Jeurissen PPT. Do private hospitals outperform public hospitals regarding efficiency, accessibility, and quality of care in the European Union? A literature review. Int J Health Plann Manag 2018 Apr;33(2):e434-e453 [FREE Full text] [doi: 10.1002/hpm.2502] [Medline: 29498430]
26. Song T, Deng N, Cui T, Qian S, Liu F, Guan Y, et al. Measuring success of patients' continuous use of mobile health services for self-management of chronic conditions: model development and validation. J Med Internet Res 2021 Jul 13;23(7):e26670 [FREE Full text] [doi: 10.2196/26670] [Medline: 34255685]
27. Meng G, McAiney C, Perlman CM, McKillop I, Tisseverasinghe T, Chen HH. Service process factors affecting patients' and clinicians' experiences on rapid teleconsultation implementation in out-patient neurology services during COVID-19 pandemic: a scoping review. BMC Health Serv Res 2022 Apr 22;22(1):534 [FREE Full text] [doi: 10.1186/s12913-022-07908-4] [Medline: 35459134]
28. Resneck JS, Abrouk M, Steuer M, Tam A, Yen A, Lee I, et al. Choice, transparency, coordination, and quality among direct-to-consumer telemedicine websites and apps treating skin disease. JAMA Dermatol 2016 Jul 01;152(7):768-775. [doi: 10.1001/jamadermatol.2016.1774] [Medline: 27180232]
29. Gong X, Hou M, Guo R, Feng XL. Investigating the relationship between consultation length and quality of tele-dermatology e-consults in China: a cross-sectional standardized patient study. BMC Health Serv Res 2022 Sep 22;22(1):1187 [FREE Full text] [doi: 10.1186/s12913-022-08566-2] [Medline: 36138410]
30. Xu D, Pan J, Dai X, Hu M, Cai Y, He H, et al. Comparing quality of primary healthcare between public and private providers in China: study protocol of a cross-sectional study using unannounced standardised patients in seven provinces of China. BMJ Open 2021 Jan 12;11(1):e040792 [FREE Full text] [doi: 10.1136/bmjopen-2020-040792] [Medline: 33436467]
31. Institute of Medicine Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
32. Das J, Holla A, Das V, Mohanan M, Tabak D, Chan B. In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. Health Aff (Millwood) 2012 Dec;31(12):2774-2784 [FREE Full text] [doi: 10.1377/hlthaff.2011.1356] [Medline: 23213162]
33. Das V, Daniels B, Kwan A, Saria V, Das R, Pai M, et al. Simulated patients and their reality: an inquiry into theory and method. Soc Sci Med 2022 May;300:114571 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114571] [Medline: 34865913]
34. Kwan A, Daniels B, Bergkvist S, Das V, Pai M, Das J. Use of standardised patients for healthcare quality research in low- and middle-income countries. BMJ Glob Health 2019 Sep 12;4(5):e001669 [FREE Full text] [doi: 10.1136/bmjgh-2019-001669] [Medline: 31565413]
35. Daniels B, Shah D, Kwan AT, Das R, Das V, Puri V, et al. Tuberculosis diagnosis and management in the public versus private sector: a standardised patients study in Mumbai, India. BMJ Glob Health 2022 Oct;7(10):e009657 [FREE Full text] [doi: 10.1136/bmjgh-2022-009657] [Medline: 36261230]
36. King JJC, Das J, Kwan A, Daniels B, Powell-Jackson T, Makungu C, et al. How to do (or not to do) … using the standardized patient method to measure clinical quality of care in LMIC health facilities. Health Policy Plan 2019 Oct 01;34(8):625-634 [FREE Full text] [doi: 10.1093/heapol/czz078] [Medline: 31424494]
37. China Health Statistical Yearbook. National Health Commission of the People's Republic of China. 2023. URL: http://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202305/6ef68aac6bd14c1eb9375e01a0faa1fb.shtml [accessed 2024-10-28]
38. Liu L, Peng J, Zeng Z, Li Q. Comparison of healthcare quality between virtual visits and office visits: based on standardized patients with postpartum depression. ResearchGate. 2021. URL: https://www.researchgate.net/publication/351226047_Comparison_of_healthcare_quality_between_virtual_visits_and_office_visits_based_on_standardized_patients_with_postpartum_depression [accessed 2024-10-28]
39. Assessing communication between patients and physicians: a manual for scoring patient-centred communication #95-2. Centre for Studies in Family Medicine. 1995. URL: https://www.schulich.uwo.ca/familymedicine/research/csfm/publications/working_papers/assessing_communication_between_patients_and_physicians_a_manual_for_scoring_patientcentred_communication.html [accessed 2024-10-28]
40. The patient perception of patient centeredness questionaire (PPPC) #04-1. Centre for Studies in Family Medicine. 2004. URL: https://www.schulich.uwo.ca/familymedicine/research/csfm/publications/working_papers/the%20patient%20perception%20of%20patient%20centerdness%20questionnaire_pppc.html [accessed 2024-10-28]

XSL•FO
RenderX

41. Chinese guidelines for the diagnosis and treatment of urticaria: 2018 update. Center for Urticaria Research. 2020 Mar 18. URL: https://journals.lww.com/ijdv/fulltext/2020/03000/chinese_guidelines_for_the_diagnosis_and_treatment.2.aspx [accessed 2024-10-28]
42. Guidelines for diagnosis and treatment of acute infectious diarrheal disease in children. National Health Commission of the People's Republic of China. 2020. URL: http://www.nhc.gov.cn/yzygj/s7659/202009/5c03bafd1db74fb68e2a74afa2ed08c1.shtml [accessed 2024-10-28]
43. Su M, Zhou Z, Si Y, Sylvia S, Chen G, Su Y, et al. Comparing the quality of primary care between public and private providers in urban China: a standardized patient study. Int J Environ Res Public Health 2021 May 11;18(10):5060 [FREE Full text] [doi: 10.3390/ijerph18105060] [Medline: 34064733]
44. Xiong X, Luo L, Zhou S, Li VJ, Zhou Y, Huo Z. A profile of patients' and doctors' perceptions, acceptance, and utilization of e-health in a deprived region in southwestern China. PLOS Digit Health 2023 Apr 25;2(4):e0000238 [FREE Full text] [doi: 10.1371/journal.pdig.0000238] [Medline: 37098005]
45. King JJC, Powell-Jackson T, Hargreaves J, Makungu C, Goodman C. Does increased provider effort improve quality of care? Evidence from a standardised patient study on correct and unnecessary treatment. BMC Health Serv Res 2023 Feb 23;23(1):190 [FREE Full text] [doi: 10.1186/s12913-023-09149-5] [Medline: 36823637]
46. Sun W, Yan M. Study the measures of regulating the prescription behavior of doctors in China by referring to foreign experience. China Health Stand Manag 2017;8(08):17-20. [doi: 10.3969/j.issn.1674-9316.2017.08.010]
47. Huang H, Yu C, Yu T, Liu S, Zhang N, Chen L. Problems and countermeasures of physician multi-sited license based on the grounded theory: a perspective of the doctors. Chin J Health Policy 2018;11(01):42-47. [doi: 10.3969/j.issn.1674-2982.2018.01.007]
48. Xie Y, He W, Wan Y, Luo H, Cai Y, Gong W, et al. Validity of patients' online reviews at direct-to-consumer teleconsultation platforms: a protocol for a cross-sectional study using unannounced standardised patients. BMJ Open 2023 May 10;13(5):e071783 [FREE Full text] [doi: 10.1136/bmjopen-2023-071783] [Medline: 37164474]
49. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. Public awareness, perception, and use of online physician rating sites. JAMA 2014 Feb 19;311(7):734-735. [doi: 10.1001/jama.2013.283194] [Medline: 24549555]
50. Ding T, Qian A, Tan Z. Comparative analysis of domestic and foreign internet medical service pricing and medical insurance payment policies. Chin Hosp 2022;26(09):10-13. [doi: 10.19660/j.issn.1671-0592.2022.9.03]
51. Hang Wang: how do Haodf.com make money. Economic Observer. 2023. URL: https://finance.sina.com.cn/wm/2023-05-27/doc-imyvfcic6822054.shtml [accessed 2024-10-28]
52. Wang Y, Wu H, Xia C, Lu N. Impact of the price of gifts from patients on physicians' service quality in online consultations: empirical study based on social exchange theory. J Med Internet Res 2020 May 05;22(5):e15685 [FREE Full text] [doi: 10.2196/15685] [Medline: 32369028]
53. Inquiry internet hospital: how is the incentive mechanism for doctors to "go online"? Workers' Daily. 2022. URL: https://politics.gmw.cn/2022-07/04/content_35857279.htm [accessed 2024-10-28]
54. Wang D, Huang Q, Huang X, Jin Y, Wang Y, Shi Y, et al. Knowledge of and compliance with guidelines in the management of non-muscle-invasive bladder cancer: a survey of Chinese urologists. Front Oncol 2021;11:735704 [FREE Full text] [doi: 10.3389/fonc.2021.735704] [Medline: 34778048]
55. Liu Y, Ren C, Shi D, Li K, Zhang X. Evaluating the social value of online health information for third-party patients: Is uncertainty always bad? Inf Proc Manag 2020 Sep;57(5):102259 [FREE Full text] [doi: 10.1016/j.ipm.2020.102259]
56. Rational understanding of doctor's role and profession. Guangming Daily. 2021. URL: https://epaper.gmw.cn/gmrb/html/2021-04/08/nw.D110000gmrb_20210408_2-02.htm [accessed 2024-10-28]
57. Bauer BS, Nguyen-Phan AL, Ong MK, Ziaeian B, Nguyen K. Cardiology electronic consultations: efficient and safe, but consultant satisfaction is equivocal. J Telemed Telecare 2020 Jul;26(6):341-348 [FREE Full text] [doi: 10.1177/1357633X19828130] [Medline: 30803322]
58. Yang B, Wen Q, Zhang Y, Wang X, Yin X, Li Q, et al. Economic value and characteristics of cloud pharmacy for children based on internet hospital in western China during the COVID-19 pandemic: cross-sectional survey study. Front Public Health 2022;10:1034450 [FREE Full text] [doi: 10.3389/fpubh.2022.1034450] [Medline: 36408041]

Abbreviations

DTC: direct to consumer
IOM: Institute of Medicine
OLS: ordinary least squares
OR: odds ratio
PCC: patient-centeredness
PPPC: Patient Perception of Patient-Centeredness
SP: standardized patient

XSL•FO
RenderX

Edited by A Schwartz; submitted 12.12.23; peer-reviewed by Y Si, L Liu; comments to author 01.04.24; revised version received 20.05.24; accepted 07.10.24; published 14.11.24.

Please cite as:
Song F, Gong X, Yang Y, Guo R
Comparing the Quality of Direct-to-Consumer Telemedicine Dominated and Delivered by Public and Private Sector Platforms in China: Standardized Patient Study
J Med Internet Res 2024;26:e55400
URL: https://www.jmir.org/2024/1/e55400
doi:10.2196/55400
PMID:

©Faying Song, Xue Gong, Yuting Yang, Rui Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telemonitoring of Active Inflammatory Bowel Disease Using the App TECCU: Short-Term Results of a Multicenter Trial of GETECCU

Mariam Aguas[1,2*], MD, PhD; Javier Del Hoyo[1,2*], MD, PhD[‡]; Raquel Vicente[3], MD; Manuel Barreiro-de Acosta[4], MD, PhD; Luigi Melcarne[5], MD; Alejandro Hernandez-Camba[6], MD, PhD; Lucía Madero[7,8], MD; María Teresa Arroyo[9], MD, PhD; Beatriz Sicilia[10], MD, PhD; María Chaparro[8,11], MD, PhD; María Dolores Martin-Arranz[12,13], MD, PhD; Ramón Pajares[14], MD; Francisco Mesonero[15], MD; Miriam Mañosa[8,16], MD, PhD; Pilar Martinez[17], MD, PhD; Silvia Chacón[18], MD, PhD; Joan Tosca[19], MD; Sandra Marín[20], MD; Luciano Sanroman[21], MD; Marta Calvo[22], MD, PhD; David Monfort[23], MD; Empar Saiz[24], MD; Yamile Zabana[8,25], MD, PhD; Ivan Guerra[26], MD, PhD; Pilar Varela[27], MD; Virginia Baydal[1], RNC; Raquel Faubel[28,29], PhD; Pilar Corsino[3], RNC, PhD; Sol Porto-Silva[4], RNC; Eduard Brunet[5], MD, PhD; Melodi González[6], RNC; Ana Gutiérrez[7], MD, PhD; Pilar Nos[1,2], MD, PhD

[1]Gastroenterology Department, La Fe University and Polytechnic Hospital, Valencia, Spain
[2]Health Research Institute La Fe, Valencia, Spain
[3]Miguel Servet University Hospital, Zaragoza, Spain
[4]University Clinical Hospital, Santiago, Spain
[5]Parc Taulí Hospital Universitari, Institut d'Investigació i Innovació Parc Taulí (I3PT-CERCA), Sabadell, Spain
[6]Nuestra Señora de la Candelaria University Hospital, Tenerife, Spain
[7]Dr Balmis General University Hospital, ISABIAL, Alicante, Spain
[8]CIBERehd, Instituto de Salud Carlos III, Madrid, Spain
[9]Lozano Blesa Clinic University Hospital, Zaragoza, Spain
[10]Burgos University Hospital, Burgos, Spain
[11]Hospital Universitario de La Princesa, Instituto de Investigación Sanitaria Princesa (IIS-Princesa), Universidad Autónoma de Madrid, Madrid, Spain
[12]La Paz University Hospital, Faculty of Medicine, Universidad Autónoma de Madrid, Madrid, Spain
[13]Instituto de Investigación Sanitaria del Hospital Universitario La Paz (IdiPAZ), Madrid, Spain
[14]Infanta Sofía University Hospital, Madrid, Spain
[15]Ramón y Cajal University Hospital, Madrid, Spain
[16]Germans Trias i Pujol University Hospital, Badalona, Spain
[17]San Cecilio Clinic University Hospital, Parque Tecnológico de la Salud, Granada, Spain
[18]Morales Meseguer General University Hospital, Murcia, Spain
[19]Clinic University Hospital, Valencia, Spain
[20]Reina Sofía University Hospital, Córdoba, Spain
[21]Hospital Alvaro Cunqueiro, Vigo, Spain
[22]Puerta de Hierro University Hospital, Madrid, Spain
[23]Consorci Sanitari Terrasa, Barcelona, Spain
[24]Xarxa Assistencial University Hospital, Manresa, Spain
[25]Mútua Terrassa University Hospital, Terrassa, Spain
[26]Fuenlabrada University Hospital, Madrid, Spain
[27]Cabueñes Universitary Hospital, Gijón, Spain
[28]Joint Research Unit in ICT Applied to Reengineering Socio-Sanitary Process, IIS La Fe-Universitat Politècnica de València, Valencia, Spain
[29]Physiotherapy in Motion Multispeciality Research Group (PTinMOTION), Department of Physiotherapy, Universitat de València, Valencia, Spain
[‡]On behalf of GETECCU
[*]these authors contributed equally

Corresponding Author:
Javier Del Hoyo, MD, PhD
Gastroenterology Department
La Fe University and Polytechnic Hospital

XSL•FO
RenderX

Avenida Fernando Abril Martorell, nº 106
Valencia, 46026
Spain
Phone: 34 34 96 124 4000
Fax: 34 34 96 124 6257
Email: delhoyo_jav@gva.es

Abstract

Background: Telemonitoring for inflammatory bowel disease (IBD) has not consistently demonstrated superiority over standard care; however, noninferiority may be an acceptable outcome if remote care proves to be more efficient.

Objective: This study aims to compare the remission time and quality of life of patients with active IBD managed through standard care versus the TECCU (Telemonitoring of Crohn Disease and Ulcerative Colitis) app.

Methods: A 2-arm, randomized, multicenter trial with a noninferiority design was conducted across 24 hospitals in Spain. The study included adult patients with IBD who were starting immunosuppressive or biological therapy. Participants were randomized into 2 groups: the telemonitoring group (G_TECCU) and the standard care group (G_Control). The follow-up schedule for the telemonitoring group (G_TECCU) was based on contacts via the TECCU app, while the control group (G_Control) adhered to standard clinical practice, which included in-person visits and telephone calls. In both groups, treatment adjustments were made based on the progression of disease activity and medication adherence, assessed using specific indices and biological markers at each check-up. The primary outcome was the duration of remission after 12 weeks, while secondary outcomes included quality of life, medication adherence, adverse events, and patient satisfaction.

Results: Of the 169 patients enrolled, 158 were randomized and 150 were analyzed per protocol: telemonitoring (n=71) and control (n=79). After 12 weeks, the time in clinical remission was not inferior in the telemonitoring group (mean 4.20, SD 3.73 weeks) compared with the control group (mean 4.32, SD 3.28 weeks), with a mean difference between arms of –0.12 weeks (95% CI –1.25 to 1.01; noninferiority *P*=.02). The mean reduction in C-reactive protein values was –15.40 mg/L (SD 90.15 mg/L; *P*=.19) in the G_TECCU group and –13.16 mg/L (SD 54.61 mg/L; *P*=.05) in the G_Control group, with no significant differences between the 2 arms (*P*=.73). Similarly, the mean improvement in fecal calprotectin levels was 832.3 mg/L (SD 1825.0 mg/L; *P*=.003) in the G_TECCU group and 1073.5 mg/L (SD 3105.7 mg/L; *P*=.03) in the G_Control group; however, the differences were not statistically significant (*P*=.96). Quality of life improved in both groups, with a mean increase in the 9-item Inflammatory Bowel Disease Questionnaire score of 13.44 points (SD 19.1 points; *P*<.001) in the G_TECCU group and 18.23 points (SD 22.9 points; *P*=.001) in the G_Control group. Additionally, the proportion of patients who adhered to their medication significantly increased from 35% (25/71) to 68% (48/71) in the G_TECCU group (*P*=.001) and from 46% (36/79) to 73% (58/79) in the G_Control group (*P*=.001). The satisfaction rate remained stable at around 90%, although noninferiority was not demonstrated for the secondary outcomes.

Conclusions: Telemonitoring patients with active IBD is not inferior to standard care for achieving and maintaining short-term remission. The TECCU app may serve as a viable alternative follow-up tool, pending confirmation of improved health outcomes and cost-effectiveness over the long-term.

Trial Registration: ClinicalTrials.gov NCT06031038; https://clinicaltrials.gov/ct2/show/NCT06031038

International Registered Report Identifier (IRRID): RR2-10.2196/resprot.9639

(J Med Internet Res 2024;26:e60966) doi:10.2196/60966

KEYWORDS

clinical trial; telemonitoring; inflammatory bowel disease; Crohn disease; ulcerative colitis; quality of life; socioeconomical and psychological end points; health outcomes; remission time

Introduction

Inflammatory bowel disease (IBD) primarily consists of Crohn disease (CD) and ulcerative colitis (UC), both of which are chronic, relapsing conditions characterized by inflammation of the gastrointestinal tract. Because of its chronic nature, IBD requires continuous and personalized monitoring to prevent medium- and long-term complications. Therefore, effective monitoring strategies must be implemented throughout the disease course to optimize the management of patients with IBD.

Unlike other chronic conditions, IBD primarily affects young individuals during their prime period of personal development. Consequently, IBD is associated with high levels of school and work absenteeism, varying degrees of disability [1], interference with social activities, and a reduced health-related quality of life (HRQoL) [2-4]. In addition, patients with IBD place significantly greater demands on health care resources compared to those with other conditions [2]. As a result, IBD has a considerable medical, social, and economic impact, further amplified by the global increase in its incidence and prevalence in recent years [3,4]. In this context, the most recent data

XSL•FO
RenderX

available for Spain show an overall IBD incidence of 16 cases per 100,000 persons per year [5].

To address the challenges in managing these patients, telemedicine applications have been developed to enhance adherence and improve clinical outcomes [6]. Telemonitoring is the primary form of telemedicine used in IBD, concentrating on the structured and continuous monitoring of clinical data that patients self-report from their usual environment. In recent years, web-based telemonitoring systems have been developed, including mobile health (mHealth) tools, which are more cost-effective than home-automated telemanagement programs [7]. Web-based telemonitoring systems are safe, feasible, and cost-effective solutions for patients with IBD [8-12]. Furthermore, their use leads to fewer outpatient visits and hospital admissions [8,9,11,13-17], resulting in reduced health care costs [8,13,14].

Telemonitoring can, in fact, address many aspects of the STRIDE-II strategy for the early detection of potential complications in IBD [18]. Symptomatic responses and remission can be assessed using validated patient-reported outcome measures. Additionally, point-of-care tests now enable the measurement of fecal calprotectin (FC) near the patient. However, to date, telemonitoring has not demonstrated superiority over standard care in terms of health outcome improvements [8,10,11,13,15,19,20], and inconsistent results have been observed across different populations [7].

Even so, noninferiority can be considered an advance if the intervention provides other benefits, such as the cost reductions previously reported with telemonitoring [8,10,11,13,14,17,21,22]. In this context, we have developed a web-based telemanagement system called TECCU (*Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa*, or Telemonitoring of Crohn Disease and Ulcerative Colitis) for the remote monitoring of patients with moderate-to-severe IBD who are starting treatment with immunosuppressants or biological agents [23]. In a pilot randomized trial, TECCU was shown to be a safe method for improving disease outcomes, with a more cost-effective profile than standard care, though the improvement in disease control was not statistically superior [20,24,25].

Therefore, telemonitoring has not been demonstrated to be either superior or noninferior to standard care in improving health outcomes, and this lack of data complicates decision-making when considering investments in mature telemedicine programs. As a result, telemonitoring has not been implemented in a structured manner in daily practice, and the reproducibility of results achieved so far has been limited to a few centers. Given these considerations, we adopted a novel noninferiority design for a multicenter trial conducted on a nationwide scale in Spain. This study aims to evaluate the time in remission and QoL of patients with IBD with moderate-to-severe activity managed through telemonitoring (G_TECCU) compared with standard care (G_Control) after 12 weeks.

Methods

Study Design

A randomized, open, multicenter, noninferiority trial was conducted across 24 hospitals in Spain. The country has 17 autonomous communities, and patients from 10 of these regions participated in the study, as illustrated in the map in Figure 1. The project was promoted by GETECCU, and the study design was discussed in successive research meetings before final protocol approval. The trial is registered on ClinicalTrials.gov with the identifier NCT06031038.

XSL•FO
RenderX

Figure 1. Geographical locations of the participating centers. TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis.

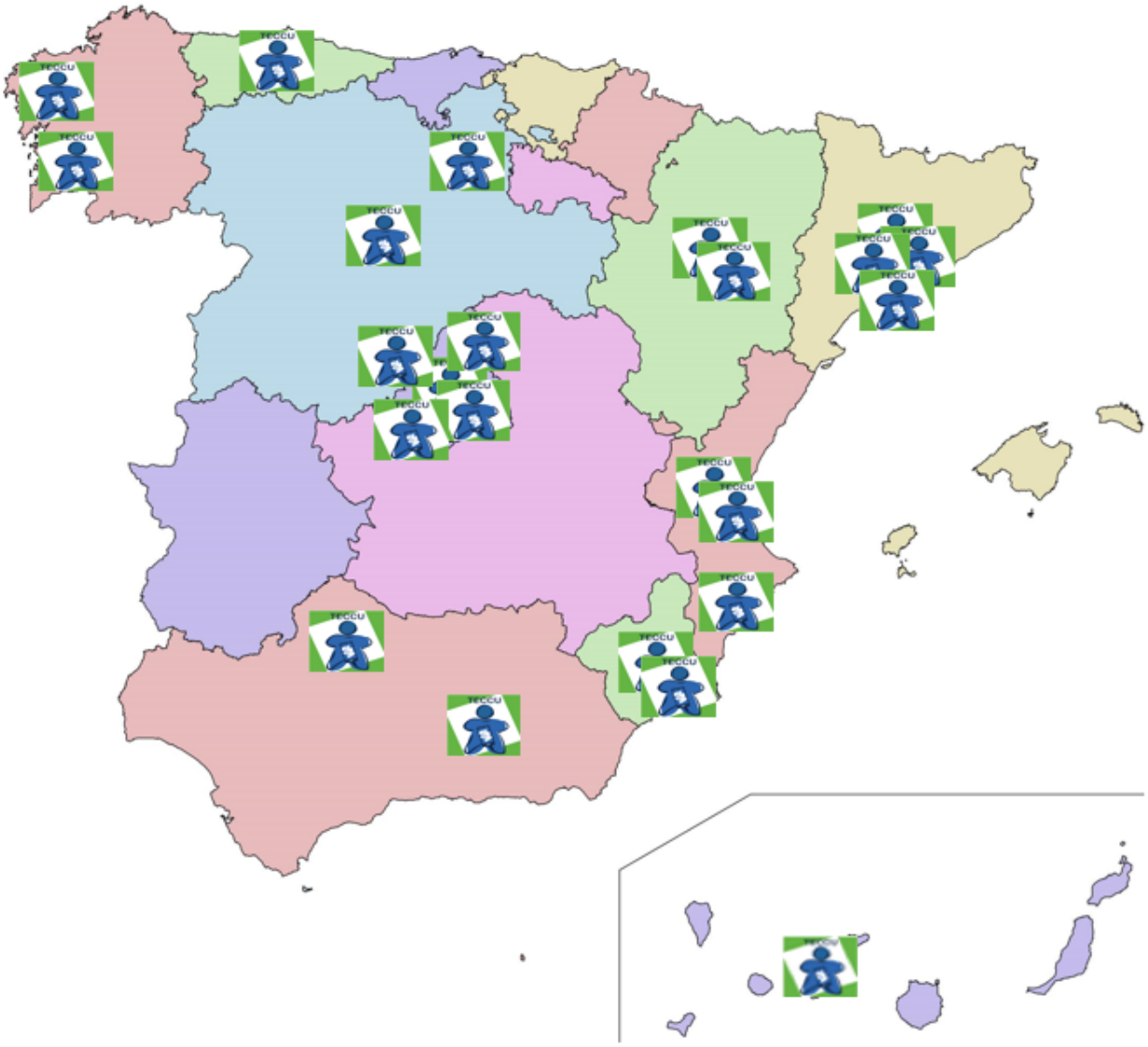

Patient Selection and Recruitment

Data Collection

Patients were recruited at the 24 participating hospitals between September 2020 and April 2023. The collected data were assessed, stored in a database, and made available for interim analysis starting March 19, 2024 [26,27]. Patients were included consecutively from the outpatient clinics of the IBD Units or the Gastroenterology wards. The inclusion and exclusion criteria (see below) were verified at each visit.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: patients aged 18 years or older diagnosed with CD or UC according to internationally accepted criteria [26,27] and initiation of therapy with immunosuppressants or biological agents (with or without steroids) or both due to disease activity occurring no more than 1 week before inclusion in the study. Active disease at the time of inclusion was defined using clinical indices, biological markers, or endoscopic activity: Harvey-Bradshaw Index (HBI) >4 for CD or Simple Clinical Colitis Activity Index (SCCAI) >2 for UC; or FC values ≥200 µg/g; and/or moderate to severe endoscopic activity (ulcers in CD/endoscopic Mayo Index ≥2 for UC).

The exclusion criteria were as follows: cognitive or sensory impairment; inability to speak or read Spanish without a legally authorized representative capable of participating in the study; transient patients; inability to manage a smartphone, tablet, or computer; lack of a telephone line; participation in other clinical trials during the inclusion period; uncontrolled medical or psychiatric conditions; presence of ileorectal or ileal pouch-anal anastomosis; recipients of a definitive ileostomy; perianal disease; patients with cancer undergoing active treatment; terminal patients or individuals receiving palliative care as defined by the Spanish Society of Palliative Care; institutionalized patients; patients or first-degree relatives who are part of the research team or staff members of the research or health centers participating in the study; and patients undergoing specific follow-up in other units (eg, hemodialysis, transplants) requiring mandatory hospital visits at least every 2 months.

XSL•FO
RenderX

Recruitment Process

Patients underwent a face-to-face interview with research staff involved in the study during outpatient visits to the IBD units or during hospitalization for an IBD flare-up. These interviews provided information about the care program and the study, and they were used to obtain written informed consent before inclusion in the study. For patients who agreed to participate, baseline data were collected using a notebook specifically designed for this purpose. Additionally, data regarding biological parameters were obtained from blood and stool samples collected at the initial study visit. If the same biological parameters had been obtained in standard clinical practice less than 2 weeks before enrollment in the study, they were accepted as baseline values.

After a 12-week follow-up, coinciding with the patient's visit to the IBD unit as part of their routine health care, a face-to-face interview was conducted to complete the case report and collect data on the following: clinical activity indices, QoL using the 9-item Inflammatory Bowel Disease Questionnaire (IBDQ-9), patient satisfaction with treatment, activity at work and productivity, medication adherence, and adverse events (AEs). Additionally, biological parameters from blood and stool samples were measured again during this visit.

Randomization

Eligible patients were randomized in a 1:1 ratio into 2 groups: 1 group underwent remote monitoring through TECCU (G_TECCU), while the other group received standard care, which consisted of in-person outpatient visits to the IBD units combined with telephone calls made by physicians or specialized nurses, following standard clinical practice (G_Control). The allocation schedule was based on computer-generated random numbers with a block size of 4 patients. Allocation concealment was ensured using a web-based open-source application for randomization in clinical trials to generate a random allocation sequence. Once a number was assigned, it could not be reassigned, and the members of the research team who were in contact with the patients did not have access to the randomization tables or lists.

The follow-up schedule was the same for both groups, consisting of an in-person visit at baseline and at 12 weeks. The differences between the groups lie in the intermediate controls: in-person visits or telephone calls for G_Control and the TECCU app for G_TECCU. Moreover, additional clinical visits and telephone calls were made at the discretion of the health care providers in both arms if necessary, based on the patient's clinical evolution. Neither the patients nor the researchers were masked to the intervention; however, the results were analyzed by an independent statistician who was blinded to the group identification.

Interventions

TECCU Telemonitoring App

For G_TECCU, monitoring and management were conducted remotely through the updated TECCU management platform, which was set up according to the patient's preferences gathered from a series of focus groups [28]. This platform addresses the needs of patients, professionals, and the organization to ensure the efficient provision of health services. It is a multiplatform and flexible solution that supports clinical decision-making and can be integrated with electronic medical records. It is configurable, allowing plans and thresholds to be customized for specific patient profiles. The platform also enables the creation of alerts using different variables, facilitating a holistic approach to patient care.

TECCU operates through a secure web page with an HTTPS app for mobile phones and tablets. During telemonitoring, patients connect to the platform via the app using their personal code and respond to various questionnaires about their disease in the form of chat messages. The questions pertain to the variables used to assess disease status at each check-up, in accordance with a preestablished schedule. Furthermore, the number of check-ups can be increased if necessary to adequately monitor disease evolution during follow-up.

In addition, patients received advice, reminders, and educational materials about their disease and its prevention. The specialized health care personnel at each IBD unit received information from the patients, which was filtered through an intelligent prioritization system to generate alerts based on an integrated intervention protocol. Upon receiving an alert, the health care personnel implemented action plans in accordance with the established intervention protocol to adjust medication and follow-up schedules as needed. These alerts were triggered based on responses to questions regarding activity indices, AEs, and fecal or blood results. Treatments were adjusted using the platform's messaging system, along with telephone calls or in-person visits when patients required training on the administration of new medications.

Standard Care Provided by the IBD Units

The G_Control patients received the usual care provided by the IBD Units (outpatient clinic) for those with moderate-to-severe disease activity, based on national and European clinical guidelines [26,27,29]. Treatment was adjusted according to the evolution of disease activity and medication adherence, measured through specific indices and biological markers during visits or via telephone calls for issues that could be monitored during intermediate check-ups. Additionally, the time in remission was assessed weekly throughout the follow-up based on the patient's self-recorded clinical activity in a home diary.

This care was complemented by ad hoc hospital care in cases of IBD flare-ups. In such instances, intensive care was maintained until the patient's condition stabilized, after which they returned to a follow-up regimen based on standard care through the IBD unit. Patients in both arms were provided with all educational materials about IBD available for the remote monitoring of patients. Questionnaires on HRQoL, satisfaction, and work productivity were completed at baseline and again at 12 weeks.

Study Outcomes

Overview

The variables measured at baseline included sociodemographic information, smoking status, disease profile and activity,

XSL•FO
RenderX

treatment received, HRQoL, work productivity, impairment in daily activities, medication adherence, and patient satisfaction.

Primary Outcome

The primary outcome of the study was to determine the time in remission after 12 weeks of follow-up. This was evaluated by assessing disease activity at baseline and at each check-up during the 12-week follow-up period established for this analysis. Clinical disease activity was evaluated using the modified HBI for patients with CD [30] and the SCCAI, also known as the Walmsley Index, for patients with UC [31]. Clinical remission was defined as an HBI of ≤4 for patients with CD or an SCCAI of ≤2 for patients with UC.

Biological markers were measured at baseline and 12 weeks after inclusion. The laboratory parameters were C-reactive protein (CRP) levels (mg/L) and FC levels (µg/g). Changes in medication were made based on these markers and clinical disease activity indices, following specific intervention plans.

Secondary Outcomes

The HRQoL of patients was evaluated at baseline and week 12 using the specific IBDQ-9, a validated tool consisting of 9 items across 4 dimensions: bowel symptoms, systemic symptoms, emotional status, and social behavior. Each item is scored on a 7-point Likert scale, yielding an overall score that ranges from 7 (lowest QoL) to 63 (highest QoL), which is then calculated as a percentage of the maximum score.

The impact of the disease on work productivity and daily activities was assessed at baseline and at week 12 using the Work Productivity and Activity Impairment (WPAI) questionnaire [32]. This questionnaire consists of 6 items that evaluate the disease's effect on work and daily activities over the past 7 days. The WPAI generates 4 scores expressed as "impairment percentages," with higher scores indicating a greater impact. The Spanish version has been validated and demonstrated reproducibility in patients with CD.

Medication adherence was evaluated using the 8-item Morisky Medication Adherence Scale (MMAS-8) [33,34], which has been utilized in clinical trials involving patients with IBD [19]. Patient satisfaction with the care received was assessed using an adapted version of the Client Satisfaction Questionnaire. In addition, patient-reported outcomes concerning health status, the presence of abdominal pain, stool frequency, and blood in the stool were also recorded. The safety of the interventions was evaluated by measuring the number of emergency department visits, unscheduled outpatient visits, hospitalizations, and AEs related to medication use.

Statistical Analysis

As this is a noninferiority study, we conducted a per-protocol analysis. The characteristics of the participants were described using the mean, SD, 95% CI, median, IQR, and range for quantitative variables, as well as absolute and relative frequencies for qualitative variables. Possible baseline differences between the 2 study groups were compared using the Student *t* test for independent samples or the Mann-Whitney *U* test for quantitative variables, and the chi-square test or Fisher exact test for qualitative variables. Time in clinical remission was summarized using the mean, SD, median, IQR, and range.

The sample size was estimated based on the results of the pilot trial, which indicated that patients in the control group remained in remission for a median of 14.3 weeks, whereas those in the G_TECCU group had a median remission time of 17.9 weeks. These differences in time to remission were evaluated with a noninferiority limit of –1.5 weeks, considering this difference to be clinically irrelevant within the 12-week follow-up period of this interim analysis. Noninferiority was assessed by calculating the mean and 95% CI for the difference in the variable of interest between the groups. This was done by comparing the limits of the CI with the predefined noninferiority margin using a Student *t* test. Accepting an α risk of .025 in a 1-sided test, a β risk of .2, and an SD of 3.2, it was determined that 71 patients were needed in each group. Furthermore, assuming a loss rate of 15%, a total of 168 patients was required for the study.

Changes in secondary outcomes from baseline were described by applying appropriate estimators based on the type of variables, as previously outlined. Possible differences in secondary outcomes between the 2 groups were compared using the Student *t* test for independent samples, the Mann-Whitney *U* test for quantitative variables, and the chi-square or Fisher exact test for qualitative variables. All analyses were conducted using SAS statistical software version 9.4 (SAS Institute), with a 1-sided significance level of .025 accepted for all analyses.

Ethical Considerations

The study protocol was approved by the local independent ethics committee at La Fe University and Polytechnic Hospital, Valencia (v0.3, 14/07/2020), and by the Spanish Agency of Medicines and Medical Devices (AEMPS/Agencia Española de Medicamentos y Productos Sanitarios: 30/07/2020). The study was conducted in accordance with the following: the "Note for Guidance on Good Clinical Practice" (CPMP/ICH/135/95, May 1, 1996); Royal Decree 223/2004 (February 2004); the Helsinki Declaration on ethical principles for medical research involving human subjects, as adopted by the General Assembly of the World Medical Association (7th revised version, Seoul, 2008); guidelines from the International Conference on Harmonization; and the official regulations imposed by the participating centers.

Results

Study Sample and Baseline Characteristics

A total of 169 patients with complex IBD were enrolled in the study between September 15, 2020, and April 28, 2023. Of these, 11 patients were excluded: 4 due to screening failure, 2 due to a lack of baseline data, and 5 who withdrew before randomization. The remaining 158 patients proceeded to follow-up, with 74 (47%) in the G_TECCU group and 84 (53%) in the G_Control group. During the study period, 8 patients (3 from G_TECCU and 5 from G_Control) did not complete the 12-week follow-up. The reasons for this are detailed in Figure 2. As such, the 71 patients in the G_TECCU group and 79 patients in the G_Control group who completed the 12-week

XSL•FO
RenderX

follow-up period were analyzed per protocol (N=150). Additionally, all randomized patients (n=158) were analyzed according to the intention-to-treat principle. The per-protocol population was used to summarize patient disposition and baseline characteristics (Figure 2; also see Multimedia Appendix 1 for the CONSORT-EHEALTH [Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth] checklist [35]).

Figure 2. Study flowchart. G_CONTROL is the group receiving standard care with in-person visits and telephone calls as usual practice; G_TECCU is the group receiving remote monitoring. IBD: inflammatory bowel disease; TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis.

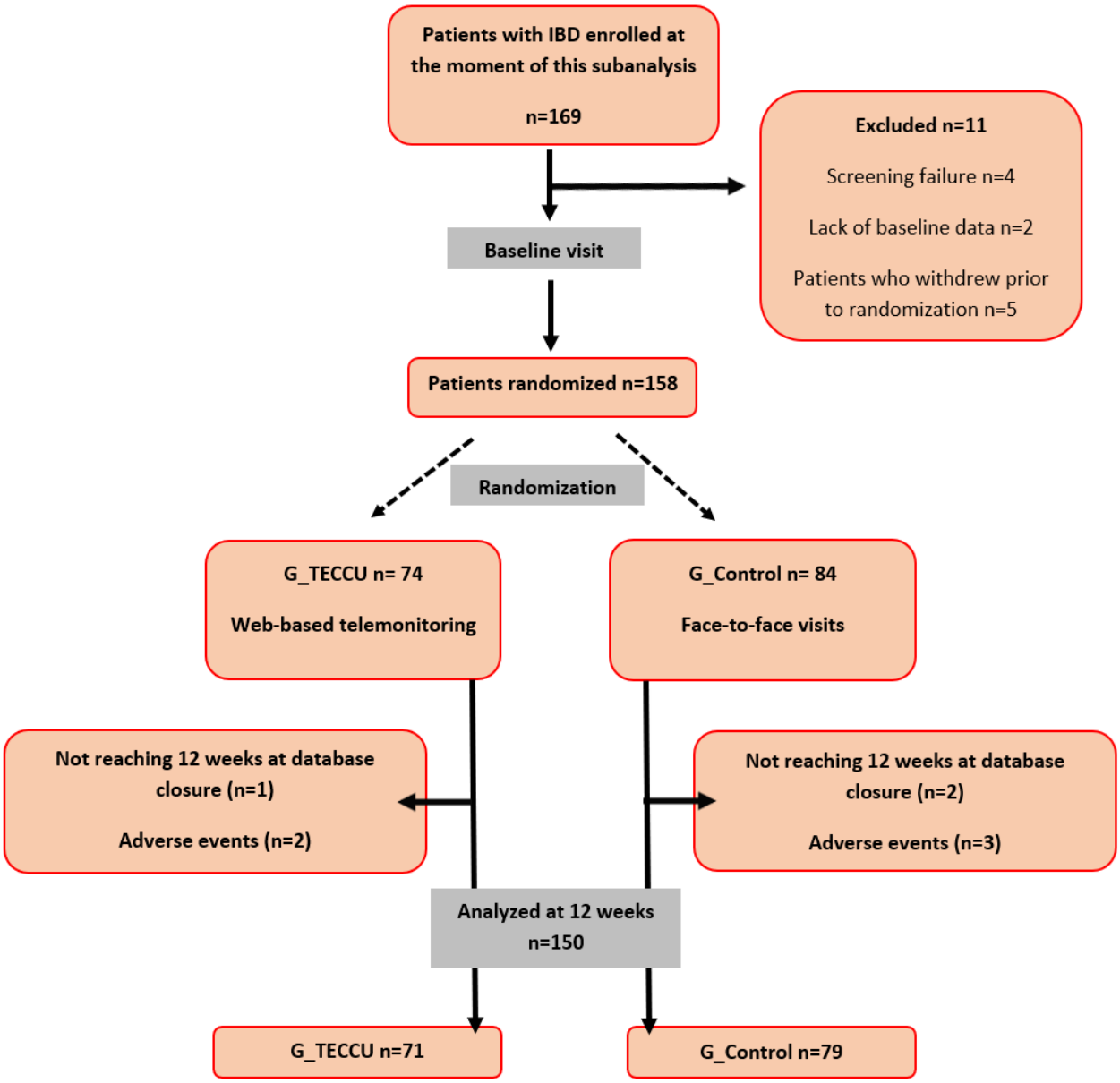

The baseline demographic and clinical characteristics are presented in Table 1. The overall mean age of the participants was 37.8 (SD 18.7) years. Of the 150 patients, 83 (55%) were male and 67 (45%) were female. Among the included patients (n=150), 58 (39%) had UC, and 92 (61%) had CD, all of whom were capable of using the technology. The education levels of the patients were as follows: primary education (12/150, 8%), secondary education (70/150, 47%), and university education (68/150, 45%). Regarding employment status, 100 (66.7%) patients were actively employed, 19 (12.7%) were students, 17 (11.3%) were unemployed, 10 (6.7%) were retired, and 4 (2.7%) were houseworkers. At the onset of the study, 109 (72.7%) patients initiated treatment with biological agents (adalimumab in 55/109, 50.5%, patients), 25 (16.7%) with immunosuppressants (azathioprine in 23/25, 92%, patients), and 16 (10.7%) with combination therapy. No significant differences were observed in the baseline characteristics of patients according to the treatment initiated (immunosuppressants, biological agents, or combination therapy).

XSL•FO
RenderX

Table 1. Baseline characteristics of the study patients.

Characteristic	Total (N=150)	TECCU[a] (n=71)	Standard care (n=79)
Age (years), mean (range)	37.76 (18-66)	38.83 (18-66)	36.80 (18-62)
Gender, n (%)			
Female	67 (45)	34 (48)	33 (42)
Male	83 (55)	37 (52)	46 (58)
Education level, n (%)			
Primary	12 (8)	5 (7)	7 (9)
Secondary	70 (47)	34 (48)	36 (46)
University	68 (45)	32 (45)	36 (46)
Smoking status, n (%)			
Current smoker	27 (18)	15 (21)	12 (15)
Former smoker	50 (33)	27 (38)	23 (29)
Nonsmoker	72 (48)	29 (41)	43 (54)
BMI (kg/m^2), mean (SD)	23.65 (4.98)	23.35 (4.01)	23.93 (5.76)
Disease profile, n (%)			
Ulcerative colitis	58 (39)	23 (32)	35 (44)
Crohn disease	92 (61)	48 (68)	44 (56)
Prior IBD[b]-related surgery[c], n (%)	3 (2)	1 (1)	2 (3)
Previous treatments for IBD, n (%)	129 (86)	61 (86)	68 (86)
Immunosuppressors	66 (44)	30 (42)	36 (46)
Biological agents	36 (24)	15 (21)	21 (27)
Steroids	96 (64)	44 (62)	52 (66)
Sulfasalazine and 5-aminosalicylates	68 (45)	30 (42)	38 (48)
Other treatments	16 (11)	6 (8)	10 (13)
No previous treatment, n (%)	21 (14)	10 (14)	11 (14)
Current treatment, n (%)			
Only immunosuppressors	25 (17)	8 (11)	17 (22)
Only biological agents	109 (73)	58 (82)	51 (65)
Combination therapy	16 (11)	5 (7)	11 (14)

[a]TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn Disease and Ulcerative Colitis (represents a group of patients with inflammatory bowel disease receiving remote monitoring).

[b]IBD: inflammatory bowel disease.

[c]Within the past 2 months before study enrollment.

Regarding their clinical status (Table 2), 58 of 150 (38.7%) patients were in clinical remission at baseline; however, they were included in the study due to FC levels exceeding 200 µg/g or the presence of moderate to severe endoscopic activity. Concerning biological markers, the mean CRP value was 18.5 (SD 73.4) mg/L, while the mean FC level was 1631.1 (SD 2722.4) µg/g. Among the 24 patients with UC who entered the study following an endoscopic evaluation, 8 (33%) had a Mayo Index of 2, indicating moderate activity, while 16 patients (67%) had a Mayo Index of 3, indicating severe activity. Among the 34 patients with CD who underwent a baseline endoscopy, 27 (79%) exhibited ulcers. Furthermore, 61 of 150 (41%) patients demonstrated good adherence to medication (25/71, 35%, in the G_TECCU group and 36/79, 46%, in the G_Control group). Patient satisfaction, as measured by specific questionnaires, yielded a score of 87.3 (SD 11.2).

XSL•FO
RenderX

Table 2. Evaluation of disease activity, impact on patient's quality of life and daily life, satisfaction with the medical care received, and medication adherence at baseline.

Characteristic	Total (N=150)	TECCU[a] (n=71)	Standard care (n=79)
Clinical activity [b]			
SCCAI[c]<3 or HBI[d]<5 (but fecal calprotectin >200 µg/g or endoscopic activity), n (%)	58 (39)	33 (46)	25 (32)
SCCAI[e], mean (SD)	6.59 (3.05)	6.78 (2.37)	6.46 (3.45)
Harvey-Bradshaw index[f], mean (SD)	4.68 (3.11)	3.94 (2.97)	5.48 (3.10)
Biological activity, mean (SD)			
C-reactive protein level (mg/L)	18.53 (73.43)	20.26 (90.96)	17.02 (54.13)
Fecal calprotectin level (µg/g)	1631.1 (2722.4)	1418.7 (1977.1)	1614.9 (3322.9)
Endoscopic activity, n/N (%)			
Mayo score of 2[g]	8/24 (33)	5/9 (56)	3/15 (20)
Mayo score of 3[h]	16/24 (67)	4/9 (44)	12/15 (80)
Patients with Crohn disease with ulcers	27/34 (79)	10/14 (71)	17/20 (85)
Quality of life [i]			
IBDQ-9[j] score, mean (SD)	54.69 (20.46)	59.01 (21.79)	50.36 (19.76)
Medication adherence[k], n (%)	61 (41)	25 (35)	36 (46)
Patient satisfaction score, mean (SD)	87.31 (11.17)	88.04 (10.44)	87.13 (11.84)
Work productivity and activity impairment, median (IQR)			
Work hours missed	7.14 (0-100)	3.34 (0-66.7)	9.72 (0-100)
Impairment while working	40.00 (0-75)	30.00 (0-70)	40.00 (10-80)
Overall work impairment	50.00 (10-90)	41.21 (0-82)	50.00 (14.9-93)
Activity impairment	50.00 (20-70)	40.00 (20-70)	50.00 (20-70)

[a]TECCU: *Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa* or Telemonitoring of Crohn Disease and Ulcerative Colitis (represents a group of patients with inflammatory bowel disease receiving remote monitoring).

[b]Clinical remission was defined as HBI ≤4 for patients with Crohn disease or Walmsley score ≤2 for patients with ulcerative colitis.

[c]SCCAI: Simple Clinical Colitis Activity Index.

[d]HBI: Harvey-Bradshaw Index.

[e]Patients with ulcerative colitis: higher scores indicate worse clinical conditions.

[f]Patients with Crohn disease: score ranges from 0 to 19, where the higher the score the worse the clinical condition.

[g]Moderate disease.

[h]Severe disease.

[i]The higher the score, the better quality of life.

[j]IBDQ-9: 9-item Inflammatory Bowel Disease Questionnaire.

[k]Score ranges from 0 to 8, with scores of 8 reflecting high adherence.

Efficacy Assessment: Disease Activity

The mean time in clinical remission after the 12-week follow-up was 4.26 (SD 3.49) weeks for the entire study cohort. Based on the noninferiority statistics, the follow-up with TECCU was not inferior to standard care, with a mean time in remission of 4.20 (SD 3.73) weeks for the G_TECCU patients compared with 4.32 (SD 3.28) weeks for the G_Control patients (noninferiority *P*=.02; see Table 3). The mean difference in clinical remission between the 2 groups was –0.12 weeks, with a 95% CI of –1.25 to 1.01 (G_TECCU vs G_Control, *P*=.02; see Figure 3A).

XSL•FO
RenderX

Figure 3. (A) Differences in the time in clinical remission between the 2 arms by the PP population. (B) Differences in the time in clinical remission between the 2 arms by the ITT population. (C) Differences in the time in clinical remission between the 2 arms by the PP population in multiplicity adjustments. (D) Differences in the time in clinical remission between the 2 arms by the ITT population in multiplicity adjustments. The data are shown for the whole cohort of patient analyzed and for specific subgroup (patients with UC or CD) analysis. Control: group of IBD patients receiving standard care with in-person visits; TECCU: group of IBD patients receiving remote monitoring. The dashed vertical line represents the non-inferiority margin of -1.5 weeks. CD: Crohn disease; ITT: intention to treat; PP: per-protocol; TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis; UC: ulcerative colitis.

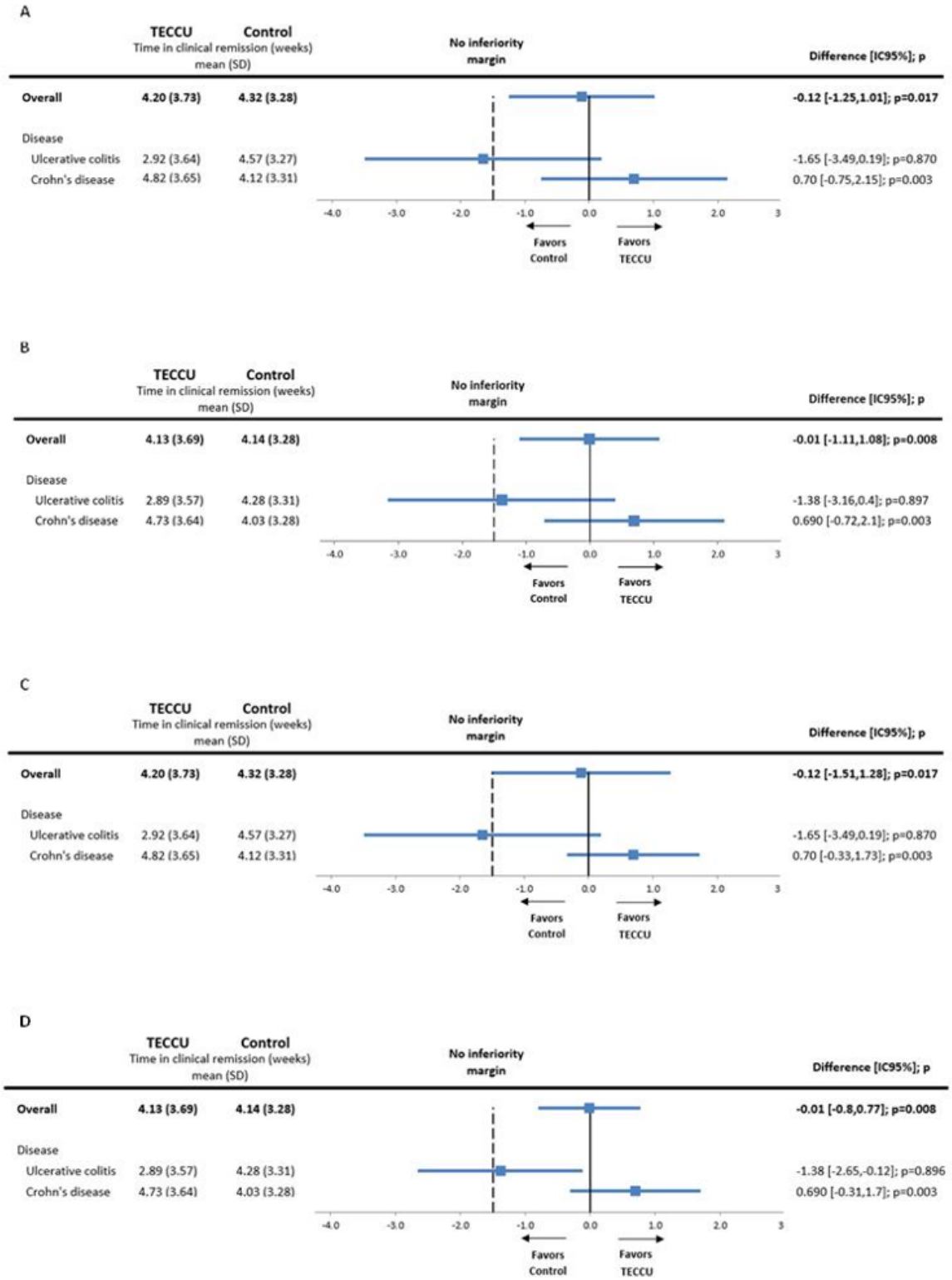

XSL•FO
RenderX

Table 3. Time in clinical remission by the per-protocol population.

Time in clinical remission (weeks)	Participants, n	Mean (SD)	Median (IQR)	Range	Noninferiority test *P* value[a]
Overall					.02[b]
Total	150	4.26 (3.49)	4.42 (0.45-6.67)	0.00-14.29	
TECCU[c]	71	4.20 (3.73)	4.33 (0.16-8.00)	0.00-14.29	
Control[d]	79	4.32 (3.28)	4.61 (0.57-6.33)	0.00-12.43	
Colitis ulcerative					.87
Total	58	3.92 (3.49)	4.00 (0.01-7.60)	0.00-10.86	
TECCU	23	2.92 (3.64)	0.45 (0.00-4.78)	0.00-10.00	
Control	35	4.57 (3.27)	4.71 (0.49-7.86)	0.00-10.86	
Crohn disease					.003[b]
Total	92	4.48 (3.49)	4.57 (0.64-6.53)	0.00-14.29	
TECCU	48	4.82 (3.65)	4.71 (0.69-8.08)	0.00-14.29	
Control	44	4.12 (3.30)	4.41 (0.58-6.07)	0.00-12.43	

[a]Noninferiority limit=–1.5.

[b]Significant.

[c]TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn Disease and Ulcerative Colitis (represents a group of patients with inflammatory bowel disease receiving remote monitoring).

[d]Control represents a group of patients with inflammatory bowel disease receiving standard care with in-person visits.

The subgroup analysis for patients with CD revealed a mean difference in clinical remission of +0.70 weeks for G_TECCU compared with G_Control (95% CI –0.75 to 2.15; *P*=.003). Given the statistical significance, the noninferiority of TECCU was confirmed in this subgroup. By contrast, the mean difference in clinical remission for patients with UC was –1.65 weeks between G_TECCU and G_Control (95% CI –3.49 to 0.19; *P*=.87), indicating that noninferiority of TECCU could not be confirmed in this cohort (Figure 3A). Similar results were observed in the intention-to-treat population and after adjusting for multiple tests (see Tables 4-6 and Figure 3B-3D).

After 12 weeks, the mean SCCAI significantly improved in both the G_TECCU (mean change –3.78, SD 3.16, *P*<.001) and G_Control patients with UC (mean change –4.11, SD 4.31; *P*<.001). Similarly, the mean HBI value also showed a significant improvement in G_TECCU patients (mean change –1.42, SD 2.69; *P*<.001) and G_Control patients (mean change –2.07, SD 4.06; *P*=.002). However, the improvements in SCCAI (*P*=.27) and HBI (*P*=.26) scores were not significantly different between the 2 intervention arms.

Disease activity was also assessed based on CRP and FC levels (Figure 4). The mean CRP value significantly decreased after 12 weeks compared with baseline in both groups, with a mean reduction of –15.40 (SD 90.15) mg/L (*P*=.19) in the G_TECCU group and –13.16 (SD 54.61) mg/L (*P*=.05) in the G_Control group. However, this improvement in CRP did not significantly differ between the 2 groups (G_TECCU vs G_Control: *P*=.73). Similarly, the mean FC levels were significantly lower after 12 weeks compared with baseline in both study groups. The G_TECCU group experienced a mean reduction of 832.3 (SD 1825.0) mg/L (*P*=.003), while the G_Control group had a mean reduction of 1073.5 (SD 3105.7) mg/L (*P*=.03). However, there were no significant differences between the 2 groups (*P*=.96).

XSL•FO
RenderX

Table 4. Time in clinical remission by the intention-to-treat population.

Time in clinical remission (weeks)	Participants, n	Mean (SD)	Median (IQR)	Range	Noninferiority test *P* value[a]
Overall					.008[b]
Total	158	4.14 (3.47)	4.33 (0.31-6.33)	0.00-14.29	
TECCU[c]	74	4.13 (3.69)	4.23 (0.16-7.70)	0.00-14.29	
Control[d]	84	4.14 (3.28)	4.41 (0.42-6.07)	0.00-12.43	
Colitis ulcerative					.90
Total	62	3.74 (3.45)	4.00 (0.00-5.86)	0.00-10.86	
TECCU	24	2.89 (3.57)	0.60 (0.00-4.60)	0.00-10.00	
Control	38	4.28 (3.31)	4.65 (0.29-7.60)	0.00-10.86	
Crohn disease					.003[b]
Total	96	4.39 (3.47)	4.53 (0.61-6.36)	0.00-14.29	
TECCU	50	4.73 (3.64)	4.71 (0.67-8.00)	0.00-14.29	
Control	46	4.03 (3.28)	4.33 (0.57-6.01)	0.00-12.43	

[a]Noninferiority limit=–1.5.

[b]Significant.

[c]TECCU: *Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa* or Telemonitoring of Crohn Disease and Ulcerative Colitis (represents a group of patients with inflammatory bowel disease receiving remote monitoring).

[d]Control represents a group of patients with inflammatory bowel disease receiving standard care with in-person visits.

Table 5. Time in clinical remission by the per-protocol population in multiplicity adjustments.

Time in clinical remission (weeks)	Participants, n	Mean (SD)	Median (IQR)	Range	Noninferiority test *P* value[a]
Overall					.02[b]
Total	150	4.26 (3.49)	4.42 (0.45-6.67)	0.00-14.29	
TECCU[c]	71	4.20 (3.73)	4.33 (0.16-8.00)	0.00-14.29	
Control[d]	79	4.32 (3.28)	4.61 (0.57-6.33)	0.00-12.43	
Colitis ulcerative					.87
Total	58	3.92 (3.49)	4.00 (0.01-7.60)	0.00-10.86	
TECCU	23	2.92 (3.64)	0.45 (0.00-4.78)	0.00-10.00	
Control	35	4.57 (3.27)	4.71 (0.49-7.86)	0.00-10.86	
Crohn disease					.003[b]
Total	92	4.48 (3.49)	4.57 (0.64-6.53)	0.00-14.29	
TECCU	48	4.82 (3.65)	4.71 (0.69-8.08)	0.00-14.29	
Control	44	4.12 (3.30)	4.41 (0.58-6.07)	0.00-12.43	

[a]Noninferiority limit=–1.5.

[b]Significant.

[c]TECCU: *Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa* or Telemonitoring of Crohn Disease and Ulcerative Colitis (represents a group of patients with inflammatory bowel disease receiving remote monitoring).

[d]Control represents a group of patients with inflammatory bowel disease receiving standard care with in-person visits.

XSL•FO
RenderX

Table 6. Time in clinical remission by the intention-to-treat population by multiplicity adjustments.

Time in clinical remission (weeks)	Participants, n	Mean (SD)	Median (IQR)	Range	Noninferiority test *P* value[a]
Overall					.008[b]
Total	158	4.14 (3.47)	4.33 (0.31-6.33)	0.00-14.29	
TECCU[c]	74	4.13 (3.69)	4.23 (0.16-7.70)	0.00-14.29	
Control[d]	84	4.14 (3.28)	4.41 (0.42-6.07)	0.00-12.43	
Colitis ulcerative					.90
Total	62	3.74 (3.45)	4.00 (0.00-5.86)	0.00-10.86	
TECCU	24	2.89 (3.57)	0.60 (0.00-4.60)	0.00-10.00	
Control	38	4.28 (3.31)	4.65 (0.29-7.60)	0.00-10.86	
Crohn disease					.003[b]
Total	96	4.39 (3.47)	4.53 (0.61-6.36)	0.00-14.29	
TECCU	50	4.73 (3.64)	4.71 (0.67-8.00)	0.00-14.29	
Control	46	4.03 (3.28)	4.33 (0.57-6.01)	0.00-12.43	

[a]Noninferiority limit=–1.5.

[b]Significant.

[c]TECCU: *Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa* or Telemonitoring of Crohn Disease and Ulcerative Colitis (represents a group of patients with inflammatory bowel disease receiving remote monitoring).

[d]Control represents a group of patients with inflammatory bowel disease receiving standard care with in-person visits.

Figure 4. Evolution of the mean of (A) C-reactive protein and (B) fecal calprotectin levels over the study period in the 2 arms. Control: group receiving standard care with in-person visits; TECCU: group receiving remote monitoring. TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis.

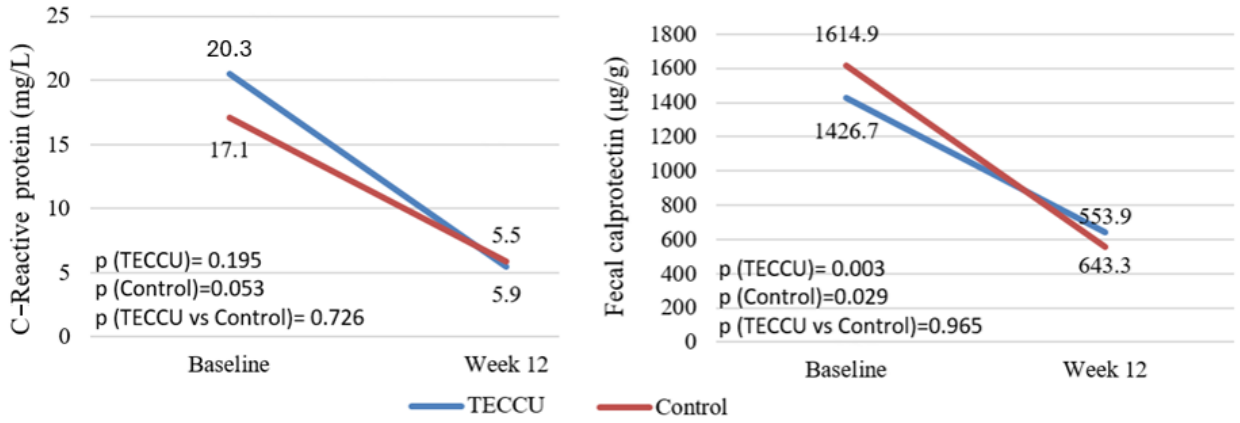

The difference in the mean change in CRP levels between the groups after 12 weeks was –2.25 mg/L for G_TECCU compared with G_Control (95% CI –28.19 to 23.69); however, noninferiority of TECCU could not be demonstrated (noninferiority *P*=.73). Similarly, noninferiority was not demonstrated for the reduction in FC levels among G_TECCU patients, with a mean change difference of 241.2 µg/g compared with G_Control (95% CI –806.9 to 1289.2, noninferiority *P*=.96).

Health-Related Quality of Life

The mean IBDQ-9 scores increased from 59.01 (SD 21.8) to 72.44 (SD 20.3) in G_TECCU patients, indicating a mean improvement of 13.44 points (SD 19.1, *P*<.001). In G_Control patients, scores improved from 50.36 (SD 19.8) to 68.58 (SD 20.2), reflecting a mean improvement of 18.23 points (SD 22.9; *P*=.001; Figure 5). However, because the mean difference between the 2 groups was –4.79 (95% CI –12.9 to 3.3; *P*=.90) and the CI crossed the noninferiority margin, the noninferiority of TECCU in improving the IBDQ-9 score could not be demonstrated.

XSL•FO
RenderX

Figure 5. Evolution of the IBDQ-9 score over the study period in the 2 arms. Control: group receiving standard care with in-person visits; TECCU: group receiving remote monitoring. IBDQ-9: 9-item Inflammatory Bowel Disease Questionnaire; TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis.

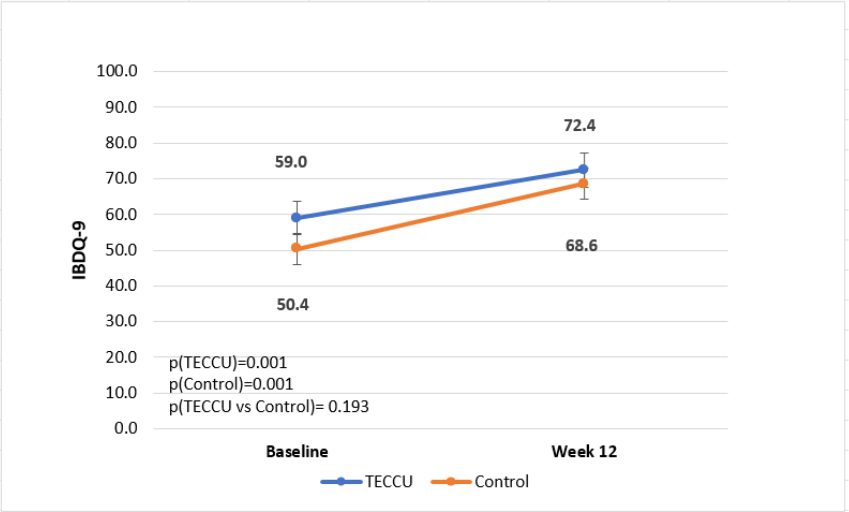

Work Productivity and Activity Impairment

Regarding the impact of the disease on work productivity and daily activities, a significant improvement in WPAI scores was observed at 12 weeks compared with baseline in the G_Control arm (*P*=.001 for all scores). By contrast, TECCU only demonstrated a significant improvement in impairment while working (mean improvement –17.74, SD 39.65; *P*=.003) and in activity impairment (mean improvement –14.49, SD 30.69; *P*=.002). However, no significant differences between the 2 arms were found in work hours missed due to health problems (*P*=.21), impairment while working (*P*=.38), overall work impairment (*P*=.19), or activity impairment (*P*=.11).

Medication Adherence

After 12 weeks, the proportion of patients adhering to their medication increased significantly in the G_TECCU group (rising from 25/71, 35%, to 48/71, 68%; *P*=.001) and in the G_Control group (increasing from 36/79, 46%, to 58/79, 73%; *P*=.001). However, this improvement in medication adherence did not differ significantly between the 2 arms (G_TECCU vs G_Control: *P*=.46). Additionally, the MMAS-8 score in G_TECCU patients improved significantly after 12 weeks, with a mean baseline value increasing from 7.07 (SD 1.03) to 7.50 (SD 0.86), representing a mean improvement of 0.48 (SD 1.6, *P*=.002). Similarly, the MMAS-8 score improved significantly in the G_Control group after 12 weeks, rising from a mean baseline value of 7.25 (SD 1.1) to 7.57 (SD 1.2), reflecting a mean improvement of 0.29 (SD 1.3, *P*=.09; Figure 6). The difference in the mean change of the MMAS-8 score between the 2 arms was 0.19 (95% CI –0.24 to 0.63; *P*=.39), indicating a trend toward statistical significance regarding the noninferiority of G_TECCU compared with G_Control (*P*=.07).

Changes in medication were made for 24 of the 150 patients (16%) during the study period, with no significant differences observed in the number of patients who changed their medication between the groups (G_TECCU: 12/71, 17%, patients vs G_Control: 12/79, 15%, patients; *P*=.77).

Figure 6. Evolution of the MMAS-8 score over the study period in the 2 arms. Control, group receiving standard care with in-person visits; TECCU, group receiving remote monitoring. MMAS-8: 8-item Morisky Medication Adherence Scale; TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis.

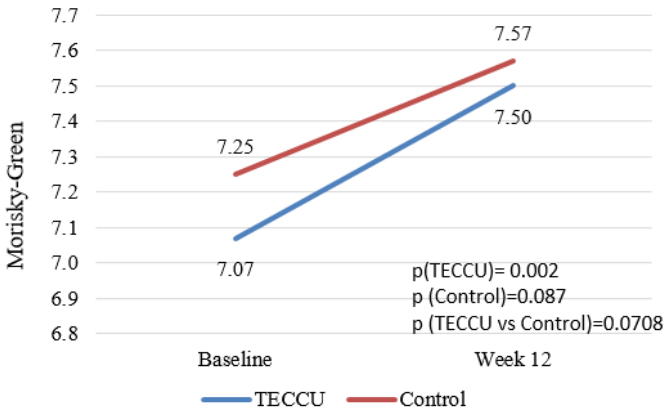

XSL•FO
RenderX

Patient Satisfaction

In the G_TECCU group, the mean satisfaction questionnaire score improved from 88.04 (SD 10.3) to 90.51 (SD 9.7; *P*=.04), while in the G_Control group, it increased from 87.13 (SD 12.3) to 89.57 (SD 8.5; *P*=.12). The mean improvement in satisfaction scores was 2.47 (SD 8.6) for G_TECCU (*P*=.04) and 2.43 (SD 11.2) for G_Control (*P*=.12); however, the noninferiority test between the groups was not significant (*P*=.97; Figure 7).

Figure 7. Evolution of the patient satisfaction score over the study period in the 2 arms. Control: group receiving standard care with in-person visits; TECCU: group receiving remote monitoring. CSQ: Client Satisfaction Questionnaire; TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn's Disease and Ulcerative Colitis.

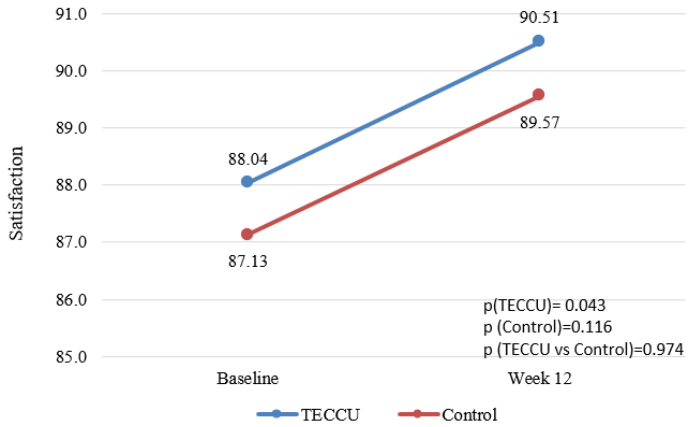

Use of Health Care Resources and Safety

Emergency department visits were significantly higher in the G_TECCU group compared with the G_Control group (13/71, 18%, G_TECCU vs 5/79, 6%, G_Control; *P*=.02). However, the intervention did not result in significant differences in unscheduled outpatient visits (19/71, 27%, G_TECCU vs 16/79, 20%, G_Control; *P*=.35) or hospitalizations (5/71, 7%, G_TECCU vs 2/79, 3%, G_Control; *P*=.19). A total of 70 AEs were recorded in the G_TECCU group in 39 of 71 (55%) patients, while 60 AEs were noted in the G_Control group in 31 of 79 (39%) patients. However, the differences between groups did not reach statistical significance (*P*=.054).

Among these AEs, 5 resulted in patient loss during the follow-up period, with 2 in the G_TECCU group and 3 in the G_Control group. A total of 17 AEs were related to medication, comprising 8 of 71 (11%) patients in the G_TECCU group and 9 of 79 (11%) patients in the G_Control group, with no significant differences between the groups (*P*=.98). Notably, none of the patients died during this period.

Discussion

Principal Findings and Comparison With Prior Work

To the best of our knowledge, this is the first trial using a noninferiority design to compare telemonitoring with standard care for patients with moderate-to-severe IBD. In previous studies, telemonitoring was shown to be safe and to improve both health outcomes and costs [8,9,11,13-15]. In fact, telemonitoring can meet the treat-to-target strategy by addressing many variables of the STRIDE-II recommendations [18], but it failed to demonstrate superiority over standard care in many studies [8,10,11,13,14,17]. Nonetheless, noninferiority would also be an acceptable goal if, in return, telemonitoring improves the efficiency of care. In light of the positive results obtained with TECCU in a pilot trial carried out on patients with active IBD [20], here we demonstrate on a nationwide scale that telemonitoring with the TECCU app is not inferior to standard care in inducing and maintaining remission over 12 weeks.

The main outcome of this study was the time patients remained in remission, which is relevant because we evaluated the evolution of clinical activity throughout the follow-up period, not just the clinical status at the end of the trial, as reported thus far [13,15,19,20,36]. The mean period in remission with telemonitoring was 4.20 weeks, a relatively short time that was probably related to the selection of patients during the induction phase of treatment. Then, in this work, we evaluated the first 12 weeks after initiating biological agents or immunosuppressants to manage persistent disease activity or a flare-up. Indeed, the mean time in remission was essentially the same as that achieved through standard care, and its distribution in each group was sufficiently similar to demonstrate noninferiority.

A subgroup analysis confirmed the noninferiority of TECCU for patients with CD, but it was inconclusive for patients with UC. Furthermore, the improvement in clinical activity was consistent with the reduction in the markers of inflammation (CRP and FC) over time. These data imply that patients with active IBD benefit from both telemonitoring and standard care, and while telemonitoring may not improve disease activity compared with in-person visits, it may enhance other health outcomes, reduce costs, decrease time off work, and lower the number of hospital visits, among other benefits.

In this sense, the TECCU app previously demonstrated greater cost-effectiveness than standard care in a similar subset of patients [24,25], as subsequently reproduced in other settings [37,38]. Therefore, telemonitoring stands out as an interesting alternative to standard care in the follow-up of patients with IBD, whether they are experiencing active disease or not. Either

XSL•FO
RenderX

of these options could be used complementarily during a patient's follow-up, based on the preferences of both patients and health care providers in each case. In this regard, telemonitoring apps can leverage the advancements achieved during the COVID-19 pandemic [39].

As witnessed here, enhanced QoL and medication adherence are generally inversely correlated with disease activity [7]. The IBDQ-9 and the MMAS-8 medication adherence scores improved in both the TECCU and control groups after 12 weeks, indicating a trend toward statistical significance in the noninferiority of G_TECCU compared with G_Control in medication adherence. The improvement in HRQoL could be attributed to the reduction in disease activity in both groups following the escalation of medication, rather than solely to the follow-up method itself. Furthermore, the enhancement in HRQoL was similarly linked to a significant improvement in work productivity and a reduction in activity impairment (as measured by the WPAI) at 12 weeks compared with baseline in both groups.

Another notable finding was that a higher proportion of patients in the G_TECCU group had emergency department visits. However, there were no significant differences between the groups concerning unscheduled outpatient visits, hospitalizations, or AEs. Taken together, these results indicate that the telemonitoring follow-up schedule should be tailored to each patient, and the learning curve for implementing this approach may vary between centers. This aspect will be further analyzed after the completion of the 12-month follow-up period of this trial.

The efficacy of telemonitoring lies in its capacity to deliver interventions tailored to the varying clinical scenarios presented by patients who are open to using remote care, rather than solely focusing on how these interventions are provided. In fact, various patient profiles have been identified as predictors of failure in a telemedicine context, partly due to the diverse demographic and socioeconomic characteristics of the patients [40]. In this context, patient satisfaction remained stable in both the G_TECCU and G_Control groups in our trial, with values around 90% after 12 weeks. It is important to note that the G_TECCU group had received standard care before being enrolled in the study, and the switch to telemonitoring did not diminish their level of satisfaction, even though their disease activity was uncontrolled at that time.

Strengths and Limitations

The main strength of this study is that it was designed as a randomized controlled trial, allowing us to assess the impact of a web-based telemonitoring system on disease outcomes in the follow-up of IBD and to compare this with standard care. Additionally, we selected patients during the induction phase of biological agents or those initiating immunosuppressants due to active IBD or a flare-up, a specific population of IBD that has been underrepresented in previous studies [8,13,15,19,41].

Moreover, the use of validated indices of activity and biological markers (CRP and FC) that correlate well with endoscopic activity [42-45] added value to the assessment of remote monitoring. Additionally, the follow-up schedule was consistent across both arms and was adapted to daily practices in the management of IBD, which enhances the reproducibility of the results obtained. Finally, this nationwide study includes patients recruited from 24 hospitals across Spain, further enhancing the external validity of our findings.

Among the limitations of this study, the follow-up period was only 12 weeks. While this time frame may be adequate to evaluate the induction phase of therapies, it may be too short to fully assess the maintenance period, especially because patients and physicians need some time to learn how to use the telemonitoring platform. Therefore, trials with longer follow-up periods will be necessary to confirm the efficacy of this web-based system in improving long-term disease outcomes. Indeed, the follow-up of the patients enrolled in this study continues, and there are intentions to extend this period to 52 weeks for as many of these patients as possible.

Besides, the sample size may be too small to compare disease outcomes in specific patient subgroups (eg, patients with UC or CD and those receiving different medications). In addition, given the nature of the interventions assessed, neither the patients nor the researchers were blinded to the intervention; however, the data obtained were analyzed by an independent statistician who was blind to group identification. Finally, we excluded patients with active perianal disease, ileo-rectal/ileo-pouch anal anastomosis, or stoma; therefore, the impact of TECCU on surgical patients, in whom physical examination is very important, remains unclear.

Conclusions

This Spanish multicenter trial conducted on behalf of GETECCU studied patients with active IBD who began using biological agents or immunosuppressant drugs, demonstrating that the TECCU app is not inferior to standard care for achieving and maintaining remission in the short term. Noninferiority could represent a meaningful advancement if, as seen in other studies, telemonitoring improves additional health outcomes and reduces costs. The TECCU app is also associated with improved medication adherence and HRQoL; however, a reduction in costs still needs to be demonstrated over a longer follow-up period.

Acknowledgments

This work was supported by the Instituto de Salud Carlos III through project No. PI18/00593 (co-funded by the European Regional Development Fund "A way to make Europe") and an Investigator Initiated Study of J&J No. CNTO1275CRD4026. The authors thank the patients and specialized IBD nurses who participated in this study. We have obtained permission to use the MMAS-8 (8-item Morisky Medication Adherence Scale) for this study from the scale developer, Dr Donald Morisky, and his team. © MMAS 2006 www.adherence.cc.

XSL•FO
RenderX

Data Availability

The data presented in this article will be shared on reasonable request to the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT E-HEALTH checklist.
[PDF File (Adobe PDF File), 1249 KB - jmir_v26i1e60966_app1.pdf]

References

1. Høivik ML, Moum B, Solberg IC, Henriksen M, Cvancarova M, Bernklev T, IBSEN Group. Work disability in inflammatory bowel disease patients 10 years after disease onset: results from the IBSEN Study. Gut 2013 Mar;62(3):368-375. [doi: 10.1136/gutjnl-2012-302311] [Medline: 22717453]
2. Kappelman MD, Porter CQ, Galanko JA, Rifas-Shiman SL, Ollendorf DA, Sandler RS, et al. Utilization of healthcare resources by U.S. children and adults with inflammatory bowel disease. Inflamm Bowel Dis 2011 Jan;17(1):62-68 [FREE Full text] [doi: 10.1002/ibd.21371] [Medline: 20564532]
3. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet 2017 Dec 23;390(10114):2769-2778. [doi: 10.1016/S0140-6736(17)32448-0] [Medline: 29050646]
4. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012 Jan;142(1):46-54.e42; quiz e30. [doi: 10.1053/j.gastro.2011.10.001] [Medline: 22001864]
5. Chaparro M, Garre A, Núñez Ortiz A, Diz-Lois Palomares MT, Rodríguez C, Riestra S, The EpidemIBD Study Group Of Geteccu. Incidence, clinical characteristics and management of inflammatory bowel disease in Spain: large-scale epidemiological study. J Clin Med 2021 Jun 29;10(13):2885 [FREE Full text] [doi: 10.3390/jcm10132885] [Medline: 34209680]
6. Aguas Peris M, Del Hoyo J, Bebia P, Faubel R, Barrios A, Bastida G, et al. Telemedicine in inflammatory bowel disease: opportunities and approaches. Inflamm Bowel Dis 2015 Feb;21(2):392-399. [doi: 10.1097/MIB.0000000000000241] [Medline: 25437818]
7. Del Hoyo J, Millán M, Garrido-Marín A, Aguas M. Are we ready for telemonitoring inflammatory bowel disease? A review of advances, enablers, and barriers. World J Gastroenterol 2023 Feb 21;29(7):1139-1156 [FREE Full text] [doi: 10.3748/wjg.v29.i7.1139] [Medline: 36926667]
8. Elkjaer M, Shuhaibar M, Burisch J, Bailey Y, Scherfig H, Laugesen B, et al. E-health empowers patients with ulcerative colitis: a randomised controlled trial of the web-guided 'Constant-care' approach. Gut 2010 Dec;59(12):1652-1661. [doi: 10.1136/gut.2010.220160] [Medline: 21071584]
9. Carlsen K, Jakobsen C, Houen G, Kallemose T, Paerregaard A, Riis LB, et al. Self-managed eHealth disease monitoring in children and adolescents with inflammatory bowel disease: a randomized controlled trial. Inflamm Bowel Dis 2017 Mar;23(3):357-365. [doi: 10.1097/MIB.0000000000001026] [Medline: 28221247]
10. Carlsen K, Houen G, Jakobsen C, Kallemose T, Paerregaard A, Riis LB, et al. Individualized infliximab treatment guided by patient-managed eHealth in children and adolescents with inflammatory bowel disease. Inflamm Bowel Dis 2017 Sep;23(9):1473-1482. [doi: 10.1097/MIB.0000000000001170] [Medline: 28617758]
11. Heida A, Dijkstra A, Muller Kobold A, Rossen JW, Kindermann A, Kokke F, et al. Efficacy of home telemonitoring versus conventional follow-up: a randomized controlled trial among teenagers with inflammatory bowel disease. J Crohns Colitis 2018 Mar 28;12(4):432-441. [doi: 10.1093/ecco-jcc/jjx169] [Medline: 29228230]
12. Yin AL, Hachuel D, Pollak JP, Scherl EJ, Estrin D. Digital health apps in the clinical care of inflammatory bowel disease: scoping review. J Med Internet Res 2019 Aug 19;21(8):e14630 [FREE Full text] [doi: 10.2196/14630] [Medline: 31429410]
13. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, Becx MC, Maljaars JP, Cilissen M, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. Lancet 2017 Sep 02;390(10098):959-968. [doi: 10.1016/S0140-6736(17)31327-2] [Medline: 28716313]
14. Johnson MW, Lithgo K, Price T. OC-080 Ibd-Sshamp (Supported, Self help and Management Programme); UK's first internet based remote management system for managing stable IBD. Gut 2013 Jun 04;62(Suppl 1):A34.3-A3A35. [doi: 10.1136/gutjnl-2013-304907.079]
15. Cross RK, Langenberg P, Regueiro M, Schwartz DA, Tracy JK, Collins JF, et al. A randomized controlled trial of TELEmedicine for patients with inflammatory bowel disease (TELE-IBD). Am J Gastroenterol 2019 Mar;114(3):472-482. [doi: 10.1038/s41395-018-0272-8] [Medline: 30410041]

XSL•FO
RenderX

16. Zhen J, Marshall JK, Nguyen GC, Atreja A, Narula N. Impact of digital health monitoring in the management of inflammatory bowel disease. J Med Syst 2021 Jan 15;45(2):23-28 [FREE Full text] [doi: 10.1007/s10916-021-01706-x] [Medline: 33449213]
17. van Deen WK, Spiro A, Burak Ozbay A, Skup M, Centeno A, Duran N, et al. The impact of value-based healthcare for inflammatory bowel diseases on healthcare utilization: a pilot study. Eur J Gastroenterol Hepatol 2017 Mar;29(3):331-337. [doi: 10.1097/MEG.0000000000000782] [Medline: 27926663]
18. Turner D, Ricciuto A, Lewis A, D'Amico F, Dhaliwal J, Griffiths AM, International Organization for the Study of IBD. STRIDE-II: An Update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology 2021 Apr;160(5):1570-1583. [doi: 10.1053/j.gastro.2020.12.031] [Medline: 33359090]
19. Cross RK, Cheevers N, Rustgi A, Langenberg P, Finkelstein J. Randomized, controlled trial of home telemanagement in patients with ulcerative colitis (UC HAT). Inflamm Bowel Dis 2012 Jun;18(6):1018-1025 [FREE Full text] [doi: 10.1002/ibd.21795] [Medline: 21688350]
20. Del Hoyo J, Nos P, Faubel R, Muñoz D, Domínguez D, Bastida G, et al. A web-based telemanagement system for improving disease activity and quality of life in patients with complex inflammatory bowel disease: pilot randomized controlled trial. J Med Internet Res 2018 Nov 27;20(11):e11602 [FREE Full text] [doi: 10.2196/11602] [Medline: 30482739]
21. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012 Dec 26;308(24):2594-2604. [doi: 10.1001/jama.2012.87802] [Medline: 23268518]
22. Jaikumar V. Non-inferiority trials: understanding the concepts - tutorials and fundamentals. Stud 4 Best Evid. URL: https://s4be.cochrane.org/blog/2022/03/18/understanding-non-inferiority-trials/ [accessed 2023-07-21]
23. Aguas M, Del Hoyo J, Faubel R, Muñoz D, Domínguez D, Bastida G, et al. A web-based telemanagement system for patients with complex inflammatory bowel disease: protocol for a randomized controlled clinical trial. JMIR Res Protoc 2018 Dec 21;7(12):e190 [FREE Full text] [doi: 10.2196/resprot.9639] [Medline: 30578197]
24. Del Hoyo J, Nos P, Bastida G, Faubel R, Muñoz D, Garrido-Marín A, et al. Telemonitoring of Crohn's disease and ulcerative colitis (TECCU): cost-effectiveness analysis. J Med Internet Res 2019 Sep 13;21(9):e15505 [FREE Full text] [doi: 10.2196/15505] [Medline: 31538948]
25. Del Hoyo J, Aguas M. Cost-effectiveness of telemedicine-directed specialized vs standard care for patients with inflammatory bowel diseases in a randomized trial. Clin Gastroenterol Hepatol 2021 Jan;19(1):206-207. [doi: 10.1016/j.cgh.2020.06.016] [Medline: 33339564]
26. Magro F, Gionchetti P, Eliakim R, Ardizzone S, Armuzzi A, Barreiro-de Acosta M, European Crohn'sColitis Organisation [ECCO]. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders. J Crohns Colitis 2017 Jun 01;11(6):649-670. [doi: 10.1093/ecco-jcc/jjx008] [Medline: 28158501]
27. Gomollón F, Dignass A, Annese V, Tilg H, Van Assche G, Lindsay JO, et al. 3rd European Evidence-based Consensus on the Diagnosis and Management of Crohn's Disease 2016: part 1: diagnosis and medical management. J Crohns Colitis 2017 Jan;11(1):3-25 [FREE Full text] [doi: 10.1093/ecco-jcc/jjw168] [Medline: 27660341]
28. Del Hoyo J, Nos P, Faubel R, Bastida G, Muñoz D, Valero-Pérez E, et al. Adaptation of TECCU app based on patients´ perceptions for the telemonitoring of inflammatory bowel disease: a qualitative study using focus groups. Int J Environ Res Public Health 2020 Mar 13;17(6):1871 [FREE Full text] [doi: 10.3390/ijerph17061871] [Medline: 32183103]
29. Sicilia B, García-López S, González-Lama Y, Zabana Y, Hinojosa J, Gomollón F, et al. GETECCU 2020 guidelines for the treatment of ulcerative colitis. Developed using the GRADE approach. Gastroenterol Hepatol 2020 Aug;43 Suppl 1:1-57. [doi: 10.1016/j.gastrohep.2020.07.001] [Medline: 32807301]
30. Harvey RF, Bradshaw JM. A simple index of Crohn's-disease activity. Lancet 1980 Mar 08;1(8167):514. [doi: 10.1016/s0140-6736(80)92767-1] [Medline: 6102236]
31. Walmsley RS, Ayres RC, Pounder RE, Allan RN. A simple clinical colitis activity index. Gut 1998 Jul;43(1):29-32 [FREE Full text] [doi: 10.1136/gut.43.1.29] [Medline: 9771402]
32. Vergara M, Montserrat A, Casellas F, Villoria A, Suarez D, Maudsley M, et al. A new validation of the Spanish Work Productivity and Activity Impairment Questionnaire-Crohn's disease version. Value Health 2011;14(6):859-861 [FREE Full text] [doi: 10.1016/j.jval.2011.02.1179] [Medline: 21914506]
33. Berlowitz DR, Foy CG, Kazis LE, Bolin LP, Conroy MB, Fitzpatrick P, SPRINT Research Group. Effect of intensive blood-pressure treatment on patient-reported outcomes. N Engl J Med 2017 Aug 24;377(8):733-744 [FREE Full text] [doi: 10.1056/NEJMoa1611179] [Medline: 28834483]
34. Bress AP, Bellows BK, King JB, Hess R, Beddhu S, Zhang Z, SPRINT Research Group. Cost-effectiveness of intensive versus standard blood-pressure control. N Engl J Med 2017 Aug 24;377(8):745-755 [FREE Full text] [doi: 10.1056/NEJMsa1616035] [Medline: 28834469]
35. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]

XSL•FO
RenderX

36. Cross RK, Finkelstein J. Feasibility and acceptance of a home telemanagement system in patients with inflammatory bowel disease: a 6-month pilot study. Dig Dis Sci 2007 Feb;52(2):357-364. [doi: 10.1007/s10620-006-9523-4] [Medline: 17211702]
37. de Jong MJ, Boonen A, van der Meulen-de Jong AE, Romberg-Camps MJ, van Bodegraven AA, Mahmmod N, et al. Cost-effectiveness of telemedicine-directed specialized vs standard care for patients with inflammatory bowel diseases in a randomized trial. Clin Gastroenterol Hepatol 2020 Jul;18(8):1744-1752. [doi: 10.1016/j.cgh.2020.04.038] [Medline: 32335133]
38. Yao J, Fekadu G, Jiang X, You JHS. Telemonitoring for patients with inflammatory bowel disease amid the COVID-19 pandemic-a cost-effectiveness analysis. PLoS One 2022;17(4):e0266464 [FREE Full text] [doi: 10.1371/journal.pone.0266464] [Medline: 35390064]
39. Ciencia YT, sociedad DLI. 6. 4 Población que usa Internet (en los últimos tres meses). URL: https://tinyurl.com/bdhwfz3t [accessed 2023-09-27]
40. Gellad ZF, Diamond S, Crockett SD, Cross RK. AGA clinical practice update on telemedicine in gastroenterology: commentary. Gastroenterology 2023 Apr;164(4):690-695. [doi: 10.1053/j.gastro.2022.12.043] [Medline: 36775722]
41. McCombie A, Walmsley R, Barclay M, Ho C, Langlotz T, Regenbrecht H, et al. A noninferiority randomized clinical trial of the use of the smartphone-based health applications IBDsmart and IBDoc in the care of inflammatory bowel disease patients. Inflamm Bowel Dis 2020 Jun 18;26(7):1098-1109. [doi: 10.1093/ibd/izz252] [Medline: 31644793]
42. Higgins PDR, Schwartz M, Mapili J, Zimmermann EM. Is endoscopy necessary for the measurement of disease activity in ulcerative colitis? Am J Gastroenterol 2005 Feb;100(2):355-361 [FREE Full text] [doi: 10.1111/j.1572-0241.2005.40641.x] [Medline: 15667493]
43. Lin J, Chen J, Zuo J, Yu A, Xiao Z, Deng F, et al. Meta-analysis: fecal calprotectin for assessment of inflammatory bowel disease activity. Inflamm Bowel Dis 2014 Aug;20(8):1407-1415. [doi: 10.1097/MIB.0000000000000057] [Medline: 24983982]
44. Zittan E, Kabakchiev B, Kelly OB, Milgrom R, Nguyen GC, Croitoru K, et al. Development of the Harvey-Bradshaw Index-pro (HBI-PRO) Score to assess endoscopic disease activity in Crohn's disease. J Crohns Colitis 2017 May 01;11(5):543-548. [doi: 10.1093/ecco-jcc/jjw200] [Medline: 28453763]
45. Marín-Jiménez I, Nos P, Domènech E, Riestra S, Gisbert JP, Calvet X, et al. Diagnostic performance of the simple Clinical Colitis Activity Index self-administered online at home by patients with ulcerative colitis: CRONICA-UC study. Am J Gastroenterol 2016 Feb;111(2):261-268. [doi: 10.1038/ajg.2015.403] [Medline: 26753886]

Abbreviations

AE: adverse event
CD: Crohn disease
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CRP: C-reactive protein
FC: fecal calprotectin
HBI: Harvey-Bradshaw Index
HRQoL: health-related quality of life
IBD: inflammatory bowel disease
IBDQ-9: 9-item Inflammatory Bowel Disease Questionnaire
mHealth: mobile health
MMAS-8: 8-item Morisky Medication Adherence Scale
SCCAI: Simple Clinical Colitis Activity Index
TECCU: Telemonitorización de la Enfermedad de Crohn y Colitis Ulcerosa or Telemonitoring of Crohn Disease and Ulcerative Colitis
UC: ulcerative colitis
WPAI: Work productivity and activity impairment

XSL•FO
RenderX

Edited by G Eysenbach, A Mavragani, A Coristine; submitted 03.06.24; peer-reviewed by X Ma; comments to author 19.07.24; revised version received 08.08.24; accepted 21.08.24; published 18.11.24.

Please cite as:
Aguas M, Del Hoyo J, Vicente R, Barreiro-de Acosta M, Melcarne L, Hernandez-Camba A, Madero L, Arroyo MT, Sicilia B, Chaparro M, Martin-Arranz MD, Pajares R, Mesonero F, Mañosa M, Martinez P, Chacón S, Tosca J, Marín S, Sanroman L, Calvo M, Monfort D, Saiz E, Zabana Y, Guerra I, Varela P, Baydal V, Faubel R, Corsino P, Porto-Silva S, Brunet E, González M, Gutiérrez A, Nos P Telemonitoring of Active Inflammatory Bowel Disease Using the App TECCU: Short-Term Results of a Multicenter Trial of GETECCU J Med Internet Res 2024;26:e60966
URL: https://www.jmir.org/2024/1/e60966
doi:10.2196/60966
PMID:39189160

©Mariam Aguas, Javier Del Hoyo, Raquel Vicente, Manuel Barreiro-de Acosta, Luigi Melcarne, Alejandro Hernandez-Camba, Lucía Madero, María Teresa Arroyo, Beatriz Sicilia, María Chaparro, María Dolores Martin-Arranz, Ramón Pajares, Francisco Mesonero, Miriam Mañosa, Pilar Martinez, Silvia Chacón, Joan Tosca, Sandra Marín, Luciano Sanroman, Marta Calvo, David Monfort, Empar Saiz, Yamile Zabana, Ivan Guerra, Pilar Varela, Virginia Baydal, Raquel Faubel, Pilar Corsino, Sol Porto-Silva, Eduard Brunet, Melodi González, Ana Gutiérrez, Pilar Nos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telemedicine Integrated Care Versus In-Person Care Mode for Patients With Short Stature: Comprehensive Comparison of a Retrospective Cohort Study

Yipei Wang[1*], MBA, PhD; Pei Zhang[2*], ME; Yan Xing[3], MD; Huifeng Shi[4], PhD; Yunpu Cui[3], MD; Yuan Wei[4], MD; Ke Zhang[5,6], MD; Xinxia Wu[6,7], MM; Hong Ji[2], PhD; Xuedong Xu[6,7], MBA; Yanhui Dong[8], PhD; Changxiao Jin[1,9], MS

[1]Institute of Hospital Management, Peking University Third Hospital, Beijing, China
[2]Information Management and Big Data Center, Peking University Third Hospital, Beijing, China
[3]Department of Paediatrics, Peking University Third Hospital, Beijing, China
[4]Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China
[5]Department of Otolaryngology, Peking University Third Hospital, Beijing, China
[6]Office of Internet Hospital, Peking University Third Hospital, Beijing, China
[7]Department of Medical Affairs, Peking University Third Hospital, Beijing, China
[8]Institute of Child and Adolescent Health, School of Public Health, Peking University, Beijing, China
[9]Party Committee Office and Dean's Office, Peking University Third Hospital, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Changxiao Jin, MS
Institute of Hospital Management
Peking University Third Hospital
North Huayuan Road 49
Beijing, 100191
China
Phone: 86 010 82266543
Email: jinchangxiao2023@163.com

Abstract

Background: Telemedicine has demonstrated efficacy as a supplement to traditional in-person care when treating certain diseases. Nevertheless, more investigation is needed to comprehensively assess its potential as an alternative to in-person care and its influence on access to care. The successful treatment of short stature relies on timely and regular intervention, particularly in rural and economically disadvantaged regions where the disease is more prevalent.

Objective: This study evaluated the clinical outcomes, health-seeking behaviors, and cost of telemedicine integrated into care for children with short stature in China.

Methods: Our study involved 1241 individuals diagnosed with short stature at the pediatric outpatient clinic of Peking University Third Hospital between 2012 and 2023. Patients were divided into in-person care (IPC; 1183 patients receiving only in-person care) and telemedicine integrated care (TIC; 58 patients receiving both in-person and virtual care) groups. For both groups, the initial 71.43% (average of 58 percentages, with each percentage representing the ratio of patients in the treatment group) of visits were categorized into the pretelemedicine phase. We used propensity score matching to select individuals with similar baseline conditions. We used 7 variables such as age, gender, and medical insurance for the 1:5 closest neighbor match. Eventually, 115 patients in the IPC group and 54 patients in the TIC group were selected. The primary clinical outcome was the change in the standard height percentage. Health-seeking behavior was described by visit intervals in the pre- and post-telemedicine phases. The cost analysis compared costs both between different groups and between different visit modalities of the TIC group in the post-telemedicine phase.

Results: In terms of clinical effectiveness, we demonstrated that the increase in height among the TIC group (Δz_{TIC}=0.74) was more substantial than that for the IPC group (Δz_{IPC}=0.51, P=.01; paired t test), while no unfavorable changes in other endpoints such as BMI or insulin-like growth factor 1 (IGF-1) levels were observed. As for health-seeking behaviors, the results showed

XSL•FO
RenderX

that, during the post-telemedicine phase, the IPC group had a visit interval of 71.08 (IQR 50.75-90.73) days, significantly longer than the prior period (51.25 [IQR 34.75-82.00] days, $P<.001$; *U* test), whereas the TIC group's visit interval remained unchanged. As for the cost per visit, there was no difference in the average cost per visit between the 2 groups nor between the pre- and post-telemedicine phases. During the post-telemedicine phase, within the TIC group, in-person visits had a higher average total cost, elevated medical and labor expenses, and greater medical cost compared with virtual visits.

Conclusions: We contend that the rise in medical visits facilitated by integrating telemedicine into care effectively restored the previously constrained number of medical visits to their usual levels, without increasing costs. Our research underscores that administering prompt treatment may enable physicians to seize a crucial treatment opportunity for children with short stature, thus attaining superior results.

(J Med Internet Res 2024;26:e57814) doi:10.2196/57814

KEYWORDS

telemedicine; telemedicine integrated care mode; short stature; clinical outcomes; health-seeking behaviors; cost analysis; in-person care; mobile health; mHealth; telehealth; eHealth; video virtual visit; access to care; children; pediatrics; China; accessibility; temporal; spatial constraints; chronic disease

Introduction

From the original transmission of electrocardiograms to the current use of asynchronous and synchronous audio and video virtual visits, telemedicine has become a service approach that provide patients with prevention, diagnosis, treatment, and continuing education [1-3]. Since the successful implementation of the Space Technology Applied to Rural Papago Advanced Health Care project, telemedicine has demonstrated its value and potential as a useful tool for physicians to provide quality health care to people in a remote area [4]. Integrating telemedicine into care has demonstrated equivalent or superior clinical outcomes to traditional in-person care when treating certain diseases, including diabetes [5,6] and hypertension [7,8]. However, in numerous studies, telemedicine was mostly used as a supplementary tool for conventional care, serving as a platform for patient-physician interactions and a channel for patient education [9-12]. To address the need for health care in remote areas, further investigation is required to determine the therapeutic efficacy of telemedicine-enabled virtual visits as a substitute for in-person visits.

Telemedicine not only delivers clinical effectiveness but also enhances accessibility. Telemedicine transcends temporal and spatial constraints [13], mitigating the risk of delayed or missed medical consultations for patients with chronic diseases [8,14,15]. Consequently, it curtails the financial burden associated with untimely interventions [16,17]. The assessment of access mostly relies on the quantification of medical consultations (ie, number of visits in a given period). Research on telemedicine's impact on improving access can be categorized into 2 distinct groups. The first type of research centered around telemedicine shows that it helps increase health care delivery for patients living in remote areas [18,19]. Another stream of study is based on the premise that the lockdown resulting from the COVID-19 pandemic led to a decrease in in-person outpatient appointments, which was partially compensated by the utilization of telemedicine [20-22]. Although the limits on medical resources resulting from COVID-19 have been alleviated, people in remote and underdeveloped regions continue to grapple with the issue of inadequate medical resources. The significance of telemedicine's contribution to enhancing accessibility thus remains a subject of importance.

Short stature is a medical term used to describe a condition in which a person's height is at least 2 SD below the average height for their age, sex, and population group [23]. Height is a personal attribute that is determined by a combination of hereditary and environmental variables [24]. The prevalence of short stature in children ranges from 0.5% to 11.3% [23]. Although the prevalence varies depending on country, people, and gender, studies have revealed that rural areas exhibit a higher prevalence than urban areas, while less economically developed regions have a higher prevalence than more economically developed regions [25]. In these rural or lower-income regions, monetary resources alone are insufficient: Using technological methods to modify the service model is critical to address the issue of short stature [26]. Intervention has a specific time frame within which it is most effective; hence, prompt and consistent intervention might improve outcomes [27]. Furthermore, short stature is listed as one of the main reasons for pediatric consultations [28]. Hence, comprehensive comparison of telemedicine integrated into care with in-person care for children with short stature provides an example for future technological interventions and impact evaluations in remote and underprivileged regions.

Currently, there is a lack of evaluation of telemedicine considering both clinical outcomes and access to care. We hypothesized that telemedicine improves accessibility by allowing patients to receive more timely visits than in-person care, leading to improved outcomes (Figure 1). To investigate this theory, we used short stature as an example and conducted a retrospective cohort study including patients who only had in-person visits (in-person care [IPC] group), and patients who had both virtual and in-person visits (telemedicine integrated care [TIC] group), which started to occur when certain individuals pursued virtual care because of concerns such as avoiding infection [29]. We performed a comprehensive comparison between the 2 cohorts, examining clinical outcomes, health-seeking behaviors, and average cost per visit associated with the 2 treatment modalities. The "time of switch" indicates the ratio of the number of visits related to the first virtual visit to the total number of visits in the TIC group. Each in-person

XSL•FO
RenderX

visit involves 4 categories of costs: medical, labor, transportation, and loss of productivity. Every virtual visit only involves the first 2 categories of costs.

Figure 1. Illustration of conventional in-person and telemedicine integrated care modes.

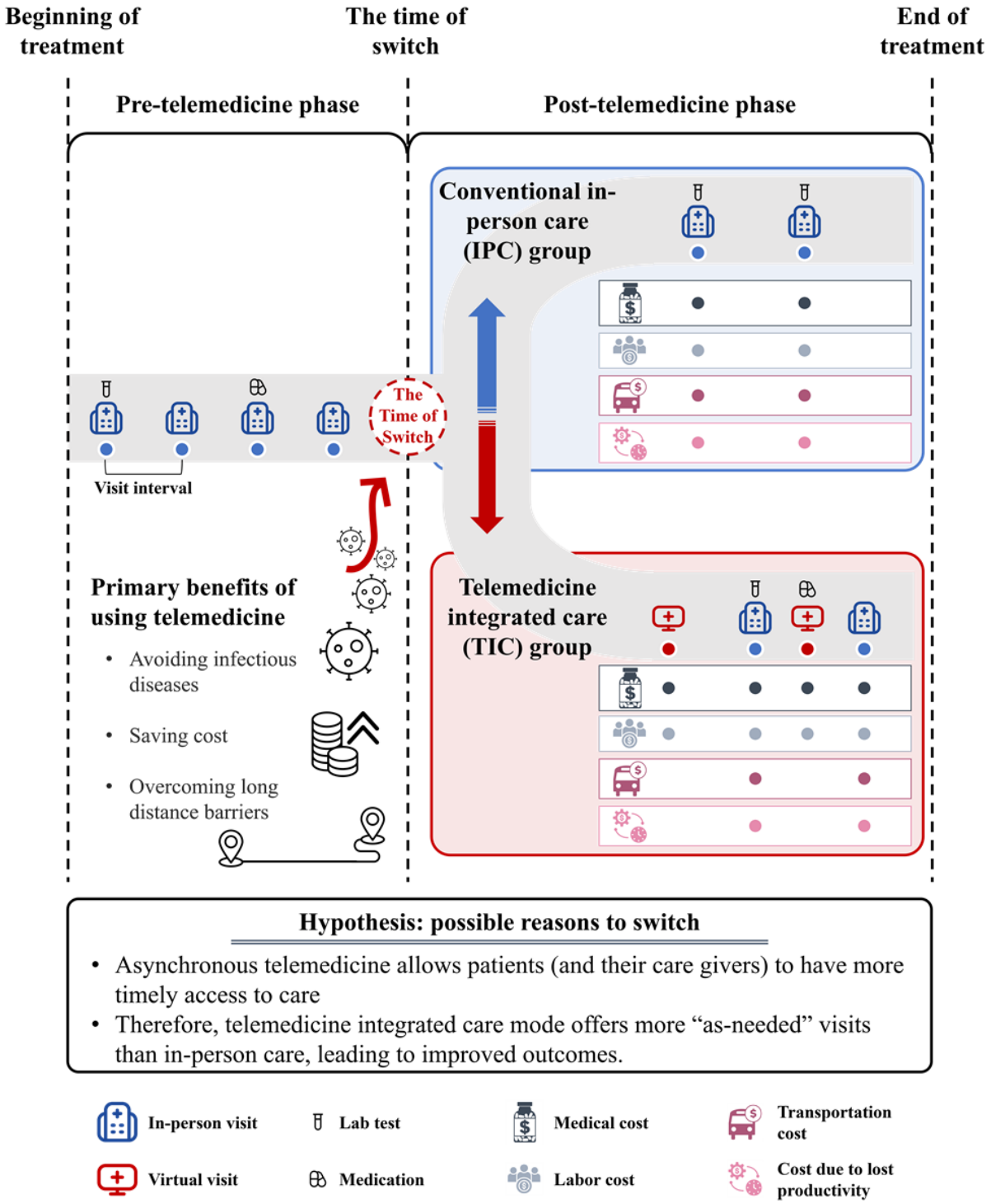

Methods

Telemedicine Integrated Care Mode

The Department of Paediatrics at Peking University Third Hospital provides comprehensive services for patients with short stature, including both in-person and virtual visits. The virtual visits are done asynchronously, meaning the physicians respond to the patients' inquiries at their own convenience [30]. Similar to in-person appointments, virtual visits enable physicians to prescribe medications and request laboratory tests for patients. After the initial in-person visit with a pediatrician, patients have the option to either continue with conventional in-person visits or transition to the integrated telemedicine mode (Figure 1). However, with regards to medical insurance reimbursement, the reimbursement policy permits online direct settlement exclusively for medical service fees while excluding prescription fees and charges related to inspections and testing.

XSL•FO
RenderX

Patients who choose virtual visits are required to bear the financial responsibility for any medication or test fees incurred should they wish to make online payments for such expenses. In order to mitigate this phenomenon, we used the patient's specific medical insurance coverage as a variable when assessing its potential impact on health-seeking behaviors.

Study Design and Data Collection

Our study focused on patients who underwent therapy for short stature in the pediatrics outpatient department at Peking University Third Hospital from March 1, 2012, to March 31, 2023. The implementation of the electronic medical record (EMR) system occurred in the year 2012. Prior to that, the storage of medical records relied solely on paper-based systems, which posed difficulties for accessing precise and comprehensive medical information. The study did not include patients who had visit records for short stature between April 2023 and December 2023, as these recent visit records indicated incomplete treatment cycles. Patients under screening were divided into 2 cohorts based on their visit modalities. Patients who had only in-person visits were in the IPC group, and those who had both virtual and in-person care were in the TIC group (Figure 1).

During each appointment, demographic information such as age, gender, and insurance type, as well as behavioral data related to the visit, including the date, type (in-person or virtual), and diagnosis, were collected using the hospital information system. The initial occurrence of a diagnosis indicating short stature was designated as the date of diagnosis. Simultaneously, the age of the patient at the time of diagnosis was denoted as the age of diagnosis. The treatment duration was operationally defined as the temporal interval between the date of the patient's most recent diagnosis of short stature and the date of diagnosis. The visit interval was defined as the number of days between 2 consecutive visits. The mean visit interval for each individual patient was calculated, followed by computing the average interval across all patients within each group. Every patient was given equal weight, regardless of the number of visit intervals they had. Clinical outcomes, including height and levels of insulin-like growth factor 1 (IGF-1) and 25-hydroxyvitamin D (25-OH VitD3), were retrieved from the EMR to assess clinical effectiveness.

The cost data encompassed medical expenses, labor expenditures, transportation fees, and costs associated with lost production. The medical expenses associated with each visit were obtained from the EMR system. The labor cost was determined by multiplying the average physician compensation per unit of time by the average service time. The salary was derived by comparing the incomes of physicians in 3 different positions: chief physician, deputy chief physician, and attending physician. The hospital financial department submitted the data. The average service time for each in-person visit was calculated by dividing the total time spent in one outpatient unit by the average number of patient visits per outpatient unit. The outpatient unit had a length of 4 hours. In order to ascertain the length of virtual visits, which posed difficulties for measurement due to their asynchronous character, we conducted interviews with 19 physicians, who accounted for 95% (19/20) of the entire population in the Department of Paediatrics that took part in virtual visits. The calculation of transportation expenses relied on the patients' location data. The cost of a single taxi ride spanning a distance of 10 kilometers for patients living in Beijing was calculated as ¥30 (US $4.18) [31], leading to a total transportation cost of ¥60 (US $8.36) for each in-person visit. To calculate the cost for patients who lived outside of Beijing, we determined the transportation cost by considering the least expensive price for a second-class train ticket from their location to Beijing. The cost of lost productivity was calculated based on the mean salary of urban workers in Beijing during the year 2019 [32]. It was assumed that the duration of in-person visits was 4 hours and 8 hours for patients residing in Beijing and in other cities, respectively. Engaging in virtual visits did not result in any costs related to lost productivity.

Propensity Score Matching (PSM)

Of the total sample size of 1241 individuals, 1183 participants opted to proceed with in-person visits, constituting the IPC group. Conversely, 58 individuals chose to explore telemedicine as an alternative, thereby forming the TIC group. The 2 groups underwent propensity score matching (PSM) to mitigate the potential influence of unobserved confounding variables. The matching and quality assessment were conducted using the *psmatch* and *pstest* functions in Stata.

For the purpose of matching, various factors like age at diagnosis, date of diagnosis, gender, and insurance type were considered as covariances. The 2 most critical variables in characterizing patient health-seeking behavior, namely visit intervals in the pre-telemedicine phase and treatment duration, were also chosen as criteria for the PSM. The 1:5 closest neighbor with replacement technique was used for its ability to provide high-quality matches, with biases of less than 10%, across all 6 attributes (Figure 1).

Statistical Analysis

The paired *t* test was used to examine the presence of significant alterations in the standard height percentage value (*z* score) associated with the height and BMI of patients within each cohort before and after the treatment (*z* score). An independent sample *t* test was used to examine if there was a significant difference in *z* scores of height and BMI between the 2 groups. The analyses were conducted using the Stata command *ttest*. The reference ranges for IGF-1, free thyroxine (FT4), and thyroid-stimulating hormone (TSH) were used to determine normal values. By identifying individuals whose measurements fell within these normal ranges, the Stata command *ranksum* was used to assess the consistency between the 2 groups in the pre- and post-telemedicine phases. The Wilcoxon rank sum test was also used to assess if there was a statistically significant difference in 25-OH VitD3 concentration between the 2 groups. Because the distribution of visit intervals and cost did not conform to a normal distribution, the median and IQR of visit intervals and cost per visit in both the IPC and TIC groups were analyzed using the Wilcoxon rank sum test with the Stata command *ranksum*.

XSL•FO
RenderX

Ethical Considerations

This study conducted a secondary analysis of existing research data, for which informed consent was obtained during the original data collection. Consequently, the Peking University Third Hospital Medical Science Research Ethics Committee permitted this secondary analysis without requiring additional patient consent. The IRB approval number is IRB00006761-M2023394.

All research data involving human subjects in this study have been anonymized to safeguard participant privacy and data security. This research did not offer any compensation to the participants.

Results

Cohort Characteristics

From March 2012 to March 2023, a total of 1241 patients presenting with short stature sought medical treatment at the Department of Paediatrics in Peking University Third Hospital. The majority of the 1241 participants in the study opted for in-person visits with pediatricians (n=1183), whereas a smaller subset of 58 participants decided to use telemedicine at some point in the treatment (Figure 2).

The 2 patient groups were comparable in age at diagnosis, date of diagnosis, composition of medical insurance types, visits intervals, and total costs prior to the adoption of care integrating telemedicine by the TIC group. However, there was a statistically significant higher proportion of female patients in the TIC group (P=.03; chi-square). These patients also exhibited a longer median treatment length of 650 (IQR 508-1144) days, as compared with a median treatment duration of 434 (IQR 224-749) days for patients who had in-person visits only (P<.001; U test; Table 1).

XSL•FO
RenderX

Figure 2. Study flow diagram. 25-OH VitD3: 25-hydroxyvitamin D3; FT4: free thyroxine; IGF-1: insulin-like growth factor 1; TSH: thyroid stimulating hormone.

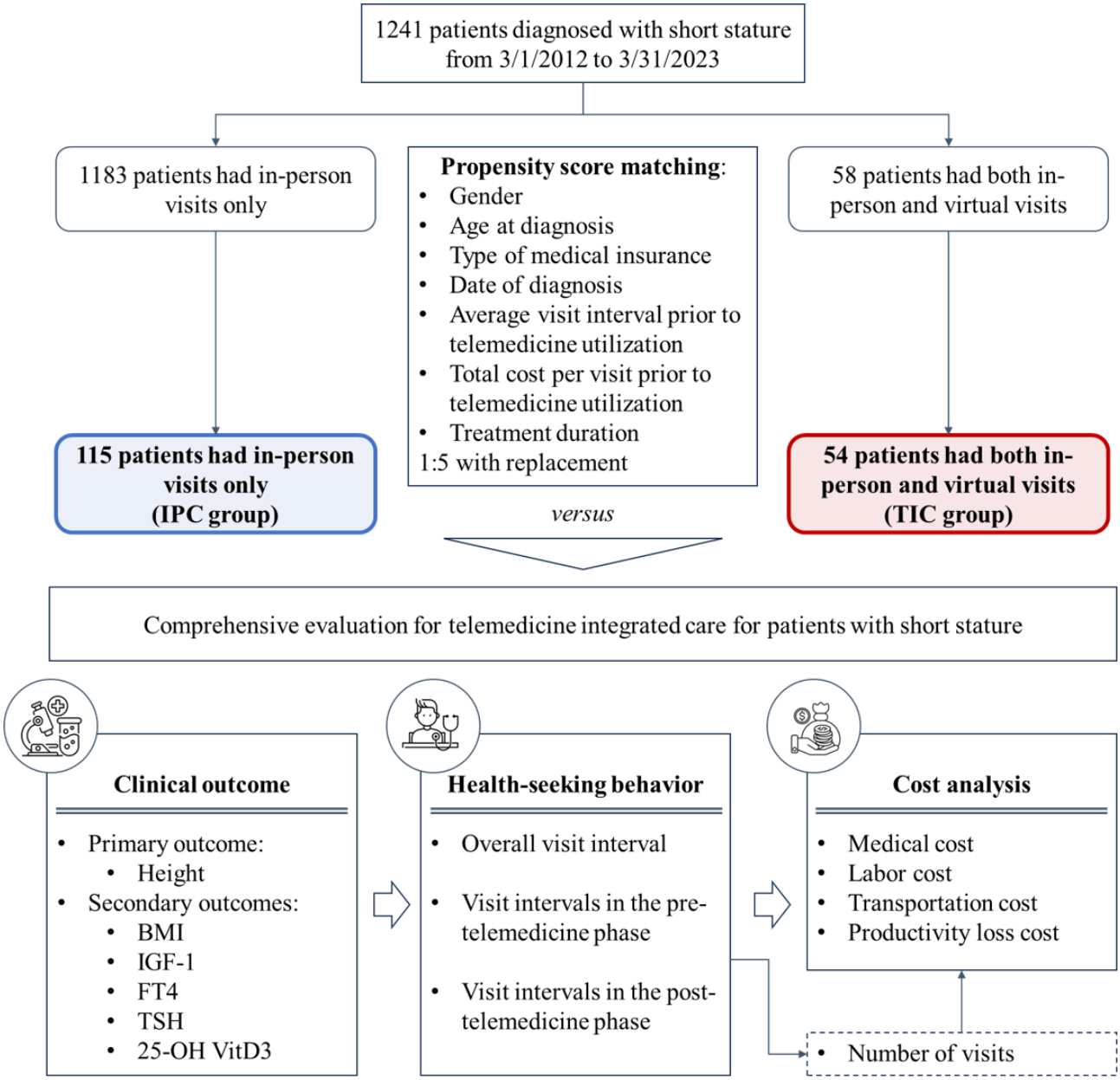

XSL•FO
RenderX

Table 1. A summary of the patients' demographic information.

Characteristics	Total sample		*P* value	PSM[a] subsample		*P* value
	IPC[b] group (n=1183)	TIC[c] (n=58)		IPC group (n=115)	TIC group (n=54)	
Female, n (%)	581 (49.1)	37 (63.8)	.03	71 (61.7)	34 (63)	.88
Age at diagnosis (years), median (IQR)	10 (8-12)	10 (8-11)	.11	10 (7-11)	10 (8-11)	.74
Medical insurance, n (%)			.22			.12
Beijing Urban and Rural Medical Insurance	990 (83.7)	52 (89.7)		111 (96.5)	49 (90.7)	
Out of pocket	189 (16)	6 (10.3)		4 (3.5)	5 (9.3)	
Cross-provincial medical insurance	4 (0.3)	0		0	0	
Date of diagnosis, n (%)			.25			.38
2013-2017	61 (5.2)	5 (8.6)		11 (9.6)	3 (5.6)	
2018-2022	1122 (94.8)	53 (91.4)		104 (90.4)	51 (94.4)	
Average visit interval in the pretelemedicine phase (days), median (IQR)	52.50 (33.80-83.89)	56.95 (36.16-76.73)	.60	51.25 (34.75-82.00)	58.13 (36.32-77.45)	.59
Treatment duration (days), median (IQR)	434 (224-749)	650 (508-1144)	<.001	652 (434-1280)	650 (508-1070)	.59
Total cost per visit in the pretelemedicine phase (¥)[d], median (IQR)	3541.93 (2308.50-5704.29)	3856.62 (2618.48-5698.33)	.23	3098.99 (2176.71-5266.23)	3412.62 (2907.62-4705.71)	.10

[a]PSM: propensity score matching.

[b]IPC: in-person care only.

[c]TIC: telemedicine integrated care, including both in-person and virtual care.

[d]A currency exchange rate of ¥1=US $0.14 is applicable.

The Time of Switch

To examine the impact of care integrating telemedicine, it is important to select patients who had comparable characteristics before the use of telemedicine. A notion called "the time of switch" was developed to divide the treatment course into pre-telemedicine and post-telemedicine phases. The time of switch refers to the ratio of the number of visits related to the first virtual visit to the total number of visits by the TIC group. In our calculation, the time of switch was 71.43% (average of 58 percentages, with each percentage representing the ratio of patients in the treatment group). As for those who only had in-person care, due to the absence of telemedicine use, we allocated the same proportion of visits as observed for the patients using both types of visits to the patients who only had in-person care. Specifically, we assigned the initial 71.43% of visits to the pretelemedicine phase and the rest to the post-telemedicine phase. During this phase, a total of 107 virtual visits were recorded. Of these, 44 patients (81.5%) subsequently had in-person visits. It is worth noting that only 1 patient had both virtual and in-person visits on the same day (Figure 1).

PSM

Following the establishment of pre- and post-telemedicine phases in both groups, we used PSM to select patients with similar baseline conditions to evaluate the clinical outcomes, change in health-seeking behaviors, and related costs. The matching process was based on 7 initial parameters in the pre-telemedicine phase, including gender, age at diagnosis, type of medical insurance, date of diagnosis, visit interval, treatment duration, and total cost per visit. A total of 54 patients constituted the TIC group and were paired with 115 patients in the IPC group. The 2 matched cohorts exhibited no statistical differences in all 7 factors; hence, we considered this matching process acceptable. The matched groups were largely female patients with Beijing Urban Residents Medical Insurance who were diagnosed at the age of 10 years during 2018 to 2022. The treatment usually lasts for about 650 days, during which the patients had in-person visits once every 51 or 58 days, with a total cost of over ¥3000 (US $471.82) per visit (Table 1). These patients were included for further analysis of clinical outcomes, health-seeking behaviors, and costs.

Clinical Outcomes

Alteration in Height

The primary outcome to assess the impact of interventions targeting short stature is the measurement of height. Height was measured at the beginning and end of treatment using the standard height percentage value, or *z* score, established by the

XSL•FO
RenderX

World Health Organization (WHO) for children aged 0 years to 18 years. The height *z* scores for patients improved from –0.556 to –0.044 in the IPC group and –0.698 to 0.047 in the TIC group. The *z* score in both groups exhibited a statistically significant increase, yet the improvement for patients in the TIC group was more pronounced (*P*=.01; paired *t* test; Figure 3A).

Figure 3. Pre- and post-telemedicine z score distributions of (A) height and (B) BMI, with the dashed lines representing the mean value of the z scores for each cohort. IPC: in-patient care only; TIC: telemedicine integrated care.

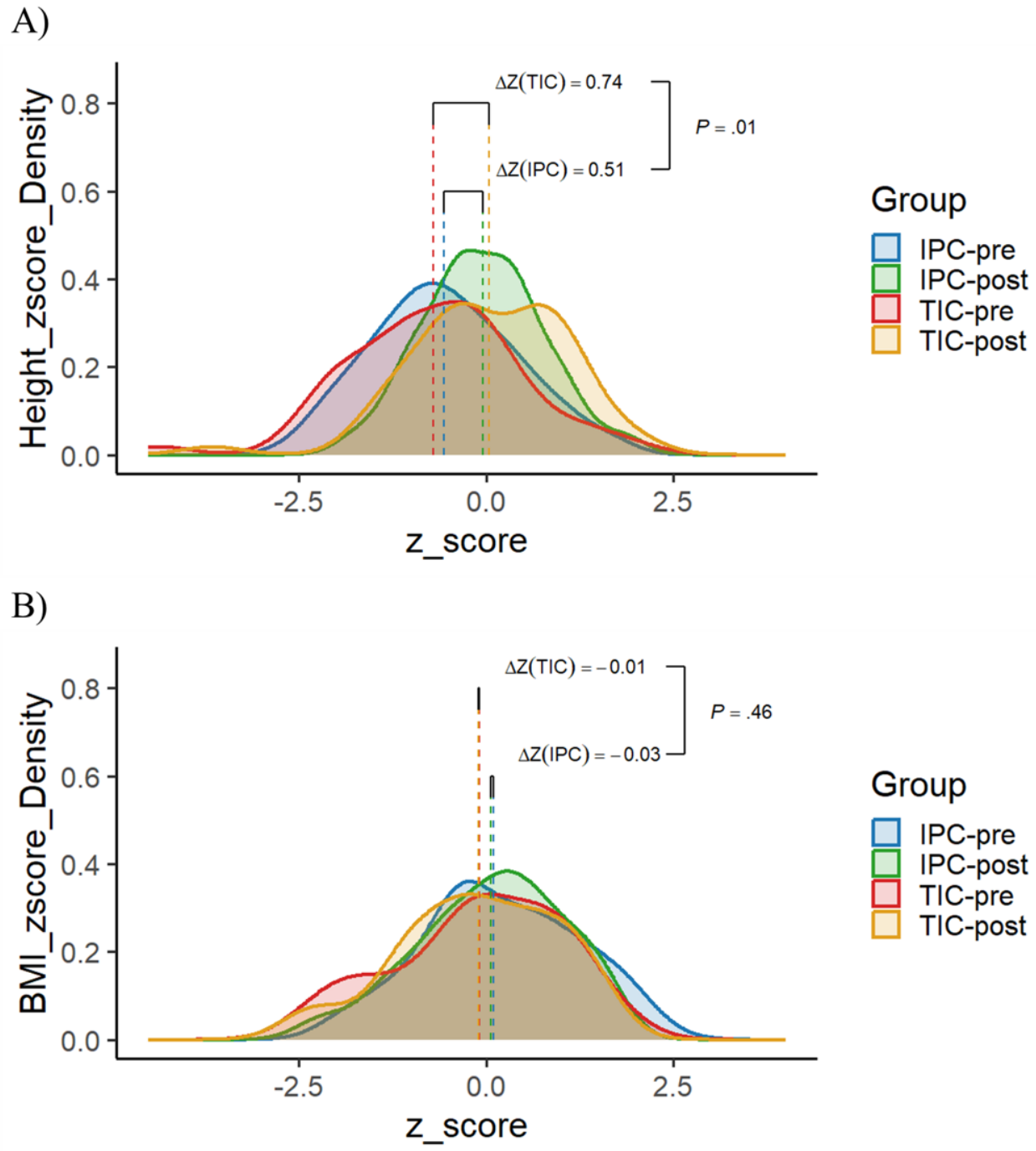

Alteration in BMI

Comparing BMI with the appropriate age- and sex-specific growth chart enables pediatric health care providers to monitor growth and identify potential health- or nutrition-related problems. Similar to height, BMI was measured at the beginning and end of treatment and evaluated according to WHO standards. As the height of both groups experienced a substantial increase, the BMI *z* scores of both groups exhibited a decrease. Specifically, the *z* score of the IPC group decreased by 0.03, while that of the TIC group decreased by 0.01. However, it is worth noting that the BMI of both cohorts remained within the normal range and there was no significant disparity in the extent of decrease (Figure 3B).

XSL•FO
RenderX

Alterations in Other Clinical Indicators

Change in IGF-1

IGF-1 is important for the processes of skeletal and general growth and development. Consequently, it is frequently investigated in the context of therapeutic interventions for individuals with short stature. This study observed a high percentage (53/54, 98%) of patients using care integrating telemedicine who had their IGF-1 levels measured both at the outset and toward the end of their treatment. At the beginning of treatment, the majority of patients (IPC group: 50/94, 53%; TIC group: 33/53, 62%) had normal IGF-1 levels. However, subsequent analysis revealed a decrease in this proportion, with the IPC group experiencing a reduction to 38% (36/94) and the TIC group experiencing a reduction to 49% (26/54). Nevertheless, there was no statistically significant difference observed between the 2 groups in any of the parameters. In both groups, the combined measurements accounted for about 95% of the length of treatment (IPC: median 92.4%; TIC: median 97.2%), thus representing the entirety of the treatment process well (Figure 4A).

Figure 4. Changes in the secondary clinical outcomes between pre-telemedicine and post-telemedicine: (A) insulin-like growth factor 1 (IGF-1; normal range: 64-345 ng/mol), (B) free thyroxine (FT4; normal range: 0.89-1.8 pmol/L), thyroid-stimulating hormone (TSH; normal range: 0.65-6.27 pmol/L), (D) 25-hydroxyvitamin D3 (25-OH VitD3; ≥20 µg/L). pts: patients.

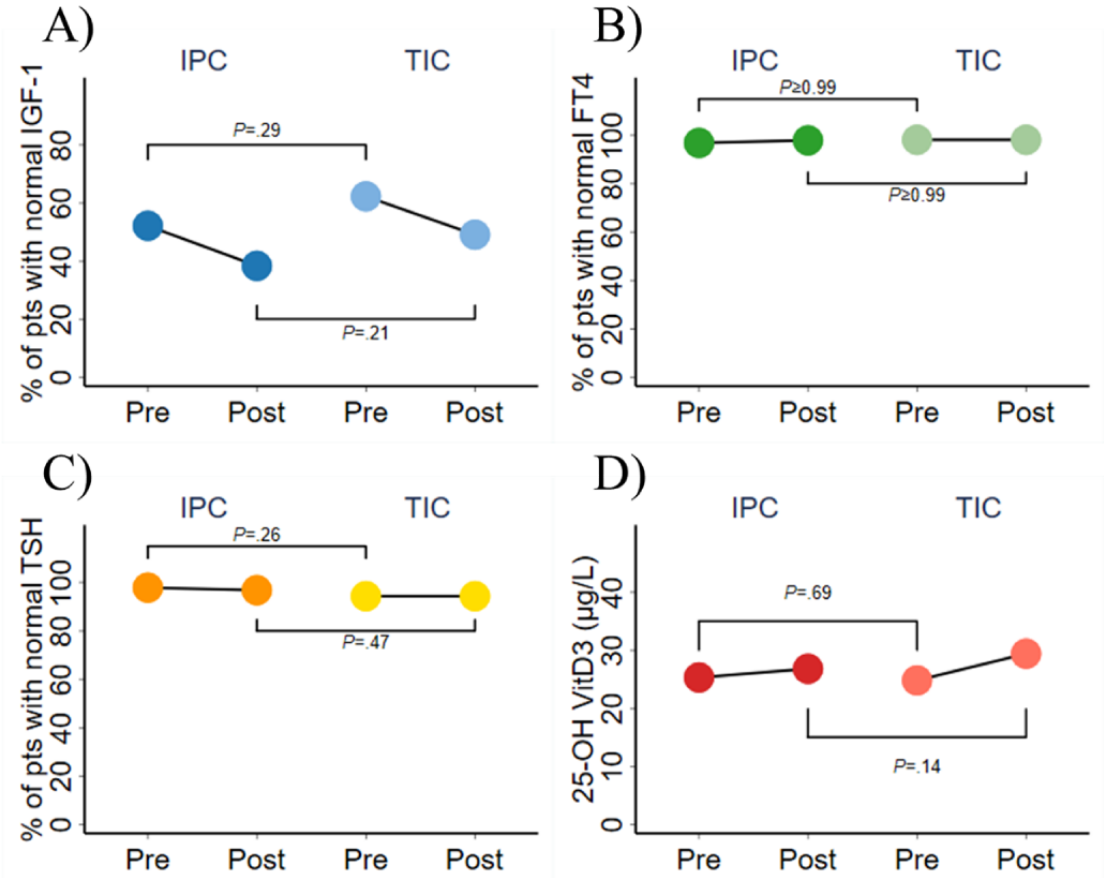

Change in FT4 and TSH

The thyroid gland is a vital endocrine gland responsible for the secretion of thyroid hormones. Thyroid hormones play a crucial role in facilitating growth and development and maintaining skeletal health. A significantly higher number of patients in the TIC group possessed comprehensive records of FT4 and TSH measurements compared with those in the IPC group. Throughout the treatment course, the percentage of patients with FT4 values within the normal range in both groups did not significantly change. Furthermore, there was no statistically significant difference between the 2 groups. Both groups exhibited over 90% of the total duration between the 2 assessments (IPC: median 91.9%; TIC: median 95.5%), effectively capturing the entirety of the treatment course (Figures 4B and 4C).

Change in 25-OH VitD3

Vitamin D has a crucial role in the processes of calcium metabolism and bone growth. The predominant form of vitamin D kept in blood is 25-OH VitD3. By the conclusion of the treatment, both groups exhibited an increase in 25-OH VitD3 levels. The concentration in the IPC group increased from a median 25.30 (IQR 17.20-30.10) µg/L to 26.80 (IQR 22.80-34.40) µg/L, whereas the concentration in the TIC group increased from a median 24.80 (IQR 18.70-31.01) µg/L to 29.40 (IQR 25.30-35.60) µg/L. However, there was no statistically significant difference between the TIC group and IPC group at either time point. In both cohorts, the 2 measurements collectively represented more than 90% of the treatment time (IPC: median 92.7%; TIC: median 96.2%), encompassing a significant portion of the entire treatment duration (Figure 4D).

XSL•FO
RenderX

Health-Seeking Behaviors

Visit Intervals

The median visit interval for the IPC group was 65.20 (IQR 51.50-85.67) days. For the TIC group, the median visit interval was 61.04 (IQR 45.08-76.11) days. There was no significant difference in the visit intervals between the 2 groups (*P*=.17; *U* test; Figure 5).

Figure 5. Bean plots of the distribution of the visit intervals for the in-person care (IPC) group and telemedicine integrated care (TIC) group, with differences assessed using Wilcoxon rank sum tests and the solid black lines representing the median value of the variable being analyzed.

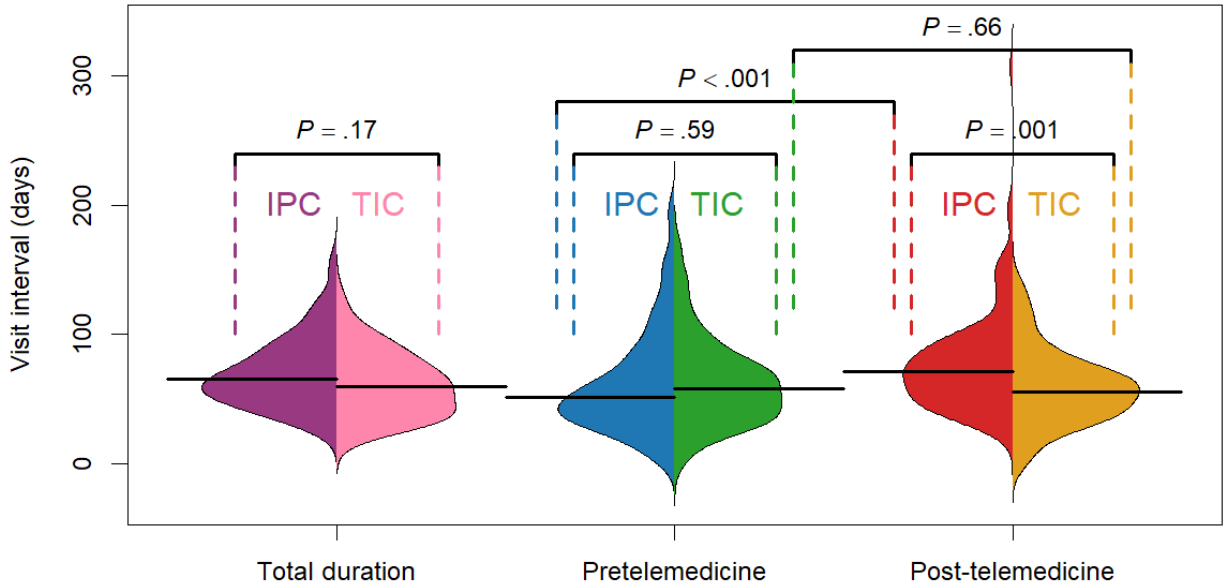

Visit Intervals in the Pretelemedicine Phase

In this phase, the IPC and TIC groups had intervals between visits of 51.25 (IQR 34.75-82.00) days and 58.13 (IQR 36.32-77.45) days, respectively. There was no statistically significant difference between the 2 groups (*P*=.59; *U* test; Figure 5).

Visit Intervals in the Post-Telemedicine Phase

After adopting care integrating telemedicine, the TIC group had average visit intervals of 55.54 (IQR 38.86-70.25) days, which was comparable to that in the preceding phase. Nevertheless, the IPC group had a substantial increase in the median average interval, amounting to 71.08 (IQR 50.75-90.73) days (*P*<.001; *U* test), which was also significantly longer than that for the TIC group (*P*=.001; *U* test; Figure 5).

Cost Analysis

Total Cost Per Visit

The total cost consists of medical costs, labor costs, transportation costs, and lost productivity costs. In the pre-telemedicine phase, the median total cost per visit was ¥3089.99 (IQR 2176.71-5266.23; US $430.35 [IQR 303.16-655.37]) and ¥3412.62 (IQR 2907.62-4705.71; US $475.28 [IQR 404.95-655.37]) in the IPC and TIC groups, respectively. There was no significant disparity in the total cost per visit between the 2 cohorts. In the post-telemedicine phase, the median total costs per visit were ¥3239.73 (IQR 1808.45-6331.82; US $451.20 [IQR 251.87-881.85]) and ¥4169.14 (IQR 1726.21-6042.05; US $580.64 [IQR 240.41-841.49]) in the IPC and TIC groups, respectively. The average total cost per visit was not statistically different between the 2 groups within each phase nor before and after the use of telemedicine within each group (Figure 6A).

XSL•FO
RenderX

Figure 6. Bean plots showing the costs per visit: (A) total cost per visit compared between the in-person care (IPC) and telemedicine integrated care (TIC) groups in the pre- and post-telemedicine phases, (B) medical and labor costs per visit compared between the IPC and TIC groups in the pre- and post-telemedicine phases, (C) medical costs per visit compared between the IPC and TIC groups in the pre- and post-telemedicine phases, (D) total cost per visit for the TIC group compared between the in-person (I) and virtual (V) visits, (E) medical and labor costs per visit compared between the I and V visits, (F) medical costs per visit compared between the I and V visits, with differences assessed using Wilcoxon rank sum tests and the solid black lines representing the median value of the variable being analyzed.

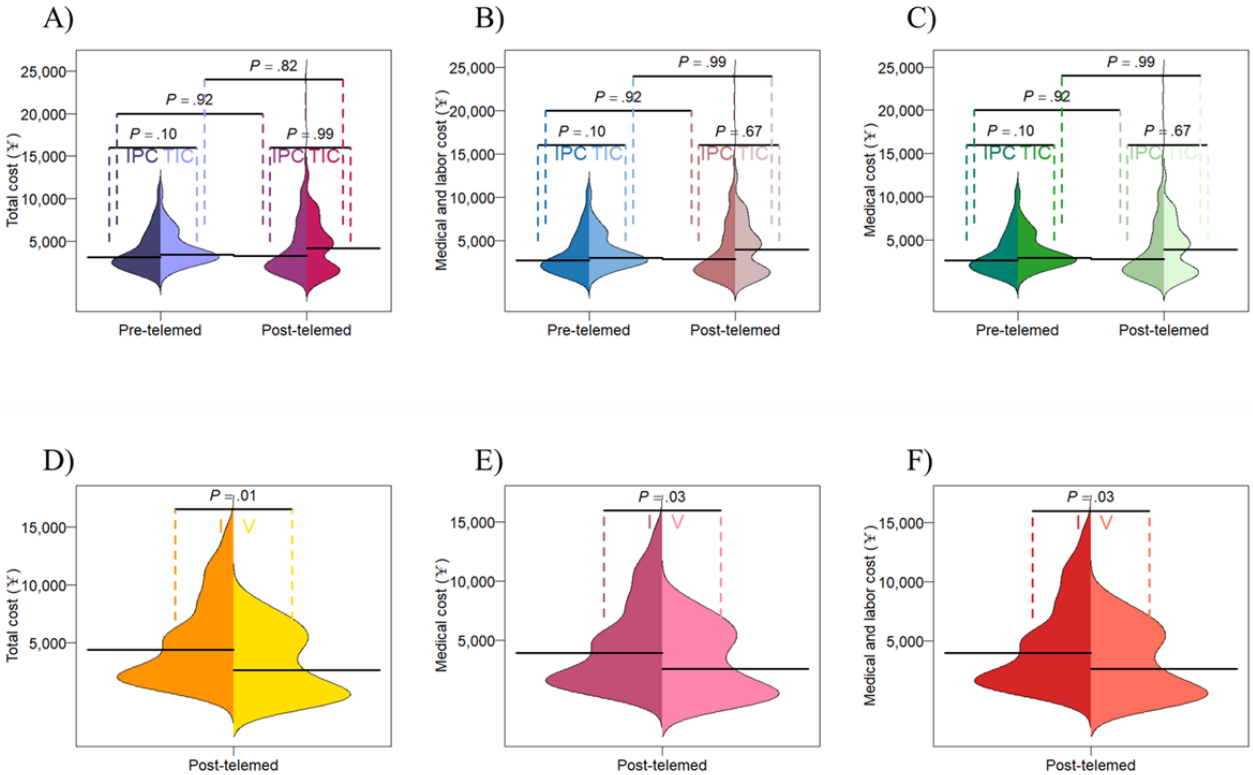

Medical and Labor Costs Per Visit

Although virtual consultations eliminate expenses such as costs of transportation and lost productivity, there are still direct costs to consider, such as medical and labor costs. In the pretelemedicine phase, the medical and labor costs were ¥2692.58 (IQR 1770.30-4859.82; US $375.00 [IQR 246.55-676.84]) for the IPC group and ¥3006.21 (IQR 2501.21-4299.30; US $418.68 [IQR 348.35-598.77]) for the TIC group. During the post-telemedicine phase, the medical and labor costs for the 2 groups amounted to ¥2833.32 (IQR 1402.04-5925.41; US $394.60 [IQR 195.27-825.24]) and ¥3923.60 (IQR 1455.27-5798.21; US $546.45 [IQR 202.68-807.53]), respectively. There were no discernible changes for each group between the pre- and post-telemedicine phases nor any differences between the 2 groups within the same phase (Figure 6B).

Medical Cost Per Visit

Typically, pediatricians allot an average of 7.84 minutes for in-person visits, and they dedicate approximately 10 minutes for each virtual visit, resulting in an average of ¥12.46 (US $1.74) per visit in labor costs due to different treatment modalities. To specifically assess the influence of telemedicine on costs, we performed another analysis focusing solely on medical costs, which included prescriptions, laboratory tests, and registration fees. In the pretelemedicine phase, the medical costs were ¥2647.34 (IQR 1725.06-4814.59; US $368.70 [IQR 240.25-670.54]) for the IPC group and ¥2960.97 (IQR 2455.97-4254.06; US $412.38 [IQR 342.05-592.47]) for the TIC group. During the post-telemedicine phase, the medical costs for the 2 groups were ¥2788.09 (IQR 1356.80-5880.18; US $388.30 [IQR 188.96-818.94]) and ¥3873.43 (IQR 1405.88-5747.99; US $539.46 [IQR 195.80-800.53]), respectively. There were no discernible changes for each group between the pre- and post-telemedicine phases nor any differences between the 2 groups within the same phase (Figure 6C).

Cost Per Visit in the Post-Telemedicine Phase for the TIC Group

To facilitate a more natural comparison of costs across different medical visit channels, we conducted an analysis of various costs incurred by types of visits for the same group of patients at the same phase. Our findings revealed that, in the post-telemedicine phase, the total cost, medical and labor costs, and medical costs associated with the in-person visits were all much greater than these costs generated by virtual visits in the TIC group (Figures 6D, 6E, and 6F).

Discussion

Principal Findings

We conducted a rigorously planned retrospective cohort study to thoroughly examine the disparities in clinical outcomes, health-seeking behaviors, and average expenses per visit between care integrating telemedicine and in-person care only

XSL•FO
RenderX

using short stature as an example. Our analysis revealed that the short stature patients in the TIC group had more significant growth than the IPC group. Meanwhile, this novel approach did not result in any negative impacts on other metrics such as BMI and IGF-1 level. Furthermore, our observations demonstrated that the patients in the TIC group consistently adhered to the same visit frequency throughout the whole treatment cycle, whereas patients in the IPC group prolonged their visits by 38.68% of the time during the post-telemedicine phase. Last, we observed that the average cost per visit in the TIC group was indistinguishable from that of the IPC group, yet the average cost per visit of virtual visits was cheaper than that of the usual in-person visits. By conducting a thorough investigation of 2 service models, we successfully demonstrated the effective implementation of telemedicine in China. This serves as compelling proof for the use of telemedicine in clinical treatment.

Telemedicine is used in domains such as diabetes [5,6] and hypertension [7,8]. These therapeutic domains have 3 attributes in common. First, all necessitate regular and consistent checkups over an extended period [33]. Second, access to care has an impact on the results [34,35]. Third, treatment is subject to a specific timeframe, meaning the timelier the intervention, the more pronounced the impact [36]. The characteristics of short stature in children align with these criteria; however, there is a lack of research examining the potential impact of telemedicine on improving results in this domain. We addressed this deficiency in our research. Given that children in lower-income regions have a higher prevalence of short stature [25], we considered not only the clinical efficacy but also the patients' health-seeking behaviors and financial burden to comprehensively evaluate the feasibility of implementing telemedicine in the perspective of value-based medicine [37].

Our research on health-seeking behaviors has shown a novel finding: There were variations in the visit intervals depending on the different modalities of treatment. This finding strongly supports the notion that telemedicine can improve clinical outcomes by enhancing accessibility. By comparing the visit intervals between the 2 groups in the post-telemedicine phase, we conclude that telemedicine enhanced accessibility. The 2 propensity score–matched cohorts should exhibit the same health-seeking behaviors, yet in the post-telemedicine phase, the visit intervals for patients in the TIC group remained the same as in the previous phase, whereas patients receiving in-person care had prolonged intervals. This indicates that telemedicine is a more efficient means of ensuring regular visits, making it a valuable instrument in situations when medical availability is restricted. For instance, during the pandemic, patients deferred or canceled their scheduled appointments due to concerns regarding coronavirus infection [38,39]. Telemedicine thereby demonstrated its benefits of allowing patients to receive health care service from afar, making it easier to adhere to the treatment schedule [40,41]. By comparing the visit intervals in the pre- and post-telemedicine phases in the TIC group, we deduced that telemedicine did not promote unnecessary medical visits. Since the visit intervals of the TIC group remained the same across the 2 phases, it is believed that the visit frequency stayed consistent. This suggests that the use of telemedicine did not result in a reduction of the interval nor did it result in excessive use of medical resources, as some researchers have expressed concern about [42]. Additionally, the clinical findings illustrate that the growth in height in the TIC group was more significant, while no statistically significant difference was observed in other clinical parameters between the 2 groups. This demonstrates that the efficacy of integrating telemedicine into care was superior and did not yield any unfavorable effects. Regarding the conclusion that telemedicine "triggered" a higher number of visits, our findings align with the research conducted by Bavafa et al [43]. Another conclusion that telemedicine did not reduce the visit interval also aligns with previously reported results [44]. Nevertheless, our work builds upon prior research and aims to comprehensively examine the mechanism of telemedicine effectiveness by integrating clinical efficacy with health-seeking behaviors. In conclusion, we contend that, during the post-telemedicine phase, the rise in medical visits facilitated by integrating telemedicine into care effectively restored the previously constrained number of medical visits to their usual levels. Our results suggest that telemedicine enhances the efficacy of short stature treatment by enhancing accessibility of care.

Our study also examined the costs associated with different modalities in the new application to short stature. We found that the average cost per visit of patients revealed no significant difference between the 2 groups or within the pre- and post-telemedicine phases. The average cost was calculated based on the patient's cohort affiliation, meaning it encompassed the cost associated with both in-person visits and virtual visits. The comparable average cost per visit suggests homogeneity of the 2 groups. Subsequently, we consider virtual sessions as a replacement for in-person visits, rather than as a supplementary method like patient education or a social network support [9-12], in order to delve deeper into the correlation between visit modalities and costs. The analysis revealed that the average total cost per visit was lower for virtual visits. Telemedicine has the potential to reduce the total cost per visit, as this novel care mode omits transportation and lost productivity costs [45-47], leaving only medical and labor costs. When comparing these 2 cost categories, our research showed that costs associated with virtual visits remained lower. This is consistent with previous research [48]. Our study also found that virtual visits are associated with lower medical costs. A possible explanation is that the frequency of prescribing laboratory tests was lower during telemedicine visits than during in-person visits [49]. Considering the benefits of telemedicine in terms of accessibility and affordability, virtual care will be a financially efficient method of delivering health care services in remote and impoverished regions with inadequate medical infrastructure, as well as in situations with restricted resources due to wars.

Limitations

In this study, we showcased successful telemedicine implementation in China. By comprehensively examining 2 service models in the 2 cohorts, our research not only holds significant clinical implications but also provides strong evidence for the cost-effectiveness of telemedicine. However, it is important to acknowledge that our research possesses specific constraints. One constraint pertains to the use of a

XSL•FO
RenderX

retrospective methodology, which is an observational approach that seeks to examine the association between occurrences using historical data. Its inherent nature makes it difficult to establish causal relationships [50]. To address the first problem, we used the PSM approach. PSM offers a solution to address selection bias by effectively managing confounding variables. Given the restricted sample size and data availability, our study used 7 variables pertaining to the patients' characteristics, health-seeking behaviors, and costs in the PSM. Additional factors that may also contribute to the change in visit modality, such as socioeconomic status, digital literacy, and the willingness to adopt new technologies, could not be included. Nonetheless, PSM remains a significant means for probing the mechanisms through which telemedicine improves outcomes. In addition, the treatment course is a natural sequential process. Some patients opt to switch to approaches integrating telemedicine after in-person treatment. The evidence maximizes the strength of a causal relationship between telemedicine use and improved treatment outcomes as well as changes in health-seeking behaviors.

Another issue is that some data could not be collected since the patients had finished their treatment course prior to our study. As a result, there were 2 repercussions. The first was our inability to evaluate the factors contributing to patient conversion from virtual to in-person appointments. A previous study found that 2% of individuals went back to physical visits because of language barriers [51]. Although we lacked the necessary data for a comparative analysis, we may deduce that situations in which virtual and in-person visits occur on the same day may involve certain needs that cannot be effectively addressed remotely. Only 1 patient had an in-person visit 1 hour and 22 minutes after having a virtual visit. Although the patient conversion rate we observed is comparable to that in prior research, additional research is necessary to elucidate patients' inclinations to modify their health-seeking behaviors. The second repercussion was that we had to estimate transport costs and lost productivity costs using the patients' demographic information and previously published work [31].

Last, we could not measure patient satisfaction data in this work. For in-person visits, given that the satisfaction was collected anonymously, it was challenging to correlate satisfaction ratings with the patients included in this study. For virtual visits, despite the fact that patients had to reveal their real names, the rate of satisfaction reports was merely 3.8%. With only 107 virtual visits, the TIC group only had 4 reviews, an inadequate sample size for further research. Although this issue may impact optimization of the 2 service models in the future, it does not impact the conclusion of this paper.

Conclusions

Our study reveals a possible mechanism for how care integrating telemedicine could lead to improved clinical outcomes compared with in-person care. Telemedicine serves as a viable alternative in situations where access is restricted, without causing abuse of medical resources in terms of the visit frequency and average cost. Our research underscores the transformative potential of telemedicine integrated into care for addressing health care challenges, exemplified by the enhanced efficacy observed in managing short stature. As we forge ahead, the intersection of big data, large language models, and digital medicine holds immense promise for shaping the future landscape of health care delivery. The integration of these cutting-edge technologies offers exciting avenues for personalized, efficient, and accessible health care solutions, paving the way for novel research directions that can revolutionize patient outcomes and propel the field toward unprecedented advancements.

Acknowledgments

We thank Dr. Tracy Xiao Liu, Shu Wang, and Dr. Jinjin Ma for their helpful discussion on data analysis. This work was supported by the National Natural Science Foundation (72204005, 72342032, 82103865, 82373593) and Capital Foundation of Medical Development (2022-2G-4099).

Data Availability

The data sets used in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Y Wang and PZ are co-first authors of this manuscript and contributed equally to this research. Y Wang, YX, and YC planned and designed the study. PZ, HS, CZ, and HJ wrote the code and collected the data. Y Wang, PZ, and HS analyzed the data. Y Wang, PZ, and YD wrote the manuscript. Y Wei, KZ, XW, XX, and JC contributed the administrative and technical support of the study. YD and JC are co-corresponding authors of this research. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[DOCX File , 33 KB - jmir_v26i1e57814_app1.docx]

XSL•FO
RenderX

References

1. WHO Group Consultation on Health Telematics. A Health Telematics Policy. World Health Organization. 1997 Dec. URL: https://apps.who.int/iris/bitstream/handle/10665/63857/WHO_DGO_98.1.pdf [accessed 2024-11-09]
2. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
3. Lovejoy JF, Read M. Telemedicine in Orthopedic Surgery and Sports Medicine. Cham, Switzerland: Springer Nature; 2020.
4. STARPAHC systems report. Volume 2: Operational performance. NTRS - NASA Technical Reports Server. 2013 Sep 3. URL: https://ntrs.nasa.gov/citations/19780006834 [accessed 2024-11-09]
5. Timpel P, Oswald S, Schwarz PEH, Harst L. Mapping the evidence on the effectiveness of telemedicine interventions in diabetes, dyslipidemia, and hypertension: an umbrella review of systematic reviews and meta-analyses. J Med Internet Res 2020 Mar 18;22(3):e16791 [FREE Full text] [doi: 10.2196/16791] [Medline: 32186516]
6. Friedman RSC, Carpenter DM, Shaver JM, McDermott SC, Voelkel J. Telemedicine familiarity and post-disaster utilization of emergency and hospital services for ambulatory care sensitive conditions. Am J Prev Med 2022 Jul;63(1):e1-e9 [FREE Full text] [doi: 10.1016/j.amepre.2021.12.025] [Medline: 35300889]
7. Zhou H, Wang X, Yang Y, Chen Z, Zhang L, Zheng C, et al. Effect of a multicomponent intervention delivered on a web-based platform on hypertension control: a cluster randomized clinical trial. JAMA Netw Open 2022 Dec 01;5(12):e2245439 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.45439] [Medline: 36477479]
8. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR Mhealth Uhealth 2018 Jul 23;6(7):e10723 [FREE Full text] [doi: 10.2196/10723] [Medline: 30037787]
9. Chen F, Su W, Becker SH, Payne M, Castro Sweet CM, Peters AL, et al. Clinical and economic impact of a digital, remotely-delivered intensive behavioral counseling program on Medicare beneficiaries at risk for diabetes and cardiovascular disease. PLoS One 2016;11(10):e0163627 [FREE Full text] [doi: 10.1371/journal.pone.0163627] [Medline: 27706216]
10. Giani E, Laffel L. Opportunities and challenges of telemedicine: observations from the wild west in pediatric type 1 diabetes. Diabetes Technol Ther 2016 Jan;18(1):1-3 [FREE Full text] [doi: 10.1089/dia.2015.0360] [Medline: 26756102]
11. Gurupur VP, Miao Z. A brief analysis of challenges in implementing telehealth in a rural setting. Mhealth 2022;8:17 [FREE Full text] [doi: 10.21037/mhealth-21-38] [Medline: 35449506]
12. Cunningham SG, Stoddart A, Wild SH, Conway NJ, Gray AM, Wake DJ. Cost-utility of an online education platform and diabetes personal health record: analysis over ten years. J Diabetes Sci Technol 2023 May;17(3):715-726 [FREE Full text] [doi: 10.1177/19322968211069172] [Medline: 34986658]
13. Chua V, Koh JH, Koh CHG, Tyagi S. The willingness to pay for telemedicine among patients with chronic diseases: systematic review. J Med Internet Res 2022 Apr 13;24(4):e33372 [FREE Full text] [doi: 10.2196/33372] [Medline: 35416779]
14. Kim BY, Lee J. Smart devices for older adults managing chronic disease: a scoping review. JMIR Mhealth Uhealth 2017 May 23;5(5):e69 [FREE Full text] [doi: 10.2196/mhealth.7141] [Medline: 28536089]
15. Strubbia C, Levack WMM, Grainger R, Takahashi K, Tomori K. Use of technology in supporting goal setting in rehabilitation for adults: a scoping review. BMJ Open 2020 Nov 10;10(11):e041730 [FREE Full text] [doi: 10.1136/bmjopen-2020-041730] [Medline: 33172946]
16. Zhou X, Siegel KR, Ng BP, Jawanda S, Proia KK, Zhang X, et al. Cost-effectiveness of diabetes prevention interventions targeting high-risk individuals and whole populations: a systematic review. Diabetes Care 2020 Jul;43(7):1593-1616. [doi: 10.2337/dci20-0018] [Medline: 33534726]
17. Arrieta A, Woods JR, Qiao N, Jay SJ. Cost-benefit analysis of home blood pressure monitoring in hypertension diagnosis and treatment: an insurer perspective. Hypertension 2014 Oct;64(4):891-896. [doi: 10.1161/HYPERTENSIONAHA.114.03780] [Medline: 25024284]
18. Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Public Health 2021 Apr 01;42:463-481 [FREE Full text] [doi: 10.1146/annurev-publhealth-090519-093711] [Medline: 33798406]
19. Delana K, Deo S, Ramdas K, Subburaman GB, Ravilla T. Multichannel delivery in healthcare: the impact of telemedicine centers in southern India. Management Science 2023 May;69(5):2568-2586. [doi: 10.1287/mnsc.2022.4488]
20. Hossain M, Dean EB, Kaliski D. Using administrative data to examine telemedicine usage among Medicaid beneficiaries during the coronavirus disease 2019 pandemic. Med Care 2022 Jul 01;60(7):488-495 [FREE Full text] [doi: 10.1097/MLR.0000000000001723] [Medline: 35679172]
21. McAlister FA. Virtual care and emergency department use. CMAJ 2023 Mar 14;195(10):E376 [FREE Full text] [doi: 10.1503/cmaj.148052-l] [Medline: 36918179]
22. Pendrith C, Nayyar D, Chu C, O'Brien T, Lyons OD, Agarwal P, et al. Outpatient visit trends for internal medicine ambulatory care sensitive conditions after the COVID-19 pandemic: a time-series analysis. BMC Health Serv Res 2022 Feb 14;22(1):198 [FREE Full text] [doi: 10.1186/s12913-022-07566-6] [Medline: 35164751]
23. Rani D, Shrestha R, Kanchan T, Krishan K. Short Stature. Treasure Island, FL: StatPearls Publishing; 2023.

XSL•FO
RenderX

24. Pagani S, Radetti G, Meazza C, Bozzola M. Analysis of growth hormone receptor gene expression in tall and short stature children. J Pediatr Endocrinol Metab 2017 Apr 01;30(4):427-430. [doi: 10.1515/jpem-2016-0355] [Medline: 28301320]
25. Wang Q, Liu D, Yang L, Liu Y, Chen X. The epidemic characteristics of short stature in school students. Ital J Pediatr 2015 Dec 29;41(1):99 [FREE Full text] [doi: 10.1186/s13052-015-0207-6] [Medline: 26714873]
26. Tornincasa V, Dixon D, Le Masne Q, Martin B, Arnaud L, van Dommelen P, et al. Integrated digital health solutions in the management of growth disorders in pediatric patients receiving growth hormone therapy: a retrospective analysis. Front Endocrinol (Lausanne) 2022;13:882192 [FREE Full text] [doi: 10.3389/fendo.2022.882192] [Medline: 35846336]
27. Mavinkurve M, Azriyanti AZ, Jalaludin MY. The short child: Importance of early detection and timely referral. Malays Fam Physician 2021 Nov 30;16(3):6-15 [FREE Full text] [doi: 10.51866/rv1157] [Medline: 34938388]
28. Corripio-Collado R, Fernández-Ramos C, González-Casado I, Moreno-Macián F, López-Siguero JP, Labarta-Aizpún JI. Delphi consensus on the diagnosis and treatment of patients with short stature in Spain: GROW-SENS study. J Endocrinol Invest 2022 Apr;45(4):887-897 [FREE Full text] [doi: 10.1007/s40618-021-01696-0] [Medline: 34791604]
29. Berry CA, Kwok L, Massar R, Chang JE, Lindenfeld Z, Shelley DR, et al. Patients' perspectives on the shift to telemedicine in primary and behavioral health care during the COVID-19 pandemic. J Gen Intern Med 2022 Dec 27;37(16):4248-4256 [FREE Full text] [doi: 10.1007/s11606-022-07827-4] [Medline: 36167954]
30. Desruisseaux M, Stamenova V, Bhatia RS, Bhattacharyya O. Channel management in virtual care. NPJ Digit Med 2020;3:44 [FREE Full text] [doi: 10.1038/s41746-020-0252-4] [Medline: 32219185]
31. Jiao Y, Qin M, Hu B, Pan H, Sun G, Chen S, et al. Cost analysis of launching telemedicine services by Peking Union Medical College Hospital during the pandemic of coronavirus disease 2019. Medical Journal of Peking Union Medical College Hospital 2021;12(1):49-53. [doi: 10.12290/xhyxzz.20200168]
32. Beijing Statistical Yearbook 2020. Beijing Municipal Bureau of Statistics. URL: https://nj.tjj.beijing.gov.cn/nj/main/2020-tjnj/zk/indexch.htm [accessed 2023-11-07]
33. Kotiso KS, Degemu N, Gebremedhin S, Taye M, Petros A, Belayneh F, et al. Determinants of hypertension among patients with type 2 diabetes mellitus on follow-up at Tikur Anbessa Specialized Hospital, Addis Ababa: A case-control study. PLoS One 2021;16(8):e0256399 [FREE Full text] [doi: 10.1371/journal.pone.0256399] [Medline: 34424924]
34. Kuehn BM. Hypertension rates in rural areas outpace those in urban locales. JAMA 2020 Jun 23;323(24):2454. [doi: 10.1001/jama.2020.9382] [Medline: 32573675]
35. Mercado CI, McKeever Bullard K, Gregg EW, Ali MK, Saydah SH, Imperatore G. Differences in U.S. rural-urban trends in diabetes ABCS, 1999-2018. Diabetes Care 2021 Aug;44(8):1766-1773 [FREE Full text] [doi: 10.2337/dc20-0097] [Medline: 34127495]
36. de Boer IH, Bangalore S, Benetos A, Davis AM, Michos ED, Muntner P, et al. Diabetes and hypertension: a position statement by the American Diabetes Association. Diabetes Care 2017 Sep;40(9):1273-1284. [doi: 10.2337/dci17-0026] [Medline: 28830958]
37. Teisberg E, Wallace S, O'Hara S. Defining and implementing value-based health care: a strategic framework. Acad Med 2020 May;95(5):682-685 [FREE Full text] [doi: 10.1097/ACM.0000000000003122] [Medline: 31833857]
38. Anderson KE, McGinty EE, Presskreischer R, Barry CL. Reports of forgone medical care among US adults during the initial phase of the COVID-19 pandemic. JAMA Netw Open 2021 Jan 04;4(1):e2034882 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.34882] [Medline: 33475757]
39. Lazzerini M, Barbi E, Apicella A, Marchetti F, Cardinale F, Trobia G. Delayed access or provision of care in Italy resulting from fear of COVID-19. Lancet Child Adolesc Health 2020 May;4(5):e10-e11 [FREE Full text] [doi: 10.1016/S2352-4642(20)30108-5] [Medline: 32278365]
40. Chu C, Nayyar D, Bhattacharyya O, Martin D, Agarwal P, Mukerji G. Patient and provider experiences with virtual care in a large, ambulatory care hospital in Ontario, Canada during the COVID-19 pandemic: observational study. J Med Internet Res 2022 Oct 25;24(10):e38604 [FREE Full text] [doi: 10.2196/38604] [Medline: 36194862]
41. Morris NP. Staying apart during a pandemic. JAMA Intern Med 2020 Aug 01;180(8):1047-1048. [doi: 10.1001/jamainternmed.2020.2505] [Medline: 32597924]
42. Dixit RA, Ratwani RM, Bishop JA, Schulman K, Sharp C, Palakanis K, et al. The impact of expanded telehealth availability on primary care utilization. NPJ Digit Med 2022 Sep 09;5(1):141 [FREE Full text] [doi: 10.1038/s41746-022-00685-8] [Medline: 36085158]
43. Bavafa H, Hitt LM, Terwiesch C. The impact of e-visits on visit frequencies and patient health: evidence from primary care. Manage Sci 2018 Dec;64(12):5461-5480 [FREE Full text] [doi: 10.1287/mnsc.2017.2900] [Medline: 33033417]
44. Cao Y, Chen D, Smith M. Use telehealth as needed: telehealth substitutes in-person primary care and associates with the changes in unplanned events and follow-up visits. BMC Health Serv Res 2023 May 03;23(1):426 [FREE Full text] [doi: 10.1186/s12913-023-09445-0] [Medline: 37138327]
45. Snoswell CL, Taylor ML, Comans TA, Smith AC, Gray LC, Caffery LJ. Determining if telehealth can reduce health system costs: scoping review. J Med Internet Res 2020 Oct 19;22(10):e17298 [FREE Full text] [doi: 10.2196/17298] [Medline: 33074157]

XSL•FO
RenderX

46. Haider Z, Aweid B, Subramanian P, Iranpour F. Telemedicine in orthopaedics during COVID-19 and beyond: A systematic review. J Telemed Telecare 2022 Jul 06;28(6):391-403 [FREE Full text] [doi: 10.1177/1357633X20938241] [Medline: 32762270]
47. Sinha N, Cornell M, Wheatley B, Munley N, Seeley M. Looking through a different lens: patient satisfaction with telemedicine in delivering pediatric fracture care. J Am Acad Orthop Surg Glob Res Rev 2019 Sep;3(9):e100 [FREE Full text] [doi: 10.5435/JAAOSGlobal-D-19-00100] [Medline: 31773080]
48. Chaiyachati KH, Snider CK, Mitra N, Huffenberger AM, McGinley S, Bristow R, et al. Economics of a health system's direct-to-consumer telemedicine for its employees. Am J Manag Care 2023 Jun;29(6):284-290 [FREE Full text] [doi: 10.37765/ajmc.2023.89369] [Medline: 37341975]
49. Alexander GC, Tajanlangit M, Heyward J, Mansour O, Qato DM, Stafford RS. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. JAMA Netw Open 2020 Oct 01;3(10):e2021476 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.21476] [Medline: 33006622]
50. Wang X, Kattan MW. Cohort studies: design, analysis, and reporting. Chest 2020 Jul;158(1S):S72-S78. [doi: 10.1016/j.chest.2020.03.014] [Medline: 32658655]
51. Cerchia E, Della Corte M, Fiori C, Mandaletti M, Ruggiero E, Catti M, et al. The digital era of pediatric urological clinical care: telemedicine for management of lower urinary tract symptoms and nocturnal enuresis - a comprehensive tertiary center insight. Minerva Urol Nephrol 2024 Jun;76(3):1. [doi: 10.23736/s2724-6051.24.05846-4]

Abbreviations

25-OH VitD3: 25-hydroxyvitamin D3
EMR: electronic medical record
FT4: free thyroxine
IGF-1: insulin-like growth factor 1
IPC: in-person care
PSM: propensity score matching
TIC: telemedicine integrated care
TSH: thyroid-stimulating hormone
WHO: World Health Organization

Edited by A Coristine; submitted 27.02.24; peer-reviewed by C Ma, M Della Corte; comments to author 01.08.24; revised version received 09.08.24; accepted 27.10.24; published 19.11.24.

Please cite as:
Wang Y, Zhang P, Xing Y, Shi H, Cui Y, Wei Y, Zhang K, Wu X, Ji H, Xu X, Dong Y, Jin C
Telemedicine Integrated Care Versus In-Person Care Mode for Patients With Short Stature: Comprehensive Comparison of a Retrospective Cohort Study
J Med Internet Res 2024;26:e57814
URL: https://www.jmir.org/2024/1/e57814
doi:10.2196/57814
PMID:

©Yipei Wang, Pei Zhang, Yan Xing, Huifeng Shi, Yunpu Cui, Yuan Wei, Ke Zhang, Xinxia Wu, Hong Ji, Xuedong Xu, Yanhui Dong, Changxiao Jin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quantitative Impact of Traditional Open Surgery and Minimally Invasive Surgery on Patients' First-Night Sleep Status in the Intensive Care Unit: Prospective Cohort Study

Chen Shang[1*], MD; Ya Yang[2*], PhD; Chengcheng He[1], BSCN; Junqi Feng[1], MM; Yan Li[1], BSCN; Meimei Tian[1], PhD; Zhanqi Zhao[3,4,5], PhD; Yuan Gao[1], MD, PhD; Zhe Li[1], MD, PhD

[1]Department of Critical Care Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Infection Control, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
[3]School of Biomedical Engineering, Guangzhou Medical University, Guangzhou, China
[4]Department of Critical Care Medicine, Peking Union Medical College Hospital, Beijing, China
[5]Institute of Technical Medicine, Furtwangen University, Villingen-Schwenningen, Germany
[*]these authors contributed equally

Corresponding Author:
Zhe Li, MD, PhD
Department of Critical Care Medicine
Renji Hospital
School of Medicine, Shanghai Jiao Tong University
No. 160, Pujian Road
Pudong New District
Shanghai, 200127
China
Phone: 86 68383162
Fax: 86 68383162
Email: slamy1987@126.com

Abstract

Background: The sleep status of patients in the surgical intensive care unit (ICU) significantly impacts their recoveries. However, the effects of surgical procedures on sleep are rarely studied.

Objective: This study aimed to investigate quantitatively the impact of traditional open surgery (TOS) versus minimally invasive surgery (MIS) on patients' first-night sleep status in a surgical ICU.

Methods: Patients transferred to the ICU after surgery were prospectively screened. The sleep status on the night of surgery was assessed by the patient- and nurse-completed Richards-Campbell Sleep Questionnaire (RCSQ) and Huawei wearable sleep monitoring wristband. Surgical types and sleep parameters were analyzed.

Results: A total of 61 patients were enrolled. Compared to patients in the TOS group, patients in the MIS group had a higher nurse-RCSQ score (mean 60.9, SD 16.9 vs mean 51.2, SD 17.3; P=.03), self-RCSQ score (mean 58.6, SD 16.2 vs mean 49.5, SD 14.8; P=.03), and Huawei sleep score (mean 77.9, SD 4.5 vs mean 68.6, SD 11.1; P<.001). Quantitative sleep analysis of Huawei wearable data showed a longer total sleep period (mean 503.0, SD 91.4 vs mean 437.9, SD 144.0 min; P=.04), longer rapid eye movement sleep period (mean 81.0, 52.1 vs mean 55.8, SD 44.5 min; P=.047), and higher deep sleep continuity score (mean 56.4, SD 7.0 vs mean 47.5, SD 12.1; P=.001) in the MIS group.

Conclusions: MIS, compared to TOS, contributed to higher sleep quality for patients in the ICU after surgery.

(J Med Internet Res 2024;26:e56777) doi:10.2196/56777

KEYWORDS

sleep quality; wearable sleep monitoring wristband; intensive care unit; minimally invasive surgery; traditional open surgery

XSL•FO
RenderX

Introduction

Impaired sleep quality is a common and distressing problem among surgical patients, particularly among those admitted to the intensive care unit (ICU) [1,2]. Sleep disruption can lead to delirium, compromised immune function, and extended wound recovery, ultimately resulting in adverse outcomes [2].

Several surgical factors can impair sleep quality, including the magnitude of the surgical intervention, physical pain, stress related to the operation, and concerns about pathology results [1,3]. For instance, patients who underwent major abdominal surgery may lose up to 80% of their total sleep time during the first few postoperative nights [4]. Minimally invasive surgery (MIS), which involves smaller incisions and fewer complications, has recently become a viable alternative to traditional open surgery (TOS), as recommended by the Enhanced Recovery After Surgery guidelines [5,6]. However, the impact of MIS on sleep status remains poorly understood.

The accurate measurement of sleep in ICU settings is hindered by various technical obstacles. While polysomnography (PSG) is widely regarded as the most reliable method for assessing sleep, its utilization is constrained by several factors. First, PSG requires special equipment, as well as complex operational procedures, which can be challenging to implement in an ICU setting. Additionally, artifacts caused by other monitoring devices and medical interventions can interfere with PSG recordings, making it difficult to obtain accurate sleep data. Furthermore, the interpretation of PSG results requires skilled personnel, which may not always be readily available in an ICU environment [7,8]. Lastly, PSG is often uncomfortable for patients and can even exacerbate sleep disturbances [9]. Consequently, it is not surprising that questionnaires, such as the Richards-Campbell Sleep Questionnaire (RCSQ), are frequently used for sleep assessment [8]. However, the accuracy and reliability of these questionnaires are undermined by a range of factors. Patients in the ICU may have cognitive, communicational, and physical limitations that hinder their ability to accurately respond to the questions. On the other hand, nurses may face difficulties in accurately assessing patients' sleep-wake states due to their workload or other reasons [3,10]. In practice, there is a pressing clinical need for a convenient and objective technique to enhance the validity of bedside sleep assessment.

The field of sleep monitoring has witnessed rapid advancements in consumer technology, with wearable devices becoming increasingly popular among the general public. Previous studies have demonstrated the promising validity of wearable wristbands in identifying patients' sleep patterns compared to PSG [11]. Moreover, it is conceivable that sleep monitoring wristbands could serve as potential tools for evaluating sleep in the ICU [12]. Empirical wristbands, equipped with accelerometers (actigraphy), offer a noninvasive and affordable solution for continuous sleep monitoring [13]. However, there is a risk that wristbands may overestimate total sleep time in patients who are sedated [13]. Modern wearable sleep monitoring devices integrate cardiopulmonary coupling technique and analysis software, enabling more accurate identification of sleep stages, including sleep depth, rapid eye movement (REM) sleep, deep sleep, and light sleep [14]. This offers the opportunity to assess postoperative sleep patterns in patients after they wake up from surgery.

This study investigates postoperative ICU patients' sleep patterns on the surgical night using wearable wristbands and RCSQs to quantify the impact of TOS and MIS procedures on sleep status. The findings will provide an objective basis for tailoring sleep management strategies in the surgical ICU.

Methods

Study Population

This prospective cohort study was conducted in an 18-bed surgery ICU at Renji Hospital, Shanghai, China. Patients aged 18 years or older who underwent gastrointestinal, biliary and pancreatic, urology, or orthopedic surgery and were transferred to the ICU from June 2022 to September 2022 were screened. Patients with 1 or more of the following conditions were excluded: neurological or mental disorders, dementia, chronic sleep disorders, failure to achieve postanesthesia resuscitation and to be extubated before 6:00 PM, use of remifentanil exceeding 0.03 μg/kg/min, hemodynamic instability (defined as an irreversible systolic blood pressure below 90 mm Hg), or at a high risk of reoperation during the study period as determined by the attending physician.

Ethical Considerations

This study was approved by the Ethics Committee of Renji Hospital (KY2022-145-A). All data have been anonymized to protect patient privacy, and participants benefit from long-term follow-up care as compensation. Written informed consent was obtained from patients or their legal representatives.

Sleep Monitoring

The sleep monitoring period was from 9:00 PM on the night of the surgery to 7:00 AM on the following day. A wearable wristband (HONOR BAND 6, Huawei Technologies Co., Ltd) was used to collect sleep data. At 7:00 AM the next day, the bedside nurse removed the wristband and transferred the encrypted sleep data to a customized application. The nurse-RCSQ was then completed by the bedside nurse, while a nonbedside nurse interviewed the patient and filled out the self-RCSQ.

Data Collection

Patient demographic data, medication history, operation information, ICU treatments, and sleep parameters were collected. Both patient- and nurse-reported RCSQ scores were rated from 0 (Bad) to 100 (Good) on a visual analog scale [15] (Multimedia Appendix 1). Sleep data from wearable wristbands were automatically translated into sleep parameters, including Huawei sleep score, light sleep, deep sleep, REM, and wakefulness, with customized, built-in software that translates cardiopulmonary coupling and accelerometry data into sleep-wake periods [14].

XSL•FO
RenderX

Study Group

Patients were divided into two study groups based on the surgical approach: (1) MIS group—patients who underwent laparoscopic or robotic-assisted techniques via small skin incisions [5] and (2) TOS group—patients who underwent traditional, open surgical techniques.

Outcome

The primary outcome was the sleep scores of patients who underwent MIS or TOS. The secondary outcomes included the consistency of sleep status assessed by wearable wristbands and RCSQs, as well as quantitative sleep parameters from the wristbands.

Statistical Analysis

Based on preliminary data, and based on a .05 α level and 80% power, we calculated a sample size of 25 per group to observe an average increase of 10 points in Huawei and RCSQ sleep scores in patients who underwent MIS compared to patients who underwent TOS [16]. All values are presented as mean (SD), median (IQR), number, and percentage (%) as appropriate. Continuous data were analyzed using 2-tailed *t* tests or the Mann-Whitney *U* test, as appropriate. Bland-Altman concordance analysis was performed to assess the agreement between the 2 assessments. Multivariable linear regression was conducted to estimate the effects of surgical mode on sleep-related indexes. A 2-sided *P* value <.05 was considered statistically significant. Anthropometric data and measurements were analyzed using R software (version 4.10; R Foundation for Statistical Computing).

Results

Characteristics

A total of 61 patients received sleep assessments; 28 (46%) patients underwent MIS, while 33 (54%) patients underwent TOS. Table 1 presents the demographic and perioperative characteristics between the MIS and TOS groups. The MIS group had a lower proportion of upper abdominal surgeries and a higher proportion of lower abdominal and pelvic surgeries compared to the TOS group.

Table 1. Demographic and perioperative characteristics of enrolled patients.

Variables	MIS[a] (n=28)	TOS[b] (n=33)
Age (year), mean (SD)	65.1 (11.2)	69.7 (8.0)
Male sex, n (%)	17 (61)	21 (64)
APACHE[c] score, median (IQR)	15.5 (15.0-18.0)	15 (13.8-16.0)
Sedation and analgesia[d], n (%)	10 (36)	12 (36)
Drainage tube number, median (IQR)	7 (6-8)	7 (6-8)
Medical interventions[e], median (IQR)	10.5 (9.0-12.0)	11.0 (10.0-12.0)
Surgical site, n (%)		
Upper abdomen	10 (36)	24 (73)
Lower abdomen and pelvis	12 (43)	2 (6)
Orthopedic	6 (21)	7 (21)

[a]MIS: minimally invasive surgery.

[b]TOS: traditional open surgery.

[c]APACHE: Acute Physiology and Chronic Health Evaluation.

[d]Sedation and analgesia: use of dexmedetomidine and remifentanil during the postoperative period.

[e]Medical interventions: noninvasive blood pressure, blood drawing and injection, and other medical procedures.

Comparison of Postoperative Sleep Scores

An average increase of 10 points was observed in the nurse-RCSQ score (mean 60.9, SD 16.9 vs mean 51.2, SD 17.3; *P*=.03), self-RCSQ score (mean 58.6, SD 16.2 vs mean 49.5, SD 14.8; *P*=.03), and Huawei sleep score (mean 77.9, SD 4.5 vs mean 68.6, SD 11.1; *P*<.001) in patients who underwent MIS compared to patients who underwent TOS (Figure 1). After adjusting for baseline information including the surgical site, the effect of surgical mode (MIS or TOS) was significant for the self-RCSQ score (*P*=.03) and Huawei sleep score (*P*<.001), as shown in Table 2. Furthermore, the Bland-Altman analysis demonstrated a high level of consistency, with over 95% agreement between the Huawei sleep score and self-RCSQ (58/61, 95%) and nurse-RCSQ scores (59/61, 97%; Figure 2).

XSL•FO
RenderX

Figure 1. Violin plots displaying the distribution of individual values of the nurse-RCSQ, self-RCSQ, and Huawei sleep scores; differences among groups were assessed by 2-tailed t tests. *P<.05, ***P<.001. MIS: minimally invasive surgery; RCSQ: Richards-Campbell Sleep Questionnaire; TOS: traditional open surgery.

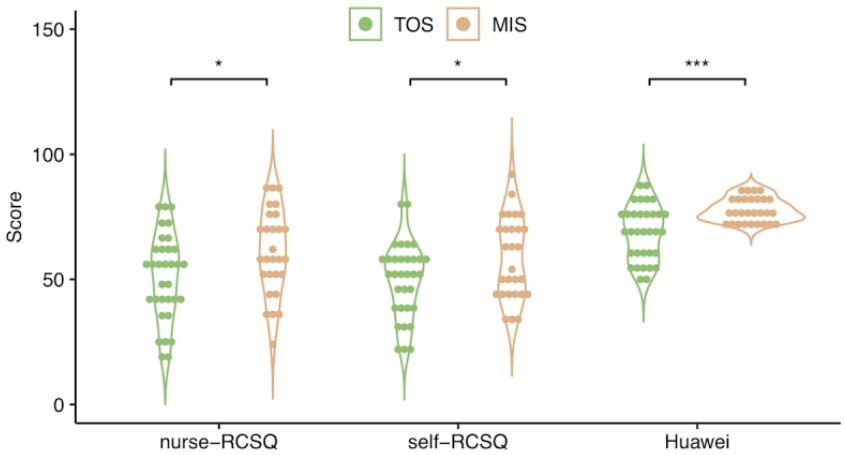

Table 2. Multivariate linear regression analysis of surgical mode on nurse-RCSQ[a], self-RCSQ, and Huawei sleep scores.

Score	Adjusted β[b]	SE	*t* statistic (*df*)	*P* value
Nurse-RCSQ score	10.012	5.416	1.849 (50)	.07
Self-RCSQ score	11.382	4.929	2.309 (50)	.03
Huawei sleep score	10.970	2.740	4.003 (50)	<.001

[a]RCSQ: Richards-Campbell Sleep Questionnaire

[b]Models were adjusted for age, sex, Acute Physiology and Chronic Health Evaluation (APACHE) score, use of sedation and analgesia, number of drainage tubes and medical interventions, and surgical site.

Figure 2. Bland-Altman analysis between Huawei and RCSQ sleep scores. Horizontal line are drawn at the mean difference and at the mean difference plus and minus 1.96 SD of the differences, with the x-axis reporting average levels of sleep scores. (A) Huawei sleep monitoring wristband versus self-RCSQ assessment of total sleep score, with 95% (58/61) of the points within the clinically acceptable consistency boundary. (B) Huawei sleep monitoring wristband versus nurse-RCSQ assessment of total sleep score, with 97% (59/61) of the points within the clinically acceptable consistency boundary. RCSQ: Richards-Campbell Sleep Questionnaire.

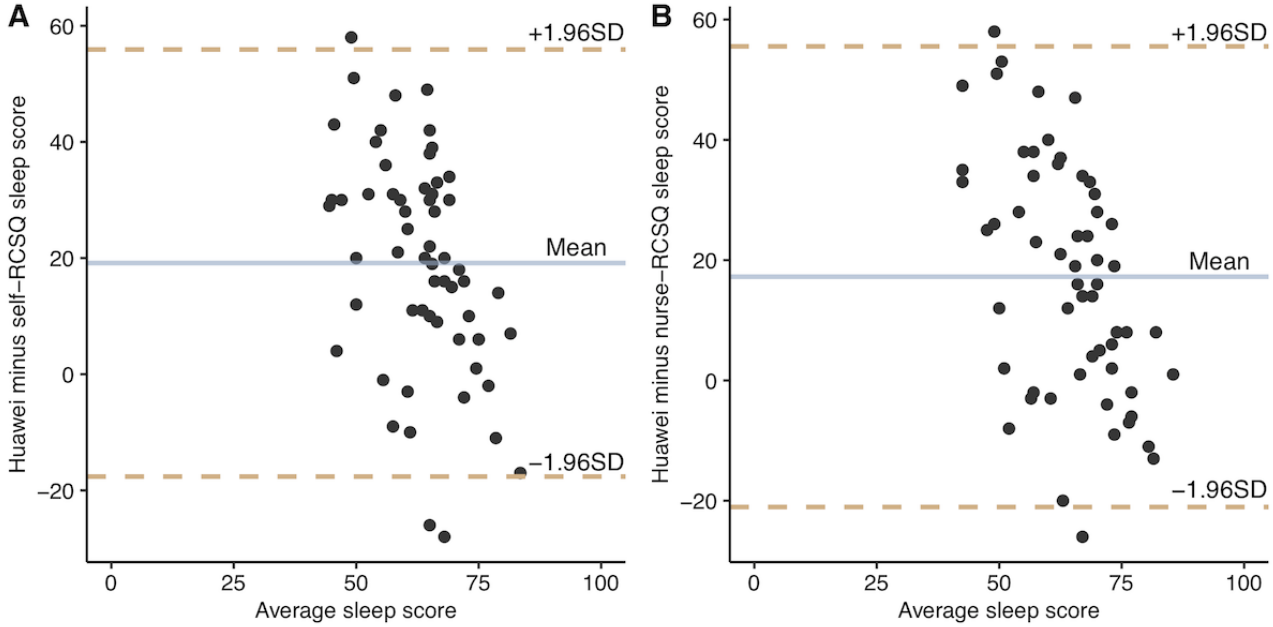

XSL•FO
RenderX

Quantitative Impact of Surgical Mode on Sleep Status

Quantitative analysis on the duration and depth of sleep using Huawei software is presented in Table 3. The MIS group had a 1-hour longer total sleep time (mean 503.0, SD 91.4 vs mean 437.9, SD 144.0 min; *P*=.04) and a 25-minute longer REM sleep time (mean 81.0, SD 52.1 vs mean 55.8, SD 44.5 min; *P*=.047) compared to the TOS group. Similarly, patients who underwent MIS had a higher deep sleep continuity score (mean 56.4, SD 7.0 vs mean 47.5, SD 12.1; *P*=.001) compared to patients who underwent TOS. However, the deep sleep time, light sleep time, deep sleep ratio, light sleep ratio, REM sleep ratio, and awakening frequency showed no significant differences in patients who underwent MIS versus TOS(all *P*>.05). Multiple linear regression showed that the effects of the surgical mode (MIS or TOS) on Huawei sleep score, total sleep time (*P*=.03), REM sleep time (*P*=.02), and deep sleep continuity score (*P*=.02) were all significant after adjustment for age, sex, Acute Physiology and Chronic Health Evaluation (APACHE) score, use of sedation and analgesia, number of drainage tubes and medical interventions, and surgical site (Table 4).

Table 3. Quantitative sleep analysis of Huawei wearable data.

Variables	MIS[a] (n=28), mean (SD)	TOS[b] (n=33), mean (SD)	*P* value[c]
Total sleep time (min)	503.0 (91.4)	437.9 (144.0)	.04
Deep sleep time (min)	145.7 (55.6)	125.6 (79.7)	.25
Light sleep time (min)	282.0 (92.3)	255.1 (94.5)	.27
REM[d] sleep time (min)	81.0 (52.1)	55.8 (44.5)	.047
Deep sleep ratio (%)	29.7 (7.8)	27.2 (12.4)	.33
Light sleep ratio (%)	55.8 (11.2)	61.1 (17.0)	.15
REM sleep ratio (%)	14.5 (6.9)	11.7 (7.3)	.13
Deep sleep continuity score	56.4 (7.0)	47.5 (12.1)	.001
Awakening frequency	2.4 (2.0)	2.5 (2.3)	.87

[a]MIS: Minimally invasive surgery.

[b]TOS: Traditional open surgery.

[c]Differences among groups were assessed using 2-tailed *t* tests.

[d]REM: rapid eye movement

Table 4. Multivariate linear regression analysis of surgical mode on Huawei sleep indexes.

Variables	Adjusted β[a]	SE	*t* statistic (*df*)	*P* value
Total sleep time	83.314	37.356	2.230 (50)	.03
REM[b] sleep time	36.900	15.501	2.380 (50)	.02
Deep sleep continuity score	8.126	3.363	2.417 (50)	.02

[a]Models were adjusted for age, sex, Acute Physiology and Chronic Health Evaluation (APACHE) score, use of sedation and analgesia, number of drainage tubes and medical interventions, and surgical site.

[b]REM: rapid eye movement.

Discussion

Principal Findings

This study reveals that patients who underwent MIS exhibit better sleep quality, characterized by longer total sleep time, increased REM sleep duration, and higher deep sleep continuity compared to those who underwent TOS. Furthermore, wearable sleep monitoring wristbands provided quantitative sleep data that align with RCSQ assessments. These findings provide initial support for the use of modern sleep monitoring technology in managing sleep in patients in surgical ICUs.

Disruptions to regular sleep-wake cycles can lead to a range of complications, including anxiety, immune disorders, prolonged hospitalization, and social reintegration difficulties [17]. MIS techniques reduced invasiveness and promoted faster recovery, even in critically ill patients. Our results show a significant 10-point improvement in sleep scores for patients who underwent MIS compared to TOS, consistent with previous findings indicating that major surgeries can exacerbate sleep disturbances compared to minor ones [18]. Our findings can be attributed to a combination of physical and psychological factors. Laparoscopic repair is associated with reduced postoperative pain and faster recovery while minimizing psychological stress, as shown by lower anxiety and stress scores [19,20]. Consequently, patients who underwent MIS require fewer narcotics, including opioids, on postoperative days 0 and 1, which reduces sleep disturbances caused by anesthesia use [18,21]. At the molecular level, MIS minimizes mechanically induced damages and immune responses, leading to reduced systemic release of acute phase proteins, leukocytosis, and interleukin-6 [22]. This reduction in inflammatory responses

XSL•FO
RenderX

may affect circadian rhythm and sleep quality by impacting vagal projections, sympathetic ganglia, and the blood-brain barrier [23,24]. Our local quality control report in 2021 also found that patients who underwent MIS had higher scores for sleep and overall in-hospital satisfaction compared to patients who underwent TOS.

Our Huawei data show that patients who underwent MIS experienced a significant reduction in sleep disruption, with an additional hour of total sleep time, an extra 25 minutes of REM sleep, and improved deep sleep continuity. REM sleep is a critical component of sleep quality, characterized by active brain activity, vivid dreaming, and relaxed skeletal muscles [25]. Major surgeries often disrupt REM sleep, with reports of significant suppression following open heart surgery [26] and a decline of REM sleep from 18% to 0% after open cholecystectomy on the night of the surgery [27]. Sleep continuity is a key indicator of sleep quality, as disruptions can lead to fragmented sleep, impaired memory consolidation, and cognitive dysfunction [28]. In our study, both groups had relatively low sleep continuity, consistent with previous research showing that even sufficient sleep duration can be fragmented in postoperative patients in the ICU [29]. Assessing sleep continuity may be a clinically relevant and reproducible method for evaluating sleep disruption in ICU settings [30].

Quantifying the sleep status of patients in the ICU is challenging due to the difficulties in implementing PSG measurement. Laboratory studies have shown that wristbands exhibit high sensitivity (0.965) and accuracy (0.863) compared to PSG [31]. In critically ill patients, wristbands have been validated in correlating with PSG to identify total sleep time and wakefulness [12,32]. However, patient selection in previous studies had affected monitoring accuracy, as wristbands monitoring has much less specificity for patients who are ventilated compared with those who are not (51% vs 83.7%), possibly due to sedation and constraints [33]. Our study focuses on patients who have regained consciousness after anesthesia and found that Huawei wristband sleep scores show high consistency (>95%) with both self-RCSQ and nurse-RCSQ scores. Our results suggest that this technology is suitable for monitoring sleep patterns in patients after they wake up in the ICU.

Wearable sleep monitoring devices offer a distinct advantage in clinical and research settings for patients in the ICU as they provide both objective, quantitative information on sleep states and continuous monitoring capabilities for future analysis. For instance, these devices can facilitate the investigation of sleep disorders and their correlation with adverse outcomes, such as delirium [13,34]. They offer a valuable early warning system for sleep-related complications and serve as a tool to assess the impact of ICU interventions and medications on sleep quality and patient outcomes. Additionally, long-term sleep monitoring, spanning over several days, and its variability (night-to-night variability) can have profound implications on patient outcomes and the well-being of health care associates [35]. The collaboration between medical and engineering professionals enables the utilization of extensive clinical data to guide the development of hardware and software for these devices, evolving them from wearable health monitors into professional medical devices.

This study, conducted at a single center, has several limitations. First, our sleep monitoring period (between 9 PM and 7 AM) was limited by local sleep and clinical work habits, which may introduce bias. In future research, it would be desirable to conduct continuous sleep monitoring in the ICU to examine sleep patterns over a longer period. Second, the higher absolute value of the Huawei sleep score in our study suggests that if we aim for large-scale clinical use, a validated wearable device scoring system with clear and simple standards is necessary. Finally, a larger, multicenter study is necessary to validate the findings through continuous recruitment.

Conclusion

Compared to TOS, MIS contributed to a higher sleep quality for patients in the ICU after surgery, manifested as longer sleep time, longer REM sleep time, and better continuity of deep sleep. Wearable monitoring wristbands hold the potential for quantified sleep assessment in ICU settings.

Acknowledgments

The study was supported by the Wu Jieping Medical Foundation (320.6750.2024-2-4) and the Shanghai Hospital Development Center Foundation (SHDC12024628). Portions of this research have been accepted by ESICM2024 in a poster form.

Data Availability

The data in this study are available from the corresponding author on reasonable request.

Authors' Contributions

ZL, YG, and MT contributed to the conception of the study. CS, CH, and JF were involved in resources, data curation, and formal analysis. CS, YY, and ZL wrote the original draft. ZZ, ZL, YG, MT, and YL validated, reviewed, and edited the manuscript. YY and ZL were responsible for funding acquisition. All authors have approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Richards-Campbell Sleep Questionnaire.
[DOCX File , 17 KB - jmir_v26i1e56777_app1.docx]

References

1. Yilmaz M, Sayin Y, Gurler H. Sleep quality of hospitalized patients in surgical units. Nurs Forum 2012;47(3):183-192. [doi: 10.1111/j.1744-6198.2012.00268.x] [Medline: 22861655]
2. Devlin JW, Skrobik Y, Gélinas C, Needham DM, Slooter AJC, Pandharipande PP, et al. Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. Crit Care Med 2018 Sep;46(9):e825-e873. [doi: 10.1097/CCM.0000000000003299] [Medline: 30113379]
3. Frisk U, Nordström G. Patients' sleep in an intensive care unit--patients' and nurses' perception. Intensive Crit Care Nurs 2003 Dec;19(6):342-349. [doi: 10.1016/s0964-3397(03)00076-4] [Medline: 14637294]
4. Rosenberg-Adamsen S, Kehlet H, Dodds C, Rosenberg J. Postoperative sleep disturbances: mechanisms and clinical implications. Br J Anaesth 1996 Apr;76(4):552-559 [FREE Full text] [doi: 10.1093/bja/76.4.552] [Medline: 8652329]
5. Mohiuddin K, Swanson SJ. Maximizing the benefit of minimally invasive surgery. J Surg Oncol 2013 Oct 23;108(5):315-319. [doi: 10.1002/jso.23398] [Medline: 24037974]
6. Nimmo SM, Foo IT, Paterson HM. Enhanced recovery after surgery: pain management. J Surg Oncol 2017 Oct 05;116(5):583-591. [doi: 10.1002/jso.24814] [Medline: 28873505]
7. Pisani MA, Friese RS, Gehlbach BK, Schwab RJ, Weinhouse GL, Jones SF. Sleep in the intensive care unit. Am J Respir Crit Care Med 2015 Apr 01;191(7):731-738 [FREE Full text] [doi: 10.1164/rccm.201411-2099CI] [Medline: 25594808]
8. Bourne RS, Minelli C, Mills GH, Kandler R. Clinical review: Sleep measurement in critical care patients: research and clinical implications. Crit Care 2007;11(4):226 [FREE Full text] [doi: 10.1186/cc5966] [Medline: 17764582]
9. Elías MN. Assessment and monitoring of sleep in the intensive care unit. Crit Care Nurs Clin North Am 2021 Jun;33(2):109-119 [FREE Full text] [doi: 10.1016/j.cnc.2021.01.008] [Medline: 34023079]
10. Richardson A, Crow W, Coghill E, Turnock C. A comparison of sleep assessment tools by nurses and patients in critical care. J Clin Nurs 2007 Sep 24;16(9):1660-1668 [FREE Full text] [doi: 10.1111/j.1365-2702.2005.01546.x] [Medline: 17459137]
11. Morgenthaler T, Lee-Chiong T, Alessi C, Friedman L, Aurora RN, Boehlecke B, et al. Practice parameters for the clinical evaluation and treatment of circadian rhythm sleep disorders. an American Academy of Sleep Medicine report. Sleep 2007 Nov;30(11):1445-1459 [FREE Full text] [doi: 10.1093/sleep/30.11.1445] [Medline: 18041479]
12. Kakar E, Priester M, Wessels P, Slooter AJ, Louter M, van der Jagt M. Sleep assessment in critically ill adults: a systematic review and meta-analysis. J Crit Care 2022 Oct;71:154102 [FREE Full text] [doi: 10.1016/j.jcrc.2022.154102] [Medline: 35849874]
13. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak C. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003 May 01;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
14. Thomas R, Mietus J, Peng C, Goldberger A. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep 2005 Sep;28(9):1151-1161. [doi: 10.1093/sleep/28.9.1151] [Medline: 16268385]
15. Richards KC, O'Sullivan PS, Phillips RL. Measurement of sleep in critically ill patients. J Nurs Meas 2000;8(2):131-144. [Medline: 11227580]
16. Kumar A, Dogra S, Kaur A, Modi M, Thakur A, Saluja S. Approach to sample size calculation in medical research. Curr Med Res Pract 2014;4(2):87-92. [doi: 10.1016/j.cmrp.2014.04.001]
17. Cilingir D, Hintistan S, Ergene O. Factors affecting the sleep status of surgical and medical patients at a University Hospital of Turkey. J Pak Med Assoc 2016 Dec;66(12):1535-1540. [Medline: 27924961]
18. Ellis BW, Dudley HA. Some aspects of sleep research in surgical stress. J Psychosom Res 1976 Jan;20(4):303-308. [doi: 10.1016/0022-3999(76)90081-7] [Medline: 1003351]
19. Rosen M, Ponsky J. Minimally invasive surgery. Endoscopy 2001 Apr 31;33(4):358-366. [doi: 10.1055/s-2001-13689] [Medline: 11315900]
20. Luo K, Li JS, Li LT, Wang KH, Shun JM. Operative stress response and energy metabolism after laparoscopic cholecystectomy compared to open surgery. World J Gastroenterol 2003 Apr;9(4):847-850 [FREE Full text] [doi: 10.3748/wjg.v9.i4.847] [Medline: 12679946]
21. Scheib SA, Thomassee M, Kenner JL. Enhanced recovery after surgery in gynecology: a review of the literature. J Minim Invasive Gynecol 2019 Feb;26(2):327-343. [doi: 10.1016/j.jmig.2018.12.010] [Medline: 30580100]
22. Kehlet H. Surgical stress response: does endoscopic surgery confer an advantage? World J Surg 1999 Aug 13;23(8):801-807. [doi: 10.1007/s002689900583] [Medline: 10415206]
23. Gögenur I, Bisgaard T, Burgdorf S, van Someren E, Rosenberg J. Disturbances in the circadian pattern of activity and sleep after laparoscopic versus open abdominal surgery. Surg Endosc 2009 May 2;23(5):1026-1031. [doi: 10.1007/s00464-008-0112-9] [Medline: 18830755]
24. Irwin MR. Sleep and inflammation: partners in sickness and in health. Nat Rev Immunol 2019 Nov 9;19(11):702-715. [doi: 10.1038/s41577-019-0190-z] [Medline: 31289370]

XSL•FO
RenderX

25. Peever J, Fuller PM. The biology of REM sleep. Curr Biol 2017 Nov 20;27(22):R1237-R1248 [FREE Full text] [doi: 10.1016/j.cub.2017.10.026] [Medline: 29161567]
26. Orr WC, Stahl ML. Sleep disturbances after open heart surgery. Am J Cardiol 1977 Feb;39(2):196-201. [doi: 10.1016/s0002-9149(77)80191-4] [Medline: 299975]
27. Knill R, Moote C, Skinner M, Rose E. Anesthesia with abdominal surgery leads to intense REM sleep during the first postoperative week. Anesthesiology 1990 Jul;73(1):52-61 [FREE Full text] [doi: 10.1097/00000542-199007000-00009] [Medline: 2360740]
28. Wilcox ME, McAndrews MP, Van J, Jackson JC, Pinto R, Black SE, et al. Sleep fragmentation and cognitive trajectories after critical illness. Chest 2021 Jan;159(1):366-381. [doi: 10.1016/j.chest.2020.07.036] [Medline: 32717265]
29. Friese RS, Diaz-Arrastia R, McBride D, Frankel H, Gentilello L. Quantity and quality of sleep in the surgical intensive care unit: are our patients sleeping? J Trauma 2007 Dec;63(6):1210-1214. [doi: 10.1097/TA.0b013e31815b83d7] [Medline: 18212640]
30. Drouot X, Bridoux A, Thille AW, Roche-Campo F, Cordoba-Izquierdo A, Katsahian S, et al. Sleep continuity: a new metric to quantify disrupted hypnograms in non-sedated intensive care unit patients. Crit Care 2014 Nov 25;18(6):628 [FREE Full text] [doi: 10.1186/s13054-014-0628-4] [Medline: 25420997]
31. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 2013 Nov 01;36(11):1747-1755 [FREE Full text] [doi: 10.5665/sleep.3142] [Medline: 24179309]
32. Schwab KE, Ronish B, Needham DM, To AQ, Martin JL, Kamdar BB. Actigraphy to evaluate sleep in the intensive care unit. a systematic review. Ann Am Thorac Soc 2018 Sep;15(9):1075-1082 [FREE Full text] [doi: 10.1513/AnnalsATS.201801-004OC] [Medline: 29944386]
33. Delaney LJ, Litton E, Melehan KL, Huang HC, Lopez V, van Haren F. The feasibility and reliability of actigraphy to monitor sleep in intensive care patients: an observational study. Crit Care 2021 Jan 29;25(1):42 [FREE Full text] [doi: 10.1186/s13054-020-03447-8] [Medline: 33514414]
34. Jaiswal SJ, Bagsic SRS, Takata E, Kamdar BB, Ancoli-Israel S, Owens RL. Actigraphy-based sleep and activity measurements in intensive care unit patients randomized to ramelteon or placebo for delirium prevention. Sci Rep 2023 Jan 26;13(1):1450 [FREE Full text] [doi: 10.1038/s41598-023-28095-0] [Medline: 36702822]
35. Morgenthaler T, Alessi C, Friedman L, Owens J, Kapur V, Boehlecke B, et al. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. Sleep 2007 Apr;30(4):519-529. [doi: 10.1093/sleep/30.4.519] [Medline: 17520797]

Abbreviations

ICU: intensive care unit
MIS: minimally invasive surgery
PSG: polysomnography
RCSQ: Richards-Campbell Sleep Questionnaire
REM: rapid eye movement
TOS: traditional open surgery

Edited by T de Azevedo Cardoso; submitted 19.02.24; peer-reviewed by TI Oh, D Yang, CA Austin; comments to author 04.06.24; revised version received 26.07.24; accepted 29.10.24; published 22.11.24.

Please cite as:
Shang C, Yang Y, He C, Feng J, Li Y, Tian M, Zhao Z, Gao Y, Li Z
Quantitative Impact of Traditional Open Surgery and Minimally Invasive Surgery on Patients' First-Night Sleep Status in the Intensive Care Unit: Prospective Cohort Study
J Med Internet Res 2024;26:e56777
URL: https://www.jmir.org/2024/1/e56777
doi:10.2196/56777
PMID:

©Chen Shang, Ya Yang, Chengcheng He, Junqi Feng, Yan Li, Meimei Tian, Zhanqi Zhao, Yuan Gao, Zhe Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Telehealth on Health Disparities Associated With Travel Time to Hospital for Patients With Recurrent Admissions: 4-Year Panel Data Analysis

Youyou Tao[1], PhD; Ace Vo[1], PhD; Dezhi Wu[2], PhD; Junyuan Lin[1], PhD; Kala Seal[1], PhD; Abhay Mishra[3], PhD; Arindam Brahma[1], PhD

[1]Loyola Marymount University, Los Angeles, CA, United States
[2]University of South Carolina, Columbia, SC, United States
[3]Iowa State University, Ames, IA, United States

Corresponding Author:
Youyou Tao, PhD
Loyola Marymount University
1 LMU Drive
Los Angeles, CA, 90045
United States
Phone: 1 2069722430
Email: youyou.tao@lmu.edu

Abstract

Background: Geographic, demographic, and socioeconomic differences in health outcomes persist despite the global focus on these issues by health organizations. Barriers to accessing care contribute significantly to these health disparities. Among these barriers, those related to travel time—the time required for patients to travel from their residences to health facilities—remain understudied compared with others.

Objective: This study aimed to explore the impact of telehealth in addressing health disparities associated with travel time to hospitals for patients with recurrent hospital admissions. It specifically examined the role of telehealth in reducing in-hospital length of stay (LOS) for patients living farther from the hospital.

Methods: We sourced the data from 4 datasets, and our final effective sample consisted of 1,600,699 admissions from 536,182 patients from 63 hospitals in New York and Florida in the United States from 2012 to 2015. We applied fixed-effect models to examine the direct effects and the interaction between telehealth and patients' travel time to hospitals on LOS. We further conducted a series of robustness checks to validate our main models and performed post hoc analyses to explore the different effects of telehealth across various patient groups.

Results: Our summary statistics show that, on average, 22.08% (353,396/1,600,699) of patients were admitted to a hospital with telehealth adopted, with an average LOS of 5.57 (SD 5.06) days and an average travel time of about 16.89 (SD 13.32) minutes. We found that telehealth adoption is associated with a reduced LOS ($P<.001$) and this effect is especially pronounced as the patients' drive time to the hospital increases. Specifically, the coefficient for drive time is –0.0079 ($P<.001$), indicating that for every additional minute of driving time, there is a decrease of 0.0079 days (approximately 11 minutes) in the expected LOS. We also found that telehealth adoption has a larger impact on patients frequently needing health services, patients living in high internet coverage areas, and patients who have high virtualization potential diseases.

Conclusions: Our findings suggest that telehealth adoption can mitigate certain health disparities for patients living farther from hospitals. This study provides key insights for health care practitioners and policy makers on telehealth's role in addressing distance-related disparities and planning health care resources. It also has practical implications for hospitals in resource-limited countries that are in the early stages of implementing telehealth.

(J Med Internet Res 2024;26:e63661) doi:10.2196/63661

KEYWORDS

telehealth; health disparities; travel time; length of stay; recurrent admissions; health care access; virtualization potential; virtual care

XSL•FO
RenderX

Introduction

The World Health Organization has stated that health services quality should not vary by factors such as ethnicity, gender, socioeconomic status, and geographic location [1]. Unfortunately, geographic, demographic, and socioeconomic differences exist in health outcomes despite the focus on such issues by health organizations worldwide [2,3]. Such disparities have adverse effects on particular groups, as evidenced by disproportionately worse medical outcomes [2,3]. The barriers to access care for these populations contribute to significant health disparities, defined as health differences related to economic, social, or environmental disadvantages [4]. Compared with other access barriers, access barriers related to travel time (ie, the time required for patients to travel from their residences to health facilities) remain understudied [5].

This paper focuses on this particular access barrier, travel time. Previous studies found that travel time to the hospital can cause considerable health disparity [6,7]. Recent literature has explored the effect of travel time on patients' in-hospital length of stay (LOS)—the number of days a patient stays in the hospital from admission to discharge—and found mixed results. For example, Simpson et al [8] found no difference in LOS based on travel time, while Jackson et al [9] discovered a positive association between these 2 variables. Rana et al [10] reported that the mean difference for LOS among patients with chronic obstructive pulmonary disease traveling short (0-15 km) and intermediate (>15-50 km) distances was statistically insignificant. However, patients traveling long distances (>50 km) had a significantly shorter LOS compared with both short- and intermediate-distance patients. This study suggested that this could be due to the distance decay phenomenon, where the interaction between 2 locations diminishes as the distance between them increases [10]. Specifically, patients who live farther from the hospital might experience a shorter LOS compared with those residing closer [10]. This distance decay phenomenon is also supported by other studies, which have shown that compared with patients living closer to health care facilities, those living farther away had worse health outcomes such as lower survival rates in hospitals and higher rates of nonattendance at follow-up due to the longer distance [6,7].

While documenting the existence of health disparities that are associated with travel time to the hospital, scholars suggested that future studies should focus on examining how travel-related access barriers can cause adverse patient outcomes and further propose feasible interventions that could mitigate these health disparities [11]. Telehealth, defined as "the use of electronic information and telecommunication technologies to support long-distance clinical health care, patient and professional health-related education, health administration, and public health [12]," can help by providing virtual care that allows patients to manage their conditions and communicate with health care providers without traveling [13-15]. A 2022 report from the American Medical Association's Physician Practice Benchmark Surveys shows that 74.4% of physicians reported using telehealth in their practice [16]. Telehealth is commonly used in specialties such as cardiology, emergency medicine, family medicine, hematology, general internal medicine, neurology, obstetrics and gynecology, oncology, and pediatrics [17]. However, even with the growing adoption of telehealth in clinical settings, the specific impact of telehealth on health outcomes related to travel time remains underexplored.

Among various health outcome measures, our study aims to investigate how telehealth can reduce disparities in in-hospital LOS. Reducing LOS can not only alleviate the financial burden on patients but also contribute to the broader impact of decreasing societal health care costs [18]. Various health information technology (HIT) interventions, including telehealth, have been shown in the literature to statistically significantly reduce LOS [18-20]. Telehealth, in particular, has been shown to improve health outcomes, including LOS reduction, through two primary mechanisms: (1) real-time interventions that require real-time interaction between patients and their health providers, such as video conferencing, telephone communication, and telemonitoring equipment, facilitating immediate patient-provider interaction and timely care management [21-23], and (2) asynchronous interventions that do not require immediate interaction between patients and their health providers, including remote monitoring systems, interactive platforms, and in-home secure messaging through a patient portal using either a phone or a computer, which provide access to patient educational resources and enable continuous patient monitoring, with patients' data reviewed at a later time for medical treatment [15,24-26]. Through these mechanisms, telehealth, alongside other HIT applications, plays a significant role in shortening LOS by enabling early intervention through the efficient gathering of useful information [27] and facilitating consultation, diagnosis, treatment, care management, education, and patients' self-care [28]. Studies have shown that telehealth decreased LOS for patients in progressive care units without significant cost increases [29], for patients with vigorous chronic obstructive pulmonary disease after 7 months of monitoring [27], and for patients in rural and small hospitals through real-time telehealth interventions [30]. However, to the best of our knowledge, no study has specifically explored how telehealth could mitigate health disparities associated with travel time to hospitals in terms of LOS.

Given the pressing need to address health disparities associated with travel time to hospitals, our study aims to comprehensively examine how health disparity in patients' in-hospital LOS caused by long travel time to the hospital could be mitigated through telehealth implementation [27,29,30]. Our overarching research question is, "How can telehealth adopted by hospitals mitigate the adverse effects of travel time to the hospital, reduce health disparity, and influence patients' health outcomes in terms of LOS?"

Methods

We present the *Methods* section following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines, and we also provided a checklist of items for the STROBE statement in Multimedia Appendix 1 [31].

XSL•FO
RenderX

Study Design and Study Setting

The data applied in this study were sourced from 4 sources. First, we used data from the Healthcare Cost and Utilization Project's State Inpatient Databases (HCUP-SID) from 2012 to 2015 to obtain patient characteristics, admission, and clinical information [32]. The patient's zip code information was part of this dataset. Second, we applied data from the American Hospital Association (AHA) annual surveys from 2012 to 2015 to obtain hospital characteristic data [33] including hospital location information such as latitude and longitude. Using this locational data and patient zip code, we computed "drive time to hospital" for each patient. Third, we used AHA's information technology (IT) supplement files for 2011-2014 to obtain HIT implementation data [34]. Considering the lagged effects of HIT adoption, we followed the method that has been widely adopted to map the datasets with HIT variables lagged by 1 year [35]. Fourth, we collected additional census-based data, such as English proficiency level at the census tract level. We then matched the reference data between the census tracts and the United States Postal Service's zip codes to allocate each census tract data proportionally.

We merged data from the Healthcare Cost and Utilization Project (HCUP), AHA surveys, and AHA IT files with unique hospital identifications before merging the datasets with census data using patients' zip codes from HCUP-SID. Furthermore, we filtered out data points beyond the 99th percentile for LOS, patient visit frequency, and travel time. To remove the outliers, we only included patients with an in-hospital stay of fewer than 35 days, fewer than 18 hospital visits over 4 years, and a travel time of less than 103.46 minutes in our sample data. To examine the effect of telehealth on recurrent hospital admissions, we only kept patients with 2 or more admissions in our sample. To ensure a consistent follow-up period for each patient, we excluded those whose first admission occurred in the dataset's final year (2015), applying a 1-year cutoff for the follow-up duration. Our final effective sample consists of 1,600,699 admissions originating from 536,182 patients within 63 hospitals in New York and Florida in the United States from 2012 to 2015. This final dataset contains no missing values. Our data processing procedure is visualized in Figure 1. We chose datasets from New York and Florida because the high volume of hospital admissions in these 2 states and the availability of linkage variables that allowed us to connect patients with their hospital admissions so that we can examine telehealth's impact on patients with recurrent admissions.

Figure 1. Data preparation method. AHA: American Hospital Association; freq.: frequency; HCUP-SID: Healthcare Cost and Utilization Project's State Inpatient Datasets; IT: information technology; LOS: length of stay.

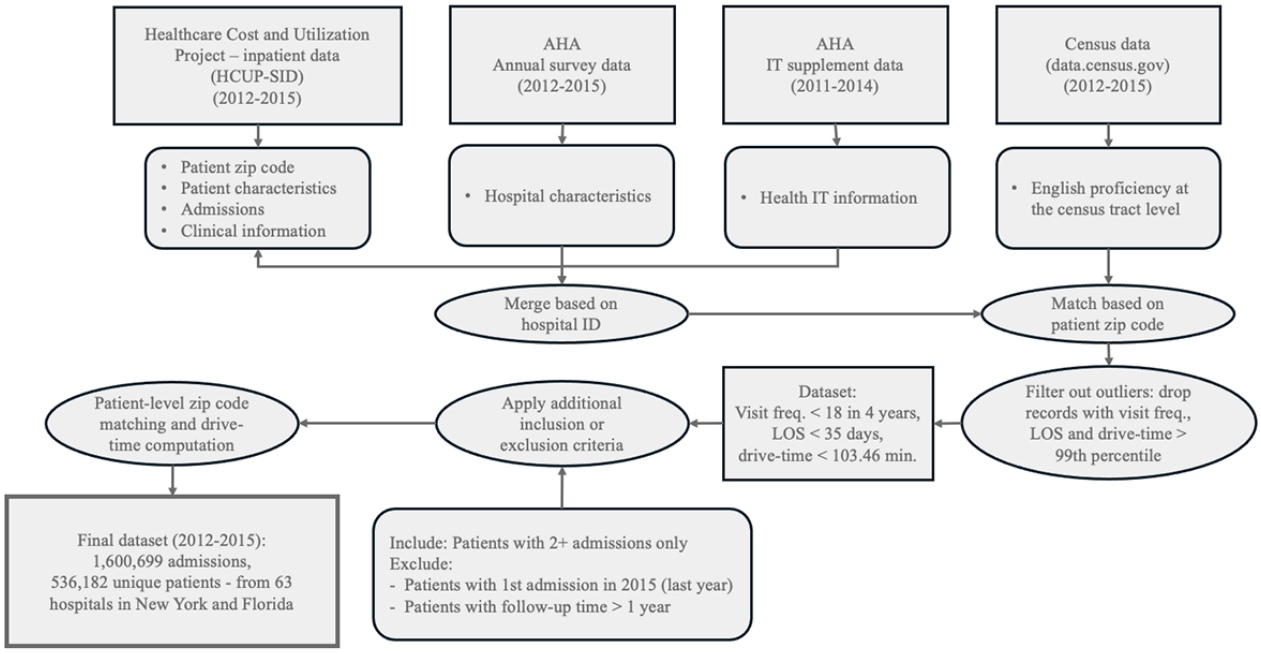

Study Outcome

We applied in-hospital LOS to investigate telehealth's impact on patients' health outcomes [20,36]. In-hospital LOS, measured as the number of days of hospital stays from admission to discharge, serves as a proxy for hospital operations efficiency measures [20,36].

Study Variables

Our main study variables were telehealth implementation and travel time to the hospital. We treated telehealth implementation in hospitals as a categorical variable, where 1 indicates telehealth is adopted by the hospital—meaning the hospital has a computerized system that supports telehealth—and 0 otherwise [33]. Telehealth implementation measured by this item includes health care services provided at a distance through various modalities such as remote monitoring, video conferencing, wireless communications, and electronic consults [37].

Travel time to the hospital was calculated using the ArcGIS Network Analyst (Environmental Systems Research Institute) extension with the origin-destination cost matrix. The matrix

XSL•FO
RenderX

provides both travel time and travel distance along the shortest route from the origins—patients' locations, represented by the centroids of their zip codes—to the destinations, that is, hospitals' physical addresses. To obtain travel time and travel distance, the origin-destination cost matrix was solved within ArcGIS Pro version 3.3.1 using the native Network Analyst solver. This solver is a subscription service from Environmental Systems Research Institute, which maintains and regularly updates the network solver. The process of obtaining drive time is as follows:

1. Geocode (transforming text addressed to geospatial data point) the hospital addresses as points on the map, retaining hospital ID as identifiers
2. Obtain New York and Florida zip code shapefiles from the census (which are equivalent to the United States Postal Service's zip code)
3. Convert the zip code shapefiles, which are polygons (spatial areas), to their respective centroids (midpoints of the polygons) as points on the map, retaining zip codes as identifiers
4. Link patients' zip codes to the respective points on the map. The points will serve as proxies of the actual patients' addresses
5. For each state (New York and Florida), assign the unique patients' zip code locations as origins and hospital addresses as destinations, then calculate the origin-destination cost matrix to obtain pair-wise travel time and travel distance for each origin and destination
6. Link the travel time and travel distance from the origin-destination cost matrix to the respective encounter data using patient zip codes and hospital IDs

We also included community-level, hospital-level, patient-level, and admission-level variables as control variables to account for other factors that may influence in-hospital LOS. At the community level, we included the English proficiency level, measured by the percentage of the population that speaks English less than well in each zip code. Studies showed that compared with patients without a language barrier, patients with a language barrier experience more health care disparity in care, including longer LOS [38]. For patients to use telehealth services from home, the first requirement is to have internet access, and recent studies show that internet access can empower patients, influence health outcomes, and change their health service usage behaviors [39,40]. Thus, we also controlled the percentage of internet coverage at the zip code level.

We applied hospital characteristics at the hospital level that may influence in-hospital LOS, such as hospital bed size, profit status, in a health system or not, and market share. Market share is measured by the Herfindahl-Hirschman Index. We also included HIT variables to control for the other HIT effects on LOS, which are clinical decision support system, computerized provider order entry, electronic clinical documentation, results viewing, and health information exchange (refer to Multimedia Appendix 2 for the detailed list of technologies and coding for HIT control variables).

At the patient level, we included age at discharge, gender, and race. At the admission level, we included insurance types, including Medicare, Medicaid, and private insurance; the total number of comorbidities; chronic diagnoses; diagnoses; procedure; total charge at the current admission; the number of times the patient visited the hospitals since the first hospital visit; and the number of days between the current visit and the previous visit. We used 17 different types of body systems (refer to "Disease types" in Multimedia Appendix 3 for details) and a dummy variable that records whether the patient's primary diagnosis is in the same body system at the current admission compared with the last admission. We presented the summary statistics of control variables and additional variables for post hoc analysis in Multimedia Appendix 3.

Ethical Considerations

The HCUP data used in this study are classified as a limited dataset, as defined by the HCUP website [41]. According to the website [41], "HCUP databases are limited data sets. HCUP databases conform to the definition of a limited data set. A limited data set is healthcare data in which 16 direct identifiers, specified in the Privacy Rule, have been removed. Under HIPAA, review by an institutional review board is not required for use of limited data sets." The AHA's annual surveys and IT supplement files used in this study were collected at the hospital level and did not involve human participants. This study was granted exemption from institutional review board review by Loyola Marymount University's Institutional Review Board (LMU IRB 2024 FA12-R).

Statistical Analysis

After conducting descriptive analyses, we used STATA/SE 18.0 (StataCorp) for the subsequent data analysis. In this study, fixed-effect regression models were applied to assess the impact of telehealth on in-hospital LOS [42]. The model is specified in equations 1 and 2.

$$LOS_{ijtv} = \beta_0 + \beta_1 Telehealth_{j,t-1} + \beta_2 Drive_Time_{ijtv} + Control_{ijtv}\eta + \lambda_t + \alpha_v + \gamma_j + \varepsilon_{itv} \quad (1)$$

In equation 1, LOS_{ijtv} is the dependent variable that represents the in-hospital LOS for patient i's vth hospital visit in hospital j at year t. $Telehealth_{j,t-1}$ indicates whether telehealth was adopted by hospital j in the previous year ($t-1$). $Drive_Time_{ijtv}$ represents the drive time for patient i's vth hospital visit to hospital j in year t. The focal parameters of interest in this model are β_1 and β_2, which capture the direct effect of telehealth implementation by hospital j in the previous year ($t-1$) and the direct effect of drive time for patient i to hospital j for the vth visit at year t. $Control_{ijtv}$ represents community-level, hospital-level, patient-level, and admission-level factors as described above that may influence patient in-hospital LOS. We include γ_j in the model to account for the hospital fixed effects that absorb unobserved hospital characteristics that do not vary over time, λ_t to absorb the change over time captured by the year fixed effect, α_v to account for the fixed effects for hospital visit number that absorb the unobserved factors that could vary with each hospital visit, and ε_{itv} as the idiosyncratic error term.

XSL•FO
RenderX

$$LOS_{ijtv} = \beta_0 + \beta_1 Telehealth_{j,t-1} + \beta_2 Drive_Time_{ijtv} + \beta_3 Telehealth_{j,t-1} \times Drive_Time_{ijtv} + Control_{ijtv}\eta + \lambda_t + \alpha_v + \gamma_j + \varepsilon_{itv} \quad (2)$$

In equation 2, the focal parameter of interest is β_3, which captures the interaction effect between telehealth and drive time to capture how the effect of telehealth on LOS varies with patients' drive time to the hospital. In both models, robust SEs were applied to account for potential heteroskedasticity, and the errors were further clustered at the patient level to account for the possible within-patient correlations.

Results

Impact of Telehealth and Travel Time

Table 1 presents summary statistics of our main variables. Our summary statistics show that, on average, 22.08% (353,396/1,600,699) of patients were admitted to a hospital with telehealth adopted, the average LOS was 5.57 (SD 5.06) days, and the average travel time was about 16.89 (SD 13.32) minutes.

Table 1. Summary statistics for main variables.

Variable name	Variable, mean (SD)	Variable, minimum	Variable, maximum
Telehealth[a]	0.22 (0.42)	0	1
Length of stay (in days)	5.569 (5.06)	0	34
Travel time (in minutes)	16.89 (13.32)	0.89	103.43

[a]Dummy variables with 2 values (ie, 0 and 1).

To explore whether telehealth mitigates the effect of travel time to the hospital on patient in-hospital LOS, we first mapped the driving time to the hospital in minutes in Florida and New York in Figures 2A and 2B, respectively. From Figure 2, we observed that people with longer drive time to hospitals were distributed in both the north and south of Florida and on the edges of New York.

Figure 2. Travel time to hospital in minutes: (A) Florida and (B) New York.

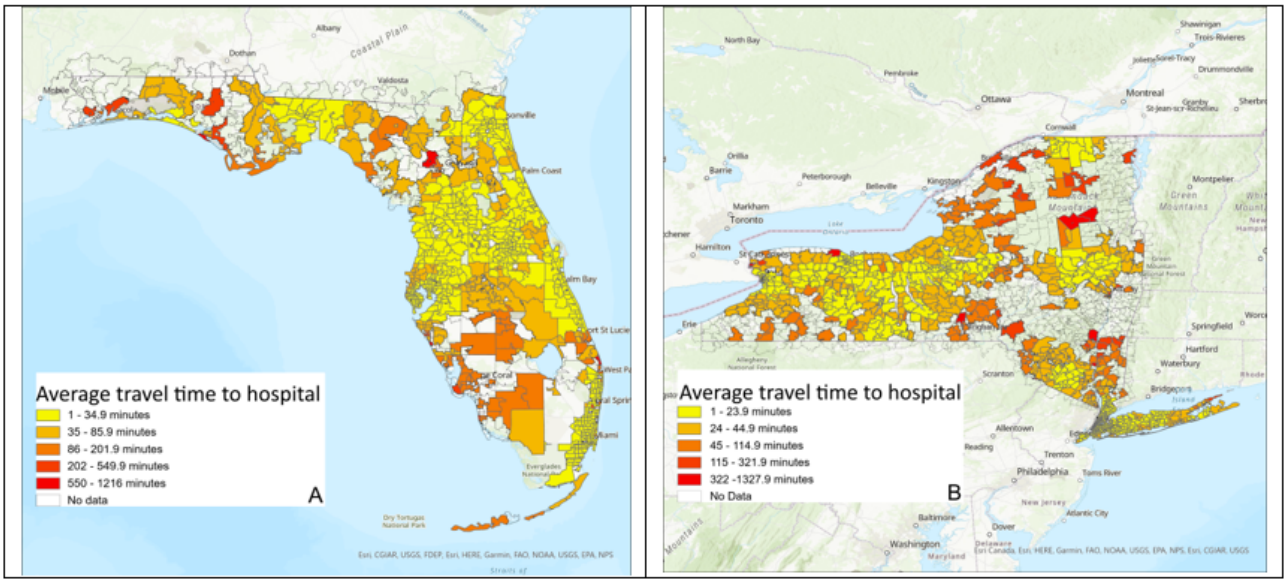

Next, we present our analysis result in Table 2. In Table 2, model 1, we can infer that both the direct effects of the implementation of telehealth and drive time are associated with a decreased LOS at the $P<.001$ level. Specifically, patients admitted to hospitals that have adopted telehealth services experience a reduction in in-hospital LOS by 0.0431 days (approximately 62 min) compared with those admitted to hospitals without telehealth implementation. Given the average hospital-adjusted expense per inpatient per day was US $3025 in the United States in 2022 [43], telehealth implementation can save around US $130.38 in hospital cost per patient stay.

XSL•FO
RenderX

Table 2. Effect of telehealth adoption and drive time to hospital on length of stay[a,b].

Models	Main models		Robustness check					
	Main effect	Interaction effect added	Comorbidities added	Dynamic model	Patient-level FE[c] model	NB[d] model	Log (LOS[e])	PSM[f]
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Telehealth								
Coefficient	*–0.0431*[g]	0.0056	0.0103	0.0035	0.0092	0.0009	0.0056	0.0310
P value	*<.001*	.71	.50	.85	.65	.73	.01	.25
Robust SE	*0.01*	0.02	0.02	0.02	0.02	0.003	0.002	0.03
Drive time								
Coefficient	*–0.0079*	–0.0073	–0.0067	–0.0060	–0.0065	–0.0009	–0.0008	–0.0057
P value	*<.001*	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Robust SE	*0.000*	0.000	0.000	0.000	0.001	0.000	0.000	0.001
Telehealth×Drive time								
Coefficient	__[h]	*–0.0026*	*–0.0026*	*–0.0029*	*–0.0023*	*–0.0003*	*–0.0003*	*–0.0023*
P value	—	*<.001*	*<.001*	*<.001*	*.01*	*<.001*	*<.001*	*.03*
Robust SE	—	*0.001*	*0.001*	*0.000*	*0.000*	*0.000*	*0.000*	*0.001*
LOS_{v-1}[i]								
Coefficient	—	—	—	0.0564	—	—	—	—
P value	—	—	—	<.001	—	—	—	—
Robust SE	—	—	—	0.001	—	—	—	—
Admission-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patient-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hospital-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Community-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year and hospital fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Comorbidities	No	No	Yes	No	No	No	No	No
Observations, n	1,600,699	1,600,699	1,600,699	1,070,901	1,600,699	1,600,699	1,600,699	542,089
Patients, n	536,182	536,182	536,182	462,234	536,182	536,182	536,182	260,139
Adjusted R^2	0.582	0.582	0.588	0.594	0.569	—	0.486	0.581

[a]Robust SEs clustered by patients.
[b]Constant is included in each of the models.
[c]FE: fixed effect.
[d]NB: negative binomial.
[e]LOS: length of stay.
[f]PSM: propensity score matching.
[g]Italicized values represent [insert reason].
[h]Not applicable.
[i]LOS_{v-1}: length of stay for patient's *v*–1st visit to the hospital.

From Table 2, model 1, we also found that the coefficient for drive time is –0.0079, with a *P* value <.001, indicating that longer drive times to the hospital are significantly associated with a slight decrease in LOS. For every additional minute of driving time, there is a decrease of 0.0079 days (approximately 11 min) in the expected LOS. From Table 2, model 2, we found that the interaction effect of the telehealth implementation and the drive time is associated with a decreased LOS at the *P*<.001 level, suggesting that telehealth may be particularly effective

XSL•FO
RenderX

at reducing LOS for patients who live farther away from the hospital.

To visualize the impact of telehealth adoption on the LOS for patients with various travel times to the hospital, we plotted this effect in the interaction plot in Figure 3. We plotted 2 separate lines, 1 for hospitals that implemented telehealth and 1 for those that did not. We found that for patients admitted to hospitals with telehealth, there is a steeper decrease in the LOS as the drive time increases from 0 to 100 minutes (indicated by the solid red line with triangle markers). Conversely, for patients admitted to hospitals without telehealth, the LOS decreases as the drive time increases, but the decrease is not as steep (indicated by the dashed blue line with square markers). This result indicates telehealth implementation is more effective in reducing LOS for patients who live farther from the hospital.

Figure 3. The impact of telehealth implementation on length of stay by drive time.

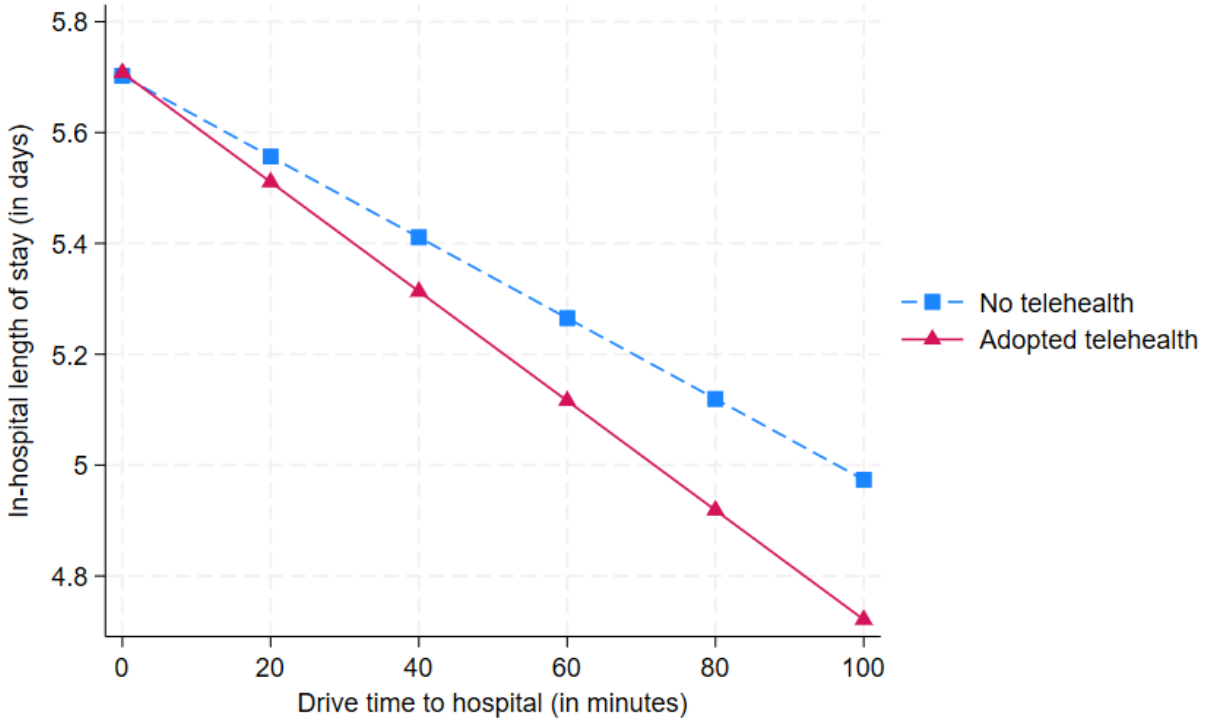

To examine the robustness of our results, we conducted 5 sets of tests, the results of which are presented in Table 2, models 3-8. We first included an additional 29 types of comorbidities (refer to "Comorbidity type" in Multimedia Appendix 3 for more details) to control for the severity of the patient's condition (result in model 3). Next, we incorporated the patients' lagged LOS from the previous admission as an independent variable to account for the potential impact of past LOS on the current LOS (result in model 4). In our main models, we accounted for the hospital-level fixed effects but not the patient-level fixed effects. Accounting for hospital-level fixed effects allows us to absorb unobserved hospital characteristics that do not vary over time. However, by not including patient-level fixed effects, we may overlook unobserved patient characteristics that also do not vary over time. To address this in model 5, we considered patient-level fixed effects by introducing patient-specific intercepts in this model to control for unobserved, patient-specific characteristics that are time-invariant. We found the results from models 3-5 consistent with our main model.

Furthermore, in our main model, we applied LOS as our dependent variable for better result interpretation. According to the literature, despite the right-skewness of LOS, linear regression is validated as statistically sound for large data samples, even when the dependent variable is not normally distributed [44-46]. The variance of the LOS is 25.66 days, which exceeds the mean of 5.57 days, indicating overdispersion in the data. To further test the robustness of our results, we applied a negative binomial model to account for skewness and over-dispersion (result in model 6). Furthermore, we examined the impact of telehealth adoption and drive time to the hospital on the logarithmically transformed LOS (result in model 7). We found the results from both models 6 and 7 to be consistent with our main model.

In model 8, we used propensity score matching to construct a balanced comparison between the treatment group (ie, patients admitted to hospitals that have implemented telehealth) and the control group (ie, those admitted to hospitals without telehealth). We first calculated propensity scores using a logistic regression model with hospital-level factors (such as the implementation level of clinical decision support system, computerized provider order entry, electronic clinical documentation, results viewing, and health information exchange) and patient-level and admission-level (such as number of comorbidities, age, race, insurance type, current visit number, and body system categorizations). After matching patients using the k-nearest neighbors approach, we assessed the balance of hospital-level and patient-level variables between the treatment and control groups and found the standardized difference did not exceed a

XSL•FO
RenderX

10% threshold, indicating that treatment and control groups share similar statistical properties [47]. Postmatching analysis was conducted on the weighted and matched sample and clustered by patients to account for within-group correlations. The results, presented in model 8, are consistent with our main model. Overall, the results in models 3-8 show that our main results are robust and consistent among various robustness tests.

Post Hoc Analysis

To explore the nuanced effects of the direct effect of telehealth adoption and the interaction effects between telehealth adoption and drive time to hospital for subgroups of patients, we conducted post hoc analyses and present the result in Table 3. We first explored whether telehealth adoption has a larger impact on patients who need frequent hospital visits and have a high level of comorbidity by dividing the patients into groups 1 and 2 based on their frequency of hospital visits and comorbidity counts. Furthermore, we investigated whether telehealth adoption has a larger impact on patients living in areas with high internet coverage compared with those in areas with low coverage by dividing the patients into groups 3 and 4 based on median internet coverage. The literature suggests that internet coverage plays a significant role in delivery of various types of health care and broadband access is indeed a social determinant of health [48]. Access to broadband-based internet can allow specialists to collaborate with doctors in rural areas for saving patients with stroke or improving access to mental health care for the veterans [49]. These are a few examples of the myriad ways that internet access, particularly access to broadband internet, can serve as an enabler of delivery of telehealth across the population. In fact, research further indicates that a lack of broadband access can create a barrier for accessing telehealth services [50]. For our research, we surmise that while driving time itself can be a determinant for the level of use of telehealth services, there is a prerequisite of the presence of broadband within the community for accessing and using the telehealth services. In other words, the internet coverage itself would play a moderating role in determining the access and effectiveness of telehealth services (as measured by LOS) and we are trying to capture that in our post hoc analyses.

Table 3. Post hoc patient group description.

Post hoc patient group	Visit frequency (median=2)	Inclusion condition	Comorbidity (median=3)	Internet coverage (median=80.04%)
Group 1: High frequency and high comorbidity	>2	AND	>3	—[a]
Group 2: Low frequency or low comorbidity	≤2	OR	≤3	—
Group 3: High internet coverage	—	—	—	≥80.04%
Group 4: Low internet coverage	—	—	—	<80.04%

[a]Not applicable.

In the subgroup analysis (Table 4), both the direct telehealth implementation effect and the interaction effects consistently reduce LOS for groups 1 and 2 (see the italicized results in Table 4), with a more pronounced effect for group 1, high frequency and high comorbidity group, indicating that telehealth could be effectively integrated into the care plans for patients who live farther from the hospitals, frequently engage with health care systems, and have complex health needs.

Furthermore, we found that telehealth decreases LOS for the high internet coverage group (group 3) but not in the low internet coverage group (group 4). We also found that the interaction effects between telehealth and drive time are consistently associated with reduced LOS for both groups, but the effect size is larger for patients living in high internet coverage areas (see the italicized results in Table 4). This highlights the importance of high internet coverage as a prerequisite to fully benefit from telehealth.

XSL•FO
RenderX

Table 4. Effect of telehealth adoption and drive time to hospital on length of stay by patient groups[a-d].

Subgroup and effects	Telehealth	Drive time	Telehealth×Drive time	Observations, n
Group 1: High frequency and high comorbidity				308,365
Direct effect				
Coefficient	*–0.0844*[e]	–0.0051	—[f]	
P value	*.002*	<.001	—	
Robust SE	*0.03*	0.001	—	
Interaction effects added				
Coefficient	–0.0093	–0.0038	*–0.0044*	
P value	.80	<.001	*.009*	
Robust SE	0.04	0.001	*0.002*	
Group 2: Low frequency or low comorbidity				1,292,334
Direct effect				
Coefficient	*–0.0483*	–0.0082	—	
P value	*<.001*	<.001	—	
Robust SE	*0.01*	0.000	—	
Interaction effects added				
Coefficient	–0.0000	–0.0076	*–0.0025*	
P value	.999	<.001	*<.001*	
Robust SE	0.02	0.000	*0.001*	
Group 3: High internet coverage				851,569
Direct effect				
Coefficient	*–0.0500*	–0.0105	—	
P value	*.001*	<.001	—	
Robust SE	*0.02*	0.000	—	
Interaction effects added				
Coefficient	0.0161	–0.0095	*–0.0039*	
P value	.42	<.001	*<.001*	
Robust SE	0.020	0.000	*0.001*	
Group 4: Low internet coverage				*749,130*
Direct effect				
Coefficient	*0.0095*	–0.0031	—	
P value	*.62*	<.001	—	
Robust SE	*0.02*	0.001	—	
Interaction effects added				
Coefficient	0.0705	–0.0024	*–0.0027*	
P value	.008	<.001	*.004*	
Robust SE	0.03	0.001	*0.001*	

[a]Robust SEs clustered by patients.

[b]Constant is included in each of the models.

[c]Control variables including community-level, hospital-level, patient-level, and admission-level are included in the models.

[d]Year and hospital fixed effects are included in the models.

[e]Italicized values represent [insert reason].

[f]Not applicable.

XSL•FO
RenderX

From Table 4, we found that telehealth adoption has a larger impact on groups 1 and 3. Thus, following these findings, in our next set of post hoc tests, we focus on these 2 patient groups and study the impact of telehealth adoption among patients with high versus low virtualization potential diseases, using a classification based on the degree of process virtualization [26]—where physical interactions between people and objects are replaced by virtual interactions. Telehealth is likely more effective among patients with diseases with high virtualization potential (DHVP) compared with those with diseases with low virtualization potential (DLVP) [26]. Based on the classification suggested by Ayabakan et al [26], we categorized patients into 2 groups—DHVP and DLVP. DHVP include endocrine, nutritional, and metabolic diseases, mental illnesses, diseases of the skin and subcutaneous tissue, and musculoskeletal system diseases. All other disease categories were classified as DLVP. Our results, as shown in Table 5, indicate that the direct impact of telehealth adoption is more effective in reducing LOS for patients with DHVP compared with those with DLVP. These findings underscore how the effectiveness of telehealth in reducing LOS is influenced by disease type.

Table 5. Effect of telehealth adoption and drive time to hospital on length of stay by disease virtualization potential[a-d].

Group	Direct effect of telehealth	Robust SE	*P* value	Observations, n
Group 5: High frequency, high comorbidity, and DHVP[e]	*–0.1736*[f]	0.03	.02	39,361
Group 6: High frequency, high comorbidity, and DLVP[f]	*–0.0731*	0.03	.01	269,004
Group 7: High internet coverage and DHVP	*–0.0825*	0.04	.04	136,252
Group 8: High internet coverage and DLVP	*–0.0491*	0.02	.001	715,317

[a]Robust SEs clustered by patients.

[b]Constant is included in each of the models.

[c]Control variables including community-level, hospital-level, patient-level, and admission-level are included in the models.

[d]Year and hospital fixed effects are included in the models.

[e]DHVP: diseases with high virtualization potential.

[f]Italicized values represent [insert reason].

[g]DLVP: diseases with low virtualization potential.

Discussion

Principal Findings

Overall, this study is among the first to comprehensively examine how telehealth mitigates health disparities associated with travel time to hospital for patients with recurrent admissions using a large, inpatient, longitudinal dataset in the United States. We make 2 main contributions. First, our study contributes to the health disparity literature by identifying the important role of telehealth in mitigating health disparities generated by distance barriers for patients with recurrent admissions. Previous literature pointed out that distance barriers are associated with decreased health outcomes [6,7] and called for studies to identify evidence for interventions that mitigate these adverse effects [11,51]. Our study answers this call and finds evidence that telehealth plays a significant role in bridging health disparities generated by distance barriers. Our results contribute to the literature by providing further evidence of the direct effect of drive distance on LOS, showing that the relationship is statistically significant and negative. One possible explanation for this phenomenon is that longer drive times might encourage using telehealth to better discharge planning and care coordination after discharge to avoid waste of resources [52], which reduces the duration of hospital stays. Furthermore, our result indicates that telehealth could be particularly effective in reducing in-hospital LOS for patients who live farther away from hospitals, potentially leading to more efficient use of hospital resources and better patient outcomes.

Second, our study examined the effects of telehealth on in-hospital LOS for patients with 2 or more admissions and various health conditions and diagnoses, thereby providing an in-depth understanding of the effects of telehealth. The result shows that telehealth adoption is not only associated with shortened in-hospital LOS but also reveals that the impact of telehealth is different for different patient groups. For example, telehealth has a larger impact on reducing in-hospital LOS for patients with complex health needs, needing frequent health services, and having DHVP. This may be due to telehealth facilitating better outpatient services [22,23] and medication management [53]. Such effects may be stronger for patients who frequently engage with health care systems and have complex health needs, potentially reducing their need for prolonged LOS further. Our results indicate telehealth could be effectively integrated into the care plans for those patient groups. Furthermore, the literature on how internet coverage influences the effects of telehealth on health outcomes is relatively sparse. Our study addresses this gap and finds that telehealth reduces in-hospital LOS for patients in high internet coverage areas and has no significant impact on those in low coverage areas. For patients living in high internet coverage areas, telehealth adoption is more effective for patients with DHVP. Future research should explore the underlying mechanisms of these findings.

The research also has practical implications. First, it informs health care practitioners and hospitals that telehealth can benefit patients who require regular monitoring and intervention and need more resources, such as patients with frequent care use and many comorbidities. Second, this research offers policy

XSL•FO
RenderX

makers evidence that telehealth can mitigate certain health disparities caused by distance barriers for patients with recurrent admissions. Third, our results highlight the importance of patients living in areas with high internet coverage in order to fully reap the benefits of telehealth. This is particularly important for rural residents, where internet services are limited. Several independent nonprofit research groups have repeatedly called for extending broadband access in rural communities to enable telehealth access [50,54,55]. Furthermore, the Federal Communications Commission has stated that broadband access is a "super determinant of health" [48], underscoring the importance of internet access and its impact on public health. Thus, we recommend decision makers consider the internet coverage factor when making telehealth implementation–related decisions. Fourth, our findings have practical implications for hospitals and policy makers in resource-limited countries that are beginning to adopt telehealth. For example, our data indicate that telehealth is more effective in reducing LOS for patients with frequent hospital visits, multiple comorbidities, those living far from medical facilities, and those with patients with DHVP. Hospitals serving such demographics are likely to benefit substantially from telehealth. Finally, the results of our study might encourage actionable process changes and software implementations for the providers leading to improved patient care, cost savings, and better resource use. We present a framework for such implementations and process changes in Figure 4. This would be particularly beneficial for providers with high rehospitalization rates and longer average LOS and patients with relatively long travel times to the health care facility. Some of the software and process changes we propose based on our findings in this research are to (1) incorporate additional data such as travel time and internet coverage data during admissions and patient intake process to the electronic health record system; (2) develop a data extraction module based on patient readmission rates, LOS, travel time, and internet coverage and apply analytics to identify patients with potential high telehealth impact; (3) develop promotional and incentivizing strategies to encourage enrollment and use of telehealth and integrate such strategies with patient reach out and communication channels; (4) integrate telehealth enrollment and use data with the electronic health record system (if not already integrated); and (5) develop analytics and predictive models for improved capacity planning, use, and cost savings.

Figure 4. Proposed process changes and software implementation framework. EHR: electronic health record; LOS: length of stay.

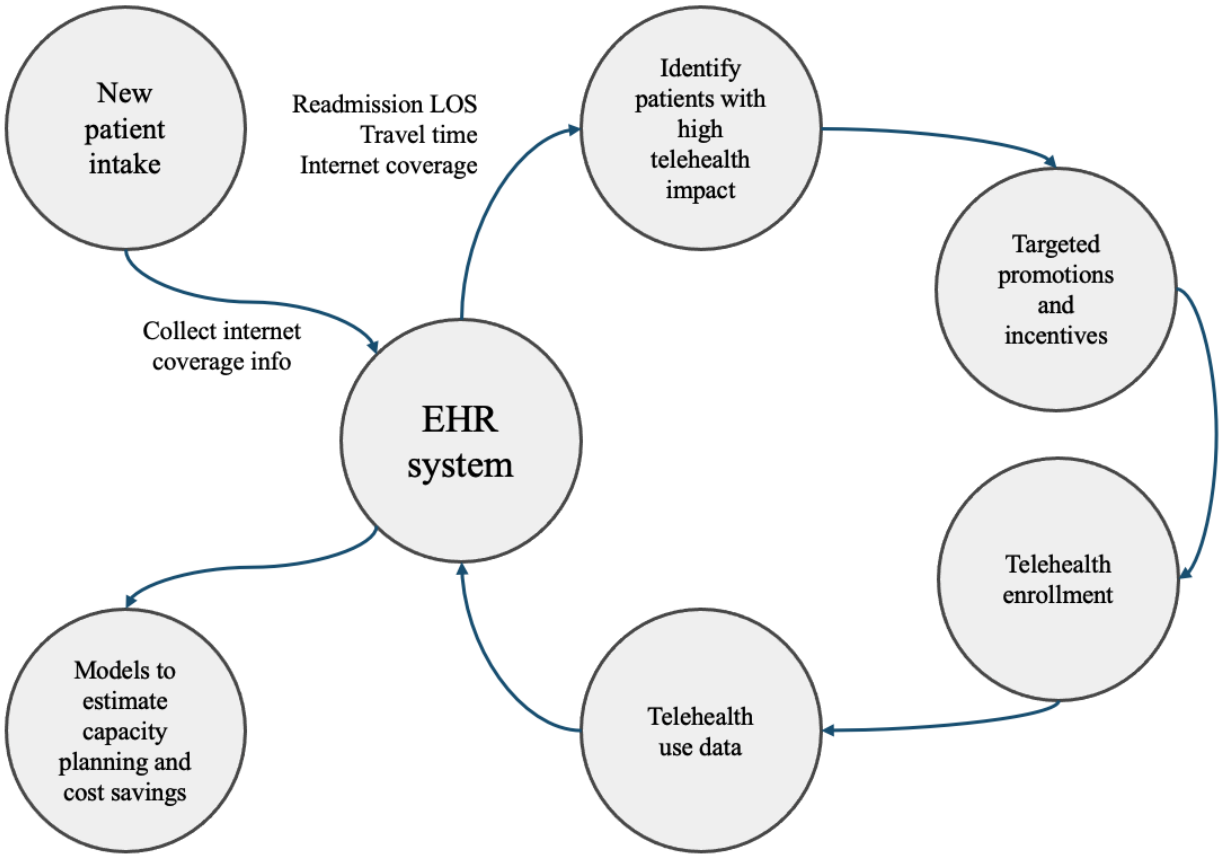

Limitations and Future Research

Our study has a few limitations that also present opportunities for future research. First, our study is limited to data from the states of New York and Florida from 2012 to 2015, when telehealth services were not as pervasive as today. Future research could investigate whether the effects of telehealth observed in this study are consistent across other US states with more recent longitudinal data. Second, due to the broad definition of telehealth in our dataset, it is challenging to pinpoint which specific telehealth interventions contribute to the observed outcomes. Future research should focus on

XSL•FO
RenderX

differentiating the effects of various telehealth modalities to better understand how they address health care delivery challenges, particularly those related to travel-time barriers. Finally, our analysis focuses on hospital-level telehealth adoption, which, while comprehensive, may be too broad to inform specific decisions for particular health care settings or hospital units that provide various treatments for specific diseases or conditions. Future studies could further examine the impact of telehealth at the departmental level or within particular medical contexts.

Conclusions

Using a large panel dataset, this study contributes to the HIT and health disparity literature, providing strong empirical evidence that telehealth may mitigate certain health disparities generated by travel time–related barriers, with stronger effects for patients who visit hospitals more frequently, have a higher number of comorbidities, and live in high internet coverage area. Furthermore, it offers insights on guiding health care practitioners to allocate more resources to improve their telehealth access and service quality, especially for hospitals serving a large demographic of patients who live farther away from the hospital.

Acknowledgments

The authors are grateful to Le Family for their support on this study. This work was partially funded by the Le Family Endowed Faculty Opportunity Fund at the College of Business Administration, Loyola Marymount University, California, United States. This fund had no influence, either direct or indirect, on the outcomes or direction of the research.

Data Availability

Due to restricted usage imposed by the American Hospital Association (AHA) data use agreement and the Healthcare Cost and Utilization Project (HCUP) data use agreement, access to the datasets used in this study is limited. The dataset applied in this study can be accessed with permission from the AHA [33] and Agency for Healthcare Research and Quality [32].

Authors' Contributions

YT contributed to conceptualization, data curation, methodology, formal analysis, visualization, writing—original draft, and writing—review and editing. AV contributed to conceptualization, methodology, visualization, writing—original draft, and writing—review and editing. DW contributed to conceptualization, methodology, writing—original draft, and writing—review and editing. JL contributed to conceptualization, methodology, and writing—review and editing. KS contributed to conceptualization, methodology, and writing—review and editing. ANM contributed to methodology, project administration, and writing—review and editing. AB contributed to methodology, visualization, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist of items for the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.
[DOCX File , 22 KB - jmir_v26i1e63661_app1.docx]

Multimedia Appendix 2
Health information technology items and coding.
[DOCX File , 17 KB - jmir_v26i1e63661_app2.docx]

Multimedia Appendix 3
Summary statistics for control variables and additional variables.
[DOCX File , 26 KB - jmir_v26i1e63661_app3.docx]

References

1. Quality of care. World Health Organization. 2024. URL: https://www.who.int/health-topics/quality-of-care#tab=tab_1 [accessed 2024-03-23]
2. Holmes L, Enwere M, Williams J, Ogundele B, Chavan P, Piccoli T, et al. Black-White risk differentials in COVID-19 (SARS-COV2) transmission, mortality and case fatality in the United States: translational epidemiologic perspective and challenges. Int J Environ Res Public Health 2020;17(12):4322 [FREE Full text] [doi: 10.3390/ijerph17124322] [Medline: 32560363]
3. Mohapatra S. Health inequity and health outcome: a causal linkage study of low and middle income countries. Qual Quant 2016 Aug 27;51(6):2475-2488. [doi: 10.1007/s11135-016-0404-4]

XSL•FO
RenderX

4. Secretary's Advisory Committee on National Health Promotion and Disease Prevention Objectives for 2020: phase I report: recommendations for the framework and format of Healthy People 2020. U.S. Department of Health and Human Services. 2020. URL: https://health.gov/our-work/national-health-initiatives/healthy-people/healthy-people-2020/secretarys-advisory-committee-2020 [accessed 2024-10-17]
5. Labban M, Chen CR, Frego N, Nguyen DD, Lipsitz SR, Reich AJ, et al. Disparities in travel-related barriers to accessing health care from the 2017 National Household Travel Survey. JAMA Netw Open 2023 Jul 03;6(7):e2325291 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.25291] [Medline: 37498602]
6. Kelly C, Hulme C, Farragher T, Clarke G. Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? A systematic review. BMJ Open 2016 Nov 24;6(11):e013059 [FREE Full text] [doi: 10.1136/bmjopen-2016-013059] [Medline: 27884848]
7. Målqvist M, Sohel N, Do TT, Eriksson L, Persson L. Distance decay in delivery care utilisation associated with neonatal mortality. A case referent study in northern Vietnam. BMC Public Health 2010 Dec 13;10(1):762. [doi: 10.1186/1471-2458-10-762] [Medline: 21144058]
8. Simpson RE, Wang CY, House MG, Zyromski NJ, Schmidt CM, Nakeeb A, et al. Travel distance affects rates and reasons for inpatient visits after pancreatectomy. HPB (Oxford) 2019 Jul;21(7):818-826 [FREE Full text] [doi: 10.1016/j.hpb.2018.10.013] [Medline: 30595461]
9. Jackson KL, Glasgow RE, Mone MC, Sheng X, Mulvihill SJ, Scaife CL. Does travel distance influence length of stay in elective pancreatic surgery? HPB (Oxford) 2014;16(6):543-549 [FREE Full text] [doi: 10.1111/hpb.12180] [Medline: 24245982]
10. Rana R, Gow J, Moloney C, King A, Keijzers G, Beccaria G, et al. Does distance to hospital affect emergency department presentations and hospital length of stay among chronic obstructive pulmonary disease patients? Intern Med J 2022;52(3):403-410. [doi: 10.1111/imj.15014] [Medline: 32786063]
11. Turrentine FE, Buckley PJ, Sohn MW, Williams MD. Travel time influences readmission risk: geospatial mapping of surgical readmissions. Am Surg 2017;83(6):573-582. [doi: 10.1177/000313481708300621] [Medline: 28637558]
12. What is telehealth? Health Resources & Services Administration. 2022. URL: https://www.hrsa.gov/rural-health/telehealth/what-is-telehealth [accessed 2024-03-27]
13. Hirko KA, Kerver JM, Ford S, Szafranski C, Beckett J, Kitchen C, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020;27(11):1816-1818 [FREE Full text] [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
14. O'Connor M, Asdornwised U, Dempsey M, Huffenberger A, Jost S, Flynn D, et al. Using telehealth to reduce all-cause 30-day hospital readmissions among heart failure patients receiving skilled home health services. Appl Clin Inform 2017 Dec 16;07(02):238-247. [doi: 10.4338/aci-2015-11-soa-0157]
15. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Hirani S, et al. Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. BMJ 2012 Jun 21;344:e3874. [doi: 10.1136/bmj.e3874] [Medline: 22723612]
16. Policy research perspectives. American Medical Association. 2022. URL: https://www.ama-assn.org/system/files/2022-prp-telehealth.pdf [accessed 2024-10-17]
17. Harris Y, Goldwater J, Lloyd S, Tutty M, Mishra V. Future of health: advancing the appropriate use of telehealth within medical specialties. American Medical Association. 2024. URL: https://www.ama-assn.org/system/files/future-health-telehealth-use.pdf [accessed 2024-10-17]
18. Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. PLoS One 2018;13(4):e0195901. [doi: 10.1371/journal.pone.0195901] [Medline: 29652932]
19. Singh R, Mathiassen L, Stachura ME, Astapova EV. Dynamic capabilities in home health: IT-enabled transformation of post-acute care. JAIS 2011;12(2):163-188. [doi: 10.17705/1jais.00257]
20. Oh JH, Zheng Z, Bardhan IR. Sooner or later? Health information technology, length of stay, and readmission risk. Production and Operations Management 2018 Nov 01;27(11):2038-2053. [doi: 10.1111/poms.12748]
21. Morguet AJ, Kühnelt P, Kallel A, Jaster M, Schultheiss HP. Impact of telemedical care and monitoring on morbidity in mild to moderate chronic heart failure. Cardiology 2008;111(2):134-139. [doi: 10.1159/000119701] [Medline: 18376125]
22. Sammour Y, Spertus JA, Austin BA, Magalski A, Gupta SK, Shatla I, et al. Outpatient management of heart failure during the COVID-19 pandemic after adoption of a telehealth model. JACC Heart Fail 2021 Dec;9(12):916-924. [doi: 10.1016/j.jchf.2021.07.003] [Medline: 34857175]
23. Lieneck C, Weaver E, Maryon T. Outpatient telehealth implementation in the United States during the COVID-19 global pandemic: a systematic review. Medicina (Kaunas) 2021 May 09;57(5):462 [FREE Full text] [doi: 10.3390/medicina57050462] [Medline: 34065050]
24. Scherr D, Kastner P, Kollmann A, Hallas A, Auer J, Krappinger H, et al. Effect of home-based telemonitoring using mobile phone technology on the outcome of heart failure patients after an episode of acute decompensation: randomized controlled trial. J Med Internet Res 2009 Aug 17;11(3):e34 [FREE Full text] [doi: 10.2196/jmir.1252] [Medline: 19687005]

XSL•FO
RenderX

25. Dendale P, de Keulenaer G, Troisfontaines P, Weytjens C, Mullens W, Elegeert I, et al. Effect of a telemonitoring-facilitated collaboration between general practitioner and heart failure clinic on mortality and rehospitalization rates in severe heart failure: the TEMA-HF 1 (Telemonitoring in the Management of Heart Failure) study. Eur J Heart Fail 2012 Mar;14(3):333-340 [FREE Full text] [doi: 10.1093/eurjhf/hfr144] [Medline: 22045925]
26. Ayabakan S, Bardhan IR, Zheng Z. Impact of telehealth and process virtualization on healthcare utilization. Information Systems Research 2024 Mar;35(1):45-65. [doi: 10.1287/isre.2023.1220]
27. Segrelles Calvo G, Gómez-Suárez C, Soriano J, Zamora E, Gónzalez-Gamarra A, González-Béjar M, et al. A home telehealth program for patients with severe COPD: the PROMETE study. Respir Med 2014 Mar;108(3):453-462 [FREE Full text] [doi: 10.1016/j.rmed.2013.12.003] [Medline: 24433744]
28. Rodriguez JA, Saadi A, Schwamm LH, Bates DW, Samal L. Disparities in telehealth use among California patients with limited English proficiency. Health Aff (Millwood) 2021 Mar;40(3):487-495. [doi: 10.1377/hlthaff.2020.00823] [Medline: 33646862]
29. Armaignac DL, Saxena A, Rubens M, Valle CA, Williams LMS, Veledar E, et al. Impact of telemedicine on mortality, length of stay, and cost among patients in progressive care units: experience from a large healthcare system. Crit Care Med 2018 May;46(5):728-735 [FREE Full text] [doi: 10.1097/CCM.0000000000002994] [Medline: 29384782]
30. Gutierrez J, Moeckli J, Holcombe A, O'Shea AM, Bailey G, Rewerts K, et al. Implementing a telehospitalist program between Veterans Health Administration hospitals: outcomes, acceptance, and barriers to implementation. J Hosp Med 2021 Mar;16(3):156-163. [doi: 10.12788/jhm.3570] [Medline: 33617436]
31. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
32. Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality. 2024. URL: https://www.ahrq.gov/data/hcup/index.html [accessed 2024-10-30]
33. AHA Annual Survey Database. American Hospital Association. 2024. URL: https://www.ahadata.com/aha-annual-survey-database [accessed 2024-04-30]
34. AHA Annual Survey IT Database. American Hospital Association. 2024. URL: https://www.ahadata.com/aha-healthcare-it-database [accessed 2024-07-01]
35. Lin YK, Lin M, Chen H. Do electronic health records affect quality of care? evidence from the HITECH Act. Information Systems Research 2019 Mar;30(1):306-318. [doi: 10.1287/isre.2018.0813]
36. DesRoches CM, Campbell EG, Vogeli C, Zheng J, Rao SR, Shields AE, et al. Electronic health records' limited successes suggest more targeted uses. Health Aff (Millwood) 2010 Apr;29(4):639-646. [doi: 10.1377/hlthaff.2009.1086] [Medline: 20368593]
37. Fact sheet: telehealth. American Hospital Association. 2019. URL: https://www.aha.org/factsheet/telehealth [accessed 2024-10-17]
38. John-Baptiste A, Naglie G, Tomlinson G, Alibhai SMH, Etchells E, Cheung A, et al. The effect of English language proficiency on length of stay and in-hospital mortality. J Gen Intern Med 2004 Mar;19(3):221-228. [doi: 10.1111/j.1525-1497.2004.21205.x] [Medline: 15009776]
39. van Parys J, Brown ZY. Broadband internet access and health outcomes: patient and provider responses in Medicare. International Journal of Industrial Organization 2024 Jul;95:103072. [doi: 10.1016/j.ijindorg.2024.103072]
40. Duplaga M. The association between Internet use and health-related outcomes in older adults and the elderly: a cross-sectional study. BMC Med Inform Decis Mak 2021 May 06;21(1):150 [FREE Full text] [doi: 10.1186/s12911-021-01500-2] [Medline: 33957909]
41. Healthcare Cost and Utilization Project data use agreement course. Agency for Healthcare Research and Quality. 2021. URL: https://hcup-us.ahrq.gov/DUA/dua_508/DUA508version.jsp [accessed 2024-10-17]
42. Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT press; 2010.
43. Hospital adjusted expenses per inpatient day. KFF. 2024. URL: https://www.kff.org/health-costs/state-indicator/expenses-per-inpatient-day/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D [accessed 2024-10-16]
44. Kim M, Holton M, Sweeting A, Koreshe E, McGeechan K, Miskovic-Wheatley J. Using health administrative data to model associations and predict hospital admissions and length of stay for people with eating disorders. BMC Psychiatry 2023 May 10;23(1):326. [doi: 10.1186/s12888-023-04688-x] [Medline: 37165320]
45. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-169. [doi: 10.1146/annurev.publhealth.23.100901.140546] [Medline: 11910059]
46. Li X, Wong W, Lamoureux EL, Wong TY. Are linear regression techniques appropriate for analysis when the dependent (outcome) variable is not normally distributed? Invest Ophthalmol Vis Sci 2012 May 01;53(6):3082-3083. [doi: 10.1167/iovs.12-9967] [Medline: 22618757]
47. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011 May;46(3):399-424. [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]

XSL•FO
RenderX

48. Advancing broadband connectivity as a social determinant of health. Federal Communications Commission. 2022. URL: https://www.fcc.gov/health/SDOH [accessed 2022-02-07]
49. Commissioner Clyburn remarks on mapping broadband health. Federal Communications Commission. 2016. URL: https://www.fcc.gov/document/commissioner-clyburn-remarks-mapping-broadband-health [accessed 2024-10-17]
50. "Spotty at Best": lack of rural broadband limits access to CARES Act telehealth services. American Telemedicine Association. 2020. URL: https://www.americantelemed.org/ata-news/spotty-at-best-lack-of-rural-broadband-limits-access-to-cares-act-telehealth-services/ [accessed 2024-10-17]
51. Chou S, Deily ME, Li S. Travel distance and health outcomes for scheduled surgery. Med Care 2014 Mar;52(3):250-257. [doi: 10.1097/MLR.0000000000000082] [Medline: 24374426]
52. Gu W, Fan N, Liao H. Evaluating readmission rates and discharge planning by analyzing the length-of-stay of patients. Ann Oper Res 2018 Jun 28;276(1-2):89-108. [doi: 10.1007/s10479-018-2957-1]
53. Badowski ME, Walker S, Bacchus S, Bartlett S, Chan J, Cochran KA, et al. Providing comprehensive medication management in telehealth. Pharmacotherapy 2018 Mar;38(2):e7-e16. [doi: 10.1002/phar.2071] [Medline: 29239004]
54. Barriers to telehealth in rural areas. Rural Health Information Hub. 2023. URL: https://www.ruralhealthinfo.org/toolkits/telehealth/1/barriers [accessed 2024-10-17]
55. AMA: equity must be baked into expansion of telehealth. American Medical Association. 2021. URL: https://www.ama-assn.org/press-center/press-releases/ama-equity-must-be-baked-expansion-telehealth [accessed 2021-06-21]

Abbreviations

AHA: American Hospital Association
DHVP: diseases with high virtualization potential
DLVP: diseases with low virtualization potential
HCUP: Healthcare Cost and Utilization Project
HCUP-SID: Healthcare Cost and Utilization Project's State Inpatient Datasets
HIT: health information technology
IT: information technology
LOS: length of stay
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by G Eysenbach, N Cahill; submitted 26.06.24; peer-reviewed by EP Greenup, B Kragen; comments to author 26.07.24; revised version received 19.08.24; accepted 24.09.24; published 25.11.24.

Please cite as:
Tao Y, Vo A, Wu D, Lin J, Seal K, Mishra A, Brahma A
Impact of Telehealth on Health Disparities Associated With Travel Time to Hospital for Patients With Recurrent Admissions: 4-Year Panel Data Analysis
J Med Internet Res 2024;26:e63661
URL: https://www.jmir.org/2024/1/e63661
doi:10.2196/63661
PMID:

©Youyou Tao, Ace Vo, Dezhi Wu, Junyuan Lin, Kala Seal, Abhay Mishra, Arindam Brahma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Benefits of Using Active Remote Patient Management for Enhanced Heart Failure Outcomes in Rural Cardiology Practice: Single-Site Retrospective Cohort Study

William Craig[1], MD; Suzanne Ohlmann[2], BSN, MFA

[1]Craig Cardiovascular Center, Seguin, TX, United States

[2]Craig Cardiovascular Center, San Antonio, TX, United States

Corresponding Author:
William Craig, MD
Craig Cardiovascular Center
1348 E. Walnut St
Seguin, TX, 78155
United States
Phone: 1 830 433 4648
Email: wcraig@cvdrs.com

Abstract

Background: Rural populations have a disproportionate burden of heart failure (HF) morbidity and mortality, associated with socioeconomic and racial inequities. Multiple randomized controlled trials of remote patient monitoring (RPM) using both direct patient contact and device-based monitoring have been conducted to assess improvement in HF outcomes, with mixed results.

Objective: We aimed to assess whether a novel digital health care platform designed to proactively assess and manage patients with HF improved patient outcomes by preventing HF re-exacerbations, thus reducing emergency room visits and HF hospitalizations.

Methods: This was a single-site, retrospective cohort study using electronic medical record (EMR) data gathered from 2 years prior to RPM initiation and 2 years afterward. In January 2017, this single center began enrolling New York Heart Association (NYHA) class II and class III patients with HF prone to HF exacerbation into an RPM program using the Cordella HF system. By July 2022, 93 total patients had been enrolled in RPM. Of these patients, 87% lived in rural areas. This retrospective review included 40 of the 93 patients enrolled in RPM. These 40 were selected because they had 2 years of established EMR data prior to initiation of RPM and 2 years of post-RPM data; each consented to this Sterling IRB–approved study.

Results: We included 40 patients with at least 4 years of follow-up, including 2 years prior to RPM initiation and 2 years after RPM initiation. In the 2 years after RPM initiation, check-up calls increased 519%, medication change calls increased 519%, and total calls increased by 519%. Emergency room visits for HF fell 93%, heart failure hospitalizations fell 83%, and all other cardiovascular hospitalizations fell 50%. Additionally, the total number of office visits declined by 15% after RPM, and unscheduled or urgent office visits declined by 73%.

Conclusions: Daily monitoring of trends in vital sign data between engaged patients and a collaborative team of clinicians, incorporated into daily clinical workflow, enhanced patient interactions and allowed timely response or intervention when HF decompensation occurred, resulting in a reduction of outpatient and inpatient clinical use over more than 2 years of follow-up.

(J Med Internet Res 2024;26:e49710) doi:10.2196/49710

KEYWORDS

rural; remote patient monitoring; heart failure; heart failure hospitalizations; office visits; rural health inequalities; telehealth

Introduction

Rural populations have a disproportionate burden of disease, particularly obesity, diabetes, smoking, hypertension, and coronary heart disease, all of which are associated with poor health outcomes [1,2]. This is specifically reflected in the fact that rural populations have worse outcomes for heart failure (HF) morbidity and mortality and is compounded by racial and socioeconomic inequities [3-5]. Rural patients also receive less outpatient cardiac management and are more likely to present to emergency care or be hospitalized in the first year of HF

XSL•FO
RenderX

diagnosis compared to urban patients [6]. Remote patient monitoring (RPM) of patients with HF to reduce HF hospitalizations and mortality has been studied in multiple randomized clinical trials, with mixed results [7-11]. These studies used a variety of RPM systems for HF management that included some combination of structured telephone calls and daily weight, blood pressure, heart rate, and symptom questionnaires. RPM trial meta-analyses have failed to show any consistent benefit from RPM when compared with the usual standard of care [12-16]. Many of the studies may not have used optimal monitoring equipment or were unable to incorporate appropriate, efficient clinical workflows into the process of RPM. Patient compliance may have also been negatively impacted by using automated telephone menus for patient data reporting rather than direct, clinician-to-patient communication for feedback, medication adjustment, and monitoring of patient acceptance and engagement in treatment modalities. Here, we report results of a single-center study on the use of a novel RPM system in a rural population with HF. We hypothesize that active patient management with RPM, when integrated into a standard cardiology workflow, enhances patient care, improves patient satisfaction, and reduces emergency room (ER) visits and HF hospitalizations.

Methods

Overview

This is a retrospective cohort review using electronic medical record (EMR) data from a group of New York Heart Association (NYHA) class II and III patients with HF who were followed over a 4-year period at a single institution, including 2 years prior to and 2 years following the initiation of a novel digital system to assess the impact of RPM on HF management.

Between January 2017 and July 2022, this single site began enrolling patients with HF who showed demonstrated difficulty maintaining clinical HF compensation into the Cordella HF System (Endotronix Inc) program. Over that period, a total of 93 patients were enrolled, 87% of whom lived in rural areas (defined as more than 1 hour from the city limits of an urban center with tertiary care).

Of the 93 total patients in this study, we included 40 who had both 2 years of pre-RPM EMR data established within the practice and 2 years of post-RPM EMR data. We excluded 53 of the 93 total patients who had been enrolled in Cordella due to the absence of sufficient data: either they lacked 2 years of established EMR data prior to joining the RPM program, or they had not accrued the full 2 years of data after initiation at the time of this review. Additionally, 3 patients in the excluded group dropped out of the RPM program prior to completing 2 years of enrollment, and 4 patients died prior to completion of 2 years of RPM. Of note, of the 4 deceased patients, 2 would not have had sufficient EMR data within the practice prior to enrolling in RPM.

The Cordella HF System is a novel, free-standing, digital health care platform designed specifically to proactively manage outpatients with HF and the multiple factors that lead to HF re-exacerbation and readmission. The system uses a wireless/cellular tablet computer to receive data from multiple Bluetooth-enabled peripherals to assess vital signs. The tablet computer, weight scale, blood pressure cuff, and pulse oximeter were provided to patients for home use with online and telephone technical support. In this mostly rural population, even older patients with limited technical experience were able to follow telephone instructions to set up the equipment in their homes. Patients were instructed to use these devices daily and to answer questions on their current clinical status via their tablets. The data were then securely transmitted to an online clinician portal and reviewed remotely from the physician's office as either tabular or trend data (Figure 1).

XSL•FO
RenderX

Figure 1. Examples of 2 clinical case applications of the Cordella Remote Patient Monitoring System in outpatients with heart failure at a single heart failure clinic lasting 2 years. (A) Vital signs over time (blue horizontal lines) and heart failure medication changes (orange vertical lines) for a case in which the system was used to lower heart rate. (B) Vital signs over time and diuretic medication changes for a case that used the system to reduce elevated weight indicative of edema. For a higher-resolution version of this figure, see Multimedia Appendix 1.

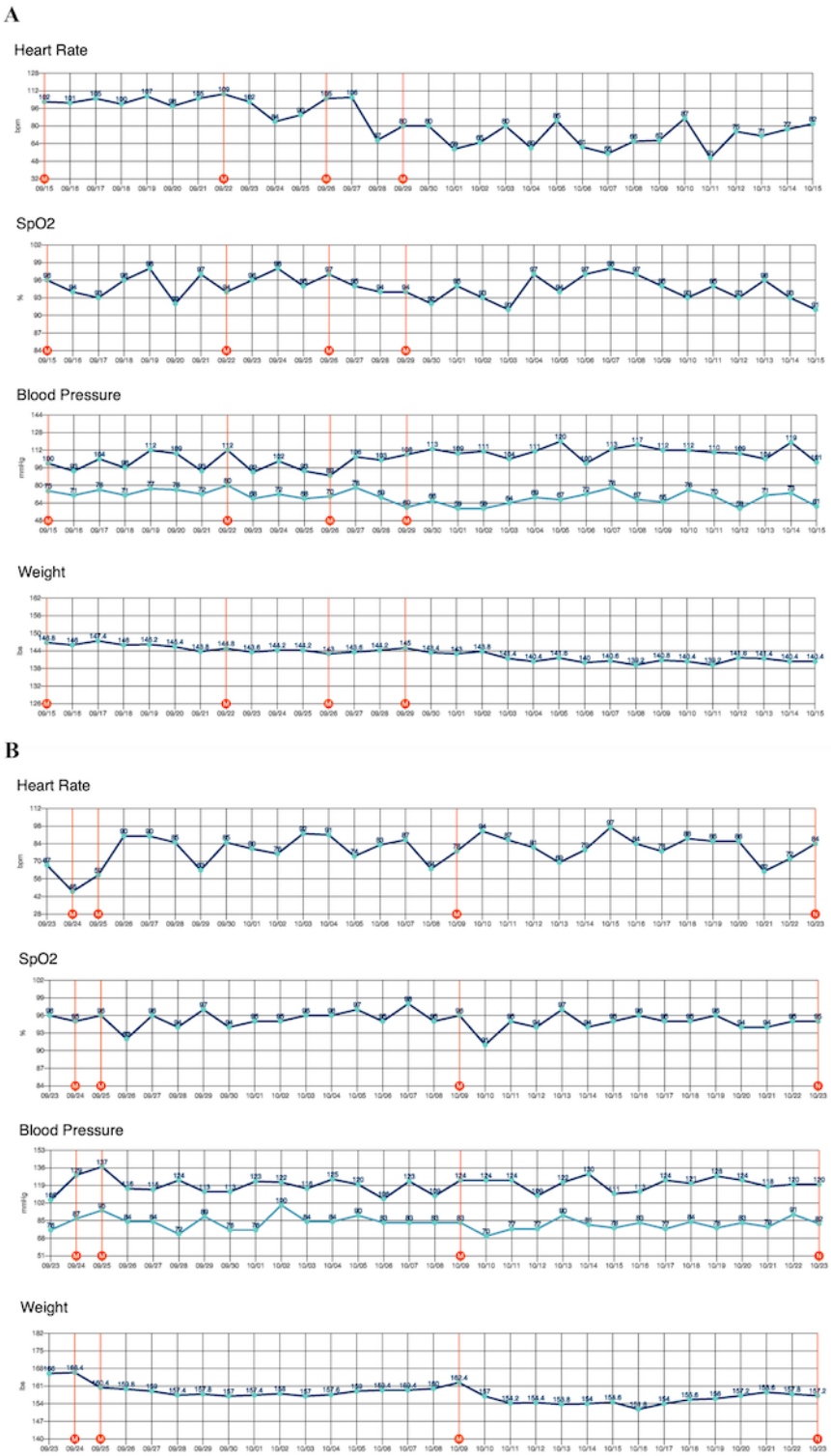

Importantly, the workflow for daily RPM for these patients was incorporated into usual clinical practices. The same interdisciplinary team reviewed the daily trends for each patient as part of an integrated workflow in the rural practice. A full-time clinic nurse was dedicated to assisting review of the patients' data and interacting with the patients by telephone and the tablet messaging system as often as necessary. The cardiologist and nurse discussed problems identified in the daily trends, and the nurse then contacted patients if additional information was needed. When a lifestyle or medication intervention became necessary, the team contacted the patient to make that change. The patient's response was then monitored on subsequent daily RPM readings with further adjustments made as needed.

XSL•FO
RenderX

All discussions between the cardiologist and the nurse were documented within the Cordella Patient Management Portal (PMP). Telephone calls to patients to discuss their status and recommend medication and lifestyle modifications were also documented in the PMP. Medication changes were annotated on the trend plots to mark the response. Permanent medication changes were documented in the patient's EMR, as the PMP is not currently integrated with the EMR. The contact time spent in daily review of patient data, discussion of changes noted in data trends, and interaction with the patient for further information or medication change was synchronously captured by the PMP. All patients were contacted at least monthly to discuss their clinical status, note visits with their other physicians, document ER visits or hospitalizations, and ensure that no changes had been made to pertinent cardiac medications.

All cardiovascular hospitalizations, ER visits, and HF hospitalizations were assessed in the 2-year period prior to initiation of RPM and the 2-year period following RPM initiation. Similarly, the total number of telephone calls made monthly to patients by the clinical team, noting the reason for the calls, was assessed in the 2-year period prior to initiation of RPM and the 2-year period following RPM initiation, as well as any unplanned clinic visits. Data on each end point were compared for the period before and after the initiation of RPM using Wilcoxon signed-rank tests. Patient compliance, defined as data transmission to the portal 5 of 7 days per week, and clinician compliance, defined as clinicians checking the PMP daily, were also assessed.

Ethical Considerations

This study was a retrospective cohort analysis of electronic patient data accrued over a 4-year period. The study received institutional review board approval (Sterling IRB; 8276-WECraig), and each participant signed informed consent documentation prior to having their data collected for daily Cordella review, as well as for the purposes of this study. Participant privacy and anonymity were achieved through elimination of any patient identifiers such as name or dates of birth, which were replaced by an anonymous, numeric system for data tabulation and analyses. The participants did not receive compensation for their participation in the study.

Results

This study included 40 patients with at least 4 years of follow-up, including 2 years prior to RPM initiation and 2 years following RPM initiation. The patients had a mean age of 76.9 (SD 9.5) years. A majority were female (n=23, 58%) and had a high comorbidity burden, including 40% (n=16) with chronic kidney disease, 55% (n=22) with atrial fibrillation, 15% (n=6) with aortic stenosis, and 68% (n=27) with mitral regurgitation. Most patients were NYHA class III (n=32, 80%) and 42% (n=17) had an ejection fraction ≥50%.

In the 2 years prior to RPM initiation, the average number of check-up calls per month to all 40 patients was 7.5 (SD 0.1), and the average number of medication calls per month to all 40 patients was 4.3 (SD 0.2), with a total average number of calls per month numbering 11.8 (SD 0.3). In the 2 years that followed RPM initiation, check-up calls increased 519% (to 46.4, SD 0.8 per month; $P<.001$), medication change calls increased 519% (to 26.6, SD 0.7 per month; $P<.001$), and total calls increased 519% (to 73, SD 1.3 per month; $P<.001$).

In the 2 years prior to RPM initiation, there were 48 unplanned office visits, 15 ER visits for HF, and 36 HF hospitalizations; all other cardiovascular hospitalizations numbered 38. In the 2 years that followed RPM initiation, ER visits for HF fell 93% (to 1 visit; $P=.006$), HF hospitalizations fell 83% (to 6 visits; $P=.003$), all other cardiovascular hospitalizations fell 50% (to 19 visits; $P=.06$), and unplanned office visits fell 73% (to 13 visits, $P=.001$) (Table 1). Patient compliance was 88% at 6 months, 90% at 1 year, and 95% at 2 years. Clinician compliance was 100% throughout the study period.

Table 1. Monthly calls and total visits in the 24 months before remote patient monitoring (RPM) and 24 months after RPM.

Call/visit type	24 months before RPM	24 months after RPM	Percentage change	*P* value
Monthly calls, mean (SD)				
Medication change	4.3 (0.2)	26.6 (0.7)	+519%	<.001
Check-up	7.5 (0.1	46.4 (0.8)	+519%	<.001
Total	11.8 (0.3)	73 (1.3)	+519%	<.001
Total visits, n				
Unplanned office visit	48	13	–73%	.001
ER[a] visit for HF[b]	15	1	–93%	.006
HF hospitalization	36	6	–83%	.003
All other CV[c] visits	38	19	–50%	.06

[a]ER: emergency room.

[b]HF: heart failure.

[c]CV: cardiovascular.

XSL•FO
RenderX

Discussion

Principal Findings

This retrospective study shows a dramatic decrease in hospitalizations, ER visits, and urgent office visits after active management with daily RPM was integrated into standard cardiology care delivery. This management approach used daily RPM in conjunction with frequent patient contact and timely adjustment of medication and lifestyle issues. Management of patients with advanced HF is a dynamic process. Each patient's unique comorbidities, diet, and lifestyle contribute to decompensations between regularly scheduled office visits. Swift identification and treatment of the precursors to decompensation, a hallmark of this study's success, significantly reduced hospitalizations, ER visits, and urgent office visits.

This study identified multiple factors that contributed to the successful reduction in HF decompensations. Most relevant, daily RPM using this model facilitates frequent clinician-patient interaction and care adjustments. Monitoring trend plot flow sheets combined with patient symptom reporting reduces clinician apprehension when making care adjustments remotely.

Patient response to medication changes was assessed quickly via vital sign trends, telephone calls, and further medication adjustment, decreasing the need for in-office follow-up.

Equally important are the challenges of patient engagement and education, both of which contribute to the success of decreasing HF decompensations. Frequent interaction with the cardiologist and nurse team facilitated numerous discussions of fluid and sodium intake, better understanding of the complex nature of HF disease and treatment and the benefits of monitoring data trends through the shared portal, and increased adherence to the current medication regimen.

Close monitoring with RPM and frequent patient interactions within this rural setting, reflected by the increase in telephone calls, improved patients' clinical well-being and was associated with a dramatic reduction in ER visits, HF hospitalizations, and unplanned office visits.

The Cordella HF system allowed patients to observe and track their own health data, provided the foundation for patient education, and facilitated shared decision-making with the clinical team. This connection between clinician and patient via shared data with frequent, personalized clinical interventions facilitated patient education and led to the observed increase in patient reporting and compliance over time.

The role of a dedicated nurse for RPM is an equally important factor in the success of this study. The HF nurse monitors the daily patient data and calls patients to assess symptoms and explain medication changes. The role evolved to include coaching for support, education, and feedback regarding the multiple factors involved in HF management. Over time, the HF nurse mediated in both clinical and lifestyle issues pertinent to HF management, providing both encouragement and intervention when needed and reassurance when the patient was stable. While this study included 40 patients, our practice now uses this system for more than 70 patients with no further disruption in clinic workflow and no additional staff, a positive indicator of the scalability of this technology.

The workflow used for RPM in the study was easily incorporated into a standard outpatient cardiology practice. The managing cardiologist maintained a busy daily clinic schedule and hospital rounds. The dedicated HF nurse was employed by the practice, and for patients eligible for chronic care management (CCM), CCM billing not only covered the cost of the Cordella HF system and the dedicated nurse but was also cost-effective, generating revenue for the practice.

Limitations and Future Research

The major limitation of the study is its retrospective nature. However, all patients with a full 24 months of RPM were included. Those patients who had enrolled for monitoring but dropped out prior to completing 24 months, or who had not completed 24 months of monitoring at the time of data analysis, were excluded. A second of the study's limitations lies in the intense nature of monitoring and frequent interaction with patients. Due to the relatively small cohort of patients, this intensive effort was feasible within the rural cardiology practice, which mirrors many of those delivering care in the rural setting. It remains to be seen if this same workflow could be expanded to a larger clinical population with HF.

Further exploration into reimbursement is merited—for both Medicare and non-Medicare patients—as well as a cost-benefit analysis to include projected savings due to the decrease in ER visits and HF hospitalizations. This workflow also offers the opportunity to explore specific combinations of data and alerts that can improve early identification of changes in status. This information may also enhance HF management by expanding the multidisciplinary team caring for these patients.

Conclusion

Active HF management with RPM and a focused multidisciplinary team reduced ER visits for HF by 93%, HF hospitalizations by 83%, and unplanned office visits by 73% in this retrospective study. Daily monitoring of trends in vital sign data between engaged patients and clinicians, when incorporated into the daily clinical workflow, allowed for immediate response to the precursors of HF decompensation and intervention. The success of this study's proactive model of HF management using the Cordella HF system stands in great contrast to the current reactive model of HF urgent office visits, ER visits, and HF hospitalizations.

Data Availability

All data generated or analyzed during this study are included in this published article.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of 2 clinical case applications of the Cordella Remote Patient Monitoring System in outpatients with heart failure at a single heart failure clinic lasting 2 years. (A) Vital signs over time (blue horizontal lines) and heart failure medication changes (orange vertical lines) for a case in which the system was used to lower heart rate. (B) Vital signs over time and diuretic medication changes for a case that used the system to reduce elevated weight indicative of edema.
[PNG File , 5019 KB - jmir_v26i1e49710_app1.png]

References

1. Michael MA, Alana KM, Gilbert T. The 2014 update of the rural-urban chartbook. Rural Health Research Gateway. 2014. URL: https://www.ruralhealthresearch.org/ [accessed 2022-02-01]
2. Harrington RA, Califf RM, Balamurugan A, Brown N, Benjamin RM, Braund WE, et al. Call to action: rural health: a presidential advisory from the American Heart Association and American Stroke Association. Circulation 2020 Mar 10;141(10):e615-e644 [FREE Full text] [doi: 10.1161/CIR.0000000000000753] [Medline: 32078375]
3. Howard G, Kleindorfer DO, Cushman M, Long DL, Jasne A, Judd SE, et al. Contributors to the excess stroke mortality in rural areas in the United States. Stroke 2017 Jul;48(7):1773-1778 [FREE Full text] [doi: 10.1161/STROKEAHA.117.017089] [Medline: 28626048]
4. Kamin Mukaz D, Dawson E, Howard VJ, Cushman M, Higginbotham JC, Judd SE, et al. Rural/urban differences in the prevalence of stroke risk factors: A cross-sectional analysis from the REGARDS study. J Rural Health 2022 Jun;38(3):668-673 [FREE Full text] [doi: 10.1111/jrh.12608] [Medline: 34270125]
5. Turecamo SE, Xu M, Dixon D, Powell-Wiley TM, Mumma MT, Joo J, et al. Association of rurality with risk of heart failure. JAMA Cardiology 2023 Mar 01;8(3):231-239 [FREE Full text] [doi: 10.1001/jamacardio.2022.5211] [Medline: 36696094]
6. Gamble J, Eurich DT, Ezekowitz JA, Kaul P, Quan H, McAlister FA. Patterns of care and outcomes differ for urban versus rural patients with newly diagnosed heart failure, even in a universal healthcare system. Circ Heart Fail 2011 May;4(3):317-323. [doi: 10.1161/CIRCHEARTFAILURE.110.959262] [Medline: 21430285]
7. Koehler F, Winkler S, Schieber M, Sechtem U, Stangl K, Böhm M, Telemedical Interventional Monitoring in Heart Failure Investigators. Impact of remote telemedical management on mortality and hospitalizations in ambulatory patients with chronic heart failure: the telemedical interventional monitoring in heart failure study. Circulation 2011 May 03;123(17):1873-1880. [doi: 10.1161/CIRCULATIONAHA.111.018473] [Medline: 21444883]
8. Koehler J, Stengel A, Hofmann T, Wegscheider K, Koehler K, Sehner S, et al. Telemonitoring in patients with chronic heart failure and moderate depressed symptoms: results of the Telemedical Interventional Monitoring in Heart Failure (TIM-HF) study. Eur J Heart Fail 2021 Jan;23(1):186-194 [FREE Full text] [doi: 10.1002/ejhf.2025] [Medline: 33063412]
9. Tütüncü S, Honold M, Koehler K, Deckwart O, Koehler F, Haeusler KG. Non-invasive telemedical care in heart failure patients and stroke: post hoc analysis of TIM-HF and TIM-HF2 trials. ESC Heart Fail 2020 Jun;7(3):884-891 [FREE Full text] [doi: 10.1002/ehf2.12679] [Medline: 32163672]
10. Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, et al. Telemonitoring in patients with heart failure. N Engl J Med 2010 Dec 09;363(24):2301-2309 [FREE Full text] [doi: 10.1056/NEJMoa1010029] [Medline: 21080835]
11. Ong MK, Romano PS, Edgington S, Aronow HU, Auerbach AD, Black JT, Better Effectiveness After Transition–Heart Failure (BEAT-HF) Research Group. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the Better Effectiveness After Transition -- Heart Failure (BEAT-HF) randomized clinical trial. JAMA Intern Med 2016 Mar;176(3):310-318 [FREE Full text] [doi: 10.1001/jamainternmed.2015.7712] [Medline: 26857383]
12. Pandor A, Thokala P, Gomersall T, Baalbaki H, Stevens JW, Wang J, et al. Home telemonitoring or structured telephone support programmes after recent discharge in patients with heart failure: systematic review and economic evaluation. Health Technol Assess 2013 Aug;17(32):1-207, v [FREE Full text] [doi: 10.3310/hta17320] [Medline: 23927840]
13. Gorst SL, Armitage CJ, Brownsell S, Hawley MS. Home telehealth uptake and continued use among heart failure and chronic obstructive pulmonary disease patients: a systematic review. Ann Behav Med 2014 Dec;48(3):323-336 [FREE Full text] [doi: 10.1007/s12160-014-9607-x] [Medline: 24763972]
14. Purcell R, McInnes S, Halcomb EJ. Telemonitoring can assist in managing cardiovascular disease in primary care: a systematic review of systematic reviews. BMC Fam Pract 2014 Mar 07;15:43 [FREE Full text] [doi: 10.1186/1471-2296-15-43] [Medline: 24606887]
15. Kitsiou S, Paré G, Jaana M. Effects of home telemonitoring interventions on patients with chronic heart failure: an overview of systematic reviews. J Med Internet Res 2015 Mar 12;17(3):e63 [FREE Full text] [doi: 10.2196/jmir.4174] [Medline: 25768664]

XSL•FO
RenderX

16. Sharif F, Rosenkranz S, Bartunek J, Kempf T, Assmus B, Mahon NG, et al. Safety and efficacy of a wireless pulmonary artery pressure sensor: primary endpoint results of the SIRONA 2 clinical trial. ESC Heart Fail 2022 Oct;9(5):2862-2872 [FREE Full text] [doi: 10.1002/ehf2.14006] [Medline: 35686479]

Abbreviations

CCM: chronic care management
CV: cardiovascular
EMR: electronic medical record
ER: emergency room
HF: heart failure
IRB: institutional review board
NYHA: New York Heart Association
PMP: Patient Management Portal
RPM: remote patient monitoring

Edited by A Mavragani; submitted 06.06.23; peer-reviewed by J Sharp, JB Park; comments to author 25.10.23; revised version received 21.01.24; accepted 19.09.24; published 26.11.24.

Please cite as:
Craig W, Ohlmann S
The Benefits of Using Active Remote Patient Management for Enhanced Heart Failure Outcomes in Rural Cardiology Practice: Single-Site Retrospective Cohort Study
J Med Internet Res 2024;26:e49710
URL: https://www.jmir.org/2024/1/e49710
doi:10.2196/49710
PMID:39589775

©William Craig, Suzanne Ohlmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Preferences of Individuals With Obesity for Online Medical Consultation in Different Demand Scenarios: Discrete Choice Experiments

Yaolin Hu[1], PhD; Jian Wang[2], Prof Dr; Jianbo Zhou[1], Prof Dr; Yuanyuan Gu[3], Prof Dr; Stephen Nicholas[4,5], Prof Dr; Elizabeth Maitland[6], Prof Dr

[1]School of Economics, Peking University, Beijing, China

[2]Dong Fureng Institute of Economic and Social Development, Wuhan University, Wuhan, China

[3]Macquarie University Centre for the Health Economy, Macquarie Business School & Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

[4]Health Services Research and Workforce Innovation Centre, Newcastle Business School, University of Newcastle, Newcastle, Australia

[5]Australian National Institute of Management and Commerce, Sydney, Australia

[6]School of Management, University of Liverpool, Liverpool, United Kingdom

Corresponding Author:
Yaolin Hu, PhD
School of Economics
Peking University
No. 5 Yiheyuan Road
Haidian District
Beijing, 100871
China
Phone: 86 13522596266
Email: huyaolin@pku.edu.cn

Abstract

Background: Obesity is a unique chronic disease, with China having the largest number of people living with overweight and obesity in the world. There has been little research from the demand perspective for online medical consultation (OMC) by individuals living with obesity. With the growing demand for obesity OMC, especially due to the emergence of new pharmacotherapies, such as glucagon-like peptide-1 receptor agonists, individuals living with obesity are seeking both advice on obesity management and the prescription of obesity drugs. Therefore, our demand scenarios defined 2 OMC motivations to manage obesity: “For-Drugs” use and “For-Advice” use.

Objective: This study aims to assess and compare the preferences for For-Drugs and For-Advice OMC among individuals living with obesity in China.

Methods: Following the International Society for Pharmacoeconomics and Outcomes Research’s checklist and comprising 400 participants assigned to the For-Drugs scenario and 400 to the For-Advice scenario, the For-Drugs and For-Advice preferences were estimated through discrete choice experiments. The groups in the 2 scenarios followed a similar distribution, and the 2 different demand scenarios shared the same discrete choice experiment design, comprising 16 choice sets with 6 representative attributes. Mixed logit modeling was used to estimate the willingness to pay and relative importance scores.

Results: Doctors with well-known and general expert titles, versus ordinary doctors; doctors from high-level, provincial, tertiary, and municipal hospitals, versus lower-level county hospitals; less waiting time; and lower OMC fees were preferred in both the For-Drugs and For-Advice scenarios. The differences between the 2 scenarios lay in the consultation format, consultation duration, and the relative importance of consultation duration versus waiting time. The For-Advice group preferred telephone consultations, while the For-Drugs group did not; the For-Drugs group preferred longer consultation duration (β=.029), while the For-Advice group preferred shorter consultation duration (β=–.030); and the For-Drugs group rated consultation duration higher than waiting time, while the For-Advice group rated the waiting time as more important than consultation duration. Combined with our qualitative research, the differences can be explained by the different consultation needs in the 2 scenarios, where longer patient consultations were preferred by the For-Drugs patients who sought detailed advice on drug side effects, while quick and direct responses were preferred by the For-Advice participants.

XSL•FO
RenderX

Conclusions: By revealing user preferences on costs, doctors' titles and hospital level, wait time, and consultation duration and format, our research informs OMC platforms, OMC regulators, and doctors on market segmentation and service differentiation strategies.

(J Med Internet Res 2024;26:e53140) doi:10.2196/53140

KEYWORDS

online medical consultation; obesity; discrete choice experiments; telehealth; telemedicine

Introduction

Seeking Web-Based Obesity Consultation

Obesity is a global public health problem [1], and China has the largest number of people living with overweight and obesity in the world [2]. With individuals living with overweight and obesity increasingly seeking medical support, there is an urgent need to better understand and improve obesity management in China, including the adoption of electronic health technologies [3]. In China's telemedicine industries, online medical consultation (OMC) is one of the most recognized and accepted web-based medical services. Promoted during the COVID-19 pandemic, OMC offers private and confidential communication, which can help relieve stigma and other stress burdens on individuals living with obesity and offer increased obesity treatment adherence relative to offline consultations [4,5]. OMC also reduces offline medical consultation costs, such as transportation and forgone work time costs [6].

OMC is also popular for its convenient prescriptions [7], a feature especially appealing to individuals living with obesity. Not only is obesity a chronic disease, but it is also correlated with comorbidities, such as diabetes, heart disease, hypertension, and cancer [8], which means individuals living with obesity have a high demand for multiple long-term medications. OMC prescriptions offer significantly time-saving experiences for individuals living with overweight or obesity who previously could only refill their medicines at hospitals. In China, OMC regulations allow patients to buy medications through web-based platforms after providing evidence of a prescription.

Recently, web-based obesity medicine prescriptions have gained a new level of popularity in China. Promising findings of glucagon-like peptide-1 receptor agonists, such as semaglutide and beinaglutide [9,10], have spurred many individuals living with obesity to adopt new pharmacotherapies, leading this group to increasingly access OMC to get prescriptions for weight-loss drugs. This new demand by individuals living with obesity has driven some Chinese OMC providers to channel patients into different consultation modes, such as "For-Advice" and "For-Drugs." The demand for the latter is so large that some OMC platforms have become increasingly dependent on digital medication sales [11,12].

In this study, we assess the preferences for OMC among Chinese adults living with obesity. Given there are 2 different motivations for OMC, we set out 2 distinct demand scenarios—"For-Advice" and "For-Drugs"; designed 2 discrete choice experiments (DCEs) to elicit consumer preferences under each scenario; and compared the "For-Advice" and "For-Drugs" groups [3]. This study provides new insights into the future development of OMC for telehealth providers, doctors, and regulators [12].

Research Motivation

Previous studies on China's OMC failed to distinguish between different diseases, with any results subject to selection bias [13,14], and failed to distinguish between different types of demand, such as For-Advice and For-Drugs [15]. By focusing on the needs of individuals living with overweight and obesity for OMC, more detailed findings from the demand perspective can be revealed. Focusing on the preferences of individuals with obesity for OMC also avoids disease selection bias and provides disease-specific advice for medical industry organizations and public health regulators [16,17].

Methods

Ethical Considerations

The experiments were ethically reviewed, and the study was approved by the Wuhan University, Faculty of Dong Fureng Institute of Economic and Social Development, Research Ethics Committee (dfr202201). The study was conducted according to the principles established by the Declaration of Helsinki, and data were anonymized. Participants were provided with information regarding the study and asked to sign an informed consent form before taking part. Participants who completed the experiment received compensation.

Definition of Different Demand Scenarios

We used DCE, a widely used method in health economics research, to elicit the preferences of individuals living with obesity [18,19]. With the help of the community workers from the community health centers in Wuhan, we undertook a 1-hour focus group discussion with 3 individuals living with overweight or obesity in July 2021 by using nonprobability convenience sampling. Facilitated by the corresponding author, the 3 interviewees in their 20s, 40s, and 50s discussed their obesity management aims, OMC experience, OMC costs, and OMC versus offline treatments. The focus group discussion confirmed our hypothesis of 2 different OMC demand scenarios (For-Drugs and For-Advice), which had been neglected by previous research.

After the focus group discussions, we consulted 6 obesity experts on OMC, obesity treatments, and DCE design through telecommunication in July 2021, comprising 2 industry experts in obesity treatment, 2 industry experts in OMC, and 2 academic professors focused on telemedicine.

Finally, we conducted 2 pilot surveys in Wuhan from July 2021 to August 2021. With the help of Wuhan's community health

XSL•FO
RenderX

workers, we selected individuals living with obesity who visited the health center by using nonprobability convenience sampling. First, 20 individuals living with overweight or obesity assessed the accuracy and appropriateness of the questionnaires and scenarios. After receiving the feedback from the first pilot survey, we modified the questionnaires and undertook the second pilot survey with another 20 individuals living with overweight or obesity as a final check.

The "For-Drugs" OMC demand scenario represents the case where OMC is for a prescription for specific weight-loss drugs or other drugs for obesity-related comorbidities. Based on the results of our focus group study and pilot surveys, this demand scenario was confirmed for 3 reasons. First, the purchase of medicines is one of the most demanded services in outpatient hospitals. Second, patients usually incur offline transportation, consultation waiting time, and lost work income costs to update prescriptions in hospitals. Third, most OMC platforms in China provide pharmaceutical e-commerce services, with some platforms mainly providing pharmaceutical e-commerce for drug sales as their profit motive rather than consultation services. For adults living with overweight or obesity, high-frequent medication demand via OMC is mainly supported by 2 motivations. The first motivation is to lose weight where doctors' prescriptions are required. The second motivation is to acquire long-term medications to treat other obesity-related chronic diseases, such as hypertension, diabetes, and hyperlipidemia. Complying with the legal requirements for web-based drug purchasing in China, the final description of the "For-Drugs" demand scenario is shown in Multimedia Appendix 1.

The second "For-Advice" demand scenario represents the case where OMC is for medical advice only. One of the reasons why the Chinese government promotes OMC has been to allocate scarce medical resources more optimally and allow people living remotely to access high-quality medical services. Further, losing weight is a long process, requiring frequent consultation on health management, diet, and exercise [20], which OMC can provide at a low cost. OMC also provides a high level of privacy, helping to overcome psychological barriers and poor treatment compliance for patients living with obesity. For these reasons, seeking medical advice is a major reason for using OMC. The final description of the "For-Advice" demand scenario is shown in Multimedia Appendix 2.

The final DCE surveys were conducted via a web-based platform, mainly based on 2 reasons. First, research suggests that in some situations, well-designed digital DCEs can better elicit respondents' preferences compared with offline experiments [21,22]. Second, during our offline pilot surveys, we found that respondents living with overweight or obesity felt pressure and stigma to talk freely about their true preferences.

As shown in Figure 1, we assigned respondents to group 1 ("For-Drugs" scenario) and group 2 ("For-Advice" scenario). The assignment of every respondent to different scenarios followed the same sample frame to ensure the 2 groups of respondents followed the same distribution characteristics of the obesity epidemic population in China. Respondents were not allowed to respond to both demand scenarios [23,24]. In our pilot surveys, respondents felt completing 2 DCEs under 2 scenarios were overwhelming and often got confused when they needed to imagine they were in 2 sets of scenarios [25].

Figure 1. Flowchart of our web-based survey. DCE: discrete choice experiment.

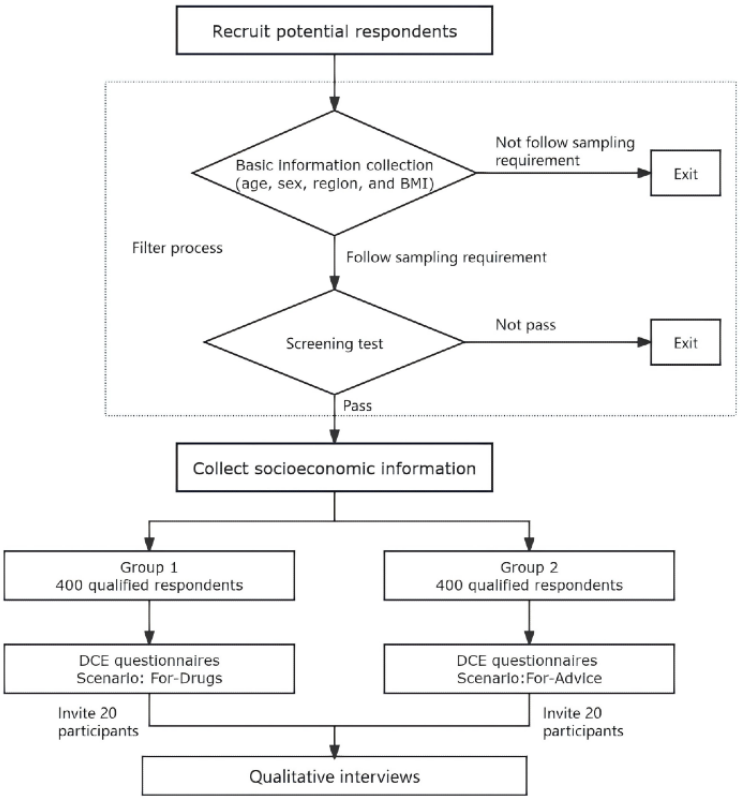

XSL•FO
RenderX

Sample Size and Data Collection

Obesity treatment involves both individuals living with overweight or obesity. We used the standards for adults living with overweight or obesity defined by China's National Health Commission, with adults living with overweight having 24≤BMI<28 and adults living with obesity having BMI≥28 [26]. Based on the prevalent rate of overweight and obesity in China [27-29] and the sample size calculation formula [30]:

where n is the sample size; Z is the Z statistic for a level of confidence (1.96 for 95% CI); P is the expected prevalence; and d is precision, we set the precision at 0.1 P, which yielded a minimum national sample size of 400 individuals living with overweight or obesity.

We used quota sampling to recruit participants, with the prevalent rates of age, sex, and regions of participants defining the quotas in our survey [31]. The proportion of overweight and obesity in Chinese adults was derived from the National Report on Nutrition and Chronic Diseases in Chinese Residents [27] while the survey of 441 thousand adults by Zhang et al [28] provided the sex, age, and regional distribution parameters for adults living with overweight or obesity. To ensure the representativeness and quality of our sample, we designed a filter process before the experiments to recruit 800 qualified participants.

The filter process had 2 steps (Figure 1). The first step was to collect information on respondents' age, sex, region, and BMI to ensure the participants recruited would follow the epidemiological distribution characteristics of obesity in China. BMI is a vital variable in our survey. In the pilot survey, we found that many potential participants did not know the meaning of BMI but knew their height and weight. We therefore collected participants' height and weight information, automatically calculating participants' BMI through our web-based experiment platform. Contracting with a commercial medical research firm with extensive contacts with doctors and hospitals across China, 2000 potential participants meeting the first filtering step were recruited. By email, phone, and in-person interviews, the commercial research firm verified the authenticity of the participants' information with the patients and their doctors. The second step was a screening test to check if the potential participants understood the DCE choice tasks as intended. Those who failed the test were screened out. Through the screening test, the academic research team selected 800 participants.

DCE Design

As shown in Figure 2, we conducted our DCE following the practice recommended by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) [19]; the checklist is shown in Multimedia Appendix 3.

The DCE attributes and levels were developed through a literature search, qualitative research on China's OMC [12], and focus groups involving the OMC industry and academic professionals. We also undertook 2 rounds of pilot study to test the attributes and levels. The DCE attributes and levels are listed in Table 1.

Figure 2. Processes of DCEs following the checklist recommended by ISPOR (International Society for Pharmacoeconomics and Outcomes Research). DCE: discrete choice experiment.

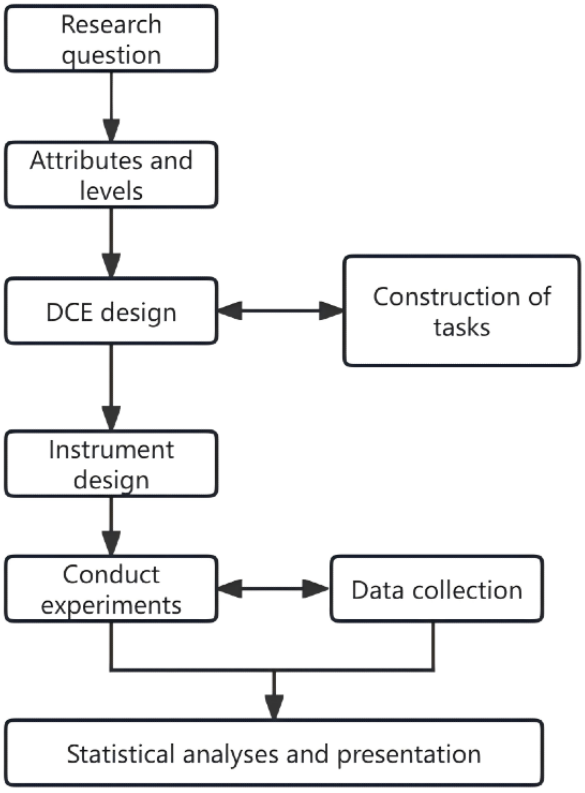

XSL•FO
RenderX

Table 1. Attributes and levels in the DCE[a] survey.

Attributes	Levels
Doctor level	• Ordinary doctors • General experts • Well-known experts
Hospital level	• County-level hospitals • Municipal-level hospitals • Provincial tertiary hospitals
Out-of-pocket cost (RMB[b])	• 25 • 50 • 80 • 150
Waiting time (minutes)	• 15 • 30 • 60 • 180
Consultation format	• Text consultation • Telephone consultation
Consultation duration (minutes)	• 10 • 15 • 20

[a]DCE: discrete choice experiment.

[b]The conversion rate is approximated at 1 RMB ≈ US $0.14 on September 27, 2024.

The description of doctor levels and hospital levels uses the common terminology among the population and follows the regulations of China's hospital management. The professional description of Chinese doctors' titles is resident physician, attending physician, deputy chief physician, and chief physician. The higher the title, the higher the experience. However, in the pilot experiments, we found participants could not clearly define the differences between titles. We revised the description of the doctors' level by using ordinary doctors (equal to resident physicians and attending physicians), general experts (deputy chief physicians), and well-known experts (chief physicians). In our web-based support, we provided an explanation of doctors' titles. A similar process was also applicable to the confirmation of the description of the hospital levels. After searching the obesity OMC platforms, we found there were few doctors providing video obesity consultation, which may be explained by the stigma around obesity treatment. Text consultation and telephone consultation were the 2 consultation formats almost all the web-based doctors provided. In text consultation, users could send text, pictures, and voice messages to doctors and the doctors could respond later. Telephone consultation was more direct. Using the telephone system on OMC platforms, doctors accepted a user's consultation request and then directly responded to the user [12].

We constructed the choice sets using Ngene software (ChoiceMetrics) and a D-efficient design was used to generate 16 choice sets, divided into For-Drugs and For-Advice participants. Each choice set was composed of 2 options with 1 more opt-out option using a dual-response design [32].

To assist participants, we provided easily available links to explanations of the attributes and levels in the web-based survey and our survey team also offered offline support. To encourage participants to complete the survey, only those completing the experiment were remunerated. We also designed the digital experiment to identify potential invalid responses. For example, when respondents chose the same choice option 3 times in a row, the digital experiment automatically reminded the respondents to make choices consistent with their true preference.

After the web-based DCE, we invited 5% (n=40) of our participants to take part in qualitative interviews. We mainly invited 2 categories of participants, one with typical responses and one with diverse choices. The 40 qualitative interviewees had representative characteristics in terms of age, region, and socioeconomic characteristics. Since the participants were distributed across China, we used telecommunication to interview the participants one by one. Each interview was completed with one interviewer asking questions based on the interview questionnaire and the other interviewer taking notes. Every interview lasted at least 10 minutes.

Statistical Analysis

Based on random utility theory [31], the data were analyzed in a mixed logit (ML) model [33], wherein the utility that respondent i obtained from choosing alternative j in the choice set s is given by $U_{isj} = V_{isj} + \varepsilon_{isj} = \beta_i X_{isj} + \varepsilon_{isj}$ where vector V_{isj} is the observable systematic part and equals vector $\beta_i X_{isj}$, with vector X_{isj} representing the alternative specific constant (ASC) and attributes of alternative j and vector β_i representing

XSL•FO
RenderX

associated preference parameters. The ASC represents preferences that are inherent and independent of specific attribute values. Vector ε_{isj} is the unobservable random component and is independently and identically distributed as a type 1 extreme value [34]. The willingness to pay (WTP) [35] and relative importance scores (RIS) [36] were also computed based on the ML model. All the analyses were conducted with STATA 16 (StataCorp).

Qualitative Analysis

Involving 40 DCE participants, the postexperiment qualitative interviews explored the participants' feelings about the experiment, their preferences for and opinions about OMC, and their views about the different scenarios. One interviewer took notes, which were reviewed by the corresponding author, the note-taking interviewer, and one member of the academic research team to develop major themes and issues. The main themes and issues identified by each note assessor were compared and discussed.

Results

Table 2 shows the participants' socioeconomic information, including marital status, education, and income. Even though we used BMI, age, sex, and region to construct our representative samples [28], the distribution of marital status, education, and income between the 2 groups was also broadly comparable.

Table 3 shows the results of preferences and WTP space based on the ML model for the For-Drugs scenario and Table 4 shows the results for the For-Advice scenario. Similar preferences were found between the 2 scenarios. Participants in both scenarios preferred doctors with higher titles and from higher-level hospitals. However, the coefficients in the For-Advice scenario were higher than the For-Drugs scenario coefficients, implying For-Advice participants rated the requirements and their expectations of doctors higher than the For-Drugs participants when seeking professional advice. As for the consultation format, participants in the For-Advice scenario had a significantly higher preference for telephone consultations over text consultations (P=.002). Shorter waiting time and lower OMC fees were preferred in both scenarios. For-Drugs participants preferred longer consultation time (β=.029; P<.001), while For-Advice participants preferred shorter consultation time (β=–.030; P=.001).

Table 5 compares the differences in the 2 scenarios directly. Participants in the For-Advice scenario were willing to pay more for doctors with a higher title; doctors from higher-level hospitals; and RMB 9 (the conversion rate is approximated at 1 RMB ≈ US $0.14 on September 27, 2024) more for telephone consultation than text consultation. Participants in both scenarios were not WTP more for a longer consultation duration. Relative importance scores were also calculated based on the ML model, and the 3 most influential attributes were cost, hospital level, and doctor level in descending order, with costs accounting for more than 48% of the explanatory power in both scenarios. For For-Drugs participants, the importance of consultation duration was greater than waiting time, but For-Advice participants rated waiting time as more important than consultation duration. RIS estimation also revealed that in the For-Drugs scenario, the importance of the consultation form did not contribute to their choice but was significantly important (β=.029; P<.001) for the For-Advice participants.

Based on our post-DCE interviews with 40 participants, our qualitative feedback confirmed the conclusions drawn from the statistical analysis. Four main themes were identified: For-Advice and For-Drugs represented 2 distinct reasons individuals living with obesity selected OMC; the level of the hospital was more important than the title of the doctor; slow responses from web-based doctors can reduce users' willingness to use and pay for OMC; and longer patient consultations were preferred by the For-Drugs patients who sought detailed advice on drug side effects, while quick and direct responses were preferred by the For-Advice participants.

XSL•FO
RenderX

Table 2. Characteristics of participants following distribution characteristics of obesity epidemic in China.

Variable	Overweight distribution (24≤BMI<28; n=271), n (%)	Obesity distribution (BMI≥28; n=129), n (%)	Overall distribution (n=400), n (%)
Scenario: For-Drugs			
Sex			
Male	145 (53.51)	63 (48.84)	208 (52)
Female	126 (46.49)	66 (51.16)	192 (48)
Age group (years)			
18-34	46 (16.97)	25 (19.38)	71 (17.75)
35-54	96 (35.42)	45 (34.88)	141 (35.25)
55-74	93 (34.32)	44 (34.11)	137 (34.25)
75 and older	36 (13.28)	15 (11.63)	51 (12.75)
Region			
North China	46 (16.97)	28 (21.71)	74 (18.50)
Northeast China	45 (16.61)	25 (19.38)	70 (17.50)
East China	38 (14.02)	15 (11.63)	53 (13.25)
Central China	36 (13.28)	19 (14.73)	55 (13.75)
South China	30 (11.07)	6 (4.65)	36 (9)
Southwest China	37 (13.65)	17 (13.18)	54 (13.50)
Northwest China	39 (14.39)	19 (14.73)	58 (14.50)
Marital status			
Married	226 (83.39)	103 (79.84)	329 (82.25)
Not married	45 (16.61)	26 (20.16)	71 (17.75)
Highest education			
Primary and below	23 (8.49)	10 (7.75)	33 (8.25)
Junior high	60 (22.14)	30 (23.26)	90 (22.50)
Senior high	35 (12.92)	19 (14.73)	54 (13.50)
Junior college	41 (15.13)	26 (20.16)	67 (16.75)
University	83 (30.63)	34 (26.36)	117 (29.25)
Master's or doctorate degree	29 (10.70)	10 (7.75)	39 (9.75)
Monthly income before tax (RMB[a])			
≤2000	26 (9.59)	20 (15.50)	46 (11.50)
2001-6000	116 (42.80)	47 (36.43)	163 (40.75)
6001-12,000	100 (36.90)	54 (41.86)	154 (38.50)
12,001-35,000	26 (9.59)	7 (5.43)	33 (8.25)
≥35,001	3 (1.11)	1 (0.78)	4 (1)
Scenario: For-Advice			
Sex			
Male	145 (53.51)	63 (48.84)	208 (52)
Female	126 (46.49)	66 (51.16)	192 (48)
Age group (years)			
18-34	46 (16.97)	24 (18.60)	70 (17.50)
35-54	98 (36.16)	46 (35.66)	144 (36)
55-74	93 (34.32)	44 (34.11)	137 (34.25)

XSL•FO
RenderX

Variable	Overweight distribution (24≤BMI<28; n=271), n (%)	Obesity distribution (BMI≥28; n=129), n (%)	Overall distribution (n=400), n (%)
75 and older	34 (12.55)	15 (11.63)	49 (12.25)
Region			
North China	46 (16.97)	30 (23.26)	76 (19)
Northeast China	45 (16.61)	25 (19.38)	70 (17.50)
East China	38 (14.02)	15 (11.63)	53 (13.25)
Central China	35 (12.92)	19 (14.73)	54 (13.50)
South China	32 (11.81)	10 (7.75)	42 (10.50)
Southwest China	37 (13.65)	15 (11.63)	52 (13)
Northwest China	38 (14.02)	15 (11.63)	53 (13.25)
Marital status			
Married	234 (86.35)	114 (88.37)	348 (87)
Not married	37 (13.65)	15 (11.63)	52 (13)
Highest education			
Primary and below	15 (5.54)	10 (7.75)	25 (6.25)
Junior high	74 (27.31)	33 (25.58)	107 (26.75)
Senior high	45 (16.61)	18 (13.95)	63 (15.75)
Junior college	51 (18.82)	23 (17.83)	74 (18.50)
University	64 (23.62)	37 (28.68)	101 (25.25)
Master's or doctorate degree	22 (8.12)	8 (6.20)	30 (7.50)
Monthly income before tax (RMB)			
≤2000	24 (8.86)	19 (14.73)	43 (10.75)
2001-6000	129 (47.60)	59 (45.74)	188 (47)
6001-12,000	91 (33.58)	39 (30.23)	130 (32.50)
12,001-35,000	20 (7.38)	11 (8.53)	31 (7.75)
≥35,001	7 (2.58)	1 (0.78)	8 (2)

[a]The conversion rate is approximated at 1 RMB ≈ US $0.14 on September 27, 2024.

XSL•FO
RenderX

Table 3. Preference and WTP[a] space for OMC[b] in the For-Drugs scenario.

ASC[c] or attribute and levels	Coefficient			SD		
	β	*P* value	95% CI	β	*P* value	95% CI
Preference estimates						
ASC	2.994	<.001	2.323-3.665	3.177	<.001	2.400-3.954
Doctor level						
Ordinary doctors	Reference	Reference	Reference	Reference	Reference	Reference
Well-known experts	1.230	<.001	1.044-1.417	.565	<.001	0.278-0.853
General experts	.396	<.001	0.246-0.546	.225	.35	0.247-0.697
Hospital level						
County-level hospitals	Reference	Reference	Reference	Reference	Reference	Reference
Provincial tertiary hospitals	1.464	<.001	1.258-1.671	.944	<.001	0.693-1.195
Municipal-level hospitals	.946	<.001	0.770-1.122	.557	.001	0.237-0.876
Consultation format						
Text consultation	Reference	Reference	Reference	Reference	Reference	Reference
Telephone consultation	.131	.06	–0.004 to 0.265	.161	.54	0.674-0.353
Consultation duration	.029	<.001	0.013-0.045	.005	.80	0.034-0.044
Waiting time	–.003	<.001	–0.005 to –0.002	.016	.054	0.000-0.032
Out-of-pocket cost	–.026	<.001	–0.041 to –0.012	.046	.14	0.015-0.106
WTP space estimates						
ASC	178.159	<.001	138.117-218.201	192.313	<.001	228.542-156.084
Doctor level						
Ordinary doctors	Reference	Reference	Reference	Reference	Reference	Reference
Well-known experts	67.127	<.001	56.573-77.681	28.927	<.001	44.883-12.970
General experts	18.496	<.001	9.049-27.943	1.640	.89	21.107-24.387
Hospital level						
County-level hospitals	Reference	Reference	Reference	Reference	Reference	Reference
Provincial tertiary hospitals	86.542	<.001	74.972-98.112	52.905	<.001	39.701-66.108
Municipal-level hospitals	57.800	<.001	47.169-68.432	28.406	.009	49.696-7.117
Consultation format						
Text consultation	Reference	Reference	Reference	Reference	Reference	Reference
Telephone consultation	6.556	.11	–1.452 to 14.564	8.485	.63	25.988-42.957
Consultation duration	–1.105	.26	–3.012 to 0.802	1.709	<.001	0.856-2.563
Waiting time	–3.642	<.001	–4.918 to –2.366	2.462	<.001	1.838-3.086

[a]WTP: willingness to pay.

[b]OMC: online medical consultation.

[c]ASC: alternative specific constant.

XSL•FO
RenderX

Table 4. Preference and WTP[a] space for OMC[b] in the For-Advice scenario.

ASC[c] or attribute and levels	Coefficient			SD		
	β	*P* value	95% CI	β	*P* value	95% CI
Preference estimates						
ASC	3.106	<.001	2.426-3.786	3.080	<.001	2.445-3.714
Doctor level						
Ordinary doctors	Reference	Reference	Reference	Reference	Reference	Reference
Well-known experts	1.362	<.001	1.142-1.582	.969	<.001	0.697-1.241
General experts	.418	<.001	0.249-0.587	.612	<.001	0.313-0.912
Hospital level						
County-level hospitals	Reference	Reference	Reference	Reference	Reference	Reference
Provincial tertiary hospitals	1.549	<.001	1.309-1.789	1.269	<.001	0.992-1.546
Municipal-level hospitals	1.095	<.001	0.905-1.285	.471	.02	0.083-0.858
Consultation format						
Text consultation	Reference	Reference	Reference	Reference	Reference	Reference
Telephone consultation	.234	.002	0.084-0.384	.322	.20	0.169-0.813
Consultation duration	–.030	.001	–0.047 to –0.013	.005	.78	0.030-0.040
Waiting time	–.004	<.001	–0.005 to –0.003	.010	<.001	0.003-0.018
Out-of-pocket cost	–.031	<.001	–0.038 to –0.024	.066	<.001	0.035-0.097
WTP space estimates						
ASC	200.153	<.001	154.694-245.612	208.949	<.001	163.235-254.663
Doctor level						
Ordinary doctors	Reference	Reference	Reference	Reference	Reference	Reference
Well-known experts	72.786	<.001	61.445-84.128	49.111	<.001	62.566-35.656
General experts	18.684	<.001	9.545-27.823	1.567	.91	29.625-26.491
Hospital level						
County-level hospitals	Reference	Reference	Reference	Reference	Reference	Reference
Provincial tertiary hospitals	87.429	<.001	74.820-100.039	71.251	<.001	58.043-84.459
Municipal-level hospitals	65.553	<.001	54.633-76.472	28.456	.01	50.370-6.542
Consultation format						
Text consultation	Reference	Reference	Reference	Reference	Reference	Reference
Telephone consultation	9.962	.02	1.277-18.646	0.302	.98	21.572-22.176
Consultation duration	–1.306	.29	–3.707 to 1.095	1.660	.001	0.639-2.680
Waiting time	–1.948	<.001	–2.566 to –1.330	0.928	<.001	0.532-1.324

[a]WTP: willingness to pay.
[b]OMC: online medical consultation.
[c]ASC: alternative specific constant.

XSL•FO
RenderX

Table 5. Comparison of preference, RIS[a], and WTP[b] for OMC[c] in different scenarios.

Attribute and levels	For-Drugs			For-Advice		
	Preference coefficient (*P* value)	RIS (*P* value)	WTP coefficient (*P* value)	Preference coefficient (*P* value)	RIS (*P* value)	WTP coefficient (*P* value)
Doctor level		0.179 (<.001)			0.170 (<.001)	
Ordinary doctors	Reference	Reference	Reference	Reference	Reference	Reference
Well-known experts	1.230 (<.001)	N/A[d]	67.127 (<.001)	1.362 (<.001)	N/A	72.786 (<.001)
General experts	0.396 (<.001)	N/A	18.496 (<.001)	0.418 (<.001)	N/A	18.684 (<.001)
Hospital level		0.214 (<.001)			0.193 (<.001)	
County-level hospitals	Reference	Reference	Reference	Reference	Reference	Reference
Provincial tertiary hospitals	1.464 (<.001)	N/A	86.542 (<.001)	1.549 (<.001)	N/A	87.429 (<.001)
Municipal-level hospitals	0.946 (<.001)	N/A	57.8 (<.001)	1.095 (<.001)	N/A	65.553 (<.001)
Consultation format		N/A			0.029 (.002)	
Text consultation	Reference	Reference	Reference	Reference	Reference	Reference
Telephone consultation	0.131 (.06)	N/A	6.556 (.11)	0.234 (.002)	N/A	9.962 (.02)
Consultation duration	0.029 (<.001)	0.082 (<.001)	–1.105 (.256)	–0.030 (.001)	0.037 (<.001)	–1.306 (.29)
Waiting time	–0.003 (<.001)	0.043 (.001)	–3.642 (<.001)	–0.004 (<.001)	0.083 (<.001)	–1.948 (<.001)
Out-of-pocket cost	–0.026 (<.001)	0.482 (<.001)	N/A	–0.031 (<.001)	0.487 (<.001)	N/A

[a]RIS: relative importance score.

[b]WTP: willingness to pay.

[c]OMC: online medical consultation.

[d]N/A: not available.

Discussion

Principal Findings

This is one of the first studies of the demand by individuals living with overweight or obesity for OMC in China. We identified 2 typical demand scenarios, For-Drugs and For-Advice, and used DCE to compare patient preferences in the 2 scenarios. Our goal is to reveal the different preferences of individuals living with overweight or obesity for OMC in the different demand scenarios.

Our results demonstrated similar preferences for OMC in both scenarios. Participants in both the For-Advice and For-Drugs scenarios preferred doctors with higher-level titles and from higher-level hospitals, with less waiting time and lower cost. In both scenarios, the RIS estimation revealed that cost had the highest influence on participants' choices (over 48%), followed by hospital level and doctor level. The similarities in patients' For-Advice and For-Drugs preferences identified key OMC competitive advantages, mainly pricing and providing doctors with higher-level titles and from higher-level hospitals, especially tertiary hospitals.

The DCE results showed that the For-Drugs scenario patients preferred longer consultation time (β=.029; *P*<.001), while For-Advice participants preferred shorter consultation time (β=–.030; *P*=.001). This demand difference has been neglected in previous OMC studies [37]. Combined with our qualitative interviews conducted after the DCE, the time differences suggest that For-Drugs consultations were not deemed urgent by patients, but patients wanted more time to discuss their problems, especially any drug side effects. In our interviews, we found the medication needs of adults living with overweight or obesity mainly fell into 2 categories: one for long-term chronic disease management and one for weight loss, which also has long-term effects. The long-term medication needs of For-Drugs participants explain their preference for longer consultation time.

Our DCE results revealed telephone consultation was preferred in the For-Advice scenario ($\beta_{preference}$=.0234 and β_{WTP}=9.962) compared to the For-Drugs scenario ($\beta_{preference}$ and β_{WTP} not significant). For-Advice participants viewed consultations as more urgent than For-Drugs patients, with For-Advice patients seeking a quick and direct response to their questions, preferring telephone consultations. We recommend that OMC platforms distinguish the different time preferences and consultation types for users living with overweight or obesity with different demands. For example, it is possible to provide a specific consultation channel to address medication issues where text and longer consultations are provided. Our qualitative interviews also found that adults individuals living with overweight or obesity cared about the long-term effects of medication. We recommend OMC platforms provide long-term consultation service packages for For-Drugs users to achieve better medication adherence management. By comparing the answers of participants from the 2 scenarios in our qualitative interview, we found that compared to drug-use consultations, the preferences for medical advice emphasized urgent responses, with adults living with overweight or obesity seeking quick and

XSL•FO
RenderX

direct consultations, which was consistent with our DCE estimation. We recommend the OMC platforms provide more telephone consultations for For-Advice patients, with doctors responding clearly and directly.

The different RIS estimates in the 2 scenarios revealed that in the For-Drugs scenario, consultation duration was relatively more important than waiting time, but in the For-Advice scenario, waiting time was more important than consultation duration. The RIS estimation also revealed that the consultation format was not an attribute that contributed to respondents' choice in the For-Drugs scenario; however, in the For-Advice scenario, the consultation format contributed 2.9% to the choice made by respondents. RIS estimation verified our explanation that the For-Advice scenario patients needed more urgent and direct consultation. Although high-end medical resources are scarce in China, we recommend the OMC platforms, supported by regulators, allocate more higher-level doctors, short and efficient telephone consultations, satisfying For-Advice users' preferences and maximizing the use of scarce high-end doctor services.

The WTP estimation showed that respondents in the For-Advice scenario were willing to pay more for doctors with higher-level titles and from higher-level hospitals and telephone consultations, which supports our recommendation that OMC platforms allocate more high-end consultation resources to the For-Advice patients in obesity treatment. Since For-Drugs patients prefer longer consultations, while For-Advice want quick replies, we also recommend doctors schedule long text medication consultations for their For-Drugs patients, but not for their For-Advice patients who prefer telephone consultations. To increase OMC performance, doctors should categorize patients as For-Advice and For-Drugs, with different consultation formats, different mix of telephone and text messaging, and different waiting times. Regulators should allow and encourage OMC platforms to differentiate web-based medical services to maximize the use of medical resources.

Comparison With Prior Work and Strengths

This study is among the first to study obese individuals' preferences for OMC, and our findings provided more detailed implications compared with previous studies [38]. Previous studies on OMC neglected the different demands of patients or users [39-41]. By identifying 2 OMC demand scenarios, we drew detailed findings and implications to inform the OMC industry compared to previous studies [40,41]. Second, previous OMC studies lumped together the demands of patients with different diseases, which may cause selection bias. Given the worldwide obesity epidemic, by focusing on obesity, we designed our experiments to obtain specific recommendations for OMC obesity treatment. Third, chronic diseases have become a major global health threat. Obesity is a major chronic disease and is correlated with the causes of many other chronic diseases. Medication is a vital step in the treatment of most chronic diseases. By comparing the preferences of individuals living with overweight or obesity in For-Advice and For-Drugs scenarios, our research findings have implications for other chronic disease treatments. Fourth, the sampling number in previous OMC studies using DCE methods was seldom over 200 patients. By using the quota sampling method, our research recruited 800 respondents nationally following the epidemiological characteristics of China's adults living with overweight or obesity, which ensured the representativeness of our research.

Limitations

Our study also has a number of limitations. First, we defined OMC as a paid service without considering free web-based consultations, so our results only apply to paid OMC platforms. Free OMC is not comparable to paid OMC in the range and quality of services, and free OMC uses the service as a point of sale for other services. Separate studies should design surveys for free OMC services. Second, while we noted the popularity of weight-loss drugs, such as semaglutide, our For-Drugs DCE does not provide insights into preferences for any special medicine. Limiting the DCE to a specific weight-loss drug would introduce choice bias, and alternative approaches, such as qualitative interviews, are better suited to research on specific weight-loss drugs. Third, in our experiments, OMC service providers were limited to doctors without considering other professions, such as dietitians and fitness instructors, who were also obesity consultants. Alternative research designs should investigate nonmedical technology interventions in obesity treatment [42]. Finally, we used quota sampling based on the national survey in 2020 to recruit our participants. Future studies should investigate alternative quota databases providing the latest epidemiological characteristics of China's overweight and individuals living with obesity.

Future Directions

The future development of OMC depends highly on users' choices. Not all diseases are suited to OMC and patients with different diseases may have different preferences for web-based consultation. Using DCE, future OMC research should investigate the OMC users' demand preferences for other diseases. We also recommend more comparative research on the treatment effects between web-based consultation and offline outpatient services.

Conclusions

Obesity is a unique chronic disease. We identified different For-Drugs and For-Advice demand scenarios in obesity OMC use. Based on the ML model to estimate preferences, WTP, and RIS, we found respondents in both scenarios preferred doctors with higher titles and from higher-level hospitals, less waiting time, and low cost. Consultation duration was relatively more important than waiting time for respondents in the For-Drugs scenario, while shorter waiting time was a more important preference than longer consultation duration in the For-Advice scenario. We also found the group in the For-Advice scenario would pay more for doctors with higher titles and telephone consultations. In the For-Drugs scenario, patient consultation was more important than a quick consultation, while For-Advice patients sought prompt telephone responses. We recommend OMC platforms provide different services catering to the different preferences and WTP for individuals living with obesity. We also recommend doctors differentiate service priorities for For-Advice and For-Drugs patients living with

XSL•FO
RenderX

overweight or obesity. Regulators should allow and encourage OMC platforms to differentiate web-based medical services, including pricing, by user need to allocate scarce high-end medical resources efficiently.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Discrete choice experiment explanation of the For-Drugs scenario.
[PNG File , 117 KB - jmir_v26i1e53140_app1.png]

Multimedia Appendix 2
Discrete choice experiment explanation of the For-Advice scenario.
[PNG File , 96 KB - jmir_v26i1e53140_app2.png]

Multimedia Appendix 3
A checklist for discrete choice experiments conducted among Chinese individuals living with obesity.
[PDF File (Adobe PDF File), 108 KB - jmir_v26i1e53140_app3.pdf]

References

1. Chooi YC, Ding C, Magkos F. The epidemiology of obesity. Metabolism 2019;92:6-10. [doi: 10.1016/j.metabol.2018.09.005] [Medline: 30253139]
2. Pan XF, Wang L, Pan A. Epidemiology and determinants of obesity in China. Lancet Diabetes Endocrinol 2021;9(6):373-392. [doi: 10.1016/S2213-8587(21)00045-0] [Medline: 34022156]
3. Zeng Q, Li N, Pan XF, Chen L, Pan A. Clinical management and treatment of obesity in China. Lancet Diabetes Endocrinol 2021;9(6):393-405. [doi: 10.1016/S2213-8587(21)00047-4] [Medline: 34022157]
4. Wang Y, Min J, Khuri J, Xue H, Xie B, A Kaminsky L, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. JMIR Mhealth Uhealth 2020;8(4):e15400. [doi: 10.2196/15400] [Medline: 32343253]
5. Oh B, Yi GH, Han MK, Kim JS, Lee CH, Cho B, et al. Importance of active participation in obesity management through mobile health care programs: substudy of a randomized controlled trial. JMIR Mhealth Uhealth 2018;6(1):e2. [doi: 10.2196/mhealth.8719] [Medline: 29298749]
6. O'Hara VM, Johnston SV, Browne NT. The paediatric weight management office visit via telemedicine: pre- to post-COVID-19 pandemic. Pediatr Obes 2020;15(8):e12694. [doi: 10.1111/ijpo.12694] [Medline: 32627434]
7. Al-Mahdi I, Gray K, Lederman RM. Online medical consultation: a review of literature and practice. 2015 Presented at: 8th Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2015); January 27-30, 2015; Sydney, Australia URL: https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV164Al-Mahdi.pdf
8. Bray GA. Medical consequences of obesity. J Clin Endocrinol Metab 2004;89(6):2583-2589. [doi: 10.1210/jc.2004-0535] [Medline: 15181027]
9. Weghuber D, Barrett T, Barrientos-Pérez M, Gies I, Hesse D, Jeppesen OK, et al. Once-weekly semaglutide in adolescents with obesity. N Engl J Med 2022;387(24):2245-2257 [FREE Full text] [doi: 10.1056/NEJMoa2208601] [Medline: 36322838]
10. Rubino DM, Greenway FL, Khalid U, O'Neil PM, Rosenstock J, Sørrig R, et al. Effect of weekly subcutaneous semaglutide vs daily liraglutide on body weight in adults with overweight or obesity without diabetes: the STEP 8 randomized clinical trial. JAMA 2022;327(2):138-150. [doi: 10.1001/jama.2021.23619] [Medline: 35015037]
11. Velayati F, Ayatollahi H, Hemmat M, Dehghan R. Telehealth business models and their components: systematic review. J Med Internet Res 2022;24(3):e33128. [doi: 10.2196/33128] [Medline: 35348471]
12. Yu D, Hu Y, Wang J, Nicholas S, Maitland E. Online medical consultation in China: evidence from obesity doctors. Digit Health 2023;9:20552076231182789 [FREE Full text] [doi: 10.1177/20552076231182789] [Medline: 37342094]
13. Li Y, Yan X, Song X. Provision of paid web-based medical consultation in China: cross-sectional analysis of data from a medical consultation website. J Med Internet Res 2019;21(6):e12126. [doi: 10.2196/12126] [Medline: 31162129]
14. Chiu YL, Wang JN, Yu H, Hsu YT. Consultation pricing of the online health care service in China: hierarchical linear regression approach. J Med Internet Res 2021;23(7):e29170. [doi: 10.2196/29170] [Medline: 34259643]
15. Byaruhanga J, Atorkey P, McLaughlin M, Brown A, Byrnes E, Paul C, et al. Effectiveness of individual real-time video counseling on smoking, nutrition, alcohol, physical activity, and obesity health risks: systematic review. J Med Internet Res 2020;22(9):e18621. [doi: 10.2196/18621] [Medline: 32915156]
16. Baulch J, Chester A, Brennan L. Treatment alternatives for overweight and obesity: the role of online interventions. Behav Change 2012;25(1):1-14. [doi: 10.1375/bech.25.1.1]

XSL•FO
RenderX

17. Puhl RM, Heuer CA. Obesity stigma: important considerations for public health. Am J Public Health 2010;100(6):1019-1028. [doi: 10.2105/AJPH.2009.159491] [Medline: 20075322]
18. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics 2008;26(8):661-677. [doi: 10.2165/00019053-200826080-00004] [Medline: 18620460]
19. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health—a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health 2011;14(4):403-413. [doi: 10.1016/j.jval.2010.11.013] [Medline: 21669364]
20. Tenbult N, Kraal J, Brouwers R, Spee R, Eijsbouts S, Kemps H. Adherence to a multidisciplinary lifestyle program for patients with atrial fibrillation and obesity: feasibility study. JMIR Form Res 2022;6(4):e32625. [doi: 10.2196/32625] [Medline: 35486435]
21. Choi SH, Templin T, Marrocco A. Development of an online discrete choice experiment to elicit preferences for tobacco treatment among college students. J Am Coll Health 2023:1-8. [doi: 10.1080/07448481.2023.2227715] [Medline: 37437187]
22. Determann D, Lambooij MS, Steyerberg EW, de Bekker-Grob EW, de Wit GA. Impact of survey administration mode on the results of a health-related discrete choice experiment: online and paper comparison. Value Health 2017;20(7):953-960. [doi: 10.1016/j.jval.2017.02.007] [Medline: 28712625]
23. Lehmann N, Sloot D, Ardone A, Fichtner W. The limited potential of regional electricity marketing—results from two discrete choice experiments in Germany. Energy Economics 2021;100:105351. [doi: 10.1016/j.eneco.2021.105351]
24. Arize I, Onwujekwe O. Acceptability and willingness to pay for telemedicine services in Enugu state, southeast Nigeria. Digit Health 2017;Jan-Dec(3). [doi: 10.1177/2055207617715524] [Medline: 29942606]
25. Bellé N, Cantarelli P. The role of motivation and leadership in public employees' job preferences: evidence from two discrete choice experiments. Int Public Manag J 2018;21(2):191-212. [doi: 10.1080/10967494.2018.1425229]
26. Adult weight determination [Article in Chinese]. National Health Commission of the People's Republic of China. 2013. URL: http://www.nhc.gov.cn/wjw/yingyang/201308/a233d450fdbc47c5ad4f08b7e394d1e8.shtml [accessed 2013-08-08]
27. Huang L, Wang Z, Wang H, Zhao L, Jiang H, Zhang B. Nutrition transition and related health challenges over decades in China. Eur J Clin Nutr 2021;75(2):247-252. [doi: 10.1038/s41430-020-0674-8] [Medline: 32620907]
28. Zhang L, Wang Z, Wang X, Chen Z, Shao L, Tian Y, et al. Prevalence of overweight and obesity in China: results from a cross-sectional study of 441 thousand adults, 2012-2015. Obes Res Clin Pract 2020;14(2):119-126. [doi: 10.1016/j.orcp.2020.02.005] [Medline: 32139330]
29. Zhan S. Epidemiology. 8th Edition. Beijing, China: People's Medical Publishing House; 2020:1-43.
30. Naing L, Nordin RB, Abdul Rahman H, Naing YT. Sample size calculation for prevalence studies using Scalex and ScalaR calculators. BMC Med Res Methodol 2022;22(1):209. [doi: 10.1186/s12874-022-01694-7] [Medline: 35907796]
31. Louviere JJ, Hensher DA, Swait JD. Stated Choice Methods: Analysis and Applications. Cambridge, United Kingdom: Cambridge University Press; 2000.
32. Rose JM, Hess S. Dual-response choices in pivoted stated choice experiments. Transportation Res Rec J Transportation Res Board 2009;2135(1):25-33. [doi: 10.3141/2135-04]
33. McFadden D, Train K. Mixed MNL models for discrete response. J Appl Econometrics 2000;15(5):447-470. [doi: 10.1002/1099-1255(200009/10)15:5<447::aid-jae570>3.0.co;2-1]
34. Si L, Tu L, Xie Y, Palmer AJ, Gu Y, Zheng X, et al. Chinese patients' preference for pharmaceutical treatments of osteoporosis: a discrete choice experiment. Arch Osteoporos 2019;14(1):85. [doi: 10.1007/s11657-019-0624-z] [Medline: 31367860]
35. Hole AR, Kolstad JR. Mixed logit estimation of willingness to pay distributions: a comparison of models in preference and WTP space using data from a health-related choice experiment. Empir Econ 2011;42(2):445-469. [doi: 10.1007/s00181-011-0500-1]
36. Lancsar E, Louviere J, Flynn T. Several methods to investigate relative attribute impact in stated preference experiments. Soc Sci Med 2007;64(8):1738-1753. [doi: 10.1016/j.socscimed.2006.12.007] [Medline: 17257725]
37. Cao B, Huang W, Chao N, Yang G, Luo N. Patient activeness during online medical consultation in China: multilevel analysis. J Med Internet Res 2022;24(5):e35557. [doi: 10.2196/35557] [Medline: 35622403]
38. Chang J, Savage SJ, Waldman DM. Estimating willingness to pay for online health services with discrete-choice experiments. Appl Health Econ Health Policy 2017;15(4):491-500. [doi: 10.1007/s40258-017-0316-z] [Medline: 28290107]
39. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020;22(2):e16407. [doi: 10.2196/16407] [Medline: 32130131]
40. Mozes I, Mossinson D, Schilder H, Dvir D, Baron-Epel O, Heymann A. Patients' preferences for telemedicine versus in-clinic consultation in primary care during the COVID-19 pandemic. BMC Prim Care 2022;23(1):33 [FREE Full text] [doi: 10.1186/s12875-022-01640-y] [Medline: 35193509]
41. Snoswell CL, Smith AC, Page M, Caffery LJ. Patient preferences for specialist outpatient video consultations: a discrete choice experiment. J Telemed Telecare 2023;29(9):707-715. [doi: 10.1177/1357633X211022898] [Medline: 34142895]

XSL•FO
RenderX

42. Lehmann M, Jones L, Schirmann F. App engagement as a predictor of weight loss in blended-care interventions: retrospective observational study using large-scale real-world data. J Med Internet Res 2024;26:e45469. [doi: 10.2196/45469] [Medline: 38848556]

Abbreviations

ASC: alternative specific constant
DCE: discrete choice experiment
ISPOR: International Society for Pharmacoeconomics and Outcomes Research
ML: mixed logit
OMC: online medical consultation
RIS: relative importance score
WTP: willingness to pay

Edited by N Cahill; submitted 30.09.23; peer-reviewed by P Li, P Wang; comments to author 10.05.24; revised version received 28.05.24; accepted 30.09.24; published 27.11.24.

Please cite as:
Hu Y, Wang J, Zhou J, Gu Y, Nicholas S, Maitland E
Preferences of Individuals With Obesity for Online Medical Consultation in Different Demand Scenarios: Discrete Choice Experiments
J Med Internet Res 2024;26:e53140
URL: https://www.jmir.org/2024/1/e53140
doi:10.2196/53140
PMID:

©Yaolin Hu, Jian Wang, Jianbo Zhou, Yuanyuan Gu, Stephen Nicholas, Elizabeth Maitland. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Teledentistry Applied to Health and Education Outcomes: Evidence Gap Map

Júlia Nascimento da Silva Mulder[1], DDS[‡]; Marcelo Ramos Pinto[1], DDS; Isabelle Aníbal[1], DDS, MDS; Ana Paula Dornellas[1], DDS, PhD; Deise Garrido[1], DDS, PDS; Camila Huanca[1], DDS, PhD; Ana Estela Haddad[1], DDS, PhD, Prof Dr; Carmen Verônica Mendes Abdala[2], MPA, MLS

[1]Department of Orthodontics and Pediatric Dentistry, University of São Paulo - School of Dentistry, São Paulo, Brazil

[2]Latin American and Caribbean Center on Health Sciences Information, São Paulo, Brazil

[‡]Telehealth and Teledentistry Center FOUSP-SAITE

Corresponding Author:
Júlia Nascimento da Silva Mulder, DDS
Department of Orthodontics and Pediatric Dentistry
University of São Paulo - School of Dentistry
Avenida Professor Lineu Prestes, 2227
São Paulo, 05508-000
Brazil
Phone: 55 11 3091 7854
Email: julia.mulder@usp.br

Abstract

Background: Teledentistry is a field of activities that comprises information and communication technologies (ICTs) applied to dentistry, including the exchange of clinical information, patient care, and the use of educational strategies across remote distances. Its use has grown progressively over the past decades—intensified by the COVID-19 pandemic—and has been improving the provision of dental services and educational strategies ever since.

Objective: This evidence gap map (EGM) study aims to present a collection of systematic reviews (SRs) with meta-analyses to answer the question “What are the applications of teledentistry in dental services and dental education?” by identifying gaps and current evidence on the improvement of health care and education.

Methods: The EGM methodology has been developed by the Latin American and Caribbean Center on Health Sciences Information and is based on the concept created by the International Initiative for Impact Evaluation. Embase, PubMed, and Virtual Health Library databases were used for the literature research, using terms for teledentistry associated with eHealth, dental education, and oral health care. The data obtained from the included studies were then characterized in a Microsoft Excel spreadsheet, with a matrix containing 8 intervention groups (combined interventions, e-learning and tele-education, teleconsultation and teleservice, telemonitoring, telediagnosis, telescreening, ICTs, and artificial intelligence) and 8 outcome groups (diagnosis accuracy, education and professional training, user behavior, clinical practice, patient-centered outcomes, clinical outcomes, health services management, and access to health services). The quality of the studies was assessed using AMSTAR2 (A Measurement Tool to Assess Systematic Reviews). The visual analytics platform Tableau (Salesforce) was used to graphically display the confidence level, number of reviews, health outcomes, and intervention effects.

Results: The confidence level obtained by the criteria applied was high for 28% (19/68) of the studies, moderate for 6% (4/68), low for 15% (10/68), and critically low for 51% (35/68). Among the interventions, the ICT group stood out with 182 (36.8%) out of 494 associations, followed by interventions with e-learning and tele-education (n=96, 19.4% of associations), telediagnosis (n=67, 13.6%), and combined interventions (n=53, 10.7%). Most of the outcomes were aimed at education and professional training (97/494, 19.6% of associations), patient-centered outcomes (74/494, 15%), and health services management (60/494, 12.1%).

Conclusions: This EGM presents an overview of the contributions of teledentistry in patient care, health services, clinical practice, and education. The study results may help guide future research and policy decisions and serve as a convenient virtual tool for accessing valuable evidence-based information on teledentistry.

(J Med Internet Res 2024;26:e60590) doi:10.2196/60590

XSL•FO
RenderX

KEYWORDS

teledentistry; systematic review; dental education; dentistry; telemedicine; research design; health information technologies; mobile phone

Introduction

Background

The World Health Organization defines telehealth as the exchange of health information mediated by information and communication technologies (ICTs) for the purposes of health care, education, research, and management of health care services or networks [1]. Although telehealth was already adopted in many countries, its use and applications experienced exponential growth during the COVID-19 pandemic, as health services worldwide were partially or completely interrupted [2-4]. The concept of telehealth applied to dentistry is labeled as teledentistry, and through the use of ICTs, it has been qualifying the provision of dental services directed at improving patient care as well as professional education and training [5]. However, despite teledentistry being a current and global reality, with a great range of reported applications and experiences, evidence gaps for the consolidation of educational and health strategies still exist [6-8].

Considering the importance of dentistry at both individual and population levels, it is logical and advisable that teledentistry is incorporated into health care and, consequently, into teaching [2,9,10]. With the expansion of care integrated into the teaching-service axis, teledentistry favors the dynamics of dental care systems and training, allowing new educational opportunities [10,11]. However, for teledentistry to fulfill its maximum potential within teaching and health care, legal support and the involvement of all policy makers are necessary to enable the optimization and expansion of services, care, and patients' referrals to specialists as well as to assist students through interprofessional interaction, experience, tool development, and remote supervision of patients [7,11-14].

The evidence gap map (EGM) methodology aims to identify current evidence and knowledge gaps on a given subject of study. The given results provide guidance for new research as well as public policy making [15]. The EGM methodology was created by the International Initiative for Impact Evaluation—a global initiative established in 2008 that seeks to develop scientific evidence for transformation effects on the lives of populations in low- and middle-income countries. The Latin American and Caribbean Center for Health Sciences Information of the Pan American Health Organization (BIREME/OPAS) uses the International Initiative for Impact Evaluation strategy to support investments directed at the production of more evidence or the synthesis of existing evidence [15-17].

This Study

This EGM presents an overview of the evidence on the effects of teledentistry for a variety of health and education outcomes, such as diagnosis accuracy, education and professional training, user behavior, clinical practice, patient-centered outcomes, clinical outcomes, health services management, and access to health services [15-17].

Methods

Overview

This study was based on the application of the EGM methodology, which consists of graphically representing the characteristics and findings of the evidence analyzed in systematic reviews (SRs), associating interventions with the outcomes analyzed in the reviews, and linking them with the reported effects of the interventions (with the population and focus countries of the primary studies included in the reviews).

On the map, associations are represented using bubbles of different colors, which represent the confidence level of the reported evidence and the effects of the intervention for each different health outcome. The bubble size is equivalent to the number of studies that analyzed each association. All bubbles lead to a list of study titles, with a link to the full text.

SRs—with or without meta-analysis—that could answer the research question "What are the applications of teledentistry in dental services and dental education?" were eligible for inclusion in the EGM.

For the development of this EGM, five different phases respecting the proposed methodology were applied: literature bibliographic search (phase 1), reference management (phase 2), study characterization (phase 3), quality assessment and report (phase 4), and infographics (phase 5).

Search Strategy

A bibliographic search (EGM phase 1) was conducted using Medical Subject Headings and DeCS terms in English, Portuguese, and Spanish within the scope of telehealth, dentistry, and teledentistry. This search was applied to PubMed (MEDLINE) and 2 additional databases—Virtual Health Library (LILACS, BBO, WHOLIS, and Pan American Health Organization IRIS) and Embase (Elsevier)—excluding results already obtained from PubMed. The full list of terms used in the search can be found in Multimedia Appendix 1. A filter for SRs and meta-analysis on teledentistry was applied to the consulted databases, and the results respected the following PICO strategy:

- P=patients, students, and health professionals
- I=teledentistry
- C=conventional dentistry
- O=health and education

A team of calibrated experts from the Telehealth and Teledentistry Center FOUSP-SAITE (NuTes-FOUSP, University of São Paulo, SP, Brazil), along with researchers and librarians from BIREME/OPAS, was consulted to cross-review the search strategy before establishing the inclusion and exclusion criteria. No disagreement was found between the examiners during this phase. The detailed search strategy can be found in Multimedia Appendix 1.

XSL•FO
RenderX

It is important to point out that the specific term "evidence gap map" (or "evidence map") is not yet included in the Medical Subject Headings and DeCS list till the date of submission of this paper. Furthermore, the expression "research design" was used as the most similar research term.

The term ICT was kept as the main reference, as it was once a part of the majority of articles and is quoted this way in the EGM. However, the researchers acknowledge the existence of the term "digital information and communication technologies," which is being used more frequently to differentiate the forms of technology applied in dentistry (nondigital and digital).

Inclusion Criteria

The inclusion criteria for the studies were as follows: clinical applications of teledentistry; educational applications of teledentistry; applications directly in services; a variety of ICTs with applications in dentistry; SRs and meta-analysis; and studies written in Portuguese, English, or Spanish.

Exclusion Criteria

The exclusion criteria for the studies were as follows: ICTs with no clinical or educational applications, clinical approaches without the use of ICT, clinical studies based on strictly laboratory or laboratory research in the experimental phase, research outside the scope of teledentistry, and studies that were not SRs or meta-analyses.

Data Extraction and Risk of Bias Assessment

Examination of search results was performed through Rayyan (Rayyan Company) reference management software (an artificial intelligence [AI]–assisted systematic literature review tool, from Rayyan Company) by 2 independent and blinded literature reviewers (phase 2) [18]. In case of disagreement among researchers regarding including or excluding an article (based on the inclusion and exclusion criteria), a third researcher of the group, who was still "blinded," made the final decision. Then, the duplicate articles were removed. This platform manages the risk of bias related to influence between individuals and offers a reasonable interface, with a dashboard that informs the overall status about inclusion decisions (undecided, maybe, included, excluded, or conflict) as well as possible duplicates (unresolved, resolved, not duplicates, or deleted).

After full-text reading, for the further evaluation and confidence level classification of studies (equivalent to EGM's phase 4), the AMSTAR2 (A Measurement Tool to Assess Systematic Reviews) checklist was used. This was also established by 2 independent reviewers as a strategy for risk of bias management. AMSTAR2 is aimed at evidence-based health care and establishes 16 questions for the appraisal of SRs to be scored as "yes," "partially yes," or "no" depending on the question and the detailed data retrieved from the review. The questions cover various aspects, including adherence to protocol statements, inclusion criteria, search strategy methods, duplication during study selection and data extraction, justification for excluded studies, detailed descriptions of included studies, risk of bias assessment, funding information, statistical analysis, management of heterogeneity, and conflicts of interest, among others. From these 16 questions, 7 (%) are considered "critical" for quality evaluation. If a study presents no critical weaknesses and up to 1 noncritical weakness, it can be considered a "high" quality review; for studies with no critical weakness but >1 noncritical weakness, it can be considered a "moderate" quality review; for studies with 1 critical weakness (with or without other noncritical weaknesses), the study is to be labeled as a "low" quality review; and studies with >1 critical weakness are labeled "critically low" quality review [19,20]. The list of the included studies can be found in Table 1. The numbered references are presented in Multimedia Appendix 2 [21-88], and the list of the excluded studies after full-text appraisal are presented in Multimedia Appendix 3.

XSL•FO
RenderX

Table 1. List of studies included in the teledentistry evidence gap map. For details about the included studies, please refer to Multimedia Appendix 2.

Reference number of the included study, quoted in Multimedia Appendix 2	Focus countries	Publication year	AMSTAR2[a] confidence level	Main study focus
1	Not specified	2023	Low	Management of pain and anxiety
2	Europe, the United States, Australia, Japan, and India	2018	Critically low	Teledentistry evaluation
3	South Korea, Egypt, Australia, and the United Kingdom	2023	Low	Education and learning
4	Not specified	2021	Critically low	Management of pain and anxiety
5	Not specified	2022	Critically low	Management of pain and anxiety
6	The United States, the United Kingdom, Brazil, Portugal, Australia, and Germany	2018	Critically low	Teledentistry evaluation
7	North America, Latin America, Europe, and Asia	2015	Critically low	Diagnosis
8	India, Brazil, and Malaysia	2023	High	Diagnosis
9	Not specified	2019	Critically low	Education and learning
10	Not specified	2022	Critically low	Education and learning
11	Germany, Saudi Arabia, and the United States	2023	High	Patient care
12	China, Japan, Finland, Portugal, the United States, Germany, and Israel	2023	Low	Virtual reality and artificial intelligence
13	Saudi Arabia, India, Brazil, Australia, the United States, France, Malaysia, and South Korea	2020	Critically low	Dental practice and training
14	The United States, Europe, and Asia	2019	Critically low	Management of pain and anxiety
15	Pakistan, Colombia, India, and Saudi Arabia	2022	Critically low	Teledentistry during the COVID-19 pandemic
16	Iran, Brazil, the United States, Saudi Arabia, India, and Australia	2022	High	Patient care
17	Saudi Arabia, Australia, the United States, Iran, and India	2023	High	Patient care
18	Australia, France, and Germany	2020	Critically low	Access to care and health systems
19	Greece, Philippines, the United States, China, India, South Korea, Germany, Pakistan, and Vietnam	2023	High	Diagnosis
20	Portugal, the United Kingdom, the Netherlands, Turkey, Australia, the United States, and Brazil	2016	Critically low	Diagnosis
21	The United States, India, Italy, Australia, Canada, Iran, China, and Turkey	2020	Critically low	Management of pain and anxiety
22	The United Kingdom, the United States, North America, and Malaysia	2020	Critically low	Education and learning
23	Not specified	2023	High	Education and learning
24	The United States, the United Kingdom, Brazil, Germany, Australia, Greece, Iran, and Saudi Arabia	2021	Critically low	Education and learning
25	The United States and Spain	2023	Low	Education and learning

XSL•FO
RenderX

Reference number of the included study, quoted in Multimedia Appendix 2	Focus countries	Publication year	AMSTAR2[a] confidence level	Main study focus
26	Iraq, Saudi Arabia, the United States, Italy, Peru, Pakistan, India, and China	2019	Critically low	Patient care
27	Not specified	2023	Moderate	Dental monitoring
28	Brazil, Canada, China, Costa Rica, France, Nepal, Peru, Serbia, and the United States	2016	Moderate	Education and learning
29	Not specified	2023	Low	Education and learning and dental practice and training
30	The United States, Germany, and the United Kingdom	2019	Critically low	Education and learning
31	Iran, India, Greece, Syria, Turkey, and Jordan	2023	High	Patient care
32	The United Kingdom, the United States, Australia, the Netherlands, Brazil, Thailand, Spain, France, Hong Kong, Germany, China, India, Serbia, Iran, Japan, Hungary, Sweden, South Korea, and Saudi Arabia	2022	Critically low	Education and learning
33	India, Iran, Turkey, and China	2023	High	Management of pain and anxiety
34	Thailand, Mexico, Iran, Saudi Arabia, Egypt, India, and Syria	2020	Moderate	Management of pain and anxiety
35	The Netherlands, India, the United States, Italy, and China	2021	Critically low	Dental monitoring
36	Not specified	2023	High	Teledentistry during the COVID-19 pandemic and education and learning
37	The United States, France, Japan, Brazil, China, Australia, the United Kingdom, the Netherlands, and Thailand	2023	High	Education and learning
38	Bahrain, Australia, the United Kingdom, Africa, Finland, Canada, Spain, United Arab Emirates, and the United States	2023	Low	Education and learning and patient care
39	Not specified	2023	High	Dental monitoring
40	Not specified	2020	Critically low	Management of pain and anxiety
41	Not specified	2022	Critically low	Patient care
42	Saudi Arabia, the Netherlands, Germany, Romania, Brazil, Italy, Turkey, Bulgaria, Canada, and Israel	2022	Low	Dental monitoring
43	The United States, Germany, the United Kingdom, Australia, Spain, Sweden, Canada, Brazil, China, the Netherlands, Switzerland, Belgium, Finland, Thailand, and Taiwan	2014	High	Education and learning and patient care
44	The United Kingdom, Italy, India, Iran, Saudi Arabia, Peru, Brazil, and Malaysia	2023	High	Patient care
45	Canada, the United Kingdom, Finland, Spain, India, the United States, Italy, and Australia	2017	Low	Patient care

XSL•FO
RenderX

Reference number of the included study, quoted in Multimedia Appendix 2	Focus countries	Publication year	AMSTAR2[a] confidence level	Main study focus
46	The United States, Israel, China, Italy, Jordan, and Austria	2020	Critically low	Dental practice and training and teledentistry during the COVID-19 pandemic
47	India	2022	Critically low	Access to care and health systems
48	The United Kingdom, the United States, Italy, Australia, India, and China	2022	High	Dental practice and training
49	Iran, India, Italy, the Netherlands, Australia, England, Brazil, Niger, and Syria	2022	Critically low	Patient care
50	Not specified	2021	Critically low	Dental practice and training
51	Not specified	2021	Critically low	Dental practice and training
52	Not specified	2023	High	Dental monitoring
53	Iraq, Saudi Arabia, Germany, Portugal, the United States, Italy, Belgium, Pakistan, India, China, Brazil, and the Netherlands	2021	Low	Oral health promotion
54	Brazil, India, and the United Kingdom	2020	Critically low	Diagnosis
55	India, Brazil, Malaysia, the United States, and Italy	2023	High	Diagnosis
56	Not specified	2013	Critically low	Dental practice and training
57	The United States, Europe, and middle-income countries	2013	Critically low	Teledentistry evaluation
58	Spain and Sweden	2023	Moderate	Diagnosis and dental monitoring
59	Not specified	2023	Low	Dental practice and training
60	The United States, Italy, Taiwan, and the Netherlands	2022	High	Dental monitoring
61	The United Kingdom, Saudi Arabia, Germany, England, the Netherlands, Turkey, and Italy	2022	High	Patient care
62	Brazil, the United States, Greece, Germany, Portugal, and the United Kingdom	2021	High	Teledentistry during the COVID-19 pandemic
63	The United States	2022	Critically low	Patient care
64	Not specified	2020	Critically low	Management of pain and anxiety
65	The United States, the United Kingdom, Serbia, Brazil, Sweden, Spain, Taiwan, Portugal, Germany, Australia, Finland, Canada, and Austria	2018	Critically low	Patient care
66	Syria, Iran, the United States, Malaysia, Finland, India, the United Kingdom, Türkiye, Canada, and Australia	2021	Critically low	Virtual reality and artificial intelligence
67	Iran and England	2021	Critically low	Education and learning and patient care
68	Not specified	2022	Critically low	Virtual reality and artificial intelligence

[a]AMSTAR2: A Measurement Tool to Assess Systematic Reviews.

XSL•FO
RenderX

Regarding data safety and protection protocols, the research team only worked with digital information and files (and not physical or printed documents). All digital files were stored on dedicated servers at the University of São Paulo and BIREME, with access restricted to researchers through a password.

Ethical Considerations

As EGM studies apply no direct interventions on individuals or animals, there was no need for study submission to the ethics committee. No sensitive individual (general or health) information was directly generated by the researchers. The use of AI tools was indirect and restricted to the ones already embodied in the platforms or applications used by the researchers; there was no direct application of deep learning and natural language processing (chatbots, generative pretrained transformers, or similar engines) to compose the writing or the revision of this paper.

Data Synthesis and Characterization

A team of 7 dentists who were also researchers from the Telehealth and Teledentistry Center FOUSP-SAITE (NuTes-FOUSP, University of São Paulo, São Paulo, Brazil) worked in collaboration with a team of 3 researchers from the Latin American and Caribbean Center for Health Sciences Information (BIREME/OPAS/World Health Organization) to analyze the data of the included studies.

After full-text analysis, a matrix on Microsoft Excel was developed based on the evidence map methodology to organize the data extracted from the included studies to summarize and measure the results found for all the teledentistry actions mentioned and to create graphic elements to be inserted later on in the interactive map (phase 3). For each row of the selected studies, the detailed matrix included the following: title, intervention group, the individual intervention to be measured, the outcome group, the individual outcome that the intervention seeks to achieve, the effect obtained by the association of both individual intervention and outcome, the focus population of the intervention, the database used to obtain the full text and the ID, the confidence level obtained through the AMSTAR2 criteria, the type of review, the study design of the primary studies included in the review, the review design, the focus and publication countries, the publication year, and a link to the full text. A tab with the possible results for the type of review, review design, study design, confidence level, effects, and populations was created for the automatic count, while interventions and outcomes were grouped and coded before automatic counting. The specific data for each study was counted using the filters available in Microsoft Excel. Table 2 presents the list of outcome groups and the number of associations found for each individual outcome, while Table 3 presents the list of intervention groups and number of associations found for each individual intervention.

The associations and results found were then counted through the use of Microsoft Excel's filters (following the manual check) for further creation and publication of the interactive evidence map using the Tableau platform in an area created for BIREME (Tableau Public [89]), graphically demonstrating the panorama of consolidated evidence and the gaps identified in the selected studies through the inclusion criteria and entered into the characterization matrix (phase 5, all available via reference 21 in the reference list).

The interactive map is divided into three main sections: (1) general, (2) confidence level, and (3) intervention effect. In the first one, same-colored bubbles are presented to illustrate the volume of SRs published for a certain association (outcome vs intervention crossing). For confidence level, these bubbles were either green, yellow, red, or blue, indicating associations belonging to reviews that have achieved high, moderate, low, or critically low confidence levels, respectively, on AMSTAR2. Regarding the intervention effect section, associations will be illustrated by dark green (for positive effects), light green (for potential positive), orange (potential negative), red (negative), pink (inconclusive), light blue (no effect), or gray (no reported effect). By directing the cursor over the bubble in any section, a pop-up appears, presenting the number of reviews and publication countries that have analyzed this particular association effect. A list of the included studies can be found right below the map to provide open access to the reviews. The map also presents a summary of the main findings and information about the authors and collaborators [89,90].

Table 2 presents the relationship between the outcome groups and individual outcomes, followed by the prevalence of associations found for each of these.

Table 3 presents the relationship between the intervention groups and individual interventions, followed by the prevalence of associations found for each of these.

XSL•FO
RenderX

Table 2. List of outcome groups and the number of associations found for each individual outcome.

Outcome groups and outcomes	Associations found, n	Associations by outcome group, n	Equivalence to the total number of associations (%)
Access to health services		21	4
Convenience	2		
Service quality	10		
Reduction of unnecessary traveling	1		
Wait reduction	4		
Line reduction	2		
Surgery access	2		
Diagnosis accuracy		44	9
Carious lesions	10		
Enamel defects	2		
Oral conditions	3		
Trauma	2		
Diagnosis in general	8		
Erosion	2		
Fluorosis	2		
Cancer	7		
Potentially malignant oral lesions	7		
Urgency detection	1		
Clinical outcomes		66	13
Gingival inflammation reduction	10		
Plaque index reduction	11		
Probing depth reduction	2		
Probing bleeding reduction	2		
Oral disease prevention	19		
Cross infection Prevention	1		
Clinical outcomes in general	7		
Prognosis improvement	13		
Orthodontic relapse	1		
Education and professional training		97	20
Learning	26		
Assessment performance	6		
Professional training	21		
Clinical skills improvement	24		
Increase in student's confidence	5		
Student's stress management	4		
Pedagogical model	6		
Educational path	4		
Student's convenience	1		
Health services management		60	12
Health care extension	21		
Patient monitoring	5		

XSL•FO
RenderX

Outcome groups and outcomes	Associations found, n	Associations by outcome group, n	Equivalence to the total number of associations (%)
Dental emergencies	2		
Time optimization	18		
Regulation and health care systems	3		
Cost reduction	7		
Decrease on appointments	4		
Patient-centered outcomes		74	15
Distress relief	1		
Health behavior change	14		
Psychological impact relief	1		
Health education	16		
Safety	2		
Behavior along treatment	6		
Anxiety relief	13		
Pain perception relief	12		
Change in hygiene habits	9		
Clinical practice		75	15
Clinical training	9		
Decision-making	13		
Communication among professionals	4		
Professional and patient communication	2		
Procedure planning	4		
Applied to restorative dentistry	2		
Applied to radiology	2		
Applied to orthodontics	14		
Applied to oral surgery	5		
Digital planning	8		
Human relations	5		
Control of events and complications	7		
User behavior		57	12
Adherence to ICT[a]	4		
Treatment adherence	10		
ICT acceptability	9		
ICT reliability	1		
ICT perception	8		
ICT satisfaction	11		
Tool use	11		
Integration with other ICT	2		
Treatment fit	1		
71 outcomes organized into 8 groups	494	494	100

[a]ICT: information and communication technology.

XSL•FO
RenderX

Table 3. List of intervention groups and number of associations found for each individual intervention.

Intervention groups and interventions	Associations found, n	Associations by intervention group, n	Equivalence to the total number of associations (%)
Combined interventions		53	11
Tele-education and virtual reality	11		
Telescreening and telediagnosis	15		
Multiple approaches in teledentistry	25		
Surgical supplementation	2		
Teleconsultation and teleservice		36	7
Telecare	8		
Personalized teleservice through videoconferencing	1		
Teleconsultation	18		
Tele-interconsultation	9		
Telediagnosis		67	14
Telediagnosis (through smartphone)	16		
Telediagnosis (through clinical photography)	15		
Telediagnosis (through digital cameras)	6		
Telediagnosis in general	30		
e-Learning and tele-education		96	19
e-Learning	19		
Digital approaches in dental education	27		
Virtual learning environments	13		
Tele-education (hybrid teaching)	16		
Teleducação (emergency basis)	7		
Tele-education (through apps and SMS)	2		
Tele-education (through videos)	4		
Tele-education (through MOOCs[a])	2		
Telementoring	6		
ICTs [b]		182	37
Digital distraction and virtual reality	24		
Technology-based interventions	54		
eHealth technologies	10		
Educational games (through app)	11		
Educational games (through social media)	2		
mHealth[c]	32		
Social media	21		
Virtual reality	15		
Videos (YouTube and other media)	2		
Digital information and communication technologies	10		
Relaxing music	1		
Telemonitoring		44	9
Telemonitoring in general	27		
Telemonitoring through app	12		

XSL•FO
RenderX

Intervention groups and interventions	Associations found, n	Associations by intervention group, n	Equivalence to the total number of associations (%)
Telemonitoring through email	2		
Telemonitoring through voice message	1		
Telemonitoring through WhatsApp	2		
Telescreening		7	1
Telescreening	5		
Telescreening through smartphone	2		
Artificial intelligence		9	2
Diagnosis via a professional camera picture	1		
Diagnosis via an intraoral camera picture	1		
Diagnosis via a smartphone picture	1		
Artificial intelligence	6		
43 interventions organized into 8 groups	494	494	100

[a]MOOC: Massive Open Online Course.

[b]ICT: information and communication technology.

[c]mHealth: mobile health.

Results

Overview

The comprehensive bibliographic search, conducted in December 2023 to ensure the most up-to-date inventory of articles, retrieved a total of 504 records from the PubMed, Virtual Health Library, and Embase databases (Multimedia Appendix 2). After retrieving the records on Rayyan, we excluded duplicates and screened the remaining papers based on inclusion and exclusion criteria. Full texts of 118 studies were read. Following this, 50 studies were excluded (based on the same criteria) by 2 independent reviewers. The excluded studies and reasons for exclusion can be found in Multimedia Appendix 3. After excluding duplicates from different sources and analyzing titles, abstracts, and full texts, a total of 68 studies were included in the teledentistry EGM. Although the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol is recommended for SRs—and the EGM concept aims for a complementary overall design direction—it was used as a methodological step to enforce the quality of the SRs selected, followed by the application of inclusion and exclusion criteria and full-text readings to check for the eligibility of the studies (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram based on the search strategy and inclusion and exclusion criteria.

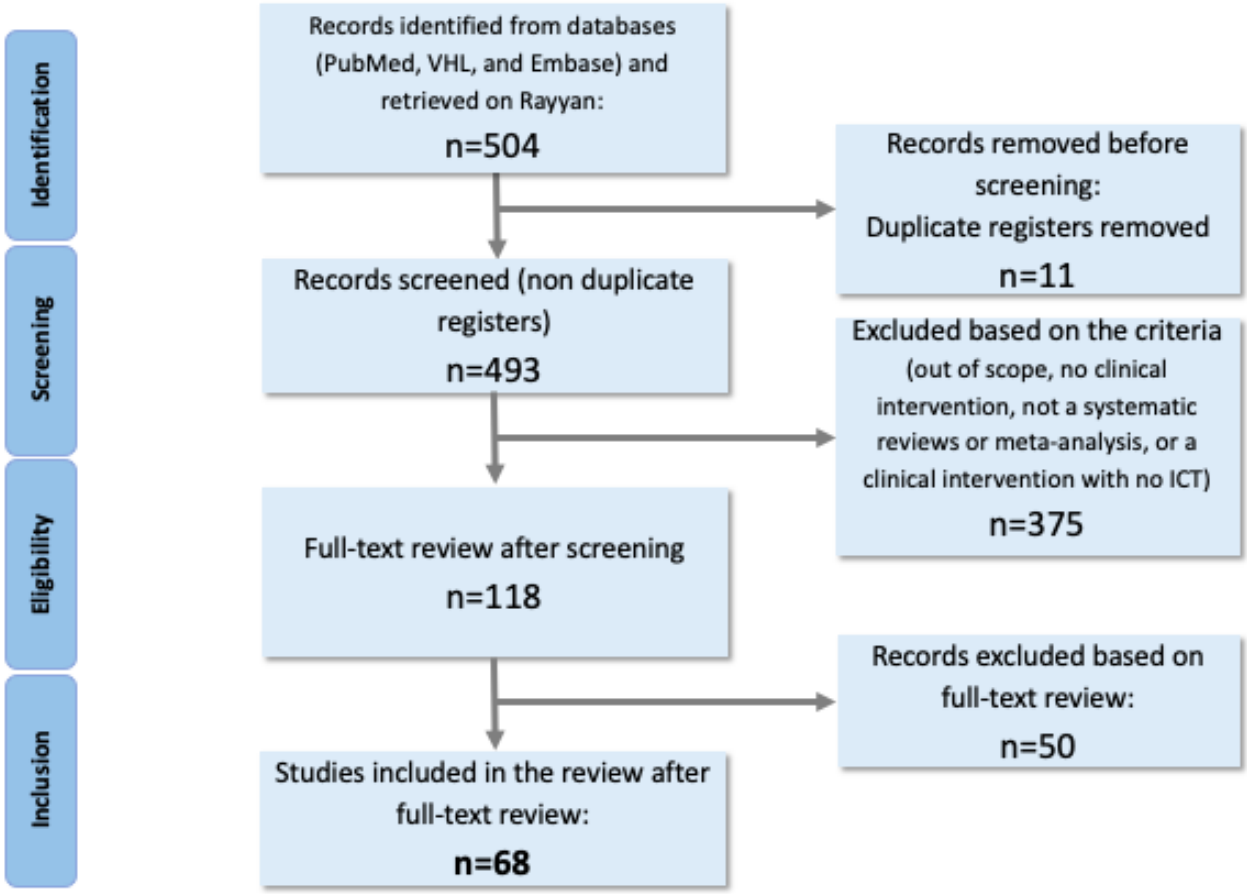

Main Results

The map presents an overview of the evidence on the applicability and effects of teledentistry for health and education outcomes. From a comprehensive bibliographic search, 68 reviews were included in the map (SRs: n=51, 75%; SRs with meta-analysis: n=16, 24%; and meta-analysis: n=1, 1%). Table 4 presents the main positive and potentially positive associations found in the high confidence–level reviews included in the map. Table 4 presents the reviews with a high level of confidence included in the map (following the AMSTAR2 criteria) along with focus populations and main positive and potentially positive effects.

In total, 494 associations were found between the combinations of teledentistry+health and teledentistry+education outcomes. The effect reported by the review study was included for each intervention-outcome association. For most associations, the positive or potentially positive effect of teledentistry was reported (434/494, 87.8% associations total). For other associations, an inconclusive or no effect was reported (39/494, 7.9% associations total). The negative or potentially negative effect of teledentistry was reported for only 0.1% (5/494) of the associations. The effect was not reported for 3.2% (16/494) of the associations.

Figure 2 presents a partial visual representation of the EGM, with the distribution of associations between teledentistry interventions and health and education outcomes. The following outcomes were highlighted: ICT, e-learning, and tele-education.

The map represents the evidence that analyzed the effect of 43 different interventions in teledentistry organized into 8 groups: combined interventions, e-learning and tele-education, teleconsultation and teleservice, telemonitoring, telediagnosis, telescreening, ICTs, and AI—for at least 1 of the 71 health and education outcomes. Furthermore, these outcomes were organized into 8 different groups: diagnosis accuracy, education and professional training, user behavior, clinical practice, patient-centered outcomes, clinical outcomes, health services management, and access to health services. The interactive map—with confidence level, effects, findings, and development information—can be assessed in the study by Portella et al [17].

XSL•FO
RenderX

Table 4. Reviews with a high level of confidence included in the map along with focus populations and main positive and potentially positive effects.

Reference number of the included study, quoted in Multimedia Appendix 2	Focus population	Main positive and potentially positive associations
16	Parents	mHealth[a] was found to have a potentially positive effect on parents' health education
43	Undergraduate students	The combined action of tele-education and virtual reality was found to be potentially positive to the learning process of undergraduate students and led to cost reduction within health services management
48	Dental surgeons and patients receiving orthodontic treatment	Technology-based interventions were found to be positive for patient monitoring and different user behavior–related outcomes (adherence to ICT[b], ICT satisfaction, and tool use) and potentially positive for the management of health emergencies and a number of clinical practice outcomes (when applied to oral surgery, communication among professionals, and professional-patient communication) as well as ICT reliability.
60	Patients with cancer	Telemonitoring via application was proven to improve service quality and access to health services among patients with cancer and to increase the levels of adherence to ICT, ICT satisfaction, and tool use in user-related outcomes.
61	Patients (orthodontic)	mHealth was found to be positive for the expansion of health care and adherence to orthodontic treatment. Potentially positive effects were seen when mHealth was applied to the improvement of clinical outcomes in general, health behavior change among patients, and when applied to oral surgery.
62	Dental students	Tele-education on an emergency basis was proven to positively improve learning and both perception and satisfaction with ICT. It has also shown the potential of improving pedagogical models and the tool use experience.
8	Patients (adults and children)	Multiple approaches in teledentistry have shown to be potentially positive on the regulation and management of health care systems. Telediagnosis through clinical photography and digital cameras was found to have a potentially positive effect on the diagnosis of potentially malignant oral lesions, allied to the potential of telemonitoring on improving patients' prognosis.
11	Parents and guardians	Technology-based interventions were found to be effective in the reduction of gingival inflammation and plaque index levels and the improvement of patients' health education.
17	Patients (children, adolescents, adults, and older individuals)	Either through a smartphone or clinical photography, telediagnosis was shown to have a positive effect on children, adolescents, adults, and older individuals.
19	Patients (children and adults)	The use of artificial intelligence in the diagnosis of carious lesions was found to be positive through both professional camera pictures and intraoral cameras. A potentially positive effect was found for the same application via smartphone pictures.
23	Dental students and dentists	The use of virtual reality reflected positive outcomes for the improvement of clinical skills and educational paths as well as for applications on restorative dentistry and ICT satisfaction levels.
31	Patients (children and adolescents)	e-Learning interventions were found to have a positive effect on the reduction of plaque index and probing depth as well as on patients' health behavior change and education.
33	Patients (children)	The use of digital distraction through virtual reality had a positive effect on children in terms of treatment behavior, anxiety relief, and pain perception reduction.
36	Dental students	Digital approaches in dental education were found to be positive for the following outcomes: learning, assessment performance, professional training, students' stress management, and students' convenience. The level of ICT satisfaction was also positive.
39	Patients receiving orthodontic treatment (children and adolescents)	The results have shown positive effects for eHealth technologies and telemonitoring through apps directed at the improvement of clinical outcomes in general, decreasing the number of appointments and procedure planning in orthodontic treatments. Telemonitoring, in general, has proven to be potentially positive for these same outcomes.
44	Patients	The use of telecare, technology-based interventions, and social media was found to improve oral disease prevention, promote health behavior change, and increase treatment adherence among patients.
55	Patients (adults)	The combination of telescreening with telediagnosis, the use of multiple approaches in teledentistry, telediagnosis (through smartphone or clinical photography), and the use of social media have shown positive effects on the diagnosis of cancerous and potentially malignant oral lesions, on the prevention of oral diseases, in prognosis improvement, and in decision-making in clinical practice.

XSL•FO
RenderX

Reference number of the included study, quoted in Multimedia Appendix 2	Focus population	Main positive and potentially positive associations
52	Patients receiving orthodontic treatment	Telemonitoring in general was found to have a positive effect when applied to oral surgery and on treatment adherence. Technology-based interventions have shown positive potential on changing patients' hygiene habits and potentially positive on treatment adherence, while the use of social media was found to have a positive effect on professional-patient communication and potentially positive on treatment adherence.
37	Patients (adults) and dental students	Digital approaches in dental education, hybrid teaching within tele-education, and technology-based interventions were proven to enhance learning, improve professional training, and increase clinical skills and students' confidence.

[a]mHealth: mobile health.

[b]ICT: information and communication technology.

Figure 2. Partial visual representation of the evidence gap map, with the main groups for both teledentistry interventions and health and education outcomes. The bubble size represents the number of associations found between intervention and outcome groups.

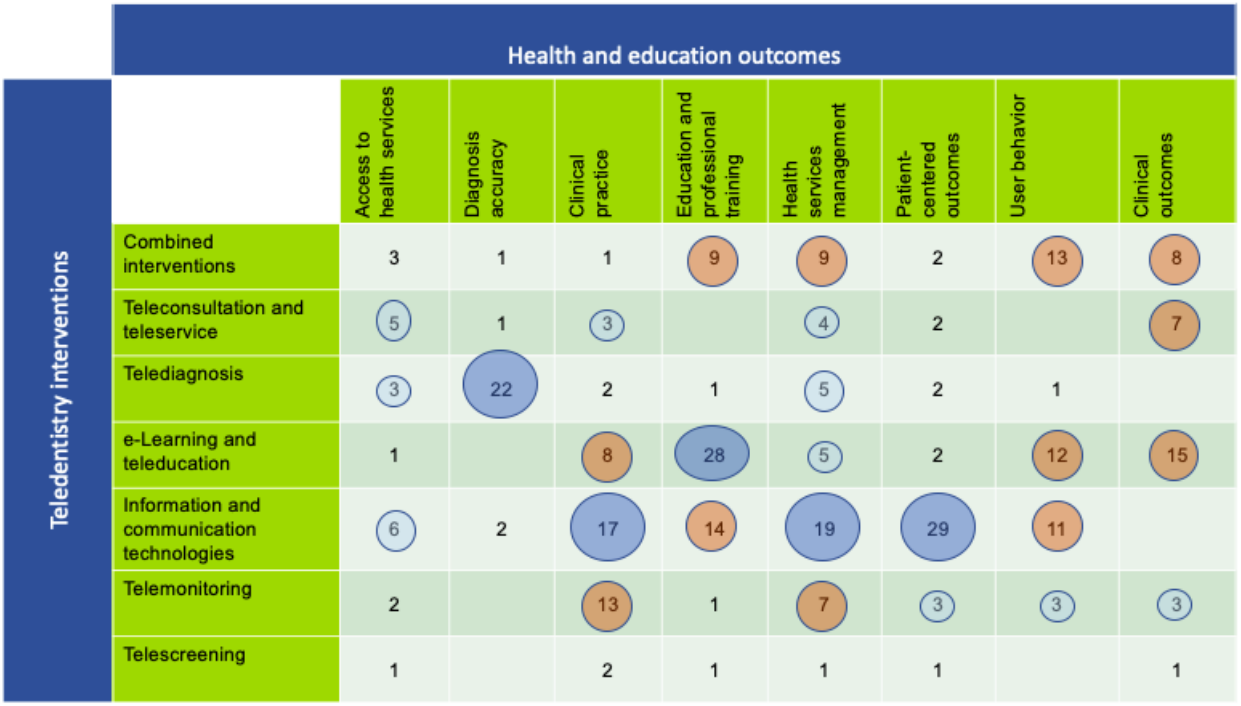

Among the groups of outcomes, the education and professional training group stands out (97/494, 19.6%), followed by the health services management group (60/494, 12.1%) and patient-centered outcomes (74/494, 15%). The following outcomes were highlighted: learning (26/494, 5.3%), clinical skills improvement (24/494, 4.8%), health care extension (21/494, 4.2%), professional training (21/494, 4.2%), oral disease prevention (19/494, 3.8%), time optimization (18/494, 3.6%), and health education (16/494, 3.2%).

As for the 494 associations, the ICTs group stands out with 182 (37%) associations. Of the 494 associations, interventions with e-learning or tele-education, telediagnosis, and combined interventions appear with 96 (19.4%), 67 (13.5%), and 53 (10.7%) associations, respectively. A smaller number of associations were observed for telemonitoring (44/494, 8.9%), teleconsultation and teleservice (36/494, 7.3%), AI (9/494, 1.8%), and telescreening (7/494, 1.4%) interventions.

Regarding the publication countries (in which the publication journal is based), most of the SRs were from the United States (22/68, 32%), Switzerland (12/68, 18%), United Kingdom (10/68, 15%), England (9/68, 13%), along with 13% (9/68) from other countries. The focus countries (the countries and populations where the primary studies included in the SRs were conducted) retrieved from the SRs were mostly concentrated in Europe (42/68, 62%), Asia (40/68, 59%), and North America (36/68, 53%). In total, 28% (19/68) of the SRs did not specify the focus country.

As for the study populations, of the 68 studies, most studies had their interventions aimed at patients: 19 (28%) SRs focused on children and pediatric patients, 17 (25%) on adults, 13 (34%) on patients or "general population," 7 (10%) on adolescents, and 3 (4%) on older adults. Other studies specified the group of patients based on their clinical condition: 1% (1/64) of the studies included patients with cancer, 3% (2/68) of the studies included patients who have undergone surgery, and 10% (7/68) of the studies included patients receiving orthodontic treatment.

XSL•FO
RenderX

Of the 68 studies, dental students were the chosen population in 19 (28%) studies; dental surgeons or dentists in 11 (16%) studies; health professionals and provisioners in 5 (7%) studies; mothers, parents, relatives, or caregivers in 5 (7%) studies; and teachers and academicians in 1 (1%) study.

The most prevalent study designs included in the SRs were observational studies (32/68, 47%), including cohort, case series, cross-sectional, and case-control studies, and randomized controlled trials (30/68, 44%), followed by non-randomized controlled trials (10/68, 15%), experimental or quasi-experimental studies (8/68, 12%), and qualitative studies (5/68, 7%).

Interventions for Outcome Groups

Overview

Among the groups of outcomes, out of 494 associations, the education and professional training group stands out with 97 (20%) associations, followed by the clinical practice group with 75 (15%) associations, and patient-centered outcomes group with 74 (15%) associations. Among the specific outcomes, highlights include learning and clinical skills improvement included in the educational and professional training group, with 26 (5.3%) and 26 (5.3%) associations, respectively; health care extension included in the health services management group, with 21 (4.3%) associations; oral disease prevention included in the clinical outcomes group, with 19 (3.8%) associations; and health education, included in the patient-centered outcomes group, with 16 (%) associations.

As for the interventions, the ICT group stands out with 182 (37%) associations from the 494, with emphasis on the following specific interventions: technology-based interventions (n=54,10.9% associations), mobile health (n=32, 6.5% associations), and digital distraction and virtual reality (n=24, 4.9% associations). The e-learning and tele-education group comes second with 96 (19%) associations out of 494 associations. Telediagnosis in general, telemonitoring in general, and digital approaches in dental education are the other highlights among the specific interventions, with 30 (6%) associations for the first and 27 (5.5%) associations for the 2 others. Figure 3 presents the distribution of outcomes for the teledentistry intervention groups included in the map.

Figure 3. Distribution of outcomes among the 8 teledentistry intervention groups.

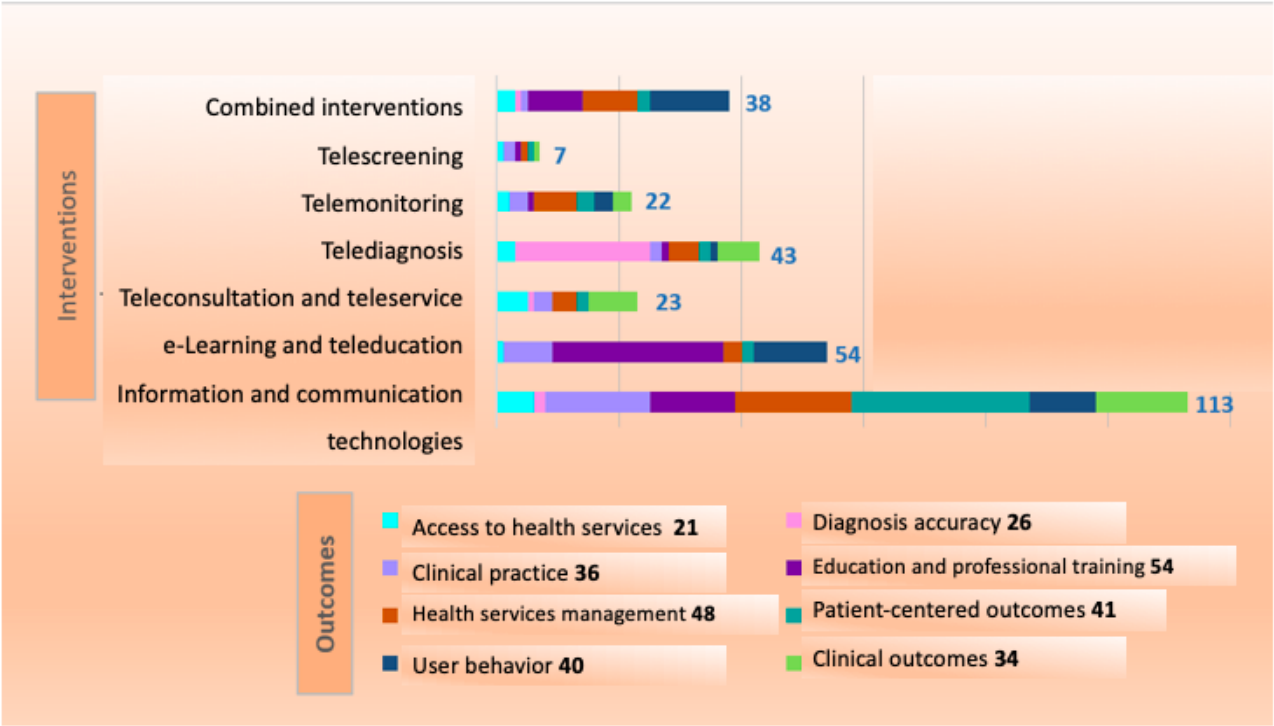

Group 1: Access to Health Services

The 6 outcomes of the access to health services group received 21 (4.3%) associations of the total 494 associations. The most common outcome in the group was *service quality* (with 10/494, 2% associations), and no intervention prevalence was found, indicating a great variety of interventions aimed at the improvement of service provision (eg, teleconsultation, telescreening, telediagnosis, and social media). Of these 10 associations for service quality, 8 (80%) reported a positive or potentially positive effect (confidence level: 7 critically low and 1 high) and 2 (20%) did not present reported effects.

Group 2: Clinical Outcomes

The 9 specific outcomes of the clinical outcomes group received 66 (13.4%) associations of the total 494 associations. The following outcomes were highlighted: oral disease prevention (19/494, 3.8% associations), prognosis improvement (13/494, 2.6% associations), plaque index reduction (11/494, 2.2% associations), and gingival inflammation reduction (10/494, 2% associations). These prevalent outcomes sum up 53 associations of which 51 (96%) present positive or potentially positive effects (1 inconclusive and 1 no effect) and 18 (34%) belong to high confidence–level reviews based on the AMSTAR2 criteria. The most common intervention group associated with clinical

XSL•FO
RenderX

outcomes was ICT, with 27 (5.5%) associations out of 494—22 (81%) of those presenting positive or potentially positive effects and 9 (27%) of those belonging to high confidence–level reviews.

Group 3: Clinical Practice

The 12 specific outcomes of the clinical practice group received 75 (15.2%) associations of the total 494 associations. The most common outcome was applied to orthodontics (14/494, 2.8% associations), followed by decision-making (13/494, 2.6% associations), clinical training (9/494, 1.8% associations), and digital planning (8/494, 1.6% associations). These 4 outcomes are responsible for 44 associations of multiple confidence level reviews (18 critically low, 13 low, 8 high, and 5 moderate). The main intervention was ICT, with 27 associations, of which 23 (85%) reported a positive or potentially positive effect (confidence level: 9 critically low, 8 high, 5 low, and 1 moderate).

Group 4: Diagnosis Accuracy

The 10 outcomes of the diagnosis accuracy group received 44 (8.9%) associations of the total 494 associations. The following outcomes were highlighted: carious lesions (10/494, 2% associations), diagnosis in general (8/494, 1.6% associations), cancer (7/494, 1.2% associations), and potentially malignant lesions (7/494, 1.2% associations). The main intervention was, naturally, telediagnosis, with 30 (6%) associations out of 494, of which 26 (87%) reported a positive or potentially positive effect (confidence level: 18 critically low and 8 high), 2 (7%) were inconclusive (both critically low), and 2 (7%) associations presented a negative report (both critically low).

Group 5: Education and Professional Training

The 9 outcomes delivered by the education and professional training group received 97 (19.6%) associations of the total 494 associations. A prevalence of the following specific outcomes was perceived: learning (26/494, 5.3% associations), clinical skills improvement (24/494, 4.9% associations), and professional training (21/494, 4.3% associations). Although the most common intervention was e-learning and tele-education, with 58 (11.7%) out of 494 associations, a substantial number of associations occurred for the ICT group (24/494, 4.9% associations). By combining these 2, a total of 74 (90%) of the 82 associations reported a positive or potentially positive effect (confidence level: 31 critically low, 21 high, 21 low, and 1 moderate), 7 (9%) were found to be inconclusive (confidence level: 5 critically low, 1 low, and 1 moderate), and 1 (1%) did not present any effects (confidence level: critically low).

Group 6: Health Services Management

The 7 outcomes of the health services management group received 60 (12%) associations of the total 494 associations. The following outcomes were highlighted: health care extension (21/494, 4.3% associations), time optimization (18/494, 3.6% associations), and cost reduction (7/494, 1.4% associations). The main intervention groups were ICT—with 21 (4.3%) associations—and telemonitoring—with 12 (2.4%) associations. Among these 33 associations, 26 (79%) have presented a positive or potentially positive effect (confidence level: 18 critically low, 6 high, and 2 moderate), 3 (9%) have proven to have no effect (confidence level: 2 moderate and 1 high), 2 (6%) have no reported effect (both critically low confidence), and 2 (6%) were found to be inconclusive (both low confidence).

Group 7: Patient-Centered Outcomes

The 9 outcomes of the patient-centered outcomes group received 74 (15%) associations of the total 494 associations. A combination of different specific outcomes was found; of the 74 associations, there were 16 (22%) associations linked to health education, 14 (19%) to health behavior change, 13 (18%) aimed at anxiety reduction, 12 (16%) at pain perception reduction, and 9 (12%) at hygiene habits change. The outstanding intervention group linked to this group was ICT, with 50 (10.1%) out of the 494 associations. In total, 74% (37/50) of these have positive or potentially positive effects (confidence level: 18 critically low, 11 high, 4 moderate, and 1 low), 5 (10%) were inconclusive (confidence level: 3 high and 1 critically low), 5 (10%) have no reported effect (confidence level: 4 critically low and 1 high), and 3 (6%) did not show any effect (confidence level: 2 high and 1 low).

Group 8: User Behavior

The 8 outcomes of the user behavior group received 57 (12%) associations of the total 494 associations. A prevalence of the following specific outcomes within the user behavior group was found: ICT satisfaction (11/57, 19% associations), tool use (11/57, 19% associations), treatment adherence (10/57, 18% associations), and ICT acceptability (9/57, 16% associations). Associations linked to the intervention groups were sparse and mostly distributed between ICT (23/57, 40% associations), e-learning and tele-education (13/57, 23% associations), and combined interventions (13/57, 23% associations). These 3 groups resulted in 49 (9.9%) associations out of 494, of which 44 (89.8%) were positive or potentially positive (confidence level: 24 critically low, 15 high, 4 low, and 1 moderate), 3 (6%) have shown no effect (confidence level: all high), 1 (2%) did not report an effect (confidence level: critically low), and 1 (2%) has presented a potentially negative effect (confidence level: high).

Interventions With Teledentistry

The teledentistry evidence map brings together studies of interventions with 4 combined interventions, 9 e-learning and tele-education actions, 11 ICT approaches aimed at teledentistry, 4 teleconsultation and teleservice actions, 4 telediagnosis, 5 telemonitoring, 2 telescreening, and 2 AI distinct approaches. The effect of these teledentistry interventions was analyzed on 494 associations for 71 health and education outcomes, and the group with the highest number of associations was the ICT, with 182 associations (37%).

The most researched approach on teledentistry reported by the map was technology-based interventions, with 54 unit interventions aimed at a great range of outcomes but mostly at user behavior, with 12 (22%) associations (11 of those reporting positive or potentially positive effects); health services management, with 10 (19%) associations (9 of which presenting positive or potentially positive effects); education and professional training, with 8 (15%) associations (all 8 of those reporting positive effect); and clinical practice, with 8 (15%)

XSL•FO
RenderX

associations (all 8 reporting positive or potentially positive effects). The second most researched approach on teledentistry was mobile health, with 32 unit interventions also directed at a considerable variety of outcomes but more substantially focused on clinical outcomes, with 14 (44%) associations (10 reporting positive or potentially positive effects), and patient-centered outcomes, with 9 (28%) associations (5 of those reporting positive or potentially positive effects, 1 with no given effect, 1 not reported, and 2 inconclusive).

Most of the studies (67/68, 99%) have presented at least 1 positive or potentially positive association with teledentistry interventions. The only exception refers to the inconclusive results of a meta-analysis focused on digital technology distraction for acute pain in children, to which the given AMSTAR2 confidence level was critically low. Among the studies that have achieved a high level of confidence on AMSTAR2, all have reported at least 1 positive or potentially positive association with regard to teledentistry interventions.

Discussion

Principal Findings

This EGM is based on 68 (SRs: n=52, 77%; SRs with meta-analysis: n=15, 22%; and metanalysis: n=1, 2%) studies that met the established inclusion criteria, and it provides a broad overview of available evidence on teledentistry interventions, applied to 64 outcomes in health and education. The majority of focus countries identified in the SRs are located in Europe (42/68, 62%), Asia (40/68, 59%), and North America (36/68, 53%). This suggests a need to study regions where online dental services could be beneficial, to address potential social and economic limitations in those areas.

While a great variety of study designs was found among the primary studies presented in the included reviews, there was a substantial prevalence of randomized controlled trials (31/68, 46%), followed by a range of observational studies (20/68, 29%). This finding endorses a higher level of design reliability, ultimately giving more scientific strength to the EGM itself. The map shows the volume of available research and highlights areas where the interventions have presented positive or potentially positive effects, signaling a broad range of not only research but also service implementation possibilities [89].

By restricting the results obtained by “high confidence level” reviews (based on the AMSTAR2 criteria) and graphically represented in the map, the incorporation of teledentistry into the teaching-service axis proposed by Chen et al [10] becomes a strong evidence-based material, especially because these high confidence–level reviews pointed to 121 (89%) out of 136 positive or potentially positive associations for teledentistry interventions [89]. The use of the AMSTAR2 tool refines the selection of articles through its specific and rigorous methodology, which focuses on the structural aspects of study designs and protocols. While this may exclude some articles with relatively good methodology and results, using AMSTAR2 is still crucial as it strengthens the overall EGM. Therefore, the EGM methodology enables the transformation of the qualified evidence found into practice in a markable accessible format, creating an important conceptual connection with evidence-based best practices. Its interface tends to be more visually compact and dynamic when compared to other systematic study designs, especially for professionals outside the academic and research realm, such as policy makers and service managers, thereby enhancing outcomes and creating cross-references that serve as a reliable index to understanding and planning optimized service offering to the population.

For students, teledentistry applications enable the improvement of the learning process, convenience, clinical skills, and professional training as well as increase the levels of perception, confidence, and satisfaction with ICT. Furthermore, teledentistry enhances assessment performance and improves pedagogical models and dental curricula. At the professional level, it helps dentists refine and optimize their clinical practice, manage emergencies, and plan procedures. Evidence indicates distinct benefits in orthodontics, restorative dentistry, and oral surgery practices. Better time management and team organization (generating less burden) are also among the operational benefits for dental services [21-25].

For patients and caregivers, teledentistry improves health education and promotion as well as the quality of service provision (from cost reduction to professional-patient communication). It can increase the levels of adherence to dental treatments and clinical outcomes in general, endorsing health behavior change, managing anxiety and pain perception, and favoring better prognosis for all types of patients [26-38]. Within health services, teledentistry promotes reductions in cost and number of appointments, improves time and emergency management, and facilitates the access to specialists [22,35,39].

Updating and improving requirements and compliances become essential for the consistent and broad development of teledentistry and for improving oral health provision [91]. As for the education and training of dental students, Kröplin et al [92] endorsed the need for digital health topics to be taught at universities in a transdisciplinary manner, developing the curricula and enabling the proper use of digital health platforms by creating an adequate scenario for the progressive use of ICTs in general.

Online learning methodology attached to a hybrid approach to care is an attractive, flexible, and convenient option for learners, with lower costs, larger geographic coverage, and easy access to information. These gains apply for undergraduate students, graduate students, and dentists in continuous medical education programs, endorsing tele-education as an alternative to keeping dental professionals up to date with the latest advancements in their field and, thus, benefiting the patients they serve [92].

The data collected demonstrate that this experience is a safe practice tool that enhances both educational and professional development. It aligns with the growing digital use and literacy in society as well as the trend in health services toward hybrid care (synchronized in-person and online offerings). This approach provides a more accessible and inclusive pathway to evidence-based practices [16].

Another important aspect to be considered is the regulatory mainframe. The speed of social, cultural, and health changes in

XSL•FO
RenderX

society is usually faster than the effort of legislators and policy makers to address new forms of offering health and dental services, which is understandable, due to the time needed for a coherent decision path, based on collecting reliable data, analysis, broad discussion, and finally document writing and approval. However, security concerns and the demand for guidelines and protocols must be taken into account, as stated by Hung et al [91]. Those points can lead to some friction in the patient-provider relationship. Therefore, private offices and public services need to be aware of the impacts and consequences of offering virtual service solutions, following consistent operational protocols, and using safe virtual environments. Thus, to provide and extend dental care through teledentistry, clinician and patient training must rely on national laws and regulations; challenges such as the inability to perform diagnostic tests and the potential for misdiagnosis due to variations in image and video quality must be carefully considered [91].

In addition, it is important to consider the complex landscape across each continent, where various countries—and even their provinces and states—interact with one another. This creates a paradox of having a wide range of laws while also addressing specific local needs, along with somewhat independent health care systems. This is observed, for instance, in Europe (in the European Union and each member country) as well as in the United States (that has a good level of legal independence among its states). In this context, it is observed that the major updated national laws (or regulations) in several parts of the world address individual and health data management but in a more generalized approach, such as the Health Insurance Portability and Accountability Act (United States), the General Data Protection Regulation Act (European Union), the Data Protection General Law (Brazil), and the Act on the Protection of Personal Information (Japan), among others. Specific segments of the health care system (including dentistry) have to create their own guidelines in an effort to align with those bigger legislations [91,93-102]. One of Pan American Health Organization's primary goals in developing and publishing the EGM is to provide consistent and clear information for policy makers and public service managers. This includes the ability to filter results by different countries and populations, with the aim of improving dental services for citizens and communities, ultimately creating tangible benefits for society as a whole [103].

From the patient's perspective, it is a simpler train of thought, focused on the expectation that a reliable, swift, and effective teledentistry treatment could be offered. There is also a perception that health care providers and governments already observed and considered all legal and safety aspects. In summary, the high-confidence SRs included in this EGM (cited in Table 4 and referenced in Multimedia Appendix 2), demonstrate the benefits of teledentistry interventions in favor of the fulfillment of needs perceived by several different populations and focus countries. Positive (or potentially positive) associations regarding the focus populations (parents, patients, and dental students) and outcomes, such as greater adherence to oral cancer treatments, gingival inflammation reduction, better orthodontic treatment outcome, diagnosis of oral lesions, and learning and skills improvement among dental students are presented in Table 4. These examples demonstrate the real impact of teledentistry through evidence-based data, especially toward clinical and oral health benefits in individuals (patients), serving as an important reference to policy makers and dental service managers.

Limitations

The results of this study are aligned with those of Yan et al [34], Uhrin et al [37], Erdilek et al [21], Dedeilia et al [24], and Bandiaky et al [25] regarding the effectiveness of teledentistry interventions toward health and education outcomes. As the methodology does not exclude duplicates among the primary studies individually included in the SRs, it is important to acknowledge the risk of amplification of evidence reported for a specific association (with a consequent increase in the bubble size represented in the interactive map, working as a nonintentional reinforcement visualization). Moreover, although a growth in the AMSTAR2 confidence level is perceived in most recently published SRs, the prevalence of "critically low" quality studies included in this EGM has to be considered to guide the development of actions and policies as well as the criteria of new teledentistry-themed reviews. Protocols attending to data privacy and global legal support are also essential to strengthen the great range of teledentistry applications and to acknowledge the importance of maintaining face-to-face health provision [95-98,101,104].

The reviews analyzed in this study either did not explore regulation or did not consider it as a core aspect of analysis. Therefore, more studies should be conducted in the future, specifically to understand those interferences and outcomes. Furthermore, although SRs provide a gold-standard panorama for a certain subject, other study designs and recent primary studies can reflect different experiences with teledentistry that have not been taken into consideration for the map development.

Conclusions

The results and graphic representations found in this EGM serve as a user-friendly instrument and a dynamic interface for the comprehension of the applications of teledentistry toward health and education outcomes, thus illustrating the answers to the proposed research question. Navigating through the interactive map allows users to understand where the reliable associations can be found or where a lack of high-quality evidence is perceived, directing the search to specific interventions, outcomes, focus populations, focus countries, etc, as needed.

The publication and further analysis of the teledentistry EGM results provide an accessible pathway to evidence-based policy making, pedagogical enhancement, and patient care. The great variability found between interventions and health and educational outcomes is noticeable and could endorse the inclusion of ICT in modern dentistry and teaching. This scenario underscores the potential of teledentistry as a comprehensive tool for improving the teaching-learning process and enhancing clinical skills within dental curricula as well as for qualifying health service provision.

XSL•FO
RenderX

Acknowledgments

This study was carried out with financial support from the National Council for Scientific and Technological Development (Cnpq) – Federal funds from the Brazilian government granted by the Process: 422896/2021-7; CNPq/MCTI/FNDCT Call No. 18/2021—Track A—Emerging Groups. The main author (JNdSM) also acknowledges the financial support from the Coordination of Superior Level Staff Improvement, Brazil. This map applies the evidence gap map methodology supported by the Latin American and Caribbean Center on Health Sciences (BIREME/OPAS/OMS), in partnership with the Telehealth Center FOUSP-SAITE (School of Dentistry, University of São Paulo).

Authors' Contributions

JNdSM coordinated the group as first author, defined the bibliographic Medical Subject Headings terms and inclusion and exclusion criteria, worked on the selection of studies through Rayyan, worked on the characterization matrix of the studies on Microsoft Excel, worked on the AMSTAR2 (A Measurement Tool to Assess Systematic Reviews) analysis of the included studies, and in the development of the conclusion drawn by the evidence gap map. MRP supported group coordination, coworked on the selection of studies through Rayyan, coworked on the characterization matrix of the studies on Microsoft Excel, coworked on the AMSTAR2 analysis of the included studies, and coworked in the discussion section and revision of the final text. IA coworked on the selection of studies through Rayyan, coworked on the characterization matrix of the studies on Microsoft Excel, and coworked on the AMSTAR2 analysis of the included studies. DG coworked on the selection of studies through Rayyan, coworked on the characterization matrix of the studies on Microsoft Excel, and coworked on the AMSTAR2 analysis of the included studies. CH coworked on the selection of studies through Rayyan, coworked on the characterization matrix of the studies on Microsoft Excel, and coworked on the AMSTAR2 analysis of the included studies. AEH coordinated between NuTes and Latin American and Caribbean Center for Health Sciences Information (BIREME) and general project supervision. CVMA received methodological and technical support from BIREME, coordination and development of the bibliographic search and establishment of Medical Subject Headings terms, coordination of the studies' inclusion in Rayyan, development of the pilot Microsoft Excel matrix, and development of the interactive map through Tableau.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 30 KB - jmir_v26i1e60590_app1.docx]

Multimedia Appendix 2
References for the included studies.
[DOCX File , 48 KB - jmir_v26i1e60590_app2.docx]

Multimedia Appendix 3
Studies excluded after full-text reading and reasons for exclusion.
[DOCX File , 36 KB - jmir_v26i1e60590_app3.docx]

References

1. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016 Dec 15. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2024-01-17]
2. WHO-ITU global standard for accessibility of telehealth services. World Health Organization. 2022. URL: https://iris.who.int/bitstream/handle/10665/356160/9789240050464-eng.pdf?sequence=1 [accessed 2024-01-17]
3. McLaren P. Telemedicine and telecare: what can it offer mental health services? Adv Psychiatr Treat 2018 Jan 02;9(1):54-61. [doi: 10.1192/apt.9.1.54]
4. Haddad AE. Experiência Brasileira do Programa Nacional Telessaude Brasil. In: Monteiro A, editor. Goldbook: Experiências em Telemedicina e Telessaúde. Rio de Janeiro, Brazil: Universidade do Estado do Rio de Janeiro; 2012:12-42.
5. Inovação e tecnologia para soluções integradas em telessaúde, teleodontologia e saúde digital. Núcleo de Telessaúde FOUSP-SAITE. 2024. URL: https://sites.usp.br/telessaude/ [accessed 2024-01-20]
6. WHO Global Observatory for eHealth. Telemedicine: opportunities and developments in Member States: report on the second global survey on eHealth. World Health Organization. 2010. URL: https://iris.who.int/handle/10665/44497 [accessed 2024-11-07]

XSL•FO
RenderX

7. Haddad AE, Skelton-Macedo MC. Teleodontologia na formação dos profissionais de saúde. In: Monteiro A, editor. Goldbook: Experiências de Telemedicina e Telessaúde. Rio de Janeiro, Brazil: Universidade do Estado do Rio de Janeiro; 2012:173-206.
8. Jampani ND, Nutalapati R, Dontula BS, Boyapati R. Applications of teledentistry: a literature review and update. J Int Soc Prev Community Dent 2011 Jul;1(2):37-44 [FREE Full text] [doi: 10.4103/2231-0762.97695] [Medline: 24478952]
9. Bavaresco CS, Hauser L, Haddad AE, Harzheim E. Impact of teleconsultations on the conduct of oral health teams in the Telehealth Brazil Networks Programme. Braz Oral Res 2020;34:e011 [FREE Full text] [doi: 10.1590/1807-3107bor-2020.vol34.0011] [Medline: 32130361]
10. Chen JW, Hobdell MH, Dunn K, Johnson KA, Zhang J. Teledentistry and its use in dental education. J Am Dent Assoc 2003 Mar;134(3):342-346. [doi: 10.14219/jada.archive.2003.0164] [Medline: 12699048]
11. Caldarelli PG, Haddad AE. Teleodontologia em consonância com as Diretrizes Curriculares Nacionais no desenvolvimento de competências profissionais. Rev ABENO 2016 Jul 12;16(2):25-32. [doi: 10.30979/rev.abeno.v16i2.264]
12. Amin M, Lai JY, Lindauer PA, McPherson K, Qari H. Should dental schools adopt teledentistry in their curricula? Two viewpoints. J Dent Educ 2021 Jul 15;85(7):1238-1244. [doi: 10.1002/jdd.12614] [Medline: 33860537]
13. Torres-Pereira C, Possebon RS, Simões A, Bortoluzzi MC, Leão JC, Giovanini AF, et al. Email for distance diagnosis of oral diseases: a preliminary study of teledentistry. J Telemed Telecare 2008 Dec 01;14(8):435-438. [doi: 10.1258/jtt.2008.080510] [Medline: 19047454]
14. Olival AR, Curvino M, Faria M, Groisman S. Teleodontologia: um novo horizonte. Perionews 2008;2(4):311-315 https://pesquisa.bvsalud.org/portal/resource/pt/lil-688172.
15. Evidence gap maps. International Initiative for Impact Evaluation. URL: https://www.3ieimpact.org/evidence-hub/evidence-gap-maps [accessed 2024-01-17]
16. BIREME/OPAS lança mapas de evidências sobre aplicabilidade clínica das práticas integrativas e complementares de saúde. Conselho Nacional de Secretários de Saúde. 2020 Nov 6. URL: https://www.conass.org.br/bireme-opas-lanca-mapas-de-evidencias-sobre-aplicabilidade-clinica-das-praticas-integrativas-e-complementares-de-saude/ [accessed 2024-01-20]
17. Portella CF, Ghelman R, Abdala CV, Schveitzer MC. Evidence map on the contributions of traditional, complementary and integrative medicines for health care in times of COVID-19. Integr Med Res 2020 Sep;9(3):100473 [FREE Full text] [doi: 10.1016/j.imr.2020.100473] [Medline: 32766112]
18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
19. What is AMSTAR. AMSTAR. URL: https://amstar.ca/About_Amstar.php [accessed 2024-01-20]
20. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
21. Erdilek D, Gümüştaş B, Güray Efes B. Digitalization era of dental education: a systematic review. Dent Med Probl 2023 Sep 25;60(3):513-525 [FREE Full text] [doi: 10.17219/dmp/156804] [Medline: 37747703]
22. George PP, Papachristou N, Belisario JM, Wang W, Wark PA, Cotic Z, et al. Online eLearning for undergraduates in health professions: a systematic review of the impact on knowledge, skills, attitudes and satisfaction. J Glob Health 2014 Jun;4(1):010406 [FREE Full text] [doi: 10.7189/jogh.04.010406] [Medline: 24976965]
23. Santos GN, da Silva HE, Leite AF, Mesquita CR, Figueiredo PT, Stefani CM, et al. The scope of dental education during COVID-19 pandemic: a systematic review. J Dent Educ 2021 Jul 23;85(7):1287-1300 [FREE Full text] [doi: 10.1002/jdd.12587] [Medline: 33759214]
24. Dedeilia A, Papapanou M, Papadopoulos AN, Karela NR, Androutsou A, Mitsopoulou D, et al. Health worker education during the COVID-19 pandemic: global disruption, responses and lessons for the future-a systematic review and meta-analysis. Hum Resour Health 2023 Feb 24;21(1):13 [FREE Full text] [doi: 10.1186/s12960-023-00799-4] [Medline: 36829158]
25. Bandiaky ON, Lopez S, Hamon L, Clouet R, Soueidan A, Le Guehennec L. Impact of haptic simulators in preclinical dental education: a systematic review. J Dent Educ 2024 Mar 03;88(3):366-379. [doi: 10.1002/jdd.13426] [Medline: 38044266]
26. de Lima TM, Moura AB, Bezerra PM, Valença AM, Vieira TI, Santiago BM, et al. Accuracy of remote examination for detecting potentially malignant oral lesions: a systematic review and meta-analysis. Telemed J E Health 2024 Feb 01;30(2):381-392. [doi: 10.1089/tmj.2023.0096] [Medline: 37651222]
27. Wang K, Yu KF, Liu P, Lee GH, Wong MC. Can mHealth promotion for parents help to improve their children's oral health? A systematic review. J Dent 2022 Aug;123:104185 [FREE Full text] [doi: 10.1016/j.jdent.2022.104185] [Medline: 35691452]
28. da Silva HE, Santos GN, Ferreira Leite A, Mesquita CR, de Souza Figueiredo PT, Miron Stefani C, et al. The feasibility of telehealth in the monitoring of head and neck cancer patients: a systematic review on remote technology, user adherence, user satisfaction, and quality of life. Support Care Cancer 2022 Oct 06;30(10):8391-8404. [doi: 10.1007/s00520-022-07109-z] [Medline: 35524146]

XSL•FO
RenderX

29. Al-Moghrabi D, Alkadhimi A, Tsichlaki A, Pandis N, Fleming PS. The influence of mobile applications and social media-based interventions in producing behavior change among orthodontic patients: a systematic review and meta-analysis. Am J Orthod Dentofacial Orthop 2022 Mar;161(3):338-354. [doi: 10.1016/j.ajodo.2021.09.009] [Medline: 34736817]
30. Ajay K, Azevedo LB, Haste A, Morris AJ, Giles E, Gopu BP, et al. App-based oral health promotion interventions on modifiable risk factors associated with early childhood caries: a systematic review. Front Oral Health 2023 Mar 10;4:1125070 [FREE Full text] [doi: 10.3389/froh.2023.1125070] [Medline: 36968137]
31. Priyank H, Verma A, Zama Khan DU, Prakash Rai N, Kalburgi V, Singh S. Comparative evaluation of dental caries score between teledentistry examination and clinical examination: a systematic review and meta-analysis. Cureus 2023 Jul;15(7):e42414 [FREE Full text] [doi: 10.7759/cureus.42414] [Medline: 37637546]
32. Moharrami M, Farmer J, Singhal S, Watson E, Glogauer M, Johnson AE, et al. Detecting dental caries on oral photographs using artificial intelligence: a systematic review. Oral Dis 2024 May;30(4):1765-1783. [doi: 10.1111/odi.14659] [Medline: 37392423]
33. Atif M, Tewari N, Saji S, Srivastav S, Rahul M. Effectiveness of various methods of educating children and adolescents for the maintenance of oral health: a systematic review of randomized controlled trials. Int J Paediatr Dent 2024 May 15;34(3):229-245. [doi: 10.1111/ipd.13125] [Medline: 37840214]
34. Yan X, Yan Y, Cao M, Xie W, O'Connor S, Lee JJ, et al. Effectiveness of virtual reality distraction interventions to reduce dental anxiety in paediatric patients: a systematic review and meta-analysis. J Dent 2023 May;132:104455. [doi: 10.1016/j.jdent.2023.104455] [Medline: 36842625]
35. Torres DK, Santos MC, Normando D. Is teledentistry effective to monitor the evolution of orthodontic treatment? A systematic review and meta-analysis. Dental Press J Orthod 2023;28(4):e2322195 [FREE Full text] [doi: 10.1590/2177-6709.28.4.e2322195.oar] [Medline: 37729285]
36. de Oliveira Júnior AJ, Oliveira JM, Bretz YP, Mialhe FL. Online social networks for prevention and promotion of oral health: a systematic review. Can J Dent Hyg 2023 Jun;57(2):83-97 [FREE Full text] [Medline: 37464997]
37. Uhrin E, Domokos Z, Czumbel LM, Kói T, Hegyi P, Hermann P, et al. Teledentistry: a future solution in the diagnosis of oral lesions: diagnostic meta-analysis and systematic review. Telemed J E Health 2023 Nov 01;29(11):1591-1600 [FREE Full text] [doi: 10.1089/tmj.2022.0426] [Medline: 36976779]
38. Wafaie K, Rizk MZ, Basyouni ME, Daniel B, Mohammed H. Tele-orthodontics and sensor-based technologies: a systematic review of interventions that monitor and improve compliance of orthodontic patients. Eur J Orthod 2023 Jul 31;45(4):450-461. [doi: 10.1093/ejo/cjad004] [Medline: 37132630]
39. Rouanet F, Masucci C, Khorn B, Oueiss A, Dridi SM, Charavet C. [Relevance of teleorthodontic tools: a systematic review of the literature]. Orthod Fr 2022 Dec 01;93(4):353-375. [doi: 10.1684/orthodfr.2022.104] [Medline: 36718756]
40. Hao T, Pang J, Liu Q, Xin P. A systematic review and network meta-analysis of virtual reality, audiovisuals and music interventions for reducing dental anxiety related to tooth extraction. BMC Oral Health 2023 Sep 22;23(1):684 [FREE Full text] [doi: 10.1186/s12903-023-03407-y] [Medline: 37735362]
41. Estai M, Kanagasingam Y, Tennant M, Bunt S. A systematic review of the research evidence for the benefits of teledentistry. J Telemed Telecare 2018 Apr;24(3):147-156. [doi: 10.1177/1357633X16689433] [Medline: 28118778]
42. Chau RC, Thu KM, Chaurasia A, Hsung RT, Lam WY. A systematic review of the use of mHealth in oral health education among older adults. Dent J (Basel) 2023 Aug 08;11(8):189 [FREE Full text] [doi: 10.3390/dj11080189] [Medline: 37623285]
43. Cunningham A, McPolin O, Fallis R, Coyle C, Best P, McKenna G. A systematic review of the use of virtual reality or dental smartphone applications as interventions for management of paediatric dental anxiety. BMC Oral Health 2021 May 07;21(1):244 [FREE Full text] [doi: 10.1186/s12903-021-01602-3] [Medline: 33962624]
44. Dreesmann NJ, Su H, Thompson HJ. A systematic review of virtual reality therapeutics for acute pain management. Pain Manag Nurs 2022 Oct;23(5):672-681 [FREE Full text] [doi: 10.1016/j.pmn.2022.05.004] [Medline: 35868974]
45. Alabdullah JH, Daniel SJ. A systematic review on the validity of teledentistry. Telemed J E Health 2018 Aug;24(8):639-648. [doi: 10.1089/tmj.2017.0132] [Medline: 29303678]
46. Inês Meurer M, Caffery LJ, Bradford NK, Smith AC. Accuracy of dental images for the diagnosis of dental caries and enamel defects in children and adolescents: a systematic review. J Telemed Telecare 2015 Dec;21(8):449-458. [doi: 10.1177/1357633X15605225] [Medline: 26377122]
47. Botelho MG, Agrawal KR, Bornstein MM. An systematic review of e-learning outcomes in undergraduate dental radiology curricula-levels of learning and implications for researchers and curriculum planners. Dentomaxillofac Radiol 2019 Jan;48(1):20180027 [FREE Full text] [doi: 10.1259/dmfr.20180027] [Medline: 30028185]
48. AbuSalim S, Zakaria N, Islam MR, Kumar G, Mokhtar N, Abdulkadir SJ. Analysis of deep learning techniques for dental informatics: a systematic literature review. Healthcare (Basel) 2022 Sep 28;10(10):1892 [FREE Full text] [doi: 10.3390/healthcare10101892] [Medline: 36292339]
49. Monaghesh E, Negahdari R, Samad-Soltani T. Application of virtual reality in dental implants: a systematic review. BMC Oral Health 2023 Aug 28;23(1):603 [FREE Full text] [doi: 10.1186/s12903-023-03290-7] [Medline: 37641060]
50. Fortich Mesa N, Hoyos Hoyos V. Applications of teledentistry in dental practice: a systematic review. Rev Fac Odontol Univ Antioq 2020 Jul 27;32(1). [doi: 10.17533/udea.rfo.v32n1a8]

XSL•FO
RenderX

51. Gujjar KR, van Wijk A, Kumar R, de Jongh A. Are technology-based interventions effective in reducing dental anxiety in children and adults? A systematic review. J Evid Based Dent Pract 2019 Jun;19(2):140-155. [doi: 10.1016/j.jebdp.2019.01.009] [Medline: 31326046]
52. Lin GS, Koh SH, Ter KZ, Lim CW, Sultana S, Tan WW. Awareness, knowledge, attitude, and practice of teledentistry among dental practitioners during COVID-19: a systematic review and meta-analysis. Medicina (Kaunas) 2022 Jan 15;58(1):130 [FREE Full text] [doi: 10.3390/medicina58010130] [Medline: 35056438]
53. Lima MS, Tonial FG, Basei E, Brew MC, Grossmann E, Haddad AE, et al. Effectiveness of the distance learning strategy applied to orthodontics education: a systematic literature review. Telemed J E Health 2019 Dec;25(12):1134-1143. [doi: 10.1089/tmj.2018.0286] [Medline: 31566489]
54. Aquilanti L, Santarelli A, Mascitti M, Procaccini M, Rappelli G. Dental care access and the elderly: what is the role of teledentistry? A systematic review. Int J Environ Res Public Health 2020 Dec 04;17(23):9053 [FREE Full text] [doi: 10.3390/ijerph17239053] [Medline: 33291719]
55. Estai M, Bunt S, Kanagasingam Y, Kruger E, Tennant M. Diagnostic accuracy of teledentistry in the detection of dental caries: a systematic review. J Evid Based Dent Pract 2016 Sep;16(3):161-172. [doi: 10.1016/j.jebdp.2016.08.003] [Medline: 27855831]
56. Gates M, Hartling L, Shulhan-Kilroy J, MacGregor T, Guitard S, Wingert A, et al. Digital technology distraction for acute pain in children: a meta-analysis. Pediatrics 2020 Feb;145(2):e20191139. [doi: 10.1542/peds.2019-1139] [Medline: 31969473]
57. Zitzmann NU, Matthisson L, Ohla H, Joda T. Digital undergraduate education in dentistry: a systematic review. Int J Environ Res Public Health 2020 May 07;17(9):3269 [FREE Full text] [doi: 10.3390/ijerph17093269] [Medline: 32392877]
58. Patano A, Cirulli N, Beretta M, Plantamura P, Inchingolo AD, Inchingolo AM, et al. Education technology in orthodontics and paediatric dentistry during the COVID-19 pandemic: a systematic review. Int J Environ Res Public Health 2021 Jun 04;18(11):6056 [FREE Full text] [doi: 10.3390/ijerph18116056] [Medline: 34199882]
59. Quek LH, Tan AJ, Sim MJ, Ignacio J, Harder N, Lamb A, et al. Educational escape rooms for healthcare students: a systematic review. Nurse Educ Today 2024 Jan;132:106004 [FREE Full text] [doi: 10.1016/j.nedt.2023.106004] [Medline: 37924674]
60. Toniazzo MP, Nodari D, Muniz FW, Weidlich P. Effect of mHealth in improving oral hygiene: a systematic review with meta-analysis. J Clin Periodontol 2019 Mar;46(3):297-309. [doi: 10.1111/jcpe.13083] [Medline: 30761580]
61. Sangalli L, Alessandri-Bonetti A, Dalessandri D. Effectiveness of dental monitoring system in orthodontics: a systematic review. J Orthod 2024 Mar;51(1):28-40. [doi: 10.1177/14653125231178040] [Medline: 37278017]
62. Santos GN, Leite AF, Figueiredo PT, Pimentel NM, Flores-Mir C, de Melo NS, et al. Effectiveness of e-learning in oral radiology education: a systematic review. J Dent Educ 2016 Sep;80(9):1126-1139. [Medline: 27587580]
63. Patil S, Bhandi S, Awan KH, Licari FW, Di Blasio M, Ronsivalle V, et al. Effectiveness of haptic feedback devices in preclinical training of dental students-a systematic review. BMC Oral Health 2023 Oct 10;23(1):739 [FREE Full text] [doi: 10.1186/s12903-023-03410-3] [Medline: 37817151]
64. Moussa R, Alghazaly A, Althagafi N, Eshky R, Borzangy S. Effectiveness of virtual reality and interactive simulators on dental education outcomes: systematic review. Eur J Dent 2022 Feb;16(1):14-31 [FREE Full text] [doi: 10.1055/s-0041-1731837] [Medline: 34428851]
65. Custódio NB, Costa FD, Cademartori MG, da Costa VP, Goettems ML. Effectiveness of virtual reality glasses as a distraction for children during dental care. Pediatr Dent 2020 Mar 15;42(2):93-102. [Medline: 32276674]
66. Choi EM, Park BY, Noh HJ. Efficacy of mobile health care in patients undergoing fixed orthodontic treatment: a systematic review. Int J Dent Hyg 2021 Feb;19(1):29-38. [doi: 10.1111/idh.12459] [Medline: 32794341]
67. Chan SC, Choa G, Kelly J, Maru D, Rashid MA. Implementation of virtual OSCE in health professions education: a systematic review. Med Educ 2023 Sep;57(9):833-843. [doi: 10.1111/medu.15089] [Medline: 37080907]
68. Rosa A, Pujia AM, Docimo R, Arcuri C. Managing dental phobia in children with the use of virtual reality: a systematic review of the current literature. Children (Basel) 2023 Oct 30;10(11):1763 [FREE Full text] [doi: 10.3390/children10111763] [Medline: 38002854]
69. Montasser MA. Mobile app and social media-based interventions may have a positive influence on the behavior of orthodontic patients. J Evid Based Dent Pract 2022 Dec;22(4):101784. [doi: 10.1016/j.jebdp.2022.101784] [Medline: 36494105]
70. Velásquez Ron B, Mosquera Cisneros V, Pazmiño Troncoso P, Rodríguez Tates M, Alvares Lalvay E, Chauca Bajaña L, et al. Monitoring of awake bruxism by intelligent app. F1000Res 2022;11:479 [FREE Full text] [doi: 10.12688/f1000research.110673.2] [Medline: 36606116]
71. Emami E, Harnagea H, Shrivastava R, Ahmadi M, Giraudeau N. Patient satisfaction with e-oral health care in rural and remote settings: a systematic review. Syst Rev 2022 Oct 29;11(1):234 [FREE Full text] [doi: 10.1186/s13643-022-02103-2] [Medline: 36309732]
72. Mahdi SS, Ahmed Z, Allana R, Peretti A, Amenta F, Nadeem Bijle M, et al. Pivoting dental practice management during the COVID-19 pandemic-a systematic review. Medicina (Kaunas) 2020 Nov 25;56(12):644 [FREE Full text] [doi: 10.3390/medicina56120644] [Medline: 33255716]

XSL•FO
RenderX

73. Dwivedi R, Mehrotra D, Chandra S. Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: a systematic review. J Oral Biol Craniofac Res 2022;12(2):302-318 [FREE Full text] [doi: 10.1016/j.jobcr.2021.11.010] [Medline: 34926140]
74. Sharma S, Mohanty V, Balappanavar AY, Chahar P, Rijhwani K. Role of digital media in promoting oral health: a systematic review. Cureus 2022 Sep;14(9):e28893 [FREE Full text] [doi: 10.7759/cureus.28893] [Medline: 36225421]
75. Grall P, Ferri J, Nicot R. Surgical training 2.0: a systematic approach reviewing the literature focusing on oral maxillofacial surgery - part I. J Stomatol Oral Maxillofac Surg 2021 Sep;122(4):411-422. [doi: 10.1016/j.jormas.2021.01.006] [Medline: 33524605]
76. Grall P, Ferri J, Nicot R. Surgical training 2.0: a systematic approach reviewing the literature focusing on oral maxillofacial surgery - part II. J Stomatol Oral Maxillofac Surg 2021 Sep;122(4):423-433. [doi: 10.1016/j.jormas.2020.11.010] [Medline: 33301948]
77. Fernández CE, Maturana CA, Coloma SI, Carrasco-Labra A, Giacaman RA. Teledentistry and mHealth for promotion and prevention of oral health: a systematic review and meta-analysis. J Dent Res 2021 Aug;100(9):914-927. [doi: 10.1177/00220345211003828] [Medline: 33769123]
78. Flores AP, Lazaro SA, Molina-Bastos CG, Guattini VL, Umpierre RN, Gonçalves MR, et al. Teledentistry in the diagnosis of oral lesions: a systematic review of the literature. J Am Med Inform Assoc 2020 Jul 01;27(7):1166-1172 [FREE Full text] [doi: 10.1093/jamia/ocaa069] [Medline: 32568392]
79. Daniel SJ, Wu L, Kumar S. Teledentistry: a systematic review of clinical outcomes, utilization and costs. J Dent Hyg 2013 Dec;87(6):345-352. [Medline: 24357563]
80. Mariño R, Ghanim A. Teledentistry: a systematic review of the literature. J Telemed Telecare 2013 Jun;19(4):179-183. [doi: 10.1177/1357633x13479704] [Medline: 23512650]
81. Abdul NS, Kumari M, Shenoy M, Shivakumar GC, Herford AS, Cicciù M, et al. Telemedicine in the diagnosis and management of temporomandibular disorders: a systematic review conducted according to PRISMA guidelines and the Cochrane Handbook for Systematic Reviews of Interventions. J Oral Rehabil 2023 Nov;50(11):1340-1347. [doi: 10.1111/joor.13546] [Medline: 37349872]
82. Sereewisai B, Chintavalakorn R, Santiwong P, Nakornnoi T, Neoh SP, Sipiyaruk K. The accuracy of virtual setup in simulating treatment outcomes in orthodontic practice: a systematic review. BDJ Open 2023 Aug 28;9(1):41 [FREE Full text] [doi: 10.1038/s41405-023-00167-3] [Medline: 37640693]
83. Zaheer HA, Jabir AR, Yang K, Othman S, Kaleem SZ, McKinnon BJ. The use of mobile devices in oculoplastic and oral and maxillofacial surgery: a systematic review. Am J Otolaryngol 2022;43(3):103282. [doi: 10.1016/j.amjoto.2021.103282] [Medline: 34893349]
84. López-Valverde N, Muriel-Fernández J, López-Valverde A, Valero-Juan LF, Ramírez JM, Flores-Fraile J, et al. Use of virtual reality for the management of anxiety and pain in dental treatments: systematic review and meta-analysis. J Clin Med 2020 Sep 24;9(10):3086 [FREE Full text] [doi: 10.3390/jcm9103086] [Medline: 32987885]
85. Irving M, Stewart R, Spallek H, Blinkhorn A. Using teledentistry in clinical practice as an enabler to improve access to clinical care: a qualitative systematic review. J Telemed Telecare 2018 Apr;24(3):129-146. [doi: 10.1177/1357633X16686776] [Medline: 28092220]
86. Kılıç A, Brown A, Aras I, Hui R, Hare J, Hughes LD, et al. Using virtual technology for fear of medical procedures: a systematic review of the effectiveness of virtual reality-based interventions. Ann Behav Med 2021 Oct 27;55(11):1062-1079 [FREE Full text] [doi: 10.1093/abm/kaab016] [Medline: 33821879]
87. Loy F, Underwood B, Stevens C. Watch and learn? A systematic review comparing oral health educational videos with written patient information aimed at parents/carers or children. Br Dent J 2021 Nov 23:1-6 [FREE Full text] [doi: 10.1038/s41415-021-3616-5] [Medline: 34815480]
88. Thurzo A, Urbanová W, Novák B, Czako L, Siebert T, Stano P, et al. Where is the artificial intelligence applied in dentistry? Systematic review and literature analysis. Healthcare (Basel) 2022 Jul 08;10(7):1269 [FREE Full text] [doi: 10.3390/healthcare10071269] [Medline: 35885796]
89. BIREME/PAHO/WHO. Evidence map - Teledentistry. Tableau Public.: BIREME URL: https://public.tableau.com/app/profile/bireme/viz/teleodontologia-en/evidence-map [accessed 2024-01-20]
90. BIREME/PAHO/WHO. BIREME. Tableau Public. URL: https://public.tableau.com/app/profile/bireme/vizzes [accessed 2024-01-20]
91. Hung M, Lipsky MS, Phuatrakoon TN, Nguyen M, Licari FW, Unni EJ. Teledentistry implementation during the COVID-19 pandemic: scoping review. Interact J Med Res 2022 Jul 21;11(2):e39955 [FREE Full text] [doi: 10.2196/39955] [Medline: 35862174]
92. Kröplin J, Maier L, Lenz JH, Romeike B. Knowledge transfer and networking upon implementation of a transdisciplinary digital health curriculum in a unique digital health training culture: prospective analysis. JMIR Med Educ 2024 Apr 15;10:e51389 [FREE Full text] [doi: 10.2196/51389] [Medline: 38632710]
93. European Commission: Directorate-General for Health and Food Safety. European health data space. Publications Office of the European Union. 2024. URL: https://data.europa.eu/doi/10.2875/269514 [accessed 2024-01-17]

XSL•FO
RenderX

94. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). European Union. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2024-01-17]
95. Nienaber McKay AG, Brand D, Botes M, Cengiz N, Swart M. The regulation of health data sharing in Africa: a comparative study. J Law Biosci 2024;11(1):lsad035 [FREE Full text] [doi: 10.1093/jlb/lsad035] [Medline: 38259628]
96. United Nations Economic and Social Commission for Asia and the Pacific. Regulation of digital health and health data in the Asia Pacific Region. United Nations ESCAP. 2023 Dec 27. URL: https://repository.unescap.org/handle/20.500.12870/6645 [accessed 2024-01-20]
97. Lupiáñez-Villanueva F, Gunderson L, Vitiello S, Febrer N, Folkvord F, Chabanier L, et al. Study on health data, digital health and artificial intelligence in healthcare. European Commission: Directorate-General for Health and Food Safety. 2023 Jul 02. URL: https://data.europa.eu/doi/10.2875/702007 [accessed 2024-01-17]
98. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html [accessed 2024-01-17]
99. Brasil. Lei Geral de Proteção de Dados Pessoais (LGPD). Ministério do Esporte, Brazil. URL: https://www.gov.br/esporte/pt-br/acesso-a-informacao/lgpd [accessed 2024-01-20]
100. Brasil. General Personal Data Protection Act (LGPD). lgpd-brazil.info. 2018 Aug. URL: https://lgpd-brazil.info/ [accessed 2024-01-17]
101. Brasil. LEI Nº 14.510, DE 27 de dezembro de 2022. Presidência da República Secretaria-Geral Subchefia para Assuntos Jurídicos. URL: https://planalto.gov.br/ccivil_03/_ato2019-2022/2022/Lei/L14510.htm [accessed 2024-01-20]
102. Brasil. Brazilian national digital health strategy 2020-2028. Brazilian Ministry of Health. 2020. URL: https://bvsms.saude.gov.br/bvs/publicacoes/strategy_health_digital_brazilian.pdf [accessed 2024-01-20]
103. Bulletin n° 70 - BIREME develops Evidence Maps Portal. BIREME/PAHO/WHO. 2022 Jul 28. URL: https://boletin.bireme.org/en/2022/07/28/bireme-develops-evidence-maps-portal/ [accessed 2024-01-20]
104. Gurgel-Juarez N, Torres-Pereira C, Haddad AE, Sheehy L, Finestone H, Mallet K, et al. Accuracy and effectiveness of teledentistry: a systematic review of systematic reviews. Evid Based Dent 2022 Jul 08:1-8 [FREE Full text] [doi: 10.1038/s41432-022-0257-8] [Medline: 35804195]

Abbreviations

AI: artificial intelligence
AMSTAR2: A Measurement Tool to Assess Systematic Reviews
BIREME/OPAS: Latin American and Caribbean Center for Health Sciences Information of the Pan American Health Organization
EGM: evidence gap map
ICT: information and communication technology
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SR: systematic review

Edited by K Williams; submitted 15.05.24; peer-reviewed by F Niknam, F Carrer; comments to author 10.07.24; revised version received 03.08.24; accepted 19.09.24; published 27.11.24.

Please cite as:
Nascimento da Silva Mulder J, Ramos Pinto M, Aníbal I, Dornellas AP, Garrido D, Huanca C, Haddad AE, Mendes Abdala CV
Teledentistry Applied to Health and Education Outcomes: Evidence Gap Map
J Med Internet Res 2024;26:e60590
URL: https://www.jmir.org/2024/1/e60590
doi:10.2196/60590
PMID:39602783

©Júlia Nascimento da Silva Mulder, Marcelo Ramos Pinto, Isabelle Aníbal, Ana Paula Dornellas, Deise Garrido, Camila Huanca, Ana Estela Haddad, Carmen Verônica Mendes Abdala. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Opportunities to Address Specialty Care Deserts and the Digital Divide through the Veterans Health Administration's Telehealth Hub-and-Spoke Cardiology Clinic: Retrospective Cohort Study

Rebecca Lauren Tisdale[1,2], MD, MPA; Colin Purmal[3,4], MD; Neil Kalwani[1,2], MD, MPP; Alexander Sandhu[2], MD, MS; Paul Heidenreich[1,2], MD, MS; Donna Zulman[1,2], MD, MS; Tanvir Hussain[3,4], MD, MBA, MSc, MHS

[1]Center for Innovation to Implementation (Ci2i), Health Services Research, VA Palo Alto Health Care System, Menlo Park, CA, United States

[2]Department of Medicine, School of Medicine, Stanford University, Stanford, CA, United States

[3]VA San Francisco Healthcare System, San Francisco, CA, United States

[4]Department of Medicine, School of Medicine, University of California, San Francisco, CA, United States

Corresponding Author:
Rebecca Lauren Tisdale, MD, MPA
Center for Innovation to Implementation (Ci2i)
Health Services Research
VA Palo Alto Health Care System
795 Willow Rd,152-MPD
Bldg 324
Menlo Park, CA, 94025
United States
Phone: 1 6502695193
Email: rebecca.tisdale2@va.gov

Abstract

Background: To address geographic barriers to specialty care access for services such as cardiology, the Veterans Health Administration (VA) has implemented a novel, regionalized telehealth care hub. The Clinical Resource Hub (CRH) model extends care, including cardiology services, to individuals in low-access communities across the region. Little is known, however, about the reach of such programs.

Objective: This study aimed to describe the initial CRH program implementation in terms of growth in users and clinical encounters, as well as the association between user characteristics and the use of CRH cardiology care, in VA's Sierra Pacific region (Northern California, Nevada, and the Pacific Islands).

Methods: We compared patients who used CRH cardiology services (CRH users) to those using non-CRH cardiology services (CRH nonusers) in the Sierra Pacific region between July 15, 2021, and March 31, 2023. After characterizing changes in the numbers of CRH users and nonusers and clinical encounters over the study period, we used multivariable logistic regression to estimate the association between patient-level factors and the odds of being a CRH user.

Results: There were 804 CRH users over the study period, with 1961 CRH encounters concentrated at 3 main CRH sites. The CRH program comprised a minority of cardiology users and encounters in the region, with 19,583 CRH nonusers with 83,489 encounters. The numbers of CRH patients and encounters both increased at a steady-to-increasing rate over the study period, with increases of 37% (n=292 vs n=213) in users and 64% (n=584 vs n=356) in encounters in the first quarter of 2023 compared with the last quarter of 2022. Among CRH users, 8.3% (67/804) were female and 41.4% (333/804) were aged ≥75 years, compared with 4.3% (840/19,583) and 49% (9600/19,583), respectively, among CRH nonusers. The proportions of rural (users: 205/804, 25.5%; nonusers: 4936/19,583, 25.2%), highly disabled (users: 387/804, 48.1%; nonusers: 9246/19,583, 47.2%), and low-income (users: 165/804, 20.5%; nonusers: 3941/19,583, 20.1%) veterans in both groups were similar. In multivariable logistic models, adjusted odds ratios of using CRH were higher for female veterans (1.70, 95% CI 1.29-2.24) and lower for older veterans (aged ≥75 years; 0.33, 95% CI 0.23-0.47). Rural veterans also had a higher adjusted odds ratio of using CRH (1.19, 95% CI 1.00-1.42; P=.046).

Conclusions: The VA's Sierra Pacific CRH cardiology program grew substantially in its first 2 years of operation, serving disproportionately more female and rural veterans and similar proportions of highly disabled and low-income veterans compared to conventional VA care. This model appears to be effective for overcoming specialty care access barriers for certain individuals,

XSL•FO
RenderX

although targeted efforts may be required to reach older veterans. While this study focuses on a single region, specialty, and health care system, lessons from implementing regionalized telehealth hub models may be applicable to other settings.

(J Med Internet Res 2024;26:e53932) doi:10.2196/53932

KEYWORDS

telehealth; specialty care; cardiovascular disease; telemedicine; cardiology; veterans; low income; digital divide; access; Veterans Health Administration; VA; VHA; rural; clinical resource hub; CRH

Introduction

Access to specialty care varies widely across US geographic regions, a pattern that poses problems for the delivery of cardiology care within the Veterans Health Administration (VA) [1-3]. Given the high prevalence of cardiovascular disease and associated morbidity and mortality among veterans [4], maintaining access to cardiology care is essential. As there are unique and disproportionately high risks of dual use of VA and community care for veterans with cardiovascular disease [5], maintaining access to VA-based cardiology care is a particular priority. Beyond the VA, patient subpopulations across the general United States, such as women [6] and those living in rural areas [7,8], similarly face barriers to accessing specialty cardiology care.

Telehealth (new or follow-up patient visits, delivered by phone or video) expanded significantly during the COVID-19 pandemic in cardiology [9] across the VA [10] and other health care systems. In this post–public health emergency phase of the pandemic, patient familiarity with these modalities of care provides opportunities for new ways of using telehealth, including improving access to specialty care in nonemergency settings.

The VA's Clinical Resource Hub (CRH) model of care offers mostly telehealth care to individuals in low-access regions. The CRH cardiology program was first implemented in July 2021 in VA's Sierra Pacific region, which serves Northern California, Nevada, and the Pacific Islands (Textbox 1).

Textbox 1. Veterans Health Administration (VA) Clinical Resource Hub (CRH) program.

Given the concentration of specialists in urban settings, the CRH model leverages specialist availability at an urban site to serve as the regional hub, where these providers extend their services (largely) through telehealth to distant sites. The Palo Alto VA site serves as the hub for specialty care, including cardiology, for the VA's CRH program in the Sierra Pacific region. Spoke site–specific contracts known as telehealth service agreements detail the relationship between the Palo Alto–based CRH clinical team and other individual sites; the spoke sites implementing CRH cardiology in 2021-2023 in this region were VA's Sierra Nevada (Reno), Southern Nevada (Las Vegas), and Northern California (Sacramento) sites. The services available through CRH at each spoke site (eg, which subspecialty clinics, such as heart failure or women's health cardiology) depend on the site's needs and service gaps, and the program employs physicians, nurses, pharmacists, and administrative staff.

While the initial implementation and usage of the CRH model has been described in primary and mental health [11,12] and for certain specialties [13,14], expansion of the program for cardiology specialty care has yet to be characterized. In this study, we have analyzed the initial implementation of the VA Sierra Pacific Region CRH cardiology program, describing the sociodemographic characteristics of users, program growth, and different modalities of care across the program.

Methods

Overview

This analysis was conducted as a clinical operations quality improvement project through the VA Sierra Pacific Region CRH leadership team. Using VA's Corporate Data Warehouse, we constructed a cohort of all patients with at least 1 evaluation and management encounter (visit) in any CRH or conventional VA-based cardiology clinic in VA's Sierra Pacific region (encompassing Northern California, Nevada, and the Pacific Islands) between July 15, 2021, when the first CRH site was first implemented in this region, and March 31, 2023.

Patient Characteristics

Patient characteristics were extracted for the first available data during the study period (from July 15, 2021, to March 31, 2023). We included the following patient-level sociodemographic data: age, sex, race, and ethnicity (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Pacific Islander, White, or unknown), rurality (highly rural, rural, or urban), and home site for receiving VA care. We also included VA enrollment priority as a proxy for social need, based on the VA's enrollment priority classification system [15]; this program categorizes VA patients according to military service–related disability and income and influences whether patients pay copays and what services they can access within VA. As in previous published literature [9,10], we condensed these into four enrollment priority categories: (1) high disability, corresponding to enrollment priority groups 1 and 4; (2) low-moderate disability, including priority groups 2, 3, and 6; (3) low-income, including priority group 5; and (4) no disability nor low-income status, wherein patients pay copays for VA care, including priority groups 7-8. Due to the hierarchical nature of these groups, veterans assigned to high- or low-moderate disability groups may also be low income.

We captured cardiovascular diagnoses based on primary diagnoses at cardiology visits, grouping these into several categories representing the most commonly-coded primary diagnoses: heart failure, ischemic heart disease, valvular heart disease, and atrial fibrillation or flutter. *International Classification of Diseases, 10th Revision* (*ICD-10*) codes are shown in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Encounter Characteristics

We captured the primary diagnosis group assigned to a given encounter as described above, as well as the encounter hub and spoke sites. We also collected data on encounter modality: phone, video (either direct to the patient's home through the VA's video platform, or from the cardiology team to a local clinic), or in person.

Statistical Analysis

For our descriptive analysis, we calculated the proportions of CRH users and nonusers with a given sociodemographic characteristic for all the characteristics outlined above. We calculated standardized mean differences with Cramer V for categorical variables and through Cohen *d* for continuous variables and binary categorical variables to contextualize differences between groups; standardized mean differences are more informative than *P* values when analyses have a large sample size, as even small differences between groups are statistically significant.

We constructed a logistic regression model with the primary outcome of adjusted odds of being a CRH user, with all the covariates outlined above as well as a fixed effect for the patient's assigned primary care site and robust SE. Because the VA's expansion of telehealth has prioritized video visits [16], we then constructed a separate logistic regression model of adjusted odds of being a video care user with every use of CRH as an additional covariate. All analyses were performed in Stata 18 (StataCorp LLC) and used a *P* value level of .05 to assess statistical significance. For all models, we treated missing data for a given characteristic as a separate category. The proportion of missing data was low (highest for ethnicity: 1231/20,387, 6%) for all characteristics except race, where 10.1% (2066/20,387) of patients were missing race information.

Ethical Considerations

This analysis was carried out as nonresearch quality improvement by VA program office partners in the VA Veterans Integrated Service Network 21 CRH operations team and was thus considered nonhuman participants' research. Therefore, it was exempt from institutional review board approval. Informed consent was waived, and patients were not compensated as data were collected in the course of normal clinical operations and used for quality improvement. Data analysis took place on a secure server to ensure privacy and confidentiality protection.

Results

Patients

There were 804 CRH users over the study period with a total of 4315 ambulatory cardiology encounters, 1961 of which were CRH encounters. Just over half of CRH users (403/804, 50.1%) had non-CRH cardiology encounters in addition to CRH encounters. In addition, there were 19,583 CRH nonusers with 83,489 ambulatory encounters, meaning CRH users comprised 3.9% (804/20,387) of the total patients using ambulatory cardiology services in the region over the study period.

Among CRH users, 8.3% (67/804) were female and 41.4% (333/804) were aged ≥75 years, compared with 4.3% (840/19,583) and 49% (9600/19,583), respectively, among CRH nonusers (Table 1). Similar proportions in both groups were rural or highly rural (CRH users: 205/804, 25.5%; CRH nonusers: 4936/19,583, 25.2%), highly disabled according to VA enrollment categorization (CRH users: 387/804, 48.1%; CRH nonusers: 9246/19,583, 47.2%), and low-income (CRH users: 165/804, 20.5%; CRH nonusers: 3941/19,583, 20.1%). Somewhat higher proportions of CRH users had diagnoses of atrial fibrillation or flutter (CRH users: 316/804, 39.3% and CRH nonusers: 6300/19,583, 32.2%), heart failure (CRH users: 149/804, 18.5% and CRH nonusers: 3405/19,583, 17.4%), and valvular heart disease (CRH users: 158/804, 19.7% and CRH nonusers: 2530/19,583, 12.9%); however, the proportion of CRH nonusers with a diagnosis of ischemic heart disease was higher than for CRH users (CRH nonusers: 8137/19,583, 41.6% and CRH users: 256/19,583, 31.8%). Similar proportions of CRH users and nonusers (CRH users: 205/804, 25.5% and CRH nonusers: 5393/19,583, 27.5%) had none of these diagnoses, that is, had a primary diagnosis other than atrial fibrillation or flutter, heart failure, valvular heart disease, or ischemic heart disease.

Figure 1 shows the number of CRH patients seen quarterly over time at the 3 sites with the most CRH encounters (Sierra Nevada, or Reno; Southern Nevada, or Las Vegas; and Northern California, or Sacramento). Quarterly patients generally increased over time (Figure 1), with some sites' growth rates picking up more abruptly (eg, Northern California) and others demonstrating a steadier increase (Sierra Nevada, Southern Nevada). Initial CRH encounters for these sites were July 29, 2021 (Southern Nevada); September 8, 2021 (Sierra Nevada); and September 21, 2021 (Northern California). Total unique patients increased over the previous quarter by 37% (n=292 vs n=213) in both the fourth quarter of 2022 and the first quarter of 2023 compared with the previous quarters. This trend was largely driven by rapid growth at the Northern California site, where patients more than tripled between the last quarter of 2022 and the first quarter of 2023.

XSL•FO
RenderX

Table 1. Characteristics of cardiology patients, clinical resource hub users (n=804), and nonusers (n=19,583).

	CRH[a] users (n=804)	CRH nonusers (n=19,583)	Standard mean difference[b]
Used video care[c] , mean (SD)	438 (54.5)	1633 (8.3)	0.29
Age (years), mean (SD)	69.5 (12.8)	72.7 (11.1)	–0.05
Age group (years), n (%)			0.25
18-44	50 (6.2)	515 (2.6)	—[d]
45-64	168 (20.9)	2998 (15.3)	—
65-74	253 (31.5)	6470 (33.0)	—
≥75	333 (41.4)	9600 (49.0)	—
Race, n (%)			0.02
American Indian or Alaska Native	15 (1.9)	275 (2.6)	—
Asian	20 (2.5)	765 (3.9)	—
Black or African American	79 (9.8)	2011 (10.3)	—
Native Hawaiian or other Pacific Islander	28 (3.5)	503 (2.6)	—
Unknown/Missing	87 (10.8)	1979 (10.1)	—
White	574 (71.4)	14,036 (71.7)	—
Ethnicity, n (%)			0.01
Hispanic or Latino	55 (6.8)	1569 (8.0)	—
Not Hispanic or Latino	695 (86.4)	16,837 (86.0)	—
Unknown/Missing	54 (6.7)	1177 (6.0)	—
Sex, n (%)			–0.04
Female	67 (8.3)	840 (4.3)	—
Male	737 (91.7)	18,743 (95.7)	—
Rurality, n (%)			0.02
Urban	592 (73.6)	14,574 (74.4)	—
Rural	197 (24.5)	4745 (24.2)	—
Highly Rural	8 (1.0)	191 (1.0)	—
Missing	7 (0.9)	74 (0.4)	—
Enrollment priority, n (%)			0.01
No special priority	113 (14.1)	2650 (13.5)	—
Low/moderate disability	137 (17.0)	3644 (18.6)	—
High disability	387 (48.1)	9246 (47.2)	—
Low income	165 (20.5)	3941 (20.1)	—
Missing	2 (0.2)	102 (0.5)	—
Diagnoses, n (%)			
Atrial fibrillation or flutter	316 (39.3)	6300 (32.2)	0.03
Heart failure	149 (18.5)	3405 (17.4)	0.01
Ischemic heart disease	256 (31.8)	8137 (41.6)	–0.04
Valvular heart disease	158 (19.7)	2530 (12.9)	0.04
Other diagnosis	205 (25.5)	5393 (27.5)	–0.01

[a]CRH: Clinical Resource Hub.

[b]Standardized mean differences calculated through Cramer V for categorical variables and through Cohen *d* for continuous variables and binary categorical variables. Both were calculated in Stata 18.

[c]Indicates use of video care during the study period.

XSL•FO
RenderX

[d]Not applicable.

Figure 1. Clinical Resource Hub patients seen at implementing sites over time (total unique patients: n=804; unique patients across sites depicted: n=688). Q: quarter.

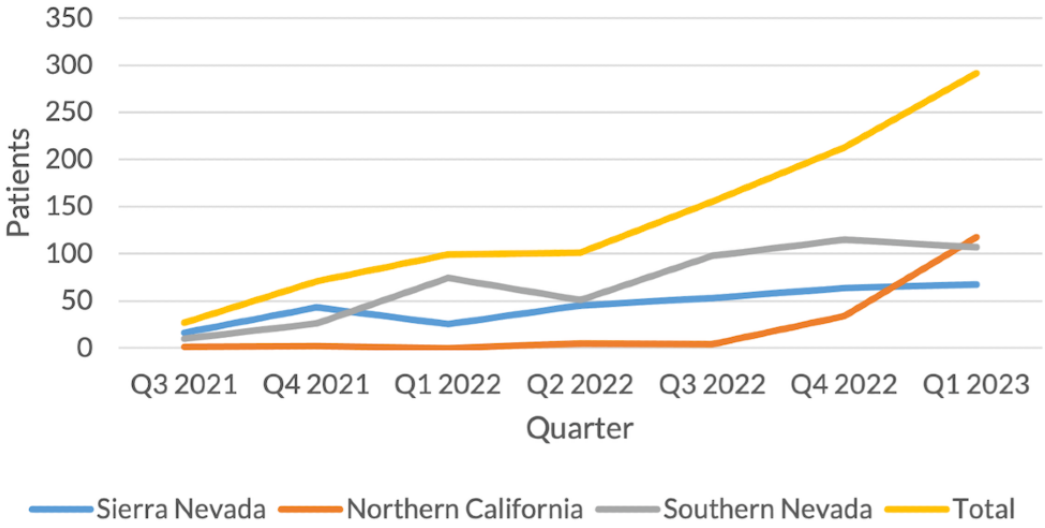

Encounters

The total number of ambulatory cardiology encounters in the VA Sierra Pacific region remained approximately constant over the study period, including all CRH and non-CRH VA-based encounters (Figure 2). A slight uptick in total encounters took place in the first quarter of 2023. This was mostly due to an increase of 762 encounters (n=8003 vs n=7241; 11% growth) in in-person encounters compared with the fourth quarter of 2022, though telephone and video visits also increased over this period by 474 visits (n=5630 vs n=5156; 9% growth) and 224 visits (n=937 vs n=713; 31% growth), respectively.

Figure 3 shows the growth in CRH cardiology encounters over time at the 3 sites with the most CRH encounters. These curves closely reflect patient-level trends over time at these sites, with 33% (n=356 vs n=268) growth in the number of quarterly CRH encounters between the third and fourth quarters of 2022 and 64% (n=584 vs n=356) growth between the fourth quarter of 2022 and the first quarter of 2023. By Q1 2023, CRH encounters comprised approximately 3.5% of regional cardiology encounters. All CRH sites depicted increased their encounters substantially between the fourth quarter of 2022 and the first quarter of 2023, but at different rates: encounters for patients based at the Sierra Nevada site increased by 11% (n=134 vs n=121) over this period, whereas encounters for patients based at the Northern California site increased more than 4-fold over the same period from 51 to 212.

A total of 714 (36.4%) of the 1961 CRH encounters were conducted through video, with the remainder (1247/1961, 63.6%) conducted through telephone. For non-CRH encounters, 4.6% (3830/83,489) were conducted through video; 41.6% (34,757/83,489) through telephone; and just over half, or 53.8% (44,902/83,489), occurred in person.

XSL•FO
RenderX

Figure 2. Cardiology encounters over time by encounter modality, Veterans Health Administration Sierra Pacific Region.

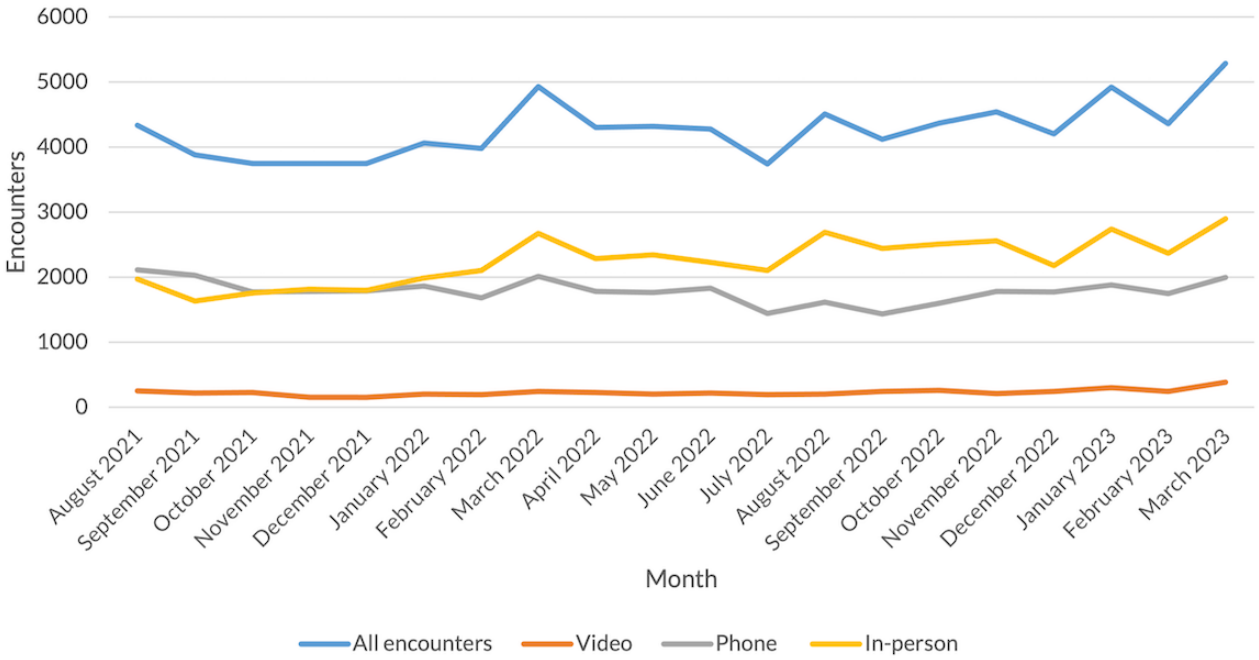

Figure 3. Clinical Resource Hub cardiology encounters over time.

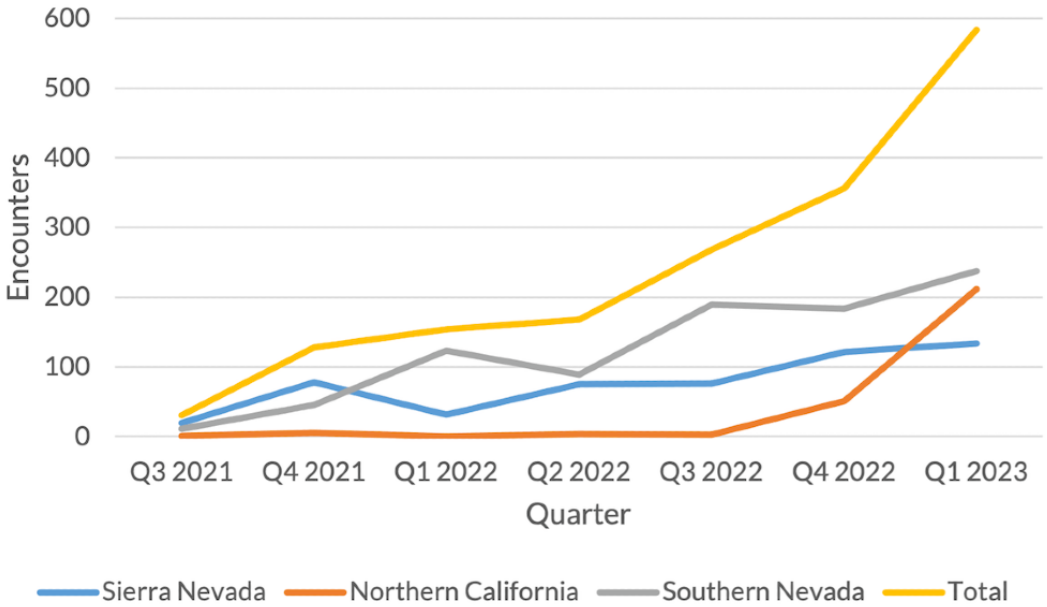

Adjusted Analyses

In our main multivariable logistic model, adjusted odds of using CRH were lower for older veterans (aged ≥75 years; adjusted odds ratio [OR] 0.33, 95% CI 0.23-0.47) and higher for female veterans (adjusted OR 1.70, 95% CI 1.30-2.24; Figure 4). Rural veterans had higher adjusted odds of using CRH (adjusted OR 1.19, 95% CI 1.00-1.42; *P*=.046) compared with the reference group of urban veterans, although the 95% CI for rural veterans overlapped to a great degree with the CI for highly rural and missing-rurality veterans. There were few significant differences by race or ethnicity, although patients of Native Hawaiian or other Pacific Islander race had higher odds of using CRH (adjusted OR 1.57, 95% CI 1.02-2.42). The enrollment priority group was not significantly associated with CRH use. Having a diagnosis of atrial fibrillation or flutter or valvular heart disease was associated with higher adjusted odds of using CRH (adjusted OR 1.54, 95% CI 1.32-1.81 and adjusted OR 1.93, 95% CI 1.59-2.34, respectively), whereas a diagnosis of ischemic heart disease was associated with lower odds (adjusted OR 0.79, 95% CI 0.68-0.93). Missing or unknown rurality was not pictured, which had an adjusted OR of 3.48 (95% CI 1.37-8.82).

XSL•FO
RenderX

Figure 4. Adjusted odds of being a Clinical Resource Hub cardiology user. OR: odds ratio.

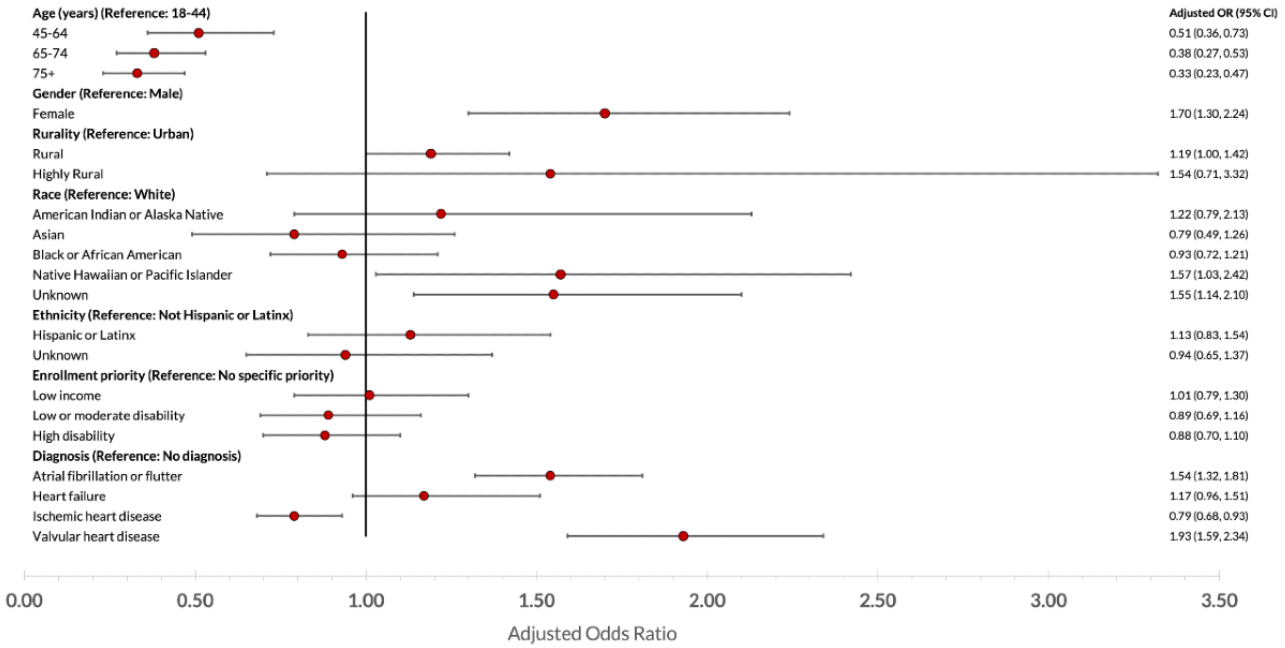

Table S2 in Multimedia Appendix 1 presents marginal probabilities of ever using CRH. These marginal probabilities varied for age group from 0.03 (95% CI 0.03-0.04) for veterans aged 75 years or older to 0.09 (95% CI 0.07-0.11) for those aged 18-44 tears. Female veterans' marginal probability of using CRH was 0.06 (95% CI 0.05-0.07), compared with 0.04 (95% CI 0.04-0.04) for male veterans. The marginal probability of using CRH was higher for veterans with valvular heart disease, at 0.06 (95% CI 0.05-0.07), than without (0.04, 95% CI 0.03-0.04). For other characteristics, there were no significant differences in marginal probabilities.

In our secondary analysis assessing the association of ever using CRH and visit modality, ever using CRH was associated with much higher adjusted odds of ever using video care (adjusted OR 29.73, 95% CI 24.43-36.18; refer to Table 2). Age was associated with lower adjusted odds of video care use according to a gradient, with an adjusted OR of 0.44 (95% CI 0.33-0.58) for veterans aged 75 years or older. Living in a rural location was associated with higher adjusted odds of video care use (adjusted OR 1.27, 95% CI 1.13-1.42), although this finding was not significant for those in highly rural locations (adjusted OR 1.30, 95% CI 0.83-2.03). Asian and Black or African American veterans had lower odds of video care use compared with White veterans (adjusted OR 0.68, 95% CI 0.50-0.91 and 0.78, 95% CI 0.65-0.93, respectively).

Marginal probabilities of video care use were higher for CRH users than nonusers at 0.64 (95% CI 0.62-0.68) and 0.08 (95% CI 0.08-0.09), respectively (Table S3 in Multimedia Appendix 1). Marginal probabilities for video care use varied significantly by age, from 0.09 (95% CI 0.08-0.09) for veterans aged 75 years or older to 0.16 (95% CI 0.13-0.19) for those aged 18-44 years. This figure was slightly higher for rural veterans than urban dwellers (0.12, 95% CI 0.11-0.12 for rural veterans compared to 0.10, 95% CI 0.09-0.10 for urban veterans). The marginal probability of using video care was significantly higher for veterans with valvular heart disease, at 0.14 (95% CI 0.13-0.15), than without (0.10, 95% CI 0.09-0.10). For other characteristics, there were no significant differences in marginal probabilities.

XSL•FO
RenderX

Table 2. Adjusted odds of using video care during the study period among all cohort patients (N=20,387).

	Adjusted odds ratio (95% CI)
CRH[a] user status	
Never used CRH	Reference category
Used CRH	29.73 (24.43-36.18)
Age group (years)	
18-44	Reference category
45-64	0.80 (0.61-1.06)
65-74	0.57 (0.43-0.75)
≥75	0.44 (0.33-0.58)
Race	
White	Reference category
American Indian or Alaska Native	1.27 (0.85-1.88)
Asian	0.68 (0.50-0.91)
Black or African American	0.78 (0.65-0.93)
Native Hawaiian or other Pacific Islander	0.84 (0.58-1.20)
Unknown or Missing	0.90 (0.74-1.09)
Ethnicity	
Not Hispanic or Latino	Reference category
Hispanic or Latino	0.91 (0.75-1.10)
Unknown or Missing	1.12 (0.89-1.42)
Sex	
Male	Reference category
Female	1.20 (0.96-1.50)
Rurality	
Urban	Reference category
Rural	1.27 (1.13-1.42)
Highly Rural	1.30 (0.83-2.03)
Missing	1.81 (0.74-4.46)
Enrollment priority	
No special priority	Reference category
Low or moderate disability	1.03 (0.86-1.23)
High disability	1.13 (0.97-1.31)
Low income	0.81 (0.68-0.97)
Missing	0.54 (0.23-1.24)
Diagnoses	
Atrial fibrillation or flutter	1.23 (1.10-1.38)
Heart failure	1.37 (1.21-1.55)
Ischemic heart disease	1.25 (1.13-1.39)
Valvular heart disease	1.69 (1.48-1.94)

[a]CRH: Clinical Resource Hub.

XSL•FO
RenderX

Discussion

Principal Findings

The VA has implemented this regionalized hub-and-spoke, primarily telehealth cardiology clinic to extend specialty care services to individuals in low-access communities across the region. In VA's Sierra Pacific region, the CRH program served 800 veterans hailing from across the region in nearly 2000 telehealth encounters for evaluation and management of cardiovascular disease in its first 2 years of operation, and numbers of both patients and encounters increased at a steady-to-increasing rate. The CRH program reached women and rural-dwelling veterans at higher rates and highly disabled and low-income veterans at similar proportions compared with conventional VA-based cardiology clinics in the same region; conversely, the CRH patient population skewed younger than the conventional VA clinic population. This suggests that such a predominately telehealth, regionalized model of specialty care may be an effective method for accessing care for many high-need groups, although more targeted efforts may be required to reach older individuals. As more than half of rural-dwelling Americans live more than 20 km from the nearest cardiologist, and 95% live more than 20 km from a heart failure specialist [8], exploring telehealth-predominant care models like the CRH to expand access to these specialists is a priority both within and beyond the VA.

The reporting of these early results coincides with a shift in telehealth use from effectively a requirement during the national emergency phase of the COVID-19 pandemic, to an option for patients and clinicians alike [17]. This evolution brings both an opportunity and a mandate for rigorous study of how, when, and for whom telehealth should be used, and how telehealth visits affect the quality of care, resource use, and health outcomes [18]. This study is formative, with a focus on examining patterns of use; this lays a foundation for follow-up studies delving into the latter set of questions.

The concern of the digital divide [19] is ever present when considering the use of telehealth: will a primarily telehealth-based model of care inadvertently exclude groups frequently falling on the wrong side of the divide, such as those who are rural dwelling or have low income? Based on these findings, this particular program has reached historically marginalized groups in VA, such as women, racial or ethnic minority veterans, or those who are highly disabled or have low income, at similar or higher rates than the conventional model. A notable exception is among older individuals, who used CRH at much lower rates than their younger counterparts. The majority of older individuals in the United States are interested in conducting visits via telehealth [20], yet disparities in use by age have been widely demonstrated in VA both in general and in cardiology [9,10]. Establishing the source or sources of this discrepancy—whether due to true differences in interest in receiving care via a primary telehealth model, lower rates of offering the CRH program to older individuals, familiarity with navigating telehealth technologies, or other factors—will be an important focus of follow-up work.

We found that patients with diagnoses of atrial fibrillation/flutter or valvular heart disease had higher adjusted odds of being CRH users, unlike patients with diagnoses of heart failure or ischemic heart disease. This finding may reflect program-specific offerings (for example, clinics or physicians in the hub site with particular expertise in managing these conditions, or a perceived lack of capacity to manage them at spoke sites), or a sense that these conditions are more amenable to primarily telehealth management. Planned qualitative work, including interviews with program clinicians and administrators, will help to differentiate between these possible drivers.

For telehealth models designed to improve patient access to a given service, it is essential to establish whether that model offloads the conventional model, as intended, or simply induces more demand (eg, patients whose cardiovascular diseases would have otherwise been cared for in a primary care setting are instead referred for cardiology care). While this study does not aim to definitively answer this question, the fact that total cardiology encounters remained constant in the region over the study period suggests that there was not a strong demand-creation effect of the CRH model. However, to date, CRH patients comprise only a small fraction of total regional patients using cardiology services, so continued attention to this question will be important as the program grows.

Limitations

Within the current data and study design, we are limited in the interpretation of various aspects of our findings. For example, although we can capture which CRH users have also used conventional care, our data lack the granularity to understand how and when this is the case; subsequent qualitative work will further elucidate these care patterns. At present, our data are limited to encounters within VA and do not extend to VA-purchased care in the community, meaning we cannot fully conclude whether CRH affects consumption of this costly form of care. This question, and characterization of other important facets of care associated with the program, such as patient, caregiver, and clinician satisfaction, clinical outcomes, and drivers of more and less successful program implementation, are left for future work. Finally, this analysis focused on a particular region and health care system and, therefore, may not be fully generalizable to other health care settings, although as Burnett et al [11] note in their publication on the early CRH implementation experience that "...some CRH design elements and experiences are unique to the [VA] system, [but] overall experience with telehealth hubs—including attempts to improve capacity for service provision, increase access, and deployment of telehealth services—is likely highly relevant to other health care systems." In particular, the patient population in the VA includes fewer women and skews older than the general US population. We do note that female and younger veterans represent 2 subpopulations that used this program at disproportionately high rates compared to male and older veterans, respectively.

Conclusions

The cardiology CRH program represents a telehealth-predominant, regionalized model of care that has extended specialty cardiology services to approximately 800

XSL•FO
RenderX

patients with low access to date in the VA region serving Northern California, Nevada, and the Pacific Islands. These data from the first 2 years of program implementation suggest that the program reached many of the most historically marginalized subpopulations of veterans, including female, rural-dwelling, and low-income veterans, at similar or higher rates compared with conventional cardiology care in the region. A notable exception was older individuals, who used CRH care at much lower rates; further work will examine the extent to which patient preference versus other factors drove this dynamic.

Many Americans live long distances from cardiologists, particularly in rural areas, prompting calls for the exploration of telehealth interventions to improve access [8]. This program represents one example of a regionalized, telehealth-predominant model that could be replicated in other regions and health care systems to address this widespread need.

Acknowledgments

The authors thank the Virtual Access QUERI team for supporting this evaluation, including Cindie Slightam, MPH, and Camila Chaudhary, MPH, from the Veterans Health Administration (VA) Palo Alto Health Care System, for providing project management support, and Liberty Greene, MEd, and James Van Campen, MS, for data consulting. Preliminary results of this work were presented at the American Heart Association Scientific Sessions, on November 13, 2023.

This work was supported by the VA Office of Academic Affairs Advanced Fellowship in Health Services Research (RLT) and by a pilot grant from the VA Palo Alto Center for Innovation to Implementation (RLT). Views expressed are those of the authors and the contents do not represent the views of the US Department of Veterans Affairs or the United States Government. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to US Department of Veterans Affairs regulations and our ethics agreements, which do not permit data to leave the Veterans Health Administration (VA) firewall without a Data Use Agreement. This limitation is consistent with other studies based on VA data. However, VA data are made freely available to researchers with an approved VA study protocol. For more information, please contact the VA Information Resource Center at VIReC@va.gov.

Conflicts of Interest

None declared.

Multimedia Appendix 1
International Classification of Diseases, 10th Revision codes for cardiovascular disease groups, and marginal probabilities of ever using Clinical Resource Hub and video care.
[DOCX File , 21 KB - jmir_v26i1e53932_app1.docx]

References

1. Gurewich D, Shwartz M, Beilstein-Wedel E, Davila H, Rosen AK. Did access to care improve since passage of the veterans choice act?: differences between rural and urban veterans. Med Care 2021;59(Suppl 3):S270-S278 [FREE Full text] [doi: 10.1097/MLR.0000000000001490] [Medline: 33976076]
2. Mattocks KM, Kroll-Desrosiers A, Kinney R, Elwy A, Cunningham K, Mengeling M. Understanding VA's use of and relationships with community care providers under the MISSION act. Med Care 2021;59(Suppl 3):S252-S258 [FREE Full text] [doi: 10.1097/MLR.0000000000001545] [Medline: 33976074]
3. Ohl ME, Carrell M, Thurman A, Weg MV, Hudson T, Mengeling M, et al. "Availability of healthcare providers for rural veterans eligible for purchased care under the veterans choice act". BMC Health Serv Res 2018;18(1):315 [FREE Full text] [doi: 10.1186/s12913-018-3108-8] [Medline: 29807536]
4. Hinojosa R. Veterans' likelihood of reporting cardiovascular disease. J Am Board Fam Med 2019;32(1):50-57 [FREE Full text] [doi: 10.3122/jabfm.2019.01.180148] [Medline: 30610141]
5. Axon RN, Gebregziabher M, Everett CJ, Heidenreich P, Hunt KJ. Dual health care system use is associated with higher rates of hospitalization and hospital readmission among veterans with heart failure. Am Heart J 2016;174:157-163. [doi: 10.1016/j.ahj.2015.09.023] [Medline: 26995383]
6. Balla S, Gomez SE, Rodriguez F. Disparities in cardiovascular care and outcomes for women from racial/ethnic minority backgrounds. Curr Treat Options Cardiovasc Med 2020;22(12):75 [FREE Full text] [doi: 10.1007/s11936-020-00869-z] [Medline: 33223802]

XSL•FO
RenderX

7. Loccoh EC, Joynt Maddox KE, Wang Y, Kazi DS, Yeh RW, Wadhera RK. Rural-urban disparities in outcomes of myocardial infarction, heart failure, and stroke in the United States. J Am Coll Cardiol 2022;79(3):267-279 [FREE Full text] [doi: 10.1016/j.jacc.2021.10.045] [Medline: 35057913]
8. Motairek I, Chen Z, Makhlouf MHE, Deo S, Salerno PRVO, Mentias A, et al. Mapping geographic proximity to cardiologists across the United States. Circ Cardiovasc Qual Outcomes 2023;16(10):e010133. [doi: 10.1161/CIRCOUTCOMES.123.010133] [Medline: 37698011]
9. Tisdale RL, Ferguson J, van Campen J, Greene L, Sandhu AT, Heidenreich PA, et al. Disparities in virtual cardiology visits among veterans health administration patients during the COVID-19 pandemic. JAMIA Open 2022;5(4):ooac103 [FREE Full text] [doi: 10.1093/jamiaopen/ooac103] [Medline: 36531138]
10. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the veterans health administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc 2021;28(3):453-462 [FREE Full text] [doi: 10.1093/jamia/ocaa284] [Medline: 33125032]
11. Burnett K, Stockdale SE, Yoon J, Ragan A, Rogers M, Rubenstein LV, et al. The clinical resource hub initiative: first-year implementation of the veterans health administration regional telehealth contingency staffing program. J Ambul Care Manage 2023;46(3):228-239 [FREE Full text] [doi: 10.1097/JAC.0000000000000468] [Medline: 37079357]
12. Kleinfeld S, Burgos-Chapman I, Filips J, Gould C. Utilizing telehealth to meet the needs of rural veterans: introduction the VA's clinical resource hubs geriatric mental health programs. Am J Geriatr Psychiatry 2023;31(3):S3-S4. [doi: 10.1016/j.jagp.2022.12.271]
13. Serper M, Agha A, Garren PA, Taddei TH, Kaplan DE, Groeneveld PW, et al. Multidisciplinary teams, efficient communication, procedure services, and telehealth improve cirrhosis care: a qualitative study. Hepatol Commun 2023;7(6):e0157 [FREE Full text] [doi: 10.1097/HC9.0000000000000157] [Medline: 37219845]
14. Chun VS, Whooley MA, Williams K, Zhang N, Zeidler M, Atwood C, et al. Veterans health administration teleSleep enterprise-wide initiative 2017-2020: bringing sleep care to our nation's veterans. J Clin Sleep Med 2023;19(5):913-923 [FREE Full text] [doi: 10.5664/jcsm.10488] [Medline: 36708262]
15. Wang ZJ, Dhanireddy P, Prince C, Larsen M, Schimpf M, Pearman G. 2019 Survey of Veteran Enrollees' Health and Use of Health Care. Monterey, CA: Advanced Survey Design; 2020:207.
16. Heyworth L, Kirsh S, Zulman D, Ferguson JM, Kizer KW. Expanding access through virtual care: the VA's early experience with Covid-19. NEJM Catal 2020:11 [FREE Full text]
17. Goettsche K, DellaCava E. Telehealth as an option, not a requirement-who may benefit, and how do we decide? Am J Geriatr Psychiatry 2023;31(3):S3. [doi: 10.1016/j.jagp.2022.12.269]
18. Wray CM. Post-emergency department virtual care-more questions than answers. JAMA Netw Open 2022;5(10):e2237790. [doi: 10.1001/jamanetworkopen.2022.37790] [Medline: 36282506]
19. Rogers EM. The digital divide. Convergence: The International Journal of Research into New Media Technologies 2001;7(4):96-111. [doi: 10.1177/135485650100700406]
20. Li KY, Marquis LB, Malani PN, Solway E, Kirch M, Singer D, et al. Perceptions of telehealth among older U.S. adults during the COVID-19 pandemic: a national survey. J Telemed Telecare 2023:1357633X231166031 [FREE Full text] [doi: 10.1177/1357633X231166031] [Medline: 37021477]

Abbreviations

CRH: Clinical Resource Hub
ICD-10: International Classification of Diseases, 10th Revision
OR: odds ratio
VA: Veterans Health Administration

Edited by A Coristine; submitted 24.10.23; peer-reviewed by C Der-Martirosian, B Tyl, CL Wheat; comments to author 27.04.24; revised version received 30.05.24; accepted 25.09.24; published 28.11.24.

Please cite as:
Tisdale RL, Purmal C, Kalwani N, Sandhu A, Heidenreich P, Zulman D, Hussain T
Opportunities to Address Specialty Care Deserts and the Digital Divide through the Veterans Health Administration's Telehealth Hub-and-Spoke Cardiology Clinic: Retrospective Cohort Study
J Med Internet Res 2024;26:e53932
URL: https://www.jmir.org/2024/1/e53932
doi:10.2196/53932
PMID:

XSL•FO
RenderX

©Rebecca Lauren Tisdale, Colin Purmal, Neil Kalwani, Alexander Sandhu, Paul Heidenreich, Donna Zulman, Tanvir Hussain. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perspectives of Clients and Health Care Professionals on the Opportunities for Digital Health Interventions in Cerebrovascular Disease Care: Qualitative Descriptive Study

Henna Härkönen[1], MHS; Kirsi Myllykangas[1], MHS; Mikko Kärppä[2,3], MD; Kirsi Maaria Rasmus[4], DMD; Julius Francis Gomes[5], PhD; Milla Immonen[6], PhD; Piia Hyvämäki[1], MHS; Miia Jansson[1,7], PhD

[1]Research Unit of Health Sciences and Technology (HST), University of Oulu, Oulu, Finland
[2]Neurocenter, Department of Neurology, Oulu University Hospital, Oulu, Finland
[3]Research Unit of Clinical Medicine, University of Oulu, Oulu, Finland
[4]Faculty of Medicine, University of Oulu, Oulu, Finland
[5]Oulu Business School, University of Oulu, Oulu, Finland
[6]VTT Technical Research Centre of Finland Ltd, Oulu, Finland
[7]College of Science, Technology, Engineering and Mathematics (STEM), RMIT University, Melbourne, Australia

Corresponding Author:
Henna Härkönen, MHS
Research Unit of Health Sciences and Technology (HST)
University of Oulu
PO BOX 8000
Oulu, FI-90014
Finland
Phone: 358 504689544
Fax: 358 8344064
Email: henna.harkonen@oulu.fi

Abstract

Background: Cerebrovascular diseases (CVDs) are a major and potentially increasing burden to public health. Digital health interventions (DHIs) could support access to and provision of high-quality health care (eg, outcomes, safety, and satisfaction), but the design and development of digital solutions and technologies lack the assessment of user needs. Research is needed to identify opportunities to address health system challenges and improve CVD care with primary users of services as the key informants of everyday requirements.

Objective: This study aims to identify opportunities for DHIs from clients' and health care professionals' perspectives to address health system challenges and improve CVD care.

Methods: This study used a qualitative, descriptive approach. Semistructured, in-person interviews were conducted with 22 clients and 26 health care professionals in a single tertiary-level hospital in Finland between August 2021 and March 2022. The data were analyzed using a deductive and inductive content analysis.

Results: Identified opportunities for DHIs in CVD care were organized according to clients, health care professionals, and data services and classified into 14 main categories and 27 generic categories, with 126 subcategories of requirements. DHIs for clients could support the long-term management of health and life changes brought on by CVD. They could provide access to personal health data and offer health information, support, and communication possibilities for clients and their caregivers. Health care professionals would benefit from access to relevant patient data, along with systems and tools that support competence and decision-making. Intersectoral and professional collaboration could be promoted with digital platforms and care pathways. DHIs for data services could enhance care planning and coordination with novel predictive data and interoperable systems for data exchange.

Conclusions: The combined study of client and health care professional perspectives identified several opportunities and requirements for DHIs that related to the information, availability, quality, acceptability, utilization, efficiency, and accountability challenges of health systems. These findings provide valuable social insights into digital transformation and the emerging design,

XSL•FO
RenderX

development, and use of user-centered technologies and applications to address challenges and improve CVD care and health care.

(J Med Internet Res 2024;26:e52715) doi:10.2196/52715

KEYWORDS

cerebrovascular disease; stroke; digitalization; interventions; health care professional; client; patient; mHealth; mobile health; application; digital health; smartphones; health system; qualitative; descriptive study; brain; blood vessel disease; cerebrovascular disorder; Finland; interviews; efficiency; information; quality; accountability; neurology; neuroscience; brain injury; mobile phone

Introduction

Cerebrovascular diseases (CVDs) are a group of neurological conditions such as strokes (brought on by ischemia or hemorrhage) and transient ischemic attacks (TIAs) [1]. CVDs are one of the major burdens to public health globally as strokes are a leading cause of mortality and long-term disability [2]. The average lifetime costs of stroke are around $66,956-$74,605 [3]. The total costs of CVD are, however, estimated to rise by the year 2047 due to improved stroke survival rates and the aging population [4].

The use of digital health interventions (DHIs) has increased rapidly during the last decade [5] to tackle health system challenges, with the potential to increase the quality, accessibility, and availability of health services [6]. Current health care services face challenges in delivering high-quality primary [7] and secondary prevention [8], rehabilitation [9], and long-term support [10-12] for CVD. The European Stroke Action Plan emphasizes the research and development of digital systems, especially to support patients' self-management in CVD to reduce the risk of strokes and increase the quality of care [13].

DHIs have shown promise in the primary prevention of CVD with their growing acceptability and use by clients and health care professionals [14]. They can increase the availability of customized health and risk information and increase the self-management capabilities (eg, mobility, positive emotions, and behavior) of patients with CVD [15]. Pre- and in-hospital specialized telestroke services (eg, remote neurological examinations, triage, and case management) are efficient at improving acute CVD care and subsequent health outcomes [16]. In addition, individually tailored delivery of telerehabilitation, with the opportunity to view their own progress, can engage patients with an improved sense of control, motor skills, and activity [17].

Clients usually have positive attitudes and acceptance toward DHIs [5]. Patients with CVD value the provided benefits of the availability of quality information and access to their own health data, supportive supervision of rehabilitation [18], connection to people, and managing everyday life [19]. However, the impact of DHIs is affected by internal (eg, motivation and social challenges) and external (eg, practical and technical barriers) factors [17], as well as cognitive and functional challenges hindering usage [19].

Health care professionals are members of the health workforce who deliver health services [20]. They generally value DHIs as improving communication with clients and other health care professionals, with improvements for care planning, coordination, and workflow management [21]. DHIs that meet health care professional requirements are also vital, as poor usability and functionality add stress and reduce motivation for use [22]. Health care professionals do see the necessity of digitalization for improving CVD care [23], but little is known about the various requirements they have for DHIs to deliver high-quality care, especially in the acute CVD context.

Previous research has identified that DHIs are seldom designed from the primary user perspective to address specific health needs [24], which can lead to potential health disparities and reduce benefits gained [25]. Therefore, more qualitative research is required to identify how technology can address health system challenges from both the client and health care professional perspectives.

The aim of this study was to identify opportunities for DHIs from the client and health care professional perspectives to address health system challenges and improve care in CVD. The research question was as follows: What kind of requirements for DHIs do clients and health care professionals have in CVD care?

Methods

Study Design

A qualitative, descriptive, phenomenological approach was used with semistructured interviews and content analysis techniques to describe client and health care professional primary user perspectives of DHIs in CVD care [26]. The study was conducted in one of the major tertiary-level hospitals in Finland providing comprehensive acute CVD care. The hospital is located in the geographically largest hospital district area for specialized medical care in Finland [27] with a high prevalence of CVD [28]. Finland is one of the leading countries in digitalization, with high national eHealth maturity in health care [29] and a national information system service (Kanta Services) with archiving of electronic patient data [30].

Recruitment

The participants were recruited with convenience sampling. The inclusion criteria for the patients were as follows: adults (aged 18 years and older) with diagnosed ischemic stroke or TIA (I63.0-63.9, G45.0-G45.9, I60, or I61.0-61.9), an in-patient at the acute neurological ward, and able to give an informed consent. Patients were first assessed for eligibility by the ward staff. If eligible, patients were approached by a study nurse (RL)

XSL•FO
RenderX

after obtaining preliminary informed consent. Patient refusals (n=62) were related to the patient's assessment of not being able to answer the questions or not feeling well enough, length of interview, or lack of time due to an upcoming hospital discharge.

The inclusion criteria for the health care professionals were adults (aged 18 years or older) directly involved in CVD care at the acute neurological ward or at a CVD diagnostic and rehabilitation ward. The criteria were broad to ensure the representation of the various professionals involved in CVD care. The health care professionals were contacted face-to-face or by email for recruiting by the study nurse. The number of refusals was not documented, but high workload was the most common reason for refusals.

Data Collection

Data were collected between August 2021 and March 2022 to recruit enough patients and health care professionals to comprehensively answer the research question. Participants were recruited until data saturation [26]. Semistructured individual interviews were conducted once in Finnish by an experienced study nurse (RL, registered nurse) in a private room at the hospital. The study nurse had no prior relationship with the participants. Separate topic guides were used for participants based on previous studies [31-33]. The interviews were recorded, with no additional field notes made. The duration of the clients' interviews varied from 19 to 72 minutes and the duration of the health care professional interviews varied from 11 to 89 minutes.

Data Analysis

The interview transcripts were coded and categorized using NVivo qualitative data analysis software (version 1.6.1; QRR International Pty Ltd). The data were first sorted into open codes with a unit of analysis being a thought pattern [34]. In accordance with the World Health Organization (WHO) categorization [20], opportunities for DHIs were deductively organized according to overarching groupings based on the targeted primary user (clients, health care providers, health system managers, and data services) and classified according to main categories and generic categories [35]. The subcategories (ie, requirements for DHIs) were then defined inductively [34]. A categorization example is presented in Table 1. The analysis with all aforementioned stages was conducted by the corresponding author (HH) and verified by authors (KM and MJ) with methodological and research expertise.

Table 1. Example of the content analysis categorization and related quotations from the interview data.

Main category and generic category	Subcategory	Examples of quotations
Targeted client communication		
Health information	Accessibility	• H20: *…video content and all that. Reading all that text and picking things up and adopting them can be so limited, that it would be easier if someone would speak it. And it would be easily available at that moment when you are alert…* • H25: *…An app or a computer system, where you can straightforwardly move things forward. That will guide it step by step. But I think they should be in a figurative form with only a few things at a time. Because CVD patients often have that difficulty in comprehension…simple with also that picture as a clue, not necessarily only verbal. I think, it is right for that building of everyday life, especially at the hospital discharge phase it would be important.* • P19: *They could come to your phone, for example a counselling video to your e-mail or something. That would be easy to watch it from there then. And would be easier to read than some centimeter thick booklet on the same matter that could be seen on video in 5 to 10 minutes.*
Alerts and reminders	Daily tasks	• H16: *Of course, I wish that there could also be those types of daily tasks and all, like have you thought about this and done that…eat this and a tip of the day and stretch. Do not drink alcohol today.* • P9: *Of course, the application could be like, it is 16 and a half degrees warm, sun is shining, go for a walk.* • P1: *You could always add some kind of exercise application, a pedometer and so forth…that nags if you have not exercised enough.*

Ethical Considerations

The study was conducted according to the national guidelines for good ethical conduct [36]. Approvals were obtained from the Ethics Committee of Northern Ostrobothnia Hospital District (Decision No: 46/2021) and from the Northern Ostrobothnia Hospital District Board (Decision No: 30/2021). Written consent was obtained prior to the interviews. Anonymity was ensured by using ID numbers. All research data were stored in digital password-protected files. Any data used in this paper have been lawfully acquired in accordance with the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization to the Convention on Biological Diversity.

Rigor and Reflexivity

The participants were purposefully selected and recruited until data saturation to improve credibility [26] with an appropriate sample size of informants concerning the phenomenon. The analysis was verified by 2 researchers with results discussed in the research group, thus establishing dependability. In addition, the first author had access to only ID-numbered transcripts with no prior connection to the participants, which reduced bias. Confirmability and dependability were ensured by comparing

XSL•FO
RenderX

transcripts with the findings to ensure the representation of the participants' views and verify the process. Several quotations from various participants were provided to represent authenticity. The Consolidated Criteria for Reporting Qualitative Research (COREQ) [37] has been used to provide reliable reporting (Multimedia Appendix 1). The demographic characteristics of participants are provided to improve transferability (Table 2).

Table 2. Demographic characteristics of the study participants.

Characteristic	Clients (n=22), n (%)	Health care professionals (n=26), n (%)
Sex		
Female	7 (32)	21 (81)
Male	15 (68)	5 (19)
Age group (years)		
20-39	2 (9)	14 (54)
40-59	8 (36)	12 (46)
60-79	7 (32)	0 (0)
80 and older	5 (23)	0 (0)
Highest academic degree		
Primary education	8 (36)	0 (0)
Vocational training	9 (41)	3 (12)
Undergraduate degree	5 (23)	10 (38)
Graduate degree	0 (0)	13 (50)
Working status		
Student	1 (5)	N/A[a]
Employee or entrepreneur	9 (41)	N/A
Retired	12 (55)	N/A
Profession		
Practical care nurses	N/A	4 (15)
Registered nurses	N/A	9 (35)
Physiotherapist or speech therapists	N/A	3 (12)
Physician or psychologists	N/A	10 (38)
Prior cerebrovascular symptoms		
Yes	6 (27)	N/A
No	16 (73)	N/A

[a]N/A: not applicable or available.

Results

Characteristics of Participants

A total of 46 participants were interviewed (Table 2). Most (16/22, 73%) of the clients have recently experienced their first CVD. The health care professionals were physicians or psychologists who specialized in neurology or neuropsychology, registered nurses, practical care nurses, and physiotherapists or speech therapists.

Identified opportunities for DHIs in CVD were organized according to the clients, health care professionals, and data services, and they were classified into 14 main categories, 27 generic categories, and 126 subcategories. The identified opportunities for DHIs were related to information, availability, quality, acceptability, utilization, efficiency, and accountability of health systems.

Clients

The opportunities of DHIs for clients were classified under 5 main categories: targeted communication, untargeted communication, peer communication, personal health tracking, and on-demand information services (Table 3).

XSL•FO
RenderX

Table 3. Identified opportunities and requirements of digital health interventions for clients in cerebrovascular disease care.

Main category and generic category	Subcategory
Targeted client communication	
Health information	• Evidence-based information • Availability • Accessibility • Adaptability • Tailorability • Visualization of care pathway • Multimedia content • Personalized tasks and activities • Convenient timing • Involving caregivers • Multimodality
Alerts and reminders	• Contacting client groups • Color coding of reference values • Motivation support • Games • Customized message intervals • Personalization • Daily tasks • Informing caregivers
Diagnostic results	• Mobile access • Long-term follow-up
Untargeted client communication	
Untargeted health information	• Multimodality
Client-to-client communication	
Peer groups for clients	• Remote self-management counselling groups • Support platforms • Networks • Disease and symptom targeted groups • Caregiver targeted groups
Personal health tracking	
Access to own medical records	• Mobile access to EMR[a] or Kanta Services • Access for caregivers
Self-monitoring of health or diagnostic data	• Wearables and personal health tracking devices • Connectivity to mobile client applications • Feedback
Active data capture or documentation	• Mobile or computer client application • Checklists • Diaries • Symptom checkers • Instruments
On-demand information services	
Look-up of health information	• Mobile scanners • Color coding of reference values • Symptom checker and guidance • Individual goals • Activity tracking • Tasks • Feedback • Rewarding

[a]EMR: electronic medical record.

XSL•FO
RenderX

Targeted Client Communication

The subcategories of requirements for targeted client communication were inductively formed under 3 generic categories: health information, alerts and reminders, and diagnostic results. The DHIs should provide long-term access to multimodal (client applications, websites, or digital care pathways), evidence-based, CVD-related health, well-being, and service information before (primary prevention) and after CVD (follow-up). The health care professionals could assign clients (based on the personal health tracking data) to a suitable modality; prescribe tasks and activities; and provide counseling, education, and guidance to support the clients' self-management, care, and rehabilitation. The DHIs should enable self-monitoring with visual feedback (eg, graphs and numeric data). The information should be provided conveniently and adaptively based on choices, as described by 2 participants.

> *You could use an app and one part of it could be counselling using video material. It would be good if that app could be used before getting sick and afterwards, so it could be versatile in that way.* [Health care professional 4]

> *…You could decide yourself when, at what moment you would use the app. Or when you have energy for rehabilitation...If you are tired when somebody is coming to counsel you, then you may not get everything out of it. You could time it better.* [Patient 8]

The transmission of health information should be tailored according to individual needs (eg, disease, personal challenges, risk factors, functioning capabilities, and stage of care pathway) and include audiovisual content (eg, pictures, comics, movies, and speech), as demonstrated by 1 participant.

> *If you want multimedia content, then some form of electronic application or a web page or something could be good in that sense. It could have videos and some spoken content on a subject… It would be good for those people whose comprehension has taken a hit… CVD can include visual field defects, or with older patients, there could be a cataract or something else…so something you can listen to would be really good.* [Health care professional 13]

Involving caregivers was important, with information to support and guide them on the CVD care pathway. Information should be provided on CVD-related changes in the patient's mental health, personality, and functioning, as well as information on the availability of services (eg, peer-support networks). Caregivers should have information on how to support the patient, but also their own well-being.

Alerts and reminders should inform, recommend, and guide clients about their health status, risk factors, self-management (eg, medication, diet, exercise, and appointments), rehabilitation, and emergency situation (eg, acute CVD) assessment. This could be done by messaging color-coded reference values (eg, traffic lights) for their health data and visualization of the relation of their actions to risk reduction. The information could be personalized based on patient data in the form of games and personal tasks provided with customizable messaging intervals based on the client's needs, as demonstrated by 2 participants.

> *It could have information such as CVD signs and reminders. It could be game-like, so for example it could test whether you identify things, and it could have tips on what to do if something happens.* [Health care professional 5]

Additionally, mobile access to long-term diagnostic results could help clients to follow their self-management. In addition, the DHIs should include alerts and reminders to inform caregivers of the client's situation and of emergencies.

Untargeted Communication

The subcategories of requirements for untargeted communication were inductively formed under 1 generic category: untargeted health information. DHIs should provide general CVD-related information (eg, prevention, signs and symptoms) to the public via different multimodalities of 1-way communication (television, social media, and web-based materials).

Client-to-Client Communication

The subcategories of requirements for client-to-client communication were inductively formed under 1 generic category: peer groups for clients. Peer groups could be in the form of remote self-management (eg, smoking and weight) counseling groups, peer networks, and on platforms that enable the sharing of experiences. Peer groups could be targeted based on the client's disease or symptoms to create group unison with their own peer support for caregivers.

Personal Health Tracking

The subcategories of requirements for personal health tracking were inductively formed under 3 generic categories: access to the patient's own medical records, self-monitoring of health or diagnostic data, and active data capture and documentation. Access to the patient's own health data at Kanta Services was important, but the DHIs should also provide mobile access to electronic medical records (EMRs) for clients and caregivers. The self-monitoring of health and diagnostic data (eg, heart rhythm, neurological status, sleep, physical activity and movement, blood pressure, and blood glucose) could be performed by wearables or personal health tracking devices with connectivity to their own client applications. Feedback from client applications could help the assessment and reaction to the patient's own condition as described by 1 participant.

> *We should identify atrial fibrillations and other atrial arrhythmias…I think there could some kind of smart watch applications, that could have an alarm for those patients with an irregular heart rate.* [Health care professional 5]

Active data capture or documentation of personal health monitoring and self-management (eg, physical exercises, CVD risks, symptoms, pain, and sleep) data should be done using different client applications, checklists, diaries, symptom checkers, and instruments (ie, questionnaires) for clients to evaluate their situation, as described by 1 client.

XSL•FO
RenderX

> *…If there was some kind of easy to use application, for example a food diary, then you could follow it yourself, and check if you have stayed on it or not.* [Patient 19]

On-Demand Information Services

The subcategories of requirements for on-demand information services were inductively formed under 1 generic category of looking up health information. DHIs could provide support for clients' decision-making and self-management. Mobile product scanners could be used to identify medications and foods with additional color coding (ie, traffic lights) indicating recommended food consumption. Features such as individual goal setting, personalized feedback, and rewards (points or material goods) for tasks and activities could support making better lifestyle choices (eg, sleep and exercise). These DHIs could also be useful to caregivers as family members might often have similar activity habits. Clients could use Q&A services with a health care professional or chatbot contact, and symptom checkers with visualization tools and alerts to support clients and caregivers in their help-seeking (eg, acute CVD and physical or mental health conditions), as demonstrated by 1 participant.

> *For the caregiver, there could be an application, for instance, showing whether there are signs of limb function, whether the mouth is drooping, or if there are signs of dizziness…It should be in plain language, with a design that immediately shows you the values as our NEWS score and what to do… For instance, it might tell you to immediately call an emergency number and so on.* [Health care professional 4]

Health Care Professionals

The opportunities of DHIs for clients were classified under 6 main categories: client health records, decision support, telemedicine, training, prescription and medication management, and laboratory and diagnostics imaging management (Table 4).

XSL•FO
RenderX

Table 4. Identified opportunities and requirements of digital health interventions for health care professionals in CVD[a] care.

Main category and generic category	Subcategory
Client health records	
Longitudinal tracking of clients' health status and services	• Shared EMR[b] for levels of care • EMR integrated CVD registry • Indicator data collection
Manage structured clinical records	• Monitoring data transfer directly from devices to EMR
Manage unstructured clinical records	• Capsulizing description
Routine health indicator data collection and management	• Health measurement kiosks • Electronic questionnaires • Systematic data collection • Targeted queries
Decision support	
Prompts and alerts	• Reminders • Notifications • Guidelines • Data visualization • Data combination • EMR integrations • Mobile and computer access • Scoring systems with alerts • Real-time alerts
Checklist	• Mobile and computer client applications • EMR integration • Part of patient data visualization • Reminders
Screening	• Triage • Treatment selection • Data combining • Scoring tools • Audio-visual recording and mediating • Movement data with wearables • Remote examination • Mobile technology
Telemedicine	
Remote consultations	• Videoconference • Collaborative digital platforms • Real-time chat • Remote visits for caregivers • Personal emergency response systems with client applications or remote alarm wristbands
Remote monitoring	• Personal health monitoring device and wearable integrations • Data transfer to EMR and/or cloud service • Audio-visual recording • Alerts to health care professional • Automatic booking
Transmission of medical data	• Low threshold client communication system • Caregiver client communication system • Referral requests • Need assessment requests • Appointment booking • Messaging • Health and activity data documentation • Symptom diaries • Digital care pathways and platforms

XSL•FO
RenderX

Main category and generic category	Subcategory
Case management between providers	• Digital care pathways • Networks • Shared calendar • Computer and mobile access • Unofficial documentation platform • Hybrid appointments • Databases
Training	
Training content and reference material	• Learning and training system • Web-based learning • Audio-visual material
Prescription and medication management	
Transmit or track prescription orders	• Transmit purchase data from pharmacies to EMR
Laboratory and diagnostics imaging management	
Capture diagnostic results from digital devices	• Mobile CT[c] scan

[a]CVD: cerebrovascular disease.

[b]EMR: electronic medical record.

[c]CT: computer tomography.

Client Health Records

The subcategories of requirements for client health records were inductively formed under 4 generic categories: longitudinal tracking of clients' health status and services, managing structured clinical records, managing unstructured clinical records, and routine health indicator data collection and management. Longitudinal tracking of the client's health status and services could be achieved by using a shared EMR on levels of care with an integrated CVD register and quality data collection as described by 1 participant.

> *An electronic medical record with an integrated CVD registry, where on follow up visits, you would enter data…* [Health care professional 2]

Structured clinical records would benefit from direct health monitoring data transfer from devices to EMRs, reducing manual and duplicate documentation. Unstructured clinical records could be managed with agreed upon capsulizing descriptions (eg, signs of CVD) to increase the comparability between documentation.

Routine health indicator data collection and management would benefit from patient-reported outcomes and self-management data (eg, lifestyle, well-being, and risk factors). The data collection method could include public health measurement kiosks and electronic questionnaires integrated with EMRs or delivered to smartphones. DHIs could ensure systematic data collection from targeted client groups, and this would also enable health care professionals to react to needs, as described by 1 participant.

> *A questionnaire that goes to certain patient group… for instance, for the next five years, or once a year. Also, if the users somehow report that things are not going well, then someone would receive that information and offer help. For example, if there has been something like six months or a year and the patient realizes that they are not getting better, then that would be a dangerous moment. That would be a good moment to offer help.* [Health care professional 16]

Decision Support

The subcategories of requirements for decision support were inductively formed under 3 generic categories: prompts and alerts, checklists, and screening by risk or other health status. Prompts and alerts could be provided via an EMR user interface on a computer or smartphone or to separate applications to health care professionals. Prompts and alerts could support planning, monitoring (eg, examination, CVD signs, and help seeking), and care and rehabilitation based on evidence-based guidelines and protocols and CVD service information, as described by 1 participant.

> *You would not have to rely on your own self-confidence. You would get support from the information technology…because sometimes it seems that people are scared to call in…There could be a protocol that says you need to call someone.* [Health care professional 18]

They could also be in the form of visual presentations or combinations of monitoring data and EMR data. Alert functions in real-time based on monitoring data, scoring systems, documentation, or combinations of these could be provided to support acute decision-making.

Checklists could be part of the user interface data visualizations in the EMR or separate mobile and computer client applications, and they could support patient counseling, evaluation (eg, for medical statements), care (eg, monitoring, examination,

XSL•FO
RenderX

consultations, laboratory- and imaging needs, care tasks) and work shift–related duties, as described by 1 participant.

> *The insecurity, that have I done everything? What could be good is that, for example in the cardiac intensive unit, they have a feature in their own system, so that if you open a certain view, you get a list, where you tick everything off. You take into account everything that needs to be done for the patient's recovery…* [Health care professional 24]

DHIs with tools for screening could combine data from various sources (eg, Kanta Services; monitoring, audiovisual, health, and indicator data; and data from scoring systems) for population or individual screening. They could be used for triage (eg, preventive care, follow-up, immediate diagnosis, treatment, and rehabilitation), to select between treatment or service options, and evaluate progression. Screening features could also include different mobile and wearable technologies to record and mediate audio-visual and movement data for remote examination, triaging and monitoring symptom progression (eg, aphasia or apraxia, neglect, and limb weakness) for decision-making in acute care, and rehabilitation in CVD care, as described by 1 participant.

> *Perhaps in places where a neurologist or a doctor with a neurology specialization are not available, then maybe some form of video imaging could be useful for assessing symptoms and deciding on the right treatment forms.* [Health care professional 1]

Telemedicine

The subcategories of requirements for telemedicine were inductively formed under 4 generic categories: remote consultations, remote monitoring, transmission of medical data, and case management between health care professionals. Remote consultations could be used to provide various CVD-related services synchronously to clients and caregivers such as emergency services, follow-up appointments, rehabilitation (eg, neurological testing and neuropsychological), counseling (eg, coaching and adaptation training), education (eg, CVD information and lectures) and social work services, as explained by 1 participant.

> *Perhaps there could be a videoconference option so that you could see the patient and the doctor or nurse could run the neurological tests that we do here. Perhaps then they could also be followed up via video, so there is no need to come here and wait.* [Health care professional 18]

In addition, caregivers from long distances could remotely visit clients and participate in ward rounds. Remote consultations could be in the form of videoconferencing, real-time chat services, collaborative digital platforms for filling and submitting administrative forms with assistance, or personal emergency response systems with mobile client applications or wristbands with remote alarms.

Remote monitoring DHIs could be used to transfer data (eg, on vital functions, blood glucose, and activity or movement) from personal or health care professionals prescribed health monitoring devices and wearables directly using integrations with cloud services or EMR. The wearables could also be sent for later analysis (ie, home telemetry) as a service. Remote monitoring could also be in the form of audio-visual monitoring combined with wearable data (eg, electrocardiogram and video monitoring of the facial area). With abnormal values, the DHI would automatically send an alert to the health care professional or emergency services or book an appointment for a follow-up, as described by 1 participant.

> *If somebody is paralyzed, they cannot move. There would be a device, for instance, that could send an alarm if the patient does not move for an hour. The patient would have given permission to use the device* [Health care professional 5]

The transmission of medical data included asynchronous messaging, referral and need assessment requests, appointment booking, and documentation of health data and activities (eg, self-monitoring values, symptom diaries) by the client from different health care professionals assigned client applications, platforms, and care pathways. The health care professional could contact relevant client groups (eg, based on risk factors and previous acute CVD) easily, to collect health and well-being data, and monitor the client's health status and progress with the provision of information and guidance based on received data (see the *Clients* section). The possibility for a low threshold communication system for clients and caregivers to health care professionals was important, as described by 2 participants.

> *One idea could be a program and a connection to us. Then the patient could ask us questions, and the answer would come from here. There could be a specific time reserved to defuse and answer the questions.* [Health care professional 9]

> *There are many kinds of CVD-patients, and some might have communication difficulties and so on. In such cases, the caregiver could be involved as well. There could be a system for it, so that, if necessary, the patient or the caregiver could send a message, and someone could answer it…* [Health care professional 14]

The DHIs could enable better case management between health care professionals by adding mutual knowledge of available services and about the health care professional roles in the care and rehabilitation of CVD. Communication between health care professionals could be supported by computer and mobile access platforms (eg, shared digital calendars, networks, e-mails, and Microsoft Teams meetings). Health care professionals could jointly monitor and communicate on client progress, with possibilities for documenting unofficial information (eg, thoughts and conversations with clients) and hybrid appointments (eg, for functionality assessment), with other health care professionals participating remotely. There is a potential need for a national database for sharing knowledge on DHIs (ie, rehabilitation) and a potentially national CVD care pathway with the interprofessional collaboration of different health care professionals (organizations and levels of care), as described by 1 participant.

> *There is little information, that may not have updated, on the condition of the rehabilitation patient at the*

XSL•FO
RenderX

moment, so digital ways could be utilized with some care pathways. Professionals from different units could message better how the pathway has progressed. [Health care professional 21]

Training

The subcategories of requirements for training were inductively formed under 1 main category: training content and reference material. The DHIs could provide continuous education on patient counseling via learning and training systems with web-based learning and audio-visual material to support the competence of health care professionals.

Prescription and Medication Management

The subcategories of requirements for prescription and medication management were inductively formed under 1 main category: transmitting or tracking prescription orders. DHIs could provide interoperability between pharmacy systems and EMRs in order to track prescribed medication purchases and provide indirect data on their consumption.

Laboratory and Diagnostics Imaging Management

The subcategories of requirements for laboratory and diagnostics imaging management were inductively formed under 1 main category: capturing diagnostic results from digital devices. DHIs could be in the form of mobile computer tomography scans.

Data Services

The opportunities of DHIs for data services were classified under 3 main categories: data collection, management, and use; location mapping; and data exchange and interoperability (Table 5).

Table 5. Identified opportunities and requirements of digital health interventions for data services in cerebrovascular disease care.

Main category and generic category	Subcategory
Data collection, management, and use	
Data synthesis and visualizations	• Data mining from databases and EMRa • Graphic presentations • Compressed presentations • Defined data presentation order • Preference visualizations
Automated analysis of data	• Speech analysis • Integration to emergency services line • Video analysis • Audio-visual video modification • Interpretation of CTb-scan images • Prediction from data • Data combination from databases and EMR
Location mapping	
Clients and households	• Wearable sensors for location tracking
Data exchange and interoperability	
Data exchange across systems	• Interoperability • Applications • Data exchange between governmental databases • Interoperability between Kanta Services and local databases

[a]EMR: electronic medical record.

[b]CT: computer tomography.

Data Collection, Management, and Use

The subcategories of requirements for data collection, management, and use were inductively formed under two categories: data synthesis and visualizations, and the automated analysis of data. Compressed and combined data visualizations could be created for the planning and assessment of care and rehabilitation with data mining from relevant databases and from the EMR sections to the user interface. The visualizations could provide data in a specific order (eg, chronological, clinical specialties), or based on health care professional preferences (eg, recanalization therapy indicators and contraindications, disease specificity, CVD signs, therapy evaluations) with quick retrieval. The DHI could also combine data (eg, medication and vital functions) for the assessment of treatment effects, as described by 1 participant.

The blood pressure and given medication could be presented side by side in the same view along with the administration route used. Then you could easily see what has led to a certain reaction, or what has lowered the blood pressure, what medication has been most effective, or if some medication has lowered the pulse too much, or whether it has been unsuitable for the patient. [Health care professional 1]

The DHIs could provide automated analyses (eg, via machine learning or artificial intelligence [AI]) of different data sources to create new information to support decision-making. A DHI could analyze the patients' speech to detect aphasia and

XSL•FO
RenderX

dysarthria via integration with an emergency service line in acute CVD situations. The automated analysis of EKG (eg, for arrhythmias) and CT images could increase the quality of their interpretation. This could also combine and analyze data with predictive recommendations for care such as recovery prognoses, ideal rehabilitation timeframes, and risk assessments (postthrombolysis hemorrhage and deterioration). Predictive recommendations could also help in the treatment selection (eg, recanalization therapy, medication, and fluid therapy), as described by 1 participant.

> *A form of support could be something where you enter Stroke Scale scores and patient's symptoms, and it would also ask contraindications and all that. The doctor would enter them there, and based on those, it would recommend thrombolysis or not.* [Health care professional 2]

Video analysis could be used to monitor patients' progress in terms of CVD symptoms or rehabilitation effects. Modification of facial audio-visual video recordings could be used in client applications to visualize the effects of CVD signs for the client to increase awareness.

Location Mapping

The subcategories of requirements for location mapping were inductively formed under 1 main category: clients and households. DHIs could be in the form of wearable sensors that could be used to track the location of missing clients.

Data Exchange and Interoperability

The subcategories of requirements for data exchange and interoperability were inductively formed under 1 main category: data exchange across systems. The DHI should enable forms of interoperability between all health care providers on all levels, sectors, and districts of care, but also between governmental services (eg, tracking clients' driving licenses and traffic violations). Better exchange of client and health data (eg, illnesses, recent treatments, medication, memory and cognition, motion and functioning abilities), rehabilitation feedback, and current caregivers' contact information would be needed to keep them up to date and available at relevant stages of care. Interoperability was also found to be needed between Kanta Services and local databases of service providers.

Discussion

Principal Findings

This study identified several opportunities for DHIs from the primary users' perspectives to address health system challenges in CVD, resulting in 14 main categories, 27 generic categories, and 126 subcategories, respectively. The identified opportunities for DHIs were related to information, availability, quality, acceptability, utilization, efficiency, and accountability of health systems.

The client and health care professional perspectives were studied together, with similar and complementary findings. The client and health care professional primary user categories both included opportunities and requirements for DHIs to support communication, data sharing (eg, monitoring with wearables, interoperability between systems, applications, and digital devices), and decision-making (eg, checklists and symptom checkers). Both categories also included requirements to support the delivery of information, such as data visualization, reminders, and alerts. These findings provide information for the development of DHIs with functionalities that support primary user-centered and beneficial health systems.

Comparison With Prior Work

For clients, the identified opportunities for DHIs were related to information (eg, lack of quality or reliable data, lack of access to information or data, and insufficient utilization of data and information), quality (eg, inadequate supportive supervision), acceptability (eg, programs that do not address individual beliefs and practices), and accountability (eg, insufficient patient engagement). In line with the prior literature, there is a lack of quality or reliable data from both clients' and caregivers' perspectives [11,12], while client applications and digital care pathways could improve the access to information or data [18] to support self-care (eg, tracking and decision-making) from primary to secondary prevention and beyond. However, attention should be paid to the provision (eg, tailoring, timing, accessibility, and multimedia) of information to address individual health needs and practices [15,17]. The culture, language, and literacy of clients should also be considered when deploying DHIs in CVD care to increase acceptability [25].

Clients with CVD need access to EMRs to manage their own health data [18]. This study elaborates that clients need to be able to access and use information (eg, laboratory values) as well as behavioral, health, and well-being data, along with decision-support tools (eg, symptom checkers and checklists). Furthermore, there is a need for self-monitoring and documentation supporting engagement for self-management before and after an acute CVD to reduce the occurrence [38], recurrence [39], health care costs [7] of CVD, and resource use [40] in health care.

In line with the previous literature, there is a lack of supportive supervision for caregivers [11,12]. Correspondingly, client applications and communication systems are rarely used by caregivers [41]. According to this study, DHIs could offer information and communication systems to caregivers, highlighting the need to support caregivers in digitalization. Interestingly, the need for mutual client and caregiver DHIs was also expressed, which could be beneficial as psychoeducational interventions engaging both client groups have been found to relieve the caregiver burden, and improve the patient's quality of life, functional independence, family functioning, and health care service use [42], revealing a valid area for future research and development.

For health care professionals, the identified opportunities for DHIs were related to information (eg, lack of quality or reliable data, communication roadblocks), quality (eg, insufficient continuity of care, health worker competence, poor adherence to guidelines), efficiency (eg, inadequate workflow management, poor planning, coordination of care), and accountability (eg, poor accountability between the levels of the health sector).

XSL•FO
RenderX

There is a lack of quality data, where registries and electronic questionnaires could be used to collect data (eg, patient-reported outcome and self-management data), to support benchmarking of CVD care for the monitoring and improvement of stroke services nationally [43] and increase accountability. In addition, health kiosks [44] could be implemented for non–tech-savvy individuals (lack of device or digital literacy) to collect clinical measurements for population information. Remote monitoring (eg, wearables and integrations) of traditional and novel measurement values (eg, oxygen saturation, electroencephalogram, and movement) with functions to alert health care professionals, could improve access to reliable health data, increase the quality of follow-up and thus, reduce health care use, hospitalization, and mortality [45,46].

In line with previous research [11,21], there were communication roadblocks between clients, caregivers, and health care professionals due to traveling distances and lack of contact information, along with the lack of planning and coordination of care after acute CVD [8,9]. Telemedicine (eg, telerehabilitation, follow-up, and digital counseling) could provide low-threshold client communication systems as well as intersectoral and professional collaboration platforms to increase coordination and continuity of care. According to the participants, digital care pathways could provide combined systems for monitoring, communication, and bringing together the relevant health care professionals for service provision in CVD care.

The need to support competence and adherence to guidelines along with improvements in workflow management were identified. In line with the previous literature [47-49], the participants in this study identified the need for digital decision-support tools, especially in the prehospital and acute phase of CVD to screen (triage), perform acute diagnoses, select treatment, and monitor symptoms and conditions. Supportive features such as information provision, instructional guidance, and client-specific data were included in the decision-support tools which are associated with successful technological interventions [50]. Despite their acceptability, there exists a lack of evidence on the effectiveness of support systems for clinical decision-making, warranting further research [51].

For data services, the identified opportunities were related to information (eg, insufficient utilization use of information or data) and efficiency (eg, poor planning and coordination, and delayed provision of care). Poor planning and coordination were identified in CVD care, with requirements of DHIs highlighting the importance of interoperable EMRs and health management information systems on all levels and sectors of health care, as functioning health information exchange can secure the coordination of care and improve the quality of care and health outcomes [52,53]. In addition, traditional and novel (eg, video and speech analysis) data sources could be used with AI, machine learning, and data mining technologies to create predictive data on the course and outcomes of CVD [54,55]. This could accelerate decision-making and planning for the efficient provision of care. As there is limited data on the clinical applicability of AI solutions in CVD care, the primary user requirements of this study could be used along with rigorous validation research to support need-based development, as AI has the potential to improve health outcomes [54].

Strengths and Limitations

This study has some limitations. The results may not be transferable to all CVD care contexts, as the clients in this study were patients with ischemic stroke and TIA, and health care professionals worked mainly in the acute hospital care of these patients. The clients were also interviewed in acute phases of their CVD care which could affect their cognition and perceptions. DHIs were not identified for all of the WHO categories, and especially not for the health system managers' primary user group. This could be due to the study not targeting health care managers and that the interview topic guides were not designed according to the WHO categorization. On the other hand, the results of this study were similar to previous studies, adding credibility. The foremost strength of this study is the combination of the client and health care professional perspectives to identify DHIs suited for primary user needs.

Conclusions

This study identified several opportunities for DHIs to address health system challenges in CVD care. The identified opportunities were related to information, availability, quality, acceptability, utilization, efficiency, and accountability. DHIs are needed to provide long-term reliable health information, support, and communication possibilities to clients and their caregivers. Health care professionals can benefit from DHIs that provide access to relevant data to inform care and clinical decision-making with the support of interdisciplinary work in the care of CVD on all levels and sectors of health care. These opportunities and requirements provide valuable insights into the digital transformation and thus, the emerging use of digital technologies for health in cerebrovascular care.

Acknowledgments

The authors are grateful to the clients and health care professionals who agreed to participate in this research project and all the people who contributed to the project. This study was financially supported by Business Finland, for which we are sincerely grateful.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to research permit restrictions, but the corresponding author can be contacted for more information about the study.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist.
[DOCX File , 26 KB - jmir_v26i1e52715_app1.docx]

References

1. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJB, Culebras A, et al. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013;44:2064-2089 [FREE Full text] [doi: 10.1161/STR.0b013e318296aeca] [Medline: 23652265]
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease study 2019. Lancet Neurol 2021;20(10):795-820 [FREE Full text] [doi: 10.1016/S1474-4422(21)00252-0] [Medline: 34487721]
3. Strilciuc S, Grad DA, Radu C, Chira D, Stan A, Ungureanu M, et al. The economic burden of stroke: a systematic review of cost of illness studies. J Med Life 2021;14(5):606-619 [FREE Full text] [doi: 10.25122/jml-2021-0361] [Medline: 35027963]
4. Wafa HA, Wolfe CDA, Emmett E, Roth GA, Johnson CO, Wang Y. Burden of stroke in Europe: thirty-year projections of incidence, prevalence, deaths, and disability-adjusted life years. Stroke 2020;51(8):2418-2427 [FREE Full text] [doi: 10.1161/STROKEAHA.120.029606] [Medline: 32646325]
5. Eze ND, Mateus C, Hashiguchi TCO. Telemedicine in the OECD: an umbrella review of clinical and cost-effectiveness, patient experience and implementation. PLoS One 2020;15(8):e0237585 [FREE Full text] [doi: 10.1371/journal.pone.0237585] [Medline: 32790752]
6. From innovation to implementation: eHealth in the WHO European Region. World Health Organization. 2016. URL: https://www.who.int/europe/publications/i/item/9789289051378 [accessed 2024-12-23]
7. Owolabi MO, Thrift AG, Mahal A, Ishida M, Martins S, Johnson WD, et al. Primary stroke prevention worldwide: translating evidence into action. Lancet Public Health 2022;7(1):e74-e85 [FREE Full text] [doi: 10.1016/S2468-2667(21)00230-9] [Medline: 34756176]
8. Aquino MRJR, Mullis R, Moore C, Kreit E, Lim L, McKevitt C, et al. "It's Difficult, There's No Formula": qualitative study of stroke related communication between primary and secondary healthcare professionals. Int J Integr Care 2020;20(4):11 [FREE Full text] [doi: 10.5334/ijic.5465] [Medline: 33250676]
9. Chen L, Xiao LD, de Bellis A. First-time stroke survivors and caregivers' perceptions of being engaged in rehabilitation. J Adv Nurs 2016;72(1):73-84. [doi: 10.1111/jan.12819] [Medline: 26399942]
10. Boehme C, Toell T, Lang W, Knoflach M, Kiechl S. Longer term patient management following stroke: a systematic review. Int J Stroke 2021;16(8):917-926 [FREE Full text] [doi: 10.1177/17474930211016963] [Medline: 33949269]
11. Pindus DM, Mullis R, Lim L, Wellwood I, Rundell AV, Abd Aziz NA, et al. Stroke survivors' and informal caregivers' experiences of primary care and community healthcare services—a systematic review and meta-ethnography. PLoS One 2018;13(2):e0192533 [FREE Full text] [doi: 10.1371/journal.pone.0192533] [Medline: 29466383]
12. Zawawi NSM, Aziz NA, Fisher R, Ahmad K, Walker MF. The unmet needs of stroke survivors and stroke caregivers: a systematic narrative review. J Stroke Cerebrovasc Dis 2020;29(8):104875. [doi: 10.1016/j.jstrokecerebrovasdis.2020.104875] [Medline: 32689648]
13. Norrving B, Barrick J, Davalos A, Dichgans M, Cordonnier C, Guekht A, et al. Action plan for stroke in Europe 2018-2030. Eur Stroke J 2018;3(4):309-336 [FREE Full text] [doi: 10.1177/2396987318808719] [Medline: 31236480]
14. Feigin VL, Owolabi M, Hankey GJ, Pandian J, Martins SC. Digital health in primordial and primary stroke prevention: a systematic review. Stroke 2022;53(3):1008-1019 [FREE Full text] [doi: 10.1161/STROKEAHA.121.036400] [Medline: 35109683]
15. Hwang NK, Park JS, Chang MY. Telehealth interventions to support self-management in stroke survivors: a systematic review. Healthcare 2021;9(4):472 [FREE Full text] [doi: 10.3390/healthcare9040472] [Medline: 33921183]
16. Lazarus G, Permana AP, Nugroho SW, Audrey J, Wijaya DN, Widyahening IS. Telestroke strategies to enhance acute stroke management in rural settings: a systematic review and meta-analysis. Brain Behav 2020;10(10):e01787 [FREE Full text] [doi: 10.1002/brb3.1787] [Medline: 32812380]
17. Chen Y, Abel KT, Janecek JT, Chen Y, Zheng K, Cramer SC. Home-based technologies for stroke rehabilitation: a systematic review. Int J Med Inform 2019;123:11-22 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.12.001] [Medline: 30654899]
18. Bally ELS, Cheng D, van Grieken A, Ferri Sanz M, Zanutto O, Carroll A, et al. Patients' perspectives regarding digital health technology to support self-management and improve integrated stroke care: qualitative interview study. J Med Internet Res 2023;25:e42556 [FREE Full text] [doi: 10.2196/42556] [Medline: 37014677]

XSL•FO
RenderX

19. Gustavsson M, Ytterberg C, Nabsen Marwaa M, Tham K, Guidetti S. Experiences of using information and communication technology within the first year after stroke—a grounded theory study. Disabil Rehabil 2018;40(5):561-568 [FREE Full text] [doi: 10.1080/09638288.2016.1264012] [Medline: 27976926]
20. Classification of digital health interventions v 1.0. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2024-02-23]
21. Odendaal WA, Anstey Watkins J, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020;3(3):CD011942 [FREE Full text] [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
22. Hyvämäki P, Kääriäinen M, Tuomikoski A, Pikkarainen M, Jansson M. Registered nurses' and medical doctors' experiences of patient safety in health information exchange during interorganizational care transitions: a qualitative review. J Patient Saf 2022;18(3):210-224. [doi: 10.1097/PTS.0000000000000892] [Medline: 34419989]
23. Baatiema L, de-Graft Aikins A, Sarfo FS, Abimbola S, Ganle JK, Somerset S. Improving the quality of care for people who had a stroke in a low-/middle-income country: a qualitative analysis of health-care professionals' perspectives. Health Expect 2020;23(2):450-460 [FREE Full text] [doi: 10.1111/hex.13027] [Medline: 31967387]
24. Gorski I, Bram JT, Sutermaster S, Eckman M, Mehta K. Value propositions of mHealth projects. J Med Eng Technol 2016;40(7-8):400-421. [doi: 10.1080/03091902.2016.1213907] [Medline: 27687907]
25. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
26. Polit DF, Beck CT. Essentials of Nursing Research: Appraising Evidence for Nursing Practice. 9th ed. Philadelphia, PA: Wolters Kluwer; 2018.
27. Healthcare and social welfare services in the Oulu University Hospital catchment area of highly specialised medical care in 2019. Finnish Institute of Health and Welfare. 2020. URL: https://urn.fi/URN:ISBN:978-952-343-564-3 [accessed 2024-11-05]
28. Cerebrovascular disease index, age-standardised (up to 2019). Sotkanet. 2023. URL: https://sotkanet.fi/sotkanet/en/kartta?indicator=szYOAwA=&region=s041tTbRMwQA&year=sy6rAgA=&gender=t [accessed 2024-11-05]
29. Haverinen J, Keränen N, Tuovinen T, Ruotanen R, Reponen J. National development and regional differences in eHealth maturity in Finnish public health care: survey study. JMIR Med Inform 2022;10(8):e35612 [FREE Full text] [doi: 10.2196/35612] [Medline: 35969462]
30. What are Kanta Services? Kanta. URL: https://www.kanta.fi/en/professionals/what-are-kanta-services [accessed 2023-06-01]
31. Jansson MM, Harjumaa M, Puhto AP, Pikkarainen M. Patients' satisfaction and experiences during elective primary fast-track total hip and knee arthroplasty journey: a qualitative study. J Clin Nurs 2020;29:567-582. [doi: 10.1111/jocn.15121] [Medline: 31769559]
32. Jansson MM, Harjumaa M, Puhto AP, Pikkarainen M. Healthcare professionals' proposed eHealth needs in elective primary fast-track hip and knee arthroplasty journey: A qualitative interview study. J Clin Nurs 2019;28:4434-4446. [doi: 10.1111/jocn.15028] [Medline: 31408555]
33. Jansson MM, Harjumaa M, Puhto AP, Pikkarainen M. Healthcare professionals' perceived problems in fast-track hip and knee arthroplasty: results of a qualitative interview study. J Orthop Surg Res 2019 04;14(1):294 [FREE Full text] [doi: 10.1186/s13018-019-1334-3] [Medline: 31484536]
34. Kyngäs H. Inductive content analysis. In: Kyngäs H, Mikkonen K, Kääriäinen M, editors. The Application of Content Analysis in Nursing Science Research. 1st ed. Cham, Switzerland: Springer; 2020:13-21.
35. Kyngäs H, Kaakinen P. Deductive content analysis. In: Kyngäs H, Mikkonen K, Kääriäinen M, editors. The Application of Content Analysis in Nursing Science Research. 1st ed. Cham, Switzerland: Springer; 2020:23-30.
36. Guidelines for ethical review in human sciences. Finnish National Board on Research Integrity TENK. 2019. URL: https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-human-sciences [accessed 2023-06-06]
37. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
38. Hankey GJ. Population impact of potentially modifiable risk factors for stroke. Stroke 2020;51(3):719-728. [doi: 10.1161/STROKEAHA.119.024154] [Medline: 32078497]
39. Singh RJ, Chen S, Ganesh A, Hill MD. Long-term neurological, vascular, and mortality outcomes after stroke. Int J Stroke 2018;13(8):787-796. [doi: 10.1177/1747493018798526] [Medline: 30160619]
40. Anderson G, Rega ML, Casasanta D, Graffigna G, Damiani G, Barello S. The association between patient activation and healthcare resources utilization: a systematic review and meta-analysis. Public Health 2022;210:134-141. [doi: 10.1016/j.puhe.2022.06.021] [Medline: 35970015]
41. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019;19(1):400 [FREE Full text] [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
42. Mou H, Wong MS, Chien WT. Effectiveness of dyadic psychoeducational intervention for stroke survivors and family caregivers on functional and psychosocial health: a systematic review and meta-analysis. Int J Nurs Stud 2021;120:103969. [doi: 10.1016/j.ijnurstu.2021.103969] [Medline: 34052538]

XSL•FO
RenderX

43. Cadilhac DA, Kim J, Lannin NA, Kapral MK, Schwamm LH, Dennis MS, et al. National stroke registries for monitoring and improving the quality of hospital care: a systematic review. Int J Stroke 2016;11(1):28-40. [doi: 10.1177/1747493015607523] [Medline: 26763018]
44. Maramba ID, Jones R, Austin D, Edwards K, Meinert E, Chatterjee A. The role of health kiosks: scoping review. JMIR Med Inform 2022;10(3):e26511 [FREE Full text] [doi: 10.2196/26511] [Medline: 35348457]
45. Bashi N, Karunanithi M, Fatehi F, Ding H, Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews. J Med Internet Res 2017;19(1):e18 [FREE Full text] [doi: 10.2196/jmir.6571] [Medline: 28108430]
46. Piccini JP, Mittal S, Snell J, Prillinger JB, Dalal N, Varma N. Impact of remote monitoring on clinical events and associated health care utilization: a nationwide assessment. Heart Rhythm 2016;13(12):2279-2286. [doi: 10.1016/j.hrthm.2016.08.024] [Medline: 27544748]
47. Decker C, Chhatriwalla E, Gialde E, Garavalia B, Summers D, Quinlan ME, et al. Patient-centered decision support in acute ischemic stroke: qualitative study of patients' and providers' perspectives. Circ Cardiovasc Qual Outcomes 2015;8:S109-S116. [doi: 10.1161/CIRCOUTCOMES.115.002003] [Medline: 26515198]
48. Dunn Lopez K, Gephart SM, Raszewski R, Sousa V, Shehorn LE, Abraham J. Integrative review of clinical decision support for registered nurses in acute care settings. J Am Med Inform Assoc 2017;24(2):441-450 [FREE Full text] [doi: 10.1093/jamia/ocw084] [Medline: 27330074]
49. Winburn AS, Brixey JJ, Langabeer J, Champagne-Langabeer T. A systematic review of prehospital telehealth utilization. J Telemed Telecare 2018;24(7):473-481. [doi: 10.1177/1357633X17713140] [Medline: 29278996]
50. Keyworth C, Hart J, Armitage CJ, Tully MP. What maximizes the effectiveness and implementation of technology-based interventions to support healthcare professional practice? A systematic literature review. BMC Med Inform Decis Mak 2018;18(1):93 [FREE Full text] [doi: 10.1186/s12911-018-0661-3] [Medline: 30404638]
51. Varghese J, Kleine M, Gessner SI, Sandmann S, Dugas M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. J Am Med Inform Assoc 2018;25(5):593-602 [FREE Full text] [doi: 10.1093/jamia/ocx100] [Medline: 29036406]
52. Hyvämäki P, Sneck S, Meriläinen M, Pikkarainen M, Kääriäinen M, Jansson M. Interorganizational health information exchange-related patient safety incidents: a descriptive register-based qualitative study. Int J Med Inform 2023;174:105045 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105045] [Medline: 36958225]
53. Menachemi N, Rahurkar S, Harle CA, Vest JR. The benefits of health information exchange: an updated systematic review. J Am Med Inform Assoc 2018;25(9):1259-1265 [FREE Full text] [doi: 10.1093/jamia/ocy035] [Medline: 29718258]
54. Akay EMZ, Hilbert A, Carlisle BG, Madai VI, Mutke MA, Frey D. Artificial intelligence for clinical decision support in acute ischemic stroke: a systematic review. Stroke 2023;54(6):1505-1516 [FREE Full text] [doi: 10.1161/STROKEAHA.122.041442] [Medline: 37216446]
55. Stinear CM, Smith MC, Byblow WD. Prediction tools for stroke rehabilitation. Stroke 2019;50(11):3314-3322. [doi: 10.1161/STROKEAHA.119.025696] [Medline: 31610763]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
CVD: cerebrovascular disease
DHI: digital health intervention
EMR: electronic medical record
TIA: transient ischemic attack
WHO: World Health Organization

Edited by A Mavragani; submitted 13.09.23; peer-reviewed by M Cozad, V Pérez Jover; comments to author 23.02.24; revised version received 04.04.24; accepted 03.10.24; published 02.12.24.

Please cite as:
Härkönen H, Myllykangas K, Kärppä M, Rasmus KM, Gomes JF, Immonen M, Hyvämäki P, Jansson M
Perspectives of Clients and Health Care Professionals on the Opportunities for Digital Health Interventions in Cerebrovascular Disease Care: Qualitative Descriptive Study
J Med Internet Res 2024;26:e52715
URL: https://www.jmir.org/2024/1/e52715
doi:10.2196/52715
PMID:39622027

XSL•FO
RenderX

©Henna Härkönen, Kirsi Myllykangas, Mikko Kärppä, Kirsi Maaria Rasmus, Julius Francis Gomes, Milla Immonen, Piia Hyvämäki, Miia Jansson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Accuracy, Reproducibility, and Responsiveness to Treatment of Home Spirometry in Cystic Fibrosis: Multicenter, Retrospective, Observational Study

Martinus C Oppelaar[1], MD; Hanneke AC van Helvoort[2], PhD; Michiel AGE Bannier[3], MD, PhD; Monique HE Reijers[2], MD, PhD; Hester van der Vaart[4], MD, PhD; Renske van der Meer[5], MD, PhD; Josje Altenburg[6], MD, PhD; Lennart Conemans[7,8], MD; Bart L Rottier[9,10], MD, PhD; Marianne Nuijsink[11], MD, PhD; Lara S van den Wijngaart[1], MD, MSc, PhD; Peter JFM Merkus[1], MD, PhD; Jolt Roukema[1], MD, PhD

[1]Department of Pediatric Pulmonology, Amalia Children's Hospital, Radboud University Medical Center, Nijmegen, Netherlands

[2]Department of Pulmonary Diseases, Radboud University Medical Center, Nijmegen, Netherlands

[3]Department of Paediatric Pulmonology, MosaKids Children's Hospital, Maastricht University Medical Centre+, Maastricht, Netherlands

[4]Department of Pulmonary Diseases, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[5]Department of Pulmonology, Haga Teaching Hospital, The Hague, Netherlands

[6]Department of Respiratory Medicine, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, Netherlands

[7]Department of Respiratory Medicine, Maastricht University Medical Centre+, Maastricht, Netherlands

[8]Division of Respiratory & Age-related Health, Department of Respiratory Medicine, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht, Netherlands

[9]Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[10]Groningen Research Institute for Asthma and COPD, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[11]Juliana Children's Hospital, Haga Teaching Hospital, The Hague, Netherlands

Corresponding Author:
Martinus C Oppelaar, MD
Department of Pediatric Pulmonology
Amalia Children's Hospital
Radboud University Medical Center
Geert Grooteplein Zuid 10
Nijmegen, 6500 HB
Netherlands
Phone: +31 243614430
Email: marc.oppelaar@radboudumc.nl

Abstract

Background: Portable spirometers are increasingly used to measure lung function at home, but doubts about the accuracy of these devices persist. These doubts stand in the way of the digital transition of chronic respiratory disease care, hence there is a need to address the accuracy of home spirometry in routine care across multiple settings and ages.

Objective: This study aimed to assess the accuracy, reproducibility, and responsiveness to the treatment of home spirometry in long-term pediatric and adult cystic fibrosis care.

Methods: This retrospective observational study was carried out in 5 Dutch cystic fibrosis centers. Home spirometry outcomes (forced expiratory volume in one second [FEV_1], and forced vital capacity [FVC]) for 601 anonymized users were collected during 3 years. For 81 users, data on clinic spirometry and elexacaftor/tezacaftor/ivacaftor (ETI) use were available. Accuracy was assessed using Bland-Altman plots for paired clinic-home measurements on the same day and within 7 days of each other (nearest neighbor). Intratest reproducibility was assessed using the American Thoracic Society/European Respiratory Society repeatability criteria, the coefficient of variation, and spirometry quality grades. Responsiveness was measured by the percentage change in home spirometry outcomes after the start of ETI.

Results: Bland-Altman analysis was performed for 86 same-day clinic-home spirometry pairs and for 263 nearest neighbor clinic-home spirometry pairs (n=81). For both sets and for both FEV_1 and FVC, no heteroscedasticity was present and hence the

XSL•FO
RenderX

mean bias was expressed as an absolute value. Overall, home spirometry was significantly lower than clinic spirometry (mean $\Delta FEV_{1clinic\text{-}home}$ 0.13 L, 95% CI 0.10 to 0.19; mean $\Delta FVC_{clinic\text{-}home}$ 0.20 L, 95% CI 0.14 to 0.25) and remained lower than clinic spirometry independent of age and experience. One-way ANOVA with post hoc comparisons showed significantly lower differences in clinic-home spirometry in adults than in children (Δmean 0.11, 95% CI –0.20 to –0.01) and teenagers (Δmean 0.14, 95% CI –0.25 to –0.02). For reproducibility analyses, 2669 unique measurement days of 311 individuals were included. Overall, 87.3% (2331/2669) of FEV_1 measurements and 74.3% (1985/2669) of FVC measurements met reproducibility criteria. Kruskal-Wallis with pairwise comparison demonstrated that for both FVC and FEV_1, coefficient of variation was significantly lower in adults than in children and teenagers. A total of 5104 unique home measurements were graded. Grade E was given to 2435 tests as only one home measurement was performed. Of the remaining 2669 tests, 43.8% (1168/2669) and 43.6% (1163/2669) received grade A and B, respectively. The median percentage change in FEV_1 from baseline after initiation of ETI was 19.2% after 7-14 days and remained stable thereafter (n=33).

Conclusions: Home spirometry is feasible but not equal to clinic spirometry. Home spirometry can confirm whether lung functions remain stable, but the context of measurement and personal trends are more relevant than absolute outcomes.

(J Med Internet Res 2024;26:e60892) doi:10.2196/60892

KEYWORDS

telemonitoring; digital health; telespirometry; remote monitoring; cystic fibrosis; pediatrics; reliability; mobile phone; hereditary; chronic pulmonary inflammation; pulmonary infections; morbidity; mortality; chronic respiratory disease

Introduction

Cystic fibrosis (CF) is an autosomal recessive hereditary condition caused by a range of genetic defects in the gene coding for the cystic fibrosis transmembrane conductance regulator (CFTR) protein [1]. The defect leads to increased sputum viscosity which causes chronic pulmonary inflammation and predisposes to recurrent pulmonary infections. Early detection and treatment of pulmonary deterioration is imperative for people with CF to prevent morbidity and mortality [1,2]. The forced expiratory volume in one second (FEV_1) measured by spirometry is the most reliable parameter for early detection of this pulmonary deterioration [1,2].

Traditionally, spirometry is performed in a clinic guided by trained pulmonary function technicians, but small portable spirometers are also increasingly used by people with CF at home. Many of these portable spirometers have been validated and cleared by regulatory bodies such as the Federal Drug Agency and European Union authorities. However, both the unsupervised setting as well as the quality of these devices cause persistent doubts about the reliability of home spirometry among both health care professionals and people with CF [3]. These doubts stand in the way of the digital transition of CF care as well as of the wider field of chronic respiratory disease care.

Several recent studies show that home spirometry is consistently lower than clinic spirometry and has more intertest variability during stable periods [4-8]. In contrast, other studies argue that unsupervised home spirometry can yield similarly acceptable and reliable results as clinic spirometry [9-11]. Importantly, accuracy has almost exclusively been studied in prospective studies or within trials that are not reflective of real-life clinical practice. This might have introduced selection bias of skilled and motivated participants, and these studies might have missed changes in accuracy as home spirometry becomes repetitive for users over time.

This study aims to assess the accuracy, reproducibility, and responsiveness to treatment of home spirometry in long-term regular CF care in 5 Dutch CF centers between April 2020 and December 2022.

Methods

Study Design

This was a retrospective, multicenter, observational study on accuracy, reproducibility, and responsiveness to treatment of home spirometry in regular CF care performed in 5 Dutch CF centers (Radboud University Medical Center, Nijmegen; University Medical Center Groningen; Maastricht University Medical Center; Amsterdam University Medical Center; and Haga Hospital, The Hague).

We combined a large dataset of anonymous home spirometry outcomes dataset 1 (DS1) with a smaller dataset of clinical outcomes from participants in a remote monitoring study dataset 2 (DS2). DS1 consisted of all home spirometry outcomes of all users of the program (601 people with CF) from April 2020 until December 2022. DS2 consisted of clinical outcomes, including clinic spirometry, of 81 people with CF spread across the 5 participating centers who provided informed consent for the collection of their clinical data. Participants within DS2 could be coupled to their home spirometry outcomes in DS1 for accuracy analyses. Textbox 1 provides a summary of which datasets are used in separate analyses.

XSL•FO
RenderX

Textbox 1. Summary of datasets used and total eligible individuals per analysis

- Accuracy: Home spirometry measurements from dataset 1 combined with clinic spirometry measurements from dataset 2. Total eligible participants: n=81.
- Reproducibility and quality grading: All same-day home spirometry measurements from dataset 1. Total eligible participants: n=601
- Responsiveness: Home spirometry measurements from dataset 1 combined with individual elexacaftor/tezacaftor/ivacaftor start dates from dataset 2. Total eligible participants: n=81.

Portable Spirometer

The CE-marked Spirobank Smart portable spirometer (Medical International Research, MIR) is used in our population. The spirometer is connected to a dedicated smartphone app (Android, iOS). Spirometers were donated by the Dutch Cystic Fibrosis Foundation. Details and screenshots of the remote monitoring program (RMP) and spirometry module can be found in a previous publication [3]. The spirometers capture FEV_1, forced vital capacity (FVC), peak expiratory flow (PEF), FEV_1-to-FVC ratio, and forced expiratory flow at 25%-75%. The spirometer uses an algorithm that only allows submission of lung function tests that meet the 2005 American Thoracic Society/European Respiratory Society (ATS/ERS) spirometry criteria (volume max 10 L; volume accuracy ±3% or 0.05 L; flow range 960 L/minute; flow accuracy ±5% or 10.2 L/minute) [12,13].

CF centers carried responsibility for the implementation of the RMP and portable spirometers within their own needs and capacities [3]. In general, people with CF who were deemed able to reliably perform spirometry at home were eligible to receive a portable spirometer. People with CF were trained to use the device in the clinic, but home measurements itself were unsupervised. People with CF were generally instructed to perform 3 acceptable home spirometry maneuvers during every lung function measurement. The frequency of lung function measurements differed between hospitals and measurements were mostly patient-initiated, hence home spirometry was often used on indication (eg, during changing symptoms) [3].

Statistical Analyses

All statistical analyses were performed in SPSS Statistics (version 27; IBM Corp). Statistical methods are described per study objective.

Accuracy

For accuracy analyses, we paired clinic spirometry measurements from DS2 with home spirometry measurements from DS1 for participants in both datasets. Accuracy was assessed for home and clinic spirometry outcomes on the same day and with the nearest neighbor (NN) approach (home spirometry was compared with clinic spirometry performed within a maximum of 7 days from each other) [5]. For both pairs, Bland-Altman plots with 95% limits of agreement were used for FEV_1, FVC, PEF, and FEV_1/FVC ratio [14]. Differences between home and clinic FEV_1 ($\Delta FEV_{1clinic\text{-}home}$) and FVC ($\Delta FVC_{clinic\text{-}home}$) were considered clinically significant when they exceeded 150mL as defined by the limit of acceptable repeatability in the ATS/ERS spirometry consensus statement [15].

One-way ANOVA was used to compare $\Delta FEV_{1clinic\text{-}home}$ and $\Delta FVC_{clinic\text{-}home}$ between age groups (6-12 years, 12-18 years, >18 years). $\Delta FEV_{1clinic\text{-}home}$ and $\Delta FVC_{clinic\text{-}home}$ were also plotted by age groups, by the time of use in days, and by the percentage of individual participants' contribution to the overall amount of paired measurements (ie, less than 1 percent, between 1 and 3 percent, more than 3 percent).

Reproducibility

All anonymous home spirometry outcomes from DS1 were used to assess the reproducibility of same-day home spirometry measurements (intratest reproducibility). For measurements with 2 or more home spirometry maneuvers, the 2 to 3 best FEV_1 or FVC maneuvers were assessed separately. Reproducibility was assessed using repeatability criteria of the ATS/ERS consensus statement for FEV_1 and FVC which states that measurements are repeatable when the difference between the largest and second-largest FEV_1 or FVC is smaller or equal to 0.150 L [15]. We also calculated the coefficient of variation (CV; intratest SD/intratest mean 100). CV between age categories was analyzed using the Kruskal-Wallis test. To examine reproducibility over time, the reproducibility of home spirometry was also assessed in 3 subgroups: measurements performed between 0-100 days after initiation, between 300-400 days after initiation, and between 600-700 days after initiation.

Quality Grading

All home spirometry tests in DS1 were graded according to the ATS/ERS technical statement [15,16]. In short, grade A includes ≥3 acceptable tests within 0.150 L; grade B includes 2 acceptable tests within 0.150 L; grade C includes ≥2 acceptable tests within 0.200 L; grade D includes ≥2 acceptable tests within 0.250 L; and grade E includes ≥2 acceptable tests within more than 0.250 L or one acceptable test [15,16]. Grading criteria were applied to the highest FEV_1 and FVC results separately. Grade F (ie, no acceptable tests) tests could not be assessed as the Spirobank Smart only allows submission of maneuvers that meet ATS/ERS quality criteria.

Responsiveness to Treatment

Individual start dates of a new CFTR modulator drug elexacaftor/tezacaftor/ivacaftor (ETI) were available for participants in DS2. The absolute percentage change from baseline in home spirometry outcomes in DS1 of these participants was analyzed. Baseline FEV_1 or FVC was defined as the average FEV_1 or FVC of all home spirometry outcomes during the 2 months before the start of ETI. Absolute percentage change in FEV_1 and FVC from baseline were then analyzed during multiple time periods after initiation of ETI: 0-7 days

XSL•FO
RenderX

after start, 7-14 days after start, 14-21 days after start, 21-28 days after start, and 28 days to 2 months after start.

Ethical Considerations

This study was part of a larger study of remote monitoring in regular CF care for which approval of the Medical Research Involving Human Subjects Act was waived by the local ethical committees [3] (file number for local ethical committee Arnhem-Nijmegen region: 2021-13214). All participants have provided informed e-consent for the use of their data to answer the research questions of this study. Data were pseudonymized using an encrypted code to which only treating health care professionals had access.

Results

Demographics

Demographics for participants in DS2 are presented in Table 1. Demographics for participants in DS1 were limited considering data of participants and centers was anonymized. The median age in DS1 was 26 (IQR 16-37) years, age was missing for 9 participants. The sex distribution in DS1 was 51% (307/601) male, 44% (263/601) female, and 5% (31/601) unspecified. The median number of participants per center in DS1 was 127 (IQR 82-153).

Table 1. Demographics of people with cystic fibrosis in dataset 2.

People with CF[a] (n=81)	Values
Male sex, n (%)	40 (49.4)
Age at enrolment on the RMP[b], year, median (IQR)	27 (12-40.5)
Age distribution at enrolment on the RMP, n (%)	
<12 years	17 (21)
12 to <18 years	9 (11)
18 to <30 years	19 (24)
≥30 years	36 (44)
Age range (years) at enrolment on the RMP, range	5 to 59
Hospital, n (%)	
Radboud University Medical Center	14 (17)
HagaZiekenhuis	16 (20)
Maastricht University Medical Center+	26 (32)
University Medical Center Groningen	12 (15)
Amsterdam University Medical Center	13 (16)
CFTR [c] genotype, n (%)	
F508del homozygous	49 (61)
F508del heterozygous	27 (33)
Other	5 (6)
Highest FEV_1 [d] in 2020, (n=77)	
z score FEV_1, median (IQR)	–2.0 (–4.4 to –0.9)
FEV_1 percentage predicted, median (IQR)	75.1 (58.2-89.6)

[a]CF: cystic fibrosis.
[b]RMP: remote monitoring program.
[c]CFTR: cystic fibrosis transmembrane conductance regulator.
[d]FEV_1: forced expiratory volume in one second.

Accuracy

In total, 86 and 263 pairs of home and clinic spirometry were made for same-day and NN measurements, respectively. For same-day pairs, 49 unique individuals had a median of 1 (IQR 1-2) pair. For NN pairs, 69 unique individuals had a median of 3 (IQR 1-5) pairs. Bland-Altman plots showed no visual heteroscedasticity (ie, variability of the differences remained similar as the magnitude of the mean value increased [14]) for NN pairs (Figure 1) nor same-day pairs (Multimedia Appendix 1) for all outcomes. Therefore, mean differences should be expressed as absolutes rather than percentages [14]. On average, home FEV_1 and FVC were significantly lower than clinic spirometry for both NN and same-day pairs (Table 2). Limits of agreements were slightly narrower for same-day FEV_1 and

XSL•FO
RenderX

FVC pairs compared with NN pairs but remained similar for PEF and FEV_1/FVC (Table 2).

To assess the effect of repeated measures, separate Bland-Altman plots using the participants' individual contributions to the total spirometry pairs were performed. Within the contribution plots, individual participants were well mixed with each other, and variation in $\Delta FEV_{1clinic\text{-}home}$ and $\Delta FVC_{clinic\text{-}home}$ remained similar meaning an aggregate Bland-Altman analysis was deemed appropriate (Multimedia Appendix 2). In separate plots for age categories, adults had more $\Delta FEV_{1clinic\text{-}home}$ and $\Delta FVC_{clinic\text{-}home}$ pairs below the line of equality (FEV_1: 34.8%, 40/115; FVC 28.7%, 33/115) than children (FEV_1: 15.2%, 15/99; FVC: 23.2%, 23/99) and adolescents (FEV_1: 24.5%, 12/49; FVC: 20.4%, 10/49) meaning clinic spirometry was more frequently lower than home spirometry in adults than in children or teenagers (Multimedia Appendix 3). Figure 2 shows Bland-Altman plots for NN FEV_1 and FVC comparing $\Delta FEV_{1clinic\text{-}home}$ and $\Delta FVC_{clinic\text{-}home}$ by the number of days of experience users had with the portable spirometer.

One-way ANOVA showed a significant difference between age categories and NN $\Delta FEV_{1clinic\text{-}home}$ ($F_{2,121.66}$=5.61, P<.01) but not for NN $\Delta FVC_{clinic\text{-}home}$, same day $\Delta FVC_{clinic\text{-}home}$, and same day $\Delta FEV_{1clinic\text{-}home}$. Gabriel post hoc test showed that adults (≥18 years, n=115) had on average lower $\Delta FEV_{1clinic\text{-}home}$ than children (6-12 years, n=99) (Δmeans=0.11 L, 95% CI –0.20 to –0.01, P=.02) and teenagers (12-18 years, n=49; Δmeans=0.14, 95% CI –0.25 to –0.02, P=.01). There was no statistically significant difference between children and teenagers (P=.93).

Figure 1. Bland-Altman plots of clinic measurements (n=263) compared with nearest neighbor home measurement (±7 days) for spirometry outcomes. From upper left to lower right: FEV1, FVC, PEF, and FEV1/FVC. X-axis: mean of clinic and home measurement; y-axis: clinic minus home measurement. N=263. Dotted line: reference line of equality (ie, Δclinic-home=0); Solid line: average bias (ie, average Δclinic-home); Dashed line: 95% limits of agreement. FEV1: forced expiratory volume in 1 second; FVC: forced vital capacity; PEF: peak expiratory flow.

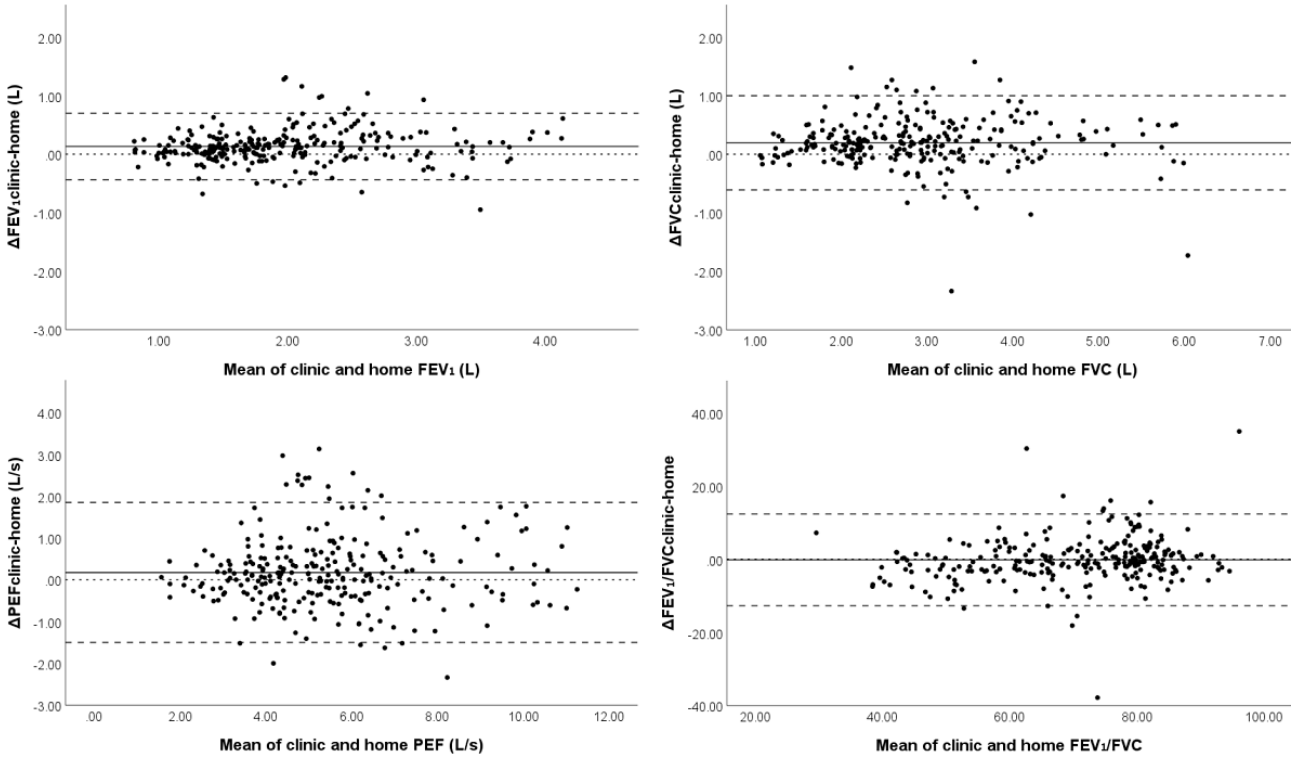

XSL•FO
RenderX

Table 2. Accuracy outcomes for forced expiratory volume in 1 second, forced vital capacity, and peak expiratory flow for paired home and clinic spirometry measurements on the same day and within 7 days from each other (nearest neighbor).

	Mean bias (sample mean Δclinic-home)	Estimated 95% CI of population mean Δclinic-home	95% limits of agreement of mean Δclinic-home
Forced expiratory volume in 1 second			
Same day (n=86)	0.10 L[a]	0.05 to 0.14[a]	–0.32 to 0.52
Nearest neighbor (n=263)	0.13 L[a]	0.10 to 0.17[a]	–0.44 to 0.70
Forced vital capacity			
Same day (n=86)	0.22 L[a]	0.15 to 0.30[a]	–0.48 to 0.92
Nearest neighbor (n=263)	0.20 L[a]	0.14 to 0.25[a]	–0.61 to 1.00
Peak expiratory flow			
Same day (n=86)	0.12 L	–0.06 to 0.31	–1.60 to 1.85
Nearest neighbor (n=263)	0.17 L[a]	0.07 to 0.28[a]	–1.50 to 1.85
Forced expiratory volume in 1 second/forced vital capacity (Tiffeneau-index)			
Same day (n=86)	–0.53	–1.84 to 0.78	–12.49 to 11.43
Nearest neighbor (n=263)	–0.15	–0.92 to 0.63	–12.70 to 12.41

[a]95% CI of mean absolute difference is outside the line of equality in Bland-Altman plots.

Figure 2. Bland-Altman plots of the absolute difference between clinic spirometry and nearest neighbor home spirometry (±7 days) by the number of days after the individual participants' start of the home spirometer. Upper plot: FEV1; Lower plot: FVC. The color of points indicates age categories at the time of measurement. Light gray: 6 to 12 years old; gray 12 to 18 years old; and black: 18 years or older. N=263. Dotted line: reference line of equality (ie, Δclinic-home=0); Solid line: average bias (ie average Δclinic-home); Dashed line: 95% limits of agreement. FEV1: forced expiratory volume in 1 second; FVC: forced vital capacity.

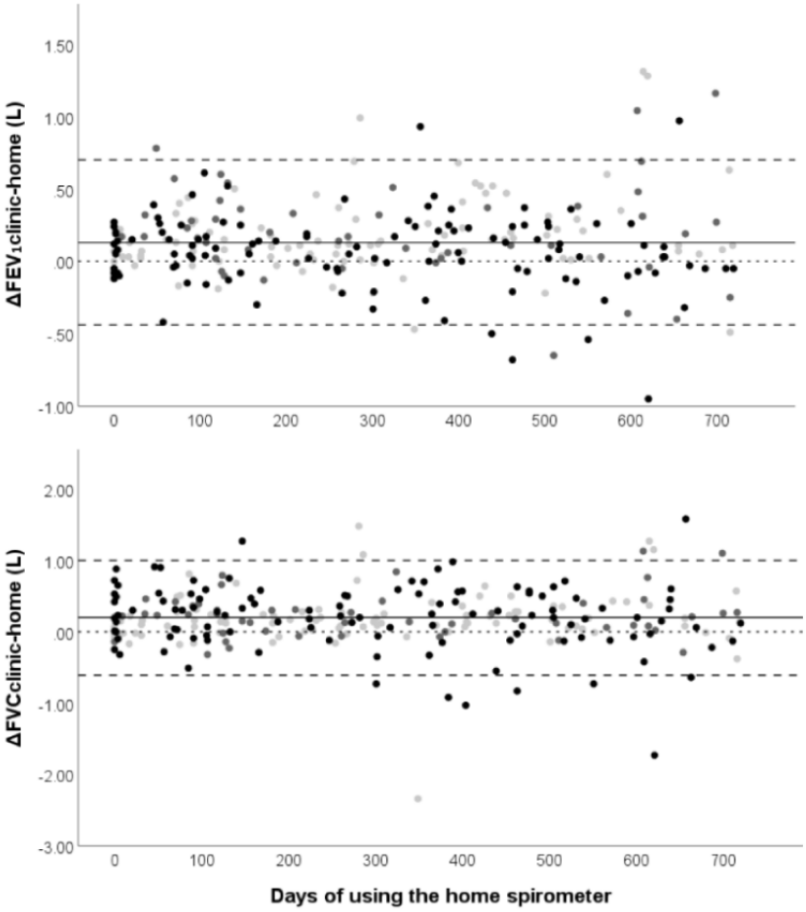

XSL•FO
RenderX

Reproducibility

In total, there were 2669 unique measurement days of 311 people with CF with 2 or more home spirometry measurements on the same day. The median of unique measurement days per participant was 4 (IQR 2–10; range 1-265). Overall, 87.3% (2331/2669) of unique measurement days met repeatability criteria (ie, the difference between the largest and second-largest FEV_1 or FVC ≤0.150L) for FEV_1 compared with 74.3% (1985/2669) for FVC. In a sensitivity analysis, only the last measurement of every individual was included: 81% (252/311) of unique measurement days met repeatability criteria for FEV_1 compared with 62.4% (194/311) for FVC. Median intraperson repeatability rates were 100% (IQR 89%-100%) for FEV_1 and 72% (IQR 33%-96%) for FVC. The median intraperson CV was 2.80% (IQR 1.93%-4.69%) for FEV_1 and 3.92% (IQR 2.45%-6.43%) for FVC.

Repeatability rates and CV between age groups are summarized in Table 3. Kruskal-Wallis with pairwise comparison demonstrated that both for FEV_1 and FVC, CV was significantly affected by age categories (FEV_1: $H(2)=140.94$, $P<.001$; FVC: $H(2)=64.28$, $P<.001$). For FEV_1, pairwise comparisons with adjusted P values showed that there were small statistically significant differences in CV among all age categories: children versus adults ($P<.001$, $r=0.24$), teenagers versus adults ($P<.001$, $r=0.16$), and children versus teenagers ($P<.001$, $r=0.10$). For FVC, pairwise comparisons also revealed small significant differences in CV between children and adults ($P<.001$; $r=0.16$) and teenagers and adults ($P<.001$; $r=0.10$). There were no significant differences between children and teenagers ($P=.06$; $r=0.07$).

Table 3. Reproducibility outcomes within subgroups.

	Acceptable repeatability		Coefficient of variation	
	FEV_1[a], n (%)	FVC[b], n (%)	FEV_1 (%), median (IQR)	FVC (%), median (IQR)
Age categories				
Children (6-12 years, n=492)	424 (86.2)	392 (79.7)	3.49 (1.88-6.33)	3.74 (1.82-6.20)
Teenagers (12-18 years, n=504)	386 (76.6)	348 (69)	2.82 (1.34-5.78)	3.02 (1.56-5.67)
Adults (18 years or older, n=1583)	1437 (90.8)	1174 (74.2)	1.91 (1.05-3.46)	2.41 (1.27-4.53)
Days of using the home spirometer				
0-100 days (n=676)	574 (84.9)	468 (69.2)	2.80 (1.36-5.29)	3.31 (1.60-6.30)
300-400 days (n=311)	283 (91)	247 (79.4)	2.13 (1.18-3.79)	2.36 (1.40-4.54)
600-700 days (n=178)	155 (87.1)	131 (73.6)	2.23 (1.18-4.01)	2.53 (1.32-5.23)

[a]FEV_1: forced expiratory volume in 1 second.

[b]FVC: forced vital capacity.

Quality Grading

Overall, 47.7% (2435/5104) home spirometry tests received grade E for both FEV_1 and FVC because only one measurement was performed. For tests with more than one measurement, FEV_1 grades were A (43.8%, 1168/2669), B (43.6%, 1163/2269), C (4.5%, 119/2669), D, (2.1%, 56/2669), and E (6.1%, 163/2669). FVC grades were A (28%, 748/2669), B (46.3%, 1237/2669), C (7.6%, 203/2669), D (4.9%, 132/2669), and E (13.1%, 349/2669). For individuals' final tests only, FEV_1 grades were A (27.7%, 86/311), B (53.4%, 166/311), C (6.8%, 21/311), D (2.9%, 9/311); and E (9.3%, 29/311). Final FVC grades were A (15.4%, 48/311), B (46.9%, 146/311), C (8.7%, 27/311), D (6.4%, 20/311), and E (22.5%, 70/311).

Responsiveness to Treatment

A total of 73 people with CF in DS2 had started ETI during the study period, of which 33 had at least one baseline home and follow-up home spirometry measurement. Median absolute percentage changes from baseline after initiation of ETI were as follows: FEV_1: 11.3% (IQR 4.2%-19.2%), FVC: 4.5% (IQR –2.2%-15.8%) after 0-7 days; FEV_1: 19.2% (IQR 7.9%-34.8%), FVC: 14.3% (IQR 3.3%-19.1%) after 7-14 days; and thereafter FVC and FEV_1 remained relatively stable (Figure 3). After 2 weeks, 66.7% (12/18) of users who had measured their lung function improved more than 10% in home FEV_1.

XSL•FO
RenderX

Figure 3. Boxplots visualizing the absolute percentage change from baseline for FEV1 (upper panel) and FVC (lower panel) after initiation of ETI. Dashed line: minimal clinically important difference for FEV1 and FVC (10% change). The total amount of participants is 33. Baseline FEV1 and FVC were calculated as the average FEV1 and FVC within the first 2 months before initiation of ETI. ETI: elexacaftor/tezacaftor/ivacaftor; FEV1: forced expiratory volume in 1 second; FVC: forced vital capacity.

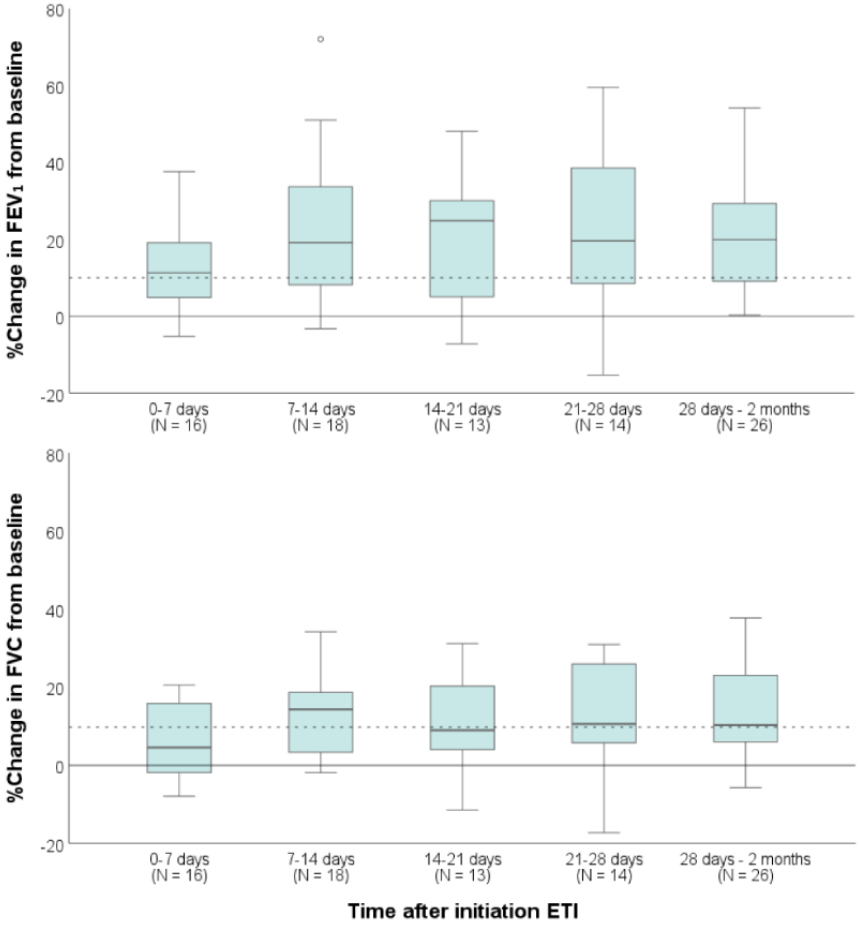

Discussion

Principal Findings

This study aimed to assess the accuracy, reproducibility, and responsiveness to the treatment of home spirometry in long-term CF care. Our findings show that home spirometry outcomes are significantly lower than clinic spirometry in both children and adults with broad 95% limits of agreement that surpass repeatability criteria. For $\Delta FVC_{clinic\text{-}home}$ but not $\Delta FEV_{1clinic\text{-}home}$, the mean difference also surpassed the 150mL limit of repeatability [15,16]. For all assessed outcomes, there was no heteroscedasticity in our population meaning that the mean bias and limits of agreement are similar for low and high absolute values, which has important implications for people with CF with lower lung volumes. The reproducibility of home spirometry was high, especially for FEV_1, but the quality of tests could be improved by increasing the amount of maneuvers during home spirometry measurements. Finally, home spirometry was responsive to ETI treatment and showed clinically significant changes in lung function in 2 thirds of participants.

Comparison to Previous Work

Our findings closely resemble those by Edmondson et al [7] who presented a similar mean bias of 0.14 L with limits of agreement of ±0.40 L for unsupervised home FEV_1 in children using a Vitalograph portable spirometer without training effects after 6 months. They also found a significant systemic underread of the portable spirometers as tested with calibrated syringes [7]. Bell et al [10] found that virtual supervision of home spirometry with Air-Next or Spirohome spirometers was not different from unsupervised FEV_1 home measurements. Taken together, multiple sources of bias and combinations could be underlying lower home spirometry outcomes. First, portable spirometers could have a systematic underread bias but it is unclear whether this is present in all different models, devices, and applications, and how this is affected by a lack of daily calibration such as performed in the clinic. Future research should assess whether periodical calibration of home spirometers during outpatient visits is possible and whether it improves long-term accuracy. Second, a lack of supervision could still cause lower home spirometry results for some people with CF as studies that compare unsupervised and supervised measurements are constrained by Hawthorne effects and selection biases. Finally, our findings show increased variability of home spirometry after multiple years of use. People with CF,

XSL•FO
RenderX

especially children, have previously reported that home spirometry was more difficult than clinic spirometry and long-term use might introduce imprudence [3]. Therefore, repeated training might be necessary to prevent further deviation rather than improving underread biases, but this should be subject to further investigation.

We found that children and teenagers more often had somewhat lower at-home FEV_1 and FVC than adults compared to those measured by in-clinic spirometry. Only one study has previously compared children with adults and reported that adults and teenagers had less variable and more reliable home spirometry than children [11]. Other studies have also argued that children perform better than teenagers, which has been ascribed to more supervision from parents [6,7]. However, these findings were not replicated in our study. We previously found that children have more difficulty with home spirometry as the technique is different and the mouthpiece is relatively large for them [3] Moreover, children and teenagers appeared less patient and motivated for spirometry within their private setting than adults, especially when parents press them to perform measurements at home [3]. Therefore, impatience and imprudence, but also issues with technique or the large mouthpiece offer alternative explanations for less reliable results in children compared to adults.

Overall, the intratest reproducibility of home spirometry was good for FEV_1 with statistical but clinically irrelevant differences between age categories. The experience did not seem to impact intratest reproducibility outcomes. FVC had less favorable intratest reproducibility and quality grades than FEV_1, possibly because it requires a more challenging maneuver. For FEV_1, quality grades in our study were similar to those found earlier in adults, but quality grades for home FVC in CF have not yet been reported [10]. The most important improvement in quality would be better instruction to include more maneuvers in measurements as most measurements only included one maneuver.

Finally, we found that within 2 weeks after initiation of ETI, a median increase of 20% in FEV_1 and 15% in FVC from baseline could be quantified with home spirometry in a subgroup of users. Unfortunately, we were unable to calculate *z* scores or percentage predictive values for these home measurements in this study which made statistical analyses and comparison with literature difficult [17,18]. Nevertheless, 2 thirds of users improved more than 10% in FEV_1, which is a clinically relevant change in lung function [2,15,16]. Considering that home spirometers consistently underestimate clinic spirometry rather than overestimate, a clinically significant improvement (ie, >10%) in short-term home spirometry is still highly relevant and will likely be even larger in the clinic. This may make home spirometers better suited to confirm short-term treatment responses remotely than to rule out treatment responses. The use of home spirometry in treatment responses should be subject to further study.

Strengths and Limitations

To our knowledge, this is the largest study on accuracy, reproducibility, and responsiveness of home spirometry that was performed within a multicenter and long-term setting and also included participants from all age categories. Therefore, our findings will likely more closely resemble actual home spirometry accuracy as there was less selection bias or Hawthorne effects compared to prospective studies. This study had 2 important limitations. First, the use of anonymous data meant there was no information on pulmonary exacerbations or the use of therapies such as antibiotic courses or short-acting beta-agonist therapies which might cause differences between home and clinic measurements. We aimed to mitigate these effects by analyzing both NN and same-day accuracy. As the same day and NN findings closely resembled each other, we are confident that the effect of this was inconsequential in our study. Second, a small group of individuals had missing data on age which could not be corrected due to the use of anonymous data and these individuals had to be excluded from age group comparisons.

Clinical Implications

Home spirometry, especially home FEV_1, can be performed acceptably at home but is not equal to clinic spirometry. FVC home measurements might pose more challenges as a remotely measured outcome considering the average bias surpassed repeatability criteria, 95% limits of agreement were broad, and acceptability grades were lower. Consequently, home spirometry can offer an alternative to clinic spirometry in the follow-up of people with CF or to provide point-of-care lung function data remotely, but home spirometry should not be used to make diagnoses or to detect subtle changes in pulmonary function.

Home spirometry can be used to confirm that a patient's lung function is stable when outcomes are comparable to recent clinic spirometry. However, the context of a measurement as well as the deviation from a personal baseline might be more clinically relevant than the actual home spirometry values [3] Therefore, education of patients about the possibility of lower values and the need for home spirometry measurements with more than one maneuver on regular intervals is essential. This will prevent concerns or disappointments of patients and enable the assessment of deterioration or improvement from personal baselines. Finally, home spirometers should regularly be taken into the clinic during routine outpatient visits to assess individual deviation from clinic measurements and to evaluate the technique by professional pulmonary function technicians.

Conclusions

Home spirometry offers a feasible opportunity to collect lung function outcomes remotely but is not equal to clinic spirometry. Home spirometry can be offered to people with CF who are familiar with pulmonary function tests as a tool for remote follow-up during periods of stability, for remote point-of-care lung function outcomes during deterioration, or to enable and evaluate home treatments. Considering the consistent underread, home spirometry might be well suited to rule out deterioration when values remain stable but should be interpreted with caution when values are lower than expected. Future research should focus on the role of training and user familiarity with portable spirometry devices on the long-term reliability of home spirometry and on the role of home spirometry in assessing treatment responses in CF and other chronic respiratory diseases.

XSL•FO
RenderX

Acknowledgments

We would like to thank the Dutch CF Foundation for donating the portable spirometers. We would like to thank the Dutch CF Foundation and the Steering Group for supplying the Dutch CF Registry data used during our evaluation. No generative artificial intelligence was used at any stage of this study.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in the conceptualization of this study and the inclusion of participants. MCO, HvH, JR, and PJFMM were responsible for the study design. MCO managed the project administration, visualization of results, writing of the manuscript, and quantitative analyses supervised by HvH, PJFMM, and JR. All authors were involved in reviewing the manuscript.

Conflicts of Interest

Authors MCO, LSvdW, JR, and PJFMM are involved in the development of the software used in this study. The software is the property of the Radboud University Medical Center and hence does not generate any profits for the authors.

Multimedia Appendix 1
Bland-Altman plots for same day clinic and home spirometry pairs. From upper left to lower right: FEV1, FVC, PEF, and FEV1/FVC.
[DOCX File , 62 KB - jmir_v26i1e60892_app1.docx]

Multimedia Appendix 2
Bland-Altman plots for nearest neighbor FEV1 and FVC with individual clinic and home spirometry pairs categorized by the participants' individual contributions to the total amount of spirometry pairs in the dataset. Blue: the participant contributed less than one percent of the overall amount of spirometry pairs; Green: the participant contributed between one and three percent of spirometry pairs; Red: the participant contributed more than three percent of spirometry pairs. Panel 1: FEV1; panel 2: FVC.
[DOCX File , 89 KB - jmir_v26i1e60892_app2.docx]

Multimedia Appendix 3
Bland-Altman plots for nearest neighbor FEV1 with individual clinic and home spirometry pairs.
[DOCX File , 270 KB - jmir_v26i1e60892_app3.docx]

References

1. Shteinberg M, Haq IJ, Polineni D, Davies JC. Cystic fibrosis. The Lancet 2021;397(10290):2195-2211. [doi: 10.1016/s0140-6736(20)32542-3]
2. Castellani C, Duff AJA, Bell SC, Heijerman HGM, Munck A, Ratjen F, et al. ECFS best practice guidelines: the 2018 revision. J Cyst Fibros 2018;17(2):153-178 [FREE Full text] [doi: 10.1016/j.jcf.2018.02.006] [Medline: 29506920]
3. Oppelaar MC, Emond Y, Bannier MAGE, Reijers MHE, van der Vaart H, van der Meer R, et al. Potential, pitfalls, and future directions for remote monitoring of chronic respiratory diseases: multicenter mixed methods study in routine cystic fibrosis care. J Med Internet Res 2024;26:e54942 [FREE Full text] [doi: 10.2196/54942] [Medline: 39106098]
4. Thornton CS, Magaret AS, Carmody LA, Kalikin LM, Simon RH, LiPuma JJ, et al. Quantifying variation in home spirometry in people with cystic fibrosis during baseline health, and associations with clinical outcomes. J Cyst Fibros 2024;23(2):321-328. [doi: 10.1016/j.jcf.2023.05.011] [Medline: 37244842]
5. Paynter A, Khan U, Heltshe SL, Goss CH, Lechtzin N, Hamblett NM. A comparison of clinic and home spirometry as longtudinal outcomes in cystic fibrosis. J Cyst Fibros 2022;21(1):78-83 [FREE Full text] [doi: 10.1016/j.jcf.2021.08.013] [Medline: 34474987]
6. Gerzon F, Jöbsis Q, Bannier M, Winkens B, Dompeling E. Discrepancy between lung function measurements at home and in the hospital in children with asthma and CF. J Clin Med 2020;9(6):1617 [FREE Full text] [doi: 10.3390/jcm9061617] [Medline: 32466623]
7. Edmondson C, Westrupp N, Short C, Seddon P, Olden C, Wallis C, et al. Unsupervised home spirometry is not equivalent to supervised clinic spirometry in children and young people with cystic fibrosis: results from the CLIMB-CF study. Pediatr Pulmonol 2023;58(10):2871-2880. [doi: 10.1002/ppul.26602] [Medline: 37503909]

XSL•FO
RenderX

8. Davis J, Ryan M, Marchetti P, Dahlberg SE, Greenberg J, Bacon C, et al. Real-world feasibility of short-term, unsupervised home spirometry in CF. Pediatr Pulmonol 2022;57(12):3129-3135. [doi: 10.1002/ppul.26147] [Medline: 36124390]
9. Fettes E, Riley M, Brotherston S, Doughty C, Griffiths B, Laverty A, et al. "You're on mute!" Does pediatric CF home spirometry require physiologist supervision? Pediatr Pulmonol 2022;57(1):278-284. [doi: 10.1002/ppul.25708] [Medline: 34581507]
10. Bell JM, Sivam S, Dentice RL, Dwyer TJ, Jo HE, Lau EM, et al. Quality of home spirometry performance amongst adults with cystic fibrosis. J Cyst Fibros 2022;21(1):84-87 [FREE Full text] [doi: 10.1016/j.jcf.2021.10.012] [Medline: 34774443]
11. Beaufils F, Enaud R, Gallode F, Boucher G, Macey J, Berger P, et al. Adherence, reliability, and variability of home spirometry telemonitoring in cystic fibrosis. Front Pediatr 2023;11:1111088 [FREE Full text] [doi: 10.3389/fped.2023.1111088] [Medline: 36911035]
12. Medical International Research. Spirobank Smart. User Manual Rev 2.4 2018:1-27.
13. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, ATS/ERS Task Force. Standardisation of spirometry. Eur Respir J 2005;26(2):319-338 [FREE Full text] [doi: 10.1183/09031936.05.00034805] [Medline: 16055882]
14. Giavarina D. Understanding bland altman analysis. Biochem Med (Zagreb) 2015;25(2):141-151 [FREE Full text] [doi: 10.11613/BM.2015.015] [Medline: 26110027]
15. Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, et al. Standardization of spirometry 2019 update. An Official American thoracic society and European respiratory society technical statement. Am J Respir Crit Care Med 2019;200(8):e70-e88. [doi: 10.1164/rccm.201908-1590st]
16. Culver BH, Graham BL, Coates AL, Wanger J, Berry CE, Clarke PK, et al. Recommendations for a standardized pulmonary function report. An Official American thoracic society technical statement. Am J Respir Crit Care Med 2017;196(11):1463-1472. [doi: 10.1164/rccm.201710-1981st]
17. Sutharsan S, McKone EF, Downey DG, Duckers J, MacGregor G, Tullis E, et al. Efficacy and safety of elexacaftor plus tezacaftor plus ivacaftor versus tezacaftor plus ivacaftor in people with cystic fibrosis homozygous for F508del-CFTR: a 24-week, multicentre, randomised, double-blind, active-controlled, phase 3b trial. The Lancet Respiratory Medicine 2022;10(3):267-277. [doi: 10.1016/s2213-2600(21)00454-9]
18. Mall MA, Brugha R, Gartner S, Legg J, Moeller A, Mondejar-Lopez P, et al. Efficacy and safety of elexacaftor/tezacaftor/ivacaftor in children 6 through 11 years of age with cystic fibrosis heterozygous for and a Minimal Function Mutation: A Phase 3b, Randomized, Placebo-controlled Study. Am J Respir Crit Care Med 2022;206(11):1361-1369. [doi: 10.1164/rccm.202202-0392oc]

Abbreviations

ATS/ERS: American Thoracic Society/European Respiratory Society
CF: cystic fibrosis
CFTR: cystic fibrosis transmembrane conductance regulator
CV: coefficient of variation
DS1: dataset 1
DS2: dataset 2
ETI: elexacaftor/tezacaftor/ivacaftor
FEV1: forced expiratory volume in one second
FVC: forced vital capacity
NN: nearest neighbor
PEF: peak expiratory flow
RMP: remote monitoring program

Edited by A Coristine; submitted 24.05.24; peer-reviewed by JF Heidenreich, L Wucherpfennig, SH Donaldson; comments to author 15.08.24; revised version received 30.08.24; accepted 30.09.24; published 03.12.24.

Please cite as:
Oppelaar MC, van Helvoort HAC, Bannier MAGE, Reijers MHE, van der Vaart H, van der Meer R, Altenburg J, Conemans L, Rottier BL, Nuijsink M, van den Wijngaart LS, Merkus PJFM, Roukema J
Accuracy, Reproducibility, and Responsiveness to Treatment of Home Spirometry in Cystic Fibrosis: Multicenter, Retrospective, Observational Study
J Med Internet Res 2024;26:e60892
URL: https://www.jmir.org/2024/1/e60892
doi:10.2196/60892
PMID:

XSL•FO
RenderX

©Martinus C Oppelaar, Hanneke AC van Helvoort, Michiel AGE Bannier, Monique HE Reijers, Hester van der Vaart, Renske van der Meer, Josje Altenburg, Lennart Conemans, Bart L Rottier, Marianne Nuijsink, Lara S van den Wijngaart, Peter JFM Merkus, Jolt Roukema. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Postimplementation Evaluation in Assisted Living Facilities of an eHealth Medical Device Developed to Predict and Avoid Unplanned Hospitalizations: Pragmatic Trial

Jacques-Henri Veyron[1], MSc; François Deparis[2], MD; Marie Noel Al Zayat[1], MS; Joël Belmin[3,4], PhD, MD; Charlotte Havreng-Théry[1,3], PhD

[1]Presage, Paris, France

[2]Arpavie, Issy-les-Moulineaux, France

[3]Laboratoire Informatique Médicale et Ingénierie des Connaissances en e-santé (UMRS 1142), Institut National de la Santé et de la Recherche Médicale, Sorbonne Université, Paris, France

[4]Hôpital Charles Foix, Assistance Publique-Hôpitaux de Paris, Ivry-sur-Seine, France

Corresponding Author:
Jacques-Henri Veyron, MSc
Presage
112-114 rue la Boétie
Paris, 75008
France
Phone: 33 0622152004
Email: jhveyron@presage.care

Abstract

Background: The proportion of very old adults in the population is increasing, representing a significant challenge. Due to their vulnerability, there is a higher frequency of unplanned hospitalizations in this population, leading to adverse events. Digital tools based on artificial intelligence (AI) can help to identify early signs of vulnerability and unfavorable health events and can contribute to earlier and optimized management.

Objective: This study aims to report the implementation in assisted living facilities of an innovative monitoring system (Presage Care) for predicting the short-term risk of emergency hospitalization. We describe its use and assess its performance.

Methods: An uncontrolled multicenter intervention study was conducted between March and August 2022 in 7 assisted living facilities in France that house very old and vulnerable adults. The monitoring system was set up to provide alerts in cases of a high risk of emergency hospitalization. Nurse assistants (NAs) at the assisted living facilities used a smartphone app to complete a questionnaire on the functional status of the patients, comprising electronic patient-reported outcome measures (ePROMs); these were analyzed in real time by a previously designed machine learning algorithm. This remote monitoring of patients using ePROMs allowed notification of a coordinating nurse or a coordinating NA who subsequently informed the patient's nurses or physician. The primary outcomes were the acceptability and feasibility of the monitoring system in the context and confirmation of the effectiveness and efficiency of AI in risk prevention and detection in practical, real-life scenarios. The secondary outcome was the hospitalization rate after alert-triggered interventions.

Results: In this study, 118 of 194 (61%) eligible patients were included who had at least 1 follow-up visit. A total of 38 emergency hospitalizations were documented. The system generated 92 alerts for 47 of the 118 (40%) patients. Of these 92 alerts, 46 (50%) led to 46 health care interventions for 14 of the 118 (12%) patients and resulted in 4 hospitalizations. The other 46 of the 92 (50%) alerts did not trigger a health care intervention and resulted in 25 hospitalizations ($P<.001$). Almost all hospitalizations were associated with a lack of alert-triggered interventions ($P<.001$). System performance to predict hospitalization had a high specificity (96%) and negative predictive value (99.4%).

Conclusions: The Presage Care system has been implemented with success in assisted living facilities. It was well accepted by coordinating nurses and performed well in predicting emergency hospitalizations. However, its use by NAs was less than expected. Overall, the system performed well in terms of performance and clinical impact in this setting. Nevertheless, further work is needed to improve the moderate use rate by NAs.

Trial Registration: ClinicalTrials.gov NCT05221697; https://clinicaltrials.gov/study/NCT05221697

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55460) doi:10.2196/55460

KEYWORDS

digital technology; unplanned hospitalization; machine learning; predictive tool; assisted living facility; eHealth; pragmatic trial; artificial intelligence; AI; gerontology; older people; aging; quality of life; uncontrolled multicenter trial; France; smartphone; app; telehealth; telemonitoring; remote monitoring of patients; electronic patient-reported outcome measure; ePROM

Introduction

The aging population is increasing at an unprecedented rate, which is a major issue for many developed countries, particularly France [1,2]. The global population aged over 60 years is predicted to increase more than 4-fold by 2050, reaching 2 billion, an increase from its current figure of 605 million, and it will then account for 16% of the world's total population [3]. This demographic shift represents a significant challenge for health care systems and public policies. They must adapt to address the unique requirements of this aging population, who are affected by frailty, dependency, and comorbidities more frequently than others. This calls for the involvement of a wide range of health care professionals, caregivers, and family members [4]. Moreover, the number of emergency department (ED) visits and unplanned hospitalizations in this population has lately increased significantly [5]. In France, for example, over 3 million individuals aged 70 and older are hospitalized each year [6]. Hospitalization of the older population leads to many adverse events, such as confusion, falls, drug iatrogenic reactions, nosocomial infections, and hospitalization-related dependence, resulting in loss of autonomy in daily living activities [7,8].

Although providing medical assistance to this population is currently challenging, there are initiatives to expand the range of services available to them, such as assisted living facilities. These communities provide support to older adults throughout their lives by adopting a strategy to prevent functional decline [9]. In France, more than 120,000 older adults reside in assisted living facilities [10]. However, residents in assisted living facilities have a higher risk of hospitalization and ED visits [11]. A study conducted in Canada found that hospitalization rates in assisted living facilities (36.1%) were greater than hospitalization rates in nursing homes (10.7%) [12]. Efforts have been made for several years to improve the organization of assisted living facilities to reduce the risk of hospitalization, facilitate the involvement of professionals, and coordinate care pathways of patients [13].

Prevention of unplanned hospital admissions is a major concern for the care of older adults, as a large proportion of ED visits could be avoided [14]. Digital tools based on artificial intelligence are used in the health care industry, such as motion sensors and real-time monitoring devices. These technologies can detect early signs of frailty, such as reduced mobility, sleep disorders, and alterations in eating habits. They have the potential to improve the quality of life and predict the risk of hospitalization for older adults [15]. Several systems have been developed in Europe to assist frail older adults and people with disabilities [16,17].

Efforts are currently underway to promote the development and diffusion of digital technology for identifying and preventing frailty among older adults in assisted living facilities. We developed Presage Care (PC), a remote monitoring system for patients based on electronic patient-reported outcome measures (ePROMs) with a machine learning algorithm to predict the risk of emergency hospitalizations and a prediction window of 7-14 days [18] (Multimedia Appendix 1). PC is a Conformité Européene–marked medical device. Sensitivity and specificity are 83% and 86%, respectively [19]. The alerts provided by this system enable health care professionals to act at the right time, before health conditions deteriorate significantly. When necessary, a hospitalization can be scheduled in an appropriate department. Hospitalization prediction is a real step forward, given the serious consequences of emergency hospitalization for older people and the increased risk of morbidity and mortality [6].

Few studies in the literature have investigated the importance of community-based strategies in preventing avoidable hospitalizations of older adults [20-22].

The purpose of this pragmatic study was to evaluate the feasibility of this remote monitoring system for patients using ePROMs in an assisted living facility. Additionally, we evaluated the effectiveness and impact of the system in reducing and predicting emergency hospitalizations. The aim was also to evaluate alert-triggered health interventions (ATHIs).

Methods

Study Design and Recruitment

This multicenter, uncontrolled pragmatic trial was conducted in 7 assisted living facilities (part of the Arpavie Group), located in France, from March 1, 2022, to August 31, 2022.

To be eligible for the PC system, participants had to be aged 65 years or older, living in one of the participating assisted living facilities, and receiving the help of a health professional. The demographic data of the participants (age and gender) were collected by the nurses and health care assistants.

Intervention

The intervention is summarized in Figure 1. The nursing assistants (NAs) at the assisted living facilities were equipped with a smartphone app. They were invited to fill out an ePROM-type questionnaire after home visits on a weekly basis via the smartphone app. The list of the 23 ePROMs recorded by the NAs is reported in Multimedia Appendix 2. Regular training was provided on the use of the app. The questionnaire focused on functional and clinical autonomy (ie, activities of daily life), possible medical symptoms (fatigue, falls, and pain), changes in behavior (recognition and aggressiveness), and communication with the NA, nurse, or family members. This

XSL•FO
RenderX

questionnaire was composed of very simple and easy-to-understand questions, giving a global view of the person's condition. For each of the 23 questions, a yes/no answer was requested. Data recorded by NAs were sent in real time to a secure server to be analyzed by our machine learning algorithm, which predicted the risk level within 14 days and displayed it on a secure web-based medical device.

Importantly, when the algorithm predicted a high risk level, an alert was displayed in the form of a notification on the screen to the coordinating nurse or the coordinating NA at the assisted living facility and to the physician. This risk notification was accompanied by information about recent changes in the patients' functional status, identified from the NAs' records, to assist the coordinating nurse or NA in interacting with other health professionals and family caregivers.

In case of an alert, the coordinating nurse or NA called the professional to inquire about recent changes in the patient's health condition and to remove any doubts, and the NA could then decide to ask for a health intervention according to a health intervention model developed before the start of the study. This ATHI consisted of calling the patient's physician and informing him or her about a worsening of the patient's functional status and the potential risk of an ED visit in the next few days according to the eHealth system algorithm. The ATHI was performed with the available resources of the health system and not by the physicians or nurses who participated in the study (Figure 1). After its completion, a nurse reported the ATHI on a dedicated electronic reporting platform. These interventions were classified into 3 categories: social, paramedical, and health. The social category included actions such as social assistance, reassessment, reinforcement of home help services, and the provision of workshops in partnership with the beneficiary. The paramedical category included occupational therapy as well as exchanges between the home nursing service and private nurses. The health category included home visits, scheduled hospitalizations, and the geriatric mobile team. Nurses could also report if the ATHI was not relevant.

The alerts were only presented when they were trustworthy (ie, when there were enough data); however, to minimize a possible loss of opportunity, the person's condition and risks (geriatric related and health related) were passed on to the coordinating nurse, who could then assess the gravity of the person's situation.

Figure 1. Organization of the system.

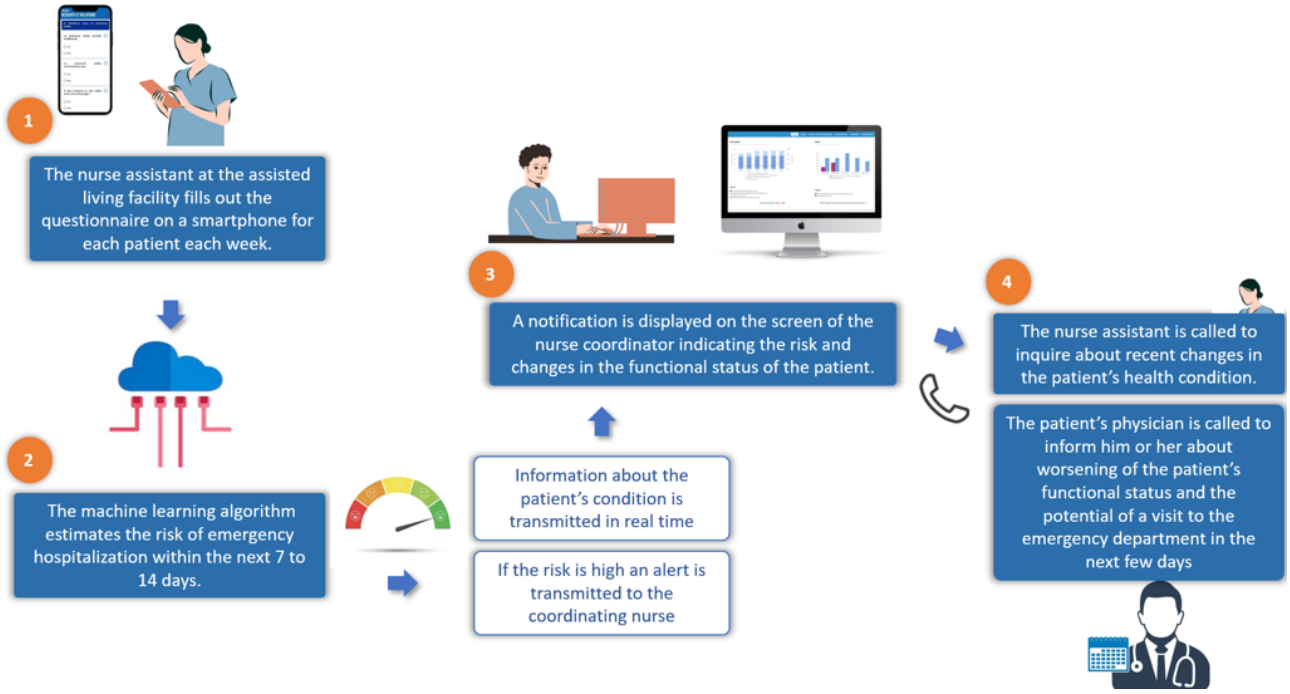

Outcomes

Feasibility and Organizational Aspects

The primary outcomes were related to use of the system and are described in Textbox 1 [22,23].

The outcome targets were defined according to a previous study that highlighted the acceptability of the device to health care professionals and its effectiveness in real-life use by older adults living at home (Textbox 2) [19].

XSL•FO
RenderX

Textbox 1. Primary outcomes.

Proportion of eligible professionals who used the system

Eligible professionals were defined as the total number of trained people at the home nursing service; a rate of use of more than 70% was considered to be acceptable.

Median completion time

A completion time under 2 minutes was considered to be acceptable.

Proportion of alerts that led to an alert-triggered health intervention

If 70% of alerts led to an intervention, it was considered to be acceptable.

Textbox 2. Outcome targets.

Regularity (proportion of professionals with regular use according to the standards)

- A result of 100% for nurse assistants meant that they all collected questionnaire data over 4 consecutive weeks, while 0% meant that none collected data over 4 consecutive weeks.
- A result of 100% for coordinating nurses meant that they used the Presage Care web application at least 3 times a week over 4 consecutive weeks.

Performance (effectiveness and efficiency of artificial intelligence in risk prevention and detection in practical, real-life scenarios)

- We monitored the eHealth system alerts' ability to predict emergency hospitalizations by assessing the true negative rate (ie, specificity) and false negative rate. The true positive rate (ie, sensitivity) was also monitored.

Delay between the alert and the intervention

- A delay of less than 4 days was considered to be acceptable [24].

Clinical Effect

The secondary outcome was the rate of emergency hospitalizations after an ATHI compared to the rate of emergency hospitalizations after no ATHI .

Statistical Analysis

Quantitative data were described using the mean and SD, or the median and IQR in cases of nonnormal distribution. Categorical data were presented as frequencies and percentages. A 2-tailed Student *t* test or Wilcoxon nonparametric test were used to compare continuous variables, whereas the χ^2 test or Fisher exact test were used for categorical variables. For each alert (by the index test), negative likelihood ratios were estimated relative to an emergency hospitalization, which was considered the target condition of the reference standard. *P* values <.05 were considered statistically significant. Statistical analyses were performed using Stata (version 16; StataCorp LLC).

Ethical Considerations

The research protocol was approved by the national French ethics committee for biomedical research, the Comité de Protection des Personnes, and the French Agency for the Safety of Health Products (2021-A02131-40–CPP 1-21-072/21.02093.000019). The protocol was presented to professional and patient representatives from the Arpavie Group (the Conseil des Résidents et des Familles). Participants, the health care professionals, and the assisted living facilities' managers were informed about the nature and purpose of the study and provided their written consent accordingly. Data storage was certified according to the ISO/IEC 27001 standard. The data were not accessible to the NAs, and the nurse coordinator and physcians could access the data only after validation with a one-time password. The patients were permitted to benefit from the service during the study; however, they received no additional compensation.

Results

The Inclusion of Centers and Patients

Among the 194 older adults in assisted living facilities who were eligible for medical device follow-up, 118 (118/194, 60.8%) were enrolled in the study and followed up at least once. The proportion of participants included in the study from each inclusion site varied by site; for example, participants from one of the sites represented 10.3% (20/194) of the total, while those from another site represented 44.8% (87/194); membership thus differed according to the facility, showing that it was based on the free choice of the older adults.

Characteristics of the Health Professionals and Participants

The mean age of the older adults was 82.55 (SD 9.19) years, with 84 of 118 (71.2%) of them being women. Out of a total of 33 eligible professional caregivers, 27 were trained (82%) and filled out questionnaires at least 1 time. Furthermore, all (20/20, 100%) coordinating nurses were trained.

Feasibility of the Implementation of the Device in an Assisted Living Facility

Completion Among Professionals and Nursing Coordinators

Among the 27 health care professionals, 7 were (26%) active during the study and used the system regularly. Regularity of use was 100% in this group. Among the 20 coordinating nurses,

XSL•FO
RenderX

9 (45%) performed regular analyses. All 9 of these coordinating nurses used the system with 100% regularity. All the facilities had some health care professionals who did not participate, and several of them had in fact left their facility during the study for another job.

Completion Time

A total of 1789 follow-ups were monitored, with a median duration for completion of 102 (IQR 86) seconds. The minimum time was 45 seconds and the maximum was 432 seconds (Table 1).

Table 1. Duration of follow-ups: about 80% of follow-ups were completed in less than 150 seconds.

Duration, s	Follow-ups (n=1789), n (%)
45-79	513 (28.7)
80-114	547 (30.6)
115-149	322 (18)
150-184	107 (6)
185-219	84 (4.7)
220-254	80 (8.5)
255-298	90 (5)
290-324	9 (0.5)
325-359	18 (1)
360-394	17 (1)
395-429	1 (0.1)
430-464	1 (0.1)

Performance of the eHealth System

During the study, a total of 92 alerts were generated by the medical device, corresponding to 5.1% (92/1789) of the follow-ups. Out of 38 emergency hospitalizations, 9 (24%) did not have any alerts in the preceding 14 days, whereas 29 (76%) were preceded by alerts within that timeframe (*P*<.001; Table 2).

The sensitivity and specificity of the alerts to predict emergency hospitalizations occurring within 14 days after the alerts were 76% and 96%, respectively. The positive and negative likelihood ratios were 17.2 and 0.36, respectively. The true positive rate was 30% and the true negative rate was 99.4%.

Table 2. Performance of the predictive system within 14 days after a regular follow-up. There were 92 alerts generated for 1789 follow-ups (5.1%).

Characteristics	Alert	No alert	*P* value
Emergency hospitalizations within 14 days after follow-up, n (%)			<.001
Yes (n=38)	29 (76.3)	9 (24.7)	
No (n=1751)	63 (3.6)	1688 (96.4)	

Health Interventions After an Alert Display

The system generated 92 alerts for 47 (47/118, 39.8%) patients. Out of these alerts, 46 (46/92, 50%) resulted in health care interventions for 14 (14/118, 11.9%) patients. The social category represented 41% (19/46) of the interventions, the paramedical category represented 24% (11/46), and the health category represented 35% (16/46). The average response time for an alert was 2 days. There was 1 alert (1/92, 1%) that was evaluated as not being an alert after a call to the nurse.

Clinical Effect of ATHIs: Health Care Intervention and Hospitalization

The total of 46 ATHIs resulted in 4 (4/46, 9%) hospitalizations, and the other 46 alerts, which did not trigger health care interventions, resulted in 25 (25/46, 54%) hospitalizations (*P*<.001) (Table 3, Figure 2). The patient's nurse or general practitioner carried out these health care interventions. Of the 47 patients who received alerts, 21 were hospitalized, 19 of whom (90%) did not receive any ATHI. When there was no ATHI, 58% of patients (19/33) were hospitalized compared to 14% of patients when there was an ATHI (2/14; *P*<.001).

XSL•FO
RenderX

Table 3. Hospitalizations of patients that occurred within 14 days of alerts generated by the eHealth system, according to whether there was an alert-generated health intervention (ATHI) among the 47 patients for which Presage Care system provided alerts. The results are expressed as number of events and number of patients, since many patients experienced several alerts and several hospitalizations during the study.

	No ATHI, n (%)	ATHI, n (%)	*P* value
Events			<.001
Alerts (n=92)	46 (50)	46 (50)	
Emergency hospitalizations (n=29)	25 (86)	4 (14)	
Patients			<.001
Patients with alerts (n=47)	33 (70)	14 (30)	
Patients with emergency hospitalization (n=21)	19 (90)	2 (10)	

Figure 2. Alert-triggered interventions and hospitalizations after alerts. We monitored patient pathways after each of the 92 alerts: of the 29 hospitalizations (marked by crosses), 25 (86%) did not receive any intervention support from nurses (labeled in orange). For 42 of 46 (91%) alerts that were followed by an intervention, there was no subsequent hospitalization (labeled in blue).

Discussion

Principal Results

The aim of this study was to evaluate the feasibility of the PC system in reducing emergency hospitalizations among older adults in assisted living facilities. The study showed that the system was accepted and used by a part of the independent-living professionals, with a use rate of 26% (7/27) for NAs and 45% (9/20) for coordinating nurses. For those who were using the system, regularity of use was excellent, with a rate of 100% for NAs and coordinating nurses. The median completion time to fill out the questionnaire was 102 seconds, which is acceptable. The results showed that 5% (89/1789) of follow-up visits had a completion time of 5 minutes or more; this much longer time was because of a professional having a discussion with a participant during the process of collecting an observation. System performance was very good, as specificity was 96% and the true negative rate was 99.4%.

The rate of interventions after an alert was moderate (46/92, 50%), and in half of the cases, an intervention was absent; we need to understand the reasons for this. However, the implementation of interventions after receiving alerts resulted in the successful prevention of hospitalization in over 91% of instances. Hospitalizations were strongly correlated with the absence of health care interventions (25/29, 86% of hospitalizations; *P*<.001). Moreover, the diagnostic capability of PC in assisted living facilities was high. When no intervention was made after an alert, 58% of the older adults were hospitalized.

These results suggest that greater completion rates by coordinating nurses would have a major clinical impact, helping to avoid as many emergency hospitalizations as possible.

The PC medical device successfully bridges the access gap associated with this type of device. The predictive performance of the device is consistent with prior published real-life studies on this device [19].

XSL•FO
RenderX

Comparison With Prior Studies

The times for filling out the questionnaire and the completion rate are consistent with a previous study on this system; however, they surpass the outcomes in other studies that highlighted issues such as long completion times and poor usability of ePROMs for caregivers and clinicians [25].

Early identification of symptoms enables "more effective monitoring and more appropriate management, leading to very concrete and sometimes significant clinical impacts, such as prolonging patients' lives and improving their quality of life" [26]. This effect was described in particular by a meta-analysis of 28 clinical trials, which found an overall improvement in the survival of older participants who participated in a gerontological assessment with long-term follow-up by specialized teams [27]. This meta-analysis, published in the *Lancet*, showed that, depending on the type of program proposed, these interventions could reduce mortality by 14% and hospitalization rates by 12%. There was also a 26% increase in home survival, a 41% improvement in cognitive function, and a 72% improvement in functional status [28].

Compared with the literature, our medical device seems to have achieved particularly good results in preventing emergency hospitalization in older adults.

Limitations of the Study

The study encountered several limitations: it was conducted during a short period and was not a randomized controlled trial with control and intervention groups. Moreover, no analysis was carried out to assess the impact of the type of professional (nurse, home aide) on the rate of completion of the questionnaire and the accuracy of the information. The relatively low participation rate among NAs was a major limitation of this study, which may be due in part to high staff turnover, and it is likely that people with short-term plans to leave the facility had little motivation to participate. In future studies, this point should be taken into account and could be ameliorated by renewed teaching for newly arrived staff or by offering incentives, rewards, or reminders. In addition, this study does not precisely describe the impact of the delay between the alert and the health intervention. Finally, the study was not controlled, and a large-scale randomized control trial is needed to measure the added value of the system in preventing emergency hospitalizations among frail older adults.

Perspectives

A randomized controlled study should be done to demonstrate how health care professionals use the new information provided by this system and examine its clinical efficacy and cost-effectiveness. The safety profile of the medical device could be assessed with a comparative analysis of adverse events or complications observed prior to and following its implementation, with the aim of identifying potential risks and side effects associated with the device. Further analysis is needed to assess the project's sustainability in a context where professional turnover is high.

Conclusion

This study demonstrates the efficacy of the PC system in minimizing unplanned hospitalizations of older adults in assisted living and assessed the predictability of the alerts generated by the system. Improvements in the regularity of use of the system by health care professionals should be feasible in future research. These encouraging results promise to be extendable to a larger population.

Acknowledgments

The authors thank the assisted living facilities (Les Grands Chênes, Chatou; La Sablonnière, Deuil-la-Barre; Le Béguinage, Lisses; Le Chêne Rouge, Chevilly-Larue; Le Bocage, Neuilly-sur-Marine; Ile de Lérins-Cannes, Cannes; Maryse-Bastie, Maison-Alfort; all located in France) and all the staff for their contributions to the study.

Data Availability

The data that support the findings of this study are available from the corresponding author, JHV, upon reasonable request. Restrictions apply to the availability of the data collected by the nursing assistants and to the algorithm that was used in this study, which has been patented.

Authors' Contributions

Conceptualization: JB, JHV, FD, CH-T
Acquisition of data: JHV, CH-T
Interpretation of the results: JB, JHV, CH-T
Formal analysis: JB; JHV, CH-T
Drafting and revising: JB, JHV, CH-T, FD
All the authors approved the final version of the manuscript.

Conflicts of Interest

JHV and CHT are employees of Presage.

Multimedia Appendix 1

XSL•FO
RenderX

Presage Care, a remote monitoring system for patients using electronic patient-reported outcome measures with a machine learning algorithm to predict the risk of emergency hospitalizations with a prediction window of 7-14 days.
[DOCX File , 15 KB - jmir_v26i1e55460_app1.docx]

Multimedia Appendix 2
List of electronic patient-reported outcome measure items recorded and their completeness.
[DOCX File , 16 KB - jmir_v26i1e55460_app2.docx]

References

1. Country case study on the integrated delivery of long-term care: Portugal. World Health Organization. 2020. URL: https://iris.who.int/handle/10665/352857 [accessed 2024-11-14]
2. Tableaux de l'économie française : Édition 2020. L'Institut national de la statistique et des études économiques. 2020 Feb 2. URL: https://www.insee.fr/fr/statistiques/4277619?sommaire=4318291 [accessed 2024-11-12]
3. Ageing and health. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2023-07-01]
4. Boisguérin B, Mauro L. Les personnes âgées aux urgences : une patientèle au profil particulier. Études et résultats 2017 Mar;1007:1-6 [FREE Full text] [doi: 10.1016/j.refiri.2017.11.005]
5. Duong HV, Herrera LN, Moore JX, Donnelly J, Jacobson KE, Carlson JN, et al. National characteristics of emergency medical services responses for older adults in the United States. Prehosp Emerg Care 2018;22(1):7-14 [FREE Full text] [doi: 10.1080/10903127.2017.1347223] [Medline: 28862480]
6. Prévenir la dépendance iatrogène liée à l'hospitalisation chez les personnes âgées. Haute Autorité de Santé. 2017 Sep. URL: https://www.has-sante.fr/upload/docs/application/pdf/2017-10/prevenir_la_dependance_iatrogene_liee_a_lhospitalisation_chez_les_personnes_agees_-_fiche_points_cles.pdf [accessed 2024-11-12]
7. Becker P, McVey L, Saltz C, Feussner J, Cohen H. Hospital-acquired complications in a randomized controlled clinical trial of a geriatric consultation team. JAMA 1987 May 01;257(17):2313-2317. [Medline: 3553627]
8. Covinsky KE, Pierluissi E, Johnston CB. Hospitalization-associated disability: "She was probably able to ambulate, but I'm not sure". JAMA 2011 Oct 26;306(16):1782-1793. [doi: 10.1001/jama.2011.1556] [Medline: 22028354]
9. Carder P, O'Keeffe J, O'Keeffe C. Compendium of Residential Care and Assisted Living Regulations and Policy: 2015 Edition. U.S. Department of Health & Human Services. URL: https://aspe.hhs.gov/sites/default/files/private/pdf/110391/15alcom.pdf [accessed 2024-11-12]
10. Miron de L'Espinay A, Roy D. Perte d'autonomie : à pratiques inchangées, 108 000 seniors de plus seraient attendus en Ehpad d'ici à 2030. Études et Résultats 2020 Dec;1172:1-5 [FREE Full text]
11. Hullick C, Conway J, Higgins I, Hewitt J, Dilworth S, Holliday E, et al. Emergency department transfers and hospital admissions from residential aged care facilities: a controlled pre-post design study. BMC Geriatr 2016 May 12;16:102 [FREE Full text] [doi: 10.1186/s12877-016-0279-1] [Medline: 27175921]
12. Maxwell CJ, Amuah JE, Hogan DB, Cepoiu-Martin M, Gruneir A, Patten SB, et al. Elevated hospitalization risk of assisted living residents with dementia in Alberta, Canada. J Am Med Dir Assoc 2015 Jul 01;16(7):568-577 [FREE Full text] [doi: 10.1016/j.jamda.2015.01.079] [Medline: 25717011]
13. Simmons SF, Coelho CS, Sandler A, Schnelle JF. A system for managing staff and quality of dementia care in assisted living facilities. J Am Geriatr Soc 2018 Aug;66(8):1632-1637. [doi: 10.1111/jgs.15463] [Medline: 30019756]
14. Uscher-Pines L, Pines J, Kellermann A, Gillen E, Mehrotra A. Emergency department visits for nonurgent conditions: systematic literature review. Am J Manag Care 2013 Jan;19(1):47-59 [FREE Full text] [Medline: 23379744]
15. Gaspar AGM, Lapão LV. A digital health service for elderly people with balance disorders and risk of falling: a design science approach. Int J Environ Res Public Health 2022 Feb 07;19(3):1855 [FREE Full text] [doi: 10.3390/ijerph19031855] [Medline: 35162877]
16. Linn N, Goetzinger C, Regnaux J, Schmitz S, Dessenne C, Fagherazzi G, et al. Digital health interventions among people living with frailty: a scoping review. J Am Med Dir Assoc 2021 Sep;22(9):1802-1812.e21 [FREE Full text] [doi: 10.1016/j.jamda.2021.04.012] [Medline: 34000266]
17. Zulfiqar A, Lorenzo-Villalba N, Zulfiqar O, Hajjam M, Courbon Q, Esteoulle L, et al. e-Health: future solution for optimized management of elderly patients. GER-e-TEC™ project. Medicines (Basel) 2020 Jul 23;7(8):41 [FREE Full text] [doi: 10.3390/medicines7080041] [Medline: 32717937]
18. Veyron J, Friocourt P, Jeanjean O, Luquel L, Bonifas N, Denis F, et al. Home care aides' observations and machine learning algorithms for the prediction of visits to emergency departments by older community-dwelling individuals receiving home care assistance: A proof of concept study. PLoS One 2019;14(8):e0220002 [FREE Full text] [doi: 10.1371/journal.pone.0220002] [Medline: 31408458]

XSL•FO
RenderX

19. Belmin J, Villani P, Gay M, Fabries S, Havreng-Théry C, Malvoisin S, et al. Real-world implementation of an eHealth system based on artificial intelligence designed to predict and reduce emergency department visits by older adults: pragmatic trial. J Med Internet Res 2022 Sep 08;24(9):e40387 [FREE Full text] [doi: 10.2196/40387] [Medline: 35921685]
20. Fournaise A, Lauridsen JT, Nissen SK, Gudex C, Bech M, Mejldal A, et al. Structured decision support to prevent hospitalisations of community-dwelling older adults in Denmark (PATINA): an open-label, stepped-wedge, cluster-randomised controlled trial. Lancet Healthy Longev 2023 Apr;4(4):e132-e142 [FREE Full text] [doi: 10.1016/S2666-7568(23)00023-5] [Medline: 37003272]
21. Huntley AL, Chalder M, Shaw ARG, Hollingworth W, Metcalfe C, Benger JR, et al. A systematic review to identify and assess the effectiveness of alternatives for people over the age of 65 who are at risk of potentially avoidable hospital admission. BMJ Open 2017 Aug 01;7(7):e016236 [FREE Full text] [doi: 10.1136/bmjopen-2017-016236] [Medline: 28765132]
22. Gayot C, Laubarie-Mouret C, Zarca K, Mimouni M, Cardinaud N, Luce S, et al. Effectiveness and cost-effectiveness of a telemedicine programme for preventing unplanned hospitalisations of older adults living in nursing homes: the GERONTACCESS cluster randomized clinical trial. BMC Geriatr 2022 Dec 22;22(1):991 [FREE Full text] [doi: 10.1186/s12877-022-03575-6] [Medline: 36550496]
23. Cartographie des impacts organisationnels pour l'évaluation des technologies de santé. Haute Autorité De Santé. 2020 Dec 31. URL: https://www.has-sante.fr/jcms/c_2902770/fr/cartographie-des-impacts-organisationnels-pour-l-evaluation-des-technologies-de-sante [accessed 2024-11-12]
24. Cash E, Cash R, Dupilet C, Pinilo M, Richard T. Qualitative survey prior to setting up a statistical system to measure waiting times for access to care. Qualitative survey prior to setting up a statistical system to measure waiting times for access to care,.: DREES; 2015. URL: https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-10/dter133.pdf [accessed 2024-11-12]
25. Duncanson E, Bennett PN, Viecelli A, Dansie K, Handke W, Tong A, et al. Feasibility and acceptability of e-PROMs data capture and feedback among patients receiving haemodialysis in the Symptom monitoring WIth Feedback Trial (SWIFT) pilot: protocol for a qualitative study in Australia. BMJ Open 2020 Nov 06;10(11):e039014 [FREE Full text] [doi: 10.1136/bmjopen-2020-039014] [Medline: 33158824]
26. Belmin J. La prévention gérontologique basée sur la fragilité : quels résultats ? Bull Acad Natl Med 2022 Aug;206(7):863-871. [doi: 10.1016/j.banm.2022.02.017]
27. Macdonald SH, Travers J, Shé ÉN, Bailey J, Romero-Ortuno R, Keyes M, et al. Primary care interventions to address physical frailty among community-dwelling adults aged 60 years or older: A meta-analysis. PLoS One 2020;15(2):e0228821 [FREE Full text] [doi: 10.1371/journal.pone.0228821] [Medline: 32032375]
28. Stuck A, Siu A, Wieland G, Adams J, Rubenstein LZ. Comprehensive geriatric assessment: a meta-analysis of controlled trials. Lancet 1993 Oct 23;342(8878):1032-1036. [doi: 10.1016/0140-6736(93)92884-v] [Medline: 8105269]

Abbreviations

ATHI: alert-triggered health intervention
ED: emergency department
ePROM: electronic patient-reported outcome measure
FNR: false negative rate
NA: nursing assistant
PC: Presage Care
TPR: true negative rate

Edited by A Mavragani; submitted 13.12.23; peer-reviewed by K Kornas, F Denis; comments to author 29.05.24; revised version received 30.08.24; accepted 11.10.24; published 10.12.24.

Please cite as:
Veyron JH, Deparis F, Al Zayat MN, Belmin J, Havreng-Théry C
Postimplementation Evaluation in Assisted Living Facilities of an eHealth Medical Device Developed to Predict and Avoid Unplanned Hospitalizations: Pragmatic Trial
J Med Internet Res 2024;26:e55460
URL: https://www.jmir.org/2024/1/e55460
doi:10.2196/55460
PMID:

XSL•FO
RenderX

©Jacques-Henri Veyron, François Deparis, Marie Noel Al Zayat, Joël Belmin, Charlotte Havreng-Théry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Contact Tracing Through Bluetooth and GPS Surveillance Data: Simulation-Driven Approach

Weicheng Qian[1], PhD; Aranock Cooke[1]; Kevin Gordon Stanley[1], PhD; Nathaniel David Osgood[1,2,3], PhD

[1]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

[2]Department of Community Health & Epidemiology, University of Saskatchewan, Saskatoon, SK, Canada

[3]Bioengineering Division, University of Saskatchewan, Saskatoon, SK, Canada

Corresponding Author:
Weicheng Qian, PhD
Department of Computer Science
University of Saskatchewan
110 Science Place
Saskatoon, SK, S7N 5C9
Canada
Phone: 1 3069661947
Email: weicheng.qian@usask.ca

Abstract

Background: Accurate and responsive epidemiological simulations of epidemic outbreaks inform decision-making to mitigate the impact of pandemics. These simulations must be grounded in quantities derived from measurements, among which the parameters associated with contacts between individuals are notoriously difficult to estimate. Digital contact tracing data, such as those provided by Bluetooth beaconing or GPS colocating, can provide more precise measures of contact than traditional methods based on direct observation or self-reporting. Both measurement modalities have shortcomings and are prone to false positives or negatives, as unmeasured environmental influences bias the data.

Objective: We aim to compare GPS colocated versus Bluetooth beacon–derived proximity contact data for their impacts on transmission models' results under community and types of diseases.

Methods: We examined the contact patterns derived from 3 data sets collected in 2016, with participants comprising students and staff from the University of Saskatchewan in Canada. Each of these 3 data sets used both Bluetooth beaconing and GPS localization on smartphones running the Ethica Data (Avicenna Research) app to collect sensor data about every 5 minutes over a month. We compared the structure of contact networks inferred from proximity contact data collected with the modalities of GPS colocating and Bluetooth beaconing. We assessed the impact of sensing modalities on the simulation results of transmission models informed by proximate contacts derived from sensing data. Specifically, we compared the incidence number, attack rate, and individual infection risks across simulation results of agent-based susceptible-exposed-infectious-removed transmission models of 4 different contagious diseases. We have demonstrated their differences with violin plots, 2-tailed *t* tests, and Kullback-Leibler divergence.

Results: Both network structure analyses show visually salient differences in proximity contact data collected between GPS colocating and Bluetooth beaconing, regardless of the underlying population. Significant differences were found for the estimated attack rate based on distance threshold, measurement modality, and simulated disease. This finding demonstrates that the sensor modality used to trace contact can have a significant impact on the expected propagation of a disease through a population. The violin plots of attack rate and Kullback-Leibler divergence of individual infection risks demonstrated discernible differences for different sensing modalities, regardless of the underlying population and diseases. The results of the *t* tests on attack rate between different sensing modalities were mostly significant ($P<.001$).

Conclusions: We show that the contact networks generated from these 2 measurement modalities are different and generate significantly different attack rates across multiple data sets and pathogens. While both modalities offer higher-resolution portraits of contact behavior than is possible with most traditional contact measures, the differential impact of measurement modality on the simulation outcome cannot be ignored and must be addressed in studies only using a single measure of contact in the future.

(J Med Internet Res 2024;26:e38170) doi:10.2196/38170

XSL•FO
RenderX

KEYWORDS

smartphone-based sensing; proximity contact data; transmission models; agent-based simulation; health informatics; mobile phone

Introduction

Sensing Modality of Colocation for Disease Transmission Models

Infectious diseases have imposed a heavy burden on the global population throughout human history [1,2]. The COVID-19 pandemic has brought the threat of contagious diseases into sharp focus. With 6.95 million deaths; 769 million confirmed cases globally as of August 8, 2023 [3]; and an estimated >US $16 trillion in lost economic activity [4] for the United States alone, the COVID-19 pandemic has been one of the defining global crises of the 21st century [5].

Epidemiological models date back over a century [6-8] but have become more useful through leveraging sophisticated algorithms [9,10] and increasing computing power [11,12]. Epidemiological models to predict, plan, and respond to pandemics and outbreaks can inform decision-making to mitigate the impact of infectious diseases [6-8]. Epidemiological models require well-grounded physical and behavioral factors to provide reasonable estimates of disease spread [13]. Linking a population's spatial behavior to infectious events that aids or inhibits the spread of disease is particularly difficult.

For airborne contagious diseases such as measles [14] or COVID-19 [5], a key enabler for disease spread is colocation (being in the same location at the same time). The effective spatial volume for COVID-19 is determined by aerosol dynamics and is often approximated at 2 m (6.56 ft) [15]. Measuring colocation can be conducted by self-reporting, as is commonly used in classic contact tracing and direct observation and counts [16], or by electronic means [17-19].

A total of 2 primary modalities for determining colocation using electronic devices exist: measurements based on estimating the distance from one person to another directly (beaconing) and measurements based on estimating the location of each person of interest within a coordinate system and calculating distances (localizing). Devices can be bespoke, such as the sociometric badge [20-22], or can leverage existing technologies such as Bluetooth(BT)–enabled phones, beacons, or dongles [23-26]. Localization techniques use systems such as GPS to place every user at a specific location at a specific time [27,28] and can be piggybacked on existing smartphones or mined from some social media platforms [29-31].

To estimate the probabilities of disease transmission, the total number of interactions, dwell times, and spatial proximity must be measured to properly baseline the parameter estimates. Techniques from companies such as Ethica Data (Avicenna Research) [32] and other companies made possible by a Google-Apple partnership [33] can be used to obtain these data for target populations under transparent and ethical data acquisition practices. However, the underlying physical processes and mathematical treatment of beaconing and localization data are substantially different and have different failure modes. Previous research had not elucidated the disparities in the estimated contact patterns for the same population between techniques. It is simple to hypothesize that colocating and beaconing will yield different contact patterns, but it is less apparent how the differences will interact with disease dynamics and impact the overall simulation outcomes.

In this study, we examine the contact patterns derived from 3 previously collected data sets using both BT beaconing and GPS localization on smartphones running the Ethica Data app. We demonstrated that while the underlying contact patterns generated from colocating and beaconing are broadly similar, they contain salient differences. For each of the 4 pathogens marked by different dynamics, we compared the results of an agent-based simulation of a communicable disease outbreak for the pathogen parameterized with beaconing- and localization-derived contact patterns. The results demonstrated that the method used to estimate contact patterns can result in significant differences between estimates of the key outbreak parameters. We showed that GPS-based contact patterns estimate significantly fewer and less-severe outbreaks than BT-derived contact patterns for the same participant and device.

Literature Review

Transmission models for communicable diseases are based on the characterization of natural history of a condition and contact networks [13]. In addition to traditional population-based nonspatial approaches, agent-based transmission models can use individual-level contact records and behaviors to identify emergent patterns using a bottom-up approach [9].

Real-world proximity tracking has applications in contact tracing, location-based risk assessment, mobility tracking, and outbreak detection [34]. Deriving real-world proximity contact mainly falls into 2 categories: calculating the delta of measured absolute positions—with, for example, GPS-assisted and Wi-Fi network–assisted locationing [35,36]—and directly measuring the relative distance with, for example, BT [19,37] or radio frequency identification [38].

Exemplars of each of these 2 approaches—GPS and BT—have been studied for digital contact tracing and transmission simulation [15,34]. Recent comparisons between GPS- and BT-inferred proximity contact collection approaches focus on privacy preservation, adoption, and compliance rates [15]. In contrast, the accuracy of simulations with GPS- and BT-derived proximity contacts is yet to be quantified across different underlying populations and pathogens [37].

Advances in digital contact tracing have also contributed to disease parameter estimation. For example, at the beginning of the COVID-19 pandemic, researchers focused on estimating the basic reproduction number R_0 from limited and highly regionally dependent infection data. As the pandemic spread, data collection and reporting standards enabled the daily reporting of incident cases, active cases, and mortality for

XSL•FO
RenderX

various geographic scales over time, allowing the estimation of the effective reproduction number R_e [39-42].

Background

BT Proximity

BT is a short-range communications protocol incorporated into most smartphones and is commonly used to pair with devices such as wireless headsets. By default, BT is configured to be in a quiescent state, not advertising its presence and only communicating with devices that have been paired. Before 2020, it was possible to lock an Android phone into a more active discovery mode, where a device would beacon approximately every 8 seconds, advertising its presence to other devices. While this functionality was intended to provide ease of initial device pairing, it could be repurposed to detect the proximity of 2 devices by registering when 1 device receives a discovery ping from another.

Studies [43,44] have investigated the use of BT to estimate the spatial proximity between devices representing people. The simplest methodology will be to create a proximity event between 2 devices if 1 device detects a discovery ping from the other or vice versa. The distance between the devices is a relevant parameter for determining a valid proximity event or contact in many applications. Researchers have typically used the Received Signal Strength Indicator (RSSI) as a proxy for distance [44-46], assuming an exponential falloff of signal strength with distance [47,48]. This approximation is confounded by reflections or transmissions off or through objects, meaning that RSSI cannot be strictly interpreted as distance, except in all but the most controlled conditions. RSSI values can plausibly be used to filter out contacts that are either far away or on the other side of a barrier, such as a wall.

The RSSI measures signal strength in decibel-milliwatts (dBm), where RSSI=0 is defined by a “Golden Receiver Power Range,” whose lower threshold level corresponds to a received power between −56 dBm and 6 dB above the actual sensitivity of the receiver, and whose upper threshold level is 20 dB above the lower threshold level to an accuracy of 6 dB. Beyond the lower and upper threshold, any positive RSSI value indicates how many dB the RSSI is above the upper limit, and any negative value indicates how many dB the RSSI is below the lower limit. Usually, a stronger signal strength (higher RSSI) indicates closer distances between 2 BT devices; however, orientation, barriers, and interference can attenuate the signal strength beyond what the distance would suggest [49]. Young [50] and the Android Beacon Library [51] contributed an RSSI to the distance function based on the Nexus 4 and Apple’s iBeacon performance, which is often used as a first approximation for similar location awareness services on modern smartphones:

Where $RSSI_0$ is the RSSI value at a 1-m (3.3 ft) distance.

GPS and Location Proximity

GPS receivers are standard on smartphones, enabling location-based services and route finding. Consumer-grade GPS receivers typically have a nominal accuracy of 10 m (32.8 ft) but can be subject to substantially larger errors due to environmental factors. Neither iOS nor Android uses pure GPS localization in their location estimation services. Both additionally use initial estimates from cell tower locations (assisted GPS) as well as fingerprinting-based localization using databases of detected Wi-Fi routers. As GPS receivers often take several seconds to obtain a position lock, even with assisted GPS, smartphone localization services tend to default to Wi-Fi–based localization initially and then switch to GPS as better location estimates become available. For simplicity of presentation, the term GPS refers to location estimation in this paper, regardless of whether it was obtained through GPS, assisted GPS, Wi-Fi fingerprinting, or some combination thereof.

Given location records, a dichotomous notion of proximity can be defined, in which 2 agents are considered proximate if they are in the same place at the same time. The precision and accuracy of the measurements and the context of the definition of proximity determine how close, in time and space, agents must be to be considered proximate or in contact. When using commodity smartphone localization hardware and services, accuracy below 5 m (16.4 ft) is rare [28], so spatial proximity has a strong lower resolution limit. Temporal resolution is substantially better—on the order of seconds—and is more likely to be limited by the measurement regime or application requirements. Elevation estimates are even less reliable than spatial estimates; therefore, commodity GPS receivers are often projected onto a 2D plane, introducing the potential for erroneous connections between people at the same location but on different floors of a building, for example.

While both GPS and BT can provide higher fidelity estimates of proximity and contact than traditional surveys or diaries, both are prone to false positives and negatives. Given 2 devices separated by a mutually proximate wall, ceiling, or floor, BT can still report contacts because the attenuation of RSSI will be such that they appear in contact but farther away. GPS is prone to false positives for detecting the proximity of communicable pathogens because the distance over which transmission can occur is smaller than the accuracy threshold for commodity devices. GPS proximity can only be interpreted as close enough that contact was possible, given the error in measurement, and not that contact actually occurred. BT can produce false negatives if the beaconing and listening cycles of the devices are misaligned, such that 1 device is beaconing while the other is asleep. GPS can lose signal or accuracy when indoors, causing false negative contacts by either having no location reported for an agent or exhibiting position inaccuracies that render inaccurate colocation calculations. While the underlying true contact dynamics for the same devices are identical, the differing failure modes of GPS and BT mean that data drawn from those data collection modalities may generate different contact networks, thereby suggesting different contact dynamics and ultimately different outbreak dynamics.

Agent-Based Susceptible-Exposed-Infectious-Removed Models

The susceptible-exposed-infectious-removed (SEIR) disease state model is a classic model used to characterize pathogen

XSL•FO
RenderX

transmission and the natural history of infection across a range of communicable diseases. Disease state transitions are unidirectional in the order of susceptible, exposed, infectious, and removed. The initial state of the model specifies the amount of population in each disease state, and the rate of transition between disease states is subject to both disease-characteristic parameters (such as latent period and infectious period) and the contact network (such as preferential mixing and average contact rate). It is common for a specific disease, given surveillance data, to obtain more detailed models. For example, there are models of COVID-19 splitting the SEIR states into more states and rerouting transitions in states [52,53]. As our goal was to probe the impact of contact measurement modality, in accordance with the Occam's razor [54], which recommends a parsimonious model with the fewest assumptions that are necessary, we chose the simplest SEIR model.

Agent-based models (ABMs) incorporate individual interactions and track the state and state transitions through which each individual progresses. Unlike a stock and flow model, which uses differential equations to model the flow of individuals from one state to another in aggregate, an ABM knows the state of every agent individually at any time step of the simulation, and aggregate statistics, for example, on infections, are queried and computed during postprocessing. An agent-based SEIR model captures both individual disease state transitions based on disease-specific parameters such as the latent period, the infectious period, R_0, as well as some abstraction of the contact behavior of the population. As the simulation of an infectious disease can capture emerging patterns in a bottom-up manner [9] and more faithfully reflect dynamics due to the proximity contact network than compartmental models, ABMs provide higher fidelity at the cost of computation when compared to stock and flow models. As an ABM can directly use a contact pattern as part of the simulation, it is the logical choice for examining the sensitivity of simulations to the contact detection methodology.

Methods

Data Set Description

For this study, we used 3 previously collected data sets, all of which were collected from the city of Saskatoon, a city in the midwestern Canadian province of Saskatchewan. In all these data sets, additional sensor modalities (eg, accelerometer, gyroscope, and Wi-Fi traces) were also collected, but only the BT traces, GPS traces, and battery data were used in this study. Battery data were used to identify gaps in data collection. If the phone is on and Ethica is running, then battery data were recorded, providing a more reliable way to assess the continuity of data collection than is possible with GPS, where signals can be obscured by the built environment but where the phone is still actively recording. The Saskatchewan Human Ethology Datasets (SHEDs) are a collection of pilot projects and technical trials taking place during the iEpi project—the academic precursor for the Ethica Data system—and associated postprocessing and methodological outcomes [55,56]. The SHEDs were exclusively collected from populations at the University of Saskatchewan in Saskatoon. The SHED7 data set was collected between July 11 and August 8, 2016, and included 61 students. The SHED8 data set was collected between September 25 and October 25, 2016, and included 74 students. The SHED9 data set was collected between October 28 and December 9, 2016, and included 88 students. These participants were part of a social science student study pool that included both undergraduate and graduate students and was weighted toward undergraduates.

Ethical Considerations

Data collection and analysis were conducted under written approval (BEH-14-203) from the University of Saskatchewan Human Behavioral Ethics Review Board. All data were collected with the informed consent of the participants and under the oversight of the University of Saskatchewan Human Behavioral Ethics Review Board.

No experimental manipulations were conducted during data collection. The studies did not undertake stratified sampling according to ethnicity, grade, or gender. The study did not proscribe participation by those connected with the department or research laboratories involved, and the study team informed colleagues in laboratories and the Department of Computer Science first. The awareness of potential study involvement can be assumed to have spread across social networks. For each study, participants joined using their own phones to install the Ethica app and consented to have sensor data collected over the study period. Although for all these 3 studies, both Android and iPhone users were welcome, because BT beaconing did not work reliably on an iPhone due to security settings, iPhone users were removed from the analysis, and all participants reported here are Android users. Each participant receives approximately a CAD $50 (US $38.6) honorarium at the end of a study, and the exact amount varies by study length.

All 3 SHED studies stored sensor data anonymously, and the sensor data associated with a participant were identified by a device-identity number. Despite measures such as encryptions being used to protect sensor data, high-velocity GPS data can allow a skilled practitioner to determine salient information about participants, such as place of work, residence, and daily habits. We circumvented this issue by committing to our own research ethics board and the participants that researchers who access the data must commit to writing requests subject to the review and approval of our ethics boards.

Sensor Data Processing

Overview

To evaluate the performance of each sensor in real-world scenarios, we needed to account for the impact of participant compliance. We defined the active period of a study with the start day as the first day when we have ≥80% of the participants' battery reading and the end day as the first day with all following days having <80% of the participants' battery reading. We retained participants who had at least 50% of the daily battery data. The descriptive statistics are presented in Table 1.

XSL•FO
RenderX

Table 1. Sensor data table for the 3 data sets (Saskatchewan Human Ethology Dataset [SHED]7, SHED8, and SHED9). These data sets are collected in 2016 with participants comprises students and staff of University of Saskatchewan. Each of these data sets is collected with the Ethica Data app with Bluetooth and GPS sensors with duty cycle about every 5 minutes over a month.

	SHED7	SHED8	SHED9
Participants, N	61	74	88
Retained participants, n	61	74	78
Total days in studies, n	35	31	41
Active days in studies, n	28	30	38
Bluetooth-inferred contacts (distance threshold: 8 m), n	37,804	34,400	20,597
GPS-inferred contacts (distance threshold: 10 m; accuracy: 10 m), n	4338	6784	5064

Ethica's multisensor sensing requests that the Android operating system perform sensor scanning and reading periodically. We call our requested period length, that is, each of the repeated 5-minute time windows, a duty cycle. For the location and BT contact data used in this study, Ethica records for 1 minute, starting every 5 minutes.

The BT discovery record from the Android application programming interface includes the discovered BT device's MAC address and RSSI. After linking such discovery records to participant IDs in the smartphone BT MAC address table collected after consent and before the experiment started, we created a table of BT discovery records for eligible participants. Those RSSI values were filtered to include records associated with an RSSI stronger than the RSSI values associated with the desired distance thresholds and then aggregated maximum RSSI values for unique tuples of discovered participant and duty cycle (data collection epoch), resulting in the final BT contact record table. Although BT discovery records are directional, our use of unique tuples will consider a pair of participants potentially in contact if at least one's BT device discovers that of the other.

Starting with raw GPS readings for each participant, we first discarded GPS readings with an accuracy radius larger than 10 m (32.8 ft) as being too inaccurate to allocate even approximate colocation estimates. For each participant, we used the median of their GPS readings within a duty cycle as the estimated geolocation of that participant. We then mapped the estimated GPS coordinates of latitude and longitude onto the Universal Transverse Mercator coordinates as the northing and easting with units of meters. For the sake of estimating interparticipant proximity, we used the Euclidean distance between the estimated geolocation for all pairs of participants within the same duty cycle as the estimated distance between pairs of participants. For each duty cycle, participants who lacked GPS readings within that duty cycle were considered as not being in contact with any of the other participants for the duration of that cycle.

Agent-Based SEIR Model

An agent-based SEIR simulation model was used to characterize pathogen transmission and describe the natural history of infection. The model assumed the following:

- There is no reinfection during the simulation period.
- The population is closed, and no births, deaths, and migrations occur during the simulation time horizon.
- The latent periods for diseases under consideration are similar to the incubation periods.
- During the infectious period, an infectious patient will have a constant hazard rate of transmission to every one of their currently contacted persons, normalizing passive shedding from active spread (eg, sneezing) over a contact period.
- There are no behavior changes in participants during the simulation period, conditional on the contact patterns measured. For small outbreaks, this is reasonable, but the COVID-19 pandemic has demonstrated the importance and magnitude of changes that can occur in hygienic personal protective behavior (eg, mask use) over the course of a pandemic.

We made use of a 4-fold duplication and concatenation of both GPS- and BT-inferred proximity contact data, such as successively replaying a movie, to allow the outbreak to run its course without running out of contact data.

All participants who connected to at least 1 other participant after filtering were included in the simulation (Table 2). Each simulation starts with 1 initially exposed participant. We conducted multiple simulations with different random seeds to account for stochastics. Each simulation began with a single participant with an infection of the corresponding disease. All active participants were the initially exposed participants in turn, for 50 realizations each.

During the initialization of each simulation realization, incubation period and infectious period were drawn uniformly from the minimum-maximum range of corresponding parameters, as presented in Table 3.

All diseases listed in Table 3 are investigated from a historical perspective, where estimates of the corresponding parameters are available.

XSL•FO
RenderX

Table 2. Number of participants with at least 1 contact within the 3 data sets (Saskatchewan Human Ethology Dataset [SHED]7, SHED8, AND SHED9). Proximity contacts are inferred from sensing data and vary by sensing modality and distance threshold.

	SHED7	SHED8	SHED9
BT8[a], n	58	71	76
BT20, n	58	71	76
GPS8, n	49	63	66
GPS20, n	58	71	74

[a]The sensing modality of Bluetooth and GPS are combined with distance thresholds of 8 m (26.2 ft) and 20 m (65.6 ft) in rows. For example, BT8 stands for proximity contacts inferred from Bluetooth-sensed distance within the threshold of 8 m (26.2 ft).

Table 3. Disease parameter.

	R_0	Incubation period, minimum to maximum	Infectious period minimum to maximum
COVID-19 (nonvariant)	2.2[a] [57,58]	5.6-7.7 [57]	3-7 [57]
Influenza	3[b] [59]	1-4 [60]	3-5 [60]
Norovirus	1.75 [61]	0.5-2 [62]	2-3 [62]
Measles	15[b] [63]	10-12 [64]	8-11 [64]

[a]Derived as midpoint of reported range.

[b]Derived range from different reports.

Simulation Configuration

For each SHED study, after preprocessing, we obtained GPS- and BT-inferred proximity contact data with distance-equivalent RSSI thresholds of −80 dBm (corresponding to approximately 8 m or 26.2 ft) and −90 dBm (approximately 20 m or 65.6 ft). A group of simulations for each of the 4 diseases—namely, influenza, COVID-19, measles, and norovirus—was run. Within each group of simulations sharing the same derived proximity contact data and disease parameters, we iterated each of the active participants as the initially exposed patient with 50 realizations, where each realization has a different predetermined random seed, resulting in 170,400 realizations across all data sets and conditions.

In our agent-based SEIR model, at any given time during the simulation, each agent resides in one of the 4 disease states (susceptible, exposed, infectious, or removed). At the start of the simulation, all agents, except the initially exposed agent, are susceptible. The transition from susceptible to exposed has a probability p when exposed to proximity to an infectious agent. Such occurrences of exposure are characterized by a Poisson process with a mean interarrival time of 5 minutes. The value assumed for p is derived from the disease-specific R_0 and the average empirically observed frequency of population contacts. The timing, the duration, and the pair of agents involved in each proximity contact are given by the proximity contact data fed to the simulation. The transitions of exposed-to-infectious and infectious-to-removed are timeouts with timers set as the corresponding latent period and infectious period as initialized for each individual.

Simulations were run on 2 servers, each with an Intel Xeon CPU E5-2690 v2 and 503 GB memory. Models were created in AnyLogic (version 8.1.0; The AnyLogic Company) and exported to a stand-alone Java application with OpenJDK (version 1.8.0_252; The OpenJDK Community) as the runtime environment. Analysis was conducted in R software (version 4.0.2; R Foundation for Statistical Computing) with major packages, including *tidyverse* (version 1.3.0), *ggprah* (version 2.0.5), and *igraph* (version 1.2.6), and in Python (version 3.8.0; Python Software Foundation) with major packages, including *pandas* (version 1.2.0), *numpy* (version 1.20.2), and *scipy* (version 1.6.1).

Evaluate Impacts on Transmission Models

Overview

We used the attack ratio as the metric to evaluate the impact of proximate contact data on transmission models. The attack ratio is the proportion of the total population that gets infected throughout the simulation. Although the ABM-SEIR can produce many estimates for different disease parameters given proximate contact data, the attack rate and individual risk of infection was chosen for simplicity, accessibility, and to serve as single summary statistics [65,66].

Attack Rates

The core research question of this study was whether and to what extent the differences in GPS- and BT-based proximity detection would alter the contact network and therefore the implied attack rate. We considered the attack rate θ defined as in the following equation:

Where $I(t)$ is the number of infectious persons at time t, T is the end time of the simulation instance, and N is the population size. The attack rate θ denotes the proportion of the population that is infected throughout the simulation instance. As the response variable (denoted as Θ) to the controlled variables of the disease or pathogen M, the initial infectious individual ν ∈

XSL•FO
RenderX

$V = \{v_1,v_2,...,v_n\}$, $n = \| V \|$ and collected proximity contact data $D(\omega,\varepsilon,V)$. For the proximity contact data $D(\omega,\varepsilon,V)$, $\omega \in \{BT, GPS\}$ is the sensor type, $\varepsilon \in \{8, 20\}$ is the distance threshold of proximate contacts, and V is the underlying population. Therefore, with the ABM-SEIR model as $P(\cdot)$ for a specific disease M and underlying population V, we can sample $\Theta \sim P(\Theta = \theta \mid \omega,\varepsilon,M,V)$ with simulation realizations. While the initial infectious individual ν has been known to impact the attack rate Θ, investigation of that impact lies outside the scope of this paper.

Welch *t* Test

Assuming disease M and an underlying population V, the choice of an initial infectious individual ν is independent of the data collection configuration (sensor type ω and proximate distance threshold ε). We were interested in the marginal probability, defined as in the following equation:

Limited by our knowledge of $P(\nu|M,V)$, we assumed the initial infectious individual ν is chosen with uniform probability from the underlying population V, that is, $P(\nu|M,V) = 1 / \| V \|$. Consider θ as the sample mean from a sample, $X_i \sim P(\Theta|\nu = v_i,\omega,\varepsilon,M,V)$, $i = 1,..., \| V \|$, and we sampled by repeating $\| V \|$ simulations iterating every individual of the population V as the initial infectious individual. According to the central limit theorem, samples of $\theta \sim P(\Theta|\omega,\varepsilon,M,V)$ tend to be normally distributed to suffice the assumption of the Welch t test (2-tailed).

Pairwise *t* Test

Without assuming that the initial infectious individual ν is homogeneous among the underlying population V, that is, $P(\nu|M,V)=1/\| V \|$, we could construct a pairwise t test by pairing the samples of attack rate having the same initial infectious individual μ, given sensor type ω and distance threshold ε, for each pair of disease M and underlying population V. In this case, we assumed the pairwise differences in the attack rate, such as for $\Theta_i^{BT8\text{-}GPS20} = \Theta^{BT8} - \Theta^{GPS20}$, are normally distributed, where $\Theta^{BT8} \sim P(\Theta|\nu = v_i,\omega = BT,\varepsilon= 8,M,V)$ and $\Theta^{GPS20} \sim P(\Theta|\nu = v_i,\omega = GPS,\varepsilon= 20,M,V)$.

Kullback-Leibler Divergence of Individual Infection Risks

Given the sensor type, proximate distance threshold, disease, and underlying population, we estimated the individual infection risk based on the Laplacian-smoothed rate of being infected across realizations, denoted by $\rho_{v\in V}(\omega,\varepsilon,M,V)$. The likelihood of being the most likely infected individual for an individual $v\in V$ follows $P(v|\omega,\varepsilon,M,V)$, which can be estimated by normalizing vector $\rho=\{\rho_v|v\in V\}$. The Kullback-Leibler (KL) divergence was used to summarize the differences between pairs of sensor type and proximate distance threshold (ω,ε) within blocks by disease and the underlying population. For disease M and the underlying population V, we have

between sensing configurations (w_1,e_1) and (w_2,e_2), where

Results

While the agent-based simulation uses dynamic contacts, some insight can be gained by examining the aggregate contact network of participants in each study. Figure 1 shows the aggregate contact networks for SHED7, SHED8, and SHED9 using BT and GPS at 8- and 20-m thresholds. If a connection ever occurred between 2 nodes given the protocol, a corresponding edge is drawn in the network, with the color of the edge proportional to the total contact duration over the course of the experiment between those nodes. Reflecting the Pareto-like distribution of contact duration, colors move from blue (weakly connected) to red (strongly connected) on a logarithmic scale, consistent with other human network observations [67,68]. As expected, most nodes appear to have weak connections compared to the highly connected dyads and triads in the network. The BT networks are denser and more highly connected than their GPS counterparts, implying a greater potential for disease spread. There is a greater preponderance of weak edges in the BT data sets than in their corresponding GPS counterparts. There is a modest increase in the number of edges between the 8- and 20-m thresholds for each data set.

XSL•FO
RenderX

Figure 1. Stress layout of aggregated weighted contact network by underlying population and data source, with edges colored in log scale by weights. BT: Bluetooth; SHED: Saskatchewan Human Ethology Dataset.

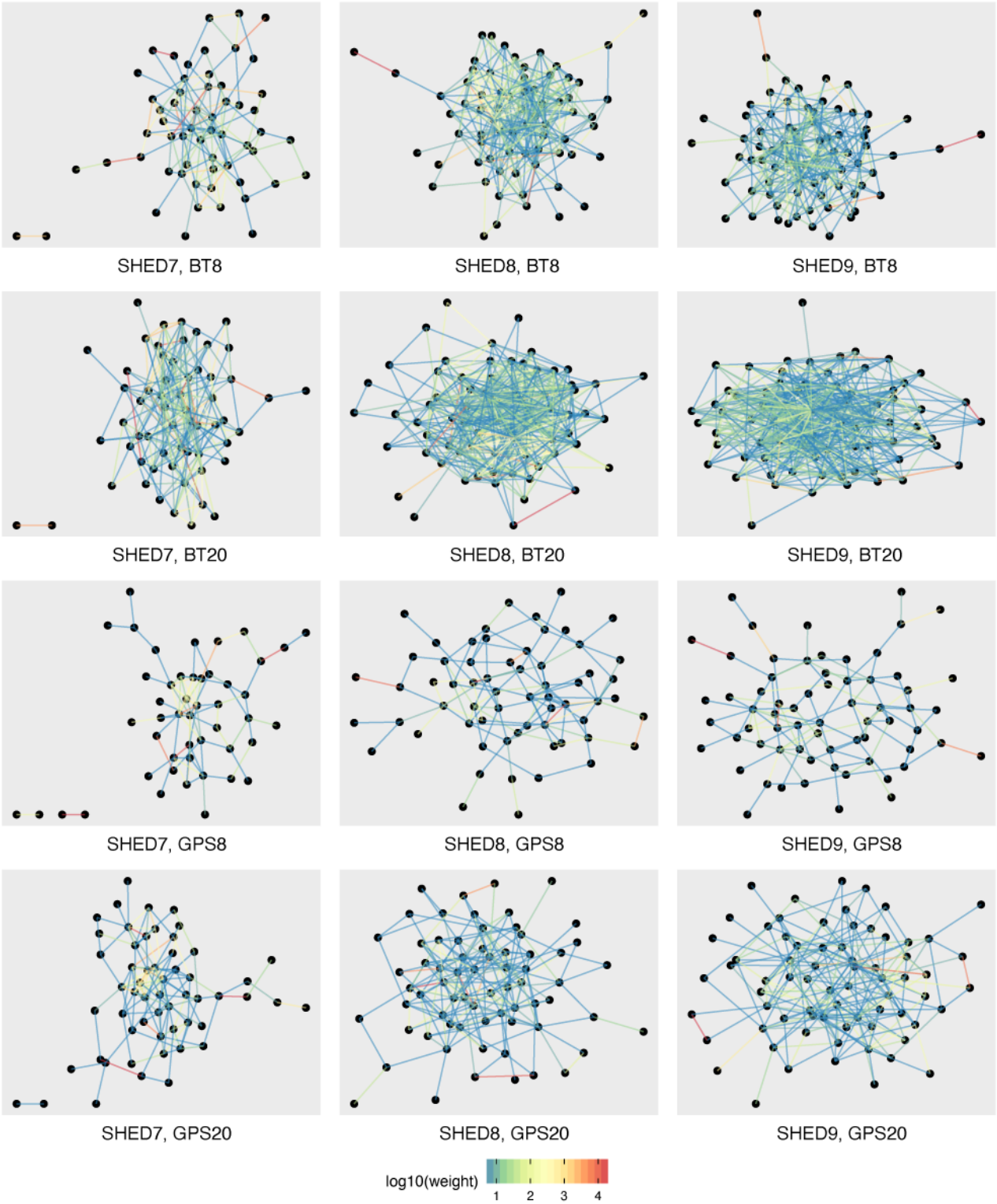

Contact frequency (the rate at which contacts occur) and intercontact time (the time between contacts) are common aggregate distributions used to characterize contact data sets. Similar to many other data sets, both the BT and GPS demonstrate power law decay for the probability of contact duration and intercontact time (Figure 2). GPS-based contact detection tends to infer more and shorter-duration contacts but exhibits truncated tails. In SHED7 and SHED9, the tail truncation leads to fewer long-duration contacts (>600 min) than BT. The intercontact times are similar for all data sets, but the BT distributions are skewed more heavily toward longer intercontact times than in the case of GPS. In contrast, for SHED8 and SHED9, BT tracking detects notably fewer moderately long contacts (those in the range of 50 min to 600 min). This may be due to localization noise–induced false

XSL•FO
RenderX

positives in the GPS data set skewing the apparent contact durations higher.

Figure 2. Empirical complementary cumulative distribution function of contact duration and intercontact time with different sources and distance thresholds. BT: Bluetooth; SHED: Saskatchewan Human Ethology Dataset.

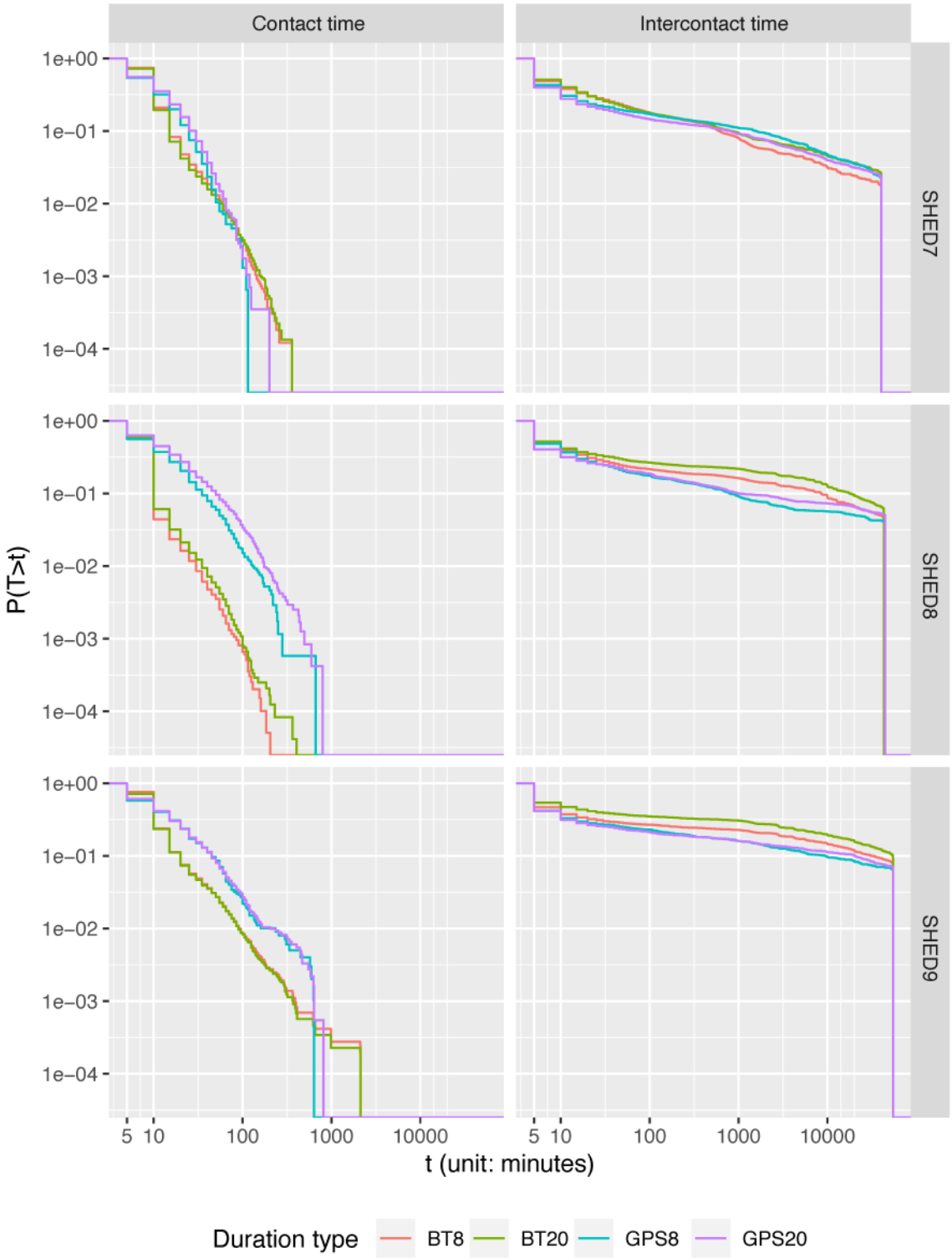

After filtering the connections for the appropriate distance threshold (8 m and 20 m or approximately 26.2 ft and 65.6 ft), the agent-based simulation was run according to the protocol described in simulation configuration. Many runs do not produce an outbreak, with the initially exogenously infected individual being the only member of the network infected. This results in a zero-heavy bimodal distribution of cumulative infection counts per realization, with a Poisson spike at 0 cumulative endogenous infections (1 exogenous infection) and a second distribution describing the probability of an outbreak of a given size conditional on outbreak occurrence (ie, the probability of at least one endogenous infection). A stacked bar plot showing the ratio of runs in which further incidences beyond the initial infectious individual did or did not occur is shown in Figure 3.

XSL•FO
RenderX

The figure clearly shows a higher likelihood of an outbreak occurring with the BT data, as expected from the aggregate network diagrams and aggregate contact duration and frequency plots. The consistent difference in the probability of outbreak occurrence between the 2 conditions is our first substantial indication that the two means of measuring contact are not equivalent. To determine the impact of each dynamic contact pattern on the outbreaks themselves, the trials in which no endogenous infection occurred were removed, and statistical analysis was conducted on the distribution of outbreak severity conditional on outbreak occurrence.

Figure 3. Number of realizations with or without further infections beyond initial infections. BT: Bluetooth; SHED: Saskatchewan Human Ethology Dataset.

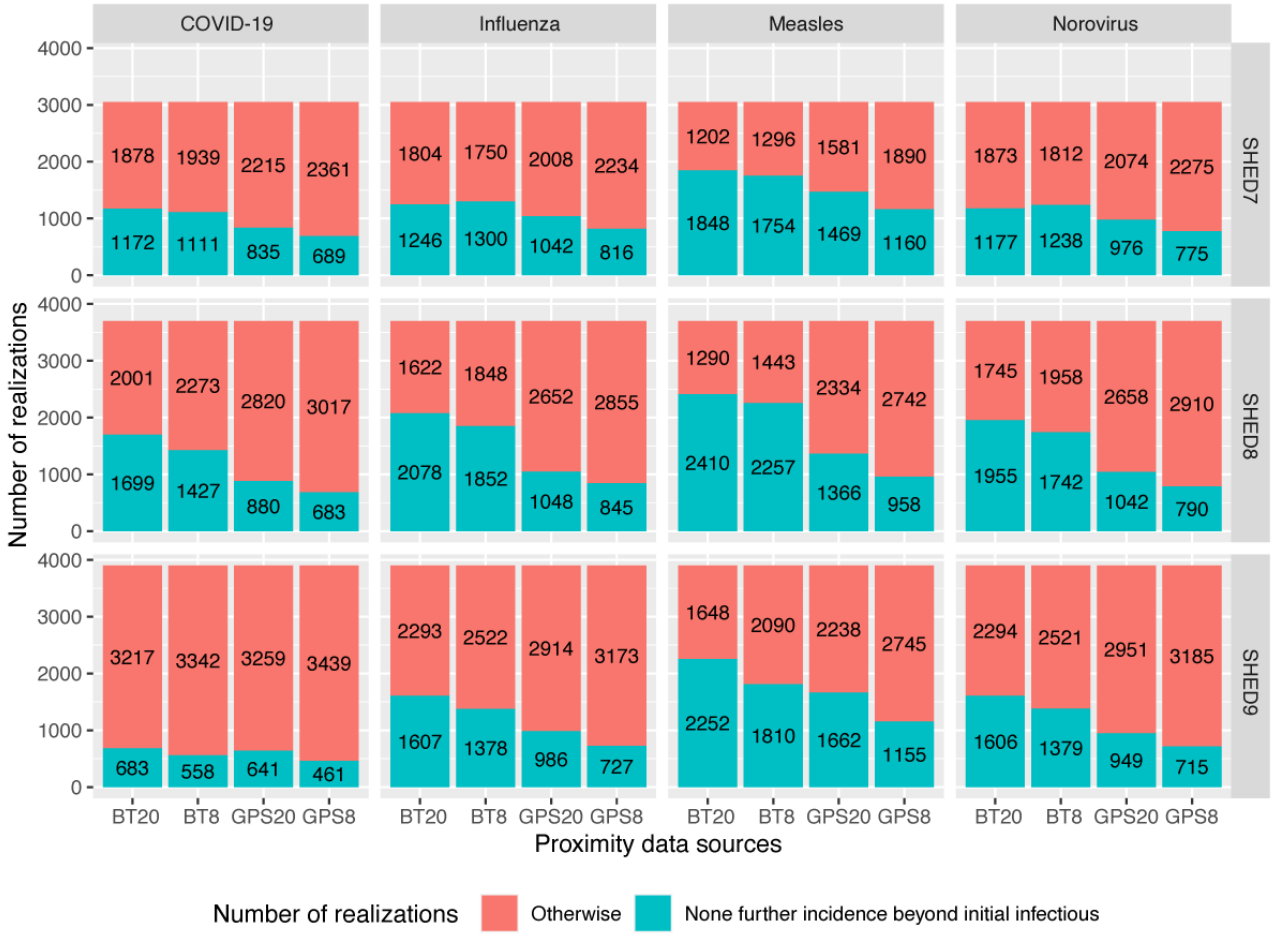

The Bonferroni-corrected Shapiro-Wilk test was passed for each of the 50 samples of attack rate, denoted by θ, for every pair of disease M and the underlying population V, except for COVID-19, with contact records collected via GPS using a distance threshold of 20 over SHED8. The results of the Bonferroni-corrected Welch t test are presented in Table 4.

XSL•FO
RenderX

Table 4. Bonferroni-corrected Welch *t* test for incidence number.

	P value					
	BT8-BT20[a]	BT8-GPS8	BT8-GPS20	BT20-GPS8	BT20-GPS20	GPS8-GPS20
Norovirus						
SHED7	<.001	.05	<.001	<.001	<.001	<.001
SHED8	<.001	<.001	<.001	<.001	<.001	<.001
SHED9	<.001	<.001	<.001	<.001	.001	<.001
Influenza						
SHED7	<.001	.12	<.001	<.001	<.001	<.001
SHED8	<.001	<.001	<.001	<.001	<.001	.001
SHED9	<.001	<.001	<.001	<.001	<.001	<.001
COVID-19						
SHED7	<.001	.85	<.001	<.001	<.001	<.001
SHED8	.04	<.001	<.001	<.001	<.001	.04
SHED9	.001	<.001	>.99	<.001	<.001	<.001
Measles						
SHED7	<.001	.003	<.001	<.001	>.99	<.001
SHED8	<.001	<.001	<.001	<.001	<.001	<.001
SHED9	<.001	<.001	<.001	<.001	<.001	<.001

[a]BT: Bluetooth.

Our predetermined α level was .05.

The results of Bonferroni-corrected pairwise *t* tests [69] between observed attack rates (having filtered out scenarios with 0 endogenous infections) across all simulation runs for a condition are presented in Table 5.

Table 5. Bonferroni-corrected pairwise *t* test for attack rate.

	P value					
	BT8-BT20[a]	BT8-GPS8	BT8-GPS20	BT20-GPS8	BT20-GPS20	GPS8-GPS20
Norovirus						
SHED7	<.001	.25	<.001	<.001	<.001	<.001
SHED8	.44	<.001	<.001	<.001	<.001	>.99
SHED9	<.001	.01	.005	<.001	.008	<.001
Influenza						
SHED7	<.001	>.99	<.001	<.001	<.001	<.001
SHED8	<.001	<.001	<.001	<.001	<.001	>.99
SHED9	<.001	.004	.02	<.001	<.001	<.001
COVID-19						
SHED7	<.001	.73	<.001	<.001	<.001	<.001
SHED8	.07	<.001	<.001	<.001	<.001	>.99
SHED9	<.001	<.001	>.99	<.001	<.001	<.001
Measles						
SHED7	<.001	>.99	<.001	<.001	.49	<.001
SHED8	<.001	<.001	<.001	<.001	<.001	.002
SHED9	<.001	<.001	<.001	<.001	<.001	<.001

[a]BT: Bluetooth.

XSL•FO
RenderX

These results confirm our hypothesis that BT- and GPS-based contact histories induce significantly different estimates of total disease burden across multiple simulated realizations. The primary comparisons are the BT8-GPS8 and the BT20-GPS20, with the others included for completeness. For SHED7 BT8-GPS8, the results are not significant. For all other diseases and data sets, the results are statistically significantly different. In the case of BT20-GPS20, all results are significantly different, with the exception of the SHED7 measles. While we suspected that the infectiousness of the disease would impact simulated outcomes, the results seem to be dominated by differences in the data set and contact measurement modality. Looking at the impact of resolution, some combinations of data set and disease are not significantly different, but for the most part, increasing the threshold increases the number of contacts, driving differences in simulated outcomes. The exception to this general rule seems to be SHED8 GPS8-GPS20, where increasing the threshold did not significantly alter the outcomes for most diseases and only marginally for measles.

Figure 4 shows the violin plots for the attack rates over each realization across all simulated conditions and provides insight into the statistical results from Table 5. SHED7 consistently has lower attack rates for all diseases, with a smaller variance and mean than other data sources. The limited attack rate likely drives the similarity between the BT and the GPS. The denser SHED8 and SHED9 networks have substantially larger variance, leading to significant differences between the measurement modality conditions. The highly contagious measles virus, in particular, exhibits marked differences within the SHED8 and SHED9 data sets. In general, BT contact patterns have longer tails, indicating a greater possibility of larger outbreaks throughout the population. In cases where a substantial probability mass is contained in the tail, the median is also drawn higher, as in SHED8 with BT20 for measles.

Figure 4. Distribution of the attack rate (filtered out 0) for data collections and diseases. BT: Bluetooth; SHED: Saskatchewan Human Ethology Dataset.

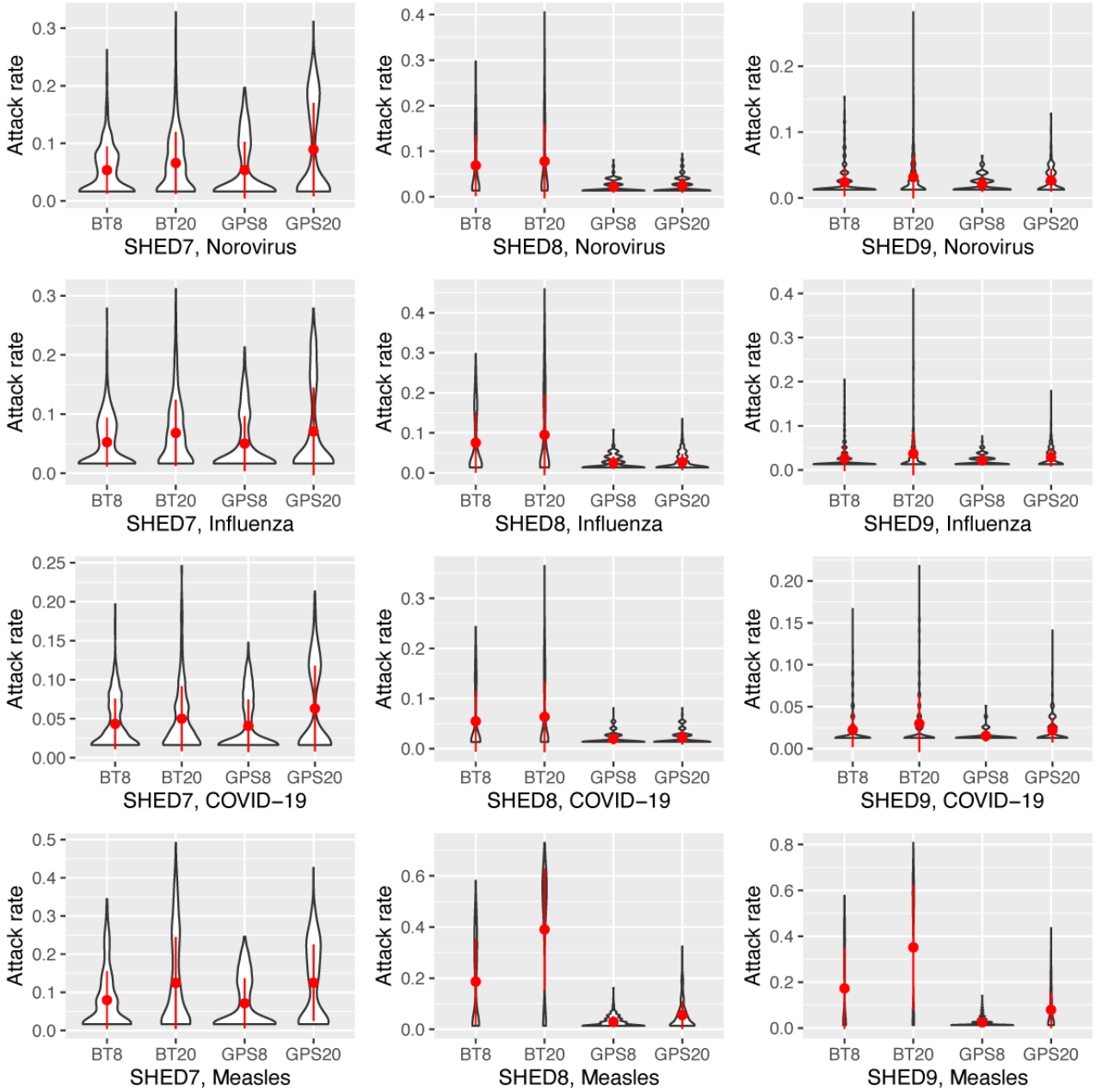

XSL•FO
RenderX

Figure 5 shows the KL divergence on individual infection risks within blocks of disease and the underlying population. The individual infection risks are reflected by the likelihood of being the most likely infected individual between different sensing configurations, where a sensing configuration is a pair of selected sensor types and proximate distance thresholds. The distance threshold of proximate contact does not appear to impact GPS-collocated inferred proximity contacts in terms of individual infection risks, regardless of the underlying population. This invariance to distance thresholds suggests that the primary bottleneck lies in the GPS-colocation method's inability to identify exact proximity contacts among a group of collocated individuals. Meanwhile, the BT-beaconing method may capture proximity contacts at certain distance thresholds (such as for SHED7 and SHED8), which can be important when considering droplet-based pathogen transmission. There seems to be a lower magnitude of KL divergence for BT8-BT20 and BT20-BT8. The KL divergence among pairs of different sensor types is similar regardless of the distance thresholds of proximate contact, suggesting that BT beaconing and GPS colocating collect different proximity contacts regardless of the resolution of the distance thresholds of proximate contacts. The magnitude of asymmetric $|D_{KL}(p \parallel q) - D_{KL}(q \parallel p)|$ shown in red lines is lower than either $D_{KL}(p \parallel q)$ or $D_{KL}(q \parallel p)$, indicating that the asymmetry of the KL divergence is not impairing our aforementioned analyses.

Figure 5. Kullback-Leibler divergence of individual infection risks.

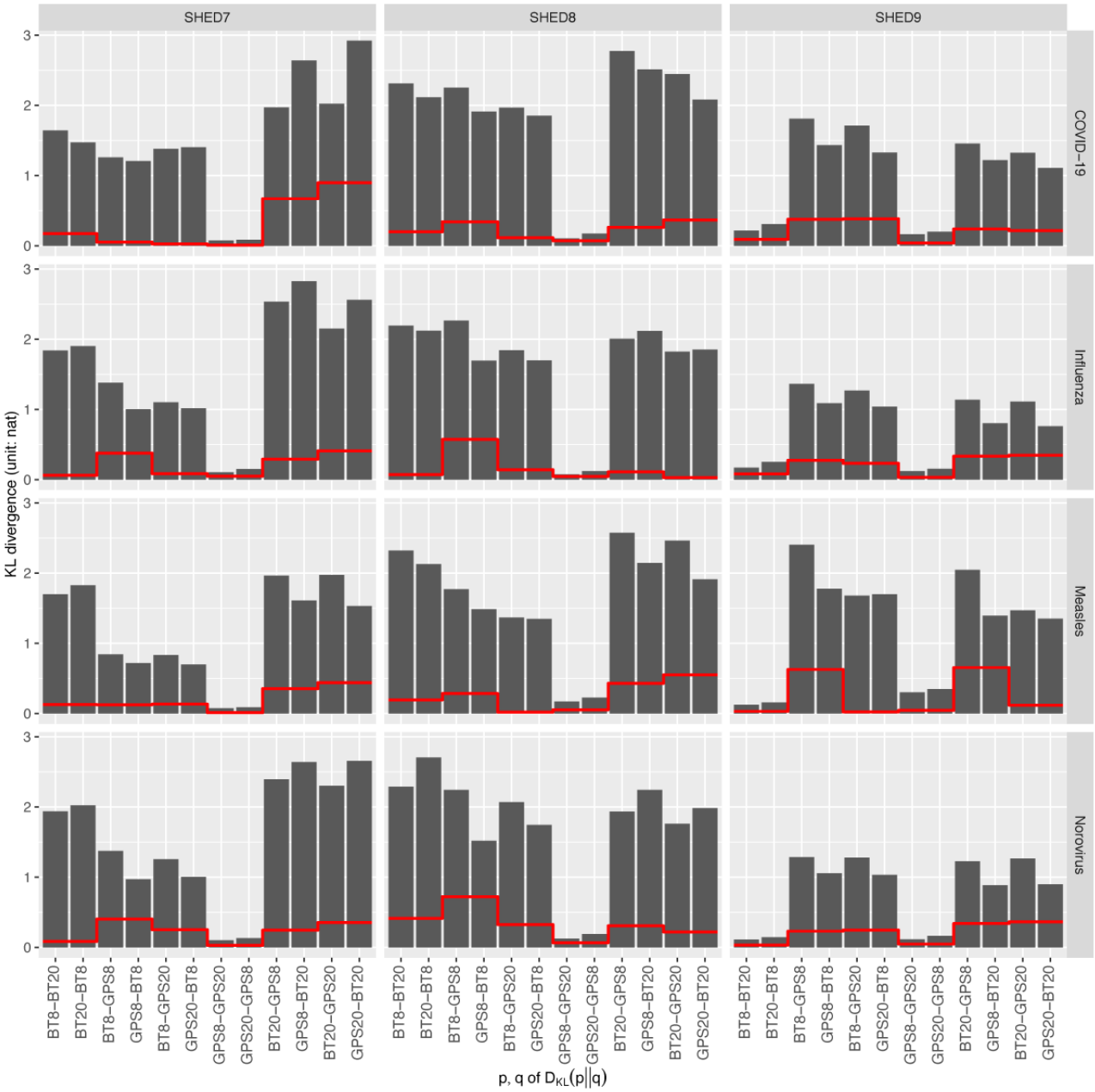

XSL•FO
RenderX

Discussion

Disparate Results From GPS- and BT-Based Contact Tracking

Our results clearly indicate that GPS- and BT-based contact tracking yield disparate results for the same cohort under measurement. The ground truth contact network, while unknown, was the same for each data set—it was the same set of participants carrying a single phone measuring both quantities. Both BT- and GPS-derived contact measurements are estimates of the underlying contact pattern, admitting false positives (eg, BT contacts through a wall) and negatives (eg, a missed GPS contact because it occurred in an area of poor satellite reception). GPS-based contact tracking identifies fewer shorter contacts, leading to a significant decrease in expected outbreak intensity and the number of outbreaks, potentially because both participants need to have a sufficiently good location fix to estimate colocation. The denser contacts reported by BT-based contact tracking led to a higher probability of an outbreak and larger outbreaks, resulting in significantly different attack rates for most data sets and diseases. While there were conditions under which no significant differences were observed across the data collection modalities (particularly for SHED7 BT8-GPS8), differences were often significant enough to encourage caution in the uptake and interpretation of these sensed contact networks. GPS8 tends to underestimate the attack rate relative to the others (BT8, BT20, and GPS20), indicating the general inability of GPS colocating to capture proximate contacts within a short distance. Sensing configurations tend to estimate similar attack rates for infectious diseases without a comparatively high R_0 in a more distant underlying population, except for SHED9-measles. Our study cannot conclusively determine if the higher outbreak frequency and size in BT-derived networks is due to false positives in BT or false negatives in GPS, but based on the precision of commodity GPS receivers and their propensity to lose signal in large buildings, we suspected that the observed disparities are predominantly driven by GPS false negatives. If this suspicion is warranted, GPS location–based proximity measurement should be used in epidemiological simulations with caution and in a fashion that anticipates and accounts for the fact that the data collection modality used may be systemically underestimating contact. This is particularly true for the short contacts outside of normal contact networks that drive mixing.

The significance results were relatively insensitive to differences in simulated disease impacting differences in GPS and BT, but the data collection modality induced fewer differences in the results for less contagious diseases, such as seasonal influenza, than for more contagious diseases, such as measles. It is possible that weakly contagious diseases might not demonstrate differences, as outbreaks would be rare and limited in both GPS and BT networks. These findings hold for both a nominal 8-m and 20-m threshold for determining if contact has occurred. The thresholds chosen are already judicious and indicate participants being close enough during a measured portion to have come into close contact during a sensor sleep period, rather than explicitly detecting close contact. Comparing the within-sensor outcomes, the contact threshold impacted the simulated attack rate for most cases, with the exception of SHED8.

We used a stylized, agent-based SEIR model to determine the attack rate using both BT- and GPS-inferred temporal contact patterns. The stylized nature of the simulation implies that the results should be generally correct, but that more detailed models may diverge in the magnitude of the differences observed. SHED7, SHED8, and SHED9 are interesting data sets due to the multiple sensor modalities, but they are also highly biased, being drawn from a university social science participant pool comprised primarily of undergraduate students in the social and physical sciences. GPS or BT data from other demographics will almost certainly have different contact patterns, leading to different outcomes. At one extreme, institutionalized individuals (eg, in incarceration facilities or care homes) have limited mobility and would be expected to have much more convergent GPS and BT contact patterns. Perhaps not surprisingly, some of the worst COVID-19 outbreaks occurred in these institutional settings. Similarly, we analyzed 4 relatively contagious diseases and ignored diseases where a specific type of contact initiates infection, such as sexually transmitted or blood-borne diseases, or where disease propagation is slow or exhibits prolonged latent periods, such as with tuberculosis. As the definition of contact for such excluded diseases is substantially different from those analyzed in this study, the difference between GPS and BT contact patterns may be more or less pronounced. The process we have used to evaluate the differences should generalize to any contagious disease or measured contact pattern and can be used to evaluate the impact of novel contact detection algorithms or other novel diseases such as COVID-19 variants of concern.

While this study has made several meaningful contributions to the literature, particularly in highlighting divergent attack rates for GPS and BT measurements of the same underlying contact network, it is subject to notable limitations. We used 3 data sets drawn from a social sciences participant pool at our institution. These data sets included individuals who were often unknown to each other and likely produced more diffuse data sets than would have been expected had we used snowball or respondent-driven sampling or other socially connected recruiting techniques. Running a similar analysis on other data sets could provide more broadly generalizable or representative results. However, for reasonable privacy reasons, public data sets containing both GPS and BT records are not available, requiring additional measurement effort to extend this analysis. We used an agent-based SEIR model because it provided the most direct link between the data and the simulated diseases. We chose the stylized SEIR model to emphasize the role of evolving contact networks in other disease dynamics. These results could be extended to include more sophisticated disease models and compared against compartmental transmission models grounded in aggregate representations of the underlying contact network. The COVID-19 pandemic has driven innovation in contact tracing, and new measurement techniques based on dongles, beacons, or badges are now readily available. A similar analysis including these data sources could be valuable. Finally, we constrained our analysis to 4 canonical contagious diseases from a historical perspective with relatively

XSL•FO
RenderX

well-parameterized behaviors. However, novel diseases will have novel disease parameters. An exploratory simulation study that outlined how diseases might be expected to behave over these contact networks using, for example, a random-walk through parameter space might be valuable in predicting new variants, existing diseases, or new diseases emerging from animal reservoirs.

Conclusions

Epidemiological models of disease propagation are an important tool in controlling and containing epidemic outbreaks. These models rely on the accurate measurement of key biological and behavioral parameters to ground the simulation results. Quantifying the characteristics of dynamic contact networks is a particularly challenging aspect of grounding these simulations. The significant differences in the predicted outcomes for contact networks demonstrated here between GPS- and BT-based contact tracking highlight the difficulty of grounding these simulations. Because of the nature of our data, we know that the contact networks being sought via measurement by BT and GPS should have been identical, as they corresponded to the same device held by the same individual as they went about their lives. The fact that the resulting contact networks and predicted attack rates were different indicates that these modalities are not interchangeable and that caution should be exercised by modelers employing these measures. While BT and GPS data provide more precise measurements than traditional surveys, they are still prone to error and disparate estimates of the underlying network structure and dynamics.

Acknowledgments

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for funding and Dr Regan Mandryk for her timely statistical advice. The authors would also like to thank Greg Oster and Mohammad Hashemian for their technical consulting and assistance. Finally, NDO would like to extend his gratitude to SYK for supporting his contributions.

Data Availability

Data cannot be made publicly because participant informed consent does not allow disclosure and because anonymized longitudinal location data can be reidentified through crosslinking by those skilled in spatial analysis. Data can be made available through a joint application to both the requester's and the University of Saskatchewan ethics boards, outlining the intended secondary use of the data and processes for ensuring that reidentification does not occur. The latest contact information for the research ethics board (REB) can be found on the website of the University of Saskatchewan. Currently, the web page for the REB is available [70].

Conflicts of Interest

None declared.

References

1. Satcher D. Emerging infections: getting ahead of the curve. Emerg Infect Dis 1995 Mar;1(1):1-6 [FREE Full text] [doi: 10.3201/eid0101.950101] [Medline: 8903147]
2. McNeill WH. Disease in history. Soc Sci Med (1967) 1978 Apr;12(2B):79-84. [doi: 10.1016/0160-7987(78)90012-1] [Medline: 358401]
3. WHO Coronavirus (COVID-19) dashboard. World Health Organization. URL: https://covid19.who.int/ [accessed 2023-08-08]
4. Cutler DM, Summers LH. The COVID-19 pandemic and the $16 trillion virus. JAMA 2020 Oct 20;324(15):1495-1496 [FREE Full text] [doi: 10.1001/jama.2020.19759] [Medline: 33044484]
5. Coronavirus disease (COVID-19): situation report. World Health Organization. URL: https://apps.who.int/iris/handle/10665/333832 [accessed 2023-08-08]
6. Ross R, Hudson HP. An application of the theory of probabilities to the study of a priori pathometry—part II. Proc R Soc Lond 1997 Jan 03;93(650):212-225 [FREE Full text] [doi: 10.1098/rspa.1917.0014]
7. Ross R. An application of the theory of probabilities to the study of a priori pathometry—part I. Proc R Soc Lond 1997 Jan;92(638):204-230 [FREE Full text] [doi: 10.1098/rspa.1916.0007]
8. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc Lond 1997 Jan;115(772):700-721. [doi: 10.1098/rspa.1927.0118]
9. Perez L, Dragicevic S. An agent-based approach for modeling dynamics of contagious disease spread. Int J Health Geogr 2009;8(1):50. [doi: 10.1186/1476-072x-8-50]
10. Osgood N, Liu J. Towards closed loop modeling: evaluating the prospects for creating recurrently regrounded aggregate simulation models using particle filtering. In: Proceedings of the 2014 Winter Simulation Conference. 2014 Presented at: WSC '14; 2014; Savannah, GA p. 829-841 URL: https://ieeexplore.ieee.org/document/7019944 [doi: 10.1109/wsc.2014.7019944]
11. Chowell G, Sattenspiel L, Bansal S, Viboud C. Mathematical models to characterize early epidemic growth: a review. Phys Life Rev 2016 Sep 27;18(5):66-97 [FREE Full text] [doi: 10.1016/j.plrev.2016.07.005] [Medline: 27451336]

XSL•FO
RenderX

12. Bobashev GV, Goedecke DM, Yu F, Epstein JM. A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches. In: Proceedings of the 2007 Winter Simulation Conference. 2007 Presented at: WSC '07; December 9-12, 2007; Washington, DC p. 1532-1537 URL: https://ieeexplore.ieee.org/document/4419767 [doi: 10.1109/wsc.2007.4419767]
13. Keeling MJ, Eames KT. Networks and epidemic models. J R Soc Interface 2005 Sep 22;2(4):295-307 [FREE Full text] [doi: 10.1098/rsif.2005.0051] [Medline: 16849187]
14. Sarkar S, Zlojutro A, Khan K, Gardner L. Measles resurgence in the USA: how international travel compounds vaccine resistance. Lancet Infect Dis 2019 Jul;19(7):684-686. [doi: 10.1016/s1473-3099(19)30231-2]
15. Rodríguez P, Graña S, Alvarez-León EE, Battaglini M, Darias FJ, Hernán MA, et al. A population-based controlled experiment assessing the epidemiological impact of digital contact tracing. Nat Commun 2021 Jan 26;12(1):587 [FREE Full text] [doi: 10.1038/s41467-020-20817-6] [Medline: 33500407]
16. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A systematic review of social contact surveys to inform transmission models of close-contact infections. Epidemiology 2019 Sep;30(5):723-736 [FREE Full text] [doi: 10.1097/EDE.0000000000001047] [Medline: 31274572]
17. Ahmed N, Michelin RA, Xue W, Ruj S, Malaney R, Kanhere SS, et al. A survey of COVID-19 contact tracing apps. IEEE Access 2020;8:134577-134601. [doi: 10.1109/access.2020.3010226]
18. Grantz KH, Cummings DA, Zimmer S, Vukotich C, Galloway D, Schweizer ML, et al. Age-specific social mixing of school-aged children in a US setting using proximity detecting sensors and contact surveys. Sci Rep 2021 Jan 27;11(1):2319 [FREE Full text] [doi: 10.1038/s41598-021-81673-y] [Medline: 33504823]
19. Hashemian M, Knowles D, Calver J, Qian W, Bullock MC, Bell S, et al. iEpi: an end to end solution for collecting, conditioningutilizing epidemiologically relevant data. In: Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare. 2012 Presented at: MobileHealth '12; June 11, 2012; Hilton Head Island, SC p. 3-8 URL: https://dl.acm.org/doi/10.1145/2248341.2248345 [doi: 10.1145/2248341.2248345]
20. Yeung WJ, Ng JK. An enhanced wireless LAN positioning algorithm based on the fingerprint approach. In: Proceedings of the 2006 IEEE Region 10 International Conference. 2006 Presented at: TENCON '06; November 14-17, 2006; Hong Kong, China p. 1-4 URL: https://ieeexplore.ieee.org/document/4142274 [doi: 10.1109/tencon.2006.343696]
21. Chen L, Pei L, Kuusniemi H, Chen Y, Kröger T, Chen R. Bayesian fusion for indoor positioning using bluetooth fingerprints. Wireless Pers Commun 2012 Aug 9;70(4):1735-1745. [doi: 10.1007/s11277-012-0777-1]
22. Pe'rez Iglesias HJ, Barral V, Escudero CJ. Indoor person localization system through RSSI bluetooth fingerprinting. In: Proceedings of the 19th International Conference on Systems, Signals and Image Processing. 2012 Presented at: IWSSIP '12; April 11-13, 2012; Vienna, Austria p. 40-43 URL: https://ieeexplore.ieee.org/document/6208163
23. Ganguly A, Reddy C, Hao Y, Panahi I. Improving sound localization for hearing aid devices using smartphone assisted technology. In: Proceedings of the 2016 IEEE International Workshop on Signal Processing Systems. 2016 Presented at: SiPS '16; October 26-28, 2016; Dallas, TX p. 165-170 URL: https://ieeexplore.ieee.org/document/7780091 [doi: 10.1109/access.2020.3021605]
24. Ens A, Höflinger F, Wendeberg J, Hoppe J, Zhang R, Bannoura A, et al. Acoustic self-calibrating system for indoor smart phone tracking. Int J Navig Obs 2015 Feb 26;2015:1-15. [doi: 10.1155/2015/694695]
25. Höflinger F, Zhang R, Hoppe J, Bannoura A, Reindl LM, Wendeberg J, et al. Acoustic self-calibrating system for indoor smartphone tracking (ASSIST). In: Proceedings of the 2012 International Conference on Indoor Positioning and Indoor Navigation. 2012 Presented at: IPIN '12; November 13-15, 2012; Sydney, Australia p. 1-9 URL: https://ieeexplore.ieee.org/document/6418877 [doi: 10.1109/ipin.2012.6418877]
26. Mackey A, Spachos P, Song L, Plataniotis KN. Improving BLE beacon proximity estimation accuracy through Bayesian filtering. IEEE Internet Things J 2020 Apr;7(4):3160-3169. [doi: 10.1109/jiot.2020.2965583]
27. Navstar GPS user equipment introduction (Public Release Version). U.S. Coast Guard. 1996. URL: https://www.navcen.uscg.gov/pubs/gps/gpsuser/gpsuser [accessed 2024-02-29]
28. Kjærgaard MB, Blunck H, Godsk T. Indoor positioning using GPS revisited. In: Proceedings of the 8th International Conference on Pervasive Computing. 2010 Presented at: Pervasive '10; May 17-20, 2010; Helsinki, Finland p. 56 URL: https://link.springer.com/chapter/10.1007/978-3-642-12654-3_3 [doi: 10.1007/978-3-642-12654-3_3]
29. Han B, Cook P, Baldwin T. Geolocation prediction in social media data by finding location indicative words. In: Proceedings of COLING 2012. 2012 Presented at: COLING '12; December 12-15, 2012; Mumbai, India p. 1045-1062 URL: https://aclanthology.org/C12-1064
30. Hasan S, Ukkusuri SV, Zhan X. Understanding social influence in activity location choice and lifestyle patterns using geolocation data from social media. Front ICT 2016 Jun 20;3:10. [doi: 10.3389/fict.2016.00010]
31. Rizwan M, Wan W, Gwiazdzinski L. Visualization, spatiotemporal patterns, and directional analysis of urban activities using geolocation data extracted from LBSN. ISPRS Int J Geo-Inf 2020 Feb 24;9(2):137-151 [FREE Full text] [doi: 10.3390/ijgi9020137]
32. Ethica data: empower your research with smartphones and big data. Ethica Data. 2021. URL: https://ethicadata.com/ [accessed 2023-08-08]

XSL•FO
RenderX

33. Michael K, Abbas R. Behind COVID-19 contact trace apps: the Google–Apple partnership. IEEE Consumer Electron Mag 2020 Sep 1;9(5):71-76 [FREE Full text] [doi: 10.1109/MCE.2020.3002492]
34. Wirth FN, Johns M, Meurers T, Prasser F. Citizen-centered mobile health apps collecting individual-level spatial data for infectious disease management: scoping review. JMIR Mhealth Uhealth 2020 Nov 10;8(11):e22594 [FREE Full text] [doi: 10.2196/22594] [Medline: 33074833]
35. Vazquez-Prokopec GM, Bisanzio D, Stoddard ST, Paz-Soldan V, Morrison AC, Elder JP, et al. Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. PLoS One 2013;8(4):e58802 [FREE Full text] [doi: 10.1371/journal.pone.0058802] [Medline: 23577059]
36. Carlotto A, Parodi M, Bonamico C, Lavagetto F, Valla M. Proximity classification for mobile devices using wi-fi environment similarity. In: Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments. 2008 Presented at: MELT '08; September 19, 2008; San Francisco, CA p. 43-48 URL: https://dl.acm.org/doi/10.1145/1410012.1410023 [doi: 10.1145/1410012.1410023]
37. Reichert L, Brack S, Scheuermann B. A survey of automatic contact tracing approaches using Bluetooth low energy. ACM Trans Comput Healthcare 2021 Mar 17;2(2):1-33. [doi: 10.1145/3444847]
38. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH. A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci U S A 2010 Dec 21;107(51):22020-22025 [FREE Full text] [doi: 10.1073/pnas.1009094108] [Medline: 21149721]
39. Adam D. A guide to R - the pandemic's misunderstood metric. Nature 2020 Jul 03;583(7816):346-348. [doi: 10.1038/d41586-020-02009-w] [Medline: 32620883]
40. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R). Emerg Infect Dis 2019 Jan;25(1):1-4 [FREE Full text] [doi: 10.3201/eid2501.171901] [Medline: 30560777]
41. Nishiura H, Chowell G. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: Chowell G, Hyman JM, Bettencourt LM, Castillo-Chavez C, editors. Mathematical and Statistical Estimation Approaches in Epidemiology. Cham, Switzerland: Springer; 2009:103-121.
42. Aronson JK, Brassey J, Mahtani KR. “When will it be over?”: an introduction to viral reproduction numbers, R0 and Re. The Centre for Evidence-Based Medicine. 2020. URL: https://www.cebm.net/covid-19/when-will-it-be-over-an-introduction-to-viral-reproduction-numbers-r0-and-re/ [accessed 2024-02-29]
43. Salonidis T, Bhagwat P, Tassiulas L. Proximity awareness and fast connection establishment in Bluetooth. In: Proceedings of the 2000 First Annual Workshop on Mobile and Ad Hoc Networking and Computing. 2001 Presented at: MOBHOC '00; August 11, 2000; Boston, MA p. 141-142 URL: https://ieeexplore.ieee.org/document/869227 [doi: 10.1109/mobhoc.2000.869227]
44. Shi Y, Shi W, Liu X, Xiao X. An RSSI classification and tracing algorithm to improve trilateration-based positioning. Sensors (Basel) 2020 Jul 30;20(15):4244 [FREE Full text] [doi: 10.3390/s20154244] [Medline: 32751485]
45. Khalil L, Jung P. Scaled unscented Kalman filter for RSSI-based indoor positioning and tracking. In: Proceedings of the 9th International Conference on Next Generation Mobile Applications, Services and Technologies. 2020 Presented at: NGMAS '20; September 9-11, 2020; Cambridge, UK p. 132-137 URL: https://ieeexplore.ieee.org/document/7373231 [doi: 10.1109/access.2020.3021605]
46. Xue W, Qiu W, Hua X, Yu K. Improved Wi-Fi RSSI measurement for indoor localization. IEEE Sens J 2017 Apr 1;17(7):2224-2230. [doi: 10.1109/jsen.2017.2660522]
47. Al Alawi R. RSSI based location estimation in wireless sensors networks. In: Proceedings of the 17th IEEE International Conference on Networks. 2011 Presented at: ICON '11; December 14-16, 2011; Singapore, Singapore p. 118-122 URL: https://ieeexplore.ieee.org/document/6168517 [doi: 10.1177/0002764289032005003]
48. Harun A, Ramli M, Kamarudin LM, Ndzi DL, Shakaff AY, Zakaria A, et al. Comparative performance analysis of wireless RSSI in wireless sensor networks motes in tropical mixed-crop precision farm. In: Proceedings of the 3rd International Conference on Intelligent Systems Modelling and Simulation. 2012 Presented at: ISMS '12; February 8-10, 2012; Kota Kinabalu, Malaysia p. 606-610. [doi: 10.1109/isms.2012.57]
49. Gu Y, Ren F. Energy-efficient indoor localization of smart hand-held devices using bluetooth. IEEE Access 2015;3:1450-1461. [doi: 10.1109/access.2015.2441694]
50. Young DG. Empedoclean epic: how far can you go? Dictynna. 2021. URL: https://web.archive.org/web/20220317222342/http://www.davidgyoungtech.com/2020/05/15/how-far-can-you-go [accessed 2022-03-17]
51. Networks R. Android beacon library: an android library providing APIs to interact with beacons. GitHub. 2019. URL: https://altbeacon.github.io/android-beacon-library/distance-calculations.html [accessed 2022-03-17]
52. Arenas A, Cota W, Gómez-Gardeñes J, Gómez S, Granell C, Matamalas JT, et al. Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions. Phys Rev X 2020 Dec 18;10(4):041055-1-04105521. [doi: 10.1103/physrevx.10.041055]
53. Tomochi M, Kono M. A mathematical model for COVID-19 pandemic-SIIR model: effects of asymptomatic individuals. J Gen Fam Med 2021 Jan;22(1):5-14 [FREE Full text] [doi: 10.1002/jgf2.382] [Medline: 33457150]
54. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Occam’s razor. Inf Process Lett 1987;24(6):377-380 [FREE Full text] [doi: 10.1016/0020-0190(87)90114-1]

XSL•FO
RenderX

55. Stanley K, Bell S, Kreuger LK, Bhowmik P, Shojaati N, Elliott A, et al. Opportunistic natural experiments using digital telemetry: a transit disruption case study. Int J Geogr Inf Sci 2016 Mar 21;30(9):1853-1872. [doi: 10.1080/13658816.2016.1145224]
56. Knowles DL, Stanley KG, Osgood ND. A field-validated architecture for the collection of health-relevant behavioural data. In: Proceedings of the 2014 IEEE International Conference on Healthcare Informatics. 2014 Presented at: ICHI '14; September 15-17, 2014; Washington, DC p. 79-88 URL: https://dl.acm.org/doi/10.1109/ICHI.2014.18 [doi: 10.1109/ichi.2014.18]
57. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Health 2020 May;5(5):e261-e270 [FREE Full text] [doi: 10.1016/S2468-2667(20)30073-6] [Medline: 32220655]
58. Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19). World Health Organization. 2020. URL: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf [accessed 2022-03-17]
59. Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infect Dis 2014 Sep 04;14(1):480 [FREE Full text] [doi: 10.1186/1471-2334-14-480] [Medline: 25186370]
60. Influenza: questions and answers. Coalition IA. 2020. URL: https://www.immunize.org/catg.d/p4208.pdf [accessed 2022-03-17]
61. Steele MK, Wikswo ME, Hall AJ, Koelle K, Handel A, Levy K, et al. Characterizing norovirus transmission from outbreak data, United States. Emerg Infect Dis 2020 Aug;26(8):1818-1825 [FREE Full text] [doi: 10.3201/eid2608.191537] [Medline: 32687043]
62. Lee RM, Lessler J, Lee RA, Rudolph KE, Reich NG, Perl TM, et al. Incubation periods of viral gastroenteritis: a systematic review. BMC Infect Dis 2013 Sep 25;13(1):446 [FREE Full text] [doi: 10.1186/1471-2334-13-446] [Medline: 24066865]
63. Fine P. Herd immunity: history, theory, practice. Epidemiol Rev 1993;15(2):265-302. [doi: 10.1093/oxfordjournals.epirev.a036121] [Medline: 8174658]
64. Hamborsky J, Kroger A. Epidemiology and prevention of vaccine-preventable diseases. Centers for Disease Control and Prevention. 2015. URL: https://www.cdc.gov/vaccines/pubs/pinkbook/index.html [accessed 2022-03-17]
65. Gao D, Lou Y, He D, Porco TC, Kuang Y, Chowell G, et al. Prevention and control of zika as a mosquito-borne and sexually transmitted disease: a mathematical modeling analysis. Sci Rep 2016 Jun 17;6(1):28070 [FREE Full text] [doi: 10.1038/srep28070] [Medline: 27312324]
66. Belik V, Geisel T, Brockmann D. Natural human mobility patterns and spatial spread of infectious diseases. Phys Rev X 2011 Aug 8;1(1):011001. [doi: 10.1103/physrevx.1.011001]
67. Amaral LA, Scala A, Barthelemy M, Stanley HE. Classes of small-world networks. Proc Natl Acad Sci U S A 2000 Oct 10;97(21):11149-11152 [FREE Full text] [doi: 10.1073/pnas.200327197] [Medline: 11005838]
68. Barabási AL, Bonabeau E. Scale-free networks. Sci Am 2003 May;288(5):60-69. [doi: 10.1038/scientificamerican0503-60] [Medline: 12701331]
69. Cabin RJ, Mitchell RJ. To Bonferroni or not to Bonferroni: when and how are the questions. Bull Ecol Soc Am 2000;81:246 [FREE Full text]
70. Human ethics. The University of Saskatchewan. URL: https://vpresearch.usask.ca/ethics/human-ethics.php#FrequentlyAskedQuestions [accessed 2024-03-05]

Abbreviations

ABM: agent-based model
BT: Bluetooth
dBm: decibel-milliwatts
KL: Kullback-Leibler
RSSI: Received Signal Strength Indicator
SEIR: susceptible-exposed-infectious-removed
SHED: Saskatchewan Human Ethology Dataset

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.04.22; peer-reviewed by K Xu, I Zachary, A Angelucci, S Pesälä; comments to author 28.07.23; revised version received 15.11.23; accepted 27.02.24; published 17.04.24.

Please cite as:
Qian W, Cooke A, Stanley KG, Osgood ND
Comparing Contact Tracing Through Bluetooth and GPS Surveillance Data: Simulation-Driven Approach
J Med Internet Res 2024;26:e38170
URL: https://www.jmir.org/2024/1/e38170
doi:10.2196/38170
PMID:38422493

©Weicheng Qian, Aranock Cooke, Kevin Gordon Stanley, Nathaniel David Osgood. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating Health and Well-Being Challenges Faced by an Aging Workforce in the Construction and Nursing Industries: Computational Linguistic Analysis of Twitter Data

Weicong Li[1], DPhil; Liyaning Maggie Tang[2], DPhil; Jed Montayre[3,4], DPhil; Celia B Harris[1], DPhil; Sancia West[5], DPhil; Mark Antoniou[1], DPhil

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, Australia
[2]School of Architecture and Built Environment, The University of Newcastle, Callaghan, Australia
[3]Centre of Evidence-based Practice for Health Care Policy, The Hong Kong Polytechnic University, Hung Hom, China (Hong Kong)
[4]School of Nursing and Midwifery, Western Sydney University, Penrith, Australia
[5]Centre for Work Health and Safety, New South Wales Government, Gosford, Australia

Corresponding Author:
Mark Antoniou, DPhil
The MARCS Institute for Brain, Behaviour and Development
Western Sydney University
Locked Bag 1797
Penrith, 2751
Australia
Phone: 61 61 2 97726673
Email: m.antoniou@westernsydney.edu.au

Abstract

Background: Construction and nursing are critical industries. Although both careers involve physically and mentally demanding work, the risks to workers during the COVID-19 pandemic are not well understood. Nurses (both younger and older) are more likely to experience the ill effects of burnout and stress than construction workers, likely due to accelerated work demands and increased pressure on nurses during the COVID-19 pandemic. In this study, we analyzed a large social media data set using advanced natural language processing techniques to explore indicators of the mental status of workers across both industries before and during the COVID-19 pandemic.

Objective: This social media analysis aims to fill a knowledge gap by comparing the tweets of younger and older construction workers and nurses to obtain insights into any potential risks to their mental health due to work health and safety issues.

Methods: We analyzed 1,505,638 tweets published on Twitter (subsequently rebranded as X) by younger and older (aged <45 vs >45 years) construction workers and nurses. The study period spanned 54 months, from January 2018 to June 2022, which equates to approximately 27 months before and 27 months after the World Health Organization declared COVID-19 a global pandemic on March 11, 2020. The tweets were analyzed using big data analytics and computational linguistic analyses.

Results: Text analyses revealed that nurses made greater use of hashtags and keywords (both monograms and bigrams) associated with burnout, health issues, and mental health compared to construction workers. The COVID-19 pandemic had a pronounced effect on nurses' tweets, and this was especially noticeable in younger nurses. Tweets about health and well-being contained more first-person singular pronouns and affect words, and health-related tweets contained more affect words. Sentiment analyses revealed that, overall, nurses had a higher proportion of positive sentiment in their tweets than construction workers. However, this changed markedly during the COVID-19 pandemic. Since early 2020, sentiment switched, and negative sentiment dominated the tweets of nurses. No such crossover was observed in the tweets of construction workers.

Conclusions: The social media analysis revealed that younger nurses had language use patterns consistent with someone experiencing the ill effects of burnout and stress. Older construction workers had more negative sentiments than younger workers, who were more focused on communicating about social and recreational activities rather than work matters. More broadly, these findings demonstrate the utility of large data sets enabled by social media to understand the well-being of target populations, especially during times of rapid societal change.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e49450) doi:10.2196/49450

KEYWORDS

social media; construction; nursing; aging; health and well-being; Twitter

Introduction

Background

Construction and nursing are 2 critical industries for the Australian economy and worldwide. Healthy construction workers are needed to build homes, commercial stores, and public infrastructure, affecting transport, health care services, commerce, recreation, and all aspects of daily life. A healthy and functioning nursing workforce is an essential resource for public health, as recognized by the World Health Organization (WHO). In the recent *2022 Skills Priority List* [1], which identified the most in-demand professions of the coming decade, both professions were listed in the top 10; specifically, construction managers claimed the top spot, whereas registered nurses secured the fourth position. Although both industries undeniably serve critical functions, there are immediately apparent differences between the 2 industries. These include the nature of the work, size of the organizations (60% of construction businesses are sole traders and 98.6% have <20 employees), and demographics of the 2 workforces (most construction workers are male, and most nurses are female) [2]. Despite these differences, the 2 industries face common challenges. They both have aging workforces, face worldwide labor shortages, involve work that is physically demanding and stressful, and need to retain and transfer older workers' expert knowledge. Prior work suggests that workers across these 2 industries face different work-related pressures and stress (particularly during the COVID-19 pandemic), which in turn are likely to impact their job performance, health, and career decisions [3]. The goal of this paper is to empirically examine social media use in the construction and nursing industries to provide insights into the challenges workers are facing and to use big data analytics to compare issues between industries, across age groups, and within Australia versus overseas.

Research Using Text Data From Social Media Platforms

Social media platforms provide workers with channels to express their experiences and feelings [4]. Social media networks may provide insights into the state of mind and the experiences of users. Such insights are typically obtained by conducting surveys or interviews with the samples drawn from the population or populations of interest. Although social media data analytics do not replace traditional face-to-face data collection methods such as surveys or interviews, they complement such approaches and make it possible to explore issues in ways that are not possible using more traditional approaches. First, because social media platforms are digitized, large data sets can be obtained for empirical research. Therefore, trawling social media data has become one of the best ways to analyze and predict trends within industries [5]. Second, social media makes it possible to compare large data sets both across time and large geographic distances. Third, the data can be subjected to state-of-the-art computational linguistic and natural language processing techniques to understand the pressures and risks that workers face within a given industry at a particular time and whether these challenges differ for subgroups within an industry (eg, younger vs older workers). The knowledge gained from such analyses can inform the design of interventions and policies to safeguard workers from potential harm. For these reasons, social media data are being used to explore issues related to work pressure, performance, career advancement, and work health and safety.

Traditionally, the construction industry has not been closely associated with the use of social media platforms, but construction companies are increasingly using social media to improve visibility and build brand awareness [6], share safety knowledge [7], enhance daily operations [8], measure students' attitudes [9], and gauge public opinion of projects. Although natural language processing is a growing area within the construction industry [10], only a few studies have applied computational linguistic techniques to social media data sets [5,11]. A recent systematic review of the literature on big data studies within construction called for research on how social media big data analytics can be used to prevent threats such as safety issues, injury, or mental illness caused by work-related stress [12]. This study directly addresses this need.

In nursing, numerous studies have examined the use of social media by nurses for a variety of purposes [13], including delivering health care to patients [14,15], training nursing students [16], and dissemination of communications during conferences [17]. However, the possibility of analyzing nurses' activity on social media platforms such as Twitter (subsequently rebranded as X) to understand their experiences of workplace stress and risks to their work health and safety has received surprisingly little attention to date.

An analysis of 53,063 tweets from January 2019 to December 2020 revealed that nurses experienced more frequent and intense negative emotions (eg, decreased joy and increased sadness, fear, and disgust) than in the year preceding the COVID-19 pandemic [18]. Notably, fear preceded the COVID-19 pandemic waves by 2 weeks, suggesting that frontline workers are finely attuned to increases in work pressure even in an already high-stress work environment. This has implications for preventing fatigue, burnout, and mental health disorders related to unhealthy or unsustainable work conditions via sensitive detection of changes to well-being at the population level.

An analysis of 4.5 million tweets posted by US and UK nurses revealed that health care providers in the 2 countries experienced common challenges concerning public health (eg, policy and COVID-19–related pressures), social values (related to aspects of health), and political issues related to the COVID-19 pandemic (more positive in the United Kingdom) [19]. However, the experiences in the 2 countries reflected local sociopolitical trends and the cultural norms regarding emotional display (more accepted in the United States and more reserved in the United

XSL•FO
RenderX

Kingdom). Both countries showed sharp increases in fear and sadness during the first wave of the COVID-19 pandemic and when there was an increase in the virus reproduction rates. Fear gradually reduced with time, but sadness was maintained. Anger was experienced by both groups in response to a rise in the number of COVID-19 deaths.

This raises the possibility that nurses are particularly vulnerable to poor mental health (eg, concerns, fear, and anxiety) due to the nature of their work, characterized by long working hours, burnout, loneliness, fatigue, and occupational stigma [20,21]. These risks are magnified when needing to manage a long-lasting emergency situation such as the COVID-19 pandemic, often with a lack of resources and under more demanding conditions [22]. While nurses have been praised for their frontline efforts in the care and treatment of patients with COVID-19, very little is known regarding how nurses have responded to the current emergency and how this is reflected in the language of their social media posts [23]. Furthermore, nurses have not been compared to workers in other industries. This study will directly address this knowledge gap.

Comparing Age Groups and Industries

Studies comparing work health and safety issues across different subgroups of workers (eg, younger vs older workers) and industries with diverse demographical profiles (eg, nursing and construction) are scarce. To our knowledge, this is the first study to compare construction and nursing industries with the goal of developing a detailed understanding of both the common and unique challenges faced by workers in these critical industries and how these challenges and risks interact with aging.

Age may be a key differentiator between workers that impacts their mental health and response to stress. In general, among health care workers [24], older adults compared to younger adults show better mental health and emotion regulation [25], including during the COVID-19 pandemic [26]. This suggests that although older workers are frequently depicted as frail and vulnerable, broader evidence from aging research suggests they could be more resilient and better able to adapt and accommodate periods of instability than younger workers. In prior work, we conducted a scoping literature review to understand the physical and mental factors that affect older construction workers' work ability [2]. We found that the literature was dominated by studies focusing on physical health factors (eg, hearing loss, muscle pain, respiratory issues, and conditions resulting from prolonged work in and around construction sites), and although there was evidence of mental health risks and harms [27], these were less well understood, including their interaction with physically debilitating conditions (eg, diminished mental health due to work-related stress is a precursor of physical injuries and chronic pain).

In contrast to the paucity of research on the mental health of construction workers, many reviews have been published on the work health and safety of nurses, including reviews focusing specifically on older nurses [28,29]. However, most of these reviews focused on macrolevel issues that affect workers at the organizational level. Consequently, in a systematic review of evaluated programs and interventions intended to support the health, well-being, and retention of aging nurses, we discovered that the interventions were often mismatched to the needs of the nurses the program was trying to address [30]. For example, wellness interventions that focus on healthy eating or yoga are unlikely to address the underlying systemic workplace challenges that nurses encounter as they age at work. Furthermore, support programs and interventions for nurses need to be both sustainable at the organizational level as well as adaptable to the workers' changing circumstances, as determined by their health and aging.

In addition, a recent survey of construction workers and nurses within New South Wales revealed that nurses were far more likely to experience the effects of stress, burnout, and workload pressures (LM Tang et al, unpublished data, September 2021). These effects were exacerbated by the COVID-19 pandemic in nurses, whereas construction workers reported being far less affected by the pandemic. Age differences were also observed.

Taken together, the findings from the abovementioned studies provide converging evidence from which a picture is beginning to emerge concerning the pressures under which workers in each industry are having to operate.

This Study

In this study, we conducted a detailed linguistic analysis of a large data set of social media posts from younger and older individuals working in the construction and nursing industries. Industry membership was determined by sourcing data from industry-based groups and Twitter hashtags and then filtering the data. Big data analytics and computational linguistics were used to examine patterns of word use relating to industry, age, location, and mental and physical health. Content and sentiment analyses allowed us to determine common issues discussed on social media across individuals working in nursing and construction industries and how they are associated with age. On the basis of the emerging picture of the work pressures within the nursing and construction industries reviewed in the earlier sections, we hypothesized that our analysis of tweets would detect greater effects of stress, burnout, and workload pressures in nurses, and these mental health effects would be exacerbated by the COVID-19 pandemic. An additional consideration was whether any observed barriers to well-being would vary across ages (ie, younger vs older workers) and locations (ie, Australia vs overseas).

Methods

Ethical Considerations

This study was conducted in full compliance with the National Statement on Ethical Conduct in Human Research and approved by the Western Sydney University Human Ethics Committee (approval number: H14518).

Platform

Twitter is a social media platform that allows users to publish short posts, called tweets, consisting of up to 280 characters about any topic, including their thoughts, daily activities, political opinions, and news. Twitter enables users to follow others without requiring confirmation, making tweets public. In total, >240 million active users access the Twitter service.

XSL•FO
RenderX

Approximately 6000 tweets are posted every second, which equates to 350,000 tweets per minute, 500 million tweets per day, and a staggering 200 billion tweets per year [31]. It has become one of the most important social media platforms in the world. At the time, Twitter had a publicly available application programming interface that developers and researchers could access to download Twitter data that fit a specific combination of criteria. For this study, data collection and processing were performed as illustrated in Figure 1. Notably, there were 2 steps: first, the data were harvested and filtered, and second, we performed data analytics, processing, and mining knowledge from the data.

Figure 1. Data collection, processing, and analysis workflow. LIWC: Linguistic Inquiry and Word Count.

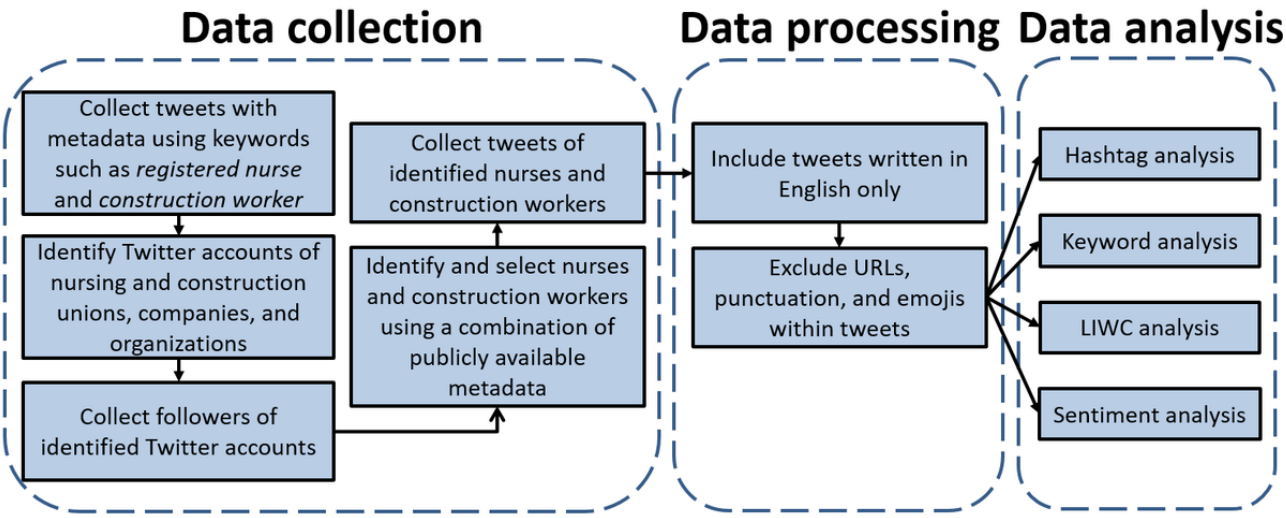

Population

To identify workers within construction and nursing professions, we first collected tweets together with their metadata using keywords such as *registered nurse* and *construction worker*. From these tweets and metadata, we identified a list of Twitter accounts of the relevant unions, companies, and organizations for each industry. Then, we collected the large lists of followers of these accounts. Next, we identified which of these followers were construction workers and nurses and used a combination of their publicly available metadata to determine their location (in or outside Australia) and their age category (eg, using their name and short biographical profile). This process was partially automated but also involved a manual component, particularly for determining the age category (>45 years or <45 years). We adopted the definition provided by the Australian Department of Employment and Workplace Relations [32], which categorizes mature-age workers as those aged >45 years. Determining the age of users involved some combination of looking at the user's profile picture, name or handle, cross-referencing with other social media accounts (eg, LinkedIn [Microsoft Corp] or Facebook [Meta Platforms, Inc]) or professional profile pages (eg, company, hospital, and university websites) that provide age data, or some proxy for estimating age (eg, the year a professional qualification was awarded). Users for whom we could not confidently determine their age category were excluded.

Data Collection

Data were collected from the Twitter platform using the Twitter application programming interface and were subject to the privacy policy regarding the release of the data held by the social media platforms to the public. Furthermore, keywords such as *construction worker* and *registered nurse* were used to retrieve potential nurses and construction workers and their tweets as an initial step. Metadata including username; name; user ID; language; hashtags; tweet time; and number of likes, replies, and retweets were collected simultaneously using *Twint* and *Tweepy*, Twitter scraping packages written in Python (Python Software Foundation) that allow for data collection from Twitter profiles.

Data Filtering and Preparation

Only tweets written in English were included in the analyses. URLs, punctuation marks, symbols, and emojis were removed. *Stanza* developed by Stanford Natural Language Processing Group and Python *Natural Language Toolkit* package were used for tokenization (ie, breaking a sentence into small units for subsequent analysis), removing stop words (eg, are, is, an, this, and that; the complete list is available on GitHub [33]), sentiment analysis, and monogram and bigram keyword analysis (ie, 1 word vs 2 words that co-occur, eg, bigrams involving mask included "wear mask" vs "anti mask").

Analysis

Hashtag Topics

A hashtag is a word or combination of words preceded by a # symbol, which is used to index topics on Twitter (eg, #construction, #nursing, and #COVID). This function was created on Twitter and allows people to easily follow topics they are interested in. We took advantage of this functionality to conduct a series of data visualizations and quantitative analyses to explore trending topics within and across industries and age categories.

Keywords

In addition to the abovementioned hashtag analysis, we analyzed the keywords contained in tweets (ie, the content words that remained after stop words, punctuation marks, and so on were

XSL•FO
RenderX

removed). This analysis provided a more detailed exploration of the themes raised by the hashtag topic analyses.

Word Counts Analysis

The Linguistic Inquiry and Word Count (LIWC) application [34] is the most widely used corpus of dictionaries for computational linguistic analyses of text data. Numerous studies on mental health topics using LIWC have shown strong evidence of particular patterns of language use that are highly relevant to certain mental health issues. Notably, LIWC compares words appearing in an input text file with the words listed in its customizable dictionary. Then, LIWC uses its algorithms to sort words into predefined and psychologically meaningful categories and performs a series of calculations. The end result is a summary of the statistical distribution of words within a given text into those that fall into LIWC categories, which include function words, pronouns, impersonal pronouns, verbs, auxiliary verbs, and past-tense words.

For this study, we used the latest version of LIWC (LIWC-22; Pennebaker Conglomerates, Inc) and the built-in English dictionary. Tweets containing <10 words were not included in the LIWC analysis, as they are unlikely to satisfy the requirements of LIWC and will not produce meaningful results.

Sentiment Analysis

Sentiment analysis was conducted using *Stanza*, the Python package of Stanford Core Natural Language Processing Group [35,36], which is based on deep learning and has been shown to have state-of-the-art performance [37]. Sentiment analysis is a powerful technique used to understand the public opinion of social media users and is suitable for use with short text posts such as tweets [38]. Using this approach, text can be classified as either positive, negative, or neutral sentiments. Examples of positive, negative, and neutral sentiments are as follows:

> *Great choice and important area that needs exploring. I would be confident even if you found someone exploring similar, you would be able to provide unique new knowledge. The anxiety element is real and recognised though! Good luck and feel free to DM.* [Positive sentiment]
>
> *Nope, so much bad energy used to devalue and treat ordinary Australians like cattle. Fascism, led by the religious right, will destroy this country. There hasn't been a time, in my life, when AU wellbeing is so low.* [Negative sentiment]
>
> *I was watching a show on tv last night re improving mental health which was timely given the day I had yesterday. It mentioned effects of social media on mental health.* [Neutral sentiment]

Results

Data Summary

The data set consisted of 1,505,638 tweets from 395 Twitter users, as presented in Multimedia Appendix 1. The data were collected across a period spanning 54 months, from January 2018 to June 2022, which equates to approximately 27 months before and 27 months after the WHO declared COVID-19 a global pandemic on March 11, 2020. Data were separated into those that originated in Australia versus those that originated in other parts of the world because tweets are known to reflect local sociopolitical trends and cultural norms (eg, regarding the display of emotions [19]). Furthermore, we divided users into younger and older workers to explore whether the challenges that workers face interact with age.

Hashtag Topics

The large data set contained a wide variety of hashtag topics. Overall, younger workers tended to use a wider variety of hashtag topics than older workers, and this pattern held for both industries. Comparing industries, nurses tended to use more hashtag topics than construction workers (Multimedia Appendix 2). Our large data set contained thousands of hashtag topics, with a very long tail in their distribution (ie, some obscure hashtags had only 1 or 2 times). To reduce the data for analysis, we grouped the top 200 hashtag topics into groups based on themes, and 10 unique thematic categories emerged (Multimedia Appendix 3).

Political hashtags were popular in both industries. In particular, older workers made greater use of political hashtags than younger workers, and this difference was especially pronounced in construction workers. Younger construction workers were by far the most likely to tweet about "life," including various entertainment activities, such as sports, eating out, camping, television shows, shopping, hobbies, and pets. This suggests the presence of a work-life balance disparity across the 2 industries. Delving deeper, we found that construction workers had a work-life topics ratio >1, whereas the ratio for nurses was <1, indicating that work was taking a disproportionate amount of nurses' time and attention, and this could be interpreted as an indicator of nurses being overworked. This pattern was exacerbated in the younger groups of workers: the work and life topics ratio of younger construction workers was 5 times higher than that of younger nurses.

Consistent with the expectations, nurses used COVID-19–related hashtags (eg, #COVID-19, #coronavirus, #mask, #vaccination, #lockdown, and #ppe) to a far greater extent than construction workers. This likely reflects their practice scope, expertise, and daily experiences in the course of performing their professional duties. Furthermore, nurses were more likely to include hashtags in their tweets falling within the "campaign" and "organization" categories, which included social media calls for action to improve working conditions (not relating to political purposes) and to voice support. Nurses, and especially older nurses, were more likely to use "positive words" in their hashtags, which consisted of SMS text messages voicing support for colleagues and others facing challenging circumstances in their industry (eg, #kind, #kindnessmatters, #respect, #grateful, #empathy, and #teamwork).

As shown in Table 1, nurses were far more likely to tweet hashtags about their well-being; mental health issues; and the deleterious effects of burnout, fatigue, and tiredness. Younger nurses tended to focus on mental health issues and burnout or fatigue, whereas older nurses were more likely to tweet about well-being. Older construction workers were more likely to tweet about pains, aches, and physical health issues.

XSL•FO
RenderX

Table 1. Hashtag topics used by younger and older nurses and construction workers directly related to health, well-being, and the COVID-19 pandemic (n=349,790).

Hashtag topic	Nurses (n=217,510, 62.18%), n (%)		Construction workers (n=132,460, 37.82%), n (%)	
	Younger (n=152,570, 70.14%)	Older (n=64,940, 29.86%)	Younger (n=77,929 58.83%)	Older (n=54,531, 41.67%)
Burnout, fatigue, tiredness, and sleep issues	93 (0.06)	37 (0.06)	10 (0.01)	6 (0.01)
Pain, ache, and physical health	42 (0.03)	10 (0.01)	10 (0.03)	52 (0.10)
Mental health	479 (0.31)	100 (0.15)	26 (0.03)	14 (0.03)
Well-being	210 (0.14)	189 (0.29)	17 (0.02)	2 (0.004)
COVID-19 pandemic	7031 (4.61)	1723 (2.65)	1488 (1.91)	1132 (2.08)

Tweets with hashtag topics had a higher average number of likes, replies, and retweets than topics without hashtags (Multimedia Appendix 4). Furthermore, tweets containing hashtag topics related to health and well-being had a higher average number of likes and retweets than those with hashtag topics other than health and well-being. Overall, tweets with hashtag topics received more attention and engagement, especially those about health and well-being topics.

Keywords

Monogram Keywords

Monogram keywords refer to individual content words within tweets that convey meaning. Compared to hashtags, keywords are more indicative of the content of the tweet, in that they better reflect the author's intended meaning. A summary of the monogram keyword data is provided in Multimedia Appendix 5.

Consistent with what was observed for hashtags, monogram keyword counts revealed that nurses made greater use of keywords associated with burnout, health issues, and mental health compared to construction workers (Table 2). Compounding the above, COVID-19 also had a huge effect on nurses, and this was especially pronounced for younger nurses.

Table 2. Monogram keywords used by younger and older nurses and construction workers directly related to health, well-being, and the COVID-19 pandemic (n=9,076,641).

Monogram keywords	Nurses (n=5,621,926, 61.94%), n (%)		Construction workers (n=3,454,715, 38.06%), n (%)	
	Younger (n=3,979,076, 70.78%)	Older (n=1,642,850, 29.22%)	Younger (n=2,167,081, 62.73%)	Older (n=1,287,634, 37.27%)
Burnout, fatigue, tiredness, and sleep issues	6066 (0.15)	1466 (0.09)	1643 (0.08)	724 (0.06)
Pain, ache, and physical health	6989 (0.18)	2410 (0.15)	1063 (0.05)	670 (0.05)
Mental health	7113 (0.18)	2256 (0.14)	1682 (0.08)	791 (0.06)
Well-being	622 (0.02)	802 (0.05)	112 (0.005)	57 (0.004)
COVID-19 pandemic	34,084 (0.86)	13,516 (0.82)	4484 (0.21)	4132 (0.32)

An analysis of likes, replies, and retweets revealed that tweets containing health and well-being keywords were more likely to gain the attention of other Twitter users, resulting in more engagement (Multimedia Appendix 6).

Monogram keywords directly related to physical and mental health and well-being were visualized as word clouds to illustrate the differences between industries and the age subgroups (Multimedia Appendix 7). Within each word cloud, the size of each word represents its frequency of occurrence, and the density of the cloud depicts the number of different keywords used. Differing patterns were observed for the 4 subgroups. Younger workers in both industries had word clouds that were denser, indicating that they posted about a wider variety of topics and problems. Younger nurses tweeted most frequently about "mental" and "sleep," whereas older nurses mentioned "pain" and "well-being" to a far greater extent and placed much less focus on "sleep." Among construction workers, both younger and older workers mentioned "sleep," "pain," and "tired." The older construction workers had the sparsest word cloud.

Bigram Keywords

Bigram keywords refer to pairs of content words that co-occur within tweets and convey additional context and the author's intended meaning (eg, the monogram "mask" has different connotations in the bigrams "wear mask" vs "anti mask"). A summary of bigrams is presented in Multimedia Appendix 8. Nurses consistently used more bigrams than construction workers.

Consistent with what was observed for hashtag topics and monograms, nurses made greater use of bigram keywords associated with burnout and physical and mental health issues

XSL•FO
RenderX

(Table 3). Furthermore, the COVID-19 pandemic disproportionately affected nurses, and this effect was magnified for younger nurses.

Table 3. Bigram keywords used by younger and older nurses and construction workers directly related to health, well-being, and the COVID-19 pandemic (n=3,528,198).

Monogram keywords	Nurses (n=2,155,270, 61.09%), n (%)		Construction workers (n=1,372,928, 38.91%), n (%)	
	Younger (n=1,490,264, 69.15%)	Older (n=665,006, 30.85%)	Younger (n=891,042, 64.9%)	Older (n=481,886, 35.1%)
Burnout, fatigue, tiredness, and sleep issues	2093 (0.14)	551 (0.08)	614 (0.07)	315 (0.06)
Pain, ache, and physical health	3371 (0.23)	1038 (0.16)	591 (0.07)	236 (0.05)
Mental health	4882 (0.33)	1510 (0.23)	1023 (0.11)	564 (0.12)
Well-being	498 (0.03)	447 (0.07)	49 (0.005)	24 (0.005)
COVID-19 pandemic	12,281 (0.82)	4036 (0.61)	1571 (0.18)	1520 (0.32)

Bigram word clouds related to physical and mental health issues and well-being are presented in Multimedia Appendix 9. The bigrams reveal additional details that were obscured in the monogram clouds. “Mental health” was the most frequently occurring bigram for all groups. “Mental illness” featured prominently for younger nurses and both groups of construction workers, but “mental well-being” occurred more frequently for older nurses. Furthermore, older nurses posted about physical well-being more than the other groups. Notably, density increased for the older groups relative to their monogram clouds. Younger nurses still had the densest cloud, and the bigrams within indicated that they were experiencing the most distress of all the groups (eg, “fall asleep,” “I am tired,” “not sleep,” “sleep tonight,” and “pain relief”). Older construction workers made frequent reference to specific medical conditions such as cardiovascular disease. Younger construction workers posted about nerve pain and neuropathic pain, likely related to the physical nature of construction work.

LIWC Results

For the LIWC analyses, tweets were grouped into those that did or did not contain health and well-being keywords. In Multimedia Appendix 10, we present the basic LIWC summary variables. Authenticity scores suggest that tweets were equally truthful regardless of whether the tweet was about a health-related topic or other topics. Analytical thinking, clout, and positive emotional tone were greater in tweets about non–health-related topics.

As presented in Multimedia Appendix 11, overall, tweet authors used more first-person singular pronouns and affect words when posting about health and well-being. Increased use of first-person singular pronouns is a reliable predictor of elevated levels of stress, depression, and suicidal ideation [39,40]. Furthermore, we found more affect words in health-related tweets. As expected, non–health-related tweets contained more words associated with the social category.

LIWC analyses of positive versus negative emotions and tone revealed that tweets about health and well-being were far more likely to convey negative emotions, perhaps reflecting fear, despair, sadness, or anger (Multimedia Appendix 12). Central to our research question and aims, we observed that words falling within the LIWC categories of *health*, *fatigue*, *illness*, *wellness*, and *mental* were far more likely to occur in tweets about health and well-being.

Then, we further examined the LIWC indicators for the tweets with health and well-being keywords. Comparing younger and older workers across industries, we observed that younger nurses showed language use patterns, suggesting that they were experiencing the deleterious effects of suboptimal work conditions (Multimedia Appendix 13). Younger nurses used words within the *fatigue*, *illness*, and *mental* categories more often than their older colleagues. Older nurses made greater use of words in the *wellness* category. These differences all reached statistical significance (Table 4). In contrast, no differences were found between the younger and older construction workers. In general, construction workers spent less time posting about health and well-being topics, and no differences were observed between younger and older construction workers.

XSL•FO
RenderX

Table 4. ANOVA tests for the Linguistic Inquiry and Word Count (LIWC) indicators health, fatigue, illness, wellness, and mental comparing younger versus older workers in the nursing and construction industries.

LIWC indicators	Younger vs older nurses		Younger vs older construction workers	
	F test (*df*)	*P* value	*F* test (*df*)	*P* value
Health	1.18 (1, 16,183)	.28	2.72 (1, 4138)	.10
Fatigue	21.82 (1, 16,183)	<.001	0.33 (1, 4138)	.56
Illness	17.12 (1, 16,183)	<.001	0.52 (1, 4138)	.47
Wellness	228 (1, 16,183)	<.001	0.07 (1, 4138)	.79
Mental	18.58 (1, 16,183)	<.001	0.08 (1, 4138)	.78

Sentiment Analysis

Sentiment analysis revealed that, overall, older construction workers made more tweets with negative sentiment (Figure 2). Younger construction workers were the most neutral. Older nurses tended to be more positive. Other than that, the differences between the younger and older nurses were relatively minor.

Figure 2. Percentage of tweets with negative, neutral, and positive sentiments among younger and older nurses and construction workers.

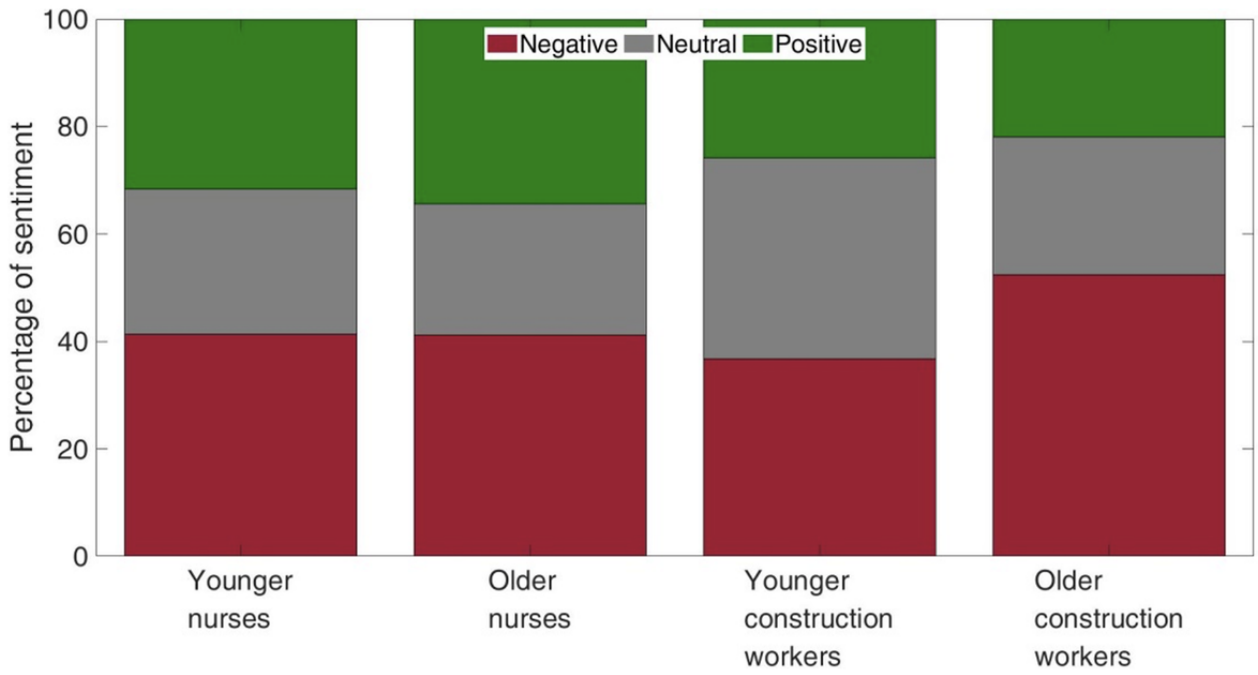

Next, we conducted a sentiment analysis of the tweets that did or did not contain keywords related to health and well-being. The right panel of Multimedia Appendix 14 presents the tweets without health and well-being keywords and serves as a baseline. The left panel of Multimedia Appendix 14 shows an increase in negative sentiment when tweeting about health or well-being topics. Younger nurses and older construction workers were the most negative, whereas older nurses were the most positive.

Australian Workers Versus Non-Australian Workers (LIWC Analysis)

To explore whether our observations were specific to the Australian context or reflected trends observable in other nations, we compared LIWC categories in Australian and non-Australian workers (Multimedia Appendix 15). In total, 51.5% (53/103) of the non-Australian construction workers come from Europe and North America, whereas 48.5% (50/103) come from Africa, Asia, and Latin America. Regarding nurses, 75% (78/104) of the non-Australian workers come from Europe and North America, and 25% (26/104) come from Africa, Asia, and Latin America. We observed that Australian workers in both industries used more words falling within the *mental* category, which was significant as confirmed by ANOVA tests (Table 5).

XSL•FO
RenderX

Table 5. ANOVA tests for the Linguistic Inquiry and Word Count (LIWC) indicators health, fatigue, illness, wellness, and mental comparing younger versus older workers in the nursing and construction industries.

LIWC indicators	Australian nurses vs non-Australian nurses				Australian construction workers vs non-Australian construction workers			
	Younger nurses		Older nurses		Younger construction workers		Older construction workers	
	F test (*df*)	*P* value	*F* test (*df*)	*P* value	*F* test (*df*)	*P* value	*F* test (*df*)	*P* value
Health	2.69 (1, 12329)	.10	4.79 (1, 3852)	.03	0.1 (1, 2732)	.75	59.95 (1, 1404)	<.001
Fatigue	3.64 (1, 12329)	.06	2.17 (1, 3852)	.14	8.04 (1, 2732)	.005	15.8 (1, 1404)	<.001
Illness	1.42 (1, 12329)	.23	46.05 (1, 3852)	<.001	14.98 (1, 2732)	<.001	22.59 (1, 1404)	<.001
Wellness	0.59 (1, 12329)	.44	64.73 (1, 3852)	<.001	18.59 (1, 2732),	<.001	32.78 (1, 1404)	<.001
Mental	23.85 (1, 12329)	<.001	55.57 (1, 3852)	<.001	30.99 (1, 2732)	<.001	23.44 (1, 1404)	<.001

This could be a sign of greater stress in Australian workers, potentially leading to future mental health concerns. Furthermore, Australians made more frequent reference to *health* and *illness* (with the exception of younger construction workers), suggesting that their health had been adversely affected, or at least that it was on their mind. Use of *health*-, *fatigue*-, *illness*-, and *wellness*-related words by younger nurses within and outside Australia were similar, whereas older Australian nurses made significantly more use of *health*- and *illness-related* words and less use of *wellness-related* words than their peers outside Australia. Similar results were found for older Australian construction workers compared to those outside Australia (except for more use of wellness-related words). On the other hand, younger Australian construction workers had less use of *fatigue*- and *illness-related* words, which might suggest that they are in a better condition compared to their colleagues worldwide.

The COVID-19 Pandemic Influence

The average number of tweets per user over the 54-month period between January 2018 and June 2022 increased more for both younger and older nurses than construction workers, especially since the COVID-19 pandemic. Furthermore, we observed marked differences in the 4 subgroups regarding sentiment at baseline and changes that occurred during the COVID-19 pandemic (Figure 3). Overall, nurses had a higher proportion of positive sentiment in their tweets than construction workers. However, this changed markedly in early 2020 as the positive and negative sentiments crossed over in the months leading up to the WHO's declaration of COVID-19 as a global pandemic. Since that time, negative sentiment dominated the tweets of nurses. No such crossover was observed among construction workers, in part because both younger and older construction workers consistently had more negative sentiments dating back to January 2018.

Figure 3. Percentage of tweets with negative, neutral, and positive sentiment changes by month for (A) younger nurses, (B) older nurses, (C) younger construction workers, and (D) older construction workers. The dashed vertical line indicates the announcement of the COVID-19 pandemic by the World Health Organization on March 11, 2020.

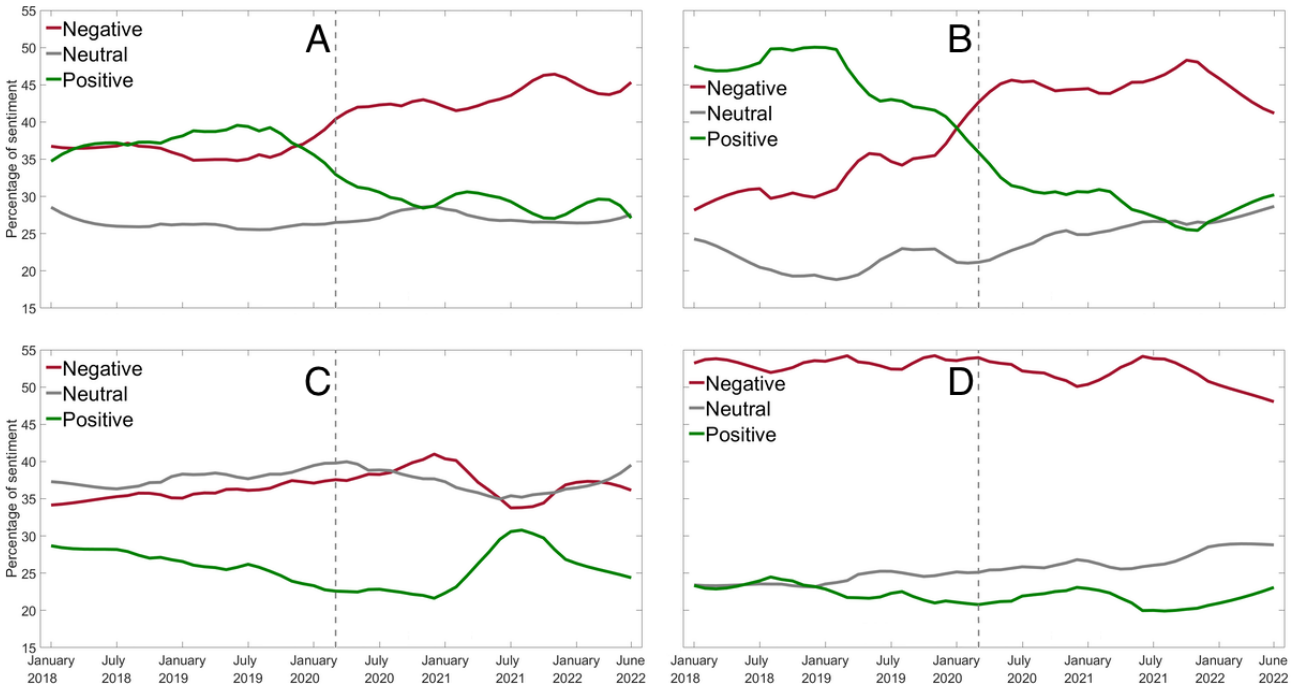

XSL•FO
RenderX

For all groups, LIWC for *illness*, *mental*, and *fatigue* categories were the most prominent, whereas the LIWC for the *wellness* category either declined or remained low across the 54-month period (Figure 4). For the older nurses, LIWC for the *wellness* category decreased but showed a bump that peaked in mid-2021, although the peak was considerably lower than the pre–COVID-19 pandemic levels. Younger nurses showed a substantial increase in *mental* words leading up to the WHO's COVID-19 pandemic declaration in March 2020. Younger construction workers were the only group for whom the *fatigue* category had a higher percentage than the *mental* category, indicating that their focus was on their physical health rather than mental health during the sampling period.

Figure 4. Changes in the Linguistic Inquiry and Word Count (LIWC) indicators illness, wellness, mental, and fatigue by month for (A) younger nurses, (B) older nurses, (C) younger construction workers, and (D) older construction workers. The dashed vertical line indicates the announcement of the COVID-19 pandemic. The calculation of LIWC indicators used only the tweets containing health and well-being keywords.

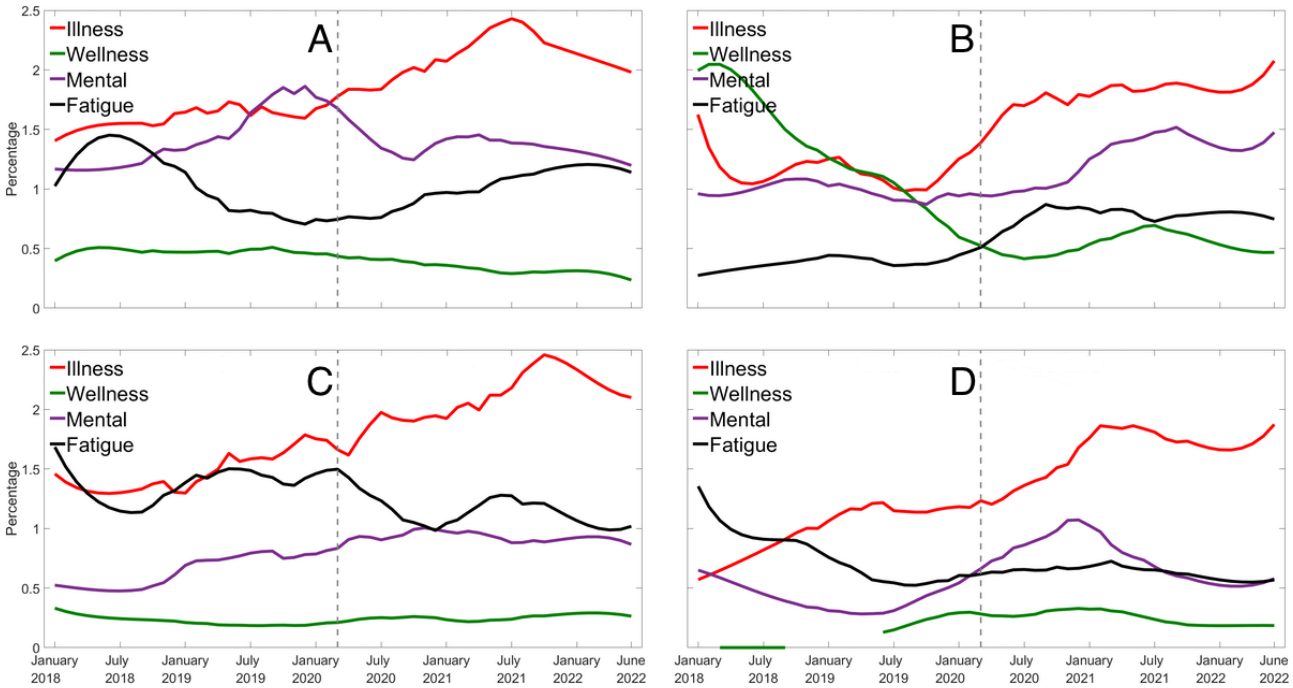

Discussion

Principal Findings

This is the first social media analysis to compare the tweets of younger and older workers in the construction and nursing industries with a view to understanding the common and unique work health and safety challenges they face and subsequent risks to their mental health. A robust observation across each type of analysis was that younger nurses were facing very challenging working conditions (eg, understaffing, high workload, and poor management and communication) that presented barriers to their mental health and well-being. Older nurses faced similar challenges but tended to place greater focus on promoting well-being in their tweets than their younger counterparts. Overall, nurses were unquestionably negatively impacted by the COVID-19 pandemic. In contrast, construction workers showed minimal impacts due to the COVID-19 pandemic. Older construction workers tended to post more negative tweets; however, construction workers in general tended to post less about work-related topics and more about life-related topics, indicating a better work-life balance than was observed in nurses. Younger construction workers posted mostly about life events, suggesting that they faced fewer mental health risks than the other groups.

Comparison With Prior Work

The most concerning observation that was borne out in each of our different social media analyses was that nurses (and especially younger nurses) were under more work pressure and stress and experienced greater and more intense negative emotions. Moreover, all these emotions were exacerbated by the COVID-19 pandemic. This puts younger nurses at increased risk of burnout and mental harm that may lead to attrition. This pattern of findings is consistent with a recent report that found a 131.5% increase in job vacancies under "health care and social assistance" industries between February 2020 and May 2022 [41]. Although there is a high attrition rate among our older nursing population through retirement, there may be higher attrition among the next generation of nurses through burnout. Previous research has shown that nurses' mental state plays a critical role in how they respond to situations, patients, and colleagues, as well as their clinical judgment and communication, potentially affecting patient safety [42].

Furthermore, we observed more negative sentiment in tweets about health and well-being topics. This does not align with prior reports that tweets with health care hashtags expressed more positivity and more action-oriented language than non–health care–initiated hashtags [19,43]. It has been suggested that nurses may be reluctant to share negative events on social media due to societal expectations of professionalism from medical experts [13]. It is possible that the discrepancy between

XSL•FO
RenderX

prior findings and our results may reflect Australian cultural norms and the extreme pressure that Australian health care workers were experiencing during the study period [13]. Notably, an analysis of tweets comparing UK and US health care workers revealed that while they experienced common challenges during the COVID-19 pandemic, nurses' experiences in the 2 countries reflected local sociopolitical trends and cultural norms regarding emotional display [19]. Future work should systematically investigate such processes in the Australian context.

Similar to their younger colleagues, older nurses faced challenging work conditions. Furthermore, although they showed signs of elevated distress and negative effects, we found evidence that older nurses possess more effective tools and strategies for dealing with periods of increased uncertainty, stress, and accelerated work demands (eg, emotional regulation and big picture thinking). A common assumption is that older workers are frail and vulnerable; however, evidence suggests that this is an inaccurate stereotype. A lifetime of experience seems to have equipped older workers with resilience and the ability to take a long-term view that permits them to better accommodate periods of instability. These suggestions in our data are consistent with broader research on cognitive aging, which suggests that older adults are skilled at emotion regulation [25]. Moreover, the findings are consistent with prior work on the effects of the COVID-19 pandemic on health care workers, which suggested that younger workers were most at risk of mental illness [26].

Overall, construction workers were less affected by the COVID-19 pandemic. Younger construction workers were mostly focused on posting about life events, including socializing, participating in hobbies, and engaging in recreational activities. Their tweets suggested that they had the best work-life balance, were not preoccupied with work, and showed minimal signs of mental distress. We did find evidence of concern with physical pains and injuries, which is notable given their young age. Older construction workers were more likely to post with more negative sentiment. However, these tweets tended to be regarding issues and events outside of their work, such as politics and current events, rather than about issues concerning their job role or working conditions (in contrast to what was observed for the nurses).

Unlike prior reports [44-46], we did not find evidence of widespread ageism (or reverse ageism) in either industry, although aspects of workplace culture could be improved. Instances of bullying were observed in both industries, originating from various sources, but they did not seem to be widespread. Differences in well-being were primarily related to age, job role, and job type. The findings of this study are compatible with the abovementioned observations and complement the data to provide a detailed insight into the mental health challenges faced by workers in each industry.

This work introduced several innovations. Our team applied social media–based big data analysis to inform our understanding of work health and safety issues. Furthermore, this was the first study to pay particular attention to issues concerning older workers' mental health by comparing findings and trends across 2 industries as well as comparing patterns in Australia versus overseas.

Implications

Computational linguistic tools and algorithms are able to reliably predict risks of future mental illness [47,48]. The patterns observed in this study strongly suggested that nurses were experiencing stress and negative emotions, and this was more extreme in younger nurses. A well-developed literature has established which linguistic indicators are reliable predictors of mental ill health (eg, first-person pronouns and negative emotion words) [39]. Language analysis technologies represent an important advancement in mental health care for the prevention and early diagnosis of mental health problems. These tools may aid professionals in identifying at-risk individuals and the follow-up and prognosis of patients [49].

Limitations

There were several limitations inherent to social media analysis when attempting to extrapolate to real-world work settings. First, not everyone (especially older workers) uses social media; therefore, we cannot assume that the comments, word use patterns, or sentiments of social media users accurately capture those of all construction workers or nurses. To address this limitation, we advocate not relying on any individual tool or method when attempting to develop a deep and holistic understanding of the issues facing an industry or workforce. It is for this reason that, in the present line of research, our team has conducted multiple literature reviews, a survey study, and now a social media analysis before attempting to design an intervention to improve the mental well-being of workers in these 2 industries.

A second limitation is that people whose data were included in this study tend to be those either who have sufficient free time or those who have strong opinions on the subject being studied. This limitation is somewhat circumvented by conducting a big data analysis on tweets that were already publicly available and freely expressed in different contexts. Furthermore, the fact that our observations converge with what we observed in our prior work gives us a high level of confidence that the conclusions drawn from these data are valid.

Conclusions

This study has advanced our understanding of the well-being of younger and older construction workers and nurses in Australia. The analyses have revealed that nurses (both younger and older) more often communicated about burnout, stress, and ill effects due to poor working conditions during the COVID-19 pandemic than construction workers. However, older nurses tended to promote well-being in their tweets more than their younger counterparts, suggesting that they could be more resilient and better able to adapt and accommodate periods of instability. Older construction workers had more negative sentiments than younger workers, who were more focused on communicating about social and recreational activities rather than work matters. These findings will inform the development of interventions that will protect the mental health of workers in highly demanding work environments such as nursing and construction. In a recent strategic report, the WHO [50] advised

XSL•FO
RenderX

that social media accounts (eg, Twitter) could be used for monitoring and provisioning training and educational materials for health care professionals and the evaluation of action plans. Our findings support the value of analyzing social media posts to gain insights into the well-being of workers in target industries and during key social challenges. These tools, and the evidence they generate, should be used to improve the lives of workers, especially those in critical industries.

Acknowledgments

This study was supported by funding from the New South Wales Government Centre for Work Health and Safety (tender SAFE/1848—CWHS_RP_091).

Data Availability

The data sets generated and analyzed during this study are available in the Western Sydney University Research Direct repository [51].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of users and tweets collected during the study period (January 2018 to June 2022).
[DOCX File , 13 KB - jmir_v26i1e49450_app1.docx]

Multimedia Appendix 2
Summary of the hashtag topics used by younger and older workers in the nursing and construction industries.
[DOCX File , 13 KB - jmir_v26i1e49450_app2.docx]

Multimedia Appendix 3
Distribution of the top 200 hashtag topics used by nurses (top panel) and construction workers (bottom panel) expressed as a percentage of the 10 most popular categories.
[PNG File , 122 KB - jmir_v26i1e49450_app3.png]

Multimedia Appendix 4
Average (SE) number of likes, replies, and retweets for tweets with hashtag topics related to health and well-being, with other hashtag topics and topics without hashtags.
[DOCX File , 13 KB - jmir_v26i1e49450_app4.docx]

Multimedia Appendix 5
Summary of the monogram keywords used in tweets by younger and older nurses and construction workers.
[DOCX File , 13 KB - jmir_v26i1e49450_app5.docx]

Multimedia Appendix 6
Average (SE) number of likes, replies, and retweets for tweets with and without health and well-being keywords.
[DOCX File , 13 KB - jmir_v26i1e49450_app6.docx]

Multimedia Appendix 7
Word clouds summarizing the monogram keywords related to physical and mental health issues and well-being for younger nurses (top left), older nurses (top right), younger construction workers (bottom left), and older construction workers (bottom right).
[DOCX File , 661 KB - jmir_v26i1e49450_app7.docx]

Multimedia Appendix 8
Summary of the bigram keywords used by younger and older nurses and construction workers.
[DOCX File , 13 KB - jmir_v26i1e49450_app8.docx]

Multimedia Appendix 9

XSL•FO
RenderX

Word clouds summarizing the bigram keywords related to physical and mental health issues and well-being for younger nurses (top left), older nurses (top right), younger construction workers (bottom left), and older construction workers (bottom right).
[DOCX File , 790 KB - jmir_v26i1e49450_app9.docx]

Multimedia Appendix 10
Descriptive statistics of the Linguistic Inquiry and Word Count scores for the basic summary variables: analytical thinking, clout, authenticity, and emotional tone. Average scores and SE are calculated based on the text from each tweet.
[DOCX File , 93 KB - jmir_v26i1e49450_app10.docx]

Multimedia Appendix 11
Percentage of words falling within the Linguistic Inquiry and Word Count categories: first-person singular pronouns, drives, cognition, affect, and social. Average percentage and SE are calculated based on the text from each tweet.
[DOCX File , 91 KB - jmir_v26i1e49450_app11.docx]

Multimedia Appendix 12
Percentage of words falling within the Linguistic Inquiry and Word Count categories: positive emotional tone, negative emotional tone, positive emotion, negative emotion, health, fatigue, illness, wellness, and mental. Average percentage and SE are calculated based on the text from each tweet.
[PNG File , 133 KB - jmir_v26i1e49450_app12.png]

Multimedia Appendix 13
Percentage of words falling within the Linguistic Inquiry and Word Count categories: health, fatigue, illness, wellness, and mental. Average percentage and SE are calculated based on the text from each tweet containing health and well-being keywords. Error bars show SE values.
[PNG File , 112 KB - jmir_v26i1e49450_app13.png]

Multimedia Appendix 14
Percentage of tweets with negative, neutral, and positive sentiments for younger and older nurses and construction workers, grouped by with and without keywords related to health and well-being in the tweets.
[DOCX File , 103 KB - jmir_v26i1e49450_app14.docx]

Multimedia Appendix 15
Percentage of words falling within the Linguistic Inquiry and Word Count categories: health, fatigue, illness, wellness, and mental. Average percentage and SE are calculated based on the text from each tweet containing health and well-being keywords.
[PNG File , 189 KB - jmir_v26i1e49450_app15.png]

References

1. 2022 Skills priority list: key findings report. National Skills Commission, Australian Government. 2022 Oct 6. URL: https://www.jobsandskills.gov.au/sites/default/files/2023-12/2022%20Skills%20Priority%20List%20Key%20Findings%20Report_0.pdf [accessed 2023-10-06]
2. Ranasinghe U, Tang LM, Harris C, Li W, Montayre J, de Almeida Neto A, et al. A systematic review on workplace health and safety of ageing construction workers. Saf Sci 2023 Nov;167:106276. [doi: 10.1016/j.ssci.2023.106276]
3. Chen Y, Ingram C, Downey V, Roe M, Drummond A, Sripaiboonkij P, et al. Employee mental health during COVID-19 adaptation: observations of occupational safety and health/human resource professionals in Ireland. Int J Public Health 2022;67:1604720. [doi: 10.3389/ijph.2022.1604720] [Medline: 36016962]
4. boyd D. Social media: a phenomenon to be analyzed. Soc Media Soc 2015 May 11;1(1):205630511558014. [doi: 10.1177/2056305115580148]
5. Tang L, Zhang Y, Dai F, Yoon Y, Song Y, Sharma RS. Social media data analytics for the U.S. construction industry: preliminary study on Twitter. J Manage Eng 2017 Sep 12;33(6). [doi: 10.1061/(ASCE)ME.1943-5479.0000554]
6. Hendricks D. How construction companies are leveraging social media. Socialnomics. 2015 Jan 29. URL: https://socialnomics.net/2015/01/29/how-construction-companies-are-leveraging-social-media/ [accessed 2023-05-12]
7. Yao Q, Li RY, Song L, Crabbe MJ. Construction safety knowledge sharing on Twitter: a social network analysis. Saf Sci 2021 Nov;143:105411. [doi: 10.1016/j.ssci.2021.105411]
8. Azhar S, Riaz Z, Robinson D. Integration of social media in day-to-day operations of construction firms. J Manag Eng 2018 Oct 15;35(1). [doi: 10.1061/(asce)me.1943-5479.0000660]
9. Yılmaz İ. Discussing future role of social media on construction and project management through evaluation of students attitudes. Artium 2020;8(2):128-134.

XSL•FO
RenderX

10. Ding Y, Ma J, Luo X. Applications of natural language processing in construction. Autom Constr 2022 Apr;136:104169. [doi: 10.1016/j.autcon.2022.104169]
11. Tang L, Zhang Y, Dai F, Yoon Y, Song Y. What construction topics do they discuss in social media? a case study of Weibo in China. In: Proceedings of the Construction Research Congress 2018. 2018 Presented at: CRC 2018; April 2-4, 2018; New Orleans, LA. [doi: 10.1061/9780784481264.060]
12. Madanayake UH, Egbu C. Critical analysis for big data studies in construction: significant gaps in knowledge. Built Environ Proj Asset Manag 2019 Jul 17;9(4):530-547. [doi: 10.1108/BEPAM-04-2018-0074]
13. Wilson R, Ranse J, Cashin A, McNamara P. Nurses and Twitter: the good, the bad, and the reluctant. Collegian 2014;21(2):111-119. [doi: 10.1016/j.colegn.2013.09.003] [Medline: 25109209]
14. Househ M, Borycki E, Kushniruk A. Empowering patients through social media: the benefits and challenges. Health Informatics J 2014 Mar;20(1):50-58 [FREE Full text] [doi: 10.1177/1460458213476969] [Medline: 24550564]
15. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res 2016 Aug 26;16(1):442 [FREE Full text] [doi: 10.1186/s12913-016-1691-0] [Medline: 27562728]
16. Jones R, Kelsey J, Nelmes P, Chinn N, Chinn T, Proctor-Childs T. Introducing Twitter as an assessed component of the undergraduate nursing curriculum: case study. J Adv Nurs 2016 Jul;72(7):1638-1653. [doi: 10.1111/jan.12935] [Medline: 26861572]
17. Ferguson C, Inglis SC, Newton PJ, Cripps PJ, MacDonald PS, Davidson PM. Social media: a tool to spread information: a case study analysis of twitter conversation at the Cardiac Society of Australia and New Zealand 61st annual scientific meeting 2013. Collegian 2014;21(2):89-93. [doi: 10.1016/j.colegn.2014.03.002] [Medline: 25109206]
18. Elyashar A, Plochotnikov I, Cohen IC, Puzis R, Cohen O. The state of mind of health care professionals in light of the COVID-19 pandemic: text analysis study of Twitter discourses. J Med Internet Res 2021 Oct 22;23(10):e30217 [FREE Full text] [doi: 10.2196/30217] [Medline: 34550899]
19. Slobodin O, Plochotnikov I, Cohen IC, Elyashar A, Cohen O, Puzis R. Global and local trends affecting the experience of US and UK healthcare professionals during COVID-19: Twitter text analysis. Int J Environ Res Public Health 2022 Jun 04;19(11):6895 [FREE Full text] [doi: 10.3390/ijerph19116895] [Medline: 35682477]
20. Apisarnthanarak A, Apisarnthanarak P, Siripraparat C, Saengaram P, Leeprechanon N, Weber DJ. Impact of anxiety and fear for COVID-19 toward infection control practices among Thai healthcare workers. Infect Control Hosp Epidemiol 2020 Sep;41(9):1093-1094 [FREE Full text] [doi: 10.1017/ice.2020.280] [Medline: 32507115]
21. Thatrimontrichai A, Weber DJ, Apisarnthanarak A. Mental health among healthcare personnel during COVID-19 in Asia: a systematic review. J Formos Med Assoc 2021 Jun;120(6):1296-1304 [FREE Full text] [doi: 10.1016/j.jfma.2021.01.023] [Medline: 33581962]
22. Adams JG, Walls RM. Supporting the health care workforce during the COVID-19 global epidemic. JAMA 2020 Apr 21;323(15):1439-1440. [doi: 10.1001/jama.2020.3972] [Medline: 32163102]
23. Dagyaran I, Risom SS, Berg SK, Højskov IE, Heiden M, Bernild C, et al. Like soldiers on the front - a qualitative study understanding the frontline healthcare professionals' experience of treating and caring for patients with COVID-19. BMC Health Serv Res 2021 Jul 07;21(1):666 [FREE Full text] [doi: 10.1186/s12913-021-06637-4] [Medline: 34229686]
24. Awano N, Oyama N, Akiyama K, Inomata M, Kuse N, Tone M, et al. Anxiety, depression, and resilience of healthcare workers in Japan during the coronavirus disease 2019 outbreak. Intern Med 2020;59(21):2693-2699 [FREE Full text] [doi: 10.2169/internalmedicine.5694-20] [Medline: 33132305]
25. Gross JJ. Emotion regulation: current status and future prospects. Psychol Inq 2015 Mar 09;26(1):1-26. [doi: 10.1080/1047840X.2014.940781]
26. Varma P, Junge M, Meaklim H, Jackson ML. Younger people are more vulnerable to stress, anxiety and depression during COVID-19 pandemic: a global cross-sectional survey. Prog Neuropsychopharmacol Biol Psychiatry 2021 Jul 13;109:110236 [FREE Full text] [doi: 10.1016/j.pnpbp.2020.110236] [Medline: 33373680]
27. Ingram C, Downey V, Roe M, Chen Y, Archibald M, Kallas KA, et al. COVID-19 prevention and control measures in workplace settings: a rapid review and meta-analysis. Int J Environ Res Public Health 2021 Jul 24;18(15):7847 [FREE Full text] [doi: 10.3390/ijerph18157847] [Medline: 34360142]
28. Denton J, Evans D, Xu Q. Older nurses and midwives in the workplace: a scoping review. Collegian 2021 Apr;28(2):222-229. [doi: 10.1016/j.colegn.2020.06.004]
29. Markowski M, Cleaver K, Weldon SM. An integrative review of the factors influencing older nurses' timing of retirement. J Adv Nurs 2020 Sep;76(9):2266-2285. [doi: 10.1111/jan.14442] [Medline: 32500926]
30. Montayre J, Knaggs G, Harris C, Li W, Tang LM, de Almeida Neto A, et al. What interventions and programmes are available to support older nurses in the workplace? a literature review of available evidence. Int J Nurs Stud 2023 Mar;139:104446 [FREE Full text] [doi: 10.1016/j.ijnurstu.2023.104446] [Medline: 36746012]
31. Twitter usage statistics. Internet Live Stats. URL: https://www.internetlivestats.com/twitter-statistics/ [accessed 2023-10-26]
32. Mature age hub. Department of Employment and Workplace Relations, Australian Government. URL: https://www.dewr.gov.au/mature-age-hub [accessed 2023-10-27]
33. NLTK's list of english stopwords. GitHub. URL: https://gist.github.com/sebleier/554280 [accessed 2023-05-04]

XSL•FO
RenderX

34. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2023-05-04]
35. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014 Presented at: ACL 2014; June 22-27, 2014; Baltimore, MD. [doi: 10.3115/v1/p14-5010]
36. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020 Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-demos.14]
37. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 Presented at: EMNLP 2013; October 18-21, 2013; Washington, DC.
38. Kiritchenko S, Zhu X, Mohammad SM. Sentiment analysis of short informal text. J Artif Intell Res 2014 Aug 20;50:723-762. [doi: 10.1613/jair.4272]
39. Brockmeyer T, Zimmermann J, Kulessa D, Hautzinger M, Bents H, Friederich HC, et al. Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. Front Psychol 2015 Oct 09;6:1564 [FREE Full text] [doi: 10.3389/fpsyg.2015.01564] [Medline: 26500601]
40. Antoniou M, Estival D, Lam-Cassettari C, Li W, Dwyer A, Neto AA. Predicting mental health status in remote and rural farming communities: computational analysis of text-based counseling. JMIR Form Res 2022 Jun 21;6(6):e33036 [FREE Full text] [doi: 10.2196/33036] [Medline: 35727623]
41. Job vacancies, Australia. Australian Bureau of Statistics. 2022 May. URL: https://www.abs.gov.au/statistics/labour/jobs/job-vacancies-australia/may-2022 [accessed 2023-05-04]
42. Heyhoe J, Birks Y, Harrison R, O'Hara JK, Cracknell A, Lawton R. The role of emotion in patient safety: are we brave enough to scratch beneath the surface? J R Soc Med 2016 Feb;109(2):52-58 [FREE Full text] [doi: 10.1177/0141076815620614] [Medline: 26682568]
43. Ojo A, Guntuku SC, Zheng M, Beidas RS, Ranney ML. How health care workers wield influence through Twitter hashtags: retrospective cross-sectional study of the gun violence and COVID-19 public health crises. JMIR Public Health Surveill 2021 Jan 06;7(1):e24562 [FREE Full text] [doi: 10.2196/24562] [Medline: 33315578]
44. Gillespie GL, Grubb PL, Brown K, Boesch MC, Ulrich DL. “Nurses eat their young”: a novel bullying educational program for student nurses. J Nurse Educ Pract 2017;7(7):11-21. [doi: 10.5430/jnep.v7n7p11]
45. Kagan SH, Melendez-Torres GJ. Ageism in nursing. J Nurs Manag 2015 Jul;23(5):644-650. [doi: 10.1111/jonm.12191] [Medline: 24238082]
46. Wilson DM, Errasti-Ibarrondo B, Low G. Where are we now in relation to determining the prevalence of ageism in this era of escalating population ageing? Ageing Res Rev 2019 May;51:78-84. [doi: 10.1016/j.arr.2019.03.001] [Medline: 30858070]
47. Teferra BG, Borwein S, DeSouza DD, Simpson W, Rheault L, Rose J. Acoustic and linguistic features of impromptu speech and their association with anxiety: validation study. JMIR Ment Health 2022 Jul 08;9(7):e36828 [FREE Full text] [doi: 10.2196/36828] [Medline: 35802401]
48. Teferra BG, Rose J. Predicting generalized anxiety disorder from impromptu speech transcripts using context-aware transformer-based neural networks: model evaluation study. JMIR Ment Health 2023 Mar 28;10:e44325 [FREE Full text] [doi: 10.2196/44325] [Medline: 36976636]
49. Loch AA, Lopes-Rocha AC, Ara A, Gondim JM, Cecchi GA, Corcoran CM, et al. Ethical implications of the use of language analysis technologies for the diagnosis and prediction of psychiatric disorders. JMIR Ment Health 2022 Nov 01;9(11):e41014 [FREE Full text] [doi: 10.2196/41014] [Medline: 36318266]
50. Effective risk communication for environment and health: a strategic report on recent trends, theories and concepts. World Health Organization. 2021 Nov 21. URL: https://www.who.int/europe/publications/i/item/WHO-EURO-2021-4208-43967-61972 [accessed 2023-05-12]
51. Li W, Tang L, Montayre J, Harris C, West S, Antoniou M. Investigating health and wellbeing challenges facing an ageing workforce in the construction and nursing industries: Twitter data set. Western Sydney University. 2023. URL: https://research-data.westernsydney.edu.au/published/da3f70c07d1511eea5c4c9c26f8c077b/ [accessed 2024-05-21]

Abbreviations

LIWC: Linguistic Inquiry and Word Count
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 29.05.23; peer-reviewed by R Fang, Y Chen; comments to author 18.10.23; revised version received 07.11.23; accepted 09.02.24; published 05.06.24.

Please cite as:
Li W, Tang LM, Montayre J, Harris CB, West S, Antoniou M
Investigating Health and Well-Being Challenges Faced by an Aging Workforce in the Construction and Nursing Industries: Computational Linguistic Analysis of Twitter Data
J Med Internet Res 2024;26:e49450
URL: https://www.jmir.org/2024/1/e49450
doi:10.2196/49450
PMID:38838308

©Weicong Li, Liyaning Maggie Tang, Jed Montayre, Celia B Harris, Sancia West, Mark Antoniou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact Analysis of COVID-19 Pandemic on Hospital Reviews on Dianping Website in Shanghai, China: Empirical Study

Weixue Huo[1*], Dr med; Mengwei He[1*], Dr med; Zhaoxiang Zeng[2], Dr med; Xianhao Bao[1], Dr med; Ye Lu[1], Dr med; Wen Tian[1], Dr med; Jiaxuan Feng[3], Dr med; Rui Feng[1], Dr med

[1]Department of Vascular Surgery, Shanghai General Hospital, Shanghai Jiaotong University, Shanghai, China

[2]Department of Vascular Surgery, Changhai Hospital, Navy Medical University, Shanghai, China

[3]Vascular Surgery Department, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Rui Feng, Dr med
Department of Vascular Surgery
Shanghai General Hospital
Shanghai Jiaotong University
650 New Song Jiang Road
Songjiang District
Shanghai, 201600
China
Phone: 86 19174227674
Email: rui.feng@shgh.cn

Abstract

Background: In the era of the internet, individuals have increasingly accustomed themselves to gathering necessary information and expressing their opinions on public web-based platforms. The health care sector is no exception, as these comments, to a certain extent, influence people's health care decisions. During the onset of the COVID-19 pandemic, how the medical experience of Chinese patients and their evaluations of hospitals have changed remains to be studied. Therefore, we plan to collect patient medical visit data from the internet to reflect the current status of medical relationships under specific circumstances.

Objective: This study aims to explore the differences in patient comments across various stages (during, before, and after) of the COVID-19 pandemic, as well as among different types of hospitals (children's hospitals, maternity hospitals, and tumor hospitals). Additionally, by leveraging ChatGPT (OpenAI), the study categorizes the elements of negative hospital evaluations. An analysis is conducted on the acquired data, and potential solutions that could improve patient satisfaction are proposed. This study is intended to assist hospital managers in providing a better experience for patients who are seeking care amid an emergent public health crisis.

Methods: Selecting the top 50 comprehensive hospitals nationwide and the top specialized hospitals (children's hospitals, tumor hospitals, and maternity hospitals), we collected patient reviews from these hospitals on the Dianping website. Using ChatGPT, we classified the content of negative reviews. Additionally, we conducted statistical analysis using SPSS (IBM Corp) to examine the scoring and composition of negative evaluations.

Results: A total of 30,317 pieces of effective comment information were collected from January 1, 2018, to August 15, 2023, including 7696 pieces of negative comment information. Manual inspection results indicated that ChatGPT had an accuracy rate of 92.05%. The F1-score was 0.914. The analysis of this data revealed a significant correlation between the comments and ratings received by hospitals during the pandemic. Overall, there was a significant increase in average comment scores during the outbreak ($P<.001$). Furthermore, there were notable differences in the composition of negative comments among different types of hospitals ($P<.001$). Children's hospitals received sensitive feedback regarding waiting times and treatment effectiveness, while patients at maternity hospitals showed a greater concern for the attitude of health care providers. Patients at tumor hospitals expressed a desire for timely examinations and treatments, especially during the pandemic period.

Conclusions: The COVID-19 pandemic had some association with patient comment scores. There were variations in the scores and content of comments among different types of specialized hospitals. Using ChatGPT to analyze patient comment content represents an innovative approach for statistically assessing factors contributing to patient dissatisfaction. The findings of this

XSL•FO
RenderX

study could provide valuable insights for hospital administrators to foster more harmonious physician-patient relationships and enhance hospital performance during public health emergencies.

(J Med Internet Res 2024;26:e52992) doi:10.2196/52992

KEYWORDS

patient satisfaction; physician-patient relationship; ChatGPT; patient concern; COVID-19

Introduction

The Tense Relationship Between Chinese Patients and Health Care Service Providers

In recent years, medical issues have been a highly focused and hot topic in China [1,2]. In 2016, 2017, and 2019, China experienced severe conflicts between patients and health care providers, including extreme cases of violence where doctors were harmed or even killed by patients or their family members [3,4]. The occurrence of this series of events highlights the challenging state of patient-health care service provider relationships in China. Moreover, the significance of these relationships cannot be understated. On the one hand, doctor-patient relationships have a profound impact. They influence patients' satisfaction with health care services and their understanding of medical information. Additionally, these relationships affect patients' behavior in response to illness and their adherence to prescribed medication, ultimately impacting their overall quality of life and health outcomes [5].

Therefore, it is important for us to improve the current state of patient-health care service provider relationships. Against this backdrop, the COVID-19 pandemic has further exacerbated the complexity of this issue [6]. The outbreak of SARS-CoV-2, causing COVID-19, has been an unprecedented event affecting all sectors of society, particularly the health care sector. The spread of the pandemic has resulted in a strained allocation of medical resources, potentially leading to shortages in certain health care services. Hospitals are required to prioritize patients based on varying levels of urgency and may need to implement isolation measures. These measures have the potential to impact the mental well-being of both health care workers and patients to varying degrees [7,8]. These factors could potentially lead to an even more strained relationship between Chinese patients and health care service providers.

Web-Based Comments and ChatGPT

In the era of the internet, individuals are becoming increasingly accustomed to expressing their viewpoints on public web-based platforms and using public comments as a means to inform their own rational decision-making processes [9,10]. The field of health care is no exception, as previous research on quality management has indicated that, due to factors such as complex complaint processes and a lack of transparency in the resolution, only a third of patients opt to make a formal complaint when they encounter unsatisfactory service [11-13]. However, with the increasing prevalence of internet technology, public web-based platforms have provided patients with a more efficient platform to express their personal opinions. When faced with unsatisfactory service, the option of web-based reviews provides an alternative approach for patients to address their concerns. Through internet platforms, they can discuss their negative medical experiences and share them with others [14]. Research has demonstrated that the public dissemination of negative reviews has the potential to undermine or tarnish the image of physicians and amplify patient dissatisfaction with doctors or hospitals [15]. Undoubtedly, these comments also have the potential to exert an underlying influence on subsequent patients seeking medical care. Hence, the identification and documentation of negative patient comments, followed by targeted improvements, are of paramount importance in enhancing the quality of health care services and fostering positive physician-patient relationships [16]. Over the past year, considerable attention has been drawn to a novel AI model called ChatGPT due to its remarkable proficiency in accomplishing varied natural language tasks [17-21]. ChatGPT is a general large language model recently developed by OpenAI. Many studies have shown that ChatGPT has demonstrated impressive analytical and observational abilities, approaching or surpassing the threshold required to simulate the United States Medical Licensing Examination multiple-choice questions [17,22-25]. Leveraging the analytical capabilities of ChatGPT, this study aims to conduct content analysis and organize publicly available negative reviews to explore the specific reasons and composition of patients' negative feedback.

Current research findings indicate that during the COVID-19 pandemic, the rating of doctor-patient relationships among patients (mean 4.18, SD 0.51) has increased compared to the prepandemic period (mean 3.86, SD 0.67), suggesting an improvement in the doctor-patient relationship. However, this study did not differentiate between specific types of hospitals. General hospitals and specialized hospitals may exhibit differences in doctor-patient relationships due to variances in the types of patients they treat. Additionally, when collecting rating data before the pandemic, this study relied on participants' retrospective memory for scoring, which could introduce certain biases and deviate from the actual circumstances. Moreover, a patient's experience during a hospital visit is composed of multiple factors; in addition to doctor-patient relationships, the quality of service provided by nurses and other hospital staff should also be taken into consideration. Even the cleanliness of the hospital environment and the quality of service offered by automated service equipment can affect the patient's experience during a hospital visit [26,27].

This study goes beyond the conventional doctor-patient relationship. Instead, it starts by exploring and analyzing the patients' public comments about the hospitals. By examining the assessments published on public platforms, we aim to investigate the impact of COVID-19 factors and hospital categorization differences on the health care experiences of

XSL•FO
RenderX

patients in China, taking into account the perspectives of the patients. The research questions posed in this survey encompass three aspects. Research question (RQ) 1: During the pandemic, how have patient ratings for hospitals changed? RQ 2: Since the outbreak of the pandemic, what changes have occurred in the composition of negative reviews from patients about hospitals? RQ 3: What kind of negative feedback have different types of hospitals—such as children's, maternity, and tumor hospitals—received, and what are the characteristics of the negative reviews for these various institutions?

Methods

Sample and Data Collection

This study relied on the open business review platform website Dianping [28]. Established in Shanghai in April 2003, this platform is the earliest independent third-party review website globally, with an active user base of 40.987 million people. The sample for this study consisted of a total of 88 hospitals, including those ranked in the top 50 nationally by Fudan University and those nominated in the Fudan University Specialist Hospital rankings for children's hospitals, maternity hospitals, and tumor hospitals. The collection of comment information was limited to the period from January 1, 2018, to August 15, 2023. The collected content included the hospital's region, type, name, date, score (ranging from 1 to 5), content, and user ID (Figure 1A). In order to ensure patient privacy, the user IDs were only used to judge whether the comment information met the requirements and do not appear in this paper.

XSL•FO
RenderX

Figure 1. (A) Negative comment samples from the Dianping website. (B) ChatGPT handling demonstration.

Comment Inclusion Criteria

It was necessary to conduct preliminary screening, labeling, and exclusion of meaningless comment content to ensure the quality of the included comment information. Comment information with a total character count less than 10 and comment information with fewer than 10 different Chinese characters was directly deleted. Comment information corresponding to user IDs that appeared more than 5 times underwent manual review to confirm its authenticity and validity before being included.

Processing and Analysis of Comment Scores

Based on the comment dates, the comment information was categorized as follows: (1) before the pandemic (January 1, 2018, to April 30, 2018), (2) during the pandemic (January 1, 2020, to April 30, 2020), and (3) after the pandemic (January 1, 2023, to April 30, 2023). The Kruskal-Wallis test was used to compare the comment score data among the three categories: (1) before the pandemic, (2) during the pandemic, and (3) after the pandemic. Subsequently, pairwise Mann-Whitney *U* tests were conducted to compare the 3 groups, and the statistical

XSL•FO
RenderX

analysis results were recorded. The significance level was set at $P<.05$ (2-tailed).

The comment information was categorized based on hospital location: (1) Shanghai region and (2) other regions. A pairwise Mann-Whitney *U* test was used to compare the comment score data between hospitals in the Shanghai region and hospitals in other regions, and the statistical analysis results were recorded. The significance level was set at $P<.05$ (2-tailed).

The comment information was also categorized based on hospital type: (1) general hospitals, (2) children's hospitals, (3) maternity hospitals, and (4) tumor hospitals. A Kruskal-Wallis test was used to compare the comment score data among children's hospitals, maternity hospitals, and tumor hospitals. Subsequently, pairwise Mann-Whitney *U* tests were conducted to compare the 3 groups, and the statistical analysis results were recorded. The significance level was set at $P<.05$ (2-tailed).

Processing and Analysis of Comment Content

Three clinical experts (RF, JF, and ZZ) were initially assigned to randomly review 500 comments with scores less than or equal to 3. After a thorough discussion among the experts, a compilation of 10 reasons was derived. This compilation was named the hospital-patient factor (HPF).

Using ChatGPT, a comprehensive analysis and extraction of all comments with scores less than or equal to 3 was conducted. For each negative comment, up to 2 reasons behind the low scores were extracted from the patient's feedback. The instructional prompts provided to ChatGPT were as follows (Figure 1B):

> *Hello, I have a collection of patient dissatisfaction comment tags regarding the hospital. I will provide you with a patient's comment and you will assign appropriate tags indicating the reasons for their dissatisfaction. Please note that you can only select from a predefined set of ten tags. (HPF-A: Long waiting times ,HPF-B: Dissatisfaction with the hospital's self-service equipment ,HPF-C: Dissatisfaction with the treatment effect ,HPF-D: Poor attitude of nurses ,HPF-E: Cumbersome medical treatment process ,HPF-F: Concerns about overtreatment ,HPF-G: Perception of understaffed medical personnel ,HPF-H: Poor hospital environment ,HPF-I: Poor attitude of doctors ,HPF-J: Poor attitude of other hospital staff) Please select the two most relevant tags. And please respond with only the tags, without any additional content. Here is the comment: "+ [comment content]."*

After processing all the data, comments that could not be tagged with specific HPF labels by ChatGPT were deleted. A systematic sampling survey of the ChatGPT output was conducted. Out of every 100 comments, 10 were extracted and given to 3 clinical experts for manual review in order to assess the authenticity and validity of the ChatGPT results.

Ethical Considerations

The data collected for this study consisted of deidentified information that has been publicly released on the web, adhering to ethical standards while fully respecting and safeguarding the rights of all participants. Furthermore, this study underwent a rigorous review and received approval from the Ethics Committee of Shanghai General Hospital (20240417094101248).

Results

Overview

The general flow of comment information from collection to processing is shown in Figure 2. In this study, a total of 32,085 pieces of comment data were gathered, and after careful screening, 1768 comments that did not meet the inclusion criteria were excluded. This resulted in a final collection of 30,317 valid comment records. (Multimedia Appendix 1). Among them, 7696 were categorized as negative comments (with scores less than or equal to 3). After undergoing processing through ChatGPT, the negative comments yielded a total of 15,588 HPF tags, and after further scrutiny, 7 invalid tags were removed.

XSL•FO
RenderX

Figure 2. Flowchart depicting the process of data collection and organization in this study. HPF: hospital-patient factor.

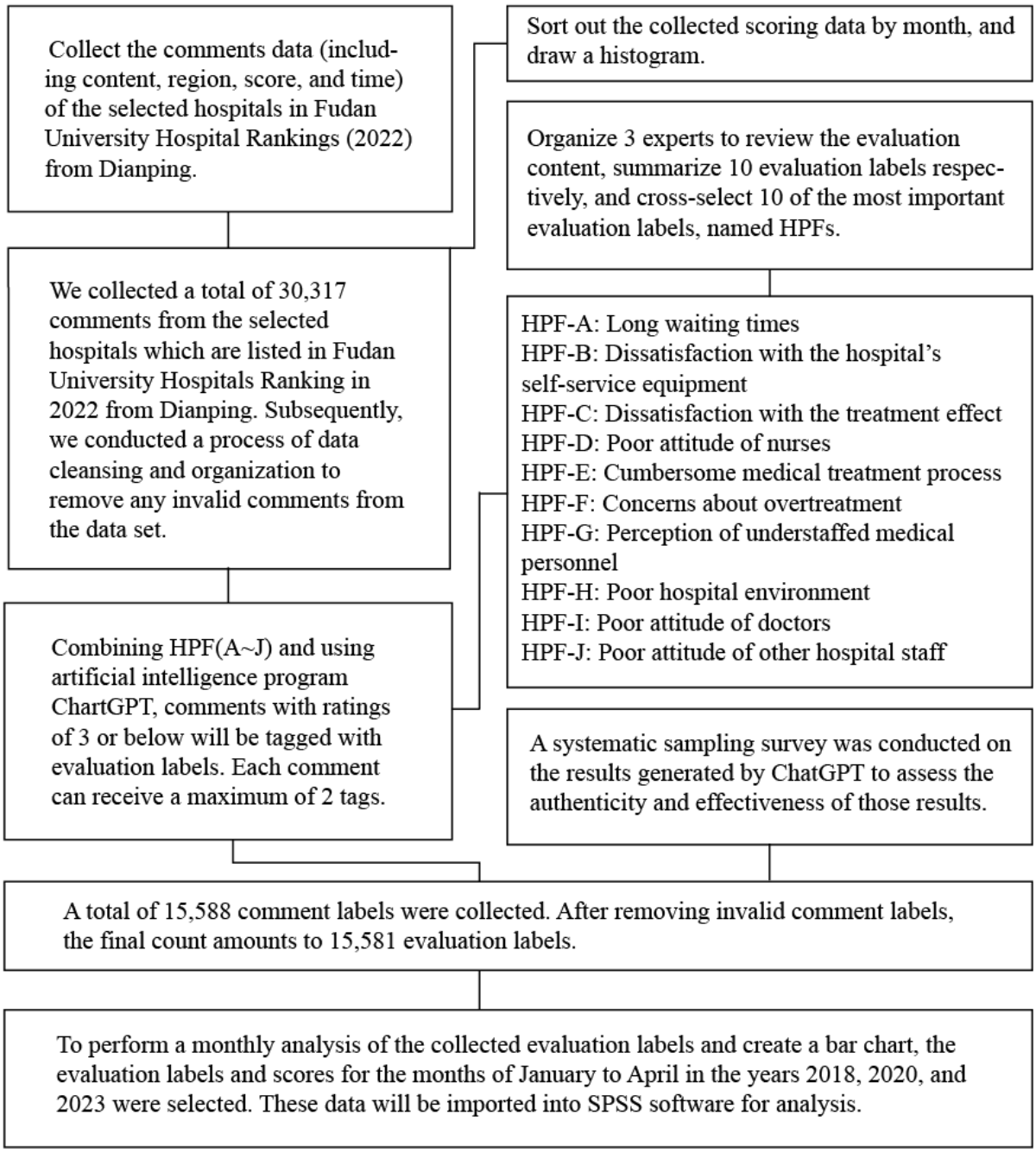

A systematic random sampling survey was conducted on the remaining output results to confirm their authenticity and effectiveness. A total of 1519 labels were manually reviewed, and the results show that the accuracy of ChatGPT's identifications reached 92.05%, and a confusion matrix for ChatGPT's automatic classification assessment was plotted (Figure 3). Based on this, the performance of GPT in text classification was evaluated using manually encoded data. Among them, the macro F_1-score was 0.914 (Table 1) [29,30]. Ultimately, 15,581 valid HPF tags were obtained. An overview of the collected data is presented in Table 2.

XSL•FO
RenderX

Figure 3. Confusion matrix: manual inspection of ChatGPT's automatic classification results. HPF: hospital-patient factor.

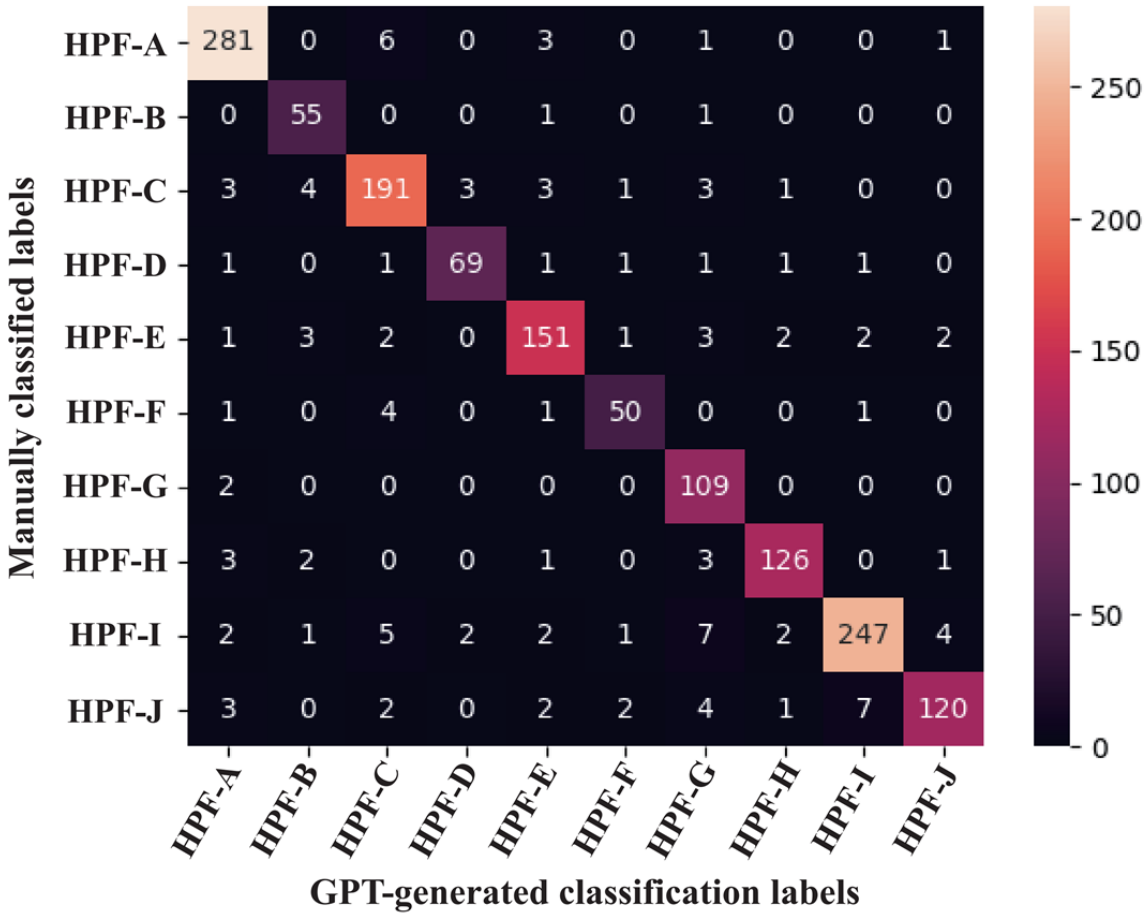

Table 1. Results of the GPT text classification performance evaluation.

Categorization of HPF[a]	Precision	Recall	F_1-score
Single HPF			
HPF-A	0.946	0.962	0.954
HPF-B	0.846	0.965	0.902
HPF-C	0.905	0.914	0.909
HPF-D	0.932	0.907	0.921
HPF-E	0.915	0.904	0.909
HPF-F	0.893	0.877	0.885
HPF-G	0.826	0.982	0.897
HPF-H	0.947	0.926	0.937
HPF-I	0.957	0.905	0.931
HPF-J	0.938	0.851	0.892
Aggregate			
Macro	0.911	0.919	0.914
Micro	0.921	0.921	0.921

[a]HPF: hospital-patient factor.

XSL•FO
RenderX

Table 2. Overview of hospital grouping, number of hospitals and comment count.

Categorization of hospitals	Number of hospitals (n=88), n (%)	Number of comments (n=30,317), n (%)
Area		
All hospitals	88 (100)	30,317 (100)
Hospitals in Shanghai	19 (22)	14,511 (47,9)
Hospitals in other areas	69 (78)	15,806 (52.1)
Period		
Before the pandemic (January 2018 to April 2018)	88 (100)	1904 (6.3)
During the outbreak of COVID-19 (January 2020 to April 2020)	88 (100)	1949 (6.3)
After the pandemic (January 2023 to April 2023)	88 (100)	1016 (3.4)
Type of hospital		
Children's hospital	17 (19)	4910 (16.2)
Maternity hospital	13 (15)	6685 (22.1)
Tumor hospital	13 (15)	978 (3.2)

Trend of Comment Scores in all Hospitals

The results indicate that the comment scores of all hospitals experienced a gradual increase from January 2018 to January 2019 and remained at a high level from February 2019 to November 2019. After the outbreak of the pandemic in December 2019, the comment scores of all hospitals rose again and reached a peak in February 2020. Subsequently, from March 2020 to March 2022, the comment scores of all hospitals gradually declined over a period of 2 years, reaching the lowest value since 2018 in March 2022. However, from April 2022 to January 2023, the comment scores experienced a sudden increase followed by a continuous decline. In October 2022, the comment scores sharply rose again, only to plummet after reaching their peak in January 2023. During the period from April 2022 to January 2023, the comment scores of all hospitals exhibited 2 drastic fluctuations, forming an M-shaped waveform. Subsequently, from January 2023 to August 2023, the comment scores remained at a lower level (Figure 4A).

XSL•FO
RenderX

Figure 4. (A)-(C): Trend chart of score changes for all hospitals, hospitals in Shanghai, and hospitals in other areas from January 2018 to August 2023. (D) and (E): The M-shaped fluctuation graph of hospitals in Shanghai and hospitals in other areas.

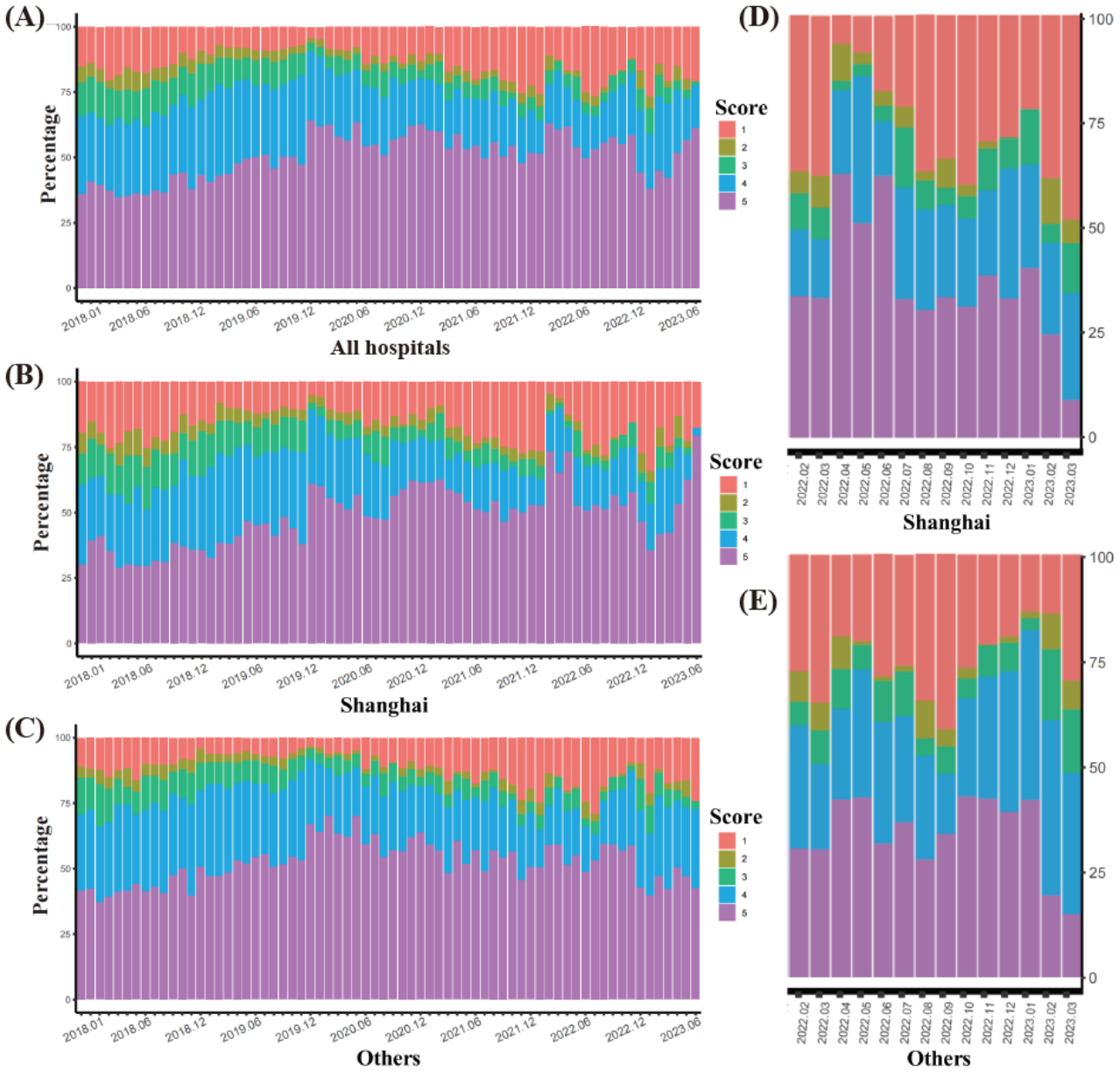

The Comment Scores Trend of Hospitals in Shanghai and Other Regions

The comment score trends of hospitals in Shanghai and other regions follow a similar pattern to the average score changes from all hospitals, encompassing a sequence of "gradual increase-maintenance-sudden increase-decline–M-shaped waveform-low score maintenance" (Figures 4B and 4C). During the M-shaped waveform period from April 2022 to January 2023, the M-shaped waveform in Shanghai was steeper and experienced more drastic fluctuations compared to other regions (Figures 4D and 4E). We conducted a separate analysis of the comment scores and content of comments for hospitals in Shanghai and other regions during the M-shaped waveform period (April 2022 to January 2023). We found that there were significant differences in the comment scores between hospitals in Shanghai and other regions during this period (P=.03). However, there was no significant difference in the composition of the comment content (P=.59; Table 3; Figures S1, S2, and S3 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 3. Score forms for hospitals in the Shanghai region and other areas.

Characteristics	Values
Shanghai (score), mean (SD)	4.027 (1.49)
Hospitals with each rank, n (%)	
1	74 (15)
2	19 (4)
3	24 (5)
4	67 (14)
5	297 (62)
Total	481 (100)
Others (score), mean (SD)	3.874 (1.54)
Hospitals with each rank, n (%)	
1	140 (19)
2	21 (3)
3	46 (6)
4	132 (18)
5	413 (54)
Total	752 (100)
P value	.03

The Comparison Between the Period Before and After the Outbreak of the Pandemic and During its Peak

Given the significant impact of the sudden outbreak of the pandemic on hospitals, we selected three periods for comparison. (1) Period A, before the outbreak (January 1, 2018, to April 30, 2018); (2) period B, during the outbreak (January 1, 2020, to April 30, 2020); and (3) period C, after the outbreak (January 1, 2023, to April 30, 2023). We compared the comment scores and analyzed the composition of negative feedback during these periods. There were clear significant differences in the scores between periods A, B, and C ($P<.001$). Pairwise Mann-Whitney *U* tests further confirmed significant differences among all groups ($P=.007$). The highest average score was observed in period B, at 4.274. Period A had an average score of 3.647, while period C had an average score of 3.737 (Figures 5A-5C, 5K, and 5L).

XSL•FO
RenderX

Figure 5. (A)-(C) Hospitals' score diagrams for the months of January to April in 2018, 2020, and 2023. (D)-(F) Composition chart of hospital evaluation content for January-April in 2018, 2020, and 2023. (G) and (H) Comparison chart of hospital evaluation content composition for January to April in 2018 and 2020. (I) and (J) Comparison chart of hospital evaluation content composition for January to April in 2020 and 2023. (K) Average hospital scores for January to April in 2018, 2020, and 2023. (L) *P* value. HPF: hospital-patient factor.

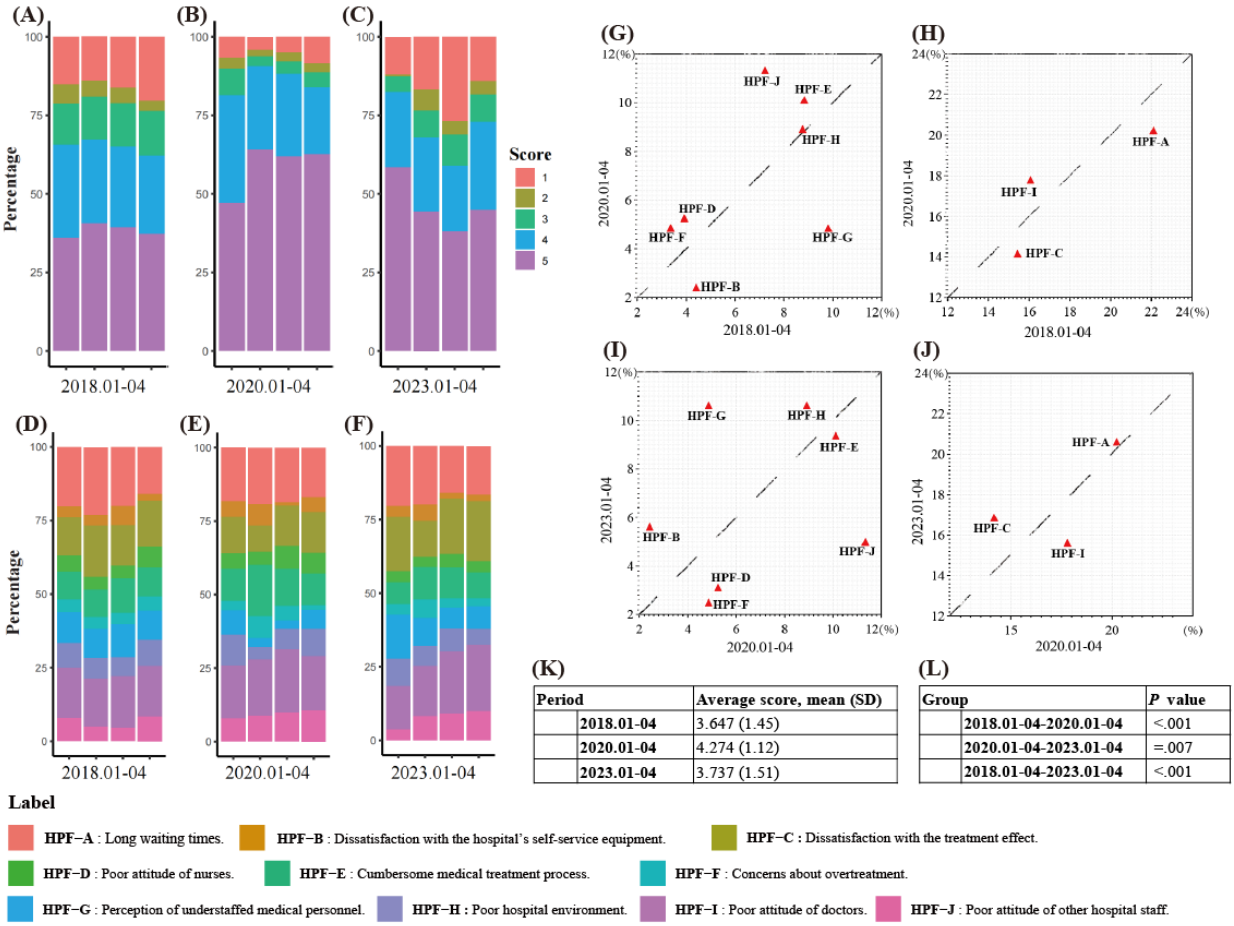

We conducted an analysis of the composition of negative feedback during periods A, B, and C, revealing significant differences among the 3 groups (*P*=.03). Pairwise comparisons using the chi-square test for 2 proportions showed no significant difference between periods A and C (*P*=.20). However, there was a relatively significant difference between periods A and B (*P*=.04), as well as between periods C and B (*P*=.07; Figures 5D-5F; Figure S4 in Multimedia Appendix 2). Period B, which corresponds to the outbreak period, exhibited a relatively higher proportion of negative comments categorized as HPF-D, E, F, I, and J (Figures 5G-5J). During periods A and C, which correspond to nonoutbreak periods, there was a relatively higher proportion of negative comments classified as HPF-B, G, and C.

Comparison of Children's Hospitals, Maternity Hospitals, and Tumor Hospitals

We compared the comment scores of these children's hospitals, maternity hospitals, and tumor hospitals and found significant differences in the comment scores among these 3 distinct types of specialized medical institutions (*P*<.001). Furthermore, pairwise Wilcoxon rank-sum tests were conducted for each pair of data sets, demonstrating significant differences between any 2 groups (*P*<.001). Maternity hospitals had the highest average score, at 4.121. Children's hospitals were next, with an average score of 3.812. Tumor hospitals had the lowest average score, at 3.572. Additionally, unlike hospitals of other types, tumor hospitals received lower comment scores during the outbreak compared to before the outbreak. On the other hand, children's hospitals and maternity hospitals received the highest comment scores during the outbreak (Table 4; Figures S5, S6, and S7 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 4. Score forms for children's hospitals, maternity hospitals, and tumor hospitals.

Characteristics	Scores, mean (SD)
Before the pandemic (January 2018 to April 2018)	
Children's hospital	3.847 (1.48)
Maternity hospital	3.958 (1.31)
Tumor hospital	3.797 (1.42)
During the outbreak of COVID-19 (January 2020 to April 2020)	
Children's hospital	4.003 (1.35)
Maternity hospital	4.461 (1.19)
Tumor hospital	3.404 (1.61)
After the pandemic (January 2023 to April 2023)	
Children's hospital	3.757 (1.49)
Maternity hospital	4.241 (1.34)
Tumor hospital	3.394 (1.51)

We conducted an analysis of the composition of negative comments for children's hospitals maternity hospitals, and tumor hospitals, and found significant differences in the composition of comments among the 3 groups (P<.001). Further pairwise comparisons using chi-square tests revealed significant differences in the composition of negative comments between children's hospitals and maternity hospitals, as well as between children's hospitals and tumor hospitals (P<.001; Figures S8, S9, S10, and S11 in Multimedia Appendix 2).

Compared to other hospitals, maternity hospitals tended to have a relatively higher proportion of negative comments related to HPF-D, J, and I, while children's hospitals tended to have a relatively higher proportion of negative comments related to HPF-A, C, F, and G (Figure 6).

Figure 6. (A) and (B) Comparison chart of the composition of evaluation content between children's hospitals and maternity hospitals. HPF: hospital-patient factor.

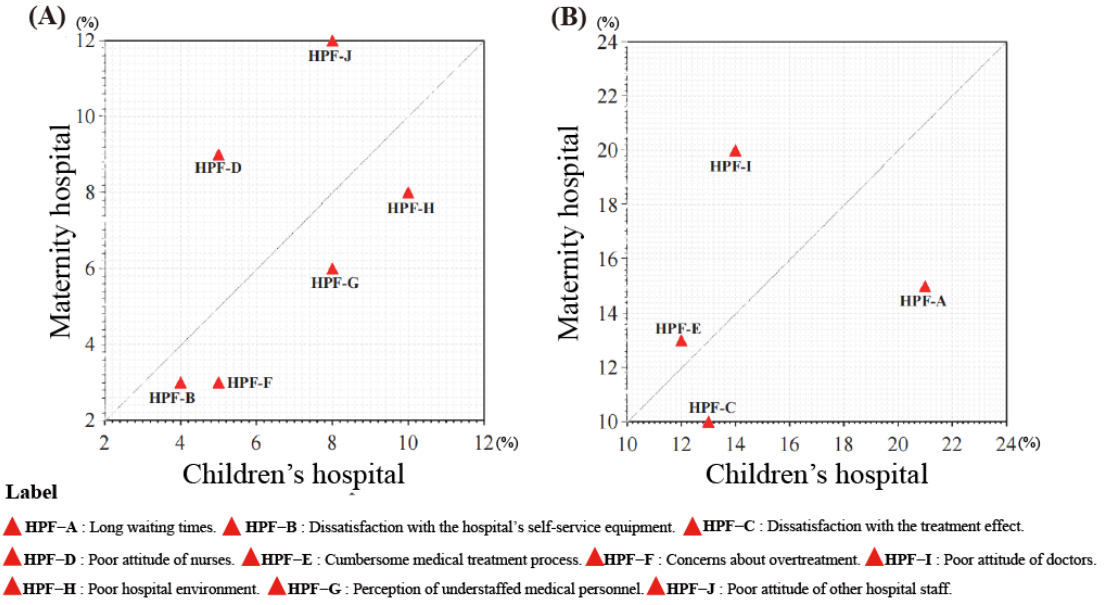

Discussion

Principal Findings

In this study, we compared the trends in comment scores and composition of comments exhibited by different hospitals in different regions and of different types during the periods before, during, and after the pandemic. We found that at the early stages of the COVID-19 outbreak, there was a certain increase in the patient satisfaction scores for hospitals (average score: 4.274; P<.001). However, as the pandemic continued, there was a gradual decline in the hospitals' satisfaction scores from patients (average score: 3.737; P<.001). Moreover, during the pandemic, patients were more likely to express dissatisfaction with the attitudes of doctors, nurses, and other hospital staff.

In addition, in our survey of specialized hospitals, we discovered that, unlike other hospitals, patients at tumor hospitals tended to give lower scores during the pandemic (average score: 3.404; P<.001). This may be related to the anxiety that patients with

XSL•FO
RenderX

cancer feel about their treatments. The negative feedback provided by patients at maternity hospitals and children's hospitals also had their own distinctive characteristics.

The Comment Score Trends for all Hospitals, Hospitals in Shanghai, and Hospitals in Other Regions

Our research demonstrates that during the COVID-19 pandemic in China, which spanned a 3-year period, patient comments on hospitals exhibited distinct characteristics across different stages. Existing findings indicate that during the peak of the outbreak (January to June 2020), the physician-patient relationship showed signs of improvement, reaching its optimal state in February 2020. Positive media coverage of health care workers, the adoption of telemedicine, and supportive health care policies had a significant and positive impact on the physician-patient relationship during the pandemic [6,31-33]. Physicians have perceived an increased level of trust from patients in their clinical practice compared to the past. For instance, patients voluntarily turn their heads away when sneezing and display greater respect toward doctors during conversations.

However, the general public does not share the belief that the physician-patient relationship has improved during the pandemic. According to a survey, 61% of the population expressed opposition to the idea of any improvement in the physician-patient relationship during this time [34]. The variations in attitudes and behaviors can be explained by several factors. First, the Chinese government has undertaken the entire cost of treating patients with COVID-19, relieving them of financial concerns and subsequently fostering improved attitudes toward seeking medical care and willingness to undergo treatment. This has contributed to a reduction in the sense of mistrust between doctors and patients [35,36]. Health care providers may perceive an improvement in the physician-patient relationship due to these factors. However, the general public, which is not affected by COVID-19, does not benefit from waived medical expenses. On the contrary, the scarcity of medical resources during the pandemic makes it even more challenging for the general population to seek medical care.

Positive media coverage and the positive impacts of health care policies during the pandemic may not necessarily be sustained in the long run [37,38]. Over a 2-year period from March 2020 to March 2022, patient ratings of hospitals gradually declined, reaching their lowest point in March 2022. Multiple factors contributed to this decline. According to Roubille et al [39], as the duration of the pandemic became prolonged, the confidence of the general public may have waned or suffered, ultimately leading to a deterioration in physician-patient relationships due to the adverse effects of COVID-19.

M-Shaped Pattern

As shown in Figure 4A, during the period of April 2022 to January 2023, the comment scores of all hospitals experienced two significant fluctuations, forming a distinctive M-shaped pattern. Analyzing the timing, research indicates that the formation of the 2 peaks of the M-shaped pattern correlates with the outbreak in Shanghai in the first half of 2022 and the subsequent comprehensive relaxation of COVID-19 control measures in China toward the end of 2022. We have observed that these 2 incidents have caused varying degrees of strain on medical resources [40], while simultaneously leading to higher comment scores from patients toward hospitals. This finding aligns with previous research in this area [6,31-33].

As shown in Figures 4D and 4E, the M-shaped waveform formed during the outbreak in Shanghai from April to July 2022 was steeper, with the highest comment scores reached in April 2022. In contrast to the Shanghai region, hospitals in other areas exhibited a smoother M-shaped waveform, with scores lower than the peak scores in Shanghai. During this period, both Shanghai and other regions implemented the same health care policies, and the media coverage received by the public was consistent. In contrast, the situation in Shanghai is characterized by a more severe outbreak and a greater strain on medical resources.

Based on these observations, we speculate that, under similar circumstances, the scarcity of medical resources and the severity of the epidemic may lower patients' expectations for medical outcomes. Consequently, they would value the precious medical resources even more, leading to an increase in comment scores and a decrease in negative reviews.

Comparison Between the Outbreak and Before and After the COVID-19 Pandemic

We selected data from January to April 2020, as well as data from January to April 2018 and 2023, representing the periods of the pandemic outbreak, prepandemic, and postpandemic, respectively. The results reveal that hospitals received the highest comment scores during the outbreak, with an average score of 4.274. In comparison, the average scores for the prepandemic and postpandemic periods were 3.647 and 3.737, respectively. We have already discussed these findings in earlier sections (Figures 5A-5C, 5K, and 5L). Furthermore, through further analysis of the composition of negative comments during these 3 periods, we found certain differences between the prepandemic and pandemic outbreak periods, as well as the postpandemic and pandemic outbreak periods. During the pandemic outbreak, there was a relatively higher proportion of HPF-D, E, F, I, and J in negative comments, while HPF-B, G, and C showed a relatively lower proportion compared to the prepandemic and postpandemic periods (Figures 5D-5J; Figure S4 in Multimedia Appendix 2).

HPF-D, HPF-I, and HPF-J, respectively, represent patient dissatisfaction with nurses, physicians, and other hospital staff. The highest proportion of negative comments was "dissatisfaction with physician attitude." For a long time, "dissatisfaction with physician attitude" has been a significant factor in the occurrence of doctor-patient disputes. These disputes not only directly lead to conflicts between doctors and patients but also erode the level of trust between them, resulting in a stalemate in the diagnosis and treatment process and affecting the quality and standard of health care [41,42]. The reasons for this situation are multifaceted, encompassing significant information asymmetry between doctors and patients, the authoritative status of doctors, poor doctor-patient communication, and a lack of empathy toward patients [43]. During the pandemic, the limited availability of medical resources has led to reduced communication time between

XSL•FO
RenderX

doctors and each patient. Consequently, effective communication skills have become crucial in such circumstances. Numerous studies have demonstrated that fostering a positive doctor-patient communication environment can enhance patient trust, contribute to superior treatment experiences, and ultimately improve patient satisfaction and treatment outcomes [42-44]. Hospitals should also strengthen training in clinical and communication skills [44]. When dealing with critically ill patients, physicians must promptly communicate the prognosis to the patients. It is important to alleviate any unrealistic expectations the patients may have regarding the outcome and share the decision-making process with them in order to prevent medical disputes and the deterioration of the doctor-patient relationship [45,46].

Similarly, nurses and other hospital staff members are crucial components of the health care system. During the pandemic, there has been a notable increase in patients' dissatisfaction with nurses and other staff members. Hospital administrators should prioritize their care and training to enhance service quality. Patient-centered nursing care can provide a better treatment experience for patients [47]. Moreover, there has been an increase in HPF-E during the pandemic, which emphasizes the need for hospitals to optimize complex health care processes to minimize negative patient experiences. This result reminds us that simplifying complicated medical procedures during a pandemic may help reduce the generation of negative emotions in patients.

Differences Between Children's Hospitals, Maternity Hospitals, and Tumor Hospitals

We selected 2 time points, December 2019, and November 2022, as reference periods to categorize the comment data into 3 stages: prepandemic, during the pandemic, and postpandemic. It is noteworthy that while most hospitals experienced an improvement in their comment scores during the pandemic, the scores for cancer hospitals instead showed a decline. From a rating of 3.797 before the pandemic, it decreased to 3.404 during the pandemic (Table 4; Figure S7 in Multimedia Appendix 2). This anomalous phenomenon may be attributed to the unique characteristics of patients with cancer. Compared to ordinary patients, patients with cancer often bear a heavier psychological burden, manifested through severe depressive and anxiety symptoms [48].

Furthermore, during the pandemic, the implementation of containment policies and the scarcity of health care resources may have significantly disrupted the diagnosis and treatment of patients with cancer. Research by Terashima et al [49] demonstrates that the COVID-19 pandemic, with its strain on health care resources, can result in delayed diagnoses, leaving patients with cancer already in advanced stages by the time they have the opportunity for diagnosis. The pandemic may also lead to treatment delays for patients with cancer, such as postponed surgeries due to circumstances beyond their control [49-51]. Given the circumstances described above, the impact of encountering the COVID-19 pandemic can be disastrous for patients with cancer who are in urgent need of diagnosis and treatment. Research by Ye et al [52] has elucidated that during the COVID-19 outbreak, patients with cancer may experience heightened psychological distress due to treatment delays. This suggests that the consequences for patients with cancer, when faced with the dual challenge of seeking timely diagnosis and navigating the pandemic, can be profound and deeply distressing [52,53]. This could be a contributing factor to the lower comment scores of cancer hospitals during the pandemic.

Therefore, for cancer hospitals, ensuring timely diagnosis and treatment for patients with cancer in the face of potential future outbreaks is crucial for enhancing health care service quality.

When analyzing the composition of negative comments, children's hospitals and maternity hospitals exhibit prominent characteristics. The categories of HPF-A, C, F, and G in negative comments are relatively high in children's hospitals (Figures 6A and 6B). These categories respectively represent "excessive waiting time," "dissatisfaction with treatment effectiveness," "perceived excessive medical procedures and examinations," and "insufficient health care staff." On one hand, the negative comments of "excessive waiting time" and "insufficient health care staff" authentically reflect the current shortage of pediatric doctors in China [54,55]. Based on the basic data from the Chinese Pediatric Medical Resources White Paper, China has approximately 260 million children aged between 0 and 14 years, but currently, there are only around 100,000 pediatric doctors nationwide. On average, each pediatric doctor is responsible for the care of 2000 children. This indicates a shortage of at least 200,000 pediatric doctors in China [56]. The shortage of pediatric doctors is a significant factor contributing to negative comments about children's hospitals. Furthermore, the negative comments of "dissatisfaction with treatment effectiveness" and "perceived excessive medical procedures and examinations" reflect the anxiety and lack of trust that parents of sick children commonly experience toward doctors [55].

Given the aforementioned context, addressing the shortage of pediatric doctors and alleviating the negative emotions of parents of sick children have emerged as 2 crucial focal points for enhancing the quality of health care services in children's hospitals. Focusing on the aforementioned dimensions, hospital administrators should prioritize training programs for pediatric doctors. Additionally, it is essential to enhance training for pediatric doctors and improve their salary levels without compromising their passion and dedication to caring for children [56,57].

When examining the negative comments of obstetric and gynecological hospitals, it becomes apparent that the proportions of HPF-D, J, and I are relatively high (Figures 6A and 6B). These refer respectively to "poor nurse attitude," "poor doctor attitude," and "poor attitude of other hospital staff." It is intriguing to observe that negative assessments in the field of obstetrics and gynecology mainly revolve around "poor attitude." This phenomenon can perhaps be explained from 2 perspectives. On one hand, during the outbreak of a pandemic, the number of patients has increased significantly, while the number of doctors has not kept pace. As a result, the amount of time and energy they can allocate to each patient is either reduced or can only be sustained by extending their working hours. A study revealed that 73% of obstetrics and gynecology

XSL•FO
RenderX

practitioners work over 50 hours per week. In such long and intense work environments, the occupational burnout rate in the field of obstetrics can even reach 56.6% [58]. This will make obstetricians feel irritable and reduce their patience with the objects they serve or contact [59]. On the other hand, due to the abnormal hormones in pregnant women, fear of childbirth, and serious anxiety, these will make pregnant women more sensitive and uneasy [60-62].

In conclusion, in order to provide a positive medical experience for pregnant women and reduce the likelihood of negative comments, hospital administrators can address the following 2 issues: first, they can decrease the working hours of each health care professional, increase the number of specialized health care personnel, and improve the salary and benefits for health care workers. This will help alleviate the occupational burnout among obstetricians and gynecologists. Second, timely psychological interventions should be provided for pregnant women with a fear of childbirth, including evidence-based treatments such as psychoeducation, cognitive restructuring, and relaxation exercises. These interventions are believed to contribute to reducing the rate of cesarean section deliveries and enhancing overall satisfaction with the childbirth experience [63,64].

The Limitations of This Study

This study has certain limitations. First, the selection of hospitals primarily focused on those with high national rankings, which are mainly located in economically developed cities. Therefore, the conclusions may not be applicable to hospitals in smaller cities. Second, unlike conventional surveys, this study only captured publicly available patient comment information, thus lacking demographic data on the individuals who provided these comments. Third, it was not possible to completely eliminate the possibility of some negative comments being maliciously posted by competitors as false information.

Conclusion

The COVID-19 pandemic has some association with patient comment scores. In this study, comment scores of general hospitals, children's hospitals, and maternity hospitals increased during the pandemic, excluding tumor hospitals. The scores and content of patient comments varied among different specialized hospitals. Analyzing patient comment content using ChatGPT is an innovative method for identifying factors contributing to patient dissatisfaction and provides a user-friendly approach. This study has certain reference value for hospital administrators in establishing harmonious relationships between health care providers and patients and in improving hospital performance during public health emergencies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China project (82270505) and the Shanghai Municipal Hospital Development Center (SHDC12022107).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original data: 30,317 pieces of hospital review information collected from the Dianping website.
[XLSX File (Microsoft Excel File), 9279 KB - jmir_v26i1e52992_app1.xlsx]

Multimedia Appendix 2
Supplementary figures.
[PDF File (Adobe PDF File), 1045 KB - jmir_v26i1e52992_app2.pdf]

References

1. Xiao Y, Liu L, Chen TT, Zhu SY. Rebuilding trust in the physician-patient relationship: addressing causes and implementing reforms in China. Fam Pract 2023:cmad104. [doi: 10.1093/fampra/cmad104] [Medline: 37942660]
2. Duan G, Liao X, Yu W, Li G. Classification and prediction of violence against Chinese medical staff on the sina microblog based on a self-organizing map: quantitative study. J Med Internet Res 2020;22(5):e13294 [FREE Full text] [doi: 10.2196/13294] [Medline: 32348253]
3. Peng W, Ding G, Tang Q, Xu L. Continuing violence against medical personnel in China: a flagrant violation of Chinese law. Biosci Trends 2016;10(3):240-243 [FREE Full text] [doi: 10.5582/bst.2016.01094] [Medline: 27245351]
4. Zhang L, Wang A, Xie X, Zhou Y, Li J, Yang L, et al. Workplace violence against nurses: a cross-sectional study. Int J Nurs Stud 2017;72:8-14. [doi: 10.1016/j.ijnurstu.2017.04.002] [Medline: 28412581]

XSL•FO
RenderX

5. Kaplan SH, Greenfield S, Ware JE. Assessing the effects of physician-patient interactions on the outcomes of chronic disease. Med Care 1989;27(3 Suppl):S110-S127. [Medline: 2646486]
6. Li J, Pang PC, Xiao Y, Wong D. Changes in doctor-patient relationships in China during COVID-19: a text mining analysis. Int J Environ Res Public Health 2022;19(20):13446 [FREE Full text] [doi: 10.3390/ijerph192013446] [Medline: 36294022]
7. Zhang J, Lu H, Zeng H, Zhang S, Du Q, Jiang T, et al. The differential psychological distress of populations affected by the COVID-19 pandemic. Brain Behav Immun 2020;87:49-50 [FREE Full text] [doi: 10.1016/j.bbi.2020.04.031] [Medline: 32304883]
8. Parsons Leigh J, Fiest K, Brundin-Mather R, Plotnikoff K, Soo A, Sypes EE, et al. A national cross-sectional survey of public perceptions of the COVID-19 pandemic: self-reported beliefs, knowledge, and behaviors. PLoS One 2020;15(10):e0241259 [FREE Full text] [doi: 10.1371/journal.pone.0241259] [Medline: 33095836]
9. Howard M, Fleming ML, Parker E. Patients do not always complain when they are dissatisfied: implications for service quality and patient safety. J Patient Saf 2013;9(4):224-231 [FREE Full text] [doi: 10.1097/PTS.0b013e3182913837] [Medline: 24257066]
10. Phau I, Baird M. Complainers versus non-complainers retaliatory responses towards service dissatisfactions. Marketing Intelligence & Planning 2008;26(6):587-604 [FREE Full text] [doi: 10.1108/02634500810902848]
11. Gurung G, Derrett S, Gauld R, Hill PC. Why service users do not complain or have 'voice': a mixed-methods study from Nepal's rural primary health care system. BMC Health Serv Res 2017;17(1):81 [FREE Full text] [doi: 10.1186/s12913-017-2034-5] [Medline: 28122552]
12. Jiang Y, Ying X, Zhang Q, Tang SR, Kane S, Mukhopadhyay M, HESVIC team authorship. Managing patient complaints in China: a qualitative study in Shanghai. BMJ Open 2014;4(8):e005131 [FREE Full text] [doi: 10.1136/bmjopen-2014-005131] [Medline: 25146715]
13. Skålén C, Nordgren L, Annerbäck EM. Patient complaints about health care in a Swedish County: characteristics and satisfaction after handling. Nurs Open 2016;3(4):203-211 [FREE Full text] [doi: 10.1002/nop2.54] [Medline: 27708831]
14. Zhang W, Deng Z, Hong Z, Evans R, Ma J, Zhang H. Unhappy patients are not alike: content analysis of the negative comments from China's good doctor website. J Med Internet Res 2018;20(1):e35 [FREE Full text] [doi: 10.2196/jmir.8223] [Medline: 29371176]
15. Gordon HS. Consider embracing the reviews from physician rating websites. J Gen Intern Med 2017;32(6):599-600 [FREE Full text] [doi: 10.1007/s11606-017-4036-8] [Medline: 28299603]
16. Hsieh SY. A system for using patient complaints as a trigger to improve quality. Qual Manag Health Care 2011;20(4):343-355. [doi: 10.1097/QMH.0b013e318222e73b] [Medline: 21971031]
17. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
18. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
19. Li J. Security implications of AI chatbots in health care. J Med Internet Res 2023;25:e47551 [FREE Full text] [doi: 10.2196/47551] [Medline: 38015597]
20. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023;21(1):269 [FREE Full text] [doi: 10.1186/s12967-023-04123-5] [Medline: 37076876]
21. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
22. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Intern Med 2023;183(9):1028-1030. [doi: 10.1001/jamainternmed.2023.2909] [Medline: 37459090]
23. Haman M, Školník M. Testing ChatGPT's capabilities for social media content analysis. Aesthetic Plast Surg 2023. [doi: 10.1007/s00266-023-03607-5] [Medline: 37605022]
24. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
25. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023;25:e51300 [FREE Full text] [doi: 10.2196/51300] [Medline: 37943581]
26. Conroy RM, Teehan M, Siriwardena R, Smyth O, McGee HM, Fernandes P. Attitudes to doctors and medicine: the effect of setting and doctor-patient relationship. Br J Health Psychol 2002;7(Pt 1):117-125. [doi: 10.1348/135910702169312] [Medline: 14596722]
27. Yan X, Li J. Better doctor-patient relationships start with the small things. BMJ 2023;383:2935. [doi: 10.1136/bmj.p2935] [Medline: 38101920]

XSL•FO
RenderX

28. Webpage in Chinese. Shanghai Hantao Information Consulting Co, Ltd. URL: https://www.dianping.com/review/1317174326 [accessed 2024-01-23]
29. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. JMIR Med Inform 2023;11:e48933 [FREE Full text] [doi: 10.2196/48933] [Medline: 38015610]
30. Domingos P. A few useful things to know about machine learning. Commun. ACM 2012;55(10):78-87. [doi: 10.1145/2347736.2347755]
31. Zhou Y, Ma Y, Yang WFZ, Wu Q, Wang Q, Wang D, et al. Doctor-patient relationship improved during COVID-19 pandemic, but weakness remains. BMC Fam Pract 2021;22(1):255 [FREE Full text] [doi: 10.1186/s12875-021-01600-y] [Medline: 34937550]
32. Zhou Y, Yang WFZ, Ma Y, Wu Q, Yang D, Liu T, et al. Doctor-patient relationship in the eyes of medical professionals in China during the COVID-19 pandemic: a cross-sectional study. Front Psychiatry 2021;12:768089 [FREE Full text] [doi: 10.3389/fpsyt.2021.768089] [Medline: 34777069]
33. Zhou Y, Chen S, Liao Y, Wu Q, Ma Y, Wang D, et al. General perception of doctor-patient relationship from patients during the COVID-19 pandemic in China: a cross-sectional study. Front Public Health 2021;9:646486 [FREE Full text] [doi: 10.3389/fpubh.2021.646486] [Medline: 34295863]
34. Gao B, Dong J. Does the impact of COVID-19 improve the doctor-patient relationship in China? Am J Med Sci 2020;360(3):305-306 [FREE Full text] [doi: 10.1016/j.amjms.2020.05.039] [Medline: 32563570]
35. Semenov AV, Pshenichnaya NY. Born in Wuhan: lessons from COVID-19 epidemic in China. Russian Journal of Infection and Immunity 2020;10(2):210-220. [doi: 10.15789/2220-7619-biw-1453]
36. Liu S, Dai L, Xu J. Tell health stories comprehensively and accurately: a case study of health edition of people's daily. Int J Nurs Sci 2020;7(Suppl 1):S46-S51 [FREE Full text] [doi: 10.1016/j.ijnss.2020.07.005] [Medline: 32995379]
37. Jiang L, Broome ME, Ning C. The performance and professionalism of nurses in the fight against the new outbreak of COVID-19 epidemic is laudable. Int J Nurs Stud 2020;107:103578 [FREE Full text] [doi: 10.1016/j.ijnurstu.2020.103578] [Medline: 32446015]
38. Gan Y, Chen Y, Wang C, Latkin C, Hall BJ. The fight against COVID-19 and the restoration of trust in Chinese medical professionals. Asian J Psychiatr 2020;51:102072 [FREE Full text] [doi: 10.1016/j.ajp.2020.102072] [Medline: 32334408]
39. Roubille C, Ribstein J, Hurpin G, Fesler P, Fiat E, Roubille F. Confidence vanished or impaired until distrust in the doctor-patient relationship because of COVID-19: confidence vanished or impaired until distrust: "COVID" in relationship. Rev Med Interne 2021;42(1):58-60 [FREE Full text] [doi: 10.1016/j.revmed.2020.10.007] [Medline: 33168352]
40. Lyu B, Xu M, Lu L, Zhang X. Burnout syndrome, doctor-patient relationship and family support of pediatric medical staff during a COVID-19 local outbreak in Shanghai China: a cross-sectional survey study. Front Pediatr 2023;11:1093444 [FREE Full text] [doi: 10.3389/fped.2023.1093444] [Medline: 36861079]
41. Jiang MM, Wu ZY, Tu AX. Research on the cooperative governance path of multiple stakeholders in doctor-patient disputes under the environment of information asymmetry. Int J Environ Res Public Health 2023;20(2):1597 [FREE Full text] [doi: 10.3390/ijerph20021597] [Medline: 36674353]
42. Wang Y, Wu H, Lei X, Shen J, Feng Z. The influence of doctors' online reputation on the sharing of outpatient experiences: empirical study. J Med Internet Res 2020;22(12):e16691 [FREE Full text] [doi: 10.2196/16691] [Medline: 33306028]
43. Drossman DA, Chang L, Deutsch JK, Ford AC, Halpert A, Kroenke K, et al. A review of the evidence and recommendations on communication skills and the patient-provider relationship: a Rome foundation working team report. Gastroenterology 2021;161(5):1670-1688 [FREE Full text] [doi: 10.1053/j.gastro.2021.07.037] [Medline: 34331912]
44. Tavakoly Sany SB, Behzhad F, Ferns G, Peyman N. Communication skills training for physicians improves health literacy and medical outcomes among patients with hypertension: a randomized controlled trial. BMC Health Serv Res 2020;20(1):60 [FREE Full text] [doi: 10.1186/s12913-020-4901-8] [Medline: 31973765]
45. Liu Y, Wang P, Bai Y. The influence factors of medical disputes in Shanghai and implications - from the perspective of doctor, patient and disease. BMC Health Serv Res 2022;22(1):1128 [FREE Full text] [doi: 10.1186/s12913-022-08490-5] [Medline: 36071431]
46. Labrie N, Schulz PJ. Does argumentation matter? A systematic literature review on the role of argumentation in doctor-patient communication. Health Commun 2014;29(10):996-1008. [doi: 10.1080/10410236.2013.829018] [Medline: 24359318]
47. Liang Z, Xu M, Liu G, Zhou Y, Howard P. Patient-centred care and patient autonomy: doctors' views in Chinese hospitals. BMC Med Ethics 2022;23(1):38 [FREE Full text] [doi: 10.1186/s12910-022-00777-w] [Medline: 35395761]
48. Morrison EJ, Novotny PJ, Sloan JA, Yang P, Patten CA, Ruddy KJ, et al. Emotional problems, quality of life, and symptom burden in patients with lung cancer. Clin Lung Cancer 2017;18(5):497-503. [doi: 10.1016/j.cllc.2017.02.008] [Medline: 28412094]
49. Terashima T, Tsutsumi A, Iwami E, Kuroda A, Nakajima T, Eguchi K. Delayed visit and treatment of lung cancer during the coronavirus disease 2019 pandemic in Japan: a retrospective study. J Int Med Res 2022;50(5):3000605221097375 [FREE Full text] [doi: 10.1177/03000605221097375] [Medline: 35579175]

XSL•FO
RenderX

50. Liu C, Piao H, Zhang T, Yang D, Li X, Tang X. Delayed diagnosis and treatment of cancer patients during the COVID-19 pandemic in Henan, China: an interrupted time series analysis. Front Public Health 2022;10:881718 [FREE Full text] [doi: 10.3389/fpubh.2022.881718] [Medline: 35685763]
51. Ribeiro R, Wainstein AJA, de Castro Ribeiro HS, Pinheiro RN, Oliveira AF. Perioperative cancer care in the context of limited resources during the COVID-19 pandemic: Brazilian Society of Surgical Oncology Recommendations. Ann Surg Oncol 2021;28(3):1289-1297 [FREE Full text] [doi: 10.1245/s10434-020-09098-x] [Medline: 32980998]
52. Ye Y, Wang J, Cai S, Fu X, Ji Y. Psychological distress of cancer patients caused by treatment delay during the COVID-19 pandemic in China: a cross-sectional study. Psychooncology 2022;31(9):1607-1615 [FREE Full text] [doi: 10.1002/pon.5946] [Medline: 35506550]
53. Bolger JC, Donlon NE, Butt W, Neary C, Al Azzawi M, Brett O, et al. Successful maintenance of process and outcomes for oesophageal cancer surgery in Ireland during the first wave of the COVID-19 pandemic. Ir J Med Sci 2022;191(2):831-837 [FREE Full text] [doi: 10.1007/s11845-021-02597-0] [Medline: 33728528]
54. Hu KJ, Sun ZZ, Rui YJ, Mi JY, Ren MX. Shortage of paediatricians in China. Lancet 2014;383(9921):954. [doi: 10.1016/S0140-6736(14)60482-7] [Medline: 24629297]
55. Zhang Y, Huang L, Zhou X, Zhang X, Ke Z, Wang Z, et al. Characteristics and workload of pediatricians in China. Pediatrics 2019;144(1):e20183532. [doi: 10.1542/peds.2018-3532] [Medline: 31253739]
56. Hu Y, Xu J, Dong W, Yuan Z, Sun X. Strategies to correct the shortage of paediatricians in China. Lancet 2018;392(10145):385. [doi: 10.1016/S0140-6736(18)31192-9] [Medline: 30102171]
57. Deng W, Feng Z, Yao X, Yang T, Jiang J, Wang B, et al. Occupational identity, job satisfaction and their effects on turnover intention among Chinese paediatricians: a cross-sectional study. BMC Health Serv Res 2021;21(1):6 [FREE Full text] [doi: 10.1186/s12913-020-05991-z] [Medline: 33397391]
58. Ye J, Wang H, Wu H, Ye L, Li Q, Ma XY, et al. Burnout among obstetricians and paediatricians: a cross-sectional study from China. BMJ Open 2019 25;9(1):e024205 [FREE Full text] [doi: 10.1136/bmjopen-2018-024205] [Medline: 30782730]
59. Rosenstein AH. Physician stress and burnout: what can we do? Physician Exec 2012;38(6):22-26, 28, 30. [Medline: 23888648]
60. Fenwick J, Toohill J, Creedy DK, Smith J, Gamble J. Sources, responses and moderators of childbirth fear in Australian women: a qualitative investigation. Midwifery 2015;31(1):239-246. [doi: 10.1016/j.midw.2014.09.003] [Medline: 25440298]
61. Jokić-Begić N, Zigić L, Nakić Radoš S. Anxiety and anxiety sensitivity as predictors of fear of childbirth: different patterns for nulliparous and parous women. J Psychosom Obstet Gynaecol 2014;35(1):22-28. [doi: 10.3109/0167482X.2013.866647] [Medline: 24328559]
62. Koc AE, Colak S, Colak GV, Pusuroglu M, Hocaoglu C. Investigating fear of childbirth in pregnant women and its relationship between anxiety sensitivity and somatosensory amplification. J Obstet Gynaecol 2021;41(2):217-223. [doi: 10.1080/01443615.2020.1732894] [Medline: 32314633]
63. Rouhe H, Salmela-Aro K, Toivanen R, Tokola M, Halmesmäki E, Saisto T. Obstetric outcome after intervention for severe fear of childbirth in nulliparous women - randomised trial. BJOG 2013;120(1):75-84. [doi: 10.1111/1471-0528.12011] [Medline: 23121002]
64. Watt MC, Stewart SH, Lefaivre MJ, Uman LS. A brief cognitive-behavioral approach to reducing anxiety sensitivity decreases pain-related anxiety. Cogn Behav Ther 2006;35(4):248-256. [doi: 10.1080/16506070600898553] [Medline: 17189242]

Abbreviations

HPF: hospital-patient factor
RQ: research question

Edited by A Mavragani; submitted 22.09.23; peer-reviewed by D Carvalho, U Sander; comments to author 11.01.24; revised version received 24.01.24; accepted 21.05.24; published 02.07.24.

Please cite as:
Huo W, He M, Zeng Z, Bao X, Lu Y, Tian W, Feng J, Feng R
Impact Analysis of COVID-19 Pandemic on Hospital Reviews on Dianping Website in Shanghai, China: Empirical Study
J Med Internet Res 2024;26:e52992
URL: https://www.jmir.org/2024/1/e52992
doi:10.2196/52992
PMID:

XSL•FO
RenderX

©Weixue Huo, Mengwei He, Zhaoxiang Zeng, Xianhao Bao, Ye Lu, Wen Tian, Jiaxuan Feng, Rui Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Determining an Appropriate Sample Size for Qualitative Interviews to Achieve True and Near Code Saturation: Secondary Analysis of Data

Claudia M Squire[1], MS; Kristen C Giombi[1], PhD; Douglas J Rupert[1], MPH; Jacqueline Amoozegar[1], MSPH; Peyton Williams[1], MPH

RTI International, Research Triangle Park, NC, United States

Corresponding Author:
Claudia M Squire, MS
RTI International
3040 East Cornwallis Road
Research Triangle Park, NC, 27709-2194
United States
Phone: 1 9195416613
Email: cms@rti.org

Abstract

Background: In-depth interviews are a common method of qualitative data collection, providing rich data on individuals' perceptions and behaviors that would be challenging to collect with quantitative methods. Researchers typically need to decide on sample size a priori. Although studies have assessed when saturation has been achieved, there is no agreement on the minimum number of interviews needed to achieve saturation. To date, most research on saturation has been based on in-person data collection. During the COVID-19 pandemic, web-based data collection became increasingly common, as traditional in-person data collection was possible. Researchers continue to use web-based data collection methods post the COVID-19 emergency, making it important to assess whether findings around saturation differ for in-person versus web-based interviews.

Objective: We aimed to identify the number of web-based interviews needed to achieve true code saturation or near code saturation.

Methods: The analyses for this study were based on data from 5 Food and Drug Administration–funded studies conducted through web-based platforms with patients with underlying medical conditions or with health care providers who provide primary or specialty care to patients. We extracted code- and interview-specific data and examined the data summaries to determine when true saturation or near saturation was reached.

Results: The sample size used in the 5 studies ranged from 30 to 70 interviews. True saturation was reached after 91% to 100% (n=30-67) of planned interviews, whereas near saturation was reached after 33% to 60% (n=15-23) of planned interviews. Studies that relied heavily on deductive coding and studies that had a more structured interview guide reached both true saturation and near saturation sooner. We also examined the types of codes applied after near saturation had been reached. In 4 of the 5 studies, most of these codes represented previously established core concepts or themes. Codes representing newly identified concepts, other or miscellaneous responses (eg, "in general"), uncertainty or confusion (eg, "don't know"), or categorization for analysis (eg, correct as compared with incorrect) were less commonly applied after near saturation had been reached.

Conclusions: This study provides support that near saturation may be a sufficient measure to target and that conducting additional interviews after that point may result in diminishing returns. Factors to consider in determining how many interviews to conduct include the structure and type of questions included in the interview guide, the coding structure, and the population under study. Studies with less structured interview guides, studies that rely heavily on inductive coding and analytic techniques, and studies that include populations that may be less knowledgeable about the topics discussed may require a larger sample size to reach an acceptable level of saturation. Our findings also build on previous studies looking at saturation for in-person data collection conducted at a small number of sites.

(J Med Internet Res 2024;26:e52998) doi:10.2196/52998

XSL•FO
RenderX

KEYWORDS

saturation; sample size; web-based data collection; semistructured interviews; qualitative; research methods; research methodology; data collection; coding; interviews; interviewing; in-depth

Introduction

Background

In-depth interviews are commonly used to collect qualitative data for a wide variety of research purposes across many subject matter areas. These types of interviews are an ideal approach for examining individuals' perceptions and behaviors at a level of depth, complexity, and richness that would be challenging to achieve with quantitative data collection methods. Typically, trained interviewers conduct interviews using a guide designed to address the study's key research aims by asking a series of questions and probes ordered by topic. These interview guides can range from highly structured to completely unstructured (eg, loosely organized conversations). Following the completion of data collection, interview notes and transcripts generated from audio recordings of the interviews are analyzed to assess for patterns in responses among the interviewees or subsets of the participants [1,2].

During the COVID-19 pandemic, web-based data collection became increasingly common, as traditional in-person data collection was not possible, and researchers continue to use web-based data collection methods post the COVID-19 emergency, citing advantages such as accessing marginalized populations, achieving greater geographic diversity, being able to offer a more flexible schedule, and saving on travel expenses [3]. Potential concerns about web-based data collection, such as the inability to build rapport and data richness, have been largely unfounded [3,4].

While we do not expect web-based data collection to supplant in-person research, it continues to show signs of growth. To date, much of the research on qualitative methods has focused on in-person data collection. Consequently, it will be important to conduct research to determine if previous widely accepted findings hold true for web-based data collection.

Researchers typically make a priori decisions about the number of interviews to conduct with the aim of balancing the need for sufficient data with resource limitations and respondent burden. The concept of saturation is frequently used to justify the study's rigor with respect to the selected sample size. To provide empirically based recommendations on adequate minimum sample sizes, researchers have conducted studies to assess when saturation occurs. However, multiple types of saturation exist—such as theoretical, thematic, code, and meaning—and within each type of saturation, the definitions and measurement approaches used by investigators vary substantially, as does the level of detail researchers report in publications about their methods for achieving and assessing saturation [5].

This study aimed to examine the number of interviews needed to obtain code saturation for 5 recently conducted studies funded by the Food and Drug Administration [6] involving web-based interviews. Specifically, how many web-based interviews are needed to obtain true code saturation (ie, the use of 100% of all codes applied in the study) and how many web-based interviews are needed to achieve near code saturation (ie, the use of 90% of all codes applied in the study)?

Literature Review

Multiple authors have defined saturation as the point during data collection and analysis, at which no new additional data are found that reveal a new conceptual category [7-13] or theme related to the research question—an indicator that further data collection is redundant [11]. Additionally, Coenen et al [14] specified that no new second-level themes are revealed in 2 consecutive focus groups or interviews.

Other authors have distinguished between various types of saturation. One of the most common types of saturation mentioned in the literature is theoretical saturation, which emerges from grounded theory and occurs when the concepts of a theory are fully reflected in the data and no new insights, themes, or issues are identified from the data [5,11,12,15-18]. Hennink et al [17] expanded this definition, adding that all relevant conceptual categories should have been identified, thus emphasizing the importance of sample adequacy over sample size. Guest et al [15] operationalized the concept of theoretical saturation as the point in data collection and analysis when new information produces little or no change to the codebook, and van Rijnsoever [19] operationalized it as being when all the codes have been observed once in the sample.

Some authors have defined theoretical saturation, thematic saturation, and data saturation as the same concept [16,18], whereas others have defined these terms differently [12,20]. For example, some authors have defined thematic saturation as the point where no new codes or themes are emerging from the data [12,21]. For thematic saturation to be achieved, data should be collected until nothing new is generated [20,22]. Data saturation has been defined as the level to which new data are repetitive of the data that have been collected [12,23,24].

Furthermore, Hennink et al [17] distinguished between code saturation and meaning saturation. Code saturation is based on primary or parent codes and relates to the quantity of the data ("hearing it all"). Meaning saturation is based on sub or child codes and relates to the quality or richness of the data ("understanding it all"). Constantinou et al [7] made the point that it is the categorization of the raw data, rather than the data, that are saturated.

The literature reflects multiple methods that have been used to determine saturation [7-10,13-18,21,25]. Sim et al [26] discussed the four general approaches that have been used to determine sample size for qualitative research: (1) rules of thumb, based on a combination of methodological considerations and past experience; (2) conceptual models, based on specific characteristics of the proposed study; (3) numerical guidelines derived from the empirical investigation; and (4) statistical approaches, based on the probability of obtaining a sufficient sample size.

XSL•FO
RenderX

For example, Galvin [9] used a statistical approach based on binomial logic to establish the relationship between identifying a theme in a particular sample and within the larger population; for example, number of chances of detecting a theme if that theme exists within number of the population. Using the probability equation, the researcher can determine the number of interviews needed for a stated level of confidence that all relevant themes held by a certain proportion of the population will occur within the interview sample. This method assumes the researcher knows in advance the emergent themes from the study and at what rate they may occur.

Constantinou et al [7] used the comparative method for themes saturation, which relies on both a deductive and an inductive approach to generate codes (keywords extracted from the participants' words) and themes (codes that fall into similar categories). Themes are compared across interviews, and theme saturation is reached when the next interview does not produce any new themes. The sequence of interviews is reordered multiple times to check for order-induced error. When exploring the various methods for determining saturation, researchers reached different conclusions on when saturation was achieved (findings on saturation by other authors are present in Multimedia Appendix 1) [7-10,13-17,21,25,27,28].

Most studies assessing saturation focused on in-person data collection or did not specify the data collection method. Given recent increases in web-based data collection, studies assessing saturation for web-based interviews are critical to ensure that recommendations regarding sample size are tailored to the mode of data collection [4]. While there is evidence to suggest that the content of data coded from in-person as compared with web-based interviews is conceptually similar [29], this is a relatively new area of exploration. Rapport may be higher with in-person as compared with web-based interviews [30], which may impact the amount and type of content generated. Additionally, participants in web-based data collection studies are more geographically diverse and may be more likely to be non-White, less educated, and less healthy than participants in in-person data collection studies [31].

Methods

Study Design

This study was based on analyses from data collected for 5 Food and Drug Administration–funded studies conducted using web-based platforms, such as Zoom (Zoom Video Communications) and Adobe Connect (Adobe Systems), and focused on patients with underlying medical conditions or on health care providers who provide primary or specialty care to patients. All platforms used for these interviews offered audio and video components and allowed for the sharing of stimuli on screen. A brief description of each study is provided in Table 1. Each study's data had been coded and stored using NVivo software (version 11; QSR International).

XSL•FO
RenderX

Table 1. Description of studies included in analysis of code saturation: sample size, eligibility criteria, topics covered, length of interview, number of questions, and regions and states covered.

Study name	Sample size, n	General eligibility criteria	Primary objectives	Summary of topics	Length of interview (minutes)	Number of interview questions	Regions and states covered
Study A	30	Patients diagnosed with a condition treated by biologic medications (eg, cancer, inflammatory bowel disease, and diabetes)	Obtain feedback on multimedia educational materials about biosimilar biologic medications	• Biosimilar awareness • Feedback on educational materials (eg, comprehension, main message, and format) • Behavioral intentions	90	• 37 main questions • Questions identified as high, average, and low priority	• Regions: Northeast, Midwest, South, and West • States: 14
Study B	48	Patients diagnosed with vulvovaginal atrophy or type 2 diabetes	Explore how patients use boxed warnings when making decisions about prescription drugs and how well the warnings meet patients' information needs	• Prescription drug information needs • Boxed warning awareness, interpretation, and perceptions • Behavioral intentions	30	• 13 main questions	• Regions: Northeast, Midwest, South, and West • States: not available
Study C	70	Primary care physicians or specialists who write at least 50 prescriptions per week	Assess how primary care physicians and specialists access, understand, and use prescription drug labeling information, including information on labels for drugs that have multiple indications.	• Resources to find information about prescription drugs • Background on prescribing information • Interpretation of language in the prescribing information	60	• 36 main questions	• Regions: Northeast, Midwest, South, and West • States: 26
Study D	35	Patients diagnosed with type 2 diabetes	Understand how patients weigh the potential benefits against possible risks and side effects, dosage and administration characteristics, and costs when selecting treatments for chronic health conditions.	• Background information on condition • Treatment decisions and discussion of attributes • Ranking attributes • Condition-specific statements about attributes • Market claims	60	• 20 main questions	• Regions: Northeast, Midwest, South, and West • States: 9
Study E	35	Patients diagnosed with psoriasis	Understand how patients weigh the potential benefits against possible risks and side effects, dosage and administration characteristics, and costs when selecting treatments for chronic health conditions.	• Background information on condition • Treatment decisions and discussion of attributes • Ranking attributes • Condition-specific statements about attributes • Market claims	60	• 21 main questions	• Regions: Northeast, Midwest, South, and West • States: 9

Ethical Considerations

This project was determined to not research with human participants by Research Triangle Institute's institutional review board (STUDY00021985). The original 5 studies that this project is based on were reviewed by Research Triangle Institute's institutional review board and were determined to be exempt under category 2ii. Participants in these studies were provided information about measures used to protect their privacy and the confidentiality of their data in the study's consent forms. All participants were provided compensation for their time (the amount and type varied by study).

XSL•FO
RenderX

Data Preparation and Analysis

We established and applied a systematic approach to analyze all 5 study data sets. Our analytic approach was organized into 2 stages—data preparation and data analysis.

Data Preparation

First, because previous interviews sometimes influence moderator probes—for example, the moderator asks a follow-up question based on something they heard in a previous interview—we sorted interviews from each study by interview order. We then extracted code- and interview-specific data from the NVivo databases—including transcript name, code name, number of files coded, number of associated parent and child codes, and number of coding references—and compiled these data in an Excel (Microsoft Corp) file. We then updated the Excel file with important code and interview characteristics, including the order in which interviews were conducted, whether each code was directly (ie, child codes) or indirectly (ie, parent codes) applied to transcripts (in a tiered coding scheme, direct codes are those that have no child codes, whereas indirect codes function as "parents" that have additional codes nested beneath them), and the point at which each code was first applied to an interview. Finally, we created pivot tables within each Excel file to compile the data.

Data Analysis

Once the data were compiled, the data summaries were examined to determine when true saturation and near saturation occurred during data collection. True saturation was defined as 100% of all applied codes being used; near saturation was defined as 90% of all applied codes being used. We calculated saturation separately for each study's data set, and we calculated saturation separately for all codes (ie, parent and child codes) as compared with direct codes (ie, child codes only). True saturation and near saturation points were identified by calculating the cumulative percentage of new codes for each interview, flagging when 100% and 90% of applied codes had been used.

Results

True and Near Saturation

The number of web-based interviews used across the 5 studies ranged from 30 to 70 (Table 2). True saturation (100% use of all applied codes) was reached in the final or near final interview (Figure 1), suggesting that, even with a large sample size, additional interviews are likely to continue uncovering a small number of new codes or findings.

Table 2. Interviews needed to reach true and near saturation by study.

Study	Total interviews, n	Coding: total codes in codebook, n	True saturation: interviews needed, n (%)	Near saturation: interviews needed, n (%)
Study A	30	657	30 (100)	18 (60)
Study B	48	313	47 (98)	21 (44)
Study C	70	362	67 (96)	23 (33)
Study D	35	205	33 (94)	15 (43)
Study E	35	200	32 (91)	15 (43)

Figure 1. Illustration of cumulative percentage of new codes applied by study to reach true and near saturation.

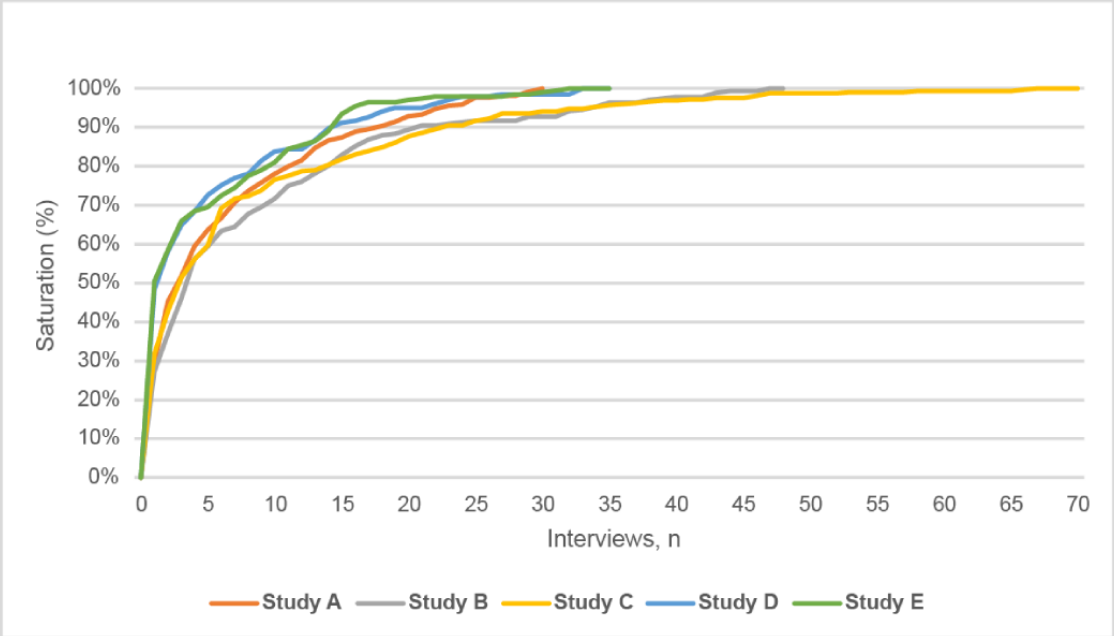

XSL•FO
RenderX

Across all studies, near saturation (90% use of all applied codes) was reached near—and often before—the midpoint of data collection. In other words, only a small number of new codes or findings were uncovered once the first half of the sample had been interviewed. In terms of absolute numbers, the point at which near saturation was reached occurred between 33% and 60% (n=15-23) of planned interviews (Table 2). Despite the participants being more geographically, and possibly demographically, diverse compared with typical in-person participants, our findings were similar to previous studies on saturation [10,15,17].

We examined the types of codes applied after near saturation had been reached. In 4 of the 5 studies, most of these codes (n=8-33, 57%-62%) represented previously established core concepts or themes, such as a trusted source of information, a behavioral intention, or a recommended change to educational material. Codes representing newly identified concepts (n=2-8, 10%-15%), other miscellaneous responses (eg, “in general”; n=6-9, 13%-41%), uncertainty or confusion (eg, “don’t know”; n=0-6, 0%-11%), or categorization for analysis (eg, “correct as compared with incorrect”; n=0-3, 0%-4%) were less commonly applied after near saturation had been reached.

The overwhelming majority of codes applied after near saturation (n=9-41, 73%-82%) had already been established in study codebooks before analysis. Only a small number of codes applied after this point (n=4-20, 18%-27%) were conceptually distinct enough to merit updating the study codebooks by including them. Likewise, most of the codes used after near saturation (n=11-35, 44%-64%) were applied to only a single interview. Far fewer codes were applied to 2 interviews (n=0-13, 0%-27%), 3 interviews (n=0-6, 0%-21%), or 4 or more interviews (n=0-12, 0%-21%).

Study B was an outlier in terms of codes applied after near saturation. This study had fewer codes representing core established concepts (n=8, 28%) and more codes representing newly identified concepts (n=7, 24%) or providing categorization for analysis (n=3, 10%) than other studies. The study also had a much higher proportion of new codes (n=20, 69%) that were added to the study codebook during analysis. These differences may be because the study sampled 2 populations with very different medical conditions (ie, type 2 diabetes as compared with vulvovaginal atrophy), leading to a broader range of applied codes.

In examining the relationship between the number of codes in the codebook for each study, the study with the most codes (study A: 657 codes) required the largest number of interviews to reach both true saturation and near saturation. However, this pattern did not hold true for the remainder of the studies. The study with the next highest number of codes (study C: 362 codes) was third to reach true saturation and last to reach near saturation.

Parent and Child Codes

All 5 study codebooks included both parent (ie, top-level codes) and child codes (ie, subcodes). We examined saturation using two analytic lenses—(1) all codes (parent and child) and (2) parent codes only—to determine if there were differences in when saturation was reached. We found no differences in when true saturation was reached. However, near saturation was reached slightly later (ie, after an additional 3 to 4 interviews) when examining only parent codes (Figure 2).

Figure 2. Illustration of cumulative percentage of new codes applied by study to reach true and near saturation (parent codes only).

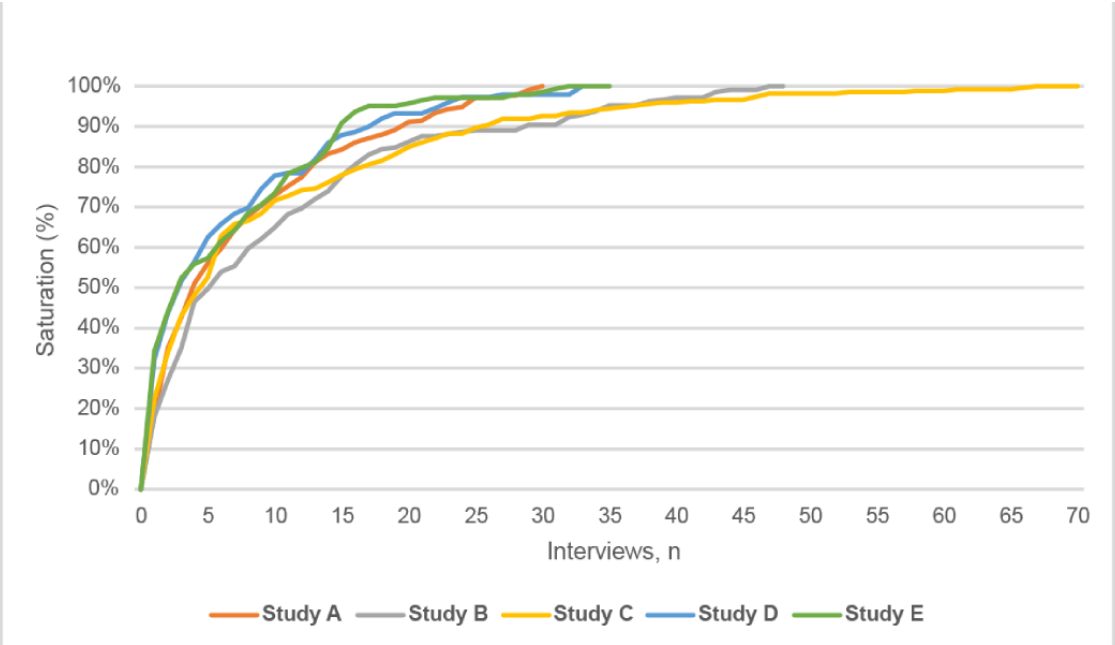

Differences by Study

In total, 3 of the studies had codebooks that consisted almost entirely of deductive (ie, concept-driven) codes, whereas the codebooks in the remaining 2 studies contained a mix of both deductive and inductive (ie, data-driven) codes. Although the results were largely consistent across the 5 studies, as expected, the studies that relied heavily on deductive coding reached both true saturation and near saturation sooner. This finding suggests

XSL•FO
RenderX

that studies using more inductive coding and analytic techniques may require slightly larger sample sizes to reach saturation.

Structure of an Interview Guide

Although all the studies used a semistructured interview guide, the level of structure varied across studies. The 3 studies (ie, studies C, D, and E) that had a more structured interview guide (eg, questions for which participants were asked their preference among discrete choices or the range of likely answers was limited) reached both true saturation and near saturation sooner. In fact, the study with the most structured guide reached near saturation the soonest, although it fell in the middle for true saturation. This finding suggests that studies using a less structured interview guide may need to conduct more interviews to reach an acceptable level of saturation.

Discussion

Principal Findings

Although true saturation was not reached until the final interview or close to the final interview, near saturation was reached much sooner, ranging from just below to just above the midpoint of data collection, with most of the studies falling just below the midpoint. Although additional interviews conducted after near saturation may result in new information, our findings suggest there may be diminishing returns relative to the resources expended. We have identified several study characteristics that researchers can consider when making decisions on sample size for web-based interviews.

Although our findings were mostly consistent across the 5 studies we examined, near saturation was reached sooner on the studies that consisted of largely deductive codes compared with those that had a greater number of inductive codes. Consequently, researchers should consider their analytic approach when determining sample size. Studies that intend for the coding scheme to be iterative throughout the coding process may want to err on the side of having a slightly higher sample size than if the codebook is expected to consist largely of deductive codes tied to the interview guide.

These studies ranged in length from 30 to 90 minutes, and a majority (n=3) lasted 60 minutes. Although the 90-minute study reached both true saturation and near saturation at the latest point, the shortest interview (at 30 minutes) required the second-highest number of interviews to reach both saturation points. Although the length of the interview may be a minor consideration, the level of structure of the interview guide and the types of codes used seem to be larger drivers.

Our findings point to the need for a slightly higher number of interviews to reach an acceptable level of saturation—categorized by us as near code saturation—than what has been found in other studies. For example, Guest et al [15] found that 6 interviews were enough to get high-level themes, reaching a plateau at 10 to 12 interviews. Similarly, Young and Casey [27] found that near code saturation was reached at 6 to 9 interviews.

Our findings also build on previous studies looking at saturation for in-person data collection conducted at a small number of sites. Data from our studies included participants from all US Census Bureau regions, which provides support that these findings may be more generalizable than previous studies.

Limitations

Our study had several limitations. First, our analysis was conducted on a sample of 5 studies that had similarities. All the studies were related to the medical field, and our study populations (patients with an identified medical condition and health care providers) were knowledgeable about the topics discussed. Second, all the studies were conducted using semistructured interview guides that leaned toward being more structured (ie, interviewers largely stuck to scripted probes as compared with guides that allow for unscripted follow-up probes and unstructured conversations). Additionally, all the studies used a similar approach to coding by using a mix of both deductive and inductive codes (though to varying extents). Consequently, studies with a less structured approach to both the interview and coding process may yield different results. Finally, all our studies are broadly classified as social science research. The findings for other fields of inquiry, such as economic or medical studies, may differ.

Conclusions

Saturation is an important consideration in planning and conducting qualitative research, yet, there is no definitive guidance on how to define and measure saturation, particularly for web-based data collection, which allows for data to be collected from a more geographically diverse sample. Our study provides support that near saturation may be a sufficient measure to target and that conducting additional interviews after that point may result in diminishing returns. Factors to consider in determining how many interviews to conduct include the structure and type of questions included in the interview guide, the coding structure, and the population being studied. Studies with less structured interview guides, studies that rely heavily on inductive coding and analytic techniques, and studies that include populations that may be less knowledgeable about the topics discussed may require a larger sample size to reach an acceptable level of saturation. Rather than trying to reach a consensus on the number of interviews needed to achieve saturation in qualitative research overall, we recommend that future research should explore saturation within different types of studies, such as different fields of inquiry, subject matter, and populations being studied. Creating a robust body of knowledge in this area will allow researchers to identify the guidance that best meets the needs of their work.

Acknowledgments

Research Triangle Institute–affiliated authors received support for the development of this manuscript from the RTI Fellow's program under RTI Fellow, Leila Kahwati, MPH, MD. All studies included in the analyses were funded by the Food and Drug

XSL•FO
RenderX

Administration. The authors would like to thank the following Food and Drug Administration staff for their contribution to this research: Kit Aikin, Kevin Betts, Amie O'Donoghue, and Helen Sullivan.

Data Availability

The data sets analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Achieving saturation in interviews: saturation type, methods for achieving saturation, and findings by other authors.
[DOCX File , 25 KB - jmir_v26i1e52998_app1.docx]

References

1. Miles MB, Michael HA, Saldana J. Qualitative Data Analysis: A Methods Sourcebook. Los Angeles, CA: SAGE Publications; 2014.
2. Trochim W, Donnelly J. The Research Methods Knowledge Base. 3rd Edition. Mason, OH: Atomic Dog Publishing; 2008.
3. Keen S, Lomeli-Rodriguez M, Joffe H. From challenge to opportunity: virtual qualitative research during COVID-19 and beyond. Int J Qual Methods 2022;21:16094069221105075 [FREE Full text] [doi: 10.1177/16094069221105075] [Medline: 35692956]
4. Roberts J, Pavlakis A, Richards M. It's more complicated than it seems: virtual qualitative research in the COVID-19 era. Int J Qual Methods 2021 Mar 25;20(20):160940692110029-160940692110020. [doi: 10.1177/16094069211002959]
5. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018;18(1):148 [FREE Full text] [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]
6. Lobe B, Morgan DL, Hoffman K. A systematic comparison of in-person and video-based online interviewing. Int J Qual Methods 2022 Sep 15;21:160940692211270. [doi: 10.1177/16094069221127068]
7. Constantinou CS, Georgiou M, Perdikogianni M. A comparative method for themes saturation (CoMeTS) in qualitative interviews. Qual Res 2017 Jan 23;17(5):571-588. [doi: 10.1177/1468794116686650]
8. Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health 2010 Dec;25(10):1229-1245. [doi: 10.1080/08870440903194015] [Medline: 20204937]
9. Galvin R. How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge? J Build Eng 2015 Mar;1:2-12. [doi: 10.1016/j.jobe.2014.12.001]
10. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
11. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
12. Sebele-Mpofu FY, Serpa S. Saturation controversy in qualitative research: complexities and underlying assumptions. A literature review. Cogent Soc Sci 2020 Oct 23;6(1):1-17. [doi: 10.1080/23311886.2020.1838706]
13. Turner-Bowker DM, Lamoureux RE, Stokes J, Litcher-Kelly L, Galipeau N, Yaworsky A, et al. Informing a priori sample size estimation in qualitative concept elicitation interview studies for clinical outcome assessment instrument development. Value Health 2018 Jul;21(7):839-842 [FREE Full text] [doi: 10.1016/j.jval.2017.11.014] [Medline: 30005756]
14. Coenen M, Stamm TA, Stucki G, Cieza A. Individual interviews and focus groups in patients with rheumatoid arthritis: a comparison of two qualitative methods. Qual Life Res 2012 Mar;21(2):359-370. [doi: 10.1007/s11136-011-9943-2] [Medline: 21706128]
15. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2016 Jul 21;18(1):59-82. [doi: 10.1177/1525822X05279903]
16. Hagaman AK, Wutich A. How many interviews are enough to identify metathemes in multisited and cross-cultural research? Another perspective on Guest, Bunce, and Johnson's (2006) landmark study. Field Methods 2016 Jul 24;29(1):23-41. [doi: 10.1177/1525822x16640447]
17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608 [FREE Full text] [doi: 10.1177/1049732316665344] [Medline: 27670770]
18. Lowe A, Norris AC, Farris AJ, Babbage DR. Quantifying thematic saturation in qualitative data analysis. Field Methods 2018 Jan 22;30(3):191-207. [doi: 10.1177/1525822x17749386]
19. van Rijnsoever FJ. (I Can't Get No) Saturation: a simulation and guidelines for sample sizes in qualitative research. PLoS One 2017;12(7):e0181689 [FREE Full text] [doi: 10.1371/journal.pone.0181689] [Medline: 28746358]

XSL•FO
RenderX

20. O'Reilly M, Parker N. 'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. Qual Res 2012 May 17;13(2):190-197. [doi: 10.1177/1468794112446106]
21. Weller SC, Vickers B, Bernard HR, Blackburn AM, Borgatti S, Gravlee CC, et al. Open-ended interview questions and saturation. PLoS One 2018;13(6):e0198606 [FREE Full text] [doi: 10.1371/journal.pone.0198606] [Medline: 29924873]
22. Green J, Thorogood N. Chapter 4: In-depth interviews. In: Qualitative Methods for Health Research. 2nd Edition. London: Sage Publications; 2004:198-202.
23. Fusch PI, Ness LR. Are we there yet? Data saturation in qualitative research. Qual Rep 2015:1408-1416. [doi: 10.46743/2160-3715/2015.2281]
24. Bowen GA. Naturalistic inquiry and the saturation concept: a research note. Qual Res 2008 Feb 01;8(1):137-152. [doi: 10.1177/1468794107085301]
25. Fugard AJB, Potts HW. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. Int J Soc Res Methodol 2015 Feb 10;18(6):669-684. [doi: 10.1080/13645579.2015.1005453]
26. Sim J, Saunders B, Waterfield J, Kingstone T. Can sample size in qualitative research be determined a priori? Int J Soc Res Methodol 2018 Mar 27;21(5):619-634. [doi: 10.1080/13645579.2018.1454643]
27. Young DS, Casey EA. An examination of the sufficiency of small qualitative samples. Soc Work Res 2019;43(1):53-58. [doi: 10.1093/swr/svy026]
28. Ryan GW, Bernard HR. Techniques to identify themes. Field Methods 2016 Jul 24;15(1):85-109. [doi: 10.1177/1525822X02239569]
29. Namey E, Guest G, O'Regan A, Godwin CL, Taylor J, Martinez A. How does mode of qualitative data collection affect data and cost? Findings from a quasi-experimental study. Field Methods 2019 Dec 16;32(1):58-74. [doi: 10.1177/1525822x19886839]
30. Namey E, Guest G, O'Regan A, Godwin CL, Taylor J, Martinez A. How does qualitative data collection modality affect disclosure of sensitive information and participant experience? Findings from a quasi-experimental study. Qual Quant 2022;56(4):2341-2360 [FREE Full text] [doi: 10.1007/s11135-021-01217-4] [Medline: 34493878]
31. Rupert DJ, Poehlman JA, Hayes JJ, Ray SE, Moultrie RR. Virtual versus in-person focus groups: comparison of costs, recruitment, and participant logistics. J Med Internet Res 2017 Mar 22;19(3):e80 [FREE Full text] [doi: 10.2196/jmir.6980] [Medline: 28330832]

Edited by A Mavragani; submitted 22.09.23; peer-reviewed by K Kelly, G Guest; comments to author 24.10.23; revised version received 30.01.24; accepted 09.05.24; published 09.07.24.

Please cite as:
Squire CM, Giombi KC, Rupert DJ, Amoozegar J, Williams P
Determining an Appropriate Sample Size for Qualitative Interviews to Achieve True and Near Code Saturation: Secondary Analysis of Data
J Med Internet Res 2024;26:e52998
URL: https://www.jmir.org/2024/1/e52998
doi:10.2196/52998
PMID:38980711

©Claudia M Squire, Kristen C Giombi, Douglas J Rupert, Jacqueline Amoozegar, Peyton Williams. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Equivalence of Alcohol Use Disorder Symptom Assessments in Routine Clinical Care When Completed Remotely via Online Patient Portals Versus In Clinic via Paper Questionnaires: Psychometric Evaluation

Theresa E Matson[1,2], MPH, PhD; Amy K Lee[3], MPH; Malia Oliver[1], BA; Katharine A Bradley[1,2,4], MPH, MD; Kevin A Hallgren[1,2,5], PhD

[1]Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States
[2]Department of Health Systems and Population Health, University of Washington, Seattle, WA, United States
[3]Mental Health and Wellness Department, Kaiser Permanente Washington, Seattle, WA, United States
[4]Department of Medicine, School of Medicine, University of Washington, Seattle, WA, United States
[5]Department of Psychiatry and Behavioral Sciences, School of Medicine, University of Washington, Seattle, WA, United States

Corresponding Author:
Theresa E Matson, MPH, PhD
Kaiser Permanente Washington Health Research Institute
1730 Minor Ave
Suite 1360
Seattle, WA, 98101
United States
Phone: 1 2062872699
Email: tessa.e.matson@kp.org

Abstract

Background: The National Institute on Alcohol Abuse and Alcoholism (NIAAA) recommends the paper-based or computerized Alcohol Symptom Checklist to assess alcohol use disorder (AUD) symptoms in routine care when patients report high-risk drinking. However, it is unknown whether Alcohol Symptom Checklist response characteristics differ when it is administered online (eg, remotely via an online electronic health record [EHR] patient portal before an appointment) versus in clinic (eg, on paper after appointment check-in).

Objective: This study evaluated the psychometric performance of the Alcohol Symptom Checklist when completed online versus in clinic during routine clinical care.

Methods: This cross-sectional, psychometric study obtained EHR data from the Alcohol Symptom Checklist completed by adult patients from an integrated health system in Washington state. The sample included patients who had a primary care visit in 2021 at 1 of 32 primary care practices, were due for annual behavioral health screening, and reported high-risk drinking on the behavioral health screen (Alcohol Use Disorder Identification Test–Consumption score ≥7). After screening, patients with high-risk drinking were typically asked to complete the Alcohol Symptom Checklist—an 11-item questionnaire on which patients self-report whether they had experienced each of the 11 AUD criteria listed in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) over a past-year timeframe. Patients could complete the Alcohol Symptom Checklist online (eg, on a computer, smartphone, or tablet from any location) or in clinic (eg, on paper as part of the rooming process at clinical appointments). We examined sample and measurement characteristics and conducted differential item functioning analyses using item response theory to examine measurement consistency across these 2 assessment modalities.

Results: Among 3243 patients meeting eligibility criteria for this secondary analysis (2313/3243, 71% male; 2271/3243, 70% White; and 2014/3243, 62% non-Hispanic), 1640 (51%) completed the Alcohol Symptom Checklist online while 1603 (49%) completed it in clinic. Approximately 46% (752/1640) and 48% (764/1603) reported ≥2 AUD criteria (the threshold for AUD diagnosis) online and in clinic (P=.37), respectively. A small degree of differential item functioning was observed for 4 of 11 items. This differential item functioning produced only minimal impact on total scores used clinically to assess AUD severity, affecting total criteria count by a maximum of 0.13 criteria (on a scale ranging from 0 to 11).

XSL•FO
RenderX

Conclusions: Completing the Alcohol Symptom Checklist online, typically prior to patient check-in, performed similarly to an in-clinic modality typically administered on paper by a medical assistant at the time of the appointment. Findings have implications for using online AUD symptom assessments to streamline workflows, reduce staff burden, reduce stigma, and potentially assess patients who do not receive in-person care. Whether modality of *DSM-5* assessment of AUD differentially impacts treatment is unknown.

(J Med Internet Res 2024;26:e52101) doi:10.2196/52101

KEYWORDS

alcohol; alcohol use disorder; assessment; symptom checklist; electronic health record; patient portal; item response theory; differential item functioning; alcohol use; patient portals; in-clinic; psychometric evaluation; alcoholism; cross-sectional; United States

Introduction

Nearly 14% of adults in the United States meet criteria for alcohol use disorder (AUD) [1]. However, AUD is underdiagnosed in medical settings [2,3], which decreases opportunities for treatment. Brief screening for unhealthy alcohol use is recommended in primary care, but follow-up assessment of AUD symptoms is necessary to accurately diagnose AUD [4].

The NIAAA recommends brief population-based screening (eg, Alcohol Use Disorder Identification Test–Consumption [AUDIT-C]) and follow-up assessment using the Alcohol Symptom Checklist for patients with high-risk drinking [5]. Previous studies support the validity [6], reliability [7], and clinical utility [8] of the Alcohol Symptom Checklist administered in clinic as part of routine care. However, with increasing virtual care, it is important to test whether measurement characteristics differ across modes of administration. This study builds upon prior psychometric evaluations of the Alcohol Symptom Checklist [6,7] to evaluate whether it had similar measurement characteristics when completed online versus in clinic as part of routine clinical care.

Methods

Setting, Design, and Sample

This cross-sectional study used secondary clinical data from electronic health records (EHRs) capturing the Alcohol Symptom Checklist as it was completed during routine care at Kaiser Permanente Washington (KPWA), an integrated health care system in Washington state. KPWA conducts annual behavioral health screening for alcohol, drug use, and depression in primary care [8,9]. Clinical workflows at KPWA for population-based screening and assessment typically occur as follows, although variation in clinical practice is expected: All patients who have not completed screening in the past year are sent a behavioral health screener through the health system's online patient portal before scheduled primary care appointments as part of electronic check-in. If high-risk drinking is reported during online screening, the Alcohol Symptom Checklist automatically displays in the portal (with exceptions; Multimedia Appendix 1). Patients who do not complete screening online are typically prompted by check-in staff or medical assistants (MAs) to complete screening in clinic on paper during primary care appointments, with results entered into EHRs by MAs. If high-risk drinking is reported in clinic, the EHR prompts MAs to give patients a paper-based version of the Alcohol Symptom Checklist to complete in clinic [9]. Patients complete in-clinic screening and assessment for several reasons: some may not have set up their online patient portal, others may need translated versions in a language other than English, while others choose not to complete the checklist online. During the study period, approximately 90% of all patients with a primary care appointment completed alcohol screening and approximately 80% of primary care patients reporting high-risk drinking completed the follow-up Alcohol Symptom Checklist.

Adult patients (aged ≥18 years) were included in this study if they completed a primary care appointment (phone, video, office) during 2021 and were due for annual screening, reported high-risk drinking on the alcohol screening measure (AUDIT-C score ≥7 [10-13]; Multimedia Appendix 2), and completed the Alcohol Symptom Checklist. Checklists with missing items were excluded (Multimedia Appendix 1). If patients completed ≥1 checklist in clinical care, a single checklist was randomly selected for analysis.

Ethical Considerations

This deidentified data-only study was approved by the KPWA Health Research Institute Institutional Review Board (reference 1481289-1) with waivers of consent and Health Insurance Portability and Accountability Act authorization.

Measures

The Alcohol Symptom Checklist is a valid [6] and reliable [7] 11-item questionnaire assessing AUD criteria over a past-year timeframe, mirroring the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*; Multimedia Appendix 3) [14]. Patients self-reported whether they experienced each of 11 *DSM-5* AUD criteria (yes/no). Summed scores ranged from 0 to 11, reflecting the number of AUD criteria endorsed and informing the severity of AUD (per the *DSM-5*, mild=2-3, moderate=4-5, and severe=6-11 criteria) [14].

Modality of administration was coded as online or in clinic (Multimedia Appendix 1).

EHR-documented demographics included age, sex, race, ethnicity, and need for a language interpreter.

Analyses

Descriptive statistics characterized the demographics of patients completing online and in-clinic checklists. Logistic regression

XSL•FO
RenderX

models estimated the prevalence of each AUD criterion and of symptoms consistent with mild, moderate, or severe AUD across modalities, adjusted for demographic characteristics that significantly differed between groups.

Main analyses used item response theory [15] to evaluate differential item functioning (DIF) on checklist items across modalities (Multimedia Appendix 4 [6,14-20]).

DIF can be statistically significant without having a clinically meaningful impact on determination of AUD severity [15]. Clinicians typically make diagnostic and treatment decisions based on total scores rather than individual items because the number of criteria informs whether AUD is likely present (≥2 criteria) and its severity [14]. We evaluated the cumulative impact of DIF by estimating how total scores would be expected to change due to DIF if a patient with the same level of latent AUD completed checklists online vs in clinic (Multimedia Appendix 4 [6,14-20]).

Results

Among 3243 patients who completed the Alcohol Symptom Checklist, 1640 (51%) completed the checklist online, while 1603 (49%) completed the checklist in clinic. Among those excluded were 1 online and 63 in-clinic checklists that were missing items (Multimedia Appendix 1). Patients were predominantly middle-aged, male, White, and non-Hispanic. There were small but significant differences across modalities, with more diversity observed among those who completed checklists in clinic (Table 1).

After adjustment for demographic characteristics, item endorsement ranged from 12.4% (hazardous use) to 38.5% (physical or psychological problems) on checklists completed online and 8.3% (hazardous use) to 39.4% (larger and longer) on checklists completed in clinic. Item endorsement differed significantly by modality for 5 items (Table 2). However, prevalence of AUD and AUD severity levels did not differ by modality (P=.37 and P=.39, respectively).

DIF for online versus in-clinic modalities was observed for 4 items (Table 3). However, this had minimal impact on total scores: for patients with equal levels of latent AUD severity, estimated differences in the number of AUD criteria endorsed were expected to differ by <0.13 criteria (out of 11) due to DIF (Figure 1). In other words, DIF did not substantially impact criteria counts used by clinicians for diagnosing.

Table 1. Patient characteristics overall and across modality of Alcohol Symptom Checklist administration.

Patient characteristics	Total (N=3243), n (%)	In clinic (n=1603), n (%)	Online portal (n=1640), n (%)	*P* value
Age (years)				.001
18-24	254 (7.8)	148 (9.2)	106 (6.5)	
25-44	1528 (47.1)	707 (44.1)	821 (50.1)	
45-64	1152 (35.5)	596 (37.2)	556 (33.9)	
≥65	309 (9.5)	152 (9.5)	157 (9.6)	
Sex				.36
Female	930 (28.7)	448 (27.9)	482 (29.4)	
Male	2313 (71.3)	1155 (72.1)	1158 (70.6)	
Race				<.001
American Indian/Alaska Native	64 (2)	30 (1.9)	34 (2.1)	
Asian	193 (6)	89 (5.6)	104 (6.3)	
Black	151 (4.7)	100 (6.2)	51 (3.1)	
Native Hawaiian/Pacific Islander	53 (1.6)	30 (1.9)	23 (1.4)	
White	2271 (70)	1104 (68.9)	1167 (71.2)	
Other	88 (2.7)	58 (3.6)	30 (1.8)	
Unknown	423 (13)	192 (12)	231 (14.1)	
Ethnicity				.01
Hispanic	209 (6.4)	124 (7.7)	85 (5.2)	
Not Hispanic	2014 (62.1)	977 (60.9)	1037 (63.2)	
Unknown	1020 (31.5)	502 (31.3)	518 (31.6)	
Needs interpreter				<.001
Yes	40 (0.3)	35 (2.2)	5 (0.3)	
No	86 (2.7)	47 (2.9)	39 (2.4)	
Unknown	2117 (96.1)	1521 (94.9)	1596 (97.3)	

XSL•FO
RenderX

Table 2. Prevalence of Alcohol Symptom Checklist item endorsement across modality of administration, adjusted for demographic characteristics that differed between groups[a].

	In-clinic (n=1603), n (%; 95% CI)	Online portal (n=1640), n (%; 95% CI)	*P* value
Alcohol Symptom Checklist items			
1. Tolerance	413 (25.8; 23.7-28.0)	464 (28.2; 26.1-30.4)	.12
2. Withdrawal	245 (15.4; 13.6-17.1)	258 (15.7; 13.9-17.4)	.82
3. Larger/longer	547 (34.1; 31.8-36.4)	647 (39.4; 37.1-41.8)	.002[b]
4. Quit/control	428 (26.6; 24.4-28.7)	348 (21.3; 19.3-23.3)	<.001[b]
5. Time spent	341 (21.4; 19.4-23.4)	394 (23.9; 21.9-25.9)	.09
6. Physical/psychological problems	613 (38.5; 36.1-40.9)	502 (30.4; 28.2-32.6)	<.001[b]
7. Neglect roles	202 (12.6; 10.9-14.2)	185 (11.3; 9.8-12.9)	.28
8. Hazardous use	200 (12.4; 10.8-14.0)	135 (8.3; 6.9- 9.6)	<.001[b]
9. Social/interpersonal problems	383 (23.6; 21.5-25.7)	273 (16.9; 15.1-18.7)	<.001[b]
10. Craving	463 (29.0; 26.8-31.3)	439 (26.6; 24.5-28.8)	.13
11. Activities given up	284 (17.8; 15.9-19.7)	278 (16.9; 15.1-18.7)	.52
AUD[c] severity based on criteria endorsed			.39
No AUD (0-1 criteria)	839 (52.3; 49.9-54.8)	888 (54.1; 51.7-56.5)	
Mild AUD (2-3 criteria)	302 (18.8; 16.9-20.8)	310 (18.9; 17.0-20.8)	
Moderate AUD (4-5 criteria)	176 (10.9; 9.4-12.5)	185 (11.3; 9.8-12.9)	
Severe AUD (≥6 criteria)	286 (17.9; 16.0-19.8)	257 (15.6; 13.9-17.4)	
Number of AUD criteria[d], mean (95% CI)	2.6 (2.4-2.7)	2.4 (2.3-2.5)	.08

[a]Adjusted for patient age, race, ethnicity, and need for an interpreter.

[b]Significant at α level of .05.

[c]AUD: alcohol use disorder.

[d]Estimated using adjusted linear regression.

XSL•FO
RenderX

Table 3. Differential item functioning across modality of Alcohol Symptom Checklist administration. Discrimination parameters *a* characterize how well each item differentiates higher versus lower alcohol use disorder (AUD) severity, and severity parameters *b* characterize where, along the continuum of latent AUD severity, the item best discriminates (Multimedia Appendix 4 [6,14-20] provides additional analytic details). For the full test parameters, the latent mean and latent variance in clinic were 0.00 and 1.00, while for the online portal the values were –0.05 and 0.90. Latent means and variances were fixed to 0 and 1, respectively, for the reference group and were freely estimated for nonreference groups.

Alcohol Symptom Checklist item	In clinic[a]		Online portal[b]	
	a	*b*	*a*	*b*
1. Tolerance	1.46	0.99	__[c]	0.81
2. Withdrawal	2.19	1.25	—	—
3. Larger/longer	2.18	0.52	—	0.28
4. Quit/control	2.68	0.73	—	0.87
5. Time spent[d]	2.8	0.84	—	—
6. Physical/psychological problems[d]	2.74	0.44	—	—
7. Neglect roles[d]	3.26	1.29	—	—
8. Hazardous use	1.54	1.86	—	—
9. Social/interpersonal problems	3.08	0.81	2.65	1.07
10. Craving	2.51	0.68	—	—
11. Activities given up	3.51	1	—	—

[a]Indicates reference group.

[b]Item parameters that significantly differed (α=.0045) from the reference group are presented in the table.

[c]Item parameters that did not significantly differ from the reference group were fixed to equality to improve the power of significance testing for remaining items. These and items that were fixed as anchoring items are indicated with dashes (—).

[d]Indicates anchor item.

XSL•FO
RenderX

Figure 1. Item response theory analysis results displaying the expected number of alcohol use disorder (AUD) criteria endorsed on the Alcohol Symptom Checklist (y-axis) for each assessment modality (online versus in-clinic; separate lines) based on an individual's latent AUD severity (x-axis). Vertical distances between the plotted lines indicate the expected differences in the number of AUD criteria endorsed for a person with the same level of latent AUD severity who completed the Alcohol Symptom Checklist online vs in person. Note that the lines nearly overlap, which is reflective of the small cumulative impact of differential item functioning between the 2 assessment modalities on expected Alcohol Symptom Checklist total scores (Multimedia Appendix 4 [6,14-20] provides additional analytic details).

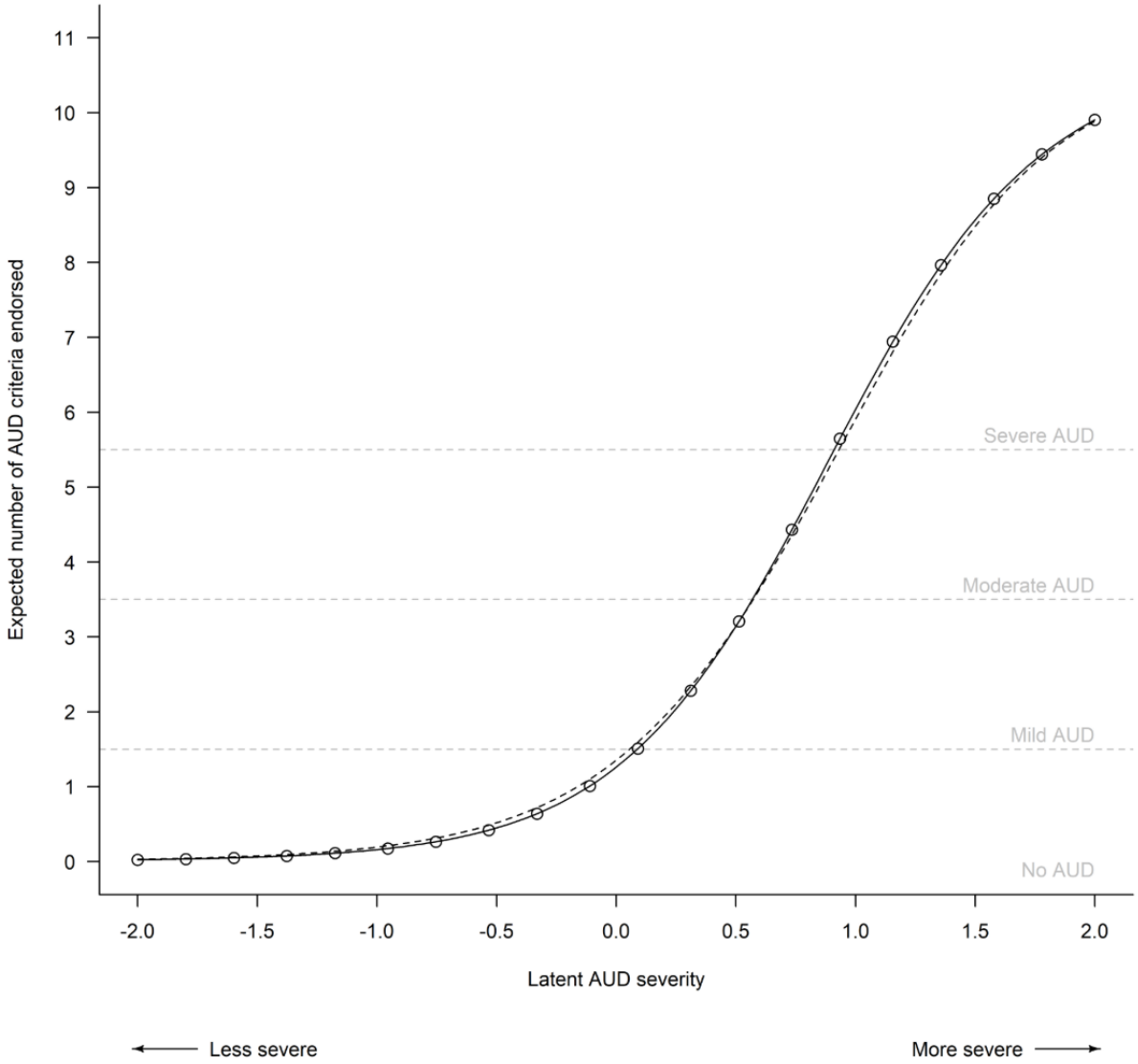

Discussion

Principal Findings

The Alcohol Symptom Checklist administered online performed similarly to the checklist completed in clinic. Small differences in the prevalence of item endorsement (<8%) did not result in differential AUD or AUD severity across modalities. While DIF across modalities was identified for 4 items, it did not meaningfully impact total scores, which are used by clinicians to diagnose AUD. The lack of clinically meaningful impact of DIF across modalities supports the use of online portals or in-clinic paper forms for administering the Alcohol Symptom Checklist in primary care to facilitate accurate AUD diagnosis.

Our findings are consistent with prior work suggesting pragmatic symptom-based checklists have high test-retest reliability [7] and perform similarly across patient demographic characteristics [6,20] and treatment status [21]. Our study adds to the growing body of evidence supporting the validity and reliability of the Alcohol Symptom Checklist completed in real-world routine care settings.

Our findings have important implications for implementing assessment in health care. In-clinic administration may be preferable for clinics that lack online resources (eg, web developers) or online infrastructure (eg, patient portals) to support integrating questionnaires into EHRs and remote workflows. When these resources are available, online administration may be preferable to help streamline workflows, reduce staff burden, and ensure more complete responses (eg, fewer missed items).

Both assessment modalities allow patients to directly report AUD symptoms (rather than staff asking the questions), which decreases stigma and improves screening quality, screening feasibility, and patient comfort [22]. Assessment of AUD symptoms remotely via an online portal may further reduce

XSL•FO
RenderX

stigma and increase comfort in self-reporting alcohol-related concerns [23].

The heterogeneity of patients who completed the Alcohol Symptom Checklist online versus in clinic is an important finding that underscores the need for both modalities to avoid inequitable identification of AUD. Young adults (aged 18-24 years), older adults (aged ≥65 years), patients of minoritized racial and ethnic backgrounds, and patients who needed an interpreter were more likely to complete checklists in clinic. As digital tools become increasingly common in health care [24], offering both modalities may help minimize inequities. For example, online portals may be more accessible to patients who do not access in-person services (eg, due to logistical, financial, health status, or other factors), while the in-clinic modality may have greater reach for patients without reliable internet or digital devices, who have lower digital literacy, who need translated materials, who do not use patient portals, and who may be more comfortable responding when they can ask questions or explain responses to a health care provider [25].

Limitations

This study has limitations. Patients may have underestimated or underreported their alcohol use and AUD symptoms. Racial and ethnic diversity were limited. The sample included patients who completed annual screening at a primary care appointment and thus represent the population that uses care but may not represent other groups that do not receive primary care. Results may not generalize to other populations or health systems. The study has noteworthy strengths. While processes for completing the Alcohol Symptom Checklist online versus in clinic differed, these differences reflect actual clinical assessment procedures, providing high external validity regarding how checklists perform under real-world conditions. The sample size was large and powered to detect small amounts of measurement invariance. Future studies should assess whether modality of AUD assessment differentially impacts diagnosis and treatment.

Conclusion

Remote completion of the Alcohol Symptom Checklist online by primary care patients performed similarly to a valid [6] and reliable [7] paper-based version completed in clinic. Future studies should evaluate methods to support health care providers in following and treating patients with AUD through in-person and virtual care.

Acknowledgments

The authors wish to thank Megan Addis, research project manager at Kaiser Permanente Washington Health Research Institute, who provided organizational and administrative support that was integral to this study. The authors also wish to acknowledge the clinicians and patients at Kaiser Permanente Washington who contributed data to this study. Research reported in this publication was supported by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health (NIH) under award numbers R21AA028073 and R33AA028073. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAAA or NIH.

Authors' Contributions

TEM conceptualized the study design and methods, led formal data analysis, and was involved in all aspects of results interpretation, manuscript preparation and writing; AKL contributed to results interpretation, and provided revisions to the manuscript. MO led data acquisition and programming, contributed to results interpretation, and provided revisions to the manuscript. KAB acquired funding for the study, contributed to study conceptualization, contributed to results interpretation, and provided revisions to the manuscript. KAH acquired funding for the study, contributed to study conceptualization, consulted on methods, contributed to results interpretation, helped with manuscript preparation and writing, and provided revisions to the manuscript. All authors have approved the final article for submission. The views and opinions expressed in this manuscript are those of the authors only and do not necessarily represent the views, official policy or position of Kaiser Permanente, the US Department of Health and Human Services or any of its affiliated institutions or agencies.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study sample selection.
[DOCX File , 42 KB - jmir_v26i1e52101_app1.docx]

Multimedia Appendix 2
Alcohol Use Disorders Identification Test-Consumption (AUDIT-C).
[DOCX File , 141 KB - jmir_v26i1e52101_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Alcohol Symptom Checklist.
[DOCX File , 259 KB - jmir_v26i1e52101_app3.docx]

Multimedia Appendix 4
Additional analytical details.
[DOCX File , 30 KB - jmir_v26i1e52101_app4.docx]

References

1. Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry 2015 Aug;72(8):757-766 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.0584] [Medline: 26039070]
2. Williams EC, Rubinsky AD, Lapham GT, Chavez LJ, Rittmueller SE, Hawkins EJ, et al. Prevalence of clinically recognized alcohol and other substance use disorders among VA outpatients with unhealthy alcohol use identified by routine alcohol screening. Drug Alcohol Depend 2014 Feb 01;135:95-103. [doi: 10.1016/j.drugalcdep.2013.11.016] [Medline: 24360928]
3. Hallgren KA, Witwer E, West I, Baldwin L, Donovan D, Stuvek B, et al. Prevalence of documented alcohol and opioid use disorder diagnoses and treatments in a regional primary care practice-based research network. J Subst Abuse Treat 2020 Mar;110:18-27 [FREE Full text] [doi: 10.1016/j.jsat.2019.11.008] [Medline: 31952624]
4. US Preventive Services Task Force, Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, et al. Screening and behavioral counseling interventions to reduce unhealthy alcohol use in adolescents and adults: US Preventive Services Task Force Recommendation statement. JAMA 2018 Nov 13;320(18):1899-1909. [doi: 10.1001/jama.2018.16789] [Medline: 30422199]
5. National Institute on Alcohol Abuse and Alcoholism. Qeios. URL: https://www.qeios.com/read/YFAE2Y [accessed 2024-06-26]
6. Hallgren KA, Matson TE, Oliver M, Witkiewitz K, Bobb JF, Lee AK, et al. Practical assessment of alcohol use disorder in routine primary care: performance of an alcohol symptom checklist. J Gen Intern Med 2022 Jun;37(8):1885-1893 [FREE Full text] [doi: 10.1007/s11606-021-07038-3] [Medline: 34398395]
7. Hallgren KA, Matson TE, Oliver M, Caldeiro RM, Kivlahan DR, Bradley KA. Practical assessment of DSM-5 alcohol use disorder criteria in routine care: High test-retest reliability of an Alcohol Symptom Checklist. Alcohol Clin Exp Res 2022 Mar;46(3):458-467 [FREE Full text] [doi: 10.1111/acer.14778] [Medline: 35275415]
8. Lee AK, Bobb JF, Richards JE, Achtmeyer CE, Ludman E, Oliver M, et al. Integrating alcohol-related prevention and treatment into primary care: a cluster randomized implementation trial. JAMA Intern Med 2023 Apr 01;183(4):319-328 [FREE Full text] [doi: 10.1001/jamainternmed.2022.7083] [Medline: 36848119]
9. Sayre M, Lapham GT, Lee AK, Oliver M, Bobb JF, Caldeiro RM, et al. Routine assessment of symptoms of substance use disorders in primary care: prevalence and severity of reported symptoms. J Gen Intern Med 2020 Apr;35(4):1111-1119 [FREE Full text] [doi: 10.1007/s11606-020-05650-3] [Medline: 31974903]
10. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Alcohol Use Disorders Identification Test. Arch Intern Med 1998 Sep 14;158(16):1789-1795. [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
11. Bradley KA, Bush KR, Epler AJ, Dobie DJ, Davis TM, Sporleder JL, et al. Two brief alcohol-screening tests from the Alcohol Use Disorders Identification Test (AUDIT): validation in a female Veterans Affairs patient population. Arch Intern Med 2003 Apr 14;163(7):821-829. [doi: 10.1001/archinte.163.7.821] [Medline: 12695273]
12. Bradley KA, DeBenedetti AF, Volk RJ, Williams EC, Frank D, Kivlahan DR. AUDIT-C as a brief screen for alcohol misuse in primary care. Alcohol Clin Exp Res 2007 Jul;31(7):1208-1217. [doi: 10.1111/j.1530-0277.2007.00403.x] [Medline: 17451397]
13. Frank D, DeBenedetti AF, Volk RJ, Williams EC, Kivlahan DR, Bradley KA. Effectiveness of the AUDIT-C as a screening test for alcohol misuse in three race/ethnic groups. J Gen Intern Med 2008 Jun;23(6):781-787 [FREE Full text] [doi: 10.1007/s11606-008-0594-0] [Medline: 18421511]
14. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychological Association; 2013.
15. Chalmers RP, Counsell A, Flora DB. It might not make a big DIF: improved differential test functioning statistics that account for sampling variability. Educ Psychol Meas 2016 Feb;76(1):114-140 [FREE Full text] [doi: 10.1177/0013164415584576] [Medline: 29795859]
16. Chalmers RP. mirt: A Multidimensional Item Response Theory Package for the R Environment. J Stat Soft 2012;48(6):1-29. [doi: 10.18637/JSS.V048.i06]
17. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7(1):23-35 [FREE Full text] [doi: 10.1007/s40271-013-0041-0] [Medline: 24403095]
18. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000 Sep;38(9 Suppl):II28-II42 [FREE Full text] [doi: 10.1097/00005650-200009002-00007] [Medline: 10982088]

XSL•FO
RenderX

19. Kopf J, Zeileis A, Strobl C. Anchor selection strategies for DIF analysis: review, assessment, and new approaches. Educ Psychol Meas 2015 Feb;75(1):22-56 [FREE Full text] [doi: 10.1177/0013164414529792] [Medline: 29795811]
20. Matson TE, Hallgren KA, Lapham GT, Oliver M, Wang X, Williams EC, et al. Psychometric performance of a substance use symptom checklist to help clinicians assess substance use disorder in primary care. JAMA Netw Open 2023 May 01;6(5):e2316283 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.16283] [Medline: 37234003]
21. Williams EC, Matson TE, Hallgren KA, Oliver M, Wang X, Bradley KA. Assessing substance use disorder symptoms with a checklist among primary care patients with opioid use disorder and/or long-term opioid treatment: an observational study. J Gen Intern Med 2024 Jul 01. [doi: 10.1007/s11606-024-08845-0] [Medline: 38954321]
22. Bodner E, Palgi Y, Wyman MF. Ageism in mental health assessment and treatment of older adults. In: Ayalon L, Tesch-Romer C, editors. Contemporary Perspectives on Ageism. Cham, Switzerland: Springer; 2018.
23. Choy MA, Sturgiss E, Goodyear-Smith F, Smith GJ. Digital health tools and patients with drug use disorders: qualitative patient experience study of the Electronic Case-Finding and Help Assessment Tool (eCHAT). J Med Internet Res 2020 Sep 14;22(9):e19256 [FREE Full text] [doi: 10.2196/19256] [Medline: 32924959]
24. Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC. Digital transformation in healthcare: Analyzing the current state-of-research. J Bus Res 2021 Feb;123:557-567. [doi: 10.1016/j.jbusres.2020.10.030]
25. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]

Abbreviations

AUD: alcohol use disorder
AUDIT-C: Alcohol Use Disorder Identification Test–Consumption
DIF: differential item functioning
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
EHR: electronic health record
KPWA: Kaiser Permanente Washington
MA: medical assistant
NIAAA: National Institute on Alcohol Abuse and Alcoholism

Edited by S Ma; submitted 23.08.23; peer-reviewed by V Palzes, M Johansson, G Humphreys; comments to author 11.01.24; revised version received 23.02.24; accepted 07.06.24; published 22.07.24.

Please cite as:
Matson TE, Lee AK, Oliver M, Bradley KA, Hallgren KA
Equivalence of Alcohol Use Disorder Symptom Assessments in Routine Clinical Care When Completed Remotely via Online Patient Portals Versus In Clinic via Paper Questionnaires: Psychometric Evaluation
J Med Internet Res 2024;26:e52101
URL: https://www.jmir.org/2024/1/e52101
doi:10.2196/52101
PMID:

©Theresa E Matson, Amy K Lee, Malia Oliver, Katharine A Bradley, Kevin A Hallgren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Contextual Acceptance of COVID-19 Mitigation Mobile Apps in the United States: Mixed Methods Survey Study on Postpandemic Data Privacy

Yuanyuan Feng[1], PhD; Brad Stenger[1], MS, MSJ; Shikun Zhang[2], PhD

[1]Department of Computer Science, University of Vermont, Burlington, VT, United States
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States

Corresponding Author:
Yuanyuan Feng, PhD
Department of Computer Science
University of Vermont
85 South Prospect Street
Burlington, VT, 05405
United States
Phone: 1 802 656 3475
Email: yuanyuan.feng@uvm.edu

Abstract

Background: The COVID-19 pandemic gave rise to countless user-facing mobile apps to help fight the pandemic ("COVID-19 mitigation apps"). These apps have been at the center of data privacy discussions because they collect, use, and even retain sensitive personal data from their users (eg, medical records and location data). The US government ended its COVID-19 emergency declaration in May 2023, marking a unique time to comprehensively investigate how data privacy impacted people's acceptance of various COVID-19 mitigation apps deployed throughout the pandemic.

Objective: This research aims to provide insights into health data privacy regarding COVID-19 mitigation apps and policy recommendations for future deployment of public health mobile apps through the lens of data privacy. This research explores people's contextual acceptance of different types of COVID-19 mitigation apps by applying the privacy framework of contextual integrity. Specifically, this research seeks to identify the factors that impact people's acceptance of data sharing and data retention practices in various social contexts.

Methods: A mixed methods web-based survey study was conducted by recruiting a simple US representative sample (N=674) on Prolific in February 2023. The survey includes a total of 60 vignette scenarios representing realistic social contexts that COVID-19 mitigation apps could be used. Each survey respondent answered questions about their acceptance of 10 randomly selected scenarios. Three contextual integrity parameters (attribute, recipient, and transmission principle) and respondents' basic demographics are controlled as independent variables. Regression analysis was performed to determine the factors impacting people's acceptance of initial data sharing and data retention practices via these apps. Qualitative data from the survey were analyzed to support the statistical results.

Results: Many contextual integrity parameter values, pairwise combinations of contextual integrity parameter values, and some demographic features of respondents have a significant impact on their acceptance of using COVID-19 mitigation apps in various social contexts. Respondents' acceptance of data retention practices diverged from their acceptance of initial data sharing practices in some scenarios.

Conclusions: This study showed that people's acceptance of using various COVID-19 mitigation apps depends on specific social contexts, including the type of data (attribute), the recipients of the data (recipient), and the purpose of data use (transmission principle). Such acceptance may differ between the initial data sharing and data retention practices, even in the same context. Study findings generated rich implications for future pandemic mitigation apps and the broader public health mobile apps regarding data privacy and deployment considerations.

(J Med Internet Res 2024;26:e57309) doi:10.2196/57309

XSL•FO
RenderX

KEYWORDS

data privacy; health privacy; COVID-19; mobile apps; contextual integrity; respiratory; infectious; pulmonary; pandemic; mobile app; app; apps; digital health; digital technology; digital intervention; digital interventions; smartphone; smartphones; mobile phone

Introduction

Background

To combat the COVID-19 pandemic, countless user-facing mobile apps ("COVID-19 mitigation apps") have been developed and deployed around the world [1,2]. These apps provide various functionalities, including contact tracing [3], symptom self-checking [4], test result reporting (eg, iHealth COVID-19 Test [iHealth Labs Inc]), and proof of vaccination (eg, Excelsior [New York State]). However, low adoption rates often undermine the real-world impact of these apps [5], partially caused by people's data privacy concerns around how these apps handle various sensitive personal data, including medical records and phone-based location data [6,7].

Public health priorities have changed throughout the COVID-19 pandemic [8], resulting in evolving needs for pandemic mitigation apps. Some apps deployed early in the pandemic with narrow functionality (eg, symptom self-checking) became obsolete, while other apps are likely to remain useful after the pandemic (eg, test result reporting). The United States ended the COVID Public Health Emergency on May 11, 2023 [9]. It is time to comprehensively examine user privacy across different types of COVID-19 mitigation apps developed in this pandemic. Specifically, we aim to investigate 3 types of COVID-19 mitigation apps with broad utility in future public health crises, namely, contact tracing, test result reporting, and proof of vaccination.

We conducted a vignette survey study that applied the privacy framework of contextual integrity [10] to examine people's acceptance of how COVID-19 mitigation apps could handle their personal health data in various social contexts. This research informs the design and deployment of future public health apps and relevant mandates around these apps from a privacy-centric perspective. Insights from this research could encourage the adoption of future public health apps by addressing people's data privacy concerns in diverse social contexts.

Prior Work: Data Privacy and the Adoption of COVID-19 Mitigation Apps

Research examined the adoption of mobile technologies for pandemic mitigation worldwide, which primarily focused on contact tracing apps [3,11-14]. Generally, adoption rates are high for COVID-19 mitigation apps backed by government mandates, such as Aarogya Setu (National Informatics Centre, Government of India) and HaMagen (Health Ministry, Israel), but low for apps relying on users' voluntary installation (eg, COVID-19 exposure alerts in the United States). A multicountry study shows that people in collective cultures are more likely to adopt contact tracing apps than those in strong individualistic cultures [15].

Among all the factors impacting user adoption of COVID-19 mitigation apps [13,14,16-18], data privacy stood out as a major concern [3,6,7,15,19,20]. Recent qualitative and mixed methods studies also unpacked individual users' complex decision-making process, including data privacy considerations when adopting COVID-19 mitigation apps [21,22].

However, most studies heavily focused on initial adoption but overlooked the continued use or future use of these apps after the pandemic. In addition, many studies only examined a single app type [3,6,7,21,22], neglecting the fact that multiple types of apps can be used together to fight the pandemic. Our study bridges these gaps by examining 3 major types of COVID-19 mitigation apps with long-term utility in future public health crises.

Another gap in the literature was the retention and reuse of personal data collected by COVID-19 mitigation apps. Data retention is central to privacy law discussion [23,24] and affects users' willingness to share personal information [25]. Currently, there is no clear guidance on how various data collected by these apps should be retained, while controversial reuse of such data by law enforcement has been reported [26,27]. Misuse of retained data from COVID-19 mitigation apps can have dire privacy consequences. Our study provides a new understanding of users' attitudes toward COVID-19 mitigation apps' data retention.

Theory: The Privacy Framework of Contextual Integrity

People's data privacy attitudes toward computing technologies differ in various contexts [28-30], including who collects the data, the purpose of the data collection, and the specific ways the collected data will be used, processed, or shared. Hargittai et al [18] found that US participants' willingness to adopt contact tracing apps differs by the providers of the apps; Zhang et al [31] revealed that people's acceptance of vaccination certificates varies in diverse real-world scenarios. However, traditional privacy theories (eg, public or private dichotomy) fail to consider the contexts for computing technologies' complex data practices [32,33].

Privacy as contextual integrity, conceptualized by Nissenbaum [10], is a rising privacy framework to examine computing technologies' highly complex data practices. Contextual integrity defines privacy as the appropriate flow of information that follows established information norms in a particular society or culture. According to contextual integrity, information norms can be captured in 5 contextual integrity parameters: sender (who sends the data), recipient (who receives the data), subject (whose data), attribute (type of information), and transmission principle (TP thereafter, conditions of the flow). Empirically, we can observe information norms from people's privacy attitudes toward different information flows exhibited in various data practices.

XSL•FO
RenderX

Compared to prior privacy theories, the contextual integrity framework is advantageous in identifying potential privacy violations typically when 1 or more of the parameters deviate from an established information norm. For example, it might be considered appropriate for a store owner (recipient) to collect vaccination information (attribute) from a customer (sender and subject) before letting them into the store (TP). However, if the business owner were to use this information for advertising purposes or keep the data indefinitely, the resulting flows, with different TPs, would deviate from the established information norms.

Built upon insights from prior studies leveraging contextual integrity to examine the information flows for public health apps [15,31,34], this study rigorously applies the contextual integrity framework to examine people's acceptance of COVID-19 mitigation apps in highly diverse contexts.

Research Questions and Hypotheses

We aim to answer 2 research questions (RQs) and 4 specific hypotheses by considering applicable contextual integrity parameters and respondents' demographics.

- RQ1: What factors impact people's acceptance of sharing data via COVID-19 mitigation apps?
  - Null hypothesis 1.1 (NH1.1): There is no difference in people's acceptance of sharing health data via COVID-19 mitigation apps, even if contextual integrity parameters are different.
  - Null hypothesis 1.2 (NH1.2): There is no difference in people's acceptance of sharing data via COVID-19 mitigation apps, even if they have different demographic backgrounds.
- RQ2: What factors impact people's acceptance of their data being retained via COVID-19 mitigation apps?
  - Null hypothesis 2.1 (NH2.1): There is no difference in people's acceptance of their data being retained via COVID-19 mitigation apps, even if contextual integrity parameters are different.
  - Null hypothesis 2.2 (NH2.1): There is no difference in people's acceptance of their data being retained via COVID-19 mitigation apps, even if they have different demographic backgrounds.

Methods

Variables and Measurements

We designed a contextual integrity–based survey instrument according to our RQ and hypotheses. The outcome variable is "perceived acceptance," a proxy to measure people's privacy attitudes toward COVID-19 mitigation apps. For RQ1, we use a 5-point Likert scale question to capture different acceptance levels because people's privacy preferences are often nonbinary [35]. For RQ2, we chose not to specify a retention time due to the evolving COVID-19 guidelines (eg, how long a person should isolate after getting COVID-19). Instead, we measure acceptance using a 3-level categorical variable to capture necessary nuances.

The independent variables are applicable contextual integrity parameters and some demographic features self-reported by respondents. According to contextual integrity, it is crucial to include all 5 contextual integrity parameters to comprehensively evaluate the appropriateness of information flow, so we follow a standard contextual integrity–based template shown in Table 1 to construct survey questions including all parameters. We stabilized 2 contextual integrity parameters ([sender] and [subject] are "you" and "your," which refer to the survey respondent) and varied the values of 3 contextual integrity parameters ([attribute], [recipient], and [TP]) as our independent variables.

Table 1. Example survey questions for 1 vignette scenario.

Question numbers	Example survey questions
Contextual integrity–based template	Would it be acceptable or unacceptable for [Sender] to share [Subject's] [Attribute] with [Recipient] for [Transmission Principle (TP)]?
Question 1	Via a smartphone mobile app, would it be acceptable or unacceptable for *you* [Sender] to share *your* [Subject's] *recent COVID test result* [Attribute] with *your employer* [Recipient] to work in person [TP]?
Question 1 follow-up	Please briefly explain your choice to the previous question (1-2 sentences).
Question 2	In the same scenario above, if *you* [Sender] shared *your* [Subject] *recent COVID test result* [Attribute] via a smartphone mobile app, would it be acceptable or unacceptable for *your employer* [Recipient] to *keep your data* [TP]?
Question 2 follow-up	Please briefly explain your choice to the previous question (1-2 sentences).

Survey Design

We designed a mixed methods survey using vignette scenarios, which are effective to study people's beliefs [36] and conduct survey-based experiments [37]. Vignette scenarios are widely used in privacy research to capture people's contextual privacy attitudes toward digital technologies [30,38-41], enabling us to examine the diverse use cases of COVID-19 mitigation apps in real-world contexts. To craft realistic vignette scenarios, we conducted a technology review of COVID-19 mitigation apps available in the United States and gathered news articles about other COVID-19 migration apps deployed around the world. We used the scenarios from our recently published research [31] and slightly modified them to match the scope of this study. Next, we chose a subset of scenarios that would apply to contact tracing, proof of vaccination, and test result reporting apps and then finalized all the values of the 3 contextual integrity parameters, as shown in Textbox 1.

XSL•FO
RenderX

We then applied a full factorial design across the 3 contextual integrity parameters in Textbox 1 to generate 60 distinct scenarios. Each scenario contains 2 main questions corresponding to the 2 RQs: the first question presents the scenario to gauge acceptance and the second question examines the acceptance toward data retention by the recipient in the scenario. Note that contextual integrity considers data retention as a type of TP, so both questions conform to the standard contextual integrity–based template.

Textbox 1. All contextual integrity parameter values in our survey (text in italics format indicates baseline values in our models).

Attribute: (1) *Recent COVID-19 test results*, (2) up-to-date COVID-19 vaccination records, and (3) COVID-19 exposure status from phone-based contact tracing.

Recipient: (1) Employers, (2) *essential stores (grocery stores, etc)*, (3) nonessential stores (department stores, etc), (4) dine-in locations (restaurants, bars, etc), (5) large public venues (stadium, music festival, etc), (6) health care providers (hospitals, clinics, specialists, etc), (7) airlines operating international flights, (8) airlines operating domestic flights, (9) domestic long-distance transportation operators (trains, Greyhound, Megabus, etc), and (10) local public transit operators (subways, metros, commuter trains, buses, etc).

Transmission principle: (1) *Public health purpose* and (2) scenario-specific purpose (eg, gain access to [location] or use [service]).

Survey Questionnaire

The survey questionnaire began with a consent page, which confirmed participant eligibility and obtained informed consent. Then, the questionnaire displayed 2 information pages, including plain language definitions of COVID-19–related terminology and examples of some COVID-19 mitigation apps deployed in the United States. At the end of the information pages, we asked respondents to confirm their understanding of the information. Failure to confirm did not disqualify respondents, but their responses were excluded from data analysis.

Afterward, the questionnaire displayed 10 randomly selected vignette scenarios to limit the study time to around 10 minutes to minimize survey fatigue [42]. As shown in Table 1, each scenario has 2 contextual integrity–based questions. In 2 of 10 vignette scenarios, we inserted 2 follow-up free-text questions to gather additional qualitative data to understand respondents' rationale. Finally, the questionnaire ended with a set of culturally responsible demographic questions for the US population and additional background questions on their COVID-19 experience.

For attention check, we avoided conventional attention check questions that are irrelevant to the survey because they are ineffective and may undermine researcher-participant trust [43]. Instead, we used multiple metrics as attention and accuracy checks, including the time spent on each scenario and the quality of free-text responses.

Ethical Considerations

This survey study was approved by the institutional review board at the University of Vermont under the exempt category, where the written consent form is waived. We obtained informed consent through a consent page displayed at the beginning of the survey. All survey responses were deidentified by the Prolific platform Academic Ltd and used for data collection. Except obvious survey abuse, all respondents were each compensated US $2 via the Prolific platform (average compensation rate=US $10.43 per hour). To ensure research credibility, we preregistered this study in the Open Science Framework (Center for Open Science, Inc) before data collection, which is available to the public [44].

Recruitment and Data Collection

We implemented the survey questionnaire in the Qualtrics XM platform (Qualtrics, LLC) for web-based distribution and chose the Prolific platform (Prolific Academic Ltd) to recruit respondents to maximize data collection speed and quality. We ran 3 small pilot surveys in December 2022 and January 2023, which led to minor refinements of the questionnaire. Next, we distributed the finalized survey in February 2023 and recruited a US representative sample via Prolific [45]. We received 694 completed survey responses and excluded 2 responses due to obvious survey abuse. The remaining 692 survey respondents were each compensated US $2 according to our study protocol.

We performed a 2-step data cleaning procedure using a customized Python script, followed by manual inspection. This led to the exclusion of 18 responses due to the lack of substance in the free-text answers (n=12) and the nonsensical content of the free-text answers (n=6). We included the remaining 674 survey responses in the final analysis.

Statistical Analysis

Overview

Because our outcome variables for both RQs are ordinal data, multilevel ordered logistic regression is suitable to examine the effects of multiple predators. We used the cumulative link mixed models (clmm) in the ordinal package in R (R Foundation for Statistical Computing) [46] with random effects to account for repeated measures in the survey.

Predictors

We selected the most promising variables as predictors to construct our models, as follows:

Main Effect Predictors

We include all values of attribute, recipient, and TP, as well as 6 demographic variables (age, gender, political leaning, living areas, education, and income), and 4 COVID-19 experience questions (installation of contact tracing apps, being tested positive for COVID-19, being vaccinated against COVID-19, and knowing someone who got seriously ill or passed away from COVID-19) as main effect predictors.

XSL•FO
RenderX

Effect Modifiers

We account for the interactions among contextual integrity parameters by including effect modifications to the estimates of individual main effects. We include all pairwise combinations of the 3 contextual integrity parameters (ie, attribute*recipient, recipient*TP, and TP*attribute) as effect modifiers in our models. These effect modifiers enable us to interpret special cases when 2 specific contextual integrity parameters are present.

Predictor Baselines

For each predictor and effect modifier, we set the baseline value to the most common option according to the US population's familiarity. For example, we chose COVID-19 test results as the baseline for attribute because app-based contact tracing and digital vaccination records were less commonly adopted in the United States. The baseline values for contextual integrity parameters are set in italics format in Table 1.

Final Models

We constructed 2 ordered logistic regression models for 2 RQs. Our models are inherently complex due to the large number of main effect predictors and effect modifiers. To ensure model convergence, we combined data categories in some demographic variables (eg, age groups and political leaning) to reduce model complexity. We also converted "not sure" and "prefer not to answer" responses to the "not available (N/A)" placeholder value to maximize model convergence. Our models used the "ucminf" optimizer in place of the package default optimizer "nlminb." Both are commonly used optimizer algorithms to maximize the marginal likelihood function.

Results Reporting and Interpretation

We followed the best practices outlined in Taylor et al [47] to report and interpret our model results. For statistically significant predictors ($P<.05$), we reported the odds ratios (ORs) of the effect, where OR is the natural exponent of the model estimate (β) for a predictor. We also report the 95% CIs of ORs.

For main effect interpretation, if a predictor value has an OR of 2, it means respondents are twice as likely to accept the scenario when compared to the baseline value of the predictor. To interpret the interactions among 2 contextual integrity parameters, we must calculate the marginal ORs of the main effect predictors and effect modifiers using the formula: $e^{(\beta_{main_effect_predictor}+\beta_{effect_modifier})}$. For example, if the estimate for a main effect predictor value is 2 (OR 7.39), and the estimate of the effect modifier containing a second predictor value is 0.5, then the marginal OR is $e^{(2+0.5)}=12.18$. This means the presence of the second predictor value increased the OR of the main effect predictor value from 7.39 to 12.18 compared to the main effect predictor baseline value. We also calculated the marginal 95% CIs for these interactions.

Each participant answered free-text follow-up questions for 2 randomly selected scenarios. After excluding low-quality free-text responses, we collected 2679 responses in total: 1340 for question 1 (44.7 per scenario) and 1339 for question 2 (44.6 per scenario). We performed a simplified qualitative analysis to synthesize notable themes from these responses and then selected representative quotes to further explain or support our statistical results.

Results

Respondents and Descriptive Statistics

We recruited a simple US representative sample (N=674) via Prolific that resembles US Census data across age, gender, and race or ethnicity (Multimedia Appendix 1). We present the overall descriptive statistics of question 1 and question 2 responses in Figures 1-6.

Figures 1-3 depict respondents' acceptance levels for sharing their data via COVID-19 mitigation apps in all the vignette scenarios. Generally, data sharing is mostly acceptable, as respondents found data sharing "acceptable" or "somewhat acceptable" in 61.5% (4147/6740) of all scenarios (referred to as "overall acceptance" thereafter). Across all attributes, sharing is most acceptable when recipients are health care providers (538/657, 81.9%), employers (484/692, 69.9%), and airlines operating international flights (515/675, 76.3%). Regarding attributes, respondents were less comfortable sharing COVID-19 exposure status from contact tracing (1284/2233, 57.5% overall acceptance) than sharing vaccination records (1435/2283, 62.9% overall acceptance) or test results (1428/2224, 64.2% overall acceptance).

Figures 4-6 show respondents' acceptance levels for their health data being retained by recipients in all vignette scenarios. Across all attributes and recipients, considerable numbers of respondents felt data retention was "unacceptable" (2947/6740, 43.7% of all scenarios) or "acceptable, only for a limited amount of time" (3010/6740, 44.7% of all scenarios). Only in 11.6% (783/6740) of all scenarios, respondents reported recipients retaining their health data as "acceptable, no matter how long it will be kept" (or acceptable without a time limit). Across all recipients, health care providers retaining health data is most accepted, but only in 37.1% (244/657) of health care provider scenarios, respondents felt data retention was acceptable without a time limit. Similarly, small percentages of respondents found data retention without a time limit acceptable across all attributes: 9.6% (214/2233) of scenarios for contact tracing exposure status, 13.8% (314/2283) of scenarios for vaccination records, and 11.4% (255/2224) of scenarios for test results. The only outliner is health care providers retaining vaccination records, where 46.6% (103/221) responses to this scenario were acceptable without a time limit. Overall, question 2 results are largely consistent with question 1 results in terms of recipients, while retention time differentiates respondents' acceptance levels drastically.

XSL•FO
RenderX

Figure 1. Descriptive statistics of question 1 responses: acceptance of sharing recent COVID-19 test results in different vignette scenarios organized by recipients.

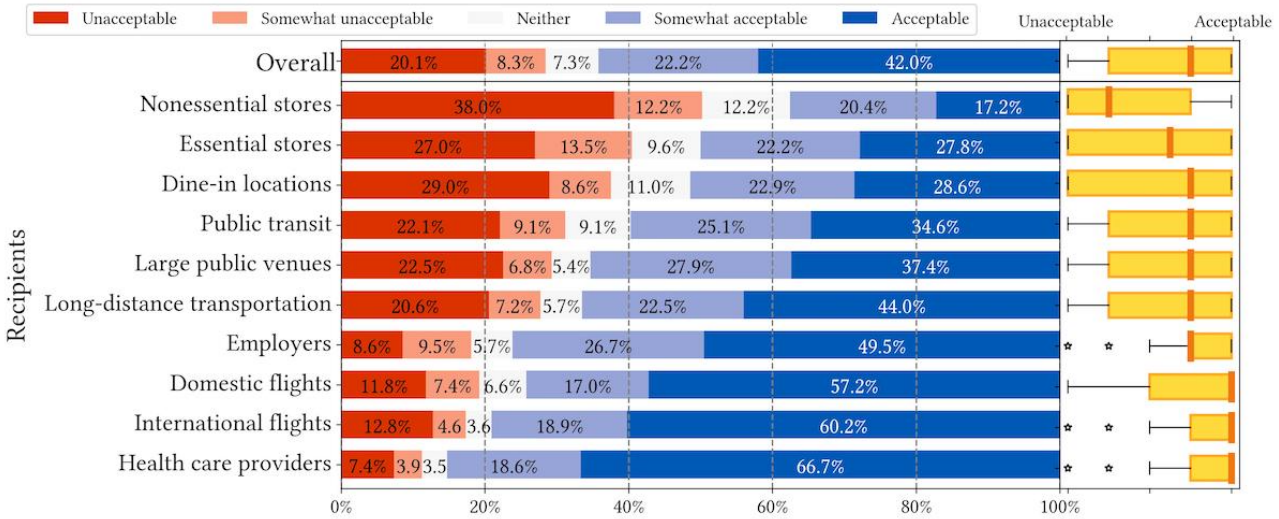

Figure 2. Descriptive statistics of question 1 responses: acceptance of sharing up-to-date COVID-19 vaccination records in different vignette scenarios organized by recipients.

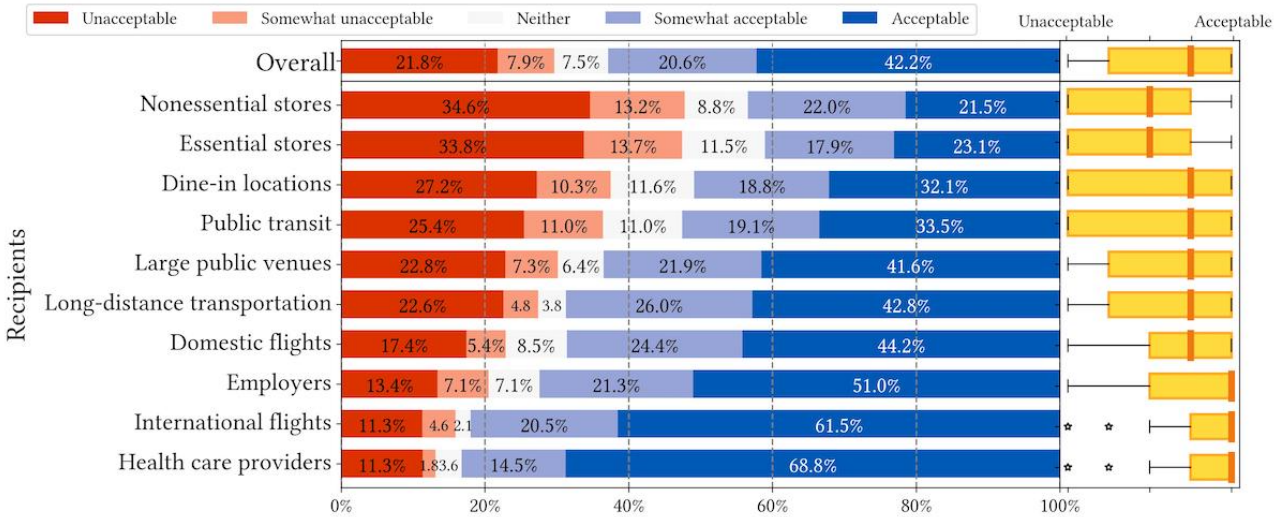

Figure 3. Descriptive statistics of question 1 responses: acceptance of sharing COVID-19 exposure status from phone-based contact tracing in different vignette scenarios organized by recipients.

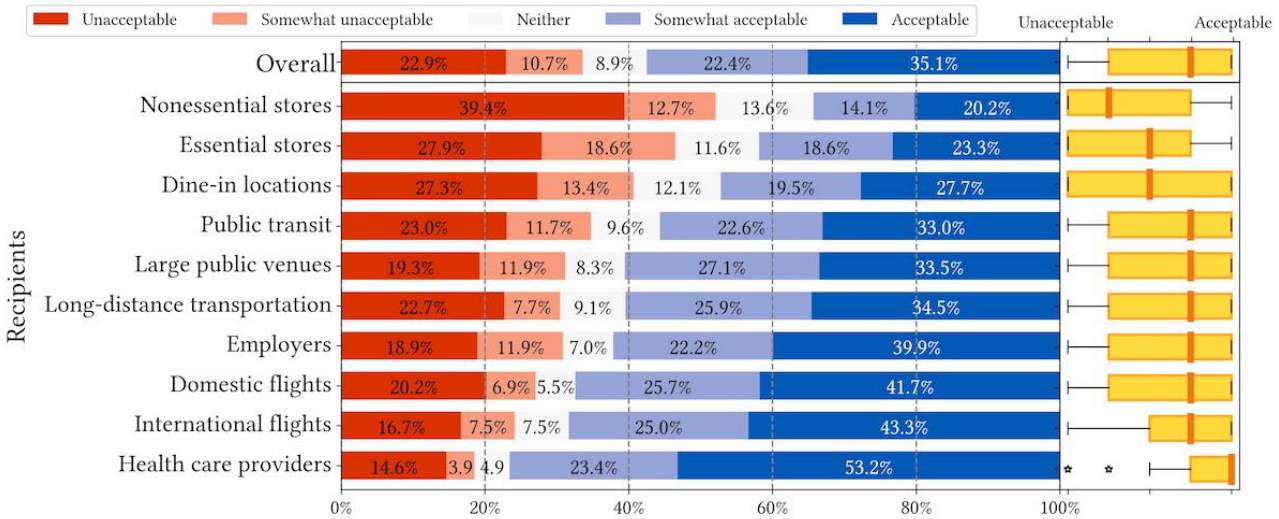

XSL•FO
RenderX

Figure 4. Descriptive statistics of question 2 responses: acceptance of recent COVID-19 test results being retained in different vignette scenarios organized by recipients.

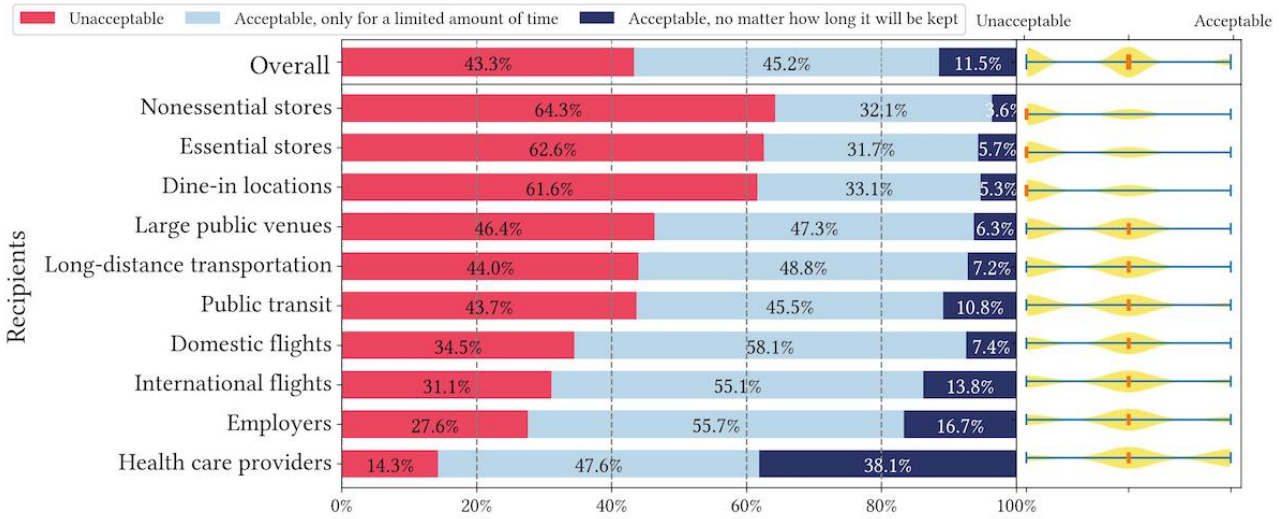

Figure 5. Descriptive statistics of question 2 responses: acceptance of up-to-date COVID-19 vaccination records being retained in different vignette scenarios organized by recipients.

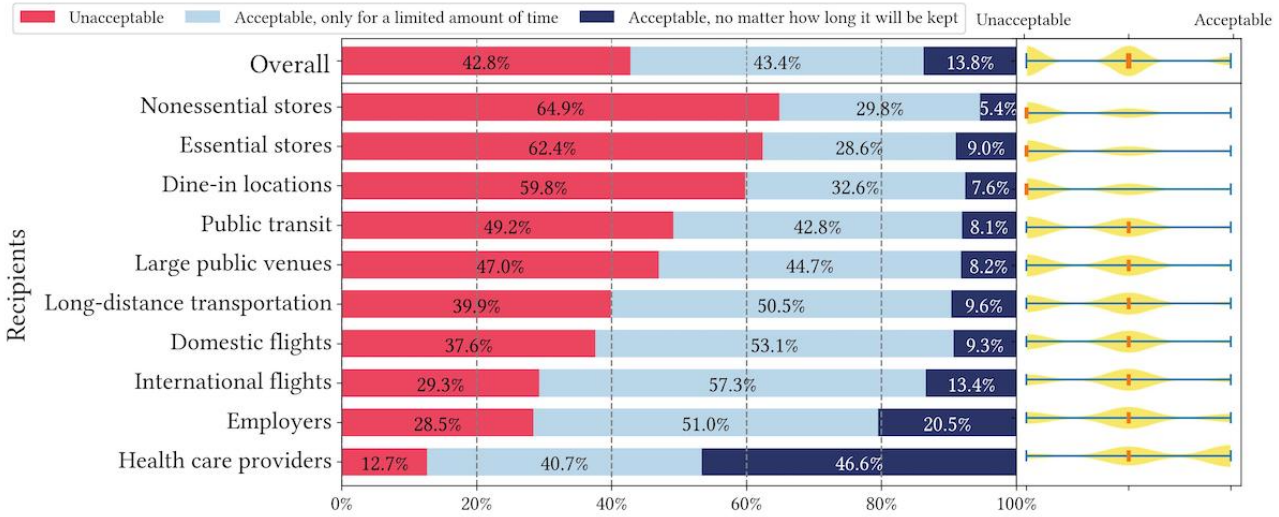

Figure 6. Descriptive statistics of question 2 responses: acceptance of COVID-19 contact tracing exposure status being retained in different vignette scenarios organized by recipients.

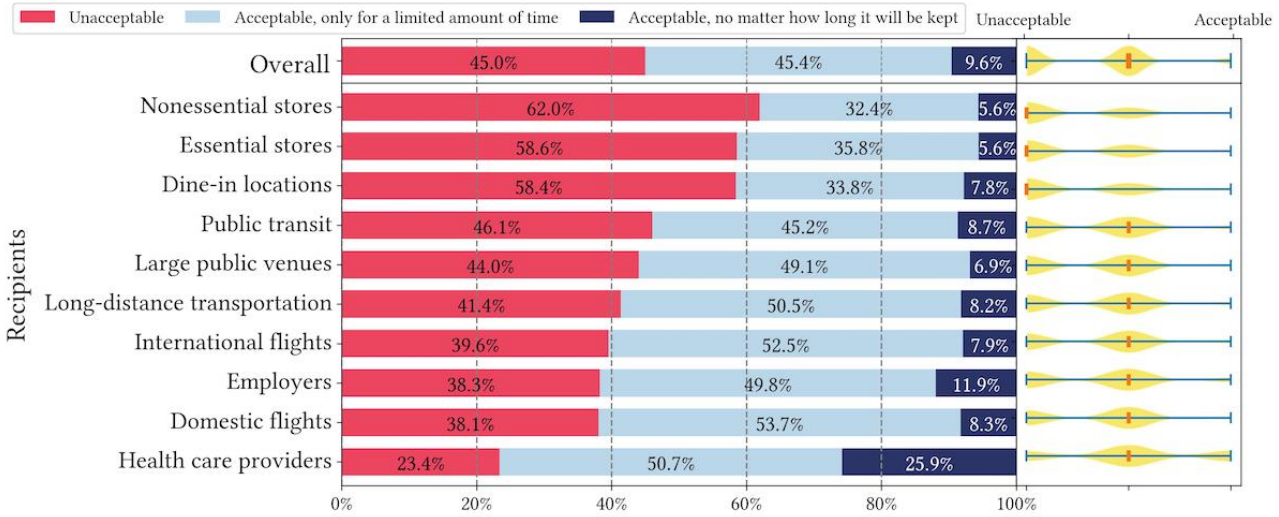

XSL•FO
RenderX

RQ1: Acceptance of Data Sharing Through COVID-19 Mitigation Apps

Contextual Integrity Parameters

As shown in Table 2, recipient is the most powerful contextual integrity parameter to predict respondents' acceptance levels toward COVID-19 mitigation apps. The recipients with the largest effects were health care providers (OR 146.36, 95% CI 78.06-274.42; *P*<.001), employers (OR 10.94, 95% CI 6.37-18.78; *P*<.001), and airlines operating domestic (OR 11.34, 95% CI 6.53-19.69; *P*<.001) and international flights (OR 16.03, 95% CI 8.88-28.93; *P*<.001), where respondents were more willing to use mobile apps to share their personal health information with them compared to the baseline recipient of essential stores. Respondents' free-text responses revealed that the most accepted recipients are in scenarios that could critically impact the health of others and themselves. For example, respondents commented that sharing data with health providers "is necessary to keep track of the virus ... also necessary to keep myself and others healthy," sharing data with airlines "would help keep travelers safe," and sharing data with employers "is important for my coworkers to feel safe when I come to work." Additionally, large public venues (OR 4.13, 95% CI 2.45-6.97; *P*<.001), local public transit operators (OR 3.12, 95% CI 1.84-5.27; *P*<.001), and long-distance transportation operators (OR 5.87, 95% CI 3.42-10.09; *P*<.001) were also more tolerated than the baseline. In contrast, sharing data via mobile apps with nonessential stores (OR 0.35, 95% CI 0.21-0.59; *P*<.001) was less accepted than the baseline. Respondents' comments explained the reasons: "Nonessential stores do not need this data." "That's too drastic. You can't deny people access to food and supplies." Only dine-in locations had no statistically significant effect.

However, no attribute or TP value has a significant effect. Respondents did not report significant differences in acceptance levels toward sharing contact tracing exposure status or proof of vaccination from test results. Similarly, scenario-specific purposes are not statistically different from generic public health purposes. Note that we group all scenario-specific purposes into 1 TP value to ensure model convergence, which means our model may miss some TP nuances in specific scenarios.

Table 2. Research question 1 (RQ1) model coefficients table of main effect predictors (contextual integrity parameters).

RQ1 main effect predictors (contextual integrity parameters)	Estimate β	OR[a] (95% CI)	*P* value
Attribute			
Test results	.00	1.00 (—[b])	—
Exposure status	.07	1.08 (0.68-1.72)	.76
Vaccination records	–.17	0.84 (0.53-1.34)	.47
Transmission principles			
Public health	.00	1.00 (—)	—
Scenario-specific	.08	1.09 (0.73-1.62)	.69
Recipient			
Essential stores	.00	1.00 (—)	—
Dine-in locations	.28	1.32 (0.80-2.21)	.28
Domestic flights	2.42	11.34 (6.53-19.68)	<.001
Employers	2.39	10.94 (6.37-18.78)	<.001
Health care providers	4.98	146.36 (78.09-274.32)	<.001
International flights	2.77	16.03 (8.88-28.92)	<.001
Large public venues	1.41	4.13 (2.45-6.97)	<.001
Long-distance transportation	1.77	5.87 (3.42-10.09)	<.001
Nonessential stores	–1.05	0.35 (0.21-0.59)	<.001
Public transit	1.13	3.12 (1.84-5.27)	<.001

[a]OR: odds ratio.

[b]Not available for baseline values.

Interactions Among Contextual Integrity Parameters

Our model reveals notable interactions among contextual integrity parameters, where some combinations of 2 contextual integrity parameter values yield worth-noting results in certain special cases, as shown in Table 3.

When considering recipient and attribute values pairwise, although respondents are willing to share their health information across all attributes with health care providers, the OR dropped from 146.36 to 28.22 when the attribute is contact tracing exposure status (marginal OR 28.22, 95% CI 11.57-68.84; *P*<.001). There was a similar drop in OR when

XSL•FO
RenderX

sharing contact tracing exposure status with airlines operating international flights (marginal OR 6.82, 95% CI 2.85-16.32; *P*=.01). In the free-text responses of these scenarios, most respondents expressed that sharing exposure status was useful to keep health care facilities and other passengers safe, but a few doubted its necessity at the late stage of this pandemic since "now we have vaccines." The qualitative data indicate relatively low acceptance of sharing contact tracing exposure status via mobile apps when other effective pandemic mitigation measures are available, even with the most acceptable recipients.

The combination of attribute and TP values shows that respondents are slightly less likely to share contract tracing exposure status than test results (marginal OR 0.73, 95% CI 0.71-0.75; *P*=.01) when TP is scenario-specific purposes (eg, entering a place or using a service). Many comments pointed out the questionable accuracy of phone-based contact tracing exposure status, as one respondent wrote: "I don't trust the standards developed to determine exposure, I don't trust the network designed to measure exposure, and I don't trust the data on the other end of the transaction (others I might have been exposed to)." The qualitative data further support respondents' overall reluctance to share contact tracing data, especially for nonpublic health purposes.

XSL•FO
RenderX

Table 3. Research question 1 (RQ1) model coefficients table of effect modifiers (interactions).

RQ1 effector modifiers (interactions)	Estimate β	OR[a] (95% CI)	*P* value
Attribute×recipient			
Exposure status×dine-in locations	.18	1.21 (0.65-2.25)	.56
Vaccination records×dine-in locations	.40	1.50 (0.80-2.79)	.21
Exposure status×domestic flights	–.40	0.68 (0.35-1.30)	.24
Vaccination records×domestic flights	.15	1.16 (0.61-2.21)	.64
Exposure status×employers	–.61	0.54 (0.29-1.04)	.07
Vaccination records×employers	.55	1.75 (0.91-3.36)	.10
Exposure status×health care providers	–1.64	0.20 (0.10-0.40)	<.001
Vaccination records×health care providers	–.30	0.74 (0.35-1.57)	.44
Exposure status×international flights	–.85	0.43 (0.22-0.85)	.02
Vaccination records×international flights	.35	1.43 (0.71-2.88)	.32
Exposure status×large public venues	–.27	0.77 (0.41-1.46)	.42
Vaccination records×large public venues	.01	1.01 (0.54-1.91)	.97
Exposure status×long-distance transportation	–.58	0.56 (0.29-1.08)	.09
Vaccination records×long-distance transportation	.12	1.13 (0.58-2.18)	.72
Exposure status×nonessential stores	.17	1.20 (0.63-2.28)	.59
Vaccination records×nonessential stores	.32	1.39 (0.73-2.62)	.32
Exposure status×public transit	–.14	0.87 (0.46-1.64)	.67
Vaccination records×public transit	.18	1.20 (0.65-2.24)	.56
Attribute×TP[b]			
Exposure status×scenario-specific	–.39	0.68 (0.51-0.92)	.01
Vaccination records×scenario-specific	–.29	0.75 (0.55-1.01)	.06
TP×recipient			
Scenario-specific×dine-in locations	–.02	0.98 (0.59-1.63)	.94
Scenario-specific×domestic flights	.02	1.02 (0.60-1.73)	.94
Scenario-specific×employers	.06	1.07 (0.63-1.81)	.81
Scenario-specific×health care providers	–.53	0.59 (0.33-1.07)	.08
Scenario-specific×international flights	.54	1.72 (0.99-2.98)	.05
Scenario-specific×large public venues	–.02	0.99 (0.59-1.67)	.97
Scenario-specific×long-distance transportation	–.05	0.96 (0.56-1.64)	.88
Scenario-specific×nonessential stores	.38	1.46 (0.86-2.47)	.16
Scenario-specific×public transit	.01	1.02 (0.61-1.71)	.94

[a]OR: odds ratio.

[b]TP: transmission principle.

Demographics and COVID-19 Experiences

As shown in Table 4, political leaning is the only statistically significant demographic predictor. Compared to self-identified political moderates, self-identified liberals (OR 4.85, 95% CI 2.44-9.61; *P*<.001) are more likely to accept using COVID-19 mitigation apps. Age, living areas, education, and income did not significantly affect respondents' acceptance levels. Interestingly, some prior COVID-19 experiences also turned out to be significant. Respondents who have installed COVID-19 contact tracing apps (OR 5.01, 95% CI 2.34-10.73; *P*<.001) and who have received COVID-19 vaccination (OR 10.54, 95% CI 4.92-22.56; *P*<.001) were generally more acceptable about data sharing via COVID-19 mitigation apps. These indicate that people's familiarity with COVID-19 mitigation apps and their attitude toward vaccination may play a role in their acceptance of COVID-19 mitigation apps.

XSL•FO
RenderX

Table 4. Research question 1 (RQ1) model coefficients table of demographic and COVID-19 predictors.

RQ1 demographic and COVID-19 predictors	Estimate β	OR[a] (95% CI)	*P* value
Sex			
Female	.00	1.00 (—[b])	—
Male	–.17	0.85 (0.48-1.50)	.57
Nonbinary	–.09	0.92 (0.12-7.06)	.94
Age (years)			
60+	.00	1.00 (—)	—
18-29	.16	1.18 (0.48-2.88)	.72
30-39	.13	1.14 (0.47-2.81)	.77
40-49	.55	1.75 (0.70-4.34)	.23
50-59	–.20	0.82 (0.35-1.90)	.64
Politics			
Moderate	.00	1.00 (—)	—
Conservative	–.79	0.46 (0.20-1.03)	.06
Liberal	1.57	4.85 (2.45-9.61)	<.001
Living area			
Town or suburb	.00	1.00 (—)	—
City	.25	1.30 (0.71-2.39)	.40
Rural area	.44	1.55 (0.65-3.72)	.32
Education			
High school	.00	1.00 (—)	—
College	–.76	0.47 (0.21-1.08)	.07
Grad school	–.30	0.75 (0.26-2.18)	.59
Income (US $)			
50,000-99,999	.00	1.00 (—)	—
25,000-49,999	.22	1.26 (0.59-2.67)	.55
Less than 25,000	.52	1.68 (0.72-3.93)	.28
More than 100,000	.64	1.91 (0.89-4.12)	.10
COVID -19 test			
No	.00	1.00 (—)	—
Yes	.39	1.48 (0.82-2.66)	.19
COVID -19 app			
No	.00	1.00 (—)	—
Yes	1.61	5.01 (2.34-10.73)	<.001
COVID -19 vaccination			
No	.00	1.00 (—)	—
Yes	2.35	10.54 (4.93-22.55)	<.001
COVID -19 illness			
No	0.00	1.00 (—)	—
Yes	.54	1.72 (0.98-3.03)	.06

[a]OR: odds ratio.

[b]Not available for baseline values.

XSL•FO
RenderX

RQ2: Acceptance of Data Retention Through COVID-19 Mitigation Apps

Contextual Integrity Parameters

As shown in Table 5, we included 2 contextual integrity parameters (recipient and attribute) as predictors in the RQ2 model because data retention practices are considered TP. Recipient remains the most impactful contextual integrity parameter for respondents' acceptance of data retention by COVID-19 mitigation apps. Compared to the baseline of essential stores, all other recipients are more accepted if they retain respondents' health data from various COVID-19 mitigation apps. The largest effect remains in health care providers (OR 251.77, 95% CI 172.17-368.16; *P*<.001), followed by employers (OR 24.34, 95% CI 24.21-24.48; *P*<.001) and airlines operating international flights (OR 16.36, 95% CI 10.89-24.57; *P*<.001). These results resemble those of RQ1, showing that respondents' acceptance of data retention practices primarily depends on recipients. Notably, the free-text responses reveal nuanced opinions on retention time. For example, even with the most accepted recipient health care providers, one respondent felt sharing test results was "only acceptable for the duration of the medical treatment," and another commented on retaining exposure status: "After a point, it will become outdated and useless, and should be deleted for privacy reasons." The qualitative data suggest that the acceptance of data retention practices is associated with the necessity of specific scenarios.

Regarding attribute, respondents expressed slightly greater acceptance of having their vaccination records (OR 1.69, 95% CI 1.68-1.70; *P*<.001) and their contact tracing exposure status (OR 1.57, 95% CI 1.56-1.58; *P*<.001) retained by recipients compared to the baseline of COVID-19 test results, which differs from RQ1 results. However, the effect is small, and there were no notable themes in the free-text responses to clearly explain such differences.

Table 5. Research question (RQ2) model coefficients table of main effect predictors (contextual integrity parameters).

RQ2 main effect predictors (contextual integrity parameters)	Estimate β	OR[a] (95% CI)	*P* value
Attribute			
Test results	.00	1.00 (—[b])	—
Exposure status	.45	1.57 (1.56-1.58)	<.001
Vaccination records	.52	1.69 (1.68-1.70)	<.001
Recipient			
Essential stores	.00	1.00 (—)	—
Dine-in locations	.53	1.71 (1.15-2.54)	.008
Domestic flights	2.25	9.50 (9.45-9.55)	<.001
Employers	3.19	24.34 (24.21-24.48)	<.001
Health care providers	5.52	251.77 (172.17-368.16)	<.001
International flights	2.79	16.36 (10.89-24.57)	<.001
Large public venues	1.36	3.93 (2.66-5.81)	<.001
Long-distance transportation	2.03	7.67 (5.08-11.59)	<.001
Nonessential stores	.17	1.20 (1.19-1.20)	<.001
Public transit	1.78	5.96 (5.92-5.99)	<.001

[a]OR: odds ratio.

[b]Not available for baseline values.

Interactions Among Contextual Integrity Parameters

As shown in Table 6, many combinations of attribute and recipient values show that respondents' acceptance of data retention through COVID-19 mitigation apps dropped significantly, demonstrating the potential inappropriateness for certain recipients to retain their phone-based contact tracing exposure status and their vaccination records.

When attribute is phone-based contact tracing exposure status, respondents' acceptance of data retention significantly decreased for 7 of 10 recipients, with the largest drop in the most accepted recipients including health care providers (marginal OR 48.91, marginal 95% CI 29.04-82.41; *P*<.001), employers (marginal OR 6.11, marginal 95% CI 3.64-10.25; *P*<.001), airliners operating international flights (marginal OR 6.82, marginal 95% CI 3.72-11.56; *P*<.001). In the free-text responses, one respondent found it acceptable for employers to retain exposure status data when "the virus was transmittable," while another believed exposure status "is not needed now in 2023" for air travel. There are smaller acceptance decreases for airlines operating domestic flights (marginal OR 5.47, marginal 95% CI 3.22-9.29; *P*<.001), long-distance transportation operators (marginal OR 3.53, marginal 95% CI 2.02-6.16; *P*=.008), and local public transit operators (marginal OR 3.49, marginal 95% CI 2.05-5.95; *P*<.001). Notably, the acceptance of nonessential stores retaining data (marginal OR 0.91, marginal 95% CI 0.52-1.62; *P*<.001) dropped even below that of the baseline as one respondent commented: "I don't see the need for

XSL•FO
RenderX

nonessential stores to keep my private data." These results resonate with RQ1: people were less comfortable with data practices around phone-based contact tracing data due to lack of necessity.

When attribute is COVID-19 vaccination records, data retention acceptance decreased for 4 recipients, with large drops for local public transit operators (marginal OR 3.19, marginal 95% CI 1.89-5.38; *P*=.001), large public venues (marginal OR 2.25, marginal 95% CI 1.32-3.82; *P*=.043), and nonessential stores (marginal OR 0.64, marginal 95% CI 0.36-1.12; *P*=.005) and a small drop for employers (marginal OR 22.87, marginal 95% CI 13.80-37.92; *P*<.001). The free-text responses revealed that many respondents believed vaccination records were medical records and "shouldn't be disclosed and even less stored or kept." Others worried about potential data breaches after data retention that "leaked information about a person's medical history and choices." Another respondent found it acceptable for large public venues to keep it, but "it should be deleted after the event has ended and resubmitted for every event." This qualitative evidence suggests the inappropriateness for some everyday services and venues to retain people's vaccination records due to the medical nature of the data.

Table 6. Research question (RQ2) model coefficients table of effect modifiers (interactions).

RQ2 effect modifiers (interactions)	Estimate β	OR[a] (95% CI)	*P* value
Attribute×recipient			
Exposure status×dine-in locations	–.28	0.76 (0.44-1.31)	.32
Vaccination records×dine-in locations	–.37	0.69 (0.40-1.19)	.19
Exposure status×domestic flights	–.55	0.58 (0.58-0.58)	<.001
Vaccination records×domestic flights	–.18	0.84 (0.58-1.21)	.35
Exposure status×employers	–1.38	0.25 (0.25-0.25)	<.001
Vaccination records×employers	–.06	0.95 (0.94-0.95)	<.001
Exposure status×health care providers	–1.63	0.20 (0.12-0.33)	<.001
Vaccination records×health care providers	–.01	1.00 (0.59-1.69)	.99
Exposure status×international flights	–.95	0.39 (0.23-0.66)	<.001
Vaccination records×international flights	–.35	0.71 (0.42-1.20)	.20
Exposure status×large public venues	–.28	0.76 (0.45-1.29)	.31
Vaccination records×large public venues	–.55	0.58 (0.34-0.98)	.04
Exposure status×long-distance transportation	–.77	0.47 (0.27-0.82)	.008
Vaccination records×long-distance transportation	–.34	0.71 (0.41-1.24)	.23
Exposure status×nonessential stores	–.26	0.77 (0.77-0.78)	<.001
Vaccination records×nonessential stores	–.62	0.54 (0.35-0.83)	.005
Exposure status×public transit	–.53	0.59 (0.59-0.59)	<.001
Vaccination records×public transit	–.62	0.54 (0.37-0.79)	.001

[a]OR: odds ratio.

Demographics and COVID-19 Experiences

As shown in Table 7, several demographic predictors have significant effects in the RQ2 model. Similar to RQ1, self-identified liberals (OR 2.03, 95% CI 2.01-2.04; *P*<.001) are slightly more likely to accept recipients to retain their health data via a mobile app. We also observe a significant but small effect for respondents aged 30 to 39 years (OR 1.03, 95% CI 1.02-1.03; *P*<.001) compared to those aged 60 years and older. In the opposite direction, compared to those with high school education, respondents with college education are less acceptable toward data retention practices (OR 0.77, 95% CI 0.77-0.78; *P*<.001). Similarly, we found a significant but smaller effect for respondents with self-reported annual income of US $25,000 to US $49,999 (OR 0.99, 95% CI 0.99-1.00; *P*=.003) compared to those with self-reported annual income of US $50,000 to US $99,999. Regarding COVID-19 experiences, respondents who have installed COVID-19 contact tracing apps (OR 3.39, 95% CI 1.77-6.48; *P*<.001), who have received COVID-19 vaccination (OR 4.74, 95% CI 4.71-4.77; *P*<.001), and who knew someone who got seriously ill or passed away (OR 1.84, 95% CI 1.16-2.90; *P*=.009) are more likely to accept data retention practices by various recipients.

XSL•FO
RenderX

Table 7. Research question (RQ2) model coefficients table of demographic and COVID-19 predictors.

RQ2 demographic and COVID-19 predictors	Estimate β	OR[a] (95% CI)	*P* value
Sex			
Female	.00	1.00 (—[b])	—
Male	.15	1.17 (0.72-1.89)	.53
Nonbinary	.64	1.91 (0.33-11.04)	.47
Age (years)			
60+	.00	1.00 (—)	—
18-29	.60	1.82 (0.95-3.50)	.07
30-39	.02	1.03 (1.02-1.03)	<.001
40-49	.50	1.65 (0.82-3.32)	.16
50-59	–.18	0.84 (0.44-1.57)	.58
Politics			
Moderate	.00	1.00 (—)	—
Conservative	–.50	0.61 (0.33-1.12)	.11
Liberal	.70	2.03 (2.01-2.04)	<.001
Living			
Town or suburb	.00	1.00 (—)	—
City	–.10	0.91 (0.54-1.52)	.71
Rural area	.05	1.06 (0.50-2.25)	.88
Education			
High school	.00	1.00 (—)	—
College	–.26	0.77 (0.77-0.78)	<.001
Grad school	.16	1.18 (0.60-2.29)	.64
Income (US $)			
50,000-99,999	.00	1.00 (—)	—
25,000-49,999	–.01	0.99 (0.99-1.00)	.003
Less than 25,000	–.15	0.87 (0.46-1.63)	.66
More than 100,000	–.02	0.98 (0.54-1.80)	.96
COVID -19 test			
No	.00	1.00 (—)	—
Yes	.14	1.16 (0.70-1.91)	.57
COVID -19 app			
No	.00	1.00 (—)	—
Yes	1.22	3.39 (1.77-6.48)	<.001
COVID -19 vaccination			
No	.00	1.00 (—)	—
Yes	1.55	4.74 (4.71-4.77)	<.001
COVID -19 illness			
No	.00	1.00 (—)	—
Yes	.60	1.84 (1.16-2.90)	.009

[a]OR: odds ratio.

[b]Not available for baseline values.

XSL•FO
RenderX

Summary of Results

For RQ1, we discovered that the contextual integrity parameter recipient has the most significant effect on how willing respondents are to share data through COVID-19 mitigation apps and that certain combinations of recipient, attribute, and TP values also influence their willingness. These results, taken together, reject NH1.1. Among demographic variables, self-reported political leaning and prior COVID-19 experiences impact respondents' overall acceptance, rejecting NH1.2.

For RQ2, the contextual integrity parameter recipient greatly influences respondents' acceptance of data retention practices via COVID-19 mitigation apps. The contextual integrity parameter attributes also turned out to be significant, as respondents felt that retaining their COVID-19 test results was less acceptable. Moreover, many combinations of attribute and recipient values further impact respondents' acceptance of data retention practices via COVID-19 mitigation apps. These results, taken together, reject NH2.1. Regarding demographics, besides respondents' political leaning and prior COVID-19 experiences as in RQ1 results, age, self-reported education, and annual income also have significant effects on respondents' acceptance of data retention, thereby rejecting NH2.2.

Discussion

Contributions of This Study

By surveying a US representative sample, this study confirms prior research that data privacy is a key factor impacting the adoption of COVID-19 mitigation apps [3,6,7,15,19,20] and that the specific contexts impact people's privacy attitudes toward these apps [15,31,34]. Different from prior studies that only examined user-centered data privacy for one type of COVID-19 mitigation apps [3,6,7,21,22], our study revealed US respondents' varying acceptance of different types of COVID-19 mitigation apps that collect different types of personal health data. This highlights the importance of holistically examining multiple COVID-19 mitigation apps after the pandemic dust settles, providing insights into the deployment of different types of pandemic mitigation apps to combat future public health crises.

Besides exploring the general acceptance of sharing personal health data via COVID-19 mitigation apps, this study dived into people's attitudes toward data retention practices via these apps—a critical consideration in privacy research [23-25]. In many scenarios, we found US respondents' acceptance differed from their acceptance of initial data sharing, highlighting the importance of thoroughly considering the data retention practices of public health mobile apps.

Finally, this study contributes to the broader application of the contextual integrity privacy framework to analyze how contextual data privacy impacts the acceptance of COVID-19 mitigation apps, showing the framework's suitability to gauge people's acceptance of pandemic mitigation apps and other public health technologies.

Limitations and Future Research

We acknowledge several limitations of the study and suggest future research directions. First, this study inherits the limitations of the web-based survey methodology, where we relied on respondents' self-reported data. We mitigated this by focusing on respondents' self-reported attitudes rather than using attitudes to predict actual behaviors. Future research could explore data on people's actual behaviors with COVID-19 mitigation apps to triangulate our study findings.

Second, we minimized the survey sampling bias by recruiting a US representative sample on Prolific. Still, our findings cannot fully represent the US population's perspective or generalize beyond the United States. To provide comparative analyses, we encourage researchers to replicate our contextual integrity–based survey methods to examine people's acceptance of COVID-19 mitigation apps in other countries or regions. This will elucidate potentially different contextual integrity information norms regarding pandemic mitigation apps across cultures.

Third, though this survey depicts US respondents' contextual acceptance of various COVID-19 mitigation apps at the end of the pandemic, it does not show longitudinal trends about people's privacy attitudes toward these apps. However, researchers can quickly adjust and deploy the contextual integrity–based survey methods developed in this study longitudinally, should future public health research needs arise. This way, future research could generate longitudinal insights into the acceptance of public health technologies using a consistent contextual integrity–based survey instrument.

Furthermore, we only evaluated the most important set of contextual factors according to the contextual integrity framework and the app types being examined, due to the quantitative nature of the survey and the goal of generating meaningful statistical results. In addition, though the qualitative data collected from the free-text questions enhanced our statistical results, the short survey completion time limited the depth of such data. We encourage researchers to explore a richer set of contextual factors and consider alternative research methods to generate deeper qualitative findings.

Deployment Strategies for Future Pandemic Mitigation Apps

This study provides a cross-sectional overview of people's contextual acceptance of 3 major types of COVID-19 mitigation apps after the pandemic in the United States, which informs the deployment strategies for future pandemic mitigation apps. Our results backed up prior contextual integrity–inspired studies [15,31,34] that contextual integrity parameters, especially recipient, attribute, and TP, can predict people's acceptance of COVID-19 mitigation apps in different situations. This survey's fully factorial design enabled us to extend these prior studies by investigating interactions among contextual integrity parameters. We found the recipient to be the most influential contextual integrity parameter, while the interplay among multiple contextual integrity parameters in specific scenarios complicated people's acceptance levels. This implies that there are no one-size-fits-all data privacy norms for COVID-19 mitigation apps and that the specific contexts for data practices,

XSL•FO
RenderX

often determined by a combination of contextual integrity parameters, matter. In the case of a future pandemic, we recommend public health policy makers proactively collect evidence-based data about the general public's contextual acceptance of using various pandemic mitigation apps to inform their decisions on when, where, and how to deploy these apps.

Our statistical models for both RQs revealed that political leaning and prior COVID-19 experiences also influenced people's acceptance of COVID-19 mitigation apps. Specifically, respondents who self-identified as liberals, have downloaded COVID-19 mitigation apps, and have been vaccinated against COVID-19 generally reported higher acceptance levels. The RQ2 model also yielded a few significant demographic predictors (eg, age, income, and education) with smaller effects. These results suggest that public health policy makers must consider how population demographics could affect the deployment of future pandemic mitigation apps. Additionally, other public health efforts, such as general education on vaccination and publicity for public health technologies, may boost the overall acceptance of future pandemic mitigation apps.

Implications for Future Public Health Mobile Apps

By articulating the contextual acceptance of various COVID-19 mitigation apps deployed in this pandemic, this study yielded rich implications for future public health mobile apps. Built upon prior studies focusing on one type of COVID-19 mitigation app [3,6,7,21,22], this study compared 3 types of apps (ie, contact tracing, test result reporting, and proof of vaccination) that have long-term utility in future public health crises. Our results suggested that people were least comfortable with apps that perform phone-based contact tracing, contrasting the relatively high acceptance and adoption of contact tracing apps early in the pandemic [3]. One explanation is that, as more effective mitigation tools (such as vaccination) become widely available in the United States near the end of the pandemic, people may not perceive phone-based contact tracing as necessary or appropriate, especially given its data privacy implications. This highlights the need for policy makers to strategize what public health mobile apps to promote according to the changing pandemic stages and available mitigation methods.

Our findings also shed light on the nuanced implications of the data retention practices of COVID-19 mitigation apps. This study provided initial evidence that people's acceptance of data retention diverged from their acceptance of initial data sharing in specific scenarios, where many respondents felt data should not be retained indefinitely even for the most accepted recipients (eg, health care providers). As we are phasing out the COVID-19 mitigation apps, it is critical to re-evaluate how these apps retain data collected during the pandemic to minimize downstream privacy harms, such as breaches of retained data.

Last but not least, our findings informed how to appropriately deploy future public health mobile apps in the face of emergencies and crises. The different acceptance levels across our vignette scenarios suggest the importance of considering people's contextual acceptance of public health apps' data practices. Deployment of public health apps should start with necessary scenarios that the general public finds acceptable and avoid controversial scenarios.

Conclusions

This study systematically applies the contextual integrity framework to examine people's contextual data privacy attitudes toward multiple types of COVID-19 mitigation apps that emerged during the COVID-19 pandemic. It confirmed that contextual integrity parameters can help predict people's acceptance of using these apps in various realistic scenarios, yield novel evidence on the acceptance of data retention practices by these apps, and generate rich implications for deploying future public health mobile apps.

Acknowledgments

This research is funded by the University of Vermont through the Office of the Vice President for Research Express grant. The authors thank Yan Shvartzshnaider and Maria Sckolnick for their guidance.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic distributions of the survey sample.
[DOCX File , 23 KB - jmir_v26i1e57309_app1.docx]

References

1. Lee B, Ibrahim SA, Zhang T. Mobile apps leveraged in the COVID-19 pandemic in East and South-East Asia: review and content analysis. JMIR Mhealth Uhealth 2021 Nov 11;9(11):e32093 [FREE Full text] [doi: 10.2196/32093] [Medline: 34748515]
2. Pandit JA, Radin JM, Quer G, Topol EJ. Smartphone apps in the COVID-19 pandemic. Nat Biotechnol 2022 Jul;40(7):1013-1022. [doi: 10.1038/s41587-022-01350-x] [Medline: 35726090]
3. Kolasa K, Mazzi F, Leszczuk-Czubkowska E, Zrubka Z, Péntek M. State of the art in adoption of contact tracing apps and recommendations regarding privacy protection and public health: systematic review. JMIR Mhealth Uhealth 2021 Jun 10;9(6):e23250 [FREE Full text] [doi: 10.2196/23250] [Medline: 34033581]

XSL•FO
RenderX

4. Schmeelk S, Davis A, Li Q, Shippey C, Utah M, Myers A, et al. Monitoring symptoms of COVID-19: review of mobile apps. JMIR Mhealth Uhealth 2022 Jun 01;10(6):e36065 [FREE Full text] [doi: 10.2196/36065] [Medline: 35609313]
5. Jenniskens K, Bootsma MCJ, Damen JAAG, Oerbekke MS, Vernooij RWM, Spijker R, et al. Effectiveness of contact tracing apps for SARS-CoV-2: a rapid systematic review. BMJ Open 2021;11(7):e050519 [FREE Full text] [doi: 10.1136/bmjopen-2021-050519] [Medline: 34253676]
6. Zimmermann BM, Fiske A, Prainsack B, Hangel N, McLennan S, Buyx A. Early perceptions of COVID-19 contact tracing apps in German-speaking countries: comparative mixed methods study. J Med Internet Res 2021;23(2):e25525 [FREE Full text] [doi: 10.2196/25525] [Medline: 33503000]
7. Seto E, Challa P, Ware P. Adoption of COVID-19 contact tracing apps: a balance between privacy and effectiveness. J Med Internet Res 2021;23(3):e25726 [FREE Full text] [doi: 10.2196/25726] [Medline: 33617459]
8. Salomon JA, Reinhart A, Bilinski A, Chua EJ, La Motte-Kerr W, Rönn MM, et al. The US COVID-19 trends and impact survey: continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. Proc Natl Acad Sci USA 2021;118(51):e2111454118 [FREE Full text] [doi: 10.1073/pnas.2111454118] [Medline: 34903656]
9. End of the federal COVID-19 Public Health Emergency (PHE) Declaration. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/coronavirus/2019-ncov/your-health/end-of-phe.html [accessed 2024-01-29]
10. Nissenbaum HF. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Redwood City, CA: Stanford University Press; 2010.
11. Blom AG, Wenz A, Cornesse C, Rettig T, Fikel M, Friedel S, et al. Barriers to the large-scale adoption of a COVID-19 contact tracing app in Germany: survey study. J Med Internet Res 2021;23(3):e23362 [FREE Full text] [doi: 10.2196/23362] [Medline: 33577466]
12. Jacob S, Lawarée J. The adoption of contact tracing applications of COVID-19 by European governments. Policy Design and Practice 2021;4(1):44-58 [FREE Full text] [doi: 10.1080/25741292.2020.1850404]
13. Sharma S, Singh G, Sharma R, Jones P, Kraus S, Dwivedi YK. Digital health innovation: exploring adoption of COVID-19 digital contact tracing apps. IEEE Transactions on Engineering Management 2024;71:12272-12288. [doi: 10.1109/tem.2020.3019033]
14. Walrave M, Waeterloos C, Ponnet K. Adoption of a contact tracing app for containing COVID-19: a health belief model approach. JMIR Public Health Surveill 2020;6(3):e20572 [FREE Full text] [doi: 10.2196/20572] [Medline: 32755882]
15. Utz C, Becker S, Schnitzler T, Farke F, Herbert F, Schaewitz L, et al. Apps against the spread: privacy implications and user acceptance of COVID-19-related smartphone apps on three continents. 2021 Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445517]
16. Oyibo K, Sahu KS, Oetomo A, Morita PP. Factors influencing the adoption of contact tracing applications: protocol for a systematic review. JMIR Res Protoc 2021;10(6):e28961 [FREE Full text] [doi: 10.2196/28961] [Medline: 33974551]
17. Kaptchuk G, Goldstein DG, Hargittai E, Hofman J, Redmiles EM. How good is good enough? Quantifying the impact of benefits, accuracy, and privacy on willingness to adopt COVID-19 decision aids. Digit Threats Res Pract 2022 Mar 26;3(3):1-18. [doi: 10.1145/3488307]
18. Hargittai E, Redmiles EM, Vitak J, Zimmer M. Americans' willingness to adopt a COVID-19 tracking app. First Monday 2020;25(11):1-21. [doi: 10.5210/fm.v25i11.11095]
19. Trestian R, Xie G, Lohar P, Celeste E, Bendechache M, Brennan R, et al. Privacy in a time of COVID-19: how concerned are you? IEEE Secur Privacy 2021;19(5):26-35. [doi: 10.1109/msec.2021.3092607]
20. Redmiles EM. User concerns & tradeoffs in technology-facilitated COVID-19 response. Digit Gov Res Pract 2020;2(1):1-12. [doi: 10.1145/3428093]
21. Jamieson J, Yamashita N, Epstein DA, Chen Y. Deciding if and how to use a COVID-19 contact tracing app: influences of social factors on individual use in Japan. 2021 Oct 18 Presented at: Proceedings of the ACM on Human-Computer Interaction; October 18, 2021; New York, NY, United States p. 1-30 URL: https://dl.acm.org/doi/abs/10.1145/3479868 [doi: 10.1145/3479868]
22. Nurgalieva L, Ryan S, Balaskas A, Lindqvist J, Doherty G. Public views on digital COVID-19 certificates: a mixed methods user study. 2022 Apr 28 Presented at: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 28, 2022; New Orleans, LA, United States p. 1-28 URL: https://dl.acm.org/doi/10.1145/3491102.3502066 [doi: 10.1145/3491102.3502066]
23. Blanchette J, Johnson DG. Data retention and the panoptic society: the social benefits of forgetfulness. Inf Soc 2011;18(1):33-45. [doi: 10.1080/01972240252818216]
24. Crump C. Data retention: privacy, anonymity, and accountability online. Stanford Law Rev 2003;56(1):191-299 [FREE Full text]
25. Leon PG, Ur B, Wang Y, Sleeper M, Balebako R, Shay R, et al. What matters to users?: Factors that affect users' willingness to share information with online advertisers. 2013 Jul 24 Presented at: SOUPS '13: Proceedings of the Ninth Symposium on Usable Privacy and Security; July 24, 2013; Newcastle, United Kingdom p. 1-12 URL: https://dl.acm.org/doi/10.1145/2501604.2501611 [doi: 10.1145/2501604.2501611]

XSL•FO
RenderX

26. Sheahan M. German restaurants object after police use COVID data for crime-fighting. Reuters. 2020. URL: https://www.reuters.com/article/us-health-coronavirus-germany-privacy/german-restaurants-object-after-police-use-covid-data-for-crime-fighting-idUSKCN24W2K6 [accessed 2024-07-19]
27. Ducket C. Western Australia finally thinks about quarantining COVID check-in info from cops. ZDNET. 2021. URL: https://www.zdnet.com/article/western-australia-finally-thinks-about-quarantining-covid-check-in-info-from-cops [accessed 2024-07-19]
28. Emami-Naeini P, Bhagavatula S, Habib H, Degeling M, Bauer L, Cranor L, et al. Privacy expectations and preferences in an IoT world. 2017 Presented at: Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security; July 12-14, 2017; Santa Clara, CA, United States p. 399-412 URL: https://dl.acm.org/doi/10.5555/3235924.3235956
29. Smullen D, Feng Y, Zhang SA, Sadeh NM. The best of both worlds: mitigating trade-offs between accuracy and user burden in capturing mobile app privacy preferences. Proc Priv Enh Technol 2020;2020(1):195-215 [FREE Full text] [doi: 10.2478/popets-2020-0011]
30. Zhang S, Feng Y, Bauer L, Cranor L, Das A, Sadeh N. “Did you know this camera tracks your mood?”: understanding privacy expectations and preferences in the age of video analytics. Proc Priv Enh Technol 2021;2021(2):282-304.
31. Zhang S, Shvartzshnaider Y, Feng Y, Nissenbaum H, Sadeh N. Stop the spread: a contextual integrity perspective on the appropriateness of COVID-19 vaccination certificates. 2022 Jun 21 Presented at: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; June 21-24, 2022; Seoul, Republic of Korea p. 1657-1670 URL: https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533222 [doi: 10.1145/3531146.3533222]
32. Lacey N. Theory into practice? Pornography and the public/private dichotomy. J Law Soc 1993;20(1):93. [doi: 10.2307/1410114]
33. Nissenbaum H. Protecting privacy in an information age: the problem of privacy in public. Law Philos 1998;17(5/6):559-596 [FREE Full text] [doi: 10.2307/3505189]
34. Gerdon F, Nissenbaum H, Bach RL, Kreuter F, Zins S. Individual acceptance of using health data for private and public benefit: changes during the COVID-19 pandemic. Harvard Data Sci Rev 2021(Special Issue 1 - COVID-19: Unprecedented Challenges and Chances):1-24 [FREE Full text] [doi: 10.1162/99608f92.edf2fc97]
35. Preibusch S. Guide to measuring privacy concern: review of survey and observational instruments. Int J Hum-Comput Stud 2013;71(12):1133-1143. [doi: 10.1016/j.ijhcs.2013.09.002]
36. Finch J. The vignette technique in survey research. Sociology 1987;21(1):105-114. [doi: 10.1177/0038038587021001008]
37. Atzmüller C, Steiner PM. Experimental vignette studies in survey research. Methodology 2010;6(3):128-138 [FREE Full text] [doi: 10.1027/1614-2241/a000014]
38. Apthorpe N, Shvartzshnaider Y, Mathur A, Reisman D, Feamster N. Discovering smart home Internet of Things privacy norms using contextual integrity. 2018 Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,; July 5, 2018; New York, NY, United States p. 1-23. [doi: 10.1145/3214262]
39. Apthorpe N, Varghese S, Feamster N. Evaluating the contextual integrity of privacy regulation: parents' IoT toy privacy norms versus COPPA. 2004 May Presented at: 28th USENIX Security Symposium (USENIX Security 19); 2019; Santa Clara, CA, United States p. 123-140 URL: https://www.usenix.org/conference/usenixsecurity19/presentation/apthorpe [doi: 10.1109/msp.2004.9]
40. Martin KE, Nissenbaum H. Measuring privacy: using context to expose confounding variables. SSRN Electronic J 2015:1-40 [FREE Full text] [doi: 10.2139/ssrn.2709584]
41. Shvartzshnaider Y, Tong S, Wies T, Kift P, Nissenbaum H, Subramanian L, et al. Learning privacy expectations by crowdsourcing contextual informational norms. 2016 Sep 21 Presented at: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP); September 21, 2016; Austin, TX, United States p. 209-218 URL: https://ojs.aaai.org/index.php/HCOMP/article/view/13271 [doi: 10.1609/hcomp.v4i1.13271]
42. Brown RF, John AS, Hu Y, Sandhu G. Differential electronic survey response: does survey fatigue affect everyone equally? J Surg Res 2024;294:191-197. [doi: 10.1016/j.jss.2023.09.072] [Medline: 37913726]
43. Abbey JD, Meloy MG. Attention by design: using attention checks to detect inattentive respondents and improve data quality. J Ops Manag 2017;53-56(1):63-70. [doi: 10.1016/j.jom.2017.06.001]
44. Contexual acceptance of COVID-19 mitigation mobile apps. Open Science Framework. URL: https://osf.io/fpsjy/?view_only=60adfb83677841fa949d54c4aff282cf [accessed 2024-08-12]
45. Representative samples. Prolific. URL: https://researcher-help.prolific.co/hc/en-gb/articles/360019236753-Representative-Samples-on-Prolific [accessed 2024-01-29]
46. Christensen RHB. ordinal: Regression Models for Ordinal Data. URL: https://cran.r-project.org/web/packages/ordinal/index.html [accessed 2024-07-19]
47. Taylor JE, Rousselet GA, Scheepers C, Sereno SC. Rating norms should be calculated from cumulative link mixed effects models. Behav Res Methods 2023;55(5):2175-2196 [FREE Full text] [doi: 10.3758/s13428-022-01814-7] [Medline: 36103049]

XSL•FO
RenderX

Abbreviations

NH: null hypothesis
OR: odds ratio
RQ: research question
TP: transmission principle

Edited by T Leung, YH Lin; submitted 16.02.24; peer-reviewed by N Apthorpe, A Rusibamayila, A Hassan; comments to author 28.04.24; revised version received 14.06.24; accepted 14.06.24; published 29.08.24.

Please cite as:
Feng Y, Stenger B, Zhang S
Contextual Acceptance of COVID-19 Mitigation Mobile Apps in the United States: Mixed Methods Survey Study on Postpandemic Data Privacy
J Med Internet Res 2024;26:e57309
URL: https://www.jmir.org/2024/1/e57309
doi:10.2196/57309
PMID:

©Yuanyuan Feng, Brad Stenger, Shikun Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Response Results of an mHealth-Based Patient Experience Survey Among People Receiving HIV Care in Lusaka, Zambia: Cohort Study

Jacob Mutale[1], MSc; Kombatende Sikombe[2], MSc; Boroma Mwale[3], BS; Mwansa Lumpa[1], MSc; Sandra Simbeza[2], MSc; Chama Bukankala[1], DS; Njekwa Mukamba[4], MSc; Aaloke Mody[5], MD; Laura K Beres[6], MPH, PhD; Charles B Holmes[7], MD, MPH; Carolyn Bolton Moore[2,8], MBChB, MSc; Elvin H Geng[5], MD, MPH; Izukanji Sikazwe[2], MBChB, MPH; Jake M Pry[2,9], MPH, PhD

[1]Data Unit, Centre for Infectious Disease Research, Lusaka, Zambia
[2]Implementation Science Unit, Centre for Infectious Disease Research, Lusaka, Zambia
[3]Analysis Unit, Centre for Infectious Disease Research, Lusaka, Zambia
[4]Social Science Research Unit, Centre for Infectious Disease Research, Lusaka, Zambia
[5]School of Medicine, Washington University, St. Louis, MO, United States
[6]School of Public Health, Johns Hopkins University, Baltimore, MD, United States
[7]School of Medicine, Georgetown University, Washington, DC, DC, United States
[8]School of Medicine, University of Alabama, Birmingham, AL, United States
[9]School of Medicine, University of California, Davis, Sacramento, CA, United States

Corresponding Author:
Jake M Pry, MPH, PhD
School of Medicine
University of California, Davis
4610 X Street
Sacramento, CA, 95817
United States
Phone: 1 916 734 4800
Email: jmpry@ucdavis.edu

Abstract

Background: This pilot study evaluates the effectiveness of mobile talk-time incentives in maintaining participation in a longitudinal mobile health (mHealth) data collection program among people living with HIV in Lusaka, Zambia. While mHealth tools, such as mobile phone surveys, provide vital health feedback, optimal incentive strategies to ensure long-term engagement remain limited. This study explores how different incentive levels affect response rates in multiple survey rounds, providing insights into effective methods for encouraging ongoing participation, especially in the context of Zambia's prepaid mobile system and multi-SIM usage, a common practice in sub-Saharan Africa.

Objective: This study aimed to assess the response rate success across multiple invitations to participate in a care experience survey using a mobile phone short codes and unstructured supplementary service data (USSD) model among individuals in an HIV care setting in the Lusaka, Zambia.

Methods: Participants were recruited from 2 study clinics–1 in a periurban setting and 1 in an urban setting. A total of 2 rounds of survey invitations were sent to study participants on a 3-month interval between November 1, 2018, and September 23, 2019. Overall, 3 incentive levels were randomly assigned by participant and survey round: (1) no incentive, (2) 2 Zambian Kwacha (ZMW; US $0.16), and (3) 5 ZMW (US $0.42). Survey response rates were analyzed using mixed-effects Poisson regression, adjusting for individual- and facility-level factors. Probability plots for survey completion were generated based on language, incentive level, and survey round. We projected the cost per additional response for different incentive levels.

Results: A total of 1006 participants were enrolled, with 72.3% (727/1006) from the urban HIV care facility and 62.4% (628/1006) requesting the survey in English. We sent a total of 1992 survey invitations for both rounds. Overall, survey completion across both surveys was 32.1% (637/1992), with significantly different survey completion between the first (40.5%, 95% CI 37.4-43.6%) and second (23.7%, 95% CI 21.1-26.4) invitations. Implementing a 5 ZMW (US $0.42) incentive significantly increased the

XSL•FO
RenderX

adjusted prevalence ratio (aPR) for survey completion compared with those that received no incentive (aPR 1.35, 95% CI 1.11-1.63). The cost per additional response was highest at 5 ZMW, equivalent to US $0.42 (72.8 ZMW [US $5.82] per 1% increase in response).

Conclusions: We observed a sharp decline of almost 50% in survey completion success from the initial invitation to follow-up survey administered 3 months later. This substantial decrease suggests that longitudinal data collection potential for a care experience survey may be limited without additional sensitization and, potentially, added survey reminders. Implementing a moderate incentive increased response rates to our health care experience survey. Tailoring survey strategies to accommodate language preferences and providing moderate incentives can optimize response rates in Zambia.

Trial Registration: Pan African Clinical Trial Registry PACTR202101847907585; https://pactr.samrc.ac.za/TrialDisplay.aspx?TrialID=14613

(J Med Internet Res 2024;26:e54304) doi:10.2196/54304

KEYWORDS

mHealth; mobile health; survey; incentives; HIV; Zambia; airtime; USSD; unstructured supplementary service data; HIV care; pilot study; mobile phone; public health service; mobile health; urban; rural; regression model; longitudinal; mobile; patient feedback

Introduction

The increasing availability of mobile devices and mobile network coverage in sub-Saharan Africa presents a growing opportunity for mobile health (mHealth) interventions. By 2025, projections estimate that 50% of the sub-Saharan African population will own a mobile phone [1]. Zambia exemplifies this trend, with rapid mobile network expansion leading to an estimated 21.1 million mobile connections, reaching 91.4% of the population [2]. Mobile device usage in Zambia is shifting from voice calls to mobile internet, with users increasing from 9.1 million in 2019 to 12.5 million in 2023 [2,3]. In addition, the use of mobile money through local mobile network providers has risen rapidly, already 2.7 times higher in 2023 compared with 2021.

In much of sub-Saharan Africa, including Zambia, mobile network operators offer prepaid service plans. These plans allow users to purchase “talk-time,” providing finite minutes or data, as dedicated. With multiple competing mobile networks vying for market share, promotions and incentives for holding multiple phone numbers across the 3 main networks in Zambia are common. This can create challenges for mHealth engagement and response.

The rising level of accessibility to mobile devices allows for mHealth tools, such as mobile phone surveys, to capture important health-related feedback without substantially limiting participant eligibility. The use of the mHealth tools enables health care providers and health facility stakeholders to improve patient monitoring, enable mobile clinical decision support, and address contextual challenges in sub-Saharan Africa [4,5]. This technological advancement has the potential to greatly enhance health care outcomes by facilitating a better understanding of patient needs and improving service delivery in sub-Saharan Africa [6-8].

HIV care has now become chronic, with people making HIV care clinic visits every 3 months at the time of this research study in Zambia, providing an opportunity to regularly collect care experience measures. However, there is limited evidence in sub-Saharan Africa concerning the durability of an initial enrollment across multiple invitations and the role of incentives on survey response rates, especially in an era where mobile technology and mobile networks have become highly accessible. Previous work done in 4 countries in sub-Saharan Africa reports varying response rates to a general population SMS survey, with the highest response in Uganda (14.2%) and the lowest in Nigeria (0.3%) [9]. In 2016, a study in Kenya offering information on family planning methods, through SMS text messages, found that incentives varying between US $0.5 and US $1.0 were found to have no association with SMS survey response [10]. Another study in Zambia used SMS notifications as an option for linkage to care for individuals who are HIV positive [11]. Our study in Zambia is the first to examine the prospective outcomes of health-related surveys based on unstructured supplemental service data (USSD) at varying levels of talk-time incentives [12-14].

Using pilot data from a large stepped-wedge trial, we assess the pattern of responses across multiple rounds of invitation to participate in an mHealth model. Specifically, we investigated the association between talk-time incentive and rate of response to a brief, 4-question, USSD-based patient experience survey in Lusaka, Zambia. We collaborated with all the mobile network operators in Zambia to establish a unique USSD code (also called a short code) for the Person-Centered Public Health for HIV Treatment in Zambia (PCPH) study. We analyzed 3 levels of talk-time incentive, preferred language for survey administration, and clinic setting (periurban or urban) on the response rate of the USSD survey. Our goal was to evaluate the overall uptake, and response rate, to the survey after an HIV care visit in Lusaka, Zambia. Although this application was limited to HIV care, these findings may provide valuable insights for future initiatives involving mHealth in Zambia and sub-Saharan Africa.

Methods

Study Setting

This study was conducted as part of the pilot phase of the larger PCPH study, a cluster-randomized stepped-wedge trial to assess the impact of a patient-centeredness intervention on patient

XSL•FO
RenderX

experiences of HIV care and outcomes, which ran from August 2018 to November 2021 in Lusaka Province (Pan African Clinical Trial Registry PACTR202101847907585) [15-17].

Data presented here were collected during the pilot phase, which was designed to develop and optimize a mobile phone–based survey for the main PCPH study, from November 1, 2018, to June 25, 2019. Pilot study clinics are run by the Zambian Ministry of Health (MoH) and receive various levels of technical support through the Centre for Infectious Disease Research in Zambia, an implementing partner of the MoH since 2011 and the primary implementer of the PCPH trial.

Recruitment Procedure

Potential survey participants were identified by research assistants at 1 of 2 (1 in an urban setting, 1 in a periurban setting) pilot study HIV care facilities using convenience sampling; participants were selected following reception and triage procedures, which included routine data collection (eg, temperature, weight, and blood pressure). Eligibility for the study participation was limited to individuals who were enrolled as a recipient of care at the study facility, were aged 18 years or older, self-reported as literate in 1 of 4 survey languages, had a mobile phone during enrollment, and voluntarily consented to participate in the study. We confirmed that the mobile phone number reported by the potential participant belonged to the individual through a registration process using the unique PCPH short code, *744#, on the potential participant's phone at enrollment. The registration process confirmed the ability to initiate and complete a USSD session, collected the language preference, and recorded the facility where the enrolment took place.

Data Collection

Surveys

Survey participants were presented with 4 questions related to their experience at their most recent HIV care visit. The opening question was the same for all participants, the next 3 questions were drawn, using a systematic algorithm, from a bank of 9 questions (USSD survey question sets are shown in Multimedia Appendix 1). The survey questions were subject to a cognitive interviewing process to increase acceptability, understanding, and appropriateness before piloting. At enrollment, participants selected their preferred survey language from English, Bemba (spoken by 33.5% of the Zambian population), Nyanja (spoken by 14.8% of the Zambian population), or Tonga (spoken by 11.4% of the Zambian population). Although only 2% of the population have English as their first language, it is the most spoken second language [18].

Survey Administration

A total of 2 rounds of surveys were sent to those enrolled. The first round of survey invitations was sent through SMS text messages between November 1, 2018, and June 25, 2019, within 14 days of enrollment. The second round of invitations was sent 3 months after the first invitation, between February 1, 2019, and September 23, 2019. Survey spacing of 3 months was established through a review of median visit interval in the previous year at pilot study clinics. This second round was intended to capture the experience during the next routine HIV care visit, commonly scheduled quarterly (every 3 months). An incentive was systematically allocated to a participant once at registration for both survey rounds. The incentive allocation algorithm created a queue with values of 0 Zambian Kwacha (ZMW) or no incentives, 2 ZMW (US $0.16), and 5 ZMW (US $0.42), assigned in the order of enrollment. This queue was cycled by the incentive allocation algorithm, which allocated each incentive level before beginning at 0 at the end of each cycle. A currency exchange rate of 12.90 ZMW=US $1 was applicable and, at the time of the first round of the survey, 12.51 ZMW=US $1 was applicable [19]. In Zambia, talk-time is typically used to purchase bundles of minutes for voice calls, data, and sending and receiving SMS text messages. These bundles are valid for a day, 1 week, or 1 month. A 5 ZMW (US $0.42) talk-time can purchase a daily bundle with an average of 25 minutes for voice calls, 25 MB of data, and 127 SMS text messages. The survey invitation informed the survey participants of the incentive amount that they would receive upon survey completion. These incentive amounts were automatically sent to the participants through a third-party vendor immediately after the completion of all 5 survey questions. The money was sent as a talk-time top-up to study participants who all had a prepaid plan through their mobile network. Participants who did not start or complete the survey received a reminder through SMS text messages 24 hours after previous invitation and again 7 days after that invitation if the survey remained incomplete.

Data Analysis

We assessed the association between survey completion and incentive amount for a USSD-based survey using a mixed-effects Poisson regression model, allowing random effects at both the individual and facility levels. A probability model with 95% CIs was illustrated for incentive level and language chosen by the participant. In addition, we projected the cost for additional response by incentive level based on 1000 invitations to the survey, by comparing the cost of additional response between 2 ZMW (US $0.16) and no incentives and between 5 ZMW (US $0.42) and no incentives. All analyses were completed using Stata BE (StataCorp LLC).

Ethical Considerations

Ethical approval was granted by the University of Zambia Biomedical Research Ethics Committee (reference 008-03-19), the National Health Research Authority in Zambia, and the University of Alabama at Birmingham (300003282). The National Health Research Authority and the Zambian Ministry of Health provided authorization to conduct the research. The study participant data were deidentified, and a unique identifier was generated to identify participants' responses across survey rounds. All the participants provided individual written informed consent before being enrolled in the study.

Results

A total of 1006 participants were enrolled, of which 986 (98%) received a second invitation to participate 3 months after the initial invitation, for a total of 1992 observations (Figure 1). At the beginning of the trial, the urban HIV care facility had a total

XSL•FO
RenderX

of 7159 individuals receiving HIV care, while a periurban HIV care facility had a total of 3112 individuals receiving HIV care.

We found an overall response rate across both surveys to be 32% (637/1992) with significantly different response rates between the first (40.5%, 407/1006) and second round (23.3%, 230/986) surveys ($\chi^2 P<.001$; Tables S1 and S2 in Multimedia Appendix 1). The majority of those enrolled visited the urban-based HIV care facility (727/1006, 72.3%) and the survey was most often requested in English (628/1006, 62.4%; Table 1).

The second round of survey invitation had a significantly lower adjusted prevalence (adjusted prevalence ratio [aPR] 0.49, 95% CI 0.44-0.53) of response compared with the first round. Implementing a 5 ZMW (US $0.42) incentive significantly increased the prevalence ratio for survey completion compared with those that received no incentives (aPR 1.35, 95% CI 1.11-1.63; Table 2). Those enrolled at a visit to the urban study clinic had a significantly higher aPR at 1.26 (95% CI 1.03-1.52) compared with those at the periurban study clinic (Table 2).

Figure 1. Participant response outcomes flow diagram by incentive level and survey round. Round 1 invitation to participate occurred within 14 days of enrollment in the study and round 2 occurred 3 months after round 1. ZMW: Zambian Kwacha.

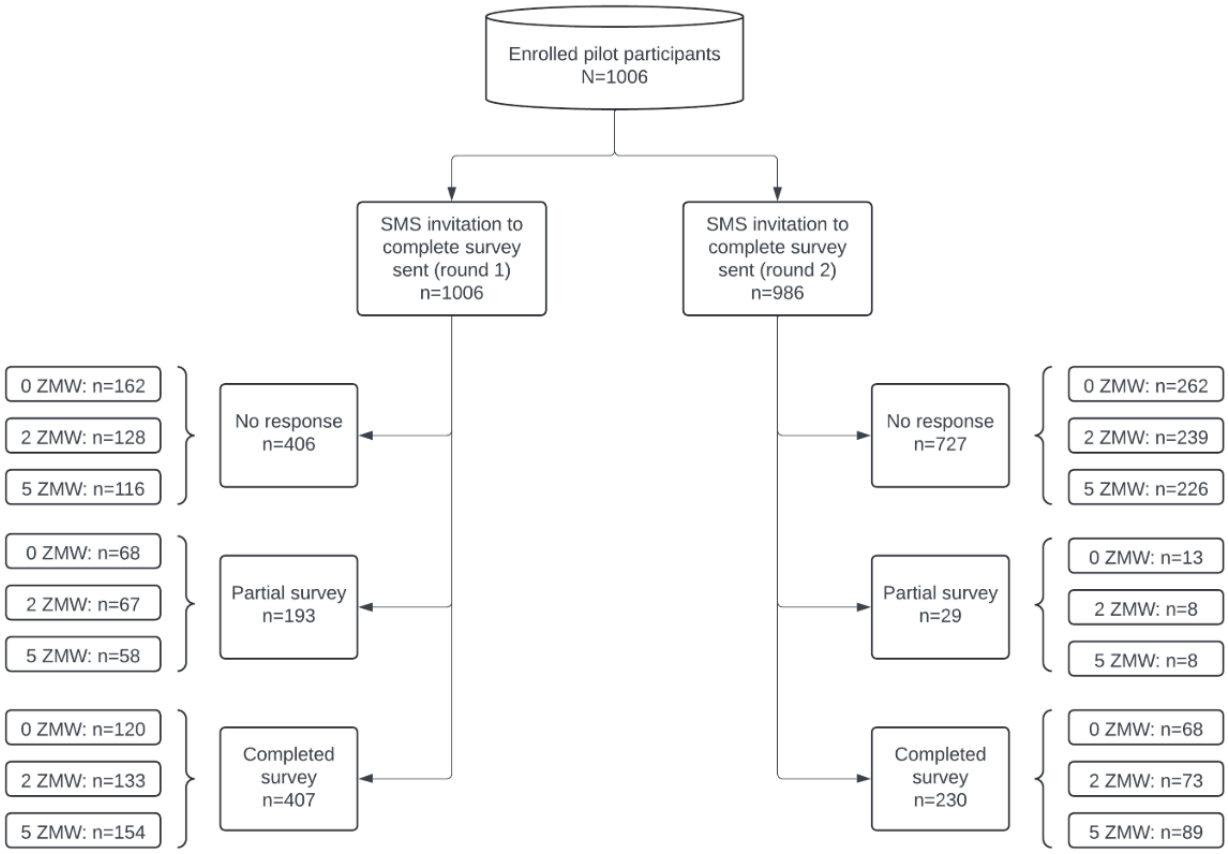

We observed a significant decrease in survey completion between survey 1 and the follow-up survey sent 3 months later. The probability of completing the survey was lowest for those who received no incentive in round 2 for which 20.6% completed the survey (95% CI 17.1-25.2; Figure 2; Table S3 in Multimedia Appendix 1).

We found a statistically significant association between language and setting for response to both survey invitations compared with just the first invitations. Individuals completing at least 1 survey in Bemba have a significantly lower adjusted prevalence (aPR 0.87, 95% CI 0.87-0.87) of completing both survey invitations compared with those completing the survey in English. In addition, individuals in urban settings have a significantly lower prevalence (aPR: 0.90, 95% CI 0.87-0.94 of responding to both survey invitations compared with individuals in a rural setting (Table 3).

Using observed response rates, we projected cost per additional response to be highest, at the 5 ZMW (US $0.42) incentive level, at 72.8 ZMW (US $5.82) per additional response as compared with the 0 ZMW incentive level considering 1000 invitations (Table 4).

XSL•FO
RenderX

Table 1. Participant characteristics by round of invitation ($N_{individuals}$=1006).

Factors and their levels	Total invites (N=1992)	First round (n=1006)	Second round (n=986)
Incentive, n (%)			
0 ZMW[a]	693 (34.8)	350 (34.8)	343 (34.8)
2 ZMW (US $0.16)	648 (32.5)	328 (32.6)	320 (32.5)
5 ZMW (US $0.42)	651(32.7)	328 (32.6)	323 (32.8)
Response status, n (%)			
No response	1133 (56.9)	406 (40.4)	727 (73.7)
Completed	637 (32.0)	407 (40.5)	230 (23.3)
Partial	222 (11.1)	193 (19.2)	29 (2.9)
Facility, n (%)			
Periurban	556 (27.9)	279 (27.7)	277 (28.1)
Urban	1436 (72.1)	727 (72.3)	709 (71.9)
Survey language, n (%)			
English	1240 (62.2)	628 (62.4)	612 (62.1)
Nyanja	524 (26.3)	263 (26.1)	261 (26.5)
Bemba	202 (10.1)	102 (10.1)	100 (10.1)
Tonga	26 (1.3)	13 (1.3)	13 (1.3)

[a]ZMW: Zambian Kwacha.

Table 2. Crude and adjusted predictors of unstructured supplementary service data response. The model was adjusted for survey round, language, and setting; both unadjusted and adjusted models allow for random effects at individual and facility levels.

Factors and their levels	PR[a] (95% CI)	aPR[b], (95% CI)
Survey round		
1	ref[c]	ref
2	0.48 (0.44-0.52)	0.49 (0.44-0.53)
Incentive amount		
0 ZMW[d]	ref	ref
2 ZMW (US $0.16)	1.17 (0.96-1.43)	1.17 (0.96-1.43)
5 ZMW (US $0.42)	1.35 (1.12-1.64)	1.35 (1.11-1.63)
Language		
English	ref	ref
Nyanja	0.77 (0.63-0.94)	0.75 (0.56-0.99)
Bemba	0.80 (0.60-1.06)	0.66 (0.43-1.03)
Tonga	0.98 (0.49-1.96)	1.08 (0.39-2.98)
Setting		
Periurban	ref	ref
Urban	1.28 (1.05-1.55)	1.26 (1.03-1.52)

[a]PR: unadjusted prevalence ratio.
[b]aPR: adjusted prevalence ratio.
[c]ref: reference.
[d]ZMW: Zambian Kwacha.

XSL•FO
RenderX

Figure 2. (A) Marginal probability of completing survey by language and survey round. (B) Marginal probability of completing the survey by incentive level and survey round. Round 1 occurred within 7 days of enrollment and round 2 occurred 3 months after the initial survey invitation of round 1. ZMW: Zambian Kwacha.

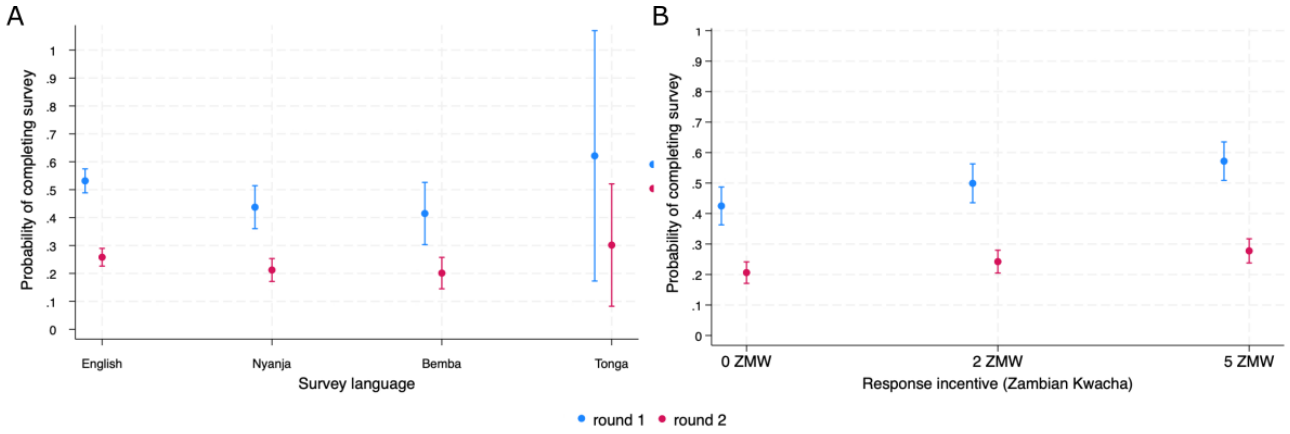

Table 3. Predictors of responding to both survey invitations, compared with the first invite only (n=803).

Factors and their levels	aPR[a] (95% CI)	*P* value
Incentive amount		
0 ZMW[b]	ref[c] (1.00)	—[d]
2 ZMW (US $0.16)	0.96 (0.77-1.21)	.75
5 ZMW (US $0.42)	1.03 (0.91-1.15)	.66
Language		
English	ref (1.00)	—
Nyanja	1.01 (0.92-1.11)	.86
Bemba	0.87 (0.87-0.87)	<.001
Tonga	0.97 (0.43-2.19)	.95
Setting		
Rural	ref (1.00)	—
Urban	0.90 (0.87-0.94)	<.001

[a]aPR: adjusted prevalence ratio.

[b]ZMW: Zambian Kwacha.

[c]ref: reference.

[d]Not applicable.

Table 4. Cost for additional observations by incentive level based on 1000 invitations. Proportion increase compared with 0 ZMW[a] incentive.

Incentive levels	Projected percent completing survey (N=1000), n (%)	Projected total cost (ZMW)	Projected percent completion increase	Projected price (ZMW) per percent increase given 1000 invitations
0 ZMW	341 (34.1)	0	0 (reference)	0 (reference)
2 ZMW (US $0.16)	403 (40.3)	258	6.2	41.6 ZMW (US $11.65) per 1% increase in response
5 ZMW (US $0.42)	464 (46.4)	750	12.3	72.8 ZMW (US $5.82) per 1% increase in response

[a]ZMW: Zambian Kwacha.

Finally, we estimated the probability of survey completion by language and incentive level. English had a significantly higher completion rate than Bemba and Nyanja at all levels of the incentives with wide CIs around the Tonga point estimate (Figure 2). While not significant, survey participants who requested the survey in Tonga and were assigned 5 ZMW (US $0.42) had the highest estimated point probability of completing the survey, at 51.6% (95% CI 13.9-89.2), while those who spoke Bemba and received 0 ZMW had the lowest estimated point probability of completing the survey at 37.78% (95% CI 27.4-48.2; Figure 3; Table S4 in Multimedia Appendix 1). Incentive value was significantly associated with the response

XSL•FO
RenderX

rate for those requesting the survey in English, while Bemba, Nyanja, and Tonga did not have a significant difference in survey completion probability by incentive (Figure 3).

Figure 3. Marginal probability of responding to the survey by incentive level and language. ZMW: Zambian Kwacha.

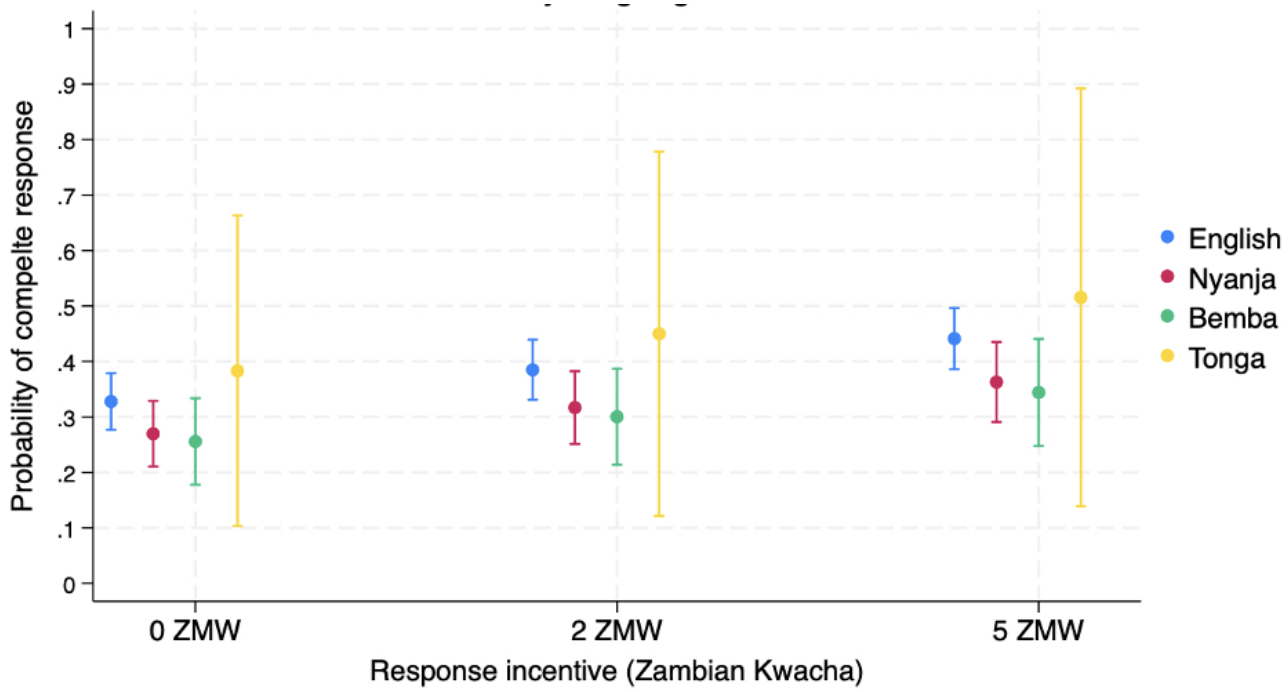

Discussion

Principal Findings

We found a significant reduction in survey response rate across different languages and incentive levels 3 months after enrollment. Interestingly, the significant association between incentive level and survey response holds only for the first round of survey invitations, while there is no significant difference across incentive levels for the second round. Overall, the response rate across both surveys is highest (32.1%) among participants who received a 5 ZMW (US $0.42) incentive.

In Australia, longitudinal data collection through SMS text messages, with presurvey notifications, has been shown to be effective for monitoring postoperative pediatric pain, providing hope that our mHealth model may be an efficient mechanism to collect data longitudinally in our setting as well [20,21]. However, in our population of adults in HIV care, we found that subsequent invitations to participants were significantly less successful than the initial invitation. This restriction does not totally exclude the tool's usage as a longitudinal data collection instrument. Some contributors to this response attrition may include (1) a high rate of phone number turnover, (2) phone discontinuance (phone sold, lost, stolen, etc), (3) no interest in the survey, (4) do not need incentive or incentive insufficient, and (5) no recent care experience to report in the event the participant has not had a return visit for HIV care. We suspect that phone number turnover is a chief contributor to the lower response rate to the second invitation given 1.1 mobile numbers per person (107.6 mobile network activations per 100 people) in Zambia in 2023 [2]. Unlike more expensive smartphones in many well-resourced settings, where mobile devices increasingly store vital information, such as credit cards and identifiable information, mobile devices in resource-limited settings tend to be more affordable [3]. In these settings, obtaining phone numbers often through prepaid schemes is inexpensive and readily accessible, and sometimes even provided for free through promotions by mobile network operators. Poor network quality and high cost of services are 2 other reasons leading to mobile customer churn in urban and periurban settings [22]. Although subsequent invitation response rates declined, it may remain a valuable method for collecting data at multiple time points, remotely (a vital aspect during health shocks like the COVID-19 pandemic) [23,24]. It is also possible that subsequent invitation response rates could be improved, despite phone number turnover, through routine updating of participant contact information.

While similar incentive-for-participation projects have been conducted in Zambia, such as those related to voluntary medical male circumcision, which used a 2 ZMW (US $0.16) incentive provided upon SMS survey completion, However, the response rate was not reported [25]. In our study, the provision of incentives was significantly associated with the response rate. There was a dose-response relationship observed between the incentive level and response rate; however, only the 5 ZMW (US $0.42) level was significantly different compared with no incentive. This is consistent with previous literature in Uganda, where incentivization for a general population survey on demographics, socioeconomic status, and technology was found to be significantly associated with response rate; however, incentive level appears to play a stronger role in response rate in our study than in the Uganda work [10]. In South Africa, a survey targeting randomly selected phone numbers from 4 mobile networks found that over 50% of participants opted for a prize raffle draw as an incentive compared with talk-time [26]. This is also the case in South Africa, where an SMS-based data

XSL•FO
RenderX

collecting tool is used as part of a national surveillance system [27]. Future research may consider offering different types of incentives to participants who may not find talk-time a strong enough incentive to complete a survey.

Monetary incentives have been found to increase response rates for both web-based and paper-based surveys compared with no incentives in the United States [28]. It has also been shown that both pre- and postsurvey incentive provisions can yield similar participation [29]. In the United Kingdom, offering unconditional incentives before survey completion was found to be more effective than providing an incentive after survey completion [30]. Then, it is possible that the completion rates observed in this study could be further improved through unconditional survey incentives; however, further research is needed to confirm that the unconditional incentive completion results in the United Kingdom hold true in Zambia.

The pilot was conducted in an urban area at a clinic located within the capital city of Lusaka with full network coverage. In this setting, each participant would likely have more than 1 SIM card to take advantage of promotions offered on other networks. Residents in this community typically work within the city limits and would be able to receive survey participation prompt texts shortly after they are sent. On the other hand, our study clinic in the periurban area may have limited network coverage, being outside the city limits. Typically, families in these communities often work in more agricultural industries, which may lead to movement in and out of network coverage as they search for and conduct work in varying places. In this setting, participants likely switch between networks, and each with a different phone number, to use the network with available reception depending on location which may, in part, drive the poorer response rates we observe in the periurban clinic compared with the urban clinic.

We observed a significant difference in response rate across periurban and urban settings, which may be related to mobile device availability and literacy where individuals in urban settings may have increased exposure to mobile technology compared with those in periurban settings. Periurban settings, especially in resource-limited countries, have the greatest potential to benefit from mHealth efforts. However, sensitization to such technologies may need to be tailored to this population, which may have reduced familiarity with or access to mobile phone technology [12,31].

One strategy that can be used to increase response rates to these types of surveys is to offer incentives to participants who do not initially respond to the survey. There is evidence that this can work in a longitudinal survey conducted in the United States to people exposed to the 9/11 terrorist attacks [32]. This approach would result in a reduction in the cost of survey implementation. Gift cards have also shown evidence of a nonresponse conversation strategy in the United States [33].

In Zambia, English has been the medium of instruction in schools at all levels from primary, secondary, and tertiary education. However, it was only in 2013 when the new language policy came into effect that local languages started being used as a medium of instruction from grades 1 to 4 [34]. This explains why most survey participants chose to receive the survey in English. Literacy in English often means one can read in their native language.

The survey was offered in the 4 most common languages in Zambia with the lowest response rate for the first survey observed among the group requesting the survey in Nyanja (aPR: 0.75, 95% CI 0.56-0.99); however, the prevalence ratio for completing both surveys was higher, although not significantly so, for those requesting the survey in Nyanja (aPR: 1.01, 95% CI 0.43-2.19) compared with English. Although the majority opted to complete the survey in English (62.4%), a large proportion of responses among Nyanja and Bemba languages suggests that this is a useful tool to collect information from those who cannot, or prefer not to, read English in Zambia.

Limitations

This study is subject to several limitations, the most substantial being that it is limited only to those with a mobile device. Although mobile technologies are becoming increasingly available in Zambia, there are still some that cannot yet be reached with this methodology. It is possible, because recruitment was based on convenience sampling, that the selection of participants was biased. However, given that we recruited 9.8% (1006/10,271) of the pilot clinic population and recruitment spanned almost 8 months (November 1, 2018, to June 25, 2019), we suspect the impact of potential selection bias to be minimal. Finally, we are limited in our model adjustment set as we did not collect participant demographic information, such as age, sex, and marital status, which may act as confounders and limit causal inference estimates. Our primary interest in the parent study was an ecological measure at the clinic level of patient experience and not the demographics of participants providing experience information.

Conclusions

The study findings demonstrate the waning response rates over time, indicating a potential benefit in regular contact updates, as well as the differential effectiveness of incentives in improving survey response rates in the context of a mHealth platform for individuals in HIV care. In addition, they suggest that it is important to accommodate varying language preferences. The cost analysis also highlights the need for careful evaluation of the cost-effectiveness of different incentive levels, especially when baseline response rates without incentives are relatively high. Future research and larger-scale studies are necessary to assess the long-term effectiveness of longitudinal, incentive-based approaches in mHealth platforms, aiming to optimize health outcomes for individuals in HIV care and other health care populations.

XSL•FO
RenderX

Acknowledgments

We are grateful to the participants without whom this work would not have been possible. We would also like to thank everyone at the Centre of Infectious Disease Research Zambia and the Zambian Ministry of Health for their tireless efforts to improve HIV care in Zambia. Finally, we would like to thank the Zambia Information and Communication Technology Authority for their guidance on the implementation of this pilot study. This work was funded by the Bill and Melinda Gates Foundation (OPP1166485 and K24 AI134413/AI/NIAID NIH HHS/United States).

Data Availability

The datasets generated during and/or analyzed during this study are available in the Data Dryad repository [35].

Authors' Contributions

This manuscript is the result of collaborative efforts from multiple authors. JM contributed to data collection, data analysis, and manuscript writing. BM was involved in data collection and data curation. KS, AM, CH, CBM, and ML contributed to the conceptualization. SS participated in data collection. CB played a role in data framing. LKB was involved in conceptualization and manuscript writing. CH, EHG, and IS contributed to conceptualization and overall project supervision. Finally, JMP contributed to conceptualization, data curation, led and carried out data analysis, data framing, project supervision, and manuscript writing. Together, these authors have contributed their expertise and efforts to bring this manuscript to fruition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Unstructured supplementary service data (USSD) survey question sets and supplemental tables.
[DOCX File , 23 KB - jmir_v26i1e54304_app1.docx]

References

1. The mobile economy 2021. GSMA. 2021. URL: https://www.gsma.com/mobileeconomy/wp-content/uploads/2021/07/GSMA_MobileEconomy2021_3.pdf [accessed 2023-07-03]
2. 2023 Annual market report for the information and communication technologies sector. ZICTA. 2023. URL: https://www.zicta.zm/market-reports/2023_mid_year_market_report.pdf [accessed 2024-08-09]
3. Kemp S. Digital 2022: Zambia. DataReportal.: Kepios; 2022 Feb 16. URL: https://datareportal.com/reports/digital-2022-zambia [accessed 2024-04-01]
4. Adepoju IO, Albersen BJA, De Brouwere V, van Roosmalen J, Zweekhorst M. mHealth for clinical decision-making in sub-Saharan Africa: a scoping review. JMIR Mhealth Uhealth 2017;5(3):e38 [FREE Full text] [doi: 10.2196/mhealth.7185] [Medline: 28336504]
5. Betjeman TJ, Soghoian SE, Foran MP. mHealth in sub-Saharan Africa. Int J Telemed Appl 2013;2013:482324 [FREE Full text] [doi: 10.1155/2013/482324] [Medline: 24369460]
6. Campbell BR, Ingersoll KS, Flickinger TE, Dillingham R. Bridging the digital health divide: toward equitable global access to mobile health interventions for people living with HIV. Expert Rev Anti Infect Ther 2019;17(3):141-144 [FREE Full text] [doi: 10.1080/14787210.2019.1578649] [Medline: 30721103]
7. Lepère P, Touré Y, Bitty-Anderson AM, Boni SP, Anago G, Tchounga B, et al. Exploring the patterns of use and acceptability of mobile phones among people living with HIV to improve care and treatment: cross-sectional study in three francophone West African countries. JMIR Mhealth Uhealth 2019;7(11):e13741 [FREE Full text] [doi: 10.2196/13741] [Medline: 31719023]
8. Yé M, Kagoné M, Sié A, Bagagnan C, Sanou H, Millogo O, et al. Promoting access equity and improving health care for women, children and people living with HIV/AIDS in Burkina Faso through mHealth. J Public Health (Oxf) 2018;40(suppl_2):ii42-ii51 [FREE Full text] [doi: 10.1093/pubmed/fdy196] [Medline: 30551129]
9. Lau CQ, Lombaard A, Baker M, Eyerman J, Thalji L. How representative are SMS surveys in Africa? Experimental evidence from four countries. Inter J Public Opi Res 2018;31(2):309-330. [doi: 10.1093/ijpor/edy008]
10. Johnson D. Collecting data from mHealth users via SMS surveys: a case study in Kenya. Surv Pract 2016;9(1):1-8. [doi: 10.29115/sp-2016-0004]
11. Chipungu J, Bosomprah S, Zanolini A, Thimurthy H, Chilengi R, Sharma A, et al. Understanding linkage to care with HIV self-test approach in Lusaka, Zambia - a mixed method approach. PLoS One 2017;12(11):e0187998 [FREE Full text] [doi: 10.1371/journal.pone.0187998] [Medline: 29149194]
12. Bervell B, Al-Samarraie H. A comparative review of mobile health and electronic health utilization in sub-Saharan African countries. Soc Sci Med 2019;232:1-16. [doi: 10.1016/j.socscimed.2019.04.024] [Medline: 31035241]

XSL•FO
RenderX

13. Saito S, Duong YT, Metz M, Lee K, Patel H, Sleeman K, et al. Returning HIV-1 viral load results to participant-selected health facilities in national population-based HIV impact assessment (PHIA) household surveys in three sub-Saharan African countries, 2015 to 2016. J Int AIDS Soc 2017;20(Suppl 7):e25004 [FREE Full text] [doi: 10.1002/jia2.25004] [Medline: 29171193]
14. Sutcliffe CG, Thuma PE, van Dijk JH, Sinywimaanzi K, Mweetwa S, Hamahuwa M, et al. Use of mobile phones and text messaging to decrease the turnaround time for early infant HIV diagnosis and notification in rural Zambia: an observational study. BMC Pediatr 2017;17(1):66 [FREE Full text] [doi: 10.1186/s12887-017-0822-z] [Medline: 28270134]
15. Beres L, Simbeza S, Holmes C, Mwamba C, Mukamba N, Sharma A, et al. Human-centered design lessons for implementation science: improving the implementation of a patient-centered care intervention. J Acquir Immune Defic Syndr 2019;82(3):S230-S243 [FREE Full text] [doi: 10.1097/QAI.0000000000002216] [Medline: 31764259]
16. Mukamba N, Mwamba C, Redkar S, Foloko M, Lumbo K, Nyirenda H, et al. Patterns of person-centred communications in public HIV clinics: a latent class analysis using the roter interaction analysis system. J Int AIDS Soc 2023;26 Suppl 1(Suppl 1):e26119 [FREE Full text] [doi: 10.1002/jia2.26119] [Medline: 37408449]
17. Mwamba C, Beres LK, Mukamba N, Jere L, Foloko M, Lumbo K, et al. Provider perspectives on patient-centredness: participatory formative research and rapid analysis methods to inform the design and implementation of a facility-based HIV care improvement intervention in Zambia. J Int AIDS Soc 2023;26(Suppl 1):e26114 [FREE Full text] [doi: 10.1002/jia2.26114] [Medline: 37408458]
18. 2010 Census. Zambia Statistics Agency. 2012. URL: https://www.zamstats.gov.zm/census/ [accessed 2024-04-02]
19. USD/ZMW - US dollar Zambia kwacha. Investing.com. 2023. URL: https://www.investing.com/currencies/usd-zmk-historical-data [accessed 2023-07-13]
20. Dal Grande E, Chittleborough CR, Campostrini S, Dollard M, Taylor AW. Pre-survey text messages (SMS) improve participation rate in an Australian mobile telephone survey: an experimental study. PLoS One 2016;11(2):e0150231 [FREE Full text] [doi: 10.1371/journal.pone.0150231] [Medline: 26919558]
21. Drake-Brockman TFE, Smallbone HE, Sommerfield D, von Ungern-Sternberg BS. Remote after-care using smartphones: a feasibility study of monitoring children's pain with automated SMS messaging. Paediatr Anaesth 2022;32(8):954-960 [FREE Full text] [doi: 10.1111/pan.14481] [Medline: 35569012]
22. Banda PK, Tembo S. Application of system dynamics to mobile telecommunication customer churn management. Journal of Telecommunication, Electronic and Computer Engineering 2017;9(3):101-104 [FREE Full text]
23. Drake AL, Rothschild C, Jiang W, Ronen K, Unger JA. Utility of short message service (SMS) for remote data collection for HIV in low- and middle-income countries. Curr HIV/AIDS Rep 2020;17(6):654-662 [FREE Full text] [doi: 10.1007/s11904-020-00534-x] [Medline: 33010003]
24. Hensen B, Mackworth-Young CRS, Simwinga M, Abdelmagid N, Banda J, Mavodza C, et al. Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. Health Policy Plan 2021;36(3):360-368 [FREE Full text] [doi: 10.1093/heapol/czaa158] [Medline: 33881138]
25. Leiby K, Connor A, Tsague L, Sapele C, Kaonga A, Kakaire J, et al. The impact of SMS-based interventions on VMMC uptake in Lusaka province, Zambia: a randomized controlled trial. J Acquir Immune Defic Syndr 2016;72(Suppl 4):S264-S272 [FREE Full text] [doi: 10.1097/QAI.0000000000001040] [Medline: 27404007]
26. Broich C. Offline data collection in sub-Saharan Africa using SMS surveys: lessons learned. Sample Solutions Preprint posted online on May 26, 2015 [FREE Full text]
27. Singh Y, Jackson D, Bhardwaj S, Titus N, Goga A. National surveillance using mobile systems for health monitoring: complexity, functionality and feasibility. BMC Infect Dis 2019;19(Suppl 1):786 [FREE Full text] [doi: 10.1186/s12879-019-4338-z] [Medline: 31526387]
28. Jacob RT, Jacob B. Prenotification, incentives, and survey modality: an experimental test of methods to increase survey response rates of school principals. Journal of Research on Educational Effectiveness 2012;5(4):401-418. [doi: 10.1080/19345747.2012.698375]
29. Millar MM, Olson LM, VanBuren JM, Richards R, Pollack MM, Holubkov R, et al. Incentive delivery timing and follow-up survey completion in a prospective cohort study of injured children: a randomized experiment comparing prepaid and postpaid incentives. BMC Med Res Methodol 2021;21(1):233 [FREE Full text] [doi: 10.1186/s12874-021-01421-8] [Medline: 34706653]
30. Juszczak E, Hewer O, Partlett C, Hurd M, Bari V, Bowler U, et al. Evaluation of the effectiveness of an incentive strategy on the questionnaire response rate in parents of premature babies: a randomised controlled study within a trial (SWAT) nested within SIFT. Trials 2021;22(1):554 [FREE Full text] [doi: 10.1186/s13063-021-05515-y] [Medline: 34419121]
31. Anderson-Lewis C, Darville G, Mercado RE, Howell S, Di Maggio S. mHealth technology use and implications in historically underserved and minority populations in the United States: systematic literature review. JMIR Mhealth Uhealth 2018;6(6):e128 [FREE Full text] [doi: 10.2196/mhealth.8383] [Medline: 29914860]
32. Yu S, Alper HE, Nguyen A, Brackbill RM, Turner L, Walker DJ, et al. The effectiveness of a monetary incentive offer on survey response rates and response completeness in a longitudinal study. BMC Med Res Methodol 2017;17(1):77 [FREE Full text] [doi: 10.1186/s12874-017-0353-1] [Medline: 28446131]

XSL•FO
RenderX

33. Coopersmith J, Vogel L, Bruursema T, Feeney K. Effects of incentive amount and type of web survey response rates. Survey Practice 2016;9(1):1-10. [doi: 10.29115/sp-2016-0002]
34. Ndeleki B. Teachers' perceptionson the use of local languages as medium of instruction from grades 1-4 in selected private schools of Lusaka [Dissertation]. University of Zambia. 2015. URL: https://dspace.unza.zm/server/api/core/bitstreams/fc53b99d-af5e-4ff0-94ab-49a5584cb609/content [accessed 2024-08-17]
35. Data Dryad. URL: https://doi.org/10.5061/dryad.2z34tmpv8 [accessed 2024-09-19]

Abbreviations

aPR: adjusted prevalence ratio
mHealth: mobile health
MoH: Ministry of Health
PCPH: Person-Centered Public Health for HIV Treatment in Zambia
PR: prevalence ratio
SSA: sub-Saharan Africa
USSD: unstructured supplementary service data
Zambian Kwacha: ZMW

Edited by A Mavragani; submitted 26.11.23; peer-reviewed by S Rupcic, R Saifi; comments to author 28.02.24; revised version received 02.04.24; accepted 12.06.24; published 30.09.24.

Please cite as:
Mutale J, Sikombe K, Mwale B, Lumpa M, Simbeza S, Bukankala C, Mukamba N, Mody A, Beres LK, Holmes CB, Bolton Moore C, Geng EH, Sikazwe I, Pry JM
Assessing the Response Results of an mHealth-Based Patient Experience Survey Among People Receiving HIV Care in Lusaka, Zambia: Cohort Study
J Med Internet Res 2024;26:e54304
URL: https://www.jmir.org/2024/1/e54304
doi:10.2196/54304
PMID:39348170

©Jacob Mutale, Kombatende Sikombe, Boroma Mwale, Mwansa Lumpa, Sandra Simbeza, Chama Bukankala, Njekwa Mukamba, Aaloke Mody, Laura K Beres, Charles B Holmes, Carolyn Bolton Moore, Elvin H Geng, Izukanji Sikazwe, Jake M Pry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Computed Tomography–Based Model for Noninvasive Prediction of the T Stage in Gastric Cancer: Multicenter Retrospective Study

Jin Tao[1,2*], MD; Dan Liu[3*], PhD; Fu-Bi Hu[4], PhD; Xiao Zhang[5], PhD; Hongkun Yin[6], PhD; Huiling Zhang[6], PhD; Kai Zhang[6], PhD; Zixing Huang[3*], PhD; Kun Yang[1,2*], PhD

[1]Department of General Surgery and Laboratory of Gastric Cancer, State Key Laboratory of Biotherapy/Collaborative Innovation Center of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, China

[2]Gastric Cancer Center, West China Hospital, Sichuan University, Chengdu, China

[3]Department of Radiology, West China Hospital, Sichuan University, Chengdu, China

[4]Department of Radiology, The First Affiliated Hospital of Chengdu Medical College, Chengdu, China

[5]Department of Radiology, People's Hospital of Leshan, Leshan, China

[6]Institute of Advanced Research, Infervision Medical Technology Co Ltd, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Kun Yang, PhD
Department of General Surgery and Laboratory of Gastric Cancer
State Key Laboratory of Biotherapy/Collaborative Innovation Center of Biotherapy and Cancer Center
West China Hospital, Sichuan University
Guo Xue street
Chengdu, 610041
China
Phone: 86 18980606729
Email: yangkun068@163.com

Abstract

Background: As part of the TNM (tumor-node-metastasis) staging system, T staging based on tumor depth is crucial for developing treatment plans. Previous studies have constructed a deep learning model based on computed tomographic (CT) radiomic signatures to predict the number of lymph node metastases and survival in patients with resected gastric cancer (GC). However, few studies have reported the combination of deep learning and radiomics in predicting T staging in GC.

Objective: This study aimed to develop a CT-based model for automatic prediction of the T stage of GC via radiomics and deep learning.

Methods: A total of 771 GC patients from 3 centers were retrospectively enrolled and divided into training, validation, and testing cohorts. Patients with GC were classified into mild (stage T1 and T2), moderate (stage T3), and severe (stage T4) groups. Three predictive models based on the labeled CT images were constructed using the radiomics features (radiomics model), deep features (deep learning model), and a combination of both (hybrid model).

Results: The overall classification accuracy of the radiomics model was 64.3% in the internal testing data set. The deep learning model and hybrid model showed better performance than the radiomics model, with overall classification accuracies of 75.7% (P=.04) and 81.4% (P=.001), respectively. On the subtasks of binary classification of tumor severity, the areas under the curve of the radiomics, deep learning, and hybrid models were 0.875, 0.866, and 0.886 in the internal testing data set and 0.820, 0.818, and 0.972 in the external testing data set, respectively, for differentiating mild (stage T1~T2) from nonmild (stage T3~T4) patients, and were 0.815, 0.892, and 0.894 in the internal testing data set and 0.685, 0.808, and 0.897 in the external testing data set, respectively, for differentiating nonsevere (stage T1~T3) from severe (stage T4) patients.

Conclusions: The hybrid model integrating radiomics features and deep features showed favorable performance in diagnosing the pathological stage of GC.

(J Med Internet Res 2024;26:e56851) doi:10.2196/56851

XSL•FO
RenderX

KEYWORDS

gastric cancer; computed tomography; radiomics; T stage; deep learning; cancer; multicenter study; accuracy; binary classification; tumor; hybrid model; performance; pathological stage

Introduction

Gastric cancer (GC) is one of the most prevalent cancers and ranks as the fourth most common cause of cancer-related death globally [1]. Although systemic therapy can increase the survival rate and improve quality of life, the prognosis remains poor due to diagnosis at an advanced stage [2]. Currently, GC staging is performed according to the Union for International Cancer Control and American Joint Committee on Cancer tumor - node - metastasis (TNM) staging system [3]. This system could be used to stratify cancer prognosis. As part of the TNM staging system, T staging based on the tumor depth of GC is crucial for developing treatment plans. For patients with clinical stage T1, endoscopic procedures have been considered the first choice, while for those with clinical stages T2, T3, or T4, surgery and perioperative chemotherapy have been recommended [4]. However, the clinical and pathological diagnosis of tumor depths can sometimes vary [5]. In preoperative T staging, computed tomography (CT) has a sensitivity and specificity of 80% to 90%, respectively, for discriminating between early- and advanced-stage GCs [6], and precise prediction of preoperative T staging using CT plays a vital role in GC treatment.

A retrospective study [7] involving 244 patients with GC reported that the performance of single-phase CT radiomics models is favorable in differentiating between T2- and T3/T4-stage tumors [7]. By extracting intricate information that is imperceptible to the human eye in medical imaging and transforming it into quantitative data, radiomics and deep learning approaches have shown potential in improving the diagnostic capability of current imaging. In radiomics, medical images are transformed into mineable high-dimensional data, which can be used to quantify lesion heterogeneity that cannot be observed in the images. Furthermore, deep learning is the state-of-the-art machine learning approach that uses multiple processing layers and connections to learn complex relationships between input data and desired outputs from a large number of labeled examples, which could provide clinicians with decision support and improve diagnostic and treatment accuracy and efficiency [8,9]. Previous studies have constructed a deep learning model based on CT imaging radiomics signatures to predict the number of lymph node metastases and survival in patients with resected GC [10,11]. Our team has previously studied occult peritoneal metastasis of GC by using imaging radiomics and deep learning [12,13]. In addition, the feasibility and promising performance of machine learning approaches in assessing T staging in lung cancer has been demonstrated [14]. However, few studies have reported the combination of deep learning and radiomics in predicting T staging in GC.

To improve the diagnostic accuracy of T staging to better develop treatment strategies, this study focused on preoperative differentiation among pathological T stages using deep learning and combining single-phase CT radiomics models with deep learning parameters, which can potentially be helpful and assist in guiding clinicians in regard to personalized medicine.

Methods

Patient Enrollment and Eligibility Criteria

The inclusion criteria were as follows: (1) primary gastric adenocarcinoma diagnosed via endoscopy-biopsy pathology; (2) venous images of the whole abdomen (with a slice thickness of 2 mm) obtained preoperatively followed by laparoscopy or surgery performed within 2 weeks; (3) no typical peritoneal metastasis findings, such as omental nodules or omental cake, extensive ascites, or irregular thickening with high peritoneal enhancement, on CT; and (4) no indications of distant metastasis or other tumors.

The exclusion criteria were as follows: (1) previous abdominal surgery, (2) previous abdominal malignancies or inflammatory diseases, (3) insufficient distention of the stomach, (4) poor imaging quality due to artefacts, and (5) indiscernible primary GC tumor on CT images.

Ethical Considerations

This study was granted approval by the ethics committee of West China Hospital, Sichuan University (2019-1158). Given the retrospective nature of the study and the anonymous analysis of all data, the need for informed consent was waived.

Segmentation and Preprocessing of Tumor Regions

Given the T stage in GC, we divided the images into 3 gradings for this study: mild (stages T1 and T2), moderate (stage T3), and severe (stages T4a and T4b) groups. The open-source software ITK-SNAP (version 3.8.0; University of North Carolina) was used for image segmentation. The outline for the tumor regions of interest (ROIs) was manually drawn by a junior radiologist with 6 years of experience in reading abdominal CT images and revised by a senior radiologist with 17 years of experience. When delineating the tumor ROIs, the radiologist referred to the results of gastroscopy to determine the location of the tumor. Besides, 40 CT images of GC lesions were randomly selected and delineated again by the junior radiologist to assess test-retest reliability. Blinded to segmentations delineated by the junior radiologist, these 40 CT images of GC lesions were delineated again by the senior radiologist to assess the intraclass correlation coefficient (ICC). Features with an interobserver ICC of >0.85 were retained, and the ICC values of the selected radiomics features are presented in Table S2 in Multimedia Appendix 1.

Extraction and Selection of the Radiomics Features

The radiomics features were automatically extracted from the ROIs on the maximum cross-section layer of the noncontrast CT images by using the PyRadiomics package (version 3.0). A set of filters (Wavelet, Square, SquareRoot, and Logarithm) were applied to highlight certain image properties [15]. Finally, a total of 1183 radiomics features were extracted from each of

XSL•FO
RenderX

the patients. Principal component analysis was used for dimension reduction of the features, to reduce computation complexity, and to prevent overfitting. Principal component analysis is an unsupervised method that transforms complex, high-dimensional original radiomics data into a dimensionally reduced set of uncorrelated features named "principal components." Principal components were computed based on singular value decomposition of the standardized radiomics features, and the top 15 components that explained most of the variance in the training data set were finally selected for model development.

CT Image Acquisition

The details of the CT protocol are presented in Section 2 and Table S1 in the Multimedia Appendix 1. Segmentation and preprocessing of tumor ROIs, Extraction and selection of the radiomics features are shown in Multimedia Appendix 1. To reduce the radiologists' workload of depicting the ROIs, this study attempted to apply partial markers to automatically generate 3D patches. The radiologists manually delineated the ROIs in the initial image, the final image, and maximum cross-section layers with the largest tumor boundaries. Based on information regarding the ROIs of the 3 key layers, 3D cubes including the whole tumor and the surrounding structures were automatically generated (Figure 1).

Figure 1. Example of the key slice–based automatic generation of a 3D cube including the whole gastric cancer region. CT: computed tomography; ROI: region of interest.

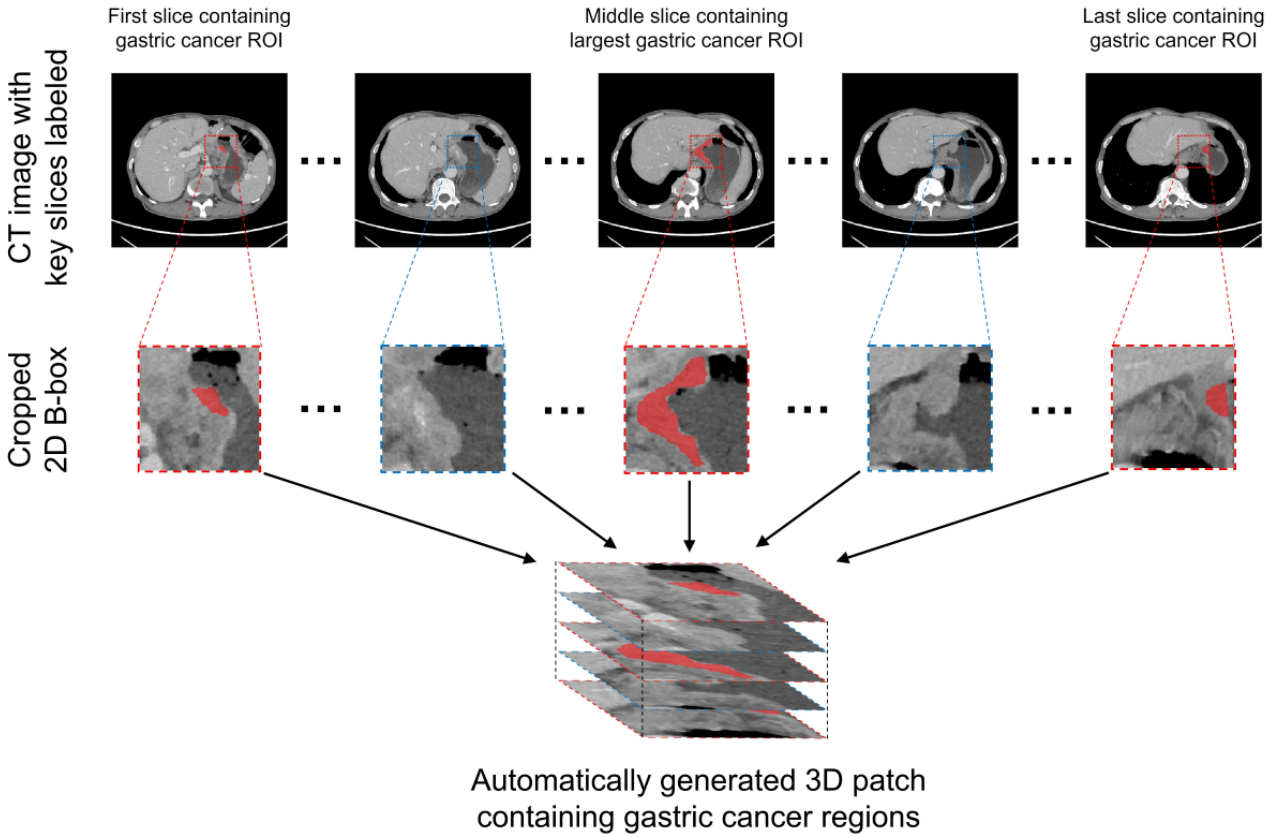

Development of the Radiomics Model

Based on the selected radiomics features, a radiomics model was constructed by using the support vector machine classifier. As a supervised learning method that was very effective in linear or nonlinear classification tasks, the support vector machine classifier has been widely used in radiomics analysis. The radiomics model was developed and validated on the InferScholar (version 3.5) platform, and the parameters were set as follows: C=1.0, kernel="Sigmoid," gamma="auto," Tol=0.001, class_weight="balanced," and other parameters were set as default.

Development of the Deep Learning Models

Vision Transformer–Based Deep Learning Model

Vision transformers (ViTs), derived by pure transformers, were first proposed for natural language processing by Vaswani et al [16]. In this study, we applied a ViT to grade GC severity based on CT images. The current implementation is inspired by the work of Alexey Dosovitskiy [17] and his team, who applied ViT to the ImageNet data set and showed excellent performance in image classification compared with state-of-the-art convolutional neural networks. Due to its advantage of analyzing overall relationships by modeling the interrelationships among different parts of an image, ViT showed strength in global image classification, which is very suitable for our case [18]. The conceptual architecture of the ViT used in our study is illustrated in Figure 2. Briefly, the 3D

XSL•FO
RenderX

patches containing GC regions were cropped from CT images. After resizing, the 3D patch was split into small patches, which, in turn, were converted to a sequence of patch embeddings via flattening and linear projection. Then, the patch embeddings and the positional embedding were fed into the transformer encoder to obtain the final representation. Consequently, the learnable features in the inputted images were fed into the classifier head to identify GC severity. In this study, the embedding dimension of the ViT models was set at 768, the numbers of both encoder layers and the attention heads were 12, and the dimensionality of the expanded representation in the predicting head was 3072.

Figure 2. Overview of the vision transformer (ViT) for the pathological severity classification of gastric cancer on noncontrast computed tomography (CT) images.

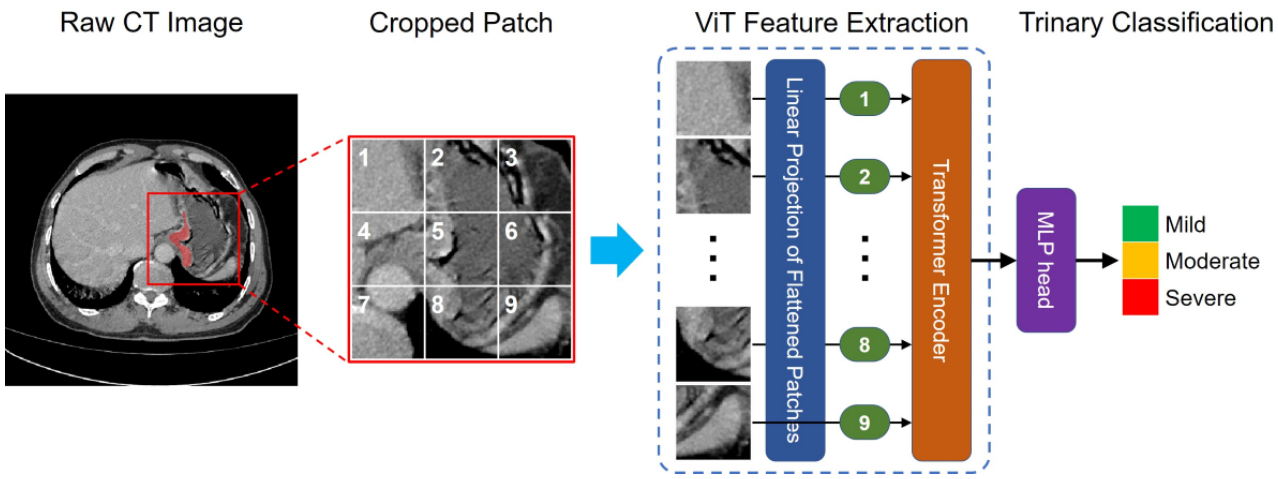

Deep Learning–Based Hybrid Model

To integrate radiomics features and deep learning features, a ViT architecture–based hybrid model was proposed in this study, namely, the ViT-Radiomics model [19,20], which combined both ViT and radiomics models for grading GC severity. In brief, the deep features extracted by the ViT were first transformed into a 1280-bit vector, while the selected radiomics features were also transformed into a 512-bit vector by using a vectorization transform approach. Then, the vectorized deep features and radiomics features were further concatenated into a 1792-bit vector and used to predict the severity of gastric cancer (mild, moderate, or severe).

Data Augmentation

To meet the high demand for large amounts of training data in ViT-based deep learning models, several sophisticated data augmentation techniques were applied in the training data set to avoid the overfitting problem and promote prediction performance during the model training process. Data augmentation techniques in this study mainly referred to a series of geometric transformations including scaling (0.9 times or 1.1 times the image size), translation (up and down, left and right, and front and back), mirroring, rotation, and flip (horizontal and vertical) on the generated 3D patches of the original images. After data augmentation, the sample size in the training set increased to 5 times that of the original training data set.

Training of the Deep Learning Models

The deep learning models (deep learning and hybrid models) were trained by maximizing their identification performance (ie, accuracy) for grading GC severity and minimizing the categorical cross-entropy loss. This study used stochastic gradient descent with a momentum of 0.95, a weight decay of 0.0001, and an initial learning rate of 0.001 to optimize the model's parameters. The number of modeling epochs was set to 150, and the minibatch size was set to 32. The model development process was implemented using 4 GeForce RTX 2080ti GPUs on Ubuntu 18.04.4 LTS, Python 3.7.11, and PyTorch 1.7.1. No samples overlapped at the patient level in the development and independent data sets.

Evaluation of Model Performance

The trinary diagnostic capability of the predictive models was assessed on the basis of overall accuracy and the Cohen κ coefficient [21], and the chi-square test was used to compare the accuracies of different predictive models. Because of the imbalanced distribution of the 3 GC severity categories, the per-class F_1-score and weighted average F_1-score were also calculated [22].

In addition, two clinically important subtasks were also evaluated: (1) binary classification of mild (stages T1~T2) GCs and nonmild GCs (stages T3~T4) and (2) binary classification of nonsevere GCs (stages T1~T3) and severe GCs (stage T4). The discriminative efficacy of binary classification was evaluated using receiver operating characteristic analysis using the area under the curve (AUC). The sensitivity, specificity, positive predictive value, and negative predictive value were also calculated under the optimal threshold in accordance with the maximum Youden index [23]. Furthermore, decision curve analysis was used to evaluate the clinical utility of the radiomics, deep learning, and hybrid models for binary classification of mild/nonmild and nonsevere/severe GCs by comparing the net benefit across a range of threshold probabilities in the training and validation data sets [24].

XSL•FO
RenderX

Statistical Analysis

Statistical analysis was performed using R (version 3.3.1; The R Foundation), SPSS (version 23.0; IBM Corp), and MedCalc (version 20.0; MedCalc). The differences in continuous variables with normal or nonnormal distributions were evaluated using a 2-tailed Student *t* test or the Mann-Whitney *U* test, respectively. Categorical variables were compared using the chi-square test. Differences between the 2 AUCs of different models were assessed by using the DeLong test [25]. A decision curve was plotted using the "rmda" package in R. A 2-sided *P* value less than .05 was considered statistically significant.

Results

Patient Characteristics

Ultimately, a total of 706 patients with GC were enrolled in this retrospective study and randomized 4:1 into a development data set (n=566) and an independent testing data set (n=140). In total, 65 patients with GC from 2 external centers were also recruited as the external testing data set (Figure S1 in Multimedia Appendix 1).

Clinical characteristics in the development, internal, and external sets are listed in Table 1. Age, peritoneum metastasis, and carbohydrate antigen 19-9 (CA19-9) differed significantly among the 3 data sets. No significant difference was observed among the 3 data sets in terms of sex, size, site, Boorman type, grade, Lauren type, adjacent tissue invasion, pathological T stage, TNM stage, serum cancer antigen 72-4 (CA72-4) levels, serum cancer antigen 125 (CA125) levels, and carcinoembryonic antigen (CEA) levels.

XSL•FO
RenderX

Table 1. Characteristics of the enrolled patients with gastric cancer in the development, internal, and external data sets.

Variables	Training (n=566)	Testing (n=140)	External (n=65)	*P* value
Sex, n (%)				.85
Male	372 (65.72)	93 (66.43)	45 (69.23)	
Female	194 (34.28)	47 (33.57)	20 (30.77)	
Age (years), median (IQR)	60 (51-67)	58 (49-65)	66 (57-70)	.001
Tumor size (cm), median (IQR)	5 (3-7)	5 (3-7)	5 (3-6)	.69
Tumor location, n (%)				.30
U	93 (16.43)	19 (13.57)	9 (13.85)	
M	73 (12.90)	16 (11.43)	7 (10.77)	
L	300 (53.00)	74 (52.86)	35 (53.85)	
U+E	7 (1.24)	9 (6.43)	3 (4.62)	
UM/MU	17 (3.00)	6 (4.29)	3 (4.62)	
LM/ML	45 (7.95)	11 (7.86)	4 (6.15)	
L+D	7 (1.24)	1 (0.71)	0 (0)	
LMU/MLU/MUL/UML	24 (4.24)	4 (2.86)	4 (6.15)	
Peritoneal metastasis, n (%)				.03
Negative	363 (64.13)	95 (67.86)	52 (80.00)	
Positive	203 (35.87)	45 (32.14)	13 (20.00)	
Borrmann classification, n (%)				.11
Early stage	94 (16.61)	22 (15.71)	1 (1.54)	
I	19 (3.36)	6 (4.29)	3 (4.62)	
II	157 (27.74)	45 (32.14)	18 (27.69)	
III	235 (41.52)	52 (37.14)	34 (52.31)	
IV	61 (10.78)	15 (10.71)	9 (13.85)	
Grade, n (%)				.14
Unknown	14 (2.48)	6 (4.29)	3 (4.62)	
1	217 (38.34)	54 (38.57)	20 (30.8)	
2	325 (57.42)	78 (55.71)	37 (56.92)	
3	1 (0.18)	0 (0)	0 (0)	
4	9 (1.59)	2 (1.43)	5 (7.7)	
Lauren histological type, n (%)				.67
Intestinal	170 (30.04)	40 (28.57)	15 (23.08)	
Diffuse	249 (43.99)	62 (44.29)	28 (43.08)	
Mixed	147 (25.97)	38 (27.14)	22 (33.85)	
Adjacent tissues invaded by tumors, n (%)				.28
Negative	502 (88.69)	127 (90.71)	54 (83.08)	
Positive	64 (11.31)	13 (9.29)	11 (16.92)	
T stage, n (%)				.22
T1a	57 (10.07)	14 (10.00)	2 (3.08)	
T1b	57 (10.07)	15 (10.71)	3 (4.62)	
T2	102 (18.02)	24 (17.14)	8 (12.30)	
T3	108 (19.08)	27 (19.29)	11 (16.92)	
T4a	187 (33.04)	49 (35.00)	34 (52.31)	

XSL•FO
RenderX

Variables	Training (n=566)	Testing (n=140)	External (n=65)	*P* value
T4b	55 (9.72)	11 (7.86)	7 (10.77)	
TNM[a] stage, n (%)				.88
IA	33 (5.83)	9 (6.43)	5 (7.69)	
IB	59 (10.42)	15 (10.71)	7 (10.77)	
IIA	77 (13.60)	12 (8.57)	5 (7.69)	
IIB	75 (13.25)	25 (17.86)	8 (12.31)	
IIIA	56 (9.89)	11 (7.86)	7 (10.77)	
IIIB	80 (14.13)	20 (14.29)	8 (12.31)	
IIIC	94 (16.61)	24 (17.14)	10 (15.38)	
IV	92 (16.25)	24 (17.14)	15 (23.08)	
Serum cancer antigen 72-4 level (ng/mL), median (IQR)	3.55 (1.78-8.58)	3.46 (1.89-7.59)	3.04 (1.20-10.55)	.70
Carbohydrate antigen 19-9 level (kU/L), median (IQR)	10.50 (6.44-20.71)	9.50 (5.88-18.09)	13.20 (8.45-27.20)	.03
Serum cancer antigen level (U/mL), median (IQR)	11.50 (8.65-17.23)	12.12 (8.16-19.48)	10.00 (7.13-13.82)	.07
Carcinoembryonic antigen level (ng/mL), median (IQR)	2.00 (1.32-3.28)	2.19 (1.40-4.15)	1.77 (1.06-4.15)	.45

[a]TNM: tumor - node - metastasis.

Performance Evaluation of the Trinary Classification of GC Severity

The 3-class confusion matrices of the predictive models in the development, internal testing, and external testing data sets are shown in Figure 3. The overall accuracies of the radiomics, deep learning, and hybrid models were 69.8%, 75.1%, and 83.7% in the development data set; 64.3%, 75.7%, and 81.4% in the internal testing data set; and 53.8%, 70.8%, and 81.5% in the external testing data set, respectively. The accuracy of the deep learning and hybrid models were significantly different from those of the radiomics model in the development testing (deep learning vs radiomics: P=.046; hybrid vs radiomics: P<.001), internal testing (deep learning vs radiomics: P=.04; hybrid vs radiomics: P=.001), and external testing (deep learning vs radiomics: P=.047; hybrid vs radiomics: P<.001) data sets. The overall accuracy of the hybrid model was significantly different from that of the deep learning model in the development data set (P<.001); however, no significant difference was observed in the internal (P=.25) or external (P=.15) testing data sets. Similarly, the predicted results of both the deep learning and hybrid models showed high agreement with the pathologically confirmed ground truth in the development (Cohen κ=0.660 for the deep learning model and 0.767 for the hybrid model), internal testing (Cohen κ=0.665 for the deep learning model and 0.735 for the hybrid model), and external testing (Cohen κ=0.689 for the hybrid model) data sets, while the agreement for the radiomics model was only moderate (Cohen κ=0.579 for the development, 0.518 for the internal testing, and 0.400 for the external testing data sets). These results, as well as the per-class F_1-score and weighted average F_1-score, are presented in Table 2.

XSL•FO
RenderX

Figure 3. Trinary classification performance of the predictive models. (A)-(C) The 3-class confusion matrix of the radiomics, deep learning, and hybrid models in the development data set. (D)-(F) The 3-class confusion matrix of the radiomics, deep learning, and hybrid models in the internal testing data set. (G)-(I) The 3-class confusion matrix of the radiomics, deep learning, and hybrid models in the external testing data set.

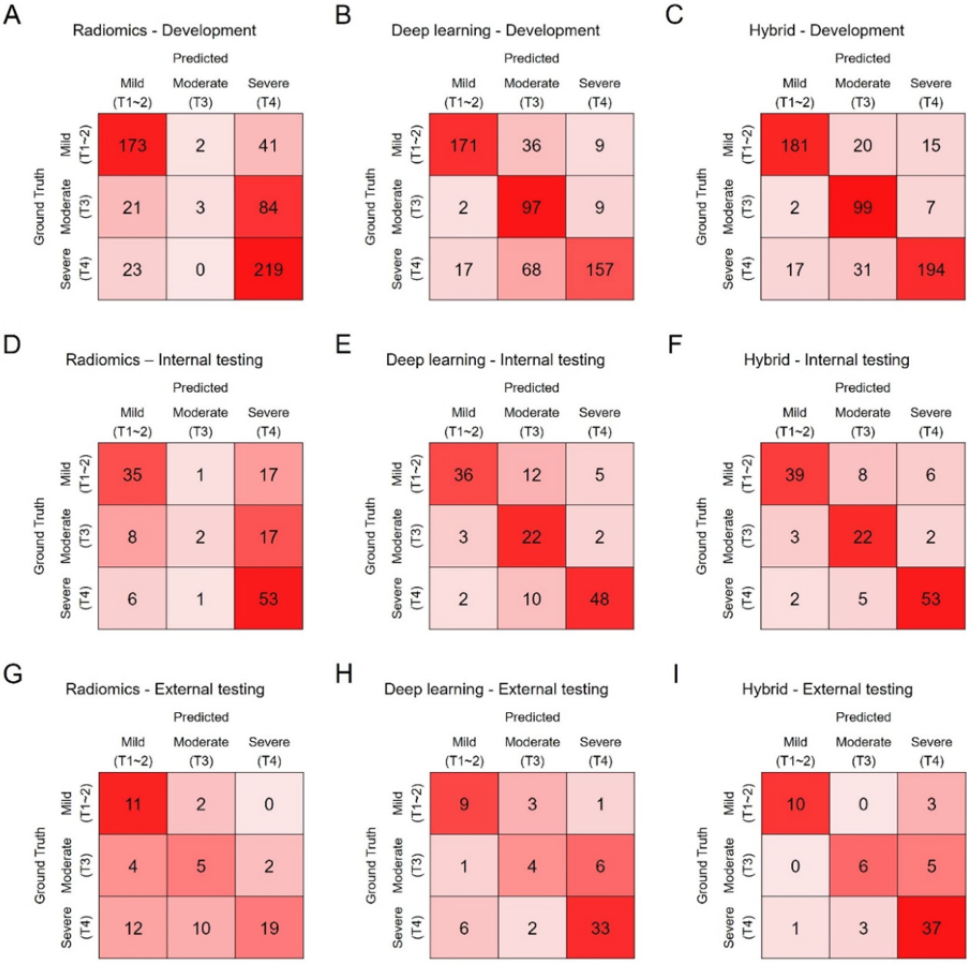

Table 2. Trinary classification performance of the models in predicting the T stage of gastric cancer.

Data set and model	Accuracy, %	F_1-score (mild)	F_1-score (moderate)	F_1-score (severe)	F_1-score (average)	Cohen κ
Development data set						
Radiomics	69.8	0.799	0.053	0.747	0.635	0.579
Deep learning	75.1	0.842	0.628	0.753	0.763	0.660
Hybrid	83.7	0.870	0.767	0.847	0.841	0.767
Internal testing data set						
Radiomics	64.3	0.686	0.129	0.721	0.594	0.518
Deep learning	75.7	0.766	0.620	0.835	0.767	0.665
Hybrid	81.4	0.804	0.710	0.876	0.817	0.735
External testing data set						
Radiomics	53.8	0.550	0.357	0.613	0.557	0.400
Deep learning	70.8	0.621	0.400	0.815	0.706	0.544
Hybrid	81.5	0.833	0.600	0.860	0.811	0.689

Evaluation on 2 Binary Classification Subtasks

As depicted in Figure 4, the binary classification performance of the predictive models for mild versus nonmild and nonsevere versus severe GCs was evaluated in both internal and external testing data sets. For binary classification of mild (stage T1~T2) and nonmild (stage T3~T4) GCs, the radiomics, deep learning, and hybrid models showed similar performance, with AUCs of 0.875 (95% CI 0.809-0.925), 0.866 (95% CI 0.799-0.918), and 0.886 (95% CI 0.822-0.934) in the internal testing data set,

XSL•FO
RenderX

respectively (Delong test; *P*>.05 for all). Meanwhile, The AUC values of the hybrid, radiomics, and deep learning models in the external testing data set were 0.972 (95% CI 0.897-0.997), 0.820 (95% CI 0.704-0.904; *P*=.002), and 0.818 (95% CI 0.703-0.903; *P*=.03), respectively. For binary classification of nonsevere (stage T1~T3) and severe (stage T4) GCs, the AUC values of deep learning and hybrid models were 0.892 (95% CI 0.829-0.938) and 0.894 (95% CI 0.831-0.940), respectively. Furthermore, the AUC value of the radiomics model was 0.815 (95% CI 0.740-0.875; radiomics vs deep learning: *P*=.02; radiomics vs hybrid: *P*=.03). In addition, the AUC values for the deep learning, hybrid, and radiomics models were 0.808 (95% CI 0.691-0.895), 0.897 (95% CI 0.797-0.959), and 0.685 (95% CI 0.558-0.795), respectively (radiomics vs deep learning: *P*=.03; radiomics vs hybrid: *P*=.002), in the external testing data set, and no significant difference was observed between them (*P*=.10). The detailed performances of these models in the internal and external testing data sets are listed in Tables 3 and 4, respectively.

Figure 4. Comparison between the predictive models on 2 binary subtasks. AUC: area under the curve.

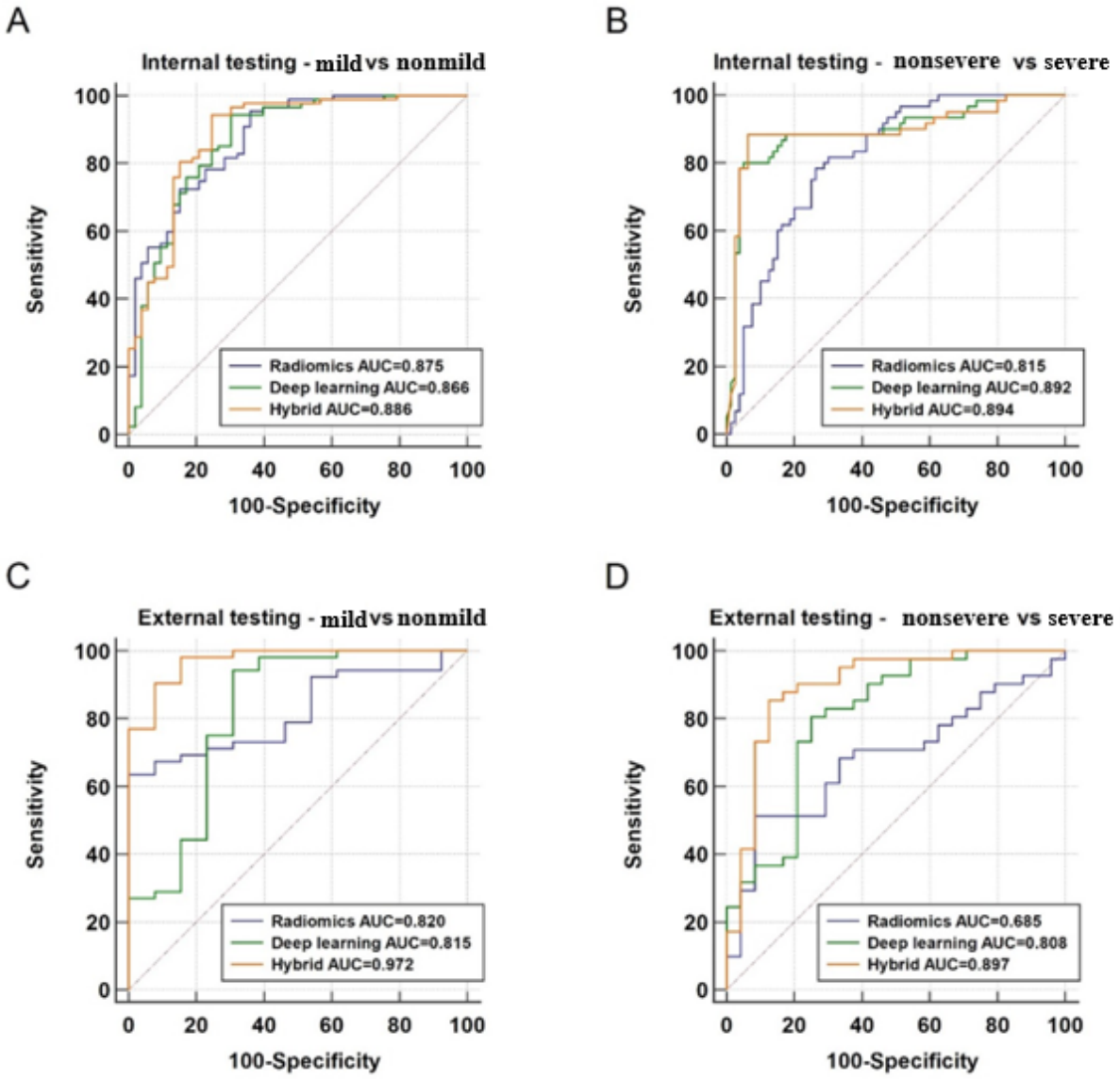

XSL•FO
RenderX

Table 3. Comparison of the performance of the predictive models for binary classification of mild and nonmild gastric cancers in the internal and external testing data sets.

Data set and model	Area under the curve (95% CI)	Sensitivity, %	Specificity, %	Positive predictive value, %	Negative predictive value, %
Internal testing data set					
Radiomics	0.875 (0.809-0.925)	95.4	64.2	81.4	89.5
Deep learning	0.866 (0.799-0.918)	94.3	69.8	83.7	88.1
Hybrid	0.886 (0.822-0.934)	94.3	75.5	86.3	88.9
External testing data set					
Radiomics	0.820 (0.704-0.904)	63.5	100	100	40.6
Deep learning	0.818 (0.703-0.903)	94.2	69.2	92.5	75.0
Hybrid	0.972 (0.897-0.997)	90.4	92.3	97.9	70.6

Table 4. Comparison of the performance of the predictive models for binary classification of nonsevere and severe gastric cancers in the internal and external testing data sets.

Data set and model	Area under the curve (95% CI)	Sensitivity, %	Specificity, %	Positive predictive value, %	Negative predictive value, %
Internal testing data set					
Radiomics	0.815 (0.740-0.875)	78.3	73.8	69.1	81.9
Deep learning	0.892 (0.829-0.938)	80.0	95.0	92.3	86.4
Hybrid	0.894 (0.831-0.940)	88.3	93.8	91.4	91.5
External testing data set					
Radiomics	0.685 (0.558-0.795)	51.2	91.7	91.3	52.4
Deep learning	0.808 (0.691-0.895)	80.5	75.0	84.6	69.2
Hybrid	0.897 (0.797-0.959)	85.4	87.5	92.1	77.8

Clinical Utility Analysis

Decision curve analyses for the 2 binary classification subtasks of the different predictive models in the internal and external data sets are presented in Figure 5. During binary classification of mild (stage T1~T2) and nonmild (stage T3~T4) GCs, the hybrid model had a slightly higher overall net benefit than the radiomics and deep learning models across the majority range of reasonable threshold probabilities in the internal testing data set. For binary classification of nonsevere (stage T1~T3) and severe (stage T4) GCs, the deep learning and hybrid models showed a markedly higher net benefit than the radiomics model across the majority range of reasonable threshold probabilities in the internal testing data set. A similar tendency was also observed in the external data set (Figures 5C and 5D). These results are consistent with those of the receiver operating characteristic analysis.

XSL•FO
RenderX

Figure 5. Decision curve analysis (DCA) for the predictive models in the internal testing data set on (A) the binary classification subtask 1 (stage T1~T2 vs stage T3~T4) and (B) the binary classification subtask 2 (stage T1~T3 vs stage T4) and for the predictive models on binary classification subtask 1 (C) and 2 (D) in the external testing data set. The net benefit is depicted on the y-axis. The gray line and black line represent situations in which all patients and no patients underwent biopsy/surgery, respectively.

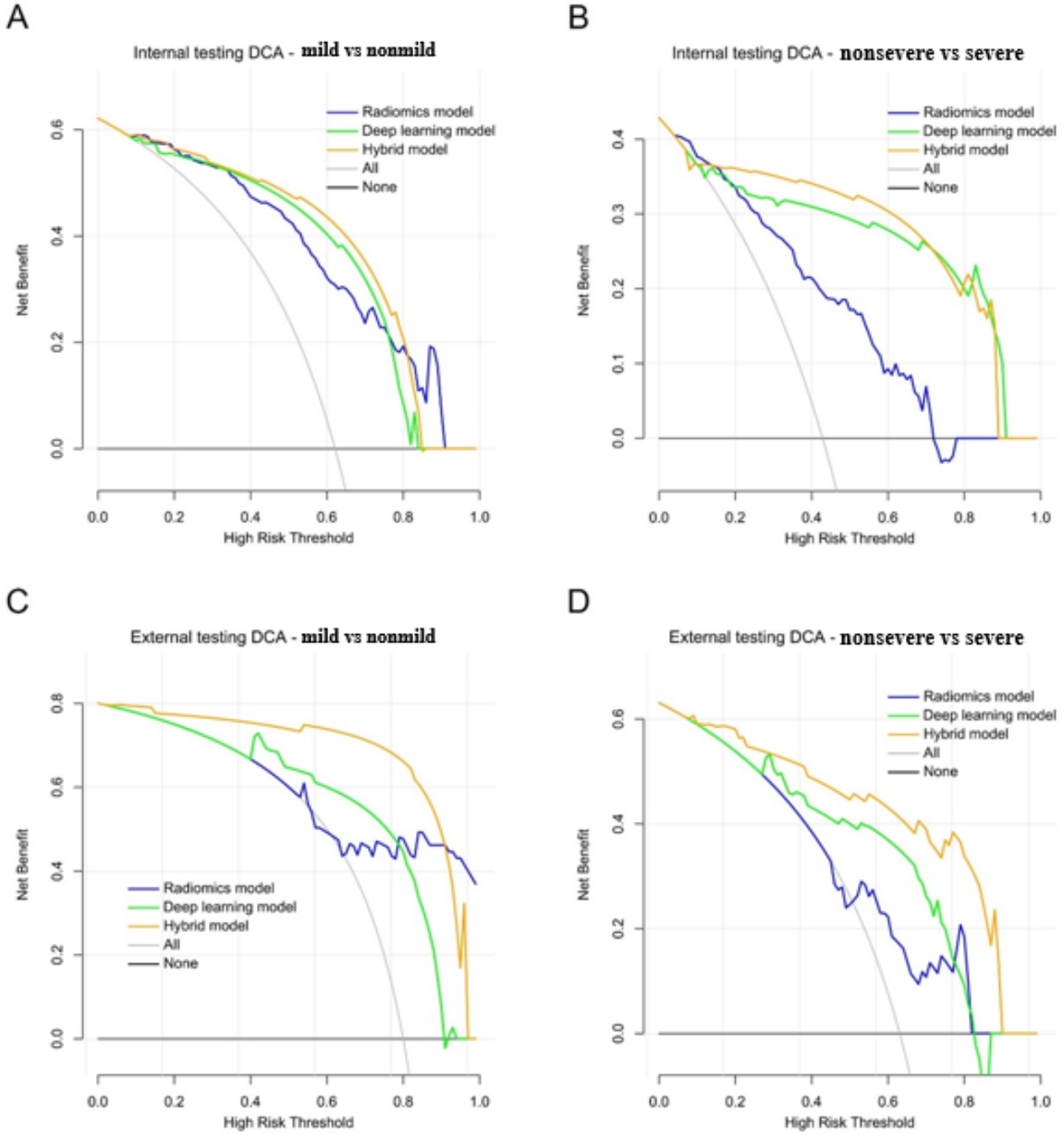

Visualization of the Internal Features Learned by the Neural Networks

The internal features learned by the hybrid model were examined in the internal testing data set by using the *t*-distributed stochastic neighbor embedding method [26]. Each point represented an input noncontrast CT image of a patient projected from the high-dimensional vector of the neural network's last hidden layer into 2 dimensions. The mild (blue point cloud) and severe (red point cloud) groups showed clear clustering patterns and were split across the moderate group (green point cloud; Figure 6).

XSL•FO
RenderX

Figure 6. Visualization of the internal representations of the hybrid model for three severity classes by t-distributed stochastic neighbor embedding. Colored point clouds represented the different severity categories, showing how the neural network clustered the diseases.

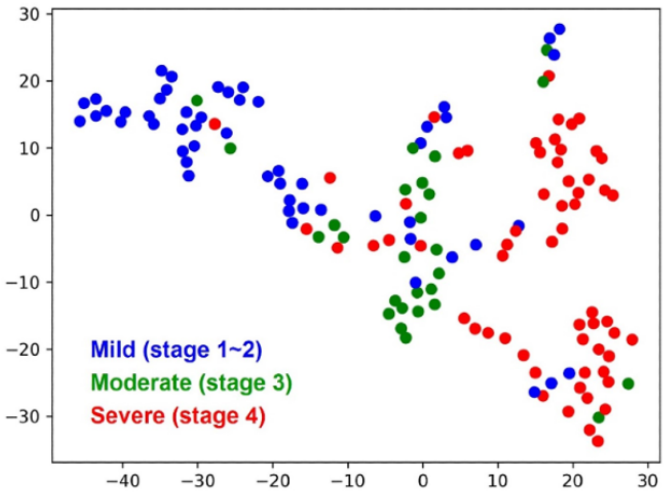

Discussion

Principal Findings and Comparison With Prior Work

Preoperative evaluation of tumor invasion depth determines individual treatment plans for GC. CT has been the first choice for preoperative GC evaluation and is important for clinical practice. However, routine preoperative determination is still not accurate enough, especially when radiologists carry out naked-eye evaluations. This study makes extensive use of the CT radiomics signature and the signature extracted by deep learning to generate a better prediction model to overcome the heterogeneity resulting from naked-eye evaluations. To our knowledge, this is the first study to demonstrate a CT-based model using radiomics and deep learning techniques for automated prediction of the T stage of GC. Our hybrid model demonstrated superior clinical utility to that of both the radiomics and deep learning models individually. Moreover, the hybrid model exhibited promising diagnostic performance in determining the pathological stage of GC.

Over the past decade, radiomics research has gained increasing attention, and a growing number of studies reported the use of radiomics in oncology [27]. At the very least, this could be an important complement to subjective evaluation by radiologists. Previous studies have illustrated that CT-based radiomics used alone or with deep learning or other statistical methods could predict the number of lymph node metastases [11,28], chemotherapy treatment responses, and overall survival [10,29] among patients with GC. Radiomics features are mathematically defined descriptors, while deep learning features are less intuitive due to the complexity of deep neural networks. The predictive values of the two are different and stackable [30]. A pilot study conducted by Sun et al [31] based on radiomics analysis reported that the model had a predictive AUC of 0.852 for the diagnostic T stage in rectal cancer. Yang et al [32] demonstrated that CT radiomics signatures exhibited favorable predictive performance for the T stage in esophageal carcinoma, with an AUC of 0.86. Furthermore, our study shows that the hybrid (83.7%) and deep learning (75.1%) models had higher accuracy than the radiomics model (69.8%) in evaluating the T stage based on 3 grades (mild, moderate, and severe groups) in the development, internal training, and external training data sets. In addition, the hybrid and deep learning models showed better performance than the radiomics model, with high agreement between the model's prediction and ground truth. Thus, using the hybrid and deep learning models for pathological prediction is eligible and reliable. Furthermore, the hybrid model had higher accuracy than the deep learning model in the development data set ($P<.001$); however, no significant difference was observed in the internal ($P=.25$) or external ($P=.15$) testing data sets, which was possibly due to the limited sample sizes therein. By using the hybrid model, the diagnostic accuracy of the T stage could be improved, which can help clinicians make more precise treatment plans for patients with GC to avoid treatment delays or mistreatment. Furthermore, deep learning can reduce the workload of doctors and improve work efficiency.

Additionally, we aimed to better explore the predictive models constructed by the deep learning and hybrid models. The tumor invasion depth was evaluated separately in mild versus nonmild and nonsevere versus severe GCs. The radiomics, deep learning, and hybrid models displayed good performance in differentiating between mild and nonmild GC and severe and nonsevere GC in both the internal and external training data sets. Meanwhile, the hybrid model (AUC 0.972, 95% CI 0.897-0.997) outperformed both the radiomics (AUC 0.820, 95% CI 0.704-0.904; $P=.002$) and deep learning (AUC 0.818, 95% CI 0.703-0.903; $P=.03$) models in the external testing data set. Based on the hybrid model, for patients with stage T1 or T2 lesions, endoscopic resection or surgery was considered in combination with other examinations, while for patients with T3 or T4 lesions, adjuvant therapy was recommended. One retrospective study involving 572 patients with GC diagnosed at pathological stage T3 or T4 showed that a radiomics model based on CT images with deep learning is effective in discriminating serosa invasion in GC [33], which is consistent with our results. Patients with GC with stage T1~T3 disease have a lower risk of peritoneal metastasis than those with stage

XSL•FO
RenderX

T4 disease, while in those with stage T4 disease, the hybrid model could assist clinicians in improving the detection accuracy of peritoneal metastasis, especially for regions where staging laparoscopy was not applicable widely. Furthermore, decision curve analysis revealed that the deep learning and hybrid models showed obviously higher net benefit than the radiomics model across the majority of the range of reasonable threshold probabilities, implying that the deep learning and hybrid models have certain clinical applicability. Furthermore, *t*-distributed stochastic neighbor embedding analysis revealed that different GC risk stratifications have their own distinct clusters. Based on the above results, the deep learning and hybrid models have good performance in distinguishing T stages in GC. Additionally, the hybrid model might perform better in distinguishing T stages than the deep learning and radiomics models.

Clinical Implications

With respect to patient outcomes, the hybrid model can accurately predict the T stage directly from conventional CT images through automated processes. Furthermore, the hybrid model shows promise in aiding clinicians by offering a more dependable and accurate preoperative T staging diagnosis, which can influence real-world clinical decision-making. For example, by providing accurate T staging, neoadjuvant therapy can be arranged for patients with advanced GC. Through pretreatment evaluation of T stages, the hybrid model could help clinicians choose the correct treatment method. Following surgical procedures, we offer individualized treatment plans for patients through precise pathological diagnosis and selection of suitable adjuvant therapies.

Ethical Implications

A prior investigation identified 3 key domains of ethical concerns related to the use of artificial intelligence (AI): algorithms, data, and practices [34]. These domains highlight the importance of obtaining informed consent and establishing data use agreements between data providers and third-party data aggregators. Furthermore, ensuring the quality of data used in AI algorithms, especially in the context of enhancing patient treatment decisions, is a significant area of concern. Furthermore, the integration of AI technology should not be viewed as a substitute for the collaborative decision-making process that is integral to optimal patient care. It is important to juxtapose these findings with the standard clinical practices currently used for T staging in GC.

Limitations

There are several limitations in the study. First, this was a retrospective study that had a selection bias, and the limited sample size of images also contributed to a bias in model construction. Hence, for further validation and to enhance the generalizability of our results, a future prospective study with a large sample size is necessary. Second, our method requires manual segmentation of the tumor, which is time-consuming. Automated segmentation methods, as well as fully automated models, could be valuable in the future. Additionally, when tumor ROIs were outlined manually, there was some heterogeneity in the radiologists' experience, and the radiologists needed to be calibrated. In addition, our study population included only Asian individuals; future studies from other centers are necessary to confirm the validity of this model prior to its potential clinical implementation. Furthermore, AI lacks the capacity to participate in complex dialogues with patients, as well as the ability to establish the necessary trust and empathy essential for fostering a therapeutic alliance that is crucial to the patient-physician relationship and to favorable treatment outcomes.

Conclusions

The hybrid model can potentially distinguish T stages of GC more effectively than the deep learning or radiomics model and could be applied in clinical practice.

Acknowledgments

This study has received funding from the National Natural Science Foundation of China (No 82472926), and the Foundation of Science & Technology Department of Sichuan Province (2023YFS0060), and the 1. 3. 5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYJC21006).

Data Availability

The data and code are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[PDF File (Adobe PDF File), 184 KB - jmir_v26i1e56851_app1.pdf]

References

XSL•FO
RenderX

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May 04;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Ajani JA, D'Amico TA, Bentrem DJ, Chao J, Cooke D, Corvera C, et al. Gastric cancer, version 2.2022, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2022 Feb;20(2):167-192. [doi: 10.6004/jnccn.2022.0008] [Medline: 35130500]
3. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 2010 Jun 24;17(6):1471-1474. [doi: 10.1245/s10434-010-0985-4] [Medline: 20180029]
4. Joshi SS, Badgwell BD. Current treatment and recent progress in gastric cancer. CA Cancer J Clin 2021 May 16;71(3):264-279 [FREE Full text] [doi: 10.3322/caac.21657] [Medline: 33592120]
5. Nanishi K, Shoda K, Kubota T, Kosuga T, Konishi H, Shiozaki A, et al. Diagnostic accuracy of the gastric cancer T-category with respect to tumor localization. Langenbecks Arch Surg 2020 Sep 26;405(6):787-796 [FREE Full text] [doi: 10.1007/s00423-020-01971-3] [Medline: 32851434]
6. Ahn HS, Lee H, Yoo M, Kim SG, Im JP, Kim SH, et al. Diagnostic accuracy of T and N stages with endoscopy, stomach protocol CT, and endoscopic ultrasonography in early gastric cancer. J Surg Oncol 2009 Jan 01;99(1):20-27. [doi: 10.1002/jso.21170] [Medline: 18937292]
7. Wang Y, Liu W, Yu Y, Liu J, Jiang L, Xue H, et al. Prediction of the depth of tumor invasion in gastric cancer: potential role of CT radiomics. Acad Radiol 2020 Aug;27(8):1077-1084. [doi: 10.1016/j.acra.2019.10.020] [Medline: 31761666]
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
9. Chan H, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020;1213:3-21 [FREE Full text] [doi: 10.1007/978-3-030-33128-3_1] [Medline: 32030660]
10. Jiang Y, Jin C, Yu H, Wu J, Chen C, Yuan Q, et al. Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. Ann Surg 2021 Dec 01;274(6):e1153-e1161. [doi: 10.1097/SLA.0000000000003778] [Medline: 31913871]
11. Dong D, Fang M, Tang L, Shan X, Gao J, Giganti F, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Ann Oncol 2020 Jul;31(7):912-920 [FREE Full text] [doi: 10.1016/j.annonc.2020.04.003] [Medline: 32304748]
12. Liu D, Zhang W, Hu F, Yu P, Zhang X, Yin H, et al. A bounding box-based radiomics model for detecting occult peritoneal metastasis in advanced gastric cancer: a multicenter study. Front Oncol 2021 Dec 3;11:777760 [FREE Full text] [doi: 10.3389/fonc.2021.777760] [Medline: 34926287]
13. Huang Z, Liu D, Chen X, He D, Yu P, Liu B, et al. Deep convolutional neural network based on computed tomography images for the preoperative diagnosis of occult peritoneal metastasis in advanced gastric cancer. Front Oncol 2020 Nov 2;10:601869 [FREE Full text] [doi: 10.3389/fonc.2020.601869] [Medline: 33224893]
14. Kirienko M, Sollini M, Silvestri G, Mognetti S, Voulaz E, Antunovic L, et al. Convolutional neural networks promising in lung cancer T-parameter assessment on baseline FDG-PET/CT. Contrast Media Mol Imaging 2018 Oct 30;2018:1382309 [FREE Full text] [doi: 10.1155/2018/1382309] [Medline: 30510492]
15. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017 Nov 01;77(21):e104-e107 [FREE Full text] [doi: 10.1158/0008-5472.CAN-17-0339] [Medline: 29092951]
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 6000-6010.
17. Alexey D. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. Preprint posted online October 22, 2020 2020.
18. Kim S, Nam J, Ko BC. Facial expression recognition based on squeeze vision transformer. Sensors (Basel) 2022 May 13;22(10):3729 [FREE Full text] [doi: 10.3390/s22103729] [Medline: 35632135]
19. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 2023 Jan 1;45(1):87-110. [doi: 10.1109/tpami.2022.3152247]
20. Leamons R, Cheng H, Al Shami A. Vision Transformers for Medical Images Classifications. In: Lecture Notes in Networks and Systems, vol 544. 2022 Presented at: IntelliSys 2022: Intelligent Systems and Applications; 2022; Amsterdam. [doi: 10.1007/978-3-031-16075-2_22]
21. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/bm.2012.031]
22. Zhao W, Yang J, Sun Y, Li C, Wu W, Jin L, et al. 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. Cancer Res 2018 Dec 15;78(24):6881-6889. [doi: 10.1158/0008-5472.CAN-18-0696] [Medline: 30279243]
23. Yin J, Samawi H, Tian L. Joint inference about the AUC and Youden index for paired biomarkers. Stat Med 2022 Jan 15;41(1):37-64. [doi: 10.1002/sim.9222] [Medline: 34964512]

XSL•FO
RenderX

24. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2016 Sep 05;26(6):565-574. [doi: 10.1177/0272989x06295361]
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837. [doi: 10.2307/2531595]
26. van der Maaten V, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9(2605):2579-2605.
27. Chen Q, Zhang L, Liu S, You J, Chen L, Jin Z, et al. Radiomics in precision medicine for gastric cancer: opportunities and challenges. Eur Radiol 2022 Sep 22;32(9):5852-5868. [doi: 10.1007/s00330-022-08704-8] [Medline: 35316364]
28. Jin C, Jiang Y, Yu H, Wang W, Li B, Chen C, et al. Deep learning analysis of the primary tumour and the prediction of lymph node metastases in gastric cancer. Br J Surg 2021 May 27;108(5):542-549. [doi: 10.1002/bjs.11928] [Medline: 34043780]
29. Zhang L, Dong D, Zhang W, Hao X, Fang M, Wang S, et al. A deep learning risk prediction model for overall survival in patients with gastric cancer: a multicenter study. Radiother Oncol 2020 Sep;150:73-80. [doi: 10.1016/j.radonc.2020.06.010] [Medline: 32540334]
30. Hao D, Li Q, Feng Q, Qi L, Liu X, Arefan D, et al. Identifying prognostic markers from clinical, radiomics, and deep learning imaging features for gastric cancer survival prediction. Front Oncol 2021 Feb 2;11:725889 [FREE Full text] [doi: 10.3389/fonc.2021.725889] [Medline: 35186707]
31. Sun Y, Hu P, Wang J, Shen L, Xia F, Qing G, et al. Radiomic features of pretreatment MRI could identify T stage in patients with rectal cancer: preliminary findings. J Magn Reson Imaging 2018 Feb 13;48(3):615-621. [doi: 10.1002/jmri.25969] [Medline: 29437279]
32. Yang M, Hu P, Li M, Ding R, Wang Y, Pan S, et al. Computed tomography-based radiomics in predicting T stage and length of esophageal squamous cell carcinoma. Front Oncol 2021 Oct 14;11:722961 [FREE Full text] [doi: 10.3389/fonc.2021.722961] [Medline: 34722265]
33. Sun R, Fang M, Tang L, Li X, Lu Q, Dong D, et al. CT-based deep learning radiomics analysis for evaluation of serosa invasion in advanced gastric cancer. Eur J Radiol 2020 Nov;132:109277. [doi: 10.1016/j.ejrad.2020.109277] [Medline: 32980726]
34. Kohli M, Geis R. Ethics, artificial intelligence, and radiology. J Am Coll Radiol 2018 Sep;15(9):1317-1319. [doi: 10.1016/j.jacr.2018.05.020] [Medline: 30017625]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CA19-9: carbohydrate antigen 19-9
CA72-4: serum cancer antigen 72-4
CA125: serum cancer antigen 125
CEA: carcinoembryonic antigen
CT: computed tomography
GC: gastric cancer
ICC: intraclass correlation coefficient
ROI: region of interest
TNM: tumor - node - metastasis
ViT: vision transformer

Edited by T de Azevedo Cardoso; submitted 31.01.24; peer-reviewed by MB Mourato, A Shour, GK Gupta; comments to author 13.03.24; revised version received 07.04.24; accepted 02.08.24; published 09.10.24.

Please cite as:
Tao J, Liu D, Hu FB, Zhang X, Yin H, Zhang H, Zhang K, Huang Z, Yang K
Development and Validation of a Computed Tomography–Based Model for Noninvasive Prediction of the T Stage in Gastric Cancer: Multicenter Retrospective Study
J Med Internet Res 2024;26:e56851
URL: https://www.jmir.org/2024/1/e56851
doi:10.2196/56851
PMID:39382960

©Jin Tao, Dan Liu, Fu-Bi Hu, Xiao Zhang, Hongkun Yin, Huiling Zhang, Kai Zhang, Zixing Huang, Kun Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.10.2024. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Lessons From 3 Longitudinal Sensor-Based Human Behavior Assessment Field Studies and an Approach to Support Stakeholder Management: Content Analysis

Johanna Kallio[1], MSc, PhD; Atte Kinnula[1], MSc, PhD; Satu-Marja Mäkelä[1], MSc; Sari Järvinen[1], MSc; Pauli Räsänen[1], MSc; Simo Hosio[2], MSc, PhD; Miguel Bordallo López[1,2], MSc, PhD

[1]VTT Technical Research Centre of Finland Ltd, Oulu, Finland

[2]Center for Machine Vision and Signal Analysis (CMVS), Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland

Corresponding Author:
Johanna Kallio, MSc, PhD
VTT Technical Research Centre of Finland Ltd
Kaitoväylä 1
Oulu, 90571
Finland
Phone: 358 50 527 0180
Email: johanna.kallio@vtt.fi

Abstract

Background: Pervasive technologies are used to investigate various phenomena outside the laboratory setting, providing valuable insights into real-world human behavior and interaction with the environment. However, conducting longitudinal field trials in natural settings remains challenging due to factors such as low recruitment success and high dropout rates due to participation burden or data quality issues with wireless sensing in changing environments.

Objective: This study gathers insights and lessons from 3 real-world longitudinal field studies assessing human behavior and derives factors that impacted their research success. We aim to categorize challenges, observe how they were managed, and offer recommendations for designing and conducting studies involving human participants and pervasive technology in natural settings.

Methods: We developed a qualitative coding framework to categorize and address the unique challenges encountered in real-life studies related to influential factor identification, stakeholder management, data harvesting and management, and analysis and interpretation. We applied inductive reasoning to identify issues and related mitigation actions in 3 separate field studies carried out between 2018 and 2022. These 3 field studies relied on gathering annotated sensor data. The topics involved stress and environmental assessment in an office and a school, collecting self-reports and wrist device and environmental sensor data from 27 participants for 3.5 to 7 months; work activity recognition at a construction site, collecting observations and wearable sensor data from 15 participants for 3 months; and stress recognition in location-independent knowledge work, collecting self-reports and computer use data from 57 participants for 2 to 5 months. Our key extension for the coding framework used a stakeholder identification method to identify the type and role of the involved stakeholder groups, evaluating the nature and degree of their involvement and influence on the field trial success.

Results: Our analysis identifies 17 key lessons related to planning, implementing, and managing a longitudinal, sensor-based field study on human behavior. The findings highlight the importance of recognizing different stakeholder groups, including those not directly involved but whose areas of responsibility are impacted by the study and therefore have the power to influence it. In general, customizing communication strategies to engage stakeholders on their terms and addressing their concerns and expectations is essential, while planning for dropouts, offering incentives for participants, conducting field tests to identify problems, and using tools for quality assurance are relevant for successful outcomes.

Conclusions: Our findings suggest that field trial implementation should include additional effort to clarify the expectations of stakeholders and to communicate with them throughout the process. Our framework provides a structured approach that can be adopted by other researchers in the field, facilitating robust and comparable studies across different contexts. Constantly managing the possible challenges will lead to better success in longitudinal field trials and developing future technology-based solutions.

(J Med Internet Res 2024;26:e50461) doi:10.2196/50461

XSL•FO
RenderX

KEYWORDS

field trial; behavioral research; sensor data; machine learning; pervasive technology; stakeholder engagement; qualitative coding; mobile phone

Introduction

Background and Objectives

Human and environmental monitoring in natural settings is essential in behavioral research. Pervasive technologies such as sensors and machine learning enable the unobtrusive assessment of how individuals behave and interact with their natural environment [1]. Application of such technologies has shown great potential for accurately recognizing emotions and activities in laboratory settings [2,3]. However, brief emotional or activity stimuli induced during an experimental protocol may not fully capture the emotional responses or behaviors occurring in real-world situations. Further studies in authentic environments are required to investigate and validate people's behavior and the feasibility of novel technologies in naturalistic contexts [4]. Therefore, transitioning research from the laboratory to real-life settings is essential, but it poses additional challenges due to the lack of control over contextual factors and human behaviors [5].

Field studies conducted in naturalistic settings are often regarded as more ecologically valid than laboratory studies, providing a more realistic representation of people's behavior [4]. These studies frequently involve collecting and analyzing data on people and their surroundings over an extended period to understand the interaction between the users, technology, and contexts and evaluate technologies in uncontrolled situations [6]. However, field trials tend to face additional challenges related to people, changing contexts, and new technology, which are often underreported or dispersed across studies.

To address this shortcoming, this work aims to provide insights and lessons on the challenges and their mitigations that impact the success of real-world trials.

This analysis involves 3 independent longitudinal field studies, each focusing on a different aspect of human behavior and environmental monitoring. Although these field studies are not directly related to one another, they collectively contribute to a broader understanding of the challenges and solutions in longitudinal field trials by answering the following research questions:

1. What kinds of unanticipated challenges arise in longitudinal, sensor-based field trials studying human behavior or activity?
2. How can these challenges be managed?

The main contributions of this work are the categorization of the challenges encountered across the 3 longitudinal field studies and recommendations for conducting real-life studies that assess humans and their environment using pervasive technology.

Challenges in Field Trials Monitoring Humans and the Environment

A field trial involves collecting data from participants and possibly their environment, typically over a long period. The collected data can be used to develop new technologies and interventions to improve human well-being. The main phases in designing and implementing such a study include defining the research question and objectives, identifying and selecting the study population and sample, designing the study, obtaining ethics approval and informed consent, collecting and analyzing data, and interpreting the findings [7].

Data on human behavior, activity, and environmental factors related to human well-being can be collected through various means such as surveys, interviews, observations, and sensing technologies [4]. There are different applicable sensing methods such as wearables, camera- or radar-based technology, virtual sensors (referring to a software program that simulates the functionality of a physical sensor), and ambient environmental sensors [8,9]. Examples of recent real-life studies on sensor-based monitoring varying from 1 to 10 months in duration include mood, stress, and health prediction [10,11]; daily stress detection [12,13]; and thermal comfort and behavior assessment [14]. However, several factors such as a low recruitment or high dropout rate, technical difficulties, data analysis and interpretation, and ethical reflections often challenge these types of field trials.

Although study participants are initially interested in contributing to the study by providing information or using sensor devices, their willingness can wane over time. Ecological momentary assessments (EMAs; self-reported labels), where participants report their feelings, experiences, and behaviors, are a common way of collecting annotations, that is, ground truth labels [15]. Regular and long-lasting EMAs have been reported to suffer from respondent fatigue [16,17]. Some studies have indicated that EMA scheduling and duration may affect perceived participant burden and dropouts [18,19]. On the other hand, a recent meta-analysis [20] suggests that the number of self-assessments does not appear to correlate with compliance or dropout rates; however, financial incentives have been shown to increase the compliance rates across studies, depending on compensation criteria (eg, providing compensation when at least 70% of self-assessments have been completed).

Furthermore, getting participants to wear or otherwise use the sensor devices consistently can be challenging, which can lead to incomplete data [21]. In addition, participants may feel uncomfortable being constantly monitored, and there may be legal or ethical issues in collecting and storing personal data [17,22]. Assessing and mitigating participant and stakeholder influence on field trial success is important. A recent systematic scoping review [23] identified 21 reports discussing stakeholder assessment approaches for planning stages, but approaches for field trials remained ill-established.

The typical technical challenges of wireless sensor deployment have been related to energy consumption (battery life), communication issues, and protecting devices, which may be especially relevant in outdoor settings [9]. However, these challenges also apply to personal devices such as wearables and

XSL•FO
RenderX

smartphones. For example, smartphones have become popular for gathering data in field trials, but continuous data collection may quickly drain the battery [24]. Moreover, collecting, managing, and storing large amounts of data can be costly and time-consuming. The data collection methods must be flexible enough to accommodate changes or unexpected events that may occur during the study [25].

Furthermore, real-world studies are susceptible to various sources of bias such as selection bias and measurement bias, which can affect the validity of the study. For instance, machine learning models require ground truth labels against which their accuracy can be compared. EMAs are subject to cognitive bias, as participants may not always be aware of their state, and to social desirability bias, where participants present themselves in a certain way they perceive favorably [26]. External observers can also label behavior, but this can be costly and time-consuming, and the observer may not always interpret the state of the participant correctly.

Interpreting human behavior is difficult due to the unstructured nature of real-world data and the lack of a consistent theoretical background; hence, behavioral modeling is often based on simplified theoretical concepts [27]. Moreover, modeling human activity and behavior is particularly challenging because emotions, expressions, and movements are highly distinctive for every individual [28]. This means that general models may be inaccurate for individual users because they are not tailored to the characteristics of a specific individual and context [29]. While personalized models can be tailored to the characteristics of a particular individual and environment, they may not be well generalizable and thus less widely applicable.

Frameworks for Field Trials

To tackle challenges related to the capabilities of wireless sensors, Booth et al [25] presented a framework for orchestrating sensor data streams through various pathways in real-life studies. The framework consolidates multimodal sensor data in one backend location, facilitating automatic stream monitoring and participant feedback systems. The authors also discussed criteria and factors for sensor selection and testing, including study site preparations and data collection. Moreover, the authors emphasized the importance of conducting a sensor survey that considers data quality and data access logistics, costs, technical requirements, ethics, and participant engagement before testing and data collection.

The perception and interpretation of context are essential in all interactions between humans or a human and technology [30]. Groh and Picard [31] have developed a conceptual framework for categorizing the context in automated affect detection using 7 factors. An ambient sensory environment refers to sensed information regarding the immediate surroundings. Situational constraints are caused by the activity or the immediate environment [32]. Furthermore, factors related to people, such as social relationships and cultural orientation, are defined as sociocultural dimensions. A temporal dimension refers to the time and dynamics of an event, such as an emotional expression or a specific behavior. Personalization includes all individual characteristics such as demographics, personality traits, abilities, or disabilities [33]. Methods of measurement refer to annotating emotions and human behavior. Furthermore, measuring human emotions and behavior also requires a semantic representation referring to a way of representing the meaning of a state. Overall, the framework developed by Groh and Picard [31] highlights the importance of considering all these contextual factors when detecting emotions and behavior to achieve more accurate results.

Methods

Research Process

To answer the research questions, we adopted a case study approach and looked for suitable types of field trials from our previous research using the following criteria: (1) carried out in real-life context, (2) investigated human behavior or activity, (3) were longitudinal with a continuous duration of ≥1 month for the data collection, and (4) used sensors for collecting primary data. After reviewing the field trials, 3 field studies met the criteria and were selected for further analysis—stress and environmental assessment in an office and a school, work activity recognition on a construction site, and stress recognition in location-independent knowledge work. Although these field studies were conducted independently, collectively they provide valuable insights into the challenges and solutions in longitudinal field trials by identifying similarities and differences across diverse contexts.

In the first phase, we developed a qualitative coding framework to systematically code and categorize the challenges faced in each field study. This novel framework was designed to include multiple dimensions of field trials such as influential factor identification, stakeholder management, data harvesting and management, and analysis and interpretation. By integrating and expanding on the existing frameworks, our approach offers a holistic tool that researchers can use to navigate the complexities of longitudinal sensor-based field trials. We used inductive reasoning to analyze written documentation of the selected field studies, including data management plans, technical reports, and internal memoranda from weekly meetings stored in project archives. The primary objective was to identify unexpected difficulties faced during the studies, mitigation actions to address the challenges, and lessons for future studies. Inductive reasoning allowed us to identify prominent, frequent, or substantial themes from the data without constraints from structured methods [34].

In the second phase, we synthesized the findings to develop a comprehensive understanding of the problems encountered and lessons learned. The synthesis enabled us to compare the types and extent of challenges across the field studies, providing valuable insights into their similarities and differences. In the third phase, we adopted methods proposed by Ballejos and Montagna [35] to identify and evaluate the influence of different stakeholders on the success of the field study. Analysis was done by researchers directly involved in the field studies, using the coding framework to ensure comparability. The findings were triangulated by having the researchers first do the analysis separately and then jointly review the findings to ensure consistency in interpretation. Finally, the most important

XSL•FO
RenderX

findings were summarized as recommendations provided in the discussion section of this paper.

Coding Framework

We used 2 established frameworks that categorize themes noted in challenges in field trials for human and environmental monitoring as a starting point for the coding framework: one by Booth et al [25] that pertains to sensor selection, deployment, and management and the other by Groh and Picard [31] that focuses on affect recognition. We also evaluated the challenges and recommendations reported in other studies (eg, L'Hommedieu et al [17]) to further identify the key categories that must be considered when planning and conducting these trials. On the basis of the existing frameworks, reported challenges, and our own experiences, we identified 4 main categories to consider when designing and conducting field trials: influential factor identification, stakeholder management, data harvesting and management, and analysis and interpretation.

Influential factor identification is essential to understand the background factors that may impact the success of a trial. For this category, we identified two key themes that should be considered during the planning and execution phases: (1) phenomenon and (2) context and duration. Phenomenon refers to the specific human behavior or activity being monitored and the characteristics of this phenomenon that are relevant to the trial [31]. Context refers to the environmental, sociocultural, and situational factors that can impact the success of a trial, such as physical settings, organizational norms and resources, and social interaction [31,32]. Duration refers to the measurement period and its potential impact on the trial.

Stakeholder management is critical for the success of longitudinal field trials involving human and environmental monitoring [17]. This category covers (1) participant engagement, (2) partner management, and (3) stakeholder engagement. Participant engagement involves recruiting and engaging individuals or groups whose data are being collected during the trial, including obtaining informed consent for data collection and use [17]. Partner management refers to collaborating with other organizations or individuals involved in the trial, such as academic institutions or research groups. Finally, other stakeholder engagement refers to the engagement of stakeholders who are not participants or partners but have a vested interest in the trial's execution or outcome, such as the management of the trial organization, local communities, government agencies, and other organizations the trial may impact.

Data harvesting and management is critical to any field trial using sensor technologies. This category builds mainly upon the work of Booth et al [25] and covers three essential themes: (1) measurements and annotations, (2) data transmission and connectivity, and (3) data storage and management. Measurements and annotations involve using sensor technologies to gather data during the trial and annotations or additional information that may be required to understand the data [31]. Data transmission and connectivity is needed to connect data sources and transfer data from the trial location to a central repository, while also considering data security and privacy [36]. Data storage and management includes data backup, access and control, and archiving [37].

Analysis and interpretation includes (1) data validation and (2) model development and validation. Data validation evaluates data quality, including statistical and computational bias [38]. Model development relates to the selection and use of machine learning and statistical methods for feature extraction and algorithm development [31]. Model validation includes approaches for explainability and techniques for model validation, including the identification of influencing factors for analysis.

Ethical Considerations

In all cases, the study plan included considerations regarding accountability, fairness, and responsible communication to promote freedom of choice and avoid potential harm when conducting field studies using novel sensing technology. The need for ethical permissions was evaluated by the Ethics Committee of VTT Technical Research Centre of Finland Ltd (henceforth VTT) against the guidelines from the Finnish National Board on Research Integrity [39]. In case A, case B, and the second pilot of case C, specific ethical permission was not required. However, permission was needed for the first pilot of case C, and the Ethics Committee of Hospital District of Helsinki and Uusimaa approved it (HUS/2536/220).

In all field studies, informed consent was obtained according to the European research integrity and General Data Protection Regulation (GDPR) guidelines, including for research objectives, study process and duration, benefits and potential disadvantages of the research, bodily integrity of the participant, voluntariness and withdrawal from the study, publishing of the research results, data collection, processing and archiving, data subject's rights, and contact information. The consent states that the data collected will be used for research that extends beyond the current projects, allowing secondary analysis without requiring a new consent.

The studies followed the GDPR in setting up data protection measures. Only data relevant to the research questions were collected. Data were pseudonymized during collection and anonymized at the time of archiving. Data are locally stored, protected by firewall, and accessible only by the researchers involved. We also adhered to other GDPR requirements such as data remaining within the boundaries of the European Union.

In the first field study, no compensation was provided to the human participants. A set of workwear was provided in the second field study. In the first pilot in the third case, personal analysis of the physiological data was provided after the pilot, and a gamification approach was used in the second pilot where participants were offered small snacks as rewards based on their reporting activity. The information contained in this paper does not contain any identifiers of participants in the 3 field studies.

Empirical Data

This section presents the 3 field studies selected for further analysis. Certain regulatory restrictions and ethical requirements were common to all field studies deploying sensors and cloud platform–based measurements in real-life settings.

XSL•FO
RenderX

Case A: Stress and Indoor Environmental Quality Assessment Using Environmental Sensor Data

Work stress is a complex phenomenon influenced by various psychosocial and physical factors, including poor environmental conditions. Exposure to environmental stressors, such as elevated carbon dioxide levels, has been shown to impair cognitive performance [40] and cause discomfort [41]. Thus, one of our field studies focused on stress detection, especially related to environmentally induced stress, and how to mitigate risk factors to improve well-being. We conducted 2 long-term indoor environmental quality (IEQ) data collections at school and office environments to study the applicability of IEQ sensor data for continuous stress and IEQ assessment of employees during the fall of 2018 and spring of 2019 [42].

The field study lasted for 3.5 months at the school and 12 months at the office facility. We recruited participants from an elementary school and a knowledge work organization in Northern Finland. The inclusion criteria for research participants were as follows: (1) uses an Android phone for self-reporting, (2) commits to wearing a wearable device during the workday, and (3) commits to answering self-reports daily. There were 27 voluntary and eligible research participants: 4 teachers and 23 office workers. The mean age for teachers was 43.5 (SD 15.6) years, and for the office workers, it was 41.6 (SD 8.0) years. All teachers (4/4, 100%) and 48% (11/23) of the office workers were women. All participants gave written informed consent.

While planning the monitoring of teachers, knowledge workers, and their environment, we engaged with several stakeholders, including the facility owners, elementary school principal, research management, and human resource (HR) departments. These stakeholders agreed upon the installation and use of a continuous monitoring system. In the elementary school setting, proper consideration to what could be measured in the presence of susceptible individuals, that is, children, was essential. Guardians of the schoolchildren were also informed about the measurements, but no data were collected directly from the schoolchildren.

Regarding data harvesting, we collected diverse data streams with varying levels of granularity to produce representative data for modeling, including annotations. The teachers reported their perceptions of IEQ, stress, and productivity on weekdays for 3.5 months and the office workers reported the same for 3.5 to 7 months via an Android self-reporting app developed by us. In addition, 2 time-triggered quantitative questionnaires on perceived IEQ, stress level, and productivity were scheduled daily, in the mornings and afternoons. Moreover, the research participants wore a wrist device to monitor physiological data (continuous activity and heart rate) during working hours.

The sensor data were gathered with a set of customized sensor devices integrated in a TinyNode hardware platform developed by VTT. The hardware platform included commercially available sensors to measure temperature (in °C), relative humidity (calculated in %), air pressure (in hPa), carbon dioxide (in ppm), and activity level based on passive infrared sensors (sum of detected motion events; 0-12). The passive infrared sensor data reflected movement for each seating place. The TinyNode operates on a cell battery and communicates wirelessly over Bluetooth Low Energy. Data samples were obtained once per minute for all parameters. The positioning of IEQ sensor devices followed the national legislative recommendations [43]. The gateway devices in school study rooms used a commercial router with 4G mobile network capabilities for internet access, while the gateway devices in the office rooms used the available wireless or wired internet connections. The gateways run on a Raspberry Pi single-board computer, and data were sent to the Microsoft Azure cloud platform via a message queuing telemetry transport protocol over transmission control protocol/IP. Data retrieval from Azure used a published application programming interface for Microsoft Azure TableStorage.

For IEQ, stress, and productivity analysis, we performed feature selection, classification model training, and evaluation separately for each research participant due to the dependence of IEQ, stress, and productivity interpretation on individual perceptions. Time windows of predetermined length were used to extract features with the Python library “tsfresh.” Feature selection involved calculating correlations between extracted feature values and respective self-reports. A support vector machine algorithm was used for classification.

Case B: Construction Site Safety Monitoring With Internet of Things Sensors

The labor-intensive construction industry has many occupational health and safety challenges such as accidents and nonergonomic work conditions [44]. Recently, there has been a shift toward using sensor-based and data-driven solutions to improve safety on construction sites [39]. In 2020, we conducted a field study to evaluate the potential of these solutions. Our study used a Bluetooth-based positioning system and inertial sensors to identify safe and potentially unsafe activities of workers in laboratory and field settings [45].

The case was implemented on a construction site in the capital region of Finland for 15 weeks. Site access was limited for safety reasons and to minimize disruptions to the work. Initially, 15 construction workers with different occupations such as painters; heating, ventilation, and air conditioning mechanics; and cleaners were recruited to participate in the field study. The inclusion criteria were that participants should be of working age and employed full time, but detailed demographic data were not collected because they were not purposeful for the study. All participants provided written informed consent. Two research consortium partners provided material for the field study: the workwear with special sensor pockets and the Internet of Things (IoT) sensor units. Other stakeholders included the subcontractor companies of the recruited workers because 11 out of 15 study participants were subcontracted workers on the construction site. In addition, the trade union for construction workers was recognized as an influencing stakeholder, and active discussions and a survey on new technologies for construction site safety were carried out with them.

For the activity data collection, the construction workers wore smart workwear with 3 special sensor pockets and 3 sensors for 4 days a week, from Tuesdays to Fridays. Mondays were reserved for sensor maintenance, including data download for storage, battery replacement, and checking sensor operational

XSL•FO
RenderX

status. Two of the sensors were IoT units primarily for motion detection, and one sensor was for positioning based on Bluetooth technology. The IoT units were 10-axis inertial measurement units composed of a 3-axis accelerometer, a 3-axis gyroscope, a 3-axis magnetometer, and a barometer, with fixed sampling rates of 103 Hz for the accelerometer, 97 Hz for the gyro and magnetometer, and 0.033 Hz for the barometer. IoT unit data were saved on the on-device flash memory. On Tuesday mornings, the construction workers fetched the sensors from a locker in the personnel room and placed them in the correct pockets and in the right orientation. They were expected to return the sensors to the locker every Friday for maintenance. The IoT sensor platform used in the study was a new version of a commercial product and had not gone through extensive testing. The positioning service was a commercial product, and the data were accessed via a web service.

The data collection system running on the construction site generated large amounts of unlabeled data with varying quality. The collected data were analyzed and interpreted using the activity recognition models created with laboratory-collected training data. Although the laboratory results could not be verified in the field study, they showed promising results that wearable IoT sensor data and machine learning could be applied to recognize different activities and provide valuable information about different work activities, working poses, and phases on construction sites.

Case C: Work Stress Detection Using a Computer, a Keyboard, and Mouse Dynamics

A quarter of the European Union employees have experienced frequent or constant stress at work [46]; thus, feasible and validated solutions are needed to detect early signs of prolonged work stress and support occupational well-being. Analyzing behavioral data is a way to detect stress as stress is manifested in human behavior [47]. Analyzing computer use, that is, digital behavior, is a potential approach, but previous studies with this approach have been conducted in laboratory conditions or in contexts where tasks are very uniform, such as student examinations [48]. We studied the applicability of computer use data for stress monitoring in knowledge work by conducting long-term field experiments in 2 organizations in 2021 and 2022.

The study was carried out in 2 pilots lasting 2 to 5 months. The first pilot had 38 participants from 2 knowledge work organizations, and the second pilot had 19 participants from 3 organizations. The mean age for the participants in the first pilot was 44.5 (SD 9.9) years, and 82% (31/38) were women. The mean age for second pilot participants was 38.4 (SD 11.6) years, and 53% (10/19) were women. The inclusion criteria for the participants were that they (1) were working regularly and (2) had a work contract covering the study period, and for the second pilot, (3) they had a fluent understanding of English. Recruitment was carried out locally in the organizations, but participants had a central point of contact during the study. The exclusion criterion was acute health conditions. All participants provided written informed consent.

Apart from the study participants, several other key stakeholders were involved. Another research organization was involved in the study, HR teams were engaged in recruiting participants, and deploying monitoring software in several separate organizations required negotiations with the respective IT or security teams.

Data were harvested using a virtual sensor, that is, a monitoring software installed on the work computers. The virtual sensor generated logs of mouse trajectories, keyboard typing tempo, and application switching and sent the data daily to the cloud backend. Actual keystrokes were not logged, except for Del and Backspace, for content security reasons. Contents of the screen were not captured. Applications were grouped into 5 broad categories, namely communication (eg, Microsoft Teams and Skype); web (all browsers); documents (eg, Microsoft Word and Microsoft PowerPoint); other (everything not falling into the above groups, such as software development tools); and unknown (rare cases when the operating system did not or could not return application information, such as system-level tasks). We did not gather data about application content, such as websites, or other computer use data, such as video or voice. The original data collection plan also included motion and IEQ sensors.

The data labels were collected via a self-report application provided for Windows, Android, and iOS platforms, which sent the daily reports to the cloud backend. Participants reported their perceived stress levels (stress, calm, and positive stress); work content (simple, need to do, interesting, and challenging); skills needed (little, average, lots, and need learning); productivity (on a 3-level scale); and optional work-related reasons for experienced stress from Monday to Friday. On Fridays, participants also answered an additional question on the overall stress level of the work week. The backend was located on the Microsoft Azure platform, accessible through a transmission control protocol/IP connection. Authentication for both virtual sensor and self-reporting data was done with the OAuth2 method. No actual names, IP addresses, or other identifiers were recorded in the transfer process. Data retrieval from Azure uses a published MySQL application programming interface for Microsoft Azure TableStorage. The backend provides role-based access control to data, with data encrypted when at rest, to minimize risks if data security was breached.

The analysis phase was ongoing at the time of writing this paper. The analysis started during the data harvesting phase, using data from the first participants, with more data added as each next batch of participants finished their data collection. Various feature extraction, feature selection, and classification methods are being tested using the research organization's internal computing servers.

Results

Identified Challenges From Field Studies

The analysis results from applying the coding framework themes to the selected field studies are presented in Tables 1-3. The tables present the challenges and actions carried out during the project to counter the challenge (ie, mitigations) or lessons learned during the project for future studies.

XSL•FO
RenderX

Table 1. Challenges identified in case A: stress and indoor environmental quality (IEQ) assessment using environmental sensor data.

Theme	Challenges and possible mitigation measures or lessons
Phenomenon	
Challenge	Showing the causality between IEQ, stress and productivity in natural settings with a small set of labels
Mitigation	Lowering the target level to find associations instead of causality was more feasible within the schedule and resources of the study
Context and duration	
Challenge	Labeling duration varied from 3.5 to 7 months and did not cover all seasons, yet, seasonal variation has a substantial impact on IEQ in the region
Lesson	Reflect if seasonal variation influences the phenomenon measured
Challenge	Relocations in the office due to unexpected facility renovation caused additional work and loss of participants and data
Lesson	Consult in advance with the stakeholders regarding potential changes in the study environment and schedule; explore alternative study locations
Participant engagement	
Challenge	Difficult to enroll participants; many potential participants were unwilling to commit to 4-6 months of measurements
Lesson	Use incentives (material or nonmaterial, such as personalized analysis reports) to engage participants and reduce dropouts
Challenge	A dropout rate of 30% due to loss of interest or relocation at the office to a location not allowing sensors
Lesson	Prepare for potential dropouts when planning
Challenge	Some participants expected that the data would validate their perceived symptoms, such as fatigue, and reveal the stress factors, leading to unrealistic expectations of the results
Mitigation	An additional information session was organized
Challenge	The number of self-reports decreased because of personal restrictions such as hurry, holidays, and business trips
Lesson	Offer incentives to increase compliance rates and simplify self-reporting as much as possible
Partner management	No challenges were identified from the data for this category
Other stakeholder engagement	
Challenge	Some other stakeholders had unrealistic expectations for the study results; for instance, facility owners wanted the data to confirm that the facility did not have air quality problems
Mitigation	An additional information session was organized
Measurements and annotations	
Challenge	Android app alerts failed for some participants, possibly due to variations in operation system versions
Lessons	Recognize common mobile app issues; providing a web application and using calendar reminders can help; monitor data accumulation to detect problems
Challenge	The wearable device to measure heart rate was not worn regularly or tightened enough, leading to inconsistent use and loss of data
Lesson	Monitor data quality and provide instructions; offer incentives for compliance
Data transmission and connectivity	
Challenge	Data loss (total of 24 hours) due to interruptions in internet connectivity and problems reconnecting to the Azure database
Mitigation	A fault recovery system for database connection solved this issue
Challenge	Electromagnetic interference caused a loss of 0.5%-1.5% of Bluetooth packets per day in most offices
Lesson	Estimate possible data loss when defining the target for the amount of data
Challenge	Long-term jitter in the timer of the sensor firmware and the lack of intersystem synchronization causing missing samples
Lesson	Perform a sufficiently long field test to detect possible data loss
Data storage and management	
Challenge	Relying on external commercial cloud platforms for backend operation for 3.5 years resulted in relatively high final costs
Lesson	Prepare for data storage expenses in the budget
Data validation	

XSL•FO
RenderX

Theme	Challenges and possible mitigation measures or lessons
Challenge	The physiological data were unusable due to quality problems (motion artifacts), although the participants were carefully instructed on how to wear the wearable device
Lesson	Test beforehand, monitor data quality, and provide instructions
Challenge	Out of the 23 participants, data from only 15 (65%) were eligible for analysis due to lost sensor data and incomplete self-reports
Lesson	Prepare for data quality issues by recruiting more participants and monitoring constantly
Model development and validation	
Challenge	The participants reported daily work stress, productivity, and IEQ perceptions on a 5-level scale, but the limited number of reports made it impossible to classify the responses into 5 categories
Mitigation	With an average of around 75 self-reports per participant, binary classification models were possible to train
Challenge	As the data collection activities required more resources than expected, testing of multiple model development approaches was not possible, potentially leading to suboptimal results; models were tailored closely to participants, possibly reducing general applicability
Lesson	Allocate resources for analyses

XSL•FO
RenderX

Table 2. Challenges identified in case B: construction site safety monitoring with Internet of Things (IoT) sensors.

Theme	Challenges and possible mitigation measures or lessons
Phenomenon	
Challenge	Identifying and modeling motion features leading to accidents and safety issues, which are short term and rare
Mitigation	Changing the modeling target toward nonergonomic poses and accident-prone working activities (eg, climbing a ladder) rather than motions
Context and duration	
Challenge	Participants left the study because they did not work on the same construction site for 15 weeks
Lesson	Consult beforehand with stakeholders regarding the ways of working and schedule
Challenge	The construction site was a constantly changing environment posing challenges to maintaining the positioning sensor infrastructure; the Bluetooth anchors were removed without informing the researchers when the anchors prevented the work (eg, painting the wall where the anchors were attached)
Lesson	Consult beforehand with stakeholders regarding potential changes in the environment
Challenge	Safety and security policy prohibited researchers' free access to the construction site; researchers felt that their visits and communication were a nuisance for everyday work on-site
Mitigation	Minimizing the number of visits and planning carefully to maximize their value
Participant engagement	
Challenge	Recruitment went through the site manager (no direct contact with participants) without visibility to the subcontractor chain; researchers met the participants only twice, once in the selection of needed workwear and then in the study kick-off delivering the workwear and explaining the study protocol; this took place in a coffee room where it was difficult to understand who the actual participants were
Lesson	Organize a dedicated session only with participants
Challenge	Many dropouts due to workers leaving the site and the study protocol being too tedious; researchers were not informed about the dates when the participants left the site, but data showed the first 5 weeks to be active in data collection.; after 6 weeks, there were only 7 participants left
Lesson	Prepare for dropouts by recruiting more participants; offer incentives
Challenge	Participants forgot to fetch or return the sensor, causing problems with data collection
Mitigation	Clear instructions were given to participants, also attached by the locker where sensors were stored, and the site manager was asked to remind the study participants; explicit study material provided for the participants' employers (including subcontractors)
Partner management	
Challenge	Workwear provider wanted to promote their offering, causing issues in clothing selection and delivery; the technology setup was under development, and the number of sensors integrated to the workwear was not fully clear to the provider in the planning phase
Mitigation	Researchers actively discussed, informed, and documented the study planning, and the material was available on the research project's document platform
Other stakeholder engagement	No challenges were identified from the data for this category
Measurements and annotations	
Challenge	To minimize disruptions to work, the workers could not be asked to annotate their daily work; over a 15-week period, observation was limited to following 3 workers on two separate occasions for 2 hours each; this severely limited the amount of labeled data and challenged the subsequent model development
Lesson	Understand and prepare for the limitations due to the environment
Challenge	Due to study requirements, a higher data sampling rate for position data was used than the commercial system was designed for, leading to unexpected technical problems and unusable data
Lesson	Perform a sufficiently long field test to detect issues
Challenge	In some cases, especially when the participants forgot to return the sensors, the IoT sensor units ran out of battery, leading to data loss
Mitigation	The site manager was asked to remind participants
Data transmission and connectivity	

XSL•FO
RenderX

Theme	Challenges and possible mitigation measures or lessons
Challenge	The IoT sensor units collected data to an internal memory card, which failed if the battery ran out, corrupting the memory card and causing the loss of all data; also, if the sensor memory was not erased, it caused problems for the following week's data collection
Mitigation	A firmware update was created and applied to the IoT units
Challenge	Data transfer to storage was a manual and complicated act, where even a slightest mistake could lead to data being erased and lost
Lesson	Prepare for problems with experimental setups. Avoid using experimental sensors in data collection for artificial intelligence development
Data storage and management	No challenges were identified from the data for this category
Data validation	
Challenge	Location data could not be used due to quality issues (see the Measurement and annotations category in this table); data gaps also existed in the IoT data (see the Data transmission and connectivity and Measurements and annotations categories in this table); a drift in sensor timestamping was noticed during the study period, which also caused validation problems
Mitigation	To minimize drifting, postprocessing of timestamps and alignment of signals was performed
Model development and validation	
Challenge	Validation of data analysis results was difficult due to the minimal number of annotated data samples
Mitigation	A separate data collection exercise in the laboratory mimicking the real world was performed to support model development; however, using models based on laboratory data was challenging due to the diversity of work activities performed in real sites; another data annotation (observation) period was organized at the end of the field study

XSL•FO
RenderX

Table 3. Challenges identified in case C: work stress detection using computer, keyboard, and mouse dynamics.

Theme	Challenges and possible mitigation measures or lessons
Phenomenon	
Challenge	Stress is a physiological but also partially subjective phenomenon. There is no consistent and accurate method to establish ground truth for stress
Mitigation	Research used self-reports as the ground truth
Context and duration	
Challenge	The COVID-19 pandemic prevented studies according to initial plans. Research on environmental stressors had to be dropped
Mitigation	A context-independent virtual sensor to measure computer use behavior
Challenge	Study organization had security-critical projects, and IT saw the use of the virtual sensor (keylogger technology) as too risky
Mitigation	An external security audit of the virtual sensor
Challenge	A loss of technical know-how due to a researcher leaving the project led to a mistake, risking a loss of a substantial amount of data; significant manual effort was required to recover the data
Lesson	At least 2 persons should be familiar with the technical setup and the setup should be carefully documented
Participant engagement	
Challenge	Recruitment was significantly less successful than anticipated; enough participants were eventually recruited, but required extra effort and complicated the study setup
Mitigation	Continued recruitment and extending the recruitment to other organizations; the field study ran the pilot in “waves” until the recruitment target was met
Challenge	Commonplace practical challenges, such as people not reading the instructions, or losing or forgetting their passwords, causing additional workload to the research team
Mitigation	Increased communication and support effort, especially at the early stages of the measurement for the second pilot study
Partner management	
Challenge	Cross-organizational study coordination and achieving mutual understanding for fluent cooperation in a cross-disciplinary (behavioral science, data science, and information and communication technology professionals) group was challenging and required more communication effort than anticipated
Mitigation	Additional communication and joint weekly meetings
Other stakeholder engagement	
Challenge	Research organization’s DPO[a] had GDPR[b] compliance concerns for voluntariness in recruitment and data sensitivity; negotiations with DPO delayed the project start, and agreed measures increased study management complexity and caused extra costs
Mitigation	Arrangements to ensure participation anonymity, additional data security measures to the backend
Challenge	The field study organization’s IT and security team considered the virtual sensor a security risk; providing documentation instead of having direct discussions was a mistake, resulting in an initial prohibition and an additional 5 months of negotiations, leading to a delay and unanticipated costs
Lesson	Actively engage with IT and security teams to get approval already when planning to use tools that may cross their domain
Challenge	Candidate study organizations were asked to test technical solutions very late in the study planning cycle; one candidate organization withdrew from the study due to technical problems that were discovered too close to the start
Lesson	Perform operative testing early enough to identify technical problems
Measurements and annotations	
Challenge	Physiological sensors were considered unreliable ground truth due to motion artifacts; self-reports of stress were feasible but subjective and prone to lapses
Mitigation	Self-reports to be treated as ground truth; gamification used to entice people to keep reporting
Data transmission and connectivity	
Challenge	Automated data harvesting application authentication failed at times, requiring repeated reauthentication from users, causing frustration among participants and added workload for the technical staff due to password reset requests
Lesson	Develop a streamlined method for reauthentication (such as browsers’ password storage feature) when using digital reporting tools

XSL•FO
RenderX

Theme	Challenges and possible mitigation measures or lessons
Challenge	Data transmission had to be tailored for each study organization's information system network, causing deployment delays
Lesson	Allocate budget and time for technical implementation
Data storage and management	
Challenge	Locally developed solutions required active maintenance and specialized knowledge, increasing technical workload and budget use beyond estimated (25% of the effort budget)
Lesson	Use commercial options when available or allocate substantial budget and time for technical implementation
Challenge	Relying on external, commercial cloud platforms for backend operation carries relatively high final costs
Mitigation	Sufficient budget was allocated for cloud expenses
Data validation	
Challenge	Features from prior art fail to correlate well with ground truth labels, possibly due to it being derived from short (few hours maximum) measurements typically done in a laboratory context
Mitigation	Extending the project duration to allow more experimentation
Challenge	Stress behavior can vary significantly from one person to another; features that work for one person do not necessarily work for another (no universal feature); limited number of labels does not allow reliable feature selection based on labeling
Mitigation	Extending the project duration to allow experiments with different features
Model development and validation	No challenges to report at the time of writing this paper

[a]DPO: data protection officer.

[b]GDPR: General Data Protection Regulation.

Summary of the Case Findings

The case-specific findings demonstrated various challenges relevant to all viewpoints in the established coding structure, some of which were mitigated during the project, while others were taken as lessons for future projects. Some phenomena are difficult to detect in real-world settings, and longitudinal studies are likely to face changes in the environment and conditions in the field. Consulting stakeholders to understand what changes might happen is critical, but some things cannot be fully anticipated, and the ability to adjust is needed. Problematic changes can also occur in the research team itself, and the more dependent the field study is on a person's expertise, the more critical it is to document carefully and have another person who can take over.

Stakeholders, including study participants and collaborators, constitute another major issue area. Achieving a common understanding can be challenging; there may be needs, interests, and expectations that the study is not intended or is unable to meet or may even conflict with, and the interest may wane over time. A particularly challenging case is dealing with stakeholders with influence but without direct interface with the research team. Communication is the key in this respect—frequent and clear, with an initiative to discuss issues as early as possible. Another essential tool is to find incentives that help to maintain the support for the study.

While pervasive technology is a significant enabler for conducting field studies on human behavior, it also presents many problems. The reliability of the technology in the field is rarely 100%, and environment or even personnel changes can cause additional issues. The outcome is often data loss, which in turn causes problems in data analysis and interpretation. Technology trials in realistic settings before deployment, designs that include fault recovery systems, and monitoring data accumulation and quality are important means that help. However, some data loss or data quality issues are likely unavoidable in field trials; therefore, extending the study duration beyond what would be needed if everything goes perfectly is necessary to ensure enough data for analysis.

Stakeholder Mapping

In the analysis phase, the importance of stakeholders and their management was raised as a significant issue as a key enabler for the study's success and a frequent source of challenges. Therefore, we analyzed this aspect further, as there was evident variation in the related challenges and roles, as well as similarities among the stakeholders. This resulted in identifying the main stakeholder groups and allowed mapping out their relevance and importance in the field studies, as presented in Figure 1. Most of these groups were present in all 3 field studies. The key difference in importance in our mapping was their impact on the study success, with those that can stop or enable the study, and even influence the data quality, thus having a major impact on study success being in the center, and those that can hinder or support the study being in the outer area.

XSL•FO
RenderX

Figure 1. Stakeholder mapping in field studies: (A) stress and indoor environmental quality assessment using environmental sensor data; (B) construction site safety monitoring with Internet of Things sensors; and (C) work stress detection using a computer, keyboard, and mouse dynamics.

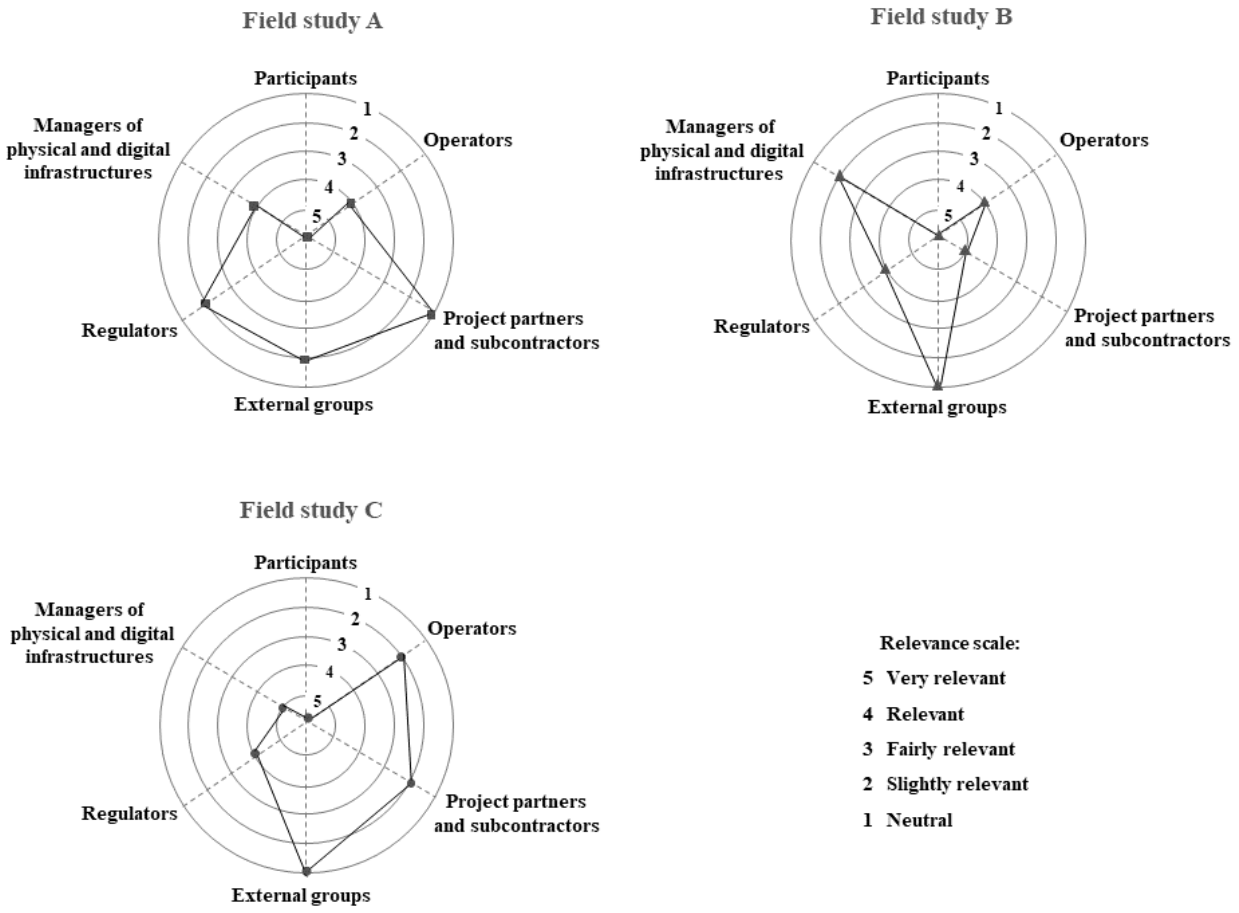

The primary stakeholder group includes the *participants of the study*, as the essential data are harvested from them. The next key group is the *operators* of the study environment or their representatives, such as participants' employers or their HR departments. Next are the *project partners and their potential subcontractors*, including service providers that the project uses to assist in a particular aspect of the study, for instance, positioning service or cloud platform facilities. The third and fourth groups are the *managers of the physical and digital infrastructures* in the study context, for instance, office facility management and IT and security departments. The fifth important group is the *regulators*, as requirements, for instance, those governing safety issues at construction site or dictating how IEQ measurement needs to be done, or through roles, such as the data protection officer for GDPR issues. The sixth and final group is the *external groups* with a vested interest, especially toward the research participants, for instance, parents of the children at school or trade unions, who can influence the study either by supporting or hindering it.

Discussion

Summary of Findings of the Specific Field Studies

Overview

Transitioning research from the laboratory to the real-world setting brings significant challenges, and the success of a study depends on overcoming them. This study aimed to provide new information on what challenges researchers will likely face when conducting sensor-based field trials over a continuous longitudinal duration and how those challenges could be managed. To answer our research questions regarding the unpredictable challenges faced in sensor-based field trials studying human behavior or activity and possible solutions, we analyzed and synthesized the challenges and lessons of 3 longitudinal field studies.

Each independent field study highlighted specific lessons that can inform future research and practical applications in similar settings.

Case A: Stress and IEQ Assessment

Case findings revealed that aiming for associations rather than causality might be more feasible within the constraints of such studies. Seasonal variations impact IEQ and stress, underscoring the need to account for these factors in longitudinal field research. Effective communication with stakeholders about potential changes and realistic outcomes is essential. Offering incentives can improve participant compliance and reduce dropouts. Monitoring data quality and addressing technical issues promptly can mitigate data loss.

XSL•FO
RenderX

Case B: Construction Site Safety Monitoring With IoT Sensors

In the dynamic environment of a construction site, modeling should focus on broader activity categories such as nonergonomic poses and accident-prone activities. Direct participant interaction and incentives can improve engagement and data quality. The operator on-site needs to commit to actively support the study implementation. Pretesting experimental setups thoroughly is necessary to minimize technical issues during the study. Robust methods for data annotation should supplement limited observation opportunities.

Case C: Work Stress Detection Using Computer Keyboard and Mouse Dynamics

Given the subjective nature of stress, self-reports were used as ground truth data. The study demonstrated the importance of flexibility in adapting to unforeseen events such as the COVID-19 pandemic. Early engagement with IT and security teams is crucial to address concerns and ensure compliance. Maintaining thorough documentation and cross-training team members can mitigate risks associated with personnel changes. Continuous recruitment and extended study periods are necessary to meet participant targets and ensure sufficient data collection.

Principal Findings and Comparison With Prior Work

The findings of case studies demonstrated challenges in 4 categories: *influential factor identification*, *data harvesting and management*, *stakeholder management*, and *analysis and interpretation*.

Concerning influential factor identification, sensor-based monitoring of humans and their environments requires understanding the phenomenon under investigation [27], as typically established through previous research and applied to define the main study objective and methods. In many cases, the environment changes due to explicit human operation or implicit use of the space (eg, different modes in office work). The need to tailor the protocols and solutions for the specific context and build in flexibility to adapt to a changing environment is standard when conducting research in a natural setting and should be reflected in the schedule and budget.

In sensor data harvesting and management, we faced technical issues with wireless connectivity and data transmission despite our testing efforts. This highlights the importance of conducting operative testing early and for a sufficient duration to detect technical problems, as well as implementing an automatic fault recovery system. In general, deploying pervasive solutions increases the scale of data harvesting and management, which means that technology setups require IT and software expertise, as well as maintenance work. Furthermore, behavior data analysis becomes intractable without annotations. Although making daily self-assessments with reminders effortless for the participants, case 1 suffered respondent fatigue without incentives, which aligns with the findings of previous studies [20,38]. Even in the immaterial form of personal feedback, the rewards improved the compliance rate in case 3. In addition to annotations provided by participants, it is worthwhile exploring modern AI-based tools for labeling large amounts of data, such as video data.

Regarding data analysis and interpretation, data quality impacts the success of learning and adaptive applications. However, real-world sensor-based measurements often produce imperfect output data. This implies that while already advanced, sensing technology still needs to improve reliability and accuracy for field study purposes. Tools for data quality inspection (eg, the Agile Data Engine [49]) would be valuable in research-oriented field studies, but budget and resource constraints hinder the implementation of comprehensive data quality management tools. Furthermore, the influence of study participants on data amount and quality is inevitable because wearing measurement devices and providing annotations is usually required. When creating models to make decisions or predictions about human behavior based on limited datasets, it is essential to consider the generalizability and biases of the models, for instance, using existing frameworks and guidelines [37].

Our study confirmed the importance of stakeholder engagement, especially for field trials of extensive duration [16,17,25], and demonstrated the utility of the coding framework in systematically addressing this and other critical challenges. The framework not only structures the identification and management of challenges but also facilitates the comparison of results across different studies, thereby contributing to the development of more cohesive knowledge in this field.

Compared with earlier studies, our study revealed 2 new stakeholder perspectives. First, the challenge in engaging with diverse stakeholder groups with varying expectations and concerns. Second, the presence of stakeholders who are not in direct contact with the research team and who have no vested interest in the study yet whose work, responsibilities or sphere of operation is somehow affected or involved. The former requires careful handling, especially if groups have conflicting or opposing desires regarding the study outcomes. The latter requires additional effort to identify, as they may not be considered relevant by the study organization, even though they can impact the study on a practical level.

During the planning phase, it is essential to identify the actors and assess their importance and needs. One possible approach is to conduct a survey to determine the expectations and worries of each group. Stakeholder mapping is a valuable tool for assessing the significance and relevance of each group, thus enabling a better understanding of where to invest resources. It allows for a clear identification of stakeholders who have influential positions and can directly impact the implementation of the study. On the other hand, stakeholders positioned toward the outer edges of the circle can either provide support or pose potential obstacles to the research, but they do not stop it entirely. After a successful stakeholder assessment phase, active stakeholder engagement is needed.

The stakeholders need to be informed about how the research meets their expectations and manages their responsibilities in the process. The critical stakeholders must especially commit to collaborating with the researchers. However, it can be challenging to satisfy all expectations, given the experimental nature of field research. Moreover, identifying all stakeholder

XSL•FO
RenderX

expectations and concerns can be tricky because their knowledge may gradually develop over time. Thus, the study organizers can ensure that the outcomes benefit all parties by implementing communication strategies that consider the distinctive needs of stakeholder groups, keep them informed, and ensure their active involvement throughout the process. The communication should also include sharing the study results with the interested stakeholders and collecting feedback to improve the stakeholder engagement process in future studies.

In conclusion, based on the mitigations and lessons from the field studies and focusing on the issues relevant to the actual research rather than project management, we provide recommendations (Textbox 1) to be considered in planning and managing a real-world study setup when collecting human behavioral data, primarily through sensors but also including self-reported material.

Textbox 1. Summary of the lessons learned.

Study setting and plan

1. Prepare for changes in the conditions and environment with flexibility in the schedule and budget.
2. Define and focus on the main objective and ensure your study design choices support it.
3. Select measurement methods and approaches to balance between accuracy and reliability versus field study feasibility.
4. Plan for a sufficiently long field test to detect issues and phenomena.
5. Check as early as possible if you can capture the phenomenon with the data from the field study (fail fast).
6. Make user-provided annotations (ecological momentary assessments) with reminders effortless for the participants.
7. Plan for dropouts by increasing the number of participants and offering incentives (financial or immaterial rewards) to improve compliance rate.

System and data engineering aspects

8. Assure resources for maintaining the technical setup.
9. Conduct sufficiently long field tests to identify technical problems in advance.
10. Apply reliable tools for regular data inspection and quality assurance.
11. Understand the generalizability limitations and potential biases when creating human behavior models based on data collected from real-life settings.

Stakeholder management

12. Identify all possible stakeholders and assess their importance and varying needs, for instance, with a stakeholder mapping tool.
13. Identify and prepare to manage external stakeholders not in direct contact with the research team, such as maintenance, security, management, or other administrative personnel of involved organizations.
14. Recognize the most impactful stakeholders and assure their commitment.
15. Inform the stakeholders in advance how the research meets their expectations and manages their concerns and responsibilities.
16. Tailor communication strategies to different stakeholder groups, keep them informed, and ensure their active involvement.
17. Include sharing the results after the study with the interested stakeholders and collecting feedback.

Limitations

While this study provided practical contributions to tackle challenges in longitudinal field trials monitoring humans and the environment using pervasive technology, it is essential to note some limitations. First, the purpose of the established coding framework was primarily to provide a practical structure for identifying and mitigating potential challenges that may arise when conducting long-term technology-oriented research in natural settings. As such, it can be considered a tool for field study risk management. Moreover, our stakeholder evaluation criteria were constrained to serve the scope of analysis. Thus, the framework and especially the stakeholder evaluation would benefit from further study. Second, our cross-analysis was qualitative, which is always vulnerable to subjective bias. Nonetheless, our framework and cross-study analysis provide valuable information to consider while designing and conducting a field trial for sensor-based human behavior or activity monitoring. Our future research will focus on enhancing and applying the coding framework more widely for planning longitudinal sensor-based field trials in various domains and identifying and mitigating possible challenges.

Conclusions

Field trials are a practical method for studying human behavior and the environment and evaluating various aspects of technology in uncontrolled situations. Nevertheless, conducting field studies can be challenging, as unforeseen difficulties can occur due to the complexity of human behavior, changing contexts, and technological factors. This study introduced a novel qualitative coding framework that provides structured insights into the challenges and lessons learned from 3 field studies monitoring stress, accident-prone work activities, and environmental factors in natural settings.

According to our results, managing participants and other stakeholders, especially those interacting via a proxy or with contradictory expectations and various concerns, can be

XSL•FO
RenderX

challenging. Our results imply that field trials should not only put additional effort to identify stakeholders and clarify their expectations and concerns already in the design phase but also communicate with stakeholders throughout the study process. Continuously monitoring and mitigating the possible challenges will ultimately lead to better success in longitudinal trials, improved data quality, and future technology-based solutions. Our analysis highlighted that field trials involving sensor-based human behavior assessment face unique challenges that require careful planning, robust technical solutions, and effective stakeholder engagement. By focusing on broader activity categories, maintaining flexibility, and ensuring continuous communication with stakeholders, future research can better navigate these challenges and enhance the success of longitudinal field trials.

Acknowledgments

The authors acknowledge the considerable contributions of the study participants. They are appreciative of the VTT Technical Research Centre of Finland Ltd, Business Finland and the European Union's Horizon 2020 program for funding the ESTABLISH, SCOTT, ConIoT, and Mad@Work projects and the project teams and collaborators for their willingness to impart their knowledge. This research was partially funded by the InSecTT project that has received funding from the ECSEL Joint Undertaking under grant agreement 876038. The joint undertaking receives support from the European Union's Horizon 2020 research and innovation program and Austria, Sweden, Spain, Italy, France, Portugal, Ireland, Finland, Slovenia, Poland, the Netherlands, and Turkey. This paper reflects only the authors' view, and the commission is not responsible for any use that may be made of the information it contains.

Data Availability

The data used in the study cannot be openly published because of clauses on project consortium agreements and informed consent.

Conflicts of Interest

None declared.

References

1. Davies N, Clinch S. Pervasive data science. IEEE Pervasive Comput 2017 Jul 27;16(3):50-58. [doi: 10.1109/mprv.2017.2940956]
2. Dzedzickis A, Kaklauskas A, Bucinskas V. Human emotion recognition: review of sensors and methods. Sensors (Basel) 2020 Jan 21;20(3):592 [FREE Full text] [doi: 10.3390/s20030592] [Medline: 31973140]
3. Minh Dang L, Min K, Wang H, Jalil Piran M, Hee Lee C, Moon H. Sensor-based and vision-based human activity recognition: a comprehensive survey. Pattern Recognition 2020 Dec;108:107561. [doi: 10.1016/j.patcog.2020.107561]
4. Rogers Y, Marshall P. Research in The Wild. Cham, Switzerland: Springer; 2017.
5. Kjeldskov J, Skov MB. Was it worth the hassle?: ten years of mobile HCI research discussions on lab and field evaluations. In: Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services. 2014 Presented at: MobileHCI '14; September 23-26, 2014; Toronto, ON p. 43-52 URL: https://dl.acm.org/doi/10.1145/2628363.2628398 [doi: 10.1145/2628363.2628398]
6. Brown B, Reeves S, Sherwood S. Into the wild: challenges and opportunities for field trial methods. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011 Presented at: CHI '11; May 7-12, 2011; Vancouver, BC p. 1657-1666 URL: https://dl.acm.org/doi/10.1145/1978942.1979185 [doi: 10.1145/1978942.1979185]
7. Smith PG, Morrow RH, Ross DA. Field Trials of Health Interventions: A Toolbox. 3rd edition. Oxford, UK: Oxford Academic Press; 2015.
8. Fernandes JM, Silva JS, Rodrigues A, Boavida F. A survey of approaches to unobtrusive sensing of humans. ACM Comput Surv 2022 Jan 18;55(2):1-28. [doi: 10.1145/3491208]
9. Parkinson T, Parkinson A, de Dear R. Continuous IEQ monitoring system: context and development. Build Environ 2019 Feb;149:15-25. [doi: 10.1016/j.buildenv.2018.12.010]
10. Taylor SA, Jaques N, Nosakhare E, Sano A, Picard R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Trans Affect Comput 2020;11(2):200-213 [FREE Full text] [doi: 10.1109/TAFFC.2017.2784832] [Medline: 32489521]
11. Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, et al. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. J Med Internet Res 2018 Jun 08;20(6):e210 [FREE Full text] [doi: 10.2196/jmir.9410] [Medline: 29884610]
12. Maxhuni A, Hernandez-Leal P, Sucar LE, Osmani V, Morales EF, Mayora O. Stress modelling and prediction in presence of scarce data. J Biomed Inform 2016 Oct;63:344-356 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.023] [Medline: 27592309]

XSL•FO
RenderX

13. Vildjiounaite E, Huotari V, Kallio J, Kyllönen V, Mäkelä SM, Gimel'farb G. Unobtrusive assessment of stress of office workers via analysis of their motion trajectories. Pervasive Mob Comput 2019 Aug;58:101028. [doi: 10.1016/j.pmcj.2019.05.009]
14. Langevin J, Gurian PL, Wen J. Tracking the human-building interaction: a longitudinal field study of occupant behavior in air-conditioned offices. J Environ Psychol 2015 Jun;42:94-115. [doi: 10.1016/j.jenvp.2015.01.007]
15. Schmidt P, Reiss A, Dürichen R, Van Laerhoven K. Wearable-based affect recognition-a review. Sensors (Basel) 2019 Sep 20;19(19):4079 [FREE Full text] [doi: 10.3390/s19194079] [Medline: 31547220]
16. Aranki D, Kurillo G, Yan P, Liebovitz DM, Bajcsy R. Real-time tele-monitoring of patients with chronic heart-failure using a smartphone: lessons learned. IEEE Trans Affective Comput 2016 Jul 1;7(3):206-219. [doi: 10.1109/TAFFC.2016.2554118]
17. L'Hommedieu M, L'Hommedieu J, Begay C, Schenone A, Dimitropoulou L, Margolin G, et al. Lessons learned: recommendations for implementing a longitudinal study using wearable and environmental sensors in a health care organization. JMIR Mhealth Uhealth 2019 Dec 10;7(12):e13305 [FREE Full text] [doi: 10.2196/13305] [Medline: 31821155]
18. Eisele G, Vachon H, Lafit G, Kuppens P, Houben M, Myin-Germeys I, et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment 2020 Sep 10:1073191120957102. [doi: 10.1177/1073191120957102] [Medline: 32909448]
19. van Berkel N, Shalawadi S, Evans MR, Visuri A, Hosio S. A longitudinal analysis of real-world self-report data. In: Proceedings of the 19th IFIP TC13 International Conference on Human-Computer Interaction. 2023 Presented at: INTERACT '23; August 28, 2023; York, UK p. 611-632 URL: https://link.springer.com/chapter/10.1007/978-3-031-42286-7_34 [doi: 10.1007/978-3-031-42286-7_34]
20. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr 11;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
21. Mishra V, Pope G, Lord S, Lewia S, Lowens B, Caine K, et al. Continuous detection of physiological stress with commodity hardware. ACM Trans Comput Healthc 2020 Apr 15;1(2):1-30 [FREE Full text] [doi: 10.1145/3361562] [Medline: 32832933]
22. Carneiro D, Novais P, Augusto JC, Payne N. New methods for stress assessment and monitoring at the workplace. IEEE Trans Affective Comput 2019 Apr 1;10(2):237-254 [FREE Full text] [doi: 10.1109/TAFFC.2017.2699633]
23. Franco-Trigo L, Fernandez-Llimos F, Martínez-Martínez F, Benrimoj SI, Sabater-Hernández D. Stakeholder analysis in health innovation planning processes: a systematic scoping review. Health Policy 2020 Oct;124(10):1083-1099. [doi: 10.1016/j.healthpol.2020.06.012] [Medline: 32829927]
24. Bin Qaim W, Ometov A, Molinaro A, Lener I, Campolo C, Lohan ES, et al. Towards energy efficiency in the internet of wearable things: a systematic review. IEEE Access 2020;8:175412-175435. [doi: 10.1109/access.2020.3025270]
25. Booth BM, Mundnich K, Feng T, Nadarajan A, Falk TH, Villatte JL, et al. Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: framework for sensor selection, deployment, and management. J Med Internet Res 2019 Aug 20;21(8):e12832 [FREE Full text] [doi: 10.2196/12832] [Medline: 31432781]
26. Iida M, Shrout PE, Laurenceau JP, Bolger N. Using diary methods in psychological research. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology, Vol. 1. Foundations, Planning, Measures, and Psychometrics. New York, NY: American Psychological Association; 2012:277-305.
27. Picard RW. Affective computing: challenges. Int J Hum Comput Stud 2003 Jul;59(1-2):55-64. [doi: 10.1016/S1071-5819(03)00052-1]
28. Koldijk S, Neerincx MA, Kraaij W. Detecting work stress in offices by combining unobtrusive sensors. IEEE Trans Affective Comput 2018 Apr 1;9(2):227-239. [doi: 10.1109/taffc.2016.2610975]
29. Smets E, De Raedt W, Van Hoof C. Into the wild: the challenges of physiological stress detection in laboratory and ambulatory settings. IEEE J Biomed Health Inform 2019 Mar;23(2):463-473. [doi: 10.1109/jbhi.2018.2883751]
30. Gross T, Specht M. Awareness in context-aware information systems. In: Oberquelle H, Oppermann R, Krause J, editors. Mensch & Computer. Cham, Switzerland: Springer; 2001:173-182.
31. Groh M, Picard R. Context in automated affect recognition. In: Proceedings of the 35th Conference on Neural Information Processing Systems. 2021 Presented at: NeurIPS '21; December 6-14, 2021; Sydney, Australia p. 1-6 URL: https://mattgroh.com/pdfs/context_automated_affect_recognition.pdf
32. Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol Sci Public Interest 2019 Jul 17;20(1):1-68 [FREE Full text] [doi: 10.1177/1529100619832930] [Medline: 31313636]
33. Greenaway KH, Kalokerinos EK, Williams LA. Context is everything (in emotion research). Soc Pers Psychol Compass 2018 May 17;12(6):e12393. [doi: 10.1111/spc3.12393]
34. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006 Jun 01;27(2):237-246. [doi: 10.1177/1098214005283748]
35. Ballejos LC, Montagna JM. Method for stakeholder identification in interorganizational environments. Requirements Eng 2008 Sep 9;13(4):281-297. [doi: 10.1007/S00766-008-0069-1]

XSL•FO
RenderX

36. Nweke HF, Teh YW, Mujtaba G, Al-garadi MA. Data fusion and multiple classifier systems for human activity detection and health monitoring: review and open research directions. Inf Fusion 2019 Mar;46:147-170. [doi: 10.1016/j.inffus.2018.06.002]
37. Yang F, Hua Y, Li X, Yang Z, Yu X, Fei T. A survey on multisource heterogeneous urban sensor access and data management technologies. Meas Sens 2022 Feb;19:100061. [doi: 10.1016/j.measen.2021.100061]
38. Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a standard for identifying and managing bias in artificial intelligence. National Institute of Standards and Technology. 2022. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf [accessed 2024-10-06]
39. The ethical principles of research with human participants and ethical review in the human sciences in Finland. Finnish National Board on Research Integrity. URL: https://tenk.fi/sites/default/files/2021-01/Ethical_review_in_human_sciences_2020.pdf [accessed 2024-10-06]
40. Wargocki P, Wyon DP. Ten questions concerning thermal and indoor air quality effects on the performance of office work and schoolwork. Build Environ 2017 Feb;112:359-366. [doi: 10.1016/j.buildenv.2016.11.020]
41. Al horr Y, Arif M, Katafygiotou M, Mazroei A, Kaushik A, Elsarrag E. Impact of indoor environmental quality on occupant well-being and comfort: a review of the literature. Int J Sustain Built Environ 2016 Jun;5(1):1-11. [doi: 10.1016/j.ijsbe.2016.03.006]
42. Kallio J, Vildjiounaite E, Koivusaari J, Räsänen P, Similä H, Kyllönen V, et al. Assessment of perceived indoor environmental quality, stress and productivity based on environmental sensor data and personality categorization. Build Environ 2020 May;175:106787. [doi: 10.1016/j.buildenv.2020.106787]
43. Decree of the Ministry of Social Affairs and Health on Health-related conditions of housing and other residential buildings and qualification requirements for third-party experts. Finnish Ministry of Social Affairs and Health. URL: https://www.finlex.fi/en/laki/kaannokset/2015/en20150545.pdf [accessed 2024-10-06]
44. Winge S, Albrechtsen E. Accident types and barrier failures in the construction industry. Saf Sci 2018 Jun;105:158-166. [doi: 10.1016/j.ssci.2018.02.006]
45. Mäkela SM, Lämsä A, Keränen JS, Liikka J, Ronkainen J, Peltola J, et al. Introducing VTT-ConIot: a realistic dataset for activity recognition of construction workers using IMU devices. Sustainability 2021 Dec 26;14(1):220. [doi: 10.3390/su14010220]
46. Introduction: the rise of the social enterprise: 2018 global human capital trends. Deloitte. 2018. URL: https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2018/introduction.html [accessed 2024-10-06]
47. Sharma S, Singh G, Sharma M. A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans. Comput Biol Med 2021 Jul;134:104450. [doi: 10.1016/j.compbiomed.2021.104450] [Medline: 33989896]
48. Pepa L, Sabatelli A, Ciabattoni L, Monteriu A, Lamberti F, Morra L. Stress detection in computer users from keyboard and mouse dynamics. IEEE Trans Consumer Electron 2021 Feb;67(1):12-19. [doi: 10.1109/tce.2020.3045228]
49. Resilient data that just works. Agile Data Engine. URL: https://www.agiledataengine.com/ [accessed 2024-10-06]

Abbreviations

EMA: ecological momentary assessment
GDPR: General Data Protection Regulation
HR: human resource
IEQ: indoor environmental quality
IoT: Internet of Things

Edited by A Mavragani; submitted 01.07.23; peer-reviewed by J Zheng, PS Johnson; comments to author 06.12.23; revised version received 31.01.24; accepted 16.09.24; published 31.10.24.

Please cite as:
Kallio J, Kinnula A, Mäkelä SM, Järvinen S, Räsänen P, Hosio S, Bordallo López M
Lessons From 3 Longitudinal Sensor-Based Human Behavior Assessment Field Studies and an Approach to Support Stakeholder Management: Content Analysis
J Med Internet Res 2024;26:e50461
URL: https://www.jmir.org/2024/1/e50461
doi:10.2196/50461
PMID:

XSL•FO
RenderX

©Johanna Kallio, Atte Kinnula, Satu-Marja Mäkelä, Sari Järvinen, Pauli Räsänen, Simo Hosio, Miguel Bordallo López. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Health Survey Data Cost and Quality Between Amazon's Mechanical Turk and Ipsos' KnowledgePanel: Observational Study

Patricia M Herman[1], ND, PhD; Mary E Slaughter[1], PhD; Nabeel Qureshi[1], MPH, MPhil; Tarek Azzam[2], PhD; David Cella[3], PhD; Ian D Coulter[1], PhD; Graham DiGuiseppi[4], PhD; Maria Orlando Edelen[5,6], PhD; Arie Kapteyn[7], PhD; Anthony Rodriguez[6], PhD; Max Rubinstein[4], PhD; Ron D Hays[8], PhD

[1]RAND, Santa Monica, CA, United States

[2]Department of Education, University of California Santa Barbara, Santa Barbara, CA, United States

[3]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[4]RAND, Pittsburgh, PA, United States

[5]Brigham and Women's Hospital, Boston, MA, United States

[6]RAND, Boston, MA, United States

[7]Center for Economic and Social Research, Dana and David Dornsife School of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, United States

[8]Division of General Internal Medicine and Health Services Research, Department of Medicine, University of California Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Patricia M Herman, ND, PhD
RAND
1776 Main Street
Santa Monica, CA, 90407
United States
Phone: 1 3103930411 ext 7129
Email: pherman@rand.org

Abstract

Background: Researchers have many options for web-based survey data collection, ranging from access to curated probability-based panels, where individuals are selectively invited to join based on their membership in a representative population, to convenience panels, which are open for anyone to join. The mix of respondents available also varies greatly regarding representation of a population of interest and in motivation to provide thoughtful and accurate responses. Despite the additional dataset-building labor required of the researcher, convenience panels are much less expensive than probability-based panels. However, it is important to understand what may be given up regarding data quality for those cost savings.

Objective: This study examined the relative costs and data quality of fielding equivalent surveys on Amazon's Mechanical Turk (MTurk), a convenience panel, and KnowledgePanel, a nationally representative probability-based panel.

Methods: We administered the same survey measures to MTurk (in 2021) and KnowledgePanel (in 2022) members. We applied several recommended quality assurance steps to enhance the data quality achieved using MTurk. Ipsos, the owner of KnowledgePanel, followed their usual (industry standard) protocols. The survey was designed to support psychometric analyses and included >60 items from the Patient-Reported Outcomes Measurement Information System (PROMIS), demographics, and a list of health conditions. We used 2 fake conditions ("syndomitis" and "chekalism") to identify those more likely to be honest respondents. We examined the quality of each platform's data using several recommended metrics (eg, consistency, reliability, representativeness, missing data, and correlations) including and excluding those respondents who had endorsed a fake condition and examined the impact of weighting on representativeness.

Results: We found that prescreening in the MTurk sample (removing those who endorsed a fake health condition) improved data quality but KnowledgePanel data quality generally remained superior. While MTurk's unweighted point estimates for demographics exhibited the usual mismatch with national averages (younger, better educated, and lower income), weighted MTurk data matched national estimates. KnowledgePanel's point estimates better matched national benchmarks even before

XSL•FO
RenderX

poststratification weighting. Correlations between PROMIS measures and age and income were similar in MTurk and KnowledgePanel; the mean absolute value of the difference between each platform's 137 correlations was 0.06, and 92% were <0.15. However, correlations between PROMIS measures and educational level were dramatically different; the mean absolute value of the difference across these 17 correlation pairs was 0.15, the largest difference was 0.29, and the direction of more than half of these relationships in the MTurk sample was the opposite from that expected from theory. Therefore, caution is needed if using MTurk for studies where educational level is a key variable.

Conclusions: The data quality of our MTurk sample was often inferior to that of the KnowledgePanel sample but possibly not so much as to negate the benefits of its cost savings for some uses.

International Registered Report Identifier (IRRID): RR2-10.1186/s12891-020-03696-2

(J Med Internet Res 2024;26:e63032) doi:10.2196/63032

KEYWORDS

data collection; probability panel; convenience sample; data quality; weighting; back pain; misrepresentation; Amazon; Mechanical Turk; MTurk; convenience panel; KnowledgePanel

Introduction

Background

Web-based panel data collection may offer cost, timing, and data quality benefits over traditional modes such as mail and telephone surveys [1,2]. However, these benefits depend on the web-based platform one uses, its available respondents, and the quality of the data obtained. These platforms range from access to curated probability-based panels, where individuals are selectively invited to join based on their membership in a representative population, to convenience panels, which are open for anyone to join [3]. The mix of respondents available on each platform can vary greatly in terms of representation of the population that is the focus of the investigation and in motivation to provide thoughtful and accurate responses.

Despite the additional labor required of the researcher to achieve a dataset ready for analysis, convenience panels are much less expensive to use than probability-based panels. However, it is important to understand what may be given up in terms of data quality for those cost savings. Researchers want to collect survey data that are representative, reliable, and valid [4]. They want to be sure that the responses gained are reasonable measures of the topic of interest (construct validity) [5]. In addition, they may want to generalize from the sample to a target population (external validity) [6]. Important questions for survey researchers are whether the likely cost saving from using a convenience panel is worth the potentially lower quality of the data obtained and under what conditions is that trade-off appealing.

Although respondents from all sources may provide distracted responses [7], convenience panels can attract respondents who use the platform as a main source of income [8-10]. This creates incentives both for careless or inattentive responses due to the desire to complete a survey quickly and for misrepresentation when respondents make false claims to qualify for a study [8,11]. Although careless or inattentive responses can both attenuate and increase expected correlations and can affect estimated factor structures [12], fraudulent or dishonest responses can pose an even greater threat to a study's integrity by introducing systematic bias [5,8,13].

Convenience panels are also by design made up of individuals who are self-motivated to participate, and these individuals may not constitute a representative sample of the targeted underlying population [8]. A key technical requirement for standard statistics (eg, CIs) is that all members of the population of interest have a known, nonzero probability of being assigned to a survey [3,14,15]. This requirement is not met with a convenience sample and is of most concern when precise point estimates of population values are required [16,17]. Representativeness may also be important for unbiased estimates of relationships between variables of interest [16]. However, there are also study types (eg, explorative, methodological, and psychometric research) that may not require representative samples and benefit more from diversity [16].

Objectives

In this study, we compared the cost of and data gathered using one of the most well-known and widely used convenience panels (Amazon's Mechanical Turk [MTurk]) [18,19] to the cost of and data gathered using a nearly identical survey with a high-quality probability-based panel (KnowledgePanel) [17]. Given the different motivations and incentives facing each panel's members, we first compared the quality of the data collected from MTurk and KnowledgePanel participants in terms of various measures of self-report accuracy, including whether additional data cleaning, especially in the MTurk sample, could improve data quality. Then, given any differences observed between the platforms in terms of sample demographic composition, we examined whether weighting could adequately address any differences observed and improve point estimates. Finally, we examined whether relationships between variables were similar across platforms and ended with a discussion of the situations under which a researcher would want to use each platform for data collection.

Methods

To evaluate the advantages and disadvantages of using a convenience panel for web-based data collection versus using a probability-based panel, we fielded essentially the same survey using MTurk (August 31, 2021, to November 2, 2021) and then using KnowledgePanel (September 22, 2022, to October 2, 2022). We report our findings following both the Strengthening

XSL•FO
RenderX

the Reporting of Observational Studies in Epidemiology statement [20] and the Checklist for Reporting Results of Internet E-Surveys [21].

Data Sources

In this study, we used Amazon's MTurk as an example of a well-known, inexpensive, fast, easily accessible convenience panel that can be used for data collection. As we note later, although data collection methods on MTurk have improved since, we used the methods recommended at the start of our data collection (August 2021). Therefore, hereafter, what we have labeled as MTurk data should be considered to represent a somewhat lower quality than what is possible now. MTurk was launched in 2005 [18,19,22]. Anyone who is aged ≥18 years and has a computing device connected to the internet is eligible to become an MTurk worker by creating a worker account. Once they have an account, they can select and complete any of the human intelligence tasks (HITs) available to them. Researchers can request a variety of HIT types from workers, such as identifying photo images, transcription, and responding to surveys [23]. Workers search for HITs using a search interface that shows them which ones they qualify for, the title of the HIT, the requester-generated description of the HIT, and the payment rate [10,24]. Workers are paid after successfully completing an HIT; the recommended pay rate is the federal minimum wage, although many HITs pay an even lower rate [6,25,26]. MTurk is an international panel with an estimated 226,500 workers in the United States, of whom 81,000 to 86,000 completed at least one HIT in 2016 to 2018 [27]. Because 25% to 42% of workers use MTurk as a main source of income [6,9,10] (see the examples on the website Finance Over Fifty [28] and Gigworker [29]), and to prevent automated methods to complete HITs, there are limits placed on the number of HITs to which a worker can respond per day [30].

KnowledgePanel [31] is a high-quality [17] probability-based panel founded in 1999 by Knowledge Networks and now owned by Ipsos Public Affairs [16,32]. Its >55,000 members are recruited using an address-based sampling methodology that uses the latest delivery sequence file of the US Postal Service. This probability-based sampling methodology improves population coverage, particularly for hard-to-reach individuals such as young adults and minority subgroups [31]. Most KnowledgePanel members have their own internet access and computers, but those who do not (approximately 5% [17]) are provided with a device and access as needed to ensure that the panel is representative of all adults in the United States regardless of phone, electronic device, and internet access status. Surveys are assigned to a random sample of panel members who meet desired sample criteria. Once assigned to a survey, the panel member receives an email notification, and reminders for response are sent via email followed by phone calls as needed. In general, panelists respond to an average of 2 to 3 surveys per month and receive a modest incentive through a point system (eg, 5000 points are worth approximately US $5) for each survey they screen in for and complete. Panelists do not receive payments for the screening surveys themselves; instead, those who do not screen in are entered into a sweepstakes. Panelists are also paid for maintaining their panel status (approximately US $4 to US $6 per month). Another study that used KnowledgePanel to gather health-related data found that the opportunity to learn health information about oneself was a strong motivation to participate [33]. Over 80% of KnowledgePanel participants are not a part of any other survey panel, and new panel members are recruited throughout the year to make up for attrition [34].

Ethical Considerations

This study was approved by the RAND Human Subject Protection Committee (approval number 2019-0651-AM02). Respondents on each platform first faced a consent screen that laid out the study's purpose; that their participation was voluntary, they could choose not to answer any question, and they could stop at any time; and that, responses would remain anonymous, only be used for research purposes, and only be reported for groups. Respondents gave consent by clicking to join the survey. The identities of the KnowledgePanel respondents were only known to Ipsos, and we had no access to MTurk respondents' identities. The respondents to both surveys were compensated according to the usual procedures and amounts used for each platform.

Comparison of Data Sources

Researcher access to platforms with probability-based panels requires a contract with the organization that owns the panel. Ipsos formatted and programmed our survey instrument, pretested it and then fielded it to 1 adult each from a nationally representative sample of households with email reminders sent every 3 days, delivered a fully formatted dataset containing the survey data with variable and value labels, created poststratification (nonresponse) statistical weights, and provided KnowledgePanel respondents' standard demographic profile variables (ie, demographic data that Ipsos has on file for each panel member).

In contrast, a convenience panel such as MTurk is more of a self-service platform [9,18]. Access to MTurk workers is available to anyone with an Amazon Web Services account for the cost of the incentives paid to workers and a fee to Amazon. Requesters (researchers) must develop and format the survey in some generally accessible program (we used SelectSurvey [ClassApps, Inc]) and post it (much like a job advertisement) on the MTurk site available to workers looking for HITs, and they are responsible for releasing and monitoring the survey, as well as for data download, cleaning, and weighting.

We applied several quality assurance steps to enhance the data quality achieved using the MTurk platform [6,8,22,35-37]. Ipsos, the owner of KnowledgePanel, followed their usual (industry standard) protocols in fielding the survey. Table 1 provides a comparison of how the survey was fielded on each platform.

XSL•FO
RenderX

Table 1. Characteristics of Mechanical Turk (MTurk) and KnowledgePanel and the fielding of the survey on each platform.

Characteristic	MTurk	KnowledgePanel
Background on each data collection platform		
Requirements for joining the panel	Platform is open to anyone aged ≥18 years with a computing device connected to the internet	Panelists must be invited to join based on a residential address sample from the US Postal Service; Ipsos provides computer or internet service if needed
Financial agreement	Workers are paid based on the HITs[a] they complete	Panelists receive points redeemable for cash per survey they screen in for and complete; entry into a sweepstakes, but no payment, is offered for the screening
Motivation	Many use MTurk for income	Panelists do not earn enough to make income a motivation
Fielding of this survey		
Respondent requirements	95% approval rate and ≥500 HITs; unique US IP address required[b]	Surveys are assigned to a random sample of panelists
Monitoring respondent forums	Monitor worker forums for mention of the survey	N/A[c]
Survey timing—how it is released to respondents	Microbatches of 9 surveys each released every hour until the sample size is achieved	Released to all assigned panelists at once
Survey timing— time spent in the field for the baseline survey	August 31, 2021, to November 2, 2021 (63 days)	September 22, 2022, to October 2, 2022 (10 days)
Survey timing—time limits on responses	After someone starts the survey, they must complete it within 48 hours, or they will be dropped	Survey is available for panelists to complete for up to 10 days
Incentives for this survey	US $1.50 for general health survey plus US $2.00 for back pain survey	Respondents to the general health survey were entered into a monthly sweepstakes for prizes; those who qualified for and completed the back pain survey received 5000 points (approximately US $5)
Survey characteristics		
Language	English only	English only
Total number of items in the general health survey	101	100
Differences in items between surveys	Included 8 additional demographic items that were available as profile variables from KnowledgePanel	Added EQ-5D-5L (n=6) and PROMIS[d] social isolation (n=4) items
Fake conditions (n=2)	Yes	Yes
Advancing through the survey	1 item per page; required to click “Next” to advance	Sets of items with identical response categories were presented on 1 page in an “accordion” format, with each item’s response categories becoming visible as a response was given for the previous item; all other items were offered 1 per page, with clicking on “Next” required to advance
Review step	No review step, but workers could use a back button to revise their answers	No review step, but respondents could use a back button to revise their answers
Impact on researcher workload		
IRB[e] implications—respondent identifiers	All anonymous but can be contacted for follow-up	Identity only known to Ipsos
Formatting and programming the survey	Researcher builds it	Ipsos builds it with researcher input
Survey initiation	Under researcher control	Ipsos schedules it with researcher input
Data cleaning and quality control	Entirely the responsibility of the researcher; requires a higher level of effort	Ipsos does some quality control and cleaning, so there is a lower level of effort for the researcher

[a]HIT: human intelligence task.

[b]These qualifications have been recommended by a number of studies [6,8,22,36,37].

[c]N/A: not applicable.

[d]PROMIS: Patient-Reported Outcomes Measurement Information System.

XSL•FO
RenderX

[e]IRB: institutional review board.

Survey Design

The survey was designed to enable psychometric evaluation and estimation of links and crosswalks between commonly used patient-reported outcome measures [38]. The first part of the survey was fielded as a survey of general health to all respondents [13,25,26] (ie, made available to all US-based MTurk workers and assigned to a nationally representative random sample of KnowledgePanel members). This portion of the survey contained >60 items from the Patient-Reported Outcomes Measurement Information System (PROMIS), demographics, and a list of health conditions that the respondent endorsed as "ever been told by a doctor or other health professional that" they had (14 items) or that they "currently" had (10 items). The second part of the survey was only offered to those who indicated current back pain and contained several established instruments used to measure back pain impact. The survey was fielded in English on both data collection platforms. Those who qualified for and completed the back pain survey at baseline were asked to complete that survey again at 3 and 6 months on the MTurk platform and were assigned that survey again at 6 months on KnowledgePanel.

Although early studies using MTurk found the platform to produce data of a quality equal to or better than those produced by many other sources and across several types of data [8,18], there have also been concerns about data quality [22,39]. To obtain the best quality data possible from MTurk, we followed several recommendations from the literature at that time. We limited the MTurk workers eligible for the survey to those who were located in the United States and had a good reputation [6,8,9,13,22,36,37] (ie, they had completed >500 HITs with a >95% approved-for-payment rating). MTurk workers are approved by their "employers" (researchers or requesters) using any criteria that the employers set up. In this study, workers were approved for payment if they reached the end of the survey and submitted the survey code available there. To ensure response consistency, we also eliminated from the sample anyone who completed the survey in an unrealistically short time [8,12,18,39]—less than 1 second per item—and we eliminated from the analytic file those who did not complete at least half of their assigned study items.

Survey Implementation

As HITs are available to MTurk workers on a first-come-first-served basis, we limited any time-of-day or day-of-the-week bias by using microbatching—automatically releasing 9 surveys per hour, 24 hours per day, until our target sample was achieved (ie, over 2 months) [40]. An application called CloudResearch (formerly TurkPrime; Prime Research Solutions LLC) was used to accomplish this microbatching [41]. CloudResearch connects with the MTurk programming interface, enabling greater control over the survey process. CloudResearch has a tool that breaks a larger survey into microbatches of <10 participants each, ensuring sampling from individuals who are online at different times throughout the day and different days throughout the week. Releasing <10 surveys at a time also reduces the Amazon fee by 50%. In addition, CloudResearch allows for exclusion of people who have already completed the study (to prevent duplicate submissions) and enables anonymous emails to workers (allowing longitudinal data to be collected).

Some MTurk workers misrepresent their health status to increase their chances of being chosen for a survey [5,11,25,26,39]. Therefore, we used a recommended screening step to identify desired respondents for the back pain survey [5,7,25,26]. All respondents received a survey of general health that contained a list of health conditions that they indicated they had "ever" or "currently" had. This list included current back pain, but no indication was given that this was our target condition for the follow-up survey. Our pilot study work revealed that a substantial number of MTurk respondents (20%) endorsed having all or essentially all conditions listed—raising the question of whether some of these were fraudulent responses [5]. To identify those more likely to be honest respondents, we embedded 2 fake conditions in the list ("syndomitis" and "chekalism") [8,13,42]. Those who endorsed either of these conditions were not offered the back pain survey even if they indicated that they had back pain. We used the same procedure to identify patients with back pain on the KnowledgePanel platform. For the MTurk platform, we also monitored online worker forums (eg, [43-46]) [26] to see whether there was any chatter about our survey and our fake conditions but found none.

The full text of all survey instruments fielded on each platform can be found on the Inter-university Consortium for Political and Social Research data repository (OPENICPSR-198049) [47].

Analyses

The analyses used the data from the survey of general health. To create an analytic file for each data collection platform, we cleaned the data (eg, checking for out-of-range variables) and removed data from respondents whose response times were <1 second per item or who answered less than half of the items and created derived variables. Completion rates were calculated for each analytic sample, and the expenses and labor required for data collection on each platform were recorded. We then identified the respondents who endorsed one or both fake conditions and examined the quality of each platform's data using several recommended metrics (eg, evidence of straight-lining and reliability; see the following paragraphs) including and excluding those respondents.

We used the κ statistic to assess the consistency of responses to 5 item pairs that addressed similar topics and had identical response categories [42,48] (refer to Multimedia Appendix 1 for the item pairs compared). Landis and Koch [49] provided a rule of thumb for the interpretation of κ—values of <0 indicated poor agreement, values of 0.00 to 0.20 indicated slight agreement, values of 0.21 to 0.40 indicated fair agreement, values of 0.41 to 0.60 indicated moderate agreement, values of 0.61 to 0.80 indicated substantial agreement, and values of 0.81 to 1.00 indicated almost perfect agreement. We also evaluated cases in which responses were too consistent [8,23,50] (ie, "straight-lining," where respondents gave identical responses to consecutive sets of items [9,51]). We calculated the number

XSL•FO
RenderX

and percentage of respondents who (1) chose the same response category for all sets of consecutive, same-response-category items; (2) chose the same response category for a set of 6 items on physical function where it would be unlikely that identical answers would make sense (ie, identical responses of "without any difficulty" [ratings of 5] were allowed, but identical responses of other response categories were not); and (3) chose the same response category for all items in one or both sets of 3 items on sleep where one item in each set was asked in a positive way and 2 were asked in a negative way (ratings of 3 [*somewhat* or *sometimes*] on a scale from 1 to 5 were allowed). We also calculated mean root of pairs [51] for each of these 3-item sleep sets—numbers closer to 1.0 indicated more straight-lining. Finally, we compared internal consistency reliability (Cronbach α [52]) for the 7 domain scales on the 29-item PROMIS (PROMIS-29) profile (physical function, fatigue, pain interference, depressive symptoms, anxiety, ability to participate, and sleep disturbance) plus the 2-item cognitive function scale using a χ^2 test for independent samples [9,39,53-55].

Institutional review boards often require that respondents be allowed to skip or not answer any item in the survey. In this study, MTurk workers were only required to get to the end of the survey and submit a completion code to be approved and paid. KnowledgePanel panelists agreed upon joining the panel to answer all survey questions truthfully unless they felt uncomfortable doing so. Nevertheless, respondents on both platforms could leave many items unanswered, and the amount of missing data could affect conclusion validity and the generalizability of the study findings. Our data cleaning procedures removed those missing responses to more than half of the items from the analytic sample, and then we calculated the proportion of respondents in that sample for each platform who completed all items in the survey of general health. Finally, we reported the response burden for each platform in terms of the Winsorized average and median and ranges of time it took to complete the survey of general health, in each case after capping the top 2.5% of durations to the 97.5th percentile value [56].

All the aforementioned data quality checks were performed on the full sample for the survey of general health. However, that section of the survey did not include any open-ended questions where free-text responses could be examined [23,39]. The back pain survey (only offered to those who endorsed current back pain and did not endorse a fake condition) included an open-ended question—"What does chronic pain mean to you?" Therefore, we compared responses from MTurk and KnowledgePanel in terms of whether (1) the response was nonsensical (not related to the question, eg, "good" or "text") and (2) the response was copied from a common source (ie, contained a string of at least 10 words that made up at least 75% of the response and were identical to those found in one or more other open-ended responses). We also reported the number of those who completed the back pain survey who went on to complete the 3- and 6-month follow-up surveys on MTurk and the 6-month follow-up survey on KnowledgePanel.

We examined the representativeness of the MTurk and KnowledgePanel samples by comparing the point estimates from each to US national estimates first including and then excluding those who endorsed one or both fake conditions and again after applying weights to those who did not endorse a fake condition [57]. Estimates for the demographic variables that were used for weighting (age, gender, race and ethnicity, income, educational level, and region), the PROMIS-29 scales [58] and physical and mental health summary scores [59], and disease prevalence were compared.

The KnowledgePanel sample was selected to match the population of adults in the United States using a probability-proportional-to-size procedure that used a set of design weights as measures of size. Therefore, even before any additional weighting, we would expect that sample to be closer to national estimates (March 2022 supplement of the Current Population Survey) than the MTurk sample. After the KnowledgePanel survey data were collected, these design weights were adjusted using poststratification weights, which can help account for any differential nonresponse [34]. Ipsos used an iterative proportional fitting (raking) procedure with trimming and scaling to produce the final weights. For MTurk, we followed a similar procedure and created weights to account for both the original composition of respondents and for any nonresponse and examined how useful weighting was in improving the accuracy of the point estimates from each platform [57,60].

Finally, some researchers are more interested in relationships between variables and argue that results of multivariate analyses are more similar between convenience samples and probability-based samples than those of univariate analyses [14,57,61-63]. Correlations measure the strength and direction of bivariate relationships. We provided a correlation matrix (using unweighted data excluding those who endorsed a fake condition and including 95% CIs) to compare platforms in terms of the linear relationships between the variables measured in the survey of general health—the 7 PROMIS-29 domain scale T-scores [58], the mental and physical health summary T-scores both from the PROMIS-29 [59] and from the PROMIS Global Health items [64], the Impact Stratification Score [65], age, educational level, and income. The coefficients for each dataset are shown above and below each other in the matrix to allow for "ocular" comparisons. Although we provide some estimates of the differences observed (eg, means and maximum absolute differences), we leave it to the reader to judge the extent to which these coefficients are similar enough for their analytic needs.

Results

Comparison of Labor Days and Dollar Expenditures Required for Use of Each Platform

Table 2 shows the study team labor days and dollar expenditures involved in each step of generating the data and analytic datasets gathered from each platform for this study. Although the labor needs for the study team were lower, the dollar expenditures to obtain these data from KnowledgePanel were substantially (approximately 9 times) higher. Of course, gathering smaller

XSL•FO
RenderX

samples would have cost less, but the relative size of the expenses would increase because, although the MTurk expenses were solely determined by sample size, the KnowledgePanel expenses included components relatively insensitive to sample size.

Table 2. Labor and expenditures to create the final analytic datasets from each platform.

	Mechanical Turk		KnowledgePanel	
	Labor days, n	Expenditures (US $)	Labor days, n	Expenditures (US $)
Formatting and programming of survey instrument	3.3	0	0	236,178[a]
Pretests and then full fielding of the survey	2.8	0	0	236,178[a]
Incentive payments	1.7	27,146	0	236,178[a]
Fully formatted SAS[b] dataset with appropriate variable and value labels	1.7	0	0	236,178[a]
Weighting	1.5	0	0	236,178[a]
Creation of final analytic dataset	15.8	0	7.8	0
Total	26.7	27,146	7.8	236,178[a]

[a]These tasks did not individually cost US $236,178. This amount is the total across all these tasks.

[b]SAS: Statistical Analysis System.

Comparison of the Numbers of Respondents and Quality of the Data From Each Platform

Tables 3 and 4 presents the comparison of various measures of data quality between platforms. Neither platform had any respondents with response times of <1 second per item. Analytic datasets were created by removing those who did not answer at least half of the items they were assigned (designated as incompletes); the percentage of incomplete surveys removed from the MTurk dataset was almost 9 times that of incomplete surveys removed from the KnowledgePanel dataset. Our completion rates based on the analytic datasets were 49.6% (6750/13,608) for MTurk (with 13,608 being the number of surveys available [63 days × 24 hours × 9 surveys released per hour]) and 57.2% (4134/7224) for KnowledgePanel (with 7224 being the number of panelists assigned to the survey).

Table 3. Numbers of respondents and completion rates for each platform.

	Mechanical Turk	KnowledgePanel
Surveys fielded, n	13,608	7224
Final total upon field close, n	6997	4149
Surveys with too short response times (<1 second per item), n (%)	0 (0)	0 (0)
Incomplete surveys (missing more than half of the items), n/N (%)	247/6997 (3.5)	15/4149 (0.4)
Analytic dataset, n	6750	4134
Completion rate, n/N (%)	6750/13,608 (50)	4134/7224 (57)
Participants who endorsed fake conditions, n/N (%)	975/6750 (14.4)	19/4134 (0.5)

XSL•FO
RenderX

Table 4. Indicators of quality for the survey data collected on each platform[a].

	All (n=6750)	No fake conditions (n=5775)	All (n=4134)	No fake conditions (n=4115)
Response consistency, κ (quadratically weighted for agreement; see Multimedia Appendix 1 for the item pairs)				
2 items on pain interference	0.83	0.83	0.91	0.91
2 items on trouble doing	0.72	0.73	0.77	0.77
2 items on problems with sleep	0.73	0.72	0.67	0.67
2 items on ability to concentrate or focus	0.46	0.47	0.71	0.71
2 items on memory	0.55	0.56	0.70	0.70
Measures of straight-lining				
Participants who straight-lined all sets of same-response-category items, n (%)	20 (0.3)	11 (0.2)	6 (0.1)	5 (0.1)
Participants who straight-lined the 6-item physical function set (no ratings of 5), n (%)	98 (1.5)	55 (1)[b]	34 (0.8)	34 (0.8)
Participants who straight-lined one or both sleep item sets (no ratings of 3), n (%)	611 (9.1)	441 (7.6)[c]	237 (5.7)	233 (5.7)[d]
Root of pairs (first sleep set; larger=more), mean (SD)[e]	0.43 (0.26)	0.41 (0.25)[d]	0.39 (0.25)	0.39 (0.24)[d]
Root of pairs (second sleep set; larger=more), mean (SD)[e]	0.45 (0.28)	0.43 (0.27)[d]	0.42 (0.26)	0.42 (0.26)
Internal consistency reliability				
Physical function, Cronbach α for PROMIS[f] scales	0.881	0.894[d]	0.938	0.938[d]
Fatigue, Cronbach α for PROMIS scales	0.919	0.924[b]	0.942	0.943[d]
Pain interference, Cronbach α for PROMIS scales	0.940	0.938	0.963	0.963[d]
Depressive symptoms, Cronbach α for PROMIS scales	0.918	0.922	0.939	0.940[d]
Anxiety, Cronbach α for PROMIS scales	0.900	0.901	0.906	0.907
Ability to participate, Cronbach α for PROMIS scales	0.922	0.923	0.947	0.947[d]
Sleep disturbance, Cronbach α for PROMIS scales	0.775	0.840[j]	0.875	0.875[d]
Cognitive function (2 items), Cronbach α for PROMIS scales	0.759	0.770	0.844	0.844[d]
Cognitive function (5 items), Cronbach α for PROMIS scales	0.856	0.862	0.905	0.905[d]
Participants who completed all items in the general health survey, n (%)	6152 (91.1)	5379 (93.1)[d]	3345 (80.9)	3335 (81)[d]
Response burden (minutes)—general health survey, Winsorized mean (SD)	21 (14)	20 (13)	11 (7)	11 (7)
Response burden (minutes)—general health survey, median (range)	17 (3-55)	15 (3-55)[d]	10 (2-38)	10 (2-38)[d]
Only asked of back pain survey respondents				
Eligible for back pain survey (endorsed back pain but no fake conditions), n (%)	—[g]	2307 (39.9)[h]	—	1533 (37.3)[i]
Responded to back pain survey (back pain analytic file), n (%)	—	1972 (34.1)[h]	—	1531 (37.2)[i]
Response burden (minutes)—general health survey+back pain survey, Winsorized mean (SD)	—	32 (13)	—	25 (14)
Response burden (minutes)—general health survey+back pain survey, median (range)	—	30 (5-58)	—	21 (3-75)[d]
Nonsense responses to “what is chronic?” n (%)	—	32 (1.6)[j]	—	4 (0.3)[d,k]
Copied or identical text in response to “what is chronic?” n (%)	—	212 (10.8)[j]	—	0 (0.0)[d,k]
Responded to 3-month follow-up surveys, n (%)	—	1077 (54.6)[j]	—	—

XSL•FO
RenderX

	All (n=6750)	No fake conditions (n=5775)	All (n=4134)	No fake conditions (n=4115)
Responded to 6-month follow-up surveys, n (%)	—	845 (42.8)[j]	—	1256 (82.0)[k]

[a]Indicators of statistical strength of differences between Mechanical Turk—all and Mechanical Turk—no fake conditions are shown in the Mechanical Turk no fake conditions column and between Mechanical Turk—no fake conditions and KnowledgePanel—no fake conditions are shown in the KnowledgePanel no fake conditions column. We used 2-tailed *t* tests for comparisons of means and χ^2 tests for comparisons of frequencies and for the comparisons of α coefficients.

[b]*P*<.05.

[c]*P*<.01.

[d]*P*<.001.

[e]Mean root of pairs=the mean of the root of the absolute differences between all pairs of items in a battery, rescaled to range from 0 (least straight-lining) to 1 (most straight-lining).

[f]PROMIS: Patient-Reported Outcomes Measurement Information System.

[g]Not applicable.

[h]n=5775.

[i]n=4115.

[j]n=1972.

[k]n=1531.

The rate of respondents endorsing one or both fake conditions was almost 30 times higher on the MTurk than the KnowledgePanel platform. As the incidence of fake condition endorsement was so low in the KnowledgePanel dataset, removing those who endorsed these conditions had little impact on its data quality. However, removing those who endorsed a fake condition from the MTurk analytic dataset generally improved its data quality, in some cases substantially [35]. Removing those who endorsed a fake condition had little impact on response consistency (ie, κ values between similar pairs of items barely increased) but did reduce the incidence of straight-lining (ie, respondents giving identical responses to consecutive sets of similar items); improve internal consistency reliability, especially for a 4-item scale with reverse-coded items (sleep disturbance) and for physical function; and increase the proportion of respondents who completed all items. On average, KnowledgePanel respondents took a little more than half the time that MTurk respondents took to complete the general health survey. However, despite this response speed, and even after removing those who endorsed the fake conditions from the MTurk dataset, the KnowledgePanel dataset showed better response consistency—for the concentration/focus and memory pairs, estimated κ values from KnowledgePanel were “substantial” versus “moderate” for MTurk. KnowledgePanel respondents exhibited less straight-lining, better internal consistency reliability, and fewer nonsense and copied responses for the open-ended back pain survey item and had a substantially (39 percentage points) higher completion rate for the 6-month follow-up survey. Nevertheless, significantly fewer KnowledgePanel respondents than MTurk respondents completed all items in the survey of general health.

Comparison of the Characteristics of Respondents on Each Platform

Table 5 shows the demographic characteristics of respondents for the 2 platforms. The biggest impact of removing those who endorsed a fake condition in the MTurk sample was that the proportion of respondents who identified as Hispanic individuals dropped from 19.54% (1319/6750) to 14.06% (812/5775) in the sample; 52% (507/975) of those who endorsed a fake condition identified as Hispanic individuals. Before weighting and when comparing those who did not endorse a fake condition, MTurk respondents were younger, with most respondents in the age category of 30 to 44 years, whereas the age category with the most respondents in KnowledgePanel was ≥60 years, better matching national estimates. More MTurk respondents were male, fewer identified as non-Hispanic Black individuals, and more identified as Hispanic individuals than those from KnowledgePanel. While both platforms had similar proportions of respondents with a master’s degree or higher, KnowledgePanel respondents included a larger proportion of those with an educational level of high school or lower than MTurk (1369/4115, 33.27% vs 480/5775, 8.31%), and MTurk had a much larger proportion of respondents with a bachelor’s degree (2807/5775, 48.61% vs 907/4115, 22.04%). MTurk had more respondents in the income categories of <US $100,000 per year, and KnowledgePanel had more in the category of ≥US $100,000 per year (4984/5775, 86.3% vs 1740/4115, 42.28%). Similar proportions of MTurk and KnowledgePanel respondents resided in each of the 4 US census regions.

XSL•FO
RenderX

Table 5. Comparison of the demographic characteristics before and after weighting between the Mechanical Turk (MTurk) and KnowledgePanel samples and national estimates.

	MTurk			KnowledgePanel			National estimates (%)[a]
	All (n=6750)	No fake conditions (n=5775)	Weighted (n=5775)	All (n=4134)	No fake conditions (n=4115)	Weighted (n=4098)	
Age (years), mean (SD)	39 (12)	40 (12)	46 (15)	52 (18)	52 (18)	48 (18)	47.5
Age (years), n (%)							
18-29	1268 (18.8)	1130 (19.6)	1158 (20.1)	560 (13.5)	559 (13.6)	817 (19.9)	19.9
30-44	3514 (52.1)	2913 (50.4)	1484 (25.7)	951 (23)	945 (23)	1059 (25.8)	25.9
45-60	1392 (20.6)	1204 (20.8)	1349 (23.4)	914 (22.1)	909 (22.1)	980 (23.9)	24
≥60	499 (7.4)	466 (8.1)	1706 (29.5)	1709 (41.3)	1702 (41.4)	1242 (30.3)	30.3
Gender, n (%)							
Female	2931 (43.4)	2617 (45.3)	2882 (49.9)	2040 (49.3)	2033 (49.4)	2096 (51.1)	50.7
Male	3695 (54.7)	3047 (52.8)	2754 (47.7)	2055 (49.7)	2044 (49.7)	1962 (47.9)	49.3
Race and ethnicity, n (%)							
Hispanic	1319 (19.5)	812 (14.1)	979 (17)	499 (12.1)	495 (12)	702 (17.1)	17.2
Multiracial	124 (1.8)	121 (2.1)	80 (1.4)	136 (3.3)	136 (3.3)	57 (1.4)	1.4
Non-Hispanic Black	531 (7.9)	477 (8.3)	686 (11.9)	411 (9.9)	411 (10)	495 (12.1)	12
Non-Hispanic White	4362 (64.6)	3968 (68.7)	3537 (61.2)	2894 (70)	2880 (70)	2542 (62)	62
Non-Hispanic other	330 (4.9)	328 (5.7)	418 (7.2)	194 (4.7)	193 (4.7)	301 (7.3)	7.3
Educational attainment, n (%)							
No high school diploma or GED[b]	18 (0.3)	18 (0.3)	540 (9.4)	278 (6.7)	277 (6.7)	394 (9.6)	9.6
High school graduate or GED	466 (6.9)	462 (8)	1650 (28.6)	1100 (26.6)	1092 (26.5)	1192 (29.1)	29.2
Some college or associate's degree	1405 (20.8)	1384 (24)	1495 (25.9)	1088 (26.3)	1083 (26.3)	1083 (26.4)	26.4
Bachelor's degree	3380 (50.1)	2807 (48.6)	1269 (22)	909 (22)	907 (22)	784 (19.1)	34.8[c]
Master's degree or higher	1385 (20.5)	1024 (17.7)	698 (12.1)	759 (18.4)	756 (18.4)	644 (15.7)	34.8[c]
Household income (US $), n (%)							
<10,000	275 (4.1)	249 (4.3)	302 (5.2)	121 (2.9)	121 (2.9)	144 (3.5)	3.6
10,000-49,999	2812 (41.7)	2438 (42.2)	1764 (30.5)	1007 (24.4)	999 (24.3)	1022 (24.9)	24.9
50,000-99,999	2821 (41.8)	2297 (39.8)	1597 (27.7)	1258 (30.4)	1255 (30.5)	1172 (28.6)	28.6
≥100,000	762 (11.3)	727 (12.6)	2037 (35.3)	1748 (42.3)	1740 (42.3)	1760 (42.9)	42.9
Region of the country, n (%)							
Northeast	1125 (16.7)	1012 (17.5)	992 (17.2)	751 (18.2)	745 (18.1)	710 (17.3)	17.4
Midwest	1292 (19.1)	1134 (19.6)	1176 (20.4)	900 (21.8)	896 (21.8)	845 (20.6)	20.6
South	2670 (39.6)	2283 (39.5)	2183 (37.8)	1522 (36.8)	1514 (36.8)	1568 (38.3)	38.3
West	1582 (23.4)	1281 (22.2)	1351 (23.4)	961 (23.2)	960 (23.3)	974 (23.8)	23.7

[a]National estimates on this table are all from the Current Population Survey Annual Social and Economic Supplement (March 2022) for persons aged ≥18 years in the United States.

[b]GED: General Educational Development.

[c]This value reflects the combined sum of both rows.

XSL•FO
RenderX

Both samples were weighted using the demographic variables shown in Table 5 (age, gender, race and ethnicity, educational level, income, and region). Weighting these samples resulted in datasets that generally well matched the national estimates. The KnowledgePanel weights (maximum weight of 2.8) brought those data completely in line with national estimates, whereas weighting the MTurk data allowing for a maximum weight of 30 brought those data within a total absolute imbalance of 0.01 across the 6 demographic variables used to construct the weights.

Table 6 shows other characteristics of the samples from each platform. In the full dataset, the MTurk sample generally had worse PROMIS T-scores than the KnowledgePanel sample and national estimates for all scales. Removing respondents who endorsed a fake condition generally brought the average PROMIS T-score on each scale for the MTurk respondents closer to those of KnowledgePanel and to national estimates. Only considering PROMIS scores that differed by >2 T-scores (ie, a “small” effect size) compared to national estimates after weighting, the MTurk sample had more anxiety, better ability to participate in social roles and activities, and a better mental health summary score than national estimates. The KnowledgePanel sample had less fatigue and better ability to participate in social roles and activities, cognitive function, and mental and physical health summary scores than national estimates.

XSL•FO
RenderX

Table 6. Characteristics of the samples from each platform.

	Mechanical Turk			KnowledgePanel			National estimates
	All (n=6750)	No fake conditions (n=5775)	Weighted (n=5775)	All (n=4134)	No fake conditions (n=4115)	Weighted (n=4098)	
PROMIS-29+2[a] scale T-scores, mean (SD)							
Physical function	48 (8)	49 (8)	50 (8)	51 (8)	51 (8)	51 (8)	50 (10)[b]
Pain interference	53 (10)	51 (9)	50 (9)	49 (9)	49 (9)	49 (9)	50 (10)[b]
Fatigue	51 (10)	50 (10)	49 (11)	48 (10)	48 (10)	48 (10)	50 (10)[b]
Depressive symptoms	54 (10)	53 (10)	51 (10)	48 (9)	48 (9)	49 (9)	50 (10)[b]
Anxiety	56 (10)	54 (10)	52 (10)	49 (9)	49 (9)	49 (9)	50 (10)[b]
Ability to participate	52 (10)	53 (9)	55 (10)	56 (9)	56 (9)	56 (9)	50 (10)[b]
Sleep disturbance	50 (9)	50 (9)	49 (10)	49 (9)	49 (9)	49 (9)	50 (10)[b]
Cognitive function	49 (9)	50 (9)	51 (9)	52 (9)	52 (9)	52 (9)	50 (10)[b]
Mental health summary score	48 (9)	50 (9)	52 (10)	53 (9)	53 (9)	53 (9)	50 (10)[b]
Physical health summary score	48 (9)	49 (9)	51 (9)	51 (9)	51 (9)	52 (9)	50 (10)[b]
Number of health conditions endorsed (out of 24 possible conditions)	6.0 (6.0)	3.8 (3.4)	3.7 (3.1)	4.0 (3.2)	4.0 (3.2)	3.8 (3.1)	__[c]
Disease prevalence, n (%)							
Anxiety (ever)	2293 (33.97)	1618 (28.02)	1520 (26)	814 (19.69)	806 (19.59)	845 (21)	33.7[d]
Depression (ever)	2737 (40.55)	2005 (34.72)	1688 (29)	829 (20.05)	820 (19.93)	832 (20)	18.4[e]
Hypertension (ever)	2423 (35.9)	1578 (27.32)	1786 (31)	1581 (38.24)	1570 (38.15)	1381 (34)	47.3[f]
Asthma (ever)	1561 (23.13)	889 (15.39)	770 (13)	536 (12.97)	532 (12.93)	547 (13)	13.5[g]
Diabetes (ever)	1360 (20.15)	678 (11.74)	542 (9)	554 (13.4)	548 (13.32)	496 (12)	13[f]
Heart disease (ever had any of these: heart attack, CHD[h], or angina—3 different)	1428 (21.16)	538 (9.32)	414 (7)	301 (7.28)	295 (7.17)	237 (6)	5.5[g]
COPD[i] (ever)	923 (13.67)	293 (5.07)	369 (6)	193 (4.67)	191 (4.64)	165 (4)	5.6[j]
Cancer (ever)	948 (14.04)	295 (5.11)	459 (8)	421 (10.18)	417 (10.13)	325 (8)	7.5[g]
Stroke (ever)	921 (13.64)	254 (4.4)	133 (2)	109 (2.64)	106 (2.58)	88 (2)	2.8[g]
Neck pain (currently)	2077 (30.77)	1393 (24.12)	1136 (20)	816 (19.74)	811 (19.71)	781 (19)	15.7[k]
Back pain (currently)	3035 (44.96)	2307 (39.95)	2136 (37)	1541 (37.28)	1533 (37.25)	1467 (36)	39[g]
Responded to the back pain survey	—	1972 (34.15)	1872 (32)	—	1531 (37.21)	1467 (36)	—
Nonspecific back pain	—	1471 (74.59)[l]	1228 (66)	—	983 (64.21)[m]	981 (67)	—
Chronic (>3-month duration)	—	1539 (78.04)[l]	1593 (85)	—	1379 (90.07)[m]	1312 (90)	—
Chronic (RTF[n] definition)	—	1174 (59.53)[l]	1169 (62)	—	868 (56.69)[m]	826 (57)	—
Chronic (provider identified)	—	495 (25.1)[l]	624 (33)	—	552 (36.05)[m]	498 (34)	—

XSL•FO
RenderX

	Mechanical Turk			KnowledgePanel			National estimates
	All (n=6750)	No fake conditions (n=5775)	Weighted (n=5775)	All (n=4134)	No fake conditions (n=4115)	Weighted (n=4098)	
Chronic (patient identified)	—	1156 (58.62)[l]	1245 (67)	—	865 (56.5)[m]	801 (55)	—
Chronic (any definition)	—	1719 (87.17)[l]	1718 (92)	—	1400 (91.44)[m]	1330 (91)	—

[a]PROMIS-29+2: 29-item Patient-Reported Outcomes Measurement Information System profile plus 2 cognitive items.

[b]Patient-Reported Outcomes Measurement Information System T-scores have a mean of 50 and an SD of 10 in a national sample.

[c]There is no national estimate available for this characteristic.

[d]National Comorbidity Survey Replication.

[e]Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report [66].

[f]National Health and Nutrition Examination Survey 2017 to 2018.

[g]2019 National Health Interview Survey data.

[h]CHD: coronary heart disease.

[i]COPD: chronic obstructive pulmonary disease.

[j]2017 Behavioral Risk Factor Surveillance System.

[k]2013 to 2015 National Health Interview Survey data.

[l]n=1972.

[m]n=1531.

[n]RTF: National Institutes of Health Pain Consortium's 2012 research task force.

Removing MTurk respondents who endorsed a fake condition substantially reduced the total number of health conditions endorsed and the number endorsing each condition, and weighting these data further reduced the number claiming to have each condition except for hypertension and cancer. Removing those who endorsed a fake condition had little effect on the KnowledgePanel results, but weighting those data generally resulted in slightly lower condition prevalence, with the greatest reduction in the prevalence of hypertension. The weighted data showed that the MTurk respondents had more anxiety and depression and slightly more back and neck pain, whereas the KnowledgePanel respondents had more hypertension and diabetes. Both platforms underestimated anxiety and hypertension compared to available national estimates and overestimated depression and neck pain. After removing those who endorsed a fake condition, the prevalence of back pain was remarkably similar between the samples. However, MTurk respondents who were eligible for the back pain survey were asked whether they wanted to go on to take that survey, and 15% (335/2307) opted out, dropping back pain prevalence in that sample substantially below national averages. KnowledgePanel respondents did not receive that opt-out question. After weighting, the proportions with nonspecific back pain were similar between samples, but the prevalence of each type of chronic back pain differed. Compared to KnowledgePanel, the proportion of the MTurk sample who said that they thought their back pain is chronic was 12 percentage points higher.

Comparison of Correlations Between Variables Measured on Each Platform

Finally, correlation coefficients estimated using the data from each platform showed many similarities but some important differences. In the correlation matrix using the full sample (Multimedia Appendix 2), the correlation coefficients for the PROMIS measure variables (ie, all but the last 3 rows) for each dataset were all in the same direction and tended to be similar in magnitude, with the KnowledgePanel correlation coefficient being larger 55% of the time. Of those 105 correlation pairs, the mean absolute value of the difference between each platform's correlations was 0.05, with differences of >0.05 in 45% of the correlations, >0.10 in 24% of the correlations, and >0.15 in 7% of the correlations. Only one correlation pair (between the Impact Stratification Score and PROMIS global mental health summary score) had an absolute value difference of >0.20 (0.202). The MTurk sample's correlation coefficients between age, income, and educational level and PROMIS measures were larger than those of KnowledgePanel 85% of the time. However, the differences among the coefficients for PROMIS measures, age, and income followed a similar pattern to that observed between the PROMIS measures alone. It is notable that the 17 correlation coefficient pairs between educational level and the other variables differed markedly by platform. More than half (9/17, 53%) of the correlation pairs had opposite signs (with the MTurk coefficient indicating a negative relationship between educational level and better health as measured using the PROMIS, which is contrary to theory [67]), and the absolute difference among 7 correlation coefficient pairs was >0.20, the average difference across all 17 pairs was 0.15, and the largest difference was 0.29.

Discussion

Principal Findings

Overview

KnowledgePanel and MTurk are examples of 2 types of web-based platforms that differ markedly in terms of the cost and investigator effort required for data collection but also in terms of the respondents and the quality of the data obtained

XSL•FO
RenderX

[3]. KnowledgePanel members are experienced survey respondents, and new members are recruited to the platform as needed to maintain a nationally representative panel. Ipsos then randomly draws a sample from this panel and assigns them to a survey. In contrast, anyone aged ≥18 years with a computing device connected to the internet can become an MTurk worker, and when they want to, they can search for and respond to HITs such as a survey. On KnowledgePanel, a random representative sample is assigned to a survey; on MTurk and similar convenience opt-in panels, although one can take some steps to improve data quality, one generally takes what one can get. As expected, the quality of the data obtained from KnowledgePanel respondents was measurably better than that of the data obtained from MTurk, but the KnowledgePanel sample analyzed in this study cost thousands of dollars more. This study provides information with which to weigh the benefits of KnowledgePanel's representative, assigned sample against the cost savings afforded by platforms such as MTurk.

There are three main considerations when comparing the data received from different sources: (1) accurate responses in terms of both potential misrepresentation and thoughtful/careful answers, (2) quality of point estimates, and (3) comparisons of multivariate analyses. In the following sections, we compare the results for our MTurk and KnowledgePanel samples (each gathered according to the criteria described in Table 1) with regard to these considerations.

Misrepresentation and Thoughtful/Careful Answers

In this study, our prescreening, including the use of fake conditions to identify and remove respondents who were misrepresenting themselves (or at least careless), helped improve data quality in the MTurk sample. MTurk respondents endorsed a fake health condition at almost 30 times the rate observed in KnowledgePanel (975/6750, 14.4% vs 19/4134, 0.5%). However, even after those who endorsed a fake condition were removed, the quality of the KnowledgePanel data remained somewhat superior. We also offered fake conditions in the follow-up surveys, and according to another study of these data, it is possible that almost a quarter of the MTurk sample were misrepresenting themselves or careless [35]—an estimate that is within the range of misrepresentation observed in other MTurk surveys [13]. Our longitudinal data made it possible to estimate the percentage of misrepresenters more accurately than a single administration. However, some misrepresenters likely remained unidentified. The comparison of the relative data quality of MTurk and KnowledgePanel could be different if we knew with certainty those who were misrepresenting themselves.

Removing respondents who endorsed a fake condition substantially reduced the number and prevalence of the health conditions endorsed by MTurk respondents, but it also reduced the number of respondents who identified as Hispanic individuals. It is unclear why more than half of those who endorsed a fake condition also identified as being Hispanic. It could be that this is a further instance of erroneous reporting, or it could be that respondents thought that claiming Hispanic ethnicity would increase their chances of being chosen for further surveys. In any case, a systematic review of 54 MTurk studies found that racial and ethnic minority groups were more likely to be excluded using a variety of screening criteria [7].

It is unclear why a larger proportion of MTurk respondents (247/6997, 3.5%) than of KnowledgePanel respondents (15/4149, 0.4%) were removed from the analytic dataset for answering less than half the questions but more remaining MTurk respondents (6152/6750, 90.1% and 5379/5775, 93.1% before and after excluding those who endorsed a fake condition, respectively) than KnowledgePanel respondents (3345/4134, 80.9% and 3335/4115, 81%) answered all the questions in the main survey. The consent screen for both sets of respondents included that their participation was voluntary, they could choose not to answer any question, and they could stop at any time. However, once started, the MTurk workers may have felt an implicit obligation to complete the task fully to receive a good rating.

One result that was surprising is that, compared to the MTurk respondents, the KnowledgePanel respondents provided higher-quality data according to several measures while being substantially faster—11 versus 20 minutes for those who only completed the survey of general health and 25 versus 32 minutes for both the general health and back pain surveys combined. There could be many reasons for this, including that KnowledgePanel respondents tend to be long-term, experienced members of a panel paid to respond to surveys or that MTurk respondents may keep multiple HITs open longer than necessary to avoid being penalized for responding too fast. However, it also could be that the MTurk survey showed respondents 1 item per page and required them to click "Next" to move to the next item, which can take longer [68]. If they chose to not answer an item, it also required that they validate their choice before moving forward. This same format was used for many of the items in the KnowledgePanel survey. However, when there were series of items all with the same response categories, Ipsos used an "accordion" format that showed all items in a set on 1 page and revealed the response categories for each as an answer for the previous item was selected. This is a format familiar to KnowledgePanel respondents and may have allowed them to move more quickly through these items.

Quality of Point Estimates

When the goal of data collection is to determine point estimates, it should be noted that standard statistics (eg, CIs) require that all members of the population of interest have a known, nonzero probability of being assigned to a survey [3,14,15]. Because workers on MTurk and similar opt-in panels select themselves into that panel and then find the HIT on their own rather than being selected to join a panel based on a known sampling frame and being randomly assigned to a survey as is done in KnowledgePanel, there is uncertainty about the inferences that can be made from the data [3,14]. Of course, low response rates can also impair probability-based panels' ability to meet this known, nonzero probability of selection requirement, especially if nonresponders cannot be shown to be missing at random [15].

The KnowledgePanel completion rate (4134/7224, 57.2%) was higher than MTurk's (6750/13,608, 49.6%); however, both are within the range observed in a large meta-analysis of online surveys [69] and higher than that study's average response rate

XSL•FO
RenderX

(44%). Nevertheless, the difference in completion rates between KnowledgePanel and MTurk was not as large as might be expected given that KnowledgePanel members were assigned the survey and the MTurk workers had to find it themselves through a posting. KnowledgePanel also had a higher completion rate on the 6-month follow-up survey (1205/1531, 82%) than MTurk (1077/1972, 55% at 3 months and 854/1972, 43% at 6 months). Part of this difference was likely because eligible KnowledgePanel respondents were assigned the 6-month follow-up survey, whereas the MTurk workers, known to have a high turnover rate [70], had to respond to an email inviting them again.

Bias due to self-selection and nonresponse can be addressed through adjustments such as weighting [3]. However, there is no guarantee that weighting will be successful; bias will only be reduced if the proper weighting variables are used, and they are often unknown [3,57,71,72]. We used the same variables for weighting the MTurk data as were used for the KnowledgePanel data with the exception of KnowledgePanel subdividing their regions into metro and nonmetro areas. As shown elsewhere [14,57] and as expected given its random assignment of surveys from a probability-based panel, the KnowledgePanel unweighted point estimates were closer to national benchmarks than the unweighted estimates from MTurk. However, even though MTurk's unweighted point estimates for demographics exhibited the usual mismatch with national averages (younger, better educated, and lower income [6,10,15,18,73]), we were able to weight the MTurk data to match national estimates with a total absolute imbalance close to 0 (0.01). This weighting also improved MTurk's PROMIS T-scores, resulting in estimates closer to national averages than those observed in KnowledgePanel. Weighting the MTurk sample also reduced the proportion endorsing each condition and generally moved prevalence closer to what was observed in KnowledgePanel. However, the maximum weight required for weighting the MTurk data was 30, which is quite high. Educational level was by far the most difficult variable to weight, which may be due to the unusual relationship between educational level and other variables observed in our correlation analyses (more on this in the following section). Nevertheless, as with any weighting procedure, this weighting may or may not have truly improved all point estimates [57,71,72,74].

Even after weighting, more KnowledgePanel than MTurk respondents had household incomes of >US $100,000 per year. Because there are fewer individuals with incomes of >US $100,000 than there are households with incomes of >US $100,000 [75], we checked the household size in each sample. The proportion of MTurk respondents reporting single-person households was lower than that for KnowledgePanel (880/5775, 15% vs 692/4115, 17%), making household size an unlikely explanation for the difference observed in household income levels.

Comparisons of Multivariate Analyses

The results of multivariate analyses might be more similar across platforms than univariate results [14,57,61-63]. In this study, we estimated and compared correlation coefficients for each platform. Overall, the correlation pairs demonstrated many similarities but also some important differences. Of the 105 correlation pairs between PROMIS measure variables, the mean absolute difference was 0.05, and the mean absolute difference between the additional 31 coefficients involving age and income was 0.07. However, the 17 correlation coefficient pairs involving educational level were dramatically different by platform—more than half had opposite signs (with MTurk correlations being the opposite from what would be expected from theory). In addition, the absolute difference among 7 correlation coefficient pairs was >0.20, and the largest difference was 0.29. These results are indicators of systematic differences between the 2 samples at least in terms of the relationship between educational level and PROMIS measures. Educational level also stood out as a problematic predictor of beliefs, attitudes, and knowledge in another study using Qualtrics Panels (a nonprobability panel selected using a variety of opt-in methods), suggesting that educational level might function differently in nonprobability panels than in nationally representative samples [15].

Similar to what was found in another study [76], 85% (41/48) of all the correlation coefficients in the full sample between PROMIS measures and demographics were larger in absolute value in MTurk than KnowledgePanel, suggesting an increased risk of false positives in those relationships. Slightly more (58/105, 55%) KnowledgePanel coefficients were larger than in MTurk when only considering correlations between PROMIS measures.

We only examined correlations between the variables we measured. Other researchers have suggested that the results of multivariate models, where the effects of several variables can be estimated at once, might differ less between probability and convenience samples, especially if the dependent variable is more "concrete" [15,77]. It has also been suggested that randomization to experimental conditions might also minimize biases observed in MTurk samples and offer internal validity at least as good as that observed in undergraduate student pools [15].

Strengths and Limitations

This study benefited from large sample sizes and the opportunity to field essentially the same survey on 2 very different (convenience vs probability-based panel) platforms. However, there are some limitations.

First, the MTurk data were gathered earlier in the COVID-19 pandemic (August 31, 2021, to November 2, 2021) than the KnowledgePanel data (September 22, 2022, to October 2, 2022). The stage of the pandemic during which we gathered our data could have had an impact. However, as our health questions were about back pain and not about infectious disease, the main impact would likely be due to the inability to exercise or visit providers, especially earlier in the pandemic. This may have contributed to the slightly higher physical function and slightly lower pain interference T-scores in the KnowledgePanel data than in the MTurk data. However, other studies have also found worse health in nonprobability samples [72], and it is unclear whether this was a significant effect or whether these were the only effects of our timing.

XSL•FO
RenderX

Second, the KnowledgePanel sample included some respondents (5% [17]) who were provided with a device or internet connection to enable their participation, whereas everyone in the MTurk sample was required to have both [14,57].

Third, we only fielded the surveys in English.

Fourth, it is not likely that the use of the same fake conditions ("syndomitis" and "chekalism") will continue to be as effective for prescreening a population as information about their use could spread among future respondents. This is similar to what happened over time to the use of "Bindro" as a fake drug name in surveys [78]. It is also possible that detecting careless respondents will be more challenging in the future as more sophisticated bots are used.

Fifth, other quality steps regarding the selection of workers on the MTurk platform may have resulted in different outcomes. For example, we chose to use workers with a good reputation (ie, those who had completed >500 HITs with a >95% approval rating). However, this may have limited us to the savvier survey takers. Other researchers have recommended recruiting naïve workers [27], and some studies are now recommending using those with a 99% approval rating [79] or using workers in CloudResearch's Approved Participants group to improve MTurk data quality [80].

Sixth, a short initial prescreening survey could also be used to restrict an MTurk sample to better represent a segment of the US population (eg, younger adults [15]), or using CloudResearch to target workers may have resulted in a more nationally representative sample even though these respondents have been found to be less attentive and provide less reliable answers [18]. Nevertheless, MTurk and other nonprobability panels have also been found to be sufficiently accurate for certain types of studies. For example, a nonprobability sample should be sufficient if a researcher is investigating a phenomenon believed to be universal (ie, everyone would behave similarly) or one believed to be appropriately distributed in any large population so that the specific makeup of the sample is unlikely to affect conclusions [14,81].

Conclusions

Similar to what has been shown in other studies [57,71], we found that a nationally representative probability-based sample resulted in higher-quality, more representative estimates compared to a convenience/nonprobability sample. Therefore, if cost is not a consideration, obtaining data from a nationally representative probability-based sample is recommended. However, given its substantially lower costs, interest in using platforms such as MTurk is likely to continue, and data quality was not so far from that observed for KnowledgePanel in our study as to negate any use of these convenience panels [15]. There are cases in which lower quality would be acceptable. For example, even if researchers can afford a probability-based panel, MTurk may be appropriate to generate hypotheses and pilot studies before preregistration and fielding [22].

Although the likelihood that a random sample is representative of a population could be determined if certain conditions are met, the accuracy of a nonprobability sample is nearly impossible to completely assess [14,15]. Nevertheless, a number of steps can improve the data collected from platforms such as MTurk. These improvements include making the survey available only once to high-quality (≥500 HITs and ≥95% approval rating) local (US IP addresses) workers in microbatches across the day and week, using a prescreening step (paid initial survey) that masks the existence of and eligibility criteria for any following more targeted surveys and includes fake or bogus items to identify those who might be misrepresenting themselves to qualify for further work, and weighting to match national benchmarks.

With appropriate prescreening and weighting, nonprobability samples can often be used but always with caution and awareness of methods to mitigate their shortcomings [15].

Acknowledgments

The authors would like to acknowledge and thank Rebecca Weir for cleaning and preparing the Mechanical Turk and KnowledgePanel data for analysis and the rest of this project's advisory council (Dagmar Amtmann, Penney Cowan, Kurt Kroenke, and Michael Von Korff) for their support and input on the project as a whole since its inception in 2020. No generative artificial intelligence was used in any portion of manuscript writing. The research reported in this publication was supported by the National Center for Complementary and Integrative Health of the National Institutes of Health under award R01AT010402. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated during and/or analyzed during this study are available in the Inter-university Consortium for Political and Social Research repository (OPENICPSR-198049) [47].

Authors' Contributions

PMH, RDH, and MOE conceptualized the project and acquired the funding. MES contributed to data curation, and PMH, RDH, and MES covered methodology and the formal analysis. PMH was responsible for project administration, and she and RDH supervised the project. PMH wrote the original draft with significant input from RDH and MES (tables). All coauthors participated in writing in terms of review and editing and have approved the final version of this manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Item pairs used for consistency checks.
[DOCX File , 14 KB - jmir_v26i1e63032_app1.docx]

Multimedia Appendix 2
Correlations between various Patient-Reported Outcomes Measurement Information System scores and age, educational level, and income.
[DOCX File , 23 KB - jmir_v26i1e63032_app2.docx]

References

1. Chang L, Krosnick JA. National surveys via RDD telephone interviewing versus the internet: comparing sample representativeness and response quality. Public Opin Q 2009;73(4):641-678. [doi: 10.1093/poq/nfp075]
2. Wyatt JC. When to use web-based surveys. J Am Med Inform Assoc 2000 Jul 01;7(4):426-429. [doi: 10.1136/jamia.2000.0070426] [Medline: 10887170]
3. Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA, Lavrakas PJ. Online panel research: history, concepts, applications and a look at the future. In: Callegaro M, Baker R, Bethlehem J, Goritz AS, Krosnick JA, Lavrakas PJ, editors. Online Panel Research: A Data Quality Perspective. New York, NY: John Wiley & Sons; 2014:1-22.
4. Trochim WM. The Research Methods Knowledge Base. Cincinnati, OH: Atomic Dog Publishing; 2001.
5. Chandler JJ, Paolacci G. Lie for a Dime: when most prescreening responses are honest but most study participants are impostors. Soc Psychol Personal Sci 2017 Apr 27;8(5):500-508. [doi: 10.1177/1948550617698203]
6. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgm Decis Mak 2023 Jan 01;5(5):411-419. [doi: 10.1017/S1930297500002205]
7. Thomas KA, Clifford S. Validity and mechanical turk: an assessment of exclusion methods and interactive experiments. Comput Human Behav 2017 Dec;77:184-197. [doi: 10.1016/j.chb.2017.08.038]
8. Agley J, Xiao Y, Nolan R, Golzarri-Arroyo L. Quality control questions on Amazon's Mechanical Turk (MTurk): a randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. Behav Res Methods 2022 Apr 06;54(2):885-897 [FREE Full text] [doi: 10.3758/s13428-021-01665-8] [Medline: 34357539]
9. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022 Aug 29;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
10. Hitlin P. Research in the crowdsourcing age: a case study. Pew Research Center. 2016. URL: http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/ [accessed 2023-05-08]
11. Sharpe Wessling K, Huber J, Netzer O. MTurk character misrepresentation: assessment and solutions. J Consum Res 2017;44(1):211-230 [FREE Full text] [doi: 10.1093/jcr/ucx053]
12. Wood D, Harms PD, Lowman GH, DeSimone JA. Response speed and response consistency as mutually validating indicators of data quality in online samples. Soc Psychol Personal Sci 2017 Jun 07;8(4):454-464. [doi: 10.1177/1948550617703168]
13. MacInnis CC, Boss HC, Bourdage JS. More evidence of participant misrepresentation on Mturk and investigating who misrepresents. Pers Individ Dif 2020 Jan;152:109603. [doi: 10.1016/j.paid.2019.109603]
14. Cornesse C, Blom AG, Dutwin D, Krosnick JA, De Leeuw ED, Legleye S, et al. A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. J Surv Stat Methodol 2020;8(1):4-36. [doi: 10.1093/jssam/smz041]
15. Zack ES, Kennedy J, Long JS. Can nonprobability samples be used for social science research? A cautionary tale. Surv Res Methods 2019;13(2):215-227 [FREE Full text]
16. Hays RD, Liu H, Kapteyn A. Use of Internet panels to conduct surveys. Behav Res Methods 2015 Sep 14;47(3):685-690. [doi: 10.3758/s13428-015-0617-9] [Medline: 26170052]
17. Bradley VC, Kuriwaki S, Isakov M, Sejdinovic D, Meng XL, Flaxman S. Unrepresentative big surveys significantly overestimated US vaccine uptake. Nature 2021 Dec 08;600(7890):695-700 [FREE Full text] [doi: 10.1038/s41586-021-04198-4] [Medline: 34880504]
18. Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L. Online panels in social science research: expanding sampling methods beyond Mechanical Turk. Behav Res Methods 2019 Oct 11;51(5):2022-2038. [doi: 10.3758/s13428-019-01273-7] [Medline: 31512174]
19. Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk Platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med 2018 Apr 4;33(4):533-538. [doi: 10.1007/s11606-017-4246-0] [Medline: 29302882]

XSL•FO
RenderX

20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/s0140-6736(07)61602-x]
21. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
22. Kennedy R, Clifford S, Burleigh T, Waggoner PD, Jewell R, Winter NJ. The shape of and solutions to the MTurk quality crisis. Political Sci Res Methods 2020 Apr 24;8(4):614-629. [doi: 10.1017/psrm.2020.6]
23. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods 2012 Mar 30;44(1):1-23. [doi: 10.3758/s13428-011-0124-6] [Medline: 21717266]
24. Ipeirotis PG. Analyzing the Amazon Mechanical Turk marketplace. The ACM Magazine for Students. 2010 Dec. URL: https://dl.acm.org/doi/10.1145/1869086.1869094 [accessed 2024-04-29]
25. Kim HS, Hodgins DC. Are you for real? Maximizing participant eligibility on Amazon's Mechanical Turk. Addiction 2020 Oct 20;115(10):1969-1971. [doi: 10.1111/add.15065] [Medline: 32309892]
26. Hydock C. Assessing and overcoming participant dishonesty in online data collection. Behav Res Methods 2018 Aug 22;50(4):1563-1567. [doi: 10.3758/s13428-017-0984-5] [Medline: 29274007]
27. Robinson J, Rosenzweig C, Moss AJ, Litman L. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. PLoS One 2019 Dec 16;14(12):e0226394 [FREE Full text] [doi: 10.1371/journal.pone.0226394] [Medline: 31841534]
28. How to make $50 a day on MTurk: the ultimate guide in 2023. Finance Over Fifty. URL: https://financeoverfifty.com/how-to-make-50-a-day-on-mturk-ultimate-guide/ [accessed 2024-04-29]
29. Definitive guide on how to make money on Amazon MTurk in 2024. Gigworker. 2023. URL: https://gigworker.com/amazon-mturk/ [accessed 2024-04-29]
30. FAQs. Amazon Mechanical Turk Worker. URL: https://www.mturk.com/worker/help [accessed 2024-04-29]
31. Bertoni N. KnowledgePanel. Ipsos. URL: https://www.ipsos.com/en-us/solutions/public-affairs/knowledgepanel [accessed 2024-04-29]
32. Crowther M, Thomas N. Ipsos acquires GfK's KnowledgePanel. Daily Research News Online. 2019. URL: https://www.mrweb.com/drno/news27826.htm [accessed 2024-04-29]
33. Kaufman DJ, Baker R, Milner LC, Devaney S, Hudson KL. A survey of U.S adults' opinions about conduct of a nationwide Precision Medicine Initiative® cohort study of genes and environment. PLoS One 2016 Aug 17;11(8):e0160461 [FREE Full text] [doi: 10.1371/journal.pone.0160461] [Medline: 27532667]
34. KnowledgePanel®: a methodological overview. Ipsos. 2023. URL: https://www.ipsos.com/sites/default/files/ipsosknowledgepanelmethodology.pdf [accessed 2024-04-29]
35. Hays RD, Qureshi N, Herman PM, Rodriguez A, Kapteyn A, Edelen MO. Effects of excluding those who report having "syndomitis" or "chekalism" on data quality: longitudinal health survey of a sample from Amazon's Mechanical Turk. J Med Internet Res 2023 Aug 04;25:e46421 [FREE Full text] [doi: 10.2196/46421] [Medline: 37540543]
36. Goodman JK, Cryder CE, Cheema A. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. J Behav Decis Mak 2012 Apr 02;26(3):213-224. [doi: 10.1002/bdm.1753]
37. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behav Res Methods 2014 Dec 20;46(4):1023-1031. [doi: 10.3758/s13428-013-0434-y] [Medline: 24356996]
38. Herman PM, Edelen MO, Rodriguez A, Hilton LG, Hays RD. A protocol for chronic pain outcome measurement enhancement by linking PROMIS-29 scale to legacy measures and improving chronic pain stratification. BMC Musculoskelet Disord 2020 Oct 10;21(1):671 [FREE Full text] [doi: 10.1186/s12891-020-03696-2] [Medline: 33038933]
39. Chmielewski M, Kucker SC. An MTurk crisis? Shifts in data quality and the impact on study results. Soc Psychol Personal Sci 2019 Oct 10;11(4):464-473. [doi: 10.1177/1948550619875149]
40. Walters K, Christakis DA, Wright DR. Are Mechanical Turk worker samples representative of health status and health behaviors in the U.S.? PLoS One 2018 Jun 7;13(6):e0198835 [FREE Full text] [doi: 10.1371/journal.pone.0198835] [Medline: 29879207]
41. Litman L, Robinson J, Abberbock T. TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav Res Methods 2017 Apr 12;49(2):433-442. [doi: 10.3758/s13428-016-0727-z] [Medline: 27071389]
42. Meade AW, Craig SB. Identifying careless responses in survey data. Psychol Methods 2012 Sep;17(3):437-455. [doi: 10.1037/a0028085] [Medline: 22506584]
43. mturk forum. XenForo™. URL: https://mturkforum.com/index.php [accessed 2024-04-29]
44. Welcome to MTurk Crowd. MTurk Crowd. URL: https://mturkcrowd.com/ [accessed 2024-04-29]
45. The Mechanical Turk (mTurk) hub. mTurk Forum by TurkerView. URL: https://forum.turkerview.com/#the-mechanical-turk-mturk-hub.1 [accessed 2024-04-29]
46. r/HITsWorthTurkingFor. Reddit. URL: https://www.reddit.com/r/HITsWorthTurkingFor [accessed 2024-04-29]
47. Find data. Institute for Social Research (ICPSR). URL: https://www.icpsr.umich.edu/web/pages/ICPSR/index.html [accessed 2024-04-29]

XSL•FO
RenderX

48. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr 01;20(1):37-46. [doi: 10.1177/001316446002000104]
49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
50. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/bm.2012.031]
51. Kim Y, Dykema J, Stevenson J, Black P, Moberg DP. Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. Soc Sci Comput Rev 2018 Feb 20;37(2):214-233. [doi: 10.1177/0894439317752406]
52. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/bf02310555]
53. Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to study clinical populations. Clin Psychol Sci 2013 Jan 31;1(2):213-220. [doi: 10.1177/2167702612469015]
54. Feldt LS, Woodruff DJ, Salih FA. Statistical inference for coefficient alpha. Appl Psychol Meas 1987 Mar 01;11(1):93-103. [doi: 10.1177/014662168701100107]
55. Diedenhofen B, Musch J. cocron: a web interface and R package for the statistical comparison of Cronbach's Alpha coefficients. Int J Inf Sci 2016;11:51-60 [FREE Full text]
56. Wilcox R. A foundation for robust methods. In: Wilcox R, editor. Introduction to Robust Estimation and Hypothesis Testing: A volume in Statistical Modeling and Decision Science. New York, NY: Academic Press; 2012:25-42.
57. Callegaro M, Villar A, Yeager D, Krosnick JA. A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In: Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA, Lavrakas PJ, editors. Online Panel Research. New York, NY: John Wiley & Sons; 2014:23-53.
58. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, et al. PROMIS adult health profiles: efficient short-form measures of seven health domains. Value Health 2019 May;22(5):537-544 [FREE Full text] [doi: 10.1016/j.jval.2019.02.004] [Medline: 31104731]
59. Hays RD, Spritzer KL, Schalet BD, Cella D. PROMIS-29 v2.0 profile physical and mental health summary scores. Qual Life Res 2018 Jul 22;27(7):1885-1891. [doi: 10.1007/s11136-018-1842-3] [Medline: 29569016]
60. Goel S, Obeng A, Rothschild D. Non-representative surveys: fast, cheap, and mostly accurate. Work Pap 2015:1 [FREE Full text]
61. Berrens RP, Bohara AK, Jenkins-Smith H, Silva C, Weimer DL. The advent of internet surveys for political research: a comparison of telephone and internet samples. Polit Anal 2017 Jan 04;11(1):1-22. [doi: 10.1093/pan/11.1.1]
62. Pasek J. When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence. Int J Public Opin Res 2015 Aug 08;28(2):269-291. [doi: 10.1093/ijpor/edv016]
63. Groves RM. Survey Errors and Survey Costs. Hoboken, NJ: John Wiley & Sons; 2004.
64. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009 Sep 19;18(7):873-880. [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
65. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Report of the NIH Task Force on research standards for chronic low back pain. Pain Med 2014 Aug 01;15(8):1249-1267. [doi: 10.1111/pme.12538] [Medline: 25132307]
66. Morbidity and mortality weekly report (MMWR): national, state-level, and county-level prevalence estimates of adults aged ≥18 years self-reporting a lifetime diagnosis of depression — United States, 2020. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/mmwr/volumes/72/wr/mm7224a1.htm?s_cid=mm7224a1_w#contribAff [accessed 2024-04-29]
67. Zajacova A, Lawrence EM. The relationship between education and health: reducing disparities through a contextual approach. Annu Rev Public Health 2018 Apr 01;39:273-289 [FREE Full text] [doi: 10.1146/annurev-publhealth-031816-044628] [Medline: 29328865]
68. Hays RD, Bode R, Rothrock N, Riley W, Cella D, Gershon R. The impact of next and back buttons on time to complete and measurement reliability in computer-based surveys. Qual Life Res 2010 Oct 15;19(8):1181-1184. [doi: 10.1007/s11136-010-9682-9] [Medline: 20552282]
69. Wu MJ, Zhao K, Fils-Aime F. Response rates of online surveys in published research: a meta-analysis. Comput Hum Behav Rep 2022 Aug;7:100206. [doi: 10.1016/j.chbr.2022.100206]
70. Stewart N, Ungemach C, Harris AJ, Bartels DM, Newell BR, Paolacci G, et al. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgm Decis Mak 2023 Jan 01;10(5):479-491. [doi: 10.1017/s1930297500005611]
71. Yeager DS, Krosnick JA, Chang L, Javitz HS, Levendusky MS, Simpser A, et al. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. Public Opin Q 2011;75(4):709-747. [doi: 10.1093/poq/nfr020]

XSL•FO
RenderX

72. Schnell R, Klingwort J. Health estimate differences between six independent web surveys: different web surveys, different results? BMC Med Res Methodol 2024 Jan 27;24(1):24 [FREE Full text] [doi: 10.1186/s12874-023-02122-0] [Medline: 38281040]
73. Qureshi N, Edelen M, Hilton L, Rodriguez A, Hays RD, Herman PM. Comparing data collected on Amazon's Mechanical Turk to national surveys. Am J Health Behav 2022 Oct 17;46(5):497-502. [doi: 10.5993/ajhb.46.5.1]
74. MacInnis B, Krosnick JA, Ho AS, Cho MJ. The accuracy of measurements with probability and nonprobability survey samples: replication and extension. Public Opin Q 2018;82(4):707-744. [doi: 10.1093/poq/nfy038]
75. Current population survey annual social and economic supplement. US Census Bureau. URL: https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.html [accessed 2024-04-29]
76. Pasek J, Krosnick JA. Measuring intent to participate and participation in the 2010 census and their correlates and trends: comparisons of RDD telephone and non-probability sample internet survey data. United States Census Bureau. URL: https://www.census.gov/library/working-papers/2010/adrm/ssm2010-15.html [accessed 2024-04-29]
77. Simmons AD, Bobo LD. Can non-full-probability internet surveys yield useful data? A comparison with full-probability face-to-face surveys in the domain of race and social inequality attitudes. Sociol Methodol 2015 Mar 11;45(1):357-387. [doi: 10.1177/0081175015570096]
78. Bindro by Corbyn. Urban Dictionary. 2005. URL: https://www.urbandictionary.com/define.php?term=Bindro [accessed 2024-04-29]
79. Matherly T. A panel for lemons? Positivity bias, reputation systems and data quality on MTurk. Eur J Mark 2019 Feb 11;53(2):195-223. [doi: 10.1108/ejm-07-2017-0491]
80. Hauser DJ, Moss AJ, Rosenzweig C, Jaffe SN, Robinson J, Litman L. Evaluating CloudResearch's approved group as a solution for problematic data quality on MTurk. Behav Res Methods 2023 Dec 03;55(8):3953-3964. [doi: 10.3758/s13428-022-01999-x] [Medline: 36326997]
81. Coppock A, Leeper TJ, Mullinix KJ. Generalizability of heterogeneous treatment effect estimates across samples. Proc Natl Acad Sci U S A 2018 Dec 04;115(49):12441-12446. [doi: 10.1073/pnas.1808083115] [Medline: 30446611]

Abbreviations

HIT: human intelligence task
MTurk: Mechanical Turk
PROMIS: Patient-Reported Outcomes Measurement Information System
PROMIS-29: 29-item Patient-Reported Outcomes Measurement Information System

Edited by A Coristine; submitted 07.06.24; peer-reviewed by M DeBell, S Stadtmüller ; comments to author 01.10.24; revised version received 14.10.24; accepted 18.10.24; published 29.11.24.

Please cite as:
Herman PM, Slaughter ME, Qureshi N, Azzam T, Cella D, Coulter ID, DiGuiseppi G, Edelen MO, Kapteyn A, Rodriguez A, Rubinstein M, Hays RD
Comparing Health Survey Data Cost and Quality Between Amazon's Mechanical Turk and Ipsos' KnowledgePanel: Observational Study
J Med Internet Res 2024;26:e63032
URL: https://www.jmir.org/2024/1/e63032
doi:10.2196/63032
PMID:

©Patricia M Herman, Mary E Slaughter, Nabeel Qureshi, Tarek Azzam, David Cella, Ian D Coulter, Graham DiGuiseppi, Maria Orlando Edelen, Arie Kapteyn, Anthony Rodriguez, Max Rubinstein, Ron D Hays. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Resting Heart Rate and Associations With Clinical Measures From the Project Baseline Health Study: Observational Study

Kent Y Feng[1*], MD; Sarah A Short[2*], MPH; Sohrab Saeb[2*], PhD; Megan K Carroll[2], MS; Christoph B Olivier[3], MD; Edgar P Simard[2], MPH, PhD; Susan Swope[1], MS; Donna Williams[1], MPH, MSN; Julie Eckstrand[4], RPh; Neha Pagidipati[4], MD, MPH; Svati H Shah[4], MD, MHS; Adrian F Hernandez[4], MD; Kenneth W Mahaffey[1], MD

[1]Stanford Center for Clinical Research, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

[2]Verily Life Sciences, South San Francisco, CA, United States

[3]Cardiovascular Clinical Research Center, Department of Cardiology and Angiology, University Heart Center Freiburg, Bad Krozingen, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[4]Duke University School of Medicine, Durham, NC, United States

[*]these authors contributed equally

Corresponding Author:
Sarah A Short, MPH
Verily Life Sciences
269 E Grand Ave
South San Francisco, CA, 94080
United States
Phone: 1 650 495 7100
Email: sarahshort@verily.com

Abstract

Background: Though widely used, resting heart rate (RHR), as measured by a wearable device, has not been previously evaluated in a large cohort against a variety of important baseline characteristics.

Objective: This study aimed to assess the validity of the RHR measured by a wearable device compared against the gold standard of ECG (electrocardiography), and assess the relationships between device-measured RHR and a broad range of clinical characteristics.

Methods: The Project Baseline Health Study (PHBS) captured detailed demographic, occupational, social, lifestyle, and clinical data to generate a deeply phenotyped cohort. We selected an analysis cohort within it, which included participants who had RHR determined by both ECG and the Verily Study Watch (VSW). We examined the correlation between these simultaneous RHR measures and assessed the relationship between VSW RHR and a range of baseline characteristics, including demographic, clinical, laboratory, and functional assessments.

Results: From the overall PBHS cohort (N=2502), 875 (35%) participants entered the analysis cohort (mean age 50.9, SD 16.5 years; n=519, 59% female and n=356, 41% male). The mean and SD of VSW RHR was 66.6 (SD 11.2) beats per minute (bpm) for female participants and 64.4 (SD 12.3) bpm for male participants. There was excellent reliability between the two measures of RHR (ECG and VSW) with an intraclass correlation coefficient of 0.946. On univariate analyses, female and male participants had similar baseline characteristics that trended with higher VSW RHR: lack of health care insurance (both $P<.05$), higher BMI (both $P<.001$), higher C-reactive protein (both $P<.001$), presence of type 2 diabetes mellitus (both $P<.001$) and higher World Health Organization Disability Assessment Schedule (WHODAS) 2.0 score (both $P<.001$) were associated with higher RHR. On regression analyses, within each domain of baseline characteristics (demographics and socioeconomic status, medical conditions, vitals, physical function, laboratory assessments, and patient-reported outcomes), different characteristics were associated with VSW RHR in female and male participants.

Conclusions: RHR determined by the VSW had an excellent correlation with that determined by ECG. Participants with higher VSW RHR had similar trends in socioeconomic status, medical conditions, vitals, laboratory assessments, physical function, and patient-reported outcomes irrespective of sex. However, within each domain of baseline characteristics, different characteristics were most associated with VSW RHR in female and male participants.

Trial Registration: ClinicalTrials.gov NCT03154346; https://clinicaltrials.gov/study/NCT03154346

XSL•FO
RenderX

(J Med Internet Res 2024;26:e60493) doi:10.2196/60493

KEYWORDS

resting heart rate; wearable devices; remote monitoring; physiology; PBHS; Project Baseline Health Study; Verily Study Watch; heart rate; observational study; cohort study; wearables; electrocardiogram; regression analyses; socioeconomic status; medical condition; vital signs; laboratory assessments; physical function; electronic health; eHealth

Introduction

Resting heart rate (RHR) has been extensively studied in healthy individuals and those with specific disease states such as cardiovascular disease (CVD) [1,2]. Increasing RHR is linked to the development of CVD risk factors such as diabetes mellitus and hypertension and is implicated as an important prognostic factor in those with CVD and cancer [3,4]. Due to these links with important clinical outcomes such as the development of disease and mortality, RHR and RHR trends are of high interest to clinicians and patients alike and have become highly accessible, particularly with the recent ubiquity of wearable devices capable of recording heart rate (HR) and even detecting concerning arrhythmias such as atrial fibrillation [5].

Traditionally, RHR is determined through clinical measurements during physical examinations as well as electrocardiography (ECG), and ambulatory devices. In the recent decade, wearable devices have become increasingly popular; many have the capability to track fitness levels with a variety of metrics, including steps, HR, and sleep. Commercially available devices have been shown to be accurate in measuring HR and steps, and studies suggest that wearable devices may improve physical activity [6-8].

The Project Baseline Health Study (PBHS) was a prospective, multicenter, longitudinal cohort study launched in 2017 to establish a comprehensive reference health state using a wide range of modalities, evaluate different technologies in measuring disease trajectory and participant diversity, and share this information with both scientists and participants. The PBHS enrolled 2502 participants to include a broad range of healthy individuals with varying disease risks (specifically CVD, breast/ovarian cancer, and lung cancer), as well as those with known disease diagnoses. The PBHS provides an opportunity to describe and assess RHR using a wearable device (Verily Study Watch [VSW]) in a contemporary population and to do so in a comprehensive and more continuous manner than previously done [9]. Previous studies have limited comparisons with clinical measurements or have small sample sizes focused on specific disease states [10-12]. The design of the PBHS allows for an extensive analysis of RHR as they relate to multimodal clinical data collected from remote and in-person visits in a deeply phenotyped cohort, allowing a unique opportunity to explore potentially significant relationships. In this exploratory study, we aimed to (1) identify an analysis cohort within the PBHS and compare baseline characteristics with the overall study cohort at large, (2) validate the VSW's determination of RHR (VSW RHR) by comparing against the gold standard of RHR by ECG, and (3) assess the relationships between VSW RHR and a broad range of baseline clinical characteristics.

Methods

Overview

The design of the PBHS has been previously described [9].

Participants

PBHS participants were selected from an online registry in which participants entered basic demographic data so that the initial target cohort could be adequately established [9]. Ultimately, 2502 participants were included. The inclusion criteria for the registry were age ≥18 years, residency in the United States, ability to speak and read English, willingness to provide health information, and ability to interact with certain study activities using a personal smartphone/device. As one of the overarching goals of PBHS is to understand disease progression in the United States, the cohort was designed so that 60% of the enrolled population in each age strata had ~60% higher risk relative to the participants of the same age and sex for atherosclerotic cardiovascular disease, lung cancer, and/or breast or ovarian cancer.

Measurements and Definitions

Study Assessments

PBHS participants underwent a deep phenotyping process, with extensive multimodal assessments during enrollment to measure their health characteristics, including demographics, vitals, laboratory, functional testing, imaging, surveys, and wearable sensor data from the VSW, an investigational medical device used in medical research and clinical care. For this study, baseline characteristics, as listed in Tables S1-S3 in Multimedia Appendix 1, were selected for each participant and were chosen in this exploratory work due to their ubiquity in clinical practice and physiological relevance to RHR.

Resting Heart Rate

Baseline RHR measurements were determined with 2 different techniques: in-clinic ECG RHR and VSW RHR. During the enrollment study site visit, a 12-lead ECG was recorded (Mortara ELI 250/250C), and HR from the computerized interpretation of the ECG was computed as the ECG RHR. An ECG was considered "Excellent" or "Good" when all 12 leads were analyzable, and either no noise/artifact or minimal noise/artifact (respectively) were noted; only ECG readings that met these criteria were considered.

VSW RHR was determined using a proprietary study wrist-wearable device, which was an integral part of the continuous assessments of PBHS. Participants were encouraged to wear it consistently during the entire study duration. The VSW captures biological signals through several sensors, including photoplethysmography (PPG) at 30 Hz and accelerometry at 30 Hz. It also provides several derived metrics

XSL•FO
RenderX

using proprietary algorithms that process these signals. In this study, we use the following derived metrics:

- PPG Interbeat Intervals (IBI), which measure the time interval between PPG-derived heartbeats in milliseconds. The IBIs are calculated at each heartbeat, and each IBI value is also accompanied by a binary quality metric ("good" vs "bad" quality). To determine the quality of IBIs, in this study, we use the "jump distance" metric, which is defined as the following for each sample i: where I_i is the IBI value in milliseconds at sample *i*. When the jump distance is smaller than 100 milliseconds, we label that IBI as having "good" quality and otherwise as having "bad" quality. The reason is that very high jump distance values indicate the presence of artifacts or the failure of the PPG peak detection algorithm. The threshold value of 100 milliseconds was chosen as the optimal value in a trade-off between heart rate error and coverage on an internally collected dataset.
- Actigraphy counts, which estimate the level of physical activity and are calculated every 30 seconds.
- On-wrist states, which indicate whether the VSW was worn or not, are computed every 1 minute and every time the on-wrist state changes.

Since the goal of this analysis was to compare the RHR estimated by VSW to ECG RHR, we used the VSW sensor data captured during the ECG RHR measurement in order to evaluate the performance of the VSW RHR. Thus, we gathered VSW data using a 2-minute measurement window centered at the middle of the ECG acquisition period, as shown in Figure 1A.

Figure 1. Verily Study Watch resting heart rate (VSW RHR) determination during the Project Baseline Health Study (PBHS) procedures. (A) Relative placement in time of the 2-minute VSW data acquisition window against the backdrop of the 10-second ECG acquisition window. (B) Flowchart showing the processing steps to calculate the VSW RHR for each participant. ECG: electrocardiogram; SW: study watch; PPG: photoplethysmography; IBI: interbeat interval; VSW: Verily Study Watch; RHR: resting heart rate.

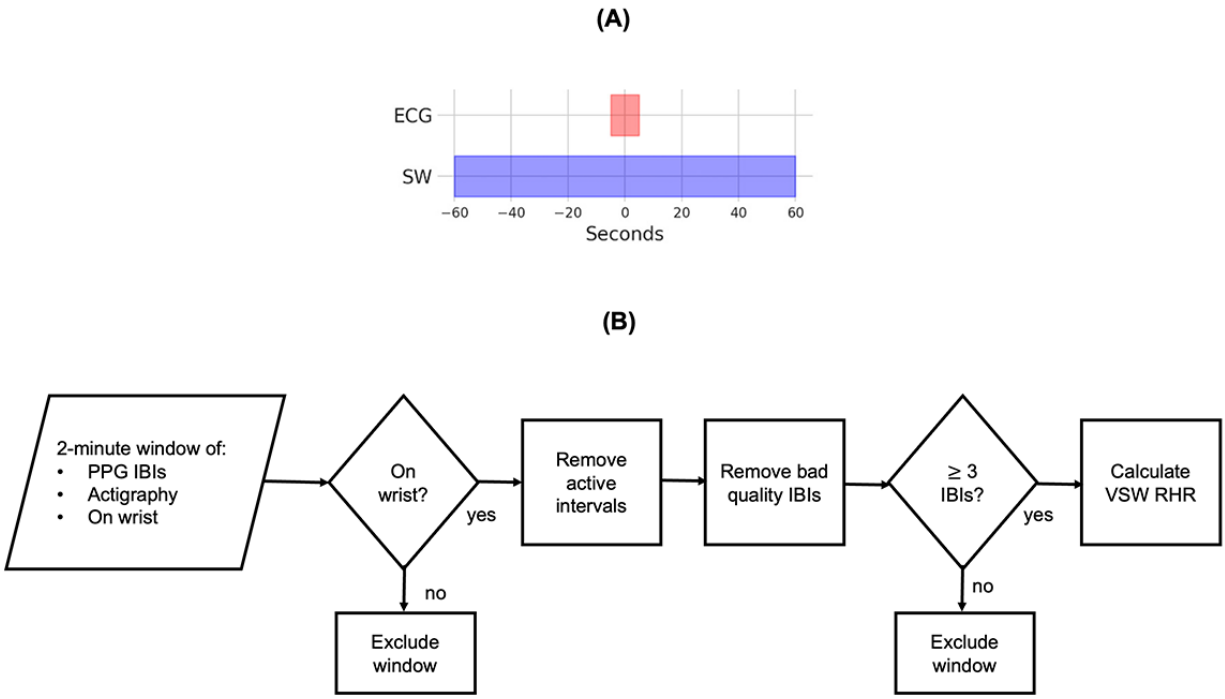

The processing steps to calculate the VSW RHR for each participant are shown in Figure 1B. First, we gathered PPG IBIs, actigraphy counts, and on-wrist states in the 2-minute window mentioned above. Then, we excluded 2-minute windows containing any off-wrist states, and we removed the IBIs associated with active intervals from the window (defined as any 30-second interval with a non-zero actigraphy count value, which we define as "Active"). Those intervals for which there was a zero actigraphy count were defined as "Still." Finally, we removed the "bad" quality IBIs from the 2-minute window. If the remaining number of IBIs was less than 3, we excluded the participant; otherwise, we calculated the VSW RHR from the remaining IBIs as the following:

Where I_i is the *i*th IBI value (milliseconds) in the 2-minute window, and N is the number of IBI values in the window.

Mean Daily Steps in the First 30 Days

Mean daily steps in the first 30 days of the study were calculated for each participant using previously validated step counts captured by the VSW RHR [13,14]. Specifically, daily step

XSL•FO
RenderX

count values were averaged across the 30 days following enrollment, only considering the days during which the participant wore the VSW for at least 10 hours.

Analysis Cohort

For this analysis, the cohort included only participants who fulfilled the following criteria: (1) recorded ECG RHR during the initial onsite visit and (2) concurrent RHR as recorded by the VSW. Additional exclusion criteria were applied during the VSW RHR calculation procedure, as described in the previous section. Inclusion and exclusion criteria are further described in Figure 2.

Figure 2. Analysis cohort flowchart. This flow chart details the creation of the eventual analysis cohort (n=875), originating from the full PBHS cohort. PBHS: Project Baseline Health Study; ECG: electrocardiogram; PPG: photoplethysmography; IBI: interbeat interval; VSW: Verily Study Watch.

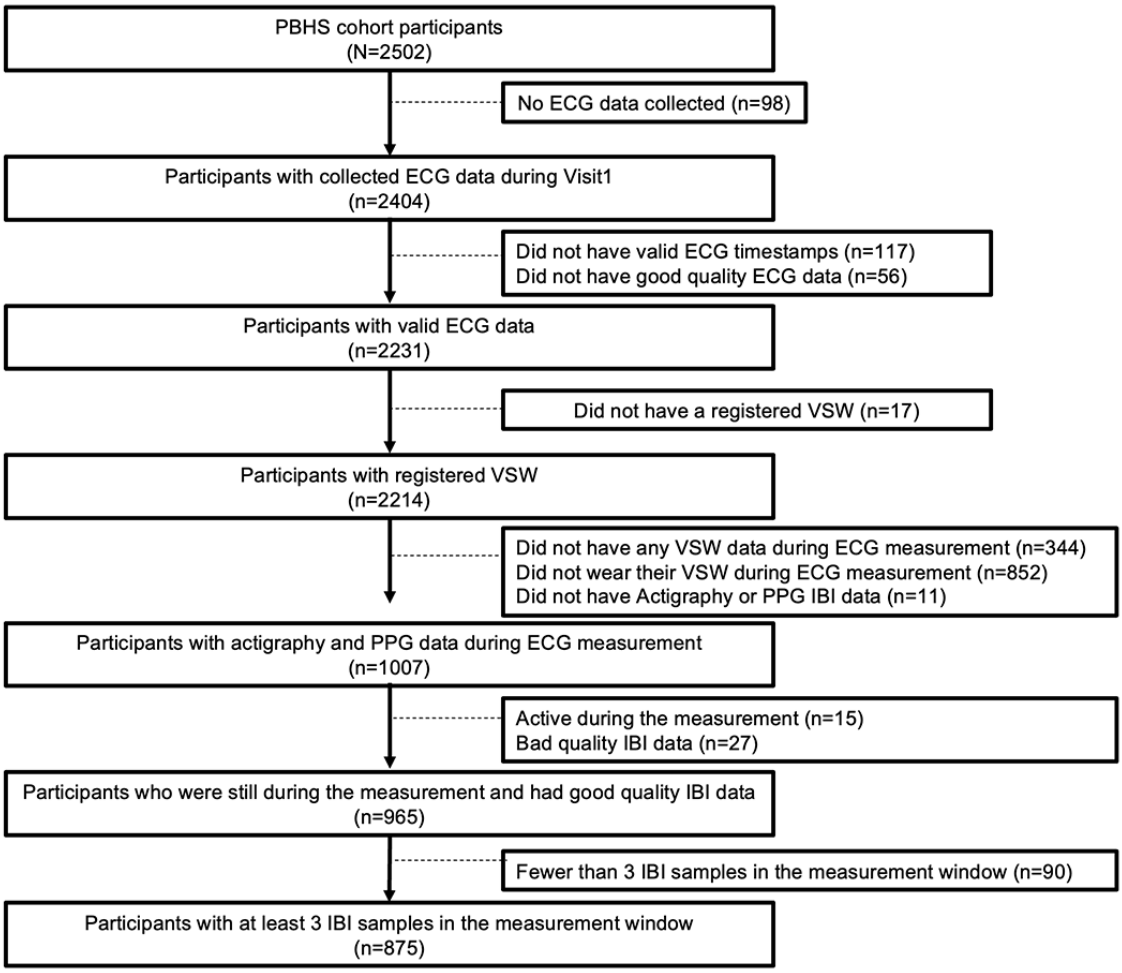

Study Watch Validity Analysis

To evaluate the validity of VSW RHR measurements, we first compared them to ECG RHR using the intraclass correlation (ICC) coefficient across the participants [15]. In addition, we calculated the bias in VSW RHR compared with ECG RHR, which is defined as the mean of the difference between the two measurements (VSW RHR and ECG RHR) across all participants. To determine how bias changed as a function of ECG RHR, we fitted a linear model to predict the measurement difference using ECG RHR as input, and we measured the slope of the fitted model. For both bias and slope values, we calculated the 95% CIs using bootstrapping (1000 bootstraps). We repeated these analyses for each of the male and female subgroups.

Statistical Analysis

Descriptive statistics were calculated for selected demographics and other baseline characteristics. Categorical variables were reported as the number of participants with corresponding percentages, and continuous variables were reported as mean and SD.

For use in statistical testing and regression modeling, categorical variables were translated into a series of 1/0 “dummy” variables, representing each level of each predictor variable versus all other levels as the reference.

Tests for trend were used to evaluate the relationship between each characteristic and ordinal category of VSW RHR,

XSL•FO
RenderX

separately for males and females. Analyses were stratified by sex due to well-established baseline differences in RHR by sex and to document sex-related differences in overall baseline characteristics. Baseline characteristic differences across RHR percentiles were not statistically compared between males and females. In addition, 3 VSW RHR categories were created using sex-specific percentile cut points: 0-25th, 25th-75th, and 75th-100th. The Cochran-Armitage Trend Test was used to evaluate binary variables (including the "dummy" indicator variables created for each level of categorical variables), and Spearman rank correlation was used to evaluate continuous variables.

Associations with VSW RHR among candidate baseline characteristics were identified using multivariable linear regression models. Before modeling, missing data were imputed using 5 rounds of multiple imputation using chained equations methods with predictive mean matching. Box-Cox transformations were used to approximate a normal distribution for continuous variables (laboratory values, vitals, and physical function measures). In addition to observed age, age-squared was added to the list of baseline variables to account for the inverted U-shaped relationship between age and VSW RHR.

All baseline characteristics (more details in Tables S1-S3 in Multimedia Appendix 1) were included as candidate variables in the least absolute shrinkage and selection operator (LASSO) regression models; no characteristics captured beyond the baseline period were included. Baseline characteristics were grouped into domains: (1) demographics and socioeconomic status (SES), (2) medical conditions, (3) vitals and physical function, (4) laboratory assessments, and (5) patient-reported outcomes (PROs); separate models were built for each domain and separately for male and female. Elastic net (ENET) regularization methods were used to fit regression models. In order to address the multiply-imputed data, a stacked objective function (sENET) method was used, with 5-fold cross-validation to penalize and select regression coefficients [16,17]. Due to limitations in computational power, ENET alpha values were restricted to 0.5 or 1, where $\alpha=1$ equates to a LASSO regression. All predictors were standardized before use in modeling as required for LASSO and ENET methods.

Ethical Considerations

The study was approved by a central institutional review board (IRB; Western IRB: approval tracking number 20170163, work order number 1-1506365-1) and the IRB at each of the participating institutions (Stanford University, Duke University, and the California Health and Longevity Institute). The PBHS was registered in ClinicalTrials.gov (identifier NCT03154346).

Informed consent was obtained from all participants enrolled in PBHS. Participants received small compensation for study visit–related time and expenses. This report is based on analyses of deidentified data.

Results

Analysis Cohort Compared With the Overall PBHS Cohort

Using the criteria as described in Figure 2, the analysis cohort consisted of 875 participants: 519 (59%) female and 356 (41%) male. Selected baseline characteristics of the analysis cohort and the PBHS cohort are shown in Table 1.

XSL•FO
RenderX

Table 1. Selected baseline characteristics of the PBHS cohort and the analysis cohort. Overall, the cohorts are similar in baseline characteristics.

	PBHS[a] cohort (N=2502)	Analysis cohort (n=875)
Demographics		
Mean age at enrollment, years (SD)	50 (17.2)	50.9 (16.5)
Female, n (%)	1375 (55)	519 (59.3)
Race, n (%)		
White	1582 (63.2)	575 (65.7)
Black	400 (16)	138 (15.8)
Asian	260 (10.4)	80 (9.1)
Other	259 (10.4)	82 (9.4)
Hispanic, n (%)	290 (11.6)	98 (11.2)
Socioeconomic status, n (%)		
Married	1116 (44.6)	433 (49.5)
Employed	1523 (60.9)	528 (60.3)
Current or former smoker	881 (35.2)	331 (37.8)
Medical conditions, n (%)		
Asthma	371 (14.8)	124 (14.2)
Diabetes, type 2	276 (11.0)	112 (12.8)
Generalized anxiety disorder	327 (13.1)	121 (13.8)
GERD[b]	424 (16.9)	176 (20.1)
Hypertension	675 (27)	262 (29.9)
Hypercholesterolemia	314 (12.5)	118 (13.5)
Major depressive disorder	354 (14.1)	142 (16.2)
Migraines	306 (12.2)	116 (13.3)
Osteoarthritis	477 (19.1)	179 (20.5)
Sleep apnea	245 (9.8)	88 (10.1)
Vitals		
Mean systolic BP[c] (SD)	123.4 (16)	125 (15.5)
Mean diastolic BP (SD)	75.9 (9.9)	77.4 (9.9)
Mean BMI (SD)	28.4 (6.9)	29.4 (7.1)
Physical performance		
Mean 6-minute walk distance, meters (SD)	474.5 (82.7)	475.4 (88.2)
Mean left ventricular ejection fraction (SD)	58.7 (4.2)	58.6 (4.5)
Laboratory findings		
Hemoglobin A_{1c}, mean (SD)	5.7 (1)	5.8 (1.1)
Hemoglobin (g/dL), mean (SD)	14.2 (1.3)	14.1 (1.3)
White blood cell count (thousand/mcL), mean (SD)	6.4 (1.9)	6.6 (1.9)
MDRD[d] (eGFR[e]), mean (SD)	88.3 (20.4)	87.5 (21.1)
C-reactive protein (mg/L), mean (SD)	2.9 (5.9)	3.4 (7.2)
Patient-reported outcomes		
PHQ-9[f] score, mean (SD)	3.7 (4.2)	3.9 (4.3)
GAD-7[g] score, mean (SD)	3.2 (4.1)	3.3 (4.2)

XSL•FO
RenderX

[a]PBHS: Project Baseline Health Study.
[b]GERD: gastro-esophageal reflux disease.
[c]BP: blood pressure.
[d]MDRD: modification of diet in renal disease.
[e]eGFR: estimated glomerular filtration rate.
[f]PHQ-9: Patient Health Questionnaire-9.
[g]GAD-7: Generalized Anxiety Disorder-7 Scale.

Study Watch Validity

The comparison of the RHR by ECG with VSW is shown in Figure 3A. There was excellent reliability between the 2 measures, with an intraclass correlation coefficient of 0.946 (Figure 3A). This reliability remained excellent within each of the male and female subgroups (Figure 3B). An agreement plot between RHR by ECG and VSW of all participants also showed high consistency between the two measures (Figure S1 in Multimedia Appendix 1).

Figure 3. Correlation between baseline ECG-based and Study Watch measured RHR in this analysis cohort (within the PBHS) for (A) all participants and (B) male (left) and female (right) participants separately. Each dot corresponds to one participant. There is excellent overall reliability between ECG RHR and VSW RHR (ICC=0.946) and within each of the male (0.942) and female (0.949) subgroups. BPM: beats per minute; ECG: electrocardiogram; ICC: intraclass correlation coefficient; RHR: resting heart rate; SW: study watch.

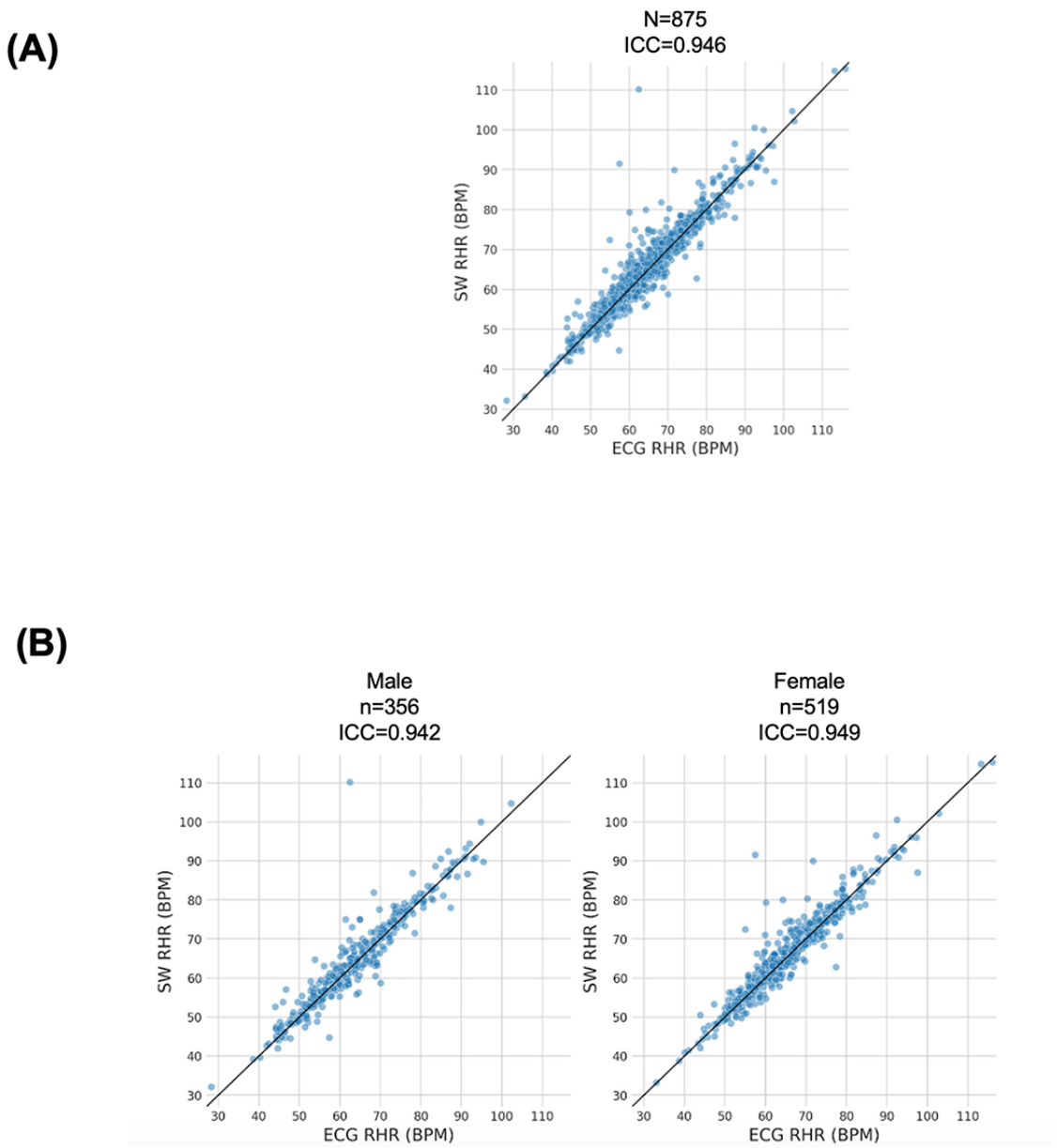

XSL•FO
RenderX

We also estimated the bias in the VSW measurement of RHR compared to the ECG RHR. The overall bias was 0.76 BPM (95% CI 0.52-1.00), which indicated a small but significant positive bias, meaning that VSW was slightly overestimating the RHR when compared with ECG-based RHR. We had a similar result in male and female subgroups, with a bias of 0.70 (95% CI 0.29-1.14) and 0.80 (95% CI 0.51-1.13), respectively.

Finally, we evaluated how bias changed as a function of ECG RHR using the slope of a fitted linear model (Figure S2A in Multimedia Appendix 1). The resulting slope was –0.029 (95% CI –0.047 to –0.010), which indicated a small but significant negative slope, meaning that for the ECG RHR values on the lower end, VSW overestimated the RHR, while on the higher end, it underestimated the RHR. We had a similar result for each of the male and female subgroups, with a slope of –0.030 (95% CI –0.057 to –0.003) and –0.028 (95% CI –0.056 to –0.003), respectively (Figure S2B in Multimedia Appendix 1).

Analysis Cohort Baseline Characteristics by Sex and Resting Heart Rate (Age-Adjusted)

The VSW RHR as a function of age and sex is shown in Figure 4, with both curves demonstrating the expected upside-down U-shaped relationship [10]. The mean and SD of VSW RHR was 66.6 (SD 11.2) beats per minute (bpm) for female participants and 64.4 (SD 12.3) bpm for male participants. For ECG RHR, the mean and SD were 65.8 (SD 10.9) bpm for females and 63.7 (SD 12.0) bpm for males.

Figure 4. Baseline study watch resting heart rate by age and sex in this analysis cohort (within the PBHS). A U-shaped curve was observed for both female and male participants when Verily Study Watch resting heart rate (VSW RHR) was plotted against age. The lines show fitted quadratic models for female and male data separately. The shaded areas show the 95% CIs of the models. BPM: beats per minute; RHR: resting heart rate.

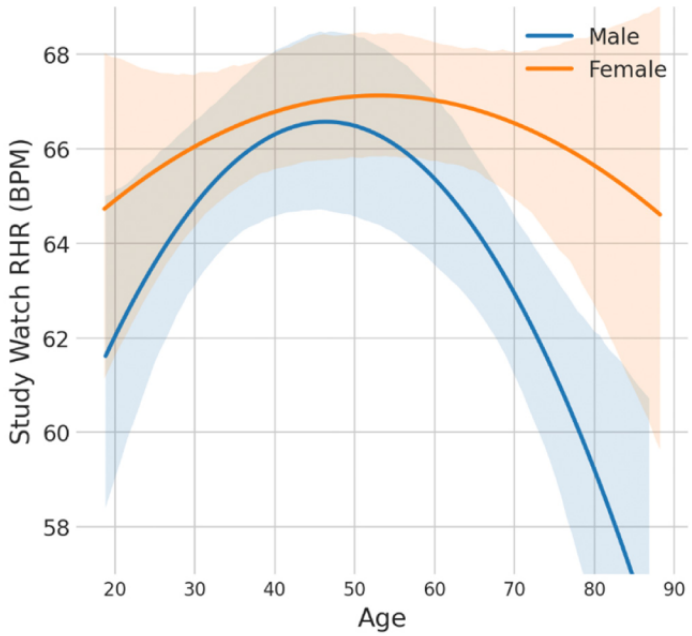

The study cohort was then separated by sex and stratified by VSW RHR to assess for trends of selected baseline characteristics: demographics and SES (Tables 2 and 3; vitals, physical function, and laboratory assessments (Tables 4 and 5); and medical conditions and participant-reported outcomes (PROs; Tables 6 and 7). Full variable comparison lists are included in Tables S1-S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Analysis cohort, female participants (within the PBHS): Selected demographics and socioeconomic status at baseline, stratified by Verily Study Watch resting heart rate (VSW RHR) percentile.

	VSW[a] RHR[b] percentile[c]			
	0-25th (n=130)	25-75th (n=259)	75-100th (n=130)	*P* value
Mean age, years (SD)	48.7 (16.3)	51.1 (16.1)	49.3 (15.8)	.76
Race, n (%)				
White	93 (71.5)	164 (63.3)	83 (63.8)	.19
Black	18 (13.8)	44 (17)	26 (20)	.19
Asian	10 (7.7)	27 (10.4)	6 (4.6)	.37
Other (NHPI[d], AIAN[e], Other)	9 (6.9)	24 (9.3)	15 (11.5)	.20
Hispanic ethnicity, n (%)	14 (10.8)	40 (15.4)	17 (13.1)	.59
Education level, n (%)				
High school or less	11 (9.7)	21 (9.5)	24 (20.9)	.01
Any college	65 (57.5)	143 (64.7)	66 (57.4)	.98
Graduate degree or higher	37 (32.7)	57 (25.8)	25 (21.7)	.06
Income (US $), n (%)				
<100,000	58 (51.3)	123 (55.7)	82 (71.3)	.002
>100,000	46 (40.7)	82 (37.1)	25 (21.7)	.003
Marital status, n (%)				
Married	63 (55.8)	131 (59.3)	50 (43.5)	.06
Divorced or separated	21 (18.6)	36 (16.3)	22 (19.1)	.91
Single	24 (21.2)	43 (19.5)	33 (28.7)	.17
Widowed	3 (2.7)	7 (3.2)	10 (8.7)	.03
Employment status, n (%)				
Employed or homemaker	92 (74.2)	164 (68.3)	69 (57.5)	.01
Unemployed	6 (4.8)	17 (7.1)	24 (20)	<.001
Retired	19 (15.3)	52 (21.7)	23 (19.2)	.44
Student	5 (4)	4 (1.7)	2 (1.7)	.21
Insured (health insurance, yes), n (%)	109 (96.5)	202 (91.4)	101 (89.4)	.048
Smoking status, n (%)				
Current smoker	12 (9.2)	41 (15.8)	31 (23.8)	.001
Former smoker	27 (20.8)	54 (20.8)	24 (18.5)	.64
Never smoker	91 (70)	164 (63.3)	75 (57.7)	.04

[a]VSW: Verily study watch.

[b]RHR: resting heart rate.

[c]Percentile cut points for females: 25th=59.38 bpm; 75th=73.66 bpm. Shading for significant observations. To generate *P* values for tests for trend, the Cochran-Armitage was used to evaluate binary variables, including 'dummy' indicator variables created for each level of categorical variables, and Spearman Rank Correlation was used to evaluate continuous variables.

[d]NHPI: Native Hawaiian, and Pacific Islander.

[e]AIAN: American Indians and Alaska Natives.

XSL•FO
RenderX

Table 3. Analysis cohort, male participants (within the PBHS): Selected demographics and socioeconomic status at baseline, stratified by Verily Study Watch resting heart rate (VSW RHR) percentile.

	VSW[a] RHR[b] percentile[c]			
	0-25th (n=89)	25-75th (n=178)	75-100th (n=89)	*P* value
Mean age, years (SD)	55.5 (16.6)	50.7 (18.5)	51.6 (14.3)	.11
Race, n (%)				
White	71 (79.8)	107 (60.1)	57 (64.0)	.03
Black	11 (12.4)	24 (13.5)	15 (16.9)	.39
Asian	4 (4.5)	24 (13.5)	9 (10.1)	.22
Other (NHPI[d], AIAN[e], Other)	3 (3.4)	23 (12.9)	8 (9.0)	.20
Hispanic ethnicity, n (%)	6 (6.7)	12 (6.7)	9 (10.1)	.40
Education level, n (%)				
High school or less	7 (9.7)	16 (10.7)	12 (15.4)	.28
Any college	34 (47.2)	74 (49.7)	48 (61.5)	.08
Graduate degree or higher	31 (43.1)	59 (39.6)	18 (23.1)	.01
Income (US $), n (%)				
<100,000	34 (47.2)	80 (53.7)	41 (52.6)	.53
>100,000	35 (48.6)	64 (43.0)	28 (35.9)	.12
Marital status, n (%)				
Married	57 (79.2)	86 (57.7)	46 (59.0)	.01
Divorced or separated	5 (6.9)	11 (7.4)	7 (9.0)	.64
Single	9 (12.5)	49 (32.9)	22 (28.2)	.04
Widowed	0 (0.0)	1 (0.7)	2 (2.6)	.28
Employment status, n (%)				
Employed or homemaker	43 (53.8)	103 (65.6)	57 (67.1)	.08
Unemployed	9 (11.2)	9 (5.7)	11 (12.9)	.67
Retired	28 (35.0)	40 (25.5)	16 (18.8)	.02
Student	0 (0.0)	4 (2.5)	1 (1.2)	.80
Insured (health insurance, yes), n (%)	69 (95.8)	137 (91.9)	67 (85.9)	.03
Smoking status, n (%)				
Current smoker	16 (18.0)	22 (12.4)	17 (19.1)	.84
Former smoker	21 (23.6)	41 (23.0)	25 (28.1)	.49
Never smoker	52 (58.4)	115 (64.6)	47 (52.8)	.44

[a]VSW: Verily study watch.

[b]RHR: resting heart rate.

[c]Percentile cutpoints for males: 25th=55.50 bpm; 75th=72.25 bpm. Shading for significant observations. To generate *P* values for tests for trend, the Cochran-Armitage was used to evaluate binary variables, including 'dummy' indicator variables created for each level of categorical variables, and Spearman Rank Correlation was used to evaluate continuous variables.

[d]NHPI: Native Hawaiian, and Pacific Islander.

[e]AIAN: American Indians and Alaska Natives.

XSL•FO
RenderX

Table 4. Analysis cohort, female participants (within the PBHS): Selected vitals, physical function, and labs at baseline, stratified by Verily Study Watch resting heart rate (VSW RHR) percentile.

	VSW[a] RHR[b] percentile[c]			
	0-25th (n=130)	25-75th (n=259)	75-100th (n=130)	*P* value
Vitals				
Systolic blood pressure, mean (SD)	119.5 (15.2)	122.6 (15.5)	126.5 (15.6)	<.001
Diastolic blood pressure, mean (SD)	73.2 (8.3)	76.0 (9.2)	81.1 (10.0)	<.001
Waist circumference (cm), mean (SD)	85.3 (14.4)	89.8 (15.9)	98.5 (18.0)	<.001
BMI, mean (SD)	27.1 (6.4)	28.5 (6.7)	32.9 (8.4)	<.001
Physical function				
6-minute walk distance (m), mean (SD)	498.1 (82.7)	469.3 (81.9)	433.8 (93.0)	<.001
10-meter walk speed (seconds), mean (SD)	2.0 (0.6)	1.9 (0.4)	1.8 (0.5)	.001
Handgrip, mean (SD)	28.9 (6.9)	28.1 (6.9)	27.4 (7.0)	.23
Leg balance time (seconds), mean (SD)	44.3 (20.6)	39.8 (22.1)	37.8 (23.1)	.02
Sit-rise score, mean (SD)	7.5 (2.3)	6.9 (2.5)	7.0 (2.4)	.11
30-second chair stand, mean (SD)	14.8 (4.7)	13.9 (5.0)	12.9 (4.3)	.002
Ejection fraction at rest (%), mean (SD)	59.0 (3.6)	59.4 (4.3)	58.5 (5.4)	.33
Coronary calcium score, mean (SD)	66.6 (214.3)	60.9 (250.0)	76.6 (249.1)	.03
Ankle brachial index abnormal, n (%)	4 (3.1)	10 (3.9)	3 (2.5)	.79
FEV_1/FVC[d], mean (SD)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	.25
Daily steps in the first 30 days, mean (SD)	8360 (2990)	8040 (3187)	6865 (3243)	.0001
Laboratory values				
Hemoglobin (g/dL), mean (SD)	13.5 (1.0)	13.5 (1.2)	13.7 (1.2)	.10
Serum creatinine (mg/dL), mean (SD)	0.8 (0.1)	0.8 (0.2)	0.8 (0.2)	.19
HDL[e] (mg/dL), mean (SD)	66.4 (18.4)	64.3 (20.5)	57.2 (14.4)	<.001
LDL[f] (mg/dL), mean (SD)	96.0 (29.1)	105.5 (33.9)	105.8 (30.4)	.02
HbA_{1c} (%), mean (SD)	5.4 (0.7)	5.6 (0.8)	6.0 (1.5)	<.001
C-reactive protein (mg/L), mean (SD)	2.3 (4.1)	3.3 (5.1)	5.6 (8.2)	<.001
Blood glucose (mg/dL), mean (SD)	88.7 (19.0)	94.1 (27.2)	108.9 (54.6)	<.001
Hematocrit (%), mean (SD)	41.2 (2.9)	41.2 (3.4)	42.0 (3.5)	.047
Platelet count (per µL), mean (SD)	249,836 (56,779)	259,269 (61,124)	278,790 (61,889)	<.001
WBC[g] count (thousand/µL), mean (SD)	6.2 (1.6)	6.5 (1.8)	7.5 (2.2)	<.001
Sodium (mEq/L), mean (SD)	139.0 (1.8)	138.8 (2.1)	138.6 (2.2)	.15
GFR MDRD[h] (mL/min), mean (SD)	86.2 (19.2)	87.5 (20.1)	91.0 (24.9)	.19
TSH[i] (mIU/L), mean (SD)	1.5 (0.8)	1.6 (1.2)	1.6 (0.9)	.35

[a]VSW: Verily study watch.

[b]RHR: resting heart rate.

[c]Percentile cutpoints for females: 25th=59.38 bpm; 75th=73.66 bpm. Shading for significant observations. To generate *P* values for tests for trend, the Cochran-Armitage was used to evaluate binary variables, including 'dummy' indicator variables created for each level of categorical variables, and Spearman Rank Correlation was used to evaluate continuous variables.

[d]FEV1/FVC=forced expiratory volume in 1 s /forced vital capacity.

[e]HDL: high-density lipoprotein.

[f]LDL: low-density lipoprotein.

[g]WBC: white blood cell.

[h]GFR MDRD: glomerular filtration rate, modification of diet in renal disease.

XSL•FO
RenderX

[i]TSH: thyroid-stimulating hormone.

Similar trends were seen in female and male participants. For instance, from an SES standpoint, those with higher baseline VSW RHR were more likely to have lower household income, less likely to be married, less likely to have health care insurance, and more likely to be smokers.

Medical conditions such as major depressive disorder, type 2 diabetes mellitus, hypertension, and sleep apnea were also more common in those with higher VSW RHR.

Participants with higher VSW RHR tended to have higher systolic and diastolic blood pressures, BMI, and waist circumference.

In terms of laboratory assessments, those with higher VSW RHR tended to have hemoglobin A_{1c} %, C-reactive protein levels, and white blood cell counts.

Participants with higher VSW RHR had shorter 6-minute walk distances and fewer mean daily steps, as recorded by the VSW.

From a PRO standpoint, participants with higher VSW RHR had higher Patient Health Questionnaire-9 (PHQ-9) scores and World Health Organization Disability Assessment Schedule (WHODAS 2.0) scores.

XSL•FO
RenderX

Table 5. Analysis cohort, male participants (within the PBHS): Selected vitals, physical function, and labs at baseline, stratified by Verily Study Watch resting heart rate (VSW RHR) percentile.

	VSW[a] RHR[b] percentile[c]			
	0-25th (n=89)	25-75th (n=178)	75-100th (n=89)	*P* value
Vitals				
Systolic blood pressure, mean (SD)	127.7 (16.2)	128.1 (14.1)	129.3 (14.3)	.37
Diastolic blood pressure, mean (SD)	76.1 (10.4)	78.3 (9.7)	81.2 (10.1)	.002
Waist circumference (cm), mean (SD)	95.5 (12.3)	98.7 (15.7)	108.4 (18.5)	<.001
BMI, mean (SD)	27.6 (4.5)	29.3 (5.7)	32.5 (8.5)	<.001
Physical function				
6-minute walk distance (m), (SD)	501.5 (83.1)	490.2 (89.8)	465.2 (84.6)	.002
10-meter walk speed, mean (SD)	2.1 (0.6)	2.1 (0.6)	1.9 (0.5)	.021
Handgrip, mean (SD)	46.0 (9.4)	44.5 (10.6)	42.4 (10.3)	.088
Leg balance time (seconds), mean (SD)	38.4 (22.8)	37.7 (23.1)	30.6 (22.6)	.01
Sit-rise score, mean (SD)	7.5 (2.1)	7.0 (2.3)	6.7 (2.3)	.02
30-second chair stand, mean (SD)	15.4 (5.3)	14.9 (5.5)	13.4 (4.4)	.002
Ejection fraction at rest (%), mean (SD)	58.2 (3.7)	57.7 (4.7)	58.7 (4.4)	.76
Coronary calcium score, mean (SD)	361.8 (1012.1)	254.6 (653.9)	209.5 (632.9)	.10
Ankle brachial index abnormal, n (%)	3 (3.5%)	7 (3.9%)	2 (2.3%)	.65
FEV_1/FVC[d], mean (SD)	0.7 (0.1)	0.8 (0.1)	0.8 (0.1)	.003
Daily steps in the first 30 days, mean (SD)	8970 (3994)	8565 (3537)	7869 (4120)	.07
Laboratory findings				
Hemoglobin (g/dL), mean (SD)	14.8 (0.9)	14.9 (1.0)	15.1 (1.1)	.02
Serum creatinine (mg/dL), mean (SD)	1.0 (0.2)	1.0 (0.3)	1.1 (0.5)	.95
HDL[e] (mg/dL), mean (SD)	54.3 (17.4)	48.3 (15.4)	43.9 (12.8)	<.001
LDL[f] (mg/dL), mean (SD)	93.3 (36.3)	95.0 (33.1)	101.6 (38.5)	.22
HbA_{1c} (%), mean (SD)	5.5 (0.5)	5.7 (1.0)	6.5 (1.9)	.003
C-reactive protein (mg/L), mean (SD)	3.0 (14.4)	2.5 (4.5)	4.6 (7.3)	<.001
Blood glucose (mg/dL), mean (SD)	92.2 (12.4)	102.0 (35.9)	130.1 (72.8)	<.001
Hematocrit (%), mean (SD)	44.7 (2.9)	45.2 (3.1)	45.8 (3.2)	.01
Platelets (per µL), mean (SD)	212,529 (48,991)	228,567 (53,411)	248,264 (68,746)	<.001
WBC[g] count (thousand/µL), mean (SD)	6.2 (1.9)	6.2 (1.5)	6.9 (1.9)	.001
Sodium (mEq/L), mean (SD)	139.4 (1.7)	138.9 (2.1)	138.6 (2.4)	.02
GFR MDRD[h] (mL/min), mean (SD)	84.3 (16.5)	88.2 (20.7)	86.6 (25.2)	.38
TSH[i] (mIU/L), mean (SD)	1.8 (1.0)	1.9 (1.1)	1.7 (0.9)	.71

[a]VSW: Verily study watch.

[b]RHR: resting heart rate.

[c]Percentile cutpoints for males: 25th=55.50 bpm; 75th=72.25 bpm. Shading for significant observations. To generate *P* values for tests for trend, the Cochran-Armitage was used to evaluate binary variables, including 'dummy' indicator variables created for each level of categorical variables, and Spearman Rank Correlation was used to evaluate continuous variables.

[e]HDL: high-density lipoprotein.

[f]LDL: low-density lipoprotein.

[g]WBC: white blood cell.

[h]GFR MDRD: glomerular filtration rate, modification of diet in renal disease.

[i]TSH: thyroid-stimulating hormone.

XSL•FO
RenderX

Table 6. Analysis cohort, female participants (within the PBHS): Selected medical conditions and participant-reported outcomes (PROs) at baseline, stratified by Verily Study Watch resting heart rate (VSW RHR) percentile.

	VSW[a] RHR[b] percentile[c]			
	0-25th (n=130)	25-75th (n=259)	75-100th (n=130)	*P* value
Medical history, n (%)				
Asthma	16 (12.3)	35 (13.5)	23 (17.7)	.21
Cataracts	12 (9.2)	38 (14.7)	22 (16.9)	.07
Colon polyps	7 (5.4)	26 (10)	10 (7.7)	.50
Major depressive disorder	15 (11.5)	43 (16.6)	29 (22.3)	.02
Diabetes type 2	6 (4.6)	27 (10.4)	26 (20)	<.001
GERD[d]	20 (15.4)	42 (16.2)	36 (27.7)	.01
Hypertension	27 (20.8)	70 (27)	40 (30.8)	.07
Hypercholesterolemia	14 (10.8)	40 (15.4)	13 (10)	.85
Osteoarthritis	25 (19.2)	49 (18.9)	32 (24.6)	.28
Sleep apnea	6 (4.6)	16 (6.2)	11 (8.5)	.20
PRO scores [e], mean (SD)				
Sheehan Disability Scale	2.9 (4.5)	2.7 (4.8)	5.0 (7.6)	.10
PHQ-9[f]	3.4 (3.6)	3.6 (4)	5.4 (4.8)	<.001
GAD-7[g]	3.2 (3.9)	3.4 (4.1)	4.1 (4.9)	.28
WHODAS[h] 2.0	2.2 (3.3)	3.0 (4.4)	5.0 (6.7)	<.001
BRFSS ACE[i]	2.2 (2.2)	2.4 (2.3)	2.7 (2.6)	.24
PROMIS[j] pain intensity	6.0 (2.3)	6.1 (2.3)	7.0 (2.8)	.02
PROMIS pain interference	10.2 (5.1)	10.5 (5.2)	12.2 (6.3)	.03
PANAS[k] positive affect	34.7 (6.4)	34.8 (7.0)	33.0 (7.3)	.08
PANAS negative affect	15.6 (6.7)	15.5 (6.5)	15.0 (6.1)	.45
Subjective happiness	21.6 (4.8)	21.7 (4.8)	20.9 (4.3)	.13
Satisfaction with life	26.1 (6.6)	25.9 (6.3)	24.2 (7.4)	.04
Perceived social support	70.1 (11.8)	66.8 (15.2)	66.9 (13.9)	.08
AUDIT-C[l]	2.0 (1.5)	2.0 (1.9)	1.9 (1.8)	.35

[a]VSW: Verily study watch.

[b]RHR: resting heart rate.

[c]Percentile cutpoints for females: 25th=59.38 bpm; 75th=73.66 bpm. Shading for significant observations. To generate *P*-values for tests for trend, the Cochran-Armitage was used to evaluate binary variables, including 'dummy' indicator variables created for each level of categorical variables, and Spearman Rank Correlation was used to evaluate continuous variables.

[d]GERD: gastroesophageal reflux disease.

[e]PROs: patient-reported outcomes.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]GAD-7: Generalized Anxiety Disorder-7 Scale.

[h]WHODAS: World Health Organization Disability Assessment Schedule.

[i]BRFSS ACE: Behavioral Risk Factor Surveillance System Adverse Childhood Experience.

[j]PROMIS: Patient-Reported Outcomes Measurement Information System.

[k]PANAS: Positive and Negative Affect Schedule.

[l]AUDIT-C: Alcohol Use Disorders Identification Test-Concise.

XSL•FO
RenderX

Table 7. Analysis cohort, male participants (within the PBHS): Selected medical conditions and participant-reported outcomes (PROs) at baseline, stratified by Verily Study Watch resting heart rate (VSW RHR) percentile.

	VSW[a] RHR[b] percentile[c]			
	0-25th (n=89)	25-75th (n=178)	75-100th (n=89)	*P* value
Medical history, n (%)				
Asthma	12 (13.5)	22 (12.4)	16 (18)	.39
Cataracts	14 (15.7)	28 (15.7)	5 (5.6)	.046
Colon polyps	14 (15.7)	23 (12.9)	6 (6.7)	.07
Major depressive disorder	8 (9)	28 (15.7)	19 (21.3)	.02
Diabetes type 2	3 (3.4)	23 (12.9)	27 (30.3)	<.001
GERD[d]	21 (23.6)	37 (20.8)	20 (22.5)	.86
Hypertension	29 (32.6)	61 (34.3)	35 (39.3)	.35
Hypercholesterolemia	15 (16.9)	27 (15.2)	9 (10.1)	.20
Osteoarthritis	21 (23.6)	33 (18.5)	19 (21.3)	.71
Sleep apnea	9 (10.1)	29 (16.3)	17 (19.1)	.10
PRO scores [e], mean (SD)				
Sheehan Disability Scale	3.3 (6.6)	3.1 (5.6)	4.2 (5.7)	.06
PHQ-9[f]	3.2 (4.1)	3.8 (4.4)	4.6 (4.4)	.01
GAD-7[g]	2.1 (3.2)	2.9 (4)	3.9 (4.8)	.02
WHODAS 2.0[h]	2.2 (4.5)	3.2 (5.2)	4.4 (5.4)	<.001
BRFSS ACE[i]	1.6 (1.9)	1.9 (2.2)	2.7 (2.6)	.01
PROMIS[j] pain intensity	6.1 (2.7)	6.4 (2.3)	6.5 (2.5)	.18
PROMIS pain interference	10.0 (5.7)	10.6 (4.7)	12.4 (6.1)	.003
PANAS[k] positive affect score	35.5 (7.7)	33.7 (7.6)	32.1 (8.1)	.01
PANAS negative affect	14.9 (5.7)	15.2 (5.9)	15.1 (5.8)	.93
Subjective happiness	22.0 (4.3)	20.7 (4.8)	18.9 (4.6)	<.001
Satisfaction with life	26.1 (6.6)	24.7 (6.7)	22.5 (6.2)	<.001
Perceived social support	65.8 (13.8)	61.6 (16.7)	59.3 (14.5)	.003
AUDIT-C[l]	2.3 (1.7)	2.1 (1.8)	1.8 (1.8)	.02

[a]VSW: Verily study watch.

[b]RHR: resting heart rate.

[c]Percentile cutpoints for males: 25th=55.50 bpm; 75th=72.25 bpm. Shading for significant observations. To generate *P* values for tests for trend, the Cochran-Armitage was used to evaluate binary variables, including 'dummy' indicator variables created for each level of categorical variables, and Spearman Rank Correlation was used to evaluate continuous variables.

[d]GERD: gastroesophageal reflux disease.

[e]PROs: patient-reported outcomes.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]GAD-7: Generalized Anxiety Disorder-7 Scale.

[h]WHODAS: World Health Organization Disability Assessment Schedule.

[i]BRFSS ACE: Behavioral Risk Factor Surveillance System Adverse Childhood Experience.

[j]PROMIS: Patient-Reported Outcomes Measurement Information System.

[k]PANAS: Positive and Negative Affect Schedule.

[l]AUDIT-C: Alcohol Use Disorders Identification Test-Concise.

XSL•FO
RenderX

Associations With VSW RHRs by Domain

The results of the sex-stratified sENET regression models are presented in Figure 5. Penalized regression coefficients reflect the relative strength and direction of each association based on standardized predictors. Within each domain of baseline characteristics (demographics and SES, medical conditions, vitals, physical function, laboratory assessments, and PROs), analyses showed that different characteristics were associated with VSW RHR in female and male participants. For instance, in the demographics and SES domain, unemployment had the highest association with VSW RHR in females, whereas lack of health insurance had the highest association in male participants. This was the case in the medical conditions, laboratory assessments, and PRO domains as well. For the vitals and physical function domain, diastolic blood pressure was the most associated characteristic with VSW RHR for both sexes.

XSL•FO
RenderX

Figure 5. LASSO (least absolute shrinkage and selection operator) regressions. Regression analyses performed in this cohort (within the PBHS) suggest that, within each domain, different baseline characteristics are most associated with resting heart rate. All continuous measurements, ie, laboratory, vital, and physical function variables, were transformed by the Box-Cox method before analysis. All variables were standardized as required for penalized regression methodology. Within each domain of baseline characteristics, there were some characteristics that were more associated with RHR in female participants and others that were more associated with male participants. A higher-resolution version of this image is available in Multimedia Appendix 2.

XSL•FO
RenderX

Discussion

Principal Findings

These analyses from a large, deeply phenotyped population show (1) strong agreement between ECG-determined RHR and a proprietary VSW-determined RHR, (2) significant trends of VSW RHR with clinically important baseline characteristics, and (3) clinical baseline characteristics highly associated with VSW RHR. These findings demonstrate that, in a relatively heterogeneous cohort of participants, RHR can be measured easily and accurately using a wearable device and may be used in light of strong associations with clinically relevant baseline characteristics.

In the last decade, the use of consumer wearables with the ability to detect HR and arrhythmias such as atrial fibrillation has become increasingly available [5]. Despite their high accuracy in measuring HR at rest and the ubiquity of these devices in modern life, an in-clinic ECG is still the gold standard method to determine RHR [18]. In clinical research, a variety of methodologies are used depending on the availability of data and clinical feasibility [10,11,19,20]. In this study, we investigated the viability of the VSW in determining RHR by comparing it with RHR determined by ECG. Using PPG data combined with actigraphy data, we isolated periods of time when the participant wearing the VSW was not in motion during the time of ECG recording, thus allowing us to estimate RHR values from VSW data. With this method, we demonstrated that there is excellent agreement between RHR determined by ECG and VSW, suggesting that the VSW is capable of determining a reliable RHR. In a world where telehealth is increasingly used, reliable wearable device-based data such as this may be useful to clinicians, providing them with clinical information that would otherwise be more cumbersome to obtain [21].

While most studies in the past have focused on analyzing the relationship of RHR with objective, laboratory-based measurements, we also extensively evaluated the relationship of RHR with participants' well-being and quality of life, including psychosocial and socioeconomic aspects. In the univariate analyses, we demonstrated that participants who had higher education were married, had health care insurance, and had lower PHQ-9 scores were more likely to have a lower VSW RHR. Furthermore, we found similar and significant associations in our regression models when stratified by sex: lack of health care insurance, psychiatric conditions (major depressive disorder and generalized anxiety disorder), and higher WHO-DAS 2.0 scores were significantly associated with higher HR. These findings are consistent with previous studies suggesting that more difficult SES and psychosocial circumstances were associated with higher chronic stress and higher HR [22-28]. However, in our analyses, we also found that there were differences by sex in which baseline characteristics were most associated with RHR. For instance, within the demographics and SES domain, unemployment was most significantly associated with higher RHR for female participants but lack of health care insurance was the most significantly associated with higher RHR for male participants. Similarly, within the PRO domain, a higher WHODAS 2.0 score was most associated with higher RHR for female participants, but the Behavioral Risk Factor Surveillance System Adverse Childhood Experience score was most associated with higher RHR for male participants. This may be the result of a multitude of factors, including physiological differences between the 2 sexes, societal influences, and diverse cultural and personal experiences that could impact HR [29-32]. In the laboratory setting, it has been demonstrated that there are sex differences in HR responses to physical and mental stressors [33-35]. A recent study using a contemporary wearable device investigated the effects of occupational stressors in the real world and found that female participants, compared with male participants, had a higher maximum HR and greater changes in HR when confronted with a moderate stressor during a work shift in a retail store [36]. Future studies will be needed to elucidate the relationships and mechanisms underlying how different clinical characteristics affect RHR in females and males.

We observed significant trends of VSW RHR with objective clinical measurements in both our univariate and regression analyses. Higher VSW RHR was associated with higher blood pressure, BMI, and waist circumference, all previously established in the literature [37,38]. Laboratory findings of higher C-reactive protein and platelet counts in those participants with higher SW RHR were also consistent with the literature [39]. Analyses of physical function showed significant trends with VSW RHR. Lower VSW RHR was significantly correlated with a higher 6-minute walk distance, an important clinical surrogate for fitness [40]. It has been demonstrated previously that HR profiles determined by wrist-worn devices can predict 6-minute walk distances in patients with mitral or aortic valve disease [11]. Another more commonplace measure of physical activity and fitness is step count, a measure that has been associated with mortality [41]. We observed that participants with lower VSW RHR had significantly higher step counts, consistent with previous studies demonstrating a negative relationship between VSW RHR and physical fitness [28,42,43]. Though causality cannot be determined from these analyses, the relationship between VSW RHR and step count is of high interest to clinicians and patients alike, given step count and other surrogates of physical fitness are integral elements of wearable devices that are often promoted as a method of remote monitoring. Interestingly, the relationship demonstrated in our study was of VSW RHR and future step count, suggesting that even a single RHR measurement could be indicative of a person's future physical activity and, therefore, may identify a population with higher RHR for targeted interventions aimed to improve physical fitness. Future studies will need to longitudinally track both RHR and physical activity levels to determine if their long-term trends are indeed correlated.

There are several limitations to our analysis. Our cohort may have a slight healthy user bias, given it was derived from the PBHS registry, and this potential self-selection bias may have had an impact on some of our findings, such as the differences across sexes. The analysis cohort was also more limited in size than expected, primarily due to a lack of procedural consistency (wearing the VSW at the time of ECG recording) during the participant enrollment visit, resulting in a loss of ~50% of participants from the DPC cohort. It is possible that this can

XSL•FO
RenderX

contribute further to a healthy user bias, as those who are more proactive in wearing the VSW may also have been healthier. Future studies will need to ensure more rigid protocols to ensure less variability due to procedural issues. In this study, hard clinical outcomes such as mortality and hospitalizations were not assessed but would be highly valuable for future studies, particularly those that evaluate not only associations of RHR with clinical outcomes but also of "free-living" HR with clinical outcomes. Other studies have examined the validity of using wearable devices to measure HR under free-living conditions, which is currently under investigation in the PBHS [44,45]. Finally, there was a positive bias of 0.76 BPM (95% CI 0.52-1.00) in VSW RHR measurements compared with the reference ECG RHR. This bias changed significantly as a function of ECG RHR, with a negative slope of –0.029 (95% CI –0.047 to –0.010). The most likely cause for the bias is the relative noisiness of PPG signals (measured by SW) compared with ECG, which occasionally results in the detection of false beats. While both bias and slope values are statistically significant, they are unlikely to be clinically meaningful.

Conclusion

In conclusion, VSW RHR correlates strongly with RHR obtained using resting ECG. VSW RHR has significant trends with important clinical characteristics that closely mirror those already established in the literature. Further investigations will be needed to inform clinicians and patients alike on how to use wearable technologies that perform noninvasive measurements—not only of RHR—in conjunction with other clinical measurements to potentially detect disease or enhance their shared decision-making process for behavioral change.

Acknowledgments

The Project Baseline Health Study and this analysis were funded by Verily Life Sciences, South San Francisco, California. Verily Life Sciences is the funding source for the PBHS and is responsible for data collection. The authors were fully responsible for the data analysis and interpretation presented herein and the writing of this article. The following individuals, SAS, SS, and MKC, had access to the raw data. The authors had access to the full dataset for the study and reviewed and approved the final manuscript for submission. The authors wish to thank the Project Baseline Health Study investigators, study sites, and participants.

Data Availability

The deidentified PBHS data corresponding to this study are available upon request for the purpose of examining its reproducibility. Interested investigators should direct requests to jsaiz@verily.com. Requests are subject to approval by PBHS governance.

Authors' Contributions

KYF, SAS, SS, and KWM contributed to the study concept and design. Verily Life Sciences, Stanford Center for Clinical Research, Duke University contributed to data collection. KYF, SAS, SS, and MKC performed data analysis and interpretation. All authors contributed to the draft writing and review. All authors approved the draft for submission.

Conflicts of Interest

SAS, SS, MKC, and EPS report employment and equity ownership in Verily Life Sciences. KWM reports research grants from Verily, American Heart Association, Apple Inc., Bayer, the California Institute of Regenerative Medicine, Eidos, Gilead, Idorsia, Johnson & Johnson, Luitpold, Pac-12, Precordior, Sanifit; consulting fees from Amgen, Applied Therapeutics, BMS, BridgeBio, Elsevier, Lexicon, Moderna, Sanofi; equity ownership in Precordior, Regencor. The rest of the authors report no relevant disclosures.

Multimedia Appendix 1
Methods and results.
[DOCX File , 540 KB - jmir_v26i1e60493_app1.docx]

Multimedia Appendix 2
High-resolution version of Figure 5.
[PNG File , 2684 KB - jmir_v26i1e60493_app2.png]

References

1. Umetani K, Singer DH, McCraty R, Atkinson M. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. J Am Coll Cardiol 1998;31(3):593-601 [FREE Full text] [doi: 10.1016/s0735-1097(97)00554-8] [Medline: 9502641]
2. Fox K, Borer JS, Camm AJ, Danchin N, Ferrari R, Lopez Sendon JL, Heart Rate Working Group. Resting heart rate in cardiovascular disease. J Am Coll Cardiol 2007;50(9):823-830 [FREE Full text] [doi: 10.1016/j.jacc.2007.04.079] [Medline: 17719466]

XSL•FO
RenderX

3. Aune D, Ó Hartaigh B, Vatten LJ. Resting heart rate and the risk of type 2 diabetes: a systematic review and dose--response meta-analysis of cohort studies. Nutr Metab Cardiovasc Dis 2015;25(6):526-534 [FREE Full text] [doi: 10.1016/j.numecd.2015.02.008] [Medline: 25891962]
4. Aune D, Sen A, ó'Hartaigh B, Janszky I, Romundstad PR, Tonstad S, et al. Resting heart rate and the risk of cardiovascular disease, total cancer, and all-cause mortality - a systematic review and dose-response meta-analysis of prospective studies. Nutr Metab Cardiovasc Dis 2017;27(6):504-517 [FREE Full text] [doi: 10.1016/j.numecd.2017.04.004] [Medline: 28552551]
5. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019;381(20):1909-1917 [FREE Full text] [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
6. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
7. Hannan AL, Harders MP, Hing W, Climstein M, Coombes JS, Furness J. Impact of wearable physical activity monitoring devices with exercise prescription or advice in the maintenance phase of cardiac rehabilitation: systematic review and meta-analysis. BMC Sports Sci Med Rehabil 2019;11:14 [FREE Full text] [doi: 10.1186/s13102-019-0126-8] [Medline: 31384474]
8. Brickwood KJ, Watson G, O'Brien J, Williams AD. Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019;7(4):e11819 [FREE Full text] [doi: 10.2196/11819] [Medline: 30977740]
9. Arges K, Assimes T, Bajaj V, Balu S, Bashir MR, Beskow L, et al. The project baseline health study: a step towards a broader mission to map human health. NPJ Digit Med 2020;3:84 [FREE Full text] [doi: 10.1038/s41746-020-0290-y] [Medline: 32550652]
10. Quer G, Gouda P, Galarnyk M, Topol EJ, Steinhubl SR. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: retrospective, longitudinal cohort study of 92,457 adults. PLoS One 2020;15(2):e0227709 [FREE Full text] [doi: 10.1371/journal.pone.0227709] [Medline: 32023264]
11. Schubert C, Archer G, Zelis JM, Nordmeyer S, Runte K, Hennemuth A, et al. Wearable devices can predict the outcome of standardized 6-minute walk tests in heart disease. NPJ Digit Med 2020;3:92 [FREE Full text] [doi: 10.1038/s41746-020-0299-2] [Medline: 32665977]
12. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, et al. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. NPJ Digit Med 2018;1:27 [FREE Full text] [doi: 10.1038/s41746-018-0032-6] [Medline: 31304309]
13. Kowahl N, Shin S, Barman P, Rainaldi E, Popham S, Kapur R. Accuracy and reliability of a suite of digital measures of walking generated using a wrist-worn sensor in healthy individuals: performance characterization study. JMIR Hum Factors 2023;10:e48270 [FREE Full text] [doi: 10.2196/48270] [Medline: 37535417]
14. Popham S, Burq M, Rainaldi EE, Shin S, Dunn J, Kapur R. An algorithm to classify real-world ambulatory status from a wearable device using multimodal and demographically diverse data: validation study. JMIR Biomed Eng 2023;8:e43726 [FREE Full text] [doi: 10.2196/43726] [Medline: 38875664]
15. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19(1):3-11. [doi: 10.2466/pr0.1966.19.1.3] [Medline: 5942109]
16. Du J, Boss J, Han P, Beesley LJ, Kleinsasser M, Goutman SA, et al. Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods. J Comput Graph Stat 2022;31(4):1063-1075 [FREE Full text] [doi: 10.1080/10618600.2022.2035739] [Medline: 36644406]
17. Github. miselect: variable selection for multiply imputed data. URL: https://github.com/umich-cphds/miselect [accessed 2024-04-03]
18. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2020;3:18 [FREE Full text] [doi: 10.1038/s41746-020-0226-6] [Medline: 32047863]
19. Lee JE, Lee DH, Oh TJ, Kim KM, Choi SH, Lim S, et al. Clinical feasibility of monitoring resting heart rate using a wearable activity tracker in patients with thyrotoxicosis: prospective longitudinal observational study. JMIR Mhealth Uhealth 2018;6(7):e159 [FREE Full text] [doi: 10.2196/mhealth.9884] [Medline: 30006328]
20. Jiang C, Faroqi L, Palaniappan L, Dunn J. Estimating personal resting heart rate from wearable biosensor data. : IEEE; 2019 Presented at: Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); May 19-22, 2019; Chicago, IL, USA p. 1-4. [doi: 10.1109/bhi.2019.8834554]
21. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
22. McGrath JJ, Matthews KA, Brady SS. Individual versus neighborhood socioeconomic status and race as predictors of adolescent ambulatory blood pressure and heart rate. Soc Sci Med 2006;63(6):1442-1453. [doi: 10.1016/j.socscimed.2006.03.019] [Medline: 16733078]

XSL•FO
RenderX

23. Chaix B, Jouven X, Thomas F, Leal C, Billaudeau N, Bean K, et al. Why socially deprived populations have a faster resting heart rate: impact of behaviour, life course anthropometry, and biology--the RECORD cohort study. Soc Sci Med 2011;73(10):1543-1550. [doi: 10.1016/j.socscimed.2011.09.009] [Medline: 22000762]
24. Latvala A, Kuja-Halkola R, Rück C, D'Onofrio BM, Jernberg T, Almqvist C, et al. Association of resting heart rate and blood pressure in late adolescence with subsequent mental disorders: a longitudinal population study of more than 1 million men in sweden. JAMA Psychiatry 2016;73(12):1268-1275. [doi: 10.1001/jamapsychiatry.2016.2717] [Medline: 27784035]
25. Califf RM, Wong C, Doraiswamy PM, Hong DS, Miller DP, Mega JL, Baseline Study Group. Biological and clinical correlates of the patient health questionnaire-9: exploratory cross-sectional analyses of the baseline health study. BMJ Open 2022;12(1):e054741 [FREE Full text] [doi: 10.1136/bmjopen-2021-054741] [Medline: 34983769]
26. Kim G, Lee YH, Jeon JY, Bang H, Lee BW, Kang ES, et al. Increase in resting heart rate over 2 years predicts incidence of diabetes: a 10-year prospective study. Diabetes Metab 2017;43(1):25-32. [doi: 10.1016/j.diabet.2016.09.002] [Medline: 27745827]
27. Aladin AI, Al Rifai M, Rasool SH, Keteyian SJ, Brawner CA, Michos ED, et al. The association of resting heart rate and incident hypertension: the henry ford hospital exercise testing (FIT) project. Am J Hypertens 2016;29(2):251-257. [doi: 10.1093/ajh/hpv095] [Medline: 26112864]
28. Jensen MT, Suadicani P, Hein HO, Gyntelberg F. Elevated resting heart rate, physical fitness and all-cause mortality: a 16-year follow-up in the Copenhagen male study. Heart 2013;99(12):882-887 [FREE Full text] [doi: 10.1136/heartjnl-2012-303375] [Medline: 23595657]
29. Agelink MW, Malessa R, Baumann B, Majewski T, Akila F, Zeit T, et al. Standardized tests of heart rate variability: normal ranges obtained from 309 healthy humans, and effects of age, gender, and heart rate. Clin Auton Res 2001;11(2):99-108. [doi: 10.1007/BF02322053] [Medline: 11570610]
30. McCrory C, Finucane C, O'Hare C, Frewen J, Nolan H, Layte R, et al. Social disadvantage and social isolation are associated with a higher resting heart rate: evidence from the irish longitudinal study on ageing. J Gerontol B Psychol Sci Soc Sci 2016;71(3):463-473. [doi: 10.1093/geronb/gbu163] [Medline: 25481923]
31. Colangelo LA, Yano Y, Jacobs DR, Lloyd-Jones DM. Association of resting heart rate with blood pressure and incident hypertension over 30 years in black and white adults: the CARDIA study. Hypertension 2020;76(3):692-698 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.120.15233] [Medline: 32783760]
32. Hill LK, Hoggard LS, Richmond AS, Gray DL, Williams DP, Thayer JF. Examining the association between perceived discrimination and heart rate variability in African Americans. Cultur Divers Ethnic Minor Psychol 2017;23(1):5-14 [FREE Full text] [doi: 10.1037/cdp0000076] [Medline: 28045306]
33. Tracy LM, Koenig J, Georgiou-Karistianis N, Gibson SJ, Giummarra MJ. Heart rate variability is associated with thermal heat pain threshold in males, but not females. Int J Psychophysiol 2018;131:37-43. [doi: 10.1016/j.ijpsycho.2018.02.017] [Medline: 29505850]
34. Adjei T, Xue J, Mandic DP. The female heart: sex differences in the dynamics of ECG in response to stress. Front Physiol 2018;9:1616 [FREE Full text] [doi: 10.3389/fphys.2018.01616] [Medline: 30546313]
35. Hnatkova K, Šišáková M, Smetana P, Toman O, Huster KM, Novotný T, et al. Sex differences in heart rate responses to postural provocations. Int J Cardiol 2019;297:126-134 [FREE Full text] [doi: 10.1016/j.ijcard.2019.09.044] [Medline: 31611089]
36. Lucas B, Grayson S, Hamidu H, Han A, No S, Varghese A, et al. Sex differences in heart rate responses to occupational stress. Stress 2020;23(1):13-18. [doi: 10.1080/10253890.2019.1621282] [Medline: 31144579]
37. Gillum RF. The epidemiology of resting heart rate in a national sample of men and women: associations with hypertension, coronary heart disease, blood pressure, and other cardiovascular risk factors. Am Heart J 1988;116(1 Pt 1):163-174. [doi: 10.1016/0002-8703(88)90262-1] [Medline: 2969182]
38. Khan H, Kunutsor S, Kalogeropoulos AP, Georgiopoulou VV, Newman AB, Harris TB, et al. Resting heart rate and risk of incident heart failure: three prospective cohort studies and a systematic meta-analysis. J Am Heart Assoc 2015;4(1):e001364 [FREE Full text] [doi: 10.1161/JAHA.114.001364] [Medline: 25589535]
39. Whelton SP, Narla V, Blaha MJ, Nasir K, Blumenthal RS, Jenny NS, et al. Association between resting heart rate and inflammatory biomarkers (high-sensitivity C-reactive protein, interleukin-6, and fibrinogen) (from the Multi-Ethnic Study of Atherosclerosis). Am J Cardiol 2014;113(4):644-649 [FREE Full text] [doi: 10.1016/j.amjcard.2013.11.009] [Medline: 24393259]
40. Rikli R, Jessie JC. The reliability and validity of a 6-minute walk test as a measure of physical endurance in older adults. J Aging Phys Act Internet 1998;6(4):363-375. [doi: 10.1123/japa.6.4.363]
41. Saint-Maurice PF, Troiano RP, Bassett DR, Graubard BI, Carlson SA, Shiroma EJ, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020;323(12):1151-1160 [FREE Full text] [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
42. Rennie KL, Hemingway H, Kumari M, Brunner E, Malik M, Marmot M. Effects of moderate and vigorous physical activity on heart rate variability in a British study of civil servants. Am J Epidemiol 2003;158(2):135-143. [doi: 10.1093/aje/kwg120] [Medline: 12851226]

XSL•FO
RenderX

43. Avram R, Tison GH, Aschbacher K, Kuhar P, Vittinghoff E, Butzner M, et al. Real-world heart rate norms in the Health eHeart study. NPJ Digit Med 2019;2:58 [FREE Full text] [doi: 10.1038/s41746-019-0134-9] [Medline: 31304404]
44. Gorny AW, Liew SJ, Tan CS, Müller-Riemenschneider F. Fitbit charge HR wireless heart rate monitor: validation study conducted under free-living conditions. JMIR Mhealth Uhealth 2017;5(10):e157 [FREE Full text] [doi: 10.2196/mhealth.8233] [Medline: 29055881]
45. Bai Y, Tompkins C, Gell N, Dione D, Zhang T, Byun W. Comprehensive comparison of apple watch and fitbit monitors in a free-living setting. PLoS One 2021;16(5):e0251975 [FREE Full text] [doi: 10.1371/journal.pone.0251975] [Medline: 34038458]

Abbreviations

CVD: cardiovascular disease
ECG: electrocardiography
ENET: Elastic net
HR: heart rate
IBI: interbeat interval
IRB: institutional review board
ICC: intraclass correlation coefficient
LASSO: least absolute shrinkage and selection operator
PBHS: Project Baseline Health Study
PHQ-9: patient health questionnaire-9
PPG: photoplethysmography
PRO: patient-reported outcome
RHR: resting heart rate
SES: socioeconomic status
VSW: Verily Study Watch
WHODAS: World Health Organization Disability Assessment Schedule

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by GE Ennis, R Zhang; comments to author 23.07.24; revised version received 13.09.24; accepted 22.10.24; published 20.12.24.

Please cite as:
Feng KY, Short SA, Saeb S, Carroll MK, Olivier CB, Simard EP, Swope S, Williams D, Eckstrand J, Pagidipati N, Shah SH, Hernandez AF, Mahaffey KW
Resting Heart Rate and Associations With Clinical Measures From the Project Baseline Health Study: Observational Study
J Med Internet Res 2024;26:e60493
URL: https://www.jmir.org/2024/1/e60493
doi:10.2196/60493
PMID:

©Kent Y Feng, Sarah A Short, Sohrab Saeb, Megan K Carroll, Christoph B Olivier, Edgar P Simard, Susan Swope, Donna Williams, Julie Eckstrand, Neha Pagidipati, Svati H Shah, Adrian F Hernandez, Kenneth W Mahaffey. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Willingness to Use Digital Health Screening and Tracking Tools for Public Health in Sexual Minority Populations in a National Probability Sample: Quantitative Intersectional Analysis

Wilson Vincent[1], MPH, PhD

Department of Psychology and Neuroscience, Temple University, Philadelphia, PA, United States

Corresponding Author:
Wilson Vincent, MPH, PhD
Department of Psychology and Neuroscience, Temple University
Weiss Hall
1701 N 13th St
Philadelphia, PA, 19122
United States
Phone: 1 404 200 0203
Email: wilson.vincent@temple.edu

Abstract

Background: Little is known about sexual minority adults' willingness to use digital health tools, such as pandemic-related tools for screening and tracking, outside of HIV prevention and intervention efforts for sexual minority men, specifically. Additionally, given the current cultural climate in the United States, heterosexual and sexual minority adults may differ in their willingness to use digital health tools, and there may be within-group differences among sexual minority adults.

Objective: This study compared sexual minority and heterosexual adults' willingness to use COVID-19–related digital health tools for public health screening and tracking and tested whether sexual minority adults differed from each other by age group, gender, and race or ethnicity.

Methods: We analyzed data from a cross-sectional, national probability survey (n=2047) implemented from May 30 to June 8, 2020, in the United States during the height of the public health response to the COVID-19 pandemic. Using latent-variable modeling, heterosexual and sexual minority adults were tested for differences in their willingness to use digital health tools for public health screening and tracking. Among sexual minority adults, specifically, associations with age, gender, and race or ethnicity were assessed.

Results: On average, sexual minority adults showed greater willingness to use digital health tools for screening and tracking than heterosexual adults (latent factor mean difference 0.46, 95% CI 0.15-0.77). Among sexual minority adults, there were no differences by age group, gender, or race or ethnicity. However, African American (*b*=0.41, 95% CI 0.19-0.62), Hispanic or Latino (*b*=0.36, 95% CI 0.18-0.55), and other racial or ethnic minority (*b*=0.54, 95% CI 0.31-0.77) heterosexual adults showed greater willingness to use digital health tools for screening and tracking than White heterosexual adults.

Conclusions: In the United States, sexual minority adults were more willing to use digital health tools for screening and tracking than heterosexual adults. Sexual minority adults did not differ from each other by age, gender, or race or ethnicity in terms of their willingness to use these digital health tools, so no sexual orientation-based or intersectional disparities were identified. Furthermore, White heterosexual adults were less willing to use these tools than racial or ethnic minority heterosexual adults. Findings support the use of digital health tools with sexual minority adults, which could be important for other public health-related concerns (eg, the recent example of mpox). Additional studies are needed regarding the decision-making process of White heterosexual adults regarding the use of digital health tools to address public health crises, including pandemics or outbreaks that disproportionately affect minoritized populations.

(J Med Internet Res 2024;26:e47448) doi:10.2196/47448

KEYWORDS

COVID-19; demographics; digital health; disparities; intersectionality; mHealth; mobile health; psychometric properties; sexual minority populations

XSL•FO
RenderX

Introduction

Despite the economic, health, and mortality impacts of COVID-19 on the population as a whole, there has been great variability in the public's willingness to participate in public health efforts to address the pandemic, including wearing masks and getting vaccinated [1,2]. Additionally, the COVID-19 pandemic has had an especially devastating impact on specific groups in the United States, with disproportionate mortality affecting older adults, men, and African American, and Hispanic American individuals [3-5]. Studies have examined the willingness to engage in preventive behaviors of American adults by age, gender, and race or ethnicity to curtail the COVID-19 pandemic, including the use of digital health tools for screening and tracking [6-11]. Although the COVID-19 pandemic has had a significant impact on sexual minority populations (ie, people who identify as gay, lesbian, bisexual, or other nonheterosexual sexual orientation identities) [12-14], less is known about sexual minority populations' preventive responses to the COVID-19 pandemic. In the current climate of increased medical mistrust, the corresponding unwillingness to follow public health recommendations often occurs along demographic lines [15-17].

Some public health efforts for screening and tracking COVID-19, such as mobile health (mHealth), which includes the use of smartphones and related digital technologies to assess or address health, and other digital health tools (eg, patient portals and web-based patient questionnaires), have previously proven effective, acceptable, and feasible for HIV prevention for sexual minority men [18-21]. Generally, research shows that digital technologies such as mHealth applications improve the feasibility of delivering health care to sexual minority patients [22]. Additionally, studies indicate that tracking mental and physical health-related information through mHealth applications was associated with better mental health status for sexual minority people to a greater extent during the COVID-19 pandemic than before the pandemic [23]. Despite the potential for significant benefits, research has yet to examine sexual minority populations' willingness to use mHealth tools for COVID-19.

For COVID-19, mobile apps created specifically in response to the COVID-19 pandemic have been the most frequently used type of mHealth tool [24]. Such mHealth tools aid in the screening, monitoring, and treatment of COVID-19 [24]. These mHealth-based approaches have had the advantage of easing the burden of in-person activities on COVID-19 testing and public health infrastructure (eg, avoiding supply chain issues, limiting the possibility of viral exposure) [18,25]. Previous research is mixed on whether there are no demographic differences by age, gender, or race or ethnicity in willingness to use digital health tools for COVID-19 screening and tracking. Although some studies showed no differences [26-28], others found evidence of greater willingness or support among younger adults, women, and racial and ethnic minority people [29-31]. However, these studies did not examine sexual orientation diversity. Heterosexual and sexual minority adults in the United States may differ in their willingness to participate in pandemic-related mHealth approaches, and these differences may vary based on other demographic characteristics such as age, gender, and race or ethnicity.

In addition to potential differences between heterosexual and sexual orientation–diverse populations, there may be notable differences among sexual minority populations. For example, the intersectionality theory asserts that seemingly independent yet intersecting social identities along social hierarchies based on dimensions such as race or ethnicity, gender, and sexual orientation jointly shape human experiences [32-35]. The intersectionality framework suggests that sexual minority people's multiple intersecting identities must be considered simultaneously (eg, sexual minority people who are also African American), rather than treating each identity as a mutually exclusive category [36,37]. A recent review [38] and commentaries [39,40] have emphasized the urgent need to use intersectionality theory in the conceptualization and methodology of examining digital health disparities.

Although research has yet to explore sexual minority people's willingness to use mHealth tools for COVID-19 based on their intersecting identities, such as age, gender, or race or ethnicity, the sexual minority health and HIV literatures illustrate how minoritized identities that intersect with sexual orientation–minoritized identities change the social position of individuals in ways that increase their risk of oppression and resulting adverse health outcomes. For example, Black sexual minority men experience both sexual orientation– and race-based stigma, and they experience more race- or ethnicity-based stigma in gay spaces than other groups [41,42]. Also, although both Black and White sexual minority men experience stereotypes about their sexual behaviors, assumptions may be more extreme for Black sexual minority men given the added layer of stereotypes about Black male sexuality [43,44]. The confluence of racism and antigay attitudes contributes to the increased risk of HIV for Black sexual minority men, including through social marginalization within communities of sexual minority men and late detection of HIV by medical and public health establishments, despite these men having no greater frequency or extent of sexual risk behavior to explain elevated HIV risk [42,45,46]. Additionally, Black sexual minority men may not benefit from or see the usefulness of mHealth tools for HIV prevention given that these tools may inadvertently stigmatize them through their "targeted" sexual health messages [47], or the tools may be viewed as a subpar offering in place of clinicians and public health professions "doing their jobs" [48]. Thus, although mHealth apps for HIV prevention have been found to be acceptable and feasible for implementation for sexual minority men in general [18-21], results may vary depending on intersecting racial, ethnic, or other identities.

This study examined the extent to which heterosexual adults and sexual minority adults differed in their willingness to participate in public health digital screening and tracking efforts to address the COVID-19 pandemic in a nationally representative sample of adults living in the United States. Also, for sexual minority participants, the author assessed differences within sexual minority populations in mean levels of willingness for COVID-19–related digital screening and tracking based on age group, gender, and race or ethnicity. Thus, the focus was on whether a sexual orientation–based disparity adversely

XSL•FO
RenderX

affected sexual minority adults' willingness to use digital health tools for screening and tracking and whether there were also intersectional disparities based on age, gender, and racial or ethnic categories.

Methods

Overview

This study was conducted and reported in accordance with the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE) guidelines [49,50]. Specific study methods are provided in subsequent sections.

Data

The COVID Impact Survey (CIS) is a national probability survey of US households designed to provide estimates for preventative behaviors and the impact of the COVID-19 pandemic; the data are publicly available [51]. The author used data from the last of 3 waves of cross-sectional data collection in the CIS, which occurred from May 30 to June 8, 2020 (n=2047). All 3 waves occurred between April 20 and June 8, 2020. These data were collected using the AmeriSpeak Panel, a probability-based panel distributed by NORC (formerly the National Opinion Research Center) at the University of Chicago.

US households were sampled with a known, nonzero probability of selection based on the NORC National Sample Frame, which was extracted from the US Postal Service Delivery Sequence File. Households were contacted by US mail, email, telephone, and field interviewers. The data are representative of noninstitutionalized adults who reside in the United States when weighted using sampling weights provider by the CIS. The CIS was funded by the Data Foundation. The NORC Institutional Review Board approved the CIS protocol to protect human participants (FWA00000142).

Measures

Willingness for Public Health Digital Screening and Tracking for COVID-19

Participants responded to questions asking about their likelihood of COVID-19–related testing (ie, "Testing you for COVID-19 infection using a Q-tip to swab your cheek or nose" and "Testing you for immunity or resistance to COVID-19 by drawing a small amount of blood") and digital screening and tracking (eg, "Installing an app on your phone that asks you questions about your own symptoms and provides recommendations about COVID-19" and "Installing an app on your phone that tracks your location and sends push notifications if you might have been exposed to COVID-19"). Response options ranged from (1) "extremely likely" to (5) "not at all likely." Items were reverse-coded such that higher scores reflected a greater perceived likelihood for screening and tracking. Participants had the option to respond with (88) "Already done this," and these cases were excluded using listwise deletion.

In a sample that included mostly heterosexual participants from Wave 2 of the CIS (manuscript under review), the measure showed construct validity in its positive correlations with participants having engaged in other protective behaviors to prevent COVID-19 infection (eg, "worn a face mask" and "avoided public or crowded places"). Additionally, participants who engaged in more frequent digital communications with friends and family before the public health response to the COVID-19 pandemic in the United States in March 2020 scored higher in willingness to use pandemic-related mHealth tools than participants who used digital communications with friends and family less frequently. The measure also showed measurement invariance across age groups, genders, and categories of race or ethnicity based on Wave 3. Based on Wave 1 of the CIS, the measure has demonstrated high internal consistency (Cronbach α=.90).

Demographic Characteristics

Participants self-reported their sexual orientation identity (ie, gay, lesbian, or bisexual, straight, something else, and I don't know). Sexual orientation identity was dichotomized to reflect heterosexual status and nonheterosexual sexual-minority status, respectively. The following additional demographic characteristics were assessed for measurement invariance: age, gender, and race or ethnicity. Additionally, participants reported their current age, which the CIS categorized (ie, 18-24 years, 25-34 years, 45-54 years, 55-64 years, 65-74 years, and ≥75 years) to help anonymize the data set; gender (female coded 1, male coded 0); and self-identified race or ethnicity (eg, Black or African American, Hispanic or Latino, White, multiple other races and ethnicities, such as Asian, Indian, and Native Hawaiian). Transgender and nonbinary identities were not options on the CIS.

Data Analysis Plan

This study tested the extent to which heterosexual and sexual minority adults differed in their willingness to use digital health tools for public health screening and tracking, a latent variable, and whether sexual minority adults' willingness to use these COVID-19–related digital health tools was associated with age, gender, and race or ethnicity. Measurement invariance (ie, whether the measure means and assesses the same thing across groups) was tested across heterosexual and sexual minority adults.

Descriptive statistics and Cronbach α were computed using Stata (version 16; StataCorp) [52], and coefficient ω and all analyses of associations and latent variables were conducted using M*plus* (version 8; Muthén & Muthén) [53]. Weighted least squares estimation with Delta parameterization was used to estimate model parameters [53]. This estimation method uses a diagonal weight matrix with SEs and mean- and variance-adjusted chi-square test statistics that rely on a full weight matrix (ie, ESTIMATOR=WLSMV in M*plus*) [53]. It is particularly appropriate for ordinal and nominal data [53]. Model fit was assessed with several fit indices based on any 2 of the following 3 criteria: a root-mean-square standard error of approximation (RMSEA) value of 0.06, a comparative fit index (CFI) value of at least 0.95, and a standardized root-mean-square residual (SRMR) criterion of 0.08 or less [54,55].

The author tested the extent to which the 5-item measure was invariant across the 2 sexual orientation categories. The 3 levels

XSL•FO
RenderX

of measurement invariance—configural, metric, and scalar—were tested to determine if the 5-item measure was invariant across sexual-orientation categories. For ordinal variables and weighted least squares estimation methods with Delta parameterization, configural invariance (ie, pattern invariance), the least strict form of invariance, shows that each group has the same indicators loading onto the same factors in the same direction (ie, positive versus negative). To model configural invariance: (1) factor loadings are free to vary across groups, (2) thresholds are free to vary across groups, (3) scale factors are fixed to 1 in all groups, (4) factor means are fixed to 0 for all groups, and (5) factor variances are free to vary across groups [53]. Metric invariance (ie, weak invariance) indicates the invariance of factor loadings across groups, wherein (1) factor loadings are constrained to be equal across groups, (2) the first threshold of each item is constrained to be equal across groups, (3) the second threshold of the item that sets the metric of the factor is constrained to be equal across groups, (4) scale factors are fixed to 1 in 1 group and free to vary in the other groups, (5) factor means a Muthén & Muthénre fixed to 0 in 1 group and free to vary in the other groups, and (6) factor variances remain free to vary across groups [53]. Scalar invariance (ie, strong invariance) indicates equivalence of item intercepts or thresholds, in the case of categorical or ordinal variables, across groups and is the minimum needed to proceed with using a measure to test for differences in latent factor means between groups [56,57]. Scalar invariance is the same as metric invariance, except that thresholds are constrained to be equal across groups [53].

To compare invariance models, the author used a difference in CFI (ΔCFI) equal to or greater than 0.01 to indicate noninvariance [56]. Thus, a lack of worsened model fit with increased constraints indicates measurement invariance. Although scaled chi-square difference tests scaled for the weighted least squares estimator were conducted, this test may detect small discrepancies in ways that are not practically or theoretically meaningful in sample sizes greater than 200 [56-58].

Upon determining measurement invariance, the latent factor mean difference between heterosexual adults and sexual minority adults in the underlying factor of willingness to use digital health tools for public health screening and tracking was tested. Specifically, the latent variable for willingness to use digital health tools was standardized such that its mean was fixed to 0 and SD set to 1. The factor mean remained 0 for the reference group, heterosexual individuals, but the factor mean was freely estimated for the comparison group, sexual minority individuals. Thus, the resulting mean for sexual minority individuals reflected the difference in the mean from the reference group on a standardized metric, or in SD units. To identify correlates of willingness to use digital health tools for public health screening and tracking among sexual minority populations, specifically, willingness to use digital health tools was regressed on sexual minority adults' age group, gender, and race or ethnicity, respectively.

Within an intersectionality-informed analytic framework, as described by Jackson et al [59], we can use additive measures of interaction to test for joint, referent, or excess intersectional disparities. Using the present analyses as a guiding example, the outcome variable would be recoded such that higher scores reflect a more adverse or disparity-oriented outcome (ie, less willingness to use COVID-19 screening and tracking tools). The predictor, gender (women coded 1), and the moderator, sexual orientation identity (sexual minority identity coded 1), would be coded such that the reference category (coded 0) is the nonminoritized group in this instance (ie, men and heterosexual adults) and the active category (coded 1) is the minoritized group (ie, women and sexual minority adults). Thus, the code of 1 reflects an adverse social position. For gender and sexual orientation, the original equation in the primary analyses before recoding and not including other covariates would be:

Given that the outcome should reflect a negative outcome to identify a disparity, the analyses would be repeated with the outcome variable recoded to reflect an unwillingness to use digital health tools rather than a willingness to use these tools. For gender and sexual orientation, the equation after recoding and not including other covariates would then be:

The joint disparity compares outcomes from the cell or group at the intersection of 2 minoritized identities, in this case, sexual minority women, to the group at the intersection of the 2 corresponding nonminoritized identities, in this instance, heterosexual men. In our example, $b_1 + b_2 + b_3$ equals the joint disparity in unwillingness to use COVID-19 screening and tracking tools comparing sexual minority women to heterosexual men. Referent disparities are those that affect only 1 minoritized population or identity, in this case, women compared with men among heterosexual adults or heterosexual adults compared with sexual minority adults among men. It describes the disparity based on gender as a proxy for sexism or sexual minority identity as a proxy for heterosexism or homonegativity, but not both. Specifically, b_1 equals the referent gender disparity in unwillingness to use COVID-19 digital screening among heterosexual adults, and b_2 equals the referent sexual minority disparity among men. Finally, the excess intersectional disparity focuses on the intersection of minoritized identities and describes the extent to which the joint disparity exceeds the 2 individual referent disparities. Suppose it is greater than 0, or statistically significant. In that case, the strength of the association indicates the disparity at the intersection of minoritized gender and sexual orientation, that is, women who are also sexual minority adults, and b_3 equals this excess intersectional disparity. A more detailed explanation can be found in Jackson et al [59] and VanderWeele and Tchetgen Tchetgen [60].

Disparities are indicated if the regression coefficients are positive, reflecting direct associations (ie, disadvantages for the minoritized groups) as opposed to inverse associations (ie, advantages for the more minoritized group). An advantage on an outcome for a relatively disadvantaged group that otherwise disproportionately and systematically experiences worse health

XSL•FO
RenderX

outcomes and greater health risks would not meet established definitions of a disparity [61,62].

Given the complex nature of these survey data, analyses were adjusted using a sampling weight based on the inverse of the probability of selection in the sample. These analyses also accounted for stratification using pseudostrata based on census tracts. The data producer, NORC, used pseudostrata to preserve confidentiality. Per NORC, they did not include cluster variables because there were negligible cluster effects, and excluding these variables better preserved confidentiality (personal communication; Jennifer Benz, May 14, 2021). Descriptive statistics for the present sample accounted for weighting and stratification to reflect the complex survey design and national representativeness of the sample along key raking variables (ie, age, gender, and race or ethnicity). Latent factor mean differences (ΔM) and regression coefficients (*b*) are presented with their 95% CIs. Missing data, which were up to 3.7% missing across analyses, were handled using listwise deletion.

Ethical Considerations

Temple University's institutional review board determined that the present analyses, which used deidentified publicly available data, did not require institutional approval for human participants research (contact the corresponding author for documentation).

Results

Sample Characteristics

Of the total sample of 1928 adults, 161 were sexual minority individuals. Other sample characteristics are listed in Tables 1 and 2. The sample size was reduced from 2047 due to missing data on sexual orientation (6.2%).

XSL•FO
RenderX

Table 1. Descriptive statistics based on raking variables (ie, age, gender, and race or ethnicity) given complex design and weighting and sexual orientation identities represented in the sample (N=1928). Proportions may not sum to 1, and percentages may not sum to 100 due to rounding. Subcategories (eg, sexual orientation subcategories) may not sum to 1928 due to missing data. Sexual orientation was significantly associated with age group; the pattern of results indicates that younger adults were more likely to self-identify as sexual minority individuals than older adults.

Variable	Total				Estimated proportion		Comparison linearized SE	*P* value
	n	Estimated proportion (%)	Linearized SE	Design effect	Heterosexual	Sexual minority		
Age (years)								<.001
18-29	311	0.211	0.015	2.48	0.882	0.118	0.025	
30-44	594	0.263	0.013	1.62	0.916	0.085	0.015	
45-59	435	0.238	0.013	1.74	0.945	0.055	0.012	
≥60	588	0.288	0.013	1.675	0.966	0.034	0.008	
Gender								.19
Male	914	0.478	0.015	1.82	0.940	0.060	0.009	
Female	1014	0.522	0.015	1.82	0.921	0.079	0.012	
Race or ethnicity								.72
Asian or Asian American	62	0.056	0.008	2.47	0.956	0.044	0.027	
Black or African American	227	0.124	0.010	1.83	0.920	0.080	0.026	
Hispanic or Latino	324	0.166	0.012	2.01	0.919	0.081	0.019	
White or European American	1127	0.623	0.015	1.84	0.934	0.066	0.010	
Other races and ethnicities	103	0.031	0.004	0.892	0.906	0.095	0.029	
Sexual orientation								__[a]
Homosexual	58	3.0	—	—	—	—	—	
Bisexual	69	3.6	—	—	—	—	—	
Heterosexual	1767	91.7	—	—	—	—	—	
Other: “Something else”	25	1.3	—	—	—	—	—	
Other: “I don’t know the answer”	9	0.5	—	—	—	—	—	

[a]Not available.

XSL•FO
RenderX

Table 2. Descriptive statistics based on sexual orientation and willingness to use mHealth tools for pandemic-related digital screening and tracking ("If these options were available to you, how likely would you be to participate in them?"; N=1928). Proportions may not sum to 1, and percentages may not sum to 100 due to rounding. Subcategories (eg, sexual orientation subcategories) may not sum to 1928 due to missing data. Sexual orientation was significantly associated with willingness to install "an app on your phone that tracks your location and sends push notifications if you might have been exposed to COVID-19"; the pattern of results indicates that sexual minority adults reported being more likely to install such an app than heterosexual adults. Additionally, sexual orientation was also significantly associated with willingness to "use a website to log your symptoms and location and get recommendations about COVID-19"; the pattern of findings indicates that sexual minority adults reported being more likely to use such a website than heterosexual adults.

Variable	Total				Estimated proportion		Comparison linearized SE	*P* value
	N	Estimated proportion (%)	Linearized SE	Design effect	Heterosexual	Sexual minority		
Installing an app on your phone that asks you questions about your own symptoms and provides recommendations about COVID-19								.06
Extremely likely	146	0.075	0.008	1.75	0.866	0.134	0.044	
Very likely	212	0.110	0.010	1.95	0.898	0.102	0.028	
Moderately likely	373	0.203	0.013	2.04	0.927	0.073	0.018	
Not too likely	404	0.219	0.013	1.85	0.937	0.063	0.014	
Not likely at all	757	0.377	0.015	1.73	0.950	0.050	0.009	
Already done this	27	0.017	0.004	1.89	0.893	0.107	0.062	
Installing an app on your phone that tracks your location and sends push notifications if you might have been exposed to COVID-19								<.001
Extremely likely	171	0.086	0.008	1.69	0.826	0.174	0.045	
Very likely	232	0.130	0.011	1.88	0.889	0.111	0.029	
Moderately likely	365	0.202	0.013	1.95	0.947	0.053	0.011	
Not too likely	342	0.180	0.013	2.17	0.958	0.042	0.012	
Not likely at all	790	0.394	0.015	1.71	0.943	0.057	0.010	
Already done this	15	0.007	0.002	1.54	0.948	0.052	0.052	
Using a website to log your symptoms and location and get recommendations about COVID-19								.008
Extremely likely	129	0.074	0.009	2.40	0.842	0.158	0.048	
Very likely	216	0.112	0.009	1.72	0.906	0.095	0.025	
Moderately likely	441	0.240	0.014	2.00	0.917	0.083	0.016	
Not too likely	403	0.204	0.012	1.72	0.953	0.047	0.014	
Not likely at all	710	0.631	0.012	1.73	0.949	0.051	0.010	
Already done this	15	0.009	0.003	1.81	1.00	0.000	0.000	

Psychometric Properties of Measure of Willingness to Use Digital Health Tools for COVID-19–Related Screening and Tracking

The measure of willingness to use digital health tools for COVID-19–related screening and tracking showed internal consistency and reliability (Cronbach α=.89 and coefficient ω=0.93). Additionally, as shown in Table 3, the measure was invariant by sexual orientation. Configural invariance was indicated by all factor loadings being significant and in the expected direction for each group. The configural model had no global fit statistics, as it was a fully saturated model. Next, the author tested a metric invariance model with factor loadings constrained to be equal across groups, and metric invariance was evident (ΔCFI<0.01; $\Delta\chi^2_2$=2.30; *P*=.32). Thus, the metric model had an equivalent model fit with the configural invariance model; the nonsaturated metric model fit the data, per the RMSEA, CFI, and SRMR (Table 3). Finally, the author tested a scalar invariance model, and scalar invariance was shown (ΔCFI<0.01; $\Delta\chi^2_8$=6.44; *P*=.60).

XSL•FO
RenderX

Table 3. Measurement invariance by sexual orientation for a measure of pandemic-related psychological distress (N=1928). Sexual orientation categories are heterosexual and sexual minority. The measure showed measurement in variance based on the criterion of ΔCFI<.01 [54] between the configural and the metric model and between the metric and the scalar model.

	Fit index				
	Chi-square (*df*)	*P* value	RMSEA[a]	CFI[b]	SRMR[c]
Invariance models					
Configural	0.00 (0)	<.001	0.00	1.00	0.00
Metric	2.30 (2)	.32	0.01	1.00	<0.01
Scalar	8.51 (10)	.58	<0.01	1.00	0.01

[a]RMSEA: root-mean-square standard error of approximation.

[b]CFI: comparative fit index.

[c]SRMR: standardized root-mean-square residual.

Mean Difference by Sexual Orientation on Willingness to Use Digital Health Tools for COVID-19–Related Screening and Tracking

Given scalar invariance, factor means for willingness to use COVID-19–related digital screening and tracking tools differed between heterosexual and sexual minority adults. Specifically, willingness to use digital health tools was nearly half an SD greater for sexual minority adults than for heterosexual adults (ΔM=0.46, 95% CI 0.15-0.77).

Associations Between Demographic Characteristics and Willingness to Use Digital Health Tools for COVID-19–Related Screening and Tracking Among Sexual Minority Adults

Within the population of sexual minority adults, no differences were detected by age group, gender, or race or ethnicity in their willingness to use digital health tools for COVID-19–related screening and tracking. Specifically, as detailed in Table 4, for each increase in age by group, there was no change in willingness to use digital health tools. Also, men and women did not differ in their willingness to use digital health tools. Finally, African American individuals, Hispanic or Latino individuals, and people of other races or ethnicities did not differ from White individuals (the reference group) in their willingness to use digital health tools.

Table 4. The independent associations of age, gender, and race and ethnicity with willingness to use digital health tools for public health screening and tracking (latent variable) in separate structural equation models in a national probability sample of sexual minority people (N=161).

Variable	Structural equation models					
	b [a]	95% CI	R^{2}[b]	RMSEA[c]	CFI[d]	SRMR[e]
Age (years)						
Age group[f]	0.04	–0.12 to 0.21	0.01	0.17	0.99	0.04
Gender						
Female (reference: male)	–0.14	–0.66 to 0.38	0.01	0.00	1.00	0.02
Race or ethnicity						
Race or ethnicity (reference: White or European American)			0.01	0.00	1.00	0.03
Black or African American	0.03	–0.69 to 0.75				
Hispanic or Latino	0.09	–0.45 to 0.64				
Other race or ethnicity	–0.20	–1.02 to 0.62				

[a]*b*=unstandardized regression coefficient.

[b]R^2=coefficient of determination.

[c]RMSEA: root-mean-square standard error of approximation.

[d]CFI: comparative fit index.

[e]SRMR: standardized root-mean-square residual.

[f]Age group is an ordinal variable with levels as follows: (1) 18-24 years, (2) 25-34 years, (3) 35-44 years, (4) 45-54 years, (5) 55-64 years, (6) 65-74 years, and (7) ≥75 years.

Additional models were tested for interactions of sexual minority status as a moderator with the other demographic characteristics as respective predictors in their associations with willingness to use digital health tools for screening and tracking. None of the interaction terms reached statistical significance in the models (Table 5). However, with the inclusion of the interaction terms, the main effects of being African American, Hispanic or Latino, and another racial or ethnic identity reached statistical

XSL•FO
RenderX

significance. Specifically, among heterosexual adults, being Black or African American was associated with 41% of an SD greater willingness to use digital health screening and tracking tools (*b*=0.41, CI 0.19-0.62), being Hispanic or Latino was associated with 36% of an SD greater willingness to use these tools (*b*=0.36, CI 0.18-0.55), and being of another racial or ethnic minority group was associated with 54% of an SD greater willingness to use these tools (*b*=0.54, CI 0.31-0.77).

Table 5. The independent associations of age, gender, and race or ethnicity and their respective interactions by sexual minority status with willingness to use digital health tools for public health screening and tracking (latent variable) in separate structural equation models in a national probability sample of sexual minority people (N=1928).

Variable	Structural equation models					
	b [a]	95% CI	R^{2b}	RMSEA[c]	CFI[d]	SRMR[e]
Age (years)			0.02	0.03	1.00	0.01
Age group[f]	–0.02	–0.06 to 0.02				
Sexual minority (reference: heterosexual)	0.28	–0.32 to 0.88				
Age group × sexual minority	0.06	–0.10 to 0.22				
Gender			0.02	0.00	1.00	0.01
Female (reference: male)	0.05	–0.08 to 0.19				
Sexual minority	0.58	0.25 to 0.91				
Female × sexual minority	–0.19	–0.68 to 0.31				
Race or ethnicity						
Race or ethnicity (reference: White or European American)			0.05	0.01	1.00	0.01
Black or African American	0.41	0.19 to 0.62				
Hispanic or Latino	0.36	0.18 to 0.55				
Other race or ethnicity	0.54	0.31 to 0.77				
Sexual minority	0.61	0.28 to 0.94				
Race or ethnicity × sexual minority						
Black/African American × sexual minority	–0.35	–1.05 to 0.35				
Hispanic or Latino × sexual minority	–0.21	–0.77 to 0.35				
Other race or ethnicity × sexual minority	–0.75	–1.57 to 0.07				

[a]*b*=unstandardized regression coefficient.

[b]R^2=coefficient of determination.

[c]RMSEA: root-mean-square standard error of approximation.

[d]CFI: comparative fit index.

[e]SRMR: standardized root-mean-square residual.

[f]Age group is an ordinal variable with levels as follows: (1) 18-24 years, (2) 25-34 years, (3) 35-44 years, (4) 45-54 years, (5) 55-64 years, (6) 65-74 years, and (7) ≥75 years.

The models were re-run with the outcome variable recoded to identify referent, excess intersectional, and joint disparities. Given the direction of the significant associations for each racial or ethnic minority group (Black or African American: *b*=–0.41, CI –0.62 to –0.19; Hispanic or Latino: *b*=–0.36, CI –0.55 to –0.18; and other racial or ethnic minority group: *b*=–0.54, CI –0.77 to –0.31), referent racial or ethnic or sexual orientation-based disparities were not detected. Also, no excess intersectional disparity was detected (Black or African American × sexual minority: *b*=–0.35, CI –0.35 to 1.05; Hispanic or Latino × sexual minority: *b*=0.21, CI –0.35 to 0.77; and other racial or ethnic minority group × sexual minority: *b*=–0.75, CI –0.07 to 1.57). Overall, no joint disparity was identified.

Discussion

Overview

Studies rarely examine the willingness of sexual minority populations to use mHealth and related digital health tools in the context of pandemic-related or non-HIV prevention. However, such mHealth tools have been acceptable and effective when used to fight the HIV epidemic for sexual minority men, specifically [18-21]. This study examined the use of digital health tools for screening and tracking for the COVID-19 pandemic, focusing on a broader, more diverse range of the sexual minority population in the United States. In particular,

XSL•FO
RenderX

the study used the conceptual [32,33] and methodological [34,35,39] frameworks of intersectionality theory to determine the presence of disparities between heterosexual and sexual minority adults across various intersections of identity (ie, age, gender, and race or ethnicity) in their willingness to use digital health tools for screening and tracking.

Findings indicated that sexual minority adults were significantly more willing to use digital health tools for screening and tracking than heterosexual adults in the United States. The greater willingness to use digital health tools among sexual minority adults compared with heterosexual adults might be explained partly by the familiarity of many sexual minority individuals with the use of mHealth and other digital health methods for outreach and other public health efforts, particularly sexual minority men in the context of HIV prevention [18-21].

Despite the difference between heterosexual adults and sexual minority adults in this study, there were no within-group demographic differences among sexual minority adults in their willingness to use digital health tools for screening and tracking. These findings from the United States are consistent with findings from a large survey of registered National Health Service users in the United Kingdom, in which there were no differences by age or gender in terms of willingness to participate in contact tracing through a mobile phone app in the adult population as a whole [63]. Interestingly, although sexual minority men are often more likely to be the focus of mHealth interventions for HIV [20,21], they were no more likely than sexual minority women to express a willingness to use digital health tools for screening and tracking in this study.

Based on the established definitions of a disparity [61,62] and the analytic framework of intersectionality theory [32,33,59], this study detected no referent disparity based on sexual orientation and no joint or excess joint disparity at the intersection of sexual orientation identity and other demographic characteristics. A previous study that used the same publicly available data without testing for sexual minority status as a predictor or moderator found no significant associations that would indicate an age-related, gender, or racial or ethnic disparity [28]. In contrast to other studies, which showed mixed findings for race or ethnicity and other demographics as a predictor without considering sexual minority status [26,27,29-31], the significant main effects of race or ethnicity in this study occurred among heterosexual adults in the presence of interactions of race or ethnicity with sexual minority status.

The differences between racial or ethnic minority adults and White adults may be explained, in part, by political ideology affecting attitudes toward the public health establishment during COVID-19. For example, a study found that moderate- and conservative-leaning respondents showed less support for using COVID-19–related digital health tools than liberal-leaning respondents in the same model in which racial and ethnic minorities showed greater support for using COVID-19–related digital health tools than White Americans or non-Hispanic Americans [31]. Additionally, studies indicate that some White Americans may be increasingly voting conservative [64-66]. To the extent that these political ideologies are also tied to public health mistrust, there may be noteworthy consequences. For example, there is evidence of excess deaths for conservative-voting adults compared with less conservative-voting adults in Florida and Ohio during the COVID-19 pandemic [67]. The CIS did not include questions regarding political beliefs. As such, this study did not test whether demographic factors interacted with political ideology, which would have helped to determine whether political ideology mattered within each racial or ethnic category. Additional studies are needed to examine the decision-making process of White heterosexual adults regarding their use of digital health tools for screening and tracking during public health emergencies.

The present psychometric evaluation indicated that the COVID-19–related psychological distress measure was assessing the same construct in heterosexual participants and sexual minority participants. A previous study has already validated the psychometric properties of the present items (eg, construct validity and internal consistency) and demonstrated measurement invariance across age groups, genders, and races and ethnicities [28]. Other studies on willingness to use digital health tools have typically used a single item [63] or several items treated as separate measures [68,69] rather than a single, validated scale. In terms of scales that use any variation on willingness (eg, intentions and perceived usefulness), 1 study used a 15-item measure [70] and another study used 2 measures of 32 items each [71]; these are notably longer than the 3-item measure of this study. Some studies have used items with binary yes-or-no responses [63,68,69], which may not capture sufficient gradation in response if the goal is to understand the degree of willingness.

Strengths and Limitations

This study has multiple strengths. For example, the study used a national probability sample to represent the population of noninstitutionalized adults in the United States. Additionally, the study used innovative methods to conceptualize and quantitatively identify disparities within the framework of intersectionality theory. In addition, the study established measurement invariance between heterosexual and sexual orientation–diverse adults for a measure of willingness to use a digital health screening and tracking tool that was previously validated by age, gender, and race or ethnicity. The measure can be adapted for screening and tracking in response to future public health events.

Additionally, several limitations must be noted. Specifically, the cross-sectional study design precludes definitive causal conclusions. In addition, the author did not attempt to draw conclusions about the temporal associations among the variables. Moreover, the sample was imbalanced with respect to the proportion of sexual minority adults compared with heterosexual adults in the sample; the number of sexual minority adults was much smaller. Limitations also include a lack of questions measuring sexual and gender identity that follow best practice [72], including the lack of transgender-inclusive gender questions and the lack of sexual orientation measures that distinguish different sexual orientation groups beyond sexual minority status by gender (ie, sexual minority men, including gay and bisexual men, and women, including lesbians and

XSL•FO
RenderX

bisexual women, which were accounted for in this study). As a result, the analyses were not more nuanced regarding gender and sexual orientation identity.

Implications

This study has several research and applied implications. For instance, additional research can oversample sexual minority adults to provide balanced samples for comparisons between heterosexual and sexual minority adults. Additionally, studies can examine sexual minority individuals' willingness to use digital health tools for other non-COVID-19–related health issues beyond HIV, including specific mental health diagnoses (eg, depression and substance use) and chronic illnesses (eg, diabetes and hypertension). These studies should consider intersections of identities among sexual minority people, such as underrepresented racial and ethnic minority people among sexual minority populations. Recently, monkey pox has emerged among sexual minority men, in particular [73], and digital health approaches may be useful in such circumstances. Additionally, studies are needed to further examine the decision-making process of White heterosexual adults regarding their use of digital health tools in response to public health emergencies.

Regarding applied implications, public health professionals and clinicians should consider screening sexual minority adults for their willingness to use digital health tools as they continue to use telehealth during COVID-19 and post–COVID-19 times. Such screening is particularly needed for sexual minority adults who contend with intersecting systems of oppression and identities and, thus, have elevated levels of medical mistrust (eg, underrepresented ethnic and sexual minority people) [15,17,74]. Policy changes and other structural interventions are needed to provide access to digital health technologies in cases in which willingness to use these technologies does not appear to be the issue. In this study, racial and ethnic minority heterosexual adults seemed particularly willing to use these technologies.

Conclusions

This study is responsive to recent calls in the literature to address the pronounced dearth of intersectionality theory-informed research investigating disparities related to digital health [38-40]. As we strive to narrow the digital divide, or the disparities in technology and internet access and use [75-77], we must understand disparities in willingness to use digital technologies for health-related purposes even when these tools are available. This study detected no disparities based on sexual minority status or intersections of identity among sexual minority adults along age, gender, or race or ethnicity. As such, for sexual minority populations, including intersections that are at joint or compound risk of experiencing adverse health outcomes, the issue is not willingness to use digital health tools compared to heterosexual adults. Sexual minority populations require culturally responsive digital health approaches to address their needs, as opposed to motivational enhancement or other interventions to increase their willingness during public health events. The willingness of sexual minority adults across intersecting identities to use pandemic-related digital health tools, including mobile health apps, is noteworthy given the potential promise of digital health tools for other public health-related concerns, such as the recently ended mpox outbreak [78,79], which disproportionately affected sexual minority men [80-82], and obesity [83,84] and cardiovascular disease [85,86], which disproportionately affects sexual minority women [87,88]. Additionally, White heterosexual adults demonstrated a disproportionately low willingness to use digital health tools, and this may become an issue in the event that this population is adversely affected by a public health concern that can benefit from digital health technologies.

Acknowledgments

The author would like to thank the NORC (formerly the National Opinion Research Center) at the University of Chicago. The data analyzed are freely available to the public through NORC at the University of Chicago. WV was supported by a National Institute of Mental Health grant (K23-MH111402).

Data Availability

The data sets analyzed during this study are available in the NORC at the University of Chicago repository [51].

Conflicts of Interest

None declared.

References

1. Daly M, Robinson E. Willingness to vaccinate against COVID-19 in the U.S.: representative longitudinal evidence from april to october 2020. Am J Prev Med 2021;60(6):766-773 [FREE Full text] [doi: 10.1016/j.amepre.2021.01.008] [Medline: 33773862]
2. Iboi E, Richardson A, Ruffin R, Ingram D, Clark J, Hawkins J, et al. Impact of public health education program on the novel coronavirus outbreak in the United States. Front Public Health 2021;9:630974 [FREE Full text] [doi: 10.3389/fpubh.2021.630974] [Medline: 33791268]
3. Truman BI, Chang MH, Moonesinghe R. Provisional COVID-19 age-adjusted death rates, by race and ethnicity—United States, 2020-2021. MMWR Morb Mortal Wkly Rep 2022;71(17):601-605 [FREE Full text] [doi: 10.15585/mmwr.mm7117e2] [Medline: 35482556]

XSL•FO
RenderX

4. Itzhak N, Shahar T, Moskovich R, Shahar Y. The impact of US county-level factors on COVID-19 morbidity and mortality. J Urban Health 2022;99(3):562-570 [FREE Full text] [doi: 10.1007/s11524-021-00601-7] [Medline: 35378717]
5. McGowan VJ, Bambra C. COVID-19 mortality and deprivation: pandemic, syndemic, and endemic health inequalities. Lancet Public Health 2022;7(11):e966-e975 [FREE Full text] [doi: 10.1016/S2468-2667(22)00223-7] [Medline: 36334610]
6. Berg MB, Lin L. Prevalence and predictors of early COVID-19 behavioral intentions in the United States. Transl Behav Med 2020;10(4):843-849 [FREE Full text] [doi: 10.1093/tbm/ibaa085] [Medline: 32893867]
7. Bowman L, Kwok KO, Redd R, Yi Y, Ward H, Wei WI, et al. Comparing public perceptions and preventive behaviors during the early phase of the COVID-19 pandemic in Hong Kong and the United Kingdom: cross-sectional survey study. J Med Internet Res 2021;23(3):e23231 [FREE Full text] [doi: 10.2196/23231] [Medline: 33539309]
8. Gerace A, Rigney G, Anderson JR. Predicting attitudes towards easing COVID-19 restrictions in the United States of America: the role of health concerns, demographic, political, and individual difference factors. PLoS One 2022;17(2):e0263128 [FREE Full text] [doi: 10.1371/journal.pone.0263128] [Medline: 35196316]
9. Huang Q, Abad N, Bonner KE, Baack B, Petrin R, Hendrich MA, et al. Explaining demographic differences in COVID-19 vaccination stage in the United States—April-May 2021. Prev Med 2023;166:107341 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107341] [Medline: 36372280]
10. Islam JY, Vidot DC, Havanur A, Camacho-Rivera M. Preventive behaviors and mental health-related symptoms among immunocompromised adults during the COVID-19 pandemic: an analysis of the COVID impact survey. AIDS Res Hum Retroviruses 2021;37(4):304-313 [FREE Full text] [doi: 10.1089/AID.2020.0302] [Medline: 33626959]
11. Latkin CA, Dayton L, Yi G, Konstantopoulos A, Boodram B. Trust in a COVID-19 vaccine in the U.S.: a social-ecological perspective. Soc Sci Med 2021;270:113684 [FREE Full text] [doi: 10.1016/j.socscimed.2021.113684] [Medline: 33485008]
12. Flentje A, Obedin-Maliver J, Lubensky ME, Dastur Z, Neilands T, Lunn MR. Depression and anxiety changes among sexual and gender minority people coinciding with onset of COVID-19 pandemic. J Gen Intern Med 2020;35(9):2788-2790 [FREE Full text] [doi: 10.1007/s11606-020-05970-4] [Medline: 32556877]
13. The lives and livelihoods of many in the LGBTS+ community are at risk amidst the COVID-19 crisis. Human Rights Campaign Foundation. Washington, DC; 2020. URL: https://tinyurl.com/54xjy5b8 [accessed 2023-09-22]
14. Salerno JP, Pease M, Devadas J, Nketia B, Fish JN. COVID-19-related stress among LGBTQ+ University students: results of a U.S. National Survey. University of Maryland Digital Repository at the University of Maryland. College Park, MD; 2020. URL: https://doi.org/10.13016/zug9-xtmi [accessed 2023-09-22]
15. Charura D, Hill AP, Etherson ME. COVID-19 vaccine hesitancy, medical mistrust, and mattering in ethnically diverse communities. J Racial Ethn Health Disparities 2023;10(3):1518-1525 [FREE Full text] [doi: 10.1007/s40615-022-01337-z] [Medline: 35641735]
16. Smith AC, Woerner J, Perera R, Haeny AM, Cox JM. An investigation of associations between race, ethnicity, and past experiences of discrimination with medical mistrust and COVID-19 protective strategies. J Racial Ethn Health Disparities 2022;9(4):1430-1442 [FREE Full text] [doi: 10.1007/s40615-021-01080-x] [Medline: 34117633]
17. Thompson HS, Manning M, Mitchell J, Kim S, Harper FWK, Cresswell S, et al. Factors associated with racial/ethnic group-based medical mistrust and perspectives on COVID-19 vaccine trial participation and vaccine uptake in the US. JAMA Netw Open 2021;4(5):e2111629 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11629] [Medline: 34042990]
18. Hall EW, Luisi N, Zlotorzynska M, Wilde G, Sullivan P, Sanchez T, et al. Willingness to use home collection methods to provide specimens for SARS-CoV-2/COVID-19 research: survey study. J Med Internet Res 2020;22(9):e19471 [FREE Full text] [doi: 10.2196/19471] [Medline: 32790639]
19. Valentine-Graves M, Hall E, Guest JL, Adam E, Valencia R, Shinn K, et al. At-home self-collection of saliva, oropharyngeal swabs and dried blood spots for SARS-CoV-2 diagnosis and serology: post-collection acceptability of specimen collection process and patient confidence in specimens. PLoS One 2020;15(8):e0236775 [FREE Full text] [doi: 10.1371/journal.pone.0236775] [Medline: 32756585]
20. MacCarthy S, Mendoza-Graf A, Wagner Z, Barreras JL, Kim A, Giguere R, et al. The acceptability and feasibility of a pilot study examining the impact of a mobile technology-based intervention informed by behavioral economics to improve HIV knowledge and testing frequency among Latinx sexual minority men and transgender women. BMC Public Health 2021;21(1):341 [FREE Full text] [doi: 10.1186/s12889-021-10335-5] [Medline: 33579242]
21. Ybarra ML, Prescott TL, Philips GL, Bull SS, Parsons JT, Mustanski B. Iteratively developing an mHealth HIV prevention program for sexual minority adolescent men. AIDS Behav 2016;20(6):1157-1172 [FREE Full text] [doi: 10.1007/s10461-015-1146-3] [Medline: 26238038]
22. Whaibeh E, Vogt EL, Mahmoud H. Addressing the behavioral health needs of sexual and gender minorities during the COVID-19 pandemic: a review of the expanding role of digital health technologies. Curr Psychiatry Rep 2022;24(9):387-397. [doi: 10.1007/s11920-022-01352-1] [Medline: 35841471]
23. Drydakis N. M-health apps and physical and mental health outcomes of sexual minorities. J Homosex 2023;70(14):3421-3448. [doi: 10.1080/00918369.2022.2095240] [Medline: 35904851]
24. Asadzadeh A, Kalankesh LR. A scope of mobile health solutions in COVID-19 pandemics. Inform Med Unlocked 2021;23:100558 [FREE Full text] [doi: 10.1016/j.imu.2021.100558] [Medline: 33842688]

XSL•FO
RenderX

25. Siegler AJ, Hall E, Luisi N, Zlotorzynska M, Wilde G, Sanchez T, et al. Willingness to seek diagnostic testing for SARS-CoV-2 with home, drive-through, and clinic-based specimen collection locations. Open Forum Infect Dis 2020;7(7):ofaa269 [FREE Full text] [doi: 10.1093/ofid/ofaa269] [Medline: 32704517]
26. Maytin L, Maytin J, Agarwal P, Krenitsky A, Krenitsky J, Epstein RS. Attitudes and perceptions toward COVID-19 digital surveillance: survey of young adults in the United States. JMIR Form Res 2021;5(1):e23000 [FREE Full text] [doi: 10.2196/23000] [Medline: 33347420]
27. Sorkin DH, Janio EA, Eikey EV, Schneider M, Davis K, Schueller SM, et al. Rise in use of digital mental health tools and technologies in the United States during the COVID-19 pandemic: survey study. J Med Internet Res 2021;23(4):e26994 [FREE Full text] [doi: 10.2196/26994] [Medline: 33822737]
28. Vincent W. Developing and evaluating a measure of the willingness to use pandemic-related mHealth tools using national probability samples in the United States: quantitative psychometric analyses and tests of sociodemographic group differences. JMIR Form Res 2023;7:e38298 [FREE Full text] [doi: 10.2196/38298] [Medline: 36689545]
29. Camacho-Rivera M, Islam JY, Rivera A, Vidot DC. Attitudes toward using COVID-19 mHealth tools among adults with chronic health conditions: secondary data analysis of the COVID-19 impact survey. JMIR Mhealth Uhealth 2020;8(12):e24693 [FREE Full text] [doi: 10.2196/24693] [Medline: 33301415]
30. Jansen-Kosterink S, Hurmuz M, den Ouden M, van Velsen L. Predictors to use mobile apps for monitoring COVID-19 symptoms and contact tracing: survey among Dutch citizens. JMIR Form Res 2021;5(12):e28416 [FREE Full text] [doi: 10.2196/28416] [Medline: 34818210]
31. Grande D, Mitra N, Marti XL, Merchant R, Asch D, Dolan A, et al. Consumer views on using digital data for COVID-19 control in the United States. JAMA Netw Open 2021;4(5):e2110918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10918] [Medline: 34009347]
32. Crenshaw KW. Demarginalizing the intersection of race and sex: a Black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. In: Bartlett KT, Kennedy R, editors. Feminist Legal Theory: Reading in Law and Gender. New York: Routledge; 1991:57-80.
33. Collins PH. Black Sexual Politics: African Americans, Gender, and the New Racism. New York: Routledge; 2004.
34. Bowleg L, Malekzadeh AN, AuBuchon KE, Ghabrial MA, Bauer GR. Rare exemplars and missed opportunities: intersectionality within current sexual and gender diversity research and scholarship in psychology. Curr Opin Psychol 2023;49:101511. [doi: 10.1016/j.copsyc.2022.101511] [Medline: 36586378]
35. Bauer GR, Churchill SM, Mahendran M, Walwyn C, Lizotte D, Villa-Rueda AA. Intersectionality in quantitative research: a systematic review of its emergence and applications of theory and methods. SSM Popul Health 2021;14:100798 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100798] [Medline: 33997247]
36. Bowleg L. When Black + Lesbian + Woman? Black lesbian woman: the methodological challenges of qualitative and quantitative intersectionality research. Sex Roles 2008;59(5):312-325. [doi: 10.1007/s11199-008-9400-z]
37. Bowleg L. "Once you've blended the cake, you can't take the parts back to the main ingredients": Black gay and bisexual men's descriptions and experiences of intersectionality. Sex Roles 2012;68(11-12):754-767. [doi: 10.1007/s11199-012-0152-4]
38. Husain L, Greenhalgh T, Hughes G, Finlay T, Wherton J. Desperately seeking intersectionality in digital health disparity research: narrative review to inform a richer theorization of multiple disadvantage. J Med Internet Res 2022;24(12):e42358 [FREE Full text] [doi: 10.2196/42358] [Medline: 36383632]
39. Queen R, Courtney KL, Lau F, Davison K, Devor A, Antonio MG. What's next for modernizing gender, sex, and sexual orientation terminology in digital health systems? Viewpoint on research and implementation priorities. J Med Internet Res 2023;25:e46773 [FREE Full text] [doi: 10.2196/46773] [Medline: 37490327]
40. Figueroa CA, Luo T, Aguilera A, Lyles CR. The need for feminist intersectionality in digital health. Lancet Digit Health 2021;3(8):e526-e533 [FREE Full text] [doi: 10.1016/S2589-7500(21)00118-7] [Medline: 34325855]
41. McConnell EA, Janulis P, Phillips G, Truong R, Birkett M. Multiple minority stress and LGBT community resilience among sexual minority men. Psychol Sex Orientat Gend Divers 2018;5(1):1-12 [FREE Full text] [doi: 10.1037/sgd0000265] [Medline: 29546228]
42. Raymond HF, McFarland W. Racial mixing and HIV risk among men who have sex with men. AIDS Behav 2009;13(4):630-637 [FREE Full text] [doi: 10.1007/s10461-009-9574-6] [Medline: 19479369]
43. Calabrese SK, Earnshaw VA, Magnus M, Hansen NB, Krakower DS, Underhill K, et al. Sexual stereotypes ascribed to Black men who have sex with men: an intersectional analysis. Arch Sex Behav 2018;47(1):143-156 [FREE Full text] [doi: 10.1007/s10508-016-0911-3] [Medline: 28224313]
44. Petsko CD, Bodenhausen GV. Racial stereotyping of gay men: can a minority sexual orientation erase race? J Exp Soc Psychol 2019;83:37-54. [doi: 10.1016/j.jesp.2019.03.002]
45. Millett GA, Peterson JL, Wolitski RJ, Stall R. Greater risk for HIV infection of Black men who have sex with men: a critical literature review. Am J Public Health 2006;96(6):1007-1019 [FREE Full text] [doi: 10.2105/AJPH.2005.066720] [Medline: 16670223]
46. Millett GA, Peterson JL, Flores SA, Hart TA, Jeffries WL, Wilson PA, et al. Comparisons of disparities and risks of HIV infection in black and other men who have sex with men in Canada, UK, and USA: a meta-analysis. Lancet 2012;380(9839):341-348. [doi: 10.1016/S0140-6736(12)60899-X] [Medline: 22819656]

XSL•FO
RenderX

47. Fields EL, Long A, Dangerfield DT, Morgan A, Uzzi M, Arrington-Sanders R, et al. There's an app for that: using geosocial networking apps to access young Black gay, bisexual, and other MSM at risk for HIV. Am J Health Promot 2020;34(1):42-51 [FREE Full text] [doi: 10.1177/0890117119865112] [Medline: 31359764]
48. Dangerfield Ii DT, Anderson JN, Wylie C, Arrington-Sanders R, Bluthenthal RN, Beyrer C, et al. Refining a multicomponent intervention to increase perceived HIV risk and PrEP initiation: focus group study among Black sexual minority men. JMIR Form Res 2022;6(8):e34181 [FREE Full text] [doi: 10.2196/34181] [Medline: 35947442]
49. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
50. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg 2014;12(12):1500-1524 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.014] [Medline: 25046751]
51. Wozniak A, Willey J, Benz J, Hart N. Version 1 dataset. COVID Impact Survey. Chicago, IL: National Opinion Research Center URL: https://www.covid-impact.org/results [accessed 2024-01-10]
52. StataCorp. Stata Statistical Software: Release 16. 2019. URL: https://www.stata.com/stata16/ [accessed 2024-01-16]
53. Muthén LK, Muthén B. Mplus User's Guide: Statistical Analysis with Latent Variables. Los Angeles, CA: Muthén & Muthén; 2017.
54. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling: Multidiscip J 1999;6(1):1-55. [doi: 10.1080/10705519909540118]
55. Bentler PM. On the fit of models to covariances and methodology to the Bulletin. Psychol Bull 1992;112(3):400-404. [doi: 10.1037/0033-2909.112.3.400] [Medline: 1438635]
56. Schmitt N, Kuljanin G. Measurement invariance: review of practice and implications. Hum Resour Manag Rev 2008;18(4):210-222. [doi: 10.1016/j.hrmr.2008.03.003]
57. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Dev Rev 2016;41:71-90 [FREE Full text] [doi: 10.1016/j.dr.2016.06.004] [Medline: 27942093]
58. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling: Multidiscip J 2007;14(3):464-504. [doi: 10.1080/10705510701301834]
59. Jackson JW, Williams DR, VanderWeele TJ. Disparities at the intersection of marginalized groups. Soc Psychiatry Psychiatr Epidemiol 2016;51(10):1349-1359 [FREE Full text] [doi: 10.1007/s00127-016-1276-6] [Medline: 27531592]
60. VanderWeele TJ, Tchetgen Tchetgen EJ. Attributing effects to interactions. Epidemiology 2014;25(5):711-722 [FREE Full text] [doi: 10.1097/EDE.0000000000000096] [Medline: 25051310]
61. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health 2006;27:167-194. [doi: 10.1146/annurev.publhealth.27.021405.102103] [Medline: 16533114]
62. National Institute of Minority Health and Health Disparities. Minority Health and Health Disparities Definitions. United States Department of Health and Human Services. 2023. URL: https://tinyurl.com/2wa4w6np [accessed 2024-01-10]
63. Bachtiger P, Adamson A, Quint JK, Peters NS. Belief of having had unconfirmed Covid-19 infection reduces willingness to participate in app-based contact tracing. NPJ Digit Med 2020;3(1):146 [FREE Full text] [doi: 10.1038/s41746-020-00357-5] [Medline: 33299071]
64. Buyuker B, D'Urso AJ, Filindra A, Kaplan NJ. Race politics research and the American presidency: thinking about white attitudes, identities and vote choice in the Trump era and beyond. J Race Ethn Polit 2020;6(3):600-641. [doi: 10.1017/rep.2020.33]
65. Knuckey J, Kim M. The politics of white racial identity and vote choice in the 2018 midterm elections. Soc Sci Q 2020;101(4):1584-1599. [doi: 10.1111/ssqu.12809]
66. Reny TT, Collingwood L, Valenzuela AA. Vote switching in the 2016 election: how racial and immigration attitudes, not economics, explain shifts in white voting. Public Opin Q 2016;83(1):91-113. [doi: 10.1093/poq/nfz011]
67. Wallace J, Goldsmith-Pinkham P, Schwartz JL. Excess death rates for republican and democratic registered voters in Florida and Ohio during the COVID-19 pandemic. JAMA Intern Med 2023;183(9):916-923 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1154] [Medline: 37486680]
68. Chen S, Sun G, Cen X, Liu J, Ye J, Chen J, et al. Characteristics and requirements of hypertensive patients willing to use digital health tools in the Chinese community: a multicentre cross-sectional survey. BMC Public Health 2020;20(1):1333 [FREE Full text] [doi: 10.1186/s12889-020-09462-2] [Medline: 32873273]
69. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018;6(3):e65 [FREE Full text] [doi: 10.2196/mhealth.9131] [Medline: 29549071]
70. Palos-Sanchez PR, Saura JR, Martin MAR, Aguayo-Camacho M. Toward a better understanding of the intention to use mHealth apps: exploratory study. JMIR Mhealth Uhealth 2021;9(9):e27021 [FREE Full text] [doi: 10.2196/27021] [Medline: 34499044]
71. Bennett BL, Goldstein CM, Gathright EC, Hughes JW, Latner JD. Internal health locus of control predicts willingness to track health behaviors online and with smartphone applications. Psychol Health Med 2017;22(10):1224-1229 [FREE Full text] [doi: 10.1080/13548506.2017.1317354] [Medline: 28415852]

XSL•FO
RenderX

72. Committee on Lesbian, Gay, Bisexual, and Transgender Health Issues and Research Gaps and Opportunities, Institute of Medicine (US). The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding. Washington, DC: The National Academies Press; 2011.
73. Delaney KP, Sanchez T, Hannah M, Edwards OW, Carpino T, Agnew-Brune C, et al. Strategies adopted by gay, bisexual, and other men who have sex with men to prevent monkeypox virus transmission-United States, august 2022. MMWR Morb Mortal Wkly Rep 2022;71(35):1126-1130 [FREE Full text] [doi: 10.15585/mmwr.mm7135e1] [Medline: 36048582]
74. Eaton LA, Driffin DD, Kegler C, Smith H, Conway-Washington C, White D, et al. The role of stigma and medical mistrust in the routine health care engagement of Black men who have sex with men. Am J Public Health 2015;105(2):e75-e82 [FREE Full text] [doi: 10.2105/AJPH.2014.302322] [Medline: 25521875]
75. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021;11(10):e053440 [FREE Full text] [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
76. Lythreatis S, Singh SK, El-Kassar AN. The digital divide: a review and future research agenda. Technol Forecast Soc Change 2022;175:121359. [doi: 10.1016/j.techfore.2021.121359]
77. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
78. Chan J, DiTullio DJ, Pagan Pirallo P, Foote M, Knutsen D, Kottkamp AC, et al. Implementation and early outcomes of a telehealth visit model to deliver tecovirimat for mpox infection in New York City. J Telemed Telecare 2023:1357633X231194796. [doi: 10.1177/1357633X231194796] [Medline: 37632124]
79. Shepherd T, Robinson M, Mallen C. Online health information seeking for mpox in endemic and nonendemic countries: Google trends study. JMIR Form Res 2023;7:e42710 [FREE Full text] [doi: 10.2196/42710] [Medline: 37052999]
80. Liu Q, Fu L, Wang B, Sun Y, Wu X, Peng X, et al. Clinical characteristics of human mpox (Monkeypox) in 2022: a systematic review and meta-analysis. Pathogens 2023;12(1):146 [FREE Full text] [doi: 10.3390/pathogens12010146] [Medline: 36678494]
81. McCollum AM, Shelus V, Hill A, Traore T, Onoja B, Nakazawa Y, et al. Epidemiology of human mpox—worldwide, 2018-2021. MMWR Morb Mortal Wkly Rep 2023;72(3):68-72 [FREE Full text] [doi: 10.15585/mmwr.mm7203a4] [Medline: 36656790]
82. Mitjà O, Alemany A, Marks M, Lezama Mora JI, Rodríguez-Aldama JC, Torres Silva MS, et al. Mpox in people with advanced HIV infection: a global case series. Lancet 2023;401(10380):939-949 [FREE Full text] [doi: 10.1016/S0140-6736(23)00273-8] [Medline: 36828001]
83. Hinchliffe N, Capehorn MS, Bewick M, Feenie J. The potential role of digital health in obesity care. Adv Ther 2022;39(10):4397-4412 [FREE Full text] [doi: 10.1007/s12325-022-02265-4] [Medline: 35925469]
84. Patel ML, Wakayama LN, Bennett GG. Self-monitoring via digital health in weight loss interventions: a systematic review among adults with overweight or obesity. Obesity (Silver Spring) 2021;29(3):478-499. [doi: 10.1002/oby.23088] [Medline: 33624440]
85. Redfern J, Coorey G, Mulley J, Scaria A, Neubeck L, Hafiz N, et al. A digital health intervention for cardiovascular disease management in primary care (CONNECT) randomized controlled trial. NPJ Digit Med 2020;3:117 [FREE Full text] [doi: 10.1038/s41746-020-00325-z] [Medline: 32964140]
86. Jiang X, Ming WK, You JH. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019;21(6):e13166 [FREE Full text] [doi: 10.2196/13166] [Medline: 31210136]
87. Simoni JM, Smith L, Oost KM, Lehavot K, Fredriksen-Goldsen K. Disparities in physical health conditions among lesbian and bisexual women: a systematic review of population-based studies. J Homosex 2017;64(1):32-44 [FREE Full text] [doi: 10.1080/00918369.2016.1174021] [Medline: 27074088]
88. Potter EC, Patterson CJ. Health-related quality of life among lesbian, gay, and bisexual adults: the burden of health disparities in 2016 behavioral risk factor surveillance system data. LGBT Health 2019;6(7):357-369. [doi: 10.1089/lgbt.2019.0013] [Medline: 31526230]

Abbreviations

CFI: comparative fit index
CIS: COVID Impact Survey
mHealth: mobile health
NORC: National Opinion Research Center
RMSEA: root-mean-square standard error of approximation
SRMR: standardized root-mean-square residual
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.03.23; peer-reviewed by A Asadzadeh, J Jabson Tree; comments to author 19.07.23; revised version received 29.11.23; accepted 20.12.23; published 08.03.24.

Please cite as:
Vincent W
Willingness to Use Digital Health Screening and Tracking Tools for Public Health in Sexual Minority Populations in a National Probability Sample: Quantitative Intersectional Analysis
J Med Internet Res 2024;26:e47448
URL: https://www.jmir.org/2024/1/e47448
doi:10.2196/47448
PMID:38457790

©Wilson Vincent. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Health Information Seeking and Preventative Health Actions: Cross-Generational Online Survey Study

Jayati Sinha[1*], BSc, MBA, PhD; Nuket Serin[2*], BSc, MSc, PhD

[1]Department of Marketing & Logistics, College of Business, Florida International University, Miami, FL, United States

[2]W. Fielding Rubel School of Business, Bellarmine University, Louisville, KY, United States

[*]all authors contributed equally

Corresponding Author:
Jayati Sinha, BSc, MBA, PhD
Department of Marketing & Logistics
College of Business
Florida International University
11200 SW 8th Street
Miami, FL, 33146
United States
Phone: 1 3053487573
Email: jsinha@fiu.edu

Abstract

Background: The popularity of online health information seeking (OHIS) has increased significantly owing to its accessibility and affordability. To facilitate better health management, it is essential to comprehend the generational differences in OHIS behavior and preventative health actions after seeking online health information (OHI).

Objective: This study investigates the variations in OHIS and engagement in preventative health actions between 2 generations based on their technology use (digital natives [aged 18-42 years] and digital immigrants [aged ≥43 years]). Additionally, this research explores the mediating role of OHIS types on the generational effect on preventative health actions and the moderating role of OHI search frequency, gender, and the presence of chronic diseases on the generational effect on OHIS types and preventative health actions.

Methods: A preregistered online survey was conducted on the Prolific online data collection platform using stratified sampling of 2 generations (digital natives and digital immigrants) from the United States in November 2023. Overall, 3 types of OHIS were collected: health wellness information search, health guidance information search, and health management information search. A 1-way analysis of covariance tested the generational differences in types of OHIS and preventative health actions, and a 2-way analysis of covariance tested the moderating role of OHIS search frequency, gender, and the presence of chronic diseases using 7 control variables. The PROCESS Macro Model 4 was used to conduct mediation analyses, testing OHI search types as mediators. Linear regression analyses tested age as a predictor of OHIS and preventative health actions.

Results: The analysis of 1137 responses revealed generational differences in OHIS. Digital natives searched for health wellness information more frequently ($P<.001$), whereas digital immigrants searched for health guidance ($P<.001$) and health management information ($P=.001$) more frequently. There were no significant differences between the 2 generations regarding preventative health actions ($P=.85$). Moreover, all 3 types of OHIS mediated the relationship between generational differences and preventative health actions. Furthermore, as people aged, they searched for significantly less health wellness information ($P<.001$) and more health guidance ($P<.001$), and health management information ($P=.003$). Age was not a significant predictor of preventative health actions ($P=.48$). The frequency of OHI searches did not moderate the effect of generations on OHIS types and preventative health actions. Gender only moderated the relationship between generation and health guidance information search ($P=.02$), and chronic diseases only moderated the relationship between generation and health wellness information search ($P=.03$).

Conclusions: To the best of our knowledge, this study is the first to explore how 2 digital generations vary in terms of searching for OHI and preventative health behaviors. As the older adult population grows, it is crucial to understand their OHIS behavior and how they engage in preventative health actions to enhance their quality of life.

(J Med Internet Res 2024;26:e48977) doi:10.2196/48977

XSL•FO
RenderX

KEYWORDS

digital natives; digital immigrants; online health information seeking; preventative health actions; mobile phone

Introduction

Background

Online health information seeking (OHIS) has become popular owing to its convenience and low-cost access [1]. Although the internet remains a popular source of health information [1], an age-based digital divide exists in OHIS behavior—young generations (such as Millennials) are more likely to seek online health information (OHI) than older generations (such as Baby Boomers) [2-7]. The gap between younger and older Americans has widened in recent years—the number of Americans aged ≥55 years grew by 27% between 2010 and 2020, whereas the number of Americans aged <55 years increased by only 1.3% during that time [8]. As the US population ages, the demand for health care and medical information increases for older generations [1,7]. However, older generations' OHIS behavior, specifically types of health-related information searches, is not well understood [9], and it is unclear to what extent older and younger Americans engage in preventative health behaviors after seeking OHI [9-11]. Moreover, the increased availability of online medical information has made patient-centered medical care more common, allowing health consumers to make informed decisions about their health. Thus, understanding the generational differences in OHI search behavior and subsequent health-related actions is crucial for designing more effective communication methods to improve and maintain health. This study explored how 2 generations with different experiences with technology (digital natives and digital immigrants) differ in their OHI search activities and preventative health actions.

Digital Natives and Digital Immigrants

Generations are defined by the years of birth. Generational cohorts are groups of people who share similar cultural, historical, and technological experiences growing up and tend to develop similar beliefs, values, and behaviors as they age [12,13]. Generations are often defined by how they use and grow up with technology [14]. Indeed, the Technology Model of Generations [15] suggests that technology is the root cause of most generational differences and indirectly affects other sociocultural forces that shape each generation (such as a shift from collectivism to individualism and the introduction of a slow life strategy). Prensky [16] defined 2 generational groups based on their preferences for technology or digital devices: digital natives (born from 1980 onward) and digital immigrants (born before 1980). Digital natives were raised in a technology-driven world (such as smartphones, tablets, and laptops as commonplace items in daily life), whereas digital immigrants did not have the same experiences with technologies [15-20].

Exposure to digital environments (or lack thereof) affects individuals' use, fluency, comfort level, literacy, and engagement with technology [21-25] as well as how they think and behave [26,27]. Digital natives tend to process digital information quickly, multitask more, and depend on technology to communicate and access information and are usually more comfortable with technology than their digital immigrant counterparts [28,29]. As digital technologies have been integrated into the lives of digital natives since the early days, they feel more at ease with OHI searches than digital immigrants [30,31]. Indeed, nearly 48% of Americans aged <30 years use the internet constantly compared with 22% of Americans aged 50 to 64 years and 8% of Americans aged >65 years [32].

Researchers have suggested studying 2 generations along a continuum based on user technology proficiency, digital literacy, or engagement with technology rather than the age-based classification of digital natives and digital immigrants [22,25]. Although differences in computer efficacy, socioeconomic status, health literacy, and technology accessibility might help explain some differences between digital natives and digital immigrants [33-35], a generation gap exists in attitude and behavior toward technology use [16,17]. Research shows that younger people prefer the internet as a source of health information [3,7,36], and with increased age, the information processing capacity for technology-enabled tasks diminishes [29]. Previous research has shown that negative attitudes toward internet use as well as privacy concerns about using technology decrease older adults' OHIS behavior [9,35]. Given that digital natives grew up with digital technology and are familiar with it from a young age, their OHIS behavior is likely to differ from that of digital immigrants as each generation uses technology differently [32]. Accordingly, digital natives and immigrants may differ in their search for various types of OHI and engagement in preventive efforts after retrieving such information. Thus, this study focused on 2 generations that differ in their level of technology immersion: digital natives and digital immigrants [16].

Seeking OHI and Preventative Behaviors

Health consumers seek OHI for various reasons, including gaining knowledge about health issues, managing health problems, and making health decisions [37-39]. Research has shown that OHI seekers frequently search for disease-related information [40]. In addition, people seek information about symptoms, medication, treatment, exercise and fitness, and nutrition or diet [41-43]. However, only a few studies have examined what types of OHI older and younger generations seek [9,43,44]. For example, Chinese OHI seekers most often search for information about health science popularization to improve their health literacy, whereas medical concerns are the least explored type of health information [43]. A recent review of the literature on health information seeking indicates that older adults seek various types of OHI, including information about specific diseases, medications and treatments, disease symptoms, nutrition and exercise, medical resources, and interpersonal advice [9]. Compared with older adults, adolescents and young adults seek online information on physical and psychological well-being, sexual health, and culturally sensitive topics [9]. In addition, OHI searches are lower for German and Australian older adults aged >65 years [7,36]. Moreover, young adults seek specific health information online by joining social media groups or following relevant

XSL•FO
RenderX

pages [44]. It has also been found that the frequency of OHIS is higher among female individuals [43,45,46] and younger people [3,7,36], whereas people with chronic diseases are more likely than healthy people to perform OHI searches [47,48]. Although racial disparities exist in OHIS behaviors [49-51], not all racial groups seek out information equally, and chronic diseases have been found to be more important than race or ethnicity in OHI search and use [52]. Moreover, political affiliation is associated with various health outcomes, including health information searches, vaccine adoption, and preventative health behaviors [53-56].

Despite the availability of a large amount of health information online [57], some OHI can be inaccurate or misleading, causing unnecessary distress and anxiety among health consumers [58,59]. The inaccuracy of OHI raises concerns about its quality [60]. For example, health consumers, especially young people, increasingly rely on social media as their first resource for health information. However, online content is often unverified and can be misleading [61,62]. Specifically, research has shown that older adults tend to be concerned about the credibility of OHI sources [63]. Moreover, low health literacy prevents people from accurately determining the credibility of OHI on social media [34,35,64,65]. In addition, health consumers' trust in OHI is associated with their self-efficacy beliefs and behavioral outcomes [66], and their confidence in OHIS skills is positively related to the use of the retrieved information [67]. Previous studies have identified that credibility, usefulness, and trust most frequently affect older adults' OHIS and health behaviors [9,39].

Access to OHI has become critical for managing health and disease, and it generally leads to positive outcomes such as improved health outcomes and empowered patients [68,69]. Research shows that health information seeking improves patient involvement and satisfaction with medical decision-making and communication between patients and informal care providers [70,71]. Despite the potential of OHIS to affect health behaviors across different age groups [47,72], little research has examined whether different generations based on their technology adoption engage in preventative behaviors after seeking OHI in the same way. As a result, there is a clear need to understand what types of health information different digital generations seek online [9,11]. Thus, this study focused on generational differences based on the use of technology or digital devices in seeking out types of health information online and engaging in preventative health actions afterward.

Objectives

This study addressed the limitations of previous research on information seeking from a generational perspective. It focused on 2 generations that differ in technology proficiency (digital natives and digital immigrants) to examine how they differ in their OHI search behavior. This study aimed to determine the types of OHI that digital natives and digital immigrants seek as well as whether each generation varies in their engagement in preventative health actions after seeking different types of OHI. The following three research questions guided this study:

1. How do generational differences influence different types of OHIS and preventative health actions?
2. Do types of OHIS behaviors mediate the generational effect on preventative health actions?
3. Do the frequency of OHI searches, gender, or the presence of a chronic disease moderate the generational effect on OHIS types and preventative health actions?

Methods

Study Design

We conducted an online survey to understand how OHIS behavior influences preventative health behaviors across 2 generations (digital natives and digital immigrants) and whether the frequency of OHI searches, gender, and the presence of a chronic disease moderate the effect. We used a 2-factor mixed factorial design where the first factor was a between-subject moderator (2 digital generations: natives vs immigrants) and the second factor was a within-subject moderator (frequency of OHI searches [frequent vs infrequent], gender [male vs female], or presence of a chronic disease [yes vs no]).

Ethical Considerations

The Bellarmine University Institutional Review Board (IRB) approved this study under the *exempted research* category (IRB 1092) before data collection, and general information about the nature of this study was included at the beginning of the survey as a means of informed consent. No formal informed consent was obtained. This study was preregistered on the AsPredicted platform (Wharton Credibility Lab) [73]. The online platform Prolific was used for data collection. Researchers using Prolific only had access to a user ID, and the respondents' identities were not revealed. An online survey was designed using the Qualtrics survey design software (Qualtrics International Inc), which did not capture any identifying information. We provided respondents with a monetary reward (US $0.40) as an incentive for participation.

Data Collection

Study data were collected in November 2023 from Prolific, an online data collection platform, using samples of adults from the United States. We selected Prolific as an online data collection platform because of its high data quality compared with Amazon Mechanical Turk and CloudResearch [74]. We used a power analysis using the G*Power software (version 3.1; Heinrich-Heine-Universität Düsseldorf) to identify the appropriate sample size [75]. The power analysis results revealed that, for a small (0.15) effect size with α=.05 and a statistical power of 0.95, the total minimum sample size was approximately 768 respondents (4 groups and 7 covariates). We used a similar stratified sampling method [76] based on age, sex, and political affiliation to understand the OHIS behavior of adults in the United States. Specifically, we aimed to collect an age-, sex-, and political affiliation–stratified sample of 1200 respondents to ensure an adequate sample size after eliminating respondents based on the data exclusion criteria outlined in the preregistration. On the basis of the age and US political affiliation criteria, we aimed to recruit 300 digital natives (aged between 18 and 42 years) of Democratic affiliation, 300 digital natives of Republican affiliation, 300 digital immigrants (aged between 43 and 99 years) of Democratic affiliation, and 300

XSL•FO
RenderX

digital immigrants of Republican affiliation from Prolific for this study (a total of 1200). To ensure a sample representative of both sexes, we used *balanced sample* criteria for an even distribution of male and female respondents. Thus, an age-, sex-, and political affiliation–stratified sample was used to ensure an equal number of respondents in the 2 digital generation groups (digital natives vs digital immigrants), with an even distribution of gender (male and female) and political affiliation (Republican and Democratic). In addition, we recruited respondents from the United States using *location* criteria as well as those with a 100% approval rate from Prolific to ensure data quality. The final recorded responses were from 299 digital immigrants who were Democrats, 311 digital immigrants who were Republicans, 310 digital natives who were Democrats, and 309 digital natives who were Republicans (a total of 1229 responses were used for the analysis after merging 4 surveys). This deviation was due to the Prolific software and was outside the researchers' control.

Data Exclusion

On the basis of the exclusion criteria outlined in the preregistration, we excluded 92 respondents in the following order—17 (18%) failed the attention check question, 8 (9%) did not complete the survey, 26 (28%) responded "never" for the frequency of OHIS, and 7 (8%) participants' age and 34 (37%) participants' political affiliation did not match for the respective surveys—from the sample, leaving 1137 (n=571, 50.22% digital natives and n=566, 49.78% digital immigrants) responses for analysis. The operationalization and coding of the focal study variables are shown in Table 1.

Table 1. The operationalization and coding of study variables.

Variables	Operationalization and coding	References
Generation	Digital natives=aged between 18 and 42 years; digital immigrants=aged ≥43 years	Ransdell et al [18], 2011; Roth-Cohen et al [77], 2021
Frequency of OHIS[a]	"How frequently did you search for online health information in the past 12 months?" ("frequent," "infrequent," and "never"; 1=frequent; 2=infrequent; respondents who selected "never" were excluded)	Xiong et al [43], 2021
HWIS[b]	The sum of 3 types (sports and fitness, nutrition and diet, and general health knowledge; 1=yes; 0=no)	Neter and Brainin [78], 2012; Xiong et al [43], 2021
HGIS[c]	The sum of 2 types (medication guidance and disease consulting; 1=yes; 0=no)	Neter and Brainin [78], 2012; Xiong et al [43], 2021
HMIS[d]	The sum of 2 types (managing health conditions and participating in an online support group; 1=yes; 0=no)	Neter and Brainin [78], 2012; Xiong et al [43], 2021
PHA[e]	The sum of 3 items ("After seeking health information and finding the information, did your health behavior change for the better? After seeking health information and finding the information, did you see a doctor? After seeking health information and finding the information, did you monitor your health yourself for any changes?"; 1=yes; 0=no)	Taylor and Humphrey [79], 2022
Gender	1=male; 2=female	N/A[f]
Marital status[g]	1=married; 2=single	Xiong et al [43], 2021
Employment[h]	1=unemployed; 2=employed	Xiong et al [43], 2021
Education level	1=high school graduate or lower; 2=high school graduate; 2=some college or associate degree; 4=bachelor's degree; 5=graduate degree or higher	Gordon and Crouch [80], 2019
Income	1=<US $25,000; 2=US $25,000-$34,999; 3=US $35,000-$49,999; 4=US $50,000-$64,999; 5=US $65,000-$79,999; 6=US $80,000-$99,999; 7=≥US $100,000	Gordon and Crouch [80], 2019
Race	1=White; 2=Black or African American; 3=American Indian or Alaska Native; 4=Asian; 5=Native Hawaiian or other Pacific Islander; 6=some other race	Jensen et al [81], 2021
Ethnicity	1=Hispanic or Latino; 2=not Hispanic or Latino	Jensen et al [81], 2021
Political affiliation	1=Democrat; 2=Republican; 3=independent; 4=other	Naeim et al [55], 2021

[a]OHIS: online health information seeking.

[b]HWIS: health wellness information search.

[c]HGIS: health guidance information search.

[d]HMIS: health management information search.

[e]PHA: preventative health actions.

[f]N/A: not applicable.

[g]Marital status was divided into 2 categories: single (unmarried, divorced, or widowed) and married.

[h]Employment was recorded as 2 groups: unemployed (unemployed, students, and retired) and employed.

XSL•FO
RenderX

Study Measures

We adapted the online survey measures based on previous research. Multimedia Appendix 1 contains a list of all the measures in the order in which they appeared in the survey.

Digital Generations: Natives and Immigrants

On the basis of previous research [18,77], we operationalized 2 digital generations based on age: digital natives (Generation Z and Millennials; aged between 18 and 42 years) and digital immigrants (Generation X and Baby Boomers; aged ≥43 years).

Frequency of OHIS

The frequency of OHIS was examined using a subjective assessment of how frequently respondents engaged in OHIS in the previous 12 months on a 3-point scale (*frequent*, *infrequent*, and *never*; adapted from the study by Xiong et al [43]). We excluded respondents who selected *never*. Those who chose *frequent* were grouped as frequent OHI seekers, and those who indicated *infrequent* were grouped as infrequent OHI seekers.

Health-Related Information

For health-related information, we asked respondents the following: "Please indicate whether you have searched for the following health-related information during the past 12 months." We presented 7 types of health-related information with a dichotomized (*yes* or *no*) response (adapted from the studies by Xiong et al [43] and Neter and Brainin [78]). We grouped them into 3 search categories: health wellness information search (HWIS; sports and fitness, nutrition and diet, and general health knowledge), health guidance information search (HGIS; medication guidance and disease consulting), and health management information search (HMIS; managing health conditions and participating in an online support group). We created 3 scores by adding the responses in the 3 search categories (HWIS, HGIS, and HMIS).

Preventative Health Actions

Preventative health actions were measured using 3 items with a dichotomized (*yes* or *no*) response ("After seeking health information and finding the information, did your health behavior change for the better? After seeking health information and finding the information, did you see a doctor? After seeking health information and finding the information, did you monitor your health yourself for any changes?"; adapted from the study by Taylor and Humphrey [79]). We computed a preventative health actions score by summing the responses to the 3 items.

Control Variables

We tested 7 control variables. We controlled for information quality, trust, and search skills (adapted from the studies by Miller and Bell [30], Diviani et al [60], and Xiao et al [82]): the *quality* of OHI ("How would you rate the quality of health-related information on the Internet?" [1=*very good quality*; 7=*very poor quality*]), *trust* in OHI ("How much would you trust health-related information on the Internet?" [1=*a lot*; 7=*not at all*]), *trust* in US health care institutions ("How much would you trust healthcare institutions in the US?" [1=*a lot*; 7=*not at all*]), and respondents' online health-related information *search skills* ("Please rate your health-related information search skills on the Internet." [1=*very good*; 7=*very poor*]). In addition, we used 3 other control variables: *race* ("What is your race?" [*White*, *Black or African American*, *American Indian or Alaska Native*, *Asian*, *Native Hawaiianor other Pacific Islander*, and *some other race*]; adapted from the study by Jensen et al [81]), *ethnicity* ("What is your ethnicity?" [*Hispanic or Latino* and *not Hispanic or Latino*]; adapted from the study by Jensen et al [81]), and *political affiliation* ("What is your political affiliation?" [*Democrat*, *Republican*, *independent*, and *other*]; adapted from the study by Naeim et al [55]). We used these 3 variables as covariates in all analyses.

Sociodemographic Information

Demographic information included age, gender, annual household income, education, marital status, and employment status. We also asked respondents whether they went for a regular annual checkup ("Do you go for annual checkup regularly?" [*yes* or *no*]), whether they had health insurance ("Do you have health insurance?" [*yes* or *no*]), and whether they had at least one chronic disease ("Do you have at least 1 chronic disease?" [*yes* or *no*]), and descriptive statistics were reported for these measures.

Attention Check

To test whether respondents paid attention while completing the online survey, we asked the following: "If you are reading this, please do not answer this question and leave it blank." (1=*not at all true of me*, 2=*slightly true of me*, 3=*moderately true of me*, 4=*very true of me*, and 5=*extremely true of me*), adapted from the study by Chugani and Irwin [83]. We excluded respondents who failed the attention check.

Statistical Analysis

We used SPSS (version 29; IBM Corp) to analyze the data. We conducted 5 sets of statistical analyses outlined in the preregistration while controlling for 7 factors (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation). Specifically, we conducted 4 one-way analyses of covariance (ANCOVAs) using generation as an independent variable, 3 types of health information searches (HWIS, HGIS, and HMIS) and preventative health behavior (preventative health actions) as dependent variables, and 7 covariates. In addition, to test the effect of age (a continuous measure instead of 2 generations) on OHIS types (HWIS, HGIS, and HMIS) and preventative health actions, we performed 4 linear regression analyses using age as an independent variable, OHIS types (HWIS, HGIS, and HMIS) and preventative health actions as dependent variables, and 7 covariates.

To test whether search frequency moderated the effect of generation, we conducted 4 two-way ANCOVAs using generation and search frequency as independent variables, the 4 focal variables (HWIS, HGIS, HMIS, and preventative health actions) as dependent variables, and 7 control variables. To test whether gender moderated the effect of generation on the same 4 focal variables, we conducted 4 two-way ANCOVAs using generation and gender as independent variables and 7 control variables. In addition, to explore whether the presence of chronic diseases moderated the effect of generation on the 4 focal

XSL•FO
RenderX

variables, we conducted 4 two-way ANCOVAs using generation and the presence of chronic diseases as independent variables and 7 control variables.

To test the types of health information searches as mediators, we used the bootstrapping approach as applied in the SPSS PROCESS (version 3.4) macro [84]. We also performed a descriptive analysis of sociodemographic characteristics.

Results

Sociodemographic Characteristics of Respondents

Table 2 presents the detailed sociodemographic characteristics of the study sample by generation (digital natives and digital immigrants). The study sample comprised 49.52% (563/1137) female individuals and had an average age of 44.2 (SD 13.8) years. Most respondents were single (614/1137, 54%) and employed (838/1137, 73.7%). Approximately 31% (351/1137, 30.87%) of respondents had some college education or associate degrees, 13.1% (149/1137) were high school graduates, 40.37% (459/1137) had bachelor's degrees, and 15.04% (171/1137) had graduate degrees or higher. Approximately 11.96% (136/1137) of respondents earned <US $25,000 a year, 9.32% (106/1137) earned between US $25,000 and US $34,999 a year, 15.04% (171/1137) earned between US $35,000 and US $49,999 a year, 12.31% (140/1137) earned between US $50,000 and US $64,999 a year, 11.79% (134/1137) earned between US $65,000 and US $79,999 a year, 10.91% (124/1137) earned between US $80,000 and US $99,999 a year, and 28.67% (326/1137) earned ≥US $100,000 a year. Most respondents reported being White (931/1137, 81.88%) and of non-Hispanic or non-Latino ethnicity (1041/1137, 91.56%). In addition, 50.22% (571/1137) of respondents reported Democrat as their political affiliation, 91.47% (1040/1137) had health insurance, 67.37% (766/1137) went for regular annual checkups, and 42.22% (480/1137) had at least one chronic disease.

XSL•FO
RenderX

Table 2. Sociodemographic characteristics of the respondents by generation (N=1137).

Sociodemographic characteristic	Overall	Generation	
		Digital natives (n=571)	Digital immigrants (n=566)
Age (years), mean (SD)	44.2 (13.8)	32.5 (6.0)	56.0 (8.3)
Gender, n (%)			
Male	574 (50.5)	288 (50.4)	286 (50.5)
Female	563 (49.5)	283 (49.6)	280 (49.5)
Marital status, n (%)			
Married	523 (46)	211 (37)	312 (55.1)
Single	614 (54)	360 (63)	254 (44.9)
Employment, n (%)			
Unemployed	299 (26.3)	142 (24.9)	157 (27.7)
Employed	838 (73.7)	429 (75.1)	409 (72.3)
Income (US $), n (%)			
<25,000	136 (12)	70 (12.3)	66 (11.7)
25,000-34,999	106 (9.3)	55 (9.6)	51 (9)
35,000-49,999	171 (15)	82 (14.4)	89 (15.7)
50,000-64,999	140 (12.3)	86 (15.1)	54 (9.5)
65,000-79,999	134 (11.8)	64 (11.2)	70 (12.4)
80,000-99,999	124 (10.9)	71 (12.4)	53 (9.4)
≥100,000	326 (28.7)	143 (25)	183 (32.3)
Educational level, n (%)			
Lower than high school graduate	7 (0.6)	5 (0.9)	2 (0.4)
High school graduate	149 (13.1)	81 (14.2)	68 (12)
Some college or associate degree	351 (30.9)	167 (29.2)	184 (32.5)
Bachelor's degree	459 (40.4)	250 (43.8)	209 (36.9)
Graduate degree or higher	171 (15)	68 (11.9)	103 (18.2)
Race, n (%)			
American Indian or Alaska Native	8 (0.7)	7 (1.2)	1 (0.2)
Asian	68 (6)	52 (9.1)	16 (2.8)
Black or African American	97 (8.5)	41 (7.2)	56 (9.9)
Native Hawaiian or other Pacific Islander	2 (0.2)	2 (0.4)	0 (0)
White	931 (81.9)	445 (77.9)	486 (85.9)
Some other race	31 (2.7)	24 (4.2)	7 (1.2)
Ethnicity, n (%)			
Hispanic or Latino	96 (8.4)	67 (11.7)	29 (5.1)
Not Hispanic or Latino	1041 (91.6)	504 (88.3)	537 (94.9)
Political affiliation, n (%)			
Democrat	571 (50.2)	289 (50.6)	282 (49.8)
Republican	566 (49.8)	282 (49.4)	284 (50.2)
Do you go for an annual checkup regularly?, n (%)			
No	371 (32.6)	237 (41.5)	134 (23.7)
Yes	766 (67.4)	334 (58.5)	432 (76.3)
Do you have health insurance?, n (%)			

XSL•FO
RenderX

Sociodemographic characteristic	Overall	Generation	
		Digital natives (n=571)	Digital immigrants (n=566)
No	97 (8.5)	64 (11.2)	33 (5.8)
Yes	1040 (91.5)	507 (88.8)	533 (94.2)
Do you have at least one chronic disease?, n (%)			
No	657 (57.8)	395 (69.2)	262 (46.3)
Yes	480 (42.2)	176 (30.8)	304 (53.7)
Online health information search frequency, n (%)			
Frequent	615 (54.1)	279 (48.9)	336 (59.4)
Infrequent	522 (45.9)	292 (51.1)	230 (40.6)

Health Information Searches by Generation

Three 1-way ANCOVAs using 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation) were conducted to determine whether there were significant differences in the 3 types of OHIS between the 2 generations (digital natives and digital immigrants). Table 3 presents the ANCOVA results. The 2 generations differed significantly in the 3 types of OHIS behaviors. Specifically, digital natives engaged in significantly more (P<.001) searches for health wellness information compared with digital immigrants. In contrast, digital immigrants carried out considerably more searches for health guidance information (P<.001) and health management information (P=.001) compared with digital natives.

Table 3. Summary statistics of outcome variables as a function of 2 generations[a].

Outcome variable	Generation		F test (df)	P value	η^2
	Digital natives (n=571), mean (SD)	Digital immigrants (n=566), mean (SD)			
HWIS[b]	2.41 (0.75)	2.13 (0.82)	28.22 (1, 1128)	<.001	0.024
HGIS[c]	0.76 (0.78)	1.02 (0.81)	29.11 (1, 1128)	<.001	0.025
HMIS[d]	0.70 (0.66)	0.83 (0.61)	10.45 (1, 1128)	.001	0.009
PHA[e]	1.94 (0.89)	1.94 (0.90)	0.04 (1, 1128)	.85	0.000

[a]A total of 4 analyses of covariance were conducted using 7 covariates: information quality, trust in online health information, trust in health care institutions, search skills, race, ethnicity, and political affiliation.

[b]HWIS: health wellness information search.

[c]HGIS: health guidance information search.

[d]HMIS: health management information search.

[e]PHA: preventative health actions.

Preventative Health Actions by Generation

A 1-way ANCOVA using generation as an independent variable and the same 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation) was conducted to determine whether there were significant differences in preventative health actions between the 2 generations. Table 3 presents the ANCOVA results. The 2 generations—digital natives and digital immigrants—did not differ significantly in their steps toward preventative health actions (P=.85).

We also tested whether OHI searches mediated the generational effects on preventative health actions and conducted 3 mediation analyses (one for each type of OHIS, namely, HWIS, HGIS, and HMIS). We selected the PROCESS Macro Model 4 for mediation analysis (simple mediation), with 20,000 bootstraps and 95% CIs with generation as an independent variable, preventative health actions as a dependent variable, OHI search types as the mediator, and 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation). The relationship between generation and preventative health actions was mediated (as the 95% CI did not include 0 for all 3 OHIS types) by each type of OHIS. Table 4 presents the mediation analysis results.

XSL•FO
RenderX

Table 4. Mediation analysis results[a].

Indirect effect	Coefficient (SE)	Bootstrapping 95% CI
HWIS[b]		
Total indirect effect	–0.08 (0.02)	–0.12 to –0.05
Generation and HWIS	–0.25 (0.05)	–0.35 to –0.16
HWIS and PHA[c]	0.31 (0.03)	0.25 to 0.38
HGIS[d]		
Total indirect effect	0.08 (0.02)	0.05 to 0.11
Generation and HGIS	0.26 (0.05)	0.17 to 0.36
HGIS and PHA	0.29 (0.03)	0.23 to 0.36
HMIS[e]		
Total indirect effect	0.06 (0.02)	0.02 to 0.11
Generation and HMIS	0.12 (0.04)	0.05 to 0.20
HMIS and PHA	0.52 (0.04)	0.44 to 0.59

[a]A total of 3 mediation analyses (PROCESS Macro Model 4) were conducted using 7 covariates: information quality, trust in online health information, trust in health care institutions, search skills, race, ethnicity, and political affiliation.

[b]HWIS: health wellness information search.

[c]PHA: preventative health actions.

[d]HGIS: health guidance information search.

[e]HMIS: health management information search.

Age as a Predictor of Health Information Searches and Preventative Health Actions

To test the effect of age on OHIS types (HWIS, HGIS, and HMIS) and preventative health actions, we conducted 4 linear regression analyses using age as an independent variable, OHIS types (HWIS, HGIS, and HMIS) and preventative health actions as dependent variables, and 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation). Table 5 presents the linear regression results. Age was a significant negative predictor of HWIS behavior; as people aged, they searched for significantly less health wellness information ($P<.001$). In contrast, age was a significant positive predictor of HGIS and HMIS behaviors; as people aged, they searched for significantly more health guidance information ($P<.001$) and health management information ($P=.003$). However, age was not a significant predictor of preventative health actions ($P=.48$).

XSL•FO
RenderX

Table 5. Linear regression results.

	Unstandardized coefficient (SE)	*t* test	*P* value
HWIS [a,b]			
Age (years)	–0.01 (0.002)	–7.69	<.001
Information quality	–0.04 (0.03)	–1.33	.18
Trust in online health information	0.01 (0.03)	0.40	.69
Trust in health care institutions	0.01 (0.02)	0.27	.78
Search skills	0.01 (0.02)	0.62	.53
Race	0.06 (0.02)	2.91	.004
Ethnicity	–0.11 (0.09)	–1.29	.20
Political affiliation	0.05 (0.05)	1.15	.25
Constant	2.95 (0.21)	14.13	<.001
HGIS[c,d]			
Age (years)	0.01 (0.002)	5.08	<.001
Information quality	–0.04 (0.03)	–1.13	.26
Trust in online health information	0.02 (0.03)	0.67	.51
Trust in health care institutions	0.01 (0.02)	0.59	.56
Search skills	–0.01 (0.02)	–0.64	.53
Race	0.04 (0.02)	1.52	.13
Ethnicity	0.03 (0.09)	0.32	.75
Political affiliation	0.01 (0.05)	0.24	.81
Constant	0.43 (0.22)	2.00	.045
HMIS[e,f]			
Age (years)	0.004 (0.001)	2.96	.003
Information quality	0.01 (0.03)	0.47	.64
Trust in online health information	–0.06 (0.03)	–2.20	.03
Trust in health care institutions	0.02 (0.02)	1.21	.23
Search skills	–0.02 (0.02)	–0.99	.32
Race	0.02 (0.02)	1.25	.21
Ethnicity	0.08 (0.07)	1.11	.27
Political affiliation	–0.01 (0.04)	–0.26	.79
Constant	0.56 (0.17)	3.25	.001
PHA[g,h]			
Age (years)	–0.001 (0.002)	–0.71	.48
Information quality	–0.004 (0.04)	–0.10	.92
Trust in online health information	–0.04 (0.04)	–1.24	.22
Trust in health care institutions	–0.02 (0.02)	–0.78	.44
Search skill	–0.04 (0.02)	–1.69	.09
Race	0.02 (0.03)	0.77	.44
Ethnicity	0.09 (0.10)	0.95	.35
Political affiliation	–0.07 (0.05)	–1.27	.21
Constant	2.24 (0.24)	9.37	<.001

[a]HWIS: health wellness information search.

XSL•FO
RenderX

[b]Model summary: R^2=0.07; $F_{8,1128}$=11.34; P<.001.

[c]HGIS: health guidance information search.

[d]Model summary: R^2=0.03; $F_{8,1128}$=4.00; P<.001.

[e]HMIS: health management information search.

[f]Model summary: R^2=0.02; $F_{8,1128}$=2.75; P=.005.

[g]PHA: preventative health actions.

[h]Model summary: R^2=0.02; $F_{8,1128}$=3.01; P=.002.

Search Frequency as a Moderator

To examine whether the frequency of OHIS (frequent vs infrequent) moderated the effect of generation on OHIS type and preventative health actions, we conducted four 2-way ANCOVAs using generation, search frequency, and interaction term as independent variables and 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation). Table 6 presents the ANCOVA results. As expected, frequent OHI seekers, compared with infrequent OHI seekers, performed significantly more searches for all 3 types of health information (P<.001 for HWIS, HGIS, and HMIS) and engaged in significantly more preventative health behaviors (preventative health actions: P<.001). However, the interaction effects of generation and search frequency were not significant for all analyses. Thus, the frequency of OHIS did not moderate the relationship between generation and both type of OHIS and preventative health actions.

XSL•FO
RenderX

Table 6. Summary statistics of outcome variables as a function of 2 generations by online health information (OHI) search frequency[a].

Outcome variables	Frequent, mean (SD)	Infrequent, mean (SD)	*F* test (*df*)	*P* value	η^2
HWIS[b]					
Generation	N/A[c]	N/A	45.24 (1, 1126)	<.001	0.04
OHI search frequency	N/A	N/A	100.99 (1, 1126)	<.001	0.08
Generation × OHI search frequency	N/A	N/A	2.18 (1, 1126)	.14	0.002
Digital natives	2.62 (0.62)	2.22 (0.82)	N/A	N/A	N/A
Digital immigrants	2.34 (0.72)	1.81 (0.86)	N/A	N/A	N/A
HGIS[d]					
Generation	N/A	N/A	18.37 (1, 1126)	<.001	0.02
OHI search frequency	N/A	N/A	200.91 (1, 1126)	<.001	0.15
Generation × OHI search frequency	N/A	N/A	0.59 (1, 1126)	.44	0.001
Digital natives	1.10 (0.78)	0.44 (0.64)	N/A	N/A	N/A
Digital immigrants	1.26 (0.78)	0.67 (0.73)	N/A	N/A	N/A
HMIS[e]					
Generation	N/A	N/A	3.68 (1, 1126)	.06	0.003
OHI search frequency	N/A	N/A	211.86 (1, 1126)	<.001	0.16
Generation × OHI search frequency	N/A	N/A	0.35 (1, 1126)	.56	0.000
Digital natives	0.97 (0.64)	0.45 (0.57)	N/A	N/A	N/A
Digital immigrants	1.03 (0.56)	0.53 (0.55)	N/A	N/A	N/A
PHA[f]					
Generation	N/A	N/A	2.25 (1, 1126)	.13	0.002
OHI search frequency	N/A	N/A	125.21 (1, 1126)	<.001	0.10
Generation × OHI search frequency	N/A	N/A	0.50 (1, 1126)	.48	0.000
Digital natives	2.22 (0.78)	1.68 (0.90)	N/A	N/A	N/A
Digital immigrants	2.19 (0.80)	1.57 (0.92)	N/A	N/A	N/A

[a]A total of 4 analyses of covariance were conducted using 7 covariates: information quality, trust in OHI, trust in health care institutions, search skills, race, ethnicity, and political affiliation.

[b]HWIS: health wellness information search.

[c]N/A: not applicable.

[d]HGIS: health guidance information search.

[e]HMIS: health management information search.

[f]PHA: preventative health actions.

Gender as a Moderator

To examine whether gender (male vs female) moderated the effect of generation on OHIS type (HWIS, HGIS, and HMIS) and preventative health actions, we conducted four 2-way ANCOVAs using generation, gender, and interaction term as independent variables and 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation). Table 7 presents the ANCOVA results. The interaction effect of generation and gender was significant only for HGIS (*P*=.02); among digital natives, female individuals, compared with male individuals, searched for significantly more health guidance information, whereas no significant differences among digital immigrants were observed between male and female individuals for HGIS. Moreover, the interaction effect of generation and gender was not significant for HWIS, HMIS, or preventative health actions. Thus, gender only moderated the relationship between generation and HGIS.

XSL•FO
RenderX

Table 7. Summary statistics of outcome variables as a function of 2 generations by gender[a].

Outcome variables	Male, mean (SD)	Female, mean (SD)	*F* test (*df*)	*P* value	η^2
HWIS[b]					
Generation	N/A[c]	N/A	28.68 (1, 1126)	<.001	0.03
Gender	N/A	N/A	8.54 (1, 1126)	.004	0.01
Generation × gender	N/A	N/A	0.00 (1, 1126)	>.99	0.000
Digital natives	2.49 (0.73)	2.34 (0.77)	N/A	N/A	N/A
Digital immigrants	2.19 (0.85)	2.06 (0.80)	N/A	N/A	N/A
HGIS[d]					
Generation	N/A	N/A	29.71 (1, 1126)	<.001	0.03
Gender	N/A	N/A	12.18 (1, 1126)	<.001	0.01
Generation × gender	N/A	N/A	5.38 (1, 1126)	.02	0.01
Digital natives	0.63 (0.78)	0.90 (0.77)	N/A	N/A	N/A
Digital immigrants	0.99 (0.79)	1.05 (0.83)	N/A	N/A	N/A
HMIS[e]					
Generation	N/A	N/A	10.69 (1, 1126)	.001	0.01
Gender	N/A	N/A	10.51 (1, 1126)	.001	0.01
Generation × gender	N/A	N/A	1.65 (1, 1126)	.20	0.001
Digital natives	0.62 (0.66)	0.79 (0.66)	N/A	N/A	N/A
Digital immigrants	0.79 (0.60)	0.86 (0.61)	N/A	N/A	N/A
PHA[f]					
Generation	N/A	N/A	0.04 (1, 1126)	.85	0.000
Gender	N/A	N/A	1.13 (1, 1126)	.29	0.001
Generation × gender	N/A	N/A	0.69 (1, 1126)	.41	0.001
Digital natives	1.90 (0.89)	1.99 (0.88)	N/A	N/A	N/A
Digital immigrants	1.93 (0.91)	1.95 (0.89)	N/A	N/A	N/A

[a]A total of 4 analyses of covariance were conducted using 7 covariates: information quality, trust in online health information, trust in health care institutions, search skills, race, ethnicity, and political affiliation.

[b]HWIS: health wellness information search.

[c]N/A: not applicable.

[d]HGIS: health guidance information search.

[e]HMIS: health management information search.

[f]PHA: preventative health actions.

Presence of Chronic Diseases as a Moderator

To examine whether the presence of chronic diseases (yes vs no) moderated the effect of generation on OHIS types (HWIS, HGIS, and HMIS) and preventative health actions, we conducted four 2-way ANCOVAs using generation, presence of chronic diseases, and interaction term as independent variables and 7 covariates (quality of OHI, trust in OHI, trust in health care institutions, OHI search skills, race, ethnicity, and political affiliation). Table 8 presents the ANCOVA results. The interaction effects of generation and presence of chronic diseases were significant only for HWIS (*P*=.03). Specifically, digital natives with at least one chronic disease searched for more health wellness information than those without chronic diseases, whereas no significant differences among digital immigrants were observed between the presence or absence of a chronic disease for HWIS. Moreover, the interaction effects of generation and presence of chronic diseases were not significant for HGIS, HMIS, or preventative health actions. Thus, the presence of chronic diseases only moderated the relationship between generation and HWIS.

XSL•FO
RenderX

Table 8. Summary statistics of outcome variables as a function of 2 generations by presence of chronic diseases[a].

Outcome variables	Chronic disease (no), mean (SD)	Chronic disease (yes), mean (SD)	*F* test (*df*)	*P* value	η^2
HWIS[b]					
Generation	N/A[c]	N/A	33.26 (1, 1126)	<.001	0.03
Presence of chronic diseases	N/A	N/A	1.95 (1, 1126)	.16	0.002
Generation × presence of chronic diseases	N/A	N/A	5.06 (1, 1126)	.03	0.004
Digital natives	2.36 (0.77)	2.53 (0.70)	N/A	N/A	N/A
Digital immigrants	2.15 (0.85)	2.11 (0.81)	N/A	N/A	N/A
HGIS[d]					
Generation	N/A	N/A	7.83 (1, 1126)	.01	0.01
Presence of chronic diseases	N/A	N/A	124.97 (1, 1126)	<.001	0.10
Generation × presence of chronic diseases	N/A	N/A	0.89 (1, 1126)	.35	0.001
Digital natives	0.58 (0.72)	1.16 (0.78)	N/A	N/A	N/A
Digital immigrants	0.76 (0.79)	1.24 (0.77)	N/A	N/A	N/A
HMIS[e]					
Generation	N/A	N/A	0.62 (1, 1126)	.43	0.001
Presence of chronic diseases	N/A	N/A	112.81 (1, 1126)	<.001	0.09
Generation × presence of chronic diseases	N/A	N/A	0.28 (1, 1126)	.60	0.000
Digital natives	0.57 (0.65)	1.00 (0.60)	N/A	N/A	N/A
Digital immigrants	0.62 (0.57)	1.00 (0.59)	N/A	N/A	N/A
PHA[f]					
Generation	N/A	N/A	1.50 (1, 1126)	.22	0.001
Presence of chronic diseases	N/A	N/A	14.88 (1, 1126)	<.001	0.01
Generation × presence of chronic diseases	N/A	N/A	1.30 (1, 1126)	.26	0.001
Digital natives	1.85 (0.88)	2.15 (0.86)	N/A	N/A	N/A
Digital immigrants	1.86 (0.85)	2.01 (0.94)	N/A	N/A	N/A

[a]A total of 4 analyses of covariance were conducted using 7 covariates: information quality, trust in online health information, trust in health care institutions, search skills, race, ethnicity, and political affiliation.

[b]HWIS: health wellness information search.

[c]N/A: not applicable.

[d]HGIS: health guidance information search.

[e]HMIS: health management information search.

[f]PHA: preventative health actions.

Discussion

Principal Findings

In this study, we examined how the experiences with technology of 2 generations (digital natives and digital immigrants) affected their OHIS and preventative health behaviors using 7 control variables (information quality, trust in OHI, trust in health care institutions, search skills, race, ethnicity, and political affiliation). We found that digital natives searched more for health wellness information than digital immigrants, whereas digital immigrants searched more for health management and health guidance information than digital natives. However, after seeking OHI, the 2 generations did not differ significantly in their steps toward preventative health actions. We also found that OHI search types served as a mediator between generational differences and preventive health outcomes, whereas the search frequency of OHI did not moderate the relationship between generation and type of OHIS. In addition, gender only moderated the relationship between generation and HGIS; in contrast, the

XSL•FO
RenderX

presence of chronic diseases only moderated the relationship between generation and HWIS. Further analysis (using age as a continuous measure) revealed that, as respondents' ages increased, they sought significantly less health and wellness information but significantly more health management and health guidance information. However, age was not found to be a significant predictor of preventative health actions.

Comparison With Prior Work

Previous research on OHIS has found that younger people are more prone to seeking OHI than older people [2-7]. The findings of this study are consistent with those of previous research indicating that younger generations (digital natives; aged between 18 and 42 years) are more likely to search for OHI than older generations (digital immigrants; aged ≥43 years). In addition, this study revealed generational differences in the types of OHI search behaviors.

In total, 3 types of OHIS (HWIS, HGIS, and HMIS) were used to analyze generational differences in search behavior. The findings revealed that digital natives (the younger generation) conducted more HWISs than digital immigrants (the older generation). The opposite was true of HGIS and HMIS—digital immigrants (the older generation) conducted more of these searches than digital natives (the younger generation). Some of these findings confirm previous research, whereas others contradict it. For example, research has shown that older adults tend to search for health information related to specific diseases, medications, and treatment guidance [9]. Similarly, this study found that digital immigrants conducted more HGI searches (namely, medication guidance and disease consulting) than digital natives. In addition, the results revealed that digital immigrants conducted more HMISs (such as managing health conditions and participating in online support groups) than digital natives. However, unlike previous research indicating that older adults tend to seek information related to nutrition and exercise [9], this study found that younger generations conducted more HWISs (namely, sports and fitness, nutrition and diet, and general health knowledge). These findings deviate from those of previous research, which shows that young adults seek health information online by joining social media groups and following pages relevant to their condition [44].

Previous findings have highlighted that OHIS is associated with positive health behaviors and health outcomes [85-91]. Although digital natives and digital immigrants did not differ in their preventative behaviors, this study found that the type of health information searches drove the effect of generational differences on preventive health outcomes by demonstrating the indirect link between OHIS and preventative health behaviors after retrieving all 3 types of OHI. This finding is significant as both generations reported engaging in preventive actions after seeking specific types of OHI. In addition, previous studies have established that the frequency of OHIS is higher among younger people [3,7,67]. However, the frequency of OHIS did not moderate the relationship between younger and older generations and types of OHI searches. The findings indicate that both frequent and infrequent OHI searchers from both generations in our sample conducted similar levels of 3 types of health information searches online, a departure from previous findings.

Furthermore, previous research has shown gender to be a significant predictor of OHIS behavior [92]. However, our analysis shows that gender only moderated the relationship between generation and searches for health guidance information (HGIS) but not searches related to health management information (HMIS) and health wellness information (HWIS). Similar to previous studies showing the association between people facing a health threat and health information–seeking behavior [93], we also found that the presence of at least one chronic disease moderated the relationship between generation and searches for health guidance information (HGIS) but not searches related to health management information (HMIS) and health wellness information (HWIS).

Although existing studies show that age, gender, ethnic background, political affiliation, and health conditions are significant predictors of OHIS behavior [3,7,92-96], there is a lack of research on whether generational differences affect the types of OHIS behavior or preventative behavior afterward [9]. Thus, our findings add to the OHIS behavior literature by illustrating that a single factor—generational cohorts that differ in their level of technology immersion—influences information-seeking and preventative health behaviors. This study furthers the understanding of OHI search behavior by differentiating between types of OHI search behavior across 2 generations.

Limitations and Future Directions

This research is limited in several ways, indicating directions for future inquiry. First, we used stratified sampling to recruit 2 generations of American adults using an online data collection platform (Prolific). Specifically, we collected a stratified sample based on age, sex, and political affiliation using the United States as the location criterion. Although online data collection methods have many advantages, they are less likely to be used by those who do not have internet access. Thus, our findings are limited to digital natives and immigrants in this population sample. Future research could use in-person recruitment and focus groups to test the differences in OHIS and preventative health behaviors among different generations. Although we ensured an equal number of respondents in 2 digital generation groups (digital natives vs digital immigrants) with an even distribution of gender (male and female) and political affiliation (Republican and Democratic) and used 7 relevant control variables (information quality, trust in OHI, trust in health care institutions, search skills, race, ethnicity, and political affiliation) to show the effect of 2 generations on OHIS and preventative health behaviors, future research could use a more representative sample to test the validity of the relationships.

Second, this is the first study to our knowledge that has examined the differences in health information–seeking behaviors among older and younger Americans. This study focused on 2 generations that differ in their perception and use of technology: digital natives (born from 1980 onward) and digital immigrants (born before 1980). Although the use of 2 generational categories can simplify comparisons based on the ubiquity of technology in formative years, it has the potential to attenuate key variations within generational cohorts. Although our sample included nearly an equal number of digital natives

XSL•FO
RenderX

and immigrants by age, sex, and political affiliation, most of our respondents were White, non-Hispanic or Latino Americans and skewed toward higher income brackets (nearly 40% made <US $50,000 a year; 413/1137, 36.32%) and educational levels (approximately 55% had a bachelor's degree or higher; 630/1137, 55.41%). As such, our findings may overstate differences based on 2 generational cohorts, and future research could focus on recruiting more racial and ethnic minority groups and low-income individuals to assess the validity of our findings.

Third, we performed additional analyses using age as a continuous variable. Although using respondents' age as a continuous variable shows results similar to those of categorizing them by generation, future research should test whether OHIS and preventative health behaviors vary among different generations, including microgenerations (such as Zillennials [14]). Thus, comparing how different generations in Western and non-Western countries differ in their OHIS and preventative health behaviors would be an exciting avenue to pursue. Moreover, technological exposure and norms [9] are likely to vary within 2 generations (digital immigrants and digital natives). Although we controlled for respondents' self-reported online information search skills, other factors could be potential barriers that older adults face during OHIS, including self-efficacy and experience with information and communications technologies, perceived ease of use, and subjective norm perception [9,91,94].

Fourth, the statistical models used in the analyses did not account for some key variables. For example, previous research indicates that factors such as (but not limited to) health consumers' perceived trust in health care professionals [97], ethnic background [3,94], and political ideology [95,96] are important predictors of OHIS and health behaviors. In addition, health information–seeking behavior has been shown to vary according to the perception of the credibility of information sources [34,35,64,65]. As our research did not account for these differences, the models used for the analyses may overstate differences based on generational cohorts.

Fifth, this study used self-reported OHIS and preventative health behavior measures, which might include biased responses, such as recall bias and social desirability bias [98,99]. Future research could benefit from using objective measures rather than relying on self-reports. For example, logged data of the OHI search history and daily diary study designs [100] would allow future researchers to measure types of health information searches and engagement in preventative health behaviors.

Finally, recent research suggests that some disadvantaged groups (older people and individuals of a lower socioeconomic status) lack the skills to use OHI and are often digitally marginalized [101]. Although this study did not explore digital marginalization, future research should address how lower technology skills might limit socially disadvantaged groups from participating in OHIS and managing their health.

Conclusions

This study used a generational cohort lens to examine how older and younger generations differ in their OHI search and preventative health behaviors. This study provides insights into differences in OHI search behavior between the 2 digital generations, suggesting that, compared with digital immigrants, digital natives tend to search for more wellness information and are less likely to search for health management or guidance information. Furthermore, this research identified that the types of OHI searches mediated the generational effect on preventive health outcomes. Although the study of OHI search behavior is a well-established research domain, older adults' OHI searches and preventative health behaviors are poorly understood [9]. As the older adult population and the demand for health care and online medical information increase, it is important to understand older adults' OHIS and preventative health behaviors. This research could also serve as a foundation for future studies to understand how these results may apply across various population groups, including disadvantaged and digitally marginalized groups.

Acknowledgments

The authors thank the study respondents. This research was supported in part by the Healthcare MBA Faculty Research Award. The funders had no role in any part of the study (conceptualization, design, analyses, and writing of the manuscript).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions.
[DOCX File , 20 KB - jmir_v26i1e48977_app1.docx]

References

1. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. Cogent Social Sciences 2017 Mar 13;3(1):1302785. [doi: 10.1080/23311886.2017.1302785]

XSL•FO
RenderX

2. Brodie M, Flournoy RE, Altman DE, Blendon RJ, Benson JM, Rosenbaum MD. Health information, the internet, and the digital divide. Health Aff (Millwood) 2000 Nov 01;19(6):255-265. [doi: 10.1377/hlthaff.19.6.255] [Medline: 11192412]
3. Din HN, McDaniels-Davidson C, Nodora J, Madanat H. Profiles of a health information-seeking population and the current digital divide: cross-sectional analysis of the 2015-2016 California health interview survey. J Med Internet Res 2019 May 14;21(5):e11931 [FREE Full text] [doi: 10.2196/11931] [Medline: 31094350]
4. Levy H, Janke AT, Langa KM. Health literacy and the digital divide among older Americans. J Gen Intern Med 2015 Mar;30(3):284-289 [FREE Full text] [doi: 10.1007/s11606-014-3069-5] [Medline: 25387437]
5. Nguyen A, Mosadeghi S, Almario CV. Persistent digital divide in access to and use of the internet as a resource for health information: results from a California population-based study. Int J Med Inform 2017 Jul;103:49-54. [doi: 10.1016/j.ijmedinf.2017.04.008] [Medline: 28551001]
6. Rao N, Tighe EL, Feinberg I. The dispersion of health information-seeking behavior and health literacy in a state in the Southern United States: cross-sectional study. JMIR Form Res 2022 Jun 15;6(6):e34708 [FREE Full text] [doi: 10.2196/34708] [Medline: 35704357]
7. Weber W, Reinhardt A, Rossmann C. Lifestyle segmentation to explain the online health information-seeking behavior of older adults: representative telephone survey. J Med Internet Res 2020 Jun 12;22(6):e15099 [FREE Full text] [doi: 10.2196/15099] [Medline: 32530433]
8. Frey WH. What the 2020 census will reveal about America: stagnating growth, an aging population, and youthful diversity. Brookings. 2021. URL: https://www.brookings.edu/research/what-the-2020-census-will-reveal-about-america-stagnating-growth-an-aging-population-and-youthful-diversity/ [accessed 2022-11-01]
9. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790 [FREE Full text] [doi: 10.2196/34790] [Medline: 35171099]
10. Berkowitz EN, Schewe CD. Generational cohorts hold the key to understanding patients and health care providers: coming-of-age experiences influence health care behaviors for a lifetime. Health Mark Q 2011 Apr 16;28(2):190-204. [doi: 10.1080/07359683.2011.572029] [Medline: 21590564]
11. Ma X, Liu Y, Zhang P, Qi R, Meng F. Understanding online health information seeking behavior of older adults: a social cognitive perspective. Front Public Health 2023 Mar 3;11:1147789 [FREE Full text] [doi: 10.3389/fpubh.2023.1147789] [Medline: 36935731]
12. Ryder NB. The cohort as a concept in the study of social change. Am Sociol Rev 1965 Dec;30(6):843. [doi: 10.2307/2090964]
13. Wyatt S. Technological determinism is dead; long live technological determinism. In: Hackett EJ, Amsterdamska O, Lynch M, Wajcman J, editors. The Handbook of Science and Technology Studies. 3rd edition. Cambridge, MA: The MIT Press; 2008:165-180.
14. Ward T. Zillennials: the newest micro-generation has a name. CNN Health. 2023. URL: https://www.cnn.com/2023/05/10/health/what-are-zillennials-wellness/index [accessed 2024-01-29]
15. Twenge JM. Generations: the real differences between Gen Z, Millennials, Gen X, Boomers, and Silents—and what they mean for America's future. Perspect Sci Christ Faith 2023 Dec;75(3):212-214. [doi: 10.56315/pscf12-23Twenge]
16. Prensky M. Digital natives, digital immigrants part 1. Horiz 2001 Sep;9(5):1-6. [doi: 10.1108/10748120110424816]
17. Prensky M. Digital natives, digital immigrants part 2: do they really think differently? Horiz 2001 Nov;9(6):1-6. [doi: 10.1108/10748120110424843]
18. Ransdell S, Kent B, Gaillard-Kenney S, Long J. Digital immigrants fare better than digital natives due to social reliance. Br J Educ Technol 2010;42(6):931-938 [FREE Full text] [doi: 10.1111/j.1467-8535.2010.01137.x]
19. Akçayır M, Dündar H, Akçayır G. What makes you a digital native? Is it enough to be born after 1980? Comput Human Behav 2016 Jul;60:435-440. [doi: 10.1016/j.chb.2016.02.089]
20. Kopalle PK, Kumar V, Subramaniam M. How legacy firms can embrace the digital ecosystem via digital customer orientation. J Acad Mark Sci 2019 Nov 20;48(1):114-131 [FREE Full text] [doi: 10.1007/s11747-019-00694-2]
21. Baham C, Kalgotra P, Nasirpouri Shadbad F, Sharda R. Generational differences in handling technology interruptions: a qualitative study. Eur J Inf Syst 2022 May 10;32(5):858-878. [doi: 10.1080/0960085X.2022.2070557]
22. Kesharwani A. Do (how) digital natives adopt a new technology differently than digital immigrants? A longitudinal study. Inf Manag 2020 Mar;57(2):103170 [FREE Full text] [doi: 10.1016/j.im.2019.103170]
23. Lai KW, Hong KS. Technology use and learning characteristics of students in higher education: do generational differences exist? Br J Educ Technol 2014 May 02;46(4):725-738 [FREE Full text] [doi: 10.1111/bjet.12161]
24. Martins P, Picoto WN, Bélanger F. The effects of digital nativity on nonvolitional routine and innovative usage. Inf Technol People 2022 Nov 08;36(7):2804-2825 [FREE Full text] [doi: 10.1108/itp-12-2021-0971]
25. Vodanovich S, Sundaram D, Myers M. Research commentary—digital natives and ubiquitous information systems. Inf Syst Res 2010 Dec;21(4):711-723 [FREE Full text] [doi: 10.1287/isre.1100.0324]
26. Colbert A, Yee N, George G. The digital workforce and the workplace of the future. Acad Manage J 2016 Jun;59(3):731-739 [FREE Full text] [doi: 10.5465/amj.2016.4003]
27. Filho EJ, Gammarano ID, Barreto IA. Technology-driven consumption: digital natives and immigrants in the context of multifunctional convergence. J Strateg Mark 2019 Aug 22;29(3):181-205. [doi: 10.1080/0965254x.2019.1656278]

XSL•FO
RenderX

28. Bennett S, Maton K, Kervin L. The 'digital natives' debate: a critical review of the evidence. Br J Educ Technol 2008 Aug 19;39(5):775-786. [doi: 10.1111/j.1467-8535.2007.00793.x]
29. Tams S. Helping older workers realize their full organizational potential: a moderated mediation model of age and it-enabled task performance (Open Access). MIS Q 2022 Mar 01;46(1):1-34 [FREE Full text] [doi: 10.25300/MISQ/2022/16359]
30. Miller LM, Bell RA. Online health information seeking: the influence of age, information trustworthiness, and search challenges. J Aging Health 2012 Apr;24(3):525-541. [doi: 10.1177/0898264311428167] [Medline: 22187092]
31. Pang PC, Chang S, Pearce J, Verspoor K. Online health information seeking behaviour: understanding different search approaches. In: proceedings of the 18th Pacific Asia Conference on Information Systems. 2014 Presented at: PACIS '14; June 24-28, 2014; Chengdu, China p. 229 URL: https://aisel.aisnet.org/pacis2014/229
32. Faverio M. Share of those 65 and older who are tech users has grown in the past decade. Pew Research Center. URL: https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/ [accessed 2022-10-27]
33. Kennedy GE, Judd TS, Churchward A, Gray K, Krause KL. First year students' experiences with technology: are they really digital natives? Australas. J Educ Technol 2008 Jan 16;24(1):1233 [FREE Full text] [doi: 10.14742/ajet.1233]
34. Feufel MA, Stahl SF. What do web-use skill differences imply for online health information searches? J Med Internet Res 2012 Jun 13;14(3):e87 [FREE Full text] [doi: 10.2196/jmir.2051] [Medline: 22695686]
35. Park E, Kwon M. Health-related internet use by children and adolescents: systematic review. J Med Internet Res 2018 Apr 03;20(4):e120 [FREE Full text] [doi: 10.2196/jmir.7731] [Medline: 29615385]
36. Zajac IT, Flight IH, Wilson C, Turnbull D, Cole S, Young G. Internet usage and openness to internet-delivered health information among Australian adults aged over 50 years. Australas Med J 2012;5(5):262-267 [FREE Full text] [doi: 10.4066/AMJ.2012.1065] [Medline: 22848321]
37. El Sherif R, Pluye P, Ibekwe F. Contexts and outcomes of proxy online health information seeking: mixed studies review with framework synthesis. J Med Internet Res 2022 Jun 24;24(6):e34345 [FREE Full text] [doi: 10.2196/34345] [Medline: 35749210]
38. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. The influence of online health information on health decisions: a systematic review. Patient Educ Couns 2021 Apr;104(4):770-784. [doi: 10.1016/j.pec.2020.11.016] [Medline: 33358253]
39. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
40. Ramsey I, Corsini N, Peters MD, Eckert M. A rapid review of consumer health information needs and preferences. Patient Educ Couns 2017 Sep;100(9):1634-1642. [doi: 10.1016/j.pec.2017.04.005] [Medline: 28442155]
41. Griffin RJ, Dunwoody S, Neuwirth K. Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. Environ Res 1999 Feb;80(2 Pt 2):S230-S245. [doi: 10.1006/enrs.1998.3940] [Medline: 10092438]
42. Madrigal L, Escoffery C. Electronic health behaviors among US adults with chronic disease: cross-sectional survey. J Med Internet Res 2019 Mar 05;21(3):e11240 [FREE Full text] [doi: 10.2196/11240] [Medline: 30835242]
43. Xiong Z, Zhang L, Li Z, Xu W, Zhang Y, Ye T. Frequency of online health information seeking and types of information sought among the general Chinese population: cross-sectional study. J Med Internet Res 2021 Dec 02;23(12):e30855 [FREE Full text] [doi: 10.2196/30855] [Medline: 34860676]
44. Lim MS, Molenaar A, Brennan L, Reid M, McCaffrey T. Young adults' use of different social media platforms for health information: insights from web-based conversations. J Med Internet Res 2022 Jan 18;24(1):e23656 [FREE Full text] [doi: 10.2196/23656] [Medline: 35040796]
45. Renahy E, Parizot I, Chauvin P. WHIST: a web-based survey on health information seeking on Internet in France, 2007. AMIA Annu Symp Proc 2007 Oct 11:1090-1091. [Medline: 18694188]
46. Baumann E, Czerwinski F, Reifegerste D. Gender-specific determinants and patterns of online health information seeking: results from a representative German health survey. J Med Internet Res 2017 Apr 04;19(4):e92 [FREE Full text] [doi: 10.2196/jmir.6668] [Medline: 28377367]
47. Lee K, Hoti K, Hughes JD, Emmerton LM. Consumer use of "Dr Google": a survey on health information-seeking behaviors and navigational needs. J Med Internet Res 2015 Dec 29;17(12):e288 [FREE Full text] [doi: 10.2196/jmir.4345] [Medline: 26715363]
48. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. J Med Internet Res 2022 Nov 02;24(11):e42447 [FREE Full text] [doi: 10.2196/42447] [Medline: 36322124]
49. Rooks RN, Wiltshire JC, Elder K, BeLue R, Gary LC. Health information seeking and use outside of the medical encounter: is it associated with race and ethnicity? Soc Sci Med 2012 Jan;74(2):176-184. [doi: 10.1016/j.socscimed.2011.09.040] [Medline: 22154611]
50. Richardson A, Allen JA, Xiao H, Vallone D. Effects of race/ethnicity and socioeconomic status on health information-seeking, confidence, and trust. J Health Care Poor Underserved 2012 Nov;23(4):1477-1493. [doi: 10.1353/hpu.2012.0181] [Medline: 23698662]

XSL•FO
RenderX

51. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625 [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
52. Rooks RN, Kapral CG, Mathis AL. Chronic conditions may be more important than race or ethnicity in relation to health information seeking and use. J Aging Health 2019 Apr;31(4):611-630. [doi: 10.1177/0898264317744643] [Medline: 29254425]
53. Kannan VD, Veazie PJ. Political orientation, political environment, and health behaviors in the United States. Prev Med 2018 Sep;114:95-101. [doi: 10.1016/j.ypmed.2018.06.011] [Medline: 29940293]
54. Imburgia TM, Hendrix KS, Donahue KL, Sturm LA, Zimet GD. Predictors of influenza vaccination in the U.S. among children 9-13 years of age. Vaccine 2017 Apr 25;35(18):2338-2342. [doi: 10.1016/j.vaccine.2017.03.060] [Medline: 28359619]
55. Naeim A, Baxter-King R, Wenger N, Stanton AL, Sepucha K, Vavreck L. Effects of age, gender, health status, and political party on COVID-19-related concerns and prevention behaviors: results of a large, longitudinal cross-sectional survey. JMIR Public Health Surveill 2021 Apr 28;7(4):e24277 [FREE Full text] [doi: 10.2196/24277] [Medline: 33908887]
56. Gyorda JA, Lekkas D, Price G, Jacobson NC. Evaluating the impact of mask mandates and political party affiliation on mental health internet search behavior in the united states during the COVID-19 pandemic: generalized additive mixed model framework. J Med Internet Res 2023 Mar 03;25:e40308 [FREE Full text] [doi: 10.2196/40308] [Medline: 36735836]
57. Prestin A, Vieux SN, Chou WY. Is online health activity alive and well or flatlining? Findings from 10 years of the health information national trends survey. J Health Commun 2015;20(7):790-798. [doi: 10.1080/10810730.2015.1018590] [Medline: 26042588]
58. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002 May 22;287(20):2691-2700. [doi: 10.1001/jama.287.20.2691] [Medline: 12020305]
59. Starcevic V, Berle D. Cyberchondria: towards a better understanding of excessive health-related internet use. Expert Rev Neurother 2013 Feb;13(2):205-213. [doi: 10.1586/ern.12.162] [Medline: 23368807]
60. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 07;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
61. Antheunis ML, Tates K, Nieboer TE. Patients' and health professionals' use of social media in health care: motives, barriers and expectations. Patient Educ Couns 2013 Sep;92(3):426-431. [doi: 10.1016/j.pec.2013.06.020] [Medline: 23899831]
62. Levac JJ, O'Sullivan T. Social media and its use in health promotion. Interdiscip J Health Sci 2010 Feb 10;1(1):47. [doi: 10.18192/riss-ijhs.v1i1.1534]
63. Robertson-Lang L, Major S, Hemming H. An exploration of search patterns and credibility issues among older adults seeking online health information. Can J Aging 2011 Dec;30(4):631-645. [doi: 10.1017/S071498081100050X] [Medline: 22085455]
64. Chen X, Hay JL, Waters EA, Kiviniemi MT, Biddle C, Schofield E, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734 [FREE Full text] [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
65. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
66. Ye Y. A path analysis on correlates of consumer trust in online health information: evidence from the health information national trends survey. J Health Commun 2010 Dec 10;15 Suppl 3(sup3):200-215. [doi: 10.1080/10810730.2010.522687] [Medline: 21154094]
67. Dutta MJ, Bodie GD. Web searching for health: theoretical foundations and connections to health related outcomes. In: Web Search: Multidisciplinary Perspectives. Cham, Switzerland: Springer; 2008:235-254.
68. Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL. Access to care and use of the Internet to search for health information: results from the US National Health Interview Survey. J Med Internet Res 2015 Apr 29;17(4):e106 [FREE Full text] [doi: 10.2196/jmir.4126] [Medline: 25925943]
69. Pluye P, El Sherif R, Gonzalez-Reyes A, Turcotte E, Schuster T, Bartlett G, et al. Outcomes of equity-oriented, web-based parenting information in mothers of low socioeconomic status compared to other mothers: participatory mixed methods study. J Med Internet Res 2020 Nov 10;22(11):e22440 [FREE Full text] [doi: 10.2196/22440] [Medline: 33170125]
70. de Looper M, van Weert JC, Schouten BC, Bolle S, Belgers EH, Eddes EH, et al. The influence of online health information seeking before a consultation on anxiety, satisfaction, and information recall, mediated by patient participation: field study. J Med Internet Res 2021 Jul 05;23(7):e23670 [FREE Full text] [doi: 10.2196/23670] [Medline: 34255657]
71. Chen YY, Li CM, Liang JC, Tsai CC. Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. J Med Internet Res 2018 Feb 12;20(2):e47 [FREE Full text] [doi: 10.2196/jmir.9370] [Medline: 29434017]
72. Ayers SL, Kronenfeld JJ. Chronic illness and health-seeking information on the internet. Health (London) 2007 Jul;11(3):327-347. [doi: 10.1177/1363459307077547] [Medline: 17606698]

XSL•FO
RenderX

73. AsPredicted - Medical Information Search. Wharton Credibility Lab. URL: https://aspredicted.org/TPZ_21P [accessed 2024-02-19]
74. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022 Aug 29;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
75. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/BF03193146]
76. Link E, Baumann E. A comparison of women's and men's web-based information-seeking behaviors about gender-related health information: web-based survey study of a stratified German sample. J Med Internet Res 2023 May 17;25:e43897 [FREE Full text] [doi: 10.2196/43897] [Medline: 37195743]
77. Roth-Cohen O, Rosenberg H, Lissitsa S. Are you talking to me? Generation X, Y, Z responses to mobile advertising. Convergence 2021 Oct 19;28(3):761-780. [doi: 10.1177/13548565211047342]
78. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
79. Taylor KA, Humphrey Jr WF. Impact of medical blog reading and information presentation on readers' preventative health intentions: mixed methods, multistudy investigation. J Med Internet Res 2021 Dec 22;23(12):e23210 [FREE Full text] [doi: 10.2196/23210] [Medline: 34941543]
80. Gordon NP, Crouch E. Digital information technology use and patient preferences for internet-based health education modalities: cross-sectional survey study of middle-aged and older adults with chronic health conditions. JMIR Aging 2019 Apr 04;2(1):e12243 [FREE Full text] [doi: 10.2196/12243] [Medline: 31518291]
81. Jensen E, Jones N, Orozco K, Medina L, Perry M, Bolender B, et al. Measuring racial and ethnic diversity for the 2020 census. United States Census Bureau. 2021. URL: https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html [accessed 2023-11-15]
82. Xiao N, Sharman R, Rao HR, Upadhyaya S. Factors influencing online health information search: an empirical analysis of a national cancer-related survey. Decis Support Syst 2014 Jan;57:417-427. [doi: 10.1016/j.dss.2012.10.047]
83. Chugani S, Irwin JR. All eyes on you: the social audience and hedonic adaptation. Psychol Mark 2020 Aug 26;37(11):1554-1570. [doi: 10.1002/mar.21401]
84. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis, First Edition: A Regression-Based Approach. New York, NY: The Guilford Press; 2018.
85. Chisolm DJ. Does online health information seeking act like a health behavior?: a test of the behavioral model. Telemed J E Health 2010 Mar;16(2):154-160. [doi: 10.1089/tmj.2009.0102] [Medline: 20156127]
86. Dutta-Bergman MJ. Health attitudes, health cognitions, and health behaviors among internet health information seekers: population-based survey. J Med Internet Res 2004 May 28;6(2):e15 [FREE Full text] [doi: 10.2196/jmir.6.2.e15] [Medline: 15249264]
87. Bass SB, Ruzek SB, Gordon TF, Fleisher L, McKeown-Conn N, Moore D. Relationship of internet health information use with patient behavior and self-efficacy: experiences of newly diagnosed cancer patients who contact the National Cancer Institute's Cancer Information Service. J Health Commun 2006 Mar;11(2):219-236. [doi: 10.1080/10810730500526794] [Medline: 16537289]
88. Cocco AM, Zordan R, Taylor DM, Weiland TJ, Dilley SJ, Kant J, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. Med J Aust 2018 Oct 15;209(8):342-347. [doi: 10.5694/mja17.00889] [Medline: 30107763]
89. Bert F, Gualano MR, Brusaferro S, De Vito E, de Waure C, La Torre G, et al. Pregnancy e-health: a multicenter Italian cross-sectional study on internet use and decision-making among pregnant women. J Epidemiol Community Health 2013 Dec 01;67(12):1013-1018 [FREE Full text] [doi: 10.1136/jech-2013-202584] [Medline: 24072743]
90. Abdul-Muhsin H, Tyson M, Raghu S, Humphreys M. The informed patient: an analysis of information seeking behavior and surgical outcomes among men with benign prostatic hyperplasia. Am J Mens Health 2017 Jan 08;11(1):147-153 [FREE Full text] [doi: 10.1177/1557988315617527] [Medline: 26614445]
91. Stellefson M, Chaney B, Barry AE, Chavarria E, Tennant B, Walsh-Childers K, et al. Web 2.0 chronic disease self-management for older adults: a systematic review. J Med Internet Res 2013 Feb 14;15(2):e35 [FREE Full text] [doi: 10.2196/jmir.2439] [Medline: 23410671]
92. Mirzaei A, Aslani P, Luca EJ, Schneider CR. Predictors of health information-seeking behavior: systematic literature review and network analysis. J Med Internet Res 2021 Jul 02;23(7):e21680 [FREE Full text] [doi: 10.2196/21680] [Medline: 33979776]
93. Link E, Baumann E, Klimmt C. Explaining online information seeking behaviors in people with different health statuses: German representative cross-sectional survey. J Med Internet Res 2021 Dec 10;23(12):e25963 [FREE Full text] [doi: 10.2196/25963] [Medline: 34890348]
94. Sedrak MS, Soto-Perez-De-Celis E, Nelson RA, Liu J, Waring ME, Lane DS, et al. Online health information-seeking among older women with chronic illness: analysis of the women's health initiative. J Med Internet Res 2020 Apr 09;22(4):e15906 [FREE Full text] [doi: 10.2196/15906] [Medline: 32271152]

XSL•FO
RenderX

95. Jin X. Political ideology and differences in seeking COVID-19 information on the internet: examining the comprehensive model of information seeking. Online Inf Rev 2023 Mar 13;47(7):1280-1301. [doi: 10.1108/OIR-08-2022-0436]
96. Baumgaertner B, Carlisle JE, Justwan F. The influence of political ideology and trust on willingness to vaccinate. PLoS One 2018 Jan 25;13(1):e0191728 [FREE Full text] [doi: 10.1371/journal.pone.0191728] [Medline: 29370265]
97. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 2014 Dec 02;16(12):e262 [FREE Full text] [doi: 10.2196/jmir.3706] [Medline: 25470306]
98. Baker M, Stabile M, Deri C. What do self-reported, objective, measures of health measure? J Hum Resour 2004 Oct 01;XXXIX(4):1067-1093. [doi: 10.3368/jhr.XXXIX.4.1067]
99. Onur I, Velamuri M. The gap between self-reported and objective measures of disease status in India. PLoS One 2018 Aug 27;13(8):e0202786 [FREE Full text] [doi: 10.1371/journal.pone.0202786] [Medline: 30148894]
100. Cho S, Kim S. Does a healthy lifestyle matter? A daily diary study of unhealthy eating at home and behavioral outcomes at work. J Appl Psychol 2022 Jan;107(1):23-39. [doi: 10.1037/apl0000890] [Medline: 33764080]
101. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020 Aug 13;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]

Abbreviations

ANCOVA: analysis of covariance
HGIS: health guidance information search
HMIS: health management information search
HWIS: health wellness information search
OHI: online health information
OHIS: online health information seeking

Edited by A Mavragani; submitted 14.05.23; peer-reviewed by S Neely; comments to author 13.06.23; revised version received 31.07.23; accepted 08.02.24; published 11.03.24.

Please cite as:
Sinha J, Serin N
Online Health Information Seeking and Preventative Health Actions: Cross-Generational Online Survey Study
J Med Internet Res 2024;26:e48977
URL: https://www.jmir.org/2024/1/e48977
doi:10.2196/48977
PMID:38466985

©Jayati Sinha, Nuket Serin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Digital Health Experience for Patients in Primary Care: Mixed Methods Study

Melinda Ada Choy[1,2], BMed, MMed, DCH, MD; Kathleen O'Brien[1], BSc, GDipStats, MBBS, DCH; Katelyn Barnes[1,2], BAPSC, MND, PhD; Elizabeth Ann Sturgiss[3], BMed, MPH, MForensMed, PhD; Elizabeth Rieger[1], BA, MClinPsych, PhD; Kirsty Douglas[1,2], MBBS, DipRACOG, Grad Cert HE, MD

[1]School of Medicine and Psychology, College of Health and Medicine, The Australian National University, Canberra, Australia

[2]Academic Unit of General Practice, Office of Professional Leadership and Education, ACT Health Directorate, Canberra, Australia

[3]School of Primary and Allied Health Care, Monash University, Melbourne, Australia

Corresponding Author:
Melinda Ada Choy, BMed, MMed, DCH, MD
School of Medicine and Psychology
College of Health and Medicine
The Australian National University
Canberra
Australia
Phone: 61 51244947
Email: melinda.choy@gmail.com

Abstract

Background: The digital health divide for socioeconomic disadvantage describes a pattern in which patients considered socioeconomically disadvantaged, who are already marginalized through reduced access to face-to-face health care, are additionally hindered through less access to patient-initiated digital health. A comprehensive understanding of how patients with socioeconomic disadvantage access and experience digital health is essential for improving the digital health divide. Primary care patients, especially those with chronic disease, have experience of the stages of initial help seeking and self-management of their health, which renders them a key demographic for research on patient-initiated digital health access.

Objective: This study aims to provide comprehensive primary mixed methods data on the patient experience of barriers to digital health access, with a focus on the digital health divide.

Methods: We applied an exploratory mixed methods design to ensure that our survey was primarily shaped by the experiences of our interviewees. First, we qualitatively explored the experience of digital health for 19 patients with socioeconomic disadvantage and chronic disease and second, we quantitatively measured some of these findings by designing and administering a survey to 487 Australian general practice patients from 24 general practices.

Results: In our qualitative first phase, the key barriers found to accessing digital health included (1) strong patient preference for human-based health services; (2) low trust in digital health services; (3) high financial costs of necessary tools, maintenance, and repairs; (4) poor publicly available internet access options; (5) reduced capacity to engage due to increased life pressures; and (6) low self-efficacy and confidence in using digital health. In our quantitative second phase, 31% (151/487) of the survey participants were found to have never used a form of digital health, while 10.7% (52/487) were low- to medium-frequency users and 48.5% (236/487) were high-frequency users. High-frequency users were more likely to be interested in digital health and had higher self-efficacy. Low-frequency users were more likely to report difficulty affording the financial costs needed for digital access.

Conclusions: While general digital interest, financial cost, and digital health literacy and empowerment are clear factors in digital health access in a broad primary care population, the digital health divide is also facilitated in part by a stepped series of complex and cumulative barriers. Genuinely improving digital health access for 1 cohort or even 1 person requires a series of multiple different interventions tailored to specific sequential barriers. Within primary care, patient-centered care that continues to recognize the complex individual needs of, and barriers facing, each patient should be part of addressing the digital health divide.

(J Med Internet Res 2024;26:e50410) doi:10.2196/50410

XSL•FO
RenderX

KEYWORDS

digital health; eHealth; primary care; general practice; digital divide; health inequities; health inequality; disparities; digital cost; financial cost; health technology; mixed methods; barriers; barrier

Introduction

The Promise of eHealth

The rapid growth of digital health, sped up by the COVID-19 pandemic and associated lockdowns, brings the promise of improved health care efficiency, empowerment of consumers, and health care equity [1]. Digital health is the use of information and communication technology to improve health [2]. eHealth, which is a type of digital health, refers to the use of internet-based technology for health care and can be used by systems, providers, and patients [2]. At the time of this study (before COVID-19), examples of eHealth used by patients in Australia included searching for web-based health information, booking appointments on the web, participating in online peer-support health forums, using mobile phone health apps (mobile health), emailing health care providers, and patient portals for electronic health records.

Digital health is expected to improve chronic disease management and has already shown great potential in improving chronic disease health outcomes [3,4]. Just under half of the Australian population (47.3%) has at least 1 chronic disease [5]. Rates of chronic disease and complications from chronic disease are overrepresented among those with socioeconomic disadvantage [6]. Therefore, patients with chronic disease and socioeconomic disadvantage have a greater need for the potential benefits of digital health, such as an improvement in their health outcomes. However, there is a risk that those who could benefit most from digital health services are the least likely to receive them, exemplifying the inverse care law in the digital age by Hart [7].

Our Current Understanding of the Digital Health Divide

While the rapid growth of digital health brings the promise of health care equity, it may also intensify existing inequities [8]. The digital health divide for socioeconomic disadvantage describes a pattern in which patients considered socioeconomically disadvantaged who are already marginalized through poor access to traditional health care are additionally hindered through poor access to digital health [9]. In Australia, only 67.4% of households in the lowest household income quintile have home internet access, compared to 86% of the general population and 96.9% of households in the highest household income quintile [10]. Survey-based studies have also shown that even with internet access, effective eHealth use is lower in populations considered disadvantaged, which speaks to broader barriers to digital health access [11].

The ongoing COVID-19 global pandemic has sped up digital health transitions with the rapid uptake of telephone and video consultations, e-prescription, and the ongoing rollout of e-mental health in Australia. These have supported the continuation of health care delivery while limiting physical contact and the pandemic spread; however, the early evidence shows that the digital health divide remains problematic. A rapid review identified challenges with reduced digital access and digital literacy among the older adults and racial and ethnic minority groups, which are both groups at greater health risk from COVID-19 infections [12]. An Australian population study showed that the rapid uptake of telehealth during peak pandemic was not uniform, with the older adults, very young, and those with limited English language proficiency having a lower uptake of general practitioner (GP) telehealth services [13].

To ensure that digital health improves health care outcome gaps, it is essential to better understand the nature and nuance of the digital health divide for socioeconomic disadvantage. The nature of the digital health divide for socioeconomic disadvantage has been explored primarily through quantitative survey data, some qualitative papers, a few mixed methods papers, and systematic reviews [11,14-16]. Identified barriers include a lack of physical hardware and adequate internet bandwidth, a reduced inclination to seek out digital health, and a low ability and confidence to use digital health effectively [16]. The few mixed methods studies that exist on the digital health divide generally triangulate quantitative and qualitative data on a specific disease type or population subgroup to draw a combined conclusion [17,18]. These studies have found digital health access to be associated with education, ethnicity, and gender as well as trust, complementary face-to-face services, and the desire for alternative sources of information [17,19].

What This Work Adds

This project sought to extend previous research by using an exploratory mixed methods design to ensure that the first step and driver of our survey of a larger population was primarily shaped by the experiences of our interviewees within primary care. This differs from the triangulation method, which places the qualitative and quantitative data as equal first contributors to the findings and does not allow one type of data to determine the direction of the other [18]. We qualitatively explored the experience of digital health for patients with socioeconomic disadvantage and chronic disease and then quantitatively measured some of the qualitative findings via a survey of the Australian general practice patient population. Our key objective was to provide comprehensive primary mixed methods data, describing the experience and extent of barriers to accessing digital health and its benefits, with a focus on the digital health divide. We completed this research in a primary care context to investigate a diverse community-based population with conceivable reasons to seek digital help in managing their health. Findings from this mixed methods study were intended to provide health care providers and policy makers with a more detailed understanding of how specific barriers affect different aspects or steps of accessing digital health. Ultimately, understanding digital health access can influence the future design and implementation of digital health services by more effectively avoiding certain barriers or building in enablers to achieve improved digital health access not only for everyone but also especially for those in need.

XSL•FO
RenderX

Methods

Study Design

We conducted a sequential exploratory mixed methods study to explore a complex phenomenon in depth and then measure its prevalence. We qualitatively explored the experience of digital health for patients with chronic disease and socioeconomic disadvantage in the first phase. Data from the first phase informed a quantitative survey of the phenomenon across a wider population in the second phase [18]. Both stages of research were conducted before the COVID-19 pandemic in Australia.

Recruitment

Qualitative Phase Participants

The eligibility criteria for the qualitative phase were as follows: English-speaking adults aged ≥18 years with at least 1 self-reported chronic disease and 1 marker of socioeconomic disadvantage (indicated by ownership of a Health Care Card or receiving a disability pension, unemployment, or a user of public housing). A chronic disease was defined to potential participants as a diagnosed long-term health condition that had lasted at least 6 months (or is expected to last for at least 6 months; examples are listed in Multimedia Appendix 1). The markers of socioeconomic disadvantage we used to identify potential participants were based on criteria typically used by local general practices to determine which patients can have lower or no out-of-pocket expenses. Apart from unemployment, the 3 other criteria to identify socioeconomic disadvantage are means-tested government-allocated public social services [20]. Qualitative phase participants were recruited from May to July 2019 through 3 general practices and 1 service organization that serve populations considered socioeconomically disadvantaged across urban, regional, and rural regions in the Australian Capital Territory and South Eastern New South Wales. A total of 2 recruitment methods were used in consultation with and as per the choice of the participating organizations. Potential participants were either provided with an opportunity to engage with researchers (KB and MAC) in the general practice waiting room or identified by the practice or organization as suitable for an interview. Interested participants were given a detailed verbal and written description of the project in a private space before providing written consent to be interviewed. All interview participants received an Aus $50 (US $32.68) grocery shopping voucher in acknowledgment of their time.

Quantitative Phase Participants

Eligibility for the quantitative phase was English-speaking adults aged ≥18 years. The eligibility criteria for the quantitative phase were deliberately broader than those for the qualitative phase to achieve a larger sample size within the limitations of recruitment and with the intention that the factors of socioeconomic disadvantage and having a chronic disease could be compared to the digital health access of a more general population. The quantitative phase participants were recruited from November 2019 to February 2020. Study information and paper-based surveys were distributed and collected through 24 general practices across the Australian Capital Territory and South Eastern New South Wales regions, with an option for web-based completion.

Ethical Considerations

Qualitative and quantitative phase research protocols, including the participant information sheet, were approved by the Australian Capital Territory Health Human Research Ethics Committee (2019/ETH/00013) and the Australian National University Human Research Ethics Committee (2019/ETH00003). Qualitative phase participants were given a verbal and written explanation of the study, including how and when they could opt out, before they provided written consent. All interview participants received an Aus $50 (US $32.68) grocery shopping voucher in acknowledgment of their time. Quantitative participants were given a written explanation and their informed consent was implied by return of a completed survey. Participants in both phases of the study were told that all their data was deidentified. Consent was implied through the return of a completed survey.

Qualitative Data Collection and Analysis

Participants were purposively sampled to represent a range in age, gender, degree of socioeconomic disadvantage, and experience of digital health. The sampling and sample size were reviewed regularly by the research team as the interviews were being completed to identify potential thematic saturation.

The interview guide was developed by the research team based on a review of the literature and the patient dimensions of the framework of access by Levesque et al [21]. The framework by Levesque et al [21] is a conceptualization of health care access comprising 5 service and patient dimensions of accessibility and ability. The patient dimensions are as follows: (1) ability to perceive, (2) ability to seek, (3) ability to reach, (4) ability to pay, and (5) ability to engage [21]. The key interview topics included (1) digital health use and access, including facilitators and barriers; (2) attitudes toward digital health; and (3) self-perception of digital health skills and potential training. The interview guide was reviewed for face and content validity by the whole research team, a patient advocate, a digital inclusion charity representative, and the general practices where recruitment occurred. The questions and guide were iteratively refined by the research team to ensure relevance and support reaching data saturation. The interview guide has been provided as Multimedia Appendix 1. The interviews, which took 45 minutes on average, were taped and transcribed. An interview summary sheet and reflective journal were completed by the interviewer after each interview to also capture nonverbal cues and tone.

Interview transcriptions were coded and processed by inductive thematic analysis. Data collection and analysis were completed in parallel to support the identification of data saturation. Data saturation was defined as no significant new information arising from new interviews and was identified by discussion with the research team [22]. The 2 interviewers (MAC and KB) independently coded the first 5 transcripts and reflected on them with another researcher (EAS) to ensure intercoder validity and reliability. The rest of the interviews were coded independently by the 2 interviewers, who regularly met to reflect on emerging

XSL•FO
RenderX

themes and thematic saturation. Data saturation was initially indicated after 15 interviews and subsequently confirmed with a total of 19 interviews. Coding disagreements and theme development were discussed with at least 1 other researcher (EAS, ER, and KD). Thematic saturation and the final themes were agreed upon by the entire research team.

Quantitative Survey Development

The final themes derived in the qualitative phase of the project guided the specific quantitative phase research questions. The final themes were a list of ordered cumulative barriers experienced by participants in accessing digital health and its benefits (Figure 1). The quantitative survey was designed to test the association between barriers to access and the frequency of use of digital health as a proxy measure for digital health access.

Figure 1. Barriers to digital health access learned in phase 1.

In the survey, the participants were asked about their demographic details, health and chronic diseases, knowledge, use and experience of digital health tools, internet access, perception of digital resource affordability, trust in digital health and traditional health services, perceived capability, health care empowerment, eHealth literacy, and relationship with their GP. Existing scales and questions from the literature and standardized Australian-based surveys were used whenever possible. We used selected questions and scales from the Australian Bureau of Statistics standards, the eHealth Literacy Scale (eHEALS), the eHealth Literacy Questionnaire, and the Southgate Institute for Health Society and Equity [17,23-26]. We adapted other scales from the ICEpop Capability Measure

XSL•FO
RenderX

for Adults, the Health Care Empowerment Inventory (HCEI), the Patient-Doctor Relationship Questionnaire, and the Chao continuity questionnaire [23,27-29]. Where an existing scale to measure a barrier or theme did not exist, the research team designed the questions based on the literature. Our questions around the frequency of digital health use were informed by multiple existing Australian-based surveys on general technology use [30,31]. Most of the questions used a Likert scale. Every choice regarding the design, adaptation, or copy of questions for the survey was influenced by the qualitative findings and decided on by full agreement among the 2 researchers who completed and coded the interviews. A complete copy of the survey is provided in Multimedia Appendix 2.

Pilot-testing of the survey was completed with 5 patients, 2 experts on digital inclusion, and 3 local GPs for both the paper surveys and web-based surveys via Qualtrics Core XM (Qualtrics LLC). The resulting feedback on face and content validity, functionality of the survey logic, and feasibility of questionnaire completion was incorporated into the final version of the survey.

The survey was offered on paper with a participant information sheet, which gave the patients the option to complete the web-based survey. The survey was handed out to every patient on paper to avoid sampling bias through the exclusion of participants who could not complete the web-based survey [32].

Quantitative Data Treatment and Analysis

Data were exported from Qualtrics Core XM to an SPSS (version 26; IBM Corp) data set. Data cleaning and screening were undertaken (KB and KO).

Descriptive statistics (number and percentage) were used to summarize participant characteristics, preference measures, and frequency of eHealth use. Significance testing was conducted using chi-square tests, with a threshold of $P<.05$; effect sizes were measured by the ϕ coefficient for 2×2 comparisons and Cramer V statistic for all others. Where the cells sizes were too small, the categories were collapsed for the purposes of significance testing. The interpretation of effect sizes was as per the study by Cohen [33]. The analysis was conducted in SPSS and SAS (version 9.4; SAS Institute).

Participant Characteristics

Participants' self-reported characteristics included gender, indigenous status, income category, highest level of education, marital status, and language spoken at home.

Age was derived from participant-reported year of birth and year of survey completion as of 2019 and stratified into age groups. The state or territory of residence was derived from the participant-reported postcode. The remoteness area was derived using the postcode reported by the participants and mapped to a modified concordance from the Australian Bureau of Statistics. Occupation-free text responses were coded using the Australian Bureau of Statistics Census statistics level 1 and 2 descriptors. The country of birth was mapped to Australia, other Organisation for Economic Cooperation and Development countries, and non–Organisation for Economic Cooperation and Development countries.

Frequency of eHealth Use

A summary measure of the frequency of eHealth use was derived from the questions on the use of different types of eHealth.

Specifically, respondents were asked if they had ever used any form of web-based health ("eHealth") and, if so, to rate how often (never, at least once, every now and then, and most days) against 6 types of "eHealth" (searching for health information online, booking appointments online, emailing health care providers, using health-related mobile phone apps, accessing My Health Record, and accessing online health forums). The frequency of eHealth use was then classified as follows:

- High user: answered "most days" to at least 1 question on eHealth use OR answered "every now and then" to at least 2 questions on eHealth use
- Never user: answered "no" to having ever used any form of eHealth OR "never" to all 6 questions on eHealth use
- Low or medium user: all other respondents.

The frequency of eHealth use was reported as unweighted descriptive statistics (counts and percentages) against demographic characteristics and for the elements of each of the themes identified in phase 1.

Themes

Overview of Key Themes

Data were reported against the 6 themes from the phase 1 results of preference, trust, cost, structural access, capacity to engage, and self-efficacy. Where the components of trust, cost, capacity to engage, and self-efficacy had missing data (for less than half of the components only), mean imputation was used to minimize data loss. For each theme, the analysis excluded those for whom the frequency of eHealth use was unknown.

Preference

Preference measures (survey section D1 parts 1 to 3) asked participants to report against measures with a 4-point Likert scale (strongly disagree, disagree, agree, and strongly agree). Chi-square tests were conducted after the categories were condensed into 2 by combining strongly disagree and as well as combining strongly agree and agree.

Trust

Summary measures for trust were created in 4 domains: trust from the eHealth Literacy Questionnaire (survey section D1 parts 4 to 8), trust from Southgate—GPs, specialists, or allied health (survey section D2 parts 1 to 5), trust from Southgate—digital health (survey section D2 parts 6, 7, 9, and 10), and trust from Southgate—books or pamphlets (survey section D2 part 8). The data were grouped as low, moderate, and high trust based on the assigned scores from the component data. Chi-square tests were conducted comparing low-to-moderate trust against high trust for GP, specialists, or allied health and comparing low trust against moderate-to-high trust for book or pamphlet.

XSL•FO
RenderX

Cost

Summary measures for cost were created from survey item C10. To measure cost, participants were asked about whether they considered certain items or services to be affordable. These included cost items mentioned in the qualitative phase interviews relating to mobile phones (1 that connects to the internet, 1 with enough memory space to download apps, downloads or apps requiring payment, repairs, and maintenance costs), having an iPad or tablet with internet connectivity, a home computer or laptop (owning, repairs, and maintenance), home fixed internet access, and an adequate monthly data allowance. These 9 items were scored as "yes definitely"=1 or 0 otherwise. Chi-square tests were conducted with never and low or medium eHealth users combined.

Structural Access

Structural access included asking where the internet is used by participants (survey section C8) and factors relating to internet access (survey section C8 parts 1-3) reporting against a 4-point Likert scale (strongly disagree, disagree, agree, and strongly agree). Chi-square tests were conducted with strongly disagree, disagree, agree, or strongly agree, and never, low, or medium eHealth use combined.

Capacity to Engage

Summary measures for capacity to engage were created from survey section E1. To measure the capacity to engage, participants were asked about feeling "settled and secure," "being independent," and "achievement and progress" as an adaptation of the ICEpop Capability Measure for Adults [27], reporting against a 4-point Likert-like scale. Responses were scored from 1 ("I am unable to feel settled and secure in any areas of my life") to 4 ("I am able to feel settled and secure in all areas of my life").

The summary capacity measure was derived by the summation of responses across the 3 questions, which were classified into 4 groups, A to D, based on these scores. Where fewer than half of the responses were missing, mean imputation was used; otherwise, the record was excluded. Groups A and B were combined for significance testing.

Self-Efficacy

Summary measures for self-efficacy were adapted from the eHEALS (E3) and the HCEI (E2) [23,24].

Survey section E3—eHEALS—comprised 8 questions, with participants reporting against a 5-point Likert scale for each (strongly disagree, disagree, neither, agree, and strongly agree). These responses were assigned 1 to 5 points, respectively. The summary eHEALS measure was derived by the summation of responses across the 8 questions, which were classified into 5 groups, A to E, based on these scores. Where fewer than half of the responses were missing, mean imputation was used; otherwise, the record was excluded. Groups A to C and D to E were combined for significance testing.

Survey section E2—HCEI—comprised 5 questions, with participants reporting against a 5-point Likert scale for each (strongly disagree, disagree, neither, agree, and strongly agree). Strongly disagree and disagree and neither were combined, and similarly agree and strongly agree were combined for significance testing.

Results

Qualitative Results

Overview

The demographic characteristics of the patients that we interviewed are presented in Table 1.

The key barriers found to accessing digital health included (1) strong patient preference for human-based health services; (2) low trust in digital health services; (3) high financial costs of necessary tools, maintenance, and repairs; (4) poor publicly available internet access options; (5) reduced capacity to engage due to increased life pressures; and (6) low self-efficacy and confidence in using digital health.

Rather than being an equal list of factors, our interviewees described these barriers as a stepped series of cumulative hurdles, which is illustrated in Figure 1. Initial issues of preference and trust were foundational to a person even when considering the option of digital health, while digital health confidence and literacy were barriers to full engagement with and optimal use of digital health. Alternatively, interviewees who did use digital health had been enabled by the same factors that were barriers to others.

XSL•FO
RenderX

Table 1. Phase 1 participant characteristics (N=19).

Characteristic	Value, n (%)
Gender	
Female	10 (53)
Male	9 (47)
Age (years)	
18-24	0 (0)
25-44	2 (11)
45-64	13 (68)
65-74	3 (16)
>75	1 (5)
Country of birth	
Australia	13 (68)
Other	6 (32)
Marital status	
Never married	6 (32)
Married	4 (21)
Widowed	2 (11)
Separated, not divorced	2 (11)
Divorced	4 (21)
Prefer not to say	1 (5)
Highest educational qualification	
Year ≤11	6 (32)
Year 12	4 (21)
Trade certification	4 (21)
Diploma	3 (16)
Bachelor's degree	1 (5)
Prefer not to say	1 (5)
Occupation	
Not employed	8 (42)
Home duties	2 (11)
Pension	5 (26)
Retired	3 (16)
Other	1 (5)
Presence of chronic disease	
Yes	19 (100)
No	0 (0)
Number of appointments with GP[a]/year	
<5	5 (26)
5-10	6 (32)
>10	8 (42)
Eligibility criteria for socioeconomic disadvantage	
User of public housing	6[b] (32)

XSL•FO
RenderX

Characteristic	Value, n (%)
Disability support pension	10[b] (53)
Health care card	8[b] (42)
Unemployed	3[b] (16)

[a]GP: general practitioner.

[b]Multiple answers per respondent.

Strong Patient Preference for Human-Based Health Services

Some patients expressed a strong preference for human-based health services rather than digital health services. In answer to a question about how digital health services could be improved, a patient said the following:

> *Well, having an option where you can actually bypass actually having to go through the app and actually talk directly to someone.* [Participant #10]

For some patients, this preference for human-based health services appeared to be related to a lack of exposure to eHealth. These patients were not at all interested in or had never thought about digital health options. A participant responded the following to the interviewer's questions:

> *Interviewer: So when...something feels not right, how do you find out what's going on?*
>
> *Respondent: I talk to Doctor XX.*
>
> *Interviewer: Do you ever Google your symptoms or look online for information?*
>
> *Respondent: No, I have never even thought of doing that actually.* [Participant #11]

For other patients, their preference for human-based health care stemmed from negative experiences with technology. These patients reported actively disliking computers and technology in general and were generally frustrated with what they saw as the pitfalls of technology. A patient stated the following:

> *If computers and internet weren't so frigging slow because everything is on like the slowest speed network ever and there's ads blocking everything. Ads, (expletive) ads.* [Participant #9]

A patient felt that he was pushed out of the workforce due his inability to keep up with technology-based changes and thus made a decision to never own a computer:

> *But, you know, in those days when I was a lot younger those sorts of things weren't about and they're just going ahead in leaps and bounds and that's one of the reasons why I retired early. I retired at 63 because it was just moving too fast and it's all computers and all those sorts of things and I just couldn't keep up.* [Participant #17]

Low Trust in Digital Health Services

Several patients described low trust levels for digital and internet-based technology in general. Their low trust was generally based on stories they had heard of other people's negative experiences. A patient said the following:

> *I don't trust the internet to be quite honest. You hear all these stories about people getting ripped off and I've worked too hard to get what I've got rather than let some clown get it on the internet for me.* [Participant #11]

Some of this distrust was specific to eHealth. For example, some patients were highly suspicious of the government's motives with regard to digital health and were concerned about the privacy of their health information, which made them hesitant about the concept of a universal electronic health record. In response to the interviewer's question, a participant said the following:

> *Interviewer: Are there any other ways you think that eHealth might help you?*
>
> *Respondent: I'm sorry but it just keeps coming back to me, Big Brother.* [Participant #7]

Another participant said the following:

> *I just would run a mile from it because I just wouldn't trust it. It wouldn't be used to, as I said, for insurance or job information.* [Participant #16]

High Financial Costs of the Necessary Tools, Maintenance, and Repairs

A wide variety of patients described affordability issues across several different aspects of the costs involved in digital health. They expressed difficulty in paying for the following items: a mobile phone that could connect to the internet, a mobile phone with enough memory space to download apps, mobile phone apps requiring extra payment without advertisements, mobile phone repair costs such as a broken screen, a computer or laptop, home internet access, and adequate monthly data allowance and speeds to functionally use the internet. Current popular payment systems, such as plans, were not feasible for some patients. A participant stated the following:

> *I don't have a computer...I'm not in the income bracket to own a computer really. Like I could, if I got one on a plan kind of thing or if I saved up for x-amount of time. But then like if I was going on the plan I'd be paying interest for having it on like lay-buy kind of thing, paying it off, and if it ever got lost or stolen I would still have to repay that off, which is always a hassle. And yeah. Yeah, I'm like financially not in the state where I'm able to...own a computer right now as I'm kind of paying off a number of debts.* [Participant #9]

XSL•FO
RenderX

Poor Publicly Available Internet Access Options

Some patients described struggling without home internet access. While they noted some cost-free public internet access points, such as libraries, hotel bars, and restaurants, they often found these to be inconvenient, lacking in privacy, and constituting low-quality options for digital health. A patient stated the following:

> *...it's incredibly slow at the library. And I know why...a friend I went to school with used to belong to the council and the way they set it up, they just got the raw end of the stick and it is really, really slow. It's bizarre but you can go to the X Hotel and it's heaps quicker.* [Participant #15]

In response to the interviewer's question, a participant said the following:

> *Interviewer: And do you feel comfortable doing private stuff on computers at the library...?*
>
> *Respondent: Not really, no, but I don't have any other choice, so, yeah.* [Participant #9]

Reduced Capacity to Engage Due to Increased Life Pressures

When discussing why they were not using digital health or why they had stopped using digital health, patients often described significant competing priorities and life pressures that affected their capacity to engage. An unemployed patient mentioned that his time and energy on the internet were focused primarily on finding work and that he barely had time to focus on his health in general, let alone engage in digital health.

Other patients reported that they often felt that their ability to learn about and spend time on digital health was taken up by caring for sick family members, paying basic bills, or learning English. Some patients said that the time they would have spent learning digital skills when they were growing up had been lost to adverse life circumstances such as being in jail:

> *So we didn't have computers in the house when I was growing up. And I didn't know I've never...I've been in and out of jail for 28 odd years so it sort of takes away from learning from this cause it's a whole different… it's a whole different way of using a telephone from a prison.* [Participant #11]

Low Self-Efficacy and Confidence in Starting the Digital Health Process

Some patients had a pervasive self-perception of being slow learners and being unable to use technology. Their stories of being unconfident learners seemed to stem from the fact that they had been told throughout their lives that they were intellectually behind. A patient said the following:

> *The computer people...wouldn't take my calls because I've always been dumb with that sort of stuff. Like I only found out this later on in life, but I'm actually severely numerically dyslexic. Like I have to triple-check everything with numbers.* [Participant #7]

Another patient stated the following:

> *I like went to two English classes like a normal English class with all the kids and then another English class with about seven kids in there because I just couldn't I don't know maybe because I spoke another language at home and they sort of like know I was a bit backward.* [Participant #6]

These patients and others had multiple missing pieces of information that they felt made it harder to engage in digital health compared to "easier" human-based services. A patient said the following:

> *Yeah I've heard of booking online but I just I don't know I find it easier just to ring up. And I'll answer an email from a health care provider but I wouldn't know where to start to look for their email address.* [Participant #11]

In contrast, the patients who did connect with digital health described themselves as independent question askers and proactive people. Even when they did not know how to use a specific digital health tool, they were confident in attempting to and asking for help when they needed it. A patient said the following:

> *I'm a "I will find my way through this, no matter how long it takes me" kind of person. So maybe it's more my personality...If I have to ask for help from somewhere, wherever it is, I will definitely do that.* [Participant #3]

Quantitative Results

Participant Characteristics

A total of 487 valid survey responses were received from participants across 24 general practices. The participant characteristics are presented in detail in Table S1 in Multimedia Appendix 3.

The mean age of the participants was approximately 50 years (females 48.9, SD 19.4 years; males 52.8, SD 20.0 years), and 68.2% (332/487) of the participants identified as female. Overall, 34.3% (151/439) of respondents reported never using eHealth, and 53.8% (236/439) reported high eHealth use.

There were statistically significant ($P<.05$) differences in the frequency of eHealth use in terms of age group, gender, state, remoteness, highest level of education, employment status, occupation group, marital status, and language spoken at home, with effect sizes being small to medium. Specifically, high eHealth characteristics were associated with younger age, being female, living in an urban area, and being employed.

Preference

Table 2 presents the frequency of eHealth use against 3 internet preference questions.

Preference for using the internet and technology in general and for health needs in particular were significantly related to the frequency of eHealth use ($P<.05$ for each), with the effect sizes being small to medium.

XSL•FO
RenderX

Table 2. Preference measures.

	Frequency of eHealth use[a], n (%)				*P* value[b]	Effect size
	Never	Low or medium	High	Total		
I am interested in using the internet and technology in general					<.001	0.30
Strongly disagree	13 (72.2)	1 (5.6)	4 (22.2)	18 (100)		
Disagree	17 (70.8)	4 (16.7)	3 (12.5)	24 (100)		
Agree	64 (37.2)	22 (12.8)	86 (50)	172 (100)		
Strongly agree	36 (18.9)	22 (11.6)	132 (69.5)	190 (100)		
I am interested in using the internet and technology for health needs					<.001	0.40
Strongly disagree	18 (75)	1 (4.2)	5 (20.8)	24 (100)		
Disagree	38 (65.5)	8 (13.8)	12 (20.7)	58 (100)		
Agree	61 (27.9)	27 (12.3)	131 (59.8)	219 (100)		
Strongly agree	13 (12.6)	13 (12.6)	77 (74.8)	103 (100)		
It is important for me to be able to access health resources (eg, information) on the internet					<.001	0.4
Strongly disagree	14 (87.5)	0 (0)	2 (12.5)	16 (100)		
Disagree	35 (63.6)	8 (14.5)	12 (21.8)	55 (100)		
Agree	65 (30.8)	30 (14.2)	116 (55)	211 (100)		
Strongly agree	15 (12.6)	11 (9.2)	93 (78.2)	119 (100)		

[a]Excludes those for whom frequency of eHealth use is unknown.

[b]Chi-square tests conducted with strongly disagree and disagree combined, and agree and strongly agree combined.

Trust

Table 3 presents the frequency of eHealth use against 4 measures of trust.

The degree of trust was not statistically significantly different for the frequency of eHealth use for any of the domains.

XSL•FO
RenderX

Table 3. Trust measures.

	Frequency of eHealth use[a], n (%)				*P* value	Effect size
	Never	Low or medium	High	Total		
Trust (as per eHLQ)[b,c]					.15	0.09
Low trust	16 (43.2)	6 (16.2)	15 (40.5)	37 (100)		
Moderate trust	56 (27.1)	25 (12.1)	126 (60.9)	207 (100)		
High trust	54 (35.3)	18 (11.8)	81 (52.9)	153 (100)		
Trust[d] (as per Southgate)—general practitioner, specialist, or allied health					.42[e]	0.07
Low trust	5 (83.3)	0 (0)	1 (16.7)	6 (100)		
Moderate trust	11 (23.9)	9 (19.6)	26 (56.5)	46 (100)		
High trust	118 (33.3)	39 (11)	197 (55.6)	354 (100)		
Trust[f] (as per Southgate): digital health					.49	0.07
Low trust	37 (30.8)	19 (15.8)	64 (53.3)	120 (100)		
Moderate trust	53 (29.4)	21 (11.7)	106 (58.9)	180 (100)		
High trust	33 (35.5)	8 (8.6)	52 (55.9)	93 (100)		
Trust[g] (as per Southgate): book or pamphlet					.18[h]	0.09
Low trust	40 (38.5)	11 (10.6)	53 (51)	104 (100)		
Moderate trust	71 (28.3)	35 (13.9)	145 (57.8)	251 (100)		
High trust	12 (31.6)	2 (5.3)	24 (63.2)	38 (100)		

[a]Excludes those for whom frequency of eHealth use is unknown.

[b]eHLQ: eHealth Literacy Questionnaire.

[c]Derived from survey question D1, parts 4 to 8. Mean imputation used where ≤2 responses were missing. If >2 responses were missing, the records were excluded.

[d]Derived from survey question D2, parts 1 to 5. Mean imputation used where ≤2 responses were missing. If >2 responses were missing, the records were excluded.

[e]Chi-square test conducted comparing low-to-moderate trust against high trust.

[f]Derived from survey question D2, parts 6, 7, 9, and 10. Mean imputation used where ≤2 responses were missing. If >2 responses were missing, the records were excluded.

[g]Derived from survey question D2 part 8.

[h]Chi-square test conducted comparing low trust against moderate-to-high trust.

Cost

Affordability of items and services was reported as *No cost difficulty* or *Cost difficulty.* eHealth frequency of use responses were available for 273 participants; among those with *no cost difficulty*, 1% (2/204) were never users, 14.2% (29/204) were low or medium users, and 84.8% (173/204) were high users of eHealth; among those with *cost difficulty*, 1% (1/69) were never users, 26% (18/69) were low or medium users, and 73% (50/69) were high users. There was a statistically significant difference in the presence of cost as a barrier between never and low or medium eHealth users compared to high users (χ^2_1=5.25; *P*=.02), although the effect size was small.

Structural Access

Table 4 presents the frequency of eHealth use for elements of structural access.

Quality of internet access and feeling limited in access to the internet were significantly associated with frequency of eHealth use (*P*<.05), although the effect sizes were small.

XSL•FO
RenderX

Table 4. Structural access.

	Frequency of eHealth use[a], n (%)				*P* value	Effect size
	Never	Low or medium	High	Total		
Where do you use the internet most?					N/A[b]	N/A
Home	2 (0.8)	45 (18.3)	199 (80.9)	199 (100)		
Work	0 (0)	4 (10.8)	33 (89.2)	33 (100)		
Library or other public Wi-Fi	1 (100)	0 (0)	0 (0)	1 (100)		
Family or friend's house	1 (100)	0 (0)	0 (0)	1 (100)		
Overall, my quality of internet access is adequate to suit my needs					.002[c]	0.18
Strongly disagree	1 (6.7)	6 (40)	8 (53.3)	15 (100)		
Disagree	0 (0)	6 (30)	14 (70)	20 (100)		
Agree	2 (1.3)	26 (16.9)	126 (81.8)	154 (100)		
Strongly agree	1 (1.1)	10 (10.8)	82 (88.2)	93 (100)		
I am comfortable dealing with health information online in a public place					.26[c]	0.07
Strongly disagree	1 (2.8)	9 (25)	26 (72.2)	36 (100)		
Disagree	0 (0)	14 (18.7)	61 (81.3)	75 (100)		
Agree	1 (0.8)	22 (18.2)	98 (81)	121 (100)		
Strongly agree	1 (2.2)	3 (6.7)	41 (91.1)	45 (100)		
I feel limited in my access to the internet					.02[c]	−0.13
Strongly disagree	1 (0.7)	22 (15.2)	122 (84.1)	145 (100)		
Disagree	1 (1)	18 (17.3)	85 (81.7)	104 (100)		
Agree	0 (0)	8 (38.1)	13 (61.9)	21 (100)		
Strongly agree	1 (12.5)	1 (12.5)	6 (75)	8 (100)		

[a]Excludes those for whom frequency of eHealth use is unknown.

[b]N/A: not applicable (cell sizes insufficient for chi-square test).

[c]Chi-square tests conducted with strongly disagree and disagree combined, agree and strongly agree combined, and never and low or medium categories combined.

Capacity to Engage

Table 5 presents the frequency of eHealth use against respondents' capacity to engage.

Capacity to engage was not significantly different for the frequency of eHealth use (*P*=.54).

Table 5. Capacity to engage.

	Frequency of eHealth use[a], n (%)				*P* value	Effect size
	Never	Low or medium	High	Total		
Capacity to engage[b]					.54[c]	0.06
Group A: 1-3	2 (50)	0 (0)	2 (50)	4 (100)		
Group B: 3-6	8 (25.8)	7 (22.6)	16 (51.6)	31 (100)		
Group C: 6-9	54 (32.5)	17 (10.2)	95 (57.2)	166 (100)		
Group D: 10-12	72 (34.6)	23 (11.1)	113 (54.3)	208 (100)		

[a]Excludes those for whom frequency of eHealth use is unknown.

[b]Derived from survey item E1. Where 1 response was missing, the mean imputation was used. If >1 response was missing, the record was excluded.

[c]Chi-square tests conducted with groups A and B combined.

XSL•FO
RenderX

Self-Efficacy

Table 6 presents the frequency of eHealth use for elements of self-efficacy.

Statistically significant results were observed for the relationship between self-efficacy by eHEALS (moderate effect size) and frequency of eHealth use as well as for some of the questions from the HCEI (reliance on health professionals or others to access and explain information; small effect size; $P<.05$).

XSL•FO
RenderX

Table 6. Self-efficacy.

	Frequency of eHealth use[a], n (%)				*P* value	Effect size
	Never	Low or medium	High	Total		
Self-efficacy by eHEALS[b,c]					<.001[c]	0.39
Group A: 1-8	3 (75)	0 (0)	1 (25)	4 (100)		
Group B: 9-16	15 (71.4)	3 (14.3)	3 (14.3)	21 (100)		
Group C: 17-24	47 (58.8)	11 (13.8)	22 (27.5)	80 (100)		
Group D: 25-32	55 (24.2)	25 (11)	147 (64.8)	227 (100)		
Group E: 33-40	13 (18.1)	8 (11.1)	51 (70.8)	72 (100)		
I prefer to get as much information as possible about treatment options					.08[d]	0.11
Strongly disagree	3 (60)	0 (0)	2 (40)	5 (100)		
Disagree	3 (50)	0 (0)	3 (50)	6 (100)		
Neither	10 (41.7)	6 (25)	8 (33.3)	24 (100)		
Agree	58 (35.4)	21 (12.8)	85 (51.8)	164 (100)		
Strongly agree	59 (28.8)	21 (10.2)	125 (61)	205 (100)		
I try to get my health care providers to listen to my preferences for my treatment					.06[d]	0.12
Strongly disagree	3 (75)	0 (0)	1 (25)	4 (100)		
Disagree	2 (40)	1 (20)	2 (40)	5 (100)		
Neither	21 (34.4)	12 (19.7)	28 (45.9)	61 (100)		
Agree	66 (35.5)	17 (9.1)	103 (55.4)	186 (100)		
Strongly agree	40 (27.2)	18 (12.2)	89 (60.5)	147 (100)		
I am very active in my health care					.052[d]	0.12
Strongly disagree	3 (75)	1 (25)	0 (0)	4 (100)		
Disagree	5 (25)	3 (15)	12 (60)	20 (100)		
Neither	22 (28.2)	15 (19.2)	41 (52.6)	78 (100)		
Agree	71 (37.8)	18 (9.6)	99 (52.7)	188 (100)		
Strongly agree	31 (27.4)	11 (9.7)	71 (62.8)	113 (100)		
I take my commitment to my treatment seriously					.30[d]	0.08
Strongly disagree	4 (80)	1 (20)	0 (0)	5 (100)		
Disagree	0 (0)	0 (0)	3 (100)	3 (100)		
Neither	8 (23.5)	7 (20.6)	19 (55.9)	34 (100)		
Agree	69 (33.8)	24 (11.8)	111 (54.4)	204 (100)		
Strongly agree	52 (32.5)	16 (10)	92 (57.5)	160 (100)		
I rely on health professionals or others to access the information for me and then explain it to me					<.001[d]	0.22
Strongly disagree	3 (33.3)	2 (22.2)	4 (44.4)	9 (100)		
Disagree	3 (7.3)	4 (9.8)	34 (82.9)	41 (100)		
Neither	18 (21.7)	11 (13.3)	54 (65.1)	83 (100)		
Agree	68 (37.6)	22 (12.2)	91 (50.3)	181 (100)		
Strongly agree	41 (45.1)	9 (9.9)	41 (45.1)	91 (100)		

[a]Excludes those for whom frequency of eHealth use is unknown.

[b]eHEALS: eHealth Literacy Scale.

[c]eHEALS derived from item E3 (8 parts). Where ≤ 4 responses were missing, mean imputation was used. If >4 responses were missing, the records were excluded. Groups A to C as well as groups D to E were combined for the chi-square test.

XSL•FO
RenderX

[d]Strongly disagree, disagree, neither, and agree or strongly agree combined for significance testing.

Discussion

Principal Findings

This paper reports on the findings of a sequential exploratory mixed methods study on the barriers to digital health access for a group of patients in Australian family medicine, with a particular focus on chronic disease and socioeconomic disadvantage.

In the qualitative first phase, the patients with socioeconomic disadvantage and chronic disease described 6 cumulative barriers, as demonstrated in Figure 1. Many nonusers of digital health preferred human-based services and were not interested in technology, while others were highly suspicious of the technology in general. Some digitally interested patients could not afford quality hardware and internet connectivity, a barrier that was doubled by low quality and privacy when accessing publicly available internet connections. Furthermore, although some digitally interested patients had internet access, their urgent life circumstances left scarce opportunity to access digital health and develop digital health skills and confidence.

In our quantitative second phase, 31% (151/487) of the survey participants from Australian general practices were found to have never used a form of digital health. Survey participants were more likely to use digital health tools frequently when they also had a general digital interest and a digital health interest. Those who did not frequently access digital health were more likely to report difficulty affording the financial costs needed for digital access. The survey participants who frequently accessed digital health were more likely to have high eHealth literacy and high levels of patient empowerment.

Comparison With Prior Work

In terms of general digital health access, the finding that 31% (151/487) of the survey participants had never used one of the described forms of eHealth is in keeping with an Australian-based general digital participation study that found that approximately 9% of the participants were nonusers and 17% rarely engaged with the internet at all [34]. With regard to the digital health divide, another Australian-based digital health divide study found that increased age, living in a lower socioeconomic area, being Aboriginal or Torres Strait Islander, being male, and having no tertiary education were factors negatively associated with access to digital health services [17]. Their findings correspond to our findings that higher-frequency users of eHealth were associated with younger age, being female, living in an urban area, and being employed. Both studies reinforce the evidence of the digital health divide based on gender, age, and socioeconomic disadvantage in Australia.

With regard to digital health barriers, our findings provide expanded details on the range of digital health items and services that present a cost barrier to consumers. Affordability is a known factor in digital access and digital health access, and it is measured often by general self-report or relative expenditure on internet access to income [30]. Our study revealed the comprehensive list of relevant costs for patients. Our study also demonstrated factors of cost affordability beyond the dollar value of an item, as interviewees described the struggle of using slow public internet access without privacy features and the risks involved in buying a computer in installments. When we reflected on the complexity and detail of the cost barrier in our survey, participants demonstrated a clear association between cost and the frequency of digital health use. This suggests that a way to improve digital health access for some people is to improve the quality, security, and accessibility of public internet access options as well as to provide free or subsidized hardware, internet connection, and maintenance options for those in need, work that is being done by at least 1 digital inclusion charity in the United Kingdom [35].

Many studies recognize the factors of eHealth literacy and digital confidence for beneficial digital health access [36]. Our interviews demonstrated that some patients with socioeconomic disadvantage have low digital confidence, but that this is often underlined by a socially reinforced lifelong low self-confidence in their intellectual ability. In contrast, active users, regardless of other demographic factors, described themselves as innately proactive question askers. This was reinforced by our finding of a relationship between health care empowerment and the frequency of eHealth use. This suggests that while digital health education and eHealth literacy programs can improve access for some patients, broader and deeper long-term solutions addressing socioeconomic drivers of digital exclusion are needed to improve digital health access for some patients with socioeconomic disadvantage [8]. The deep permeation of socially enforced low self-confidence and lifelong poverty experienced by some interviewees demonstrate that the provision of free hardware and a class on digital health skills can be, for some, a superficial offering when the key underlying factor is persistent general socioeconomic inequality.

The digital health divide literature tends to identify the digital health divide, the factors and barriers that contribute to it, and the potential for it to widen if not specifically addressed [16]. Our findings have also identified the divide and the barriers, but what this study adds through our qualitative phase in particular is a description of the complex interaction of those barriers and the stepped nature of some of those barriers as part of the individual's experience in trying to access digital health.

Strengths and Limitations

A key strength of this study is the use of a sequential exploratory mixed methods design. The initial qualitative phase guided a phenomenological exploration of digital health access experiences for patients with chronic disease and socioeconomic disadvantage. Our results in both study phases stem from the patients' real-life experiences of digital health access. While some of our results echo the findings of other survey-based studies on general digital and digital health participation, our method revealed a greater depth and detail of some of these barriers, as demonstrated in how our findings compare to prior work.

As mentioned previously, the emphasis of this study on the qualitative first phase is a strength that helped describe the

XSL•FO
RenderX

interactions between different barriers. The interviewees described their experiences as cumulative unequal stepped barriers rather than as producing a nonordered list of equal barriers. These findings expand on the known complexity of the issue of digital exclusion and add weight to the understanding that improving digital health access needs diverse, complex solutions [17]. There is no panacea for every individual's digital health access, and thus, patient-centered digital health services, often guided by health professionals within the continuity of primary care, are also required to address the digital health divide [37].

While the sequential exploratory design is a strength of the study, it also created some limitations for the second quantitative phase. Our commitment to using the qualitative interview findings to inform the survey questions meant that we were unable to use previously validated scales for every question and that our results were less likely to lead to a normal distribution. This likely affected our ability to demonstrate significant associations for some barriers. We expect that further modeling is required to control for baseline characteristics and determine barrier patterns for different types of users.

One strength of this study is that the survey was administered to a broad population of Australian family medicine patients with diverse patterns of health via both paper-based and digital options. Many other digital health studies use solely digital surveys, which can affect the sample. However, we cannot draw conclusions from our survey about patients with chronic disease due to the limitations of the sample size for these subgroups.

Another sample-based limitation of this study was that our qualitative population did not include anyone aged from 18 to 24 years, despite multiple efforts to recruit. Future research will hopefully address this demographic more specifically.

While not strictly a limitation, we recognize that because this research was before COVID-19, it did not include questions about telehealth, which has become much more mainstream in recent years. The patients may also have changed their frequency of eHealth use because of COVID-19 and an increased reliance on digital services in general. Future work in this area or future versions of this survey should include telehealth and acknowledge the impact of COVID-19. However, the larger concept of the digital health divide exists before and after COVID-19, and in fact, our widespread increased reliance on digital services makes the digital divide an even more pressing issue [12].

Conclusions

The experience of digital health access across Australian primary care is highly variable and more difficult to access for those with socioeconomic disadvantage. While general digital interest, financial cost, and digital health literacy and empowerment are clear factors in digital health access in a broad primary care population, the digital health divide is also facilitated in part by a stepped series of complex and cumulative barriers.

Genuinely improving digital health access for 1 cohort or even 1 person requires a series of multiple different interventions tailored to specific sequential barriers. Given the rapid expansion of digital health during the global COVID-19 pandemic, attention to these issues is necessary if we are to avoid entrenching inequities in access to health care. Within primary care, patient-centered care that continues to recognize the complex individual needs of, and barriers facing, each patient should be a part of addressing the digital health divide.

Acknowledgments

The authors are thankful to the patients who shared their experiences with them via interview and survey completion. The authors are also very grateful to the general practices in the Australian Capital Territory and New South Wales who kindly gave their time and effort to help organize interviews, administer, and post surveys in the midst of the stress of day-to-day practice life and the bushfires of 2018-2019. The authors thank and acknowledge the creators of the eHealth Literacy Scale, the eHealth Literacy Questionnaire, the ICEpop Capability Measure for Adults, the Health Care Empowerment Inventory, the Patient-Doctor Relationship Questionnaire, the Chao continuity questionnaire, and the Southgate Institute for Health Society and Equity for their generosity in sharing their work with the authors [17,19-25]. This study would not have been possible without the support of the administrative team of the Academic Unit of General Practice. This project was funded by the Royal Australian College of General Practitioners (RACGP) through the RACGP Foundation IPN Medical Centres Grant, and the authors gratefully acknowledge their support.

Data Availability

The data sets generated during this study are not publicly available due to the nature of our original ethics approval but are available from the corresponding author on reasonable request.

Authors' Contributions

MAC acquired the funding, conceptualized the project, and organized interview recruitment. MAC and KB conducted interviews and analyzed the qualitative data. EAS, ER, and KD contributed to project planning, supervision and qualitative data analysis. MAC, KB and KO wrote the survey and planned quantitative data analysis. MAC and KB recruited practices for survey administration. KO and KB conducted the quantitative data analysis. MAC and KO, with KB drafted the paper. EAS, ER, and KD helped with reviewing and editing the paper.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Phase 1 interview guide.
[DOCX File , 31 KB - jmir_v26i1e50410_app1.docx]

Multimedia Appendix 2
Phase 2 survey: eHealth and digital divide.
[DOCX File , 461 KB - jmir_v26i1e50410_app2.docx]

Multimedia Appendix 3
Phase 2 participant characteristics by frequency of eHealth use.
[DOCX File , 28 KB - jmir_v26i1e50410_app3.docx]

References

1. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
2. Iyawa GE, Herselman M, Botha A. Digital health innovation ecosystems: from systematic literature review to conceptual framework. Procedia Comput Sci 2016;100:244-252 [FREE Full text] [doi: 10.1016/j.procs.2016.09.149]
3. Berrouiguet S, Baca-García E, Brandt S, Walter M, Courtet P. Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health. J Med Internet Res 2016 Jun 10;18(6):e135 [FREE Full text] [doi: 10.2196/jmir.5066] [Medline: 27287668]
4. Shen H, van der Kleij RM, van der Boog PJ, Chang X, Chavannes NH. Electronic health self-management interventions for patients with chronic kidney disease: systematic review of quantitative and qualitative evidence. J Med Internet Res 2019 Nov 05;21(11):e12384 [FREE Full text] [doi: 10.2196/12384] [Medline: 31687937]
5. Australia's health 2018. Australian Institute of Health and Welfare. 2018. URL: https://www.aihw.gov.au/reports/australias-health/australias-health-2018/contents/table-of-contents [accessed 2024-04-04]
6. Australian Institute of Health and Welfare. Chronic Diseases and Associated Risk Factors in Australia, 2006. Canberra, Australia: Australian Institute of Health and Welfare; 2006.
7. Hart JT. The inverse care law. The Lancet 1971 Feb 27;297(7696):405-412. [doi: 10.1016/S0140-6736(71)92410-X]
8. Davies AR, Honeyman M, Gann B. Addressing the digital inverse care law in the time of COVID-19: potential for digital technology to exacerbate or mitigate health inequalities. J Med Internet Res 2021 Apr 07;23(4):e21726 [FREE Full text] [doi: 10.2196/21726] [Medline: 33735096]
9. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 02;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
10. Household use of information technology. Australian Bureau of Statistics. 2018. URL: https://tinyurl.com/4efm6u92 [accessed 2024-03-24]
11. Kontos E, Blake KD, Chou WY, Prestin A. Predictors of eHealth usage: insights on the digital divide from the health information national trends survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
12. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440 [FREE Full text] [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
13. Butler DC, Joshy G, Douglas KA, Sayeed MS, Welsh J, Douglas A, et al. Changes in general practice use and costs with COVID-19 and telehealth initiatives: analysis of Australian whole-population linked data. Br J Gen Pract 2023 Apr 27;73(730):e364-e373 [FREE Full text] [doi: 10.3399/BJGP.2022.0351] [Medline: 37105730]
14. Arsenijevic J, Tummers L, Bosma N. Adherence to electronic health tools among vulnerable groups: systematic literature review and meta-analysis. J Med Internet Res 2020 Feb 06;22(2):e11613 [FREE Full text] [doi: 10.2196/11613] [Medline: 32027311]
15. Kontos EZ, Bennett GG, Viswanath K. Barriers and facilitators to home computer and internet use among urban novice computer users of low socioeconomic position. J Med Internet Res 2007 Oct 22;9(4):e31 [FREE Full text] [doi: 10.2196/jmir.9.4.e31] [Medline: 17951215]
16. Latulippe K, Hamel C, Giroux D. Social health inequalities and eHealth: a literature review with qualitative synthesis of theoretical and empirical studies. J Med Internet Res 2017 Apr 27;19(4):e136 [FREE Full text] [doi: 10.2196/jmir.6731] [Medline: 28450271]

XSL•FO
RenderX

17. Foley K, Freeman T, Ward P, Lawler A, Osborne R, Fisher M. Exploring access to, use of and benefits from population-oriented digital health services in Australia. Health Promot Int 2021 Aug 30;36(4):1105-1115. [doi: 10.1093/heapro/daaa145] [Medline: 33367568]
18. Cresswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: SAGE Publications; 2007.
19. Tappen RM, Cooley ME, Luckmann R, Panday S. Digital health information disparities in older adults: a mixed methods study. J Racial Ethn Health Disparities 2022 Feb;9(1):82-92 [FREE Full text] [doi: 10.1007/s40615-020-00931-3] [Medline: 33415705]
20. Who can get a card. Services Australia. URL: https://www.servicesaustralia.gov.au/who-can-get-health-care-card?context=21981 [accessed 2023-11-03]
21. Levesque JF, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int J Equity Health 2013 Mar 11;12:18 [FREE Full text] [doi: 10.1186/1475-9276-12-18] [Medline: 23496984]
22. Bryant A, Charmaz K. The SAGE Handbook of Grounded Theory, Paperback Edition. Thousand Oaks, CA: SAGE Publications; 2010.
23. Johnson MO, Rose CD, Dilworth SE, Neilands TB. Advances in the conceptualization and measurement of health care empowerment: development and validation of the health care empowerment inventory. PLoS One 2012;7(9):e45692 [FREE Full text] [doi: 10.1371/journal.pone.0045692] [Medline: 23029184]
24. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
25. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
26. Standards. Australian Bureau of Statistics. URL: https://www.abs.gov.au/statistics/standards [accessed 2024-04-04]
27. Al-Janabi H, Flynn TN, Coast J. Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. Qual Life Res 2012 Feb;21(1):167-176 [FREE Full text] [doi: 10.1007/s11136-011-9927-2] [Medline: 21598064]
28. Van der Feltz-Cornelis CM, Van Oppen P, Van Marwijk HW, De Beurs E, Van Dyck R. A patient-doctor relationship questionnaire (PDRQ-9) in primary care: development and psychometric evaluation. Gen Hosp Psychiatry 2004;26(2):115-120. [doi: 10.1016/j.genhosppsych.2003.08.010] [Medline: 15038928]
29. Chao J. Continuity of care: incorporating patient perceptions. Fam Med 1988;20(5):333-337. [Medline: 3266158]
30. Wilson CK, Thomas J, Barraket J. Measuring digital inequality in Australia: the Australian digital inclusion index. JTDE 2019 Jun 30;7(2):102-120. [doi: 10.18080/jtde.v7n2.187]
31. Digital participation: a view of Australia's online behaviours. Australia Post. 2017 Jul. URL: https://auspost.com.au/content/dam/auspost_corp/media/documents/white-paper-digital-inclusion.pdf [accessed 2024-04-04]
32. Poli A, Kelfve S, Motel-Klingebiel A. A research tool for measuring non-participation of older people in research on digital health. BMC Public Health 2019 Nov 08;19(1):1487 [FREE Full text] [doi: 10.1186/s12889-019-7830-x] [Medline: 31703655]
33. Cohen J. Statistical Power Analysis for the Behavioral Sciences Second Edition. London, UK: Routledge; 1988.
34. Borg K, Smith L. Digital inclusion and online behaviour: five typologies of Australian internet users. Behav Inf Technol 2018 Feb 15;37(4):367-380. [doi: 10.1080/0144929x.2018.1436593]
35. Mathers A, Richardson J, Vincent S, Joseph C, Stone E. Good Things Foundation COVID-19 response report. Good Things Foundation. 2020. URL: https://tinyurl.com/2peu3kak [accessed 2024-04-04]
36. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
37. Neves AL, Burgers J. Digital technologies in primary care: implications for patient care and future research. Eur J Gen Pract 2022 Dec 11;28(1):203-208 [FREE Full text] [doi: 10.1080/13814788.2022.2052041] [Medline: 35815445]

Abbreviations

eHEALS: eHealth Literacy Scale
GP: general practitioner
HCEI: Health Care Empowerment Inventory

XSL•FO
RenderX

Edited by T Leung; submitted 03.07.23; peer-reviewed by T Freeman, H Shen; comments to author 16.08.23; revised version received 30.11.23; accepted 31.01.24; published 11.04.24.

Please cite as:
Choy MA, O'Brien K, Barnes K, Sturgiss EA, Rieger E, Douglas K
Evaluating the Digital Health Experience for Patients in Primary Care: Mixed Methods Study
J Med Internet Res 2024;26:e50410
URL: https://www.jmir.org/2024/1/e50410
doi:10.2196/50410
PMID:38602768

©Melinda Ada Choy, Kathleen O'Brien, Katelyn Barnes, Elizabeth Ann Sturgiss, Elizabeth Rieger, Kirsty Douglas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating Patient Use and Experience of Online Appointment Booking in Primary Care: Mixed Methods Study

Helen Atherton[1], BSc, MSc, MPH, PhD; Abi Eccles[2], BA, DPhil; Leon Poltawski[3], BSc, PhD; Jeremy Dale[2], MBBS, MA; John Campbell[3], BSc, MBChB, MD, FRCGP; Gary Abel[3], BSc, MSc, PhD

[1]Primary Care Research Centre, School of Primary Care, Population Science, and Medical Education, University of Southampton, Southampton, United Kingdom

[2]Unit of Academic Primary Care, Warwick Medical School, University of Warwick, Coventry, United Kingdom

[3]University of Exeter Medical School, Exeter, United Kingdom

Corresponding Author:
Helen Atherton, BSc, MSc, MPH, PhD
Primary Care Research Centre
School of Primary Care, Population Science, and Medical Education
University of Southampton
Aldermoor Health Centre
Southampton, SO16 5ST
United Kingdom
Phone: 44 023 8059 5000
Email: h.atherton@soton.ac.uk

Abstract

Background: Online appointment booking is a commonly used tool in several industries. There is limited evidence about the benefits and challenges of using online appointment booking in health care settings. Potential benefits include convenience and the ability to track appointments, although some groups of patients may find it harder to engage with online appointment booking. We sought to understand how patients in England used and experienced online appointment booking.

Objective: This study aims to describe and compare the characteristics of patients in relation to their use of online appointment booking in general practice and investigate patients' views regarding online appointment booking arrangements.

Methods: This was a mixed methods study set in English general practice comprising a retrospective analysis of the General Practice Patient Survey (GPPS) and semistructured interviews with patients. Data used in the retrospective analysis comprised responses to the 2018 and 2019 GPPS analyzed using mixed-effects logistic regression. Semistructured interviews with purposively sampled patients from 11 general practices in England explored experiences of and views on online appointment booking. Framework analysis was used to allow for comparison with the findings of the retrospective analysis.

Results: The retrospective analysis included 1,327,693 GPPS responders (2018-2019 combined). We conducted 43 interviews with patients with a variety of experiences and awareness of online appointment booking; of these 43 patients, 6 (14%) were from ethnic minority groups. In the retrospective analysis, more patients were aware that online appointment booking was available (581,224/1,288,341, 45.11%) than had experience using it (203,184/1,301,694, 15.61%). There were deprivation gradients for awareness and use and a substantial decline in both awareness and use in patients aged >75 years. For interview participants, age and life stage were factors influencing experiences and perceptions, working patients valued convenience, and older patients preferred to use the telephone. Patients with long-term conditions were more aware of (odds ratio [OR] 1.43, 95% CI 1.41-1.44) and more likely to use (OR 1.65, 95% CI 1.63-1.67) online appointment booking. Interview participants with long-term conditions described online appointment booking as useful for routine nonurgent appointments. Patients in deprived areas were clustered in practices with low awareness and use of online appointment booking among GPPS respondents (OR for use 0.65, 95% CI 0.64-0.67). Other key findings included the influence of the availability of appointments online and differences in the registration process for accessing online booking.

Conclusions: Whether and how patients engage with online appointment booking is influenced by the practice with which they are registered, whether they live with long-term conditions, and their deprivation status. These factors should be considered in designing and implementing online appointment booking and have implications for patient engagement with the wider range of online services offered in general practice.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51931) doi:10.2196/51931

KEYWORDS

appointment; patient appointments; online systems; primary health care; general practice; qualitative research; secondary data analysis; mobile phone

Introduction

Background

Booking an appointment or service online is widespread internationally, with most sectors, including travel, entertainment, and hospitality, offering this facility. Health care has adopted online appointment booking with varying levels of patient uptake and engagement [1-3]. In England, most primary care is delivered by National Health Service (NHS) general practices to registered lists of patients. All NHS general practices must make appointments available for online booking [4,5], with appointments booked via web platform or app. This is part of the contract that general practices have with NHS England, the body responsible for leading the NHS in England. Despite the availability of online appointment booking being mandatory since 2012, uptake by patients has been slow and variable, with national-level data showing that, in March 2023, a total of 43% of patients were registered to use an online appointment booking service [6].

The use of online appointment booking is intended to lead to reduction in reception staff workload and increased patient satisfaction. It can offer flexibility, convenience, and time saving [7-13] and allow carers of older adults to make and track health care appointments [14]. It may reduce the likelihood of patients missing appointments [11]. It is known that people who use online appointment booking tend to be female, educated to degree level, and more frequent users of the internet [9,10,15,16]; use is lower in non-White patients, in lower socioeconomic groups, and in those with poorer health [9,10,15,16]. Patients from medically underserved and vulnerable populations are less likely than other patient groups to access and use online health technologies [17]. Specific barriers to engagement with online services for these groups include a lack of experience, knowledge, and skills when using the internet [17-19]; lack of technical support; and lower health literacy [17,20].

The routine use of online services in general practice settings, such as online triage platforms and video consultation [21-23], increased during the COVID-19 pandemic. Understanding who benefits and who does not benefit from the use of online appointment booking and why is important given the potential for inequality of access to primary care services to be exacerbated when services move online [19]. As an online service with a defined purpose and broad geographic availability, online appointment booking potentially provides a marker for how patients may engage with the wider range of online services within modern primary care. Evidence that seeks to understand how patients interact with online health care services is needed to shape such services to suit patients and their needs as we adjust to a new normal for primary care delivery. With levels of uptake of online appointment booking low in general practice and a lack of evidence on why this is the case, we conducted this mixed methods study.

Objectives

We sought to describe and compare the characteristics of patients in relation to their use of online appointment booking to explore how patients' use relates to their experience of care and investigate patients' views on online appointment booking arrangements in general practice.

Methods

Approach

We conducted a retrospective analysis of data on awareness and use of online appointment booking from the annual national General Practice Patient Survey (GPPS) conducted in English general practices. This was complemented by a semistructured interview study with patients from a range of general practices in England with differing levels of uptake of online appointment booking. The interviews were conducted both before and during the COVID-19 pandemic.

In UK general practice, patients can book an appointment online directly with the general practice they are registered with. Patients can do this using the NHS website, app, or via a web platform with app option provided by their general practice. General practices use a web platform often with app option, of their own choosing to offer online appointment booking, and there are multiple suppliers of these systems in the UK market. Patients can use these services via an internet-enabled device (eg, smartphone, tablet, or computer) according to their preference. These systems are not necessarily linked to the electronic health record, which is also chosen by the general practice from a range of available systems. This study examined online appointment booking across all platform and associated app types, and we did not set out to compare platforms as all offer online appointment booking in line with NHS requirements.

Retrospective Analysis of GPPS

Data Source

We conducted secondary analyses of data from the English GPPS [24]. The GPPS is a postal and online survey of patients' experience of primary care in England. The survey is sent annually (January-March) to approximately 2.2 million adult patients, with the findings published each July [25]. A stratified sample of patients aged >16 years is drawn from the practice list of each general practice in England. Patients from practices known from previous surveys to have low response rates are oversampled to ensure an adequate number of responses per practice. We used data from the 2018 and 2019 surveys making use of 2 questions, one asking about awareness of online services in the respondents' general practice and another asking about

XSL•FO
RenderX

their use of online services in the previous 12 months. These were used in conjunction with data from the survey on demographic characteristics and health status [26].

Analysis

Descriptive analyses and logistic regression were used to investigate the percentage of patients who reported awareness of their general practice offering online appointment booking and recent use of online appointment booking. These analyses were restricted to patients reporting trying to make an appointment at their general practice in the last 12 months. This is important because patients may not have used online appointment booking (or been aware of the option) due to not needing an appointment.

Multivariate logistic regression was used to examine associations between both awareness and use of online appointment booking and age, gender, ethnicity, deprivation (based on the Index of Multiple Deprivation [IMD] [27] corresponding to the respondents' postcode of residence), presence of a long-term condition, long-term sickness, and being deaf.

Comparison of models accounting and not accounting for clustering by practice (using a random effect) was used to illustrate the extent to which disparities reflect the clustering of certain types of patients in practices where awareness and use of online appointment booking is high or low for all patient groups. The random effect from these models was used to quantify the variability in the odds of patients being aware of or using online appointment booking between practices.

Further models augmented those described previously to include other GPPS report and evaluation items (as predictors of online appointment booking awareness and use) regarding how easy practice websites were to use, whether respondents had a preferred general practitioner (GP), how helpful receptionists were, ease of getting through on the phone, and use of online prescription ordering. A final set of regression models examined the extent to which awareness and use of online appointment booking were associated with patients' experiences of making an appointment, choice of appointment, ability to see their preferred GP, and their overall experience. These models were adjusted for the same patient factors included in the previous models as well as including a random effect for practice.

Full statistical details are provided in Multimedia Appendix 1. Analysis was conducted using Stata (version 12; StataCorp).

Semistructured Interview Study

Sampling and Recruitment

Interview participants were recruited from 11 general practices selected for maximum variation in levels of online booking, list size, rurality, deprivation, and ethnic diversity. There were three recruitment waves: (1) a scoping survey was distributed to patients to gather data on their demographic characteristics (gender, age, educational level, employment status, health status, and ethnicity), booking behaviors and awareness of online booking, and willingness to participate in an interview; (2) interview volunteers were purposively sampled according to background and booking behavior and invited to take part in an interview; and (3) finally, aiming to fill demographic gaps within the sample, potential participants were identified based on their background via general practices and invited directly to interview. Patients who were aged <18 years or at end of life, lacked capacity to provide informed consent, or had not recently booked a general practice appointment were excluded from participation. The COVID-19 pandemic meant that recruitment was paused as the funder made the decision to halt all research activity not related to COVID-19, allowing practices to focus on their COVID-19 response. Recruitment was resumed in June 2020, at which time we made use of direct invites to interviews (wave 3 detailed previously) to make up the remainder of the sample.

Data Collection

A total of 43 interviews were carried out via telephone, audio recorded, transcribed, and anonymized.

During the interviews, we explored in-depth patients' experiences of and views on booking appointments online, including how they did this and, for those who had not used it, whether they knew about it or not. The topic guide can be found in Multimedia Appendix 2.

Data Analysis

We used the framework method to conduct a comparative thematic analysis of the interview data relevant to the key findings from the retrospective analysis of GPPS data. The framework method is commonly used in health research in which research questions are relevant to policy. It has 5 key stages: familiarization, identifying a thematic framework, indexing, charting, and mapping and interpretation [28]. A matrix was used with "cases" arranged along one axis and "codes" along the other. This allowed data to be organized systematically to compare differing experiences and perspectives on key areas from participants' accounts. The NVivo software (version 12; QSR International) was used to support this process.

Patient and Public Involvement

We involved members of the public at various stages of the study. A lay coapplicant joined the study team and attended meetings, provided advice throughout, and helped facilitate workshops. Public workshops were held to refine our protocol (before funding was awarded), gain input on the participant-facing documentation, and discuss the implications of the study findings. We piloted the scoping survey and telephone interviews with 3 contributors. In total, 2 contributors also provided their interpretation of the qualitative data.

Ethical Considerations

We obtained NHS research ethics approval and Health Research Authority approval for the semistructured interview study from the West Midlands – Solihull Research Ethics Committee (study reference: 19/WM/0272). The secondary analysis of GPPS data did not require ethics or Health Research Authority approval.

Respondents to the GPPS are anonymous with their identifying data not collected at initial data collection. In the information provided with the survey, potential respondents are informed that their responses may be shared with NHS approved researchers. Providing a response to the survey is taken as

XSL•FO
RenderX

implied consent to agree to the conditions laid out in the information provided.

Informed consent was obtained from all participants in the semistructured interview study.

Interview transcripts were deidentified, allocated an identifying code and stored separately from participant information to ensure anonymity. All data were stored in password protected secure electronic storage. Participants in the semistructured interview study were given a GBP £10 (US $12.7) shopping voucher as a thank you for the time taken in participating. Interview quotes are used with consent, and basic demographic information is supplied with the quotes for context.

Results

Retrospective Analysis of GPPS

Overview

Of the 1,327,693 GPPS responders (2018-2019 combined) who reported attempting to make an appointment within the previous 12 months, 45.11% (581,224/1,288,341) were aware that their practice offered online appointment booking, and 15.61% (203,184/1,301,694) reported using online appointment booking. We found very strong evidence ($P<.001$) of associations between both awareness and use of online appointment booking and all variables considered in both unadjusted and adjusted models except for the effect of rurality in the adjusted model, including a random effect for practice when considering awareness (Tables 1 and 2).

There were only small differences between the unadjusted and adjusted models, with consistent direction of association except for awareness among patients aged 25 to 44 years and for rural versus urban location (Tables 1 and 2). Findings from the adjusted model included a random effect for practice and so can be interpreted as differences experienced by patients registered with the same general practice.

XSL•FO
RenderX

Table 1. Awareness of online booking of general practice appointments (N=1,288,341).

Variable	Frequency[a]		Unadjusted[b], OR[c] (95% CI)	Adjusted[b], OR (95% CI)	Adjusted+RE[b,d], OR (95% CI)
	Total respondents, n (%)	Respondents aware of online booking (n=581,224), n (%)			
Sex					
Male	587,137 (45.57)	245,470 (41.81)	Reference	Reference	Reference
Female	688,801 (53.46)	331,432 (48.12)	1.24 (1.23-1.25)	1.25 (1.24-1.26)	1.27 (1.26-1.28)
Age group (y)					
16-24	122,879 (9.54)	50,573 (41.16)	0.75 (0.73-0.76)	0.89 (0.87-0.90)	0.86 (0.84-0.88)
25-34	208,897 (16.21)	95,574 (45.75)	0.94 (0.93-0.95)	1.13 (1.11-1.15)	1.10 (1.08-1.12)
35-44	214,993 (16.69)	99,001 (46.05)	0.96 (0.95-0.97)	1.13 (1.11-1.15)	1.09 (1.08-1.11)
45-54	230,452 (17.89)	108,454 (47.06)	1.00 (0.99-1.01)	1.11 (1.09-1.12)	1.09 (1.08-1.10)
55-64	203,277 (15.78)	99,884 (49.14)	1.07 (1.06-1.08)	1.13 (1.12-1.14)	1.13 (1.12-1.15)
65-74	168,014 (13.04)	80,836 (48.11)	Reference	Reference	Reference
75-84	94,412 (7.33)	34,188 (36.21)	0.62 (0.62-0.63)	0.61 (0.60-0.62)	0.58 (0.57-0.59)
≥85	34,431 (2.67)	8915 (25.89)	0.40 (0.39-0.41)	0.38 (0.37-0.39)	0.34 (0.33-0.35)
IMD[e] quintile					
1—least deprived	249,352 (19.35)	128,348 (51.47)	Reference	Reference	Reference
2	253,343 (19.66)	121,617 (48.00)	0.87 (0.86-0.88)	0.87 (0.86-0.88)	0.91 (0.90-0.92)
3	256,313 (19.89)	117,260 (45.75)	0.78 (0.77-0.78)	0.77 (0.77-0.78)	0.85 (0.84-0.86)
4	262,693 (20.39)	113,058 (43.04)	0.69 (0.68-0.69)	0.69 (0.68-0.69)	0.77 (0.76-0.78)
5—most deprived	265,750 (20.63)	100,578 (37.85)	0.54 (0.53-0.55)	0.54 (0.53-0.55)	0.66 (0.65-0.67)
Rurality					
Urban	1,112,590 (86.36)	502,382 (45.15)	Reference	Reference	Reference
Rural	175,751 (13.64)	78,843 (44.86)	1.08 (1.07-1.09)	0.96 (0.95-0.97)	1.01 (0.96-1.06)
Ethnic group					
Asian	107,325 (8.33)	44,346 (41.32)	0.81 (0.80-0.82)	0.87 (0.86-0.89)	0.89 (0.88-0.91)
Black	41,896 (3.25)	16,085 (38.39)	0.73 (0.71-0.74)	0.83 (0.81-0.85)	0.79 (0.77-0.81)
Mixed	19,296 (1.5)	8,557 (44.35)	0.94 (0.91-0.97)	0.96 (0.93-1.00)	0.91 (0.88-0.95)
White	1,076,912 (83.59)	497,261 (46.17)	Reference	Reference	Reference
Other	24,836 (1.93)	8306 (33.44)	0.59 (0.57-0.60)	0.65 (0.63-0.68)	0.63 (0.61-0.65)
Long-term condition					
Yes	689,896 (53.55)	328,180 (47.57)	1.15 (1.15-1.62)	1.38 (1.36-1.39)	1.43 (1.41-1.44)
No	531,919 (41.29)	228,690 (42.99)	Reference	Reference	Reference
Occupation					
Other	1,187,020 (92.14)	540,342 (45.52)	Reference	Reference	Reference
Sick, disability, or both	50,585 (3.93)	21,152 (41.81)	0.84 (0.82-0.85)	0.80 (0.78-0.81)	0.80 (0.79-0.82)
Deafness and sign language use					
Yes	5508 (0.43)	1933 (35.09)	0.65 (0.62-0.69)	0.78 (0.73-0.84)	0.76 (0.71-0.81)
No	1,266,842 (98.33)	573,583 (45.28)	Reference	Reference	Reference
Period					
2018	641,337 (49.78)	277,392 (43.25)	Reference	Reference	Reference
2019	647,004 (50.22)	303,832 (46.69)	1.17 (1.16-1.18)	1.18 (1.17-1.19)	1.19 (1.18-1.20)

XSL•FO
RenderX

[a]Frequency numbers are weighted and rounded. Regression results are not weighted.

[b]*P* values from joint Wald test were computed for the adjusted and unadjusted models. All *P* values are <.001 except that corresponding to rurality in the mixed-effects logistic regression model.

[c]OR: odds ratio.

[d]RE: random effect.

[e]IMD: Index of Multiple Deprivation.

XSL•FO
RenderX

Table 2. Variation in use of online booking of general practice appointments (N=1,301,694).

Variable	Frequency[a]		Unadjusted[b], OR[c] (95% CI)	Adjusted[b], OR (95% CI)	Adjusted+RE[b,d], OR (95% CI)
	Total respondents, n (%)	Respondents aware of online booking (n=203,184), n (%)			
Sex					
Male	592,522 (45.52)	89,505 (15.11)	Reference	Reference	Reference
Female	696,669 (53.52)	112,135 (16.1)	1.04 (1.03-1.05)	1.04 (1.03-1.05)	1.04 (1.02-1.05)
Age group (y)					
16-24	124,337 (9.55)	17,548 (14.11)	0.94 (0.91-0.96)	1.14 (1.11-1.18)	1.13 (1.09-1.16)
25-34	211,051 (16.21)	36,138 (17.12)	1.21 (1.18-1.23)	1.52 (1.49-1.55)	1.48 (1.45-1.51)
35-44	217,280 (16.69)	37,319 (17.18)	1.22 (1.20-1.24)	1.47 (1.44-1.49)	1.43 (1.40-1.46)
45-54	232,622 (17.87)	39,142 (16.83)	1.19 (1.17-1.21)	1.34 (1.32-1.37)	1.34 (1.31-1.36)
55-64	205,150 (15.76)	34,604 (16.87)	1.16 (1.14-1.18)	1.23 (1.21-1.25)	1.24 (1.22-1.26)
65-74	169,731 (13.04)	25,833 (15.22)	Reference	Reference	Reference
75-84	95,635 (7.35)	9003 (9.41)	0.59 (0.58-0.60)	0.57 (0.56-0.59)	0.56 (0.55-0.57)
≥85	34,840 (2.68)	2217 (6.36)	0.41 (0.40-0.43)	0.37 (0.36-0.39)	0.35 (0.34-0.37)
IMD[e] quintile					
1—least deprived	251,435 (19.32)	45,936 (18.27)	Reference	Reference	Reference
2	255,837 (19.65)	42,270 (16.52)	0.88 (0.87-0.90)	0.87 (0.86-0.89)	0.90 (0.89-0.92)
3	258,902 (19.89)	39,996 (15.45)	0.79 (0.78-0.81)	0.77 (0.76-0.79)	0.83 (0.81-0.84)
4	265,735 (20.41)	39,569 (14.89)	0.75 (0.74-0.76)	0.70 (0.69-0.72)	0.75 (0.74-0.77)
5—most deprived	268,897 (20.73)	35,270 (13.12)	0.62 (0.61-0.63)	0.57 (0.56-0.58)	0.65 (0.64-0.67)
Rurality					
Urban	1,124,219 (86.37)	178,247 (15.86)	Reference	Reference	Reference
Rural	177,475 (13.63)	24,937 (14.05)	0.97 (0.95-0.98)	0.91 (0.90-0.93)	0.92 (0.88-0.97)
Ethnic group					
Asian	108,829 (8.36)	18,500 (17.00)	1.10 (1.08-1.12)	1.11 (1.08-1.13)	1.02 (1.00-1.05)
Black	42,387 (3.26)	5666 (13.37)	0.83 (0.81-0.86)	0.89 (0.86-0.92)	0.76 (0.73-0.78)
Mixed	19,511 (1.5)	3,475 (17.81)	1.19 (1.14-1.25)	1.15 (1.10-1.21)	1.03 (0.98-1.09)
White	1,087,517 (83.55)	169,378 (15.57)	Reference	Reference	Reference
Other	25,204 (1.94)	3777 (14.99)	1.00 (0.96-1.04)	1.02 (0.97-1.06)	0.92 (0.87-0.96)
Long-term condition					
Yes	697,247 (53.56)	120,347 (17.26)	1.24 (1.23-1.25)	1.59 (1.57-1.61)	1.65 (1.63-1.67)
No	537,166 (41.27)	74,212 (13.82)	Reference	Reference	Reference
Occupation					
Other	1,198,858 (92.1)	189,142 (15.78)	Reference	Reference	Reference
Sick, disability, or both	51,258 (3.94)	7646 (14.92)	0.94 (0.92-0.97)	0.85 (0.82-0.87)	0.86 (0.84-0.89)
Deafness and sign language use					
Yes	5611 (0.43)	1046 (18.64)	1.18 (1.10-1.27)	1.23 (1.13-1.33)	1.21 (1.11-1.32)
No	1,280,135 (98.34)	200,171 (15.64)	Reference	Reference	Reference
Period					
2018	647,454 (49.74)	93,728 (14.48)	Reference	Reference	Reference

XSL•FO
RenderX

Variable	Frequency[a]		Unadjusted[b], OR[c] (95% CI)	Adjusted[b], OR (95% CI)	Adjusted+RE[b,d], OR (95% CI)
	Total respondents, n (%)	Respondents aware of online booking (n=203,184), n (%)			
2019	654,240 (50.26)	109,456 (16.73)	1.18 (1.17-1.20)	1.19 (1.18-1.20)	1.19 (1.18-1.21)

[a]Frequency numbers are weighted and rounded. Regression results are not weighted.

[b]*P* values from joint Wald test were computed for the adjusted and unadjusted models. All *P* values are <.001.

[c]OR: odds ratio.

[d]RE: random effect.

[e]IMD: Index of Multiple Deprivation.

Age

There was little variability in both awareness and use among those aged <65 years but a substantial drop in both for those aged >75 years (odds ratio [OR] for awareness [those aged ≥85 years vs 65-74 years] 0.34, 95% CI 0.33-0.37; Tables 1 and 2 and Figure 1).

Figure 1. Age and awareness and use of online appointment booking. OR: odds ratio.

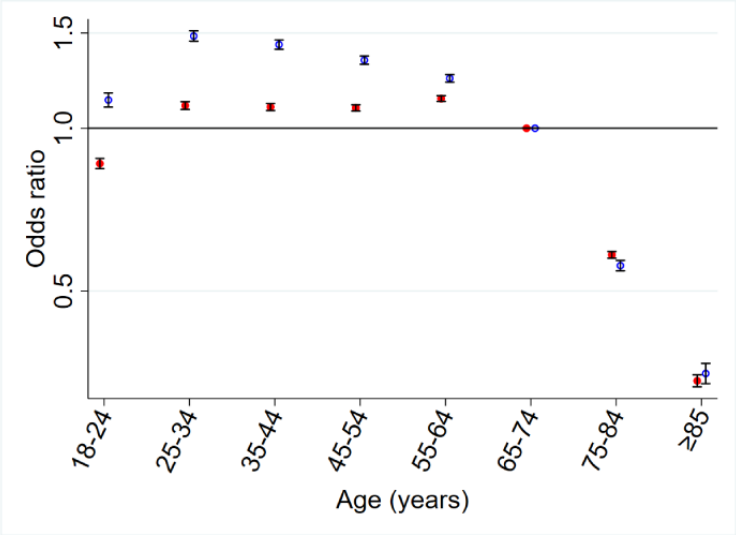

Deprivation Status

There was a strong deprivation gradient, with more deprived patients being less likely to use or be aware of online appointment booking (OR for use [most vs least deprived] 0.65, 95% CI 0.64-0.67; Tables 1 and 2 and Figure 2).

Figure 2. Deprivation and awareness and use of online appointment booking. IMD: Index of Multiple Deprivation.

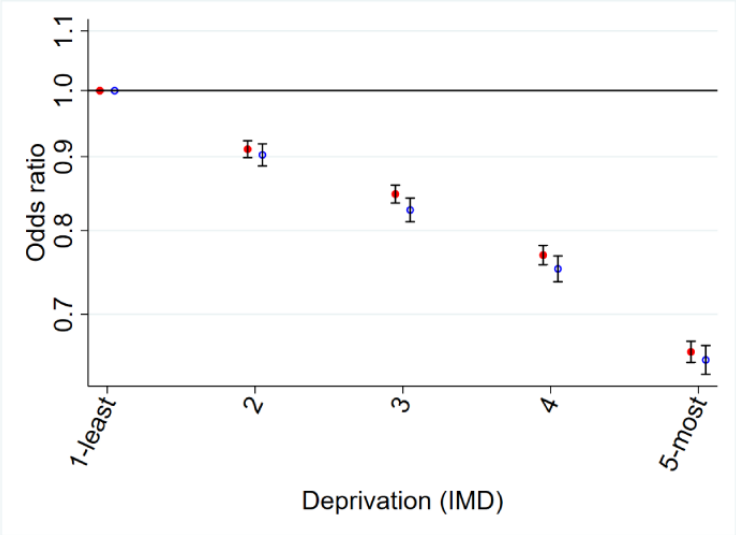

XSL•FO
RenderX

Ethnicity

A more complex relationship was observed for ethnicity. All patients from ethnic minority groups were less likely to be aware of online appointment booking than White patients. However, only patients in the Black and *Other* ethnic groups were less likely to use online appointment booking than White patients, and mixed and Asian patients were somewhat more likely to use online appointment booking (Tables 1 and 2 and Figure 3).

Figure 3. Ethnicity and awareness and use of online appointment booking.

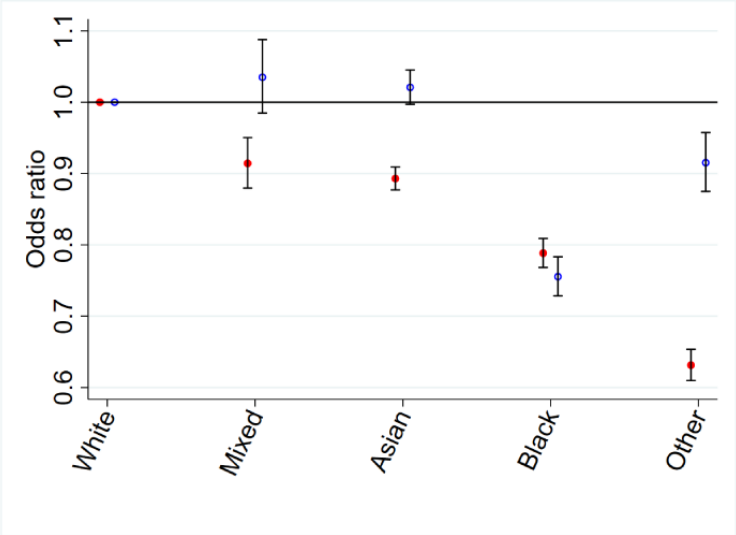

Location

Patients living in rural areas were less likely to be aware of or use online appointment booking than patients living in urban areas (OR for use [rural vs urban] 0.92, 95% CI 0.88-0.97; Tables 1 and 2 and Figure 4).

Figure 4. Location and long-term conditions (LTCs) and awareness, and use, of online appointment booking. LTS: long-term sickness.

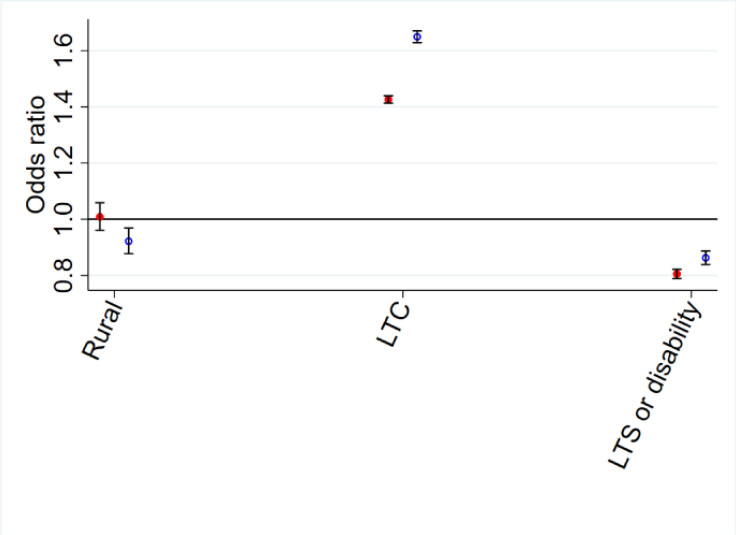

Health Status

Patients who were permanently sick or had a disability were less likely to be aware of or use online appointment booking than those living in urban areas and patients who were not sick or patients without disabilities (OR for awareness [sick or disability vs not sick or no disability] 0.80, 95% CI 0.78-0.82; Tables 1 and 2 and Figure 4).

Patients with long-term conditions were more likely to be aware of and use online appointment booking than patients without them (OR for awareness [long-term condition vs no long-term condition] 1.43, 95% CI 1.41-1.44; Tables 1 and 2 and Figure 4). Interestingly, patients who were deaf and used sign language were less likely to be aware of but more likely to use online appointment booking than those without such an impairment (OR for use [deaf vs not deaf] 1.21, 95% CI 1.11-1.32).

Year

Comparing the GPPS data from the 2018 and 2019 surveys, both awareness and use increased across this period (OR for awareness in 2019 vs 2018, 1.19, 95% CI 1.18-1.20; OR for use in 2019 vs 2018, 1.19, 95% CI 1.18-1.21).

XSL•FO
RenderX

Practice

When comparing models that did and did not account for the patients' registered practice, the patient groups that were less likely to use or be aware of online appointment booking remained the same between models (except for rural patients' awareness, where the effect was small in both cases). The strongest factor in determining awareness and use of online appointment booking was the practice the patient was registered with. For all factors apart from deprivation, the changes in OR magnitudes were small (Tables 1 and 2). The deprivation gradient in awareness and use of online appointment booking was noticeably stronger in models that did not incorporate adjustment for practice, which implies that patients living in deprived areas are more likely to be registered at practices where awareness and use is low for all patients (Tables 1 and 2).

The OR covering 95% of practices (comparing the highest with the lowest) estimated from the random-effects models was 18.1 (95% CI 17.2-19.0) for awareness and 16.7 (95% CI 15.8-17.7) for use. These are considerably larger than the ORs associated with any patient factor, suggesting that the practice a patient was registered with was the strongest predictor of awareness and use of online appointment booking when compared with a wide range of other potentially important variables.

Patient Experience

Patients reporting their practice website being easy to use were far more likely to be aware of and use online appointment booking (OR for use ["Very easy to use" vs "Not easy at all"] 6.34, 95% CI 6.06-6.64; Table 3). Having a preferred GP was associated with somewhat higher awareness and use, whereas ease of getting through to a practice on the phone was associated with lower use and somewhat lower awareness. Patients who reported finding practice receptionists helpful were more likely to be aware of online booking but less likely to use it. Patients who had ordered repeat prescriptions online were also more likely to be aware of and use online appointment booking than those who had not (OR 3.58, 95% CI 3.50-3.62).

Table 3. Odds ratio (OR) for awareness and use of online booking adjusted for demographic factors, presence of long-term condition, occupation, deafness, and period and for clustering of patients within practices.

Variable	Awareness		Use	
	OR (95% CI)	*P* value[a]	OR (95% CI)	*P* value[a]
Have you ordered repeat prescriptions online?		<.001		<.001
No	Reference		Reference	
Yes	1.84 (1.80-1.87)		3.58 (3.50-3.62)	
How easy is it to get through to someone at your GP[b] on the phone?		<.001		<.001
Very easy	0.85 (0.82-0.88)		0.63 (0.61-0.65)	
Fairly easy	0.97 (0.94-1.00)		0.80 (0.77-0.82)	
Not very easy	0.99 (0.96-1.02)		0.90 (0.87-0.93)	
Not at all easy	Reference		Reference	
How helpful do you find the receptionist at your GP?		<.001		<.001
Very helpful	1.15 (1.09-1.21)		0.72 (0.68-0.76)	
Fairly helpful	1.12 (1.06-1.18)		0.83 (0.79-0.88)	
Not very helpful	1.03 (0.97-1.08)		0.91 (0.86-0.97)	
Not at all helpful	Reference		Reference	
Is there a particular general practitioner you usually prefer to see or speak to?		<.001		<.001
Yes, for all appointments	1.19 (1.16-1.21)		1.61 (1.58-1.64)	
Yes, for some appointments	1.29 (1.26-1.32)		1.56 (1.53-1.59)	
No	Reference		Reference	
How easy is it to use your GP website?		<.001		<.001
Very easy	4.34 (4.19-4.51)		6.34 (6.06-6.64)	
Fairly easy	2.77 (2.68-2.86)		3.68 (3.52-3.85)	
Not very easy	1.58 (1.53-1.64)		1.81 (1.73-1.90)	
Not easy at all	Reference		Reference	

[a]*P* values from joint Wald tests.

[b]GP: general practice.

XSL•FO
RenderX

A final set of models (Table 4) showed that, regardless of whether patients had tried online booking, those who were aware of it were more likely than patients who were not aware of it to have a positive overall experience of their general practice, see or speak with their preferred GP, be offered a choice of appointment, or report a positive overall experience of making an appointment. Furthermore, patients who were aware of but had not used online appointment booking were more likely than unaware patients to be satisfied with their appointment type. Surprisingly, patients who were aware of and used online appointment booking were less likely than patients who were not aware of it to be satisfied with their appointment type.

Table 4. Odds ratio (OR) for different outcome measures adjusted for demographic factors, presence of long-term condition, occupation, deafness, and period and for clustering of patients within practices.

Awareness and use of online appointment booking	OR (95% CI)	*P* value[a]
Outcome—overall experience of general practice[b]		<.001
Unaware	Reference	
Aware but not used	1.34 (1.32-1.36)	
Aware and used	1.27 (1.25-1.30)	
Outcome—see or speak with preferred GP[c,d]		<.001
Unaware	Reference	
Aware but not used	1.25 (1.23-1.27)	
Aware and used	1.43 (1.40-1.45)	
Outcome—offered a choice of appointment[e]		<.001
Unaware	Reference	
Aware but not used	1.17 (1.16-1.18)	
Aware and used	1.42 (1.40-1.44)	
Outcome—satisfied with appointment type[f]		<.001
Unaware	Reference	
Aware but not used	1.16 (1.14-1.17)	
Aware and used	0.95 (0.93-0.96)	
Outcome—overall experience of making an appointment[g]		<.001
Unaware	Reference	
Aware but not used	1.22 (1.20-1.23)	
Aware and used	1.07 (1.06-1.09)	

[a]*P* values from joint Wald tests.
[b]Overall, how would you describe your experience of your general practice?
[c]GP: general practitioner.
[d]How often do you see or speak to your preferred GP when you would like to?
[e]On this occasion, were you offered a choice of appointment?
[f]Were you satisfied with the type of appointment you were offered?
[g]Overall, how would you describe your experience of making an appointment?

Semistructured Interview Study

Overview

We conducted 43 semistructured qualitative interviews with individuals who had recently booked general practice appointments (Tables 5 and 6). Participants were aged 18 to ≥85 years and had a range of educational levels and employment statuses. In total, 47% (20/43) of the participants reported having at least one long-term condition. Of the 43 participants, 18 (42%) had tried online appointment booking, 16 (37%) were aware that they could access online appointment booking but had not tried it, and 9 (21%) were unaware that online appointment booking was available.

XSL•FO
RenderX

Table 5. Characteristics of participating patients (N=43).

Characteristic	Participants, n (%)
Sex	
Male	22 (51)
Female	21 (49)
Age (y)	
18-24	2 (5)
25-34	5 (12)
35-44	6 (14)
45-54	5 (12)
55-64	9 (21)
65-74	6 (14)
75-84	6 (14)
≥85	3 (7)
Educational level	
None	2 (5)
Secondary education	6 (14)
Further education (16-18 years)	4 (9)
Higher education (18+ years)	24 (56)
Other or do not know	5 (12)
Employment status	
Full time	13 (30)
Part time	3 (7)
Student	2 (5)
Unemployed	1 (2)
Sick, disability, or both	5 (12)
Retired	11 (26)
Home or family	2 (5)
Other	5 (12)
Condition	
Long-term condition or conditions	20 (47)
No long-term condition	6 (14)
Unknown	1 (2)
Ethnicity	
White—English, Welsh, Scottish, Northern Irish, or British	34 (79)
White other	3 (7)
Other ethnic group	6 (14)
Experience with online booking	
Awareness	
Tried online booking	18 (42)
Aware but not tried online booking	16 (37)
Unaware	9 (21)
Use	
Never use online booking	29 (67)

XSL•FO
RenderX

Characteristic	Participants, n (%)
Sometimes use online booking	3 (7)
Frequently use online booking	11 (26)

Table 6. Demographic details on participating general practices in the interview study (N=11).

	Practices, n (%)
Online booking use	
High	4 (36)
Medium	6 (55)
Low	1 (9)
Rural	3 (27)
Semirural	3 (27)
Urban	5 (45)
List size	
Small	3 (27)
Medium	4 (36)
Large	4 (36)
IMD[a]	
Below average	6 (55)
Above average	5 (45)
Coventry and Warwickshire	7 (64)
Devon	4 (36)

[a]IMD: Index of Multiple Deprivation.

The framework analysis identified 3 key areas in which there was comprehensive qualitative evidence that matched the data identified in the quantitative element of the study. These were (1) general practice–mediated factors, (2) impact of age and life stage, and (3) specific experiences of those with a long-term condition. Patients also shared information about their differing levels of confidence and experience using online services in general.

General Practice–Mediated Factors

How the general practice organized online appointment booking systems influenced patient experience.

Patients had varied awareness and understanding of the online appointment booking system available to them. Routes to learn about online appointment booking included advertisements on noticeboards, SMS text messages, letters, or being told directly by practice staff. Despite all participating practices having online appointment booking available, many participants were not aware of it and described not having seen or heard any information about it:

> *The first time I went the reception, what do you call it the receptionist? She turned round and said, "Well, you can also book online."* [Male participant; aged 55-64 years; practice 4; White British]

> *I mean, you could be on hold for up to half an hour trying to call through. And the thing is, it's never been advertised to us that you could book online. Like, I've never seen it anywhere or been informed in any communication that you can do it online, so that's why that's always been my way of booking it.* [Male participant; aged 25-34 years; practice 4; White British]

The practices varied in the processes that patients described having to follow to register for online appointment booking. The most common requirement was to attend the practice in person to collect a form to complete and then return with identification before they could be registered. In contrast, other participants stated that their practice had not required them to show any identification.

Providing identification required the patient to be proactive in seeking registration. There were several examples of participants describing an intention to register and then not completing the process:

> *I got as far as bringing, bringing the form home and filling it in. But then because I don't go past the surgery on a daily basis I never took it, I never took it back.* [Male participant; aged 55-64 years; practice 7; White British]

There was also variation among general practices in the appointments made available to book online, with some patients describing that their practice appeared to make very few appointments available. A lack of appointments that could be

XSL•FO
RenderX

booked urgently or soon was experienced as a key barrier to use in cases in which patients needed this kind of appointment:

> *Right, I have tried to book online...and I saw that there was the option. [...] And that thrill was met with disappointment when I saw the lack of slots available online.* [Female participant; aged 25-34 years; practice 6; White European]

Where participants found that they could book appointments on the same day or within the week, this was a key facilitator to using and continuing to use online appointment booking:

> *I much prefer it, it's, it's quite, quite easy to learn how to do it, it's, is simple to use, it's fast, and it's reliable. [...] I've managed to get an appointment within a day or two sometimes.* [Male participant; aged ≥85 years; practice 12; White British]

Overall, online appointment booking was seen as fitting into a modern society, which relies heavily on online services:

> *It's probably very, very good for some people, but it's not for me. I appreciate that we are in a digital age and we, you know, we're relying on these, the computer more, and online, and e-mails and stuff, but no, I'm, I really don't think it is an advantage in some cases. It's not for me.* [Female participant; aged 65-74 years; practice 10; White Irish]

Patients, including those who did not use online appointment booking, were accepting of its value in the context of it being available as one of several options for accessing an appointment in general practice. Across the sample, online appointment booking was seen as a tool that was a good fit for some patients in some circumstances:

> *Well, I think in this day and age most people can do it online [...] I think a range of all three, I mean, going forward I suppose with staffing and stuff, you know, they're obviously looking to minimise the face-to-face contact, but I still think there's a, a need for that at, at some, some level for everybody.* [Female participant; aged 55-64 years; practice 12; White British]

Age and Life Stage

The convenience of online appointment booking was particularly appealing for those in the workforce. Booking an appointment via telephone could be problematic due to having to call while at work or the lack of privacy in an employment setting:

> *If you're at work then you, you need permission off your managers that you're going to be on the phone, especially if you work in a very busy environment and there's just no way you can just be on the phone for ages. You know, like, you might be Number 23 [...] in the, in the order, so you're going to wait and wait, so, you know, sometimes it's a bit difficult. [...] And then you, waiting, what, when you do get on the phone then they will ask you what the problem is and sometimes you don't really want to say, it's quite confidential and personal.* [Female participant; aged 55-64 years; practice 4; Asian British]

Online appointment booking was advantageous when childcare or caring responsibilities made it difficult for patients to call the general practice when telephone lines opened in the morning:

> *When she was at primary school and my eldest was at primary school, I wouldn't, I wouldn't be able to ring at half eight because I'd be in the car, or I'd be stood in the playground. You know, it, it doesn't work.* [Female participant; aged 45-54 years; practice 11; White British]

Some patients regarded their age as leading to a reluctance to use online appointment booking. Online appointment booking was regarded as something new to learn that was unfamiliar. Some participants did not use the internet regularly, did not use smartphones to access such services, and saw switching on a computer as inconvenient:

> *People that are younger than me and just do this automatically online, you know, whereas I don't do it automatically.* [Female participant; aged 75-84 years; practice 12; White British]

Requiring help to use online appointment booking was seen as disempowering. Older patients were accustomed to booking an appointment using the analog telephone and did not explore the possibility of using another route. Sometimes, using online appointment booking meant relying on family members or a partner to do the "internet admin" in whatever format this might take.

Differing Experiences of Those With a Long-Term Condition

Patients in the sample who had experience using online appointment booking and had long-term conditions found that the service was a good fit for them. Participants with long-term conditions who used online appointment booking said that they had only become aware of and subsequently used online appointment booking when accessing the online system to obtain their repeat prescriptions for their long-term condition:

> *I actually first used the app to get a repeat prescription [...] Then I noticed that you could also book appointments on there.* [Female participant; aged 25-34 years; practice 4; White British]

Accessing repeat prescriptions provided a gateway to use other online services, including online appointment booking, and this was also demonstrated in the quantitative findings.

Patients with long-term conditions were used to booking preplanned routine appointments for the management of their condition. Online appointment booking allowed them to do this at a time that was convenient and in a way that was convenient, be this via smartphone or computer:

> *It was for a regular, you know, prescription update thing, so it wasn't like I needed it on the day, it was something I could plan.* [Female participant; aged 35-44 years; practice 1; White British]

This reflected the nonurgent nature of the appointments they were booking, which fit with how online appointment booking was being offered at their practice, where appointments could be more readily booked some time ahead than on the day.

XSL•FO
RenderX

Discussion

Principal Findings

Awareness of online appointment booking is much higher than use, although still under half of patients are aware of online appointment booking. Patients, including those who did not use it, were accepting of its value as a convenient option for accessing an appointment.

There was little variability in awareness and use among those aged <65 years but a substantial decline in those aged >75 years. Age and life stage were key factors influencing experiences and perceptions. Working patients and those with caring responsibilities found it particularly convenient relative to the telephone, but older patients in the interview sample preferred to use the telephone as it was familiar and did not require the help of others.

There was a strong deprivation gradient, with deprived patients being less likely to use or be aware of online appointment booking. Findings related to ethnicity were more complex, with all minority groups less likely to be aware of online appointment booking than White patients, whereas patients in the "Black" or "other" ethnic groups were less likely to use online appointment booking than White patients and those in the "mixed" or "Asian" ethnic groups were more likely to have used them. Location was influential, with patients in rural areas less likely to be aware of or use online appointment booking. While the interview study sampled patients from practices with a range of IMD scores (6/11, 55% of practices were below the average IMD), from different ethnic groups, and from a range of rural and urban areas, we did not observe any findings related directly to these factors.

Health status was a key factor. Those who were permanently sick or had a disability were less likely to be aware of or use online appointment booking. Patients with long-term conditions found that it was a good fit for them, being useful for booking routine nonurgent appointments for their condition. The only group where there was a disparity between awareness and use was in deaf patients who used sign language, who were less likely to be aware of but more likely to use online appointment booking. This was the only group in which there was such disparity between awareness and use (compared with the wider population), suggesting that, although deaf people may find online appointment booking particularly useful, they are unlikely to be aware that it is an option.

The strongest factor in determining awareness and use of online appointment booking across the study was the general practice the patient was registered with. The influence of general practice–mediated factors on patient experience was a key finding in the interview study. There was variation in how general practices organized their appointments and the registration process required for online appointment booking, both of which impacted experience.

Patient experience was a key focus of this study and impacted awareness and use of online appointment booking. Patients who found the practice website easy to use, patients with a preferred GP, and patients who had ordered repeat prescriptions online were more likely to be aware of and use online appointment booking. Patients reported that accessing repeat prescriptions was their gateway to use other online services, including online appointment booking. Patients reporting that it was easier to get through to the general practice on the phone had lower use and somewhat lower awareness, and patients reporting receptionists as helpful were more likely to be aware of but less likely to use online appointment booking. Regardless of whether they had tried it, patients who were aware of online appointment booking were more likely to have had a positive experience with their general practice when arranging and obtaining an appointment than those who were not aware of it.

Limitations

Our study was conducted on data from the 2018 to 2019 GPPS, and much has changed in health care delivery since. However, the 2022 GPPS showed that the number of patients booking appointments online had only risen to 21%, representing a change of just 5% over the period since 2018 to 2019. The 2022 GPPS also asked how patients chose to book an appointment, and just 16% of patients booked online.

The GPPS is a large national survey; it includes responses from patients registered with every general practice in England, and provides a high generalizability of findings. Like similar surveys, the GPPS has a relatively modest response rate (34% in 2018 and 33% in 2019). While low response rates may lead to substantial biases in absolute percentages, they are less likely to result in biased estimates of associations [29].

The interview study was disrupted by the COVID-19 pandemic. Recruitment was paused for 3 months, when the funder made the decision to halt all research activity not related to COVID-19, allowing general practices to focus on their COVID-19 response. Interviews were conducted via telephone, and therefore, we could interview participants already recruited before mid-March 2020 and adhere to social distancing restrictions. Recruitment was resumed in June 2020, and we revised our approach (as outlined in the Methods section) using a purposive approach to ensure that our sample was as diverse as possible.

Online appointment booking was unavailable to patients from the onset of the COVID-19 pandemic, and general practices used telephone and online triage to control the availability of face-to-face appointments. Online appointment booking resumed later in 2020 and has been available since. This may have influenced the responses to the interviews that were conducted after the onset of the COVID-19 pandemic, although we were not able to identify anything specific related to this.

We aimed for a maximally variable sample in the interview study. This allowed us to provide qualitative findings contextualizing the quantitative results according to age and presence of long-term conditions. However, the proportion of participants from ethnic minority groups was slightly lower than those in the population of England and Wales in the 2021 census (14% in our sample vs 18.3% in England and Wales) and was not reflective of the populations of the included practices in all cases. Underrecruitment to health research studies of people from ethnic minority groups is a long-standing concern

XSL•FO
RenderX

within the research community and can be attributed to researchers not facilitating access to research studies, with rapport with participants being a key factor in increasing recruitment [30,31]. This was a challenge during the COVID-19 pandemic. Further exploration of the views and experience of patients from ethnic minority groups is needed.

Our interview data did not provide insights into the impact of deprivation as identified in the retrospective analysis. This may be due to our sample, with a smaller number of interview participants having lower levels of education (a proxy for deprivation levels). The interview participants were not sampled for location (rural or urban) as we did not realize the significance of this distinction ahead of conducting the study. Future studies should consider these factors.

Our study was conducted in the United Kingdom, where there is a national health service and health care is free at the point of use. Findings may differ in countries where health systems differ, particularly those where primary care is accessed on a fee-for-service basis.

Comparison With Prior Work

Our study included online appointment booking via any online interface for a general practice appointment. This included the NHS app, where appointments can be booked. The NHS app had 18 million registrations during the COVID-19 pandemic due to its use for vaccination passports [32]. A study examined the use of the NHS app between January 2019 and May 2021. Those researchers examined total number of downloads, registrations, prescriptions ordered, records accessed, and appointments booked. Using time-series analysis, they calculated that, 12 months after the first UK COVID-19 lockdown, there were 21,606 fewer appointments booked online than might have been expected if the COVID-19 lockdown had not occurred [33], indicating that the COVID-19 pandemic had a detrimental effect on the use of online appointment booking.

A US observational study examining a patient portal [34] offering online appointment booking found that use of the portal was low and use was lowest in persons of lower socioeconomic status and without broadband internet. Another observational study in the United States examined the use of patient portal functions by patients, including appointment booking [35]. As in our study, the study demonstrated that use of the appointment booking function decreased with age, with the sharpest decrease in those aged >65 years. A cross-sectional survey of general practice patients in the West Midlands, England, looked at use of online appointment booking services before the pandemic [16]. The study observed similar associations to those in our study between awareness and use and deprivation level and between awareness and use and long-term conditions. Our study reinforces these findings as well as adding explanations for some of the patterns identified within the qualitative findings. Patterns related to age and socioeconomic status, as well as ethnicity, which we could say less about, have been identified as structural barriers to the equitable use of digital health by those examining the political economy of digital health equity [36]. This has implications for those designing and devising online services in health care.

Future Research

Experience with the general practice was a key factor in awareness and use of online appointment booking, and this warrants further research into exactly how practices can best enable patients to access and use online services and the different types of support that may be needed for different patient groups. General practices in England offer online services through portals that are provided by various third-party companies. Alongside these portals is the NHS app. Future research should explore how the different interfaces impact experience and use of online services and what that means for uptake. A survey exploring patient views on what supports uptake and use of web-based services in primary care found that there were varied factors impacting their use of these services, including poor design of online interfaces and lack of support from the general practice, along with personal factors such as preference for human interaction [37].

Older patients used online appointment booking less and, in some cases, found it difficult to use online services or chose not to when the telephone was available and they were allowed to book an appointment through it. Services may wish to focus on supporting those older people who do want to use online services while maintaining a telephone service for those who cannot or do not wish to use online routes. Research has demonstrated that understanding the characteristics of GPs, including their personalities, can influence the digital maturity levels in a given general practice, and so examining the GPs in a practice may contribute to a deeper understanding of how and why patients choose to use or not use online services [38].

Our qualitative interviews showed that using an online repeat prescription service was a gateway to using other online services. Researchers and those designing services may wish to explore why this service is relatively highly accessed when compared to other online services to determine what makes an online service usable. In the case of online appointment booking, accessing repeat prescriptions raised awareness of online appointment booking services, and we have demonstrated that most patients are not aware of this service.

Our quantitative findings identified that deaf patients who use sign language may find online appointment booking particularly useful but they are unlikely to be aware that it is an option. We did not recruit any deaf patients to the interview study, and their views are likely to be of importance in future research.

We identified a clear deprivation gradient, and this may exacerbate inequitable access for patients. The risk of excluding patients who cannot access online services should be a key consideration in the design of the wider-access systems within general practices and a key factor examined in future research. To ensure equitable access, it is important to understand how particular groups (patients who are older, more deprived, or from ethnic minority groups) who were shown to be less likely to be aware of and use online booking can be supported should they wish to use these services and not disadvantaged should they choose not to.

Patient awareness of online appointment booking may be influenced by the actions of the general practice and how they

XSL•FO
RenderX

promote these services. A systematic review examining how primary care supports patients to become aware of and access online services highlighted the lack of evidence in this area but concluded that existing evidence indicated that there was the potential for such support to increase awareness and use of services, with more research needed to understand what this support would look like [39].

Conclusions

Awareness and use of online appointment booking by patients in general practice is associated with the characteristics of the practice they attend, their age, the level of deprivation associated with their general practice, and whether they have a long-term condition. There are clear contextual explanations behind why online booking is more appealing for those with long-term conditions and those aged <75 years. Other factors such as why deprivation or ethnicity mediates use of online booking remain less understood. The findings have implications for patient engagement with the wider range of online services offered in general practice, how these are delivered, and for future research in this area.

Acknowledgments

The authors would like to thank Mayam Gomez-Cano, who conducted statistical analysis for this study. They would like to thank Graham Roberts, their patient and public coinvestigator, who provided advice and support throughout the study. This manuscript presents independent research funded by the National Institute for Health and Care Research under its Research for Patient Benefit program (grant reference PB-PG-1217-20033). The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Conflicts of Interest

In addition to his university role, JD is director of the Clinical Knowledge Unit at Advanced Health and Care Ltd. HA, AE, LP, JC, and GA declare no conflict of interest.

Multimedia Appendix 1
Detailed statistical methods.
[DOCX File , 22 KB - jmir_v26i1e51931_app1.docx]

Multimedia Appendix 2
Interview topic guide.
[DOCX File , 20 KB - jmir_v26i1e51931_app2.docx]

References

1. Online appointment technology. The Royal Australian College of General Practitioners. URL: https://www.racgp.org.au/running-a-practice/technology/business-technology/online-appointment-technology [accessed 2023-08-17]
2. What GP practices need to do. National Health Service England. URL: https://www.england.nhs.uk/gp-online-services/about-the-prog/gp-practices/ [accessed 2023-08-17]
3. e-Booking: a relatively simple concept that can make a big difference. Canada Health Infoway. 2014 Apr 10. URL: https://www.infoway-inforoute.ca/en/news-events-blog/blog/digital-health/e-booking-a-relatively-simple-concept-that-can-make-a-big-difference-part-1-of-2 [accessed 2024-06-25]
4. Standard general medical services contract – November 2022. National Health Service England. 2022. URL: https://www.england.nhs.uk/publication/standard-general-medical-services-contract-nov-22/ [accessed 2023-08-17]
5. Personal medical services agreement 2021/22. National Health Service England. 2022. URL: https://www.england.nhs.uk/publication/personal-medical-services-agreement-2021-22/ [accessed 2023-08-17]
6. Patient online management information (POMI). National Health Service England. 2023. URL: https://digital.nhs.uk/GP-data-hub/patient-online-management-information [accessed 2023-08-17]
7. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017 Apr 26;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
8. Paré G, Trudel MC, Forget P. Adoption, use, and impact of e-booking in private medical practices: mixed-methods evaluation of a two-year showcase project in Canada. JMIR Med Inform 2014 Sep 24;2(2):e24 [FREE Full text] [doi: 10.2196/medinform.3669] [Medline: 25600414]
9. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015 Mar;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]

XSL•FO
RenderX

10. de Lusignan S, Mold F, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic interpretative review. BMJ Open 2014 Sep 08;4(9):e006021 [FREE Full text] [doi: 10.1136/bmjopen-2014-006021] [Medline: 25200561]
11. Graham TA, Ali S, Avdagovska M, Ballermann M. Effects of a web-based patient portal on patient satisfaction and missed appointment rates: survey study. J Med Internet Res 2020 May 19;22(5):e17955 [FREE Full text] [doi: 10.2196/17955] [Medline: 32427109]
12. Mohammed MA, Montague J, Faisal M, Lamming L. The value of a Patient Access Portal in primary care: a cross-sectional survey of 62,486 registered users in the UK. Univ Access Inf Soc 2019 Oct 22;19(4):855-872. [doi: 10.1007/s10209-019-00693-8]
13. Denkovski V, Gavrilov G. E-appointments at primary care physicians during Covid-19 pandemic: viewpoint and satisfaction of medical consumers. EAI Endorsed Trans Perv Health Tech 2022 May 23;8(31):e2. [doi: 10.4108/eetpht.v8i31.1129]
14. Latulipe C, Quandt SA, Melius KA, Bertoni A, Miller DPJ, Smith D, et al. Insights into older adult patient concerns around the caregiver proxy portal use: qualitative interview study. J Med Internet Res 2018 Nov 02;20(11):e10524 [FREE Full text] [doi: 10.2196/10524] [Medline: 30389654]
15. Zhang M, Zhang C, Sun Q, Cai Q, Yang H, Zhang Y. Questionnaire survey about use of an online appointment booking system in one large tertiary public hospital outpatient service center in China. BMC Med Inform Decis Mak 2014 Jun 09;14:49 [FREE Full text] [doi: 10.1186/1472-6947-14-49] [Medline: 24912568]
16. Bryce C, O'Connell MD, Dale J, Underwood M, Atherton H. Online and telephone access to general practice: a cross-sectional patient survey. BJGP Open 2021 Aug;5(4):BJGPO.2020.0179 [FREE Full text] [doi: 10.3399/BJGPO.2020.0179] [Medline: 33910917]
17. Huh J, Koola J, Contreras A, Castillo AK, Ruiz M, Tedone KG, et al. Consumer health informatics adoption among underserved populations: thinking beyond the digital divide. Yearb Med Inform 2018 Aug;27(1):146-155 [FREE Full text] [doi: 10.1055/s-0038-1641217] [Medline: 30157518]
18. Reiners F, Sturm J, Bouw LJ, Wouters EJ. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019 Feb 21;16(4):645 [FREE Full text] [doi: 10.3390/ijerph16040645] [Medline: 30795623]
19. Peeters JM, Krijgsman JW, Brabers AE, Jong JD, Friele RD. Use and uptake of eHealth in general practice: a cross-sectional survey and focus group study among health care users and general practitioners. JMIR Med Inform 2016 Apr 06;4(2):e11 [FREE Full text] [doi: 10.2196/medinform.4515] [Medline: 27052805]
20. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
21. Solans O, Vidal-Alaball J, Roig Cabo P, Mora N, Coma E, Bonet Simó JM, et al. Characteristics of citizens and their use of teleconsultations in primary care in the Catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629 [FREE Full text] [doi: 10.2196/28629] [Medline: 33970867]
22. Due TD, Thorsen T, Andersen JH. Use of alternative consultation forms in Danish general practice in the initial phase of the COVID-19 pandemic - a qualitative study. BMC Fam Pract 2021 Jun 02;22(1):108 [FREE Full text] [doi: 10.1186/s12875-021-01468-y] [Medline: 34078281]
23. Murphy M, Scott LJ, Salisbury C, Turner A, Scott A, Denholm R, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021;71(704):e166-e177 [FREE Full text] [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
24. General practice patient survey 2022 results. National Health Service England. 2022 Jul 14. URL: https://www.gov.uk/government/statistics/gp-patient-survey-2022-results [accessed 2023-08-17]
25. GP patient survey 2017. National Health Service England. 2017 Jul 6. URL: https://www.gov.uk/government/statistics/gp-patient-survey-2017 [accessed 2023-08-17]
26. GP patient survey 2019. National Health Service England. 2019 Jul 11. URL: https://www.england.nhs.uk/statistics/2019/07/11/gp-patient-survey-2019/ [accessed 2023-08-17]
27. English indices of deprivation 2019. Ministry of Housing, Communities & Local Government, United Kingdom Government. 2019 Sep 26. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2023-08-17]
28. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
29. Saunders CL, Elliott MN, Lyratzopoulos G, Abel GA. Do differential response rates to patient surveys between organizations lead to unfair performance comparisons?: evidence from the English cancer patient experience survey. Med Care 2016 Jan;54(1):45-54 [FREE Full text] [doi: 10.1097/MLR.0000000000000457] [Medline: 26595223]
30. Nielsen AL, Jervelund SS, Villadsen SF, Vitus K, Ditlevsen K, TØrslev MK, et al. Recruitment of ethnic minorities for public health research: an interpretive synthesis of experiences from six interlinked Danish studies. Scand J Public Health 2017 Mar;45(2):140-152. [doi: 10.1177/1403494816686267] [Medline: 28081666]

XSL•FO
RenderX

31. Macneill V, Nwokoro C, Griffiths C, Grigg J, Seale C. Recruiting ethnic minority participants to a clinical trial: a qualitative study. BMJ Open 2013;3(4):e002750 [FREE Full text] [doi: 10.1136/bmjopen-2013-002750] [Medline: 23572193]
32. NHS App turns three with 22 million users. National Health Service England. 2021 Dec 31. URL: https://digital.nhs.uk/news/2021/nhs-app-turns-three-with-22-million-users [accessed 2023-08-17]
33. KC S, Reidy C, Laverty AA, Papoutsi C, Powell J, Tewolde S, et al. Adoption and use of the NHS App in England: a mixed-methods evaluation. Br J Gen Pract 2023 Jul 21;73(suppl 1):bjgp23X733737. [doi: 10.3399/bjgp23X733737]
34. Perzynski AT, Roach MJ, Shick S, Callahan B, Gunzler D, Cebul R, et al. Patient portals and broadband internet inequality. J Am Med Inform Assoc 2017 Sep 01;24(5):927-932 [FREE Full text] [doi: 10.1093/jamia/ocx020] [Medline: 28371853]
35. Zhong X, Park J, Liang M, Shi F, Budd PR, Sprague JL, et al. Characteristics of patients using different patient portal functions and the impact on primary care service utilization and appointment adherence: retrospective observational study. J Med Internet Res 2020 Feb 25;22(2):e14410 [FREE Full text] [doi: 10.2196/14410] [Medline: 32130124]
36. Shaw J, Glover W. The political economy of digital health equity: structural analysis. J Med Internet Res 2024 Mar 26;26:e46971 [FREE Full text] [doi: 10.2196/46971] [Medline: 38530341]
37. Khan N, Pitchforth E, Winder R, Abel G, Clark CE, Cockcroft E, et al. What helps patients access web-based services in primary care? Free-text analysis of patient responses to the Di-Facto questionnaire. BMC Prim Care 2024 Jan 10;25(1):20 [FREE Full text] [doi: 10.1186/s12875-023-02257-5] [Medline: 38200431]
38. Weik L, Fehring L, Mortsiefer A, Meister S. Big 5 personality traits and individual- and practice-related characteristics as influencing factors of digital maturity in general practices: quantitative web-based survey study. J Med Internet Res 2024 Jan 22;26:e52085 [FREE Full text] [doi: 10.2196/52085] [Medline: 38252468]
39. Leach B, Parkinson S, Gkousis E, Abel G, Atherton H, Campbell J, et al. Digital facilitation to support patient access to web-based primary care services: scoping literature review. J Med Internet Res 2022 Jul 14;24(7):e33911 [FREE Full text] [doi: 10.2196/33911] [Medline: 35834301]

Abbreviations

GP: general practitioner
GPPS: General Practice Patient Survey
IMD: Index of Multiple Deprivation
NHS: National Health Service
OR: odds ratio

Edited by T de Azevedo Cardoso, K Williams; submitted 17.08.23; peer-reviewed by H Veldandi, A Mahnke, M Barbosa; comments to author 24.04.24; revised version received 16.05.24; accepted 18.05.24; published 08.07.24.

Please cite as:
Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G
Investigating Patient Use and Experience of Online Appointment Booking in Primary Care: Mixed Methods Study
J Med Internet Res 2024;26:e51931
URL: https://www.jmir.org/2024/1/e51931
doi:10.2196/51931
PMID:

©Helen Atherton, Abi Eccles, Leon Poltawski, Jeremy Dale, John Campbell, Gary Abel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sociodemographic Factors Associated With Using eHealth for Information Seeking in the United States: Cross-Sectional Population-Based Study With 3 Time Points Using Health Information National Trends Survey Data

Christian Elias Vazquez[1], MSW, PhD; Rebecca L Mauldin[1], MSW, PhD; Denise N Mitchell[2], MA; Faheem Ohri[1], MSW

[1]School of Social Work, The University of Texas at Arlington, Arlington, TX, United States
[2]Department of Sociology, University of North Carolina, Chapel Hill, NC, United States

Corresponding Author:
Christian Elias Vazquez, MSW, PhD
School of Social Work
The University of Texas at Arlington
501 W Mitchell St
Arlington, TX, 76010
United States
Phone: 1 8052596963
Email: christian.vazquez@uta.edu

Abstract

Background: Despite the potential benefits of using eHealth, sociodemographic disparities exist in eHealth use, which threatens to further widen health equity gaps. The literature has consistently shown age and education to be associated with eHealth use, while the findings for racial and ethnic disparities are mixed. However, previous disparities may have narrowed as health care interactions shifted to web-based modalities for everyone because of the COVID-19 pandemic.

Objective: This study aims to provide an updated examination of sociodemographic disparities that contribute to the health equity gap related to using eHealth for information seeking using 3 time points.

Methods: Data for this study came from the nationally representative 2018 (n=3504), 2020 (n=3865), and 2022 (n=6252) time points of the Health Information National Trends Survey. Logistic regression was used to regress the use of eHealth for information seeking on race and ethnicity, sex, age, education, income, health status, and year of survey. Given the consistent association of age with the dependent variable, analyses were stratified by age cohort (millennials, Generation X, baby boomers, and silent generation) to compare individuals of similar age.

Results: For millennials, being female, attaining some college or a college degree, and reporting an annual income of US $50,000-$74,999 or >US $75,000 were associated with the use of eHealth for information seeking. For Generation X, being female, having attained some college or a college degree, reporting an annual income of US $50,000-$74,999 or >US $75,000, better self-reported health, and completing the survey in 2022 (vs 2018; odds ratio [OR] 1.80, 95% CI 1.11-2.91) were associated with the use of eHealth for information seeking. For baby boomers, being female, being older, attaining a high school degree, attaining some college or a college degree, reporting an annual income of US $50,000-$74,999 or >US $75,000, and completing the survey in 2020 (OR 1.56, 95% CI 1.15-2.12) and 2022 (OR 4.04, 95% CI 2.77-5.87) were associated with the use of eHealth for information seeking. Among the silent generation, being older, attaining some college or a college degree, reporting an annual income of US $50,000-$74,999 or >US $75,000, and completing the survey in 2022 (OR 5.76, 95% CI 3.05-10.89) were associated with the use of eHealth for information seeking.

Conclusions: Baby boomers may have made the most gains in using eHealth for information seeking over time. The race and ethnicity findings, or lack thereof, may indicate a reduction in racial and ethnic disparities. Disparities based on sex, education, and income remained consistent across all age groups. This aligns with health disparities literature focused on individuals with lower socioeconomic status, and more recently on men who are less likely to seek health care compared to women.

(J Med Internet Res 2024;26:e54745) doi:10.2196/54745

XSL•FO
RenderX

KEYWORDS

health information seeking; eHealth use; disparities; sex; age; education; mobile phone

Introduction

Background

When paired with appropriate eHealth literacy, using eHealth can help improve individuals' management of their health [1,2]. eHealth has been defined as the use of information and communication technologies for health and can include the use of email, SMS text messaging, websites, and mobile apps [3]. eHealth use has been associated with better physical and mental health, healthier behaviors, fewer health risk factors, knowledge of medical issues, and more informed health care decision-making [4-7]. Despite its potential benefits, some groups are less likely to adopt eHealth and as a result may experience related health or health care disparities.

It is well documented both within and outside of the United States across time that older adults use eHealth at lower rates than younger adults [8-10]. However, regarding eHealth use among racial and ethnic groups, the literature has produced mixed findings in recent years [9,11,12]. Even in the few studies that use nationally representative data from the United States, the association of race and ethnicity with eHealth use is inconsistent [10,12-15].

A study from 2013 using data from California found that using eHealth for information seeking was significantly lower for Hispanic, Black, and Asian adults than White adults [16]. In a 2014 nationally representative sample of 1336 US residents aged ≥54 years, White respondents were more likely to report using eHealth for information seeking than Black or Hispanic respondents [17]. However, for adults aged ≥75 years, there were no significant racial or ethnic differences in using eHealth for information seeking. For younger adults in this same sample, there were no ethnic or racial differences in using eHealth for information seeking (eg, SMS text messages, mobile apps, and emails), but significant differences emerged beginning at the age of 62 years [17]. These findings indicate that racial and ethnic differences in using eHealth for information seeking may be most appropriately examined within subgroups by age.

The literature also shows significant trends in other key sociodemographic characteristics that are critical to health equity. In addition to age, race, and ethnicity, a recent meta-analysis of studies examining correlates of using eHealth for information seeking identified income, education, sex, and health status as key characteristics to monitor [18]. Income and education were both positively associated with eHealth use, while better self-reported health and being female were also associated with eHealth use, and this is commonly reported both within and outside of the United States [18-23]. A previous study using the same data as this study, Health Information National Trends Survey (HINTS), from 2008 to 2017, found that using eHealth for information seeking was associated with an annual household income >US $20,000 [15]. The health status finding only appears to be significant in nonrepresentative samples, further suggesting that previous findings must be reproduced with representative samples [18]. To support efforts to identify and reduce health disparities across a variety of domains, regularly monitoring these factors in addition to age and race and ethnicity as potential correlates to eHealth use is important.

It is acknowledged that using eHealth for information seeking alone may not necessarily lead to being able to manage or improve health; a common issue is an inability to identify misinformation or understand the information received (ie, having insufficient eHealth literacy) [1,2]. However, systematic reviews of cross-sectional studies and independent intervention studies provide evidence that prior eHealth use is one of the strongest predictors of eHealth literacy mastery [18,24,25]. Using eHealth should precede eHealth literacy training or happen in tandem, which has been effective with a diverse set of participants [24,25]. eHealth literacy is associated with positive outcomes such as the identification of accurate information and better health management (ie, adherence) [26,27]. Thus, identifying those who are less likely to use eHealth—groups that historically may have been excluded from achieving health equity—and targeting them for intervention is critical to address health equity, given the potential benefits that effective eHealth use can have for individuals' health.

This Study

As eHealth continues to evolve at a rapid pace, identification of disparities must be kept up to date to every few years, helping interventionists and practitioners understand which groups may need help to maximize the benefits of eHealth. This study will (1) provide updated evidence of sociodemographic disparities across these domains related to the use of eHealth for information seeking in the United States using the most recent publicly available nationally representative relevant data at the time of analysis and (2) use 3 time points to provide evidence of possible growth in eHealth use over time, stratified by adult age cohorts (ie, millennials, Generation X, baby boomers, and silent generation).

In addition to the literature indicating that results may vary by age group, the literature on each generation's experiences with technology adds to the rationale for stratifying by age cohort. For example, baby boomers and the silent generation grew up without the internet, Generation X is considered the first technologically literate generation, and millennials were teenagers when the internet was launched [28,29]. Thus, these 4 generations had dramatically different experiences with computers, internet, and technology. eHealth interventionists who work with older adults document the need of tailoring interventions based on age, given the relationship between age and prior experience with technology [24,25].

The literature has documented an increase in eHealth use since 2008 [12,15,18]. Though statistically significant changes are not seen yearly or even biennially, they are typically seen over longer periods and have increased only steadily. Given the large shift in how people interacted with health care or accessed health care information due to the COVID-19 pandemic, it is expected that eHealth use has increased in recent years; however, it is

XSL•FO
RenderX

essential for public health science to document this with evidence. The last update with multiple time points of nationally representative data used data up to 2017. Thus, this is the rationale for using 2018 as the baseline for this study. The expected evidence showing increases in eHealth use could provide further justification for studies continuing to refine eHealth use engagement and literacy.

This study adds to the literature in several ways. First, it uses the most recent (at the time of analysis) publicly available nationally representative data that asks about eHealth use behaviors. Second, while some studies have used multiple data time points, to our knowledge, no studies to date have examined sociodemographic differences within multiple adult age cohorts with multiple data time points. Third, the pooling of data from 2018, 2020, and 2022 provides evidence across multiple time points to monitor health disparities in using eHealth for information seeking. Collectively, these contributions may help in the identification of groups to target for interventions that assist individuals to maximize the benefits of eHealth.

Research Questions

This study is driven by two overarching research questions. Research question 1: Aggregating 3 time points, what sociodemographic factors are associated with using eHealth for information seeking in the United States, stratified by age cohort? Research question 2: Do the data indicate that individuals within each age cohort are, on average, more likely to use eHealth for information seeking in more recent time points (eg, 2018 vs 2020 vs 2022), controlling for key sociodemographic factors?

Methods

Sample and Data Collection

Data for this study came from 3 time points (2018, 2020, and 2022) of the HINTS. HINTS is an assessment of the American public's access to and use of information about cancer across the cancer care continuum, spanning from prevention and early detection to diagnosis, treatment, and survivorship. In addition to this standard HINTS content, each cycle also includes questions relating to general eHealth information-seeking behaviors not specific to cancer information. HINTS is a nationally representative survey of the noninstitutionalized adult population (aged >18 years) in the United States, conducted every few years since 2003, and from 2020, it is consistently administered biennially [30-32]. Respondents were sampled using a complex sampling design to achieve a nationally representative sample. Data collection for 2018 started on January 26, 2018, and concluded on May 2, 2018 (n=3504), and had a response rate of 32.9%. Data collection for 2020 started on February 24, 2020, and concluded on June 15, 2020 (n=3865), and had a response rate of 37%. Data collection for 2022 started on March 7, 2022, and concluded on November 8, 2022 (n=6252), and had a response rate of 28.1%. Data were collected using a self-administered survey sent by mail for 2018 and 2020. For 2022, two different data collection modes were available: self-administered web-based surveys and self-administered surveys sent by mail. Surveys were available in English and Spanish. More information about the design, procedures, measures, and weighting of HINTS cycles can be found elsewhere [30-32]. Final analytic samples by time point were n=2612 for 2018, n=2962 for 2020, and n=4971 for 2022.

Ethical Considerations

Information about informed consent and participant compensation from original HINTS data collection periods are published elsewhere [30-32]. The original data collection was approved by the Westat institutional review board. Ethics review and approval for this analysis were waived due to the study being a secondary analysis of deidentified publicly available data [33]. The authors registered with the HINTS website to use the publicly available data.

Variables

Dependent Variable

The dependent variable conceptually measured whether the respondents used eHealth for information seeking in the past 12 months. This was operationalized by creating a dichotomous variable based on the respondent's answer to 3 yes or no questions that (1) captured various types of eHealth information seeking and (2) were consistent across the 3 time points. The final variable was coded 1 if they answered yes to any of the 3 questions and 0 if they answered no to all questions. There is a caveat with the abovementioned reason number 2 because the question changed slightly in 2022; however, the researchers of this study suggest the changes capture the same concept, and this is also stated as a limitation. The questions for 2018 and 2020 asked: "In the past 12 months, have you used a computer, smartphone, or other electronic means to do any of the following?" (1) "Looked for health or medical information for yourself?" (2) "Used email or the internet to communicate with a doctor or doctor's office?" and (3) "Looked up medical test results." For 2022, the question asked: "In the past 12 months, have you used the internet to take care of any of the following health-related needs?" (1) "Look for health or medical information"; (2) "Send a message to a health care provider or a health care provider's office"; and (3) "View medical test results." The response options were no (2) and yes (1).

Independent Variables

Sex

The sex question asked respondents to report their gender at birth as either male or female. Although the item uses the term "gender," we believed it aligned more closely with our conceptualization of sex, based on the wording and response options.

Age Cohort

Given the consistent and well-documented association between older age and reduced use of eHealth for information seeking, this study stratified analyses by age cohort to assess differences among respondents of similar age. The age cohorts of interest were millennials, Generation X, baby boomers, and silent generation. Respondents were grouped based on cohort ages defined by the Pew Research Center [34]. The age range for millennials in 2018 was 22 to 37 years, in 2020 was 24 to 39 years, and in 2022 was 26 to 41 years. The age range for Generation X in 2018 was 38 to 53 years, in 2020 was 40 to 55

XSL•FO
RenderX

years, and in 2022 was 42 to 57 years. The age range for baby boomers in 2018 was 54 to 72 years, in 2020 was 56 to 74 years, and in 2022 was 58 to 76 years. The age range for the silent generation in 2018 was 73 to 89 years, in 2020 was 75 to 91 years, and in 2022 was 77 to 93 years. Age was also included as a continuous variable within each stratified analysis. Gen Z was also present in the data; however, they were excluded from the sample given that a portion of this group was of college age and may not have had a chance to complete college or enter the workforce full time, conflicting with 2 other key sociodemographic factors (ie, education and income).

Race and Ethnicity

The analysis included all racial and ethnic groups in the sample, including those who identified as Asian, Black, Hispanic, other, and White. White was the reference group. The Asian group included those identifying as Asian Indians, Chinese, Filipino, Japanese, Korean, Vietnamese, and other Asian. The Hispanic group included those identifying as Mexican, Puerto Rican, Cuban, other Hispanic, and multiple Hispanic ethnicities. The "other" group included those identifying as Native Hawaiian, Guamanian or Chamorro, Samoan, or other Pacific Islander.

Education

Education was originally coded as 4 categories (less than high school, high school graduate, some college, and college graduate). Three dummy coded variables were created with less than high school as the reference.

Annual Income

The data included an income variable with 5 categories, including <US $20,000, US $20,000-$34,999, US $35,000-$49,999, US $50,000-$74,999, and ≥US $75,000. Dummy codes were created with <US $20,000 as the reference. This variable was kept in this form because it was used in the same way in a previous study using HINTS data, with eHealth information-seeking behaviors as the outcome [15].

Health Status

The general health question asked, "In general, would you say your health is..." and was self-reported as "excellent," "very good," "good," "fair," and "poor." This was used as a continuous variable and recoded with poor=1 and excellent=5.

Year

The year of survey was included to determine differences between time points. This was coded as 1=2018, 2=2020, and 3=2022.

Statistical Analyses

The analyses were performed using SAS (version 9.3; SAS Institute) [35]. Complete case analysis was used. Thus, only cases without missing data for all included variables were used. This study used PROC SURVEYLOGISTIC with replicate weights and the jackknife approach as suggested by the HINTS documentation to provide population-level results adjusted for the sampling methods [30-32]. Because we combined 3 time points to test for trends across time points, we created a new set of 150 replicate weights (50 for each time point) using the Rizzo method [36]. Sample descriptive characteristics were examined using frequencies, percentages, means, and SDs. This was done with PROC SURVEYFREQ and SURVEYMEANS with the final sample weight. Chi-square and ANOVA tests were also conducted to assess potential differences of variables across time points.

This study used binary logistic regressions stratified by 4 age cohorts to examine sociodemographic associations with eHealth use for information seeking and to assess differences in use across the years 2018, 2020, and 2022. All variables of interests were added to the model. The reference categories were White race, male sex, less than high school graduate education level, annual income <US $20,000, and the year 2018. The 4-category age cohort variable was included in the DOMAIN option in SAS for stratification. The results for the full sample were also reported for a snapshot of the pooled time points 2018, 2020, and 2022. Odds ratios (ORs) and CIs were used to assess significance.

Results

Descriptive Statistics

Table 1 shows the descriptive characteristics of the study sample as well as *P* value for tests of differences across time points. Slightly more female respondents competed the survey in 2022 (2484/4188, 51.45%) compared to 2020 (1651/2927, 50.58%) and 2018 (1478/2576, 49.18%; *P*=.04). The tests of differences resulted in a *P*=.02; however, the order of largest to smallest groups remained the same: White (largest), Hispanic, Black, Asian, and other (smallest). For education, there was a statistically significant difference across time points (*P*<.001), and the order of largest to smallest groups also remained the same: some college, college graduate, high school graduate, and less than high school. The pattern of similar ordering was also evident with the income variable: ≥US $75,000, US $50,000-$74,999, <US $20,000, US $35,000-$49,999, and US $20,000-$34,999, where a significant difference was also observed across time points (*P*<.001). There was not a statistically significant difference in self-reported health status across time points (*P*=.06), with the average rating being "good" health. There was a significant difference in the size of cohort groups across time point (*P*≤.001), but the order of the size of the groups remained the same: Generation X, baby boomers, millennials, and silent generation. There was a significantly (*P*≤.001) smaller proportion of respondents who reported using eHealth in the past 12 months in 2018 and 2020 (1993/2576, 78.89% and 2386/2927, 82.38%) compared to 2022 (3877/4188, 91.31%). When examining differences in eHealth use among age cohorts within each time point, the proportion of respondents who reported using eHealth decreased with increasing age across all time points (*P*≤.001 for all 3 time points, not shown in Table 1).

XSL•FO
RenderX

Table 1. Descriptive characteristics of the study sample from the Health Information National Trends Survey for 2018, 2020, and 2022.

Characteristics	Full sample (N=9691)	2018 (n=2576)	2020 (n=2927)	2022 (n=4188)	*P* value (chi-square or ANOVA)
Sex, n (%)					.04
Male	4078 (49.62)[a]	1098 (50.82)	1276 (49.41)	1704 (48.55)	
Female	5613 (50.38)	1478 (49.18)	1651 (50.58)	2484 (51.45)	
Age (years), n (%); mean (SD)					<.001
Millennials[b]	2035 (28.36); 32.7 (4.60)	443 (25.88); 30.6 (4.31)	559 (28.16); 32.1 (4.37)	1033 (31.25); 34.9 (4.42)	
Generation X[c]	2595 (37.08); 48.4 (4.97)	681 (38.64); 46.1 (4.70)	760 (36.11); 48.0 (4.75)	1154 (36.5); 50.0 (4.69)	
Baby boomers[d]	4079 (27.97); 65.0 (5.27)	1080 (27.52); 63.3 (5.01)	1296 (28.85); 64.8 (5.15)	1703 (27.46); 66.3 (5.19)	
Silent generation[e]	982 (6.59); 80.0 (4.46)	372 (7.96); 78.8 (4.64)	312 (6.88); 70.1 (4.31)	298 (4.79); 81.4 (3.93)	
Race and ethnicity, n (%)					.02
Asian	474 (4.99)	108 (4.52)	140 (5.44)	226 (4.99)	
Black	1363 (10.86)	346 (10.68)	389 (11.37)	628 (10.49)	
Hispanic	1487 (14.82)	365 (15.9)	479 (14.68)	643 (13.82)	
White	6019 (65.77)	1652 (65.85)	1820 (65.85)	2547 (65.58)	
Other	348 (3.55)	105 (3.04)	99 (2.65)	144 (5.11)	
Education, n (%)					<.001
Less than high school graduate	499 (6.42)	165 (8.32)	181 (6.50)	153 (4.29)	
High school graduate	1484 (19.69)	419 (20.1)	486 (20.73)	579 (18.07)	
Some college	2805 (38.53)	760 (38.26)	850 (38.60)	1195 (38.75)	
College graduate	4903 (35.36)	1232 (33.32)	1410 (34.17)	2261 (38.89)	
Income (US $), n (%)					<.001
<20,000	1360 (12.54)	426 (14.82)	438 (12.66)	496 (9.95)	
20,000-34,999	1205 (10.81)	351 (11.76)	366 (10.95)	488 (9.63)	
35,000-49,999	1278 (12.26)	341 (12.82)	397 (12.78)	540 (11.07)	
50,000-74,999	1746 (18.28)	486 (18.21)	510 (18.29)	750 (18.34)	
≥75,000	4102 (46.12)	972 (42.38)	1216 (45.32)	1914 (51.01)	
Health status, mean (SD)	3.42 (0.94)	3.45 (0.95)	3.44 (0.94)	3.40 (0.93)	.06
eHealth use in the past year, n (%)[f]					<.001
No	1435 (15.99)	583 (21.11)	541 (17.62)	311 (8.69)	
Yes	8256 (84.01)	1993 (78.89)	2386 (82.38)	3877 (91.31)	

[a]Values were estimated using final sample weights to account for complex sampling.

[b]The age range for millennials in 2018 was 22 to 37 years, in 2020 was 24 to 39 years, and in 2022 was 26 to 41 years.

[c]The age range for Generation X in 2018 was 38 to 53 years, in 2020 was 40 to 55 years, and in 2022 was 42 to 57 years.

[d]The age range for baby boomers in 2018 was 54 to 72 years, in 2020 was 56 to 74 years, and in 2022 was 58 to 76 years.

[e]The age range for the silent generation in 2018 was 73 to 89 years, in 2020 was 75 to 91 years, and in 2022 was 77 to 93 years.

[f]eHealth use in the past year was coded as yes if the respondent answered yes to one of the three questions asking about using internet and devices to (1) look up health information, (2) communicate with health care provider, and (3) look up medical test results.

Pooled Sample

The pooled sample is intended to provide a snapshot of the average of the 3 time points. Among the pooled sample, women had higher odds of using eHealth for information seeking compared to men (OR 1.70, 95% CI 1.41-2.05). Older age was associated with lower odds of using eHealth for information seeking (OR 0.97, 95% CI 0.97-0.98). Compared to respondents

XSL•FO
RenderX

with less than a high school education, respondents with some college (OR 2.64, 95% CI 1.81-3.85) or a college degree (OR 5.26, 95% CI 3.58-7.72) had higher odds of using eHealth for information seeking. Compared to an annual household income of <US $20,000, reporting an annual income of US $35,000-$49,999 (OR 1.55, 95% CI 1.09-2.21), US $50,000-$74,999 (OR 2.62, 95% CI 1.94-3.54), or ≥US $75,000 (OR 3.60, 95% CI 2.49-5.19) was associated with higher odds of using eHealth for information seeking. Better self-reported health status was associated with lower odds of using eHealth for information seeking (OR 0.82, 95% CI 0.74-0.92). Compared to respondents who completed the survey in 2018, respondents who completed the survey in 2020 (OR 1.26, 95% CI 1.03-1.55) or 2022 (OR 2.59, 95% CI 1.99-3.36) had higher odds of using eHealth for information seeking.

Millennials

Table 2 provides the logistic regression results for using eHealth for information seeking for the pooled sample and by each age cohort. Among millennials, women had higher odds of using eHealth for information seeking compared to men (OR 2.35, 95% CI 1.32-4.16). Compared to respondents with less than a high school education, respondents with some college (OR 4.63, 95% CI 1.67-12.84) or a college degree (OR 5.39, 95% CI 1.73-16.78) had higher odds of using eHealth for information seeking. Compared to an annual household income of <US $20,000, reporting an annual income of US $50,000-$74,999 (OR 3.75, 95% CI 1.47-9.56) or ≥US $75,000 (OR 5.55, 95% CI 1.79-17.22) was associated with higher odds of using eHealth for information seeking.

Table 2. Logistic regression for using eHealth for information seeking, stratified by age cohort, Health Information National Trends Survey for 2018 (n=2576), 2020 (n=2927), and 2022 (n=4188).

	Pooled sample, OR[a] (95% CI)	Millennials[b], OR (95% CI)	Generation X[c], OR (95% CI)	Baby boomers[d], OR (95% CI)	Silent generation[e], OR (95% CI)
Female[f]	*1.70 (1.41-2.05)*[g]	*2.35 (1.32-4.16)*	*1.79 (1.17-2.73)*	*1.68 (1.32-2.13)*	1.38 (0.81-2.33)
Age	*0.97 (0.97-0.98)*	1.03 (0.94-1.12)	0.97 (0.93-1.01)	*0.96 (0.93-0.99)*	*0.93 (0.89-0.99)*
Asian[h]	0.85 (0.51-1.40)	0.90 (0.26-3.06)	0.87 (0.39-1.93)	0.80 (0.37-1.74)	1.02 (0.23-4.53)
Black[h]	0.97 (0.69-1.38)	1.00 (0.31-3.28)	1.01 (0.58-1.77)	0.83 (0.57-1.20)	1.17 (0.61-2.23)
Hispanic[h]	0.95 (0.70-1.30)	0.98 (0.45-2.14)	0.89 (0.52-1.53)	0.80 (0.54-1.17)	1.69 (0.76-3.78)
Other[h]	1.37 (0.78-2.39)	0.98 (0.25-3.89)	1.68 (0.54-5.18)	1.22 (0.56-2.65)	2.21 (0.35-14.11)
High school graduate[i]	1.29 (0.86-1.92)	1.14 (0.39-3.34)	1.11 (0.58-2.12)	*2.23 (1.25-3.99)*	1.12 (0.42-2.99)
Some college[i]	*2.64 (1.81-3.85)*	*4.63 (1.67-12.84)*	*2.00 (1.07-3.72)*	*3.23 (1.79-5.84)*	*2.77 (1.11-6.88)*
College graduate[i]	*5.26 (3.58-7.72)*	*5.39 (1.73-16.78)*	*5.32 (2.60-10.89)*	*7.70 (3.96-14.98)*	*3.76 (1.28-11.03)*
US $20,000-$34,999[j]	1.38 (1.00-1.90)	2.49 (0.98-6.31)	1.24 (0.64-2.40)	1.26 (0.81-1.96)	1.35 (0.68-2.70)
US $35,000-$49,999[j]	*1.55 (1.09-2.21)*	2.19 (0.69-7.00)	1.44 (0.79-2.64)	1.35 (0.83-2.21)	1.62 (0.76-3.47)
US $50,000-$74,999[j]	*2.62 (1.94-3.54)*	*3.75 (1.47-9.56)*	*4.63 (2.45-8.76)*	*1.54 (1.01-2.34)*	*2.63 (1.25-5.51)*
≥US $75,000[j]	*3.60 (2.49-5.19)*	*5.55 (1.79-17.22)*	*4.09 (2.24-7.46)*	*2.71 (1.69-4.34)*	*2.92 (1.30-6.58)*
Health status	*0.82 (0.74-0.92)*	0.74 (0.53-1.02)	*0.67 (0.54-0.84)*	0.94 (0.81-1.09)	1.08 (0.83-1.40)
2020[k]	*1.26 (1.03-1.55)*	0.65 (0.32-1.31)	1.45 (0.93-2.27)	*1.56 (1.15-2.12)*	1.37 (0.87-2.14)
2022[k]	*2.59 (1.99-3.36)*	1.42 (0.68-2.94)	*1.80 (1.11-2.91)*	*4.04 (2.77-5.87)*	*5.76 (3.05-10.89)*

[a]OR: odds ratio.

[b]The age range for millennials in 2018 was 22 to 37 years, in 2020 was 24 to 39 years, and in 2022 was 26 to 41 years.

[c]The age range for Generation X in 2018 was 38 to 53 years, in 2020 was 40 to 55 years, and in 2022 was 42 to 57 years.

[d]The age range for baby boomers in 2018 was 54 to 72 years, in 2020 was 56 to 74 years, and in 2022 was 58 to 76 years.

[e]The age range for the silent generation in 2018 was 73 to 89 years, in 2020 was 75 to 91 years, and in 2022 was 77 to 93 years.

[f]Reference category: male (sex).

[g]Statistically significant values are italicized.

[h]Reference category: White (race and ethnicity).

[i]Reference category: less than high school graduate (education).

[j]Reference category: <US $20,000 (income).

[k]Reference category: 2018 (year).

XSL•FO
RenderX

Generation X

Among Generation X, women had higher odds of using eHealth for information seeking compared to men (OR 1.79, 95% CI 1.17-2.73). Compared to respondents with less than a high school education, respondents with some college (OR 2.00, 95% CI 1.07-3.72) or a college degree (OR 5.32, 95% CI 2.60-10.89) had higher odds of using eHealth for information seeking. Compared to an annual household income of <US $20,000, reporting an annual income of US $50,000-$74,999 (OR 4.63, 95% CI 2.45-8.76) or ≥US $75,000 (OR 4.09, 95% CI 2.24-7.46) was associated with higher odds of using eHealth for information seeking. Better self-reported health status was associated with lower odds of using eHealth for information seeking (OR 0.67, 95% CI 0.54-0.84). Compared to respondents who completed the survey in 2018, respondents who completed the survey in 2022 (OR 1.80, 95% CI 1.11-2.91) had higher odds of using eHealth for information seeking.

Baby Boomers

Among baby boomers, women had higher odds of using eHealth for information seeking compared to men (OR 1.68, 95% CI 1.32-2.13). Older age was associated with lower odds of using eHealth for information seeking (OR 0.96, 95% CI 0.93-0.99). Compared to respondents with less than a high school education, respondents with a high school education (OR 2.23, 95% CI 1.25-3.99), some college (OR 3.23, 95% CI 1.79-5.84), or a college degree (OR 7.70, 95% CI 3.96-14.98) had higher odds of using eHealth for information seeking. Compared to an annual household income of <US $20,000, reporting an annual income of US $50,000-$74,999 (OR 1.54, 95% CI 1.01-2.34) or ≥US $75,000 (OR 2.71, 95% CI 1.69-4.34) was associated with higher odds of using eHealth for information seeking. Compared to respondents who completed the survey in 2018, respondents who completed the survey in 2020 (OR 1.56, 95% CI 1.15-2.12) or 2022 (OR 4.04, 95% CI 2.77-5.87) had higher odds of using eHealth for information seeking.

Silent Generation

Among the silent generation, older age was associated with lower odds of using eHealth for information seeking (OR 0.93, 95% CI 0.89-0.99). Compared to respondents with less than a high school education, respondents with some college (OR 2.77, 95% CI 1.11-6.88) or a college degree (OR 3.76, 95% CI 1.28-11.03) had higher odds of using eHealth for information seeking. Compared to an annual household income of <US $20,000, reporting an annual income of US $50,000-$74,999 (OR 2.63, 95% CI 1.25-5.51) or ≥US $75,000 (OR 2.92, 95% CI 1.30-6.58) was associated with higher odds of using eHealth for information seeking. Compared to respondents who completed the survey in 2018, respondents who completed the survey in 2022 (OR 5.76, 95% CI 3.05-10.89) had higher odds of using eHealth for information seeking.

Discussion

Principal Findings

This study set out to (1) provide updated evidence on sociodemographic disparities across these domains related to the use of eHealth for information seeking in the United States using the most recent publicly available nationally representative relevant data at the time of analysis and (2) use 3 time points to provide evidence of possible growth in eHealth use over time, stratified by adult age cohorts, at the population level. eHealth and its use are continuously evolving given the growing ubiquity of technology in health care and should be monitored every few years to keep researchers, practitioners, and policy makers up to date. At the time of analysis, the HINTS 2022 was the most up-to-date publicly available nationally representative data set available for reporting eHealth information-seeking behaviors in the United States. These findings are important to help interventionists understand what populations to target and how this may have changed over time. These findings can be used as a reference for future updates.

Given the consistent association between age and eHealth use and each generation's historical experience with technology, analyses were stratified by age cohort to assess differences among individuals from the same age group. Though being younger has historically been associated with higher eHealth use, the findings of this study highlight sex, education, and income disparities even within the youngest age groups. For Generation X, in addition to the disparities seen in the millennial cohort, those with better self-reported health were also less likely to use eHealth for information seeking. For baby boomers, those who were male, older, had less than a high school education, and reported income <US $50,000 were less likely to use eHealth for information seeking. For the silent generation, those who were older, had less than some college education, and reported income <US $50,000 were less likely to use eHealth for information seeking. The findings from the pooled analysis align with the findings from previous studies with pooled adult population analysis; however, the stratified findings from this study highlight the need to examine the findings by age group [9,11,12,17,18]. For example, no significant differences appear to exist by race and ethnicity for any group, which have been seen in previous work. This study also shows how the educational disparities related to eHealth use are mostly consistent with a slight variation for baby boomers. In addition, the results indicate that baby boomers and silent generation may have made the largest gains in using eHealth for information seeking over time, with Generation X also making gains, which is something that was not observed within the millennial cohort. This is a promising finding for older adults in the United States because it aligns with the literature from other countries, which suggests that older adults are ready to engage with eHealth and eHealth interventions [37].

The education and income findings show strong consistency. Similarly, the sex finding is consistent, with the exception of the silent generation. The lack of statistically significant racial and ethnic differences across all cohorts and time points is consistent in this study that uses nationally representative data; however, there continues to be mixed findings in this area of the literature overall, with some previous studies suggesting that this only exists with older adults and others reporting that significant findings for racial and ethnic differences tend to appear in analyses using nonrepresentative samples [17,18]. Other pooled studies using older HINTS data mostly found Black adults to use eHealth less than White adults [12,15] or

XSL•FO
RenderX

that health information seeking was higher for White adults than Hispanic adults [38]. Those studies did not stratify by age and use data before 2018; however, they may help put the findings of this study in the context of time. Thus, the findings of this study indicate that Hispanic and Black adults may have made gains in using eHealth for information seeking compared to White adults. It will be interesting to follow this divide as well as the age disparity that exists in the oldest age cohorts, given that millennials and Generation X both currently use eHealth at higher rates and will have had more time to adjust to eHealth use compared to baby boomers and the silent generation. The health status finding is also interesting because eHealth use has been associated with better health, but as the findings of this study suggest, those with better self-reported health also tend to be less likely to use eHealth for information seeking; though this only appears for the pooled and Generation X samples.

Overall, the millennial cohort had the least number of disparities with the silent generation following them, though these 2 groups were on opposite ends in regard to use across all time points. The largest number of disparities were observed in the baby boomer cohort. The education finding has been one of the most consistent correlates across the literature [9,11,12,17,18]. The result for less than high school education compared to high school graduate for baby boomers was a result that was unique to this group, even when looking at the pooled analysis. Addressing educational disparities is particularly important as eHealth may be most useful to those groups considered potentially vulnerable such as older adults, with respect to technology. The sex findings are concerning as men continue to have lower life expectancy than women, and the literature suggests that men are less likely to seek out health care in general compared to women [39,40]. In the younger age groups, it may also be the case that men have fewer reproductive health information needs than women.

Implications

We acknowledge earlier literature suggesting that, similar to general internet use, using eHealth for information seeking can have both positives and negatives, with misinformation and scams or fraud being some of the main negatives [1,2]. While this study did not assess misinformation or being scammed or defrauded or eHealth literacy, it is important to note that using eHealth with a sufficient level of eHealth literacy can help address some of these issues. Furthermore, the groups considered most vulnerable may be the most susceptible to misinformation and scams or fraud. This is relevant to this study because researchers who may want to focus on increasing eHealth use among those who are less likely to use it, based on the findings of this study (eg, those with lower education and low income), should take caution to pair any intervention that promotes uptake with appropriate eHealth literacy to help prevent negative outcomes of using eHealth. eHealth literacy interventions have proven to be effective at increasing eHealth literacy up to 61% for a diverse range of individuals [24]. In addition, relevant to this study, interventionists have seen differential dropout rates based on race, ethnicity, sex, and prior computer experience, which led other researchers to develop machine learning–based approaches that can provide useful information for predicting retention [24,41]. Studies with up-to-date population-based data can contribute to this and other types of artificial intelligence approaches to provide initial clues as to whom and how to target with appropriate strategies.

On the positive end, once individuals know how to use eHealth effectively, they can take advantage of technologies that allow health care providers to assist in health management. Adults may upload food logs, blood sugar levels, and drugs taken, which the health care providers can check in real time [41]. Other tools for older adults include ones that can detect changes in daily activities, such as falls, and devices that send notifications for one to exercise or take prescriptions [42]. However, these eHealth tools are only useful if one is using eHealth and is eHealth literate. Thus, using eHealth for information seeking is an ideal place to start this journey. Using eHealth effectively can also help reduce stress, such as navigating a web-based health portal to save time or being able to check a physician's message on their own time instead of waiting for a phone call. As telehealth becomes more common, when appropriate, individuals can take appointments at home, avoiding travel and schedule conflicts.

Our results may inform health educators when selecting intervention target populations and sites. For example, the consistent finding that those with a high school education or less are less likely to use eHealth for information seeking suggests that intervention at the middle school or high school level could be beneficial. Health education is common in school systems, and adding eHealth information seeking and literacy to the curriculum is relevant and could help reduce the educational disparities in eHealth information seeking. Similarly, sites with older adult classes, groups, and activities such as older adult centers, assisted living communities, and retirement communities could also be valuable locations for eHealth and literacy education.

However, it is important to note that access to internet services and devices is also imperative for eHealth use. Public policies that increase access to broadband or freely available Wi-Fi hot spots could be seen as having public health implications by reducing barriers to eHealth information seeking. Indeed, as a public health intervention, landing pages for free public internet services could be set to default to reputable health information websites as a way to promote eHealth information-seeking behaviors. Similarly, as part of their business practices, technology companies in the private sector could engage in social well-being efforts by expanding access to low-cost internet services and technological devices and pairing these efforts with campaigns that promote eHealth information seeking and literacy. The importance of these efforts was highlighted during the COVID-19 pandemic, as in-person information channels were restricted and those most underserved and underresourced, such as those with characteristics identified in this study, may have benefitted greatly.

Limitations

A strength of this study is that it uses a nationally representative data set. However, a limitation to using this secondary data was that the measure for eHealth information seeking was a dichotomous composite of 3 items that varied slightly across

XSL•FO
RenderX

time points and could have been interpreted differently, particularly in 2022 compared to 2018 and 2020. Furthermore, this variable only partially captures the complex phenomenon of easily accessing health information on the internet. Measures of general literacy, health care literacy, and English proficiency, which may be related to the information seeking experience, were not available in the HINTS data sets that were analyzed. Another potential limitation is bias in survey completion, as those who are most underrepresented and underserved may face barriers that prevent their participation in this type of survey. For example, those who do not speak English or Spanish, those who may be displaced (ie, no physical address) and do not have internet or phone access, and those who may have literacy issues not related to speaking a different language are people who represent some of the most underserved and underrepresented populations in health research. An inevitable limitation of any technology study is that technology evolves rapidly, rendering some components outdated. Continuous updating of intervention components and outcome measures will be necessary for future studies. The HINTS longitudinal data set will allow for future comparisons. It is important to acknowledge that the data collection period for the 2020 data overlapped with the COVID-19 social distancing measures when people were at home and potentially more likely to be using eHealth; however, the proportions of respondents who reported using eHealth in the past 12 months in 2018 and 2020 (1993/2576, 78.89% and 2386/2927, 82.38%) were closer in size compared to 2022 (3877/4188, 91.31%). It is also possible that respondents could have still been practicing certain social distancing measures during part of the 2022 reporting period (ie, in the past 12 months). The findings from this study align with findings from samples in other countries during this period [26,43,44]. It is also important to note that HINTS is a trend study and not a panel study. As such, the observed trends over time do not represent within-respondent changes but rather larger trends in the population as a whole.

Conclusions

eHealth information seeking is an important step in maximizing eHealth use in general, which promotes individual health and well-being. As technology rapidly evolves, disparities in eHealth information behaviors also shift. Recognizing changing disparities over time is a necessary step for closing gaps in eHealth information seeking for members of groups considered vulnerable or underserved groups and may be an important component in public health efforts to reduce overall health disparities. Results from this study support a growing body of literature that racial and ethnic gaps in eHealth information seeking may be closing. This promising trend provides hope for future efforts to reduce other types of disparities in eHealth information seeking. With regular monitoring and interventions tailored to populations in need, the public health community may also realize success in reducing disparities for adults with lower education levels.

Data Availability

The data sets generated during and analyzed during this study are available in the Health Information National Trends Survey repository [45].

Conflicts of Interest

None declared.

References

1. Choi M. Association of eHealth use, literacy, informational social support, and health-promoting behaviors: mediation of health self-efficacy. Int J Environ Res Public Health 2020 Oct 28;17(21):7890 [FREE Full text] [doi: 10.3390/ijerph17217890] [Medline: 33126469]
2. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
3. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://apps.who.int/iris/handle/10665/44607 [accessed 2024-07-10]
4. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the internet and e-mail for health care information: results from a national survey. JAMA 2003 May 14;289(18):2400-2406. [doi: 10.1001/jama.289.18.2400] [Medline: 12746364]
5. McCully SN, Don BP, Updegraff JA. Using the internet to help with diet, weight, and physical activity: results from the Health Information National Trends Survey (HINTS). J Med Internet Res 2013 Aug 01;15(8):e148 [FREE Full text] [doi: 10.2196/jmir.2612] [Medline: 23906945]
6. Weaver JB3, Mays D, Weaver SS, Hopkins GL, Eroglu D, Bernhardt JM. Health information-seeking behaviors, health indicators, and health risks. Am J Public Health 2010 Aug;100(8):1520-1525. [doi: 10.2105/AJPH.2009.180521] [Medline: 20558794]
7. Ziebland S, Chapple A, Dumelow C, Evans J, Prinjha S, Rozmovits L. How the internet affects patients' experience of cancer: a qualitative study. BMJ 2004 Mar 06;328(7439):564 [FREE Full text] [doi: 10.1136/bmj.328.7439.564] [Medline: 15001506]
8. Arcury TA, Sandberg JC, Melius KP, Quandt SA, Leng X, Latulipe C, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020 Feb;39(2):141-150 [FREE Full text] [doi: 10.1177/0733464818807468] [Medline: 30353776]

XSL•FO
RenderX

9. AshaRani PV, Jue Hua L, Roystonn K, Siva Kumar FD, Peizhi W, Ying Jie S, et al. Readiness and acceptance of ehealth services for diabetes care in the general population: cross-sectional study. J Med Internet Res 2021 Sep 02;23(9):e26881 [FREE Full text] [doi: 10.2196/26881] [Medline: 34473062]
10. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
11. Kim HS, Kim HJ, Juon HS. Racial/ethnic disparities in patient-provider communication and the role of e-health use. J Health Commun 2021 Mar 04;26(3):194-203. [doi: 10.1080/10810730.2021.1919248] [Medline: 33899688]
12. Kontos E, Blake KD, Chou WY, Prestin A. Predictors of eHealth usage: insights on the digital divide from the health information national trends survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
13. Walker DM, Hefner JL, Fareed N, Huerta TR, McAlearney AS. Exploring the digital divide: age and race disparities in use of an inpatient portal. Telemed J E Health 2020 May;26(5):603-613 [FREE Full text] [doi: 10.1089/tmj.2019.0065] [Medline: 31313977]
14. Reiners F, Sturm J, Bouw LJ, Wouters EJ. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019 Feb 21;16(4):645 [FREE Full text] [doi: 10.3390/ijerph16040645] [Medline: 30795623]
15. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625 [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
16. Yoon H, Jang Y, Vaughan PW, Garcia M. Older adults' internet use for health information: digital divide by race/ethnicity and socioeconomic status. J Appl Gerontol 2020 Jan;39(1):105-110. [doi: 10.1177/0733464818770772] [Medline: 29661052]
17. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 09;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
18. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
19. Basnyat I, Nekmat E, Jiang S, Lin J. Applying the modified comprehensive model of information seeking to online health information seeking in the context of India. J Health Commun 2018;23(6):563-572. [doi: 10.1080/10810730.2018.1493058] [Medline: 29979921]
20. Kose T, Oymak C. E-health in Turkey: an analysis of consumer activities. Health Technol 2018 Sep 5;9(2):113-121 [FREE Full text] [doi: 10.1007/s12553-018-0256-0]
21. Makowsky MJ, Jones CA, Davachi S. Prevalence and predictors of health-related internet and digital device use in a sample of South Asian adults in Edmonton, Alberta, Canada: results from a 2014 community-based survey. JMIR Public Health Surveill 2021 Jan 08;7(1):e20671 [FREE Full text] [doi: 10.2196/20671] [Medline: 33416506]
22. Hallyburton A, Evarts LA. Gender and online health information seeking: a five survey meta-analysis. J Consum Health Internet 2014 Jun 04;18(2):128-142. [doi: 10.1080/15398285.2014.902268]
23. Tong V, Raynor DK, Aslani P. Gender differences in health and medicine information seeking behaviour : a review. J Malta Coll Pharm Pract 2014;20:14-16 [FREE Full text]
24. Vazquez CE, Xie B, Shiroma K, Charness N. Individualistic versus collaborative learning in an eHealth literacy intervention for older adults: quasi-experimental study. JMIR Aging 2023 Feb 09;6:e41809 [FREE Full text] [doi: 10.2196/41809] [Medline: 36757773]
25. Xie B. Improving older adults' e-health literacy through computer training using NIH online resources. Libr Inf Sci Res 2012 Jan 01;34(1):63-71 [FREE Full text] [doi: 10.1016/j.lisr.2011.07.006] [Medline: 22639488]
26. Guo Z, Zhao SZ, Guo N, Wu Y, Weng X, Wong JY, et al. Socioeconomic disparities in eHealth literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021 Apr 14;23(4):e24577 [FREE Full text] [doi: 10.2196/24577] [Medline: 33784240]
27. Wynn R, Gabarron E, Johnsen JA, Traver V. Special issue on e-health services. Int J Environ Res Public Health 2020 Apr 22;17(8):2885 [FREE Full text] [doi: 10.3390/ijerph17082885] [Medline: 32331273]
28. McMullin JA, Duerden Comeau T, Jovic E. Generational affinities and discourses of difference: a case study of highly skilled information technology workers. Br J Sociol 2007 Jun;58(2):297-316. [doi: 10.1111/j.1468-4446.2007.00152.x] [Medline: 17610624]
29. Mcmillan SJ, Morrison M. Coming of age with the internet: a qualitative exploration of how the internet has become an integral part of young people's lives. New Media Soc 2006;8(1):73-95 [FREE Full text] [doi: 10.1177/1461444806059871]
30. Health information national trends survey 5 (HINTS 5): cycle 4 methodology report. Health Information National Trends Survey. 2020 Dec. URL: https://hints.cancer.gov/docs/methodologyreports/HINTS5_Cycle4_MethodologyReport.pdf [accessed 2024-07-11]
31. Health information national trends survey 5 (HINTS 5): cycle 2 methodology report. Health Information National Trends Survey. 2018 Jul. URL: https://hints.cancer.gov/docs/methodologyreports/HINTS5_Cycle_2_Methodology_Report.pdf [accessed 2024-07-11]

XSL•FO
RenderX

32. Health information national trends survey 6 (HINTS 6): HINTS 6 methodology report. Health Information National Trends Survey. 2023 Apr 28. URL: https://hints.cancer.gov/docs/methodologyreports/HINTS_6_MethodologyReport.pdf [accessed 2024-07-11]
33. Guidance on secondary analysis of existing data sets. University of Connecticut. URL: https://ovpr.uconn.edu/services/rics/irb/researcher-guide/secondary-analysis-of-data-sets/ [accessed 2024-07-10]
34. Dimock M. Defining generations: where millennials end and generation Z begins. Pew Research Center. 2019 Jan 17. URL: https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/ [accessed 2024-07-10]
35. SAS 9.1.3 documentation. SAS Institute Inc. URL: http://support.sas.com/documentation/onlinedoc/91pdf/index_913.html [accessed 2024-07-11]
36. Rizzo L, Moser RP, Waldron W, Wang Z, Davis WW. Analytic methods to examine changes across years using hints 2003 and 2005 data. National Cancer Institute. URL: https://hints.cancer.gov/docs/HINTS_Data_Users_Handbook-2008.pdf [accessed 2024-07-10]
37. Buyl R, Beogo I, Fobelets M, Deletroz C, Van Landuyt P, Dequanter S, et al. e-Health interventions for healthy aging: a systematic review. Syst Rev 2020 Jun 03;9(1):128 [FREE Full text] [doi: 10.1186/s13643-020-01385-8] [Medline: 32493515]
38. Fareed N, Swoboda CM, Jonnalagadda P, Walker DM, Huerta TR. Differences between races in health information seeking and trust over time: evidence from a cross-sectional, pooled analyses of HINTS data. Am J Health Promot 2021 Jan;35(1):84-92. [doi: 10.1177/0890117120934609] [Medline: 32588638]
39. Chatmon BN. Males and mental health stigma. Am J Mens Health 2020;14(4):1557988320949322 [FREE Full text] [doi: 10.1177/1557988320949322] [Medline: 32812501]
40. Harper S, Riddell CA, King NB. Declining life expectancy in the United States: missing the trees for the forest. Annu Rev Public Health 2021 Apr 01;42:381-403 [FREE Full text] [doi: 10.1146/annurev-publhealth-082619-104231] [Medline: 33326297]
41. He Z, Tian S, Singh A, Chakraborty S, Zhang S, Lustria ML, et al. A machine-learning based approach for predicting older adults' adherence to technology-based cognitive training. Inf Process Manag 2022 Sep;59(5):103034 [FREE Full text] [doi: 10.1016/j.ipm.2022.103034] [Medline: 35909793]
42. Telehealth: technology meets health care. Mayo Clinic. URL: https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/telehealth/art-20044878 [accessed 2024-07-10]
43. Zakar R, Iqbal S, Zakar MZ, Fischer F. COVID-19 and health information seeking behavior: digital health literacy survey amongst university students in Pakistan. Int J Environ Res Public Health 2021 Apr 11;18(8):4009 [FREE Full text] [doi: 10.3390/ijerph18084009] [Medline: 33920404]
44. Dadaczynski K, Okan O, Messer M, Leung AY, Rosário R, Darlington E, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Jan 15;23(1):e24097 [FREE Full text] [doi: 10.2196/24097] [Medline: 33395396]
45. Download data. NIH National Cancer Institute. URL: https://hints.cancer.gov/data/download-data.aspx [accessed 2024-07-10]

Abbreviations

HINTS: Health Information National Trends Survey
OR: odds ratio

Edited by A Mavragani; submitted 20.11.23; peer-reviewed by P AshaRani, R Moser; comments to author 08.03.24; revised version received 14.04.24; accepted 04.06.24; published 14.08.24.

Please cite as:
Vazquez CE, Mauldin RL, Mitchell DN, Ohri F
Sociodemographic Factors Associated With Using eHealth for Information Seeking in the United States: Cross-Sectional Population-Based Study With 3 Time Points Using Health Information National Trends Survey Data
J Med Internet Res 2024;26:e54745
URL: https://www.jmir.org/2024/1/e54745
doi:10.2196/54745
PMID:

©Christian Elias Vazquez, Rebecca L Mauldin, Denise N Mitchell, Faheem Ohri. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study

Dianna Vidas[1], BPsycSc, MSc, PhD; Romina Carrasco[1,2], BA, MA, PhD; Ryan M Kelly[1,3], BSc, MSc, PhD; Jenny Waycott[1], BA, PhD; Jeanette Tamplin[4], BMus (Hons), MMus, PhD; Kate McMahon[4], BA, MMusThrpy, PhD; Libby M Flynn[4], BMus, MMusThrpy, PhD; Phoebe A Stretton-Smith[4], BMus, MMusThrpy; Tanara Vieira Sousa[4], BA, MSc, PhD; Felicity A Baker[4,5], BMus, GDip, MMus, RMT, PhD

[1]School of Computing and Information Systems, The University of Melbourne, Parkville, Australia
[2]Facultad de Comunicación y Artes Audiovisuales, University of Las Américas Ecuador, Quito, Ecuador
[3]School of Computing Technologies, RMIT University, Melbourne, Australia
[4]Faculty of Fine Arts and Music, The University of Melbourne, Melbourne, Australia
[5]Centre for Research in Music and Health, Norwegian Academy of Music, Oslo, Norway

Corresponding Author:
Dianna Vidas, BPsycSc, MSc, PhD
School of Computing and Information Systems
The University of Melbourne
Grattan Street
Parkville, 3010
Australia
Phone: 61 3 90355511
Email: dianna.vidas@unimelb.edu.au

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e65977

Abstract

Background: Music has long been identified as a nonpharmacological tool that can provide benefits for people with dementia, and there is considerable interest in designing technologies to support the use of music in dementia care. However, to ensure that music technologies are appropriately designed for supporting caregivers and people living with dementia, there remains a need to better understand how music is currently used in everyday dementia care at home.

Objective: This study aims to understand how people living with dementia and their caregivers use music and music technologies in everyday caring, as well as the challenges they experience using music and technology.

Methods: This study used a mixed methods design. First, a survey was administered to 13 people living with dementia and 64 caregivers to understand their use of music and technology. Subsequently, 18 survey respondents (family caregivers: n=12, 67%; people living with dementia: n=6, 33%) participated in focus groups regarding their experiences of using music and technology in care. Interview transcripts were analyzed using reflexive thematic analysis.

Results: Most of the survey respondents (people living with dementia: 9/13, 69%; family caregivers: 47/63, 75%) reported using music *often* or *very often* in their daily lives. Participants reported a range of technologies used for listening to music, such as CDs, radio, and streaming services. Focus groups highlighted the benefits and challenges of using music and music technologies in everyday care. Participants identified using music and music technologies to regulate mood, provide joy, facilitate social interaction and connection, encourage reminiscence, provide continuity of music use before and after the dementia diagnosis, and make caregiving easier. The challenges of using music technology in everyday caring included difficulties with staying up to date with evolving technology and low self-efficacy with technology for people living with dementia.

Conclusions: This study shows that people with a dementia diagnosis and their caregivers already use music and music technologies to support their everyday care needs. The results suggest opportunities to design technologies that enable easier

XSL•FO
RenderX

access to music and to support people living with dementia with recreational and therapeutic music listening as well as music-based activities.

(J Med Internet Res 2024;26:e54186) doi:10.2196/54186

KEYWORDS

dementia; dementia care; technology; music technology; mobile phone

Introduction

Background

Worldwide, >55 million people are currently living with dementia, making dementia one of the leading causes of disability among older adults globally [1]. While many people living with dementia are aged ≥65 years, younger onset dementia can affect people of different ages. As well as difficulties with cognition, people living with dementia may experience a range of behavioral and psychological symptoms, such as agitation, anxiety, irritability, and depression [2]. These difficulties place considerable demands on family and other "informal" caregivers, who provide the majority of care for people with dementia living in the community [3]. Although informal caregivers frequently report that they need additional support and information [4], they are often reluctant, or lack knowledge, to fully use services [5], leaving many of them feeling burdened and experiencing depressive symptoms [6]. As a result, caregivers and people with dementia need accessible strategies to support their everyday care needs [7-10].

Music is a nonpharmacological tool that can provide numerous benefits for people with dementia [11-14]. Music therapy, where music is used as an intervention by a trained therapist, has long been shown to assist in managing symptoms of dementia, such as agitation and depression, and can support interpersonal connection between people with dementia and their caregivers [11,15-19]. While music therapy is a promising avenue for improving the lives of people with dementia and their caregivers, it may not be readily accessible for all people with dementia, particularly those who are less able to participate in activities outside of the home [20]. Caregiver training programs informed by music therapy, where caregivers are trained by music therapists in intentional uses of music, have received attention as a method of providing quality informal care [21]. While experiences informed by music therapy can strategically target specific care needs for people living with dementia, music activities not informed by therapeutic approaches can still be valuable for a range of purposes across the continuum from therapeutic to recreational experiences [11].

In the absence of a music therapist, recreational music experiences such as music listening or singing can also provide important everyday support for improving quality of life [11,22]. In the aged care setting, recreational music listening can provide people with dementia with enjoyment, relaxation, social connectedness, entertainment, and mood regulation [13,23]. Similar benefits of music use have been identified for older adults with dementia living at home, with music reducing agitation, improving cognition, and enhancing social well-being and connection [19,24]. However, much of the research on this topic has focused on identifying benefits gained from participating in organized music activities, such as community singing groups [18]. There has been less research investigating how informal caregivers use music and music technologies in everyday home-based care [25]. One exception, a study by Elliott et al [24], used a case study approach to examine the musical experiences of 3 people living with dementia and their caregivers, but it did not examine the technologies used to access music. As such, there remains a need for further research to understand the role of music in everyday dementia care as well as the role of technology in facilitating access to music for people with dementia and their caregivers.

While technology can support both therapeutic and recreational uses of music in dementia care, existing research has focused primarily on using technology to support recreational uses of music, primarily in formal care settings. Several technologies have been developed and trialed to support music listening and music-based activities for people living with dementia [26,27]. Researchers have designed and trialed technologies that aim to support music for dementia care by bringing enjoyment and social connection to people with dementia through both music and sound [28,29]; for example, Hodge et al [30] created a virtual reality environment of a concert experience, tailored for the specific interests of 1 person living with dementia.

Music-based technologies in the aged care context have been explored as a way to promote an enjoyable, active form of engagement with music that may facilitate connection and shared movement [31-33]. In another example, researchers created a musical interface that aimed to facilitate collaborative music making to foster partnership between aged care residents and professional caregivers [29]. Interactive approaches to designing for musical interaction for people living with dementia extend to mobile apps, which can facilitate collaborative music making [34]. In addition, music therapists use a range of technologies to engage their clients in communication, connection, and music making [35]. This suggests that music technologies have utility for recreational and therapeutic music experiences for people living with dementia; yet, their use by informal caregivers remains understudied. There are substantial differences between formal and informal care settings. Formal care often involves group activities, facilitated by trained caregivers. Conversely, informal care typically takes place in the private home, with family caregivers who may not have training in using specific care strategies in their care. While benefits shown in formal care settings likely apply across contexts, informal caregivers' existing everyday care practices with mainstream technologies, without training or formal intervention, remain unclear. Furthermore, understanding everyday uses of music technologies in caregiving can help to identify what works well and where opportunities may lie for technology innovation.

XSL•FO
RenderX

Objectives

In this paper, we aim to understand the role of music and music technologies in the everyday lives of people with dementia and their family caregivers. Here, we define music technologies broadly as technologies used to facilitate listening to (eg, music streaming services and CDs) or playing (eg, a piano app) music. We consider the potential of music technologies both for recreation and for managing the behavioral and psychological symptoms of dementia. Music is now often accessed through technologies such as streaming services, but there may be barriers to using these technologies in the complex environment of family-based dementia care. To create new technologies that support the use of music for people living with dementia at home, we must understand how music technologies are currently used in dementia care and what challenges these groups face in using music with current technologies.

This study used a mixed methods design with 2 phases. An initial survey with family caregivers and people with dementia aimed to investigate their use of music and technology. Survey participants were then invited to take part in a focus group study, which further discussed current practices with music and music technologies in dementia care. We aimed to understand how people with dementia and their family caregivers use music and music technologies as part of their everyday care practices. We further aimed to explore the perceived benefits of using music and music technologies, as well as the perceived challenges that caregivers and people living with dementia face in accessing music through technology.

Methods

Ethical Considerations

All procedures were approved by the University of Melbourne Human Research Ethics Committee (2020-14155-13119-4). A plain language statement and consent form were shown at the beginning of the survey, and respondents had to provide consent before proceeding with the survey. Survey respondents could also indicate their interest in participating in focus groups for the main study. Participants who completed the survey were entered into a draw to win an iPad Mini. Survey respondents who were willing to participate in the focus groups were sent an electronic plain language statement and consent form indicating that they could be contacted via phone or email. Participants who provided consent were then contacted by a member of the research team who invited them to take part in web-based focus group discussions. During this contact, the researcher provided an overview of the study and an opportunity for participants to ask questions. Participants could attend the sessions alone or as a dyad (caregiver and person living with dementia attending together). All participants were offered a voucher worth Aus $30 (approximately US $23) for their time.

Participant Recruitment

Survey

An anonymous web-based survey was distributed through Facebook advertisements and StepUp for Dementia Research (a web-based platform that connects volunteer research participants with dementia researchers in Australia) [36], calling for responses from Australian caregivers and people living with dementia. There were no strict eligibility criteria, and participants could self-identify as a person living with dementia or a caregiver.

Focus Groups

The focus groups did have eligibility criteria: participants were asked to self-identify that they were literate, able to speak English, and had access to stable internet as well as a laptop or tablet computer to participate in the web-based sessions via Zoom (Zoom Video Communications, Inc).

Procedure

Survey

Study data were collected and managed with REDCap (Research Electronic Data Capture; Vanderbilt University), a web-based platform that supports data capture for research studies [37,38]. After the participants provided consent, the survey asked them for demographic information (age, gender, and level of education). Next, participants were asked about how they currently use music in their daily life, whether music has been an important or meaningful part of their life in the past, their musical preferences, and the technologies used to listen to music. Participants were asked about their experience with digital technologies, such as computers, smartphones, music apps, and dementia-related apps. Finally, participants were asked about their openness to using mobile apps to support daily activities and the ways in which a music-based mobile app might be able to support daily activities. For some questions, caregivers answered both for themselves and the person they cared for, while people with dementia responded for themselves and their caregiver, if relevant. The survey (Multimedia Appendix 1) took approximately 15 minutes to complete.

Focus Groups

Five focus group sessions were organized, and participants could choose which session suited them. Each participant attended 1 focus group. All sessions were conducted via Zoom to enable the remote participation of people in different regions of Australia and to mitigate COVID-19 infection risk. Of the 18 participants, only 1 (6%; a caregiver) attended the first focus group, as the other registered participants withdrew from the study at the last minute due to unforeseen circumstances. The second and third focus groups were conducted with caregivers only. The fourth focus group was conducted with caregivers and 1 dyad (a person with dementia and their caregiver), and the fifth focus group was conducted with a mix of dyads and people living with dementia who attended the session independently (without a caregiver). The focus groups, facilitated by 2 members of the research team, were led by a researcher in human-computer interaction, who was supported by a music therapist.

The focus groups followed a protocol divided into 4 phases (Multimedia Appendix 1). Each focus group started with an introduction (5-10 min), with participants and researchers introducing themselves. The second phase for caregivers involved discussing their role as caregivers, their experiences of using music and technology in care, the types of music

XSL•FO
RenderX

technologies currently used, and how they believed digital technologies may support care. People living with dementia were similarly asked about their experiences with music and how they use technology to listen to music. The third phase involved discussing how potential future music technologies could respond to their needs, leading into a wrap-up and conclusion (phase 4). The questions were developed by the research team to align with the broader goals of the project, which aimed to understand what caregivers and people with dementia wanted from a mobile music therapy–informed app. The focus groups ranged from 46 to 71 minutes in duration. All focus group sessions were video and audio recorded.

Data Analysis

Survey responses were analyzed using descriptive statistics; we tabulated frequency data for responses to each question using Stata SE (version 16.1; StataCorp LLC). The audio recordings of the focus groups were transcribed by the researchers. We analyzed the transcripts using reflexive thematic analysis [39,40]. Consistent with the assumptions of thematic analysis, we adopted a constructionist approach, meaning that our interpretation considered the socially constructed meaning of participants' experiences alongside our own interpretations [41]. The data were coded using an inductive approach. Researcher subjectivity is central to reflexive thematic analysis, which is typically led by a single coder [42]. This analysis was led by the first author, who has training in the fields of human-computer interaction and music psychology as well as experience conducting research with people living with dementia and their caregivers.

After familiarization with the transcripts, the first author generated initial codes and developed preliminary themes, which included 6 themes grouped under the category of benefits of music and technology and 5 themes grouped under the category of barriers to use. Initial themes were discussed among a subset of the authors with expertise in human-computer interaction. Initial themes were subsequently defined and named and further refined, using an iterative approach when outlining themes during the writing process. During this phase, several themes were removed, combined, or restructured based on discussions between authors and comments on draft iterations. When a full draft was complete, further discussion with the remaining authors (with expertise in music therapy), and reflection on the presentation of findings led to additional revisions. Discussions and comments led to further narrowing the focus and framing the findings around music technologies, acknowledging that music use is almost always facilitated by technology. These revisions led to the finalization of the themes presented here.

Results

Survey

Overview

The survey was launched on the web in December 2020 and was open during the first quarter of 2021. There were 77 completed surveys (people living with dementia: n=13, 17%; caregivers: n=64, 83%). Of the 13 people living with dementia, 9 (69%) consented to be contacted for the focus groups, while 24 (38%) of the 64 caregivers gave their consent to be contacted for the focus groups.

Survey Respondents' Characteristics

Table 1 outlines the survey respondents' characteristics.

In total, 13 people living with dementia responded to the survey and, where applicable, responded on behalf of their caregiver (n=8, 62% participants). Respondents living with dementia were aged between 49 and 78 (mean 65.3, SD 9.0) years and were predominantly female (9/13, 69%). Of the 13 people living with dementia, 8 (62%) responded that they had someone providing care for them. Respondents living with dementia reported that their caregivers ranged in age from 36 to 77 (mean 55.1, SD 15.1) years and were predominantly female (5/8, 63%). Most of the caregivers were a partner or a spouse (6/8, 75%). People living with dementia were primarily diagnosed with either Alzheimer disease (3/11, 27%) or frontotemporal dementia (3/11, 27%). The remainder lived with mixed dementia (2/11, 18%), another type not specified (2/11, 18%), or vascular dementia (1/11, 9%).

In addition, 64 caregivers responded to the survey and, where applicable, completed survey details on behalf of the person with dementia for whom they were caring. The caregivers ranged in age from 29 to 86 (mean 63.0, SD 11.2) years, and the majority (46/64, 72%) were female. They reported that the people they cared for ranged in age from 45 to 93 (mean 78.1, SD 9.9) years and that more than half were female (36/64, 56%). Almost half of the caregivers (30/63, 48%) were a spouse or a partner of someone living with dementia, while a little more than a third (22/63, 35%) were adult children caring for a parent with dementia.

XSL•FO
RenderX

Table 1. Survey respondents' sociodemographic characteristics. Persons living with dementia answered questions about themselves. They were also asked questions about their caregiver for them to answer. Caregivers of a person living with dementia answered questions about themselves. They also answered questions about the person living with dementia for whom they were caring (n=77)[a].

	Persons living with dementia	Caregivers	Total
Persons with dementia: age (y), mean (SD)	65.3 (9.0)	78.1 (9.9)	76.0 (10.8)
Persons with dementia: gender, n/N (%)			
Female	9/13 (69)	36/64 (56)	45/77 (58)
Male	4/13 (31)	28/64 (44)	32/77 (42)
Dementia diagnosed, n/N (%)	11/13 (84.6)	56/64 (88)	67/77 (87)
Dementia type, n/N (%)			
Alzheimer disease	3/11 (27)	32/56 (57)	35/67 (52)
Vascular dementia	1/11 (9)	5/56 (10)	6/67 (9)
Frontotemporal dementia	3/11 (27)	5/56 (10)	8/67 (12)
Mixed dementia	2/11 (18)	2/56 (4)	4/67 (6)
Other	2/11 (18)	12/56 (214)	14/67 (21)
Time since diagnosis (years), mean (SD)	4.2 (2.1)	4.2 (2.9)	4.2 (2.7)
Persons with dementia with at least 1 caregiver, n/N (%)	8/13 (62)	64/64 (100)	72/77 (94)
Persons with dementia with >1 caregiver, n/N (%)	3/8 (38)	31/64 (55)	34/72 (47)
Dyad relationship, n/N (%)			
Spouse or partner	6/8 (75)	30/63 (48)	36/71 (51)
Parent and child	1/8 (12)	23/63 (37)	24/71 (34)
Other relative or friend	0/8 (0)	3/63 (5)	3/71 (4)
Paid carer	1/8 (12)	7/63 (11)	8/71 (11)
Caregivers: age (years), mean (SD)	55.1 (15.1)	63.0 (11.2)	62.1 (11.9)
Caregivers: gender, n/N (%)			
Female	5/8 (62)	46/64 (72)	51/72 (71)
Male	3/8 (38)	18/64 (29)	21/72 (29)

[a]Total N values differ between questions due to missing data and differences in questions displayed resulting from the participants' responses (eg, if a person living with dementia did not respond for their caregiver).

Survey Respondents' Use of Music and Technology

With regard to music use, 55% (42/76) of the people living with dementia (including those whose caregiver responded) and 75% (47/63) of the caregivers used music *often* or *very often* in their daily lives (Table 2). Furthermore, the majority of the participants (people living with dementia: 39/74, 53%; caregivers: 38/63, 60%) felt that music was *definitely* an important or meaningful part of their lives in the past. The respondents accessed music via streaming services (people living with dementia: 15/65, 23%; caregivers: 20/57, 35%), CDs (people living with dementia: 23/65, 35%; caregivers: 27/57, 47%), and radio (people living with dementia: 21/65, 32%; caregivers: 23/57, 40%). When asked about dementia-specific apps, only 15% (11/75) of all respondents had previously used one, and 62% (47/76) were willing to use a music app to support their daily activities.

When asked "What ways, if any, do you think a music-based mobile app might be able to support you in your daily activities?" people living with dementia reported that an app could help with supporting calming and relaxation (5/12, 42%) and providing entertainment or enrichment (3/12, 25%), as well as a combination of these. Caregivers responded with similar benefits, with calming and relaxation (11/57, 19%) and entertainment and enrichment (7/57, 12%) as common responses; however, 25% (14/57) of the caregivers were not certain about the benefits and suggested that a new app would have to be distinct from other music technologies to be useful to support daily activities.

XSL•FO
RenderX

Table 2. Survey respondents' use of music and technology (n=77).

	Persons living with dementia, n/N (%)	Caregivers, n/N (%)	
	Respondents answering for themselves	Respondents answering for the person they care for	Respondents answering for themselves
Current use of music			
Never	1/13 (8)	8/63 (13)	4/63 (6)
Rarely	0/13 (0)	10/63 (16)	7/63 (11)
Sometimes	3/13 (23)	12/63 (19)	5/63 (8)
Often	3/13 (23)	12/63 (19)	19/63 (30)
Very often	6/13 (46)	21/63 (33)	28/63 (44)
Music was important in the past			
Not at all	2/13 (15)	4/61 (7)	5/63 (8)
Somewhat	4/13 (31)	20/61 (33)	20/63 (32)
Definitely	7/13 (54)	32/61 (52)	38/63 (60)
Don't know	0/13 (0)	5/61 (8)	0/63 (0)
How they currently listen to music			
CDs	2/12 (17)	21/53 (40)	27/57 (47)
Radio	5/12 (42)	16/53 (30)	23/57 (40)
Music streaming services (eg, Spotify)	5/12 (42)	10/53 (19)	20/57 (35)
Video streaming services (eg, YouTube)	5/12 (42)	6/53 (11)	19/57 (33)
Digital music library (eg, iTunes library)	1/12 (8)	9/53 (17)	13/57 (23)
Television	2/12 (17)	4/53 (8)	4/57 (7)
Wireless speaker system	0/12 (0)	2/53 (4)	4/57 (7)
Vinyl records	1/12 (8)	4/53 (8)	2/57 (4)
Cassettes	1/12 (8)	2/53 (4)	0/57 (0)
Sound therapy app	1/12 (8)	0/53 (0)	0/57 (0)
Dementia-specific music player	0/12 (0)	1/53 (2)	0/57 (0)
Own a smartphone currently	11/13 (85)	—[a]	60/64 (94)
Own a tablet device currently	9/13 (69)	—	47/64 (73)
Experienced in the use of smartphone apps[b]	7/12 (58)	—	52/64 (81)
Experienced in the use of music apps[b]	5/12 (42)	—	36/64 (56)
Experienced in the use of dementia apps[b]	2/12 (17)	—	9/63 (14)
Experienced in the use of sound devices[b]	6/12 (50)	—	32/62 (52)
"Very" open to using a mobile app (or a new app) as part of daily activities[c]	8/13 (62)	—	39/63 (62)
Ways in which a music-based smartphone app might be able to support daily activities			
For calming and relaxation	5/12 (42)	—	11/57 (19)
For entertainment, enrichment, and reminiscence	3/12 (25)	—	7/57 (12)
To improve or maintain mood	0/12 (0)	—	3/57 (5)
Multiple uses	1/12 (8)	—	2/57 (4)
General benefits	0/12 (0)	—	6/57 (11)
New apps must be different from other music streaming services to be useful	1/12 (8)	—	14/57 (25)

XSL•FO
RenderX

	Persons living with dementia, n/N (%)	Caregivers, n/N (%)	
	Respondents answering for themselves	Respondents answering for the person they care for	Respondents answering for themselves
Not sure or none	2/12 (17)	—	14/57 (25)

[a]Not applicable.

[b]The answer options for this question were *no experience*, *some experience*, *quite experienced*, *very experienced*, and *expert user*; for this table, we considered *experienced users* to be those who reported the last 3 options.

[c]Respondents answered the question "How open would you be to using a mobile app (or new app) as part of your daily activities?" with *not open to the idea at all*, *a little open to the idea*, *very open to the idea*, or *undecided*.

Focus Groups

Focus Group Participants' Characteristics

Of the survey respondents willing to participate further in the study, 18 took part in the focus groups. Two people living with dementia and 2 caregivers also invited their respective care partner to participate in the focus group sessions. The focus group participants included 9 caregivers of people with dementia (partners or spouses: n=5, 56%; adult children: n=4, 44%), 3 dyads (people living with dementia: n=3, 50%; caregivers: n=3, 50% [participating together]), and 3 people currently living with dementia who attended independently. Participants ranged in age from 50 to 86 (people living with dementia: mean 72.5, SD 3.5; caregivers: mean 70.3, SD 12.1) years. Of the 18 participants, 13 (72%) identified as female, and 5 (28%) identified as male. Table 3 provides a summary of the participants' characteristics.

XSL•FO
RenderX

Table 3. Focus group participants' characteristics.

Group and pseudonym	Role	Relationship to person with dementia	Dementia type	Music technologies discussed
Group 1				
Anita[a]	Caregiver	Partner	Alzheimer disease	CDs, music streaming services (with dedicated music device), vinyl records, and iPad (with music from CDs)
Group 2				
Betty	Caregiver	Child	Memory problems	Radio and music streaming services (on smartphone)
Caroline	Caregiver	Partner	Mixed	CDs
David	Caregiver	Partner	Alzheimer disease	CDs
Group 3				
Elizabeth	Caregiver	Child	Alzheimer disease	MP3 player
Fred	Caregiver	Partner	Alzheimer disease	Music streaming services (on smartphone and tablet device) and wireless speaker system
Geraldine	Caregiver	Partner of Grace	Primary progressive aphasia	Wireless speaker system, MP3 player, video streaming services, digital music library, and music streaming services
Grace	Person with dementia	Partner of Geraldine	Primary progressive aphasia	__[b]
Group 4				
Harriet	Caregiver	Child	Alzheimer disease	Vinyl records and music streaming services (on smartphone)
Ingrid[c]	Caregiver	Child	Alzheimer disease	Radio, CDs, cassettes, and music streaming services (on smartphone and tablet device connected to a Bluetooth speaker)
Ian[c]	Caregiver	Partner	Alzheimer disease	Radio and CDs
Group 5				
James	Caregiver	Partner of Julia	Alzheimer disease	CDs, radio, music streaming services (on smartphone), video streaming services, and piano apps
Julia	Person with dementia	Partner of James	Alzheimer disease	—
Karen	Caregiver	Partner of Kevin	Lewy body	—
Kevin	Person with dementia	Partner of Karen	Lewy body	Radio and music streaming services
Lucy	Person with dementia	Attended independently	Mixed	CDs and MP3 player
Mary	Person with dementia	Attended independently	Mixed	Radio and CDs
Naomi	Person with dementia	Attended independently	Frontotemporal	Radio and vinyl records

[a]Focus group 1 was conducted with 1 participant because 2 others who were scheduled to join withdrew due to unforeseen circumstances.

[b]Not applicable.

[c]Caregivers Ian and Ingrid are father and daughter, both caring for the same person—Ian's partner and Ingrid's mother—who did not participate in the study.

Benefits of Music Technologies: How Caregivers and People Living With Dementia Use Music as Part of Their Everyday Care Practices

Caregivers and people living with dementia described four benefits of using music technologies in everyday care: (1) to regulate mood and provide joy, (2) to facilitate music-based social activities, (3) to provide continuity of music use before and after a clinical diagnosis, and (4) to make caregiving easier.

Using Music for Regulating Mood and Providing Joy

Caregivers said that they often used music and music technologies as a tool for regulating the mood of their loved ones in their everyday lives and in care practices. Caregivers believed that music was valuable for its ability to help manage the behavioral and psychological symptoms of dementia and provided joy for both caregivers and people living with dementia. Caregivers chose to listen to music out loud through shared technologies (eg, speakers) because they perceived these

XSL•FO
RenderX

potential benefits for their loved ones; for example, Harriet (all names used in the quotes herein are pseudonyms) found that listening to vinyl records could be valuable in managing her mother's mood:

> *I would say [music] lifts her mood...she does get a bit down...gets a bit paranoid sometimes, but I just think music has a calming effect and it does definitely lift her mood.*

For Fred and his wife, music had a relaxing effect even when "choosing different playlists [that are] not specific for calming."

Ingrid, who cared for her mother during the day, noted the benefits of music listening as part of a routine and at difficult times of day, notably "sundowning":

> *And if the light is changing, and I know that there's still two to three hours before Papa arrives to collect her...I say: "Shall I put some music on?" And she said: "Oh yes, please." And I find some classical works extremely well, and anything that's calming...gentle jazz.*

As well as specific emotional impacts of music, caregivers and people with dementia discussed the benefits of using music to manage the behavioral symptoms of dementia; for example, Ian, also speaking about Ingrid's mother (his partner), stated as follows:

> *[She is] a bit prickly at the moment...and then occasionally she will lash out at those closest to her, so I use music as a pacifier.*

In terms of the technologies used to access music, their portability provided benefits such as allowing music to be used when and where it was needed as part of emergent care practices for both the person living with dementia and their caregivers. Lucy, a person living with dementia, said that she found it useful to keep a small MP3 player close by to help her manage a particularly troubling symptom of dementia: hallucinations that occurred at night. Lucy found that using her MP3 player to listen to music helped her calm down during these hallucinations.

Furthermore, the portability of some music technologies such as MP3 players or streaming services such as Spotify has made music an accessible means of emotion regulation outside of the home. Anita, caring for her husband, discussed the benefits of streaming music from a mobile device:

> *It's also then something that I can take with us, you know...there are occasions when we have to go to appointments and things like that and if there is waiting time then you know I can use it there...if I feel him getting a bit agitated or anything.*

While music listening offers a range of practical benefits for people with dementia and their caregivers, participants also noted the fun and enjoyment that comes from actively engaging in music activities. Active forms of engagement such as singing, playing instruments, and moving to music were discussed. In multiple examples, participants with dementia and caregivers spoke of the joy they experienced when singing along to music; for example, Naomi, a person living with dementia, stated as follows:

> *Now because I'm having struggle speaking, but I have the joy of being able to sing...it's so much easier to sing now than actually talk some days.*

Other enjoyable activities included moving and dancing to music. Geraldine, a caregiver, spoke of how she and her partner could enjoy music and dancing together by streaming music through wireless speakers in their home:

> *We have a playlist that we both like and every now and then...one of us might go to the other person's room wherever they're cleaning you know and have a dance because it's...particularly memorable for us.*

These examples suggest that music activities beyond listening can offer benefits for people with dementia and their caregivers. Participants described how every day and spontaneous activities, such as singing and dancing to music, can be supported using technologies that are available within the home (eg, wireless speakers).

Technologies Can Facilitate Music-Based Social Activities in Care

Music and music technologies were reported to encourage social interaction. Participants shared examples that showed how participating in organized music activities facilitated social bonding for those still living at home. These activities were supported by technology that provided access to music resources beyond the setting of the organized activity.

In the home care context, Geraldine, a caregiver, described attending a social musical activity for people with dementia and their caregivers:

> *[W]e had another session which was a karaoke sing along for people and it was...through a care group so it wasn't necessarily all for people with dementia, but it was for people caring for other people...but any kind of interaction I think brings so much more to a person's life, if they are involved in doing something while the music is playing.*

These music activities were sometimes extended with technologies used in the home; for example, James, who cared for his wife at home, shared how they used a music app to assist with singing in their local choir:

> *I use certain apps to do the choir sections...when you sing in a choir, it's very hard to get your pitch right. So, you use your app, that's why I've got a keyboard to try and play the parts of the choir. So, you download, basically using YouTube and the facilities there they give you choir sections which break up the score into alto, soprano and so on.*

In other words, James and his wife used resources available through YouTube to practice their parts for the choir.

Music technologies were important because they not only enabled participants to share music with others—for example, by listening to music together—but also made it possible to experience music and a sense of companionship independently.

XSL•FO
RenderX

Ingrid, who cares for her mother with dementia and also hosts a radio show, highlighted the role that music technologies, such as radio, can play in fulfilling social needs:

> *So, radio again...it's very personal...literally there is somebody in the room there and you don't have to recognize them. You can recognize the voice, for example, you can hear, say, certain voices are absolutely unmistakable when you hear them on radio.*

Music Technologies Provide Continuity of Music Use Before and After the Diagnosis

One of the key benefits caregivers shared was that music technologies enabled continuity in music listening throughout the different stages of their experience with dementia. Caregivers and people living with dementia described their existing relationships with music and explained that listening to music provided a link between experiences before and after the dementia diagnosis.

Participants living with dementia primarily preferred listening to older styles of music from their youth; yet, caregivers also noted the potential benefits of other styles. Music from earlier decades was often familiar, and despite experiencing difficulties with memory, people with dementia still remembered songs from their youth and enjoyed singing along to them, as noted by Kevin, who has Lewy body dementia:

> *With the Lewy body, it doesn't affect your long-term memory really, it's the short-term memory. So, a lot of the old songs, you can remember. It just comes out and you sing.*

While there was substantial emphasis on jazz and classical music as genres people with dementia "used to like listening to" (Caroline), participants also noted that music taste was not fixed and that other genres could provide benefits, as noted by Elizabeth, who cared for her mother:

> *I had this...idea that, you know, mum loves Ella Fitzgerald and all those old female singers and Frank Sinatra...and I sort of had this, right it's gotta be all those oldies or Peter Allen 'cause she loves him and those songs from the '70s. But I had some success with...like one of my kids would play Taylor Swift or something...and that sort of calmed her.*

Music collections were predominantly gathered in the years before the dementia diagnosis, and music from the individuals' formative years was prominent. The technologies used were diverse across participants, often reflecting participants' music use before their diagnosis; for example, Harriet noted that her mother had "a big record collection," while Anita's husband had "CDs from...singers he would have grown up listening to." Radio was popular with both people living with dementia and caregivers, particularly stations that played classical music; for example, Lucy shared that radio "kept [her] sane" during the COVID-19 pandemic, while caregivers, such as James and Ingrid, discussed turning on the radio regularly because it was a familiar technology, as Ingrid explained:

> *So music in radio is very, very powerful because that's also a medium, in actual fact it was the medium, apart from cinema, for my parents' generation growing up.*

In addition, Fred continued to use music streaming services, with "playlists that [were] put together some years ago before the dementia set in."

Notably, despite the positive effects of music listening, participants also recalled negative experiences. In particular, music could be associated with memories that were not always welcome for the person living with dementia. After trialing an MP3 player and headphones with her mother, Elizabeth noted as follows:

> *For my mum it [music] triggered a lot of memories and she wanted to shut them down, and music upset her very much.*

For many caregivers, playing music was used as an important way to facilitate reminiscence and, as Ian described, "a different form of bringing back reality" for their loved ones; for example, Ingrid (speaking about her mother, Ian's partner) had a sense that music was special for her mother, sharing that they often played the radio for her:

> *Sometimes she doesn't recognize me but recognizes my voice. So, I do a jazz program on community radio every 3-4 weeks...and dad...will have the radio on and she will recognize my voice, and...[she] will recognize the music and the pieces, particularly if they are ones that I was familiar with growing up...So...there will be like a little trigger there.*

Participants highlighted how streaming services and MP3 players can be used to create tailored playlists that foster reminiscence for people living with dementia, but these playlists need to be used carefully, given that music can also provoke painful memories, as highlighted by Elizabeth. While the link between music and reminiscence is well known, these examples show that family caregivers are informally using music technologies to connect with their loved ones, including being mindful of both the continuity and the positive link with past experiences that music can evoke and the risk of triggering difficult memories when listening to music.

Music Technologies Help Make Caregiving Easier

As well as using music to directly support their loved ones, caregivers also used music technologies to address their own needs. By providing entertainment and relaxation for the person they were caring for, music freed up time for the caregiver and helped to make caring easier for participants and others providing care for their loved ones. In both situations, new music technologies made it easier for caregivers to use music strategically in their caregiving.

When used to provide an enjoyable distraction for the person living with dementia, music freed up caregivers' time to attend to other activities. Participants explained that newer technologies (eg, streaming services) had advantages over older modalities (eg, vinyl records) in this regard, because they could be left to play for longer, as Anita explained:

XSL•FO
RenderX

> *We have lots of old records...but I find that if I play the music there, I've no sooner got it on than it seems to stop...[it] interrupts another activity that I might be involved in so I find the newer technologies leave me freer to do other things here at home for a longer time.*

Music was used not only by the primary caregiver but also by others in the care network. Fred had set up music technologies in the home before his wife's diagnosis. This allowed visiting caregivers to use these technologies to play music when caring for Fred's wife:

> *There's a carer in each day...and if I'm out they can still tap into [the Wi-Fi] and play music that's on their [phone]...That's another big win now, exploring with that so they can get music going for [his wife] and either their own music, their own playlists, individual songs, etcetera.*

Finally, caregivers described using music technology to entertain their loved one. David mentioned putting on music in the background so that his wife was not sitting in silence. Anita said that she frequently plays CDs at home to fill spare time:

> *I just use it in those times, when there's really nothing else [to do] apart from sleep and just sitting staring.*

Barriers for Music Technology in Care

Despite the potential advantages of using music-listening technologies in care, several issues with technology use were raised by the participants. Barriers to use centered around a difficulty for caregivers and people living with dementia to stay up to date with constantly evolving technologies as well as issues with confidence and ability to use digital technologies.

Difficulty Staying Up to Date With Evolving Technology

Participants described difficulties staying up to date with evolving technology in terms of smartphone apps and other music technologies. Even if familiar, these technologies sometimes became more difficult for people living with dementia and caregivers to use over time.

As a self-described older caregiver, James said he experienced difficulties with the planned obsolescence common to new digital technologies:

> *To try to adapt as you're older to the phones, and they're designed to be out of date within six years and you gotta find a new one and it all changes, I just find it...very difficult to negotiate.*

Caregivers faced similar barriers with smartphone apps. Geraldine mentioned an instance when she and her partner with dementia stopped using an app that they had previously found useful. The app was practical until they changed from one smartphone brand to another where the app displayed differently, and they encountered errors:

> *[Grace] had an iPhone at the time and then changed over to Samsung, and it was an Apple app. The Samsung version is not nearly as user friendly...the Apple one was just visually clearer...seemed to be easier to upload [photos]...this one...got stuck...it wouldn't let me [upload]...So Grace doesn't use that so much now.*

When discussing other music technologies, Harriet said her mother still relies on vinyl records, while David, caring for his wife, stated that he still uses CDs and does not know how to use newer forms of music technology:

> *I would prefer CDs I think because they're controllable...and there are different ones. Presumably you can do that all digitally but I don't really know how to do that.*

Similarly, Lucy, a person living with dementia, often listened to the radio but believed that other digital technologies might make music listening easier if she could learn how to use them. Lucy did use an MP3 player but faced difficulties downloading music onto the device:

> *It's like, I know music, listening to it in those little...iPods...because it's handy in the night if I feel like the traumas are coming up or some hallucinations. Immediately put that [on], it will at least calm me down. But I don't have the know-how, how to do that...the downloading...I'm not very good at.*

Participants With Dementia May Have Low Self-Efficacy With Technology

While older technologies tended to be more familiar, people with dementia still faced challenges with using some technologies and expressed low self-efficacy with technology; for example, Harriet found that her mother often had trouble putting on a vinyl record:

> *I think she panics. She panics with anything with any buttons. If she calms down enough, she'll do it. But if she starts overthinking it, then she just panics and can't do it.*

Mary, a person living with dementia, stated that it was common for people with dementia to feel overwhelmed when talking about smartphone technologies:

> *As you talk "app," nine-tenths of the people I know with dementia are going to back away because that's beyond them.*

As a result of these challenges, several caregivers discussed how important it was that apps and smartphone-based technologies were easy to use, as explained by Harriet:

> *I'm just all about functionality: easiness. Basically, the easier the better. The more basic, the better.*

The accessibility of smartphone technology was also considered at length, with participants discussing hearing issues (Betty) and vision impairment; for example, James said that "[a] bigger screen is definitely helpful rather than small screens." There were also physical issues with using mobile technology, as noted by Ingrid:

> *[I]f I'm typing [on the phone]...I really have to be very careful, how I type, and...that's just me [a younger person] so for somebody who no longer has*

XSL•FO
RenderX

agility, let alone mobility with fingers, that could be an issue.

In addition, some participants perceived smartphone apps as difficult to use for people living with dementia; for example, Betty was concerned that setting up a music app for her mother with dementia would require "knowledge, information or training for the person setting it up, which probably may well be the carer."

Discussion

Principal Findings

Using a mixed methods design, this study aimed to understand how people living with dementia and their family caregivers use music and music technologies in their everyday caring, as well as the challenges they experience using music technologies. In our survey, we found that the majority of participants living with dementia (9/13, 69%) and caregivers (47/63, 75%) use music *often* or *very often* in their daily lives, with music having been an important or meaningful part of their lives in the past. Consistent with prior work, participants reported using a range of methods and devices for accessing music, with high rates of use for music and video streaming services, as well as the use of more traditional music technologies such as CDs and radio [43]. Participants were interested in new ways of using music to support their care, with a willingness to try new music-based mobile apps, and considered a range of potential benefits such an app could provide. These findings, alongside previous work supporting the important role that music can play in the everyday lives of people living with dementia [24], suggest that new music technologies can be designed to complement and extend existing uses. New technology-based solutions, if designed appropriately to ensure ease of use and adding no further burden to caregivers, may provide valuable benefits for this group.

Through the focus groups, we found several perceived benefits of using music technologies, alongside the challenges that caregivers and people living with dementia face in using music and technology in these contexts. We found that people living with dementia and their caregivers use music for regulating mood, providing joy, and encouraging social connection, as well as to make caregiving easier, provide continuity, and promote reminiscence. These uses align with benefits identified in research evaluating music interventions with people living with dementia [11,23,25,28,31]; however, we additionally highlight how these benefits are demonstrated in everyday caregiving practices. We also identified barriers to using music technologies, predominantly due to participants with dementia having low self-efficacy with technology and challenges for both caregivers and people living with dementia in staying up to date with evolving technology. Here, we consider how future music technologies can be designed to capitalize on the existing uses of music technology among caregivers of people living with dementia and to empower caregivers to use music strategically and therapeutically at home [7].

People living with dementia and caregivers predominantly used music during everyday care for mood regulation, such as relaxation, to lift mood, and to provide joy. This extends prior research findings that caregivers and people living with dementia often use music for specific functions, such as relaxation or calming [23]; for example, Ingrid listened to music with her mother to manage behaviors such as late-day confusion and restlessness. In addition, participants spoke fondly of music's ability to simply provide joy, echoing previous work where music and sound technologies have been shown to enrich the lives of people living with dementia and provide enjoyable moments at home [29]. This highlights the need for music technologies and interventions to consider the enjoyable and recreational aspects of music as well as its potential therapeutic benefits [11]. Importantly, digital technologies provide an avenue for these benefits to be translated to situations outside of the home, with Anita and Lucy suggesting that music could be a convenient and portable mood regulation tool.

The role of music and music technologies for supporting social connection, both within a care dyad and with others, was evidenced in our study and aligns with prior research [25,29,31-33]. Participating in music activities outside of the home fostered social connections for people living with dementia, as did music activities within residential aged care settings. Extending prior work, our findings show how everyday technologies were important to these interactions; for example, James discussed how mobile apps supported him and his partner in practicing their choir singing.

We found that caregivers and people living with dementia described existing relationships with music, outlining some continuity between experiences before and after the dementia diagnosis. Music technologies facilitated this continuity, enabling caregivers to encourage the use of familiar technologies, such as CDs and vinyl records, as well as personalized playlists. Music use was often associated with reminiscence, with people living with dementia reporting that they could sing along to old favorites, even if they experienced challenges with memory or other forms of verbal communication. Consistent with the idea of the "reminiscence bump," where individuals tend to favor the music of their youth [44-46], many people living with dementia preferred older styles of music. Existing relationships with music were also connected with collections of vinyl records, CDs, and favorite radio stations, with our findings also illustrating the importance of radio for older adults during everyday care [47,48].

Finally, consistent with previous work [25,49], music technologies were important for making caregiving easier, such as accessing music to fill time or silence. We extend these findings, showing that modern music technologies, such as wireless speakers, enabled music to be played for the person living with dementia not only by the primary carer but also by others in the broader care network, such as visiting caregivers, with minimal setup. This enhanced everyday care for our participants.

Barriers and Design Opportunities

While previous research has identified the uses of music in residential aged care settings [13,23] and designed tools for music listening, sound, and music activities in dementia care [27-30,33,34], there has been limited research investigating the everyday uses of music technologies in care practices by family caregivers of people living with dementia. To date, no research

XSL•FO
RenderX

has clearly identified opportunities for music technologies to play a role in the therapeutic use of music at home. Building on our findings, ensuring the ease of use of music technologies is likely crucial for engaging caregivers and people living with dementia in music use, both for its therapeutic benefits and for recreational uses of music. This is especially pertinent given that the major barriers to music technology use in everyday care included difficulties with self-efficacy around technology use and challenges staying up to date with evolving technology. Previous research has established the importance of digital literacy and self-efficacy in technology use among older adults [50], indicating that this is an important consideration when designing technologies for people living with dementia. Therefore, if caregivers and people living with dementia are not confident using existing technologies to support music use, co-designing new technologies that capitalize on the everyday ways in which these groups use music technologies should be a priority.

Our findings suggest several design opportunities for future technologies. New technologies should be designed to align with how music is used to regulate mood and capitalize on the ways in which music technologies can be used as a portable emotion regulation tool. Music technologies can also be used to foster social interaction between people living with dementia and their caregivers, with videoconferencing tools and apps designed to support web-based music activities, such as choirs and dancing. In addition, in designing music technologies for people living with dementia and their caregivers, access to personally preferred music is crucial. Often, this may be from the reminiscence bump (ie, music from a person's youth); however, music taste should not be assumed nor limited to older styles. These considerations for music selection and the affordances of personalized playlist creation highlight the benefits that streaming platforms might have over physical music collections, as long as they are easy to access for both people living with dementia and caregivers. One way this could be achieved is through playlists configured to play on a smart speaker with an easy-to-remember prompt that activates the desired playlist. Research has shown that voice interaction devices such as smart speakers (eg, Amazon Alexa and Google Home) have accessibility advantages over screen-based technologies for some users [51-54]. Music technologies can also be designed to make it easy for caregivers to play recorded music and use music in intentional ways to relieve pressure in family caregivers [6]. This could extend to co-designing new music technologies that train caregivers to use music therapeutically in their everyday care [7,55]. Focusing on such factors and ensuring that people living with dementia and their caregivers are comfortable using such technologies is a crucial first step to supporting the use of music technologies in care.

Strengths, Limitations, and Future Directions

This is the first study to our knowledge that provides an in-depth understanding of how caregivers and people living with dementia use music and music technologies in their everyday caregiving practices, outside the context of a formal intervention. Understanding these everyday uses of music technologies in caregiving highlights opportunities for new technological innovation. We intend to use these findings to inform the design of future music-based technologies for this setting, capitalizing on these existing uses of music technologies. A further strength of this study was the involvement of both caregivers and people living with dementia in our survey and focus groups, recognizing the importance of understanding the perspectives of both groups.

One limitation of this study was that due to the COVID-19 pandemic at the time of recruitment, the complexity of this care setting, and the aim to recruit participants across Australia, recruitment and data collection took place entirely on the web. As such, our sample may not be representative of people living with dementia and caregivers with less technology experience. Future research may consider recruiting participants with a more diverse experience with digital technologies. Furthermore, our findings may not be generalizable outside of the Australian context.

Conclusions

Informal caregivers of people with dementia report that they need additional support and information [4]; yet, they are often reluctant, or lack knowledge, to fully use services [5]; in addition, many family caregivers experience high levels of stress and burden [6]. As a result, caregivers and people living with dementia need accessible, everyday strategies to draw on to support their needs, and technologically supported music use is a key avenue for future research [7,9,27]. This study draws attention to the ways in which these groups already use music technologies to support their needs, outside the context of organized activities or targeted music interventions. Caregivers and people living with dementia believe that music technologies may be a useful way to access music for regulating mood, providing joy, encouraging social connection, providing continuity, promoting reminiscence, and making caregiving easier. The technologies used to play music are diverse across users, and people living with dementia and their caregivers face barriers to using new digital music technologies, despite their acknowledged advantages over more familiar technologies. These findings will contribute to understanding how new music technologies and interventions may complement existing practices. There is a clear opportunity to design technologies to capitalize on existing music use in these groups to enable easier access to specific music activities to support people living with dementia with recreational and therapeutic music listening as well as music-based activities.

Authors' Contributions

The overall project was developed by FAB, with contributions from TVS, JT, and JW. RC, TVS, JT, JW, and FAB designed the study. RC managed focus group data collection. KM assisted RC in focus group data collection, and PAS-S contacted participants for the focus groups. Survey data were summarized by TVS, with further refinement by DV during writing. TVS and RC transcribed

XSL•FO
RenderX

the focus groups. The transcripts of the focus groups were checked by DV and coded by DV using an inductive approach. After familiarization with the transcripts, DV generated initial codes and developed initial themes, with JW and RMK reviewing and refining themes into the final set. DV, RMK, and JW defined and named the final themes, further refining them with the rest of the authors during writing. All authors edited and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions and semistructured interview guide.
[DOCX File , 33 KB - jmir_v26i1e54186_app1.docx]

References

1. Dementia. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2022-12-13]
2. Cerejeira J, Lagarto L, Mukaetova-Ladinska EB. Behavioral and psychological symptoms of dementia. Front Neurol 2012;3:73 [FREE Full text] [doi: 10.3389/fneur.2012.00073] [Medline: 22586419]
3. Support for informal carers under the spotlight at Royal Commission into aged care. Dementia Australia. 2019. URL: https://www.dementia.org.au/node/88901 [accessed 2023-03-27]
4. Peeters JM, Van Beek AP, Meerveld JH, Spreeuwenberg PM, Francke AL. Informal caregivers of persons with dementia, their use of and needs for specific professional support: a survey of the National Dementia Programme. BMC Nurs 2010 Jun 07;9(1):9 [FREE Full text] [doi: 10.1186/1472-6955-9-9] [Medline: 20529271]
5. Brodaty H, Thomson C, Thompson C, Fine M. Why caregivers of people with dementia and memory loss don't use services. Int J Geriatr Psychiatry 2005 Jun;20(6):537-546. [doi: 10.1002/gps.1322] [Medline: 15920707]
6. Papastavrou E, Kalokerinou A, Papacostas SS, Tsangari H, Sourtzi P. Caring for a relative with dementia: family caregiver burden. J Adv Nurs 2007 Jun;58(5):446-457. [doi: 10.1111/j.1365-2648.2007.04250.x] [Medline: 17442030]
7. Carrasco R, Baker FA, Bukowska AA, Clark IN, Flynn LM, McMahon K, et al. Empowering caregivers of people living with dementia to use music therapeutically at home: design opportunities. In: Proceedings of the 32nd Australian Conference on Human-Computer Interaction. 2020 Presented at: OzCHI '20; December 2-4, 2020; Sydney, Australia p. 198-209 URL: https://dl.acm.org/doi/10.1145/3441000.3441082 [doi: 10.1145/3441000.3441082]
8. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. J Med Internet Res 2020 Sep 09;22(9):e19468 [FREE Full text] [doi: 10.2196/19468] [Medline: 32902388]
9. Bosch LB, Kanis M. Design opportunities for supporting informal caregivers. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2016 Presented at: CHI EA '16; May 7-12, 2016; San Jose, CA p. 2790-2797 URL: https://dl.acm.org/doi/10.1145/2851581.2892354 [doi: 10.1145/2851581.2892354]
10. Meiland F, Innes A, Mountain G, Robinson L, van der Roest H, García-Casal JA, et al. Technologies to support community-dwelling persons with dementia: a position paper on issues regarding development, usability, effectiveness and cost-effectiveness, deployment, and ethics. JMIR Rehabil Assist Technol 2017 Jan 16;4(1):e1 [FREE Full text] [doi: 10.2196/rehab.6376] [Medline: 28582262]
11. Clements-Cortés A. Understanding the Continuum of Musical Experiences for People With Dementia. Oxford, UK: Oxford University Press; 2019.
12. Dingle GA, Sharman LS, Bauer Z, Beckman E, Broughton M, Bunzli E, et al. How do music activities affect health and well-being? A scoping review of studies examining psychosocial mechanisms. Front Psychol 2021 Sep 8;12:713818 [FREE Full text] [doi: 10.3389/fpsyg.2021.713818] [Medline: 34566791]
13. McDermott O, Orrell M, Ridder HM. The importance of music for people with dementia: the perspectives of people with dementia, family carers, staff and music therapists. Aging Ment Health 2014 Jan 13;18(6):706-716 [FREE Full text] [doi: 10.1080/13607863.2013.875124] [Medline: 24410398]
14. Lee YE, Sousa TV, Stretton-Smith PA, Gold C, Geretsegger M, Baker FA. Demographic and clinical profile of residents living with dementia and depressive symptoms in Australian private residential aged care: Data from the Music Interventions for Dementia and Depression in ELderly care (MIDDEL) cluster-randomised controlled trial. Australas J Ageing 2022 Dec 08;41(4):e387-e396 [FREE Full text] [doi: 10.1111/ajag.13104] [Medline: 35801957]
15. Tamplin J, Clark IN, Lee YE, Baker FA. Remini-Sing: a feasibility study of therapeutic group singing to support relationship quality and wellbeing for community-dwelling people living with dementia and their family caregivers. Front Med (Lausanne) 2018 Aug 31;5:245 [FREE Full text] [doi: 10.3389/fmed.2018.00245] [Medline: 30234118]
16. Zhang Y, Cai J, An L, Hui F, Ren T, Ma H, et al. Does music therapy enhance behavioral and cognitive function in elderly dementia patients? A systematic review and meta-analysis. Ageing Res Rev 2017 May;35:1-11. [doi: 10.1016/j.arr.2016.12.003] [Medline: 28025173]

XSL•FO
RenderX

17. van der Steen JT, van Soest-Poortvliet MC, van der Wouden JC, Bruinsma MS, Scholten RJ, Vink AC. Music-based therapeutic interventions for people with dementia. Cochrane Database Syst Rev 2017 May 02;5(5):CD003477 [FREE Full text] [doi: 10.1002/14651858.CD003477.pub3] [Medline: 28462986]
18. Stedje K, Kvamme TS, Johansson K, Sousa TV, Odell-Miller H, Stensæth KA, et al. The influence of home-based music therapy interventions on relationship quality in couples living with dementia-an adapted convergent mixed methods study. Int J Environ Res Public Health 2023 Feb 06;20(4):2863 [FREE Full text] [doi: 10.3390/ijerph20042863] [Medline: 36833562]
19. McMahon K, Clark IN, Stensæth K, Wosch T, Odell Miller H, Bukowska A, et al. A qualitative systematic review of the experiences of sharing music for people living with dementia and their family care partners: the thread of connection. Arts Health 2023 Oct 12;15(3):229-256. [doi: 10.1080/17533015.2022.2128381] [Medline: 36224535]
20. Clark IN, Tamplin JD, Baker FA. Community-dwelling people living with dementia and their family caregivers experience enhanced relationships and feelings of well-being following therapeutic group singing: a qualitative thematic analysis. Front Psychol 2018;9:1332 [FREE Full text] [doi: 10.3389/fpsyg.2018.01332] [Medline: 30104994]
21. Baker FA, Bloska J, Braat S, Bukowska A, Clark I, Hsu MH, et al. HOMESIDE: home-based family caregiver-delivered music and reading interventions for people living with dementia: protocol of a randomised controlled trial. BMJ Open 2019 Nov 19;9(11):e031332 [FREE Full text] [doi: 10.1136/bmjopen-2019-031332] [Medline: 31748300]
22. Garrido S, Dunne L, Stevens CJ, Chang E. Music playlists for people with dementia: trialing a guide for caregivers. J Alzheimers Dis 2020;77(1):219-226 [FREE Full text] [doi: 10.3233/JAD-200457] [Medline: 32741829]
23. Krause AE, Davidson JW. A qualitative exploration of aged-care residents' everyday music listening practices and how these may support psychosocial well-being. Front Psychol 2021;12:585557 [FREE Full text] [doi: 10.3389/fpsyg.2021.585557] [Medline: 33746821]
24. Elliott M, Gardner P, Narushima M, McCleary L. Music lessons: exploring the role and meaning of music for older adults with dementia. Can J Aging 2020 Jan 28;39(4):586-599. [doi: 10.1017/s071498081900076x]
25. Elliott M, Gardner P. The role of music in the lives of older adults with dementia ageing in place: a scoping review. Dementia (London) 2018 Feb 18;17(2):199-213. [doi: 10.1177/1471301216639424] [Medline: 26993049]
26. Creech A. Using music technology creatively to enrich later-life: a literature review. Front Psychol 2019 Jan 30;10:117 [FREE Full text] [doi: 10.3389/fpsyg.2019.00117] [Medline: 30761052]
27. MacRitchie J, Floridou GA, Christensen J, Timmers R, de Witte L. The use of technology for arts-based activities in older adults living with mild cognitive impairment or dementia: a scoping review. Dementia (London) 2023 Jan 04;22(1):252-280 [FREE Full text] [doi: 10.1177/14713012221127359] [Medline: 36194002]
28. Houben M, Brankaert R, Kenning G, Bongers I, Eggen B. Designing for everyday sounds at home with people with dementia and their partners. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-15 URL: https://dl.acm.org/doi/10.1145/3491102.3501993 [doi: 10.1145/3491102.3501993]
29. Houben M, Lehn B, Van Den Brink N, Diks S, Verhoef J, Brankaert R. Turnaround: exploring care relations in dementia through design. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI EA '20; April 25-30, 2020; Honolulu, HI p. 1-8 URL: https://dl.acm.org/doi/10.1145/3334480.3382846 [doi: 10.1145/3334480.3382846]
30. Hodge J, Balaam M, Hastings S, Morrissey K. Exploring the design of tailored virtual reality experiences for people with dementia. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC p. 1-13 URL: https://dl.acm.org/doi/10.1145/3173574.3174088 [doi: 10.1145/3173574.3174088]
31. Morrissey K, Mccarthy J. Creative and opportunistic use of everyday music technologies in a dementia care unit. In: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition. 2015 Presented at: C&C '15; June 22-25, 2015; Glasgow, UK p. 295-298 URL: https://dl.acm.org/doi/10.1145/2757226.2757228 [doi: 10.1145/2757226.2757228]
32. Morrissey K, McCarthy J, Pantidi N. The value of experience-centred design approaches in dementia research contexts. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO p. 1326-1338 URL: https://dl.acm.org/doi/10.1145/3025453.3025527 [doi: 10.1145/3025453.3025527]
33. Morrissey K, Wood G, Green D, Pantidi N, McCarthy J. 'I'm a rambler, I'm a gambler, I'm a long way from home': the place of props, music, and design in dementia care. In: Proceedings of the 2016 ACM Conference on Designing Interactive Systems. 2016 Presented at: DIS '16; June 4-8, 2016; Brisbane, Australia p. 1008-1020 URL: https://dl.acm.org/doi/10.1145/2901790.2901798 [doi: 10.1145/2901790.2901798]
34. Favilla S, Pedell S. Touch screen ensemble music: collaborative interaction for older people with dementia. In: Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration. 2013 Presented at: OzCHI '13; November 25-29, 2013; Adelaide, Australia p. 481-484 URL: https://dl.acm.org/doi/10.1145/2541016.2541088 [doi: 10.1145/2541016.2541088]
35. Baglione AN, Clemens MP, Maestre JF, Min A, Dahl L, Shih PC. Understanding the technological practices and needs of music therapists. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-25. [doi: 10.1145/3449107]

XSL•FO
RenderX

36. Jeon Y, Shin M, Smith A, Beattie E, Brodaty H, Frost D, et al. Early implementation and evaluation of stepup for dementia research: an Australia-wide dementia research participation and public engagement platform. Int J Environ Res Public Health 2021 Oct 28;18(21):11353 [FREE Full text] [doi: 10.3390/ijerph182111353] [Medline: 34769871]
37. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
38. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
39. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
40. Terry G, Hayfield N. Essentials of Thematic Analysis. Washington, DC: American Psychological Association; 2021.
41. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2021 Jun 26;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
42. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022 Feb 13;9(1):3-26. [doi: 10.1037/qup0000196]
43. Krause AE, North AC, Hewitt LY. Music-listening in everyday life: devices and choice. Psychol Music 2013 Aug 14;43(2):155-170. [doi: 10.1177/0305735613496860]
44. Jakubowski K, Eerola T, Tillmann B, Perrin F, Heine L. A cross-sectional study of reminiscence bumps for music-related memories in adulthood. Music Sci (Lond) 2020 Oct 23;3:205920432096505. [doi: 10.1177/2059204320965058]
45. Rubin DC, Wetzler SE, Nebes RD. Autobiographical memory across the lifespan. In: Rubin DC, editor. Autobiographical Memory. Cambridge, MA: Cambridge University Press; 1986:202-222.
46. Zimprich D, Wolf T. The distribution of memories for popular songs in old age: an individual differences approach. Psychol Music 2015 Apr 01;44(4):640-657. [doi: 10.1177/0305735615578708]
47. Krause AE. The role and impact of radio listening practices in older adults' everyday lives. Front Psychol 2020 Dec 18;11:603446 [FREE Full text] [doi: 10.3389/fpsyg.2020.603446] [Medline: 33391122]
48. Krause AE, Fletcher H. The purpose of radio and how it supports older adults' wellbeing. Front Comput Sci 2023 Jun 20;5:1111806 [FREE Full text] [doi: 10.3389/fcomp.2023.1111806]
49. Lewis V, Bauer M, Winbolt M, Chenco C, Hanley F. A study of the effectiveness of MP3 players to support family carers of people living with dementia at home. Int Psychogeriatr 2015 Mar;27(3):471-479. [doi: 10.1017/S1041610214001999] [Medline: 25247750]
50. Hill R, Betts LR, Gardner SE. Older adults' experiences and perceptions of digital technology: (Dis)empowerment, wellbeing, and inclusion. Computers in Human Behavior 2015 Jul;48:415-423. [doi: 10.1016/j.chb.2015.01.062]
51. Kowalski J, Jaskulska A, Skorupska K, Abramczuk K, Biele C, Kope? W, et al. Older adults and voice interaction: a pilot study with Google Home. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI EA '19; May 4-9, 2019; Glasgow, UK p. 1-6 URL: https://dl.acm.org/doi/10.1145/3290607.3312973 [doi: 10.1145/3290607.3312973]
52. Jakob D. Voice controlled devices and older adults – a systematic literature review. In: Proceedings of the 8th International Conference, ITAP 2022, Held as Part of the 24th HCI International Conference on Human Aspects of IT for the Aged Population, Design, Interaction and Technology Acceptance. 2022 Presented at: HCII '22; June 26-July 1, 2022; Virtual Event p. 175-200 URL: https://link.springer.com/chapter/10.1007/978-3-031-05581-2_14 [doi: 10.1007/978-3-031-05581-2_14]
53. Cuadra A, Bethune J, Krell R, Lempel A, Hänsel K, Shahrokni A, et al. Designing voice-first ambient interfaces to support aging in place. In: Proceedings of the 2023 ACM Designing Interactive Systems Conference. 2023 Presented at: DIS '23; July 10-14, 2023; Pittsburgh, PA p. 2189-2205 URL: https://dl.acm.org/doi/10.1145/3563657.3596104 [doi: 10.1145/3563657.3596104]
54. McCloud R, Perez C, Bekalu MA, Viswanath K. Using smart speaker technology for health and well-being in an older adult population: pre-post feasibility study. JMIR Aging 2022 May 09;5(2):e33498 [FREE Full text] [doi: 10.2196/33498] [Medline: 35532979]
55. Vidas D, Thompson Z, Kelly RM, Waycott J, Tamplin J, Sousa T, et al. Family caregiver experiences of using a mobile app for music-based training to support dementia care. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. 2024 Presented at: CHI '24; May 11-16, 2024; Honolulu, HI p. 1-16 URL: https://dl.acm.org/doi/10.1145/3613904.3642308 [doi: 10.1145/3613904.3642308]

Abbreviations

REDCap: Research Electronic Data Capture

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 01.11.23; peer-reviewed by M Shin, B Waters, G Goodall; comments to author 12.03.24; revised version received 03.05.24; accepted 28.06.24; published 27.08.24.

Please cite as:
Vidas D, Carrasco R, Kelly RM, Waycott J, Tamplin J, McMahon K, Flynn LM, Stretton-Smith PA, Sousa TV, Baker FA
Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study
J Med Internet Res 2024;26:e54186
URL: https://www.jmir.org/2024/1/e54186
doi:10.2196/54186
PMID:

©Dianna Vidas, Romina Carrasco, Ryan M Kelly, Jenny Waycott, Jeanette Tamplin, Kate McMahon, Libby M Flynn, Phoebe A Stretton-Smith, Tanara Vieira Sousa, Felicity A Baker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mediating Effects of Information Access on Internet Use and Multidimensional Health Among Middle-Aged and Older Adults: Nationwide Cross-Sectional Study

Liping Fu[1,2*], Prof Dr; Caiping Liu[1,2*], PhD; Yongqing Dong[1], Prof Dr; Xiaodong Ma[2], Prof Dr; Quanling Cai[1,2], PhD; Dongli Li[1,3], PhD; Kaisheng Di[1,2], PhD

[1]College of Management and Economics, Tianjin University, Tianjin, China

[2]College of Politics and Public Administration, Qinghai Minzu University, Xining, China

[3]College of Chunming, Hainan University, Haikou, China

[*]these authors contributed equally

Corresponding Author:
Caiping Liu, PhD
College of Management and Economics
Tianjin University
Weijin Road Campus No. 92
Xuefu Street, Nankai District
Tianjin, 300072
China
Phone: 86 13997138522
Email: liucaiping@tju.edu.cn

Abstract

Background: With the exacerbation of population aging, the health issues of middle-aged and older adults have increasingly become a focus of attention. The widespread use of the internet has created conditions for promoting the health of this demographic. However, little is known about the effects of information access in promoting the relationship between internet use and the health of middle-aged and older adults.

Objective: This study aims to examine the relationship between internet use and multidimensional health in middle-aged and older adults, as well as the mediating effect of information access. Moreover, this study will explore the relationship between other dimensions of internet use (purposes and frequency) and health.

Methods: Data were sourced from the China General Social Survey conducted in 2018. Health outcomes, including self-rated, physical, and mental health, were assessed using the 5-level self-rated health scale, the 5-level basic activities of daily living scale, and the 5-level depression scale, respectively. The ordinal logistic regression model was used to examine the relationship between internet use and health among middle-aged and older adults. Additionally, the Karlson-Holm-Breen decomposition method was used to examine the mediation effect of information access. To address endogeneity issues, the two-stage least squares approach was applied.

Results: In our sample, nearly half (n=3036, 46.3%) of the respondents use the internet. Regression analyses revealed that internet use was positively associated with self-rated health (odds ratio [OR] 1.55, 95% CI 1.39-1.74; $P<.001$), physical health (OR 1.39, 95% CI 1.25-1.56; $P<.001$), and mental health (OR 1.33, 95% CI 1.19-1.49; $P<.001$) of middle-aged and older adults. Various dimensions of internet use positively contribute to health. In addition, information access significantly mediated the relationship between internet use and self-rated health (β=.28, 95% CI 0.23-0.32), physical health (β=.40, 95% CI 0.35-0.45), and mental health (β=.16, 95% CI 0.11-0.20). Furthermore, there were significant differences in the relationship between internet use and health among advantaged and disadvantaged groups.

Conclusions: The study showed that different dimensions of internet use are associated with better self-rated health, better physical health, and better mental health in middle-aged and older adults. Information access mediates the relationship between internet use and health. This result emphasizes the significance of promoting internet access as a means to enhance the health of middle-aged and older adults in China.

(J Med Internet Res 2024;26:e49688) doi:10.2196/49688

XSL•FO
RenderX

KEYWORDS

internet use; health; middle-aged and older adults; information access; mediation analysis

Introduction

Background

With the extension of human life expectancy and the decline in birth rates, the global trend of population aging is becoming increasingly pronounced [1]. This phenomenon poses significant challenges to health care and social security systems, making the promotion of healthy aging crucial for socioeconomic development and realizing global health governance [2]. Health is the foundation for enhancing life expectancy, improving quality of life, and achieving healthy aging [3].

In today's world, with the advancement of information technology, the internet is integrating into various aspects of society, profoundly changing people's ways of production and living. According to Digital 2022: Global Digital Reports, as of January 2022, global internet users have reached 4.95 billion, with an internet penetration rate of 62.5% [4]. Meanwhile, as of December 2022, China's internet users have surpassed 1.067 billion, including 153 million older people internet users [5]. The rapid spread and development of the internet have had a significant influence on the field of health. For instance, it enables access to health information [6], intervenes in health behaviors [7], monitors public health [8], formulates health policies [9], and provides medical services [10].

Evidence is mixed on the influence of internet use on health in middle-aged and older adults. One perspective suggests that the internet positively influences health and is considered an effective tool for improving health status [11-13]. Evidence indicates that middle-aged and older adults view the internet as a communication platform for expressing emotions and maintaining social relationships [14]. The internet has established novel social connections, expanding the external social networks of older individuals and helping them regain social opportunities [15]. Furthermore, the internet has strengthened intergenerational family bonds [16]. It has substituted face-to-face communication methods, reconstructing a web-based family space for older individuals. Adult children can provide telecare services for older adults, thereby partially fulfilling the expectations of intimacy that older individuals have toward their children [17]. In summary, the internet has created a new social space for older adults not limited by physical boundaries, significantly enhancing their overall physical and mental health and positively influencing intergenerational relationships [18,19]. In addition, the internet's web-based consultation and telemedicine have enhanced the accessibility of health care services for middle-aged and older adults [20,21]. However, another perspective suggests that internet use may adversely affect health [22-24]. Research from the Stanford Institute for the Quantitative Study of Society and the Kennedy School of Government has revealed that internet use does not reduce social isolation [25,26]. It may even be associated with a higher risk of severe social isolation in certain situations [27]. Similarly, some studies indicate that excessive internet use can lead to internet addiction, causing distress to individuals' psychological, social, and occupational well-being [28]. Other research also suggests that excessive internet use can lead to reduced rest time, which may harm the health of middle-aged and older adults [29].

A possible mediator between internet use and health association is information access. The health and wellness model supports this viewpoint, indicating that individuals with health consciousness can better understand their health status, including disease risks, preventative measures, and treatment options, through the acquisition of accurate health information [30,31]. This contributes to heightened health awareness and encourages people to engage in proactive health behaviors such as regular checkups, healthy eating, and exercise [32]. Additionally, information retrieval can provide psychological support and coping strategies, aiding in managing stress and anxiety, and thereby improving mental health [33,34]. In summary, information access plays a key role in maintaining and enhancing health conditions, facilitating individuals in effectively managing and preserving their health [35,36]. However, Chinese scholars have not yet given sufficient consideration to interpreting the influence of internet use on the health of middle-aged and older adults from an information perspective. Research on this issue is not only important for understanding the factors influencing the health of middle-aged and older adults but also for gaining insights into the socioeconomic consequences of internet use.

Similarly, the literature debates the relationship between internet use and social inequality. Some studies conclude that during the early and middle stages of internet adoption, advantaged groups with higher income, higher education, and more excellent technological proficiency are likelier to access the internet and enjoy its benefits [37-39]. Conversely, disadvantaged groups faced information disadvantages due to limitations in cognitive abilities and lower digital literacy, contributing to the digital divide [40,41]. However, other studies have concluded that the widespread availability of the internet has lowered the barriers to entry, increasing the possibility for all social groups, especially disadvantaged groups, to access information resources and technology, thus promoting upward mobility among society's marginalized populations [42,43]. According to the law of diminishing marginal utility, the benefits of internet use are limited in populations with substantial existing information resources. Instead, the benefits of the internet may be more concentrated among disadvantaged groups, leading to digital dividends [44,45]. So, whether the internet's effect on health brings about a digital divide or digital dividend and whether it narrows health inequalities between advantaged and disadvantaged groups or widens disparities among different populations needs to be empirically tested in this study.

There are several significant research gaps in the existing literature. First, although there is increasing interest in the potential influence of the internet on health [46], there is still a relative shortage of in-depth research explicitly targeting middle-aged and older adults in China. More comprehensive

XSL•FO
RenderX

studies are required to investigate the relationship between multidimensional internet use and multidimensional health. Furthermore, limited existing literature categorizes middle-aged and older adults into advantaged and disadvantaged groups based on multiple dimensions within the same sample, and investigates the relationship between internet use and their health. In the diverse social context of China, it is essential to conduct in-depth research to investigate whether there are differences in the influence of internet use on the health of advantaged and disadvantaged groups. Such research can help us better understand the potential of the internet to reduce or widen health inequalities, thus providing a scientific basis for formulating more equitable and effective public health policies. In summary, there is significant room for further exploration in studying internet use and its relationship with health among middle-aged and older adults in China. More in-depth and comprehensive exploration is needed to address the current research gaps in this field.

Objectives

Based on the analysis provided in the background section, this study aims to analyze the relationship between internet use and the multidimensional health of middle-aged and older adults using a nationally representative survey in China. Additionally, it investigates the relationship between the purposes and frequency of internet use and multidimensional health. Furthermore, this study extends the existing literature by examining the mediating effects of information access on the association between information use and multidimensional health. Moreover, the study will categorize middle-aged and older adults into advantaged and disadvantaged groups based on 4 criteria: income level, educational attainment, urban-rural classification, and region. Subsequently, this study will unveil potential heterogeneous influences for different subgroups within the advantaged and disadvantaged groups.

Methods

Data and Sample

The China General Social Survey (CGSS) is China's earliest, nationally representative, and comprehensive ongoing academic survey initiative [47]. The project uses a stratified, multistage probability proportional to size method to collect data, comprehensively gathering data at multiple levels including society, communities, households, and individuals. It covers 28 provinces, autonomous regions, and municipalities across China [48]. This study used nationwide survey data from CGSS2018. Following the age classification standards of the World Health Organization and considering the situation in China, individuals aged 45 years and older were defined as middle-aged and older adults [49]. Among the participants in CGSS2018, a total of 12,787 individuals were surveyed. After excluding 4344 individuals younger than 45 years, and additionally removing cases with missing data, a final valid sample of 6562 individuals was obtained. Figure 1 presents the specific data processing procedures.

Figure 1. Participant screening flowchart. CGSS: China General Social Survey.

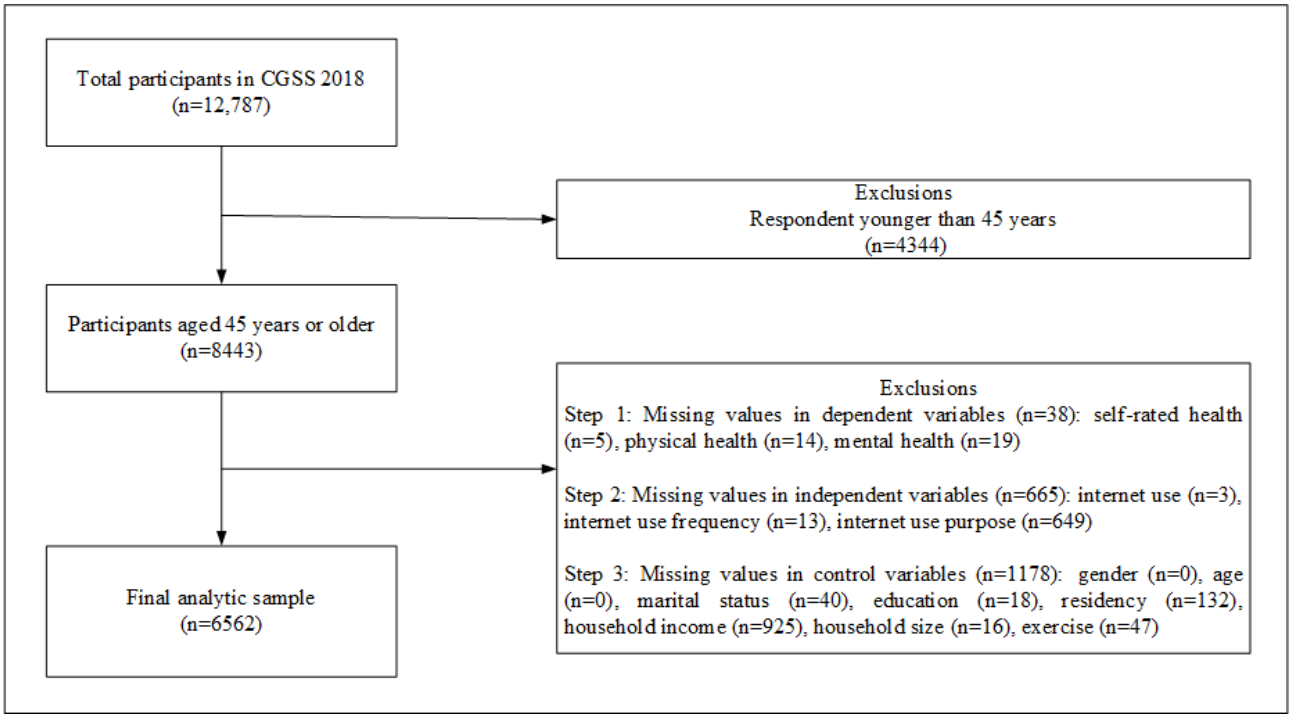

To examine whether internet use widens the digital divide between the advantaged and disadvantaged groups or brings more information welfare to the disadvantaged group, this study classified middle-aged and older adults into advantaged and disadvantaged groups based on 4 criteria. The criteria include income level, educational attainment, urban-rural classification, and region. The disadvantaged group refers to middle-aged and older adults with income below the median, an education level at or below junior high school, residing in rural areas, or living in western regions of China. Conversely, those not meeting these criteria are categorized as the advantaged group. This

XSL•FO
RenderX

differentiation between advantage and disadvantage is measured using objective criteria and has no discriminatory nature.

Ethical Considerations

CGSS was a survey approved by the Ethical Review Committee of Renmin University of China. The survey data was also anonymous, and the answers were protected by privacy law. Each participant provided signed informed consent at the time of participation. There was no requirement for additional ethics approval for approved data users.

Measures

Dependent Variable

Similar to prior research [50], health was measured using 3 indicators: self-rated health, physical health, and mental health.

Self-rated health, although simple, has been demonstrated to predict mortality and disability rates successfully [51]. Self-rated health ("How do you rate your current health?") was assessed using a 5-point scale (1=very unhealthy, 2=relatively unhealthy, 3=average, 4=relatively healthy, and 5=very healthy). The higher the value, the healthier.

Physical health ("In the past 4 weeks, how frequently has your work or other daily activities been affected by health problems?") was assessed using a 5-point scale (1=always, 2=often, 3=sometimes, 4=rarely, and 5=never), with higher values indicating better physical health.

Mental health ("How frequently have you felt symptoms of depression or sadness in the past 4 weeks?") is assessed using a 5-point scale (1=always, 2=often, 3=sometimes, 4=rarely, and 5=never), with higher values indicating better mental health status.

Independent Variables

The core independent variable in this study is internet use. Respondents were asked whether they use the internet (yes=1 and no=0). Additionally, the CGSS questionnaire includes questions regarding internet use frequency (never=1, rarely=2, sometimes=3, often=4, and very frequently=5), with higher values indicating more frequent internet use. Moreover, the CGSS questionnaire includes questions about the purposes of internet use, such as web-based payment (yes=1 and no=0), web-based socializing (yes=1 and no=0), web-based learning (yes=1 and no=0), and web browsing (yes=1 and no=0).

Mediating Variable

The mediating variable is information access. Information access was measured using the question "Is the internet (including mobile internet) your main source of information?" with "yes" coded as 1 and "no" as 0.

Control Variables

According to the health capital demand theory and existing research literature [52,53], this study controlled for individual sociodemographic and economic variables. These variables include gender (male=1 and female=0), age (continuous variable), educational level (illiterate=1, primary school or below=2, junior high school=3, high school=4, and college or above=5), marital status (married=1 and unmarried=0), residence (rural=0 and urban=1), household annual income (continuous variable), household size (continuous variable), and whether to exercise (yes=1 and no=0).

Instrumental Variables

We used 2 instrumental variables to address endogeneity issues. The first instrumental variable is the mean of the frequency of internet use by other community residents (excluding the respondent's internet use). The second instrumental variable is the broadband download rate of the respondents' province. Table S1 in Multimedia Appendix 1 presents the definition and coding of variables.

Statistical Analysis

All statistical analyses were conducted using Stata software (version 17.0; StataCorp LLC), with P values <.05 considered statistically significant (2-tailed).

We performed a descriptive analysis of the sample, summarizing continuous data using means and SDs, and presenting categorical data as frequencies and percentages. In the context of univariate analysis, we examined the relationship between internet use and health. The chi-square test was applied for categorical variables, while the Mann-Whitney U test was used for continuous variables.

Additionally, we used an ordinal logistic regression model to examine the odds ratios (ORs) and 95% CI for the association between internet use and health. Three models were constructed: model 1 included internet use as the independent variable and self-rated health, physical health, and mental health as the dependent variables, without any additional control variables. Model 2 incorporated sociodemographic variables such as gender, age, education, residency, marital status, and exercise. Model 3 introduced family variables, including family size and family income, on the basis of model 2. Models 2 and 3 both included provincial fixed effects to adjust for unobserved provincial-level factors.

To provide a comprehensive understanding of the research, we conducted the following analyses: first, we explored the relationship between internet use and health, the purposes of internet use and health, as well as the relationship between the frequency of internet use and health. All 3 sets of regressions used an ordinal logistic regression model. Second, we used the Karlson-Holm-Breen (KHB) decomposition method to examine whether information access mediated the relationship between internet use and health. Third, to address potential endogeneity problems, we introduced instrumental variable methods, specifically using a two-stage least squares regression. Fourth, middle-aged and older adults were categorized into advantaged and disadvantaged groups based on 4 criteria. Further analysis was conducted using grouped regression analysis for the advantaged and disadvantaged subgroups. Finally, it should be noted that all our regression analyses incorporated provincial fixed effects to control for provincial disparities.

XSL•FO
RenderX

Results

Descriptive Statistics

Table 1 summarizes the sociodemographic characteristics of the study participants (n=6562). From a demographic perspective, the average age of the participants was 61.41 (SD 10.69) years. The gender ratio between males and females (n=3392, 51.69% vs n=3170, 48.31%) was approximately balanced. Most participants (n=5342, 81.41%) had partners, while a minority (n=1220, 18.59%) were single. The educational level of the participants was relatively low (n=4875, 74.29%), with approximately three-quarters having yet to receive education beyond high school. Regarding geographical distribution, there were more participants from rural areas than urban areas (n=4067, 61.98% vs n=2495, 38.02%). In terms of family characteristics, the average household size was 2.55 (SD 1.33), indicating that at least two people lived together in each household, with few individuals living alone.

XSL•FO
RenderX

Table 1. Descriptive statistics of variables (n=6562).

Characteristics and variables	Total sample (n=6562)	Internet use (n=3036)	Internet nonuse (n=3526)	*P* value
Self-rated health, n (%)				<.001
Very unhealthy	301 (4.59)	59 (1.94)	242 (6.86)	
Relatively unhealthy	1372 (20.91)	391 (12.88)	981 (27.82)	
Average	1719 (26.20)	802 (26.42)	917 (26.01)	
Relatively healthy	2355 (35.89)	1282 (42.23)	1073 (30.43)	
Very healthy	815 (12.41)	502 (16.53)	313 (8.88)	
Physical health, n (%)				<.001
Very unhealthy	286 (4.36)	50 (1.65)	236 (6.69)	
Relatively unhealthy	834 (12.71)	200 (6.59)	634 (17.98)	
Average	1245 (18.97)	436 (14.36)	809 (22.94)	
Relatively healthy	2060 (31.39)	1094 (36.03)	966 (27.40)	
Very healthy	2137 (32.57)	1256 (41.37)	881 (24.99)	
Mental health, n (%)				<.001
Very unhealthy	85 (1.30)	23 (0.76)	62 (1.76)	
Relatively unhealthy	533 (8.12)	161 (5.30)	372 (10.55)	
Average	1561 (23.79)	594 (19.57)	967 (27.43)	
Relatively healthy	2495 (38.02)	1192 (39.26)	1303 (36.95)	
Very healthy	1888 (28.77)	1066 (35.11)	822 (23.31)	
Sex, n (%)				.003
Male	3392 (51.69)	1510 (49.74)	1882 (53.37)	
Female	3170 (48.31)	1526 (50.26)	1644 (46.63)	
Age (years), mean (SD)	61.41 (10.69)	56.92 (9.17)	65.27 (10.39)	<.001
Marital status, n (%)				<.001
Partnered	5342 (81.41)	2662 (87.68)	2680 (76.01)	
Single	1220 (18.59)	374 (12.32)	846 (23.99)	
Education, n (%)				<.001
Illiterate	1227 (18.70)	186 (6.13)	1.41 (29.52)	
≤Primary school	1813 (27.63)	501 (16.50)	1312 (37.21)	
Middle school	1835 (27.96)	1.38 (34.19)	797 (22.6)	
High school	1150 (17.53)	834 (27.47)	316 (8.96)	
≥College	537 (8.18)	477 (15.71)	60 (1.70)	
Residency, n (%)				<.001
Urban	2495 (38.02)	1657 (54.58)	838 (23.77)	
Rural	4067 (61.98)	1379 (45.42)	2688 (76.23)	
Exercise, n (%)				<.001
Yes	3122 (47.58)	1811 (59.65)	1311 (37.18)	
No	3440 (52.42)	1225 (40.35)	2215 (62.82)	
Family size, mean (SD)	2.55 (1.33)	2.58 (1.24)	2.53 (1.39)	<.001
Family income, mean (SD)	10.10 (2.19)	10.80 (1.66)	9.49 (2.40)	<.001

Regarding health status, among the participants, 35.89% (n=2355) considered their self-rated health relatively healthy, and 12.41% (n=815) considered it to be very healthy. In terms of physical health, 31.39% (n=2060) reported themselves as relatively healthy, while 32.57% (n=2137) reported themselves as very healthy. Concerning mental health, 38.02% (n=2495)

XSL•FO
RenderX

rated their mental health as relatively healthy, while 28.77% (n=1888) rated it as very healthy. Overall, the health status of middle-aged and older adults was relatively good.

Among participants who rated their self-rated health as unhealthy (including very unhealthy and relatively unhealthy), 450 (6.86%) individuals had used the internet, while 1223 (18.64%) individuals had not used the internet. In the case of participants who assessed their physical health as unhealthy (including very unhealthy and relatively unhealthy), 250 (3.80%) individuals had used the internet, while 870 (13.26%) individuals had not used the internet. Likewise, among participants who considered their mental health as unhealthy (including very unhealthy and relatively unhealthy), 184 (2.80%) individuals had used the internet, while 434 (6.61%) individuals had not used the internet. The descriptive statistics showed that middle-aged and older adults with poor self-rated, physical, and mental health are more likely to not use the internet (*P*<.001). Additionally, females (*P*<.001), older individuals (*P*<.001), those with less than a high school education (*P*<.001), and those without a partner (*P*<.001) are more inclined not to use the internet. Furthermore, individuals with lower household income (*P*<.001), smaller household size (*P*<.001), and those living in rural areas (*P*<.001) were also more likely to refrain from using the internet.

Table 2 displays the frequency and purpose of internet use. In the middle-aged and older adult population, internet use frequency is relatively low. Over half of the individuals (n=3526, 53.73%) have never used the internet, while only 15.33% (n=1006) use it often, and 13.2% (n=866) use it very frequently. A total of 27.52% (n=1806) of middle-aged and older adults use the internet for information access, 41.31% (n=2711) for web-based learning, 44.93% (n=2948) for web-based socializing, 17.16% (n=1126) for web-based payment, and 44.38% (n=2912) for web browsing.

Table 2. Internet use frequency and purpose by respondents in the sample (n=6562).

Variable	Participants, n (%)
Internet use frequency	
Never	3526 (53.73)
Rarely	542 (8.26)
Sometime	622 (9.48)
Often	1006 (15.33)
Very frequently	866 (13.20)
Internet use purpose (reference: yes)	
Information access	1806 (27.52)
Web-based learning	2711 (41.31)
Web-based socializing	2948 (44.93)
Web-based payment	1126 (17.16)
Web browsing	2912 (44.38)

Internet Use and Multidimensional Health

Table 3 presents the results of the regression analysis, illustrating the relationship between internet use and multidimensional health. Internet use demonstrated a significant positive association with self-rated health, physical health, and mental health. The findings indicated that middle-aged and older adults who used the internet were more likely to report good self-rated health compared to nonusers (OR 1.55, 95% CI 1.39-1.74; *P*<.001). Additionally, individuals who used the internet had significantly higher odds of better physical health than nonusers (OR 1.39, 95% CI 1.25-1.56; *P*<.001). Moreover, individuals who used the internet were more likely to have positive mental health compared to nonusers (OR 1.33, 95% CI 1.19-1.49; *P*<.001). This section presents only the results of model 3, while the regression results of models 1 and 2 can be found in Table S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Regression results for health among study participants (n=6562).

Variable	Self-rated health, OR[a] (95% CI)	*P* value	Physical health, OR (95% CI)	*P* value	Mental health, OR (95% CI)	*P* value
Internet use						
Yes (reference: no)	1.55 (1.39-1.74)	<.001	1.39 (1.25-1.56)	<.001	1.33 (1.19-1.49)	<.001
Sex						
Male (reference: female)	1.28 (1.17-1.41)	<.001	1.26 (1.14-1.38)	<.001	1.29 (1.18-1.42)	<.001
Age	0.98 (0.97-0.98)	<.001	0.98 (0.97-0.98)	<.001	1.01 (1.00-1.01)	.009
Marital status						
Married (reference: unmarried)	1.07 (0.95-1.21)	.27	1.16 (1.02-1.31)	.02	1.26 (1.11-1.43)	<.001
Education						
≤Primary school (reference: illiterate)	1.06 (0.93-1.22)	.38	1.05 (0.92-1.20)	<.001	1.08 (0.94-1.23)	.29
Middle school (reference: illiterate)	1.18 (1.01-1.37)	.04	1.35 (1.16-1.57)	.51	1.28 (1.11-1.49)	.001
High school (reference: illiterate)	1.37 (1.15-1.63)	<.001	1.56 (1.31-1.86)	<.001	1.34 (1.12-1.60)	.001
≥College (reference: illiterate)	1.37 (1.10-1.71)	.006	1.60 (1.28-2.01)	<.001	1.21 (0.97-1.52)	.10
Exercise						
Yes (reference: no)	1.57 (1.43-1.73)	<.001	1.52 (1.38-1.68)	<.001	1.36 (1.24-1.50)	<.001
Residency						
Urban (reference: rural)	0.85 (0.75-0.96)	.009	1.02 (0.90-1.15)	.80	1.11 (0.98-1.25)	.09
Family size	1.02 (0.99-1.06)	.24	1.03 (0.99-1.07)	.14	1.02 (0.98-1.06)	.28
Family income	1.09 (1.06-1.11)	<.001	1.06 (1.04-1.09)	<.001	1.05 (0.02-1.07)	<.001

[a]OR: odds ratio.

Frequency and Purpose of Internet Use and Health

Table 4 displays the associations between the frequency and purpose of internet use and health. From the analysis results, it becomes evident that the higher the frequency of internet use among middle-aged and older adults, the better their self-rated health (very frequently: OR 1.48, 95% CI 1.26-1.74; *P*<.001), physical health (very frequently: OR 1.76, 95% CI 1.48-2.08; *P*<.001), and mental health (very frequently: OR 1.52, 95% CI 1.28-1.79; *P*<.001) tend to be. Regarding the purposes of internet use, except for web-based payment (OR 1.08, 95% CI 0.94-1.24; *P*=.30), which had no relationship on the physical health of middle-aged and older adults, all other purposes of internet use showed significant associations with health. Compared to the control group, middle-aged and older adults who used the internet showed improvements in self-rated, physical, and mental health.

XSL•FO
RenderX

Table 4. Associations between health and internet use frequency and purpose (n=6562).

Independent variables: internet use	Self-rated health, OR[a] (95% CI)	*P* value	Physical health, OR (95% CI)	*P* value	Mental health, OR (95% CI)	*P* value
Frequency (reference: never)						
Rarely	1.64 (1.38-1.94)	<.001	1.25 (1.05-1.47)	.01	1.27 (1.07-1.51)	.006
Sometimes	1.41 (1.19-1.66)	<.001	1.28 (1.08-1.51)	<.004	1.20 (1.02-1.42)	.03
Often	1.67 (1.44-1.94)	<.001	1.44 (1.23-1.67)	<.001	1.39 (1.19-1.62)	<.001
Very frequently	1.48 (1.26-1.74)	<.001	1.76 (1.48-2.08)	<.001	1.52 (1.28-1.79)	<.001
Purpose (reference: no)						
Web-based payment	1.24 (1.08-1.42)	.002	1.08 (0.94-1.24)	.30	1.22 (1.06-1.41)	.005
Web-based learning	1.49 (1.33-1.67)	<.001	1.41 (1.25-1.58)	<.001	1.29 (1.15-1.45)	<.001
Web-based socializing	1.51 (1.34-1.70)	<.001	1.58 (1.41-1.78)	<.001	1.48 (1.3-1.66)	<.001
Web browsing	1.40 (1.25-1.57)	<.001	1.39 (1.24-1.55)	<.001	1.29 (1.15-1.45)	<.001

[a]OR: odds ratio.

Mediation Effect of Information Access

Table 5 reports the mediation effect of information access using the KHB decomposition method. The results revealed that information access had a significant mediating effect (β=.28, 95% CI 0.23-0.32) on the relationship between internet use and self-rated health. Additionally, information access also had a significant mediating effect (β=.40, 95% CI 0.35-0.45) on the relationship between internet use and physical health. Moreover, the KHB results showed that the indirect effect of information access on internet use and mental health was significant (β=.16, 95% CI 0.11-0.20). Therefore, the mediating effect of information access on internet use and each dimension of health among middle-aged and older adults was significant.

Table 5. Mediation of the relationship between middle-aged and older adults' health and internet use by information access (n=6562).

Mediation by information access	Self-rated health, β (95% CI)	*P* value	Physical health, β (95% CI)	*P* value	Mental health, β (95% CI)	*P* value
Total effect	.52 (0.47-0.57)	<.001	.63 (0.58-0.68)	<.001	.33 (0.29-0.38)	<.001
Direct effect	.24 (0.18-0.31)	<.001	.23 (0.16-0.30)	<.001	.18 (0.12-0.24)	<.001
Indirect effect	.28 (0.23-0.32)	<.001	.40 (0.35-0.45)	<.001	.16 (0.11-0.20)	<.001

Treatment: Instrumental Variables Approach

Internet use as a form of individual decision-making may lead to endogeneity issues due to reverse causality or omitted variables. While significant efforts have been made in this paper to carefully select control variables to mitigate omitted variable issues, endogeneity concerns may still exist. To address this issue, the study sought instrumental variables for the internet use of middle-aged and older adults to mitigate potential bias and nonconsistency resulting from endogeneity. Ultimately, 2 instrumental variables were chosen, and a two-stage estimation was performed using the product of these two instrumental variables as instruments. The first instrumental variable is the mean internet use frequency among residents in the respondent's community (excluding the respondent's internet use). The second instrumental variable is the broadband download rate of the province. Average bandwidth data for each province is sourced from the 19th edition of the "China Broadband Speed Report," published by the Broadband Development Alliance.

Table S3 in Multimedia Appendix 1 displays the two-stage least squares regression. In the first-stage regression of the endogenous variables, the instrumental variables were statistically significant at the 1% significance level. The results from the second stage of the regression still suggested that individuals who used the internet had a higher likelihood of better self-rated health (β=.31, 95% CI 0.03-0.59; *P*=.03, better physical health (β=.55, 95% CI 0.25-0.86; *P*<.001), and better mental health (β=.33, 95% CI 0.08-0.59; *P*=.01) compared to nonusers. This suggested that the conclusions drawn earlier remained robust.

Heterogeneity Analysis

In existing research [41,45], some scholars argue that the internet may create a digital divide between advantaged and disadvantaged groups, while others believe it may bring about digital dividends. To address this controversy, this study categorizes middle-aged and older adults into advantaged and disadvantaged groups with 4 criteria: income level, educational attainment, urban-rural classification, and region.

Table S4 in Multimedia Appendix 1 demonstrates that internet use had a significant facilitative effect on self-rated health (urban: OR 1.37, 95% CI 1.34-1.65; *P*=.001 vs rural: OR 1.67, 95% CI 1.45-1.93; *P*<.001), mental health (urban: OR 1.20, 95% CI 0.99-1.45; *P*=.06 vs rural: OR 1.40, 95% CI 1.22-1.62; *P*<.001) for both rural and urban samples, as well as physical health of the rural group (OR 1.57, 95% CI 1.36-1.81; *P*<.001). However, it had no significant effect on the physical health of the urban group (OR 1.12, 95% CI 0.93-1.36; *P*=.24). Therefore,

XSL•FO
RenderX

for those in rural areas, the health effects of internet use were much larger.

Table S5 in Multimedia Appendix 1 reveals that the association between internet use and self-rated health (eastern: OR 1.45, 95% CI 1.24-1.69; *P*<.001 vs western: OR 2.03, 95% CI 1.60-2.57; *P*<.001), physical health (eastern: OR 1.21, 95% CI 1.03-1.41; *P*=.02 vs western: OR 2.14, 95% CI 1.69-2.71; *P*<.001), and mental health (eastern: OR 1.40, 95% CI 1.20-1.64; *P*<.001 vs western: OR 1.29, 95% CI 1.02-1.63; *P*=.04), for both western and eastern region samples were also significantly positive. However, for those in western areas, the effect was much larger.

Table S6 in Multimedia Appendix 1 displays that the association between internet use and self-rated health (high-income: OR 1.37, 95% CI 1.17-1.61; *P*<.001 vs low-income: OR 1.74, 95% CI 1.47-2.05; *P*<.001), physical health (high-income: OR 1.17, 95% CI 1.00-1.38; *P*=.05 vs low-income: OR 1.61, 95% CI 1.37-1.90; *P*<.001), and mental health (high-income: OR 1.16, 95% CI 0.98-1.36; *P*=.08 vs low-income: OR 1.49, 95% CI 1.26-1.76; *P*<.001) for both high-income and low-income samples were also significantly positive. However, for those in the low-income group, the effect was much larger.

Finally, as indicated in Table S7 in Multimedia Appendix 1, the association between internet use and self-rated health (high education: OR 1.56, 95% CI 1.34-1.80; *P*<.001 vs low education: OR 1.61, 95% CI 1.36-1.92; *P*<.001), physical health (high education: OR 1.49, 95% CI 1.28-1.73; *P*<.001 vs low-education: OR 1.29, 95% CI 1.09-1.53; *P*=.004), and mental health (high education: OR 1.35, 95% CI 1.17-1.57; *P*<.001 vs low education: OR 1.27, 95% CI 1.07-1.52; *P*=.007) for both high education and low-education samples were also significantly positive.

Discussion

Principal Findings

This study used data from China's first nationwide publicly available database CGSS to examine the relationship between internet use and multidimensional health in middle-aged and older adults. We initially examined the significant relationships between internet use and self-rated, physical, and mental health. Subsequently, we conducted further analyses to explore the links between the purposes and frequency of internet use and multidimensional health. Then, we investigated the partial mediating role of information access in the relationship between internet use and health. Finally, we explored whether there were significant differences in internet use and health among advantaged and disadvantaged groups.

Our research confirmed that internet use was currently influencing the daily lives of middle-aged and older adults: 46.3% (n=3036) of middle-aged and older adults use the internet in our sample. This indicated that over half of the older people in our sample had not yet accessed the internet and were unable to benefit from the advancements in digital technology. This was still a significant gap compared to the internet penetration rates among older people in other countries, which ranged from 76% to 82% [54,55]. The reasons behind this phenomenon are multifaceted. First, the availability of smart devices and services for older adults is not yet comprehensive [56], and their digital skills are generally inadequate [57], resulting in fundamental operational obstacles during their use [58]. Second, the ownership of internet access devices among older adults is lower than younger generations [59]. Third, older adults have weaker information processing abilities and lower discernment skills, making them more susceptible to false information and telecommunications fraud [60]. These factors make some older adults unwilling to use the internet.

In this study, it was found that internet use was positively associated with the health of middle-aged and older adults. This contribution addressed the ongoing debate in the literature [18,22], as other surveys presented contradictory results. This could be related to the results of other studies [61] that the internet enables people to access health-related knowledge, gradually develop healthy lifestyle habits, and improve overall health. The internet provides middle-aged and older adults access to more health resources and convenient methods for managing their health, which positively affects their overall health [62,63]. For instance, middle-aged and older adults can use health apps and digital tools for medication management, health data recording, and health trend tracking, facilitating better self-management of health-related issues [64]. Furthermore, the internet offers crucial support for mental health, allowing middle-aged and older adults to access emotional support and resources to address psychological concerns through digital therapy [65]. Additionally, individuals' internet use is associated with better financial and health care decision-making [66,67]. People who use the internet may have more financial resources, which can contribute to their health [38].

We found that information access significantly mediated the relationship between internet use and each dimension of health. The internet provides middle-aged and older adults access to a wealth of health-related information, from disease prevention to treatment recommendations, to understand better and manage their health [68,69]. They can enhance their health knowledge through web-based resources and adopt healthier lifestyles such as regular exercise and balanced diets [70]. Furthermore, information access reduces uncertainty in medical decision-making as individuals consult health care professionals through web-based platforms to seek treatment advice [71]. In summary, the information access feature of the internet enhances middle-aged and older adults' understanding and management of their health, granting them greater autonomy and control over their well-being, ultimately resulting in a positive influence on overall health [72].

This study also confirmed that there were significant differences in the relationship between internet use and the health of middle-aged and older adults among advantaged and disadvantaged groups. Specifically, for those in rural and western areas, the effect was much larger. As pointed out in previous research [73], China faces an imbalance in the allocation of health care resources among regions, with a concentration of high-quality medical resources in urban and eastern regions [74]. In contrast, rural and western regions have relatively limited health care resources, resulting in insufficient

XSL•FO
RenderX

accessibility to health care services [75]. Therefore, middle-aged and older adults in rural and western regions are more likely to improve their health through internet use [76]. This study also indicated that the association between internet use and the health of middle-aged and older adults was significantly positive for both higher-income and lower-income samples. However, the effect was greater for those in the lower-income group. Low-income disadvantaged group often faces challenges in accessing health care services than advantaged groups [77]. However, the internet provides them with abundant health information resources, including web-based consultations and medical advice, which help alleviate health care access issues linked to economic constraints [78]. In summary, internet use brings more benefits to disadvantaged groups [79]. This conclusion provides essential insights into how to mitigate health inequalities among advantaged and disadvantaged groups.

Contributions

This study makes significant contributions to the existing literature from 4 perspectives. First, unlike previous research that often examined the relationship between internet use and health from a single perspective, we adopt an innovative research approach. We categorize internet use into 3 dimensions (whether to use the internet, internet use frequency, and internet use purpose) to provide a multidimensional consideration of internet use. Second, the statistical analysis in existing studies often overlooks endogeneity issues and sample selection bias. This paper addresses these concerns by selecting broadband download rates and internet use frequency among community residents as instrumental variables to ensure the reliability of the research results. Third, our findings provide empirical evidence for the first time, demonstrating that information access serves as a crucial mediating variable in the relationship between internet use and the health of middle-aged and older adults. This not only supports the theoretical proposition that internet use improves the health status of middle-aged and older adults but also advocates for the promotion of active aging policies. Fourth, to our knowledge, limited literature categorizes middle-aged and older adults into advantaged and disadvantaged groups based on multiple criteria and investigates the relationship between internet use and multidimensional health. Our results emphasize that internet use tends to benefit disadvantaged groups more. These insights are of significant importance and provide evidence for narrowing the health and welfare gap between different population groups.

Limitations

Certainly, our study acknowledges specific limitations. First, due to data constraints, we have to rely on the question “How do you rate your current physical health?” to measure physical health. This may not comprehensively measure their health status. In future research, we will select more comprehensive physical health indicators, such as activities of daily living and instrumental activities of daily living, to better reflect the health status of individuals. Second, given the complexity of the relationship between internet use and health, future research should consider longitudinal analysis to understand better the causal relationships and temporal dynamics. Nonetheless, our research offers valuable evidence for low-income countries seeking a deeper understanding of the relationship between internet use and health.

Conclusions

In this study, it was found that internet use was positively associated with better self-rated health, physical health, and mental health after controlling for key determinants of health. Additionally, information access mediated the relationship between internet use and the health of middle-aged and older adults in China. Furthermore, there were significant differences in the relationship between internet use and health among advantaged and disadvantaged groups. This study suggests collaborative efforts between the government and the market to further enhance internet accessibility for the older people population in China. Additionally, society should strive to lower the technological and income barriers for middle-aged and older adults in using the internet so that more seniors can enjoy the digital dividends brought by the internet.

Acknowledgments

This study was funded by the Provincial Philosophy and Social Science Project of Qinghai Province (23ZCQ089), and the Qinghai Minzu University Foundation (39D2023003). We thank all members in LF’s laboratory at Tianjin University for their assistance with the manuscript review.

Data Availability

The data sets generated during or analyzed during this study are available in the China General Social Survey repository [80].

Authors' Contributions

LF was responsible for the study concept, and supervision. CL was responsible for the study design, data analysis, interpretation of the data, and writing of the manuscript. YD and XM were responsible for supervision. QC, DL, and KD reviewed the manuscript. All authors have read and approved the published version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Additional tables for the study.
[DOCX File , 69 KB - jmir_v26i1e49688_app1.docx]

References

1. Vollset SE, Goren E, Yuan CW, Cao J, Smith AE, Hsiao T, et al. Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study. Lancet 2020;396(10258):1285-1306 [FREE Full text] [doi: 10.1016/S0140-6736(20)30677-2] [Medline: 32679112]
2. Cesari M, Sumi Y, Han ZA, Perracini M, Jang H, Briggs A, et al. Implementing care for healthy ageing. BMJ Glob Health 2022;7(2):e007778 [FREE Full text] [doi: 10.1136/bmjgh-2021-007778] [Medline: 35185014]
3. Abud T, Kounidas G, Martin KR, Werth M, Cooper K, Myint PK. Determinants of healthy ageing: a systematic review of contemporary literature. Aging Clin Exp Res 2022;34(6):1215-1223 [FREE Full text] [doi: 10.1007/s40520-021-02049-w] [Medline: 35132578]
4. Digital 2022: Global Overview Report. We are Social & Hootsuite. 2022. URL: https://datareportal.com/reports/digital-2022-global-overview-report [accessed 2024-05-03]
5. The 51st statistical report on internet development in China. China Internet Network Information Center. 2023. URL: https://www.cnnic.cn/NMediaFile/2023/0807/MAIN169137187130308PEDV637M.pdf [accessed 2024-05-03]
6. Buntin MB, Burke MF, Hoaglin MC, Blumenthal D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Aff (Millwood) 2011;30(3):464-471. [doi: 10.1377/hlthaff.2011.0178] [Medline: 21383365]
7. McDool E, Powell P, Roberts J, Taylor K. The internet and children's psychological wellbeing. J Health Econ 2020;69:102274 [FREE Full text] [doi: 10.1016/j.jhealeco.2019.102274] [Medline: 31887480]
8. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 2;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
9. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
10. Wald HS, Dube CE, Anthony DC. Untangling the web—the impact of internet use on health care and the physician-patient relationship. Patient Educ Couns 2007;68(3):218-224. [doi: 10.1016/j.pec.2007.05.016] [Medline: 17920226]
11. Allcott H, Braghieri L, Eichmeyer S, Gentzkow M. The welfare effects of social media. Am Econ Rev 2020;110(3):629-676. [doi: 10.1257/aer.20190658]
12. Duplaga M, Szulc K. The association between internet use and health outcomes in the population of older adults. Eur J Public Health 2020:165. [doi: 10.1093/eurpub/ckaa165.443]
13. Amini R, Chee KH, Mendieta M, Parker S. Online engagement and cognitive function among older adults. Geriatr Gerontol Int 2019;19(9):918-923 [FREE Full text] [doi: 10.1111/ggi.13749] [Medline: 31368165]
14. Burke M, Kraut RE. The relationship between Facebook use and well-being depends on communication type and tie strength. J Comput-Mediat Comm 2016;21(4):265-281. [doi: 10.1111/jcc4.12162]
15. Schrempft S, Jackowska M, Hamer M, Steptoe A. Associations between social isolation, loneliness, and objective physical activity in older men and women. BMC Public Health 2019;19(1):74 [FREE Full text] [doi: 10.1186/s12889-019-6424-y] [Medline: 30651092]
16. Wu HY, Chiou AF. Social media usage, social support, intergenerational relationships, and depressive symptoms among older adults. Geriatr Nurs 2020;41(5):615-621. [doi: 10.1016/j.gerinurse.2020.03.016] [Medline: 32268948]
17. Cruz J, Brooks D, Marques A. Home telemonitoring in COPD: a systematic review of methodologies and patients' adherence. Int J Med Inform 2014;83(4):249-263. [doi: 10.1016/j.ijmedinf.2014.01.008] [Medline: 24529402]
18. Duplaga M. The association between internet use and health-related outcomes in older adults and the elderly: a cross-sectional study. BMC Med Inform Decis Mak 2021;21(1):150 [FREE Full text] [doi: 10.1186/s12911-021-01500-2] [Medline: 33957909]
19. Bekalu MA, McCloud RF, Viswanath K. Association of social media use with social well-being, positive mental health, and self-rated health: disentangling routine use from emotional connection to use. Health Educ Behav 2019;46(2_suppl):69-80. [doi: 10.1177/1090198119863768] [Medline: 31742462]
20. Kamal SA, Shafiq M, Kakria P. Investigating acceptance of telemedicine services through an extended technology acceptance model (TAM). Technol Soc 2020;60:101212. [doi: 10.1016/j.techsoc.2019.101212]
21. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res 2016;16(1):442 [FREE Full text] [doi: 10.1186/s12913-016-1691-0] [Medline: 27562728]
22. Braghieri L, Levy R, Makarin A. Social media and mental health. Am Econ Rev 2022;112(11):3660-3693. [doi: 10.1257/aer.20211218]

XSL•FO
RenderX

23. Huang C. Internet use and psychological well-being: a meta-analysis. Cyberpsychol Behav Soc Netw 2010;13(3):241-249. [doi: 10.1089/cyber.2009.0217] [Medline: 20557242]
24. Błachnio A, Przepiórka A, Pantic I. Internet use, Facebook intrusion, and depression: results of a cross-sectional study. Eur Psychiatry 2015;30(6):681-684. [doi: 10.1016/j.eurpsy.2015.04.002] [Medline: 25963476]
25. Nie NH. Sociability, interpersonal relations, and the internet: reconciling conflicting findings. Am Behav Sci 2016;45(3):420-435. [doi: 10.1177/00027640121957277]
26. Stepanikova I, Nie NH, He X. Time on the internet at home, loneliness, and life satisfaction: evidence from panel time-diary data. Comput Hum Behav 2010;26(3):329-338. [doi: 10.1016/j.chb.2009.11.002]
27. Meshi D, Cotten SR, Bender AR. Problematic social media use and perceived social isolation in older adults: a cross-sectional study. Gerontology 2020;66(2):160-168. [doi: 10.1159/000502577] [Medline: 31522178]
28. Tavares AI. Self-assessed health among older people in Europe and internet use. Int J Med Inform 2020 Sep;141:104240 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104240] [Medline: 32739610]
29. Kokka I, Mourikis I, Nicolaides NC, Darviri C, Chrousos GP, Kanaka-Gantenbein C, et al. Exploring the effects of problematic internet use on adolescent sleep: a systematic review. Int J Environ Res Public Health 2021;18(2):760 [FREE Full text] [doi: 10.3390/ijerph18020760] [Medline: 33477410]
30. McCloud RF, Okechukwu CA, Sorensen G, Viswanath K. Beyond access: barriers to internet health information seeking among the urban poor. J Am Med Inform Assoc 2016;23(6):1053-1059 [FREE Full text] [doi: 10.1093/jamia/ocv204] [Medline: 27206459]
31. Pandey SK, Hart JJ, Tiwary S. Women's health and the internet: understanding emerging trends and implications. Soc Sci Med 2003;56(1):179-191. [doi: 10.1016/s0277-9536(02)00019-9] [Medline: 12435560]
32. McCully SN, Don BP, Updegraff JA. Using the internet to help with diet, weight, and physical activity: results from the health information national trends survey (HINTS). J Med Internet Res 2013;15(8):e148 [FREE Full text] [doi: 10.2196/jmir.2612] [Medline: 23906945]
33. Chopik WJ. The benefits of social technology use among older adults are mediated by reduced loneliness. Cyberpsychol Behav Soc Netw 2016;19(9):551-556 [FREE Full text] [doi: 10.1089/cyber.2016.0151] [Medline: 27541746]
34. Magsamen-Conrad K, Dillon JM, Billotte Verhoff C, Faulkner SL. Online health-information seeking among older populations: family influences and the role of the medical professional. Health Commun 2019;34(8):859-871 [FREE Full text] [doi: 10.1080/10410236.2018.1439265] [Medline: 29474125]
35. Jiang S, Street RL. Pathway linking internet health information seeking to better health: a moderated mediation study. Health Commun 2017;32(8):1024-1031. [doi: 10.1080/10410236.2016.1196514] [Medline: 27464036]
36. Rowlands IJ, Loxton D, Dobson A, Mishra GD. Seeking health information online: association with young Australian women's physical, mental, and reproductive health. J Med Internet Res 2015;17(5):e120 [FREE Full text] [doi: 10.2196/jmir.4048] [Medline: 25986630]
37. Oyedemi TD. Digital inequalities and implications for social inequalities: a study of internet penetration amongst university students in South Africa. Telemat Inf 2012;29(3):302-313. [doi: 10.1016/j.tele.2011.12.001]
38. Van DAJAM, Helsper E. The third-level digital divide: who benefits most from being online? Commun Inf Technol 2015;10:29-52. [doi: 10.1108/s2050-206020150000010002]
39. Vassilakopoulou P, Hustad E. Bridging digital divides: a literature review and research agenda for information systems research. Inf Syst Front 2023;25(3):955-969 [FREE Full text] [doi: 10.1007/s10796-020-10096-3] [Medline: 33424421]
40. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. Lancet 2013;381(9868):752-762 [FREE Full text] [doi: 10.1016/S0140-6736(12)62167-9] [Medline: 23395245]
41. Bauer JM. The internet and income inequality: socio-economic challenges in a hyperconnected society. Telecommun Policy 2018;42(4):333-343. [doi: 10.1016/j.telpol.2017.05.009]
42. Mesch G, Mano R, Tsamir J. Minority status and health information search: a test of the social diversification hypothesis. Soc Sci Med 2012;75(5):854-858. [doi: 10.1016/j.socscimed.2012.03.024] [Medline: 22633160]
43. van Deursen AJ, van Dijk JAGM, Peters O. Rethinking internet skills: the contribution of gender, age, education, internet experience, and hours online to medium-and content-related internet skills. Poetics 2011;39(2):125-144. [doi: 10.1016/j.poetic.2011.02.001]
44. Chen L, Liu W. The effect of internet access on body weight: evidence from China. J Health Econ 2022;85:102670. [doi: 10.1016/j.jhealeco.2022.102670] [Medline: 36055079]
45. Lustria MLA, Smith SA, Hinnant CC. Exploring digital divides: an examination of eHealth technology use in health information seeking, communication and personal health information management in the USA. Health Inf J 2011;17(3):224-243 [FREE Full text] [doi: 10.1177/1460458211414843] [Medline: 21937464]
46. Hunsaker A, Hargittai E. A review of internet use among older adults. New Media Soc 2018;20(10):3937-3954. [doi: 10.1177/1461444818787348]
47. Chyi H, Mao S. The determinants of happiness of China‘s elderly population. J Happiness Stud 2011;13(1):167-185. [doi: 10.1007/s10902-011-9256-8]
48. Wang J, Wang C, Li S, Luo Z. Measurement of relative welfare poverty and its impact on happiness in China: evidence from CGSS. China Economic Review 2021;69:101687. [doi: 10.1016/j.chieco.2021.101687]

XSL•FO
RenderX

49. World Report on Ageing and Health. Geneva, Switzerland: World Health Organization; 2015:246.
50. Boardman JD. Self-rated health among U.S. adolescents. J Adolesc Health 2006;38(4):401-408 [FREE Full text] [doi: 10.1016/j.jadohealth.2005.01.006] [Medline: 16549301]
51. Franks P, Gold MR, Fiscella K. Sociodemographics, self-rated health, and mortality in the US. Soc Sci Med 2003;56(12):2505-2514. [doi: 10.1016/s0277-9536(02)00281-2] [Medline: 12742613]
52. Cotten SR, Ford G, Ford S, Hale TM. Internet use and depression among retired older adults in the United States: a longitudinal analysis. J Gerontol B Psychol Sci Soc Sci 2014;69(5):763-771. [doi: 10.1093/geronb/gbu018] [Medline: 24671896]
53. Leibowitz AA. The demand for health and health concerns after 30 years. J Health Econ 2004;23(4):663-671. [doi: 10.1016/j.jhealeco.2004.04.005] [Medline: 15587693]
54. Crouch E, Gordon NP. Prevalence and factors influencing use of internet and electronic health resources by middle-aged and older adults in a US health plan population: cross-sectional survey study. JMIR Aging 2019;2(1):e11451 [FREE Full text] [doi: 10.2196/11451] [Medline: 31518256]
55. Zajac IT, Flight IHK, Wilson C, Turnbull D, Cole S, Young G. Internet usage and openness to internet-delivered health information among Australian adults aged over 50 years. Australas Med J 2012;5(5):262-267 [FREE Full text] [doi: 10.4066/AMJ.2012.1065] [Medline: 22848321]
56. Hong A, Nam C, Kim S. What will be the possible barriers to consumers' adoption of smart home services? Telecommun Policy 2020;44(2):101867. [doi: 10.1016/j.telpol.2019.101867]
57. Nimrod G. Changes in internet use when coping with stress: older adults during the COVID-19 pandemic. Am J Geriatr Psychiatry 2020;28(10):1020-1024 [FREE Full text] [doi: 10.1016/j.jagp.2020.07.010] [Medline: 32771312]
58. Beaunoyer E, Dupéré S, Guitton MJ. COVID-19 and digital inequalities: reciprocal impacts and mitigation strategies. Comput Human Behav 2020 Oct;111:106424 [FREE Full text] [doi: 10.1016/j.chb.2020.106424] [Medline: 32398890]
59. Seifert A, Cotten SR, Xie B. A double burden of exclusion? Digital and social exclusion of older adults in times of COVID-19. J Gerontol B Psychol Sci Soc Sci 2021;76(3):e99-e103 [FREE Full text] [doi: 10.1093/geronb/gbaa098] [Medline: 32672332]
60. Seo H, Blomberg M, Altschwager D, Vu HT. Vulnerable populations and misinformation: a mixed-methods approach to underserved older adults' online information assessment. New Media Soc 2020;23(7):2012-2033 [FREE Full text] [doi: 10.1177/1461444820925041]
61. Arcury TA, Sandberg JC, Melius KP, Quandt SA, Leng X, Latulipe C, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020;39(2):141-150 [FREE Full text] [doi: 10.1177/0733464818807468] [Medline: 30353776]
62. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015;17(2):e52 [FREE Full text] [doi: 10.2196/jmir.3951] [Medline: 25803266]
63. Al Bassam N, Hussain SA, Al Qaraghuli A, Khan J, Sumesh EP, Lavanya V. IoT based wearable device to monitor the signs of quarantined remote patients of COVID-19. Inform Med Unlocked 2021;24:100588 [FREE Full text] [doi: 10.1016/j.imu.2021.100588] [Medline: 33997262]
64. Giustini D, Ali SM, Fraser M, Boulos MNK. Effective uses of social media in public health and medicine: a systematic review of systematic reviews. Online J Public Health Inform 2018;10(2):e215 [FREE Full text] [doi: 10.5210/ojphi.v10i2.8270] [Medline: 30349633]
65. Forsman AK, Nordmyr J. Psychosocial links between internet use and mental health in later life: a systematic review of quantitative and qualitative evidence. J Appl Gerontol 2017;36(12):1471-1518. [doi: 10.1177/0733464815595509] [Medline: 26245208]
66. Schilbach F, Schofield H, Mullainathan S. The psychological lives of the poor. Am Econ Rev 2016;106(5):435-440. [doi: 10.1257/aer.p20161101] [Medline: 29547249]
67. James BD, Boyle PA, Yu L, Bennett DA. Internet use and decision making in community-based older adults. Front Psychol 2013;4:605 [FREE Full text] [doi: 10.3389/fpsyg.2013.00605] [Medline: 24578696]
68. Deng Z, Mo X, Liu S. Comparison of the middle-aged and older users' adoption of mobile health services in China. Int J Med Inform 2014;83(3):210-224. [doi: 10.1016/j.ijmedinf.2013.12.002] [Medline: 24388129]
69. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics J 2015;21(3):173-194 [FREE Full text] [doi: 10.1177/1460458213512220] [Medline: 24670899]
70. Leung L, Chen C. E-health/m-health adoption and lifestyle improvements: exploring the roles of technology readiness, the expectation-confirmation model, and health-related information activities. Telecommun Policy 2019;43(6):563-575. [doi: 10.1016/j.telpol.2019.01.005]
71. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
72. Fischer SH, David D, Crotty BH, Dierks M, Safran C. Acceptance and use of health information technology by community-dwelling elders. Int J Med Inform 2014;83(9):624-635 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.005] [Medline: 24996581]
73. Tang S, Meng Q, Chen L, Bekedam H, Evans T, Whitehead M. Tackling the challenges to health equity in China. Lancet 2008;372(9648):1493-1501 [FREE Full text] [doi: 10.1016/S0140-6736(08)61364-1] [Medline: 18930531]

XSL•FO
RenderX

74. Evandrou M, Falkingham J, Feng Z, Vlachantoni A. Individual and province inequalities in health among older people in China: evidence and policy implications. Health Place 2014;30:134-144 [FREE Full text] [doi: 10.1016/j.healthplace.2014.08.009] [Medline: 25262491]
75. Wang Y, Li Y, Qin S, Kong Y, Yu X, Guo K, et al. The disequilibrium in the distribution of the primary health workforce among eight economic regions and between rural and urban areas in China. Int J Equity Health 2020;19(1):28 [FREE Full text] [doi: 10.1186/s12939-020-1139-3] [Medline: 32102655]
76. Berg T, Winterton R, Petersen M, Warburton J. 'Although we're isolated, we're not really isolated': the value of information and communication technology for older people in rural Australia. Australas J Ageing 2017;36(4):313-317. [doi: 10.1111/ajag.12449] [Medline: 28639293]
77. Jacobs B, Ir P, Bigdeli M, Annear PL, Van Damme W. Addressing access barriers to health services: an analytical framework for selecting appropriate interventions in low-income Asian countries. Health Policy Plan 2012;27(4):288-300. [doi: 10.1093/heapol/czr038] [Medline: 21565939]
78. Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unützer J, Patel V, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. Lancet Psychiatry 2017;4(6):486-500 [FREE Full text] [doi: 10.1016/S2215-0366(17)30096-2] [Medline: 28433615]
79. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
80. Chinese General Social Survey. URL: http://cgss.ruc.edu.cn/ [accessed 2024-08-26]

Abbreviations

CGSS: China General Social Survey
KHB: Karlson-Holm-Breen
OR: odds ratio

Edited by T de Azevedo Cardoso; submitted 06.06.23; peer-reviewed by E Brainin, W He; comments to author 26.07.23; revised version received 08.10.23; accepted 02.08.24; published 09.09.24.

Please cite as:
Fu L, Liu C, Dong Y, Ma X, Cai Q, Li D, Di K
Mediating Effects of Information Access on Internet Use and Multidimensional Health Among Middle-Aged and Older Adults: Nationwide Cross-Sectional Study
J Med Internet Res 2024;26:e49688
URL: https://www.jmir.org/2024/1/e49688
doi:10.2196/49688
PMID:

©Liping Fu, Caiping Liu, Yongqing Dong, Xiaodong Ma, Quanling Cai, Dongli Li, Kaisheng Di. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceptions and Satisfaction With the Use of Digital Medical Services in Urban Older Adults of China: Mixed Methods Study

Ning Wang[1*], BA, MA; Siyu Zhou[2*], BA, MA, PhD; Zhuo Liu[1], BE, MA; Ying Han[1], BA, MA, PhD

[1]School of Management, Shanxi Medical University, Jinzhong, China

[2]School of Public Health, Hangzhou Normal University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Ying Han, BA, MA, PhD
School of Management
Shanxi Medical University
Number 98, University Street
Yuci District
Jinzhong, 030600
China
Phone: 86 13603513812
Email: hanying0520@sxmu.edu.cn

Abstract

Background: In an aging and information-driven society, older adults have distinct perceptions of and specific demands for digital medical services. It is essential for society to understand these needs and develop a more thoughtful approach to digital health care.

Objective: This study aims to evaluate the behavioral intention and satisfaction of older adults with digital medical services by identifying the perceived factors and the pathways through which these factors influence their behavior.

Methods: This study used a mixed methods approach, combining qualitative and quantitative analyses. A focus group interview was conducted with 30 randomly selected older adults, and the interviews were transcribed verbatim and coded using grounded theory. In addition, 876 valid questionnaires were collected to describe older adults' perceptions of and satisfaction with digital medical care. Then, t tests and ANOVA were used to explore differences among various demographic groups, while hierarchical multiple regression was conducted to identify the factors most closely related to satisfaction. Structural equation modeling was used to identify multiple mediating effects.

Results: The qualitative study identified the core category of "medical service relief and transformation paths for older adults in the context of digital reform." Quantitative analysis revealed that more than half of the older adults were satisfied with digital medical services, and behavioral intentions were higher among those with higher incomes and education levels. Structural equation modeling confirmed that external variables, such as digital skills training, positively influenced perceived ease of use (β=.594, P<.001), perceived usefulness (β=.544, P<.001), and promoted digital medical behavioral intentions (β=.256, P<.001), while also reducing perceived risk (β=–.295, P<.001). Additionally, perceived ease of use (β=.168, P<.001) and perceived usefulness (β=.508, P<.001) positively impacted behavioral intention, whereas perceived risk (β=–.05, P=.037) exerted a negative influence. Furthermore, behavioral intention (β=.641, P<.001) significantly and positively affected older adults' satisfaction with digital medical care. The mediation test identified 4 significant paths: (1) external variables → perceived ease of use → behavioral intention (effect size of 13.9%); (2) external variables → perceived usefulness → behavioral intention (effect size of 38.4%); (3) external variables → perceived ease of use → perceived usefulness → behavioral intention (effect size of 10.1%); and (4) a direct effect (35.5%) from external variables to behavioral intention.

Conclusions: Based on the study's findings, addressing the needs of older adults and enhancing perceived usefulness are the most effective ways to encourage the use of digital health care devices. Community support plays a crucial role in helping older adults integrate into digital health care, and adapting the design of services and products to suit their needs improves their perceptions of digital health care. This, in turn, promotes usage behavior and satisfaction, while the negative impact of perceived risk remains minimal.

(J Med Internet Res 2024;26:e48654) doi:10.2196/48654

XSL•FO
RenderX

KEYWORDS

digital medical services; older adults; Technology Acceptance Model; perception; behavioral intention; satisfaction

Introduction

Background

According to data from the seventh national census, China has 260 million people aged 60 or older, accounting for 18.7% of the country's population, which is significantly higher than the global average of 12.8%. Aging has become a key concern for society, and the ongoing transformation of this issue continues to place significant pressure and challenges on China. With the rapid advancement of internet technology, "Internet Plus Service" has been applied across various sectors, creating a synergistic effect that benefits the public. When integrated with traditional health care models, the digital approach of "Internet Plus Healthcare," combined with resident services, has achieved significant success in multiple areas of society. However, older adults have increasingly become "information islands," marginalized from the digital society [1].

In the current era of rapidly advancing information and communications technology, older adults are increasingly digitally excluded [2]. In China, individuals over 60 years of age make up 41.9% of all noninternet users [3], and this group represents a marginal audience with a low willingness to use digital medical services [4]. Their reluctance is often driven by fear of the unfamiliar features of digital and mobile products. The combined influence of various factors, including personal, family, social, and technological aspects [5], has resulted in older adults struggling to efficiently access health information and services, contributing to the emergence of the "health digital divide" [6]. Compared with younger age groups, older adults must invest more effort and time to overcome the greater challenges they face when familiarizing themselves with internet technologies, such as digital medical services [7]. A study found that digital medical utilization in Jordan is significantly influenced by demographic factors, particularly age (due to physiological changes) and gender [8]. Previous research also suggests that healthy older adults are more likely to use the internet compared with those with medical conditions [7]. As body functions deteriorate with age, older adults often experience physical impairments such as hearing and vision loss [9,10], memory decline [11], reduced fine motor control [12], and cognitive decline [13], all of which affect their willingness to adopt digital medical services [14]. Concerns about the security of internet information further diminish older adults' willingness to use it. They face numerous barriers when attempting to utilize digital medical services [15], leading to technology-related anxiety [16]. According to Heart and Kalderon [17], the factors influencing health information technology utilization among aging users can be summarized into 6 dimensions: usefulness, ease of use, technology-related issues, user characteristics, social factors, and convenience measures [17].

Currently, the digital medical health care and service system for older adults in China is not yet well developed. Older adults who lack digital literacy are unable to benefit from the convenience of digital services, which in turn reduces the efficiency of medical treatment for this group. Age-related declines in physical functions, such as vision, hearing, and mobility, contribute to weaker adaptability and lower sensitivity to digital health care. Additionally, older adults may experience behavioral disorientation and diminished perception of health services during medical visits. Declining health care satisfaction among older adults has inevitably led to digital health care inequities [18]. The growing injustice in digital health, known as the "health digital divide," has significantly contributed to the underutilization of health care services by older adults in their daily care processes. To address this issue, the State Council of the People's Republic of China has emphasized the importance of "accelerating the construction of a digital society." This initiative aims to promote the use of digital services in medical, health care, and pension services, with the goal of continuously improving the public's sense of access. A series of policies have been introduced, focusing on high-frequency health care issues for older adults in the digital environment, with an emphasis on enhancing daily medical care as a key aspect of improving people's livelihoods.

This study focuses on Hangzhou City in Zhejiang Province, known as the "first city of digital governance in China." As a rapidly maturing megacity, Hangzhou boasts a high level of internet development [19]. Numerous internet technology enterprises, spearheaded by Alibaba, have supported Hangzhou's digital growth. Additionally, Hangzhou's strong influence has contributed to a level of digital governance in the Yangtze River Delta region that is significantly higher than the national average [20]. Numerous studies on digital health care and digital infrastructure development have cited Hangzhou as a model for both national and global urban digital governance [21,22].

Hangzhou City, recognized as a national role model for older adult care services, is actively implementing the national strategy to address population aging by making smart products more intelligent, practical, accessible, and suitable for older adults. While Hangzhou has made significant strides in exploring medical innovations, the system for digitally accessing medical services for older adults still requires further development. The purpose of this study is to investigate the current perception and satisfaction with digital medical services among older adults in Hangzhou through a survey. It aims to explore the primary perceptual reasons why older adults struggle to adapt to digital medical services amid the existing "health digital divide" and to propose strategies to help them overcome these challenges. This research holds significant practical value for advancing digital health care transformation, providing insights for digital reform in other areas, and enhancing its impact on digital development in China and neighboring countries. Additionally, this study integrates perceived risk and satisfaction into the Technology Acceptance Model (TAM), examining the potential relationships between these variables, which could contribute theoretically to the further refinement of the model.

XSL•FO
RenderX

Theories and Hypotheses

The TAM, depicted in Figure 1, is the most established and robust theoretical model in the field of information systems. Developed from the Theory of Reasoned Action and first proposed by Davis in 1989 [23], the TAM has undergone revisions in 1993 and 1996, evolving into the most widely used and generalized model. The modified TAM includes 3 core concepts: perceived usefulness, perceived ease of use, and behavioral intention to use.

Figure 1. The Technology Acceptance Model revised by Davis in 1996.

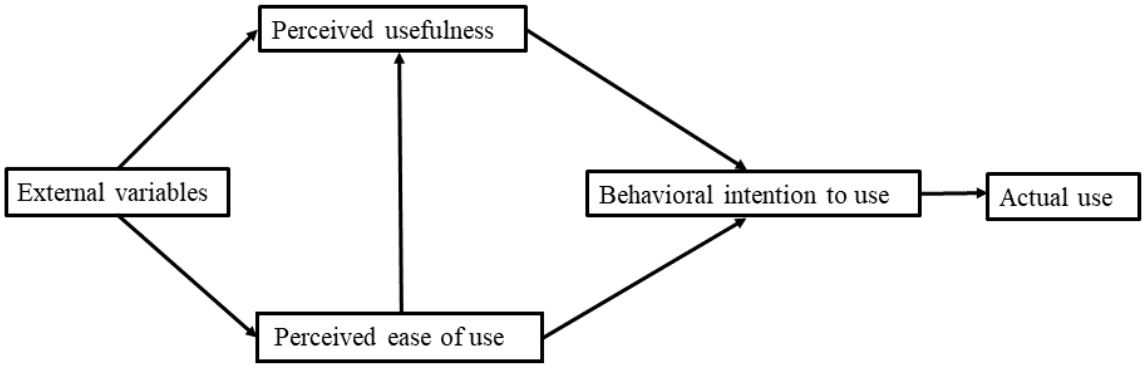

The TAM provides strong predictive and explanatory power regarding users' acceptance behavior. This study focuses on behavioral intention and satisfaction factors influencing the use of digital medical services among the aging population, including new technologies such as online booking and e-payment. Given that this population often harbors varying degrees of distrust toward digital technology, it is essential to consider the perceived risks associated with new technologies when developing new models [24]. Therefore, building on the TAM, this paper integrates the perceived risk associated with digital medical services for older adults into the model, using their satisfaction with these services as the outcome variable. The study examines the interactive relationships between perceptual elements, analyzes how each element impacts satisfaction, and aims to elucidate the influence of perceived health care utilization among urban older adults from the demand perspective. The final model is presented in Figure 2.

Figure 2. Satisfaction model of digital medical service utilization for older adults.

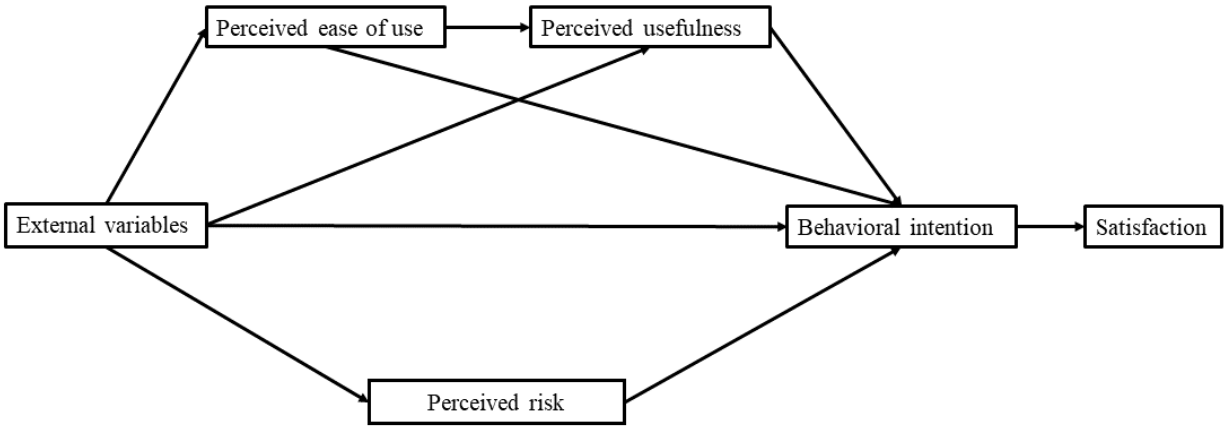

External variables include system characteristics, interface features, individual differences, and other factors [25]. Systematic assistance from digital devices can facilitate easier access to useful information, and the more straightforward this access, the more convenient it is for the user [26]. Additionally, well-designed interface features of electronic devices can reduce the user's burden, thereby influencing perceived ease of use, which in turn affects behavioral intentions. If older adults perceive digital devices as highly convenient for their lives, their desire to access digital health care is further enhanced [27]. Additionally, if they receive education about digital medical services and guidance from volunteers, family, and friends during the use of these services, their perceived usefulness will improve significantly, thereby increasing their behavioral intention to use digital medical services. This study considers external variables such as social environment factors, including training older adults in the use of digital equipment and the distribution of promotional materials. The use of online applications or self-service machines for registration and bill payment can reduce queuing time and provide additional conveniences. Individual differences are used as a control variable. According to Davis [23], external variables influence users' perceived ease of use of new technologies, and both external variables and perceived ease of use together determine perceived usefulness.

In this study, perceived usefulness refers to older adults' belief that using digital health services will enhance their access to health care or facilitate obtaining health-related information. Perceived ease of use refers to how simple older adults find it

XSL•FO
RenderX

to operate digital medical services, such as accessing instructions for using digital devices. Perceived ease of use has a significant and positive impact on perceived usefulness [23,28]. Martin et al [29] conducted a study on the adoption of mobile health (mHealth) apps among patients in community health service centers and found that both perceived ease of use and perceived usefulness positively influenced the behavioral intention to adopt these technologies. Lee's [30] study confirmed this mediating effect, demonstrating that perceived ease of use influences the attitude toward using digital services by affecting perceived usefulness.

Bauer [31] defined perceived risk as "a combination of uncertainty plus seriousness of outcome involved." In this study, perceived risk refers to older adults' concerns about safety when using new technologies for digital medical services, which directly affects their widespread adoption. Older adults may worry about risks such as personal information leakage and insecure payments when using these services [24,32]. Without adequate digital health care training and publicity from governments and organizations, mistrust of digital medical services among older adults may increase. This heightened perception of risk can diminish their motivation to engage in digital health care.

Behavioral intention refers to the willingness of older adults to use digital medical services, reflecting their subjective intent. Satisfaction, by contrast, pertains to the level of contentment with the services before, during, and after a digital medical visit. This study posits that external variables, such as social environment and convenience, influence the behavioral intention of older adults by shaping their perceptions of digital health care. These perceptions, in turn, affect their satisfaction with digital health care services. Based on this hypothesis, the following propositions are made, informed by theory and existing research.

- H1: External variables positively influence perceived ease of use and perceived usefulness, as well as behavioral intention, while negatively influencing perceived risk.
- H2: Perceived ease of use and perceived usefulness positively affect behavioral intention, while perceived risk negatively affects behavioral intention.
- H3: Perceived ease of use, perceived usefulness, and perceived risk have parallel multiple mediating effects between external variables and behavioral intention.
- H4: Behavioral intention is influenced through a chain mediation process, where perceived ease of use affects perceived usefulness, which in turn influences perceived risk, ultimately impacting behavioral intention.
- H5: Behavioral intention positively influences satisfaction with digital health care among older adults.

Methods

Participants

In this study, a multistage sampling method was applied. First, 4 administrative districts were randomly selected from the 8 main urban areas of Hangzhou. Next, a community health service center was randomly chosen from each of these districts. Finally, older adults aged 60 and over were numbered using the eHealth medical record database of each selected community health service center. As many as 5-8 individuals were randomly selected by a computer program, and these older adults were contacted by a team of family physicians. In each community, at least five participants aged 60 years or older, who had experience with digital health care, were included in focus group interviews. Additionally, for the questionnaire survey, the computer program randomly selected 250 older adults with self-directed health care behaviors from each community health service center. Before the official survey, a presurvey was conducted with 50 older adults. Based on their feedback, the questionnaire was refined to ensure clarity and ease of understanding, with adjustments made to the questions and page layout to better suit older adults (Multimedia Appendix 1).

All survey participants were trained in advance to familiarize themselves with the interview outline and questionnaire content, ensuring a thorough understanding of their meaning. They explained the purpose and significance of the study to the participants and conducted the questioning and survey only after obtaining informed consent. Ideally, the questionnaires were completed by the older adults themselves. For those who had difficulty filling out the questionnaires, the surveyors read each question aloud and recorded the participants' responses to ensure that the questionnaires were completed accurately and thoroughly.

Semistructured Interviews

In this study, the interview outline was initially developed based on literature research and then revised through expert consultation to finalize the outline. Grounded theory was used as the methodological framework, and NVivo 11.0 (Lumivero) was used for data analysis, including open coding, axial coding, and selective coding of the raw data. Memos and interview outlines were also prepared as part of the analysis process (Textbox 1).

XSL•FO
RenderX

Textbox 1. Interview outlines.

- Basic digital health perception
  - What are your health needs? Is it possible to solve this through digital methods?
- Digital medical use experience
  - Which functions of digital devices have you used, and can you tell us about your experience with and feelings about using them?
  - What are the advantages and disadvantages of digital medical services compared with manual services?
- Digital medical use dilemma
  - Do you have any difficulties using it?
  - Do you have any concerns or worries about using digital equipment for registration, and billing, among others?
  - Have you received help from family, friends, the community, volunteers, or medical personnel?
- Recommendations for digital medical services
  - What can be improved in digital medical services?

We conducted semistructured interviews with 30 older adults aged 60 years and over. This sample size met the requirements of grounded theory research, and data saturation was deemed achieved when the interviewer began hearing repeated narratives. The interview recordings and textual materials were compiled into 110,000 words of raw data. Two-thirds of these data were randomly selected for analysis, while the remaining one-third were used to test theoretical saturation. Following grounded theory [33], the study conducted open coding and conceptualization of the original text, analyzing and categorizing each word. This process resulted in 668 conceptual labels from the open coding of the data. After eliminating invalid and repetitive concepts, 125 valid concepts and 18 categories were identified. To further clarify the relationships between these categories, axial coding was used to group and recategorize the main and subcategories based on their logical relationships. The core categories were summarized from the main categories, and storylines were developed to capture the key findings within a broader theoretical context, verifying these relationships with all available information. After repeatedly comparing and analyzing the connections among the major categories, the core category was refined to present a comprehensive view of the case, titled “Medical Service Relief and Transformation for Older Adults in the Context of Digital Reform.” This analysis focused on the barriers to accessing and using digital health care from a perceptual perspective, aiming to bridge the “health digital divide” and promote health equity. Figure 3 illustrates the storyline, main categories, and subcategories. Theoretical saturation occurs when no new concepts or categories emerge during the data analysis process [34]. By continuing to code the remaining one-third of the interview texts, we found that no new concepts or relationships were generated. This indicates that the coding process for this study is nearly complete, and theoretical saturation has been achieved, providing a certain degree of realistic explanatory power. The qualitative research effectively examines the influencing factors, satisfaction, suggestions, expectations, and future prospects for digital medical service utilization among older adults in Hangzhou, offering valuable insights for questionnaire design.

XSL•FO
RenderX

Figure 3. The results of 3-level coding.

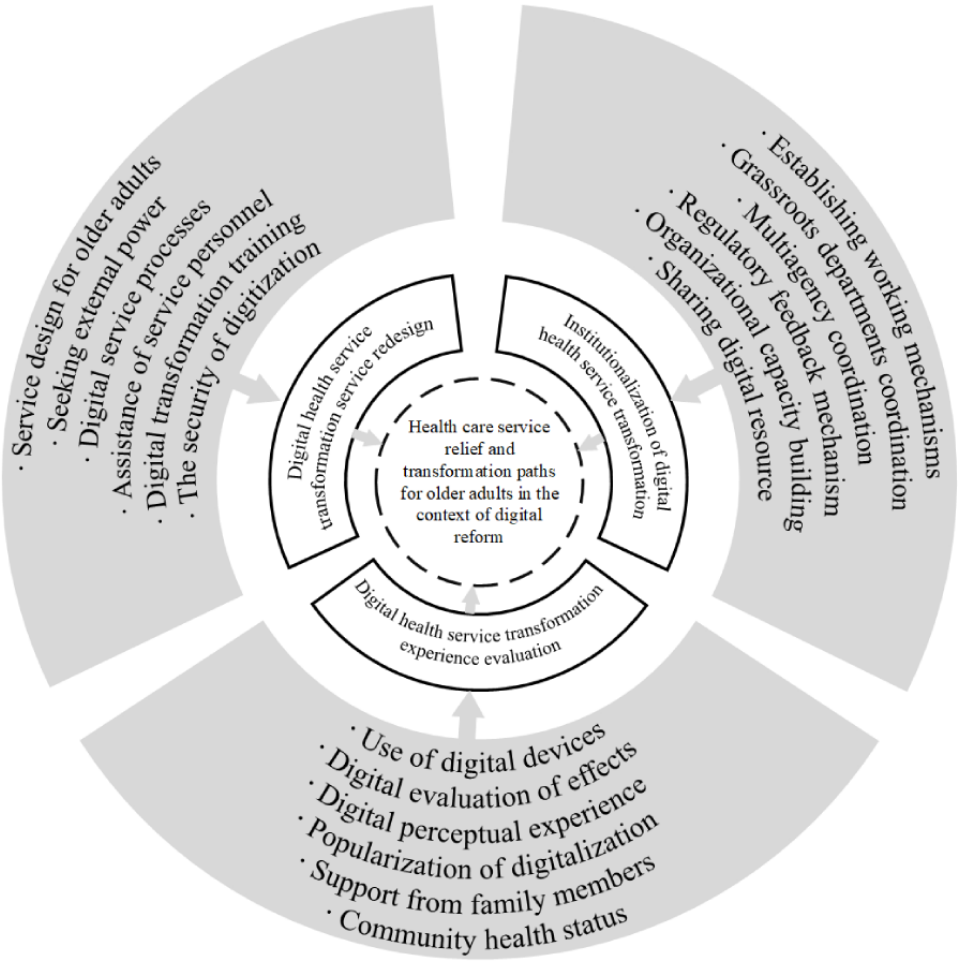

Quantitative Analysis

Following the qualitative research, a questionnaire was developed and administered by trained investigators who adhered to uniform protocols. The purpose and significance of the study were explained to eligible senior citizens, and their consent was obtained before beginning the survey. To ensure accuracy, each respondent's records were reviewed individually. We then analyzed the collected data, removed invalid questionnaires, and addressed dummy variables for categorical data. Reliability and validity tests were conducted to ensure the quality of the data. The Cronbach α coefficient for the overall scale in this study was 0.914, and the Cronbach α values for all dimensions were greater than 0.8, indicating that the reliability of the questionnaire was excellent.

First, the scale designed in this study demonstrates good content validity, based on references to numerous studies and modifications made to the questionnaire content following expert consultation. Second, the Bartlett test and the Kaiser-Meyer-Olkin test were conducted, with a Kaiser-Meyer-Olkin value of 0.933 (P<.001), indicating that the data are highly suitable for factor analysis. Third, confirmatory factor analysis was performed using Amos 26.0 (IBM Corp.). The overall fit indices of the model were acceptable, suggesting good structural validity (χ^2/df=4.957, df=303, root-mean-square error of approximation 0.067, normed fit index 0.926>0.9, relative fit index 0.914>0.9, incremental fit index 0.940>0.9, Tucker-Lewis index 0.930>0.9, and comparative fit index 0.940>0.9). Fourth, the average variance extracted for each latent variable was greater than 0.5, and the composite reliability was greater than 0.8, indicating ideal convergent validity (Table 1; also see Textbox 1). Fifth, as shown in Table 2, most correlation coefficients were less than the square root of the average variance extracted, suggesting that there is a sufficient level of differentiation between the latent variables. This indicates that the discriminant validity of the scale's data is acceptable. In summary, the questionnaire demonstrates good validity.

XSL•FO
RenderX

Table 1. Questionnaire items and results of reliability and validity tests.

Dimensions[a]	Cronbach α	Average variance extracted	Composite reliability	References
External variables	0.856	0.503	0.847	[35-37]
Perceived ease of use	0.889	0.787	0.917	[28,38-40]
Perceived usefulness	0.896	0.742	0.896	[28,38-40]
Perceived risk	0.841	0.549	0.827	[24,41,42]
Behavioral intention	0.943	0.884	0.942	[28,38,39]
Satisfaction	0.946	0.662	0.940	[40,43,44]

[a]See Textbox 2 for scale items.

XSL•FO
RenderX

Textbox 2. Scale items of the questionnaire.

- External variables
  - T1: If digital devices improve friendliness (eg, longer operation time, simplified interaction interface and flow, and enlarged fonts), it will increase my willingness to use.
  - T2: The community offers training on the use of digital medical devices, and I would like to attend.
  - T3: I would like to read and study the digital medical information provided by the community.
  - T4: I think self-service registration (self-service machines or cell phones) is more convenient than manual registration.
  - T5: I think it is more convenient to make an appointment than to wait for a consultation on-site.
  - T6: I think self-service payment (self-service machines or cell phones) is more convenient than tollbooth.
- Perceived ease of use
  - T7: I can use my smartphone proficiently.
  - T8: I can make online appointments and payments.
  - T9: I can operate the self-service machines in the hospital.
- Perceived usefulness
  - T10: Digital access can save time.
  - T11: Digital access helps me get more information about medical care in advance.
  - T12: It is easier to adapt to society if you have digital medical skills.
- Perceived risk
  - T13: I am concerned that third-party platforms used in the digital medical process (eg, WeChat, various apps) may disclose personal information.
  - T14: I am concerned about the security of online and self-service machine payment.
  - T15: I am concerned that there is a haphazard fee for doctor's office billing.
  - T16: I am concerned about the accuracy of the self-service report retrieval.
- Behavioral intention
  - T17: I am willing to accept digital access to medicine.
  - T18: I would like to learn about the use of digital medical equipment.
  - T19: I would like to use digital medical equipment.
- Satisfaction
  - T20: Various ways of digital appointment registration.
  - T21: Guidance on the use of digital medical equipment.
  - T22: Online doctor appointment.
  - T23: Call the information indicator screen in the waiting room.
  - T24: Pay for medical treatment in the doctor's office without paying at the tollbooth.
  - T25: Self-service payment (mobile phone/self-service machine).
  - T26: Self-service report retrieval (mobile phone/self-service machine).
  - T27: The whole process of digital medical treatment.

XSL•FO
RenderX

Table 2. Discriminant validity[a].

Dimensions	External variables	Perceived ease of use	Perceived risk	Perceived usefulness	Behavioral intention	Satisfaction
External variables	*0.503*	__[b]	—	—	—	—
Perceived ease of use	0.594	*0.787*	—	—	—	—
Perceived risk	–0.295	–0.175	*0.549*	—	—	—
Perceived usefulness	0.688	0.565	–0.203	*0.742*	—	—
Behavioral intention	0.721	0.617	–0.258	0.79	*0.844*	—
Satisfaction	0.462	0.395	–0.165	0.506	0.641	*0.662*
Square root of average variance extracted	0.709	0.887	0.741	0.861	0.919	0.814

[a]The diagonal cell value (italicized) is the average variance extracted.

[b]Not available.

The questionnaire uses a 5-point Likert scale, where 1 denotes "strongly disagree," 2 denotes "disagree," 3 denotes "neutral," 4 denotes "agree," and 5 denotes "strongly agree." This scale is used to assess older adults' feedback on digital medical services, as shown in Table 2.

The core formula for scoring for each dimension is as follows:

The questionnaire includes the following: (1) basic demographic characteristics, such as gender, age, education level, marital status, residency, former occupation, and frequency of medical treatment, to provide descriptive statistical analysis of the sample for the empirical study; (2) a survey scale for digital health care service utilization among urban older adults, covering external variables, perceived ease of use, perceived usefulness, perceived risk, and behavioral intentions; and (3) satisfaction evaluation of digital medical services, which includes ratings for each stage of digital access to care.

We identified the primary concerns of older adults and synthesized findings from previous studies and qualitative research to design questionnaire items corresponding to each dimension. The "external variables" were designed to assess the impact of external factors on the willingness to use and satisfaction with digital medical devices among older adults. These factors include age-adapted equipment (T1), community guidance on use (T2 and T3), and convenience conditions (T4, T5, and T6). The "perceived ease of use" dimension examines difficulties encountered with devices (smartphones T8 and kiosks T9) and processes (registration and payment T4) during medical care, which may affect willingness to use and satisfaction. The "perceived usefulness" dimension evaluates the role and value of digital health care services for older adults from 3 perspectives: time saving (T10), knowledge acquisition (T11), and social adaptation (T12). This helps in identifying areas for service improvement and upgrading. The "perceived riskiness" dimension investigates the risks that concern older adults, including privacy protection (T13), payment security (T14), cost settlement (T15), and report accuracy (T16). Addressing these risks in software design can enhance trust in digital health care. Finally, the "behavioral intentions" dimension analyzes acceptance (T15), learning (T16), and usage (T17) processes to identify barriers to digital health care service utilization by older adults. This provides insights for future intervention development to improve engagement and satisfaction.

According to the sample estimation method proposed by Kendall [45], which suggests having 10-20 samples per questionnaire item, and accounting for the questionnaire recovery rate and efficiency, the sample size should be increased by 20%. Using the formula 27 questions items × 20 × (1 + 20%) = 648, the study requires a minimum of 648 samples. As a result of the complex analysis of the data using multivariate statistical methods, the sample size was increased to 1000 to ensure more robust statistical parameters. Data entry was performed by 2 individuals using 2 separate machines. Out of the 1000 questionnaires distributed, 876 valid questionnaires were recovered, resulting in an effective recovery rate of 87.60%.

In this study, descriptive analyses were performed to understand the basic demographic information and perceptions of digital access among the older adults interviewed. Additionally, difference analyses were conducted to determine whether there were variations in core explanatory variables across different genders, ages, educational levels, and other demographic characteristics. Hierarchical regression analysis was used to determine whether including core explanatory variables—such as external variables, perceived ease of use, perceived usefulness, perceived risk, and behavioral intention—was meaningful for refining the model. Finally, structural equation modeling was used to test the multiple mediating effects of perceived usefulness, perceived ease of use, and perceived risk on the relationship between external variables and behavioral intention, as well as the chain mediating effects of perceived ease of use and perceived usefulness.

Ethics Approval

This study was approved by the Hangzhou Normal University Ethics Committee (research ethics committee number 2021-1147). Participants were informed about the use of their data before both the interview and questionnaire began and were given the right to be informed and to withdraw at any time. All data were anonymized to protect participant privacy. As the

XSL•FO
RenderX

survey posed no potential harm to respondents, we provided facial tissues, towels, and other household items as tokens of appreciation for their participation and cooperation.

Results

Participant Characteristics

Descriptive analyses were conducted to assess the demographic characteristics, perceptions of digital medical care, and current satisfaction levels among the surveyed older adults. Table 3 presents the demographic information of the respondents. The majority had an education level of junior high school (366/876, 41.8%). Additionally, over one-half (478/876, 54.6%) of older adults reported that declining physical function impacted their use of digital devices. Their preferred registration methods were, in order, manual window registration, self-service machine registration, and online registration. Feedback from older adults indicated several issues with digital medical equipment, including overly complex interfaces and processes, short time limits, small font sizes, inconsistent system updates across hospitals, and frequent equipment failures. These issues increase the difficulty of using digital health care devices and may serve as significant barriers to access and utilization for older adults, exacerbating gaps in access to health information and widening the “health digital divide.”

XSL•FO
RenderX

Table 3. Participant demographic data (N=876).

Characteristics	Participants n (%)
Gender	
Male	358 (40.9)
Female	518 (59.1)
Age (years)	
60-69	433 (49.4)
70-79	282 (32.2)
80-89	151 (17.2)
≥90	10 (1.1)
Education	
Primary and below	125 (14.3)
Junior high school	366 (41.8)
Technical secondary school/senior high school	243 (27.7)
Junior college/bachelor and above	142 (16.2)
Marriage	
Married	724 (82.6)
Unmarried	19 (2.2)
Bereaved spouse	133 (15.2)
Living conditions	
Living with spouse	400 (45.7)
Living with spouse and children	288 (32.9)
Living alone	141 (16.1)
Others	47 (5.4)
Monthly disposable income (CNY[a])	
<1000	19 (2.2)
1001-3000	124 (14.2)
3001-5000	483 (55.1)
5001-7000	179 (20.4)
>7000	71 (8.1)
Previous occupations	
Government offices	64 (7.3)
Enterprises and businesses	575 (65.6)
Private enterprise	127 (14.5)
Individuals	62 (7.1)
Others	48 (5.5)
Frequency of medical treatment in the past 6 months	
0	166 (18.9)
1-2	306 (34.9)
3-4	135 (15.4)
>5	269 (30.7)
Health insurance	

XSL•FO
RenderX

Characteristics	Participants n (%)
Urban employee medical insurance	431 (49.2)
Urban residents' medical insurance	357 (40.8)
New rural cooperative medical care	35 (4.0)
Commercial health insurance	10 (1.1)
None	41 (4.7)
Others	2 (0.2)
Deterioration of physical function (eg, vision loss, hearing loss) that affects your ability to use digital medical devices	
Totally disagree	69 (7.9)
Comparatively disagree	95 (10.8)
Generally	234 (26.7)
Comparatively agree	395 (45.1)
Totally agree	83 (9.5)
Preferred appointment method	
Online	157 (17.9)
Self-service machines	323 (36.9)
Manual window	334 (38.1)
Doctor makes the next clinic appointment	19 (2.2)
Make a call	28 (3.2)
Others	15 (1.7)
The problems of digital medical equipment	
Time limit is too short	106 (12.1)
Too complicated interface and process	469 (53.5)
Font size too small	92 (10.5)
Others	209 (23.9)

[a]1 CNY=US $0.14.

Participants' Perceptions and Behavioral Intentions of Digital Medical Care

Table 4 presents the descriptive analysis of scores for each influencing factor related to digital medical service utilization among urban older adults. The SD for all variables was above 0.5, indicating variability in the responses among the survey participants. After compiling the statistical data, the scores for the dimensions influencing the perception of digital medical service utilization among older adults, ranked from the highest to the lowest, were as follows: perceived usefulness (mean 3.78, SD 0.86), external variables (mean 3.68, SD 0.74), behavioral intention (mean 3.61, SD 0.97), perceived ease of use (mean 2.94, SD 1.01), and perceived risk (mean 2.80, SD 0.87). This indicates that while most older adults perceive digital health care access as beneficial, challenges such as operational difficulty and perceived risks hinder their effective use of digital access devices.

Table 4. Descriptive statistical analysis of influencing factors[a].

Influencing factors	Range	Mean (SD)
External variables	1.00-5.00	3.68 (0.74)
Perceived usefulness	1.00-5.00	3.78 (0.86)
Perceived ease of use	1.00-5.00	2.94 (1.01)
Perceived risk	1.00-5.00	2.80 (0.87)
Behavioral intention	1.00-5.00	3.61 (0.97)

[a]Number of valid cases: 876.

XSL•FO
RenderX

Participant Satisfaction With Each Aspect of Digital Medical Care

Table 5 shows the satisfaction evaluation of digital medical services and equipment at various stages of the diagnosis process. More than 60% of respondents expressed satisfaction with aspects such as registration methods (570/876, 65.1%), digital medical guidance (550/876, 62.8%), online doctor appointments (527/876, 60.2%), call information indicator screens in waiting rooms (584/876, 66.7%), payment at the doctor's office (545/876, 62.2%), self-service payment (545/876, 62.2%), and self-service report retrieval (558/876, 63.7%).

Table 5. The constituent ratio of digital medical satisfaction (N=876).

Variables	Highly dissatisfied, n (%)	Dissatisfied, n (%)	Common, n (%)	Satisfied, n (%)	Highly satisfied, n (%)
Various ways of digital appointment registration	10 (1.1)	47 (5.4)	249 (28.4)	443 (50.6)	127 (14.5)
Guidance on the use of digital medical equipment	8 (0.9)	62 (7.1)	256 (29.2)	412 (47.0)	138 (15.8)
Online doctor appointment	11 (1.3)	51 (5.8)	287 (32.8)	396 (45.2)	131 (15.0)
Call information indicator screen in the waiting room	7 (0.8)	38 (4.3)	247 (28.2)	444 (50.7)	140 (16.0)
Pay for medical treatment in the doctor's office without paying at the tollbooth	9 (1.0)	44 (5.0)	278 (31.7)	401 (45.8)	144 (16.4)
Self-service payment (mobile phone/self-service machine)	12 (1.4)	61 (7.0)	258 (29.5)	415 (47.4)	130 (14.8)
Self-service report retrieval (mobile phone/self-service machine)	13 (1.5)	63 (7.2)	242 (27.6)	429 (49.0)	129 (14.7)
The whole process of digital medical treatment	9 (1.0)	45 (5.1)	238 (27.2)	464 (53.0)	120 (13.7)

Differences in Perceptions, Behavioral Intentions, and Satisfaction Between Groups

The *t* test (2-tailed) and ANOVA were used to explore differences in external variables, perceived ease of use, perceived usefulness, perceived risk, behavioral intention, and satisfaction among individuals with different characteristics (as shown in Table 6). Gender differences were not statistically significant in external variables (P=.97), perceived ease of use (P=.42), perceived usefulness (P=.11), perceived risk (P=.98), behavioral intention (P=.27), and satisfaction (P=.83). Older age and lower levels of education were associated with lower scores on external variables, perceived ease of use, perceived usefulness, behavioral intention, and satisfaction. However, the differences in age (P=.10) and education (P=.14) in perceived risk were not statistically significant. Married older adults had significantly higher scores on external variables (P<.001) and overall satisfaction (P=.03). Those living with their children scored higher in perceived ease of use and perceived risk of digital medical devices. This may be because living with children provides easier access to guidance, which reduces the difficulty of using digital devices, while reminders from children about fraud may heighten their awareness of potential risks. Conversely, older adults living alone with their spouses scored higher in the remaining dimensions. Older adults with higher incomes reported lower perceived risk and higher scores across other dimensions. This trend may be linked to higher education levels associated with greater income. Additionally, older adults with previous work experience in government positions scored higher in perceived usefulness and behavioral intention, potentially due to increased confidence in trying new technologies gained from their professional backgrounds. Those with frequent access to health care had higher scores on external variables, perceived usefulness, behavioral intention, and satisfaction, and lower scores on perceived risk. This likely results from their greater familiarity with digital access devices, leading to increased proficiency and a better appreciation of the convenience digital health care offers.

XSL•FO
RenderX

Table 6. Difference analysis of characteristics of older adults.

Variables	External variables	Perceived ease of use	Perceived usefulness	Perceived risk	Behavioral intention	Satisfaction
Gender						
Male, mean (SD)	3.68 (0.69)	3.00 (1.08)	3.72 (0.89)	2.80 (0.90)	3.57 (0.98)	3.71 (1.08)
Female, mean (SD)	3.69 (0.77)	2.97 (1.08)	3.82 (0.83)	2.80 (0.85)	3.64 (0.96)	3.70 (0.71)
t test[a] (*df*)	–0.031 (874)	–0.802 (874)	–1.597 (874)	0.018 (874)	–1.112 (874)	0.216 (874)
P value	.97	.42	.11	.98	.27	.83
Age (years)						
60-69, mean (SD)	3.81 (0.76)	3.35 (1.04)	3.90 (0.44)	2.75 (0.86)	3.79 (0.97)	3.82 (0.74)
70-79, mean (SD)	3.64 (0.63)	2.68 (0.95)	0.72 (0.43)	2.78 (0.88)	3.58 (0.84)	3.67 (0.65)
80-89, mean (SD)	3.47 (0.75)	2.33 (0.95)	0.88 (0.72)	2.94 (0.90)	3.22 (1.07)	3.49 (0.67)
>90, mean (SD)	2.63 (0.94)	2.27 (1.15)	0.80 (0.25)	3.05 (0.82)	2.53 (0.65)	3.41 (0.78)
F test (*df*)	12.611 (872)	51.127 (872)	7.809 (872)	2.087 (872)	20.613 (872)	9.298 (872)
P value	<.001	<.001	<.001	.10	<.001	<.001
Education						
Primary and below, mean (SD)	3.41 (0.87)	2.37 (1.07)	3.64 (0.82)	2.92 (0.89)	3.20 (1.15)	3.55 (0.78)
Junior high school, mean (SD)	3.66 (0.69)	2.85 (0.99)	3.72 (0.84)	2.83 (0.85)	3.55 (0.90)	3.68 (0.70)
Technical secondary school/senior high school, mean (SD)	3.78 (0.71)	3.19 (0.99)	3.86 (0.86)	2.75 (0.85)	3.75 (0.93)	3.78 (0.65)
Junior college/bachelor and above, mean (SD)	3.81 (0.74)	3.28 (1.20)	3.93 (0.89)	2.70 (0.94)	3.91 (0.91)	3.78 (0.74)
F test (*df*)	7.348 (872)	21.357 (872)	4.156 (872)	1.845 (872)	12.516 (872)	3.787 (872)
P value	<.001	<.001	.006	.14	<.001	.01
Marriage						
Married, mean (SD)	3.73 (0.71)	3.02 (1.06)	3.81 (0.84)	2.80 (0.87)	3.67 (0.92)	3.74 (0.70)
Unmarried, mean (SD)	3.54 (0.91)	3.04 (1.28)	3.65 (1.13)	2.74 (1.04)	3.68 (1.12)	3.49 (1.15)
Bereaved spouse, mean (SD)	3.44 (0.81)	2.53 (1.07)	3.64 (0.86)	2.80 (0.88)	3.28 (1.14)	3.57 (0.69)
F test (*df*)	9.647 (873)	11.827 (873)	2.270 (873)	0.049 (873)	7.007 (873)	3.837 (873)
P value	<.001	<.001	.10	.95	.002	.03
Living conditions						
Living only with spouse, mean (SD)	3.81 (0.66)	2.98 (1.09)	3.90 (0.83)	2.72 (0.87)	3.77 (0.90)	3.78 (0.68)
Living with spouse and children, mean (SD)	3.65 (0.70)	3.05 (0.99)	3.67 (0.82)	2.90 (0.84)	3.56 (0.90)	3.74 (0.66)
Living alone, mean (SD)	3.52 (0.90)	2.66 (1.16)	3.75 (0.91)	2.74 (0.96)	3.45 (1.12)	3.59 (0.84)
Others, mean (SD)	3.32 (0.82)	2.82 (1.13)	3.48 (0.99)	3.05 (0.70)	3.09 (1.20)	3.27 (0.74)
F test (*df*)	9.383 (872)	4.584 (872)	6.526 (872)	4.475 (872)	7.895 (872)	7.716 (872)
P value	<.001	.003	<.001	.005	<.001	<.001
Monthly disposable income (CNY[b])						

XSL•FO
RenderX

Variables	External variables	Perceived ease of use	Perceived usefulness	Perceived risk	Behavioral intention	Satisfaction
<1000, mean (SD)	3.07 (1.28)	2.56 (1.14)	3.28 (1.21)	3.01 (1.12)	2.72 (1.10)	3.02 (1.15)
1001-3000, mean (SD)	3.54 (0.73)	2.74 (1.06)	3.56 (0.80)	2.98 (0.86)	3.37 (0.94)	3.62 (0.77)
3001-5000, mean (SD)	3.69 (0.69)	2.87 (1.03)	3.76 (0.82)	2.81 (0.83)	3.63 (0.92)	3.71 (0.63)
5001-7000, mean (SD)	3.75 (0.74)	3.20 (1.04)	3.97 (0.82)	2.70 (0.92)	3.74 (0.98)	3.73 (0.74)
>7000, mean (SD)	3.92 (0.74)	3.26 (1.32)	3.97 (1.02)	2.62 (0.95)	3.84 (1.09)	3.96 (0.75)
F test (*df*)	4.389 (871)	5.984 (871)	6.316 (871)	2.737 (871)	8.000 (871)	4.078 (871)
P value	.003	<.001	<.001	.03	<.001	.004
Previous occupations						
Government offices, mean (SD)	3.75 (0.77)	3.01 (1.11)	3.88 (0.85)	2.62 (0.90)	3.88 (0.78)	3.65 (0.74)
Enterprises and businesses, mean (SD)	3.73 (0.71)	2.98 (1.09)	3.86 (0.82)	2.77 (0.86)	3.68 (0.96)	3.77 (0.68)
Private enterprise, mean (SD)	3.61 (0.66)	2.74 (1.03)	3.63 (0.84)	2.83 (0.86)	3.43 (0.97)	3.70 (0.65)
Individuals, mean (SD)	3.65 (0.77)	3.06 (1.04)	3.61 (0.92)	3.07 (0.85)	3.52 (1.02)	3.58 (0.74)
Others, mean (SD)	3.25 (0.98)	2.89 (1.07)	3.33 (1.02)	2.97 (0.92)	2.98 (0.98)	3.21 (0.91)
F test (*df*)	3.508 (871)	1.570 (871)	6.512 (871)	2.946 (871)	8.666 (871)	7.882 (871)
P value	.009	.18	<.001	.02	<.001	<.001
Frequency of medical treatment						
0 times, mean (SD)	3.60 (0.85)	3.01 (1.17)	3.67 (0.94)	2.83 (0.87)	3.50 (1.05)	3.62 (0.79)
1-2 times, mean (SD)	3.67 (0.70)	2.87 (0.99)	3.69 (0.82)	2.92 (0.87)	3.65 (0.88)	3.66 (0.68)
3-4 times, mean (SD)	3.61 (0.72)	2.87 (1.10)	3.80 (0.84)	2.82 (0.85)	3.59 (0.94)	3.63 (0.70)
5 times and more, mean (SD)	3.79 (0.69)	3.02 (1.10)	3.94 (0.83)	2.63 (0.86)	3.64 (1.03)	3.85 (0.68)
F test (*df*)	3.216 (872)	1.406 (872)	5.119 (872)	5.636 (872)	0.865 (872)	5.606 (872)
P value	.02	.24	.002	.001	.46	.001
Health insurance						
Urban employee medical insurance, mean (SD)	3.74 (0.68)	3.04 (1.06)	3.89 (0.80)	2.76 (0.86)	3.74 (0.95)	3.78 (0.66)
Urban residents' medical insurance, mean (SD)	3.64 (0.69)	2.80 (1.05)	3.64 (0.83)	2.80 (0.85)	3.49 (0.91)	3.67 (0.69)
New rural cooperative medical care, mean (SD)	3.34 (1.09)	2.59 (1.13)	3.63 (1.15)	2.89 (1.15)	3.10 (1.24)	3.28 (0.97)
Commercial health insurance, mean (SD)	3.00 (1.14)	3.03 (1.13)	3.07 (1.40)	3.75 (0.87)	3.10 (1.35)	3.26 (1.16)
None, mean (SD)	3.89 (1.04)	3.41 (1.23)	4.09 (0.87)	2.80 (0.82)	3.86 (1.08)	3.74 (0.90)
Others, mean (SD)	3.67 (0.47)	3.83 (0.24)	3.50 (0.71)	2.88 (0.53)	3.50 (0.71)	3.94 (0.09
F test (*df*)	2.042 (870)	4.705 (870)	4.350 (870)	2.637 (870)	3.708 (870)	4.182 (870)
P value	.16	<.001	.02	.02	.04	.02

[a]2-tailed.

[b]1 CNY=US $0.14.

Factors Associated With Satisfaction of Digital Medical Care

Hierarchical multiple regression analysis was conducted to evaluate the impact of each dimension of perception factors on older adults' satisfaction with digital health care (as shown in Table 7). This analysis also assessed whether the effects of control variables on the explanatory variables were significant. The findings from this regression analysis provided valuable insights for constructing the subsequent structural equation model. Using demographic variables such as gender, age, and education as control variables, the model's significance level was less than .001, and each variance inflation factor was below 5, indicating no multicollinearity among the variables and that the findings were statistically significant. All dimensions

XSL•FO
RenderX

significantly influenced satisfaction with digital health care access, supporting hypothesis H5, which states that "behavioral intention positively influences satisfaction." Demographic variables did not have a significant effect and were therefore excluded from the structural equation model construction. Regression analysis was conducted to confirm the reliability of the results after removing the control variables (model 3). The findings showed that the significance and direction of the coefficients for the core independent variables remained consistent, indicating that the model passed the robustness test and confirming a generally reliable relationship between these variables and the dependent variable (Table 7).

Table 7. Hierarchical multiple regression results.

Variables	Satisfaction		
	Model 1	Model 2	Model 3
Constant	3.931[a]	1.808[a]	1.570[a]
Gender	0.012	–0.041	—[b]
Age	–0.181[a]	0.008	—
Education	0.032	–0.042	—
Marriage	0.036	0.033	—
Living conditions	–0.110[a]	–0.035	—
Monthly disposable income	0.080[c]	–0.003	—
Previous occupations	–0.094[a]	–0.029	—
Frequency of medical treatment	0.055[c]	0.021	—
Health insurance	–0.021	–0.027	—
External variables	—	0.346[a]	0.354[a]
Perceived usefulness	—	0.110[a]	0.124[a]
Perceived ease of use	—	0.136[a]	0.114[a]
Perceived risk	—	–0.069[a]	–0.076[a]
Behavioral intention	—	0.069[c]	0.066[c]
F test (*df*)	10.893[a] (866)	130.349[a] (861)	159.84[a] (870)
R^2	0.102	0.489	0.479
ΔR^2	0.102	0.387	—

[a]*P*<.001.
[b]Not applicable.
[c]*P*<.05.

Structural Equation Modeling of Perceptions and Satisfaction of Digital Medical Care for Older Adults

To determine the path coefficient relationships between the latent variables and the multiple mediation effects in this study, structural equation modeling was conducted using Amos 26.0 software (Figure 4). The significance of the paths was analyzed using *t* test results (Table 8). The test results are as follows:

- The significant effects of external variables on the dimensions, in descending order, were perceived ease of use (β=.594, *P*<.001), perceived usefulness (β=.544, *P*<.001), perceived risk (β=–.295, *P*<.001), and behavioral intention (β=.256, *P*<.001). This supports hypothesis H1.
- Perceived usefulness (β=.508, *P*<.001), perceived ease of use (β=.168, *P*<.001), and external variables (β=.256, *P*<.001) have significant positive effects on behavioral intention, while perceived risk (β=–.05, *P*=.04) has a weak negative effect. The effect of perceived usefulness is notably greater than that of the other dimensions. Therefore, hypothesis H2 is supported.
- Behavioral intention significantly and positively affects the satisfaction with digital health care access among older adults, with a path coefficient of 0.641 (*P*<.001). Therefore, hypothesis H5 is supported.

XSL•FO
RenderX

Figure 4. The structural equation model.

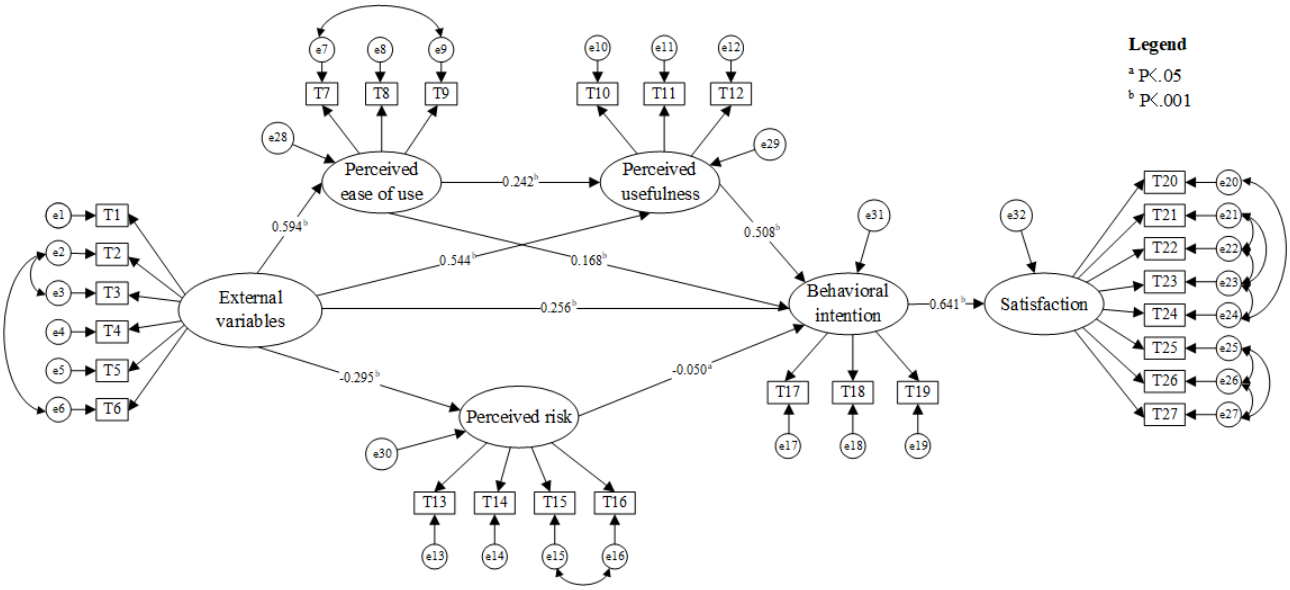

Table 8. Scalar estimates of the research model.

Outcome variables and predictive variables	R^2	β	SEs	Critical ratio	*P* value
Perceived ease of use					
External variables	0.353	.594	0.042	18.801	<.001
Perceived usefulness					
External variables	0.511	.544	0.038	14.352	<.001
Perceived ease of use	__[a]	.242	0.028	6.714	<.001
Perceived risk					
External variables	0.087	–.295	0.042	–7.945	<.001
Behavioral intention					
External variables	0.703	.256	0.042	7.179	<.001
Perceived ease of use	—	.168	0.025	5.87	<.001
Perceived usefulness	—	.508	0.041	14.281	<.001
Perceived risk	—	–.05	0.025	–2.087	.04
Satisfaction					
Behavioral intention	0.41	.641	0.023	20.703	<.001

[a]Not available.

Mediating Role Played by Perceived Ease of Use, Perceived Usefulness, and Perceived Risk

To verify the multiple mediating roles of perceived ease of use, perceived usefulness, and perceived risk between external variables and behavioral intention (H3) and the chained mediating roles of perceived ease of use and perceived usefulness (H4), the bias-corrected percentile bootstrap procedure test for the significance of each path of the multimediation model was performed using Amos 26.0, with 5000 repeated extractions to obtain 95% CI [46]. As can be seen from Table 9, the bootstrap (95% CI) of ind3 is [0-0.033], which contains 0, indicating that the mediation effect of this path is not significant, and the bootstrap (95% CI) of ind1, ind2, and ind4 does not contain 0, which has a significant mediation effect, and the value of the mediation effect is ind2>ind1>ind4. Thus, hypothesis H3 partially holds, the mediating effects of perceived ease of use and perceived usefulness are significant, the mediation hypothesis of perceived risk does not hold, and hypothesis H4 of chained mediation effects is valid.

XSL•FO
RenderX

Table 9. Results of distal mediation test.

Effect types	Effect	Boot SE	Boot lower level of confidence interval	Boot upper level of confidence interval	Ratio of indirect to total effect, %	Ratio of indirect to direct effect, %
Total effect	0.721	0.027	0.662	0.771	__[a]	—
Direct effect	0.256	0.045	0.169	0.345	35.5	—
Total indirect effect	0.465	0.033	0.403	0.537	64.5	—
ind1: external variables → perceived ease of use → behavioral intention	0.1	0.019	0.066	0.142	13.9	39.1
ind2: external variables → perceived usefulness → behavioral intention	0.277	0.034	0.213	0.348	38.4	108.2
ind3: external variables → perceived risk → behavioral intention	0.015	0.008	0	0.033	2.1	5.9
ind4: external variables → perceived ease of use → perceived usefulness → behavioral intention	0.073	0.016	0.044	0.11	10.1	28.5
C1: ind1 – ind2	–0.176	0.044	–0.266	–0.092	—	—
C2: ind1 – ind3	0.086	0.021	0.047	0.129	—	—
C3: ind1 – ind4	0.027	0.028	–0.027	0.081	—	—
C4: ind2 – ind3	0.262	0.036	0.197	0.336	—	—
C5: ind2 – ind4	0.203	0.041	0.124	0.286	—	—
C6: ind3 – ind4	–0.058	0.019	–0.098	–0.023	—	—

[a]Not available.

Discussion

Principal Findings

This study used a mixed methods approach, combining qualitative and quantitative analyses. Both methods revealed that while older adults recognize the convenience of digital access, their willingness to use digital devices is hindered by several factors. These include low literacy, vision and hearing impairments, inadequate assistance and guidance, and apprehension about new technologies. Additionally, there is a desire for digital devices to be adapted to the needs of older adults to enhance their satisfaction. The qualitative research provided insights into the digital challenges faced by older adults and highlighted their expectations for future digital development based on their personal needs, offering a scientific basis for designing the questionnaire. Meanwhile, the quantitative study utilized the questionnaire data to verify whether these challenges affected older adults' willingness to use digital technologies and to assess the pathways and extent of this impact, thereby validating the hypotheses.

The current situation and trend in digital medical care for older adults reveal low awareness, low utilization, and high demand [47]. This study, based on the AM, examined the current state of digital medical services for older adults and explored the factors influencing their behavioral intention and satisfaction. The following conclusions, consistent with previous studies, were drawn:

Each dimension significantly affects behavioral intention. Jiao [48] argued that perceived ease of use has the greatest impact on the attitude toward continued use of online appointment platforms, while Shang and Wang [49] found that perceived usefulness positively influences the behavioral intention to use digital health care services among older adults. In this study, perceived usefulness emerges as the most critical factor influencing behavioral intention toward digital health care platforms and devices among older adults. This suggests that if digital health care services are designed to effectively serve and facilitate older adults, many other challenges can be overcome. Therefore, attention should be focused on transforming and upgrading digital health care platforms and devices to enhance service convenience and effectiveness. To encourage older adults to use digital health care devices, it is essential to balance the improvement of external variables—such as social support and the user-friendliness of the technology—with enhancing perceived usefulness and reducing perceived risk. This involves institutionalizing support, redesigning services, and refining the digital health care experience to accommodate the aging population effectively.

Pavlou [36] argued that perceived risk negatively affects the willingness to use technology, and this study similarly validates that finding. However, older adults perceive digital access to health care as low risk and believe it has minimal impact on their behavioral intentions. This study suggests the following reasons for this phenomenon: First, older adults often lack awareness of the potential risks associated with digitalization in health care, leading to limited concern about these risks.

XSL•FO
RenderX

Second, most digital health care services are provided by public hospitals, and older adults' trust in government-run health care organizations extends to their online platforms and self-service devices. Third, digital access platforms and devices are subject to uniform and stringent regulatory standards, which reduces negative press and fosters greater trust among older adults. Fourth, the convenience offered by digitalization sufficiently outweighs its potential risks, leading older adults to remain willing to experiment with and adapt to digital health care services.

Numerous studies have confirmed the positive effect of satisfaction on behavioral intention [50,51], which aligns with the common understanding that older adults' satisfaction with various aspects of digital health care access influences their behavioral intentions. However, this study also explored whether older adults' internal acceptance of digital health care predominantly determines their satisfaction with these services. This study argues that fundamentally improving digital health care services for older adults must begin with enhancing their willingness to accept and use digital health care platforms and devices. This approach aims to realize the positive, cumulative effects of behavioral intention and satisfaction in a cyclic manner.

Based on the study's results, we propose strategies to enhance the willingness and satisfaction of urban older adults with digital health care services. These recommendations offer valuable insights for optimizing digital medical services in Hangzhou and for the development of similar health care initiatives in other cities.

Improve Accessibility of Digital Medical Information and Assistance

The study results indicate that external variables, such as social environment factors and convenience, are significantly positively correlated with the perceived usefulness and ease of use of digital medical services among older adults. Convenience and related factors notably influence older adults' health care behavior choices [52]. Therefore, it is crucial for various departments to collaborate in expanding and enhancing channels that improve the convenience and accessibility of digital health care knowledge and guidance for older adults. Government departments should leverage both traditional and new media channels to disseminate information about digital health care, enhancing public awareness and trust in government-provided services [53]. This approach can help reduce older adults' perceived risks associated with digital health care, improve their perception of its usefulness, and thereby boost behavioral intentions. Additionally, community organizations should support these efforts by promoting local digital access platforms, acting as an auxiliary force to further engage and assist older adults. Older adults who have used intelligent platform terminals for senior care services should be encouraged to share their experiences and feelings, promoting positive perceptions of these platforms. Communities should organize digital access training sessions to provide more opportunities for older adults to become familiar with and utilize digital platforms in their daily lives. Teaching older adults how to identify beneficial information and basic cybersecurity skills will boost their confidence in using digital medical devices and improve their perceived ease of use [54]. Health care institutions should recruit more trained volunteers to offer digital access support, addressing the reluctance of older adults to use self-service machines due to a lack of timely assistance and guidance. While promoting digital access, it is crucial to accommodate the specific needs of older adults by providing alternative options such as green channels for manual services or one-on-one assistance from volunteers. In addition, fostering "interaction with relatives" and "interaction with friends" is essential, as these individuals can play an irreplaceable role in helping older adults overcome the "health digital divide" [55]. Such measures aim to enhance older adults' perceptions of digital health care through external support, thereby encouraging the use of digital services, sharing the benefits of digital health care, and bridging the "digital divide in health."

Promote Digital Access Platform More Suitable for Older Adults

The most direct aspects of digital medical treatment that users experience are the interactive interface, operation process, and server response speed of apps and self-service machines. The quality of both hardware and software significantly impacts users' ability to effectively utilize digital technologies [56]. Older adults may abandon the use of digital health services if they do not receive timely assistance when issues arise, exacerbating the health digital divide. Departments should focus on distributing resources for digital medical services more equitably and adapt mobile platforms and self-service machines to better meet the needs of older adults. To improve perceived ease of use and reduce perceived risk, the government should facilitate the unification of hospitals onto a single official platform for registration, payment, and other medical activities. At the same time, the credibility of an official platform will also boost older adults' willingness to use digital medical services. Developers of smart terminals should optimize product design to cater to the characteristics of older adults, creating more user-friendly electronic products, apps, and websites. Programs should minimize unnecessary upgrades, be accessible to users with low education levels and older adults, enhance perceived ease of use, and ultimately improve behavioral intention and satisfaction.

Increasing the Acceptance of Digital Health Services

Based on the findings from the questionnaire and interviews, the study identified that the barriers older adults face in digital medical processes are not limited to technical, social, and literacy issues, but also include personal and cognitive factors [1]. Addressing these barriers is essential for bridging the health digital divide for this population. As a result of financial, physical, and ideological factors, older adults often resist digital medical care [48]. Therefore, older adults should actively embrace technological advancements, shift their mindset away from fear and reluctance, and take the initiative to learn and master digital knowledge. By overcoming their hesitance and engaging with digital tools, they can foster a positive change in their ability to use these technologies effectively. Older adults should actively participate in digital training activities to enhance their skills in screening information, managing risks,

XSL•FO
RenderX

and reducing perceived risk. This engagement will help improve their behavioral intentions and increase their satisfaction with digital medical care.

Limitations

This study has several limitations. First, due to constraints related to time and resources, the survey focused solely on older adults aged 60 years and above in urban areas of Hangzhou. Future research should aim to increase the sample size and broaden the geographic scope to encompass a national population, thereby enhancing the representativeness of the findings. Second, while this study focused on 5 key factors—external variables, perceived usefulness, perceived ease of use, perceived risk, and behavioral intention—many other factors could influence willingness and satisfaction with digital medical services. Future studies could explore additional hypotheses and contributing factors to provide a more comprehensive understanding of these variables.

Ongoing advancements in internet technology and aging transformation are expected to shift older adults' digital medical behaviors toward online health management, chronic disease management, online hospitals, and various other health services. This research lays a foundational basis for future studies on a broader range of digital health topics, aiming to help older adults bridge the digital health gap. For example, efforts can be directed toward assisting older adults with mobility impairment, disability, and incapacitation in accessing online medical consultations, purchasing medication, booking home appointments, and enhancing their ability to participate in and benefit from the digital society. Such initiatives are crucial for advancing the concept of active aging and ensuring that older adults globally can fully engage with and benefit from digital advancements.

Conclusions

In summary, this study constructed a structural equation model to elucidate the effects and mechanisms of external variables, perceived usefulness, perceived ease of use, and perceived risk on behavioral intention and satisfaction with digital health care among older adults. To enhance the behavioral intention to use digital health care and increase satisfaction, it is crucial to optimize external influencing factors, improve perceived usefulness and ease of use, and reduce perceived risk. Achieving these goals requires concerted efforts from the government, various sectors of society, and the older adults themselves. By doing so, digital health care services can be better tailored to the needs of older adults, thereby improving their satisfaction, ensuring equitable access to health care services, and bridging the digital health divide.

Acknowledgments

We thank Dr. Le Yang for her linguistic assistance during the preparation of this manuscript. This study would not have been possible without the support of the Hangzhou Normal University and Shanxi Medical University.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study questionnaire.
[DOCX File , 29 KB - jmir_v26i1e48654_app1.docx]

References

1. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. Inquiry 2022;59:469580221096272. [doi: 10.1177/00469580221096272] [Medline: 35471138]
2. Neves BB, Waycott J, Malta S. Old and afraid of new communication technologies? Reconceptualising and contesting the 'age-based digital divide'. Journal of Sociology 2018 Apr 04;54(2):236-248. [doi: 10.1177/1440783318766119]
3. China Internet Network Information Center (CNNIC). The 52nd statistical report on China's internet development. CNNIC. 2023. URL: https://www3.cnnic.cn/n4/2023/0828/c88-10829.html [accessed 2023-11-21]
4. Wang L, He L, Dai K, Zhao L, Xu J, Liu X, et al. In the "internet +" environment, the needs and countermeasures of the information community of urban and rural elderly in Tianjin are home-based. Chinese Journal of Gerontology 2018 Jun 10;38(11):2778-2780. [doi: 10.3969/j.issn.1005-9202.2018.11.083]
5. Lythreatis S, Singh SK, El-Kassar A. The digital divide: a review and future research agenda. Technological Forecasting and Social Change 2022 Feb 01;175:121359. [doi: 10.1016/j.techfore.2021.121359]
6. Yanbin Y, Chengyu M. A review on health digital divide across the world. Documentation, Information & Knowledge 2024 Jul 04;41(3):154-166 [FREE Full text]

XSL•FO
RenderX

7. Choi N. Relationship between health service use and health information technology use among older adults: analysis of the US National Health Interview Survey. J Med Internet Res 2011 Apr 20;13(2):e33 [FREE Full text] [doi: 10.2196/jmir.1753] [Medline: 21752784]
8. Faqih KM, Jaradat MRM. Assessing the moderating effect of gender differences and individualism-collectivism at individual-level on the adoption of mobile commerce technology: TAM3 perspective. Journal of Retailing and Consumer Services 2015 Jan 22;22:37-52. [doi: 10.1016/j.jretconser.2014.09.006]
9. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han H. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018 Aug;36(8):376-382 [FREE Full text] [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
10. Nymberg VM, Bolmsjö BB, Wolff M, Calling S, Gerward S, Sandberg M. 'Having to learn this so late in our lives…' Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care. Scand J Prim Health Care 2019 Mar;37(1):41-52 [FREE Full text] [doi: 10.1080/02813432.2019.1570612] [Medline: 30732519]
11. Zibrik L, Khan S, Bangar N, Stacy E, Novak Lauscher H, Ho K. Patient and community centered eHealth: exploring eHealth barriers and facilitators for chronic disease self-management within British Columbia's immigrant Chinese and Punjabi seniors. Health Policy and Technology 2015 Dec;4(4):348-356. [doi: 10.1016/j.hlpt.2015.08.002]
12. Mishuris RG, Stewart M, Fix GM, Marcello T, McInnes DK, Hogan TP, et al. Barriers to patient portal access among veterans receiving home-based primary care: a qualitative study. Health Expect 2015 Dec;18(6):2296-2305 [FREE Full text] [doi: 10.1111/hex.12199] [Medline: 24816246]
13. Rasche P, Wille M, Bröhl C, Theis S, Schäfer K, Knobe M, et al. Prevalence of Health App Use Among Older Adults in Germany: National Survey. JMIR Mhealth Uhealth 2018 Jan 23;6(1):e26 [FREE Full text] [doi: 10.2196/mhealth.8619] [Medline: 29362211]
14. Whyte WF, Greenwood DJ, Lazes P. Participatory action research. American Behavioral Scientist 2016 Jul 27;32(5):513-551 [FREE Full text] [doi: 10.1177/0002764289032005003]
15. Deng Z, Hong Z, Ren C, Zhang W, Xiang F. What predicts patients' adoption intention toward mHealth services in China: empirical study. JMIR Mhealth Uhealth 2018 Aug 29;6(8):e172 [FREE Full text] [doi: 10.2196/mhealth.9316] [Medline: 30158101]
16. Pywell J, Vijaykumar S, Dodd A, Coventry L. Barriers to older adults' uptake of mobile-based mental health interventions. Digit Health 2020;6:2055207620905422 [FREE Full text] [doi: 10.1177/2055207620905422] [Medline: 32110429]
17. Heart T, Kalderon E. Older adults: are they ready to adopt health-related ICT? Int J Med Inform 2013 Nov;82(11):e209-e231. [doi: 10.1016/j.ijmedinf.2011.03.002] [Medline: 21481631]
18. Liew HP. Healthcare satisfaction among older adults. Am J Health Behav 2018 Jan 01;42(1):99-108. [doi: 10.5993/AJHB.42.1.10] [Medline: 29320343]
19. Yang K, Li Y, Qi H. Determinants of and willingness to use and pay for digital health technologies among the urban elderly in Hangzhou, China. Risk Manag Healthc Policy 2023;16:463-478 [FREE Full text] [doi: 10.2147/RMHP.S393767] [Medline: 37007299]
20. Hangzhou's digital governance index ranks first in China. Hangzhou Daily News. 2020. URL: https://www.hangzhou.gov.cn/art/2020/8/19/art_812262_54590710.html [accessed 2023-10-19]
21. Yan X, Li T. Construction and application of urban digital infrastructure—practice of "Urban Brain" in facing COVID-19 in Hangzhou, China. ECAM 2022 Aug 29;30(8):3123-3141. [doi: 10.1108/ecam-10-2021-0935]
22. Guanghui Z, Jingchao H. Following the "smart governance" to solve the predicaments of grassroots social public opinion work ——reflecting the values and future of China's digital reform from the case of Hangzhou. Zhejiang Social Sciences 2023(01):46-50. [doi: 10.14167/j.zjss.2023.01.019]
23. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
24. Featherman MS, Pavlou PA. Predicting e-services adoption: a perceived risk facets perspective. International Journal of Human-Computer Studies 2003 Oct;59(4):451-474. [doi: 10.1016/S1071-5819(03)00111-3]
25. Kai Z, Kai G, Zhao L. Review on factors affecting of the adoption behavior of mobile library users based on the TAM. Library 2018 Jun 13(06):93-100.
26. Park N, Roman R, Lee S, Chung JE. User acceptance of a digital library system in developing countries: an application of the Technology Acceptance Model. International Journal of Information Management 2009 Jun;29(3):196-209. [doi: 10.1016/j.ijinfomgt.2008.07.001]
27. Jeong H. An investigation of user perceptions and behavioral intentions towards the e-library. Library Collections, Acquisitions, & Technical Services 2013 Dec 03;35(2-3):45-60. [doi: 10.1080/14649055.2011.10766298]
28. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified View. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
29. Martin J, Mortimer G, Andrews L. Re-examining online customer experience to include purchase frequency and perceived risk. Journal of Retailing and Consumer Services 2015 Jul;25:81-95. [doi: 10.1016/j.jretconser.2015.03.008]

XSL•FO
RenderX

30. Lee JK. The roles of individual differences in time perspective, promotion focus, and innovativeness: testing technology acceptance model. Curr Psychol 2022 Nov 17:1-13 (forthcoming) [FREE Full text] [doi: 10.1007/s12144-022-04016-8] [Medline: 36415452]
31. Bauer R. Consumer behavior as risk taking. In: Cox DF, editor. Risk Taking and Information Handling in Consumer Behavior. Boston, MA: Division of Research, Graduate School of Business Administration, Harvard University; 1967:33-34.
32. Gatto SL, Tak SH. Computer, internet, and e-mail use among older adults: benefits and barriers. Educational Gerontology 2008 Aug 22;34(9):800-811. [doi: 10.1080/03601270802243697]
33. Glaser B, Strauss A. Discovery of Grounded Theory Strategies for Qualitative Research. New York, NY: Barney Glaser, Anselm Strauss; 2017.
34. Bowen GA. Naturalistic inquiry and the saturation concept: a research note. Qualitative Research 2008 Feb 01;8(1):137-152. [doi: 10.1177/1468794107085301]
35. Fishbein M, Ajzen I. Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research. Reading, MA: Addison-Wesley; 1975.
36. Pavlou PA. Integrating trust in electronic commerce with the technology acceptance model: model development and validation. International Journal of Human-Computer Studies 2001;59(4):451-474. [doi: 10.1016/s1071-5819(03)00111-3]
37. Sung YP. An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. Journal of Educational Technology & Society 2009;12(3):150-162 [FREE Full text]
38. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Management Science 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
39. Ratna P, Mehra S. Exploring the acceptance for e-learning using technology acceptance model among university students in India. IJPMB 2015;5(2):194-210. [doi: 10.1504/ijpmb.2015.068667]
40. Jimenez J, Del Rio A, Berman AN, Grande M. Personalizing digital health: adapting health technology systems to meet the needs of different older populations. Healthcare (Basel) 2023 Jul 26;11(15):2140 [FREE Full text] [doi: 10.3390/healthcare11152140] [Medline: 37570380]
41. Grewal D, Gotlieb J, Marmorstein H. The moderating effects of message framing and source credibility on the price-perceived risk relationship. J Consum Res 1994 Jun;21(1):145-153. [doi: 10.1086/209388]
42. Mitchell V. Understanding consumers' behaviour: can perceived risk theory help? Management Decision 1992;30(3):13050 [FREE Full text] [doi: 10.1108/00251749210013050]
43. Hider S, Muller S, Gray L, Manning F, Brooks M, Heining D, et al. Digital exclusion as a potential cause of inequalities in access to care: a survey in people with inflammatory rheumatic diseases. Rheumatol Adv Pract 2023;7(1):rkac109 [FREE Full text] [doi: 10.1093/rap/rkac109] [Medline: 36632437]
44. Suggala S, Pathak B, Thomas S. Determinants of customer relationship building on digital healthcare networks: an extension of social media engagement theory from emerging market context. IJEMR 2023;14(2):123. [doi: 10.1504/ijemr.2023.129948]
45. Kendall MGMG. Kendall's advanced theory of statistics. London, UK: Edward Arnold; 2004.
46. Lau RS, Cheung GW. Estimating and comparing specific mediation effects in complex latent variable models. Organizational Research Methods 2010 Dec 03;15(1):3-16. [doi: 10.1177/1094428110391673]
47. Sun X, Yu Z, Xue Z, Zhang X. Research on present situation and service demand of internet combination with home-based treatment-care service among the elderly. Chinese Nursing Research 2020;34(2):318-422. [doi: 10.12102/j.issn.1009-6493.2020.02.029]
48. Jiao L. Study of influencing factors of sustained use behavior of network medical reservation platform users. Medicine and Society 2017;30(10):81-83. [doi: 10.13723/j.yxysh.2017.10.025]
49. Shang S, Wang Y. Continuous use of information behavior by Weibo users based on TAM model. Shanxi Library Journal 2021;21(4):23-31.
50. Zhang Y, Lian N. Research on the influencing factors of institutional intelligent old-age service satisfaction based on the empirical evidence of Baotou City. China Academic Journal Publishing House 2019;191(02):114.
51. Wang F, Ding H. Research on the public's willingness to continue to use mobile government affairs: taking the government service APP as an example. E-Government 2019;12:65-74. [doi: 10.16582/j.cnki.dzzw.2019.12.007]
52. Zhou M, Zhao L, Kong N, Campy KS, Qu S, Wang S. Factors influencing behavior intentions to telehealth by Chinese elderly: an extended TAM model. Int J Med Inform 2019 Jun;126:118-127. [doi: 10.1016/j.ijmedinf.2019.04.001] [Medline: 31029253]
53. Shi Y, Ma D, Zhang J, Chen B. In the digital age: a systematic literature review of the e-health literacy and influencing factors among Chinese older adults. Z Gesundh Wiss 2023;31(5):679-687 [FREE Full text] [doi: 10.1007/s10389-021-01604-z] [Medline: 34104627]
54. Chen Y, Cao M. Discussion on mobile health education for the elderly based on Roy adaptation model health research. Health Research 2018;38(2):148-151.
55. Schreurs K, Quan-Haase A, Martin K. Problematizing the digital literacy paradox in the context of older adults' ICT use: aging, media discourse, and self-determination. Canadian Journal of Communication 2017 May 23;42(2):3130 [FREE Full text] [doi: 10.22230/cjc.2017v42n2a3130]

XSL•FO
RenderX

56. Carney F, Kandt J. Health, out-of-home activities and digital inclusion in later life: implications for emerging mobility services. Journal of Transport & Health 2022 Mar;24:101311. [doi: 10.1016/j.jth.2021.101311]

Abbreviations

mHealth: mobile health
TAM: Technology Acceptance Model

Edited by T de Azevedo Cardoso; submitted 02.05.23; peer-reviewed by X Xiao, J Shi, L Zhao; comments to author 11.10.23; revised version received 21.11.23; accepted 29.07.24; published 20.09.24.

Please cite as:
Wang N, Zhou S, Liu Z, Han Y
Perceptions and Satisfaction With the Use of Digital Medical Services in Urban Older Adults of China: Mixed Methods Study
J Med Internet Res 2024;26:e48654
URL: https://www.jmir.org/2024/1/e48654
doi:10.2196/48654
PMID:

©Ning Wang, Siyu Zhou, Zhuo Liu, Ying Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Associated With the Use of Digital Technology Among Youth in Zimbabwe: Findings From a Cross-Sectional Population-Based Survey

Kevin Martin[1,2], MBBS, MSc; Rachel Wei Chun Peh[1], MSc; Mandikudza Tembo[2,3,4], PhD; Constancia Vimbayi Mavodza[2,5], PhD; Aoife M Doyle[2,3], PhD; Chido Dziva Chikwari[2,3], PhD; Ethel Dauya[2], MPH; Tsitsi Bandason[2], MSc; Steven Azizi[2,3], MSc; Victoria Simms[2,3], PhD; Rashida A Ferrand[1,2], PhD

[1]Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, United Kingdom

[2]The Health Research Unit Zimbabwe, Biomedical Research and Training Institute, Harare, Zimbabwe

[3]MRC International Statistics and Epidemiology Group, London School of Hygiene & Tropical Medicine, London, United Kingdom

[4]Department of Global Health and Development, Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, United Kingdom

[5]Department of Public Health, Environments and Society, London School of Hygiene & Tropical Medicine, London, United Kingdom

Corresponding Author:
Kevin Martin, MBBS, MSc
Department of Clinical Research
London School of Hygiene & Tropical Medicine
Keppel Street
London, WC1E 7HT
United Kingdom
Phone: 44 2076368636
Email: kevin.martin@lshtm.ac.uk

Abstract

Background: Globally, the increasing use of digital technologies such as mobile phones and the internet has allowed for the development of innovative mobile health interventions, particularly for reaching and engaging with youth. However, there is a risk that using such technologies may exclude those who lack access to them.

Objective: In this study, we investigated the sociodemographic factors associated with mobile phone ownership, internet use, and social media use among youth in Zimbabwe.

Methods: A population-based prevalence survey was conducted in 24 urban and periurban communities across 3 provinces of Zimbabwe (Harare, Mashonaland East, and Bulawayo). Youths aged 18 to 24 years resident in randomly selected households in the study communities completed an interviewer-administered questionnaire. The primary outcomes were mobile phone ownership and current internet and social media use. A household wealth indicator was developed using principal components analysis, based on household asset ownership. Multivariable logistic regression was used to investigate the factors associated with each primary outcome. Age, sex, and province were considered a priori confounders. Household wealth, marital status, education level, employment status, time lived at current address, and HIV status were included in the final multivariable model if there was an age-, sex-, and province-adjusted association with a primary outcome on univariable analysis at a significance level of $P<.10$.

Results: Of the 17,636 participants assessed for the primary outcome, 16,370 (92.82%) had access to a mobile phone, and 15,454 (87.63%) owned a mobile phone. Among participants with access to a mobile phone, 58.61% (9594/16,370) and 57.79% (9460/16,370), respectively, used internet and social media at least weekly. Older age (adjusted odds ratio [aOR] 1.76, 95% CI 1.55-2.00), increasing wealth (ranging from aOR 1.85, 95% CI 1.58-2.16, for wealth quintile 2 to aOR 3.80, 95% CI 3.00-4.80, for wealth quintile 5, with quintile 1 as reference), and higher education level (secondary: aOR 1.96, 95% CI 1.60-2.39; tertiary: aOR 8.36, 95% CI 5.29-13.20) were associated with mobile phone ownership. Older age, male sex, increasing wealth, having never been married, higher education level, being in education or formal employment, and having lived at the same address for ≥2 years were associated with higher levels of internet and social media use.

XSL•FO
RenderX

Conclusions: While mobile phone ownership was near-universal, over one-third of youths in urban and periurban settings did not have access to the internet and social media. Access to the internet and social media use were strongly associated with household wealth and education level. Mobile health interventions must ensure that they do not amplify existing inequalities in access to health care. Such interventions must be accompanied by alternative strategies to engage and enroll individuals without internet or social media access to prevent the exclusion of young people by sex and socioeconomic status.

(J Med Internet Res 2024;26:e52670) doi:10.2196/52670

KEYWORDS

Zimbabwe; youth; digital technology; household wealth

Introduction

Background

Mobile phone use is rapidly increasing in many countries in the Global South, where this is expanding faster than other infrastructures [1]. Mobile technology is increasingly being deployed to support health care and improve access to public health interventions, including health promotion and facilitating engagement with and access to particular health services [2]. Mobile health (mHealth) may be implemented through mobile phones, internet access, and social media. mHealth approaches range from simple reminders via text message to improve medication compliance [3], locating study participants or delivery of health messaging through social media or other applications [4,5], to more complex methods such as wearable technologies [6]. This transition to increasing use of digital health technologies has been accelerated by the COVID-19 pandemic, when use of such technologies was essential to deliver public health interventions and maintain routine services that could not be delivered face to face [7].

Digital technologies may be particularly suitable for young people, who have grown up in an age where mobile phone use is largely normalized and where access and availability to the internet is higher than before [1]. However, we must recognize that the distribution of mobile phone ownership, including smart phone ownership, and internet connectivity is not equal, neither across nor within countries [1,8]. Data from the International Telecommunication Union show that in 2022, although almost all young people (aged 15-24 years) in high-income countries had access to the internet, only 67% and 39% in lower–middle-income and low-income countries, respectively, used the internet [9]. The *Lancet* and Financial Times Commission on governing health futures 2030 has also reported that although internet uptake is growing globally, the digital divide between sexes is also widening, which may exacerbate the existing inequalities [1]. Although there has been significant growth in mobile technologies globally, a 2018 survey across 18 “advanced” economies and 9 “emerging” economies found that a median of 76% and 45% of respondents within each category owned a smartphone, respectively [8].

Disparities in mobile phone use are also mirrored in the mHealth literature. Although the number of studies assessing the effectiveness of mHealth interventions in promoting health outcomes or modifying behavior is growing globally, with an overall positive albeit weak effect demonstrated, data are still very sparse for the Global South [10,11]. A 2021 systematic review of studies assessing mHealth apps for health promotion and disease management found 172 studies meeting their inclusion criteria. However, only 8 studies were performed outside of high-income settings [10]. One particular example of an mHealth intervention in Zimbabwe is a digital mental health intervention called “Inuka,” delivered via a chat-based application, which demonstrated high levels of acceptability, feasibility, and appropriateness [12]. However, the barriers to the use of the app and to potential integration into routine care, did include challenges with connectivity, which negatively affected some of the clients’ experiences. Given the poorer connectivity in high-density suburbs and rural areas than in low-density suburbs and urban areas, the authors noted that this would affect the feasibility of delivering an equitable rollout of such an intervention in this setting [13].

Due to the heterogeneity in the distribution of mobile technology and of ownership and use across settings, it is critical to understand the context and digital landscape before digital technologies are deployed, as there is potential for these to amplify inequalities [12]. Exploring access to mobile phone ownership, internet access, and social media use is necessary to understand who would and would not potentially benefit from particular mHealth interventions. Furthermore, it would allow for the barriers to be preempted and for alternate strategies to be developed. This is particularly pertinent for Zimbabwe, where data on technology use are limited. The 2015 Zimbabwe Demographic and Health Survey reported that 87% of households owned a mobile phone [14]. However, this does not necessarily mean that young people within these households have access. This is demonstrated by Doyle et al [15] who conducted a relatively small survey of young people aged 13 to 24 years in urban communities in Zimbabwe and found that 63% of the young people surveyed owned a mobile phone and 65% had access to the internet [15]. More data are needed to confirm these findings and to further explore factors associated with technology use.

Objective

The objectives of this study were to investigate the levels of mobile phone ownership, internet use, and social media use and to understand the factors associated with mobile phone ownership, internet use, and social media use among youth in Zimbabwe.

Methods

Study Design and Setting

The study used data from a cross-sectional population-based survey conducted to ascertain the outcome of a cluster

XSL•FO
RenderX

randomized trial investigating the impact of providing a community-based package of integrated HIV and sexual and reproductive health services on population-level HIV outcomes. The study protocol with details of the intervention and the survey has been previously published [16].

Briefly, the 2-arm CHIEDZA trial (*Community-based interventions to improve HIV outcomes in youth: a cluster randomised trial in Zimbabwe*; trial registration NCT03719521) was conducted in 24 communities in 3 provinces (Harare, Mashonaland East and Bulawayo), with each province containing 8 clusters randomized 1:1 to 4 intervention and 4 standard of care (routine, existing services) clusters. A *cluster* was defined as a geographically demarcated area in a community containing a primary health care clinic and a community center from which services are delivered. The 3 provinces were chosen to include areas representing both main ethnic groups (Bulawayo is predominantly Ndebele, and Harare and Mashonaland East are predominantly Shona) and to include both urban and periurban settings. Rural settings were excluded, as population density was too low to make the delivery of the CHIEDZA intervention feasible [16]. Individuals aged 16 to 24 years living within an intervention cluster were eligible to receive a package of integrated HIV and sexual and reproductive health services delivered from the intervention cluster community center. This package included HIV testing, HIV treatment and adherence support, contraception, pregnancy testing, syndromic management of sexually transmitted infections, menstrual health information and products, condoms, and general health counseling in addition to existing health services (standard of care). Service delivery was accompanied by peer outreach to promote CHIEDZA and engage youth to achieve high coverage. Outreach teams consisted of 16-to 24-year-old cluster residents who had previously engaged with CHIEDZA. Outreach activities included flyer distribution, information dissemination, and in-field live demonstrations of CHIEDZA products (such as reusable pads, menstrual cups, and condoms), alongside door-to-door sensitization within the cluster [16].

Study Procedures

The postintervention prevalence survey was conducted after a 30-month intervention period, between October 2021 and June 2022, in both intervention and control clusters. Mapping of the streets within clusters was conducted by checking the OpenStreetMap road network against satellite imagery to ensure all streets were mapped. The streets were manually split into segments with ArcGIS software (Esri) using either junctions or features such as school grounds. Each segment was assigned a number, and segments were randomly selected. Following community sensitization, all *households* (defined as a person or group of related or unrelated persons who live together in the same dwelling or units of a dwelling, who acknowledge 1 adult male or female as the head of the household, who share the same housekeeping arrangements, and who are considered a single unit) in each dwelling in the randomly selected street segments were enumerated. All individuals aged 18 to 24 years residing in the enumerated households were eligible to participate. If a potentially eligible individual was not available at the time of enumeration, up to 3 repeat visits were made to enroll the individual. Following consent, data were collected using an interviewer-administered questionnaire, using SurveyCTO (Dobility Inc) on tablets. The survey was piloted in youth before being used in the prevalence survey. The relevant survey questions and response categories for this study are shown in Multimedia Appendix 1.

Outcomes and Statistical Analysis

Sample size calculations for the survey were based on having sufficient power to ascertain the primary outcome of the CHIEDZA trial and are described in the trial protocol [16]. STATA (version 17.0; StataCorp) was used for data analysis. The primary outcomes in this study were mobile phone ownership, internet use, and social media use. *Mobile phone ownership* was defined as participants who have their own mobile phone, therefore excluding those who share a phone. *Internet use* was defined as participants who use the internet at least once per week on any device, for any reason (including social media). *Social media use* was defined as use of at least one of Twitter, Facebook, Instagram, TikTok, LinkedIn, YouTube, or Snapchat. WhatsApp was excluded from the internet use and social media definitions, as its function is largely similar to text messaging. In Zimbabwe, stand-alone WhatsApp data bundles that exclude internet use are much more affordable than either an internet connection or other data packages. Access to both the internet and social media will be largely influenced by access to a mobile phone [15]. As a result, internet use and social media use were only examined in participants who either owned or had access to a mobile phone. Secondary outcome variables included mobile phone functions (call, text, WhatsApp, internet browsing); frequency of internet used; internet access point; and types of social media platform used.

Principal component analysis was used to develop a household wealth indicator, based on the presence of 6 functioning household assets (fridge, bicycle, car or truck, television, radio, microwave; Multimedia Appendix 2). Univariable logistic regression was conducted to examine associations between variables and each primary outcome. For multivariable logistic regression, age, sex, and province were considered as a priori confounders. Other variables with an age-, sex-, and province-adjusted association with a primary outcome with a P value of <.10 were included in the final multivariable model. Clustering was adjusted for at the street segment level.

Ethical Considerations

Ethical approval was obtained from the Medical Research Council of Zimbabwe (MRCZ/A/2387), the Biomedical Research and Training Institute Institutional Review Board (AP149/2018), and the London School of Hygiene and Tropical Medicine Ethics Committee (16,948). Consent was documented electronically on a tablet, with participants retaining a signed paper copy for their records.

Results

Participant Characteristics

The CONSORT (Consolidated Standards for Reporting Trials) diagram is shown in Figure 1. Of the 18,721 enumerated individuals, 17,682 (94.45%) were enrolled, of whom 17,636

XSL•FO
RenderX

(99.71%) were assessed for the primary outcome, with approximately one-third recruited from each province. Overall, participants had a mean age of 20.6 (SD 2.1) years, and 60.82% (n=10,726) were female. Three-quarters of the participants (n=13,279, 75.29%) had never been married, almost half (8711/17,636, 49.39%) were neither in employment nor education, and 87.11% (15,363/17,636) of the participants' highest education level was secondary (Table 1).

Figure 1. CONSORT (Consolidated Standards for Reporting Trials) diagram showing study recruitment of youth aged 18 to 24 years in a population-based survey in Zimbabwe to ascertain the outcome of a cluster randomized trial (community-based interventions to improve HIV outcomes in youth: a cluster randomised trial in Zimbabwe [CHIEDZA]).

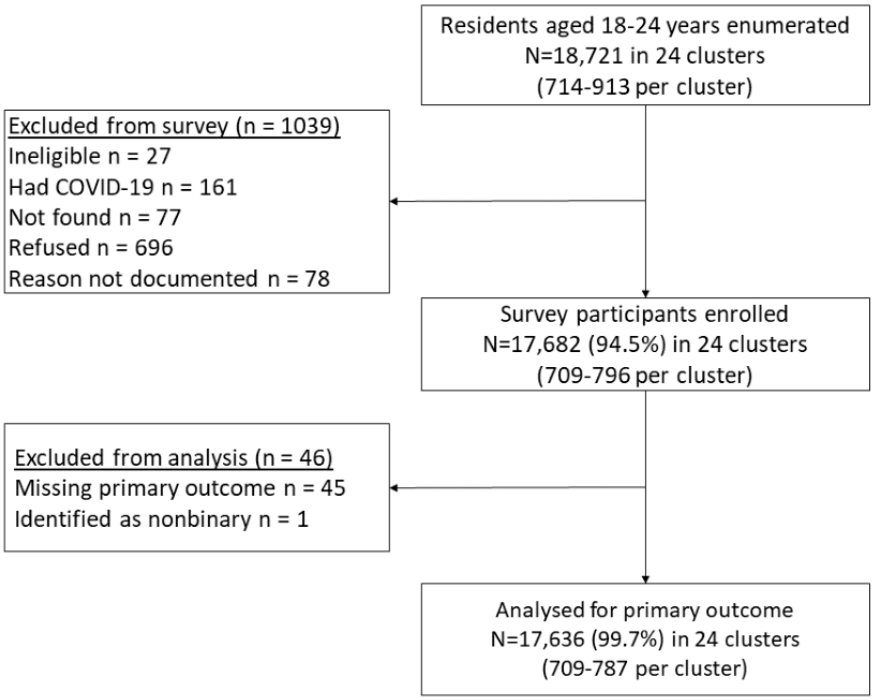

XSL•FO
RenderX

Table 1. Characteristics of youth in Zimbabwe who were recruited into the population-based survey (n=17,636 unless otherwise stated).

Variable	Value, n (%)
Sex	
Female	10,726 (60.82)
Male	6910 (39.18)
Age (y)	
18-20	9239 (52.39)
21-24	8397 (47.61)
Province	
Harare	5845 (33.14)
Bulawayo	5929 (33.62)
Mashonaland East	5862 (33.24)
Household wealth index (n=17 , 628)	
1 (poorest)	4161 (23.6)
2	3430 (19.46)
3	3658 (20.75)
4	3664 (20.79)
5 (richest)	2715 (15.4)
Marital status	
Never married	13,279 (75.29)
Married or living together as if married	3559 (20.18)
Divorced, widowed, or separated (and currently unmarried)	798 (4.52)
Highest completed education level	
None or primary level	867 (4.92)
Form 1-4	12,638 (71.66)
Form 5-6	2725 (15.45)
Tertiary level	1406 (7.97)
Employment status	
In education	4962 (28.14)
Formal employment	832 (4.72)
Informal employment	3151 (17.87)
Not in education or employment	8711 (49.39)
Time lived at current address (years)	
<1	4241 (24.05)
1-2	1701 (9.65)
2-3	1685 (9.55)
>3	10,009 (56.75)
Previous residence	
Always lived at this address	6866 (38.93)
Same suburb but different address	3747 (21.25)
Same town or city but different suburb	3748 (21.25)
Lived outside this town or city	3275 (18.57)
Average monthly household income (US $)	
<50	2616 (14.83)

XSL•FO
RenderX

Variable	Value, n (%)
50-100	4457 (25.27)
101-200	4740 (26.88)
201-500	2697 (15.29)
501-900	436 (2.47)
>900	157 (0.89)
Don't know or don't want to say	2533 (14.36)
HIV status (n=17 , 503)	
Positive	1190 (6.8)
Negative	16,313 (93.2)

Use of Digital Technologies

Of the 17,636 participants, 16,370 (92.82%) had access to a mobile phone, with 15,454 (87.63%) owning a mobile phone, 842 (4.77%) sharing a phone with a family member, and 74 (0.42%) sharing a phone with someone other than a family member. Among mobile phone owners, 98.89% (15,282/15,454), 97.75% (15,106/15,454), 86.51% (13,369/15,454), and 63.55% (9821/15,454) of mobile phones were equipped to call, text, use WhatsApp, and browse the internet, respectively.

Among participants (n=16,370) with access to a mobile phone, 58.61% (n=9594) reported internet use. Frequency of internet use ranged from never (n=6776, 41.39%), once or twice a week (n=4004, 24.46%), most days of the week (n=1969, 12.03%), every day (n=3086, 18.85%), to several times per day (n=535, 3.27%). Among internet users, this was most often on a personal phone (9039/9594, 94.22%), as opposed to on someone else's phone or using a different type of device.

Among participants with mobile phone access (n=16,370), social media use was 57.79% (n=9460). The most accessed types of social media, in descending order, were Facebook (n=8785, 53.67%), Instagram (n=4396, 26.85%), YouTube (n=4315, 26.36%), TikTok (n=2857, 17.45%), Twitter (n=2257, 13.79%), and LinkedIn (n=583, 3.56%). WhatsApp was used regularly by 85.85% (n=14,054) of participants with mobile access.

Among participants who owned a WhatsApp-enabled phone, 98.62% (13,185/13,369) accessed WhatsApp regularly (Table S1 in Multimedia Appendix 3). For participants who owned an internet-enabled phone, 85.17% (8365/9821) reported internet use (Table S2 in Multimedia Appendix 3), in comparison to 17.34% (977/5633) of those who owned a mobile phone not equipped to access the internet.

Factors Associated With Mobile Phone Ownership, Internet Use, and Social Media Use

Univariable associations with mobile phone ownership, internet use, and social media use are shown in Tables 2, 3, and 4, respectively. Multimedia Appendix 4 also shows each primary outcome by age, education level, and duration of residence, disaggregated to a higher degree than shown in Tables 2-4.

Higher wealth index, being unmarried, higher education level, being in education or formal employment, having lived at their current address for at least 2 years, having always lived at their address, not living with HIV, and being in Harare (compared to Mashonaland East) were all associated with higher levels of mobile phone ownership, internet use, and social media use. Being aged 21 to 24 years (compared to 18-20 years) was associated with mobile phone ownership and internet use, but not with social media use, whereas male sex was associated with internet and social media use, but not with mobile phone ownership.

In multivariable analysis, mobile phone ownership was associated with being aged 21 to 24 (adjusted odds ratio [aOR] 1.76, 95% CI 1.55-2.00), being in Harare (capital city) compared to both Bulawayo (aOR 0.88, 95% CI 0.76-1.01) and Mashonaland East (aOR 0.81, 95% CI 0.71-0.96), increasing household wealth, and increased education level (Table 2).

Internet use was associated with being aged 21 to 24 years (aOR 1.21, 95% CI 1.11-1.33), male sex (aOR 1.43, 95% CI 1.30-1.57), increasing household wealth, having never been married, increasing education level, being in education (aOR 2.19, 95% CI 1.97-2.44) or formal employment (aOR 1.55, 95% CI 1.26-1.89), having lived at their current address for >2 years (aOR 1.35, 95% CI 1.21-1.51), and having always lived at their current address (Table 3).

Social media use was associated with being aged 21 to 24 years (aOR 1.39, 95% CI 1.27-1.52), male sex (aOR 1.41, 95% CI 1.29-1.55), being in Harare compared to both Bulawayo (aOR 0.76, 95% CI 0.67-0.85) and Mashonaland East (aOR 0.82, 95% CI 0.72-0.92), increasing household wealth, having never been married, increasing education level, being in education (aOR 1.97, 95% CI 1.77-2.20) or formal employment (aOR 1.62, 95% CI 1.35-1.95), having lived at their current address for >2 years (aOR 1.21, 95% CI 1.09-1.35), and having always lived at their current address (Table 4).

XSL•FO
RenderX

Table 2. Univariable and multivariable associations between exposure variables and mobile phone ownership among youth recruited into a population-based survey in Zimbabwe (n=17,636).

Variable	Mobile phone ownership, n (%)	Unadjusted OR[a] (95% CI)	Age-, sex-, and province-adjusted OR (95% CI)	Final adjusted OR (95% CI)[b]
Age (y)		*P*<.001	*P*<.001	*P*<.001
18-20	7841/9239 (84.87)	1.00	1.00	1.00
21-24	7613/8397 (90.66)	1.77 (1.59-1.97)	1.78 (1.60-1.99)	1.76 (1.55-2.00)
Sex		*P*=.24	*P*=.09	*P*=.10
Female	9364/10,726 (87.3)	1.00	1.00	1.00
Male	6090/6910 (88.13)	1.08 (0.95-1.22)	1.11 (0.98-1.26)	0.90 (0.79-1.02)
Province		*P*<.001	*P*<.001	*P*=.03
Harare	5227/5845 (89.42)	1.00	1.00	1.00
Bulawayo	5231/5929 (88.23)	0.92 (0.79-1.06)	0.95 (0.82-1.09)	0.88 (0.76-1.01)
Mashonaland East	4996/5862 (85.23)	0.66 (0.56-0.77)	0.67 (0.57-0.79)	0.83 (0.71-0.96)
Household wealth index		*P*<.001	*P*<.001	*P*<.001
1 (poorest)	3274/4161 (78.68)	1.00	1.00	1.00
2	2994/3430 (87.29)	1.97 (1.69-2.30)	2.00 (1.71-2.33)	1.85 (1.58-2.16)
3	3267/3658 (89.31)	2.51 (2.11-2.98)	2.57 (2.16-3.05)	2.30 (1.92-2.75)
4	3371/3664 (92)	3.36 (2.82-4.00)	3.49 (2.93-4.16)	3.04 (2.53-3.66)
5 (richest)	2541/2715 (93.59)	4.46 (3.57-5.56)	4.64 (3.71-5.80)	3.80 (3.00-4.80)
Marital status		*P*<.001	*P*<.001	*P*=.47
Never married	11,737/13,279 (88.39)	1.00	1.00	1.00
Married or living together as if married	3011/3559 (84.6)	0.77 (0.68-0.87)	0.61 (0.52-0.71)	0.93 (0.78-1.11)
Divorced, widowed, or separated (and currently unmarried)	706/798 (88.47)	1.01 (0.78-1.32)	0.80 (0.61-1.06)	1.10 (0.82-1.47)
Highest completed education level		*P*<.001	*P*<.001	*P*<.001
None or primary level	647/867 (74.63)	1.00	1.00	1.00
Secondary level	13,432/15,363 (87.43)	2.26 (1.87-2.72)	2.36 (1.94-2.86)	1.96 (1.60-2.39)
Tertiary level	1375/1406 (97.8)	17.3 (11.00-27.06)	14.98 (9.52-23.58)	8.36 (5.29-13.20)
Employment status		*P*<.001	*P*<.001	*P*<.001
Not in education or employment	7372/8711 (84.63)	1.00	1.00	1.00
Informal employment	2831/3151 (89.84)	1.71 (1.48-1.99)	1.58 (1.34-1.85)	1.54 (1.31-1.81)
In education	4462/4942 (90.29)	1.63 (1.43-1.87)	1.81 (1.57-2.08)	1.20 (1.03-1.39)
Formal employment	789/832 (94.83)	3.64 (2.62-5.05)	3.13 (2.24-4.37)	2.30 (1.65-3.21)
Time lived at current address (y)		*P*<.001	*P*=.001	*P*=.58
<2	5099/5942 (85.81)	1.00	1.00	1.00
>2	10,355/11,694 (88.55)	1.28 (1.15-1.42)	1.24 (1.11-1.38)	1.04 (0.91-1.17)
Previous residence		*P*<.001	*P*<.001	*P*=.01
Always lived at this address	6119/6866 (89.12)	1.00	1.00	1.00
Same suburb but different address	3260/3747 (87)	0.81 (0.70-0.94)	0.81 (0.70-0.94)	1.01 (0.86-1.19)
Same town or city but different suburb	3295/3748 (87.91)	0.92 (0.78-1.08)	0.93 (0.79-1.09)	1.11 (0.93-1.32)
Lived outside this town or city	2780/3275 (84.89)	0.67 (0.58-0.77)	0.69 (0.60-0.81)	0.83 (0.70-0.98)
HIV status		*P*=.02	*P*=.004	*P*=.17

XSL•FO
RenderX

Variable	Mobile phone ownership, n (%)	Unadjusted OR[a] (95% CI)	Age-, sex-, and province-adjusted OR (95% CI)	Final adjusted OR (95% CI)[b]
Negative	14,327/16,313 (87.83)	1.00	1.00	1.00
Positive	1007/1190 (84.62)	0.79 (0.64-0.96)	0.73 (0.59-0.90)	0.85 (0.68-1.07)

[a]OR: odds ratio.

[b]Adjusted for age, sex, province, household wealth, marital status, highest completed education level, employment status, and time lived at current address and previous residence.

XSL•FO
RenderX

Table 3. Univariable and multivariable associations between exposure variables and internet use among youth recruited into a population-based survey in Zimbabwe (n=16,370).

Variable	Internet use, n (%)	Unadjusted OR[a] (95% CI)	Age-, sex-, and province-adjusted OR (95% CI)	Final adjusted OR (95% CI)[b]
Age (y)		*P*=.002	*P*=.07	*P*<.001
18-20	5031/8423 (59.73)	1.00	1.00	1.00
21-24	4563/7947 (57.42)	0.89 (0.83-0.96)	0.93 (0.87-1.01)	1.21 (1.11-1.33)
Sex		*P*<.001	*P*<.001	*P*<.001
Female	5138/9978 (51.49)	1.00	1.00	1.00
Male	4456/6392 (69.71)	2.05 (1.89-2.24)	2.05 (1.89-2.23)	1.43 (1.30-1.57)
Province		*P*<.001	*P*<.001	*P*=.06
Harare	3189/5557 (57.39)	1.00	1.00	1.00
Bulawayo	3548/5554 (63.88)	1.26 (1.11-1.43)	1.18 (1.04-1.34)	0.95 (0.83-1.09)
Mashonaland East	2857/5259 (54.33)	0.77 (0.66-0.89)	0.73 (0.62-0.85)	0.84 (0.72-0.97)
Household wealth index		*P*<.001	*P*<.001	*P*<.001
1 (poorest)	1367/3622 (37.74)	1.00	1.00	1.00
2	1677/3196 (52.47)	1.92 (1.68-2.18)	1.81 (1.58-2.06)	1.53 (1.33-1.75)
3	2152/3438 (62.59)	2.88 (2.52-3.28)	2.63 (2.31-3.00)	2.02 (1.76-2.32)
4	2452/3497 (70.12)	4.19 (3.65-4.82)	3.84 (3.35-4.40)	2.64 (2.30-3.04)
5 (richest)	1943/2610 (74.44)	5.23 (4.44-6.16)	4.68 (3.98-5.49)	2.78 (2.33-3.30)
Marital status		*P*<.001	*P*<.001	*P*<.001
Never married	8098/12,322 (65.72)	1.00	1.00	1.00
Married or living together as if married	1215/3309 (36.72)	0.28 (0.26-0.32)	0.32 (0.29-0.36)	0.56 (0.50-0.63)
Divorced, widowed, or separated (and currently unmarried)	281/739 (38.02)	0.33 (0.27-0.39)	0.37 (0.31-0.44)	0.60 (0.50-0.74)
Highest completed education level		*P*<.001	*P*<.001	*P*<.001
None or primary level	200/728 (27.47)	1.00	1.00	1.00
Secondary level	8169/14,248 (57.33)	3.76 (3.07-4.62)	3.58 (2.89-4.43)	2.63 (2.14-3.22)
Tertiary level	1225/1394 (87.88)	19.19 (14.17-25.99)	19.30 (14.15-26.34)	7.47 (5.45-10.22)
Employment status		*P*<.001	*P*<.001	*P*<.001
Not in education or employment	3827/7917 (48.34)	1.00	1.00	1.00
Informal employment	1544/2975 (51.9)	1.12 (1.00-1.26)	1.00 (0.89-1.13)	0.99 (0.87-1.12)
In education	3656/4668 (78.32)	3.85 (3.46-4.28)	3.61 (3.24-4.01)	2.19 (1.97-2.44)
Formal employment	567/810 (70)	2.47 (2.02-3.03)	2.07 (1.70-2.53)	1.55 (1.26-1.89)
Time lived at current address (y)		*P*<.001	*P*<.001	*P*<.001
<2	2468/5473 (45.09)	1.00	1.00	1.00
>2	7126/10,897 (65.39)	2.17 (1.98-2.37)	1.89 (1.72-2.07)	1.35 (1.21-1.51)
Previous residence		*P*<.001	*P*<.001	*P*<.001
Always lived at this address	4406/6374 (69.12)	1.00	1.00	1.00
Same suburb but different address	1867/3480 (53.65)	0.52 (0.46-0.59)	0.60 (0.54-0.68)	0.95 (0.82-1.09)
Same town or city but different suburb	1902/3521 (54.02)	0.55 (0.49-0.63)	0.64 (0.56-0.72)	0.92 (0.79-1.07)
Lived outside this town or city	1419/2995 (47.38)	0.43 (0.38-0.48)	0.49 (0.43-0.55)	0.70 (0.61-0.80)
HIV status		*P*=.002	*P*=.03	*P*=.78
Negative	8952/15,168 (59.02)	1.00	1.00	1.00
Positive	568/1076 (52.79)	0.75 (0.63-0.90)	0.82 (0.68-0.98)	0.97 (0.79-1.19)

XSL•FO
RenderX

[a]OR: odds ratio.

[b]Adjusted for age, sex, province, household wealth, marital status, highest completed education level, employment status, and time lived at current address and previous residence.

XSL•FO
RenderX

Table 4. Univariable and multivariable associations between exposure variables and social media use among youth recruited into a population-based survey in Zimbabwe (n=16,370).

Variable	Social media use, n (%)	Unadjusted OR[a] (95% CI)	Age-, sex-, and province-adjusted OR (95% CI)	Final adjusted OR (95% CI)[b]
Age (y)		*P*=.86	*P*=.19	*P*<.001
18-20	4843/8423 (57.5)	1.00	1.00	1.00
21-24	4617/7947 (58.1)	1.01 (0.94-1.08)	1.05 (0.98-1.13)	1.39 (1.27-1.52)
Sex		*P*<.001	*P*<.001	*P*<.001
Female	5078/9978 (50.89)	1.00	1.00	1.00
Male	4382/6392 (68.55)	2.03 (1.87-2.21)	2.05 (1.89-2.24)	1.41 (1.29-1.55)
Province		*P*<.001	*P*<.001	*P*<.001
Harare	3259/5557 (58.65)	1.00	1.00	1.00
Bulawayo	3349/5554 (60.3)	1.03 (0.92-1.16)	0.97 (0.87-1.09)	0.76 (0.67-0.85)
Mashonaland East	2852/5259 (54.23)	0.75 (0.65-0.86)	0.71 (0.62-0.82)	0.82 (0.72-0.92)
Household wealth index		*P*<.001	*P*<.001	*P*<.001
1 (poorest)	1366/3622 (37.71)	1.00	1.00	1.00
2	1631/3196 (51.03)	1.86 (1.65-2.11)	1.82 (1.61-2.05)	1.51 (1.34-1.72)
3	2053/3438 (59.71)	2.60 (2.30-2.94)	2.49 (2.21-2.81)	1.89 (1.67-2.14)
4	2402/3497 (68.69)	3.81 (3.33-4.35)	3.68 (3.24-4.17)	2.51 (2.21-2.85)
5 (richest)	2006/2610 (76.86)	6.04 (5.17-7.06)	5.71 (4.90-6.65)	3.45 (2.95-4.03)
Marital status		*P*<.001	*P*<.001	*P*<.001
Never married	7973/12,322 (64.71)	1.00	1.00	1.00
Married or living together as if married	1200/3309 (36.26)	0.29 (0.26-0.32)	0.29 (0.26-0.32)	0.50 (0.44-0.56)
Divorced, widowed, or separated (and currently unmarried)	287/739 (38.84)	0.37 (0.31-0.43)	0.37 (0.31-0.45)	0.60 (0.50-0.71)
Highest completed education level		*P*<.001	*P*<.001	*P*<.001
None or primary level	192/728 (26.37)	1.00	1.00	1.00
Secondary level	8042/14,248 (56.44)	3.75 (2.89-4.86)	3.56 (2.72-4.66)	2.59 (2.03-3.30)
Tertiary level	1226/1394 (87.95)	22.49 (16.27-31.08)	21.62 (15.48-30.19)	8.38 (6.06-11.57)
Employment status		*P*<.001	*P*<.001	*P*<.001
Not in education or employment	3792/7917 (47.9)	1.00	1.00	1.00
Informal employment	1536/2975 (51.63)	1.19 (1.07-1.31)	1.02 (0.92-1.14)	1.01 (0.91-1.13)
In education	3557/4668 (76.2)	3.49 (3.15-3.87)	3.34 (3.00-3.71)	1.97 (1.77-2.20)
Formal employment	575/810 (70.99)	2.71 (2.26-3.25)	2.21 (1.85-2.65)	1.62 (1.35-1.95)
Time lived at current address (y)		*P*<.001	*P*<.001	*P*<.001
<2	2499/5473 (45.66)	1.00	1.00	1.00
>2	6961/10,897 (63.88)	2.01 (1.85-2.18)	1.79 (1.65-1.95)	1.21 (1.09-1.35)
Previous residence		*P*<.001	*P*<.001	*P*<.001
Always lived at this address	4340/6374 (68.09)	1.00	1.00	1.00
Same suburb but different address	1848/3480 (53.1)	0.53 (0.47-0.60)	0.59 (0.53-0.66)	0.88 (0.77-1.01)
Same town or city but different suburb	1903/3521 (54.05)	0.58 (0.52-0.65)	0.65 (0.57-0.73)	0.89 (0.77-1.02)
Lived outside this town or city	1359/2995 (45.38)	0.41 (0.36-0.46)	0.45 (0.40-0.51)	0.59 (0.52-0.67)
HIV status		*P*=.002	*P*=.02	*P*=.55

XSL•FO
RenderX

Variable	Social media use, n (%)	Unadjusted OR[a] (95% CI)	Age-, sex-, and province-adjusted OR (95% CI)	Final adjusted OR (95% CI)[b]
Negative	8830/15,168 (58.21)	1.00	1.00	1.00
Positive	555/1076 (51.58)	0.75 (0.62-0.90)	0.80 (0.66-0.96)	0.94 (0.76-1.16)

[a]OR: odds ratio.

[b]Adjusted for age, sex, province, household wealth, marital status, highest completed education level, employment status, and time lived at current address and previous residence.

Relationships Between Mobile Phone Ownership, Internet Use, and Social Media Use

Individuals who owned a mobile phone had a 4.17 (95% CI 3.51-4.94; *P*<.001) times increase in the odds of using the internet (9342/15,454, 60.45%) compared to individuals who did not own a mobile phone (252/916, 27.51%). Similarly, individuals who owned a mobile phone had a 4.42 (95% CI 3.69-5.30; *P*<.001) times increase in the odds of using any social media (9227/15,454, 59.71%) compared to individuals who did not own a mobile phone (233/916, 25.44%).

Discussion

Principal Findings

Overall, this study demonstrates high levels of ownership of mobile phones among young people aged 18 to 24 years in urban and periurban areas of Zimbabwe. Nearly all these phones had both call and text functions, while 86.51% had WhatsApp and 63.55% had access to internet browsing on their device. This has important ramifications for the potential modalities for digital technology interventions, with the need for internet browsing likely to exclude a large proportion of potential users.

Where low levels of internet access are problematic, WhatsApp may potentially offer a compromise. Although WhatsApp may have reduced accessibility to young people compared to text messaging, this may be offset by less-stringent character-limit restrictions and reduced cost of message delivery if a project has Wi-Fi access, which may be more important considerations for an intervention. This may be particularly relevant where 2-way messaging with youth is required. However, if only “out” messaging to youth is needed, then text messaging may allow access to more individuals.

There were strong associations between household wealth and mobile phone ownership, internet use, and social media use, even after adjusting for a range of potential confounders. This highlights that the use of digital technologies may likely exacerbate inequality along socioeconomic lines. As a result, if these technologies are utilized, implementers must provide alternate strategies for individuals who do not have access. This could include loaning such digital technologies to individuals without access [17] or offering a nondigital alternative for engagement. In particular, choice of technology is crucial. For example, only a fifth of participants (3621/16,370, 22.12%) with a mobile phone accessed the internet every day or more often. Any technologies requiring regular internet access would therefore only be appropriate for this small subset of individuals.

In addition, although there was no association between sex and mobile phone ownership, male participants were more likely to use the internet or social media. Possible reasons include such technologies being considered “male” and being more socially acceptable for men to access. Future qualitative work is needed to explore this further. Importantly, the sex imbalance may indicate that such technologies are particularly appropriate to engage men, who are often considered “difficult” to reach and retain in services, particularly in relation to the HIV care cascade [18]. However, counter to this is that deployment of mHealth interventions may risk women being disenfranchised, unless provision can be made for this eventuality. Another interesting finding was the strong association between being married or previously married and not accessing the internet or social media. No association was found between marital status and mobile phone ownership. Social media use within marriage may affect quality of offline relationships, so married individuals may make a more active effort to disengage from social media [19]. Conversely, use of the internet and social media may also be seen as a means to seek out new relationships, perhaps contributing to higher levels of use in unmarried individuals.

In a survey of 634 young people aged 13 to 24 years in urban and periurban communities in Zimbabwe, Doyle et al [15] reported that 63% owned a phone, with a further 4.3% having access to a shared phone. Disaggregated by age, Doyle et al [15] found that 71.5% of the individuals aged 18 to 19 years and 84.7% of those aged 20 to 24 years owned a mobile phone. Our survey found higher proportions of mobile phone ownership of 83.22% (5657/6798) and 90.39% (9797/10,838) in these age groups, respectively. However, internet access was reported by 73.6% of individuals aged 18 to 19 years and 77.2% of those aged 20 to 24 years in the study by Doyle et al [15], in comparison to only 59.34% (3636/6127) and 58.17% (5958/10,243) in this study, respectively. This is despite this variable being restricted to individuals with access to a mobile phone in our study. Given that Doyle et al [15] collected data in 2018, it is expected that a higher proportion of individuals in these comparable age groups would own a mobile phone, due to increasing levels of access globally. However, it is less clear why there would be reduced levels of internet use. Possible reasons include variation between selected communities in both studies or differing case definitions. For example, the question related to internet use in our survey considered use generally, with options including never, once or twice a week, most days of the week, every day, or several times a day. However, Doyle et al [15] defined *internet access* as “if they reported accessing the internet once or more in the last 3 months, including on a device belonging to a family member or employer.” This broader definition may therefore have included some individuals who

XSL•FO
RenderX

only access the internet occasionally in the Doyle et al [15] survey, but not our survey. Across both studies, mobile phone ownership was associated with increased age and higher education level, among young people in Zimbabwe, but not sex [15]. Additionally, both studies found that internet access was associated with male sex, increased age, higher education levels, and longer duration of residence at their current address [15].

Importantly, the strong associations between mobile phone ownership and internet or social media use suggest that mobile phones are a key entry point for use of other digital technologies. Similarly, among internet users, the vast majority of internet access was through a personal mobile phone. Together, this suggests that increasing smart phone ownership may allow for increased access to not only basic phone functions such as text and calls but also the internet. This may be a more cost-effective strategy to promote access to digital technologies compared to the provision of tablets or computers. It is also important to note that some digital interventions will simply not be deliverable by text or phone call, such as those involving videos, music, or peer-led social media communities [20-22]. However, access to technology alone is not sufficient. For example, 14.83% (1456/9821) of those who owned mobile phones capable of accessing the internet did not regularly use the internet. Identification of the additional barriers faced by these individuals in accessing the internet is key, determining if this was an active choice or if other factors such as cost of data were prohibitive. There is also a need to go beyond simply investigating where disparities exist. There needs to be an active effort to address any disparities in order to build digital readiness at a community level [23]. This will require significant investment in local, national, and international infrastructure, alongside lowering the individual barriers to access such as the cost of hardware and internet access [15,23]. Economic evaluations on the potential benefit of subsidies for smartphones or communal Wi-Fi networks would be prudent to consider whether such interventions may be cost-effective. One must also consider the possibility that digital technologies may not be the way forward in all settings. In addition to the expense required to upgrade current infrastructure, it may not be acceptable or feasible to local populations. Within the CHIEDZA trial, a substudy assessed the use of a novel application called ITHAKA to support HIV self-testing among youth, either on a tablet on site or on their mobile phone off-site. However, only 5.8% (128/2181) youth opted for HIV self-testing over provider testing. Furthermore, of those who performed HIV self-testing off-site, less than half (9/19, 47%) completed their testing journey [24]. Although some issues were technology focused, such as lack of phone ownership, limited functionality, and erratic network coverage, other factors included a lack of agency or private space and a desire for provider support during testing. As a result, even addressing the digital factors may not have led to acceptability and feasibility.

Strengths and Limitations

Strengths of this study include a population-based sample with a very large sample size, random sampling, and the fact that digital technologies were not required to access the study. Limitations include the potential for social desirability bias, especially given that mobile phones may be perceived as a symbol of wealth and status. Of note, 151 mobile phone owners had phones that could not call or text. Of them, 110 (72.8%) were also neither equipped to use WhatsApp or browse the internet. It is interesting that these individuals still considered themselves mobile phone owners, despite their phones not having any of the expected core functionalities. Social desirability bias may have influenced this self-classification. Although we were able to construct a household wealth index from 6 household assets, the reluctance of some participants to disclose information related to estimated household income demonstrates the difficulties related to collecting data on wealth. Importantly, we are unable to demonstrate causality from this cross-sectional study. Therefore, although we hypothesized that, for example, being wealthier increases the odds of owning a phone or accessing the internet, the inverse may account for at least some of the relationship. For example, owning a mobile phone may allow a young adult to engage more dynamically with business opportunities, thus increasing wealth. Selection bias may also have influenced the results. Of the 1039 individuals enumerated but not enrolled, 696 (67%) declined participation and a further 77 (7.4%) were not found. These participants may have different characteristics from those enrolled. Finally, another key limitation is generalizability, with all participants recruited from urban or periurban areas, which are likely to have different levels of access to digital technologies compared to rural areas.

Conclusions

To our knowledge, this is the largest population-based survey examining the factors associated with digital technology use among young people in Africa. It demonstrates that a high proportion of young adults in the urban or periurban areas studied own or at least have access to a mobile phone. However, to fully make use of digital technology in Zimbabwe, more effort must be made to increase access further, targeted at those with lower levels of wealth, in order that interventions do not exacerbate existing socioeconomic inequality. Before the implementation of digital technology–based interventions, whether urban or rural, a key component of formative work must be focused on the use of and accessibility to the required technologies.

Acknowledgments

The authors thank the Wellcome Trust, which funded the CHIEDZA trial (senior fellowship to RAF: 206316/Z/17/Z).

XSL•FO
RenderX

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request. Additionally, individual, anonymized participant data and a data dictionary will be available through The London School of Hygiene & Tropical Medicine repository (Data Compass) 12 months after the publication of results.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions and response categories.
[DOCX File , 21 KB - jmir_v26i1e52670_app1.docx]

Multimedia Appendix 2
Construction of wealth index using principal component analysis.
[DOCX File , 40 KB - jmir_v26i1e52670_app2.docx]

Multimedia Appendix 3
WhatsApp and internet use according to phone functionality among mobile phone owners.
[DOCX File , 15 KB - jmir_v26i1e52670_app3.docx]

Multimedia Appendix 4
Univariable associations between age, education level, and time at current address and mobile phone ownership, internet use, and social media use.
[DOCX File , 17 KB - jmir_v26i1e52670_app4.docx]

References

1. Kickbusch I, Piselli D, Agrawal A, Balicer R, Banner O, Adelhardt M, et al. The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. Lancet 2021 Nov 06;398(10312):1727-1776. [doi: 10.1016/S0140-6736(21)01824-9] [Medline: 34706260]
2. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023 Mar 23;25:e43542 [FREE Full text] [doi: 10.2196/43542] [Medline: 36951896]
3. Ibeneme SC, Ndukwu SC, Myezwa H, Irem FO, Ezenwankwo FE, Ajidahun AT, et al. Effectiveness of mobile text reminder in improving adherence to medication, physical exercise, and quality of life in patients living with HIV: a systematic review. BMC Infect Dis 2021 Aug 23;21(1):859 [FREE Full text] [doi: 10.1186/s12879-021-06563-0] [Medline: 34425789]
4. Ubrihien A, Stone AC, Byth K, Davies SC. The impact of Grindr advertising on attendance and HIV testing by men who have sex with men at a sexual health clinic in northern Sydney. Int J STD AIDS 2020 Sep 10;31(10):989-995. [doi: 10.1177/0956462420927815] [Medline: 32772689]
5. Goedel WC, Duncan DT. Geosocial-networking app usage patterns of gay, bisexual, and other men who have sex with men: survey among users of Grindr, a mobile dating app. JMIR Public Health Surveill 2015 May 08;1(1):e4 [FREE Full text] [doi: 10.2196/publichealth.4353] [Medline: 27227127]
6. Huhn S, Axt M, Gunga HC, Maggioni MA, Munga S, Obor D, et al. The impact of wearable technologies in health research: scoping review. JMIR Mhealth Uhealth 2022 Jan 25;10(1):e34384 [FREE Full text] [doi: 10.2196/34384] [Medline: 35076409]
7. Gunasekeran DV, Tseng RM, Tham YC, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021 Feb 26;4(1):40 [FREE Full text] [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
8. Silver L. Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. 2019 Feb 5. URL: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ [accessed 2024-08-28]
9. Measuring digital development: facts and figures 2022. International Telecommunication Union. 2022. URL: https://www.itu.int/hub/publication/d-ind-ict_mdd-2022/ [accessed 2024-08-28]
10. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021 Jan 11;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
11. Hurt K, Walker RJ, Campbell JA, Egede LE. mHealth interventions in low and middle-income countries: a systematic review. Glob J Health Sci 2016 Sep 01;8(9):54429 [FREE Full text] [doi: 10.5539/gjhs.v8n9p183] [Medline: 27157176]

XSL•FO
RenderX

12. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]
13. Dambi J, Norman C, Doukani A, Potgieter S, Turner J, Musesengwa R, et al. A digital mental health intervention (Inuka) for common mental health disorders in Zimbabwean adults in response to the COVID-19 pandemic: feasibility and acceptability pilot study. JMIR Ment Health 2022 Oct 07;9(10):e37968 [FREE Full text] [doi: 10.2196/37968] [Medline: 35960595]
14. Zimbabwe National Statistics Agency and ICF International. Zimbabwe demographic and health survey 2015: final report. Zimbabwe National Statistics Agency (ZIMSTAT) and ICF International. 2016. URL: https://dhsprogram.com/publications/publication-fr322-dhs-final-reports.cfm [accessed 2024-08-28]
15. Doyle AM, Bandason T, Dauya E, McHugh G, Grundy C, Dringus S, et al. Mobile phone access and implications for digital health interventions among adolescents and young adults in Zimbabwe: cross-sectional survey. JMIR Mhealth Uhealth 2021 Jan 13;9(1):e21244 [FREE Full text] [doi: 10.2196/21244] [Medline: 33439136]
16. Dziva Chikwari C, Dauya E, Bandason T, Tembo M, Mavodza C, Simms V, et al. The impact of community-based integrated HIV and sexual and reproductive health services for youth on population-level HIV viral load and sexually transmitted infections in Zimbabwe: protocol for the CHIEDZA cluster-randomised trial. Wellcome Open Res 2023 Nov 7;7:54. [doi: 10.12688/wellcomeopenres.17530.2]
17. Yang WE, Spaulding EM, Lumelsky D, Hung G, Huynh PP, Knowles K, et al. Strategies for the successful implementation of a novel iPhone loaner system (iShare) in mHealth interventions: prospective study. JMIR Mhealth Uhealth 2019 Dec 16;7(12):e16391 [FREE Full text] [doi: 10.2196/16391] [Medline: 31841115]
18. Male engagement in HIV testing, treatment and prevention in eastern and southern Africa — a framework for action. Joint United Nations Programme on HIV/AIDS. 2022 Apr 7. URL: https://www.unaids.org/en/resources/documents/2022/male-engagement-hiv-testing-treatment-prevention-eastern-southern-africa [accessed 2024-08-28]
19. Konlan I, Abdulai M, Ibrahim H. Exploring the effects of social media on marriages in northern Ghana. Hu Arenas 2023 Apr 12. [doi: 10.1007/s42087-023-00333-x]
20. Young SD, Cumberland WG, Nianogo R, Menacho LA, Galea JT, Coates T. The HOPE social media intervention for global HIV prevention in Peru: a cluster randomised controlled trial. The Lancet HIV 2015 Jan;2(1):e27-e32. [doi: 10.1016/s2352-3018(14)00006-x]
21. Bastien S. Reflecting and shaping the discourse: the role of music in AIDS communication in Tanzania. Soc Sci Med 2009 Apr;68(7):1357-1360. [doi: 10.1016/j.socscimed.2009.01.030] [Medline: 19232806]
22. Capurro D, Cole K, Echavarría MI, Joe J, Neogi T, Turner AM. The use of social networking sites for public health practice and research: a systematic review. J Med Internet Res 2014 Mar 14;16(3):e79 [FREE Full text] [doi: 10.2196/jmir.2679] [Medline: 24642014]
23. Lennon MR, Bouamrane MM, Devlin AM, O'Connor S, O'Donnell C, Chetty U, et al. Readiness for delivering digital health at scale: lessons from a longitudinal qualitative evaluation of a national digital health innovation program in the United Kingdom. J Med Internet Res 2017 Feb 16;19(2):e42 [FREE Full text] [doi: 10.2196/jmir.6900] [Medline: 28209558]
24. Larsson L, Chikwari CD, McHugh G, Koris A, Bandason T, Dauya E, et al. Feasibility and usability of mobile technology to assist HIV self-testing in youth in Zimbabwe: a mixed-methods study. J Adolesc Health 2023 Sep;73(3):553-560 [FREE Full text] [doi: 10.1016/j.jadohealth.2023.05.011] [Medline: 37389521]

Abbreviations

aOR: adjusted odds ratio
CONSORT: Consolidated Standards for Reporting Trials
mHealth: mobile health

Edited by A Mavragani; submitted 14.09.23; peer-reviewed by L Kuhns, E Bantum; comments to author 27.02.24; revised version received 18.03.24; accepted 08.07.24; published 23.09.24.

Please cite as:
Martin K, Peh RWC, Tembo M, Mavodza CV, Doyle AM, Dziva Chikwari C, Dauya E, Bandason T, Azizi S, Simms V, Ferrand RA Factors Associated With the Use of Digital Technology Among Youth in Zimbabwe: Findings From a Cross-Sectional Population-Based Survey
J Med Internet Res 2024;26:e52670
URL: https://www.jmir.org/2024/1/e52670
doi:10.2196/52670
PMID:39312390

XSL•FO
RenderX

©Kevin Martin, Rachel Wei Chun Peh, Mandikudza Tembo, Constancia Vimbayi Mavodza, Aoife M Doyle, Chido Dziva Chikwari, Ethel Dauya, Tsitsi Bandason, Steven Azizi, Victoria Simms, Rashida A Ferrand. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

From Digital Inclusion to Digital Transformation in the Prevention of Drug-Related Deaths in Scotland: Qualitative Study

Hadi Daneshvar[1,2], BSc, MSc, PhD; Hannah Carver[2], MA (Hons), MSc, PhD; Graeme Strachan[2], BSc, MSc; Jessica Greenhalgh[2], BSc, MRES; Catriona Matheson[2], BSc, MSc, PhD

[1]School of Health and Social Care, Edinburgh Napier University, Edinburgh, United Kingdom

[2]Faculty of Social Sciences, University of Stirling, Stirling, United Kingdom

Corresponding Author:
Hadi Daneshvar, BSc, MSc, PhD
School of Health and Social Care
Edinburgh Napier University
Sighthill Campus
9 Sighthill Court
Edinburgh, EH11 4BN
United Kingdom
Phone: 44 (0)131455 2929
Email: h.daneshvar@napier.ac.uk

Abstract

Background: Globally, drug-related deaths (DRDs) are increasing, posing a significant challenge. Scotland has the highest DRD rate in Europe and one of the highest globally. The Scottish Government launched the Digital Lifelines Scotland (DLS) program to increase the provision of digital technology in harm reduction services and other support services. Digital technology responses to DRDs can include education through digital platforms, improved access to treatment and support via telehealth and mobile apps, analysis of data to identify risk factors, and the use of digital tools for naloxone distribution. However, digital technology should be integrated into a comprehensive approach that increases access to services and addresses underlying causes. Digital transformation could enhance harm reduction service and support, but challenges must be addressed for successful implementation. The DLS program aims to enhance digital inclusion and improve health outcomes for people who use or are affected by drug use to reduce the risk of DRDs.

Objective: This study aims to explore the role of digital technology as an enabler and supporter in enhancing existing services and innovating new solutions, rather than being a stand-alone solution. Specifically focusing on individuals who use drugs, the research investigates the potential of digital inclusion and technology provision for preventing DRDs within the context of the DLS program.

Methods: Semistructured interviews were conducted with 47 people: 21 (45%) service users, 14 (30%) service providers, and 12 (26%) program staff who were all involved in DLS. Interviews were audio recorded, transcribed, and then coded. Analysis was done in three phases: (1) thematic analysis of interview data to identify the benefits of digital technologies in this sector; (2) identification of the challenges and enablers of using digital technologies using the Technology, People, Organizations, and Macroenvironment conceptual framework; and (3) mapping digital technology provision to services offered to understand the extent of digital transformation of the field.

Results: Participants identified increased connectivity, enhanced access to services, and improved well-being as key benefits. Digital devices facilitated social connections, alleviated loneliness, and fostered a sense of community. Devices enabled engagement with services and support workers, providing better access to resources. In addition, digital technology was perceived as a preventive measure to reduce harmful drug use. Lack of technical knowledge, organizational constraints, and usability challenges, including device preferences and security issues, were identified.

Conclusions: The study found that digital inclusion through the provision of devices and connections has the potential to enhance support in the harm reduction sector. However, it highlighted the limitations of existing digital inclusion programs in achieving comprehensive digital transformation. To progress, there is a need for sustained engagement, cultural change, and economic considerations to overcome barriers.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52345) doi:10.2196/52345

KEYWORDS

digital inclusion; digital transformation; digital health; drug-related death; digitalization; drug overdose; drug overdose death; harm reduction; mobile phone; digital divide

Introduction

Background

The world is grappling with a drug-related death (DRD) epidemic, affecting many countries globally. However, specific statistics for DRDs worldwide are not readily available, although there were an estimated 600,000 deaths related to drug use worldwide in 2019 [1,2]. While data vary across countries and regions, it is evident that DRDs pose a significant challenge on a global scale. Scotland has been facing a significant issue with DRDs in recent years. In 2022, a total of 1051 DRDs were reported in Scotland [3]. The DRD rate in Scotland is several times higher than in the rest of the United Kingdom and is also the highest in Europe. High DRD rates worldwide, including in Scotland, are due to complex and multifaceted issues, including poverty, a complex drug market, the proliferation of unregulated drugs, the aging population of people who use drugs, social inequality, stigma, and a lack of adequate treatment and harm reduction services [4-7].

The Scottish Government has recognized the severity of the problem [8]. It declared DRDs a public health emergency and established a Drug Deaths Taskforce in 2019 to propose solutions [9]. The Drug Deaths Taskforce and Scottish Government recognized the potential of digital technology, to provide connection and support to people at risk of overdose. This gave rise to the Digital Lifelines Scotland (DLS) program [10] which provided funding to support the provision of devices and data to enable digital inclusion and support between 2021 and 2023.

The objective of the DLS program was to enhance digital inclusion and develop digital solutions that effectively cater to the specific needs of individuals, thereby improving the health outcomes for people who use drugs or their families and thereby reducing their risk of DRDs. Digital inclusion is defined as “the activities necessary to ensure that all individuals and communities, including the most disadvantaged, have access to and use of information and communication technologies” [11]. The DLS program specifically offered different digital solutions encompassing digital devices, assistance, education, well-being services, and innovative approaches to people at risk of drug-related harm. Following the launch of the DLS program, we evaluated the impact of digital technology availability and use on service users and service provision.

Different types of digital technology have the potential to play a significant role in addressing DRDs [12]. Daneshvar et al [13] considered that digital technology has the potential to be used to help combat the problem by providing education, raising awareness about the risks of drug use, promoting harm reduction strategies, and providing educational resources to individuals and communities. Social media campaigns, websites, mobile apps, and web-based communities can all be effective tools for disseminating information and reaching a wide audience. Second, telehealth and telemedicine can improve access to addiction treatment and mental health support, particularly in rural or underserved areas where there may be a shortage of health care providers [14-16]. Telemedicine can facilitate remote consultations, medication-assisted treatment, counseling, and follow-up care. Mobile apps also have the potential to act as support, monitoring, and self-management tools for individuals experiencing drug use problem by offering features, such as tracking substance use, setting goals, connecting with support networks, accessing educational materials, and providing immediate access to crises helplines [12]. In addition, by analyzing large datasets and using predictive modeling techniques, digital technology can help identify patterns and risk factors associated with DRDs [12,17,18]. This information can assist policy makers and health care professionals in targeting interventions and allocating resources more effectively. Finally, digital tools, such as distribution apps for naloxone (an opioid overdose reversal medication) or wearable devices, can help prevent or reverse overdoses by providing timely information and resources. These technologies can facilitate the dissemination of naloxone and provide instructions on its use during an emergency, while wearable devices can detect overdose situations [12]. Overall, by promoting harm reduction, facilitating access to support services, enhancing communication and engagement, and enabling personalized interventions, digital technology has the potential to empower service users and improve their overall well-being [12].

This Study

While digital technology holds promise, it is not a stand-alone solution. It is rather an enabler and supporter for enhancing the existing service or innovating new services [19]. This paper examines the potential of digital inclusion and provision of technology for people who use or used drugs (called service users in this study) and identifies the barriers and enablers in the use of digital technologies for preventing DRDs, as part of the larger study of the evaluation of the DLS program. To do this, we address the following research questions: (1) What are the perceived benefits of the provision of digital technology for service users and what type of services (existing or new) are offered because of the provision of digital technologies? and (2) What are the barriers and enablers for using digital technologies for prevention of DRDs?

We have adopted a qualitative research approach to examine the views of service users and providers and those delivering the DLS program.

Methods

Overview

This paper draws on data from a wider study that evaluated the DLS program using semistructured interviews and surveys [20].

XSL•FO
RenderX

The study was informed by the Technology, People, Organization, and Macroenvironment (TPOM) framework [21]. This framework underscores the interconnectedness between various technological and social dimensions, illustrating how diverse viewpoints influence the implementation and adoption of emerging technologies. Qualitative data relating to technology were considered.

Participants and Recruitment

This study used purposive sampling [22] to recruit participants from 3 groups: (1) service users who had access to digital solutions via the DLS program, (2) service providers providing digital solutions through DLS funding, and (3) members of the DLS program team or board responsible for the delivery of the program. We collected data from 3 groups of participants to provide insight into the DLS program from a range of viewpoints rather than covering a single perspective.

The inclusion criteria of participants were as follows: first, service users who were using illicit drugs (or had done so in the last 12 months) who had wider health and social challenges (multiple and complex needs) and received a digital technology–based innovation funded by the DLS program. Second, the criteria included service providers who were managers, frontline staff, and volunteers working in third-sector organizations that provided digital solutions funded by the DLS program, including harm reduction services, counseling, and housing support. In the context of the United Kingdom, third-sector organizations operate distinctively from both the public and private sectors. This classification encompasses a variety of voluntary and community groups, including registered charities, as well as other forms, such as associations, self-help groups, and community-based organizations. In addition, it includes social enterprises, mutuals, and cooperatives [23]. Third, the criteria included the program delivery team and the program board, responsible for the delivery of DLS. Through purposive sampling, we ensured that participants were recruited from a range of organizations and initiatives funded by DLS (eg, types of technology and training and service delivery) with different genders, roles, and diverse types of technology received. The exclusion criteria were people aged <18 years, unable to provide informed consent, unable to speak or understand English, unable to take part due to severe mental health or behavioral problems, those not currently living in Scotland, and those not involved in DLS. We used the DLS program to identify participants in the study.

Service users were recruited via service providers, who were emailed and provided with information about the study by the DLS program team and asked to pass this on to relevant individuals. Staff in the organizations then arranged interviews with the study researchers at a suitable date or time. Service providers were recruited via email from the DLS program team, who provided them with information about the study and asked them to contact the research team to arrange an interview. The program team or program board was emailed by the research team and invited to participate, with interviews being arranged directly with the researchers.

Semistructured Interviews

The interview guide was developed by HD, CM, and HC (Multimedia Appendix 1), incorporating relevant areas of the TPOM and drawing on insights from the broader DLS program delivery work and a prior user need study [20]. The topic guide for service user participants primarily addressed the social or human (People) domains, examining how the technology was used and its impact on service users and their relationships with service providers.

Participants were provided with an information sheet and had the opportunity to ask questions. Written or verbal informed consent was provided before each interview. All interviews were conducted between August 2022 and February 2023. Interviews took place either in person, on the web, or over phone by one of the researchers (HD, GS, and JG) and lasted an average of 32 minutes. All interviews were audio recorded. Interviews with service users explored their experiences of receiving support as part of DLS; service provider interviews covered the infrastructure of digital services and current availability and levels of digital literacy. Interviews with the program team explored their views of the impact of the program and gaps, barriers, and enablers at an organizational and macroenvironment level. Participants were provided with a debrief sheet that outlined information about the study and sources of support.

Data Analysis

The TPOM provided a broad conceptual framework for analysis [21]. This framework is designed to cover areas of evaluation that allows for provision of both formative and summative evaluation for studying technology in health and social care settings [21]. The interviews were transcribed verbatim by an external transcriber and identifiable information was removed, with the Scottish dialect retained. Data analysis was completed using NVivo (version 12; Lumivero), which streamlines qualitative research by allowing researchers to organize, manage, and code data efficiently, including text, thereby saving time, and enhancing accuracy [24]. NVivo was chosen as it is the software regularly used by members of the research team. GS coded 3 transcripts to develop the initial coding framework, which was reviewed by HC. Because no structural changes were necessary, GS and HD proceeded to code the remaining transcripts.

Data analysis had 3 phases. In phase 1, to understand the promises of digital technologies and their transformative outcomes, we used thematic analysis [25] within the technology domain of the TPOM to identify categories of perceived benefits. The coding started in a line-by-line manner on data, and it involved understanding the “meanings” of the messages, rather than the words used to communicate [26]. We then searched for common subthemes across codes in an iterative process where we moved back and forth between the codes to identify commonalities.

In phase 2, we coded the data related to the challenges and enablers of using digital technologies for the provision of harm reduction and services. Then, in an iterative process, we grouped the codes into subthemes around challenges. In this phase, we

XSL•FO
RenderX

identified the enablers in response to challenges as shown in Figure 1, where the green boxes are added to the original TPOM framework.

Figure 1. Adapted Technology, People, Organization, and Macroenvironment (TPOM) framework diagram.

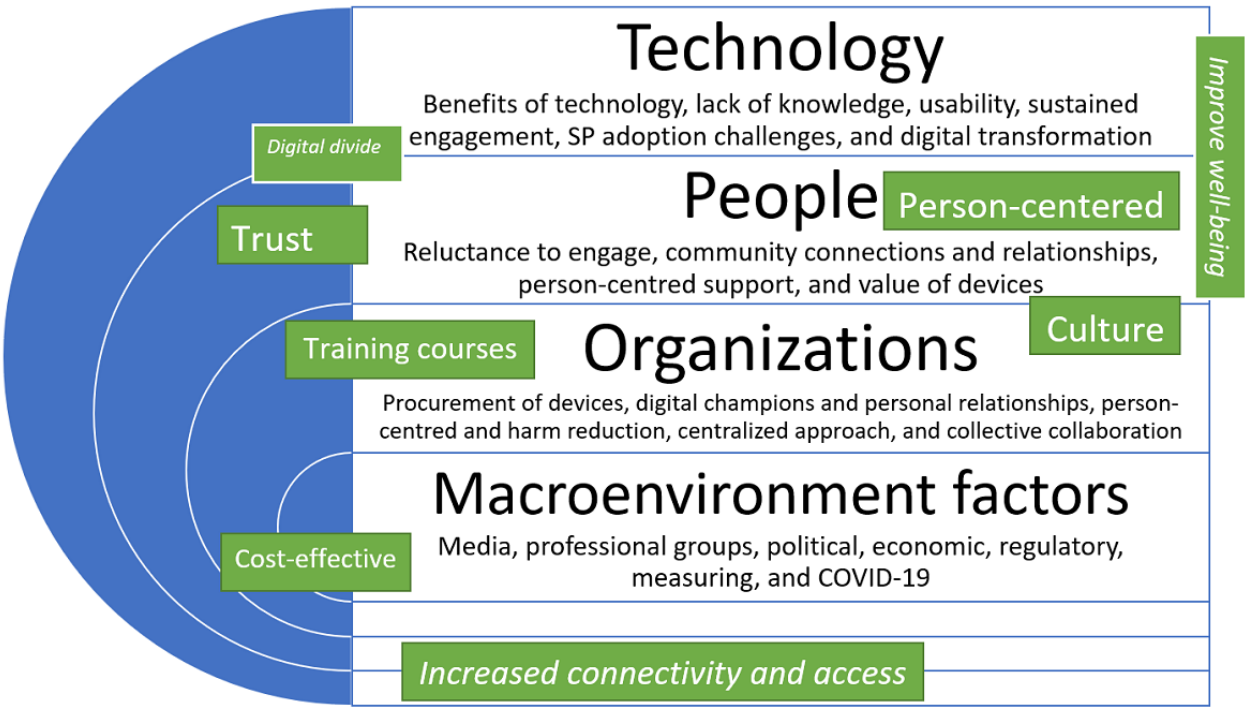

In phase 3, we reviewed findings against a conceptual framework (Figure 2) by Daneshvar et al [27] to explore the relationship between existing and new digital technology and the provision of existing and new services. We used this model to categorize the potential of different types of technologies. This model encompasses 2 main components: the integration of digital tools into provision of existing services, and the creation of novel services through adoption and use of established or emerging digital tools.

XSL•FO
RenderX

Figure 2. Digital technology and service providers.

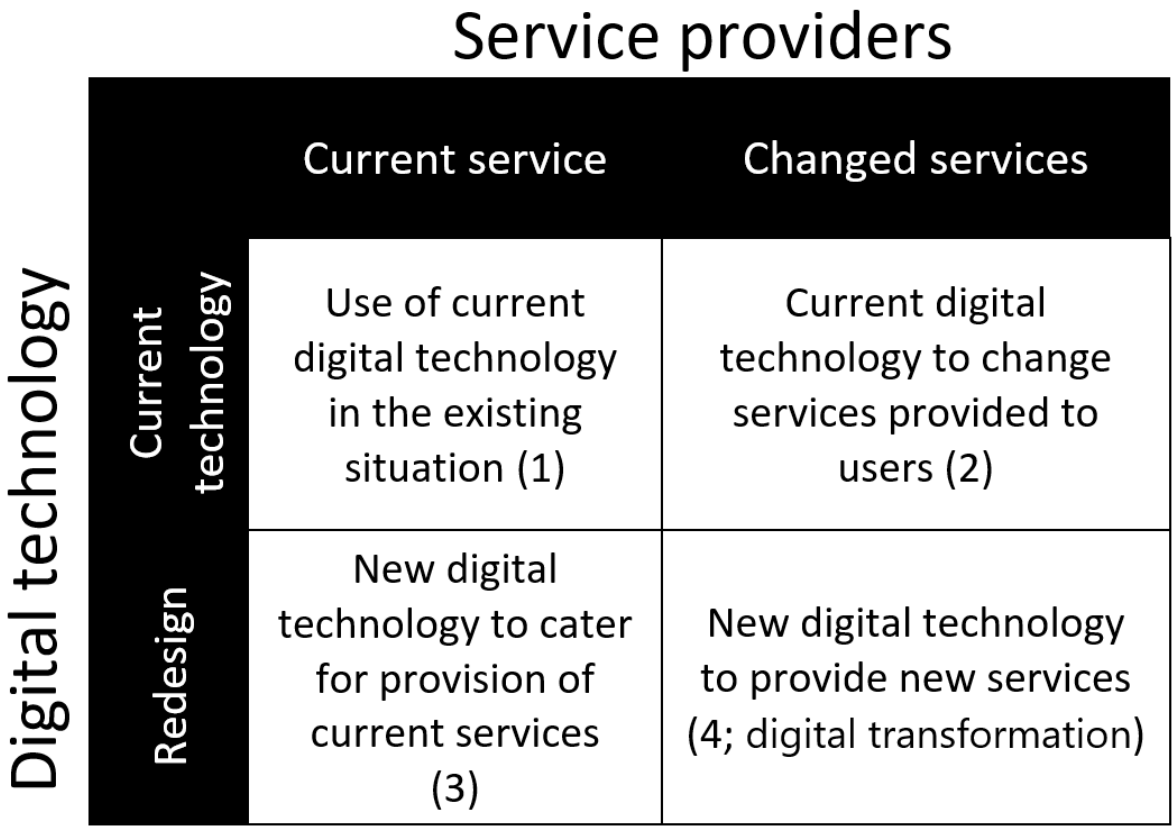

Ethical Considerations

This study adhered to the guidelines of the Declaration of Helsinki and received approval from the University of Stirling's General University Ethics Panel (7799—August 2022), the Ethics Subgroup of the Research Coordinating Council of the Salvation Army, Turning Point Scotland, and Shine Mentoring. All participants provided informed consent either through a consent form or verbally. All data and quotes, including secondary data, were pseudonymized and attributed to each participant; identifiable names, places, or people were removed.

Service user participants received a £10 (US $13) shopping voucher as an honorarium for their participation. All data protection and General Data Protection Regulation aspects of the project and data privacy have been reviewed and approved by the Data Protection Office of Stirling University, and a Data Protection Impact Assessment form has been issued.

Results

Overview

The DLS program provided diverse digital solutions, including devices, support, education, and innovative services, to individuals at risk of drug-related harm. A total of 534 devices were distributed by 13 (81%) of the 16 organizations involved in the DLS program: 265 (49.6%) smartphones, 145 (27.2%) data packs, 75 (14%) tablets, and 49 (9.2%) laptops.

In total, 47 interviews were conducted by researchers, with 21 (26%) service users, 14 (30%) service providers, and 12 (45%) program team or board members. Table 1 shows the breakdown of the number of service users, service providers, and program team or board participants. The differences in the total number of participants from each organization are due to the different sizes of organizations. The demographics of participants were not collected, but they were similar to those in our related papers [20,28].

XSL•FO
RenderX

Table 1. Participation numbers and organizations.

Breakdown of interview participants (n=47) by organization, program team, or board member	Participation, n (%)
Program team or board members (n=12, 26%)	
Program organization 1	4 (9)
Program organization 2	2 (4)
Program organization 3	2 (4)
Program organization 4	2 (4)
Program organization 5	1 (2)
Program organization 6	1 (2)
Service provider (n=14, 30%)	
Service provider 1	5 (11)
Service provider 2	2 (4)
Service provider 3	2 (4)
Service provider 4	1 (2)
Service provider 5	1 (2)
Service provider 6	1 (2)
Service provider 7	1 (2)
Service provider 8	1 (2)
Service user (n=21, 45%)	
Service provider 1	9 (19)
Service provider 5	8 (17)
Service provider 2	3 (6)
Service provider 4	1 (2)

The findings are presented as 3 themes and 7 subthemes as presented in Textbox 1. Then, they are brought together to represent the journey of digital transformation in the sector.

Textbox 1. Themes and subthemes of the findings.

Themes and subthemes

- The perceived and actual benefits of digital technology for service users
  - Increased connectivity for informal support
  - Increased access to service
  - Improved well-being
- Challenges of and enablers for using digital technology
  - Technological
  - People
  - Organizational
  - Macroenvironment
- The journey to digital transformation

XSL•FO
RenderX

The Perceived and Actual Benefits of Digital Technology for Service Users

The benefits of technology provision described by participants were (1) increased connectivity for informal support, (2) increased access to services, and (3) improved well-being.

Increased Connectivity for Informal Support

The first category of benefits of digital devices as identified by participants was making connections with others to receive some type of support. This support could range from overcoming loneliness to creating friendships and a sense of community building:

> *Before I got the iPad...I wasn't really connecting with people...But it [iPad] opened up a lot of doors for me, aye [yes]...it did. It made me feel more connected with people...I think more and more people are getting there...I know people in the community are getting connected.* [Service user participant 1]

This benefit was discussed most frequently by both service users and service providers, with many providing examples of connectivity to family and friends:

> *I stay on my own just now, my oldest son he's in the Army, he's 27 so I can Zoom him. My grandson, I can Zoom him...I can Zoom my other son, he's in prison. He's done it all the time. He's 25...in prison, they've got these [computers] in the visit room. They started giving virtual visits during COVID, so we got a visit with him with Zoom.* [Service user participant 21]

> *We are asking people what they want, and mostly people want access to community. They want to be able to get on to social networks. They want to be able to interact with friends and family digitally, because their friends and family might have left [city/town], or they might have moved here from somewhere else, and they don't have that community.* [Service provider participant 5]

Increased connectivity was also mentioned to be of high importance for those who had recently completed drug treatment recovery programs or those who were released from prison who were at risk of overdose if they started using drugs again:

> *The first kind of most important thing that they use it for is that initial contact, and that's then how we work with them and do the project really. So, I think without a phone, and without that kind of line of contact, it's pretty difficult. You know, I've had a few women that have come out (of prison) and not had contact numbers. And quite a lot of the time it means, you know, disengagement and you find them further along down the line but it's really important initially.* [Service provider participant 14]

Several participants noted that if it were not for the provision of technology, many would lose any benefit obtained from having a digital connection:

> *See, until I get a new smartphone, I am not getting any contact wi' [with] ma [my] house, my Mum hasnae [hasn't] been. She had my phone number, so she can see me.* [Service user participant 13]

Increased Access to Services

The use of digital devices improved access to existing services. This was perceived as a benefit to both service users and service providers:

> *I do suspect as well that, again, if we're giving people decent enough phones, they're going to need to use them for other aspects of their life anyway and they'll actually get quite used to them, you know. So, whether it's speaking to your daughter, or speaking to your dealer, the phone's quite important to you. But if at the same time, we can utilize that to be able to access those services when they're ready and able to do so, then that for me is a win-win scenario.* [Program team participant 6]

An increase in access to services was likely due to several factors. First, people who were originally excluded due to lack of access to technology had an increased chance of being engaged as technology was provided to them:

> *We can cover things that are happening here. So, that access to those devices was quite helpful for engaging with people who normally wouldn't get a chance to engage.* [Service provider participant 4]

As technology provision increased, access to support was made available to individuals, and therefore service users were better engaged with support workers:

> *Eighty-five percent of people in the project said that they engaged more with their support worker since having digital [devices]. And the reason for that is that the person is available to be contacted at any time and if they are in a moment of need, you know, we're available to them and they're available to us so it's just that easy way of staying in contact. And that's massive because not only is it like obviously that extra connection between the service and the person, it also means that when there is a moment of need, we're able to help.* [Service provider participant 12]

Another reason that access to services was increased was that new offerings such as upcoming events could be more easily identified by service users:

> *[I find out about things] on Facebook and that, it keeps me connected digitally and also, I know where I can go and meet people, like events are advertised and what not.* [Service user participant 1]

One participant highlighted the connections between DLS and several wider policy programs of social inclusion. They highlighted that such digital inclusion programs give people more choices regarding access to resources and information to support their own self-management and preventive care:

> *The benefit, I guess, is to improve access and ideally outcomes for people in being able to connect in with services and support...I think it should bring in more choice to people about when and how they want to*

XSL•FO
RenderX

> *access services. So, rather than being told, 'well, this is where you have to be at a particular time,' you've more choice. I think it also should empower people more in terms of self-management and early intervention. So, rather than wait for somebody to tell you what's wrong, you're able to get data and information in real time that's giving you some information around choices or decisions as you can make, and I suppose bring things into more real time.* [Program team participant 10]

Program team members and service providers discussed benefits from several perspectives, contrasting short- and longer-term benefits and objective versus more subjective aspects. One noted that participation in the project had led to directly intervening in several potentially fatal overdoses. Hence, delivery of prevention services was enabled through the use of technology:

> *We have intervened in six scenarios where there was a kind of near-death experience and we've managed to turn that around. That, for me, would be a great story to tell but whether we're going to be able to tell it or not, I don't know, we're obviously still recruiting [participants for study].* [Program team participant 6]

The same participant discussed "softer," but still "life-changing," benefits that included service users being able to connect with services and other supports at key points in their recovery or rehousing journey:

> *So, you know, we've got a number of impacts that we've already [seen?], some of them are very kind of softer impacts though, you know, the stuff that's coming out of the digital inclusion piece that [service providers] have been leading on? You know, it's even just things like, you know, speaking to my housing officer. You know, connecting into services that were very difficult for you to connect into before. Or settle in quicker when you're discharged from prison and maybe not linking up with your previous networks. You know, so some of those kind of softer life changing stories I think are equally valuable from a policy perspective.* [Program team participant 6]

Participants suggested that while not everyone who receives devices will be at risk of overdose, these devices could still have a positive impact by preventing individuals from engaging in harmful behaviors associated with drug use in the long term:

> *I think it's okay that as well like a lot of people who get devices will not be someone that may have overdosed, but the device may be the thing that has stopped them taking that longer term pathway...We just need to kind of cut into like why are we doing this? Who's it for? What are the projected long-term benefits? And equally I think being really bold about the fact that some of this stuff is a total unknown.* [Program team participant 1]

Improved Well-Being

Digital technologies are seen by participants to have significant potential in improving the well-being of the community. While the original intention of the provision of technology was the prevention of DRDs, one additional potential that was observed was improving people's lives, which could result in reduced drug use and related harms:

> *I also think, to state the obvious, it shouldn't just look at preventing drug death. It should look at supporting people's wellbeing and ways into the community from drug use.* [Service provider participant 5]

For instance, when individuals were given access to technology, some tried to use it to make positive life changes, such as using the technology to learn new skills to find employment:

> *I think it's just opening doors for people...for example, one guy, he came along and he's really trying to get his life on track. He's come along to the computer group on a Tuesday and he's just really wanting to change, really wanting to do something. He's in a bit of better place, he's more stable. He came along and he's wanting to look for jobs so that's his benefit.* [Service provider participant 9]

The well-being of individuals was also perceived as improving, by enhancing their confidence through the use of technology. This in turn could lead to other benefits, such as connecting to a support network that was previously unavailable to them:

> *So, everyone's really receptive of why we're needing skills with computers, tablets, mobiles. And it's just building their confidence as well. And it gives them that option to have that support network that they wouldn't really have before.* [Service provider participant 2]

Another benefit identified in terms of well-being was that access to high-quality technology and the provision of web-based services could reduce people's feelings of stress because they could access support more quickly than they might have previously:

> *So, when you compare that to only having a two-hour session, you're spending 15, 20 minutes booting [up the equipment]. I think having good equipment and having an IT suite that actually works as it should enables people to come in and quickly do the things they want to do. That has made a massive difference. Because, anecdotally, prior to that, we had equipment that people would wait ages for things to happen, which would create a stressful environment, which would add stress on to individuals that don't need stress adding on.* [Service provider participant 2]

One program team member noted the importance of and challenges in measuring the softer impacts. They also reflected that this project was relatively novel and that measures of success should be considered emergent:

> *...We just need to kind of cut into like why are we doing this? Who's it for? What are the projected long-term benefits? And equally I think being really bold about the fact that some of this stuff is a total unknown. So, it's like you can't just give devices to like, you know, 25 people who are supported by the [service provider] and saying that is going to cut drug*

XSL•FO
RenderX

> *deaths by ten percent. There might be a point where we have some evidence base that really succinctly shows that, but it's very, very difficult...* [Program team participant 1]

Challenges of and Enablers for Using Digital Technologies

Overview

While the program intended to offer a range of benefits to service users by provision of technologies, not all perceived benefits were realized. This was due to numerous barriers at the individual (eg, lack of technical knowledge), organizational (eg, lack of resources), and macroenvironmental (eg, negative media impact) levels. Some of these barriers led to slow adoption of existing technologies for the services, while others led to limited innovations, which restricted new service provision. The subthemes are presented under technological, people, and organizational challenges.

Technological Challenges

Although benefits associated with technology provision were realized, the usability of the devices, digital technologies, and apps raised concerns. There were different preferences and feelings expressed toward different devices offered by the program. Some devices were perceived to be easier to use than others:

> *If somebody offered me a tablet, I'd definitely take it, because it would help me with my college work. And it would be easier, probably, to use a tablet than the Chromebook.* [Service user participant 21]

There were different factors affecting the usability of devices, such as age, level of education, and preexisting conditions. For instance, one participant explained how their dyslexia made it more challenging to work with laptops than tablets:

> *I guess tablets are just the same as the phones. I mean it's still kind of like touchscreen. But again, with my dyslexia I'd have to get someone to set it up because my son-in-law had to set this up 'cos I've not got a clue how to set phones up. But I would prefer the tablets to like the laptops as well...Because I find they are a lot easier for me to like use because there's no buttons. And I get confused on what buttons I am pressing.* [Service user participant 14]

Another point that was highlighted was how "digitally savvy" an individual was. Smartphones and tablets were identified as more learnable and usable to those who were less proficient, as opposed to laptops and Chromebooks (Google LLC):

> *So, we did a trial last year where we did basically the same program but with 100 people. That told us that the two most popular devices for the type of people we're supporting would be a smartphone or a tablet. Very few people would need or have use for a Chromebook really, or a laptop. The people we're supporting tend to be at the start of their digital journey.* [Service provider participant 10]

Apart from devices, the operating system on the devices seemed to have an impact on the usability of technology:

> *I tried to [offer] a bit of mix of phones and tablets, because the operating software on the tablet, which tends to be an iPad, is quite user-friendly.* [Service provider participant 4]

Moreover, the need to use passwords and other security features was seen to negatively impact the usability of the technologies provided:

> *Sometimes logging into things can be a bit hard when you forget your passcode and that.* [Service user participant 18]

People Challenges

While many benefits were envisaged and realized by our study participants, the lack of technical knowledge among service users was the most evident challenge highlighted. This could lead to low engagement or short intervals of engagement with technology:

> *There are people that engage for a little while. There are people that come and say they don't know how to work this. They can barely get online. They don't even have an email address. They want to know how to log in to Zoom, all those kinds of stuff.* [Service provider participant 4]

There are also other service users who are more confident but have limited technology skills. They reported being more engaged in trying to find ways to gain the technical knowledge needed:

> *I'm confident to use [digital technology] but there's certain parts like sending emails I'm not good at, I can receive emails but not send them.* [Service user participant 21]

To overcome the lack of knowledge, some service providers offered different types of training. They ranged from setting up email addresses to using social media:

> *It's giving them the chance to set up things like email accounts, social media accounts as well, so they can keep in touch with friends and family and stuff like that. So, it is having a massive effect.* [Service provider participant 2]

Training could boost the confidence of service users, which could further improve their engagement with technology:

> *It's just that initial fear that, "oh, I can't do this." But then when you show them, it's like, "if you press that, that and that your whole world lights up." And just by doing that you see the reaction on their faces. Like, "oh, this isn't quite as hard as I expected it to be."* [Service provider participant 2]

Some service users suggested that it would be beneficial for free training to be widely available for people to further develop their skills in using digital technology:

> *[They should] advertise free courses to show people how to use digital devices. Like the Job Centre, you*

XSL•FO
RenderX

> *can go in and use their computers and all that now, right? But see this scheme, where I got the tablet and stuff...* [Service user participant 15]

Technology provision without appropriate training could lead to a lack of engagement or even individuals selling their devices:

> *But the other thing with that is that we've found that if you just give people a smartphone and they don't know how to use it and there's no support and there's no connectivity, of course, they're going to sell it. It's useless, it's a lump of metal in their pocket and they can get £100 for it. If you teach somebody the value of digital [technology], and people start to realize that what digital is bringing to their lives in terms of the social connection, the benefits, the finances, the health, all of it, that is worth so much more to them than the second-hand resale price of what they would get going to a Cash Converters [pawn shop] and getting £40 for a second-hand phone.* [Service provider participant 12]

We identified that one of the important and challenging issues in technology provision and adoption programs is the sustainability of engagement after initial technology provision by the program:

> *And so, we're now looking at, well, actually what's the next iteration of Connecting Scotland? And if we were to do something similar, hand out kit connectivity, what else do we need to put in place around it? Now, unfortunately we don't have the funding that we did because it was Covid funding, and lots of it has dried up. But actually, we've still got all that research and evidence to say, well, there's other things you need to do around it to help support, and it's more than just that kind of kit connectivity and then skills and support...that support word goes a lot further than I think we envisaged previously.* [Program team participant 11]

"Connecting Scotland" is a program aimed at supporting every citizen in Scotland to get on the web, thereby bridging the digital divide. It was launched in response to the COVID-19 pandemic, recognizing the importance of internet access in keeping people connected to their friends, family, and essential services, as well as enabling learning, work, and access to public services.

This means avoiding the disconnection of service users from the digital community and to keep them connected for the long term to ensure improvements in well-being and other related areas of their lives could be sustained:

> *...we should be looking for something a bit longer-term that's right for them...And that might not be a continuation of a SIM card, that might be home broadband using social tariffs or something like that. It's difficult because everyone is different, but what we're really conscious of is not allowing people after the 12 months to fall off a cliff and come back round and start from square one again, because that would just be a year wasted in a way.* [Service provider participant 10]

This sustained engagement can lead to increased inclusion in the community and reduced DRDs. Technology was seen as part of a larger suite of options to keep people included in society:

> *I think the more options that we give people, the better. I think, you know, we say this all the time, but digital [technology] isn't the way to solve drug deaths, but it's one small part of the picture... Like the more that we can connect people, the better.* [Service provider participant 7]

Sustainability of engagement, however, requires an understanding of the need for cultural changes in the sector. One participant gave an example of how improved understanding of technology use can help sustainability:

> *It's that ability to think critically and that ability to really reflect that we need to be giving people or helping people to understand. Now, not everybody will feel comfortable with it, so it's finding ways into it because if you can do that, actually their ability to then use digital tools becomes much greater because they're much more consciously aware of what they're doing. And ultimately, then, what you would do is you would have an impact on culture and society.* [Program team participant 11]

Another challenge highlighted was the economic challenges faced by individuals, leading to unintended consequences, such as the need for food, leading to the potential selling of devices:

> *A sort of £80 resale probably isn't really worth it, whereas a £200 resell is worth it. You know, if you've got nothing to eat and you've got no electricity, and I think there's a real challenge that we face probably within Digital Lifelines and the broader digital connection work around the cost-of-living crisis...I had that a couple of weeks ago with a woman who came into our access hub on the Thursday and she wasn't getting paid until the Tuesday and she didn't want to sell her device, but the reality was she had absolutely no money. She had no money. No food. So, we were able to get her like an emergency food package to keep her going over the weekend and I said this at the last Digital Lifelines meetings, like we almost need like a personal budget to go alongside the devices. So that if people are at absolute crisis point that organizations have a little bit of like wiggle room where they can do say, let's buy you a weeks' worth of shopping...in order to keep hold of your device.* [Service provider participant 3]

However, the program team highlighted that while the selling of devices was identified as a challenge, only 10% of service users engaged in DLS either lost, sold, or had their devices stolen. Of those who did sell their devices, most sold them at pawn shops and later bought them back.

Organizational Challenges

A range of challenges related to the organizational adoption of technologies by service providers require careful consideration.

XSL•FO
RenderX

As digital innovation is occurring fast, it is the adoption into practice that is perceived as a considerable challenge:

> *So, you bring out the technology, you develop apps or digital services that can do things. And the industry's absolutely fantastic, they can develop anything. The challenge is actually to integrate it into our service delivery models and that means working with people, changing their roles, and the way they go about doing what they do, and that is very, you know, again, the human being factor is the bit that always is the most challenging but also the most rewarding.* [Program team participant 6]

One reason for these adoption issues is that service providers are already under pressure, lacking time and resources to spend on planning for the effective adoption of technologies:

> *Another of the challenges is just how much work there is going on in the field of drug death prevention at the moment and lots of frontline organizations feel completely overwhelmed...we're a small fish in part of the whole of that pool. So, actually organizations have had the absolute best of intentions, but they've been pulled off from focusing on the digital inclusion element from other aspects.* [Program team participant 6]

This further leads to a lack of provision of digital services. While digital devices are made available to the service users, services for those at risk of drug-related harm are not effectively offered on the web:

> *I mean probably the best example is like we're giving people devices, but very few of these organizations have what you would really call a true digital service...generally the public want access to an addiction service or a homelessness service and can go on and find it and I can self-refer. I can speak to someone right now when I'm in that service, I know what you're doing with my data I know where I'm at in the pathway.* [Program team participant 1]

Moreover, there existed a notable gap in knowledge of service providers relating to data security concerns in the use of technologies within their organizations. For instance, the use of Google services posed potential risks to the security of service user data, and there was a lack of governance adherence to the General Data Protection Regulation.

The Journey to the Digital Transformation of the Sector

Participants across all groups described various uses and benefits of different types of technologies provided by the program. The technologies provided included different types of devices, such as smartphones, tablets, and laptops, and internet connectivity tools, with service users generally preferring smartphones:

> *In terms of the device, it's usually either a smartphone or a tablet. We have given out some Chromebooks...But I know the majority is mobile phones, probably about 60% and then 40% tablets.* [Service provider participant 12]

None of the participants described any type of new technology provided; that is, the technologies provided were laptops, smartphones, internet connections, and tablets; for example, new data-driven innovative apps were not provided. In addition, despite the provision of digital technologies, most benefits were described around existing services, for instance, making the existing knowledge more readily available, connecting more easily to their support workers, or being informed about different existing events. Only in cases of service users identifying new unintended use, such as learning new skills or identifying new support groups, resulted in some extent of innovative service delivery observed. Overall, by analyzing our findings against the model in Figure 2, we find that in the context of preventing DRD, the digital inclusion program offered current technologies mainly serving existing services, and only occasionally leading to changes in services.

Overall, we observed that the digital inclusion program used current technologies to enhance the provision of existing services. Interestingly, some participants highlighted that such programs were effective in their emphasis on teaching and promoting fundamental digital skills. While impactful, it does not cater for the transformation of existing services and practices, hence it is not a program of digital transformation:

> *I think because the Digital Lifelines work, you know, it's about basic digital skills. It's not necessarily about, you know, a big digital transformation or developing digital services. A lot of it is quite simple, on the surface.* [Program team participant 1]

Digital transformation requires rethinking and redesigning of processes, not adding a digital layer to existing cumbersome processes [29], and as the example mentioned subsequently indicates, this change of processes was not evident in many settings, neither drug-related nor other areas of health care provision:

> *Last time I needed to make a phone call to the doctor to try and get an appointment for my daughter, I made 135, I think, repeat phone calls...which is bonkers. And actually, what happened in that process was the receptionist took my details and said doctor will phone you back. Doctor then phoned me back within a couple of hours...But actually, what we've done there is instead of really thinking about a smart digital approach, is we've actually layered up two telephone calls and a face-to-face and increased costs rather than decreased costs...So, let's think about that in a different way. Quite a lot of adults are now used to doing some form of digital or telehealth care.* [Program team participant 11]

By pointing to other areas of health and care provision and highlighting examples of transformations, participants pointed to the need for similar actions in the field of prevention of DRD. In this process, COVID-19 was seen as an accelerator of transformation:

> *Yes, I think because of COVID, things pushed through a lot quicker than they maybe have before because it was needs must, and you couldn't see your GP. So, things like digital conferencing for appointments, and*

XSL•FO
RenderX

> *stuff like that, really ramped up throughout that as well. I think people through COVID have had to become acquainted with technology in a lot of ways...So, it's how do we keep plodding ahead where appropriate and utilizing the momentum?* [Program team participant 12]

Some participants were aware of the need for transformative changes to service provision, while others were mainly concerned about improving the day-to-day activities using digital technologies:

> *You'd go into the travel agents, and you'd sit down, and you'd look through the brochures...'this is the holiday that I really fancy' and then you'd sit down with somebody. Actually, doing all of that online has changed it dramatically and you need to feel confident about doing that. And the supports and the people that you could have talked to aren't always there when you need them to be there. So, that's a complete transformation of how the travel industry has shifted its way of working. So, if we're trying to do a transformation of how our health and care system works then we're going to go through similar changes when we're introducing digital into that. So, it's not just a matter of me and you talking over a video instead of sitting down and, you know, having a coffee in a café, and blethering about it. This is a really simple digital interaction because we're doing what we always do, we're just using a different channel to do it. When you're actually looking at digital services, they can completely change the way that you access something and how you actually use it, and we've not even scratched the surface of that within Digital Lifelines at this stage.* [Program team participant 6]

Hence, digital transformation needs new value creation, creating new digital services that did not exist previously in models of health and care. Such services have started to increase more recently as digital technologies are made available to those, for instance, in rural areas:

> *Those areas like, for example, the people in recovery, the people in the [rural area] that I was talking to you about. We weren't able to help them in a digital way whereas we can now because we've got the offer of the devices and people can take them away and we can help them in that way. So, that's a massive difference.* [Service provider participant 9]

As participants highlighted, digital transformation does not have to be about radical changes. Instead, even a focus on a particular niche as part of that bigger picture of transformation can improve the lives of people:

> *So, as somebody's kind of working through that element of it are there particular parts of that, that by using digital services it makes it better and it makes it easier? Do you know whether it's chat rooms or do, you know, what kinds of things it is that might actually help people at those particular little glitches, those problem areas, that they experience as they go along? So, that's the bit I'm kind of interested in and where I think we can make a contribution.* [Program team participant 6]

Digital transformation may not necessarily be achieved easily. Apart from having a clear understanding and strategic planning for transformation, other aspects of society and organizations also need to go through change. For instance, digital transformation requires particular skills; staff either need to be trained or new employees need to be hired:

> *I think part of it is that a lot of frontline staff don't even have the digital skills to pass on to the people they're working with...I think a lot of charities don't really have a grasp of the digital skill set of their staff. If you are a big social care charity with 2000 staff, we really need to get to grips with your status of digital skills and competence and where it's lacking. How would you support them and how do you train them?* [Program team participant 1]

Digital transformation also requires cultural and structural changes in organizations:

> *I think there are those cultural issues and change management within services, and again, ensuring that the staff that are supporting people feel confident and equipped to use digital.* [Program team participant 10]

Discussion

Principal Findings

This study explored the benefits and challenges of the provision of digital technologies for digital inclusion, which aimed to support people at risk of DRDs through digital technology. Findings indicate that the provision of digital technologies offers many benefits to service users and providers. The technologies examined ranged from smartphones, tablets, and laptops to internet connectivity tools. We identified 3 main categories of perceived benefits for these devices: increased connectivity for informal support, increased access to services, and improved well-being.

This research found that provision of such devices and tools allowed service users to connect with others, overcome loneliness, create friendships, and build a sense of community, thus reducing the digital divide. Participants provided examples of connectivity to family and friends, as well as the importance of connectivity to a range of harm reduction and support services. Increased access to services was another benefit of digital devices, as described by both service users and providers, as people who were originally excluded from receiving information and support, due to a lack of access to technology, had an increased chance of being engaged in receiving such services. In most cases, they described better access to existing services, such as more readily available information or access to support workers, and better identification of existing events through web-based news, email, and media.

A novel finding was that digital technologies were seen to have significant potential in improving the well-being of the community by improving mental health and emotional well-being. Participants noted that access to digital technology

XSL•FO
RenderX

provided a way to improve self-esteem and confidence. Some also noted that they used digital technologies provided by the program to learn new skills, which appeared to improve their mental well-being.

In general, the provision of digital technologies led to improved access to existing services and sources of support and information. At the time of this research, only in occasional cases, new services were designed, or existing services were somewhat transformed to web-based services. However, such cases had significant potential; for instance, as a participant highlighted because of new service provisions, 6 potential cases of prevention of overdose death were reported.

While the provision of technology has many potential benefits, we also identified technological, people-related, organizational, and macroenvironmental challenges that led to limited adoption and effective use of technology. Examples of the most evident challenges were inadequate digital literacy of service users (which could have other negative knock-on effects such as short intervals of engagement and periods of stress), technology usability issues, service providers being under pressure and being underresourced leading to limited planning for the offering of new services and effective adoption of technologies, and negative public and media perceptions of the program due to the particular context—drug use.

Comparison With Prior Work

These findings contribute novel information in several ways. Our findings contribute to the literature on the potential benefits of digital technologies to tackle drug overdose by identifying 3 categories of benefits. Previous studies, such as McClure et al [30] and Ozga et al [31], found that there are still many concerns about the digital divide for marginalized populations, and our study shows that national programs that provide technologies can help mitigate some of these problems by providing better access to informal support, better access to existing formal services, and offering new or transformed services. This supports studies that show that digital technologies can be effective in treating drug use disorders [32,33]. While adoption of digital technologies is useful, our study found that implementation of technology alone is not sufficient for tackling the immense problems of drug overdose. Participants noted that many organizations were using some digital technology for providing activities and support to service users, but that was not transforming their services. The findings highlight the need for new value creation to enable new pathways such as new digital services. For instance, digital technologies can be used to redesign processes to streamline the existing ones or allow organizations to offer new services that are not offered if digital technologies were unavailable, such as remote consultations. Hence, digital transformation, a change process in organizations or societies driven by the use of digital technologies [34], is needed to better tackle such a complex issue. Digital transformation is not just about adopting digital technologies but also about redesigning and rethinking the existing processes to offer new and improved services [34]. Our evaluation confirmed that in drug harm reduction services, the digital transformation process requires particular skills, cultural and structural changes in organizations [29], and the transformation of society as a whole [35].

Digital transformation is the process of integrating digital technologies into an organization's operations, strategies, and services to drive innovation, efficiency, and improved outcomes [19]. For drug service providers, digital transformation involves leveraging digital tools and technologies to enhance the delivery of addiction treatment and support services. We integrated a conceptual framework provided by Daneshvar et al [27] to explore the relationship between digital technology and the services provided to determine how far along the journey the sector is in relation to digital transformation. We show that currently, in most cases, existing digital tools are used to deliver existing services (in a digital form). Only very few new services were offered because of adopting digital tools (both existing and new digital tools). Hence, we propose that in the field of prevention of overdose death, we are only at the early stages of digital transformation. There are several reasons for this as either identified in previous research or by our study. For instance, many organizations do not have the required digital infrastructures [36], there are challenges with regard to data security and privacy [37-39], limited digital literacy [40], and technology usability in health and care [41,42]. Our study contributes to these areas of concern by indicating ways in which these challenges can be addressed. For example, training programs offered by service providers (eg, setting up accounts and email addresses to use social media) can be a very effective approach to boost the confidence of service users, which further improves their engagement with technology. We found different preferences and feelings expressed toward different devices offered by the digital inclusion program. Smartphones and tablets were considered more learnable and usable to those who were less digitally proficient, as opposed to laptops and Chromebooks. Thus, the usability of digital technologies was important. Hence, digital transformation programs need to consider these aspects when designing or transforming services.

Implications for Policy, Practice, and Research

The findings underscore the necessity for policies and programs aimed at addressing digital inclusion to ameliorate the digital divide among marginalized populations. Provision of access to devices and internet connectivity can facilitate social connections and access to support services for people who use drugs. Government policies are essential to ensure the secure and ethical use of digital technologies in the health care sector, encompassing protections for sensitive health data. Regulatory frameworks surrounding privacy and consent should be fortified. Harm reduction organizations could integrate digital technologies, such as telehealth, mobile health apps, and web-based tools, into service delivery models where appropriate. Digital modalities can enhance convenience, accessibility, and engagement. Service providers could necessitate training in both using new technologies and imparting digital skills to service users. Supporting digital literacy and usability is imperative for successful adoption. Practitioners must consider user preferences and be cautious of a one-size-fits-all approach to technology.

XSL•FO
RenderX

There is a need for more rigorous studies and formative evaluations to establish the efficacy of specific digital health interventions for substance use disorders, particularly newer mobile health apps and wearable biosensors. Research should investigate how technology influences intended treatment outcomes. Qualitative data on user experiences are also invaluable. Cost-effectiveness studies would bolster decision-making regarding investment in digital modalities versus traditional delivery methods. Further exploration of barriers to digital transformation of addiction services could steer stakeholders toward transitioning from simple digitization of existing services to more innovative care models. Quantitative studies, such as randomized controlled trials, might be possible in the future to compare outcomes of those with particular digital technologies. In addition, a mixed methods study could investigate whether the availability of technology might lead to unintended consequences, such as facilitating connections with drug providers.

Strengths and Limitations

This work is timely and coincides with the urgent impetus to respond to DRD and significant political appetite in Scotland, as well as the enhanced availability of digital health care funding because of COVID-19 [20,36]. To the knowledge of the authors, this is the first study to attempt to combine primary data regarding the digital inclusion for harm reduction and prevention of overdose death from the perspective of service users, service providers, and the program team. The specific strengths of this study lie in those areas that were identified as complementing academic literature. For example, the themes presented here were intentionally selected for their ability to elucidate the sociotechnical task faced by technologists and intervention implementation teams. The subtleties of the challenge, such as the delicate balances required to achieve acceptability, may be highly specific to this population. The opportunity presented to support the development of the TPOM for this sector and sociotechnical theory more broadly was also welcomed. Another strength of the work is the total number of interviews (n=47) conducted for this study, which provides a comprehensive perspective on the topic across stakeholders.

A limitation of the work is that our sample population consisted of people who participated in the program and therefore provided with access and support to digital tech. While we tried to speak to people who declined to participate in DLS, there was no engagement from them. Participation from individuals and groups involved in DRD prevention at the level of policy and elected officials would have added finer details concerning the macroenvironment and, thus, greater strategic insight into the overall feasibility of digital programs.

Conclusions

Digital inclusion programs can offer a platform for engagement of individuals at risk of drug-related harm to offer them a space for accessing informal support, formal existing services, or, in some cases, experiencing and benefiting from new services. However, challenges remain, including service users' concerns regarding data security, digital literacy, and technology usability issues. Addressing these concerns necessitates a training program for both service providers and users to enable them to fully embrace the opportunities provided by digital connections, as well as initiating a cultural and societal shift toward the development and use of technologies to tackle DRDs. Moving beyond device provision, such programs should focus on the development of new digital services. Digital transformation is complex and may take a long time to achieve, but it can have a significant impact on preventing DRDs. In the modern landscape of health and social care, access to digital technology for all is no longer a luxury but an essential necessity.

Acknowledgments

The authors would like to thank the Digital Lifelines Scotland and Technology Enabled Care program of the Scottish Government, who funded this project. They would like to thank all the participants across services, the program team, and the delivery team, as well as those who helped with recruitment. The authors also want to thank Tessa Parkes, Josh Dumbrell, and Joe Schofield, who supported other aspects of the study.

Authors' Contributions

HD and CM conceptualized, designed, and developed this part of the analysis. HD wrote the first draft of the paper. CM and HC led the design of the wider study. GS, HD, and JG carried out all data collection. GS and HD performed data coding, while GS, HD, HC, and CM conducted data analysis and prepared the write-up. Our team comprised members of mixed gender, all of whom are educated to the master's or PhD level and have extensive experience working with people who use drugs and service staff. GS has relevant lived experience. We discussed the data collection process and analysis as a team, ensuring our findings were grounded in participants' experiences. All authors participated in reviewing, editing, and finalizing the manuscript, and all approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topic guide.
[DOCX File , 18 KB - jmir_v26i1e52345_app1.docx]

XSL•FO
RenderX

References

1. World drug report 2021. United Nations. 2021. URL: https://www.unodc.org/unodc/en/data-and-analysis/wdr2021.html [accessed 2024-09-11]
2. McAuley A, Matheson C, Robertson JR. From the clinic to the street: the changing role of benzodiazepines in the Scottish overdose epidemic. Int J Drug Policy 2022 Feb;100:103512. [doi: 10.1016/j.drugpo.2021.103512] [Medline: 34753047]
3. Drug-related deaths in Scotland in 2022. National Records of Scotland. 2023. URL: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/deaths/drug-related-deaths-in-scotland/2022 [accessed 2023-10-24]
4. Jalali MS, Botticelli M, Hwang RC, Koh HK, McHugh RK. The opioid crisis: a contextual, social-ecological framework. Health Res Policy Syst 2020 Aug 06;18(1):87 [FREE Full text] [doi: 10.1186/s12961-020-00596-8] [Medline: 32762700]
5. Cicero TJ, Ellis MS, Kasper ZA. Increased use of heroin as an initiating opioid of abuse. Addict Behav 2017 Nov;74:63-66. [doi: 10.1016/j.addbeh.2017.05.030] [Medline: 28582659]
6. Heroin research report. National Institute on Drug Abuse. URL: https://nida.nih.gov/publications/drugfacts/heroin [accessed 2023-06-28]
7. Opioid overdose. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/opioid-overdose [accessed 2023-06-28]
8. National mission on drug deaths: plan 2022-2026. The Scottish Government. 2022 Aug. URL: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2022/08/national-drugs-mission-plan-2022-2026/documents/national-mission-drug-deaths-plan-2022-2026/national-mission-drug-deaths-plan-2022-2026/govscot%3Adocument/national-mission-drug-deaths-plan-2022-2026.pdf [accessed 2024-09-11]
9. Drug Deaths Taskforce response: a cross Government approach. The Scottish Government. 2023 Jan. URL: https://www.drugsandalcohol.ie/37914/ [accessed 2024-09-11]
10. Welcome to Digital Lifelines Scotland. Digital Lifelines Scotland. URL: https://digitallifelines.scot/ [accessed 2024-09-11]
11. Reisdorf B, Rhinesmith C. Digital inclusion as a core component of social inclusion. Soc Incl 2020 May 14;8(2):132-137. [doi: 10.17645/si.v8i2.3184]
12. Oteo A, Daneshvar H, Baldacchino A, Matheson C. Overdose alert and response technologies: state-of-the-art review. J Med Internet Res 2023 Feb 15;25:e40389 [FREE Full text] [doi: 10.2196/40389] [Medline: 36790860]
13. Daneshvar H, Oteo A, Schofield J, Parkes T, Matheson C, Baldacchino A. Mapping of overdose detection and alert technologies: a summary. Universities of Stirling and St Andrews. 2022 Feb. URL: https://digitallifelines.scot/media/1115/digital-lifelines-odart-mapping-summary.pdf [accessed 2024-09-11]
14. Molfenter T, Roget N, Chaple M, Behlman S, Cody O, Hartzler B, et al. Use of telehealth in substance use disorder services during and after COVID-19: online survey study. JMIR Ment Health 2021 Feb 08;8(2):e25835 [FREE Full text] [doi: 10.2196/25835] [Medline: 33481760]
15. Mimigiannis C, Gallos P, Mantas J. The drug addicts' usage of information and communication technologies. Stud Health Technol Inform 2020 Nov 23;275:236-237. [doi: 10.3233/SHTI200734] [Medline: 33227780]
16. Ganesh R, Rao R, Deb KS, Bhad R, Yadav D. Digital capacity and interest in mHealth interventions among individuals on opioid agonist maintenance treatment: a cross-sectional community-based study. Indian J Psychol Med 2022 Jul;44(4):354-358 [FREE Full text] [doi: 10.1177/02537176211027239] [Medline: 35949629]
17. Carrà G, Crocamo C, Humphris G, Tabacchi T, Bartoli F, Neufeind J, et al. Engagement in the Overdose Risk Information (ORION) e-health tool for opioid overdose prevention and self-efficacy: a preliminary study. Cyberpsychol Behav Soc Netw 2017 Dec;20(12):762-768. [doi: 10.1089/cyber.2016.0744] [Medline: 29148826]
18. Roth AM, Tran NK, Cocchiaro B, Mitchell AK, Schwartz DG, Hensel DJ, et al. Wearable biosensors have the potential to monitor physiological changes associated with opioid overdose among people who use drugs: a proof-of-concept study in a real-world setting. Drug Alcohol Depend 2021 Dec 01;229(Pt A):109138 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.109138] [Medline: 34781097]
19. Hess T, Matt C, Benlian A, Wiesböck F. Options for formulating a digital transformation strategy. MIS Q Exec 2016 Jun;15(2):123-139. [doi: 10.4324/9780429286797-7]
20. Evaluation of the Digital Lifelines Scotland Programme, March 2023. Digital Lifelines Scotland. 2023 Jul 31. URL: https://digitallifelines.scot/news-events/evaluation-of-the-digital-lifelines-scotland-programme-march-2023/ [accessed 2024-09-11]
21. Cresswell K, Williams R, Sheikh A. Developing and applying a formative evaluation framework for health information technology implementations: qualitative investigation. J Med Internet Res 2020 Jun 10;22(6):e15068 [FREE Full text] [doi: 10.2196/15068] [Medline: 32519968]
22. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
23. What are third sector organisations? National Audit Office. URL: https://www.nao.org.uk/successful-commissioning/introduction/what-are-civil-society-organisations-and-their-benefits-for-commissioners/ [accessed 2024-02-22]
24. Zamawe FC. The implication of using NVivo software in qualitative data analysis: evidence-based reflections. Malawi Med J 2015 Mar;27(1):13-15 [FREE Full text] [doi: 10.4314/mmj.v27i1.4] [Medline: 26137192]

XSL•FO
RenderX

25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Miles MB, Huberman AM. Qualitative Data Analysis: An Expanded Sourcebook. Thousand Oaks, CA: SAGE Publications; 1994.
27. Daneshvar H, Anderson S, Williams R, Mozaffar H. How can social media lead to co-production (co-delivery) of new services for the elderly population? A qualitative study. JMIR Hum Factors 2018 Feb 12;5(1):e5 [FREE Full text] [doi: 10.2196/humanfactors.7856] [Medline: 29434014]
28. Strachan G, Daneshvar H, Carver H, Greenhalgh J, Matheson C. Using digital technology to reduce drug-related harms: a targeted service users' perspective of the digital lifelines Scotland programme. Harm Reduct J 2024 Jul 01;21(1):128 [FREE Full text] [doi: 10.1186/s12954-024-01012-y] [Medline: 38951880]
29. Matt C, Hess T, Benlian A. Digital transformation strategies. Bus Inf Syst Eng 2015 Aug 4;57:339-343. [doi: 10.1007/s12599-015-0401-5]
30. McClure EA, Acquavita SP, Harding E, Stitzer ML. Utilization of communication technology by patients enrolled in substance abuse treatment. Drug Alcohol Depend 2013 Apr 01;129(1-2):145-150 [FREE Full text] [doi: 10.1016/j.drugalcdep.2012.10.003] [Medline: 23107600]
31. Ozga JE, Paquette C, Syvertsen JL, Pollini RA. Mobile phone and internet use among people who inject drugs: implications for mobile health interventions. Subst Abus 2022;43(1):592-597 [FREE Full text] [doi: 10.1080/08897077.2021.1975871] [Medline: 34491889]
32. Monarque M, Sabetti J, Ferrari M. Digital interventions for substance use disorders in young people: rapid review. Subst Abuse Treat Prev Policy 2023 Feb 17;18(1):13 [FREE Full text] [doi: 10.1186/s13011-023-00518-1] [Medline: 36805783]
33. Marsch LA. Digital health and addiction. Curr Opin Syst Biol 2020 Apr;20:1-7 [FREE Full text] [doi: 10.1016/j.coisb.2020.07.004] [Medline: 32885099]
34. Vial G. Understanding digital transformation: a review and a research agenda. J Strateg Inf Syst 2019 Jun;28(2):118-144. [doi: 10.1016/j.jsis.2019.01.003]
35. Braa J, Sahay S, Monteiro E. Design theory for societal digital transformation: the case of digital global health. J Assoc Inf Syst 2023;24(6):1645-1669. [doi: 10.17705/1jais.00816]
36. Matheson C, Carver H, Parkes T, Daneshvar H, Connell C, Dumbrell J, et al. Digital inclusion to prevent drug related deaths: scoping user needs. Digital Lifelines Scotland. 2022. URL: https://www.stir.ac.uk/research/hub/publication/1825538#panel_1_1 [accessed 2024-09-11]
37. Molfenter T, Brown R, O'Neill A, Kopetsky E, Toy A. Use of telemedicine in addiction treatment: current practices and organizational implementation characteristics. Int J Telemed Appl 2018 Mar 11;2018:3932643 [FREE Full text] [doi: 10.1155/2018/3932643] [Medline: 29713341]
38. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 06;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
39. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010 Oct;19 Suppl 3(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
40. Pillai NM, Mohan A, Gutjahr G, Nedungadi P. Digital literacy and substance abuse awareness using tablets in indigenous settlements in Kerala. In: Proceedings of the IEEE 18th International Conference on Advanced Learning Technologies. 2018 Presented at: ICALT 2018; July 9-13, 2018; Mumbai, India. [doi: 10.1109/icalt.2018.00026]
41. Holden RJ, Campbell NL, Abebe E, Clark DO, Ferguson D, Bodke K, et al. Usability and feasibility of consumer-facing technology to reduce unsafe medication use by older adults. Res Social Adm Pharm 2020 Jan;16(1):54-61 [FREE Full text] [doi: 10.1016/j.sapharm.2019.02.011] [Medline: 30853507]
42. de Joode E, van Heugten C, Verhey F, van Boxtel M. Efficacy and usability of assistive technology for patients with cognitive deficits: a systematic review. Clin Rehabil 2010 Aug;24(8):701-714. [doi: 10.1177/0269215510367551] [Medline: 20543021]

Abbreviations

DLS: Digital Lifelines Scotland
DRD: drug-related death
TPOM: Technology, People, Organization, and Macroenvironment

XSL•FO
RenderX

Edited by N Cahill; submitted 01.09.23; peer-reviewed by S Hinder, M Barratt, H Veldandi; comments to author 23.01.24; revised version received 23.02.24; accepted 31.07.24; published 24.09.24.

Please cite as:
Daneshvar H, Carver H, Strachan G, Greenhalgh J, Matheson C
From Digital Inclusion to Digital Transformation in the Prevention of Drug-Related Deaths in Scotland: Qualitative Study
J Med Internet Res 2024;26:e52345
URL: https://www.jmir.org/2024/1/e52345
doi:10.2196/52345
PMID:

©Hadi Daneshvar, Hannah Carver, Graeme Strachan, Jessica Greenhalgh, Catriona Matheson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Digitalization of Social Care in England and Implications for Older, Unpaid Carers: Constructionist Thematic Analysis

Anastasia Rousaki[1], PhD; Efpraxia D Zamani[2], PhD; Laura Sbaffi[1], PhD; Kate Hamblin[3], PhD; Rachael Black[3], PhD

[1]Information School, University of Sheffield, Sheffield, United Kingdom

[2]Durham University Business School, Durham, United Kingdom

[3]Centre for Care, University of Sheffield, Sheffield, United Kingdom

Corresponding Author:
Laura Sbaffi, PhD
Information School
University of Sheffield
The Wave
2 Whitham Rd
Sheffield, S10 2SJ
United Kingdom
Phone: 44 0114 222 268
Email: l.sbaffi@sheffield.ac.uk

Abstract

Background: Globally, populations are aging, generating concerns about the sustainability of health and social care provision. In terms of the public provision of social care in particular, unpaid carers provide much of the support to people with disabilities and older people. In addition, there is an increased onus in many countries on digital transformation projects, in the hope that the digitalization of services can create efficiencies and savings in both costs and care labor. In England, the focus of this paper, the shift to digital services is also framed as a means to enhance choice and control for older unpaid carers, while being part of a broader offering that includes nondigital alternatives and support to mitigate digital exclusion.

Objective: This study examines the impact of digitalization on older, unpaid carers—a group more likely to be both expected to engage digitally with services and at risk of digital exclusion—in England, focusing on their lived experiences in terms of caring and access to social care.

Methods: We used a constructionist approach to thematic analysis, where data from 48 older unpaid carers collected through focus groups were analyzed using thematic analysis, resulting in 4 prevailing themes.

Results: Our findings indicated that while unpaid carers largely acknowledge the benefits of digitalization, they also highlight several points of failure, whereby engagement with digital spaces is experienced as coercive and exacerbates feelings of exclusion. These are further worsened by government failures to address issues of connectivity, imposing additional financial burdens and complicating tasks such as benefit applications.

Conclusions: In this study, we have highlighted the need for greater involvement in shaping both policy and technological solutions, which in turn will be more inclusive and aligned to the aspirations and circumstances of older carers.

(J Med Internet Res 2024;26:e60056) doi:10.2196/60056

KEYWORDS

social care; England; digitalization; digital transformation; unpaid care; mobile phone

Introduction

There are concerns globally regarding the sustainability of health and social care systems as populations are aging and growing numbers of individuals aged ≥65 years are facing unmet care needs [1]. Unpaid or *informal* care has long been part of "care diamonds" (the configuration of care provision among states, markets, families, and the third sector [2]). Indeed, the role of unpaid care is increasing in the provision of care, as even in nations where statutory provision had been fairly comprehensive, retrenchment is "bringing the family in [to caring arrangements] through the back door" [3]. Therefore, unpaid carers provide a large majority of all care.

XSL•FO
RenderX

In the United Kingdom—and specifically England, which is the empirical focus of this paper—unpaid care is an important societal phenomenon. Census data indicate there are 5.7 million unpaid carers [4], but based on other analyses and due to challenges around self-identification, the numbers seem to be closer to 10.6 million [5]. Aging and unpaid care are closely related, with older age groups providing the highest number of hours of care [4]. Estimates also put the value of care provided by unpaid carers in England and Wales annually at £162 billion (US $211 billion), exceeding the annual expenditure on the National Health Service (NHS) [6]. However, in the past decade, there have been significant cutbacks to social care funding and provision, with such austerity measures often being underpinned by neoliberal ideologies [7]. This financialized, technology-oriented policy agenda places the onus of care provision and financing onto families and communities while limiting their capacity to provide care via the imposition of austerity [8]. This is particularly the case in the United Kingdom, with care-related austerity policies impacting mainly susceptible individuals, namely minority ethnic groups, women, and people with disabilities [7].

The combination of such austerity policies and the increased need for care has resulted in the increasing use of digital interventions in social care, where these primarily aim toward cost reduction [9]. Digitalization has now become a global phenomenon impacting both professional and personal life through the increasing prevalence of digital technologies for the development and provision of mobile health, health information systems and technologies, wearable devices, and telehealth. Within public discourse, such interventions are often promoted by key political stakeholders, who suggest that digitalization will improve care almost deterministically by creating efficiencies [10]. In relation to social care, technologies have been espoused by the Secretaries of State for Health and Social Care in England as “transformative” and as a means to create efficiencies and “free up” care labor and financial resources [11].

The digitalization of care, or the shift to the use of technologies in the provision of care including through web-based services, is also framed as a means to enhance choice and control for people in receipt of support and unpaid carers. The NHS App and the NHS website aim to “become platforms through which people, their families and unpaid carers can access more services and resources proactively have more control over their care benefit from more personalised and preventative offers” [12]. The underpinning notion is that digital options will increase both choice and allow for “informed decision-making” through access to data and a wider range of services. While policy documents also indicate that nondigital options and services are to be retained for those who cannot or do not wish to use web-based services [12], technological visions often lead to exclusions and further marginalization [13]. The reasons are diverse as several factors contribute toward digital exclusion, including gender, age, gender identity, disabilities, and ethnicity, to name only a few [14], as well as unpaid care [15]. Among these, age is a significant factor. Until recently, older adults aged ≥65 years constituted only 5.5% of the global digital population [16], and consequently, the implications of digitalizing essential tasks for older adults can be particularly pronounced: many everyday activities nowadays are either solely or primarily web-based (eg, benefits, tickets, and banking), and these are known to pose difficulties for older adults, thus further reinforcing existing exclusions [17]. In addition, digital exclusion can be accentuated through the confluence of roles and identities. For instance, older adults are more likely to be engaged in unpaid care (in 2021, two-thirds of unpaid carers were older adults) [18]. Therefore, digital exclusion of older adults, coupled with their likelihood to be unpaid carers, has spurred growing concerns [19].

Given the substantial interrelation among aging, digital exclusion, care, and socioeconomic factors [20], we argue that it is imperative to elucidate the effects of digitalization of care on older, unpaid carers. In this study, we asked, “What is the impact of the digitalization of care on older adults and particularly those with care-related responsibilities?” and built on previous work by examining unintended consequences and empirically interrogating the promissory discourse of digital health [21] and, in this case, social care.

We contextualized our study within England, United Kingdom, due to its distinctive social care landscape and the policy onus on encouraging widespread adoption of care-related technologies [1,10]. In subsequent sections, we have first provided an overview of our methods and then presented our findings in detail. This is followed by a critical analysis and discussion of our findings in relation to the digitalization of care and the connections and tensions between unpaid carers’ experiences and the prevailing political focus on digitizing social care.

Methods

Overview

We conducted focus groups in care centers located in Liverpool, Devon, Leeds, Derby, and Bath and virtually via Google Meets (Google LLC) between September 2023 and February 2024. The location choice constitutes a contextual exploration of the effects of digitalization in relation to the north-south divide. The divide is a postwar Britain concept, characterizing the bifurcation of the varying socioeconomic conditions across England [22], which often negatively affect the northern regions.

Focus groups allow participants to coconstruct meaning and explicate normative discourses and how inequalities are worked up via discussions [23]. These focus groups were designed with the help of story completion prompts [24] that entailed stories relating to digital poverty, digitalization policies, and web safety. To provide a sense of detachment, we used gender-neutral names and third-person story prompts.

Recruitment

Recruitment was conducted via carer centers and our research partners. This outreach strategy involved advertising through carer centers’ mailing lists and newsletters, displaying posters in the carer centers, and promoting the research during their coffee days. Details regarding our focus group participants can be found in Table 1.

XSL•FO
RenderX

Table 1. Focus groups and participant information (all participants were aged ≥50 years).

Location	Participants (N=48), n (%)	Women (n=35), n (%)	Men (n=8), n (%)
Liverpool	13 (27)	9 (19)	4 (8)
Devon	14 (29)	14 (29)	0 (0)
Leeds	7 (15)	6 (13)	1 (2)
Bath	5 (10)	3 (6)	2 (4)
Derby	5 (10)	__[a]	—
Virtual focus group	4 (8)	3 (6)	1 (2)

[a]Participants preferred not to disclose their identity.

Data Analysis

The lived experiences of care, as well as its accompanying sociocultural structures, institutions, and apparatuses, constitute a complex topic; this is particularly the case when studying it through the lens of digitalization [25]. Consequently, we used a constructionist thematic analysis that focused on the construction of ideological subjects [26,27]. We began by immersing ourselves in the data, whereby we repeatedly and iteratively reviewed the collected data to independently identify codes and themes; document our observations; and organize the data into codes, subthemes, and themes. We began coding line by line to familiarize ourselves, and we then revisited the preliminary set of codes to identify overlaps, similarities, and differences. This resulted in merging codes together and eventually forming subthemes. These were then cross-checked between members of the research team again for cohesiveness and consistency. Finally, initial themes were organized and discussed; they were then reevaluated within the entire dataset, refined, and labeled.

Throughout the analysis and write-up, we considered questions such as the implications of the themes and potential justifications of participants' constructions within the sociopolitical context. Our analysis resulted in 4 prevalent themes, which are discussed in subsequent sections (Multimedia Appendix 1).

Ethical Considerations

The research received ethics approval from the University of Sheffield, United Kingdom (approval 053260), and due to the sensitive nature of our research, the focus groups were not audio recorded and we kept handwritten notes. No personal information was collected from participants, and all potentially identifiable information was anonymized or removed from the data. Before the workshops, participants received a comprehensive information sheet describing the project's purpose in lay terms and signed a consent form. The consent form outlined their right to withdraw from the study at any time. Compensation was provided to each participant in the form of a £40 (US $52) shopping voucher.

Results

Overview

In what follows, we present the findings of our thematic analysis. In brief, our findings indicate that older, unpaid carers and people needing care perceive the digitalization of social care as an experience of them being pushed *online* (intended here as being forced to engage with the digital), without however having the required support or the provisions necessary for meaningfully interacting with the digital world. This in turn is a great digital disconnect element, whereby they feel as if the lack of connectivity and the associated costs, coupled with the digital push, requires them to negotiate such support and provisions in market (or economy) terms to make their case. Consequently, while there are benefits in the use of digital technologies in social care (eg, accessibility), there are structural challenges that prohibit the fulfillment of digitalization's potential.

Being "Pushed" Online

Overview

Carers constructed their engagement with digital technologies as compulsory and reported feeling coerced into *online* engagement, describing the experience as being "pushed" toward it. This pressure exacerbated their sense of exclusion, as viable offline alternatives or simpler systems were hard to access or completely substituted. They further described this by highlighting that their engagement with web-based care services is based on pseudoconsent, characterized by the lack of informed decision-making or alternative options. Participants advocated for personalized, human-centric approaches and preserving face-to-face interactions, particularly for future generations who may encounter similar challenges. Within this theme, we found the subthemes explored in subsequent sections.

Lack of Agency and Pseudoconsent

Our participants described the digitalization of care in ways that appear ideologically hegemonic [28], whereby such efforts are driven by assumptions of equal access to web-based communications. They often feel obligated to navigate digital services and the digital world in general while experiencing challenges with web-based transactions. The pressure to exclusively adopt digital processes was a common theme (eg, "I feel like I'm being pushed online"); however, participants' constructions were in tension with such societal assumptions, prompting reflections on the absence of choice within a democratic society. In many ways, participants resisted the current status quo with remarks such as "I want people to have an alternative," providing clear arguments as to why this was important.

XSL•FO
RenderX

They emphasized the problematic nature of the expectation to conduct all activities digitally, questioning the lack of consent regarding their involvement in digitized activities. Participants questioned the data-retention practices of companies, as well as their lengthy terms and conditions (T&Cs), with participants being uncertain about the consequences of agreeing to confusing disclaimer notices ("These documents are a legal contract, meaning they can't get the blame for anything"). In turn, such feelings of coercion further increased, due to the sense of inevitability in agreeing to T&Cs, as failure to do so would result in the denial of essential services ("You have to agree, or you don't get the services"). Such power imbalances were constructed as inherently related to the nature of caring; participants perceived consenting to anything as a binary, a choice between contributing to the well-being of their loved one or having to face the potential consequences of not accessing such services.

Other assumptions they challenged relate to those regarding web-based safety and security and confidentiality. Carers expressed apprehension about providing information on the web, citing unclear T&Cs and the intimidating experience of having to navigate numerous pages. They further noted that there is added complexity in comprehending content, often due to the absence of accessible language ("Deliberately written to be non-informative!"). App permissions (eg, permission to access the computer's camera) raised further inquiries, with 1 participant querying, "Why, what else are they looking for?"

Maintaining Alternative Options

In contradiction to the policy statements regarding web-based services being a part of "multichannel offers" [12], participants highlighted concerns about the diminishing availability of traditional support systems, particularly noting the post–COVID-19 era's dearth of in-person options and telephone helplines. Accessing services remotely can be challenging, as often it is impossible to contact anyone for further support. They thus resisted the transition of care-related companies to digital platforms, which necessitate web-based access and the navigation of digital apps. Instead, they highlighted that reading information on paper is more straightforward.

The importance of being able to choose and having options was framed as a matter of inclusivity ("We have to be inclusive, we can't make people do it"), emphasizing the importance of ensuring that every organization provides alternatives. Participants suggested that having a telephone consultation in person would be helpful, as it takes the same amount of time and maintains a "human" element ("People need personalisation and a human touch and technology takes that away").

Furthermore, carers voiced the need for a round-the-clock helpline where they can interact with a live person instead of navigating automated prompts ("I don't want to hear horrible stupid inanimate words," expressing a preference for speaking to real people over automated systems). Stressing the necessity for 24/7 availability, carers argued that nighttime incidents pose particular challenges, as they are characterized by limited support, increased isolation and vulnerability, and by carers being more fatigued and stressed.

Digital Alienation and Exclusion

Participants vividly described technology as "dangerous" and challenging ("As soon as someone tells me, 'I send you the link,' my heart sinks!"), in terms of contributing to their marginalization, even in seemingly positive cases ("Technology is amazing but it excludes so many people"). Feelings of alienation and a lack of support were common discussion points (eg, "nobody cares"). For example, if any information is missing, it is often considered the carers' fault ("you feel like you are up against it all the time"). This lack of support further feeds into experiencing neglect (eg, "abandoned in the wilderness" and "who is gonna care for us"), even in cases where they need to use blood pressure monitors without assistance. Another element that contributes to feelings of alienation is when they are left facing black screens and computers ("It makes you feel more isolated").

Feelings of exclusion were further communicated in terms of generational disparities, with reference to the unfulfilled promises of technology. Society is perceived as depriving older individuals of the choice to engage with technology at their own pace or to learn according to their preferences ("Society is not allowing older people to have a choice regarding using technology or to learn at their own pace"), which is interpreted as abandonment and lack of caring ("Nobody cares about me"; "People are not kind to older people"). Participants felt like outsiders compared with technologically proficient younger generations, citing anxiety caused by technological complexity. For example, while talking about the rapid pace of technological advancement, 1 participant constructed their predicament by using an example of a computer purchased just a few years ago that is already deemed "vintage" and said that "you are fighting the up-to-date-ness," emphasizing the struggle to stay current.

Supporting Carers to "Digitalize"

Overview

Carers highlighted the lack of assistance to access digital and web-based support as a barrier to the shift toward "digitalization." In their discussions, they stressed the pervasive requirement for assistance in their everyday lives and juxtaposed this against the lack of formal instruction in navigating web-based systems and training (eg, in terms of privacy and safety), which could alleviate some of their fears regarding data security. They constructed digitalization as inherently ableist and advocated regarding addressing ableism and accessibility concerns via inclusive policy initiatives and the development of tailored assistive technologies.

Skills Gap and Training Needs

Pushing individuals toward digital services without providing guidance was heavily critiqued. Participants highlighted the lack of accessible resources for queries, using rhetorical questions such as "Am I doing this right?" with humor. They reported poor explanations in instructional videos, making it challenging for those without internet access. The public library, once a place for help, is no longer reliable due to cutbacks. Similar concerns were raised about general practitioner (GP) surgeries. In relation to this, participants resisted the notion of digitalization as "common sense" by challenging the occasional

XSL•FO
RenderX

supply of free devices, stating that it does not tackle the need for assistance to use devices in practice.

Further reflecting on the nature of their personal circumstances, participants framed digitalization as social exclusion toward older carers. Indeed, the need for assistance was omnipresent in various aspects of daily life, including banking, health care, and other services. Participants expressed concern about the diminishing availability of in-person help compared with the pre–COVID-19 era and emphasized the need for training in the broader community and among care-related professionals, advocating for community-based training because, while support may be available, accessing it requires either using the internet or commuting. Some of their suggestions included drop-in centers for carers to seek assistance, emphasizing the urgency for policies addressing the assistance gap, possibly with tailored training modules and continuous support available.

Concerns Regarding Privacy and Surveillance

One of the major areas in which participants indicated the need for support and training is that of internet safety and privacy, as the digitalization of essential services brings forward concerns regarding data (eg, “My safety is at risk”), which can be particularly problematic when one is not aware of how their data are handled (“Unknown lurking behind the door”).

These concerns were raised as part of broader skepticism toward emerging technologies. Participants constructed a sense of panoptical surveillance, where their data and information can be easily accessible, leading to concerns and constructions about the private domain becoming public (“Nothing is accurate or private; everything is virtual and everything is fake”; “I find that worrying, that anybody can tap into what you are saying in your own home”; “Hospital at homes...everything to do with your treatment is public knowledge”). This sense of panoptical surveillance was described as a contemporary societal phenomenon (eg, “...trivialised our generation...we would never go around announcing things to 100 people...telling people all our personal details”).

Data protection is thus a highly sensitive topic, giving rise to questions regarding the security of technologies used in social care (“How secure is all of this?” “How many people read that?” and “I don’t want my stuff on there”). Yet, such perspectives are further influenced by a lack of understanding regarding the purposes of data collection and use (“People are worried about where data go and what happens to their information”; “we don’t know what is happening to this information”). Considering the domain of social care, which often intersects with health care, carers are concerned that while technology is used for efficiency, providers (including the British NHS) are still susceptible to cybersecurity issues (“I get worried because things get lost, all records can go missing”; “I worry about being hacked”). Such concerns are not unsubstantiated: multiple participants shared personal experiences of losing money and falling victim to internet scams or struggling to identify which web-based activities were scams and which were safe (“I got scammed last Friday...they know more about me than I do”). In this context, technology designers and policy makers were framed as indifferent to issues of cybersecurity and safety in the name of profit (“There’s obviously a flaw, should have been to policy makers before they released that”; “Designers are here to make money; they don’t care for you”).

Issues of Ableism and Accessibility

Participants emphasized the need to address ableism. They often reported that the person they care for needs help with technology or with daily tasks in general. Most often, carers themselves live with disabilities, making it difficult for them to use technology, and they emphasized the need for more accessible devices (“Access is important...we need accessible things because it is good stuff, but what’s the point if it’s not accessible?”). Participants shared their struggles with concentration issues, needing multiple screens, and facing memory or vision problems when accessing web-based tasks (“You are given something you forget how to do, and you feel embarrassed”). These aspects are exacerbated by the different conditions older adults often experience. For example, a participant with hearing loss reported experiencing diminished communication abilities and a consequent fear of being perceived as less intelligent than in the past. This can have a profound effect, as when alternative support is available, it is typically via customer support–automated call centers that one must navigate before connecting to a human agent.

Crucially, and throughout the focus groups, carers constructed digitalization as ideologically hegemonic and expressed concerns about its abruptness. They resisted this technological prevalence by questioning the societal assumption that everyone has knowledge of digital technologies and that everyone should be actively using the internet (eg, “I felt very guilty and I still do” for lacking technical knowledge).

The Great Digital Disconnect

Overview

The “great digital disconnect” refers to the failure of digital services and processes to address digital poverty, a situation where the individual is unable to interact with the digital world in full when, where, and how they want or need to [14]. As participants were older adults, most of whom are state pension recipients, digital poverty issues pose additional financial strain. For them, this lack of support is counterintuitive because supporting unpaid carers could ultimately benefit the state. This in turn suggests that they feel the need to construct their labor within a neoliberal discursive framework to legitimize their need for financial support.

Lack of Connectivity

Concerns surfaced regarding the dearth of connectivity, resulting in extensive time spent on paperwork due to persistent connectivity issues. The shift from traditional landlines to digital platforms further heightened participants’ apprehensions, more pronounced among participants from the Southern and Midland regions of England and particularly those residing in rural settings.

The repercussions of inadequate connectivity impacted the material circumstances of the participants, with some resorting to significant expenditures for connectivity via their mobile phones. Carers depicted the lack of connectivity as frustrating (“I tried that 4-5 times, and I rang because it wasn’t working...I

XSL•FO
RenderX

gave up on it"), exacerbating financial or other care-related challenges, such as benefit or grant applications disappearing without an option to save progress ("it disappears, and you have to start all over again"). This deficiency in connectivity was portrayed as compounding the shortage of time ("What should take you one hour takes you all day"), a perennial issue faced by carers, and was construed as mentally taxing, adding to the "invisible" labor already shouldered by carers.

Throughout our discussions, participants consistently raised concerns regarding the impending transition to an all-digital landscape. The prevailing sentiment was one of unease concerning the growing dependence on digital communications, underpinned by scenarios involving power outages, prompting questions about the reliability of digital systems during electricity disruptions. The consequences of increased technological reliance are even more threatening, whereby a lack of connectivity may result in irreversible failures (eg, for individuals dependent on medical devices).

In many cases, participants indicated that internet providers have increased responsibility in addressing some of their connectivity issues. As explained earlier, slow internet speeds impede access to essential services, which can be even more significant for those benefiting from social tariffs, as connections are described as notably "sluggish."

Cost of Connectivity

Participants described navigating the financial aspects of technology and internet-related expenses as a "nightmare." The cost of broadband services was labeled as a "cutthroat business," with participants expressing a lack of awareness regarding variations in costs and services, pointing out the financial strain on their state pension. One participant noted that, with providers, the internet is bundled with other services, limiting their ability to shop around. They stressed the importance of balancing internet costs with desired services, stating, "You get to balance what you pay for the internet against what you want to have."

Technology was described as "fancy ideas," but participants emphasized the need for significant financial resources, stating that "there is a need for lots of money for crap." The discussion revolved around laptop prices and the increased cost of owning one. The cumulative costs of technology, such as electricity to power a laptop, were also mentioned, which is disadvantageous to those who may not fully understand these expenses.

Affordability concerns led some participants to limit internet use to their homes, avoiding data use outside. Pay-as-you-go options were preferred, with some participants lacking awareness of potential savings when purchasing a dedicated phone for internet use. A participant shared the financial strain caused by background apps "eating away" at phone credit, despite infrequent phone use. Carers highlighted the need for financial assistance and education on cost-effective technology use.

Participants found it challenging that the devices often needed to be up-to-date to perform certain tasks. For instance, an iPhone (Apple Inc) was mentioned as necessary for diabetes-related devices and apps, but it had to be the latest version to use the required app. However, carers suggested they were not concerned about having the latest technology; thus, such demands led them to spend money updating their phones. Moreover, participants highlighted that apps often charge for access, adding to the overall cost, and emphasized the need for funding for computers.

Some participants noted that social tariffs were available only to those on benefits, leaving others to cover all expenses themselves and potentially incur debt. The perceived inadequacy of social tariffs, with slow internet speeds not justifying the cost, was raised. Applying for Carers' Allowance requires internet access and a printer, creating barriers. Navigating available entitlements and benefits was described as a "nightmare." Carer Allowance can be claimed until pension age, creating additional financial challenges as pensions are considered income, yet carers suggested they need the allowance on top of that.

Mobilizing Neoliberal Ideals and Mirroring Neoliberal Discourses to Leverage Support

Participants brought up concerns regarding devices and overall funding, not just for training but also for health care and digital inclusion. They discursively resisted and challenged the ideologically hegemonic role of technology, emphasizing that not everyone possesses computers or phones nor can they afford broadband. Participants discussed that funders should be aware that relying solely on basic pension retention may only cover essential needs such as food, leaving individuals unable to afford other necessities. Therefore, they emphasized the importance of making technology accessible, suggesting that refurbished devices could be distributed through carers' trusts or voluntary organizations, saying, "What do you do about that...tablets, free SIM cards, other things." In addition, carers suggested allocating funds within organizations for community services specifically targeting digital exclusion, such as providing devices, data, and staffing, saying, "Give them money, give them laptops."

The need for funding was also discussed in the context of training. Participants proposed that training requires financial support and assistance to individuals requiring help with equipment and access. They used inclusive rhetoric via stressing the importance of viewing each person as an individual, stating that maintaining someone's well-being comes with significant costs. They often suggested that not only volunteers but also paid workers could play a role in providing technological training and support.

Against this background, it is of analytic interest that carers negotiate that they are worthy of funding and technological inclusivity by framing their value in market terms, often discussing the financial surplus they create through their labor and the social care expenditure the government avoids as a result of unpaid care ("Carers save millions for the government"; "we save the government millions of pounds...the stress we are under, anxiety"). In many ways, carers constructed themselves as the idealized neoliberal individual, not due to wanting to be one but to legitimize their need to be digitally included ("in the end, the more savvy we are, the more money we save"), perhaps pronouncing the ideological hegemony of neoliberal constructions. Indeed, a lack of funding can impact health care and, in turn, the overall economy ("funding has to be brought

XSL•FO
RenderX

somewhere otherwise everything will cost more because our health will be worse”).

The Unfulfilled Potential of Digitalization

Overview

This theme concerned the unfulfilled potential of digitalization. Participants in the study acknowledged the numerous benefits of technology yet frequently framed digitalization as a promise that remained unfulfilled, expressing disappointment in the unrealized potential of technology. Technological innovations were perceived as overlooking their own requirements and lacked lived experiences. Participants raised the necessity for greater agency in informing the design of technological solutions to ensure that the potential for technology to aid and facilitate materializes. They also expressed a desire for increased agency in shaping policies and technologies aimed at carers, emphasizing the importance of research informed by lived experiences and coproduction methods. This theme resulted in numerous subthemes, as outlined in subsequent sections.

Benefits of Technology and the Unfulfilled Potential of Digitalization

Despite numerous challenges, participants expressed a belief in the potential of technology to enhance their lives. Some adopted an optimistic perspective, encouraging others to persevere, asserting that embracing technology will eventually lead to a better quality of life for those who are willing. Drawing parallels to historical revolutions that simplified various aspects of life, participants pointed to the ongoing digitalization as a comparable transformation. They illustrated this point by referencing the impact of machines on domestic labor, providing an example of how technology (eg, washing machines) historically made life easier.

Participants highlighted the positive impact of technology in assisting them, that is, a “bonus of technology.” For instance, they mentioned services that identify hearing difficulties and use voice recognition technology. This included technologies such as cameras in the homes of older adults that allow families to monitor and ensure the well-being of their older members, offering a sense of security; videoconferencing tools (eg, Zoom [Zoom Video Communications]), which were described as instrumental in reducing loneliness, enabling individuals to engage in physical exercises, socialize, and attend events remotely; and other technologies used by doctors to communicate medical information before appointments, saving travel time through web-based appointments and receiving text reminders for medical appointments.

Yet, and despite these positive aspects, technologies and digitalization of services more broadly were framed as currently inaccessible with unfulfilled potential, creating a contrast between the envisioned benefits and the realities experienced by participants (“There is technology, I’ve seen it, people who need it don’t get it”).

Lack of Interoperability in Services, Institutions, and Within the NHS

A prevailing theme revolved around failings in the implementation of digital services and systems, spanning functionality and interoperability between services and departments, which leads to further complications when accessing essential services (“She could not see my account details”). On the one hand, this questions the effectiveness of urging technology use while facing ineffective implementation (“if the surgery can’t do it, what hope do you have”; “The GP/health services are telling us to use the services, but they don’t use them themselves. They send text messages we cannot even reply to and ask us to go on a website, but we are missing out on services”). On the other hand, it draws attention to the very design of these services, which are perceived as non–user-friendly, challenging to navigate, characterized by complex interfaces, and lacking in direct communication options such as telephone numbers and one-to-one interactions (“DWP [Department for Work and Pensions], this is actually serious and it’s threatening, and I cannot even call them out...when they call their number comes up as unknown and you have to wait for weeks and weeks”). Relatedly, text messages sent by the GP or the NHS system lacking reply options and instructions to visit websites posed challenges, especially for older individuals.

This lack of interoperability was often intertwined with intricate and dysfunctional bureaucratic processes within institutions. Participants were critical of the inflexibility of the systems in place, noting that deviating from what is perceived as the “right answer” could lead to requests being ignored or deemed invalid. This was particularly challenging for individuals with dyslexia, as the questions posed were difficult, and selecting the correct response became a hurdle. In addition, participants shared accounts of malfunctions and the necessity for prolonged phone calls, sometimes lasting up to an hour, especially in cases where individuals lacked smartphones (“And if you call you end up on the phone for hours because you don’t have a smartphone”).

Interestingly, we observed a relationship between poor functionality and lack of interoperability, whereby they can exacerbate each other, and preexisting difficulties, together. For example, NHS digital systems were described as “difficult” and “surreal.” One participant commented, “How many people really do this?” and that “Every new app makes my brain shudder.” They voiced confusion over accessing health care via various methods, leading to uncertainty (“I don’t know whether I am using the right thing”). In many cases, this results in challenges when individuals need to navigate multiple web pages or hop between apps, especially when the individual has multiple responsibilities (eg, caring for a child and a parent). Mostly, these issues lead to significant delays (“The nurse can’t help me”; “Two weeks from chemist to doctor, online 1-to-1 consultancy and still cannot get my medication”), but they also exacerbate carers’ lack of agency and highlight power imbalances. For example, for benefits claimants, participants expressed wariness and suggested that the forms and processes were designed to be confusing, so individuals in need of financial support would not be able to apply to claim it (“you can easily think that doesn’t apply to me”).

Research Coproduction for Reclaiming Agency

Throughout conversations, participants highlighted the need for robust carer representation, specifically, including individuals

XSL•FO
RenderX

from underrepresented backgrounds in the creation of policy more broadly and digital "solutions" more specifically. They specifically discussed that policy making needs to be based on engagement and collaboration with diverse carers, in ways that acknowledge their lived experience of caring for and accessing care. Volunteers at carers' forums and parent carers were deemed essential, and participants emphasized the need for continuous listening to carers at every stage of policy making. They expressed frustration with policies that have been drafted for them but without them (ie, no consultation beforehand) because this has led to initiatives that are not helpful nor needed ("it's like going to the hotel...it says it's disabled friendly and it's absolutely not").

The same principles underpinned conversations relating to technology design and designers and underlined the importance of co-designing with prospective users of technology rather than merely for them. Interestingly, these elements were considered in terms of designing and testing, as well as advertising, so that features and promises are aligned with reality and factual needs.

In short, collaboration with carers was constructed as essential for effective interventions, and policy makers were urged to invest in research on digital inclusion, following a holistic approach that values the input and experiences of carers in shaping policies and technology solutions. In doing so, one interesting finding was that participants highlighted that more coproduced qualitative research is needed, as this allows them to reclaim their agency and communicate their needs clearly, and this was contrasted to surveys, which are often closed-ended, leaving little room to properly engage with lived experiences.

Discussion

Principal Findings

In this study, we focused on the implications of the digitalization of social care on unpaid carers. Our aim was to explicate the perceptions and the lived experience of unpaid older carers, in particular, concerning digitalization, with the view to unpack and theorize around the implications of accelerated digitalization within a highly sensitive domain.

Our findings indicate that individuals who provide or receive care currently experience being "pushed" on the web, whereby they feel as if they are being coerced to access care services digitally, with limited alternative options. Their experiences are far removed from the ideal presented in the Department for Health and Social Care and the NHS's "A Plan for Digital Health and Social Care" [12]:

> *Sarah can choose a telephone, video, or in-person meeting for her first IAPT appointment, during which she will be assessed. During her video assessment, Sarah chooses a clinician-guided digital therapy app that she can log onto in her own time, day or night, and is easy to use as it is interactive. She can also choose to have her sessions in her preferred language via chat, audio or video.* [12]

The pressure to use web-based services to access support is met with not only a barrier in terms of the will or motivation of unpaid carers to do so but also a lack of meaningful support to enable this in practice. Such support would take the form of digital skills training, including privacy and safety, as well as service design that is accessible and free of ableist assumptions. Currently, however, our insights suggest that such support is considerably lacking, which results in what we term the "great digital disconnect": a situation whereby the digitalization interventions and programs seem to ignore, by omission or commission, those exposed to digital poverty and who are most likely to need social care support, that is, the disconnect between goals and actuality. However, this "disconnect" in service design is not an entirely new concept. For example, Stirling and Burgess [29], when examining smart telecare provision, uncovered disparities between intended goals and actual outcomes due to challenges arising from organizational and operational complexities. We extend this by arguing that such disparities may also exist due to ignorance of or indifference toward the contextual conditions of those who engage with the social care system. It is noteworthy that the described constructions and participants' experiences, characterized by significant waiting times and dehumanizing conditions, resonate with Weber's [30] notion of bureaucracy as an "iron cage" where individuals are subjected to challenging and counterproductive rules, leading to dehumanization and dysfunction within organizational structures, ultimately disregarding the needs of those seeking assistance.

Our findings indicate that older, unpaid carers wish to resist prevalent assumptions of universal access very clearly that portray everyone, irrespective of circumstances and choices, as being willing and able to access web-based services. This is congruent with earlier research. Choudrie et al [31], for example, found that older adults, in many cases, are willing and able to join the digital world, yet they do not do so because the available services do not offer any added value in their everyday lives. Similarly, Wyatt [32] criticized such universal assumptions regarding access and usefulness and went further on to highlight that nonusers' situations and needs should also be factored in when thinking about technology design.

Our findings echo the aforementioned findings but further explain that such assumptions and the accelerated replacement of physical solutions with digital ones lead to a power imbalance between the state and beneficiaries, whereby unpaid carers are left with very few choices and little or no alternatives. This results in pseudoconsent and coercion and reinforced exclusions, where individuals experience powerlessness in sharing their data. Lutz et al [33] described this as privacy cynicism, and indeed, data capitalism involves the extraction of personal data for extensive monetization, presenting users—in this case unpaid carers—with dilemmas between privacy and social connection or accessing services and benefits. Therefore, the digitalization of social care becomes ideologically dilemmatic [34] where older carers provide superficial consent, leading to passivity not only due to full acceptance of dominant ideologies but also due to a lack of effective dissent expression [28]. Maglaras [35] noted limited resistance possibilities due to the dominance of the ruling class's ideology. In this respect, we argue that unpaid carers' experiences illustrate the increasing dominance of digitalization and the ways imposed digitalization links to what

XSL•FO
RenderX

is called "surveillance realism" [36], whereby unpaid carers simply resign or give up, rather than truly consent in terms of parting with their data. However, there is still the possibility of recognizing the potential of alternative solutions—although rendered practically impossible—suggesting the digitalization of care has not yet reached a common sense status [37] or developed complete "ideological power" [38].

Against this background, we wish to highlight the implications of surveillance realism, unpaid carers' resignations, and the neoliberalization of social care. Participants criticized the paradoxical relationship between the lack of services and social care support and the efficiencies and savings unpaid caring produces for the state. On the one hand, they challenged the notion of being solely responsible for acquiring equipment and broadband connections, which is reminiscent of the neoliberal ideals of individual responsibility [28,39], which see the individual solely responsible for their welfare. Participants highlighted the inadequacy of governmental benefits–related funding, echoing societal norms within neoliberalism that valorize volunteering and unpaid labor while deeming social care too costly for the market. On the other hand, however, power imbalances and dominant market logics lead unpaid carers to resign, adopt the same logic, and position themselves as idealized neoliberal individuals, framing their caregiving labor as a form of capital production that benefits the government. To an extent, this reflects the erosion of the welfare state under neoliberalism, where volunteering and unpaid labor become normalized to facilitate private service delivery [40]. In this respect, our study highlights how, under surveillance realism and digital capitalism, unpaid carers adopt market-centric logic to make sense of their experience and as a form of negotiating their worth and greater access to social care.

In this context, we introduce alienation to theorize the digital exclusion experienced by carers. Alienation manifests as a disconnection from one's agency, resulting in a sense of estrangement from various aspects of life, including nature, labor, and social connections. Baines et al [41] discussed alienation of the social care worker under capitalistic constraints, but our study shows that unpaid carers, too, are constrained by a production system where they are treated as instruments rather than recognized as social beings, perpetuating a cycle of alienation from their labor and potential as human beings. In digital societies, production chains aim to shape individuals, further exacerbating this alienation [42].

Limitations

We note that our study comes with certain limitations. However, the implications of the digitalization of social care can vary depending on the sociodemographics of the unpaid carer, such as sexual identity, race, disability, and socioeconomic status [20]. As such, these implications need to be further examined through an intersectional lens to uncover a more nuanced understanding. With regard to the methods used, our participants commented on the use of the story completion material and the inclusive nature we adopted in preparing them, as they were able to "see" themselves in these stories and identify with the developed personas without necessarily sharing personal details openly. Participants valued our method as it allowed them to distance themselves from the points raised during discussions. Yet, they also expressed that they would have liked for the activities to be longer (each focus group lasted about 2 hours) as they felt they had more to share with the group.

Conclusions

Over recent years, in England and elsewhere, more social care services are becoming increasingly digital or moving to web-based provision, often framed as a means to increase the capacity of the social care system and create necessary efficiencies [10]. Digitalization is also aligned in policy discourse with offering people greater choice, control, power, and agency. While this is certainly possible, often, the digitalization of social care services risks excluding and disempowering those in need of support. Our findings and the concerns raised regarding the complexity of the technology and its exclusionary nature were often communicated with emotional language that conveyed a sense of abandonment. In addition, regarding the untapped potential of digitalization, while advantages were recognized, participants often felt let down as their needs were overlooked and stressed the importance of having a greater say in shaping technological solutions to ensure that these actually deliver on their potential benefits. Therefore, we highlight the importance of involving caregivers in the design of policies and technologies, emphasizing the need for research informed by lived experience and based on coproduction to allow for the meaningful integration of expertise and experience and for enabling person-centered care [43]. This can further enhance accessibility and usability by design but also requires challenging entrenched habits and assumptions [44].

Acknowledgments

This work was supported by the funded project "The impact of the digitalisation of social care on older informal carers" (2023, Higher Education Innovation Funding [HEIF] Fund, Research England) and the Economic and Social Research Council (ESRC) Centre for Care. The Centre for Care is funded by the ESRC (award ES/W002302/1), with contributions from the Department of Health and Social Care's National Institute for Health and Care Research (NIHR) and partner universities. The views expressed are those of the author or authors and are not necessarily those of the ESRC, UK Research and Innovation, National Health Service, or NIHR. The authors would like to extend their thanks and gratitude to the participants, Carers UK, and the Good Thing Foundation, for their support—without whom this research would not have been possible.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding structure.
[DOCX File , 20 KB - jmir_v26i1e60056_app1.docx]

References

1. Whitfield G, Hamblin K. Technology in social care: spotlight on the English policy landscape, 2019-2022. Centre for Care. URL: https://centreforcare.ac.uk/wp-content/uploads/2022/12/Technology-in-social-care-report-Dec-2022_FINAL.pdf [accessed 2024-04-29]
2. Razavi S. The political and social economy of care in a development context conceptual issues, research questions and policy options. United Nations Research Institute for Social Development. 2007. URL: https://cdn.unrisd.org/assets/library/papers/pdf-files/razavi-paper.pdf [accessed 2024-04-29]
3. Kodate N, Timonen V. Bringing the family in through the back door: the stealthy expansion of family care in Asian and European long-term care policy. J Cross Cult Gerontol 2017 Jun 9;32(3):291-301. [doi: 10.1007/S10823-017-9325-5]
4. Unpaid care by age, sex and deprivation, England and Wales: census 2021. Office for National Statistics. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/populationexploreragehealthandunpaidcare/2020-04-08 [accessed 2024-04-15]
5. Key facts and figures about caring. Carers UK. URL: https://www.carersuk.org/policy-and-research/key-facts-and-figures/ [accessed 2024-04-19]
6. Petrillo M, Bennett M. Valuing carers 2021 England and Wales. Carers UK. URL: https://www.carersuk.org/media/2d5le03c/valuing-carers-report.pdf [accessed 2023-05-25]
7. Pearson R. A feminist analysis of neoliberalism and austerity policies in the UK. Soundings 2019 Apr 01;71(71):28-39. [doi: 10.3898/soun.71.02.2019]
8. Čakardić A. Who cares: neoliberalism, informal labour, and life-making. Sociologija 2022;64(4):503-518. [doi: 10.2298/soc2204503c]
9. Hirnoven H, Tammelin M, Wouters EJ, Hanninen R. Introduction. In: Hirvonen H, Tammelin M, Hänninen R, Wouters EJ, editors. Digital Transformations in Care for Older People: Critical Perspectives. New York, NY: Routledge; 2022:3-14.
10. Hamblin KA. Technology in care systems: displacing, reshaping, reinstating or degrading roles? New Technol Work Employ 2022 Feb 06;37(1):41-58. [doi: 10.1111/ntwe.12229]
11. DHSC. Using the power of technology to make the world a safer and healthier place. Government of UK. URL: https://www.gov.uk/government/speeches/using-the-power-of-technology-to-make-the-world-a-safer-and-healthier-place [accessed 2024-04-19]
12. A plan for digital health and social care. Department of Health & Social Care, NHS England, Government of UK. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2024-04-19]
13. Fiske A, Buyx A, Prainsack B. The double-edged sword of digital self-care: physician perspectives from Northern Germany. Soc Sci Med 2020 Sep;260:113174 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113174] [Medline: 32659512]
14. Zamani ED, Vannini S. Understanding digital poverty in south Yorkshire - research note. The University of Sheffield. URL: https://www.sheffield.ac.uk/office-for-data-analytics/digital-poverty [accessed 2024-04-29]
15. Blackburn C, Read J, Hughes N. Carers and the digital divide: factors affecting internet use among carers in the UK. Health Soc Care Community 2005 May;13(3):201-210. [doi: 10.1111/j.1365-2524.2005.00547.x] [Medline: 15819741]
16. Distribution of internet users worldwide as of 2021, by age group. Statista. URL: https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/ [accessed 2024-03-20]
17. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. Inquiry 2022;59:469580221096272 [FREE Full text] [doi: 10.1177/00469580221096272] [Medline: 35471138]
18. Peytrignet S, Grimm F, Tallack C. Understanding unpaid carers and their access to support. The Health Foundation. URL: https://www.health.org.uk/publications/long-reads/understanding-unpaid-carers-and-their-access-to-support [accessed 2024-04-22]
19. New inquiry: the rights of older people. Women and Equalities Committee, UK Parliament. URL: https://committees.parliament.uk/work/7930/the-rights-of-older-people/news/197544/new-inquiry-the-rights-of-older-people/ [accessed 2024-04-22]
20. Savage B, Barringer MN. It matters who does the (unpaid) work: gender and the provision of informal caregiving by lesbian, gay, bisexual, and transgender (LGBT) aging adults. Sex Gend Divers Soc Serv 2023 Jul 31;36(1):105-132. [doi: 10.1080/10538720.2023.2241029]
21. Ziebland S, Hyde E, Powell J. Power, paradox and pessimism: on the unintended consequences of digital health technologies in primary care. Soc Sci Med 2021 Nov;289:114419. [doi: 10.1016/j.socscimed.2021.114419] [Medline: 34619631]

XSL•FO
RenderX

22. Baker AR, Billinge M. Geographies of England: The North-South Divide, Material and Imagined. Cambridge, UK: Cambridge University Press; 2004.
23. Smithson J. Using and analysing focus groups: limitations and possibilities. Int J Soc Res Methodol 2000 Jan;3(2):103-119. [doi: 10.1080/136455700405172]
24. Braun V, Clarke V, Hayfield N, Frith H, Malson H, Moller N, et al. Qualitative story completion: possibilities and potential pitfalls. Qual Res Psychol 2018 Nov 15;16(1):136-155. [doi: 10.1080/14780887.2018.1536395]
25. Charitsis V, Lehtiniemi T. Data ableism: ability expectations and marginalization in automated societies. Telev New Media 2022 Feb 22;24(1):3-18. [doi: 10.1177/15274764221077660]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Hayfield N, Wood M. Looking heteronormatively good! Combining story completion with Bitstrips to explore understandings of sexuality and appearance. Qual Res Psychol 2018 Nov 28;16(1):115-135. [doi: 10.1080/14780887.2018.1536390]
28. Gramsci A. Prison Notebooks. Volume 2. New York, NY: Columbia University Press; 2011.
29. Stirling P, Burgess G. The promises and the reality of smart, digital telecare in England. Hous Care Support 2021 Oct 03;24(2):65-79. [doi: 10.1108/hcs-03-2021-0010]
30. Weber M. Economy and Society: An Outline of Interpretive Sociology. Berkeley, CA: University of California Press; 1978.
31. Choudrie J, Zamani E, Obuekwe C. Bridging the digital divide in ethnic minority older adults: an organisational qualitative study. Inf Syst Front 2021 May 15;24(4):1355-1375. [doi: 10.1007/s10796-021-10126-8]
32. Wyatt S. Non-users also matter: the construction of users and non-users of the internet. In: Oudshoorn N, Pinch T, editors. How Users Matter: The Co-Construction of Users and Technology. Boston, MA: MIT Press; 2003:67-79.
33. Lutz C, Hoffmann CP, Ranzini G. Data capitalism and the user: an exploration of privacy cynicism in Germany. New Media Soc 2020 Jul 22;22(7):1168-1187. [doi: 10.1177/1461444820912544]
34. Billig M, Condor S, Edwards D, Gane M, Middleton D, Radley A. Ideological Dilemmas: A Social Psychology of Everyday Thinking. Thousand Oaks, CA: Sage Publications; 1988.
35. Maglaras V. Consent and submission: aspects of Gramsci's theory of the political and civil society. Sage Open 2013 Jan 24;3(1):215824401247234. [doi: 10.1177/2158244012472347]
36. Dencik L, Cable J. Digital citizenship and surveillance: the advent of surveillance realism: public opinion and activist responses to the Snowden leaks. Int J Commun 2017;11:19.
37. Fairclough N. The appropriacy of `appropriateness'. In: Fairclough N, editor. Critical Language Awareness. New York, NY: Routledge; 1992.
38. Lukes S. Power: a radical view. In: Scott J, editor. Power: Critical Concepts. New York, NY: Routledge; 1994:233-268.
39. Houghton E. Becoming a neoliberal subject. Ephemera Theory Politi Organ 2019;19(3):615-626 [FREE Full text]
40. Hawksley C, Georgeou N. Gramsci 'makes a difference': volunteering, neoliberal 'common sense', and the sustainable development goals. Third Sect Rev 2019;25(2):27-56 [FREE Full text] [doi: 10.3316/ielapa.929306512882392]
41. Baines D, Charlesworth S, Daly T. Underpaid, unpaid, unseen, unheard and unhappy? Care work in the context of constraint. J Ind Relat 2016 Aug 04;58(4):449-454. [doi: 10.1177/0022185616655981]
42. Nygren KG, Gidlund KL. The pastoral power of technology: rethinking alienation in digital culture. In: Fuchs C, Mosco V, editors. Marx in the Age of Digital Capitalism. New York, NY: BRILL; 2016:396-412.
43. Park (Ethan) S. Beyond patient-centred care: a conceptual framework of co-production mechanisms with vulnerable groups in health and social service settings. Public Manag Rev 2019 Jul 01;22(3):452-474. [doi: 10.1080/14719037.2019.1601241]
44. van Grunsven J, IJsselsteijn W. Confronting ableism in a post-COVID world: designing for world-familiarity through acts of defamiliarization. In: Dennis MJ, Ishmaev G, Umbrello S, van den Hoven J, editors. Values for a Post-Pandemic Future. Cham, Switzerland: Springer; 2022:185-200.

Abbreviations

GP: general practitioner
NHS: National Health Service
T&Cs: terms and conditions

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 30.04.24; peer-reviewed by X Vassilakopoulou, N Steils; comments to author 27.06.24; revised version received 11.07.24; accepted 24.09.24; published 24.10.24.

Please cite as:
Rousaki A, Zamani ED, Sbaffi L, Hamblin K, Black R
The Digitalization of Social Care in England and Implications for Older, Unpaid Carers: Constructionist Thematic Analysis
J Med Internet Res 2024;26:e60056
URL: https://www.jmir.org/2024/1/e60056
doi:10.2196/60056
PMID:

©Anastasia Rousaki, Efpraxia D Zamani, Laura Sbaffi, Kate Hamblin, Rachael Black. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Racial and Demographic Disparities in Susceptibility to Health Misinformation on Social Media: National Survey-Based Analysis

Ranganathan Chandrasekaran[1], PhD; Muhammed Sadiq T[2], MLiS; Evangelos Moustakas[3], PhD

[1]University of Illinois at Chicago, Chicago, IL, United States
[2]Indian Institute of Technology Madras, Chennai, India
[3]Canadian University Dubai, Dubai, United Arab Emirates

Corresponding Author:
Ranganathan Chandrasekaran, PhD
University of Illinois at Chicago
601 S Morgan Street
Chicago, IL, 60607
United States
Phone: 1 3129962847
Email: ranga@uic.edu

Abstract

Background: Social media platforms have transformed the dissemination of health information, allowing for rapid and widespread sharing of content. However, alongside valuable medical knowledge, these platforms have also become channels for the spread of health misinformation, including false claims and misleading advice, which can lead to significant public health risks. Susceptibility to health misinformation varies and is influenced by individuals' cultural, social, and personal backgrounds, further complicating efforts to combat its spread.

Objective: This study aimed to examine the extent to which individuals report encountering health-related misinformation on social media and to assess how racial, ethnic, and sociodemographic factors influence susceptibility to such misinformation.

Methods: Data from the Health Information National Trends Survey (HINTS; Cycle 6), conducted by the National Cancer Institute with 5041 US adults between March and November 2022, was used to explore associations between racial and sociodemographic factors (age, gender, race/ethnicity, annual household income, marital status, and location) and susceptibility variables, including encounters with misleading health information on social media, difficulty in assessing information truthfulness, discussions with health providers, and making health decisions based on such information.

Results: Over 35.61% (1740/4959) of respondents reported encountering "a lot" of misleading health information on social media, with an additional 45% (2256/4959) reporting seeing "some" amount of health misinformation. Racial disparities were evident in comparison with Whites, with non-Hispanic Black (odds ratio [OR] 0.45, 95% CI 0.33-0.6, $P<.01$) and Hispanic (OR 0.54, 95% CI 0.41-0.71, $P<.01$) individuals reporting lower odds of finding deceptive information, while Hispanic (OR 1.68, 95% CI 1.48-1.98, $P<.05$) and non-Hispanic Asian (OR 1.96, 95% CI 1.21-3.18, $P<.01$) individuals exhibited higher odds in having difficulties in assessing the veracity of health information found on social media. Hispanic and Asian individuals were more likely to discuss with providers and make health decisions based on social media information. Older adults aged ≥75 years exhibited challenges in assessing health information on social media (OR 0.63, 95% CI 0.43-0.93, $P<.01$), while younger adults (18-34) showed increased vulnerability to health misinformation. In addition, income levels were linked to higher exposure to health misinformation on social media: individuals with annual household incomes between US $50,000 and US $75,000 (OR 1.74, 95% CI 1.14-2.68, $P<.01$), and greater than US $75,000 (OR 1.78, 95% CI 1.20-2.66, $P<.01$) exhibited greater odds, revealing complexities in decision-making and information access.

Conclusions: This study highlights the pervasive presence of health misinformation on social media, revealing vulnerabilities across racial, age, and income groups, underscoring the need for tailored interventions.

(J Med Internet Res 2024;26:e55086) doi:10.2196/55086

KEYWORDS

health misinformation; digital divide; racial disparities; social media; national survey-based analysis; health information; interventions

XSL•FO
RenderX

Introduction

Background

In today's digitally connected world, social media has emerged as a fertile platform for the rapid dissemination of health-related content. Platforms such as Facebook [1], Instagram [2], YouTube [3], and Twitter [4] offer users the ability to share personal health experiences, disseminate medical news, and engage in online communities centered on health and wellness. This quick and widespread distribution of health information has the potential to elevate public health discourse, fostering awareness and encouraging dialogue, as was witnessed during the COVID-19 pandemic [5]. However, alongside the abundance of genuine and reliable health information, there is also a growing concern regarding the proliferation of misinformation on social media platforms [6,7]. Health-related misinformation, including rumors, pseudoscience, and personal anecdotes, can rapidly spread, often influencing unsuspecting users [8]. This misinformation can range from misleading claims about miracle cures to false opinions about vaccines, leading individuals to potentially harmful or dangerous practices [9,10].

Susceptibility to misinformation refers to an individual's tendency or vulnerability to believe, accept, or be influenced by false or misleading information, often despite evidence to the contrary. This susceptibility can stem from various factors, including cognitive biases, lack of critical thinking skills, emotional responses, social influence, and the persuasive techniques used by purveyors of misinformation [11]. Extant research suggests that certain segments of the population, such as individuals with lower health literacy, higher educational attainment, or those seeking alternative therapies [12], may be particularly susceptible to health-related misinformation on social media [13-16]. Given the far-reaching impact of social media on individuals' perceptions and behaviors, understanding the vulnerabilities that make certain individuals more prone to accepting and sharing health misinformation is crucial. It becomes important to identify the most vulnerable groups and target specific interventions to combat health misinformation [17].

Current studies on health misinformation on social media have documented the diffusion patterns [18,19], nature, and type of health misinformation and have suggested negative outcomes from such misinformation spread [17,20,21], including distorted interpretations of medical evidence, negative impacts on mental well-being, misallocation of health care resources, and a rise in vaccine hesitancy [22]. Furthermore, studies have also proposed advanced machine learning and artificial intelligence–based models that could detect health misinformation on social media [23-25]. However, limited research has delved into understanding the characteristics of individuals who are prone to accepting and using health misinformation [26]. Our study addresses this gap by documenting the nature and type of individuals who are more susceptible to health misinformation on social media.

Our research aims to explore 2 key questions. First, we seek to understand what kinds of individuals are more prone to accepting and believing health misinformation on social media. We specifically explore whether there are racial differences and other disparities in the susceptibility of individuals to health fake news content on social media. Recognizing unequal vulnerabilities and disparities in health information exposure and decision-making is essential for ensuring equitable access to accurate health information and promoting health equity.

Second, we intend to investigate how individuals react to health misinformation encountered on social media. We assess if they include such information in their discussions with providers or how such information affects their decisions pertaining to their health. Understanding the cognitive and emotional responses individuals have when presented with false or misleading health information is essential to comprehending the potential impact on their decision-making processes, health behaviors, and overall well-being.

Theoretical Underpinnings

Selective and superficial information processing represent 2 prominent theoretical frames within the realm of cognitive psychology, offering insights into how individuals perceive, evaluate, and act upon information encountered in various contexts. These 2 theoretical frames provide valuable perspectives for understanding susceptibility to misinformation, particularly in the context of health information dissemination on social media platforms.

Selective information processing posits that individuals tend to selectively attend to and interpret information in a manner that is congruent with their existing beliefs, attitudes, and motivations [27,28]. This selective attention and interpretation of information allow individuals to maintain consistency within their cognitive frameworks and protect their preexisting beliefs from cognitive dissonance [29,30]. Selective information processing is driven by various factors, including cognitive biases, emotional responses, social influences, and motivational goals [31]. Individuals engaging in selective information processing may exhibit confirmation bias, wherein they prioritize information that confirms their preconceptions while discounting or ignoring contradictory evidence [32,33].

Contrary to selective processing, superficial information processing involves the use of cognitive shortcuts or heuristics to evaluate information, often leading to quick and intuitive judgments rather than using thorough analysis [34,35]. This mode of information processing is characterized by minimal cognitive effort and reliance on mental shortcuts, such as heuristics, to make judgments or decisions [36]. Superficial processing is influenced by factors such as cognitive load, time constraints, and cognitive fluency [37,38]. Individuals engaging in superficial information processing may be susceptible to misinformation due to their tendency to overlook nuances, inconsistencies, or inaccuracies in the information presented [11].

The concepts of selective and superficial information processing guided us to choose specific variables for assessing individuals' susceptibility to health misinformation on social media, as explained in the next section.

XSL•FO
RenderX

Methods

This study relies on data sourced from the Health Information National Trends Survey (HINTS)— a comprehensive national survey conducted by the National Cancer Institute. HINTS is designed to provide nationally representative data, offering insights into the health-related needs, social media access and use, perceptions, and knowledge of adults aged ≥18 years in the United States [39]. HINTS 6, the latest iteration of this survey, includes a comprehensive set of questions pertaining to social media use, health misinformation on social media, and demographic factors. HINTS 6 follows a biennial data collection schedule, using both paper and web-based survey modes. The data collection for HINTS 6 commenced on March 7, 2022, and concluded on November 8, 2022. The final data set of HINTS 6 encompasses responses from 6252 participants, rendering it a robust resource for our research endeavors. Control totals in HINTS 6 underwent adjustments by drawing insights from the 2021 American Community Survey (ACS) carried out by the US Census Bureau. These adjustments factored in a range of variables, including age, gender, educational attainment, marital status, as well as racial and ethnic backgrounds. This was instrumental in harmonizing the demographic attributes within the HINTS 6 data set to closely mirror the population characteristics identified by the Census Bureau so that population-level estimates can be computed.

In the HINTS survey, weighted frequencies are calculated by applying survey weights to each respondent's data. These weights adjust for differences in the probability of selection and nonresponse rates across demographic groups, ensuring that the survey results are representative of the target population. The weights are derived from statistical procedures that consider sampling design, survey nonresponse, and demographic distributions from external sources such as census data. Applying these weights helps provide accurate estimates of population parameters in the survey results. Additional information is available in the HINTS methodology report [40].

Variables

We used a set of 4 questions from the HINTS 6 survey to gauge individuals' susceptibility to health misinformation on social media platforms. First, respondents were asked to assess the credibility of health information encountered on social media by rating the question, "How much of the health information that you see on social media do you think is false or misleading?" with response options ranging from 1 (none) to 4 (a lot). Second, respondents were asked to rate the level of difficulty they encountered in discerning the truthfulness of health information on social media using the question, "I find it hard to tell whether health information on social media is true or false," with response options ranging from "strongly disagree" to "strongly agree." Third, we sought to understand participants' agreement or disagreement with statements related to their use of health information on social media through 2 questions: one inquiring whether they used social media data for making personal health-related decisions ("I use information from social media to make decisions about my health"), another regarding the use of social media information in their discussions with health care providers ("I use information from social media in discussions with my healthcare provider"). Respondents indicated their agreement with these statements using a scale ranging from "strongly disagree" to "strongly agree."

The 4 facets used in this study–perception of the frequency of false health information on social media, difficulty in discerning the truthfulness of health information, use of social media data for personal health decisions, and use of social media information in discussions with health care providers–can be linked to the theoretical concepts of selective and superficial information processing.

Perception of false health information on social media and use of social media data for personal health decisions reflect individuals' engagement in selective information processing. Individuals who perceive a significant portion of health information on social media as false or misleading and rely heavily on social media data for personal health decisions are likely to selectively attend to information that confirms their preexisting beliefs or biases. Similarly, the use of social media information in discussions with health care providers may also be driven by selective processing, wherein individuals prioritize information that aligns with their beliefs or preferences while disregarding contradictory evidence.

On the other hand, difficulty in discerning the truthfulness of health information may indicate engagement in superficial information processing. Individuals experiencing great difficulty in evaluating the truthfulness of health information on social media may resort to cognitive shortcuts or heuristics, leading to superficial judgments that overlook the accuracy or validity of the information presented.

To assess variations in susceptibility to health misinformation on social media due to sociodemographic disparities, we chose a set of variables for our analysis. These factors encompassed age, which was categorized into 5 groups: 18-34, 35-49, 50-64, 65-74, and ≥75 years. We also considered birth gender (male or female) and race/ethnicity, which we classified as non-Hispanic White, non-Hispanic Black/African American, Hispanic, non-Hispanic Asian, or other. Household income was divided into 5 categories: <US $20,000, US $20,000-<US $35,000, US $35,000-<US $50,000, US $50,000-<US $75,000, and ≥US $75,000. Education levels were grouped as less than high school, high school graduate, some college, or college graduate or more. Marital status was categorized as married or otherwise. In addition, we classified the nature of residence using the National Center for Health Statistics (NCHS) 2013 Rural-Urban Commuting Area (RUCA) system, which assigns RUCA codes to specific census tracts based on factors such as population density, urbanization, and commuting patterns. These RUCA codes encompassed various residential categories, including large metro, large fringe metro, medium metro, small metro, nonmetropolitan micropolitan, and nonmetropolitan noncore, as developed by the NCHS to assess the level of rurality in residential areas.

Analysis

We conducted survey-weighted descriptive analyses to explore the socioeconomic demographics of HINTS 6 respondents, their

XSL•FO
RenderX

susceptibility to social media misinformation, and their responses to health information on social media. This analysis presented the frequencies (n), weighted estimates of the sample, and weighted percentages (%) of various variables. To account for the complex survey design of HINTS, we incorporated jackknife replication weights to adjust for standard errors. In addition, we used ordinal logistic regression analysis to investigate the relationship between predictor variables and respondents' susceptibility to health misinformation on social media. We tested 4 models that examined the associations between 4 variables pertaining to susceptibility and the sociodemographic predictors. As shown in Table 3, Model 1 assessed the extent of exposure to false or misleading health information on social media. Model 2 explored whether respondents used social media information in their health decision-making process. Model 3 examined whether individuals discussed information from social media with health care providers. Model 4 investigated the difficulty individuals faced in assessing health information on social media. The socioeconomic and demographic variables were included as predictors in these models. Respondents with missing values for any of the items were excluded from our analysis. All statistical analyses were carried out using STATA software (version 17.0; StataCorp LLC).

Ethical Considerations

HINTS data is collected by the National Cancer Institute and is available to scholars for research purposes. No separate approvals were sought for the use of this secondary data as all the data, survey instruments, and methodology reports are available on the HINTS website. Participation in the survey was voluntary. Participants were offered incentives ranging from US $2 in the initial mailing to US $10 or US $30 (for specific groups) to improve response rates. Additional details about the informed consent process, incentives, and methodology are available on the HINTS website [40].

Results

Sample Demographics

Our final analysis included 5041 adults (out of the 5922 survey respondents) who indicated using social media. Table 1 presents frequencies, weighted frequencies, and weighted percentages of the demographic characteristics of the study population. The respondents were 51.87% (n=2910) female, 59.77% (n=2553) non-Hispanic White, 11.3% (n=720) non-Hispanic Black, 17.45% (n=835) Hispanic, 5.94% (n=246) non-Hispanic Asian, and 5.54% (n=164) other non-Hispanic adults. Age distribution revealed that 29.62% (n=919) were between 18 and 34 years old, 27.57% (n=1168) were in the 35-49 age bracket, 26.49% (n=1469) were aged 50-64, 10.69% (n=966) were between 65 and 75 years old, and 5.62% (n=464) were 75 years old or older. In addition, 56% (n=2518) were either married or cohabiting. In terms of education, 34.61% (n=2116) held a bachelor's or postbaccalaureate degree, while 38.98% (n=1366) had completed "some college." In addition, 20.28% (n=755) had graduated from high school, and 6.12% (n=249) had education levels below high school. Household income distribution indicated that over 46% (n=1873) reported an annual income above $75,000, 18% (n=776) fell within the $50,000 to $75,000 range, 11.38% (n=577) had incomes between $35,000 and $50,000, 10.87% (n=569) earned between $20,000 and $35,000, while 13.63% (n=688) reported incomes less than $20,000. In addition, 31.88% (n=1962) resided in a large metropolitan area, while 46.34% (n=2103) were from either large fringe areas or medium metros. A smaller percentage, 9.85% (n=352), came from small metros, and 11.93% (n=624) hailed from nonmetro, micropolitan, or noncore regions (Table 1).

XSL•FO
RenderX

Table 1. Study population demographics.

	Sample, n	Weighted count	Percentage, %
Birth gender			
Male	1771	98790072.75	48.13
Female	2910	106475217.3	51.87
Marital status			
Not married	2159	90332590.72	44.04
Married or cohabitating	2518	114794809.3	55.96
Age group (years)			
18-34	919	64862315.77	29.62
35-49	1168	60382516.95	27.57
50-64	1469	58018272.31	26.49
65-74	966	23416142.89	10.69
75+	464	12300699.69	5.62
Education			
Less than high school	249	12492492.71	6.12
High school graduate	755	41366798.43	20.28
Some college	1366	79521119.41	38.98
Bachelor's degree	1372	40856012.93	20.03
Postbaccalaureate degree	944	29746253.58	14.58
Race\ethnicity			
Non-Hispanic White	2553	119961238.7	59.77
Non-Hispanic Black	720	22673290.6	11.30
Hispanic	835	35015805.29	17.45
Non-Hispanic Asian	246	11931031.66	5.94
Non-Hispanic other	164	11122270.69	5.54
Household income			
<$20,000	688	26819275.74	13.63
$20,000 to < $35,000	569	21398873.45	10.87
$35,000 to < $50,000	577	22396522.45	11.38
$50,000 to < $75,000	776	35433041.84	18.01
$75,000 or more	1873	90747802.43	46.11
Location			
Metropolitan: large metro	1962	70436347.45	31.88
Metropolitan: large fringe metro	1108	54876171.34	24.84
Metropolitan: medium metro	995	47496064.98	21.50
Metropolitan: small metro	352	21765167.73	9.85
Non-Metropolitan: Micropolitan	369	15577808.38	7.05
Non-Metropolitan: Noncore	255	10792811.39	4.88

Susceptibility to Health Misinformation in Social Media

About 35.61% (1740/4959) of respondents reported finding a substantial amount of health information on social media to be misleading or false. Meanwhile, 45.77% (2256/4959) felt that only some of this information was misleading, while a smaller portion, 17.09% (855/4959), considered 'a little' information misleading. Surprisingly, only 1.54% (108/4959) believed that none of the health-related information on social media was misleading. In evaluating health information found on social media, a significant majority, 66.56% (3225/4928), somewhat

XSL•FO
RenderX

or strongly agreed that they encounter difficulty in assessing health information on social media. However, 15.59% (818/4928) strongly disagreed, while 17.85% (885/4928) somewhat disagreed that they had difficulty in assessing the truthfulness of health information on social media (Table 2).

Among the respondents, 16.94% (811/4939) either somewhat or strongly affirmed their use of health information from social media in their personal health decisions. Contrastingly, 60.84% (3090/4939) strongly disagreed with this practice, while an additional 22.21% (1038/4939) somewhat disagreed. When asked about using health information from social media in discussions with their health care provider, 59.18% (2947/4935) strongly disagreed. 20.71% (977/4935) somewhat disagreed, while 20.11% (1011/4935) somewhat or strongly agreed, indicating that they discuss such information with their providers.

Table 2. Susceptibility to health misinformation on social media.

	Sample, n	Weighted count	Percentage, %
How much of the health information in social media you see is false or misleading?			
None	108	3342519.182	1.54
A little	855	37194084.68	17.09
Some	2256	99611868.69	45.77
A lot	1740	77499724.54	35.61
Difficulty in assessing health information on social media.			
Strongly disagree	818	33749756.19	15.59
Somewhat disagree	885	38662314.23	17.85
Somewhat or strongly agree	3225	144133135.4	66.56
Use information from social media to make decisions about health.			
Strongly disagree	3090	132073094.1	60.84
Somewhat disagree	1038	48218662.57	22.21
Somewhat or strongly agree	811	36,776,697	16.94
Use information from social media in discussions with health care provider.			
Strongly disagree	2947	128438034.5	59.18
Somewhat disagree	977	44951772.36	20.71
Somewhat or strongly agree	1011	43650679.74	20.11

Racial and Demographic Disparities

Table 3 displays the outcomes of 4 ordinal logistic regression models that investigated the relationships between susceptibility to social media health misinformation and various demographic variables, including age group, marital status, race and ethnicity, gender, education, annual household income, and location.

In investigating the prevalence of health misinformation on social media (Model 1), non-Hispanic Black (odds ratio [OR] 0.45, 95% CI 0.33-0.60) and Hispanic (OR 0.54, 95% CI 0.41-0.71) individuals, when compared to non-Hispanic White individuals, exhibited lower odds of encountering a substantial amount of health misinformation on social media. This suggests that these 2 racial groups may be more susceptible to health misinformation on social media channels. Individuals aged 75 years or older (OR 0.63, 95% CI 0.43-0.93), in comparison to those in the 18-34 age group, exhibited reduced odds of encountering a significant amount of health misinformation on social media. Furthermore, US adults with annual household incomes ranging from US $50,000 to US $75,000 (OR 1.74, 95% CI 1.14-2.68), as well as those earning more than $75,000 (OR 1.78, 95% CI 1.20-2.66), demonstrated notably higher odds of encountering a substantial volume of health misinformation through social media channels.

In our examination of the difficulties associated with assessing the accuracy of health information on social media (Model 4), we observed an increase in the odds with age. Particularly, individuals in the age groups of 65-74 (OR 1.66, 95% CI 1.11-2.51) and 75 and above (OR 1.86, 95% CI 1.10-3.14), when compared with the younger 18-34 group, displayed higher odds of finding it challenging to assess the truthfulness of health information obtained from social media channels. Furthermore, Hispanic (OR 1.68, 95% CI 1.48-1.98) and non-Hispanic Asian (OR 1.96, 95% CI 1.21-3.18) individuals, when contrasted with non-Hispanic White individuals, exhibit significantly higher odds of encountering challenges in assessing health information sourced from social media.

XSL•FO
RenderX

Table 3. Results from ordinal logistic regression analysis.

	Model 1: How much of the health information in social media you see is false or misleading?		Model 2: Use information from social media to make decisions about health.		Model 3: Use information from social media in discussions with health care provider.		Model 4: Difficulty in assessing health information on social media.	
	OR[a] (95 % CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Birth gender (Ref[b]: Male)								
Female	1.03 (0.82-1.28)	.82	0.95 (0.76-1.19)	.68	1.17 (0.94-1.46)	.15	1.09 (0.89-1.33)	.42
Marital status (Ref: Not married)								
Married or cohabitating	0.95 (0.75-1.18)	.62	1.00 (0.80-1.26)	.99	1.03 (0.81-1.32)	.80	0.93 (0.71-1.2)	.56
Age group (years; Ref: 18-34)								
35-49	0.91 (0.67-1.23)	.52	0.55 (0.41-0.77)	<.001	0.65 (0.51-0.85)	<.001	0.85 (0.62-1.15)	.29
50-64	0.98 (0.70-1.38)	.90	0.67 (0.50-0.91)	.01	0.89 (0.66-1.19)	.42	1.17 (0.87-1.59)	.29
65-74	1.06 (0.76-1.46)	.74	0.61 (0.44-0.85)	<.001	0.70 (0.55-0.89)	.01	1.66 (1.11-2.51)	.02
75+	0.63 (0.43-0.93)	.02	0.57 (0.40-0.81)	<.001	0.67 (0.44-1.03)	.07	1.86 (1.10-3.14)	.02
Education (Ref: less than high school)								
High school graduate	0.88 (0.53-1.46)	.61	0.77 (0.41-1.46)	.42	0.68 (0.36-1.31)	.25	1.22 (0.61-2.46)	.58
Some college	1.13 (0.68-1.87)	.63	0.90 (0.47-1.75)	.76	0.92 (0.5-1.7)	.79	1.26 (0.67-2.37)	.46
Bachelor's degree	0.86 (0.53-1.41)	.55	0.89 (0.48-1.63)	.69	1.01 (0.55-1.85)	.99	0.93 (0.47-1.85)	.83
Postbaccalaureate degree	1.18 (0.69-2.01)	.54	0.79 (0.43-1.44)	.43	1.02 (0.54-1.92)	.96	0.60 (0.31-1.19)	.14
Race\ethnicity (Ref: non-Hispanic White)								
Non-Hispanic Black	0.45 (0.33-0.6)	<.001	1.21 (0.90-1.62)	.20	1.01 (0.72-1.41)	.98	0.86 (0.65-1.14)	.29
Hispanic	0.54 (0.41-0.71)	<.001	1.64 (1.13-2.39)	.01	1.25 (0.97-1.62)	.08	1.68 (1.48-1.98)	.04
Non-Hispanic Asian	0.91 (0.55-1.5)	.71	2.19 (1.18-4.1)	.01	2.09 (1.15-3.82)	.02	1.96 (1.21-3.18)	.01
Non-Hispanic other	1.10 (0.54-2.26)	.79	0.51 (0.28-0.93)	.03	0.79 (0.45-1.4)	.42	0.68 (0.39-1.2)	.18
HHInc[c] (US $; Ref: <20,000)								
20,000 to <35,000	1.37 (0.86-2.17)	.18	0.65 (0.41-1.03)	.07	0.55 (0.34-0.89)	.02	0.95 (0.56-1.62)	.85
35,000 to <50,000	1.18 (0.80-1.75)	.39	0.54 (0.36-0.84)	.01	0.51 (0.32-0.79)	<.001	0.91 (0.57-1.45)	.68
50,000 to <75,000	1.74 (1.14-2.68)	.01	0.60 (0.36-1.02)	.06	0.57 (0.32-1.04)	.07	0.89 (0.59-1.36)	.60
75,000 or more	1.78 (1.20-2.66)	.01	0.62 (0.40-0.99)	.04	0.54 (0.32-0.89)	.02	1.18 (0.79-1.75)	.42
Location (Ref: Large metro)								
Large fringe metro	1.13 (0.86-1.49)	.39	1.05 (0.75-1.45)	.78	1.01 (0.72-1.43)	.94	1.08 (0.79-1.48)	.60
Medium metro	1.34 (0.99-1.81)	.06	1.11 (0.80-1.52)	.53	1.26 (0.93-1.69)	.13	1.02 (0.78-1.33)	.90
Small metro	1.34 (0.97-1.84)	.08	0.78 (0.46-1.34)	.37	0.91 (0.54-1.54)	.73	1.01 (0.69-1.48)	.97
Micropolitan	1.27 (0.87-1.85)	.21	1.11 (0.73-1.69)	.62	1.16 (0.74-1.81)	.50	1.37 (0.78-2.41)	.27
Noncore	0.94 (0.53-1.67)	.82	0.73 (0.42-1.27)	.26	0.95 (0.6 -1.52)	.83	0.83 (0.45-1.55)	.55

[a]OR: odds ratio.

[b]Ref: reference group.

[c]HHInc: annual household income.

In the analysis of health-related decision-making based on information from social media (Model 2), individuals within the age groups of 35-49 (OR 0.55, 95% CI 0.41-0.77), 50-64 (OR 0.67, 95% CI 0.50-0.91), 65-74 (OR 0.61, 95% CI 0.44-0.85), and 75 years and older (OR 0.57, 95% CI 0.40-0.81), in contrast to those in the 18- to 34-year age group, exhibited significantly diminished odds of basing health decisions on information derived from social media. This suggests that younger adults in the 18-34 age group are more likely to integrate social media–based health information into their decision-making processes regarding personal health matters. Our analysis also revealed significant associations between

XSL•FO
RenderX

race/ethnicity and health decision-making based on information from social media. When compared with non-Hispanic White individuals, Hispanic (OR 1.64, 95% CI 1.13-2.39) and non-Hispanic Asian (OR 2.19, 95% CI 1.18-4.10) individuals displayed significantly elevated odds of making health-related decisions based on information obtained from social media. Conversely, non-Hispanic individuals categorized as "others" (OR 0.51, 95% CI 0.28-0.93) exhibited lower odds in this regard. The likelihood of making health decisions based on information observed on social media exhibits a diminishing trend with higher income levels. As indicated in Table 3, the odds decrease as annual household income levels increase.

In regard to incorporating social media–based health information in discussions with health care providers (Model 3), our analysis reveals that non-Hispanic Asian (OR 2.09, 95% CI 1.15-3.82) and Hispanic (OR 1.25, 95% CI 0.97-1.62; *P*=.08) individuals exhibit higher propensities compared to non-Hispanic White individuals. In addition, the odds appear to decrease with both age and income levels. Gender, marital status, location, and educational levels did not emerge as significant in any of the 4 models that we examined.

Discussion

The pervasiveness of misleading or false health information on social media poses a significant threat to public health. In our study, a good proportion of respondents (35.61%, 1740/4959) reported encountering a lot of misleading or false health information on social media, underscoring the widespread presence of health misinformation on these platforms. We also found considerable racial disparities when examining susceptibility to health misinformation on social media. Interestingly, non-Hispanic Black and Hispanic individuals reported lower odds of finding deceptive information on social media. Previous studies have documented considerable circulation of health misinformation and susceptibility of Black communities [41,42], primarily due to the lack of access to health resources, low levels of health literacy, and barriers due to social determinants of health [43,44]. Susceptibility to health misinformation on social media for Black individuals may not vary much as compared with that of non-Hispanic White individuals, as we did not detect any significant differences in 3 of the 4 models tested. However, our findings also indicated that Hispanic and non-Hispanic Asian groups find it difficult to assess the truthfulness of health information in social media channels. In addition, Hispanic and Asian individuals are more likely to discuss social media–based health information with their provider and also make health-related decisions based on what they find in social media. Previous studies have found Hispanic individuals to face a heightened risk of adverse health outcomes, including obesity, diabetes, and HIV [45]. Previous research has recorded the vulnerability of Asian American individuals to health-related misinformation and its impact on their decision-making regarding health. During the COVID-19 pandemic, misinformation about the pandemic was disseminated among Asian American individuals, including Chinese, Korean, Vietnamese, and South Asian American individuals, and this influenced their vaccine decision-making [46]. Exposure to health misinformation, which suggests that smoking or alcohol consumption can protect against COVID-19, was associated with self-reported increases in tobacco and alcohol use among Chinese individuals [47]. Hispanic communities have also been found to be susceptible to misinformation regarding the safety of COVID-19 vaccines and the efficacy of "alternate treatments" [48]. Further, considerable differences exist in how Hispanic individuals and non-Hispanic White individuals use social media platforms [49]. Taken together, our findings indicate considerable vulnerabilities among Hispanic and non-Hispanic Asian groups, highlighting the need for targeted interventions to address these disparities. As racial groups are more prone to making health-related decisions based on information found on social media, health care providers should specifically engage with individuals from these communities to educate them on health misinformation on social media platforms.

Our findings indicate that individuals of different age groups exhibit varying levels of susceptibility to health misinformation on social media. Older adults aged 65 years and older often face challenges in assessing the credibility of health information encountered on social media. This can be attributed to several factors, including limited exposure to digital literacy education, unfamiliarity with social media platforms, and cognitive changes associated with aging. Furthermore, older adults are less likely to discuss social media–based health information with their health care providers. This may be due to concerns about appearing uninformed or a reluctance to trust information on social media. Consequently, health care providers may miss opportunities to address potential health concerns stemming from misinformation exposure. Our findings also note that older adults, as compared with younger ones, are less likely to make health-related decisions based on information from social media. On the other hand, younger adults aged between 18 and 34 seem more vulnerable and susceptible. The younger adults are likely to spend more time on social media channels and, hence, could get more exposure to misleading health information that could even factor into their health decision-making processes. Studies have shown that younger adults tend to miss physician visits [50], engage in fewer discussions and interactions with their health care providers [51], and are more highly influenced by external sources [52]. This combination of factors makes them particularly vulnerable to the influence of health misinformation on social media. If younger adults continue to rely heavily on social media for health information and fail to seek guidance from health care professionals, they may make uninformed decisions that could jeopardize their health and well-being. This highlights the urgent need for interventions that address the issue of health misinformation on social media and promote informed decision-making among younger adults.

Our study identified a notable connection between income levels and susceptibility to health misinformation on social media. Individuals with higher annual household incomes are prone to encountering a greater amount of misleading information on social media. Paradoxically, despite their exposure, they are less inclined to engage in discussions with health care providers or base critical health decisions on information from social media. Conversely, individuals with lower incomes may be more susceptible to health misinformation, highlighting potential disparities in information access and decision-making resources.

XSL•FO
RenderX

Our findings regarding positive associations between income levels and misinformation susceptibility are similar to a few studies [53,54] but differ from others who have recorded a negative association [55,56]. Some researchers have also suggested a curvilinear relationship between income levels and misinformation susceptibility [57]. Individuals with higher incomes may be more susceptible to medical misinformation on social media due to their greater access to technology and trust in authoritative sources without critical evaluation. Busy lifestyles and reliance on social networks could limit their ability to thoroughly assess information, while targeted marketing may exploit their purchasing power. Overconfidence in personal knowledge and peer influence further exacerbate susceptibility, highlighting the need for interventions promoting critical thinking and digital literacy across socioeconomic groups.

Limitations

This study has some important limitations. First, our focus was on racial and sociodemographic disparities that are likely to be associated with susceptibility to health misinformation. There could be several other factors, like health conditions, health literacy, and technological know-how, which could influence an individual's behavior toward social media–based health information. Second, data obtained from HINTS rely on self-reported responses that could be biased. Third, the cross-sectional nature of data collection relies on data at one specific point in time. Given the evolving nature of the social media landscape, individual behavior on social media could change with time, and additional data collection across different time periods could throw light on the longitudinal behavioral dynamics. Finally, we were also limited by the use of a predefined set of measures included in the HINTS survey. A self-developed set of measures that is based on theoretical frameworks from cognitive psychology could capture the multidimensional nature of susceptibility to misinformation.

Implications and Recommendations

Our research highlights the intricate interplay of sociodemographic and racial/ethnic factors in shaping individuals' interactions with health information on social media. To address the multifaceted challenges identified, targeted health literacy campaigns should be developed, particularly focusing on younger age groups, Hispanic individuals, and non-Hispanic Asian individuals. These campaigns should emphasize the cultivation of critical evaluation skills for assessing the authenticity of health information disseminated through social media platforms. In addition, there is a pressing need to acknowledge and address racial and ethnic disparities by tailoring health communication strategies to different demographic groups. This tailored approach can enhance the relevance and effectiveness of public health interventions.

Furthermore, our study emphasizes the importance of equipping health care providers with the necessary skills and knowledge to navigate discussions about social media–based health information with patients [58]. This not only fosters a collaborative and informed health care decision-making process but also positions health care providers as key advocates in increasing awareness about health misinformation on social media. They can provide specific guidance, especially to vulnerable groups, to navigate the influx of health information online. Health care providers should acknowledge that patients frequently seek medical guidance from the Internet, friends, peers, and family members. Discouraging patients entirely from using alternative sources is unlikely to succeed. To combat misinformation, providers should empower patients by providing accurate information that meets their needs for self-education [59,60].

Another tactic to handle health misinformation dissemination that has emerged has been proactive efforts by social media platforms to label inaccurate content or to lower the prominence of such content in their presentation algorithms. Many individuals have reported encountering interventions such as credibility labels that are assigned to social media posts [61]. Accuracy nudges involving subtle clues or prompts and post hoc corrections of misinformation have helped improve the discerning behavior of individuals toward health misinformation on social media [62,63]. However, evidence regarding whether such efforts affect misinformation spread and acceptance is mixed. Exposure to fact-checking labels has been found to discourage misinformation sharing [64] as well as encourage positive health decision-making behaviors [65], whereas evidence for effects such as reduced acceptance of inaccurate claims is also contested [66]. While the impact of fact-checking labels may be uncertain, community engagement approaches centered on building relationships over time show potential in establishing trusted communication networks. These networks can play a crucial role in combating the spread of health misinformation among specific racial groups and communities. For instance, approaches involving the use of social media influencers, volunteers, and celebrities to spread pro-vaccine messaging on social media platforms have proved to be effective in Hispanic communities [67]. Educational campaigns to combat health misinformation on social media have also been effective [68].

Finally, health care institutions and agencies have a crucial role to play in combating health misinformation. They can contribute by offering fact-check mechanisms, partnering with social media companies to contain the spread of misinformation, and implementing strategies to enhance the digital health literacy of the broader population. These collaborative efforts are vital in creating a robust defense against the unchecked dissemination of misleading health information on social media platforms.

Theoretical Implications

The theoretical implications of this study are significant, particularly in the context of information processing theories that deal with superficial processing and selective processing. The findings underscore the prevalence of health misinformation on social media platforms, implicating individuals' susceptibility to misleading information. Specifically, the observed racial disparities suggest variations in how different demographic groups engage with and evaluate health information on social media. For instance, non-Hispanic Black and non-Hispanic Asian individuals may demonstrate a tendency towards selective processing, wherein they are less likely to encounter deceptive information but may face challenges in assessing the truthfulness

XSL•FO
RenderX

of health content on social media. Conversely, Hispanic and non-Hispanic Asian individuals may exhibit tendencies toward superficial processing, as evidenced by their increased likelihood to discuss health information with providers and make health decisions based on social media content, potentially without thorough scrutiny.

Furthermore, the age-related patterns observed in the study align with theories of information processing, with older adults exhibiting challenges in assessing health information on social media, possibly due to cognitive factors or limited digital literacy. Conversely, younger adults demonstrate increased vulnerability to health misinformation, suggesting a propensity toward superficial processing or susceptibility to persuasive content on social media. In addition, the linkage between income levels and higher exposure to health misinformation suggests complexities in decision-making and information access, potentially influenced by factors such as digital literacy, access to reputable sources, and trust in information on social media.

Conclusion

In conclusion, addressing health misinformation on social media requires a nuanced and comprehensive approach that considers the diverse demographic factors influencing individuals' interactions with online health content. By understanding these complexities, public health initiatives can better tailor strategies to foster informed decision-making and mitigate the impact of misinformation on individual and community health.

Data Availability

Data used in this study is available on the Health Information National Trends Survey website.

Authors' Contributions

CR and EM contributed to the conceptualization. CR and MST contributed to the formal analysis. EM contributed to project administration. CR and EM contributed to writing-original draft. CR, EM, and MST contributed to writing-review and editing.

Conflicts of Interest

None declared.

References

1. Sharma M, Yadav K, Yadav N, Ferdinand KC. Zika virus pandemic-analysis of Facebook as a social media health information platform. Am J Infect Control 2017;45(3):301-302. [doi: 10.1016/j.ajic.2016.08.022] [Medline: 27776823]
2. Seltzer EK, Horst-Martz E, Lu M, Merchant RM. Public sentiment and discourse about Zika virus on Instagram. Public Health 2017;150:170-175. [doi: 10.1016/j.puhe.2017.07.015] [Medline: 28806618]
3. Osman W, Mohamed F, Elhassan M, Shoufan A. Is youTube a reliable source of health-related information? A systematic review. BMC Med Educ 2022 May;22(1):382 [FREE Full text] [doi: 10.1186/s12909-022-03446-z] [Medline: 35590410]
4. Hönings H, Knapp D, Nguy n BC, Richter D, Williams K, Dorsch I, et al. Health information diffusion on twitter: the content and design of WHO tweets matter. Health Info Libr J 2022;39(1):22-35. [doi: 10.1111/hir.12361] [Medline: 33682996]
5. Neely S, Eldredge C, Sanders R. Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: survey study. J Med Internet Res 2021;23(6):e29802 [FREE Full text] [doi: 10.2196/29802] [Medline: 34043526]
6. Chou WYS, Oh A, Klein WMP. Addressing Health-Related misinformation on social media. JAMA 2018;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
7. Melchior C, Oliveira M. Health-related fake news on social media platforms: a systematic literature review. New Media & Society 2021;24(6):1500-1522. [doi: 10.1177/14614448211038762]
8. Ferrara E, Cresci S, Luceri L. Misinformation, manipulation, and abuse on social media in the era of COVID-19. J Comput Soc Sci 2020;3(2):271-277 [FREE Full text] [doi: 10.1007/s42001-020-00094-5] [Medline: 33251373]
9. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of Health-related misinformation on social media. Soc Sci Med 2019;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
10. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
11. Pennycook G, Rand DG. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences 2019;116(7):2521-2526 [FREE Full text] [doi: 10.1073/pnas.1806781116] [Medline: 30692252]
12. Scherer LD, McPhetres J, Pennycook G, Kempe A, Allen LA, Knoepke CE, et al. Who is susceptible to online health misinformation? A test of four psychosocial hypotheses. Health Psychol 2021;40(4):274-284. [doi: 10.1037/hea0000978] [Medline: 33646806]

XSL•FO
RenderX

13. Oh HJ, Lee H. When do people verify and share health rumors on social media? the effects of message importance, health anxiety, and health literacy. J Health Commun 2019;24(11):837-847. [doi: 10.1080/10810730.2019.1677824] [Medline: 31609678]
14. Xue H, Taylor L. When do people believe, check, and share health rumors on social media? Effects of evidence type, health literacy, and health knowledge. J Health Psychol 2023;28(7):607-619. [doi: 10.1177/13591053221125992] [Medline: 36168751]
15. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
16. Nan X, Wang Y, Thier K. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. Social Science & Medicine 2022;314:115398 [FREE Full text] [doi: 10.1016/j.socscimed.2022.115398] [Medline: 36327631]
17. Sylvia Chou WY, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health 2020;110(S3):S273-S275. [doi: 10.2105/AJPH.2020.305905] [Medline: 33001722]
18. Daradkeh M. Analyzing sentiments and diffusion characteristics of COVID-19 vaccine misinformation topics in social media: a data analytics framework. International Journal of Business Analytics (IJBAN) 2022;9(3):22. [doi: 10.4018/ijban.292056]
19. Tao ZY, Chu G, McGrath C, Hua F, Leung YY, Yang WF, et al. Nature and diffusion of COVID-19-related oral health information on Chinese social media: analysis of tweets on weibo. J Med Internet Res 2020;22(6):e19981 [FREE Full text] [doi: 10.2196/19981] [Medline: 32501808]
20. Walter N, Brooks JJ, Saucier CJ, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a Meta-Analysis. Health Commun 2021;36(13):1776-1784. [doi: 10.1080/10410236.2020.1794553] [Medline: 32762260]
21. Bautista JR, Zhang Y, Gwizdka J. US Physicians' and Nurses' motivations, barriers, and recommendations for correcting health misinformation on social media: qualitative interview study. JMIR Public Health Surveill 2021;7(9):e27715 [FREE Full text] [doi: 10.2196/27715] [Medline: 34468331]
22. Borges do Nascimento IJ, Pizarro AB, Almeida JM, Azzopardi-Muscat N, Gonçalves MA, Björklund M, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022;100(9):544-561 [FREE Full text] [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
23. Mahara T, Josephine VLH, Srinivasan R, Prakash P, Algarni AD, Verma OP. Deep vs. shallow: a comparative study of machine learning and deep learning approaches for fake health news detection. IEEE Access 2023;11:79330-79340 [FREE Full text] [doi: 10.1109/access.2023.3298441]
24. Khan S, Hakak S, Deepa N, Prabadevi B, Dev K, Trelova S. Detecting COVID-19-Related fake news using feature extraction. Front Public Health 2021;9:788074 [FREE Full text] [doi: 10.3389/fpubh.2021.788074] [Medline: 35059379]
25. Di Sotto S, Viviani M. Health misinformation detection in the social web: an overview and a data science approach. Int J Environ Res Public Health 2022;19(4):2173 [FREE Full text] [doi: 10.3390/ijerph19042173] [Medline: 35206359]
26. Scherer LD, Pennycook G. Who is susceptible to online health Misinformation? Am J Public Health 2020;110(S3):S276-S277. [doi: 10.2105/AJPH.2020.305908] [Medline: 33001736]
27. Kunda Z. The case for motivated reasoning. Psychol Bull 1990;108(3):480-498. [doi: 10.1037/0033-2909.108.3.480] [Medline: 2270237]
28. Kahan DM. Climate-Science Communication and the Measurement Problem. Political Psychology 2015;36(S1):1-43 [FREE Full text] [doi: 10.1111/pops.12244]
29. Festinger L. A Theory of Cognitive Dissonance. Redwood City, California: Stanford University Press; 1962.
30. Morvan C, O-Connor A. An Analysis of Leon Festinger?s A Theory of Cognitive Dissonance. Boca Raton, Florida: CRC Press; 2017.
31. Lewandowsky S, Ecker UKH, Cook J. Beyond misinformation: understanding and coping with the “post-truth” era. Journal of Applied Research in Memory and Cognition 2017;6(4):353-369. [doi: 10.1016/j.jarmac.2017.07.008]
32. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Review of General Psychology 1998;2(2):175-220. [doi: 10.1037/1089-2680.2.2.175]
33. Pohl RF. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory. United Kingdom: Psychology Press; 2012.
34. Kahneman D, Slovic P, Tversky A. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press 1982;185(4157):1124-1131 [FREE Full text] [doi: 10.1017/cbo9780511809477]
35. Morvan C, Jenkins W. An analysis of amos tversky and daniel kahneman's judgment under uncertainty. In: Heuristics and Biases. Boca Raton, Florida: CRC Press; 2017.
36. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
37. Alter AL, Oppenheimer DM. Uniting the tribes of fluency to form a metacognitive nation. Pers Soc Psychol Rev 2009;13(3):219-235. [doi: 10.1177/1088868309341564] [Medline: 19638628]

XSL•FO
RenderX

38. Unkelbach C, Greifeneder R. The Experience of Thinking: How Feelings From Mental Processes Influence Cognition and Behaviour. Germany: Psychology Press; 2013.
39. Hesse BW, Greenberg AJ, Peterson EB, Chou WYS. The health information national trends survey (HINTS): a resource for consumer engagement and health communication research. Stud Health Technol Inform 2017;240:330-346. [Medline: 28972526]
40. Health Information National Trends Survey. HINTS 6 Methodology Report. 2023. URL: https://hints.cancer.gov/data/methodology-reports.aspx [accessed 2024-10-12]
41. Laurencin CT, Valantine H, Yancy C, Jones CP, Bright C. The COVID-19 vaccine and the black community: addressing the justified questions. J Racial Ethn Health Disparities 2021;8(4):809-820 [FREE Full text] [doi: 10.1007/s40615-021-01082-9] [Medline: 34143380]
42. Dodson K, Mason J, Smith R. Covid-19 vaccine misinformation and narratives surrounding Black communities on social media. Media Democracy Fund. 2021. URL: https://firstdraftnews.org/wp-content/uploads/2021/12/COVID-19_VACCINE_MISINFORMATION_BLACK_COMMUNITIES.pdf [accessed 2024-10-12]
43. Tucker-Seeley RD. Social determinants of health and disparities in cancer care for black people in the United States. JCO Oncol Pract 2021;17(5):261-263. [doi: 10.1200/OP.21.00229] [Medline: 33974819]
44. Egede LE, Walker RJ. Structural racism, social risk factors, and Covid-19 - a dangerous convergence for black Americans. N Engl J Med 2020;383(12):e77 [FREE Full text] [doi: 10.1056/NEJMp2023616] [Medline: 32706952]
45. Hudnut-Beumler J, Po'e E, Barkin S. The use of social media for health promotion in hispanic populations: a scoping systematic review. JMIR Public Health Surveill 2016;2(2):e32 [FREE Full text] [doi: 10.2196/publichealth.5579] [Medline: 27400979]
46. Chong SK, Ali SH, Đoàn LN, Yi SS, Trinh-Shevrin C, Kwon SC. Social media use and misinformation among Asian Americans during COVID-19. Front Public Health 2021;9:764681 [FREE Full text] [doi: 10.3389/fpubh.2021.764681] [Medline: 35096736]
47. Luk TT, Zhao S, Weng X, Wong JYH, Wu YS, Ho SY, et al. Exposure to health misinformation about COVID-19 and increased tobacco and alcohol use: a population-based survey in Hong Kong. Tob Control 2021;30(6):696-699. [doi: 10.1136/tobaccocontrol-2020-055960] [Medline: 32855353]
48. Longoria J, Acosta D, Urbani S, Smith R. A limiting lens: how vaccine misinformation has influenced hispanic conversations online. Regulation, Policy, & Platform Governance. 2021. URL: https://firstdraftnews.org/wp-content/uploads/2021/12/COVID-19_VACCINE_MISINFORMATION_HISPANIC_COMMUNITIES.pdf [accessed 2024-10-12]
49. Wang W, Segev S, Liu Y. Use of social networking sites for product communication: a comparative study of hispanics and non-Hispanic whites. Computers in Human Behavior 2015;49:86-93 [FREE Full text] [doi: 10.1016/j.chb.2015.02.051]
50. Rand CM, Goldstein NPN. Patterns of primary care physician visits for US adolescents in 2014: implications for vaccination. Acad Pediatr 2018;18(2S):S72-S78. [doi: 10.1016/j.acap.2018.01.002] [Medline: 29502641]
51. Santelli JS, Grilo SA, Klein JD, Liu Y, Yan H, Li B, et al. The unmet need for discussions between health care providers and adolescents and young adults. J Adolesc Health 2020;67(2):262-269. [doi: 10.1016/j.jadohealth.2020.01.019] [Medline: 32169522]
52. Viktorsson L, Törnvall E, Falk M, Wåhlin I, Yngman-Uhlin P. Young adults' healthcare utilisation and healthcare needs: perceptions and experiences of healthcare providers. Health Expect 2022;25(1):245-253 [FREE Full text] [doi: 10.1111/hex.13370] [Medline: 34624154]
53. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
54. Bridgman A, Merkley E, Loewen PJ, Owen T, Ruths D, Teichmann L, et al. The causes and consequences of COVID-19 misperceptions: understanding the role of news and social media. HKS Misinfo Review 2020. [doi: 10.37016/mr-2020-028]
55. Pan W, Liu D, Fang J. An examination of factors contributing to the acceptance of online health misinformation. Front Psychol 2021;12:630268 [FREE Full text] [doi: 10.3389/fpsyg.2021.630268] [Medline: 33732192]
56. Sallam M, Dababseh D, Yaseen A, Al-Haidar A, Taim D, Eid H, et al. COVID-19 misinformation: mere harmless delusions or much more? A knowledge and attitude cross-sectional study among the general public residing in Jordan. PLoS One 2020;15(12):e0243264 [FREE Full text] [doi: 10.1371/journal.pone.0243264] [Medline: 33270783]
57. Wang T, Yu W. Alternative sources use and misinformation exposure and susceptibility: the curvilinear moderation effects of socioeconomic status. Telematics and Informatics 2022;70:101819. [doi: 10.1016/j.tele.2022.101819]
58. Mordell L, Bloomgarden E, Jain S, Arora V. Medical Professionals Using Social Media to Combat Misinformation. In: Keselman A, Arnott Smith C, Amanda Wilson J, editors. Combating Online Health Misinformation: A Professional's Guide to Helping the Public. London: Rowman & Littlefield; 2022:167.
59. Southwell BG, Wood JL, Navar AM. Roles for health care professionals in addressing Patient-Held misinformation beyond fact correction. Am J Public Health 2020;110(S3):S288-S289. [doi: 10.2105/AJPH.2020.305729] [Medline: 33001723]
60. Oktavianus J, Bautista JR. Motivating healthcare professionals to correct online health misinformation: the roles of subjective norm, third-person perception, and channel differences. Computers in Human Behavior 2023;147:107839 [FREE Full text] [doi: 10.1016/j.chb.2023.107839]

XSL•FO
RenderX

61. Saltz E, Barari S, Leibowicz C, Wardle C. Misinformation interventions are common, divisive, and poorly understood. HKS Misinfo Review 2021. [doi: 10.37016/mr-2020-81]
62. Pennycook G, McPhetres J, Zhang Y, Lu JG, Rand DG. Fighting COVID-19 misinformation on social media: experimental evidence for a scalable Accuracy-Nudge intervention. Psychol Sci 2020;31(7):770-780 [FREE Full text] [doi: 10.1177/0956797620939054] [Medline: 32603243]
63. Roozenbeek J, van der Linden S. How to combat health misinformation: a psychological approach. Am J Health Promot 2022;36(3):569-575. [doi: 10.1177/08901171211070958] [Medline: 35164546]
64. Mena P. Cleaning up social media: the effect of warning labels on likelihood of sharing false news on facebook. Policy & Internet 2019;12(2):165-183 [FREE Full text] [doi: 10.1002/poi3.214]
65. Zhang J, Featherstone JD, Calabrese C, Wojcieszak M. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. Prev Med 2021;145:106408. [doi: 10.1016/j.ypmed.2020.106408] [Medline: 33388335]
66. Oeldorf-Hirsch A, Schmierbach M, Appelman A, Boyle MP. The ineffectiveness of Fact-Checking labels on news memes and articles. Mass Communication and Society 2020;23(5):682-704. [doi: 10.1080/15205436.2020.1733613]
67. Silesky MD, Panchal D, Fields M, Peña AS, Diez M, Magdaleno A, et al. A multifaceted campaign to combat COVID-19 misinformation in the hispanic community. J Community Health 2023;48(2):286-294 [FREE Full text] [doi: 10.1007/s10900-022-01170-9] [Medline: 36399238]
68. Leininger LJ, Albrecht SS, Buttenheim A, Dowd JB, Ritter AZ, Simanek AM, et al. Fight like a nerdy girl: the dear pandemic playbook for combating health misinformation. Am J Health Promot 2022;36(3):563-567 [FREE Full text] [doi: 10.1177/08901171211070956] [Medline: 35164543]

Abbreviations

ACS: American Community Survey
HINTS: Health Information National Trends Survey
NCHS: National Center for Health Statistics
OR: odds ratio
RUCA: Rural-Urban Commuting Area

Edited by N Cahill; submitted 01.12.23; peer-reviewed by H Xie, Y Wang, M Meenu; comments to author 13.02.24; revised version received 16.04.24; accepted 29.09.24; published 06.11.24.

Please cite as:
Chandrasekaran R, Sadiq T M, Moustakas E
Racial and Demographic Disparities in Susceptibility to Health Misinformation on Social Media: National Survey-Based Analysis
J Med Internet Res 2024;26:e55086
URL: https://www.jmir.org/2024/1/e55086
doi:10.2196/55086
PMID:

©Ranganathan Chandrasekaran, Muhammed Sadiq T, Evangelos Moustakas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Technology Acceptance Among Low-Income Asian American Older Adults: Cross-Sectional Survey Analysis

Pauline DeLange Martinez[1], PhD; Daniel Tancredi[2], PhD; Misha Pavel[3], PhD; Lorena Garcia[4], DrPH; Heather M Young[1], PhD

[1]Family Caregiving Institute, Betty Irene Moore School of Nursing, University of California, Davis, Sacramento, CA, United States

[2]Department of Pediatrics, University of California, Davis, Sacramento, CA, United States

[3]Khoury College of Computer Science, Northeastern University, Boston, MA, United States

[4]Department of Public Health Sciences, University of California, Davis, Sacramento, CA, United States

Corresponding Author:
Pauline DeLange Martinez, PhD
Family Caregiving Institute
Betty Irene Moore School of Nursing
University of California, Davis
2570 48th St
Sacramento, CA, 95817
United States
Phone: 1 916 426 2862
Email: pdmartinez@ucdavis.edu

Abstract

Background: Studies show that the use of information and communications technologies (ICTs), including smartphones, tablets, computers, and the internet, varies by demographic factors such as age, gender, and educational attainment. However, the connections between ICT use and factors such as ethnicity and English proficiency, especially among Asian American older adults, remain less explored. The technology acceptance model (TAM) suggests that 2 key attitudinal factors, perceived usefulness (PU) and perceived ease of use (PEOU), influence technology acceptance. While the TAM has been adapted for older adults in China, Taiwan, Singapore, and Korea, it has not been tested among Asian American older adults, a population that is heterogeneous and experiences language barriers in the United States.

Objective: This study aims to examine the relationships among demographics (age, gender, educational attainment, ethnicity, and English proficiency), PU, PEOU, and ICT use among low-income Asian American older adults. Two outcomes were examined: smartphone use and ICT use, each measured by years of experience and current frequency of use.

Methods: This was a secondary data analysis from a cross-sectional baseline survey of the Lighthouse Project, which provided free broadband, ICT devices, and digital literacy training to residents living in 8 affordable senior housing communities across California. This analysis focused on Asian participants aged ≥62 years (N=392), specifically those of Korean, Chinese, Vietnamese, Filipino, and other Asian ethnicities (eg, Hmong and Japanese). Hypotheses were examined using descriptive statistics, correlation analysis, and hierarchical regression analysis.

Results: Younger age, higher education, and greater English proficiency were positively associated with smartphone use (age: β=–.202; P<.001; education: β=.210; P<.001; and English proficiency: β=.124; P=.048) and ICT use (age: β=–.157; P=.002; education: β=.215; P<.001; and English proficiency: β=.152; P=.01). Male gender was positively associated with PEOU (β=.111; P=.047) but not with PU (β=–.031; P=.59), smartphone use (β=.023; P=.67), or ICT use (β=.078; P=.16). Ethnicity was a significant predictor of PU ($F_{4,333}$=5.046; P<.001), PEOU ($F_{4,345}$=4.299; P=.002), and ICT use ($F_{4,350}$=3.177; P=.01), with Chinese participants reporting higher levels than Korean participants, who were the reference group (β=.143; P=.007). PU and PEOU were positively correlated with each other (r=0.139, 95% CI=0.037-0.237; P=.007), and both were significant predictors of smartphone use (PU: β=.158; P=.002 and PEOU: β=.166; P=.002) and ICT use (PU: β=.117; P=.02 and PEOU: β=0.22; P<.001), even when controlling for demographic variables.

Conclusions: The findings support the use of the TAM among low-income Asian American older adults. In addition, ethnicity and English proficiency are significant predictors of smartphone and ICT use among this population. Future interventions should consider heterogeneity and language barriers of this population to increase technology acceptance and use.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52498) doi:10.2196/52498

KEYWORDS

aged; older adults; Asian American; immigrant; vulnerable populations; internet; information and communications technology; ICT; digital divide; technology acceptance model; mobile phone

Introduction

Background

Older adults increasingly use information and communications technologies (ICTs), including smartphones, tablets, computers, and the internet, to manage their health and finances, seek information, access services, and stay socially connected [1]. However, there are significant disparities in ICT use among older adults, particularly among those with low income [2]. According to the most recent study of intersectionality and technology use, only 46% of low-income adults aged ≥65 years in the United States used the internet as compared to 90% and 94% of mid-income and high-income older adults, respectively [3]. In addition, the ownership of smartphones, computers, and broadband subscriptions is lower among low-income older adults [4].

This digital divide is exacerbated among Asian American older adults, who experience higher rates of poverty (9.3%) as compared to the general older adult population in the United States (8.9%) [5]. In an analysis of the California Health Interview Survey (CHIS), the combination of Asian ethnicity and low income had an interactive, negative effect on ICT use [6]. In fact, Asian American older adults in the lowest income category were 14 times less likely to use the internet for health information compared to older non-Hispanic White individuals in the highest income category [6].

The COVID-19 pandemic highlighted the impact of limited digital access and skills among older adults. Compared to older adults with ICT access and proficiency, older adults who were not digitally connected faced health disparities, including social isolation [7,8], difficulties in accessing timely COVID-19–related information [9], and barriers to health care [10].

Asian American individuals are the fastest growing racial or ethnic group among adults aged ≥65 years in the United States [11,12]. In 2019, Asian Americans made up 4.6% of older adults in the United States, and they are projected to comprise approximately 8% of US older adults by 2060, at 7.9 million [13]. However, few studies have focused on technology use among Asian American older adults, despite indications that this population is among the least likely to use ICTs [6,14,15]. The 2011 National Health and Aging Trends Study indicated that Asian American older adults were less likely to send emails, SMS text messages, or conduct web-based tasks as compared to non-Hispanic White older adults [15]. Furthermore, a CHIS analysis between 2011 and 2016 found that internet use among non-Hispanic White older adults increased significantly from 66% to 73%; however, during this same period, internet use did not significantly change among Asian, Latinx, or Black older adults [14].

Asian Americans are frequently lumped together in research, but this population is heterogeneous, including >40 ethnicities, each with diverse cultural backgrounds, languages, and immigration histories [11,12,16]. Asian American ethnic groups vary significantly in income, educational attainment, English proficiency, and health status [17]. Few studies compare ICT use among Asian American ethnicities, though one exception is the study by Gordon and Hornbrook [18]. This analysis of a stratified, randomized survey of patients from a major health system in California revealed that Chinese older adults were significantly more likely to be using ICTs than Filipino older adults. While Chinese and non-Hispanic White older adults were equally likely to use the internet and access patient portals, Filipino older adults were less likely to own computers or smartphones, use the internet and email, or be willing and able to use digital technology to perform health-related tasks, including seeking health information [18]. Specifically, 84% of non-Hispanic White adults and 79% of Chinese adults aged 65 to 79 years reported being able to use the internet alone or with some help, as compared to only 53% of Filipino older adults. In addition, although 80% of non-Hispanic White older adults and Chinese older adults were able to send and receive email, only 60% of Filipino older adults were able to do so, even with help [18].

Limited English proficiency (LEP) is another characteristic that is negatively associated with ICT use. In a cross-sectional analysis of data from the 2011 National Health and Aging Trends Study, LEP was a significant, independent predictor negatively associated with using email and SMS text messaging, conducting web-based personal tasks, and seeking health information on the internet [15]. Furthermore, in an analysis of the CHIS, older adults with LEP were 53% less likely to report using the internet to seek health information as those who spoke English well or very well [6]. While these studies were not specific to Asian American older adults, LEP is a pertinent characteristic to consider when understanding technology use among Asian American older adults. According to the National Asian Pacific Center on Aging, >60% of Asian Americans aged ≥65 years have LEP, including >85% of Vietnamese older adults and >67% of Korean and Chinese older adults [11].

Theoretical Framework

Over the past 4 decades, researchers have developed several models to describe the process of technology acceptance. The first models were based on the theory of planned behavior (TPB) by Ajzen [19], developed in 1980, which posits that one will not change their behavior until they have the psychological intent. The TPB suggests that 3 key factors, attitudes, subjective norms, and perceived behavioral control, influence intention. While the TPB is broadly about any behavior change, the technology acceptance model (TAM) proposed by Davis [20] adapted the TPB to describe factors that lead to technology use. The TAM proposes that perceived usefulness (PU) and

XSL•FO
RenderX

perceived ease of use (PEOU) shape attitudes toward technology, which in turn shapes intention to use technology. PU refers to whether one perceives technology to be useful for what they hope to achieve, while PEOU refers to how much effort one expects to need to make to learn to use a new technology. PEOU has been operationalized to measure feelings of confusion, frustration, ease, predictability, intuitiveness of the system, or frequency of making mistakes [20].

While the TAM focused on understanding technology acceptance in the workplace [20], it has been adapted and tested with various subpopulations, in various geographies, and with a variety of technology applications. Later adaptations operationalized the variables using different measures, and they incorporated demographic factors as control variables in the model, including age, gender, and educational attainment [21,22].

Age has consistently been a significant, negative predictor of ICT use. According to 2017 Pew Research, 82% of adults aged 65 to 69 years in the United States reported using the internet; this number dropped to 75% for adults aged 70 to 74 years, 60% for adults aged 75 to 79 years, and 44% for adults aged ≥80 years [3]. Age is also negatively associated with smartphone ownership and home broadband subscription [3]. Baby boomers are more likely to have encountered ICTs in the workforce, leading to greater ICT acceptance and stronger digital skills [23-25]. In contrast, compared to baby boomers, adults aged ≥75 years experience greater technology anxiety, have lower technology proficiency, engage in less diverse web-based activities, and require more support to use ICTs [26-28].

Although gender is sometimes included as a control variable in the TAM [21,22], research examining the relationship between gender and ICT use has shown mixed results. While many studies found that older men are more likely to use the internet than older women [29-31], some recent studies have found either a shrinking gap between genders [29] or no difference [32]. Notably, a recent analysis of data from the Health and Retirement Study demonstrated that women aged ≥50 years were more likely to access the internet than men, although this difference decreased with age [33].

Educational attainment and ICT use are consistently, positively associated. In 2017, Americans aged ≥65 years who were college graduates were significantly more likely to use the internet, have a smartphone, and subscribe to home broadband (92%, 65%, and 82%, respectively), as compared to those who had a high school degree or less (49%, 27%, and 30%, respectively) [3]. In a cross-sectional analysis of a nationally representative sample of Medicare beneficiaries, education was one of the strongest predictors of internet use [34]. Compared to college graduates, high school graduates had 18% odds of using the internet, and those with less than a high school education had 8% odds of using the internet. In addition, those with a high school degree or less were likely to use the internet exclusively for emails and texting, while those with a college degree were more likely to engage in web-based shopping, banking, or health-related internet tasks [34].

While the TAM and its subsequent adaptations have been used to understand technology use among older adults living in China, Taiwan, Singapore, and Korea, the TAM has not been used to understand technology acceptance among Asian American older adults. Compared to Asian individuals living in their countries of origin, Asian American individuals are unique in their ethnic diversity. In addition, the majority of Asian American older adults in the United States experience language barriers [11]. While researchers have incorporated demographic variables such as age, gender, and educational attainment into the TAM, ethnicity and English proficiency are novel predictors that may increase the TAM's ability to predict ICT use among Asian American older adults.

Objectives

This cross-sectional survey explored the relationships among age, gender, educational attainment, English proficiency, ethnicity, PU, PEOU, and ICT use among low-income Asian American older adults. This study examined 2 outcomes: smartphone use and ICT use. In addition, this study tested a simplified version of the TAM, as shown in Figure 1. In this adapted model, we removed the mediators from the original model (attitude and behavioral intention to use) [20]. Furthermore, this adaptation involved revising the TAM to shift the focus from workplace usefulness to perceptions of technology among older adults.

XSL•FO
RenderX

Figure 1. The technology acceptance model adapted for low-income Asian American older adults. ICT: information and communications technology.

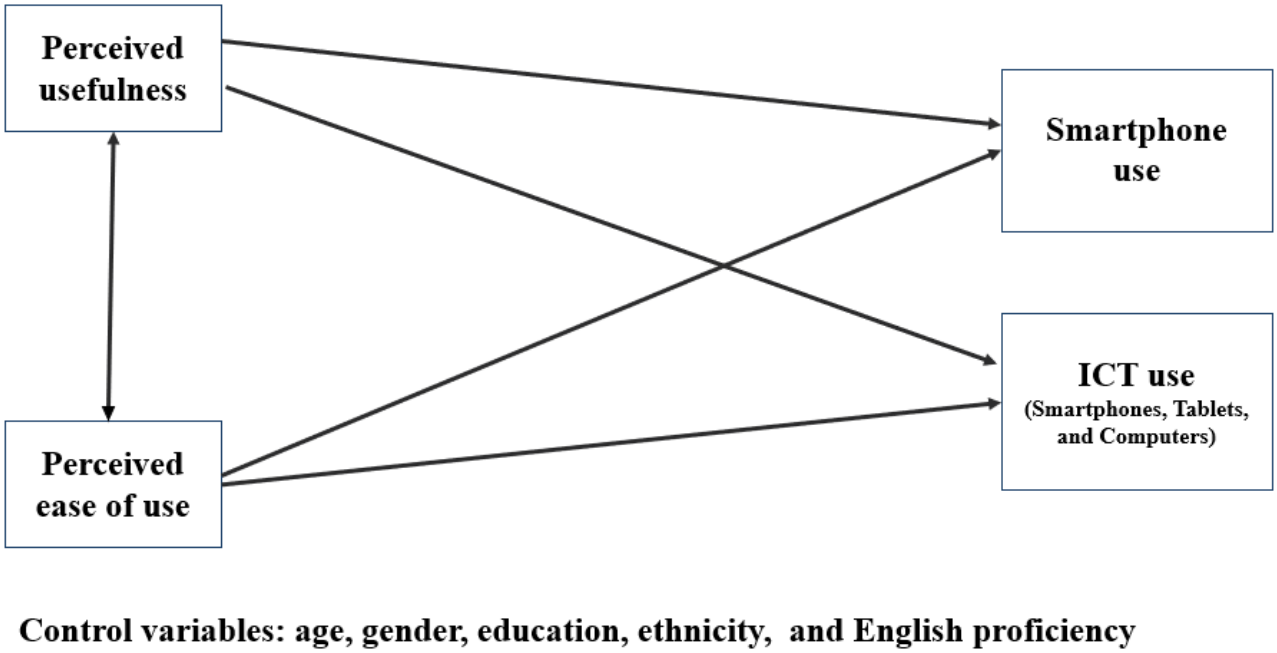

This study tested 4 hypotheses.

- Hypothesis 1: PEOU will be positively associated with PU.
- Hypothesis 2: age, gender, educational attainment, English proficiency, and ethnicity will be associated with PU and PEOU. We expect that age will be negatively associated with PU and PEOU; male gender, educational attainment, and English proficiency will be positively associated with PU and PEOU; and there will be heterogeneity in associations with PU and PEOU across Asian ethnicities.
- Hypothesis 3: age, male gender, educational attainment, English proficiency, and ethnicity will be associated with smartphone use and ICT use. We expect that age will be negatively associated with smartphone use and ICT use; male gender, educational attainment, and English proficiency will be positively associated with smartphone use and ICT use; and there will be heterogeneity in associations with smartphone use and ICT use across Asian ethnicities.
- Hypothesis 4: PU and PEOU will be positively associated with smartphone use and ICT use, even when accounting for age, gender, educational attainment, English proficiency, and ethnicity.

Methods

Ethical Considerations

This study was a secondary data analysis of the Lighthouse Project. On the basis of the HRP-210 Determination Request, the University of California, Davis, institutional review board determined that this research is exempt as it did not directly involve human participants and used deidentified secondary data (1938286-1).

The Lighthouse Project

In partnership with 2 senior housing providers (Front Porch and Eskaton), the Lighthouse Project provided 8 affordable senior housing communities (open to residents aged >62 years) across California with high-speed broadband access, ICTs, and digital literacy training between July 2021 and July 2022 [35]. All residents (N=1050) were invited to participate. Recruitment occurred over a period of 1 to 6 months at each community during Wi-Fi installation. Onsite staff distributed flyers about the Lighthouse Project, set up an informational table in lobby areas or community rooms, and held social events where residents could test out a variety of ICT devices to inform device selection for the project and gather input. The Lighthouse Project participants completed surveys at entry and again at 30 and 90 days, received an ICT device, and were asked to attend a minimum of 1 of a series of digital literacy classes offered onsite at each community.

Hard copy surveys were distributed to residents' apartments. Surveys were available in several languages; they were self-administered and collected by staff at each community. Staff were available to provide clarification and help with completing surveys, if needed. Most staff spoke English and at least 1 additional language, although on occasion, Google Translate or pocket translators were used to communicate with residents. Survey data were entered by affordable senior housing staff and verified by researchers from the University of California, Davis.

To address our study hypotheses, this study consisted of a secondary, cross-sectional analysis of a subset of Lighthouse Project baseline surveys. Participants were included if they reported being of Asian ethnicity and were aged ≥62 years. This age was selected based on the housing eligibility criteria. Participants were excluded if they had missing responses to any survey items that contributed to the 2 outcome variables (smartphone use and ICT use).

XSL•FO
RenderX

Measures

Overview

In the Lighthouse Project, an evaluation expert and a gerontologist developed an evidence-informed survey that included demographic items and self-reported assessments of physical and emotional health, risk of depression, social isolation, and loneliness. It also featured technology-specific questions that examined attitudes toward ICTs, frequency and preferred use of devices, and available technology support. The survey was adapted for low-literacy levels, translated, and pilot-tested with affordable senior housing staff and residents across 2 communities representing 5 different languages (Chinese, Korean, Vietnamese, Spanish, and English). To minimize participant burden and increase accessibility, surveys were limited to 5 pages (front and back) and used large font with generous spacing.

Outcome Variables: Smartphone Use and ICT Use

In this study, smartphone use was operationalized as the standardized sum of 2 survey items. The first measured frequency of smartphone use ("How often do you use a smartphone [iPhone or Android]?"). Answer choices were coded as 0=never or I do not own, 1=once or less than once per week, 2=2-4 times per week, or 3=about once per day. The second item measured years of experience using ICTs ("How long have you been using technology, such as a computer, laptop, tablet or smartphone?"). Answer choices were coded as 0 (I have never used these), 1 (<1 year), 2 (1-2 years), or 3 (>2 years).

ICT use was operationalized as the sum of 4 survey items. Three separate survey items asked about frequency of use of computers, tablets, or smartphones ("How often do you use a desktop or laptop computer?" "How often do you use a tablet or iPad?" and "How often do you use a smartphone [iPhone or Android]?"). The fourth item inquired about years of experience using ICTs ("How long have you been using technology, such as a computer, laptop, tablet or smartphone?").

Operationalization of outcome variables was assessed by 4 researchers with expertise in survey methodology with older adults, and data were reviewed by a biostatistician and a nurse researcher. When reviewing Predicted Probability plots, we observed that the residuals followed a normal distribution, closely aligning with the plots' diagonal line. In addition, scatterplots of the residuals displayed no discernible pattern, with points evenly distributed both above and below 0 on the x-axis and to the left and right of 0 on the y-axis.

Attitudinal Factors: PU and PEOU

When Davis [20] developed the original TAM in the workplace, PU referred to usefulness in the work environment (eg, productivity and efficiency). As this study is focused on older adults, PU was adapted to encompass perceptions of technology as being useful for social connection, information seeking, and learning new skills [36]. Specifically, PU was operationalized as the standardized sum of 2 items, "Technology helps me be connected with family and friends" and "Technology helps me learn new information and skills." Response categories for all statements ranged on a scale of 1 (strongly disagree) to 4 (strongly agree).

PEOU was operationalized as the standardized sum of 6 survey items combining 2 validated subscales. The Attitudes Toward Computers Questionnaire—Comfort Subscale [37] includes 4 items ("I feel comfortable with technology" [reverse scored], "Technology makes me nervous," "I don't feel confident about my ability to use technology," and "Technology is confusing."). The Senior Technology Acceptance Model—Technology Anxiety Subscale [22] includes 2 items ("I feel apprehensive about using technology" and "I hesitate to use technology for fear of making mistakes I cannot correct."). Response categories for all statements ranged on a scale from 1 (strongly agree) to 4 (strongly disagree).

Demographic Factors

Demographic information included age (in years, centered at 80), gender (0=female, 1=male, and 2=other), and educational attainment (coded as –2=never attended school, –1=some high school, 0=completed high school or obtained a General Educational Development certificate, 1=some college, 2=college degree, or 3=graduate degree).

English proficiency was assessed by asking, "How well do you speak English?" Answer responses were coded as –1=not at all, 0=not well, 1=well, or 2=very well.

Five binary ethnicity variables were included in this study, including Korean, Chinese, Vietnamese, Filipino, or other Asian. Responses were mutually exclusive.

Statistical Analysis

All data analysis was conducted using SPSS Statistics (version 28; IBM Corp). First, we computed descriptive statistics, including demographic characteristics, attitudes toward technology, and smartphone and ICT use. We assessed the distribution and internal consistency of the attitudinal variables, PU and PEOU, and the 2 composite outcome variables (smartphone use and ICT use).

Next, we used Pearson correlation analysis to examine relationships among all dependent and independent variables. To test hypothesis 1, we examined the unadjusted correlation between PU and PEOU.

To assess the other 3 hypotheses, hierarchical multiple regression was conducted using stepwise blocks [38]. To test hypothesis 2, outcome variables included PU and PEOU; to test hypotheses 3 and 4, outcome variables were smartphone use and ICT use. The sequence of inclusion of independent variables was established based on previous studies [22]. Demographic variables such as age, male gender, education level, English proficiency, and ethnicity (with Korean ethnicity as the reference group) were added in the first step (model 1). Next, attitudes toward technology (PU and PEOU) were added (model 2); finally, interaction terms were included in a forward stepwise regression (model 3). Interaction terms involving all pairwise combinations of demographic variables (except for ethnicity) with attitudinal variables (PU and PEOU) were also included. In addition, we evaluated the interaction of age and education, reasoning that female Asian Americans had systematic lower

XSL•FO
RenderX

access to education in their native Asian country, and this may result in gender-specific adjusted associations of education with outcomes [39]. At each step, we assessed goodness of fit (adjusted R^2), that is, the percentage of variability in the dependent variable that could be accounted for by the predictors. With each new set of terms, the change in R^2 was calculated to quantify the change in the predictive power. The α level for testing significance was set to .05.

Results

Participation

In the Lighthouse Project, across the 8 affordable senior housing communities, 58% (609/1050) of the residents participated and 68% (414/609) of the Lighthouse participants were Asian. Two of the Lighthouse communities did not have Asian residents and were therefore excluded from this study.

Residents opted out of participating in the Lighthouse Project for several reasons, including fear or perceived burden of learning a new device or attending training classes; challenges related to vision, hearing, mobility, or cognitive decline; already having a device; or not finding the Lighthouse devices relevant.

After applying the inclusion and exclusion criteria, the final sample included 392 Asian residents aged ≥62 years, living in 6 affordable senior housing complexes located across California. Participants' ages ranged from 62 to 97 years, with a mean of 79.1 (SD 6.95) years. Participant demographics are described in Table 1.

XSL•FO
RenderX

Table 1. Participant demographics (N=392)[a].

Characteristic	Participants, n (%)
Age (y)	
62-74	105 (27.6)
75-84	192 (50)
>85	84 (22)
Gender	
Female	266 (68.2)
Male	124 (31.8)
Other	0 (0)
Ethnicity	
Chinese	73 (18.6)
Filipino	12 (3.1)
Korean	274 (69.9)
Other Asian (eg, Japanese or Hmong)	14 (3.6)
Vietnamese	19 (4.8)
English proficiency	
Very well	8 (2.1)
Well	63 (16.3)
Not well	198 (51.3)
Not at all	117 (30.3)
Educational attainment	
Never attended school	22 (5.9)
Some high school	133 (35.4)
Completed high school or obtained a General Educational Development certificate	76 (20.2)
Some college	62 (16.5)
College degree	60 (16)
Graduate degree	23 (6.1)
Years of experience using ICTs[b]	
>2	227 (57.9)
1-2	45 (11.5)
<1	30 (7.7)
I have never used these	90 (23)
Computer use	
About once/d	89 (22.7)
2-4 times/wk	30 (7.7)
Once or less than once/wk	27 (6.9)
Never	246 (62.8)
Tablet use	
About once/d	106 (27)
2-4 times/wk	35 (8.9)
Once or less than once/wk	20 (5.1)
Never	231 (58.9)

XSL•FO
RenderX

Characteristic	Participants, n (%)
Smartphone use	
About once/d	243 (62)
2-4 times/wk	40 (10.2)
Once or less than once/wk	17 (4.3)
Never	92 (23.5)
Number of ICT device types used (computer, tablet, and smartphone)	
No devices	78 (19.9)
1 device	129 (32.9)
2 devices	77 (19.6)
3 devices	108 (27.6)

[a]Missing data: 11 participants did not report their age, 2 participants did not report their gender, 6 participants did not report English proficiency, and 16 participants did not report their education level.

[b]ICT: information and communications technology.

Internal Consistency of ICT Use and PEOU

We measured internal consistency of the items comprising the measures for ICT use and PEOU. The 4 items that made up ICT use had strong internal consistency and reliability, with a Cronbach α of 0.74. The 6 items that made up PEOU also had strong internal consistency and reliability, with a Cronbach α of 0.89.

Correlation Analysis

Correlation analysis was used to examine relationships among demographic characteristics, attitudes toward technology, and smartphone and ICT use (Table 2).

XSL•FO
RenderX

Table 2. Correlation analysis (Pearson *r* and 2-tailed *P* value) among the research variables (N=392).

Variable	Smartphone use	ICT[a] use	PU[b]	PEOU[c]	Age	Male gender	Educational attainment	English proficiency
Smartphone use								
r	1	0.846[d]	0.234[d]	0.276[d]	–0.262[d]	0.127[d]	0.295[d]	0.256[d]
P value	__[e]	<.001	<.001	<.001	<.001	.01	<.001	<.001
ICT use								
r	0.846[d]	1	0.218[d]	0.327[d]	–0.225[d]	0.187[d]	0.321[d]	0.260[d]
P value	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001
PU								
r	0.234[d]	0.218[d]	1	0.139[d]	–0.087	0.019	0.129[d]	0.064
P value	<.001	<.001	—	.007	.10	.71	.01	.22
PEOU								
r	0.276[d]	0.327[d]	0.139[d]	1	–0.141[d]	0.157[d]	0.186[d]	0.240[d]
P value	<.001	<.001	.007	—	.006	.002	<.001	<.001
Age (y)								
r	–0.262[d]	–0.225[d]	–0.087	–0.141[d]	1	–0.009	–0.106	–0.194[d]
P value	<.001	<.001	.10	.006	—	.86	.04	<.001
Male gender								
r	0.127[d]	0.187[d]	0.019	0.157[d]	–0.009	1	0.331[d]	0.189[d]
P value	.01	<.001	.71	.002	.86	—	<.001	<.001
Educational attainment								
r	0.295[d]	0.321[d]	0.129[d]	0.186[d]	–0.106	0.331[d]	1	0.403[d]
P value	<.001	<.001	.01	<.001	.04	<.001	—	<.001
English proficiency								
r	0.256[d]	0.260[d]	0.064	0.240[d]	–0.194[d]	0.189[d]	0.403[d]	1
P value	<.001	<.001	.22	<.001	<.001	<.001	<.001	—
Korean								
r	0.002	–.095	–0.203[d]	–.131[d]	0.003	0.014	–0.015	0.020
P value	.97	.06	<.001	.01	.96	.78	.77	.70
Chinese								
r	–0.045	.051	0.171[d]	0.058	0.102	0.030	–0.048	–0.280[d]
P value	.38	.31	.001	.25	.047	.56	.36	<.001
Vietnamese								
r	0.066	.100	0.049	–0.014	–0.113	–0.001	0.002	0.079
P value	.19	.049	.35	.79	.03	.98	.97	.12
Filipino								
r	0.109	.104	0.091	0.220[d]	–0.077	–0.090	0.172[d]	0.329[d]
P value	.03	.04	.08	<.001	.13	.08	.001	<.001
Other Asian								
r	–0.088	–.084	–0.002	0.011	–0.012	–0.013	–0.028	0.142
P value	.08	.10	.97	.83	.81	.79	.59	.005

XSL•FO
RenderX

[a]ICT: information and communications technology.

[b]PU: perceived usefulness.

[c]PEOU: perceived ease of use.

[d]The correlation is significant at a significance level of .01 (2-tailed).

[e]Not applicable.

PU and PEOU were positively, significantly associated with one another (r=0.139, 95% CI 0.037 to –0.237; P=.007). Hence, hypothesis 1 (PEOU will be positively associated with PU) was supported; we reject the null hypothesis.

Regression Analysis

Results of hierarchical regression are presented in Tables 3 and 4. Tolerance values of all independent variables were >0.01, and variance inflation factor values were <5, indicating lack of multicollinearity [40].

Table 3. Results of hierarchical regression: perceived usefulness (PU) and perceived ease of use (PEOU) as dependent variables (N=392)[a].

Model 1: independent variables	PU		PEOU	
	β	*P* value	β	*P* value
Age	–.084	.12	–.099	.06
Male gender	–.031	.59	.111	.047
Education	.121	.047	.045	.44
English proficiency	.039	.55	.148	.02
Ethnicity				
Chinese	.231	<.001	.127	.02
Filipino	.074	.20	.179	.001
Korean	Reference	Reference	Reference	Reference
Other Asian	–.021	.70	.001	.99
Vietnamese	.068	.20	–.019	.71
Adjusted R^2	.058	—	.098	—

[a]β values are standardized regression coefficients.

XSL•FO
RenderX

Table 4. Results of hierarchical regression: smartphone and information and communications technology (ICT) use as dependent variables (N=392)[a].

Model and independent variables	Smartphone use		ICT use	
	β	*P* value	β	*P* value
1				
Age	–.202	<.001	–.157	.002
Male gender	.023	.67	.078	.16
Education	.210	<.001	.215	<.001
English proficiency	.124	.048	.152	.01
Ethnicity				
Chinese	.016	.76	.143	.007
Filipino	.017	.75	.019	.72
Korean	Reference	Reference	Reference	Reference
Other Asian	–.124	.02	–.089	.08
Vietnamese	.025	.63	.077	.13
Adjusted R^2	.137	—	.155	—
2				
Age	–.171	.001	–.124	.01
Male gender	.010	.86	.057	.29
Education	.184	.001	.191	.001
English proficiency	.096	.12	.118	.05
Ethnicity				
Chinese	–.041	.44	.089	.09
Filipino	–.025	.65	–.030	.58
Korean	Reference	Reference	Reference	Reference
Other Asian	–.118	.02	–.083	.09
Vietnamese	.018	.73	.074	.13
PU	.158	.002	.117	.02
PEOU	.166	.002	.221	<.001
Adjusted R^2	.184	—	.211	—
ΔR^2	.047	<.001	.056	<.001
3			—	—
Age	–.166	.001	—[b]	—[b]
Male gender	.009	.86	—[b]	—[b]
Education	.168	.004	—[b]	—[b]
English proficiency	.111	.07	—[b]	—[b]
Ethnicity				
Chinese	–.057	.29	—[b]	—[b]
Filipino	.019	.74	—[b]	—[b]
Korean	Reference	Reference	—[b]	—[b]
Other Asian	–.110	.03	—[b]	—[b]
Vietnamese	.006	.91	—[b]	—[b]

XSL•FO
RenderX

Model and independent variables	Smartphone use		ICT use	
	β	*P* value	β	*P* value
PU	.183	.001	__[b]	__[b]
PEOU	.157	.003	__[b]	__[b]
English proficiency×PEOU	–.120	.03	__[b]	__[b]
Adjusted R^2	.193	—	__[b]	__[b]
ΔR^2	.009	.03	__[b]	__[b]

[a]β values are standardized regression coefficients.

[b]No interaction terms were kept in a stepwise regression.

Testing Hypothesis 2 (Model 1: PU and PEOU as Outcome Variables)

As expected, educational attainment was positively, significantly associated with PU when controlling for age, gender, English proficiency, and ethnicity. In addition, we found heterogeneity in associations with PU across Asian ethnicities; the omnibus *F* test for assessing the adjusted association of the 5-level ethnicity classification with PU was statistically significant ($F_{4,333}$=5.046; *P*<.001). Specifically, Chinese ethnicity was significantly, positively associated with PU as compared to the reference level (Korean ethnicity). However, contrary to expectation, age, gender, and English proficiency were not significant predictors of PU.

Male gender and English proficiency were each significant, positive predictors of PEOU when controlling for age, educational attainment, and ethnicity. In addition, we found heterogeneity in association with PEOU across Asian ethnicities; the omnibus *F* test for assessing the adjusted association of the 5-level ethnicity classification with PEOU was statistically significant ($F_{4,345}$=4.299; *P*=.002). Contrary to expectation, age and educational attainment were not significant predictors of PEOU. Therefore, hypothesis 2 (age, gender, educational attainment, English proficiency, and ethnicity will be associated with PU and PEOU) was partially supported.

Testing Hypothesis 3 (Model 1: Smartphone Use and ICT Use as Outcome Variables)

As expected, age was significantly, negatively associated with smartphone use, and educational attainment and English proficiency were significantly, positively associated with smartphone use, when controlling for other demographic variables. Contrary to hypothesis 3, male gender was not a significant predictor of smartphone use. In addition, ethnicity was not a significant predictor of smartphone use; the omnibus *F* test for assessing the adjusted association of the 5-level ethnicity classification with smartphone use was not statistically significant ($F_{4,340}$=1.619; *P*=.17). Nevertheless, indicator variables for ethnicity were still included in the overall model (with Koreans as the reference category), and regression analysis revealed that participants of “other Asian” ethnicity were significantly less likely than Koreans to be using smartphones.

Likewise, age was significantly, negatively associated with ICT use, and educational attainment and English proficiency were significantly, positively associated with ICT use, when controlling for other demographic variables. While ethnicity was a significant predictor of ICT use when controlling for age, gender, education, and English proficiency, significance was lost when we added PU and PEOU to the model. The omnibus *F* test for assessing the adjusted association of the 5-level ethnicity classification in the final model (Table 4: model 3) was $F_{4,340}$=2.087; *P*=.08. Finally, contrary to hypothesis 3, male gender was not a significant predictor of ICT use. Therefore, overall, hypothesis 3 (age, male gender, educational attainment, English proficiency, and ethnicity will be associated with smartphone use and ICT use) was partially supported.

Testing Hypothesis 4 (Models 2 and 3: Smartphone Use and ICT Use as Outcome Variables)

Adding PU and PEOU to the model significantly strengthened the predictive power of smartphone use and ICT use. Hypothesis 4 was fully supported; PU and PEOU were significant, positive predictors of smartphone use and ICT use, when accounting for age, gender, educational attainment, English proficiency, and ethnicity.

There was a significant interaction between PEOU and English proficiency in the model predicting smartphone use. None of the interaction terms tested were significant in the model predicting ICT use.

Discussion

This was the first study to test the TAM [20] among low-income Asian American older adults. We sought to understand what demographic factors predict smartphone use and ICT use in this population.

Smartphone Use

Approximately three-quarters of participants (300/392, 76.5%) reported that they use a smartphone, with most of these (283/300, 94.3%) reporting that they use a smartphone at least twice per week. These findings are high compared to a nationally representative study from 2022 that reported that 61% of Americans aged ≥65 years (of all incomes) owned a smartphone [41]. We would expect our sample to have lower smartphone ownership due to their income [42].

However, it is also possible that participants in our study confused smartphones and regular mobile phones, leading to

XSL•FO
RenderX

overreporting smartphone use. According to 2017 Pew Research, 80% of Americans aged ≥65 years owned a cellphone of any kind (including smartphones and nonsmartphones) [3], and similarly, Gordon and Hornbrook [18] found that 81% of patients aged 65 to 79 years from a large health plan in California reported owning a cell phone of any kind. These percentages are similar to the prevalence of smartphone use that we found in our study (300/392, 76.5%). While our survey defined smartphones as "an iPhone or Android," this verbiage may not have been sufficient. During the Lighthouse Project (the parent study for this analysis), affordable senior housing residents expressed confusion about terminology such as "tablet" and "Wi-Fi." As older adults did not grow up with the technology available today, it is particularly important to use clear and explicit language to avoid confusion and inferences [43]. In addition, translating technology terms to other languages may contribute to further misunderstandings.

Tablet Use

Tablet ownership in our sample was similar to the national average, with 41.1% (161/392) of the participants reporting that they use a tablet and 87.3% (141/161) of the tablet users saying that they use their tablet at least twice per week. In comparison, in 2021, in total 44% of Americans aged ≥65 years reported owning a tablet [41]. However, we expected to see lower tablet ownership in our study population because all participants had low income, while the previous study included individuals from all income levels [42].

Computer Use

A little more than one-third of the participants (146/392, 37.2%) reported that they use a computer (desktop or laptop), with 81.6% (119/146) of these reporting that they use a computer at least twice per week. In comparison, in the study by Gordon and Hornbrook [18] including patients aged 65 to 79 years from a large health system in California, 81.5% reported having access to a desktop, laptop, or notebook computer.

While computer use appears to be low in our population, these studies are difficult to compare as our study asked about "use," while the previous study focused on "access." This is a common challenge across the research literature related to technology acceptance; researchers use a variety of measurement constructs, and outcomes can vary widely, including "intention to use," "access," and "actual use" [26,44]. Furthermore, most studies depend on self-reported data, which limits their validity.

Demographic Associations With PU, PEOU, Smartphone Use, and ICT Use

Age and Education

Our regression analysis revealed that among low-income Asian American adults aged ≥62 years, younger age and greater educational attainment are independently, positively associated with smartphone use and ICT use. This is consistent with studies conducted with older adults living in the United States [3,27,45], China [46], and Korea [47]. Although these other studies evaluated various outcomes, they consistently found positive relationships between younger age, higher education, and constructs related to ICT use. For example, in a nationally representative survey of Americans aged ≥65 years, there were independent, positive associations between younger age, educational attainment, smartphone ownership, and internet use [3]. In a cross-sectional survey of New Englanders aged ≥65 years, age and educational attainment predicted ICT use [45]. Another survey study of 500 Americans aged ≥60 years found significant associations among age, education, ICT access, and internet skills [27]. In a study in China, age and education were significant factors predicting smartphone acceptance among older adults [46]. Finally, in a study in Korea, educational attainment was a significant predictor of technology acceptance among older adults [47].

Although age was a predictor of smartphone and ICT use, in our study, it did not emerge as a significant predictor of PU or PEOU when controlling for other variables. We also found that education was positively, significantly associated with PU, but not PEOU, when accounting for age, gender, educational attainment, English proficiency, and ethnicity. These findings are consistent with those of Mitzner et al [48], who also found that among low-income older adults, age was not significantly associated with PU or PEOU, and education was not associated with PEOU. In contrast, other studies that engaged a broader range of age groups (not only older adults) found significant relationships between age and PU, PEOU, computer anxiety, and computer self-efficacy [49,50].

Gender

We did not find significant associations between gender and smartphone use or ICT use. These findings are consistent with several other studies [29,32,45,48,51,52]. Interestingly, male gender was a significant, positive predictor of PEOU, although it was not associated with PU. In a systematic review focused on gender differences in technology use, men were more technologically adept than women [53]. Other studies suggest that women and men use ICTs in different ways, particularly when it comes to maintaining social relationships [53,54]. Some researchers explored the impact of the cultural construct of masculinity in Asian cultures on technology acceptance [55,56]. According to Hofstede [57] and Hofstede and Bond [58], masculinity-femininity refers to the preference for ambition or material success. Masculine cultures stress greater gender roles, and this may lead Asian men to invest more time in learning to use technology as compared to Asian women, possibly increasing men's PEOU [55,56].

English Proficiency

English proficiency had significant adjusted associations with smartphone use and ICT use, as well as with the mediator PEOU, but not PU. Further investigation showed that English proficiency remained associated with ICT use when adjusting for the mediators PU and PEOU. For the smartphone use outcome, English proficiency modified the association of the mediator PEOU. These findings suggest that participants with LEP feel less comfortable using ICTs, and this in turn may lead to lower smartphone and ICT use. Our findings align with those of others who found that LEP was a significant, independent predictor negatively associated with using email and SMS text messaging, conducting web-based personal tasks, and seeking health information on the internet in a nationally representative

XSL•FO
RenderX

sample of older adults [15]. It is important to note that, in our study, participants who were Filipino or in the "other Asian" category were more likely to be proficient in English. This is expected for the Filipino population, as English is one of the official languages in the Philippines.

Language barriers can hinder the ability to access relevant information on the internet, navigate user interfaces, and communicate effectively. Many websites, apps, and user interfaces are limited in non-English languages [59]. Most online content is in English, and other Asian languages, such as Vietnamese, are significantly less represented on the internet [60]. Though there are an estimated 6000 languages used around the globe today, Google Search is only available in 130 different languages [59]. Availability of mobile apps also varies by language. The iOS App Store offers mobile apps in only 40 languages, and many popular apps are only offered in a handful of languages [61]. In addition, popular voice user interfaces, such as Google Voice and Alexa, operate in limited languages; for example, Google Assistant does not function in Cantonese [62]. Individuals who speak English as a second language face multiple challenges when learning to use Google Home smart speakers; they struggled with using structured voice commands, choosing the "right" words to activate the device, and pronouncing instructions in a manner that the Google voice interface can recognize [63].

Other studies have focused on LEP in the context of health-related ICT use. Health portals, websites, mobile health apps, and digital health interventions are often challenging for those with LEP to use [64-66]. In addition, Californian older adults with LEP were 53% less likely to report using the internet to seek health information compared to those who indicated they spoke English well or very well [6].

Ethnicity

Ethnicity was a significant predictor of PU, PEOU, and ICT use, although it was not a significant predictor of smartphone use. Specifically, with Korean participants as the reference group, Chinese ethnicity was positively associated with PU, PEOU, and ICT use when adjusting for age, gender, educational attainment, and English proficiency. In addition, Filipino ethnicity was associated with greater PEOU compared to Korean ethnicity.

Few previous studies have compared technology acceptance across Asian ethnicities. One exception was the finding that Chinese older adults were significantly more likely to use ICTs than Filipino older adults in California [18]. Another study found that low-income older adults living in Korea had greater comfort and use of ICTs as compared to low-income older adults living in the United States [67]. In Korea, internet is a prerequisite for all types of daily activities; broadband is cheap and accessible; and Korean apps (such as Kakao Talk) are ubiquitous, even among low-income older adults [67]. The year in which someone immigrated from Korea to the United States may impact their technology acceptance.

Yoon et al [68] suggested that acculturation also plays a role in technology acceptance. In a cross-sectional survey analysis of Korean Americans aged ≥60 years, acculturation was a significant predictor of computer use and computer anxiety. Acculturation was defined as the degree to which a person from another culture learned the new language, customs, and behaviors of the host culture [68].

Testing the TAM

Overall, our regression analysis supported the TAM [20]. We found that PU and PEOU were positively, significantly associated with one another, and PU and PEOU were significantly, positively associated with our 2 outcome variables—smartphone use and ICT use. This remained true when accounting for age, gender, educational attainment, English proficiency, and ethnicity. Our findings align with subsequent TAM adaptations, such as the unified theory of acceptance and use of technology [21], TAM2 [69], TAM3 [70], and unified theory of acceptance and use of technology 2 [71], as well as studies conducted among older adults in China [71,72], Taiwan [73], and Korea [47,74,75].

In contrast, Chen and Chan [22] found that PU and PEOU were not significant predictors of technology use among a sample of Hong Kong Chinese adults aged ≥55 years. Instead, health and social factors were better predictors of technology use [22]. Because of their findings, Chen and Chan [22] proposed the senior technology acceptance model, which added health conditions, cognitive ability, physical function, and social relationships as independent variables predicting technology acceptance. While we did not account for health and social factors in this study, in future studies, we plan to explore how health and social factors impact PU, PEOU, smartphone use, and ICT use among Asian American older adults.

Limitations

Our study had limitations. We conducted a secondary analysis of data from the Lighthouse Project, using surveys that were administered before the design of this study. While we operationalized variables to match concepts from the original TAM, our measures differ from other TAM studies, making it difficult to compare findings. This is a common problem across research studies examining technology acceptance among older adults [26,44].

The measures used in this study were evidence based and translated from English to Korean, Chinese, and Vietnamese, although the translated items were not validated in all these languages. Vocabulary used and subsequent translation may have impacted data validity. In addition, study participants reported low educational attainment, and qualitatively, staff shared that some residents had low literacy.

Survey results were further limited by response bias and self-report. According to Deng et al [51], older adults tend to overestimate their smartphone use. Future studies seeking to understand ICT acceptance among older adults would benefit from collecting objective ICT use data.

This study engaged a convenience sample; our sample was not representative of Asians across California or the United States. Most of our participants were Korean, and the sample sizes representing Vietnamese and Filipino ethnicities were small. In addition, although surveys were collected across 6

XSL•FO
RenderX

communities, 70.7% (277/392) of our surveys were collected from residents living in 2 communities, which house primarily Korean residents and are both located in Koreatown, Los Angeles. Residents in these communities may experience less linguistic isolation than the monolingual Asian residents in the other Lighthouse communities. Finally, many other factors potentially impact ICT use among Asian American older adults, such as health status, access to Wi-Fi, availability of devices, digital literacy training, and technology support.

Conclusions

This study contributed to our insights of factors that influence ICT acceptance and use among community-dwelling Asian American older adults, specifically those of Korean, Chinese, Vietnamese, and Filipino ethnicities (N=392). By increasing the understanding of the associations among age, gender, educational attainment, English proficiency, ethnicity, attitudes toward technology, and ICT use, our findings can inform digital interventions aimed at addressing disparities among Asian American older adults. For example, our study suggests that older age, lower educational attainment, and LEP are negatively associated with smartphone and ICT use. To minimize disparities, interventions could be targeted toward Asian older adults with these characteristics.

Furthermore, PU and PEOU are potentially modifiable factors that can be addressed through group-based or one-on-one digital literacy training and support to new learners [76-79]. The technology industry could address access barriers by creating tailored user interfaces including voice user interfaces that minimize the burden for individuals with low literacy [80,81]. Our study found that female gender, LEP, and Korean ethnicity were each independently, negatively associated with PEOU, and limited educational attainment and Korean ethnicity were negatively associated with PU. Future research could explore how interventions influence attitudes such as PU or PEOU in this subpopulation to increase ICT acceptance and use.

There is a need for validated measures of ICT acceptance and use targeted toward older adults; these measures should be brief to minimize participant burden, continuously updated to keep pace with technology innovation, accessible for those with low literacy, and validated in multiple languages to allow greater understanding of technology acceptance among older adults with LEP.

Future studies should involve larger, more representative samples of older Asian Americans, including diverse ethnic groups. As highlighted, Asian Americans comprise >40 ethnicities, each with unique cultures, immigration backgrounds, and languages [11,12,16], and our findings demonstrate significant variations in technology acceptance and use across these groups.

The impact of English proficiency on ICT use warrants deeper exploration to identify strategies for enhancing equity and access for individuals with primary languages other than English. Further research is needed to understand the interplay of age, education, English proficiency, smartphone use, and ICT adoption, particularly considering the distinct challenges this combination poses for older adults using technology.

Interventional studies are needed to explore effective strategies to overcome resistance to technology, support novice users in adopting new devices, and broaden the diversity of technology use among comfortable users beyond basic functions. Qualitative and ethnographic studies could further elucidate the barriers and facilitators of technology adoption. Finally, future research should explore how other factors, such as health and social factors, influence ICT acceptance and use among Asian American older adults.

Acknowledgments

This work was funded by the Healthy Aging in a Digital World Initiative at the University of California, Davis, the Center for Information Technology Research in the Interest of Society (CITRIS), and the Banatao Institute. The Lighthouse surveys were designed by the Evaluation Unit of the Office of Research at the University of California, Davis, School of Medicine. This team included Rebeca Giacinto, Pauline DeLange Martinez, Stuart Henderson, and Julie Rainwater, who also oversaw the data collection and entry process. The authors would like to thank all Lighthouse participants for their time and effort to complete the evaluation surveys. The authors would also like to thank the staff at each site for their support with participant recruitment, data collection, and entry. Lighthouse Project leadership included the Front Porch Center for Innovation and Wellbeing (Davis Park, Kari Olson, and Jennifer Lee), Eskaton (Therese Ten Brinke and Sheri Peifer), and CITRIS (David Lindeman and Gale Berkowitz). Finally, we thank Lisa Miller and Ladson Hinton for contributing to the conceptualization of our research design.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

PDM led conceptual design, data analysis, and authorship. HMY made substantial conceptual and design contributions, providing invaluable input along all stages of manuscript preparation. DT provided oversight for operationalization of key variables, data analysis, reporting results, and interpretation of the data. MP and LG helped to critically revise the manuscript and agreed to be accountable for the accuracy of the work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Fang ML, Canham SL, Battersby L, Sixsmith J, Wada M, Sixsmith A. Exploring privilege in the digital divide: implications for theory, policy, and practice. Gerontologist 2019 Jan 09;59(1):e1-15. [doi: 10.1093/geront/gny037] [Medline: 29750241]
2. Chan DY, Lee SW, Teh PL. Factors influencing technology use among low-income older adults: a systematic review. Heliyon 2023 Sep;9(9):e20111 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e20111] [Medline: 37809586]
3. Anderson M, Perrin A. Tech adoption climbs among older adults. Pew Research Center. 2017. URL: https://www.pewresearch.org/internet/2017/05/17/tech-adoption-climbs-among-older-adults/ [accessed 2021-03-20]
4. Vogels EA. Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. 2021. URL: https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/ [accessed 2021-12-30]
5. Tran V. Asian American seniors are often left out of the national conversation on poverty. Urban Institute. 2017. URL: https://www.urban.org/urban-wire/asian-american-seniors-are-often-left-out-national-conversation-poverty [accessed 2022-01-21]
6. Yoon H, Jang Y, Vaughan PW, Garcia M. Older adults' internet use for health information: digital divide by race/ethnicity and socioeconomic status. J Appl Gerontol 2020 Jan;39(1):105-110. [doi: 10.1177/0733464818770772] [Medline: 29661052]
7. Horst BR, Sixsmith A, Simeonov D, Mihailidis A. Demographic and psychographic factors of social isolation during the COVID-19 pandemic: the importance of technology confidence. Front Public Health 2021;9:749515 [FREE Full text] [doi: 10.3389/fpubh.2021.749515] [Medline: 34778182]
8. Seifert A, Cotten SR, Xie B. A double burden of exclusion? Digital and social exclusion of older adults in times of COVID-19. J Gerontol B Psychol Sci Soc Sci 2021 Feb 17;76(3):e99-103 [FREE Full text] [doi: 10.1093/geronb/gbaa098] [Medline: 32672332]
9. Martins Van Jaarsveld G. The effects of COVID-19 among the elderly population: a case for closing the digital divide. Front Psychiatry 2020 Nov 12;11:577427 [FREE Full text] [doi: 10.3389/fpsyt.2020.577427] [Medline: 33304283]
10. Hsiao V, Chandereng T, Lankton RL, Huebner JA, Baltus JJ, Flood GE, et al. Disparities in telemedicine access: a cross-sectional study of a newly established infrastructure during the COVID-19 pandemic. Appl Clin Inform 2021 May;12(3):445-458 [FREE Full text] [doi: 10.1055/s-0041-1730026] [Medline: 34107542]
11. Blacher K. Asian Americans and Pacific Islanders in the United States aged 65 years and older: population, nativity, and language. National Asian Pacific Center on Aging. 2013. URL: https://www.napca.org/wp-content/uploads/2017/10/65-population-report-FINAL.pdf [accessed 2021-12-30]
12. Lopez G, Ruiz N, Patten E. Key facts about asian americans, a diverse and growing population. Pew Research Center. 2017. URL: https://www.pewresearch.org/fact-tank/2017/09/08/key-facts-about-asian-americans/ [accessed 2021-12-30]
13. 2018 profile of Asian Americans age 65 and over. Administration on Aging. 2019. URL: https://acl.gov/sites/default/files/Aging%20and%20Disability%20in%20America/2018AsA_OAProfile.pdf [accessed 2021-12-30]
14. Yoon H, Jang Y, Kim S, Speasmaker A, Nam I. Trends in internet use among older adults in the United States, 2011-2016. J Appl Gerontol 2021 May;40(5):466-470. [doi: 10.1177/0733464820908427] [Medline: 32131670]
15. Kim J, Lee HY. Older minority immigrants' digital technology use: does English proficiency matter? Gerontologist 2016;56(Issue Suppl_3):139 [FREE Full text] [doi: 10.1093/geront/gnw162.557]
16. Jang Y, Kim MT. Limited English proficiency and health service use in Asian Americans. J Immigr Minor Health 2019 Apr 24;21(2):264-270 [FREE Full text] [doi: 10.1007/s10903-018-0763-0] [Medline: 29797103]
17. Mui P, Bowie JV, Juon HS, Thorpe Jr RJ. Ethnic group differences in health outcomes among Asian American men in California. Am J Mens Health 2017 Sep;11(5):1406-1414 [FREE Full text] [doi: 10.1177/1557988316664508] [Medline: 27530820]
18. Gordon NP, Hornbrook MC. Differences in access to and preferences for using patient portals and other eHealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. J Med Internet Res 2016 Mar 04;18(3):e50 [FREE Full text] [doi: 10.2196/jmir.5105] [Medline: 26944212]
19. Ajzen I. From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J, editors. Action Control: From Cognition to Behavior. Cham, Switzerland: Springer; 1985:11-39.
20. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
21. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
22. Chen K, Chan AH. Gerontechnology acceptance by elderly Hong Kong Chinese: a senior technology acceptance model (STAM). Ergonomics 2014 Mar 24;57(5):635-652. [doi: 10.1080/00140139.2014.895855] [Medline: 24655221]
23. Berkowsky RW, Rikard RV, Cotten SR. Signing off: predicting discontinued ICT usage among older adults in assisted and independent living. In: Proceedings of the 2015 International Conference on Human Aspects of IT for the Aged Population

XSL•FO
RenderX

on Design for Everyday Life. 2015 Presented at: ITAP '15; August 2-7, 2015; Los Angeles, CA p. 389-398 URL: https://link.springer.com/chapter/10.1007/978-3-319-20913-5_36 [doi: 10.1007/978-3-319-20913-5_36]
24. Cotten SR. Examining the roles of technology in aging and quality of life. J Gerontol B Psychol Sci Soc Sci 2017 Sep 01;72(5):823-826. [doi: 10.1093/geronb/gbx109] [Medline: 28817928]
25. Rikard RV, Berkowsky RW, Cotten SR. Discontinued information and communication technology usage among older adults in continuing care retirement communities in the United States. Gerontology 2018;64(2):188-200 [FREE Full text] [doi: 10.1159/000482017] [Medline: 29130976]
26. Cotten SR. Technologies and aging: understanding use, impacts, and future needs. In: Ferraro KF, Carr D, editors. Handbook of Aging and the Social Sciences: A volume in Handbooks of Aging. 9th edition. New York, NY: Academic Press; 2021:373-392.
27. Hargittai E, Piper AM, Morris MR. From internet access to internet skills: digital inequality among older adults. Univ Access Inf Soc 2018 May 3;18(4):881-890. [doi: 10.1007/S10209-018-0617-5]
28. Hargittai E, Dobransky K. Old dogs, new clicks: digital inequality in skills and uses among older adults. Can J Commun 2017 May 23;42(2):5. [doi: 10.22230/cjc.2017v42n2a3176]
29. König R, Seifert A, Doh M. Internet use among older Europeans: an analysis based on SHARE data. Univ Access Inf Soc 2018 Jan 19;17(3):621-633. [doi: 10.1007/S10209-018-0609-5]
30. Selwyn N, Gorard S, Furlong J, Madden L. Older adults' use of information and communications technology in everyday life. Ageing Soc 2003 Sep 02;23(5):561-582. [doi: 10.1017/S0144686X03001302]
31. van Deursen AJ, Helsper EJ. A nuanced understanding of Internet use and non-use among the elderly. Eur J Commun 2015 Apr 09;30(2):171-187. [doi: 10.1177/0267323115578059]
32. Friemel TN. The digital divide has grown old: determinants of a digital divide among seniors. New Media Soc 2014 Jun 12;18(2):313-331. [doi: 10.1177/1461444814538648]
33. Yu RP, Ellison NB, McCammon RJ, Langa KM. Mapping the two levels of digital divide: internet access and social network site adoption among older adults in the USA. Inf Commun Soc 2015 Nov 19;19(10):1445-1464. [doi: 10.1080/1369118x.2015.1109695]
34. Choi NG, Dinitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. J Med Internet Res 2013 May;15(5):e97 [FREE Full text] [doi: 10.2196/jmir.2333] [Medline: 23681083]
35. Lighthouse for older adults: a CITRIS and University of California initiative, brings technology-enabled health and well-being to low-income California seniors during COVID crisis. CITRIS and the Banatao Institute. URL: https://citris-uc.org/lighthouse-for-older-adults/ [accessed 2020-09-16]
36. Sims T, Reed AE, Carr DC. Information and communication technology use is related to higher well-being among the oldest-old. J Gerontol B Psychol Sci Soc Sci 2017 Sep 01;72(5):761-770. [doi: 10.1093/geronb/gbw130] [Medline: 27702839]
37. Jay GM, Willis SL. Influence of direct computer experience on older adults' attitudes toward computers. J Gerontol 1992 Jul;47(4):P250-P257. [doi: 10.1093/geronj/47.4.p250] [Medline: 1624702]
38. IBM SPSS statistics software: hierarchical linear regression. IBM Corp. URL: https://methods.sagepub.com/dataset/howtoguide/hierarchical-linear-regression-prison-inmates#:~:text=A%20hierarchical%20linear%20regression%20is,improves%20a%20model's%20ability%20to [accessed 2022-05-24]
39. Masci D. About one-fifth of adults globally have no formal schooling. Pew Research Center. 2017. URL: https://www.pewresearch.org/fact-tank/2017/01/11/about-one-fifth-of-adults-globally-have-no-formal-schooling/ [accessed 2023-04-13]
40. Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate Data Analysis. 4th edition. Upper Saddle River, NJ: Prentice Hall; 2010.
41. Faverio M. Share of those 65 and older who are tech users has grown in the past decade. Pew Research Center. 2022. URL: https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/ [accessed 2023-05-01]
42. Anderson M, Kumar M. Digital divide persists even as lower-income Americans make gains in tech adoption. Per Research Center. 2019. URL: https://www.pewresearch.org/fact-tank/2019/05/07/digital-divide-persists-even-as-lower-income-americans-make-gains-in-tech-adoption/ [accessed 2021-06-21]
43. LoBuono DL, Leedahl SN, Maiocco E. Teaching technology to older adults: modalities used by student mentors and reasons for continued program participation. J Gerontol Nurs 2020 Jan 01;46(1):14-20. [doi: 10.3928/00989134-20191118-02] [Medline: 31895957]
44. Yap YY, Tan SH, Choon SW. Elderly's intention to use technologies: a systematic literature review. Heliyon 2022 Jan;8(1):e08765 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e08765] [Medline: 35128090]
45. Vroman KG, Arthanat S, Lysack C. “Who over 65 is online?”: older adults’ dispositions toward information communication technology. Comput Human Behav 2015 Feb;43:156-166. [doi: 10.1016/j.chb.2014.10.018]
46. Ma Q, Chan AH, Chen K. Personal and other factors affecting acceptance of smartphone technology by older Chinese adults. Appl Ergon 2016 May;54:62-71. [doi: 10.1016/j.apergo.2015.11.015] [Medline: 26851465]

XSL•FO
RenderX

47. Ha J, Park HK. Factors affecting the acceptability of technology in health care among older Korean adults with multiple chronic conditions: a cross-sectional study adopting the senior technology acceptance model. Clin Interv Aging 2020;15:1873-1881 [FREE Full text] [doi: 10.2147/CIA.S268606] [Medline: 33061336]
48. Mitzner TL, Rogers WA, Fisk AD, Boot WR, Charness N, Czaja SJ, et al. Predicting older adults' perceptions about a computer system designed for seniors. Univers Access Inf Soc 2016 Jun 7;15(2):271-280 [FREE Full text] [doi: 10.1007/s10209-014-0383-y] [Medline: 31186624]
49. Porter CE, Donthu N. Using the technology acceptance model to explain how attitudes determine internet usage: the role of perceived access barriers and demographics. J Bus Res 2006 Sep;59(9):999-1007. [doi: 10.1016/j.jbusres.2006.06.003]
50. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
51. Deng T, Kanthawala S, Meng J, Peng W, Kononova A, Hao Q, et al. Measuring smartphone usage and task switching with log tracking and self-reports. Mob Media Commun 2018 Apr 25;7(1):3-23. [doi: 10.1177/2050157918761491]
52. Macedo IM. Predicting the acceptance and use of information and communication technology by older adults: an empirical examination of the revised UTAUT2. Comput Human Behav 2017 Oct;75:935-948. [doi: 10.1016/j.chb.2017.06.013]
53. Goswami A, Dutta S. Gender differences in technology usage—a literature review. Open J Bus Manag 2016;04(01):51-59. [doi: 10.4236/ojbm.2016.41006]
54. Kim J, Lee HY, Christensen MC, Merighi JR. Technology access and use, and their associations with social engagement among older adults: do women and men differ? J Gerontol B Psychol Sci Soc Sci 2017 Sep 01;72(5):836-845. [doi: 10.1093/geronb/gbw123] [Medline: 28073816]
55. Metallo C, Agrifoglio R, Lepore L, Landriani L. Explaing users' technology acceptance through national cultural values in the hospital context. BMC Health Serv Res 2022 Jan 17;22(1):84 [FREE Full text] [doi: 10.1186/s12913-022-07488-3] [Medline: 35039014]
56. Srite M. Culture as an explanation of technology acceptance differences: an empirical investigation of Chinese and US users. Australas J Inf Syst 2006 Nov 01;14(1):5-26. [doi: 10.3127/ajis.v14i1.4]
57. Hofstede G. The cultural relativity of organizational practices and theories. J Int Bus Stud 1983 Jun 1;14(2):75-89. [doi: 10.1057/palgrave.jibs.8490867]
58. Hofstede G, Bond MH. The Confucius connection: from cultural roots to economic growth. Organ Dyn 1988 Mar;16(4):5-21. [doi: 10.1016/0090-2616(88)90009-5]
59. Young H. The digital language divide: how does the language you speak shape your experience of the internet? The Guardian. 2013. URL: http://labs.theguardian.com/digital-language-divide/ [accessed 2022-03-18]
60. Internet world users by language: top 10 languages. Miniwatts Marketing Group. URL: https://www.internetworldstats.com/stats7.htm [accessed 2022-03-18]
61. Expanding your apps to new markets. Apple Developer. 2021. URL: https://developer.apple.com/localization/ [accessed 2022-03-18]
62. Kinsella B. Google assistant now supports Cantonese on Androidwithin actions on Google SDK. Voicebot.ai. 2019. URL: https://voicebot.ai/2019/10/23/google-assistant-now-support-cantonese-on-android-and-within-actions-on-google-sdk/ [accessed 2022-03-18]
63. Pyae A, Scifleet P. Investigating the role of user's English language proficiency in using a voice user interface: a case of Google Home smart speaker. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI EA '19; May 4-9, 2019; Scotland, UK p. 1-6 URL: https://dl.acm.org/doi/10.1145/3290607.3313038 [doi: 10.1145/3290607.3313038]
64. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
65. Sarkar U, Gourley GI, Lyles CR, Tieu L, Clarity C, Newmark L, et al. Usability of commercially available mobile applications for diverse patients. J Gen Intern Med 2016 Dec;31(12):1417-1426. [doi: 10.1007/s11606-016-3771-6] [Medline: 27418347]
66. Barros AC, Leitão R, Ribeiro J. Design and evaluation of a mobile user interface for older adults: navigation, interaction and visual design recommendations. Procedia Comput Sci 2014;27:369-378. [doi: 10.1016/j.procs.2014.02.041]
67. Lee OE, Kim DH, Beum KA. Factors affecting information and communication technology use and eHealth literacy among older adults in the US and South Korea. Educ Gerontol 2020 Jul 09;46(9):575-586. [doi: 10.1080/03601277.2020.1790162]
68. Yoon H, Jang Y, Xie B. Computer use and computer anxiety in older Korean Americans. J Appl Gerontol 2016 Sep 09;35(9):1000-1010. [doi: 10.1177/0733464815570665] [Medline: 25698717]
69. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
70. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decis Sci 2008 May 09;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
71. Zhao S, Kinshuk, Yao Y, Ya N. Adoption of mobile social media for learning among Chinese older adults in senior citizen colleges. Educ Technol Res Dev 2021 Sep 27;69(6):3413-3435 [FREE Full text] [doi: 10.1007/s11423-021-10048-x] [Medline: 34602799]

XSL•FO
RenderX

72. Pan S, Jordan-Marsh M. Internet use intention and adoption among Chinese older adults: from the expanded technology acceptance model perspective. Comput Human Behav 2010 Sep;26(5):1111-1119. [doi: 10.1016/j.chb.2010.03.015]
73. Liu CH, Wu ZY, Lin CP. A study of intention and behavior of using mobile communication software: the case of seniors. In: Proceedings of the 2018 IEEE International Conference on Applied System Invention. 2018 Presented at: ICASI '18; April 13-17, 2018; Chiba, Japan p. 13-17 URL: https://ieeexplore.ieee.org/document/8394443 [doi: 10.1109/icasi.2018.8394443]
74. Han S, Nam SI. Creating supportive environments and enhancing personal perception to bridge the digital divide among older adults. Educ Gerontol 2021 Oct 17;47(8):339-352. [doi: 10.1080/03601277.2021.1988448]
75. Shin HR, Um SR, Yoon HJ, Choi EY, Shin WC, Lee HY, et al. Comprehensive senior technology acceptance model of daily living assistive technology for older adults with frailty: cross-sectional study. J Med Internet Res 2023 Apr 10;25:e41935 [FREE Full text] [doi: 10.2196/41935] [Medline: 37036760]
76. Arthanat S. Promoting information communication technology adoption and acceptance for aging-in-place: a randomized controlled trial. J Appl Gerontol 2021 May 29;40(5):471-480 [FREE Full text] [doi: 10.1177/0733464819891045] [Medline: 31782347]
77. Lagana L, Garcia JJ. The mental health impact of computer and internet training on a multi-ethnic sample of community-dwelling older adults: results of a pilot randomised controlled trial. Int J Biomed Sci 2013 Sep;9(3):135-147 [FREE Full text] [Medline: 24151452]
78. González A, Ramírez MP, Viadel V. ICT learning by older adults and their attitudes toward computer use. Curr Gerontol Geriatr Res 2015;2015:849308 [FREE Full text] [doi: 10.1155/2015/849308] [Medline: 26346158]
79. Senteio CR. Investigating the enduring impact of a community-based health education program to promote African American elders' use of technology designed to support chronic disease self-management. Geriatrics (Basel) 2018 Oct 13;3(4):70 [FREE Full text] [doi: 10.3390/geriatrics3040070] [Medline: 31011105]
80. Czaja SJ, Boot WR, Charness N, Rogers WA, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018 May 08;58(3):467-477 [FREE Full text] [doi: 10.1093/geront/gnw249] [Medline: 28201730]
81. Lor M, Bowers BJ. Feasibility of audio-computer-assisted self-interviewing with color-coding and helper assistance (ACASI-H) for Hmong older adults. Res Nurs Health 2017 Aug 21;40(4):360-371 [FREE Full text] [doi: 10.1002/nur.21796] [Medline: 28431187]

Abbreviations

CHIS: California Health Interview Survey
ICT: information and communications technology
LEP: limited English proficiency
PEOU: perceived ease of use
PU: perceived usefulness
TAM: technology acceptance model
TPB: theory of planned behavior

Edited by T de Azevedo Cardoso; submitted 05.09.23; peer-reviewed by F Regalado; comments to author 29.02.24; revised version received 09.04.24; accepted 17.09.24; published 22.11.24.

Please cite as:
DeLange Martinez P, Tancredi D, Pavel M, Garcia L, Young HM
Technology Acceptance Among Low-Income Asian American Older Adults: Cross-Sectional Survey Analysis
J Med Internet Res 2024;26:e52498
URL: https://www.jmir.org/2024/1/e52498
doi:10.2196/52498
PMID:

©Pauline DeLange Martinez, Daniel Tancredi, Misha Pavel, Lorena Garcia, Heather M Young. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Relationship Between Internet Use and Cognitive Function Among Middle-Aged and Older Chinese Adults: 5-Year Longitudinal Study

Bowen Chen[1], MD; Chun Yang[2], PhD; Shanshan Ren[3], MD; Penggao Li[2*], PhD; Jin Zhao[1*], MD

[1]Department of Hospital Epidemiology and Infection Control, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[2]School of Public Health, Capital Medical University, Beijing, China
[3]Department of Clinical Nutrition, Beijing Hospital, Chinese Academy of Medical Sciences, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Jin Zhao, MD
Department of Hospital Epidemiology and Infection Control
Sir Run Run Shaw Hospital
Zhejiang University School of Medicine
3 East Qingchun Road
Hangzhou, 310016
China
Phone: 86 571 86006435
Fax: 86 571 86006435
Email: enqizhao@zju.edu.cn

Abstract

Background: Cognitive decline poses one of the greatest global challenges for health and social care, particularly in China, where the burden on the older adult population is most pronounced. Despite the rapid expansion of internet access, there is still limited understanding of the long-term cognitive impacts of internet use among middle-aged and older adults.

Objective: This study aims to explore the association between internet use and age-related cognitive decline among middle-aged and older Chinese adults. To gain a more comprehensive understanding of the effects of internet use, we also focused on assessing the impact of both the frequency of internet use and the types of internet devices on cognition. Moreover, we assessed the mediating role of internet use on cognitive function for characteristics significantly linked to cognition in stratified analysis.

Methods: We analyzed data based on 12,770 dementia-free participants aged ≥45 years from the China Health and Retirement Longitudinal Study. We used a fixed effects model to assess the relationship between internet use and cognitive decline and further validated it using multiple linear regression analysis, generalized estimating equations, propensity score matching, inverse probability of treatment weighting, and overlap weighting. We further examined the varying effects of internet device type and frequency on cognitive function using fixed effects models and Spearman rank correlations. The Karlson-Holm-Breen method was used to estimate the mediating role of internet use in the urban-rural cognitive gap.

Results: Participants using the internet (n=1005) were younger, more likely to be male, more educated, married, retired and living in an urban area and had higher cognitive assessment scores than nonusers (n=11,765). After adjusting for demographic and health-related risk factors, there was a positive correlation between internet use and cognitive function (β=0.551, 95% CI 0.391-0.710). Over the follow-up period, persistent internet users had a markedly lower 5-year incidence of neurodegenerative diseases, at 2.2% (15/671), compared with nonusers, at 5.3% (379/7099; $P<.001$). The negative impact of aging (>50 years) on cognitive function was consistently less pronounced among internet users than among nonusers. Furthermore, increased frequency of internet use was associated with greater cognitive benefits for middle-aged and older adults ($r_s=0.378$, $P<.001$). Among digital devices used for internet access, cell phones (β=0.398, 95% CI 0.283-0.495) seemed to have a higher level of cognitive protection than computers (β=0.147, 95% CI 0.091-0.204). The urban-rural disparity in cognitive function was partially attributed to the disparity in internet use (34.2% of the total effect, $P<.001$).

Conclusions: This study revealed that the use of internet by individuals aged 45 years and older is associated with a reduced risk of cognitive decline. Internet use has the potential to be a viable, cost-effective, nonpharmacological intervention for cognitive decline among middle-aged and older adults.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57301) doi:10.2196/57301

KEYWORDS

aging; cognitive function; internet use; longitudinal study; fixed effects model

Introduction

With the rapid increase in the aging population and the extension of life expectancy, age-related cognitive impairment has emerged as an influential global public health concern [1]. Cognitive decline is an irreversible cerebral pathological and physiological progression, serving as both a precursor to dementia and a potential catalyst for physical disability and mortality [2]. A nationwide survey conducted in 2019 revealed that China has become the country with the highest number of individuals with dementia, as the population aged 60 years and older has soared to a staggering 253 million people, among which 6.0% have dementia [3,4]. Cognitive impairment typically emerges in individuals aged 45 years to 60 years and not only imposes a significant burden on the patients' families but also presents formidable challenges to the health care system and society [5]. Hence, the discovery of nonpharmacological strategies to intervene for middle-aged and older adults is crucial to reducing the incidence of cognitive decline, slowing the progression of cognitive impairment, and improving health and well-being.

Regarding the prevention of dementia, establishing a healthy lifestyle and implementing nonpharmacological interventions are critically important. Based on randomized clinical trials and meta-analyses, consistent research-based evidence indicates that social and cognitively stimulating activities can delay cognitive decline, particularly in middle-aged and older adults who are most susceptible to cognitive impairment [6-8].

With the rapid development of the internet and the widespread use of smartphones, digital literacy is increasing in China, particularly among middle-aged and older adults. According to the 52nd Statistical Report on Internet Development in China, released by the China Internet Network Information Center [9], the number of internet users aged 60 years and older reached a staggering 140 million as of June 2023. It signifies a profound penetration of internet use within the older adult demographic, bringing about substantial transformations in daily life. Preliminary analysis of the English Longitudinal Study of Aging indicated that individuals who frequently use the internet or email exhibit a 3.1% improvement in delayed recall compared with those who do not use the internet or email [10]. An Australian study revealed that men aged 70 years and older who use the internet had a 50% relative risk reduction for dementia, even after controlling for age, education, social relationships, and physical health status [11]. The study by Klimova [12] suggested that the use of the internet, specifically online cognitive training programs, may exert a beneficial effect on enhancing the cognitive function of healthy older adults.

The mechanisms through which internet use enhances cognitive abilities are highly complex, and current research has not fully elucidated them yet. A cross-sectional study showed that internet searches can enhance the activation of neural circuits in the frontal pole, anterior temporal region, anterior and posterior cingulate, and hippocampus areas among people older than 55 years [13]. Numerous previous studies have indicated that the internet can provide people with more social contacts, social support, and positive attitudes toward life, enhancing social connections while reducing loneliness and social isolation, which can have a positive impact on cognitive function [14-16]. The study by Firth et al [17] suggested that the use of the internet may influence cognitive function through 3 key factors: attentional capacity, memory processes, and social cognition.

However, not all research findings consistently demonstrate the cognitive benefits of internet use. A randomized controlled trial conducted in the Netherlands with a small sample of individuals aged 64 years to 75 years revealed that 12 months of computer and internet use did not have a significant impact on the cognitive function of older adults [18]. Overall, despite the common perception that internet use can improve cognitive function in older adults, the current literature provides varying evidence, necessitating studies with larger sample sizes to elucidate the correlation.

The China Health and Retirement Longitudinal Study (CHARLS) is the first nationally representative survey of middle-aged and older adults in China, with a sample collected from 28 provinces and 150 regions across the country [19]. This project aims to construct a high-quality public microdatabase collecting multidimensional information encompassing socioeconomic status and health conditions to meet the needs of gerontological research. Therefore, the CHARLS questionnaire includes several aspects such as demographic background, family structure, health status, health insurance, work, retirement, pensions, income, expenditures, assets, and house property. To ensure randomization and the representativeness of the sample, CHARLS adopts multistage stratified probability proportional to size sampling. Meanwhile, CHARLS conducts face-to-face interviews facilitated by well-trained surveyors and implements effective quality control through methods such as data verification, audio playback, and the use of a computer-assisted personal interview system.

Our primary aim was to assess the independent association between internet use and cognitive function in middle-aged and older adults by collecting survey data from the CHARLS study. To gain deeper insights into the cognitive effects of internet use, we also evaluated the age-related impacts on cognitive function in internet users and nonusers as one of our secondary objectives. Additionally, to explore the most effective means for enhancing cognitive function, we analyzed the correlations between cognitive function and both internet device use and frequency. We further intended to evaluate the mediating role of internet use on cognitive function for participant characteristics for which a significant association between internet use and cognitive function was observed in stratified analysis.

XSL•FO
RenderX

Methods

Participants

Participants aged 45 years and older were enrolled in the CHARLS study and included in subsequent secondary analyses, with their socioeconomic and health status being systematically recorded. The CHARLS survey was initiated in 2011 (Wave 1), followed by 4 consecutive waves of regular follow-up questionnaires conducted in 2013 (Wave 2), 2015 (Wave 3), 2018 (Wave 4), and 2020 (Wave 5). More comprehensive information on the study design and sampling strategies used in CHARLS has been previously described [19]. Given the insufficient number of internet users in Waves 1 and 2, we conducted data collection for the subsequent waves, including Wave 3, Wave 4, and Wave 5.

Measurements

Internet Use

The primary exposure and mediating variable was internet use, which was assessed based on a binary response (yes/no) to the question regarding internet access. With regard to the frequency of internet use, participants' questionnaire responses were classified into 4 categories: never, infrequently, weekly, and daily. Additionally, the devices used for internet access were categorized into 2 groups: computers (including desktop computers, laptop computers, and tablet computers) and cell phones. The specific wording of the questionnaire survey on internet use is detailed in Multimedia Appendix 1.

Cognitive Function

We used 2 established composite measures, mental intactness and episodic memory, to evaluate global cognitive function. These measures are similar to the assessment indicators used by the Health and Retirement Study in the United States [20] and have been validated among the Chinese population [21]. Mental intactness is determined based on 3 indicators of cognitive status, which comprise time orientation (naming month, day, year, week, and season), numerical ability (subtracting 7 consecutively from 100, 5 times), and visual and spatial abilities (the task of redrawing 2 overlapping pentagons). The mental intactness score for the 3 elements ranges from 0 to 11. Episodic memory can be evaluated through the recall of commonly used phrases, including immediate word memory (score range 0-10) and delayed word memory (score range 0-10, with a total score of 20). Cognitive function is constructed from these 2 components (Multimedia Appendix 1), with higher scores suggesting better cognitive function. Furthermore, to gain a more comprehensive understanding of cognitive benefits, we analyzed the incidence of neurodegenerative diseases, such as Alzheimer disease and Parkinson disease, among both internet users and nonusers during the follow-up period.

Covariates

This study accounted for multiple potential confounding factors that are known risk factors for cognitive function [22], in order to provide more compelling evidence. In our study, we took into account the following covariates: baseline or prior cognitive score (continuous variable), gender (male or female), age (continuous variable), marital status (yes [married or partnered] or no [unmarried, separated, divorced, widowed]), residency (urban or rural), retirement status (yes or no), education level (illiterate, low education level [primary or junior high school], mid education level [senior high school], or high education level [college or above]), smoking status (current smoker, former smoker, never smoked), alcohol consumption (never drank, drink less than once a month, drink more than once a month), diabetes (yes or no), hypertension (yes or no), and per capita household expenditure (quartiles).

Statistical Analysis

To comprehensively assess the relationship between internet use and cognitive function, we used both longitudinal and cross-sectional analytical approaches, thereby aiming to complement the limitations inherent in each research method. We used fixed effects analysis to analyze longitudinal data with time-varying variables and performed multiple linear regression analyses on cross-sectional data from each wave. Furthermore, we used the restricted cubic spline regression method to investigate the differences in cognitive performance between internet users and nonusers across various ages. We used the Spearman rank correlation test to analyze the correlation between internet use frequency and cognitive function. We regarded baseline characteristics as covariates and conducted stratified and interaction analyses to assess the potential differences in the impact of internet use among different subgroups. After stratifying by baseline characteristics, we constructed a fixed effects model and incorporated the cross-product terms (internet use × baseline characteristics) within the same model to evaluate the interaction between internet use and baseline characteristics.

The mediation analysis was performed using the Karlson-Holm-Breen (KHB) method [23], which dissects the influence of urban-rural areas on cognitive functionality into its distinct direct and indirect constituents. In particular, it measures the extent to which the intermediate variable (internet use) mediates the association between the independent variable (urban-rural status) and the dependent variable (cognitive function).

In sensitivity analysis, we conducted propensity score matching (PSM) [24] with a 1:1 ratio to adjust for covariates and analyze the impact of internet use on cognitive function. Second, we also used inverse probability of treatment weighting (IPTW) [25] and overlap weighting (OW) [26] methods to weight the balance based on covariates, aiming to validate our main results. To obtain robust results and further enhance the sensitivity analysis, we additionally selected individuals who had been continuous internet users and continuous nonusers during a 5-year follow-up period to investigate their impact on cognitive function using generalized estimating equations (GEE) and multiple linear regression analysis. Continuous variables are shown as means and SDs, while categorical variables are represented as numbers (proportions). All analyses were performed using R (version 4.0.2) and Stata (version 14.1).

XSL•FO
RenderX

Ethical Considerations

The Biomedical Ethics Review Committee of Peking University approved the ethical application (IRB00001052-11014 and IRB00001052-11015) for collecting data from human participants in all waves, as well as for secondary analysis of CHARLS data. Written informed consent was obtained from all participants prior to the questionnaire survey, and the original informed consent allows secondary analysis without additional consent. All respondents who completed the questionnaire received cash as compensation according to the unified standard amount set by the CHARLS in advance. In order to protect the privacy of the respondents, all the data in this study are anonymous.

Results

Descriptive Statistics

A total of 12,770 Chinese adults aged 45 years and older were ultimately included in this study, with the process of selection and exclusion detailed in Figure 1. In addition, Table 1 presents the baseline (ie, Wave 3) demographic characteristics of the study participants. At baseline, internet users accounted for 7.9% (1005/12,770) of the participants, and nonusers accounted for 92.1% (11,765/12,770) of the participants. Participants engaging in internet use had a significantly lower mean age of 53.04 (SD 7.33) years in contrast to the 59.32 (SD 9.16) years among nonusers (P<.001). More men were internet users (583/1005, 58%) than nonusers (5507/11,765, 46.8%). In terms of educational attainment, the internet-using population had a markedly lower illiteracy rate of 0.6% (6/1005) than the nonuser population (2269/11,765, 19.3%). This trend was consistent across educational strata, with 11.2% (113/1005), 71.9% (722/1005), and 16.3% (164/1005) of internet users classified as having low, mid, and high levels of education, respectively, as opposed to 44.1% (5189/11,765), 35.3% (4159/11,765), and 1.3% (148/11,765), respectively, of the nonuser population. Marital status among internet users was predominantly characterized by marriage, with 93.9% (944/1005) of this group being married, which was a higher proportion than the 90.3% (10,623/11,765) of nonusers who were married (P<.001). The proportion of retired individuals was also higher within the internet-using group, at 21% (211/1005) compared with 9.1% (1072/11,765) among nonusers (P<.001). Residence in rural areas was less common among internet users, with 46.9% (471/1005) living in such regions, a significantly lower proportion than the 83.5% (9825/11,765) of nonusers (P<.001). Notably, in terms of cognitive function, the cognitive assessment scores of internet users across different age groups (45-54 years, 55-64 years, 65-74 years, and 75 years) were significantly higher than those of nonusers (P<.001).

XSL•FO
RenderX

Figure 1. Flowchart of participant selection.

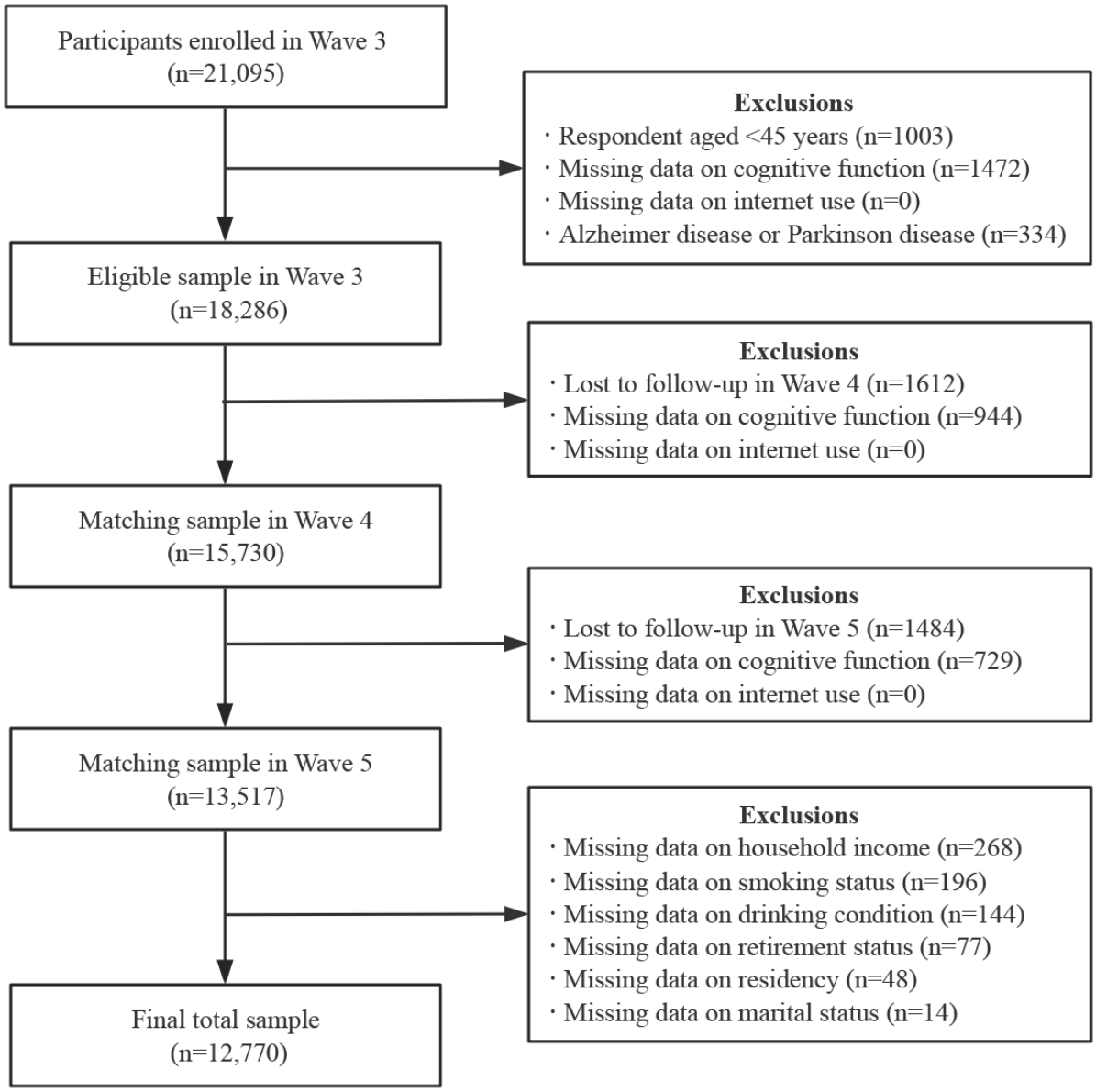

XSL•FO
RenderX

Table 1. Baseline characteristics of the study population.

Characteristic	Nonusers (n=11,765)	Internet users (n=1005)	*P* value
Gender , n (%)			<.001
Male	5507 (46.8)	583 (58)	
Female	6258 (53.2)	422 (42)	
Age (years), mean (SD)	59.32 (9.16)	53.04 (7.33)	<.001
Age group (years), n (%)			<.001
45-54	4384 (37.3)	692 (68.8)	
55-64	3917 (33.3)	219 (21.8)	
65-74	2754 (23.4)	84 (8.4)	
≥75	710 (6)	10 (1)	
Household income per capita[a], n (%)			<.001
Quartile 1	2779 (23.6)	137 (13.6)	
Quartile 2	2877 (24.4)	146 (14.5)	
Quartile 3	3091 (26.3)	250 (24.9)	
Quartile 4	3018 (25.7)	472 (47)	
Marital status, n (%)			<.001
Yes	10,623 (90.3)	944 (93.9)	
No	1142 (9.7)	61 (6.1)	
Residency, n (%)			<.001
Rural	9825 (83.5)	471 (46.9)	
Urban	1940 (16.5)	534 (53.1)	
Educational attainment, n (%)			.001
Illiterate	2269 (19.3)	6 (0.6)	
Low education	5189 (44.1)	113 (11.2)	
Mid education	4159 (35.3)	722 (71.9)	
High education	148 (1.3)	164 (16.3)	
Retirement status, n (%)			<.001
No	10,693 (90.9)	794 (79)	
Yes	1072 (9.1)	211 (21)	
Smoking status, n (%)			<.001
Never smoked	4533 (38.5)	294 (29.3)	
Quit	2234 (19)	195 (19.4)	
Still smoke	4998 (42.5)	516 (51.3)	
Drinking status, n (%)			<.001
Never drank	7579 (64.4)	432 (43)	
Drink less than once a month	996 (8.5)	195 (19.4)	
Drink more than once a month	3190 (27.1)	378 (37.6)	
Hypertension, n (%)			.045
No	8353 (71)	744 (74)	
Yes	3412 (29)	261 (26)	
Dyslipidemia, n (%)			.12
No	8526 (72.5)	752 (74.8)	
Yes	3239 (27.5)	253 (25.2)	

XSL•FO
RenderX

Characteristic	Nonusers (n=11,765)	Internet users (n=1005)	*P* value
Diabetes, n (%)			.004
No	8387 (71.3)	760 (75.6)	
Yes	3378 (28.7)	245 (24.4)	
Cognitive assessment score by age (years), mean (SD)			
45-54	16.00 (4.85)	19.99 (3.61)	<.001
55-64	14.50 (4.98)	19.69 (3.30)	<.001
65-74	13.49 (5.20)	19.24 (3.67)	<.001
≥75	11.46 (5.25)	18.00 (4.71)	<.001

[a]Divided into 4 quartiles based on the 25th percentile (¥1643.60 [US $226.89]), 50th percentile (¥4497.50 [US $620.86]), and 75th percentile (¥10,959.80 [US $1512.95]) of the overall data.

Association Between Internet Use and Cognitive Function

This study demonstrated a positive association between internet use and cognitive function among middle-aged and older adults (Figure 2) based on the multiple linear regression analysis to evaluate cross-sectional data from Wave 3, Wave 4, and Wave 5 and the fixed effects model to analyze the longitudinal data across the entire population. Model 1 was adjusted for baseline or prior cognitive score; Model 2 was additionally adjusted for age, gender, marital status, educational attainment, residency, and retirement status; Model 3 was additionally adjusted for smoking, drinking, hypertension, and diabetes; and Model 4 was additionally adjusted for household income per capita. There were significant differences between internet users and nonusers in both mental intactness score and episodic memory score (*P*<.001; Multimedia Appendix 2). In the fixed effects model adjusting only for prior cognitive scores, internet use was significantly associated with cognitive function (model 1: β=0.573, 95% CI 0.414 to 0.732). The relationship remained statistically significant for model 2 (β=0.583, 95% CI 0.424 to 0.743) and model 3 (β=0.555, 95% CI 0.395 to 0.714). In the fully adjusted model (model 4), which additionally controlled for household income per capita, internet use remained positively associated with cognitive function (β=0.551, 95% CI 0.391 to 0.710). To further confirm the cognitive benefits of internet use, we conducted a multiple linear regression analysis on the waves of each cross-sectional data. Even after adjusting for all covariates, the results remained consistent with our expectations (Wave 3: β=1.592, 95% CI 1.289 to 1.895; Wave 4: β=0.956, 95% CI 0.720 to 1.193; Wave 5: β=1.154, 95% CI 0.907 to 1.402).

To further investigate the association between internet use and specific clinical manifestations of cognitive function, we analyzed the incidence of neurodegenerative diseases (Alzheimer disease or Parkinson disease) among internet users and nonusers (Multimedia Appendix 3). Notably, the 5-year incidence of neurodegenerative diseases was significantly lower in persistent internet users (15/671, 2.2%) than in nonusers (379/7099, 5.3%; *P*<.001), as detailed in Multimedia Appendix 3. This implies a potential association between internet use and the prevention of neurodegenerative diseases.

Figure 2. Associations between cognitive function and internet use. *Crude model.

	Wave 3 (2015)		Wave 4 (2018)		Wave 5 (2020)		Total	
	ß (95% CI)	*P*	ß (95% CI)	*P*	ß (95% CI)	*P*	ß (95% CI)	*P*
Model 1	5.200 (4.874−5.525)*	<.001	2.367 (2.122−2.613)	<.001	2.232 (2.009−2.456)	<.001	0.573 (0.414−0.732)	<.001
Model 2	1.647 (1.346−1.949)	<.001	0.935 (0.699−1.171)	<.001	1.184 (0.936−1.431)	<.001	0.583 (0.424−0.743)	<.001
Model 3	1.614 (1.313−1.916)	<.001	0.942 (0.705−1.178)	<.001	1.150 (0.902−1.397)	<.001	0.555 (0.395−0.714)	<.001
Model 4	1.592 (1.289−1.895)	<.001	0.956 (0.720−1.193)	<.001	1.154 (0.907−1.402)	<.001	0.551 (0.391−0.710)	<.001

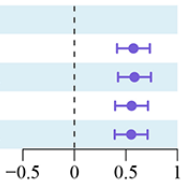

Impact of Internet Use on Cognitive Function Across Age

In order to further investigate the effects of internet use on cognitive function across different ages, we used restricted cubic splines to fit the curves that influence cognitive function (Figure 3). Our results indicated that age becomes a risk factor for cognitive function in internet nonusers after the age of 50.02 years, whereas for internet users, it occurs after the age of 49.02 years. Interestingly, we observed that the impact of age on cognitive decline was consistently minor among internet users compared with among nonusers after the age of 50 years. In other words, internet use has the potential to slow age-related cognitive decline. The results also showed that the older the age, the larger the gap in the effect of age on cognitive function between internet users and nonusers. This implies that the protective effect of internet use on cognitive decline is increasingly evident with age.

XSL•FO
RenderX

Figure 3. The influence of internet use on cognitive function across different ages using restricted cubic splines with 3 knots.

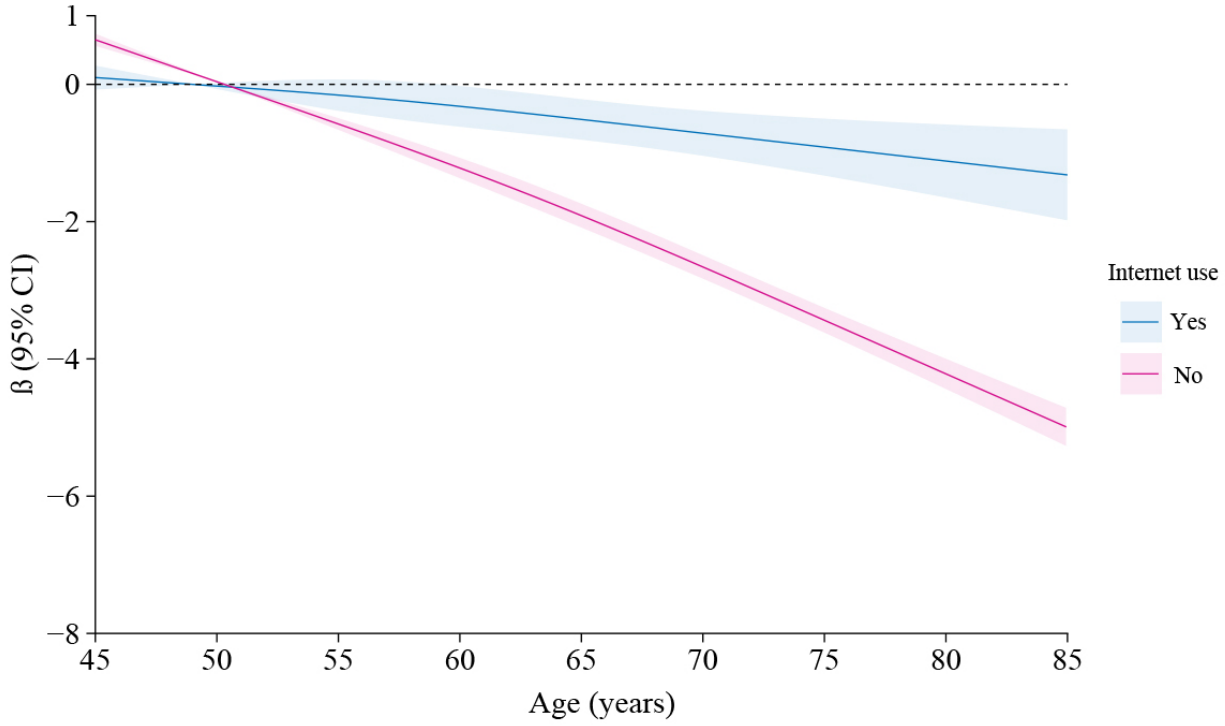

Impact of Internet Device and Frequency on Cognitive Function

The fixed effects model (Figure 4A) suggests that, among middle-aged and older adults, internet users who use cell phones experience cognitive benefits (model 1: β=0.406, 95% CI 0.300 to 0.512, *P*<.001; model 2: β=0.412, 95% CI 0.306 to 0.519, *P*<.001; model 3: β=0.392, 95% CI 0.287 to 0.499, *P*<.001; model 4: β=0.398, 95% CI 0.283 to 0.495, *P*<.001). Furthermore, internet use on computer devices also has a positive influence on cognitive function (model 1: β=0.153, 95% CI 0.097 to 0.210, *P*<.001; model 2: β=0.158, 95% CI 0.101 to 0.214, *P*<.001; model 3: β=0.149, 95% CI 0.092 to 0.205, *P*<.001; model 4: β=0.147, 95% CI 0.091 to 0.204, *P*<.001), though to a lesser extent than cell phone–based internet use. This indicates that using cell phones as internet devices results in greater cognitive benefits than using computers. The proportion of mobile phone users among internet users increased dramatically over the past 5 years, from 64.5% (648/1005) to 99.3% (5463/5503), as detailed in Multimedia Appendix 4. However, the proportion of daily internet users only slightly increased, from 74.3% (747/1005) to 79.9% (4399/5503), as shown in Multimedia Appendix 4. Additionally, we used a Spearman rank correlation test to explore the association between internet use frequency and cognitive function. The results demonstrated a meaningful positive correlation between the frequency of internet use and cognitive function (r_s=0.378, *P*<.001; Figure 4B). Moreover, individuals who reported using the internet infrequently, weekly, or daily scored higher on cognitive tests than those who never used it (*P*<.001). This suggests that, among middle-aged and older adults, the more frequent their internet use, the more beneficial it is for their cognitive function.

Figure 4. The influence on cognitive function of (A) internet device types (presented as beta coefficients and 95% CIs) and (B) frequency of internet use (assessed using Spearmans rank correlation). **P*<.001, compared with individuals who never used the internet.

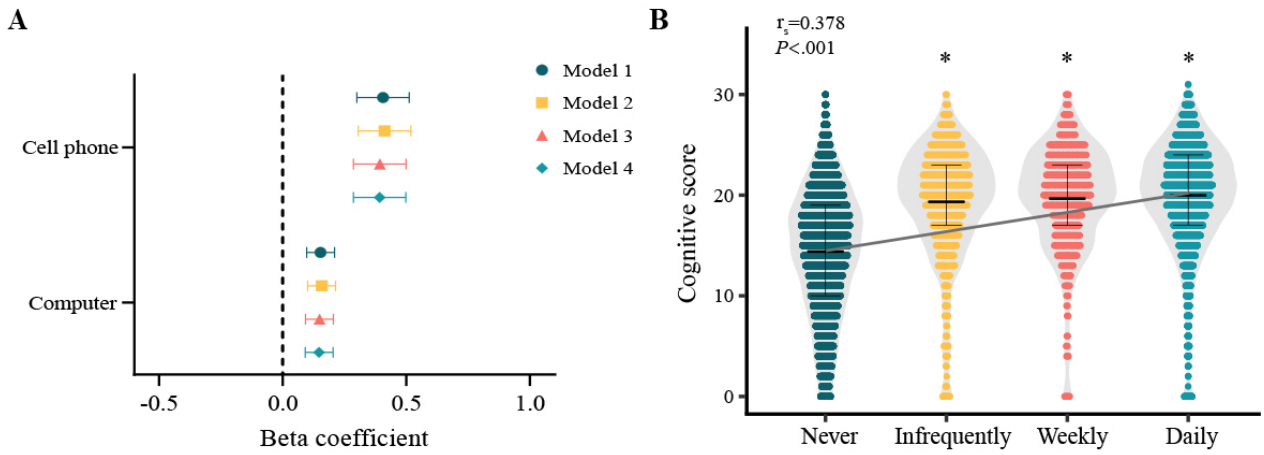

XSL•FO
RenderX

Stratified and Interaction Analyses

Stratified analysis revealed that the positive correlation between internet use and cognitive function was similar across subgroups differentiated by age, gender, residency, and retirement status (Figure 5). In addition, internet use significantly affected individuals who were married (*P*<.001), while among those who were unmarried, there was only a trend present without a significant difference (*P*=.18). Interestingly, among middle-aged and older adults, there seemed to be a positive association between lower levels of educational attainment and higher levels of cognitive benefits gained from internet use (illiterate: β=0.807, 95% CI 0.209 to 1.407, *P*=.008; low education level: β=0.545, 95% CI 0.280 to 0.810, *P*<.001; mid education level: β=0.389, 95% CI 0.160 to 0.617, *P*=.001; high education level: β=0.160, 95% CI –0.215 to 0.537, *P*=.40). As the majority of middle-aged and older adults in China currently have low levels of education [27], there may be tremendous potential for cognitive benefits from internet use.

Figure 5. Association between internet use and cognitive function stratified by participant characteristics and adjusted for all other factors (prior cognitive score, age, gender, marital status, educational attainment, residency, retirement status, smoking, drinking, hypertension, diabetes, household income per capita), and the *P* value for the interaction was evaluated by including the variables' cross-product term (internet use × baseline characteristics) in the same model.

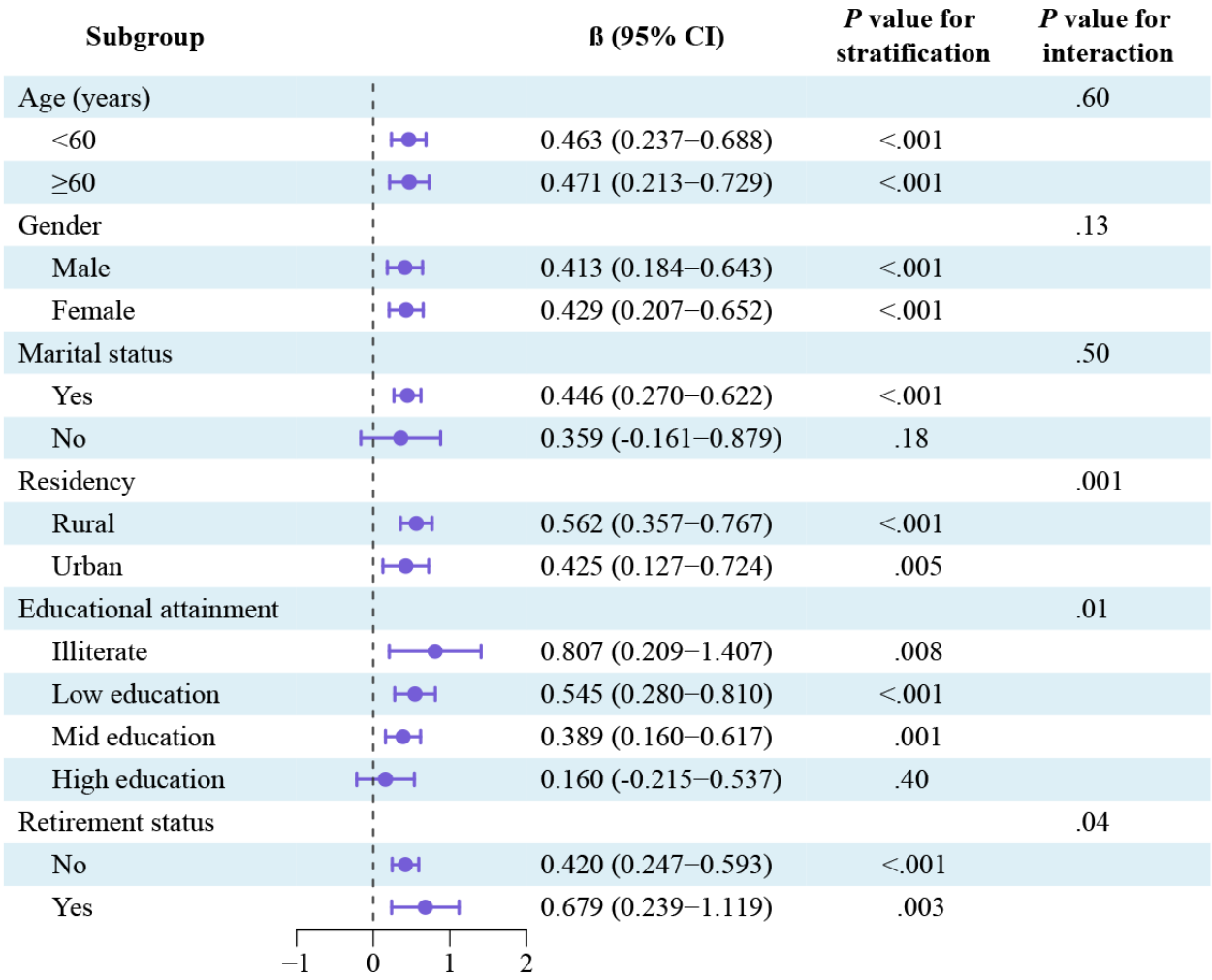

Mediation Analyses

Different areas of residence (urban or rural) among middle-aged and older adults resulted in notable differences in cognitive function (*P*<.001; Multimedia Appendix 5). In order to further explore the mediating effect of internet use on cognitive function in different residential areas, we used the KHB method for mediation analysis. As a result, internet use mediated 34.2% of the impact of residency on cognitive function (β=1.053, *P*<.001; Figure 6). This suggests that the differences in cognitive function between urban and rural areas can be partially attributed to variations in internet use.

XSL•FO
RenderX

Figure 6. The mediation effect of internet use on the relationship between residential location and cognitive function using the Karlson-Holm-Breen (KHB) method.

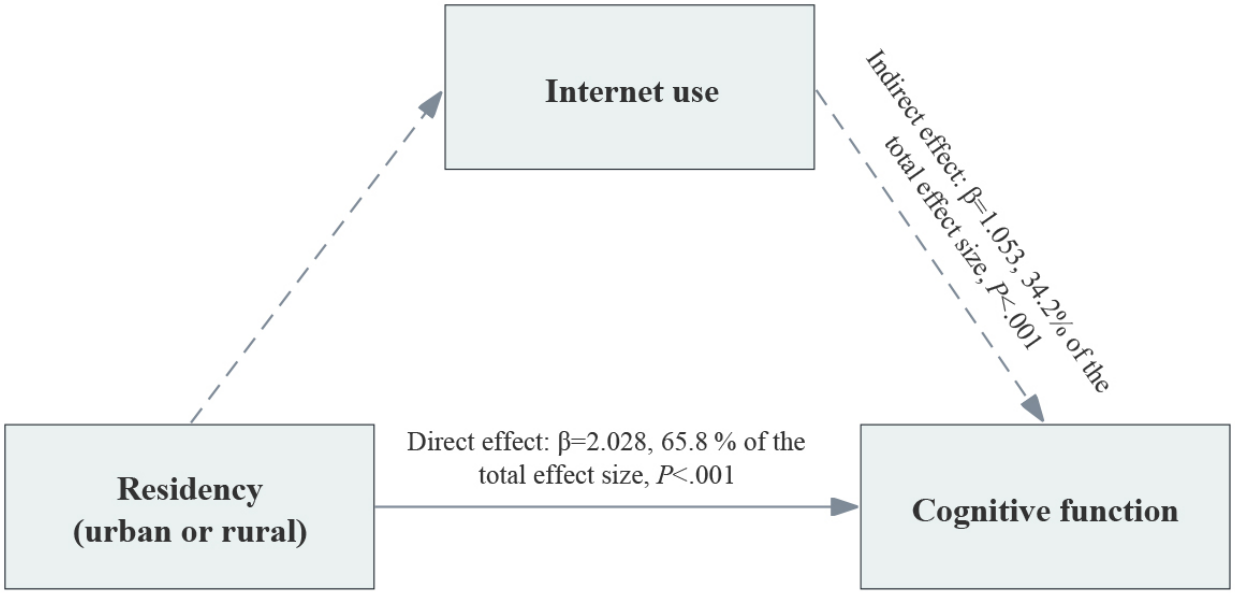

Sensitivity Analyses

Sensitivity analyses proved that the main findings of our study were robust. Following the implementation of PSM to achieve a balanced distribution of covariates (standardized mean difference<0.1) between internet users and nonusers, we obtained similar results in line with the fixed effects model (Figure 7). Moreover, we used IPTW to adjust for covariates, and the results were broadly consistent with the main analysis (Multimedia Appendix 6; model 1: β=0.559, 95% CI 0.142 to 0.976, *P*=.009; model 2: β=0.551, 95% CI 0.181 to 0.922, *P*=.004; model 3: β=0.514, 95% CI 0.167 to 0.861, *P*=.004; model 4: β=0.500, 95% CI 0.152 to 0.848, *P*=.005). In addition, we further used the OW to validate our main results (Multimedia Appendix 6; model 1: β=0.526, 95% CI 0.286 to 0.766, *P*<.001; model 2: β=0.551, 95% CI 0.312 to 0.790, *P*<.001; model 3: β=0.544, 95% CI 0.305 to 0.783, *P*<.001; model 4: β=0.542, 95% CI 0.303 to 0.781, *P*<.001). Last, when analyzing individuals who consistently used or did not use the internet for 5 years, the results of the GEE still demonstrated a positive relationship between internet use and cognitive function (Multimedia Appendix 7).

XSL•FO
RenderX

Figure 7. The effect of internet use on cognitive function after controlling covariate balance using propensity score matching (PSM), with each model adjusted after PSM.

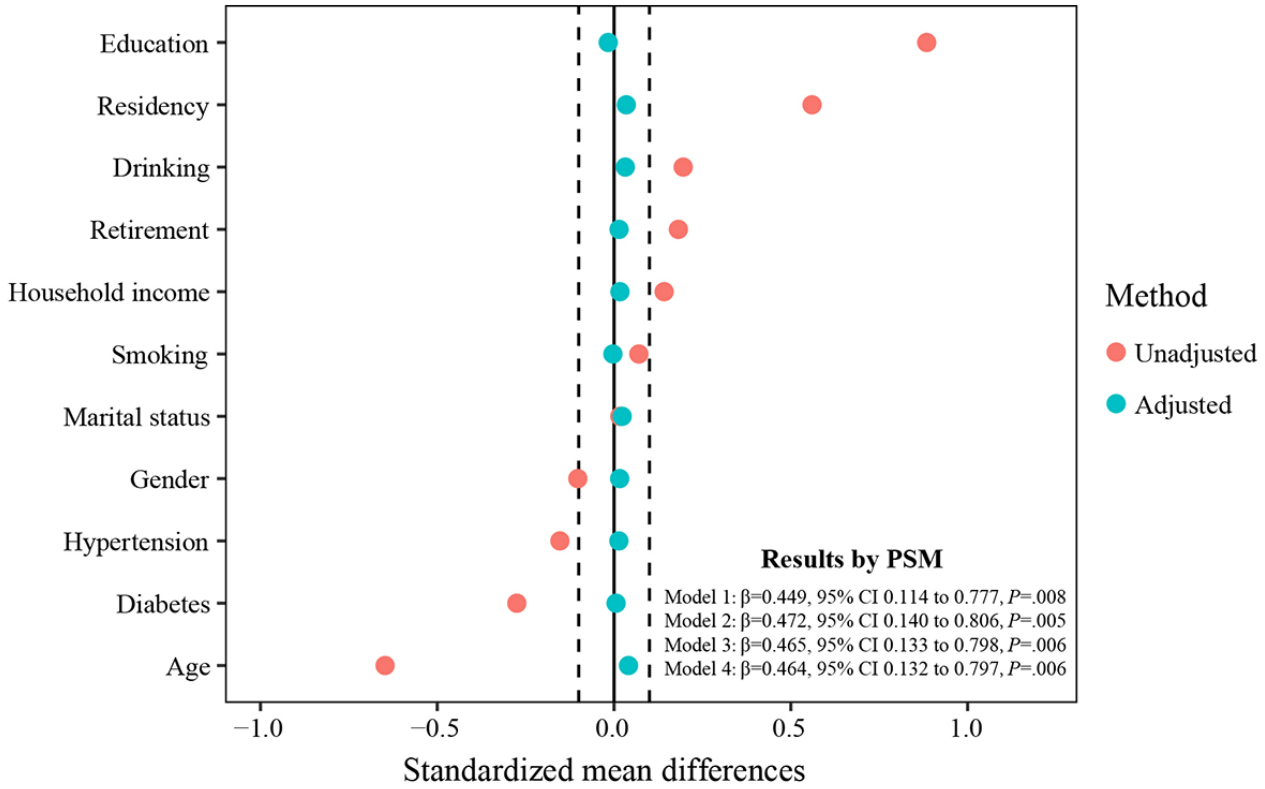

Discussion

Principal Findings

In this nationwide cohort study focusing on individuals aged 45 years and older, we found a positive relationship between internet use and cognitive function. Moreover, a higher frequency of internet use was associated with a lower risk of cognitive decline. Previous studies have suggested that using the internet can reduce social isolation, increase social interactions (both face to face and virtual or online), stimulate the acquisition of new knowledge, enhance cognitive demand for mastering novel skills using digital tools, and facilitate access to health and cultural information [28]. Individuals who use the internet may improve attention and psychomotor skills and enhance cognitive reserve through stimulation and reduction of depressive symptoms [29]. For middle-aged and older adults who use the internet, going online may help their brains use more efficient cognitive networks to counteract dementia [30]. Another explanation is that internet use is not only a form of cognitive stimulation but also a mode of social engagement. The sense of belonging created by online activities and the formation of social networks facilitate the brain's evolution and functional development [31]. These mechanisms partially account for the beneficial cognitive effects of internet use observed in our study.

Our study further explores the trends in the impact of aging on cognitive function among internet users and nonusers. In contrast to internet nonusers, we found that internet users showed a delay in age-related cognitive decline. Although we cannot ascertain a causal relationship from this observational study, we have reasons to speculate that internet use, as a cognitive-stimulating activity, may lead to beneficial physiological changes in the brain. With the progression of age, cognitive abilities such as attention, memory, executive function, language, and visuospatial skills prominently decline [32]. Additionally, the physiological changes in brain structure that accompany aging encompass alterations in neuronal structure, synaptic loss, and dysfunction of neuronal networks [33-35]. Research on brain plasticity provides strong support for the social modulation of cognition, as activity-related stimuli can induce angiogenesis, synaptogenesis, and neurogenesis in the brain [36,37]. Previous studies have not fully elucidated the cognitive benefits of internet use. Future studies could incorporate more detailed assessments of cognitive domains, neuroimaging data, or biomarkers to better understand the pathways through which internet use may influence cognitive function.

Our findings indicate that both computer-based and mobile-based internet use provide cognitive-enhancing effects, which is consistent with previous studies [38-40]. The research findings by Jin et al [38] revealed that middle-aged and older adults who own a cell phone or desktop computer at home experience a delay in cognitive decline. However, this study was based on the assessment of cell phone and desktop computer ownership, which is not as straightforward as our study in which we inquired about the specific devices used for internet access. Interestingly, our research findings suggest that middle-aged and older adults who use cell phones to access the internet have higher cognitive scores than those who use computers. This outcome might be explained by the fact that mobile phones have more advantages regarding operation and accessibility and are

XSL•FO
RenderX

more user-friendly for older adults, so mobile phone users used the internet at a higher frequency and for longer durations. With the rapid development of social information technology, some mobile apps (such as WeChat and TikTok) have become increasingly vital to the daily lives of middle-aged and older adults, as they not only facilitate social communications and connections but also provide leisure and entertainment [41]. Meanwhile, these apps have lower operational barriers, making them more user-friendly. A previous study [42] showed that, as of 2012, older adults aged 65 years and older were more likely to own a mobile phone (69%) rather than a desktop computer (48%) or a laptop (32%). The study conducted by Liu et al [43] also suggested that the absence of using mobile social media platforms, such as WeChat and mobile payment systems, may be positively associated with cognitive impairments. Therefore, the impact of cell phone use on age-related cognitive decline among middle-aged and older adults deserves more attention.

The cognitive benefits of internet use among middle-aged and older adults with different educational levels seem to vary. The results of the stratified analysis suggest that internet use has a greater extent of cognitive benefits for those with lower educational attainment. Based on previous research, there continues to be controversy regarding the impact of education on age-related cognitive decline. Some research reports that education may slow the rate of age-related cognitive decline [44-46], while others [47,48] suggest that education has no significant effect on the rate of cognitive decline. Moreover, there have been studies indicating a relationship between higher levels of education and an accelerated cognitive decline, particularly in the domains of verbal memory, processing speed, and verbal fluency [49,50]. Although the impact of education on cognitive decline remains uncertain, internet use still holds enormous potential for improving cognition in individuals with low educational levels, particularly in China where 71.1% of older adults either have no formal education or have only received primary schooling [27]. On the other hand, although middle-aged and older adults with lower educational attainment receive great cognitive benefits from surfing the internet, they may face greater potential risks or unintended consequences (such as exposure to misinformation, online scams, and privacy concerns). Hence, ensuring internet safety for middle-aged and older adults, especially those with lower levels of education, constitutes a major challenge for developing a digital society in the future.

The Spearman rank correlation results in this study indicate that, among middle-aged and older adults, higher frequencies of internet use are associated with better cognitive performance. Although higher use frequency means better cognitive benefits, the impact of overuse of the internet cannot be ignored. Since specific durations of internet use for each individual could not be observed in this study, analyzing the cognitive impacts resulting from internet overuse was not feasible. More research is needed to understand the relationship between cognitive function and both the frequency and duration of internet use, especially in the case of overuse.

The phenomenon of urban-rural disparity in cognitive performance among middle-aged and older adults is prevalent in both lower-income and higher-income countries [51,52]. Previous research has indicated that cognitive functional disparities between urban and rural areas may be caused by various factors, including health care service resources, public education resources, early-life cognitive reserves, race, social class, social activities, and social engagement [53,54]. Our research revealed that internet use, as a tool for social participation, exerts a noteworthy mediating role in the cognitive function disparities between urban and rural areas. This indicates that promoting the use of the internet in rural regions could help to reduce the urban-rural gap in cognitive performance.

Limitations

This study is not without limitations, and caution should be exercised when interpreting the findings. First, the study participants may not fully represent the middle-aged and older adult population in China, as the questionnaire survey was conducted only within the community and did not include public institutions such as nursing homes and hospitals. Therefore, the generalizability of our study findings primarily applies to healthy middle-aged and older adults, rather than frail older adults. Second, our analysis was based on a follow-up interval of approximately 5 years, and we cannot ascertain whether this correlation will endure over a more extended period. Third, the age of data and the changes in internet use over the past decade are factors that must be considered when interpreting the results of this study, as the pace of internet development and the expansion of internet access among older adults over the past 10 years may differ from what is expected in the next decade. Fourth, our approach to assessing cognitive function was not entirely based on specialized screening scales for cognitive impairment, such as the Mini-Mental State Examination or the Montreal Cognitive Assessment Scale. Therefore, caution must be exercised when interpreting the effects of internet use on mild cognitive impairment or Alzheimer disease. Fifth, due to the presence of loss to follow-up and unmeasured confounders, there may be bias in the association estimation. However, to ensure the robustness of the research data, we used various sensitivity analyses, including multiple linear regression analysis, GEE, PSM, IPTW, and OW, to validate the cognitive benefits of internet use. Sixth, our study did not directly establish a causal relationship between internet use and cognitive decline. However, the lower incidence of neurodegenerative diseases in persistent internet users indicated that internet use may prevent neurodegenerative diseases. Therefore, in order to prove the causal relationship between internet use and cognitive function, more research needs to be carried out, such as physiological mechanism research, instrumental variable analysis, and a Mendelian randomization study. Last, although the positive association between internet use and cognitive function was statistically significant after considering various potential confounders, there may still be some relevant variables, such as the apolipoprotein E4 genotype, that cannot be obtained from the survey questionnaire but could potentially influence the results.

Conclusions

In conclusion, this study presents valuable evidence supporting a longitudinal protective association between internet use and

XSL•FO
RenderX

cognitive function in middle-aged and older adults, with a higher frequency of internet use having a greater extent of cognitive benefit. As digital devices for internet access, cell phones may offer a potentially superior level of cognitive protection than computers. The cognitive gap among middle-aged and older adults in urban and rural areas can be partially attributed to internet use. This study suggests that internet use may be a promising cognitive stimulation activity that mitigates age-related cognitive decline and prevents neurodegenerative diseases. Although further research is needed to confirm the causal relationship, the findings highlight the potential benefits of promoting internet use among middle-aged and older adults, especially in rural areas, as a strategy to support cognitive health and potentially reduce the risk of dementia.

Acknowledgments

The authors thank the China Health and Retirement Longitudinal Study (CHARLS) research team and all respondents for their contribution. This study was supported by grants from the National Natural Science Foundation of China (number 81703216).

Data Availability

The data sets used and analyzed during the study are available on the China Health and Retirement Longitudinal Study (CHARLS) website [55].

Authors' Contributions

BC and JZ designed the research. BC conducted the statistical analysis and wrote the manuscript drafts. PL supervised the use of statistical methods, validated the data, and replicated the results. SR and CY conducted the data visualization. PL and JZ revised the manuscript. All authors read and approved the submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The questionnaire section on internet use and cognitive function in China Health and Retirement Longitudinal Study (CHARLS).
[DOCX File , 16 KB - jmir_v26i1e57301_app1.docx]

Multimedia Appendix 2
Cognitive scores of internet users and non-users at different waves.
[DOCX File , 13 KB - jmir_v26i1e57301_app2.docx]

Multimedia Appendix 3
The prevalence and incidence of neurodegenerative diseases (Alzheimer disease or Parkinson disease) in internet users and non-users during the 5-year period.
[DOCX File , 17 KB - jmir_v26i1e57301_app3.docx]

Multimedia Appendix 4
Changes in the proportion of internet users with different internet access devices and frequencies.
[DOCX File , 50 KB - jmir_v26i1e57301_app4.docx]

Multimedia Appendix 5
Cognitive scores of rural-urban participants at different waves.
[DOCX File , 13 KB - jmir_v26i1e57301_app5.docx]

Multimedia Appendix 6
The relationship between internet use and cognitive function in middle-aged and older adults was analyzed using inverse probability of treatment weighting (IPTW) and overlap weighting (OW) methods.
[DOCX File , 13 KB - jmir_v26i1e57301_app6.docx]

Multimedia Appendix 7
Differences in cognitive function between persistent internet users (n=671) and persistent non-users (n=7099) using generalized estimating equations (GEE) and multiple linear regression models.
[DOCX File , 14 KB - jmir_v26i1e57301_app7.docx]

XSL•FO
RenderX

References

1. Pais R, Ruano L, P Carvalho O, Barros H. Global cognitive impairment prevalence and incidence in community dwelling older adults-a systematic review. Geriatrics (Basel) 2020 Oct 27;5(4):84 [FREE Full text] [doi: 10.3390/geriatrics5040084] [Medline: 33121002]
2. Prince MJ, Wu F, Guo Y, Gutierrez Robledo LM, O'Donnell M, Sullivan R, et al. The burden of disease in older people and implications for health policy and practice. Lancet 2015 Feb 07;385(9967):549-562. [doi: 10.1016/S0140-6736(14)61347-7] [Medline: 25468153]
3. Liu J, Wang L, Tan J. Dementia in China: Current status. Neurology 2013 Sep 17;81(12):1077-1078. [doi: 10.1212/wnl.0b013e3182a4a3cb]
4. Jia L, Du Y, Chu L, Zhang Z, Li F, Lyu D, et al. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. The Lancet Public Health 2020 Dec;5(12):e661-e671. [doi: 10.1016/s2468-2667(20)30185-7]
5. Frankish H, Horton R. Prevention and management of dementia: a priority for public health. The Lancet 2017 Dec;390(10113):2614-2615. [doi: 10.1016/s0140-6736(17)31756-7]
6. Satizabal CL, Beiser AS, Chouraki V, Chêne G, Dufouil C, Seshadri S. Incidence of dementia over three decades in the Framingham Heart Study. N Engl J Med 2016 Feb 11;374(6):523-532. [doi: 10.1056/nejmoa1504327]
7. Anstey KJ, Bahar-Fuchs A, Herath P, Kim S, Burns R, Rebok GW, et al. Body brain life: A randomized controlled trial of an online dementia risk reduction intervention in middle-aged adults at risk of Alzheimer's disease. Alzheimers Dement (N Y) 2015 Jun 10;1(1):72-80 [FREE Full text] [doi: 10.1016/j.trci.2015.04.003] [Medline: 29854927]
8. García-Casal JA, Loizeau A, Csipke E, Franco-Martín M, Perea-Bartolomé MV, Orrell M. Computer-based cognitive interventions for people living with dementia: a systematic literature review and meta-analysis. Aging Ment Health 2017 May;21(5):454-467. [doi: 10.1080/13607863.2015.1132677] [Medline: 26806365]
9. The 52nd Statistical Report on China's Internet Development. China Internet Network Information Center (CNNIC). 2023 Aug. URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202311/P020231121355042476714.pdf [accessed 2024-11-21]
10. Xavier AJ, d'Orsi E, de Oliveira CM, Orrell M, Demakakos P, Biddulph JP, et al. English Longitudinal Study of Aging: can internet/e-mail use reduce cognitive decline? J Gerontol A Biol Sci Med Sci 2014 Sep 12;69(9):1117-1121. [doi: 10.1093/gerona/glu105] [Medline: 25116923]
11. Almeida OP, Yeap BB, Alfonso H, Hankey GJ, Flicker L, Norman PE. Older men who use computers have lower risk of dementia. PLoS One 2012 Aug 28;7(8):e44239 [FREE Full text] [doi: 10.1371/journal.pone.0044239] [Medline: 22937167]
12. Klimova B. Use of the Internet as a prevention tool against cognitive decline in normal aging. Clin Interv Aging 2016 Sep;11:1231-1237 [FREE Full text] [doi: 10.2147/CIA.S113758] [Medline: 27672317]
13. Small GW, Moody TD, Siddarth P, Bookheimer SY. Your brain on Google: patterns of cerebral activation during internet searching. The American Journal of Geriatric Psychiatry 2009 Feb;17(2):116-126. [doi: 10.1097/jgp.0b013e3181953a02]
14. Cresci MK, Jarosz PA, Templin TN. Are health answers online for older adults? Educational Gerontology 2012 Jan;38(1):10-19. [doi: 10.1080/03601277.2010.515890]
15. Hakkarainen P. 'No good for shovelling snow and carrying firewood': Social representations of computers and the internet by elderly Finnish non-users. New Media & Society 2012 May 07;14(7):1198-1215. [doi: 10.1177/1461444812442663]
16. Ashby FG, Isen AM, Turken AU. A neuropsychological theory of positive affect and its influence on cognition. Psychol Rev 1999 Jul;106(3):529-550. [doi: 10.1037/0033-295x.106.3.529] [Medline: 10467897]
17. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the Internet may be changing our cognition. World Psychiatry 2019 Jun 06;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]
18. Slegers K, van Boxtel M, Jolles J. Effects of computer training and internet usage on cognitive abilities in older adults: a randomized controlled study. Aging Clin Exp Res 2013 Jul 26;21(1):43-54. [doi: 10.1007/bf03324898]
19. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol 2014 Feb 12;43(1):61-68. [doi: 10.1093/ije/dys203] [Medline: 23243115]
20. Manly JJ, Jones RN, Langa KM, Ryan LH, Levine DA, McCammon R, et al. Estimating the prevalence of dementia and mild cognitive impairment in the US: the 2016 Health and Retirement Study Harmonized Cognitive Assessment protocol project. JAMA Neurol 2022 Dec 01;79(12):1242-1249 [FREE Full text] [doi: 10.1001/jamaneurol.2022.3543] [Medline: 36279130]
21. Meng Q, Wang H, Strauss J, Langa KM, Chen X, Wang M, et al. Validation of neuropsychological tests for the China Health and Retirement Longitudinal Study Harmonized Cognitive Assessment Protocol. Int Psychogeriatr 2019 Dec;31(12):1709-1719 [FREE Full text] [doi: 10.1017/S1041610219000693] [Medline: 31309907]
22. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nat Rev Neurol 2018 Nov 5;14(11):653-666. [doi: 10.1038/s41582-018-0070-3] [Medline: 30291317]
23. Karlson KB, Holm A, Breen R. Comparing regression coefficients between same-sample nested models using logit and probit. Sociological Methodology 2012 Nov 21;42(1):286-313. [doi: 10.1177/0081175012444861]
24. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983 Apr;70(1):41. [doi: 10.2307/2335942]

XSL•FO
RenderX

25. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling 1986 Jan;7(9-12):1393-1512 [FREE Full text] [doi: 10.1016/0270-0255(86)90088-6]
26. Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. JAMA 2020 Jun 16;323(23):2417-2418. [doi: 10.1001/jama.2020.7819] [Medline: 32369102]
27. Fang Y, Wang H, Ouyang Z. An overview report on the situation of elderly people in urban and rural China: 2000–2015. China popul. dev. stud 2018 Oct 9;2(3):323-345. [doi: 10.1007/s42379-018-0016-x]
28. d'Orsi E, Xavier AJ, Rafnsson SB, Steptoe A, Hogervorst E, Orrell M. Is use of the internet in midlife associated with lower dementia incidence? Results from the English Longitudinal Study of Ageing. Aging Ment Health 2018 Nov 10;22(11):1525-1533 [FREE Full text] [doi: 10.1080/13607863.2017.1360840] [Medline: 28795579]
29. Shapira N, Barak A, Gal I. Promoting older adults' well-being through Internet training and use. Aging Ment Health 2007 Sep;11(5):477-484. [doi: 10.1080/13607860601086546] [Medline: 17882585]
30. La Rue A. Healthy brain aging: role of cognitive reserve, cognitive stimulation, and cognitive exercises. Clin Geriatr Med 2010 Feb;26(1):99-111. [doi: 10.1016/j.cger.2009.11.003] [Medline: 20176296]
31. Fratiglioni L, Paillard-Borg S, Winblad B. An active and socially integrated lifestyle in late life might protect against dementia. The Lancet Neurology 2004 Jun;3(6):343-353. [doi: 10.1016/s1474-4422(04)00767-7]
32. Murman D. The impact of age on cognition. Semin Hear 2015 Aug 9;36(3):111-121. [doi: 10.1055/s-0035-1555115] [Medline: 27516712]
33. Masliah E, Mallory M, Hansen L, DeTeresa R, Terry RD. Quantitative synaptic alterations in the human neocortex during normal aging. Neurology 1993 Jan;43(1):192-197. [doi: 10.1212/wnl.43.1_part_1.192] [Medline: 8423884]
34. Dennis EL, Thompson PM. Functional brain connectivity using fMRI in aging and Alzheimer's disease. Neuropsychol Rev 2014 Mar 23;24(1):49-62. [doi: 10.1007/s11065-014-9249-6] [Medline: 24562737]
35. Morrison JH, Baxter MG. The ageing cortical synapse: hallmarks and implications for cognitive decline. Nat Rev Neurosci 2012 Mar 07;13(4):240-250. [doi: 10.1038/nrn3200] [Medline: 22395804]
36. Churchill JD, Galvez R, Colcombe S, Swain RA, Kramer AF, Greenough WT. Exercise, experience and the aging brain. Neurobiol Aging 2002 Sep;23(5):941-955. [doi: 10.1016/s0197-4580(02)00028-3] [Medline: 12392797]
37. Fillit HM, Butler RN, O'Connell AW, Albert MS, Birren JE, Cotman CW, et al. Achieving and maintaining cognitive vitality with aging. Mayo Clin Proc 2002 Jul;77(7):681-696. [doi: 10.4065/77.7.681] [Medline: 12108606]
38. Jin Y, Jing M, Ma X. Effects of digital device ownership on cognitive decline in a middle-aged and elderly population: longitudinal observational study. J Med Internet Res 2019 Jul 29;21(7):e14210 [FREE Full text] [doi: 10.2196/14210] [Medline: 31359864]
39. Tun PA, Lachman ME. The association between computer use and cognition across adulthood: use it so you won't lose it? Psychol Aging 2010 Sep;25(3):560-568. [doi: 10.1037/a0019543] [Medline: 20677884]
40. Lin L, Jing X, Lv S, Liang J, Tian L, Li H, et al. Mobile device use and the cognitive function and depressive symptoms of older adults living in residential care homes. BMC Geriatr 2020 Feb 03;20(1):41 [FREE Full text] [doi: 10.1186/s12877-020-1427-1] [Medline: 32013915]
41. Wang G, Zhang W, Zeng R. WeChat use intensity and social support: The moderating effect of motivators for WeChat use. Computers in Human Behavior 2019 Feb;91:244-251 [FREE Full text] [doi: 10.1016/j.chb.2018.10.010]
42. Joe J, Demiris G. Older adults and mobile phones for health: a review. J Biomed Inform 2013 Oct;46(5):947-954 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.008] [Medline: 23810858]
43. Liu X, Wang X, Zhang H, Pei M, Li N. Relationship between digital exclusion and cognitive impairment in Chinese adults. Front Aging Neurosci 2023 Jul 3;15:1194348 [FREE Full text] [doi: 10.3389/fnagi.2023.1194348] [Medline: 37465320]
44. Bennett DA, Wilson RS, Schneider JA, Evans DA, Mendes de Leon CF, Arnold SE, et al. Education modifies the relation of AD pathology to level of cognitive function in older persons. Neurology 2003 Jun 24;60(12):1909-1915. [doi: 10.1212/01.wnl.0000069923.64550.9f] [Medline: 12821732]
45. Evans D, Beckett L, Albert M, Hebert L, Scherr P, Funkenstein H, et al. Level of education and change in cognitive function in a community population of older persons. Ann Epidemiol 1993 Jan;3(1):71-77 [FREE Full text] [doi: 10.1016/1047-2797(93)90012-s] [Medline: 8287159]
46. Lyketsos CG, Chen L, Anthony JC. Cognitive decline in adulthood: an 11.5-year follow-up of the Baltimore Epidemiologic Catchment Area study. Am J Psychiatry 1999 Jan;156(1):58-65. [doi: 10.1176/ajp.156.1.58] [Medline: 9892298]
47. Carmelli D, Swan GE, LaRue A, Eslinger PJ. Correlates of change in cognitive function in survivors from the Western Collaborative Group Study. Neuroepidemiology 1997 Dec 3;16(6):285-295. [doi: 10.1159/000109699] [Medline: 9430128]
48. Hultsch DF, Hertzog C, Dixon RA, Small BJ. Memory change in the aged. New York, NY: Cambridge University Press; 1998.
49. Alley D, Suthers K, Crimmins E. Education and cognitive decline in older Americans: results from the AHEAD sample. Res Aging 2007 Jan 01;29(1):73-94. [doi: 10.1177/0164027506294245] [Medline: 19830260]
50. Proust-Lima C, Amieva H, Letenneur L, Orgogozo J, Jacqmin-Gadda H, Dartigues J. Gender and education impact on brain aging: a general cognitive factor approach. Psychol Aging 2008 Sep;23(3):608-620. [doi: 10.1037/a0012838] [Medline: 18808250]

XSL•FO
RenderX

51. Glauber R. Rural depopulation and the rural-urban gap in cognitive functioning among older adults. J Rural Health 2022 Sep 07;38(4):696-704. [doi: 10.1111/jrh.12650] [Medline: 35257439]
52. Zhang T, Lu B, Wang X. Urban-rural disparity in cognitive performance among older Chinese adults: explaining the changes from 2008 to 2018. Front Public Health 2022 Mar 23;10:843608 [FREE Full text] [doi: 10.3389/fpubh.2022.843608] [Medline: 35400051]
53. Wight R, Aneshensel C, Miller-Martinez D, Botticello A, Cummings J, Karlamangla A, et al. Urban neighborhood context, educational attainment, and cognitive function among older adults. Am J Epidemiol 2006 Jun 15;163(12):1071-1078. [doi: 10.1093/aje/kwj176] [Medline: 16707655]
54. Liu D, Li L, An L, Cheng G, Chen C, Zou M, et al. Urban-rural disparities in mild cognitive impairment and its functional subtypes among community-dwelling older residents in central China. Gen Psychiatr 2021 Oct 27;34(5):e100564 [FREE Full text] [doi: 10.1136/gpsych-2021-100564] [Medline: 34790888]
55. 2020 National Tracking Survey. CHARLS. URL: https://charls.charlsdata.com/pages/data/111/zh-cn.html [accessed 2024-11-21]

Abbreviations

CHARLS: China Health and Retirement Longitudinal Study
GEE: generalized estimating equations
IPTW: inverse probability of treatment weighting
KHB: Karlson-Holm-Breen method
OW: overlap weighting
PSM: propensity score matching

Edited by N Cahill, T Leung; submitted 15.02.24; peer-reviewed by A Azizan, T Soluoku; comments to author 06.05.24; revised version received 31.05.24; accepted 12.11.24; published 02.12.24.

Please cite as:
Chen B, Yang C, Ren S, Li P, Zhao J
Relationship Between Internet Use and Cognitive Function Among Middle-Aged and Older Chinese Adults: 5-Year Longitudinal Study
J Med Internet Res 2024;26:e57301
URL: https://www.jmir.org/2024/1/e57301
doi:10.2196/57301
PMID:39539034

©Bowen Chen, Chun Yang, Shanshan Ren, Penggao Li, Jin Zhao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effects of the COVID-19 Pandemic on Age-Based Disparities in Digital Health Technology Use: Secondary Analysis of the 2017-2022 Health Information National Trends Survey

Yuanbo Qiu[1], PhD; Huang Huang[2], PhD; Junjie Gai[3], MS; Gianluca De Leo[2], MBA, PhD

[1]School of Journalism and Communication, South China University of Technology, Guanzhou, China

[2]Department of Health Management, Economics, and Policy, School of Public Health, Augusta University, Augusta, GA, United States

[3]Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa city, IA, United States

Corresponding Author:
Huang Huang, PhD
Department of Health Management, Economics, and Policy
School of Public Health
Augusta University
2500 Walton Way
Science Hall, E-1031
Augusta, GA, 30904
United States
Phone: 1 859 551 9185
Email: huhuang@augusta.edu

Abstract

Background: The COVID-19 pandemic accelerated the adoption of digital health technology, but it could also impact age-based disparities as existing studies have pointed out. Compared with the pre-pandemic period, whether the rapid digitalization of the health care system during the pandemic widened the age-based disparities over a long period remains unclear.

Objective: This study aimed to analyze the long-term effects of the COVID-19 pandemic on the multifaceted landscape of digital health technology used across diverse age groups among US citizens.

Methods: We conducted the retrospective observational study using the 2017-2022 Health Information National Trends Survey to identify the influence of the COVID-19 pandemic on a wide range of digital health technology use outcomes across various age groups. The sample included 15,505 respondents, which were categorized into 3 age groups: adults (18-44 years), middle-aged adults (45-64 years), and older adults (more than 65 years). We also designated the time point of March 11, 2020, to divide the pre- and post-pandemic periods. Based on these categorizations, multivariate linear probability models were used to assess pre-post changes in digital health technology use, controlling for demographic, socioeconomic, and health-related variables among different age groups.

Results: Essentially, older adults were found to be significantly less likely to use digital health technology compared with adults, with a 26.28% lower likelihood of using the internet for health information ($P<.001$) and a 32.63% lower likelihood of using health apps ($P<.001$). The usage of digital health technology for all age groups had significantly increased after the onset of the pandemic, and the age-based disparities became smaller in terms of using the internet to look for health information. However, the disparities have widened for older adults in using the internet to look up test results (11.21%, $P<.001$) and make appointments (10.03%, $P=.006$) and using wearable devices to track health (8.31%, $P=.01$).

Conclusions: Our study reveals a significant increase in the use of digital health technology among all age groups during the pandemic. However, while the disparities in accessing online information have narrowed, age-based disparities, particularly for older adults, have widened in most areas such as looking up test results and making appointments with doctors. Therefore, older adults are more likely left behind by the rapidly digitalized US health care system during the pandemic. Policy makers and health care providers should focus on addressing these disparities to ensure equitable access to digital health resources for US baby boomers.

(J Med Internet Res 2024;26:e65541) doi:10.2196/65541

XSL•FO
RenderX

KEYWORDS

age-based disparities; health equity; digital health technology use; digital divide; health policy; COVID-19; mobile phone

Introduction

Digital health technology use is the usage of information and communication technologies such as mobile phones, computers, and wearable devices to engage individuals in health-related activities, including seeking health information, reviewing test results, scheduling appointments, and monitoring health conditions [1]. Digital health technology facilitates health information seeking and health management, and has become increasingly vital for individuals to access health information and remote care in recent decades [2,3]. The adoption and implementation of digital health technology in the US health care system have been accelerated dramatically during the COVID-19 pandemic when individual mobility and in-person contacts were greatly restricted by social distancing and quarantine orders [4,5]. Researchers record approximately 50% of the US citizens have embraced smartphones, computers, or other electronic devices to engage with health apps, track personal health metrics, and consult with health care professionals to various degrees during the pandemic [6,7].

Despite the potential benefits, heavy reliance on digital health technology may inadvertently intensify disparities in access to health information and resources across different demographic groups [8-10]. The digital divide, characterized by unequal access to digital technology in health care, was found to exist for decades and has been intensified during the pandemic due to various barriers, including disparities in internet access [11,12], broadband infrastructure [13], or digital literacy skills [14]. Consequently, the widening digital divide may render vulnerable populations less likely to access health care services and more susceptible to diseases, therefore amplifying health inequalities [15,16].

The age-based digital divide, or the gray digital divide, constitutes a significant challenge for older adults in receiving health information, accessing health services, and managing health conditions in the information era. The low adoption and limited use of digital health technology by older adults, in contrast to younger generations, have been well documented over a long period [17-19]. Entering the pandemic, older adults' greater needs for regular health consultations were constrained because of the restrictions on physical interaction, and therefore, accessing health care services remotely through digital health technology became the primary or even the only option [20]. The rapid advancement of digital health technology and infrastructure during the pandemic significantly increased telehealth use among the older adult population [21]. This trend seems to echo the observations regarding the narrowing race and ethnicity disparity in the use of digital health technology: the threat of a global pandemic and the fear of disease and death, or in other words, the coronavirus nosophobia, overpowered objective or subjective barriers of using updated digital health technology among the vulnerable populations [22].

Nevertheless, researchers also warned that the surging demands of digital health technology generated by the pandemic among older adults have not been adequately catered [23]. As several preexisting disadvantages such as the lack of digital skills, economic disparities, technological anxiety, and distrust of digital health care remained the key factors that contributed to age-based disparities [20,24,25], the older groups encountered more digital obstacles and experienced more digital isolation compared with their younger counterparts during the pandemic [26,27]. In such circumstances, failing to effectively use digital health technology may result in inadequate chronic disease management, delayed health care access, difficulty in seeking health information, disengagement in health behaviors or promotion efforts, and ultimately worse health outcomes for older generations [28-30]. If the age-based disparity in the use of digital health technology continues to widen during the COVID-19 period, the older generations left behind will inevitably become marginalized by the accelerated digitalization of the US health care system. Therefore, as Healthy People 2030 highlights the critical roles of health information techniques in health communication and promotion [31], whether the COVID-19 pandemic has narrowed or widened the age-based disparities in the long term becomes a pressing concern of policy makers and scholars.

Despite the increased significance and ample evidence of age-based disparities in the use of digital health technology before or at the onset of the pandemic, a question remains unanswered: to what extent has the age-based digital divide been reshaped by the pandemic over a long term? Researchers tend to theoretically discuss the reasons, potential harms, or feasible strategies of general disparity in digital health technology usage [12,32,33]. However, research gaps persist, including restricted study periods, lack of prepandemic baseline comparison, and small sample size. Thus, there is a pressing need for empirical evidence on the changes in digital health technology use across different generations throughout the COVID-19 pandemic. Answering this question not only helps us identify specific underused digital health technology and often-neglected vulnerable age groups at the present but also informs effective public health communication and intervention strategies for the underserved and vulnerable groups in the future.

To address this gap, we leverage the multivariate linear probability model (LPM) to analyze the large, nationwide representative sample in the Health Information National Trends Survey (HINTS) from 2017 to 2022. Our investigation delves into the multifaceted landscape of digital health technology use across diverse age groups and covers a long period from the prepandemic period, the pandemic's onset, to the late pandemic period. The findings provide empirical evidence of age-based disparities changed by the pandemic and justify further solutions and interventions to mitigate age-based disparities in digital health technology usage among the baby boomer generation.

XSL•FO
RenderX

Methods

Data

We conducted the retrospective observational study using the 2017-2022 HINTS to identify the influence of the COVID-19 pandemic on a wide range of digital health technology uses across various age groups. HINTS is a nationally representative survey administered by the National Cancer Institute (NCI) every 1 or 2 years since 2003. It targets the US civilian noninstitutionalized adults aged 18 years and older [34]. In order to guarantee a nationally representative sample, HINTS adopts a 2-stage sampling design that first generates a random stratified household address sample from the sampling frame and randomly selects one adult in the household to complete the survey. It provides detailed information including but not limited to health information searching, medical health records, health behaviors, health care access, health beliefs, and demographics, which enables health scholars to better monitor the changes in health-related knowledge and health information technology usage among the US public.

We constructed the pooled cross-sectional dataset (N=22,344) from the HINTS 5 cycles 1 to 4 (2017-2020) and HINTS 6 (2022) original datasets. A special case is that HINTS divides the 2020 sample into the pre- and post-pandemic–onset periods based on the indicator of whether the respondents completed the survey after March 11, 2020 [12]. However, because HINTS asked about the respondents' digital health technology use in the past 12 months, the records of the 2020 post-pandemic–onset sample may potentially reflect their use before the pandemic, so we decided to exclude the 2020 post-pandemic–onset sample from our analytic sample (n=2428). After listwise deletion of all observations with any missing values in the control covariates (n=4411), the final sample size was 15,505; however, due to the change of survey design and missing values in outcome variables, readers should notice the sample sizes in regression results vary by outcomes.

Measurement

Outcomes

Our outcomes of digital health technology use involved using the internet, health apps in mobile devices, and electronic wearable devices to conduct health-related activities, consistent with previous studies [18,35,36]. First, the outcome of internet use asked whether the respondents have ever used the internet to (1) look for health or medical information, (2) send a message to a health care provider or a health care provider's office, (3) view medical test results, or (4) make an appointment with a health care provider. Second, health app usage focused on whether respondents ever used the health or wellness app on a tablet or smartphone. Third, we also had interests in whether respondents used an electronic wearable device to monitor health conditions or activity. All the questions were framed within the time limit of the past 12 months. We coded all the outcomes as binary variables; in order to encompass a large group who fail to use digital health technology due to the lack of internet connection or digital devices, we also coded the respondents who reported no internet access, no digital device, or only basic cell phone as 0 in our outcomes. We believe this full range of digital health technology variables covers the majority of digital health technology that Americans currently rely on to seek health information, access health services, and maintain overall well-being.

COVID-19 Pandemic

The pandemic indicator was generated based on when the survey was completed before March 11, 2020. The prepandemic period consisted of the 2017, 2018, 2019, and 2020 prepandemic periods as flagged by HINTS, while we classified 2022 as the postpandemic period.

Age Groups

Following the definition of age group in medical subject headings [37], we classified the sample into three age groups based on their self-reported ages: (1) adults (aged 18-44 years), (2) middle-aged adults (aged 45-64 years), and (3) older adults (aged 65 years or older).

Control Covariates

Following the existing literature, 3 sets of individual-level control covariates were introduced into our study to alleviate the concerns of confounders in digital health technology uses, which were demographic, socioeconomic, and health [12,22,38,39]. First, the demographic control covariates included race and ethnicity (non-Hispanic White [reference group], Hispanic, non-Hispanic Black, and other non-Hispanic), sex (male and female), marital status, and household size. Second, socioeconomic covariates contained education level (less than high school [reference group], high school, some college, college graduate, or more), household annual income in the past year (less than US $20,000 [reference group]; US $20,000-34,999; US $25,000-49,999; US $50,000-74,999; and US $75,000 or more), residing in a metropolitan area, and census region (Northeast [reference group], Midwest, South, and West). Third, evidence indicates that individuals with different general health status tend to have different degrees of digital health technology use [40], and that adoption of digital health technologies is more prevalent among those with chronic conditions [41], including diabetes [42], hypertension [43], heart conditions [44], chronic lung disease [45], and depression [46] in certain conditions. Therefore, we considered the following variables as health-related control covariates: insured, general health status (poor [reference group], fair, good, very good, and excellent), BMI, diagnosed as diabetes, hypertension, heart conditions, chronic lung disease or depression, and mental health scores derived from a list of mental health issues in the past 2 weeks.

Analytical Approach

In the descriptive analysis, we computed the descriptive statistics for 6 outcome variables of digital health technology use and the control covariates in the overall sample and 3 subsamples of adults, middle-aged adults, and older adults. Also, we showed the unadjusted trends of digital health technology use across the pandemic by different age groups in the graph.

We used the multivariate LPM, the extension of ordinary least squares regression in the binary outcome, to detect any significant differences in the changes of outcomes due to the

XSL•FO
RenderX

pandemic across various age groups. In other words, we compared the pre-post differences in digital health technology use in the target age group, either middle-aged or older adults, with the differences in the reference group (ie, adults). Specifically, we estimated the following equation:

Y_{it} is the outcome for individual i at year t. $COVID_t$ is the dummy variable indicating whether the individual was surveyed after the onset of COVID-19 pandemic. $Middle\ Aged_{it}$ and $Old\ Adult_{it}$ are dummy indicators of age group. X_{it} captures the individual-level confounders of digital health technology use. ω_t includes year fixed effects to account for common changes in outcomes shared by each survey cohort. β_1 reflects the effect of the pandemic on the adults group. β_2 is the effect of the middle-aged population before the pandemic, while β_3 catches the impact of older adults before the pandemic. Our key parameters of interest are β_4 and β_5, which allows us to determine whether the COVID-19 pandemic affected various age groups to different extents. β_4 indicates the changes of disparity in digital health technology use between middle-aged adults and adults in the postpandemic period, relative to the prepandemic disparity between middle-aged adults and adults. If the signs of β_4 and β_2 are the same, the gap between middle-aged adults and adults expands; otherwise, the gap narrows. Similarly, we can interpret β_5 as the pre-post changes in the differences between adults and older adults and conclude whether the disparity between adults and older adults is exacerbated by the signs of β_5 and β_3.

Because pandemic-related changes of age-based digital disparities may be different across various demographic groups, we conducted several subgroup analyses to differentiate the heterogeneous effects and separately estimated the same regression model shown in equation 1 among multiple groups of race and ethnicity, marital status, sex, and rurality. In addition, the sensitivity analyses included (1) estimating the logit model which is used for the binary outcomes and comparing with the LPM model, (2) various approaches of sample selections for the 2020 sample, and (3) estimating the model in a subset sample with internet connections and digital devices in order to isolate the influence of objective barriers.

All the analyses were conducted on Stata (version 18; StataCorp) [47]. All LPM coefficients were scaled by 100 for readability. To account for the HINTS complex sample design and obtain the nationally representative estimates, we calculated the jackknife replication SE and incorporated sampling weights into descriptive and regression analyses, suggested by HINTS [48] as well as previous studies [12,18,38]. We used the weights provided by HINTS and the Stata survey command “svy” to implement this complex survey weighting process and obtain the weighted statistics. The statistical significance level was set at $P<.05$.

Ethical Considerations

Given that HINTS data collection has received approval from the institutional review board, and our secondary data analysis used deidentified, publicly available information, our study is exempt from institutional review board review.

Results

Descriptive Analysis

Among the 15,505 survey respondents across 2017-2022, a total of 4361 were adults, 6055 were middle-aged adults, and 5089 were older adults. Table 1 summarizes their digital health technology use patterns and personal characteristics. Overall, US populations tended to use the internet to look for health information (11,529/15,411, 74.81%), followed by using health apps (3073/6059, 50.71%), making appointments (5807/12,771, 45.47%), using the internet to communicate with health providers (6825/15,399, 44.32%), and looking up test results (6649/15,370, 43.26%); the lowest was using electronic wearable devices to track health, with only 32.89% (3394/10,318). Adults were most likely to use digital health technology, while older adults demonstrated the lowest likelihood. Several individual characteristics across age groups are noticeable. Older adults were more likely to report as non-Hispanic White (4011/5089, 78.82%), female (2695/5089, 52.96%), having less than high school or high school education level (1710/5089, 33.6%), and having chronic diseases except for depression (ie, diabetes: 1470/5089, 28.89%; hypertension: 3278/5089, 64.41%; heart conditions: 927/5089, 18.22%; and chronic lung diseases: 795/5089, 15.62%). These differences suggest potential confounders and the necessity of adjusted regression.

We also visualized the unadjusted trends of digital health technology use in 2017-2022 by various age groups in Figure 1. Compared with adults and middle-aged adults, older adults significantly underutilized digital tools for health-related activities. Entering the COVID-19 pandemic, all 3 groups exhibited increased use of digital health technology to some extent; however, slightly wider gaps between older adults and the other 2 groups were observed regarding communicating with health providers, looking up test results, making appointments, using health apps and using wearable devices to track health. The age-based disparity only became smaller in using the internet to look for health information. Unadjusted regression results in Table S1 in Multimedia Appendix 1 correspond with our observation in Figure 1.

XSL•FO
RenderX

Table 1. Characteristics of the sample by age groups. All statistics are weighted to reflect the complex sampling strategy of the survey. Sample size may vary for outcome variables due to missing values. Source: authors' own analyses of the 2017-2022 Health Information National Trends Survey.

Variables	Full sample (N=15,505)	Adults (18-44 years; n=4361)	Middle-aged (45-64 years; n=6055)	Older adult (≥65 years; n=5089)
Outcomes, n (%)				
Look for health information	11,529 (74.81)	3628 (83.40)	4474 (74.32)	2790 (55.35)
Communicate with health provider	6825 (44.32)	2058 (47.31)	2763 (45.94)	1680 (33.35)
Look up test results	6649 (43.26)	1932 (44.52)	2715 (45.20)	1793 (35.68)
Make appointments	5807 (45.47)	1943 (53.45)	2136 (43.36)	1328 (31.55)
Use health apps	3073 (50.71)	1064 (60.07)	1115 (50.97)	650 (30.98)
Use wearable devices to track health	3394 (32.89)	1241 (42.04)	1174 (30.39)	628 (17.91)
Control covariates				
Demographics				
Race and ethnicity, n (%)				
Non-Hispanic White	10,007 (64.54)	2486 (57)	4004 (66.13)	4011 (78.82)
Hispanic	2468 (15.92)	884 (20.27)	879 (14.52)	447 (8.79)
Non-Hispanic Black	1642 (10.59)	450 (10.31)	736 (12.16)	386 (7.58)
Other non-Hispanic	1388 (8.95)	542 (12.42)	435 (7.18)	245 (4.81)
Sex, n (%)				
Male	7771 (50.12)	2225 (51.02)	3058 (50.51)	2394 (47.04)
Female	7734 (49.88)	2136 (48.98)	2997 (49.49)	2695 (52.96)
Marital status, n (%)				
Married	8689 (56.04)	1918 (43.97)	4052 (66.92)	3032 (59.58)
Unmarried	6816 (43.96)	2443 (56.03)	2003 (33.08)	2057 (40.42)
Household size, mean (SD)	3.11 (0.17)	3.77 (0.39)	2.90 (0.03)	2.05 (0.02)
Socioeconomic				
Education level, n (%)				
Less than high school	1,081 (6.97)	265 (6.07)	440 (7.26)	432 (8.49)
High school	3,237 (20.88)	770 (17.66)	1,355 (22.37)	1,278 (25.11)
Some college	5,959 (38.43)	1,618 (37.11)	2,387 (39.43)	1,999 (39.28)
College graduate or more	5,227 (33.71)	1,708 (39.16)	1,874 (30.95)	1,381 (27.13)
Household annual income, n (%)				
Less than US $20,000	2,380 (15.35)	692 (15.86)	850 (14.04)	874 (17.18)
US $20,000-US $34,999	1,724 (11.12)	468 (10.73)	550 (9.09)	852 (16.75)
US $35,000-US $49,999	2,008 (12.95)	593 (13.59)	651 (10.75)	841 (16.52)
US $50,000-US $74,999	2,841 (18.32)	767 (17.58)	1,100 (18.16)	1,043 (20.49)
US $75,000 or more	6,552 (42.26)	1,842 (42.24)	2,904 (47.96)	1,479 (29.06)
Rurality, n (%)				
Nonmetropolitan	2002 (12.91)	477 (10.94)	812 (13.41)	837 (16.44)
Metropolitan area	13,503 (87.09)	3884 (89.06)	5243 (86.59)	4252 (83.56)
Census region, n (%)				
Northeast	2741 (17.68)	755 (17.31)	1053 (17.39)	978 (19.21)
Midwest	3304 (21.31)	942 (21.61)	1332 (22)	967 (19.01)
South	5816 (37.51)	1554 (35.63)	2321 (38.34)	2040 (40.09)
West	3644 (23.50)	1109 (25.44)	1348 (22.27)	1103 (21.68)

XSL•FO
RenderX

Variables	Full sample (N=15,505)	Adults (18-44 years; n=4361)	Middle-aged (45-64 years; n=6055)	Older adult (≥65 years; n=5089)
Health				
Health insurance status, n (%)				
Insured	14,105 (90.97)	3765 (86.33)	5594 (92.39)	5024 (98.73)
Uninsured	1400 (9.03)	596 (13.67)	461 (7.61)	65 (1.27)
General health status, n (%)				
Poor	340 (2.19)	70 (1.61)	150 (2.48)	149 (2.92)
Fair	1989 (12.83)	445 (10.21)	831 (13.73)	867 (17.03)
Good	5399 (34.82)	1422 (32.61)	2193 (36.22)	1873 (36.80)
Very good	5755 (37.12)	1730 (39.68)	2165 (35.75)	1740 (34.19)
Excellent	2020 (13.03)	693 (15.89)	715 (11.81)	461 (9.06)
BMI, mean (SD)	28.62 (0.10)	27.99 (0.20)	29.46 (0.11)	28.17 (0.11)
Chronic disease, n (%)				
Diabetes	2513 (16.21)	295 (6.76)	1239 (20.47)	1470 (28.89)
Hypertension	5462 (35.23)	719 (16.48)	2538 (41.92)	3278 (64.41)
Heart conditions	1116 (7.20)	126 (2.88)	418 (6.91)	927 (18.22)
Chronic lung diseases	1775 (11.45)	416 (9.54)	703 (11.61)	795 (15.62)
Depression	3974 (25.63)	1259 (28.87)	1567 (25.88)	880 (17.29)
Mental health scores, mean (SD)	13.75 (0.04)	13.33 (0.10)	13.92 (0.06)	14.36 (0.06)
Observations, n	15,505	4361	6055	5089

Figure 1. Unadjusted trends of digital health technology use by age groups: (A) look for health information, (B) communicate with health providers, (C) look up test results, (D) make appointments, (E) use health apps, and (F) use wearable devices to track health. All statistics are weighted to reflect the complex sampling strategy of the survey. Source: authors' own analyses of the 2017-2022 Health Information National Trends Survey.

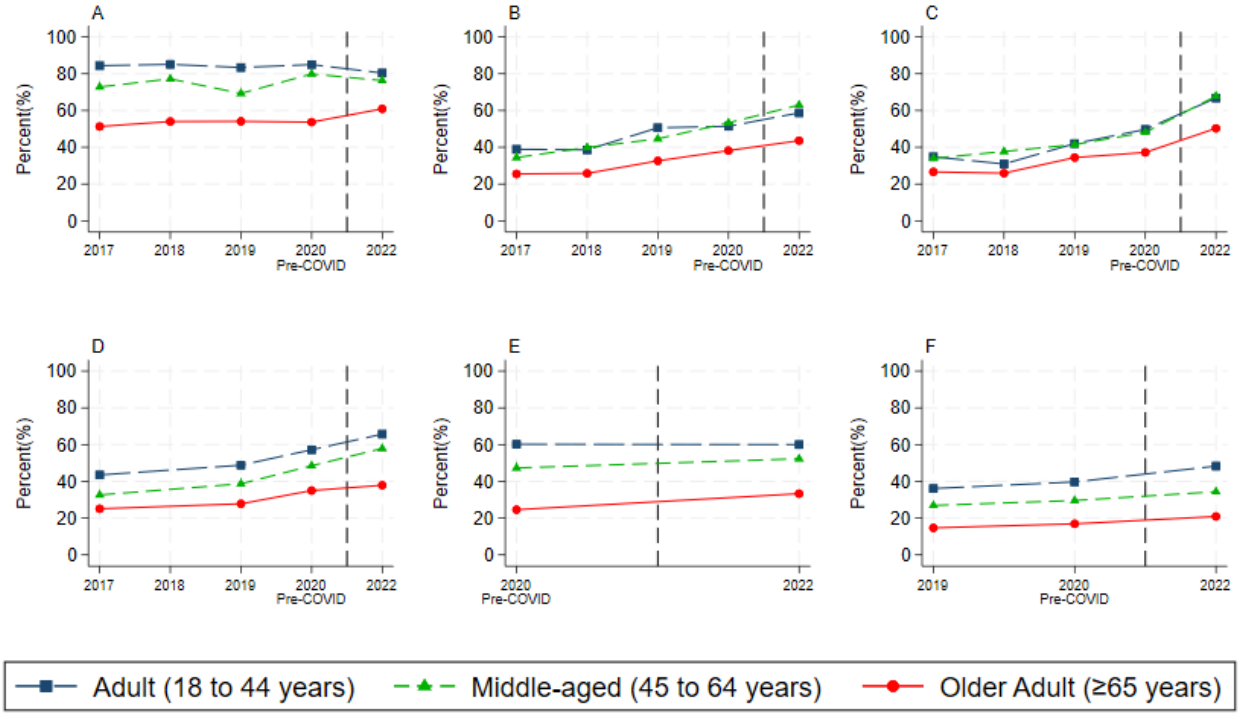

XSL•FO
RenderX

Regression Results

Table 2 reports coefficients of the COVID-19 indicator, age groups, and their interactive terms in equation 1 accounting for confounders in Table 1. Compared with adults, older adults showed greater reluctance against digital health technology, followed by middle-aged adults. Older adults were 26.28% less likely to use the internet to look for health information (*P*<.001), 13.38% less likely to communicate with health providers through the internet (*P*<.001), 7.63% less likely to use the internet to look up test results (*P*<.001), 14.97% less likely to use internet for making appointments (*P*<.001), 32.63% less likely to use health apps (*P*<.001), and 18.31% less likely to use wearable devices to track health (*P*<.001). Middle-aged adults had a lower probability of using the internet for health information seeking (–10.12%; *P*<.001), health care provider communication (–5.22%; *P*=.008), appointment-making (–9.01%; *P*<.001), and using wearable devices to track health (–9.79%; *P*=.001). The magnitudes are only half of the coefficients of older adults.

When the pandemic started, adults used digital health technology with higher probability: adults' use of the internet to communicate with health care providers increased by 18.78% (*P*<.001), using internet to look up test results increased by 32.38% (*P*<.001), using internet to make appointments increased by 21.28% (*P*<.001), and using wearable devices to track health increased by 11.4% (*P*<.001). However, there were no significant changes in using health apps (*P*=.64), and there were only marginal increases in using the internet to look for health information (*P*=.07), although not statistically significant.

The gaps between older adults and adults after the onset of the pandemic were significantly wider for using the internet to look up test results (–11.21%; *P*<.001), using internet to make appointments (–10.03%; *P*=.006), and using wearable devices to track health (–8.31%; *P*=.01). Conversely, the gap in using the internet to look for health information narrowed (9.64%; *P*=.002). The gaps between middle-aged adults and adults became narrower in both using the internet to look for health information (6.50%; *P*=.049) and communicate with health providers (7.60%; *P*=.05).

With regard to control covariates in the regression model shown in Table S2 in Multimedia Appendix 1, marital status, education level, household annual income, health insurance status, and depression were positively associated with digital health technology use, while male sex and Hispanic ethnicity were negatively associated.

The adjusted trends predicted from our regression models were much more interpretable (shown in Figure 2). We made three observations. First, regarding the probability of digital health technology use, the adults ranked highest, followed by middle-aged adults, and older adults ranked the lowest. Second, following the outbreak of the pandemic, we witnessed a prominent growth in digital health technology use among all 3 groups. Third, digital health technology use among adults increased at a more rapid pace than among older adults, so age-based disparities in digital health technology use turned out to widen during the pandemic. However, the disparities were narrowed between middle-aged adults and adults, especially for using the internet to look for health information and communicate with health providers.

Table 2. Regressions of COVID-19 pandemic influence on age-based disparities in health technology use. Regression models were estimated by linear probability model regression, and SE derived from the jackknife replication method are shown in parentheses. All linear probability model coefficients were scaled by 100 for readability. In each regression, we controlled for the demographic, socioeconomic, and health covariates shown in in addition to year fixed effects. Source: authors' own analyses of the 2017-2022 Health Information National Trends Survey.

	Look for health information		Communicate with health providers		Look up test results		Make appointments		Use health apps		Use wearable devices to track health	
	Value	*P* value	Value	*P* value	Value	*P* value	Value	*P* value	Value	*P* value	Value	*P* value
COVID-19, β (SE)	–4.46 (2.39)	.07	*18.78*[a] (3.10)	<.001	*32.38* (2.71)	<.001	*21.28* (3.07)	<.001	2.48 (5.31)	.64	*11.40* (2.96)	<.001
Age group (reference=adults), β (SE)												
Middle-aged adults	*–10.12* (1.56)	<.001	*–5.22* (1.88)	.008	*–1.55* (1.62)	.34	*–9.01* (2.47)	.001	–12.76 (7.67)	.10	*–9.79* (2.75)	.001
Older adults	*–26.28* (1.94)	<.001	*–13.38* (2.09)	<.001	*–7.63* (1.89)	<.001	*–14.97* (2.73)	<.001	*–32.63* (7.35)	<.001	*–18.3* (2.48)	<.001
COVID-19 × age group (reference=adults), β (SE)												
Middle-aged adults	*6.50* (3.23)	.049	*7.60* (3.82)	.05	–0.41 (3.36)	.90	2.39 (3.66)	.52	2.13 (8.07)	.79	–5.48 (3.64)	.14
Older adults	*9.64* (2.93)	.002	–2.57 (3.62)	.48	*–11.21* (3.32)	<.001	*–10.03* (3.50)	.006	4.58 (7.06)	.52	*–8.31* (3.19)	.01
Observations, n	15,411		15,399		15,370		12,771		6,059		10,318	
R^2	.19		.18		.19		.14		.17		.13	

[a]Significant results (*P*<.05) are shown in italics.

XSL•FO
RenderX

Figure 2. Predicted trends of digital health technology use by age groups: (A) look for health information, (B) communicate with health providers, (C) look up test results, (D) make appointments, (E) use health apps, and (F) use wearable devices to track health. All statistics are weighted to reflect the complex sampling strategy of the survey. Values are predicted by a linear probability model with SE derived from the jackknife replication method. In each regression, we control for the demographic, socioeconomic, and health covariates shown in Table 1, in addition to year fixed effects. Source: authors' own analyses of the 2017-2022 Health Information National Trends Survey.

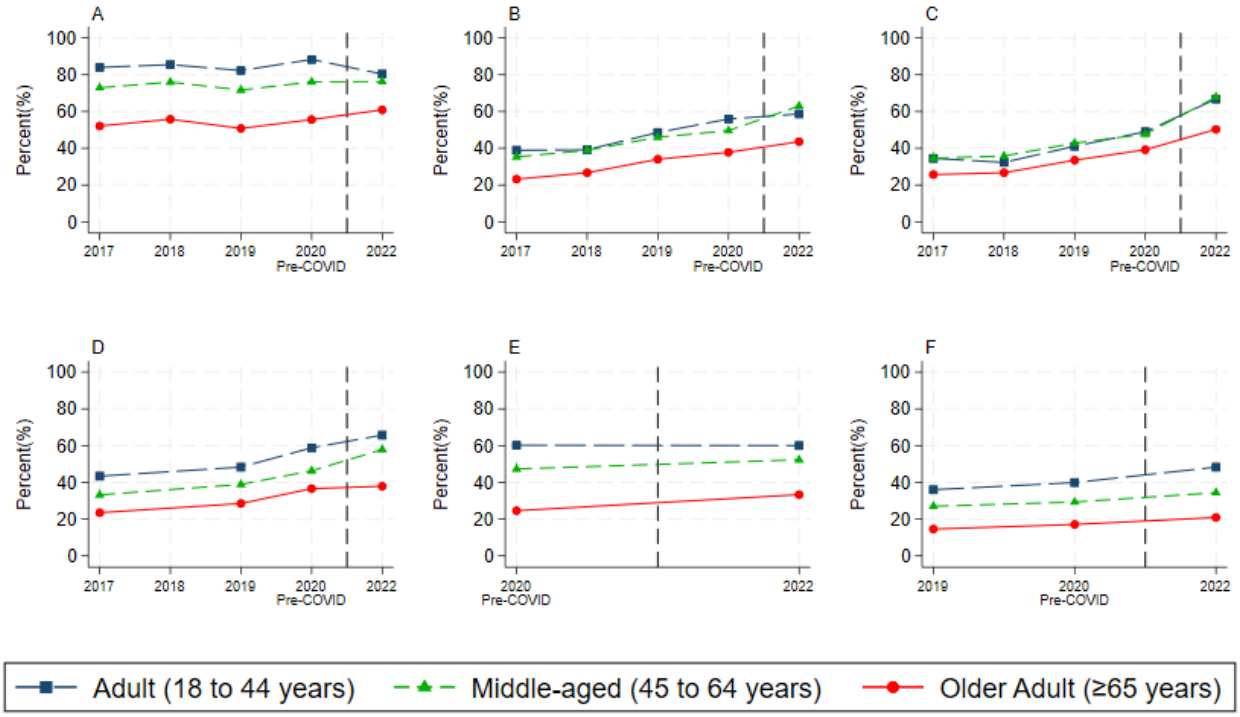

Additional Analysis

We conducted the subgroup analyses based on race and ethnicity, sex, marital status, and rurality. Results shown in Tables S3-S6 in Multimedia Appendix 1 generally support the consistent patterns across various demographic groups except for the rurality, due to the small sample size of nonmetropolitan populations. However, non-White, male, and married older adult groups exhibited greater disparities in using the internet to look up test results and using wearable devices to track health conditions after the pandemic. Meanwhile, among non-White, male, and nonmarried groups, the transitions between the negative main effect (ie, the coefficients of age group) and the positive interactive effect (ie, the coefficients of COVID-19 × age group) were significantly larger than our main estimates, suggesting that the age-based disparities in looking for health information through internet became noticeably smaller in such groups. It indicates the complexity of digital health technology use and the potentially heterogeneous effects of certain demographic factors on digital health technology use among older adults, which should capture the attention of future researchers.

We conducted 3 sensitivity analyses to ensure that our main estimates are robust to model specification and sample selection. First, we tested the robustness of the LPM model relative to the logit model. Although the outcomes were binary, we chose LPM instead of the logit model due to potential concerns of interactive terms estimated by the logit model [49,50]. However, we still estimated the logit model and compared their signs with those of our main estimates, and Table S7 in Multimedia Appendix 1 suggests both estimated effects have the same directions. Second, we were concerned about the issue that the 2020 sample spanned across the outbreak of the pandemic, so we tried 2 approaches, either eliminating the 2020 sample from the regression model (Table S8 in Multimedia Appendix 1) or including the 2020 post-pandemic–onset sample (Table S9 in Multimedia Appendix 1). The results are consistent with our main estimates with slightly different magnitudes. Finally, as the age-based disparity in digital health technology uses results from objective barriers (eg, bad internet connections or no digital devices) [11-13] and subjective barriers such as health literacy [14], we excluded the sample without internet connections or digital devices to isolate the effect of objective barriers. Results in Table S10 in Multimedia Appendix 1 show that the significant effects of the pandemic on reducing age-based disparity among middle-aged adults disappeared but the effects among older adults remained significant and unchanged.

Discussion

Principal Findings

Our study explores the pre-post changes of digital health technology use driven by the COVID-19 pandemic among different generations by analyzing the nationally representative cross-sectional sample across 2017-2022. We found that (1) the usage of digital health technology for middle-aged adults and older adults was at disproportionately lower rates than younger cohorts, (2) the usage of digital health technology for all age groups had abruptly increased after the onset of the pandemic,

XSL•FO
RenderX

and (3) age-based digital disparities significantly widened between older adults and adults in using the internet to look up test results or make appointments and using wearable devices to track health conditions.

Our study makes significant contributions to the literature on health disparities and the digital divide. There have been debates about whether the perceived health risk associated with the pandemic overpowers the existing objective or psychological disadvantages among vulnerable populations, thereby reducing the age-based disparities in digital health technology use. Using long-period nationwide datasets across the prepandemic era, the onset of the pandemic, and the late pandemic period, we provide the latest empirical evidence showing that age-based disparities expand due to the pandemic, which echoes previous literature that the digital divide is increasing during the pandemic around the globe [17,24,51]. Our study also exhibits the complexities of age-based disparities in digital health technology use, highlighting that the disparities of online health information seeking have narrowed, but disparities of advanced use of digital health technology have widened. Thus, it is needed to delve into the nuanced aspects of digital health technology use in future research.

Compared with the widening age-based disparities in using the internet to look up test results or using wearable devices to track health conditions, the expanding disparities in using the internet to make appointments should raise the attention of policy makers and professionals. Web-based medical appointment system is deemed as an advanced informational technology application with several merits, such as achieving patient-centered services, improving patient satisfaction, curbing no-show rates, and reducing waiting time [52]. In contrast to traditional methods, such as phone calls and in-person visits, web-based appointment systems have higher success rates and reduced waiting times due to after-hour access and enhanced scheduling transparency. However, considering the limited availability of time slots, older adults may be in a disadvantageous position when competing for health care service appointments against younger generations, as they rely more on the traditional appointment approaches as indicated by our findings. Furthermore, the usage of digital health technology is unevenly distributed among older adults [53]: our additional analysis in Table S11 in Multimedia Appendix 1 suggests that education, income, and rurality serve as significant predictors of limited adoption of web-based medical appointment systems in older adult population. Therefore, as the age-based disparity increases over time, older adults, especially those with socioeconomic vulnerability, may encounter restricted access to health care services when needed [54]. As US hospitals enthusiastically adopt digital health care systems [55], it is imperative that health care providers address the needs of older adults who are marginalized and left behind in this digital age.

The only reduced disparity between older adults and younger populations is the disparity in seeking health information through the internet. One factor is the unchanged trend of using the internet for health information among adults. Before the pandemic, we observed the possibility of adults seeking health information online had already reached the maximum level of about 80%, and even the pandemic hardly raised this possibility. Another attribute factor is that the pandemic not only increased the level of perceived risk but also interrupted in-person health care delivery. As older adults had substantial needs for health information, which have been previously addressed through in-person communication with health care providers, they were forced to seek such information online alternatively. Notably, despite the promising trend, only 50% of older adults obtain the necessary health-related information through the internet.

Another important finding of our study is that the often-neglected middle-aged adults, who grow up with digital technology and are supposed to have a higher level of digital literacy than the older adults, are also found to be suffering from lower rates of digital health technology use to some extent. Existing studies point out that not only older adults but also middle-aged adults can hardly use digital tools to access health information [56,57]. Our study shows that middle-aged adults have increasingly used digital health technology for searching for health information and messaging doctors without any significant differences in the adult population, but they still did not substantially use digital health technology to make appointments or track health activities through wearable devices. This means that a large group of middle-aged adults have health care demands and get access to digital health technology, but they still lack instructions and supportive services for using digital health technology to gain effective health care services and achieve better health management. Therefore, digital literacy programs, the design of digital health technology, and digital resources that previously targeted older adults should also take into account middle-aged adults.

In our sensitivity analysis, we tried to eliminate the effect of objective obstacles such as broadband, internet access, or digital devices, but the estimates remained similar, suggesting the change of age-based disparity in digital health technology use results from subjective barriers rather than infrastructure conditions. Scholars have identified several psychological factors of the digital divide, including digital literacy proficiency [58,59], technology anxiety [20,60-62], wariness of cybercrime or privacy violation, or fear of social labels as obsolete [63]. In addition, user-unfriendly interfaces may render low readability of text and usability or apps for older adults [59,64], and these issues might be more serious during the pandemic when digital interventions were intensively introduced without sufficient consultation and evaluation.

Policy and Practical Implications

During the COVID-19 pandemic, federal and state governments have made substantial efforts to improve internet connectivity [65]. For example, the Affordable Connectivity Program enabled low-income households to afford broadband [66], improved telehealth accessibility, and expanded residential broadband infrastructure. Those endeavors are critical to reach unconnected Americans, especially those in rural, remote, and underserved communities. However, additional approaches and strategies are necessary to address older adults' psychological concerns related to digital health technology usage as this study indicates.

We propose 3 approaches to increasing digital health technology usage among older adults from the perspectives of local community engagement, technology redesign, and governmental

XSL•FO
RenderX

support. First, local community and health care systems could enhance the education programs for digital literacy for older populations. For example, face-to-face training can be delivered at community-based health centers to help older adults install and use digital health technology. Second, more older adult–friendly interfaces or apps should be designed to relieve the operational burdens for such adults to use digital health technology. Researchers start to emphasize the variation of user demands and preferences during the development of digital health tools, and take the needs and preferences of such vulnerable populations into consideration, such as the user-centered design approach [67] or a participatory digital co-design approach [68]. Third, Medicare or state Medicaid programs could consider funding some projects to encourage health care providers to develop and implement interventions aimed at enhancing the use of digital health technology among older adults. For example, projects can invest in robust telehealth infrastructure to assist health care providers in easily reaching out to older adults, launch health care awareness campaigns to educate older adults on the potential benefits of using digital health technologies, and develop health education protocols and offer practical training for health care providers regarding how to communicate with older adults about using digital health tools. These approaches can help ensure these vulnerable populations are not left behind by the digitalized health care system and are adequately prepared for future public health crises.

Limitations

This study has several limitations. First, this study is a correlation study instead of a causality study. Although we included several control covariates to alleviate the confounder issue, there may be other unobservable confounding factors that bias our estimates, so readers should interpret our findings as associative instead of causal. Second, our analytic sample comes from HINTS collected by the NCI, which inevitably includes the disproportionate older population and oversamples of ethnic minority and rural populations [69]. We have weighted our inference statistics using the sample weights suggested by HINTS, but further research may use other datasets (eg, National Health Interview Survey) to re-examine the changes of age-based disparities and ensure our findings are not distorted by the sampling strategy and selected sample. Third, restricted by the HINTS survey item, we failed to examine the change of frequency of digital health technology use related to the COVID-19 pandemic. Finally, the reasons behind the low use rate for middle-aged adults and older adults are still unclear; thereby, it is difficult to develop detailed approaches for alleviating the age-based disparities. Future studies can undertake comprehensive interviews or detailed surveys to examine the underlying psychological, socioeconomic, and technological factors contributing to the disparities in the use of digital health technology.

Conclusions

According to the World Health Organization [70], the proportion of global older adults will nearly double from 12% to 22% between 2015 and 2050. Although information technology adoption in the US health care system enhances patient engagement, improves health care outcomes, and increases patient safety, facilitating digital inclusion for the aging baby boomers is still a challenging task. Our study is the first work that systematically examines the changes in age-based disparities in digital health technology use over a long period spanning the pandemic. We find that although the gap in health information seeking through the internet narrowed during the pandemic, being capable of using digital health technology to access health care providers and services remained a challenge. This empirical evidence implies that older adults may be marginalized by the US digitalized health care systems during public health crises such as the COVID-19 pandemic, and therefore health policy makers and health care providers should develop novel approaches to encourage the adoption and use of digital health technology among these vulnerable populations.

Acknowledgments

This study is not funded by any research project. The content is solely the responsibility of the authors.

Data Availability

The original datasets can be downloaded from the publicly available Health Information National Trends Survey. The datasets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information and sensitivity analysis.
[DOCX File , 134 KB - jmir_v26i1e65541_app1.docx]

References

1. What is digital health? US Food & Drug Administration. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2024-08-12]

XSL•FO
RenderX

2. Inkster B, O'Brien R, Selby E, Joshi S, Subramanian V, Kadaba M, et al. Digital health management during and beyond the COVID-19 pandemic: opportunities, barriers, and recommendations. JMIR Ment Health 2020;7(7):e19246 [FREE Full text] [doi: 10.2196/19246] [Medline: 32484783]
3. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022;5(1):13. [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
4. Chiesa V, Antony G, Wismar M, Rechel B. COVID-19 pandemic: health impact of staying at home, social distancing and 'lockdown' measures-a systematic review of systematic reviews. J Public Health (Oxf) 2021;43(3):e462-e481 [FREE Full text] [doi: 10.1093/pubmed/fdab102] [Medline: 33855434]
5. Webster P. Virtual health care in the era of COVID-19. The Lancet 2020;395(10231):1180-1181. [doi: 10.1016/s0140-6736(20)30818-7]
6. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020;22(6):e19284. [doi: 10.2196/19284] [Medline: 32501804]
7. R Niakan Kalhori S, Bahaadinbeigy K, Deldar K, Gholamzadeh M, Hajesmaeel-Gohari S, Ayyoubzadeh SM. Digital health solutions to control the COVID-19 pandemic in countries with high disease prevalence: literature review. J Med Internet Res 2021;23(3):e19473 [FREE Full text] [doi: 10.2196/19473] [Medline: 33600344]
8. Brodie M, Flournoy RE, Altman DE, Blendon RJ, Benson JM, Rosenbaum MD. Health information, the internet, and the digital divide. Health Aff (Millwood) 2000;19(6):255-265. [doi: 10.1377/hlthaff.19.6.255] [Medline: 11192412]
9. Feng Y, Xie W. Digital divide 2.0: the role of social networking sites in seeking health information online from a longitudinal perspective. J Health Commun 2015;20(1):60-68. [doi: 10.1080/10810730.2014.906522] [Medline: 25119019]
10. Eruchalu CN, Pichardo MS, Bharadwaj M, Rodriguez CB, Rodriguez JA, Bergmark RW, et al. The expanding digital divide: digital health access inequities during the COVID-19 pandemic in New York City. J Urban Health 2021;98(2):183-186 [FREE Full text] [doi: 10.1007/s11524-020-00508-9] [Medline: 33471281]
11. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021;4(1):52 [FREE Full text] [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
12. Zeng B, Rivadeneira NA, Wen A, Sarkar U, Khoong EC. The impact of the COVID-19 pandemic on internet use and the use of digital health tools: secondary analysis of the 2020 Health Information National Trends Survey. J Med Internet Res 2022;24(9):e35828 [FREE Full text] [doi: 10.2196/35828] [Medline: 36041005]
13. Khoong EC, Rivadeneira NA, Hiatt RA, Sarkar U. The use of technology for communicating with clinicians or seeking health information in a multilingual urban cohort: cross-sectional survey. J Med Internet Res 2020;22(4):e16951 [FREE Full text] [doi: 10.2196/16951] [Medline: 32250280]
14. Karnoe A, Furstrand D, Christensen KB, Norgaard O, Kayser L. Assessing competencies needed to engage with digital health services: development of the eHealth literacy assessment toolkit. J Med Internet Res 2018;20(5):e178 [FREE Full text] [doi: 10.2196/jmir.8347]
15. Spanakis P, Peckham E, Mathers A, Shiers D, Gilbody S. The digital divide: amplifying health inequalities for people with severe mental illness in the time of COVID-19. Br J Psychiatry 2021;219(4):529-531 [FREE Full text] [doi: 10.1192/bjp.2021.56] [Medline: 35048887]
16. Beaunoyer E, Dupéré S, Guitton MJ. COVID-19 and digital inequalities: reciprocal impacts and mitigation strategies. Comput Human Behav 2020;111:106424 [FREE Full text] [doi: 10.1016/j.chb.2020.106424] [Medline: 32398890]
17. Song Y, Qian C, Pickard S. Age-related digital divide during the COVID-19 pandemic in China. Int J Environ Res Public Health 2021;18(21):11285 [FREE Full text] [doi: 10.3390/ijerph182111285] [Medline: 34769801]
18. Hong YA, Cho J. Has the digital health divide widened? Trends of health-related internet use among older adults from 2003 to 2011. J Gerontol B Psychol Sci Soc Sci 2017;72(5):856-863. [doi: 10.1093/geronb/gbw100] [Medline: 27558403]
19. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022;24(2):e34790. [doi: 10.2196/34790] [Medline: 35171099]
20. Frishammar J, Essén A, Bergström F, Ekman T. Digital health platforms for the elderly? Key adoption and usage barriers and ways to address them. Technol Forecast Soc Change 2023;189:122319. [doi: 10.1016/j.techfore.2023.122319]
21. Alexander GC, Tajanlangit M, Heyward J, Mansour O, Qato DM, Stafford RS. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. JAMA Netw Open 2020;3(10):e2021476 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.21476] [Medline: 33006622]
22. Mai F, Ko DG, Shan Z, Zhang D. The impact of accelerated digitization on patient portal use by underprivileged racial minority groups during COVID-19: longitudinal study. J Med Internet Res 2023;25:e44981 [FREE Full text] [doi: 10.2196/44981] [Medline: 37384810]
23. Győrffy Z, Boros J, Döbrössy B, Girasek E. Older adults in the digital health era: insights on the digital health related knowledge, habits and attitudes of the 65 year and older population. BMC Geriatr 2023;23(1):779 [FREE Full text] [doi: 10.1186/s12877-023-04437-5] [Medline: 38012565]
24. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021;11(10):e053440 [FREE Full text] [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]

XSL•FO
RenderX

25. Dykgraaf SH, Desborough J, Sturgiss E, Parkinson A, Dut GM, Kidd M. Older people, the digital divide and use of telehealth during the COVID-19 pandemic. Aust J Gen Pract 2022;51(9):721-724. [doi: 10.31128/ajgp-03-22-6358]
26. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
27. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. Inquiry 2022;59:469580221096272. [doi: 10.1177/00469580221096272] [Medline: 35471138]
28. Hartz J, Yingling L, Powell-Wiley TM. Use of mobile health technology in the prevention and management of diabetes mellitus. Curr Cardiol Rep 2016;18(12):130. [doi: 10.1007/s11886-016-0796-8] [Medline: 27826901]
29. Bender JL. The internet as a route to health information, support, and healthcare for cancer survivors. J Natl Compr Canc Netw 2020;18(11):1586-1588. [doi: 10.6004/jnccn.2020.7672] [Medline: 33152699]
30. Gordon NP, Hornbrook MC. Older adults' readiness to engage with eHealth patient education and self-care resources: a cross-sectional survey. BMC Health Serv Res 2018;18(1):220 [FREE Full text] [doi: 10.1186/s12913-018-2986-0] [Medline: 29587721]
31. Health IT. Healthy People 2030. URL: https://health.gov/healthypeople/objectives-and-data/browse-objectives/health-it [accessed 2024-07-10]
32. Cheshmehzangi A, Zou T, Zhang Y, Chen H, Su Z, Dawodu A, et al. Commentary: reflecting on the neglected digital divide barriers of telemedicine during COVID-19. Front Public Health 2022;10:915401 [FREE Full text] [doi: 10.3389/fpubh.2022.915401] [Medline: 35669758]
33. Bakhtiar M, Elbuluk N, Lipoff JB. The digital divide: How COVID-19's telemedicine expansion could exacerbate disparities. J Am Acad Dermatol 2020;83(5):e345-e346 [FREE Full text] [doi: 10.1016/j.jaad.2020.07.043] [Medline: 32682890]
34. National Cancer Institute. Health Information National Trends Survey. URL: https://hints.cancer.gov/ [accessed 2024-07-15]
35. Mahmood A, Kedia S, Wyant DK, Ahn S, Bhuyan SS. Use of mobile health applications for health-promoting behavior among individuals with chronic medical conditions. Digit Health 2019;5:2055207619882181. [doi: 10.1177/2055207619882181] [Medline: 31656632]
36. Shan R, Ding J, Plante TB, Martin SS. Mobile health access and use among individuals with or at risk for cardiovascular disease: 2018 Health Information National Trends Survey (HINTS). J Am Heart Assoc 2019;8(24):e014390. [doi: 10.1161/jaha.119.014390]
37. Medical Subject Headings (MeSH). National Library of Medicine. 2024. URL: https://www.ncbi.nlm.nih.gov/mesh/68009273 [accessed 2024-06-10]
38. Ye L, Chen Y, Cai Y, Kao Y, Wang Y, Chen M, et al. Gender differences in the nonspecific and health-specific use of social media before and during the COVID-19 pandemic: trend analysis using HINTS 2017-2020 data. J Health Commun 2023;28(4):231-240. [doi: 10.1080/10810730.2023.2193151] [Medline: 36942570]
39. Nouri SS, Adler-Milstein J, Thao C, Acharya P, Barr-Walker J, Sarkar U, et al. Patient characteristics associated with objective measures of digital health tool use in the United States: a literature review. J Am Med Inform Assoc 2020;27(5):834-841 [FREE Full text] [doi: 10.1093/jamia/ocaa024] [Medline: 32364238]
40. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
41. Ajayi KV, Wachira E, Onyeaka HK, Montour T, Olowolaju S, Garney W. The use of digital health tools for health promotion among women with and without chronic diseases: insights from the 2017-2020 Health Information National Trends Survey. JMIR Mhealth Uhealth 2022;10(8):e39520 [FREE Full text] [doi: 10.2196/39520] [Medline: 35984680]
42. Shan R, Sarkar S, Martin SS. Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. Diabetologia 2019;62(6):877-887. [doi: 10.1007/s00125-019-4864-7] [Medline: 30963188]
43. Khoong EC, Olazo K, Rivadeneira NA, Thatipelli S, Barr-Walker J, Fontil V, et al. Mobile health strategies for blood pressure self-management in urban populations with digital barriers: systematic review and meta-analyses. NPJ Digit Med 2021;4(1):114 [FREE Full text] [doi: 10.1038/s41746-021-00486-5] [Medline: 34294852]
44. McBeath KCC, Angermann CE, Cowie MR. Digital technologies to support better outcome and experience of care in patients with heart failure. Curr Heart Fail Rep 2022;19(3):75-108. [doi: 10.1007/s11897-022-00548-z] [Medline: 35486314]
45. Kates J, Stricker CT, Rising KL, Gentsch AT, Solomon E, Powers V, et al. Perspectives from patients with chronic lung disease on a telehealth-facilitated integrated palliative care model: a qualitative content analysis study. BMC Palliat Care 2024;23(1):103 [FREE Full text] [doi: 10.1186/s12904-024-01433-3] [Medline: 38637806]
46. Eisner E, Berry N, Bucci S. Digital tools to support mental health: a survey study in psychosis. BMC Psychiatry 2023;23(1):726 [FREE Full text] [doi: 10.1186/s12888-023-05114-y] [Medline: 37803367]
47. Stata version release 18. StataCorp. URL: https://www.stata.com/ [accessed 2024-11-14]
48. HINTS methodology reports. National Cancer Institute. URL: https://hints.cancer.gov/data/methodology-reports.aspx [accessed 2024-11-14]
49. Ai C, Norton EC. Interaction terms in logit and probit models. Economics Letters 2003;80(1):123-129. [doi: 10.1016/s0165-1765(03)00032-6]
50. Norton EC, Wang H, Ai C. Computing interaction effects and standard errors in logit and probit models. Stata Journal 2004;4(2):154-167. [doi: 10.1177/1536867x0400400206]

XSL•FO
RenderX

51. Esteban-Navarro MA, García-Madurga MA, Morte-Nadal T, Nogales-Bocio AI. The rural digital divide in the face of the COVID-19 pandemic in europe—recommendations from a scoping review. Informatics 2020;7(4):54. [doi: 10.3390/informatics7040054]
52. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
53. Tappen RM, Cooley ME, Luckmann R, Panday S. Digital health information disparities in older adults: a mixed methods study. J Racial Ethn Health Disparities 2022;9(1):82-92 [FREE Full text] [doi: 10.1007/s40615-020-00931-3] [Medline: 33415705]
54. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024;26:e51931 [FREE Full text] [doi: 10.2196/51931] [Medline: 38976870]
55. Chen J, Amaize A, Barath D. Evaluating telehealth adoption and related barriers among hospitals located in rural and urban areas. J Rural Health 2021;37(4):801-811 [FREE Full text] [doi: 10.1111/jrh.12534] [Medline: 33180363]
56. Heponiemi T, Jormanainen V, Leemann L, Manderbacka K, Aalto AM, Hyppönen H. Digital divide in perceived benefits of online health care and social welfare services: national cross-sectional survey study. J Med Internet Res 2020;22(7):e17616. [doi: 10.2196/17616] [Medline: 32673218]
57. Crouch E, Gordon NP. Prevalence and factors influencing use of internet and electronic health resources by middle-aged and older adults in a US health plan population: cross-sectional survey study. JMIR Aging 2019;2(1):e11451 [FREE Full text] [doi: 10.2196/11451] [Medline: 31518256]
58. Quialheiro A, Miranda A, Garcia M, Carvalho ACD, Costa P, Correia-Neves M, et al. Promoting digital proficiency and health literacy in middle-aged and older adults through mobile devices with the workshops for online technological inclusion (OITO) project: experimental study. JMIR Form Res 2023;7:e41873 [FREE Full text] [doi: 10.2196/41873] [Medline: 36753331]
59. Chee SY. Age-related digital disparities, functional limitations, and social isolation: unraveling the grey digital divide between baby boomers and the silent generation in senior living facilities. Aging Ment Health 2024;28(4):621-632. [doi: 10.1080/13607863.2023.2233454] [Medline: 37424361]
60. Anderson M, Perrin A. Tech adoption climbs among older adults. Pew Research Center. 2017. URL: https://www.pewresearch.org/internet/2017/05/17/tech-adoption-climbs-among-older-adults/ [accessed 2016-11-29]
61. Meuter ML, Ostrom AL, Bitner MJ, Roundtree R. The influence of technology anxiety on consumer use and experiences with self-service technologies. J Bus Res 2003;56(11):899-906. [doi: 10.1016/s0148-2963(01)00276-4]
62. Wang CH, Wu CL. Bridging the digital divide: the smart TV as a platform for digital literacy among the elderly. Behav Inf Technol 2021;41(12):2546-2559. [doi: 10.1080/0144929x.2021.1934732]
63. Wu YH, Damnée S, Kerhervé H, Ware C, Rigaud AS. Bridging the digital divide in older adults: a study from an initiative to inform older adults about new technologies. Clin Interv Aging 2015;10:193-200 [FREE Full text] [doi: 10.2147/CIA.S72399] [Medline: 25624752]
64. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020;180(10):1389-1391 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
65. de Wit K. States tap federal CARES Act to expand broadband. Pewtrusts. URL: https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2020/11/states-tap-federal-cares-act-to-expand-broadband [accessed 2024-05-10]
66. Affordable Connectivity Program. Federal Communications Commission. URL: https://www.fcc.gov/acp [accessed 2024-06-01]
67. Bonet-Olivencia S, Carrillo-Leal J, Rao A, Sasangohar F. User-centered design of a diabetes self-management tool for underserved populations. J Diabetes Sci Technol 2024;18(1):22-29. [doi: 10.1177/19322968231212220] [Medline: 37978811]
68. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020;7(4):e16460. [doi: 10.2196/16460] [Medline: 32243256]
69. Nelson D, Kreps G, Hesse B, Croyle R, Willis G, Arora N, et al. The Health Information National Trends Survey (HINTS): development, design, and dissemination. J Health Commun 2004;9(5):443-460. [doi: 10.1080/10810730490504233] [Medline: 15513791]
70. Ageing and health. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2024-08-01]

Abbreviations

HINTS: Health Information National Trends Survey
LPM: linear probability model
NCI: National Cancer Institute

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.08.24; peer-reviewed by Y Luo, R Guo; comments to author 17.09.24; revised version received 10.10.24; accepted 12.11.24; published 04.12.24.

Please cite as:
Qiu Y, Huang H, Gai J, De Leo G
The Effects of the COVID-19 Pandemic on Age-Based Disparities in Digital Health Technology Use: Secondary Analysis of the 2017-2022 Health Information National Trends Survey
J Med Internet Res 2024;26:e65541
URL: https://www.jmir.org/2024/1/e65541
doi:10.2196/65541
PMID:

©Yuanbo Qiu, Huang Huang, Junjie Gai, Gianluca De Leo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Dietary Intake Assessment Using a Novel, Generic Meal–Based Recall and a 24-Hour Recall: Comparison Study

Cathal O'Hara[1,2,3*], BSc; Eileen R Gibney[1,2,3*], PhD

[1]University College Dublin Institute of Food and Health, Science Centre South, University College Dublin, Dublin, Ireland
[2]Insight Centre for Data Analytics, University College Dublin, Belfield, Dublin, Ireland
[3]School of Agriculture and Food Science, University College Dublin, Belfield, Dublin, Ireland
[*]all authors contributed equally

Corresponding Author:
Eileen R Gibney, PhD
University College Dublin Institute of Food and Health
Science Centre South
University College Dublin
Belfield
Dublin, D04 N2E5
Ireland
Phone: 353 17162819
Email: eileen.gibney@ucd.ie

Abstract

Background: Dietary intake assessment is an integral part of addressing suboptimal dietary intakes. Existing food-based methods are time-consuming and burdensome for users to report the individual foods consumed at each meal. However, ease of use is the most important feature for individuals choosing a nutrition or diet app. Intakes of whole meals can be reported in a manner that is less burdensome than reporting individual foods. No study has developed a method of dietary intake assessment where individuals report their dietary intakes as whole meals rather than individual foods.

Objective: This study aims to develop a novel, meal-based method of dietary intake assessment and test its ability to estimate nutrient intakes compared with that of a web-based, 24-hour recall (24HR).

Methods: Participants completed a web-based, generic meal–based recall. This involved, for each meal type (breakfast, light meal, main meal, snack, and beverage), choosing from a selection of meal images those that most represented their intakes during the previous day. Meal images were based on generic meals from a previous study that were representative of the actual meal intakes in Ireland. Participants also completed a web-based 24HR. Both methods were completed on the same day, 3 hours apart. In a crossover design, participants were randomized in terms of which method they completed first. Then, 2 weeks after the first dietary assessments, participants repeated the process in the reverse order. Estimates of mean daily nutrient intakes and the categorization of individuals according to nutrient-based guidelines (eg, low, adequate, and high) were compared between the 2 methods. *P* values of less than .05 were considered statistically significant.

Results: In total, 161 participants completed the study. For the 23 nutrient variables compared, the median percentage difference between the 2 methods was 7.6% (IQR 2.6%-13.2%), with *P* values ranging from <.001 to .97, and out of 23 variables, effect sizes for the differences were small for 19 (83%) variables, moderate for 2 (9%) variables, and large for 2 (9%) variables. Correlation coefficients were statistically significant (*P*<.05) for 18 (78%) of the 23 variables. Statistically significant correlations ranged from 0.16 to 0.45, with median correlation of 0.32 (IQR 0.25-0.40). When participants were classified according to nutrient-based guidelines, the proportion of individuals who were classified into the same category ranged from 52.8% (85/161) to 84.5% (136/161).

Conclusions: A generic meal–based method of dietary intake assessment provides estimates of nutrient intake comparable with those provided by a web-based 24HR but with varying levels of agreement among nutrients. Further studies are required to refine and improve the generic recall across a range of nutrients. Future studies will consider user experience including the potential feasibility of incorporating image recognition of whole meals into the generic recall.

(J Med Internet Res 2024;26:e48817) doi:10.2196/48817

XSL•FO
RenderX

KEYWORDS

meal patterns; eating behaviors; eating occasions; nutrition assessment; dietary intake assessment; 24-hour recall; relative validity

Introduction

Well-established causal relationships exist between dietary intakes and health [1]. Accurate dietary intake assessment is required to identify suboptimal intakes and devise interventions to address them [2]. Existing food-based methods of dietary intake assessment can be time-consuming and burdensome for individuals to complete [3]. In addition, not all methods provide information such as the timing of meals, different foods that are consumed in combination as part of these meals, or combinations of different meals over a day; for example, the food frequency questionnaire (FFQ) focuses on mean daily food and nutrient intakes [4].

The time and effort required for individuals to complete the existing methods of dietary intake assessment such as FFQs, 24-hour recalls (24HRs), and food diaries may limit adherence to and engagement with such methods, which are often used on web-based and mobile-based personalized nutrition platforms [3]. A survey of 2382 adults across Europe found that ease of use was the most important feature for participants when choosing a nutrition or diet app [5]. Digital versions of 24HRs and food diaries require the user to text search for a food and then select from a list of results the food that they consumed. This process is then repeated for each food in the meal and each meal in the day [6]. Intakes of whole meals can be recorded in a manner that is less burdensome than recording individual foods, providing a potentially low-burden method for dietary intake assessment [3]. For example, instead of text searching for individual food, as is required in 24HRs and food diaries, the user could be presented with images of whole meals and choose the image most similar to their meal [3]. The use of meal-based methods may also be preferred in personalized nutrition because people tend to perceive their dietary intakes in terms of the meals they have consumed rather than their daily intakes of nutrients or foods; therefore, recording dietary intakes and providing dietary advice in this manner may be more intuitive [7,8].

Although the number of studies examining meal patterns has increased in recent years [4], only 3 studies [9-11] have developed meal-based methods of dietary intake assessment rather than using existing food-based methods. Englund-Ögge et al [9] used a method in which participants reported how often they consumed various meal types (breakfast, morning snack, lunch, afternoon snack, dinner, evening snack, supper, and night meal). Wilson et al [10] used a similar approach, but instead of using meal types, they divided the day into periods and asked participants to report for each period whether they consumed nothing, a snack, a small meal, or a large meal and whether they drank nothing, alcohol, water, or something else. These approaches to meal-based dietary assessment are simple to complete and provide qualitative information about meal types and their timing. They do not, however, provide the qualitative detail necessary to identify the different combinations of foods being consumed in meals and the combinations of those meals over a day or the quantitative detail required to estimate the nutrient intakes arising from those consumptions. Murakami et al [11] developed an approach that involves participants reporting the frequency of consumption of combinations of food groups and foods at specified meal types (breakfast, morning snack, lunch, afternoon snack, dinner, and night snack), and it has been designed for use in the Japanese population. This approach allows for the identification of meal patterns and nutrient intakes but still requires individuals to report intakes at the food level. None of those studies, however, allow for the reporting of meal portion sizes or capture information from the previous 24 hours in the form of recall.

Several tools have been developed that use image recognition software for dietary intake assessment [12]. However, these methods remain food based rather than meal based. The software first segments a meal image into its constituent foods and then provides a suggested match for each food in the image. The user must then confirm whether the suggested foods are correct. For any missing or incorrect foods, the user must text search for the correct food and add it to their record [12,13]. No study that allows individuals to record their dietary intakes at the meal level by reporting intakes of whole meals rather than individual foods or food groups has been identified.

This study aimed to develop a novel, meal-based method of dietary intake assessment that would allow individuals to report their intakes of whole meals rather than reporting the individual foods that make up those meals and to compare this method with a web-based 24HR.

Methods

Ethical Considerations

The human research ethics committee of University College Dublin granted ethics approval to conduct this study (LS-21-64-OHara-Gibney). Participants were assigned a study number, and on completion of data collection any information linking this number to participants' personal data was deleted, thus de-identifying participants. No financial compensation was provided for participation in this study. However, on completion of the study, all participants received a personalized nutrition advice report based on the data they provided during the 24HRs.

Recruitment

The target sample size was 160 participants, based on a previous review of comparisons between digital and paper-based 24HRs, which found a range of sample sizes from 53 to 167 [14]. There were no studies of meal-based methods of dietary assessment on which to base the sample size.

Recruitment was conducted using local radio, local newspaper, posters, social media, and word of mouth. Researchers directed potential participants to a web page containing full details of the study requirements and contact details of the researchers for further queries, if required. After reading the study information, potential participants could indicate whether they had read and understood the material and agree or disagree to

XSL•FO
RenderX

proceed to the web-based screening questionnaire to determine their eligibility to participate in the study. Those who were eligible then completed a web-based consent form to provide electronic informed consent. People were eligible if they were aged >18 years, were fluent in English, had regular access to the internet, and were not current or former students of a degree in nutrition or dietetics (Figure 1).

Figure 1. Flow diagram of participants' journey through the study. 24HR: 24-hour recall.

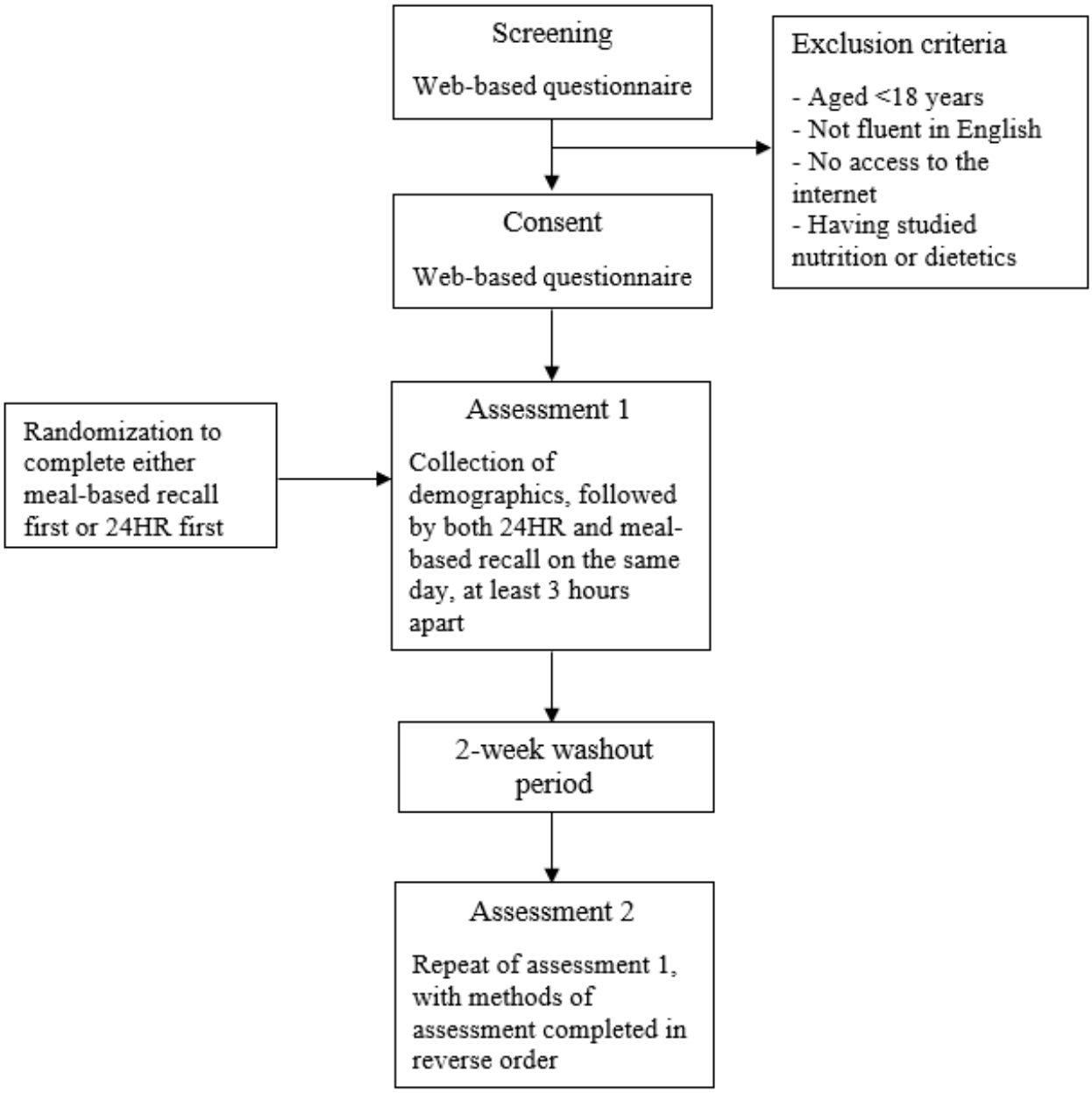

Study Design

Participants who were deemed eligible to participate and provided consent were contacted via email, which provided details about the next steps of the study and the links to the 2 web-based dietary intake assessment tools (24HR and generic meal–based recall). A crossover design was used with regard to the order in which participants completed the recalls. Participants were randomized to complete either one of the 2 methods first and then complete the second method at least 3 hours later on the same day. Participants were also randomized regarding whether they would recall a weekday or a weekend day. Then, 2 weeks after having completed the first set of recalls, participants completed the recalls again in the reverse order (compared with the order in which they were completed on the first occasion), followed by the completion of the evaluation questionnaire (Figure 1).

Overview of the Generic Meal–Based Dietary Recall

The meal-based dietary intake assessment method was administered using Qualtrics XM (Qualtrics International Inc), a web-based platform for questionnaires. Before completing the dietary intake assessment, participants provided information about their sex, age, weight, and height via a web-based questionnaire. Participants were then asked to select the meal types they had consumed on the previous day from the following list: breakfast, morning snack, lunch, afternoon snack, evening meal, evening snack, and beverage (for beverages consumed

XSL•FO
RenderX

alone without food). For each meal type selected, participants were presented with a series of images of generic meals that are associated with that meal type and asked to choose the meal image that was most similar to the meal that they had consumed (Figure 2). For each selected meal, participants were asked to choose from 3 different images of that meal, each representing a different portion size, and then asked to select whether the chosen image was smaller than the chosen portion, the same size as the portion chosen, or larger than the portion chosen (Figure 2). For beverages, participants were asked to choose from 3 different images of that beverage, representing 3 different portion sizes, and then asked to select the number of those portion sizes that they had consumed. For each meal type, participants could choose the option that none of the images presented were representative of their intake for that meal type, for example, "none of the above options are similar to what I ate for breakfast." If participants selected this option, a box appeared, in which they could enter a free-text description of what they had consumed for that meal. This allowed the researchers to determine whether a suitable generic meal could have been chosen or whether there was, in fact, no matching generic meal.

Figure 2. An example of the user interface of the generic meal–based recall, where participants are asked to choose the generic breakfast image that most represents their breakfast intake on the previous day, to specify the portion size that they had consumed for their chosen breakfast meal, and to answer the follow-up portion size question.

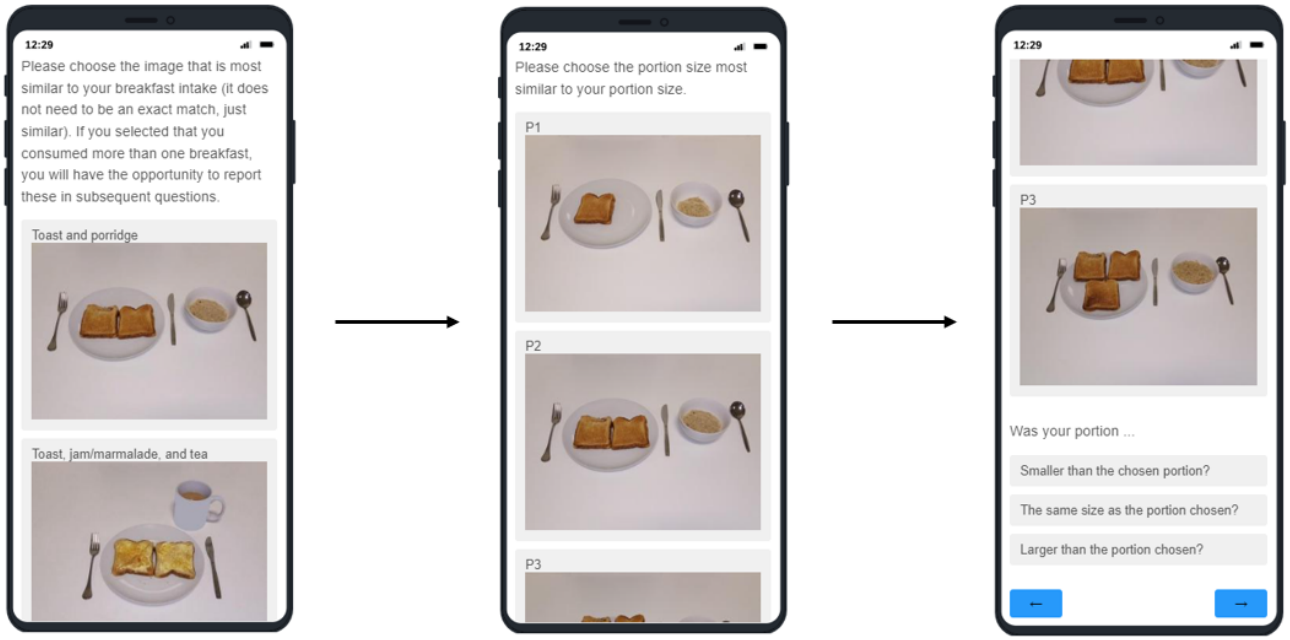

Development of the Meal-Based Dietary Recall

The process of identifying the generic meals that were presented to participants as images has been described in detail elsewhere [15]. In brief, data from the Irish National Adult Nutrition Survey (NANS; 2008-2010) [16] were used. This is a representative data set about dietary intakes of 1500 adults in Ireland, collected using 4-day weighed food diaries. The meals reported were categorized into the following meal types: breakfast, light meals, main meals, snacks, and beverages. A nutrient profiling score, the Nutrient Rich Foods (NRF) Index [17], was calculated for each meal; specifically, the NRF9.3 version of that profiling score was used. Within each meal type, meals were grouped (ie, clustered) using partitioning around the medoids clustering to identify groups of meals that had similar NRF Index scores and food group composition. These groups were defined as generic meals.

The 27,336 individual meals consumed by the participants in the NANS study were condensed to 63 generic meals; 49% (31/63) of these were consumed during the week and 51% (32/63) during the weekend. Among the 63 generic meals that were identified, there was overlap between weekday and weekend meals. That is, some weekday meals were the same as the weekend meals. When these duplicates were removed, participants were presented with 43 meal images: 5 (12%) breakfasts, 5 (12%) snacks (repeated for morning, afternoon, and evening snacks), 10 (23%) lunches, 19 (44%) dinners, and 4 (9%) beverages.

The nutrient content of a given generic meal was defined as the mean nutrient content of the individual meals that made up that generic meal per 100 g. Each generic meal was assigned 7 portion sizes by ordering each of the individual meals by weight and dividing the meals into 7 equal parts based on septile values for meal weight. The median weights for each part were assigned as the generic portion size for that meal. The nutrient composition for each of the portion sizes was calculated using the meal weight for that portion and the generic meal nutrient composition [15]. Within the meal intake assessment tool presented in this study, the second, fourth, and sixth portion sizes were used as the 3 portion size images shown to participants, with the options asking whether the image chosen was smaller than, the same as, or larger than that consumed, allowing participants to be categorized as consuming the first, third, fifth, or seventh portion size for a given meal.

24HR Method

Participants completed their 24HRs using a validated, web-based, self-administered 24HR tool called Foodbook24,

XSL•FO
RenderX

which follows the multipass recall method [18-20]. Participants first chose the meal types that they had consumed from the following list: breakfast, morning snack, lunch, afternoon snack, evening meal, and evening snack with the option to add additional snacks. For each of the selected meal types, participants added the foods and beverages they had consumed as part of those meal types by text searching from the food list using a search bar. Portion size was then reported based on the number of the food or beverage items consumed or from portion size photographs, as appropriate. Participants were then presented with the list of foods they had recorded for review, before being presented with a list of commonly forgotten foods. The food list contained food composition data from the McCance and Widdowson's Composition of Foods Integrated Dataset (CoFID) [21], with some additions relevant to dietary intakes in Ireland. The development of Foodbook24 and its food list is described in detail elsewhere [18,20].

Statistical Analysis

All analyses were performed using R (version 4.2.2, R Foundation for Statistical Computing) [22] in the RStudio integrated development environment (version 2022.07.2+576, Posit PBC) [23]. Data from 24HRs were used to identify participants likely to be misreporters of energy intake (EI), based on the ratio of estimated EI to basal metabolic rate (BMR; EI:BMR) using the BMR equations from Henry [24]. On the basis of the Goldberg equations [25], EI:BMR <0.96 was deemed indicative of underreporting, and EI:BMR >2.49 was indicative of overreporting. The analysis presented in this paper includes all participants (161/161, 100%), given the negligible differences observed when misreporters were removed; the analysis of the smaller data set with misreporters excluded is provided in Multimedia Appendix 1. *P* values of <.05 were considered statistically significant.

The Shapiro-Wilk test was used to determine whether the differences in the variables between methods were normally distributed and confirmed using visual inspection of histograms. Wilcoxon signed rank test was performed to compare nutrient intake estimates obtained from the web-based 24HR with those obtained from the generic meal–based recall. Wilcoxon effect size (*r*) was calculated. Effect size ≥0.1 and <0.3 was considered small, effect size ≥0.3 and <0.5 was considered moderate, and effect size ≥0.5 was considered large [26]. Bland-Altman analysis was also performed, whereby the mean difference between the 2 data sets and the limits of agreement (LOAs; mean difference – 1.96 SD to mean difference + 1.96 SD) for each nutrient were calculated. The correlation of nutrient intakes between the 2 methods was assessed using Spearman rank correlation coefficients. Correlation coefficient <0.20 was indicative of poor correlation, coefficient ≥0.20 and <0.50 was indicative of acceptable correlation, and coefficient ≥0.5 was indicative of good correlation [27].

Cross-classification of quartiles was performed for all nutrients (23/23, 100%) assessed. That is, nutrient intakes from both methods were divided into quartiles to determine the proportion of participants who remained in the same quartile for both methods (exact agreement), the proportion of participants who were classified in the same or adjacent quartiles (exact + adjacent), the proportion of participants who were classified 2 quartiles apart (disagreement), and the proportion of participants who were classified 3 quartiles apart (extreme disagreement). Participants were also classified according to nutrient-based dietary guidelines, separately for both methods [28-30]. For example, they were classified based on whether their nutrient intakes were low, adequate, or high according to those guidelines. The nutrients assessed were those deemed to be of public health relevance and included protein, carbohydrate, fat, monounsaturated fat, polyunsaturated fat, saturated fat, salt, dietary fiber, calcium, iron, folate, thiamin, riboflavin, and vitamin C. The proportion of individuals who were classified into the same category based on both methods was calculated for each nutrient.

Evaluation Questionnaire

Participants also completed an evaluation questionnaire at the end of the study, which was administered via Qualtrics XM (Qualtrics International Inc). Participants were asked to what extent they agreed that the meals presented in the meal-based dietary intake assessment tool were representative of what they consume, that the portion sizes presented were representative of what they consume, that the instructions were clear and easy to understand, and that overall, the tool was easy to use. The response options were agree, somewhat agree, somewhat disagree, or disagree. Participants were also asked how they would describe the ease of use of the meal-based tool compared with the 24HR with the response options being: better, somewhat better, somewhat worse, or worse. Finally, participants were asked to respond either yes or no as to whether they would consider using a similar tool to the meal-based tool in the future.

Results

Study Sample

A total of 161 participants completed both methods of dietary intake assessment at 2 time points. Most participants were female (131/161, 81.4%), the median age was 54 (IQR 39-63) years, and the median BMI was 25.3 (IQR 22.5-28.9) kg/m^2 (Table 1).

XSL•FO
RenderX

Table 1. Participant demographics and anthropometry[a].

Characteristics	Female participants (n=131, 81.4%), median (IQR)	Male participants (n=30, 18.6%), median (IQR)	Total (N=161), median (IQR)
Age (y)	54 (41-63)	54 (37-63)	54 (39-63)
Weight (kg)	67.9 (60.5-75.6)	80.8 (75.6-94.0)	70 (62-80.8)
Height (m)	1.63 (1.59-1.68)	1.8 (1.75-1.83)	1.65 (1.60-1.74)
BMI (kg/m^2)	25.2 (22.5-28.8)	26.0 (22.9-28.8)	25.3 (22.5-28.9)

[a]Weight and height were self-reported by participants, and BMI was subsequently calculated from the self-reported values.

Daily Nutrient Intakes

For the 23 variables compared, the percentage difference between the meal-based and 24HR methods ranged from 0% to 46.7%, with the median percentage difference being 7.6% and with the generic method providing a higher estimate than the 24HR for 18 nutrients. *P* values for the differences between the 2 methods ranged from <.001 to .97, with 13 (57%) of 23 comparisons reaching statistical significance ($P<.05$). Among the 23 variables, effect sizes for the differences were small for 19 (83%) variables, moderate for 2 (9%) variables (folate in µg and sodium in mg), and large for 2 (9%) variables (polyunsaturated fat in g and as % total EI [TEI]; Table 2).

XSL•FO
RenderX

Table 2. Median (IQR) daily nutrient intakes estimated using the web-based, 24-hour recall (24HR) and novel, generic meal–based recall[a,b].

Nutrient intake	Recall method, median (IQR)		Difference (%)	*P* value	Effect size (*r*)	Effect size (magnitude)
	24HR	Generic				
Energy (kcal)	1569.9 (1318.9-2005.7)	1715.3 (1440.9-1935.3)	9.3	.18	0.107	Small
Protein (g)	65.4 (51.4-86.8)	72.7 (63.2-82.1)	11.2	.69	0.032	Small
Protein (% TEI[c])	16.6 (13.8-20.1)	16.8 (15.8-18.3)	0.8	.45	0.060	Small
Carbohydrate (g)	192 (140.5-243.3)	202.8 (170.3-240)	5.6	.04	0.162	Small
Carbohydrate (% TEI)	44.5 (39.4-50.4)	45.5 (43.7-47.4)	2.3	.15	0.113	Small
Sugar (g)	71.8 (49.5-97.7)	84.8 (64.3-102.4)	18.1	.001	0.263	Small
Sugar (% TEI)	16.6 (12.8-21.4)	18.6 (16.8-21)	12.2	.002	0.249	Small
Dietary fiber (g)	17.2 (12.3-21.7)	16.7 (14.6-19.8)	−3.4	.24	0.093	Small
Total fat (g)	61.6 (47.1-77.7)	63.1 (52.8-73.5)	2.4	.89	0.011	Small
Total fat (% TEI)	34.7 (29.4-40.2)	33.8 (31.8-35.4)	−2.5	.03	0.169	Small
Saturated fat (g)	22 (15.7-30)	25.1 (21.3-29.3)	14.2	.01	0.202	Small
Saturated fat (% TEI)	12.3 (9.3-15)	13.5 (12.9-14.2)	9.8	.005	0.220	Small
Monounsaturated fat (g)	22.1 (15.4-29.3)	22.7 (19-26.3)	2.7	.97	0.004	Small
Monounsaturated fat (% TEI)	12.3 (10.4-14.7)	12.1 (11.2-12.8)	−1.5	.33	0.076	Small
Polyunsaturated fat (g)	7.6 (5.7-9.7)	11.1 (9.2-12.8)	46.7	<.001	0.598	Large
Polyunsaturated fat (% TEI)	4.2 (3.4-5)	5.9 (5.5-6.2)	41	<.001	0.723	Large
Vitamin D (μg)	2.2 (0.8-4.2)	2.3 (1.9-2.6)	7.6	.15	0.113	Small
Folate (μg)	215.1 (173.4-279.8)	198.1 (165.5-223.4)	−7.9	<.001	0.334	Moderate
Vitamin C (mg)	72.3 (32.4-123.3)	67.1 (55.1-80.2)	−7.2	.002	0.244	Small
Calcium (mg)	686 (509.4-884.2)	796.4 (670.3-944)	16.1	.002	0.250	Small
Iron (mg)	9.9 (7.7-12.6)	9.9 (8.1-12)	0	.13	0.119	Small
Potassium (mg)	2865.5 (2095.8-3462.8)	2716.2 (2200.3-3066.4)	−5.2	.01	0.204	Small
Sodium (mg)	1685.5 (1157-2142.6)	2161 (1853.1-2424.3)	28.2	<.001	0.398	Moderate

[a]*P* values were derived using Wilcoxon signed rank test, with $P<.05$ indicating statistical significance.

[b]Effect size ≥0.1 and <0.3 was considered small, effect size ≥0.3 and <0.5 was considered moderate, and effect size ≥0.5 was considered large [26].

[c]TEI: total energy intake.

Comparing the differences using the Bland-Altman analysis, the mean differences between the 2 methods for macronutrients were close to 0, whereas those for micronutrients were larger. The LOAs tended to be wide for all nutrients (23/23, 100%). The analysis identified 17 (74%) nutrients for which ≥95% of participants fell within the LOA. The proportion of individuals who fell within the LOA ranged from 92.5% (149/161) for polyunsaturated fats (% TEI) to 98.1% (158/161) for dietary fiber (g). Bland-Altman plots for energy and macronutrients are presented in Figure 3; values from the Bland-Altman analysis for the remaining nutrients are given in Table 3.

XSL•FO
RenderX

Figure 3. Bland-Altman plots of energy and macronutrient intake estimates. The middle, dashed line and associated number represents the mean difference in daily intakes between web-based 24-hour recall (24HR) and generic recall. The upper and lower dashed lines and associated numbers represent the upper and lower limits of agreement (LOAs), respectively. Each point represents an individual participant (N=161). TEI: total energy intake.

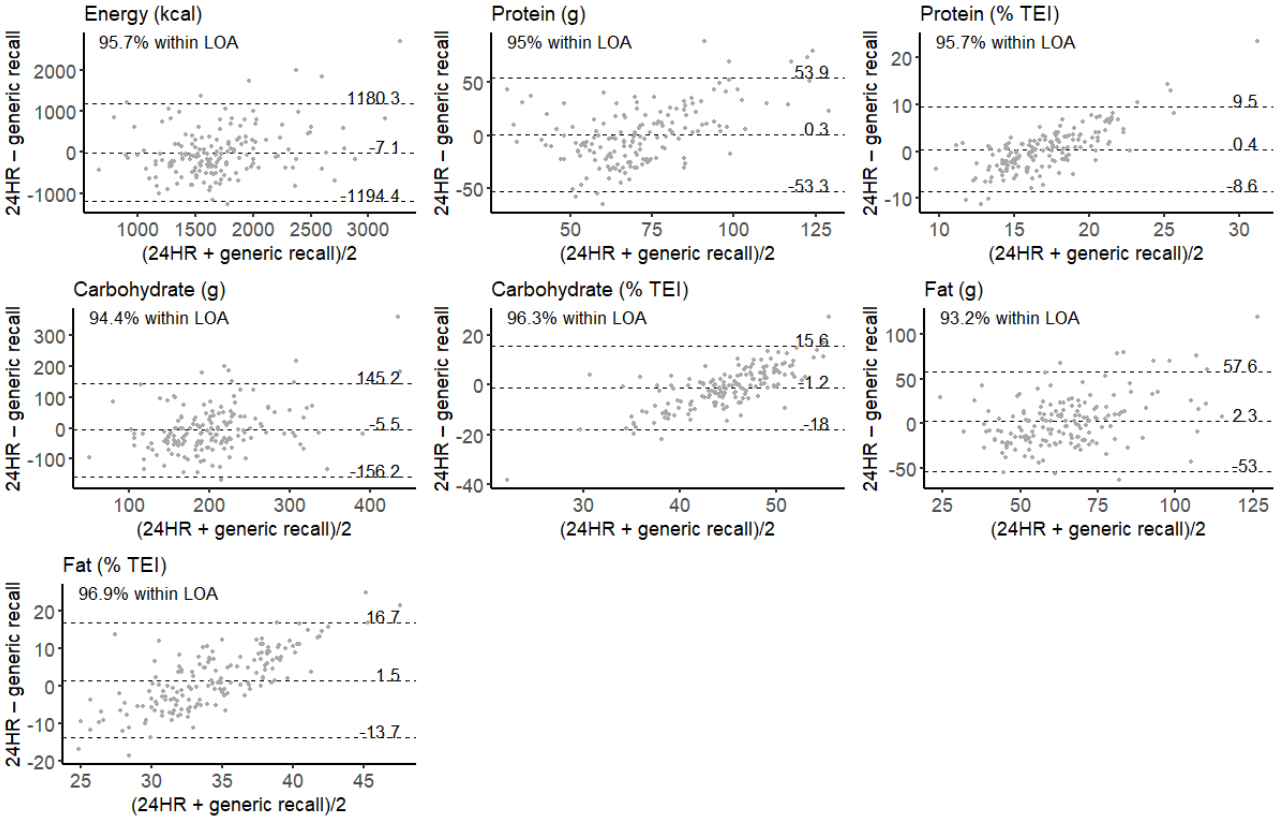

XSL•FO
RenderX

Table 3. Bland-Altman analysis of nutrient intake estimates obtained from the web-based, 24-hour recall and the generic recall[a].

Nutrient intake	Mean difference (SD)	Lower LOA[b]	Upper LOA	Participants within LOA (N=161), n (%)
Sugar (g)	−6.2	−89.6	77.1	152 (94.4)
Sugar (% TEI[c])	−1.3	−14.6	12	155 (96.3)
Dietary fiber (g)	1.4	−16	18.8	158 (98.1)
Saturated fat (g)	−1.7	−26	22.6	152 (94.4)
Saturated fat (% TEI)	−0.8	−9.3	7.7	155 (96.3)
Monounsaturated fat (g)	0.8	−22.3	24	157 (97.5)
Monounsaturated fat (% TEI)	0.4	−6.4	7.2	155 (96.3)
Polyunsaturated fat (g)	−3.0	−11.9	6	151 (93.8)
Polyunsaturated fat (% TEI)	−1.5	−4.9	1.8	149 (92.5)
Vitamin D (µg)	0.8	−5.3	6.8	154 (95.7)
Folate (µg)	43.9	−176.6	264.4	152 (94.4)
Vitamin C (mg)	26.3	−124.3	177	154 (95.7)
Calcium (mg)	−54.2	−765	656.6	155 (96.3)
Iron (mg)	0.6	−7.9	9.2	152 (94.4)
Potassium (mg)	252.2	−1733.8	2238.2	150 (93.2)
Sodium (mg)	−347.6	−2266.3	1571.2	153 (95)

[a]Differences are given as values from the 24-hour recall minus values from the generic recall.

[b]LOA: limit of agreement.

[c]TEI: total energy intake.

Correlation coefficients were statistically significant ($P<.05$) for 18 (78%) of the 23 variables. No significant correlation was identified for total fat (% TEI), monounsaturated fat (% TEI), polyunsaturated fat (% TEI), vitamin D (µg), and sodium (mg). For those where a statistically significant correlation was identified, they ranged from 0.16 for saturated fat as % TEI to 0.45 for sugar in grams, with median correlation of 0.32 (IQR 0.25-0.40; Table 4).

XSL•FO
RenderX

Table 4. Comparison of daily nutrient intakes estimated using the web-based, 24-hour recall and the novel, meal-based method, based on correlation and cross-classification of quartiles.

Nutrient intake	Correlation		Cross-classification of quartiles (N=161)			
	Spearman coefficient	*P* value	Exact agreement, n (%)	Exact + adjacent, n (%)	Disagreement, n (%)	Extreme disagreement, n (%)
Energy (kcal)	0.34	<.001	57 (35.4)	115 (71.4)	34 (21.1)	12 (7.5)
Protein (g)	0.26	.001	56 (34.8)	112 (69.6)	41 (25.5)	8 (5)
Protein (% TEI[a])	0.43	<.001	59 (36.6)	129 (80.1)	26 (16.1)	6 (3.7)
Carbohydrate (g)	0.42	<.001	62 (38.5)	127 (78.9)	25 (15.5)	9 (5.6)
Carbohydrate (% TEI)	0.33	<.001	51 (31.7)	122 (75.8)	28 (17.4)	11 (6.8)
Sugar (g)	0.45	<.001	53 (32.9)	126 (78.3)	26 (16.1)	9 (5.6)
Sugar (% TEI)	0.40	<.001	55 (34.2)	123 (76.4)	32 (19.9)	6 (3.7)
Dietary fiber (g)	0.43	<.001	57 (35.4)	125 (77.6)	26 (16.1)	10 (6.2)
Total fat (g)	0.24	.002	53 (32.9)	113 (70.2)	36 (22.4)	12 (7.5)
Total fat (% TEI)	0.14	.07	37 (23)	102 (63.4)	46 (28.6)	13 (8.1)
Saturated fat (g)	0.20	.01	53 (32.9)	112 (69.6)	38 (23.6)	11 (6.8)
Saturated fat (% TEI)	0.16	.045	41 (25.5)	110 (68.3)	38 (23.6)	13 (8.1)
Monounsaturated fat (g)	0.20	.01	53 (32.9)	112 (69.6)	36 (22.4)	13 (8.1)
Monounsaturated fat (% TEI)	0.02	.75	42 (26.1)	97 (60.2)	41 (25.5)	23 (14.3)
Polyunsaturated fat (g)	0.18	.02	52 (32.3)	108 (67.1)	41 (25.5)	12 (7.5)
Polyunsaturated fat (% TEI)	0.13	.11	47 (29.2)	112 (69.6)	32 (19.9)	17 (10.6)
Vitamin D (µg)	0.14	.07	50 (31.1)	104 (64.6)	35 (21.7)	22 (13.7)
Folate (µg)	0.31	<.001	51 (31.7)	114 (70.8)	34 (21.1)	13 (8.1)
Vitamin C (mg)	0.34	<.001	42 (26.1)	111 (68.9)	35 (21.7)	15 (9.3)
Calcium (mg)	0.26	.001	57 (35.4)	114 (70.8)	34 (21.1)	13 (8.1)
Iron (mg)	0.32	<.001	46 (28.6)	106 (65.8)	45 (28)	10 (6.2)
Potassium (mg)	0.39	<.001	42 (26.1)	106 (65.8)	45 (28)	10 (6.2)
Sodium (mg)	0.12	.12	46 (28.6)	116 (72)	33 (20.5)	12 (7.5)

[a]TEI: total energy intake.

Categorization of Daily Nutrient Intakes

Cross-classification of quartiles is also presented in Table 4. The proportion of individuals remaining in the same quartile ranged from 22.9% (37/161) for total fat (% TEI) to 39.1% (63/161) for carbohydrates (g), with median of 32.3% (IQR 28.6%-34.5%). Of the 23 nutrients, 3 (13%) nutrients had extreme disagreement for ≤5% of participants (protein in both g and % TEI and sugar as % TEI), 17 (74%) had extreme disagreement for between >5% and ≤10% of participants, and 3 (13%) had extreme disagreement for >10% of participants (monounsaturated and polyunsaturated fat as % TEI and vitamin D in µg). When participants were classified according to nutrient-based guidelines (eg, low, adequate, or high) for the 14 nutrients, the proportion of participants who were classified into the same category by both methods ranged from 52.8% (85/161) for total fat (% TEI) to 84.5% (136/161) for protein (g/kg body weight; Table 5). The median proportion of participants who were classified correctly among the 14 nutrients was 70.5% (IQR 61.8%-78.9%).

XSL•FO
RenderX

Table 5. Percentage of participants classified to the same category when their mean daily nutrient intakes estimated from both the 24-hour recall and the novel, meal-based recall were categorized according to nutrient-based guidelines.

Nutrient	Possible categories for the classification of individual nutrient intakes	Participants classified to the same category (N=161), n (%)
Protein (g/kg of BW[a])	Low, adequate, and high	136 (84.5)
Carbohydrate (% TEI[b])	Low, adequate, and high	101 (62.7)
Total fat (% TEI)	Low, adequate, and high	85 (52.8)
Monounsaturated fat (% TEI)	Low, adequate, and high	128 (79.5)
Polyunsaturated fat (% TEI)	Low, adequate, and high	96 (59.6)
Saturated fat (% TEI)	Adequate and high	114 (70.8)
Salt (g)	Adequate and high	99 (61.5)
Dietary fiber (g)	Low and adequate	122 (75.8)
Calcium (mg)	Low, adequate, and high	92 (57.1)
Iron (mg)	Low, adequate, and high	135 (83.9)
Folate (µg)	Low, adequate, and high	133 (82.6)
Thiamin (mg)	Low and adequate	124 (77)
Riboflavin (mg)	Low and adequate	113 (70.2)
Vitamin C (mg)	Low, adequate, and high	106 (65.8)

[a]BW: body weight.

[b]TEI: total energy intake.

Participant Evaluation Questionnaire

Most participants either somewhat agreed or agreed regarding generic recall that the instructions provided were clear and easy to understand (147/161, 91.3%); that the portion sizes in the tool were largely representative of the portion sizes they had consumed (137/161, 85.1%); and that, overall, the tool was easy to use (130/161, 80.7%). However, most (89/161, 55.3%) reported that the meal images were not representative of what they had actually consumed. This was anticipated, and for each meal in generic recall, participants were given the option to select that none of the meal images presented to them was similar to what they had consumed. This option was chosen for 36.04% (683/1895) of the total meals consumed, distributed across the meal types as follows: 17.4% (119/683) were breakfasts, 24.7% (169/683) were light meals, 20.4% (139/683) were main meals, 29.3% (200/683) were snacks, and 8.2% (56/683) were beverages. When these choices were reviewed by the researchers, by comparing participants' text descriptions of their meals with the possible options from generic meal images, it was determined that 86.8% (593/683) of the meals were correctly recorded by participants as not having a matching generic meal, whereas 13.2% (90/683) of the meals could have been matched to one of the generic meal images. Although most of the participants (106/161, 65.8%) reported that the ease of use of the meal-based tool was worse or somewhat worse than the 24HR, 39.1% (63/161) of the participants reported that they would consider using a similar tool to generic recall again.

Discussion

Principal Findings

This study reports about a novel, meal-based recall that allows individuals to report their intakes of whole meals rather than reporting individual food or food group intakes for each meal, as is necessary in commonly used traditional dietary intake assessment methods. Estimated nutrient intakes obtained from the generic meal–based method were comparable with those obtained from the 24HR for some but not all nutrients. Comparisons between the methods were more similar at the group level than the individual level. Participants found the meal-based method understandable and easy to use. Previous studies have identified generic meals that exist in national dietary survey data [15,31-33], but this study is the first to use images of those generic meals as a novel method of dietary intake assessment.

In previous studies of generic meals, comparisons were made between the estimated nutrient intakes obtained from the original data and from the generic data. Agreement between the original and generic data was stronger in those studies than the agreement between the 2 methods of dietary assessment presented in this study; however, this was expected because those studies compared intakes arising from generic meals with intakes arising from the original data from which those generic meals were derived [15,31-33]. In this study, comparisons were made on data collected using 2 different methods; therefore, the generic meal images presented to participants were not influenced by those participants' food intakes. Agreement between the methods in this study varied depending on the nutrient in question. In general, percentage differences and effect

XSL•FO
RenderX

sizes for differences between the 2 methods were small for most nutrients (19/23, 83%). Certain nutrients showed poorer agreement than others including total fat, polyunsaturated fats, monounsaturated fats, and vitamin D. A few features of the generic method may have given rise to differences in these fat-soluble nutrients. In the food-based 24HR, participants could specify the types of fats that they added to foods; however, in the generic method, this was not the case, as participants had to choose between predefined generic meals. Some of these nutrients are found in relatively high concentrations in relatively few foods, which may also give rise to differences between a food-based method and a generic method. These trends have also been observed previously when comparing FFQs with 24HRs and food records, as noted by Cui et al [34] in a meta-analysis of 130 such studies. In the case of polyunsaturated fats, the difference between the median intakes was considerably larger than other nutrients. This was identified as having arisen from differences in food composition between the generic data and the 24HR data. The generic meals used in the generic recall and their nutritional composition were derived from data from NANS in Ireland [15,16]. The composition data in that survey, in turn, were obtained from a variety of sources including food packaging, industry information, published papers, and various food composition tables including those from the United Kingdom, Finland, and Australia that were published between 2002 and 2010 [35]. The composition data used in the 24HR were obtained from the 2021 publication of McCance and Widdowson's CoFID [21]. Upon further examination of polyunsaturated fat content of individual foods from both NANS and CoFID, several foods were identified in NANS that had considerably higher values for polyunsaturated fat than those corresponding foods in CoFID in a manner which was not evident for other nutrients.

Although other studies have not examined the potential for dietary assessment based on individuals' reporting of whole meal intakes, comparisons can be drawn with other related methods. Murakami et al [36] developed a meal-focused method called the food combination questionnaire (FCQ). In this method, participants reported, for each meal type (breakfast, lunch, dinner, and snacks), what staple foods they had consumed, what accompanying foods they had consumed with those staple foods, and how frequently per week they had consumed them during the preceding month. Agreement regarding estimates of nutrient intake between the FCQ and a 4-day weighed food record was similar to the agreement reported in this study for the comparison between the generic recall and 24HR. The median value of the statistically significant correlation coefficients was 0.35 compared with 0.31 in this study. In that study, however, the FCQ estimated lower intakes for most nutrients compared with food record, whereas in this study, the generic recall estimated higher intakes for most nutrients (16/23, 70%) compared with the 24HR.

Another approach to dietary intake assessment reported by Katz et al [37] involves participants reporting their intakes at the dietary pattern level instead of the meal, food group, or food levels. In this format, participants are presented with 2 images at a time. Each image contains multiple different foods and meals representing different dietary patterns based on the Healthy Eating Index, and the participant must choose which image is more representative of their dietary intakes. This process of choosing between 2 different diet pattern images is repeated until a "best fit" is identified for the individual [37]. Comparison of estimated nutrient intakes obtained from this method with those obtained from 3 separate 24HRs resulted in median correlation of 0.30 for correlations that were statistically significant [38]. Similar to this study, the abovementioned diet pattern approach also tended to estimate higher nutrient intakes than the 24HR. Similar trends were also observed in the Bland-Altman analysis with low bias or systematic error for macronutrients (ie, mean differences were close to 0) but slightly greater bias for micronutrients. Wide LOAs were also observed in both studies, indicating considerable random error. Random error can be reduced through repeated measures [39]; therefore, conducting >2 generic recalls performed in this study may mitigate against the random error observed here. Another trend that is evident from the Bland-Altman plots for macronutrients expressed as % TEI is that generic recall gives higher estimates of intake than 24HR when intakes are low and gives lower estimates of intake than 24HR when intakes are high. This arises from the narrowing of the distribution of intakes (ie, reduced variance) in the generic method, where participants can only choose from a small number of generic meals compared with the practically limitless different combinations of foods that participants can choose from in 24HR. The use of portion sizes mitigates against this trend for values expressed in absolute intakes, for example, grams of macronutrient intake. However, expressing those values relative to EI effectively nullifies the increased variance introduced by portion size, giving rise to the trends observed in the Bland-Altman plots for values given as % TEI but not for those given as absolute intakes.

A version of FCQ described previously has been used to provided dietary advice in a pilot study by Murakami et al [40]. Similarly, the dietary pattern–based method of dietary assessment described previously has also been incorporated into a commercially available tool that provides nutrition recommendations to users [41]; however, the recommendation aspect of this tool has not been described in the scientific literature. It has also been proposed that a meal-based method of dietary intake assessment could provide a low-burden alternative to gathering dietary data for personalized nutrition advice [3]. This study presents the first attempt at implementing such a proposal. Its performance in estimating nutrient intakes is similar to those of other approaches that focus on meals or dietary patterns. However, these methods tend to have poorer agreement compared with that between FFQs and 24HRs, where median correlation of 0.42 has been reported for the various nutrients assessed [34], or compared with the agreement between 24HRs and diet records, where median correlations in the range of 0.45 to 0.50 have been reported [42,43]. Further studies are required to appraise the performance of these methods, including generic recall, against objective biomarkers of dietary intake.

Many of the comparisons between the 2 methods in this study are based on their ability to provide point estimates of nutrient intakes, with agreement for point estimates of nutrient intake typically stronger at the group level than at the individual level. This trend has also been observed in comparisons of other

XSL•FO
RenderX

methods of dietary intake assessment, for example, between 24HR and diet record [42], between FFQ and diet record [44], and between FFQ and 24HR [45]. In this study, this was expected, given that there are fewer generic meals than foods for participants to choose from, that is, the variance was intentionally reduced but in a manner that was systematic and consistent across meals [15]. This trend has been observed not just in comparison studies of different methods of dietary assessment but also in studies that have used generic meals as a method for secondary analysis of dietary data that have already been collected [15,31-33]. Approaches to personalized nutrition facilitated by technology, however, do not rely solely on point estimates of nutrient intake. Instead, individuals are categorized into ranges of nutrient intakes (eg, low, adequate, or high), allowing room for error in the point estimates [40,46]. In the Food4Me study of personalized nutrition, for example, participants were categorized in this manner, and dietary advice reports were tailored depending on the participants' categories for various nutrients [40,46]. This study has shown that individuals can be classified according to nutrient-based dietary guidelines using generic recall. Similarly, this approach can be used to rank individuals based on their nutrient intakes, with values from the cross-classification of quartiles being comparable with those observed in studies comparing FFQs with 24HRs, where median exact + adjacent agreement has been reported between 66% and 86.1% [47-49].

Notably, with regard to the use of generic recall in dietary assessment in personalized nutrition, most participants (130/161, 80.7%) reported in the evaluation questionnaire that the recall was easy to use. Ease of use has previously been reported as an important factor in the choice of nutrition or diet apps among a cohort of 2382 adults in Europe [5]. In contrast, two-thirds of the participants reported that the ease of use of the 24HR was better than that of the meal-based approach. This is understandable given that the web-based 24HR used in this study was developed as a stand-alone platform in collaboration with software developers [20], whereas generic recall was a concept or pilot tool implemented by the authors using a commercially available questionnaire platform and not specifically designed for user experience. Future user evaluations and collaboration with software development professionals could further enhance the user experience of generic recall, including incorporation of meal image recognition.

So far, many studies have been conducted on image-based food recognition using computer vision as a means of reducing the burden of data input for food-based methods of dietary intake assessment [13]. These image recognition tools, however, are food based, insofar as when an individual takes a photo of their meal, the software segments the image and provides a suggested match for each of the individual foods that make up the meal [13]. The user must then confirm that each of the suggested foods are correct. For any missing or incorrect foods, the user must text search for the correct food and add it to their record [12]. It is possible that a meal-based approach could be taken to image recognition in dietary assessment, removing the need to identify individual foods. Instead, the software would classify a whole meal image as one of the generic meals used in this study. Further studies are required, however, to determine the feasibility of such an approach for meal-based image recognition in nutrition.

Limitations

This study has a number of limitations. The generic meal images used are based on dietary intake data obtained from NANS (2008-2010) [16]. Although, at the time of writing, these were the most recently published intake data in Ireland, the generic meals do not account for any changes in food composition or meal intakes that may have occurred since that time. This may account for the findings in the evaluation questionnaire of this study, with 55.3% (89/161) of the participants reporting that the generic meals were not representative of their intakes, and shows the need to ensure that generic meals are revised using most recent data. This may be of particular importance if the tool was to be used with a younger cohort of participants than those who participated in this study. This study also used a convenience sample rather than one that is representative of the population in Ireland. This may result in the recruitment of participants who are more likely to have an interest in nutrition and health. The unbalanced nature of the demographics of the study participants precludes subgroup analysis in relation to demographic factors.

The 2 types of recall used in this study are influenced by measurement error and only provide an estimation of true intakes. This study aimed to compare these 2 methods, and therefore, the reported statistics should be interpreted as representing the relationship between generic recall and 24HR and not the relationship between generic recall and true intakes. A comparison study was deemed more appropriate as an initial indication of generic recall's strengths and weaknesses before considering more labor-intensive and costly objective measures of comparison such as feeding studies or biomarkers of dietary intake.

Strengths

The strengths of this study include its sample size of 161 participants. The randomization of participants regarding the recall method they would complete first, the reversal of the order of completion in their second set of recalls, and the 2-week washout period mitigates against any learning effect that participants may have experienced after completing the first recall. The comparison method used, Foodbook24, is a validated, web-based 24HR [20]. This study also captured dietary intakes on both weekend days and weekdays, accounting for the potential differences in eating habits that occur between the 2 categories [50-52].

Conclusions

A generic meal–based method of dietary intake assessment provides estimates of nutrient intake comparable with those provided by a web-based 24HR. The agreement ranges among nutrients from weak to moderate, with better agreement at the group level than the individual level. Further studies are required to improve this method of dietary assessment considering the number of recalls required, and more recent dietary intake data should be used to define the generic meals. Future studies will

XSL•FO
RenderX

determine the feasibility of taking a meal-based approach to image recognition in dietary intake assessment.

Acknowledgments

This paper has emanated from a study conducted with the financial support of Science Foundation Ireland under grant 12/RC/2289_P2. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. Science Foundation Ireland had no role in any aspect of designing or conducting the study or preparing the manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

CO was involved in conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, original draft preparation, and review and editing. ERG was involved in conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, resources, supervision, visualization, and review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analysis of the smaller data set, with misreporters excluded.
[DOCX File , 31 KB - jmir_v26i1e48817_app1.docx]

References

1. GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2019 May 11;393(10184):1958-1972 [FREE Full text] [doi: 10.1016/S0140-6736(19)30041-8] [Medline: 30954305]
2. Gandy J. Manual of Dietetic Practice. 6th edition. Hoboken, NJ: John Wiley & Sons; Jul 2019.
3. Gibney MJ, Walsh MC. The future direction of personalised nutrition: my diet, my phenotype, my genes. Proc Nutr Soc 2013 May;72(2):219-225. [doi: 10.1017/S0029665112003436] [Medline: 23360849]
4. O'Hara C, Gibney ER. Meal pattern analysis in nutritional science: recent methods and findings. Adv Nutr 2021 Jul 30;12(4):1365-1378 [FREE Full text] [doi: 10.1093/advances/nmaa175] [Medline: 33460431]
5. Vasiloglou MF, Christodoulidis S, Reber E, Stathopoulou T, Lu Y, Stanga Z, et al. Perspectives and preferences of adult smartphone users regarding nutrition and diet apps: web-based survey study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e27885 [FREE Full text] [doi: 10.2196/27885] [Medline: 34328425]
6. Eldridge AL, Piernas C, Illner AK, Gibney MJ, Gurinović MA, de Vries JH, et al. Evaluation of new technology-based tools for dietary intake assessment-an ILSI Europe dietary intake and exposure task force evaluation. Nutrients 2018 Dec 28;11(1):55 [FREE Full text] [doi: 10.3390/nu11010055] [Medline: 30597864]
7. Camargo AM, Botelho AM, Dean M, Fiates GM. Meal planning by high and low health conscious individuals during a simulated shop in the supermarket: a mixed methods study. Appetite 2020 Jan 01;144:104468. [doi: 10.1016/j.appet.2019.104468] [Medline: 31557495]
8. Bisogni CA, Falk LW, Madore E, Blake CE, Jastran M, Sobal J, et al. Dimensions of everyday eating and drinking episodes. Appetite 2007 Mar;48(2):218-231. [doi: 10.1016/j.appet.2006.09.004] [Medline: 17088011]
9. Englund-Ögge L, Birgisdottir BE, Sengpiel V, Brantsæter AL, Haugen M, Myhre R, et al. Meal frequency patterns and glycemic properties of maternal diet in relation to preterm delivery: results from a large prospective cohort study. PLoS One 2017 Mar 01;12(3):e0172896 [FREE Full text] [doi: 10.1371/journal.pone.0172896] [Medline: 28249018]
10. Wilson JE, Blizzard L, Gall SL, Magnussen CG, Oddy WH, Dwyer T, et al. An eating pattern characterised by skipped or delayed breakfast is associated with mood disorders among an Australian adult cohort. Psychol Med 2020 Dec;50(16):2711-2721 [FREE Full text] [doi: 10.1017/S0033291719002800] [Medline: 31615586]
11. Murakami K, Shinozaki N, McCaffrey TA, Livingstone MB, Sasaki S. Data-driven development of the meal-based diet history questionnaire for Japanese adults. Br J Nutr 2021 Oct 14;126(7):1056-1064. [doi: 10.1017/S0007114520004936] [Medline: 33298210]
12. Boushey CJ, Spoden M, Zhu FM, Delp EJ, Kerr DA. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. Proc Nutr Soc 2017 Aug;76(3):283-294. [doi: 10.1017/S0029665116002913] [Medline: 27938425]

XSL•FO
RenderX

13. Dalakleidi KV, Papadelli M, Kapolos I, Papadimitriou K. Applying image-based food-recognition systems on dietary assessment: a systematic review. Adv Nutr 2022 Dec 22;13(6):2590-2619 [FREE Full text] [doi: 10.1093/advances/nmac078] [Medline: 35803496]
14. Timon CM, van den Barg R, Blain RJ, Kehoe L, Evans K, Walton J, et al. A review of the design and validation of web- and computer-based 24-h dietary recall tools. Nutr Res Rev 2016 Dec;29(2):268-280. [doi: 10.1017/S0954422416000172] [Medline: 27955721]
15. O'Hara C, O'Sullivan A, Gibney ER. A clustering approach to meal-based analysis of dietary intakes applied to population and individual data. J Nutr 2022 Oct 06;152(10):2297-2308 [FREE Full text] [doi: 10.1093/jn/nxac151] [Medline: 35816468]
16. Walton J. National adult nutrition survey summary report. Irish Universities Nutrition Alliance. 2011 Mar. URL: https://irp-cdn.multiscreensite.com/46a7ad27/files/uploaded/The%20National%20Adult%20Nutrition%20Survey%20Summary%20Report%20March%202011.pdf [accessed 2024-05-20]
17. Fulgoni 3rd VL, Keast DR, Drewnowski A. Development and validation of the nutrient-rich foods index: a tool to measure nutritional quality of foods. J Nutr 2009 Aug;139(8):1549-1554 [FREE Full text] [doi: 10.3945/jn.108.101360] [Medline: 19549759]
18. Evans K, Hennessy Á, Walton J, Timon C, Gibney E, Flynn A. Development and evaluation of a concise food list for use in a web-based 24-h dietary recall tool. J Nutr Sci 2017 Aug 29;6:e46 [FREE Full text] [doi: 10.1017/jns.2017.49] [Medline: 29152250]
19. Timon CM, Evans K, Kehoe L, Blain RJ, Flynn A, Gibney ER, et al. Comparison of a web-based 24-h dietary recall tool (Foodbook24) to an interviewer-led 24-h dietary recall. Nutrients 2017 Apr 25;9(5):425 [FREE Full text] [doi: 10.3390/nu9050425] [Medline: 28441358]
20. Timon CM, Blain RJ, McNulty B, Kehoe L, Evans K, Walton J, et al. The development, validation, and user evaluation of Foodbook24: a web-based dietary assessment tool developed for the irish adult population. J Med Internet Res 2017 May 11;19(5):e158 [FREE Full text] [doi: 10.2196/jmir.6407] [Medline: 28495662]
21. Public Health England. McCance and Widdowson's The Composition of Foods. Burlington, MA: The Royal Society of Chemistry Publishing; 2002.
22. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2019. URL: https://www.R-project.org/ [accessed 2024-01-26]
23. RStudio: integrated development for R. RStudio. URL: http://www.rstudio.com/ [accessed 2024-01-26]
24. Henry CJ. Basal metabolic rate studies in humans: measurement and development of new equations. Public Health Nutr 2005 Oct;8(7A):1133-1152. [doi: 10.1079/phn2005801] [Medline: 16277825]
25. Black AE. Critical evaluation of energy intake using the Goldberg cut-off for energy intake:basal metabolic rate. A practical guide to its calculation, use and limitations. Int J Obes Relat Metab Disord 2000 Sep;24(9):1119-1130. [doi: 10.1038/sj.ijo.0801376] [Medline: 11033980]
26. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
27. Lombard MJ, Steyn NP, Charlton KE, Senekal M. Application and interpretation of multiple statistical tests to evaluate validity of dietary intake assessment methods. Nutr J 2015 Apr 22;14:40 [FREE Full text] [doi: 10.1186/s12937-015-0027-y] [Medline: 25897837]
28. Institute of Medicine. Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids. Washington, DC: The National Academies Press; 2005.
29. Joint WHO/FAO/UNU Expert Consultation. Protein and amino acid requirements in human nutrition. World Health Organ Tech Rep Ser 2007(935):1-265. [Medline: 18330140]
30. Institute of Medicine (U.S.). Subcommittee on Interpretation and Uses of Dietary Reference Intakes, Institute of Medicine (U.S.). Subcommittee on Upper Reference Levels of Nutrients, Standing Committee on the Scientific Evaluation of Dietary Reference Intakes. Dietary Reference Intakes: Applications in Dietary Assessment. Washington, DC: The National Academies Press; 2000.
31. Woolhead C, Gibney MJ, Walsh MC, Brennan L, Gibney ER. A generic coding approach for the examination of meal patterns. Am J Clin Nutr 2015 Aug;102(2):316-323 [FREE Full text] [doi: 10.3945/ajcn.114.106112] [Medline: 26085514]
32. Uzhova I, Woolhead C, Timon CM, O'Sullivan A, Brennan L, Peñalvo JL, et al. Generic meal patterns identified by latent class analysis: insights from NANS (national adult nutrition survey). Nutrients 2018 Mar 06;10(3):310 [FREE Full text] [doi: 10.3390/nu10030310] [Medline: 29509665]
33. Murakami K, Livingstone MB, Sasaki S, Hirota N, Notsu A, Miura A, et al. Applying a meal coding system to 16-d weighed dietary record data in the Japanese context: towards the development of simple meal-based dietary assessment tools. J Nutr Sci 2018 Nov 13;7:e29 [FREE Full text] [doi: 10.1017/jns.2018.21] [Medline: 30459946]
34. Cui Q, Xia Y, Wu Q, Chang Q, Niu K, Zhao Y. Validity of the food frequency questionnaire for adults in nutritional epidemiological studies: a systematic review and meta-analysis. Crit Rev Food Sci Nutr 2023;63(12):1670-1688. [doi: 10.1080/10408398.2021.1966737] [Medline: 34520300]
35. Li K, McNulty BA, Tiernery AM, Devlin NF, Joyce T, Leite JC, et al. Dietary fat intakes in Irish adults in 2011: how much has changed in 10 years? Br J Nutr 2016 May 28;115(10):1798-1809. [doi: 10.1017/S0007114516000787] [Medline: 26996058]

XSL•FO
RenderX

36. Murakami K, Shinozaki N, Livingstone MB, Kimoto N, Masayasu S, Sasaki S. Food combination questionnaire for Japanese: relative validity regarding food and nutrient intake and overall diet quality against the 4-day weighed dietary record. J Nutr Sci 2023;12:e22 [FREE Full text] [doi: 10.1017/jns.2023.7] [Medline: 36843967]
37. Katz DL, Rhee LQ, Katz CS, Aronson DL, Frank GC, Gardner CD, et al. Dietary assessment can be based on pattern recognition rather than recall. Med Hypotheses 2020 Feb 26;140:109644. [doi: 10.1016/j.mehy.2020.109644] [Medline: 32131036]
38. Turner-McGrievy G, Hutto B, Bernhart JA, Wilson MJ. Comparison of the diet ID platform to the automated self-administered 24-hour (ASA24) dietary assessment tool for assessment of dietary intake. J Am Nutr Assoc 2022;41(4):360-382 [FREE Full text] [doi: 10.1080/07315724.2021.1887775] [Medline: 33705267]
39. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML, et al. Addressing current criticism regarding the value of self-report dietary data. J Nutr 2015 Dec;145(12):2639-2645 [FREE Full text] [doi: 10.3945/jn.115.219634] [Medline: 26468491]
40. Murakami K, Shinozaki N, Masayasu S, Livingstone MB. Web-based personalized nutrition system for delivering dietary feedback based on behavior change techniques: development and pilot study among dietitians. Nutrients 2021 Sep 27;13(10):3391 [FREE Full text] [doi: 10.3390/nu13103391] [Medline: 34684392]
41. The evidence-based nutrition wellbeing toolkit. Diet ID. URL: https://www.dietid.com/home2 [accessed 2024-03-15]
42. Frankenfeld CL, Poudrier JK, Waters NM, Gillevet PM, Xu Y. Dietary intake measured from a self-administered, online 24-hour recall system compared with 4-day diet records in an adult US population. J Acad Nutr Diet 2012 Oct;112(10):1642-1647. [doi: 10.1016/j.jand.2012.06.003] [Medline: 22878341]
43. De Keyzer W, Huybrechts I, De Vriendt V, Vandevijvere S, Slimani N, Van Oyen H, et al. Repeated 24-hour recalls versus dietary records for estimating nutrient intakes in a national food consumption survey. Food Nutr Res 2011;55:1-10 [FREE Full text] [doi: 10.3402/fnr.v55i0.7307] [Medline: 22084625]
44. Araujo MC, Yokoo EM, Pereira RA. Validation and calibration of a semiquantitative food frequency questionnaire designed for adolescents. J Am Diet Assoc 2010 Aug;110(8):1170-1177. [doi: 10.1016/j.jada.2010.05.008] [Medline: 20656092]
45. Dehghan M, del Cerro S, Zhang X, Cuneo JM, Linetzky B, Diaz R, et al. Validation of a semi-quantitative food frequency questionnaire for Argentinean adults. PLoS One 2012;7(5):e37958 [FREE Full text] [doi: 10.1371/journal.pone.0037958] [Medline: 22662256]
46. Celis-Morales C, Livingstone KM, Marsaux CF, Macready AL, Fallaize R, O'Donovan CB, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4Me European randomized controlled trial. Int J Epidemiol 2017 Apr 01;46(2):578-588. [doi: 10.1093/ije/dyw186] [Medline: 27524815]
47. Ye Q, Hong X, Wang Z, Yang H, Chen X, Zhou H, et al. Reproducibility and validity of an FFQ developed for adults in Nanjing, China. Br J Nutr 2016 Mar 14;115(5):887-894. [doi: 10.1017/S0007114515005334] [Medline: 26785928]
48. Zack RM, Irema K, Kazonda P, Leyna GH, Liu E, Gilbert S, et al. Validity of an FFQ to measure nutrient and food intakes in Tanzania. Public Health Nutr 2018 Aug;21(12):2211-2220 [FREE Full text] [doi: 10.1017/S1368980018000848] [Medline: 29656731]
49. Liu L, Wang PP, Roebothan B, Ryan A, Tucker CS, Colbourne J, et al. Assessing the validity of a self-administered food-frequency questionnaire (FFQ) in the adult population of Newfoundland and Labrador, Canada. Nutr J 2013 Apr 16;12:49 [FREE Full text] [doi: 10.1186/1475-2891-12-49] [Medline: 23590645]
50. Haines PS, Hama MY, Guilkey DK, Popkin BM. Weekend eating in the United States is linked with greater energy, fat, and alcohol intake. Obes Res 2003 Aug;11(8):945-949 [FREE Full text] [doi: 10.1038/oby.2003.130] [Medline: 12917498]
51. Racette SB, Weiss EP, Schechtman KB, Steger-May K, Villareal DT, Obert KA, et al. Influence of weekend lifestyle patterns on body weight. Obesity (Silver Spring) 2008 Aug;16(8):1826-1830 [FREE Full text] [doi: 10.1038/oby.2008.320] [Medline: 18551108]
52. An R. Weekend-weekday differences in diet among U.S. adults, 2003-2012. Ann Epidemiol 2016 Jan;26(1):57-65. [doi: 10.1016/j.annepidem.2015.10.010] [Medline: 26559331]

Abbreviations

24HR: 24-hour recall
BMR: basal metabolic rate
CoFID: Composition of Foods Integrated Dataset
EI: energy intake
FCQ: food combination questionnaire
FFQ: food frequency questionnaire
LOA: limit of agreement
NANS: National Adult Nutrition Survey
NRF: Nutrient Rich Foods
TEI: total energy intake

XSL•FO
RenderX

Edited by T Leung, G Eysenbach; submitted 10.05.23; peer-reviewed by C Matthys, N Pahlavani; comments to author 31.08.23; revised version received 19.09.23; accepted 29.11.23; published 14.02.24.

Please cite as:
O'Hara C, Gibney ER
Dietary Intake Assessment Using a Novel, Generic Meal–Based Recall and a 24-Hour Recall: Comparison Study
J Med Internet Res 2024;26:e48817
URL: https://www.jmir.org/2024/1/e48817
doi:10.2196/48817
PMID:38354039

©Cathal O'Hara, Eileen R Gibney. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Validation of 3 Computer-Aided Facial Phenotyping Tools (DeepGestalt, GestaltMatcher, and D-Score): Comparative Diagnostic Accuracy Study

Alisa Maria Vittoria Reiter[1]; Jean Tori Pantel[1,2,3]; Magdalena Danyel[1,4,5], MD; Denise Horn[1], Prof Dr; Claus-Eric Ott[1], MD; Martin Atta Mensah[1,6], MD

[1]Institute of Medical Genetics and Human Genetics, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]Institute for Digitalization and General Medicine, University Hospital Aachen, Aachen, Germany

[3]Center for Rare Diseases Aachen ZSEA, University Hospital Aachen, Aachen, Germany

[4]BIH Biomedical Innovation Academy, Clinician Scientist Program, Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany

[5]Berlin Center for Rare Diseases, Charité - Universitätsmedizin Berlin corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[6]BIH Biomedical Innovation Academy, Digital Clinician Scientist Program, Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Martin Atta Mensah, MD
Institute of Medical Genetics and Human Genetics
Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin
Augustenburger Platz 1
Berlin, 13353
Germany
Phone: 49 30450569132
Fax: 49 30450569914
Email: martin-atta.mensah@charite.de

Abstract

Background: While characteristic facial features provide important clues for finding the correct diagnosis in genetic syndromes, valid assessment can be challenging. The next-generation phenotyping algorithm DeepGestalt analyzes patient images and provides syndrome suggestions. GestaltMatcher matches patient images with similar facial features. The new D-Score provides a score for the degree of facial dysmorphism.

Objective: We aimed to test state-of-the-art facial phenotyping tools by benchmarking GestaltMatcher and D-Score and comparing them to DeepGestalt.

Methods: Using a retrospective sample of 4796 images of patients with 486 different genetic syndromes (London Medical Database, GestaltMatcher Database, and literature images) and 323 inconspicuous control images, we determined the clinical use of D-Score, GestaltMatcher, and DeepGestalt, evaluating sensitivity; specificity; accuracy; the number of supported diagnoses; and potential biases such as age, sex, and ethnicity.

Results: DeepGestalt suggested 340 distinct syndromes and GestaltMatcher suggested 1128 syndromes. The top-30 sensitivity was higher for DeepGestalt (88%, SD 18%) than for GestaltMatcher (76%, SD 26%). DeepGestalt generally assigned lower scores but provided higher scores for patient images than for inconspicuous control images, thus allowing the 2 cohorts to be separated with an area under the receiver operating characteristic curve (AUROC) of 0.73. GestaltMatcher could not separate the 2 classes (AUROC 0.55). Trained for this purpose, D-Score achieved the highest discriminatory power (AUROC 0.86). D-Score's levels increased with the age of the depicted individuals. Male individuals yielded higher D-scores than female individuals. Ethnicity did not appear to influence D-scores.

Conclusions: If used with caution, algorithms such as D-score could help clinicians with constrained resources or limited experience in syndromology to decide whether a patient needs further genetic evaluation. Algorithms such as DeepGestalt could support diagnosing rather common genetic syndromes with facial abnormalities, whereas algorithms such as GestaltMatcher could suggest rare diagnoses that are unknown to the clinician in patients with a characteristic, dysmorphic face.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e42904) doi:10.2196/42904

KEYWORDS

facial phenotyping; DeepGestalt; facial recognition; Face2Gene; medical genetics; diagnostic accuracy; genetic syndrome; machine learning; GestaltMatcher; D-Score; genetics

Introduction

Background

Particularly in genetic syndromology, specific external anatomical features of patients are used to evaluate differential diagnoses [1]. Many inheritable syndromes are associated with specific patterns of dysmorphic signs, especially of the face. However, the clinical genetic description of a patient's face is complicated, and the perception and naming of abnormalities depend on the individual examiner [2]. Establishing a clinical genetic differential diagnosis is also complicated by the plethora of rare genetic syndromes [3].

Computer vision and machine learning offer scalable technologies that allow rapid and standardized recognition of relevant facial anatomical parameters and automatic assignment to a list of possible diagnoses. Several tools have been developed (eg, [4-13]) that perform computer-aided facial phenotyping using common patient photographs or even 3D patient images, often referred to as next-generation phenotyping (NGP). Ideally, NGP tools succeed in three tasks: (1) discerning patients from unaffected controls, (2) recognizing as many syndromes as possible, and (3) assigning the correct diagnosis.

One provider of such NGP tools is Face2Gene (FDNA Inc), which offers a range of applications based on the DeepGestalt [10] neural network. Initial studies have evaluated the ability of this network to correctly detect a syndrome in a patient image, suggesting good sensitivity (eg, [14-21]). It has been shown that such tools can also be trained for yet unknown syndromes by demonstrating that a new syndrome has a specific gestalt that can be distinguished from other syndromes (eg, [22-29]). DeepGestalt can only detect a specific syndrome if a minimum number of images of that syndrome is included in the training set. However, the GestaltMatcher tool—a further development of DeepGestalt—solves this problem by matching patient images against each other so that extremely rare or unknown syndromes can be detected. The last layer of the neural network is used to create a multidimensional clinical face phenotype space in which the spatial proximity reflects the similarity between 2 patient faces [30]. Among other studies, these findings suggest that facial NGP could advance the diagnostic process in medical genetics as well as support and accelerate the detection of novel syndromes. More recently, computer-aided facial phenotyping has been discussed for screening in telehealth applications, especially in resource-limited settings [13]. While DeepGestalt, in principle, is able to distinguish patients who are dysmorphic from unaffected control individuals [20,22], it by design has a limited ability to produce markedly different scores on photographs of affected and unaffected individuals [20,21]. The latest addition to the Face2Gene application suite called D-Score, however, aims to provide a discriminatory score to identify patients with dysmorphic facial features.

Objective

The recently described GestaltMatcher has not yet been benchmarked in individuals without a genetic syndrome and D-Score has not been benchmarked at all. Therefore, we evaluated the diagnostic accuracy of D-Score in individuals of different sexes, ages, and ethnic backgrounds and compared the results with those of DeepGestalt and GestaltMatcher.

Methods

Deep Learning Models

DeepGestalt

DeepGestalt is a deep convolutional neural network, trained on portrait images of individuals with genetic syndromes as described in detail previously [10]. The classes of the network are defined by sets of images of specific syndromes. When DeepGestalt analyzes the facial image of the query, it calculates and assigns a score for each of these classes, which corresponds to the degree of similarity of the query image to that class. The tool returns a list of syndrome suggestions ordered by the so-called DeepGestalt score. Here, we used DeepGestalt version 21.5.0 to obtain the top-30 syndrome matches and DeepGestalt scores for each image of the included test sets.

GestaltMatcher

GestaltMatcher is a previously described [30] encoder for portrait photos based on a deep convolutional neural network that creates a clinical face phenotype space [7] to match patient portrait images by the similarity of facial features. This tool can be used either to determine the similarity of different query images to each other or to individually embedded images, or to get syndrome suggestions for a single query image. In this study, we only tested GestaltMatcher's latter ability. For the analysis of the test sets with GestaltMatcher (version 0.8.0), the maximum possible number (n=100) of suggested syndromes in the web interface was chosen to count the number of syndromes known to this system, including syndromes supported by DeepGestalt as well as ultrarare syndromes not supported by DeepGestalt. When comparing the rank-dependent sensitivity of GestaltMatcher and DeepGestalt, however, only the first 30 suggestions of GestaltMatcher were used, since DeepGestalt only provides 30 suggestions.

D-Score

D-score is a support vector machine running on the output of a deep convolutional neural network for the syndrome classification of patient images. As a binary classifier, D-Score returns a single value per image, aiming to capture the degree of facial dysmorphism. The concept has been described previously [20]. However, D-Score was trained and validated on a larger data set and its architecture is more advanced and more complex than the model described by Pantel et al [20].

XSL•FO
RenderX

The test sets were analyzed by the beta version of D-Score (version 0.1 beta) accessed via an applied programming interface.

Test Sets

Overview

In this study, we analyzed images of 3 data sets (London Medical Database [LMD], GestaltMatcher Database [GMDB], and data previously described in Pantel et al [20]; Figure 1 [20] and Table 1) as a retrospective sample. We aimed to analyze as many cases as possible using only 1 frontal image per case and restricting the test set to those cases that are not included in the training sets of DeepGestalt, GestaltMatcher, or D-Score. For syndromic cases, the diagnosis was established either clinically or molecularly. No other information was entered into the facial analysis applications besides the frontal images.

Figure 1. Workflow of analyses. GMDB: GestaltMatcher Database; LMD: London Medical Database. *Binary classifier uses only 1 score.

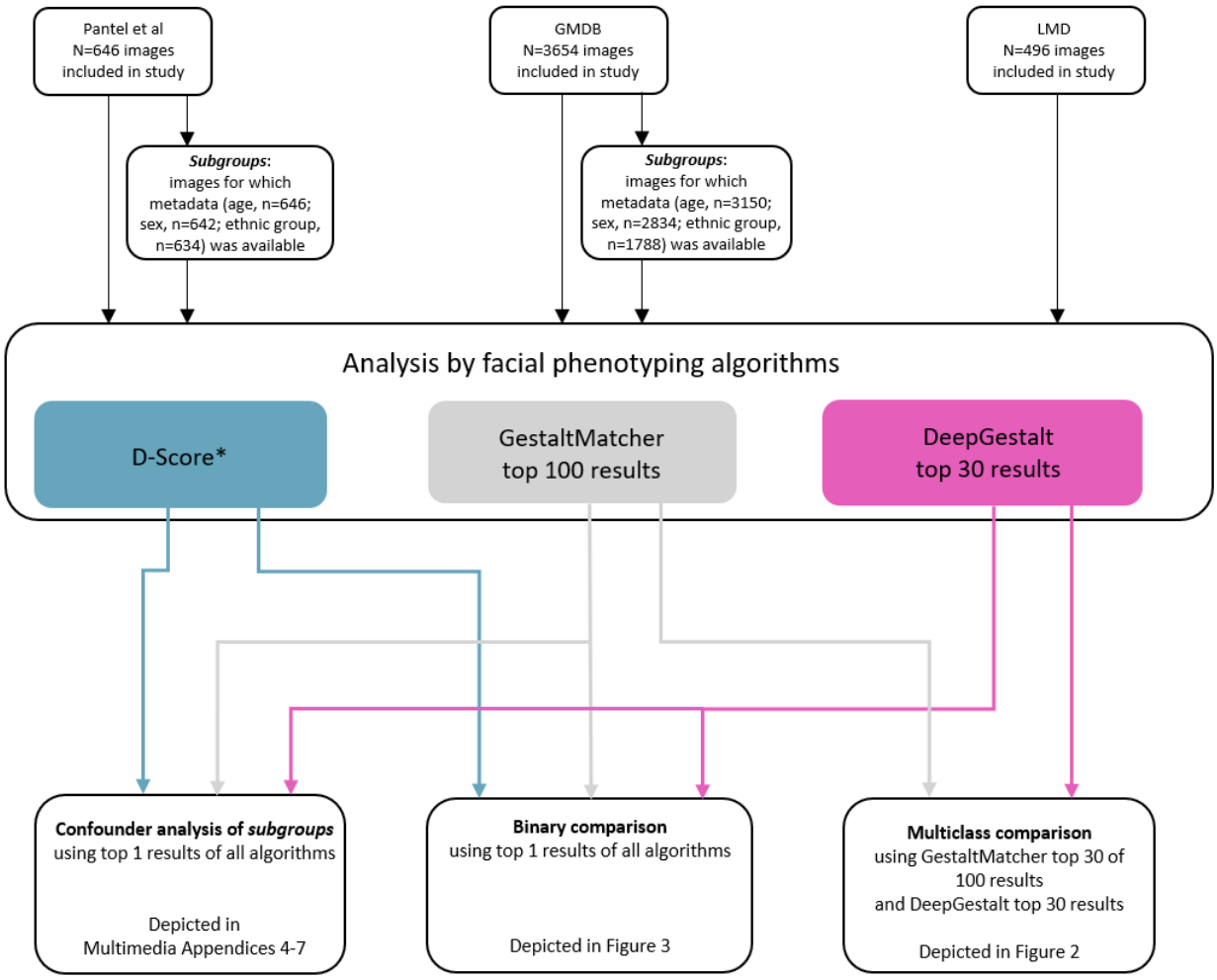

XSL•FO
RenderX

Table 1. Characteristics of the image data sets.

	Pantel et al [20] syndromic cohort, n	Pantel et al [20] healthy control cohort, n	LMD[a] test set, n	GMDB[b] test set, n	Total, N
Included images (total)	323	323	496	3654	4796
D-Score	322	323	496	3654	4795
Failed D-Score	1	0	0	0	1
DeepGestalt	323	323	496	3654	4796
GestaltMatcher	323	321	496	3654	4794
Failed GestaltMatcher	0	2	0	0	2
Number of syndromes	17	N/A[c]	199	346	486
Sex					
Female	159	159	—[d]	1299	1617
Male	162	162	—	1535	1859
Unknown	2	2	496	820	1320
Ethnicity					
White	273	273	—	1221	1767
Other	44	44	—	567	655
Unknown	6	6	496	1866	2374

[a]LMD: London Medical Database.

[b]GMDB: Gestalt Matcher Database.

[c]N/A: not applicable.

[d]Not available.

London Medical Database Set

This test set was built from 496 images of the LMD showing individuals with a genetic syndrome.

GestaltMatcher Database Set

The GMDB is a user-curated database of facial images of individuals with genetic syndromes. From this database, 3654 portrait photographs were included in our study that were not part of the training sets.

Literature Images

This data set was previously used by Pantel et al [20] comprising a total of 646 images, including 323 images of clinically or molecularly diagnosed patients (Pantel syndromic cohort) with 17 genetic syndromes (19 images for each syndrome) and 323 sex-, ethnicity-, and age-matched control images (Pantel healthy control cohort).

Ethical Considerations

This study was approved by the ethics committee of the Charité-Universitätsmedizin Berlin (EA2/190/16).

Statistical Analyses

Overview

Statistical analyses were performed using Python (version 3.8; Python Software Foundation). The code and data required to reproduce statistical analyses can be found in Multimedia Appendix 1 (see also the readme.txt for further information).

Comparison of DeepGestalt and GestaltMatcher

The algorithms performing a multiclass comparison—that is, DeepGestalt and GestaltMatcher—were compared regarding their numbers of supported syndromes, distributions and heights of yielded scores, sensitivities, and false positive rates. Since DeepGestalt's output was limited to 30 suggestions, for reason of comparability, a syndrome was also only considered to be supported if it was suggested at least once on rank 30 or better in GestaltMatcher. Top-1, top-10, and top-30 sensitivities, that is, the ability to respectively place the correct diagnosis on the first, within the first 10, or within the first 30 ranks of the results list, were calculated for each supported syndrome represented by at least 5 cases in the testing data sets. Sensitivities were additionally calculated for the images of patients in the literature image set. Mean sensitivities were averaged by the syndrome, not by the number of images, giving each syndrome the same weight. Any syndrome suggestion other than the syndrome featured by the person depicted was considered a false positive. As the tools do not support a class label "inconspicuous face," any suggestion was defined as a false positive in unaffected controls. Top-1, top-5, top-10, and top-30 false positive rates of each syndrome falsely suggested by a tool were calculated, representing the fraction of cases in which a syndrome was falsely suggested as the top-1 or within the first 5, 10, and 30 suggestions, respectively. False positive rates of patient and control images were tested for a possible correlation using a linear regression model and Pearson correlation coefficient.

XSL•FO
RenderX

Validation of the Binary Classification Performance of DeepGestalt, GestaltMatcher, and D-Score

We tested the ability of all 3 models (DeepGestalt, GestaltMatcher, and D-Score) to distinguish portrait images of individuals affected with a genetic syndrome from those of healthy controls. As they return a list of scores, only the highest scores (top-1 scores) assigned to an image by DeepGestalt and GestaltMatcher were compared with the single number output of D-Score. Mann-Whitney *U* tests were used to analyze possible differences in the yielded scores (1-sided for dysmorphic vs nondysmorphic cohorts and 2-sided for any other comparisons). The area under the receiver operating characteristic curve was calculated as a metric of accuracy. Analyses were also performed for clinically relevant patient-specific subgroups. For images for which this information was available, output data were separated by age group, sex, and ethnicity. We used the age groups 0 to 2 years (babies and toddlers), 3 to 10 years (children), 11 to 20 years (teenagers), 21 to 40 years (young adults), 41 to 60 years (middle-aged adults), and older than 60 years (older adults). As the ethnicity of most images was White (defined as being of apparently solely European descent), we built 2 groups of ethnicities: White and other. We did not define a cutoff for syndromic or inconspicuous for any of the tools.

Results

Syndrome Classification With DeepGestalt and GestaltMatcher

In total, 4796 images were used to test DeepGestalt, GestaltMatcher, and D-score. To investigate and compare the scope of diagnoses for which DeepGestalt and GestaltMatcher can be used, we counted the number of syndromes suggested by each tool. Among all matches, DeepGestalt and GestaltMatcher suggested a total of 340 and 1128 different syndromes, respectively. There was an intersection of 284 syndromes proposed by both tools (Figure 2A). A total of 56 syndromes were returned only by DeepGestalt, whereas 844 syndromes were exclusively suggested by GestaltMatcher (Figure 2A).

XSL•FO
RenderX

Figure 2. Comparison of DeepGestalt and GestaltMatcher: (A) Venn diagram showing the number of syndromes supported by DeepGestalt (purple) and GestaltMatcher (grey). (B) Histogramme of top-30 scores yielded by DeepGestalt and GestaltMatcher in patient and control images. Sensitivities of (C) DeepGestalt and (D) GestaltMatcher for syndromes featuring at least 5 analyzed images among all patients included images with mean sensitivity averaged by syndrome. Linear regressions of false positive rates of (E) DeepGestalt's and (F) GestaltMatcher's top-30 results in matched control (y-axis) and patient (x-axis) images of the literature data set. (High resolution image available in Multimedia Appendix 2). FPR: false positive rate; h: healthy control; s: patient with syndrome.

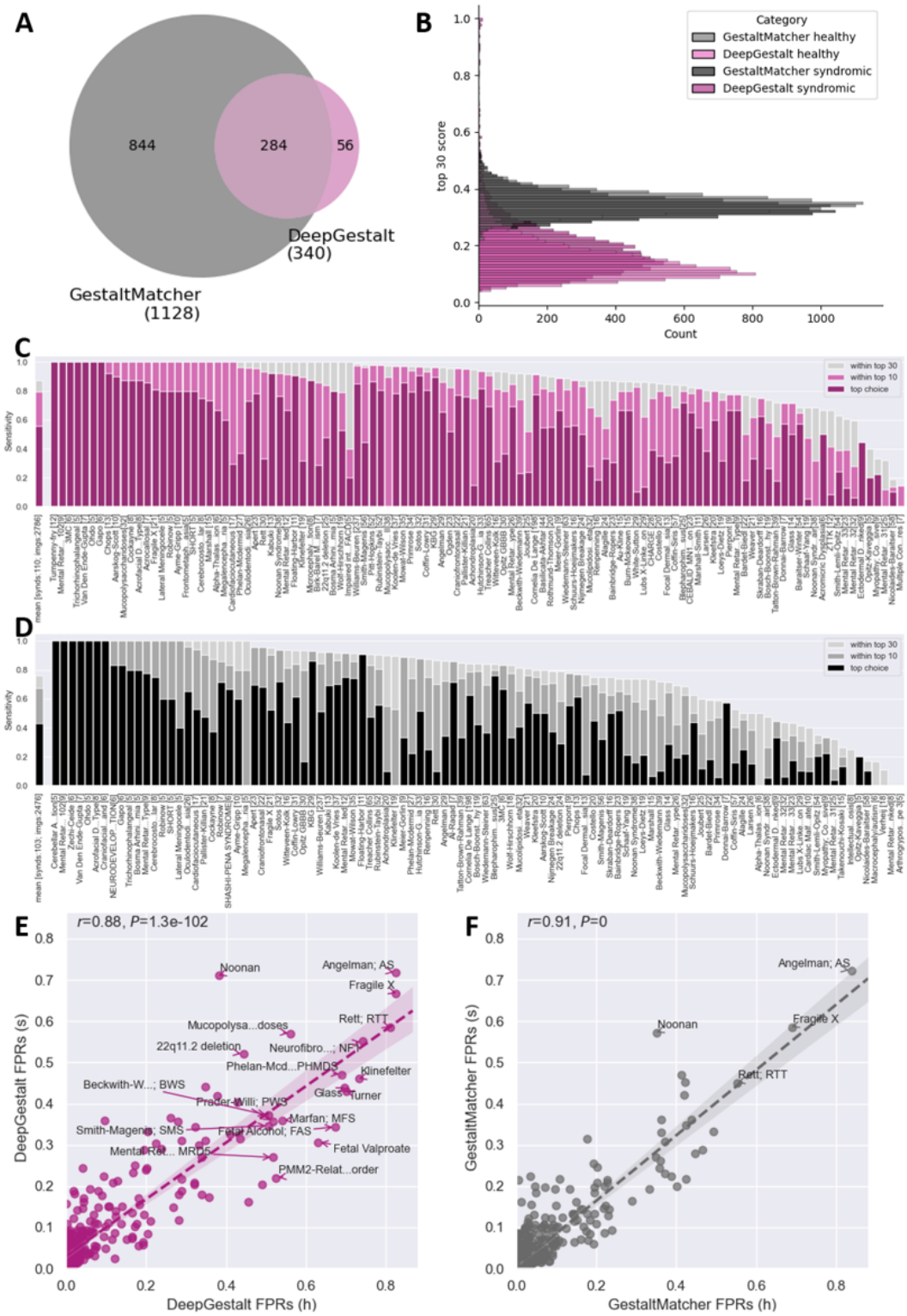

The level of a yielded score is designed to increase with the likelihood of a correct syndrome suggestion. However, this might falsely be interpreted by a user as a direct metric for the probability of a suggested syndrome to be a true positive result. To elucidate the actual distributions, we analyzed the scores returned by DeepGestalt and GestaltMatcher in the top-30 (Figure 2B). Gestalt scores assigned by DeepGestalt ranged predominantly from 0.05 to 0.5, while those assigned by GestaltMatcher were mostly between 0.2 and 0.5. We found that DeepGestalt returned higher scores for syndromic subjects

XSL•FO
RenderX

with a peak around 0.16 compared to a peak around 0.1 for the healthy cohort. In contrast, GestaltMatcher scores showed a higher degree of overlap with a peak around 0.33 for the patient cohort and a peak at even higher values around 0.35 for the controls. Overall, the values assigned by GestaltMatcher were higher than those of DeepGestalt but did not differ between patients and controls.

To measure the tools' quality, we explored their sensitivities for the syndromes included in our data sets. We were able to evaluate 110 syndromes in DeepGestalt and 103 syndromes in GestaltMatcher with at least 5 images per supported syndrome (Figure 2C and D). DeepGestalt and GestaltMatcher both ranked 7 syndromes correctly on top-1 in all cases. The syndromes that were always suggested correctly on rank 1 by both tools were Mental Retardation X-Linked 102, Van Den Ende-Gupta syndrome, and Ohdo syndrome. GestaltMatcher did not suggest the correct diagnosis in the top-30 for 2 syndromes. Overall, DeepGestalt showed higher sensitivities than GestaltMatcher, achieving an average top-1 sensitivity of 56% (SD 29%), an average top-10 sensitivity of 79% (SD 23%), and an average top-30 sensitivity of 88% (SD 18%), while the correct diagnosis was returned by GestaltMatcher at rank 1 on average in 43% (SD 30%), within the top-10 in 67% (29%), and within the top-30 in 76% (SD 26%) of patient images. To directly compare results yielded by the current version with those yielded by an earlier version of DeepGestalt, we also explored the current sensitivities of both tools exclusively on the patient cohort of the "literature image" set. Also on these images, DeepGestalt outperformed GestaltMatcher (top-1 sensitivity 58%, SD 28% vs 38%, SD 24%; top-10 sensitivity 91%, SD 10% vs 79%, SD 19%; and top-30 sensitivity 97%, SD 6% vs 93%, SD 9%; Multimedia Appendix 3).

To test the specificity of syndrome suggestions, we explored false syndrome suggestions by DeepGestalt and GestaltMatcher. The vast majority of syndromes were infrequently suggested by DeepGestalt with only 19 syndromes being suggested among the top-30 ranks in more than 50% of patient or control images (Figure 2E). False syndrome suggestion rates strongly correlated between patient and control images (r=0.88). The most frequently falsely suggested syndrome in either cohort was Angelman syndrome (healthy controls false positive rate [FPR]=0.83 and patient cohorts FPR=0.72). False positive rates of GestaltMatcher were lower with only 4 syndromes being suggested among the top-30 ranks in more than 50% of either patient or healthy control images (Figure 2F). GestaltMatcher false positive rates also showed a strong correlation between patient and control images (r=0.91) and suggested Angelman syndrome as the most frequent false positive result (healthy controls FPR=0.84 and patient cohorts FPR=0.72). False positive suggestion rates of DeepGestalt and GestaltMatcher strongly correlated in both the controls (r=0.91) and the patient cohort (r=0.91, Multimedia Appendix 4). Angelman syndrome was not only the syndrome that was most frequently suggested by either tool among the top-30 ranks, but together with Fragile X syndrome, it was also the most frequently implied among the top-10, top-5, and on the first rank (Multimedia Appendix 5).

Discriminatory Power of GestaltMatcher, DeepGestalt, and D-Score

To benchmark their ability to discern patients from controls, we examined and compared scores yielded by D-Score and first-rank scores of DeepGestalt and GestaltMatcher (Figure 3A). First-rank DeepGestalt scores were lower in the healthy control cohort (median 0.23, IQR 0.169-0.347) than in the patient cohorts (Pantel syndromic median 0.409, IQR 0.255-0.696; LMD median 0.41, IQR 0.217-0.754; and GMDB median 0.37, IQR 0.21-0.685). First-rank DeepGestalt scores showed high variances and overlapped between patient and control images with lower quartiles of the patient cohorts being between the median and upper quartile of the control cohort.

XSL•FO
RenderX

Figure 3. Accuracy of DeepGestalt, GestaltMatcher and D-Score. (A) Distributions of the scores yielded by D-Score (D, turquoise), GestaltMatcher (GM, grey), and DeepGestalt (DG, purple), in the different test sets. (B) Receiver operating characteristic curves (ROC) of D-Score (D, turquoise), GestaltMatcher (GM, grey), and DeepGestalt (DG, purple). (High resolution version available in Multimedia Appendix 6). AUC: area under the curve; GMDB: GestaltMatcher Database; LMD: London Medical Database; ROC: receiver operating characteristic.

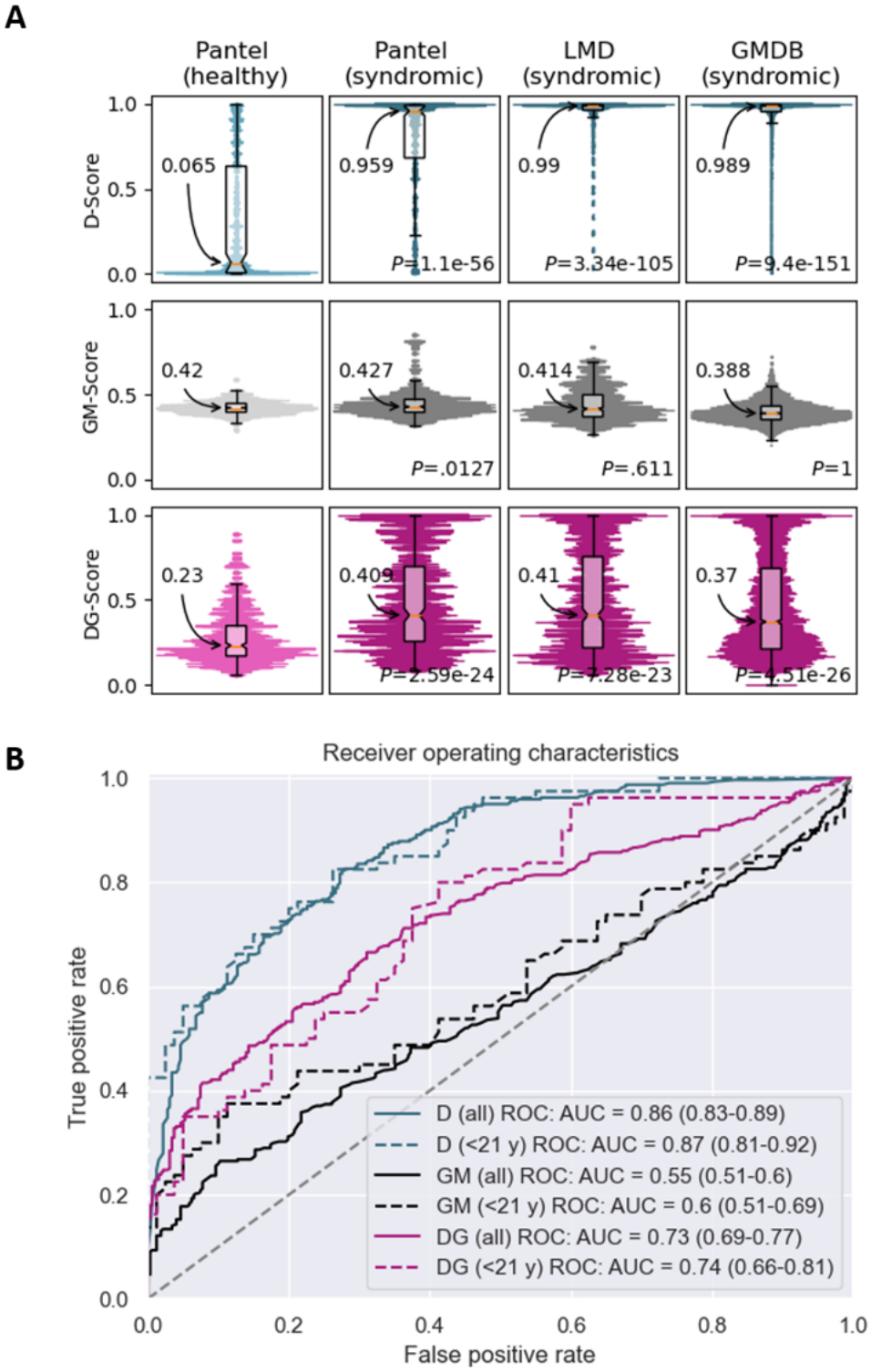

Median first-rank GestaltMatcher scores were higher than median first-rank DeepGestalt scores but similar between control and patient images (Pantel control GestaltMatcher median 0.42, Pantel syndromic GestaltMatcher median 0.427, LMD GestaltMatcher median 0.414, and GMDB GestaltMatcher median 0.388). First-rank GestaltMatcher scores were less variable than the first-rank DeepGestalt scores.

D-Score distributions differed between the control cohort (median 0.065) and the patient cohorts (Pantel syndromic median 0.959, LMD median 0.99, and GMDB median 0.989) with nonoverlapping IQR. D-Score showed little variability with control scores clustering toward 0 and patient scores toward 1.

XSL•FO
RenderX

As a major metric of the power to discern patient and control images, we investigated the receiver operating characteristic (ROCs) curves of DeepGestalt, GestaltMatcher, and D-Score (Figure 3B). D-Score yielded the highest curve with an area under the curve (AUC) of 0.86. DeepGestalt achieved the second-best outcome with an AUC of 0.73. As expected, GestaltMatcher did not achieve separability of the 2 classes with an AUC of 0.55.

Confounders of D-Score

To test for performance differences resulting from patient-specific characteristics known to possibly confound face classification systems, we also examined performance metrics on respective subsets (age, sex, ethnicity, and diagnosis) of our cohort.

Distributions of first-rank scores of DeepGestalt and GestaltMatcher were comparable for all age groups regarding both patient and control images (Multimedia Appendix 7). However, D-scores tended to increase with growing age, particularly for control images (Multimedia Appendix 7). Nevertheless, the ROC of D-scores of minors (younger than 21 years of age, AUC=0.87) was not markedly different from the ROC for all age groups (Figure 3B). Upon evaluation of the 17 syndromes included in the Pantel syndromic cohort, first-rank GestaltMatcher scores did not vary depending on a patient's diagnosis, while we found minor differences in the first-rank DeepGestalt score distributions and strong differences in D-Score distributions (Multimedia Appendix 8).

First-rank DeepGestalt and GestaltMatcher scores did not differ between female and male individuals, whereas D-scores tended to be higher in photos of male individuals than in images of female individuals for both the Pantel syndromic and control cohort. This strong difference was not seen in the larger GMDB cohort (Multimedia Appendix 9).

Strongly significant differences in score distributions of White and other individuals were only seen in the GMDB patient cohort (Multimedia Appendix 10).

Discussion

Principal Findings

We evaluated the recently developed algorithms GestaltMatcher and D-Score in comparison to the established DeepGestalt. D-Score provides an estimate for the degree of dysmorphism of a face, while DeepGestalt showed the best sensitivity by rank. GestaltMatcher offers a large number of detectable syndromes.

Capable of achieving the same task of suggesting dysmorphic syndromes based on patient images, DeepGestalt and GestaltMatcher have specific strengths and weaknesses. When comparing the ranges of scores yielded by DeepGestalt and by GestaltMatcher, we found that GestaltMatcher scores on average exceeded DeepGestalt scores. This information is crucial for use in the clinical field, as physicians must be aware that scores obtained through analysis by the new GestaltMatcher cannot be interpreted identically to scores produced by DeepGestalt. While DeepGestalt shows higher sensitivities for the syndromes it supports, the number of syndromes supported by GestaltMatcher is more than 3 times greater. The average sensitivities of both tools are sufficient for diagnostic decision support. The sensitivity of DeepGestalt version 21.5.0 tested here is comparable to that of version 19.1.7 tested in a previous study [20], suggesting that the version updates carried out so far have had a minor impact on DeepGestalt's performance. Nevertheless, users need to be aware that specificities for a few officially supported syndromes are very low, for example, GestaltMatcher failed to identify 2 syndromes (Mental Retardation X-linked and Arthrogryposis, distal type 3). Less syndromes are frequently falsely suggested by GestaltMatcher than by DeepGestalt, implying GestaltMatcher has a greater specificity. This could be expected as GestaltMatcher includes a much larger number of syndromes than DeepGestalt, lowering the chance of individual syndromes appearing as a match. The tendency of the 2 algorithms to falsely suggest certain syndromes correlated strongly. Among the most frequent false suggestions were Angelman syndrome, Fragile X syndrome, and Rett syndrome which feature rather mild facial dysmorphism. Thus, caution is needed when interpreting classification results listing such diagnoses.

First-rank GestaltMatcher scores of affected and unaffected cases overlapped largely, while surprisingly, unaffected cases received a higher median score than affected cases. Thus, GestaltMatcher can currently not be used to screen for individuals with dysmorphic syndromes. In our experiment, the scores assigned by DeepGestalt allowed for a certain separability between affected and unaffected cohorts. Nevertheless, this accuracy does not enable the identification of patients who are dysmorphic with great certainty. The best accuracy to differentiate between affected and unaffected was found for D-Score. Limiting the test set to the age group predominantly seen for syndrome evaluation (under 21 years of age) slightly improved the accuracy.

Notably, in the patient cohort of Pantel et al [20], D-Score yielded the highest scores in syndromes with a strong facial dysmorphism (highest median score in Apert syndrome) and lower scores in syndromes with mild facial features (lowest median score in Rett syndrome). D-Scores of control images increased with age, suggesting it is predominantly trained on children. We suggest that clinicians may exercise caution when interpreting scores obtained for patients older than 21 years of age, as older faces may automatically be recognized as dysmorphic due to signs of aging. D-Score assigned higher scores to images of healthy male individuals than healthy female individuals, which needs to be taken into consideration when interpreting the obtained score values. Ethnic background was not shown to influence the performance of either of the 3 algorithms tested. However, future studies ideally include ethnically more diverse test sets.

Strengths and Limitations

To our knowledge, we performed the most extensive evaluation of DeepGestalt thus far and the first comprehensive evaluations of GestaltMatcher and D-Score. Nevertheless, there are limitations to the analyses we conducted in this study. First, we restricted our analysis of GestaltMatcher to the first 30 results to match the number of results of DeepGestalt provided via

XSL•FO
RenderX

Face2Gene CLINIC. As GestaltMatcher can match a far larger number of patients and syndromes, an analysis including all possible 100 results, using a larger number of syndromes and even images of undiagnosed patients, would provide a more comprehensive assessment of its performance. Second, the number of healthy control images is limited, and we tested for possible confounders of D-Score, which constrained the number of (both patient and control) cases in some subgroups. To evaluate the influence of these and other possible confounders such as facial hair, glasses, positioning, and lighting, further research using larger test sets is necessary. Third, a direct comparison of the D-score to other algorithms, designed to discern images of patients who are dysmorphic from those of healthy controls is needed. Recently, Porras et al [13] described such a tool achieving an AUC of 0.94 on a private data set. Unfortunately, no tool other than D-score is available for the broader public for testing purposes. Thus, benchmarking different tools on the same data set has not been possible.

Future Directions

Even though the algorithms cannot replace the expertise of a physician in medical genetics, they could potentially play a complementary role in future decision-making. D-Score showed the best separation between syndromic and nonsyndromic images, giving it a potential future role as a screening tool. GestaltMatcher could be applied as a “second-line” test to analyze patients’ images that yielded no or only low-similarity matches in the DeepGestalt analysis.

Physicians must exercise caution when using tools tested in this study, as the scores assigned vary between the tools and can thus not be interpreted identically. A prospective study, as recently conducted to assess the application of DeepGestalt in the clinical field [21], is needed to further investigate the functionality of GestaltMatcher and D-Score in a point-of-care setting.

Acknowledgments

We thank Nicole Fleischer, Aviram Bar-Haim, and Guilherme Bonini of FDNA Inc for technical assistance. MD and MAM are participants in the Berlin Institute of Health (BIH) Charité (Digital) Clinician Scientist Program founded by late Duska Dragun and funded by the Charité—Universitätsmedizin Berlin and the Berlin Institute of Health.

Data Availability

Characteristics of the literature image data analyzed during this study are available in a previous publication [20]. Image data from the London Medical Database analyzed during this study are available in the Face2Gene Library [31]. Image Data of the GestaltMatcher Database analyzed during this study are available therein [32]. All data generated during this study are included in this published article and Multimedia Appendices 1-10. High resolution versions of Figures 2 and 3 can be found in Multimedia Appendices 2 and 6.

Authors' Contributions

JTP and MAM Contributed to the conceptualization; AMVR and MAM performed the investigation; AMVR and MAM contributed to the formal analysis; AMVR, CEO, and MAM contributed to writing—original draft; JTP, MD, and DH contributed to writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Code and data required to reproduce statistical results from this study.
[ZIP File (Zip Archive), 26162 KB - jmir_v26i1e42904_app1.zip]

Multimedia Appendix 2
Comparison of DeepGestalt and GestaltMatcher (high resolution).
[PNG File , 597 KB - jmir_v26i1e42904_app2.png]

Multimedia Appendix 3
Sensitivities of (A) DeepGestalt and (B) GestaltMatcher in the patient cohort of the "literature image" set. Higher bars indicate higher mean sensitivity, averaged by syndrome. Far left: mean sensitivity across all represented syndromes.
[PNG File , 234 KB - jmir_v26i1e42904_app3.png]

Multimedia Appendix 4
Linear regression of DeepGestalt’s (x-axis) and GestaltMatcher’s (y-axis) false positive rates (FPRs) in matched control (A) and patient (B) images of the “literature image” set.

XSL•FO
RenderX

[PNG File , 200 KB - jmir_v26i1e42904_app4.png]

Multimedia Appendix 5
Linear regressions of false positive rates (FPRs) of DeepGestalt's (purple) and GestaltMatcher's (grey) in patient (x-axis) matched control (y-axis) images of the literature data set. (A,B) Calculated from top-10 results; (C,D) Calculated from top-5 results; (E,F) Calculated from top-1 suggestions only.
[PNG File , 434 KB - jmir_v26i1e42904_app5.png]

Multimedia Appendix 6
Accuracy of DeepGestalt, GestaltMatcher and D-Score (high resolution).
[PNG File , 438 KB - jmir_v26i1e42904_app6.png]

Multimedia Appendix 7
Distributions of the scores yielded by D-Score (turquoise), GestaltMatcher (grey), and DeepGestalt (purple) in the different test sets, separated by age groups. h: healthy control, s: patient with syndrome, P: Pantel image set, G: GestaltMatcher Database.
[PNG File , 223 KB - jmir_v26i1e42904_app7.png]

Multimedia Appendix 8
Distributions of the scores yielded by D-Score (turquoise), GestaltMatcher (grey), and DeepGestalt (purple) in the Pantel patient set separated by diagnosis. The *P* values refer to differences to matched healthy controls.
[PNG File , 231 KB - jmir_v26i1e42904_app8.png]

Multimedia Appendix 9
Distributions of the scores yielded by D-Score (turquoise), GestaltMatcher (grey), and DeepGestalt (purple) separated by sex. Heal: healthy control, synd: patients with a syndrome, f: female, m: male, GMDB: GestaltMatcher Database.
[PNG File , 165 KB - jmir_v26i1e42904_app9.png]

Multimedia Appendix 10
Distributions of the scores yielded by D-Score (turquoise), GestaltMatcher (grey), and DeepGestalt (purple) separated by ethnicity. Heal: healthy control, synd: patients with a syndrome, Wh: White, Oth.: Other, GMDB: GestaltMatcher Database.
[PNG File , 162 KB - jmir_v26i1e42904_app10.png]

References

1. Jayaratne YSN, Zwahlen RA. Application of digital anthropometry for craniofacial assessment. Craniomaxillofac Trauma Reconstr 2014;7(2):101-107 [FREE Full text] [doi: 10.1055/s-0034-1371540] [Medline: 25050146]
2. Lumaka A, Cosemans N, Mampasi AL, Mubungu G, Mvuama N, Lubala T, et al. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. Clin Genet 2017;92(2):166-171. [doi: 10.1111/cge.12948] [Medline: 27925162]
3. Ferreira CR. The burden of rare diseases. Am J Med Genet A 2019;179(6):885-892. [doi: 10.1002/ajmg.a.61124] [Medline: 30883013]
4. Boehringer S, Vollmar T, Tasse C, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, et al. Syndrome identification based on 2D analysis software. Eur J Hum Genet 2006;14(10):1082-1089 [FREE Full text] [doi: 10.1038/sj.ejhg.5201673] [Medline: 16773127]
5. Vollmar T, Maus B, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, Wieczorek D, et al. Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces. Eur J Med Genet 2008;51(1):44-53. [doi: 10.1016/j.ejmg.2007.10.002] [Medline: 18054308]
6. Boehringer S, Guenther M, Sinigerova S, Wurtz RP, Horsthemke B, Wieczorek D. Automated syndrome detection in a set of clinical facial photographs. Am J Med Genet A 2011;155A(9):2161-2169. [doi: 10.1002/ajmg.a.34157] [Medline: 21815261]
7. Ferry Q, Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, et al. Diagnostically relevant facial gestalt information from ordinary photos. Elife 2014;3:e02020 [FREE Full text] [doi: 10.7554/eLife.02020] [Medline: 24963138]
8. Cerrolaza JJ, Porras AR, Mansoor A, Zhao Q, Summar M, Linguraru MG. Identification of dysmorphic syndromes using landmark-specific local texture descriptors. : IEEE; 2016 Presented at: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); April 13-16, 2016; Prague, Czech Republic. [doi: 10.1109/isbi.2016.7493453]
9. Tu L, Porras AR, Boyle A, Linguraru MG. Analysis of 3D facial dysmorphology in genetic syndromes from unconstrained 2D photographs. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing

XSL•FO
RenderX

and Computer Assisted Intervention—MICCAI: 21st International Conference Granada, Spain, September 16-20, 2018 Proceeding, Part I: LNCS 11070. Cham: Springer; 2018:347-355.
10. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med 2019;25(1):60-64. [doi: 10.1038/s41591-018-0279-0] [Medline: 30617323]
11. Dudding-Byth T, Baxter A, Holliday EG, Hackett A, O'Donnell S, White SM, et al. Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. BMC Biotechnol 2017;17(1):90 [FREE Full text] [doi: 10.1186/s12896-017-0410-1] [Medline: 29258477]
12. Hallgrímsson B, Aponte JD, Katz DC, Bannister JJ, Riccardi SL, Mahasuwan N, et al. Automated syndrome diagnosis by three-dimensional facial imaging. Genet Med 2020;22(10):1682-1693 [FREE Full text] [doi: 10.1038/s41436-020-0845-y] [Medline: 32475986]
13. Porras AR, Rosenbaum K, Tor-Diez C, Summar M, Linguraru MG. Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study. Lancet Digit Health 2021;3(10):e635-e643 [FREE Full text] [doi: 10.1016/S2589-7500(21)00137-0] [Medline: 34481768]
14. Pantel JT, Zhao M, Mensah MA, Hajjir N, Hsieh TC, Hanani Y, et al. Advances in computer-assisted syndrome recognition by the example of inborn errors of metabolism. J Inherit Metab Dis 2018;41(3):533-539 [FREE Full text] [doi: 10.1007/s10545-018-0174-3] [Medline: 29623569]
15. Mishima H, Suzuki H, Doi M, Miyazaki M, Watanabe A, Matsumoto T, et al. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. J Hum Genet 2019;64(8):789-794. [doi: 10.1038/s10038-019-0619-z] [Medline: 31138847]
16. Zarate YA, Bosanko KA, Gripp KW. Using facial analysis technology in a typical genetic clinic: experience from 30 individuals from a single institution. J Hum Genet 2019;64(12):1243-1245. [doi: 10.1038/s10038-019-0673-6] [Medline: 31551534]
17. Narayanan DL, Ranganath P, Aggarwal S, Dalal A, Phadke SR, Mandal K. Computer-aided facial analysis in diagnosing dysmorphic syndromes in Indian children. Indian Pediatr 2019;56(12):1017-1019 [FREE Full text] [Medline: 31884430]
18. Hsieh TC, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, et al. PEDIA: prioritization of exome data by image analysis. Genet Med 2019;21(12):2807-2814 [FREE Full text] [doi: 10.1038/s41436-019-0566-2] [Medline: 31164752]
19. Elmas M, Gogus B. Success of face analysis technology in rare genetic diseases diagnosed by whole-exome sequencing: a single-center experience. Mol Syndromol 2020;11(1):4-14 [FREE Full text] [doi: 10.1159/000505800] [Medline: 32256296]
20. Pantel JT, Hajjir N, Danyel M, Elsner J, Abad-Perez AT, Hansen P, et al. Efficiency of computer-aided facial phenotyping (DeepGestalt) in individuals with and without a genetic syndrome: diagnostic accuracy study. J Med Internet Res 2020;22(10):e19263 [FREE Full text] [doi: 10.2196/19263] [Medline: 33090109]
21. Marwaha A, Chitayat D, Meyn MS, Mendoza-Londono R, Chad L. The point-of-care use of a facial phenotyping tool in the genetics clinic: enhancing diagnosis and education with machine learning. Am J Med Genet A 2021;185(4):1151-1158. [doi: 10.1002/ajmg.a.62092] [Medline: 33554457]
22. Liehr T, Acquarola N, Pyle K, St-Pierre S, Rinholm M, Bar O, et al. Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. Clin Genet 2018;93(2):378-381. [doi: 10.1111/cge.13087] [Medline: 28661575]
23. Pascolini G, Fleischer N, Ferraris A, Majore S, Grammatico P. The facial dysmorphology analysis technology in intellectual disability syndromes related to defects in the histones modifiers. J Hum Genet 2019;64(8):721-728. [doi: 10.1038/s10038-019-0598-0] [Medline: 31086247]
24. Danyel M, Cheng Z, Jung C, Boschann F, Pantel JT, Hajjir N, et al. Differentiation of MISSLA and Fanconi anaemia by computer-aided image analysis and presentation of two novel MISSLA siblings. Eur J Hum Genet 2019;27(12):1827-1835 [FREE Full text] [doi: 10.1038/s41431-019-0469-3] [Medline: 31320746]
25. Kruszka P, Hu T, Hong S, Signer R, Cogné B, Isidor B, et al. Phenotype delineation of ZNF462 related syndrome. Am J Med Genet A 2019;179(10):2075-2082 [FREE Full text] [doi: 10.1002/ajmg.a.61306] [Medline: 31361404]
26. Weiss K, Lazar HP, Kurolap A, Martinez AF, Paperna T, Cohen L, et al. The CHD4-related syndrome: a comprehensive investigation of the clinical spectrum, genotype-phenotype correlations, and molecular basis. Genet Med 2020;22(2):389-397 [FREE Full text] [doi: 10.1038/s41436-019-0612-0] [Medline: 31388190]
27. Carli D, Giorgio E, Pantaleoni F, Bruselles A, Barresi S, Riberi E, et al. NBAS pathogenic variants: defining the associated clinical and facial phenotype and genotype-phenotype correlations. Hum Mutat 2019;40(6):721-728 [FREE Full text] [doi: 10.1002/humu.23734] [Medline: 30825388]
28. Staufner C, Peters B, Wagner M, Alameer S, Barić I, Broué P, et al. Defining clinical subgroups and genotype-phenotype correlations in NBAS-associated disease across 110 patients. Genet Med 2020;22(3):610-621 [FREE Full text] [doi: 10.1038/s41436-019-0698-4] [Medline: 31761904]
29. Mak BC, Russo RS, Gambello MJ, Fleischer N, Black ED, Leslie E, et al. Craniofacial features of 3q29 deletion syndrome: application of next-generation phenotyping technology. Am J Med Genet A 2021;185(7):2094-2101 [FREE Full text] [doi: 10.1002/ajmg.a.62227] [Medline: 33938623]

XSL•FO
RenderX

30. Hsieh TC, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nat Genet 2022;54(3):349-357 [FREE Full text] [doi: 10.1038/s41588-021-01010-x] [Medline: 35145301]
31. Introducing the pediatrician view. Face2Gene. URL: https://www.face2gene.com/ [accessed 2023-12-29]
32. Analyze human sequence variants. GeneTalk. URL: https://www.gene-talk.de/ [accessed 2023-12-29]

Abbreviations

AUC: area under the curve
AUROC: area under the receiver operating characteristic
FPR: false positive rate
GMDB: Gestalt Matcher Database
LMD: London Medical Database
NGP: next-generation phenotyping
ROC: receiver operating characteristic

Edited by G Eysenbach, T Leung; submitted 27.09.22; peer-reviewed by T Liehr, P Robinson; comments to author 16.02.23; revised version received 19.04.23; accepted 17.11.23; published 13.03.24.

Please cite as:
Reiter AMV, Pantel JT, Danyel M, Horn D, Ott CE, Mensah MA
Validation of 3 Computer-Aided Facial Phenotyping Tools (DeepGestalt, GestaltMatcher, and D-Score): Comparative Diagnostic Accuracy Study
J Med Internet Res 2024;26:e42904
URL: https://www.jmir.org/2024/1/e42904
doi:10.2196/42904
PMID:38477981

©Alisa Maria Vittoria Reiter, Jean Tori Pantel, Magdalena Danyel, Denise Horn, Claus-Eric Ott, Martin Atta Mensah. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mining Real-World Big Data to Characterize Adverse Drug Reaction Quantitatively: Mixed Methods Study

Qi-Xuan Yue[1,2*], ME; Ruo-Fan Ding[1*], PhD; Wei-Hao Chen[1,2], MS; Lv-Ying Wu[1], MS; Ke Liu[1], PhD; Zhi-Liang Ji[1,2,3], PhD

[1]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, China
[2]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[3]Fujian Provincial Key Laboratory of Chemical Biology, Xiamen University, Xiamen, China
[*]these authors contributed equally

Corresponding Author:
Zhi-Liang Ji, PhD
State Key Laboratory of Cellular Stress Biology, School of Life Sciences
Faculty of Medicine and Life Sciences
Xiamen University
Number 4221 Xiang'an South Road
Xiang'an District
Xiamen, 361102
China
Phone: 86 0592 2182897
Email: appo@xmu.edu.cn

Abstract

Background: Adverse drug reactions (ADRs), which are the phenotypic manifestations of clinical drug toxicity in humans, are a major concern in precision clinical medicine. A comprehensive evaluation of ADRs is helpful for unbiased supervision of marketed drugs and for discovering new drugs with high success rates.

Objective: In current practice, drug safety evaluation is often oversimplified to the occurrence or nonoccurrence of ADRs. Given the limitations of current qualitative methods, there is an urgent need for a quantitative evaluation model to improve pharmacovigilance and the accurate assessment of drug safety.

Methods: In this study, we developed a mathematical model, namely the Adverse Drug Reaction Classification System (ADReCS) severity-grading model, for the quantitative characterization of ADR severity, a crucial feature for evaluating the impact of ADRs on human health. The model was constructed by mining millions of real-world historical adverse drug event reports. A new parameter called *Severity_score* was introduced to measure the severity of ADRs, and upper and lower score boundaries were determined for 5 severity grades.

Results: The ADReCS severity-grading model exhibited excellent consistency (99.22%) with the expert-grading system, the Common Terminology Criteria for Adverse Events. Hence, we graded the severity of 6277 standard ADRs for 129,407 drug-ADR pairs. Moreover, we calculated the occurrence rates of 6272 distinct ADRs for 127,763 drug-ADR pairs in large patient populations by mining real-world medication prescriptions. With the quantitative features, we demonstrated example applications in systematically elucidating ADR mechanisms and thereby discovered a list of drugs with improper dosages.

Conclusions: In summary, this study represents the first comprehensive determination of both ADR severity grades and ADR frequencies. This endeavor establishes a strong foundation for future artificial intelligence applications in discovering new drugs with high efficacy and low toxicity. It also heralds a paradigm shift in clinical toxicity research, moving from qualitative description to quantitative evaluation.

(J Med Internet Res 2024;26:e48572) doi:10.2196/48572

KEYWORDS

clinical drug toxicity; adverse drug reaction; ADR severity; ADR frequency; mathematical model

XSL•FO
RenderX

Introduction

In recent years, the issue of drug toxicity has emerged as a serious concern in the fields of clinical medicine, pharmacology, and sociology. As a result, drug regulatory agencies worldwide are making continuous efforts to monitor marketed drugs and assess their potential risks in large populations [1,2]. The precision medicine projects launched around the world have drawn significant attention to the adverse effects of drug therapy. As a result, high-risk drugs have been subject to warnings or even market withdrawal. In the field of new drug discovery, the focus of drug safety evaluation has shifted from early-stage cell or animal toxicity to later-stage clinical toxicity [3].

Adverse drug reactions (ADR) represent the clinical manifestations of drug toxicity in humans. Fundamentally, information on ADRs can be obtained from the 4-phase clinical trials, primarily through postmarketing surveillance. To evaluate the benefits and risks of medicines, monitoring and reporting systems, such as the US Food and Drug Administration Adverse Event Reporting System (FAERS) and the European database of suspected adverse drug reaction reports (EudraVigilance), were established by regulatory authorities. These systems enable researchers to access real-world individual responses to drug therapy. Intelligent tools can be compiled to mine drug-ADR associations, illustrate drug toxicity mechanisms, and predict novel ADRs. In addition, some leading-edge projects like Tox21 and eTRANSAFE have been launched to develop integrative data infrastructure and innovative computational methods. These projects aim to enhance translational safety assessment during the drug development process. High-end applications often use machine learning or artificial intelligence algorithms to automatically correlate chemicals with toxicity, even in cases where the exact molecular mechanisms underlying toxicity are not known [4-6]. For example, Kuang et al [7] built machine learning models using topological information from the drug-ADR associations network, drug chemical structures, and drug Anatomical Therapeutic Chemical (ATC) classification information to discover new drug-ADR associations. Anjani and colleagues [8] constructed a convolutional neural network model solely using drug chemical structures for the prediction of ADR occurrence. However, due to the absence of high-dimensional toxicity information such as ADR severity and frequency, current ADR prediction models are insufficient to comprehensively assess the true impact of drug toxicity on human health [9]. ADR severity is a critical indicator that manifests the seriousness of the ADR's impact on human health, while ADR frequency (occurrence rate) is a quantitative parameter that reflects how often the ADR occurs in the population receiving drug therapy. These 2 parameters are crucial for accurately characterizing drug toxicity in humans. Previously, Hartwig and his colleagues [10] proposed 7 levels of ADR severity and graded 367 ADRs in 1992. Gottlieb et al [11] used nonprofessional crowdsourcing to rank the severity of 2929 ADRs. In 2010, a total of 11 French hospitals collaborated to investigate and determine the severity and frequency of ADRs resulting from self-medication among emergency department patients [12]. Ferreira et al [13] conducted a cross-sectional study to determine the severity and frequency of ADRs based on reports from drug treatments for Alzheimer disease in a Brazilian city. In 2017, the National Cancer Institute released version 5.0 of the Common Terminology Criteria for Adverse Events (CTCAE) [14], which includes 837 distinct adverse event terms observed in cancer therapy. Regrettably, many of these efforts were qualitative in nature.

Undoubtedly, qualitative descriptions of ADRs have become insufficient to support advanced computational algorithms and reduce the use of animals in preclinical toxicity testing. It also hampers the widespread adoption of quantitative methods for cautious pharmacovigilance and prospective assessment of clinical drug toxicity. Therefore, in this study, we aimed to develop mathematical models for estimating ADR severity and frequency by mining millions of historical medical reports over a 10-year period at the population level. Additionally, we created a benchmark data set of drug-ADR relations with quantitative features to support advanced computational applications in drug toxicity research.

Methods

Rationale

The rationale for quantifying ADR severity was graphically illustrated in Figure 1, and we briefly described the principles here. A cohort of patients ($M_{Di–Aj} \in M$) takes the drug D_i ($D_i \in D$, $D = \{D_1, D_2, \ldots, D_n\}$) and thus induce the ADR A_j ($A_j \in A$, $A = \{A_1, A_2, \ldots, A_m\}$) (the D_i-A_j pair). Treating the patients with A_j may have 5 clinical outcomes in FAERS: recovered, recovering, not recovered, resolved with sequelae, and fatal (corresponding to $O = \{O_1, O_2, O_3, O_4, O_5\}$, respectively). While clinical outcomes in response to ADR treatment are influenced by multiple factors, such as primary medical status, treatment location (in or outside the hospital), individual genetic variation, and others, the primary determinant remains the severity of the ADR itself. This inspires us to develop a penalized model that uses a statistical score to quantitatively characterize the primary outcome, which is the most frequently occurring, of ADR treatment in large patient populations. In the model, various outcomes of ADR treatment will be penalized differently, with the "recovered" outcome receiving the highest penalty and the "fatal" outcome receiving the lowest. The model calculates the cumulative effects of ADR treatment through penalties, with the primary outcome contributing the most. Summing up the penalties across large populations $M_{Di–Aj}$ would then reveal the primary treatment outcomes of ADR A_j induced by the drug D_i. The penalty score (later referred to as *Severity_score*) can also be used to indicate the severity of the ADR in the D_i-A_j pair. In most cases, an ADR ($A_j \in A$) can be induced by multiple drugs ($D_n \in D$); hence, A_j may have several scores (corresponding to different D_i-A_j pairs), indicating varying severities for A_j.

To grade ADR severity, we assume a congruent relationship between ADR treatment outcomes and ADR severity. Therefore, we divide the penalty scores of all D_i-A_j pairs in large populations M into 5 zones, corresponding to 5 severity grades. The upper and lower score boundaries for the 5 zones in the

XSL•FO
RenderX

entire score distribution can be theoretically determined by calculating the scores for the extreme cases where only 1 of the 5 treatment outcomes occurs. Accordingly, severity grades for A_j can be assigned by determining the zones where the penalty scores of A_j fall.

Figure 1. Schematic illustration of ADR severity quantification process. (A) Mining and processing of drug-ADR relations; (B) Estimate the ADR *Severity_score*; (C) Determine boundarise of *Severity_score*; (D) Assignment of ADR severity grade. ADR: adverse drug reaction.

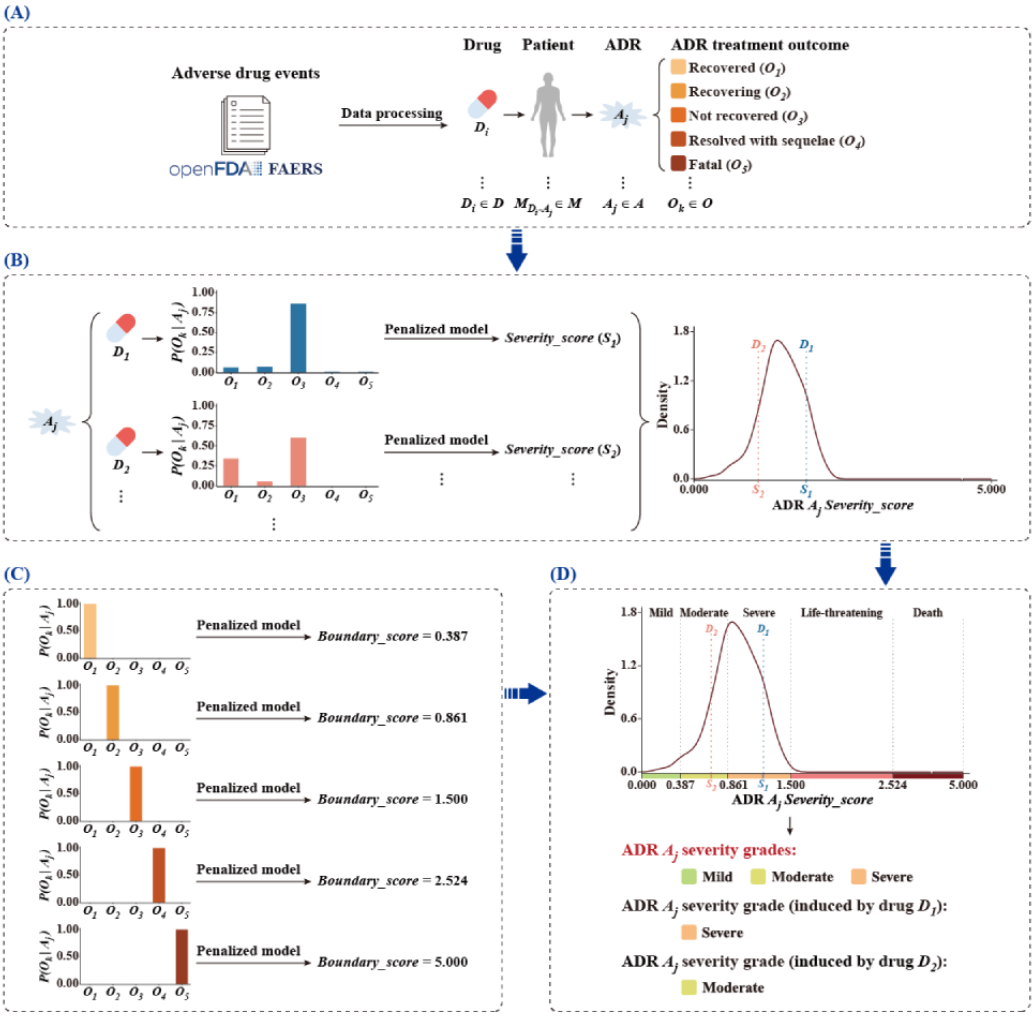

Mining, Processing, and Normalization of Drug-ADR Relations

The drug-ADR associations used for model construction were extracted from the extensive real-world historical adverse drug event (ADE) reports in the FAERS. The FAERS is a shared database maintained by the US Food and Drug Administration. It contains reports of adverse events, including ADRs, medication errors, and drug product problems. These reports are submitted by health care professionals, consumers, and manufacturers during the postmarketing use of drug and therapeutic biologic products. Since 2012, information on ADR treatment outcomes has been recorded in the FAERS. Therefore, data since 2012 can be used for the quantitative study of ADR. The FAERS data files, dating from the first quarter of 2012 to the first quarter of 2020, were downloaded in JSON format from openFDA [15]. These files included 9,146,439 ADEs.

To ensure the reliability of drug-ADR associations, the following operations were performed: (1) qualified ADE reports were obtained by excluding redundant, unreliable, and unrelated reports. Reports submitted by nonprofessionals, such as consumers and patients, were considered unreliable. Additionally, nondrug-induced ADE reports under the System Organ Class (SOC) categories of "congenital, familial and genetic disorders," "surgical and medical procedures," "social circumstances," "product issues," and "injury, poisoning, and procedural complications" were considered unrelated. Only ADE reports involving single-ingredient small molecule drugs were included in this study; (2) to reduce the data bias caused by occasional reports, misinformation, or underreporting, we excluded ADRs that were reported less than two times and drugs that were reported less than 26 times in all reports; (3) drug-ADR pairs reported in less than two separate reports were also discarded; and (4) the reporting odds ratio (*ROR*) for ADR A_j induced by drug D_i was calculated based on the two-by-two contingency table. In this study, the *ROR* was determined using the formula:

Where a represents the number of reports in which the A_j is caused by the D_i, b represents the number of reports in which

XSL•FO
RenderX

other ADRs are caused by the D_i, c represents the number of reports in which A_j is induced by other drugs, and d represents the number of reports in which other ADRs are caused by other drugs. A value of ROR>1.0 typically indicates that the drug is a risk factor for ADR, suggesting a reliable drug-ADR association. A 95% CI was used for statistical analysis, and a significance level of P<.05 was applied.

Before constructing the model, the reliable drug-ADR pairs underwent preprocessing and normalization to eliminate redundancy. The drugs were standardized based on their main active ingredients. Pharmacological and chemical data about the drugs, including drug description, indication, synonyms, structure, PubChem ID, DrugBank ID, KEGG (Kyoto Encyclopedia of Genes and Genomes) ID, ATC code, National Drug Code product code, and targets, were extracted from publicly available medical repositories. These repositories included the Unified Medical Language System [16], DrugBank [17], PubChem [18], KEGG [19], and the ATC classification system [20]. The drug names were consolidated by cross-referencing with the DrugBank database. The ADRs were standardized by associating them with the standard ADR terms in the Adverse Drug Reaction Classification System (ADReCS) [21] using self-coded scripts. When encountering novel ADR terms, we adhered to the ADR standardization protocol outlined in ADReCS (version 1.2). This involved standardizing the ADR terms, establishing an ADR hierarchy, and assigning digital IDs.

Construction of the ADReCS Severity-Grading System for ADR Severity Estimation

Constructing the Penalized Model to Estimate the ADR Severity_score

To quantitatively characterize the primary outcome of ADR treatment in large populations, we developed a penalized model. We represented the drugs in the qualified ADE reports of FAERS as $D = \{D_1, D_2, \ldots, D_n\}$, the collection of ADRs as $A = \{A_1, A_2, \ldots, A_m\}$, the distinct ADE reports as M (which was nearly equal to the size of patient populations, with different events of the same patient treated as a separate report), and the 5 outcomes of ADR treatment as $O = \{O_1, O_2, O_3, O_4, O_5\}$, corresponding to the clinical end points of recovered, recovering, not recovered, resolved with sequelae, and fatal, respectively. The model penalized A_j according to the clinical outcomes in a reciprocal manner. The "recovered" outcome was penalized the most, while the "fatal" outcome was penalized the least. We defined the penalty scheme as $Penalty(O_k) = \{5, 4, 3, 2, 1\}$, corresponding to the outcomes O one-by-one in order. Hence, for an ADR A_j ($A_j \in A$) induced by the drug D_i ($D_i \in D$) (the D_i-A_j pair), the overall penalty scores of ADR treatment outcomes can be determined by:

To eliminate potential data bias (eg, cases of A_j-O_k combinations may vary greatly) in FAERS, we introduced 2 parameters: the conditional probability of A_j with treatment outcome O_k (that is $P(O_k|A_j)$ in M which was determined by counting the number of A_j-O_k cases against all A_j-O cases, and the association strength of D_i-A_j pair (ROR) which was further normalized with the sigmoid function to a range of 0.5 to 1.0.

To emphasize severe ADRs, we further introduced a weight parameter w_k ($w_k = \{1, 2, 3, 4, 5\}$). Since the primary treatment outcomes contributed most to the sum-up penalty score, we could use the penalty score as the indicator of the primary outcome of A_j and further infer the A_j severity as well. Hence, we defined the ADR severity score, $Severity_score_{Aj}$, which can be determined by:

The *Severity_score* ranges from 0 to 5.0; the larger the score is, the more severe the ADR will be.

Assignment of ADR Severity Grade by the Severity_score

Most of the current expert systems follow the generally accepted principles or rules for ADR severity grading; however, the grades and the rules between systems are different and the rules are sometimes uneasy for experts to follow exactly. In this study, we used the CTCAE's 5-grade architecture, including mild, moderate, severe, life-threatening, and death. To determine the appropriate *Severity_score* boundaries for grading ADR severity, we assumed that there was a congruent relationship between ADR treatment outcomes and ADR severity (actually, they were not fully matched). Hence, we calculated the *Severity_score*s for the extreme cases, for instance, when the ADRs in FAERS have only 1 of 5 treatment outcomes (that is $P(O_k|A_j) = 1$, $ROR \rightarrow \infty$, and $1 + e^{-log2(ROR)} \approx 1$). Accordingly, we can denote the theoretical upper and lower boundary of *Severity_score*s for ADR severity grades:

Thereby, we determined the threshold scores for each severity grade as given in Table 1.

XSL•FO
RenderX

Table 1. Architecture comparison of ADR[a] severity-grading systems.

ADReCS[b] severity-grading system[c]	Vertigo (ADReCS)[d]	CTCAE[e] version 5.0	Vertigo (CTCAE version 5.0)[f]
Mild (0.000-0.387)	Mild (0.196-0.385)	Grade 1: mild	Mild symptoms
Moderate (0.387-0.861)	Moderate (0.394-0.839)	Grade 2: moderate	Moderate symptoms; limiting instrumental ADL[g]
Severe (0.861-1.500)	Severe (0.875-1.179)	Grade 3: severe but not immediately life-threatening	Severe symptoms; limiting self-care ADL
Life-threatening (1.5-2.524)	__[h]	Grade 4: life-threatening consequences	—
Death (2.524-5.000)	—	Grade 5: death related to AE[i]	—

[a]ADR: adverse drug reaction.

[b]ADReCS: Adverse Drug Reaction Classification System.

[c]The value in the brackets stands for the upper-lower boundaries of *Severity_score*. Values are limited to 3 decimal places.

[d]Using "Vertigo" as an example. The value in the brackets represents the range of *Severity_score* for vertigo estimation using the ADReCS severity-grading system.

[e]CTCAE: Common Terminology Criteria for Adverse Events.

[f]Using "Vertigo" as an example.

[g]ADL: activities of daily living.

[h]An em dash (—) indicates that a grade is not available.

[i]AE: adverse event.

Evaluation of the ADReCS Severity-Grading System

The applicability of the ADReCS severity-grading system in estimating ADR severity was evaluated by comparing it with the widely recognized expert-based system, CTCAE. CTCAE classifies ADR severity into 5 grades: mild, moderate, severe, life-threatening, and death. As CTCAE grades are primarily designed for cancer therapy, the evaluation focused on the mutually preferred terms (PTs) shared by both grading systems. CTCAE, version 5.0, includes 837 MedDRA Lowest Level Terms, corresponding to 729 PTs, of which 658 terms align with those in ADReCS. However, 71 terms in CTCAE could not be matched with terms in ADReCS. These unmatched ADRs often stem from clinical treatments other than drug therapy, such as pulmonary valve disease, vaccination site lymphadenopathy, and vaccination complications.

The evaluation was conducted on the mutual 658 PTs by comparing the consistency of ADR severity grades assigned by both systems. The correspondence of ADR severity grades between both systems was summarized in Table 1. Considering an ADR may have several grades subject to different drug therapies, we classified the comparison results of severity grades into three states (Figure 2A): (1) "consistent" if the CTCAE grades were fully matched or covered by the ADReCS severity grades, (2) "partially consistent" if the CTCAE grades overlapped with the ADReCS severity grades, and (3) "inconsistent" when the grades of both systems were exclusive to each other.

XSL•FO
RenderX

Figure 2. Schematic illustration of the results of ADR severity grading. (A) Evaluation of the ADReCS severity-grading system with the expert system CTCAE, along with illustrated rules and evaluation outcomes. (B) Statistics of ADRs with the information on severity grade and frequency by SOCs. (C) Density distribution of drug-ADR pairs based on Severity_scores. (D) Distribution of drug-ADR pairs categorized by ADR severity grades. ADR: adverse drug reaction; ADReCS: Adverse Drug Reaction Classification System; CTCAE: Common Terminology Criteria for Adverse Events; PT: Preferred Term; SOC: System Organ Class.

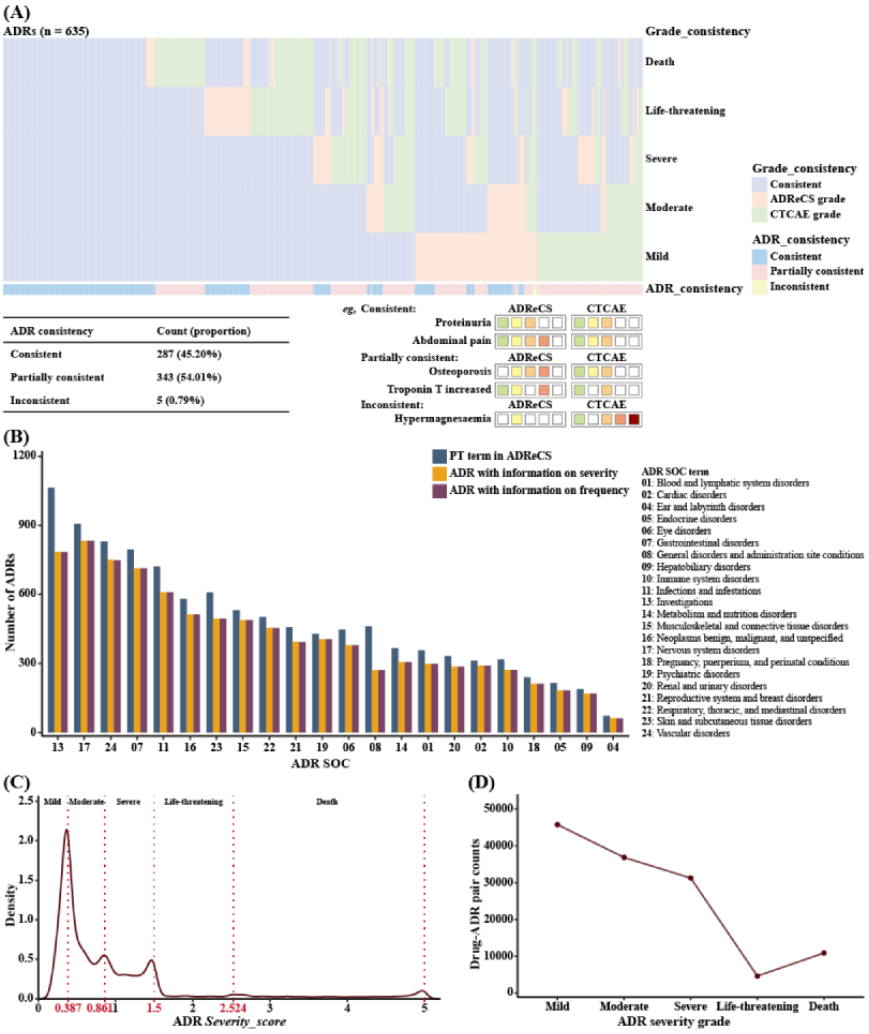

Estimation of ADR Frequency by Cross-Mining the Big Data of Historical Medical Reports

Overview

Theoretically, ADR frequency can be determined by dividing the number of reported ADRs by the number of drug prescriptions. However, the FAERS, like many other spontaneous reporting systems, faces challenges related to underreporting and bias in reporting. Typically, serious events are more likely to be reported than nonserious ones [22]. Previous studies have estimated that the average underreporting rate (URR) in FAERS is around 94%, implying that the reporting rate is only 6%. For particularly severe events, the URR drops to 77% [22,23].

Determination of the Average Annual Prescription (AAP)

To address the potential bias resulting from underreporting, we acquired real-world drug prescription data from the Medical Expenditure Panel Survey (MEPS) [24]. The MEPS is a publicly available repository that gathers information on health services and expenditures in the United States through surveys conducted among households and individuals. We downloaded the SAS/XLSX files containing household-reported prescription medicines from the MEPS, which amounted to 2,545,184 records and covered the period from 2012 to 2019. This timeframe closely aligns with the period of FAERS reports analyzed in this study. The MEPS records were preprocessed to consolidate the drug name with the FAERS, intermediated by the DrugBank ID, through three routes (Figure 3A): for each MEPS record, (1) the standard generic name, brand name, and active ingredient name of the drug were retrieved from the US Food and Drug Administration National Drug Code Directory via the drug National Drug Code record (RXNDC). Subsequently, the standard drug name was mapped to DrugBank to obtain the unique DrugBank ID; (2) in cases where the RXNDC retrieval failed, the drug name (RXNAME and RXDRGNAM) in the MEPS record underwent an exact keyword search in DrugBank for direct consolidation; and (3) for the remaining reports that failed in the previous 2 routes, the RXNAME or RXDRGNAM names underwent cleaning by removing extraneous words such as “mg,” “ophthalmic,” “tablets,” “chewable,” etc. The clean drug names were then

XSL•FO
RenderX

mapped to DrugBank again. The correspondence of the drug name to the DrugBank ID was manually checked for validation.

For drugs with confirmed DrugBank IDs, the annual prescriptions were calculated based on the respondent person weights in the MEPS records. Subsequently, the *AAPs* were determined for the period spanning 2012 to 2019. In instances where drugs were not listed in the MEPS, their *AAPs* were estimated by inferring from drugs in the same ATC class (at the 2nd level, which corresponds to the therapeutic subgroup; Figure 3B).

Figure 3. Evaluation of ADR frequency by mining historical medical records data. (A) Schematic illustration of obtaining the AAPs from the MEPS database. (B) Estimation of AAPs for drugs not mapped in the MEPS through ATC code inference. (C) Density distribution of drug-ADR pairs by ADR frequency. (D) Distribution of frequency for common and very common ADRs with ADR severity. AAP: average annual prescription; ADR: adverse drug reaction; ATC: Anatomical Therapeutic Chemical; FDA: Food and Drug Administration; MEPS: Medical Expenditure Panel Survey; NDC: National Drug Code.

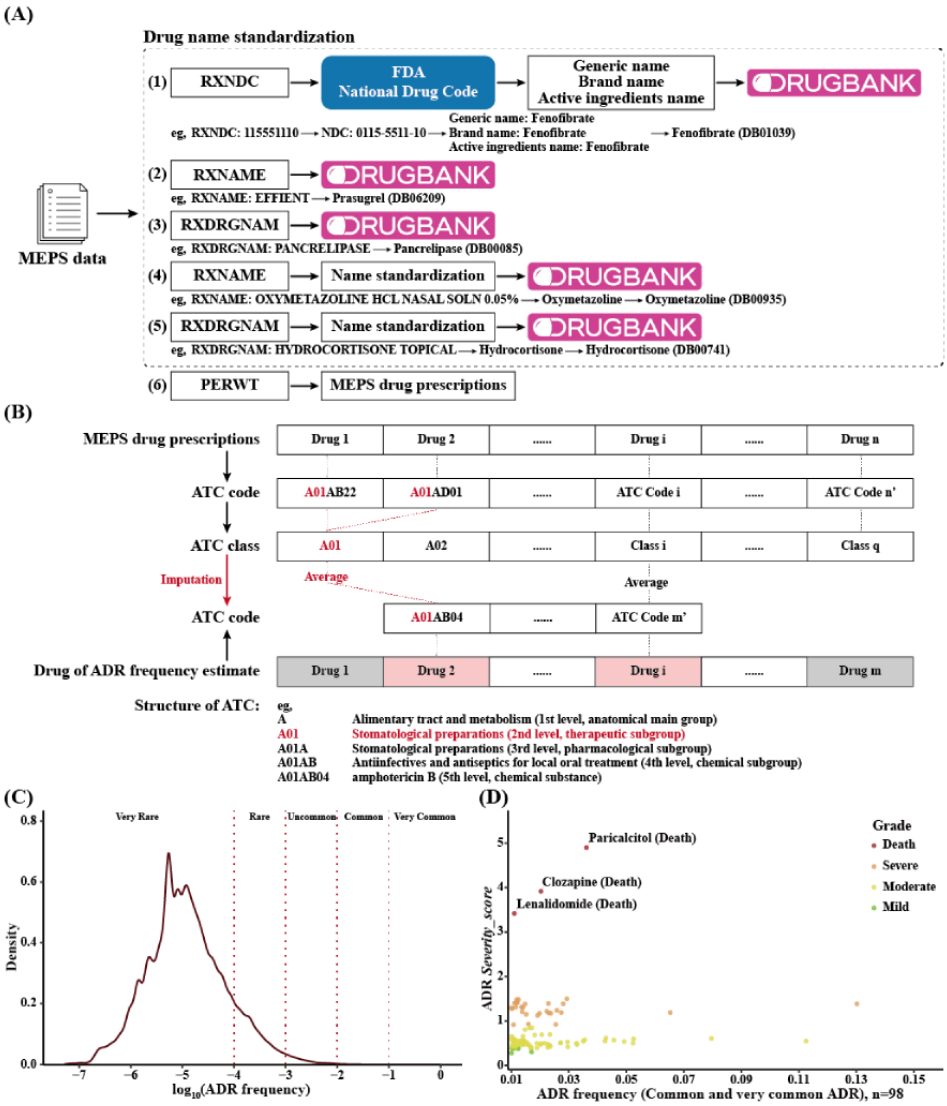

Estimation of ADR Frequency

For an ADR A_j ($A_j \in A$) induced by a drug D_i ($D_i \in D$), the frequency $Freq_{Di-Aj}$ can be calculated as follows:

The $count(D_i - A_j)$ represents the number of times that the D_i-A_j pair was reported in the ADE collection M of FAERS. "*yr*" stands for the period of this estimation, which was 8 years (corresponding to quarter 1, 2012, to quarter 1, 2020, of FEARS). The *URR* was estimated to be 77% for very serious ADRs (life-threatening and death) and 94% for other ADRs (mild, moderate, and severe).

Association Analysis Between the Targets and the ADRs

The association analysis was conducted on the 129,407 drug-ADR pairs with quantitative features. The information on therapeutic targets was obtained from the DrugBank, and the drug-ADR pairs were associated with therapeutic targets to generate a list of distinct drug-ADR–target entries. Singleton drug-ADR–target entries, where the ADR was associated with only 1 target or vice versa, were excluded from the analysis. The strength of association between the ADR/ADR group (fatal or nonfatal) (A_j) and the target (T_p) was assessed using the odds ratio (*OR*) through the construction of a two-by-two contingency table:

XSL•FO
RenderX

Where a represents the number of drug-ADR–target entries that involve both A_j and T_p, b represents the number of drug-target–ADR entries that involve T_p but not A_j, c represents the number of drug-target–ADR entries that involve A_j but not T_p, and d represents the number of drug-target–ADR entries that involve neither T_p nor A_j. A 95% CI was used for statistical analysis, and a P value of .05 was determined for significance.

Ethical Considerations

The FAERS and MEPS databases are freely available to the public, and patient information is anonymous and deidentified. Therefore, this study does not require ethical review and informed consent.

Results

Performance Evaluation of the ADReCS Severity-Grading Model

After conducting data preprocessing and normalization, we obtained 1,058,727 qualified ADE reports from FAERS. From these reports, we extracted 129,407 reliable and distinct drug-ADR relations, encompassing 774 drugs and 6277 standard ADR terms (PTs). Using these reports, we developed the ADReCS severity-grading model and introduced a new parameter, *Severity_score*, to quantify the severity of ADRs. We also determined the boundaries of the *Severity_score* parameter to classify the ADRs into 5 severity grades (Table 1 and Figure 2C). To evaluate the reliability of the ADReCS severity-grading system, we compared it with the expert system CTCAE (version 5.0). Out of a total of 635 mutual ADR PTs between the 2 systems, 287 (45.20%) terms were assigned identical severity grades, 343 (54.02%) were partially consistent, and only 5 (0.79%) had completely different grade assignments (Figure 2A). This result indicates the reliability of the ADReCS severity-grading system. Furthermore, we conducted additional comparisons of the ADReCS severity-grading system with other severity-grading systems or related works (Table 2). The ADReCS severity-grading system outperformed others in almost all aspects of data size and data width.

After applying the ADReCS severity-grading system, severity grades were determined for 6277 standard ADRs, involving a total of 129,407 drug-ADR pairs and 774 single-active ingredient drugs. The distribution of severity-graded ADRs was summarized by SOCs, as shown in Figure 2B, using all ADR standard terms in ADReCS as the background. Additionally, we counted the drug-ADR pairs based on ADR *Severity_scores*. The majority of the ADR *Severity_scores* fell within the range of 0 to 1.5, corresponding to the intervals for mild, moderate, and severe ADRs (Figures 2C and 2D). This outcome offers supporting evidence for the general consensus that the safety of most marketed drugs has undergone meticulous evaluation.

Table 2. Comparison of the ADReCS[a] severity-grading system with previous works.

	ADReCS severity-grading system	CTCAE[b] version 5.0	Ferreira et al [13]	Gottlieb et al [11]	Hartwig et al's study [10]
Number of ADRs[c]	6277	837	1149	2929	367
Method	Quantitative (model)	Qualitative (expert)	Qualitative (expert)	Qualitative (expert)	Qualitative (expert)
Architecture of grades	5 grades	5 grades	4 grades	Rank without grading	7 grades
Graded by drug-ADR pair	129,407	—[d]	—	—	—
Corresponding drugs	774	—	—	—	—
Latest update	March 2023	November 2017	December 2020	March 2015	September 1992

[a]ADReCS: Adverse Drug Reaction Classification System.

[b]CTCAE: Common Terminology Criteria for Adverse Events

[c]ADR: adverse drug reaction.

[d]—: not available.

Estimation of ADR Frequency in Large Patient Populations

Mining the MEPS data, we extracted 2,064,016 qualified prescription records of 743 drugs. Based on these records, we calculated the *AAP*s for 774 drugs in the FAERS. For 438 drugs, *AAP*s were directly computed from the MEPS data, while for 315 drugs, *AAP*s were estimated via ATC inference. However, for 21 drugs, the *AAP*s could not be estimated due to the absence of ATC classification for 17 drugs and failed ATC mapping for 4 drugs. Consequently, we obtained the frequency of 6272 ADRs, covering 127,763 drug-ADR pairs and 753 drugs.

Following conventional rules [25], we classified the 127,763 drug-ADR pairs into 5 groups based on their estimated frequency: 87.87% were classified as very rare (frequency<0.0001), 10.78% were classified as rare (0.0001<frequency<0.001), 1.27% were classified as uncommon (0.001<frequency<0.01), 0.08% were classified as common (0.01<frequency<0.1), and 0.002% were classified as very common (0.1<frequency; Figure 3C). This distribution is consistent with previous estimates indicating that most ADRs are low-probability events [5,21]. More specifically, among the 127,763 drug-ADR pairs, only 75 distinct ADRs were categorized as common and very common, and they were involved in 98 drug-ADR pairs. The majority of the very

XSL•FO
RenderX

common ADRs were mild, such as diarrhea, fatigue, headache, nausea, rash, weight gain, cough, and dizziness. It is noteworthy that 3 drugs, namely paricalcitol, clozapine, and lenalidomide, exhibited a high incidence of death (Figure 3D). Paricalcitol is a vitamin D receptor activator used to treat secondary hyperparathyroidism. Studies have reported that treatment with paricalcitol can induce life-threatening ADRs, including hypercalcemia, hyperphosphatemia, and cardiovascular diseases [26-28]. Clozapine is an antipsychotic medication used to treat treatment-resistant schizophrenia. However, it is known to cause potentially life-threatening side effects such as arrhythmias, agranulocytosis, myocarditis, seizures, and nonsuicidal death in the patient population [29,30]. Lenalidomide is an immunomodulatory and antitumor agent used to treat multiple myeloma. According to reports, lenalidomide has been associated with rare instances of severe acute liver injury or acute liver failure, which can lead to fatal outcomes [31]. In summary, the estimated ADR frequency is reasonable.

Store and Distribution of the Quantitative Features

The ADR data obtained from this study have been integrated into the ADReCS. This effort added 2831 novel ADR terms and 33,271 synonyms to the ADReCS, resulting in a total of 9375 distinct standard ADR terms and 68,067 synonyms. Meanwhile, the number of single-active ingredient drugs increased significantly by approximately 86.4%, from 1355 in the previous ADReCS version to 2526. Additionally, the number of nonredundant drug-ADR relations increased by about 6-fold, from 134,022 to 809,346. More importantly, the quantitative features of ADR severity and frequency were also incorporated. Of all 7570 ADR PTs in the updated ADReCS, approximately 82.92% (6277 ADRs) were assigned a severity grade and 82.85% (6272 ADRs) were estimated for frequency (Figure 4). This enhancement makes ADReCS the most information-rich database of drug-ADR interactions, providing unique quantitative data for multi-scale drug safety assessment and drug discovery purposes.

The quantitative parameters of ADRs can be obtained using the BROWSE or the keyword search function of ADReCS [32] (Figure 4). The complete data set of drug-ADR relations with quantitative features is available for download from the DOWNLOAD page of ADReCS [33].

Figure 4. Statistics and retrieval of the quantitative features of ADRs from the ADReCS. ADR: adverse drug reaction; ADReCS: Adverse Drug Reaction Classification System; HLT: High-Level Term; HLGT: High-Level Group Term; PT: Preferred Term; SOC: System Organ Class.

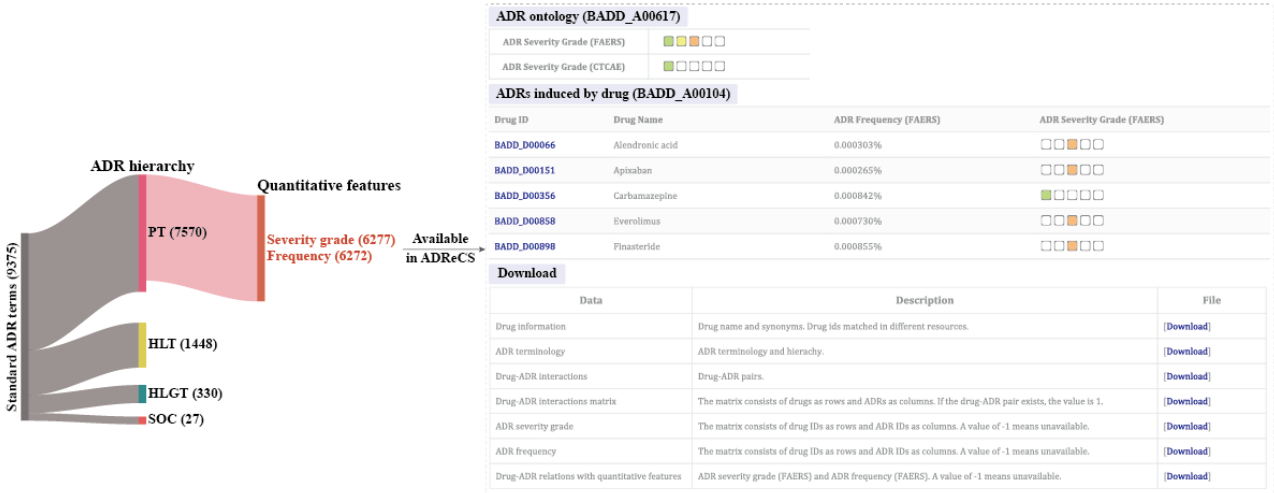

Potential Applications

Elucidation of ADR Mechanism

The quantitative features can aid in revealing potential mechanisms underlying ADRs. For instance, we conducted an association analysis between the ADRs and the therapeutic targets based on 549,670 distinct drug-target–ADR relations, involving 689 drugs, 6050 ADRs with quantitative features, and 1082 therapeutic targets. All ADR-target relations were roughly categorized into 4 zones based on the association strength (the *OR*) and the frequency of ADRs (Figure 5A). These 4 zones pretty much elucidated 4 types of ADR mechanisms: zone 1 and zone 2 included the majority (about 89.59%) of ADR-target relations, in which the ADRs were most likely induced in an on-target way or via the overdose mechanism. On-target ADRs are predictable by assessing the pharmacological activity of drugs [34]. Usually, a well-designed dosage can prevent the high occurrence of ADRs in large patient populations (zone 1). For example, cinacalcet is a calcium-sensing receptor agonist used to treat secondary hyperparathyroidism; occasionally, cinacalcet can cause hypocalcemia in a dose-dependent manner [35]. In contrast, the improper dosage regimen for normal drug therapy was prone to cause dose-dependent ADRs (zone 2). For instance, zoledronic acid, which was designed to target the farnesyl pyrophosphate synthase for the treatment of osteolytic bone disorders, is often accompanied by osteonecrosis of the jaws [36,37]. In this regard, there is still some space for optimizing the therapy dosage of zoledronic acid to gain the balance between efficacy and toxicity. A selected list of drugs with “improper” dosages is provided in Multimedia Appendix 1. This list will also suggest potential ADR mechanisms by providing the drug-target–ADR associations. Zones 3 and 4 accounted for approximately 10.41% of total ADR-target relations. The ADRs in these 2 zones were likely induced by off-target effects, with unclear underlying mechanisms. Thus,

XSL•FO
RenderX

these ADRs were usually unpredictable based on pharmacological principles, and some of them could be direct immune-mediated ADRs [34,38]. For example, clozapine-induced neutropenia has been found to be associated with the carriage of specific human leukocyte antigen risk alleles [34,39] rather than overacting on the anticipated therapeutic target of clozapine, histamine H1 receptor, used to treat psychotic diseases. Here, we also listed the drug-ADR pairs within zone 4 in Multimedia Appendix 2.

Figure 5. Potential applications of quantitative features in mechanistic understanding of ADRs. (A) Brief elucidation of ADR mechanisms via analyzing the relationship between ADR-target association strength and ADR frequency. (B) The ATC classification of 94 drugs susceptible to fatal ADRs. (C) The protein-protein interaction network of 104 risky therapeutic targets, constructed with Cytoscape (Cytoscape Consortium). (D) The KEGG pathway enrichment analysis of the risk targets. ADR: adverse drug reaction; ATC: Anatomical Therapeutic Chemical; CASR: calcium-sensing receptor; FDPS: farnesyl pyrophosphate synthase; KEGG: Kyoto Encyclopedia of Genes and Genomes; OR: odds ratio.

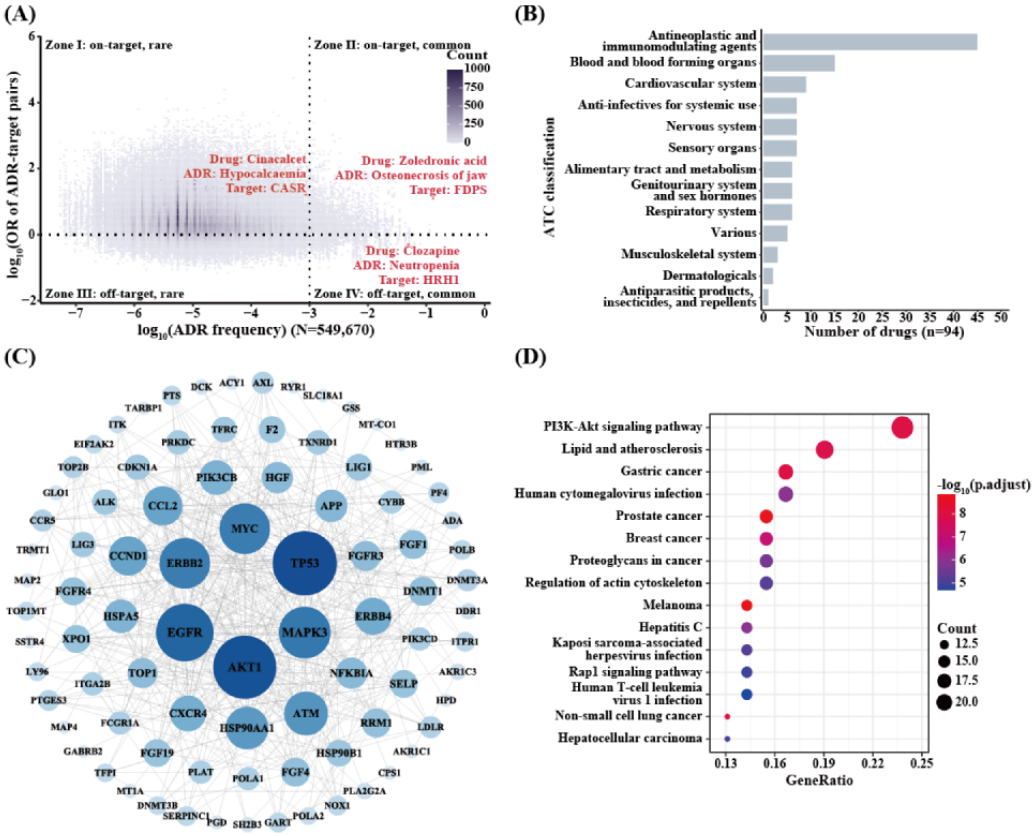

Seeking the Risky Factors of Fatal ADRs

Fatal ADRs are of paramount concern in new drug discovery. Identifying the risk factors of fatal ADRs could substantially enhance the success rate of designing "high efficacy and low toxicity" drugs. In this study, we conducted an association analysis to identify the risk factors of fatal ADRs (ADRs graded as life-threatening or death). As a result, we found 104 targets for 94 drugs that were susceptible to fatal ADRs (*OR*>2). Many of these targets were prioritized to combat complex diseases such as cancers, immunological diseases, blood diseases, and cardiovascular diseases (Figure 5B). The protein-protein interaction network analysis revealed that many of these "toxic" targets were interconnected in a highly dense subnetwork, which centered around several well-known targets such as epidermal growth factor receptor, AKT1 (RAC-alpha serine/threonine-protein kinase), and TP53 (cellular tumor antigen p53; Figure 5C). The functional analysis further specified that these targets were enriched in the PI3K-Akt signaling pathway, one of the major cell signaling pathways involved in regulating various cellular processes such as cell proliferation, growth, cell size, metabolism, and motility (Figure 5D). Actually, some PI3K/Akt/mTOR inhibitors have been reported to often induce severe ADRs such as cardiac toxicity, liver toxicity, immunosuppression, and pneumonia [40,41]. In this regard, the discovery of novel targets or drugs for safe cancer therapy remains a significant challenge. Similar approaches can be applied to discover the potential risk factors for a definite severe ADR.

Discussion

Principal Findings

ADR is more than a binary issue of occurring or not occurring; instead, it is a multidimensional concept. For instance, a drug may not always trigger a definite ADR in all cases of drug therapy. The occurrence of ADR is influenced by multiple factors such as drug dosage, treatment course, individual genetic variation, physiological and pathological states of patients, and so on. Moreover, different ADRs, or even the same ADR induced in the treatment of different diseases, may exhibit varying impacts on patient health. In drug discovery, common ADRs such as nausea and itching are often considered tolerable as they are not severe. However, when it comes to combating life-threatening diseases for which there are no available drugs,

XSL•FO
RenderX

drug candidates that may cause severe but rare ADRs may still have a chance to enter the market. Therefore, information on ADR severity and frequency is essential for fair characterization of drug toxicity in humans, and a simple counting of ADRs would be inadequate and biased for precise drug safety assessment. To break the qualitative constraints, we take the first step in measuring ADR severity quantitatively by learning from the big data of historical ADE reports in this study. Furthermore, we also estimate ADR frequency in large patient populations by cross-mining real-world prescription records. These attempts could be a significant leap for the community of clinical pharmacology and toxicology, surpassing the binary dimension of current ADR research and expanding it to a multidimensional space. Meanwhile, these multidimensional features can enrich the vectorized representation of ADRs, providing machine learning applications with richer input information on ADRs.

Limitations

This work has several limitations. The ADReCS severity-grading system is based on the assumption that the ADEs have been fully and unbiasedly reported to the FAERS. However, in real-world clinical practices, clinicians tend to report severe ADRs rather than mild ADRs. As a consequence, the severity of serious ADRs could be overestimated. For this, the incorporation of more ADE sources such as EudraVigilance for severity grading and frequency estimation will partially rectify the reporting bias. Moreover, optimization of the ADReCS severity-grading model or deployment of new algorithms is also desirable to improve the quantitative characterization of ADRs.

Conclusions

In summary, quantitative estimation of ADR severity and frequency enriches current knowledge of the clinical phenotypes caused by drug toxicity in both depth and width. It also addresses data gaps by providing high-quality data sets of drug-ADR relations for multiscale drug safety assessment and drug discovery using advanced artificial intelligence algorithms. Last but not least, it prompts current drug safety research to shift from qualitative description to quantitative analysis.

Acknowledgments

This study was funded by the National Key Research and Development Program of China (2018YFC1003601) and the National Natural Science Foundation of China (30873159).

Data Availability

The complete data set with quantitative features can be obtained [42].

Authors' Contributions

QXY, RFD, WHC, LYW, and KL conducted data collection, preprocessing, and analysis. QXY and RFD developed the algorithms. QXY and WHC prepared the figures and tables. QXY and ZLJ wrote this paper. ZLJ designed and supervised this whole study. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A selected list of drugs with “improper” dosages.
[XLSX File (Microsoft Excel File), 9 KB - jmir_v26i1e48572_app1.xlsx]

Multimedia Appendix 2
The drug-adverse drug reactions pairs within zone IV.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v26i1e48572_app2.xlsx]

References

1. de Vries ST, van der Sar MJM, Coleman AM, Escudero Y, Pascual AR, Martínez M, et al. Safety communication tools and healthcare professionals' awareness of specific drug safety issues in Europe: a survey study. Drug Saf 2018;41(7):713-724 [FREE Full text] [doi: 10.1007/s40264-018-0643-5] [Medline: 29500798]
2. Zhang JD, Sach-Peltason L, Kramer C, Wang K, Ebeling M. Multiscale modelling of drug mechanism and safety. Drug Discov Today 2020;25(3):519-534 [FREE Full text] [doi: 10.1016/j.drudis.2019.12.009] [Medline: 31899257]
3. Pognan F, Steger-Hartmann T, Díaz C, Blomberg N, Bringezu F, Briggs K, et al. The eTRANSAFE project on translational safety assessment through integrative knowledge management: achievements and perspectives. Pharmaceuticals (Basel) 2021;14(3):237 [FREE Full text] [doi: 10.3390/ph14030237] [Medline: 33800393]

XSL•FO
RenderX

4. Minerali E, Foil DH, Zorn KM, Lane TR, Ekins S. Comparing machine learning algorithms for predicting Drug-Induced Liver Injury (DILI). Mol Pharm 2020;17(7):2628-2637 [FREE Full text] [doi: 10.1021/acs.molpharmaceut.0c00326] [Medline: 32422053]
5. Xiang YP, Liu K, Cheng XY, Cheng C, Gong F, Pan JB, et al. Rapid assessment of adverse drug reactions by statistical solution of gene association network. IEEE/ACM Trans Comput Biol Bioinform 2015;12(4):844-850. [doi: 10.1109/TCBB.2014.2338292] [Medline: 26357325]
6. Santín EP, Solana RR, García MG, Suárez MDMG, Díaz GDB, Cabal MDC, et al. Toxicity prediction based on artificial intelligence: a multidisciplinary overview. WIREs Comput Mol Sci 2021;11(5):e1516. [doi: 10.1002/wcms.1516]
7. Kuang Q, Wang M, Li R, Dong Y, Li Y, Li M. A systematic investigation of computation models for predicting Adverse Drug Reactions (ADRs). PLoS One 2014;9(9):e105889 [FREE Full text] [doi: 10.1371/journal.pone.0105889] [Medline: 25180585]
8. Mantripragada AS, Teja SP, Katasani RR, Joshi P, Ramesh R, Masilamani V. Prediction of adverse drug reactions using drug convolutional neural networks. J Bioinform Comput Biol 2021;19(1):2050046. [doi: 10.1142/S0219720020500468] [Medline: 33472571]
9. Lavertu A, Hamamsy T, Altman RB. Quantifying the severity of adverse drug reactions using social media: network analysis. J Med Internet Res 2021;23(10):e27714 [FREE Full text] [doi: 10.2196/27714] [Medline: 34673524]
10. Hartwig SC, Siegel J, Schneider PJ. Preventability and severity assessment in reporting adverse drug reactions. Am J Hosp Pharm 1992;49(9):2229-2232. [Medline: 1524068]
11. Gottlieb A, Hoehndorf R, Dumontier M, Altman RB. Ranking adverse drug reactions with crowdsourcing. J Med Internet Res 2015;17(3):e80 [FREE Full text] [doi: 10.2196/jmir.3962] [Medline: 25800813]
12. Asseray N, Ballereau F, Trombert-Paviot B, Bouget J, Foucher N, Renaud B, et al. Frequency and severity of adverse drug reactions due to self-medication: a cross-sectional multicentre survey in emergency departments. Drug Saf 2013;36(12):1159-1168 [FREE Full text] [doi: 10.1007/s40264-013-0114-y] [Medline: 24163273]
13. Ferreira TR, Lopes LC, de Càssia Bergamaschi C. Frequency and severity of adverse drug reactions to medications prescribed for Alzheimer's disease in a Brazilian city: cross-sectional study. Front Pharmacol 2020;11:538095 [FREE Full text] [doi: 10.3389/fphar.2020.538095] [Medline: 33519428]
14. Trotti A, Colevas AD, Setser A, Basch E. Patient-reported outcomes and the evolution of adverse event reporting in oncology. J Clin Oncol 2007;25(32):5121-5127. [doi: 10.1200/JCO.2007.12.4784] [Medline: 17991931]
15. openFDA. URL: https://open.fda.gov/data/downloads [accessed 2020-12-10]
16. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
17. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46(D1):D1074-D1082 [FREE Full text] [doi: 10.1093/nar/gkx1037] [Medline: 29126136]
18. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021;49(D1):D1388-D1395 [FREE Full text] [doi: 10.1093/nar/gkaa971] [Medline: 33151290]
19. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353-D361 [FREE Full text] [doi: 10.1093/nar/gkw1092] [Medline: 27899662]
20. ATC. Norwegian Institute of Public Health. URL: https://www.whocc.no/atc [accessed 2022-06-05]
21. Cai MC, Xu Q, Pan YJ, Pan W, Ji N, Li YB, et al. ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. Nucleic Acids Res 2015;43(Database issue):D907-D913 [FREE Full text] [doi: 10.1093/nar/gku1066] [Medline: 25361966]
22. Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf 2017;16(7):761-767. [doi: 10.1080/14740338.2017.1323867] [Medline: 28447485]
23. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions : a systematic review. Drug Saf 2006;29(5):385-396. [doi: 10.2165/00002018-200629050-00003] [Medline: 16689555]
24. Medical Expenditure Panel Survey (MEPS). Agency for Healthcare Research and Quality. URL: https://meps.ahrq.gov/mepsweb [accessed 2022-04-08]
25. Eriksson R, Aagaard L, Jensen LJ, Borisova L, Hørlück D, Brunak S, et al. Discrepancies in listed adverse drug reactions in pharmaceutical product information supplied by the regulatory authorities in Denmark and the USA. Pharmacol Res Perspect 2014;2(3):e00038 [FREE Full text] [doi: 10.1002/prp2.38] [Medline: 25505588]
26. Hansen D, Rasmussen K, Danielsen H, Meyer-Hofmann H, Bacevicius E, Lauridsen TG, et al. No difference between alfacalcidol and paricalcitol in the treatment of secondary hyperparathyroidism in hemodialysis patients: a randomized crossover trial. Kidney Int 2011;80(8):841-850 [FREE Full text] [doi: 10.1038/ki.2011.226] [Medline: 21832979]
27. Liu Y, Liu LY, Jia Y, Wu MY, Sun YY, Ma FZ. Efficacy and safety of paricalcitol in patients undergoing hemodialysis: a meta-analysis. Drug Des Devel Ther 2019;13:999-1009 [FREE Full text] [doi: 10.2147/DDDT.S176257] [Medline: 30992658]

XSL•FO
RenderX

28. Geng X, Shi E, Wang S, Song Y. A comparative analysis of the efficacy and safety of paricalcitol versus other vitamin D receptor activators in patients undergoing hemodialysis: a systematic review and meta-analysis of 15 randomized controlled trials. PLoS One 2020;15(5):e0233705 [FREE Full text] [doi: 10.1371/journal.pone.0233705] [Medline: 32470067]
29. Nielsen J, Correll CU, Manu P, Kane JM. Termination of clozapine treatment due to medical reasons: when is it warranted and how can it be avoided? J Clin Psychiatry 2013;74(6):603-613 [FREE Full text] [doi: 10.4088/JCP.12r08064] [Medline: 23842012]
30. Rose E, Chen S, Turrion C, Jenkins C, Cardinal RN, Fernandez-Egea E. Causes of death in clozapine-treated patients in a catchment area: a 10-year retrospective case-control study. Eur Neuropsychopharmacol 2020;36:160-166. [doi: 10.1016/j.euroneuro.2020.05.011] [Medline: 32546415]
31. Hussain S, Browne R, Chen J, Parekh S. Lenalidomide-induced severe hepatotoxicity. Blood 2007;110(10):3814 [FREE Full text] [doi: 10.1182/blood-2007-06-097758] [Medline: 17984315]
32. ADReCS. URL: http://www.bio-add.org/ADReCS [accessed 2023-03-30]
33. ADReCS. URL: http://www.bio-add.org/ADReCS/download.jsp [accessed 2024-04-19]
34. Redwood AJ, Pavlos RK, White KD, Phillips EJ. HLAs: key regulators of T-cell-mediated drug hypersensitivity. HLA 2018;91(1):3-16 [FREE Full text] [doi: 10.1111/tan.13183] [Medline: 29171940]
35. Liu YJ, Sun LN, Cheng ZP, Qian Y, Ma ZQ, Zhang XH, et al. Pharmacogenetic and safety analysis of cinacalcet hydrochloride in healthy Chinese subjects. Ann Transl Med 2020;8(21):1385 [FREE Full text] [doi: 10.21037/atm-20-1329] [Medline: 33313130]
36. Günaldi M, Afsar CU, Duman BB, Kara IO, Tatli U, Sahin B. Effect of the cumulative dose of zoledronic acid on the pathogenesis of osteonecrosis of the jaws. Oncol Lett 2015;10(1):439-442 [FREE Full text] [doi: 10.3892/ol.2015.3156] [Medline: 26171047]
37. Marini F, Tonelli P, Cavalli L, Cavalli T, Masi L, Falchetti A, et al. Pharmacogenetics of bisphosphonate-associated osteonecrosis of the jaw. Front Biosci (Elite Ed) 2011;3(1):364-370 [FREE Full text] [doi: 10.2741/e251] [Medline: 21196316]
38. White KD, Chung WH, Hung SL, Mallal S, Phillips EJ. Evolving models of the immunopathogenesis of T cell-mediated drug allergy: the role of host, pathogens, and drug response. J Allergy Clin Immunol 2015;136(2):219-234 [FREE Full text] [doi: 10.1016/j.jaci.2015.05.050] [Medline: 26254049]
39. Saito T, Ikeda M, Mushiroda T, Ozeki T, Kondo K, Shimasaki A, et al. Pharmacogenomic study of clozapine-induced agranulocytosis/granulocytopenia in a Japanese population. Biol Psychiatry 2016;80(8):636-642 [FREE Full text] [doi: 10.1016/j.biopsych.2015.12.006] [Medline: 26876947]
40. Lee JJ, Loh K, Yap YS. PI3K/Akt/mTOR inhibitors in breast cancer. Cancer Biol Med 2015;12(4):342-354 [FREE Full text] [Medline: 26779371]
41. Alzahrani AS. PI3K/Akt/mTOR inhibitors in cancer: at the bench and bedside. Semin Cancer Biol 2019;59:125-132. [doi: 10.1016/j.semcancer.2019.07.009] [Medline: 31323288]
42. ADReCS drug ADR relations quantification. URL: http://www.bio-add.org/ADReCS/download/ADReCS_Drug_ADR_relations_quantification.txt.gz [accessed 2024-04-18]

Abbreviations

AAP: average annual prescription
ADE: adverse drug event
ADR: adverse drug reaction
ADReCS: Adverse Drug Reaction Classification System
AKT1: RAC-alpha serine/threonine-protein kinase
ATC: Anatomical Therapeutic Chemical
CTCAE: Common Terminology Criteria for Adverse Events
FAERS: Food and Drug Administration Adverse Event Reporting System
MEPS: Medical Expenditure Panel Survey
OR: odds ratio
PT: preferred term
ROR: reporting odds ratio
SOC: System Organ Class
URR: underreporting rate

XSL•FO
RenderX

Edited by Z Yin; submitted 02.05.23; peer-reviewed by LA Lee, SK Suthar, I Zakrocka; comments to author 10.03.24; accepted 18.03.24; published 03.05.24.

Please cite as:
Yue QX, Ding RF, Chen WH, Wu LY, Liu K, Ji ZL
Mining Real-World Big Data to Characterize Adverse Drug Reaction Quantitatively: Mixed Methods Study
J Med Internet Res 2024;26:e48572
URL: https://www.jmir.org/2024/1/e48572
doi:10.2196/48572
PMID:38700923

©Qi-Xuan Yue, Ruo-Fan Ding, Wei-Hao Chen, Lv-Ying Wu, Ke Liu, Zhi-Liang Ji. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Human-Smartphone Interactions and Actigraphy Measurements for Circadian Rhythm Stability and Adiposity: Algorithm Development and Validation Study

Hai-Hua Chuang[1,2,3,4], MD, PhD; Chen Lin[5], PhD; Li-Ang Lee[1,4,6], MD, PhD; Hsiang-Chih Chang[7], PhD; Guan-Jie She[7], MS; Yu-Hsuan Lin[7,8,9], MD, PhD

[1]College of Medicine, Chang Gung University, Taoyuan, Taiwan
[2]Department of Family Medicine, Chang Gung Memorial Hospital, Linkou Main Branch, Taoyuan, Taiwan
[3]Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, Taiwan
[4]School of Medicine, National Tsing Hua University, Hsinchu, Taiwan
[5]Department of Biomedical Sciences and Engineering, National Central University, Taoyuan City, Taiwan
[6]Department of Otorhinolaryngology - Head and Neck Surgery, Chang Gung Memorial Hospital, Linkou Main Branch, Taoyuan, Taiwan
[7]Institute of Population Health Sciences, National Health Research Institutes, Miaoli County, Taiwan
[8]Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan
[9]Department of Psychiatry, College of Medicine, National Taiwan University, Taipei, Taiwan

Corresponding Author:
Yu-Hsuan Lin, MD, PhD
Institute of Population Health Sciences
National Health Research Institutes
35 Keyan Road
Zhunan
Miaoli County, 35053
Taiwan
Phone: 886 37 206 166 ext 36383
Email: yuhsuanlin@nhri.edu.tw

Abstract

Background: This study aimed to investigate the relationships between adiposity and circadian rhythm and compare the measurement of circadian rhythm using both actigraphy and a smartphone app that tracks human-smartphone interactions.

Objective: We hypothesized that the app-based measurement may provide more comprehensive information, including light-sensitive melatonin secretion and social rhythm, and have stronger correlations with adiposity indicators.

Methods: We enrolled a total of 78 participants (mean age 41.5, SD 9.9 years; 46/78, 59% women) from both an obesity outpatient clinic and a workplace health promotion program. All participants (n=29 with obesity, n=16 overweight, and n=33 controls) were required to wear a wrist actigraphy device and install the Rhythm app for a minimum of 4 weeks, contributing to a total of 2182 person-days of data collection. The Rhythm app estimates sleep and circadian rhythm indicators by tracking human-smartphone interactions, which correspond to actigraphy. We examined the correlations between adiposity indices and sleep and circadian rhythm indicators, including sleep time, chronotype, and regularity of circadian rhythm, while controlling for physical activity level, age, and gender.

Results: Sleep onset and wake time measurements did not differ significantly between the app and actigraphy; however, wake after sleep onset was longer (13.5, SD 19.5 minutes) with the app, resulting in a longer actigraphy-measured total sleep time (TST) of 20.2 (SD 66.7) minutes. The obesity group had a significantly longer TST with both methods. App-measured circadian rhythm indicators were significantly lower than their actigraphy-measured counterparts. The obesity group had significantly lower interdaily stability (IS) than the control group with both methods. The multivariable-adjusted model revealed a negative correlation between BMI and app-measured IS (P=.007). Body fat percentage (BF%) and visceral adipose tissue area (VAT) showed significant correlations with both app-measured IS and actigraphy-measured IS. The app-measured midpoint of sleep showed a positive correlation with both BF% and VAT. Actigraphy-measured TST exhibited a positive correlation with BMI, VAT, and BF%, while no significant correlation was found between app-measured TST and either BMI, VAT, or BF%.

XSL•FO
RenderX

Conclusions: Our findings suggest that IS is strongly correlated with various adiposity indicators. Further exploration of the role of circadian rhythm, particularly measured through human-smartphone interactions, in obesity prevention could be warranted.

(J Med Internet Res 2024;26:e50149) doi:10.2196/50149

KEYWORDS

actigraphy; body composition; circadian rhythm; human-smartphone interaction; interdaily stability; obesity

Introduction

The global obesity epidemic has become a major public health concern in recent decades, and there is growing evidence to suggest that disrupted circadian rhythms and sleep disturbances are contributing factors [1,2]. Circadian rhythms are responsible for regulating energy homeostasis and adipose tissue metabolism, and disruptions to these rhythms have been associated with an increased risk of obesity [3]. Several animal studies [4], human studies with circadian gene variants [5,6], and examinations of shift workers have established this link [7]. However, most studies linking circadian disruptions and obesity have relied on self-reported questionnaires [8], which may be prone to bias and limitations. For instance, social jet lag, which is a well-known circadian rhythm indicator related to obesity, is quantified through self-reported measures of the difference in midsleep time between workdays and free days [8]. Accurate detection of circadian disturbances in clinical settings often necessitates continuous sleep-wake recordings or serial assays of timed plasma melatonin levels [9], which are more invasive and less commonly used.

Actigraphy, commonly used in research as a primary method for measuring physical activity and sleep patterns, operates on the principle that changes in body movements detected by an accelerometer are indicative of sleep onset and wakefulness. Traditionally, this approach has been seen as a “ground truth” for assessing circadian activity rhythms, especially when more invasive endogenous measurements like melatonin byproducts or urine secretion analysis are not feasible [10]. Additionally, recent studies have explored the effectiveness of consumer sleep-tracking devices in circadian measurement. Notably, devices such as the Fatigue Science Readiband (Fatigue Science), Fitbit Alta HR (Fitbit Inc), EarlySense Live (EarlySense Ltd), ResMed S+ (ResMed Inc), and SleepScore Max (SleepScore Labs) have shown comparable or superior performance to actigraphy in certain sleep or wake measures [10], underscoring their emerging relevance in circadian rhythm research. While valuable as a proxy, capturing physical activity patterns that hint at circadian activity rhythms, actigraphy primarily reflects physical movement but not the entire spectrum of circadian biological rhythms. Moreover, the advent of modern technologies such as smartphones has challenged this traditional paradigm. Although actigraphy efficiently records physical motion, it falls short in capturing cognitive engagement that may occur with minimal physical movement, such as during smartphone use [11]. This limitation suggests a potential disconnect between physical movement and cognitive engagement, particularly in the context of contemporary technology use. Cognitive activities, especially those involving minimal physical movements like smartphone interactions, can significantly impact the sleep-wake cycle. These human-smartphone interactions not only reflect physical movements but also include social engagements and digital behaviors, thus offering a more comprehensive insight into an individual’s circadian rhythm in the modern era. They indicate social rhythms and present a broader spectrum of daily routines that contribute to our understanding of modern circadian rhythms. This approach challenges the traditional reliance on actigraphy alone for sleep tracking, which primarily measures physical activity. It underscores the need to reevaluate conventional methods and acknowledge the broader implications of cognitive and social dynamics on circadian activity rhythms.

In recent years, digital footprints, such as human-smartphone interactions, have emerged as a new way to observe human circadian rhythms [11,12]. Real-time, passively collected data from these interactions can provide long-term recordings of circadian rhythms in a natural setting, potentially offering an alternative to actigraphy. However, while some mobile apps can estimate sleep time based on human-smartphone interactions [11,12], no research has yet used these patterns to determine key circadian rhythm indicators like interdaily stability (IS) and intradaily variability (IV), which have been well established through actigraphy-based analysis [13]. Additionally, previous research [11,12] has only focused on healthy participants and has not been validated in patients with sleep disturbances or disrupted circadian rhythms, which are common comorbidities in obese individuals in clinical settings [14].

In this cross-sectional study, our primary goal was to investigate the relationship between adiposity and key indicators of sleep and circadian rhythm. We measured sleep-wake cycles using wrist-worn actigraphy, a well-established method in the field. In contrast, we also collected data on human-smartphone interaction patterns through our app, Rhythm, which automatically records smartphone use data and uses an algorithm similar to actigraphy to calculate the circadian rhythm of these interactions [13,15]. Our hypothesis was that the app-based measurements might offer insights differing from those provided by traditional actigraphy, potentially revealing unique associations with adiposity indicators.

The core objective of this study was to examine and compare the associations between various adiposity indices and sleep and circadian variables as measured by 2 different methods: traditional actigraphy and our novel app-based approach. Adiposity was indicated by BMI, visceral adipose tissue (VAT) area, and body fat percentage (BF%), while the sleep variables included total sleep time, sleep timing (chronotype), and the regularity of circadian rhythm. We aimed to assess how these relationships vary when circadian rhythms are measured through actigraphy as opposed to human-smartphone interactions. This comparative analysis was crucial for understanding the relative

XSL•FO
RenderX

effectiveness of traditional and novel methods in reflecting the complex relationships between obesity, sleep, and circadian rhythms.

Methods

Study Population

A total of 78 participants (mean age 41.5, SD 9.9 years, 46/78, 59% women) were recruited from an obesity outpatient clinic and a workplace health promotion program at 7 hospitals in northern Taiwan for the study period between September 2021 and February 2023. The participants from the obesity outpatient clinics were assessed at baseline before commencing any weight reduction treatments. This is a crucial detail, as it implies that these patients have not yet started any therapeutic interventions that could potentially interfere with their adiposity or sleep, such as medications or other treatment modalities. This baseline assessment ensured that the data collected reflected their unaltered adiposity as well as their natural sleep and circadian rhythms without the influence of any treatments. Upon giving informed consent, all participants were asked to install the Rhythm app and wear a wrist actigraphy device for a minimum of 4 weeks, resulting in a total of 2182 person-days of data collection. Only participants with Android-operated smartphones were eligible to participate, with the condition that their phones were to be exclusively used by them during the study period. While participants were instructed to wear their actigraphy devices continuously, there was no specific guidance on the placement of their smartphones.

In determining the optimal sample size for this study, we focused on 2 key objectives. The primary objective was to explore the correlations between various adiposity indices, sleep, and circadian variables. The secondary objective involved comparing app-defined circadian rhythm and sleep indicators with actigraphy-derived measures across groups of individuals categorized as obese, overweight, and healthy controls. This required a detailed analysis at the level of person days.

We referred to a recent study [16] to guide our determination of sample size. That study used a similar methodology with 66 participants, including 33 patients with insomnia and 33 healthy controls, over a minimum duration of 4 weeks. This previous study [16] generated a comprehensive data set encompassing 2097 person-days and aimed to investigate the correlations between specific indicators and self-reported depressive symptoms, as well as sleep quality. It also sought to compare app-defined circadian rhythm and sleep indicators with actigraphy-derived measures.

This study consists of a total sample size of 78 participants, divided into 3 groups: obesity (n=29), overweight (n=16), and controls (n=33). The adequacy of our sample size is supported by sample size estimations for different outcome measures: When considering BMI as the main outcome and applying an *F* test with a fixed effect, omnibus, 1-way ANOVA model, with an α error probability of .05 and a statistical power of 0.90 for 3 groups, the total required sample size is 9. Similarly, if BF% is the main outcome, the same *F* test and ANOVA model, with the same α error probability and power, indicate a total required sample size of 15 for 3 groups. Additionally, when VAT serves as the main outcome, the sample size estimation under identical conditions indicates a total required sample size of 9 for 3 groups. Therefore, our current sample size for assessing adiposity-related outcomes is considered sufficient, as the effect sizes fall within the range of 0.44 to 2.59 and our statistical power exceeds 90%. This provides ample support for the suitability of this research study.

Surpassing the participant number [16] in the study we referenced, our larger sample size granted us the necessary statistical power to delve into the intricate relationships between multiple variables within our analytical framework. Additionally, the varied composition of our participant groups significantly strengthened the validity and applicability of our results. For the investigation of clinical correlations, we considered adiposity indices as dependent variables and identified several critical independent variables, including the (1) regularity of circadian rhythm, (2) physical activity level, (3) total sleep time, (4) sleep timing (chronotype), (5) age, and (6) gender. Adhering to the “1 in 10 rule,” which serves as a guideline for the number of predictor parameters in regression analysis [15], our sample size of 78 is adequately equipped to accommodate up to 6 independent variables as required by our analytical framework. This approach is validated by the principle that a sample size exceeding 60 is necessary for such analyses, confirming the adequacy of our chosen sample size.

Sleep and Circadian Rhythm Measures: the Use of the Rhythm Mobile App and Actigraphy

Actigraphy Measurement

The participants were instructed to wear a research-grade wrist actigraphy device (MiCorTM A100, MiTAC Inc) on their nondominant wrist for a minimum of 4 weeks. The actigraphy device gathered acceleration data along 3 axes, sampled at 30 Hz, and calculated the Euclidean distance of the deviations from 0. The data was then bandpass filtered from 0.5-3 Hz, and 0 values above a predefined threshold were integrated within 2 seconds. From there, activity counts were derived by averaging the integrated segments over 1 minute [17,18]. For long-term recordings, we used a predefined threshold to identify “off-wrist” periods when the wristwatch might have been removed for activities like charging or showering. These “off-wrist” epochs were excluded from the analysis to mitigate potential impacts on rest-activity patterns caused by varying wearing habits. Any daily actigraphy data containing more than 6 hours of “off-wrist” time was discarded before further analysis.

The standard Cole-Kripke algorithm [17] was applied to the activity count data with slight modifications to determine the putative sleep and wake times. The algorithm categorizes data into rest and active states based on a weighted sum of the current minute and contiguous minutes, to minimize the impact of sudden changes in activity levels that could compromise the categorization. The algorithm was run on MATLAB software (MathWorks) and used preexisting codes.

XSL•FO
RenderX

Human-Smartphone Interaction Data Collection and Analysis

The app, Rhythm, was specifically designed to collect data on smartphone use by tracking 3 key variables: screen on or off events, notifications, and the app being used [12,19]. The data were collected in the background without interfering with the smartphone's operation or significantly affecting battery life (less than 1%). To accurately capture app use behaviors in real-life scenarios and consider instances where users might temporarily leave their smartphones during activities like work or charging, we introduced the "app-count" method. This approach involves using longer durations to represent app use behaviors. Specifically, the "app-count" was defined as the sum of minute-by-minute use counts within nonoverlapping 5-minute epochs (288 epochs per day). By using these longer durations, we aimed to encompass the overall engagement with the smartphone, accounting for both periods of active use and temporary disengagement. Consequently, the number of apps used per minute was aggregated into these 5-minute epochs that did not overlap to eliminate excessive zero-count segments. These data were then used to mimic the activity data obtained from the wrist actigraphy device, estimate the near-24-hour cycle of the circadian rhythm, including the active and inactive phases, and determine sleep time during the inactive phase based on a threshold of app-counts.

Circadian Rhythm and Sleep Indicators

The acti-counts and app-counts generated from actigraphy, and smartphone use data were processed to determine 5 sleep indicators and 2 circadian rhythm indicators (Figure 1). The daily sleep indicators included sleep onset, wake time, midpoint of sleep, wake after sleep onset (WASO), and total sleep time (TST). The data were filtered to extract approximately 16 to 24 hour-long cycles, and putative sleep was assumed to occur at the half-cycle with nadir. Sleep onset was defined as a period of 8 consecutive epochs with a zero app-count, while wake time was defined as a period of 6 consecutive epochs with a nonzero app-count.

The nonparametric method was used to calculate 2 circadian rhythm indicators, IS and IV (Figure 2). IS quantified the stability of the rhythms between days, that is, the coupling strength of the rhythms to the supposedly stable environmental factors. It could vary between 0 and 1, with higher values indicating more stable daily rhythms. IV indicated the fragmentation of the rhythms, that is, the frequency and extent of transitions between rest and activity. It could vary roughly between 0 and 2, with higher values indicating higher fragmentations [13].

The formula for IS is as follows:

in which N is the total amount of data, p is the number of data point per day; x_h is the hourly means, x_i represents the individual data points, and represents the mean of all days.

In this study, we chose the nonparametric method over cosinor analysis for several key reasons. First, cosinor analysis is typically applied to physical activity data under the assumption of relatively stable circadian rhythms. However, our research focused primarily on the association between disrupted circadian rhythms and adiposity. In this context, many of our participants likely exhibited disrupted circadian rhythms, under which even the data collected through actigraphy might not conform to the norms expected in cosinor analysis. Recent studies exploring circadian rhythm as a marker for various diseases have increasingly favored nonparametric or extended-cosine methods over traditional cosinor analysis in similar scenarios [20,21].

Furthermore, our app-based measurements transformed screen events from human-smartphone interactions into app-counts, similar to actigraphy's acti-counts. However, these app-counts fundamentally differed in nature. Unlike acti-counts, which are derived from physical activity and characterized by continuous variations, app-counts did not inherently possess such continuity. Coupled with the issue of disrupted circadian rhythms in this study population, this complication made our data less suitable for cosinor analysis. The noncomparable units of most cosinor indicators and the unique, potentially nonstandard distribution of app-counts based on human-smartphone interactions reinforced our decision to opt for a nonparametric approach. Additionally, we did not use relative amplitude due to the unvaried nature of such measurements when derived from human-smartphone interactions.

XSL•FO
RenderX

Figure 1. Algorithm for app-measured circadian rhythm and sleep indicators. (A) Timestamps of smartphone events recorded by the Rhythm app. (B) Deriving usage count from screen events. (C) Computing circadian rhythms and sleep patterns through smartphone. (D) Computing circadian rhythms with app-count analysis.

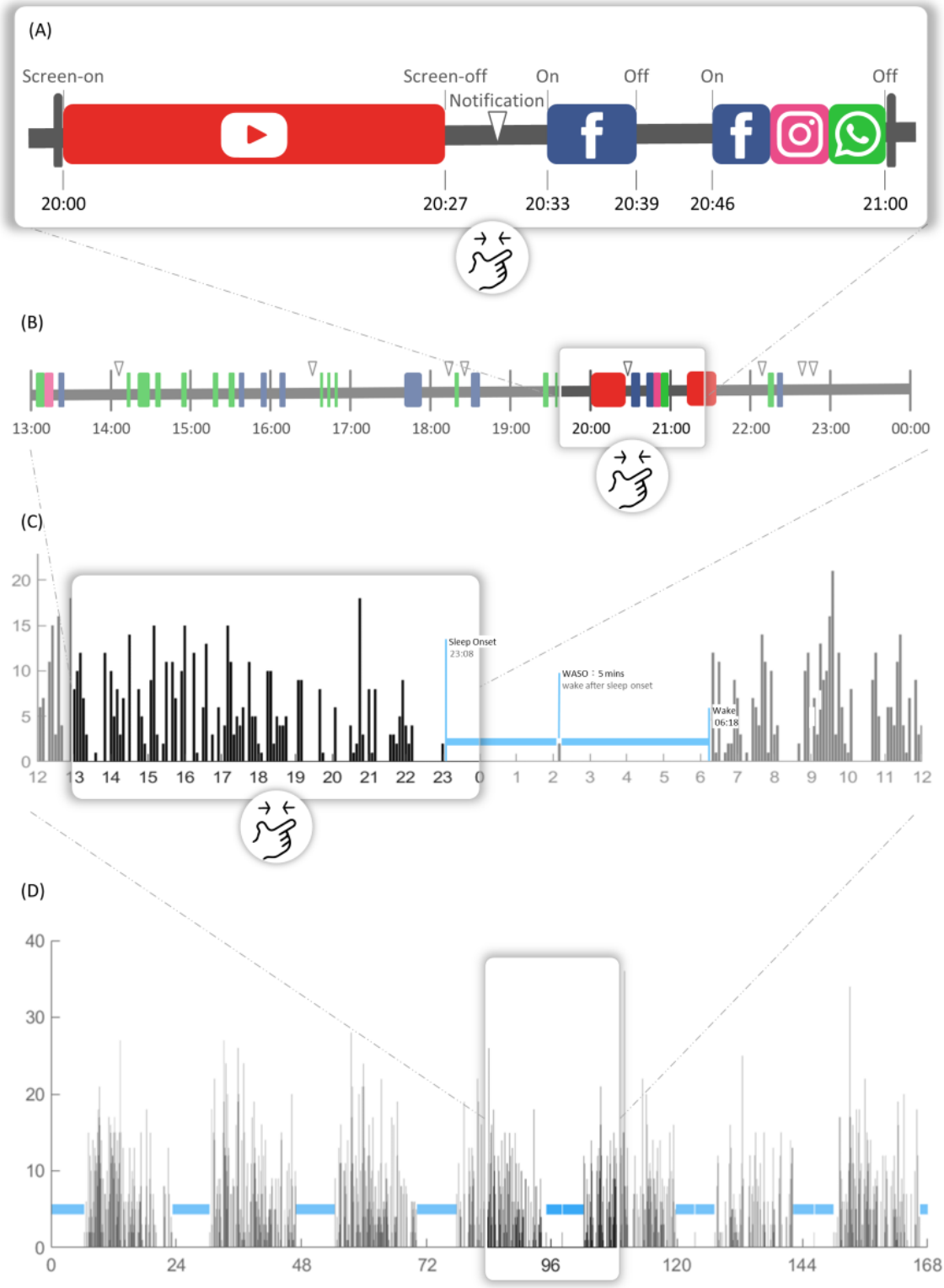

XSL•FO
RenderX

Figure 2. (A) The calculation of interdaily stability. (B) The hourly mean of the app count is calculated from the raw data and computed for each day of the recording period. (C) The mean of all hourly means across days is calculated, resulting in the grand mean.

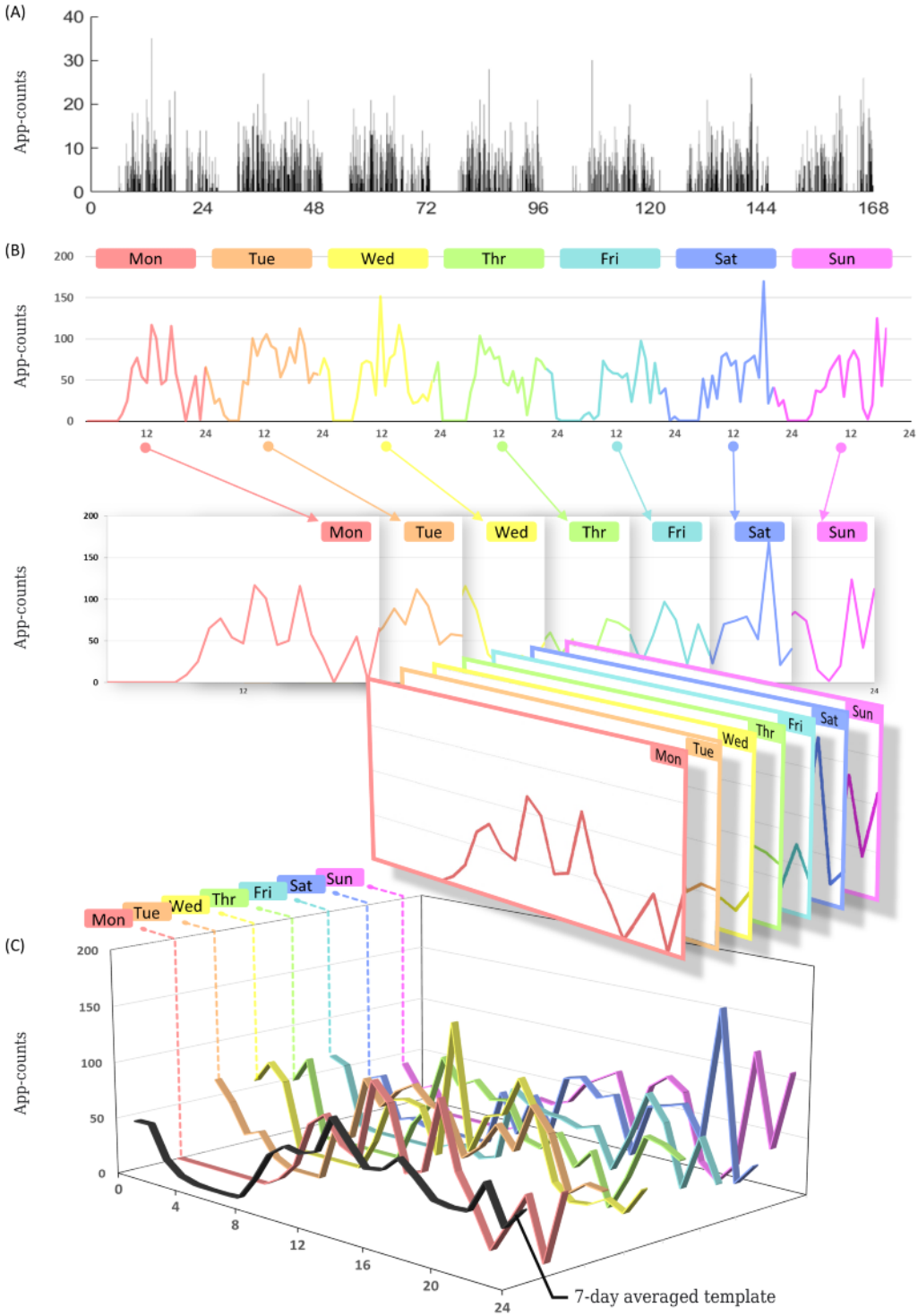

Physical Activity

Physical activity level was measured with an accelerometer to quantify the daily movements of each participant. Key features were calculated for each day, including the M10 and L5 values, representing the 10 hours of a day when the participant was most active and the 5 hours when they were least active, respectively. These are widely recognized indicators of a person's circadian activity patterns [22,23]. To determine the M10, a 10-hour moving average was used to estimate the period of the day with the highest average acceleration, which was considered to be the participant's overall physical activity level.

Pittsburgh Sleep Quality Index

The Pittsburgh Sleep Quality Index (PSQI) was used to measure the overall sleep quality of participants in one month. This index

XSL•FO
RenderX

comprises 19 items, evaluating 7 components of sleep quality, such as subjective sleep quality, sleep onset latency, total sleep duration, sleep efficiency, sleep disturbance, use of sleep medication, and daytime dysfunction. The sum of the 7 component scores yields one total score of subjective sleep quality (range 0-21); higher scores represent poorer subjective sleep quality. The cutoff score for PSQI-defined cases of poor sleep quality is a 6 or greater [24]. A Taiwanese version of the PSQI had been validated with adequate reliability [25].

Anthropometric Measurements and Weight Status

All participants had their weight (in kilograms) and height (in centimeters) measured without shoes, following standard protocols. The BMI was calculated by dividing the body weight (in kilograms) by the square of the body height (in meters), resulting in units of kg/m^2. The BMI categories used were overweight, defined as a BMI between 24 kg/m^2 and less than 27kg/m^2, and obesity, defined as a BMI equal to or greater than 27 kg/m^2 [26].

Body Composition Measurements

A total of 2 techniques were used to gather body composition measurements of the participants. For patients in the obesity outpatient clinic, dual-energy fan-beam X-ray absorptiometry (Hologic Horizon DXA system) was used with an array scan mode, following the manufacturer's protocol for body composition measurements. The Hologic APEX software (version 5.6.0.4) was used to assess the scans. Each participant was placed according to the guidelines set by the International Society for Clinical Densitometry to ensure accuracy and consistency [27]. For individuals enrolled in the workplace health promotion program, body composition was measured using bio-electrical impedance analysis, which is based on the differences in the conductivity of various components of the human body. We used the IOI-353 BC analyzer (Jawon Medical) to evaluate segmental multifrequency impedance values at 1, 5, 50, 250, 550, and 1000 kHz with a tetra-polar 8-point tactile electrode system [28]. Measurements of BF%, total body fat, total lean mass, and VAT area were obtained.

Statistical Analysis

The χ^2 test and 1-way ANOVA were used to compare demographic variables and sleep and circadian rhythm indicators among the obesity, overweight, and control groups. Fisher least significant difference was used for further post hoc tests. A 2-way ANOVA was used to compare the app-measured circadian rhythm and sleep indicators with their actigraphy counterparts. It was also used to compare these indicators among participants with obesity, overweight, and healthy controls.

The relationship between obesity, sleep, and circadian rhythm indicators was analyzed using a multivariable regression model. This approach was chosen as the evening chronotype, shorter total sleep time, and unstable circadian rhythm were found to be interconnected due to potential shared underlying factors. BMI, VAT, and BF% were used as the dependent variables, while 3 circadian rhythm and sleep indicators were used as independent variables, including (1) TST, (2) midpoint of sleep, which served as a proxy for chronotype, and (3) IS, the circadian rhythm indicator. The choice of these indicators was influenced by previous research on obesity and circadian rhythm, which has primarily focused on the concept of social jet lag. IS was specifically selected as it is more closely related to social jet lag, which is calculated from the midpoint of sleep on weekdays and weekends. Other variables included in the multivariable regression model were age, gender, and physical activity level (M10).

$P<.05$ was considered to be statistically significant. Data arrangement and statistical analysis were performed using SPSS Statistics (version 25; IBM Corp).

Ethical Considerations

The study was approved by the institutional review boards of Chang-Gung Memorial Hospital (202002452A3 and 202100434B0A3) and was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. Informed consent was obtained from each participant. Data were anonymized and stored with caution. Consent for publication will be obtained using our institutional consent form.

Results

Description of the Cohort

Table 1 presents the characteristics of the study participants. The 3 weight status groups differed significantly in BMI and body composition, including BF%, total body fat, total lean mass, and VAT area. The mean age of all participants was 41.5 (SD 9.9) years, and 59% (46/78) were women, with no significant differences in age or gender across the groups. The obesity group had significantly lower levels of physical activity on average (M10) than the overweight and healthy control groups. Both the obesity and overweight groups had average scores on the PSQI questionnaire above the cutoff point of 6, indicating poor sleep quality, whereas the healthy control group had an average PSQI score below the cutoff point.

XSL•FO
RenderX

Table 1. Demographic characteristics of participants in obesity, overweight, and control groups.

Characteristics	Obesity (n=29)	Overweight (n=16)	Control (n=33)	Chi-square (2)	*F* test (2)	*P* value	Effect Size	Comparison
Sex, n								__[a]
Male	10	7	15	0.830	—	.66	—	
Female	19	9	18	—	—	—	—	
Age (years), mean (SD)	44.4 (9.9)	40.8 (10.4)	39.4 (9.3)	—	2.104	.13	—	—
BMI (kg/m^2), mean (SD)	30.8 (2.7)	25.3 (0.8)	21.5 (1.6)	—	169.951	<.001	2.59	Ob[b]>Ow[c]>C[d]
Body fat %, mean (SD)	40.4 (6.4)	34.4(2.8)	27.6 (4.9)	—	46.850	<.001	1.16	Ob>Ow>C
Total body fat (g), mean (SD)	32,237.2 (7012.5)	22,527.3 (2023.3)	15,532.2 (3496.2)	—	88.468	<.001	2.13	Ob>Ow>C
Total lean mass (g), mean (SD)	45,804 (6636.6)	42,914.4 (4047.8)	40,324.1 (5555.2)	—	7.047	.002	0.44	Ob>C
VAT[e] area (cm^2), mean (SD)	151.3 (40.2)	84.6 (18.5)	53.9 (23)	—	83.477	<.001	1.90	Ob>Ow>C
Physical activity (M10[f]), mean (SD)	57.96 (26.58)	81.32 (17.71)	77.37 (22.55)	—	7.332	.001	0.45	Ob<Ow, C
PSQI[g] score	7.5 (4.2)	8.4 (4.6)	5.3 (2.7)	—	4.186	.02	0.31	Ow>C

[a]Not applicable.
[b]Ob: obesity group.
[c]Ow: overweight group.
[d]C: control group.
[e]VAT: visceral adipose tissue.
[f]M10: the maximum 10 hours of physical activity within a 24-hour period (derived from actigraphy data).
[g]PSQI: Pittsburgh Sleep Quality Index.

Differences Between App and Actigraphy

Table 2 presents a comparison of app-measured and actigraphy-measured circadian rhythm and sleep indicators. Results showed no significant difference between app-measured and actigraphy-measured sleep onset and wake time, as well as the midpoint of sleep. However, the app-measured WASO was found to be longer (13.5, SD 19.5 minutes) than the actigraphy-measured WASO, resulting in a longer actigraphy-measured TST of 20.2 (SD 66.7) minutes compared to the app-measured TST. The obesity group had a significantly longer TST than the overweight and control groups, regardless of the measurement method. App-measured circadian rhythm indicators, including IS and IV, were significantly lower than their actigraphy-measured counterparts. The obesity group also had significantly lower IS than the control group, whether measured by app or actigraphy.

In addition, the multivariable-adjusted model revealed a negative correlation between BMI and app-measured IS (P=.007), and a borderline significant negative correlation between BMI and actigraphy-measured IS (P=.06). BF% and VAT showed significant correlations with both app-measured IS and actigraphy-measured IS (Table 3). However, when IS was replaced with social jet lag in the model, there was no significant relationship between BMI, BF%, VAT, and either app-measured or actigraphy-measured social jet lag (Table 4). The app-measured midpoint of sleep showed a positive correlation with both BF% and VAT, whereas the actigraphy-measured midpoint did not demonstrate a significant correlation with either variable. Actigraphy-measured TST exhibited a positive correlation with BMI, VAT, and BF%, while no significant correlation was found between app-measured TST and either BMI, VAT, or BF%.

XSL•FO
RenderX

Table 2. Comparison of app-measured and actigraphy-measured circadian rhythm and sleep indicators in participants across obesity, overweight, and control groups.

Indicators	Obesity, mean (SD)	Overweight, mean (SD)	Control, mean (SD)	*P* value differences			Comparison
				Group[a]	Measurement[b]	Interaction[c]	
Sleep indicators							
Sleep onset[d]				.002	.06	.09	Ob[e], C[f]<Ow[g]
App	23.74 (1.01)	24.35 (1.14)	23.66 (0.83)				
Actigraphy	23.11 (0.73)	24 (1.08)	23.75 (0.85)				
Wake time[d]				.04	.50	.68	—[h]
App	7.37 (1.27)	7.23 (1.05)	6.91 (0.8)				
Actigraphy	7.08 (0.99)	7.34 (1.18)	6.74 (0.89)				
Midpoint of sleep[d]				.04	.17	.39	—
App	3.56 (1.09)	3.79 (1.05)	3.28 (0.69)				
Actigraphy	3.1 (0.7)	3.67 (1.05)	3.24 (0.82)				
Wake after sleep onset (minutes)				.76	<.001	.78	—
App	17.5 (27)	18.4 (16.4)	14.7 (14.7)				
Actigraphy	2.4 (2.2)	4 (4.7)	3.1 (3.1)				
Total sleep time (minutes)				<.001	.003	.03	Ob>Ow, C
App	440 (52.3)	394.3 (35.2)	420.5 (48.3)				
Actigraphy	475.8 (63)	436.5 (51.1)	416.4 (36.4)				
Circadian rhythm indicators							
Interdaily stability				.001	<.001	.95	Ob<C
App	0.17 (0.11)	0.22 (0.11)	0.25 (0.09)				
Actigraphy	0.37 (0.12)	0.42 (0.12)	0.44 (0.11)				
Intradaily variability				.54	<.001	.23	—
App	1.3 (0.3)	1.28 (0.26)	1.22 (0.23)				
Actigraphy	0.91 (0.18)	0.81 (0.22)	0.92 (0.25)				

[a]The effect of health controls versus overweight versus obesity.
[b]The effect of app versus actigraphy.
[c]The interaction of group and measurement.
[d]Time are in day decimal time, for example, 23.50=23:30 PM.
[e]Ob: obesity group.
[f]C: control group.
[g]Ow: overweight group.
[h]Not applicable.

XSL•FO
RenderX

Table 3. Results of a multivariate linear regression analysis for IS[a] on weight status and body composition indicators.

	IS	Midpoint[b]	TST[c]	M10[d]	Age	Gender	Adjusted R^2
BMI							
App							0.213
β (SE)	–0.342 (0.123)	0.188 (0.109)	–0.119 (0.119)	–0.282 (0.108)	0.196 (0.108)	0.029 (0.111)	
P value	.007	.09	.33	.01	.07	.79	
Actigraphy							0.262
β (SE)	–0.225 (0.118)	0.012 (0.107)	0.458 (0.124)	0.081 (0.135)	0.285 (0.107)	0.004 (0.096)	
P value	.06	.91	<.001	.56	.009	.97	
Body fat %							
App							0.408
β (SE)	–0.293 (0.106)	0.278 (0.094)	0.024 (0.096)	–0.252 (0.095)	0.296 (0.093)	–0.158 (0.095)	
P value	.008	.004	.82	.009	.002	>.99	
Actigraphy							0.410
β (SE)	–0.292 (0.105)	0.011 (0.099)	0.417 (0.112)	0.101 (0.121)	0.367 (0.121)	–0.206 (0.093)	
P value	.02	.91	<.001	.41	<.001	.03	
Estimate visceral adipose tissue area							
App							0.421
β (SE)	–0.305 (0.105)	0.188 (0.093)	–0.128 (0.104)	–0.364 (0.093)	0.458 (0.091)	0.087 (0.093)	
P value	.005	.05	.22	<.001	<.001	.36	
Actigraphy							0.437
β (SE)	–0.268 (0.103)	0.015 (0.098)	0.335 (0.106)	–0.039 (0.117)	0.536 (0.093)	0.062 (0.091)	
P value	.01	.88	.002	.74	<.001	.49	

[a]IS: interdaily stability.

[b]Midpoint of sleep.

[c]TST: total sleep time.

[d]M10: the maximum 10 hours of physical activity within a 24-hour period (derived from actigraphy data).

XSL•FO
RenderX

Table 4. Results of a multivariate linear regression analysis for social jet lag on weight status and body composition indicators.

	Social jet lag	Midpoint[a]	TST[b]	M10[c]	Age	Gender	Adjusted R^2
BMI							
App							0.139
β (SE)	–0.111 (0.102)	0.238 (0.113)	0.031 (0.103)	–0.272 (0.114)	0.238 (0.109)	–0.026 (0.111)	
P value	.32	.04	.78	.02	.04	.82	
Actigraphy							0.247
β (SE)	–0.152 (0.099)	0.048 (0.109)	0.452 (0.122)	–0.040 (0.131)	0.264 (0.107)	0.023 (0.106)	
P value	.14	.66	<.001	.75	.02	.83	
Body fat %							
App							0.361
β (SE)	–0.127 (0.095)	0.314 (0.097)	0.153 (0.100)	–0.238 (0.098)	0.332 (0.096)	–0.206 (0.097)	
P value	.19	.002	.12	.02	.001	.04	
Actigraphy							0.379
β (SE)	–0.181 (0.093)	0.057 (0.099)	0.411 (0.112)	–0.055 (0.110)	0.338 (0.097)	–0.185 (0.097)	
P value	.05	.57	<.001	.63	.001	.06	
Estimate visceral adipose tissue area							
App							0.370
β (SE)	–0.135 (0.094)	0.225 (0.097)	0.006 (0.102)	–0.349 (0.099)	0.497 (0.095)	0.037 (0.097)	
P value	.16	.02	.95	.001	<.001	.70	
Actigraphy							0.405
β (SE)	–0.147 (0.090)	0.057 (0.096)	0.331 (0.109)	–0.182 (0.110)	0.508 (0.095)	0.078 (0.094)	
P value	.11	.56	.003	.10	<.001	.41	

[a]Midpoint of sleep.

[b]TST: total sleep time.

[c]M10: the maximum 10 hours of physical activity within a 24-hour period (derived from actigraphy data).

Discussion

This study found that among all the circadian rhythm and sleep indicators, IS, measured by either actigraphy or human-smartphone interaction, presented with the most correlations with adiposity indicators, such as BMI, VAT, and BF%. These correlations remained significant even after adjusting for confounding factors, including TST, chronotype, physical activity level, age, and gender. These findings were consistent with previous studies that have examined the relationship between circadian rhythm, chronotype, and adiposity simultaneously, with circadian rhythm being the most significant indicator [29-31]. Some studies have measured the correlation between adiposity and circadian rhythm using IS measured through actigraphy [29,32,33], while others have relied on self-reported measures of social jet lag [8].

IS is a concept similar to social jet lag [33], but it provides a more comprehensive understanding of a person's sleep patterns compared to social jet lag (Figure 2). Unlike social jet lag, which simply calculates the difference between sleep midpoint on workdays and days off [8], IS considers the regularity of sleep and wake times within a 24-hour period for a minimum of 7 days, including both workdays and days off. IS uses the variance method to quantify the consistency of a person's sleep and wake times, while social jet lag is determined through questionnaires that reflect its limited scope. Accurately computing IS requires continuous 24-hour monitoring for at least 7 days [13], which can be done through methods such as actigraphy or mobile app use, as seen in this study. However, social jet lag can be evaluated through a questionnaire because it only requires limited information.

Although the data used to determine IS can also be used to calculate social jet lag, the reverse is not necessarily true. This study revealed that IS recorded through a mobile app had several significant correlations with the obesity index, whereas social jet lag did not, and its results were similar to those obtained through actigraphy. While the small sample size in the study may not reflect the correlation between social jet lag and obesity found in previous large-scale epidemiological studies, it highlights that when it comes to personalized and precise weight management, IS is a more precise and sensitive indicator than social jet lag.

In this study, we compared app-based measurements of circadian rhythm and sleep indicators with those obtained through actigraphy. Our findings align with previous research, showing typical results for app-based measurements [29,32-39].

XSL•FO
RenderX

However, actigraphy measurements diverged, either failing to demonstrate similar results (Table 3) or presenting inconsistent findings with established patterns in previous studies. This highlights the potential superiority of the app in capturing circadian rhythms more accurately, especially considering its consistency with established research and ability to track broader circadian dynamics.

Notably, this study revealed several key differences between the 2 methods. First, the app's internal sleep measurements showed a significant correlation with BMI, while actigraphy's IS correlation with BMI was only marginally significant (P=.06). Second, the app identified an association between the evening chronotype (measured by sleep midpoint) and both BF% and VAT, a connection not found with actigraphy. Third, TST measured through actigraphy positively correlated with BMI, VAT, and BF%, contrary to app measurements and previous studies suggesting that sufficient sleep time is generally linked with lower obesity rates [40-42]. Interestingly, this study's average sleep time was 7 hours, fitting within the substantial evidence that sleeping less than 7 hours a night can lead to adverse health consequences, including obesity [43].

Addressing concerns that the app might overestimate sleep time compared to actigraphy, our results showed no significant differences in sleep onset, wake time, or sleep midpoint between the 2 methods. The primary variance was in the WASO duration, leading to a longer TST as measured by actigraphy. However, this discrepancy had a minimal impact on circadian rhythm evaluation, as we adjusted for TST in our regression models to mitigate its influence on the association between circadian rhythm indicators and adiposity markers.

Intriguingly, while the actigraphy-measured TST was positively correlated with BMI, VAT, and BF%, the app-measured TST was not, indicating a complex relationship between sleep parameters and adiposity. Additionally, extensive research, including meta-analyses [44-46], suggests a complicated link between longer sleep duration and higher obesity risk in adults, influenced by age and other factors. Given that most sleep duration data in research is self-reported [44-46], adding another layer of complexity, our findings emphasize the importance of considering these nuances and the variability of the sleep duration-BMI relationship across different age groups and study methodologies.

The study's results reveal a significant correlation between the app-measured midpoint of sleep and both BF% and VAT, but not BMI. This correlation aligns with previous research that recognizes the role of evening chronotype and circadian disruption in excessive body fat accumulation. Our data suggest that this negative effect may be observable in the early stages of obesity development, when weight status (measured by BMI) remains unchanged, but body composition (measured by BF% and VAT) is already deteriorating. The growing body of evidence suggests that the relationship between circadian disruption and obesity is driven by alterations in the hormonal rhythmicity of melatonin, leptin, and glucocorticoids, leading to disruptions in energy homeostasis [47]. Late mealtimes, which often result from circadian misalignment [48], have also been linked to weight gain and obesity [49], with potential underlying mechanisms including decreased energy expenditure during rest and after eating and increased insulin resistance [50]. The fasting-feeding cycle plays a role in synchronizing peripheral circadian rhythms with the central clock [51], while shifting mealtimes can result in or worsen circadian disruption by uncoupling peripheral and central clocks. Recent studies suggest that circadian misalignment can cause imbalances in the gut microbiota, which plays a crucial role in energy homeostasis [52]. This misalignment, represented by lower IS in this study, can result in changes to diurnal fluctuations in the gut microbiota and lead to glucose intolerance and obesity [53]. Additionally, misaligned circadian rhythm has been linked to later mealtimes, higher consumption of calories and saturated fat [54], reduced intake of Mediterranean diet components, and ultimately, a higher BMI [55]. Compared to the actigraphy-measured midpoint of sleep, the app-measured midpoint was found to be 12.7 (SD 43.2) minutes later without statistical significance (P=.17). Despite this relatively small difference, it resulted in variations in the relationship between VAT, BF%, and the midpoint of sleep measured by app and actigraphy. This study suggested that the app-measured midpoint of sleep is more clinically sensitive in reflecting obesity indicators, as it is based on human-smartphone interactions and associated smartphone use, which affects melatonin secretion through exposure to smartphone light. Our results suggested that an app-measured chronotype, including sleep and wake times and midpoint of sleep, may provide a more representative measurement of melatonin secretion trends compared to actigraphy, which only captures physical activity. The method used in this study, which incorporates both physical activity and smartphone use, provides a measurement that is closer to melatonin secretion, considered the gold standard in measuring circadian rhythm.

This study has several limitations that should be considered when interpreting the results. First, the cross-sectional design of the study precludes the ability to make causal inferences about the relationship between rest-activity rhythm and health outcomes. Longitudinal studies are needed to further explore these associations and identify the potential mechanisms involved. Second, the methods used to measure sleep duration and circadian rhythm in this study, which are based on human-smartphone interactions, have not been validated against the gold standard measures of polysomnography and melatonin secretion, respectively. However, previous research has found that rest-activity rhythms measured by actigraphy, and mobile apps are associated with the timing of light exposure, the strongest zeitgeber of the circadian system, and the amplitude of melatonin secretion, a classic circadian phase marker [19,56]. This suggests that altered rest-activity rhythms, to some extent, reflect circadian disruption and can serve as a relevant biological marker for circadian function. Third, it should also be noted that this study did not record participants' timing and energy intake from food, as well as other well-known risk factors for obesity. Fourth, a diagnosis of obstructive sleep apnea (OSA) was found in 5 of the participants in the obesity group. Despite controlling OSA as a confounding factor, our results remained consistent. However, it is critical to acknowledge that OSA is a prevalent comorbidity of obesity and is frequently accompanied by sleep disruption and circadian rhythm misalignment [57]. Further research using larger study

XSL•FO
RenderX

populations is necessary to further clarify the role and underlying mechanisms of OSA in the relationship between obesity and circadian rhythm. Fourth, while this study demonstrated that sleep onset and wake times calculated using the Rhythm app did not significantly differ from those recorded by actigraphy, there remained a potential for misjudgment of sleep onset. This scenario could have occurred if an individual fell asleep while using their smartphone, for example, while watching Netflix. In our algorithm, we introduced the "app-count" method, where "app-count" was defined as the sum of minute-by-minute use counts within nonoverlapping 5-minute epochs. Sleep onset was determined as a period of 8 consecutive epochs with a zero app-count, whereas wake time was identified as a period of 6 consecutive epochs with a nonzero app-count. Indeed, this study found that the WASO recorded by the app was longer than that recorded by actigraphy. In future research, it may be necessary to adjust the threshold of zero app-count to achieve more accurate recordings. Finally, we faced potential interruptions in data collection from both the wrist-worn actigraphy device and the smartphone app, primarily due to issues such as low battery power or temporary nonuse by participants. To enhance the robustness and continuity of our data in future research, we plan to synergistically integrate the app with actigraphy. This integration strategy aims to mitigate the challenges associated with individual data collection methods. By using dual recording modes, we anticipate a reduction in the likelihood of concurrent data interruptions. The synergy between these 2 distinct data sources is expected to provide a more comprehensive and nuanced understanding of patient behaviors and circadian rhythms, offering a more holistic perspective in our future studies.

In conclusion, our results reveal that disruptions in the circadian rhythms are linked to higher levels of BMI, VAT, and BF%. These associations appear to hold even when taking into account factors such as sleep duration and levels of daytime motor activity. The strongest associations with obesity were seen in metrics of IS and circadian rhythm regularity, as measured by either a mobile app or actigraphy. These findings suggest that regularity in the circadian rhythm may be a promising target for obesity prevention and body weight management interventions. Future studies should further explore the mechanisms underlying these relationships and develop strategies to improve the regularity of circadian rhythms.

Acknowledgments

This research was funded by the National Health Research Institutes of Taiwan (12A1-PHSP12-014 and 13A1-PHPP14-014), the National Science and Technology Council of Taiwan (MOST 110-2314-B-182-062, 111-2314-B-182-019-MY2, 112-2218-E-008-015, and 111-2221-E-008-103-MY3), and the Chang Gung Medical Foundation of Taiwan (CMRPG1H0061-63, CORPG1L0011-13, CORPG3C0011-13, and CORPG1I0021).

Data Availability

The data sets generated or analyzed during this study are not publicly available due to restrictions imposed by informed consent from participants in this study, which has been approved by the institutional review boards of the Chang Gung Memorial Hospital, but they are available from the corresponding author on reasonable request.

Authors' Contributions

CL, HCC, and GJS performed data generation and statistical analysis. HHC, CL, LAL, and YHL drafted the manuscript, conceived and designed the study, and acquired data. All authors contributed to data analysis and interpretation and were involved in manuscript revisions and the approval of the final revision. YHL is responsible for the integrity of the work as a whole.

Conflicts of Interest

YHL is a new Associate Editor of JMIR at the time of this publication.

References

1. Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. Prevalence of healthy sleep duration among adults—United States, 2014. MMWR Morb Mortal Wkly Rep 2016;65(6):137-141 [FREE Full text] [doi: 10.15585/mmwr.mm6506a1] [Medline: 26890214]
2. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity and severe obesity among adults: United States, 2017-2018. NCHS Data Brief 2020(360):1-8 [FREE Full text] [Medline: 32487284]
3. Laermans J, Depoortere I. Chronobesity: role of the circadian system in the obesity epidemic. Obes Rev 2016;17(2):108-125. [doi: 10.1111/obr.12351] [Medline: 26693661]
4. Turek FW, Joshu C, Kohsaka A, Lin E, Ivanova G, McDearmon E, et al. Obesity and metabolic syndrome in circadian clock mutant mice. Science 2005;308(5724):1043-1045 [FREE Full text] [doi: 10.1126/science.1108750] [Medline: 15845877]
5. Bandín C, Martinez-Nicolas A, Ordovás JM, Ros Lucas JA, Castell P, Silvente T, et al. Differences in circadian rhythmicity in CLOCK 3111T/C genetic variants in moderate obese women as assessed by thermometry, actimetry and body position. Int J Obes (Lond) 2013;37(8):1044-1050 [FREE Full text] [doi: 10.1038/ijo.2012.180] [Medline: 23183326]

XSL•FO
RenderX

6. Garaulet M, Lee YC, Shen J, Parnell LD, Arnett DK, Tsai MY, et al. Genetic variants in human CLOCK associate with total energy intake and cytokine sleep factors in overweight subjects (GOLDN population). Eur J Hum Genet 2010;18(3):364-369 [FREE Full text] [doi: 10.1038/ejhg.2009.176] [Medline: 19888304]
7. Sun M, Feng W, Wang F, Li P, Li Z, Li M, et al. Meta-analysis on shift work and risks of specific obesity types. Obes Rev 2018;19(1):28-40. [doi: 10.1111/obr.12621] [Medline: 28975706]
8. Zhang R, Cai X, Lin C, Yang W, Lv F, Wu J, et al. The association between metabolic parameters and evening chronotype and social jetlag in non-shift workers: a meta-analysis. Front Endocrinol (Lausanne) 2022;13:1008820 [FREE Full text] [doi: 10.3389/fendo.2022.1008820] [Medline: 36479212]
9. Allada R, Bass J. Circadian mechanisms in medicine. N Engl J Med 2021;384(6):550-561 [FREE Full text] [doi: 10.1056/NEJMra1802337] [Medline: 33567194]
10. Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep 2021;44(5):zsaa291 [FREE Full text] [doi: 10.1093/sleep/zsaa291] [Medline: 33378539]
11. Borger JN, Huber R, Ghosh A. Capturing sleep-wake cycles by using day-to-day smartphone touchscreen interactions. NPJ Digit Med 2019;2:73 [FREE Full text] [doi: 10.1038/s41746-019-0147-4] [Medline: 31372507]
12. Lin YH, Wong BY, Pan YC, Chiu YC, Lee YH. Validation of the mobile app-recorded circadian rhythm by a digital footprint. JMIR Mhealth Uhealth 2019;7(5):e13421 [FREE Full text] [doi: 10.2196/13421] [Medline: 31099340]
13. Witting W, Kwa IH, Eikelenboom P, Mirmiran M, Swaab DF. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. Biol Psychiatry 1990;27(6):563-572. [doi: 10.1016/0006-3223(90)90523-5] [Medline: 2322616]
14. Gibson‐Moore H, Chambers L. Sleep matters: can a good night's sleep help tackle the obesity crisis? Nutr Bull 2019;44(2):123-129. [doi: 10.1111/nbu.12386]
15. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-387. [doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4] [Medline: 8668867]
16. Lin C, Chen IM, Chuang HH, Wang ZW, Lin HH, Lin YH. Examining human-smartphone interaction as a proxy for circadian rhythm in patients with insomnia: cross-sectional study. J Med Internet Res 2023;25:e48044 [FREE Full text] [doi: 10.2196/48044] [Medline: 38100195]
17. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992;15(5):461-469 [FREE Full text] [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
18. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003;26(3):342-392 [FREE Full text] [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
19. Lin YH, Wong BY, Lin SH, Chiu YC, Pan YC, Lee YH. Development of a mobile application (App) to delineate "digital chronotype" and the effects of delayed chronotype by bedtime smartphone use. J Psychiatr Res 2019;110:9-15. [doi: 10.1016/j.jpsychires.2018.12.012] [Medline: 30611008]
20. Smagula SF, Zhang G, Gujral S, Covassin N, Li J, Taylor WD, et al. Association of 24-hour activity pattern phenotypes with depression symptoms and cognitive performance in aging. JAMA Psychiatry 2022;79(10):1023-1031 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.2573] [Medline: 36044201]
21. Leng Y, Blackwell T, Cawthon PM, Ancoli-Israel S, Stone KL, Yaffe K. Association of circadian abnormalities in older adults with an increased risk of developing Parkinson disease. JAMA Neurol 2020;77(10):1270-1278 [FREE Full text] [doi: 10.1001/jamaneurol.2020.1623] [Medline: 32539075]
22. Gonçalves BSB, Adamowicz T, Louzada FM, Moreno CR, Araujo JF. A fresh look at the use of nonparametric analysis in actimetry. Sleep Med Rev 2015;20:84-91. [doi: 10.1016/j.smrv.2014.06.002] [Medline: 25065908]
23. Jones SE, van Hees VT, Mazzotti DR, Marques-Vidal P, Sabia S, van der Spek A, et al. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. Nat Commun 2019;10(1):1585 [FREE Full text] [doi: 10.1038/s41467-019-09576-1] [Medline: 30952852]
24. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
25. Tsai PS, Wang SY, Wang MY, Su CT, Yang TT, Huang CJ, et al. Psychometric evaluation of the Chinese version of the Pittsburgh Sleep Quality Index (CPSQI) in primary insomnia and control subjects. Qual Life Res 2005;14(8):1943-1952. [doi: 10.1007/s11136-005-4346-x] [Medline: 16155782]
26. Taiwan's Obesity Prevention and Management Strategy. Health Promotion Administration. 2018. URL: https://www.hpa.gov.tw/File/Attach/10299/File_11744.pdf [accessed 2024-04-09]
27. Hangartner TN, Warner S, Braillon P, Jankowski L, Shepherd J. The official positions of the international society for clinical densitometry: acquisition of dual-energy x-ray absorptiometry body composition and considerations regarding analysis and repeatability of measures. J Clin Densitom 2013;16(4):520-536. [doi: 10.1016/j.jocd.2013.08.007] [Medline: 24183641]
28. Zimny M, Starczewska M, Szkup M, Cybulska A, Grochans E. Body composition and biological functioning in polish perimenopausal women with type 2 diabetes. Int J Environ Res Public Health 2021;18(21):11422 [FREE Full text] [doi: 10.3390/ijerph182111422] [Medline: 34769939]

XSL•FO
RenderX

29. Li J, Vungarala S, Somers VK, Di J, Lopez-Jimenez F, Covassin N. Rest-activity rhythm is associated with obesity phenotypes: a cross-sectional analysis. Front Endocrinol (Lausanne) 2022;13:907360 [FREE Full text] [doi: 10.3389/fendo.2022.907360] [Medline: 35837304]
30. McMahon DM, Burch JB, Youngstedt SD, Wirth MD, Hardin JW, Hurley TG, et al. Relationships between chronotype, social jetlag, sleep, obesity and blood pressure in healthy young adults. Chronobiol Int 2019;36(4):493-509. [doi: 10.1080/07420528.2018.1563094] [Medline: 30663440]
31. Li TM, Pien LC, Cheng WJ. Direction and magnitude of social jetlag are associated with obesity: results from a sample of hospital employees. Chronobiol Int 2022;39(9):1242-1248. [doi: 10.1080/07420528.2022.2090953] [Medline: 35796193]
32. Qian J, Martinez-Lozano N, Tvarijonaviciute A, Rios R, Scheer FAJL, Garaulet M. Blunted rest-activity rhythms link to higher body mass index and inflammatory markers in children. Sleep 2021;44(5):zsaa256 [FREE Full text] [doi: 10.1093/sleep/zsaa256] [Medline: 33249510]
33. Fárková E, Schneider J, Šmotek M, Bakštein E, Herlesová J, Kopřivová J, et al. Weight loss in conservative treatment of obesity in women is associated with physical activity and circadian phenotype: a longitudinal observational study. Biopsychosoc Med 2019;13:24 [FREE Full text] [doi: 10.1186/s13030-019-0163-2] [Medline: 31673283]
34. Feliciano EMC, Quante M, Weng J, Mitchell JA, James P, Marinac CR, et al. Actigraphy-derived daily rest-activity patterns and body mass index in community-dwelling adults. Sleep 2017;40(12):zsx168 [FREE Full text] [doi: 10.1093/sleep/zsx168] [Medline: 29029250]
35. Rogers-Soeder TS, Blackwell T, Yaffe K, Ancoli-Israel S, Redline S, Cauley JA, et al. Rest-activity rhythms and cognitive decline in older men: the Osteoporotic Fractures in Men Sleep Study. J Am Geriatr Soc 2018;66(11):2136-2143 [FREE Full text] [doi: 10.1111/jgs.15555] [Medline: 30136716]
36. Luik AI, Zuurbier LA, Hofman A, Van Someren EJW, Tiemeier H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: the importance of age, lifestyle, and mental health. Chronobiol Int 2013;30(10):1223-1230. [doi: 10.3109/07420528.2013.813528] [Medline: 23971909]
37. Mitchell JA, Quante M, Godbole S, James P, Hipp JA, Marinac CR, et al. Variation in actigraphy-estimated rest-activity patterns by demographic factors. Chronobiol Int 2017;34(8):1042-1056 [FREE Full text] [doi: 10.1080/07420528.2017.1337032] [Medline: 28650674]
38. Sohail S, Yu LY, Bennett DA, Buchman AS, Lim ASP. Irregular 24-hour activity rhythms and the metabolic syndrome in older adults. Chronobiol Int 2015;32(6):802-813 [FREE Full text] [doi: 10.3109/07420528.2015.1041597] [Medline: 26061588]
39. Xiao Q, Qian J, Evans DS, Redline S, Lane NE, Ancoli-Israel S, et al. Cross-sectional and prospective associations of rest-activity rhythms with metabolic markers and type 2 diabetes in older men. Diabetes Care 2020;43(11):2702-2712 [FREE Full text] [doi: 10.2337/dc20-0557] [Medline: 32887712]
40. Zhou Q, Zhang M, Hu D. Dose-response association between sleep duration and obesity risk: a systematic review and meta-analysis of prospective cohort studies. Sleep Breath 2019;23(4):1035-1045. [doi: 10.1007/s11325-019-01824-4] [Medline: 30941582]
41. Jaiswal SJ, Quer G, Galarnyk M, Steinhubl SR, Topol EJ, Owens RL. Association of sleep duration and variability with body mass index: sleep measurements in a large US population of wearable sensor users. JAMA Intern Med 2020;180(12):1694-1696 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2834] [Medline: 32926073]
42. Patel SR, Hu FB. Short sleep duration and weight gain: a systematic review. Obesity (Silver Spring) 2008;16(3):643-653 [FREE Full text] [doi: 10.1038/oby.2007.118] [Medline: 18239586]
43. St-Onge MP, Grandner MA, Brown D, Conroy MB, Jean-Louis G, Coons M, et al. Sleep duration and quality: impact on lifestyle behaviors and cardiometabolic health: a scientific statement from the American Heart Association. Circulation 2016;134(18):e367-e386 [FREE Full text] [doi: 10.1161/CIR.0000000000000444] [Medline: 27647451]
44. Garfield V. The association between body mass index (BMI) and sleep duration: where are we after nearly two decades of epidemiological research? Int J Environ Res Public Health 2019;16(22):4327 [FREE Full text] [doi: 10.3390/ijerph16224327] [Medline: 31698817]
45. Liu W, Zhang R, Tan A, Ye B, Zhang X, Wang Y, et al. Long sleep duration predicts a higher risk of obesity in adults: a meta-analysis of prospective cohort studies. J Public Health (Oxf) 2019;41(2):e158-e168 [FREE Full text] [doi: 10.1093/pubmed/fdy135] [Medline: 30107483]
46. Grandner MA, Schopfer EA, Sands-Lincoln M, Jackson N, Malhotra A. Relationship between sleep duration and body mass index depends on age. Obesity (Silver Spring) 2015;23(12):2491-2498 [FREE Full text] [doi: 10.1002/oby.21247] [Medline: 26727118]
47. Li Y, Ma J, Yao K, Su W, Tan B, Wu X, et al. Circadian rhythms and obesity: timekeeping governs lipid metabolism. J Pineal Res 2020;69(3):e12682. [doi: 10.1111/jpi.12682] [Medline: 32656907]
48. Gifkins J, Johnston A, Loudoun R. The impact of shift work on eating patterns and self-care strategies utilised by experienced and inexperienced nurses. Chronobiol Int 2018;35(6):811-820. [doi: 10.1080/07420528.2018.1466790] [Medline: 29737884]
49. McHill AW, Phillips AJ, Czeisler CA, Keating L, Yee K, Barger LK, et al. Later circadian timing of food intake is associated with increased body fat. Am J Clin Nutr 2017;106(5):1213-1219 [FREE Full text] [doi: 10.3945/ajcn.117.161588] [Medline: 28877894]

XSL•FO
RenderX

50. Allison KC, Hopkins CM, Ruggieri M, Spaeth AM, Ahima RS, Zhang Z, et al. Prolonged, controlled daytime versus delayed eating impacts weight and metabolism. Curr Biol 2021;31(3):650-657.e3 [FREE Full text] [doi: 10.1016/j.cub.2020.10.092] [Medline: 33259790]
51. Lewis P, Oster H, Korf HW, Foster RG, Erren TC. Food as a circadian time cue—evidence from human studies. Nat Rev Endocrinol 2020;16(4):213-223. [doi: 10.1038/s41574-020-0318-z] [Medline: 32055029]
52. Wang Y, Kuang Z, Yu X, Ruhn KA, Kubo M, Hooper LV. The intestinal microbiota regulates body composition through NFIL3 and the circadian clock. Science 2017;357(6354):912-916 [FREE Full text] [doi: 10.1126/science.aan0677] [Medline: 28860383]
53. Thaiss CA, Zeevi D, Levy M, Zilberman-Schapira G, Suez J, Tengeler AC, et al. Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. Cell 2014;159(3):514-529 [FREE Full text] [doi: 10.1016/j.cell.2014.09.048] [Medline: 25417104]
54. Mota MC, Silva CM, Balieiro LCT, Fahmy WM, Crispim CA. Social jetlag and metabolic control in non-communicable chronic diseases: a study addressing different obesity statuses. Sci Rep 2017;7(1):6358 [FREE Full text] [doi: 10.1038/s41598-017-06723-w] [Medline: 28743872]
55. Zerón-Rugerio MF, Cambras T, Izquierdo-Pulido M. Social jet lag associates negatively with the adherence to the mediterranean diet and body mass index among young adults. Nutrients 2019;11(8):1756 [FREE Full text] [doi: 10.3390/nu11081756] [Medline: 31366143]
56. Quante M, Mariani S, Weng J, Marinac CR, Kaplan ER, Rueschman M, et al. Zeitgebers and their association with rest-activity patterns. Chronobiol Int 2019;36(2):203-213 [FREE Full text] [doi: 10.1080/07420528.2018.1527347] [Medline: 30365354]
57. Gottlieb DJ, Punjabi NM. Diagnosis and management of obstructive sleep apnea: a review. JAMA 2020;323(14):1389-1400. [doi: 10.1001/jama.2020.3514] [Medline: 32286648]

Abbreviations

BF%: body fat percentage
IS: interdaily stability
IV: intradaily variability
OSA: obstructive sleep apnea
PSQI: Pittsburgh Sleep Quality Index
TST: total sleep time
VAT: visceral adipose tissue area
WASO: wake after sleep onset

Edited by T de Azevedo Cardoso; submitted 21.06.23; peer-reviewed by S Jaiswal, V Sanchez Antelo, J Moreno; comments to author 11.11.23; revised version received 17.11.23; accepted 20.03.24; published 05.06.24.

Please cite as:
Chuang HH, Lin C, Lee LA, Chang HC, She GJ, Lin YH
Comparing Human-Smartphone Interactions and Actigraphy Measurements for Circadian Rhythm Stability and Adiposity: Algorithm Development and Validation Study
J Med Internet Res 2024;26:e50149
URL: https://www.jmir.org/2024/1/e50149
doi:10.2196/50149
PMID:38838328

©Hai-Hua Chuang, Chen Lin, Li-Ang Lee, Hsiang-Chih Chang, Guan-Jie She, Yu-Hsuan Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multicentric Assessment of a Multimorbidity-Adjusted Disability Score to Stratify Depression-Related Risks Using Temporal Disease Maps: Instrument Validation Study

Rubèn González-Colom[1], PhD; Kangkana Mitra[1], MSc; Emili Vela[2,3], MSc; Andras Gezsi[4], PhD; Teemu Paajanen[5], PhD; Zsófia Gál[6,7], MSc; Gabor Hullam[4,7], PhD; Hannu Mäkinen[5], PhD; Tamas Nagy[4,6,7], MSc; Mikko Kuokkanen[5,8,9], PhD; Jordi Piera-Jiménez[2,3,10], PhD; Josep Roca[1,11,12], MD, PhD; Peter Antal[4*], PhD; Gabriella Juhasz[6,7*], PhD; Isaac Cano[1,12*], PhD

[1]Fundació de Recerca Clínic Barcelona - Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain
[2]Catalan Health Service, Barcelona, Spain
[3]Digitalization for the Sustainability of the Healthcare - Institut d'Investigació Biomèdica de Bellvitge, Barcelona, Spain
[4]Department of Measurement and Information Systems, Budapest University of Technology and Economics, Budapest, Hungary
[5]Department of Public Health and Welfare, Finnish Health and Welfare Institute, Helsinki, Finland
[6]Department of Pharmacodynamics, Faculty of Pharmacy, Semmelweis University, Budapest, Hungary
[7]NAP3.0-SE Neuropsychopharmacology Research Group, Hungarian Brain Research Program, Semmelweis University, Budapest, Hungary
[8]Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine at University of Texas Rio Grande Valley, Brownsville, TX, United States
[9]Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[10]Faculty of Informatics, Telecommunications and Multimedia, Universitat Oberta de Catalunya, Barcelona, Spain
[11]Hospital Clínic de Barcelona, Barcelona, Spain
[12]Faculty of Medicine, Universitat de Barcelona, Barcelona, Spain
[*]these authors contributed equally

Corresponding Author:
Rubèn González-Colom, PhD
Fundació de Recerca Clínic Barcelona - Institut d'Investigacions Biomèdiques August Pi i Sunyer
C/Rosselló 149-153
Barcelona, 08036
Spain
Phone: 34 932275707
Email: rgonzalezc@recerca.clinic.cat

Abstract

Background: Comprehensive management of multimorbidity can significantly benefit from advanced health risk assessment tools that facilitate value-based interventions, allowing for the assessment and prediction of disease progression. Our study proposes a novel methodology, the Multimorbidity-Adjusted Disability Score (MADS), which integrates disease trajectory methodologies with advanced techniques for assessing interdependencies among concurrent diseases. This approach is designed to better assess the clinical burden of clusters of interrelated diseases and enhance our ability to anticipate disease progression, thereby potentially informing targeted preventive care interventions.

Objective: This study aims to evaluate the effectiveness of the MADS in stratifying patients into clinically relevant risk groups based on their multimorbidity profiles, which accurately reflect their clinical complexity and the probabilities of developing new associated disease conditions.

Methods: In a retrospective multicentric cohort study, we developed the MADS by analyzing disease trajectories and applying Bayesian statistics to determine disease-disease probabilities combined with well-established disability weights. We used major depressive disorder (MDD) as a primary case study for this evaluation. We stratified patients into different risk levels corresponding to different percentiles of MADS distribution. We statistically assessed the association of MADS risk strata with mortality, health care resource use, and disease progression across 1 million individuals from Spain, the United Kingdom, and Finland.

XSL•FO
RenderX

Results: The results revealed significantly different distributions of the assessed outcomes across the MADS risk tiers, including mortality rates; primary care visits; specialized care outpatient consultations; visits in mental health specialized centers; emergency room visits; hospitalizations; pharmacological and nonpharmacological expenditures; and dispensation of antipsychotics, anxiolytics, sedatives, and antidepressants ($P<.001$ in all cases). Moreover, the results of the pairwise comparisons between adjacent risk tiers illustrate a substantial and gradual pattern of increased mortality rate, heightened health care use, increased health care expenditures, and a raised pharmacological burden as individuals progress from lower MADS risk tiers to higher-risk tiers. The analysis also revealed an augmented risk of multimorbidity progression within the high-risk groups, aligned with a higher incidence of new onsets of MDD-related diseases.

Conclusions: The MADS seems to be a promising approach for predicting health risks associated with multimorbidity. It might complement current risk assessment state-of-the-art tools by providing valuable insights for tailored epidemiological impact analyses of clusters of interrelated diseases and by accurately assessing multimorbidity progression risks. This study paves the way for innovative digital developments to support advanced health risk assessment strategies. Further validation is required to generalize its use beyond the initial case study of MDD.

(J Med Internet Res 2024;26:e53162) doi:10.2196/53162

KEYWORDS

health risk assessment; multimorbidity; disease trajectories; major depressive disorder

Introduction

Background

The co-occurrence of multiple chronic diseases, known as multimorbidity [1], affects 1 in 3 adults. Its prevalence rises with age, affecting 60% of individuals aged between 65 and 74 years and escalating to 80% among those aged ≥85 years [2]. Due to its association with poor prognosis, functional impairment, and reduced quality of life, multimorbidity is considered a global health care challenge [3,4] tied to complex clinical situations, leading to increased encounters with health care professionals, hospitalizations, and pharmacological prescriptions, resulting in a substantial rise in health care costs [5]. The emergence of multimorbidity is not arbitrary and frequently aligns with shared risk factors and underlying pathophysiological mechanisms [6-8] that result from complex interactions between genetic and environmental factors throughout the life span [9]. Perceiving diseases not in isolation but as integral components of a more extensive, interconnected system within the human body has led to the emergence of network medicine [10,11]. Network medicine analyzes disease co-occurrence patterns, aiming to understand the complex connections between diseases to uncover biomarkers, therapeutic targets, and potential interventions [12,13]. The studies investigating the temporal patterns of disease concurrence, or disease trajectories [14,15], rely on a pragmatic approach to this concept to yield a better understanding of the time-dependent relationships among diseases and establish a promising landscape to identify disease-disease causal relationships.

According to this paradigm, a disease-centered approach might lead to suboptimal treatment of patients with multiple chronic conditions, triggering the need to implement new tools to enhance the effectiveness of health services [16]. In this regard, multimorbidity-adjusted health risk assessment (HRA) tools [17-21], such as the morbidity groupers, are crucial for assessing the comprehensive health needs of patients with multimorbidity [22]. HRA uses algorithms and patient data to categorize individuals by risk, aiding health care professionals in customizing interventions, optimizing resource allocation, and enhancing patient outcomes through preventive care. HRA tools facilitate efficient case-finding and screening processes [23]. Case finding targets the individuals at high risk, which is crucial for specialized health care programs, whereas patient screening detects latent illnesses early, enabling cost-effective interventions to prevent disease progression and reduce health care demands.

However, despite their widespread use, prevailing population-based HRA tools such as the Adjusted Clinical Groups [24], Clinical Risk Groups [25], or Adjusted Morbidity Groups (AMG) [4,21] still do not incorporate information on disease trajectories in their calculations. The AMG system is currently used in Catalonia (Spain; 7 million inhabitants) for health policy and clinical purposes. Adding disease-disease association information to the AMG (or other morbidity groupers) may open new avenues for implementing epidemiological impact analyses concerning clusters of interrelated diseases. In addition, it may facilitate the construction of risk groups that accurately represent probabilities of developing new associated disease conditions [26] susceptible to early prevention.

Objectives

While acknowledging current limitations, this study sought to explore the feasibility of incorporating procedures relevant to the study of disease trajectories [14,15] and novel techniques for analyzing dependency relationships between concomitant diseases [27,28] to improve the capabilities of the current morbidity groupers. This approach might better adjust the estimations of the burden of morbidity to clusters of diseases and improve the ability to anticipate the progression of multimorbidity.

We used major depressive disorder (MDD; F32-F33 in the *International Classification of Diseases, 10th Revision, Clinical Modification* [*ICD-10-CM*] [29]) as a use case due to its clinical relevance in multimorbidity management. However, this study pursued to showcase a methodology applicable beyond MDD, allowing for the assessment of the impact of multimorbidity across different clusters of diseases.

XSL•FO
RenderX

This paper describes the process of development and assessment of the Multimorbidity-Adjusted Disability Score (MADS) through an observational retrospective multicentric cohort study, showcasing a pioneering approach that integrates advanced techniques for analyzing disease associations, insights from the analysis of disease trajectories, and a comprehensive scoring method aimed at evaluating the disease burden. The MADS was designed to stratify patients with different health needs according to (1) the disease burden caused by MDD and its comorbidities on individuals and health systems and (2) the risk of morbidity progression and the onset of MDD comorbid conditions.

On the basis of the temporal disease maps among MDD and highly prevalent disease conditions [30] generated using Bayesian direct multimorbidity maps (BDMMs) [27,28], a promising method for filtering indirect disease associations in the context of the European Research Area on Personalized Medicine project "Temporal disease map based stratification of depression-related multimorbidities: towards quantitative investigations of patient trajectories and predictions of multi-target drug candidates" (TRAJECTOME) [31], we combined the probabilities of relevance (PRs) among MDD and its comorbid conditions with the disability weights (DWs) [32], documented in the 2019 revision of the Global Burden of Disease (GBD) study, to compute the MADS. We used the MADS to generate a risk pyramid and stratify the study population into 5 risk groups using different percentiles of MADS distribution. Finally, we analyzed the correspondence between the MADS risk groups and health outcomes through a cross-sectional analysis of mortality and use of health care resources and a longitudinal analysis of disease prevalence and incidence of new disease onsets. The clinical relevance of the identified risk groups was assessed through a multicentric assessment of the findings. To this end, MADS performance was analyzed using data from 3 independent European cohorts from the United Kingdom, Finland, and Spain including >1 million individuals.

Methods

Overview

The development and evaluation of the MADS involved the following steps (Figure 1).

Figure 1. Workflow for building and assessing the Multimorbidity-Adjusted Disability Score (MADS). BDMM: Bayesian direct multimorbidity map; CHSS: Catalan Health Surveillance System; DW: disability weight; ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; PR: probability of relevance; THL: Finnish Institute for Health and Welfare; UKB: UK Biobank.

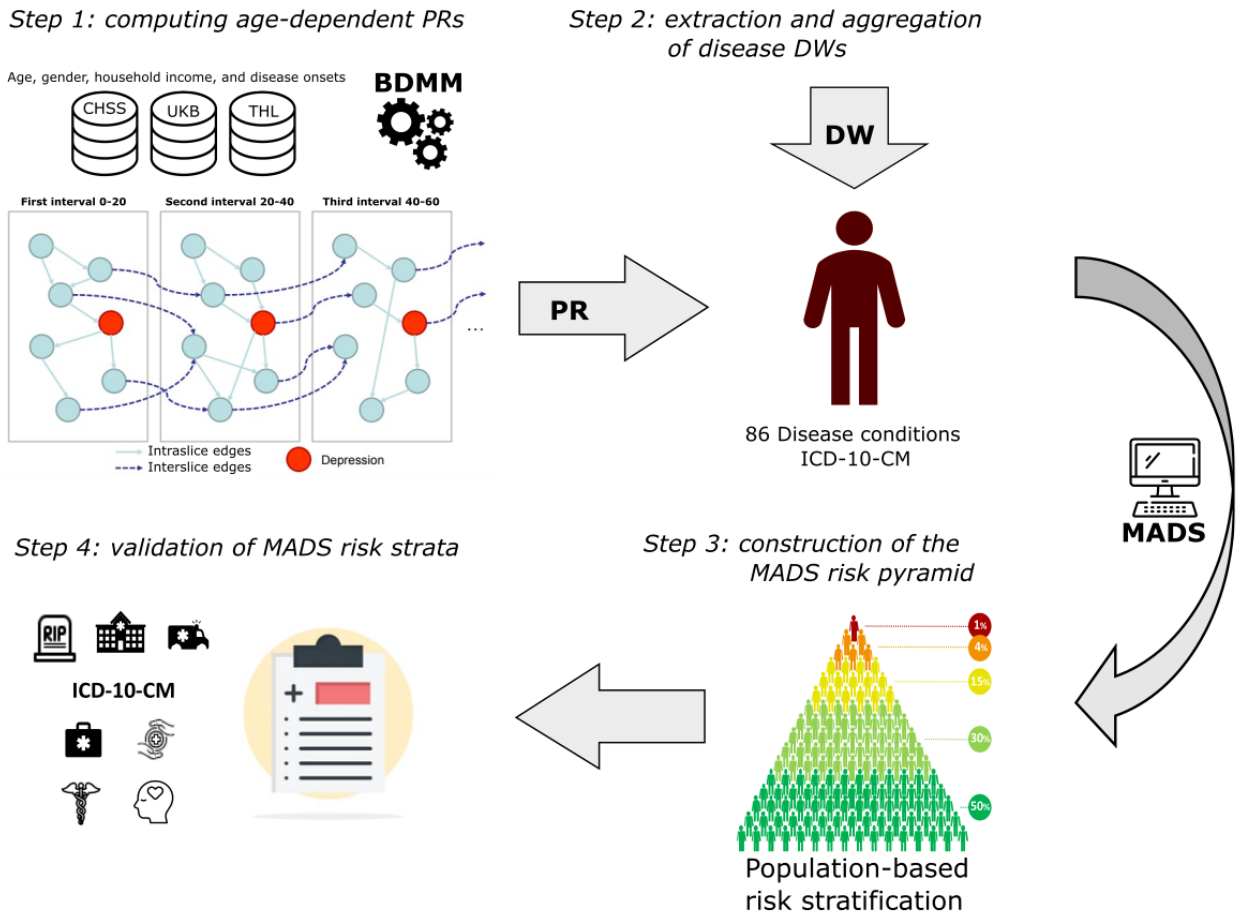

Step 1 involved computing age-dependent disease-disease PRs using the BDMM method in 4 age intervals (0-20 years, 0-40 years, 0-60 years, and 0-70 years). This analysis resulted in an inhomogeneous dynamic Bayesian network that determined the PR for MDD against the most prevalent co-occurring diseases in the 3 European cohorts considered in TRAJECTOME,

XSL•FO
RenderX

namely, the Catalan Health Surveillance System (CHSS) [33], the UK Biobank (UKB) [34], and the Finnish Institute for Health and Welfare (THL) [35] cohorts. The THL cohort amalgamates information from the FINRISK [36] 1992, 1997, 2002, 2007, and 2012; the FinHealth [37] 2017; and the Health [38] 2000 and 2011 studies.

In step 2, combining the PR of every disease condition assessed in the study with its corresponding DW extracted from the GBD 2019 study, we estimated the morbidity burden caused by MDD and its comorbid conditions. The MADS was computed following a multiplicative combination of the PR and DW of all the disease conditions present in an individual.

Step 3 involved using the MADS to stratify patients into different risk levels corresponding to different percentiles of the population-based risk pyramid of each patient cohort: (1) very low risk (percentile ≤50), (2) low risk (percentile 50 to percentile 80), (3) moderate risk (percentile 80 to percentile 90), (4) high risk (percentile 90 to percentile 95), and (5) very high risk (percentile >99).

Finally, in step 4, the correspondence between the MADS risk strata and health outcomes was analyzed through a cross-sectional analysis of use of health care resources, mortality, pharmacological burden, and health care expenditure and a longitudinal analysis of disease prevalence and incidence of new disease onsets. The results were validated through a multicentric replication of the findings in the 3 study cohorts including 1,041,014 individuals.

Step 1: Computing Age-Dependent PRs

BDMMs were used to assess direct and indirect associations between MDD and 86 potential comorbid conditions. The set of 86 disease conditions considered in the study had a prevalence of >1% in all the study cohorts. The list of diseases and their associated *ICD-10-CM* [29] codes are shown in Multimedia Appendix 1.

This step considered information on *disease diagnosis* (disease conditions were cataloged using the first 3 characters of the *ICD-10-CM* codes), age at disease onset (the age at disease onset corresponds to the first diagnosis in a lifetime for each *ICD-10-CM* code), sex, and socioeconomic status (annual average total household income [before tax with copayment exemption] as a categorical variable with 3 categories: <€18,000 [US $19,565.30], €18,000-100,000 [US $19,565.30-108,696], and >€100,000 [US $108,696]).

BDMM analysis resulted in an inhomogeneous dynamic Bayesian network, which was used to compute temporal PR, ranging from 0 (no association) to 1 (strong association), for MDD in conjunction with sex, socioeconomic status, and the set of 86 predetermined diseases [30]. To construct the trajectories, the PR was calculated in 4 different age ranges: 0 to 20 years, 0 to 40 years, 0 to 60 years, and 0 to 70 years. The PRs calculated and used for MADS computation are reported in Multimedia Appendix 1. Further details regarding the core analysis conducted in TRAJECTOME can be found in the study by Juhasz et al [30].

Step 2: Extracting and Aggregating Disease DWs

The MADS was developed by weighting the DWs of single diseases according to their estimated PRs against MDD. DWs indicate the degree of health loss based on several health outcomes and are used to indicate the total disability caused by a certain health condition or disease. Often, the DWs present specific disability scores tailored to the severity of the disease. The disease categories, severity distribution, and associated DWs used in this study were extracted from the GBD 2019 study and are reported in Multimedia Appendix 1.

DWs were extracted and aggregated as follows. First, we considered only the DWs of MDD and the set of 86 disease codes. Second, we considered the DWs of all the chronic conditions diagnosed in patients' lifetime, whereas as the disability caused by acute illnesses is transitory, the DWs for the acute diseases diagnosed >12 months before the MADS assessment were arbitrarily set to 0 (no disability). Third, due to the unavailability of information on the severity of diagnoses, we determined the DWs of each disease condition by calculating the weighted mean of the DWs associated with the disease severity categories and their prevalence. In instances in which the severity distribution was not available, we computed the arithmetic mean of the DWs of each severity category. Fourth, we weighted the DWs according to the PR of each disease condition with respect to MDD. The PRs were adjusted according to the age of disease onset, discretized in the following intervals: 0 to 20 years, 20 to 40 years, 40 to 60 years, and >60 years.

As the DWs do not account for multimorbidity in their estimates, the use of DW independently can cause inaccuracies in the burden of disease estimations, particularly in aging populations that include large proportions of persons with ≥2 disabling disease conditions [39]. Consequently, we combined the DW and the PR for all the disease conditions present in 1 individual following a multiplicative approach (equation 1) [40], aggregating several DWs in a single score that accounts for the overall disability caused by numerous concurrent chronic conditions in which every comorbid disease increases the utility loss of a patient, although it is less than the sum of the utility loss of both diseases independently.

In equation 1, "DW" stands for disability weight, "PR" stands for probability of relevance, and "n" is the number of diseases present in 1 individual.

The MADS pseudocode is reported in Multimedia Appendix 1.

Step 3: Constructing the MADS Risk Pyramid

Once calculated, the MADS was used to stratify patients in different levels of risk according to the percentiles of its distribution in the source population, producing the following risk pyramid: (1) very low risk (percentile ≤50), (2) low risk (percentile 50-percentile 80), (3) moderate risk (percentile 80-percentile 90), (4) high risk (percentile 90-percentile 95), and (5) very high risk (percentile >99).

XSL•FO
RenderX

Step 4: Evaluating MADS Risk Strata

The clinical relevance of the risk strata was assessed through two interconnected analyses: (1) a cross-sectional analysis of health outcomes and (2) a longitudinal analysis of disease prevalence and incidence of new onsets.

Cross-Sectional Analysis of Health Outcomes and Use of Health Care Resources

To validate the results of the MADS, we conducted a cross-sectional analysis of clinical outcomes within the 12 months following the MADS assessment. The burden of MDD and its comorbidities on patients and health care providers, corresponding to each risk group of the MADS risk pyramid, was assessed using the following features (the parameters evaluated in each cohort may vary depending on the availability of the requested information in the source databases):

- Prescriptions for psycholeptic and psychoanaleptic drugs (information available in all the databases)—the prescribed medication was cataloged using the first 4 characters from Anatomical Therapeutic Chemical classification [41] codes, resulting in the following categories: *antipsychotics (N05A), anxiolytics (N05B), hypnotics and sedatives (N05C), and antidepressants (N06A)*
- Cost of the pharmacological prescriptions in euros (information available only in the CHSS and THL)
- Mortality rates (information available only in the CHSS and THL)
- Contacts and encounters with health care professionals (information available only in the CHSS), encompassing (1) primary care visits, (2) specialized care outpatient visits, (3) ambulatory visits in mental health centers, (4) emergency room visits, (5) planned and unplanned hospital admissions, and (6) admissions in mental health centers
- Total health care expenditure (information available only in the CHSS), including (1) direct health care delivery costs; (2) pharmacological costs; and (3) other billable health care costs such as nonurgent medical transportation, ambulatory rehabilitation, domiciliary oxygen therapy, and dialysis

We assessed the effect of sex and age, replicating the analyses disaggregated by sex and age. The age ranges were discretized into the following categories: 0 to 20 years, 20 to 40 years, 40 to 60 years, and >60 years.

Longitudinal Analysis of Disease Prevalence and Incidence of New Onsets

To address the age dependency of disease onsets, we performed a longitudinal analysis of the prevalence of a target disease and the incidence of new diagnostics within the 5 years following the MADS assessment.

We iteratively computed the MADS in 5-year intervals throughout the patients' lives. Within each interval, the population was stratified based on the MADS distribution. Subsequently, within each risk tier, the prevalence of the target disease and the incidence of new disease onset over the subsequent 5 years were calculated. Only individuals with complete information for the next interval at each time point of the analysis were included.

In the analysis, we considered only the chronic disease conditions with a PR against MDD of ≥0.80 in at least 1 of the 4 age intervals assessed, namely, 0 to 20 years, 0 to 40 years, 0 to 60 years, and 0 to 70 years. It resulted in the following set of mental diseases—*MDD (F32-F33), schizophrenia (F20), bipolar disorder (F31), anxiety-related disorders (F40-F41), stress-related disorders (F43)*, and *mental disorders related to alcohol abuse (F10)*—and the following somatic diseases: *irritable bowel syndrome (K58), overweight and obesity (E66)*, and *gastroesophageal reflux (K21)*.

Data Sources

The study was conducted using data from 3 public health cohorts.

CHSS Cohort

The main cohort used in MADS development was extracted from the CHSS. Operated by a public single-payer system (CatSalut) [42] since 2011, the CHSS gathers information across health care tiers on the use of public health care resources, pharmacological prescriptions, and patients' basic demographic data, including registries of 7.5 million citizens from the entire region of Catalonia (Spain). Nevertheless, for MADS development purposes, we considered only registry data from citizens residing in the entire Integrated Health District of Barcelona-Esquerra between January 1, 2011, and December 31, 2019 (n=654,913). To validate the results of the MADS, we retrieved additional information from the CHSS corresponding to the 12 months after the MADS assessment, from January 1, 2020, to December 31, 2020. It should be noted that all the deceased patients, in addition to those who moved their residence outside of the Integrated Health District of Barcelona-Esquerra between 2011 and 2019, were discarded from the MADS assessment analysis; the remaining subset of patients comprised 508,990 individuals.

UKB Cohort

The UKB data considered in this study contained medical and phenotypic data from participants aged between 37 and 93 years. Recruitment was based on National Health Service patient registers, and initial assessment visits were carried out between March 3, 2006, and October 1, 2010 (n=502,504). The analyzed data included disease diagnosis and onset time, medication prescriptions, and socioeconomic descriptors.

THL Cohort

The THL cohort integrates information from the FINRISK [36] 1992, 1997, 2002, 2007, and 2012; FinHealth [37] 2017; and Health [38] 2000/2011 studies. For the consensual clustering, 41,092 participants were used from Finnish population surveys. After data cleaning, 30,961 participants remained. These participants, aged 20 to 100 years, were chosen at random from the Finnish population and represented different parts of Finland.

Demographic information on the study cohorts is shown in the Results section.

Ethical Considerations

As a multicentric study, TRAJECTOME accessed data from multiple cohorts, all subject to the legal regulations of their

XSL•FO
RenderX

respective regions of origin, and obtained the necessary approvals from the corresponding ethics committees.

For the CHSS cohort, the Ethical Committee for Human Research at Hospital Clínic de Barcelona approved the core study of TRAJECTOME on March 24, 2021 (HCB/2020/1051), and subsequently approved the analysis for the generation and the assessment of the MADS on July 25, 2022 (HCB/2022/0720).

The UKB cohort received ethics approval from the National Research Ethics Service Committee North West–Haydock (reference 11/NW/0382).

The THL cohort integrates information from the FINRISK databases (1997 [ethical committee of the National Public Health Institute; statement 38/96; October 30, 1996], 2002 [Helsinki University Hospital, ethical committee of epidemiology and public health; statement 87/2001; reference 558/E3/2001; December 19, 2001], 2007 [Helsinki University Hospital, coordinating ethics committee; Dnro HUS 229/EO/2006; June 20, 2006], and 2012 [Helsinki University Hospital, coordinating ethics committee; Dnro HUS 162/13/03/11; December 1, 2011]), the FinHealth 2017 (Helsinki University Hospital, coordinating ethics committee; 37/13/03/00/2016; March 22, 2016), and the Health 2000 to 2011 databases (ethical committee of the National Public Health Institute, 8/99/12; Helsinki University Hospital, ethical committee of epidemiology and public health, 407/E3/2000; May 31, 2000, and June 17, 2011).

The ethics committees exempted the requirement to obtain informed consent for the analysis and publication of retrospectively acquired and fully anonymized data in the context of this noninterventional study.

All the data were handled in compliance with the General Data Protection Regulation 2016/679, which safeguards data protection and privacy for all individuals in the European Union (EU). The study was conducted in conformity with the Helsinki Declaration (Stronghold Version, Brazil, October 2013) and in accordance with the protocol and the relevant legal requirements (Law 14/2007 on Biomedical Research of July 3).

Statistical Analysis

The results of the cross-sectional analysis of health outcomes and use of health care resources were evaluated through various metrics. Mortality rates were summarized as cases per 1000 inhabitants. In contrast, numeric health outcome variables were described by the average number of cases per person, per 100 inhabitants, or per 1000 inhabitants according to their prevalence. Average health care expenditures were reported in euro per person. Kruskal-Wallis tests, supplemented with Bonferroni-adjusted post hoc right-tailed Dunn tests, and pairwise Fisher exact tests were used to evaluate changes in the target outcomes across the risk pyramid tiers. Statistical significance was determined by considering a P value of <.05 in all analyses.

The results of the longitudinal analysis on disease prevalence and on the incidence of new disease onsets of MDD and 9 mental and somatic MDD-related chronic conditions (PR>0.80) were expressed in percentages and in per thousand (‰), respectively.

All the data analyses were conducted using R (version 4.1.1; R Foundation for Statistical Computing) [43]. The MADS algorithm was fully developed and tested in the CHSS database and transferred to the other sites using an R programming executable script.

The study is reported according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [23] guidelines for observational studies.

Results

Sociodemographic Characteristics of the Study Cohorts

One of the first results was the characterization of the 3 study cohorts and comparison of the sociodemographic attributes of their MADS risk groups (Table 1). All the individuals were classified into distinct risk strata based on quantiles of MADS distribution within the source population, resulting in the formation of the subsequent risk pyramid: very-low-risk tier (percentile ≤50), low-risk tier (percentile 50 to percentile 80), moderate-risk tier (percentile 80 to percentile 90), high-risk tier (percentile 90 to percentile 95), and very high–risk tier (percentile >99).

XSL•FO
RenderX

Table 1. Demographic characteristics of each stratum of the Multimorbidity-Adjusted Disability Score (MADS) risk pyramid in the 3 study cohorts: Catalan Health Surveillance System (CHSS) [33], UK Biobank (UKB) [34], and Finnish Institute for Health and Welfare (THL) [35][a].

Risk pyramid tier and demographics	CHSS	THL	UKB
All cases			
Participants, N	507,549	30,961	502,504
Age (y), mean (SD)	45.36 (23.07)	64.27 (14.28)	61.48 (9.31)
Sex, n (%)			
Male	237,598 (46.8)	14,435 (46.61)	229,122 (45.6)
Female	269,951 (53.2)	16,526 (53.39)	273,382 (54.4)
Household income, n (%)			
Low (<€18,000 [US $19,565.30])	262,753 (51.76)	11,489 (37.1)	117,737 (23.42)
Medium (€18,000-100,000 [US $19,565.30-$108,696])	223,369 (44)	10,025 (32.4)	358,492 (71.34)
High (>€100,000 [US $108,696])	21,427 (4.24)	9447 (30.5)	26,275 (5.24)
Major depressive disorder prevalence, n (%)	38,479 (7.58)	2287 (7.39)	53,466 (10.64)
Very high risk (percentile >99)			
Participants, N	5651	310	5026
Age (y), mean (SD)	55.74 (18.83)	68.83 (14.86)	61.7 (8.75)
Sex, n (%)			
Male	2322 (41.09)	129 (41.61)	2207 (43.89)
Female	3329 (58.91)	181 (58.39)	2819 (56.11)
Household income, n (%)			
Low (<€18,000 [US $19,565.30])	4343 (76.86)	191 (61.61)	2285 (45.47)
Medium (€18,000-100,000 [US $19,565.30-$108,696])	1251 (22.13)	77 (24.84)	2620 (52.13)
High (>€100,000 [US $108,696])	57 (1.01)	42 (13.55)	121 (2.4)
Major depressive disorder prevalence, n (%)	3870 (68.48)	186 (60)	4370 (86.94)
High risk (percentile 95 to percentile 99)			
Participants, N	22,894	1238	20,084
Age (y), mean (SD)	60.08 (20.00)	65.12 (15.10)	63.2 (8.74)
Sex, n (%)			
Male	7170 (31.32)	559 (45.23)	7545 (37.57)
Female	15,724 (68.68)	679 (54.77)	12,539 (62.43)
Household income, n (%)			
Low (<€18,000 [US $19,565.30])	14,568 (63.65)	690 (55.74)	7626 (37.97)
Medium (€18,000-100,000 [US $19,565.30-$108,696])	7946 (34.7)	327 (26.41)	12,003 (59.76)
High (>€100,000 [US $108,696])	380 (1.65)	221 (17.85)	455 (2.27)
Major depressive disorder prevalence, n (%)	18,368 (80.27)	734 (59.29)	19,039 (94.78)
Moderate risk (percentile 80 to percentile 95)			
Participants, N	84,371	4644	75,378
Age (y), mean (SD)	54.56 (21.87)	68.86 (14.77)	63.6 (9.02)
Sex, n (%)			
Male	34,462 (40.86)	2201 (47.4)	34,282 (45.48)
Female	49,909 (59.14)	2441 (52.6)	41,096 (54.52)
Household income, n (%)			
Low (<€18,000 [US $19,565.30])	49,818 (59.05)	2285 (49.24)	23,208 (30.77)

XSL•FO
RenderX

Risk pyramid tier and demographics	CHSS	THL	UKB
Medium (€18,000-100,000 [US $19,565.30-$108,696])	32,822 (38.9)	1437 (30.93)	49,684 (65.93)
High (>€100,000 [US $108,696])	1731 (2.05)	920 (19.83)	2486 (3.3)
Major depressive disorder prevalence, n (%)	16,241 (19.25)	1367 (29.43)	25,776 (34.2)
Low risk (percentile 50 to percentile 80)			
Participants, N	162,170	9,266	150,759
Age (y), mean (SD)	47.66 (24.20)	66.16 (14.15)	62.2 (9.39)
Sex, n (%)			
Male	77,082 (47.53)	4132 (44.58)	70,550 (46.72)
Female	85,088 (52.47)	5137 (55.42)	80,209 (53.28)
Household income, n (%)			
Low (<€18,000 [US $19,565.30])	85,936 (5300)	3623 (39.08)	36,773 (24.42)
Medium (€18,000-100,000 [US $19,565.30-$108,696])	71,429 (44.06)	3081 (33.25)	106,441 (70.59)
High (>€100,000 [US $108,696])	4805 (2.96)	2565 (27.67)	7545 (4.99)
Major depressive disorder prevalence, n (%)	0 (0)	0 (0)	2002 (1.3)
Very low risk (percentile ≤50)			
Participants, N	232,463	15,503	251,257
Age (y), mean (SD)	38.72 (20.72)	61.62 (13.55)	60.3 (9.22)
Sex, n (%)			
Male	116,562 (50.12)	7414 (47.83)	114,538 (45.59)
Female	115,901 (49.88)	8088 (52.17)	136,719 (54.41)
Household income, n (%)			
Low (<€18,000 [US $19,565.30])	108,088 (46.48)	4700 (30.32)	47,845 (19.04)
Medium (€18,000-100,000 [US $19,565.30-$108,696])	109,921 (47.30)	5103 (32.92)	187,744 (74.72)
High (>€100,000 [US $108,696])	14,454 (6.22)	5699 (36.76)	15,668 (6.24)
Major depressive disorder prevalence, n (%)	0 (0)	0 (0)	2279 (0.9)

[a]The prevalence of depression was calculated considering both F32 and F33 International Classification of Diseases, 10th Revision, Clinical Modification diagnostic codes. Kruskal-Wallis tests were used to assess changes in the target outcomes according to the risk pyramid tiers (statistical significance: $P<.05$; H0="all MADS risk groups have the same outcome distribution"; H1="at least one MADS risk group has a different outcome distribution than the others"). $P<.001$ for age, sex, household income, and major depressive disorder prevalence for all cohorts.

It is imperative to underscore the fundamental distinctions in the cohorts under study to comprehend the inherent sociodemographic disparities across them. Specifically, the THL and UKB cohorts predominantly consist of data derived from biobanks, specifically focusing on the middle-aged and older adult population. In contrast, the CHSS cohort represents a population-based sample encompassing the entire population spectrum.

It is worth noting that a common pattern was observed among all the cohorts in the age distribution of the citizens at risk. Although the MADS is an additive morbidity grouper, it did not monotonically increase with age. Remarkably, a notable proportion of high-risk cases were observed within the age range of 40 to 60 years, when depression typically manifests for the first time on average.

A divergence in the sex distribution across the risk strata was observable and especially noticeable in the CHSS and UKB cohorts, where the morbidity burden associated with depression and its related diseases was amplified in women ($P<.001$). Similarly, the disability caused by depression and its comorbidities was larger in families with fewer economic resources ($P<.001$). Overall, the prevalence of MDD was greater in the UKB cohort than in the other cohorts. However, upon analyzing the allocation of the population with depression in the risk pyramid, a total of 57.79% (22,238/38,479) of individuals diagnosed with MDD were categorized in the "high"- and "very high" risk tiers in the CHSS cohort, whereas the proportion of individuals diagnosed with MDD who were allocated to the tip of the risk pyramid was 40.22% (920/2287) in the THL cohort and 43.78% (23,409/53,466) in the UKB cohort.

Assessment of the MADS Risk Groups

Assessment of the PRs

Analyzing the relationship between MDD and the morbidities assessed in the study is essential to interpreting the MADS risk strata. This analysis revealed various relevant connections

XSL•FO
RenderX

between MDD and the diseases investigated, encompassing both acute and chronic conditions, with the latter being particularly noteworthy due to their nontransient nature. Notably, the cluster of mental and behavioral disorders showed the highest average PRs in depression. However, relevant associations also emerged among MDD and specific chronic somatic diseases affecting multiple organic systems (Figure 2).

Figure 2. Average probabilities of relevance between major depressive disorder and 45 chronic conditions used to compute the Multimorbidity-Adjusted Disability Score.

Use of Health Care Resources

The impact of MADS risk groups on health care systems was evaluated by investigating the correlation between the MADS risk categories and the use of health resources over the 12-month period following the MADS assessment within the CHSS cohort (Table 2). The results revealed significantly different distributions of the assessed outcomes across the MADS risk tiers, including primary care visits ($P<.001$), specialized outpatient visits ($P<.001$), emergency room visits ($P<.001$), hospital admissions ($P<.001$), and ambulatory visits in mental health centers ($P<.001$), as well as the pharmacological burden ($P<.001$). Furthermore, the results of the pairwise comparisons between adjacent risk tiers illustrated a substantial and gradual pattern of increased health care use as individuals progress from lower MADS risk tiers to higher MADS risk tiers, reflecting an escalation in health care needs and requirements. Overall, patients with higher MADS scores exhibited a greater likelihood of experiencing morbidity-related adverse events, which subsequently leads to recurrent interactions with health care systems across multiple levels.

XSL•FO
RenderX

Table 2. Use of health care resources over 12 months in each stratum of the Multimorbidity-Adjusted Disability Score risk pyramid for the Catalan Health Surveillance System cohort[a].

Risk pyramid tier	Primary care visits (visits per person)	Specialized outpatient visits (visits per person)	Emergency room visits (visits per 100 inhabitants)	Hospital admissions (admissions per 100 inhabitants)	Mental health visits (visits per 100 inhabitants)	Number of prescriptions (prescriptions per person)
Very high risk (percentile >99)	12.50	3.07[b]	135.00[b]	28.50[b]	554.00[b]	8.02[b]
High risk (percentile 95 to percentile 99)	11.90[b]	2.56[b]	87.20[b]	20.60[b]	136.00[b]	7.48[b]
Moderate risk (percentile 80 to percentile 95)	9.03[b]	1.82[b]	61.90[b]	14.50[b]	44.20[b]	5.11[b]
Low risk (percentile 50 to tpercentile 80)	6.21[b]	1.21[b]	42.40[b]	8.87[b]	15.10[b]	3.20[b]
Very low risk (percentile ≤50)	2.96	0.50	23.40	3.25	5.96	1.07

[a]Kruskal-Wallis tests were used to assess changes in the target outcomes according to the risk pyramid tiers (*P* value). Subsequent pairwise comparisons between each risk tier and the next level of less risk were conducted using right-tailed Dunn post hoc tests (statistical significance: *P*<.05).

[b]*P*<.001.

Mortality and Health Care Expenditure

We conducted a cross-sectional analysis investigating mortality rates and the health care expenditure within the 12 months following the MADS assessment, expressed as the average health care expenditure per capita and differentiating between pharmaceutical and nonpharmaceutical costs within the CHSS and THL cohorts (Table 3). Significant variations in mortality rates were observed across the risk pyramid tiers (*P*<.001), with rates in the high-risk strata being markedly elevated (ranging from 5 to 20 times depending on the cohort) compared to those for low-risk individuals. Furthermore, the distribution of average health care expenditures per person was significantly different among the risk tiers, with both pharmacological and nonpharmacological expenses demonstrating disparities (*P*<.001). Pairwise comparisons further indicated that individuals at the highest-risk tier incurred substantially greater health care costs than those at the lowest tier, reflecting a gradient of financial impact correlated with increased risk levels.

Table 3. Mortality rates and pharmacological and nonpharmacological health care expenditure in euro over 12 months in each stratum of the Multimorbidity-Adjusted Disability Score risk pyramid in the Catalan Health Surveillance System (CHSS) [33] and Finnish Institute for Health and Welfare (THL) [35][a].

Risk pyramid tier	Mortality (cases per 1000 inhabitants)		Pharmacological expenditure in euro per person, mean (SD)		Hospitalization expenditure in euro per person, mean (SD)		Total expenditure in euro per person—CHSS, mean (SD)
	CHSS	THL	CHSS	THL	CHSS	THL	
Very high risk (percentile >99)	46.2[b]	36.0[b]	1214[b]	966	539[b]	270	12,517[b]
High risk (percentile 95 to percentile 99)	41.5[b]	33.7[b]	772[b]	1131[b]	383[b]	340[b]	8404[b]
Moderate risk (percentile 80 to percentile 95)	25.5[b]	32.2[b]	485[b]	1077[b]	270[b]	254[b]	5209[b]
Low risk (percentile 50 to percentile 80)	11.5[b]	14.8[b]	292[b]	810[b]	165[b]	185[b]	3075[b]
Very low risk (percentile ≤50)	2.57	7.3	99	363	60	123	1192

[a]Kruskal-Wallis tests were used to assess changes in the target outcomes according to the risk pyramid tiers (*P* value). Subsequent pairwise comparisons between each risk tier and the next level of less risk were conducted using right-tailed Dunn post hoc tests. Pairwise comparisons of Fisher exact tests were used to assess changes in mortality rates. Statistical significance: *P*<.05.

[b]*P*<.001.

XSL•FO
RenderX

Pharmacological Burden

This study also examined the pharmacological burden on individuals after 12 months following the MADS assessment (Table 4). The data analysis revealed distinct patterns of medication use across the risk tiers, with significant differences in the use of antidepressants, antipsychotics, anxiolytics, and sedatives (*P*<.001 in all cases). This trend, consistently observed across the 3 cohorts, was further emphasized by pairwise comparisons between adjacent risk levels, which revealed a strong positive correlation between higher-risk strata and increased pharmaceutical consumption. This upward trend in medication use forms a clear gradient, demonstrating that individuals in progressively higher-risk tiers face substantially greater pharmaceutical needs.

Table 4. Prescription of depression-related pharmacological treatments over 12 months in each stratum of the Multimorbidity-Adjusted Disability Score risk pyramid in the Catalan Health Surveillance System (CHSS) [33], UK Biobank (UKB) [34], and Finnish Institute for Health and Welfare (THL) [35][a].

Risk pyramid tier	Antipsychotics (N05A; prescriptions per person)			Anxiolytics (N05B; prescriptions per person)			Hypnotics and sedatives (N05C; prescriptions per person)			Antidepressants (N06A; prescriptions per person)		
	CHSS	THL	UKB	CHSS	THL	UKB	CHSS	THL	UKB	CHSS	THL	UKB
Very high risk (percentile >99)	0.75[b]	0.60[b]	0.33[b]	0.47	0.21[b]	0.27[b]	0.15[b]	0.14[b]	0.24[b]	0.79[b]	0.43[b]	0.80[b]
High risk (percentile 95 to percentile 99)	0.20[b]	0.27[b]	0.18[b]	0.46[b]	0.19[b]	0.20[b]	0.10[b]	0.12[b]	0.19[b]	0.66[b]	0.41[b]	0.71[b]
Moderate risk (percentile 80 to percentile 95)	0.07[b]	0.08[b]	0.15[b]	0.28[b]	0.08[b]	0.16[b]	0.05[b]	0.10[b]	0.18[b]	0.27[b]	0.27[b]	0.54[b]
Low risk (percentile 50 to percentile 80)	0.03[b]	0.03[b]	0.13[b]	0.14[b]	0.04[b]	0.12[b]	0.02[b]	0.07[b]	0.13[b]	0.08[b]	0.11[b]	0.36[b]
Very low risk (percentile ≤50)	0.01	0.01	0.11	0.04	0.02	0.09	0.01	0.04	0.10	0.02	0.06	0.26

[a]For recurrently dispensed medication, only the first prescription was considered in the analysis. Kruskal-Wallis tests were used to assess changes in the target outcomes according to the risk pyramid tiers (*P* value). Subsequent pairwise comparisons between each risk tier and the next level of less risk were conducted using right-tailed Dunn post hoc tests. Statistical significance: *P*<.05.

[b]*P*<.001.

To evaluate the influence of age and sex on the outcomes examined in this section, we replicated all the previously presented results categorizing the outcomes by sex and age and reported them in Multimedia Appendix 1. The results suggest that the morbidity burden in individuals might be a primary driver influencing the occurrence of adverse health events and the heightened use of health care resources.

Multimorbidity Progression

We analyzed the prevalence and incidence of new MDD-associated diagnoses and the relevant comorbid conditions in 5-year intervals after the MADS assessment for depression throughout the patients' life span (Multimedia Appendix 2), allowing for a comprehensive examination of multimorbidity progression over time.

Multimedia Appendix 2 shows the current disease prevalences expressed in percentages and the incidence of new disease onsets across an interval of 5 years after the MADS assessment expressed in per thousand. Multimedia Appendix 2 also showcases the results for MDD and 9 mental and somatic MDD-related (PR>0.80) chronic conditions assessed independently in the 3 study cohorts, namely, CHSS, THL, and UKB, and in 4 time points, that is, ages of 20 years, 40 years, 60 years, and 70 years, corresponding to the intervals in which the PRs were recalculated. A continuous assessment of these outcomes is reported in Multimedia Appendix 1.

In general, both MDD and the comorbid conditions investigated in this study exhibited a positive correlation between the MADS risk tiers and the current prevalence and incidence of new disease onsets within a subsequent 5-year interval. This is evident from the table. Notably, the highest disease prevalence and incidence values consistently appeared in the high- and very-high-risk tiers. In addition, there was a discernible pattern of well-stratified values across these risk tiers within the same age ranges, underlining significantly elevated prevalence rates of the studied diseases compared to the population average within the high-risk groups. Age also emerged as a pivotal determinant influencing disease onset, delineating unique patterns across various disorders. Notably, conditions such as gastroesophageal reflux and overweight consistently exhibited ascending trends in both incidence and prevalence throughout individuals' life spans. Conversely, severe afflictions such as schizophrenia, bipolar disorder, and alcohol abuse reached their zenith in prevalence and incidence during middle-aged adulthood followed by a decline, possibly indicating an association with premature mortality. Moreover, anxiety- and stress-related disorders showed their highest incidence rates during youth and early adulthood.

XSL•FO
RenderX

The consistency of the findings illustrated in Multimedia Appendix 2 remained robust across all 3 study cohorts despite their significant demographic differences, described in Table 1. These heterogeneities resulted in disease prevalence discrepancies among cohorts, as vividly portrayed in Multimedia Appendix 2. Among the most relevant cases, there was an elevated prevalence of schizophrenia in the THL cohort in comparison with the CHSS and UKB cohorts. In this particular case, patients with schizophrenia constituted 100% of the very high–risk group in adulthood. Such differences in disease prevalence among cohorts may influence distinct health outcomes, particularly for the citizens allocated to the apex of the Finnish risk pyramid, as observed in the pharmacological and hospitalization expenditure outcomes reported in Table 3.

Discussion

Principal Findings

The MADS seems to provide a novel and more comprehensive understanding of the complex nature of depression-related multimorbidity. This approach recognizes that individuals with depression often experience a range of comorbid conditions that may manifest and evolve differently over time. By capturing this dynamic aspect, the MADS offers a nuanced assessment beyond a mere checklist of discrete disorders. The novelty of the MADS approach lies in its capability to serve as the first morbidity grouper that incorporates information on disease trajectories while improving the filtering of indirect disease associations using BDMMs.

In addition to capturing disease-disease associations, the MADS endeavors to gauge their impact within the system by leveraging well-established DWs. However, despite achieving success in fulfilling the study's objectives, it is crucial to acknowledge that this approach carries inherent limitations, as will be elaborated on in the subsequent sections of this discussion.

In this investigation, we unearthed robust correlations between the MADS risk strata and the extent of deleterious impact caused by MDD and its comorbid conditions. Such associations indicate the presence of specific health risks and an escalated use of health care resources. Furthermore, a positive association emerged between the levels of pharmacological and nonpharmacological health care expenditures and the different tiers of MADS risk. In addition, the analysis revealed an augmented risk of disease progression within the high-risk groups (high and very high risk), as indicated by a heightened incidence of new-onset depression-related illnesses within a 12-month period after the MADS assessment. Similarly, mortality rates exhibited elevated values in these high-risk groups.

The findings presented in this study are underpinned by the complementary studies conducted within the TRAJECTOME project [30] that have established a better understanding of the complex multimorbidity landscape associated with MDD across an individual's life span, encompassing modifiable and genetic risk factors.

Limitations of This Approach

Despite meeting expectations and validating the hypothesis through which the study was conceived, the authors acknowledge a series of limitations leading to suboptimal results and limited potential for adaptation and generalization that should be undertaken to bring the MADS, or an indicator derived from it, to short-term real-world implementation.

In this research, the use of estimations of mean DW [44] to assess the burden of disease conditions achieved desirable results and was conceptually justified, but it undoubtedly exhibited significant limitations. In an ideal clinical scenario, each disease diagnosis indicated in the patient's electronic medical record should be characterized by three key dimensions: (1) severity of the diagnosis, (2) rate of disease progression, and (3) impact on disability. However, the degree of maturity for characterizing the last 2 dimensions—disease progression and disability—is rather poor because of the complexities involved in their assessment. In other words, the authors acknowledge the weakness associated with the current use of DW. However, they stress the importance of incorporating such dimensions in future evolutions of the MADS.

A noteworthy aspect that should be acknowledged is that factors such as the advancements in diagnostic techniques, the digitization of medical records, and the modifications in disease taxonomy and classification over time have contributed to a more exhaustive documentation of the disease states in the most recent health records. Consequently, this fact could lead to imprecisions in estimating the disease onset ages in older individuals.

Insights and Potential Impact of the MADS in Multimorbidity Management

The results reported in this study not only reaffirm the well-established link between multimorbidity and adverse outcomes, such as a decline in functional status, compromised quality of life, and increased mortality rates [45], but also shed light on the significant burden imposed on individuals and health care systems. From the population-based HRA perspective, the strain on resource allocation and overall health care spending is a pressing concern that necessitates effective strategies for addressing and managing multimorbidity [46]. In this context, assessing individual health risks and patient stratification emerge as crucial approaches that enable the implementation of predictive and preventive measures in health care.

While population-based HRA tools such as Adjusted Clinical Groups, Clinical Risk Groups, or AMG have traditionally addressed this aspect, the MADS is designed to complement rather than replace those tools. This study aimed to test a method to refine existing HRA tools by aligning them with the principles of network medicine, thereby merging traditional HRA with the practical application of network medicine insights. This innovative approach holds the promise of unlocking new potential advantages and capabilities.

The strength of the MADS approach lies in using disease-disease associations drawn from the analysis of temporal occurrence patterns among concurrent diseases. This virtue allows the MADS to refine the analysis of the morbidity burden by

XSL•FO
RenderX

focusing on clusters of correlated diseases, which in turn can aid in developing more tailored epidemiological risk-related studies. This refined analysis might also assist in resource allocation and inform health care policies for targeted patient groups with specific needs. Moreover, this approach holds promise for potential extrapolation to other noncommunicable disease clusters such as diabetes, cardiovascular ailments, respiratory diseases, or cancer. By leveraging this targeted approach, the MADS can be adapted to other disease clusters with shared characteristics, enabling a more precise assessment of disease burden and comorbidity patterns and thereby generating multiple disease-specific indexes.

Notably, when considering information derived from disease co-occurrence patterns, the presence or absence of certain diseases seems to correlate with the risk of developing related comorbid conditions, as elucidated in Multimedia Appendix 2. This highlights the potential for a nuanced understanding of disease relationships and their impacts on health outcomes and to implement preventive interventions to mitigate their effect. Moreover, the findings of this study highlight the potential of preventive strategies targeted at mental disorders, including substance abuse disorders, depressive disorders, and schizophrenia, to reduce the incidence of negative clinical outcomes in somatic health conditions. These important implications for clinical practice call for a comprehensive and interdisciplinary approach that bridges the gap between psychiatric and somatic medicine. By developing cross-specialty preventive strategies, health care professionals can provide more holistic and effective care for individuals with complex health needs, ensuring that their mental and physical health are adequately addressed [47].

This study provided good prospects of using disease trajectories to enhance the performance of existing state-of-the-art morbidity groupers such as AMG. Recognized for its transferability across EU regions by the EU Joint Action on implementation of digitally enabled integrated person-centered care [48], AMG stands out due to its stratification capabilities, adaptability, and distribution as open-source software, providing several advantages over its commercial counterparts. The AMG system uses disease-specific weighting derived from statistical analysis incorporating mortality and health care service use data. This method addresses the primary drawback identified in the MADS approach inherent to the use of DW while enabling the development of adaptable tools that align with the unique characteristics of each health care system. Consequently, it allows for the adjustment to the impact of specific disease conditions within distinct regions and enhances the overall applicability and adaptability of the tool. In this regard, this study offered promising insights aligned with the developers' envisioned future features for integration into the AMG system. Serving as a proof of concept, it highlighted the potential improvements achievable within AMG by leveraging disease-disease associations, thereby shaping the road map for further AMG development.

MADS Integration in Precision Medicine: Advancing Toward Patient-Centric Strategies

By assessing whether the MADS is appropriate for the stratification of depression-related multimorbidity, we attempted to confirm its potential for contributing to precision medicine [49]. In the clinical arena, identifying individuals at elevated risk and customizing interventions enable health care providers to intervene proactively, potentially preventing or lessening disease progression and enhancing patient outcomes. These strategies not only yield immediate value in terms of improved patient care but also lay the foundation for the broader adoption of integrated care and precision medicine, particularly in the management of chronic conditions [50].

Incorporating systems medicine [51] methodologies and ITs has prompted significant shifts in clinical research and practice, paving the way for holistic approaches, computational modeling, and predictive tools in clinical medicine. These advancements are driving the adoption of clinical decision support systems, which use patient-specific data to generate assessments or recommendations, aiding clinicians in making informed decisions. It is well established that, to improve predictive precision and aid clinical decision-making, implementing comprehensive methodologies that consider various influencing factors from multiple sources in patient health could enhance individual prognosis estimations [52].

This integration might facilitate predictive modeling methodologies for personalized risk prediction and intervention planning. This approach, known as multisource clinical predictive modeling [53,54], enables the integration of (1) health care data and health determinants from other domains, including (2) population health registry data; (3) informal care data (including patients' self-tracking data, lifestyles, environmental and behavioral aspects, and sensors); and, ideally, (4) biomedical research omics data. In this paradigm, it is crucial to acknowledge the pivotal role that multimorbidity groupers play in capturing the clinical complexity of individuals. Previous research [53,54] has highlighted the synergy between patient clinical complexity (eg, AMG) and acute episode severity, correlating with higher risks of adverse health events. This opens avenues for further research, exploring how adjusted morbidity indicators such as the MADS can significantly contribute to predictive modeling, aiming at supporting the implementation of cost-effective, patient-centered preventive measures to manage patients with chronic diseases and potentially delay or prevent their progression to the highest-risk levels in the stratification pyramid [55].

Conclusions

The MADS showed to be a promising approach to estimate multimorbidity-adjusted risk of disease progression and measure MDD's impact on individuals and health care systems, which could be tested in other diseases. The novelty of the MADS approach lies in its unique capability to incorporate disease trajectories, providing a comprehensive understanding of depression-related morbidity burden. In this regard, the BDMM method played a crucial role in isolating and identifying true direct disease associations. The results of this study pave the way for the development of innovative digital tools to support

XSL•FO
RenderX

advanced HRA strategies. Nevertheless, clinical validation is imperative before considering the widespread adoption of the MADS.

Acknowledgments

This initiative was supported by European Research Area on Personalized Medicine (ERA PerMed) program ("Temporal disease map based stratification of depression-related multimorbidities: towards quantitative investigations of patient trajectories and predictions of multi-target drug candidates" [TRAJECTOME] project; ERAPERMED2019-108). Locally, this study was supported by the Academy of Finland under the frame of the ERA PerMed program and the Hungarian National Research, Development, and Innovation Office (2019-2.1.7-ERA-NET-2020-00005K143391, K139330 and PD 134449 grants); the Hungarian Brain Research Program 3.0 (NAP2022-I-4/2022); and the Ministry for Innovation and Technology of Hungary from the National Research, Development, and Innovation Fund under the TKP2021-EGA funding scheme (TKP2021-EGA-25 and TKP2021-EGA-02). This study was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. The authors want to acknowledge the earnest collaboration of the Digitalization for the Sustainability of the Healthcare System research group at Institut d'Investigació Biomèdica de Bellvitge (IDIBELL) for their support in the preparation of the Catalan cohort, which was extracted from the Catalan Health Surveillance System database, owned and managed by the Catalan Health Service. In addition, the authors want to acknowledge the participants and investigators of the FinnGen study and CSC–IT Center for Science, Finland, for computational resources. This research was conducted using the UK Biobank resource under application 1602. Linked health data Copyright 2019, NHS England. Reused with the permission of the UK Biobank. All rights reserved.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to patient privacy concerns. The scripts used to compute the Multimorbidity-Adjusted Disability Score are available from the corresponding author upon reasonable request.

Authors' Contributions

PA, GJ, and IC designed the study and directed the project. RG-C, KM, and IC led the design of the Multimorbidity-Adjusted Disability Score. RG-C, KM, AG, and TP executed the quantitative analysis, processed the experimental data, performed the statistical analysis, and created the figures. EV generated the Catalan Health Surveillance System database and provided statistical support. ZG, GH, HM, TN, MK, JPJ, and JR provided insightful information to the study. The manuscript was first drafted by RGC, IC, and JR and thoroughly revised by KM, EV, AG, TP, ZG, GH, HM, TN, MK, JP-J, PA, and GJ. All authors approved the final version of the manuscript and are accountable for all aspects of the work in ensuring its accuracy and integrity.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary material encompassing the tables and figures from the cross-sectional and longitudinal analyses of outcomes, along with the disability weights and probabilities of relevance, as well as the Multimorbidity-Adjusted Disability Score (MADS) pseudocodes.

[DOCX File , 2119 KB - jmir_v26i1e53162_app1.docx]

Multimedia Appendix 2

Longitudinal analysis of disease prevalence and incidence of new disease onsets in the Catalan Health Surveillance System, UK Biobank, and Finnish Institute for Health and Welfare.

[DOCX File , 47 KB - jmir_v26i1e53162_app2.docx]

References

1. Vetrano DL, Calderón-Larrañaga A, Marengoni A, Onder G, Bauer JM, Cesari M, et al. An international perspective on chronic multimorbidity: approaching the elephant in the room. J Gerontol A Biol Sci Med Sci 2018 Sep 11;73(10):1350-1356 [FREE Full text] [doi: 10.1093/gerona/glx178] [Medline: 28957993]
2. Salive ME. Multimorbidity in older adults. Epidemiol Rev 2013 Jan 31;35(1):75-83. [doi: 10.1093/epirev/mxs009] [Medline: 23372025]

XSL•FO
RenderX

3. Garin N, Olaya B, Moneta MV, Miret M, Lobo A, Ayuso-Mateos JL, et al. Impact of multimorbidity on disability and quality of life in the Spanish older population. PLoS One 2014 Nov 6;9(11):e111498 [FREE Full text] [doi: 10.1371/journal.pone.0111498] [Medline: 25375890]
4. Monterde D, Vela E, Clèries M, Garcia-Eroles L, Roca J, Pérez-Sust P. Multimorbidity as a predictor of health service utilization in primary care: a registry-based study of the Catalan population. BMC Fam Pract 2020 Feb 17;21(1):39 [FREE Full text] [doi: 10.1186/s12875-020-01104-1] [Medline: 32066377]
5. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. Ann Fam Med 2009;7(4):357-363 [FREE Full text] [doi: 10.1370/afm.983] [Medline: 19597174]
6. Rahman MH, Rana HK, Peng S, Kibria MG, Islam MZ, Mahmud SM, et al. Bioinformatics and system biology approaches to identify pathophysiological impact of COVID-19 to the progression and severity of neurological diseases. Comput Biol Med 2021 Nov;138:104859 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104859] [Medline: 34601390]
7. Taraschi A, Cimini C, Colosimo A, Ramal-Sanchez M, Moussa F, Mokh S, et al. Human immune system diseasome networks and female oviductal microenvironment: new horizons to be discovered. Front Genet 2021 Jan 27;12:795123 [FREE Full text] [doi: 10.3389/fgene.2021.795123] [Medline: 35154249]
8. Chauhan PK, Sowdhamini R. Integrative network analysis interweaves the missing links in cardiomyopathy diseasome. Sci Rep 2022 Nov 16;12(1):19670 [FREE Full text] [doi: 10.1038/s41598-022-24246-x] [Medline: 36385157]
9. Calderón-Larrañaga A, Vetrano DL, Ferrucci L, Mercer SW, Marengoni A, Onder G, et al. Multimorbidity and functional impairment-bidirectional interplay, synergistic effects and common pathways. J Intern Med 2019 Mar;285(3):255-271 [FREE Full text] [doi: 10.1111/joim.12843] [Medline: 30357990]
10. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms-disease network. Nat Commun 2014 Jun 26;5(1):4212 [FREE Full text] [doi: 10.1038/ncomms5212] [Medline: 24967666]
11. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 2015 Feb 20;347(6224):1257601 [FREE Full text] [doi: 10.1126/science.1257601] [Medline: 25700523]
12. Goh K, Choi I. Exploring the human diseasome: the human disease network. Brief Funct Genomics 2012 Nov 12;11(6):533-542. [doi: 10.1093/bfgp/els032] [Medline: 23063808]
13. Barabási AL. Network Medicine — from obesity to the “diseasome”. N Engl J Med 2007 Jul 26;357(4):404-407. [doi: 10.1056/nejme078114]
14. Murray SA, Kendall M, Boyd K, Sheikh A. Illness trajectories and palliative care. BMJ 2005 Apr 30;330(7498):1007-1011 [FREE Full text] [doi: 10.1136/bmj.330.7498.1007] [Medline: 15860828]
15. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun 2014 Jun 24;5(1):4022 [FREE Full text] [doi: 10.1038/ncomms5022] [Medline: 24959948]
16. Tinetti ME, Fried TR, Boyd CM. Designing health care for the most common chronic condition--multimorbidity. JAMA 2012 Jun 20;307(23):2493-2494 [FREE Full text] [doi: 10.1001/jama.2012.5265] [Medline: 22797447]
17. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
18. Parkerson GR, Broadhead WE, Tse CK. The Duke Severity of Illness Checklist (DUSOI) for measurement of severity and comorbidity. J Clin Epidemiol 1993 Apr;46(4):379-393. [doi: 10.1016/0895-4356(93)90153-r] [Medline: 8483003]
19. Charlson ME, Pompei P, Ales KL, MacKenzie C. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987 Jan;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
20. Starfield B, Weiner J, Mumford L, Steinwachs D. Ambulatory care groups: a categorization of diagnoses for research and management. Health Serv Res 1991 Apr;26(1):53-74 [FREE Full text] [Medline: 1901841]
21. Monterde D, Vela E, Clèries M, grupo colaborativo GMA. [Adjusted morbidity groups: a new multiple morbidity measurement of use in Primary Care]. Aten Primaria 2016 Dec;48(10):674-682 [FREE Full text] [doi: 10.1016/j.aprim.2016.06.003] [Medline: 27495004]
22. Steffen A, Nübel J, Jacobi F, Bätzing J, Holstiege J. Mental and somatic comorbidity of depression: a comprehensive cross-sectional analysis of 202 diagnosis groups using German nationwide ambulatory claims data. BMC Psychiatry 2020 Mar 30;20(1):142 [FREE Full text] [doi: 10.1186/s12888-020-02546-8] [Medline: 32228541]
23. Cerezo-Cerezo J, López CA. Good practice brief - population stratification: a fundamental instrument used for population health management in Spain. World Health Organization. 2018. URL: https://iris.who.int/bitstream/handle/10665/345586/WHO-EURO-2018-3032-42790-59709-eng.pdf?sequence=3&isAllowed=y [accessed 2024-04-29]
24. Calling all ACG system users!. Johns Hopkins ACG® System. URL: https://www.hopkinsacg.org/ [accessed 2023-08-29]
25. 3M clinical risk groups (CRGs). 3M. URL: https://www.3m.com/3M/en_US/health-information-systems-us/drive-value-based-care/patient-classification-methodologies/crgs/ [accessed 2023-08-29]
26. Uhlig K, Leff B, Kent D, Dy S, Brunnhuber K, Burgers JS, et al. A framework for crafting clinical practice guidelines that are relevant to the care and management of people with multimorbidity. J Gen Intern Med 2014 Apr 18;29(4):670-679 [FREE Full text] [doi: 10.1007/s11606-013-2659-y] [Medline: 24442332]

XSL•FO
RenderX

27. Marx P, Antal P, Bolgar B, Bagdy G, Deakin B, Juhasz G. Comorbidities in the Diseasome are more apparent than real: what Bayesian filtering reveals about the comorbidities of depression. PLoS Comput Biol 2017 Jun 23;13(6):e1005487 [FREE Full text] [doi: 10.1371/journal.pcbi.1005487] [Medline: 28644851]
28. Bolgár B, Antal P. VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization. BMC Bioinformatics 2017 Oct 04;18(1):440 [FREE Full text] [doi: 10.1186/s12859-017-1845-z] [Medline: 28978313]
29. ICD-10-CM international classification of diseases, tenth revision, clinical modification (ICD-10-CM). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/icd/icd10cm.htm [accessed 2021-07-21]
30. Juhasz G, Gezsi A, Van der Auwera S, Mäkinen H, Eszlari N, Hullam G, et al. Unique genetic and risk-factor proles in multimorbidity clusters of depression-related disease trajectories from a study of 1.2 million subjects. Research Square. Preprint posted online August 2, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3199113/v1]
31. Temporal disease map-based stratification of depression-related multimorbidities: towards quantitative investigations of patient trajectories and predictions of multi-target drug candidates. TRAJECTOME project. URL: https://semmelweis.hu/trajectome/en/ [accessed 2023-11-18]
32. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
33. Farré N, Vela E, Clèries M, Bustins M, Cainzos-Achirica M, Enjuanes C, et al. Medical resource use and expenditure in patients with chronic heart failure: a population-based analysis of 88 195 patients. Eur J Heart Fail 2016 Sep 25;18(9):1132-1140 [FREE Full text] [doi: 10.1002/ejhf.549] [Medline: 27108481]
34. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015 Mar;12(3):e1001779 [FREE Full text] [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
35. Biobank. Finnish Institute for Health and Welfare. URL: https://thl.fi/en/web/thl-biobank [accessed 2023-04-19]
36. Borodulin K, Tolonen H, Jousilahti P, Jula A, Juolevi A, Koskinen S, et al. Cohort profile: the National FINRISK Study. Int J Epidemiol 2018 Jun 01;47(3):696. [doi: 10.1093/ije/dyx239] [Medline: 29165699]
37. Borodulin K, Sääksjärvi K. FinHealth 2017 study: methods. Finnish Institute for Health and Welfare. 2019. URL: https://www.julkari.fi/bitstream/handle/10024/139084/URN_ISBN_978-952-343-449-3.pdf [accessed 2024-04-29]
38. Heistaro S. Methodology report: health 2000 survey. Kansanterveyslaitos. 2008. URL: https://www.julkari.fi/bitstream/handle/10024/78185/2008b26.pdf?sequence=1&isAllowed=y [accessed 2024-04-29]
39. Hilderink HB, Plasmans MH, Snijders BE, Boshuizen HC, Poos MJ, van Gool CH. Accounting for multimorbidity can affect the estimation of the burden of disease: a comparison of approaches. Arch Public Health 2016 Aug 22;74(1):37 [FREE Full text] [doi: 10.1186/s13690-016-0147-7] [Medline: 27551405]
40. Haagsma JA, van Beeck EF, Polinder S, Toet H, Panneman M, Bonsel GJ. The effect of comorbidity on health-related quality of life for injury patients in the first year following injury: comparison of three comorbidity adjustment approaches. Popul Health Metr 2011 Apr 24;9:10 [FREE Full text] [doi: 10.1186/1478-7954-9-10] [Medline: 21513572]
41. Guidelines for ATC classification and DDD assignment. WHO Collaborating Centre for Drug Statistics Methodology. 2021. URL: https://atcddd.fhi.no/atc_ddd_index_and_guidelines/guidelines/ [accessed 2024-04-29]
42. CatSalut. Catalan Health Service. URL: https://catsalut.gencat.cat/ca/inici/ [accessed 2023-01-21]
43. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.r-project.org/ [accessed 2024-04-29]
44. Haagsma JA, Polinder S, Cassini A, Colzani E, Havelaar AH. Review of disability weight studies: comparison of methodological choices and values. Popul Health Metr 2014 Aug 23;12(1):20 [FREE Full text] [doi: 10.1186/s12963-014-0020-2] [Medline: 26019690]
45. Makovski TT, Schmitz S, Zeegers MP, Stranges S, van den Akker M. Multimorbidity and quality of life: systematic literature review and meta-analysis. Ageing Res Rev 2019 Aug;53:100903. [doi: 10.1016/j.arr.2019.04.005] [Medline: 31048032]
46. Yamanashi H, Nobusue K, Nonaka F, Honda Y, Shimizu Y, Akabame S, et al. The role of mental disease on the association between multimorbidity and medical expenditure. Fam Pract 2020 Sep 05;37(4):453-458. [doi: 10.1093/fampra/cmaa015] [Medline: 32086514]
47. Dragioti E, Radua J, Solmi M, Gosling CJ, Oliver D, Lascialfari F, et al. Impact of mental disorders on clinical outcomes of physical diseases: an umbrella review assessing population attributable fraction and generalized impact fraction. World Psychiatry 2023 Feb 14;22(1):86-104 [FREE Full text] [doi: 10.1002/wps.21068] [Medline: 36640414]
48. Joint action on implementation of digitally enabled integrated person-centred care. JADECARE project. URL: https://www.jadecare.eu/ [accessed 2024-04-29]
49. Musker M. Treating depression in the era of precision medicine: challenges and perspectives. In: Quevedo J, Carvalho AF, Zarate CA, editors. Neurobiology of Depression: Road to Novel Therapeutics. Cambridge, MA: Academic Press; 2019:265-275.

XSL•FO
RenderX

50. Dueñas-Espín I, Vela E, Pauws S, Bescos C, Cano I, Cleries M, et al. Proposals for enhanced health risk assessment and stratification in an integrated care scenario. BMJ Open 2016 Apr 15;6(4):e010301 [FREE Full text] [doi: 10.1136/bmjopen-2015-010301] [Medline: 27084274]
51. Federoff HJ, Gostin LO. Evolving from reductionism to holism: is there a future for systems medicine? JAMA 2009 Sep 02;302(9):994-996. [doi: 10.1001/jama.2009.1264] [Medline: 19724047]
52. Cano I, Tenyi A, Vela E, Miralles F, Roca J. Perspectives on big data applications of health information. Curr Opin Syst Biol 2017 Jun;3:36-42. [doi: 10.1016/j.coisb.2017.04.012]
53. Calvo M, González R, Seijas N, Vela E, Hernández C, Batiste G, et al. Health outcomes from home hospitalization: multisource predictive modeling. J Med Internet Res 2020 Oct 07;22(10):e21367 [FREE Full text] [doi: 10.2196/21367] [Medline: 33026357]
54. González-Colom R, Herranz C, Vela E, Monterde D, Contel JC, Sisó-Almirall A, et al. Prevention of unplanned hospital admissions in multimorbid patients using computational modeling: observational retrospective cohort study. J Med Internet Res 2023 Feb 16;25:e40846 [FREE Full text] [doi: 10.2196/40846] [Medline: 36795471]
55. Baltaxe E. Population-based analysis of COPD patients in Catalonia: implications for case management. Eur Respir J 2017;50(suppl 61):PA4956. [doi: 10.1183/1393003.congress-2017.pa4956]

Abbreviations

AMG: Adjusted Morbidity Groups
BDMM: Bayesian direct multimorbidity map
CHSS: Catalan Health Surveillance System
DW: disability weight
EU: European Union
GBD: Global Burden of Disease
ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification
MADS: Multimorbidity-Adjusted Disability Score
MDD: major depressive disorder
PR: probability of relevance
THL: Finnish Institute for Health and Welfare
UKB: UK Biobank

Edited by A Mavragani; submitted 27.09.23; peer-reviewed by R Meng, C Doucet; comments to author 02.11.23; revised version received 23.11.23; accepted 23.05.24; published 24.06.24.

Please cite as:
González-Colom R, Mitra K, Vela E, Gezsi A, Paajanen T, Gál Z, Hullam G, Mäkinen H, Nagy T, Kuokkanen M, Piera-Jiménez J, Roca J, Antal P, Juhasz G, Cano I
Multicentric Assessment of a Multimorbidity-Adjusted Disability Score to Stratify Depression-Related Risks Using Temporal Disease Maps: Instrument Validation Study
J Med Internet Res 2024;26:e53162
URL: https://www.jmir.org/2024/1/e53162
doi:10.2196/53162
PMID:38913991

©Rubèn González-Colom, Kangkana Mitra, Emili Vela, Andras Gezsi, Teemu Paajanen, Zsófia Gál, Gabor Hullam, Hannu Mäkinen, Tamas Nagy, Mikko Kuokkanen, Jordi Piera-Jiménez, Josep Roca, Peter Antal, Gabriella Juhasz, Isaac Cano. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Potential of Evidence-Based Clinical Intake Tools to Discover or Ground Prevalence of Symptoms Using Real-Life Digital Health Encounters: Retrospective Cohort Study

Eden Avnat[1,2], BMedSci, MPH; Michael Samin[1,3], MD; Daniel Ben Joya[1,3], MD; Eyal Schneider[1,4], MD; Elia Yanko[5], BSc; Dafna Eshel[1], MD; Shahar Ovadia[1], MCS; Yossi Lev[1], BCS; Daniel Souroujon[1,2], MPH, MD

[1]Kahun Medical Ltd, Givatayim, Israel

[2]Faculty of Medicine, Tel-Aviv University, Tel Aviv, Israel

[3]Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel

[4]Department of Otolaryngology-Head and Neck Surgery, Samson Assuta Ashdod University Hospital, Ben Gurion University Faculty of Health Sciences, Ashdod, Israel

[5]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel

Corresponding Author:
Eden Avnat, BMedSci, MPH
Kahun Medical Ltd
4 Ariel Sharon St
Givatayim, 515909117
Israel
Phone: 972 545299622
Email: eden@kahun.com

Abstract

Background: Evidence-based clinical intake tools (EBCITs) are structured assessment tools used to gather information about patients and help health care providers make informed decisions. The growing demand for personalized medicine, along with the big data revolution, has rendered EBCITs a promising solution. EBCITs have the potential to provide comprehensive and individualized assessments of symptoms, enabling accurate diagnosis, while contributing to the grounding of medical care.

Objective: This work aims to examine whether EBCITs cover data concerning disorders and symptoms to a similar extent as physicians, and thus can reliably address medical conditions in clinical settings. We also explore the potential of EBCITs to discover and ground the real prevalence of symptoms in different disorders thereby expanding medical knowledge and further supporting medical diagnoses made by physicians.

Methods: Between August 1, 2022, and January 15, 2023, patients who used the services of a digital health care (DH) provider in the United States were first assessed by the Kahun EBCIT. Kahun platform gathered and analyzed the information from the sessions. This study estimated the prevalence of patients' symptoms in medical disorders using 2 data sets. The first data set analyzed symptom prevalence, as determined by Kahun's knowledge engine. The second data set analyzed symptom prevalence, relying solely on data from the DH patients gathered by Kahun. The variance difference between these 2 prevalence data sets helped us assess Kahun's ability to incorporate new data while integrating existing knowledge. To analyze the comprehensiveness of Kahun's knowledge engine, we compared how well it covers weighted data for the symptoms and disorders found in the 2019 National Ambulatory Medical Care Survey (NMCAS). To assess Kahun's diagnosis accuracy, physicians independently diagnosed 250 of Kahun-DH's sessions. Their diagnoses were compared with Kahun's diagnoses.

Results: In this study, 2550 patients used Kahun to complete a full assessment. Kahun proposed 108,523 suggestions related to symptoms during the intake process. At the end of the intake process, 6496 conditions were presented to the caregiver. Kahun covered 94% (526,157,569/562,150,572) of the weighted symptoms and 91% (1,582,637,476/173,4783,244) of the weighted disorders in the 2019 NMCAS. In 90% (224/250) of the sessions, both physicians and Kahun suggested at least one identical disorder, with a 72% (367/507) total accuracy rate. Kahun's engine yielded 519 prevalences while the Kahun-DH cohort yielded 599; 156 prevalences were unique to the latter and 443 prevalences were shared by both data sets.

Conclusions: ECBITs, such as Kahun, encompass extensive amounts of knowledge and could serve as a reliable database for inferring medical insights and diagnoses. Using this credible database, the potential prevalence of symptoms in different disorders

XSL•FO
RenderX

was discovered or grounded. This highlights the ability of ECBITs to refine the understanding of relationships between disorders and symptoms, which further supports physicians in medical diagnosis.

(J Med Internet Res 2024;26:e49570) doi:10.2196/49570

KEYWORDS

clinical intake tool; evidence-based medicine; big data; digital health; symptoms; prevalence

Introduction

Evidence-based medicine (EBM) has become an integral part of modern medical practice [1]. It relies on the use of systematic research to identify up-to-date and reliable data, which serves as a crucial component in making medical decisions for individual patients [2]. The use of EBM can help reduce outdated practices or implicit biases that may disrupt medical decision-making such as prioritizing White individuals over minorities in emergency care or underestimating diagnosis of coronary heart disease in females [3-5].

The emergence of the big data revolution in the medical world not only reduces some of the bias and keeps physicians updated, but also serves as an important factor in the evolution of EBM [6-8]. By using machine learning tools, implicit insights can be more easily derived from raw data, and instead of focusing on data analysis as a tool to answer questions, it is now used as a tool to find new questions which can lead to new and promising hypotheses [7].

In particular, personalized medicine has benefited from the analysis of abundant and diverse data [9]. Analyzing multi-omics data derived from large-scale cohort and population studies, combined with the study's conclusions, allows for identifying subtle differences in an individual's genetics which may lead to precise and personalized interventions [9].

Today, as the demand for personalized medicine based on the EBM approach is on the rise, health providers are seeking to bridge the gap between the EBM paradigm which formulates generalized conclusions gathered from many studies, and personalized medicine that focuses on the individual [10,11]. One solution that bridges that gap suggests that using evidence-based clinical intake tools (EBCITs), such as Kahun [11,12], can help physicians with more personalized decision-making.

Kahun is an artificial intelligence engine, encompassing more than 50,000 peer-reviewed publications and more than 30,000,000 medical relations and insights that were mapped to a knowledge graph [13]. Using its evidence-based knowledge graph, Kahun asks the patient personalized questions and then uses dynamic reasoning to generate tailored clinical assessments [12,13].

Prior studies have focused primarily on the diagnostic accuracy of Kahun and other diagnostic support tools [12,14-16]. However, one study [17] that examined COVID-19 patients and their specific symptoms showed that these tools can also provide medical insights such as extending the understanding of symptoms associated with a specific disease such as COVID-19. By collecting patient findings and diagnoses, EBCITs have the potential to become comprehensive real-world databases in their own right. Analyzing these databases using big data methods could contribute to medical knowledge by grounding the real prevalence of symptoms in medical conditions or identifying new ones while supporting the EBM approach.

This study aims to test whether EBCITs cover symptoms and diagnoses similar to physicians and thus serve as a potential reliable tool for decision-making in the clinic. Furthermore, this study demonstrated the ability of EBCITs to identify current or emerging prevalence of symptoms in disorders, thereby enhancing medical knowledge and providing additional support for physicians in clinical practice.

Methods

Study Design and Data Collection

This study includes the analysis of 3 data sets.

Kahun's Knowledge Engine

As mentioned, Kahun's data are based on more than 30,000,000 medical relations and insights from over 50,000 peer-reviewed publications that were mapped by medical experts to a knowledge graph [13]. The medical knowledge is represented by the nodes and the edges of the knowledge graph. For example, node1 represents a specific disorder, which is mapped by edge1 to node2, which represents a specific symptom. The data on edge1 represents the relationship between node1 and node2 such as prevalence. Therefore, the triple (node1, edge1, and node2) represents the prevalence of the specific symptom in the specific disorder. These data are referred to here as Kahun's knowledge engine.

Kahun-Digital Health Care Cohort

This data set was collected by the Kahun platform which assessed patients who received digital health care (DH) services from a DH provider based in the United States, specializing in doctor-patient medical visits that occur using camera-enabled smartphones or computers, between August 1, 2022, and January 15, 2023.

Patients who used the provider's services were first given a link for initial assessment by Kahun. During the Kahun assessment, questions regarding the patient's medical background, chief complaint, symptoms, and risk factors were personally generated based on the Kahun algorithms [12]. Additionally, during each assessment, the algorithm suggested a dynamic list of relevant differential diagnoses (disorders) while computing a matching probability between the findings and each disorder. All the data and metadata regarding questions, answers, and disorders were anonymized and stored in a separate part of the Kahun database,

XSL•FO
RenderX

which is not related to Kahun's knowledge engine and is referred to here as the Kahun-DH cohort.

National Ambulatory Medical Care Survey

The data were obtained from the latest (2019) National Ambulatory Medical Care Survey (NAMCS), which was conducted by the Centers for Disease Control and Prevention [18]. NAMCS 2019 was designed to provide objective information about the provision and use of ambulatory medical care services in the United States [18]. The findings in NAMCS are based on a sample of weighted visits to nonfederally employed office-based physicians, who are primarily engaged in direct patient care. The findings included "reason of visit #1-5" (coded by NAMCS internal method), "diagnosis #1-5" (*ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*] coded), and more. Since the scope of this study covered symptoms and disorders, we addressed only NAMCS symptoms related to the "SYMPTOM MODULE" and only those disorders that appeared among the NAMCS diagnoses. Findings were excluded where the reason for the visit was related to follow-up visits and the prescription of medication or diagnoses related to an encounter for a specific examination.

Study Population

Kahun-DH cohort included patients aged 16 years or older who used the provider's services and completed the assessment by Kahun. Assessments missing information regarding sex, chief complaint, and differential diagnosis were excluded.

Analysis and Variables

All statistical analyses were performed using RStudio (R version 4.2.2; Posit PBC). Categorical variables were represented by percentage while continuous variables were represented by mean and SD values if distributed normally, and otherwise median and IQR values. A "positive symptoms ratio" was determined as the ratio between the total number of new symptoms or refinements of known symptoms the patient confirmed during the assessment, and the total number of symptoms or refinements suggested to the patient during the assessment. "Total questions" included only those questions that were answered by the patients during the assessment.

Kahun's Coverage Rates

We calculated how well Kahun covers NAMCS data by multiplying the ratio of symptoms or disorders that were reported in each visit and appeared in Kahun's knowledge engine, with the relevant visit weighted score designated by NAMCS. All multiples were summed and then divided by the summed total of all the visits' weighted scores. We calculated the coverage ratio for the entire set of symptoms and disorders, for the groups of symptoms subset by NAMCS's "SYMPTOM MODULE," and for groups of disorders classified according to *ICD-10* prefixes.

Kahun's Diagnostic Ability

To evaluate Kahun's diagnostic ability, 250 sessions (10% of all sessions) were randomly selected and blindly assessed by Israeli licensed physicians. The full transcript of Kahun's sessions, including questions suggested by Kahun and answers provided by the patient, was given to the physicians. The transcript did not include Kahun's suggestions for differential diagnosis. Then, based on the transcript received, the physicians suggested up to 3 (nonordered) suitable diagnoses. The physicians' suggested diagnoses were compared with both the relevant diagnoses and the relevant number of diagnoses suggested by Kahun (eg, if only 2 diagnoses were suggested by the physicians, only 2 diagnoses suggested by Kahun were compared). Kahun's accuracy rate was set as the number of matched diagnoses suggested by the physicians and Kahun divided by the total number of diagnoses suggested by Kahun (eg, in a specific session, if Kahun suggested 3 different disorders and 2 of them were also suggested by the physician, Kahun's accuracy rate was 67%, 2/3).

Prevalence

Kahun calculated the prevalence of symptoms in different disorders based on the data from Kahun's knowledge engine and independently, based on the Kahun-DH cohort. For each disorder, symptoms that were suggested in more than 29 different assessments were selected. The prevalence of a symptom in a disorder was determined by the ratio of the total occurrences of the confirmed symptom to the total occurrences of the suggested symptom.

Only disorders and symptoms that appeared in both data sets were included. This created 2 equally dimensioned prevalence matrices: the Kahun prevalence matrix and the Kahun-DH cohort prevalence matrix. The prevalence similarity ratio was determined by dividing a prevalence from Kahun's prevalence matrix with the corresponding prevalence extracted from the Kahun-DH cohort's prevalence matrix.

Each prevalence matrix underwent hierarchical clustering that clustered disorders by symptoms using the pheatmap R package [19].

Ethical Considerations

This study analyzed anonymized data from 3 sources: Kahun's knowledge engine (published peer-reviewed papers), the Kahun-DH cohort, and the NAMCS. The data used in the Kahun-DH cohort was anonymized and deidentified according to HIPAA (Health Insurance Portability and Accountability Act) safe-harbor privacy rules. The data from NAMCS is publicly available and has been anonymized by the Centers for Disease Control and Prevention. As such, this study did not require Institutional Review Board approval, ethical review, or individual informed consent—all the data used was anonymized and deidentified, with no risk to individual privacy. Additionally, no personal or identifying information about participants was accessed or stored during the study. Appropriate measures were taken to ensure compliance with relevant privacy guidelines. No compensation was provided to participants, as this study was a secondary analysis of existing anonymized data sets.

Results

Kahun's Coverage Rates (Kahun's Knowledge Engine vs NAMCS)

Kahun covered 94% (526 million/562 million) of all weighted symptoms reported in the NAMCS 2019 data set. In 9 out of

XSL•FO
RenderX

10 different symptom groups, there was coverage for at least 85% of weighted symptoms. Notably, there was nearly complete coverage in the nervous system group (Figure 1). Additionally, Kahun covered 91% (1582 million/1734 million) of all weighted disorders reported in the NAMCS 2019 data set, with at least 88% coverage for weighted disorders in 14 out of 17 different disorder groups. Interestingly, nearly complete coverage was observed for diseases of the blood and certain disorders involving the immune mechanism, and for infectious and parasitic diseases groups (Figure 2).

Figure 1. Kahun's coverage rates for weighted symptoms reported in the National Ambulatory Medical Care Survey 2019. Purple-colored bars represent total symptoms while light-purple bars represent different groups of symptoms.

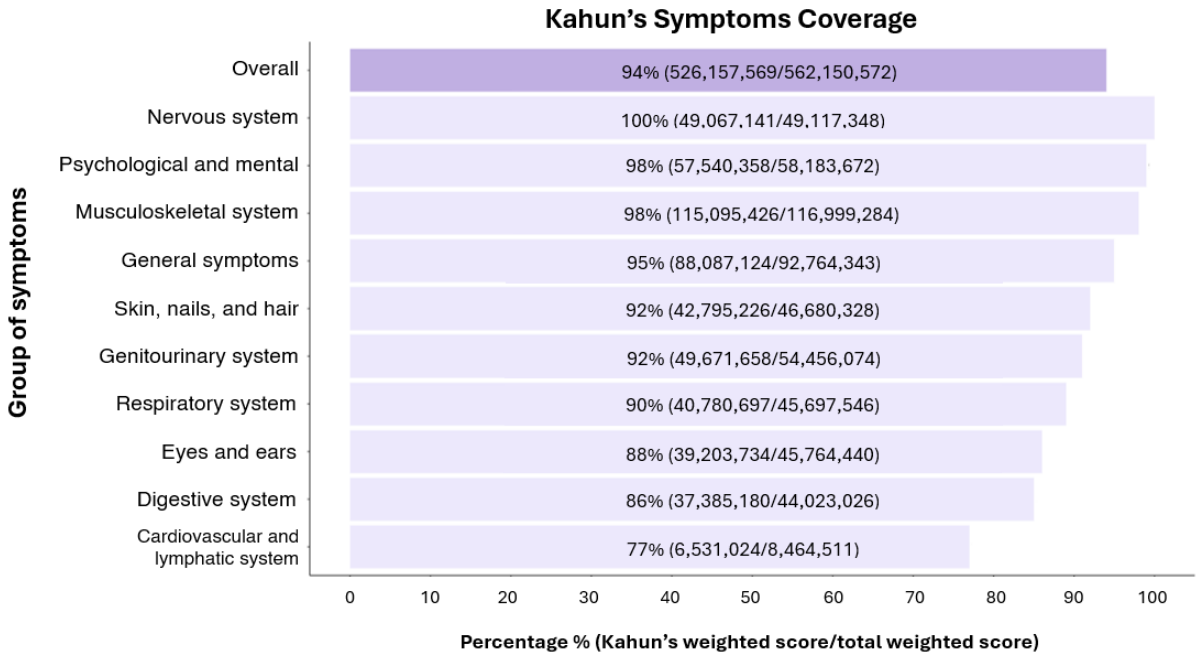

Figure 2. Kahun's coverage rates for weighted disorders reported in the National Ambulatory Medical Care Survey 2019. Purple-colored bars represent total disorders while light-purple bars represent different groups of disorders.

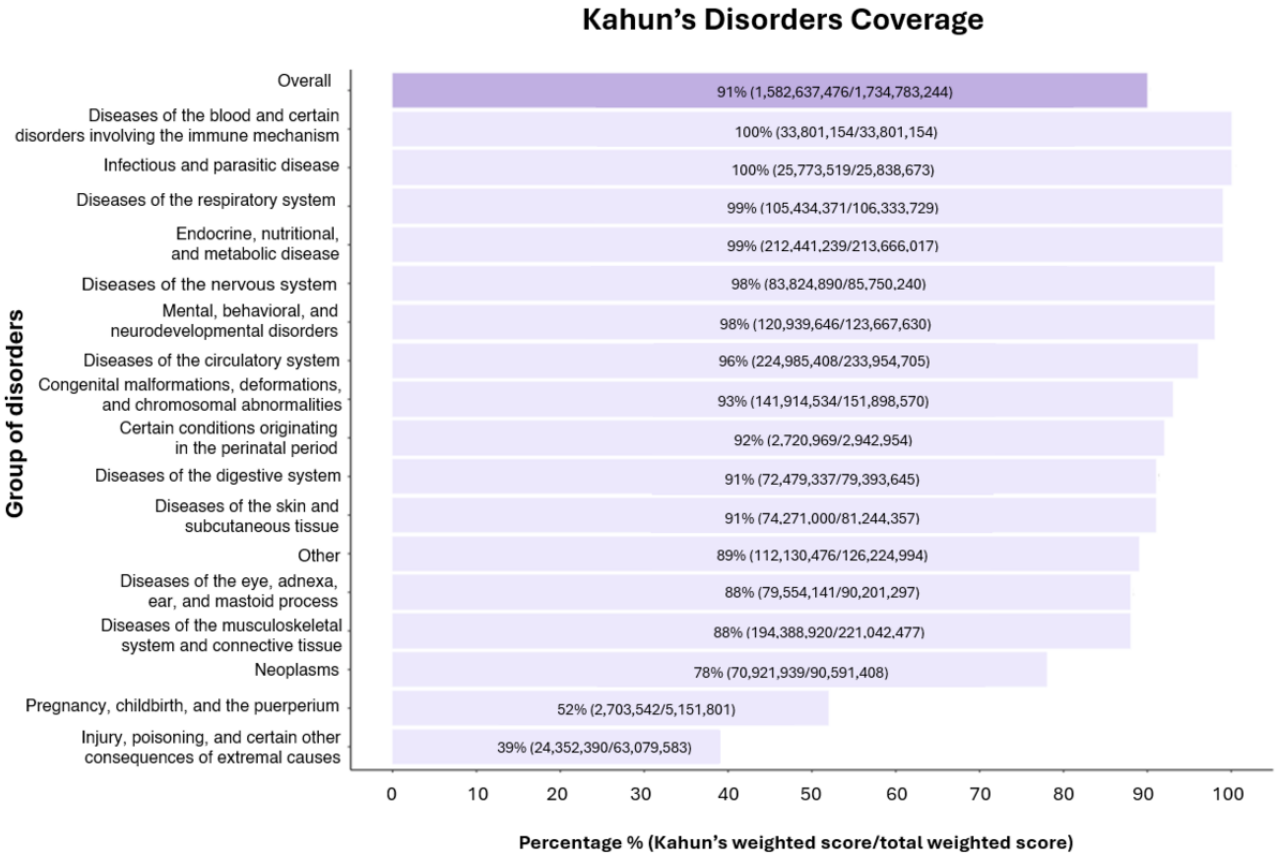

XSL•FO
RenderX

Patients' Characteristics (Kahun-DH Cohort)

Kahun-DH cohort included 1714 women (67%) and 836 men (33%) with a median age of 35 (IQR 16-90) years (Table 1). During each session, a median of 34 (IQR 11-61) questions were asked with a positive symptom ratio of 0.3 (Table 1).

During the assessments, 314 unique chief complaints were reported. Among them, the 5 most frequent chief complaints were anxiety (n=173), sore throat (n=153), cough (n=138), sinus pain (n=114), and headache (n=108). Additionally, the duration of the most frequent chief complaints was between 24 hours and 1 week for 32.3% (836/2550) of the complaints (Table 1).

Overall, Kahun generated 108,523 suggestions relating to 905 unique symptoms and provided 6496 disorders (321 unique disorders) as possible diagnoses.

Table 1. Patients' characteristics.

Characteristics	Overall (N=2550)
Age (years)	
Mean (SD)	38.3 (14.3)
Median (IQR)	35.0 (16.0-90.0)
Sex, n (%)	
Female	1714 (67.2)
Male	836 (32.8)
Duration of chief complaint, n (%)	
<24 hours	325 (12.7)
>24 hours and <1 week	824 (32.3)
>1 week and <12 weeks	481 (18.9)
>12 weeks	344 (13.5)
Missing	576 (22.6)
Total questions	
Mean (SD)	33.6 (9.20)
Median (IQR)	34.0 (11.0-61.0)
Positive symptoms ratio	
Mean (SD)	0.322 (0.201)
Median (IQR)	0.297 (0.0244-1.00)

Kahun's Diagnostic Ability (Kahun-DH Cohort)

A random sample of 250 Kahun-DH sessions was selected to assess Kahun's diagnostic ability. The sessions were blindly evaluated by physicians. In 90% (224/250) of the sessions, at least one identical disorder matched the differential diagnosis suggested by the physicians and the differential diagnosis suggested by Kahun. Additionally, 367 diagnoses that were suggested by Kahun matched the diagnoses suggested by the physicians, resulting in a 72% (367/507) accuracy rate.

Prevalence (Kahun-DH Cohort vs Kahun's Knowledge Engine)

Kahun calculated the prevalence of 60 symptoms in 28 disorders. A total of 519 prevalences were detected based on Kahun's knowledge engine, whereas 599 prevalences were detected based on the data from the Kahun-DH cohort (Figures 3 and 4, respectively). There was no statistically significant difference between the median prevalence value of Kahun's knowledge engine and the Kahun-DH cohort: 21% (IQR 5%-50%) and 23% (IQR 8%-47%), respectively (Wilcoxon P=.35).

Out of the prevalences that originated in the Kahun-HV cohort, 159 did not appear in Kahun's knowledge engine. This resulted in a median detection rate of 4 (IQR 1-10) new prevalences per disorder. Tonsillitis had the highest detection rate with 24 new prevalences, while 6 disorders had no new prevalences: disorders of the pituitary gland, hypertensive crisis, hypothyroidism, laryngitis, sexually transmitted infection diseases, and urinary tract infection diseases. Additionally, 443 prevalences were identified in both Kahun's knowledge graph and the DH patient cohort with a median prevalence similarity ratio of 1.04 (IQR 0.61-2.27). Among them, the prevalence similarity ratio of 85 (19%) prevalences ranged from 0.85 to 1.15.

Using a hierarchical clustering algorithm, 5 clusters of disorders were established based on their symptoms prevalence. Disorders with relatively similar symptom prevalence distribution were clustered together. The prevalence of symptoms that were not suggested by Kahun during the sessions in the corresponding disorder, though suggested in other disorders, was not calculated. The median number of disorders per cluster in Kahun's knowledge engine was 5, whereas the median number of disorders per cluster was 6 in the Kahun-DH cohort (Figures 3 and 4).

XSL•FO
RenderX

Figure 3. Heat map of the prevalence matrix of symptoms in disorders based on Kahun's knowledge engine. The white vertical lines cluster the disorders by their symptoms-prevalences distribution, using the hierarchical clustering algorithm. Dark red colors represent high prevalence while light yellow colors represent low prevalence. Gray squares represent symptoms that were not suggested by Kahun during sessions dealing with the corresponding disorder.

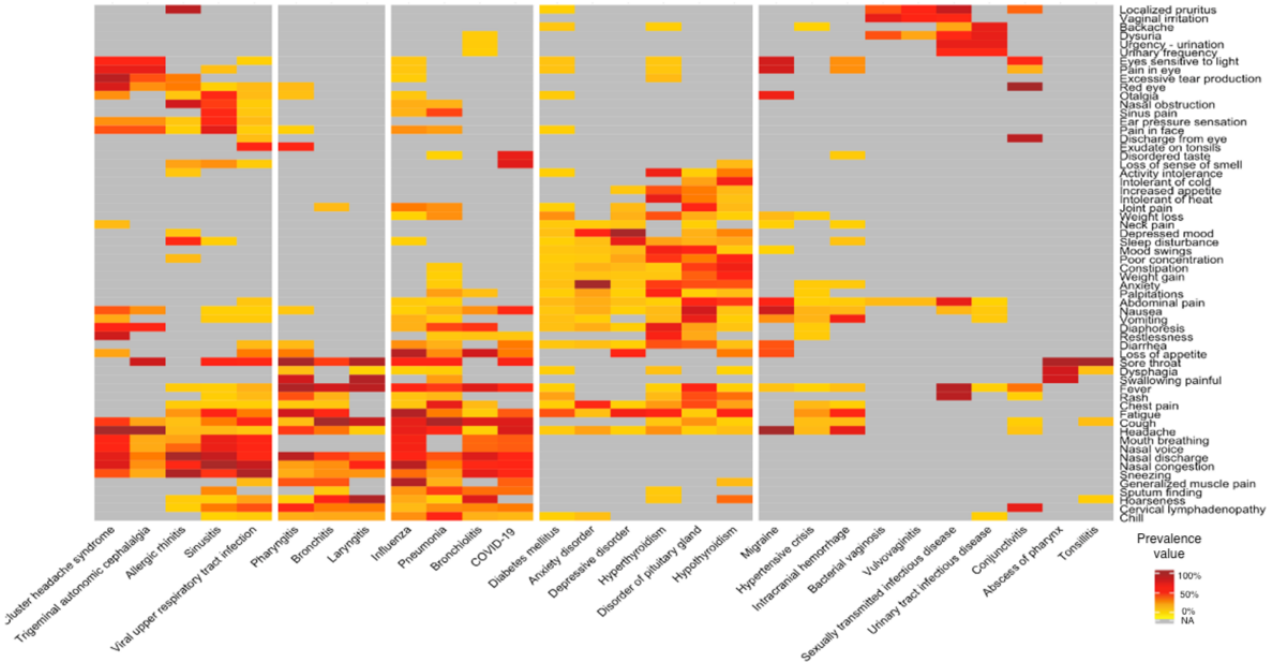

Figure 4. Heat map of the prevalence matrix of symptoms in disorders, based on Kahun-DH cohort. The white vertical lines cluster the disorders by their symptoms-prevalences distribution, using the hierarchical clustering algorithm. Dark red colors represent high prevalence while light yellow colors represent low prevalence. Gray squares represent symptoms that were not suggested by Kahun during sessions dealing with the corresponding disorder. DH: digital health care.

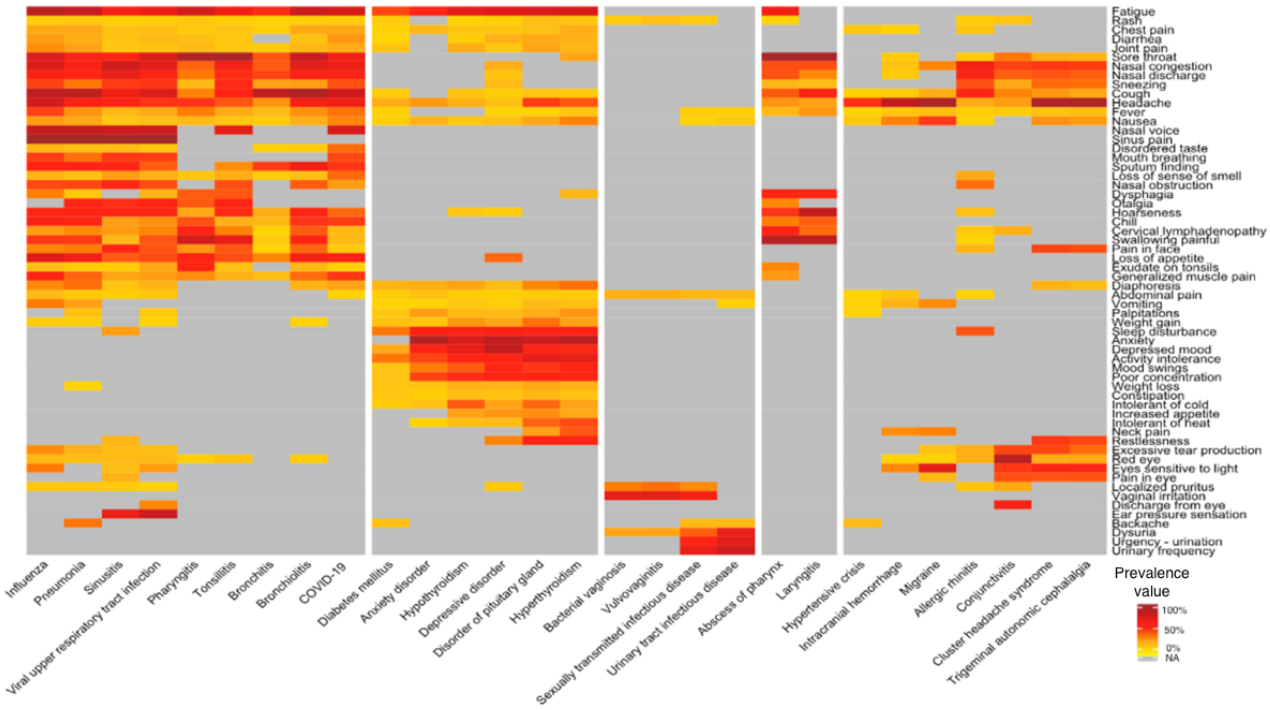

Discussion

Principal Findings

This study tested the potential of EBCITs to cover the same medical data assessed by physicians, thus acting as a reliable source for drawing medical conclusions in addition to identifying or grounding possible prevalences of symptoms in different disorders. We used the data in the Kahun EBCIT to examine our hypothesis.

Kahun's relatively comprehensive coverage of the weighted symptoms collected by NAMCS 2019 shows that although Kahun is not a human physician, it can address at least 94% of the actual symptoms that patients report during medical visits. Moreover, when the symptoms are classified into groups, Kahun

XSL•FO
RenderX

still covers most of the symptoms in each group. Kahun also generates differential diagnoses while taking into account about 91% (1,582,637,476/1,734,783,244) of the weighted disorders that physicians diagnosed in NAMCS 2019. These encouraging coverage rates indicate that data gathered by Kahun is almost identical to the data gathered by a human physician, and therefore, relatively reliable.

Moreover, while compared with human medical licensed physicians, Kahun's decent diagnostic ability indicates that the differential diagnosis suggested by Kahun and therefore its reasoning and intake process are reliable and trustable. Hence, we can infer that medical insights and assessments based on this data are valid.

To identify the prevalence of symptoms in disorders, we analyzed the Kahun-DH cohort. Similarly to other remote-care service users, most of the patients in this cohort were females and relatively young. Additionally, the total number of questions that were asked during the assessment was similar to other clinical intake tools [20]. Comparing a real physician interview and EBTICs might reveal a difference in the effectiveness of these approaches and should be explored further. While focusing on chief complaints, this study highlighted frequent chief complaints that are considered common among primary physicians, as suggested by other works [21,22]. This supports the claim that the data in the study agrees with real-world data. Additionally, more than 40% of the chief complaints lasted less than a week, presenting the relative acute characteristic of the symptoms, which although not widely studied, are supported by Lee et al [23].

The relatively young age of the patients and the fact that those patients were assessed by the DH's physicians only via digital encounters, might indicate that some of the typical brick-and-mortar clinic patients in the traditional setting of primary care are underrepresented in this cohort. Although telemedicine and digital encounters are gaining more popularity worldwide [24], we believe that evaluating EBCITs in a typical primary clinic setting is also important and should be considered in future research.

Nearly 109,000 suggestions, relating to 905 unique symptoms and approximately 6500 diagnoses of 321 distinct disorders, were included in this study, demonstrating the abundance of data gathered. Having such diverse data could enable profound research.

Moreover, the Kahun-DH cohort included only 2550 patients. Thus, increasing the use of Kahun, which was shown to have decent diagnostic accuracy [12], has the potential to assist physicians in medical assessment while contributing to extended research by enriching Kahun's database even further.

This study focused on the prevalence of symptoms in disorders. We compared the prevalences generated by Kahun's knowledge engine with prevalences from the Kahun-DH cohort. We found 156 prevalences, exclusively in the latter group. These unique prevalences emphasize that even after learning and memorizing more than 50,000 medical papers, just as Kahun did [13], medical knowledge can still be improved. Thus, consistently incorporating real-world data into EBCITs, like Kahun, could result in more precise and effective diagnoses which may help physicians in medical decision-making. An example of this assumption is partly demonstrated by the relative medical-improvement in the hierarchical clustering of the 28 analyzed disorders. The hierarchical clustering algorithm grouped together disorders with similar symptoms-prevalence distributions, which may resemble physicians' process of medical reasoning. For example, according to Figure 4, bacterial vaginosis, vulvovaginitis, sexually transmitted infection diseases, and urinary tract infection diseases were clustered together as they shared similar symptoms-prevalence distribution and can reasonably be clustered as genital-related clusters. By comparing the suggested clusters of the disorders, as shown in Figures 3 and 4, sexually transmitted infection diseases and headache-related disorders were reclustered with more appropriate disorders for each, as noted by certificated physicians. This improvement may indicate that as EBCITs are being used more, they result in better performances, thus suggesting more relevant diagnoses to the physician.

By assessing patients while taking into account a wide range of differential diagnoses and systematically storing the data, Kahun can ask unorthodox questions and less obvious prevalences may be discovered. These prevalences have the potential to serve as new leads for innovative research or medical trends. For example, a new possible prevalence of hoarseness in pharyngeal abscesses was identified by the Kahun-DH cohort, with an estimated prevalence value of 48% (19/40). Literature reviews suggest that such a prevalence exists [25-27]; although, as far as we know, an exact evaluation of this prevalence was not conducted. Using the Kahun-DH cohort, we were able to estimate this prevalence.

A practical example of the potential of EBCIT in grounding prevalences is the prevalence of nasal congestion in migraines. Although such a prevalence did not appear in Kahun's knowledge engine, according to the data from the Kahun-DH cohort, the prevalence was 28% (9/32; Figure 4). Remarkably, a study by Muehlberger et al [28] evaluated this prevalence as 25%. These relatively close estimations testify to Kahun's potential to discover prevalences accurately.

Apart from the novel prevalences that were discovered, 443 prevalences were simultaneously identified in both data sets. Of those prevalences, 81% had relative difference values (prevalence similarity ratio >±0.15). This crucial finding highlights the importance of grounding prevalences. By combining comprehensive databases, such as the ones EBCITs may provide, with the information from previous studies, these discrepancies could be resolved and more refined prevalences would be computed. A different approach for settling these discrepancies argues that each cohort represents a different population. Hence, each prevalence is valid as long it refers to its original cohort. The latter approach may be helpful in creating personalized medical prevalences.

While comparing Kahun-DH's prevalences with Kahun's knowledge engine prevalences, we included only selected symptoms and disorders, mainly to compare reasoning and sample size issues. Other prevalences were identified based on Kahun's knowledge engine and the Kahun-DH cohort but are

XSL•FO
RenderX

not presented in this analysis. The actual number of prevalences is much greater. Splitting the main cohort into smaller, more homogeneous, subcohorts, could yield prevalences that are more personalized.

Limitations

This study has some known limitations. The coverage analysis refers to only some of the symptoms and disorders, as explained in the Methods section. On one hand, because Kahun is an EBCIT and presently does not cover physical examination nor does it save personal information for follow-up visits, some valuable information was not covered in this tool by definition. Therefore, the full coverage of the NAMCS is unknown. On the other hand, many other symptoms and disorders that were not included in NAMCS exist in Kahun's knowledge engine. Therefore, we recommend further coverage analysis of other tools and reference databases. Moreover, the NAMCS and DH cohort both cover the US population. Although the US population is diverse, some of the insights from this study are less relevant to other populations.

Although Kahun's diagnostic ability showed encouraging results and improved performance compared with previous studies [13], no such tool can be considered perfect. Therefore, constant improvement of Kahun is advised, and as a result, the grounded prevalences suggested in this study will probably be refined in the future. Since the focus of this study was not on Kahun's accuracy rate, the sample that was used to determine this accuracy rate included only about 10% of the medical vignette data (250/2550). Thus, further research that focuses on EBCITs' diagnosis ability and includes more data should be considered in the future.

An additional limitation of the Kahun-DH cohort is that it consists only of patients who can afford DH services and actively seek medical care. This fact can potentially lead to selection bias and limit the generalizability of our real-world prevalences.

Moreover, since the setting of this cohort was based on digital encounters and included younger patients with fewer comorbidities, there might be an underrepresentation of older patients with more comorbidities, and the relevant prevalences might be affected. Therefore, future research involving typical primary care settings should be conducted.

Another limitation of this study is that most of the authors are employed by Kahun. This potential bias may have influenced the design and interpretation of the study results. However, we have taken steps to minimize the impact of this bias by using rigorous methods and statistical analyses and by involving other authors who are not employed by Kahun in analyzing some of the data. Despite these efforts, we recognize that the potential for bias remains and we encourage readers to carefully evaluate the evidence presented in this study and consider alternative perspectives.

Comparison With Prior Work

As far as we know, only one study, conducted by Perlman et al [17], showed the potential advantage of using clinical intake tools to infer medical insights. Specifically, the study identified relative rates and associations between disorders and symptoms. That in-depth study included 71,619 self-assessments of COVID-19 participants. The study focused only on one disorder: COVID-19. Although it also included only several predetermined symptoms, it managed to identify prominent results such as an increased probability of having COVID-19 if a loss of smell or taste was experienced. It is worth noting that the demographic characteristics of the participants in that study, which were mainly females in their 30s, were similar to another study regarding clinical intake tools [14] and similar to our results.

In our view, the pioneering work done by Perlman et al [17] is extremely significant. Because their study had a relatively narrow scope in terms of coverage, symptoms, and disorders, we believed that at least one study with a wider scope was needed to support our main hypothesis. To the best of our knowledge, this study showed, for the first time, the potential of an EBCIT to generate prevalences of many symptoms in diseases.

Conclusions

Kahun has the ability to cover most of the symptoms that patients may present while addressing most of the possible disorders. Although not perfect, Kahun has the potential to serve as a reliable medical database and provide plausible EBM diagnoses. Thus, increased use of EBCITs, such as Kahun, could help an improvement in EBCITs' ability to uncover medical insights, including discovering and grounding prevalences of symptoms in disorders, which might support physicians in medical diagnosis and expand medical knowledge.

Acknowledgments

This study was funded by Kahun Medical Ltd.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

EA, SO, and DS contributed to conceptualization. EA, SO, YL, and EY contributed to data curation and software. MS, EY, and DS contributed to the investigation. EA, MS, DBJ, and EY contributed to methodology and project administration. DBJ, ES, EY,

XSL•FO
RenderX

and DE contributed to validation. EA, MS, YL, and DS contributed to writing the original draft. EA, ES, EY, and DE contributed to reviewing and editing the paper. All authors reviewed and approved the final version of this manuscript.

Conflicts of Interest

All authors are salaried employees at Kahun Medical Ltd except EY.

References

1. Bonilla-Velez J, Small M, Urrutia R, Lomberk G. The enduring value of research in medical education. Int J Med Students 2017 Dec 26;5(1):37-44 [FREE Full text] [doi: 10.5195/ijms.2017.168]
2. Sackett D. Evidence-based medicine. Semin Perinatol 1997 Feb;21(1):3-5. [doi: 10.1016/s0146-0005(97)80013-4] [Medline: 9190027]
3. Dehon E, Weiss N, Jones J, Faulconer W, Hinton E, Sterling S. A systematic review of the impact of physician implicit racial bias on clinical decision making. Acad Emerg Med 2017 Aug;24(8):895-904 [FREE Full text] [doi: 10.1111/acem.13214] [Medline: 28472533]
4. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics 2017 Mar 01;18(1):19 [FREE Full text] [doi: 10.1186/s12910-017-0179-8] [Medline: 28249596]
5. Moore N, Detering K, Low T, Nolte L, Fraser S, Sellars M. Doctors' perspectives on adhering to advance care directives when making medical decisions for patients: an Australian interview study. BMJ Open 2019 Oct 31;9(10):e032638 [FREE Full text] [doi: 10.1136/bmjopen-2019-032638] [Medline: 31676659]
6. Au-Yong-Oliveira M, Pesqueira A, Sousa M, Dal Mas F, Soliman M. The potential of big data research in healthcare for medical doctors' learning. J Med Syst 2021 Jan 07;45(1):13 [FREE Full text] [doi: 10.1007/s10916-020-01691-7] [Medline: 33409620]
7. Mayer-Schönberger V, Ingelsson E. Big data and medicine: a big deal? J Intern Med 2018 May;283(5):418-429 [FREE Full text] [doi: 10.1111/joim.12721] [Medline: 29235174]
8. Ioannidis JPA, Khoury MJ. Evidence-based medicine and big genomic data. Hum Mol Genet 2018 May 01;27(R1):R2-R7 [FREE Full text] [doi: 10.1093/hmg/ddy065] [Medline: 29474574]
9. Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol 2019 Aug;58:161-167 [FREE Full text] [doi: 10.1016/j.copbio.2019.03.004] [Medline: 30965188]
10. Lin JZ, Long JY, Wang AQ, Zheng Y, Zhao HT. Precision medicine: in need of guidance and surveillance. World J Gastroenterol 2017 Jul 28;23(28):5045-5050 [FREE Full text] [doi: 10.3748/wjg.v23.i28.5045] [Medline: 28811702]
11. Afzal M, Riazul Islam SM, Hussain M, Lee S. Precision medicine informatics: principles, prospects, and challenges. IEEE Access 2020;8:13593-13612 [FREE Full text] [doi: 10.1109/access.2020.2965955]
12. Ben-Shabat N, Sloma A, Weizman T, Kiderman D, Amital H. Assessing the performance of a new artificial intelligence-driven diagnostic support tool using medical board exam simulations: clinical vignette study. JMIR Med Inform 2021 Nov 30;9(11):e32507 [FREE Full text] [doi: 10.2196/32507] [Medline: 34672262]
13. Kahun Ltd. URL: https://www.kahun.com/ [accessed 2023-05-24]
14. Fraser HSF, Cohan G, Koehler C, Anderson J, Lawrence A, Pateña J, et al. Evaluation of diagnostic and triage accuracy and usability of a symptom checker in an emergency department: observational study. JMIR Mhealth Uhealth 2022 Sep 19;10(9):e38364 [FREE Full text] [doi: 10.2196/38364] [Medline: 36121688]
15. Gräf M, Knitza J, Leipe J, Krusche M, Welcker M, Kuhn S, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 2022 Dec;42(12):2167-2176 [FREE Full text] [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]
16. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One 2021;16(7):e0254088 [FREE Full text] [doi: 10.1371/journal.pone.0254088] [Medline: 34265845]
17. Perlman A, Vodonos Zilberg A, Bak P, Dreyfuss M, Leventer-Roberts M, Vurembrand Y, et al. Characteristics and symptoms of app users seeking COVID-19-related digital health information and remote services: retrospective cohort study. J Med Internet Res 2020 Oct 20;22(10):e23197 [FREE Full text] [doi: 10.2196/23197] [Medline: 32961527]
18. National ambulatory medical care survey. US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/ahcd/about_ahcd.htm [accessed 2023-05-24]
19. Kolde R. Package Pheatmap. 2019. URL: https://cran.r-project.org/package=pheatmap [accessed 2024-06-12]
20. Ben-Shabat N, Sharvit G, Meimis B, Ben Joya D, Sloma A, Kiderman D, et al. Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study. Int J Med Inform 2022 Dec;168:104897 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104897] [Medline: 36306653]
21. Finley CR, Chan DS, Garrison S, Korownyk C, Kolber MR, Campbell S, et al. What are the most common conditions in primary care? systematic review. Can Fam Physician 2018 Nov;64(11):832-840 [FREE Full text] [Medline: 30429181]
22. Krüger K, Töpfner N, Berner R, Windfuhr J, Oltrogge J, Guideline group. Clinical practice guideline: sore throat. Dtsch Arztebl Int 2021 Mar 19;118(11):188-194 [FREE Full text] [doi: 10.3238/arztebl.m2021.0121] [Medline: 33602392]

XSL•FO
RenderX

23. Lee H, Choi S, Rim J, Lim H, Heo Y, Shin J, et al. Illnesses encountered during medical volunteering in Takeo province, Cambodia. Medicina (Kaunas) 2020 Jan 10;56(1):30 [FREE Full text] [doi: 10.3390/medicina56010030] [Medline: 31936867]
24. Meuter LS, Wolf KI, Pacak K. Maintaining professional encounters and enhancing telemedicine interactions with core virtual-clinical values. Endocr Pract 2021 Jan;27(1):77-79 [FREE Full text] [doi: 10.1016/j.eprac.2020.11.006] [Medline: 33475505]
25. Christoforidou A, Metallidis S, Kollaras P, Agathangelidis A, Nikolaidis P, Vital V, et al. Tuberculous retropharyngeal abscess as a cause of oropharyngeal dysphagia. Am J Otolaryngol 2012;33(2):272-274. [doi: 10.1016/j.amjoto.2011.06.006] [Medline: 21871690]
26. Lin J, Wu XM, Feng J, Chen M. Retropharyngeal abscess presenting as acute airway obstruction in a 66-year-old woman: a case report. World J Clin Cases 2019 Nov 26;7(22):3838-3843 [FREE Full text] [doi: 10.12998/wjcc.v7.i22.3838] [Medline: 31799312]
27. Mishra A, Anand A, Gupta K, Saini R. Tuberculous retropharyngeal abscess as an indiscernible cause of dysphagia. J Case Rep 2018;8(1):76-79.
28. Muehlberger T, Wormald JCR, Hachach-Haram N, Mosahebi A. Migraine: a look down the nose. J Plast Reconstr Aesthet Surg 2017 Jul;70(7):914-921. [doi: 10.1016/j.bjps.2017.02.009] [Medline: 28359726]

Abbreviations

DH: digital health care
EBCIT: evidence-based clinical intake tool
EBM: evidence-based medicine
HIPPA: Health Insurance Portability and Accountability Act
ICD-10: International Statistical Classification of Diseases, Tenth Revision
NAMCS: National Ambulatory Medical Care Survey

Edited by S Ma; submitted 17.06.23; peer-reviewed by N Slabyj, R Bidkar, S Pesälä; comments to author 17.01.24; revised version received 24.02.24; accepted 27.05.24; published 16.07.24.

Please cite as:
Avnat E, Samin M, Ben Joya D, Schneider E, Yanko E, Eshel D, Ovadia S, Lev Y, Souroujon D
The Potential of Evidence-Based Clinical Intake Tools to Discover or Ground Prevalence of Symptoms Using Real-Life Digital Health Encounters: Retrospective Cohort Study
J Med Internet Res 2024;26:e49570
URL: https://www.jmir.org/2024/1/e49570
doi:10.2196/49570
PMID:39012659

©Eden Avnat, Michael Samin, Daniel Ben Joya, Eyal Schneider, Elia Yanko, Dafna Eshel, Shahar Ovadia, Yossi Lev, Daniel Souroujon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ChatGPT for Automated Qualitative Research: Content Analysis

Rimke Bijker[1], MSc, PhD; Stephanie S Merkouris[2], PhD; Nicki A Dowling[2], PhD; Simone N Rodda[1,2], PhD

[1]Department of Psychology and Neuroscience, Auckland University of Technology, Auckland, New Zealand
[2]School of Psychology, Deakin University, Burwood, Australia

Corresponding Author:
Simone N Rodda, PhD
Department of Psychology and Neuroscience
Auckland University of Technology
90 Akoranga Drive
Auckland, 0627
New Zealand
Phone: 64 9921 9999 ext 29079
Email: simone.rodda@aut.ac.nz

Abstract

Background: Data analysis approaches such as qualitative content analysis are notoriously time and labor intensive because of the time to detect, assess, and code a large amount of data. Tools such as ChatGPT may have tremendous potential in automating at least some of the analysis.

Objective: The aim of this study was to explore the utility of ChatGPT in conducting qualitative content analysis through the analysis of forum posts from people sharing their experiences on reducing their sugar consumption.

Methods: Inductive and deductive content analysis were performed on 537 forum posts to detect mechanisms of behavior change. Thorough prompt engineering provided appropriate instructions for ChatGPT to execute data analysis tasks. Data identification involved extracting change mechanisms from a subset of forum posts. The precision of the extracted data was assessed through comparison with human coding. On the basis of the identified change mechanisms, coding schemes were developed with ChatGPT using data-driven (inductive) and theory-driven (deductive) content analysis approaches. The deductive approach was informed by the Theoretical Domains Framework using both an unconstrained coding scheme and a structured coding matrix. In total, 10 coding schemes were created from a subset of data and then applied to the full data set in 10 new conversations, resulting in 100 conversations each for inductive and unconstrained deductive analysis. A total of 10 further conversations coded the full data set into the structured coding matrix. Intercoder agreement was evaluated across and within coding schemes. ChatGPT output was also evaluated by the researchers to assess whether it reflected prompt instructions.

Results: The precision of detecting change mechanisms in the data subset ranged from 66% to 88%. Overall κ scores for intercoder agreement ranged from 0.72 to 0.82 across inductive coding schemes and from 0.58 to 0.73 across unconstrained coding schemes and structured coding matrix. Coding into the best-performing coding scheme resulted in category-specific κ scores ranging from 0.67 to 0.95 for the inductive approach and from 0.13 to 0.87 for the deductive approaches. ChatGPT largely followed prompt instructions in producing a description of each coding scheme, although the wording for the inductively developed coding schemes was lengthier than specified.

Conclusions: ChatGPT appears fairly reliable in assisting with qualitative analysis. ChatGPT performed better in developing an inductive coding scheme that emerged from the data than adapting an existing framework into an unconstrained coding scheme or coding directly into a structured matrix. The potential for ChatGPT to act as a second coder also appears promising, with almost perfect agreement in at least 1 coding scheme. The findings suggest that ChatGPT could prove useful as a tool to assist in each phase of qualitative content analysis, but multiple iterations are required to determine the reliability of each stage of analysis.

(J Med Internet Res 2024;26:e59050) doi:10.2196/59050

KEYWORDS

ChatGPT; natural language processing; qualitative content analysis; Theoretical Domains Framework

XSL•FO
RenderX

Introduction

Background

Emerging from a variety of fields such as psychology, nursing, media communication, and market research, content analysis has become one of the main methods for analyzing large qualitative data sets [1,2]. Content analysis uses systematic and replicable methods to synthesize and describe patterns in textual, visual, or audio data and facilitate understanding of the content and meaning of the data [1-4]. It can be used to analyze qualitative data obtained from interviews, focus groups, and patient records as well as naturally occurring data such as user-generated website content, discussion forums, social media, and customer databases [4,5]. Furthermore, content analysis can use quantitative methods, which focus on quantification of the data to enable calculation of prevalence data or use in statistical analyses, or qualitative methods, which focus on distilling concepts or constructs from the data to facilitate understanding [6]. Qualitative content analysis also allows for the integration of quantitative components, for instance, by summarizing the findings in frequency distributions, which can be particularly useful when working with large data sets [7].

Content analysis can be performed using an inductive, data-driven approach or a deductive, theory-driven approach. An inductive approach is suitable when the topic of interest is still emerging or lacking a firm scientific knowledge database [4,8,9]. In contrast, a deductive approach is more appropriate when there is an existing body of knowledge and the aim is to confirm hypotheses and code data into an existing framework or theory [3,10]. While the practice and philosophy of qualitative content analytic approaches may vary, they have similar tasks to be executed as part of the data analysis process, including identification of relevant data, organization of data, and data classification [6]. The systematic nature of this process provides opportunities for computer assistance. In fact, computer-assisted execution of qualitative analysis tasks has been around for some time with the aim of improving the speed and accuracy of the analysis process, especially when large data sets are involved [11].

Whether inductive or deductive, each approach to content analysis requires a preparation phase, which involves reading the text for familiarization and identifying the meaning units (ie, the words or sentences in the text that are relevant to the research questions and contained within the unit of analysis) by assigning them initial labels that reflect the key content of the meaning units [12]. It is recommended that this is repeated multiple times, whereby there is a cycle of reading and rereading the data and adjusting the labels so that they represent the data that are being coded [1]. The data are then subject to condensation, where the meaning units are summarized into short descriptions that maintain the key content of the data and are assigned codes that are grouped according to content similarity [13]. Traditional qualitative analysis software offers automated coding, but multiple studies have reported inaccuracies, suggesting that manual coding should remain the dominant approach [13,14].

Following the preparation phase, machine learning may assist in the development of a coding scheme (also called a coding frame or data dictionary) that can be used to systematically organize the data by allocating each code to that scheme. When the analysis involves large data sets, coding schemes are typically constructed on a subset of data, after which the coding scheme can be applied to the full data set [15]. For the inductive content analysis approach, the development of a coding scheme involves organizing codes and initial groupings into categories according to the research question [3,4]. Categories are continuously adjusted and are often split into subgroups to ensure that they are discrete and mutually exclusive while still adequately reflecting the data [12]. Meaning units might have to be adapted as well to ensure that each code belongs to only 1 category [2]. Multiple researchers can also develop lists of categories that are combined into 1 coding scheme via negotiation [16]. For each category, labels are developed to define the category, and commonly, a separate coding manual is composed to specify coding rules [2].

For the deductive content analysis approach, codes are similarly organized into categories; however, the categories are predefined in a coding matrix that is informed by a theory or framework [3,4]. The matrix can then be made into a coding scheme that reflects the data being coded by labeling and relabeling the categories in line with the current data set [3]. This unconstrained deductive approach results in a coding scheme in which different categories may emerge that still reflect the data and are consistent with the theory or framework. Alternatively, a structured deductive approach can test categories, concepts, and hypotheses by coding data directly into predefined categories of the coding matrix [3]. When using a structured coding matrix, the categories are not redefined to reflect the data being coded, thereby giving the researcher the opportunity to focus on the degree of alignment between the matrix and the current data set [3].

Whether following an inductive or unconstrained deductive approach, the development of a coding scheme is an iterative process that involves continuously updating the coding scheme until a version is achieved that can be used to reliably code the entire data set [17]. The development and refinement of coding schemes is generally conducted by the research team and may involve either others with expertise in the framework or theory or a literature review [3,4]. There is recent evidence suggesting that machine learning and natural language processing techniques are promising for relatively straightforward tasks such as data extraction and classification of unstructured qualitative data [18,19]. However, using such techniques typically requires programming skills that are unavailable to most researchers.

After the development phase, the quality of the coding schemes can be evaluated by assessing the reliability of application. This phase involves ≥2 researchers classifying the data into the defined categories and comparing the resulting annotated data sets [2]. Manual coding may be aided by technology in terms of sorting and renaming categories (eg, using Excel [Microsoft Corp] or qualitative analysis software), but this is also prone to fatigue and error, especially in cases in which coding schemes have multiple categories and subcategories or interpretation of

XSL•FO
RenderX

meaning is part of the selected approach [3,10,13,20,21]. For example, reliability is dependent on the number of correctly coded items that may be at the higher order or subcategory level. While percentage of agreement is often reported [21], a more robust approach is through κ interrater agreement (henceforth referred to as intercoder agreement, a more appropriate term when classification relates to nominal categories) [22]. This intercoder agreement determines the degree to which the coders agree beyond what would be expected from chance alone [23]. The calculation of the κ score relies on each category in the coding scheme being mutually exclusive and the unit of analysis being independent in that it contains only 1 statement or code [23]. In cases in which the κ score indicates low intercoder agreement, the coding scheme might require further refinement until the application of the coding scheme achieves high intercoder agreement [2]. The final coding scheme can be used to classify the content of the full data set and produce frequency data for further analysis. In cases in which previous evaluation of the coding scheme was based on a data subset, reliability might be assessed again once the full data set has been annotated by multiple researchers as coding of large data sets may be prone to errors due to researcher fatigue and subtle changes in interpretation over time [2,6,13]. While reliability is commonly evaluated by intercoder agreement, an alternative option is intracoder agreement, in which the same person codes the same data on 2 separate occasions [24].

In addition to evaluating reliability, the validity of the findings from content analysis can be enhanced via triangulation, which can be achieved by combining methodologies, involving various investigators, or incorporating multiple theories [4,8,13,20,25]. In content analysis, increased validity might be achieved through transparency in reporting and potential for replication, confirmation that the categories relate to the research questions, member checks, evaluation of the appropriateness of the coding scheme by content experts, and the use of different data sources [3,4,13]. Other considerations relate to credibility (ie, how well data analysis procedures ensure the inclusion of all relevant data, which may be improved through consensus among the research team or external experts) and dependability (ie, how the analytic process may alter data over time—such as recoding data—thereby implying a need for careful logging of changes and consistent decision-making) [13].

ChatGPT

Advances in artificial intelligence (AI) and machine learning technologies have greatly impacted working life by automating manual processes and assisting in laborious tasks. One of the more prominent developments by OpenAI is ChatGPT, a large language model (LLM) that has been trained using a wide range of data, which enables it to synthesize human inputs and generate humanlike responses [26]. Inputs are provided in the form of "prompts," which contain questions or instructions that are processed to provide responses in line with the training data. According to an investigation by Altmetric, a platform that tracks web-based attention to research, the interest in ChatGPT appears to be particularly high among scientists [27]. Since its launch in November 2022 [28], scientists have enlisted the assistance of ChatGPT for a range of tasks, including conceptualizing ideas, summarizing scientific literature, generating analytic code, and writing up manuscripts [22,29]. In this field, ChatGPT's potential has primarily been discussed in relation to scientific writing and editing. For example, ChatGPT has been used to conduct rapid literature reviews, translate and edit texts, structure manuscripts to increase readability and fulfill journal guidelines, write logically sound abstracts, and provide suggestions on how to address reviewer comments [26,30-32].

Research suggests that ChatGPT may be a promising tool for qualitative analysis methods [33-36]. As a natural language processing tool that applies computation techniques and semantic similarities to analyze and synthesize natural language and speech [37], ChatGPT may enhance the streamlining of tasks and increase the efficiency of qualitative research projects. Nevertheless, interviews with those involved in qualitative analysis suggest that the time needed for prompt engineering, understanding complex responses, and organizing unstructured output, combined with concerns about the accuracy and validity of the output, may not outweigh the benefits of incorporating ChatGPT as a tool in qualitative research [38].

Study Aims

The aim of this study was to explore the utility of ChatGPT in conducting qualitative content analysis through the analysis of forum posts from people sharing their experiences on reducing their sugar consumption. Specifically, given the various analytic approaches that can be used, the aim was to explore the utility of ChatGPT in both inductive (ie, data-driven) and deductive (ie, using an existing framework) approaches. For the deductive approach, we used the Theoretical Domains Framework (TDF) as this is a commonly used framework to guide evaluation of the implementation of evidence-based practice in health care settings [39]. A secondary aim was to provide insights into the mechanisms of sugar reduction most frequently discussed in forum data by providing frequency data generated using the inductive and deductive approaches.

Methods

Overview

An overview of the study methods is presented in Figure 1. The qualitative content analysis process was broken down into several tasks—referred to as queries—that we identified as having the potential for automation using ChatGPT. Prompt engineering was used to create appropriate prompts with instructions in line with the respective tasks (ie, the queries). In an approximation of a manual coding process, ChatGPT was first used to generate a data set of condensed meaning units that were extracted from forum posts on sugar reduction. On the basis of a subset of the generated data set, ChatGPT was then used to create 10 versions of an inductive, data-driven coding scheme and 10 versions of an unconstrained deductive, theory-driven coding scheme. The unconstrained deductive coding scheme was based on the TDF. Thus, different versions of the coding schemes were developed in parallel, in contrast to a manual process in which coding schemes are developed iteratively and reflect more refined updates from previous versions. ChatGPT was then used to annotate the full data set by applying the coding schemes. ChatGPT was also used to

XSL•FO
RenderX

code the full data set into a structured coding matrix based on the TDF, in which data were coded directly into the predefined categories guided only by the training data of ChatGPT's underlying LLM. All coding scheme versions and the coding matrix were applied to the full data set 10 times in 10 new conversations, whereby each conversation was considered an independent coder. By doing so, the resulting annotated data sets could be compared to assess the interrater reliability of the coding schemes.

Figure 1. Study flow. TDF: Theoretical Domains Framework.

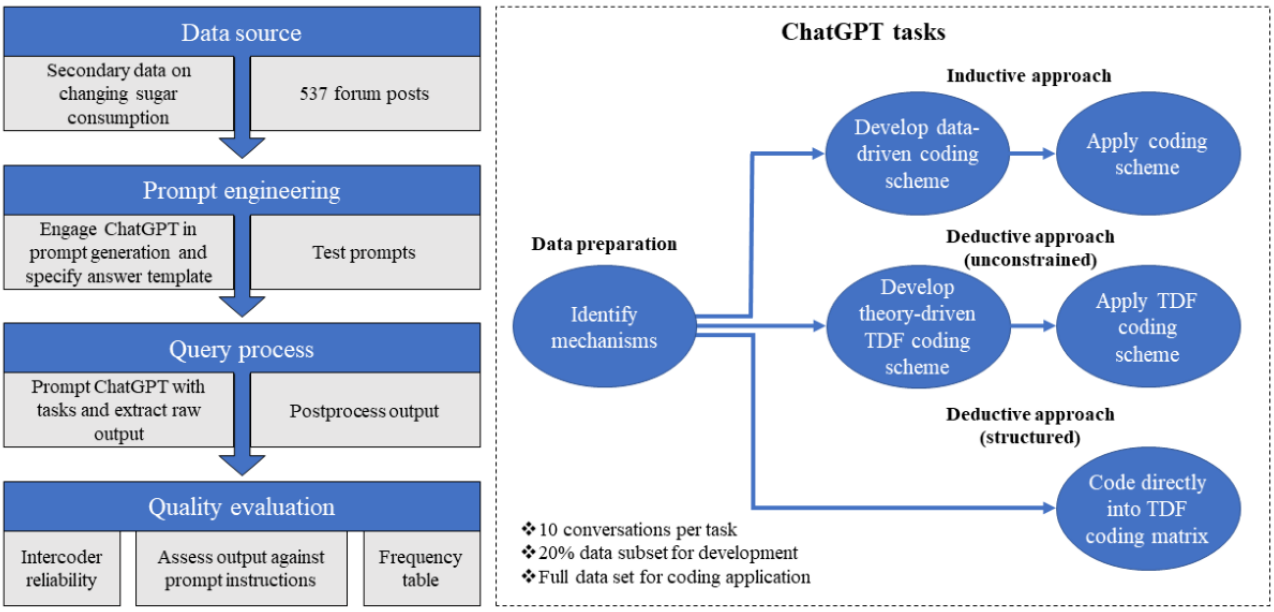

All ChatGPT queries were conducted from June 12 to 18, 2023, when ChatGPT operated under the GPT-3.5 Turbo model [40]. We opted to use the ChatGPT web application with unpaid subscriptions to ensure that our approach can be applied by anyone interested in using qualitative content analysis regardless of the programming skills and financial resources they have available to conduct their research.

Data Source

The data source of this study was a study on mechanisms of sugar reduction that has been previously described [41]. Briefly, Google searches were conducted to identify web-based content in which internet users mentioned mechanisms related to changing sugar consumption. In this search, potentially relevant posts were sourced from consumer platforms (eg, forums and message boards); popular media platforms (eg, blogs, web-based magazines, and news articles); and professional platforms (governmental and treatment provider websites). User-generated posts (including context, where applicable) were manually transferred to a Microsoft Excel spreadsheet and assigned labels and codes reflecting the change mechanisms identified in the forum posts. Mechanisms were then analyzed to create an overview of the range of change mechanisms applied to reduce sugar consumption. For our study, we retained all consumer posts, which resulted in a data set with 539 unique forum posts.

Ethical Considerations

All data were from open access sources and freely available without sign-in requirement or agreement with a specified set of terms and conditions. The retained posts did not contain any user-identifiable data. As such, this study was exempted from ethics approval in accordance with Auckland University of Technology Ethics Committee guidelines [42].

Prompt Engineering

Research has highlighted the importance of using high-quality prompts for LLMs given that different prompts that appear to reflect similar instructions may lead to highly variable responses [43,44]. Therefore, 2 researchers (RB and SSM) engaged in a systematic process of prompt engineering to generate and compare prompts. To enhance the quality of our prompts, we used a combination of various techniques that have been suggested to improve responses in conversations with LLMs such as ChatGPT [45-47]. For example, we started conversations with prompts that included information on the context of our study and then applied techniques to engage ChatGPT in an iterative prompt engineering process in which instructions referred to writing relevant prompts, providing alternative prompts, refining prompts based on additional information, listing the pros and cons of prompts, and recommending prompts given the relevant pros and cons. Through this process, we refined and selected at least 3 prompts per analysis task that were all run several times to assess which prompts contained the most optimal instructions for use in the qualitative content analysis query process. Assessment at this stage was based on the extent to which responses were in line with the prompt instructions, such as correct response format, no apparent issues in response content, and comprehensiveness of the response.

The final set of prompts for the query process had a similar construction in that each prompt started with (1) detailed instructions on the task to be executed, including notes on what to watch out for, if applicable, followed by (2) specifications on the required response format or answer template and ending with (3) information on which the response had to be based (ie, list of forum posts or change mechanisms). In some prompts, the instructions followed a step-by-step structure [48] as prompt

XSL•FO
RenderX

engineering showed that this resulted in a more comprehensive response. Textbox 1 provides a prompt example (with a step-by-step structure), and Multimedia Appendix 1 provides a complete overview of the final prompts. At the time of this research, ChatGPT had a limit on the number of tokens (ie, units that are meaningful to an LLM, such as words, word fragments, or punctuation signs) processed in each combined prompt and response. Consequently, for each query, we reprompted ChatGPT with the same instructions but different subsets of the data on which responses had to be based until all data were processed.

Textbox 1. Example of a prompt used in this study.

Example

- Below are forum posts reflecting real-life experiences related to changing sugar consumption. For each of these posts, please extract all potential mechanisms, strategies, or techniques for reducing sugar consumption. Each mechanism should be described in a brief and concise manner, using up to 10 words. If multiple mechanisms can be extracted from a single post, each should be mentioned separately. Please note that mechanisms may not be explicitly presented as such, so be sure to interpret and extract them from the forum posts accordingly.
- First format the response as:
- "Post 1:
  - mechanism 1: {max 10 word description}
  - mechanism 2: {max 10 word description}
  - etc (if applicable)
- Post 2:
  - mechanism 1: {max 10 word description}
  - mechanism 2: {max 10 word description}
  - etc (if applicable)"
- Then please reformat the response as an excel spreadsheet with in the first column the post number, the second the mechanism number, and the third the mechanism description. Please format the table as csv and put the table in code block.
- Forum posts: ""*[insert forum posts]*""

After prompt engineering, 4 new OpenAI accounts were created and used on 4 different computers to run the large number of queries. As such, the accounts were used to conduct tasks on a different computer simultaneously or repeat tasks on the same computer within an interval of 48 hours. New conversations were started for each task and ended when all data pertaining to the task in the query were processed (ie, 1 run of the task). Conversations were logically labeled to enable ease of lookup at a later stage (eg, Task1_account1_run3 and Task5_account4_run1).

Query Process

Data Preparation

Data preparation involved creating a data set of condensed meaning units based on the change mechanisms identified in the forum posts. Posts from the original published sugar study were randomized using a random number generator function in Microsoft Excel. The forum posts were included alongside step-by-step instructions to identify potential change mechanisms from the provided posts (Textbox 1). The first step included directions to extract the mechanisms for reducing sugar consumption from the posts and summarize the change mechanisms in a brief description (up to 10 words). The resulting brief descriptions functioned as the condensed meaning units to be used for coding scheme development. In the next step, ChatGPT was instructed to reformat the response into a table with post number, mechanism number (within the post), and the brief description of each identified change mechanism. All output tables in the conversation were transferred to a Microsoft Excel spreadsheet, which received a label consistent with the conversation name. The data preparation query was repeated to emulate a manual data preparation phase involving multiple coders and enabling comparison of brief descriptions across coded data sets. The query was repeated 10 times in 10 different conversations, thereby resulting in 10 data sets reflecting 10 different coders.

Comparison of the 10 data sets generated by ChatGPT with human coding was based on the change mechanisms identified in 108 forum posts (ie, a 20% data subset). In total, 3 researchers (RB, SSM, and SNR) compared the condensed meaning units with the forum posts and indicated which condensed meaning units correctly reflected a change mechanism described in the forum posts and which condensed meaning units incorrectly identified a change mechanism from the posts. Data sets were double coded, and disagreements were discussed until consensus was achieved. The data set with the highest percentage of correctly identified change mechanisms across all identified change mechanisms (ie, the best precision) was selected for further use in the analysis. Specifically, the 20% data subset was used for the development of the inductive and unconstrained deductive coding schemes, after which the developed coding schemes were applied to the full data set of change mechanisms. For the structured deductive approach, the full data set was coded directly into a TDF coding matrix. Change mechanisms

XSL•FO
RenderX

in the selected data set were given a unique identifier from 1 to *n* to enable data linkage in subsequent tasks.

Inductive Approach: Development and Application of Data-Driven Coding Schemes

The inductive approach started with a task that reflected the development of a coding scheme where categories are derived from the data [3]. The prompt developed for this task instructed ChatGPT to organize meaning units from the subset of data (20%) into categories based on underlying patterns in the data. The prompt included explicit instructions to ensure that there was no overlap between the categories. Furthermore, the instruction and answer template indicated that ChatGPT had to provide a 20-word label to define each category. The task was repeated in 10 new conversations to create 10 versions of an inductively developed coding scheme.

The next task in the inductive approach reflected the classification of the full data set into the inductively developed coding scheme. The output from the previous task was inserted into a prompt in which instructions specified that change mechanisms were to be classified under the best-matching category in accordance with the 20-word category definitions. ChatGPT was also instructed to format the output as a table that listed the mechanism unique identifier, the brief description of the change mechanism, and the best-matching category from the coding scheme. As there were 10 inductively developed coding schemes, the standard prompt was also adapted 10 times to reflect the various versions of the coding scheme. To reflect a process with 10 independent coders, each coding scheme was applied 10 times by starting 10 new conversations per coding scheme.

Unconstrained Deductive Approach: Development and Application of Theory-Driven Coding Schemes

The deductive approach started with the development of an unconstrained coding scheme that was informed by the TDF. The prompt instructed ChatGPT to identify domains (akin to categories in the inductive approach) from the TDF and redefine these domains to reflect the current data set subset. For the first step, the instructions specified grouping all mechanisms under the best-aligning domain, with explicit statements that a group could only reflect 1 domain and that each change mechanism could only be listed once. The instructions also specified listing each change mechanism under the domain in which it was grouped to ensure that each mechanism was listed once. Prompt engineering illustrated that although ChatGPT training data included sources on the TDF, prompt responses tended to incorporate fabricated or adapted domain names. Therefore, the prompt included a list of all 14 domains to remove any ambiguity in the instructions. For the second step, ChatGPT was instructed to provide a 20-word definition for each domain based on the group of change mechanisms it listed under the domain in the first step. The instructions further detailed that domains not identified in step 1 should be labeled as “N/A.” The task was repeated in 10 new conversations to create 10 versions of a deductively developed coding scheme.

The next task in the deductive approach reflected classification of the full data set into the domains of the deductively developed unconstrained coding scheme. The prompt instructed ChatGPT that change mechanisms were to be classified under the best-matching category in accordance with the 20-word category definitions. There was no reference to the TDF in the instructions to ensure that ChatGPT based its response on each provided coding scheme and to minimize the risk of unintentionally infusing the instructions with additional information based on ChatGPT’s preexisting knowledge of the TDF. The prompt was adapted 10 times (ie, once for each of the unconstrained coding schemes), and 10 new conversations were started per coding scheme (to mimic a process with 10 coders).

Structured Deductive Approach: Coding Directly Into TDF Coding Matrix

The full data set was classified directly into the domains of the TDF using a coding matrix that listed all 14 TDF domains [39] but was empty of definitions for the domains. Thus, this approach did not consider the underlying data and solely relied on ChatGPT’s preexisting knowledge of the TDF and how it would describe the mechanisms of sugar reduction. The prompt included instructions to classify each change mechanism under only the TDF domain that best matched each mechanism and specified the response format to be in the form of a table with columns for the mechanism unique identifier, short description, and the best-matching TDF domain. Direct coding into the TDF was conducted over 10 new conversations, reflective of a process with 10 coders.

Postprocessing

The output of conversations applying the coding schemes to the full data set was manually transferred to Microsoft Excel spreadsheets and labeled in accordance with the conversation names. As described previously [49], ChatGPT-generated data commonly require various postprocessing steps to clean the data and check whether responses are in line with the instructions. Checks were performed on all output to see whether the change mechanism descriptions in the output were identical to those included in the prompt. Doing so revealed instances of hallucination in which ChatGPT altered brief descriptions in the output generated after lengthier prompts. Where hallucinations were identified, ChatGPT was reprompted with the instructions followed by only those change mechanisms that were incorrectly processed. Furthermore, postprocessing involved minor adjustments to clean category and domain names in which the full label was not reported. In some conversations in which coding was directly into the TDF, the output showed instances in which >1 domain was allocated per change mechanism. In these cases, the first of the annotated domains was selected for use in data analysis. Finally, the clean data sets were aggregated by coding scheme version. This entailed creating 1 spreadsheet per coding scheme where the first column reflected the mechanism unique identifier, the second column reflected the brief descriptions, and the remaining 10 columns reflected the category or domain name allocated to the change mechanism by the conversation in which the respective coding scheme was applied.

XSL•FO
RenderX

Quality Evaluation

Results from the preparation phase were presented as number of change mechanisms identified in the subset and number that ChatGPT correctly identified against human coding. We also present the total number of change mechanisms identified in the full data set across each of the conversations. Results from the development of the inductive and unconstrained deductive coding schemes were presented as number of categories or domains identified and median and range of label word count per coding scheme version. Furthermore, the content of the labels was inspected to check whether coding schemes were in line with the instructions (eg, no overlap in categories or labels). The structured deductive approach was not evaluated on any of these metrics as the coding matrix contained all 14 TDF domains and lacked labels with domain definitions.

The reliability of the coding schemes was evaluated by calculating the Fleiss κ [50] for intercoder agreement with >2 coders. Specifically, intercoder agreement calculation was performed for each version of the coding scheme developed using ChatGPT and the unstructured coding matrix by comparing allocated codes in the aggregated data sets. κ statistics were calculated per coding scheme (ie, overall κ score) and per category (or domain) within the coding scheme (ie, category-specific κ score) to compare reliability across and within coding schemes. κ scores of <0 were interpreted as poor agreement, scores from 0.00 to 0.20 were interpreted as slight agreement, scores from 0.21 to 0.40 were interpreted as fair agreement, scores from 0.41 to 0.60 were interpreted as moderate agreement, scores from 0.61 to 0.80 were interpreted as substantial agreement, and scores from 0.81 to 1.00 were interpreted as almost perfect agreement [51].

Furthermore, frequency data were generated for each of the coding schemes and the coding matrix. To do so, the mode across the 10 ChatGPT coders per change mechanism was retrieved for each of the aggregated data sets (ie, the data sets that combined the annotated data from 10 conversations). This meant that frequency data reflected a majority agreement among the ChatGPT coders rather than a consensus decision-making approach. Thus, frequency tables were created per coding scheme after a decision-making approach where the final classification of the mechanisms was based on the mode of the annotated data per applied coding scheme. Frequency data and κ scores were obtained using Stata software (version 18.0; StataCorp).

Results

Evaluation of Preparation Phase

As shown in Table 1, the number of change mechanisms identified from forum posts varied across data sets (ie, the ChatGPT conversations from the data preparation phase), ranging from 571 to 623 condensed meaning units. Precision rates between ChatGPT and human coding ranged between 0.66 and 0.88. In total, 2 data sets yielded a precision of 0.88, which meant that a decision needed to be made on which one to bring forward to the next phase. We selected the data set that identified the larger number of change mechanisms based on the data subset (n=127) as this would enable a greater number of change mechanisms to be incorporated in the development of the coding schemes.

Table 1. Precision in identifying change mechanisms during the preparation phase.

Data set generated during data preparation	Change mechanisms identified in the full data set, n	Change mechanisms identified in the 20% data subset, n	Mechanisms correctly identified from the subset, n	Precision
Data set 1	584	128	111	0.87
Data set 2	587	119	105	0.87
Data set 3[a]	585	127	112	0.88
Data set 4	602	118	104	0.88
Data set 5	571	131	110	0.84
Data set 6	616	128	109	0.85
Data set 7	619	132	114	0.86
Data set 8	584	124	107	0.86
Data set 9	623	135	115	0.85
Data set 10	591	121	101	0.66

[a]Data set selected for the development of coding schemes.

Evaluation of the Inductive Approach

The results of the inductive approach are presented in Table 2. The inductively developed coding schemes contained a variable number of categories, ranging from 5 to 13. The category labels were overall lengthier than instructed, with a median label word count of 15 to 51 across coding schemes. Overall κ scores indicated substantial or almost perfect intercoder agreement for all coding schemes. Most definitions included specific examples of change mechanisms. Categories with labels reflecting a mixed or residual group of change mechanisms had category-specific κ scores indicating moderate or less than moderate intercoder agreement. Frequency tables based on a majority agreement decision-making approach showed that categories reflecting substituting sugary products (16%-29%) and gradually reducing

XSL•FO
RenderX

sugar consumption (12%-20%) and a joint category reflecting both substituting and reducing sugar (41%) were the most frequently mentioned categories in the final annotated data sets. Multimedia Appendix 2 provides a detailed overview of all inductively developed coding schemes, related metrics, and frequency tables.

Table 2. Intercoder agreement for the application of the coding schemes from the inductive approach.

Aggregated data sets by version of data-driven coding scheme	Categories in coding scheme, n	Label word count, median (range)	Intercoder agreement, overall κ[a]
Version 1	10	21 (15-25)	0.76
Version 2	9	26 (13-34)	0.69
Version 3	5	28 (24-29)	0.79
Version 4	10	30 (21-36)	0.84
Version 5	9	15 (12-20)	0.77
Version 6	6	28 (23-36)	0.83
Version 7	10	25 (19-31)	0.77
Version 8	10	16 (11-22)	0.72
Version 9	13	51 (10-69)	0.77
Version 10	5	17 (15-19)	0.72

[a]$P<.001$ for intercoder agreement for all versions of the coding scheme.

Coding scheme 4 had the best intercoder agreement overall (κ=0.84; $P<.001$) and was selected for further evaluation. As indicated in Table 3, intercoder agreement on the categories within this coding scheme was almost perfect (ie, κ>0.80; $P<.001$) for all but 2 categories. Particularly high category-specific intercoder agreement (κ=0.91; $P<.001$) was observed for a category that included specification of what was not included in the category (ie, “This approach contrasts with gradual reduction or moderation strategies”). There was overlap in the names of 2 categories (ie, “Sugar alternatives and substitutes” and “Substitution and replacement approaches”); however, the category definitions clearly delineated these categories, and both categories had almost perfect intercoder agreement (κ=0.84; $P<.001$).

XSL•FO
RenderX

Table 3. Inductively developed coding scheme with best overall intercoder agreement.

Category	Definition	Category-specific intercoder agreement, κ[a]	Frequency distribution—change mechanisms coded into the category[b] (n=585), n (%)
Psychological and behavioral strategies	Employing various psychological and behavioral techniques to change sugar consumption, such as seeking support, addressing addiction, recognizing the problem, planning ahead, and finding alternative distractions.	0.83	116 (19.8)
Substitution and replacement approaches	Replacing sugary foods and drinks with healthier alternatives, including options without added sugars; high-fiber carbs; fruits; nuts; or drinks like cinnamon tea, black coffee, or soda water.	0.84	111 (19)
Gradual reduction and moderation methods	Gradually reducing sugar consumption over time by reducing portion sizes, gradually decreasing sugar in tea/coffee, or incorporating a gradual process for adapting to reduced sugar. Moderation and small daily changes are emphasized.	0.85	85 (14.5)
Knowledge- and awareness-based approaches	Gaining knowledge about the harmful effects of excessive sugar consumption, checking sugar content in products, reading articles for information, and being aware of hidden sugars in various food products. Seeking resources and advice is also encouraged.	0.87	65 (11.1)
Environmental and practical strategies	Implementing changes in the environment to support reduced sugar consumption, such as changing grocery shopping habits, keeping cabinets stocked with healthy snacks, discarding carbs-filled foods, and not buying sugary items in the first place.	0.88	51 (8.7)
Health and well-being focus	Emphasizing the benefits of reduced sugar consumption on energy and overall health, incorporating nutrient-rich foods, ensuring adequate sleep, exercising to reduce stress, and considering health consequences and diabetes complications as motivation.	0.72	40 (6.8)
Elimination and cold turkey approaches	Completely quitting sugar consumption abruptly or going "cold turkey" for better health. This approach contrasts with gradual reduction or moderation strategies.	0.91	35 (6)
Support and community engagement	Seeking parental guidance, useful advice, moral support, weight loss buddies, or support from others. Sharing success stories and helping others break addiction are also emphasized.	0.95	30 (5.1)
Sugar alternatives and substitutions	Exploring and using various sugar alternatives and substitutes, including natural sweeteners such as stevia, honey, or Swerve, as well as incorporating naturally sweet options such as fruit.	0.84	28 (4.8)
Personal determination and accountability	Making a firm decision to quit sugar consumption, acknowledging weak moments and impulsive behavior, taking small steps, acknowledging the time and trial-and-error process, building willpower, and using tools such as MyFitnessPal to track sugar intake.	0.67	24 (4.1)

[a]$P<.001$ for intercoder agreement for all categories of the coding scheme.

[b]On the basis of a majority agreement decision-making approach across ChatGPT coders (ie, by selecting the mode of the codes per change mechanism).

We also examined the content of coding scheme version 2, which had the poorest overall intercoder agreement (κ=0.69; $P<.001$). The poorest category-specific agreement (κ=0.46; $P<.001$) within this coding scheme was for a miscellaneous category. Furthermore, overlap was observed in the definitions, with similar examples provided for multiple categories. For example, "seeking advice and support," "seeking support from others," and "seeking parental guidance" appeared in 3 different categories, and "avoiding purchasing sugary foods" and "changing shopping habits" appeared in 2 different categories.

Evaluation of the Unconstrained Deductive Approach

Results from the unconstrained deductive approach are presented in Table 4. None of the developed TDF coding schemes contained all 14 domains. The number of domains identified from the data subset was variable, ranging from 6 to 10. Definition word counts were in line with the prompt instructions

XSL•FO
RenderX

for all coding schemes (median between 12 and 17), with maximum definition word counts mostly being close to 20 words. The extent to which there was intercoder agreement overall was moderate or substantial across coding schemes.

Table 4. Intercoder agreement for the application of the coding schemes from the unconstrained deductive approach.

Aggregated data sets by version of theory-driven coding scheme	Domains, n	Label word count, median (range)	Overall intercoder agreement, κ[a]
Version 1	10	16 (12-19)	0.58
Version 2	8	13 (12-18)	0.62
Version 3	7	14 (13-17)	0.52
Version 4	7	16 (14-24)	0.73
Version 5	10	12 (9-13)	0.73
Version 6	10	15 (14-22)	0.53
Version 7	7	12 (11-19)	0.53
Version 8	9	14 (13-17)	0.62
Version 9	6	17 (13-19)	0.60
Version 10	10	16 (12-22)	0.58

[a]*P*<.001 for intercoder agreement for all versions of the coding scheme.

Across all TDF coding schemes, domain-specific κ scores ranged from 0.06 to 0.89 (Multimedia Appendix 3). The domains “Beliefs about consequences,” “Environmental context and resources,” and “Social influences” were identified in all versions of the coding scheme, whereas the domains “Optimism,” “Reinforcement,” and “Social and professional role and identity” were identified in none. The domain “Social influence” had the highest domain-specific intercoder agreement (κ=0.79-0.89; *P*<.001), whereas the domain “Memory, attention, and decision processes” had the lowest domain-specific intercoder agreement (κ=0.06-0.63; *P*<.001). There was no overlap in domain labels within coding schemes. Multimedia Appendix 4 provides a detailed overview of all TDF coding schemes, related metrics, and frequency tables. Of the coding schemes with the highest intercoder agreement (κ=0.73; *P*<.001), coding scheme 5 had the greatest number of TDF domains and, therefore, was selected for further evaluation. As indicated in Table 5, domain-specific intercoder agreement on 6 out of 10 domains within this coding scheme was substantial or near perfect.

XSL•FO
RenderX

Table 5. Unconstrained deductive coding scheme with best overall intercoder agreement.

Domain of the Theoretical Domains Framework	Definition	Domain-specific intercoder agreement, κ[a]	Frequency distribution—change mechanisms coded into the domain[b] (n=585), n (%)
Behavioral regulation	Implementing strategies, techniques, and habits to regulate and control sugar consumption.	0.77	369 (63.1)
Knowledge	Acquiring information and understanding the effects of sugar on health and nutrition.	0.79	54 (9.2)
Beliefs about consequences	Understanding and acknowledging the positive outcomes of reducing sugar consumption motivate behavior change.	0.75	48 (8.2)
Social influences	External factors and support from others influence sugar consumption and dietary choices.	0.87	38 (6.5)
Environmental context and resources	The impact of the physical environment and available resources on sugar consumption habits.	0.61	34 (5.8)
Beliefs about capabilities	Confidence in one's ability to change sugar consumption habits and overcome addiction.	0.57	24 (4.1)
Goals	Setting specific targets and objectives related to reducing sugar intake.	0.68	10 (1.7)
Emotion	Emotional factors that influence sugar cravings and behavior change.	0.35	5 (0.9)
Intention	Having a clear purpose and determination to change sugar consumption behavior.	0.59	3 (0.5)
Memory, attention, and decision processes	Cognitive processes that involve memory, attention, and decision-making in relation to sugar consumption.	0.13	0 (0)

[a]$P<.001$ for intercoder agreement for all domains of the coding scheme.

[b]On the basis of a majority agreement decision-making approach across ChatGPT coders (ie, by selecting the mode of the codes per change mechanism).

We compared domain labels across the coding schemes with the best and poorest intercoder agreement. This revealed that domains with better domain-specific intercoder agreement included examples of change mechanisms in their labels. An example of this was observed for the domain "Behavioral regulation," labeled as "The self-directed process of monitoring, controlling, and modifying behaviors related to sugar consumption" in one version (κ=0.43; $P<.001$) and more descriptively labeled as "Techniques and strategies used to regulate and control sugar consumption, including portion control, substitution, gradual reduction, and self-monitoring" in another version (κ=0.77; $P<.001$).

Evaluation of the Structured Deductive Approach

This approach used a structured coding matrix in which change mechanisms were coded directly into TDF domains without first specifying domain labels. This meant that, instead of 10 different coding schemes each applied using 10 different ChatGPT conversations, the structured deductive approach applied the single published coding matrix using 10 different conversations. The overall κ scores for the coding matrix indicated substantial intercoder agreement (κ=0.66; $P<.001$). As indicated in Table 6, domain-specific intercoder agreement on 7 domains of the TDF coding matrix had substantial or near-perfect agreement, and 4 domains had moderate agreement. The highest domain-specific intercoder agreement was observed for the domain "Social influences" (κ=0.85; $P<.001$), which similarly yielded almost perfect intercoder agreement in all but one of the unconstrained coding schemes (as per the previous section). The lowest domain-specific intercoder agreement was observed for the domain "Optimism" (κ=0.33; $P<.001$), which was not featured in any of the unconstrained coding schemes, and for the domain "Emotion" (κ=0.35; $P=.55$), which yielded fair to substantial intercoder agreement in the unconstrained deductively developed coding schemes.

XSL•FO
RenderX

Table 6. Intercoder agreement on the structured deductive approach using the Theoretical Domains Framework (TDF) coding matrix.

TDF domain	Domain-specific intercoder agreement, κ[a]	Frequency distribution—change mechanisms coded into the domain[b] (n=585), n (%)
Behavioral regulation	0.66	281 (48)
Beliefs about consequences	0.73	82 (14)
Environmental context and resources	0.56	82 (14)
Knowledge	0.79	46 (7.9)
Social influences	0.85	29 (5)
Memory, attention, and decision processes	0.50	19 (3.2)
Social and professional role and identity	0.73	18 (3.1)
Beliefs about capabilities	0.54	12 (2.1)
Goals	0.74	7 (1.2)
Emotion	0.36[c]	4 (0.7)
Intentions	0.85	2 (0.3)
Optimism	0.33	2 (0.3)
Skills	0.46	1 (0.2)
Reinforcement	0	0 (0)

[a]$P<.001$ for intercoder agreement for all domains of the coding scheme unless otherwise stated.

[b]On the basis of a majority agreement decision-making approach across ChatGPT coders (ie, by selecting the mode of the codes per change mechanism).

[c]$P=.51$.

Discussion

Principal Findings

This study is among the first to use ChatGPT to automate a range of tasks related to qualitative content analysis of web-based data on behavior change. Preparation for the analysis process was done by identifying relevant change mechanisms from forum data on reducing sugar consumption, which we did through ChatGPT with an estimated 88% precision rate. On the basis of a subset of change mechanisms, 10 coding schemes were developed using an inductive approach in which categories and category labels were informed by the data. Another 10 coding schemes were developed using an unconstrained deductive approach categories that reflected relevant domains of the TDF and were relabeled in line with the current data on sugar reduction. The developed coding schemes largely followed prompt instructions but were highly variable in the number of categories across and within approaches. Using ChatGPT to code the full change mechanism data set into each coding scheme showed moderate to almost perfect intercoder agreement, where the intercoder agreement of the inductively developed coding schemes was generally superior to that of the deductively developed coding schemes. A structured deductive approach was also applied by coding directly into the original TDF coding matrix without specifying domain labels to reflect the current data set. The overall intercoder agreement for this approach exceeded that for most coding schemes from the unconstrained deductive approach, but it was lower than the overall agreement observed for the inductively developed coding schemes.

Comparison With Other Studies

There have been a few exploratory studies using ChatGPT to analyze small data sets with analysis approaches that bear similarities to the inductive and deductive content analysis approaches used in our study. For example, one study used ChatGPT to conduct reflexive thematic analysis on a feature newspaper article about post–COVID-19 condition [35]. Analysis of the text was data driven as ChatGPT was prompted to generate a list of categories featured in the data, which, in contrast to the approach in our study, was then compared to a list of categories independently developed by a researcher. The findings revealed that the lists were largely similar, although the ChatGPT-generated list included more focused categories than the broader human-developed categories. Both lists were combined to refine the categories, and ChatGPT was queried to confirm the fit of the refined categories with the data, reflective of a negotiation step when combining category lists. Other research has used ChatGPT to assess problem-solving through content analysis of 40 short internet chats in which university students aimed to solve a mathematical issue [33]. Using an inductive approach, the authors prompted ChatGPT to generate a list of categories with description (comparable to the coding schemes in our study) for each chat and provide an overall problem-solving score based on the extent to which the categories were featured in the chat. The reliability of ChatGPT output over time was evaluated by repeating exact prompts at a different date in new conversations, which revealed only a moderately positive correlation between the output at different time points. Our study showed more promising results as we observed substantial to almost perfect intercoder agreement when evaluating the reliability of the inductively developed

XSL•FO
RenderX

coding schemes by repeating prompts in conversations across time points and OpenAI accounts.

The aforementioned data from problem-solving chats were also analyzed using a deductive approach [33]. To do so, the authors prompted ChatGPT with questions reflecting categories of a theory-driven coding scheme that had previously been applied by researchers to code the data. Overall problem-solving scores were calculated from the number of categories that were featured in a chat. Prompts were repeated at 2 time points to calculate intracoder agreement. The findings showed 90% intracoder agreement in output generated on separate occasions. Our study similarly evaluated the reliability of ChatGPT output by repeating prompts in conversations across time points and OpenAI accounts, albeit across 10 conversations as opposed to 2. Interestingly, our findings were more promising for the inductive approach, whereas we observed only moderate to substantial overall agreement for the coding schemes from the unconstrained deductive approach.

The deductive approach with a structured coding matrix in our study is somewhat similar to the approach taken in another study that evaluated feedback on a university course [34]. The study used ChatGPT to code 200 student comments directly into a predefined list of categories that were based on the literature but not accompanied by a label to define the categories. Researchers subsequently double checked the ChatGPT-annotated data and agreed with 85% of the allocated codes. An alternative double check was performed by prompting ChatGPT to rate how well an allocated category reflected the content of the comment. Approximately 10% of the annotated data received low ratings, suggesting incorrectly coded data. These findings do not necessarily indicate an error in ChatGPT coding; rather, they highlight specific categories of the coding matrix that might need to be refined.

Our objective was to explore the utility of ChatGPT in conducting qualitative content analysis with an emphasis on diminishing the human workload and time spent on analysis tasks and the error prone–ness related to human fatigue when analyzing large data sets. In line with this, human input in our study was focused on prompt engineering and data preparation, whereas the inductive and deductive analysis was conducted by prompting ChatGPT with a set of instructions related to those tasks. As such, it was beyond our scope to combine ChatGPT-generated coding schemes with a human-developed coding scheme or compare ChatGPT-annotated data with human-annotated data. However, as we used secondary data, it is possible to compare findings related to our secondary aim with the change mechanisms identified in the original study that relied solely on manual coding [41]. The original study identified 25 different categories that were organized into 4 phases of readiness to change. The most frequent categories were similar to the inductive approach categories “Substance substitution,” “Knowledge and information,” and “Avoidance.” However, the most frequently occurring category in the inductive and deductive approaches was related to behavioral regulation, which the previous study coded into smaller categories that reflected the study’s overarching theory and coding scheme.

Implications and Future Directions

Our study suggests that intercoder agreement using the inductive approach was superior to that achieved through deductive approaches. This finding aligns with the fundamental principles underlying each approach as the inductive approach, developed solely on the current data set, aims to describe its content. This approach is particularly valuable in novel or emerging areas of study or when preexisting data, concepts, or theories are scarce [3]. In fields such as behavior change, numerous explanatory and predictive models and theories offer insights into not only what is happening but also why. Our study indicates that coding within a model or theory yields higher intercoder agreement when the coding scheme is tailored to reflect the current data set. However, adjusting a published coding matrix poses the risk of missing opportunities to advance widely applicable models such as the TDF, which has demonstrated universal utility [39,52]. Moreover, as a framework, the TDF is more explanatory than predictive, and some of its categories overlap (eg, planning appears in 2 different categories), complicating data analysis methods such as content analysis [53]. Future research may benefit from applying the insights from our study to behavior change theories characterized by more clearly defined categories, such as the theory of planned behavior, to further explore ChatGPT utility.

The difference between intercoder agreement for inductive and deductive approaches might be further explained by prompt engineering and category development used by ChatGPT. For example, category labels from the inductive approach mostly exceeded the maximum word count specified in the instructions but also commonly included examples, which was not the case for the labels from the deductive unconstrained approach. LLMs process data based on semantic relations in text, and as illustrated by the literature [54], queries with richer semantic data (eg, synonyms) increase LLM performance. Examples included in the category labels are likely semantically related to the short descriptions of change mechanisms, which might, in turn, result in more consistent annotation of the data. Future research is warranted to confirm whether coding schemes that include examples increase intercoder reliability across ChatGPT output.

Our findings underline the importance of including a step in the analysis process in which coding schemes are refined based on data that are to be coded as opposed to using a structured coding matrix. Depending on the topic investigated, certain elements of a theory or framework may be less applicable to the data but indicate new insights into the topic [3]. This may explain why domains that were not featured in the unconstrained coding schemes showed low domain-specific agreement when applying the structured coding matrix. Without the option to code into domains not featured in the data subset, the overall intercoder agreement using this approach may have been better. A noteworthy issue with the structured deductive approach was that, despite instructions to only code mechanisms into the best-matching domain, there were still considerable data across coding schemes allocated to more than one domain. The issue persisted despite adapting instructions for the task during the prompt engineering phase. This could be an artifact of the underlying TDF knowledge, which in some settings results in

XSL•FO
RenderX

an overlap between domains [53]. It might also be that, with the lack of domain labels and, thus, limited semantic information, instructions are more ambiguous, thereby leaving more room for interpretation and resulting in less consistent output across conversations.

Through a thorough process of prompt engineering, we developed a set of structured prompts that can easily be adapted, expanded, and tailored by other researchers intending to use ChatGPT for qualitative content analysis. The extent to which prompts need adaptation and further prompt engineering depends on the research topic and approach. Replicating this study on a different topic would require an adequate description of that topic, which might be limited to a substitution of the topic keywords used in the prompts (ie, "sugar consumption" and "change mechanisms"). Where other theories are selected for the inductive approach, it is advisable to check ChatGPT's familiarity with the theory. Depending on the topic and theory, we suggest specifying instructions with other notes on what to look out for during task execution. The use of appropriate synonyms is also recommended to improve ChatGPT performance [54] as terminology may vary across research groups, disciplines, and communities. With other versions of ChatGPT, token limit may not be a restriction, and we suggest that researchers experiment with adding more context to the prompt. To further automate the preparation task, comparison with human coding may be replaced by having ChatGPT rate the accuracy of previous output, as done in previous research [34].

The approaches used in this study can be tested on other data types, such as survey responses, interviews, or focus group discussions. However, the nature of the data (eg, depth vs breadth and level of structuredness) might affect ChatGPT performance on analysis tasks, and research is warranted to refine prompt engineering and assess the reliability of the output based on other data types. Regardless of the research approach, when using tools such as ChatGPT for qualitative analysis, researchers should be aware of ethical considerations to prevent harm, such as transparency, informed consent, data privacy, and potential biases in LLM training data or disseminated results [55,56]. It is warranted that such issues are considered properly to ensure that research projects abide by ethical standards.

It should be reiterated that our study focused primarily on the reliability of ChatGPT-annotated data without addressing the validity of the findings. Assessing the validity of the findings is an important area for future investigation. Validity checks could be incorporated through triangulation of investigators, for instance, by having experienced qualitative researchers develop coding schemes and compare these with coding schemes developed using ChatGPT. Alternatively, the prompt with step-by-step instructions, which we used to create a theory-driven coding scheme, enables a less time-intensive method to check the validity of the ChatGPT-generated coding scheme. In this type of prompt, the first part of the output facilitates the possibility to assess whether the data underlying each category in the generated coding scheme accurately reflect the categories. In addition, we encourage others using ChatGPT for qualitative content analysis to experiment with coding rules that may further increase the quality of the output and prevent overlap between categories. It should also be mentioned that we presented frequency data based on the coding schemes with the best intercoder agreement. However, the selection of the most optimal coding scheme may depend on other considerations, including expert reflection of categories and labels generated by ChatGPT, number of categories within a coding scheme, and the need to incorporate certain categories or mix inductive and deductive approaches. Moreover, the appropriateness of the analysis methods and techniques used is dependent on the type of topic as well as the character and volume of the available data [19].

Limitations

The colloquial phrase "garbage in, garbage out" can be seen as a general rule in machine learning, meaning that any output retrieved is only as good as the data underlying the output [57]. As we followed a thorough process of prompt engineering and evaluation of the output did not reveal concerning deviations from the prompt instructions, we are relatively confident that the quality of the output was not substantially diminished by the instructions for the tasks. However, it should be noted that sourcing the data to be analyzed from web-based forums may have implications for the quality of the output. Forum data are user generated and, by nature, less focused than data generated though targeted questions (eg, as is the case with interview and survey data), potentially leading to convoluted data [58]. In our study, this issue was partly circumvented by using secondary data that met certain eligibility criteria to ensure that they related to lived experiences of people trying to change sugar consumption [41]. Still, using these data to generate a condensed set of change mechanisms did result in the inclusion of nontarget data, as evident from our estimate of 88% precision in identification of change mechanisms, which we considered acceptable for use in the inductive and deductive analysis approaches.

Whether precision levels should be considered acceptable is dependent on the goal of the study and is often a trade-off with other metrics of model performance, including recall and accuracy [59]. When generating a data set of change mechanisms identified from the forum posts, we focused on precision to optimize the amount of relevant data to be used in creating and applying the coding schemes. Imperfect precision could have led to distorted categories in a coding scheme as instructions specified to group all mechanisms, regardless of relevance, in discreet categories. It may also have caused the creation of a miscellaneous category, observed in various versions of the inductively developed coding schemes. Similarly, coding nontarget change mechanisms into the coding schemes may have distorted the frequency data as these were also based on all data. We included instructions to code change mechanisms as not applicable in cases in which they did not fit into any category, and these data could be further examined to build new knowledge of the TDF and how it applies to a range of different contexts.

Conclusions

AI assistance has the potential to make a massive impact on research involving qualitative content analysis. As demonstrated in this study, ChatGPT can assist with each phase of the

XSL•FO
RenderX

methodology, from condensing data and developing coding schemes to double coding and facilitating consensus meetings. While we have shown that ChatGPT can perform these tasks largely without human oversight, there are many risks associated with this approach, with the key risk being that the output may reflect the worst of ChatGPT's coding and, therefore, may not accurately reflect the data or research questions. We recommend that human involvement is necessary, but it could be reduced to just 1 or 2 researchers across each phase. Such human involvement would largely ensure that prompt engineering and interpretation remain aligned with the research goals and context. This study provides the foundations for qualitative content analysis using ChatGPT and can be tested and further developed as new AI emerges.

Acknowledgments

Funding for the study was provided by the Health Research Council of New Zealand (grant 20/733). Artificial intelligence was used as part of the study design as reported in the *Methods* section of this paper.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RB contributed to conceptualization (supporting), data curation, formal analysis, investigation, methodology, validation, visualization, and writing—original draft preparation and review and editing, while SSM contributed to formal analysis, investigation, methodology, validation, and review and editing. NAD reviewed and edited the manuscript. SNR contributed to conceptualization (lead), funding acquisition, investigation, project administration, supervision, validation, and review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final prompts.
[DOCX File , 19 KB - jmir_v26i1e59050_app1.docx]

Multimedia Appendix 2
Inductively developed coding schemes and metadata.
[DOCX File , 44 KB - jmir_v26i1e59050_app2.docx]

Multimedia Appendix 3
Domain-specific κ scores per the Theoretical Domains Framework coding scheme.
[DOCX File , 17 KB - jmir_v26i1e59050_app3.docx]

Multimedia Appendix 4
Deductively developed coding schemes and metadata.
[DOCX File , 40 KB - jmir_v26i1e59050_app4.docx]

References

1. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: SAGE Publications; 1980.
2. Weber RP. Basic Content Analysis: Issue 49. Thousand Oaks, CA: SAGE Publications; 1990.
3. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
4. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
5. Fu J, Li C, Zhou C, Li W, Lai J, Deng S, et al. Methods for analyzing the contents of social media for health care: scoping review. J Med Internet Res 2023 Jun 26;25:e43349 [FREE Full text] [doi: 10.2196/43349] [Medline: 37358900]
6. Kleinheksel AJ, Rockich-Winston N, Tawfik H, Wyatt TR. Demystifying content analysis. Am J Pharm Educ 2020 Jan;84(1):7113 [FREE Full text] [doi: 10.5688/ajpe7113] [Medline: 32292185]
7. Mayring P. Qualitative content analysis: demarcation, varieties, developments. Forum Qual Sozialforsch 2019;20(3). [doi: 10.17169/fqs-20.3.3343]
8. Mayring P. Qualitative content analysis. Forum Qual Sozialforsch 2000;1(2).

XSL•FO
RenderX

9. Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. BMJ 2000 Jan 08;320(7227):114-116 [FREE Full text] [doi: 10.1136/bmj.320.7227.114] [Medline: 10625273]
10. Lindgren BM, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020 Aug;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632] [Medline: 32505813]
11. Hahn C. Doing Qualitative Research Using Your Computer: A Practical Guide. Thousand Oaks, CA: SAGE Publications; 2008.
12. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
13. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
14. Zamawe FC. The implication of using NVivo software in qualitative data analysis: evidence-based reflections. Malawi Med J 2015 Mar;27(1):13-15 [FREE Full text] [doi: 10.4314/mmj.v27i1.4] [Medline: 26137192]
15. Brooks J, McCluskey S, Turley E, King N. The utility of template analysis in qualitative psychology research. Qual Res Psychol 2015 Apr 03;12(2):202-222 [FREE Full text] [doi: 10.1080/14780887.2014.955224] [Medline: 27499705]
16. Haney W, Russell M, Gulek C, Fierros E. Drawing on education: using student drawings to promote middle school improvement. Sch Middle 1998;7(3):38-43 [FREE Full text]
17. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
18. Young IJ, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 2019 Dec;132:103971. [doi: 10.1016/j.ijmedinf.2019.103971] [Medline: 31630063]
19. Karmegam D, Ramamoorthy T, Mappillairajan B. A systematic review of techniques employed for determining mental health using social media in psychological surveillance during disasters. Disaster Med Public Health Prep 2020 Apr;14(2):265-272. [doi: 10.1017/dmp.2019.40] [Medline: 31272518]
20. Neale J. Iterative categorization (IC): a systematic technique for analysing qualitative data. Addiction 2016 Jun;111(6):1096-1106 [FREE Full text] [doi: 10.1111/add.13314] [Medline: 26806155]
21. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922. [doi: 10.1177/1609406919899220]
22. Hutson M. Could AI help you to write your next paper? Nature 2022 Nov;611(7934):192-193. [doi: 10.1038/d41586-022-03479-w] [Medline: 36316468]
23. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
24. Joffe H, Yardley L. Content and thematic analysis. In: Research Methods for Clinical and Health Psychology. London, UK: SAGE Publications; 2003:56-68.
25. Mayring P. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Social Science Open Access Repository. 2014. URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173 [accessed 2023-07-19]
26. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Libr Hi Tech News 2023 Feb 14;40(3):26-29. [doi: 10.1108/lhtn-01-2023-0009]
27. Raman R, Lathabhai H, Diwakar S, Nedungadi P. Early research trends on ChatGPT: insights from altmetrics and science mapping analysis. Int J Emerg Technol Learn 2023 Oct 04;18(19):13-31. [doi: 10.3991/ijet.v18i19.41793]
28. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/index/chatgpt/ [accessed 2023-08-02]
29. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature. 2023 Jan 24. URL: https://www.nature.com/articles/d41586-023-00191-1 [accessed 2023-07-20]
30. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
31. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol 2023 Mar 10;74(5):570-581. [doi: 10.1002/asi.24750]
32. Moskatel LS, Zhang N. The utility of ChatGPT in the assessment of literature on the prevention of migraine: an observational, qualitative study. Front Neurol 2023 Aug 17;14:1225223 [FREE Full text] [doi: 10.3389/fneur.2023.1225223] [Medline: 37662036]
33. Siiman LA, Rannastu-Avalos M, Pöysä-Tarhonen J, Häkkinen P, Pedaste M. Opportunities and challenges for AI-assisted qualitative data analysis: an example from collaborative problem-solving discourse data. In: Proceedings of the International Conference of Innovative Technologies and Learning. 2023 Presented at: ICITL 2023; August 28-30, 2023; Porto, Portugal. [doi: 10.1007/978-3-031-40113-8_9]
34. Katz A, Wei S, Nanda G, Brinton C, Ohland M. Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy. arXiv. Preprint posted online on May 9, 2023 [FREE Full text]

XSL•FO
RenderX

35. Hitch D. Artificial intelligence augmented qualitative analysis: the way of the future? Qual Health Res 2024 Jun;34(7):595-606. [doi: 10.1177/10497323231217392] [Medline: 38064244]
36. Tabone W, de Winter J. Using ChatGPT for human-computer interaction research: a primer. R Soc Open Sci 2023 Sep 13;10(9):231053. [doi: 10.1098/rsos.231053] [Medline: 37711151]
37. Neelakantan A, Weng L, Power B, Jang J. Introducing text and code embeddings. OpenAI. 2022 Jan 25. URL: https://openai.com/index/introducing-text-and-code-embeddings/ [accessed 2023-10-10]
38. Zhang H, Wu C, Xie J, Lyu Y, Cai J, Carroll JM. Redefining qualitative analysis in the AI era: utilizing ChatGPT for efficient thematic analysis. arXiv. Preprint posted online on September 19, 2023 [FREE Full text]
39. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
40. Brockman G, Eleti A, Georges E, Jang J, Kilpatrick L, Lim R, et al. Introducing APIs for GPT-3.5 Turbo and Whisper. OpenAI. URL: https://openai.com/index/introducing-chatgpt-and-whisper-apis/ [accessed 2023-08-02]
41. Rodda SN, Booth N, Brittain M, McKean J, Thornley S. I was truly addicted to sugar: a consumer-focused classification system of behaviour change strategies for sugar reduction. Appetite 2020 Jan 01;144:104456. [doi: 10.1016/j.appet.2019.104456] [Medline: 31525418]
42. Exceptions to activities requiring AUTEC approval (6). Auckland University of Technology. URL: https://www.aut.ac.nz/research/researchethics/guidelines-and-procedures#6 [accessed 2024-05-03]
43. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023 May 18;6(1):9 [FREE Full text] [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
44. Chen S, Li Y, Lu S, Van H, Aerts HJ, Savova GK, et al. Evaluation of ChatGPT family of models for biomedical reasoning and classification. arXiv. Preprint posted online on April 5, 2023 [FREE Full text]
45. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 2023 Dec;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
46. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. Preprint posted online on February 21, 2023 [FREE Full text]
47. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. TechRxiv. Preprint posted online on May 4, 2023 2024 [FREE Full text] [doi: 10.36227/techrxiv.22683919.v2]
48. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on January 28, 2022 [FREE Full text]
49. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. Inf Fusion 2023 Nov;99:101861. [doi: 10.1016/j.inffus.2023.101861]
50. Fleiss JL, Levin B, Cho Paik M. Statistical Methods for Rates and Proportions. Hoboken, NJ: John Wiley & Sons; 2003.
51. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310]
52. Francis JJ, O'Connor D, Curran J. Theories of behaviour change synthesised into a set of theoretical groupings: introducing a thematic series on the theoretical domains framework. Implement Sci 2012 Apr 24;7:35 [FREE Full text] [doi: 10.1186/1748-5908-7-35] [Medline: 22531601]
53. Phillips CJ, Marshall AP, Chaves NJ, Jankelowitz SK, Lin IB, Loy CT, et al. Experiences of using the Theoretical Domains Framework across diverse clinical environments: a qualitative study. J Multidiscip Healthc 2015 Mar 18;8:139-146 [FREE Full text] [doi: 10.2147/JMDH.S78458] [Medline: 25834455]
54. Claveau V. Neural text generation for query expansion in information retrieval. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2021 Presented at: WI-IAT '21; December 14-17, 2021; Melbourne, Australia. [doi: 10.1145/3486622.3493957]
55. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
56. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
57. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
58. McKenna B, Myers MD, Newman M. Social media in qualitative research: challenges and recommendations. Inf Org 2017 Jun;27(2):87-99. [doi: 10.1016/j.infoandorg.2017.03.001]
59. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. J Med Internet Res 2016 Feb 26;18(2):e41 [FREE Full text] [doi: 10.2196/jmir.4738] [Medline: 26920122]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

LLM: large language model
TDF: Theoretical Domains Framework

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 31.03.24; peer-reviewed by I Kousa; comments to author 26.04.24; revised version received 08.05.24; accepted 18.06.24; published 25.07.24.

Please cite as:
Bijker R, Merkouris SS, Dowling NA, Rodda SN
ChatGPT for Automated Qualitative Research: Content Analysis
J Med Internet Res 2024;26:e59050
URL: https://www.jmir.org/2024/1/e59050
doi:10.2196/59050
PMID:

©Rimke Bijker, Stephanie S Merkouris, Nicki A Dowling, Simone N Rodda. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews

Kentaro Matsui[1,2], MD, PhD; Tomohiro Utsumi[2,3], MD; Yumi Aoki[4], PhD; Taku Maruki[5], MD; Masahiro Takeshima[6], MD, PhD; Yoshikazu Takaesu[7], MD, PhD

[1]Department of Clinical Laboratory, National Center Hospital, National Center of Neurology and Psychiatry, Kodaira, Japan
[2]Department of Sleep-Wake Disorders, National Institute of Mental Health, National Center of Neurology and Psychiatry, Kodaira, Japan
[3]Department of Psychiatry, The Jikei University School of Medicine, Tokyo, Japan
[4]Graduate School of Nursing Science, St. Luke's International University, Tokyo, Japan
[5]Department of Neuropsychiatry, Kyorin University School of Medicine, Tokyo, Japan
[6]Department of Neuropsychiatry, Akita University Graduate School of Medicine, Akita, Japan
[7]Department of Neuropsychiatry, Graduate School of Medicine, University of the Ryukyus, Okinawa, Japan

Corresponding Author:
Yoshikazu Takaesu, MD, PhD
Department of Neuropsychiatry
Graduate School of Medicine
University of the Ryukyus
207 Uehara
Nishihara
Okinawa, 903-0215
Japan
Phone: 81 98 895 3331
Email: takaesuy@med.u-ryukyu.ac.jp

Abstract

Background: The screening process for systematic reviews is resource-intensive. Although previous machine learning solutions have reported reductions in workload, they risked excluding relevant papers.

Objective: We evaluated the performance of a 3-layer screening method using GPT-3.5 and GPT-4 to streamline the title and abstract-screening process for systematic reviews. Our goal is to develop a screening method that maximizes sensitivity for identifying relevant records.

Methods: We conducted screenings on 2 of our previous systematic reviews related to the treatment of bipolar disorder, with 1381 records from the first review and 3146 from the second. Screenings were conducted using GPT-3.5 (gpt-3.5-turbo-0125) and GPT-4 (gpt-4-0125-preview) across three layers: (1) research design, (2) target patients, and (3) interventions and controls. The 3-layer screening was conducted using prompts tailored to each study. During this process, information extraction according to each study's inclusion criteria and optimization for screening were carried out using a GPT-4–based flow without manual adjustments. Records were evaluated at each layer, and those meeting the inclusion criteria at all layers were subsequently judged as included.

Results: On each layer, both GPT-3.5 and GPT-4 were able to process about 110 records per minute, and the total time required for screening the first and second studies was approximately 1 hour and 2 hours, respectively. In the first study, the sensitivities/specificities of the GPT-3.5 and GPT-4 were 0.900/0.709 and 0.806/0.996, respectively. Both screenings by GPT-3.5 and GPT-4 judged all 6 records used for the meta-analysis as included. In the second study, the sensitivities/specificities of the GPT-3.5 and GPT-4 were 0.958/0.116 and 0.875/0.855, respectively. The sensitivities for the relevant records align with those of human evaluators: 0.867-1.000 for the first study and 0.776-0.979 for the second study. Both screenings by GPT-3.5 and GPT-4 judged all 9 records used for the meta-analysis as included. After accounting for justifiably excluded records by GPT-4, the sensitivities/specificities of the GPT-4 screening were 0.962/0.996 in the first study and 0.943/0.855 in the second study. Further

XSL•FO
RenderX

investigation indicated that the cases incorrectly excluded by GPT-3.5 were due to a lack of domain knowledge, while the cases incorrectly excluded by GPT-4 were due to misinterpretations of the inclusion criteria.

Conclusions: Our 3-layer screening method with GPT-4 demonstrated acceptable level of sensitivity and specificity that supports its practical application in systematic review screenings. Future research should aim to generalize this approach and explore its effectiveness in diverse settings, both medical and nonmedical, to fully establish its use and operational feasibility.

(J Med Internet Res 2024;26:e52758) doi:10.2196/52758

KEYWORDS

systematic review; screening; GPT-3.5; GPT-4; language model; information science; library science; artificial intelligence; prompt engineering; meta-analysis

Introduction

Large language models (LLMs) with extensive parameters, honed on substantial textual data, have seen striking advancements recently. Following OpenAI's third-generation Generative Pre-trained Transformer (GPT-3), LLMs now possess advanced competencies in various natural language processing tasks [1]. Among these, ChatGPT, which is built on GPT-3.5—an iteration that improves upon GPT-3 by integrating both supervised and reinforcement learning techniques—has received particular attention [2,3]. GPT-3.5 has shown exceptional performance in the medical domain, achieving remarkable results on medical licensing examinations across different regions [4]. Furthermore, GPT-4, the successor to GPT-3.5, has exhibited superior performance [5], with its contextual understanding abilities potentially exceeding those of humans [6,7]. Beyond its use for language editing [8,9], both GPT-3.5 and GPT-4 have proven to be effective tools for analyzing and comprehending the abstracts of research papers, offering potential benefits in the screening process for systematic reviews.

Systematic reviews and subsequent meta-analyses bear crucial clinical significance. The screening of titles and abstracts is a crucial step in this process [10-13], often involving more than 1000 papers identified via targeted keyword searches [14]. This screening process can take approximately 1 hour for every 60-120 papers [10], which is a substantial drain on human and time resources. In addition, human error is inevitable in the screening process [15-17], and the number of such errors can increase as the amount of paper to be screened increases possibly due to fatigue and cognitive overload [18,19]. To mitigate this labor-intensive task, attempts have been made to use text mining and machine learning technologies [17,20-29]. Although these methods have successfully reduced the workload, they risk omitting relevant papers, which could result in a high false-negative rate. Specifically, several studies reported the exclusion of records that should have been included in the meta-analysis [20,21,23,29]. Consequently, using machine learning techniques, such as natural language processing, to assist with abstract screening has not yet become widely adopted [14,30]. For systematic reviews, maintaining high sensitivity for studies eligible for full-text assessment, ideally at 100% [10], is crucial if they are to be fully supplanted by an automated process.

With the advanced language-processing capabilities of GPT-3.5 and GPT-4 [2,5], there has been an expectation of achieving higher accuracy in screening processes. Kohandel Gargari et al [31] conducted title and abstract screening using GPT-3.5, but the sensitivity for identifying relevant papers remained at a maximum of 69%, even after attempting various prompt modifications. Khraisha et al [32] explored the use of GPT-4 across different systematic review processes and found that the sensitivity for title and abstract screening ranged between 42% and 50%. Guo et al [33] have also demonstrated the use of GPT-4 in title and abstract screenings; however, the sensitivity for relevant papers was limited to 76%, highlighting the challenge of unintentionally excluding necessary records. Notably, Tran et al [34] used GPT-3.5 for title and abstract screening with rigorous prompt adjustments, achieving a high sensitivity of 97.1% for relevant papers. While this high-sensitivity level might already be suitable for practical use in the systematic review process, its specificity was limited to 37.7% [34].

The aim of this study is to develop a title- and abstract-screening method using GPT-3.5 and GPT-4 that achieves as high a sensitivity as possible. Although the method of using GPT-3.5 by Tran et al [34] achieved high sensitivity for identifying relevant papers, we aim to maintain high sensitivity while also improving specificity through a unique approach that incorporates GPT-4. To achieve this, we subdivided the process of determining inclusion for systematic reviews [11] involving 3 layers of screening. By breaking down the screening process into multiple steps, each addressing a specific aspect, we aimed to optimize the performance of the language models. In this study, we regarded the results of human screening as the gold standard and calculated the sensitivity and specificity of the GPT-3.5 and GPT-4 screening results in comparison with them. Furthermore, we carefully examined the records that were erroneously excluded by GPT-3.5/GPT-4. This examination was conducted to assess the appropriateness of their exclusion.

Methods

Language Model Details

GPT-3.5 and GPT-4, LLMs used in this study, are accessible through ChatGPT. However, ChatGPT does not support processing multiple queries against the titles and abstracts of scholarly papers simultaneously. To address this limitation, we leveraged the application programming interfaces (APIs) of GPT-3.5 and GPT-4, known as gpt-3.5-turbo and gpt-4-turbo-preview, respectively [35].

XSL•FO
RenderX

For gpt-3.5-turbo, we used the most current model available, gpt-3.5-turbo-0125. This model could be used at a low cost of US $0.50 per 1M tokens for input and US $1.50 per 1M tokens for output, with approximately 750 tokens corresponding to 1000 words [36]. Similarly, for GPT-4, we used the latest model available, gpt-4-0125-preview, which was available at a cost of US $10.00 per 1M tokens for input and US $30.00 per 1M tokens for output [36].

Calling the GPT-3.5 and GPT-4 API

In this study, we used Google Spreadsheet and Google Apps Script to interface with the GPT-3.5 and GPT-4 APIs for batch processing. Specifically, we created the "GPT35" function to call the gpt-3.5-turbo-0125 API within Google Spreadsheet. Users can invoke this function by entering "=GPT35([prompt])" into a cell, enabling the intuitive batch processing of multiple titles and abstracts. Similarly, we established the "GPT4" function to access the gpt-4-0125-preview API.

Both the gpt-3.5-turbo-0125 and gpt-4-0125-preview have a parameter called "temperature," which introduces "variability" in the responses—the higher the temperature, the greater the randomness, with a range between 0 and 2 [37]. As described later in this study, the decision to include or exclude records was delegated to GPT-3.5 and GPT-4. At the preliminary trials, it was observed that setting the temperature above 0 resulted in varying responses from one trial to another. In addition, setting the temperature above 0 can lead to unexpected responses. When instructed to respond with either "E" (for the exclusion) or "I" (for the inclusion), if the temperature is 0, the output will be strictly "E" or "I." However, if the temperature is above 0, even if it is only 0.1, the response might be, for example, "The answer is 'E'." In light of these observations, and primarily to ensure reproducibility, this study fixed the temperature at 0 for all screenings. The Apps Script used in this study is shown in Multimedia Appendix 1.

Process of Screening and Prompt Engineering

Generally, in a systematic review, a comprehensive examination is conducted on studies that address a relevant clinical question. After a comprehensive literature search is performed to identify all potential studies for review, each record is assessed to determine whether it addresses the clinical question [11]. In this study, we used either GPT-3.5 or GPT-4 to assess the inclusion or exclusion of relevant papers at each of the following three layers: (1) research design, (2) target population, and (3) intervention and control [11]. Records not deemed for exclusion at any of these layers were classified as "included." We present the workflow of the process we conducted in Figure 1.

Figure 1. Three-layer screening process using GPT-3.5 and GPT-4 for literature review.

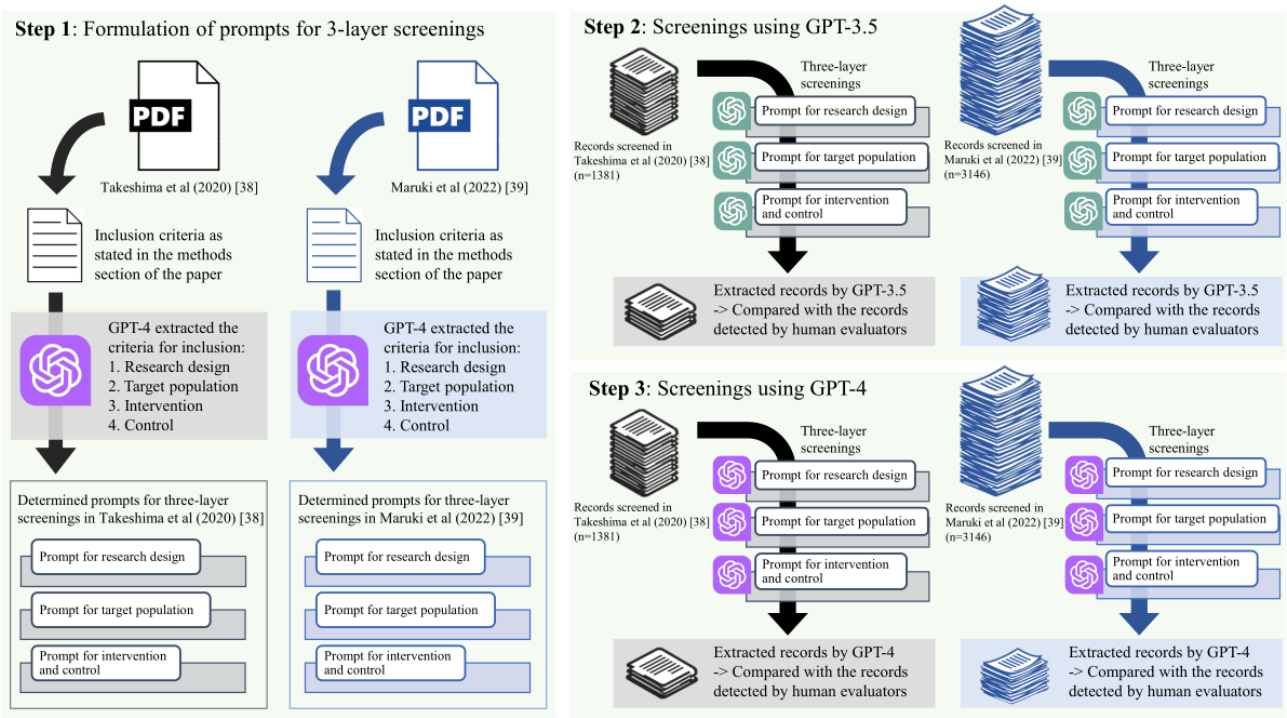

The characteristics of the 2 systematic review papers [38,39] used in this study are summarized in Table 1. The first paper by Takeshima et al [38] investigated the efficacy of bright light therapy in patients with bipolar disorder. In this study, the titles and abstracts of a total of 1381 records were initially screened in duplicate, with the task being divided between 2 pairs of independent evaluators. The first pair reviewed the initial 753 records, while the second pair assessed the remaining 628 records. Of these, 30 records were targeted for a full-text assessment, and eventually 6 records (encompassing 6 studies) were selected for meta-analysis. The second paper by Maruki et al [39] verified the difference in therapeutic effects between

XSL•FO
RenderX

the usage of 2 types: second-generation antipsychotics (SGAs) and mood stabilizers (MSs), versus the usage of either type alone, targeting patients with bipolar disorder. In this study, the titles and abstracts of a total of 3146 records were initially screened in duplicate, with the screening divided between 2 pairs of evaluators. The first pair reviewed the initial 1694 records, while the second pair evaluated the remaining 1452 records. Of these, 96 records were targeted for a full-text assessment, and eventually 9 records (encompassing 5 studies) were selected for meta-analysis. We used the data on the inclusion or exclusion decisions of each human evaluator made prior to reaching a consensus among evaluators.

Table 1. Characteristic of the 2 selected systematic review studies.

	Takeshima et al (2020) [38]	Maruki et al (2022) [39]
Clinical question	Is bright light therapy an effective and safe treatment for managing manic and depressive symptoms in patients with bipolar disorder, and can it also be used as a preventive measure for recurrent mood episodes?	Does the use of second-generation antipsychotics (SGA) or mood stabilizers (MS) as adjunctive therapy improve the efficacy and safety outcomes compared to their use as monotherapy in the treatment of bipolar depression?
Databases	Ovid MEDLINE, Cochrane Central Register of Controlled Trials, Embase, PsycINFO, and ClinicalTrials.gov	PubMed, Cochrane Central Register of Controlled Trials, and Embase
Number of records screened	1381	3146
Number of records for full-text assessment	30	96
Number of records (studies) included in quantitative synthesis	6 (6)	9 (5)

The screening process was divided into three layers: (1) research design, (2) target population, and (3) intervention and control. The prompts for each layer must be specifically tailored to each systematic review. At this point, manual prompt adjustments could lead to issues with reproducibility in future research. Therefore, in this study, we used GPT-4 (gpt-4-0125-preview, temperature=0) to automatically extract the information and generate the content for the prompts related to “research design,” “target population,” “intervention,” and “control.” The prompts used for extraction, along with the content defined for “research design,” “target population,” “intervention,” and “control,” are detailed in Textbox 1. In this study, we extracted information by inserting the text from the “inclusion criteria” paragraph of the Methods section of each paper into the specified location in the prompt (Textbox 1).

The structure of the prompts for each of the 3 layers is shown in Textbox 2. Within these prompts, we specified that if a decision cannot be made, records should be considered potentially eligible for full-text assessment and not excluded. In this study, the information supplied to GPT-3.5 and GPT-4 was limited to the titles and abstracts of the records; details such as authors, their affiliations, or journal names were not included in the prompts.

In the screening process using GPT-3.5 or GPT-4, we initially verified whether the research design of all records satisfied the inclusion criteria. For records not excluded in the first layer, we subsequently confirmed whether the target population aligned with the inclusion criteria. Moreover, for records that were not excluded in the first and second layers, we assessed whether both the intervention and control groups met the inclusion criteria (Figure 1).

XSL•FO
RenderX

Textbox 1. Prompts used for extracting information and the extracted details on research design, target population, intervention, and control from each study.

- Prompt for extracting information

  #Inclusion criteria

  [The descriptions of the inclusion criteria presented within the Methods section of each paper were inserted here]

  #Query

  The above is an inclusion criteria of a systematic review paper. Please describe the records included in the meta-analysis in a sentence: 1) research design, 2) target population, 3) intervention, and 4) control. Please list each item without including verbs.

  #Your answer

  1. Research design: [insert your answer here]
  2. Target population: [insert your answer here]
  3. Intervention: [insert your answer here]
  4. Control: [insert your answer here]

- Response to the above prompt

  Inclusion criteria of Takeshima et al (2020) [38], extracted by GPT-4

  1. Research design: Randomized controlled trials (RCTs) at the individual or cluster level, including crossover studies reporting results from the first period.
  2. Target population: Patients with a clinical diagnosis of bipolar disorder (BD), type I or type II.
  3. Intervention: Any kind of light therapy, including 'light therapy,' 'bright light therapy,' 'phototherapy,' or chronotherapy in any intensity and color.
  4. Control: Sham treatment (e.g., low-intensity light, dim red light, or negative ion) or treatment as usual (no light treatment).

  Inclusion criteria of Maruki et al (2022) [39], extracted by GPT-4

  1. Research design: Randomized controlled trials (RCTs) at the individual or cluster level, including crossover studies before crossover
  2. Target population: Participants diagnosed with bipolar I or II depression, including mixed features and/or rapid cycling.
  3. Intervention: Adjunctive therapy with second-generation antipsychotics (SGA) or mood stabilizers (MS) during baseline treatment with SGA or MS.
  4. Control: Adjunctive therapy with a placebo during baseline treatment with second-generation antipsychotics (SGA) or mood stabilizers (MS).

XSL•FO
RenderX

Textbox 2. The structure of the prompts used for each of the 3 layers in the screening process.

1. Prompt for research design

   #Title and abstract

   Title: [*Title of the record was inserted here*]

   Abstract: [*Abstract of the record was inserted here*]

   #Research design

   [*The 'research design' specified in Textbox 1 was inserted here*]

   #Query

   You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper.

   Does the paper with the above title and abstract meet the specified research design? If yes, highly suspected, or difficult to determine, answer 'I'. If not, answer 'E'.

   #Rules

   You can reply using only 'E' or 'I'.

   #Your answer:

2. Prompt for target population

   #Title and Abstract

   Title: [*Title of the record was inserted here*]

   Abstract: [*Abstract of the record was inserted here*]

   #Target population

   [*The'target population' specified in Textbox 1 was inserted here*]

   #Query

   You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper.

   Does the paper with the above title and abstract meet the specified target population? If yes, highly suspected, or difficult to determine, answer 'I'. If not, answer 'E'.

   #Rules

   You can reply using only 'E' or 'I'.

   #Your answer:

3. Prompt for intervention and control

   #Title and abstract

   Title: [*Title of the record was inserted here*]

   Abstract: [*Abstract of the record was inserted here*]

   #Intervention

   [*The 'intervention' specified in Textbox 1 was inserted here*]

   #Control

   [*The 'control' specified in Textbox 1 was inserted here*]

   #Query

   You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper.

   Does the paper with the above title and abstract meet the specified intervention and control criteria? If yes, highly suspected, or difficult to determine, answer 'I'. If not, answer 'E'.

   #Rules

   You can reply using only 'E' or 'I'.

   #Your answer:

XSL•FO
RenderX

Data Analysis

In this study, we analyzed the results from human evaluators of systematic review papers, comparing these with the records identified by GPT-3.5 or GPT-4. We considered the records included in the full-text assessment to be correct. We assessed the inclusion or exclusion decisions made by each human evaluator (before consensus was reached) against those determined by GPT-3.5 or GPT-4, focusing on sensitivity and specificity. Sensitivity was defined as the proportion of correctly identified eligible records for full-text assessment by human evaluators, GPT-3.5, or GPT-4. Formally, sensitivity is calculated as follows:

Sensitivity = True positives / (True positives + False negatives)

where:

True positives = Number of records correctly identified as eligible

False negatives = Number of records incorrectly identified as ineligible.

Similarly, specificity was defined as the proportion of correctly identified ineligible records (for full-text assessment) by human evaluators, GPT-3.5, or GPT-4. Formally, specificity is calculated as follows:

Specificity = True negatives / (True negatives + False positives)

where:

True negatives = Number of records correctly identified as ineligible

False Positives = Number of records incorrectly identified as eligible.

For records eligible for full-text assessment but excluded by either GPT-3.5 or GPT-4, we reviewed the title and the abstract to assess whether the exclusion decision was justified. Following this review, we recalculated sensitivity and specificity after adjusting for these justified exclusions. Furthermore, for records that were incorrectly excluded by GPT-3.5 or GPT-4, we conducted a narrative verification of the erroneous judgments by asking each LLM to explain the reasons behind their decisions. We modified the prompt used for screening (Textbox 2) by replacing the "#Rules" statement with "Specify the reason for your answer." This modification allowed GPT-3.5 or GPT-4 to provide their judgment results along with the underlying reasons.

Ethical Considerations

This study used only publicly available data from research papers and does not involve human subjects or personal data. Therefore, it does not require a human subject ethics review or exemption.

Results

Results on Takeshima et al Paper

Figure 2 [38] shows the number of records excluded by GPT-3.5 and GPT-4 at each layer of research design, target population, and intervention and control, applied to records in the paper by Takeshima et al [38].

Figure 2. Comparison of 3-layer screening results using GPT-3.5 and GPT-4 with human evaluation for Takeshima et al [38].

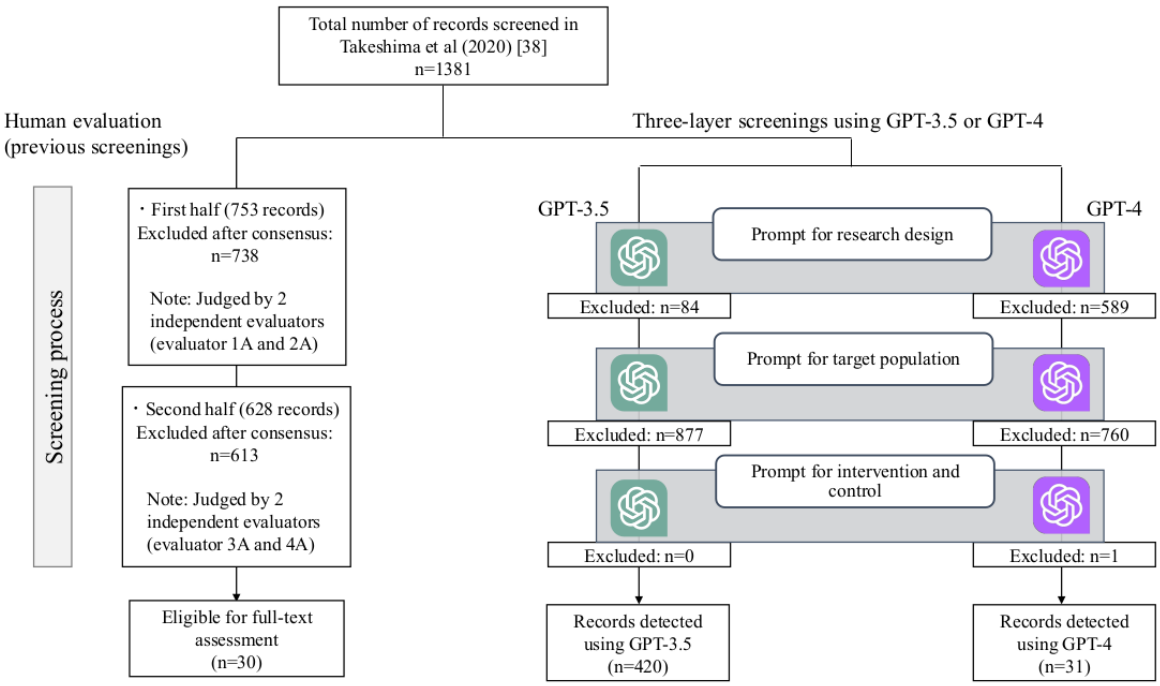

GPT-3.5 excluded 84 records at the research design layer, 877 records at the target population layer, and 0 record at the intervention and control layer, ultimately determining 420 out of 1382 records for inclusion. None of the 6 records (including 6 papers) that were included in the meta-analysis were excluded by GPT-3.5. The sensitivity for included records was 0.900 and

XSL•FO
RenderX

the specificity was 0.709. Among the eligible records for full-text assessment, GPT-3.5 classified 3 (10.0%) records as excluded. Of these, the exclusion of 2 records by GPT-3.5 was justified, while the remaining 1 (3.3%) record was deemed to require full-text assessment (Table 2). After adjustments for these justified judgments (Multimedia Appendix 2), the sensitivity improved to 0.966 and the specificity remained at 0.710. For the one record that GPT-3.5 determined to be excluded at the target population layer, it was suggested that GPT-3.5 concluded that the record "included both bipolar disorder and unipolar mood disorder, which did not match the selection criteria."

Table 2. Records for full-text assessment in the study by Takeshima et al [38] paper but were excluded by GPT-3.5 and GPT-4.

	Number of excluded records on each layer (number of those not justified)		
	Research design	Target population	Intervention and control
Number of records eligible for full-text assessment (n=30)			
Excluded by GPT-3.5	0	3 (1)[a]	0
Excluded by GPT-4	4 (1)[a]	2 (0)[a]	0

[a]Number of records for which exclusion was not justified.

GPT-4 excluded 589 records at the research design layer, 760 records at the target population layer, and 1 record at the intervention and control layer, ultimately determining 31 out of 1381 records for inclusion. None of the 6 records (including 6 papers) that were included in the meta-analysis were excluded by GPT-4. The sensitivity for included records was 0.806 and the specificity was 0.996. Among the eligible records for full-text assessment, GPT-4 classified 6 (20.0%) records as excluded. Of these, the exclusion of 5 records by GPT-4 was justified, while the remaining 1 (3.3%) record was considered to require full-text assessment (Table 2). After adjustments for these justified judgments (Multimedia Appendix 2), the sensitivity improved to 0.962 and the specificity remained at 0.996. GPT-4 included all 6 records (including 6 papers) that were included in the meta-analysis. For the one record that GPT-4 judged to be excluded at the research design layer, it was revealed that GPT-4 deduced that "although this study mentioned registration in an RCT, it investigated the associations between sleep, physical activity, and circadian rhythm indicators" (from the perspective of whether to include the study in the meta-analysis, GPT-4's judgment is likely to be correct; however, considering the purpose of the initial screening, we determined that it would be appropriate to include the study).

Results of the Paper by Maruki et al

Figure 3 [39] shows the number of records excluded by GPT-3.5 and GPT-4 at each layer of research design, target population, and intervention and control, applied to records in the Maruki et al [39] paper.

GPT-3.5 excluded 220 records at the research design layer, 126 records at the target population layer, and 10 records at the intervention and control layer, ultimately determining 2790 out of 3146 records for inclusion. None of the 9 records (including 9 papers) that were included in the meta-analysis were excluded by GPT-3.5. The sensitivity for included records was 0.958 and the specificity was 0.116. Among the eligible records for full-text assessment, GPT-3.5 classified 4 (4.2%) records as excluded. None of these records' exclusion by GPT-3.5 was justified, and all were considered to require full-text assessment (Table 3 and Multimedia Appendix 2). For the 2 records that GPT-3.5 inferred to be excluded at the research design layer, it was revealed that GPT-3.5 determined that "although they were RCTs, either the individual or cluster level was not specified" for both records. For the 2 records that GPT-3.5 deemed to be excluded at the target population layer, it was suggested that GPT-3.5 surmised that "although the records involved bipolar disorder, they did not match the selection criteria due to the presence of comorbidities (one record had generalized anxiety disorder, and the other had alcohol dependence)."

XSL•FO
RenderX

Figure 3. Comparison of 3-layer screening results using GPT-3.5 and GPT-4 with human evaluation for Maruki et al [39].

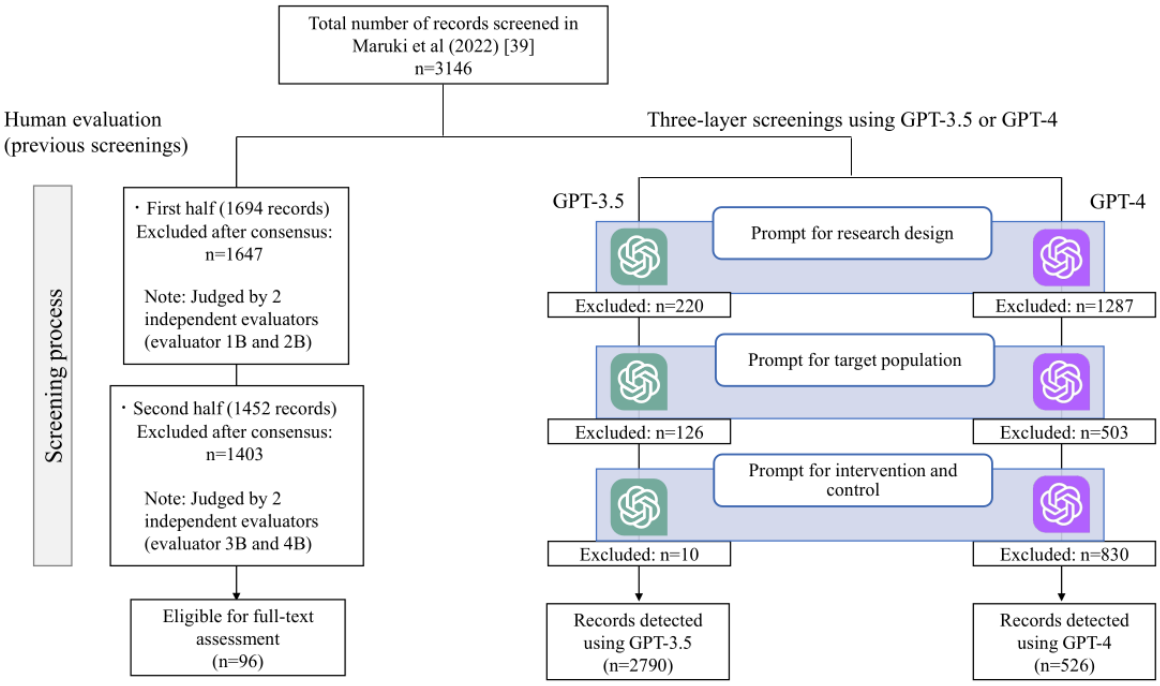

Table 3. Records for full-text assessment in the paper by Maruki et al [39] but were excluded by GPT-3.5 and GPT-4.

	Number of excluded records on each layer (number of those not justified)		
	Research design	Target population	Intervention and control
Number of records eligible for full-text assessment (n=96)			
Excluded by GPT-3.5	2 (2)[a]	2 (2)[a]	0
Excluded by GPT-4	5 (0)[a]	2 (1)[a]	5 (3)[a]

[a]Number of records for which exclusion was not justified.

GPT-4 excluded 1287 records at the research design layer, 503 records at the target population layer, and 830 records at the intervention and control layer, ultimately determining 526 out of 3146 records for inclusion. None of the 9 records (including 9 papers) that were included in the meta-analysis were excluded by GPT-4. The sensitivity for included records was 0.875 and the specificity was 0.855. Among the eligible records for full-text assessment, GPT-4 classified 12 (12.5%) records as excluded. Of these, the exclusion of 8 records by GPT-4 was justified, while the remaining 4 (4.2%) records were considered to require full-text assessment (Table 3). After adjustments for these justified judgments (Multimedia Appendix 2), the sensitivity improved to 0.943 and the specificity remained at 0.855. "For the one record that GPT-4 determined to be excluded at the target population layer, it was suggested that GPT-4 inferred that 'although the record involved bipolar disorder, it did not match the selection criteria due to the presence of a comorbidity (alcohol dependence).' For the three records that GPT-4 judged to be excluded at the Intervention and control layer, in each case, GPT-4 cited the reason for exclusion as 'the intervention criteria are the addition of either SGA or MS to SGA or MS, but this study does not mention the use of SGA.'"

In the list used in the paper by Maruki et al [39], there were a total of 355 records where part of the title and abstract were corrupted into irrelevant Chinese characters (eg, "This was an eight窶陣eek, open窶人abel, prospective study"). Despite these errors, all cases could be appropriately discerned, likely due to the context-sensitive judgment capability of GPT-3.5 and GPT-4.

Comparison of GPT-3.5, GPT-4, and Human Evaluators

Both the study by Takeshima et al [38] and the study by Maruki et al [39] involved 2 individuals conducting screening for the initial segment, while a different set of 2 individuals was responsible for the screening of the latter segment. The sensitivity and specificity of human evaluators and GPT-3.5 and GPT-4 for each segment are shown in Table 4. The adjusted results, in cases where the exclusion of GPT-3.5 or GPT-4 was justified, are shown in the numbers within parentheses (Table 4).

XSL•FO
RenderX

Table 4. Comparison of evaluation metrics: GPT-3.5, GPT-4, and human evaluators.

Screenings on Takeshima et al (2020) [38]	Human evaluators				LLMs[a]	
	1A	2A	3A	4A	GPT-3.5	GPT-4
Initial segment (n=753)						
Sensitivity	1.000	0.867	—[b]	—	0.800 (0.929)[c]	0.688 (1.000)[c]
Specificity	0.995	0.996	—	—	0.702 (0.704)[c]	0.997 (0.997)[c]
Latter segment (n=628)						
Sensitivity	—	—	1.000	1.000	1.000 (1.000)[c]	0.933 (0.933)[c]
Specificity	—	—	1.000	0.997	0.718 (0.718)[c]	0.993 (0.993)[c]
Screenings on Maruki et al (2022) [39]	Human evaluators	Human evaluators	Human evaluators	Human evaluators	LLMs	LLMs
Screenings on Maruki et al (2022) [39]	1B	2B	3B	4B	GPT-3.5	GPT-4
Initial segment (n = 1694)						
Sensitivity	0.766	0.979	—	—	0.936	0.872 (0.952)[c]
Specificity	0.998	0.998	—	—	0.129	0.886 (0.886)[c]
Latter segment (n=1452)						
Sensitivity	—	—	0.776	0.939	0.980	0.878 (0.935)[c]
Specificity	—	—	0.999	0.999	0.100	0.818 (0.819)[c]

[a]LLMs: large language models.

[b]Not applicable.

[c]Values after adjusting for cases where exclusion was justified.

Time and Cost Required for Screenings

In our Google Spreadsheet setup, both GPT-3.5 and GPT-4 managed to process approximately 110 records per minute across each of the 3 layers. Consequently, the estimated ideal completion time was between 20 and 30 minutes for the study by Takeshima et al [38], and between 60 and 80 minutes for the study by Maruki et al [39]. However, in practice, due to errors with the Google Spreadsheet and API, the screening process took about 1 hour for the study by Takeshima et al [38] and about 2 hours in total for the study by Maruki et al [39]. Furthermore, due to daily API call limits, the work had to be spread out over 3 days. The screening for these 2 studies incurred a total cost of US $59, with US $4 for calls to GPT-3.5 and US $55 for calls to GPT-4.

Discussion

Principal Findings

This study demonstrates the use of a 3-layer screening method using GPT-3.5 and GPT-4 for title and abstract screenings in systematic reviews, highlighting its remarkable speed and sensitivity comparable with that of human evaluators. However, GPT-3.5 demonstrated low specificity for relevant records, rendering it less practical. In contrast, the use of GPT-4 showed both high sensitivity and specificity, particularly where adjustments for justified exclusions led to an improvement in sensitivity. Although achieving 100% sensitivity remained unattainable, a 3-layer screening method with GPT-4 may potentially be practical for use in the systematic review process and can reduce human labor.

Previous research demonstrating the effectiveness of automated screening using text mining has encountered sensitivity issues [20-29]. Specifically, the exclusion of important studies that should have been included in their meta-analysis [20,21,23,29], a limitation not observed in our approach, hampered their application to clinical practice. False negatives in machine learning–based screening can arise from several factors: complexity in research design, characteristics of the target demographic, types of interventions, complexity in selection criteria, a significant scarcity of relevant records within the data set (leading to data imbalance), and inconsistency in the terminology used for judgment [21,23,29]. Our method using GPT-3.5 or GPT-4 was able to address issues related to data set imbalance and terminology inconsistency, as we used the same prompt across records, and assess the inclusion or exclusion one by one. In addition, previous text mining screenings may not have effectively addressed garbled text, such as “open-label” mistakenly appearing as “open窶人abel” [40], an issue that LLMs can potentially mitigate through their attention mechanisms [41]. Moreover, the outstanding knowledge base of GPT-4 [6,7] likely helped address the complexity in research design, target demographics, and intervention, as well as selection criteria—areas where GPT-3.5 might have fallen short. These distinctions possibly account for the notable differences in specificity observed between GPT-3.5 and GPT-4. Recently, Guo et al [33] conducted title and abstract screening using GPT-4. Their approach diverges from our 3-layer method; it

XSL•FO
RenderX

integrated inclusion and exclusion criteria within the context, generating decisions and reasoning through a single prompt. While we believe that our 3-layer method could potentially offer greater sensitivity than theirs, it remains difficult to definitively assert a significant improvement in sensitivity over the method by Guo et al [33], given the limited sample size and the differences in data sets. Tran and colleagues' approach [34], despite using GPT-3.5, demonstrated remarkable sensitivity. It is important to note, however, that the manual creation of their highly effective prompt raises questions regarding its replicability and broader applicability.

Both human-conducted and LLM-conducted systematic reviews have their inherent pitfalls. Errors made by humans are inevitable, with their accuracy estimated to be around 10% [15], and slightly higher for false exclusions, at approximately 13%-14% [16,17]. These values represent the performance of experts in the relevant field, and the accuracy may be lower for individuals with less expertise or shallow screening experience; therefore, guidelines have recommended piloting and training the abstract screening team [12]. In this study, we observed that human evaluation in the paper by Takeshima et al [38] exhibited slightly more false negatives than that in the paper by Maruki et al [39]. Although the reasons for the judgment discrepancies were not investigated in this study's data set, they may be attributed to the larger volume of records screened [14] and the potentially more complex and challenging research question in the paper by Maruki et al [39]. Using 2 reviewers to screen records can significantly lower the likelihood of false negatives [16] and has been recommended [11,13]. Yet, simultaneously, there has been a case that the systematic review screenings, albeit rare, are conducted by a single reviewer, because of time constraints [13,42]. Hence, the unavoidable errors and substantial time and effort required for screening represent significant drawbacks of human screening in systematic reviews [10,13].

Conversely, methods using LLMs also present several drawbacks. One primary concern is their susceptibility to misinformation and quality issues inherent in their training data [43]. Notably, in this study, the specificity of the GPT-3.5 screenings in Maruki et al [39] paper was markedly low. While the causes are not definitive, this may be attributed to an insufficient understanding of bipolar disorder, MSs, and second-generation antipsychotics. Tran and colleagues [34] incorporated relevant knowledge into their manually created prompts; it might have enhanced sensitivity but not specificity; and this could also be due to GPT-3.5's knowledge limitations. Furthermore, the decision-making processes of LLMs lack transparency, making them difficult to interpret [43]. This lack of interpretability is compounded by the "grounding problem," where LLMs struggle to grasp concrete facts and real-world scenarios due to their lack of real-world experiences and sensory input [1,44]. We attempted to verify incorrectly excluded records by querying GPT-3.5 and GPT-4 with the original screening prompts, their responses, and justifications. Our findings revealed that GPT-3.5's lower accuracy was primarily due to a lack of knowledge about the target domain, while GPT-4's incorrect exclusions were mainly due to misinterpretations of the inclusion criteria. These findings highlight the ongoing challenges in understanding and interpreting the decision-making processes of LLMs. Although GPT-4 demonstrates advancements in comprehension, factuality, specificity, and inference, it is still more susceptible to factual errors [45]. In addition, it has been suggested that LLMs' accuracy diminishes with longer prompts [46]; lengthy abstracts might have contributed to decreased accuracy in decision-making. A potential future risk is that the normalization of AI-based judgments could result in the oversight of human expert verification, potentially diminishing the quality of systematic reviews.

On the positive side, compared with the human screening time reported in previous studies [10], our method enabled remarkably faster screening. Although our approach uses a 3-layer structure, which might seem time-consuming at first glance, by limiting GPT-3.5/GPT-4 responses to "E" (Exclude) or "I" (Include), we efficiently screened a large volume of records in batch. Unlike humans, LLMs do not experience fatigue and subsequent decline in performance; moreover, they are presumed to have better reproducibility in their judgments. While using GPT-4's API comes with associated costs [36], the increased efficiency compared with human effort more than compensates for these expenses. Using LLMs for title and abstract screening could also enable screening a much larger number of records, previously deemed impractical due to time limitations. Our 3-layer method using GPT-4 exhibits high sensitivity and a useful level of specificity and yet opportunities for further refinement exist. Future studies could enhance accuracy through methods such as optimizing prompts [47] and integrating multiple LLMs for decision assessment [48], which may contribute to higher precision. In the meantime, swift advancements in LLM technology are set to continuously evolve; future breakthroughs in LLMs may readily overcome our current challenges—possibly, only by a simple prompt.

Limitations

This study has some limitations. First, the 2 systematic reviews used in this investigation [38,39] were confined to clinical studies within psychiatry, limiting the generalizability of our findings. In addition, the sample size was small, and the investigation remained exploratory, with the results lacking statistical substantiation. Future studies should aim to replicate these findings across a broader range of medical fields and specialized domains to enhance their applicability and reliability. Second, the artificial intelligence industry is progressing rapidly, with information becoming obsolete within a matter of months or even weeks. The models we used in this study, gpt-3.5-turbo-0125 and gpt-4-0125-preview, are currently the most up-to-date. However, updates to these models might alter screening outcomes. Third, to ensure consistency in our findings, we set the temperature parameter to 0. However, a temperature of 0 does not always guarantee absolute uniformity in output sentences [35]. However, our observations indicate no variation in results across multiple tests with the same model in this study. Fourth, this study did not investigate the discrepancies in screening results between GPT-3.5 and GPT-4, nor did it examine the impact of prompt variations on performance. In addition, this research did not directly compare the performance of the proposed approach with existing systematic literature

XSL•FO
RenderX

review strategies. Furthermore, this study was not designed to explore the risks associated with using LLMs for screening purposes. Finally, gpt-3.5-turbo-0125's training data include information up to September 2021, whereas gpt-4-0125-preview's training data extend to December 2023 [35]. Consequently, the systematic review paper by Takeshima et al [38] might have been incorporated into GPT-3.5's training data set, with both systematic review papers possibly included in GPT-4's data set. Nevertheless, as the study's prompts did not explicitly reference these reviews, we consider that their impact is minimal.

Conclusions

We developed a practical screening method using GPT-3.5 and GPT-4 in the title- and abstract-screening process of systematic reviews. Our 3-layer method not only achieved better sensitivity for relevant records than previous machine learning–based screening methods [20,21,23,29] but also demonstrated a remarkable potential to reduce human reviewers' workload significantly. Although GPT-3.5 showed lower specificity, which may limit its applicability, the use of GPT-4 within our method yielded sensitivity comparable with human evaluators, making it suitable for use in systematic review screenings. Despite the focus on psychiatric fields and the small sample size of our study, our findings highlight the potential for broader application. We emphasize the importance of further validation across multiple domains to establish a universal screening methodology. Concurrently, developing more effective approaches in response to the advancing capabilities of LLMs is warranted in future research.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant 22K15778). During the preparation of this work, the authors used ChatGPT (GPT-4 and GPT-4o, by OpenAI), Claude (Claude 3 Opus, by Anthropic), and Gemini (Gemini 1.5 Pro, by Google) to enhance the readability and proofread the English text. After using these services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Script for the Google Spreadsheet.
[DOCX File , 23 KB - jmir_v26i1e52758_app1.docx]

Multimedia Appendix 2
Records eligible for full paper screening but excluded by GPT-3.5 or GPT-4.
[XLSX File (Microsoft Excel File), 26 KB - jmir_v26i1e52758_app2.xlsx]

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901.
2. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730-27744. [doi: 10.48550/arXiv.2203.02155]
3. Introducing ChatGPT. URL: https://openai.com/blog/chatgpt [accessed 2023-07-01]
4. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
5. GPT-4. URL: https://openai.com/research/gpt-4 [accessed 2024-02-29]
6. Bojic L, Kovacevic P, Cabarkapa M. GPT-4 surpassing human performance in linguistic pragmatics. arXiv. Preprint posted online 2023 Dec 15.
7. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2023;1(1):AIp2300031. [doi: 10.1056/aip2300031]
8. Kim SG. Using ChatGPT for language editing in scientific articles. Maxillofac Plast Reconstr Surg 2023;45(1):13 [FREE Full text] [doi: 10.1186/s40902-023-00381-x] [Medline: 36882591]
9. Matsui K, Koda M, Yoshida K. Implications of nonhuman "Authors". JAMA 2023;330(6):566. [doi: 10.1001/jama.2023.10568] [Medline: 37552501]
10. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Searching for and selecting studies. In: Cochrane Handbook for Systematic Reviews of Interventions. New York, NY: John Wiley & Sons; 2019:67-107.
11. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions. New York, NY: John Wiley & Sons; 2019.
12. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large‐evidence systematic reviews and meta‐analyses. Res Synth Methods 2019;10(3):330-342. [doi: 10.1002/jrsm.1354]

XSL•FO
RenderX

13. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
14. O'Hearn K, MacDonald C, Tsampalieros A, Kadota L, Sandarage R, Jayawarden SK, et al. Evaluating the relationship between citation set size, team size and screening methods used in systematic reviews: a cross-sectional study. BMC Med Res Methodol 2021;21(1):142 [FREE Full text] [doi: 10.1186/s12874-021-01335-5] [Medline: 34238247]
15. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One 2020;15(1):e0227742 [FREE Full text] [doi: 10.1371/journal.pone.0227742] [Medline: 31935267]
16. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol 2020;121:20-28 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.01.005] [Medline: 31972274]
17. Wilson E, Cruz F, Maclean D, Ghanawi J, McCann SK, Brennan PM, et al. Screening for in vitro systematic reviews: a comparison of screening methods and training of a machine learning classifier. Clin Sci (Lond) 2023;137(2):181-193 [FREE Full text] [doi: 10.1042/CS20220594] [Medline: 36630537]
18. Bannach-Brown A, Przybyła P, Thomas J, Rice ASC, Ananiadou S, Liao J, et al. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst Rev 2019;8(1):23 [FREE Full text] [doi: 10.1186/s13643-019-0942-7] [Medline: 30646959]
19. Cierco Jimenez R, Lee T, Rosillo N, Cordova R, Cree IA, Gonzalez A, et al. Machine learning computational tools to assist the performance of systematic reviews: a mapping review. BMC Med Res Methodol 2022;22(1):322 [FREE Full text] [doi: 10.1186/s12874-022-01805-4] [Medline: 36522637]
20. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods 2014;5(1):31-49. [doi: 10.1002/jrsm.1093] [Medline: 26054024]
21. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev 2015;4:80 [FREE Full text] [doi: 10.1186/s13643-015-0067-6] [Medline: 26073974]
22. Olofsson H, Brolund A, Hellberg C, Silverstein R, Stenström K, Österberg M, et al. Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. Res Synth Methods 2017;8(3):275-280. [doi: 10.1002/jrsm.1237] [Medline: 28374510]
23. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018;7(1):45 [FREE Full text] [doi: 10.1186/s13643-018-0707-8] [Medline: 29530097]
24. Gartlehner G, Wagner G, Lux L, Affengruber L, Dobrescu A, Kaminski-Hartenthaler A, et al. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. Syst Rev 2019;8(1):277 [FREE Full text] [doi: 10.1186/s13643-019-1221-3] [Medline: 31727159]
25. Gates A, Gates M, Sebastianski M, Guitard S, Elliott SA, Hartling L. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage abstrackr's relevance predictions in systematic and rapid reviews. BMC Med Res Methodol 2020;20(1):139 [FREE Full text] [doi: 10.1186/s12874-020-01031-w] [Medline: 32493228]
26. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening—impact on reviewer-relevant outcomes. BMC Med Res Methodol 2020;20(1):256 [FREE Full text] [doi: 10.1186/s12874-020-01129-1] [Medline: 33059590]
27. Reddy SM, Patel S, Weyrich M, Fenton J, Viswanathan M. Comparison of a traditional systematic review approach with review-of-reviews and semi-automation as strategies to update the evidence. Syst Rev 2020;9(1):243 [FREE Full text] [doi: 10.1186/s13643-020-01450-2] [Medline: 33076975]
28. Pham B, Jovanovic J, Bagheri E, Antony J, Ashoor H, Nguyen TT, et al. Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow. Syst Rev 2021;10(1):156 [FREE Full text] [doi: 10.1186/s13643-021-01700-x] [Medline: 34039433]
29. Valizadeh A, Moassefi M, Nakhostin-Ansari A, Hosseini Asl SH, Saghab Torbati M, Aghajani R, et al. Abstract screening using the automated tool rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. BMC Med Res Methodol 2022;22(1):160 [FREE Full text] [doi: 10.1186/s12874-022-01631-8] [Medline: 35655155]
30. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev 2019;8(1):143 [FREE Full text] [doi: 10.1186/s13643-019-1062-0] [Medline: 31215463]
31. Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, Samiee R. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. BMJ Evid Based Med 2024;29(1):69-70 [FREE Full text] [doi: 10.1136/bmjebm-2023-112678] [Medline: 37989538]
32. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods 2024;15(4):616-626. [doi: 10.1002/jrsm.1715] [Medline: 38484744]

XSL•FO
RenderX

33. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res 2024;26:e48996 [FREE Full text] [doi: 10.2196/48996] [Medline: 38214966]
34. Tran VT, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. Ann Intern Med 2024;177(6):791-799. [doi: 10.7326/M23-3389] [Medline: 38768452]
35. Models. URL: https://platform.openai.com/docs/models [accessed 2024-02-29]
36. Pricing. URL: https://openai.com/pricing [accessed 2024-03-06]
37. API Reference. URL: https://platform.openai.com/docs/api-reference/ [accessed 2024-02-29]
38. Takeshima M, Utsumi T, Aoki Y, Wang Z, Suzuki M, Okajima I, et al. Efficacy and safety of bright light therapy for manic and depressive symptoms in patients with bipolar disorder: a systematic review and meta-analysis. Psychiatry Clin Neurosci 2020;74(4):247-256 [FREE Full text] [doi: 10.1111/pcn.12976] [Medline: 31917880]
39. Maruki T, Utsumi T, Takeshima M, Fujiwara Y, Matsui M, Aoki Y, et al. Efficacy and safety of adjunctive therapy to lamotrigine, lithium, or valproate monotherapy in bipolar depression: a systematic review and meta-analysis of randomized controlled trials. Int J Bipolar Disord 2022;10(1):24 [FREE Full text] [doi: 10.1186/s40345-022-00271-7] [Medline: 36269465]
40. Benchimol J, Kazinnik S, Saadon Y. Text mining methodologies with R: An application to central bank texts. Machine Learn with Appl 2022;8:100286. [doi: 10.1016/j.mlwa.2022.100286]
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process syst 2017;30:5998-6008.
42. Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, et al. Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. Cochrane Database Syst Rev 2020;4(4):CD013574 [FREE Full text] [doi: 10.1002/14651858.CD013574] [Medline: 32267544]
43. Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, et al. Pre-trained language models in biomedical domain: a systematic survey. ACM Comput Surv 2023;56(3):1-52. [doi: 10.1145/3611651]
44. Mollo DC, Millière R. The vector grounding problem. arXiv. Preprint posted online 2023 Apr 04.
45. Zheng S, Huang J. Chang KC-C. why does chatgpt fall short in providing truthful answers. arXiv. Preprint posted online 2023 Dec 03.
46. Levy M, Jacoby A, Goldberg Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. arXiv. Preprint posted online 2024 Jul 10.
47. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 2023;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
48. Li J, Zhang Q, Yu Y, Fu Q, Ye D. More agents is all you need. arXiv. Preprint posted online 2024 Feb 03.

Abbreviations

API: application programming interface
GPT: Generative Pre-trained Transformer
LLM: large language model
MS: mood stabilizers
SGA: second-generation antipsychotics

Edited by S Ma; submitted 14.09.23; peer-reviewed by D Fraile Navarro, T Nguyen, A Nakhostin-Ansari; comments to author 23.01.24; revised version received 10.03.24; accepted 25.06.24; published 16.08.24.

Please cite as:
Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y
Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews
J Med Internet Res 2024;26:e52758
URL: https://www.jmir.org/2024/1/e52758
doi:10.2196/52758
PMID:

©Kentaro Matsui, Tomohiro Utsumi, Yumi Aoki, Taku Maruki, Masahiro Takeshima, Yoshikazu Takaesu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Dual Task Ball Balancing Test and Its Association With Cognitive Function: Algorithm Development and Validation

Barry Greene[1,2], PhD; Sean Tobyne[1], PhD; Ali Jannati[1], PhD; Killian McManus[1,2], PhD; Joyce Gomes Osman[1], PhD; Russell Banks[1], PhD; Ranjit Kher[1]; John Showalter[1], MD; David Bates[1], PhD; Alvaro Pascual-Leone[1,3], MD, PhD

[1]Linus Health, Boston, MA, United States

[2]Linus Health Europe, Dublin, Ireland

[3]Hinda and Arthur Marcus Institute for Aging Research, Deanna and Sidney Wolk Center for Memory Health, Hebrew SeniorLife, Boston, MA, United States

Corresponding Author:
Killian McManus, PhD
Linus Health
280 Summer St
Boston, MA, 02210
United States
Phone: 1 851682046
Email: kmcmanus@linus.health

Abstract

Background: Dual task paradigms are thought to offer a quantitative means to assess cognitive reserve and the brain's capacity to allocate resources in the face of competing cognitive demands. The most common dual task paradigms examine the interplay between gait or balance control and cognitive function. However, gait and balance tasks can be physically challenging for older adults and may pose a risk of falls.

Objective: We introduce a novel, digital dual-task assessment that combines a motor-control task (the "ball balancing" test), which challenges an individual to maintain a virtual ball within a designated zone, with a concurrent cognitive task (the backward digit span task [BDST]).

Methods: The task was administered on a touchscreen tablet, performance was measured using the inertial sensors embedded in the tablet, conducted under both single- and dual-task conditions. The clinical use of the task was evaluated on a sample of 375 older adult participants (n=210 female; aged 73.0, SD 6.5 years).

Results: All older adults, including those with mild cognitive impairment (MCI) and Alzheimer disease–related dementia (ADRD), and those with poor balance and gait problems due to diabetes, osteoarthritis, peripheral neuropathy, and other causes, were able to complete the task comfortably and safely while seated. As expected, task performance significantly decreased under dual task conditions compared to single task conditions. We show that performance was significantly associated with cognitive impairment; significant differences were found among healthy participants, those with MCI, and those with ADRD. Task results were significantly associated with functional impairment, independent of diagnosis, degree of cognitive impairment (as indicated by the Mini Mental State Examination [MMSE] score), and age. Finally, we found that cognitive status could be classified with >70% accuracy using a range of classifier models trained on 3 different cognitive function outcome variables (consensus clinical judgment, Rey Auditory Verbal Learning Test [RAVLT], and MMSE).

Conclusions: Our results suggest that the dual task ball balancing test could be used as a digital cognitive assessment of cognitive reserve. The portability, simplicity, and intuitiveness of the task suggest that it may be suitable for unsupervised home assessment of cognitive function.

(J Med Internet Res 2024;26:e49794) doi:10.2196/49794

KEYWORDS

cognitive function; dual task; inertial sensors; mHealth; tablet; MCI; Alzheimer; dementia; motor; older adults; cognitive impairment; balance

XSL•FO
RenderX

Introduction

Recent research has suggested that up to 40% of dementia cases [1] can be delayed or prevented through early identification of impairment and adherence to recommended lifestyle modifications [2]. Furthermore, recent developments in pharmaceutical intervention suggest that the progression of Alzheimer dementia can be delayed through amyloid plaque removal [3].

An individual's cognitive and behavioral performance is a combination of brain activity and cognitive reserve. Cognitive reserve can be conceptualized as a property of the brain that allows for better than expected performance, given the degree of life-course related brain changes and brain injury or disease [4]. Cognitive reserve can be influenced by multiple genetic and environmental factors, operating at various points or continuously across the lifespan. In the presence of disease, for example, a neurodegenerative disease such as Alzheimer disease, cognitive reserve is engaged to sustain function for as long as possible and minimize symptoms and disability. Thus, individuals with more cognitive reserve manifest symptoms or disability later than those with lower cognitive reserve; symptoms are less prominent or severe than might be expected for a given amount of pathology. Low cognitive reserve makes individuals with underlying brain pathology prone to episodes of confusion, delirium, and other acute decompensations when exposed to a stressor or insult, for example, elective surgery, infection, sleep, and deprivation. Individuals with mild cognitive impairment (MCI) and higher cognitive reserve can delay the development of dementia. Thus, assessment of cognitive reserve is important to predict an individual's functional state and prognosis. In addition, cognitive reserve can be a powerful therapeutic target, as increasing cognitive reserve might reduce disability.

The brain's resource allocation capacity has been studied extensively and is thought to provide insight into cognitive reserve and depend on prefrontal function. However, the nature and causality of this relationship is not as well understood. Dual task paradigms have long been thought to unlock deficits in the allocation of prefrontal resources [5]. Recent studies [6-8] have examined the impact of a cognitive task (eg, backward counting) on a participant's gait or balance, and thus, are dependent on peripheral nerve and musculoskeletal factors often affected in older adults. Furthermore, gait and balance analysis may not be suitable or safe for use with older adults or those with comorbidities such as osteoarthritis, neuropathies, etc. A validated tool that can support objective characterization and quantitative evaluation of cognitive reserve safely and reliably in older adults, as well as early identification of cognitive decline in nonclinical settings, could be of clinical benefit in more accurately identifying those patients who would benefit most from early and targeted intervention.

We introduce a novel test of motor control, coordination, and attention—the "ball balancing" test, in which an individual is asked to maintain the position of a virtual ball in the center of a circular target area. Task performance is measured by examining the position of a virtual ball on the screen of a touchscreen tablet, estimated using the inertial sensors embedded in the tablet. The test can be easily adapted to a dual task condition, for example, by asking the individual to balance the ball while simultaneously doing a different, attention demanding task. The test can be completed comfortably and safely in a sitting position. In an initial version of a dual task paradigm, an individual's ball balancing test performance was assessed while simultaneously conducting a backward digit span test (BDST).

We aimed to examine the use of the ball balancing test under single and dual conditions [as quantified using the inertial measurement unit (IMU) sensors embedded in the target device] in assessment of cognitive reserve and identification of cognitive impairment. While this task (and other dual task paradigms) is not primarily aimed at serving as a means to classify cognitive function, one may predict there should be a loss of cognitive reserve between MCI and dementia, given that cognitive reserve would be "used up" to sustain cognitive function and ultimately be no longer sufficient to prevent progression of deficit, impact on activities of daily living (ADL), and thus transition from MCI to dementia. We report the performance of the task in classifying cognitive status according to 3 different outcome measures (consensus clinical judgment, Rey Auditory Verbal Learning, and Mini Mental State Examination). Given that the outcome measures are imperfectly mutually correlated, it can be assumed that they may contain complementary information pertinent to assessment of cognitive function, which can be leveraged to examine cognitive reserve deficits.

Methods

Ball Balancing Task

Participants were seated in a comfortable and supportive chair and asked to hold a touchscreen tablet device (iPad Pro, Apple) parallel to the ground and tilt the screen as needed to keep a virtual ball within a target area—the ball was not perturbed during the test unless the tablet was moved by the participant.

Participants were asked to balance a virtual ball on a touchscreen tablet screen, the subsequent movement is measured by the IMU sensors embedded in the tablet and used to calculate the position of the virtual ball on the tablet screen. The ball balancing test was completed under both single task (ball balancing alone) as well as under dual task conditions, with participants completing a single trial of each. The dual task involved asking the participant to complete the ball balancing test while simultaneously completing a BDST. In the BDST, the participant is played an audio sequence of 4 digits and is prompted to repeat them in reverse order. The single task was 20 seconds in duration while the dual task was 45 seconds in duration.

A custom iOS application (Swift, iOS) was developed to conduct the ball balancing test, supporting single and dual tasks. IMU and ball position data were stored within the application in JSON format and exported for offline analysis. All analyses were conducted using MATLAB (R2017b, MathWorks; Figure 1).

XSL•FO
RenderX

Figure 1. Linus Ball balancing task iOS app running on an iPad Pro (left panel). Trajectory of virtual ball position on screen over the course of a dual task ball balancing test, as calculated from tablet inertial measurement unit data (right Panel).

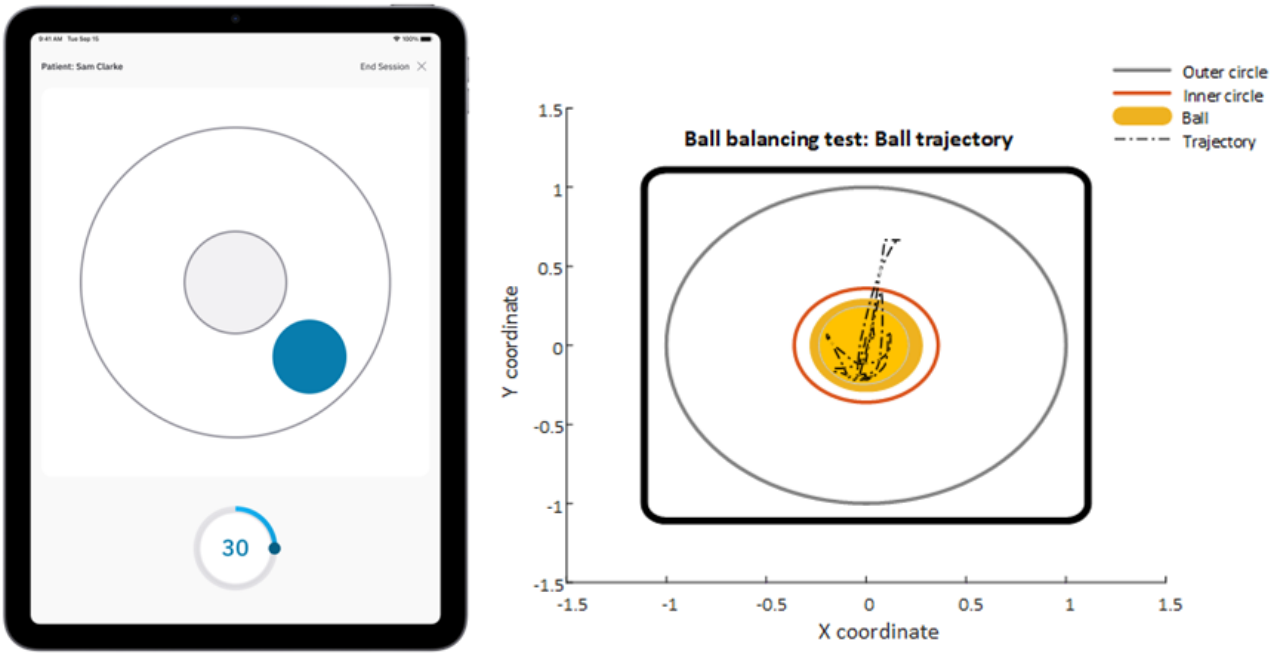

The following two distinct approaches were used to analyze the data, that is, by examining: (1) the ball position on the screen (estimated from IMU data); and (2) the IMU data from the movement of the tablet during the test.

Signal Processing: Ball Position Metrics

The position of the ball on the screen was calculated using the inertial sensor data as input to a kinematic model, which derives the ball placement on the screen using Newtonian mechanics and allowed plotting of ball displacement on the screen. The following parameters were calculated from the virtual ball displacement (values in parenthesis indicate variants of the calculated feature) (Textbox 1).

For all ball position metrics, the displacement is normalized to the range [9], where the outer edge is the radius of the outer circle, while the radius of the inner circle is calculated based on the ratio of the inner circle radius to the outer circle. The percentage of time spent within the inner circle is calculated as the proportion of time where the resultant displacement is less than the radius of the inner circle less the radius of the ball. The radial symmetry is calculated as the sum of the first difference values of the resultant displacement from the center of the circle. It is intended to measure quadrant placement of the ball within the outer circle. To examine learning effects and changes in performance over the course of each test, the percentage of time within the inner circle is calculated for each 5-s epoch within the test. The mean, standard deviation, and first difference were then calculated across all epochs per test to provide a measure of intratest performance. A number of standard center of pressure measures [10,11] were also calculated based on time and frequency domain analysis of the ball displacement. Each ball position metric was calculated for each participant under single task (ST) and dual task (DT) conditions; the dual task cost was calculated as the percentage difference between the parameter value under DT conditions and the parameter value under ST conditions and can be expressed mathematically as –100*(DT-ST)/ST [12].

A “perfect score” was achieved when the ball was found to lie within the inner circle for 100% of the test. As it was possible to achieve a perfect score by placing the tablet flat on a table, we examined if perfect score tests had any effect on the overall results to rule out the possibility that certain participants were engaging less with the task but achieving a perfect score.

XSL•FO
RenderX

Textbox 1. Features calculated from the virtual ball displacement data for each ball balancing test.

- Percentage of test time spent within inner circle
- Radial symmetry
- Percentage of time spent in the inner circle per 5-second epoch (mean, SD, and first difference)
- Median frequency of ball displacement (mean, X, and Y)
- 95% spectral edge frequency of ball displacement (mean, X, and Y)
- Sway area of ball displacement
- Mean sway frequency (mean, X, and Y)
- Mean sway distance (mean, X, and Y)
- Resultant sway distance (mean, X, and Y)
- Sway length of ball displacement (mean, X, and Y))
- Sway velocity (mean, X, and Y)

Inertial Sensor Parameters

Inertial sensor data from the tablet device under both ST and DT conditions were processed using an adapted version of a previously reported algorithm [13,14]; this approach treats the IMU data as arising from motion about a rigid plane. Figure 2 below shows the IMU (triaxial accelerometer and triaxial gyroscope) data for a dual task ball balancing test.

For each test, 1 second of data was excluded from the start and end of each recording to remove artifacts due to tablet positioning. Any recordings less than 10 seconds were discarded. IMU data were resampled to 100 Hz as iPad IMU data can be unevenly sampled, leading to distortion in frequency domain signal features [15]. Signals were bandpass filtered using a fourth order Butterworth IIR filter, in the range 0.1-40 Hz and calibrated using a standard method [16].

The following parameters were calculated from the IMU data for each ball balancing test (Textbox 2).

For each calculated parameter, the dual task cost was calculated as the percentage difference between the parameter value under dual task conditions and the parameter value under single task conditions. Figure 2 provides a 3D representation of the ball balancing test signal relative to the rigid plane.

Figure 2. Inertial measurement unit (IMU) signal obtained from an iPad during a dual task ball balancing test (left panel). 3D acceleration signal about the tablet plane during a ball balancing test (right panel).

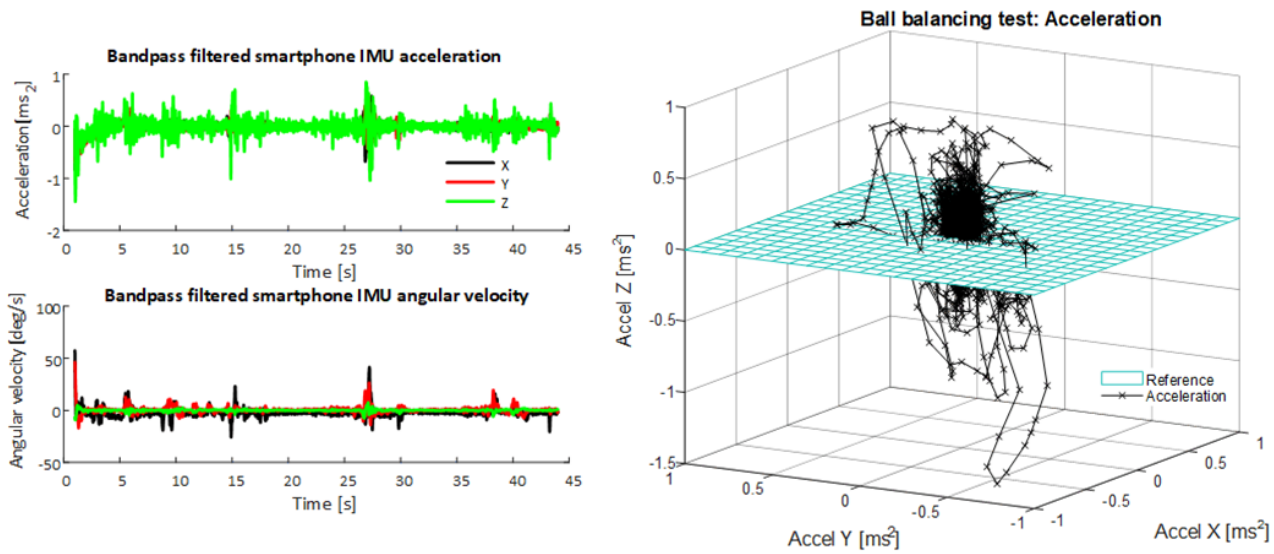

XSL•FO
RenderX

Textbox 2. Features calculated from inertial sensor data for each ball balancing test.

- Root mean square (RMS) acceleration (m/s^2)
- RMS acceleration—x-axis (m/s^2)
- RMS acceleration—y-axis (m/s^2)
- RMS acceleration—z-axis (m/s^2)
- RMS angular velocity (°/s)
- Median frequency acceleration (Hz)
- RMS angular velocity—x-axis (°/s)
- RMS angular velocity-y-axis (°/s)
- Spectral edge frequency acceleration (Hz)
- Spectral entropy acceleration
- Median frequency angular velocity (Hz)
- Spectral edge frequency angular velocity (Hz)
- Spectral entropy angular velocity
- Sway path length of acceleration—x-axis (m/s^2)
- Sway path length of acceleration—z-axis (m/s^2)
- Sway area of the acceleration
- Sway jerk of the acceleration
- Area of 95% confidence ellipse of acceleration

Statistical Analysis

To examine the association between the calculated ball balancing test parameters and cognitive function, we considered the 3 available neurocognitive measures (cohort status, Mini Mental State Examination [MMSE], and Rey Auditory Verbal Learning Test [RAVLT]), treated as either continuous variables or binary labels (eg, impaired or not impaired). Cohort status was treated as a 3-category label (healthy, MCI, and Alzheimer disease–related dementia [ADRD]). Similarly, the differences between the healthy and impaired subgroups (MCI and ADRD) were also examined using a Wilcoxon rank sum test. A Wilcoxon signed rank test was used to test for significant differences across task conditions.

The MMSE (total score) and RAVLT (long recall delay score) data were dichotomized into cognitively impaired and cognitively intact with values below a threshold of 28 for the MMSE [17] and age group thresholds for the RAVLT [18] used to identify impaired cognition.

Spearman rank correlation was used to examine the relationship between each feature with the MMSE and RAVLT, while the Wilcoxon rank sum test was used to test for differences between impaired and nonimpaired groups for each feature. A confusion matrix was calculated for each set of binary labels (impaired/nonimpaired) to see how well cohort status, MMSE-, and RAVLT-based categorization agree with each other.

To examine the association of each variable with cognitive function and allow for the effect of age, a linear mixed effects model analysis was conducted with age as a within-subjects' factor and cohort status as a categorical response variable. ANOVA was then used to examine the effect of each factor on cohort status, while controlling for age. This analysis was repeated for each variant with binary cohort status as well as impaired and nonimpaired labels obtained from MMSE and RAVLT.

In addition, we aimed to examine if any of the calculated ball balancing parameters were associated with functional impairment, independent of cognitive impairment. We conducted a one-way ANOVA for each ball balancing parameter with functional impairment (as measured by the Functional Activity Questionnaire [FAQ], with a threshold greater than or equal to 6 denoting functional impairment), controlling for MMSE and age. This analysis was then repeated when controlling for RAVLT and age.

To determine how well ball balancing parameters (features) could classify "unseen" participants according to binary cognitive status (cognitively impaired or cognitively intact), we used a logistic regression classifier model with a sequential forward feature selection procedure [19] validated using 10-fold cross-validation. Interaction terms were included in the candidate feature set and separate models were produced for each condition and feature set (ST, DT, dual task cost, all features as well as age only).

Data

A sample of 375 older adults (n=210 female; aged 73.0, SD 6.5 years). Completed a battery of cognitive and motor function tests as part of wider study on brain health. The Bio-Hermes research study is managed by the Global Alzheimer Platform

XSL•FO
RenderX

(GAP) and seeks new solutions to monitor and maintain brain health. Each participant received a clinical examination, which included the MMSE [20], the RAVLT [21] and "cohort status," which classified participants into 3 clinical categories (healthy, MCI, and ADRD), as determined by a panel of qualified clinicians. For RAVLT, 2 summary scores were examined: the RAVLT total score and the RAVLT long recall delay score. In addition, each participant completed an FAQ [9] to examine functional status including ADL.

Ethical Considerations

The Bio-Hermes research study is managed by the GAP. The study was performed in accordance with the Declaration of Helsinki and its later amendments. The study procedures were explained to participants verbally and through written informed consent that was approved by the local IRB of each site participating in the GAP consortium (see the Bio-Hermes study website [22] for a list of study sites). If, in the opinion of the site principal investigator, the participant did not have the capacity to sign the informed consent form, a legally authorized representative was used to grant consent on behalf of the participant. Ethical approval was granted by each institution participating in the GAP consortium (reference number: Pro00046018). Inclusion criteria for the study were adults 60-85 years of age, fluent in the language of the tests used and the test site, and with an MMSE score of 20-30 at Screening. Exclusion criteria were extensive and based on underlying conditions. All data collected as part of this study were deidentified to confidentiality protection. Participants in the study were compensated in order to cover any time or expense they incurred as a result of completing the study.

Results

Age was significantly different (P<.001) across cohort status groups. The mean total MMSE scores for the sample was 26.3 (SD 3.0), mean total adjusted RAVLT score was 38.8 (SD 14.3), while mean RAVLT long delay score was 5.4 (3.5). According to cohort status, 132 participants were deemed cognitively normal, 116 were considered to have MCI and 126 had probable AD (ADRD), 1 participant did not have a valid cohort status label. Combining the MCI and ADRD classes to produce 2 classes (Impaired and Intact) produced 242 participants with cognitive impairment and 132 deemed intact. Using MMSE and FAQ cut-offs of 28 and 6, respectively, along with RAVLT age group thresholds [18] to categorize participants as impaired or unimpaired, allowed a comparison of these labels against binary cohort status. MMSE agreed with cohort status with 73.8% (277/375) accuracy, RAVLT total score agreed with cohort status 45.7% (171/375), while RAVLT long recall delay score agreed with 84.8% (318/375) accuracy. Pearson's correlation coefficient between MMSE and RAVLT total score was 0.43, while correlation coefficient between MMSE and RAVLT long delay recall score was 0.60 (see Figure 3).

Figure 3. Relationship between the Mini Mental State Examination (MMSE) and Rey Auditory Verbal Learning Test (RAVLT) long recall delay score for Bio-Hermes population sample. Pearson correlation coefficient of 0.60 was observed between the 2 outcome measures. ADRD: Alzheimer disease–related dementia; MCI: mild cognitive impairment.

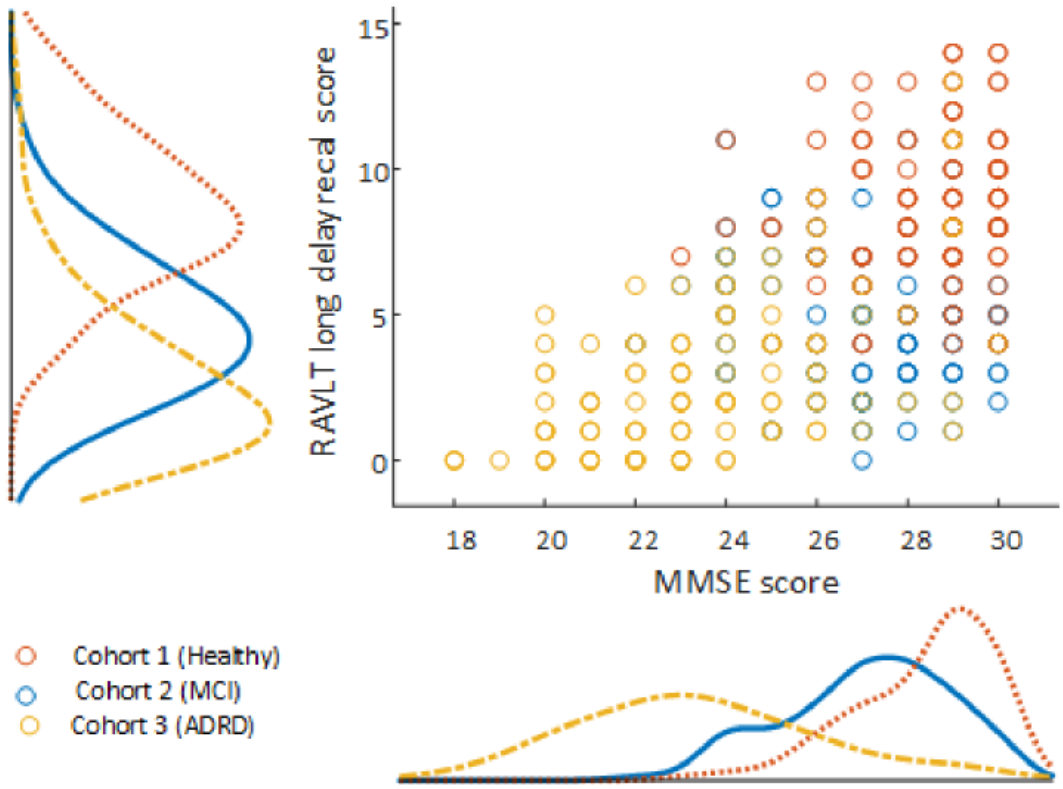

XSL•FO
RenderX

Ball Balancing Task Performance

All participants were able to complete the task under ST and DT conditions. The main metric of task performance was the percentage of time the virtual ball spent within the inner circle ("percentage time in circle"). Mean percentage time spent in the inner circle was 86.0% (SD 23.0%) and 66.1% (SD 35.8%) under ST and DT conditions, respectively, while the mean DT cost was 21.0% (SD 34.1%). Task performance was significantly different ($P<.05$) across cognitive status groups and between ST and DT conditions (Table 1). As expected, participants achieved lower performance under DT conditions with a higher mean percentage time within the inner circle and a lower proportion of "perfect score" tests (see Figures 4 and 5). Removing tests with "perfect" task performance (percentage time is circle equal to 100%) did not change this finding. Performance in the task declined with increased cognitive impairment, with best mean performance observed in the healthy group for both ST and DT and worst task performance in the ADRD group.

Task performance was statistically significantly different ($P<.05$) across group and between conditions.

Table 1. Ball balancing task performance: percentage (%) of time spent within the circle for each of the cohort groups (all participants, healthy, mild cognitive impairment [MCI], and Alzheimer disease–related dementia [ADRD]).

Group	Single task performance (%), mean (SD)	Dual task performance (%), mean (SD)	Dual task cost (%), mean (SD)
All	86.0 (23.0)	66.1 (35.8)	21.0 (34.1)
Healthy	92.6 (12.2)	70.2 (34.0)	18.8 (29.7)
MCI	87.5 (20.4)	63.7 (37.0)	21.9 (35.4)
ADRD	77.5 (30.0)	63.9 (35.9)	22.7 (37.1)

Figure 4. Ball balancing task performance measures per cohort status group and task condition.

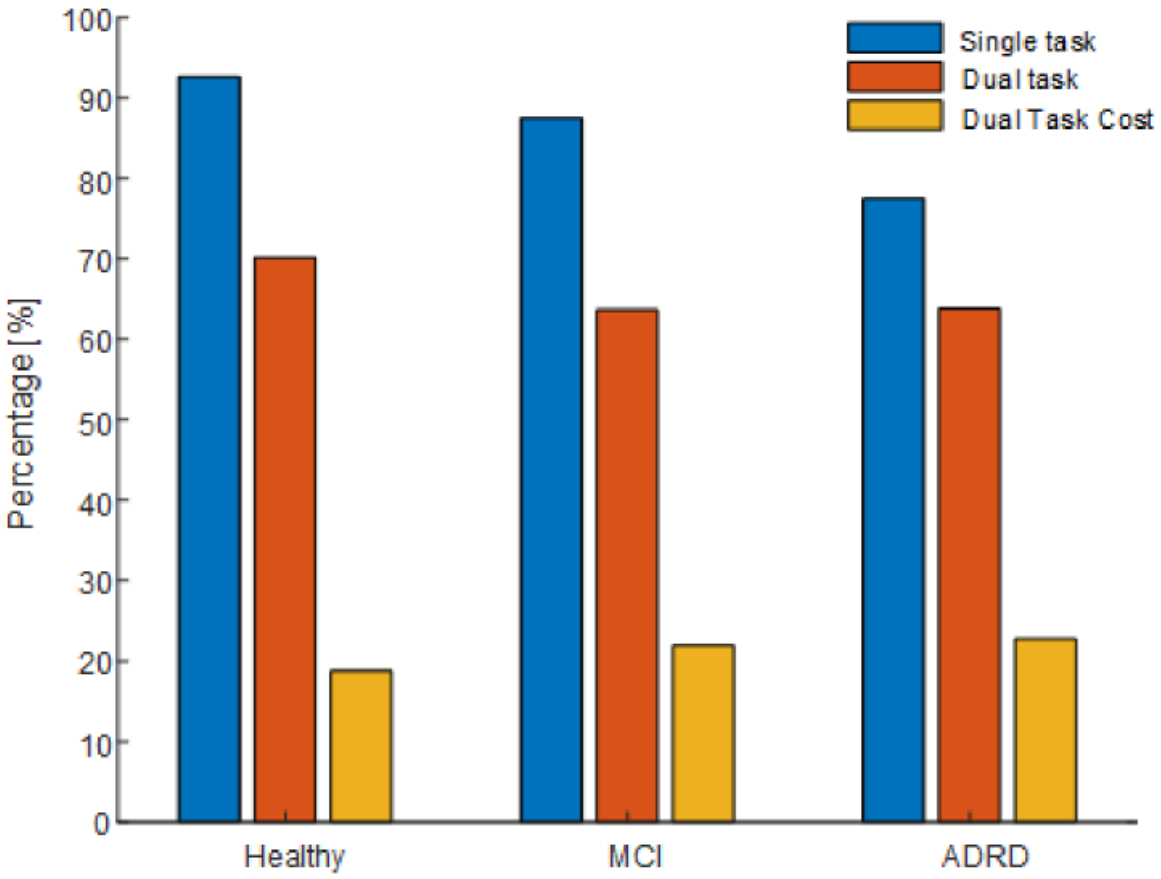

XSL•FO
RenderX

Figure 5. Box plot of ball balancing test performance (percentage of time within inner circle) under single (left panel) and dual (center panel) task conditions as well as dual task cost (right panel) for each cohort group (Healthy, mild cognitive impairment [MCI], and Alzheimer disease–related dementia [ADRD]).

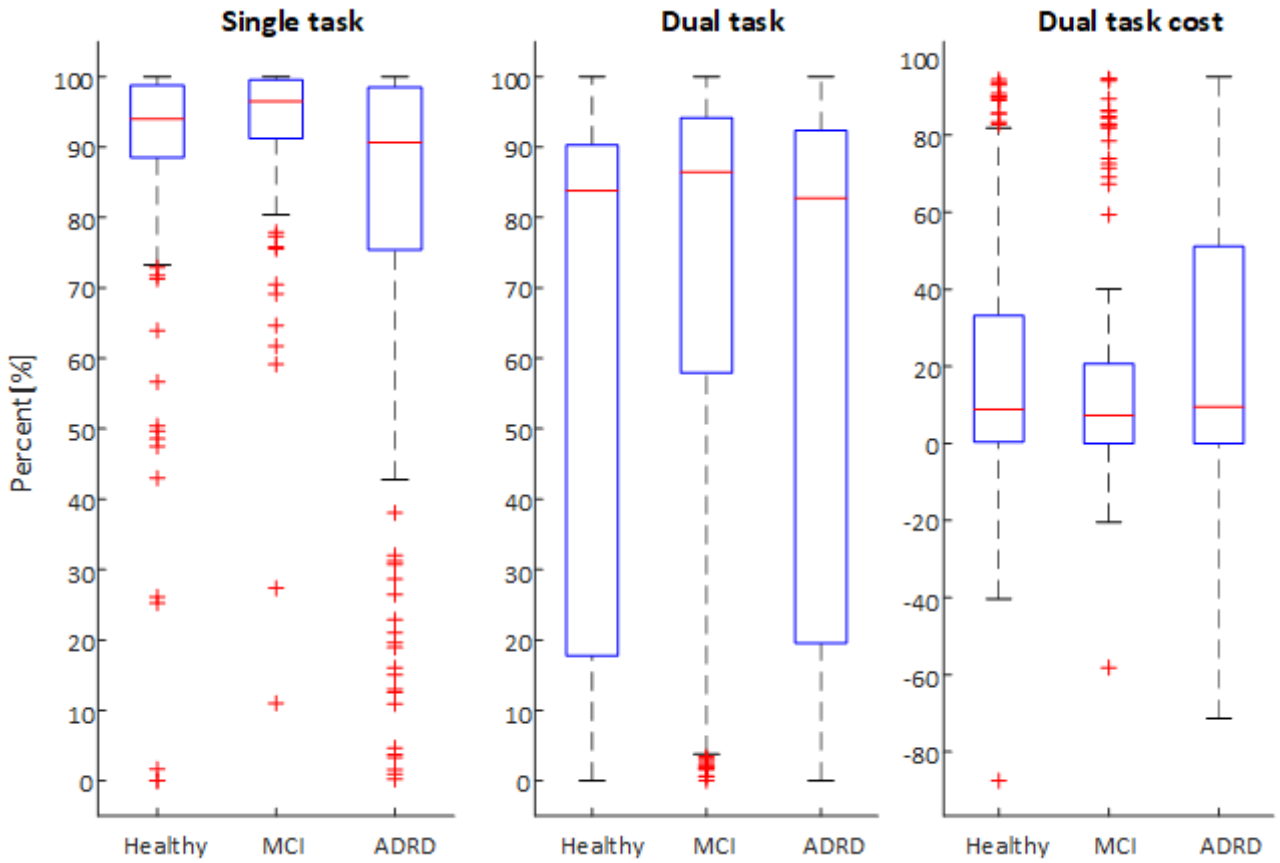

Exploratory Results for Cohort Status

Age is significantly different between impaired and nonimpaired groups. When controlling for age using ANOVA, a large number of calculated parameters below were significantly (P<.05) different on the basis of 3 category cohort status.

Similarly, when using ANOVA with a binary cohort label and correcting for age, a large number of parameters were significantly (P<.05) different on the basis of binary cognitive status. Figure 6 below details 2 IMU parameters where there were significant differences across groups when corrected for age.

Figure 6. Dual task cost of RMS acceleration and median frequency acceleration per cohort status. ADRD: Alzheimer disease–related dementia; MCI: mild cognitive impairment; RMS: root mean square.

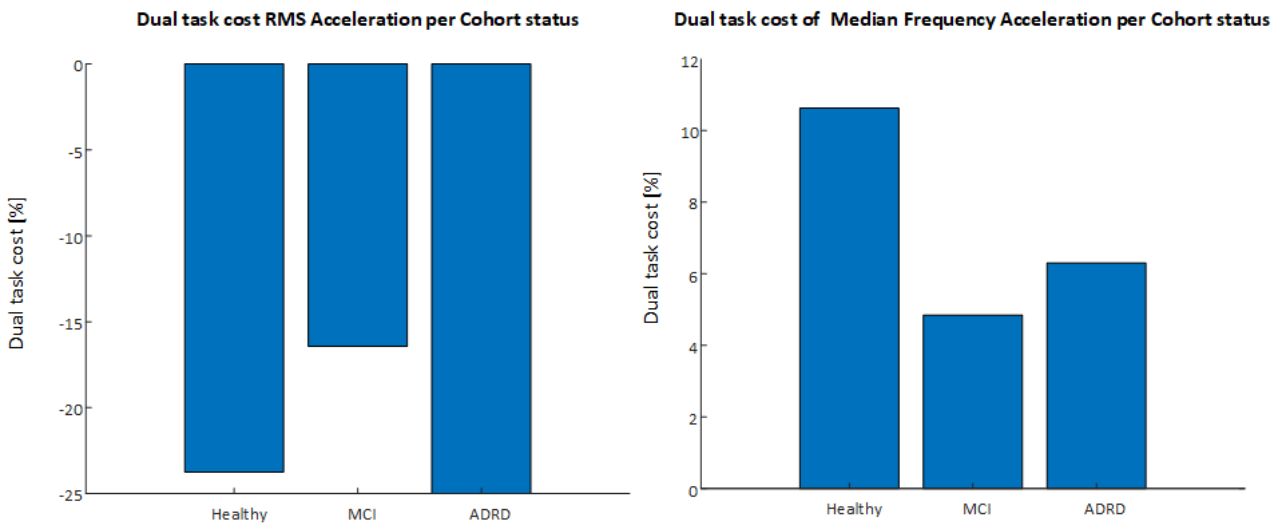

XSL•FO
RenderX

Classification Using Cohort Status

A linear logistic regression classifier model based on ball balancing measures from the DT condition (including age and gender) compared against a model obtained from age only found that ball balancing parameters could classify cognitive status with 70.5% accuracy compared to 62.3% accuracy for age alone (Table 2).

Table 2. Cross-validated logistic regression classification for models trained using cohort status.

	Ball balancing parameters				Age only			
	All	Male	Female	Mean	All	Male	Female	Mean
Accuracy (%)	66.67	72.12	68.93	*70.53* [a]	*63.47*	66.06	58.57	62.32
Sensitivity (%)	81.82	92.11	78.13	85.12	90.91	95.61	82.03	88.82
Specificity (%)	36.09	27.45	51.22	39.34	13.53	0.00	21.95	10.98
Positive predictive value (%)	69.96	73.94	71.43	72.69	65.67	68.13	62.13	65.13
Negative predictive value (%)	52.17	60.87	60.00	60.43	45.00	0.00	43.90	21.95

[a]Results are shown for a model based on ball balancing inertial measurement unit (IMU) parameters and a model based on age only. The best result per group is italicized. Results for separate gender stratified male and female models are reported as well as models based on All available data.

Exploratory Results of the FAQ

A small number of ball balancing parameters under both single and dual task conditions were significantly associated with functional impairment (as measured by the FAQ with a threshold of 6), independent of cognitive function (as measured by MMSE total score) and age. These parameters included dual task cost of task performance (percent time in circle), dual task median frequency acceleration and single task radial frequency. A similar analysis controlling for age and RAVLT long recall delay score found that several ball balancing parameters including median frequency acceleration and single task radial frequency were significantly associated with functional impairment.

RAVLT Exploratory Results

A number of ball balancing parameters under both single and dual task conditions were significantly different on the basis of cognitive status (using RAVLT long recall delay score, with age bucketed thresholds) to define cognitive impairment) and correcting for age.

Pearson correlation coefficient was used to calculate the correlation between the RAVLT (long recall delay score) and each of the calculated ball balancing parameters per task condition. Weak correlations were observed for a number of parameters.

Classification Using RAVLT Long Delay

A linear logistic regression using RAVLT long recall delay score with age bucket thresholds to denote impairment yielded a mean classification accuracy of 70.43% compared to 57.03% for age alone (Table 3).

Classification results for a model using age alone trained with the RAVLT long delay are also supplied. Results for separate genders stratified as male and female models are reported as well as models based on all available data.

Table 3. Classification results for ball balancing classifier model trained using Rey Auditory Verbal Learning Test (RAVLT) long delay score.

	Ball balancing parameters				Age only			
	All	Male	Female	Mean	All	Male	Female	Mean
Accuracy (%)	62.33	76.06	64.80	*70.43* [a]	54.67	61.21	52.86	57.03
Sensitivity (%)	40.51	30.69	47.87	39.28	74.36	100.00	14.89	57.45
Specificity (%)	30.56	35.94	31.03	33.49	33.33	0.00	83.62	41.81
Positive predictive value (%)	38.73	43.06	36.00	39.53	54.72	61.21	42.42	51.82
Negative predictive value (%)	32.16	24.73	42.35	33.54	54.55	0.00	54.80	54.80

[a]Italics are used to highlight the values most indicative of the true model accuracy.

Exploratory Results of the MMSE

A number of parameters under both single and dual task conditions were significantly different (using ANOVA and correcting for age) on the basis of cognitive status using MMSE, with a threshold of 28 to classify participants as cognitively impaired or cognitively unimpaired.

Pearson correlation coefficient was used to calculate the correlation between the MMSE (total score) and each of the

XSL•FO
RenderX

calculated ball balancing parameters per task condition. Weak correlations were observed for a number of parameters.

Classification Using the MMSE

A linear logistic regression using the MMSE total score with a threshold of 28 to denote impairment yielded a mean classification accuracy of 72.8% compared to 69.6% for age alone (Table 4).

Classification results for a model using age alone trained with the MMSE are also supplied. Results for separate genders stratified as male and female models are reported as well as models based on all available data.

Table 4. Classification results for ball balancing classifier model trained using the Mini Mental State Examination (MMSE) total score.

	Ball balancing parameters				Age only			
	All	M	F	Mean	All	M	F	Mean
Accuracy (%)	*72.78* [a]	71.52	72.86	72.19	69.60	67.27	70.48	68.87
Sensitivity (%)	90.80	94.64	96.64	95.64	99.23	98.21	97.32	97.76
Specificity (%)	28.95	22.64	14.75	18.70	1.75	1.89	4.92	3.40
Positive predictive value (%)	74.53	72.11	73.47	72.79	69.81	67.90	71.43	69.66
Negative predictive value (%)	57.89	66.67	64.29	65.48	50.00	33.33	42.86	38.10

[a]Italics are used to highlight the values most indicative of the true model accuracy.

Discussion

We introduce a novel dual task paradigm to evaluate cognitive reserve and prefrontal resource allocation that does not rely on gait and balance metrics and can, thus, be safely completed by older adults and those with falls risk. We found that older adults were able to complete the task regardless of their age or level of cognitive impairment. Even those with MCI and ADRD, as well as those with peripheral neuropathy, osteoarthritis, frailty, and other potential sources of gait and balance problems were able to complete the task reliably and safely.

A sample of 375 participants completed the dual task ball balancing test protocol. Participants ranged in age from 60 to 85 years and exhibited a wide range of cognitive ability. As predicted, participants achieved significantly higher ball balancing test performance under ST conditions (as measured by the percentage of test time, the ball was within the inner circle) compared to DT performance. Thus, along with the higher proportion of perfect tests under ST conditions, the findings confirm that participants were more challenged by the test under DT conditions and that task performance decreased with increasing cognitive impairment. We found that task performance was significantly improved in healthy individuals compared to those with MCI and that performance was worse again in those with ADRD.

A number of significant differences were observed between cognitively intact (unimpaired) and cognitively impaired participants for ball positioning and IMU parameters calculated during a ball balancing test, when correcting statistics for the effect of age, using cohort status, RAVLT long delay score, and MMSE to determine cognitive status.

Significantly decreased performance in the ball balancing test was observed during the DT compared to the ST. Similarly, decreased performance was observed for increasing levels of cognitive impairment. An interpretation of this result is that with increasing impairment, there needs to be greater reliance on cognitive reserve to sustain (or attempt to sustain) cognitive and functional performance. These results are in line with results reported in the literature for other DT paradigms, which suggest that task performance reduced during a DT as compared to an ST and that the reduction in task performance is increased with increased impairment [6,22]. As such, DT performance across different tasks becomes increasingly altered and with that increasingly correlated, while before the high DT cost suggesting impaired reserve (if present at all) might be detectable for some but not all DT conditions. Importantly, a number of ball balancing parameters, measured under both ST and DT conditions were found to be significantly associated with functional impairment (as measured by the FAQ score) independent of MMSE, RAVLT, and age. This suggests that differences observed between MCI to ADRD groups under dual task conditions are consistent with loss of cognitive reserve contributing to progression of clinical manifestation and impact on ADL [4,23]. The ball balancing dual task paradigm may, thus, offer a valuable, objective means to evaluate the risk of ADL impact and enable early detection of MCI-to-dementia transition risk [24].

Moderate classification performance (>70%) was also observed in classifying binary cognitive status using a logistic regression classifier model trained on each of the cognitive function outcome measures. This compared favorably to models based on age alone, which distinguished between impaired and unimpaired groups with ~60% accuracy. A simple linear classifier model (logistic regression) was used to obtain a baseline of classification performance; improved performance may be achieved through the addition of nonlinear interaction terms or the use of higher order classification methods (eg, support vector machines), given the wider data set and potential nonlinear statistical relationships between features. To provide an indication of how well the ball balancing test can distinguish cognitively impaired participants from cognitively intact participants, cross-validation and wrapper-based feature

XSL•FO
RenderX

selection was used. This method ensures unbiased estimate of classifier performance on previously unseen participants [25].

Three cognitive function outcome measures were considered in analyzing the use of the ball balancing test in classifying cognitive status. Each outcome measure (MMSE, RAVLT, and cohort status) contains differing and potentially complementary information about cognitive status (as evidenced by the modest mutual correlation observed between each outcome measure). In future work, we will examine the ability of a model based on the weighted combination of the 3 outcomes in longitudinally predicting cognitive impairment on a statistically independent data set. Furthermore, future work may also seek to examine the relationship of the ball balancing test parameters with blood biomarkers [26] and brain structure and pathology [5].

A limitation of this implementation of the ball balancing test is that the virtual ball is not perturbed during the test (other than by the movement of the tablet). This means that placement of the tablet on a flat, stable surface would allow the participant to achieve “perfect” task performance. However, it should be noted that the presence of “perfect score” tests were not found to affect the group-wise findings. An additional limitation is potential usability issues in using this task with an older adult population, particularly those with cognitive fine motor or visual impairments. While the current study involved participants conducting the task under supervised conditions to ensure adequate adherence to the task protocol, there may have been participants in the cognitively impaired groups who struggled to understand the instructions even with the support of the research assistant. Furthermore, impairment to fine motor skills may have prevented some participants from performing to their full capacity. Such usability issues may be exacerbated if the task were to be conducted under unsupervised conditions and would need to be carefully considered in the protocol for future studies.

The ball balancing test is a novel dual task paradigm that may have use in assessment of cognitive reserve and identification of cognitive impairment. Participants with mild or severe cognitive impairment performed less well on the task than healthy participants, particularly when a DT was introduced. A simple cross-validated classifier model used inertial sensor derived parameters obtained during the task to distinguish between cognitively impaired and cognitively intact participants with 70% accuracy. As the ball balancing test can be delivered entirely through a touchscreen tablet device, does not require a controlled environment, and is relatively simple to understand, the task may be suitable for administration by nonexpert users or for unsupervised use in the home environment and could support remote, longitudinal assessment of cognitive function.

Acknowledgments

The data that support the findings of this study were collected as part of the Bio-Hermes study and are governed by the Global Alzheimer Platform (GAP) consortium agreement. We would like to thank Dante Smith for his help in determining kinematic position of the virtual ball during the task. We would also like to thank the patients and clinicians involved in the study.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to the terms of the Global Alzheimer Platform (GAP) consortium agreement. However, the data are available from the corresponding author on reasonable request.

Authors' Contributions

Data analysis was carried out by BG. All authors contributed to manuscript preparation and analysis plan.

Conflicts of Interest

APL is a co-founder and Chief Medical Officer of Linus Health and declares ownership of shares or share options in the company. APL serves as a paid member of the scientific advisory boards for Neuroelectrics, Magstim Inc, TetraNeuron, Skin2Neuron, MedRhythms, and Hearts Radiant. All other authors are employees of Linus Health and declare ownership of shares or share options in the company.

References

1. Lee M, Whitsel E, Avery C, Hughes TM, Griswold ME, Sedaghat S, et al. Variation in population attributable fraction of dementia associated with potentially modifiable risk factors by race and ethnicity in the US. JAMA Netw Open 2022;5(7):e2219672 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.19672] [Medline: 35793088]
2. Ngandu T, Lehtisalo J, Solomon A, Levälahti E, Ahtiluoto S, Antikainen R, et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. Lancet 2015;385(9984):2255-2263. [doi: 10.1016/S0140-6736(15)60461-5] [Medline: 25771249]
3. van Dyck CH, Swanson CJ, Aisen P, Bateman RJ, Chen C, Gee M, et al. Lecanemab in early Alzheimer's disease. N Engl J Med 2023;388(1):9-21. [doi: 10.1056/NEJMoa2212948] [Medline: 36449413]

XSL•FO
RenderX

4. Stern Y, Albert M, Barnes CA, Cabeza R, Pascual-Leone A, Rapp PR. A framework for concepts of reserve and resilience in aging. Neurobiol Aging 2023;124:100-103 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2022.10.015] [Medline: 36653245]
5. Hupfeld KE, Geraghty JM, McGregor HR, Hass CJ, Pasternak O, Seidler RD. Differential relationships between brain structure and dual task walking in young and older adults. Front Aging Neurosci 2022;14:809281 [FREE Full text] [doi: 10.3389/fnagi.2022.809281] [Medline: 35360214]
6. Smith E, Walsh L, Doyle J, Greene B, Blake C. Effect of a dual task on quantitative timed up and go performance in community-dwelling older adults: a preliminary study. Geriatr Gerontol Int 2017;17(8):1176-1182. [doi: 10.1111/ggi.12845] [Medline: 27427514]
7. Yang Q, Tian C, Tseng B, Zhang B, Huang S, Jin S, et al. Gait change in dual task as a behavioral marker to detect mild cognitive impairment in elderly persons: a systematic review and meta-analysis. Arch Phys Med Rehabil 2020;101(10):1813-1821. [doi: 10.1016/j.apmr.2020.05.020] [Medline: 32553899]
8. Zhou J, Cattaneo G, Yu W, Lo O, Gouskova NA, Delgado-Gallén S, et al. The age-related contribution of cognitive function to dual-task gait in middle-aged adults in Spain: observations from a population-based study. Lancet Healthy Longev 2023;4(3):e98-e106 [FREE Full text] [doi: 10.1016/S2666-7568(23)00009-0] [Medline: 36870341]
9. Mayo AM. Use of the functional activities questionnaire in older adults with dementia. Hartford Inst Geriatr Nurs 2016;13:2 [FREE Full text]
10. Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(12):652-657. [doi: 10.1056/NEJM196803212781204] [Medline: 5637250]
11. Prieto T, Myklebust J, Hoffmann R, Lovett E, Myklebust B. Measures of postural steadiness: differences between healthy young and elderly adults. IEEE Trans Biomed Eng 1996;43(9):956-966. [doi: 10.1109/10.532130] [Medline: 9214811]
12. Somberg BL, Salthouse TA. Divided attention abilities in young and old adults. J Exp Psychol Hum Percept Perform 1982;8(5):651-663. [doi: 10.1037//0096-1523.8.5.651] [Medline: 6218227]
13. McManus K, Greene BR, Ader LGM, Caulfield B. Development of data-driven metrics for balance impairment and fall risk assessment in older adults. IEEE Trans Biomed Eng 2022;69(7):2324-2332. [doi: 10.1109/TBME.2022.3142617] [Medline: 35025734]
14. Greene BR, McManus K, Ader LGM, Caulfield B. Unsupervised assessment of balance and falls risk using a smartphone and machine learning. Sensors (Basel) 2021;21(14):4770 [FREE Full text] [doi: 10.3390/s21144770] [Medline: 34300509]
15. VanderPlas JT. Understanding the lomb–scargle periodogram. ApJS 2018;236(1):16. [doi: 10.3847/1538-4365/aab766]
16. Ferraris F, Grimaldi U, Parvis M. Procedure for effortless in-field calibration of three-axis rate gyros and accelerometers. Sens. Mater 1995;7(5):311-330.
17. Kukull W, Larson E, Teri L, Bowen J, McCormick W, Pfanschmidt M. The mini-mental state examination score and the clinical diagnosis of dementia. J Clin Epidemiol 1994;47(9):1061-1067. [doi: 10.1016/0895-4356(94)90122-8] [Medline: 7730909]
18. Mitrushina M, Boone KB, Razani J, D'Elia LF. Handbook of Normative Data for Neuropsychological Assessment, 2nd ed (Handbook of Normative Data for Neuropsychological Assessment, 2nd ed.). New York, NY, US: Oxford University Press; 2005:xxii-1029-xxii, 1029.
19. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97(1-2):273-324 [FREE Full text] [doi: 10.1016/s0004-3702(97)00043-x]
20. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
21. A R. The Clinical Examination in Psychology. Paris: Presses Universitaires de France; 1964.
22. Bio-hermes study. Global Alzheimers Platform. URL: https://globalalzplatform.org/biohermesstudy/ [accessed 2022-05-11]
23. O'Brien C, Holtzer R. Cognitive reserve moderates associations between walking performance under single- and dual-task conditions and incident mobility impairment in older adults. J Gerontol A Biol Sci Med Sci 2021;76(10):e314-e320 [FREE Full text] [doi: 10.1093/gerona/glab178] [Medline: 34153107]
24. Bahureksa L, Najafi B, Saleh A, Sabbagh M, Coon D, Mohler MJ, et al. The impact of mild cognitive impairment on gait and balance: a systematic review and meta-analysis of studies using instrumented assessment. Gerontology 2017;63(1):67-83 [FREE Full text] [doi: 10.1159/000445831] [Medline: 27172932]
25. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning. USA: Springer; 2009.
26. Montero-Odasso MM, Sarquis-Adamson Y, Speechley M, Borrie MJ, Hachinski VC, Wells J, et al. Association of dual-task gait with incident dementia in mild cognitive impairment: results from the gait and brain study. JAMA Neurol 2017;74(7):857-865 [FREE Full text] [doi: 10.1001/jamaneurol.2017.0643] [Medline: 28505243]

Abbreviations

ADL: activities of daily living
ADRD: Alzheimer disease–related dementia
BDST: backward digit span task

XSL•FO
RenderX

DT: dual task
GAP: Global Alzheimer Platform
IMU: inertial measurement unit
MCI: mild cognitive impairment
MMSE: Mini Mental State Examination
RAVLT: Rey Auditory Verbal Learning Test
ST: single task

Edited by A Mavragani; submitted 09.06.23; peer-reviewed by HF Hsieh, J Rider, B Poston; comments to author 21.11.23; revised version received 23.11.23; accepted 13.05.24; published 19.08.24.

Please cite as:
Greene B, Tobyne S, Jannati A, McManus K, Gomes Osman J, Banks R, Kher R, Showalter J, Bates D, Pascual-Leone A
The Dual Task Ball Balancing Test and Its Association With Cognitive Function: Algorithm Development and Validation
J Med Internet Res 2024;26:e49794
URL: https://www.jmir.org/2024/1/e49794
doi:10.2196/49794
PMID:39158963

©Barry Greene, Sean Tobyne, Ali Jannati, Killian McManus, Joyce Gomes Osman, Russell Banks, Ranjit Kher, John Showalter, David Bates, Alvaro Pascual-Leone. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Validity of Impressions as a Media Dose Metric in a Tobacco Public Education Campaign Evaluation: Observational Study

Kevin Davis[1], MA; Laurel Curry[2], DrPH; Brian Bradfield[1], BA; David A Stupplebeen[3], MPH, PhD; Rebecca J Williams[4], MPH, DrPH; Sandra Soria[4], CPA; Julie Lautsch[4], BA

[1]RTI International, Research Triangle Park, NC, United States

[2]RTI International, Washington, DC, United States

[3]Office of Public Health Studies, Thompson School of Social Work and Public Health, University of Hawai i at Mānoa, Honolulu, HI, United States

[4]California Tobacco Prevention Program, California Department of Public Health, Sacramento, CA, United States

Corresponding Author:
Laurel Curry, DrPH
RTI International
701 13th St. NW
Ste 750
Washington, DC, 20005
United States
Phone: 1 202 728 2086
Email: lcurry@rti.org

Abstract

Background: Evaluation research increasingly needs alternatives to target or gross rating points to comprehensively measure total exposure to modern multichannel public education campaigns that use multiple channels, including TV, radio, digital video, and paid social media, among others. Ratings data typically only capture delivery of broadcast media (TV and radio) and excludes other channels. Studies are needed to validate objective cross-channel metrics such as impressions against self-reported exposure to campaign messages.

Objective: This study aimed to examine whether higher a volume of total media campaign impressions is predictive of individual-level self-reported campaign exposure in California.

Methods: We analyzed over 3 years of advertisement impressions from the California Tobacco Prevention Program's statewide tobacco education campaigns from August 2019 through December 2022. Impressions data varied across designated market areas (DMAs) and across time. These data were merged to individual respondents from 45 waves of panel survey data of Californians aged 18-55 years (N=151,649). Impressions were merged to respondents based on respondents' DMAs and time of survey completion. We used logistic regression to estimate the odds of respondents' campaign recall as a function of cumulative and past 3-month impressions delivered to each respondent's DMA.

Results: Cumulative impressions were positively and significantly associated with recall of each of the Flavors Hook Kids (odds ratio [OR] 1.15, $P<.001$), Dark Balloons and Apartment (OR 1.20, $P<.001$), We Are Not Profit (OR 1.36, $P<.001$), Tell Your Story (E-cigarette, or Vaping, product use Associated Lung Injury; OR 1.06, $P<.05$), and Thrown Away and Little Big Lies (OR 1.05, $P<.01$) campaigns. Impressions delivered in the past 3 months were associated with recall of the Flavors Hook Kids (OR 1.13, $P<.001$), Dark Balloons and Apartment (OR 1.08, $P<.001$), We Are Not Profit (OR 1.14, $P<.001$), and Thrown Away and Little Big Lies (OR 1.04, $P<.001$) campaigns. Past 3-month impressions were not significantly associated with Tell Your Story campaign recall. Overall, magnitudes of these associations were greater for cumulative impressions. We visualize recall based on postestimation predicted values from our multivariate logistic regression models.

Conclusions: Variation in cumulative impressions for California Tobacco Prevention Program's long-term multichannel tobacco education campaign is predictive of increased self-reported campaign recall, suggesting that impressions may be a valid proxy for potential campaign exposure. The use of impressions for purposes of evaluating public education campaigns may help address current methodological limitations arising from the fragmented nature of modern multichannel media campaigns.

(J Med Internet Res 2024;26:e55311) doi:10.2196/55311

XSL•FO
RenderX

KEYWORDS

communication; public education; tobacco; media; public health

Introduction

Evaluations of mass media tobacco education campaigns have often used variation (over time and across markets) in campaign target ratings points (TRPs) or gross ratings points (GRPs) to estimate dose-response relationships between campaign advertisement buys and outcomes of interest (eg, Davis et al [1]; Farrelly et al [2]; Wakefield et al [3]). Ratings points are based on Nielsen ratings for programs during which TV or radio advertisements are shown, providing a metric of population-level reach and frequency of exposure to a given ad for specific audiences [4]. These metrics vary by designated market areas (DMAs) and over time. Because these metrics assume that individuals within a market's population have the same opportunity to view a campaign's advertisements, these metrics are only applicable for broadcast media channels [5].

An advantage of using GRPs or TRPs for media evaluation is that variation in ratings data is exogenous to study respondents and therefore not subject to potential selective attention biases that can arise from using self-reported measures of ad exposure. As such, this approach enables stronger conclusions about causality of campaign effects. However, a major limitation of this approach is that ratings data typically only capture delivery of broadcast media (TV and radio) and exclude other media channels. Therefore, this method is only applicable to evaluations of media campaigns that predominantly rely on broadcast channels in their media buys. In addition, the delivery of media content has shifted dramatically over the past decade toward a more fragmented system that simultaneously uses multiple nonbroadcast channels, including digital video, digital displays, and streaming audio. These channels are often used on their own or in combination with traditional broadcast media.

These changes to the modes and channels of media distribution complicate the evaluation of modern public health media campaigns. Traditional evaluation methods that include the use of GRPs may not adequately capture the breadth of a campaign's media delivery and related variation in that delivery. A recent review by Durkin et al [6] cataloged many of these challenges in the tobacco prevention media landscape, calling for an exploration of alternative methods and data to measure total campaign exposure across the spectrum of media channels that are used. These authors noted that studies to "validate self-reported exposure to digital campaign messages against objective digital metrics such as impressions, advertisements served, and completed video views" are needed [5-8].

Previous research has examined the relationships between digital tobacco education campaigns and audience response, including recall of advertisements and information seeking. For example, Kim et al [7] found that consumers who were verified as being exposed to a set of digital tobacco education advertisements were significantly more likely than unexposed consumers to visit websites that were promoted by the campaign. Other studies have used intentional geography-based manipulation of ad buys to show that markets receiving additional doses of digital video had higher overall campaign awareness compared with markets that received a conventional dose of standard TV and digital media [8]. Although these studies demonstrate correlation between nonbroadcast media delivery and overall campaign reach, they do not provide a basis for establishing alternatives to TRPs or ratings-based metrics for measuring overall delivery doses of a multichannel campaign. Validating the predictive relationships between an alternative comprehensive metric for ad delivery, such as impressions, and self-reported exposure to campaign advertisements requires more robust datasets with frequent collection of self-reported exposure merged with high variation in ad delivery metrics.

In this study, we use data from the California Tobacco Prevention Program's (CTPP's) tobacco education media campaign to examine the relationship between self-reported exposure to campaign messages and total campaign impressions across all channels. CTPP's statewide campaign is a dynamic multimessage campaign that targets numerous audiences using TV, radio, digital video, digital audio and display, and other nonbroadcast channels. In addition, CTPP's campaign ad buys are predominantly purchased based on guaranteed impressions units, not promised TRPs. As part of its broader campaign evaluation, CTPP tracks self-reported exposure to campaign messages and other outcomes using a large monthly survey of adults in California. Our study merges this robust survey data with detailed information on campaign impressions over time and across media markets to examine whether higher volume of total campaign impressions is predictive of individual-level self-reported campaign exposure in California. We discuss the implications and future directions of this research for using impressions as a valid exogenous proxy for total exposure to multichannel campaigns.

Methods

Survey Data

Survey data came from a monthly web-based tracking survey recruited from the Dynata consumer panel, a large commercially available panel of US adults. The data were collected in 45 waves from August 2019 through December 2022 and included respondents from California between 18 and 55 years of age (N=151,649). Dynata is a nonprobability convenience panel that uses a range of recruitment approaches, including email and marketing as well as targeted websites to yield a broad sample of US consumers. Dynata panelists are remunerated through an e-rewards system that provides redeemable points for merchandise for each completed survey. Each web survey was offered in English and Spanish to accommodate Spanish-speaking respondents who speak limited or no English or preferred to take the survey in Spanish. Sample sizes varied by wave, ranging from 2954-3576 per wave. Respondents were allowed to participate in more than 1 wave (34.9% of observations (n=52,971) were from repeat respondents). The survey data were weighted and calibrated to reflect distributions

XSL•FO
RenderX

(within Census region) of sex, age, race and ethnicity, education, smoking status (smoker vs nonsmoker), and daily internet use.

Measures

Recall of California Tobacco Prevention Program Campaigns

Recall of each of 5 CTPP campaigns that aired over the course of survey data collection from August 2019 to December 2022 were the primary outcome variables for this study: Flavors Hook Kids, We Are Not Profit, Dark Balloons and Apartment, Tell Your Story, and Thrown Away and Little Big Lies. The Flavors Hook Kids and We Are Not Profit campaigns highlighted the effects of flavored tobacco products on youth and Black populations and deceptive industry marketing tactics, among other themes. The Dark Balloons and Apartment campaigns focused on the dangers and ubiquity of secondhand smoke, particularly in outdoor and multiunit housing settings. CTPP's Tell Your Story campaign was focused on providing young adults with vaping cessation services in response to the E-cigarette or Vaping-Associated Lung Injury outbreak, while the Thrown Away and Little Big Lies campaigns highlighted the negative effects of discarded tobacco products, including the proliferation of microplastics in the environment. More information about CTPP's media campaigns can be found on the program's website [9].

Unaided Campaign Recall

We assessed unprompted general recall of the CTPP campaign to gauge overall awareness of the campaign in California. Respondents were asked "Have you recently seen or heard any California Department of Public Health antitobacco ads on TV, online, on radio, in magazines or on outdoor signs?" with response options of "yes" and "no." Respondents who answered "yes" were considered to have unaided recall of the overall California Department of Public Health antitobacco campaign.

Prompted Ad Recall

Specific ad recall was measured using an aided ad recognition protocol where after viewing videos of each campaign ad, respondents were asked "Have you seen this ad on TV or online?" Respondents could answer "yes" or "no" to each ad-specific recall question. A similar protocol was used to construct recall of radio advertisements. After listening to each radio advertisement (or reading the script if the ad playback function did not work properly), respondents were asked "Have you heard this ad on the radio?" with response options of "yes" and "no." Respondents were also asked to recall whether they had seen any outdoor or online static display-style advertisements, through the following prompt: "Next, we would like to show you an image of a few ads that may have appeared on outdoor signs in California such as on buses, billboards, gas stations, or public places, or on websites and other places online." After viewing images of these advertisements, respondents selected "yes" or "no" to indicate having seen the advertisements previously or not. We then created campaign-specific awareness variables indicating awareness of at least one ad across all ad types (video, radio, or static). These campaign-specific ad recall variables served as the dependent variables in each of our analytic models.

Exogenous Campaign Exposure Measured by Impressions

We used weekly impressions data provided by the campaign's media buyer to measure potential exposure to CTPP's paid media. Impressions are defined as the number of times a piece of media displays on a user's screen (or plays on an audio channel). These data were provided at the DMA level for the 14 DMAs in California and were calculated separately for each campaign aired by CTPP across multiple target audiences of adults in California. Impressions were included for each media channel used by the campaign (broadcast, digital, and display) and combined to provide a compilation of total potential exposure across all channels. We measured total cumulative impressions and past 3-month impressions for each of CTPP's campaigns on air during this time period: Flavors Hook Kids, Dark Balloons and Apartment, Tell Your Story, Thrown Away and Little Big Lies, and We Are Not Profit campaigns. Impression values were linked to individual survey respondents based on survey timing and respondent location in DMAs. Each survey participant was assigned the full week of impressions data for the week they took the survey, no matter which day of the week they participated. Cumulative impressions were totaled from campaign launch through the date each participant took the survey for the DMA in which they reported residing. Past 3-month impressions were similarly totaled, but from the 3 months before the respondent's survey date.

Mean values of cumulative and past 3-month impressions by campaign are summarized in Table 1. While these values reflect the market-level mean impressions, respondents in the sample can experience a range of impression levels (from low to high) based on the market where they reside.

Table 1. Mean Market-level California Tobacco Prevention Program's cumulative campaign impressions, August 2019 to December 2022.

Campaign	Cumulative impressions, mean (SD)[a]	Past 3-month impressions, mean (SD)
All campaigns	328,789,592 (586,912,213)	42,758,557 (90,205,224)
Flavors Hook Kids	152,422,867 (238,477,527)	17,052,572 (40,158,606)
Dark Balloons and Apartment	112,884,348 (208,276,945)	9,870,736 (46,101,224)
We Are Not Profit	40,224,813 (99,809,204)	9,413,460 (31,246,961)
Tell Your Story	13,277,377 (39,494,383)	2,733,164 (12,124,264)
Thrown Away and Little Big Lies	9,980,187 (38,762,599)	3,688,625 (20,496,183)

[a]There were 176 weeks and 14 unique DMAs (n=2464 weeks x DMA observations) represented in the analytic data.

XSL•FO
RenderX

Control Variables

Our analysis controlled for several factors that may influence recall of campaign advertisements. We created control variables for respondent demographics including age (18-55 years); race and ethnicity (White non-Hispanic, Black non-Hispanic, Asian non-Hispanic, other non-Hispanic, and Hispanic); gender (male, female, transgender, or other or unsure); sexual orientation (straight or do not know and LGBQ [lesbian, gay, bisexual, and queer/questioning]); education (less than high school, high school degree, some college, associate degree, bachelor's degree or higher); and income (low income—less than US $45,000, middle to high income—US $45,000 or more). We also controlled for any past 30-day use of tobacco products (including e-cigarettes) or marijuana. Finally, we controlled for background secular trends in campaign recall by measuring total quarters elapsed since the beginning of the study (August 2019).

Statistical Analysis

We began by summarizing average campaign recall for each of the CTPP campaigns described earlier. Campaign recall was estimated separately for each year in our data (2019-2022) and for all years combined. We then used logistic regression to estimate the odds of campaign-specific recall as a function of cumulative and past 3-month impressions delivered to each respondent's DMA. Impressions were scaled to produce odds ratios that reflected the increase in campaign recall for delivery of the mean past 3-month and cumulative impressions levels. This scaling provides a realistic interpretation of the relationship between impressions and campaign recall based on mean levels of impressions delivered for each campaign in the real world. In addition, because we did not expect increasing levels of paid impressions to perpetually generate linear increases in recall, we used a nonlinear functional form for impressions (square root) to capture asymptotic diminishing effects over the range of observed impressions. Each model included control variables for the respondent characteristics described earlier, as well as a linear time trend for quarters since the first survey wave of the study in August 2019.

To examine whether channel-specific impressions were more or less strongly predictive of campaign recall, we conducted the same analysis using 3 separate models: one with just video impressions, another with just audio impressions, and another with just display impressions. We then compared the odds ratios (ORs) across those 3 model types to determine whether different impression types were more strongly predictive of overall recall than others.

To further illustrate the estimated relationship between campaign impressions and recall, we calculated postestimation predicted probabilities of overall prompted campaign recall across observed values of past 3-month cumulative impressions for the combined CTPP campaigns. We chose to use past 3-month impressions as the basis for these predictions as this is a more common time frame for real-world ad buys. These predictions were generated from the multivariate logistic regression model results to provide interpretable ranges of campaign recall across the observed range of cumulative impressions with all covariates in our statistical model held constant. Because a small number of respondents in certain markets receive very high impressions levels, these predictions exclude the highest 5% of impression values. This helps ensure that our graphical representation of the dose-response relationship between impressions and ad recall reflects the experience of most respondents and is not skewed by outlier impressions values. All analyses were conducted using Stata (version 17.0, Computing Resource Center).

Ethical Considerations

RTI International's Institutional Review Board reviewed this research protocol and determined that it did not meet the definition of research with human subjects because it was for program evaluation purposes. Nevertheless, all study participants gave informed consent for panel participation in addition to assurances that all study participation is voluntary, and any question could be skipped if desired.

Results

Sample Characteristics

The combined unweighted sample included 151,649 observations on adults aged 18 to 55 years in California collected from August 2019 to December 2022. The mean unweighted age of the sample was 38.9 years and consisted of individuals across the following demographics: 69,627 White non-Hispanic (45.9%), 14,904 Black non-Hispanic (9.8%), 23,812 Asian non-Hispanic (15.7%), 8436 other non-Hispanic races (5.6%), and 34,858 Hispanic (23.0%) adults. The unweighted and weighted sample distributions of gender, sexual orientation, education, income, and tobacco and other product use are summarized in Table 2.

XSL•FO
RenderX

Table 2. Unweighted sample demographics and other characteristics.

Respondent characteristic	Unweighted percentage	Weighted percentage	Participants, n
Age (years), mean	38.92	36.27	__[a]
18-29	18.7	31	28,360
30-55	81.3	69	123,289
Race and Ethnicity			
White non-Hispanic	45.9	34.7	69,627
Black non-Hispanic	9.8	6.1	14,904
Asian non-Hispanic	15.7	17.5	23,812
Other non-Hispanic	5.6	1.4	8436
Hispanic	23.0	40.2	34,858
Gender			
Male	40.7	49.1	61,577
Female	57.9	48.8	87,595
Transgender	0.7	1.3	1091
Other	0.5	0.5	717
Unsure	0.2	0.3	267
Sexual orientation			
Straight or do not know	87.7	86.7	130,404
LGBQ[b]	12.3	13.3	18,252
Education			
Less than high school	2.9	15.4	4446
High school degree	14.8	21.6	22,455
Some college	21.6	20.7	32,750
Associate degree	11.9	11.3	18,088
Bachelor's degree or higher	48.7	31.1	73,902
Income			
Low income (less than US $45,000)	33.7	45.4	51,042
Middle to high income (US $45,000 or more)	66.3	54.6	100,471
Tobacco and other product use			
Smoked cigarettes in past 30 days	30.3	9.4	45,946
Used electronic vape products in past 30 days	25.4	13.6	38,520
Used marijuana in past 30 days	32.5	22.3	49,321
Used other tobacco products in past 30 days	23.2	14.5	35,200

[a]—: not applicable.

[b]LGBQ: lesbian, gay, bisexual, and queer/questioning.

Campaign Recall

Table 3 summarizes descriptive statistics on campaign recall by year from 2019 to 2022. Across all years in our analysis, 68.3% (53,137) of respondents indicated awareness of any ad in the CTPP Flavors Hook Kids campaign. During this time frame, recall of the We Are Not Profit campaign was 59.1% (16,920) while recall of the Dark Balloons and Apartment, Tell Your Story, and Thrown Away and Little Big Lies campaigns were 54.8% (27,748), 53.2% (16,886), and 50.8% (15,450), respectively. Overall, 65.6% (87,814) of adults aged 18 to 55 years old in California reported awareness of any CTPP campaign ad across the combined time period from 2019 to 2022. Unprompted general recall of antitobacco advertisements in California across all years was 51.5% (78,090).

XSL•FO
RenderX

Table 3. Recall of California Tobacco Prevention Program campaigns among all adults aged 18-55 years old in California, August 2019 to December 2022.

Campaign	2019 (n=19,004)	2020 (n=45,509)	2021 (n=39,584)	2022 (n=47,552)	All years (n=151,649)
Prompted Ad Recall					
All campaigns	60.9%	67.7%	68.5%	62.7%	65.6%
Flavors Hook Kids	65.8%	65.4%	74.1%	67.0%	68.3%
We Are Not Profit	__[a]	47.8%	62.4%	__[a]	59.1%
Dark Balloons and Apartment	51.4%	59.2%	49.9%	__[a]	54.8%
Tell Your Story	__[a]	__[a]	47.5%	58.8%	53.2%
Thrown Away and Little Big Lies	__[a]	49.3%	__[a]	51.5%	50.8%
Unaided campaign recall	56.5%	49.4%	49.7%	53.1%	51.5%

[a]__: not applicable.

Relationship Between Campaign Impressions and Recall

Cumulative and past 3-month impressions were widely associated with prompted campaign recall (Table 4). Cumulative impressions were positively and significantly associated with recall of each of the Flavors Hook Kids (OR 1.15, *P*<.001), Dark Balloons and Apartment (OR 1.20, *P*<.001), We Are Not Profit (OR 1.36, *P*<.001), Tell Your Story (OR 1.06, *P*<.05), and Thrown Away and Little Big Lies (OR 1.05, *P*<.01) campaigns. Impressions delivered in the past 3 months were associated with recall of the Flavors Hook Kids (OR 1.13, *P*<.001), Dark Balloons and Apartment (OR 1.08, *P*<.001), We Are Not Profit (OR 1.14, *P*<.001), and Thrown Away and Little Big Lies (OR 1.04, *P*<.001) campaigns. Past 3-month impressions were not significantly associated with Tell Your Story campaign recall. Overall, magnitudes of these associations were greater for cumulative impressions.

Similar associations between impressions and unaided campaign recall were found. Past 3-month impressions (OR 1.07, *P*<.001) and cumulative impressions (OR 1.05, *P*<.01) were both significantly associated with unaided campaign recall.

Table 4. Odds ratios and 95% CIs from logistic regressions of associations between campaign impressions and campaign recall among all adults in California aged 18-55 years, August 2019-December 2022.

Prompted ad recall	Cumulative impressions, OR (95% CI)	Past 3-month impressions, OR (95% CI)
All campaigns	1.16[a] (1.11-1.20)	1.20[a] (1.17-1.23)
Flavors Hook Kids	1.15[a] (1.09-1.22)	1.13[a] (1.10-1.17)
Dark Balloons and Apartment	1.20[a] (1.13-1.27)	1.08[a] (1.06-1.11)
We Are Not Profit	1.36[a] (1.27-1.46)	1.14[a] (1.09-1.20)
Tell Your Story	1.06[b] (1.01-1.11)	1.01 (0.98-1.04)
Thrown Away and Little Big Lies	1.05[c] (1.01-1.09)	1.04[a] (1.02-1.07)
Unaided campaign recall	1.04[c] (1.01-1.08)	1.08[a] (1.05-1.10)

[a]*P* (overall): <.001.

[b]*P* (overall): <.05.

[c]*P* (overall): <.01.

We divided the impressions variables used in the models by the overall sample mean to produce odds ratios that give the percent change in recall for a unit change in impressions where the unit change is the sample mean of the impressions. To illustrate, for cumulative impressions, going from no campaign (ie, 0 impressions) to the average campaign delivery (328 million impressions) generates a 16% increase in odds of ad recall. This scaling provides a realistic unit of change for impressions that reflects the size of a real-world ad buy for the CTPP campaign. Results from our sensitivity analysis showed similar relationships between the channel-specific impressions (video, audio, and display) and campaign awareness measured through both prompted and unaided recall (results not shown). Given that CTPP used a relatively balanced mixture of these media channels concurrently, combined impression should best reflect the overall relationship between campaign advertisements buys and reach among target audiences in California.

Multimedia Appendix 1 summarizes estimated campaign recall rates based on postestimation predicted values from our multivariate logistic regression model for overall campaign recall. Predicted recall for all campaigns combined ranged from

XSL•FO
RenderX

63.7% at 100 million past 3-month cumulative impressions to 77.0% at 800 million past 3-month cumulative impressions.

Discussion

Principal Findings

Results from this study show that variation in cumulative impressions for CTPP's long-term multichannel tobacco education campaign is predictive of increased self-reported campaign recall, suggesting that impressions may be a valid proxy for potential campaign exposure. Because impressions are generally calculated and known for all media channels, the use of impressions for purposes of evaluating public education campaigns may help address current methodological limitations arising from the fragmented nature of modern multichannel media campaigns [6]. That is, impressions can provide a single metric that captures whole-campaign exposure regardless of a campaign's media channel mixture. The use of impressions may be particularly helpful in the evaluation of campaigns that dedicate substantial portions of ad buys to nonbroadcast channels, including digital video, display, and out-of-home advertising.

Our study also provides one of the first real-world examples of how impressions data can be applied in the evaluation of a major multichannel public health awareness campaign. In our application of these data, there is a high degree of variation in impressions over a long period of time. Impressions variation, in this context, arises from differences in ad buy sizes across media markets and from time-specific accumulation of impressions. This variation in impressions is paired with high-sample, frequent (roughly monthly) surveys of individual respondents over a data collection period exceeding 3 years. These conditions likely contributed to a high degree of statistical power to detect relationships between impressions at the market-time level and recall measured at the individual level. Future evaluation research that aims to use impressions for evaluating multichannel campaigns should consider whether the available variation in both impressions and respondent data are sufficient to statistically capture relationships between impressions and respondent-level constructs.

Although our analysis was able to detect positive and statistically significant relationships between impressions and recall for each of the CTPP campaigns we measured, the pattern of predicted recall across observed impressions levels (Multimedia Appendix 1) suggests the overall magnitude of this relationship was modest. This may be due to a high level of background exposure to tobacco control media campaigns in California. Historically, California has a long track record of well-funded tobacco education media campaigns in addition to strong tobacco control policies [10,11]. These campaigns and policies have been implemented over a long period of time that precedes the time frame of our study. Given the likely high preexisting awareness of tobacco-related messages and related issues in California, the effects of ongoing awareness-raising media campaigns may be incremental. Therefore, our results should not be interpreted as a prescription of expected dose-responses for other campaigns that may occur in different contexts. Rather, the historical context of past campaign efforts and existing public awareness and knowledge of campaign-targeted messages should be considered when evaluating these types of campaigns.

Limitations

There are several limitations to this analysis that should form the basis of next steps and recommendations for future research with similar data. First, impressions measure potential, not confirmed individual-level exposure. Although our results suggest that impressions are predictive of self-reported campaign recall, future research can further validate impressions as a proxy for exposure by applying them to evaluate campaign effects on other respondent-level outcomes, including campaign-targeted attitudes, beliefs, and behaviors.

Second, we did not examine the effects of channel-specific impressions on self-reported recall and therefore cannot make conclusions about the effects of different mixtures of media channels. This leaves unanswered questions about the equivalency of impressions across different media channels. For example, do impressions for video advertisements generate greater exposure rates than the same level of impressions for digital display advertisements? Furthermore, within video advertisements, does delivery through streaming services result in different exposure levels than those delivered by broadcast? Disentangling the unique contributions of each media channel is challenging, since creative elements including visuals, main messages, and calls to action, are usually very similar across all media channels used within a single campaign. The estimation of channel-specific effects would require the production of sufficiently different creative content across channels and more extensive coordination of ad buys set up for the purposes of evaluation tests. This level of coordination may not be possible for most campaigns, given limited resources and focus on concise sets of outcomes and objectives. Therefore, for purposes of addressing channel-specific effects, we recommend that future studies at least consider examining multiple campaigns that have differing creative content in addition to different mixtures of media channels.

An additional limitation is that although impressions have a straightforward definition, they are difficult to interpret, given the volumes that are delivered for individual campaigns, which often exceed tens of millions. To put results in greater context, future studies should consider reporting additional information, when available, about the campaign ad buys such as overall investment and expenditure levels for each media channel. These additional data can potentially facilitate helpful return on investment analysis as public health programs consider options for campaign ad buys with varying levels of purchased impressions.

Finally, we note that while the survey data are weighted to reflect the demographic profile of California, the data are not recruited using a statistical probability sample. Because the data are recruited from a large opt-in consumer panel, the unweighted data is skewed toward heavier representation of some groups such as older and more educated adults. This results in increased weighting for less represented groups and should be considered when interpreting the overall results.

XSL•FO
RenderX

Conclusion

Measurement of digital media delivery and its application in the evaluation of multichannel public health campaigns is an emerging science. This study provides preliminary evidence that combined campaign impressions are predictive of overall campaign recall at the respondent level. Therefore, impressions may provide a valid alternative exposure variable that captures all media channels in evaluations of modern multichannel campaigns. Future research can build on this study by (1) exploring relationships between impressions and other campaign-targeted outcomes; (2) implementing study designs that attempt to distinguish the unique effects of each media channel; and (3) examining additional information on campaign investments to aid in the interpretation of evaluations that use impressions data.

Acknowledgments

This work was funded under contract (#18-10190) from the California Department of Public Health.

Data Availability

The dataset generated and analyzed during this study is not publicly available, but may be available from the sponsoring agency (CTPP) upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Predicted values of campaign recall by observed all-campaign impressions levels among all adults in California aged 18-55 years, August 2019-December 2022.
[PNG File , 91 KB - jmir_v26i1e55311_app1.png]

References

1. Davis KC, Patel D, Shafer P, Duke J, Glover-Kudon R, Ridgeway W, et al. Association between media doses of the Tips from Former Smokers Campaign and cessation behaviors and intentions to quit among cigarette smokers, 2012-2015. Health Educ Behav 2017 May 12;45(1):52-60 [FREE Full text] [doi: 10.1177/1090198117709316] [Medline: 28497703]
2. Farrelly MC, Duke JC, Davis KC, Nonnemaker JM, Kamyab K, Willett JG, et al. Promotion of smoking cessation with emotional and/or graphic antismoking advertising. American Journal of Preventive Medicine 2012 Nov;43(5):475-482. [doi: 10.1016/j.amepre.2012.07.023] [Medline: 23079169]
3. Wakefield MA, Spittal MJ, Yong HH, Durkin SJ, Borland R. Effects of mass media campaign exposure intensity and durability on quit attempts in a population-based cohort study. Health Education Research 2011 Jul 05;26(6):988-997 [FREE Full text] [doi: 10.1093/her/cyr054] [Medline: 21730252]
4. Farris PW, Bendle NT, Pfeifer PE, Reibstein DJ. Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. New Jersey: Pearson Education; 2010.
5. Romberg AR, Bennett M, Tulsiani S, Simard B, Kreslake JM, Favatas D, et al. Validating self-reported ad recall as a measure of exposure to digital advertising: an exploratory analysis using ad tracking methodology. IJERPH 2020 Mar 25;17(7):2185 [FREE Full text] [doi: 10.3390/ijerph17072185] [Medline: 32218265]
6. Durkin SJ, Brennan E, Wakefield MA. Optimising tobacco control campaigns within a changing media landscape and among priority populations. Tob Control 2022 Mar 03;31(2):284-290. [doi: 10.1136/tobaccocontrol-2021-056558] [Medline: 35241601]
7. Kim A, Hansen H, Duke J, Davis K, Alexander R, Rowland A, et al. Does digital ad exposure influence information-seeking behavior online? Evidence from the 2012 Tips from Former Smokers National Tobacco Prevention Campaign. J Med Internet Res 2016;18(3):e64 [FREE Full text] [doi: 10.2196/jmir.4299] [Medline: 26983849]
8. Davis KC, Shafer PR, Rodes R, Kim A, Hansen H, Patel D, et al. Does digital video advertising increase population-level reach of multimedia campaigns? Evidence from the 2013 Tips From Former Smokers Campaign. J Med Internet Res 2016;18(9):e235 [FREE Full text] [doi: 10.2196/jmir.5683] [Medline: 27627853]
9. : UNDO URL: https://www.undo.org/ [accessed 2024-10-09]
10. Rogers T. The California Tobacco Control Program: introduction to the 20-year retrospective. Tob Control 2010 Apr 09;19(Suppl 1):i1-i2 [FREE Full text] [doi: 10.1136/tc.2010.036293] [Medline: 20382644]
11. Rohrbach LA, Howard-Pitney B, Unger JB, Dent CW, Howard KA, Cruz TB, et al. Independent evaluation of the California Tobacco Control Program: relationships between program exposure and outcomes, 1996–1998. Am J Public Health 2002 Jun;92(6):975-984. [doi: 10.2105/ajph.92.6.975] [Medline: 12036792]

XSL•FO
RenderX

Abbreviations

CTPP: California Tobacco Prevention Program
DMA: designated market area
GRP: gross ratings point
LGBQ: lesbian, gay, bisexual, and queer/questioning
OR: odds ratio
TRP: target ratings point

Edited by A Mavragani; submitted 19.12.23; peer-reviewed by MN Shalaby, S Durkin; comments to author 23.02.24; revised version received 26.04.24; accepted 07.09.24; published 05.11.24.

Please cite as:
Davis K, Curry L, Bradfield B, Stupplebeen DA, Williams RJ, Soria S, Lautsch J
The Validity of Impressions as a Media Dose Metric in a Tobacco Public Education Campaign Evaluation: Observational Study
J Med Internet Res 2024;26:e55311
URL: https://www.jmir.org/2024/1/e55311
doi:10.2196/55311
PMID:

©Kevin Davis, Laurel Curry, Brian Bradfield, David A Stupplebeen, Rebecca J Williams, Sandra Soria, Julie Lautsch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identifying and Estimating Frailty Phenotypes by Vocal Biomarkers: Cross-Sectional Study

Yu-Chun Lin[1,2], MD, PhD; Huang-Ting Yan[3], PhD; Chih-Hsueh Lin[4,5*], MD, PhD; Hen-Hong Chang[1,2,6*], MD, PhD

[1]Graduate Institute of Integrated Medicine, College of Chinese Medicine, China Medical University, Taichung, Taiwan
[2]Department of Chinese Medicine, China Medical University Hospital, Taichung, Taiwan
[3]Institute of Political Science, Academia Sinica, Taipei, Taiwan
[4]School of Medicine, College of Medicine, China Medical University, Taichung, Taiwan
[5]Department of Family Medicine, China Medical University Hospital, Taichung, Taiwan
[6]Chinese Medicine Research Centre, China Medical University, Taichung, Taiwan
[*]these authors contributed equally

Corresponding Author:
Hen-Hong Chang, MD, PhD
Graduate Institute of Integrated Medicine
College of Chinese Medicine
China Medical University
No 91, Hsueh-Shih Road
North District
Taichung, 40402
Taiwan
Phone: 886 22053366 ext 3609
Email: tcmchh55@gmail.com

Abstract

Background: Researchers have developed a variety of indices to assess frailty. Recent research indicates that the human voice reflects frailty status. Frailty phenotypes are seldom discussed in the literature on the aging voice.

Objective: This study aims to examine potential phenotypes of frail older adults and determine their correlation with vocal biomarkers.

Methods: Participants aged ≥60 years who visited the geriatric outpatient clinic of a teaching hospital in central Taiwan between 2020 and 2021 were recruited. We identified 4 frailty phenotypes: energy-based frailty, sarcopenia-based frailty, hybrid-based frailty–energy, and hybrid-based frailty–sarcopenia. Participants were asked to pronounce a sustained vowel "/a/" for approximately 1 second. The speech signals were digitized and analyzed. Four voice parameters—the average number of zero crossings (A1), variations in local peaks and valleys (A2), variations in first and second formant frequencies (A3), and spectral energy ratio (A4)—were used for analyzing changes in voice. Logistic regression was used to elucidate the prediction model.

Results: Among 277 older adults, an increase in A1 values was associated with a lower likelihood of energy-based frailty (odds ratio [OR] 0.81, 95% CI 0.68-0.96), whereas an increase in A2 values resulted in a higher likelihood of sarcopenia-based frailty (OR 1.34, 95% CI 1.18-1.52). Respondents with larger A3 and A4 values had a higher likelihood of hybrid-based frailty–sarcopenia (OR 1.03, 95% CI 1.002-1.06) and hybrid-based frailty–energy (OR 1.43, 95% CI 1.02-2.01), respectively.

Conclusions: Vocal biomarkers might be potentially useful in estimating frailty phenotypes. Clinicians can use 2 crucial acoustic parameters, namely A1 and A2, to diagnose a frailty phenotype that is associated with insufficient energy or reduced muscle function. The assessment of A3 and A4 involves a complex frailty phenotype.

(J Med Internet Res 2024;26:e58466) doi:10.2196/58466

KEYWORDS

frailty phenotypes; older adults; successful aging; vocal biomarkers; frailty; phenotype; vocal biomarker; cross-sectional; gerontology; geriatrics; older adult; Taiwan; energy-based; hybrid-based; sarcopenia

XSL•FO
RenderX

Introduction

Frailty is a prevalent geriatric syndrome characterized by age-dependent declines in the functioning of multiple organ systems, resulting in elevated susceptibility to stressors and a greater likelihood of adverse health outcomes [1] and mortality [2]. In population-based studies, the global prevalence of frailty among people older than 50 years is estimated to range from 12% to 24% [3].

Frailty can be roughly categorized as physical, cognitive, and social. Researchers have developed a variety of indices to assess physical frailty, including the Frailty Index scale [4]; the Clinical Frailty Scale [5]; the Cardiovascular Health Study (CHS) index [6]; the Study of Osteoporotic Fractures index [7]; and the Fatigue, Resistance, Ambulation, Illness and Loss of weight index [8]; among others. The components of the frailty scale can be clustered. There are 2 areas of investigation that are particularly important. One area used the Frailty Index scale to identify individuals who were multifrail, cognitively and functionally frail, psychologically frail, and physiologically frail. This was accompanied by examining socioeconomic factors, immunoscenescence markers, and inflammatory biomarkers associated with distinct frailty subtypes [9,10].

Another avenue of inquiry centers on the 5 phenotypic criteria proposed by Fried et al [6] (the CHS index), which are intended to distinguish 2 [11] or 3 distinct subgroups of physical frailty [12,13]. The MacArthur Study of Successful Aging captured 2 subdimensions of the CHS phenotype, in which slower gait, weaker grip strength, and lower physical activity define the first component that can better predict cognitive impairment, disability, and mortality, while exhaustion and weight loss define the second component [11]. The I-Lan Longitudinal Aging Study and the National Institute for Longevity Sciences-Longitudinal Study of Aging distinguished among 3 subgroups: nonmobility-type, mobility-type frailty, and low physical activity; they confirmed that the mobility subtype was associated with significantly adverse outcomes [12,13]. Recent research compared the longitudinal trajectories of distinct prefrailty (PF) groups based on the CHS frailty components and reported that the PF2—defined by 1 or 2 among weakness, slowness, and low physical activity, in the absence of exhaustion and unexplained weight loss—had a higher risk of difficulty in carrying out instrumental activities of daily living and mortality than PF1—defined by exhaustion and/or unexplained weight loss, in the absence of weakness, slowness and low physical activity [14,15].

Recent research has examined several physical frailty scales with overlapping components and has distinguished these components into 2 groups: energy-based frailty (EBF) and sarcopenia-based frailty (SBF). This indicates that dietary sodium restriction is associated with a significantly increased risk of SBF, suggesting a potential pathway linking dietary sodium restriction, poor appetite, compromised nutritional status, and SBF in older adults. However, no significant association is observed between dietary sodium restriction and the likelihood of EBF [16].

Inflammation, oxidative stress and antioxidants, coagulation and platelet function, growth factors, musculoskeletal and cardiac function, amino acids and vitamins, hepatic and renal metabolism, DNA, RNA, and miRNA are potential biomarkers or alternatives of frailty [17]. A meta-analysis suggested that C-reactive protein, hemoglobin, albumin, 25-hydroxy vitamin D, and free testosterone are circulating biomarkers strongly associated with clinical frailty. However, none of them alone are sufficient for the diagnosis of frailty. That is consistent with the definition of frailty as a multidimensional syndrome caused by multiple biological alterations [18]. Digital biomarkers, such as wearable sensors recording physical activity features, were limited and may not represent all frailty phenotypes [19]. Indeed, most of these proposed biomarkers necessitate invasive, time-consuming, and relatively high-cost analytical techniques and clinical settings [20].

Voice, which is an emerging indicator of physical and mental health, has been scarcely discussed in the context of frailty. In contrast to other biomarkers, voice analysis presents a rapid, noninvasive, and cost-effective estimation tool, thereby opening up novel avenues for diagnosis, risk prediction, and remote monitoring of patients. Vocal signals feature a multisystemic physical function; they are consistent with the definition of frailty as a multidimensional syndrome and are more advanced biomarkers of frailty. Previous studies have identified acoustic features for detecting cognitive impairment [21-23], screening major depressive disorder [24-26], diagnosing diabetes mellitus [27], monitoring patients with heart failure [28,29] and pulmonary hypertension [30], and identifying patients with respiratory conditions [31]. Two recent studies have demonstrated a correlation between acoustic measures and frailty [32,33]. One study found that the most frail participants exhibited greater speech irregularity but did not have a lower voice intensity than the less frail participants [33]. Another study suggested that spectral-domain voice parameters may be potentially useful in frailty assessment, as the voice characteristics of frail older adults revealed significant variations in the first and second formant frequencies, as well as an energy increase in the low-frequency portion [32]. These results indicate that voice parameters differed according to the frailty status.

The frailty phenotype is rarely discussed in the literature on the aging voice. Nonetheless, vocal biomarkers that elucidate various physical functions and mental status may be potentially applicable in estimating diverse dimensions of frailty. Given the potential relationship and limited evidence, this study aimed to examine the potential phenotypes of frail older adults and determine their relationship with acoustic parameters.

Methods

Study Design and Sample

Participants older than 60 years at the geriatric outpatient clinic of a teaching hospital in central Taiwan between January and December 2020 were recruited. Participants with acute infections and inflammatory diseases (eg, laryngopharyngitis and upper respiratory tract infection), anatomic lesions of the laryngopharynx, gastroesophageal reflux disease, neurologic diseases associated with voice disorders (eg, Parkinson disease

XSL•FO
RenderX

and myasthenia gravis), or a surgical history involving the neck were excluded. Eligible individuals were invited to participate in an observational study focusing on vocal biomarkers of frailty at the geriatric outpatient clinic of a teaching hospital.

Ethical Considerations

The consent form was thoroughly reviewed and signed by all individuals who expressed their willingness to participate in the study, with the assistance of our research team members. Participants had the ability to opt out anytime during the research period. Collected data were deidentified; for example, the participants' names and medical record numbers were replaced by a temporary ID. The study did not provide any specific compensation due to the absence of any invasive intervention. No images or biometric identifiers of individual participants are provided. The study was carried out in accordance with tenets of the Declaration of Helsinki and was approved by the Ethics Committee of China Medical University Hospital, Taiwan (CMUH108-REC3-160).

Measurement of Frailty Phenotypes

Overview

Two frailty phenotypes were identified: EBF and SBF [16]. The study identified 2 additional and potential intermediate forms of the 2 frailty phenotypes: hybrid-based frailty–energy (HBF-E) and hybrid-based frailty–sarcopenia (HBF-S). The measures used to characterize the frailty phenotype are as follows.

EBF: the EBF Index

Weight loss was defined as unintentional weight loss of at least 4.5 kg or >5% of the body weight in the previous year. Fatigue or exhaustion was measured using the question ("In this last week, do you feel that you have less energy to do the things you want?") and categorized as 0 (a "no" answer) or 1 (a "yes" answer). Participants were considered frail if they fulfilled 2 criteria, prefrail if they fulfilled 1 criterion, and robust if no criterion was fulfilled.

SBF: the SBF Index

Low resistance was assessed by measuring the ability to rise from a chair 5 consecutive times without using the arms. Low handgrip strength was assessed by measuring handgrip strength using cutoff values (for the dominant hand) modified for Asian individuals (28 kg for men and 18 kg for women) [34]. Low walking ability was evaluated on the basis of the time spent in walking a 4-m distance, with slow gait defined as a gait speed of <1.0 m/second according to the 2019 Asian Working Group for Sarcopenia [35]. Participants who could not perform the walking test, such as wheelchair users, were classified as having low mobility.

Low physical activity was assessed on the basis of the incidence and progression of basic activities of daily living disability from an emergency geriatric assessment [36], with the following question: "In last year, do you have any deterioration in activities of daily living (feeding, hygiene, dressing, transferring, walking, toileting, and bathing)?" Participants who had difficulty performing at least one of the activities were considered not physically active. Participants were considered frail if at least 3 of the 4 criteria were fulfilled, prefrail if only 1 or 2 criteria were fulfilled, and robust if none of the criteria were fulfilled.

HBF-E

Participants were deemed frail if they were identified as frail based on the EBF index and prefrail or frail based on the SBF index.

HBF-S

Participants were deemed frail if they were identified as frail based on the SBF index and prefrail or frail based on the EBF index.

Acoustic Parameters

After relaxing for 5 minutes, participants sitting in a separate quiet room and maintaining about a 10-cm distance between their mouth and a 90° angle unidirectional stereo condenser microphone (SONY ECM-MS907) were asked to pronounce a sustained vowel "/a/" for approximately 1 second with natural speech. Then, the speech signals were recorded and digitized using a sound blaster (SB1090, Creative Labs) at a sampling rate of 44.1 kHz with an antialiasing function and analyzed using LabVIEW (National Instruments). An end point–detecting algorithm was incorporated to eliminate the leading and trailing nonspeech portions of each utterance [37]. A recording signal that was either shorter than 0.8 seconds or longer than 1.2 seconds was withdrawn. For the calibration and validation of recording equipment, the analyzing program included a standard vowel "/a/" sound, which was used to help participants pronounce a sustained vowel "/a/" correctly. Only parameters with a variability less than 15% could be further analyzed by repeating the test 3 times. Researchers also observed sound waveforms and parameters to identify inappropriate signals and asked participants to rerecord the sounds. The manufacturer conducted periodic calibrations of the device.

The study applied 4 voice parameters—the average number of zero crossings (A1), variations in local peaks and valleys (A2), variations in the first and second formant frequencies (A3), and spectral energy ratio (A4)—to analyze voice changes [37]. A1 was defined as the number of times the signal changed in value, from positive to negative, and vice versa, divided by the frame length. A2 was calculated as the average deviation of the largest (and the smallest) values for all peaks (and valleys), as a reflection of the degree of the temporal stability of vocal variations. A3 was defined as the average deviation from the mean of the first and second formant frequencies, which depend on vocal tract length and the location and narrowness of constrictions along the vocal tract. A4 was defined as the ratio of the spectral energy below 800 Hz (end frequency) to the total spectral energy.

Statistical Analysis

Statistical analyses were performed using the Stata (StataCorp) software. Descriptive statistics were used to describe the study data, including absolute and percentage frequency distributions, mean, and SD. The chi-square test and 1-way ANOVA were respectively used to compare categorical and continuous variables, with $P<.05$ indicating statistical significance. Each acoustic variable was separately evaluated in relation to the

XSL•FO
RenderX

frailty phenotype. Logistic regression was used for elucidating the prediction model. Binomial logistic regression was used in situations where the outcome of a target variable can be limited to only 2 distinct types (eg, EBF versus non-EBF). Odds ratios (ORs) and 95% CIs of the variables included in each model were calculated. The odds to predicted probabilities were also converted using the following formula: probability = odds / (1 + odds). A multinomial logistic regression model was applied to predict the probabilities of the various possible outcomes of a multicategorical dependent variable (eg, nonfrail, EBF, SBF, and both types of frailties). We reported relative risk ratios (RRRs), defined as the ratio of the probability of choosing one outcome category to that of choosing the baseline comparison group.

Results

Characteristics of Nonfrail and Frail Participants

A total of 277 older adults were assessed. Frailty, as defined by the EBF index, was associated with reduced body weight and BMI, malnutrition, and depression, whereas frailty, as defined by the SBF index, was associated with older age, dementia, and fractures. The 4 frailty phenotypes were associated with polypharmacy and falls (Table S1 in Multimedia Appendix 1). When the participants were classified into 4 categories (nonfrail, EBF, SBF, and both types of frailty) and multinomial logistic regression was used, the risk of depression was higher among those with only EBF (RRR 14.92, $P<.001$), while fractures were more likely among those with only SBF (RRR 2.81, $P=.009$; Figure 1). These results suggest differences in the characteristics between nonfrail and frail older people.

Figure 1. The association between frailty phenotypes and the probability of depression and fractures among older adults in 2020.

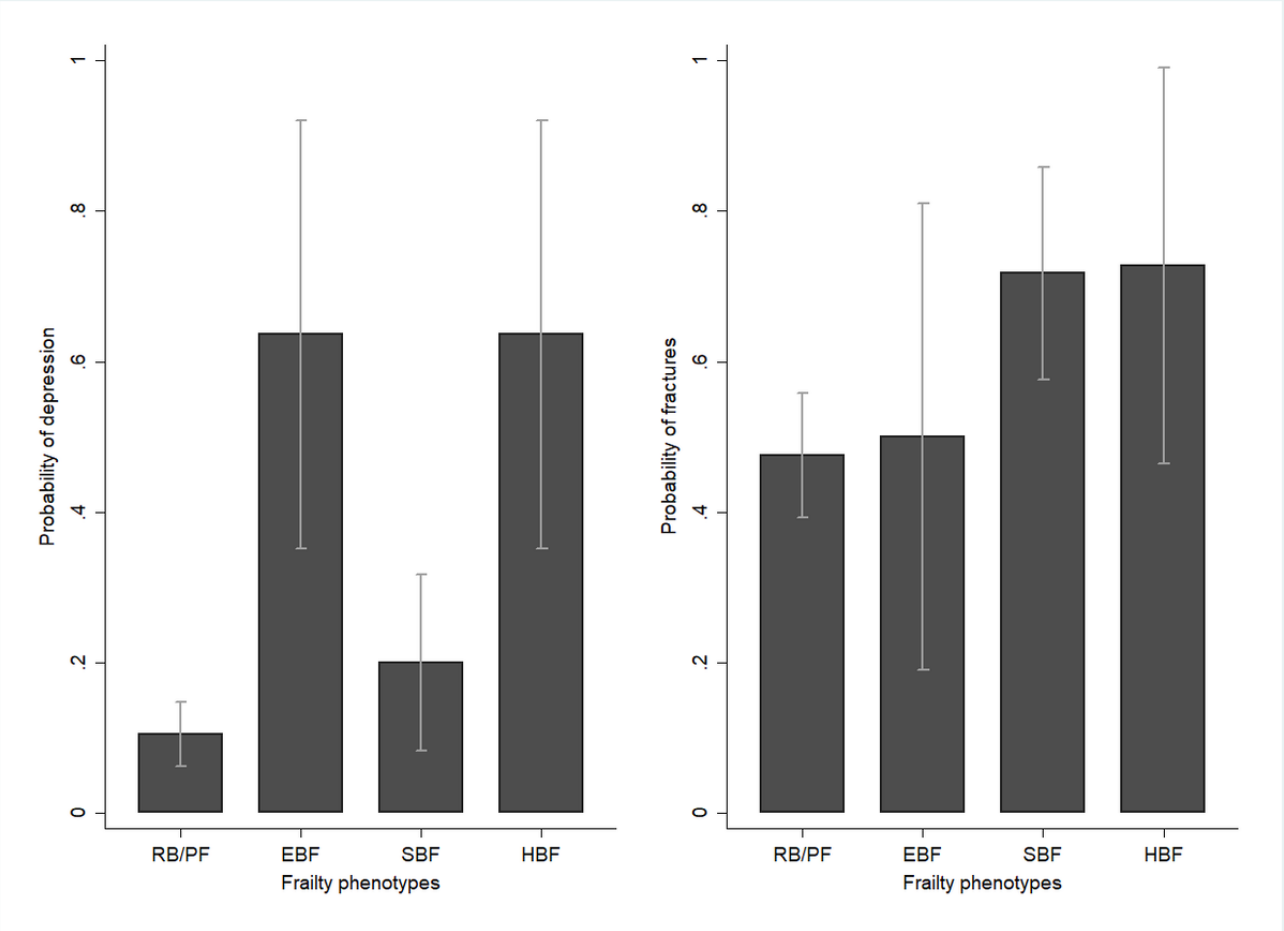

Acoustic Features and Frailty Phenotypes

The acoustic features were related to the probability of frailty as defined by the EBF and SBF indices among older adults (Figure 2). An increase in A1 values was associated with a lower likelihood of EBF (OR 0.81, 95% CI 0.68-0.96), whereas an increase in A2 values resulted in a higher likelihood of SBF (OR 1.34, 95% CI 1.18-1.52). When the participants are classified into 4 categories (nonfrail, EBF, SBF, and both types of frailty) and multinomial logistic regression was used, no significant association was found between A1 and the likelihood of SBF (RRR 0.89, 95% CI 0.66-1.20), and between A2 and the likelihood of EBF (RRR 1.01, 95% CI 0.89-1.15).

XSL•FO
RenderX

Figure 2. The association between acoustic features and the probability of energy-based frailty (EBF) and sarcopenia-based frailty (SBF) among older adults in 2020.

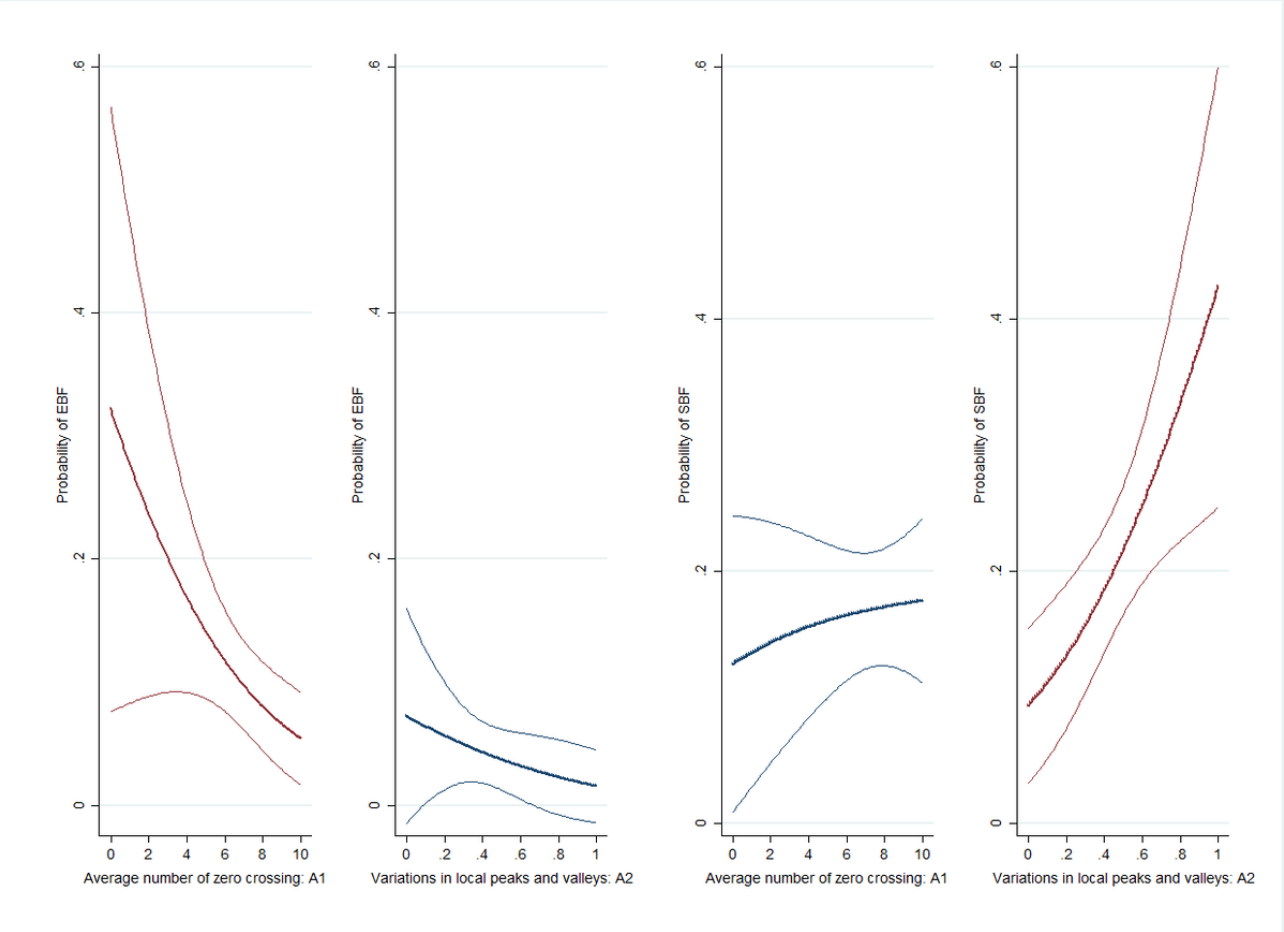

Respondents with higher A3 values had a higher likelihood of HBF-S (OR 1.03, 95% CI 1.002-1.06; Figure 3). When the same criteria for HBF-S were applied, we found that the association between A1 or A2 and the likelihood of HBF-S was not significant (A1: OR 0.97, 95% CI 0.84-1.11; A2: OR 1.09, 95% CI 0.94-1.26). Instead, respondents with larger A4 values had a higher likelihood of HBF-E (OR 1.43, 95% CI 1.02-2.01; Figure 4). Similarly, no significant association was found between A1 or A2 and the likelihood of HBF-E (A1: OR 0.93, 95% CI 0.76-1.13; A2: OR 0.92, 95% CI 0.73-1.15).

XSL•FO
RenderX

Figure 3. The association between acoustic features and the probability of hybrid-based frailty–sarcopenia (HBF-S) among older adults in 2020.

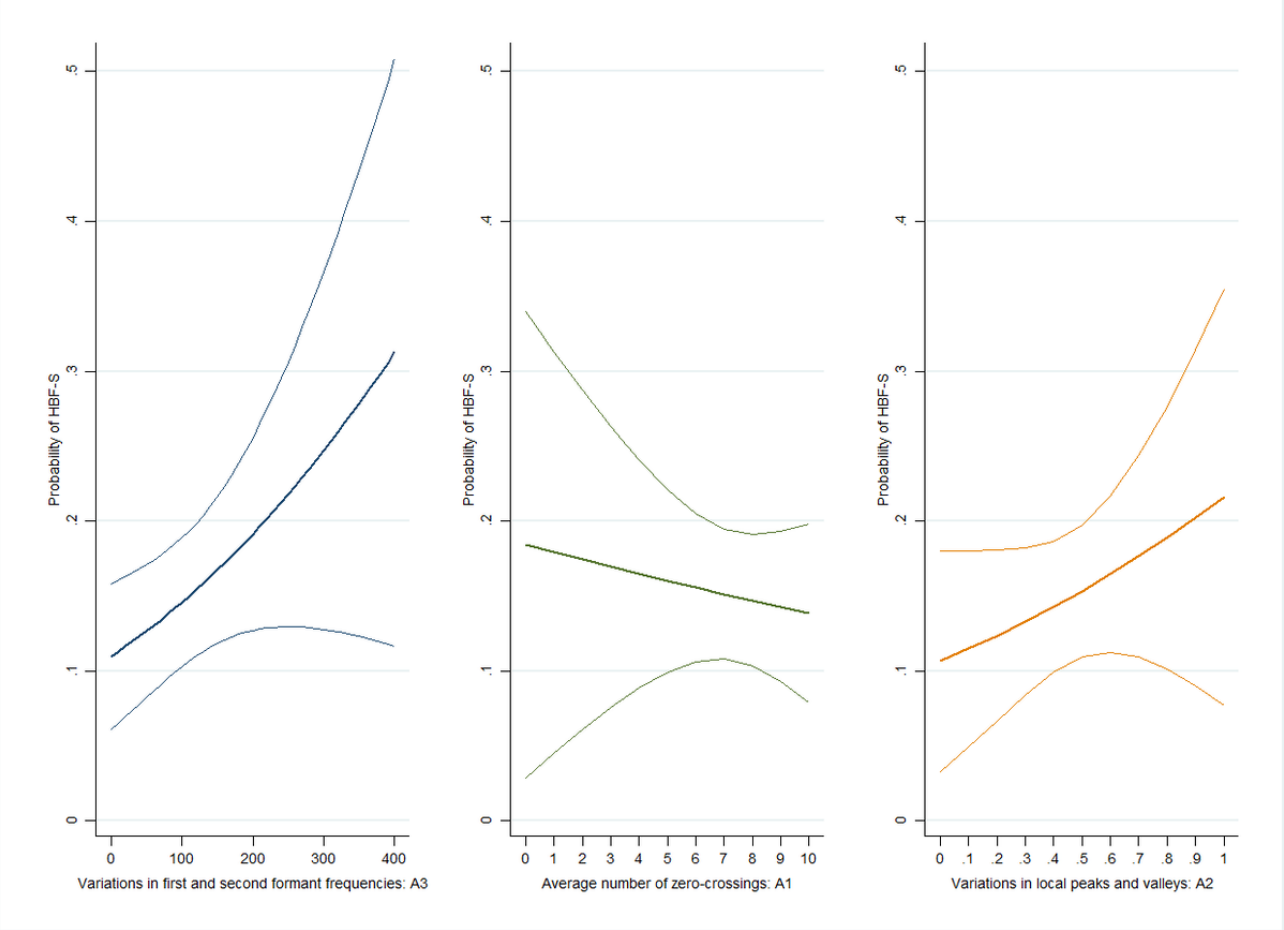

XSL•FO
RenderX

Figure 4. The association between acoustic features and the probability of hybrid-based frailty–energy (HBF-E) among older adults in 2020.

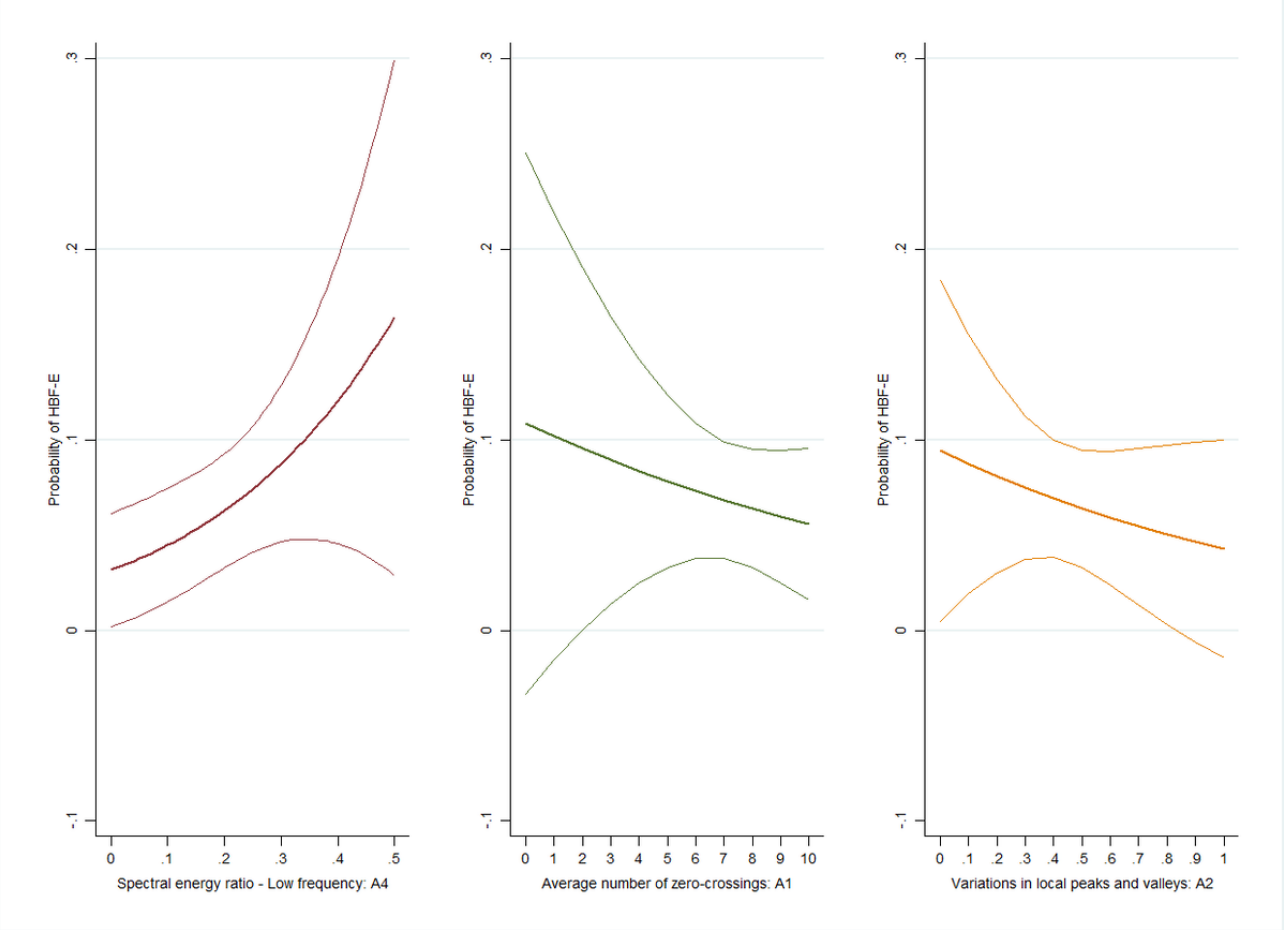

Finally, when the participants were categorized as above, we did not find a significant association between A3 and the likelihood of EBF (RRR 1.02, 95% CI 0.97-1.08) or SBF (RRR 1.00, 95% CI 0.96-1.03). Similar results were observed when A4 was analyzed (EBF: RRR 1.33, 95% CI 0.86-2.07; SBF: RRR 1.05, 95% CI 0.81-1.38).

Discussion

Principal Findings

To the best of our knowledge, this is the first study to ascertain the correlation between frailty phenotypes and acoustic parameters among older adults. Our study identified 4 distinct frailty phenotypes, with A1 being better at identifying older adults with EBF and A2 being much better at identifying older persons with SBF. A3 and A4 can serve as reliable predictors for individuals who report prefrailty or frailty in one dimension and frailty in another dimension.

A1 might reflect the airflow volume initiated by lung contraction during phonation to which the component of exhaustion and weight loss are more related [38]. Prior to speaking, individuals engage in deep breathing to expand their lung volume. Subsequently, a notable increase in esophageal pressure occurs during the expiratory phase, coinciding with the initiation of airflow for speech production [39]. Frailty and respiratory impairment are strongly linked with each other [40]. Wijnant et al [41] evaluated participants with impaired pulmonary function (revealed through spirometry) and reported that they more often developed frailty with poor reversibility. Furthermore, frailty was associated with airflow obstruction and dyspnea in patients with chronic obstructive pulmonary disease [42]. A study on the association between frailty syndrome and acoustic measures of voice quality and voice-related handicap suggested that the components of frailty most related to the Voice Handicap Index-10—which mainly measures voice disorders due to reduced vocal loudness or increased vocal effort from limited airflow [43]—were exhaustion and weight loss rather than slowness, weakness, or low physical activity [44]. In fact, exhausted older adults complain of reduced vocal loudness or increased phonatory effort [45]. These results are consistent with our findings that participants who were more prone to fatigue and body weight loss performed fewer zero crossings on average (A1).

A2 might indicate glottal flow stability during phonation, which is attributed to the control of muscle mass and strength [38]. During the initial phase of speech, the inspiratory external intercostal muscles come into play, effectively opposing the passive recoil of the thorax and lungs. However, as the respiratory cycle progresses, the expiratory internal intercostal muscles take precedence, ensuring a consistent level of air pressure within the trachea [39]. Thus, impaired aerodynamic force control causes greater expansion of the chest and lungs and necessitates more abdominal movement to increase vocal amplitude, which, in turn, results in larger variations in the local peaks and valleys (A2). Indeed, age-related dysphonia is

XSL•FO
RenderX

primarily attributed to the weakness of respiratory, laryngeal, and lingual muscles [45]. Previous research suggested that karaoke training exercises may be a key to increase activity in the respiratory muscles, thus improving pulmonary function and reducing chest stiffness [46]. Additionally, exercises strengthening suprahyoid muscles have demonstrated greater compensatory effects on swallowing and voice production [47]. Poor vocal cord control via the laryngeal muscles also plays a role in aerodynamic instability. The loss of muscle mass and strength could potentially exert a significant effect on the development of presbylarynx, that is, the loss of vocal fold tone and elasticity with aging. In fact, Santos et al [48] reported a correlation between functional impairment and presbylarynx. These results are consistent with our findings that components of SBF, including grip strength, walking speed, physical activity, and chair stands, which reflect muscle mass and strength, are more related to A2.

Frequency-domain parameters (A3 and A4) are better acoustic predictors of a complex frailty phenotype, of which A3 can evaluate frailty that is closer to the sarcopenia type, while A4 can assess frailty that is closer to the energy type. According to the source-filter theory, changes in vocal tract shape, including tongue, laryngeal, and mouth adjustments, are made to produce resonances to specific frequencies [49]. Fine muscular control as well as extensive neurological involvement are required to produce acoustic characteristics of vowels and some consonants during the phonation process [50]. Acoustic parameters—variations in the first and second formant frequencies (A3)—may be determined by both muscular and glottal flow control where the former plays a major role.

A4 may also be related to the interplay of glottal airflow and muscular control of the vocal fold. A plausible major mechanism is that reduced glottal flow causes an increase in the time spent in opening the glottis, which subsequently leads to a more dominant first harmonic in the low-frequency portion of the voice source spectrum, thus increasing the energy in the low-frequency portion of the source spectrum [51]. In breathy (hypofunctional) phonation, relatively low values of the maximum flow declination rate—the major determinant of the peak amplitude and energy portion of the produced voice—from the glottal area waveform were found [52]. Furthermore, the loss of mass and strength in thyroarytenoid muscles that control the vocal cords may result in reduced vertical thickness of the vocal cord and contribute to a longer open phase of the glottis, thereby generating more predominant low-frequency energy (A4) [38]. However, further evidence is required to support these claims.

These findings indicate that older adults with EBF are likely to report depression, whereas those with SBF are likely to have fractures. This finding is in line with previous evidence. Weight loss is associated with an increase in depressive symptoms [53]. Furthermore, older adults with involuntary weight loss are significantly more likely to experience fatigue, as physical fatigue might be partly due to a lack of energy or nutrients [54]. Fatigue is associated with a greater likelihood of deterioration in subsequent self-rated health, functional status, loneliness, depression, and physical activity level [55]. EBF is characterized by involuntary weight loss and fatigue, which suggests a potentially strong association with depressive symptoms. Several studies suggest that low handgrip strength [56], low walking ability [57], and low physical activity [58] are associated with the risk of fractures. This suggests that SBF is potentially strongly associated with fractures, as it exhibits low muscle strength, walking ability, and physical activity.

Limitations

This study has several limitations. First, some acoustic characteristics are not captured within a specific time frame, despite the fact that the sustained vowel "/a/" possesses the advantage of pronunciation without training [37], resulting in a greater overall intensity of voice and elevating the level of reproducible assessment [59], thus rendering it relatively stable for analysis. Future studies can allow participants to pronounce distinct vowels or modify the program to record their voices for different time frames to determine whether this would affect the results. Second, dynamic adjustments to voice frequency and amplitude are required, given that steady vowel utterances bear limited resemblance to natural language production [60]. Nonetheless, biomarkers A1 to A4 proved to be more advantageous in analyzing the changes in the voice than commonly used voice parameters, such as pitch frequency or pitch variations, volume of the voice, and speed of speech, given that the results were unaffected by psychological factors [37]. Third, there is a dearth of data pertaining to the mechanisms that link the frailty phenotypes to the alteration in glottal airflow and anatomical changes that result in alterations in the acoustic properties of voice. More physiological data may help determine causality. Fourth, patient recruitment from a single center may limit the external generalizability of our findings. Multicenter studies should be the focus of future research, allowing for faster recruitment, diverse population coverage, and enhanced generalizability. Fifth, confounding factors, such as recording time (in the morning or afternoon), race, and other physical and mental health conditions, may affect acoustic characteristics, which may bias the outcomes. However, no existing data are available to confirm this. Future research should examine confounding effects in the event that additional data become available. Finally, this study identified distinct frailty phenotypes mainly based on the physical dimension of frailty, thereby failing to use psychological and social components of frailty for developing a holistic typology and corresponding definitions of frailty. Future research that considers such issues would enrich our understanding of the link between diverse frailty phenotypes and voice-related measures.

Implications in Clinical Practice and Diagnostics

Two implications for clinical practice deserve consideration. First, efforts to link acoustic measures to the diagnosis of frailty phenotypes in older adults could commence with 2 crucial acoustic parameters, namely A1 and A2, which roughly indicate insufficient energy or reduced muscle function. Second, the assessment of A3 and A4 involves a complex frailty phenotype that is associated with involuntary loss of body weight, fatigue, and loss of muscle mass and function.

Remote monitoring of physical frailty is crucial for personalized care to decelerate overall deterioration and/or promote the healthy recovery of older adults following exposure to acute or

XSL•FO
RenderX

chronic stressors. For example, the monitoring of frailty status in patients with chronic conditions such as cerebrovascular accident or diabetes in home health care and long-term care institutions can improve the prognosis and treatment of chronic diseases. The application of this program in annual health checkups for older adults, geriatric integrated outpatient clinics, and cardiac and pulmonary rehabilitation workflow in clinical practice may be explored in future research. Furthermore, future research may integrate this program into smartphone apps, vocal chatbots, voice assistants, or other smart devices to record a simple vowel sound for remote monitoring in clinical practice and telemedicine [61].

Conclusions

Given that the frailty phenotype is seldom discussed in the literature on the aging voice, this study, to the best of our knowledge, is the first to determine the relationship between frailty phenotypes and acoustic parameters among older adults. As a noninvasive, instantaneous, objective, and cost-effective estimation tool, the 4 vocal biomarkers have the potential to assess distinct frailty phenotypes.

Acknowledgments

We would like to thank all the participants and patients and their general practitioners for their good collaboration. In particular, we wish to thank Prof Chuang-Chien Chiu for his invaluable guidance on human voice processing. This study was supported by the National Science and Technology Council (NSTC 112-2320-B-039-046). We express our gratitude to support from China Medical University Hospital, Taiwan (grant DMR-HHC-112-3; grant DMR-HHC-113-1); and the Chinese Medicine Research Center, China Medical University from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan (CMRC-CMA-0).

Data Availability

The datasets used and/or analyzed in this study are available from the corresponding author on reasonable request.

Authors' Contributions

Study concept and design: Y-CL, H-TY, and H-HC. Acquisition of data: Y-CL, C-HL, and H-HC. Analysis and interpretation of data: Y-CL and H-TY. Drafting the manuscript: Y-CL and H-TY. Critical revision of the manuscript for important intellectual content: all. Final approval of the version to be published: all. Co-corresponding authors: C-HL and H-HC.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of robust/prefrail and frail participants based on the energy-based frailty index, sarcopenia-based frailty index, hybrid-based frailty–energy index, and hybrid-based frailty–sarcopenia index.
[DOCX File , 22 KB - jmir_v26i1e58466_app1.docx]

References

1. Chen X, Mao G, Leng SX. Frailty syndrome: an overview. Clin Interv Aging 2014 Mar 19:433-441. [doi: 10.2147/cia.s45300]
2. Midão L, Brochado P, Almada M, Duarte M, Paúl C, Costa E. Frailty status and polypharmacy predict all-cause mortality in community dwelling older adults in Europe. Int J Environ Res Public Health 2021 Mar 30;18(7):3580 [FREE Full text] [doi: 10.3390/ijerph18073580] [Medline: 33808273]
3. Veronese N, Custodero C, Cella A, Demurtas J, Zora S, Maggi S, et al. Prevalence of multidimensional frailty and pre-frailty in older people in different settings: a systematic review and meta-analysis. Ageing Res Rev 2021 Dec;72:101498. [doi: 10.1016/j.arr.2021.101498] [Medline: 34700009]
4. Searle SD, Mitnitski A, Gahbauer EA, Gill TM, Rockwood K. A standard procedure for creating a frailty index. BMC Geriatr 2008 Sep 30;8:24 [FREE Full text] [doi: 10.1186/1471-2318-8-24] [Medline: 18826625]
5. Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005 Aug 30;173(5):489-495 [FREE Full text] [doi: 10.1503/cmaj.050051] [Medline: 16129869]
6. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar 01;56(3):M146-M157. [doi: 10.1093/gerona/56.3.M146]
7. Ensrud KE, Ewing SK, Cawthon PM, Fink HA, Taylor BC, Cauley JA, Osteoporotic Fractures in Men Research Group. A comparison of frailty indexes for the prediction of falls, disability, fractures, and mortality in older men. J Am Geriatr Soc 2009 Mar 03;57(3):492-498 [FREE Full text] [doi: 10.1111/j.1532-5415.2009.02137.x] [Medline: 19245414]

XSL•FO
RenderX

8. Morley JE, Malmstrom TK, Miller DK. A simple frailty questionnaire (FRAIL) predicts outcomes in middle aged African Americans. J Nutr Health Aging 2012 Jul;16(7):601-608 [FREE Full text] [doi: 10.1007/s12603-012-0084-2] [Medline: 22836700]
9. Yin MJ, Xiong YZ, Xu XJ, Huang LF, Zhang Y, Wang XJ, et al. Tfh cell subset biomarkers and inflammatory markers are associated with frailty status and frailty subtypes in the community-dwelling older population: a cross-sectional study. Aging (Albany NY) 2020 Feb 08;12(3):2952-2973 [FREE Full text] [doi: 10.18632/aging.102789] [Medline: 32039831]
10. Zhang Y, Xu X, Lian T, Huang L, Zeng J, Liang D, et al. Development of frailty subtypes and their associated risk factors among the community-dwelling elderly population. Aging (Albany NY) 2020 Jan 16;12(2):1128-1140 [FREE Full text] [doi: 10.18632/aging.102671] [Medline: 31951595]
11. Sarkisian CA, Gruenewald TL, John Boscardin W, Seeman TE. Preliminary evidence for subdimensions of geriatric frailty: the MacArthur study of successful aging. J Am Geriatr Soc 2008 Dec 02;56(12):2292-2297 [FREE Full text] [doi: 10.1111/j.1532-5415.2008.02041.x] [Medline: 19016933]
12. Huang ST, Tange C, Otsuka R, Nishita Y, Peng L, Hsiao F, et al. Subtypes of physical frailty and their long-term outcomes: a longitudinal cohort study. J Cachexia Sarcopenia Muscle 2020 Oct;11(5):1223-1231 [FREE Full text] [doi: 10.1002/jcsm.12577] [Medline: 32558267]
13. Liu LK, Guo CY, Lee WJ, Chen LJ, Hwang AC, Lin MH, et al. Subtypes of physical frailty: Latent class analysis and associations with clinical characteristics and outcomes. Sci Rep 2017 Apr 11;7(1):46417 [FREE Full text] [doi: 10.1038/srep46417] [Medline: 28397814]
14. Lee YS, Chen YM, Chou YY, Lin CS, Chen JP, Kuo FH, et al. Prefrailty subtypes differentially predict 5-year mortality in the functionally independent geriatric population. Arch Gerontol Geriatr 2023 Mar;106:104897. [doi: 10.1016/j.archger.2022.104897] [Medline: 36508848]
15. Romero-Ortuno R, Scarlett S, O'Halloran A, Kenny RA. Is phenotypical prefrailty all the same? A longitudinal investigation of two prefrailty subtypes in TILDA. Age Ageing 2019 Nov 11;49(1):39-45. [doi: 10.1093/ageing/afz129] [Medline: 31711148]
16. Lin YC, Yan HT. Dietary sodium restriction and frailty among middle-aged and older adults: an 8-year longitudinal study. Nutrients 2024 Feb 20;16(5):580 [FREE Full text] [doi: 10.3390/nu16050580] [Medline: 38474709]
17. Al Saedi A, Feehan J, Phu S, Duque G. Current and emerging biomarkers of frailty in the elderly. Clin Interv Aging 2019 Feb 19;14:389-398. [doi: 10.2147/cia.s168687]
18. Mailliez A, Guilbaud A, Puisieux F, Dauchet L, Boulanger É. Circulating biomarkers characterizing physical frailty: CRP, hemoglobin, albumin, 25OHD and free testosterone as best biomarkers. Results of a meta-analysis. Exp Gerontol 2020 Oct 01;139:111014. [doi: 10.1016/j.exger.2020.111014] [Medline: 32599147]
19. Park C, Mishra R, Golledge J, Najafi B. Digital biomarkers of physical frailty and frailty phenotypes using sensor-based physical activity and machine learning. Sensors (Basel) 2021 Aug 05;21(16):5289 [FREE Full text] [doi: 10.3390/s21165289] [Medline: 34450734]
20. Sepúlveda M, Arauna D, García F, Albala C, Palomo I, Fuentes E. Frailty in aging and the search for the optimal biomarker: a review. Biomedicines 2022 Jun 16;10(6):1426 [FREE Full text] [doi: 10.3390/biomedicines10061426] [Medline: 35740447]
21. Lin H, Karjadi C, Ang TFA, Prajakta J, McManus C, Alhanai TW, et al. Identification of digital voice biomarkers for cognitive health. Explor Med 2020 Dec 31;1:406-417 [FREE Full text] [doi: 10.37349/emed.2020.00028] [Medline: 33665648]
22. Mahon E, Lachman M. Voice biomarkers as possible indicators of cognitive aging. Innov Aging. 2021 2021 Dec 17;5:705. [doi: 10.1093/geroni/igab046.2639]
23. Mahon E, Lachman ME. Voice biomarkers as indicators of cognitive changes in middle and later adulthood. Neurobiol Aging 2022 Nov;119:22-35 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2022.06.010] [Medline: 35964541]
24. Lee S, Suh SW, Kim T, Kim K, Lee KH, Lee JR, et al. Screening major depressive disorder using vocal acoustic features in the elderly by sex. J Affect Disord 2021 Aug 01;291:15-23. [doi: 10.1016/j.jad.2021.04.098] [Medline: 34022551]
25. Taguchi T, Tachikawa H, Nemoto K, Suzuki M, Nagano T, Tachibana R, et al. Major depressive disorder discrimination using vocal acoustic features. J Affect Disord 2018 Jan 01;225:214-220. [doi: 10.1016/j.jad.2017.08.038] [Medline: 28841483]
26. Wang J, Zhang L, Liu T, Pan W, Hu B, Zhu T. Acoustic differences between healthy and depressed people: a cross-situation study. BMC Psychiatry 2019 Oct 15;19(1):300 [FREE Full text] [doi: 10.1186/s12888-019-2300-7] [Medline: 31615470]
27. Sidorova J, Carbonell P, Čukić M. Blood glucose estimation from voice: first review of successes and challenges. J Voice 2022 Sep;36(5):737.e1-737.e10 [FREE Full text] [doi: 10.1016/j.jvoice.2020.08.034] [Medline: 33041176]
28. Maor E, Perry D, Mevorach D, Taiblum N, Luz Y, Mazin I, et al. Vocal biomarker is associated with hospitalization and mortality among heart failure patients. JAHA 2020 Apr 01;9(7). [doi: 10.1161/jaha.119.013359]
29. Schöbi D, Zhang YP, Kehl J, Aissani M, Pfister O, Strahm M, et al. Evaluation of speech and pause alterations in patients with acute and chronic heart failure. JAHA 2022 Oct 31;11(21). [doi: 10.1161/jaha.122.027023]
30. Sara JDS, Maor E, Borlaug B, Lewis BR, Orbelo D, Lerman LO, et al. Non-invasive vocal biomarker is associated with pulmonary hypertension. PLoS One 2020 Apr 16;15(4):e0231441 [FREE Full text] [doi: 10.1371/journal.pone.0231441] [Medline: 32298301]

XSL•FO
RenderX

31. Kaur S, Larsen E, Harper J, Purandare B, Uluer A, Hasdianda MA, et al. Development and validation of a respiratory-responsive vocal biomarker-based tool for generalizable detection of respiratory impairment: independent case-control studies in multiple respiratory conditions including asthma, chronic obstructive pulmonary disease, and COVID-19. J Med Internet Res 2023 Apr 14;25:e44410 [FREE Full text] [doi: 10.2196/44410] [Medline: 36881540]
32. Lin YC, Yan HT, Lin CH, Chang HH. Predicting frailty in older adults using vocal biomarkers: a cross-sectional study. BMC Geriatr 2022 Jul 01;22(1):549 [FREE Full text] [doi: 10.1186/s12877-022-03237-7] [Medline: 35778699]
33. Rosen-Lang Y, Zoubi S, Cialic R, Orenstein T. Using voice biomarkers for frailty classification. Geroscience 2024 Feb;46(1):1175-1179 [FREE Full text] [doi: 10.1007/s11357-023-00872-9] [Medline: 37480417]
34. Auyeung TW, Arai H, Chen LK, Woo J. Letter to the editor: Normative data of handgrip strength in 26344 older adults - a pooled dataset from eight cohorts in Asia. J Nutr Health Aging 2020 Jan;24(1):125-126 [FREE Full text] [doi: 10.1007/s12603-019-1287-6] [Medline: 31886819]
35. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020 Mar;21(3):300-307.e2. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
36. Ke YT, Peng AC, Shu YM, Chung MH, Tsai KT, Chen PJ, et al. Emergency geriatric assessment: a novel comprehensive screen tool for geriatric patients in the emergency department.. Am J Emerg Med 2018 Jan;36(1):143-146 [FREE Full text] [doi: 10.1016/j.ajem.2017.07.008] [Medline: 28712643]
37. Chiu CC, Chang HH, Yang CH. Objective auscultation for traditional chinese medical diagnosis using novel acoustic parameters. Comput Methods Programs Biomed 2000 Jun;62(2):99-107. [doi: 10.1016/s0169-2607(00)00055-9] [Medline: 10764936]
38. Zhang Z. Mechanics of human voice production and control. J Acoust Soc Am 2016 Oct 14;140(4):2614-2635 [FREE Full text] [doi: 10.1121/1.4964509] [Medline: 27794319]
39. Draper MH, Ladefoged P, Whitteridge D. Expiratory pressures and air flow during speech. BMJ 1960 Jun 18;1(5189):1837-1843. [doi: 10.1136/bmj.1.5189.1837]
40. Vaz Fragoso CA, Enright PL, McAvay G, Van Ness PH, Gill TM. Frailty and respiratory impairment in older persons. Am J Med 2012 Jan;125(1):79-86 [FREE Full text] [doi: 10.1016/j.amjmed.2011.06.024] [Medline: 22195532]
41. Wijnant SRA, Benz E, Luik AI, Rivadeneira F, Voortman T, Brusselle GG, et al. Frailty transitions in older persons with lung function impairment: a population-based study. J Gerontol A Biol Sci Med Sci 2023 Feb 24;78(2):349-356 [FREE Full text] [doi: 10.1093/gerona/glac202] [Medline: 36226677]
42. Hanlon P, Guo X, McGhee E, Lewsey J, McAllister D, Mair FS. Systematic review and meta-analysis of prevalence, trajectories, and clinical outcomes for frailty in COPD. NPJ Prim Care Respir Med 2023 Jan 05;33(1):1 [FREE Full text] [doi: 10.1038/s41533-022-00324-5] [Medline: 36604427]
43. Samlan RA, Black MA, Abidov M, Mohler J, Fain M. Frailty syndrome, cognition, and dysphonia in the elderly. J Voice 2020 Jan;34(1):160.e15-160.e23. [doi: 10.1016/j.jvoice.2018.06.001] [Medline: 30055984]
44. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS, et al. The Voice Handicap Index (VHI). Am J Speech Lang Pathol 1997 Aug;6(3):66-70. [doi: 10.1044/1058-0360.0603.66]
45. Ziegler A, Verdolini Abbott K, Johns M, Klein A, Hapner ER. Preliminary data on two voice therapy interventions in the treatment of presbyphonia. Laryngoscope 2014 Aug;124(8):1869-1876. [doi: 10.1002/lary.24548] [Medline: 24375313]
46. Miyazaki A, Mori H. Frequent karaoke training improves frontal executive cognitive skills, tongue pressure, and respiratory function in elderly people: pilot study from a randomized controlled trial. Int J Environ Res Public Health 2020 Feb 24;17(4):1459 [FREE Full text] [doi: 10.3390/ijerph17041459] [Medline: 32102472]
47. Easterling C. Does an exercise aimed at improving swallow function have an effect on vocal function in the healthy elderly? Dysphagia 2008 Sep;23(3):317-326. [doi: 10.1007/s00455-008-9158-z] [Medline: 18437461]
48. Santos M, Freitas SV, Dias D, Costa J, Coutinho M, Sousa CA, et al. Presbylarynx: does body muscle mass correlate with vocal atrophy? A prospective case control study. Laryngoscope 2021 Jan;131(1):E226-E230. [doi: 10.1002/lary.28685] [Medline: 32401382]
49. Lindblom BE, Sundberg JE. Acoustical consequences of lip, tongue, jaw, and larynx movement. J Acoust Soc Am 1971 Oct;50(4):1166-1179. [doi: 10.1121/1.1912750] [Medline: 5117649]
50. Wolfe J, Garnier M, Smith J. Vocal tract resonances in speech, singing, and playing musical instruments. HFSP J 2009;3(1):6-23 [FREE Full text] [doi: 10.2976/1.2998482] [Medline: 19649157]
51. Baier H, Wanner A, Zarzecki S, Sackner MA. Relationships among glottis opening, respiratory flow, and upper airway resistance in humans. J Appl Physiol Respir Environ Exerc Physiol 1977 Oct 01;43(4):603-611. [doi: 10.1152/jappl.1977.43.4.603] [Medline: 908674]
52. Patel RR, Sundberg J, Gill B, Lã FMB. Glottal airflow and glottal area waveform characteristics of flow phonation in untrained vocally healthy adults. J Voice 2022 Jan;36(1):140.e1-140.e21. [doi: 10.1016/j.jvoice.2020.07.037] [Medline: 32868146]
53. Koster A, van Gool CH, Kempen GI, Penninx BW, Lee JS, Rubin SM, et al. Late-life depressed mood and weight change contribute to the risk of each other. Am J Geriatr Psychiatry 2010 Mar;18(3):236-244. [doi: 10.1097/jgp.0b013e3181c65837]

XSL•FO
RenderX

54. Franz K, Otten L, Müller-Werdan U, Doehner W, Norman K. Severe weight loss and its association with fatigue in old patients at discharge from a geriatric hospital. Nutrients 2019 Oct 10;11(10):2415 [FREE Full text] [doi: 10.3390/nu11102415] [Medline: 31658676]
55. Moreh E, Jacobs JM, Stessman J. Fatigue, function, and mortality in older adults. J Gerontol A Biol Sci Med Sci 2010 Aug;65(8):887-895. [doi: 10.1093/gerona/glq064] [Medline: 20418349]
56. Kunutsor SK, Seidu S, Voutilainen A, Blom AW, Laukkanen JA. Handgrip strength-a risk indicator for future fractures in the general population: findings from a prospective study and meta-analysis of 19 prospective cohort studies. Geroscience 2021 Apr;43(2):869-880 [FREE Full text] [doi: 10.1007/s11357-020-00251-8] [Medline: 32812100]
57. Luukinen H, Koski K, Laippala P, Kivelä SL. Factors predicting fractures during falling impacts among home-dwelling older adults. J Am Geriatr Soc 1997 Nov;45(11):1302-1309. [doi: 10.1111/j.1532-5415.1997.tb02928.x] [Medline: 9361654]
58. Stel VS, Pluijm SM, Deeg DJ, Smit JH, Bouter LM, Lips P. Functional limitations and poor physical performance as independent risk factors for self-reported fractures in older persons. Osteoporos Int 2004 Sep;15(9):742-750 [FREE Full text] [doi: 10.1007/s00198-004-1604-7] [Medline: 15014931]
59. Lu FL, Matteson S. Speech tasks and interrater reliability in perceptual voice evaluation. J Voice 2014 Nov;28(6):725-732. [doi: 10.1016/j.jvoice.2014.01.018] [Medline: 24841668]
60. Lortie CL, Thibeault M, Guitton MJ, Tremblay P. Effects of age on the amplitude, frequency and perceived quality of voice. Age (Dordr) 2015 Dec;37(6):117 [FREE Full text] [doi: 10.1007/s11357-015-9854-1] [Medline: 26578457]
61. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. Digit Biomark 2021 Apr 16;5(1):78-88 [FREE Full text] [doi: 10.1159/000515346] [Medline: 34056518]

Abbreviations

A1: average number of zero crossings
A2: variations in local peaks and valleys
A3: variations in first and second formant frequencies
A4: spectral energy ratio
CHS: Cardiovascular Health Study
EBF: energy-based frailty
HBF-E: hybrid-based frailty–energy
HBF-S: hybrid-based frailty–sarcopenia
OR: odds ratio
PF: prefrailty
RRR: relative risk ratio
SBF: sarcopenia-based frailty

Edited by A Schwartz; submitted 16.03.24; peer-reviewed by M Colombo, X Qiu; comments to author 22.05.24; revised version received 04.06.24; accepted 07.10.24; published 08.11.24.

Please cite as:
Lin YC, Yan HT, Lin CH, Chang HH
Identifying and Estimating Frailty Phenotypes by Vocal Biomarkers: Cross-Sectional Study
J Med Internet Res 2024;26:e58466
URL: https://www.jmir.org/2024/1/e58466
doi:10.2196/58466
PMID:

©Yu-Chun Lin, Huang-Ting Yan, Chih-Hsueh Lin, Hen-Hong Chang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Image Content on Medical Crowdfunding Success: A Machine Learning Approach

Renwu Wang[1], PhD; Huimin Xu[1], BM; Xupin Zhang[1], PhD

School of Economics and Management, East China Normal University, Shanghai, China

Corresponding Author:
Xupin Zhang, PhD
School of Economics and Management
East China Normal University
3663 North Zhongshan Road
Shanghai, 200062
China
Phone: 86 15026588918
Email: xpzhang@fem.ecnu.edu.cn

Abstract

Background: As crowdfunding sites proliferate, visual content often serves as the initial bridge connecting a project to its potential backers, underscoring the importance of image selection in effectively engaging an audience.

Objective: This paper aims to explore the relationship between images and crowdfunding success in cancer-related crowdfunding projects.

Methods: We used the Alibaba Cloud platform to detect individual features in images. In addition, we used the Recognize Anything Model to label images and obtain content tags. Furthermore, the discourse atomic topic model was used to generate image topics. After obtaining the image features and image content topics, we built regression models to investigate the factors that influence the results of crowdfunding success.

Results: Images with a higher proportion of young people (β=0.0753; *P*<.001), a larger number of people (β=0.00822; *P*<.001), and a larger proportion of smiling faces (β=0.0446; *P*<.001) had a higher success rate. Image content related to good things and patient health also contributed to crowdfunding success (β=0.082, *P*<.001; and β=0.036, *P*<.001, respectively). In addition, the interaction between image topics and image characteristics had a significant effect on the final fundraising outcome. For example, when smiling faces are considered in conjunction with the image topics, using more smiling faces in the *rest and play* theme increased the amount of money raised (β=0.0152; *P*<.001). We also examined causality through a counterfactual analysis, which confirmed the influence of the variables on crowdfunding success, consistent with the results of our regression models.

Conclusions: In the realm of web-based medical crowdfunding, the importance of uploaded images cannot be overstated. Image characteristics, including the number of people depicted and the presence of youth, significantly improve fundraising results. In addition, the thematic choice of images in cancer crowdfunding efforts has a profound impact. Images that evoke beauty and resonate with health issues are more likely to result in increased donations. However, it is critical to recognize that reinforcing character traits in images of different themes has different effects on the success of crowdfunding campaigns.

(J Med Internet Res 2024;26:e58617) doi:10.2196/58617

KEYWORDS

medical crowdfunding; visual analytics; machine learning; image content; crowdfunding success

Introduction

Background

Charitable crowdfunding has revolutionized philanthropy and social support, providing a unique and accessible platform for individuals and organizations to raise funds for their causes. In the realm of medical crowdfunding, patients often turn to web-based platforms such as GoFundMe for financial support.

A critical element of a successful crowdfunding campaign is the selection of an effective campaign cover image. This image serves as a visual representation of the project and plays a key role in attracting potential donors [1-3]. It also helps to communicate the project's core information and value

XSL•FO
RenderX

proposition, allowing potential supporters to quickly understand the project's goals and importance of the project [4,5].

In the current digital age, the campaign cover image serves as a first impression. Given people's limited attention span, it has a significant impact on the outcome of the crowdfunding campaign [6,7]. Research has shown that individuals process and retain visual information better than textual data [8-12]. The importance of visual information has been demonstrated in advertising, marketing, technology crowdfunding, and other fields [13-16]. Researchers have investigated several factors that influence the effectiveness of crowdfunding campaigns, such as project descriptions, incentives, and engagement on social media platforms [17-21]. While there is growing public awareness of using web-based crowdfunding platforms to raise medical funds for patients with cancer, there is a lack of research in academia on how specific elements of a cancer crowdfunding project's image affect its success rates.

In order to obtain a more profound understanding of the influence of images on the success of cancer crowdfunding campaigns, we performed a study of cancer-related crowdfunding campaigns on GoFundMe, which is the most extensive site for medical crowdfunding. Our hypotheses are shown in Textbox 1.

Textbox 1. Our hypotheses.

- Individual characteristics of images used for medical crowdfunding projects significantly influence the projects' fundraising outcomes.
- The thematic content of images is a significant determinant of the success of crowdfunding campaigns.
- There is a significant interaction effect between the subject matter of images and individual characteristics of crowdfunding images, which in turn affects the funding results of medical crowdfunding projects.

Related Work

Individual Characteristics of Crowdfunding Images

Scholars have confirmed the importance of character appearance and character attributes in images, including the number of faces, facial expressions, race, and gender. Some researchers have shown that the presence of faces greatly facilitates funders' decision-making [22-24], but there are also areas of research that demonstrate that the disclosure of a project initiator's real face negatively impacts crowdfunding success [25]. In addition, studies suggest that both positive expressions, including smiling, as well as negative expressions, such as sadness, can have an influence [26-28].

Compared to other types of crowdfunding projects, medical crowdfunding for cancer projects has its own specific focus and target audience. Therefore, factors that have been shown to positively or negatively influence people's success in crowdfunding in other fields [29,30] may not be applicable to cancer projects in medical crowdfunding.

The Topic of Image Content

While empirical studies of crowdfunding data have increased, there is still a lack of attention to image content, highlighting the need for broader adoption of visual data analysis in research. Some academic discourse has focused on the photographic aspects of images, such as color dynamics, composition, and figure-ground relationships [31,32]. However, for crowdfunding initiators, launching a crowdfunding project and uploading images that take into account photographic features require a certain technical threshold. The content of the image is more easily understood and judged by the general public, which can be an important factor in determining whether it is suitable for uploading. Platforms such as Indiegogo emphasize the importance of an image's style and content as critical factors for achieving crowdfunding success. Therefore, when evaluating the impact of visual elements, it is imperative to place significant emphasis on the role of image content in crowdfunding results.

Previous research has analyzed imagery used by nonprofit organizations in fundraising appeals and raised concerns about the possibility of perpetuating harmful stereotypes. The depiction of individuals as passive, impoverished, and victimized, waiting for rescue by the Western spectatorship, has been examined for its potential to reinforce racial and economic clichés [33,34]. Careful selection of image content is crucial, as potential donors rely solely on the information presented within a crowdfunding campaign due to their limited firsthand experience. The content of images plays a pivotal role in shaping their initial perceptions of individuals in need of aid.

Research has shown that image content significantly influences crowdfunding campaign donations from all perspectives. For example, images with self-help behavior content attract more donations [35]; the presence of a direct gaze action in campaign images leads to higher fundraising results [36]; and images depicting a patient's current serious medical condition are more effective than images depicting a relatively normal image of that condition [37].

Images with varying content have distinct persuasive effects when used in the domain of health care crowdfunding [24]. Emotional and credibility-based images display a positive persuasive effect, while logos-related images depict a negative persuasive effect.

The Intersection of the Factors in Crowdfunding

The convergence of factors in crowdfunding projects is influenced by a variety of elements, and the effects of the interplay between these factors on the final crowdfunding outcome are underresearched.

Previous studies have attempted to investigate the relationships between visual and textual factors. Zhao et al [38] addressed potential interactions between visual emotions and textual emotions that had previously been ignored. Their study showed that emotional congruence between images and linguistic approach can mitigate the outcomes of unsuccessful web-based charity crowdfunding campaigns. Yang et al [16] investigated

XSL•FO
RenderX

how the number of images and videos affects the relationship between text length and crowdfunding outcomes. They observed a positive correlation between text length and fundraising outcomes; however, this relationship weakened as the number of videos and images increased. Buskila and Perez [35] attempted to analyze the various features of images, focusing on whether or not the characters in the images were rescued, as well as the emotions conveyed by the images. The study concluded that victims received fewer donations due to 2 specific factors, including the use of images depicting negative emotions and the absence of self-help behavioral content.

However, due to limited research, scholars are unable to gain a deeper understanding of the interactions between the elements in the image. Based on the above research, this paper conducts further research on the interactions between various elements in crowdfunding programs.

Methods

Overview

As shown in Figure 1, this section describes the flow of the research throughout the article in 5 parts.

Figure 1. Technological route throughout the paper. AliCloud: Alibaba Cloud; DATM: discourse atomic topic modeling; RAM: Recognize Anything Model.

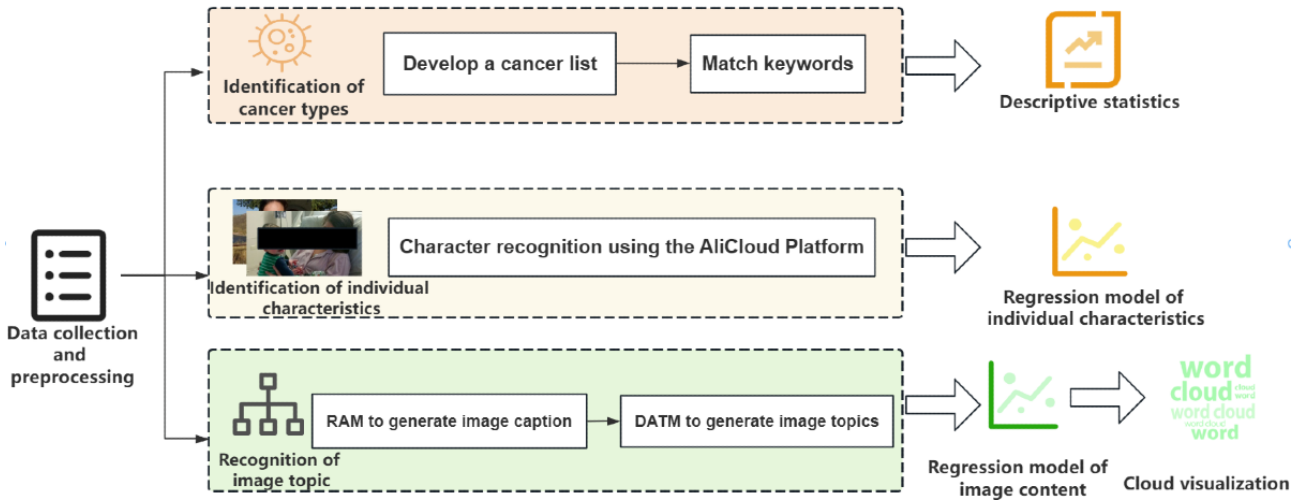

Data Collection

We used a Python-based web-scraping program to collect data on all cancer-related crowdfunding projects hosted on the platform from January 1, 2019, to July 28, 2020. Ultimately, our efforts yielded data from a total of 307,886 projects available on the site. The dataset includes essential project details such as basic information, textual content, and images. In addition, the basic information includes several key variables that are critical for later analysis, including funding amount, number of donations, description length, title length, and duration. These elements are referred to as basic features.

Given that the United States is a major player in the global crowdfunding market, with a well-established environment and individuals often demonstrating a higher level of engagement and trust in crowdfunding efforts, we focused our initial screening on US-based projects with individual crowdfunding initiators. This refinement narrowed our dataset to 243,795 medical campaigns. We then transformed the dataset to facilitate analysis. This involved converting the fundraising amounts from textual to numerical values and calculating the time elapsed since the creation of each campaign using a predetermined base date.

Procedure

Identification of Cancer Types

Different types of cancer have significant variations in disease etiology, treatment methods, and costs [39]. These differences can have a significant impact on the goals and outcomes of crowdfunding initiatives [40]. Therefore, it is essential to identify the specific cancer type associated with each crowdfunding project and analyze its characteristics using different cancer classifications. Crowdfunding initiators usually provide detailed information in the project description, including the relevant cancer type details. Therefore, identifying the cancer type through the description is crucial. We compiled a list of 238 cancer categories and their corresponding keywords to perform a keyword matching analysis of crowdfunding project descriptions. In the end, we identified 97 different cancer types after analyzing a total of 243,795 crowdfunding projects (Table 1).

XSL•FO
RenderX

Table 1. Top 10 cancers in terms of the number of projects.

Cancer type	Projects, n
Skin cancer	40,038
Childhood cancer	16,714
Brain tumors	7251
Lung cancer	6495
Breast cancer	6338
Kidney cancer	6185
Familial gastrointestinal stromal tumors	5482
Liver cancer	5237
Gastric cancer	4990
Eye cancer	4978

Identification of Image Features

In crowdfunding cover images, individuals can provide supplementary information about patients and other project aspects, allowing potential donors to better comprehend and empathize with the project, ultimately increasing support likelihood. This study used the face attribute recognition interface provided by Alibaba Cloud's Vision Intelligent Application Programming Interface Platform [41] to detect personal features in crowdfunding project images. The face detection and recognition algorithms of this platform can realize high-performance attribute recognition.

We conducted analysis of the uploaded crowdfunding project images by designing a Python program to interface with the application programming interface. The results returned by the interface included *Face Count*, *Male Count*, *Female Count*, *Age*, *Smile Count*, *Glasses Count*, and *Hat Count* for each image, as shown in the Table 2.

Table 2. Examples of image features in crowdfunding images[a].

Picture	Face Count, n	Male Count, n	Female Count, n	Smile Count, n	Glasses Count, n	Hat Count, n	Age (years)
	2	0	2	2	0	0	36-60 and 60
	2	1	1	2	0	1	18-35

[a]The people in the pictures are mosaicked to protect privacy.

Image Caption on the Recognize Anything Model

In addition to exploring the impact of image features in crowdfunding images, the cancer crowdfunding project's image content can provide additional insights into the project. To characterize the content of these images, we used the Recognize Anything Model (RAM) [42].

The RAM integrates the Segment Anything scene segmentation model from Meta [43] with the recognition capabilities of Zero Shot. This combination allows the RAM to localize and identify object categories within images. It recognizes over 6400 common labels, encompassing a broader range of pertinent categories compared to Open Images V6 [44]. More details about the RAM can be found in Multimedia Appendix 1.

Through configuring local parameters, we standardized and resized images. We then used the *ram-swin-large_14m* pretrained model to process images from cancer crowdfunding projects, generating results using the RAM.

Discourse Atomic Topic Modeling

However, the complexity of image descriptions makes them unsuitable for subsequent impact analysis on crowdfunding results. Therefore, we performed topic recognition on image descriptions to uncover commonalities hidden within the descriptive content. We then used this comprehensive integration of image content for further analysis.

In the process of recognizing image themes, we used the discourse atomic topic modeling (DATM) technique [45]. This method consists of 3 basic steps. First, the data are distilled into the most salient features. Second, we used latent variable modeling to detect potential themes in the corpus. Finally, we used sentence embedding to tightly link these themes to the specific contexts observed.

In order to reveal the topics within this embedding space, we applied K-singular value decomposition, a sparse dictionary learning algorithm introduced by Aharon et al [46] in 2006, to the word vectors. This algorithm yields a set of K vectors called "discourse atoms." Each of the V word vectors in the vocabulary can be expressed as a sparse linear combination of these atom vectors. Using the generative model outlined below (equation 1), the atom vectors are constructed as topics within the embedding space.

XSL•FO
RenderX

Equation 1 is also a simplified version of the latent variable model. The formula expresses the probability of a word w being present at some location t in the corpus. c_t is a vector in the semantic space that represents the underlying meaning of the word in the current context. The similarity between its word vector w and the latent "gist" at that point in the corpus c_t is computed. Z_{ct} is the sum of the similarities so that the distribution sums to 1.

The formula expresses the probability of a word w being present at some location t in the corpus. We used smooth inverse frequency embeddings to derive themes from the observed phrases. Smooth inverse frequency embeddings are calculated using the maximum a posteriori estimate of the combined contextual vectors for a set of contextual terms.

is a weighted average of the word vectors in the context window; words are weighted based on their corpus frequency $p(w)$. Frequent words make a smaller contribution to the estimate of . The α is a hyperparameter.

This procedure allows us to discover latent topics in a corpus and identify the topic that best matches the estimated gist of an observed context window. Finally, we used the relevant words to name each topic, resulting in corresponding topic types for each image.

Since image descriptions frequently comprise many insignificant articles and prepositions, the corpus used for clustering image topics relies on the image tags generated by the RAM. We identified the optimal combination of parameters by evaluating different combinations using root mean squared error. The best results were obtained with a configuration of 7 topics, each with a maximum of 2 atoms per topic, meaning each term could be used in up to 2 topics.

Evaluation of Topic Model Results

Evaluation methods are necessary for assessing the image descriptions recognized by the RAM and the results of topic discovery of image descriptions by DATM.

Topic coherence is a measure of the relatedness and interpretability between words within a topic. The topic coherence proposed by Newman et al [47] uses Pairwise Pointwise Mutual Information for the coherence between topic words. Coherence scores are computed by taking into account the co-occurrence of words in a given context (eg, a document or a sliding window), with the goal of evaluating whether words within a topic are "coherent," that is, whether they form meaningful associations. In our study, we selected the topics with the highest topic coherence from the generated topics by varying the number of the topics and the number of atoms.

In addition, topic diversity [48] is an indicator of variation among topics. The way to find the result for topic diversity is to determine the percentage of representative words of the current topic that do not appear in the representative words of other topics.

Variables

By collecting data and identifying project features, we had gathered specific variables that could potentially affect cancer crowdfunding projects. These variables consist of project basic information, cancer types, image features in crowdfunding images, and the topic of image content.

Funding Percentage is a continuous variable, indicating the percentage of the target amount that the current project has raised. We used *Funding Percentage* to measure the outcome of a crowdfunding project.

In the process of selecting the independent variables, we considered a combination of interdisciplinary literature [22-27] that identifies individual characteristics such as the number of faces, gender, and facial expression. While these factors have been empirically supported in other domains, their effects have yet to be validated in the specific context of health care crowdfunding images. Furthermore, given the centrality of image content in the operational strategies of crowdfunding platforms such as Indiegogo, the visual presentation of different thematic content has been shown to significantly modulate crowdfunding outcomes [35-37]. Therefore, the inclusion of "topic" as an independent variable aims to analyze how it plays a unique role in the field of health care crowdfunding (Table 3).

XSL•FO
RenderX

Table 3. Description of the variables of the regression model, including dependent variable and image characteristics.

Variable	Description
Funding Percentage	The ratio of the donation amount to the target amount
Funding Amount	The amount of donations received
Donation Count	The number of donations received
Description Length	The length of the project description for the crowdfunding project
Title Length	The length of the title description of the crowdfunding project
Time	The time interval between the creation of the crowdfunding project and the crawling of the data
Face Count	Number of individuals in each image
Female Percent	Percentage of female individuals in each image
Age <18	Percentage of individuals younger than the age of 18 years in each image
Age 18-35	Percentage of individuals aged 18-35 years in each image
Age 36-60	Percentage of individuals aged 36-60 years in each image
Age ≥60	Percentage of individuals aged over 60 years in each image
If Smile	Whether there is a smiling face in each image
Topic	The topic of image caption

Regression Model

In order to investigate the influence of image characteristics on crowdfunding results, a logistic regression model was used. In this study, regression models were built to predict crowdfunding outcomes in the presence of multicollinearity as determined by the variance inflation factor value. Additionally, the regression results were visualized as a word cloud to visually illustrate the content of different image themes and their effects by using Python's *WordCloud* package.

Ethical Considerations

The data sources for this study were information available on public platforms, and data collection did not involve direct participation of study participants. To protect privacy, all data were anonymized to ensure that individuals could not be identified from the results. Additionally, all data were stored on a secure server with access restricted to project team members, in accordance with a strict privacy policy.

Results

The Impact of Image Features in Crowdfunding Images on Crowdfunding Outcomes

Of the 243,795 images, 205,496 contained depictions of human figures, indicating a *Face Count* ≥1. The dataset was largely dominated by single-person images, accounting for 58.83% (120,893/205,496) of the total. Multiperson images included from 2 to 10 individuals. Regarding gender distribution, 42.31% (86,945/205,496) of images only contained male individuals, while images featuring only female individuals constituted 29.83% (61,230/205,496). The remaining 27.86% (57,251/205,496) featured a mix of male and female individuals. In total, 58.11% (119,413/205,496) of the total images depicted smiling faces. Moreover, 21.81% (44,818/205,496) of individuals wore hats.

To explore how the individuals' content in the images affects the funding outcome, we built regression models for the attributes. We found the following variables to have a significant impact on funding outcomes. Table 4 presents the outcomes from the regression model.

Table 4. The regression results for the attributes of individuals in the image[a].

Variables	Coefficients	*P* value
Face Count	0.00822	<.001
Female Percent	–0.0201	<.001
Age <18	0.0753	<.001
Age 18-35	0.0432	<.001
Age 36-60	0.0304	<.001
If Smile	0.0446	<.001

[a]The controlled variables included *Funding Amount*, *Donation Count*, *Description Length*, *Title Length*, and *Time*. The significance level is *P*<.001. The *Age* variable takes age ≥60 years as the reference group.

XSL•FO
RenderX

Table 4 includes 4 types of individual variables that have a significant effect on funding outcomes.

- *Face Count*: The results of our experiment indicated a positive correlation between the number of people in the image and funding outcome ($P<.001$).
- *Female Percent*: The regression model indicated that as the percentage of female individuals in the image increases, *Funding Percentage* would decrease. ($P<.001$).
- *Age*: Age ≥60 years was used as the reference group for the regression experiment, which aimed to examine the impact of character age on crowdfunding results. The findings suggested that age-specific personas play a crucial role in determining fundraising ratios, especially among younger crowdfunding initiators ($P<.001$).
- *If Smile*: We found that the presence of someone smiling in the image had a significant positive effect on funding outcome.

The Impact of Topic on Crowdfunding Outcomes

RAM Performance

We used the RAM to analyze image data from 243,795 crowdfunding projects. It obtained tags for various objects in the image and combined them to form an image caption (Table 5).

Table 5. Examples of image description[a].

Image	Tags	Caption
	baby|hospital|mother|woman|boy|hospital bed|child|bottle|person|hold|small|young|little	A woman holding a small child in a hospital bed while the baby is being held by a woman who is holding a bottle
	hair|kitchen|woman|shirt|picture|people|glass|man|couple|smile|pose|stand	A man with glasses and a plaid shirt standing next to a woman in a kitchen smiling for a picture

[a]The people in the pictures are mosaicked to protect privacy.

We analyzed the obtained image descriptions using various methods. Word frequency is determined through tag results for statistical analysis. Term frequency–inverse document frequency is a statistical method for evaluating the importance of a word to a document set or to one of the documents in a corpus [49]. Additionally, n-gram is an algorithm based on a statistical language model [50]. As can be seen in Table 6, the most critical words described in the graph in the text analysis obtained from the word frequency, term frequency–inverse document frequency, and n-gram models are all related to people.

Table 6. Keywords of the image caption.

Keyword category	Words
Top 10 words with the highest frequency	person, woman, man, picture, people, stand, girl, child, primer, young
Keywords	gadget, control, project picture, maple leave, urinal, organism, crystal, mini, dragonfly, knight
Combination of words	(person, man), (picture, people), (people, person), (person, face)

Descriptive Statistics of Image Caption

We used tags to compute word frequencies and generate word cloud visualizations, presenting the variations in content for various types of images. The analyzed image categories comprised differences between images with and without smiles and variances in images depicting individuals of varying age groups.

We categorized the images into “smiling images” and “unsmiling images” based on the presence or absence of smiles on the individuals depicted. The results of the word cloud are shown in Multimedia Appendix 2. We calculated the word frequency and filtered out the 30 words with the highest word frequency in each category of images for comparison. After comparison, we found that the “smiling images” contained more scenes with individuals standing in groups or couples. Conversely, the “unsmiling images” contained more scenes of people lying down in hospital settings.

Images were included in a sample set for each age group if the number of individuals in that age group constituted more than half of the total number of individuals in the image. A comparison of the top 30 words in each category in terms of word frequency revealed that images in the younger than 18 years age group predominantly depict family scenes, while the other age groups tend to depict couples. In addition, hospital scenes are less common in images of people younger than 18 and older than 60 years of age than in the other 2 age groups (Multimedia Appendix 3).

The images differ based on the type of patients with cancer. We selected the 3 cancer type programs with the highest number of occurrences to explore. Corresponding image content for each type of cancer was analyzed with a word cloud, resulting in Multimedia Appendix 4. By comparing the word frequencies of the things that appear in the images for each type of cancer, we found that images related to different types of cancer primarily depict family and medical situations. However, the content varies from project to project. For example, skin cancer

XSL•FO
RenderX

projects show a higher frequency of descriptions related to hand images for patients with skin cancer than other projects. Conversely, childhood cancer projects have fewer occurrences of adult features, such as beards, than other projects.

Topic Modeling and Regression Models

The use of the DATM resulted in the identification of 7 categories, each consisting of multiple keywords. Our topic coherence score was 0.85 out of 1, indicating strong performance [51]. Our topic diversity score was 0.98 out of 1, indicating that we have effectively covered a wide range of topics [52].

The categories were named based on the commonalities and relationships among the keywords. The names of the topics and their corresponding keywords are listed in Table 7.

Table 7. The topic of image caption.

Topic	Related words
Food and cooking	pan, dish, pastry, salad, tea, soup, pie, chocolate, pepper, bread, vegetable
Hair and beauty	hairstyle, earring, braid, beauty, eyebrow, hair, natural, short, lip
City life	garage, luggage, sidewalk, ramp, fire hydrant, cane, building, door, van, train
Clothing and color	white, green, black, blue, red, striped, long, purple, gray, yellow, wear, bald
Rest and play	toy, diaper, crib, bunny, monkey, dog bed, nap, blanket, teddy bear, play with
Medical and health	blood, chest, bandage, thermometer, cord, surgery, dentist, tube, cast, wrist
Diverse locations	rocket, square, site, satellite, knight, traffic light, suburb, mosque, wheat

The regression model shows that the theme of the image has a significant effect on the final fundraising percentage of the project, as shown in Multimedia Appendix 5. We extracted the descriptive words from each image category and organized them into a word cloud for each topic. We analyzed the impact of each theme on the crowdfunding outcome. Red word clouds indicate a positive impact on the crowdfunding outcome, while blue word clouds indicate a negative impact. The word clouds of each topic are ordered from the highest to the lowest degree of influence.

According to Multimedia Appendix 5, several themes have a significant impact on crowdfunding success, each with different levels and directions of influence. In total, 3 specific themes, namely *hair and beauty*, *medicine and health*, and *food and cooking*, have a positive impact on crowdfunding results, in descending order of influence. Conversely, the use of images depicting *city life*, *diverse locations*, and *rest and play* negatively impacted the crowdfunding results for 3 other themes.

The Impact of Image Features and Image Content on Crowdfunding Outcomes

The word cloud analysis from the previous section displays that images possessing diverse character traits have distinct scenarios and contents. Image content and individual traits exert significant influences on crowdfunding results, respectively. Table 8 illustrates the effects on crowdfunding results when both character traits and image content appear in an image.

The analysis revealed that for certain themes—namely *food and cooking*, *city life*, and *medical and health*—the interaction term *Topic×Face Count* showed a positive and significant coefficient (β values of 0.005, 0.0175, and 0.0118, and *P* values of .008, .007, and .005, respectively), indicating that an increase in the number of faces in these themed images correlates with a higher fundraising proportion.

XSL•FO
RenderX

Table 8. The regression results for the individual characteristic image features and the topic of image content in crowdfunding images and topics[a].

Topic	Topic×Face Count	Topic×Female Percent	Topic×If Smile
Food and cooking			
r	0.0050	–0.016	–0.0138
P value	.008	.20	.20
Hair and beauty			
r	–0.0051	–0.0381	–0.0043
P value	.06	.50	.80
City life			
r	0.0175	0.0177	–0.0441
P value	.007	.80	.002
Clothing and color			
r	0.0403	0.035	–0.0919
P value	.40	.20	.005
Rest and play			
r	–0.0024	–0.0182	0.0152
P value	.06	<.001	<.001
Medical and health			
r	0.0118	–0.0055	0.0004
P value	.005	.90	.20
Diverse locations			
r	–0.0046	0.0312	–0.0100
P value	<.001	<.001	.006

[a]The controlled variables were *Funding Amount*, *Donation Count*, *Description Length*, *Title Length*, and *Time*. The significance level is *P*<.001.

Conversely, the *Topic×Female Percent* interaction term suggested that a higher percentage of female faces intensifies the negative impact when negative topics are present, particularly in *rest and play* and *medical and health* themes. However, an increase in *Female Percent* within the *diverse locations* theme positively influenced fundraising, as evidenced by a negative coefficient for *Female Percent* (β=–0.0201; *P*<.001) being offset by the positive interaction term *Topic×Female Percent* (β=0.0312; *P*<.001), resulting in a net positive effect.

When considering the presence of a smiley face in conjunction with the image's topic, the *rest and play* theme correlated with a beneficial increase in funds raised (β=0.0152; *P*<.001). In contrast, themes such as *city life* (β=–0.0441; *P*=.002), *clothing and color* (β=–0.0919; *P*=.005), and *diverse locations* (β=–0.01; *P*=.006) were associated with a negative impact on fundraising.

Counterfactual Analysis

To better understand the causal relationship between the key variables and crowdfunding success in the above study, we used counterfactual analysis, a widely used method for comparative investigation [53].

In the counterfactual analysis, the indicator of each significant variable is changed to zero, and the predicted values of the model are calculated, comparing the growth of the predicted values before and after the change. Table 9 and Table 10 show the results of the counterfactual analysis. The effect of the variable on the dependent variable as reflected in the results of the counterfactual analysis was consistent with the results of the regression model. For example, *Face Count* (*P*<.001) and *If Smile* (*P*<.001) had a positive causal effect on crowdfunding outcomes and *Female Percent* (*P*<.001) had a negative causal effect.

XSL•FO
RenderX

Table 9. Counterfactual analysis of individual characteristic image features and the topic of image content regression[a].

Variables	Percentage of increase
Face Count	–4.79
Female Percent	2.65
Age <18	–2.62
Age 18-35	–7.16
Age 36-60	–2.90
If Smile	–7.90
Food and cooking	–0.89
Hair and beauty	–0.46
City life	0.43
Rest and play	1.22
Medical and health	–0.28
Diverse locations	0.73

[a]Percentage of increase indicates the percentage increase in the average of the predicted values after changing the data compared to before changing the data.

Table 10. Counterfactual analysis of interaction term[a].

Topic	Topic×Face Count	Topic×Female Percent	Topic×If Smile
Food and cooking	–0.15%	—[b]	—
Hair and beauty	0.09%	—	—
City life	–0.21%	—	0.06%
Clothing and color	—	—	0.03%
Rest and play	0.79%	1.46%	–1.77%
Medical and health	–0.12%	—	—
Diverse locations	2.13%	–1.86%	2.84%

[a]Percentage of increase indicates the percentage increase in the average of the predicted values after changing the data compared to before changing the data.

[b]Not applicable.

Discussion

Principal Findings

We found that certain characteristics of individuals in crowdfunding images can significantly influence the success or failure of crowdfunding activities. Specifically, the number of faces, age, gender, and facial expressions, such as smiling, were found to significantly affect the success of crowdfunding campaigns. First, previous studies have shown that the inclusion of faces in crowdfunding campaigns can significantly influence funders' decisions [38,54]. Specifically, increasing the number of faces used leads to more positive effects. Similarly, our study found that obtaining more characters also leads to more donations. Second, our study suggests that cancer campaigns have better outcomes when a higher percentage of their images feature younger people. This finding is consistent with previous studies [37] suggesting that individuals tend to be more generous in their donations to children when viewing medical crowdfunding programs. Third, several studies [26-28] have demonstrated the positive impact of displaying smiles in crowdfunding projects, and our findings are consistent with this. Such images may contribute to a stronger first impression by increasing viewers' empathy, which in turn may lead to better crowdfunding outcomes. Fourth, Elitzur and Solodoha [55] found a negative correlation between female leadership and funding success in rewards-based crowdfunding programs. In contrast, Figueroa-Armijos and Berns [56] reported that backers are more likely to support female-led projects than male-led projects. Elitzur and Solodoha [55] and Figueroa-Armijos and Berns [56] argue that the effect of gender on crowdfunding success varies according to the different patterns and stages of project development. However, our study shows that an increase in the proportion of female individuals in the image of medical crowdfunding cancer projects leads to a decrease in crowdfunding success.

For the thematic content of images, we found that images depicting the *hair and beauty*, *medical and health*, and *food and cooking* themes significantly impacted crowdfunding outcomes. The study suggests that these themes may help to awaken empathy and a sense of justice in viewers when viewing

XSL•FO
RenderX

such images [57]. For example, beautiful images and photographic narratives related to patient health can evoke hope and empathy in viewers, ultimately encouraging donations [58,59]. However, certain image themes, such as *city life*, *diverse locations*, and *rest and play*, may negatively impact the outcome of a crowdfunding project. These themes are not relevant to medical crowdfunding content and may create additional challenges for readers to understand the primary message, resulting in a negative impression [60,61]. Therefore, a more specific and intentional effort is needed to select images that are relevant to the crowdfunding promoter and help viewers understand the current status of the project. This includes ensuring that images are selected that clearly demonstrate the project's progress, the promoter's involvement, and key elements related to the crowdfunding goal. When selecting images, consider the quality, relevance, and expressiveness of the images to ensure that they provide viewers with a comprehensive and clear visual representation of the project.

Our study found that the interaction between individual features in crowdfunding images and image themes in crowdfunding visuals will exert different directional influences on the ultimate fundraising success of medical crowdfunding campaigns. For example, themes such as *medical and health* are positively influenced by an increase in the number of characters in crowdfunding results, while the presence of smiley faces in images for themes such as *city life* is detrimental to crowdfunding results. This dynamic suggests that images with higher thematic and character complexity may deter potential contributors due to the increased cognitive load required to process them, as suggested by Yang et al [16].

The application of counterfactual analysis effectively helps us to distinguish causal effects from statistical correlations [62-64], providing stronger evidence for understanding which factors directly influence the success of crowdfunding projects. In terms of the performance of the results, the characteristics of the 3 components mentioned above, including certain characteristics of individuals, certification image topic, and the interaction, are all consistent with the results of the regression models. This further confirms the causal relationship between these factors and crowdfunding outcomes.

Practical Implications

Selecting an appropriate image theme that aligns with and emphasizes the informational cues of the crowdfunding project may enhance its persuasive appeal [65] and thereby encourage donation behavior. For example, using *medical and health*–related images with more faces has been shown to increase fundraising effectiveness. This strategy not only increases a campaign's appeal but also accelerates the achievement of its financial goals.

Analyzing these dynamics can reveal important patterns in crowdfunding results. For example, it can identify which types of projects attract more support or pinpoint image features and themes that are particularly effective in attracting attention and donations. Gathering such insights is critical to refining crowdfunding strategies and developing more effective program designs. These improvements could lead to increased donor engagement and support, thereby improving overall campaign success.

In addition, these findings have practical applications for the design and functionality of crowdfunding platforms. By guiding users to select and upload images that have a higher likelihood of success, such as those depicting younger people or medical scenarios, platforms can strategically influence the visual presentation of campaigns. This targeted approach can optimize the visual impact of these campaigns, potentially increasing their success rates and making the crowdfunding experience more effective for both creators and donors.

For website designers, designers should focus on the selection and presentation of images on the project page to ensure that they are not only clear [66] but also rich in elements that can touch people's hearts. At the same time, designers can help project sponsors make better choices by creating thematic guidelines or providing a gallery of sample images. Given that different thematic images have different impacts on fundraising results, designers should encourage diversity and provide tools or suggestions to help personalize image content to meet different donor preferences while maintaining overall site consistency and professionalism.

Limitations and Future Research

There are certain limitations to our research that open the way for potential avenues of future investigation. First, our analysis focused exclusively on crowdfunding campaigns for individual cancer treatments on GoFundMe within the United States. This intentional scope precluded the examination of cancer treatment crowdfunding projects in other countries, leaving room for future studies to explore the dynamics of such campaigns on a global scale.

Second, our study of image content was limited to character features and image descriptions. To increase the comprehensiveness of future research, it is necessary to delve into more complex aspects of image features. This includes a deeper exploration of the interactive actions depicted in the images to provide a more nuanced understanding of how characters visually engage with each other.

Third, the study's conclusions are based on currently observed image features but do not sufficiently take into account the dynamic changes in audience preferences. Our study is limited to the analysis of image features, while the deeper psychology of the viewer is not sufficiently explored. There is a lack of in-depth analysis of core factors such as emotion and attention.

Fourth, this study focuses on exploring image associations using a quantitative approach, and future research will combine qualitative methods and take advantage of mixed methods research to explore deeper factors of crowdfunding success.

By addressing these limitations and broadening the scope of image analysis, future research efforts can contribute to a more comprehensive and insightful understanding of crowdfunding dynamics in the context of individual cancer treatments.

XSL•FO
RenderX

Conclusions

In summary, our research examined the influence of image content on the effectiveness of medical crowdfunding initiatives by focusing on specific image features and themes. We rigorously tested 3 hypotheses. Our findings indicate that the representation of individuals in crowdfunding images significantly affects fundraising performance. Certain aspects, such as the presence of multiple individuals, a higher proportion of younger people, and smiles, contribute positively, while a higher proportion of female individuals correlates with a negative impact. The thematic content of the images is also critical, with themes related to *hair and beauty*, *medicine and health*, and *food and cooking* showing positive effects on fundraising. Conversely, themes associated with *city life*, *diverse locations*, and *rest and play* tend to have a negative impact on fundraising efforts. In addition, there is a significant interaction between image themes and the characteristics of the people depicted, influencing the overall fundraising success of the project. This impact varies, increasing or decreasing with the intensity of the image characteristics across themes. Although this study highlights the significant impact of image features on crowdfunding outcomes, viewers' perceptions of images are constantly evolving. Further research into the underlying mechanisms that drive attention in crowdfunding is warranted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Recognize Anything Model (RAM).
[PDF File (Adobe PDF File), 247 KB - jmir_v26i1e58617_app1.pdf]

Multimedia Appendix 2
Word cloud of the smile of individuals.
[PNG File , 184 KB - jmir_v26i1e58617_app2.png]

Multimedia Appendix 3
Word cloud of the age of individuals.
[PNG File , 303 KB - jmir_v26i1e58617_app3.png]

Multimedia Appendix 4
Word cloud of the cancer type.
[PNG File , 230 KB - jmir_v26i1e58617_app4.png]

Multimedia Appendix 5
The impact of topic on funding percentage. This figure demonstrates how different image themes affect the funding outcome. The independent variables include the Funding Amount, Donation Count, Description Length, Title Length, and Time. The significance level is $P<.001$. The correlation coefficient (r) denotes the relationship between the image topic and Funding Percentage.
[PNG File , 279 KB - jmir_v26i1e58617_app5.png]

References

1. Erlandsson A, Björklund F, Bäckström M. Emotional reactions, perceived impact and perceived responsibility mediate the identifiable victim effect, proportion dominance effect and in-group effect respectively. Organ Behav Hum Decis Process 2015;127:1-14. [doi: 10.1016/j.obhdp.2014.11.003]
2. Coke JS, Batson CD, McDavis K. Empathic mediation of helping: a two-stage model. J Pers Soc Psychol 1978;36(7):752-766. [doi: 10.1037//0022-3514.36.7.752]
3. Lee S, Winterich KP, Ross WT. I'm moral, but I won't help you: the distinct roles of empathy and justice in donations. J Consum Res 2014;41(3):678-696. [doi: 10.1086/677226]
4. Tusche A, Böckler A, Kanske P, Trautwein F, Singer T. Decoding the charitable brain: empathy, perspective taking, and attention shifts differentially predict altruistic giving. J Neurosci 2016;36(17):4719-4732. [doi: 10.1523/jneurosci.3392-15.2016]
5. Obrador P, Schmidt-Hackenberg L, Oliver N. The role of image composition in image aesthetics. 2010 Presented at: IEEE International Conference on Image Processing; September 26-29, 2010; Hong Kong, China.
6. Schraven E, van Burg E, van Gelderen M, Masurel E. Predictions of crowdfunding campaign success: the influence of first impressions on accuracy and positivity. J Risk Financ Manag 2020;13(12):1-16. [doi: 10.3390/jrfm13120331]

XSL•FO
RenderX

7. Cox J, Nguyen T, Thorpe A, Ishizaka A, Chakhar S, Meech L. Being seen to care: the relationship between self-presentation and contributions to online pro-social crowdfunding campaigns. Comput Hum Behav 2018;83:45-55. [doi: 10.1016/j.chb.2018.01.014]
8. Childers TL. Memory for the visual and verbal components of print advertisements. Psychol Market 1986;3.3:137-149. [doi: 10.1002/mar.4220030303]
9. Joffe H. The power of visual material: persuasion, emotion and identification. Diogenes 2008;55(1):84-93. [doi: 10.1177/0392192107087919]
10. Majumdar A, Bose I. My words for your pizza: an analysis of persuasive narratives in online crowdfunding. Inf Manag 2018;55(6):781-794. [doi: 10.1016/j.im.2018.03.007]
11. MacInnis DJ, Price LL. The role of imagery in information processing: review and extensions. J Consum Res 1987;13(4):473-491. [doi: 10.1086/209082]
12. Yuan H, Lau RY, Xu W. The determinants of crowdfunding success: a semantic text analytics approach. Decis Support Syst 2016;91:67-76. [doi: 10.1016/j.dss.2016.08.001]
13. Bruce NI, Murthi BPS, Rao RC. A dynamic model for digital advertising: the effects of creative format, message content, and targeting on engagement. J Mark Res 2017;54(2):202-218. [doi: 10.1509/jmr.14.0117]
14. Pieters R, Wedel M. Attention capture and transfer in advertising: brand, pictorial, and text-size effects. J Mark 2004;68(2):36-50. [doi: 10.1509/jmkg.68.2.36.27794]
15. Smith RA. The effects of visual and verbal advertising information on consumers' inferences. J Advert 1991;20(4):13-24. [doi: 10.1080/00913367.1991.10673351]
16. Yang J, Li Y, Calic G, Shevchenko A. How multimedia shape crowdfunding outcomes: the overshadowing effect of images and videos on text in campaign information. J Bus Res 2020;117:6-18. [doi: 10.1016/j.jbusres.2020.05.008]
17. Ba Z, Zhao Y, Song S, Zhu Q. Understanding the determinants of online medical crowdfunding project success in China. Inf Process Manage 2021;58(2):102465. [doi: 10.1016/j.ipm.2020.102465]
18. Ba Z, Zhao Y, Zhou L, Song S. Exploring the donation allocation of online charitable crowdfunding based on topical and spatial analysis: evidence from the tencent gongYi. Inf Process Manage 2020;57(6):102322. [doi: 10.1016/j.ipm.2020.102322]
19. Belleflamme P, Lambert T, Schwienbacher A. Crowdfunding: tapping the right crowd. J Bus Ventur 2014;29(5):585-609. [doi: 10.1016/j.jbusvent.2013.07.003]
20. Liu D, Brass DJ, Lu Y, Chen D. Friendship in online peer-to-peer lending: pipes, prisms, and relational herding. MIS Q 2015;39(3):729-742. [doi: 10.25300/misq/2015/39.3.11]
21. Wu Y, Zhang X, Xiao Q. Appeal to the head and heart: the persuasive effects of medical crowdfunding charitable appeals on willingness to donate. Inf Process Manage 2022;59(1):102792. [doi: 10.1016/j.ipm.2021.102792]
22. Dey S, Duff B, Karahalios K, Fu WT. The art and science of persuasion: not all crowdfunding campaign videos are the same. 2017 Presented at: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing; February 25, 2017; Portland, OR.
23. Kaminski JC, Hopp C. Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. Small Bus Econ 2019 Jul 09;55(3):627-649. [doi: 10.1007/s11187-019-00218-w]
24. Chen S, Wang H, Fang Y, Wang W. Informational and emotional appeals of cover image in crowdfunding platforms and the moderating role of campaign outputs. Decis Support Syst 2023;171:113975. [doi: 10.1016/j.dss.2023.113975]
25. Li L, Yang L, Zhao M, Liao M, Cao Y. Exploring the success determinants of crowdfunding for cultural and creative projects: an empirical study based on signal theory. Technol Soc 2022;70:102036. [doi: 10.1016/j.techsoc.2022.102036]
26. Genevsky A, Knutson B. Neural affective mechanisms predict market-level microlending. Psychol Sci 2015;26(9):1411-1422. [doi: 10.1177/0956797615588467]
27. Wang Z, Mao H, Jessica Li Y, Liu F. Smile big or not? Effects of smile intensity on perceptions of warmth and competence. J Consum Res 2017;43(5):787-805. [doi: 10.1093/jcr/ucw062]
28. Cheng Y, Mukhopadhyay A, Williams P. Smiling signals intrinsic motivation. J Consum Res 2020;46(5):915-935. [doi: 10.1093/jcr/ucz023]
29. Duarte J, Siegel S, Young L. Trust and credit: the role of appearance in peer-to-peer Lending. Rev Financ Stud 2012;25(8):2455-2484. [doi: 10.1093/rfs/hhs071]
30. Jenq C, Pan J, Theseira W. Beauty, weight, and skin color in charitable giving. J Econ Behav Organ 2015;119:234-253. [doi: 10.1016/j.jebo.2015.06.004]
31. Zhang M, Luo L. Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from yelp. Manag Sci 2023;69(1):25-50. [doi: 10.1287/mnsc.2022.4359]
32. Zhang S, Lee D, Singh PV, Srinivasan K. What makes a good image? Airbnb demand analytics leveraging interpretable image features. Manag Sci 2022;68(8):5644-5666. [doi: 10.1287/mnsc.2021.4175]
33. Black S. Fictions of humanitarian responsibility: narrating microfinance. J Hum Rights 2013;12(1):103-120. [doi: 10.1080/14754835.2013.754293]
34. von Engelhardt J, Jansz J. Distant suffering and the mediation of humanitarian disaster. In: World Suffering and Quality of Life. Social Indicators Research Series. Berlin, Germany: Springer; 2015:75-87.

XSL•FO
RenderX

35. Buskila G, Perez D. The impact of victims' imagery on charity crowdfunding campaigns: how photos of victims doing nothing to help themselves elicit fewer donations. J Advert Res 2022;62(4):385-396. [doi: 10.2501/jar-2022-021]
36. Liu Y, Yu S, Ravichandran T. Direct gaze, story narration, and online medical crowdfunding outcomes. Tom Shohfi. URL: https://tom.shohfi.com/don/papers/Don_Shohfi_2021_Liu.pdf [accessed 2024-10-17]
37. Ren J, Raghupathi V, Raghupathi W. Understanding the dimensions of medical crowdfunding: a visual analytics approach. J Med Internet Res 2020;22(7):e18813. [doi: 10.2196/18813]
38. Zhao K, Zhou L, Zhao X. Multi-modal emotion expression and online charity crowdfunding success. Decis Support Syst 2022;163:113842. [doi: 10.1016/j.dss.2022.113842]
39. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst 2011;103(2):117-128 [FREE Full text] [doi: 10.1093/jnci/djq495] [Medline: 21228314]
40. Zhang X, Tao X, Ji B, Wang R, Sörensen S. The success of cancer crowdfunding campaigns: project and text analysis. J Med Internet Res 2023;25:e44197 [FREE Full text] [doi: 10.2196/44197] [Medline: 36692283]
41. Alibaba Cloud's Vision Intelligent Application Programming Interface Platform. Aliyun. 2022. URL: https://vision.aliyun.com/?spm=a2c4g.11174283.0.0.1a7e4924XmlIr4 [accessed 2024-10-17]
42. Zhang Y, Huang X, Ma J, Li Z, Luo Z, Xie Y, et al. Recognize anything: a strong image tagging model. 2024 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 17-21, 2024; Seattle, WA p. 1724-1732.
43. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L. Segment anything. 2023 Presented at: IEEE/CVF International Conference on Computer Vision (CVPR); June 18-22, 2023; Vancouver, BC p. 4015-4026.
44. Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, et al. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. Int J Comput Vis 2020;128(7):1956-1981. [doi: 10.1007/s11263-020-01316-z]
45. Bewong M, Wondoh J, Kwashie S, Liu J, Liu L, Li J, et al. DATM: a novel data agnostic topic modeling technique with improved effectiveness for both short and long text. IEEE Access 2023;11:32826-32841. [doi: 10.1109/access.2023.3262653]
46. Aharon M, Elad M, Bruckstein A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 2006;54(11):4311-4322. [doi: 10.1109/tsp.2006.881199]
47. Newman D, Lau JH, Grieser K, Baldwin T. Automatic evaluation of topic coherence. 2010 Presented at: HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics; June 2, 2010; Los Angeles, CA p. 100-108 URL: https://aclanthology.org/N10-1012
48. Dieng AB, Ruiz FJR, Blei DM. Topic modeling in embedding spaces. Trans Assoc Comput Linguistics 2020;8:439-453. [doi: 10.1162/tacl_a_00325]
49. Jones KS. A statistical interpretation of term specificity in retrieval. J Doc 2004;60(5):493-502.
50. Broder AZ, Glassman SC, Manasse MS, Zweig G. Syntactic clustering of the web. Comput Networks ISDN Syst 1997;29(8-13):1157-1166. [doi: 10.1016/s0169-7552(97)00031-7]
51. Syed S, Spruit M. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. 2017 Presented at: IEEE International Conference on Data Science and Advanced Analytics (DSAA); October 19-21, 2017; Tokyo, Japan. [doi: 10.1109/DSAA.2017.61]
52. Rani R, Lobiyal DK. An extractive text summarization approach using tagged-LDA based topic modeling. Multimed Tools Appl 2021;80(3):3275-3305. [doi: 10.1007/s11042-020-09549-3]
53. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature 2021;589(7840):82-87. [doi: 10.1038/s41586-020-2923-3] [Medline: 33171481]
54. Gleasure R, Feller J. Emerging technologies and the democratisation of financial services: a metatriangulation of crowdfunding research. Inf Organ 2016;26(4):101-115. [doi: 10.1016/j.infoandorg.2016.09.001]
55. Elitzur R, Solodoha E. Does gender matter? Evidence from crowdfunding. J Bus Ventur Insights 2021;16:e00268. [doi: 10.1016/j.jbvi.2021.e00268]
56. Figueroa-Armijos M, Berns JP. Vulnerable populations and individual social responsibility in prosocial crowdfunding: does the framing matter for female and rural entrepreneurs? J Bus Ethics 2022:1-18. [doi: 10.1007/s10551-020-04712-0]
57. Yazdani E, Chakravarty A, Inman J. Identifying Emotions in Images and Their Effects on Donation Behavior in Online Crowdsourcing Platforms. Atlanta, GA: SSRN; 2021.
58. Li Y, Xie Y. Is a picture worth a thousand words? An empirical study of image content and social media engagement. J Mark Res 2020;57(1):1-19. [doi: 10.1177/0022243719881113]
59. Wang X, Wu Y, Guo J, Yang L. Says what in your photos? The impacts of photographic narratives on medical crowdfunding performance. 2020 Presented at: Pacific Asia Conference on Information Systems; Dubai, UAE; May 1, 2020 URL: https://www.researchgate.net/publication/341176146_Says_what_in_your_photos_The_impacts_of_photographic_narratives_on_medical_crowdfunding_performance
60. Heckler SE, Childers TL. The role of expectancy and relevancy in memory for verbal and visual information: what is incongruency? J Consum Res 1992;18(4):475-492. [doi: 10.1086/209275]
61. Lee YH, Mason C. Responses to information incongruency in advertising: the role of expectancy, relevancy, and humor. J Consum Res 1999;26(2):156-169. [doi: 10.1086/209557]

XSL•FO
RenderX

62. Du Z, Zhang L. Home-purchase restriction, property tax and housing price in China: a counterfactual analysis. J Econom 2015;188(2):558-568. [doi: 10.1016/j.jeconom.2015.03.018]
63. Manda J, Alene AD, Tufa AH, Abdoulaye T, Wossen T, Chikoye D, et al. The poverty impacts of improved cowpea varieties in Nigeria: a counterfactual analysis. World Dev 2019;122:261-271. [doi: 10.1016/j.worlddev.2019.05.027]
64. Wang J, Luo J, Zhang X. How COVID-19 has changed crowdfunding: evidence from GoFundMe. Front Comput Sci 2022;4:893338. [doi: 10.3389/fcomp.2022.893338]
65. Brown C, Carpenter G. Why is the trivial important? A reasons - based account for the effects of trivial attributes on choice. J Consum Res 2000;26(4):372-385. [doi: 10.1086/209569]
66. Wessel A. Investigating the influence of website design and message clarity on consumers' response in the online complaint process [Thesis]. University of Twente. 2016. URL: https://essay.utwente.nl/70797/ [accessed 2024-10-28]

Abbreviations

DATM: discourse atomic topic modeling
RAM: Recognize Anything Model

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by S Bhavnani, M Saric; comments to author 05.06.24; revised version received 16.06.24; accepted 25.07.24; published 15.11.24.

Please cite as:
Wang R, Xu H, Zhang X
Impact of Image Content on Medical Crowdfunding Success: A Machine Learning Approach
J Med Internet Res 2024;26:e58617
URL: https://www.jmir.org/2024/1/e58617
doi:10.2196/58617
PMID:

©Renwu Wang, Huimin Xu, Xupin Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of an App-Based Mobile Triage System for Mass Casualty Incidents: Within-Subjects Experimental Study

Martin Schmollinger[1*], Prof Dr; Jessica Gerstner[2*], MSc; Eric Stricker[2], MSc; Alexander Muench[2], MD, Dr med; Benjamin Breckwoldt[2], MD; Manuel Sigle[3], MD, Dr med; Peter Rosenberger[2], MD, Prof Dr; Robert Wunderlich[2,4], MSc, MD, Dr med

[1]School of Informatics, Reutlingen University, Reutlingen, Germany

[2]University Department of Anesthesiology and Intensive Care Medicine, University Hospital Tübingen, Eberhard Karls University, Tübingen, Germany

[3]Department of Cardiology and Angiology, University Hospital Tübingen, Eberhard Karls University, Tübingen, Germany

[4]German Society for Disaster Medicine (Deutsche Gesellschaft für Katastrophenmedizin), Kirchseeon, Germany

[*]these authors contributed equally

Corresponding Author:
Robert Wunderlich, MSc, MD, Dr med
University Department of Anesthesiology and Intensive Care Medicine
University Hospital Tübingen
Eberhard Karls University
Hoppe-Seyler-Str.3
Tübingen, 72076
Germany
Phone: 49 70712981116
Fax: 49 7071295530
Email: Robert.Wunderlich@med.uni-tuebingen.de

Abstract

Background: Digitalization in disaster medicine holds significant potential to accelerate rescue operations and ultimately save lives. Mass casualty incidents demand rapid and accurate information management to coordinate effective responses. Currently, first responders manually record triage results on patient cards, and brief information is communicated to the command post via radio communication. Although this process is widely used in practice, it involves several time-consuming and error-prone tasks. To address these issues, we designed, implemented, and evaluated an app-based mobile triage system. This system allows users to document responder details, triage categories, injury patterns, GPS locations, and other important information, which can then be transmitted automatically to the incident commanders.

Objective: This study aims to design and evaluate an app-based mobile system as a triage and coordination tool for emergency and disaster medicine, comparing its effectiveness with the conventional paper-based system.

Methods: A total of 38 emergency medicine personnel participated in a within-subject experimental study, completing 2 triage sessions with 30 patient cards each: one session using the app-based mobile system and the other using the paper-based tool. The accuracy of the triages and the time taken for each session were measured. Additionally, we implemented the User Experience Questionnaire along with other items to assess participants' subjective ratings of the 2 triage tools.

Results: Our 2 (triage tool) × 2 (tool order) mixed multivariate analysis of variance revealed a significant main effect for the triage tool ($P<.001$). Post hoc analyses indicated that participants were significantly faster ($P<.001$) and more accurate ($P=.005$) in assigning patients to the correct triage category when using the app-based mobile system compared with the paper-based tool. Additionally, analyses showed significantly better subjective ratings for the app-based mobile system compared with the paper-based tool, in terms of both school grading ($P<.001$) and across all 6 scales of the User Experience Questionnaire (all $P<.001$). Of the 38 participants, 36 (95%) preferred the app-based mobile system. There was no significant main effect for tool order ($P=.24$) or session order ($P=.06$) in our model.

Conclusions: Our findings demonstrate that the app-based mobile system not only matches the performance of the conventional paper-based tool but may even surpass it in terms of efficiency and usability. This advancement could further enhance the potential of digitalization to optimize processes in disaster medicine, ultimately leading to the possibility of saving more lives.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e65728) doi:10.2196/65728

KEYWORDS

disaster medicine; mass casualty incidents; digitalization; triage; Germany; mobile triage app

Introduction

Background

In recent years, the number of disasters involving mass casualty incidents (MCIs) has significantly increased [1], driven by factors such as terrorism, climate change, and war. While disaster medicine procedures are well-established and effective, there remains potential for optimization through digitalization [2]. Currently, information collected manually and on paper during an MCI is often incomplete, sometimes disorganized, and potentially unreliable. This includes critical details such as the number of patients, their locations, identifying characteristics, and injury severity. These limitations can hinder decision makers' ability to obtain timely and accurate information essential for an effective response. This information, especially from rescue teams who are first on the scene, is crucial for effective mass casualty and disaster management, as it can save lives and provide valuable insights for scientific research to enhance future operations [3]. First responders must prioritize patients based on injury severity to ensure that those in critical condition receive immediate attention. Incident commanders coordinate rescue resources and communicate with hospitals about capacity, requiring a comprehensive overview of patient numbers, injury severity, and locations. Upon arrival at the incident scene, rescue teams search for casualties and categorize patients [4] using triage algorithms such as Simple Triage and Rapid Treatment (START) [5], modified Simple Triage and Rapid Treatment (mSTART) [6], and jumpSTART for children [7,8], classifying them by health status with color codes (Textbox 1).

Textbox 1. Triage categories following the modified Simple Triage and Rapid Treatment algorithm.

1. Red
   - Patients with life-threatening injuries that immediately need an adequate treatment.
2. Yellow
   - Patients that are seriously injured and not able to walk. The treatment is still urgent but can be delayed after the first group.
3. Green
   - Patients that are slightly injured, able to walk, and can be treated later. Sometimes they are also called "walking wounded."
4. Black
   - Deceased casualties or casualties that are dying and only need supportive treatment.
5. BLUE
   - Sometimes blue is used for patients that have no chance of surviving under the given circumstances considering the available resources.

Biographic data and initial treatments are recorded on handwritten cards attached to patients, aiding in communication and tracking. All information is relayed to incident commanders via radio, who coordinate with various rescue organizations and manage resources and patient transport based on the triage data [9,10]. Rescue services establish areas for admission, treatment, decontamination, and transportation, where patients are re-triaged to monitor changes in their health status [11]. This process ensures that the most critical patients receive timely and appropriate care.

Weaknesses and Problems With the Current Emergency Medical Care During MCIs

Triage data collected by rescue teams are currently communicated to the command post via radio, which must be concise and minimal due to time constraints [12]. Radio communication is not suitable for transmitting large volumes of data. Communication from rescue teams to incident commanders and dispatch centers follows a many-to-one pattern, requiring recipients to manually compile information—a process that is both time-consuming and prone to error. During triage, rescue teams gather additional data, including patient biographic information and minor treatment details. Transmitting all this information via radio is too time-consuming, so only the most critical data are communicated, resulting in incomplete information that can hinder effective disaster management [13,14]. Incident commanders often need to contact rescue teams for missing details, adding to communication time. Furthermore, verbal descriptions of patient locations are less precise than GPS tracking, complicating rapid evacuations. Rescue teams move quickly from patient to patient and can only intermittently update the command post on their findings, causing delays in situational awareness. This slow development of an overall view of the disaster scene can lead to outdated or incorrect decisions, requiring reorganization and resulting in time loss [15]. The manual, communication-intensive processes reduce work quality and increase task execution times, which can dramatically impact patient outcomes. Automating the process with a

XSL•FO
RenderX

software system can help address these issues by providing a common, centralized data source accessible to all involved parties [16].

The App-Based Mobile System "KatApp"

The weaknesses in the current rescue process stem from the absence of a common, centralized data source accessible to all involved parties. If rescue teams utilized software for triaging patients, the results could be immediately stored in a shared data source. Incident commanders connected to this data source would have access to real-time information tailored to their needs, thereby eliminating several weaknesses through process automation. This would eliminate the need for manual compilation of information and facilitate the creation of complete data sets, minimizing data incompleteness and reducing additional communication costs. The software could also automatically track GPS coordinates of triage locations and present real-time data, enhancing the overall quality and speed of the process. By removing several manual tasks, the overall quality of work would improve, reducing errors and allowing more time for critical tasks, thereby accelerating process execution. From a functional perspective, developing such a software system may seem straightforward; however, several nonfunctional requirements must also be considered, including usability, data security, data privacy, and system availability. The software must be intuitive and easy to use under stressful conditions, such as low visibility (eg, bright sunlight or darkness). Additionally, data security and privacy safeguards must be in place to prevent unauthorized access to sensitive patient information, and the system must be reliable and instantly accessible in real-time disaster scenarios.

These requirements are essential for the system to be effective in real disaster scenarios.

To address these requirements, we utilized software prototyping [17] to design and develop a prototype named "KatApp." This prototype was tested in several disaster medicine training sessions, during which we collected both quantitative data and qualitative feedback for evaluation. The system uses smartphones for triage and tablets for the command post, all connected to a shared cloud-based architecture. Applications running on the client side are part of the front-end layer and utilize services from the backend layer to read or write data, with the backend services relying on a persistence layer for permanent data storage. To connect patients with their digital data records during a disaster, rescue teams scan a patient-specific QR code before starting triage, seamlessly linking patients to their digital records (Figure 1).

Figure 1. Digital triage procedure using the KatApp Software System.

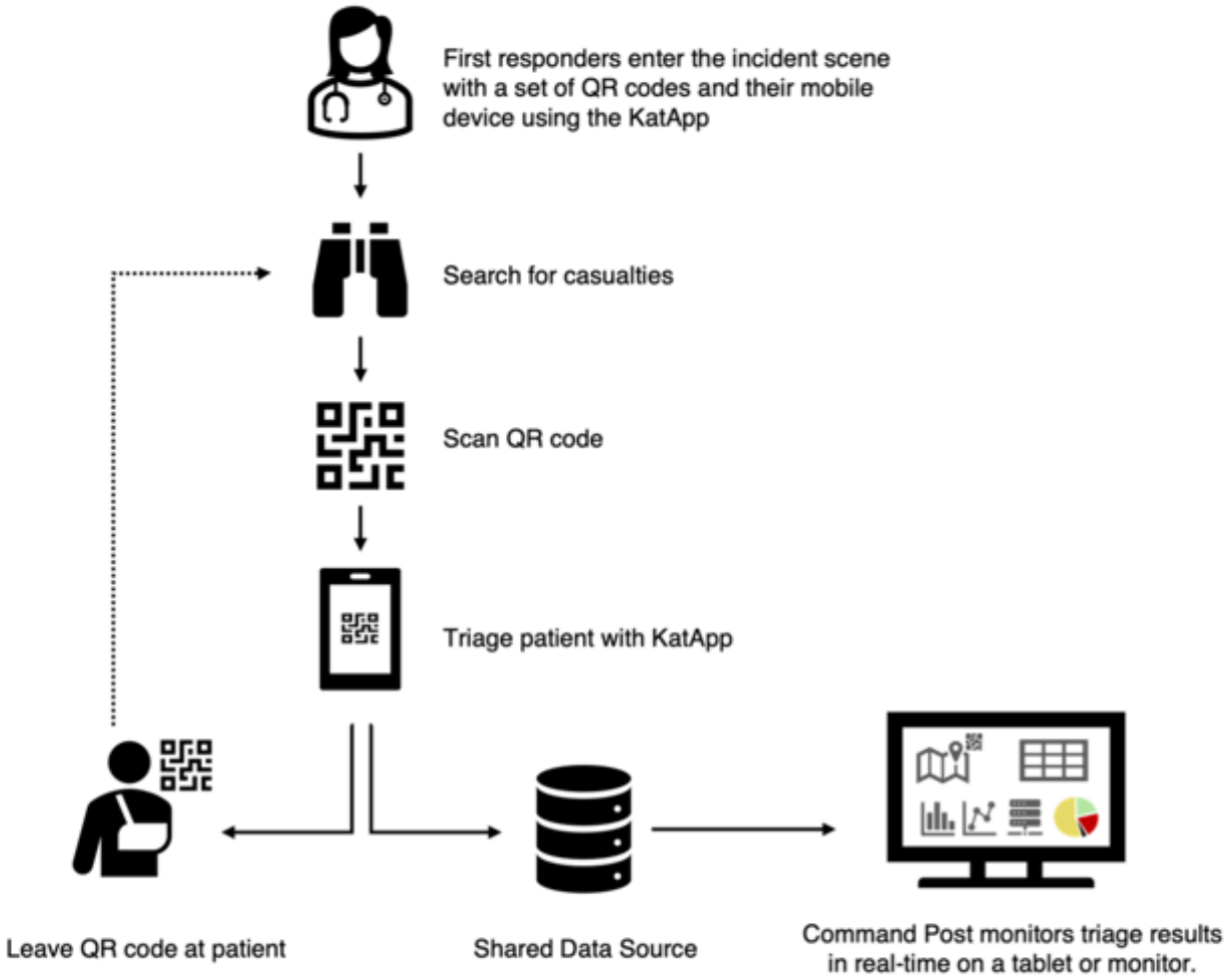

XSL•FO
RenderX

The software architecture of the KatApp system follows a modern cloud-based architecture style, integrating backend and front-end components. The front-end layer comprises 2 client applications: a native app for triaging patients, used by rescue teams, and a web application that implements a dashboard for the command post. Both client applications access services from the backend layer, with the native app using a REST (Representational State Transfer) application programming interface and the web application connecting via a web socket connection.

The triage app is developed using Google's Flutter framework [18], which supports multiple platforms from a single codebase—an essential feature for training and future use by different organizations. The app's user interface is designed for ease of use during triage, transitioning from flowchart-based designs to a page flow that represents the triage algorithms. In addition to implementing the triage algorithms (mSTART and jumpSTART), the app includes several features such as day and night modes; using the smartphone flashlight to easily scan the QR code, even at night; and an overview of completed triages. The command post dashboard was developed as a web application using the React framework [19,20]. It displays information on the exact GPS locations of patients using OpenStreetMap [21], along with a list of patients who have been triaged and the time elapsed since the start of triage handling, together with the current time.

The dashboard application connects to the backend service using a web socket connection. As soon as new triage records are received by the backend service from the rescue teams' apps, they are pushed to the dashboard application through this connection. This ensures that the command post always has the most current information without the need to manually pull data. The system's design is well-suited for serverless computing [22,23], as it is only required during real disasters or training sessions. This allows for billing based solely on active usage, which aligns with the sporadic nature of MCIs and has the potential to reduce overall operational expenses.

After prototyping KatApp, several functionality tests were conducted under realistic conditions, including the use of waterproof devices and gloves. These tests were followed by training sessions aimed at gathering qualitative feedback to refine the prototype.

Logging functionalities were subsequently added to measure triage time and accuracy. Early usability insights were incorporated, and the system was standardized for evaluation in a within-subjects experimental study comparing it with the paper-based triage method. Table 1 presents the early implementations of our prototype in pilot studies. Several usability insights for the app and dashboard were promptly added to the backlog. To standardize the evaluation of our refined prototype against the conventional paper-based triage tool, we conducted an experiment using a within-subjects design.

Based on the insights gained from these early implementations, we made several improvements to the KatApp prototype. These refinements included enhancements to the user interface, optimization of the triage algorithms, and improvements to the data logging capabilities. The resulting version of the app, which incorporated feedback from all 3 training sessions, was used in the subsequent comparative experiment.

Table 1. Early-stage implementations of the KatApp prototype.

Training session	Objective	Sample	Procedure	Results
Training 1	Obtain first qualitative user feedback	N=15 rescue service personnel, emergency physicians, and medical students	Participants triaged 60 patient cards at different locations (ie, at the location of the incident, at the treatment area, and before the release of the patient)	• 150 triages were made • More than 90% (11/12) of the app users and 100% (4/4) of the dashboard users stated that they would like to use such a system in a real disaster
Training 2	Use a prototype in an evacuation training (earthquake) and finish the evacuation within 90 minutes	N=50 medical students as first responder teams, supervised by experienced emergency physicians	Participants triaged 18 patients (played by actors)	• 48 triages were made • The scenario could be finished in time • 2 patients were triaged in the wrong category • Average time of a triage: 42 seconds
Training 3	Use prototype in an evacuation training (flood warning + gas explosion) and finish the evacuation within 75 minutes	N=50 rescue service personnel and emergency physicians	Participants triaged 23 patients (played by actors)	• 50 triages were made • All patients were triaged in the correct category • Average time of a triage: 37 seconds

Research Objective

The study aimed to evaluate the KatApp as a triage tool for emergency and disaster medicine, which was previously designed and developed by our research group, as described in the "Background" section. The objective was to determine whether the KatApp is superior to paper-based triage tools in terms of speed, accuracy, and user experience among emergency medicine experts. We hypothesized that the KatApp would be

XSL•FO
RenderX

at least as accurate as the paper-based tool, while also being faster and more user-friendly.

Methods

Ethical Considerations

The Ethics Committee of the Faculty of Medicine at Tübingen University Hospital approved the study (approval number 260/2024BO2). All methods were implemented in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

Study Design and Setting

The study was designed as a within-subjects experimental design, with the triage tool serving as a within-subjects factor (comparing the paper-based tool and KatApp). The order of the triage tools was counterbalanced to mitigate carryover effects.

The study was conducted at the German Red Cross in Kirchentellinsfurt on behalf of the University Hospital Tübingen, in collaboration with Reutlingen University, in June 2024. Participation in the study was voluntary.

Recruitment

Invitations to the triage training course and calls for participation in the study were sent to a mailing list from the Tübingen District Association of the German Red Cross, which includes approximately 500 email addresses of rescue service personnel, physicians, and medical students with German Red Cross Emergency Medical Technician (EMT) qualifications.

To determine the sample size for the study, a statistical power analysis was conducted using G*Power (version 3.1 for MacOS; Universität Düsseldorf) [24]. The analysis indicated a required sample size of 45 participants, with an effect size of Cohen d=0.5, a statistical power of 0.95, and a significance level (P value) of .05. It is important to note that, due to the limited number of similar publications, the statistical power analysis was based on general considerations regarding the trade-offs between the ability to detect medium-sized effects and the feasibility of obtaining a sufficiently large sample. To account for potential no-shows and dropouts, we decided to recruit an additional 10 participants, bringing the total to 55.

Study Procedures

After participants registered for the triage training course and study, they received instructions on how to install the app and details about the study procedure 1 week before the course.

To familiarize participants with the mSTART algorithm using both the paper-based tool and the KatApp, an introduction was provided at the beginning of the course. Participants were then asked to complete 2 triage sessions set in the context of a simulated MCI, involving a terrorist driving into the fan zone of a public screening of the UEFA (Union of European Football Associations) European Football Championship 2024. In one session, participants used the KatApp, while in the other session, they used the conventional paper-based tool. The order of the tools was randomized. Each session involved classifying 30 patient cards into 1 of 4 triage categories, with the cards distributed along a course of exactly 96.32 m for each triage tool. The 2 sessions were conducted simultaneously, with participants starting their sessions every 5 minutes. Refer to Figure 2 for a sketch of the experimental setup, Multimedia Appendix 1 for a drone photograph of the setting, and Multimedia Appendix 2 for a drone video.

After each session, participants completed a web-based questionnaire using SurveyMonkey (SurveyMonkey Inc.). Following a recovery period of approximately 30 minutes, they began the second session with the alternate triage tool.

Figure 2. Sketch of the experimental setup. Each session consisted of 30 patient cards (marked as triangles in this sketch) that were distributed along a course of 96.32 m. Emergency vehicles and a fog machine blocked the view between the 2 courses. In addition, simulated sounds of a mass casualty incident were played via connected sound boxes that were placed in the middle of each course. Immediately after the sessions, the participants were asked to scan a QR code to complete the follow-up questionnaire.

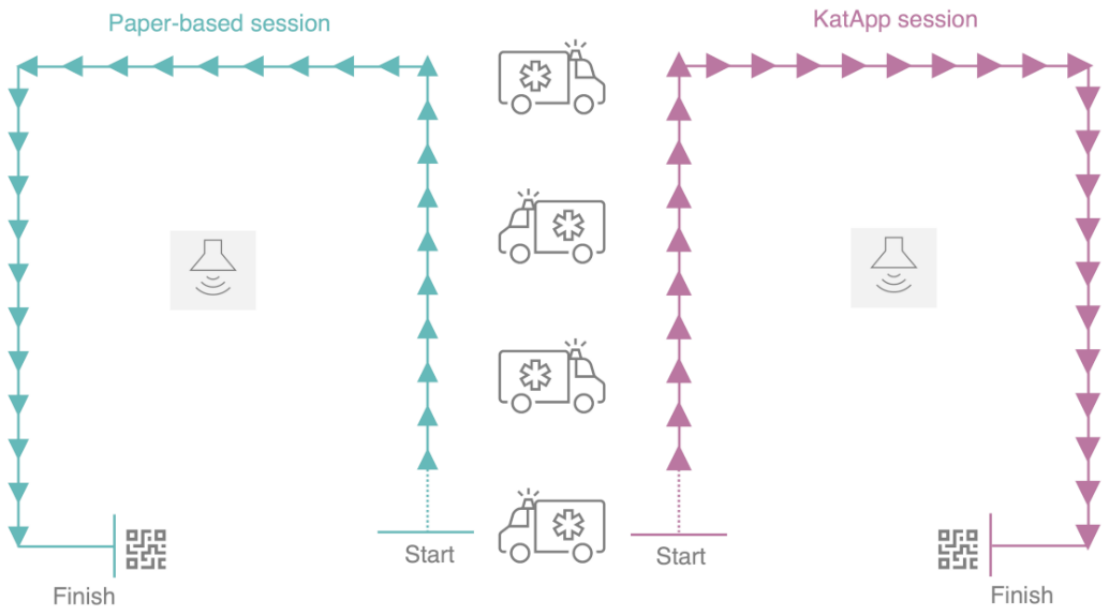

XSL•FO
RenderX

Patient Cards

The patient cards were initially developed using a preexisting artificial patient database [25,26] and subsequently adapted by the study team. Each card featured a pictogram to illustrate the injury pattern, along with details on injury severity (Abbreviated Injury Scale) and vital parameters, including the Glasgow Coma Scale, respiratory rate, systolic blood pressure, radial pulse, capillary refill time, walking ability, and age. Additionally, a short storyboard provided further context regarding the circumstances of the accident (Multimedia Appendix 3). To ensure comparability between the 2 sessions, the number of patient cards in each triage category was kept identical: red (6/30, 20%), yellow (13/30, 43%), green (9/30, 30%), and black (2/30, 7%). This categorization was based on empirical data from MCIs [27]. Additionally, we paired patient cards to match them in terms of overall injury severity (Abbreviated Injury Scale) and injury pattern, minimizing the impact of any differences in the patient cards on the outcome variables between the triage tools. The order of patient cards in each course was randomized using the online tool RANDOM.ORG (Randomness and Integrity Services Limited), ensuring that the randomization was consistent for each participant.

Triage Tools

Paper-Based Tool

When using the conventional paper-based tool, participants were required to complete a casualty card (see Multimedia Appendix 4) from their equipment. This process involved manually indicating the triage category by folding the card and attaching it to the patient card. Additionally, participants needed to record the triage category on a documentation sheet. To assist with the triage process, they were provided with an mSTART checklist [6].

KatApp

When using the KatApp, participants were required to attach a QR code from their equipment to each patient. They then scanned the QR code with the KatApp's QR code scanner and conducted the triage process as previously described (see Figure 1 and Multimedia Appendix 5). Participants used their smartphones, which were connected to the mobile network, allowing the triage results to be immediately forwarded to a central dashboard.

System Architecture Behind the KatApp

All software components are managed using repositories on GitHub (GitHub, Inc.). Automatic deployments are facilitated by GitHub Actions and custom scripts [28]. The iOS app is deployed to TestFlight (Apple Inc.), while the Android app is deployed to Firebase (Google LLC/Alphabet Inc.). Rescue teams can install the desired version from these platforms before training. The dashboard web application and the backend are deployed on Amazon Web Services (AWS; Amazon Web Services, Inc.). The dashboard functions as a web application hosted on AWS, while the backend services are implemented as AWS Lambda functions [29]. These services utilize AWS DynamoDB for user management, logging, and triage records, and AWS S3 for binary data storage. Figure 3 provides a detailed depiction of the architecture.

The KatApp system operates as a distributed system, dependent on the connectivity of its various nodes, which include the mobile devices used by rescue teams, the tablets at the command post, and the backend services. In standard operation mode, it is assumed that an internet connection is accessible at the incident scene, enabling the mobile apps used by rescue teams to access the web application and backend services hosted in the cloud. The dashboard web application is accessible through browsers on tablets or computers at the command post, allowing it to establish a web socket connection to the KatApp backend. In an ideal scenario, this operational mode would suffice. However, as highlighted by the fallacies of distributed computing [30], networks can often be unreliable, particularly during disasters, when essential communication infrastructure may be compromised or overwhelmed, or in areas with inadequate network coverage. To tackle these challenges, the KatApp system incorporates a caching mechanism within the triage app, enabling rescue teams to continue triaging patients even in the absence of an internet connection. Triage records are stored locally and are automatically transmitted once connectivity is restored. Furthermore, the KatApp system features an operational mode that functions independently of an internet connection, ensuring that critical triage processes can continue uninterrupted during emergencies. A local wireless network can be established by the command post or a technical organization on-site, featuring a server that hosts the dashboard application and backend services. Mobile devices can connect to this local network, enabling the system to operate as intended. As the backend services were developed as AWS Lambda functions, they cannot be executed locally. To address this limitation, an additional abstraction layer was implemented, allowing the use of alternative open-source or custom-built components for each AWS service. This ensures that the KatApp system remains functional and efficient, even in environments where internet connectivity is compromised.

XSL•FO
RenderX

Figure 3. Schematic overview of the KatApp architecture without deployment and identity and access management. API: application programming interface; AWS: Amazon Web Services; DB: database; SPA: single-page application.

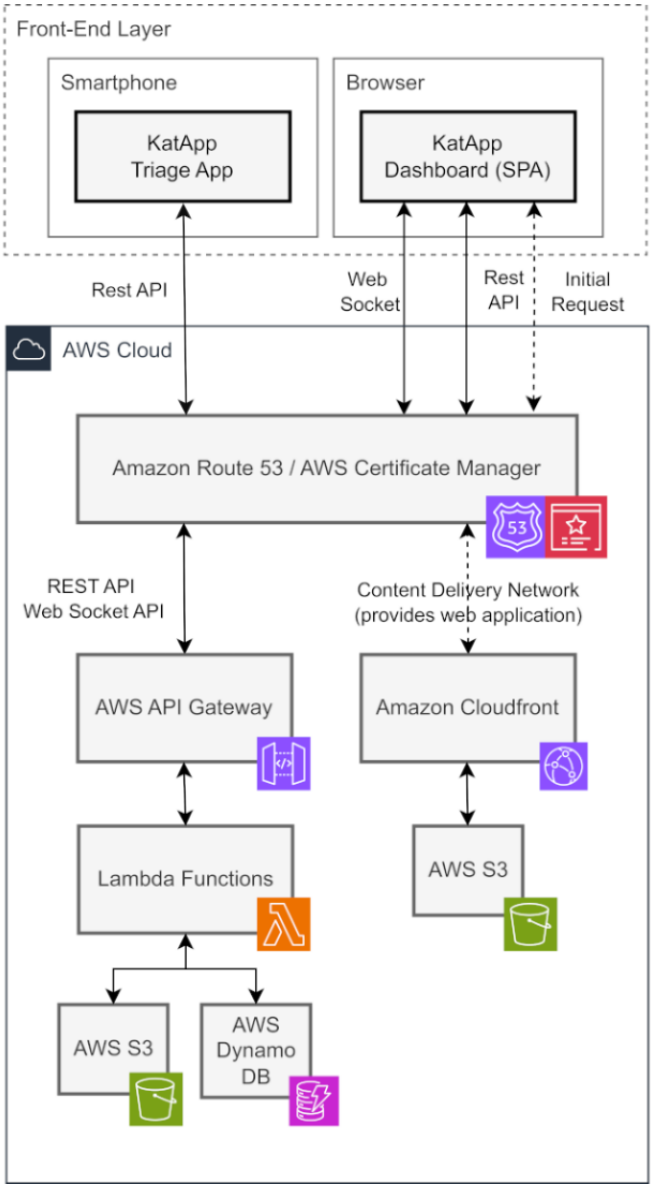

Measuring and Outcomes

Duration and Quality of Triages

The study staff recorded the time at the beginning and end of each course as participants crossed the start and finish lines. This timing encompassed the entire triage process, including walking between patients, applying QR codes or attaching paper triage cards (depending on the method used), and conducting the triage assessment itself. Additionally, the number of correct triages per session was evaluated by analyzing participants' documentation sheets from the paper-based tool, as well as the data collected by the KatApp.

Subjective Ratings

Participants were required to rate both triage tools using the German grading system, where a score of 1 represents "excellent" and a score of 6 indicates "insufficient." They were also asked to express their preference between the 2 tools through a forced-choice item.

User Experience Questionnaire

We used the German version of the User Experience Questionnaire (UEQ) [31] to evaluate participants' experiences with each triage tool. The UEQ comprises 26 items organized into 6 scales: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. Each item is presented as a semantic differential, requiring participants to rate their responses on a 7-point Likert scale, which is coded from –3 to +3 (Multimedia Appendix 6). *Attractiveness* refers to the overall impression of the product (Does a user like or dislike the product?), whereas *Perspicuity* assesses the ease with which users can get familiar with the product and learn how to use it. *Efficiency* indicates whether users can complete their tasks without unnecessary effort, while *Dependability* refers to the sense of control, security, and predictability in the interaction. *Stimulation* measures how exciting and motivating the product is to use it, while *Novelty* assesses whether the product is creative and captures the interest of users. Figure 4 illustrates the overall experimental design.

XSL•FO
RenderX

Figure 4. Within-subjects experimental design that was used for the evaluation of the KatApp compared with the conventional paper-based tool. Each participant conducted 2 sessions in which they had to triage 30 patient cards. The tool order was counterbalanced to compensate for possible carry-over effects. UEQ: User Experience Questionnaire.

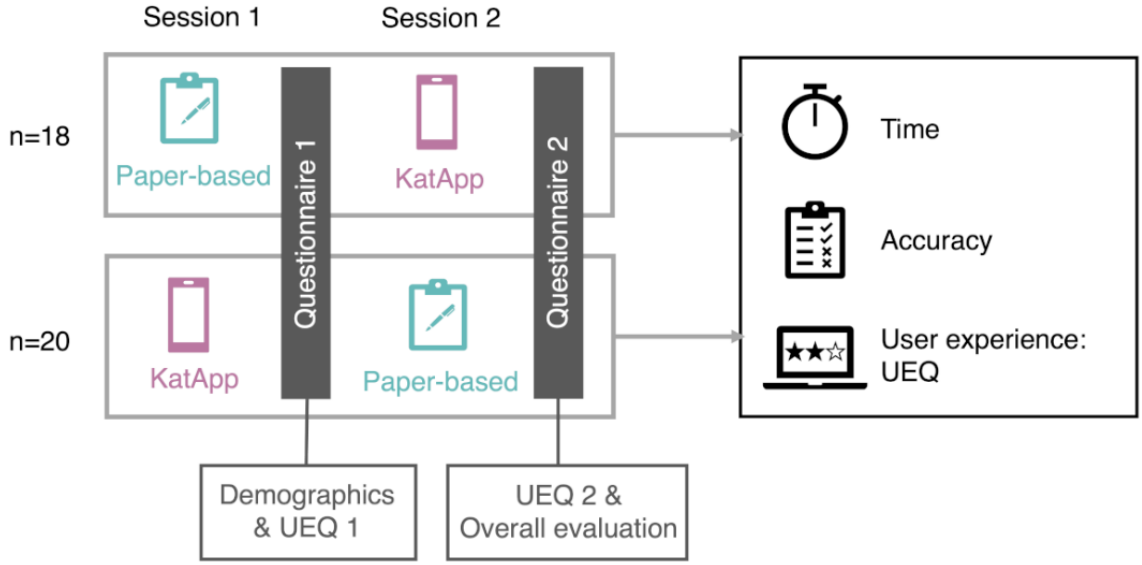

Statistical Analysis

SPSS version 29 for MacOS (IBM Corp.) was utilized for data management and statistical analyses.

To analyze the differences between the conventional paper-based tool and the KatApp, we performed a mixed multivariate analysis of variance (MANOVA) with the triage tool as a within-subjects factor (paper-based tool vs KatApp). To account for potential interactions between the triage tool and tool order, we included tool order (KatApp used in the first session vs the second session) as a between-subjects factor. As dependent variables, we included the overall time taken to complete each triage session, the number of correct triages, the participants' subjective ratings of the triage tools using the German school grading system, and the results from the 6 scales of the UEQ.

In a separate repeated-measures MANOVA, we utilized the measurement time point (first vs second session) as a within-subjects factor for the same 9 dependent variables to evaluate differences between the 2 time points.

We reported the Pillai trace as the *F*-statistic in our models and conducted post hoc univariate ANOVAs for each dependent variable.

Despite some outcome variables violating the assumptions of normal distribution, we calculated MANOVAs, as this method is relatively robust against such violations to conduct MANOVAs instead of multiple ANOVAs because MANOVAs account for the complex relationships between dependent variables, offer greater statistical power, and help mitigate the issue of alpha error accumulation compared with multiple testing.

As an exploratory analysis, we calculated nonparametric correlations (Kendall tau) between age and the outcome variables to evaluate the suitability of including age as a covariate in our model. This assessment aimed to investigate the potential influences of age on the use of the app, as younger generations may be more accustomed to using smartphones and apps compared with older individuals. However, no statistically significant correlations were found (see Multimedia Appendix 7). Additionally, we chose not to include other demographic data, such as gender or profession, in our model, as they did not provide additional value concerning our study objectives and could further decrease statistical power, as previously discussed.

We established the significance level at α=.05 for our analyses.

We created raincloud plots in R (R Foundation) using the raincloudplots package [32]. This approach visualizes differences between the 2 triage tools, incorporating raw data, probability density, and key summary statistics in a transparent format.

The study was conducted in accordance with the Template for Intervention Description and Replication (TIDieR) checklist and guideline recommendations. The completed checklist is available as Multimedia Appendix 8.

Results

Sample Size and Exclusions

In total, 55 individuals initially registered for the course, resulting in 48 participants who attended, which reflects a no-show rate of 13% (7/55). Among those who attended, 6 participants were excluded from the statistical analyses due to incomplete triage documentation (n=3 with the paper-based tool, n=2 with the KatApp, and n=1 with both tools). Additionally, 2 participants did not complete questionnaire 1 and 2 did not complete questionnaire 2. Consequently, the final sample size for further statistical analyses comprised 38 participants.

Sample Characteristics

Demographic characteristics of the sample are presented in Table 2. The average age of the participants was 32.5 (SD 9.2) years. Of the total 38 participants, 24 (63%) were male and 15

XSL•FO
RenderX

(39%) reported having previous experience with triaging. In terms of occupation, 14 (37%) were physicians, 17 (45%) worked as rescue service personnel, 4 (11%) were medical students with German Red Cross EMT qualifications, and 3 (8%) belonged to other groups (eg, standby service or honorary positions).

Table 2. Demographic characteristics of the sample.

Characteristics	Value
Age (years), mean (SD)	32.5 (9.2)
Gender, n/N (%)	
Female	14/38 (37)
Male	24/38 (63)
Previous experience, n/N (%)	
Yes	15/38 (39)
No	23/38 (61)
Occupation, n/N (%)	
Physician	14/38 (37)
Rescue service personnel	17/38 (45)
Medical students with EMT[a] qualifications	4/38 (11)
Other	3/38 (8)
Specialization (physicians), n/N(%)	
Anesthesiology	10/14 (71)
Surgery	2/14 (14)
Neurology	1/14 (7)
Pediatrics	1/14 (7)
Additional qualification in emergency medicine (physicians), n/N (%)	
Yes	11/14 (79)
No	3/14 (21)
Highest qualification (rescue service personnel), n/N (%)	
EMT-basic	12/17 (71)
EMT-intermediate	1/17 (6)
Paramedic	4/17 (24)

[a]EMT: Emergency Medical Technician.

Evaluation Results

Overall

Findings from the 2 (triage tool) × 2 (tool order) mixed MANOVA revealed a significant main effect for the within-subjects factor of the triage tool ($F_{9,28}$=70.585, P<.001, η^2_p=.958). Additionally, there was a significant interaction between the triage tool and the tool order ($F_{9,28}$=4.432, P=.001, η^2_p=.588); however, no significant main effect was found for the between-subjects factor of tool order ($F_{9,28}$=1.382, P=.24, η^2_p=.308).

Duration

The results of the post hoc univariate ANOVA indicated that participants were significantly faster when using the KatApp compared with the paper-based tool ($F_{1,36}$=229.769, P<.001, η^2_p=.865). The mean difference is illustrated in Figure 5A. In addition, a significant interaction was found between the triage tool and the tool order ($F_{1,36}$=15.052, P<.001, η^2_p=.295), suggesting that the time taken for each tool was lower in the second session, although the time was consistently less for the KatApp.

XSL•FO
RenderX

Figure 5. Results of the User Experience Questionnaire (UEQ) comparing the conventional paper-based tool with the KatApp regarding the 6 scales (A) Attractiveness, (B) Perspicuity, (C) Efficiency, (D) Dependability, (E) Stimulation, and (F) Novelty.

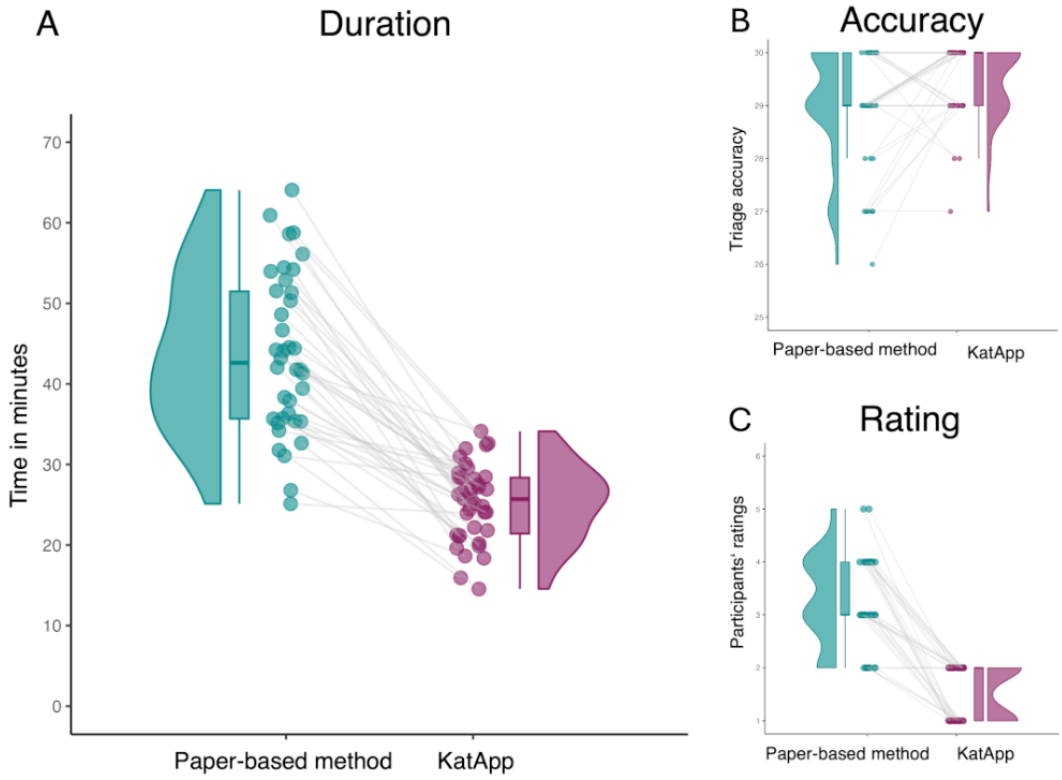

Accuracy

The post hoc univariate ANOVA also demonstrated that participants assigned a significantly greater number of patient cards to the correct category when using the KatApp compared with the paper-based tool ($F_{1,36}$=8.979, P=.005, η^2_p=.2), as illustrated in Figure 5B. Additionally, a significant interaction was observed between the triage tool and the tool order concerning triage quality ($F_{1,36}$=4.731, P=.04, η^2_p=.116). This interaction indicates that the number of correct triages was higher in the second session for both tools, with the KatApp consistently yielding a greater number of correct triages.

Subjective Ratings

When participants rated the triage tools using the German school grading system, the post hoc univariate ANOVA revealed significantly better ratings for the KatApp (rated as excellent) compared with the paper-based tool (rated as satisfactory; $F_{1,36}$=116.281, P<.001, η^2_p=.764). There was no significant interaction between the triage tool and the tool order ($F_{1,36}$=0.007, P=.93, η^2_p=.000). The mean difference is illustrated in Figure 5C.

User Experience Questionnaire

Regarding the 6 scales of the UEQ, post hoc univariate ANOVAs indicated significantly better ratings for the KatApp compared with the paper-based tool on each scale (Attractiveness: $F_{1,36}$=128.812, P<.001, η^2_p=.782; Efficiency: $F_{1,36}$=105.542, P<.001, η^2_p=.744; Perspicuity: $F_{1,36}$=26.390, P<.001, η^2_p=.423; Dependability: $F_{1,36}$=61.693, P<.001, η^2_p=.631; Stimulation: $F_{1,36}$=58.422, P<.001, η^2_p=.619; and Novelty: $F_{1,36}$=132.687, P<.001, η^2_p=.787). There was no significant interaction between the triage tool and the tool order for any of the UEQ scales (Attractiveness: $F_{1,36}$=0.214, P=.65, η^2_p=.006; Efficiency: $F_{1,36}$=0.006, P=.94, η^2_p=.000; Perspicuity: $F_{1,36}$=0.636, P=.43, η^2_p=.017; Dependability: $F_{1,36}$=1.762, P=.19, η^2_p=.047; Stimulation: $F_{1,36}$=0.434, P=.51, η^2_p=.012; Novelty: $F_{1,36}$=2.361, P=.13, η^2_p=.062). Figure 6 illustrates the differences between the tools regarding the 6 scales of the UEQ.

The mean values and corresponding statistics for the outcome measures between the 2 triage tools are presented in Table 3 and Multimedia Appendix 9 (also see [33]).

XSL•FO
RenderX

Figure 6. Comparison of the conventional paper-based tool with the KatApp regarding (A) duration to complete each session in minutes, (B) triage accuracy as number of correct triages out of 30, and (C) the participants' subjective ratings as German school grades from 1 (excellent) to 6 (insufficient). UEQ: User Experience Questionnaire.

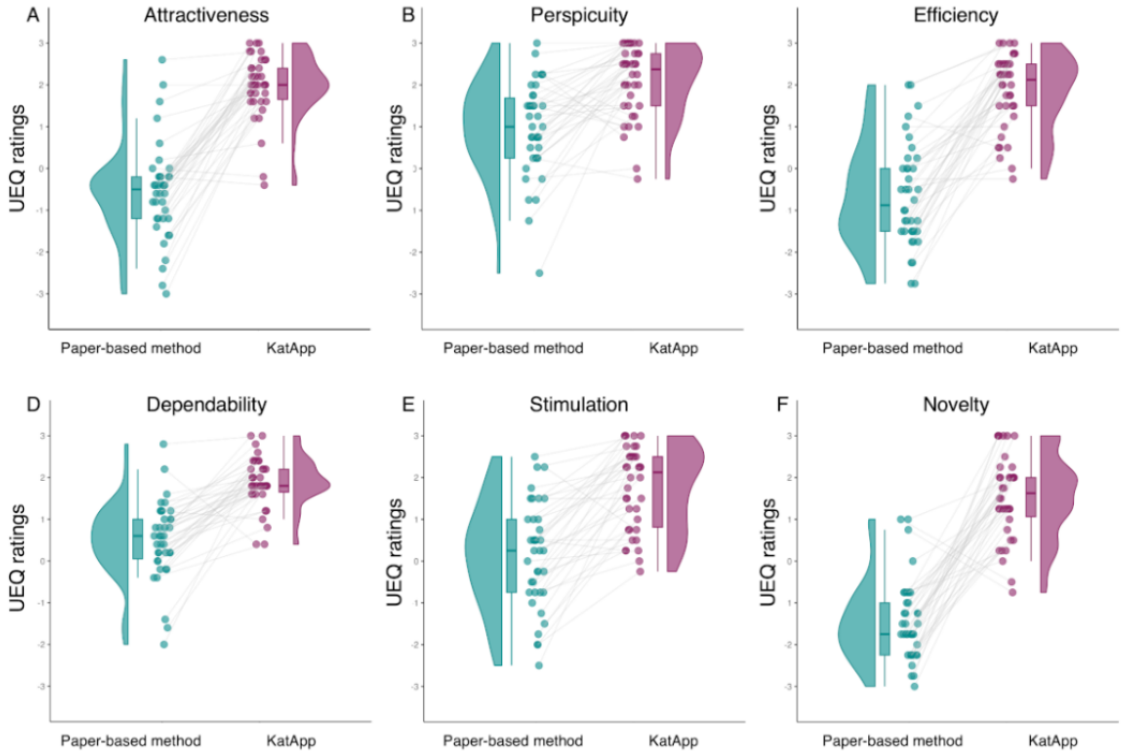

Table 3. Post hoc univariate analysis of variance comparison between the paper-based tool and the KatApp (N=38).

Outcome variable	Annotations in the paper-based tool, mean (SD)	Annotations in the KatApp, mean (SD)	*F* test (*df*)	*P* value	η^2_p
Duration (minutes)	43.71 (9.816)	25.22 (4.737)	229.769 (1, 36)	<.001	0.865
Triage accuracy	28.92 (1.099)	29.45 (0.724)	8.979 (1, 36)	.005	0.200
Subjective rating	3.26 (0.860)	1.53 (0.506)	116.281 (1, 36)	<.001	0.764
User Experience Questionnaire					
Attractiveness	–0.58 (1.175)	1.96 (0.764)	128.812 (1, 36)	<.001	.782
Efficiency	–0.65 (1.224)	1.86 (0.900)	104.542 (1, 36)	<.001	.744
Perspicuity	0.94 (1.146)	2.09 (0.861)	26.390 (1, 36)	<.001	.423
Dependability	0.49 (0.937)	1.88 (0.600)	61.693 (1, 36)	<.001	.631
Stimulation	0.14 (1.251)	1.76 (0.997)	58.422 (1, 36)	<.001	.619
Novelty	–1.45 (0.964)	1.54 (0.999)	132.687 (1, 36)	<.001	.787

Preference

When participants were asked to express their preference using a forced-choice item, an overwhelming majority (36/38, 95%) favored the KatApp, while the remaining 2 (5%) preferred the paper-based tool.

Controlling for Order Effects

The results of the 2 (triage tool) × 2 (tool order) mixed MANOVA indicated no significant main effect for the between-subjects factor of tool order ($F_{9,28}$=1.382, *P*=.24, η^2_p=.308). However, univariate ANOVAs for the between-subjects effect revealed no significant differences across the measures, except for the UEQ scale novelty ($F_{1,36}$=6.625, *P*=.01, η^2_p=.155). This finding suggests that when the KatApp was used in the second session, novelty ratings were significantly higher than when it was used first.

Additionally, the repeated measures MANOVA for outcome measures at each time point, with the session as a within-subjects factor, did not show a significant main effect for session ($F_{9,29}$=2.094, *P*=.06, η^2_p=.394). None of the post hoc analyses indicated significant differences between the first and second sessions for any of the outcome measures (Table 4).

XSL•FO
RenderX

Table 4. Post hoc univariate multivariate analysis of variance comparison between the results of the first and second sessions.

Compared variables	*F* test (*df*)	*P* value	η^2_p
Triage quality	3.357 (1, 37)	.07	.083
Time duration	3.357 (1, 37)	.26	.035
Subjective rating	0.104 (1, 37)	.75	.003
User Experience Questionnaire			
Attractiveness	0.253 (1, 37)	.62	.007
Efficiency	0.101 (1, 37)	.75	.003
Perspicuity	0.165 (1, 37)	.69	.004
Dependability	1.151 (1, 37)	.29	.030
Stimulation	0.443 (1, 37)	.51	.012
Novelty	0.190 (1, 37)	.66	.005

Discussion

Principal Findings

In this paper, we present an innovative approach to triage that replaces traditional paper-based classification with a mobile app–based system for categorizing patients in the context of MCIs. After prototyping the mobile app, we tested it in several training sessions and refined its features. The final prototype was then evaluated against the conventional paper-based tool regarding time duration, triage quality, and subjective ratings, such as user experience, using a within-subjects experimental study. Our results suggest that using this software system enables triage to be completed more quickly, easily, and accurately compared with the conventional paper-based tool. This approach addresses the time-consuming and error-prone issues associated with current procedures in disaster medicine.

The most significant and noticeable difference was the time participants required to complete the 2 triage sessions, each consisting of 30 patient cards. Using the KatApp, participants were, on average, over 18 minutes faster than when using the paper-based tool. Additionally, because command posts receive information in real time, critical decisions can be made sooner, allowing for faster patient treatment, which could ultimately save more lives. It is important to note that participants did not have to transmit their information via radio communication to the command post during the paper-based session, which would have added even more time. The requirement to fill out casualty cards manually and complete the documentation sheet likely contributed to this time difference.

Participants also found the KatApp to be more user-friendly than the conventional paper-based tool. In each of the 6 scales of the UEQ, participants rated the KatApp as “excellent” or at least “good” according to the corresponding benchmarks [33]. By contrast, the paper-based tool received ratings of “bad” or “below average.” This trend is reflected in the school grades awarded by participants: the KatApp was perceived as “excellent,” while the paper-based tool averaged a “satisfactory” grade. These results suggest that, in addition to the time advantage, the KatApp was perceived as easier to use than the paper-based tool. Furthermore, the significance of digitizing disaster medicine processes is underscored by the fact that 36 out of 38 (95%) participants indicated a preference for the KatApp over the traditional triage tool.

Surprisingly, there was also a statistically significant difference in triage accuracy between the 2 tools, indicating that participants triaged more patients correctly when using the KatApp compared with the paper-based tool. However, because the effect size for this difference was smaller than that of other measures, this result should be interpreted with caution. It is possible that the patient cards varied slightly, even though they were matched in terms of injury severity, injury pattern, and triage category. Additionally, the number of correctly assigned patient cards was high for both tools. However, some participants particularly valued the support provided by the app during triage, which may have contributed to fewer errors due to the automated application of the mSTART algorithm.

The results of our within-subjects experimental study support our hypotheses, demonstrating that the KatApp system is at least as useful and reliable as the conventional paper-based tool for triaging in disasters involving MCIs. The advantages of a digital triage tool—such as automatic forwarding of information to a shared data source and central dashboard, collection of real-time data, tracking of patients’ geolocations, and gathering of complex data to optimize rescue operations—are clearly evident. In addition to our quantitative results, we observed several additional advantages in our study. During the early implementations of the KatApp and in the experimental study, participants did not receive extensive training before using the system. Instead, they were given only a brief introduction right before the training began. This highlights the overall usability of the KatApp system.

Furthermore, qualitative feedback from the command posts after the training sessions confirmed that the information presented in the dashboard significantly simplified their work. In particular, the real-time overview of patients along with their triage categories, as well as the map displaying patient locations, was highly appreciated.

XSL•FO
RenderX

Next Steps and Further Adaptions

The next steps will include enhancing interconnectivity between smartphones when mobile networks are unavailable and incorporating emergency resources such as ambulances, helicopters, and hospitals into the backend. We also plan to conduct final tests in realistic scenarios, including the use of full personal protective equipment and simulated body fluids. Following these tests, we will initiate a pilot phase in real, nonsimulated settings with the goal of introducing the live system as quickly as possible.

Comparison With Prior Work

To our knowledge, this is the first paper addressing the challenges of current rescue operations in mass casualty disasters through the development, implementation, and experimental evaluation of a mobile triage app compared with a conventional paper-based tool. A systematic review of triage apps for health emergencies [34] indicates that there have been previous attempts to develop similar approaches. For example, in 2014, a research group developed a mobile-based system to support triage in MCIs, incorporating a Cox proportional hazard model to present patients' survival curves and aid in triage and transportation decisions [35]. Another approach involved using smart glasses along with a specific Android app to assist in triage during MCIs, which was evaluated in a randomized controlled simulation trial [36]. The TRIAGIST mobile app [37,38] was developed as a low-cost system that dispatch centers or rescue teams in Thailand could use for developmental purposes. However, none of these approaches have been further implemented for use in disaster medicine to date. Additionally, there has been a decline in studies focused on apps for catastrophe triage in recent years, and the few existing triage apps on the market that are based on scientific and academic research often lack free access [34]. In response to the call for an accessible app that provides patient geolocation and assists rescue teams by automatically applying a triage algorithm [34], our KatApp system aims to address a significant gap in the field. This contribution is of great practical importance for improving triage processes in disaster scenarios.

Limitations

In our study, we utilized patient cards instead of real patients or actors, which may limit the external validity of our findings. Our primary goal was to compare both triage tools within a standardized procedure and analyze a large number of triages, resulting in a total of 2280 triages. Involving 60 actors would have been more costly with only marginal additional benefit. The KatApp still needs to be evaluated in real disaster operations or large-scale simulations. Regarding the within-subjects design, we acknowledge that order effects, such as practice or fatigue, cannot be completely ruled out. To address this, we counterbalanced the tool order and included it as a between-subjects factor in our MANOVA. We also conducted a second repeated-measures MANOVA with measurement time points as a within-subjects factor. Our analyses revealed no significant influence of order on the primary outcomes. While practice effects were observed, participants consistently performed better in terms of time and accuracy with the KatApp. Despite these order effects, we chose a within-subjects design for its greater statistical power, better control over individual differences, and smaller sample size requirement. However, within-subjects designs may also make it easier for participants to infer the hypotheses, potentially distorting results due to heightened expectations. Finally, some assumptions for our statistical analyses, such as the normal distribution of variables, were partially violated. However, MANOVA is generally considered robust to such violations [39]. Although the overall missing data rate was high (10/48, 21%), we still reported large effect sizes in our results.

Conclusions

In this paper, we present research that addresses the weaknesses and challenges of the current triage process in MCIs through the development, implementation, and evaluation of a mobile system called KatApp.

We demonstrated that the KatApp is a mobile triage app designed for disaster medicine that not only matches the performance of the conventional paper-based tool but may even surpass it in terms of efficiency and usability. This advancement highlights the potential of digitalization to optimize processes in disaster medicine, ultimately leading to the possibility of saving more lives. While it is clear that this novel approach reduces time-consuming and error-prone manual tasks, we are currently unable to fully quantify the impact of the KatApp system on efficiency in real disaster situations, as it has only been implemented in simulated training scenarios thus far. Further research and refinement of this approach are needed to address this gap.

Acknowledgments

As the development of the KatApp system was a purely academic project, several information technology students contributed to the development of the KatApp system. We would like to mention the following: Ralf Arnold, Marvin Bermel, Niklas Digeser, Nico Döbele, Hai Yen Dang, Berend Fuchs, Diana Gilmanov, Laurenz Guth, Alexander Heinz, David Hötten, Jarne Jost, Nico Kantas, Hannes Kececi, Marvin Kotzian, Jona Krumrein, Fabian Just, Jan Meckelholt, Pascal Nerger, Sait Ok, Michael Pfeiffer, Ansam Qasem, Benjamin Reiser, Philipp Riedel, Hannes Wagner, and Tim Weiß. We also thank Dominik Neumann for cosupervising the students and the German Red Cross for hosting us at their center for the experiment. To improve language, we used Perplexity AI (Perplexity AI, Inc.) as generative artificial intelligence to rephrase and translate text passages. We acknowledge the support from the Open Access Publication Fund of the University of Tübingen.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request. The software code for the KatApp cannot be made publicly available. However, interested researchers may inquire about specific aspects of the system's functionality and architecture. We are committed to ensuring that the data supporting the findings of this study are accessible to the scientific community.

Authors' Contributions

RW, ES, and M Schmollinger initiated and coordinated the study. RW and JG coordinated the experiment, including data collection. JG and RW interpreted the data. M Schmollinger and JG wrote the initial draft of the manuscript. PR, AM, BB, and M Sigle critically reviewed and substantially improved the manuscript. All authors reviewed and approved the final version of the manuscript before submission. All authors had access to all data and analysis in the study. All authors had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Experiment overview picture.
[PNG File , 1675 KB - jmir_v26i1e65728_app1.png]

Multimedia Appendix 2
Experiment overview film.
[MOV File , 39934 KB - jmir_v26i1e65728_app2.mov]

Multimedia Appendix 3
Patient cards.
[PDF File (Adobe PDF File), 593 KB - jmir_v26i1e65728_app3.pdf]

Multimedia Appendix 4
Paper-based casualty cards.
[PDF File (Adobe PDF File), 749 KB - jmir_v26i1e65728_app4.pdf]

Multimedia Appendix 5
KatApp mock-ups.
[PDF File (Adobe PDF File), 1066 KB - jmir_v26i1e65728_app5.pdf]

Multimedia Appendix 6
German version of the User Experience Questionnaire as used in our study with English translations.
[DOCX File , 16 KB - jmir_v26i1e65728_app6.docx]

Multimedia Appendix 7
Kendall tau correlations between age and the outcome variables to check the conditions for a multivariate analysis of covariance.
[PDF File (Adobe PDF File), 258 KB - jmir_v26i1e65728_app7.pdf]

Multimedia Appendix 8
TIDieR (Template for Intervention Description and Replication) checklist.
[PDF File (Adobe PDF File), 53 KB - jmir_v26i1e65728_app8.pdf]

Multimedia Appendix 9
Post hoc univariate analysis of variance comparison between the paper-based tool and the KatApp (N=38).
[DOCX File , 16 KB - jmir_v26i1e65728_app9.docx]

References

XSL•FO
RenderX

1. Centre for Research on the Epidemiology of Disasters (CRED). 2022 Disasters in numbers. CRED. Brussels, Belgium: Institute Health and Society – UCLouvain; 2023. URL: https://cred.be/sites/default/files/2022_EMDAT_report.pdf [accessed 2024-08-06]
2. Sever MS, Remuzzi G, Vanholder R. Disaster medicine and response: optimizing life-saving potential. Am J Disaster Med 2018;13(4):253-264. [doi: 10.5055/ajdm.2018.0305] [Medline: 30821339]
3. Cuthbertson J, Archer F, Robertson A, Rodriguez-Llanes JM. Improving disaster data systems to inform disaster risk reduction and resilience building in Australia: a comparison of databases. Prehosp Disaster Med 2021 Sep 16;36(5):511-518. [doi: 10.1017/s1049023x2100073x]
4. Cheng T, Staats K, Kaji AH, D'Arcy N, Niknam K, Donofrio-Odmann JJ. Comparison of prehospital professional accuracy, speed, and interrater reliability of six pediatric triage algorithms. J Am Coll Emerg Physicians Open 2022 Feb;3(1):e12613 [FREE Full text] [doi: 10.1002/emp2.12613] [Medline: 35059689]
5. Franc JM, Kirkland SW, Wisnesky UD, Campbell S, Rowe BH. METASTART: a systematic review and meta-analysis of the diagnostic accuracy of the Simple Triage and Rapid Treatment (START) algorithm for disaster triage. Prehosp Disaster Med 2022 Feb;37(1):106-116. [doi: 10.1017/S1049023X2100131X] [Medline: 34915954]
6. Kanz KG, Hornburger P, Kay MV, Mutschler W, Schäuble W. mSTaRT-algorithmus für sichtung, behandlung und transport bei einem massenanfall von verletzten. Notfall Rettungsmed 2006 May;9(3):264-270. [doi: 10.1007/s10049-006-0821-3]
7. Romig LE. Pediatric triage. A system to JumpSTART your triage of young patients at MCIs. JEMS 2002 Jul;27(7):52-8, 60. [Medline: 12141119]
8. Smith M. Get smart: jumpSTART!. Emerg Med Serv 2001 May;30(5):46-8, 50. [Medline: 11373910]
9. Spöttl P, Ziegler A. Triage: prioritization of injured after a major accident or mass casualty incident. Med Klin Intensivmed Notfmed 2015 Feb 3;110(1):9-14. [doi: 10.1007/s00063-014-0410-z] [Medline: 25451766]
10. Dirks B. Management des massenanfalls von verletzten/erkrankten durch den leitenden notarzt. Notfall Rettungsmed 2006 May;9(3):333-346. [doi: 10.1007/s10049-006-0814-2]
11. Rauch L, Dotzer M, Händl T. [Management of a mass casualty incident according to the recommendations of the Federal Office of Civil Protection and Disaster Assistance and the principles of the Terror and Disaster Surgical Care Concept in a regional trauma center]. Anaesthesiologie 2024 Mar 27;73(3):193-203. [doi: 10.1007/s00101-024-01384-7] [Medline: 38413414]
12. Auf der Heide E. The importance of evidence-based disaster planning. Ann Emerg Med 2006 Jan;47(1):34-49. [doi: 10.1016/j.annemergmed.2005.05.009] [Medline: 16387217]
13. De Cauwer H, Barten D, Willems M, Van der Mieren G, Somville F. Communication failure in the prehospital response to major terrorist attacks: lessons learned and future directions. Eur J Trauma Emerg Surg 2023 Aug 10;49(4):1741-1750. [doi: 10.1007/s00068-022-02131-6] [Medline: 36214838]
14. Sadat SJ, Afrasiabifar A, Khorasani-Zavarehg D, Moradian MJ, Vafaeenasab M, Dehghani Tafti A, et al. Exploring barriers and facilitators of inter-organizational management in response to mass casualty traffic incidents: a qualitative study. Bull Emerg Trauma 2021 Apr;9(2):86-95 [FREE Full text] [doi: 10.30476/BEAT.2021.89416] [Medline: 34150919]
15. Bania A, Iatrellis O, Samaras N. Information communication technologies (ICTs) and disaster risk management (DRM): systematic literature review. Berlin, Germany: Springer; 2023 Presented at: CSUM 2022; August 31-September 2, 2022; Skiathos Island, Greece p. 1779-1794 URL: https://link.springer.com/book/10.1007/978-3-031-23721-8 [doi: 10.1007/978-3-031-23721-8_137]
16. Sakurai M, Murayama Y. Information technologies and disaster management – benefits and issues. Progress in Disaster Science 2019 Jul;2:100012. [doi: 10.1016/j.pdisas.2019.100012]
17. Davis A. Software prototyping. Advances in Computers 1995;40:39-63. [doi: 10.1016/S0065-2458(08)60544-6]
18. Białkowski D, Smołka J. Evaluation of Flutter framework time efficiency in context of user interface tasks. jcsi 2022 Dec 30;25:309-314. [doi: 10.35784/jcsi.3007]
19. Karić A, Durmić N. Comparison of JavaScript frontend frameworks – Angular, React, and Vue. International Journal of Innovative Science and Research Technology (IJISRT) 2024 Jul 1:1383-1390. [doi: 10.38124/ijisrt/ijisrt24jun600]
20. CACM Staff. React: Facebook's functional turn on writing JavaScript. Commun ACM 2016 Dec;59(12):56-62. [doi: 10.1145/2980991]
21. Herfort B, Lautenbach S, Porto de Albuquerque J, Anderson J, Zipf A. The evolution of humanitarian mapping within the OpenStreetMap community. Sci Rep 2021 Feb 04;11(1):3037 [FREE Full text] [doi: 10.1038/s41598-021-82404-z] [Medline: 33542423]
22. Castro P, Ishakian V, Muthusamy V, Slominski A. The rise of serverless computing. Commun ACM 2019 Nov 21;62(12):44-54. [doi: 10.1145/3368454]
23. Aluvala R, Maheswari V U. Serverless Computing Concepts, Technology and Architecture. Hershey, PA: IGI Global; Apr 2024:1-19.
24. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behavior Research Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/brm.41.4.1149]

XSL•FO
RenderX

25. Sigle M, Berliner L, Richter E, van IM, Gorgati E, Hubloue I. Artificial patient database used for the publication "Development of an anticipatory triage-ranking algorithm using dynamic simulation of the expected time course of patients with trauma: modeling and simulation study". Zenodo. 2022. URL: https://zenodo.org/records/7134900 [accessed 2024-08-06]
26. Sigle M, Berliner L, Richter E, van Iersel M, Gorgati E, Hubloue I, et al. Development of an anticipatory triage-ranking algorithm using dynamic simulation of the expected time course of patients with trauma: modeling and simulation study. J Med Internet Res 2023 Jun 15;25:e44042 [FREE Full text] [doi: 10.2196/44042] [Medline: 37318826]
27. Heller AR, Neidel T, Klotz PJ, Solarek A, Kowalzik B, Juncken K, et al. Validation of secondary triage algorithms for mass casualty incidents-a simulation-based study-German version. Anaesthesiologie 2023 Jul;72(7):467-476 [FREE Full text] [doi: 10.1007/s00101-023-01291-3] [Medline: 37318526]
28. Kinsman T, Wessel M, Gerosa M, Treude C. How do software developers use GitHub actions to automate their workflows? 2021 Presented at: The 2021 Mining Software Repositories Conference; May 17-19, 2021; Virtual URL: https://arxiv.org/abs/2103.12224 [doi: 10.1109/msr52588.2021.00054]
29. Adzic G, Chatley R. Serverless computing: economic and architectural impact. New York City, NY: Association for Computing Machinery; 2017 Presented at: ESEC/FSE'17: Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering; September 4-8, 2017; Paderborn, Germany. [doi: 10.1145/3106237.3117767]
30. Rotem-Gal-Oz A. Fallacies of distributed computing explained. ResearchGate. 2008. URL: https://www.researchgate.net/publication/322500050_Fallacies_of_Distributed_Computing_Explained [accessed 2024-08-06]
31. Laugwitz B, Held T, Schrepp M. Construction and evaluation of a user experience questionnaire. In: HCI and Usability for Education and Work. Berlin, Heidelberg: Springer; 2008:66-76.
32. Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 2019;4:63 [FREE Full text] [doi: 10.12688/wellcomeopenres.15191.2] [Medline: 31069261]
33. Schrepp M, Hinderks A, Thomaschewski J. Construction of a benchmark for the User Experience Questionnaire (UEQ). IJIMAI 2017;4(4):40. [doi: 10.9781/ijimai.2017.445]
34. Montano IH, de la Torre Díez I, López-Izquierdo R, Villamor MAC, Martín-Rodríguez F. Mobile triage applications: a systematic review in literature and play store. J Med Syst 2021 Aug 13;45(9):86 [FREE Full text] [doi: 10.1007/s10916-021-01763-2] [Medline: 34387773]
35. Tian Y, Zhou T, Wang Y, Zhang M, Li J. Design and development of a mobile-based system for supporting emergency triage decision making. J Med Syst 2014 Jun;38(6):65. [doi: 10.1007/s10916-014-0065-6] [Medline: 24898115]
36. Follmann A, Ohligs M, Hochhausen N, Beckers SK, Rossaint R, Czaplik M. Technical support by smart glasses during a mass casualty incident: a randomized controlled simulation trial on technically assisted triage and telemedical app use in disaster medicine. J Med Internet Res 2019 Jan 03;21(1):e11939 [FREE Full text] [doi: 10.2196/11939] [Medline: 30609988]
37. Sutham K, Khuwuthyakorn P, Thinnukool O. Thailand medical mobile application for patients triage base on criteria based dispatch protocol. BMC Med Inform Decis Mak 2020 Apr 09;20(1):66 [FREE Full text] [doi: 10.1186/s12911-020-1075-6] [Medline: 32272928]
38. Nimmolrat A, Sutham K, Thinnukool O. Patient triage system for supporting the operation of dispatch centres and rescue teams. BMC Med Inform Decis Mak 2021 Feb 19;21(1):68 [FREE Full text] [doi: 10.1186/s12911-021-01440-x] [Medline: 33608001]
39. Finch H. Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. Methodology 2005 Jan;1(1):27-38. [doi: 10.1027/1614-1881.1.1.27]

Abbreviations

AWS: Amazon Web Services
EMT: Emergency Medical Technician
MANOVA: multivariate analysis of variance
MCI: mass casualty incident
mSTART: modified Simple Triage and Rapid Treatment
REST: Representational State Transfer
START: Simple Triage and Rapid Treatment
TIDieR: Template for Intervention Description and Replication
UEQ: User Experience Questionnaire

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.08.24; peer-reviewed by J Murphy; comments to author 23.09.24; revised version received 30.09.24; accepted 08.10.24; published 21.11.24.

Please cite as:
Schmollinger M, Gerstner J, Stricker E, Muench A, Breckwoldt B, Sigle M, Rosenberger P, Wunderlich R
Evaluation of an App-Based Mobile Triage System for Mass Casualty Incidents: Within-Subjects Experimental Study
J Med Internet Res 2024;26:e65728
URL: https://www.jmir.org/2024/1/e65728
doi:10.2196/65728
PMID:39474975

©Martin Schmollinger, Jessica Gerstner, Eric Stricker, Alexander Muench, Benjamin Breckwoldt, Manuel Sigle, Peter Rosenberger, Robert Wunderlich. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reliable Web-Based Auditory Cognitive Testing: Observational Study

Meher Lad[1], MBBS, PhD; John-Paul Taylor[1,2], PhD; Timothy David Griffiths[3,4], MD

[1]Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, United Kingdom
[2]NIHR Newcastle Biomedical Research Centre, Newcastle University, Newcastle upon Tyne, United Kingdom
[3]Biosciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom
[4]Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom

Corresponding Author:
Meher Lad, MBBS, PhD
Translational and Clinical Research Institute
Newcastle University
Auditory Laboratory
Faculty of Medical Sciences, Newcastle University
Newcastle upon Tyne, NE2 4HH
United Kingdom
Phone: 44 0191 208 6000
Email: meher.lad@newcastle.ac.uk

Abstract

Background: Web-based experimentation, accelerated by the COVID-19 pandemic, has enabled large-scale participant recruitment and data collection. Auditory testing on the web has shown promise but faces challenges such as uncontrolled environments and verifying headphone use. Prior studies have successfully replicated auditory experiments but often involved younger participants, limiting the generalizability to older adults with varying hearing abilities. This study explores the feasibility of conducting reliable auditory cognitive testing using a web-based platform, especially among older adults.

Objective: This study aims to determine whether demographic factors such as age and hearing status influence participation in web-based auditory cognitive experiments and to assess the reproducibility of auditory cognitive measures—specifically speech-in-noise perception and auditory memory (AuM)—between in-person and web-based settings. Additionally, this study aims to examine the relationship between musical sophistication, measured by the Goldsmiths Musical Sophistication Index (GMSI), and auditory cognitive measures across different testing environments.

Methods: A total of 153 participants aged 50 to 86 years were recruited from local registries and memory clinics; 58 of these returned for web-based, follow-up assessments. An additional 89 participants from the PREVENT cohort were included in the web-based study, forming a combined sample. Participants completed speech-in-noise perception tasks (Digits-in-Noise and Speech-in-Babble), AuM tests for frequency and amplitude modulation rate, and the GMSI questionnaire. In-person testing was conducted in a soundproof room with standardized equipment, while web-based tests required participants to use headphones in a quiet room via a web-based app. The reproducibility of auditory measures was evaluated using Pearson and intraclass correlation coefficients, and statistical analyses assessed relationships between variables across settings.

Results: Older participants and those with severe hearing loss were underrepresented in the web-based follow-up. The GMSI questionnaire demonstrated the highest reproducibility (r=0.82), while auditory cognitive tasks showed moderate reproducibility (Digits-in-Noise and Speech-in-Babble r=0.55 AuM tests for frequency r=0.75 and amplitude modulation rate r=0.44). There were no significant differences in the correlation between age and auditory measures across in-person and web-based settings (all P>.05). The study replicated previously reported associations between AuM and GMSI scores, as well as sentence-in-noise perception, indicating consistency across testing environments.

Conclusions: Web-based auditory cognitive testing is feasible and yields results comparable to in-person testing, especially for questionnaire-based measures like the GMSI. While auditory tasks demonstrated moderate reproducibility, the consistent replication of key associations suggests that web-based testing is a viable alternative for auditory cognition research. However, the underrepresentation of older adults and those with severe hearing loss highlights a need to address barriers to web-based participation. Future work should explore methods to enhance inclusivity, such as remote guided testing, and address factors like digital literacy and equipment standards to improve the representativeness and quality of web-based auditory research.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58444) doi:10.2196/58444

KEYWORDS

auditory testing; hearing loss; cognitive testing; auditory; observational study; older adult; hearing; questionnaire; auditory cognitive testing; in person; web-based setting; auditory memory; Pearson; female; women; audiology

Introduction

Web-based experimentation has allowed researchers to reach a large number of participants in a short time. This has increased the statistical power of studies and the reliability of their findings. Web-based recruitment platforms like Amazon Mechanical Turk and Prolific.co have become common venues for data collection, and stimulus creation and presentation are simplified through mediums like Gorilla, PsychoPy, and Qualtrics [1,2]. This has resulted in a shift in research culture, where an increasing number of laboratories are trying to answer their research questions by using these resources. The COVID-19 pandemic accelerated this process [3].

Web-based auditory testing has also increased during this time, and a range of studies have evidenced the reliability and replicability of this method [4,5]. Auditory experiments rely on well-controlled delivery of sounds with headphone usage, as the variability of background noise during the sound presentation, equipment, and choice of device used for presenting sounds can dramatically impact the user experience. For example, in a web-based setting, it is not easy to verify whether a participant is using headphones, which attenuate background noise and ensure accurate delivery of sound requiring 2 audio channels, without visual verification. This introduces concerns regarding privacy. Many tests have been devised to overcome this problem [6]. This has led to a number of successful replications of auditory perceptual experiments [5,7]. However, a number of these studies have been conducted on young, motivated individuals who have volunteered for such experiments. People with severe hearing loss who may depend on hearing aids may not be able to physically accommodate headphones easily, as listening without hearing aids may need to compensate for individual hearing impairments [8].

Some researchers have expressed concerns about the population being sampled from web-based databases [9]. Participants tend to be less representative of local populations. In-person research has a lack of diversity in participation where people with limited literacy skills are often excluded from health research, and this may be exacerbated when digital literacy is taken into consideration [10]. Older adults may be further marginalized, although it has been shown that web-based cognitive testing has been successfully implemented with adults older than 65 years of age [11]. Nevertheless, the effectiveness of web-based testing on older adults with hearing difficulty is unclear.

There is a paucity of auditory cognitive research, using stimuli assessing speech-in-noise (SIN) hearing, for example, in older populations with a range of hearing abilities. Previous work has used verbal stimuli, which are easier to perceive than simple or complex auditory stimuli below the level of speech [12]. We have learned that participants are able to complete demanding tasks on a web-based platform with a particular experimental setup [13]. Learning effects of complex sound stimuli have also been studied in young and older adults recruited from Prolific.co successfully [14]. A combination of such stimuli with verbal stimuli has been studied in a similar participant group [15]. Whether reliable, web-based, auditory cognitive testing can be performed, using similar stimuli, in a “real-world” population has not been previously attempted to our knowledge.

This first part of this exploratory study aimed to assess if there were population characteristics, such as age and hearing status, that were more likely to lead to web-based experimentation. We then assessed the reproducibility of auditory cognitive metrics in the participants who participated in person and on the web. We asked participants to complete 2 verbal SIN perception tasks, for digits and sentences, and 2 auditory memory (AuM) tasks, for different sound features. We compared whether participants scored similarly on a self-reported questionnaire about musical sophistication, the Goldsmiths Musical Sophistication Index (GMSI), to scores obtained using computerized behavioral experiments for auditory cognition such as SIN measures and AuM [16]. We hypothesized better reproducibility of the questionnaire compared to the auditory cognitive tasks. We also assessed if there were differences in correlation coefficients between auditory measures and those that were shared across the 2 task settings, such as age and hearing thresholds.

For the second part of the study, we tested relationships between the auditory cognitive metrics that were identified previously by the in-person experimentation and were reproducible in a web-based group of participants. This included the participants from the first part and adults from the PREVENT study, a cohort of participants who were recruited to study risk factors for dementia [17]. We wished to examine whether web-based auditory cognitive experiments would reproduce findings that we had found between the GMSI and AuM for frequency precision and between SIN thresholds for sentence-in-babble perception and AuM [18,19]. During these in-person experiments, stricter auditory experimental procedures were used, and we expected participants to be more consistent, in this context, across in-person and web-based experiments compared to the web-based computerized tests.

Methods

Participants

A total of 153 older adults were recruited (from 300 invitations, equaling a response rate of 51%) from a local Newcastle University volunteer database, the Join Dementia Research registry, and from friends and family of people attending memory clinics for a dementia workup. Phone calls and emails were the methods of invitation. Inclusion criteria included the absence of a neurological or psychiatric condition at the time

XSL•FO
RenderX

of participation and age of 50 years or older. Lack of fluency in English was the only exclusion criterion.

The age range of participants was 50-86 years with a mean of 67 (SD 10) years. A total of 23 participants were active hearing aid users. All participants agreed to take part in a web-based, follow-up assessment between 3 and 6 months after their initial in-person assessment. Participants who conducted the web-based assessments did not report any subjective change in their hearing abilities or any new medical conditions affecting their cognition. In total, 58 (37.9% of the 153 in-person attendees) of these participants took part in a web-based repeat assessment. The performance of these participants was used to determine the reproducibility of auditory cognitive tests in person versus in a web-based setting. We aimed to recruit as many participants as possible, and therefore, no sample size calculation was performed.

A total of 89 additional participants (from 500 invitations, equaling a response rate of 18%) were invited from another participant cohort, the PREVENT study. This study was originally developed to establish midlife risk factors for dementia in a multicenter UK cohort of participants. The aim of the web-based study was to test whether previously published findings could be replicated on the web. These were studies showing an association between AuM and sentence-in-noise perception ability and between AuM for frequency precision and musical sophistication [17-19]. The Newcastle and the PREVENT cohort were combined to provide a dataset to analyze web-based auditory cognitive metrics. Participants were compensated with US $15 shopping vouchers for in-person testing and US $7 vouchers for web-based experimentation.

The mean age of participants in this group was 65 (SD 8) years. Of the 153 participants, 56.8% (n=87) were women, 5.2% (n=8) of participants left school at the age of 15 years, 69.9% (n=107) completed school and had a university degree, and 24.2% (n=37) had multiple degrees. In total, 11.1% (n=17) had normal hearing, 18.9% (n=87) had mild hearing loss, 54.2% (n=83) had moderate hearing loss, and 15.7% (n=24) had severe hearing loss. As shown in Figure 1, there were no participants older than 75 years of age who participated on the web. Only 4 (11%) out of 37 hearing aid users took part in the web-based experimentation session. Due to the small number of participants present in this group, further subgroup analysis was not conducted. In the study of web-based participants, across the Newcastle and PREVENT cohorts, the average age was 59 (SD 7.7) years. The sample included 75 female and 47 male participants. As these participants did not have pure-tone audiometry performed, we were unable to grade the severity of hearing loss like for the in-person participants.

Figure 1. Age in the in-person (n=153; green) versus web-based (n=58; yellow) cohort of participants. People older than 70 years of age were less likely to participate in web-based auditory cognitive experimentation.

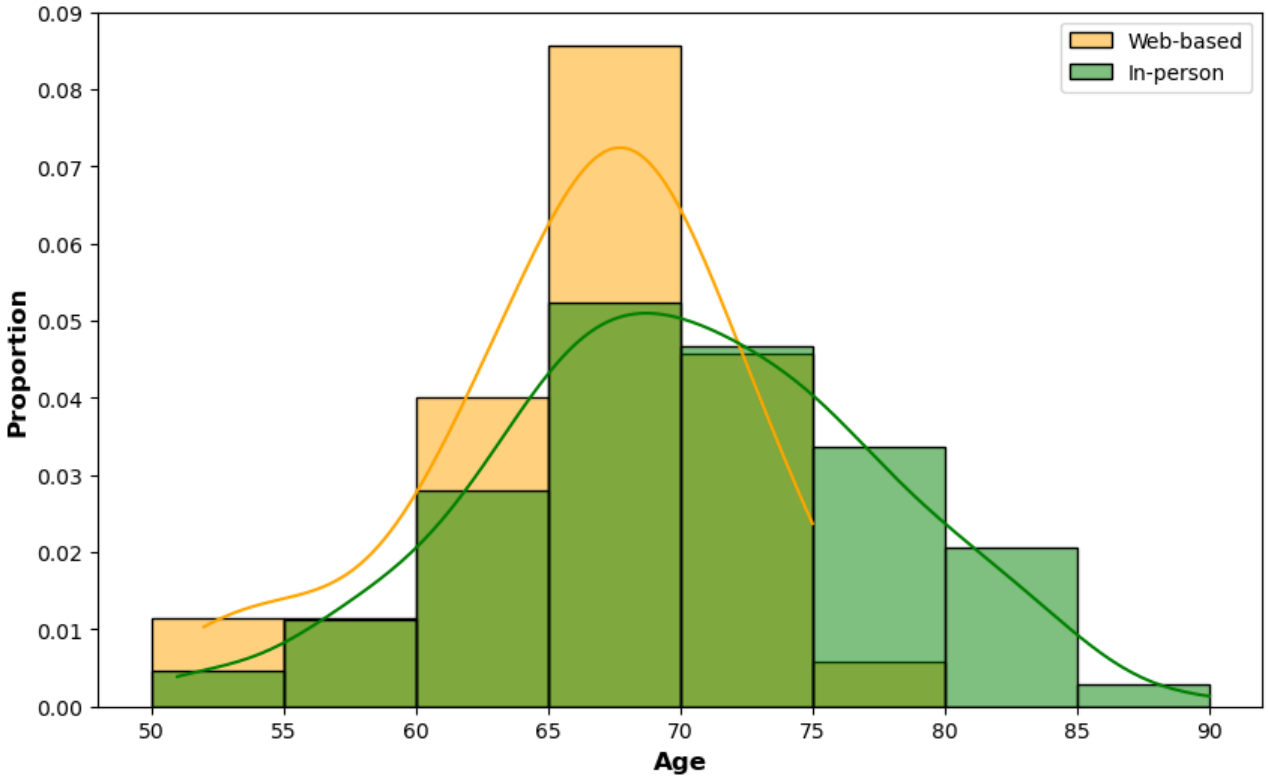

Stimuli and Equipment

In-person testing was conducted in a soundproof room using a Dell desktop computer with Sennheiser HD 201 circumaural headphones. An external sound card was used to process and deliver auditory stimuli, and acoustic stimuli were presented at 70 dBA. All auditory cognitive tasks and the GMSI questionnaire, both of which are described in detail in the next section, were coded in Javascript and displayed with a Google Chrome web browser. The web page was designed as a fully automated single-page application, where all stimulus generation and user interaction were determined by processes that occurred

XSL•FO
RenderX

"client-side." The web page was hosted by Google Cloud Platform using Firebase Hosting. A Firebase Realtime Database was used to store anonymized results that were linked to each participant ID, which they used to access the testing portal at home at the follow-up assessment. At home, participants were instructed to use a desktop, laptop, or tablet device with headphones and perform the test in a quiet room with no distractions. Each participant was sent a link that directed them to the web-based testing platform. Hearing aid users were instructed to use over-ear headphones that do not interfere with the aids, if available, or to increase the computer device's sound to a comfortable level before beginning the auditory tests.

Experimental Procedure

Each participant had a 1-hour visit to the Auditory Laboratory at the Newcastle University Medical School. Pure tone audiometry (PTA) testing was performed on both ears from 250 Hz to 8 kHz at octave intervals for air conduction using an Interacoustics AS608e screening audiometer. Tones were manually presented as short bursts twice starting at 30 dB hearing level (HL) then increased in 5 dB increments until comfortably audible if necessary. Then 5 dB HL reductions were made until the tone was not audible. This process was repeated twice, and the lowest audible volume was chosen as the value for a particular frequency. If maximum amplification at 100 dB HL could not be perceived then this was used as the ceiling value at a particular frequency. The overall mean of high-frequency values between 4 and 8 kHz for the best ear was taken as the threshold value for an individual for further analysis. This value was chosen as high-frequency thresholds are suspected to deteriorate first in age-related hearing loss and previous research from our group has suggested that PTA thresholds in this range correlate with SIN difficulties [20,21]. Hearing status was determined as normal if the mean threshold was below 20 dB HL, mild if between 20 and 40 dB HL, moderate if 40-60 dB HL hearing loss, and severe if >80 dB HL.

The Digits-in-Noise (DIN) task involved participants listening to 3 digits on a background of speech-shaped white noise, created using white noise using the long-term average speech spectrum of speech stimuli between 80 and 10,000 Hz. Participants had 2 practice trials at the beginning of the task to familiarize themselves with the stimuli at a signal-to-noise ratio (SNR) of 10 dB. An adaptive 1-up, 1-down psychophysical paradigm was implemented, whereby a correct response resulted in the SNR being reduced and an incorrect one caused the SNR to increase. The starting SNR was 0 dB and the step sizes decreased from 5 to 2 dB after 3 reversals, which then reduced to 0.5 dB after 3 more reversals. The run terminated after 10 reversals and the SNR at the last 5 reversals was averaged to calculate the DIN threshold for each participant. Lower SNR values indicated a better performance. As with the DIN task, participants had 2 practice trials at the beginning of the task to familiarize themselves with the stimuli at an SNR of 10 dB. The Speech-in-Babble (SIB) task consisted of participants listening to sentences on a background of 16-talker babble as described previously by our research group [18,21]. Target sentences had the form <name> <verb> <number> <adjective> <noun> (eg, "Alan gives four pretty flowers") and participants had to click on the correct word from a list of 5 columns (10 options for each word) shown on the screen with the same structure. The SIB threshold was determined using the same adaptive threshold procedure used in the DIN test. The starting parameters and adaptive design were exactly the same as the DIN task.

AuM was tested using nonspeech stimuli as previously described (Figure 2; Lad et al [19]). A 1-second tone or amplitude-modulated white noise stimulus was presented to a participant, after which they were asked to "find" the sound on a horizontal scale on a computer screen. Participants had to move a mouse and click on the line to produce a sound at that location. They could make as many clicks as they wanted with no set time limit. After they were satisfied with their choice, they would advance to the next trial by pressing the "Enter" key on a keyboard. Frequencies that determined the pure-tone sounds were chosen from a uniform distribution between 440 and 880 Hz, and amplitude modulation (AM) rates for the white noise stimulus were 5-20 Hz with a sinusoidal function used to apply this modulation. This parameter space, with an addition of 10% at either end of the scale, for each sound feature was mapped to the pixels of the horizontal scale on the screen during the matching phase. Hanning windows were applied to all synthetic sounds to avoid clicks and the beginning and end of the stimuli. The task consisted of 32 trials with the frequency and AM rate matching trials being interleaved. Participants had a short break after 16 trials. A Gaussian function was used to estimate the SD of the errors in each trial across the whole experiment and the inverse of this value, the precision, was used for further analysis. Thus, one obtains a precision for frequency AuM ("AuM (F)") and AM rate AuM ("AuM (A)"). Studies in vision have found that this measure better reflects the memory resource a participant can allocate in a given task [22]. Participants had 2 practice trials with each stimulus—2 for AuM (F) and 2 for AuM (A)—at the beginning of the task to familiarize themselves with the stimuli.

XSL•FO
RenderX

Figure 2. Auditory memory experiment. An auditory (pure tone or amplitude-modulated noise) stimulus is presented for 1 second; then, after a delay of 2 to 4 seconds, participants can match sounds using a horizontal scale on the screen. The scale is linked to the parameter of interest (frequency for pure tone or AM rate) that can generate the original stimulus after exploring the parameter space to "find" the stimulus. The figure shows an auditory matching trial where the participant's "final match" (rightmost dark gray marker on the scale) is shown in comparison to where the original stimulus (orange marker on the scale) is actually located. In this example, the participant first clicked on the scale to make a "first match" (which produced a sound linked to the parameter at that location), then a "second match," and then a "final match." The discrepancy between the "final match" location parameter and that of the original stimulus gives an "error" for each trial that can be used to calculate the auditory working memory "precision," the inverse of the SD of errors from a trial target, for all auditory trials. AM: amplitude modulation.

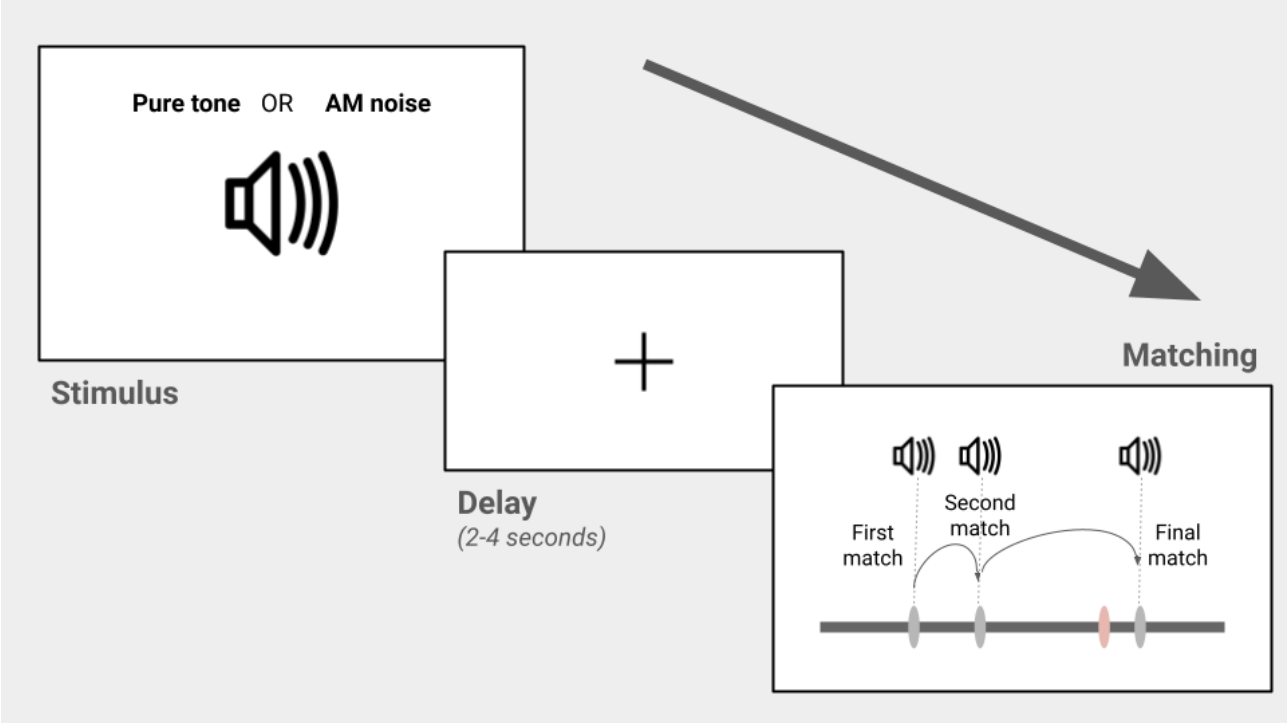

Finally, participants completed the short version of the GMSI questionnaire (Multimedia Appendix 1) consisting of 38 questions on paper as a test of musicality [16]. The GMSI is a self-report inventory that assesses individual differences in musical sophistication. It is said to measure the ability to engage with music in a comprehensive manner that does not just include musical instrumentation or training. For the purposes of this study, the GMSI was used as it has been previously used alongside the auditory cognitive measures in this study in an in-person setting [19]. We have previously shown a significant relationship between GMSI scores and AuM for frequency precision. This includes the total score of 266—the sum of the scores of the 5 domains of "Active Engagement," "Perceptual Abilities," "Musical Training," "Singing Abilities," and "Emotions." This questionnaire was used as a control of internal consistency between in-person and web-based experimentation, as we expected the change in testing modality to least affect a participant's ability to answer multiple-choice questions about their musical behaviors compared to the auditory tests.

Statistical Analysis

AuM scores were log-transformed to achieve normal distributions. All other variables were normally distributed and converted to *z* scores for further analysis. Any missing values were not imputed.

The chi-square tests were used to calculate the difference in proportions in age groups and hearing groups between in-person and web-based participants. Pearson correlation coefficients were then used to measure the strength and direction of the linear relationship between auditory metrics. To evaluate the reproducibility of web-based versus in-person testing across different age groups, we used correlation coefficients and intraclass correlation coefficients. Steiger test was used to assess whether the correlation coefficients between in-person and web-based auditory metrics with shared variables such as age and PTA thresholds were significantly different.

The relationship between 2 variables while controlling for the effect of another variable (eg, age) was performed after using linear regression to account for the influence of the control variable. This was used to assess the relationship between auditory cognitive metrics in person and then via a web-based platform. Finally, differences in correlation coefficients between these 2 sets of variables were calculated using bootstrapping. This was performed using repeated resampling of the data with replacement 1000 times. This allowed for the calculation of the mean difference and 95% CIs for the correlation differences. Significant differences between correlation coefficients were established if the CIs did not overlap. Due to the exploratory nature of this study, statistical correction for multiple corrections was not applied. All analyses were carried out using the Python (Python Software Foundation) programming language using Jupyter notebooks. *SciPy* and *Pingouin* packages were used.

Ethical Considerations

The Oxford C NHS Research Ethics Committee (21/SC/0139) approved this study on June 12, 2021. All participants gave

XSL•FO
RenderX

written consent to participate and publish data and for secondary analysis without additional consent. All study data are anonymous. In-person participants obtained US $15 vouchers and web-based participants received US $7 vouchers for participation. This study was not registered as it was not a randomized controlled trial.

Results

In-Person Versus Web-Based Auditory Cognitive Testing

Figure 2 shows the normalized age distribution of the in-person participant group and those who returned for web-based testing. The proportion of participants with different degrees of hearing loss is also shown. There was a significant difference in the proportion of participants older than 70 years of age between the web-based and in-person groups, (χ^2_1=9.40; *P*=.002), with a higher proportion of older participants in the in-person group. For the web-based participants, the proportion of those with moderate hearing loss was 54.1% (n=83), while the proportion with severe hearing loss was 16.2% (n=25). The mean hearing loss level was 45.84 (SD 18.84) dB HL for web-based participants and 52.32 (SD 23.03) dB HL for in-person participants. There was a statistically significant difference between the proportions of moderate and severe hearing loss categories among web-based participants (χ^2_1=6.02; *P*=.01). For the in-person participants, the proportion of those with moderate hearing loss was 33.5% (n=51), and the proportion with severe hearing loss was 35.4% (n=54). There was no statistically significant difference in the proportions of participants with moderate and severe hearing loss among in-person participants (χ^2_1=1.07; *P*=.30).

Table 1 shows the mean and SD values of the auditory cognitive tests in both settings. There were no statistical differences between the mean values between in-person and web-based auditory measures. We assessed the reproducibility of various auditory and cognitive measures conducted in both laboratory and web-based settings using the Pearson correlation coefficient and intraclass correlation coefficients. The results indicated moderate reproducibility for DIN (*r*=0.55, 95% CI 0.27-0.74; *P*<.001) and SIB (*r*=0.55, 95% CI 0.27-0.74; *P*<.001). AuM (F) demonstrated good reliability (*r*=0.75, 95% CI 0.56-0.87; *P*<.001), while AuM (A) showed lower reliability (*r*=0.44, 95% CI 0.14-0.67; *P*=.007). Notably, the GMSI exhibited the highest test-retest reliability among the measures (*r*=0.82, 95% CI 0.67-0.90; *P*<.001). The reproducibility of the GMSI, with a correlation coefficient (*r*) of 0.82, was significantly greater than that of the SIB (mean difference=0.25, 95% CI 0.05-0.46; *P*<.05) and AuM (A) (mean difference=0.37, 95% CI 0.08-0.69; *P*<.05). The correlation coefficient of AuM (F) was not significantly higher than AuM (A) (mean difference=0.31, 95% CI –0.02 to 0.67; *P*=.07) and the DIN was not significantly more reproducible than SIB (mean difference=–0.25, 95% CI –0.36 to 0.29; *P*=.81). The visualizations for these relationships are shown in Figure 3.

Table 1. In-person and web-based auditory cognitive performance.

	Participants (n=58)		Correlation coefficient	Intraclass correlation coefficient
	In person, mean (SD)	Web based, mean (SD)		
GMSI[a] (total score)	156 (36)	151 (23)	0.82	0.72
DIN[b] (dB)[c]	6.35 (2.8)	5.42 (3.2)	0.55	0.49
SIB[d] (dB)	2.45 (2.8)	2.01 (2.1)	0.55	0.62
AuM (F)[e] (a.u)[f]	1.52 (0.6)	1.34 (0.7)	0.75	0.67
AuM (A)[g] (a.u)	2.52 (0.4)	2.42 (0.4)	0.44	0.43

[a]GMSI: Goldsmiths Musical Sophistication Index.

[b]DIN: Digits-in-Noise task threshold.

[c]dB: decibels.

[d]SIB: Speech-in-Babble threshold.

[e]AuM (F): auditory memory precision for frequency.

[f]a.u: arbitrary units.

[g]AuM (A): auditory memory precision for amplitude modulation rate.

XSL•FO
RenderX

Figure 3. Participant (n=58) data for various auditory cognitive measures and the GMSI questionnaire in in-person and web-based settings. In-person data are shown on the x-axis, whereas web-based metrics are shown on the y-axis. AuM (A): auditory memory precision for amplitude modulation rate; AuM (F): auditory memory precision for frequency; DIN: digits-in-noise task threshold; GMSI: Goldsmiths Musical Sophistication Index score; SIB: Speech-in-Babble threshold.

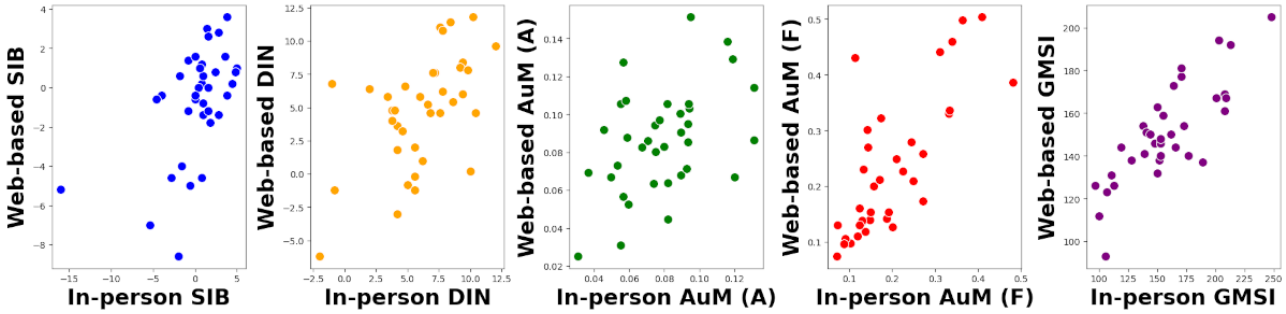

We tested whether there were differences in correlation coefficients between age, PTA thresholds measured in person, and auditory cognitive metrics measured in both settings using the Steiger test. There were no significant differences between the correlations between age and any auditory metric in the web-based and in-person settings. Specifically, these included the DIN (t_{57}=0.165; *P*=.87), SIB (t_{57}=0.162; *P*=.87), AuM (A) (t_{57}=–0.170; *P*=.87), and AuM (F) (t_{57}=0.073; *P*=.94). Additionally, there were no significant differences between the correlations between PTA thresholds and any auditory metrics across settings. These included the DIN (t_{57}=0.005; *P*=.99), SIB (t_{57}=0.031; *P*=.98), AuM (A) (t_{57}=–0.012; *P*=.99), and AuM (F) (t_{57}=0.154; *P*=.88).

Web-Based Auditory Cognitive Testing Associations

We tested whether auditory cognitive associations between variables obtained in person were reproducible on the web with a larger group of web-based participants. The latter included the addition of a new group of participants, from the PREVENT study, who had not previously performed the task. This was done to improve the statistical power of the web-based findings. We examined the reproducibility of the association between AuM for frequency and GMSI scores and between SIB and AuM when conducted in person and web-based. This has been studied previously by our group on an in-person participant group [18,19].

The correlation coefficients for the relationship between SIB and AuM (A) were statistically significant for in-person testing (*r*=0.46, 95% CI 0.34-0.56; *P*<.001) and web-based testing (*r*=0.18, 95% CI 0.0-0.34; *P*<.05; Figure 4). Bootstrapping was used to generate 95% CIs for the correlation coefficients to test if they were significantly different at the significance level with an α of .05. We did not find any differences across the 2 settings for this relationship. Similarly, the correlation coefficients for the relationship between GMSI scores and AuM (F) were statistically significant for in-person testing (*r*=0.52, 95% CI 0.42-0.61; *P*<.001) and web-based testing (*r*=0.47, 95% CI 0.32-0.6; *P*<.001). We did not find any differences across the 2 settings for this relationship. This suggested that there were no differences in the correlation of these auditory measures in in-person or web-based settings.

XSL•FO
RenderX

Figure 4. The relationship between SIB perception ability and AuM (A) in web-based (n=147; top panel: orange) and in-person (n=153; bottom panel: green) participants. Shaded areas indicate CIs for the regression lines of best fit. There is greater variability for web-based participant data compared to in-person data. AuM (A): auditory memory precision for amplitude modulation; SIB: Speech-in-Babble threshold.

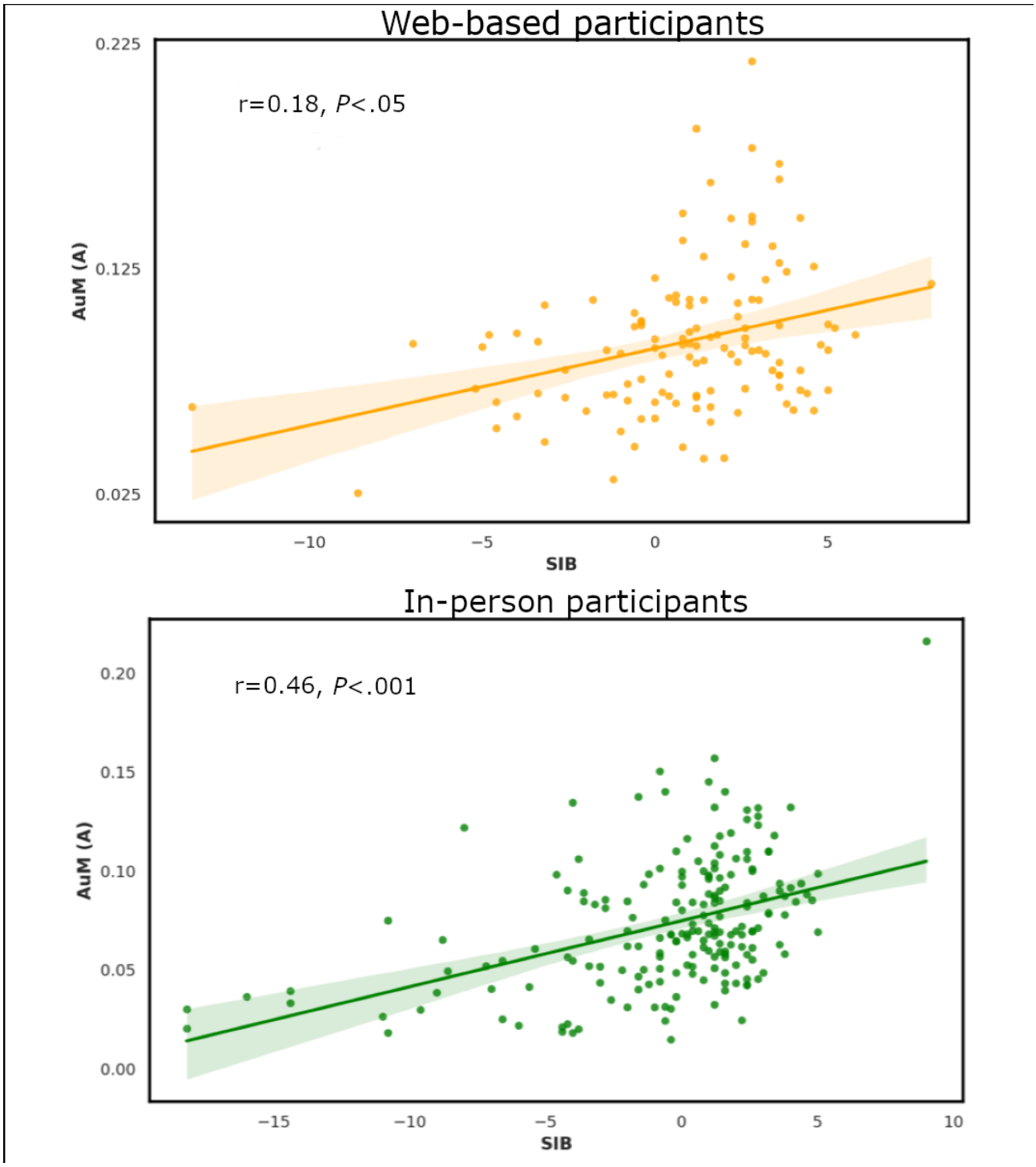

Discussion

Principal Findings

The principal findings of this study were that remote auditory testing of cognition is reliable when conducted on the web. However, web-based participation may not be fully representative of the participants who may take part in person. We found that older participants and those with severe hearing loss were less likely to participate on a web-based platform. Furthermore, we found that the questionnaire-based test, the GMSI, had the best reproducibility as compared to the auditory cognitive tasks, which did not differ significantly when conducted in person or on the web. Importantly, previously published findings were reproducible on the web suggesting that web-based testing may be a suitable avenue to test auditory cognition.

Comparison to Prior Work

Although this study has not been conducted previously, our experience with web-based auditory experimentation was similar to previous work in the field [5,14,15]. Older adults in the

XSL•FO
RenderX

"real-world" were able to successfully negotiate a self-directed, web-based auditory experimentation portal and give reliable results in comparison to in-person testing. Our study differed from the previous work in the type and range of stimuli across the same experimental paradigm. For example, we tested the perception of digits and sentences on a noisy background and sounds manipulated by frequency and temporal fluctuations. This allowed us to test different types of auditory stimuli across the same type of experiment (SIN perception and AuM). We did not find any differences in the reproducibility of any of these measures across in-person or web-based settings. Importantly, there was no significant difference when comparing the relationships of these metrics with shared variables across both settings such as age and hearing thresholds.

The focus of this study was to engage participants with a range of hearing abilities that are recruited for in-person studies. Despite an increase in web-based research methods, there may be an underrepresentation of participants who lack adequate digital literacy or find it difficult to interact with computer-generated stimuli. In particular, people with hearing loss may rely more on visual cues to understand information, which may be absent from typical auditory research performed on healthy young adults without hearing difficulty [23]. Although we did not explore the qualitative reasons behind the reduced proportion of participants with severe hearing loss in our web-based cohort, these participants were also more likely to be hearing aid users, which could have interfered with their use of headphones. Data on whether participants were using Bluetooth-enabled devices that allowed connection to their electronic devices were not collected. Additionally, to maximize participation in the study, we gave volunteers up to 6 months to complete their web-based experiment. Although this may have improved participant retention to a degree, we did not remeasure their PTA thresholds to ensure the peripheral hearing was the same as in the in-person experiment. This may have improved the reproducibility of the web-based auditory cognitive metrics.

We also found no differences in the reproducibility of auditory cognitive stimuli with particular sound parameters. There could have been better reproducibility for AuM (F) due to the invariance of sounds with a pitch to changes in timbre over different electronic devices [24]. The pitch of a sound is also less likely to be affected if it is played on headphones or speakers and due to the sound cards in electronic devices. For example, the frequency space from which the pure tones were generated was from 440 to 880 Hz, which is captured by most audio devices. Speech-based sounds and amplitude-modulated white noise stimuli would have only been limited by the bandwidth of the electronic devices and equipment. This could have affected the performance in the SIN tasks and AuM (A) task on the web and contributed to the lower correlation coefficient. However, these were not statistically different. Further work is needed to clarify the reasons for this.

The observed similarities between amplitude-modulated noise stimuli and speech in conversations highlight their importance in understanding auditory processing, particularly in relation to SIN perception ability. Both modalities exhibit temporal amplitude fluctuations critical for investigating auditory system functions without background noise [25]. Modulated noise mimics speech's natural rhythms, aiding in the study of temporal processing in speech. Cognitive engagement is essential in both, requiring attention and working memory to decode or track fluctuations; however, the presence of linguistic and phonetic information in speech is not found in the generated sounds of the AuM task. Despite these differences, the correlation between nonspeech stimuli in auditory modulation tasks and verbal SIN tasks indicates underlying auditory cognitive abilities shared across both domains, bridging the gap between speech and nonspeech auditory processing. This suggests a broader scope of auditory cognition beyond specific linguistic content, highlighting fundamental auditory processing mechanisms applicable across different auditory stimuli. The reproducibility of this relationship across the in-person and web-based settings suggests that these fundamental connections between the stimuli can be investigated on the web.

Strengths and Limitations

The home environment of a participant may affect a participant's performance in a number of ways. Although there were instructions to set the volume level to one that was comfortable in our task, participants may alter this through the course of the tasks. The home environment is also prone to background fluctuations in noise that may not be observed in strict laboratory conditions. These interactions may affect the listening ability of participants despite the sounds being at a comfortable level [26]. Finally, despite using headphones, the sound quality that may have been transmitted through them could have varied between participants resulting in any changes in sound quality mentioned. It was impossible to monitor this accurately during this study. Despite all of these challenges, it is promising for the auditory measures used in this study to display reproducibility across a range of different SIN stimuli and AuM tasks with different sound features. However, the participants in this study were recruited from a local university volunteer database, which may introduce a selection bias as these individuals are more likely to have prior experience with research participation. Therefore, while our findings provide valuable insights, they may not be fully generalizable to the broader population and this limitation should be considered when interpreting the results.

The strengths of this study include the inclusion of "real-world" older adult participants in independent cohorts, with a range of hearing abilities, to perform in-person and web-based auditory experiments. This allowed us to accurately assess the differences in within-subject parameters that are attributable to web-based testing. We also used multiple types of verbal and nonverbal stimuli to test auditory cognitive abilities, which allowed us to find differences in performance for particular sound features, like frequency versus amplitude modulation, in both settings.

Future Work

Future work could include an assessment of the reasons for which participants did not participate in the web-based session despite agreeing to participate in this initially. There may have been other factors, rather than those related to hearing, that influenced these decisions for each participant. Other improvements to the study design include remote guided testing,

XSL•FO
RenderX

where the participant performs the task on their home computer but through a screen share [27]. This would ensure that the home environment and equipment could "pass" a minimum standard to perform the tasks. Recruitment of a larger group of participants who return for web-based testing may allow further analysis to assess the reproducibility of the auditory measures across different age groups and hearing loss severity levels. This will allow us to assess if these are factors that affect the reproducibility of certain results.

Acknowledgments

The Newcastle National Institute of Health Research Biomedical Research Centre supported the study. ML has received funding from the National Institute of Health Research, Alzheimer's Research UK, the Guarantors of Brain, and Medical Research Council (MR/V006568/1) in the United Kingdom. JPT receives funding from the National Institute of Health Research Biomedical Research Centre, United Kingdom in Newcastle upon Tyne. TG receives funding from the Medical Research Council (MR/T032553/1), Wellcome Trust (WT106964MA), and National Institute on Deafness and Other Communication Disorders (DC000242 36).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ML performed the conceptualization, methodology, investigation, and writing—original draft. ML, JPT, and TG contributed to the writing—review and editing. ML and TG performed the funding acquisition and resources. ML, JPT, and TG performed the supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Goldsmiths Musical Sophistication Index (GMSI) questionnaire.
[PDF File (Adobe PDF File), 75 KB - jmir_v26i1e58444_app1.pdf]

References

1. Sauter M, Draschkow D, Mack W. Building, hosting and recruiting: a brief introduction to running behavioral experiments online. Brain Sci 2020;10(4):251 [FREE Full text] [doi: 10.3390/brainsci10040251] [Medline: 32344671]
2. Tomczak J, Gordon A, Adams J, Pickering JS, Hodges N, Evershed JK. What over 1,000,000 participants tell us about online research protocols. Front Hum Neurosci 2023;17(1228365) [FREE Full text] [doi: 10.3389/fnhum.2023.1228365]
3. Peyton K, Huber GA, Coppock A. The generalizability of online experiments conducted during the COVID-19 pandemic. J Exp Polit Sci 2021;9(3):379-394. [doi: 10.1017/xps.2021.17]
4. Eerola T, Armitage J, Lavan N, Knight S. Online data collection in auditory perception and cognition research: recruitment, testing, data quality and ethical considerations. Auditory Perception Cogn 2021;4(3-4):251-280. [doi: 10.1080/25742442.2021.2007718]
5. Zhao S, Brown CA, Holt LL, Dick F. Robust and efficient online auditory psychophysics. Trends Hear 2022;26 [FREE Full text] [doi: 10.1177/23312165221118792] [Medline: 36131515]
6. Milne AE, Bianco R, Poole KC, Zhao S, Oxenham AJ, Billig AJ, et al. An online headphone screening test based on dichotic pitch. Behav Res Methods 2021;53(4):1551-1562 [FREE Full text] [doi: 10.3758/s13428-020-01514-0] [Medline: 33300103]
7. Mok BA, Viswanathan V, Borjigin A, Singh R, Kafi H, Bharadwaj HM. Web-based psychoacoustics: hearing screening, infrastructure, and validation. Behav Res Methods 2024;56(3):1433-1448 [FREE Full text] [doi: 10.3758/s13428-023-02101-9] [Medline: 37326771]
8. Baumgartner H, Schulz A, Hein A, Holube I, Herzke T. A fitting method for headphones to compensate individual hearing impairments. : IEEE; 2009 Presented at: 3rd International Conference on Pervasive Computing Technologies for Healthcare; April 1-3, 2009; London, UK p. 1. [doi: 10.4108/icst.pervasivehealth2009.5991]
9. Stewart N, Chandler J, Paolacci G. Crowdsourcing samples in cognitive science. Trends Cogn Sci 2017;21(10):736-748 [FREE Full text] [doi: 10.1016/j.tics.2017.06.007] [Medline: 28803699]
10. Loignon C, Dupéré S, Leblanc C, Truchon K, Bouchard A, Arsenault J, et al. Equity and inclusivity in research: co-creation of a digital platform with representatives of marginalized populations to enhance the involvement in research of people with limited literacy skills. Res Involv Engagem 2021;7(1):70 [FREE Full text] [doi: 10.1186/s40900-021-00313-x] [Medline: 34610847]

XSL•FO
RenderX

11. Cyr AA, Romero K, Galin-Corini L. Web-based cognitive testing of older adults in person versus at home: within-subjects comparison study. JMIR Aging 2021;4(1):e23384 [FREE Full text] [doi: 10.2196/23384] [Medline: 33522972]
12. Slote J, Strand JF. Conducting spoken word recognition research online: validation and a new timing method. Behav Res Methods 2016;48(2):553-566. [doi: 10.3758/s13428-015-0599-7] [Medline: 25987305]
13. Bianco R, Mills G, de Kerangal M, Rosen S, Chait M. Reward enhances online participants' engagement with a demanding auditory task. Trends Hear 2021;25:23312165211025941 [FREE Full text] [doi: 10.1177/23312165211025941] [Medline: 34170748]
14. Bianco R, Hall ETR, Pearce MT, Chait M. Implicit auditory memory in older listeners: from encoding to 6-month retention. Curr Res Neurobiol 2023;5:100115 [FREE Full text] [doi: 10.1016/j.crneur.2023.100115] [Medline: 38020808]
15. Bianco R, Chait M. No link between speech-in-noise perception and auditory sensory memory - evidence from a large cohort of older and younger listeners. Trends Hear 2023;27 [FREE Full text] [doi: 10.1177/23312165231190688] [Medline: 37828868]
16. Müllensiefen D, Gingras B, Musil J, Stewart L. The musicality of non-musicians: an index for assessing musical sophistication in the general population. PLoS One 2014;9(2):e89642 [FREE Full text] [doi: 10.1371/journal.pone.0089642] [Medline: 24586929]
17. Ritchie CW, Ritchie K. The PREVENT study: a prospective cohort study to identify mid-life biomarkers of late-onset alzheimer's disease. BMJ Open 2012;2(6):e001893 [FREE Full text] [doi: 10.1136/bmjopen-2012-001893] [Medline: 23166135]
18. Lad M, Holmes E, Chu A, Griffiths TD. Speech-in-noise detection is related to auditory working memory precision for frequency. Sci Rep 2020;10(1):13997 [FREE Full text] [doi: 10.1038/s41598-020-70952-9] [Medline: 32814792]
19. Lad M, Billig AJ, Kumar S, Griffiths TD. A specific relationship between musical sophistication and auditory working memory. Sci Rep 2022;12(1):3517 [FREE Full text] [doi: 10.1038/s41598-022-07568-8] [Medline: 35241747]
20. Wiley TL, Chappell R, Carmichael L, Nondahl DM, Cruickshanks KJ. Changes in hearing thresholds over 10 years in older adults. J Am Acad Audiol 2008;19(4):281-92; quiz 371 [FREE Full text] [doi: 10.3766/jaaa.19.4.2] [Medline: 18795468]
21. Holmes E, Griffiths TD. 'Normal' hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception. Sci Rep 2019;9(1):16771 [FREE Full text] [doi: 10.1038/s41598-019-53353-5] [Medline: 31728002]
22. Ma WJ, Husain M, Bays PM. Changing concepts of working memory. Nat Neurosci 2014;17(3):347-356 [FREE Full text] [doi: 10.1038/nn.3655] [Medline: 24569831]
23. Suess N, Hauswald A, Zehentner V, Depireux J, Herzog G, Rösch S, et al. Influence of linguistic properties and hearing impairment on visual speech perception skills in the German language. PLoS One 2022;17(9):e0275585 [FREE Full text] [doi: 10.1371/journal.pone.0275585] [Medline: 36178907]
24. McPherson MJ, McDermott JH. Relative pitch representations and invariance to timbre. Cognition 2023;232:105327 [FREE Full text] [doi: 10.1016/j.cognition.2022.105327] [Medline: 36495710]
25. Goswami U. Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. Ann N Y Acad Sci 2019;1453(1):67-78. [doi: 10.1111/nyas.14137] [Medline: 31237357]
26. Ruggles D, Bharadwaj H, Shinn-Cunningham BG. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. Proc Natl Acad Sci U S A 2011;108(37):15516-15521 [FREE Full text] [doi: 10.1073/pnas.1108912108] [Medline: 21844339]
27. Leong V, Raheel K, Sim JY, Kacker K, Karlaftis VM, Vassiliu C, et al. A new remote guided method for supervised web-based cognitive testing to ensure high-quality data: development and usability study. J Med Internet Res 2022;24(1):e28368 [FREE Full text] [doi: 10.2196/28368] [Medline: 34989691]

Abbreviations

AM: amplitude modulation
AuM: auditory memory
AuM (A): auditory memory for amplitude modulation rate
AuM (F): auditory memory for frequency
DIN: Digits-in-Noise
GMSI: Goldsmiths Musical Sophistication Index
HL: hearing level
PTA: pure tone audiometry
SIB: Speech-in-Babble
SIN: speech-in-noise
SNR: signal-to-noise ratio

XSL•FO
RenderX

Edited by A Coristine; submitted 16.03.24; peer-reviewed by P Sooful, Y Zhang, A Moulin, TK Koerner; comments to author 27.05.24; revised version received 26.06.24; accepted 10.10.24; published 09.12.24.

Please cite as:
Lad M, Taylor JP, Griffiths TD
Reliable Web-Based Auditory Cognitive Testing: Observational Study
J Med Internet Res 2024;26:e58444
URL: https://www.jmir.org/2024/1/e58444
doi:10.2196/58444
PMID:

©Meher Lad, John-Paul Taylor, Timothy David Griffiths. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Research Trends on Metabolic Syndrome in Digital Health Care Using Topic Modeling: Systematic Search of Abstracts

Kiseong Lee[1*], PhD; Yoongi Chung[2], BSN; Ji-Su Kim[2*], PhD

[1]Humanities Research Institute, Chung-Ang University, Seoul, Republic of Korea

[2]Department of Nursing, Chung-Ang University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Ji-Su Kim, PhD
Department of Nursing
Chung-Ang University
84, Heukseok-ro, Dongjak-gu, Seoul, Republic of Korea
Seoul, 06974
Republic of Korea
Phone: 82 028205991
Email: jisu80@cau.ac.kr

Abstract

Background: Metabolic syndrome (MetS) is a prevalent health condition that affects 20%-40% of the global population. Lifestyle modification is essential for the prevention and management of MetS. Digital health care, which incorporates technologies like wearable devices, mobile apps, and telemedicine, is increasingly becoming integral to health care systems. By analyzing existing research trends in the application of digital health care for MetS management, this study identifies gaps in current knowledge and suggests avenues for future research.

Objective: This study aimed to identify core keywords, topics, and research trends concerning the use of digital health care in the management of MetS.

Methods: A systematic search of abstracts from peer-reviewed papers was conducted across 6 academic databases. Following eligibility screening, 162 abstracts were selected for further analysis. The methodological approach included text preprocessing, text network analysis, and topic modeling using the BERTopic algorithm.

Results: Analysis of the 162 selected abstracts yielded a keyword network comprising 1047 nodes and 34,377 edges. The top 5 core keywords were identified as “MetS,” “use,” “patient,” “health,” and “intervention.” We identified 12 unique topics, with topic 1 focusing on the use of telehealth for self-management of diabetes. The diversity of the 12 topics reflected various aspects of digital health care, including telehealth for diabetes management, electronic health records for MetS complications, and wearable devices for monitoring metabolic status. Research trends showed an expanding field of precision medicine driven by the demand for tailored interventions and the significant impact of the COVID-19 pandemic.

Conclusions: By analyzing past research trends and extracting data from scholarly databases, this study has provided valuable insights that can guide future investigations in the field of digital health care and MetS management.

(J Med Internet Res 2024;26:e53873) doi:10.2196/53873

KEYWORDS

metabolic syndrome; digital health care; topic modeling; text network analysis; research trends; topic modeling; prevention; management; telemedicine; wearable; devices; apps; applications; methodological; cardiovascular disease

Introduction

Metabolic syndrome (MetS) is reported to have a global prevalence ranging between 20% and 40% [1]. Characterized as a cluster of major interrelated risk factors for cardiovascular disease, MetS significantly exacerbates health outcomes by more than doubling the risk of cardiovascular diseases [2]. A MetS diagnosis is established when 3 or more of the following 5 risk factors are present: waist circumference indicative of obesity, low high-density lipoprotein cholesterol, hypertriglyceridemia, hypertension, and hyperglycemia [3]. As MetS is closely linked to lifestyle factors such as poor diet and

XSL•FO
RenderX

insufficient physical activity (PA) [4], it can be prevented by targeted lifestyle modifications [5]. Given the substantial individual, societal, and economic burdens MetS imposes, proactive prevention and ongoing management are imperative.

Digital health care refers to the use of information and communication technology to monitor and manage health and well-being [6]. Advances in the information and communication technology sector have facilitated health care delivery through mobile apps, wearable devices, artificial intelligence, the Internet of Things, and telemedicine [7]. Digital health care has already been extensively applied across various health care domains. For instance, smart bands, a category of wearable devices, can quantify PA metrics such as step count, walking distance, and walking intensity [8]. Studies indicate that the adoption of these wearable devices can enhance PA levels among individuals with chronic conditions [9]. Moreover, health-related mobile apps have the potential to instigate behavioral changes in users by providing educational content, personalized health counseling, and diet and exercise monitoring [10]. This approach takes genetic, environmental, and lifestyle factors into account to develop personalized treatment plans, offering potential improvements in the management of MetS through more targeted interventions [11]. In other words, it enables the application of precision medicine, a health care approach that tailors medical treatment to each patient's individual characteristics [11]. Given that digital health care is already widely used in health care settings, evidence from multiple studies [1,5,7-10] supports its applicability in managing MetS-associated risk factors, including obesity and diabetes.

To mitigate the impact of MetS and enhance long-term quality of life through ongoing management, a deep understanding of existing research trends is crucial. Previous studies on MetS within the realm of digital health care have used systematic reviews, meta-analyses [9,12-14], and scoping reviews [15]. However, these approaches have limitations, particularly in the depth to which they explore subjects, key concepts, and topics. Given the rapid advancements in digital health care research, it is imperative that researchers stay informed about the evolving knowledge structure, including emerging topics and shifting trends in MetS and digital health care research. This understanding is essential for maintaining the relevance and efficacy of ongoing and future studies in this dynamic field.

This study used text network analysis to analyze the frequency and patterns of the co-occurrence of words in textual data such as article abstracts [16,17]. This method offers the benefit of allowing a visual and intuitive understanding of the relationships between significant keywords, serving as a valuable network statistical indicator [16,17]. Furthermore, topic modeling serves as a statistical inference model that uses text mining to objectively and clearly evaluate research topics [18]. We used text mining analysis to examine trends in MetS-related digital health care research to understand the existing knowledge structure and inform the direction of future investigations. Accordingly, this study aimed to clarify the relationships between keywords by analyzing previous research on MetS within the domain of digital health care through text network analysis and to investigate research topics related to MetS within the digital health care sector.

The objective of this study was to construct a network that consolidates prior research on MetS within the field of digital health care. The study aimed to create an informed body of literature by comparing and analyzing keywords and research topics. The study had 3 primary goals: first, to identify the terms associated with MetS research in digital health care; second, to establish a text network comprising core keywords from existing literature on MetS and digital health care; and third, to identify trends in MetS research within the digital health care context through topic modeling.

Methods

Study Design

This is a descriptive study in which we used topic modeling techniques to classify keywords into topics within the realm of digital health care for MetS management and tracked topic-specific trends over time.

Search Strategy and Data Collection

We conducted a comprehensive search of 6 electronic academic databases, namely PubMed, Embase, Cochrane, CINAHL, SCOPUS, and Web of Science. The search strategy involved using combinations of various terms to filter publication titles, abstracts, and keywords. The terms used included: (Digital health OR mobile health OR ehealth OR e-health OR mhealth OR telehealth OR electronic health OR web-based OR apps OR wearables OR devices) AND (metabolic syndrome OR Metabolic Syndromes OR syndrome, metabolic OR syndromes, metabolic OR MetS OR syndrome X) (Multimedia Appendix 1).

The search and selection processes were performed in May 2023. We included all abstracts from articles published before April 20, 2023. There were no restrictions on the start date of the literature to ensure a comprehensive analysis of studies over time. Studies were excluded if they involved participants with conditions other than MetS or conditions related to subdiagnosis criteria such as obesity, diabetes, hypertension, and dyslipidemia. Additionally, studies without an abstract or with a non-English abstract were excluded.

From the initial 33,078 studies identified, 17,675 duplicate entries were removed. The remaining 15,403 studies were screened based on their title and abstract. Studies were further excluded based on the following criteria: 12,746 studies lacked keywords related to both MetS and digital health care, 2114 studies did not focus on patients, 380 studies lacked an abstract, and one was a duplicate. Subsequently, 2 researchers independently screened the identified articles' titles and abstracts. In cases of disagreement, consensus was reached through discussion and further review against the selection criteria. Ultimately, a total of 162 articles were included for data analysis (Figure 1). Data, including serial numbers, authors, titles, author keywords, and abstracts, were organized into a matrix using Microsoft Excel.

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

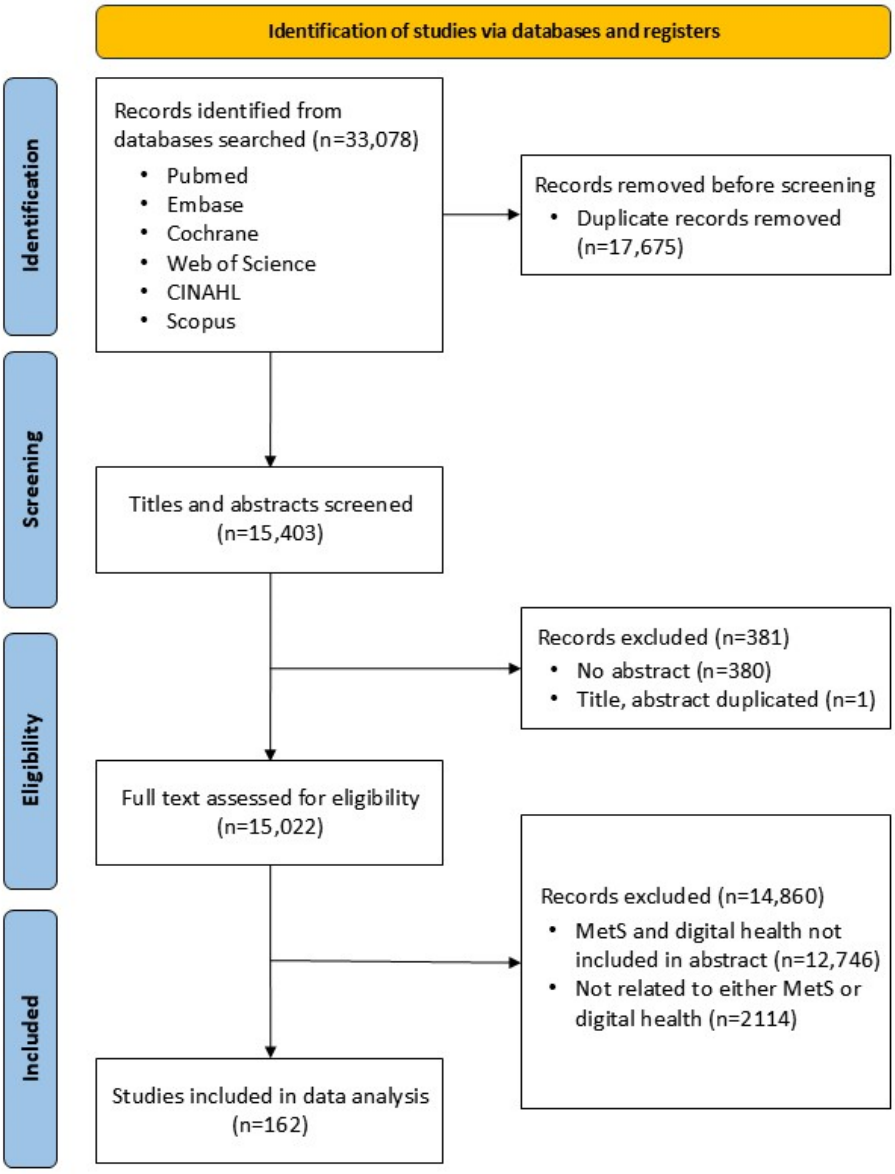

Analysis

We performed text preprocessing, topic modeling, and text network analysis on the collected abstracts. Custom code written in Python was used to execute these tasks. The code can be reviewed on GitHub [19].

Text Preprocessing

In the text preprocessing stage, we initiated the process by removing internet address formats, HTML tags, and special characters from the abstracts. Additionally, any text that was not in English was discarded. We then harmonized expressions that had similar or identical meanings; for instance, the terms "mobile health," "mhealth," and "m-health" were consolidated as "m-health." We also compiled a list of specialized terms to avoid issues with word separation because of spacing; notably, the term "randomized controlled trial" was transformed into "randomized_controlled_trial." To refine the focus of our analysis, we carefully curated a list of stop words by iteratively reviewing the text analysis results. This process allowed us to differentiate between common terms and significant terms, thus ensuring that essential information was retained. Consequently, we excluded typical English stop words as well as frequently occurring domain-specific terms such as "study," "result," and "research." The aim was to prioritize in-depth analysis over a more generalized examination. To standardize different word forms like -ing, -ed, and -s, we used lemmatization. The preprocessing stage concluded with the adoption of noun extraction as a standard text mining practice to augment semantic understanding.

Text Network Analysis

In text network analysis, words act as nodes, and their connections form edges, thereby enabling semantic analysis. In this study, we established a "co-occurrence" relationship among nouns present in the abstracts of scholarly articles, representing these relationships as edges in the network. Focusing on the specific keyword "metabolic_syndrome," we constructed a network based on the frequency of word co-occurrence, limiting edges to instances where words appeared together at least twice.

XSL•FO
RenderX

We also allocated higher weights to edges where word co-occurrence was repeated.

Our network construction resulted in a total of 1047 nodes and 34,377 edges. Using this foundational data, we evaluated the degree, closeness, and betweenness centrality for each node. Degree centrality quantifies the number of edges linked to a given node, thus indicating its direct connections to other words within the network. Closeness centrality is a measure of how quickly a node can reach all other nodes in the network, reflecting its proximity to a multitude of other terms. Betweenness centrality measures the frequency with which a node resides on the shortest paths between other nodes in the network, highlighting its function as a connector or bridge among words.

Given the substantial number of nodes and edges, the network inherently possesses the characteristics of a dense graph, complicating the task of visualization. To address this issue, we chose to visualize only the top 5% of nodes based on their degree centrality, which yielded 52 nodes, and focused our analysis on keywords strongly associated with "metabolic syndrome." For both network construction and visualization, we used the Python library network to analyze the degree centralization index.

Topic Modeling

Topic modeling is a text mining technique for elucidating the inherent semantic structure of textual data. We performed topic modeling on a refined set of abstracts to uncover the latent themes within them. We included only abstracts in our analysis because topic modeling has been shown to be particularly effective with short texts [20]. Classical algorithms for this purpose include latent semantic analysis and latent Dirichlet allocation. These traditional methods rely on statistics-based word embeddings, which often ignore the order and relationships between words within the text. More advanced models, such as the combined topic model [21] and BERTopic [22], have recently been introduced for this task. In this study, we used the BERTopic model, which uses BERT (Bidirectional Encoder Representations from Transformers), a large language model, to perform semantic-based embeddings. BERT captures the contextual relationships between words by considering their order and entire sentence context through mechanisms like attention. This allows BERTopic to extract topics that are sensitive to the relationships between words, leading to more nuanced and contextually aware topic modeling. Our implementation was conducted using the BERTopic API (application programming interface) [23].

The BERTopic API is designed for topic modeling, leveraging BERT embeddings and clustering algorithms to extract coherent topics from textual data. It uses large language models like BERT to capture the semantic relationships between words and then applies techniques such as hierarchical density–based clustering of applications with noise for clustering and term frequency-inverse document frequency for keyword extraction. This combination allows BERTopic to generate interpretable topics that reflect the underlying themes in the data, making it a powerful tool for analyzing and understanding large text corpora. We used the "all-mpnet-base-v2" large language model, which has been fine-tuned on over a billion sentences and was chosen for its top performance in sentence embeddings [24]. Using BERTopic, we conducted topic analysis on a dataset comprising 162 abstracts, which led to the extraction of 13 distinct topics. Among these, 1 topic contained outlier values that presented challenges for semantic interpretation. To address this, we amalgamated the documents associated with this outlier topic into the nearest topic cluster, resulting in a final count of 12 topics. Each topic was represented by a set of words, which we closely examined along with the papers related to each topic to assign them appropriate names.

Results

Core Keywords

A network comprising 1047 nodes (keywords) and 34,377 edges was constructed from the 162 abstracts. Table 1 presents the core keywords associated with MetS in the context of digital health care. The top 10 keywords, ranked by edge count, are as follows: "MetS" (817 edges), "use" (n=724), "patient" (n=644), "health" (n=610), "intervention" (n=555), "data" (n=536), "risk" (n=526), "group" (n=503), "increase" (n=483), and "application" (n=456). Figure 2 presents the 817 keywords that co-occur with "MetS," which has the highest degree of centrality. The degree centrality of the keywords, in descending order, are: "MetS" (0.781), "use" (0.692), "patient" (0.616), "health" (0.583), "intervention" (0.531), "data" (0.512), "risk" (0.503), "group" (0.481), "increase" (0.463), and "application" (0.436). The closeness centrality measures for the keywords are: "MetS" (0.819), "use" (0.764), "patient" (0.721), "health" (0.705), "intervention" (0.680), "data" (0.671), "risk" (0.667), "group" (0.657), "increase" (0.650), and "application" (0.639). The betweenness centrality of the keywords, in descending order, are "MetS" (0.117), "use" (0.071), "patient" (0.066), "health" (0.042), "intervention" (0.034), "data" (0.034), "risk" (0.030), "group" (0.029), "increase" (0.024), and "application" (0.020).

XSL•FO
RenderX

Table 1. Core keywords in MetS research within digital health care.

Rank	Keywords	Degree	Degree centrality	Closeness centrality	Betweenness centrality
1	MetS	817	0.781	0.819	0.117
2	use	724	0.692	0.764	0.071
3	patient	644	0.616	0.721	0.066
4	health	610	0.583	0.705	0.042
5	intervention	555	0.531	0.680	0.034
6	data	536	0.512	0.671	0.034
7	risk	526	0.503	0.667	0.030
8	group	503	0.481	0.657	0.029
9	increase	484	0.463	0.650	0.024
10	application	456	0.436	0.639	0.020
11	effect	444	0.424	0.634	0.018
12	blood pressure	432	0.413	0.629	0.017
13	disease	416	0.398	0.623	0.016
14	diabetes	403	0.385	0.619	0.015
15	factor	399	0.381	0.617	0.014
16	care	395	0.378	0.616	0.013
17	management	393	0.376	0.615	0.012
18	year	391	0.374	0.614	0.012
19	obesity	384	0.367	0.612	0.011
20	body	376	0.359	0.609	0.011
21	physical activity	376	0.359	0.609	0.010
22	control	360	0.344	0.603	0.010
23	change	348	0.333	0.599	0.009
24	outcome	343	0.328	0.597	0.009
25	population	341	0.326	0.597	0.009

XSL•FO
RenderX

Figure 2. Network of top 5% keywords based on degree centrality.

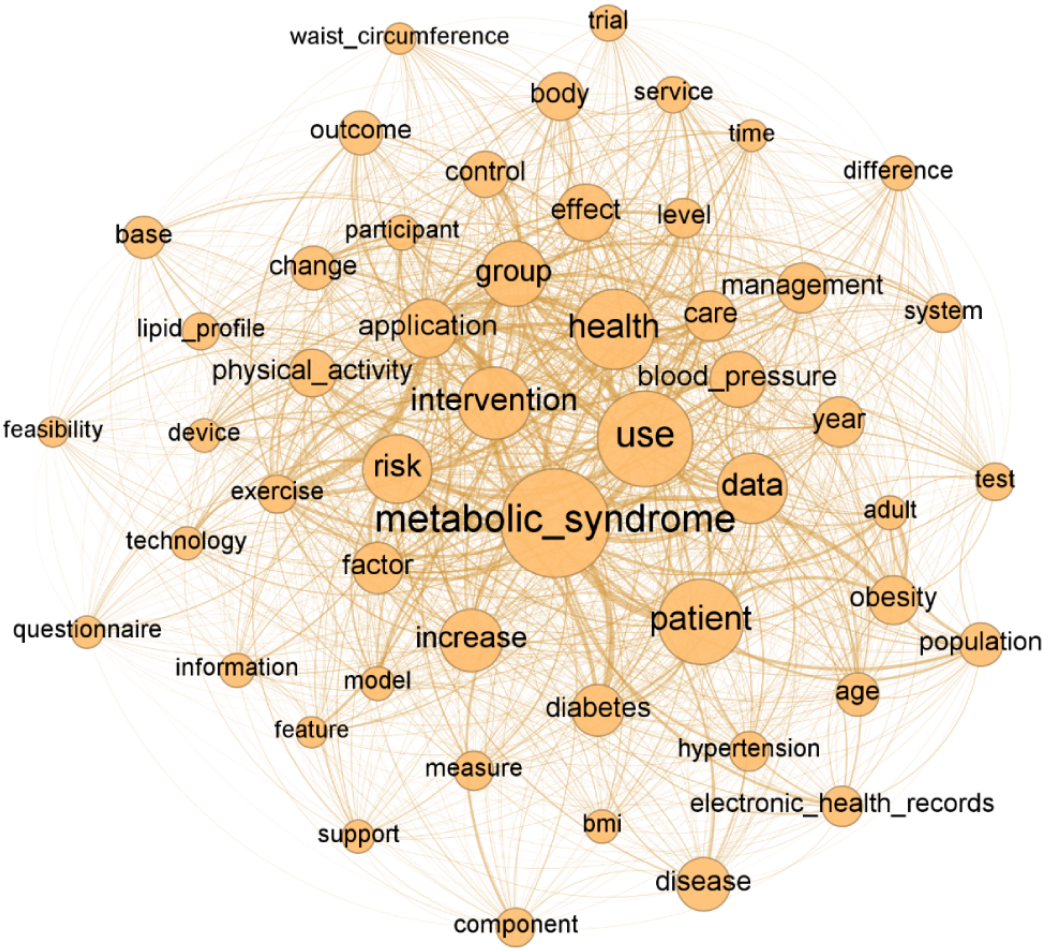

Text Network Analysis

Using the core keywords extracted from the abstracts, a text network analysis was conducted to visualize relationships among these core keywords. To enhance understanding of the keyword structure, we visualized the network of top keywords based on degree centrality, highlighting the top 5% of nodes (Figure 2). "MetS," the keyword most frequently connected to others, occupied the central position within the network. Moreover, a network of core keywords was constructed by organizing keywords in proximity to "MetS" based on their frequency of co-occurrence with it (Figure 3). The top 5 keywords frequently co-occurring with "MetS" were "risk," "patient," "diabetes," "intervention," and "use."

XSL•FO
RenderX

Figure 3. Network of core keywords related to MetS. MetS: metabolic syndrome.

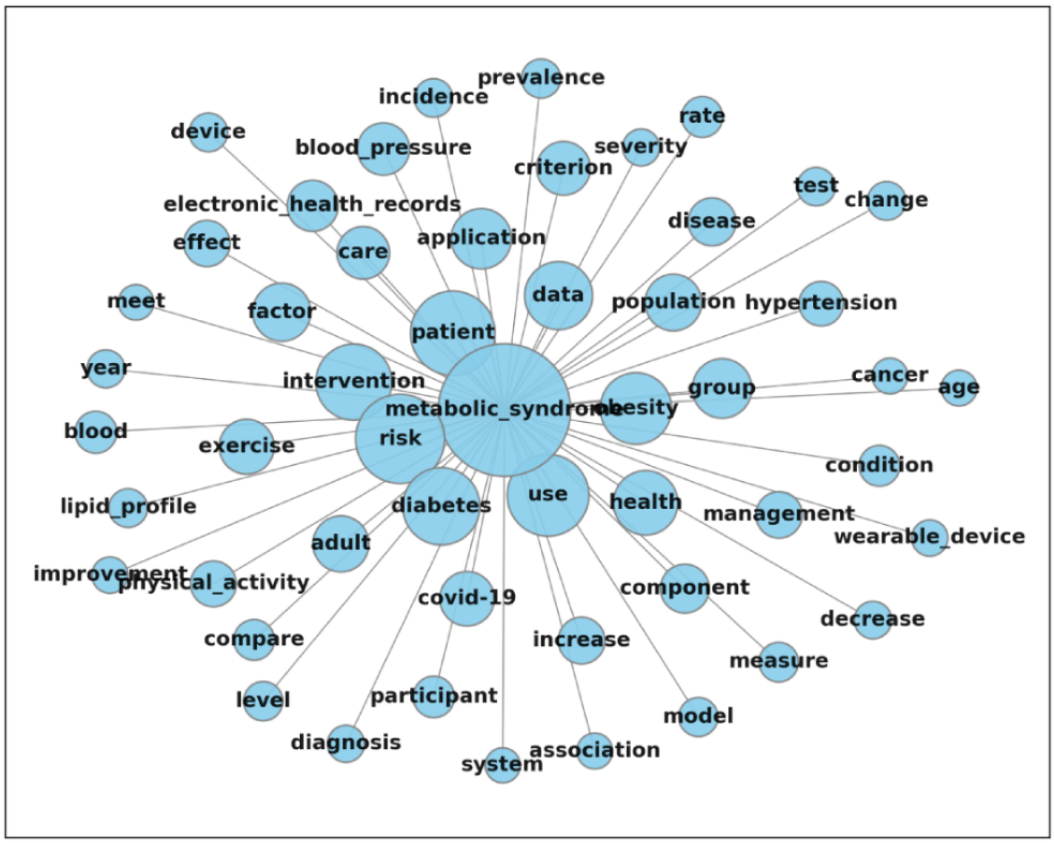

Topic Modeling

Table 2 presents the distribution of keywords related to MetS in digital health care studies, organized by topic. Each topic was identified and labeled based on its associated keywords. Topic 1 constituted 19.8% of all articles and featured the top 10 keywords, in descending order of probability, as follows: “telehealth,” “diabetes,” “control,” “intervention,” “self-management,” “care,” “effect,” “group,” “type,” and “health.” Topic 1 was labeled “use of telehealth for self-management intervention of diabetes control.” Topic 2 comprised 14.2% of the total articles, with its top 10 keywords being “MetS,” “electronic health records,” “diabetes,” “patients,” “complication,” “risk,” “criterion,” “data,” “population,” and “model.” Topic 2 was labeled “use of EHR to manage MetS complications.” Topic 3, which accounted for 12.3% of all articles, was named “use of wearable devices for monitoring metabolic status.” The top 10 keywords for this topic included “device,” “wearable device,” “use,” “sensor,” “technology,” “monitoring,” “data,” “sit,” “disease,” and “review.” Topic 4 represented 10.5% of the articles and was labeled “use of apps for weight loss intervention.” Its top 10 keywords were “obesity,” “weight loss,” “intervention,” “program,” “health,” “body,” “application,” “weight,” “effect,” and “group.” Topic 5, making up 6.8% of the articles, was named “use of digital health for health-related behavior-changing intervention.” The top 10 keywords for this topic were “intervention,” “application,” “participant,” “hospitalization,” “care,” “change,” “m-health,” “trial,” “effect,” and “behavior.” Topic 6, accounting for 6.2% of all articles, was labeled “use of digital health to assess MetS status and trend during COVID-19.” Its top 10 keywords were “COVID-19,” “diabetes,” “MetS,” “patient,” “data,” “disease,” “health,” “disorder,” “incidence,” and “decrease.” Topic 7, constituting 7.4% of the articles, was named “use of digital health for blood pressure control with physical activity.” The top 10 keywords were “blood pressure,” “factor,” “group,” “physical activity,” “risk,” “MetS,” “child,” “measure,” “intervention,” and “adult.” Topic 8, also comprising 7.4% of the articles, was labeled “use of apps for engaging in physical activity.” The top 10 keywords for this topic were “physical activity,” “application,” “intervention,” “minute,” “body,” “engagement,” “week,” “user,” “effect,” and “woman.” Topic 9, accounting for 6.2% of the articles, was named “use of apps for access to data on individual metabolic status.” The top 10 keywords were “application,” “user,” “health,” “use,” “access,” “care,” “management,” “data,” “information,” and “feature.” Topic 10, representing 4.3% of all articles, was labeled “use of m-health for diabetes control.” Its top 10 keywords were “application,” “diabetes,” “smartphone,” “field,” “modification,” “HbA_{1c},” “review,” “use,” “control,” and “search.” Topics 11 and 12 each accounted for 2.5% of the articles. Topic 11 was named “use of digital health for management of metabolic status in women of child-bearing age,” and its top 10 keywords were “woman,” “pregnancy,” “feedback,” “health,” “pilot,” “intervention,” “program,” “delivery,” “interview,” and “factor.” Topic 12, labeled “use of digital health for management of lipid profile,” featured the top 10 keywords “lipid profile,” “patient,” “training,” “optimization,” “level,” “therapy,” “health care,” “coach,” “create,” and “metabolism.”

XSL•FO
RenderX

Table 2. Grouping of topics and high-ranking keywords.

Key words	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	Telehealth	MetS[a]	Device	Obesity	Intervention	COVID-19	Blood pressure	Physical activity	Application	Application	Woman	Lipid profile
2	Diabetes	EHR[b]	Wearable device	Weight loss	Application	Diabetes	Factor	Application	User	Diabetes	Pregnancy	Patent
3	Control	Diabetes	Use	Intervention	Participant	MetS	Group	Intervention	Health	Smartphone	Feedback	Training
4	Intervention	Patients	Sensor	Program	Hospitalization	Patient	Physical activity	Minute	Use	Field	Health	Optimization
5	Self-management	Complication	Technology	Health	Care	Data	Risk	Body	Access	Modification	Pilot	Level
6	Care	Risk	Monitoring	Body	Change	Disease	MetS	Engagement	Care	HbA_{1c}	Intervention	Therapy
7	Effect	Criterion	Data	Application	mHealth[c]	Health	Child	Week	Management	Review	Program	Health care
8	Group	Data	Sit	Weight	Trial	Disorder	Measure	User	Data	Use	Delivery	Coach
9	Type	Population	Disease	Effect	Effect	Incidence	Intervention	Effect	Information	Control	Interview	Create
10	Health	Model	Review	Behavior	Behavior	Decrease	Adult	Woman	Feature	Search	Factor	Metabolism

[a]MetS: metabolic syndrome.

[b]EHR: electronic health records.

[c]mHealth: mobile health.

Topic Trend

Research into MetS within the context of digital health care began in 2003 and showed steady growth over time, with a substantial surge in publications observed after the onset of COVID-19. This trend was particularly notable between 2021 and 2022, when the number of related articles rose sharply from 20 to 43, reflecting heightened interest and activity in this field. In the early years, the focus of research was primarily on telemedicine and electronic medical records. However, the scope of studies has since broadened to include a wider array of topics, such as the development of intervention programs that leverage wearable devices and mobile apps to manage subjects' metabolic status and promote their engagement in health-related behaviors. Moreover, beginning in 2020, research has increasingly targeted specific populations, including women of childbearing age. The emergence of COVID-19 also prompted investigations into the application of digital health care in remote scenarios necessitated by the pandemic (Figure 4).

XSL•FO
RenderX

Figure 4. Annual trend in topics. BP: blood pressure; EHR: electronic health record; MetS: metabolic syndrome; PA: physical activity.

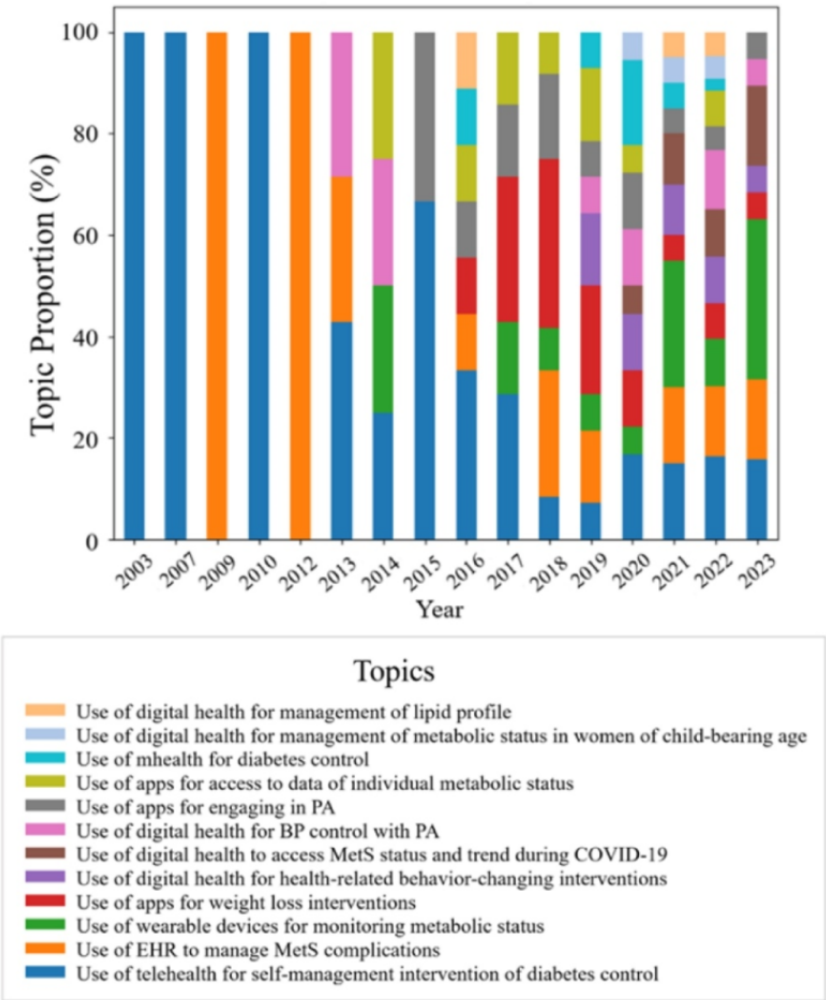

Discussion

Principal Results

Regarding key terms in the digital health care domain related to MetS, the following terms emerged as predominant: “MetS,” “use,” “patient,” “health,” “intervention,” “data,” “risk,” “group,” “increase,” “application,” and “effect.” Given that MetS was the focal point of our inquiry, it unsurprisingly ranked highest in both degree and centrality among the keywords. Our analysis further suggests that the primary research focus in this area involves the collection of data to verify the effectiveness of digital health interventions, particularly those delivered via apps targeting specific patient groups for MetS management. This observation was corroborated by text network analysis (Figures 2 and 3).

Additionally, an analysis of the number of relevant studies published annually revealed a noticeable increase in publications, particularly following the onset of the COVID-19 pandemic. While digital health care has been used for MetS management for several years, the COVID-19 crisis marked a significant turning point, enhancing both its accessibility and applicability. The term “COVID-19” itself became a key focus in topic 6 (Use of digital health to assess MetS status and trends during the COVID-19 pandemic), first emerging in the literature in 2020. Given these developments, we can reasonably anticipate an escalation in research activity surrounding MetS and digital health care applications in the coming years. This is particularly pertinent as the prevalence of MetS continues to rise, necessitating lifestyle alterations to mitigate or reverse such trends [1,25]. Moreover, as the digital health care field expands in tandem with technological advancements [7,26], we expect this upward trajectory to continue.

Furthermore, we observed a significant disparity in the representation of individual topics. Topic 1, which focuses on the use of telehealth for self-management in diabetes control, constitutes 19.8% of the studies. In contrast, topic 11, which explores the use of digital health for managing metabolic status in women of childbearing age, and topic 12, which centers on the use of digital health for lipid profile management, each make up only 2.5%. This distribution suggests that not all topics are equally represented in the literature. However, it is worth noting that topic 11 was first introduced in 2020 and topic 12 in 2016, making them relatively new areas of research. This may account for their lower representation. In comparison, topic 1 has been under consistent since 2003, indicating that the self-management of diabetes remains a long-standing primary concern in this field.

Some topics and their associated keywords align with the burgeoning field of precision medicine. Precision medicine aims

XSL•FO
RenderX

to provide personalized care to each individual by focusing on the individual's genomic structure and molecular biology. In this context, health care providers make medical decisions based on the unique characteristics of each individual rather than generalizing from average patients [27,28]. Given this focus, there is a growing need for intervention studies that target specific subjects, a need that is mirrored by the keywords identified in this study's results. For example, topic 9 (Use of apps for access to data of individual metabolic status) centers on individuals' metabolic status and emerged in the literature in 2014, with research ongoing through 2022. Keywords associated with topic 7 (Use of digital health for BP control with PA) include "child" and "adult," while those for topic 8 (Use of apps for engaging in PA) feature the term "woman." Topic 11 (Use of digital health for management of metabolic status in women of childbearing age) includes keywords like "woman," "pregnancy," and "delivery," reflecting its focus on women who are pregnant or in the process of childbirth. This targeted research contributes incrementally to the developing knowledge base for tailored interventions, which aligns with the broader medical community's emphasis on precision medicine. Nonetheless, given the small proportion of these topics, further research is essential because precision medicine relies on comprehensive population-level data to inform the selection of the most appropriate treatment for a given individual [27].

Limitations

This study presents some limitations that warrant attention. First, the dataset comprises a relatively small number of abstracts, which may impact the generalizability of the findings. To collect articles for use in our analysis, the topic search and category settings were retrieved on academic databases. It is possible that the data collected did not include all papers related to MetS in digital health care. Second, for the sake of clarity in interpretation, the analysis focused solely on nouns. The exclusion of other word types such as adjectives and adverbs may result in differing outcomes compared with analyses that consider all word types.

Conclusions

This study used topic modeling, a data-mining method, to extract salient topics related to MetS and digital health care. By analyzing historical research trends, the study aims to forecast and establish directions for future research endeavors. To this end, we systematically gathered all pertinent data from several academic databases, conducted data cleaning, implemented topic modeling, and undertook visualization and analysis of the results. This research contributes to a more comprehensive understanding of MetS management within the context of digital health care and offers valuable insights for future researchers in this domain.

Acknowledgments

Funding for JK was provided by the National Research Foundation of Korea (NRF) Funding (grant NRF-2022R1A2C1002743).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JK designed and conceptualized the study. JK and YC were responsible for screening the eligibility of the abstracts included in the study. KL conducted data analysis and executed both the text network and topic modeling processes. KL, JK, and YC contributed to the writing and revision of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search Strategy.
[DOCX File , 20 KB - jmir_v26i1e53873_app1.docx]

References

1. Patlar Akbulut F, Ikitimur B, Akan A. Wearable sensor-based evaluation of psychosocial stress in patients with metabolic syndrome. Artif Intell Med 2020;104:101824. [doi: 10.1016/j.artmed.2020.101824] [Medline: 32499003]
2. Greenfield DM, Snowden JA. Cardiovascular diseases and metabolic syndrome. In: Carreras E, Dufour C, Mohty M, Kröger N, editors. The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies. Cham: Springer; 2019:415-420.
3. Expert Panel on Detection, Evaluation,Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). J Am Med Assoc 2001;285(19):2486-2497. [doi: 10.1001/jama.285.19.2486] [Medline: 11368702]

XSL•FO
RenderX

4. Nilsson PM, Tuomilehto J, Rydén L. The metabolic syndrome - what is it and how should it be managed? Eur J Prev Cardiol 2019;26(2_suppl):33-46. [doi: 10.1177/2047487319886404] [Medline: 31766917]
5. Bae C, Kim BS, Cho K, Kim JH, Kim IH, Kim JH. Effects of customized digital health care service on metabolic syndrome status and lifestyle using a health care app: clinical trial. JMIR Form Res 2023;7:e41427 [FREE Full text] [doi: 10.2196/41427] [Medline: 36652290]
6. Iyawa GE, Herselman M, Botha A. Digital health innovation ecosystems: from systematic literature review to conceptual framework. Proc Comput Sci 2016;100:244-252. [doi: 10.1016/j.procs.2016.09.149]
7. Anthony Jnr. B. Integrating telemedicine to support digital health care for the management of COVID-19 pandemic. Int J Healthcare Manage 2021;14(1):280-289. [doi: 10.1080/20479700.2020.1870354]
8. Lugones-Sanchez C, Recio-Rodriguez JI, Agudo-Conde C, Repiso-Gento I, G Adalia E, Ramirez-Manent JI, EVIDENT 3 Investigators. Long-term effectiveness of a smartphone app combined with a smart band on weight loss, physical activity, and caloric intake in a population with overweight and obesity (Evident 3 Study): randomized controlled trial. J Med Internet Res 2022;24(2):e30416 [FREE Full text] [doi: 10.2196/30416] [Medline: 35103609]
9. Kirk MA, Amiri M, Pirbaglou M, Ritvo P. Wearable technology and physical activity behavior change in adults with chronic cardiometabolic disease: a systematic review and meta-analysis. Am J Health Promot 2019;33(5):778-791. [doi: 10.1177/0890117118816278] [Medline: 30586996]
10. Joshi A, Amadi C, Schumer H, Galitzdorfer L, Gaba A. A human centered approach to design a diet app for patients with metabolic syndrome. Mhealth 2019;5:43 [FREE Full text] [doi: 10.21037/mhealth.2019.08.13] [Medline: 31620470]
11. NA. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. In: National Research Council; Division on Earth and Life Studies; Board on Life Sciences; Committee on A Framework for Developing a New Taxonomy of Disease. Washington, DC: The National Academies Press; 2011.
12. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
13. Eland-de Kok P, van Os-Medendorp H, Vergouwe-Meijer A, Bruijnzeel-Koomen C, Ros W. A systematic review of the effects of e-health on chronically ill patients. J Clin Nurs 2011;20(21-22):2997-3010. [doi: 10.1111/j.1365-2702.2011.03743.x] [Medline: 21707807]
14. Sequi-Dominguez I, Alvarez-Bueno C, Martinez-Vizcaino V, Fernandez-Rodriguez R, Del Saz Lara A, Cavero-Redondo I. Effectiveness of mobile health interventions promoting physical activity and lifestyle interventions to reduce cardiovascular risk among individuals with metabolic syndrome: systematic review and meta-analysis. J Med Internet Res 2020;22(8):e17790 [FREE Full text] [doi: 10.2196/17790] [Medline: 32865503]
15. Tully L, Burls A, Sorensen J, El-Moslemany R, O'Malley G. Mobile health for pediatric weight management: systematic scoping review. JMIR Mhealth Uhealth 2020;8(6):e16214 [FREE Full text] [doi: 10.2196/16214] [Medline: 32490849]
16. Armstrong H, McCulloh I, Johnson A. Social Network Analysis With Applications. Hoboken, NJ: John Wiley & Sons; 2013.
17. Scott J. Social Network Analysis. London: Sage; 2017.
18. Park JD. A study on issue tracking on multi-cultural studies using topic modeling. J Kor Soc Lib Inf Sci 2019;53(3):273-289.
19. kisleepublic/mets. GitHub. URL: https://github.com/kisleepublic/mets [accessed 2024-11-12]
20. Sbalchiero S, Eder M. Topic modeling, long texts and the best number of topics. Some problems and solutions. Qual Quant 2020;54(4):1095-1108. [doi: 10.1007/s11135-020-00976-w]
21. Bianchi F, Terragni S, Hovy D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. arXiv:2004.03974 2020. [doi: 10.18653/v1/2021.acl-short.96]
22. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 2022.
23. BERTopic. GitHub. URL: https://maartengr.github.io/BERTopic/index.html [accessed 2024-11-12]
24. SBert. Pretrained Models. SBert.net. URL: https://www.sbert.net/docs/pretrained_models.html [accessed 2024-11-12]
25. Liang X, Or B, Tsoi MF, Cheung CL, Cheung BMY. Prevalence of metabolic syndrome in the United States national health and nutrition examination survey 2011-18. Postgrad Med J 2023;99(1175):985-992. [doi: 10.1093/postmj/qgad008] [Medline: 36906842]
26. Ekman B. Cost analysis of a digital health care model in Sweden. Pharmacoecon Open 2018;2(3):347-354 [FREE Full text] [doi: 10.1007/s41669-017-0059-7] [Medline: 29623633]
27. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
28. Ho D, Quake SR, McCabe ERB, Chng WJ, Chow EK, Ding X, et al. Enabling technologies for personalized and precision medicine. Trends Biotechnol 2020;38(5):497-518 [FREE Full text] [doi: 10.1016/j.tibtech.2019.12.021] [Medline: 31980301]

Abbreviations

API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers

XSL•FO
RenderX

MetS: metabolic syndrome
PA: physical activity

Edited by T de Azevedo Cardoso; submitted 23.10.23; peer-reviewed by TAR Sure, M Ghayour-mobarhan, EJ Bae, J An; comments to author 10.07.24; revised version received 21.08.24; accepted 31.10.24; published 12.12.24.

Please cite as:
Lee K, Chung Y, Kim JS
Research Trends on Metabolic Syndrome in Digital Health Care Using Topic Modeling: Systematic Search of Abstracts
J Med Internet Res 2024;26:e53873
URL: https://www.jmir.org/2024/1/e53873
doi:10.2196/53873
PMID:39666378

©Kiseong Lee, Yoongi Chung, Ji-Su Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Patient and Public Willingness to Share Personal Health Data for Third-Party or Secondary Uses: Systematic Review

Rebecca Baines[1], PhD; Sebastian Stevens[1,2], PhD; Daniela Austin[1], MSc; Krithika Anil[3], PhD; Hannah Bradwell[1], PhD; Leonie Cooper[1], PhD; Inocencio Daniel Maramba[1], MSc; Arunangsu Chatterjee[1,4], PhD; Simon Leigh[2,5], PhD

[1]Centre for Health Technology, University of Plymouth, Plymouth, United Kingdom
[2]Prometheus Health Technologies Ltd, Newquay, United Kingdom
[3]University of Plymouth, Plymouth, United Kingdom
[4]School of Medicine, University of Leeds, Leeds, United Kingdom
[5]Warwick Medical School, University of Warwick, Conventry, United Kingdom

Corresponding Author:
Rebecca Baines, PhD
Centre for Health Technology
University of Plymouth
Portland Square, Drakes Circus
Plymouth, PL4 8AA
United Kingdom
Phone: 44 07508916450
Email: rebecca.baines@plymouth.ac.uk

Abstract

Background: International advances in information communication, eHealth, and other digital health technologies have led to significant expansions in the collection and analysis of personal health data. However, following a series of high-profile data sharing scandals and the emergence of COVID-19, critical exploration of public willingness to share personal health data remains limited, particularly for third-party or secondary uses.

Objective: This systematic review aims to explore factors that affect public willingness to share personal health data for third-party or secondary uses.

Methods: A systematic search of 6 databases (MEDLINE, Embase, PsycINFO, CINAHL, Scopus, and SocINDEX) was conducted with review findings analyzed using inductive-thematic analysis and synthesized using a narrative approach.

Results: Of the 13,949 papers identified, 135 were included. Factors most commonly identified as a barrier to data sharing from a public perspective included data privacy, security, and management concerns. Other factors found to influence willingness to share personal health data included the type of data being collected (ie, perceived sensitivity); the type of user requesting their data to be shared, including their perceived motivation, profit prioritization, and ability to directly impact patient care; trust in the data user, as well as in associated processes, often established through individual choice and control over what data are shared with whom, when, and for how long, supported by appropriate models of dynamic consent; the presence of a feedback loop; and clearly articulated benefits or issue relevance including valued incentivization and compensation at both an individual and collective or societal level.

Conclusions: There is general, yet conditional public support for sharing personal health data for third-party or secondary use. Clarity, transparency, and individual control over who has access to what data, when, and for how long are widely regarded as essential prerequisites for public data sharing support. Individual levels of control and choice need to operate within the auspices of assured data privacy and security processes, underpinned by dynamic and responsive models of consent that prioritize individual or collective benefits over and above commercial gain. Failure to understand, design, and refine data sharing approaches in response to changeable patient preferences will only jeopardize the tangible benefits of data sharing practices being fully realized.

(J Med Internet Res 2024;26:e50421) doi:10.2196/50421

KEYWORDS

data sharing; personal health data; patient; public attitudes; systematic review; secondary use; third party; willingness to share; data privacy and security

XSL•FO
RenderX

Introduction

International advances in information communication, eHealth, and other digital health technologies have led to significant expansions in the collection and analysis of personal health data [1]. While the benefits of data sharing are widely recognized, including improved care quality [2-4], strengthened care coordination [3], reduced medical costs [4], and enhanced treatment development [5], critical exploration of public attitudes, perceptions, and preferences remains limited [1,3,4,6,7]. Furthermore, as reported by Braunack-Mayer et al [8], a specific focus on public preferences for sharing personal health data with third parties or for secondary uses (ie, other than their health care) remains sparse [9].

Critical exploration of such perceptions is considered vital as the optimization and successful execution of data sharing relies on public support [4,10]. Furthermore, to advance existing knowledge and understanding, health care systems are increasingly partnering with third-party organizations to provide novel data processing and analysis capabilities [2]. Moreover, several high-profile data sharing scandals, including the Cambridge Analytica scandal [11], period tracking app scandals [12], and other scandals related to the National Health System [13] have arguably heightened public concern.

Developing a sound understanding of public willingness to share personal health data with third-party organizations or for secondary uses is therefore essential if we are to leverage the full potential of data sharing practices and inform international policy [14]. In addition, following the COVID-19 pandemic and possible related changes in public willingness to share health data, an updated synthesis of available literature is urgently required [15].

Thus, while existing research has often focused on data sharing in a single area, such as research [9,16], health administrative or clinical trial data [17], or electronic health records (EHRs) [7,18], this review aims to synthesize previously siloed literature into a single corpus of information, identifying barriers and enablers that affect public willingness to share personal health data for third party or secondary uses.

Methods

Overview

A systematic review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis; Multimedia Appendix 1 [19]) guidelines [20] and guidance by Popay et al [21] on the conduct of narrative synthesis in systematic reviews was conducted. This systematic review sought to address the following review question: What factors affect public willingness to share personal health data for third-party or secondary uses? A review protocol was not registered, but the review was assessed by an information specialist according to PRESS (Peer Review of Electronic Search Strategies) guidance [22].

Search Strategy

Following the PRESS guidance [22], the search strategy was informed and approved by an information specialist. Six databases were searched: MEDLINE, Embase, PsycINFO, CINAHL, Scopus, and SocINDEX. Search terms used (Table 1) were designed to maximize sensitivity and specificity using the SPICE (Setting, Population [or perspective], Intervention, Comparator, and Evaluation) framework [23]. Database searches (conducted August 11, 2021) were also supplemented with reference list searches to ensure sufficient coverage.

Table 1. Search terms used organized according to the SPICE (Setting, Population [or perspective], Intervention, Comparator, and Evaluation) framework.

	Search terms
Setting	"health data" OR "personal health data" OR "health information exchange*" OR "wearable* data" OR "patient data" OR "personal data" OR "personal health information" OR "patient health record*" OR "patient record*" "electronic health record*" OR "personal health record*"
Perspective	consumer* OR patient* OR client* OR citizen* OR carer* OR user* OR "end user" OR public* OR stakeholder* OR service-user* OR "service user*"
Intervention	[share or sharing or transfer* or access* or "secondary us*" or link] adj3 [data or record* or information]
Comparison	N/A[a]
Evaluation	attitude* OR perspective* OR willing* OR perception* OR barrier* OR enable* OR facilitat* OR opinion* OR trust* OR confiden* OR concern* OR view* OR challenge*OR accept* OR qualitative OR interview* OR survey*

[a]N/A: not applicable.

Study Selection

Studies were selected through a 2-stage process. First, due to the large number of abstracts returned, studies were equally split among 5 reviewers (RB, SS, DA, KA, and HB; n=201, 20% each) who independently reviewed returned abstracts for study inclusion using a collaboratively agreed decision-making flowchart and predefined inclusion and exclusion criteria outlined below. To ensure consistency and decision-making rigor, 10% of each reviewer's abstracts were randomly selected to be blindly assessed and compared by a second reviewer. This process was facilitated using Rayyan (Rayyan Systems Inc), a web-based app for systematic reviews. If an inclusion decision could not be made from the abstract alone, the paper was taken through to full text. Papers included in this first stage were then read in full and independently assessed for study inclusion by the research team. Reasons for study exclusion were recorded during this process for transparency purposes. If discrepancies in inclusion or exclusion decisions could not be resolved by discussion alone, this would have been resolved by being sent

XSL•FO
RenderX

to a third reviewer until consensus was achieved. However, this process was not required in this review.

Inclusion Criteria

Papers published in the English language between 2011 and 2021 that explored patient and public attitudes toward sharing personal health data with third-party organizations or for secondary use, using any study design except protocols, conference proceedings, letters, or theses were included. Justification for the data parameters used stems from the rapidly evolving nature of technology acceptance, data sharing policies [24], and the desire to ensure only the most contemporary information was included. For clarity, if a paper discussed views from patients and the public and other related stakeholders such as researchers, the paper was included if the source from which the data were being presented was clear.

Exclusion Criteria

Protocols, conference proceedings, letters, or theses, not available in the English language, published prior to 2011 that did not explore patient and public attitudes toward sharing personal health data were excluded. Studies that discussed patient and public attitudes toward sharing personal health data by proxy, for example, from the perspective of health care professionals alone were also excluded due to the study's focus on understanding public attitudes toward sharing personal health data and not related assumptions or perspectives. If it was not possible to identify the source from which the data were being presented, that is, patient or health care professional, the paper was excluded. Due to limited resources, a sensitive interpretation of non-English papers could not be guaranteed. Non-English papers were therefore excluded, even though this may have introduced a risk of bias.

Data Extraction

To ensure consistency and standardization, 6 reviewers independently undertook data screening and extraction using a piloted data extraction form. The information extracted from the papers encompassed details such as author names, publication dates, study locations, study designs, data types, recipients of the data (whether individuals or organizations), as well as factors influencing the sharing of personal health data. For clarity, only information that was clearly identified as being from a patient or public perspective was extracted. No risk of bias assessments were undertaken due to the size of the literature reviewed.

Data Analysis and Synthesis

Review findings were analyzed using inductive thematic analysis as proposed by Braun and Clarke [25]. Defined as a method for identifying, analyzing, and interpreting patterns of meaning (themes), inductive thematic analysis embodies an organic approach to coding and theme development [25,26]. Justification for adopting an inductive thematic approach includes its provision of an accessible and systematic procedure to generating codes and themes [26], emphasis on producing rigorous analyses, and providing flexibility in identifying patterns within and across multiple data types [26].

Identified themes were synthesized using a narrative approach defined as "an approach to the systematic review and synthesis of findings from multiple studies that relies primarily on the use of words and text to summarize and explain findings of the synthesis" according to the guidance mentioned in Popay et al [21]. Other systematic reviews have adopted a similar approach to data analysis and synthesis [27,28], justifying its inclusion in this review. Simple tables of themes have also been created to summarize findings wherever possible. No sensitivity or certainty assessments were conducted.

Results

Overview

Of the 13,949 papers originally identified, 135 were included (Figure 1). Included papers discussed the opinions of 164,478 patients or public members and reviewed 173 papers. Most papers were published in the past 5 years from the United States (n=64), the United Kingdom (n=18), and Canada (n=9), with representation from 17 other countries and continents including Ghana, Taiwan, Japan, Egypt, and Australia.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis).

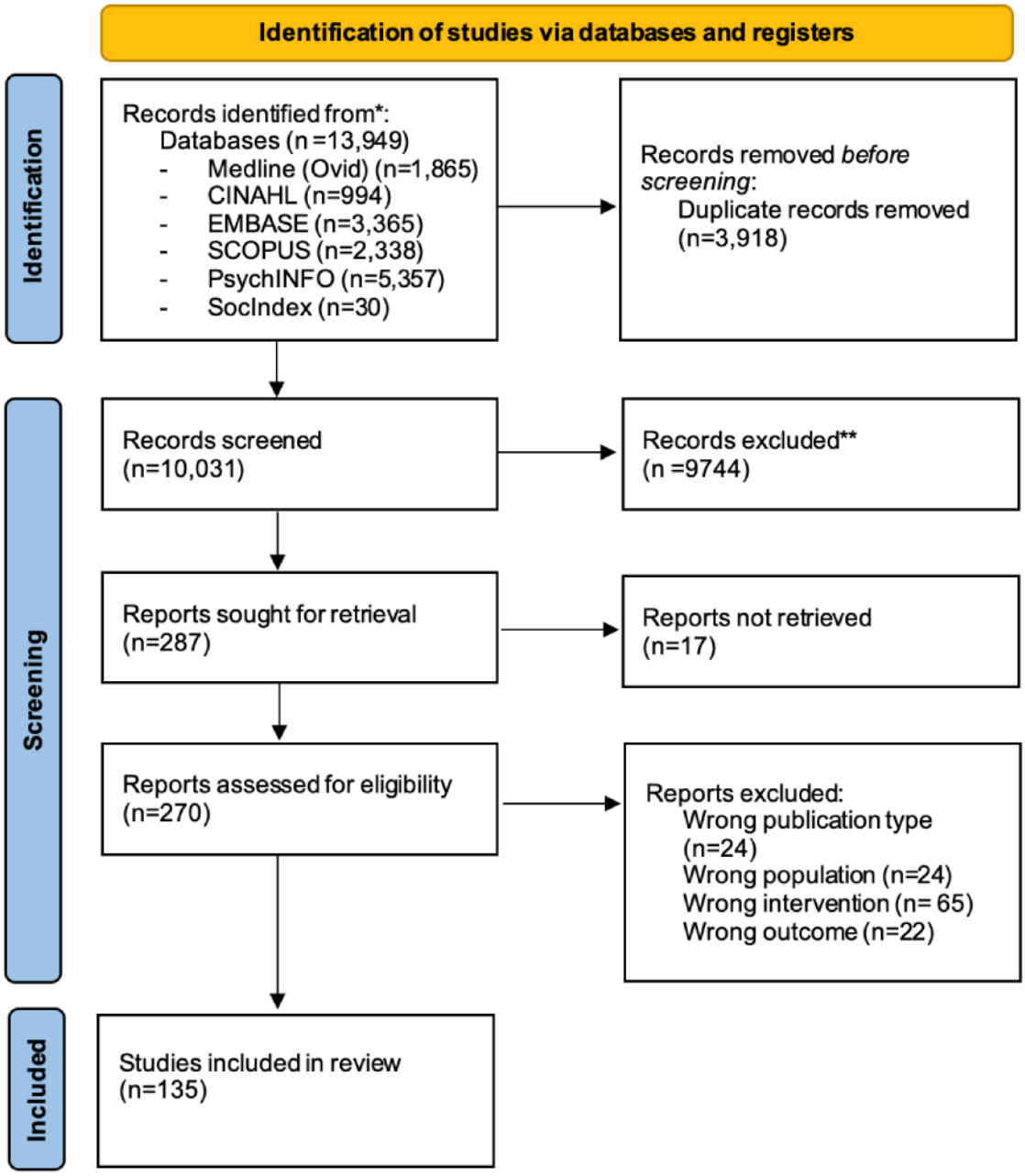

As shown in Multimedia Appendix 2 [2-4,6,7,9,10,12,14-18,24,29-147] and Multimedia Appendix 3, most papers were of a survey (n=68) or qualitative (n=41) design, with mixed methods (n=12), reviews (n=6), citizen juries (n=5), discrete choice experiments (n=1), randomized controlled trial (n=1), and experimental (n=1) study designs also used. Most papers discussed a willingness to share personal health data (n=52), EHRs (n=31), or genomic or genetic data (n=12), with either researchers (n=61), health care providers (n=27), or multiple stakeholders combined (n=38). Other types of data reviewed included app data (n=2), monitoring data (n=2), adverse drug events (n=2), and personal health data voluntarily shared digitally (n=1).

The majority of papers explored willingness to share data from a patient (n=47), patient and public (n=48), or public (n=26) perspective, with patients and caregivers (n=5), parents (n=1), next of kin (n=1), employees (n=1), and insurance customers (n=1) also represented (Multimedia Appendix 3). For clarity, only data pertaining to patients (inclusive of parents and caregivers) and public were extracted. Of the included papers reviewed, 2 had a focus on rare diseases. A table of study characteristics is provided in Multimedia Appendix 2 [2-4,6,7,9,10,12,14-18,24,29-147] with further summary tables also available in Multimedia Appendix 3.

Reflecting the review's aim, the results are synthesized according to key themes that both support (enablers) and inhibit (barriers) public willingness to share personal health data for third-party or secondary uses.

Barriers

Data Protection

Beginning with barriers, perhaps unsurprisingly, data privacy, security, and management concerns were the barriers most

XSL•FO
RenderX

commonly identified as affecting public willingness to share personal health data [3,4,9,14,16-18,29-55,148]. Such findings are concerning, as several studies highlighted the detrimental outcome of patients intentionally engaging in privacy-protective behaviors including withholding clinically relevant information from health care providers to counteract data privacy and security concerns [48,56,57]. For example, one study stated the following:

> *If individual needs for privacy and security of data exchange are not met, consumers may intentionally prefer to hide relevant health information from their healthcare providers.* [56]

In some studies, data privacy was the most influential factor affecting willingness to share [4,51,52,58], highlighting the severity of the issue at hand.

Specific data privacy and security concerns identified included confidentiality breaches [16,59] often linked to the risk of reidentification [16,60], or ineffective anonymity processes [55,61]; unauthorized, or unknown data access [18,39,42,44-46,51,62,63]; data misuse and abuse [3,15,16,34,47,55,59,63,64], particularly for stigmatizing, or sensitive health conditions [45,47,55]; data or identity theft and fraud [7,15,16,18,31,39,44,58,63-67]; and the unauthorized reuse or future use of collected data [15,16,18,42,59,60,65,68] that extends beyond the scope of originally intended and consented purposes [59,68]. The latter point indicates how data sharing practices often operate in "largely unchartered territory and, as such, new harms may emerge that we cannot yet foresee" [65], accentuating the importance of dynamic consent that enables progressive patient choice as later discussed [47].

Concerns that data sharing activities may adversely affect an individual or community [2,15,49,59,60,67] also appear to have a significant impact on public willingness to share personal health data. Several papers discussed the risk of exploitation [3,16,44,60,65], inequalities [3,44]; "fear of discrimination" [18] or stigmatization [3,15,16,18,30,44,47,60,65-67,69-71]; financial hardship including insurance premiums being increased or denied [18,59,64-67,71-73]; reduction in care quality [31] or person-centered care [3]; employers having oversight of personal information [58,59,64,74]; and feelings of judgment, shame, or guilt [14,75]. Concerns of adverse consequences appeared particularly salient when being asked to share more sensitive or stigmatizing data [24,47,59]. However, as later described, what constitutes sensitive data is largely dependent on individual perceptions, definitions, and experience.

Concerningly, confidence in existing data security systems and processes appears low [44,58,64,76,77]. For example, only 38.4% (n=19,372) of respondents agreed that health care providers are currently successful in providing effective data security [77]. The majority of US citizens do not trust organizations that store and share their personal health data [76]. The size and bureaucracy of an organization have also been linked to a perceived inability to protect personal health data [36,44], with 71.3% (n=1969) of participants stating that they could not fully trust the "big and bureaucratic" National Health System to guarantee the security of EHRs [44]. Size and status do not therefore appear to guarantee public trust and confidence. Alternative ways of nurturing public trust that align with public preferences and expectations may therefore be required, as health care systems and organizations enter a rapidly evolving era of expanded data sharing [76].

Public Awareness and Understanding

Other factors that undermine public willingness to share personal health data include limited public awareness and understanding [16,18,29,51,53,59,60,71,78-84]. Studies often reported that most patients feel they are not receiving the information required to understand data sharing opportunities [85] and thus provide informed consent. Participants also seemed to be unaware of the meaning and value of their health data to third-party organizations [86]. However, O'Brien et al [58] found that of the 693 patients who were not comfortable with sharing deidentified data, most reported that their comfort levels would increase if they were made aware and understood how their health data were protected [58]. Such findings reiterate the combined effect of increased public awareness and understanding and assurances of data privacy and security in facilitating public support.

Data Type

In addition to awareness and understanding, willingness to share personal health data also appears dependent on the type of data being shared [18,24,34,61,70], in particular, whether data are considered "sensitive" or not. While sensitive data are often clearly defined in privacy regulations such as the General Data Protection Regulation, participant definitions of sensitive data often differed between and within studies. Types of data identified by participants as particularly sensitive or stigmatizing are listed in Table 2.

XSL•FO
RenderX

Table 2. Identified types of sensitive or stigmatizing data.

Data type	Supporting references
Mental health	[18,32,33,43,59,64,67,77,87]
Sexual health including reproductive health	[18,32,33,43,59,77,88]
Genetic or genomic data	[15,16,33,66,67]
Substance use	[18,33,77,87]
Social security number and insurance ID	[51,58,89]
Financial information	[32,34,51]
Sexuality	[59]
Religion	[59]
Disability status	[64,66]

However, while similar categorizations of sensitive data types exist, diversity in patient-defined classifications and data sharing preferences persist [87], particularly regarding genetic and genomic data [18,70,90]. For example, research conducted on rare disease perspectives concludes that roughly half of individuals consider information on their disability (51/47), genetics (49/48), or physiology (48/50) to be sensitive. However, this also means that half of the respondents did not consider such information to be sensitive [66]. Furthermore, in a study exploring data sharing among children with genetic conditions, parents of typically developing children reported being uncomfortable sharing sensitive information and in some cases, completely unwilling to do so. Conversely, parents of children with Fragile X syndrome and autism spectrum disorder were willing to share their child's sensitive data including mental health and genetic information [67]. Similar variations in data sharing preferences depending on rare disease or health status results have also been reported in cancer, where patients with cancer reportedly more willing to share sensitive genetic data [91]. While the direction of this effect was opposite to that originally hypothesized, it was the inherited genetic information that often made participants with cancer more willing to share such data [91]. Thus, emerging evidence suggests individuals with inherited or rare diseases may be more inclined to share personal health data to advance existing knowledge and treatment options, with the term "informational altruists" often used [90]. A universal definition of sensitive and nonsensitive data is therefore not available given variations in peoples' perspectives, health status, and experience.

Type of Data User

Overview

In addition to data type, the type of user requesting data to be shared also appears highly influential in determining public willingness to share personal health data [82,92,148,149]. Most patients appear more willing to share health data with health care providers or organizations [24,32,92-95], researchers [32,67,70,96-98,149], including researchers at universities [14,18,32,40,55,67,70,92,99-101], although not always [59,77], researchers who identify as medical professionals [10,59,92,95] and nonprofit organizations [18,40,67,70], as they trust such entities to "produce socially valuable knowledge, products and interventions" [55] that are often meaningfully constrained by ethics and ethics boards [67].

Conversely, patients appear less willing to share personal health data with government bodies [18,59,102], private insurance companies [16,32,34,36,44,54,67,72,77,84,92,95,100] (often linked to a fear of being denied coverage [67,72,101]), pharmaceutical companies [14,16,18,40,67,68,77,84,92,96,103], and commercial or private entities [16,32,36,40,44,47,51,59,60,67,69,70,97,100,102,104-108]. However, it is often not a straightforward delineation between data user type and public willingness to share. Such relationships are often complex, dynamic, and highly variable [17,34,41,55,71,76,82,96,109,110], depending on personal attributes or personality traits, the purpose, context, perceived benefit, incentivization, and compensation of data sharing requests. As Aitken et al [59] mentioned, "there does not appear to be a clear, or static hierarchy of trusted organisations/sectors." Understanding and responding to public preferences for who, when, how, and under what circumstances they share their personal health data are therefore imperative.

Perceived Motivation of Data User

A common factor underpinning public willingness to share personal health data is the perceived motivation of the data user type [14,61,96], with the end purpose, or intended use considered to be highly influential in determining willingness to share personal health data with a third party [14,96]. This is perhaps most evident in data sharing with pharmaceutical, insurance, or commercial companies [93,96]. Specifically, the perceived prioritization of profit [14,16,54,59] over individual or collective health and well-being [59]. For example, recent research conducted by Jagsi et al [111] reported that when insurance companies were the data receiver, 79.5% (n=171) of participants felt comfortable if the purpose was to ensure patients received recommended care. However, this figure dropped to 50.9% (n=110) if the data were being used to determine insurance eligibility or reimbursement [111]. Similarly, when drug companies were the intended data users, most participants were comfortable if the information was being used to help develop new treatments (n=200, 92.6%) or understand which patients may benefit from certain drugs (n=197, 92.1%) [111].

Included papers suggest patients are increasingly recognizing the balance of working with pharmaceutical and private

XSL•FO
RenderX

companies [59]. Mamo et al [54] mentioned that participants were supportive of "corporate profits if the company developed new or modified medical technologies that would help people and if such help could be distributed equitably." Profit creation does not therefore necessarily dictate the automatic refusal of private sector involvement [2,59,107,112]. Profit creation can be seen as acceptable under certain conditions [2,107,112]. Namely, when the intended purpose is of clear public benefit [107,112] (a notion referred to as the "public benefit criterion" [107]), there is a transparent commitment to sharing or appropriately reinvesting profits to create further public or societal benefits [59,112], and any commercial gain that is accrued is secondary to ensuring public benefit [107], that is, public benefits are prioritized above profit [112]. Thus, social responsibility, evidence of equitable and reciprocal benefits, and the prioritization of public benefit over commercial gain are widely regarded as an essential prerequisite for public support [59].

Distance Between Provider and Data User

Finally, linked to perceived motivation is the perceived distance between the data requester and their ability to directly impact patient care. The further data recipients appear from having a direct impact on care, the less willing patients are to share their personal health data, often under the inference that there is no clear reason why individuals need access to such data [95]. Clearly articulating the purpose behind data sharing exercises, particularly when there is a significant "distance" between data requester and patient impact, also seems imperative.

Facilitators

Personal Attributes or Personality Traits

With regard to facilitators, several personal attributes or personality traits appear to support a willingness to share personal health data. Those most frequently identified are shown in Table 3.

Table 3. Personal attributes and personality traits that support a willingness to share personal health data.

Personal attributes and personality traits	Supporting references
Age: Although opinion appears divided on the direction of influence, with some researchers suggesting older individuals are less willing to share, others suggest they are more willing to share, with some authors reporting no difference in age and willingness to share	[2,15,16,32,33,40,44,58,60-62,66,70,76,97,99,100,102,113-117,149,150]
Educational attainment: Evidence to suggest people with a lower level of educational attainment are typically less willing to share health data, although this conclusion is again not unanimous, with evidence to suggest the impact of educational attainment may also be dependent on the purpose of any data sharing activities	[2,15,16,32,33,44,48,58,64,66,70,100,101,109,114,118-120,150]
Sex: Differing conclusions on whether men are less or more concerned than women about data sharing	[16,32,60-62,64,66,88,100,102,113,114,150]
Ethnicity: Black participants are often less willing to share personal health data than White participants	[16,44,76,90,96,100,101,109,114,118,119]
Health status: Severity and rare disease, with healthier individuals typically more willing to share personal health data	[15,40,58,62,66,113]
Issue involvement or perceived level of personal relevance and benefit: Higher issue involvement or personal relevance typically yields greater support	[4,6,7,31,39,44,49,51,52,67,69,105,115,120-123]
Privacy concerns including data misuse and fear of harm: Fewer privacy concerns associated with greater intentions to share	[2,6,40,44,52,62,63,113,115,120,122]
Religious beliefs: Lower religiosity is often associated with a greater willingness to share	[15,16,102,109]
Employment status: Being employed is often associated with a greater willingness to share	[33,61,64,76,115,150]
Household income	[35,63,115,120]
Internet or computer access and smartphone ownership or familiarity (sometimes referred to as digital health literacy): Individuals more comfortable using digital technologies are typically more willing to share personal health data	[7,35,52,62,63,100,115,124]

Other factors identified as influential included the presence of altruistic traits [49,58,120]; generalized trust [49,120] or high levels of trust in health care providers [6,7,120,122]; health service use, with more frequent users more likely to be supportive of data sharing [49,114]; patient activation [4]; patient-physician relationship [4]; nationality [15,33]; health insurance coverage [120], with individuals experiencing cost barriers to care typically less willing to share personal health data [96]; marital status [15,16,63]; parental status [15]; child health status [63], with parents of children in excellent-to-good health status less likely to share health data [63]; existing knowledge or personal experience of data sharing topic [15,100], particularly in relation to genetics or genomics [10,71,102]

XSL•FO
RenderX

(although not always [70]); and previous experience of research participation [58,89].

Often several personal attributes affected data sharing willingness simultaneously. For example, evidence suggests that Black, Asian, and minority ethnic populations [90,100,114], younger age groups [100], and those with lower education attainment [100] are generally less supportive of data sharing. However, it is important to note this is not always the case [44,66], with other factors identified including data user type and purpose also influencing intentions to share, highlighting the complexity of data sharing intentions and the subsequent importance of understanding and aligning data sharing requests with public expectations and requirements.

Clear Demonstration of Data Sharing Benefits

Another key factor in supporting data sharing willingness is the clear demonstration of personal, public, or knowledge development benefits. Table 4 outlines the potential benefits of sharing health data ordered according to the frequency in which they were reported in the included literature.

Table 4. Potential benefits of sharing personal health data ordered according to reporting frequency.

Reported benefit of data sharing	Supporting references
Personal health benefits include helping health care professionals make better decisions about their health through increased awareness and understanding	[4,9,14,16,18,29,39,44,46,52,54,58,59,67,83,106,115,122]
Contributing to the "greater good" or altruistic sharing for public benefit	[9,14,16,29,32,48,55,59,63,67,74,105,106,122,125]
Improved care quality, including enhanced treatment options or quality, personalized care provision, improved diagnosis accuracy, and speed	[16,18,41,42,48,52,54,77,115]
Facilitating innovation development, particularly making new therapies available quicker	[16,58]
Reducing resource or research wastage	[16,44]
Increased access to information including sharing of research findings	[41,105]
Reducing treatment delays and medical errors	[41,77]
Preventing health epidemics	[16]
Enabling the study of rare events	[16]
Improving research quality	[32]

Some research provides additional insights into which benefits are considered most desirable from a patient perspective. For example, research conducted by O'Brien et al [58] suggests that the highest proportion (n=3305, 94%) of patients stated that "Helping my doctor make better decisions about my health" and "Helping make new therapies available faster" were "extremely" or "very important" benefits associated with data sharing. "Helping researchers evaluate the quality of care delivered by doctors and hospitals," "Helping other patients with my primary health condition," and "Reducing the cost of doing research" were also rated as "extremely" or "very important" by the majority of patients (91% and 89%, respectively) [58]. Furthermore, patients reported the strongest level of statement agreement with "knowing the study using my data could help patients with my health condition" (mean summary score 1.44) and "knowing the study using my data could reduce health disparities" (mean summary score 1.54) [58]. Patients were less likely to agree that comfort levels would be improved by "knowing the study using my data could help patients with other health conditions" [58]. Similar results have been reported elsewhere [16,18], reiterating the importance of issue involvement or personal relevance previously discussed, suggesting data sharing intentions may increase if there is a direct association between data sharing requests and assured advances in knowledge, treatment quality, and options for the health condition individuals experience. However, it is also important to note that many individuals also appear motivated by the concept of contributing to the "greater good" or altruistic sharing for public benefit [9,14,16,29,32,48,55,59,63,67,74,105,106,122,125]. Therefore, highlighting the purported benefits of sharing health data on both individual and collective levels may be beneficial.

Trust (Specifically, How to Build It)

Overview

Underpinning many of the barriers and enablers listed above is trust, something that can be earned and irreparably damaged [112]. Many papers described the importance of trust in facilitating the sharing of health data [20,45,47,48,51,54,55,84,95,112,120,123,126,149]. However, few studies specifically described how to establish it. The following subthemes therefore report on available literature that supports the development and maintenance of trust.

Control

A core feature of trust in data sharing processes is control [16,18,32,55,66,73,74,79,84,86,92,104,127,128]. Specifically control over what data are shared, with whom, and when, with the majority of participants "overwhelmingly in favour of keeping the strictest control on their data" [66], believing "they have a right to control the use of their data" [127]. For example, in 1 study, all participants (100%) stated that they would like to be able to control what entities access their EHR [79]. Perceived control also appears to lower privacy concerns [59,62,69,104], with clear public preferences for controlling the type, duration (eg, not for the duration, or any longer than the research study [43]), and level of access depending on different data user types or individuals or organizations and data sharing

XSL•FO
RenderX

purposes [18,44,73]. Thus, as noted by Courbier et al [66], being in favor of data sharing practices and wanting more control are not contradictory but rather parallel requirements.

As alluded to below, granular or hierarchical control [79,86,104] is likely to "become more and more important in the practice of healthcare technology and services" [88]. However, it is also important to acknowledge variations in desired levels of control [79]. Adopting a personalized approach to data sharing control is therefore required. Without patients feeling in control of their data sharing practices, the likelihood of nurturing and sustaining data sharing intentions is minimal.

Choice

In addition to control, individuals also require choice in how, what, and when their data are both accessed and shared [16,24,70,79,88,104]. Both elements (choice and control) are considered "crucial for the public to be able to decide who and how to trust" [129]. However, many participants do not yet feel that they have control and choice over their own data, including sensitive data such as biomedical data [129]. Caine et al [79] provided six implications for the design of a patient-centered tool that enables individual choice in the disclosure of personal health data: (1) easy patient access, (2) an overview of current sharing permission, (3) granular, hierarchical control over data access, (4) access controls based on dates, (5) contextual privacy controls, and (6) notification of when their data have been accessed [79].

Consent

Further strengthening the relationship between trust and choice is consent [18,127], with consent often acting as a mechanism for facilitating individual control [59]. While consensus on public preferences for models of consent (eg, explicit opt-in, opt-out (often least favorable), broad, 1-time consent, project specific consent) is not yet available [54,56,119], varied, dynamic, or flexible consent [47,59,66,67] that enable individuals to change their consent preferences over time appears most desirable, recognizing public preferences can often change in response to events, project purposes, motivations, and incentivization [59,78]. Dynamic models of consent may also provide beneficial flexibility and adaptability in response to future technological and regulatory, legal changes [66]. Appropriate models of consent appear particularly important in relation to identifiable data, qualitative information, genetic research, or when a commercial entity is involved [59,98], mirroring previous barriers identified above.

Modernizing consent policies to better reflect changing public preferences is therefore also considered crucial in facilitating transparency and fostering sustained patient engagement and empowerment [90,97].

Feedback Loop

Other mechanisms for establishing trust and transparency include the provision of a feedback loop [47,67]. Such practice was often seen as a form of "compensation" or expression of appreciation and recognition [51]. However, the creation of a feedback loop can be onerous and is often dependent on the use of identifiable data. Despite this, relaying the impact of data sharing is considered "a matter of respect" [67] and a highly desirable way [15,88,93,97,104] to "improve trust and public engagement" [47]. In 1 study, 99.7% of participants positively responded that they would like to be informed about the outcome of a data sharing project they participated in [15], with the ideal frequency of being informed for the majority of respondents being once a month [66]. Developing a clear feedback loop was therefore repeatedly identified as a way to encourage and support data sharing practices [71].

Assurances of Patient Confidentiality and Data Protection

Other ways to develop trust and subsequent data sharing support include providing assurances of data privacy, security, and confidentiality [10,17,55,59,64,74,83,126,130], with assurances of confidentiality often associated with aggregated, or trust in the data's anonymization processes [16,33,47,58,59,61,63,64,113,148]. The influence of governance and safeguarding assurances on public acceptance should not be underestimated [20,43,44,59,116]. Existing literature suggests that it is insufficient to simply indicate that participant data will be kept securely. Specific information regarding what data are being collected, where the data will be stored, which safeguards are applied, who has access to the information, and how data breaches or misuses will be addressed [55,67,71,99,103] is required. Informing the public about security practices and processes can help reassure patients that data sharing is being responsibly executed in a framework of accountability [54,67,112]. More transparency and detailed information regarding data usage and protection are therefore urgently required, ensuring such information is included in consenting content.

Transparency

Overview

In addition, providing transparent information about who will benefit from data access has also been identified as the most important measure to increase trust, having been endorsed by more than 50% of participants across 20 of the 22 countries studied, followed by transparency about who is using data and why [123].

Incentives and Compensation

Finally, the provision of meaningful incentives and compensation [10,51,72,86,93,105,121,148,149] "broadly defined to include financial compensation, expressions of appreciation...recognition and sharing of results" [51] also appears integral in fostering public support. Recent research conducted by Luo et al [14] identified the following types of compensation: free treatment, money, food, cryptocurrencies, discounts on health insurance, shared research findings, and donations to a good cause, with free treatment receiving the most mentions, followed by cryptocurrencies and money. Furthermore, as noted by Navarro‐Millán et al [124], self-tracking technologies and data sharing requests may be more appealing if "coupled with opportunities to learn" about specific health issues including symptom management, medications, side effects, and opportunities to obtain social support. Perhaps unsurprisingly, the majority of participants report that they would be "more" or "much more" likely to share data if they felt appropriately compensated for their actions

XSL•FO
RenderX

[105]. For some, individuals are also incentivized by the idea of benefiting themselves or their community [121], reflecting altruistic motivations [47,51,59,67,74,82,149] previously described. However, appropriate compensation (recognizing this looks different for different people) appeared particularly influential if the third party, particularly pharmaceutical companies were making money from participant data [14], with resulting profits expected to be returned to the public or patient [108]. Providing valued, reciprocal, and equitable benefits therefore appears influential in facilitating data sharing intentions [72].

Discussion

Summary of Findings

This review sought to explore factors that affect public willingness to share personal health data for third-party or secondary uses. Perhaps unsurprisingly, data privacy, security, and management concerns were most commonly identified as the biggest barriers to sharing personal health data from a public perspective. Other factors found to influence public willingness include the type of data being collected (ie, data sensitivity); the type of data user requesting data to be shared, including their perceived motivation, profit prioritization, and ability to directly impact patient care; personal attributes or personality traits; trust established through individual choice and control over what data is shared with who, when, and for how long, supported by appropriate models of dynamic consent; the presence of a feedback loop, and evidence of perceived benefits or issue relevance including valued incentivization or compensation at both an individual and collective or societal level.

Profiting from data sharing does not necessarily dictate public rejection. Namely, when the intended purpose of sharing data is of clear public benefit, there is a transparent commitment to sharing data or appropriate reinvestment of profits to create further public or societal benefits, and any commercial gain is secondary to ensuring public benefit, concerns appear reduced but not entirely removed. Thus, social responsibility, evidence of equitable and reciprocal benefits, and the prioritization of public benefit over commercial gain are widely regarded as an essential prerequisite for public support.

Conversely, universal agreement on patient definitions of sensitive data is not yet available, with variation in patient data sharing preferences also apparent including the perceived direction of influence regarding certain personality traits or personal attributes. Such findings reiterate the complex context data sharing practices often operate within and the subsequent need for more dynamic and responsive models of consent that provide individual choice and control over who has access to what data, when, and for how long. Without patients feeling in control of their data sharing practices, the likelihood of nurturing and sustaining data sharing intentions is minimal.

Comparison With Existing Research

Many of the factors identified as influential in this review including data privacy and security, data, and user type have been reported in previous literature [16,131,151]. Thus, previous conclusions of widespread, yet conditional public support for sharing health data appear to remain true [16]. There was also no reported difference in themes identified in pre- and post–COVID-19 publications. Further exploration of such comparisons may be beneficial moving forward as post–COVID-19 literature continues to emerge and develop. However, this review provides previously unavailable insights into the factors that influence public willingness to share personal health data across historically siloed literature, with several factors (eg, privacy, value, and transparency) clearly synonymous across different patient data sharing fields.

Novel contributions of this review include its synthesis of updated literature following COVID-19 and conscious focus on trust, specifically how to build and nurture trust against a highly changeable background of public preferences and expectations. Focusing on more modern models of dynamic consent that engender individual choice and control, this review provides additional insights into the building blocks of establishing trust in the complex context of sharing personal health data for third-party or secondary uses.

Implications

The implications of this review are clear. First, there is a need to reconceptualize the public as active data owners and not as passive data subjects or providers [59,74]. Second, regular attention needs to be paid to the changeable conditions of public support, consciously designing and refining data sharing processes and related policies to reflect ongoing public preferences and requirements [48,70,86,91,98,122]. While the immediate cost of mishandling personal health data can be calculated following high-profile scandals [152,153], the longer-term impacts of failing to respond and support public willingness to share personal health data are not yet known. Third, every effort should be made to ensure the purpose and details of data sharing requests including assurances of data privacy and security, data and user type, personal and collective benefit, profit, and duration (ie, no longer than required for the originally consented activity [43]) of data sharing processes are clearly articulated [154], with individual members able to continuously choose and refine desired levels of control through accessible consenting processes. Part of this involves actively raising awareness of data sharing processes in a bid to build public trust and support [7,44,59]. However, it is imperative that awareness raising is not approached as a simple 1-way exchange of information. A more engaged approach that supports opportunities for ongoing deliberation regarding public interests, concerns, and uncertainties, as opposed to controlled dissemination, is required [59]. Providing public information does not equate to enhanced public trust [106,107,125]. As suggested by Middleton et al [71], suggested areas of information include the purpose of data sharing, why partnerships between profit and not-for-profit industries are necessary, and what the relevance of data sharing processes is to public lives. But first, we must work to understand what people want to know and how to make the subject resonate [71], ensuring opportunities afforded by data sharing practices are not miss-sold or overemphasized in a bid to engender public support, highlighting the ethical arguments that underpin such activities. Similar to seeking patient feedback about experiences

XSL•FO
RenderX

of care [155], it is perhaps unethical to ask patients to share their personal health data if no change or progress is to be achieved. Failure to use consented data sharing practices only seeks to invalidate and undermine patients' time and expertise. Understanding public needs and aspirations around these issues is arguably best achieved through meaningful involvement or co-design to ensure information value, relevance, and understanding [3,54]. As stated by Franklin et al [82],

"It is vital that we understand patient perspectives on data sharing and work with them as partners, valuing their unique contributions and attending to their preferences."

Working with patient advocacy groups may be an essential component of this process, particularly in areas where public confidence in patient-sharing initiatives may already be weakened.

Finally, public trust is widely recognized as the central pillar in supporting public willingness. However, few studies have explicitly explored how to build public trust. Emerging research suggests health care organizations and institutions may be able to increase public comfort by sharing health data with third-party or commercial companies by emphasizing patient-centered benefits [2]. However, future research including the application of trust models may further benefit from focusing on how issues and relationships of trust can be both nurtured and overcome [59].

Strengths and Limitations

Strengths of this research include its application of a recognized systematic review process, the inclusion of updated literature following COVID-19 including international literature from developing countries, responding to criticisms that existing studies have largely concentrated on developed countries only [62], and synthesis of historically disparate literature into 1 corpus of information. However, its limitations must also be acknowledged. Much of the existing literature is based on data sharing intentions, or hypothetical scenarios as opposed to real-world, or observed behavior [4,32,52,57,73,96,101,109,113]. Reported findings may not therefore translate into practice [10,71,132]. However, commonality in findings provides some level of confidence in the findings presented. Furthermore, there is evidence to suggest people are more hesitant about hypothetical scenarios than they are in reality [10,71,102,156]. Moreover, not all included studies focused on intentional behavior, with some exploring data sharing practices in real-world settings [18,118]. Finally, the literature search was restricted to the English language only. The risk of publication bias is therefore acknowledged. Further research may also benefit from a critical exploration of country-specific legal restrictions and their role in data sharing practices.

Summary

In summary, there is general, yet conditional public support for sharing personal health data for third-party or secondary use. Clarity, transparency, and individual control over who has access to what data, why, when, and for how long are widely regarded as an essential prerequisite for public data sharing support. Individual levels of control and choice need to operate within the auspices of assured data privacy and security processes, underpinned by dynamic and responsive models of consent that prioritize individual and collective benefits over and above commercial gain. Failure to acknowledge, design, and refine data sharing processes in response to patient preferences and needs will jeopardize the tangible benefits of data sharing practices being fully realized.

Acknowledgments

The authors would like to thank Mrs Lorna Burns for her invaluable expertise, insights, and experience as the information specialist for the review search strategy. There was no funder or financial support for this review.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 checklist.
[DOCX File , 32 KB - jmir_v26i1e50421_app1.docx]

Multimedia Appendix 2
Table of study characteristics.
[DOCX File , 53 KB - jmir_v26i1e50421_app2.docx]

Multimedia Appendix 3
Summary tables.
[DOCX File , 28 KB - jmir_v26i1e50421_app3.docx]

XSL•FO
RenderX

References

1. Seltzer E, Goldshear J, Guntuku SC, Grande D, Asch DA, Klinger EV, et al. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. BMC Med Inform Decis Mak 2019;19(1):157 [FREE Full text] [doi: 10.1186/s12911-019-0886-9] [Medline: 31395102]
2. Trinidad MG, Platt J, Kardia SLR. The public's comfort with sharing health data with third-party commercial companies. Humanit Soc Sci Commun 2020;7(1):149 [FREE Full text] [doi: 10.1057/s41599-020-00641-5] [Medline: 34337435]
3. Lounsbury O, Roberts L, Goodman JR, Batey P, Naar L, Flott KM, et al. Opening a "can of worms" to explore the public's hopes and fears about health care data sharing: qualitative study. J Med Internet Res 2021;23(2):e22744 [FREE Full text] [doi: 10.2196/22744] [Medline: 33616532]
4. Abdelhamid M, Gaia J, Sanders GL. Putting the focus back on the patient: how privacy concerns affect personal health information sharing intentions. J Med Internet Res 2017;19(9):e169 [FREE Full text] [doi: 10.2196/jmir.6877] [Medline: 28903895]
5. Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015;20(5):464-e20 [FREE Full text] [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
6. Cherif E, Bezaz N, Mzoughi M. Do personal health concerns and trust in healthcare providers mitigate privacy concerns? Effects on patients' intention to share personal health data on electronic health records. Soc Sci Med 2021;283:114146. [doi: 10.1016/j.socscimed.2021.114146] [Medline: 34242891]
7. Shen N, Bernier T, Sequeira L, Strauss J, Silver MP, Carter-Langford A, et al. Understanding the patient privacy perspective on health information exchange: a systematic review. Int J Med Inform 2019;125:1-12. [doi: 10.1016/j.ijmedinf.2019.01.014] [Medline: 30914173]
8. Braunack-Mayer A, Fabrianesi B, Street J, O'Shaughnessy P, Carter SM, Engelen L, et al. Sharing government health data with the private sector: community attitudes survey. J Med Internet Res 2021;23(10):e24200 [FREE Full text] [doi: 10.2196/24200] [Medline: 34596573]
9. Howe N, Giles E, Newbury-Birch D, McColl E. Systematic review of participants' attitudes towards data sharing: a thematic synthesis. J Health Serv Res Policy 2018;23(2):123-133. [doi: 10.1177/1355819617751555] [Medline: 29653503]
10. Middleton A, Milne R, Almarri MA, Anwer S, Atutornu J, Baranova EE, et al. Global public perceptions of genomic data sharing: what shapes the willingness to donate DNA and health data? Am J Hum Genet 2020;107(4):743-752. [doi: 10.1016/j.ajhg.2020.08.023] [Medline: 32946764]
11. Hinds J, Williams EJ, Joinson AN. “It wouldn't happen to me”: privacy concerns and perspectives following the Cambridge Analytica scandal. Int J Hum Comput Stud 2020;143:102498. [doi: 10.1016/j.ijhcs.2020.102498]
12. Ford A, De Togni G, Miller L. Hormonal health: period tracking apps, wellness, and self-management in the era of surveillance capitalism. Engag Sci Technol Soc 2021;7(1):48-66 [FREE Full text] [doi: 10.17351/ests2021.655] [Medline: 37155428]
13. Moats D, McFall L. In search of a problem: mapping controversies over NHS (England) patient data with digital tools. Sci Technol Hum Values 2018;44(3):478-513. [doi: 10.1177/0162243918796274]
14. Luo Y, Oh CY, St Jean B, Choe EK. Interrelationships between patients' data tracking practices, data sharing practices, and health literacy: onsite survey study. J Med Internet Res 2020;22(12):e18937 [FREE Full text] [doi: 10.2196/18937] [Medline: 33350960]
15. Brall C, Berlin C, Zwahlen M, Ormond KE, Egger M, Vayena E. Public willingness to participate in personalized health research and biobanking: a large-scale Swiss survey. PLoS One 2021;16(4):e0249141 [FREE Full text] [doi: 10.1371/journal.pone.0249141] [Medline: 33793624]
16. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics 2022;48(1):3-13 [FREE Full text] [doi: 10.1136/medethics-2019-105651] [Medline: 31719155]
17. Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of health consumer attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on privacy, trust, and transparency. Syst Rev 2020;9(1):235 [FREE Full text] [doi: 10.1186/s13643-020-01481-9] [Medline: 33036664]
18. Moon LA. Factors influencing health data sharing preferences of consumers: a critical review. Health Policy Technol 2017;6(2):169-187. [doi: 10.1016/j.hlpt.2017.01.001]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
21. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme version. Comput Sci 2006;1(1):b92 [FREE Full text]

XSL•FO
RenderX

22. Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009;62(9):944-952 [FREE Full text] [doi: 10.1016/j.jclinepi.2008.10.012] [Medline: 19230612]
23. Booth A. Clear and present questions: formulating questions for evidence based practice. Library Hi Tech 2006;243:355-368. [doi: 10.1108/07378830610692127]
24. Whiddett D, Hunter I, McDonald B, Norris T, Waldon J. Consent and widespread access to personal health information for the delivery of care: a large scale telephone survey of consumers' attitudes using vignettes in New Zealand. BMJ Open 2016;6(8):e011640 [FREE Full text] [doi: 10.1136/bmjopen-2016-011640] [Medline: 27554103]
25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Smith JA, editor. Qualitative Psychology: A Practical Guide to Research Methods, 3rd Edition. London: Sage Publications; 2015:222-248.
27. Baines R, Bradwell H, Edwards K, Stevens S, Prime S, Tredinnick-Rowe J, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022;25(4):1232-1245 [FREE Full text] [doi: 10.1111/hex.13506] [Medline: 35526274]
28. Lalani M, Baines R, Bryce M, Marshall M, Mead S, Barasi S, et al. Patient and public involvement in medical performance processes: a systematic review. Health Expect 2019;22(2):149-161 [FREE Full text] [doi: 10.1111/hex.12852] [Medline: 30548359]
29. Carson C, Hinton L, Kurinczuk J, Quigley M. 'I haven't met them, I don't have any trust in them. It just feels like a big unknown': a qualitative study exploring the determinants of consent to use Human Fertilisation and Embryology Authority registry data in research. BMJ Open 2019;9(5):e026469 [FREE Full text] [doi: 10.1136/bmjopen-2018-026469] [Medline: 31152033]
30. Cheung C, Bietz MJ, Patrick K, Bloss CS. Privacy attitudes among early adopters of emerging health technologies. PLoS One 2016;11(11):e0166389 [FREE Full text] [doi: 10.1371/journal.pone.0166389] [Medline: 27832194]
31. Cochran GL, Lander L, Morien M, Lomelin DE, Brittin J, Reker C, et al. Consumer opinions of health information exchange, e-prescribing, and personal health records. Perspect Health Inf Manag 2015;12(Fall):1e [FREE Full text] [Medline: 26604874]
32. Kim KK, Sankar P, Wilson MD, Haynes SC. Factors affecting willingness to share electronic health data among California consumers. BMC Med Ethics 2017;18(1):25 [FREE Full text] [doi: 10.1186/s12910-017-0185-x] [Medline: 28376801]
33. King T, Brankovic L, Gillard P. Perspectives of Australian adults about protecting the privacy of their health information in statistical databases. Int J Med Inform 2012;81(4):279-289. [doi: 10.1016/j.ijmedinf.2012.01.005] [Medline: 22306206]
34. Krahe M, Milligan E, Reilly S. Personal health information in research: perceived risk, trustworthiness and opinions from patients attending a tertiary healthcare facility. J Biomed Inform 2019;95:103222 [FREE Full text] [doi: 10.1016/j.jbi.2019.103222] [Medline: 31176040]
35. Gaylin DS, Moiduddin A, Mohamoud S, Lundeen K, Kelly JA. Public attitudes about health information technology, and its relationship to health care quality, costs, and privacy. Health Serv Res 2011;46(3):920-938 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01233.x] [Medline: 21275986]
36. Haddow G, Bruce A, Sathanandam S, Wyatt JC. 'Nothing is really safe': a focus group study on the processes of anonymizing and sharing of health data for research purposes. J Eval Clin Pract 2011;17(6):1140-1146. [doi: 10.1111/j.1365-2753.2010.01488.x] [Medline: 20629997]
37. Gray CS, Miller D, Kuluski K, Cott C. Tying eHealth tools to patient needs: exploring the use of eHealth for community-dwelling patients with complex chronic disease and disability. JMIR Res Protoc 2014;3(4):e67 [FREE Full text] [doi: 10.2196/resprot.3500] [Medline: 25428028]
38. Heath D, Ardestani A, Nemati H. Sharing personal genetic information: the impact of privacy concern and awareness of benefit. J Inf Commun Ethics Soc 2016;14(3):288-308. [doi: 10.1108/jices-07-2015-0025]
39. Hui SKA, Miller SM, Hazuda L, Engelman K, Ellerbeck EF. Novel method for recruiting representative at-risk individuals into cancer prevention trials: online health risk assessment in employee wellness programs. J Cancer Educ 2016;31(3):421-429 [FREE Full text] [doi: 10.1007/s13187-015-0927-8] [Medline: 26507744]
40. Goodman D, Johnson CO, Bowen D, Smith M, Wenzel L, Edwards K. De-identified genomic data sharing: the research participant perspective. J Community Genet 2017;8(3):173-181 [FREE Full text] [doi: 10.1007/s12687-017-0300-1] [Medline: 28382417]
41. Esmaeilzadeh P. Patients' perceptions of different information exchange mechanisms: an exploratory study in the United States. Methods Inf Med 2020;59(4-05):162-178. [doi: 10.1055/s-0040-1721784] [Medline: 33618421]
42. Hassan L, Dalton A, Hammond C, Tully MP. A deliberative study of public attitudes towards sharing genomic data within NHS genomic medicine services in England. Public Underst Sci 2020;29(7):702-717 [FREE Full text] [doi: 10.1177/0963662520942132] [Medline: 32664786]
43. Hentschel A, Hsiao CJ, Chen LY, Wright L, Shaw J, Du X, et al. Perspectives of pregnant and breastfeeding women on participating in longitudinal mother-baby studies involving electronic health records: qualitative study. JMIR Pediatr Parent 2021;4(1):e23842 [FREE Full text] [doi: 10.2196/23842] [Medline: 33666558]

XSL•FO
RenderX

44. Papoutsi C, Reed JE, Marston C, Lewis R, Majeed A, Bell D. Patient and public views about the security and privacy of electronic health records (EHRs) in the UK: results from a mixed methods study. BMC Med Inform Decis Mak 2015;15:86 [FREE Full text] [doi: 10.1186/s12911-015-0202-2] [Medline: 26466787]
45. Small SS, Hohl CM, Balka E. Patient perspectives on health data privacy and implications for adverse drug event documentation and communication: qualitative study. J Med Internet Res 2021;23(1):e21452 [FREE Full text] [doi: 10.2196/21452] [Medline: 33470936]
46. Smith AB, Odlum M, Sikka M, Bakken S, Kanter T. Patient perceptions of pre-implementation of personal health records (PHRs): a qualitative study of people living with HIV in New York City. J HIV AIDS Soc Serv 2012;11(4):406-423. [doi: 10.1080/15381501.2012.735166]
47. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: a qualitative study. J Med Internet Res 2016;18(4):e66 [FREE Full text] [doi: 10.2196/jmir.5011] [Medline: 27083521]
48. Platt J, Raj M, Kardia SLR. The public's trust and information brokers in health care, public health and research. J Health Organ Manag 2019;33(7/8):929-948. [doi: 10.1108/jhom-11-2018-0332]
49. Platt J, Kardia S. Public trust in health information sharing: implications for biobanking and electronic health record systems. J Pers Med 2015;5(1):3-21 [FREE Full text] [doi: 10.3390/jpm5010003] [Medline: 25654300]
50. Morrison Z, Fernando B, Kalra D, Cresswell K, Sheikh A. National evaluation of the benefits and risks of greater structuring and coding of the electronic health record: exploratory qualitative investigation. J Am Med Inform Assoc 2014;21(3):492-500 [FREE Full text] [doi: 10.1136/amiajnl-2013-001666] [Medline: 24186957]
51. Mazor KM, Richards A, Gallagher M, Arterburn DE, Raebel MA, Nowell WB, et al. Stakeholders' views on data sharing in multicenter studies. J Comp Eff Res 2017;6(6):537-547 [FREE Full text] [doi: 10.2217/cer-2017-0009] [Medline: 28805448]
52. Medford-Davis LN, Chang L, Rhodes KV. Health information exchange: what do patients want? Health Informatics J 2017;23(4):268-278 [FREE Full text] [doi: 10.1177/1460458216647190] [Medline: 27245671]
53. Mansour E. Egyptian patients' perceptions of using personal health records (PHRs): an exploratory study. Glob Knowl Mem Commun 2018;67(8/9):523-546. [doi: 10.1108/gkmc-04-2018-0034]
54. Mamo LA, Browe DK, Logan HC, Kim KK. Patient informed governance of distributed research networks: results and discussion from six patient focus groups. AMIA Annu Symp Proc 2013;2013:920-929 [FREE Full text] [Medline: 24551383]
55. Lysaght T, Ballantyne A, Xafis V, Ong S, Schaefer GO, Ling JMT, et al. "Who is watching the watchdog?": ethical perspectives of sharing health-related data for precision medicine in Singapore. BMC Med Ethics 2020;21(1):118 [FREE Full text] [doi: 10.1186/s12910-020-00561-8] [Medline: 33213433]
56. Esmaeilzadeh P, Mirzaei T. Comparison of consumers' perspectives on different Health Information Exchange (HIE) mechanisms: an experimental study. Int J Med Inform 2018;119:1-7. [doi: 10.1016/j.ijmedinf.2018.08.007] [Medline: 30342677]
57. Hwang HG, Lin Y. Evaluating people's concern about their health information privacy based on power-responsibility equilibrium model: a case of Taiwan. J Med Syst 2020;44(6):112. [doi: 10.1007/s10916-020-01579-6] [Medline: 32382934]
58. O'Brien EC, Rodriguez AM, Kum HC, Schanberg LE, Fitz-Randolph M, O'Brien SM, et al. Patient perspectives on the linkage of health data for research: insights from an online patient community questionnaire. Int J Med Inform 2019;127:9-17 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.04.003] [Medline: 31128837]
59. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016;17(1):73 [FREE Full text] [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]
60. Atkin C, Crosby B, Dunn K, Price G, Marston E, Crawford C, et al. Perceptions of anonymised data use and awareness of the NHS data opt-out amongst patients, carers and healthcare staff. Res Involv Engagem 2021;7(1):40 [FREE Full text] [doi: 10.1186/s40900-021-00281-2] [Medline: 34127076]
61. Helou S, Abou-Khalil V, El Helou E, Kiyono K. Factors related to personal health data sharing: data usefulness, sensitivity and anonymity. Stud Health Technol Inform 2021;281:1051-1055 [FREE Full text] [doi: 10.3233/SHTI210345] [Medline: 34042839]
62. Adu EK, Mills A, Todorova N. Factors influencing individuals' personal health information privacy concerns. A study in Ghana. Inf Technol Dev 2020;27(2):208-234. [doi: 10.1080/02681102.2020.1806018]
63. Bouras A, Simoes EJ, Boren S, Hicks L, Zachary I, Buck C, et al. Non-Hispanic white mothers' willingness to share personal health data with researchers: survey results from an opt-in panel. J Particip Med 2020;12(2):e14062 [FREE Full text] [doi: 10.2196/14062] [Medline: 33064096]
64. Weng C, Hao T, Friedman C, Hurdle J. Crowdsourcing public opinion for sharing medical records for the advancement of science. Stud Health Technol Inform 2019;264:1393-1397 [FREE Full text] [doi: 10.3233/SHTI190456] [Medline: 31438155]
65. Jones RD, Krenz C, Gornick M, Griffith KA, Spence R, Bradbury AR, et al. Patient preferences regarding informed consent models for participation in a learning health care system for oncology. JCO Oncol Pract 2020;16(9):e977-e990 [FREE Full text] [doi: 10.1200/JOP.19.00300] [Medline: 32352881]

XSL•FO
RenderX

66. Courbier S, Dimond R, Bros-Facer V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection—quantitative survey and recommendations. Orphanet J Rare Dis 2019;14(1):175 [FREE Full text] [doi: 10.1186/s13023-019-1123-4] [Medline: 31300010]
67. Andrews SM, Raspa M, Edwards A, Moultrie R, Turner-Brown L, Wagner L, et al. "Just tell me what's going on": the views of parents of children with genetic conditions regarding the research use of their child's electronic health record. J Am Med Inform Assoc 2020;27(3):429-436 [FREE Full text] [doi: 10.1093/jamia/ocz208] [Medline: 31913479]
68. Bosisio F, Barazzetti G, Koutaissoff D, Spencer B. Patients' decision to contribute to a biobank in the light of the patient-recruiter relationship-a qualitative study of broad consent in a hospital setting. J Community Genet 2021;12(1):15-25 [FREE Full text] [doi: 10.1007/s12687-020-00479-z] [Medline: 32779150]
69. Burstein MD, Robinson JO, Hilsenbeck SG, McGuire AL, Lau CC. Pediatric data sharing in genomic research: attitudes and preferences of parents. Pediatrics 2014;133(4):690-697 [FREE Full text] [doi: 10.1542/peds.2013-1592] [Medline: 24616359]
70. Vidgen ME, Kaladharan S, Malacova E, Hurst C, Waddell N. Sharing genomic data from clinical testing with researchers: public survey of expectations of clinical genomic data management in Queensland, Australia. BMC Med Ethics 2020;21(1):119 [FREE Full text] [doi: 10.1186/s12910-020-00563-6] [Medline: 33213438]
71. Middleton A, Milne R, Howard H, Niemiec E, Robarts L, Critchley C, et al. Members of the public in the USA, UK, Canada and Australia expressing genetic exceptionalism say they are more willing to donate genomic data. Eur J Hum Genet 2020;28(4):424-434 [FREE Full text] [doi: 10.1038/s41431-019-0550-y] [Medline: 31784701]
72. Grundstrom C, Korhonen O, Väyrynen K, Isomursu M. Insurance customers' expectations for sharing health data: qualitative survey study. JMIR Med Inform 2020;8(3):e16102 [FREE Full text] [doi: 10.2196/16102] [Medline: 32213467]
73. Miyamoto SW, Henderson S, Young HM, Pande A, Han JJ. Tracking health data is not enough: a qualitative exploration of the role of healthcare partnerships and mHealth technology to promote physical activity and to sustain behavior change. JMIR Mhealth Uhealth 2016;4(1):e5 [FREE Full text] [doi: 10.2196/mhealth.4814] [Medline: 26792225]
74. Satinsky EN, Driessens C, Crepaz-Keay D, Kousoulis AA. Mental health service users' perceptions of data sharing and data protection: a qualitative report. J Innov Health Inform 2018;25(4):239-242 [FREE Full text] [doi: 10.14236/jhi.v25i4.1033] [Medline: 30672404]
75. Litchman ML, Allen NA, Colicchio VD, Wawrzynski SE, Sparling KM, Hendricks KL, et al. A qualitative analysis of real-time continuous glucose monitoring data sharing with care partners: to share or not to share? Diabetes Technol Ther 2018;20(1):25-31. [doi: 10.1089/dia.2017.0285] [Medline: 29154685]
76. Platt JE, Jacobson PD, Kardia SLR. Public trust in health information sharing: a measure of system trust. Health Serv Res 2018;53(2):824-845 [FREE Full text] [doi: 10.1111/1475-6773.12654] [Medline: 28097657]
77. Patil S, Lu H, Saunders CL, Potoglou D, Robinson N. Public preferences for electronic health data storage, access, and sharing—evidence from a pan-European survey. J Am Med Inform Assoc 2016;23(6):1096-1106 [FREE Full text] [doi: 10.1093/jamia/ocw012] [Medline: 27107445]
78. Bak MAR, Veeken R, Blom MT, Tan HL, Willems DL. Health data research on sudden cardiac arrest: perspectives of survivors and their next-of-kin. BMC Med Ethics 2021;22(1):7 [FREE Full text] [doi: 10.1186/s12910-021-00576-9] [Medline: 33509184]
79. Caine K, Kohn S, Lawrence C, Hanania R, Meslin EM, Tierney WM. Designing a patient-centered user interface for access decisions about EHR data: implications from patient interviews. J Gen Intern Med 2015;30(Suppl 1):S7-S16 [FREE Full text] [doi: 10.1007/s11606-014-3049-9] [Medline: 25480719]
80. Esmaeilzadeh P, Mirzaei T. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. J Med Internet Res 2019;21(6):e14184 [FREE Full text] [doi: 10.2196/14184] [Medline: 31223119]
81. Horgan D, Bernini C, Thomas PPM, Morre SA. Cooperating on data: the missing element in bringing real innovation to Europe's healthcare systems. Public Health Genomics 2019;22(3-4):77-101 [FREE Full text] [doi: 10.1159/000503296] [Medline: 31634895]
82. Franklin EF, Nichols HM, House L, Buzaglo J, Thiboldeaux K. Cancer patient perspectives on sharing of medical records and mobile device data for research purposes. J Patient Exp 2020;7(6):1115-1121 [FREE Full text] [doi: 10.1177/2374373520923837] [Medline: 33457553]
83. Grande D, Marti XL, Merchant RM, Asch DA, Dolan A, Sharma M, et al. Consumer views on health applications of consumer digital data and health privacy among US adults: qualitative interview study. J Med Internet Res 2021;23(6):e29395 [FREE Full text] [doi: 10.2196/29395] [Medline: 34106074]
84. Wetzels M, Broers E, Peters P, Feijs L, Widdershoven J, Habibovic M. Patient perspectives on health data privacy and management: "where is my data and whose is it?". Int J Telemed Appl 2018;2018:3838747 [FREE Full text] [doi: 10.1155/2018/3838747] [Medline: 30631347]
85. Coathup V, Teare HJA, Minari J, Yoshizawa G, Kaye J, Takahashi MP, et al. Using digital technologies to engage with medical research: views of myotonic dystrophy patients in Japan. BMC Med Ethics 2016;17(1):51 [FREE Full text] [doi: 10.1186/s12910-016-0132-2] [Medline: 27553007]

XSL•FO
RenderX

86. Bernaerdt J, Moerenhout T, Devisch I. Vulnerable patients' attitudes towards sharing medical data and granular control in patient portal systems: an interview study. J Eval Clin Pract 2021;27(2):429-437. [doi: 10.1111/jep.13465] [Medline: 32886953]
87. Soni H, Grando A, Murcko A, Diaz S, Mukundan M, Idouraine N, et al. State of the art and a mixed-method personalized approach to assess patient perceptions on medical record sharing and sensitivity. J Biomed Inform 2020;101:103338 [FREE Full text] [doi: 10.1016/j.jbi.2019.103338] [Medline: 31726102]
88. Kim J, Kim H, Bell E, Bath T, Paul P, Pham A, et al. Patient perspectives about decisions to share medical data and biospecimens for research. JAMA Netw Open 2019;2(8):e199550 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9550] [Medline: 31433479]
89. Avram R, Marquis-Gravel G, Simard F, Pacheco C, Couture É, Tremblay-Gravel M, et al. Understanding the patient perspective on research access to national health records databases for conduct of randomized registry trials. Int J Cardiol 2018;262:110-116. [doi: 10.1016/j.ijcard.2017.12.074] [Medline: 29706388]
90. McGuire AL, Oliver JM, Slashinski MJ, Graves JL, Wang T, Kelly PA, et al. To share or not to share: a randomized trial of consent for data sharing in genome research. Genet Med 2011;13(11):948-955 [FREE Full text] [doi: 10.1097/GIM.0b013e3182227589] [Medline: 21785360]
91. Grande D, Asch DA, Wan F, Bradbury AR, Jagsi R, Mitra N. Are patients with cancer less willing to share their health information? Privacy, sensitivity, and social purpose. J Oncol Pract 2015;11(5):378-383 [FREE Full text] [doi: 10.1200/JOP.2015.004820] [Medline: 26265174]
92. Kim KK, Joseph JG, Ohno-Machado L. Comparison of consumers' views on electronic data sharing for healthcare and research. J Am Med Inform Assoc 2015;22(4):821-830 [FREE Full text] [doi: 10.1093/jamia/ocv014] [Medline: 25829461]
93. Lee K, Lim K, Jung SY, Ji H, Hong K, Hwang H, et al. Perspectives of patients, health care professionals, and developers toward blockchain-based health information exchange: qualitative study. J Med Internet Res 2020;22(11):e18582 [FREE Full text] [doi: 10.2196/18582] [Medline: 33185553]
94. Hyde ET, Omura JD, Fulton JE, Weldy A, Carlson SA. Physical activity surveillance using wearable activity monitors: are US adults willing to share their data? Am J Health Promot 2020;34(6):672-676. [doi: 10.1177/0890117119900587] [Medline: 31957481]
95. Maiorana A, Steward WT, Koester KA, Pearson C, Shade SB, Chakravarty D, et al. Trust, confidentiality, and the acceptability of sharing HIV-related patient data: lessons learned from a mixed methods study about health information exchanges. Implement Sci 2012;7:34 [FREE Full text] [doi: 10.1186/1748-5908-7-34] [Medline: 22515736]
96. Grande D, Mitra N, Shah A, Wan F, Asch DA. Public preferences about secondary uses of electronic health information. JAMA Intern Med 2013;173(19):1798-1806 [FREE Full text] [doi: 10.1001/jamainternmed.2013.9166] [Medline: 23958803]
97. Tosoni S, Voruganti I, Lajkosz K, Habal F, Murphy P, Wong RKS, et al. The use of personal health information outside the circle of care: consent preferences of patients from an academic health care institution. BMC Med Ethics 2021;22(1):29 [FREE Full text] [doi: 10.1186/s12910-021-00598-3] [Medline: 33761938]
98. Page SA, Manhas KP, Muruve DA. A survey of patient perspectives on the research use of health information and biospecimens. BMC Med Ethics 2016;17(1):48 [FREE Full text] [doi: 10.1186/s12910-016-0130-4] [Medline: 27527514]
99. Weidman J, Aurite W, Grossklags J. On sharing intentions, and personal and interdependent privacy considerations for genetic data: a vignette study. IEEE/ACM Trans Comput Biol Bioinform 2019;16(4):1349-1361. [doi: 10.1109/TCBB.2018.2854785] [Medline: 30004885]
100. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021;23(8):e26162 [FREE Full text] [doi: 10.2196/26162] [Medline: 34236994]
101. Jagsi R, Griffith KA, Sabolch A, Jones R, Spence R, De Vries R, et al. Perspectives of patients with cancer on the ethics of rapid-learning health systems. J Clin Oncol 2017;35(20):2315-2323 [FREE Full text] [doi: 10.1200/JCO.2016.72.0284] [Medline: 28537812]
102. Milne R, Morley KI, Howard H, Niemiec E, Nicol D, Critchley C, et al. Trust in genomic data sharing among members of the general public in the UK, USA, Canada and Australia. Hum Genet 2019;138(11-12):1237-1246 [FREE Full text] [doi: 10.1007/s00439-019-02062-0] [Medline: 31531740]
103. Harle CA, Golembiewski EH, Rahmanian KP, Krieger JL, Hagmajer D, Mainous AG, et al. Patient preferences toward an interactive e-consent application for research using electronic health records. J Am Med Inform Assoc 2018;25(3):360-368 [FREE Full text] [doi: 10.1093/jamia/ocx145] [Medline: 29272408]
104. Bell EA, Ohno-Machado L, Grando MA. Sharing my health data: a survey of data sharing preferences of healthy individuals. AMIA Annu Symp Proc 2014;2014:1699-1708 [FREE Full text] [Medline: 25954442]
105. Bietz MJ, Bloss CS, Calvert S, Godino JG, Gregory J, Claffey MP, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc 2016;23(e1):e42-e48 [FREE Full text] [doi: 10.1093/jamia/ocv118] [Medline: 26335984]
106. Romero RA, Young SD. Ethical perspectives in sharing digital data for public health surveillance before and shortly after the onset of the COVID-19 pandemic. Ethics Behav 2022;32(1):22-31 [FREE Full text] [doi: 10.1080/10508422.2021.1884079] [Medline: 35340741]

XSL•FO
RenderX

107. Tully MP, Hassan L, Oswald M, Ainsworth J. Commercial use of health data-a public "trial" by citizens' jury. Learn Health Syst 2019;3(4):e10200 [FREE Full text] [doi: 10.1002/lrh2.10200] [Medline: 31641688]
108. Kimura M, Nakaya J, Watanabe H, Shimizu T, Nakayasu K. A survey aimed at general citizens of the US and Japan about their attitudes toward electronic medical data handling. Int J Environ Res Public Health 2014;11(5):4572-4588 [FREE Full text] [doi: 10.3390/ijerph110504572] [Medline: 24776721]
109. Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AHM, Aufox SA, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. Am J Hum Genet 2017;100(3):414-427 [FREE Full text] [doi: 10.1016/j.ajhg.2017.01.021] [Medline: 28190457]
110. Esmaeilzadeh P. The impacts of the perceived transparency of privacy policies and trust in providers for building trust in health information exchange: empirical study. JMIR Med Inform 2019;7(4):e14050 [FREE Full text] [doi: 10.2196/14050] [Medline: 31769757]
111. Jagsi R, Griffith KA, Jones RD, Krenz C, Gornick M, Spence R, et al. Effect of public deliberation on patient attitudes regarding consent and data use in a learning health care system for oncology. J Clin Oncol 2019;37(34):3203-3211 [FREE Full text] [doi: 10.1200/JCO.19.01693] [Medline: 31577472]
112. Street J, Fabrianesi B, Adams C, Flack F, Smith M, Carter SM, et al. Sharing administrative health data with private industry: a report on two citizens' juries. Health Expect 2021;24(4):1337-1348 [FREE Full text] [doi: 10.1111/hex.13268] [Medline: 34048624]
113. Frost J, Vermeulen IE, Beekers N. Anonymity versus privacy: selective information sharing in online cancer communities. J Med Internet Res 2014;16(5):e126 [FREE Full text] [doi: 10.2196/jmir.2684] [Medline: 24828114]
114. Luchenski SA, Reed JE, Marston C, Papoutsi C, Majeed A, Bell D. Patient and public views on electronic health records and their uses in the United kingdom: cross-sectional survey. J Med Internet Res 2013;15(8):e160 [FREE Full text] [doi: 10.2196/jmir.2701] [Medline: 23975239]
115. Patel VN, Abramson E, Edwards AM, Cheung MA, Dhopeshwarkar RV, Kaushal R. Consumer attitudes toward personal health records in a beacon community. Am J Manag Care 2011;17(4):e104-e120 [FREE Full text] [Medline: 21774099]
116. Esmaeilzadeh P, Dharanikota S, Mirzaei T. The role of patient engagement in patient-centric Health Information Exchange (HIE) initiatives: an empirical study in the United States. Inf Technol People. Preprint posted online on January 25, 2021 2021 [FREE Full text] [doi: 10.1108/itp-05-2020-0316]
117. Nicholas J, Shilton K, Schueller SM, Gray EL, Kwasny MJ, Mohr DC. The role of data type and recipient in individuals' perspectives on sharing passively collected smartphone data for mental health: cross-sectional questionnaire study. JMIR Mhealth Uhealth 2019;7(4):e12578 [FREE Full text] [doi: 10.2196/12578] [Medline: 30950799]
118. Harle CA, Golembiewski EH, Rahmanian KP, Brumback B, Krieger JL, Goodman KW, et al. Does an interactive trust-enhanced electronic consent improve patient experiences when asked to share their health records for research? A randomized trial. J Am Med Inform Assoc 2019;26(7):620-629 [FREE Full text] [doi: 10.1093/jamia/ocz015] [Medline: 30938751]
119. Riordan F, Papoutsi C, Reed JE, Marston C, Bell D, Majeed A. Patient and public attitudes towards informed consent models and levels of awareness of electronic health records in the UK. Int J Med Inform 2015;84(4):237-247 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.01.008] [Medline: 25649841]
120. Platt J, Raj M, Büyüktür AG, Trinidad MG, Olopade O, Ackerman MS, et al. Willingness to participate in health information networks with diverse data use: evaluating public perspectives. EGEMS (Wash DC) 2019;7(1):33 [FREE Full text] [doi: 10.5334/egems.288] [Medline: 31367650]
121. Lucero RJ, Kearney J, Cortes Y, Arcia A, Appelbaum P, Fernández RL, et al. Benefits and risks in secondary use of digitized clinical data: views of community members living in a predominantly ethnic minority urban neighborhood. AJOB Empir Bioeth 2015;6(2):12-22 [FREE Full text] [doi: 10.1080/23294515.2014.949906] [Medline: 26101782]
122. Shen N, Sequeira L, Silver MP, Carter-Langford A, Strauss J, Wiljer D. Patient privacy perspectives on health information exchange in a mental health context: qualitative study. JMIR Ment Health 2019;6(11):e13306 [FREE Full text] [doi: 10.2196/13306] [Medline: 31719029]
123. Milne R, Morley KI, Almarri MA, Anwer S, Atutornu J, Baranova EE, et al. Demonstrating trustworthiness when collecting and sharing genomic data: public views across 22 countries. Genome Med 2021;13(1):92 [FREE Full text] [doi: 10.1186/s13073-021-00903-0] [Medline: 34034801]
124. Navarro-Millán I, Zinski A, Shurbaji S, Johnson B, Fraenkel L, Willig J, et al. Perspectives of rheumatoid arthritis patients on electronic communication and patient-reported outcome data collection: a qualitative study. Arthritis Care Res (Hoboken) 2019;71(1):80-87 [FREE Full text] [doi: 10.1002/acr.23580] [Medline: 29669191]
125. Tully MP, Bozentko K, Clement S, Hunn A, Hassan L, Norris R, et al. Investigating the extent to which patients should control access to patient records for research: a deliberative process using citizens' juries. J Med Internet Res 2018;20(3):e112 [FREE Full text] [doi: 10.2196/jmir.7763] [Medline: 29592847]
126. Stevenson F. The use of electronic patient records for medical research: conflicts and contradictions. BMC Health Serv Res 2015;15:124 [FREE Full text] [doi: 10.1186/s12913-015-0783-6] [Medline: 25886266]
127. Holm S, Kristiansen TB, Ploug T. Control, trust and the sharing of health information: the limits of trust. J Med Ethics. preprint published online on January 12, 2021 2020. [doi: 10.1136/medethics-2019-105887] [Medline: 32843438]

XSL•FO
RenderX

128. Crotty BH, Walker J, Dierks M, Lipsitz L, O'Brien J, Fischer S, et al. Information sharing preferences of older patients and their families. JAMA Intern Med 2015;175(9):1492-1497 [FREE Full text] [doi: 10.1001/jamainternmed.2015.2903] [Medline: 26147401]
129. Velarde MCR, Tsantoulis P, Burton-Jeangros C, Aceti M, Chappuis P, Hurst-Majno S. Citizens' views on sharing their health data: the role of competence, reliability and pursuing the common good. BMC Med Ethics 2021;22(1):62 [FREE Full text] [doi: 10.1186/s12910-021-00633-3] [Medline: 34006284]
130. Beckjord EB, Rechis R, Nutt S, Shulman L, Hesse BW. What do people affected by cancer think about electronic health information exchange? Results from the 2010 LIVESTRONG Electronic Health Information Exchange Survey and the 2008 Health Information National Trends Survey. J Oncol Pract 2011;7(4):237-241 [FREE Full text] [doi: 10.1200/JOP.2011.000324] [Medline: 22043188]
131. Tully MP, Bernsten C, Aitken M, Vass C. Public preferences regarding data linkage for research: a discrete choice experiment comparing Scotland and Sweden. BMC Med Inform Decis Mak 2020;20(1):109 [FREE Full text] [doi: 10.1186/s12911-020-01139-5] [Medline: 32546147]
132. Grande D, Mitra N, Shah A, Wan F, Asch DA. The importance of purpose: moving beyond consent in the societal use of personal health information. Ann Intern Med 2014;161(12):855-862 [FREE Full text] [doi: 10.7326/M14-1118] [Medline: 25506854]
133. Bond CS, Ahmed OH, Hind M, Thomas B, Hewitt-Taylor J. The conceptual and practical ethical dilemmas of using health discussion board posts as research data. J Med Internet Res 2013 Jun 07;15(6):e112 [FREE Full text] [doi: 10.2196/jmir.2435] [Medline: 23748147]
134. Hiremath A, Awan O, Mendelson D, Siegel EL. Patient perceptions of participating in the RSNA Image Share Project: a preliminary study. J Digit Imaging 2016 Apr;29(2):189-194 [FREE Full text] [doi: 10.1007/s10278-015-9832-2] [Medline: 26452494]
135. Howard S, Lang A, Sharples S, Shaw D. See I told you I was taking it! - Attitudes of adolescents with asthma towards a device monitoring their inhaler use: Implications for future design. Appl Ergon 2017 Jan;58:224-237 [FREE Full text] [doi: 10.1016/j.apergo.2016.06.018] [Medline: 27633217]
136. Huberty J, Puzia M, Eckert R, Larkey L. Cancer patients' and survivors' perceptions of the Calm app: cross-sectional descriptive study. JMIR Cancer 2020 Jan 25;6(1):e16926 [FREE Full text] [doi: 10.2196/16926] [Medline: 32039812]
137. Jia Hui K. User privacy protection behavior and information sharing in mobile health application. Int J Adv Trend Com Sci Eng 2020 Aug 25;9(4):5250-5258 [FREE Full text] [doi: 10.30534/ijatcse/2020/155942020]
138. Kim H, Bell E, Kim J, Sitapati A, Ramsdell J, Farcas C, et al. iCONCUR: informed consent for clinical data and bio-sample use for research. J Am Med Inform Assoc 2017 Mar 01;24(2):380-387 [FREE Full text] [doi: 10.1093/jamia/ocw115] [Medline: 27589942]
139. Kim J, Kim H, Bell E, Bath T, Paul P, Pham A, et al. Patient perspectives about decisions to share medical data and biospecimens for research. JAMA Netw Open 2019 Aug 02;2(8):e199550 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9550] [Medline: 31433479]
140. Klein DM, Pham K, Samy L, Bluth A, Nazi KM, Witry M, et al. The veteran-initiated electronic care coordination: a multisite initiative to promote and evaluate consumer-mediated health information exchange. Telemed J E Health 2017 Apr;23(4):264-272. [doi: 10.1089/tmj.2016.0078] [Medline: 27726644]
141. Lu C, Batista D, Hamouda H, Lemieux V. Consumers' intentions to adopt blockchain-based personal health records and data sharing: focus group study. JMIR Form Res 2020 Nov 05;4(11):e21995 [FREE Full text] [doi: 10.2196/21995] [Medline: 33151149]
142. Mählmann L, Schee Gen Halfmann S, von Wyl A, Brand A. Attitudes towards personal genomics and sharing of genetic data among older Swiss adults: A qualitative study. Public Health Genomics 2017;20(5):293-306. [doi: 10.1159/000486588] [Medline: 29414817]
143. Buckley BS, Murphy AW, MacFarlane AE. Public attitudes to the use in research of personal health information from general practitioners' records: a survey of the Irish general public. J Med Ethics 2011 Jan;37(1):50-55. [doi: 10.1136/jme.2010.037903] [Medline: 21071570]
144. Parkin L, Paul C. Public good, personal privacy: a citizens' deliberation about using medical information for pharmacoepidemiological research. J Epidemiol Community Health 2011 Feb;65(2):150-156. [doi: 10.1136/jech.2009.097436] [Medline: 19948532]
145. Pedersen H, Taylor D, Gilbert M, Achen M, Lester R, Ogilvie G. A cross-sectional survey exploring attitudes towards provincial electronic health record implementation among clients attending the Provincial Sexually Transmitted Infections Clinic in British Columbia. Sex Transm Infect 2015 Feb;91(1):44-48. [doi: 10.1136/sextrans-2014-051804] [Medline: 25480149]
146. Schairer CE, Cheung C, Kseniya Rubanovich C, Cho M, Cranor LF, Bloss CS. Disposition toward privacy and information disclosure in the context of emerging health technologies. J Am Med Inform Assoc 2019 Jul 01;26(7):610-619 [FREE Full text] [doi: 10.1093/jamia/ocz010] [Medline: 30938756]
147. Serena SS, Peddie D, Ackerley C, Hohl CM, Balka E. Context Sensitive Health Informatics: Redesigning Healthcare Work. Amsterdam, the Netherlands: IOS Press; 2017:109-114.

XSL•FO
RenderX

148. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Med Ethics 2020;46(6):367-377 [FREE Full text] [doi: 10.1136/medethics-2019-105472] [Medline: 32457202]
149. Mählmann L, Halfmann SSG, von Wyl A, Brand A. Attitudes towards personal genomics and sharing of genetic data among older Swiss adults: a qualitative study. Public Health Genomics 2017;20(5):293-306. [doi: 10.1159/000486588] [Medline: 29414817]
150. Buckley BS, Murphy AW, MacFarlane AE. Public attitudes to the use in research of personal health information from general practitioners' records: a survey of the Irish general public. J Med Ethics 2011;37(1):50-55. [doi: 10.1136/jme.2010.037903] [Medline: 21071570]
151. Chawinga WD, Zinn S. Global perspectives of research data sharing: a systematic literature review. Libr Inf Sci Res 2019;41(2):109-122. [doi: 10.1016/j.lisr.2019.04.004]
152. Kressbach M. Period hacks: menstruating in the big data paradigm. Telev New Media 2019;22(3):241-261. [doi: 10.1177/1527476419886389]
153. Healy RL. Zuckerberg, get out of my uterus! An examination of fertility apps, data-sharing and remaking the female body as a digitalized reproductive subject. J Gend Stud 2020;30(4):406-416. [doi: 10.1080/09589236.2020.1845628]
154. Baines R, Stevens S, Garba-Sani Z, Chatterjee A, Austin D, Leigh S. Attitudes toward personal health data sharing among people living with sickle cell disorder, exemplar for study of rare disease populations. J Patient Cent Res Rev 2023;10(2):68-76 [FREE Full text] [doi: 10.17294/2330-0698.2006] [Medline: 37091117]
155. Sheard L, Peacock R, Marsh C, Lawton R. What's the problem with patient experience feedback? A macro and micro understanding, based on findings from a three-site UK qualitative study. Health Expect 2019;22(1):46-53 [FREE Full text] [doi: 10.1111/hex.12829] [Medline: 30244499]
156. Oliver JM, Slashinski MJ, Wang T, Kelly PA, Hilsenbeck SG, McGuire AL. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. Public Health Genomics 2012;15(2):106-114 [FREE Full text] [doi: 10.1159/000334718] [Medline: 22213783]

Abbreviations

EHR: electronic health record
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
SPICE: Setting, Population (or Perspective), Intervention, Comparator, and Evaluation

Edited by Y Zhuang; submitted 30.06.23; peer-reviewed by A Mann, K Gierend, N Howe, S Song; comments to author 28.10.23; revised version received 01.12.23; accepted 18.12.23; published 05.03.24.

Please cite as:
Baines R, Stevens S, Austin D, Anil K, Bradwell H, Cooper L, Maramba ID, Chatterjee A, Leigh S
Patient and Public Willingness to Share Personal Health Data for Third-Party or Secondary Uses: Systematic Review
J Med Internet Res 2024;26:e50421
URL: https://www.jmir.org/2024/1/e50421
doi:10.2196/50421
PMID:38441944

©Rebecca Baines, Sebastian Stevens, Daniela Austin, Krithika Anil, Hannah Bradwell, Leonie Cooper, Inocencio Daniel Maramba, Arunangsu Chatterjee, Simon Leigh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding Symptom Self-Monitoring Needs Among Postpartum Black Patients: Qualitative Interview Study

Natalie Benda[1], PhD; Sydney Woode[2], BSc; Stephanie Niño de Rivera[1], BS; Robin B Kalish[3], MD; Laura E Riley[3], MD; Alison Hermann[4], MD; Ruth Masterson Creber[1], MSc, PhD, RN; Eric Costa Pimentel[5], MS; Jessica S Ancker[6], MPH, PhD

[1]School of Nursing, Columbia University, New York, NY, United States

[2]Department of Radiology, Early Lung and Cardiac Action Program, The Mount Sinai Health System, New York, NY, United States

[3]Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York, NY, United States

[4]Department of Psychiatry, Weill Cornell Medicine, New York, NY, United States

[5]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States

[6]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Natalie Benda, PhD
School of Nursing
Columbia University
560 West 168th Street
New York, NY, 10032
United States
Phone: 1 212 305 9547
Email: nb3115@cumc.columbia.edu

Abstract

Background: Pregnancy-related death is on the rise in the United States, and there are significant disparities in outcomes for Black patients. Most solutions that address pregnancy-related death are hospital based, which rely on patients recognizing symptoms and seeking care from a health system, an area where many Black patients have reported experiencing bias. There is a need for patient-centered solutions that support and encourage postpartum people to seek care for severe symptoms.

Objective: We aimed to determine the design needs for a mobile health (mHealth) patient-reported outcomes and decision-support system to assist Black patients in assessing when to seek medical care for severe postpartum symptoms. These findings may also support different perinatal populations and minoritized groups in other clinical settings.

Methods: We conducted semistructured interviews with 36 participants—15 (42%) obstetric health professionals, 10 (28%) mental health professionals, and 11 (31%) postpartum Black patients. The interview questions included the following: current practices for symptom monitoring, barriers to and facilitators of effective monitoring, and design requirements for an mHealth system that supports monitoring for severe symptoms. Interviews were audio recorded and transcribed. We analyzed transcripts using directed content analysis and the constant comparative process. We adopted a thematic analysis approach, eliciting themes deductively using conceptual frameworks from health behavior and human information processing, while also allowing new themes to inductively arise from the data. Our team involved multiple coders to promote reliability through a consensus process.

Results: Our findings revealed considerations related to relevant symptom inputs for postpartum support, the drivers that may affect symptom processing, and the design needs for symptom self-monitoring and patient decision-support interventions. First, participants viewed both somatic and psychological symptom inputs as important to capture. Second, self-perception; previous experience; sociocultural, financial, environmental, and health systems–level factors were all perceived to impact how patients processed, made decisions about, and acted upon their symptoms. Third, participants provided recommendations for system design that involved allowing for user control and freedom. They also stressed the importance of careful wording of decision-support messages, such that messages that recommend them to seek care convey urgency but do not provoke anxiety. Alternatively, messages that recommend they may not need care should make the patient feel heard and reassured.

Conclusions: Future solutions for postpartum symptom monitoring should include both somatic and psychological symptoms, which may require combining existing measures to elicit symptoms in a nuanced manner. Solutions should allow for varied, safe

XSL•FO
RenderX

interactions to suit individual needs. While mHealth or other apps may not be able to address all the social or financial needs of a person, they may at least provide information, so that patients can easily access other supportive resources.

(J Med Internet Res 2024;26:e47484) doi:10.2196/47484

KEYWORDS

maternal mortality; patient-reported outcomes; patient-reported outcome; health equity; qualitative research; mobile health; mHealth; qualitative; postpartum; postnatal; maternity; maternal; Black; women's health; ethnic; design need; mortality; death; decision support; information need; informational need; obstetric; obstetrics; mental health; mobile phone

Introduction

Background

This study focused on designing a culturally congruent mobile health (mHealth) app to support postpartum symptom monitoring, as the current practice does not adequately support patients in identifying the warning signs of pregnancy-related death (PRD). First, we describe the public health case for symptom monitoring and decision support for PRD, specifically among US-based, Black patients, a group that faces severe disparities [1,2]. Next, we discuss why the current mechanisms for symptom monitoring and decision support are insufficient. We then outline the existing solutions while also emphasizing the need for new interventions, particularly why those using a combination of mHealth and patient-reported outcomes (PROs) may be appropriate. Finally, we introduce a conceptual model used to accomplish our study objectives.

PRD and Associated Health Disparities

The pregnancy-related mortality ratio has increased by >200% in the United States in the past 2 decades, and in a recent review of PRDs, experts estimated that 80% of the deaths were preventable [3]. The Centers for Disease Control and Prevention (CDC) defines PRD as "the death of a woman while pregnant or within 1 year of the end of pregnancy from any cause related to or aggravated by the pregnancy" [4,5]. Mental health conditions (22.7%), hemorrhage (13.7%), cardiac and coronary conditions (12.8%), infection (9.2%), thrombotic embolism (8.7%), and cardiomyopathy (8.5%) have been cited as the most common causes for PRD [3]. Although the global maternal mortality rate has declined, the global rates are still high with 287,000 people dying following childbirth in 2020. There are significant disparities in maternal mortality based on a country's income, with almost 95% of the cases occurring in low- and middle-income countries [6]. Stark disparities in pregnancy-related outcomes in the United States, such as PRD, exist based on race. Specifically, Black or African American (henceforth, referred to as "Black") perinatal patients experience PRD 3 times more than White perinatal patients [1,2,7-10].

The disparities in maternal health outcomes experienced by Black patients in the United States are based on inequitable access to care, biased treatment, and inadequate communication, driven by systemic racism and all the cascading effects it creates. Black perinatal patients are significantly more likely to be uninsured and significantly less likely to have a usual source of medical care (eg, a primary care clinician) than White patients [7,10]. When Black patients seek care, they face implicit biases that negatively affect care quality and health outcomes [1,7,10-12]. Unsurprisingly, these biases have led to reduced trust in the health care system among Black patients [13-17]. Black patients also receive less patient-centered communication and feel that they have poorer access to communication with their medical team [10,18,19]. Our study aimed to improve the patient centeredness of information and support for Black patients in the postpartum period through a participatory design, an approach by which representative end users are involved throughout the design process [20-23]. While this study focused on Black postpartum patients in the United States, we believe that our findings may provide insights for improving perinatal support for patients from minority groups globally.

Challenges to Supporting Symptom Recognition and Treatment Seeking Post Partum

Patients encounter several challenges recognizing concerning postpartum symptoms. First, the initial postpartum visit occurs 6 weeks after birth, and 86% of PRD cases occur within the first 6 weeks post partum [24,25]. Second, most strategies for improving postpartum outcomes focus on hospital-based solutions, which rely on people recognizing symptoms and contacting a health professional [7]. Most counseling regarding the warning signs of PRD occurs during the discharge process following delivery, when people are physically exhausted from childbirth and primarily focused on infant care [24]. As such, this is a suboptimal time for patient education about postpartum risk factors. Discharge nurses report spending <10 minutes on the warning signs of postpartum issues, and most nurses could not correctly identify the leading causes of PRD, making it unlikely that their patients could recognize the warning signs [26]. There are many measures for postpartum symptom reporting, but the most common instruments focus narrowly on specific mental health issues, many of which are not specific to postpartum mental health or postpartum health–related quality of life [27]. While these are helpful measures to use in a clinic or hospital setting, they do not provide real-time decision support regarding the full spectrum of severe symptoms that may be indicative of PRD.

Suitability of Different Solutions for Supporting Symptom Monitoring

mHealth can address the need for tailored, dynamic symptom monitoring and support. The Association of Women's Health, Obstetric, and Neonatal Nurses and the CDC have developed 1-page summaries to help patients identify the warning signs of PRD, such as the Urgent Maternal Warning Signs (UWS) [28,29]. These tools represent a positive step toward improving symptom management, but these solutions do not provide real-time, tailored support. Telephone-based support staffed by

XSL•FO
RenderX

health professionals has been demonstrated to decrease postpartum depression and improve maternal self-efficacy [30-33]. However, 24-hour hotlines can be resource intensive, and people may still experience bias when accessing these services. The goal of this study was to conduct a qualitative needs assessment for the Maternal Outcome Monitoring and Support app, an mHealth system using PROs to provide decision support for postpartum symptom monitoring.

Mobile phones offer a viable, inclusive option for intervention delivery for Black people of childbearing age. In 2020, data from the Pew Research Center indicate that 83% of Black people owned smartphones, which is comparable to smartphone ownership among White people (85%). Smartphone ownership is also higher among people aged <50 years (96%), which encompasses most postpartum patients [34]. However, Black people are twice as likely as White people to be dependent on smartphones for internet access [35]. mHealth-based apps for blood pressure and weight tracking during pregnancy have demonstrated success among diverse groups, providing evidence that mHealth may be an acceptable means for symptom reporting in the target population [36-38].

Symptom education and PRO-based interventions have demonstrated success in improving knowledge, self-efficacy, and outcomes. Use of PROs has improved symptom knowledge, health awareness, communication with health care professionals, and prioritization of symptoms in patients with chronic disease and cancer [39-44]. Multiple studies have also demonstrated that educational interventions regarding expected symptoms in the postpartum period can improve self-efficacy, resourcefulness, breastfeeding practices, and mental health [12,38,45-47]. However, given the issues related to trust and disparities in patient-centered communication, it is critical to understand Black patients' perspectives about how such a system should be designed and implemented.

Conceptual Model

To study the issue of supporting symptom monitoring, we combined 2 theoretical frameworks (Figure 1): the common sense model of self-regulation (health behavior) by Diefenbach and Leventhal [48] and the model of human information processing (human factors engineering) by Wickens [49]. The model by Diefenbach and Leventhal [48] depicts patients as active problem solvers with a mental model of their conditions. Patients process their symptoms, both cognitively and emotionally, and then evaluate whether action is needed [48]. The patient's mental model of their condition, personal experiences, and sociocultural factors impact processing, evaluation, and action. In the information processing model by Wickens [49], action occurs in 2 steps—selection and execution [48]. Environmental or organizational factors also affect patients' selection of actions and whether they can execute an action. For example, a patient may suspect that they should visit the emergency room but may not go because they do not have insurance, transportation, or childcare. Our qualitative inquiry investigated how to better support symptom processing and appropriate response selection, while also uncovering the barriers to action that may need to be mitigated.

Figure 1. Theoretical model for postpartum symptom processing and self-regulation.

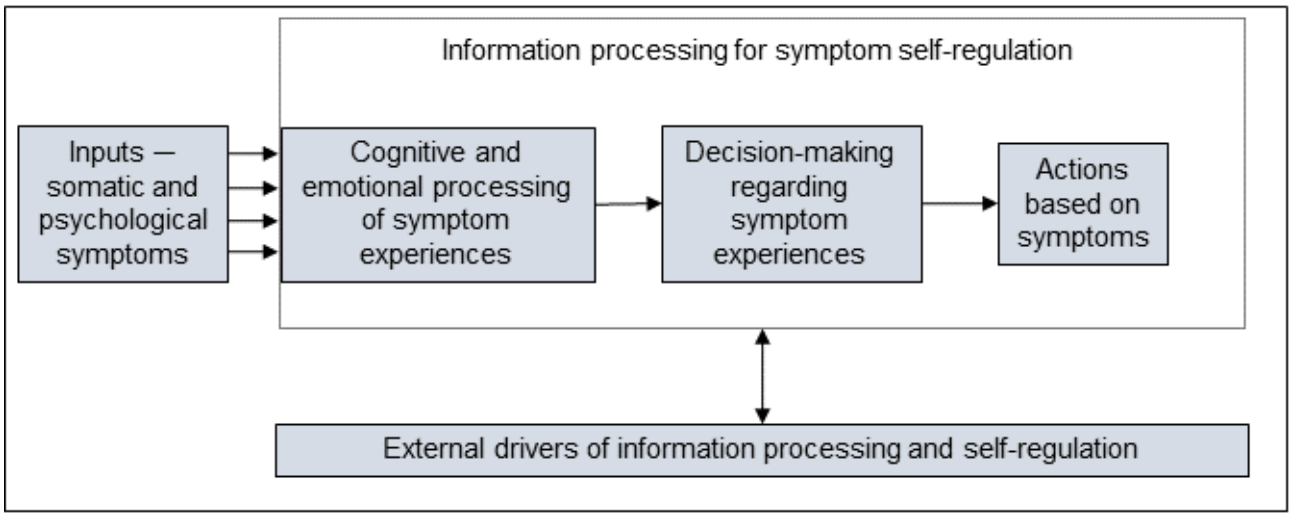

Objectives

The goal of this study was to identify the design and implementation needs of an mHealth-based symptom self-monitoring and decision-support system to support Black patients in determining when to seek care from a health professional for signs of PRD in the postpartum period. This tool will support both somatic and psychological symptoms given their complex, critical, and connected presentation. We used the described conceptual model in qualitative inquiry and pragmatic intervention design to provide contributions regarding the following: (1) relevant symptom inputs for postpartum support, (2) drivers that may affect symptom processing, and (3) how the previous 2 aspects highlight the design needs for symptom self-monitoring and patient decision support. To address our study objective, we conducted semistructured interviews with postpartum Black patients, obstetrics health professionals, and mental health professionals.

XSL•FO
RenderX

Methods

Setting

The study was conducted in 3 tertiary care hospitals and affiliated clinics within the same health system in New York City. The 3 hospitals, taken together, are involved in the delivery of >14,000 babies annually. All participants were either patients who received obstetric care in the included sites or health professionals affiliated with the sites.

Eligible patients were identified by the institutions' research informatics team using electronic health record data. First, the patients' providers consented to their patients being contacted, and patients' charts were reviewed by the primary obstetrician or designate to ensure that the patient was eligible for the study and that they had a delivery experience that would allow them to participate in the interview without undue stress. Next, the patients were sent an invitation to participate via the email address listed in their record. We also posted fliers in 2 high-risk, outpatient obstetric clinics.

Obstetric and mental health professionals were eligible if they were affiliated with one of the institutions in the obstetrics or mental health department. Brief presentations were given at relevant faculty meetings, and participants were contacted individually via email or through departmental listserves.

Interested participants from all groups used a link to schedule a time to speak with a researcher.

Ethical Considerations

The study was approved by the affiliated medical schools' institutional review board (protocol number 20-08022582). All participants provided written informed consent. Study data were coded (ie, all identifying information was removed) to protect participant privacy. Each participant was compensated US $50 for their time via a physical or electronic gift card.

Study Design and Sample

The study used semistructured interviews with 3 key stakeholder groups: recent postpartum Black patients, obstetric health professionals, and mental health professionals. Eligible patients were within 12 months post partum of a live birth, self-identified their race as Black or African American, and had at least 1 somatic or psychological high-risk feature associated with their pregnancy. High-risk features included attendance at a high-risk clinic for prenatal or postnatal care, inpatient hospitalization within 12 months post partum, a prescription of an antidepressant or benzodiazepine within 12 months of the pregnancy, or a new diagnosis of depression or anxiety within 12 months of the pregnancy. High-risk clinics treated various conditions, but the most common conditions were gestational hypertension and gestational diabetes.

We adopted an interpretivist qualitative research paradigm to study patient and health professionals' perspectives of how symptom recognition and care seeking may be better supported [50]. Our methodological orientation involved directed content analysis, adopting an abductive reasoning approach. First, we used the previously specified conceptual model to construct questions and thematically categorize responses [48]. Then, we allowed unique subthemes to inductively emerge from the data collected [51].

Interview Guide Development

Interview guides were iteratively developed by our team of researchers with expertise in obstetrics, perinatal mental health, nursing, consumer informatics, inclusive design, and qualitative methods. The guide for each stakeholder group was reviewed and piloted before enrollment of the first participant. Interview guides were tailored for patients or health professionals but followed a similar structure, based on our conceptual model (Figure 1), such that participants were first asked about barriers to and facilitators of processing symptoms cognitively and emotionally (eg, Do they notice the symptom or realize its severity?), making decisions about symptoms they are experiencing (ie, When to seek help from a health professional?), and taking action on problematic symptoms. Probing questions encouraged participants to elaborate on experiential, educational, sociocultural, organizational, environmental, or health systems–level drivers of patients' symptom management. Then, participants were asked a series of questions related to their thoughts regarding the design of the mHealth system, including how to best report symptoms, the wording of system decision support, the desired level of involvement of the obstetrics health professionals, the means for facilitating outreach to a health professional, additional information resources, and preferences for sharing information included in the system with a trusted friend or family members. During this process, obstetrics and mental health professionals were also shown a handout that outlined the draft of the symptom management algorithm for the system being developed (CDC's UWS) and asked if they would make any changes, additions, or deletions [29]. Full interview guides are included in Multimedia Appendix 1.

Data Collection

All interviewees provided consent electronically before the interview. A PhD-trained qualitative research expert (NB) completing a postdoctoral study in health informatics and population health conducted all the interviews via Zoom (Zoom Video Communications) or telephone. Participants had the option to request an in-person interview, but none of them chose this option. Interviews lasted 30 to 60 minutes and were audio recorded. We explicitly described the study objectives to each participant before the interview. Following the interview, participants completed a demographics survey electronically. All electronic survey information was collected using REDCap (Research Electronic Data Capture; Vanderbilt University).

Data Preparation and Analysis

Audio recordings were converted into transcripts using an electronic software (NVivo Transcription; QSR International) and manually checked for accuracy by a study team member who did not conduct the initial interviews. We completed all data analyses using NVivo (versions 12 and 13), but we manually analyzed the data and did not use computer-aided techniques (eg, computerized emotion detection or autocoding).

Data were analyzed using thematic analysis and the constant comparative process [51-53]. Specifically, each analyst open coded the transcripts, by coding segments that pertained to the

XSL•FO
RenderX

research questions, as opposed to coding all words and phrases. We used thematic analysis to detect the common and divergent needs for postpartum symptom monitoring. We chose this method over other approaches such as grounded theory or sentiment analysis because our needs were pragmatic to solution design, and we were not attempting to establish theory, describe phenomena, or represent collective feeling about a topic.

The first deductive analysis was conducted using an initial theoretical model derived from the common sense model by Diefenbach and Leventhal [48] and the model of human information processing by Wickens [49] (Figure 1). To promote reliability, 2 coders in addition to the interviewer were involved in the analysis, and each transcript was first analyzed independently by at least 2 people (NB, SW, or SNdR), followed by meetings to resolve discrepancies based on consensus coding. The analysis team created initial codes based on the conceptual model and added new items to the codebook inductively (ie, post hoc instead of a priori, as they arose in the data). The team used NVivo to maintain a working codebook of themes, definitions, and relevant quotes derived from the data. The codebook was periodically presented to coinvestigators with expertise in obstetrics and perinatal psychiatry to improve external validity [51,52]. The sufficiency of sample size was assessed according to the theoretical saturation of themes encountered, specifically based on the need to add additional subthemes to the codebook [54,55]. After all the transcripts had been coded, at least 2 members of the coding team reviewed the data code by code to ensure that meaning remained consistent throughout the analysis and to derive key emerging themes [51].

Results

Participant Characteristics

This study included 36 participants—15 (42%) obstetrics health professionals, 10 (28%) mental health professionals, and 11 (31%) recent postpartum Black patients. Table 1 presents the self-reported demographic information. As shown, 19% (7/36) of the health professionals and 11% (4/36) of the patients had missing data (ie, did not complete the questionnaire). Participants could also selectively choose not to answer questions. “Other” affiliations were possible for health professionals because those who had a secondary affiliation with one of the included sites but primary affiliation with another organization were eligible.

XSL•FO
RenderX

Table 1. Summary of the study participants' self-identified characteristics collected using the survey immediately following their interviews.

Characteristics	Total (N=36), n (%)	Health professionals (n=25), n (%)	Patients (n=11), n (%)
Health professional role			
Advanced practice provider (eg, nurse practitioner or midwife)	4 (11)	4 (16)	0 (0)
Attending physician	7 (18)	7 (28)	0 (0)
Licensed clinical social worker	2 (5)	2 (8)	0 (0)
N/A[a]	13 (34)	0 (0)	11 (100)
Nurse	1 (3)	1 (4)	0 (0)
Other[b]	4 (11)	4 (16)	0 (0)
Missing	5 (14)	7 (28)	0 (0)
Affiliation or point of care			
Other	1 (3)	1 (4)	0 (0)
Site 1	23 (61)	14 (56)	9 (69)
Site 2	1 (3)	1 (4)	0 (0)
Site 3	2 (5)	2 (8)	0 (0)
Missing	9 (25)	7 (28)	2 (18)
Gender			
Female	26 (68)	16 (64)	10 (77)
Male	2 (5)	2 (8)	0 (0)
Missing	8 (22)	7 (28)	1 (9)
Year of birth			
1965-1980	6 (16)	5 (20)	1 (8)
1981-1996	22 (58)	13 (52)	9 (69)
Missing	8 (22)	7 (28)	1 (9)
Race			
American Indian or Alaska Native Asian	1 (3)	0 (0)	1 (8)
Black or African American	11 (29)	2 (8)	9 (69)
Native Hawaiian or other Pacific Islander	1 (3)	1 (4)	0 (0)
White	11 (29)	11 (44)	0 (0)
Missing	12 (33)	11 (44)	1 (9)
Ethnicity			
Hispanic or Latino	1 (3)	1 (4)	0 (0)
Not Hispanic or Latino	25 (66)	16 (64)	9 (69)
Missing	10 (28)	8 (32)	2 (18)

[a]N/A: not applicable.

[b]Health professionals' self-reported role of resident psychiatrist, chief resident in psychiatry, psychologist, and patient care director was combined into the *other* category for analysis purposes.

Structure of Themes

Overview

Our initial theoretical model, derived from the common sense model by Diefenbach and Leventhal [48] and the model of human information processing by Wickens [49] (Figure 1), described that patients experience some inputs (psychological and somatic symptoms of PRD). Then, there is a series of drivers that affect how patients cognitively and emotionally process (eg, notice and realize symptom severity), make decisions about, and act on symptoms they are experiencing. The nature of these symptoms, how they are processed, how decisions are made, and how they are acted upon then drive a conversation regarding the design needs for symptom monitoring and decision support for PRD. The emerging themes were organized into the following categories: (1) symptoms of PRD; (2) drivers of

XSL•FO
RenderX

processing, decision-making, and action; and (3) design needs for a symptom-reporting and decision-support system. Quotes are labeled with study-specific identifiers: OB denotes obstetric health professional, MHP denotes mental health professional, and PT denotes patient.

Inputs: Psychological and Somatic Symptoms of PRD

Concerning and routine symptoms were reported both from a psychological and somatic perspective. Sometimes, the distinction between routine and concerning symptoms was clear. Other times, it was more challenging to differentiate routine versus concerning symptoms particularly because they were related to psychological health. Mental health professionals also noted the challenge that routine symptoms can progress to something more serious over time:

> *In my mind, like normal becomes abnormal, when there is any kind of functioning [loss] that like withstands two to three weeks.* [MHP 04]

> *We really hear a lot about postpartum depression and stuff...A lot of women think...postpartum depression is you just don't want to. You don't have it. You go into depression where you can't take care of your child and you don't want to hold your child. You don't feel connected to your child. And I learned...it can be so many different things.* [PT 09]

A clear distinction was not always present between psychological and somatic symptoms:

> *If someone...has pain in their chest or shortness of breath, the first thing you want to think about is it sort of like clots and other kind of physiologic reasons for that. Those are also very implicated and sort of obviously [associated with] panic attacks and anxiety. So, I think though those symptoms are also relevant of physical symptoms, [they] are also relevant for mental health.* [MHP 05]

Drivers of Processing, Decision-Making, and Action Based on the Symptoms Experienced

Several drivers were reported to affect symptom processing (ie, whether they noticed the symptom and its severity), patients' capacity to decide what should be done (ie, make decisions), and whether they were able to act on concerning symptoms (Table 2).

Table 2 presents exemplary quotes for emerging themes under a single driver, but many quotes were coded under multiple drivers in our analysis process. The following passage, for example, highlights how self-perception, sociocultural concerns, and the health system can overlap to present a complex set of factors that may prevent women from receiving the care they need for the symptoms they are experiencing:

> *A lot of times I think that does get overlooked because people feel like, well, you're OK, you're fine. But what research shows us is that especially for Black women, it really doesn't matter how much money you make or your income level, like our postpartum and perinatal health outcomes are the same across the board, which is really detrimental. So, yeah, I think they get overlooked because of that. I think they get overlooked or we get overlooked in the health care system. But I also think we get overlooked by our family and friends because we're the strong ones. So, if anybody can deal with this, it's you.* [MHP 10]

XSL•FO
RenderX

Table 2. Descriptions of the drivers of symptom processing, decision-making regarding symptoms, and taking action based on symptoms patients' experience in the postpartum period.

Drivers	Description of emerging themes	Example quotes
Self-perception	Health professionals and patients described the challenges in recognizing symptoms based on a lack of knowledge regarding what is normal during the perinatal period and the continuous focus on their newborn, not allowing time for self-reflection.	• "But I would say...one of the hard things I think for women is, especially for first time mothers, they don't know what it's supposed to feel like postpartum. They have nothing to compare it to...And I find that that's the biggest barrier for women seeking out care, is that they just don't know what's normal or not normal." [MHP[a] 02] • "I was very much attached to her [patient's child], but you just felt like you felt down, and you felt like, am I doing enough? So, I never really felt like I was doing enough. Um, and you're exhausted, right? So, when you're exhausted can't process too much." [PT[b] 13]
Previous experience	Patients' previous experiences could either help them know what to look for or could prevent them from seeking help, if their previous experience was negative.	• "It's kind of hard to say because I already know what to look out for because I had postpartum with my first kid. So, I kind of already know I didn't really need much education in that area." [PT 05]
Sociocultural factors	Social factors could be supportive or prohibitive based on how the patients' support system viewed seeking help, particularly for mental health. Participants also explained how stigma plays a strong role in processing symptoms and seeking help for mental health–related issues. Participants articulated that social factors may be influenced by cultural norms based on age, race, country of origin, or previous experiences with mental health. Lack of social support (ie, not having someone to help with childcare while seeking medical attention) was also described as a barrier to acting upon concerning symptoms.	• "On one hand...the patient might not have full insight into how they're doing and what's going on for them, and the partner might have a lot of insight...It could be very positive to have that information. At the same time, I have also seen partners who then either exaggerate or pathologize them or don't understand some of the changes...but are just part of becoming more preoccupied in that phase with the infant and infants care and that status parenthood. And so that can be a very...difficult and alienating experience for the mother." [MHP 06] • "I live with my son's father...He was there, so he was the one that told me to go. And I told my mom, she was like, I should go in. And that's when I called the doctors because it was a little concerning." [PT 01] • "They feel like they should be able to do it. They feel like this is, you know, other women do this and they don't have any issues. So, why am I having difficulty with this?...They feel like they're weak, they're incapable, or incompetent. And yeah, that's definitely a barrier." [MHP 02] • "I'm thinking of one patient who, like, she was really like every time was so close, had her see a therapist and had her see me...The location background, really, like grew up in an environment where like, psychiatry is just not something that you do and you should tough it out, which was so impaired, like this is like years and years and like, it just got to be more severe." [MHP 01] • "I think there's just so much stigma and shame about having difficulties in that initial period of time." [MHP 03] • "And I think that's like the most difficult thing for them to deal with, especially if, like the father of the child is not involved or like there's no family in the area. And I don't honestly see it very often. But when I when there is an issue with, like someone being admitted or someone coming in, it's usually because there's another child." [OB[c] 03]
Financial and insurance-related factors	Financial and insurance-related factors were described as potential barriers to seeking care. Cost-related issues were particularly prevalent during discussions about mental health.	• "So, I think there's a fractured system. I think we're kind of isolated and working in our silos to a large extent, and then if we really talk about mental health...the big ugly secret is we don't take insurance. Most people you go on Psychology Today and you look for perinatal, these people are not taking insurance and they're certainly not taking Medicare or Medicaid." [MHP 08] • "So, the first one is insurance and the cost of care, which is a big one, and a lot of people will choose not to go to therapy because it's not affordable." [MHP 07]

XSL•FO
RenderX

Drivers	Description of emerging themes	Example quotes
Built environment	Issues related to the patient's environment such as how far away they lived, traffic, and parking were described as barriers to seeking and receiving care.	• "Like traffic is like never ideal, you know, during certain times of the day. Parking in the area is not ideal...So, like, I can only imagine that, like especially like immediately postpartum with like a crying baby at home, like I have to drive into the city." [OB 03] • "Some patients use public transportation if they live in the boroughs. And so, they don't have access to a vehicle that they can easily get back to the hospital." [OB 04]
Health system–related factors	There were various levels of health system factors that could impact how patients processed symptoms, made decisions, and took action. Patients described individual positive experiences of how processes and their care team prepared them for the types of symptoms they should look out for. However, health professionals described anecdotes of how patients' previous experience of bias and systemic racism prevented women from seeking care.	• "When I was being discharged the first time, they were very clear about what numbers should be alarming or would be cause for concern. And so, I had my discharge papers, they were handy. But by that point, I think it was just pretty much ingrained what the numbers I should be looking for were. But...the discharge papers were very clear." [PT 02] • "Yeah, patients can tell when they're under a microscope. And especially Black women are scared when they come into the hospital. They know that Black women are more at risk of dying in the hospital...and that their needs will not be heard, that their expressions for pain or need will be ignored...or they generally just mistrust all of us." [MHP 07] • "There is one attending who is very involved in how patients are feeling and their concerns and everything and her patients happen. Her patient was Black and she just had a lot of anxiety. She just kept saying, If I go to the hospital, am I going to die?...You're supposed to go to the hospital to get the help and the care you need, but if you go to the hospital, are you going to die? You know, that's kind of really sad that a patient would feel that way, you know, especially at such an exciting time in her life...So, with that stigma out there, I wouldn't want something like that to have patients avoid getting care because they think they're better off just not going to the hospital and not going to be hurt." [OB 07]

[a]MHP: mental health professional.

[b]PT: patient.

[c]OB: obstetric health professional.

Design Needs for a Symptom-Reporting and Decision-Support System

Obstetric health professionals, mental health professionals, and patients discussed multiple needs for improved PRD symptom reporting and decision support. The key design requirements are embedded and italicized in the following text.

Participants generally agreed that although the proposed system focuses on postpartum symptoms, *it would be advantageous to introduce the system during pregnancy, particularly in the third trimester*:

> *You have to reach women before they give birth. They might look, they might not look, they might look at it and be concerned. But then they might forget about it and not have time to call. Those first six weeks are really chaotic.* [MHP 06]
>
> *I think in the third trimester would be great because often we don't really have anything to talk about in the office. It's very quick visits like blood pressure and you're still pregnant and we're just waiting. And so, I think and they start to have a lot of questions about like, well, when I get home and how's this going to go? So, I think that time is a good time. We're all kind of just waiting for labor to happen or full term to get there, and this kind of gives them something to feel like they can prepare for.* [OB 08]

Patients were open to reminders regarding entering symptoms they were experiencing, and participants described a desire for just-in-time symptom reporting and decision support, so that they could get quick feedback as they were experiencing the symptoms:

> *When people get home so much in their life has changed. And it's probably a very hectic time. So maybe I think that's a great idea reaching out again, either a few days or a week later to make sure they're really able to use it and engage with it to the extent that's helpful to them.* [OB 02]
>
> *I think it would be a good idea to have like a system where you can report whenever you want.* [PT 03]
>
> *I think for me, I would say in the moment. But then also having something at the end of every week to just, you know, to check in with yourself. I think that would be good as well.* [PT 09]

In addition to considerations about how symptoms would be recorded, participants stressed the *importance of the wording of the decision-support messages that patients receive*. For

XSL•FO
RenderX

messages that inform the patient that their symptom did not seem to require immediate medical attention, *it was important to ensure that the patient still felt heard and that they did not leave the interaction feeling stuck* with nothing to do regarding a symptom that was concerning to them:

> *Reframe the message. You know...we apologize that you were experiencing this. We just want to reassure you that this is normal.* [PT 01]
>
> *[You] don't want to make anyone feel like their feelings aren't valid because that's a horrible thing, especially in health care, especially if a person is convinced that something is wrong with them and you're telling them that it's normal and is perfectly fine. So, in that situation, I would just, depending on what the issue is, I would also share information of what to look out for.* [PT 05]
>
> *The first thing is that it's normal, but also something that you want to be able to do for comfort. For me, I don't have to do too much, especially if I'm having anxiety, like if I get a text back that says here are some things you can do in this very moment to handle it. And then also, here are some links or information that you can also look up.* [PT 09]

In the events where a concerning symptom was reported and it was recommended that the patient should reach out to a health professional, importance of conveying a sense of urgency without scaring the patient:

> *You don't want to scare people, but it's kind of hard to get around that when something is serious, and you don't want to dumb it down.* [PT 01]
>
> *Participants wanted multiple, easy-to-do methods for connecting with their health professional team, including having the number to call pop up, scheduling a time for someone to call them, and being able to start a live web-based chat:*
>
> *I like all the options, especially that form or chat you can have like, you know, those online chat where like you really chatting with someone for those who like the type. I'm the type of person I just want to make a phone call, right? So, like for me, [it] will be a call. Maybe say maybe if it's five, five or ten minutes then that will be great. Like especially, it's going to make me feel like, OK, there's someone out there that will care about my health.* [PT 06]

However, participants noted that they would prefer not to use a symptom-reporting and decision-support tool, but instead reach out directly via phone if they were experiencing issues.

Participants, particularly mental health professionals, described a *need for improved nuance or details regarding the different psychological symptoms patients could experience* that are indicative of severe mental health issues:

> *Thoughts of hurting yourself or someone else is a good one...I would say I would add difficulty bonding. It would add something about not being able to sleep, even if you could sleep, you know, like or your anxiety that doesn't go away, that changes your behavior. So, it changes the way that you interact with the baby or kind of do childcare. I guess I would want to say something about. psychotic thoughts, like fear that someone else may be hurting you or...recurrent worries or anxieties that don't go away.* [MHP 02]

Patients had differing opinions regarding whether the system should be integrated with other health technologies, particularly the patient portal:

> *I love the patient portal. I was able to be traveling to reach out to my OB, to reach out to all, you know, the nurses and stuff like that and just experience things that I needed.* [PT 09]
>
> *I feel like...it's an integral part of my medical history. So, even if it may seem somewhat insignificant for whatever reason, I would still want to have access.* [PT 09]
>
> *I didn't find it [the patient portal] very helpful...* [PT 03]

On the basis of the feedback from health professionals that it may be challenging for postpartum patients to process and recognize certain symptoms, especially those related to mental health, we explored whether patient participants would be open to sharing educational information about symptoms to expect (rather than sharing the actual symptom reports) with trusted friends or family members. Similar to other design considerations, results were mixed, but it seemed helpful to have a *patient-driven option for sharing symptom-related educational information with chosen friends or family members*:

> *I think that there's so much going on it would help to have someone with a different perspective equipped with this information.* [PT 02]
>
> *There's a lot of shame that comes with this. I'm not sure people would actually want other people to know. I can't speak for the majority, but I didn't really want people to know because I don't want the kind of energy that came with people knowing.* [PT 05]

We also discovered the competing needs of *balancing the patient's desire for their health professionals to be involved in symptom reporting with the need to avoid significant increases to health professional workload*:

> *I sort of wonder from the health care provider perspective, how involved is the provider in that in the app? Like, do they get like a PDF of all the information? Is that more work for the provider? How does the provider interpret that data?* [MHP 03]
>
> *I feel like they [the health professional] should be super involved. Especially because I'm not just going off of my experience because, you know, I don't want to feel like they're not really like I'm experiencing. And so, it's scaring me. So, I just want to know that, you know, you're hands on with everything.* [PT 01]

Finally, *the participants desired information beyond PRD symptoms to entice them to use the system*. They were supportive of including various types of information, such as breastfeeding support resources, milestones and information regarding their child, other websites and apps with trusted maternal and child

XSL•FO
RenderX

health information, further support resources for how they feel mentally, and links to social services (eg, food, housing, or other assistance).

Discussion

Principal Findings

In this qualitative study, we interviewed obstetric health professionals, mental health professionals, and Black postpartum patients. Our findings helped to identify the design and implementation needs of an mHealth-based, symptom self-monitoring and decision-support system designed to support Black patients in determining when to seek care from a health professional for signs of PRD in the postpartum period. We encountered important findings related to (1) inputs, including psychological and somatic symptoms; (2) drivers of processing, decision-making, and action based on the symptoms experienced; and (3) design needs for a symptom-reporting and decision-support system. We have discussed how our findings may be helpful to other postpartum populations as well as the implications of our study for patient decision-support in other clinical settings.

First, our findings related to symptom inputs revealed the challenges caused by the overlapping presentation of somatic and psychological symptoms. This provides support for our approach of including psychological and somatic issues in a single app, particularly given that mental health conditions are a leading cause of PRD. A 2021 review found 15 PRO measures for assessing postpartum recovery. The measures typically focused on mental health or health-related quality of life, but few included both psychological and somatic outcomes, and none were targeted for PRD, such as the system [56].

Moreover, related to symptom inputs, we found that current tools for pinpointing severe symptoms, such as the CDC's UWS did not provide sufficient nuance for concerning psychological symptoms. Symptom-reporting tools for PRD will either need to consider incorporating structured assessments, such as the Edinburgh Postnatal Depression Scale (EPDS) [56], or incorporating additional symptoms. The latter approach may have advantages as the EPDS focuses on depression (while providing subscales for anxiety) and PROs evaluated for use with anxiety disorders have limitations [57]. Furthermore, the EPDS has been validated in in-person laboratory settings but not in community settings or for web-based entry [58]. We must also consider how mistrust in the health system may lead to less truthful answers. Issues expressed around stigma related to mental health indicate that the way in which these symptoms are elicited may require further assessment to promote the normalcy of the symptoms and improve candid reporting. Technology-based approaches for supporting perinatal mental health have been described as uniformly positive but having limited evidence for use [59], suggesting that further exploration is needed in this area, also considering how adding somatic issues may be perceived by patients.

Second, there were several drivers that affected symptom processing, decision-making, and action that cannot typically be solved through a symptom-reporting and decision-support system. Challenges related to self-perception and lack of experience or expectations may be addressed based on the wording for how the symptoms are elicited and by providing concise, easy-to-understand depictions of what should be expected versus what are the causes for concern. However, many of the other issues described related to sociocultural, financial, and environmental factors and the health systems' systemic racism issues cannot be addressed directly in a simple PRO-based app and decision-support system. Directly addressing these issues will likely require more systematic, multipronged approaches. Therefore, it seems advisable to couple patient decision-support aids with other social support interventions for perinatal health [60,61].

Drivers of processing, decision-making, and action are still important contextual elements to be considered in the design of the system. Another study tailoring an mHealth app for Latina patients to support health during pregnancy also found it important to address issues related to financial barriers, social support, health care accessibility, and cultural differences [62]. Our best attempt to address these issues may be to promote information transparency and inclusive design. For example, there may be a "frequently asked questions" section of an app, where patients can explore things such as supportive resources for childcare while they seek medical attention or information they may show their friends or family members regarding postpartum symptoms of concern. The system may also use common human-computer interaction principles, such as information filtering [63] and organizing the suggested resources (eg, for mental health care) based on whether they accept the patient's insurance. The built environment can also be changed through the system, but it may offer mechanisms for remote monitoring, such as telemedicine-based support or linking the system to a blood pressure cuff, when clinically appropriate [64,65]. As noted, the system obviously cannot address issues related to systematic racism directly [66]. Instead, we used a participatory design approach, with the hope that the nature of the information presented may be more patient centered, acceptable, and better aligned with the beliefs and values of Black patients [67]. Issues related to systematic racism have commonly been described in the US health care system, but structural inequities also exist on a global scale. Future studies should investigate how our findings regarding design needs may extend to other minoritized perinatal patient groups.

A systematic review of patient decision aids for socially disadvantaged populations across clinical settings found that such tools can improve knowledge, enhance patient-clinician communication, and reduce decisional conflict [68]. However, descriptions of patient decision aids focus on the type of tool (eg, paper vs digital), how it was delivered, when it was delivered, and by whom, as opposed to describing the content the aid provides. Therefore, it is challenging to determine how other decision-support tools have addressed information regarding environmental, financial, or health system–level factors that may affect care seeking based on the decision aid. Some tools seem to address sociocultural needs by tailoring to the target population, but the aforementioned systematic review did not find differential effects on outcomes when tools were tailored versus not tailored [16]. Future studies on patient

XSL•FO
RenderX

decision aids may benefit from including non-symptom related information. Providing appropriate informational support may involve a deeper study of the systemic needs that patients may have, even if these needs may not directly be addressed by the decision aid.

Third, descriptions of the design needs for PRD symptom monitoring revealed that there is likely not a one-size-fits-all solution related to reminders, involvement of health professionals, and how the tool is incorporated with other systems (eg, the patient portal). “User control and freedom” and “flexibility of use” are two of the key items in commonly used heuristics for user interface design [69]; therefore, it is important to include options for customization and varied but safe pathways for interaction with the proposed system. For example, some participants described that they may not be likely to access the symptom-reporting system through the patient portal. Although there may be safety and convenience-related reasons for having the system as part of the patients’ medical record, if the patient chooses, the system could, on the front end, appear more like a stand-alone app than something that must be accessed through the patient portal. Patients also had varying opinions related to how they may want to reach out to a health professional if a problematic symptom was reported. These preferences may differ from instance to instance; therefore, it is helpful to ensure that patients have a choice regarding how to reach out, but system designers must also create workflows with feedback loop, so that patients who are reporting problematic symptoms are not missed (ie, if patients do not reach out themselves, they never receive attention). Patient-level customizations and options for interaction also respects patients as individuals and may promote patient-centered interactions.

Furthermore, related to design needs, participants indicated that the wording of the decision-support messages was critical. Specifically, for reports that did not include currently urgent symptoms, it was important that the message still conveyed support and validation, clarified that the patient could still reach out for help, and provided additional means for managing their symptoms, so the patient did not feel frustrated by their report [70]. Regarding messages that recommended patients to reach out to their health professional team, it was crucial to note what the symptom meant (eg, what kind of disease it could indicate), encourage the patient to reach out without increasing anxiety, and provide different avenues for easy outreach. Going forward, we plan to incorporate the aforementioned elements into the messages built into the system. We will then complete additional acceptance and comprehension testing with a larger sample of postpartum patients. These findings also indicate that care must be taken in translating such tools, and the translated materials should be reviewed with the target end user groups before implementation. This may mitigate unintended consequences or inadvertent inclusion of language that does not support the needs of minoritized groups.

Strengths and Limitations

Our study highlighted the limitations and areas that would benefit from further exploration. First, our study involved recruitment sites that were within a single health system in New York City. Second, while we achieved thematic saturation of qualitative themes (a means for determining sample sufficiency in qualitative studies) [54,55], our conclusions are based on a sample of 36 participants from 3 stakeholder groups. Third, given the documented disparities, we deliberately focused on the needs of Black postpartum patients, but this may not represent the needs of the postpartum patients of other races. Furthermore, our sample should not be viewed as encompassing the opinions of all Black postpartum patients. Our findings revealed the need for individual customization and varied interaction patterns on a case-by-case basis. Fourth, all interviews were conducted remotely (via Zoom or telephone), which can have effects on the interaction. On the one hand, it may be harder to connect with the interviewee, and on the other hand, people may feel more anonymous and comfortable with sharing information. Finally, although we attempted to promote external validity through the review of the coding scheme by a subject matter expert, we did not have the opportunity to perform triangulation of the findings by returning the results to participants. To address these limitations, it would be beneficial to survey a larger group of postpartum patients, powered to assess the differences based on race and ethnicity. This would allow us to come to a stronger consensus regarding design choices, assess whether there are differences in design needs or preferences, and gain feedback from patients in areas outside New York City. Future studies may also explore how other underserved groups, such as those with limited English proficiency, may benefit from tailored symptom self-monitoring and decision support.

Conclusions

In this qualitative study regarding postpartum symptom monitoring and decision support, we found that the current structured reporting measures do not include the combination of somatic and psychological symptoms that may be indicative of severe outcomes in the postpartum period. While not explicitly related to symptom reporting and decision support, patient decision aids, particularly those focusing on minoritized groups, should consider how the aids may be coupled with other structural support interventions or, at least, information about how other resources may be accessed. As stated in the commonly accepted design heuristics, we also found that user control and freedom unsurprisingly remain important for a patient decision-support aid for Black postpartum patients. Finally, decision aid–related phrases must take care to convey urgency without inducing anxiety when action may be indicated and consider respect and empathy for the patients’ symptoms when action may not be indicated to ensure that they do not feel unheard and are empowered to report new or worsening symptoms.

XSL•FO
RenderX

Acknowledgments

This study was supported by the National Institute on Minority Health and Health Disparities (K99MD015781; principal investigator: NB).

Data Availability

The data sets generated and analyzed during this study are not publicly available due to institutional review board regulations but are available from the corresponding author on reasonable request.

Authors' Contributions

NB conceptualized the study and acquired funding under the advisement of RBK, LER, AH, RMC, and JSA. NB collected the data. NB, SW, and SNdR analyzed the data with input from all other authors. ECP completed the literature review and descriptive analysis of participants' characteristics. NB drafted the paper and received substantial inputs from all other authors.

Conflicts of Interest

LER is an Up to Date contributor and an advisory board member for the New English Journal of Medicine, and Contemporary OB/GYN. She has also been a speaker for Medscape is an an expert reviewer for Pfizer on the RSV Vaccine. AH is an Up to Date contributor, a co-founder and medical consultant for Iris Ob Health, and a consultant for Progyny.

Multimedia Appendix 1
Semistructured interview guide questions for patients and health professionals.
[DOCX File , 26 KB - jmir_v26i1e47484_app1.docx]

References

1. Howell EA, Egorova NN, Janevic T, Brodman M, Balbierz A, Zeitlin J, et al. Race and ethnicity, medical insurance, and within-hospital severe maternal morbidity disparities. Obstet Gynecol 2020 Feb;135(2):285-293 [FREE Full text] [doi: 10.1097/AOG.0000000000003667] [Medline: 31923076]
2. Wisner KL, Sit DK, McShea MC, Rizzo DM, Zoretich RA, Hughes CL, et al. Onset timing, thoughts of self-harm, and diagnoses in postpartum women with screen-positive depression findings. JAMA Psychiatry 2013 May 01;70(5):490-498 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.87] [Medline: 23487258]
3. Pregnancy-related deaths: data from maternal mortality review committees in 36 US States, 2017-2019. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/reproductivehealth/maternal-mortality/erase-mm/data-mmrc.html [accessed 2022-11-20]
4. Pregnancy mortality surveillance system. Centers for Disease Control and Prevention. 2020. URL: https://tinyurl.com/356dwufh [accessed 2024-03-23]
5. Creanga AA, Syverson C, Seed K, Callaghan WM. Pregnancy-related mortality in the United States, 2011-2013. Obstet Gynecol 2017 Aug;130(2):366-373 [FREE Full text] [doi: 10.1097/AOG.0000000000002114] [Medline: 28697109]
6. Home page. World Health Organization. URL: https://www.who.int/ [accessed 2024-03-21]
7. Troiano NH, Witcher PM. Maternal mortality and morbidity in the United States: classification, causes, preventability, and critical care obstetric implications. J Perinat Neonatal Nurs 2018;32(3):222-231. [doi: 10.1097/JPN.0000000000000349] [Medline: 30036304]
8. Creanga AA, Berg CJ, Ko JY, Farr SL, Tong VT, Bruce FC, et al. Maternal mortality and morbidity in the United States: where are we now? J Womens Health (Larchmt) 2014 Jan;23(1):3-9 [FREE Full text] [doi: 10.1089/jwh.2013.4617] [Medline: 24383493]
9. Campbell-Grossman C, Brage Hudson D, Keating-Lefler R, Ofe Fleck M. Community leaders' perceptions of single, low-income mothers' needs and concerns for social support. J Community Health Nurs 2005 Dec;22(4):241-257. [doi: 10.1207/s15327655jchn2204_6] [Medline: 16245975]
10. New York State maternal mortality review report on pregnancy-associated deaths in 2018. New York State Department of Health. 2018. URL: https://www.health.ny.gov/community/adults/women/docs/maternal_mortality_review_2018.pdf [accessed 2024-03-23]
11. Suplee PD, Kleppel L, Santa-Donato A, Bingham D. Improving postpartum education about warning signs of maternal morbidity and mortality. Nurs Womens Health 2017 Dec;20(6):552-567. [doi: 10.1016/j.nwh.2016.10.009] [Medline: 27938796]
12. Howell EA, Bodnar-Deren S, Balbierz A, Parides M, Bickell N. An intervention to extend breastfeeding among black and Latina mothers after delivery. Am J Obstet Gynecol 2014 Mar;210(3):239.e1-239.e5 [FREE Full text] [doi: 10.1016/j.ajog.2013.11.028] [Medline: 24262719]

XSL•FO
RenderX

13. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am J Public Health 2015 Dec;105(12):e60-e76. [doi: 10.2105/ajph.2015.302903]
14. Stepanikova I, Mollborn S, Cook KS, Thom DH, Kramer RM. Patients' race, ethnicity, language, and trust in a physician. J Health Soc Behav 2006 Dec 24;47(4):390-405. [doi: 10.1177/002214650604700406] [Medline: 17240927]
15. Schwei RJ, Kadunc K, Nguyen AL, Jacobs EA. Impact of sociodemographic factors and previous interactions with the health care system on institutional trust in three racial/ethnic groups. Patient Educ Couns 2014 Sep;96(3):333-338 [FREE Full text] [doi: 10.1016/j.pec.2014.06.003] [Medline: 24973901]
16. Blair IV, Steiner JF, Fairclough DL, Hanratty R, Price DW, Hirsh HK, et al. Clinicians' implicit ethnic/racial bias and perceptions of care among Black and Latino patients. Ann Fam Med 2013;11(1):43-52 [FREE Full text] [doi: 10.1370/afm.1442] [Medline: 23319505]
17. Ayanian JZ, Zaslavsky AM, Guadagnoli E, Fuchs CS, Yost KJ, Creech CM, et al. Patients' perceptions of quality of care for colorectal cancer by race, ethnicity, and language. J Clin Oncol 2005 Sep 20;23(27):6576-6586. [doi: 10.1200/JCO.2005.06.102] [Medline: 16116149]
18. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull 2009 Nov;135(6):943-973 [FREE Full text] [doi: 10.1037/a0017327] [Medline: 19883143]
19. Language use in the United States: 2011. United States Census Bureau. 2011. URL: https://www.census.gov/library/publications/2013/acs/acs-22.html [accessed 2024-03-23]
20. Valdez RS, Brennan PF. Exploring patients' health information communication practices with social network members as a foundation for consumer health IT design. Int J Med Inform 2015 May;84(5):363-374. [doi: 10.1016/j.ijmedinf.2015.01.014] [Medline: 25704761]
21. Valdez RS, Gibbons MC, Siegel ER, Kukafka R, Brennan PF. Designing consumer health IT to enhance usability among different racial and ethnic groups within the United States. Health Technol 2012 Jul 13;2(4):225-233. [doi: 10.1007/s12553-012-0031-6]
22. Valdez RS, Holden RJ. Health care human factors/ergonomics fieldwork in home and community settings. Ergon Des 2016 Oct;24(4):4-9 [FREE Full text] [doi: 10.1177/1064804615622111] [Medline: 28781512]
23. Valdez RS, Holden RJ, Novak LL, Veinot TC. Transforming consumer health informatics through a patient work framework: connecting patients to context. J Am Med Inform Assoc 2015 Jan;22(1):2-10 [FREE Full text] [doi: 10.1136/amiajnl-2014-002826] [Medline: 25125685]
24. Bingham D, Suplee PD, Morris MH, McBride M. Healthcare strategies for reducing pregnancy-related morbidity and mortality in the postpartum period. J Perinat Neonatal Nurs 2018;32(3):241-249. [doi: 10.1097/JPN.0000000000000344] [Medline: 30036306]
25. Creanga AA, Berg CJ, Syverson C, Seed K, Bruce FC, Callaghan WM. Pregnancy-related mortality in the United States, 2006-2010. Obstet Gynecol 2015 Jan;125(1):5-12. [doi: 10.1097/AOG.0000000000000564] [Medline: 25560097]
26. Suplee PD, Bingham D, Kleppel L. Nurses' knowledge and teaching of possible postpartum complications. MCN Am J Matern Child Nurs 2017;42(6):338-344. [doi: 10.1097/NMC.0000000000000371] [Medline: 28816712]
27. O'Byrne LJ, Bodunde EO, Maher GM, Khashan AS, Greene RM, Browne JP, et al. Patient-reported outcome measures evaluating postpartum maternal health and well-being: a systematic review and evaluation of measurement properties. Am J Obstet Gynecol MFM 2022 Nov;4(6):100743 [FREE Full text] [doi: 10.1016/j.ajogmf.2022.100743] [Medline: 36087713]
28. Adler A, Conte TF, Illarraza T. Improvement of Postpartum Nursing Discharge Education Through Adaptation of AWHONN's Post-Birth Education Program. J Obstet Gynecol Neonatal Nurs 2019 Jun;48(3):S54. [doi: 10.1016/j.jogn.2019.04.093]
29. Urgent maternal warning signs. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hearher/maternal-warning-signs/index.html [accessed 2022-12-11]
30. Hannan J. APN telephone follow up to low-income first time mothers. J Clin Nurs 2013 Jan 30;22(1-2):262-270. [doi: 10.1111/j.1365-2702.2011.04065.x] [Medline: 22845337]
31. Dennis CL, Kingston D. A systematic review of telephone support for women during pregnancy and the early postpartum period. J Obstet Gynecol Neonatal Nurs 2008 May;37(3):301-314. [doi: 10.1111/j.1552-6909.2008.00235.x] [Medline: 18507601]
32. Letourneau N, Secco L, Colpitts J, Aldous S, Stewart M, Dennis CL. Quasi-experimental evaluation of a telephone-based peer support intervention for maternal depression. J Adv Nurs 2015 Jul 23;71(7):1587-1599. [doi: 10.1111/jan.12622] [Medline: 25705786]
33. Shamshiri Milani H, Azargashb E, Beyraghi N, Defaie S, Asbaghi T. Effect of telephone-based support on postpartum depression: a randomized controlled trial. Int J Fertil Steril 2015;9(2):247-253 [FREE Full text] [doi: 10.22074/ijfs.2015.4246] [Medline: 26246884]
34. Mobile fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2023-02-22]
35. Anderson M. Digital divide persists even as lower-income Americans make gains in tech adoption. Pew Research Center. URL: https://www.urbanismnext.org/resources/digital-divide-persists-even-as-lower-income-americans-make-gains-in-tech-adoption [accessed 2024-03-23]

XSL•FO
RenderX

36. Drexler K, Cheu L, Donelan E, Kominiarek M. 415: Remote self-monitoring of perinatal weight and perinatal outcomes in low-risk women. Am J Obstet Gynecol 2020 Jan;222(1):S272-S273. [doi: 10.1016/j.ajog.2019.11.431]
37. Marko KI, Ganju N, Krapf JM, Gaba ND, Brown JA, Benham JJ, et al. A mobile prenatal care app to reduce in-person visits: prospective controlled trial. JMIR Mhealth Uhealth 2019 May 01;7(5):e10520 [FREE Full text] [doi: 10.2196/10520] [Medline: 31042154]
38. Vernon MM, Yang FM. Implementing a self-monitoring application during pregnancy and postpartum for rural and underserved women: a qualitative needs assessment study. PLoS One 2022 Jul 19;17(7):e0270190 [FREE Full text] [doi: 10.1371/journal.pone.0270190] [Medline: 35853001]
39. Berry DL, Blumenstein BA, Halpenny B, Wolpin S, Fann JR, Austin-Seymour M, et al. Enhancing patient-provider communication with the electronic self-report assessment for cancer: a randomized trial. J Clin Oncol 2011 Mar 10;29(8):1029-1035 [FREE Full text] [doi: 10.1200/JCO.2010.30.3909] [Medline: 21282548]
40. Lordon RJ, Mikles SP, Kneale L, Evans HL, Munson SA, Backonja U, et al. How patient-generated health data and patient-reported outcomes affect patient-clinician relationships: a systematic review. Health Informatics J 2020 Dec 20;26(4):2689-2706 [FREE Full text] [doi: 10.1177/1460458220928184] [Medline: 32567460]
41. Todd BL, Feuerstein M, Gehrke A, Hydeman J, Beaupin L. Identifying the unmet needs of breast cancer patients post-primary treatment: the Cancer Survivor Profile (CSPro). J Cancer Surviv 2015 Jun 29;9(2):137-160. [doi: 10.1007/s11764-015-0428-0] [Medline: 25820913]
42. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
43. Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol 2016 Feb 20;34(6):557-565 [FREE Full text] [doi: 10.1200/JCO.2015.63.0830] [Medline: 26644527]
44. Denis F, Basch E, Septans A, Bennouna J, Urban T, Dueck AC, et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA 2019 Jan 22;321(3):306-307 [FREE Full text] [doi: 10.1001/jama.2018.18085] [Medline: 30667494]
45. Howell EA, Balbierz A, Wang J, Parides M, Zlotnick C, Leventhal H. Reducing postpartum depressive symptoms among black and Latina mothers: a randomized controlled trial. Obstet Gynecol 2012 May;119(5):942-949 [FREE Full text] [doi: 10.1097/AOG.0b013e318250ba48] [Medline: 22488220]
46. Ngai FW, Chan SW, Ip WY. The effects of a childbirth psychoeducation program on learned resourcefulness, maternal role competence and perinatal depression: a quasi-experiment. Int J Nurs Stud 2009 Oct;46(10):1298-1306. [doi: 10.1016/j.ijnurstu.2009.03.007] [Medline: 19361801]
47. Shorey S, Chan SW, Chong YS, He HG. A randomized controlled trial of the effectiveness of a postnatal psychoeducation programme on self-efficacy, social support and postnatal depression among primiparas. J Adv Nurs 2015 Jun 15;71(6):1260-1273. [doi: 10.1111/jan.12590] [Medline: 25496615]
48. Diefenbach MA, Leventhal H. The common-sense model of illness representation: theoretical and practical considerations. J Soc Distress Homeless 2016 Jul 07;5(1):11-38. [doi: 10.1007/bf02090456]
49. Wickens CD. Multiple resources and mental workload. Hum Factors 2008 Jun;50(3):449-455. [doi: 10.1518/001872008X288394] [Medline: 18689052]
50. Gadamer HG. Philosophical Hermeneutics. Oakland, CA: University of California Press; 1976.
51. Saldana J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: Sage Publications; 2012.
52. Huberman AM, Miles M, Saldana J. Qualitative Data Analysis: A Methods Sourcebook. Thousand Oaks, CA: Sage Publications; 2014.
53. Pope C, Ziebland S, Mays N. Analysing qualitative data. In: Pope C, Mays N, editors. Qualitative Research in Health Care. Hoboken, NJ: John Wiley & Sons; 2006:63-81.
54. Fusch PI, Ness LR. Are we there yet? Data saturation in qualitative research. Qual Rep 2015 Sep 8;20(9):1408-1416 [FREE Full text] [doi: 10.46743/2160-3715/2015.2281]
55. Morse JM. The significance of saturation. Qual Health Res 2016 Jul 01;5(2):147-149. [doi: 10.1177/104973239500500201]
56. Sultan P, Sharawi N, Blake L, Ando K, Sultan E, Aghaeepour N, et al. Use of patient-reported outcome measures to assess outpatient postpartum recovery: a systematic review. JAMA Netw Open 2021 May 03;4(5):e2111600 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11600] [Medline: 34042993]
57. O'Carroll J, Ando K, Yun R, Panelli D, Nicklin A, Kennedy N, et al. A systematic review of patient-reported outcome measures used in maternal postpartum anxiety. Am J Obstet Gynecol MFM 2023 Sep;5(9):101076. [doi: 10.1016/j.ajogmf.2023.101076] [Medline: 37402438]
58. Cox J. Thirty years with the Edinburgh postnatal depression scale: voices from the past and recommendations for the future. Br J Psychiatry 2019 Mar 18;214(3):127-129. [doi: 10.1192/bjp.2018.245] [Medline: 30774059]
59. Novick AM, Kwitowski M, Dempsey J, Cooke DL, Dempsey AG. Technology-based approaches for supporting perinatal mental health. Curr Psychiatry Rep 2022 Sep 23;24(9):419-429 [FREE Full text] [doi: 10.1007/s11920-022-01349-w] [Medline: 35870062]

XSL•FO
RenderX

60. Guy Jr GP, Adams EK, Redd SK, Dunlop AL. Effects of Georgia's Medicaid family planning waiver on pregnancy characteristics and birth outcomes. Womens Health Issues 2023 Dec 15. [doi: 10.1016/j.whi.2023.11.004] [Medline: 38103999]
61. Zimmermann K, Haen LS, Desloge A, Handler A. The role of a local health department in advancing health equity: universal postpartum home visiting in a large urban setting. Health Equity 2023 Oct 01;7(1):703-712 [FREE Full text] [doi: 10.1089/heq.2023.0104] [Medline: 37908403]
62. Castillo AF, Davis AL, Krishnamurti T. Adapting a pregnancy app to address disparities in healthcare access among an emerging Latino community: qualitative study using implementation science frameworks. Research Square Preprint posted online April 27, 2022 [FREE Full text] [doi: 10.21203/rs.3.rs-1555131/v1]
63. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: Bederson BB, Shneiderman B, editors. The Craft of Information Visualization: Readings and Reflections. Burlington, MA: Morgan Kaufmann; 2003:364-371.
64. White KM, Williamson C, Bergou N, Oetzmann C, de Angel V, Matcham F, et al. A systematic review of engagement reporting in remote measurement studies for health symptom tracking. NPJ Digit Med 2022 Jun 29;5(1):82 [FREE Full text] [doi: 10.1038/s41746-022-00624-7] [Medline: 35768544]
65. Walsh S, Golden E, Priebe S. Systematic review of patients' participation in and experiences of technology-based monitoring of mental health symptoms in the community. BMJ Open 2016 Jun 21;6(6):e008362 [FREE Full text] [doi: 10.1136/bmjopen-2015-008362] [Medline: 27329437]
66. Davidson KW, Mangione CM, Barry MJ, Cabana MD, Caughey AB, Davis EM, et al. Actions to transform US preventive services task force methods to mitigate systemic racism in clinical preventive services. JAMA 2021 Dec 21;326(23):2405-2411. [doi: 10.1001/jama.2021.17594] [Medline: 34747970]
67. Im EO, Chee W, Hu Y, Kim S, Choi H, Hamajima Y, et al. What to consider in a culturally tailored technology-based intervention? Comput Inform Nurs 2018 Sep;36(9):424-429 [FREE Full text] [doi: 10.1097/CIN.0000000000000450] [Medline: 29927767]
68. Yen RW, Smith J, Engel J, Muscat DM, Smith SK, Mancini J, et al. A systematic review and meta-analysis of patient decision aids for socially disadvantaged populations: update from the international patient decision aid standards (IDPAS). Med Decis Making 2021 Jun 21;41(7):870-896. [doi: 10.1177/0272989x211020317]
69. Nielsen J. 10 usability heuristics for user interface design. Nielsen Norman Group. 1994. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2020-11-15]
70. Ancker JS, Stabile C, Carter J, Chen LY, Stein D, Stetson PD, et al. Informing, reassuring, or alarming? Balancing patient needs in the development of a postsurgical symptom reporting system in cancer. AMIA Annu Symp Proc 2018;2018:166-174 [FREE Full text] [Medline: 30815054]

Abbreviations

CDC: Centers for Disease Control and Prevention
EPDS: Edinburgh Postnatal Depression Scale
mHealth: mobile health
PRD: pregnancy-related death
PRO: patient-reported outcome
REDCap: Research Electronic Data Capture
UWS: Urgent Maternal Warning Signs

Edited by A Mavragani; submitted 22.03.23; peer-reviewed by C Laranjeira; comments to author 15.01.24; revised version received 20.02.24; accepted 08.03.24; published 26.04.24.

Please cite as:
Benda N, Woode S, Niño de Rivera S, Kalish RB, Riley LE, Hermann A, Masterson Creber R, Costa Pimentel E, Ancker JS
Understanding Symptom Self-Monitoring Needs Among Postpartum Black Patients: Qualitative Interview Study
J Med Internet Res 2024;26:e47484
URL: https://www.jmir.org/2024/1/e47484
doi:10.2196/47484
PMID:38669066

©Natalie Benda, Sydney Woode, Stephanie Niño de Rivera, Robin B Kalish, Laura E Riley, Alison Hermann, Ruth Masterson Creber, Eric Costa Pimentel, Jessica S Ancker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Short-Term Effects of an eHealth Care Experiential Learning Program Among Patients With Type 2 Diabetes: Randomized Controlled Trial

Yu-Shan Cheng[1], MSN; Cheng-Pei Lin[2,3], PhD; Lu-Yen Anny Chen[4], PhD; Wei-Ren Hwang[5], MHA, MD; Yi-Chun Lin[5,6], MD; Yu-Chi Chen[4], PhD

[1]School of Nursing, The University of Texas at Austin, Austin, TX, United States
[2]Institute of Community Health Care, College of Nursing, National Yang Ming Chiao Tung University, Taipei, Taiwan
[3]Cicely Saunders Institute of Palliative Care, Policy and Rehabilitation, King's College London, London, United Kingdom
[4]Institute of Clinical Nursing, College of Nursing, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Rong-Yang Clinic, Taipei, Taiwan
[6]Division of Endocrinology and Metabolism, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

Corresponding Author:
Yu-Chi Chen, PhD
Institute of Clinical Nursing
College of Nursing
National Yang Ming Chiao Tung University
No155, Section 2
Li-Nong Street, Beitou District
Taipei, 112
Taiwan
Phone: 886 228267093
Email: ycchen2@nycu.edu.tw

Abstract

Background: Type 2 diabetes is a chronic disease with a significant medical burden. eHealth care integrates medicine and technology to enhance the outcomes of such patients; however, adequate eHealth literacy (eHL) is necessary for that to happen. Fostering eHL is crucial for patients with diabetes to engage with eHealth care and receive quality care and timely support. Experiential learning theory can enhance patients' eHL and skills to use eHealth care technology in their daily care.

Objective: This study explored the effectiveness of an eHealth care experiential learning program in improving eHL, patient health engagement, and eHealth care use status among patients with type 2 diabetes in 3 months.

Methods: In this randomized controlled trial, patients under case management services from various clinics in Taiwan were randomly assigned to either the intervention group receiving the 6-session eHealth care experiential learning program or the control group receiving the usual care. Data were collected using structured questionnaires at 3 time points: pretest, postintervention, and 3 months after the intervention. Descriptive data were presented using frequency distribution, percentage, mean, and SD. The outcomes were analyzed using a generalized estimating equation method by intention-to-treat analysis.

Results: A total of 92 participants (46 in each group) were recruited in this study. Of these, 86 completed the course and follow-up evaluations with a mean age of 62.38 (SD 12.91) years. After completing the intervention, the intervention group had significantly higher posttest scores in eHL (β=19.94, SE 3.52; $P<.001$), patient health engagement (β=.28, SE 0.13; $P=.04$), and eHealth use (β=3.96, SE 0.42; $P<.001$) than the control group. Furthermore, the intervention group maintained these significant improvements in eHL (β=18.19, SE 3.82; $P<.001$) and eHealth use (β=3.87, SE 0.49; $P<.001$) after 3 months.

Conclusions: Participating in the eHealth care experiential learning program resulted in significant improvements in eHL, patient health engagement, and eHealth use among patients with type 2 diabetes. Our interventional program can inform future clinical practice and policies to strengthen self-management skills and facilitate the use of health technology in caring for patients with chronic diseases.

Trial Registration: ClinicalTrials.gov NCT05180604; https://clinicaltrials.gov/ct2/show/NCT05180604

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53509) doi:10.2196/53509

KEYWORDS

diabetes; eHealth literacy; eHealth; patient engagement; experiential learning theory; experimental learning theory; type 2 diabetes; randomized controlled trial

Introduction

Diabetes is a chronic disease, and its prevalence is rapidly increasing globally. It is currently ranked among the top 10 causes of death worldwide. According to the World Health Organization, over 460 million people have diabetes, with this number expected to reach 700 million by 2045 [1,2]. Diabetes is often accompanied by several complications, leading to soaring medical expenses that impose a heavy burden on patients and their families. In the United States, the cost of direct and indirect care for diabetes is US $327 billion. Moreover, from 2012 to 2017, the economic costs of diabetes increased by 26% in the United States [3]. Therefore, preventing the progression of diabetes is a crucial public health issue.

Diabetes carries significant risks, including pathological changes, irreversible complications, and even death. To effectively manage diabetes, maintain a stable condition, and mitigate these risks, attentive, self-managed daily care by patients is necessary [4]. Patients need to engage in various self-management tasks such as monitoring their blood sugar regularly, managing medication, exercising regularly, eating a healthy diet, and monitoring their condition periodically [5,6]. These interventions can slow disease progression, prevent complications, reduce medical expenses, and enable patients to coexist with the disease [7].

With the rapid development of eHealth technology, using eHealth care systems for chronic disease management has become a growing trend. This innovative strategy allows individuals to efficiently engage in disease self-management in their daily lives and meet continuous care demands. Available applications include those offering physiological monitoring, recordkeeping for self-management at home, health consultation guidance, location and access to emergency services, and communication through social media [8,9]. However, while eHealth applications offer multiple health care solutions, they also experience potential problems, especially if patients are unfamiliar with mobile health or do not know how to use eHealth care systems. In that case, they may not be able to access the service, be apprehensive, or refuse to use such systems altogether. This can result in decreased willingness and motivation to use eHealth care for self-management, leading to deteriorating health outcomes [10,11]. Therefore, improving the engagement of patients with diabetes in the eHealth care system is crucial for continuing to provide them with high-quality care.

eHealth literacy (eHL) refers to an individual's ability to seek, find, understand, and evaluate health information obtained from electronic sources such as websites, mobile apps, and other digital platforms [12,13]. It involves the skills, knowledge, and capabilities needed to navigate, comprehend, and use digital health resources effectively to make informed decisions on health-related matters. Patients with insufficient eHL find it difficult to use eHealth care applications and respond to their requirements such as completing web-based registration and electronic card check-in, which can result in increased passive resistance to self-management. Conversely, those with higher eHL levels can navigate and operate eHealth care effectively [14-16]. Moreover, enhancing patients' engagement in self-management of their health is key to promoting their further use of smart health care systems in the eHealth era. Such patient health engagement is a continuous, dynamic process through which patients' cognition, emotions, and behaviors drive psychological change across the following 4 stages: blackout, arousal, adhesion, and eudaimonic project [17,18]. As a result, patient health engagement affects the health outcomes of chronic disease management.

Introducing smart health care applications could be a beneficial strategy for patients; however, if they cannot engage in smart health care or face serious challenges while using eHealth in their daily care, they may feel reluctant, or even refuse, to use eHealth. Thus, teaching patients to use eHealth care systems and improving their eHL based on their pace may help them engage in eHealth care. In addition, encouraging patients to engage in practical work, discuss with each other, and reflect on their health issues during the learning process can increase their knowledge, attitudes, and behaviors [19,20]. Experiential learning theory (ELT) emphasizes the role of a patient's experience in developing the skills they need to function in their daily life. This theory has been widely used in education and shown to improve learners' overall knowledge, attitudes, and behaviors [21-26]. In the eHealth era, it is crucial to improve patients' eHL to enable them to engage in self-management behaviors. Patients with higher levels of eHL are better able to effectively implement eHealth systems in their daily self-management [27,28], corresponding with the expected achievement indicators of this research. Therefore, we propose that adapting ELT as an intervention for patients' eHealth care education would enhance their eHL, allow them to engage in smart technology health care, and increase their use of health technology, thus enabling eHealth care to become a part of the lives of patients with diabetes. Ultimately, not only does it enhance self-management behavior, but it can also achieve disease control in the long run, preventing deterioration and complications.

Thus, this study aimed to explore the effectiveness of an eHealth care experiential learning program in improving eHL, patient health engagement, and eHealth care use status among patients with type 2 diabetes.

XSL•FO
RenderX

Methods

Overview

This randomized controlled trial reports on the results of a 3-month study that included 92 patients with type 2 diabetes that examined the short-term effects on eHL, patient health engagement, and eHealth care use status. The experimental group received an eHealth care experiential learning program, while the control group received the usual standard care. CONSORT (Consolidated Standards of Reporting Trials) guidelines were used to report the findings (Multimedia Appendix 1).

Study Design

This was a single-blind, randomized controlled trial, in which the research team performed random allocation by drawing slots. The study process flowchart is presented in Figure 1.

Figure 1. Study process flowchart.

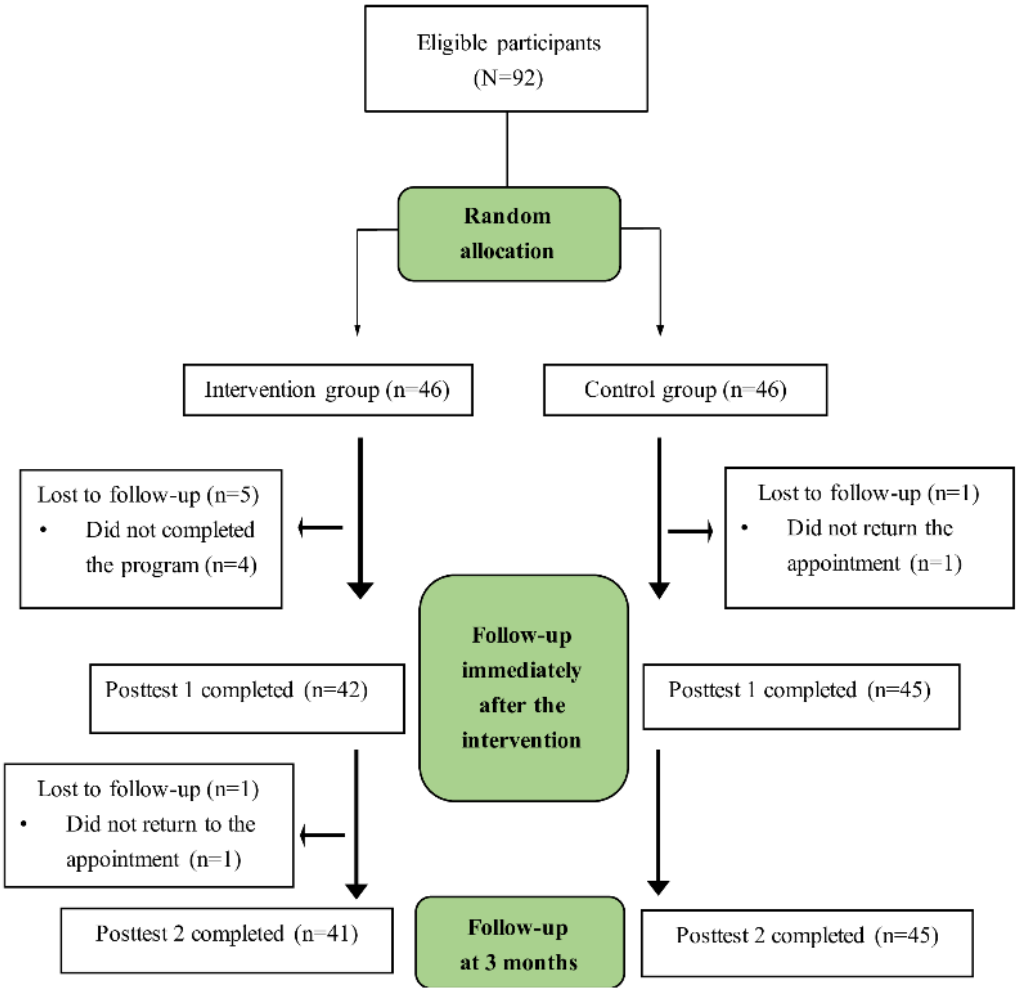

Participants

The study included participants who met the following criteria: (1) diagnosis of type 2 diabetes and receiving case management services for at least 3 months, (2) aged 20 years or older, (3) able to communicate in Mandarin or Taiwanese, and (4) possess a mobile phone or tablet with an internet connection. Patients with serious diseases such as general paralysis, mental disorders, and cognitive function abnormalities were excluded. Enrollment took place from July to September 2020, with the intervention period from September to December 2020, and the posttest completed in December 2020 and March 2021.

To detect a clinically significant effect, defined by the within or between-group interaction and follow-up time, on the outcome indicators, a sample size of 82 was calculated using repeated measures ANOVA between the factors based on an effect size of 0.25, as per Cohen guidelines [29]. This used an intercluster correlation coefficient of 0.5, a power of 0.8, and an α of .05. To account for a potential percentage dropout rate of 20%, the total sample size was increased to 92 participants. Ultimately, the study recruited 92 patients with type 2 diabetes.

Procedure

We recruited patients with type 2 diabetes from various metabolic clinics throughout northern Taiwan. These clinics are integrated into Taiwan's nationwide health insurance system to ensure all citizens have access to medical services. They are also part of the Diabetes Shared Care Program in Taiwan, which provides specialized care for individuals with metabolic disorders. In these clinics, patients receive comprehensive care from a multidisciplinary team comprising metabolic doctors, nurses, case managers, dieticians, and pharmacists. The services provided include the essential components of effective diabetes

XSL•FO
RenderX

management such as regular laboratory tests, health examinations, and health education programs.

Before starting the study, our research team collaborated with the clinic management to establish a clear and mutually agreed-upon research process. To recruit participants, the clinics displayed posters providing information about the study. Research assistants were available to thoroughly explain the research procedures to interested individuals and obtain informed consent. Participants were asked to complete a pretest only after completing the explanation process and obtaining informed consent. This ensured a thorough and ethical approach to participant recruitment and data collection.

Next, after the preassessment, those willing participants were added with a case number. The research team used a computer program to draw slots, assigning the cases to the intervention or control group using a lottery system. The intervention group participated in a 3-month, 6-session eHealth care experiential learning program, while the control group received the usual care and an eHealth care manual. In Taiwan, the usual care refers to the standard treatment and follow-up in outpatient units. Typically, physicians schedule appointments every 3 months for patients with type 2 diabetes to conduct blood tests and prescribe medications. This care often includes health consultations by case managers or nutritional counseling. Both groups underwent posttesting twice: immediately after the 6-session course and 3 months later. Patients with chronic diseases usually have regular medical appointments every 3 months in Taiwan, as mandated by the national health insurance. As such, the control group completed their regular medical appointments during the same period.

Intervention

The 6-session eHealth care experiential learning program (Multimedia Appendix 2), which focused on self-management of diabetes, was based on the eHL framework (eHLF) [30,31] and ELT cycle [20]. Over 3 years, our research team developed a program specifically designed for patients with type 2 diabetes and the challenges health care providers encounter in managing this condition. In the initial year, we began a qualitative study to gain in-depth insights into the needs and challenges associated with diabetes care. This study revealed several critical aspects for enhancing self-management among patients with diabetes. One of the pivotal insights from the study was the vital role of eHealth care for enhancing home care for patients, particularly those with chronic diseases such as diabetes, for which consistent health monitoring at home is essential. Patients' 3-monthly clinical visits further highlighted the need for effective home-based health monitoring and guidance. These insights underscored their struggle to engage in meaningful health care discussions, often due to their unfamiliarity with eHealth technologies and a lack of knowledge on self-management practices.

In the program's second year, we initially developed 8 sessions to address these identified gaps. However, after conducting a Delphi study and a pilot study to test the program's feasibility and usability, we streamlined it to 6 sessions. This refinement process intended to provide a more effective and concentrated participant learning experience focused on the program content. The program's overall development cost, including expenses for wearable devices and Bluetooth-enabled health monitors (eg, blood pressure monitors, glucose monitors, and scales), amounted to approximately US $66.

Therefore, the program followed the learning cycle, with the following activities and strategies, divided into 4 stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. First, the concrete experience stage introduced the importance of health care management and eHealth care for patients with diabetes. The program provided real-life examples and scenarios to help patients understand the significance of eHealth care in managing their condition.

Second, various methods were used in the reflective observation stage to encourage patients to reflect on their learning experiences. Videos were played to show the practical implementation of eHealth care. Group discussions were held to facilitate dialogue and share new insights among patients participating in the eHealth care program. Additionally, each session began with a review to help increase patients' familiarity with eHealth care devices. This approach ensured a blend of visual learning and active participation, which enhanced the program's overall effectiveness.

Third, the abstract conceptualization stage aimed to clarify the concepts of eHealth care using social media platforms such as LINE and Facebook and internet groups. Patients engaged in web-based discussions to deepen their understanding of eHealth care practices. Moreover, group discussions fostered dialogue among participants around their eHealth care practice experiences.

Finally, the active experimentation stage allowed patients to learn by actively engaging in related health situations. In this stage, they had the opportunity to choose suitable health technology software, hardware, and wearable devices for their specific diseases in a simulation environment.

Then, patients participated in group and individual competitions that involved using eHealth care tools effectively. They were assisted in setting practice and self-management goals and encouraged to use eHealth activities at home. Overall, this experiential learning program followed a systematic learning cycle that involved introducing concepts, facilitating reflection, deepening understanding, and encouraging active application and experimentation with eHealth care tools and techniques.

Our eHealth care experiential learning program comprised 6 biweekly sessions over 3 months, each lasting around 90 minutes. These sessions, conducted by the project investigator (YCC), combined theoretical knowledge with extensive hands-on training and at-home practice using various eHealth tools and applications commonly used in Taiwanese clinical settings. This included wearable devices such as Xiaomi, health care applications such as Health2Sync, LINE, My Health Bank, and blood pressure and glucose monitors. These hardware and software tools are designed to track various health records and statuses such as heart rates, sleep patterns, blood sugar levels, blood pressure, diet, and medication adherence. With Bluetooth connectivity, they seamlessly integrate into everyday health

XSL•FO
RenderX

management and have been widely adopted in clinical settings due to their effectiveness and ease of use.

The six sessions were as follows:

1. "A new realm of health technology care": Introduced smart health care functions, eHealth applications, and motivated patients to engage with eHealth technologies while reflecting on their health care experiences.
2. "Fighting against sugar": Emphasized the importance of regular health monitoring (blood sugar, blood pressure, and diet) in diabetes care, focusing on self-management, demonstrating diabetes health care technology, and providing patients with practice tools.
3. "Trick of the trade for chronic kidney disease prevention": Focused on diabetes complications, particularly chronic kidney disease, discussing self-management and relevant health care technologies that could be used for diabetic nephropathy.
4. "eHealth care by your side": Used competitions and hands-on activities to foster patient engagement with health care technology. Patients were encouraged to operate the technology independently and discuss any challenges they faced. This approach not only motivated active participation but also facilitated learning from real-life experiences, enabling patients to better manage their health conditions using these technologies.
5. "My smart eHealth in daily care": Focused on how health care technology can enhance patients' lives. The participants were taught to correctly operate various eHealth components and integrate smart health care into their daily self-management. This approach emphasized practical use, encouraging patients to apply smart health care technologies to manage their health and benefit from their functionality in everyday life.
6. "My eHealth care practice journey": Synthesized the key elements from the previous sessions. This final session focused on ensuring that patients could accurately operate smart health care applications, reinforcing their ability and willingness to engage in self-care management using health care technology. This comprehensive review and hands-on practice session aimed to solidify patients' understanding and proficiency in applying these technologies for their ongoing health management.

Overall, these sessions aimed to provide patients with the knowledge and skills to effectively use eHealth tools for managing their health, enhancing their outcomes through technology-driven self-management. During each session, the project investigator (YCC) began by encouraging participants to ask questions and provide feedback on the previous session for 15 minutes. Following this, the project investigator introduced the current session's content, and research assistants assisted participants in the practical operation of health care technology and tools for the remaining 15-20 minutes. After the session, participants could practice one-on-one and discuss their learning challenges with the research team. Additionally, between the 2 sessions, the research team used a social media platform to coach participants and provide feedback on any questions. Throughout the sessions, various teaching aids and equipment were used, including learning manuals, social media platforms, health applications, and wearable health devices. The participants in each session simultaneously experienced 4 ELT learning cycles. This approach enabled them to experience intelligent care and learn by practicing daily self-management using eHealth and engaging in the smart health care system.

Data Collection

The data collection was blind to grouping. A single-blind method was applied. The clinics' case managers collected the posttest data. The participants filled out the questionnaires independently during their medical appointments every 3 months. The case managers assisted participants if they required help with completing these questionnaires. All participants completed the questionnaires during admission (T0). The participants in the intervention group completed the posttest 1 (T1) questionnaires after completing the 3-month intervention program (T1) and again after a 3-month follow-up (T2) during their regular medical appointments. The participants in the control group completed the 2 posttest questionnaires during their regular medical appointments scheduled at 3 monthly intervals, 3 (T1) and 6 months (T2) after enrolling in this study.

Measurement

This study used questionnaires to collect data, which included sociodemographic data, the eHealth Literacy Questionnaire (eHLQ), the Patient Health Engagement Scale, and the eHealth Care Use Scale (Multimedia Appendix 3 [13,32-34]).

Sociodemographic Data

Sociodemographic data included age, sex, education, economic status, number of comorbidities, perceived severity of illness, and health status.

eHLQ Instrument

The 35-item eHLQ is based on the eHLF, which refers to the ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to address or solve a health problem. The eHLQ is typically structured to cover various dimensions of eHL, including technical skills, ability to understand health information, critical evaluation of web-based sources, and ability to actively engage with digital health technologies [13]. It comprises the following seven subscales: (1) using technology to process health information, (2) understanding the concept and language of health, (3) actively participating in technology service abilities, (4) feeling safe and self-controllable about personal health information, (5) motivation to participate in information technology services, (6) obtaining useful information technology services, and (7) information technology services that meet personal needs. Each item was scored on a 4-point scale (1=strongly disagree, 2=disagree, 3=agree, and 4=strongly agree). The higher the score, the better the patients' eHL. The total score ranged from 35 to 140. The original scale's reliability was 0.8, and it had good construct and discriminant validity [13]. The Chinese version of the eHLQ has been found to have a content validity of 0.97, and the Cronbach α for the entire scale was 0.98, with subscales ranging from 0.74 to 0.97 [32].

XSL•FO
RenderX

Patient Health Engagement Scale

Graffigna and Barello [33,35] developed the Patient Health Engagement Scale using the patient health engagement model. The scale investigates if patients are constantly engaging in their health [33,35]. It includes the following four positions along a continuum of engagement: (1) blackout, (2) arousal, (3) adhesion, and (4) eudaimonic project. The scale can conceptualize the psychological aspects of patient health engagement and includes 5 ordinal items. To avoid social desirability bias, 4 positions of raw patient health engagement were categorized into a 7-point scale: 1 and 2=blackout, 3 and 4=arousal, 5 and 6=adhesion, and 7=eudaimonic project. To acquire the final engagement position, the 4 positions of patient health engagement were arranged from lowest to highest. The median score corresponds to the third position, which represents the patient health engagement position. These 4 positions arise from conjoint cognitive (thinking), emotional (feeling), and conative (acting) enactment. The Cronbach α for this scale was 0.85, and its retest reliability was 0.95, indicating good reliability and validity [33]. The Chinese version of the scale translated by Zhang et al [34] had an internal consistency of 0.89 and a retest reliability between 0.52 and 0.79.

eHealth Care Use Scale

We used a self-developed scale to investigate various types of eHealth care use and monitoring items in daily disease management. The scale had high internal consistency (Cronbach α=0.92) and good content validity (content validity=0.95). The 4 types of eHealth care included were computer or internet systems, mobile apps, health monitoring systems, and wearable devices (eg, pedometers, smart bracelets, heart rate monitors, blood pressure monitors, blood glucose meters, and weight scales). The scoring method was as follows: 0=no use of any eHealth types, 1=use of 1 type, 2=use of 2 types, 3= use of 3 types, and 4=use of 4 types. Higher scores indicated a greater variety of eHealth types used.

Furthermore, the monitoring items encompassed blood pressure, blood sugar, weight, diet, sleep, heart rate, steps, and other health data. The scoring method for the monitoring items was as follows: 0=not monitoring any health data, 1=monitoring 1 type, 2=monitoring 2 types, 3=monitoring 3 types, 4=monitoring 4 types, 5=monitoring 5 types, 6=monitoring 6 types, and 7=monitoring 7 types. A higher score indicated a greater number of health data items being monitored. The total score reflected overall eHealth care use with a higher score indicating more comprehensive use.

Statistical Analysis

The data were analyzed using SPSS (version 24.0; IBM Corp). Descriptive statistics were used to report continuous data using scores, means, and SDs. Categorical data were presented as numbers and percentages. Independent sample *t* tests (2-tailed) and chi-square tests (2-tailed) were used to compare the homogeneity of the pretest data between the 2 groups. To avoid overestimating the effect of the intervention while not ruling out loss to follow-up, intention-to-treat analysis was used.

The generalized estimating equation (GEE) was used to examine the difference in effectiveness, with statistical significance set at $P<.05$. This statistical analysis method can accommodate repeated responses from each participant. The autoregressive (1) model of the GEE using a linear regression model was used to estimate whether there was a significant difference in the improvement of the outcome indicators between the 2 groups over time. To reduce interference factors, sex and economic status—both of which showed significant differences in the initial test after randomization—were included as covariates in the corrected GEE model. This adjustment addresses the potential impact of sex and economic status imbalances and the interaction between time and group. Consequently, the reliability of our findings is enhanced despite the imbalance issue following randomization.

Ethical Considerations

This study was approved by the institutional review board of the university with which the research team is affiliated (YM106120E-3). Before data collection, we provided the participants with an explanation of the research objectives, process, and questionnaire content, both verbally and in writing. Once participants completed a questionnaire at each time point—pretest, postintervention, and 3 months after the intervention—they were able to receive an NT $100 voucher (a currency exchange rate of NT $1=US $0.33 is applicable). We also assured them that their participation would be confidential and would affect neither their treatment rights nor their health care. Data were collected after the participants had signed the consent form. The participants were free to withdraw from the study at any point, even after providing consent.

Results

Participant Characteristics

A total of 86 participants completed the 2 posttests, including 41 (48%) in the intervention group and 45 (52%) in the control group. The total participant attrition rate was 7% (n=6). The rates of subject loss in the intervention and control groups were 11% (5/46) and 2% (1/46), respectively. The initial 92 participants were included in the final analysis (Table 1).

XSL•FO
RenderX

Table 1. Baseline comparison of the patients.

Variable	Total (N=92)	Intervention group (n=46)	Control group (n=46)	*t* test (*df*) or chi-square test (*df*)
Age (years), mean (SD)	62.38 (13)	64.96 (10)	59.80 (15)	1.9 (90)[a]
Sex, n (%)				8.0 (1)[b,c]
Male	33 (36)	10 (22)	23 (50)	
Female	59 (64)	36 (78)	23 (50)	
Education, n (%)				6.1 (4)[b]
Elementary school and below	18 (20)	6 (13)	12 (26)	
Junior high school	16 (17)	11 (24)	5 (11)	
High or vocational school	32 (35)	14 (30)	18 (39)	
College or university	21 (23)	13 (28)	8 (17)	
Graduate school	5 (5)	2 (4)	3 (7)	
Economic status (NT $[d]), n (%)				9.7 (4)[b,c]
<10,000	29 (32)	16 (35)	13 (28)	
10,001-20,000	24 (26)	6 (13)	18 (39)	
>20,001-30,000	19 (21)	13 (28)	6 (13)	
>30,001-40,000	15 (16)	9 (20)	6 (13)	
>40,001	5 (5)	2 (4)	3 (7)	
Number of comorbidities, mean (SD)	1.98 (1)	1.83 (1)	2.13 (1)	–1.4 (90)[a]
1 item, n (%)	38 (41)	20 (44)	18 (39)	
≧2 items, n (%)	54 (59)	26 (57)	28 (61)	
Perceived severity of illness, n (%)				0.8 (2)[b]
Not serious	28 (30)	12 (26)	16 (35)	
Normal	62 (67)	33 (72)	29 (63)	
Serious	2 (2)	1 (2)	1 (2)	
Health status, n (%)				4.3 (2)[b]
Bad	18 (20)	7 (15)	11 (24)	
Normal	54 (59)	26 (57)	28 (61)	
Good	20 (22)	13 (28)	7 (15)	
eHealth literacy, mean (SD)	93.82 (19)	90.96 (19)	96.67 (18)	–1.5 (90)[a]
Patient health engagement, mean (SD)	2.92 (1)	2.91 (1)	2.96 (1)	–0.5 (90)[a]
eHealth care use, mean (SD)	2.27 (2)	2.17 (2)	2.37 (2)	–0.4 (90)[a]

[a]Independent sample *t* test (2-tailed).

[b]Chi-square test (2-tailed).

[c]The difference between the 2 groups at a significance level of .01 (2-tailed).

[d]A currency exchange rate of NT $1=US $0.33 is applicable.

The mean age of the 92 participants was 62.38 (SD 12.9) years, and 59 (64%) were female participants, while 33 (36%) were male participants. Regarding education level, the highest proportion had a high school or higher vocational education, accounting for 32 (35%) participants. The intervention group consisted mostly of female participants, whereas the control group had an equal distribution of male and female participants. There was a significant difference in sex between the 2 groups (χ^2_1=8.0, *P*=.005). Regarding economic status, the intervention group had more participants with an income of NT $20,000-30,000, while the control group had more participants earning NT $10,000-20,000. There was a significant difference in economic status between the 2 groups (χ^2_4=9.7, *P*=.046). In terms of comorbidities, 54 (59%) individuals had at least 2 types

XSL•FO
RenderX

of chronic diseases. In contrast, 38 (41%) individuals had 1 type. Moreover, 62 (67%) participants indicated a general level of subjective disease severity. Similarly, regarding self-rated health status, 54 (59%) participants considered their health status general (Table 1).

There were no significant differences in most characteristics between the 2 groups, except for sex and economic status. Furthermore, the baseline assessment of the outcome indicators, including eHL, patient health engagement, and eHealth care use, showed no significant differences.

Effectiveness of the eHealth Care Experiential Learning Program

The results revealed significant differences in the patients' eHL, patient health engagement, and eHealth use after participating in the eHealth care experiential learning program (Table 2, Multimedia Appendix 4, and Figure 2). Moreover, there were also significant differences in the eHL subscales (Multimedia Appendix 5).

Figure 2. Effects of the eHealth care experiential learning program on eHealth literacy, patient health engagement, and eHealth care use among patients with type 2 diabetes. eHLQ: eHealth Literacy Questionnaire.

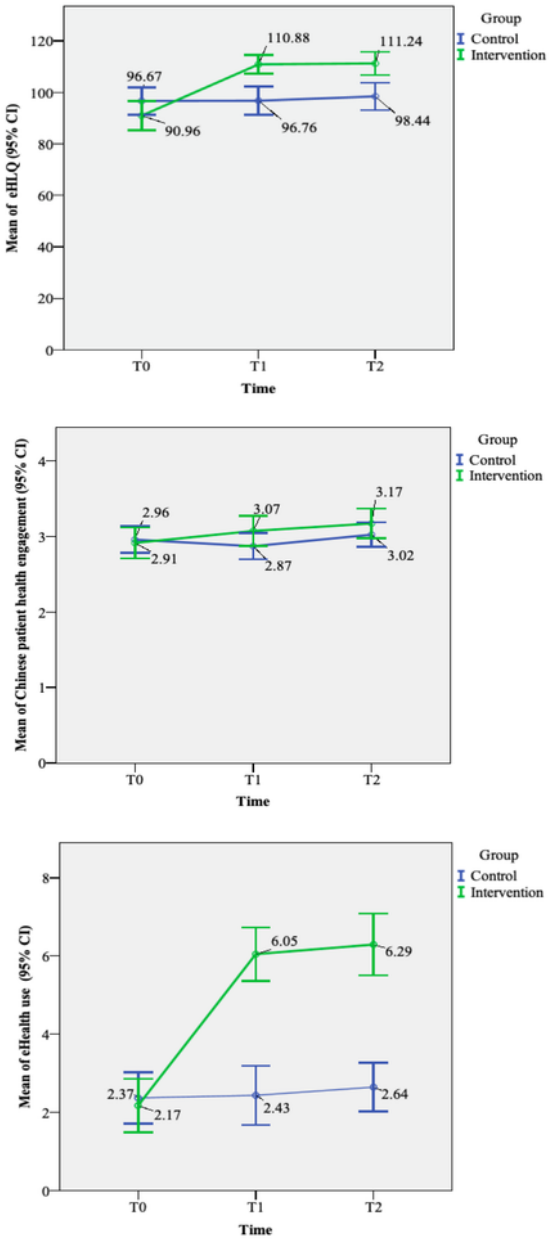

XSL•FO
RenderX

Table 2. Mean changes in eHealth literacy, patient engagement, and eHealth care use over time (N=92).

	β (SE)	*P* value
eHealth literacy		
Intercept	91.12 (4.06)	<.001[a]
Group (intervention vs control)	–6.11 (3.58)	.09
Time		
T2 vs T0	2.11 (2.03)	.30
T1 vs T0	.09 (1.39)	.95
Sex		
Female vs male	–1.19 (3.08)	.70
Economic status (NT $[b])		
>40,001 versus <10,000	13.95 (4.63)	.003
>30,001-40,000 versus <10,000	8.41 (4.08)	.04
>20,001-30,000 versus <10,000	13.55 (3.78)	<.001[a]
10,001-20,000 versus <10,000	6.08 (4.21)	.15
Group*time		
Intervention*T2 versus control*T2	18.19 (3.82)	<.001[a]
Intervention*T1 versus control*T1	19.94 (3.52)	<.001[a]
Patient health engagement		
Intercept	3.05 (0.11)	<.001[a]
Group (intervention vs control)	–.07 (0.14)	.59
Time		
T2 versus T0	.05 (0.08)	.52
T1 versus T0	–.09 (0.08)	.28
Sex		
Female versus male	–.12 (0.12)	.32
Economic status (NT $[b])		
>40,001 versus <10,000	–.43 (0.28)	.12
>30,001-40,000 versus <10,000	–.06 (0.17)	.73
>20,001-30,000 versus <10,000	.15 (0.13)	.25
10,001-20,000 versus <10,000	–.05 (0.13)	.70
Group×time		
Intervention×T2 versus control×T2	.24 (0.14)	.07
Intervention×T1 versus control×T1	.28 (0.13)	.04[c]
eHealth care use		
Intercept	1.54 (0.47)	.001
Group (intervention vs control)	–.21 (0.48)	.66
Time		
T2 versus T0	.35 (0.30)	.23
T1 versus T0	.07 (0.25)	.79
Sex		
Female versus male	.30 (0.44)	.49

XSL•FO
RenderX

	β (SE)	*P* value
Economic status		
>40,001 versus <10,000	2.16 (0.60)	<.001[a]
>30,001-40,000 versus <10,000	.70 (0.57)	.22
>20,001-30,000 versus <10,000	.97 (0.55)	.08
10,001-20,000 versus <10,000	.82 (0.55)	.14
Group × time		
Intervention×T2 versus control×T2	3.87 (0.49)	<.001[a]
Intervention×T1 versus control×T1	3.96 (0.42)	<.001[a]

[a]The difference between the 2 groups at a significance level of .001 (2-tailed).

[b]A currency exchange rate of NT $1=US $0.33 is applicable.

[c]The difference between the 2 groups at a significance level of .05 (2-tailed).

The intervention group showed a significantly higher increase in eHL scores than the control group at both posttest time points. Specifically, the improvement in the intervention group was greater than that in the control group (β=19.94, SE 3.52; *P*<.001 and β=18.19, SE 3.82; *P*<.001) at T1 and T2. These findings indicate that the intervention positively improved eHL, which persisted for up to 3 months afterward.

Similarly, patient health engagement increased following the intervention, with a higher mean score change in the intervention group than in the control group. Specifically, the improvement in the intervention group was significantly higher than in the control group (β=.28, SE 0.13; *P*=.03) at the initial posttest (T1). However, at T2 (3 months after the intervention), there was no significant difference between the 2 groups in the extent of the change (β=.25, SE 0.14; *P*=.06). The effect of the learning program on the patients' eHealth use was also examined. We compared the test scores of the intervention group at the 2 posttest time points and found that the mean score increase in the intervention group was higher than that in the control group. Specifically, the improvement in the intervention group was significantly greater than in the control group at T1 and T2 (β=3.96, SE 0.42; *P*<.001 and β=3.87, SE 0.49; *P*<.001).

Discussion

Principal Findings

This study examined the impact of an experiential learning program focused on eHealth care among individuals with type 2 diabetes. We found significant improvements in the intervention group's eHL, patient health engagement, and eHealth use compared with the control group following a series of educational sessions. The experiential learning program immersed participants in daily care activities and provided real-life experiences, resulting in a low attrition rate and positive acceptance of health care technology. These findings indicate that the program effectively enhanced participants' skills, attitudes, and behaviors. Enhancing eHL is crucial in the rapidly evolving eHealth era, in which an increasing array of eHealth devices is becoming available. When equipped with adequate eHL, patients with chronic diseases are better positioned to engage in self-management and address health inequities [36]. While many previous studies have conducted eHealth interventions, their findings on the effectiveness of eHealth care vary [37]. This theory-based eHealth intervention focuses on practical, experiential learning and the learning cycle, thus bridging this gap in the research because of its broad generalizability. After participants joined the program, we also found that increasing eHL is crucial to improving their health-related behavior. Patients became more actively involved in self-management when using the eHealth system. Consequently, this program demonstrates the potential for future application in clinical care, enabling patients to actively engage in a smart eHealth care system, implement self-management skills, and maintain disease control to improve quality of life. Moreover, validating our program with a clinical population of patients with diabetes underscores its clinical relevance and applicability, thereby demonstrating its value in real-world health care settings. The successful implementation and positive outcomes in this context indicate that the program can also be used to manage other chronic conditions.

The study found that participation in experiential learning courses effectively improved individuals' eHL, enabling them to use technology for daily self-management tasks. The course curriculum was based on the eHLF, encompassing all stages of the learning process. By engaging in experiential learning activities, participants applied learning cycles and integrated teaching strategies mirroring real-world situations in chronic disease management. This approach aligns with Kolb's theory, emphasizing learning through everyday life experiences, where knowledge is generated and transformed through practical engagement [19,20]. Learners fully engage in learning by incorporating active participation and hands-on experiences, fostering interest, encouraging questioning, and promoting reflective practices [20]. Our program comprised activities that closely resembled real-life situations, adhering to the principles of the eHLF to promote active participation among the learners. For example, during the concrete experience stage, we introduced eHealth care for the self-management of chronic diseases through lectures and videos to enhance patients' understanding of eHealth care. This approach enhances their interest in digital services, prompting them to take a more active role in managing their health, as indicated by the increased

XSL•FO
RenderX

patient health engagement shown in Figure 2. Moreover, in the active experimentation stage, we assisted patients in setting personal goals and facilitated hands-on learning by using various eHealth care devices. These activities provided them with access to appropriate digital services tailored to their health care needs and were designed to meet the unique challenges and requirements of the learners. As depicted in Figure 2, the intervention group recorded higher scores in eHealth care use than the control group. Through these experiential learning opportunities, participants gain knowledge and insights that evolve into personal skills they can integrate into their daily lives. This process empowers them to make informed decisions about their health and well-being, ultimately enhancing their eHL. By engaging in experiential learning courses, participants develop the necessary skills and confidence in the digital health domain [22,38].

Participants demonstrated enhanced overall eHL and significant progress in all 7 subscales. Specifically, notable improvements were observed in subscale 1 (using technology to process health information) and subscale 2 (understanding health concepts and language). To improve the ability to process information, the program effectively incorporated sessions that elucidated the importance of disease management, demonstrated appropriate health care technology applications, and introduced the concept of smart health care. This comprehensive learning program enabled participants to internalize these concepts and cultivate their abilities while actively engaging them in health care. Previous research has shown that explaining the principles that necessitate improvement not only enhances knowledge and skills but also diminishes barriers to implementation [20,26]. We also observed evident improvements in subscales 3, 5, 6, and 7.

To enhance their understanding of eHealth care, we introduced the patients in this study to various eHealth applications such as smart health care applications, wearable devices, and social media through case studies, lectures, and videos. We also used visual aids, questions, and case studies to enhance their knowledge about health care technology. Thus, patients could select suitable health care technologies that meet their individual care needs. Previous studies have indicated that higher health awareness makes it easier for individuals to comprehend health-related information and actively seek to improve their self-management abilities [6,39]. People can integrate knowledge through practical experience, case studies, and handouts [22,24].

In addition, the issue of information security often affects willingness to use eHealth technologies; however, subscale 4 showed clear improvements in this study. Each session provided information about health-related security, ways to preserve health information, and strategies for fraud prevention. Moreover, we encouraged patients to ask questions and express their concerns, which were clarified by the research team. Patients were informed about security measures and policies and how health care professionals use them. As certain software and tools can determine authenticity, patients' trust in technological information can be promoted through our intervention. Previous studies have suggested that asking and answering questions [40], assisting in the operation [26], and clarifying concerns [41] can help build people's trust in their abilities.

We also observed an improvement in patient health engagement. Although there was no significant difference in patient health engagement after 3 months, the level of engagement reached the arousal stage. Patient health engagement is a continuous, dynamic psychological process through which patients' cognition, emotions, and behaviors interact. This indicates that participants gradually accepted their illness and actively started to engage in disease management [35]. Previous research has highlighted the lack of knowledge on self-management of diseases and how to effectively use tools for disease management, leading to low engagement [6,42]. In this study, implementing the eHLF helped systematically organize participants' care needs and connect them to relevant eHealth applications. This enabled patients to understand, take action, and gradually internalize these practices while passing through the learning cycle. Faiola et al [42] also emphasized providing systematic mobile-based health frameworks based on patients' needs, empowering lifestyle changes, and promoting sustainable healthy behaviors.

Additionally, patient health engagement was measured on a 5-point scale, making it difficult to observe significant changes. Nevertheless, Figure 2 demonstrates the rise in patient health engagement, which can be attributed to patients learning to use health care technology daily and setting goals to track their health status. Through this program, we helped patients accumulate concrete, real-life self-care experiences, transformed abstract care concepts into usable knowledge, and assisted in operating health technology applications. Using their life experiences in case studies and addressing their questions reduced anxiety about technology use, promoted patient identification, and facilitated learning and engagement [6,43]. Furthermore, patients also participated in group discussions, which previous studies have shown to increase their sense of reality and engagement with eHealth care services [22,23]. This further supported patients' engagement by fostering a sense of community and shared experiences. As a result, patients are actively engaged with health care technology in their daily lives.

Finally, the intervention significantly improved participants' eHealth use. Our experiential learning program encouraged patients to participate in 6 sessions of the experience cycle. During these sessions, we demonstrated the operation of eHealth care devices and assisted each patient in selecting suitable devices according to their health conditions and abilities. Additionally, we provided simulation scenarios related to eHealth care, allowing patients to gradually familiarize themselves with operating these health care devices. Patients were encouraged to ask questions, and immediate answers were provided. Once taught, patients could use health technology for self-management independently. Educating individuals based on ELT can enhance their learning and skills [4,22,26]. Moreover, patients who engage in hands-on practice and receive information from simulation scenarios experience reduced fear and increased acceptance of new things [38]. Furthermore, we used LINE, a communication app, to remotely accompany patients in their daily lives. A web-based patient support group was established that allowed patients to ask questions at any

XSL•FO
RenderX

time and receive emotional support. We also addressed patients' questions individually whenever they encountered difficulties in using eHealth care. As a result, patients gained problem-solving skills and overcame any fear or hesitation in using health technology. These findings align with previous studies showing that when patients recognize the need to use such devices, they become more willing to share their needs and engage in discussions [9,43,44]. Additionally, when patients truly experience the benefits of using health technology, they integrate it as a helpful and effective part of their lives [45,46]. Therefore, our program has the potential to sustain patients' eHealth use.

Limitations

This study has several limitations. The 3-month follow-up period may not fully capture the long-term effectiveness of the intervention. Recognizing the importance of longer follow-ups for assessing the sustainability of health behaviors and disease management outcomes, future research should consider extended follow-up periods to more accurately assess health behavior and health outcome changes. Moreover, several strategies can be implemented to sustain the long-term effects of eHealth care experiential learning program. Empowering patients to participate actively in their health management is crucial. Continuous support and counseling through regular refresher sessions can help reinforce learning and maintain patient engagement. Therefore, integrating the program into routine clinical practice is essential, involving health care providers in the ongoing monitoring and supporting patients' eHealth activities to maintain the program's benefits. Providing peer support and fostering a shared responsibility for health management can also contribute to sustaining the program's impact. The extended follow-up includes collecting data on motivation (empowerment), health behaviors (such as patient health engagement, self-management, and eHealth use), and disease control indicators such as blood sugar levels, hemoglobin A_{1c} %, and lipid profiles. We will share these long-term findings in future publications. Additionally, a sex disparity was observed, with female participants being more active in health-related activities than male participants, partly due to the higher employment rates among male participants limiting their availability. Future studies should explore strategies to increase male's participation to improve external validity. Furthermore, this study's reliance on quantitative data may overlook insights that qualitative research can provide. Incorporating qualitative approaches in future research could yield a deeper understanding of user perspectives and the effectiveness of the ELT-based intervention. Addressing these limitations would enhance our understanding of the intervention's long-term impact and its applicability to a broader population of patients with chronic illnesses.

Conclusions

This study provided compelling evidence of theory-based interventions in eHealth care, displaying their effectiveness in improving eHL, patient health engagement, and eHealth use. By applying ELT, patients with type 2 diabetes demonstrated notable enhancements in their abilities and skills to use eHealth care technology. This approach clarifies the generalizability of the intervention and its components and contributes significantly to bridging the research gap in understanding the impact of eHL on patient behavior. A key finding from this study, after participants engaged with the program, was the crucial role of increasing eHL in improving patients' health-related behaviors. The study further underscored the significance of the eHLF as a fundamental knowledge base, acquired through immersive, real-world, self-management tasks using eHealth care tools within the learning cycle by demonstrating the tangible benefits of a theory-driven eHealth intervention in a clinical setting. The interventions designed to mimic patients' daily life situations closely facilitated the optimal learning outcomes and consequent changes in their health behavior. As such, integrating patient-centered care becomes imperative in the health care system to actively involve patients in eHealth care programs and promote improved disease management in the future.

Acknowledgments

This research was funded by the Ministry of Science and Technology, Taiwan (MOST-107-2314-B-010-014-MY3) and the Yen Tjing Ling Medical Foundation, Taiwan (CI-111-38). The authors thank the research assistants, case managers, and all the participants for their time and effort.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 92 KB - jmir_v26i1e53509_app1.pdf]

Multimedia Appendix 2
Content of the eHealth care experiential learning program.
[DOCX File , 30 KB - jmir_v26i1e53509_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

The study questionnaire.
[DOCX File , 19114 KB - jmir_v26i1e53509_app3.docx]

Multimedia Appendix 4
Mean changes in the measures of eHealth literacy, patient health engagement, and eHealth use over time (N=92).
[DOCX File , 21 KB - jmir_v26i1e53509_app4.docx]

Multimedia Appendix 5
Mean changes in the 7 eHealth literacy subscales over time (N=92).
[DOCX File , 23 KB - jmir_v26i1e53509_app5.docx]

References

1. The top 10 causes of death. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2022-08-24]
2. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022;183:109119 [FREE Full text] [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
3. American Diabetes Association. Economic costs of diabetes in the U.S. in 2017. Diabetes Care 2018;41(5):917-928 [FREE Full text] [doi: 10.2337/dci18-0007] [Medline: 29567642]
4. Helou N, Talhouedec D, Zumstein-Shaha M, Zanchi A. A multidisciplinary approach for improving quality of life and self-management in diabetic kidney disease: a crossover study. J Clin Med 2020;9(7):2160. [doi: 10.3390/jcm9072160] [Medline: 32650548]
5. Saghaee A, Ghahari S, Nasli-Esfahani E, Sharifi F, Alizadeh-Khoei M, Rezaee M. Evaluation of the effectiveness of persian diabetes self-management education in older adults with type 2 diabetes at a diabetes outpatient clinic in Tehran: a pilot randomized control trial. J Diabetes Metab Disord 2020;19(2):1491-1504 [FREE Full text] [doi: 10.1007/s40200-020-00684-0] [Medline: 33520849]
6. Davis J, Fischl AH, Beck J, Browning L, Carter A, Condon JE, et al. 2022 National standards for diabetes self-management education and support. Sci Diabetes Self Manag Care 2022;48(1):44-59. [doi: 10.1177/26350106211072203] [Medline: 35049403]
7. Carpenter R, DiChiacchio T, Barker K. Interventions for self-management of type 2 diabetes: an integrative review. Int J Nurs Sci 2019;6(1):70-91 [FREE Full text] [doi: 10.1016/j.ijnss.2018.12.002] [Medline: 31406872]
8. Domecq JP, Prutsky G, Elraiyah T, Wang Z, Nabhan M, Shippee N, et al. Patient engagement in research: a systematic review. BMC Health Serv Res 2014;14:89 [FREE Full text] [doi: 10.1186/1472-6963-14-89] [Medline: 24568690]
9. Chen Y, Chang P. Using eHealth in the continuity care of chronic kidney disease: opportunities and considerations. Hu Li Za Zhi 2016;63(2):12-18. [doi: 10.6224/JN.63.2.12] [Medline: 27026552]
10. Bodie GD, Dutta MJ. Understanding health literacy for strategic health marketing: eHealth literacy, health disparities, and the digital divide. Health Mark Q 2008;25(1-2):175-203. [doi: 10.1080/07359680802126301] [Medline: 18935884]
11. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. J Med Internet Res 2011;13(1):e20 [FREE Full text] [doi: 10.2196/jmir.1600] [Medline: 21345783]
12. Kayser L, Kushniruk A, Osborne RH, Norgaard O, Turner P. Enhancing the effectiveness of consumer-focused health information technology systems through eHealth literacy: a framework for understanding users' needs. JMIR Hum Factors 2015;2(1):e9 [FREE Full text] [doi: 10.2196/humanfactors.3696] [Medline: 27025228]
13. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
14. Gilstad H. Toward a comprehensive model of eHealth literacy. 2014 Presented at: Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014); May 19, 2014; Trondheim, Norway p. 63-72 URL: https://ceur-ws.org/Vol-1251/paper7.pdf [doi: 10.13140/2.1.4569.0247]
15. Kaufman DR, Mirkovic J, Chan C. eHealth literacy as a mediator of health behaviors. In: Patel VL, Arocha JF, Ancker JS, editors. Cognitive Informatics in Health and Biomedicine. Health Informatics. Cham: Springer; 2017:271-297.
16. Schulz PJ, Fitzpatrick MA, Hess A, Sudbury-Riley L, Hartung U. Effects of eHealth literacy on general practitioner consultations: a mediation analysis. J Med Internet Res 2017;19(5):e166 [FREE Full text] [doi: 10.2196/jmir.6317] [Medline: 28512081]
17. Barello S, Graffigna G, Vegni E. Patient engagement as an emerging challenge for healthcare services: mapping the literature. Nurs Res Pract 2012;2012:905934 [FREE Full text] [doi: 10.1155/2012/905934] [Medline: 23213497]

XSL•FO
RenderX

18. Menichetti J, Libreri C, Lozza E, Graffigna G. Giving patients a starring role in their own care: a bibliometric analysis of the on-going literature debate. Health Expect 2016;19(3):516-526 [FREE Full text] [doi: 10.1111/hex.12299] [Medline: 25369557]
19. Kolb D. Experiential Learning: Experience as the Source of Learning and Development. Englewood Cliffs, NJ: Prentice Hall, Inc; 1984.
20. Kolb D. Experiential Learning: Experience as the Source of Learning and Development. 2nd Edition. Upper Saddle River, NJ: Pearson FT Press; 2015.
21. Kao CC. Applications of the Kolb Experiential Learning Cycle to improve the attitude of nursing students toward older people. Macau J Nur 2018;16(1-2):1-6 [FREE Full text] [doi: 10.6729/MJN.2017.001]
22. Hsu H, Hsiao J, Yeh C, Huang C. Effects of experiential learning theory in the simulation experience course for nursing students. VGH Nur 2019;36(4):425-432 [FREE Full text] [doi: 10.6142/VGHN.201912_36(4).0011]
23. Dewi EU, Nursalam, Mahmudah, Yunitasari E. The effect of peer support psychoeducation based on experiential learning on self-care demands among breast cancer patients with post-chemotherapy. J Public Health Res 2023;12(1):22799036221146901 [FREE Full text] [doi: 10.1177/22799036221146901] [Medline: 36643605]
24. Falloon G. Using simulations to teach young students science concepts: an experiential learning theoretical analysis. Comput Educ 2019;135:138-159. [doi: 10.1016/j.compedu.2019.03.001]
25. Li C, Yang Y, Jing Y. Formulation of teaching strategies for graduation internship based on the experiential learning styles of nursing undergraduates: a non-randomized controlled trial. BMC Med Educ 2022;22(1):153 [FREE Full text] [doi: 10.1186/s12909-022-03221-0] [Medline: 35255897]
26. Rasouli S, Namnabati M. Nurses' learning of infants' venipuncture based on Kolb's learning theory. J Neonatal Nurs 2019;25(5):245-248. [doi: 10.1016/j.jnn.2019.04.002]
27. Hennemann S, Witthöft M, Bethge M, Spanier K, Beutel ME, Zwerenz R. Acceptance and barriers to access of occupational e-mental health: cross-sectional findings from a health-risk population of employees. Int Arch Occup Environ Health 2018;91(3):305-316. [doi: 10.1007/s00420-017-1280-5] [Medline: 29189895]
28. Guo SH, Hsing H, Lin J, Lee C. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR Mhealth Uhealth 2021;9(2):e18404 [FREE Full text] [doi: 10.2196/18404] [Medline: 33544088]
29. Cohen S. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of hHealth. Thousand Oaks, CA: Sage Publications, Inc; 1988:31-67.
30. Norgaard O, Furstrand D, Klokker L, Karnoe A, Batterham R, Kayser L, et al. The e-health literacy framework: a conceptual framework for characterizing e-health users and their interaction with e-health systems. KM&EL 2015;7(4):522-540 [FREE Full text] [doi: 10.34105/j.kmel.2015.07.035]
31. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
32. Chen Y, Cheng C, Osborne RH, Kayser L, Liu C, Chang L. Validity testing and cultural adaptation of the eHealth Literacy Questionnaire (eHLQ) among people with chronic diseases in Taiwan: mixed methods study. J Med Internet Res 2022;24(1):e32855 [FREE Full text] [doi: 10.2196/32855] [Medline: 35044310]
33. Graffigna G, Barello S, Bonanomi A, Lozza E. Measuring patient engagement: development and psychometric properties of the Patient Health Engagement (PHE) Scale. Front Psychol 2015;6:274 [FREE Full text] [doi: 10.3389/fpsyg.2015.00274] [Medline: 25870566]
34. Zhang Y, Graffigna G, Bonanomi A, Choi K, Barello S, Mao P, et al. Adaptation and validation of a Chinese version of Patient Health Engagement Scale for patients with chronic disease. Front Psychol 2017;8:104 [FREE Full text] [doi: 10.3389/fpsyg.2017.00104] [Medline: 28220090]
35. Graffigna G, Barello S. Modelling patient engagement in healthcare: insight for research and practice. In: Patient Engagement. Latin America: Sciendo Migration; 2015:27-43.
36. Refahi H, Klein M, Feigerlova E. e-Health literacy skills in people with chronic diseases and what do the measurements tell us: a scoping review. Telemed J E Health 2023;29(2):198-208. [doi: 10.1089/tmj.2022.0115] [Medline: 35671526]
37. Wang Y, Min J, Khuri J, Xue H, Xie B, Kaminsky LA, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. JMIR Mhealth Uhealth 2020;8(4):e15400 [FREE Full text] [doi: 10.2196/15400] [Medline: 32343253]
38. Zhang L, Liu J, Yan C, Zhao Z, Tang J, Ma X. The effect of experiential learning on attitude and willingness to insulin therapy in patients with type 2 diabetes. Int J Clin Exp Med 2016;9(9):18104-18113 [FREE Full text]
39. Chen X, Hay JL, Waters EA, Kiviniemi MT, Biddle C, Schofield E, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734 [FREE Full text] [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
40. Chmil JV, Turk M, Adamson K, Larew C. Effects of an experiential learning simulation design on clinical nursing judgment development. Nurse Educ 2015;40(5):228-232. [doi: 10.1097/NNE.0000000000000159] [Medline: 25763781]

XSL•FO
RenderX

41. Bonner GJ, Freels S, Ferrans C, Steffen A, Suarez ML, Dancy BL, et al. Advance care planning for African American caregivers of relatives with dementias: cluster randomized controlled trial. Am J Hosp Palliat Care 2021;38(6):547-556 [FREE Full text] [doi: 10.1177/1049909120916127] [Medline: 32308012]
42. Faiola A, Papautsky EL, Isola M. Empowering the aging with mobile health: a mHealth framework for supporting sustainable healthy lifestyle behavior. Curr Probl Cardiol 2019;44(8):232-266. [doi: 10.1016/j.cpcardiol.2018.06.003] [Medline: 30185374]
43. Shen H, van der Kleij R, van der Boog PJM, Wang W, Song X, Li Z, et al. Digital tools/eHealth to support CKD self-management: a qualitative study of perceptions, attitudes and needs of patients and health care professionals in China. Int J Med Inform 2022;165:104811 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104811] [Medline: 35753175]
44. Knowles M, Holton IE, Swanson R. The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development. 8th Edition. London and New York: Routledge; 2014.
45. Burner ER, Menchine MD, Kubicek K, Robles M, Arora S. Perceptions of successful cues to action and opportunities to augment behavioral triggers in diabetes self-management: qualitative analysis of a mobile intervention for low-income Latinos with diabetes. J Med Internet Res 2014;16(1):e25 [FREE Full text] [doi: 10.2196/jmir.2881] [Medline: 24476784]
46. Öberg U, Isaksson U, Jutterström L, Orre CJ, Hörnsten Å. Perceptions of persons with type 2 diabetes treated in Swedish primary health care: qualitative study on using eHealth services for self-management support. JMIR Diabetes 2018 Mar 12;3(1):e7 [FREE Full text] [doi: 10.2196/diabetes.9059] [Medline: 30291075]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
eHL: eHealth literacy
eHLF: eHealth literacy framework
eHLQ: eHealth Literacy Questionnaire
ELT: experiential learning theory
GEE: generalized estimating equation

Edited by S Ma; submitted 09.10.23; peer-reviewed by W Wei, M Savolainen, M Shen; comments to author 14.12.23; revised version received 07.02.24; accepted 21.06.24; published 16.08.24.

Please cite as:
Cheng YS, Lin CP, Chen LYA, Hwang WR, Lin YC, Chen YC
Short-Term Effects of an eHealth Care Experiential Learning Program Among Patients With Type 2 Diabetes: Randomized Controlled Trial
J Med Internet Res 2024;26:e53509
URL: https://www.jmir.org/2024/1/e53509
doi:10.2196/53509
PMID:

©Yu-Shan Cheng, Cheng-Pei Lin, Lu-Yen Anny Chen, Wei-Ren Hwang, Yi-Chun Lin, Yu-Chi Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impacts of Clinical Decision Support Systems on the Relationship, Communication, and Shared Decision-Making Between Health Care Professionals and Patients: Multistakeholder Interview Study

Florian Funer[1,2*], MD, MA, MagTheol; Diana Schneider[3*], MA; Nils B Heyen[3], PhD; Heike Aichinger[3], PhD; Andrea Diana Klausen[4], Dipl. Special Need Education; Sara Tinnemeyer[2], MEd; Wenke Liedtke[5,6], PhD; Sabine Salloch[2], PhD, MD; Tanja Bratan[3], PhD

[1]Institute for Ethics and History of Medicine, Eberhard Karls University Tuebingen, Tuebingen, Germany
[2]Institute for Ethics, History and Philosophy of Medicine, Hannover Medical School, Hannover, Germany
[3]Competence Center Emerging Technologies, Fraunhofer Institute for Systems and Innovation Research ISI, Karlsruhe, Germany
[4]Institute for Medical Informatics, University Medical Center – RWTH Aachen, Aachen, Germany
[5]Department of Social Work, Protestant University of Applied Sciences Rhineland-Westphalia-Lippe, Bochum, Germany
[6]Ethics and its Didactics, Faculty of Theology, University of Greifswald, Greifswald, Germany
[*]these authors contributed equally

Corresponding Author:
Diana Schneider, MA
Competence Center Emerging Technologies
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Straße 48
Karlsruhe, 76139
Germany
Phone: 49 0721 6809 58
Email: diana.schneider@isi.fraunhofer.de

Abstract

Background: Clinical decision support systems (CDSSs) are increasingly being introduced into various domains of health care. Little is known so far about the impact of such systems on the health care professional–patient relationship, and there is a lack of agreement about whether and how patients should be informed about the use of CDSSs.

Objective: This study aims to explore, in an empirically informed manner, the potential implications for the health care professional–patient relationship and to underline the importance of this relationship when using CDSSs for both patients and future professionals.

Methods: Using a methodological triangulation, 15 medical students and 12 trainee nurses were interviewed in semistructured interviews and 18 patients were involved in focus groups between April 2021 and April 2022. All participants came from Germany. Three examples of CDSSs covering different areas of health care (ie, surgery, nephrology, and intensive home care) were used as stimuli in the study to identify similarities and differences regarding the use of CDSSs in different fields of application. The interview and focus group transcripts were analyzed using a structured qualitative content analysis.

Results: From the interviews and focus groups analyzed, three topics were identified that interdependently address the interactions between patients and health care professionals: (1) CDSSs and their impact on the roles of and requirements for health care professionals, (2) CDSSs and their impact on the relationship between health care professionals and patients (including communication requirements for shared decision-making), and (3) stakeholders' expectations for patient education and information about CDSSs and their use.

Conclusions: The results indicate that using CDSSs could restructure established power and decision-making relationships between (future) health care professionals and patients. In addition, respondents expected that the use of CDSSs would involve more communication, so they anticipated an increased time commitment. The results shed new light on the existing discourse by demonstrating that the anticipated impact of CDSSs on the health care professional–patient relationship appears to stem less from the function of a CDSS and more from its integration in the relationship. Therefore, the anticipated effects on the relationship between health care professionals and patients could be specifically addressed in patient information about the use of CDSSs.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55717) doi:10.2196/55717

KEYWORDS

clinical decision support system; CDSS; health care professionals; patients; relationships; communication; shared decision-making; Germany

Introduction

Background

A multitude of expectations and hopes are associated with the development of technologies to support difficult decision-making situations in health care. Clinical decision support systems (CDSSs) are seen as a possible solution for managing an exploding amount of health care data and scientific evidence in the decision-making for individual patients. In particular, the combination of advanced data storage and processing capabilities enabled by artificial intelligence (AI), or rather, machine learning, have led to a rapid exploration of the application areas in almost all health domains (eg, diagnostics, prognostics, therapeutics, and research).

The use of CDSSs is associated with ethical and social issues, such as patient safety, algorithmic transparency, lack of appropriate regulation, liability, and accountability and requires the development of a sophisticated governance approach with regard to AI-empowered health care [1-4]. The use of CDSSs also adds another layer to the health care professional–patient relationship. Therefore, it will invariably be connected with certain effects, for example, on the physician–patient relationship, that are likely to involve changes in the mode of interaction [4-6]. On the one hand, the use of CDSSs is expected to reduce workload and facilitate a more empathetic relationship with patients due to the time gained [7]; by contrast, it is feared that CDSSs will undermine the relationship of trust between patients and health care professionals, for example, through the so-called computer paternalism [4,6].

There are comparatively a large number of studies that address the potential impact of CDSSs on professionals' clinical decision-making, in part by drawing on social epistemology or literature on professionals' moral responsibilities [8-15]. Due to the expected impact of CDSSs on the doctor–patient relationship, some scholars suggest that it could evolve into a "doctor–computer–patient relationship" [16-18]. Illustrating this new kind of a triadic relationship, Braun et al [13] outlined three different interaction modes that illustrate how AI-based CDSSs could influence the roles of clinical agents:

1. A conventional AI-based CDSS, which supports health care professionals by generating options from patient data in the electronic health record (EHR) for professionals to consider.
2. An integrative AI-based CDSS could instead support shared decision-making between clinicians and patients, as such systems not only collect the data sets from the EHR and other patient data (thus expanding the EHR) but also present the data analysis in a way that can be used for patient interaction. The results of the algorithmic analysis not only can constructively complement the shared decision-making process but can also exacerbate existing tensions between the agents due to the lack of transparency of the AI-based CDSSs [13].
3. A fully automated AI-based CDSS that interacts directly with patients and enables them to make decisions independently and without human clinical expertise, for instance, with regard to taking medication. So far, it is unclear whether this interaction mode can contribute to increased patient participation or create new vulnerabilities, (eg, due to a lack of data literacy and biased training data) [13].

These modes of interaction are shown in Figure 1.

Figure 1. Modes of interaction illustrating how artificial intelligence (AI)–based clinical decision support systems (CDSSs) could influence the roles of clinical agents (author's representation of the 3 types of interaction developed by Braun et al [13]).

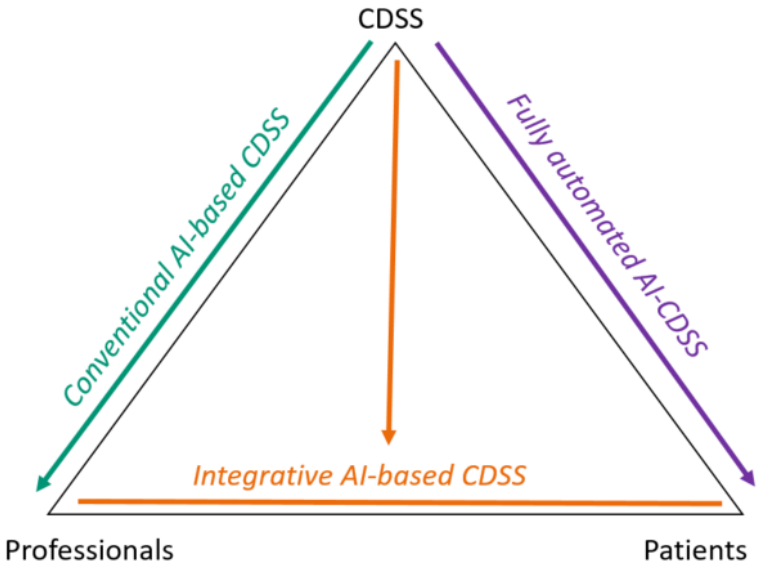

XSL•FO
RenderX

Objective

Despite this intense debate, there are only a few studies in which the empirically gathered perspectives of patients and clinicians [11,15,19-21] on a CDSS are presented together; there are exceptions such as those studies in which digital tools are intended to support the interaction between clinicians and patients, such as patient decision aids [22]. Despite the increased use of CDSSs in nursing [23-26], empirical studies with a focus on nurse–patient interactions involving CDSSs are still scarce. Therefore, the main motive for our research project was to evaluate and highlight the ethical and social aspects in terms of their importance for the various stakeholders in clinical decision-making. As part of a qualitative study combining semistructured interviews and focus groups, we explored how different stakeholders perceive the inclusion of CDSSs in decision-making processes within health care, what effects they expect for the health care professional–patient relationship, and to what extent the use of CDSSs should be addressed in information and education provided to patients. This study helps to rectify the lack of evidence in at least 3 ways: first, we explore expectations from a multi-professional perspective, interviewing both future doctors and future nurses; second, we consider the perspectives of future health care professionals and of patients; third, we include different forms of decision support for different areas of health care (ie, surgery, nephrology, and outpatient intensive care). This enables us to identify similarities and differences in the use of CDSSs in various application areas.

Methods

Overview

The study is part of the research consortium Decision Support in Routine and Emergency Health Care: Ethical and Social Implications (DESIREE). The aim of the research project is to explore key ethical, social, professional, and technical aspects of digital decision support in health care by combining empirical and theoretical approaches. The reporting was guided by the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [27].

Recruitment and Sample

The data collection was carried out between April 2021 and April 2022 as a methodological triangulation at 2 different institutions in Germany. On the one hand, individual semistructured interviews were conducted at a maximum care hospital with 15 medical students and 12 trainee nurses, all of whom were near the end of their training. In this way, individual attitudes in dealing with CDSSs could be examined in depth, and comparisons could be made between the different professions. We included people at the end of their respective training and, therefore, future health care professionals as interviewees for 2 main reasons. First, this was a relatively homogenous sample of people with the same level of medical or nursing experience, education, and familiarity with technologies in everyday life. Second, they had comparable knowledge of current technologies in the health care sector, even if only introductory, due to the content of their training. On the other hand, 3 focus groups with 18 patients (5-7 per group) were conducted by Fraunhofer Institute for Systems and Innovation Research ISI. This aimed to open up a communicative space for patients in which they could exchange opinions and viewpoints and thus foster mutual reflection as "experts in their own right."

A convenience sampling approach was used, and respondents were included if they belonged to one of the stakeholder groups: patients who are members of self-help groups or institutions, medical students (fourth and fifth year of study), and trainee nurses (second and third year of training); were aged ≥18 years; had sufficient knowledge of German; were capable of giving consent; and were, according to their own declaration, in sufficiently good health. Since the discourse within a focus group can be considered as an expression of collective knowledge and homogeneous groups (ie, individuals with similar experiences and similar diseases) are better suited for this purpose [28], study participation was promoted in selected patient self-help groups. With the exception of 1 participant (acute hospitalization), all participants (44/45, 98%) recruited for the study took part in the interview or focus group.

There was no relationship established between participants and researchers before the interviews or focus groups. In some cases, focus group respondents knew each other. Respondents were informed in advance that the interview or focus group topic was "clinical decision support systems and digitalization in health care" but not the specific questions that were to be discussed.

Data Collection

Due to the COVID-19 pandemic and mobility issues of some patients, all individual interviews and focus groups were conducted via video calls instead of face to face. Respondents were generally at home during the interview and focus group; most respondents were also alone unless they needed caregiver support. In such cases, the caregivers did not actively participate in the focus group. The interviewers used interview guides for the different stakeholders (refer to Multimedia Appendices 1 [15,29] and 2). As part of each interview and focus group, we used a stimulus including illustrations that introduced 3 different CDSS application scenarios. The CDSSs used as stimuli were (1) the AI-supported CKDNApp to assist nephrologists in prognosis assessment and therapy planning for patients with chronic kidney disease; (2) an AI-supported CDSS to assist surgeons with intraoperative navigation by providing recommendations for incision lines; and (3) the Safety Box, a rule-based alarm system designed to provide nurses and relatives with recommendations for action in the care of ventilated patients in the home environment. One of these CDSS examples was presented in each focus group, based on which a CDSS was used to treat their specific condition; medical students were presented with stimuli 1 and 2, and trainee nurses were presented with stimulus 3. The similarly structured stimuli allowed both a comparable understanding of the CDSS presented and, especially in the case of patients, abstraction from their own personal situation. By being confronted with the CDSS stimuli, respondents were asked about their hopes and concerns of implementing such technology, especially regarding the health care professional–patient relationship. They were asked whether and how the relationship might change and what information they thought was important to know before such a system could

XSL•FO
RenderX

be deployed. From the responses, different perceptions and expectations of CDSSs can be derived, which in turn may have different implications for the health care professional–patient relationship. This also results in specific information and education needs for professionals.

Audio recordings were made to document the interviews and focus groups, and recordings were transcribed ad verbatim and pseudonymized. Field notes were taken during and after the interviews or focus groups. Data collection was terminated when theoretical saturation had been reached, that is, when additional interviews or focus groups did not provide additional information with respect to the research question.

Data Analysis

The individual interviews lasted on average 51:26 (range 29:44-75:37) minutes; the focus groups lasted on average 114:17 (range 110:07-123:04) minutes, including a short break after 1 hour. The sociodemographic characteristics of the respondents are presented in Table 1.

The data were analyzed using structured qualitative content analysis according to Kuckartz and Rädiker [30]. The recording transcripts were assigned categories that were drawn both deductively from literature research and inductively from the interviews or focus groups themselves. Coding rules were explicated for the main categories, and examples were identified (refer to coding tree in Multimedia Appendix 3). The categories were developed and revised during the coding procedure. Two of the authors independently coded the collected data using the software MAXQDA 2020 (VERBI GmbH); one coded the transcripts of the individual interviews, and the other coded the transcripts of the focus groups. Ambiguous passages were discussed within the research team and then solved by consensus. During the coding, the focus of the analysis was directed toward distinct topics, such as role setting and relationships or patient information and education. Suitable exemplary codes were selected in order to answer the research question; these have been translated from German into English and included in this paper.

Table 1. Sociodemographic data of persons participating in individual interviews and focus groups (n=45).

	Medical students	Trainee nurses	Patients
Participants, n (%)	15 (33)	12 (27)	18 (40)
Age (y), mean (range)	25.5 (23-36)	25 (20-50)	45.5 (24-76)
Gender (self-reported), n (%)			
Men	7 (47)	2 (17)	6 (33)
Women	8 (53)	10 (83)	12 (67)

Ethical Considerations

This study was approved by the Research Ethics Committee of the Hannover Medical School, Germany (registration number 9805_BO_K_2021) as part of the project DESIREE. All respondents provided informed documented consent to participate in this study. At any time during the study, respondents had the opportunity to stop participating. Respondents received a small amount of monetary compensation for taking part in the interview or focus group.

Results

General Information

The empirical results will be presented in three thematic parts: (1) CDSSs and their impact on the roles of and requirements for professionals, (2) CDSSs and their impact on the relationship between health care professionals and patients (including communication requirements for shared decision-making), and (3) stakeholders' expectations for patient education and information about the CDSS and its use. In the following sections, we use the abbreviations "SI" for (medical) student interviews, "TI" for (nursing) trainee interviews, and "PFG" for patient focus group.

CDSSs and Their Impact on the Roles of and Requirements for Professionals

The respondents agreed that CDSSs should only be used to *support their decision-making*. They believed that this kind of support is provided by making expert or aggregated knowledge available to CDSS users. Some patients associated CDSSs with being able to pool expertise from different institutions (eg, across clinics), federal states (in Germany), or countries (PFG nephrology). As such, CDSSs could also be used in the training of future specialists as learning support (PFG surgery). Consequently, a CDSS is not seen as an "agent" in its own right, but only as an aid:

> *[I]t is also important for work that it's still made clear: "Okay, this is not a new agent—so to speak—an independent one, but a supplementary aid to what I am doing as a doctor at this moment."* [SI5]

In general, respondents understood support for decision-making as assistance of an individual professional within their decision-making rather than supporting a team. Patients emphasized that the use of CDSSs should not lead to the replacement of professionals who, for example, have more experience and can consider different factors and perspectives in their decision-making:

> *I think that the experience that a doctor has, I believe that nothing can replace it, not with any technological things, because everyone knows that from route guidance systems (laughing).... So I think the [the doctor's] combination of experience is very important and that you can really see it [the CDSS] as support.* [PFG surgery]

XSL•FO
RenderX

Medical students, however, expected a potential replacement in fields such as radiology and claimed, for instance, that they take such considerations into account when choosing their residency training [eg, SI9].

Especially for the Safety Box to support care in home environments, patients saw the risk that the use of CDSSs could lead to a devaluation of the nursing profession if financial resources are invested more in the acquisition of technical systems such as CDSSs rather than in the training of professionals:

> *[It is better that the nurse] is well trained and can communicate with me when I am not able to speak or that I can interpret the alarms from the ventilator correctly, that she knows what to do, than that I bring in such a [CDSS] that thinks for the nurse, but why should it think for the nurse when the nurse has a mind of her own?* [PFG home care]

One concern of the professionals is a potential information overload, as they do not have the time or competences to check all the recommendations of a CDSS comprehensively, even if the CDSS refers their base of scientific evidence, research literature, or guidelines. However, if doctors can no longer keep track of large amounts of data, then they are not in a position to check for sources of error. In this sense, 1 medical student maintained the following:

> *[T]hen I can't actually check this algorithm anymore or track it, to see if there's an error in it somewhere or if data was collected incorrectly somewhere.* [SI8]

Some students therefore feared that they may become dependent on the algorithmic support, eventually even "helplessly in the hands of the developers of the CDSS" [SI8]. Some patients were equally concerned that the use of CDSSs would provide professionals with a false sense of security and that individual characteristics of a case might be neglected because of the CDSS recommendation obtained on the basis of probability.

Therefore, respondents emphasized that the use of CDSSs places new demands on the profession. Some students spoke of a change in professional self-image, so that responsibilities and demands on the medical profession might change and call for much more in-depth education, for example, on how data are generated and how CDSSs work in general:

> *[T]hat you understand what data are, how they are processed and how a kind of neural artificial network works. That you can also weigh things up and that this weighting can also be further differentiated by machine learning and that it then produces an output, which I can use afterwards for my clinical decision-making. That would also, I think, be enough for me [to know] for now.* [SI8]

Correspondingly, patients emphasized that professionals need to understand, comprehend, and be able to communicate information that comes from the CDSSs. They insisted that the training of professionals should include basic medical, statistical, and information technological knowledge. That should enable professionals both to evaluate CDSS information in a well-founded manner and to make clinical judgments without such support. Indeed, patients would only accept the use of CDSSs in everyday practice if well-trained medical experts were present all the time, that is, inexperienced professionals should not interact with the CDSSs on their own. Moreover, some patients expected AI-based CDSSs to be used only after professionals have had intensive training (eg, after acquiring a device operator's license after a minimum number of training sessions).

One of the unique features of the focus groups with patients was the concern that the CDSSs could be used for other purposes in addition to improving the quality of care. On the one hand, patients feared that the use of CDSSs would be prioritized due to economic interests and that professionals could be influenced in their decision-making to deviate from their conviction about the most suitable option, for example, because only certain treatments are financially compensated (eg, PFG home care). On the other hand, patients recognized the potential for the use of CDSSs to lead to greater control of patients and their adherence to treatment, for example, by other stakeholders such as health insurers:

> *I was just wondering whether something like this could also arouse desires, i.e. if this data is available in such a simple form. Whether health insurance companies, for example, would come up with the idea to say: "We will only fund this medication for patients where the doctors can prove that the patients adhere well to the instructions." In other words, it [the AI-CDSS] could be detrimental to you because there is a desire for it elsewhere.* [PFG nephrology]

Patients expected that such third-party interests could reduce the range of medical services actually available to them. Other interests thus become intertwined in the decision-making and interaction between patients and professionals.

CDSSs and Their Impact on the Relationship Between Professionals and Patients

Most respondents expected changes in the relationship and communication between health care professionals and patients through the use of CDSSs. Communication was cited by the respondents as an important factor for successful treatment:

> *I see this as a double-edged sword. Because as we have already said a few times today, that this communication with the affected person, yes, that is the key. And not just in the situation where the person needs me, but maybe even also during the rest of the time.* [PFG home care]

The respondents distinguished 3 possible scenarios in their representations. First, some respondents (particularly trainee nurses) were concerned that the use of CDSSs would lead to an increasing focus on technology because professionals would "cling very much to the technology" [SI10] or "just concentrate on the screen" [TI3], leading to reduced attention to patients. Patients had similar concerns, fearing a maldistribution of attention:

> *I see it similarly to [R7] that I always say to nurses: "Look into my eyes first and see how I make gestures and perceive, and don't just worry about the beep or*

XSL•FO
RenderX

> *the device." Exactly, so I find this [people-to-people] contact much, much more important.* [PFG home care]

Some patients feared that the use of CDSSs could undermine trust and interaction; this is not limited to the cases where a CDSS has a control function (eg, nonadherence of therapy measures by patients). They feared that health care professionals' actions could be reduced mostly to basic care:

> *We do not want to be called or seen as a "living object," so to speak. We want professionals to communicate with us, maybe in a different way than at the moment, but we don't want them to just.... I mean, I don't even want to compare it to an animal. But that they just feed us and then it's enough. Then you can move away, because the technology is at work, and then I don't have the connection to caregivers anymore.* [PFG home care]

Since they linked CDSSs with the tendency toward the standardization of treatment, patients on a ventilator at home feared that patient autonomy could be impeded by the use of CDSSs. They associated standardization in nursing above all with making services more economical, which would then be less flexible in their response to individual patient needs. Therefore, they were concerned about an increase in computer paternalism:

> *There is always a very big risk of being impinged upon by such a device.* [PFG home care]

In addition, particularly home care patients emphasized that the "right to be medically unreasonable" must still exist even with the implementation of CDSSs. They associated this with the right to be able to make independent decisions that go beyond defined medical limits, in case of doubt, even against both medical necessity and the recommendation of the CDSS:

> *We also have a right to medical irrationality, and this is of course completely undermined by such a device. If the device says: You're over-breathing, you're getting too much air... and the person themselves says: "Yes, I know, but I need this because I can speak better that way." Then there is already a conflict: Am I allowed to act against the medical advice of this device? But that is also a conflict for the nurse. Okay, do I follow what the patient says or what the device says?* [PFG home care]

Second, some respondents recognized the potential for CDSSs' visualizations to contribute in a beneficial way, for example in the case of the CKDNApp for communicating prognostic statements or in the case of the Safety Box for explaining professional actions and their necessity to patients. Although the benefit of using visuals for shared decision-making is not unique to the use of CDSSs, such ideas were enough for many medical students to assume that the use of CDSSs would lead to a necessity for increased communication in the future: on the one hand, the "increased communication" brought about by the use of CDSSs can strengthen shared decision-making, since the respondents expected that communication about CDSS results would be regularly included in future discussions between patients and professionals. On the other hand, it would also lead to additional effort, since the necessity to communicate the results and their interpretation could come to the fore, even if there is limited time available. In particular, applications that potentially involve patients would require professionals to communicate more frequently with their patients about how certain CDSS results are obtained, how they are to be interpreted, and what implications they may have on the patient's circumstances.

The inclusion of CDSSs in the shared decision-making process is seen as problematic in certain situations, particularly if professionals are unable to assess the appropriateness of a specific CDSS recommendation for a patient (refer to the CDSSs and Their Impact on the Roles of and Requirements for Professionals section). If the clinical judgment of the professional does not match with the recommendation of the CDSS, some medical students see the only reasonable solution in communicating both the CDSS recommendation and their own recommendation to the patient, thereby invoking the principle of "patient autonomy" [SI6]:

> *Then I would say both, as I always think that patients should be informed, and they should be informed throughout, all the way across the spectrum, which means you should give them all their options and I think patients should be able, I mean, responsible adult patients, should be able to make their own decisions and figure out for themselves where they put their trust. So, I would say: "I did this with an app [CKDNApp] with a certain data set that stated, your prognosis is rather poor. I personally using my clinical experience would say I don't see it as pessimistically, but the bottom line is it's hard to say." So, I would list both options and then it is up to the patient...and then the patient can think for himself, whom or what he trusts more* [SI4]

This way of dealing with such challenges was not an isolated case. Indeed, most patients also expected professionals to be transparent about whether and how CDSS recommendations had been incorporated into their professional decision:

> *Yes, I think it's better if the doctor tells me, because then I know whether it comes from him/her, from his/her experience, from his/her many years of experience with patients, or simply from this app, from mathematical things, etc., from different values.* [PFG nephrology]

According to the respondents, the involvement of CDSSs in decision-making processes could also lead to a loss of trust in the health care professional–patient relationship, if it is somewhat unclear to patients where a recommendation for therapy comes from:

> *Yes, I find it a bit difficult when the doctor says that everything is fine now, for example, I think she/he has to tell me that it comes from her/him or from this app. So I don't know if the trust is still 100% there when this app is involved.* [PFG nephrology]

However, some medical students also felt that communicating divergent recommendations to patients would be inappropriate.

XSL•FO
RenderX

They feared additional confusion on the part of the patients because they would have difficulty deciding which recommendation they should follow.

While 1 medical student surmised that patients would generally trust the recommendation of the human more than that of the CDSSs, this was far from clear to patients. They argue that trusting someone is also a conscious decision, regardless of whether you trust people or technical systems. Some patients perceived the recommendations from CDSSs as a second opinion. If the professional judgment and the recommendation of the CDSS should diverge more frequently, some patients feared that they might get into a conflict of loyalty:

> *And if the doctor would say to me: "The app recommends that... but I would like to tell you something completely different."; and I would always follow the app, then I would actually get into a predicament somehow, it would be about me and the doctor as individuals. If I were to keep saying: "I don't trust you like I trust the app, so please do it the way the app says," then I think I would get into difficulties with the doctor at some point.* [PFG nephrology]

Finally, 1 patient changed their opinion during the focus group, as they felt that this complete transparency had not been provided so far either:

> *Well, for me it wouldn't be important, I'm realizing in the course of the evening. The doctor also seeks advice from studies, for example, and then doesn't tell me every time: "There is a new study on which I am now basing my opinion, where there are new findings."* [PFG nephrology]

Third, some respondents associated the use of CDSSs with the opportunity for patients to be more actively involved in decision-making and therapy planning. For example, in the case of the CKDNApp, patients assumed that their own role in the treatment process could be strengthened by contributing independently to the collection of data:

> *I think it would be good if you, as a patient, could see at least partial data [of the CDSS] yourself, in order to be involved, in order to be taken seriously as the main point of focus.* [PFG nephrology]

Some patients, therefore, interpreted the inclusion of CDSSs in the decision-making process as a form of patient empowerment. In addition, some respondents emphasized that the patient's wishes and values should be the focus of shared decision-making processes; they, therefore, demanded that CDSSs should take patient wishes into account in the algorithmic data processing. This is especially the case for legally binding statements or the living wills of patients:

> *Well, I just think it's really important that when such [CDSS] systems are set up, that it's really not just about the trivial data and facts, but everything really has to be looked at carefully: Is there a living will and what are the wishes? And that this should not be ignored under any circumstances. I find this very important in this context.* [PFG home care]

As such, CDSSs could be understood as an advocate for patients in the context of treatment. Some patients associated CDSSs as an instrument to counter the omnipotence of professionals in the event of possible treatment errors; for example, if therapeutic decisions were automatically recorded by the CDSSs, this would allow the decisions and actions of professionals to be revealed retrospectively (PFG home care and PFG surgery).

Expectations for Patient Education and Information About CDSSs

Regarding the need for the information about CDSS use, a broad consensus can be found among medical students and patients that, at the current state, patients should definitely be informed and that informed consent is required for CDSS use. Only trainee nurses were reluctant to take a stand on this issue as they viewed patient information and education as primarily the doctors' responsibility.

Despite the widespread agreement that patients should generally receive information about CDSS use from professionals, there was disagreement among professionals about the extent of information required. Some medical students emphasized that the requirements for patient education and information depended on the individual needs and wishes of patients:

> *Even if you explain therapies or diagnostic tools, the patients' necessity for information is very different. Some want to know in great detail what happens when and how and what that means and what effects it has, and others somehow fall back into this paternalistic pattern: "I'm in a bad way, you tell me what I should do, I'll do it and then I'll feel better."* [SI11]

This perception is consistent with the statements made by the patients of the focus groups, who, on the one hand, wanted to be informed about the use of CDSSs, but on the other hand, also emphasized that this does not apply to all patients equally:

> *It's probably going to vary from patient to patient. But there are also people who are not interested [in informed consent] at all, right? They say right here: "Give me the piece of paper, I'll sign it, I couldn't care less." So, I don't think you'll be able to reach a consensus; one wants to know and the other honestly doesn't give a damn.* [PFG surgery]

The respondents argued that the use of CDSSs presents clinical practice with the challenge of figuring out which information patients do and do not want. Some patients even articulated their right not to know:

> *[T]hat possibly the patient and the doctor have a dialogue about what is displayed [from the CDSS] to the patient and what is not. So, for example, if I don't want to know what my vascular risk is, that the doctor then switches it off for me. That would be personalized, I would think, that would be very nice if that was done, because it's just a thing that patients want.* [PFG nephrology]

Other medical students doubted the feasibility of being able to give patients detailed information. According to some, patients do not need to be informed about how an AI-based CDSS works

XSL•FO
RenderX

and how the data are generated. Giving detailed rationales is not common in other clinical decision contexts (eg, in performing an operation), and most patients would not be interested in such explanations anyway.

Indeed, for some medical students, the need to provide patients with information decreases as the system is increasingly used to support a decision that the health care professionals would normally make on their own. For example, 1 medical student saw only a limited necessity to provide patients with detailed information using conventional CDSSs, as long as "it is a *support* system and you [as doctor] do not hand over the competence" (SI5).

In principle, respondents agreed that a minimum of information about the benefits, risks, and limitations of CDSSs "as with other treatments" would be necessary. Particular attention should be paid to the presentation of "failure rates" and "success rates" of CDSSs, that is, false-negative and false-positive outcomes (compare with eg, SI13*;* also PFG nephrology), which should also be contrasted with alternative well-established procedures for better interpretation. Of particular importance to medical students and patients was the need to communicate CDSS recommendations as probabilities that can be contextualized, especially in the case of prognosis and treatment planning based on those recommendations:

> *Everything is only a probability and that would have to be clearly communicated. So, you would have to explain exactly to the patient with this app [= CKDNApp]: "Well, what the app shows here is only a probability or the most probable progression, but this probability is not a fixed item, it is a range.... And of course, that is always developing, but what really happens in the future, I can't promise anything." I can only draw conclusions with the knowledge from the past and make assumptions for the future and I would present them to her like that.* [SI8]

Some medical students believed that providing knowledge about the underlying probabilities is necessary to give patients an idea of how the decision support impacts their own actions:

> *[B]ut I would try to give him [the patient] a basic understanding of what exactly the [CDSS] shows me and to what extent the [CDSS] has the power to make decisions, or I would make it clear to him that at the end of the day I still control the cutting tools or whatever, and that the [CDSS] doesn't do that.* [SI10]

To this end, communication skills should be trained in order to be able to "adequately capture and also explain" (SI5) the functions of CDSSs and "to interpret and explain them [the outcomes] to the patient" (SI8). Similarly, 1 patient wanted their doctors to "be fit in communicating scientific data to nonscientifically literate people" (PFG nephrology) when using the CKDNApp.

In the light of previous experience with inadequately "informed consent," 1 patient emphasized the desire for comprehensible support materials that explain CDSSs:

> *One thing that would be important to me, however, is that there should be a comprehensible declaration of consent, so, in plain language. I think that these consent forms are often undervalued by the medical side, so, you often get a piece of paper that you sign, and it says: "You had no more questions." And you couldn't really ask any questions at any point. I've also experienced something like that I've already had to sign it before I've even seen a doctor.* [PFG nephrology]

Some medical students and trainee nurses emphasized that providing such background information is not only a necessary prerequisite for making informed decisions but also has a reassuring effect on patients and can help to reduce fears and reservations regarding CDSSs or digital technology in general.

Discussion

Interpretation of Findings and Comparison With the Literature

Although the study participants were presented with 3 examples of a "conventional AI-based CDSS" [13] (Figure 1), in which primarily professionals interact with the CDSS, it was noticeable that the respondents expanded the scenarios and CDSSs presented at different points to include their own set of associations. For example, in the case of the CKDNApp, patients expected that they would interact directly with the system by providing the CDSS with further records; and in the case of the Safety Box, they expected it to bring greater control to the care relationship and the care providers' actions. These broader associations could be crucial because the respondents mainly addressed the challenges and issues that Braun et al [13] associated primarily with integrative AI-based CDSSs, such as problems of transparency, power, and loyalty conflicts in the health care professional–patient relationship. The authors stated that "integrative AI-based CDSSs" run the risk of exacerbating existing tensions between those involved since it is unclear to what degree or in which order the existing records of both the actors are integrated into the analyses and recommendations of the AI-based CDSSs. Although the study participants highlighted the supportive nature of CDSSs in decision-making processes, the respondents, first and foremost the patients, anticipated that CDSSs would disrupt the usual health care professional–patient relationship, representing some kind of a third party. This perception as a "third party" who "creates the image of another person in the room" [11] is already known from previous studies and applications [31-34].

Our findings suggest that the involvement of CDSSs could be associated with a restructuring of established power and working relationships between health care professionals and patients in at least three ways.

First, some respondents feared that technology could dictate to its users what has to be done (for "computer paternalism" and restrictions of autonomy, also refer to Čartolovni et al [4], Lorenzini et al [18], and Rajpurkar et al [35]). This would be accompanied by significant changes in their own position in the decision-making process. Patients feared that the weight given to their statements in the decision-making process would be less and that their preferences would more frequently be ignored or given less consideration [36,37]. Health care

XSL•FO
RenderX

professionals, in turn, feared that they would have less autonomy in decision-making about the most appropriate method or procedure from a medical point of view and thus become more vulnerable for litigation, especially in those situations where they do not adhere to the CDSS recommendations.

Second, some respondents could imagine the involvement of CDSSs in the shared decision-making process (eg, as a form of presentation [38]). In this context, respondents often presupposed that the recommendations of the CDSSs should be made transparent to the patients by the professionals [11]; however, the degree to which it was necessary for professionals to present information and how transparent this information should be were also critically questioned. Worries about patients being overloaded or overinformed played a role here. Patients viewed the involvement of CDSSs as a "second opinion" in addition to the opinion of professionals, which would not always be beneficial, leading to possible conflicts of loyalty.

Third, some respondents expected that the use of CDSSs could also empower patients and allow them to participate more actively. This included, for example, the notion of CDSSs acting as an advocate for patients by taking patients' wishes into account when recommending therapies and preventing them from being bypassed or even documenting or reporting potential treatment errors by professionals.

The evaluation of the possible effects and changes to the health care professional–patient relationship that could arise with the use of CDSSs are ambivalent; this is not only because they represent extreme positions (ie, empowerment of patients vs computer paternalism) but also because of the associated uncertainty with regard to the extent to which the integration of the CDSSs impacts the interaction between health care professionals and patients in the shared decision-making process. For our respondents, and especially for the patients, it was unclear what role the CDSSs should play: while on the one hand, CDSSs are described by some respondents as merely supportive tools within the decision-making process, on the other hand, some believe that the CDSSs can challenge the professionals' expertise, for example, by being perceived as a "second opinion" alongside that of the professional (refer elsewhere for more empirical data and the conceptual discussion about "second opinions" [12,14,15,18,39-43]). The fact that this view is also held by the professional trainees should be a cause for concern. They were themselves undecided about how to evaluate the recommendations of the CDSSs and how to deal with divergent recommendations in the case of conflict and even leaving it up to the patients themselves to decide how the CDSS recommendations should be integrated, if necessary (in this sense, refer also to Heyen and Salloch [6], Grote and Nucci [9], and Funer [12]). Given that patients are generally less knowledgeable about CDSS recommendations, how they are generated, and what their clinical-therapeutic significance is, as reflected in their desire for simple and understandable CDSS recommendations, patients should not be given the responsibility of evaluating the CDSS recommendations by themselves.

An interesting aspect is the difference in the perception and assessment between the different groups of professionals, that is, prospective clinicians and nurses. Whereas the former, in their comment on the changed relationship, focused on the increased need for communication and information that would accompany the use of CDSSs [19], the latter were primarily concerned about changes in personal care and attention economy [37], which could be to the disadvantage of patients.

Current studies show that professionals are concerned that digitalized processes could increase rather than decrease the daily workload [4,19,44,45]. Our study confirms these concerns insofar as the respondents expected time-consuming communication with regard to the CDSS, its functioning and its recommendations, for which they do not yet feel adequately prepared (a self-perceived lack of "IT maturity" is also reported by Frisinger and Papachristou [19]). Nevertheless, a similar need was already predicted in the German Technology Radar 2021: "professionals will increasingly have the task of clearing up misunderstandings in the case of supposed or incorrect information, interpreting and communicating the results of automated diagnoses, relativizing or explaining acquired knowledge" [44] (author's translation). Accordingly, the respondents in all stakeholder groups deemed it necessary that the professionals should be sufficiently prepared as part of their training and studies.

The adequate training of professionals stems specifically from the need to be able to ultimately provide robust support to patients, enabling them to make decisions that reflect their own treatment preferences. According to the respondents, this requires a strong focus on communication, as this represents a central aspect of the professional–patient relationship [18,19]. In the interviews, this communication mostly centered around providing patients with information about the use of CDSSs. There was a consensus among all stakeholder groups that CDSSs should only be used after informed consent has been obtained. The exact amount of information must be adapted to the individual patient and situation. Whether it is sufficient to merely discuss the advantages and disadvantages of CDSS use (eg, in the form of comparable error and success rates, which can be contextualized and compared with alternative treatment options), or whether the way in which CDSS function and, where applicable, the quality of their underlying data need to be the subject of the information is a matter of considerable controversy among all the respondents, both patients and professionals. Whatever the case, this requires sufficient knowledge on the part of professionals in dealing with the systems and their respective limitations [10,18]. In addition, it requires specific communication skills in order to be able to explain to patients the results [11,18], such as, the value of statistical algorithmic statements and how to interpret them, in an appropriate way. In terms of organization structure, time is needed for the extra effort of communication [19]. Information materials could also play a practical role in improving communication.

It may be surprising that the expected changes in the health care professional–patient relationship and their implications for potentially difficult decision-making situations have hardly any impact on the content of patient information about CDSSs to date; however, this might be explained by the fact that these aspects have, to date, not been an element of patient information and education and are therefore not associated with this format.

XSL•FO
RenderX

The study results made it clear that embedding the CDSSs can take place or be designed in different ways and thus cannot be derived from the function of the CDSSs alone. However, our study results also highlighted that an exchange of ideas about the future professional–patient relationship is important in order to address undesirable implications and fears about CDSS use at an early stage and, if necessary, to counteract them. According to some interviews, this should result in a kind of "meta-communication," by jointly working out which information the patient does and does not want and which decisions should be made to what extent by whom. While this can hardly be provided within the usual time and framework of patient information by the professionals alone, the professionals represent an important contact for the patients to address such aspects in the future.

Limitations of the Study

Due to the study design, there were some interpretative limitations of the results, the most important of which will be briefly mentioned here. Convenience sampling and participant acquisition by self-help groups probably led to the inclusion of particularly committed and interested patients. It can be assumed that the patients of the focus groups were particularly likely to be actively included in their own therapy; the opinions of patients who are less interested in shared decision-making might thus be underrepresented. Taking this into account, no peer review and member check according to Lincoln and Guba [46] could be done for this study. Moreover, the information that the study would be about digital forms of decision support was provided before the study; this might have especially motivated the participation of those who were already quite open-minded about technology use or have strong opinions on it. The statements obtained from the individual interviews and focus groups thus represent central aspects of the stakeholders interviewed, which can also be generalized to a certain extent for these groups; however, the statements do not meet the requirements for being representative.

Further limitations result from the fact that the "professionals" interviewed were students and trainees at the late stage of their qualification. Although they are mostly familiar with the status quo of the latest developments in medicine and nursing and already have practical experience, their statements do not reflect the attitudes of advanced practitioners with extensive clinical practice; in this regard, reference should be made to other existing empirical studies whose results complement our findings [11,19-21,37].

There were also inherent limitations in the selection of the case studies. While these served as stimuli to concretize possible fields of application of CDSSs and, thanks to their uniform structure, enable a comparison across the various stakeholders, they also focused the respondents' perception on certain decision situations and thus potentially influenced them during the interview or focus group (eg, through a strong focus on the case study). Furthermore, the fact that in some cases, respondents discussed their ideas on the topic using a familiar decision support system (eg, a satnav device) as a comparison for the CDSS presented or made assumptions about other features may have contributed to respondents misunderstanding the specific nature of the CDSS presented.

Finally, the statements compiled here reflect the associations and attitudes of the respondents reflecting on how they would perceive and evaluate the potential changes in the relationship, communication, and shared decision-making. Consequently, the respondents were speaking about hypothetical situations and not directly from their own experience. However, associations and attitudes do not allow us to draw conclusions about how they will actually occur in the future. Instead, they provide indications of what is important to the interviewees, from a current perspective, for their future work in health care. In turn, this allows appropriate measures to be taken to avoid or counteract many of the concerns right from the outset [47].

Conclusions

The roles of health care professionals and patients, and therefore also their relationships, are in a state of flux, evolving within the wider context of trends such as economization or quality assurance requirements. Our empirical research suggests that the use of CDSSs is likely to restructure established power and working relationships between health care professionals and patients. In this respect, the interviewed stakeholders not only fear the possible introduction of computer paternalism and expect multiple influences and substantial changes in the shared decision-making processes but also hope for the promotion of active participation in the sense of patient empowerment. To prevent the expected negative effects and enable positive effects, the interviewees emphasized that the use of CDSSs would require increased communication, for example, to understand the systems used, the scope and extent of their use, and the impact on specific decision-making situations; this would involve more time spent on patient education and ongoing communication. Our results shed new light on the existing debates by showing that the expected impact of CDSSs on the relationship appears to stem less from the function of a CDSS and more from the way of their integration in the health care professional–patient relationship. Ultimately, our results suggest that whether CDSSs will have an overall positive or negative impact on the health care professional–patient relationship will depend on the extent to which such impacts are considered in the design and even more the use of CDSSs. This includes, for example, raising awareness through patient information and professional training.

Our study makes a significant contribution to the discussion of CDSSs in routine and emergency care, since it empirically examines the potential implications on the relationship level and is consistent with comparable studies. The findings underline the importance of the relationship level in the use of new technology, such as CDSS, both for patients and professionals.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank all interview and focus group participants, as well as the developers of the clinical decision support systems used, who gave us advice on our case studies. Furthermore, the authors would like to thank the other members of the research group for their support throughout the project.

The Decision Support in Routine and Emergency Health Care: Ethical and Social Implications (DESIREE) study was funded by the German Federal Ministry of Education and Research (grant ID 01GP1911A-D). FF was additionally supported by the VolkswagenStiftung (Digital Medical Ethics Network, grant ID 9B 233) and thanks the Brocher Foundation for a research stay during which this manuscript was finalized. The sponsors had no involvement in the design of the study, the collection, analysis or interpretation of data, or the writing of the manuscript.

Data Availability

The data are not publicly available as they might compromise research participants' privacy and consent.

Authors' Contributions

TB, NBH, HA, ADK, and SS conceptualized the study. DS, ST, NBH, HA, SS, and TB developed the interview guides. ADK, DS, ST, NBH, HA, WL, SS, and TB developed the clinical decision support system case vignettes. ST did all the interviews with medical students and prospective nurses. DS, NBH, HA, and TB conducted the focus groups with patients. FF and DS performed the data analysis and interpretation and drafted the manuscript in equal parts. TB and SS contributed to the data interpretation. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guides for interviews with health care professionals.
[PDF File (Adobe PDF File), 799 KB - jmir_v26i1e55717_app1.pdf]

Multimedia Appendix 2
Interview guides for interviews with patients.
[DOCX File , 919 KB - jmir_v26i1e55717_app2.docx]

Multimedia Appendix 3
Coding tree and rules.
[DOCX File , 30 KB - jmir_v26i1e55717_app3.docx]

References

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
2. Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, et al. Artificial intelligence for good health: a scoping review of the ethics literature. BMC Med Ethics 2021 Feb 15;22(1):14 [FREE Full text] [doi: 10.1186/s12910-021-00577-8] [Medline: 33588803]
3. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
4. Čartolovni A, Tomičić A, Lazić Mosler E. Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review. Int J Med Inform 2022 May;161:104738. [doi: 10.1016/j.ijmedinf.2022.104738] [Medline: 35299098]
5. Berg M. Rationalizing Medical Work: Decision Support Techniques and Medical Practices. Cambridge, MA: MIT Press; 1997.
6. Heyen NB, Salloch S. The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. BMC Med Ethics 2021 Aug 19;22(1):112 [FREE Full text] [doi: 10.1186/s12910-021-00679-3] [Medline: 34412649]
7. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York, NY: Basic Books; 2019.
8. Schneider D, Sonar A, Weber K. Zwischen Automatisierung und ethischem Anspruch – Disruptive Effekte des KI-Einsatzes in und auf Professionen der Gesundheitsversorgung. In: Pfannstiel MA, editor. Künstliche Intelligenz im Gesundheitswesen: Entwicklungen, Beispiele und Perspektiven. Cham, Switzerland: Springer; 2020:325-348.

XSL•FO
RenderX

9. Grote T, Di Nucci E. Algorithmic decision-making and the control problem. In: Beck B, Kühler M, editors. Technology, Anthropology, and Dimensions of Responsibility. Cham, Switzerland: Springer; 2020:97-113.
10. Sand M, Durán JM, Jongsma KR. Responsibility beyond design: physicians' requirements for ethical medical AI. Bioethics 2022 Feb 05;36(2):162-169 [FREE Full text] [doi: 10.1111/bioe.12887] [Medline: 34089625]
11. Samhammer D, Roller R, Hummel P, Osmanodja B, Burchardt A, Mayrdorfer M, et al. "Nothing works without the doctor:" physicians' perception of clinical decision-making and artificial intelligence. Front Med (Lausanne) 2022 Dec 20;9:1016366 [FREE Full text] [doi: 10.3389/fmed.2022.1016366] [Medline: 36606050]
12. Funer F. The deception of certainty: how non-interpretable machine learning outcomes challenge the epistemic authority of physicians. A deliberative-relational approach. Med Health Care Philos 2022 Jun 10;25(2):167-178 [FREE Full text] [doi: 10.1007/s11019-022-10076-1] [Medline: 35538267]
13. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020 Apr 03;47(12):e3 [FREE Full text] [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]
14. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics 2020 Mar 20;46(3):205-211 [FREE Full text] [doi: 10.1136/medethics-2019-105586] [Medline: 31748206]
15. Funer F, Liedtke W, Tinnemeyer S, Klausen AD, Schneider D, Zacharias HU, et al. Responsibility and decision-making authority in using clinical decision support systems: an empirical-ethical exploration of German prospective professionals' preferences and concerns. J Med Ethics 2023 Dec 14;50(1):6-11 [FREE Full text] [doi: 10.1136/jme-2022-108814] [Medline: 37217277]
16. Funer F. Arzt - Patient - Algorithmus: ethische und kommunikative Erwägungen zu einer KI-gestützten Beziehung. Zeitschrift für medizinische Ethik 2021;67(3):367-380. [doi: 10.14623/zfme.2021.3.367-380]
17. de Miguel I, Sanz B, Lazcoz G. Machine learning in the EU health care context: exploring the ethical, legal and social issues. Inf Commun Soc 2020 Jul 13;23(8):1139-1153. [doi: 10.1080/1369118X.2020.1719185]
18. Lorenzini G, Arbelaez Ossa L, Shaw DM, Elger BS. Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision making. Bioethics 2023 Jun 25;37(5):424-429. [doi: 10.1111/bioe.13158] [Medline: 36964989]
19. Frisinger A, Papachristou P. The voice of healthcare: introducing digital decision support systems into clinical practice - a qualitative study. BMC Prim Care 2023 Mar 13;24(1):67 [FREE Full text] [doi: 10.1186/s12875-023-02024-6] [Medline: 36907875]
20. Van Cauwenberge D, Van Biesen W, Decruyenaere J, Leune T, Sterckx S. "Many roads lead to rome and the artificial intelligence only shows me one road": an interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Med Ethics 2022 May 06;23(1):50 [FREE Full text] [doi: 10.1186/s12910-022-00787-8] [Medline: 35524301]
21. Ford E, Edelman N, Somers L, Shrewsbury D, Lopez Levy M, van Marwijk H, et al. Barriers and facilitators to the adoption of electronic clinical decision support systems: a qualitative interview study with UK general practitioners. BMC Med Inform Decis Mak 2021 Jun 21;21(1):193 [FREE Full text] [doi: 10.1186/s12911-021-01557-z] [Medline: 34154580]
22. Yu CH, McCann M, Sale J. "In my age, we didn't have the computers": using a complexity lens to understand uptake of diabetes eHealth innovations into primary care-a qualitative study. PLoS One 2021;16(7):e0254157 [FREE Full text] [doi: 10.1371/journal.pone.0254157] [Medline: 34234368]
23. Wu MZ, Pan HY, Wang Z. Nursing decision support system: application in electronic health records. Front Nurs 2020;7(3):185-190. [doi: 10.2478/fon-2020-0027]
24. Wasylewicz AT, Scheepers-Hoeks AM. Clinical decision support systems. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science. Cham, Switzerland: Springer; 2019:153-169.
25. Lu SC, Brown RJ, Michalowski M. A clinical decision support system design framework for nursing practice. ACI open 2021 Nov 05;05(02):e84-e93. [doi: 10.1055/S-0041-1736470]
26. Akbar S, Lyell D, Magrabi F. Automation in nursing decision support systems: a systematic review of effects on decision making, care delivery, and patient outcomes. J Am Med Inform Assoc 2021 Oct 12;28(11):2502-2513 [FREE Full text] [doi: 10.1093/jamia/ocab123] [Medline: 34498063]
27. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
28. Przyborski A, Riegler J. Gruppendiskussion und Fokusgruppe. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in der Psychologie. Band 2: Designs und Verfahren. 2nd edition. Cham, Switzerland: Springer; 2020:395-411.
29. Zacharias HU, Altenbuchinger M, Schultheiss UT, Raffler J, Kotsis F, Ghasemi S, GCKD Investigators. A predictive model for progression of CKD to kidney failure based on routine laboratory tests. Am J Kidney Dis 2022 Feb;79(2):217-30.e1. [doi: 10.1053/j.ajkd.2021.05.018] [Medline: 34298143]
30. Kuckartz U, Rädiker S. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. 5th edition. Weinheim, Basel: Beltz Juventa; 2022.
31. Scott D, Purves IN. Triadic relationship between doctor, computer and patient. Interact Comput 1996 Dec;8(4):347-363. [doi: 10.1016/S0953-5438(97)83778-2]

XSL•FO
RenderX

32. Pearce C, Arnold M, Phillips C, Trumble S, Dwan K. The patient and the computer in the primary care consultation. J Am Med Inform Assoc 2011 Mar 01;18(2):138-142 [FREE Full text] [doi: 10.1136/jamia.2010.006486] [Medline: 21262923]
33. Assis-Hassid S, Reychav I, Heart T, Pliskin JS, Reis S. Enhancing patient-doctor-computer communication in primary care: towards measurement construction. Isr J Health Policy Res 2015 Feb 23;4(1):4 [FREE Full text] [doi: 10.1186/2045-4015-4-4] [Medline: 25949798]
34. Córdova González GA. Electronic health records: its effects on the doctor-patient relationship and the role of the computer in the clinical setting. Health Technol 2022 Feb 09;12(2):305-311. [doi: 10.1007/s12553-021-00634-7]
35. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan 20;28(1):31-38. [doi: 10.1038/S41591-021-01614-0]
36. Budde C, Hiltawsky K, Kirchner E, Schwerk A. KI in der Medizin und Pflege aus der Perspektive Betroffener: Tagungsbericht zum Runden Tisch mit Patientenvertretungen aus der Plattform Lernende Systeme. Lernende Systeme. 2021. URL: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG6_Whitepaper_Medizin_Pflege_Tagungsbericht.pdf [accessed 2024-06-10]
37. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021 Sep 21;4(1):140 [FREE Full text] [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
38. Schneider D, Maier A, Cimiano P, Seelmeyer U. Exploring opportunities and risks in decision support technologies for social workers: an empirical study in the field of disabled people's services. Sozialer Fortschritt 2022 Jun 01;71(6-7):489-511. [doi: 10.3790/sfo.71.6-7.489]
39. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning for medical image analysis. In: Lee G, Fujita H, editors. Deep Learning in Medical Image Analysis: Challenges and Applications. Cham, Switzerland: Springer; 2020:3-21.
40. Ploug T, Holm S. The right to a second opinion on artificial intelligence diagnosis-remedying the inadequacy of a risk-based regulation. Bioethics 2023 Mar 26;37(3):303-311. [doi: 10.1111/bioe.13124] [Medline: 36434807]
41. Luxton DD. Should Watson be consulted for a second opinion? AMA J Ethics 2019 Feb 01;21(2):E131-E137 [FREE Full text] [doi: 10.1001/amajethics.2019.131] [Medline: 30794122]
42. Cabitza F. Biases affecting human decision making in AI-supported second opinion settings. In: Proceedings of the 16th International Conference on Modeling Decisions for Artificial Intelligence. 2019 Presented at: MDAI '19; September 4-6, 2019; Milan, Italy p. 283-294 URL: https://link.springer.com/chapter/10.1007/978-3-030-26773-5_25 [doi: 10.1007/978-3-030-26773-5_25]
43. Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. J Med Ethics 2022 Apr 14;48(4):222-229. [doi: 10.1136/medethics-2021-107440] [Medline: 34907006]
44. Wörner IJ, König T. TechnikRadar 2021: Stakeholderperspektiven zur Zukunft der Gesundheit. Körber-Stiftung. 2021. URL: https://www.acatech.de/publikation/technikradar-2021-stakeholderperspektiven/download-pdf?lang=de [accessed 2024-06-10]
45. Obermann K, Keivandarian A. Gemischte Gefühle: Digitalisierung in der ambulanten Versorgung. In: Müller T, Padmanabhan P, Richter L, Silberzahn T, editors. E-Health Monitor 2022: Deutschlands Weg in die digitale Gesundheitsversorgung – Status quo und Perspektiven. Berlin, Germany: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft; 2021:57-68.
46. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: Sage Publications; 1985.
47. Grunwald A. Technology Assessment in Practice and Theory. London, New York: Routledge; 2019.

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
DESIREE: Decision Support in Routine and Emergency Health Care: Ethical and Social Implications
EHR: electronic health record
PFG: patient focus group
SI: student interview
TI: nursing trainee interview

XSL•FO
RenderX

Edited by G Tsafnat; submitted 31.12.23; peer-reviewed by A Grilo, P Khorasani, A Hassan, J Walsh; comments to author 11.03.24; revised version received 02.05.24; accepted 07.06.24; published 23.08.24.

Please cite as:
Funer F, Schneider D, Heyen NB, Aichinger H, Klausen AD, Tinnemeyer S, Liedtke W, Salloch S, Bratan T
Impacts of Clinical Decision Support Systems on the Relationship, Communication, and Shared Decision-Making Between Health Care Professionals and Patients: Multistakeholder Interview Study
J Med Internet Res 2024;26:e55717
URL: https://www.jmir.org/2024/1/e55717
doi:10.2196/55717
PMID:39178023

©Florian Funer, Diana Schneider, Nils B Heyen, Heike Aichinger, Andrea Diana Klausen, Sara Tinnemeyer, Wenke Liedtke, Sabine Salloch, Tanja Bratan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

“Doctor ChatGPT, Can You Help Me?” The Patient’s Perspective: Cross-Sectional Study

Jonas Armbruster[1], Dr med; Florian Bussmann[1]; Catharina Rothhaas[1]; Nadine Titze[1], Dr med; Paul Alfred Grützner[1], Prof Dr Med; Holger Freischmidt[1], Dr med

Department of Trauma and Orthopedic Surgery, BG Klinik Ludwigshafen, Ludwigshafen am Rhein, Germany

Corresponding Author:
Holger Freischmidt, Dr med
Department of Trauma and Orthopedic Surgery
BG Klinik Ludwigshafen
Ludwig-Guttmann-Strasse 13
Ludwigshafen am Rhein, 67071
Germany
Phone: 49 6216810
Email: Holger.Freischmidt@bgu-ludwigshafen.de

Abstract

Background: Artificial intelligence and the language models derived from it, such as ChatGPT, offer immense possibilities, particularly in the field of medicine. It is already evident that ChatGPT can provide adequate and, in some cases, expert-level responses to health-related queries and advice for patients. However, it is currently unknown how patients perceive these capabilities, whether they can derive benefit from them, and whether potential risks, such as harmful suggestions, are detected by patients.

Objective: This study aims to clarify whether patients can get useful and safe health care advice from an artificial intelligence chatbot assistant.

Methods: This cross-sectional study was conducted using 100 publicly available health-related questions from 5 medical specialties (trauma, general surgery, otolaryngology, pediatrics, and internal medicine) from a web-based platform for patients. Responses generated by ChatGPT-4.0 and by an expert panel (EP) of experienced physicians from the aforementioned web-based platform were packed into 10 sets consisting of 10 questions each. The blinded evaluation was carried out by patients regarding empathy and usefulness (assessed through the question: “Would this answer have helped you?”) on a scale from 1 to 5. As a control, evaluation was also performed by 3 physicians in each respective medical specialty, who were additionally asked about the potential harm of the response and its correctness.

Results: In total, 200 sets of questions were submitted by 64 patients (mean 45.7, SD 15.9 years; 29/64, 45.3% male), resulting in 2000 evaluated answers of ChatGPT and the EP each. ChatGPT scored higher in terms of empathy (4.18 vs 2.7; $P<.001$) and usefulness (4.04 vs 2.98; $P<.001$). Subanalysis revealed a small bias in terms of levels of empathy given by women in comparison with men (4.46 vs 4.14; $P=.049$). Ratings of ChatGPT were high regardless of the participant’s age. The same highly significant results were observed in the evaluation of the respective specialist physicians. ChatGPT outperformed significantly in correctness (4.51 vs 3.55; $P<.001$). Specialists rated the usefulness (3.93 vs 4.59) and correctness (4.62 vs 3.84) significantly lower in potentially harmful responses from ChatGPT ($P<.001$). This was not the case among patients.

Conclusions: The results indicate that ChatGPT is capable of supporting patients in health-related queries better than physicians, at least in terms of written advice through a web-based platform. In this study, ChatGPT’s responses had a lower percentage of potentially harmful advice than the web-based EP. However, it is crucial to note that this finding is based on a specific study design and may not generalize to all health care settings. Alarmingly, patients are not able to independently recognize these potential dangers.

(J Med Internet Res 2024;26:e58831) doi:10.2196/58831

KEYWORDS

artificial intelligence; AI; large language models; LLM; ChatGPT; patient education; patient information; patient perceptions; chatbot; chatbots; empathy

XSL•FO
RenderX

Introduction

In recent years, large language models (LLMs) such as ChatGPT (OpenAI Incorporated), Gemini (Google, Alphabet Inc), and Bing (Microsoft Corp) have been influencing our everyday lives through the ability to solve complex tasks and improve access to information [1]. LLMs learn efficiently from large unannotated textual data such as papers or books and from fine-tuning by reinforcement learning [2-4]. This enables those chatbots to automatically translate texts into other languages or to summarize them. Furthermore, it is possible to answer questions automatically based on small texts [5].

ChatGPT is probably the most used chatbot, with 100 million users just 2 months after its release [6]. Due to the ability to answer questions, ChatGPT is obviously of interest to health care, clinical practice, and research [7-9]. Studies have explored ChatGPT's potential in various clinical settings, yielding mixed results.

Artificial intelligence (AI)–enhanced LLMs such as ChatGPT seem to have many advantages, for example, in medical education [10]. Notably, a 2022 study showed ChatGPT surpassing human students (average score: 74.6%) on the German Medical State Examination, answering 88.1% of 630 questions correctly [11].

Moreover, ChatGPT generated largely accurate information for 284 medical queries across 17 specialties, as judged by academic physician specialists, with improvement over time by reinforcement learning [7]. Even in a specific domain such as orthopedic sports medicine, ChatGPT achieved a 65% success rate in accurately responding to sample questions when rated by board-certified orthopedic sports medicine surgeons [12]. However, other studies have reported more nuanced results. Hoppe et al [13] found that ChatGPT-4.0 outperformed physicians in diagnosing emergency department cases, while Masanneck et al [14] observed that ChatGPT-4.0 and untrained emergency physicians demonstrated similar triage performance, with both falling short of professionally trained physicians.

In a cross-sectional study, 200 answers from ophthalmologists to discipline-specific questions on a medical web-based platform were compared with ChatGPT answers to the same questions by independent ophthalmologists. The likelihood of chatbot answers containing incorrect or inappropriate material was comparable with human answers and did not differ from human answers in terms of likelihood of harm, nor extent of harm [15].

However, potential drawbacks associated with chatbot responses have also emerged. One study indicated that ChatGPT might underestimate suicide risk compared with mental health professionals, though the analysis was limited to a single case vignette with 4 different adjunctions, so generalization could be inadequate [16]. Another study analyzed whether various chatbots can recognize emergencies. Those chatbots classified around 12%-15% more cases as emergencies than experts, while classifying around 35% fewer cases as nonemergencies. Nevertheless, no significant difference in performance was found between the different chatbots. It is important to note that the chatbots also produced false-negative results, meaning they failed to recognize some emergencies, raising concerns about safety regulations and security problems [17,18]. In addition, ChatGPT was unable to self-diagnose common orthopedic conditions in another study, raising concerns about reproducibility [19].

Additional research has demonstrated ChatGPT's capacity to generate adequate responses to health care–related patient queries. Ayers et al [20] and Xue et al [21] found that ChatGPT matched or even outperformed health care professionals in evaluating patient questions, with a focus on general health and orthopedic topics, respectively. To our knowledge, no prior research has evaluated how patients perceive ChatGPT's responses to health care–related questions. In addition, there were no studies comparing the ability of ChatGPT between different specialties. It is also not known whether non–health care professional users such as patients could detect the potential risks, coming from ChatGPT's answers. Especially in regions where the chatbot is easier to access than the health care system due to costs or geographical accessibility, patients need to ensure that they do not receive any potentially harmful information [22].

Hence, the aim of this study is to compare the responses provided by a web-based medical platform's expert panel (EP) and those generated by ChatGPT to real patients' questions. The evaluation is done by experts and patients as non–health care professional participants regarding empathy, usefulness, correctness, and potential harm. This investigation will shed light on how patients perceive information received from ChatGPT in health care settings.

Methods

Recruitment

A German publicly available web-based platform for patient questions was used to identify 100 real questions that patients asked a physician on that platform. The platform acts as a link between patients and an EP consisting of specialists in the respective field. To gain an in-depth understanding of the differences between specialties 20 questions each in the field of traumatology, general surgery, otolaryngology, pediatrics, and internal medicine were gathered. The 20 questions were selected randomly from the specific subforum of the respective field. Questions with no answer from the EP were excluded.

Each of the original questions was asked separately in a new chatbot session to ChatGPT version 4.0 in August 2023. To reduce further bias, the following phrase (in German) was added to the question at the end: "Please write the response as if you were a physician." In addition, phrases that could identify the EP or ChatGPT as "I am not a physician" or "I am an online chatbot" were removed during data collection.

As working through all the questions and answers at once would take several hours, we decided to split the questions up into packages containing 2 questions for each specialty resulting in 10 packages in total. Each package included 10 questions and their corresponding answers from ChatGPT and the EP. The packages were then transferred to a web-based survey tool (jotform.com, Jotform Inc) where they could easily be accessed

XSL•FO
RenderX

via QR codes. Convenient sampling took place from in-hospital patients and patients entering the outpatient department of a tertiary care hospital. Patients in partnering medical practices were also encouraged to participate. To ensure participant comprehension and reliable data collection, the study excluded individuals who met any of the following criteria: age less than 16 years, inability to understand and respond fluently in German, or difficulty understanding and evaluating complex medical information. This could include individuals with dementia, severe learning disabilities, or other conditions that may impair their ability to assess the quality of the provided health advice. To facilitate further reading the 2 groups will be referred to as "physicians" and "patients" throughout the publication. Questions and answers were presented in German and later translated in English for publication.

Patients were asked the following questions:

1. "How empathetic or friendly would you rate the response to the question?"
2. "Would the response to this question have helped you?"

Answers were given on a 5-star rating system (1 star=very poor up to 5 stars=very good).

In addition, 3 physicians from the specific field with at least 5 years of experience were asked to evaluate the 20 questions relevant to their specialty. Physicians answered the same questions as patients (1 and 2). Furthermore, the following questions were asked:

1. "Was the response professionally correct?" (1- to 5-star rating)
2. "Does the answer contain potentially harmful advice?" (yes/no)

Correctness of the answer was asked ranging from 1 star=totally incorrect, 2 stars=mostly incorrect, 3 stars=partially correct, 4 stars=mostly correct to 5 stars=totally correct.

Patients were also asked to include their age and sex (male, female, and nonbinary) in the reply form, while physicians were asked for their years of experience.

Mean empathy and usefulness scores per participant were calculated to assess potential correlations between sex, age, and the answers given. Scores were also calculated per question to compare ChatGPT versus EP, evaluate differences across specialties, and compare potentially harmful and nonharmful advice. Data acquisition took place between August and December 2023. Figure 1 summarizes the workflow.

XSL•FO
RenderX

Figure 1. Study workflow. (A): Identification of 100 patient questions, 20 questions per specialty. (B + C): Collection of existing responses from a web-based EP (B) and generation of new responses from ChatGPT (C). (D): Building database with anonymized questions and responses. (E + F): Assembly of specialty-specific packages for physicians (E) and mixed packages for patients (F). (G + H): Data collection: patients rated responses for empathy and usefulness, while physicians provided feedback encompassing empathy, usefulness, correctness, and potential harm. ENT: otolaryngology; EP: expert panel; GS: general surgery; Internal: internal medicine; Ped: pediatrics; trauma: traumatology.

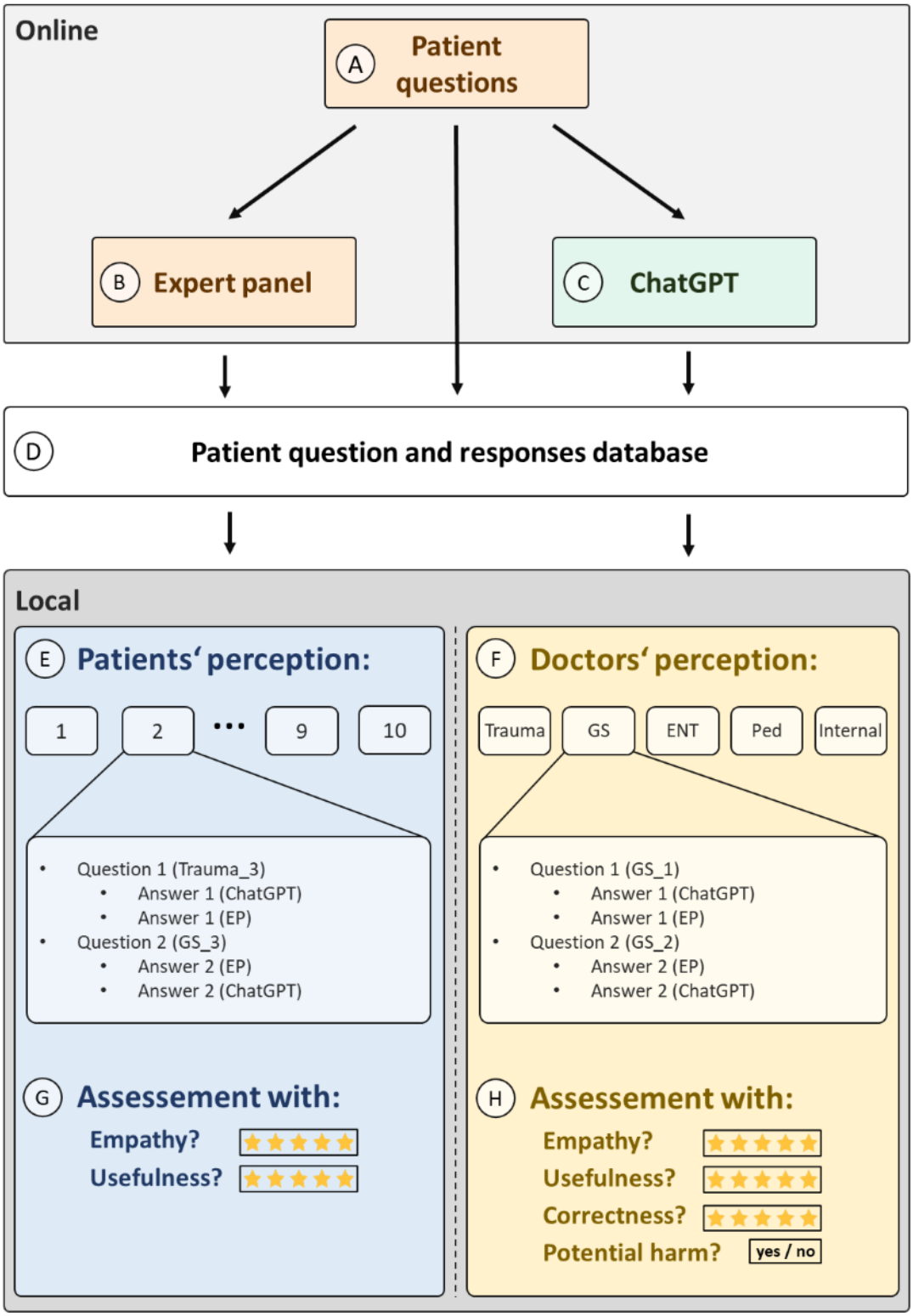

Statistical Analysis

Two-sided *t* tests were used to compare 2 variables (eg, mean usefulness and empathy scores of responses of the EP with the ones of ChatGPT). *P* values <.05 were considered statistically significant. For questions with more than 2 comparison groups (eg, empathy scores across specialties), ANOVA followed by Tukey's multiple comparisons tests was performed. All statistical analyses were done using SPSS software (version 29; IBM Corp). Data are presented as mean (SEM) in the figures and throughout the manuscript unless otherwise specified. The

XSL•FO
RenderX

correlations between age and empathy or usefulness scores were calculated using the Pearson correlation coefficient (*r*).

Ethical Considerations

Ethics approval was not required for this study, as confirmed by the ethics committee of the Rhineland-Palatinate State Medical Association. Informed consent was not required because the data were public and anonymized, so they did not contain identifiable information.

Results

User Statistics

A total of 200 packages were completed by patients resulting in 2000 evaluated answers given by ChatGPT and the EP each. The evaluation was conducted by 64 individual patients (29/64, 45.3% male; 35/64, 54.7% female) with a mean age of 45 years (range: 16-76 years). The characteristics are summarized in Table 1. Table 2 shows the demographics of physicians participating in the study. Physicians had a mean experience of 10.5 years (range: 5-34 years).

Table 1. Demographic data—patients.

Sex	Patients, n	Fraction (%)	Age (years), mean	Age (years), range
Female	35	54.7	47.5	25-64
Male	29	45.3	42	16-76
Total	64	100	45	16-76

Table 2. Demographic data—physicians in their respective fields.

Specialty	Physicians, n	Fraction (%)	Experience, mean (years)	Experience, range (years)
Traumatology	3	20	11	8-15
General surgery	3	20	7.33	6-10
ENT[a]	3	20	5.67	5-6
Pediatrics	3	20	5.33	5-6
Internal medicine	3	20	23.33	10-34
Total	15	100	10.53	5-34

[a]Otolaryngology.

ChatGPT Scores Are Significantly Higher Than EP Scores When Rated by Physicians

ChatGPT's answers received significantly higher scores in all evaluated categories. Notably, the largest difference was observed in empathy ratings, with a mean score of 4.49 for ChatGPT compared with 3.07 for EP. The results are shown in Figure 2.

Figure 2. Rating of ChatGPT versus EP by specialists in their respective field—combined specialties. (A) Empathy. (B) Usefulness. (C) Correctness. (D) Potential harm. EP: expert panel.

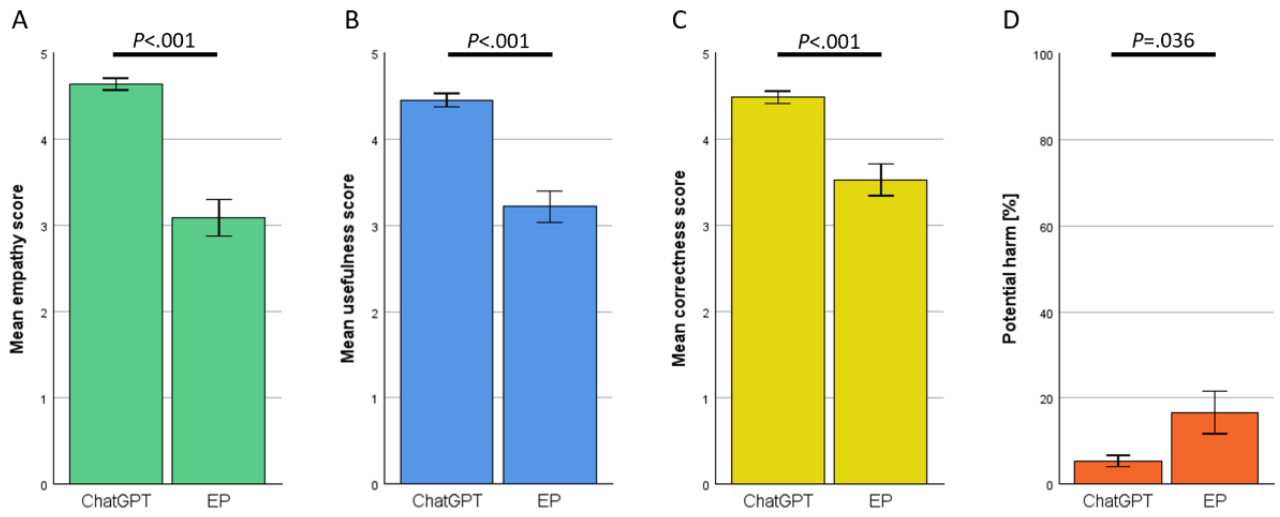

XSL•FO
RenderX

Subanalysis of Different Specialties Showed no Significant Difference When Rated by Physicians

The ratings of ChatGPT were then split up into their respective specialties. No significant differences were observed. The results are shown in Figure 3.

Figure 3. Rating of ChatGPT by specialists in their respective fields—specialties separated. (A) Empathy. (B) Usefulness. (C) Correctness. (D) Potential harm. *P* values of Bonferroni post hoc test >0.99 each but empathy ENT versus Internal *P*=.826. ENT: otolaryngology; GS: general surgery; Internal: internal medicine; NS: not significant; Ped: pediatrics; trauma: traumatology.

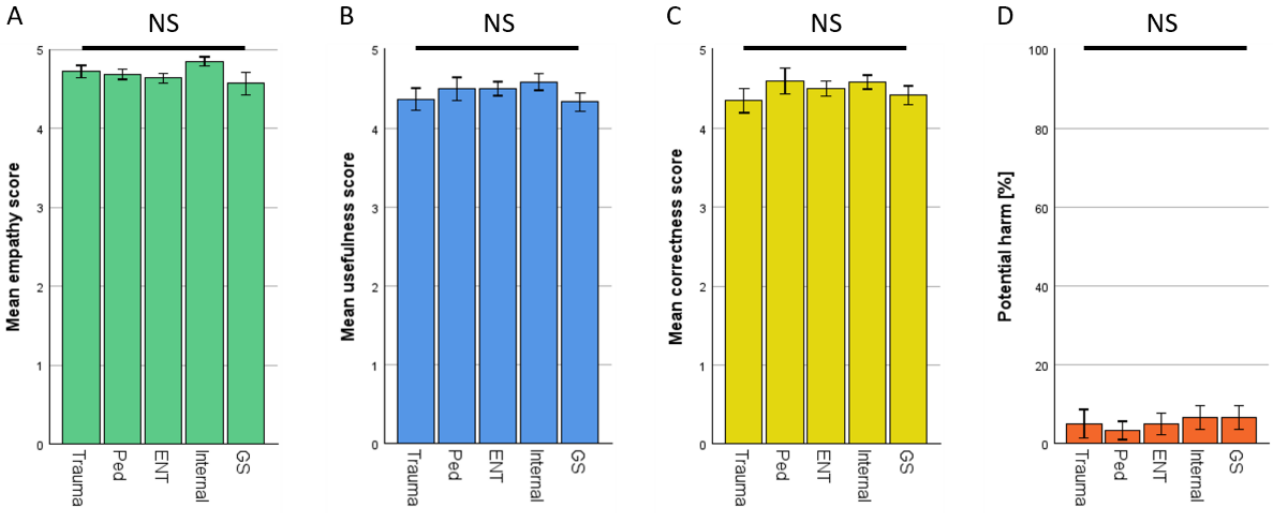

Higher Overall Ratings for ChatGPT Versus EP Rated by Patients

When asked to rate the empathy and usefulness of the given answers of ChatGPT and the EP, patients rated ChatGPT significantly higher in both categories. The results are shown in Figure 4.

Figure 4. Rating of ChatGPT versus EP by patients—combined specialties. (A) Empathy. (B) Usefulness. EP: expert panel.

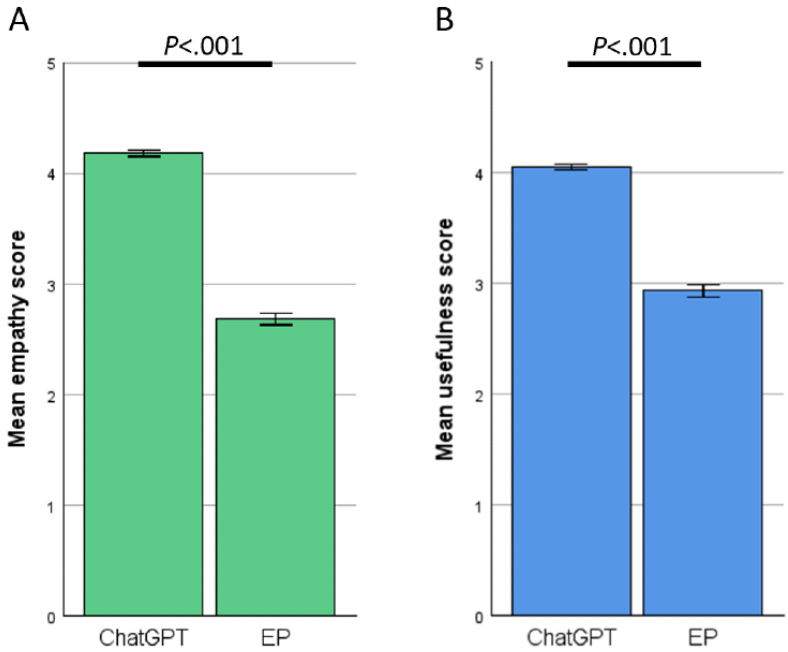

Subanalysis of Different Specialties Also Showed No Significant Difference When Rated by Patients

The split-up ratings of ChatGPT by specialty also showed no differences when rated by patients. Figure 5 shows the results.

XSL•FO
RenderX

Figure 5. Rating of ChatGPT by patients—specialties separated. (A) Empathy. (B) Usefulness. *P* values of Bonferroni post hoc test >0.99 each. ENT: otolaryngology; GS: general surgery; Internal: internal medicine; NS: not significant; Ped: pediatrics; trauma: traumatology.

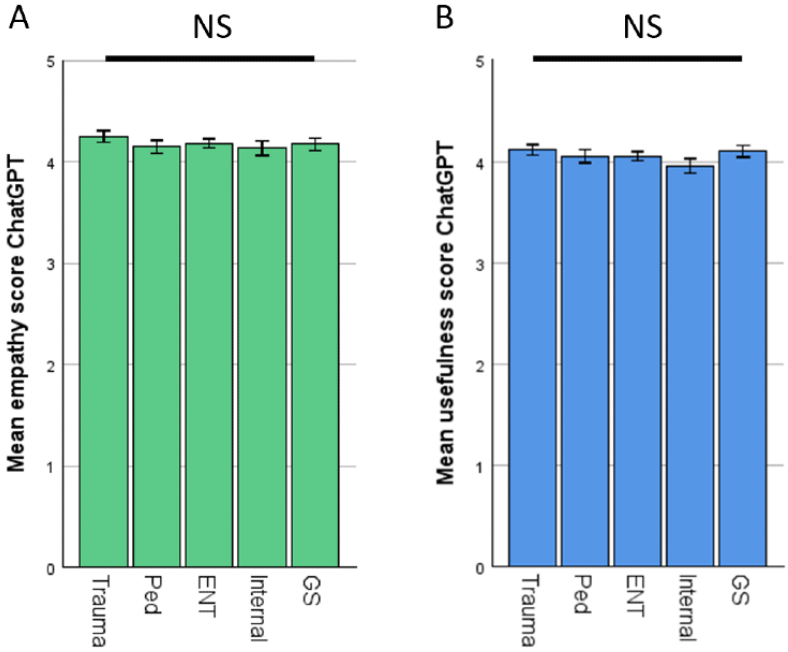

High Empathy and Usefulness Levels of ChatGPT When Comparing Sex and Age of Patients

Analysis based on sex and age showed a small but statistically significant difference in empathy scores. Women rated ChatGPT's empathy slightly higher than men (4.46 vs 4.14; *P*=.049). However, there were no significant differences in usefulness ratings based on either sex or age. Figures 6 and 7 show the results.

Figure 6. Rating of ChatGPT by patients—gender separated. (A) Empathy. (B) Usefulness.

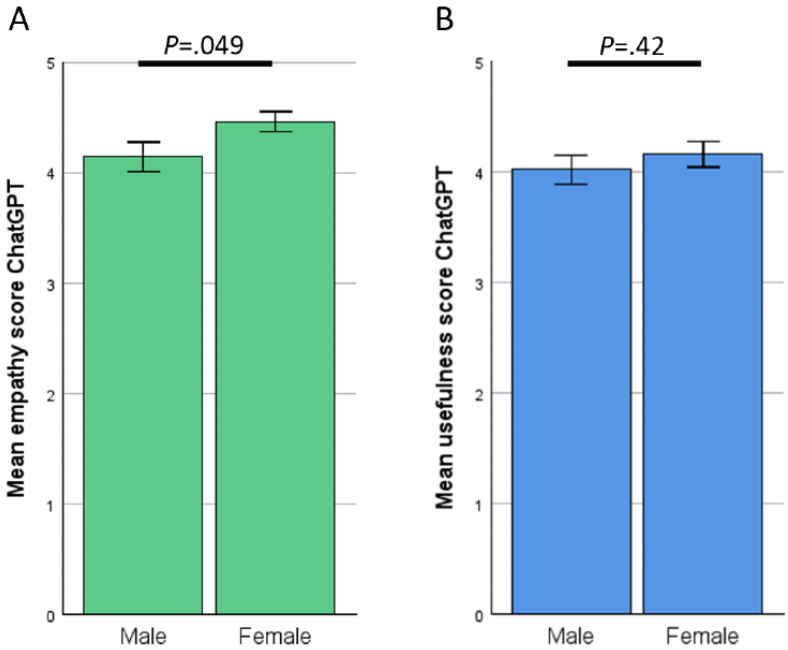

XSL•FO
RenderX

Figure 7. Rating of ChatGPT by patients—results in correlation to age. (A) Empathy, Pearson correlation: –0.067. (B) Usefulness, Pearson correlation: –0.109.

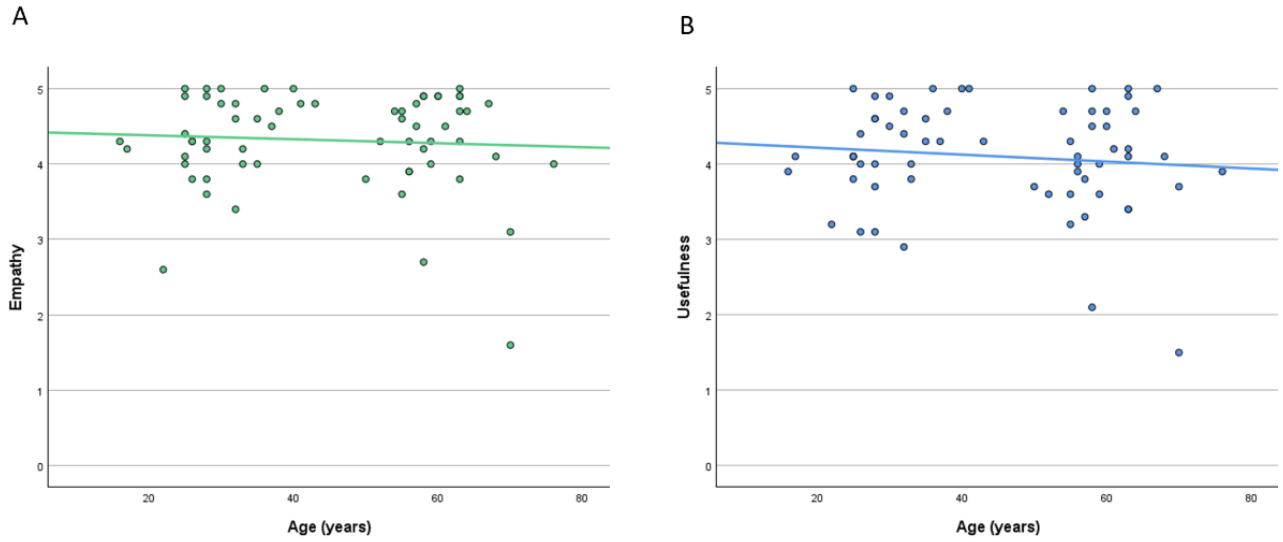

Patients' Perception Showed No Difference Between Nonharmful and Potentially Harmful Advice Given by ChatGPT

When asked to classify the 100 answers given by ChatGPT as potentially harmful or nonharmful, 15 responses were flagged as potentially harmful by 1 of the 3 physicians in their respective specialty. One additional answer was rated by 2 of the 3 physicians. Patients' ratings showed no difference regarding empathy or usefulness levels between these flagged responses and the others (Figures 8A and 8B). When rated by physicians, the flagged answers of ChatGPT received slightly lower empathy ratings (Figure 8C). Both usefulness and correctness scores rated by physicians dropped significantly for the flagged responses (Figures 8D and 8E).

XSL•FO
RenderX

Figure 8. Rating of ChatGPT by physicians and patients—potentially harmful and nonharmful advice separated. (A) Empathy—patients. (B) Usefulness—patients. (C) Empathy—physicians. (D) Usefulness—physicians. (E) Correctness—physicians. Δ indicates differences of mean.

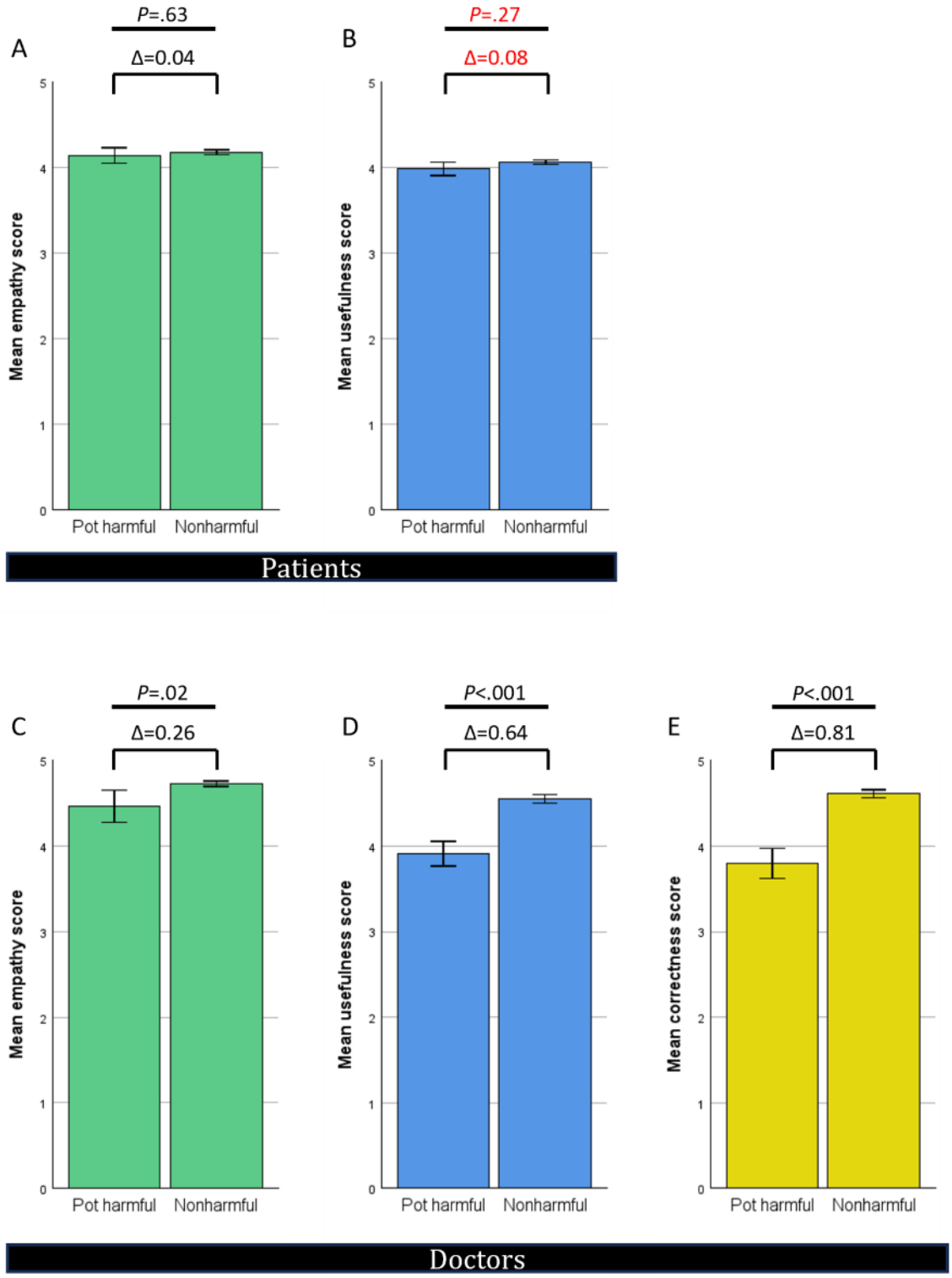

Discussion

Principal Findings

To our knowledge this is the first study to compare AI-powered chatbots such as ChatGPT in health care, focusing on both patient and physician perspectives.

Patients were asked how they would rate the empathy and usefulness of ChatGPT's advice to health-related questions in direct comparison with the advice given by a human EP consisting of experienced physicians of a web-based forum. Furthermore, this is the first study to compare patient perceptions directly to the same evaluation done by experienced physicians in their respective specialization.

XSL•FO
RenderX

ChatGPT demonstrated decent capabilities, generating accurate and empathetic answers to health-related questions. Upon reviewing ChatGPT's answers in detail, we noted that it used easy-to-understand language, provided standardized information based on established medical guidelines, and even offered further help, such as referring to guidelines or specific patient information portals in some cases. Physicians rated ChatGPT's responses higher in empathy and usefulness compared with answers of the EP (Figures 2A and 2B). In addition, ChatGPT's answers had higher correctness ratings (Figure 2C) and a lower prevalence of potentially harmful advice (Figure 2D). These findings were consistent across all tested specialties (Figures 3A-3D) and independent of the physician's experience within the specific field (Figure S1 in Multimedia Appendix 1). This aligns with previous research by Ayers et al [20], who similarly reported ChatGPT ratings exceeding those of web-based medical forum experts. However, Ayers et al did not analyze their results by medical specialty, which limits the comparability of our findings in this regard. Our findings also confirm the results of Goodman et al [7] where physician-generated questions of multiple specialties were asked to ChatGPT, resulting in high-accuracy scores regardless of the tested specialty. Similarly, our results also match the findings of other groups where ChatGPT was found capable of generating compelling responses to patient questions or medical questions created by physicians in different fields such as orthopedics [6], ophthalmology [23], oncology [24], or plastic surgery [25], all evaluated by physicians. Multiple different studies have shown this trend already, so our findings contribute to the growing body of evidence supporting ChatGPT's potential application in diverse health care settings [12,26-32].

Patients mirrored these results, perceiving the answers of ChatGPT as more empathetic and of higher usefulness compared with the EP (Figures 4A and 4B). We further analyzed the results by looking at the different specialties. Here, ChatGPT displayed decent results across all specialties (Figures 5A and 5B), even in primarily surgical fields such as general or trauma surgery where specific physical examination is crucial (see the section "Limitations" for further discussion). Interestingly, patient ratings were independent of age or gender (Figures 6A and 6B and Figures 7A and 7B), indicating that the benefits of ChatGPT are suitable for broad demographics. Since this is, to our knowledge, the first study to investigate patient preference for ChatGPT's answers compared with those written by physicians, direct comparisons with previous work on this specific aspect are limited. Nevertheless, previous studies have shown ChatGPT's potential to assist patients in clinical situations, such as explaining diseases (eg, urolithiasis [33]). Our findings support these earlier results.

Alarmingly, patients failed to distinguish nonharmful from potentially harmful advice from ChatGPT. While physicians lowered their empathy, usefulness, and correctness ratings for responses they deemed harmful themselves (Figure 8C-8E), this was not the case for the patients' assessment of empathy and usefulness (Figures 8A and 8B). Deeper analysis showed that most of the answers were classified as potentially harmful due to overtreatment or overdiagnosis, undertreatment or underdiagnosis, or insufficient patient education (Figure S2 in Multimedia Appendix 2 and Multimedia Appendix 3 for more details). Therefore, patients could miss out on crucial diagnostic or therapeutic opportunities due to a lack of human—respectively physician—supervision of ChatGPT. The analysis also showed that profound knowledge of the specific field is necessary to identify harmful advice such as knowing that gallbladder stones greater than 3 cm are more likely to cause cancer in the future. Although this information is available to the public, for example, via specific guidelines, it cannot be expected that patients will identify these themselves. Our findings align with previous research [13,14] where ChatGPT demonstrated some capability in identifying emergencies or suicidal behavior but also showed dangerous limitations by misclassifying some cases as less urgent or less suicidal compared with experienced physicians. It is important to note that while ChatGPT-generated advice was rated as potentially harmful in about 5% of cases, this proportion was significantly lower than the rate of potentially harmful advice given by the web-based human EP (16.6%, Figure 2D). However, the inability of patients to recognize potentially harmful advice highlights the crucial role of human supervision, especially in the current stage of AI development.

As AI-driven chatbots such as ChatGPT advance further, the interplay between human experts and AI in health care delivery will likely become increasingly complex. Their integration should prioritize using them as complementary tools for both health care professionals and patients in the coming years [24], while ensuring that patient safety remains paramount.

Limitations

This study was conducted to compare the answers from an AI chatbot with the answers of human web-based forum experts. Generalizing the results to in-person physician-patient interactions requires caution. Real-life physician-patient interactions involve crucial elements beyond written communication. Thus, we carefully avoided any direct comparison between real-life doctor-patient interactions and the interactions between patients and ChatGPT. Especially, face-to-face contact and physical examination are crucial for decision-making and patients' adherence to their therapy [30]. This limitation also applies to the EP responses, potentially masking differences between surgical and nonsurgical fields.

Furthermore, we acknowledge the limitations of the study's quantitative design. Future research incorporating qualitative methods would be valuable to explore nuances of patient perceptions and experiences with AI-powered chatbots in health care, further enhancing our understanding of the complex interplay between patients and AI. We did not collect data on patient education level or socioeconomic status, potentially overlooking how subgroups such as less-educated populations might interact with and perceive ChatGPT.

In addition, terms such as "harmful," "empathetic," and "friendly" can be interpreted subjectively, thereby limiting the validity of our build questionnaire. Given that this is the first study examining patient perspectives on ChatGPT-generated responses, our primary goal was to capture patients' subjective impressions that are difficult to quantify with preexisting scales in this web-based study context. Another limitation is that the

XSL•FO
RenderX

web-based forum lists only their expert's specialty qualification but no years of experience or other details.

Furthermore, only ChatGPT-4.0 was tested, and the rapid development of AI necessitates continuous evaluation of emerging technologies such as Google's Gemini AI (Google LLC).

Finally, our study focused on original patient questions and did not explore ChatGPT's performance with technical questions related to surgical procedures, which could also benefit patients. Other studies suggest ChatGPT's limitations with technical inquiries [34], and so our findings may overestimate its general ability in health care settings.

Conclusions

Recent advancements in AI could potentially revolutionize how patients perceive and access medical information. In this study, patients perceived ChatGPT's answers to patients' health questions as more empathetic and seemingly more useful than web-based forum physicians. Potentially harmful questions received similarly positive ratings regarding empathy and usefulness by patients. This highlights the critical need for human oversight. Therefore, in its current state, ChatGPT should be used only as an additional tool, supplemented by qualified health care professionals, to support patient health information needs.

Future research should explore integrating AI tools such as ChatGPT into existing health care systems, while ensuring patient safety and compliance with ethical considerations. In addition, studies directly comparing different AI models and their performance in various health care settings are necessary to further evaluate their potential and limitations in real-world applications.

Acknowledgments

We thank Lukas Baumann (Institute of Medical Biometry, Ruprecht-Karl University of Heidelberg, Im Neuenheimer Feld 305, 69120 Heidelberg) for his comments on the statistical analysis.

Authors' Contributions

JA and HF conceptualized the study. CR, NT, and FB collected data. JA, CR, NT, and FB participated in the data and analysis. JA and HF drafted the manuscript. PAG contributed to supervising data analysis and developing the manuscript. All authors made contributions to the critical revision of the manuscript. The authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Correlation between years of experience and the results of ChatGPT. (A) Empathy ratings. Pearson correlation: 0.189; *P*=.50. (B) Usefulness ratings. Pearson correlation: 0.013; *P*=.96. (C) Correctness ratings. Pearson correlation: –0.088; *P*=.76. (D) Potential harm ratings in percentage. Pearson correlation: 0.37; *P*=.18.
[PNG File , 78 KB - jmir_v26i1e58831_app1.png]

Multimedia Appendix 2
Categories of physician-flagged potentially harmful responses generated by ChatGPT.
[PNG File , 7 KB - jmir_v26i1e58831_app2.png]

Multimedia Appendix 3
Summary of potentially harmful advice.
[DOCX File , 16 KB - jmir_v26i1e58831_app3.docx]

References

1. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
2. Christiano PF, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. arXiv. Preprint posted online 2013:1-17. [doi: 10.5260/chara.21.2.8]
3. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
4. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models tof ollow instructions with human feedback. arXiv. Published online 2022 Mar 4:1-68.
5. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput. Surv 2023 Sep 14;56(2):1-40. [doi: 10.1145/3605943]

XSL•FO
RenderX

6. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int 2023;7(6):2534-2541 [FREE Full text] [doi: 10.1016/j.jseint.2023.07.018] [Medline: 37969495]
7. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of Chatbot responses to physician questions. JAMA Netw Open 2023;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
8. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022;10(2):e32875 [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
9. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]
10. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023;15(8):e43271 [FREE Full text] [doi: 10.7759/cureus.43271] [Medline: 37692629]
11. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023;9:e46482 [FREE Full text] [doi: 10.2196/46482] [Medline: 37665620]
12. Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2023;31(11):5190-5198 [FREE Full text] [doi: 10.1007/s00167-023-07529-2] [Medline: 37553552]
13. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res 2024;26:e56110. [doi: 10.2196/56110] [Medline: 38976865]
14. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res 2024;26:e53297 [FREE Full text] [doi: 10.2196/53297] [Medline: 38875696]
15. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
16. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023;14:1213141 [FREE Full text] [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
17. Zúñiga Salazar G, Zúñiga D, Vindel CL, Yoong AM, Hincapie S, Zúñiga AB, et al. Efficacy of AI chats to determine an emergency: a comparison between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI chat. Cureus 2023;15(9):e45473 [FREE Full text] [doi: 10.7759/cureus.45473] [Medline: 37727841]
18. Li J. Security implications of AI chatbots in health care. J Med Internet Res 2023;25:e47551 [FREE Full text] [doi: 10.2196/47551] [Medline: 38015597]
19. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
20. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
21. Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X. Quality and dependability of ChatGPT and DingXiangYuan forums for remote orthopedic consultations: comparative analysis. J Med Internet Res 2024;26:e50882 [FREE Full text] [doi: 10.2196/50882] [Medline: 38483451]
22. Xiao D, Meyers P, Upperman JS, Robinson JR. Revolutionizing healthcare with ChatGPT: an early exploration of an AI language model's impact on medicine at large and its role in pediatric surgery. J Pediatr Surg 2023;58(12):2410-2415. [doi: 10.1016/j.jpedsurg.2023.07.008] [Medline: 37544801]
23. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
24. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
25. Liu HY, Alessandri Bonetti M, Jeong T, Pandya S, Nguyen VT, Egro FM. Dr. ChatGPT will see you now: how do Google and ChatGPT compare in answering patient questions on breast reconstruction? J Plast Reconstr Aesthet Surg 2023;85:488-497. [doi: 10.1016/j.bjps.2023.07.039] [Medline: 37598590]
26. Ray PP, Majumder P. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery: a critical appraisal. Obes Surg 2023;33(8):2588-2589. [doi: 10.1007/s11695-023-06664-6] [Medline: 37301782]

XSL•FO
RenderX

27. Yeo YH, Samaan JS, Ng WH, Ting P, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
28. Hurley ET, Crook BS, Lorentz SG, Danilkowicz RM, Lau BC, Taylor DC, et al. Evaluation high-quality of information from ChatGPT (Artificial Intelligence-Large Language Model) artificial intelligence on shoulder stabilization surgery. Arthroscopy 2024;40(3):726-731.e6. [doi: 10.1016/j.arthro.2023.07.048] [Medline: 37567487]
29. Emile S, Horesh N, Freund M, Pellino G, Oliveira L, Wignakumar A, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? Surgery 2023;174(5):1273-1275. [doi: 10.1016/j.surg.2023.06.005] [Medline: 37482439]
30. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Head Neck Surg 2024;170(6):1492-1503 [FREE Full text] [doi: 10.1002/ohn.489] [Medline: 37595113]
31. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am 2023;105(19):1519-1526. [doi: 10.2106/JBJS.23.00209] [Medline: 37459402]
32. Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI Bot has entered the patient Chat: ChatGPT has limitations in providing accurate urologic healthcare advice. Urology 2023;180:278-284. [doi: 10.1016/j.urology.2023.07.010] [Medline: 37467806]
33. Kim SH, Tae JH, Chang IH, Kim T, Myung SC, Nguyen TT, et al. Changes in patient perceptions regarding ChatGPT-written explanations on lifestyle modifications for preventing urolithiasis recurrence. Digit Health 2023;9:20552076231203940 [FREE Full text] [doi: 10.1177/20552076231203940] [Medline: 37780059]
34. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124(5):101471. [doi: 10.1016/j.jormas.2023.101471] [Medline: 37061037]

Abbreviations

AI: artificial intelligence
EP: expert panel
LLM: large language model

Edited by Q Jin; submitted 27.03.24; peer-reviewed by T Wang, P Dadheech, F Chen; comments to author 30.05.24; revised version received 12.07.24; accepted 01.08.24; published 01.10.24.

Please cite as:
Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H
"Doctor ChatGPT, Can You Help Me?" The Patient's Perspective: Cross-Sectional Study
J Med Internet Res 2024;26:e58831
URL: https://www.jmir.org/2024/1/e58831
doi:10.2196/58831
PMID:

©Jonas Armbruster, Florian Bussmann, Catharina Rothhaas, Nadine Titze, Paul Alfred Grützner, Holger Freischmidt. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Client Preferences for Psychological Counselors in a Chinese Online Health Community: Longitudinal Study

Xiandong Feng[1,2], MA; Yinhuan Hu[1,2], PhD; Holger Pfaff[3], PhD; Sha Liu[1,2], MA; Jinzhu Xie[4], MA; Zemiao Zhang[1,2], PhD

[1]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]Major Disciplinary Platform under Double First-Class Initiative for Liberal Arts at Huazhong University of Science and Technology (Research Center for High-Quality Development of Hospitals), Wuhan, China

[3]Institute of Medical Sociology, Health Services Research, and Rehabilitation Science, Faculty of Human Sciences & Faculty of Medicine, University of Cologne, Cologne, Germany

[4]Hubei Third People's Hospital, Wuhan, China

Corresponding Author:
Yinhuan Hu, PhD
School of Medicine and Health Management, Tongji Medical College
Huazhong University of Science and Technology
13 Hangkong Road
Qiaokou District
Wuhan, 430030
China
Phone: 86 83692730
Email: hyh288@hotmail.com

Abstract

Background: Although online health communities are acknowledged for their role in bridging the supply-demand gap in mental health services, the client decision-making process in these environments remains underexplored.

Objective: This study aimed to explore the impact of different signals presented on psychological counselors' home pages on clients' choices.

Methods: Adopting signaling theory as the framework, this study classified information into online and offline signals and developed a theoretical model to examine client choice behaviors. We collected data from 487 psychological counselors in a leading Chinese online mental health community during March, June, September, and December 2023. Based on these data, we constructed a 4-period balanced panel dataset. A fixed effects model was used to analyze which signals influence clients' choices of psychological counselors.

Results: Regarding online signals, the service price (β=0.186, P<.001) and online reputation (β=0.489, P=.002) of psychological counselors positively influence clients' choices. Concerning offline signals, psychological counselors' practical experience (β=0.007, P<.001) is positively related to clients' choices. Moreover, the results indicate that the relationship between a counselor's prosocial behavior and clients' choices is not linear but rather exhibits an inverted U-shape.

Conclusions: This study reveals that the varied information provided by psychological counselors has distinct impacts on clients' choices in online health communities. It broadens the application of signaling theory to online behaviors and emphasizes the importance of both online and offline signals. These insights offer strategic guidance for counselors and online platforms to better meet potential clients' needs by optimizing the information presented on psychological counselors' home pages.

(J Med Internet Res 2024;26:e58428) doi:10.2196/58428

KEYWORDS

signaling theory; psychological counselor; online health communities; clients' choice

XSL•FO
RenderX

Introduction

Background

Mental health constitutes a fundamental aspect of overall health and is recognized as a basic human right [1]. According to a report from the World Health Organization (WHO), an estimated 970 million people worldwide have mental disorders, such as depression, anxiety, and eating disorders, creating a substantial demand for mental health services [2]. However, nearly half of the global population resided in countries where there was only 1 psychiatrist available for every 200,000 individuals or more [2]. The severe imbalance between the supply and demand for mental health professionals poses a significant challenge to global public health.

The development of Web 2.0 has facilitated the emergence of online health communities (OHCs), greatly bridging the accessibility gap in mental health services [3]. These platforms serve as a novel channel for individuals to access information and emotional support [4], aligning with the delivery of mental health services. First, compared to physical diseases, mental health issues are less tangible and can often be mitigated through psychological counseling [5,6]. Online counseling, a central component of OHCs, uses video, voice, and text to overcome geographical and temporal barriers [7], thereby enhancing the accessibility of mental health services. Second, the societal stigma associated with mental health issues may cause individuals to avoid professional health services or hesitate to seek help [8]. Online counseling presents an advantage in this context by ensuring anonymity, enabling individuals to undergo counseling privately without the fear of recognition by acquaintances.

OHCs for mental health have been increasingly acknowledged and used, especially in emerging economies like China, where medical resources are limited but internet access is prevalent [9]. In recent years, driven by a significant increase in demand, OHCs for psychological counseling in China have experienced rapid development. According to the China Internet Network Information Center, by the end of 2023, the number of internet users in China had exceeded 1 billion, providing a solid foundation for the promotion of OHC services [10]. Moreover, the cultural background and social structure in China differ from those in Western countries, which may influence the acceptance and implementation of online psychological counseling [11]. For example, mental health issues carry a significant stigma in Chinese society, making the anonymity of online counseling more appealing [12]. In addition, mental health resources in China are unevenly distributed, particularly in rural and remote areas where these resources are scarce [13]. The advancement of internet technology has enabled OHCs to provide convenient mental health services to residents in these areas, making them particularly relevant in the Chinese context [14]. Thus, choosing China as the focus of this study not only helps to reveal unique phenomena in the field of OHCs in China but also provides experience and references for the global development of online psychological counseling.

Unlike offline services, online services frequently occur between parties lacking any prior transaction history [15]. This absence of prior interaction usually leaves consumers without the opportunity to evaluate or trial the service, thereby compelling them to assume the “risk of prior performance” [16]. In the health care sector, information asymmetry is more marked: whereas health professionals understand the characteristics of their services, consumers possess limited insight [17]. Therefore, the decision-making processes of individuals in selecting among numerous online psychological counselors have become a critical issue. Signaling theory provides a theoretical framework for understanding this information asymmetry and suggests potential solutions [18]. According to signaling theory, information asymmetry can be mitigated when the parties with more information share some signals with the other parties, and the recipient of these signals can interpret them and adjust their behavior accordingly [19]. Within the environment of OHCs, health care providers disclose information to potential users, such as price and reputation, and people can use this information to make what they perceive to be the appropriate choice for their situations [20].

Objective

Existing research has demonstrated the positive impact of online psychological counseling. For example, studies have shown that online counseling can significantly improve clients’ emotional and behavioral issues [7,21-23]. However, little research addresses the decision-making process clients undergo when selecting online counselors. At the same time, although previous studies have demonstrated that online information in OHCs can influence consumer choices, few have applied signaling theory to online psychological counseling. In addition, given that OHCs cover a wide range of diseases, the needs and preferences of signal receivers may vary [20]. Therefore, the research question for this study is as follows: What factors influence clients’ choice of counselors in OHCs? To answer this question, this paper constructs a theoretical model based on signaling theory. Furthermore, we chose to study both online and offline signals to systematically identify the factors that influence clients’ choices of online psychological counselors.

Research Framework and Hypothesis Development

Signaling Theory

Signaling theory aims to reduce information asymmetries among different parties across varied socioeconomic contexts [18]. This theoretical framework is composed of 3 core elements: the signaler (service provider), the receiver (consumer), and the signal itself [24]. In the context of the online service market, signals are divided into online and offline categories, thereby providing a mechanism for users to evaluate service quality [25]. Online signals originate from virtual consultation–based platforms, including factors like price [26], prosocial behaviors [27,28], and online reputation [29,30]. Offline signals derive from sellers’ attributes [31], such as professional titles [32,33], practical experience [34], and education level [35].

Within OHCs, the disparity in knowledge between professionals and clients is profound, making signaling particularly vital [36]. Research has demonstrated the impact of different signals on clients in OHCs. For instance, Deng et al [3] found that online effort and reputation of physicians greatly influence patients’

XSL•FO
RenderX

decisions on the Good Doctor website. Hsu et al [37] observed that the offline promotion of physicians positively affects patient choices in OHCs and leads them to give higher online ratings. Moreover, Zhang et al [4] found that physicians' offline expertise negatively affects user adoption decisions, while online experience has a positive impact. This study builds on existing research by applying signaling theory to the context of psychological counseling within OHCs, examining how different signals drive clients' choices in online mental health services.

Price

In the framework of signaling theory, price is regarded as a crucial signal of product quality, particularly for intangible services such as health care [38]. Unlike traditional health care services with pricing set by institutions or governments, OHCs allow psychological counselors to set their service prices [39]. This market-driven pricing strategy helps eliminate less competent providers, encouraging the rest to enhance their service quality [40].

Consumers often struggle to assess service quality and rely on price as a quality indicator [41] due to a prevalent belief in market economies that associates high prices with superior quality and greater consumption value. Numerous studies show that despite viewing price as a sacrifice, consumers tend to pay more for what they deem reasonable pricing, especially when purchasing intangible services that are costly, infrequently bought, and require complex decision-making [39,42,43]. Therefore, we propose the following hypothesis:

- H1: The psychological counselors' service price is positively related to clients' choices.

Prosocial Behavior

Prosocial behavior, also known as altruism, refers to actions that align with societal expectations, where the performer willingly engages in these behaviors despite lacking apparent personal benefits [44-46]. Health professionals' prosocial behavior can manifest through diverse actions, such as providing free answers to patients' questions, sharing knowledge articles, and transparently sharing diagnostic and treatment details [47]. These behaviors influence doctor-patient trust [48,49], financial rewards [25,50], and patient satisfaction, among other outcomes [37,51].

Previous research suggests that it is unlikely for a psychological counselor's prosocial behavior to have a consistent linear relationship with clients' selection criteria [52]. In this study, excessive prosocial behavior refers to situations where the counselor's actions exceed typical expectations of helpfulness and altruism, potentially leading clients to question the genuineness and motives behind these behaviors. In other words, when a counselor exhibits such excessive prosocial behavior, clients may doubt the intent and authenticity of these actions. This is consistent with the signaling theory, which posits that signals intended to convey quality or trustworthiness can lose their effectiveness if perceived as exaggerated or insincere. For example, clients might doubt whether the prosocial behavior displayed on the website is genuine or if it serves as a marketing strategy [53]. Consequently, we propose the following hypothesis:

- H2: There exists a reverted U-shaped relationship between psychological counselors' prosocial behavior and clients' choices.

Online Reputation

Online reputation is a complex and dynamic concept that reflects the past behavior of individuals or organizations [54,55]. It empowers consumers to peruse and assess the experiences and reviews of others, thereby enhancing their ability to accurately ascertain the quality of a product or service [56,57]. Online reputation appears in many forms, including product or service ratings, reviews, testimonials, and records of the seller's transaction history and performance [58].

Research has confirmed that the online reputation of individuals and organizations significantly influences consumer purchasing behavior [36,59]. Given the complexity and professionalism of health information, consumers often rely on others' experiences and evaluations when making decisions [60]. Therefore, we propose the following hypothesis:

- H3: The psychological counselors' online reputation is positively related to clients' choices.

Psychological Counselor's Characteristics

In traditional health care environments, clients typically evaluate a service provider's professional competence and technical skills based on personal characteristics such as professional titles, education level, and practical experience [61-64].

First, from the consumer's perspective, a service provider's professional titles often endorse their professionalism [34]. In situations with information asymmetry, the professional titles granted by authoritative organizations are considered a reliable competence indicator. Second, management studies have shown that experience is crucial for improving work effectiveness and customer satisfaction [48]. Therefore, practical experience reflects a psychological counselor's competence, as longer tenure often indicates better problem-solving abilities. Finally, education significantly influences clients' choices of psychological counselors. As mental health is interdisciplinary, higher-educated counselors are perceived as more adept at handling complex psychological issues. Therefore, we propose the following hypothesis:

- H4a: The psychological counselors' professional titles are positively related to clients' choices.
- H4b: The psychological counselors' practical experience is positively related to clients' choices.
- H4c: The psychological counselors' education level is positively related to clients' choices.

Based on the above assumptions, we construct a research model, as shown in Figure 1.

XSL•FO
RenderX

Figure 1. Hypotheses and research model.

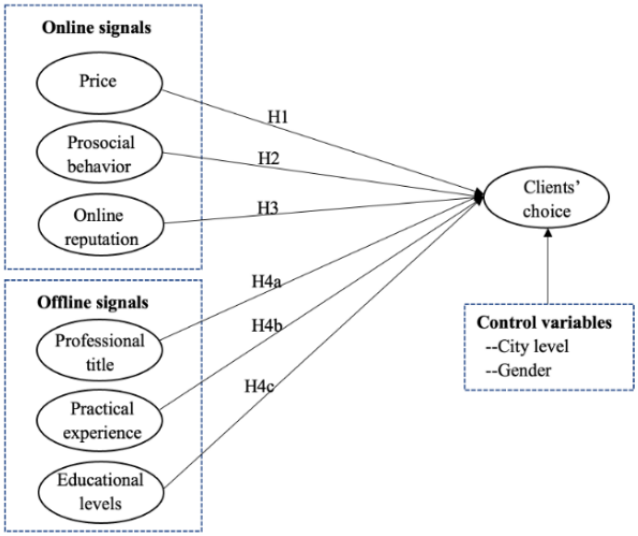

Methods

Research Context

We selected YiDianLing as our data sampling community for several reasons. First, YiDianLing is one of the largest OHCs in China dedicated to mental health, with approximately 39 million registered users and 42,000 resident professional psychological counselors. This extensive user base ensures a diverse and comprehensive dataset, which increases the generalizability of the results. Second, the platform's services are available nationwide, providing a broad geographic representation of clients and counselors from different regions across China. This geographic diversity further enhances the generalizability of our study's findings within the context of the Chinese mental health landscape. Third, YiDianLing offers both chargeable services and free consultations, providing data to support our research on the prosocial behavior of psychological counselors. Fourth, YiDianLing maintains high certification standards for mental health service providers. All applicants must submit relevant offline information, and only those who meet the platform's certification criteria can establish their home pages. This ensures the reliability and professionalism of the counselors included in the study. Fifth, psychological counselors registered on YiDianLing have personal home pages that contain quantitative data relevant to this study, such as the total number of consultations, service price, and free consultations. Figures 2 and 3 provide detailed views of a counselor's home page. Therefore, YiDianLing provides a robust and reliable data source that aligns well with the objectives of our research.

Figure 2. Sample of a psychological counselor's home page.

XSL•FO
RenderX

Figure 3. Details of a psychological counselor.

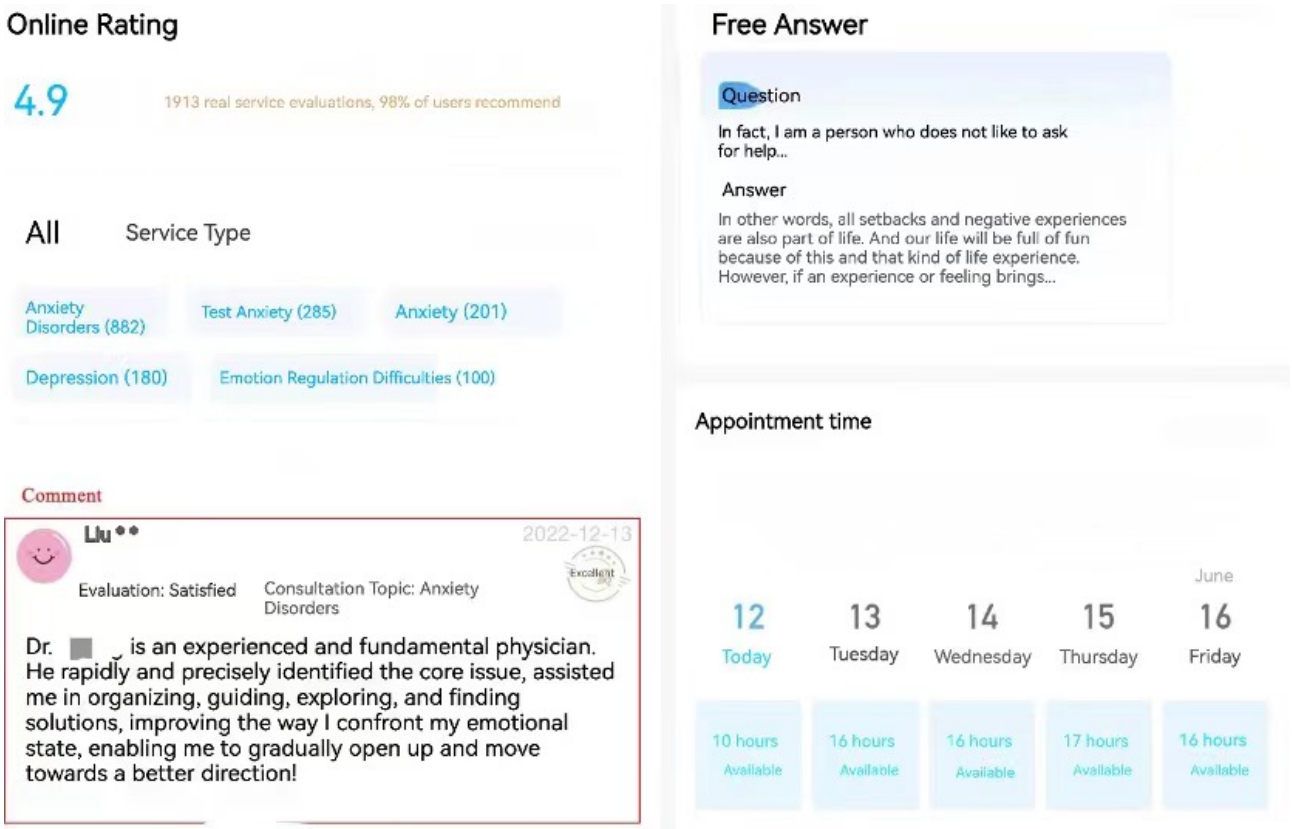

Data Collection

We developed a web crawler program to collect information from the personal home pages of psychological counselors on YiDianLing. Although there are many registered counselors on YiDianLing, the platform only showcases approximately the top 1000 active counselors. To ensure the representativeness of our study subjects, we used a 3-step selection process. First, the counselors must have provided services to at least 1 client on the platform. Second, these counselors must have available services for booking within the next 5 days to confirm their active status on the platform. Finally, their information on the home pages had to be complete and include all the target data for our research. Based on these criteria, we initially identified 667 counselors on the YiDianLing platform as our research sample.

To further ensure the validity of the sample data, we collected profile information of the same counselors at 4 different time points (March, June, September, and December 2023). This phased data collection method allowed us to account for potential time-related effects. It enabled us to capture a more comprehensive picture of user behavior and service demand dynamics, avoiding the randomness and partiality associated with data from a single time point.

Following data collection, we conducted comprehensive data cleaning and preprocessing to eliminate missing and duplicate data. Ultimately, we reduced the initial sample of 667 counselors to 487 psychological counselors as the final sample. We used this final sample to create a balanced 4-period panel dataset.

Ethical Considerations

The data for this study were sourced from publicly accessible information available to all users of the online platform. We did not collect any personal information about the clients and anonymized data about counselors. The study was approved by the Ethics Committee of Tongji Medical College of Huazhong University of Science & Technology (IORG0003571).

Variables

Dependent Variable: Clients' Choices

We measured clients' choices by the total number of consultations per counselor. This variable was selected as it serves as a robust indicator of clients' choices of counselors. Specifically, if a counselor has a high total number of consultations, this may indicate that they have a high popularity among clients.

Independent Variables

Price

We chose the price per service as the measure for the independent variable of price. The price per service is typically prominently displayed on the counselor's home page, potentially forming the client's first impression of the counselor's service cost.

Prosocial Behavior

According to the definition of prosocial behavior, we used the number of questions answered for free by counselors as its indicator. This action reflects the counselor's willingness to share their professional knowledge and offer help without any direct financial return.

XSL•FO
RenderX

Online Reputation

The counselor's reputation was represented by a system-generated rating, computed from feedback provided by all previous clients who have used the counselor's services. The online rating system operates on a scale from 0 to 5, accurate to 1 decimal place.

Professional Titles

In China, there are 3 primary professional titles associated with psychological counselors: counselor, psychotherapist, and psychiatrist. Each of these titles is further categorized into 3 levels: junior, intermediate, and senior. For the purpose of our study, we assigned numerical codes to these levels: 1 for junior, 2 for intermediate, and 3 for senior. As a counselor may hold multiple titles, the variable *professional titles* or *PT* we used in our analysis represented the sum of the codes for all the titles held by the counselor:

$$PT = PTcounselor + PTpsychotherapist + PTpsychiatrist$$

Practical Experience

Practical experience can be reflected by the number of hours of service, which is directly extracted from the counselor's home page and measured in hours. Although the number of hours of service is displayed online, it primarily reflects the service duration proof provided by the counselor when registering on the platform. Therefore, we classified practical experience as an offline signal, as it mainly showcases the counselor's professional experience outside the online platform.

Educational Level

The educational levels of counselors were classified into 4 categories and assigned corresponding numerical codes. The categories and codes used were as follows: "1 for college and below," "2 for a bachelor's degree," "3 for a master's degree," and "4 for a doctorate and above."

Control Variables

City Level

The variability in internet coverage across China's vast regions could affect clients' behavior in choosing online psychological counselors. Due to the stigma associated with mental health issues, the platform does not display clients' personal information. To control for regional differences, we included the city tier of the psychological counselors as a control variable. According to the latest Chinese city classification standards [65], we coded city level classes into 3 categories: "3 for first-tier cities," "2 for second-tier cities," and "1 for others."

Gender

Gender was coded as "1 for male" and "2 for female."

Estimation Model

The descriptive statistics for all variables (see Table 1) reveal that clients' choices, price, prosocial behavior, and practical experience exhibit significantly larger extremum and SD. Such variability might result in biased and inefficient model estimations. Thus, we choose to log-transform these variables to mitigate the impact of extreme values in our model construction.

Table 1. Summary statistics.

Variable	Value, mean (SD)	Value, minimum	Value, maximum
CC[a]	2075.21 (2569.61)	1	14,408
Pr[b]	317.38 (248.57)	10	1600
PB[c]	276.44 (223.95)	10	990
OR[d]	4.90 (0.16)	3	5
PT[e]	2.20 (0.98)	0	7
PE[f]	2084.87 (2565.08)	0	14,552
EL[g]	1.59 (0.91)	1	4
CL[h]	2.27 (0.83)	1	3
Gen[i]	1.75 (0.43)	1	2

[a]CC: clients' choices.
[b]Pr: price.
[c]PB: prosocial behavior.
[d]OR: online reputation.
[e]PT: professional titles.
[f]PE: practical experience.
[g]EL: educational level.
[h]CL: city level.
[i]Gen: gender.

XSL•FO
RenderX

In order to control for unobservable heterogeneity that might influence clients' choices, we used a dynamic panel data model for our analysis. We used the Hausman test to contrast the fixed and random effects models. Given that the resulting *P* value was below the conventional level of significance (.05), we opted for a fixed effects model for our subsequent analysis. The advantage of the fixed effects model is that it controls for time-invariant characteristics of individuals, allowing us to isolate the effects of variables that change over time. The empirical model is presented as follows:

$$log(CCit) = \beta 0 + \beta 1 log(Prit) + \beta 2 log(PBit) + \beta 3[(log(PBit))2] + \beta 4(ORit) + \beta 5(PTit) + \beta 6 log(PEit) + \beta 7(ELit) + \beta 8(CLit) + \beta 9(Genit) + \lambda t + \nu i + \varepsilon it$$

where *i* is the index of the counselor and *t* represents the time point. β_1 to β_9 are the parameters to be estimated. λ_t represents the time fixed effects, which controls for common changes in mental health needs over time (eg, seasonal variations, holidays). ν_i represents the individual fixed effect, which controls for the stable characteristics of each counselor that do not change over time (eg, registration duration on the platform, counseling style). ε is the error term. *CC* is clients' choices, *Pr* is price, *PB* is prosocial behavior, *OR* is online reputation, *PT* is professional titles, *PE* is practical experience, *EL* is educational level, *CL* is city level, and *Gen* is gender.

Results

Research Model Results

Table 2 presents the correlation coefficients of all variables for the analysis. The results show that almost all of the independent variables exhibit a correlation with the dependent variable. The correlations among the independent variables are significantly low. Moreover, the variance inflation factor values of the independent and control variables are less than 2, indicating that multicollinearity is not a concern in our model.

Table 3 presents the results of our empirical models. Throughout our analysis, we systematically incorporated variables, culminating in the construction of 4 distinct models. We observed a consistent increase in R^2 from 0.337 in Model 1 to 0.962 in Model 4 with the inclusion of variables. The significance levels of the key variables remained consistent across all models. Therefore, we chose to interpret our research results using the final model.

As shown in Model 4, in the online signals category, both price (β_1=.186, $P<.001$) and online reputation (β_4=.489, P=.002) positively affect the clients' choices, thereby supporting H1 and H3. Based on our previous assumptions, we incorporated prosocial behavior and its quadratic term into the empirical model. We find a significant inverted U-shaped relationship between prosocial behavior and the clients' choices (β_2=.190, P=.002; β_3=–.018, P=.001). After conducting the fixed effects model analysis, we used the U-shape test estimation algorithm by Lind and Mehlum [66] to verify the stability of the inverted U-shaped relationship. The procedure involved 4 main steps. First, we used the Wald test to assess the significance of the linear and nonlinear combinations of prosocial behaviors. The *P* value of the Wald test was .004, indicating that these variables significantly contributed to the model. Second, we calculated the slope of the curve at the low and high points of prosocial behavior. According to our research data, the low point was 2.40 and the high point was 6.90. The results showed that the slope of prosocial behavior at the low point was 0.102, and at the high point, it was –0.060. Third, we used the likelihood ratio test to verify whether the influence of prosocial behavior on clients' choices increased monotonically at low levels of prosocial behavior and decreased monotonically at high levels. The *P* value of the likelihood ratio test was .001, indicating that this nonlinear relationship was significant. Fourth, we used the Delta confidence interval test to verify whether the poles of the curve fell within the range of high and low values of prosocial behaviors. The results showed that the poles fell within the interval between the 2 values, indicating that the inverted U-shape was considered stable. Therefore, H2 is supported. In the offline signals category, we discovered that practical experience (β_6=.911, $P<.001$) has a positive impact on clients' choices, supporting H4b. However, the effect of professional titles and educational levels on the clients' choices was found to be insignificant; thus, H4a and H4c are not supported.

XSL•FO
RenderX

Table 2. The correlation coefficient table.

Variable	log(CC[a])	log(Pr[b])	log(PB[c])	OR[d]	PT[e]	log(PE[f])	EL[g]	CL[h]	Gen[i]
log(CC)									
r	1	0.369	0.276	0.370	0.197	0.957	–0.040	–0.011	0.170
P value	__j	<.001	<.001	<.001	<.001	<.001	.08	.64	<.001
log(Pr)									
r	0.369	1	0.094	0.103	0.083	0.397	–0.004	0.131	0.098
P value	<.001	—	<.001	<.001	<.001	<.001	.85	<.001	<.001
log(PB)									
r	0.276	0.094	1	0.061	0.055	0.247	–0.052	–0.058	0.021
P value	<.001	<.001	—	.01	.02	<.001	.02	.01	.36
OR									
r	0.37	0.103	0.061	1	0.095	0.403	–0.079	0.007	0.202
P value	<.001	<.001	0.01	—	<.001	<.001	<.001	.75	<.001
PT									
r	0.197	0.083	0.055	0.095	1	0.195	0.054	0.001	0.045
P value	<.001	<.001	.02	<.001	—	<.001	.02	.99	.049
log(PE)									
r	0.957	0.397	0.247	0.403	0.195	1	–0.047	0.001	0.200
P value	<.001	<.001	<.001	<.001	<.001	—	.04	.99	<.001
EL									
r	–0.040	–0.004	–0.052	–0.079	0.054	–0.047	1	0.042	0.076
P value	.08	.85	.02	<.001	.02	.04	—	.06	<.001
CL									
r	–0.011	0.131	–0.058	0.007	0.001	0.001	0.042	1	0.095
P value	.64	<.001	.01	.75	.99	.99	.06	—	<.001
Gen									
r	0.170	0.098	0.021	0.202	0.045	0.200	0.076	0.095	1
P value	<.001	<.001	.36	<.001	.049	<.001	<.001	<.001	—

[a]CC: clients' choices.
[b]Pr: price.
[c]PB: prosocial behavior.
[d]OR: online reputation.
[e]PT: professional titles.
[f]PE: practical experience.
[g]EL: educational level.
[h]CL: city level.
[i]Gen: gender.
[j]Not applicable.

XSL•FO
RenderX

Table 3. Empirical model results.

Variable	Model 1	*P* value	Model 2	*P* value	Model 3	*P* value	Model 4	*P* value
log(Pr[a]), mean (SE)	—[b]		0.421 (0.154)	.01	—		0.186 (0.042)	<.001
log(PB[c]), mean (SE)	—		0.148 (0.028)	<.001	—		0.190 (0.063)	.002
$(\log(PB))^2$, mean (SE)	—		—		—		–0.018 (0.006)	.01
OR[d], mean (SE)	—		9.820 (0.517)	<.001	—		0.489 (0.157)	.002
PT[e], mean (SE)	—		—		0.027 (0.023)	.24	0.024 (0.023)	.29
log(PE[f]), mean (SE)	—		—		0.920 (0.006)	<.001	0.911 (0.007)	<.001
EL[g], mean (SE)	—		—		0.005 (0.010)	.63	0.007 (0.010)	.50
CL[h], mean (SE)	–0.937 (0.488)	<.001	–0.675 (0.216)	.002	–0.291 (0.059)	<.001	–0.267 (0.059)	<.001
Gen[i], mean (SE)	–2.125 (0.488)	<.001	–2.314 (0.436)	<.001	–0.121 (0.119)	.31	–0.200 (0.119)	.09
Time fixed effect	Yes		Yes		Yes		Yes	
Individual fixed effect	Yes		Yes		Yes		Yes	
R^2	0.337		0.485		0.961		0.962	
Number of time periods, T	4		4		4		4	
Observations, N	1948		1948		1948		1948	

[a]Pr: price.
[b]Not applicable.
[c]PB: prosocial behavior.
[d]OR: online reputation.
[e]PT: professional titles.
[f]PE: practical experience.
[g]EL: educational level.
[h]CL: city level.
[i]Gen: gender.

Robustness Check

In the robustness check section, we divided the dataset into 2 groups based on service duration. These two subsamples included (1) psychological counselors with fewer than 1000 hours of service and (2) those with 1000 hours or more. As service duration is consistent within each group, we did not include this factor in the model. Table 4 shows that the findings from these subsamples correspond with the primary results from the full dataset. Therefore, research findings exhibit high robustness and reliably support our preliminary results.

XSL•FO
RenderX

Table 4. Robustness check results.

Variable	Model 5 (group 1)	*P* value	Model 6 (group 2)	*P* value
log(Pr[a]), mean (SE)	0.207 (0.333)	<.001	0.497 (0.070)	<.001
log(PB[b]), mean (SE)	1.261(0.342)	<.001	0.331 (0.146)	.02
$(\log(PB))^2$, mean (SE)	–0.140 (0.036)	<.001	–0.029 (0.014)	.04
OR[c], mean (SE)	7.866 (0.636)	<.001	4.105 (1.774)	.02
PT[d], mean (SE)	–0.139 (0.042)	.42	0.028 (0.036)	.44
EL[e], mean (SE)	0.006 (0.086)	.94	0.004 (0.015)	.04
CL[f], mean (SE)	–0.672 (0.443)	.13	–0.711 (0.101)	<.001
Gen[g], mean (SE)	–1.475 (0.539)	.01	–0.377 (0.667)	.27
Time fixed effect	Yes		Yes	
Individual fixed effect	Yes		Yes	
R^2	0.694		0.177	
Number of time periods, T	4		4	
Observations, N	800		1148	

[a]Pr: price.
[b]PB: prosocial behavior.
[c]OR: online reputation.
[d]PT: professional titles.
[e]PE: practical experience.
[f]EL: educational level.
[g]CL: city level.
[h]Gen: gender.

Discussion

Analysis of Results

In this paper, we draw on signaling theory to investigate how information on psychological counselors' home pages influences clients' choices. The findings highlight the crucial role of online and offline signals in shaping decision-making processes and demonstrate that effective information disclosure can encourage consumers to select a psychological counselor.

In the context of online signals, it has been found that both price and online reputation affect a client's choice of psychological counselor, aligning with the findings of previous relevant studies [26,42,43,67-69]. On the one hand, given that consumers might not possess the requisite expertise to evaluate professional services, clients are often willing to pay a premium in exchange for the assurance of higher service quality [39]. As prices increase, consumers expect better services from a counselor, and thus are more inclined to select that counselor. On the other hand, a strong online reputation can boost consumers' trust in a counselor's professional skills. Demonstrating an online reputation highlights the consistency and professionalism of a counselor's service quality and reveals the professional attitude and commitment they maintain during service delivery.

In the context of offline signals, it was determined that practical experience has a positive impact on a client's choice of a psychological counselor. Offline signals mainly offer insights into a counselor's personal attributes, thus often considered reflective of their skills and professional standards. Although offline signals originate from a physical environment, this study demonstrates that consumers attribute significant value to them and recognize their essential role in decision making. In other words, even in the digital sphere, the importance and impact of offline signals are preserved.

Interestingly, we discovered an inverted U-shaped relationship between a psychological counselor's prosocial behavior and clients' choices. This suggests that while clients appreciate a certain level of prosocial behavior, an overly excessive display may arouse suspicion and negatively affect their choices. Specifically, moderate prosocial behavior can improve a counselor's image and online reputation, thus enhancing client trust. However, exceeding a certain threshold of prosocial behavior may be perceived as insincere, resulting in increased client skepticism toward the counselor.

Theoretical Implications

Our study broadens the scope of theoretical research on signaling theory. First, we introduce signaling theory into the specific service domain of online psychological counseling and establish a theoretical framework to clarify the role of online and offline signals in the process of clients choosing a psychological counselor. The results of this study align with Zhou et al [70], which highlights the significance of both demonstration and description signals in online mental health markets and

XSL•FO
RenderX

demonstrates how various types of signals can influence clients' choices.

Second, this study considers the service price, online reputation, and prosocial behavior as signals and incorporates them into the research framework, which is an approach previously unexplored in previous studies. Previous studies have often focused on isolated online signals, such as online reviews and ratings [25,71,72]. Our research integrates both online signals and offline signals, offering a more comprehensive framework that enhances the understanding of how different signals interact to influence clients' choices.

Finally, our research results reveal an inverted U-shaped relationship between a psychological counselor's prosocial behavior and client's choices. This finding addresses a gap in the literature, where prosocial behavior was generally considered to have a linear positive effect [73-76]. Our study shows that there is an optimal level of prosocial behavior, beyond which client skepticism may arise, thus contributing to a more nuanced understanding of the dynamics at play.

Managerial Implications

For psychological counselors in OHCs, our study reveals the positive influence of service price on clients' choices. Consequently, psychological counselors should leverage their pricing autonomy in OHCs, using their service price as a signal of their capability to convey accurate quality signals to potential clients. Moreover, considering the significant impact of an online reputation on client's choice, it is crucial for psychological counselors to carefully manage client relationships and maintain their online reputation. This suggests that, besides providing high-quality services, psychological counselors should proactively respond to client feedback, engage in OHC activities, and enhance their online visibility to gain a competitive edge in the domain of online psychological counseling [77]. Finally, our research exposes an inverted U-shaped relationship between a psychological counselor's prosocial behavior and clients' choices. This finding suggests that psychological counselors must adeptly balance and temper their prosocial behavior in practice, aiming for an ideal equilibrium that attracts clients while maintaining a professional image.

For OHC platforms, despite operating in a digital environment, offline signals from psychological counselors still play a significant role in client selection. Therefore, platforms may consider displaying more offline personal information on the counselors' profile pages, such as their affiliated institutions and accolades, which can enhance information transparency and boost client trust. At the same time, platforms need to recognize the high level of user attention given to key signals such as price, online reputation, and prosocial behavior. As a result, when designing web and app interfaces, platforms should strategically position this information prominently, allowing users to quickly and efficiently access these details.

Limitations

We admit some limitations in our study. First, this study uses data from a single platform, YiDianLing, collected at 4 intervals throughout the year. While YiDianLing is a reliable source, relying on one website may limit the generalizability of our findings. The 4 data collection periods were chosen to cover key times of the year, providing a representative snapshot. However, future research should incorporate data from multiple platforms and extend the observation period to enhance the findings' robustness and applicability. Second, although our study on online psychological counseling in China offers valuable insights, there is significant variability in counseling services across different countries, particularly in terms of providers and content. Future research should include data from multiple countries to capture a more comprehensive view of global trends and differences in online psychological counseling. Third, this study focuses on using signaling theory to explore clients' decision-making processes when choosing online psychological counselors, emphasizing the behaviors prior to engaging in online counseling. However, psychological counseling is inherently a long-term and complex process. Future research should incorporate assessments of the long-term outcomes and effectiveness of counseling sessions to comprehensively evaluate the overall process of online psychological counseling. Additionally, the results and effectiveness of clients' previous psychological counseling sessions may also influence their choices to some extent. Due to privacy protections, YiDianLing has not disclosed data regarding these past experiences. Future research should consider addressing this limitation by finding ways to incorporate such data while ensuring privacy protection. Fourth, while we considered regional diversity in our sample selection and controlled for the city tiers of psychological counselors in our data analysis, the lack of direct data on clients' locations may impact the generalizability of our results. Under the condition of complying with relevant regulations, future research should collect and analyze data on clients' locations to enhance the study's comprehensiveness. Fifth, counselors' home pages also contain some textual data that disclose clients' experiences, but this paper only analyzed some quantitative data and future research can do deeper mining of textual data of OHCs.

Conclusions

This study uses signaling theory to investigate client preferences for psychological counselors. We segmented the information displayed on counselors' home pages into online and offline signals and established a theoretical model for client decision making using data derived from a Chinese online health community. The findings reveal that service price, online reputation, professional titles, and practical experience positively influence clients' choices. Additionally, we discovered an inverted U-shaped correlation between a psychological counselor's prosocial behavior and clients' choices. Fundamentally, this study unravels a complex client selection process influenced by different signals, providing valuable guidance for both researchers and practitioners.

XSL•FO
RenderX

Acknowledgments

This study was funded by the National Natural Science Foundation of China (grant 72274068).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XF contributed to writing the original draft, methodology, formal analysis, data curation, and conceptualization. YH was responsible for writing, review and editing, validation, supervision, methodology, and conceptualization. HP contributed to conceptualization, methodology, and writing. SL was responsible for conceptualization, methodology, and data curation. JX and ZZ contributed to data curation and writing. Generative artificial intelligence was not used in any portion of the manuscript writing. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Choudhury MD, De S. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. 2014 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; 2014 May 16; USA p. 71-80. [doi: 10.1609/icwsm.v8i1.14526]
2. Freeman M. The world mental health report: transforming mental health for all. World Psychiatry 2022;21(3):391-392 [FREE Full text] [doi: 10.1002/wps.21018] [Medline: 36073688]
3. Deng Z, Hong Z, Zhang W, Evans R, Chen Y. The effect of online effort and reputation of physicians on patients' choice: 3-wave data analysis of China's good doctor website. J Med Internet Res 2019;21(3):e10170 [FREE Full text] [doi: 10.2196/10170] [Medline: 30848726]
4. Zhang Y, Li X, Fan W. User adoption of physician's replies in an online health community: an empirical study. Asso for Info Science & Tech 2020;71(10):1179-1191. [doi: 10.1002/asi.24319]
5. Alleman JR. Online counseling: the Internet and mental health treatment. Psychotherapy: Theory, Research, Practice, Training 2002;39(2):199-209. [doi: 10.1037/0033-3204.39.2.199]
6. Feijt MA, de Kort YA, Bongers IM, IJsselsteijn WA. Perceived drivers and barriers to the adoption of eMental health by psychologists: the construction of the levels of adoption of eMental health model. J Med Internet Res 2018;20(4):e153 [FREE Full text] [doi: 10.2196/jmir.9485] [Medline: 29691215]
7. Zhou J, Liu F, Zhou T. Exploring the factors influencing consumers to voluntarily reward free health service contributors in online health communities: empirical study. J Med Internet Res 2020;22(4):e16526 [FREE Full text] [doi: 10.2196/16526] [Medline: 32286231]
8. Vogel DL, Wester SR, Larson LM. Avoidance of counseling: psychological factors that inhibit seeking help. Jour of Counseling & Develop 2011;85(4):410-422. [doi: 10.1002/j.1556-6678.2007.tb00609.x]
9. YU Y, MEI QY, WANG QH. From offline to online: how health insurance policies drive the demand for online healthcare service? 2016 Presented at: Proceedings of the 20th Pacific Asia Conference on Information Systems; 2016 Jul 1; Chaiyi, Taiwan p. 1-13.
10. The 52nd Statistical Report on China's Internet Development. China: China Internet Network Information Center; 2023. URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202311/P020231121355042476714.pdf [accessed 2024-09-12]
11. Xiang YT, Yu X, Sartorius N, Ungvari GS, Chiu HFK. Mental health in China: challenges and progress. Lancet 2012;380(9855):1715-1716. [doi: 10.1016/S0140-6736(11)60893-3] [Medline: 23158236]
12. Chen J, Zhu S. Online information searches and help seeking for mental health problems in Urban China. Adm Policy Ment Health 2016;43(4):535-545. [doi: 10.1007/s10488-015-0657-6] [Medline: 25981055]
13. Que J, Lu L, Shi L. Development and challenges of mental health in China. Gen Psychiatr 2019;32(1):e100053 [FREE Full text] [doi: 10.1136/gpsych-2019-100053] [Medline: 31179426]
14. Yang Y, Li W, Zhang Q, Zhang L, Cheung T, Xiang YT. Mental health services for older adults in China during the COVID-19 outbreak. Lancet Psychiatry 2020;7(4):e19 [FREE Full text] [doi: 10.1016/S2215-0366(20)30079-1] [Medline: 32085843]
15. Mavlanova T, Benbunan-Fich R, Koufaris M. Signaling theory and information asymmetry in online commerce. Information & Management 2012;49(5):240-247. [doi: 10.1016/j.im.2012.05.004]
16. Jøsang A, Ismail R, Boyd C. A survey of trust and reputation systems for online service provision. Decision Support Systems 2007;43(2):618-644. [doi: 10.1016/j.dss.2005.05.019]
17. Huang X, Sun P, Zhang X, Wu J. What affects patients' choice of consultant: an empirical study of online doctor consultation service. Electron Commer Res 2022:625. [doi: 10.1007/s10660-022-09617-w]

XSL•FO
RenderX

18. Connelly BL, Certo ST, Ireland RD, Reutzel CR. Signaling theory: a review and assessment. Journal of Management 2010;37(1):39-67. [doi: 10.1177/0149206310388419]
19. Li J, Tang J, Yen DC, Liu X. Disease risk and its moderating effect on the e-consultation market offline and online signals. ITP 2019;32(4):1065-1084. [doi: 10.1108/itp-03-2018-0127]
20. Zhang M, Guo X, Wu T. Impact of free contributions on private benefits in online healthcare communities. International Journal of Electronic Commerce 2019;23(4):492-523. [doi: 10.1080/10864415.2019.1655208]
21. Pulat M, Barutçu-Yıldırım F. Online counseling and ethics: A systematic review of empirical research. Online Counseling and Ethics: A Systematic Review of Empirical Research 2021:1-12.
22. Zeren G, Erus SM, Amanvermez Y, Genç AB, Duy B. Client's experiences of online counseling: satisfaction and therapeutic alliance. Cukurova University Faculty of Education Journal Cukurova University 2022;51(1):658. [doi: 10.14812/cuefd.843542]
23. Khan S, Shapka JD, Domene JF. Counsellors' experiences of online therapy. British Journal of Guidance & Counselling 2021;50(1):43-65. [doi: 10.1080/03069885.2021.1885009]
24. Shah AM, Naqvi RA, Jeong OR. The impact of signals transmission on patients' choice through e-consultation websites: an econometric analysis of secondary datasets. Int J Environ Res Public Health 2021;18(10):5192 [FREE Full text] [doi: 10.3390/ijerph18105192] [Medline: 34068291]
25. Li J, Tang J, Jiang L, Yen DC, Liu X. Economic success of physicians in the online consultation market: a signaling theory perspective. International Journal of Electronic Commerce 2019;23(2):244-271. [doi: 10.1080/10864415.2018.1564552]
26. Wu H, Deng Z, Wang B, Gupta S. How does service price influence patients' decisions? An examination of the free-market pricing mechanism in online health communities. Electron Markets 2021;31(4):877-893. [doi: 10.1007/s12525-020-00453-0]
27. Meng F, Zhang X, Liu L, Ren C. Converting readers to patients? From free to paid knowledge-sharing in online health communities. Information Processing & Management 2021;58(3):102490. [doi: 10.1016/j.ipm.2021.102490]
28. Chen S, Guo X, Wu T, Ju X. Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. Information Processing & Management 2020;57(5):102253. [doi: 10.1016/j.ipm.2020.102253]
29. Li L, Tadelis S, Zhou X. Buying reputation as a signal of quality: evidence from an online marketplace. The RAND J of Economics 2020;51(4):965-988. [doi: 10.1111/1756-2171.12346]
30. Reuber AR, Fischer E. Signalling reputation in international online markets. Strategic Entrepreneurship 2009;3(4):369-386. [doi: 10.1002/sej.79]
31. Al-Hasan A. Online social stock picking: an empirical examination. International Journal of Electronic Commerce 2018;22(1):66-97. [doi: 10.1080/10864415.2018.1396119]
32. Wan Y, Zhang Y, Yan M. What influences patients' willingness to choose in online health consultation? An empirical study with PLS–SEM. IMDS 2020;120(12):2423-2446. [doi: 10.1108/imds-11-2019-0633]
33. Guo S, Guo X, Fang Y, Vogel D. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. Journal of Management Information Systems 2017;34(2):487-519. [doi: 10.1080/07421222.2017.1334480]
34. Wang J, Liu H, Ye J. The impact of individual and team professional capital on physicians' performance in online health-care communities: a cross-level investigation. Internet Research 2022;33(1):152-177. [doi: 10.1108/intr-08-2021-0544]
35. Yang H, Du HS, Shang W. Understanding the influence of professional status and service feedback on patients' doctor choice in online healthcare markets. Internet Research 2020;31(4):1236-1261. [doi: 10.1108/intr-10-2019-0429]
36. Zhang S, Wang JN, Chiu YL, Hsu YT. Exploring types of information sources used when choosing doctors: observational study in an online health care community. J Med Internet Res 2020;22(9):e20910 [FREE Full text] [doi: 10.2196/20910] [Medline: 32936080]
37. Hsu YT, Chiu YL, Wang JN, Liu HC. Impacts of physician promotion on the online healthcare community: using a difference-in-difference approach. Digit Health 2022;8:20552076221106319 [FREE Full text] [doi: 10.1177/20552076221106319] [Medline: 35694119]
38. Campo S, Jesús YM. Exploring non-linear effects of determinants on tourists' satisfaction. International Journal of Culture, Tourism and Hospitality Research Emerald 2009;3(2):127-138. [doi: 10.1108/17506180910962131]
39. Wu H, Lu N. Service provision, pricing, and patient satisfaction in online health communities. Int J Med Inform 2018;110:77-89. [doi: 10.1016/j.ijmedinf.2017.11.009] [Medline: 29331257]
40. Li J, Hu F, Yan T, Cai X, Song X. How to charge doctors and price medicines in a two-sided online healthcare platform with network externalities? International Journal of Production Research 2022;61(9):3052-3070. [doi: 10.1080/00207543.2022.2077151]
41. Bolle S, van Weert JCM, Daams JG, Loos EF, de Haes HCJM, Smets EMA. Online health information tool effectiveness for older patients: a systematic review of the literature. J Health Commun 2015;20(9):1067-1083. [doi: 10.1080/10810730.2015.1018637] [Medline: 26165846]
42. Conrad DA, Sales A, Liang SY, Chaudhuri A, Maynard C, Pieper L, et al. The impact of financial incentives on physician productivity in medical groups. Health Serv Res 2002;37(4):885-906 [FREE Full text] [doi: 10.1034/j.1600-0560.2002.57.x] [Medline: 12236389]

XSL•FO
RenderX

43. Zeithaml VA, Berry LL, Parasuraman A. Communication and control processes in the delivery of service quality. Journal of Marketing 2018;52(2):35-48. [doi: 10.1177/002224298805200203]
44. Zhang S, Fu YN, Liu Q, Turel O, He Q. Psychological capital mediates the influence of meaning in life on prosocial behavior of university students: a longitudinal study. Children and Youth Services Review 2022;140:106600. [doi: 10.1016/j.childyouth.2022.106600]
45. Jing D, Jin Y, Liu J. The impact of monetary incentives on physician prosocial behavior in online medical consulting platforms: evidence from China. J Med Internet Res 2019;21(7):e14685 [FREE Full text] [doi: 10.2196/14685] [Medline: 31350834]
46. Feng L, Guo Q. Beneficial effect of altruism on well-being among Chinese college students: the role of self-esteem and family socioeconomic status. Journal of Social Service Research 2016;43(3):416-431. [doi: 10.1080/01488376.2016.1242449]
47. Zhou J, Zuo M, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
48. Gong Y, Wang H, Xia Q, Zheng L, Shi Y. Factors that determine a patient's willingness to physician selection in online healthcare communities: a trust theory perspective. Technol Soc 2021;64:101510 [FREE Full text] [doi: 10.1016/j.techsoc.2020.101510] [Medline: 36033357]
49. Qiu C, Zhang Y, Wang X, Gu D. Trust-based research: influencing factors of patients' medical choice behavior in the online medical community. Healthcare (Basel) 2022;10(5):938 [FREE Full text] [doi: 10.3390/healthcare10050938] [Medline: 35628075]
50. Gregg DG, Scott JE. The role of reputation systems in reducing on-line auction fraud. International Journal of Electronic Commerce 2014;10(3):95-120. [doi: 10.2753/jec1086-4415100304]
51. Lambert MJ, Barley DE. Research summary on the therapeutic relationship and psychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training 2001;38(4):357-361. [doi: 10.1037//0033-3204.38.4.357]
52. Gärtner M, Andersson D, Västfjäll D, Tinghög G. Affect and prosocial behavior: the role of decision mode and individual processing style. Judgm Decis Mak 2023;17(1):1-13. [doi: 10.1017/s1930297500008998]
53. Wang JJ, Liu H, Cui X, Ye J, Chen H. Impact of a physician's prosocial behavior on the patient's choice: an empirical investigation in online health community. Information Technology & People Emerald 2022;36(4):1703-1725. [doi: 10.1108/itp-12-2020-0878]
54. Filieri R, Mariani M. The role of cultural values in consumers' evaluation of online review helpfulness: a big data approach. International Marketing Review Emerald 2021;38(6):1267-1288. [doi: 10.1108/imr-07-2020-0172]
55. Zervas G, Proserpio D, Byers JW. A first look at online reputation on Airbnb, where every stay is above average. Mark Lett 2020;32(1):1-16. [doi: 10.1007/s11002-020-09546-4]
56. Djafarova E, Trofimenko O. 'Instafamous' – credibility and self-presentation of micro-celebrities on social media. Information, Communication & Society 2018;22(10):1432-1446. [doi: 10.1080/1369118X.2018.1438491]
57. Utz S, Matzat U, Snijders C. On-line reputation systems: the effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. International Journal of Electronic Commerce 2014;13(3):95-118. [doi: 10.2753/jec1086-4415130304]
58. Cioppi M, Curina I, Forlani F, Pencarelli T. Online presence, visibility and reputation: a systematic literature review in management studies. Journal of Research in Interactive Marketing 2019;13(4):547-577. [doi: 10.1108/jrim-11-2018-0139]
59. Qiao W, Yan Z, Wang X. Join or not: the impact of physicians' group joining behavior on their online demand and reputation in online health communities. Information Processing & Management 2021;58(5):102634. [doi: 10.1016/j.ipm.2021.102634]
60. Ba Z, Zhao Y, Song S, Zhu Q. Does the involvement of charities matter? Exploring the impact of charities' reputation and social capital on medical crowdfunding performance. Information Processing & Management 2022;59(3):102942. [doi: 10.1016/j.ipm.2022.102942]
61. Delgadillo J, Moreea O, Lutz W. Different people respond differently to therapy: a demonstration using patient profiling and risk stratification. Behav Res Ther 2016;79:15-22. [doi: 10.1016/j.brat.2016.02.003] [Medline: 26937855]
62. Evans L. Professionalism, professionality and the development of education professionals. British Journal of Educational Studies 2008;56(1):20-38. [doi: 10.1111/j.1467-8527.2007.00392.x]
63. Hemsley B, Balandin S, Togher L. Narrative analysis of the hospital experience for older parents of people who cannot speak. Journal of Aging Studies 2007;21(3):239-254. [doi: 10.1016/j.jaging.2006.12.001]
64. Tyler TR. Psychological perspectives on legitimacy and legitimation. Annu Rev Psychol 2006;57:375-400. [doi: 10.1146/annurev.psych.57.102904.190038] [Medline: 16318600]
65. List of cities in China. Wikipedia. 2024. URL: https://en.wikipedia.org/w/index.php?title=List_of_cities_in_China&oldid=1211361304 [accessed 2024-03-15]
66. Lind JT, Mehlum H. With or without U? The appropriate test for a u-shaped relationship*. Oxford Bulletin of Economics and Statistics 2010;72(1):118. [doi: 10.1111/j.1468-0084.2009.00569.x]
67. Anagnostopoulou SC, Buhalis D, Kountouri IL, Manousakis EG, Tsekrekos AE. The impact of online reputation on hotel profitability. International Journal of Contemporary Hospitality Management 2020;32(1):20-39. [doi: 10.1108/ijchm-03-2019-0247]

XSL•FO
RenderX

68. Chen T, Samaranayake P, Cen X, Qi M, Lan Y. The impact of online reviews on consumers' purchasing decisions: evidence from an eye-tracking study. Front Psychol 2022;13:865702 [FREE Full text] [doi: 10.3389/fpsyg.2022.865702] [Medline: 35756238]
69. Liu X, Guo X, Wu H, Wu T. The impact of individual and organizational reputation on physicians' appointments online. International Journal of Electronic Commerce 2016;20(4):551-577. [doi: 10.1080/10864415.2016.1171977]
70. Zhou J, Kishore R, Amo L, Ye C. Description and demonstration signals as complements and substitutes in an online market for mental health care. MIS Quarterly 2022;46(4):2055-2084. [doi: 10.25300/misq/2022/16122]
71. Chen L, Baird A, Straub D. A linguistic signaling model of social support exchange in online health communities. Decision Support Systems 2020;130:113233. [doi: 10.1016/j.dss.2019.113233]
72. Liu J, Gao L. Analysis of topics and characteristics of user reviews on different online psychological counseling methods. Int J Med Inform 2021;147:104367. [doi: 10.1016/j.ijmedinf.2020.104367] [Medline: 33401170]
73. Kawamura Y, Ohtsubo Y, Kusumi T. Effects of cost and benefit of prosocial behavior on reputation. Social Psychological and Personality Science 2020;12(4):452-460. [doi: 10.1177/1948550620929163]
74. Liu X, Guo H, Wang L, Hu M, Wei Y, Liu F, et al. Effect of prosocial behaviors on e-consultations in a web-based health care community: panel data analysis. J Med Internet Res 2024;26:e52646 [FREE Full text] [doi: 10.2196/52646] [Medline: 38663006]
75. Gherghel C, Nastas D, Hashimoto T, Takai J, Cargile AC. Culture, morality, and the effect of prosocial behavior motivation on positive affect. Ethics & Behavior 2019;30(2):126-149. [doi: 10.1080/10508422.2019.1651651]
76. Aknin LB, van de Vondervoort JW, Hamlin JK. Positive feelings reward and promote prosocial behavior. Curr Opin Psychol 2018;20:55-59. [doi: 10.1016/j.copsyc.2017.08.017] [Medline: 28837957]
77. Yang H, Guo X, Wu T, Ju X. Exploring the effects of patient-generated and system-generated information on patients' online search, evaluation and decision. Electronic Commerce Research and Applications 2015;14(3):192-203. [doi: 10.1016/j.elerap.2015.04.001]

Abbreviations

OHC: online health community
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 15.03.24; peer-reviewed by YT Hsu, YL Chiu; comments to author 24.05.24; revised version received 25.06.24; accepted 12.08.24; published 10.10.24.

Please cite as:
Feng X, Hu Y, Pfaff H, Liu S, Xie J, Zhang Z
Exploring Client Preferences for Psychological Counselors in a Chinese Online Health Community: Longitudinal Study
J Med Internet Res 2024;26:e58428
URL: https://www.jmir.org/2024/1/e58428
doi:10.2196/58428
PMID:

©Xiandong Feng, Yinhuan Hu, Holger Pfaff, Sha Liu, Jinzhu Xie, Zemiao Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Quality and Cultural Safety of Online Osteoarthritis Information for Affected Persons and Health Care Professionals: Content Analysis

Smita Dhakal[1], MPH; Shermeen Merani[1], MPH; Vandana Ahluwalia[2], MD; Marisa Battistella[1], PharmD; Cornelia M Borkhoff[3], PhD; Glen Stewart Hazlewood[4], MD, PhD; Aisha Lofters[5], MD, PhD; Deborah A Marshall[4], PhD; Crystal MacKay[6], PhD; Anna R Gagliardi[1], PhD

[1]University Health Network, Toronto, ON, Canada
[2]William Osler Health System, Brampton, ON, Canada
[3]Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[4]University of Calgary, Calgary, AB, Canada
[5]Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada
[6]West Park Healthcare Centre, North York, ON, Canada

Corresponding Author:
Anna R Gagliardi, PhD
University Health Network
200 Elizabeth Street
Toronto, ON, M5G2C4
Canada
Phone: 1 416 340 4800
Email: anna.gagliardi@uhnresearch.ca

Abstract

Background: Osteoarthritis is more prevalent and severe among women than among men, but women are less likely to access early diagnosis and first-line management, particularly racialized immigrant women. Previous research advocated for greater access to culturally safe osteoarthritis information for both diverse women and health care professionals. The internet can reduce disparities by facilitating access to health information, but online materials can vary in quality.

Objective: This study aimed to assess the quality and cultural safety of online osteoarthritis materials for persons affected by osteoarthritis and health care professionals.

Methods: Content analysis was used to describe publicly available materials on osteoarthritis first-line management developed by Canadian organizations for affected persons or health care professionals. Searching, screening, and data extraction were performed in triplicate. We identified materials by searching Google, MEDLINE, and references of osteoarthritis-relevant guidelines and policies, and consulting our research team and collaborators. We assessed quality using DISCERN (University of Oxford) and a compiled framework for affected persons and health care professionals. We compiled frameworks to assess cultural safety. We derived an overall score, categorized as low (<50%), moderate (50%-69%), or high (≥70%+) for criteria met.

Results: After screening 176 items and eliminating 129, we included 47 osteoarthritis materials published between 2013 and 2023. Of those, 43 were for persons with osteoarthritis, most were developed by charities (n=31, 72.1%), based on expert advice (n=16, 55.2%), and in the format of booklets (n=15, 34.9%) or text on web pages (n=10, 23.3%). Of those, 23.3% (10/43) low, 46.5% (20/43) moderate, and 30.2% (13/43) high scored quality; and 25.6% (11/43), 48.8% (21/43), and 25.6% (11/43) were rated low, moderate, and high cultural safety, respectively. Of the 47 included osteoarthritis materials, 4 were for health care professionals. They were developed by a consortium (2/4, 50%), a charity (1/4, 25%), and a professional society (1/4, 25%), and largely based on expert advice (3/4, 75%). The format included infographics (3/4, 75%) and text on web pages (1/4, 25%). Of those, 25% (1/4), 25% (1/4), and 50% (2/4) were rated low, moderate, and high quality, respectively; and all were rated low for cultural safety. Quality and cultural safety did not appear to be associated with the characteristics of osteoarthritis materials (eg, type of developer, development method, and format).

Conclusions: Overall, included osteoarthritis materials for persons with osteoarthritis and health care professionals were of low to moderate quality and cultural safety. These findings reveal the need for further efforts to improve existing or develop new

XSL•FO
RenderX

osteoarthritis materials for both affected persons, including ethnoculturally diverse immigrant women, and health care professionals. Further research is needed to assess the quality and cultural safety of osteoarthritis materials developed by organizations outside of Canada and to establish a framework or instrument to assess cultural safety in the osteoarthritis context.

(J Med Internet Res 2024;26:e57698) doi:10.2196/57698

KEYWORDS

osteoarthritis; women's health; equity; educational materials; internet; content analysis; Canada; persons living with osteoarthritis; healthcare professionals; OA care; ethno-culturally women; immigrant women; diverse women; online materials; health information; prevention; management; misinformation; cultural safety; educational materials

Introduction

The internet has become a commonly used way to share health information with patients and health care professionals, alike. For patients, online health information can help people learn about disease prevention or management, prepare them to interact with the formal health care system, supplement advice from health care professionals, connect them to others with a common health concern, link patients with their own health records, or support those without access to care [1]. Associated positive outcomes include increased health care visits, patient question-asking during consultations, adherence to prescribed therapies, and behavioral compliance with self-care [2]. For health care professionals, online health information can help clinicians to answer patient questions, engage in shared decision-making, and support professional development and maintenance of competencies, with a demonstrated positive impact on knowledge and behavior [3,4].

Despite these numerous benefits, the use of the internet to access health information can present challenges to both patients and health care professionals. Inaccurate or incomplete information can misinform users, and poor design or functionality can impede learning, and the problem is widespread [5]. For example, a review of 153 reviews evaluating 11,785 websites offering health information to the public found that most were rated as poor quality [6]. Another review of 282 studies identified many technological barriers to the use of online health information by health care professionals including lack of interactivity, feedback, personalization, cost, and privacy concerns [4]. Therefore, as we strive to enhance widespread access to health information, it is important to assess online sources to identify those of high quality that can be recommended to users.

The internet can be particularly helpful in supporting persons with chronic conditions that benefit from self-management. One such condition is osteoarthritis. Osteoarthritis is an incurable condition affecting millions worldwide that most often is felt in the hips, knees, hands, neck, back, and feet [7]. Osteoarthritis manifests as chronic joint pain, tenderness, swelling, and stiffness, affecting activities of daily living, employment, and quality of life, and leading to other conditions such as heart disease, diabetes, or depression [7,8]. Early intervention is key to prevent worsening and manage symptoms, and involves education about self-management, which can include physical activity, weight loss, and pain management [9,10]. Osteoarthritis is more prevalent and severe among women than among men [11]. However, women are less likely than men to receive self-management advice, particularly racialized immigrant women [12-14]. This may be due to low rates of help-seeking that result from poor mainstream health care experiences [15], and lack of knowledge and skills among clinicians for supporting osteoarthritis self-management among ethnoculturally diverse persons [16,17]. Furthermore, immigrant women have lower rates of physical activity, an important part of osteoarthritis self-management, compared with immigrant men or nonimmigrants and are therefore in need of osteoarthritis self-management advice [18].

A multicountry study showed that access to health information on the internet facilitates health care access and mitigates the negative impact of social and economic determinants of health care access to reduce disparities [19]. In our previous research, ethnoculturally diverse women (White, East Asian, South Asian, and African and Caribbean Black) aged from 40 to 70+ with one or more of hand, hip, knee, and neck osteoarthritis for ages from 5 to 30+ years from across Canada, and women and men health care professionals of various specialties (chiropractor, health care executive, family physician, nurse practitioner, occupational therapist, pharmacist, physiotherapist, and policy maker) of early, mid and later career from across Canada identified and prioritized strategies to improve access to quality of osteoarthritis care for women. Strategies included greater access to publicly available osteoarthritis information for women that accommodates cultural expressions, norms and preferences, and to continuing education for health care professionals on osteoarthritis management and how to tailor osteoarthritis advice to women's intersectional factors [20,21]. Most research on support for osteoarthritis self-management has focused on the impact of osteoarthritis self-management programs delivered directly by health care professionals [22]. Other research on internet self-management resources has focused on conditions other than osteoarthritis; for example, fatigue among patients with cancer [23,24]. Hence, knowledge is needed on the quality and relevance of online health information for diverse women and health care professionals that could support osteoarthritis self-management.

Given the lack of access to osteoarthritis self-management advice among ethnoculturally diverse women [12-18]; and the important role of online health information in supporting osteoarthritis self-management among such women [19-21], and continuing education among health care professionals [3,4], the aim of this study was to identify online osteoarthritis information to improve osteoarthritis care for ethnoculturally diverse women. More specifically, the objective of this study was to assess the quality and cultural safety of online

XSL•FO
RenderX

osteoarthritis information for patients and health care professionals. If identified, we could recommend osteoarthritis materials of high quality and cultural safety to diverse women and health care professionals. If none or few were identified, that would signal the need for de novo development.

Methods

Approach

We used content analysis, a methodological approach widely used to analyze the content of written or visual materials [25,26]. More specifically, we applied manifest content analysis, which is a way to describe explicit content without theoretical interpretation. This technique differs from thematic coding typically used to analyze transcripts of interviews or focus groups and involves both quantitative (eg, counting the number of documents that mention a concept) and qualitative (eg, extracting content relevant to a concept) methods. To optimize rigor, we complied with content analysis methods [25,26] and criteria for reporting qualitative research [27]. The research team informed data collection, analysis, and interpretation to further support rigor. The research team was comprised of a 12-member advisory group of diverse women with osteoarthritis; 6 coinvestigators including health care professionals (family physician, rheumatologists, physiotherapist, and pharmacist), and health services researchers with expertise in the topics of osteoarthritis, person-centered care, equity, and women's health; and 4 collaborators who represented organizations that could apply the findings: 2 advocacy groups for persons with osteoarthritis and 2 health care professional societies. In particular, women advisors, who varied by type of osteoarthritis, age, length of time since osteoarthritis diagnosis, ethnocultural group, and location in Canada, helped us to search for osteoarthritis materials, develop and test the use of analytic frameworks, and summarize the findings.

Eligibility

Multimedia Appendix 1 provides detailed eligibility criteria. In brief, we included publicly available, online, English-language materials (eg, video, infographic, checklist, booklet, and web page featuring text with or without other materials) that support osteoarthritis first-line management for persons with osteoarthritis or health care professionals who care for such persons; and were developed by Canadian nonprofit organizations (eg, government, academic groups, professional societies, and charities) in 2012 or later, when the Arthritis Alliance Canada issued a framework to improve arthritis prevention and care [28]. These types of organizations are inclusive of multiple sectors, represent a comprehensive list of types of organizations most likely to generate develop and share osteoarthritis materials on the internet, and because they are nonprofit, materials are likely to freely accessible. We included osteoarthritis materials developed or shared by only Canadian organizations for reasons, that (1) this study was 1 part of a multicomponent research study focused on osteoarthritis care in the Canadian setting, (2) the study was funded by the Arthritis Society Canada, and (3) the study aimed to inform policies and practice needed in the Canadian setting to enhance equitable access to quality of osteoarthritis care. Furthermore, restricting to only Canadian osteoarthritis materials enhanced study feasibility; it required considerable time and effort to evaluate the eligible Canadian osteoarthritis materials—we would not have been able to do so for the potentially hundreds or thousands of osteoarthritis materials developed by organizations around the world. We excluded clinical practice guidelines and resources that focused solely on second-line management (eg, joint replacement) given the importance of first-line management in preventing further decline and optimizing the quality of life [9,10].

Searching and Screening

SM (woman of color, Master of Public Health candidate) and SD (woman of color, Master of Public Health research coordinator) conducted searching and independently screened all items in duplicate between April 12 and May 23, 2023. ARG (woman, PhD-trained principal investigator) provided training and oversight and resolved discrepancies by independent review, and then discussion with SM and SD. SM and SD first conducted searches, capturing the organization, title, and address of potentially relevant items in a Microsoft Excel file; then revisited each item to confirm eligibility. Given no standard method of searching for "gray literature" [29,30], we used a multipronged strategy that included (1) using Google, searched for "osteoarthritis AND Canada OR <each of the provinces>," and scanned search results on up to 6 pages of 10 results per page, or before relevancy dropped off; (2) consulting members of the aforementioned research team including women advisors, coinvestigators and collaborators plus other professional societies representing health care professionals who care for persons with osteoarthritis (eg, family physicians, physiotherapists, chiropractors, occupational therapists, and pharmacists); (3) scanning the websites of the Arthritis Society, advocacy groups and professional societies; (4) reviewed the content of Canadian guidelines and policies related to osteoarthritis identified in our previous research [31,32]; and (5) Searched MEDLINE for "osteoarthritis AND Canada AND Internet AND health resources OR education, medical, continuing" for articles that reviewed or developed osteoarthritis information.

Data Extraction

SM extracted data from osteoarthritis materials for persons with osteoarthritis, SD extracted data from osteoarthritis materials for health care professionals, and then each checked the others' work, consulting with ARG to resolve discrepancies through discussion to achieve 100% agreement. From all materials, we extracted data on osteoarthritis material characteristics (developer name, type of organization, year published, title of osteoarthritis material, web address, development approach, osteoarthritis topic, and objective) and design according to the Workgroup for Intervention Development and Evaluation Research framework (content, format, and delivery) [33].

Persons With Osteoarthritis

To assess the quality of osteoarthritis materials for affected persons, we used DISCERN because it is a validated instrument and has been widely used by many others to assess the quality of consumer health information [34]. The 15-item DISCERN

XSL•FO
RenderX

instrument assesses reliability (clarity of aims, aims achieved, relevance, transparency about sources of information, date information generated or published, information balanced or unbiased, links to additional information, and identification of uncertainties) and quality (description of treatments and underlying mechanism, treatment benefits, treatment risks, likely progression with no treatment, alternative treatment options, impact of different treatments on quality of life, and support for shared decision-making), and provides detailed criteria by which to identify and assess each item.

To assess the cultural safety of osteoarthritis information, in the absence of an existing validated or widely used model or framework, we compiled a 5-item framework from the only available exploratory studies of barriers to or preferences for culturally safe specific to osteoarthritis care among ethnoculturally diverse women: use of plain language (no medical jargon, use of active voice, brief uncomplicated sentences, and words less than 3 syllables), availability of translated versions, empathetic communication (language or tone conveys respect, understanding, and support; language avoids blame and does not provoke fear), recognition of ethnoculturally diverse beliefs and norms (depicts ethnoculturally diverse images, acknowledges different cultures, and uses cultural idioms), and strategies to overcome barriers of access to care [35-41].

Health Care Professionals

To extract data on quality, in the absence of an existing validated or widely used framework, we compiled a 7-item framework from the only available studies that identified quality criteria for online or virtual continuing education given that we wished to assess the quality of online continuing education about osteoarthritis: accommodated individual preferences for learning pace, content could be tailored to meet individual learning needs, provided example cases, included interactive components, evidence-based, supported skill enhancement, developed by credible sources and provided choice of format [4,42,43]. To assess cultural safety, in the absence of an existing validated or widely used framework, we compiled a 3-item framework from the only available exploratory studies of what constitutes cultural competency when caring for ethnoculturally diverse persons with any type of health care concern: recognition of disparities in access to or quality of care, insight on ethnocultural beliefs or norms, and strategies to deliver culturally safe care [36,44-46].

Data Analysis

SM and SD independently scored all osteoarthritis materials, then met to resolve discrepancies through discussion. ARG resolved the remaining uncertainties, and discrepancies, and checked all data. DISCERN scoring involved a 5-point scale with each of the 15 items marked as met (5), partially met (2-4), or not met (1). We deemed items that were not relevant to the materials as not applicable and did not record a score (eg, not all items included treatment information). We applied a similar scoring system for all other frameworks. For all instruments and frameworks, we derived an overall score based on the total number of "met" items divided by the total number of applicable items. Applying the same rubric as DISCERN, we categorized the overall score as low (<50%), moderate (50%-69%), or high (≥70%+). We used summary statistics, referring to simple counts of numbers and percentages, to report osteoarthritis material characteristics, design, quality, and cultural safety, and used tables to summarize data along with specific examples of osteoarthritis materials.

Ethical Considerations

No ethics approval was required because all data were publicly available.

Results

Search Results

We identified 176 potentially eligible items: 122 for persons with osteoarthritis and 54 for health care professionals. We excluded 129 items that did not meet eligibility criteria. This included 79 items for persons with osteoarthritis: content not relevant to osteoarthritis (n=29), duplicate (n=27), no educational content (n=17), focused solely on joint replacement (n=5) or not produced in Canada (n=1); and 50 items for health care professionals: not relevant to osteoarthritis (n=23), blank template or form to record details for patient records (n=10), duplicate (n=8), clinical guideline or standard (n=5), or focused solely on joint replacement (n=4). Ultimately, we included 47 items: 43 for persons with osteoarthritis [47-89] and 4 for health care professionals [90-93].

Characteristics

Multimedia Appendices 2 and 3 provide data on the characteristics of included osteoarthritis materials for persons with osteoarthritis and for health care professionals, respectively. Table 1 summarizes the characteristics of the included items.

XSL•FO
RenderX

Table 1. Summary of characteristics, quality, and cultural safety of 47 included osteoarthritis materials.

Year [reference]	Developer type	Development approach	Osteoarthritis topic	Format	Quality	Cultural safety
Persons with osteoarthritis						
2023 [47]	Patient advocacy	Expert advice	General	Infographic	Low	Low
2023 [48]	Charity	Not reported	General	Text + videos	Moderate	Moderate
2023 [49]	Charity	Not reported	General	Infographic	Low	Low
2023 [50]	Charity	Not reported	General	Text + videos	Moderate	High
2023 [51]	Charity	Not reported	General	Checklist	Moderate	Moderate
2023 [52]	Charity	Not reported	General	Survey + response summary	Moderate	Moderate
2023 [53]	Charity	Not reported	General	Survey	Moderate	Moderate
2023 [54]	Charity	Expert advice	General	Survey + response summary	Moderate	Low
2023 [55]	Charity	Review	General	Survey	High	Low
2023 [56]	Charity	Expert advice	General	Text	Moderate	Low
2023 [57]	Charity	Expert advice	Hip or knee	Booklet	Moderate	High
2023 [58]	Charity	Not reported	General	Infographic	Low	Moderate
2023 [59]	Patient advocacy	Not reported	General	Survey + response summary	High	Low
2023 [60]	Patient advocacy	Key informant interviews	General	Infographic	Low	High
2022 [61]	Charity	Expert advice	General	Video	Low	High
2022 [62]	Consortium	Not reported	Hip or knee	Text	Moderate	High
2021 [63]	Charity	Expert advice	General	Text	Moderate	Moderate
2021 [64]	Charity	Expert advice	General	Text	High	Moderate
2021 [65]	Charity	Expert advice	General	Text	Moderate	Low
2021 [66]	Patient advocacy	Review + expert advice	General	Video	Low	Moderate
2020 [67]	Consortium	Not reported	General	Booklet	Moderate	High
2020 [68]	Charity	Expert advice	Hip or knee + shoulder	Video	Moderate	Moderate
2020 [69]	Charity	Expert advice	General	Video	Low	Low
2020 [70]	Charity	Expert advice	General	Booklet	Moderate	High
2020 [71]	Charity	Expert advice	General	Text	Low	Low
2020 [72]	Charity	Expert advice	General	Text	Moderate	Low
2019 [73]	Charity	Review + expert advice	General	Booklet	High	Low
2019 [74]	Consortium	Not reported	General	Booklet	Moderate	Moderate
2019 [75]	Academia	Expert advice	Hip or knee	Video	Moderate	High
2018 [76]	Charity	Review	General	Booklet	High	Moderate
2018 [77]	Charity	Review	General	Booklet	High	Moderate
2018 [78]	Charity	Review	General	Booklet	High	Moderate
2018 [79]	Charity	Review	General	Booklet	High	Moderate
2018 [80]	Charity	Review	General	Booklet	High	Moderate
2018 [81]	Charity	Review	General	Booklet	High	High
2018 [82]	Charity	Review	General	Booklet	High	Moderate
2018 [83]	Charity	Review	General	Booklet	High	Moderate
2018 [84]	Patient advocacy	Review + expert advice	General	Booklet	Moderate	Moderate

XSL•FO
RenderX

Year [reference]	Developer type	Development approach	Osteoarthritis topic	Format	Quality	Cultural safety
2018 [85]	Government	Expert advice	Hip or knee + hand	Booklet	Moderate	High
2018 [86]	Charity	Not reported	General	Checklist	Low	Moderate
2018 [87]	Government	Not reported	General	Text	High	High
2015 [88]	Charity	Not reported	General	Checklist	Moderate	Moderate
2013 [89]	Consortium	Expert advice	General	Infographic	Low	Moderate
Health care professionals						
2022 [90]	Charity	Expert advice	General	Infographic	Moderate	Low
2022 [91]	Consortium	Expert advice	Knee	Text	Low	Low
2020 [92]	Professional society	Review	General	Infographic	High	Low
2017 [93]	Consortium	Expert advice	Hip or knee + hand	Infographic	High	Low

Persons With Osteoarthritis

In total, 43 materials for persons with osteoarthritis were published between 2013 and 2023 (median 2020). Most were developed by charities (31/43, 72.1%) followed by patient advocacy groups (5/43, 11.6%), consortia (4/43, 9.3%), government (2/43, 4.7%), and academia (1/43, 2.3%). A total of 32.6% (14/43) of materials did not report development methods. Of the remaining 29 that did, most (16, 55.2%) were based on expert advice, 9 (31%) on review of published research, 3 (10.3%) on review of published research and expert advice, and 1 (3.4%) on key informant interviews. The majority (38/43, 88.4%) of materials focused on osteoarthritis in general while 11.6% (5/43) focused on hip and knee osteoarthritis with or without hand or shoulder osteoarthritis. With respect to format, the largest proportion were booklets (15/43, 34.9%) followed by text on web pages with or without other items such as videos (10/43, 23.3%), infographics (5/43, 11.6%), surveys with or without response summaries (5/43, 11.6%), and checklists (3/43, 7%).

Health Care Professionals

A total of 4 materials for health care professionals were published between 2017 and 2022 by consortia (n=2, 50%), a charity (n=1, 25%), and a professional society (n=1, 25%). Methods of development included expert advice (3/4, 75%) and review of published research (1/4, 25%). In total, 50% (2/4) of materials focused on osteoarthritis in general and 50% (2/4) on knee osteoarthritis with or without hip or hand osteoarthritis. A total of 75% (3/4) of materials were infographics, and 25% (1/4) of item consisted of text on web pages.

Quality

Multimedia Appendices 4 and 5 provide data on the quality of included osteoarthritis materials for persons with osteoarthritis and for health care professionals, respectively. Table 1 summarizes the quality of the included items. There did not appear to be associations between osteoarthritis material quality and other characteristics.

Persons With Osteoarthritis

Of the 43 materials for persons with osteoarthritis, 23.3% (n=10), 46.5% (n=20), and 30.2% (n=13) were rated as low, moderate, and high on the DISCERN scale for quality, respectively. An example of a low-scoring item was “What is osteoarthritis” with a score of 26.7% [49]. As a one-page infographic, details about symptoms, diagnosis, and treatment are sparse, with little to no detail about sources of the information, accounting for only 8 of 15 DISCERN criteria that were fully met. An example of a high-scoring item was “Daily living online learning module” with a score of 86.7% [78]. As a 22-page booklet, this item fully met 13 of 15 DISCERN criteria by providing detailed advice on how to achieve a wide range of daily activities and links to sources of information.

Health Care Professionals

Of the 4 materials for health care professionals, 1 (25%) scored low, 1 moderate (25%), and 2 (50%) high on the framework for quality. The low-scoring item, “Knee’d: What to tell patients about knee injections for osteoarthritis”, presented text information on a web page and met only 2 (self-directed and needs-based) of 8 criteria for a score of 25% [91]. An example of a high-scoring item was the “Osteoarthritis tool” with a score of 87.5% [93]. This 8-page infographic met 7 of 8 criteria (self-directed, needs-based, evidence-based, interactive, support for practice, credible, and choice of format), but was not case-based.

Cultural Safety

Multimedia Appendices 6 and 7 provide data on the cultural safety of included osteoarthritis materials for persons with osteoarthritis and for health care professionals, respectively. Table 1 summarizes the cultural safety of the included items. There did not appear to be associations between osteoarthritis material cultural safety and characteristics.

Persons With Osteoarthritis

Of the 43 materials for persons with osteoarthritis, 11 (25.6%), 21 (48.8%), and 11 (25.6%) were rated as low, moderate, and high cultural safety, respectively. An example of a low cultural safety item was “How Can An OT Help People With Arthritis?” [47]. This 1-page infographic met 2 criteria (plain language and empathetic communication) but did not address cultural beliefs and norms or offer solutions to overcome barriers and was not available in languages other than English. An example of a high cultural safety item was “Tips On How To Manage Daily

XSL•FO
RenderX

Cooking Tasks and Live Well With Arthritis" with a score of 80% [60]. This 2-part infographic met 4 of 5 criteria (plain language, empathetic communication, accommodation of cultural beliefs and norms, and solutions to overcome barriers) but was not translated to languages other than English.

Health Care Professionals

Of the 4 materials for health care professionals, all (100%) were rated low for cultural safety. For example, "Conservative OA Treatments–Examples for Providers," a 2-page infographic, did not acknowledge disparities in access to or quality of osteoarthritis care, identify related cultural issues or offer strategies to address cultural beliefs and norms [90].

Discussion

Principal Findings

We assessed the quality and cultural safety of 47 osteoarthritis materials: 43 for persons with osteoarthritis [47-89] and 4 for health care professionals [90-93]. Most were developed by charities, focused on osteoarthritis in general rather than osteoarthritis affecting a specific body part, and in the format of booklets. Many did not describe how they were developed; those that did were most often based on expert advice. Only 30% (13) patient osteoarthritis materials and 50% (2) health care professional materials scored high on the DISCERN scale for quality. Only 26% (11) of patient osteoarthritis materials scored high for cultural safety, and all health care professional osteoarthritis materials scored low for cultural safety. There did not appear to be associations between osteoarthritis material quality or cultural safety with other characteristics.

Strengths and Limitations

This study featured several strengths. We used rigorous content analysis methods and complied with qualitative research reporting criteria [25-27]. To ensure reliability, multiple individuals independently performed searching and screening, and data extraction and analysis. We also used multiple frameworks or instruments to describe osteoarthritis material characteristics [33], quality [4,34,42,43], and cultural safety [35-41,44-46]. An interdisciplinary research team that included ethno-culturally diverse women with osteoarthritis guided all aspects of the study. We must also acknowledge several limitations. The search strategy may not have identified all relevant osteoarthritis materials. Also, because we assessed only publicly available osteoarthritis materials, we may have missed osteoarthritis materials available to health care professionals through individual or organizational subscriptions. We excluded osteoarthritis materials available only in French, Canada's second official language. While we included only English-language materials, we assessed whether those materials were available in languages other than English, which was not the case for any included in this study. We included only osteoarthritis materials developed in Canada. This enhanced study feasibility, but possibly excluded osteoarthritis materials of high quality or cultural safety that may be widely accessible on the internet. However, others could use the frameworks and instruments used here to assess osteoarthritis materials developed in their own countries. With no validated instruments to assess the quality of online materials for health care professionals, and no validated instruments to assess cultural safety of materials for either patients or health care professionals, we relied on sparse available research to compile frameworks by which to evaluate included materials. Future research could expand on and validate these frameworks.

Comparison With Previous Work

Little previous research on support for first-line osteoarthritis management examined the characteristics of online osteoarthritis materials for patients or health care professionals [22-24]. One study that assessed 125 web pages on shoulder arthroplasty reported a low mean of 51.4 (SD 10.7) and a median of 51.5 (IQR 30.5-71.5) DISCERN score [94]. However, the study did not specify that arthroplasty was specific to second-line therapy for advanced osteoarthritis, and did not report which materials were for patients or health care professionals [94]. A meta-analysis of 8 randomized controlled trials showed that digital (eg, telephone, video, internet, and mobile app) self-management programs for persons with osteoarthritis significantly reduced pain and improved physical function at 12 months [95]. While this meta-analysis underscores the benefit of online osteoarthritis materials, it did not assess the quality or cultural safety of those materials. Similarly, another review of 11 randomized controlled trials found that digital (eg, internet or telephone app) self-management support programs for persons with osteoarthritis ranged from 6 to 9 months in duration, largely focused on physical activity, and resulted in program adherence of >70% in studies that reported adherence but did not assess the quality or cultural safety of programs [96]. A systematic review of 25 articles on mobile health technologies (mobile phones or other devices) and an assessment of 23 apps relevant to osteoarthritis self-management focused only on categorizing those technologies by role: self-management, decision support, or shared decision-making [97]. A previous review of online health information for the public found that most resources were of low quality but did not assess cultural safety [6]. Hence, this study is unique because we explicitly assessed online osteoarthritis materials for patients and health care professionals and found the majority of osteoarthritis materials scored low or moderate for quality and cultural safety.

Implications for Practice and Research

Given the high prevalence of osteoarthritis and its debilitating effects [7,8], lack of access to osteoarthritis care among high-risk racialized immigrant women [11-18], lack of high quality, culturally safe osteoarthritis materials for persons with osteoarthritis or health care professionals revealed by this study, and pan-Canadian recommendations for improved access to publicly available osteoarthritis materials [20,21], greater efforts are needed to develop such materials. With respect to development methods, few osteoarthritis materials were explicitly based on published research, relying instead on expert advice, but it was not clear in all cases whether "experts" included persons with osteoarthritis. This knowledge could be applied to the improvement of existing osteoarthritis materials or the development of new materials. Because of the long-standing recognized importance of evidence-informed decision-making, the rubrics we used to assess quality (eg,

XSL•FO
RenderX

DISCERN for patient materials and frameworks we compiled for health care professional material) included explicit identification of the underlying evidence and sources of data [4,34,42,43]. Furthermore, meaningfully engaging patients, family or caregivers, or members of the public in planning or developing health care policy, services, interventions, or tools is increasingly encouraged and common [98], and numerous resources are available to provide guidance on how to optimize their engagement [99]. Additional insight could be borrowed from methods used to engage persons with lived experience in developing other types of tools such as decision aids for patients and clinical guidelines for health care professionals [100,101]. Developers may want to give extra attention to quality domains that were frequently not met by osteoarthritis materials included in this study; for example, sources of information and benefits and harms of treatment for patient materials; and case-based and interactive for health care professional material. While we compiled cultural safety frameworks for both patient and health care professional osteoarthritis materials based on our research and that of others [35-41,44-46], none were specific to osteoarthritis, therefore further research is needed to establish what constitutes cultural safety in the context of osteoarthritis.

Conclusions

We aimed to assess the quality and cultural safety of online osteoarthritis informational or educational materials for persons with osteoarthritis and health care professionals. Of the 47 osteoarthritis materials (43 for persons with osteoarthritis, 4 for health care professionals), most were developed by charities, focused on osteoarthritis in general rather than osteoarthritis affecting a specific body part, and in the format of booklets. Many did not describe how they were developed; those that did were most often based on expert advice rather than explicit evidence. Few osteoarthritis materials for persons with osteoarthritis or health care professionals scored high for quality. Few osteoarthritis materials for persons with osteoarthritis and none for health care professionals scored high for cultural safety. Quality and cultural safety did not appear to be associated with osteoarthritis material characteristics. These findings reveal the need for further efforts to improve existing or develop new osteoarthritis materials for both persons with osteoarthritis and health care professionals. Further research is needed to assess the quality and cultural safety of osteoarthritis materials developed by organizations outside of Canada and to establish a framework or instrument to assess cultural safety in the osteoarthritis context.

Acknowledgments

The authors express their gratitude to the Osteoarthritis Study Women Advisory Group, consisting of AM, AN, JC, JT, KC, LR, MH, MT, SR, TM, VI, and ZL, as well as our collaborative partners for their valuable input and feedback: Arthritis Consumer Experts, Canadian Rheumatology Association, Canadian Arthritis Patient Alliance, Arthritis Health Professions Association, Neighborhood Pharmacy Association of Canada, Bone and Joint Canada, and the University Health Network Schroeder Arthritis Institute. The Arthritis Society (grant 21-0000000145) funded this research and took no part in conducting the research, drafting the manuscript, or deciding where to submit it.

Data Availability

All data generated or analyzed during this study are included in the published article and supplementary information files.

Authors' Contributions

ARG conceived and acquired funding for this study and supervised all aspects of its conduct. VA, MB, CMB, GSH, AL, CM, and DAM contributed to conceiving and planning the study and reviewed all data. SD and SM assisted in collecting and analyzing data. All authors drafted the manuscript and approved this final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Eligibility criteria.
[DOCX File , 22 KB - jmir_v26i1e57698_app1.docx]

Multimedia Appendix 2
Characteristics of materials for persons with osteoarthritis.
[DOCX File , 84 KB - jmir_v26i1e57698_app2.docx]

Multimedia Appendix 3
Characteristics of materials for health care professionals.
[DOCX File , 42 KB - jmir_v26i1e57698_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Quality of materials for persons with osteoarthritis.
[DOCX File , 48 KB - jmir_v26i1e57698_app4.docx]

Multimedia Appendix 5
Quality of materials for health care professionals.
[DOCX File , 22 KB - jmir_v26i1e57698_app5.docx]

Multimedia Appendix 6
Cultural safety of materials for persons with osteoarthritis.
[DOCX File , 21 KB - jmir_v26i1e57698_app6.docx]

Multimedia Appendix 7
Cultural safety of materials for health care professionals.
[DOCX File , 20 KB - jmir_v26i1e57698_app7.docx]

References

1. Wollmann K, der Keylen PV, Tomandl J, Meerpohl JJ, Sofroniou M, Maun A, et al. The information needs of internet users and their requirements for online health information-a scoping review of qualitative and quantitative studies. Patient Educ Couns 2021;104(8):1904-1932. [doi: 10.1016/j.pec.2021.01.020] [Medline: 33563502]
2. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. The influence of online health information on health decisions: a systematic review. Patient Educ Couns 2021;104(4):770-784. [doi: 10.1016/j.pec.2020.11.016] [Medline: 33358253]
3. Daei A, Soleymani MR, Ashrafi-Rizi H, Zargham-Boroujeni A, Kelishadi R. Clinical information seeking behavior of physicians: a systematic review. Int J Med Inform 2020;139:104144. [doi: 10.1016/j.ijmedinf.2020.104144] [Medline: 32334400]
4. Cheng C, Papadakos J, Umakanthan B, Fazelzad R, Martimianakis MAT, Ugas M, et al. On the advantages and disadvantages of virtual continuing medical education: a scoping review. Can Med Educ J 2023;14(3):41-74 [FREE Full text] [doi: 10.36834/cmej.75681] [Medline: 37465741]
5. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
6. Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the Internet. J Gen Intern Med 2019;34(9):1884-1891 [FREE Full text] [doi: 10.1007/s11606-019-05109-0] [Medline: 31228051]
7. Katz JN, Arant KR, Loeser RF. Diagnosis and treatment of hip and knee osteoarthritis: a review. JAMA 2021;325(6):568-578 [FREE Full text] [doi: 10.1001/jama.2020.22171] [Medline: 33560326]
8. Long H, Liu Q, Yin H, Wang K, Diao N, Zhang Y, et al. Prevalence trends of site-specific osteoarthritis from 1990 to 2019: findings from the global burden of disease study 2019. Arthritis Rheumatol 2022;74(7):1172-1183 [FREE Full text] [doi: 10.1002/art.42089] [Medline: 35233975]
9. Bannuru RR, Osani MC, Vaysbrot EE, Arden NK, Bennell K, Bierma-Zeinstra SMA, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. Osteoarthritis Cartilage 2019;27(11):1578-1589 [FREE Full text] [doi: 10.1016/j.joca.2019.06.011] [Medline: 31278997]
10. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American college of rheumatology/arthritis foundation guideline for the management of osteoarthritis of the hand, hip, and knee. Arthritis Care Res (Hoboken) 2020;72(2):149-162 [FREE Full text] [doi: 10.1002/acr.24131] [Medline: 31908149]
11. Goldring AE, Ashok AP, Casey EK, Mulcahey MK. Key components and potential benefits of a comprehensive approach to women's musculoskeletal health. Phys Sportsmed 2016;44(4):417-424. [doi: 10.1080/00913847.2016.1222854] [Medline: 27548840]
12. Schwartzberg HG, Roy R, Wilson K, Starring H, Leonardi C, Bronstone A, et al. Patient characteristics independently associated with knee osteoarthritis symptom severity at initial orthopedic consultation. J Clin Rheumatol 2022;28(2):e359-e362. [doi: 10.1097/RHU.0000000000001726] [Medline: 33657591]
13. Odonkor CA, Esparza R, Flores LE, Verduzco-Gutierrez M, Escalon MX, Solinsky R, et al. Disparities in health care for black patients in physical medicine and rehabilitation in the United States: a narrative review. PM R 2021;13(2):180-203. [doi: 10.1002/pmrj.12509] [Medline: 33090686]
14. Rabah NM, Knusel KD, Khan HA, Marcus RE. Are there nationwide socioeconomic and demographic disparities in the use of outpatient orthopaedic services? Clin Orthop Relat Res 2020;478(5):979-989 [FREE Full text] [doi: 10.1097/CORR.0000000000001168] [Medline: 32310622]
15. Gagliardi AR, Kim C, Jameel B. Physician behaviours that optimize patient-centred care: focus groups with migrant women. Health Expect 2020;23(5):1280-1288 [FREE Full text] [doi: 10.1111/hex.13110] [Medline: 32707600]

XSL•FO
RenderX

16. Egerton T, Diamond LE, Buchbinder R, Bennell KL, Slade SC. A systematic review and evidence synthesis of qualitative studies to identify primary care clinicians' barriers and enablers to the management of osteoarthritis. Osteoarthritis Cartilage 2017;25(5):625-638 [FREE Full text] [doi: 10.1016/j.joca.2016.12.002] [Medline: 27939622]
17. Bouma SE, van Beek JFE, Diercks RL, van der Woude LHV, Stevens M, van den Akker-Scheek I. Barriers and facilitators perceived by healthcare professionals for implementing lifestyle interventions in patients with osteoarthritis: a scoping review. BMJ Open 2022;12(2):e056831 [FREE Full text] [doi: 10.1136/bmjopen-2021-056831] [Medline: 35105598]
18. Pan-Canadian Health Inequalities Data Tool. A Joint Initiative of the Public Health Agency of Canada, the Pan-Canadian Public Health Network, Statistics Canada and the Canadian Institute of Health Information. 2017. URL: https://health-infobase.canada.ca/health-inequalities/data-tool/index [accessed 2023-05-03]
19. Yu J, Meng S. Impacts of the Internet on health inequality and healthcare access: a cross-country study. Front Public Health 2022;10:935608 [FREE Full text] [doi: 10.3389/fpubh.2022.935608] [Medline: 35757602]
20. Gagliardi AR, Abbaticchio A, Theodorlis M, Marshall D, MacKay C, Borkhoff CM, et al. Multi-level strategies to improve equitable timely person-centred osteoarthritis care for diverse women: qualitative interviews with women and healthcare professionals. Int J Equity Health 2023;22(1):207 [FREE Full text] [doi: 10.1186/s12939-023-02026-x] [Medline: 37803475]
21. Iziduh S, Abenoja A, Theodorlis M, Ahluwalia V, Battistella M, Borkhoff CM, et al. Priority strategies to reduce socio-gendered inequities in access to person-centred osteoarthritis care: delphi survey. BMJ Open 2024;14(2):e080301 [FREE Full text] [doi: 10.1136/bmjopen-2023-080301] [Medline: 38373862]
22. Uritani D, Koda H, Sugita S. Effects of self-management education programmes on self-efficacy for osteoarthritis of the knee: a systematic review of randomised controlled trials. BMC Musculoskelet Disord 2021;22(1):515 [FREE Full text] [doi: 10.1186/s12891-021-04399-y] [Medline: 34090406]
23. Huang J, Han Y, Wei J, Liu X, Du Y, Yang L, et al. The effectiveness of the Internet-based self-management program for cancer-related fatigue patients: a systematic review and meta-analysis. Clin Rehabil 2020;34(3):287-298. [doi: 10.1177/0269215519889394] [Medline: 31793340]
24. Pereira K, Phillips B, Johnson C, Vorderstrasse A. Internet delivered diabetes self-management education: a review. Diabetes Technol Ther 2015;17(1):55-63. [doi: 10.1089/dia.2014.0155] [Medline: 25238257]
25. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
26. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
27. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
28. Arthritis alliance of Canada. Joint action on Arthritis: A Framework to Improve Arthritis Prevention and care in Canada. 2012. URL: https://www.arthritisalliance.ca/en/initiativesen/joint-action [accessed 2023-05-03]
29. Benzies KM, Premji S, Hayden KA, Serrett K. State-of-the-evidence reviews: advantages and challenges of including grey literature. Worldviews Evid Based Nurs 2006;3(2):55-61. [doi: 10.1111/j.1741-6787.2006.00051.x] [Medline: 17040510]
30. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, et al. Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. Syst Rev 2016;5(1):164 [FREE Full text] [doi: 10.1186/s13643-016-0337-y] [Medline: 27686611]
31. Abuwa C, Abbaticchio A, Theodorlis M, Marshall D, MacKay C, Borkhoff CM, et al. Identifying strategies that support equitable person-centred osteoarthritis care for diverse women: content analysis of guidelines. BMC Musculoskelet Disord 2023;24(1):734 [FREE Full text] [doi: 10.1186/s12891-023-06877-x] [Medline: 37710195]
32. Abbaticchio A, Theodorlis M, Marshall D, MacKay C, Borkhoff CM, Hazlewood GS, et al. Policies in Canada fail to address disparities in access to person-centred osteoarthritis care: a content analysis. BMC Health Serv Res 2024;24(1):522 [FREE Full text] [doi: 10.1186/s12913-024-10966-5] [Medline: 38664819]
33. Albrecht L, Archibald M, Arseneau D, Scott SD. Development of a checklist to assess the quality of reporting of knowledge translation interventions using the workgroup for intervention development and evaluation research (WIDER) recommendations. Implement Sci 2013;8:52 [FREE Full text] [doi: 10.1186/1748-5908-8-52] [Medline: 23680355]
34. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
35. Alvi S, Zaidi A, Ammar N, Culbert L. A comparative and exploratory analysis of socio-cultural factors and immigrant women's mental health within a Canadian context. J Immigr Minor Health 2012;14(3):420-432. [doi: 10.1007/s10903-011-9567-1] [Medline: 22223121]
36. Brevik TB, Laake P, Bjørkly S. Effect of culturally tailored education on attendance at mammography and the Papanicolaou test. Health Serv Res 2020;55(3):457-468 [FREE Full text] [doi: 10.1111/1475-6773.13271] [Medline: 31994187]
37. de Sequeira S, Halperin I, Lipscombe LL. Culturally tailored resources for South Asian immigrant women with gestational diabetes: do they work and what's missing? A qualitative study. Can J Diabetes 2019;43(8):573-579. [doi: 10.1016/j.jcjd.2019.09.007] [Medline: 31787243]

XSL•FO
RenderX

38. Hawa RN, Underhill A, Logie CH, Islam S, Loutfy M. South Asian immigrant women's suggestions for culturally-tailored HIV education and prevention programs. Ethn Health 2019;24(8):945-959. [doi: 10.1080/13557858.2017.1378806] [Medline: 28922011]
39. Lasch KE, Wilkes G, Montuori LM, Chew P, Leonard C, Hilton S. Using focus group methods to develop multicultural cancer pain education materials. Pain Manag Nurs 2000;1(4):129-138. [doi: 10.1053/jpmn.2000.18743] [Medline: 11709866]
40. Jones CA, Mawani S, King KM, Allu SO, Smith M, Mohan S, et al. Tackling health literacy: adaptation of public hypertension educational materials for an indo-Asian population in Canada. BMC Public Health 2011;11:24 [FREE Full text] [doi: 10.1186/1471-2458-11-24] [Medline: 21223580]
41. Iziduh S, Dhakal S, Sihota R, Ye X, Friedenreich CM, Khanlou N, et al. Raising awareness about physical activity's role in reducing cancer risk: qualitative interviews with immigrant women and community agency managers. BMC Public Health 2024;24(1):2072 [FREE Full text] [doi: 10.1186/s12889-024-19612-5] [Medline: 39085836]
42. Shortt SED, Guillemette JM, Duncan AM, Kirby F. Defining quality criteria for online continuing medical education modules using modified nominal group technique. J Contin Educ Health Prof 2010;30(4):246-250. [doi: 10.1002/chp.20089] [Medline: 21171030]
43. Young KJ, Kim JJ, Yeung G, Sit C, Tobe SW. Physician preferences for accredited online continuing medical education. J Contin Educ Health Prof 2011;31(4):241-246. [doi: 10.1002/chp.20136] [Medline: 22189987]
44. Betancourt JR, Corbett J, Bondaryk MR. Addressing disparities and achieving equity: cultural competence, ethics, and health-care transformation. Chest 2014;145(1):143-148. [doi: 10.1378/chest.13-0634] [Medline: 24394825]
45. Lin CJ, Lee CK, Huang MC. Cultural competence of healthcare providers: a systematic review of assessment instruments. J Nurs Res 2017;25(3):174-186. [doi: 10.1097/JNR.0000000000000153] [Medline: 28481813]
46. Like RC. Educating clinicians about cultural competence and disparities in health and health care. J Contin Educ Health Prof 2011;31(3):196-206. [doi: 10.1002/chp.20127] [Medline: 21953661]
47. How Can an OT Help People with Arthritis? Canadian Arthritis Patient Alliance. 2023. URL: https://arthritispatient.ca/wp-content/uploads/2023/09/Occupational-Therapy-infographic.pdf [accessed 2023-05-18]
48. Osteoarthritis. Arthritis Society Canada. 2023. URL: https://arthritis.ca/about-arthritis/arthritis-types-(a-z)/types/osteoarthritis [accessed 2023-05-03]
49. What is osteoarthritis?. Arthritis Society Canada. 2023. URL: https://arthritis-society.azureedge.net/media/as/media/pdf/whatisoa_infographic.pdf [accessed 2023-05-03]
50. Complementary Therapies Guide for Arthritis. Arthritis Society Canada. 2023. URL: https://arthritis.ca/treatment/pain-management/complementary-therapies [accessed 2023-05-03]
51. Appointment Discussion Guide. Arthritis Society. 2023. URL: https://arthritis.ca/getmedia/d595f5ae-688e-44a0-8b7b-01ce2dadc698/AppointmentDiscussionGuide_EN_Jan10B.pdf [accessed 2023-05-07]
52. Arthritis Symptom Checker. Arthritis Society Canada. 2023. URL: https://arthritis.ca/about-arthritis/signs-of-arthritis/symptom-checker/ [accessed 2023-05-03]
53. Daily Symptom Tracker. Arthritis Society Canada. 2023. URL: https://arthritis.ca/AS/media/pdf/Support%20and%20Education/EN-daily-symptom-tracker.pdf?[assessed [accessed 2023-05-04]
54. Arthritis Risk Factor Assessment. Arthritis Society Canada. 2023. URL: https://arthritis.ca/about-arthritis/arthritis-risk-factors/arthritis-risk-factor-assessment [accessed 2023-05-03]
55. Joint Pain Symptom Checker. Arthritis Society Canada. 2023. URL: https://arthritis.ca/getmedia/b4fac3b9-8b48-4ec1-ac38-6bba4e2f2322/EN-joint-symptom-tracker-2023.pdf [accessed 2023-05-03]
56. Medication Reference Guide. Arthritis Society Canada. 2023. URL: https://arthritis.ca/treatment/medication/medication-reference-guide?atype=694c7b0f-c2fb-4afc-be72-0b4924263294&atype [accessed 2023-05-07]
57. Low load Activities for Osteoarthritis. Arthritis Society Canada. 2023. URL: https://arthritis.ca/getmedia/28589053-8d9d-4860-892c-b236e5ced8b3/LowLoad_Activities_EN-Jun1_small.pdf [accessed 2023-05-17]
58. Modifying activities for osteoarthritis. Arthritis Society Canada. 2023. URL: https://arthritis.ca/getmedia/a142085b-aa62-4842-be5a-41a972c4fd44/Activity_Modifications_EN-Jan20_small.pdf [accessed 2023-05-17]
59. Arthritis Screening Exam. Arthritis Consumer Expert and Arthritis Research Canada. 2023. URL: https://arthritisbroadcastnetwork.org/arthritis-screening/ [accessed 2023-05-17]
60. Tips on how to manage daily cooking and live well with arthritis. Canadian Arthritis Patient Alliance. 2023. URL: https://arthritispatient.ca/tips-on-how-to-manage-daily-cooking-tasks-and-live-well-with-arthritis/ [accessed 2023-05-17]
61. Tai Chi To Help With Arthritis. Arthritis Society Canada. 2022. URL: https://www.youtube.com/watch?v=qj0ft1CQC_4 [accessed 2023-05-18]
62. Osteoarthritis. GLA:D Canada. 2022. URL: https://gladcanada.ca/osteoarthritis [accessed 2023-05-19]
63. Osteoarthritis Symptoms and Diagnosis. Arthritis Society Canada. 2021. URL: https://arthritis.ca/about-arthritis/arthritis-types-(a-z)/types/osteoarthritis/osteoarthritis-symptoms-and-diagnosis [accessed 2023-05-17]
64. Osteoarthritis Self-Management. Arthritis Society Canada. 2021. URL: https://arthritis.ca/about-arthritis/arthritis-types-(a-z)/types/osteoarthritis/osteoarthritis-self-management [accessed 2023-05-17]
65. Osteoarthritis Treatment. Arthritis Society Canada. 2021. URL: https://arthritis.ca/about-arthritis/arthritis-types-(a-z)/types/osteoarthritis/osteoarthritis-treatment [accessed 2023-05-17]

XSL•FO
RenderX

66. Updates in osteoarthritis treatment and care with Dr. Tom Appleton. Joint Health:Arthritis Consumer Experts. 2021. URL: https://arthritisathome.jointhealth.org/?p=2286 [accessed 2023-05-01]
67. Living your best life with osteoarthritis. Arthritis Society Canada, Alberta Health Services, Alberta Bone and Joint Health Institute. 2021. URL: https://www.albertahealthservices.ca/assets/about/scn/ahs-scn-bjh-osteoarthritis-toolkit.pdf [accessed 2023-05-17]
68. Introduction to Exercise for Osteoarthritis. Arthritis Society Canada. 2021. URL: https://arthritis.ca/living-well/2020/introduction-to-exercise-for-osteoarthritis [accessed 2023-05-05]
69. 20-Minute Warm-Up for the Joints. Arthritis Society Canada. 2020. URL: https://www.youtube.com/watch?v=X15pa33QO-o&t=2s [accessed 2023-05-05]
70. OA: Patient Journey. Arthritis Society Canada. 2020. URL: https://arthritis.ca/getmedia/26457d6d-41cf-4c7f8fbbb26f75b88781/ [accessed 2023-05-03]
71. Assistive Devices Resource. Arthritis Society Canada. 2020. URL: https://arthritis.ca/treatment/pain-management/assistive-devices-resource [accessed 2023-05-15]
72. Drug Free Pain Management Tool. Arthritis Society Canada. 2020. URL: https://arthritis.ca/treatment/pain-management/drug-free-pain-management-tool [accessed 2023-05-07]
73. Medical Cannabis and Arthritis. Arthritis Society Canada. 2019. URL: https://arthritis.ca/getmedia/f43c06f9-9e25-4524-a22a-18b1666d3d09/Medical-Cannabis-Guide-EN.pdf [accessed 2023-05-04]
74. Talk to your Doctor about Joint Pain. Arthritis Society Canada. 2019. URL: https://arthritis.ca/getmedia/a2609338-226f-4bbe-8786a4fd6f35f7ec/Talk_to_Your_Doctor_about_Joint_Pain_Handout_letter_EN [accessed 2023-05-03]
75. Exercises for arthritis. University of Alberta. 2019. URL: https://www.ualberta.ca/glen-sather-clinic/patient-resources/exercises-for-arthritis.html [accessed 2023-05-03]
76. Managing Chronic Pain Online Learning Module. Arthritis Society Canada. 2018. URL: https://arthritis.ca/getmedia/a70cd03b-e2e4-4ab9-bcc6-40ddecb9882d/Managing-Chronic-Pain-Final.pdf [accessed 2023-05-04]
77. Overcoming Fatigue Online Learning Module. Arthritis Society Canada. 2018. URL: https://arthritis.ca/getmedia/175ca269-5551-42ad-8ab6-9d22bff0eea8/Overcoming-Fatigue-Final.pdf [accessed 2023-05-04]
78. Daily Living Online Learning Module. Arthritis Society Canada. 2018. URL: https://arthritis.ca/getmedia/9f3c6e33-9390-4ae7-9c70-64c786099a78/Daily-Living-Final.pdf [accessed 2023-05-03]
79. Arthritis Society Canada. Mental Health & Well-being Online Learning Module. 2018. URL: https://arthritis.ca/getmedia/1c7b3e72-cc2e-4dbd-b571-2c7482a8d849/Mental-Health-and-Wellbeing-Final.pdf [accessed 2023-05-03]
80. Arthritis Society Canada. Arthritis and Work Online Learning Module. 2018. URL: https://arthritis.ca/getmedia/0f374d04-eede-449d-99be-993a7d17b0f1/Arthritis-and-Work-Final.pdf [accessed 2023-05-03]
81. Arthritis Society Canada. Staying active Online Learning Module. 2018. URL: https://arthritis.ca/getmedia/743fa1e2-a69d-4e5b-91c1-26e7748593cc/Staying-Active-Final.pdf [accessed 2023-05-03]
82. Arthritis society Canada. Eating Well online Learning Module. 2018. URL: https://arthritis.ca/getmedia/3a67574b-d006-420e-9cbf-cb85d4ef613f/Eating-Well-Final.pdf [accessed 2023-05-04]
83. Arthritis Society Canada. Navigating your Healthcare online Learning Module?. 2018. URL: https://arthritis.ca/getmedia/0aaa21d2-bf16-472e-b96d-12cd9af04474/Navigating-your-Healthcare-Final.pdf [accessed 2023-05-04]
84. Arthritis society Canada, arthritis research Canada, arthritis consumer experts. A Guide to living with osteoarthritis. 2018. URL: https://www.jointhealth.org/pdfs/JointHealth-insight-2018-September.pdf [accessed 2023-05-17]
85. Patient Reference Guide. Health Quality Ontario. 2020. URL: https://www.hqontario.ca/Portals/0/documents/evidence/quality-standards/qs-osteoarthritis-patient-guide-en.pdf [accessed 2023-05-18]
86. Activity Diary. Arthritis Society. 2018. URL: https://arthritis.ca/AS/media/pdf/Support%20and%20Education/EN-activity-diary.pdf [accessed 2023-05-04]
87. Osteoarthritis. Alberta Health Services. 2018. URL: https://www.healthiertogether.ca/health-conditions/osteoarthritis/ [accessed 2023-05-04]
88. Healthcare Appointment Checklist?. Arthritis Society Canada. 2015. URL: https://arthritis.ca/AS/media/pdf/Support%20and%20Education/EN-appointment-checklist.pdf [accessed 2023-05-04]
89. Rheum Info. NSAIDs & Osteoarthritis: Putting Risk Into Perspective. 2013. URL: https://rheuminfo.com/docs/living-with-arthritis/NSAIDs-and-OA-Putting-Risks-Into-Perspective-RheumInfo_EN.pdf [accessed 2023-05-04]
90. Conservative OA treatments-examples for providers. Alberta Bone & Joint Institute. 2022. URL: https://toolkit.albertaboneandjoint.com/osteoarthritis/wp-content/uploads/7_Tables-of-Examples-for-the-Treatments-Toolbox-4.pdf [accessed 2023-05-04]
91. Taggart k. Knee'd: What to tell patient about knee injections for osteoarthritis. Canadian Healthcare Network. 2022. URL: https://canadianhealthcarenetwork.ca/kneed-what-tell-patients-about-knee-injections-osteoarthritis [accessed 2023-05-01]
92. Lindblad AJ, McCormack J, Korownyk CS, Kolber MR, Ton J, Perry D. PEER simplified decision aid: osteoarthritis treatment options in primary care. The College of Family Physicians of Canada. 2020. URL: https://www.cfp.ca/content/cfp/66/3/191.full.pdf [accessed 2023-05-01]
93. Arthritis Alliance of Canada, Center for effective practice, The college of Family Physicians of Canada. Osteoarthritis Tool. 2017. URL: https://jointhealth.org/pdfs/OATookKit_En.pdf [accessed 2023-04-19]

XSL•FO
RenderX

94. Fares MY, Singh J, Vadhera AS, Koa J, Boufadel P, Abboud JA. Online resources for information on shoulder arthroplasty: an assessment of quality and readability. Clin Shoulder Elb 2023;26(3):238-244 [FREE Full text] [doi: 10.5397/cise.2023.00290] [Medline: 37607858]
95. Safari R, Jackson J, Sheffield D. Digital self-management interventions for people with osteoarthritis: systematic review with meta-analysis. J Med Internet Res 2020;22(7):e15365 [FREE Full text] [doi: 10.2196/15365] [Medline: 32706657]
96. Patten RK, Tacey A, Pile R, Parker A, De Gori M, Tran P, et al. Digital self-management interventions for osteoarthritis: a systematic scoping review of intervention characteristics, adherence and attrition. Arch Public Health 2022;80(1):103 [FREE Full text] [doi: 10.1186/s13690-022-00854-x] [Medline: 35361270]
97. Choi W, Zheng H, Franklin P, Tulu B. mHealth technologies for osteoarthritis self-management and treatment: a systematic review. Health Informatics J 2019;25(3):984-1003 [FREE Full text] [doi: 10.1177/1460458217735676] [Medline: 29090628]
98. Bombard Y, Baker GR, Orlando E, Fancott C, Bhatia P, Casalino S, et al. Engaging patients to improve quality of care: a systematic review. Implement Sci 2018;13(1):98 [FREE Full text] [doi: 10.1186/s13012-018-0784-z] [Medline: 30045735]
99. Anderson NN, Gagliardi AR. Patient and Family Engagement in Hospital Planning and Improvement: A casebook.: A casebook; 2021. URL: https://www.uhnresearch.ca/sites/default/files/AGagliardi_CASEBOOK-FINAL.pdf [accessed 2023-05-23]
100. Witteman HO, Maki KG, Vaisson G, Finderup J, Lewis KB, Dahl Steffensen K, et al. Systematic development of patient decision aids: an update from the IPDAS collaboration. Med Decis Making 2021;41(7):736-754 [FREE Full text] [doi: 10.1177/0272989X211014163] [Medline: 34148384]
101. Kim C, Berta WB, Gagliardi AR. Exploring approaches to identify, incorporate and report patient preferences in clinical guidelines: qualitative interviews with guideline developers. Patient Educ Couns 2021;104(4):703-708 [FREE Full text] [doi: 10.1016/j.pec.2020.10.001] [Medline: 33059950]

Edited by N Cahill; submitted 23.02.24; peer-reviewed by J Mistry, C Lirio-Romero, A Jain; comments to author 19.04.24; revised version received 19.04.24; accepted 21.08.24; published 18.10.24.

Please cite as:
Dhakal S, Merani S, Ahluwalia V, Battistella M, Borkhoff CM, Hazlewood GS, Lofters A, Marshall DA, MacKay C, Gagliardi AR
The Quality and Cultural Safety of Online Osteoarthritis Information for Affected Persons and Health Care Professionals: Content Analysis
J Med Internet Res 2024;26:e57698
URL: https://www.jmir.org/2024/1/e57698
doi:10.2196/57698
PMID:39422989

©Smita Dhakal, Shermeen Merani, Vandana Ahluwalia, Marisa Battistella, Cornelia M Borkhoff, Glen Stewart Hazlewood, Aisha Lofters, Deborah A Marshall, Crystal MacKay, Anna R Gagliardi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Role of the Generative Pretrained Transformer (GPT) in Alzheimer's Disease Management: Comparative Study of Neurologist- and Artificial Intelligence–Generated Responses

Jiaqi Zeng[1*], BMed; Xiaoyi Zou[2,3*], MD; Shirong Li[4], MD; Yao Tang[3], MMed; Sisi Teng[3], BMed; Huanhuan Li[5], MSN; Changyu Wang[6], BSc; Yuxuan Wu[7], MMed; Luyao Zhang[8], BSN; Yunheng Zhong[8], BSN; Jialin Liu[7,9], MD; Siru Liu[10], PhD

[1]West China Medical School, Sichuan University, Chengdu, China
[2]Department of Neurology, West China Hospital, Sichuan University, Chengdu, China
[3]Department of Neurology, Chengdu Shangjin Nanfu Hospital, Chengdu, China
[4]Department of Neurology, Guizhou Provincial People's Hospital, Guiyang, China
[5]Mental Health Center, West China Hospital, Sichuan University, Chengdu, China
[6]West China College of Stomatology, Sichuan University, Chengdu, China
[7]Department of Medical Informatics, West China Medical School, Chengdu, China
[8]West China School of Nursing, Sichuan University, Chengdu, China
[9]Department of Otolaryngology-Head and Neck Surgery, West China Hospital, Sichuan University, Chengdu, China
[10]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Department of Medical Informatics
West China Medical School
No 37 Guoxue Road
Chengdu, 610041
China
Phone: 86 28 85422306
Email: Dljl8@163.com

Abstract

Background: Alzheimer's disease (AD) is a progressive neurodegenerative disorder posing challenges to patients, caregivers, and society. Accessible and accurate information is crucial for effective AD management.

Objective: This study aimed to evaluate the accuracy, comprehensibility, clarity, and usefulness of the Generative Pretrained Transformer's (GPT) answers concerning the management and caregiving of patients with AD.

Methods: In total, 14 questions related to the prevention, treatment, and care of AD were identified and posed to GPT-3.5 and GPT-4 in Chinese and English, respectively, and 4 respondent neurologists were asked to answer them. We generated 8 sets of responses (total 112) and randomly coded them in answer sheets. Next, 5 evaluator neurologists and 5 family members of patients were asked to rate the 112 responses using separate 5-point Likert scales. We evaluated the quality of the responses using a set of 8 questions rated on a 5-point Likert scale. To gauge comprehensibility and participant satisfaction, we included 3 questions dedicated to each aspect within the same set of 8 questions.

Results: As of April 10, 2023, the 5 evaluator neurologists and 5 family members of patients with AD rated the 112 responses: GPT-3.5: n=28, 25%, responses; GPT-4: n=28, 25%, responses; respondent neurologists: 56 (50%) responses. The top 5 (4.5%) responses rated by evaluator neurologists had 4 (80%) GPT (GPT-3.5+GPT-4) responses and 1 (20%) respondent neurologist's response. For the top 5 (4.5%) responses rated by patients' family members, all but the third response were GPT responses. Based on the evaluation by neurologists, the neurologist-generated responses achieved a mean score of 3.9 (SD 0.7), while the GPT-generated responses scored significantly higher (mean 4.4, SD 0.6; $P<.001$). Language and model analyses revealed no significant differences in response quality between the GPT-3.5 and GPT-4 models (GPT-3.5: mean 4.3, SD 0.7; GPT-4: mean 4.4, SD 0.5; $P=.51$). However, English responses outperformed Chinese responses in terms of comprehensibility (Chinese

XSL•FO
RenderX

responses: mean 4.1, SD 0.7; English responses: mean 4.6, SD 0.5; *P*=.005) and participant satisfaction (Chinese responses: mean 4.2, SD 0.8; English responses: mean 4.5, SD 0.5; *P*=.04). According to the evaluator neurologists' review, Chinese responses had a mean score of 4.4 (SD 0.6), whereas English responses had a mean score of 4.5 (SD 0.5; *P*=.002). As for the family members of patients with AD, no significant differences were observed between GPT and neurologists, GPT-3.5 and GPT-4, or Chinese and English responses.

Conclusions: GPT can provide patient education materials on AD for patients, their families and caregivers, nurses, and neurologists. This capability can contribute to the effective health care management of patients with AD, leading to enhanced patient outcomes.

(J Med Internet Res 2024;26:e51095) doi:10.2196/51095

KEYWORDS

Alzheimer's disease; artificial intelligence; AI; large language model; LLM; Generative Pretrained Transformer; GPT; ChatGPT; patient information

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that has emerged as a significant health challenge worldwide. AD International reports indicate that the US population aged 65 years and older is predicted to surge from 58 million in 2021 to 88 million by 2050 [1,2]. As the most common form of dementia, AD accounts for 60%-80% of cases and significantly affects patients, caregivers, and society [3-5]. Given the disease's progressive nature, efficient management of AD and related cognitive decline is crucial for enhancing the life quality of patients and caregivers [6].

Optimally managing AD often requires a multifaceted approach to address the cognitive, functional, and behavioral symptoms associated with the disorder [7]. Current pharmacological interventions, such as cholinesterase inhibitors and *N*-methyl-D-aspartate receptor antagonists, deliver moderate symptomatic relief but fail to halt disease progression [8-10]. Meanwhile, nonpharmacological interventions, such as cognitive stimulation, physical exercise, and caregiver support programs, have shown potential benefits for patients with AD [11-13]. Patients diagnosed with AD and their caregivers frequently grapple with unmet needs and inadequate awareness about managing the disease and its complications. A lack of health literacy among patients with AD and caregivers has been reported [14,15]. The internet can offer valuable health information, but the complex nature of the primary literature and potential misinformation can often lead to more confusion than clarity [16]. Studies show that online resources provided by dementia-focused organizations are often lengthy and complex, demonstrating a scarcity of easily understandable information for this demographic [17].

Improving health literacy and empowering patients and caregivers about AD require the provision of precise, accessible, and holistic resources that cover both pharmacological and nonpharmacological interventions. Customizing these resources to meet the unique needs of individuals and caregivers at different disease stages is vital, given potential variations in management strategies based on AD severity and progression [18,19]. The increasing prevalence of AD poses a significant challenge for patients, caregivers, and society at large. Effective disease management is paramount to enhance the life quality of those affected. By boosting health literacy and providing accessible, precise, and comprehensive AD management and treatment information, we can empower patients and caregivers, which will ultimately lead to better outcomes for those living with the disease [20].

The Generative Pretrained Transformer (GPT), developed by OpenAI, is an advanced natural language processing (NLP) model based on the GPT-3.5 architecture. It has been refined through supervised learning, human feedback, and reinforcement learning techniques [21]. Released on November 30, 2022, GPT has demonstrated potential in various medical applications, such as answering United States Medical Licensing Examination (USMLE) questions [22], generating simplified radiology reports for patients [23], optimizing clinical decision support [24], and several other clinical applications [25]. However, concerns exist regarding GPT's comprehension of queries and its restricted capacity to deliver detailed answers [26]. To date, there has been no research evaluating the accuracy and comprehensiveness of GPT in answering specific questions related to AD. Therefore, we collected the perception of participants (evaluator neurologists/family members) regarding the accuracy, comprehensibility, clarity, and usefulness of GPT's responses to frequently asked questions related to the management and caregiving of patients with AD.

The aims of this study were:

- To evaluate the potential of GPT's capability to respond questions related to AD management
- To compare and assess the difference between GPT's responses to AD's care queries and those provided by neurologists

Methods

Participant Selection

The recruitment of neurologists was carried out through a purposeful sampling approach, with careful consideration given to ensuring the representativeness of the selected neurologists. Factors considered included their affiliation with different hospitals, years of clinical experience, medical specialties, and age. The selection criteria for neurologists were as follows: (1) ≥5 years of clinical experience in neurology, (2) a medical doctor degree, (3) grade III hospitals (the highest level of

XSL•FO
RenderX

hospitals in China), and (4) voluntary willingness to participate in this study.

In the selection of family members to participate in the study, our aim was to ensure diversity and representativeness by including individuals of varying ages, sexes, cultural backgrounds, and educational levels who have experience in caring for a person with AD daily. The criteria for selecting these participants were as follows: (1) 18-55 years old, (2) living with or providing daily care of a person with AD for a minimum of 1 year, (3) proficiency in the language skills necessary to comprehend and evaluate the study's content, (4) primary school education or higher, and (5) voluntary willingness to participate in this study.

Questionnaire Development

In our initial phase, we systematically compiled frequently encountered inquiries pertaining to AD from authoritative publications issued by esteemed professional associations and organizations [27-35]. A study showed that most of the content of the educational materials provided by these associations is unanimously recognized, with a mean overall score of 32.33 (SD 4.66) for the applicability of their educational content out of a total score of 44 [36]. This indicates that the content is well suited for health education purposes. To foster inclusivity and encompass a broad spectrum of patients and caregivers, we expanded the initial question pool. We meticulously excluded queries that were redundant, had ambiguous connotations, displayed subjectivity susceptible to individual interpretation, and underwent precise linguistic and grammatical refinement to enhance precision. Subsequently, after 3 rounds of expert discussion, we definitively curated a comprehensive set of 14 questions related to AD. These questions comprehensively address the multifaceted concerns of patients' families and caregivers, encompassing vital subjects, such as general knowledge regarding AD, associated risks, diagnostic processes, treatment modalities, caregiving strategies, prognostic outlook, rehabilitation measures, and nursing practices (Multimedia Appendix 1). All questions used in this study are listed in Table S1 in Multimedia Appendix 2.

GPT and Response Generation

GPT (GPT-3.5, or ChatGPT) leverages reinforcement learning from human feedback or reinforcement learning from human preference approaches to generate coherent and contextually appropriate answers. Building on the foundation of the GPT-3.5 large language model, this research preview prototype demonstrates impressive capabilities in understanding and generating human-like text [36]. It underwent pretraining on an extensive data set comprising text extracted from online sources, such as websites, books, and papers up until 2021.

Questions were input into the GPT May 3 version, using the GPT-3.5 and GPT-4 models. Each question was treated as an individual prompt using the "New Chat" function. Additionally, to explore the variations in GPT's responses, each question was translated into Chinese. Figure 1 shows an example of using GPT-3.5 to generate responses. We also invited 4 respondent neurologists to answer each question to compare the quality of GPT and respondent neurologists' responses. All responses are listed in Table S2 in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. Example of using GPT-3.5 to generate a response. GPT: Generative Pretrained Transformer.

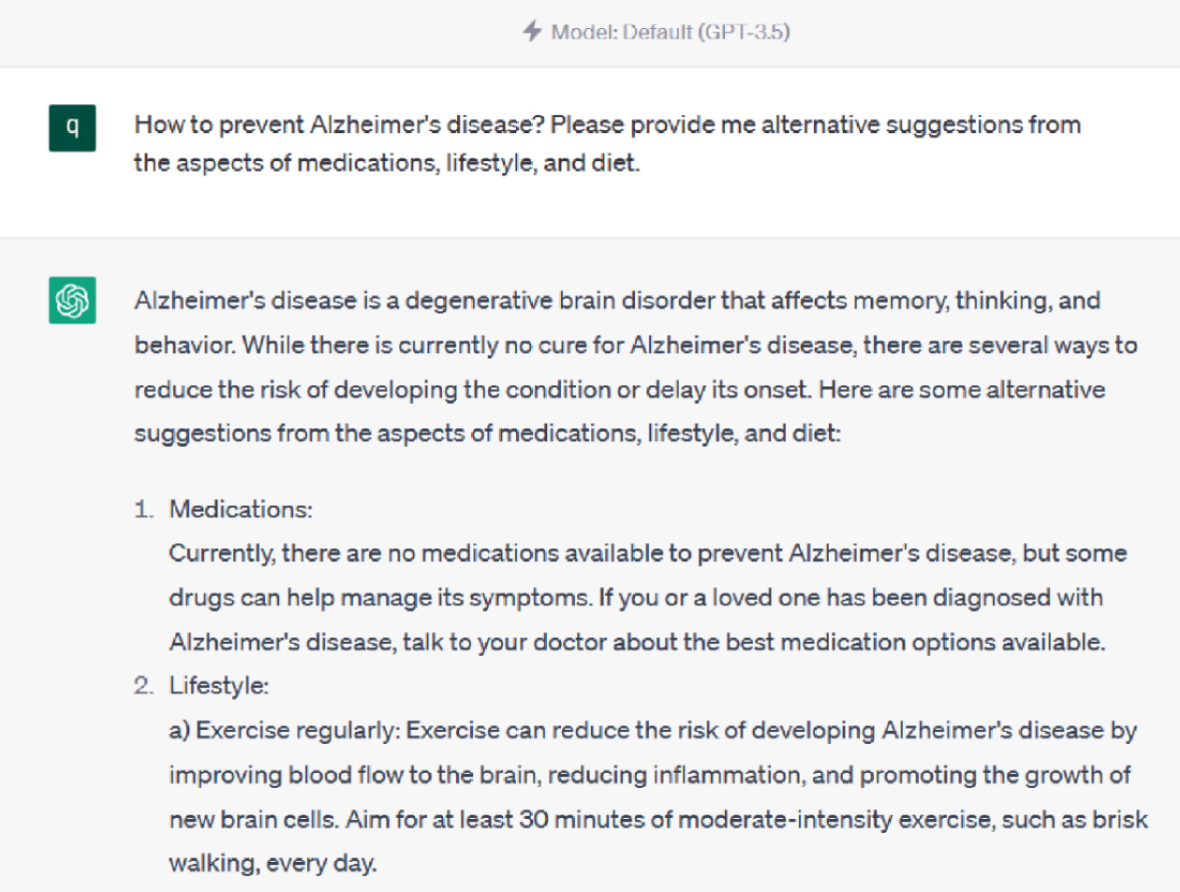

Evaluation Response

To minimize bias between the evaluators, comprising neurologists and patients' family members, in their perception of GPT and the responses from respondent neurologists, we eliminated any mentions of artificial intelligence (AI) in the GPT-generated responses. Subsequently, we used a random number generator to generate random identifiers for 8 answers. The coder remained unaware of whether these identifiers were generated by a neurologist or by GPT, ensuring a double-blind approach. The randomly coded answers were independently reviewed and scored by 5 evaluator neurologists and 5 family members of patients for each response. The evaluator neurologists were asked to provide their demographic information, including sex, age, professional title, and years of practice. Each response was assessed using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with criteria including accuracy, comprehensiveness, comprehensibility, and overall satisfaction [25,37-40]. Table S3 in Multimedia Appendix 2 shows the questionnaire items for the evaluator neurologists. To further assess the quality of the responses, optional open-ended questions were included, and the evaluator neurologists were asked to explain their scoring choices. The evaluator neurologists were also asked to select the most satisfactory and least satisfactory responses. Patients' family members were asked to provide their demographic information, including sex, age, occupation, marital status, duration of illness, stage of illness, annual household income, and education level. Each response provided was rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), focusing on criteria such as comprehensibility, usefulness, and overall satisfaction. Table S4 in Multimedia Appendix 2 details the questionnaire items used for the patients' family members. Furthermore, patients' family members were asked to select the most satisfactory and least satisfactory response from each set of responses. In this study, we implemented several measures to ensure that all participants, including patients' family members, could fully comprehend the English responses generated by GPT. All English questionnaires and responses were carefully translated into Chinese by qualified professionals and underwent a rigorous review process to ensure linguistic accuracy and cultural relevance. This translation process was aimed at minimizing any language barriers for the participants (Multimedia Appendix 3).

Statistical Analysis

Descriptive statistics were calculated for the Likert scale ratings of each response category, including means (SDs). We evaluated the distribution of continuous variables using the Shapiro-Wilk test with IBM SPSS Statistics version 26.0, which indicated nonnormal distribution (Multimedia Appendices 4 and 5). To compare the ratings of responses produced by GPT and those generated by neurologists, we conducted the Mann-Whitney *U* test. The statistical significance level was set at $P<.05$. Furthermore, we calculated intraclass correlation coefficients (ICCs) to assess the reliability among evaluators. The ICC estimates and 95% CIs were calculated with a 2-way mixed-effects model using both the mean and consistency definitions for the k rater type. The ICC values were interpreted

XSL•FO
RenderX

as follows: values<0.5 indicate low reliability, 0.5-0.74 indicate moderate reliability, 0.75-0.9 indicate good reliability, and >0.9 indicate excellent reliability [41]. All questionnaires were input into IBM SPSS Statistics version 26.0 for processing. We used Microsoft Excel version 2304 to rank the satisfaction of each response; stacked bar charts were generated by Microsoft Excel version 2304.

Ethical Considerations

All participants in this study completed the questionnaire voluntarily to ensure equal participation. The data collected do not contain any personally identifiable information, as all information was anonymized to comply with data protection regulations. According to Article 32 of the Regulations on Ethical Review of Biomedical Research Involving Human Subjects in China [42], this study was exempt from the requirement to obtain approval from an ethical review board.

Results

Study Participants

As of April 10, 2023, we enrolled 9 neurologists, 4 respondent neurologists, and 5 evaluator neurologists in the study. The respondent neurologists included 2 (50%) neurology specialists, 1 (25%) fellow in neurology, and 1 (35%) neurologist, while the evaluator neurologists included 1 (20%) neurology professor, 1 (20%) neurology specialist, 2 (40%) fellows in neurology, and 1 (20%) neurologist. The mean age of the respondent neurologists and evaluator neurologists was 38.50 (SD 10.91) and 43.80 (SD 16.83) years, respectively. The mean clinical experience of the respondent neurologists and evaluator neurologists was 12.75 (SD 8.81) and 19.60 (SD 16.83) years, respectively. The ICC value of the respondent neurologists and evaluator neurologists was 0.95 (95% CI 0.42-0.99) and 0.98 (95% CI 0.85-0.99), respectively, indicating good reliability. Table 1 shows the characteristics of the respondent neurologists and evaluator neurologists in the study.

As of April 2023, we invited 5 family members of patients with AD to participate in this study. Most participants were female (n=4, 80%), with a mean age of 31.20 (SD 14.87) years. In terms of occupation, 3 (60%) participants were students, 1 (20%) was employed, and the remaining 1 (20%) was self-employed. Their annual family income was concentrated (n=4, 80%) in the range of Chinese Yuan (CNY) 10,000-20,000 (~US $1406-$2813), and the majority of them (n=4, 80%) lived in urban areas. In addition, their educational background ranged from primary school to a master's degree. Regarding the stage of AD, 4 (80%) were in the middle stage and 1 (20%) in the late stage. The mean duration of AD was 7.0 (SD 4.9) years. The ICC value was 0.89 (95% CI 0.48-0.99), also showing good reliability. Table 2 shows the characteristics of the patients' family members who participated in the survey.

Table 1. Characteristics of neurologists participating in the study.

Characteristics	Respondent neurologists (n=4)	Evaluator neurologists (n=5)
Sex, n (%)		
Male	2 (50)	3 (60)
Female	2 (50)	2 (40)
Age (years), mean (SD)	38.50 (10.91)	43.80 (16.83)
Title of public health technician, n (%)		
Neurology professor	0	1 (20)
Neurology specialist	2 (50)	1 (20)
Fellow in neurology	1 (25)	2 (40)
Neurologist	1 (25)	1 (20)
Clinical experience (years), mean (SD)	12.75 (8.81)	19.60 (16.83)
ICC[a] (95% CI)	0.95 (0.42-0.99)	0.98 (0.85-0.99)

[a]ICC: intraclass correlation coefficient.

XSL•FO
RenderX

Table 2. Characteristics of patients' family members (n=5).

Characteristics	Value
Sex, n (%)	
Male	1 (20)
Female	4 (80)
Age group (years), n (%)	
18-29	3 (60)
30-50	1 (20)
51-52	1 (20)
Occupation, n (%)	
Student	3 (60)
Customer service	1 (20)
Other	1 (20)
Marital status, n (%)	
Married	3 (60)
Unmarried	2 (40)
Living place, n (%)	
Urban	4 (80)
Rural	1 (20)
Education, n (%)	
Primary school diploma	1 (20)
High school diploma	1 (20)
Bachelor's degree	1 (20)
Master's degree	2 (40)
Household income (CNY[a] per annum), n (%)	
10,000-20,000 (~US $1406-$2813)	4 (80)
20,000-50,000 (~US $2813-$7032)	1 (20)
Stage of AD[b], n (%)	
Middle stage	4 (80)
Late stage	1 (20)
Duration of AD (years), mean (SD)	7.0 (4.9)
ICC[c] value (95% CI)	0.89 (0.48-0.99)

[a]CNY: Chinese Yuan. An exchange rate of CNY 1= US $0.14 was applied.

[b]AD: Alzheimer's disease.

[c]ICC: intraclass correlation coefficient.

Comparison of GPT Responses with Neurologist Responses

A total of 112 questions were answered by GPT and the respondent neurologists. Specifically, GPT-3.5 and GPT-4 each provided responses to 14 questions, both in English and in Chinese. Additionally, the 4 respondent neurologists individually answered 14 questions in a manner consistent with GPT's responses. All responses are included in the questionnaire (Multimedia Appendix 2). The average length of GPT responses to each question was 668.9 (SD 218.9) characters, of which GPT-3.5 responses had an average length of 630.9 (SD 212.7) characters, GPT-4 responses had an average length of 708.2 (SD 222.2) characters, and the average length of the respondent neurologists' responses to each question was 577.6 (SD 717.3) characters. In the evaluation of the most satisfactory responses for each questionnaire, the top 4 (80%) of the top 5 responses rated by the respondent neurologists were GPT responses and the fifth was a respondent neurologist' response (Multimedia Appendix 1). The top 5 responses with the highest scores rated by the patients' family members were GPT responses (n=4, 80%), except for the third response to a respondent neurologist's

XSL•FO
RenderX

answer (Multimedia Appendix 2). Overall, GPT's responses received higher satisfaction scores from both evaluator neurologists and patients' families compared to those from respondent neurologists.

Results of Evaluator Neurologists' Review of GPT- and Neurologist-Generated Responses

In the evaluator neurologists' evaluations, on average, both GPT- and neurologist-generated responses were rated as "agree" in comprehensiveness and satisfaction. In terms of accuracy and comprehensibility, GPT-generated responses were rated as "agree," while neurologist-generated responses were rated as "neither agree nor disagree." Figure 2 shows the stacked bar graphs representing the scores of each item of the GPT- and neurologist-generated responses. The detailed information for each item is listed in Table S3 in Multimedia Appendix 2.

Compared to GPT-generated responses, neurologist-generated responses scored lower in accuracy (GPT: mean 4.3, SD 0.6; neurologists: mean 3.7, SD 0.8; P=.04), comprehensiveness (GPT: mean 4.4, SD 0.6; neurologists: mean 4.0, SD 0.7; P=.11), comprehensibility (GPT: mean 4.4, SD 0.6; neurologists: mean 3.9, SD 0.8; P<.001), and satisfaction (GPT: mean 4.4, SD 0.6; neurologists: mean 4.0, SD 0.6; P<.001). The mean overall score for neurologist-generated responses was 3.9 (SD 0.7), while that for GPT-generated responses was 4.4 (SD 0.6; P<.001). Table 3 shows the means (SDs) for each aspect, and Table S5 in Multimedia Appendix 2 shows the means (SDs) for each item.

Figure 2. Stacked bar charts of neurologists' scores of each item for GPT (GPT-3.5+GPT-4)–generated responses and neurologist-generated responses, GPT-3.5- and GPT-4-generated responses, Chinese responses, and English responses. GPT: Generative Pretrained Transformer; S1: accuracy; S2: comprehensiveness; S3, S4, S7: comprehensibility; S5, S6, S8: satisfaction.

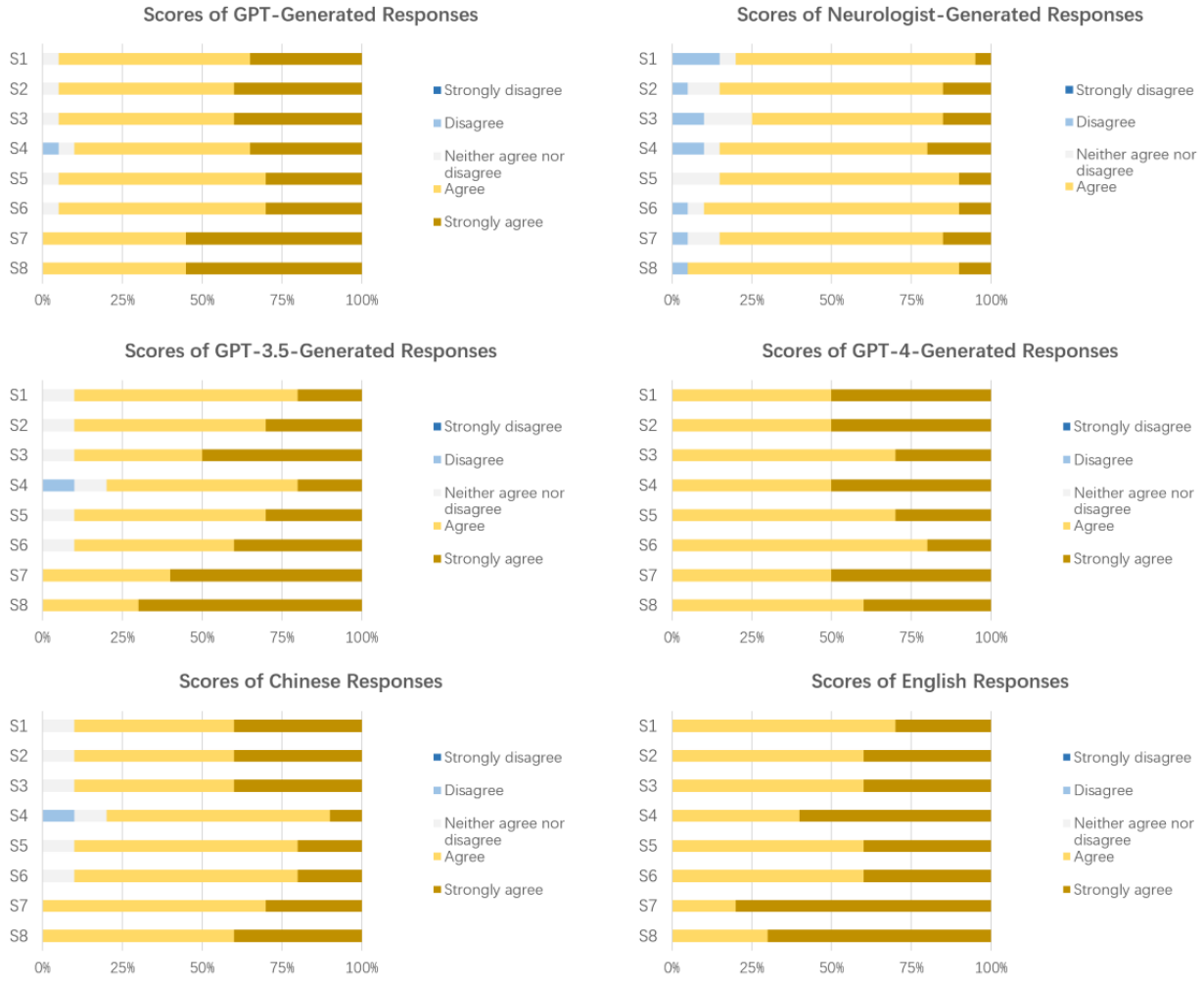

XSL•FO
RenderX

Table 3. Evaluator neurologists' ratings.

Group and response category	Accuracy	Comprehensiveness	Comprehensibility	Satisfaction	Overall
Group 1					
GPT[a,b]-generated responses, mean (SD)	4.3 (0.6)	4.4 (0.6)	4.4 (0.6)	4.4 (0.6)	4.4 (0.6)
Neurologist-generated responses, mean (SD)	3.7 (0.8)	4.0 (0.7)	3.9 (0.8)	4.0 (0.6)	3.9 (0.7)
P value	.04	.11	<.001	<.001	<.001
Group 2					
GPT-3.5-generated responses, mean (SD)	4.1 (0.6)	4.2 (0.6)	4.3 (0.7)	4.4 (0.9)	4.3 (0.7)
GPT-4-generated responses, mean (SD)	4.5 (0.5)	4.5 (0.5)	4.4 (0.5)	4.3 (0.5)	4.4 (0.5)
P value	.19	.35	.67	.36	.51
Group 3					
GPT[b] Chinese responses, mean (SD)	4.3 (0.7)	4.3 (0.7)	4.1 (0.7)	4.2 (0.8)	4.4 (0.6)
GPT[b] English responses, mean (SD)	4.3 (0.5)	4.4 (0.5)	4.6 (0.5)	4.5 (0.5)	4.5 (0.5)
P value	.91	.85	.005	.04	.002

[a]GPT: Generative Pretrained Transformer.

[b]GPT-3.5+GPT-4.

To further analyze the responses generated by GPT, we divided the responses into 2 groups according to the model that generated each response and the language of the response. We further compared the scores of each aspect (Table 3) and the scores of the items of the 2 groups of responses generated by GPT. Figure 2 shows stacked bar charts for the scores of each item. The results indicated that there were no statistically significant differences between GPT-3.5 and GPT-4 in terms of accuracy, comprehensiveness, comprehensibility, and satisfaction. However, although there were no significant differences in accuracy and comprehensiveness scores between Chinese and English responses, there were significant differences in comprehensibility (Chinese: mean 4.1, SD 0.7; English: mean 4.6, SD 0.5; *P*=.005) and satisfaction (Chinese: mean 4.2, SD 0.8; English: mean 4.5, SD 0.5; *P*=.04) scores. The mean total score for Chinese responses was 4.4 (SD 0.6), while that for English responses was 4.5 (SD 0.5; *P*=.002). This suggests that GPT performs differently in different languages, with responses in the English context being more comprehensible and satisfying.

Results of Patient Families' Review of GPT- and Neurologist-Generated Responses

In patient family ratings, on average, the comprehensibility of GPT-generated responses was rated as "agree," while practicality and satisfaction were rated as "neither agree nor disagree." Figure 3 shows stacked bar graphs representing the scores of each item of GPT- and neurologist-generated responses. The neurologist-generated responses scored similarly to the GPT-generated responses in terms of comprehensibility, usefulness, and satisfaction. The total scores for both were the same, 3.9. The means (SDs) for each aspect are shown in Table 4, the detailed information for each item is listed in Table S4 in Multimedia Appendix 2, and the means (SDs) for each item are shown in Table S8 in Multimedia Appendix 2.

As mentioned before, we divided the responses into 2 groups and compared the scores of each aspect (Table 4) and the scores of the items of the 2 groups of responses generated by GPT. We found that there were no significant differences in the comprehensibility, usefulness, and satisfaction scores between GPT-3.5 and GPT-4 and between Chinese and English responses.

XSL•FO
RenderX

Figure 3. Stacked bar charts of patient family scores for each item for GPT- and neurologist-generated responses, GPT-3.5- and GPT-4-generated responses, Chinese responses, and English responses. GPT: Generative Pretrained Transformer; I1-I3: comprehensibility; I4, I6: practicality; I5, I7, I8: satisfaction.

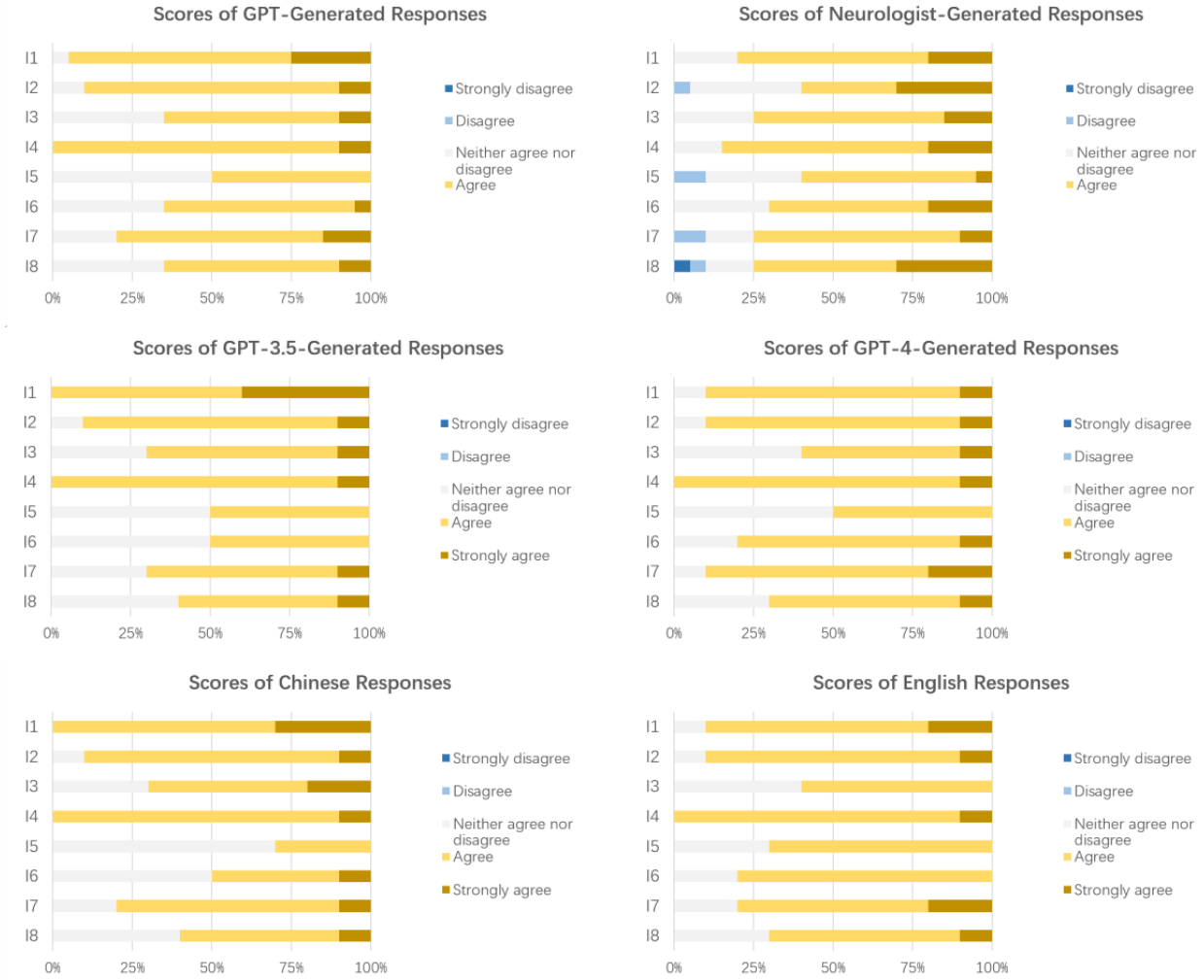

Table 4. Patients' family members' ratings.

Group and response category	Comprehensibility	Usefulness	Satisfaction	Overall
Group 1				
GPT[a,b]-generated responses, mean (SD)	4.0 (0.5)	3.9 (0.5)	3.7 (0.6)	3.9 (0.6)
Neurologist-generated responses, mean (SD)	3.9 (0.7)	4.0 (0.7)	3.7 (0.9)	3.9 (0.8)
P value	.62	.60	.57	.75
Group 2				
GPT-3.5-generated responses, mean (SD)	4.1 (0.6)	3.8 (0.5)	3.7 (0.6)	3.9 (0.6)
GPT-4-generated responses, mean (SD)	3.9 (0.5)	4.0 (0.5)	3.8 (0.6)	3.9 (0.6)
P value	.26	.34	.39	.65
Group 3				
GPT[b] Chinese responses, mean (SD)	4.1 (0.6)	3.9 (0.6)	3.6 (0.6)	3.9 (0.6)
GPT[b] English responses, mean (SD)	3.9 (0.5)	4.0 (0.4)	3.8 (0.6)	3.9 (0.5)
P value	.26	.60	.19	.62

[a]GPT: Generative Pretrained Transformer.

[b]GPT-3.5+ GPT-4.

XSL•FO
RenderX

Comments on GPT- and Neurologist-Generated Responses

In this study, the comments of the 5 evaluator neurologists on all responses were qualitatively analyzed. According to the evaluator neurologists' comments, the GPT responses had some shortcomings:

- Deviation in understanding the question. For example, 1 of the GPT responses to the question "What are the precautions for medication in AD, and how can medication-related side effects be minimized?" was "diet and lifestyle", and the advice on care and minimizing medication-related side effects was confusing.
- Omission of information. Some evaluator neurologists commented that the responses generated by GPT were incomplete. For example, in response to the question "How can I prevent AD? Please give me advice on alternative medications, lifestyle, and diet," the list of medications was incomplete.

There were more problems with the responses generated by the 4 respondent neurologists than with those generated by GPT. Problems found with respondent neurologists' responses included:

- Missing information. For example, for the question "How can I prevent AD? Please give me alternative advice on medication, lifestyle, and diet," 1 (25%) respondent neurologist did not provide a comprehensive list of risk factors.
- Template responses. For example, 2 (50%) respondent neurologists provided overly structured answers to the questions "What is the role of regular physical examinations and health assessments in preventing AD?" and "What are the causes or risk factors for AD?"
- Excessive use of medical jargon, for example, in responses to the questions "What is the role of regular health checks and health assessments in preventing AD?" and "What symptoms should a person with AD look for and go to the emergency room for?"
- Lack of relevance. For example, for the question "What are the precautions for medication in AD, and how can medication-related side effects be minimized?" 1 (25%) respondent neurologist did not mention ways to avoid medication-related side effects.
- Lack of practicality. For example, the response of 1 (25%) respondent neurologist to the question "What precautions should I take when treating people with AD with medication, and how can I minimize adverse effects related to medication?" was not practical.
- Potential for ambiguity. For example, when asked, "How do you determine the stage of progression of AD in a patient?", 1 (25%) respondent neurologist responded that there are 3 stages, with a total of 12 years, which could lead to confusion about the patient's life expectancy.

These results suggest that although GPT may suffer from misunderstanding, a lack of information, and a lack of relevance, its overall rating is still better than the individual responses of respondent neurologists.

Discussion

Principal Findings

The main finding of this research highlights the exceptional performance of GPT responses compared to those provided by evaluator neurologists, showing a significant advantage across 4 key dimensions: accuracy, comprehensiveness, comprehensibility, and overall satisfaction (Table 3). Although neurologists' responses still performed well in all aspects, they were outperformed by those of GPT. Additionally, a qualitative analysis revealed certain limitations in both GPT's and neurologists' responses, with more shortcomings identified in the neurologists' responses. Qualitative analysis can be influenced by factors such as the specific questions asked, the expertise of the neurologist, and the criteria used to score the responses. These findings align with prior research indicating the superior performance of GPT compared to human professionals in various medical specialties, particularly in cardiology, interventional radiology, and ophthalmology [43-45]. However, the application of GPT in a clinical setting must be approached with care, emphasizing ethics and transparency to safeguard patients' rights and the quality of health care [25,46].

The superior performance of GPT responses compared to those of neurologists can be attributed to GPT's remarkable potential for comprehensive data coverage [47]. GPT excels in effectively covering a wide spectrum of medical literature, allowing it to provide information that spans different medical domains [48]. In contrast, neurologists tend to have expertise focused on specific medical specialties or in dealing with individual patient histories, which tends to limit their scope of response [49]. The discrepancies observed between the responses of GPT and those of neurologists can be attributed to various factors affecting the latter. These include time constraints, the depth of expertise, and the variability inherent in individual knowledge and experience. Conversely, the limitations in GPT's performance may be due to deficiencies in training data, which may include inaccuracies and omissions [43,50]. Moreover, as an AI language model, GPT may not fully capture and communicate the contextual nuances and emotional subtleties that are often essential in medical consultations [51,52]. A remarkable limitation is its interaction style: When faced with ambiguous queries, GPT lacks the capability to request clarification through follow-up questions, such as "Did you mean home care services?" [50]. Furthermore, several studies have indicated that the responses generated by GPT may present readability challenges for a general audience, potentially impacting their comprehensibility and accessibility [50,53-56]. Therefore, it is important to maintain a balanced perspective when evaluating the effectiveness of AI models like GPT in clinical settings. This involves acknowledging their strengths in terms of data coverage and information retrieval, while also being aware of their limitations in areas where human practitioners outperform them.

In the comparative analysis, GPT-4 (score range 4.2-4.5) outperformed GPT-3.5 (score range 3.9-4.7). However, this difference did not have statistical significance ($P>.05$),

XSL•FO
RenderX

indicating that the improvements in GPT-4 may not be significant enough to substantially improve the quality of responses to AD-related queries. Nevertheless, GPT-4 has displayed superior performance in various domains, especially in the USMLE and StatPearls questions related to ophthalmology, epilepsy, and patient education. GPT-4 has statistically significantly outperformed GPT-3.5 ($P<.001$ to $<.005$) [45,57,58]. Additionally, it has proven to be effective in delivering accurate and comprehensible information to patients about medical procedures, associated risks, benefits, and recovery periods, thereby supporting informed decision-making [59]. OpenAI reports that GPT-4 is "82% less likely to respond to requests for disallowed content and 40% more likely to produce factual responses than GPT-3.5 on our internal evaluations" [60]. Moreover, GPT-4 benefits from training on more recent data, extending up to September 2021, including up-to-date information, unlike GPT-3.5, which was limited to data available before June 2021. This enhancement allows GPT-4 to generate more up-to-date responses [61]. There has been a notable increase in the consistency of GPT-4's responses over time, possibly due to its robust training and advanced sampling mechanisms that promote response stability [62]. With continued advances in AI, the GPT-4 is poised to become a valuable tool for patient management and health care delivery.

The comparison analysis showed that the responses generated by the GPT model in English (score range 4.3-4.7) significantly outperformed those in Chinese (score range 3.8-4.5; $P<.05$). This difference in performance suggests more advanced training and proficiency of the GPT model in English language contexts, indicating that responses to AD-related consultations may be more effective in English. Additionally, Hristidis et al's [50] research supported the limited capability of the GPT model in languages other than English, attributing this to a relatively lower level of training and development in non-English languages. Conversely, Takagi et al [57] provided evidence of GPT-4's effectiveness in the areas of clinical reasoning and medical knowledge within the Japanese language context. This finding provides an important perspective on the potential of GPT models in different linguistic environments, emphasizing the necessity for more extensive training and development in languages other than English to achieve optimal performance in a wider range of linguistic settings.

To better evaluate the usefulness of the GPT model's responses, we compared the responses of the GPT model to patients' family members with those provided by the neurologists. The focus was on comprehensibility, usefulness, and overall satisfaction with the patient's actual care needs. The results showed that the responses of the GPT model were similar to those of the neurologists, with a score of 3.9, and the difference was not statistically significant ($P>.05$). This suggests that the similarity in scores may be due to the limited medical expertise of the patients' families. Their ability to recognize the professional depth of the responses may be limited, leading them to prioritize the usefulness and understandability of the information over its medical accuracy and complexity [58,63,64]. This observation emphasizes a crucial aspect of using AI in patient care communication: the need to balance professional medical advice with lay understandability and applicability.

Limitations

There are several limitations of our study that need to be discussed. First, the sample size used for assessment was relatively small, consisting of only 5 neurologists and 5 family members of patients with AD as raters. Although we carefully considered the representativeness of both neurologists and family members when selecting them, the limited sample size could introduce potential biases into the results. Factors such as the clinical experience of the selected neurologists, the time constraints they faced, and the educational background and medical knowledge of the patients' family members may have influenced the outcomes. Consequently, the results of this study may not be fully reflective of the broader population of neurologists and patients' families. Second, it is important to note that GPT was trained on text data only up to 2021 and does not include information on drugs, clinical guidelines, or research developments beyond that year. Consequently, GPT's responses may contain outdated clinical recommendations, which could impact the accuracy and relevance of its suggestions. Furthermore, GPT has the capability to generate responses that sound confident but may not necessarily be accurate, including the possibility of providing incorrect answers or fabricated references [65-67]. Finally, it is essential to acknowledge that the assessment process relies on the perceptions of the participants, specifically the assessors who are neuroscientists and family members. Although we have developed normative assessment criteria, these evaluations inherently involve subjective judgments, introducing a degree of subjectivity to the assessment process. These limitations emphasize the need for caution in interpreting our findings and the importance of further research to address these limitations. Such research will contribute to a more comprehensive understanding of the potential and limitations of GPT in health care settings.

Conclusion

The results of this study suggest that GPT is a promising tool in the management of AD. The responses provided by GPT can be a valuable resource, providing useful information and support to patients and their families. However, it is important to emphasize that GPT should be used judiciously and should not be relied upon as the sole source of knowledge and expertise. Neurologists should continue to work with GPT, using their clinical expertise and judgment to ensure that patients receive accurate and tailored treatment recommendations.

Acknowledgments

We used OpenAI's generative artificial intelligence tool Generative Pretrained Transformer (GPT) to answer questions for the research survey. The original GPT transcripts are provided in Multimedia Appendix 6.

XSL•FO
RenderX

Data Availability

The input prompts and artificial intelligence outputs are available in Multimedia Appendix 2.

Authors' Contributions

JL and S Liu contributed to conceptualization; JZ, JL, and CW to data curation; JZ, XZ, JL, and S Liu to writing—original draft; and JZ, XZ, CW, JL, S Liu, YT, ST, HL, YW, LZ, JL, and S Li to formal analysis, investigation, supervision, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire development.
[DOCX File , 17 KB - jmir_v26i1e51095_app1.docx]

Multimedia Appendix 2
Input prompts and AI outputs. AI: artificial intelligence.
[DOCX File , 173 KB - jmir_v26i1e51095_app2.docx]

Multimedia Appendix 3
Translation development process.
[DOCX File , 16 KB - jmir_v26i1e51095_app3.docx]

Multimedia Appendix 4
Normal distribution test of evaluator neurologists' reviews.
[DOC File , 69532 KB - jmir_v26i1e51095_app4.doc]

Multimedia Appendix 5
Normal distribution test of patients' family members' reviews.
[DOC File , 53451 KB - jmir_v26i1e51095_app5.doc]

Multimedia Appendix 6
Original GPT transcripts. GPT: Generative Pretrained Transformer.
[TXT File , 80 KB - jmir_v26i1e51095_app6.txt]

References

1. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nat Rev Neurol 2018 Nov;14(11):653-666. [doi: 10.1038/s41582-018-0070-3] [Medline: 30291317]
2. Patterson C. World Alzheimer report 2018. The state of the art of dementia research: new frontiers. Alzheimer's Disease International. 2018. URL: https://www.alzint.org/resource/world-alzheimer-report-2018/ [accessed 2023-04-10]
3. Alzheimer's Association. 2021 Alzheimer's disease facts and figures. Alzheimers Dement 2021 Mar 23;17(3):327-406. [doi: 10.1002/alz.12328] [Medline: 33756057]
4. Wimo A, Seeher K, Cataldi R, Cyhlarova E, Dielemann JL, Frisell O, et al. The worldwide costs of dementia in 2019. Alzheimers Dement 2023 Jan 08;19(7):2865-2873 [FREE Full text] [doi: 10.1002/alz.12901] [Medline: 36617519]
5. Jia J, Wei C, Chen S, Li F, Tang Y, Qin W, et al. The cost of Alzheimer's disease in China and re‐estimation of costs worldwide. Alzheimers Dement 2018 Feb 09;14(4):483-491. [doi: 10.1016/j.jalz.2017.12.006] [Medline: 29433981]
6. Kales HC, Gitlin LN, Lyketsos CG. Assessment and management of behavioral and psychological symptoms of dementia. BMJ 2015 Mar 02;350(mar02 7):h369-h369 [FREE Full text] [doi: 10.1136/bmj.h369] [Medline: 25731881]
7. Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. Lancet 2017 Dec 16;390(10113):2673-2734. [doi: 10.1016/S0140-6736(17)31363-6] [Medline: 28735855]
8. Cummings JL, Tong G, Ballard C. Treatment combinations for Alzheimer's disease: current and future pharmacotherapy options. J Alzheimers Dis 2019 Feb 12;67(3):779-794. [doi: 10.3233/jad-180766] [Medline: 30689575]
9. Yu T, Lane H, Lin C. Novel therapeutic approaches for Alzheimer's disease: an updated review. Int J Mol Sci 2021 Jul 30;22(15):8208. [doi: 10.3390/ijms22158208] [Medline: 34360973]
10. Mangialasche F, Solomon A, Winblad B, Mecocci P, Kivipelto M. Alzheimer's disease: clinical trials and drug development. Lancet Neurol 2010 Jul;9(7):702-716. [doi: 10.1016/s1474-4422(10)70119-8] [Medline: 20610346]

XSL•FO
RenderX

11. Risk reduction of cognitive decline and dementia: WHO guidelines. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/312180/9789241550543-eng.pdf [accessed 2023-04-10]
12. Scheltens P, De Strooper B, Kivipelto M, Holstege H, Chételat G, Teunissen CE, et al. Alzheimer's disease. Lancet 2021 Apr;397(10284):1577-1590. [doi: 10.1016/s0140-6736(20)32205-4] [Medline: 33667416]
13. Forbes D, Forbes SC, Blake CM, Thiessen EJ, Forbes S. Exercise programs for people with dementia. Cochrane Database Syst Rev 2015 Apr 15;2015(4):CD006489. [doi: 10.1002/14651858.CD006489.pub4] [Medline: 25874613]
14. Yu L, Wilson RS, Schneider JA, Bennett DA, Boyle PA. Financial and health literacy predict incident Alzheimer's disease dementia and pathology. J Alzheimers Dis 2017 Feb 20;56(4):1485-1493 [FREE Full text] [doi: 10.3233/JAD-161132] [Medline: 28157101]
15. Rostamzadeh A, Stapels J, Genske A, Haidl T, Jünger S, Seves M, et al. Health literacy in individuals at risk for Alzheimer's dementia: a systematic review. J Prev Alzheimers Dis 2020;7(1):47-55. [doi: 10.14283/jpad.2019.34] [Medline: 32010926]
16. Palotti J, Zuccon G, Hanbury A. Consumer health search on the web: study of web page understandability and its integration in ranking algorithms. J Med Internet Res 2019 Jan 30;21(1):e10986 [FREE Full text] [doi: 10.2196/10986] [Medline: 30698536]
17. Werner P, Karnieli‐Miller O, Eidelman S. Current knowledge and future directions about the disclosure of dementia: a systematic review of the first decade of the 21st century. Alzheimers Dement 2012 Oct 24;9(2):e74-e88. [doi: 10.1016/j.jalz.2012.02.006] [Medline: 23098912]
18. Clare L, Wu Y, Teale JC, MacLeod C, Matthews F, Brayne C, et al. Potentially modifiable lifestyle factors, cognitive reserve, and cognitive function in later life: a cross-sectional study. PLoS Med 2017 Mar 21;14(3):e1002259 [FREE Full text] [doi: 10.1371/journal.pmed.1002259] [Medline: 28323829]
19. Gitlin LN, Winter L, Dennis MP, Hodgson N, Hauck WW. A biobehavioral home-based intervention and the well-being of patients with dementia and their caregivers: the COPE randomized trial. JAMA 2010 Sep 01;304(9):983-991 [FREE Full text] [doi: 10.1001/jama.2010.1253] [Medline: 20810376]
20. Chesser AK, Keene Woods N, Smothers K, Rogers N. Health literacy and older adults: a systematic review. Gerontol Geriatr Med 2016 Mar 15;2:2333721416630492. [doi: 10.1177/2333721416630492] [Medline: 28138488]
21. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2023 Mar 28;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
23. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2023 Oct 05;34(5):2817-2825. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
24. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
25. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
26. Bogost I. ChatGPT is dumber than you think: treat it like a toy, not a tool. The Atlantic. 2022. URL: https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/ [accessed 2023-04-10]
27. Nursing and rehabilitation. China Association for Alzheimer’s Disease. URL: http://www.caad.org.cn/col/col50/index.html [accessed 2023-04-10]
28. The global voice on dementia. Alzheimer's Disease International. URL: https://www.alzint.org/ [accessed 2023-04-10]
29. Dementia Australia is here for you. Dementia Australia. URL: https://www.dementia.org.au/ [accessed 2023-04-10]
30. The moment is now. Walk with us. Alzheimer's Association. URL: https://www.alz.org/ [accessed 2023-04-10]
31. Anna Richardson: love, loss and dementia. Alzheimer's Society. URL: https://www.alzheimers.org.uk/ [accessed 2023-04-10]
32. Latest news. Hong Kong Alzheimer's Association. URL: https://www.hkada.org.hk/ [accessed 2023-04-10]
33. Kia ora. Alzheimer New Zealand. URL: https://alzheimers.org.nz/ [accessed 2023-04-10]
34. Association for Dementia and Alzheimer’s of South Africa. URL: https://www.adasa.org.za/ [accessed 2023-04-10]
35. Dementia Singapore. Dementia Singapore. URL: https://dementia.org.sg/ [accessed 2023-04-10]
36. Xia D, Chen P, Huang M, Li X. Current status and analysis of online patient education materials released Alzheimer's Disease Associations at home and abroad. Chin J Health Educ 2020;36(02):182-185. [doi: 10.16168/j.cnki.issn.1002-9982.2020.02.020]
37. Swart ECS, Good CB, Henderson R, Manolis C, Huang Y, Essien UR, et al. Identifying outcome measures for atrial fibrillation value-based contracting using the Delphi method. Res Soc Admin Pharm 2022 Aug;18(8):3425-3431. [doi: 10.1016/j.sapharm.2021.10.008] [Medline: 34764046]
38. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]

XSL•FO
RenderX

39. Delanian Halsdorfer N, Blasquez J, Bensoussan L, Gentile S, Collado H, Viton J, et al. An assessment of patient satisfaction for a short-stay program in a physical and rehabilitation medicine day hospital. Ann Phys Rehabil Med 2011 Jun;54(4):236-247 [FREE Full text] [doi: 10.1016/j.rehab.2011.04.001] [Medline: 21570929]
40. Gavurova B, Dvorsky J, Popesko B. Patient satisfaction determinants of inpatient healthcare. IJERPH 2021 Oct 28;18(21):11337 [FREE Full text] [doi: 10.3390/ijerph182111337] [Medline: 34769856]
41. Thayaparan AJ, Mahdi E. The Patient Satisfaction Questionnaire Short Form (PSQ-18) as an adaptable, reliable, and validated tool for use in various settings. Med Educ Online 2013 Jul 23;18(1):21747 [FREE Full text] [doi: 10.3402/meo.v18i0.21747] [Medline: 23883565]
42. National Health Commission of the People's Republic of China, Ministry of Education, Ministry of Science and Technology, National Administration of Traditional Chinese Medicine. Regulations on ethical review of life sciences and medical research involving human subjects, document no. 4. China Government. 2023 Feb 18. URL: https://www.gov.cn/zhengce/zhengceku/2023-02/28/content_5743658.htm [accessed 2023-04-10]
43. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844. [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
44. Waisberg E, Ong J, Zaman N, Kamran SA, Sarker P, Tavakkoli A, et al. GPT-4 for triaging ophthalmic symptoms. Eye (Lond.) 2023 May 25;37(18):3874-3875. [doi: 10.1038/s41433-023-02595-9] [Medline: 37231187]
45. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. Cardiovasc Intervent Radiol 2023 Oct 23;47(2):245-250 [FREE Full text] [doi: 10.1007/s00270-023-03563-2] [Medline: 37872295]
46. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]
47. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
48. Huh S. Issues in the 3rd year of the COVID-19 pandemic, including computer-based testing, study design, ChatGPT, journal metrics, and appreciation to reviewers. J Educ Eval Health Prof 2023 Jan 31;20:5 [FREE Full text] [doi: 10.3352/jeehp.2023.20.5] [Medline: 36718045]
49. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. Cureus 2023 Feb;15(2):e35029 [FREE Full text] [doi: 10.7759/cureus.35029] [Medline: 36819954]
50. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 2023 Jul 25;25:e48966 [FREE Full text] [doi: 10.2196/48966] [Medline: 37490317]
51. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023 Apr 19;21(1):269 [FREE Full text] [doi: 10.1186/s12967-023-04123-5] [Medline: 37076876]
52. Ray P. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst 2023;3:121-154 [FREE Full text] [doi: 10.1016/j.iotcps.2023.04.003]
53. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad015 [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
54. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 2023 Oct 03;7(10):862-868. [doi: 10.1016/j.oret.2023.05.022] [Medline: 37277096]
55. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs 2024 Mar 12;23(2):122-126. [doi: 10.1093/eurjcn/zvad087] [Medline: 37603843]
56. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. Am J Roentgenol 2023 Nov 20;221(5):701-704. [doi: 10.2214/ajr.23.29622] [Medline: 37341179]
57. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
58. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023 Jun;15(6):e40822 [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
59. Kim H, Shin D, Kim J, Lee G, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. Seizure: Eur J Epilepsy 2024 Jan;114:1-8 [FREE Full text] [doi: 10.1016/j.seizure.2023.11.013] [Medline: 38007922]

XSL•FO
RenderX

60. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Open AI. URL: https://openai.com/product/gpt-4 [accessed 2023-12-01]
61. Martindale J, Bizzaco M. GPT-4 vs. GPT-3.5: how much difference is there? Digital Trends Media Group. 2024 Apr 4. URL: https://www.digitaltrends.com/computing/gpt-4-vs-gpt-35/ [accessed 2023-12-01]
62. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 01;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
63. Jackson DN, Trivedi N, Baur C. Re-prioritizing digital health and health literacy in Healthy People 2030 to affect health equity. Health Commun 2020 Apr 30;36(10):1155-1162. [doi: 10.1080/10410236.2020.1748828] [Medline: 32354233]
64. Bangerter LR, Griffin J, Harden K, Rutten LJ. Health information-seeking behaviors of family caregivers: analysis of the Health Information National Trends Survey. JMIR Aging 2019 Jan 14;2(1):e11237 [FREE Full text] [doi: 10.2196/11237] [Medline: 31518309]
65. ChatGPT: optimizing language models for dialogue. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-04-10]
66. Castelvecchi D. Are ChatGPT and AlphaCode going to replace programmers? Nature 2022 Dec 08:Online ahead of print. [doi: 10.1038/d41586-022-04383-z] [Medline: 36481949]
67. Stokel-Walker C. AI bot ChatGPT writes smart essays — should professors worry? Nature 2022 Dec 09:Online ahead of print. [doi: 10.1038/d41586-022-04397-7] [Medline: 36494443]

Abbreviations

AD: Alzheimer's disease
AI: artificial intelligence
CNY: Chinese Yuan
GPT: Generative Pretrained Transformer
ICC: intraclass correlation coefficients
NLP: natural language processing
USMLE: States Medical Licensing Examination

Edited by G Eysenbach, T Leung; submitted 23.07.23; peer-reviewed by L Zhu, K Giannakopoulos, A Kavadella; comments to author 16.09.23; revised version received 06.10.23; accepted 25.09.24; published 31.10.24.

Please cite as:
Zeng J, Zou X, Li S, Tang Y, Teng S, Li H, Wang C, Wu Y, Zhang L, Zhong Y, Liu J, Liu S
Assessing the Role of the Generative Pretrained Transformer (GPT) in Alzheimer's Disease Management: Comparative Study of Neurologist- and Artificial Intelligence–Generated Responses
J Med Internet Res 2024;26:e51095
URL: https://www.jmir.org/2024/1/e51095
doi:10.2196/51095
PMID:

©Jiaqi Zeng, Xiaoyi Zou, Shirong Li, Yao Tang, Sisi Teng, Huanhuan Li, Changyu Wang, Yuxuan Wu, Luyao Zhang, Yunheng Zhong, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Heuristics Identified in Health Data–Sharing Preferences of Patients With Cancer: Qualitative Focus Group Study

Anna Hermansen[1,2], BA, MSc; Samantha Pollard[1], PhD; Kimberlyn McGrail[2], PhD; Nick Bansback[2], PhD; Dean A Regier[1,2], PhD

[1]BC Cancer Research Institute, Vancouver, BC, Canada

[2]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Anna Hermansen, BA, MSc
BC Cancer Research Institute
675 W 10th Ave
Vancouver, BC, V5Z 1L3
Canada
Phone: 1 (604) 822 2772
Email: ajhermansen@gmail.com

Abstract

Background: Evaluating precision oncology outcomes requires access to real-world and clinical trial data. Access is based on consent, and consent is based on patients' informed preferences when deciding to share their data. Decision-making is often modeled using utility theory, but a complex decision context calls for a consideration of how heuristic, intuitive thought processes interact with rational utility maximization. Data-sharing decision-making has been studied using heuristic theory, but almost no heuristic research exists in the health data context. This study explores this evidence gap, applying a qualitative approach to probe for evidence of heuristic mechanisms behind the health data-sharing preferences of those who have experienced cancer. Exploring qualitative decision-making reveals the types of heuristics used and how they are related to the process of decision-making to better understand whether consent mechanisms should consider nonrational processes to better serve patient decision-making.

Objective: This study aimed to explore how patients with cancer use heuristics when deciding whether to share their data for research.

Methods: The researchers conducted a focus group study of Canadians who have experienced cancer. We recruited participants through an online advertisement, screening individuals based on their ability to increase demographic diversity in the sample. We reviewed the literature on data-sharing platforms to develop a semistructured topic guide on concerns about data sharing, incentives to share, and consent and control. Focus group facilitators led the open-ended discussions about data-sharing preferences that revealed underlying heuristics. Two qualitative analysts coded transcripts using a heuristic framework developed from a review of the literature. Transcripts were analyzed for heuristic instances which were grouped according to sociocultural categories. Using thematic analysis, the analysts generated reflexive themes through norming sessions and consultations.

Results: A total of 3 focus groups were held with 19 participants in total. The analysis identified 12 heuristics underlying intentions to share data. From the thematic analysis, we identified how the heuristics of social norms and community building were expressed through altruism; the recognition, reputation, and authority heuristics led to (dis)trust in certain institutions; the need for security prompted the illusion of control and transparency heuristics; and the availability and affect heuristics influenced attitudes around risk and benefit. These thematic relationships all had impacts on the participants' intentions to share their health data.

Conclusions: The findings provide a novel qualitative understanding of how health data–sharing decisions and preferences may be based on heuristic processing. As patients consider the extent of risks and benefits, heuristic processes influence their assessment of anticipated outcomes, which may not result in rational, truly informed consent. This study shows how considering heuristic processing when designing current consent mechanisms opens up the opportunity for more meaningful and realistic interactions with the complex decision-making context.

(J Med Internet Res 2024;26:e63155) doi:10.2196/63155

XSL•FO
RenderX

KEYWORDS

heuristics; health data sharing; cancer patients; decision-making; real-world data; altruism; trust; control; data sharing; focus group; precision medicine; clinical data; exploratory study; qualitative; Canada; thematic analysis; informed consent; patient education; information technology; healthcare; medical informatics

Introduction

IT in health care has generated a wide range of digital patient data collected during various health-related activities, including medical interactions, diagnoses, quality of life, treatments, drug adherence, and reimbursement [1-3]. Accessing real-world data allows for a detailed profile of a patient to better understand and target health issues and build evidence to understand the clinical and cost-effectiveness of personalized health technologies [4]. Access to clinical and routinely collected real-world data hinges on patient consent and legislation. As health institutions look toward data-sharing solutions that can support evidence generation for personalized care, they must consider how to best incorporate patient concerns and expectations to reach meaningful consent [5-9].

Understanding an individual's process of providing consent involves modeling attitudes, context, and risks and benefits involved in the decision [10,11]. Decision-making modeling is typically grounded in utility theory, where rational actors deliberate on the outcomes of possible alternatives and make a decision that maximizes their utility, based on well-defined and consistent preferences and using unlimited cognitive processing power [12-15]. Simon [14,16] (1957, 1979) provided an alternative framework, arguing that in a complex and uncertain decision context, individuals hit a cognitive limit to their rational processing capabilities and instead satisfice their utility. The individual's choice is based on the context-specific value placed on the potential outcomes rather than a rational assessment of their utility function [17-19].

Kahneman and Tversky [15,18] (1974) explored satisficing through heuristic theory and defined heuristics as the different rules of thumb used by individuals to substitute rational processing for quicker judgment calls. Kahneman [20] (2003) elaborated on the decision-making process as a dual-system model where system 1 rational (controlled, slow) and system 2 heuristic (intuitive, automatic) processes react during complex decision-making. The authors used this dual-system theory to develop Figure 1, which represents decision-making as an interaction among heuristics, attitudes, and the risk-benefit calculus, all of which are defined in Table 1 [21,22]. In this study, we take the position that the decision to share data is a complex process that involves the interaction of both heuristic and rational processes. In this map, the system 1 heuristic process interacts with the system 2 attitudes and risk-benefit calculus to form the decision to share data. The relationships between systems 1 and 2 are not reducible to a simple cause and effect; both systems nondirectionally influence each other to lead to a decision.

Figure 1. Conceptual model of heuristic decision-making.

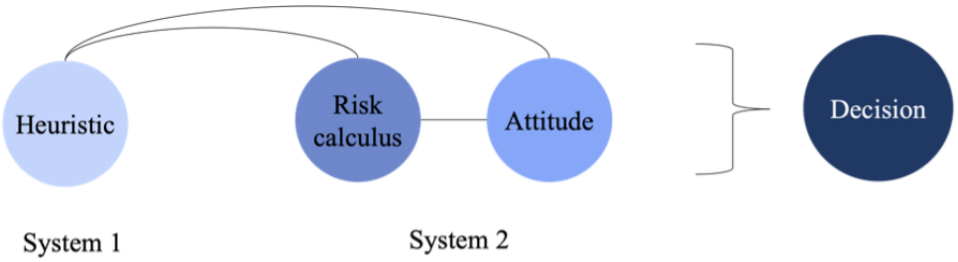

Table 1. Definitions of variables in the conceptual model of heuristic decision-making.

Variable	Definition
Heuristics	Judgment calls replacing rational processing with intuitive responses
Attitudes	The way individuals feel about a certain behavior; includes feelings, experiences, awareness, situational cues, and dispositions
Risk calculus	Rational deliberation between the risk and benefit involved in a decision
System 1	Intuitive, involuntary processing of concepts to generate impressions
System 2	Effortful, explicit processing of concepts to generate deliberate reasoning
Decision	Interaction between systems 1 and 2

Published research has found the presence of various heuristic processes in data-sharing decision-making [10,23-26]. However, very few have studied heuristics specifically used in health data sharing [7], meaning that the impact of heuristics on a patient's consent to share health data is not well understood. This study addresses this evidence gap by asking what heuristics, if any, are used when deciding whether to share personal health data for research purposes and how they impact the outcome of a stated decision. We are investigating the presence of heuristic processes to understand whether patients may be using a decision process similar to Figure 1. Capturing the use of heuristic processing in the health data–sharing context helps illuminate whether consent processes can better serve patient decision-making by considering the factors that influence

XSL•FO
RenderX

patients beyond their rational, utility-maximizing processes. These findings can support the development of consent mechanisms that represent the reality of this complex and context-dependent decision environment [5].

Methods

Overview

We conducted a qualitative focus group study with patients with cancer and survivors in Canada as part of the Canadian Network for Learning Healthcare Systems and Cost Effective 'Omics Innovation (CLEO) project, an initiative evaluating 6 precision oncology programs in Canada to inform the design of a learning health care system for cancer research and care [27]. Patients with cancer were the chosen population for this focus group study because of their relevance to the CLEO project and convenience sampling for the study authors who worked at a cancer research institute. Focus groups were chosen because they are effective for collecting data on socially construed beliefs, for exploring concepts that are not easily measurable, and for revealing underlying thought processes [24,28,29]. As compared with interviews, their social setting allows for discussion that is influenced by others' perspectives [30]. This discussion style can provide mental cues for participants, triggering memories, opinions, or feelings that represent heuristic processing.

Ethical Considerations

The protocol for data collection and analysis received ethics approval from the University of British Columbia BC Cancer Research Ethics Board (H20-00861). Participation in the focus groups was entirely voluntary. Consent for participation was collected through a consent form that described the goals of the study, the research team, the format of the focus group, expectations of participants, compensation, how confidentiality will be protected, and who to contact for more information. Consent forms were delivered by email to each participant before the focus group. The participants signed their forms as a PDF or through REDCap (Research Electronic Data Capture, Vanderbilt University), and only received the Zoom (Zoom Video Communications) link once they had completed the consent form. A copy of the consent form can be found in Multimedia Appendix 1. The participants received financial compensation of CAD $100 (US $70) after participating.

This study presented minimal risk to the participants. All demographic data collected from the participants through REDCap was stored in the REDCap platform and a locked folder on secure BC Cancer servers. We removed participant names from the transcripts and replaced them with anonymous identifiers. We stored the de-identified transcripts in a locked folder on BC Cancer servers.

Study Procedure

Individuals with a current or previous cancer diagnosis were recruited through a purposive sampling technique across provinces in Canada [30]. Participants were deliberately sampled to maximize diversity and not representativeness of the Canadian population. We recruited a diverse sample of cancer type, age, gender, and racial or cultural groups in order to hear from a wide range of perspectives. The participants were recruited from an advertisement posted on various provincial patient organizations (REACHBC, AbSPORU, and Cancer Care Ontario) and Kijiji, Craigslist, and Facebook patient groups. Oncologists at BC Cancer also supported purposive recruitment for racial and cultural diversity and cancer type. Those who expressed interest in participating were sent a screening survey that collected gender, age, racial or cultural group identity, cancer type, and general availability (a copy of the survey is mentioned in Multimedia Appendix 2). As screened candidates entered our selection process, we built a list based on those who provided variation in these demographic categories we collected, and we excluded those who did not provide demographic diversity compared with those previously selected. Those who were selected were invited to participate in a single focus group. They were sent the consent form and an introductory video on precision oncology and data-sharing platforms.

Focus groups were 90-minute sessions held and recorded over Zoom, facilitated by 2 female researchers at their workplace (SP and AH). At the time of the study, SP was a senior methodologist at BC Cancer with a PhD in epidemiology, ethics, and mixed-methods research. AH was a graduate student in public health with a bachelor's degree in political science with no prior experience conducting focus groups. No relationship was established with the participants prior to the focus groups beyond the communication required to set up their participation.

The facilitators used a semistructured question guide informed by a literature review [31] to guide the discussions on preferences for the data-sharing process (Multimedia Appendix 3). The literature review identified key topic areas and evidence gaps in the study of preferences for secured data-sharing platforms. These included motivators for and concerns about sharing data, incentives to allow access, types of institutions that can access the data, and consent and control features. The facilitators used these topics to develop and prioritize questions for the focus groups. Two researchers (AH and DAR) also developed questions to probe for heuristic processing, although the open-ended nature of the conversations was enough to naturally reveal heuristic processing. After the focus groups, the participants completed a demographic survey and were sent a CAD $100 (US $70) honorarium. Transcripts were not returned to participants.

Data Analysis

Data collection and analysis happened concurrently to achieve constant comparison [32] and followed a thematic analysis approach. Thematic analysis is a qualitative practice used for identifying and interpreting patterns across the data on participants' experiences, views, perspectives, and behaviors [33]. We chose this method because of the flexibility it offers in terms of theoretical approach and data analysis, allowing for a theoretical (deductive) application of a previously established theory to a novel research area [34]. Because this area is novel, we chose to extend our methods beyond cataloging or describing the results as topic summaries, and into the interpretation of these results, generating themes by reflexively identifying underlying motivators for the heuristic processing identified [35].

XSL•FO
RenderX

The research team worked to maintain a degree of transferability, dependability, and reflexivity in the data collection, analysis, and reporting [36]. Using the COREQ (Consolidated Criteria for Reporting Qualitative Research) standards, the explanation of the research setting and definition of heuristics increases the transferability of the study, and the documentation of our data collection and analysis processes encourages dependability of the findings [37]. This work also incorporated reflexive activities including team debriefing to review field notes after each focus group, weekly norming sessions, and working group consultations at major points along our study process [30].

We conducted a deductive analysis guided by a heuristic framework we developed from a review of literature on data-sharing heuristics. After focus group recordings were de-identified, transcribed, and checked for quality, 2 researchers (AH and SP) independently open-coded the transcripts, tagging the data for instances of the heuristic framework. Throughout the coding process, we refined and grouped heuristics to establish the final codebook. Statements could be coded with more than one heuristic. AH and SP conducted the coding process using QSR NVivo Software [38]. We completed data collection after reaching saturation, which we defined as the moment when analysis of new transcripts led to the same heuristic findings, with the same or similar impacts on the decision to share [39].

Once coding was complete, we looked for phrases in relevant passages of the transcript to identify the heuristics' corresponding attitudes and intentions to share, following the decision-making model (Figure 1). Heuristic codes were grouped into categories that were inductively created by identifying what causes the heuristic to be triggered and then defining larger sociocultural spheres of influence related to these causes. Resulting themes were derived from the data, representing each of these categories' heuristic influences on participants' intentions to share their data. The participants did not provide feedback on the heuristic findings.

Results

In total, 3 researchers (AH, SP, and DAR) conducted 3 Zoom-based focus groups from January to April 2022. Overall, 19 participants took part in the focus groups, with 5 to 8 participants per group. One candidate completed the consent form and received the Zoom link but did not attend the focus group. A total of 4 candidates confirmed their availability for the focus group but ended communication before completing the consent form. Table 2 describes participant demographic characteristics.

We coded 12 heuristics in the transcripts following the codebook (Table 3). Heuristics each have their own definitions but overlap in how and why they are triggered and what outcomes they produce. The differences between some of these heuristics are nuanced and the researchers relied on definitions from literature to maintain distinctions between them. We generated 4 themes linked to 9 heuristics. The other 3 heuristics identified were minor, nonsaturated findings (gatekeeping, novelty, and representativeness), and were not included in the thematic analysis.

XSL•FO
RenderX

Table 2. Self-reported participant demographics.

Demographics (N=19)	Values
Gender, n (%)	
Men	11 (58)
Women	8 (42)
Age (years), median (range)	63.5 (22-85)
Canadian province, n (%)	
British Columbia	8 (42)
Ontario	7 (37)
Alberta	3 (16)
Manitoba	1 (5)
Racial or cultural group, n (%)	
White	12 (79)
East Asian	4 (21)
Southeast Asian	2 (11)
Missing	1 (5)
Experience with cancer, n (%)	
Previous diagnosis	10 (53)
Current diagnosis, undergoing tx[a]	6 (32)
Current diagnosis, not undergoing tx	2 (10)
Missing	1 (5)
Cancer type[b]	
Breast	6 (33)
Prostate	3 (17)
Lung	3 (17)
Melanoma	2 (11)
Bone	1 (5)
Thyroid	1 (5)
Gastric	1 (5)
Uterine	1 (5)
Cervix	1 (5)
Endometrial	1 (5)
Duodenum	1 (5)
CLL[c]	1 (5)
Education level, n (%)	
Bachelor's or above	11 (58)
Diploma	4 (21)
University below bachelor's level	3 (16)
High school	1 (5)

[a]tx: treatment.

[b]Some participants reported more than one cancer type.

[c]CLL: chronic lymphocytic leukemia.

XSL•FO
RenderX

Table 3. Heuristic codebook.

Category and code[a]	Definition	Example from data
The personal		
Affect	Emotional response dictates assessment; objects and events are tagged with emotion and stored in a cognitive "affective pool" that is consulted when making a decision [40]	*That gives me a warm fuzzy feeling that it's very secure.* [P10]
Availability	The ease of recalling something from memory, whether a frequent occurrence or a prominent event (eg, personal experience or news story) [15]	*[W]e hear it in the news all the time with hacking.* [P09]
The social		
Social norms	How others' behavior impacts the decision, or how external peer influences impact the decision [41]	*I've seen this in other, a lot of other people that are not going ahead with [genetic] testing.* [P04]
Community building	Being part of a community influences individuals to trust and share information with each other, and to contribute to the common good [25]	*We have to be together trying to help each other to survive the best we can.* [P07]
The institutional		
Recognition	Ascribing value to and holding an inherent attitude toward an entity that is recognized or familiar [42,43]	*The only time I care [where my information is used] is when it's being used for profit, for like a pharmaceutical company.* [P11]
Representativeness	Passing judgment on something because of simple or salient cues that remind the individual of something else [44]	*This whole selling and stuff, that kind of scares me and that's sort of something like the US does.* [P04]
Reputation	Holding an inherent attitude toward an entity due to its prominence and reputation [29]	*We are dealing with medical professionals who are highly respectable and respectful* [P16]
Authority	Holding an inherent attitude toward something because of its authority over the subject [45,46]	*We opt in or we opt out, but then the legalese covers us for all the stuff that we don't know as patients.* [P10]
The informational		
Illusion of control	Believing that having personal controls over the data-sharing process will reduce the chance of risks involved, such as a data breach [47]	*I think informed consent would still be important just so you feel you have some control.* [P19]
Transparency	The degree to which data-sharing practices are known to the individual increases their sense of control and fairness [25,48]	*I don't want to say I'm 100 percent in favor of sharing data, because I don't necessarily know where it is going to be stored.* [P05]
The technological		
Novelty	Encounters with new technology, good or bad [24,28]	*in two years I guess you probably have to move [the data] to something new, a new technology, right?* [P04]
Gatekeeping	Being confident in and trusting a system that has many layers of access [24]	*a central committee or group that is a gatekeeper to what is going to be accessed or what needs to be inputted and not.* [P18]

[a]Codes are used to identify and group raw data into categories of analysis. The codes used in this study were all previously identified heuristics, which are defined as cognitive processes that substitute the rational calculation of probabilities for intuitive judgments of the decision context. We organized these heuristic codes in categories of what influences or is influenced by the heuristic.

The Social: Altruism as a Social Rule of Data Sharing

The first theme was the motivation to act altruistically when deciding to share data. The community building heuristic was found throughout the transcripts in participants' intention to benefit others. This heuristic captured the attitude of collectivity that participants felt toward others experiencing cancer, wanting to contribute to a "common good" [P07], "mankind" [P17], and the "big world" [P02]. Participant 7 explained, "everybody has their own rights. But we have to be together trying to help each other to survive the best we can."

Some participants acknowledged the benefit they had received from the contributions that had come before them and wanted to share to "make somebody's procedure better through my experience just as others have gone before me" [P04]. The community-building heuristic engendered strong intentions to share data for the benefit of others.

The social norms heuristic was also identified in instances where a decision not to share data was criticized, implying a social rule to be altruistic. For some participants, this rule was expressed as a disbelief at individuals who are unwilling to

XSL•FO
RenderX

share their data: “I can’t see how anyone would put up a fight or why they would not want it shared” [P02]. Participant 17 similarly implied that the social benefit generated by sharing data should motivate everyone: “Isn’t most of that data for the betterment of mankind?... I was just struggling to even get my head around why we wouldn’t do this.”

At times, this sentiment was found in direct contrast to discussions of risk. One participant downplayed the financial risk of being denied insurance: “human life is way more important than a mortgage, I’m sorry. We have to be part of a common good, not be selfish” [P07]. Altruism was also identified in the discussion on providing monetary incentives for sharing data, where some participants questioned why some would want compensation, and that it seemed like a “totally opposite kind of attitude” [P12]. Through their vocal belief in these behavior patterns, participants relied on the social norms heuristic in their decision processes.

The Institutional: Trust as a Measure of Legitimacy

The second theme centered on instances where participants reacted to the type of entity accessing their data. The recognition heuristic was primarily used in reaction to for-profit companies accessing health data, expressing distrust and privacy concerns. Without prompting, the participants expressed negative associations with pharmaceutical and insurance companies, such as Participant 13 stating that insurance access “immediately turns [them] off.” Participant 18 shared,

> *I don’t think I have a problem with accessing all my data. But the question is who gets to access it, and then would I be able to opt out on certain [access] like a pharmaceutical company or an insurance company.*

Some participants assumed for-profit to not have as high an ethical standard as other types of data requestors (such as academic researchers), tying this perception to the company’s morality, and whether their activity could be considered altruistic [P19]. This came out at times as a contrast between commercial and research activity, with one participant expressing that

> *…the only time I care [where my information is used] is when it’s being used for profit, for like a pharmaceutical company, that’s where I’m more hesitant … but if it’s just for research purposes, I’m happy.* [P11]

We coded trust in nonprofit and public researchers, as well as medical professionals, as the authority heuristic. During a conversation about managing data-sharing requests, Participant 18 recognized, “I don’t have the medical expertise in a lot of this,” and suggested, “I’d rather have a primary person or oncologist feeding me the information that I can trust a little bit more.” This demonstrated a reliance on their health care providers when considering whether to share data. Trust in researchers was also identified in instances of the reputation heuristic, where participants connected legitimacy and trust to the prestigiousness of public or nonprofit researchers and medical professionals. For example, one participant commented that

> *…we’re not dealing with gambling sites or anything disreputable. We are dealing with medical professionals who are highly respectable.* [P16]

Rather than examining the true risk of sharing with these types of entities, they relied on the recognition, reputation, and authority heuristics to evaluate their trust in them, which ultimately dictated their intentions to share.

The Informational: Gaining Power and Security Through Control

The third theme was the need for control over the data-sharing process. We coded the illusion of control heuristic if the participant believed that having personal controls over the data-sharing process reduced the probability of risks. Some participants expressed an interest in having controls in place to decide who gets access to what data and under what conditions to prevent a data leak. There appeared to be support for informed, detailed, and dynamic consent, which was valued by one participant to “give people some sense of control over what they want to share and what they don't want to share” [P19]. Other participants perceived a lack of control, with the understanding that “the second [your data] leaves you you’re no longer in control of that information” [P18]. In the face of this lack of control, these oversight features were important to our participants and gave them a sense of power over the data-sharing process, despite this kind of control not necessarily minimizing the risk.

This desire for controls was often connected to participants’ need to know more about the data request before allowing access, which we identified as the transparency heuristic. Some wanted to know about the storage of data:

> *I don’t want to say I’m 100 per cent in favor of sharing data, because I don’t necessarily know where it is going to be stored.* [P05]

Other participants wanted details on why and how the data was being used:

> *I need to understand the reasoning behind why on each and every aspect or piece of data. Then I would be all in.* [P13]

Being “all in” indicates a preference based on a heuristic response to the opacity of the data-sharing process, rather than on a deliberation of risks and benefits once receiving that detailed information. Whether in the form of information on or involvement in the data-sharing process, the perception of control was an important component of our participants’ decision-making processes to mitigate risk.

The Personal: Framing Risk and Benefit Through Personal Experiences

When explaining their decision-making processes, participants shared anecdotal experiences, feelings, and externally influenced beliefs tied to data sharing. Found across all focus groups was a reliance on the availability heuristic to evaluate the risk of cyberattacks. In these discussions, the participants recalled news stories that negatively impacted their intentions to share data:

XSL•FO
RenderX

> *Even just reading the paper today, you know that people are always trying to get at your data.* [P05]

This also included anecdotes that participants recalled from memory:

> *[U]nfortunately we just can't trust everybody today because look at what we see in the world, right? … if you remember that – a few years ago where somebody left important data on a bus, of all things, that was in Vancouver, you know.* [P04]

These stories led the participants to believe that data breach was an inevitability, with one expressing, "it seems there is always something somewhere that goes wrong" [P04]. The inevitability and frequency of a breach were a substantial consideration for our participants when discussing the risks of sharing data, despite none reporting to have experienced a data breach themselves.

The participants also used the availability heuristic when referencing personal and peers' experiences with and outside of cancer. Some of these experiences led to negative perceptions of sharing, in particular, those who had had or heard of "bad experiences with insurance people in the past" [P05] that made them fearful of health or employment insurance denial. Other experiences led to positive reflections on data sharing. This was evident in some participants' treatment issues that could have been mitigated if data had been shared, such as a life-threatening reaction or an adverse side effect. As Participant 13 noted, "that's what's important about data sharing, knowing who else out there has had that experience." Another participant explained their willingness to share through their own treatment experience, recalling:

> *I'm one of the ones that treatment worked. And there is some mystification around it… good data is just so important.* [P10]

These experiences were salient to decision-making processes, given the way participants used these anecdotes as rationales for deciding whether to share their data.

The affect heuristic was identified primarily through voiced emotion around for-profit involvement in the data-sharing process. These feelings included the loss aversion our participants felt toward insurance company access, such as Participant 5's previous experience with insurance "cast[ing] a negative light on the insurance industry" [P05], and Participant 13's feeling that insurance access "immediately turns [them] off" [P13]. Other participants felt uneasy about the sale of data, with one participant expressing, "this whole selling and stuff, that kind of scares me" [P04], and another who felt that "tying finance to personal data it just doesn't feel right." [P10]. Another felt "spiteful" toward for-profit companies [P11], wanting to limit their access to data, and another described for-profit activity as "sinister" [P19]. Many participants used their affective pool of feelings to explain their intentions. We found that the availability and affect heuristics triggered a more negative intention when the participants felt a distrust, privacy concern, or loss aversion, and positively when the participants found a personal benefit or collective benefit to sharing data.

Discussion

Implications of Findings for Consent

We identified qualitative patterns in the heuristics that patients with cancer use for health data sharing and how they impact decisions to share. Data-sharing themes found in our transcripts are not new to the data-sharing literature: research has found that individuals are motivated by altruism, are hesitant to share with for-profit entities, and want greater control over the process [31]. What makes our findings novel is the consideration of how heuristics play a role in the generation of these preferences. By studying underlying intentions, this study shows how the mechanism used to construct seemingly rational preferences for data sharing is influenced by heuristic processing. For example, previously studied preferences for greater transparency into the use of health data may, according to this study, be grounded in heuristic responses despite the appearance of deliberation. The heuristics analyzed in these focus groups therefore shed new light on how preferences are formed and prioritized when deciding to share data.

Heuristics, or "decision shortcuts," ultimately prove relevant in response to the complex and uncertain decision of sharing data. If participants are in fact using heuristics, then their decisions may not represent informed consent in the sense that they are not rationally deliberating on all outcomes to come to a utility-maximizing decision [49]. Instead of relying on what may be an impossible deliberation for patients, decision mechanisms can directly address these patterns of heuristic behavior by applying behavioral economic strategies, such as providing simplifying and salient information on the true risks and benefits at the time of consent. This approach can disrupt cognitive bias without asking individuals to deliberate all probabilities for a rational decision, as previous research has shown [10,23,40]. By respecting the boundedness of patients' processing capabilities, this kind of consent process can reach meaningful consent that is not necessarily a rational process but produces rational outcomes that maximize the patient's utility. As argued by Noah [49] (2016), it is important to recognize the obstacles to achieving perfect utility-maximizing consent and to instead encourage support for the "right" decision depending on the patient and the circumstances. The evidence found in this study can guide the testing of relevant behavioral economics strategies to manage the gap between the difficulty and yet necessity of giving deliberative informed consent.

Limitations

These results are subject to limitations. First, the majority of our sample self-reported as white and of higher education and income, and participants skewed toward older age. Our findings lack perspectives that may be systematically different from what we captured [50,51]. Second, we did not put a limit on the length of time since the patients' diagnosis or last treatment, which may have resulted in recall bias. Third, we chose to sample from patients with cancer and cancer survivors, who have unique experiences with their health and research in a way that may not be generalizable to individuals outside of this population [8,52,53]. Fourth, these findings are dependent on the subjectivity of the researchers' applications of the heuristic

XSL•FO
RenderX

framework. It is possible that the researchers confounded some of the heuristics, due to their similarities and the definitional nuances the researchers chose to apply. We attempted to mitigate this by having two researchers conduct independent analyses, applying a set of definitions informed by the literature, and comparing these two analyses during norming sessions. Finally, owing to study timelines and feasibility, we conducted all focus groups in English language.

Although not a limitation, we experienced some challenges with the use of Zoom for focus groups. This included the delayed and disruptive identification of one participant and another participant with poor sound quality. Following this first group, we developed a new identification process and were clearer about our expectations for sound quality from our participants, which resulted in smoother discussions in the next two groups.

Conclusion

This study investigated whether and which heuristic processes are used in intentions to share health data for research. By applying a novel heuristic perspective, this research illuminates the influence of heuristic processes in the formation of preferences around data sharing. Future research can expand this work beyond intentions to share, observing how these heuristics may or may not be evidenced in actual decision-making behavior. Our findings also have implications for future practice, where the design of consent mechanisms for sharing health data should consider the role of heuristic processes in this multifactorial decision, and how behavioral economics strategies can be used to encourage more meaningful patient engagement with their decision-making.

Acknowledgments

The CLEO project is funded by Genome British Columbia/Genome Canada (G05CHS), and AH received funding from the Canadian Institutes of Health Research's Canada Graduate Scholarship–Master's award (6556) to complete this research. The authors wish to thank the Canadian Network for Learning Healthcare Systems and Cost-Effective Omics Innovation's (CLEO) Working Group 2 members for their methodological support. AH used Covidence [54] and CitationChaser [55] during the literature reviews to streamline the review and extract reference lists, respectively. AH and SP used QSR NVivo Software to code the focus group transcripts. No authors used generative AI for any aspect of this study.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

AH, SP, DAR, KM, and NB contributed to conceptualization. SP and AH handled data curation and investigation. AH and SP performed formal analysis, investigation, validation, and handled resources. DAR and SP contributed to funding acquisition. SP, AH, and DAR handled methodology and project administration. DAR, SP, KM, and NB contributed to supervision and writing–review and editing. AH and DAR helped with visualization. AH contributed to writing–original draft.

Conflicts of Interest

AH, KM, and NB declare they have no relevant financial or nonfinancial interests to disclose. SP codirects IMPRINT Research Consulting and has received funding from Birota Economics Group. DAR has received honoraria from Roche and AstraZeneca.

Multimedia Appendix 1
Focus group consent form.
[PDF File (Adobe PDF File), 2175 KB - jmir_v26i1e63155_app1.pdf]

Multimedia Appendix 2
Demographic screening survey.
[PDF File (Adobe PDF File), 84 KB - jmir_v26i1e63155_app2.pdf]

Multimedia Appendix 3
Semistructured focus group topic guide.
[PDF File (Adobe PDF File), 98 KB - jmir_v26i1e63155_app3.pdf]

References

1. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med 2020;26(1):29-38. [doi: 10.1038/s41591-019-0727-5] [Medline: 31932803]
2. El-Galaly TC, Cheah CY, Villa D. Real world data as a key element in precision medicine for lymphoid malignancies: potentials and pitfalls. Br J Haematol 2019;186(3):409-419 [FREE Full text] [doi: 10.1111/bjh.15965] [Medline: 31140600]

XSL•FO
RenderX

3. Daniels H, Jones KH, Heys S, Ford DV. Exploring the use of genomic and routinely collected data: narrative literature review and interview study. J Med Internet Res 2021;23(9):e15739 [FREE Full text] [doi: 10.2196/15739] [Medline: 34559060]
4. Hulsen T, Jamuar S, Moody A, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. Front Med (Lausanne) 2019;6:34 [FREE Full text] [doi: 10.3389/fmed.2019.00034] [Medline: 30881956]
5. Deverka PA, Majumder MA, Villanueva AG, Anderson M, Bakker AC, Bardill J, et al. Creating a data resource: what will it take to build a medical information commons? Genome Med 2017;9(1):84 [FREE Full text] [doi: 10.1186/s13073-017-0476-3] [Medline: 28938910]
6. Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. Health Serv Res 2010;45(5 Pt 2):1442-1455 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01140.x] [Medline: 21054365]
7. Angst CM, Agarwal R. Adoption of electronic health records in the presence of privacy concerns: the elaboration likelihood model and individual persuasion. MIS Q 2009;33(2):339. [doi: 10.2307/20650295]
8. Anderson CL, Agarwal R. The digitization of healthcare: boundary risks, emotion, and consumer willingness to disclose personal health information. Inf Syst Res 2011;22(3):469-490. [Medline: 19642375]
9. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021;4(1):2 [FREE Full text] [doi: 10.1038/s41746-020-00362-8] [Medline: 33398052]
10. Kehr F, Kowatsch T, Wentzel D, Fleisch E. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. Info Systems J 2015;25(6):607-635. [doi: 10.1111/isj.12062]
11. Dinev T, McConnell AR, Smith HJ. Informing privacy research through information systems, psychology, and behavioral economics: thinking outside the APCO box. Inf Syst Res 2015;26(4):639-655. [doi: 10.1287/isre.2015.0600]
12. von NJ, Morganstern O. Theory of Games and Economic Behavior. Princeton, NJ: Princeton University Press; 1953.
13. Dinev T, Hart P. An extended privacy calculus model for e-Commerce transactions. Inf Syst Res 2006;17(1):61-80 [FREE Full text] [doi: 10.1287/isre.1060.0080]
14. Simon HA. Rational decision making in business organizations. Am Econ Rev 1979;69(4):493-513 [FREE Full text]
15. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
16. Simon HA. Models of Man. Hoboken, NJ: John Wiley & Sons; 1957.
17. Kahneman D, Tversky A. Prospect Theory: an analysis of decision under risk. Econometrica 1979;47(2):263-292 [FREE Full text] [doi: 10.2307/1914185]
18. Kahneman D. Thinking, Fast and Slow. Toronto, Canada: Anchor; 2011.
19. Simon H. Invariants of human behavior. Annu Rev Psychol 1990;41:1-20. [doi: 10.1146/annurev.ps.41.020190.000245] [Medline: 18331187]
20. Kahneman D. A perspective on judgment and choice: mapping bounded rationality. Am Psychol 2003;58(9):697-720. [doi: 10.1037/0003-066X.58.9.697] [Medline: 14584987]
21. Smith HJ, Dinev T, Xu H. Information privacy research: an interdisciplinary review. MIS Q 2011;35(4):989-1015. [doi: 10.2307/41409970]
22. Fishbein M, Ajzen I. Predicting and understanding consumer behavior: attitude-behavior correspondence. In: Ajzen I, Fishbein M, editors. Understanding Attitudes and Predicting Social Behavior. Hoboken, NJ: Prentice Hall; 1980:148-172.
23. Acquisti A, John LK, Loewenstein G. What is privacy worth? J Legal Stud 2013;42(2):249-274 [FREE Full text] [doi: 10.1086/671754]
24. Gambino A, Kim J, Sundar SS, Ge J, Rosson M. User disbelief in privacy paradox: heuristics that determine disclosure. New York, NY: Association for Computing Machinery; 2016 Presented at: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems; 2016 May 07; San Jose, CA. [doi: 10.1145/2851581.2892413]
25. Sundar SS, Kim J, Rosson MB, Molina MD. Online privacy heuristics that predict information disclosure. 2020 Presented at: Proceedings of the 2020 CHI Conference Extended Abstracts on Human Factors in Computing Systems; 2020 April 25-30; Honolulu, HI. [doi: 10.1145/3313831.3376854]
26. Adjerid I, Peer E, Acquisti A. Beyond the privacy paradox: objective versus relative risk in privacy decision making. MIS Q 2018;42(2):465-488. [doi: 10.25300/misq/2018/14316]
27. Canadian Network for Learning Healthcare Systems and Cost-Effective 'Omics Innovation (CLEO). Canada: BC Cancer URL: https://www.bccrc.ca/dept/ccr/projects/canadian-network-learning-healthcare-systems-and-cost-effective-omics-innovation-cleo [accessed 2024-10-31]
28. Marmion V, Bishop F, Millard DE, Stevenage SV. The cognitive heuristics behind disclosure decisions. In Social Informatics. SocInfo 2017;10539:591-607. [doi: 10.1007/978-3-319-67217-5_35]
29. Metzger MJ, Flanagin AJ, Medders R. Social and heuristic approaches to credibility evaluation online. J Commun 2010;60:413-439. [doi: 10.1111/j.1460-2466.2010.01488.x]
30. Patton MQ. Qualitative Research and Evaluation Methods. North America: Sage; 2015.
31. Hermansen A, Regier DA, Pollard S. Developing data sharing models for health research with real-world data: a scoping review of patient and public preferences. J Med Syst 2022;46(12):86. [doi: 10.1007/s10916-022-01875-3] [Medline: 36271208]

XSL•FO
RenderX

32. Glaser BG. The constant comparative method of qualitative analysis. Soc Probl 1965;12(4):436-445 [FREE Full text] [doi: 10.1525/sp.1965.12.4.03a00070]
33. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2017;12(3):297-298 [FREE Full text] [doi: 10.1080/17439760.2016.1262613]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
35. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. International Journal of Transgender Health 2022;24(1):1-6. [doi: 10.1080/26895269.2022.2129597]
36. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: Sage; 1985.
37. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
38. NVivo Qualitative Data Analysis Software. Australia: QSR International Pty Ltd; 2018.
39. Saunders B, Sim J, Kingston T. Saturation in qualitative researchxploring its conceptualization and operationalization. Qual Quant 2018;52:1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8]
40. Finucane ML, Alhakami A, Slovic P, Johnson SM. The affect heuristic in judgments of risks and benefits. J Behav Decis Making 2000;13:1-17 [FREE Full text] [doi: 10.1002/(sici)1099-0771(200001/03)13:1<1::aid-bdm333>3.0.co;2-s]
41. Acquisti A, John LK, Loewenstein G. The impact of relative standards on the propensity to disclose. J Mark Res 2012;49(2):160-174. [doi: 10.1509/jmr.09.0215]
42. Gigerenzer G, Todd PM, The ABC Research Group. Fast and frugal heuristics: the adaptive toolbox. In: Gigerenzer G, Todd PM, editors. Simple Heuristics That Make Us Smart. Oxford, United Kingdom: Oxford University Press; 1999:3-34.
43. Joeckel S, Dogruel L, Bowman ND. The reliance on recognition and majority vote heuristics over privacy concerns when selecting smartphone apps among German and US consumers. Inf Commun Soc 2016;20(4):621-636. [doi: 10.1080/1369118x.2016.1202299]
44. Kahneman D, Tversky A. Subjective probability: a judgment of representativeness. Cogn Psychol 1972;3(3):430-454 [FREE Full text] [doi: 10.1016/0010-0285(72)90016-3]
45. Dogruel L, Joeckel S, Vitak J. The valuation of privacy premium features for smartphone apps: the influence of defaults and expert recommendations. Comput Hum Behav 2017;77:230-239. [doi: 10.1016/j.chb.2017.08.035]
46. Sundar SS. The MAIN Model: a heuristic approach to understanding technology effects on credibility. In: Metzger MJ, Flanagin AJ, editors. Digital Media, Youth, and Credibility. Cambridge, MA: MIT Press; 2007:73-100.
47. Brandimarte L, Acquisti A, Loewenstein G, Babcock L. Privacy concerns and information disclosure: an illusion of control hypothesis. Dissertation. Carnegie Melon University. 2009. URL: https://www.researchgate.net/publication/43014896_Privacy_Concerns_and_Information_Disclosure_An_Illusion_of_Control_Hypothesis [accessed 2022-03-22]
48. Wang L, Hu HH, Yan J, Mei MQ. Privacy calculus or heuristic cues? The dual process of privacy decision making on Chinese social media. J Enterp Inf 2019;33(2):353-380. [doi: 10.1108/jeim-05-2019-0121]
49. Noah BA. The (ir)rationality of (un)informed consent. Quinnipiac Law Review 2016;34:691-704 [FREE Full text]
50. Hudson M, Garrison N, Sterling R, Caron NR, Fox K, Yracheta J, et al. Rights, interests and expectations: indigenous perspectives on unrestricted access to genomic data. Nat Rev Genet 2020;21(6):377-384. [doi: 10.1038/s41576-020-0228-x] [Medline: 32251390]
51. Priniski JH, Holyoak KJ. A darkening spring: how preexisting distrust shaped COVID-19 skepticism. PLoS One 2022;17(1):e0263191 [FREE Full text] [doi: 10.1371/journal.pone.0263191] [Medline: 35081170]
52. Hay AE, Leung YW, Pater JL, Brown MC, Bell E, Howell D, et al. Linkage of clinical trial and administrative data: a survey of cancer patient preferences. Curr Oncol 2017;24(3):161-167 [FREE Full text] [doi: 10.3747/co.24.3400] [Medline: 28680275]
53. Köngeter A, Schickhardt C, Jungkunz M, Bergbold S, Mehlis K, Winkler EC. Patients' willingness to provide their clinical data for research purposes and acceptance of different consent models: findings from a representative survey of patients with cancer. J Med Internet Res 2022;24(8):e37665 [FREE Full text] [doi: 10.2196/37665] [Medline: 36006690]
54. Covidence Systematic Review Software, Veritas Health Innovation, Melbourne, Australia. URL: https://www.covidence.org/ [accessed 2024-10-31]
55. Haddaway N, Grainger M, Gray C. citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching. Zenodo 2021:4543513. [doi: 10.5281/zenodo.4543513]

Abbreviations

CLEO: the Canadian Network for Learning Healthcare Systems and Cost Effective ‘Omics Innovation
COREQ: Consolidated Criteria for Reporting Qualitative Research
REDCap: Research Electronic Data Capture

XSL•FO
RenderX

Edited by A Coristine; submitted 12.06.24; peer-reviewed by HJ Toh, L Wreyford; comments to author 02.10.24; revised version received 22.10.24; accepted 23.10.24; published 17.12.24.

Please cite as:
Hermansen A, Pollard S, McGrail K, Bansback N, Regier DA
Heuristics Identified in Health Data–Sharing Preferences of Patients With Cancer: Qualitative Focus Group Study
J Med Internet Res 2024;26:e63155
URL: https://www.jmir.org/2024/1/e63155
doi:10.2196/63155
PMID:

©Anna Hermansen, Samantha Pollard, Kimberlyn McGrail, Nick Bansback, Dean A Regier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Environmental Impacts of Electronic Medical Records Versus Paper Records at a Large Eye Hospital in India: Life Cycle Assessment Study

Cordelia Kwon[1], BSc; Lernik Essayei[2], MSc; Michael Spencer[3], BSc; Tom Etheridge[4], PhD; Rengaraj Venkatesh[5], MD; Natrajan Vengadesan[5], MBA, MPhil; Cassandra L Thiel[6,7], PhD

[1]Department of Population Health, NYU Langone Health, New York, NY, United States
[2]NYU Wagner School of Public Service, New York, NY, United States
[3]Rausser College of Natural Resources, University of California, Berkeley, Berkeley, CA, United States
[4]EarthShift Global, Corvallis, OR, United States
[5]Aravind Eye Care System, Pondicherry, India
[6]Center for Healthcare Innovation and Delivery Science, Department of Population Health, NYU Langone Health, New York, NY, United States
[7]Department of Ophthalmology, NYU Langone Health, New York, NY, United States

Corresponding Author:
Cassandra L Thiel, PhD
Center for Healthcare Innovation and Delivery Science
Department of Population Health
NYU Langone Health
227 E. 30th St.
Room #636
New York, NY, 10016
United States
Phone: 1 6083871985
Email: cassandra.thiel@nyulangone.org

Abstract

Background: Health care providers worldwide are rapidly adopting electronic medical record (EMR) systems, replacing paper record-keeping systems. Despite numerous benefits to EMRs, the environmental emissions associated with medical record-keeping are unknown. Given the need for urgent climate action, understanding the carbon footprint of EMRs will assist in decarbonizing their adoption and use.

Objective: We aimed to estimate and compare the environmental emissions associated with paper medical record-keeping and its replacement EMR system at a high-volume eye care facility in southern India.

Methods: We conducted the life cycle assessment methodology per the ISO (International Organization for Standardization) 14040 standard, with primary data supplied by the eye care facility. Data on the paper record-keeping system include the production, use, and disposal of paper and writing utensils in 2016. The EMR system was adopted at this location in 2018. Data on the EMR system include the allocated production and disposal of capital equipment (such as computers and routers); the production, use, and disposal of consumable goods like paper and writing utensils; and the electricity required to run the EMR system. We excluded built infrastructure and cooling loads (eg. buildings and ventilation) from both systems. We used sensitivity analyses to model the effects of practice variation and data uncertainty and Monte Carlo assessments to statistically compare the 2 systems, with and without renewable electricity sources.

Results: This location's EMR system was found to emit substantially more greenhouse gases (GHGs) than their paper medical record system (195,000 kg carbon dioxide equivalents [CO_2e] per year or 0.361 kg CO_2e per patient visit compared with 20,800 kg CO_2e per year or 0.037 kg CO_2e per patient). However, sensitivity analyses show that the effect of electricity sources is a major factor in determining which record-keeping system emits fewer GHGs. If the study hospital sourced all electricity from renewable sources such as solar or wind power rather than the Indian electric grid, their EMR emissions would drop to 24,900 kg CO_2e (0.046 kg CO_2e per patient), a level comparable to the paper record-keeping system. Energy-efficient EMR equipment

XSL•FO
RenderX

(such as computers and monitors) is the next largest factor impacting emissions, followed by equipment life spans. Multimedia Appendix 1 includes other emissions impact categories.

Conclusions: The climate-changing emissions associated with an EMR system are heavily dependent on the sources of electricity. With a decarbonized electricity source, the EMR system's GHG emissions are on par with paper medical record-keeping, and decarbonized grids would likely have a much broader benefit to society. Though we found that the EMR system produced more emissions than a paper record-keeping system, this study does not account for potential expanded environmental gains from EMRs, including expanding access to care while reducing patient travel and operational efficiencies that can reduce unnecessary or redundant care.

(J Med Internet Res 2024;26:e42140) doi:10.2196/42140

KEYWORDS

carbon emissions; electronic health records; electronic medical records; environmental impact; greenhouse gases; life cycle assessment; low middle income country; medical records; paper medical records; sustainability

Introduction

Expansion of Electronic Medical Records

Health care systems are in the middle of a rapidly shifting framework, one that comes not from new medicines or new gene therapies but from the way medical information is tracked, stored, and shared through electronic health records (or electronic medical records [EMRs]). The use of EMR has been steadily increasing across the world since the early 2000s [1,2].

The COVID-19 pandemic and the subsequent rise of telemedicine propelled the expansion of EMRs. Telemedicine visits have expanded across various medical settings and are becoming more common in low- and middle-income countries (LMICs) [3]. In order to better integrate video visits and allow important information to be shared between doctor and patient remotely, patients and physicians began increasingly relying on EMRs to integrate video visits and allow important information to be virtually shared between doctor and patient, as well as among other providers in the patient's circle of care, ensuring better informed care and better outcomes [4]. EMRs also facilitated the implementation of screening processes for COVID-19, reporting and analytics of COVID-19 cases, and other outbreak management supports [5].

Beyond facilitating telemedicine, there is a sizable body of literature on the impacts, mostly beneficial, of EMR transitions on physician use, patient experience, and hospital efficiency [6-10]. EMRs are key data management tools as well as important validation tools for reducing clinician error [11]. When used in conjunction with telemedicine, they can facilitate medical care in rural and remote areas, which is particularly important in LMICs [3]. On the other hand, EMR adoption and implementation require immense institutional support, such as policy making and cultural change.

However, these considerations only represent some of the implications of implementing EMR systems. The carbon footprint of health care is 4.4% of net emissions globally [12], and therefore considering the environmental impact of implementing EMR systems is key to understanding the implications of their use. It is also worth noting that >90% of US hospitals have adopted some kind of certified EMR system to conduct 7 out of 9 patient care-related processes [13]. Although the rate of adoption of EMR systems in other countries, particularly LMICs, is slower due to the high costs associated with their purchase and maintenance, technological advancement and easier implementation have led to a steady upward trend, and it is expected that other countries' systems will soon mirror the United States. Understanding these implications is even more important when considering implementing EMR systems in the rapidly growing health care systems of LMICs. Measures to reduce carbon emissions in LMICs' health systems would not only combat climate change, which disproportionately threatens the health of the most vulnerable populations, but would also garner long-term savings [14].

To date, there is relatively little quantitative data about the environmental implications of transitioning to an EMR system, and even less in the context of LMICs. Transitioning away from paper-based records has the potential to reduce emissions from paper production and waste. However, electronic records also require infrastructure and electricity to implement and maintain, which require carbon emissions. This study will therefore try to fill this gap in knowledge by looking at how the transition to EMRs impacts a private institution in an LMIC, as well as addressing wider relevance toward sustainable and green health care goals.

Case Location

Aravind Eye Care Systems (referred to as "Aravind" henceforth) is the largest conglomerate eye care provider in India. It annually handles over 4 million outpatient visits and performs over 400,000 surgeries or laser procedures, with over 50% of them being free or steeply subsidized. Aravind's operating theaters are highly productive through efficient assembly line service operations and effective resource consumption. Cataract surgeries at Aravind are noted for their excellent outcomes, efficiency, cost-effectiveness, and environmental sustainability [15-18]. One sustainability initiative was the system's transition from paper-based health records, which was enacted at Aravind's Pondicherry-based location in 2018. Aravind-Pondicherry tertiary center is a 37,160 m^2 facility located on an 81,000 m^2 plot. It has 650 beds and caters to the needs of over 21.2 million people in the nearby districts of Tamil Nadu and Pondicherry. Aravind-Pondicherry served 568,982 patients in 2016, while still on a paper medical record system, and 538,325 in 2019, after switching to an EMR system. This

XSL•FO
RenderX

study compares the life-cycle greenhouse gas emissions (GHGs) of their paper record-keeping system in 2016 with their EMR system in 2019, 1 year after it was enacted.

Methods

Overview of Life Cycle Assessment

This study uses life cycle assessment (LCA) to estimate the GHG emissions of Aravind's paper record-keeping system (2016) compared with its EMR system (2019). LCA is a tool used to estimate the environmental emissions of a product or process throughout its life cycle, from raw material extraction, production, use, end of life, and all the transportation steps in between. It has been used to estimate emissions for a variety of products and services, including in the health care space [19,20]. Though it has been increasingly used to monitor the impact of telemedicine, this is the first time, to our knowledge, that it has been used to estimate GHGs from EMR systems.

According to ISO (International Organization for Standardization) 14040 standard [21], LCA is conducted in four steps: (1) goal and scope definition, (2) life cycle inventory (LCI), (3) life cycle impact assessment (LCIA), and (4) interpretation. The goal of this study is to compare the environmental impacts of Aravind's paper record-keeping system with their EMR system (Figures 1A and 1B), annually and on a per-patient basis. The functional unit is the creation and maintenance of 1 patient's health records. We include all production and disposal impacts from disposable and reusable supplies and equipment in both systems. We did not include heating, ventilation, and air conditioning required to maintain either system, given the difficulty in making such an allocation. Space is a major issue for Aravind's paper record-keeping system, for which this study may not accurately estimate the emissions. Aravind's policy requires them to store paper records on site for 3 years after the patient's last visit for outpatient records and 5 years for surgical records. As of 2020, Aravind-Pondicherry was maintaining 1,600,000 patient paper files in 2 hospital spaces, for a total of 237 m^2 of air-conditioned space dedicated to paper record storage. Likewise, EMRs require cooling servers for data storage, though this would take up considerably less space per patient than paper records.

XSL•FO
RenderX

Figure 1. System boundaries and included or excluded elements. (A) Aravind Eye Care System's paper record-keeping system (before 2016) and (B) Aravind's electronic medical record (EMR) system (implemented in 2018 and tested in 2019).

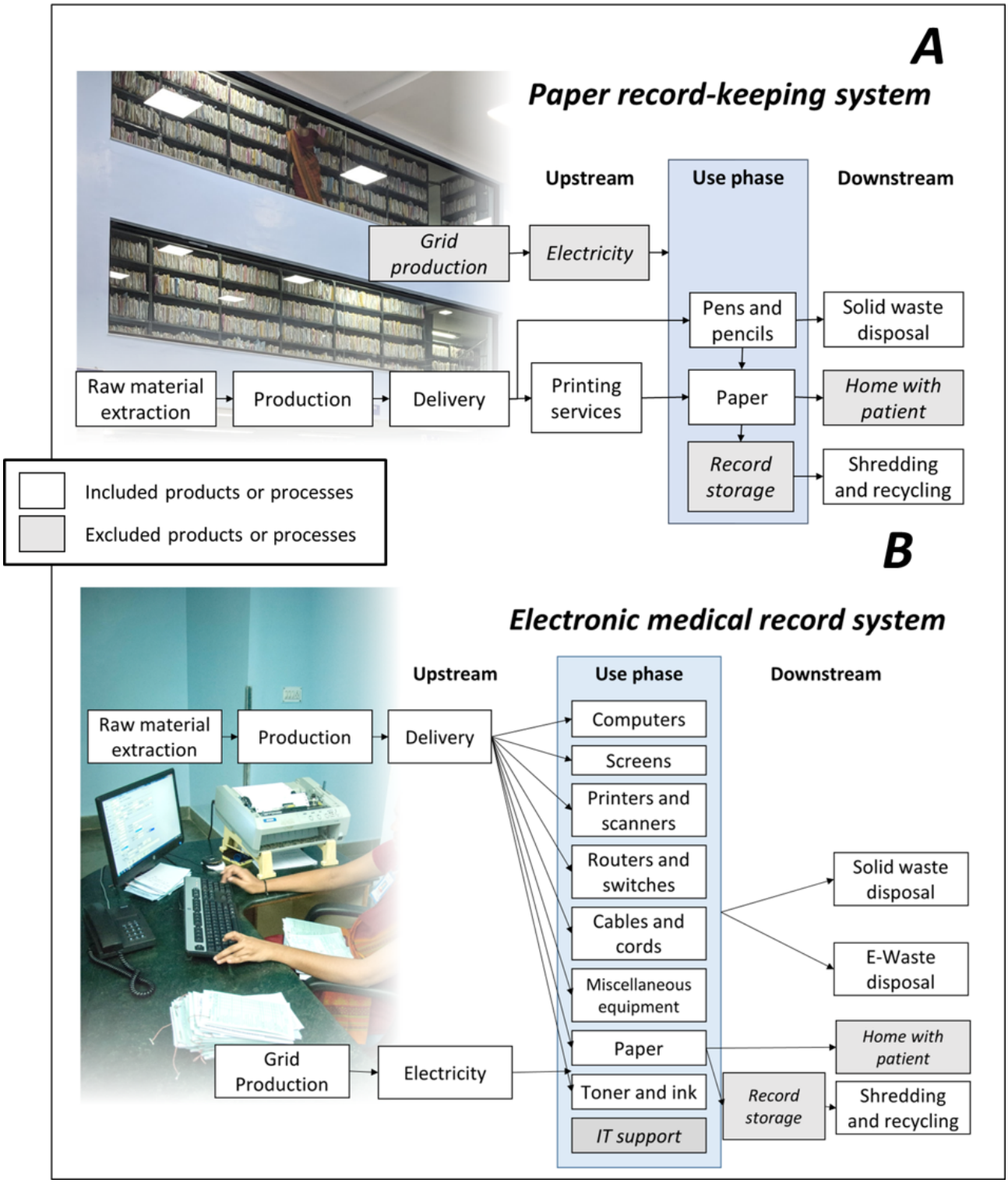

Inventory: Consumable Supplies

Aravind's paper medical records required pens, pencils, and paper forms. The study team obtained purchase records for all paper, pens, pencils, and toner purchased by Aravind in 2016 and in 2019. The excess of pens and pencils purchased in 2016 was assumed to be a requirement of paper record-keeping (ie, the quantity purchased in 2016 was subtracted from the quantity purchased in 2019 to assess the changes due to the EMR implementation). Our LCI databases did not contain unit processes for pens or pencils, so we estimated their GHG emissions from previous literature [22-24]. To estimate the environmental emissions from disposing of these items, we converted the number of items purchased into weights by directly measuring the weight of a generic pen and pencil. We assumed these products were disposed of in a sanitary landfill.

XSL•FO
RenderX

Paper purchasing data were obtained for 2016 and 2019. All forms for the paper medical record were printed by a local third party using offset printing methods. After enacting the EMR, some paper forms were still printed by Aravind in the hospital on purchased laser printers. For example, a patient consent form and medical documents are often sent home with patients. Though Aravind purchased an extra 430 toner cartridges in 2019 compared with 2016, these are not expected to be the annual purchases of toner required for the EMR system. The capacity of the cartridges purchased ranges from 1600 to 3500 pages printed, so we conservatively use the 1600-page estimate. As Aravind used 251,180 sheets of paper for their EMRs, this resulted in an estimated 157 cartridges expended.

For disposal impacts of these consumables, we did not assume any losses in weight as pens or pencils were used, but we did account for the "loss" of toner in the cartridges at their end of life by using empty cartridge specifications from the manufacturer. Aravind shreds all paper records and sells the shreds to local paper dealers for recycling. We assumed all papers were disposed of this way, even if some were taken home by patients.

Inventory: EMR Capital Outlay and Disposal

Installing the EMR system required the purchase of capital equipment and wiring infrastructure, including computers (tablets and desktops), screens, scanners, printers, routers and switches, cables and chords, other miscellaneous equipment (platforms, racks, power strips, monitor stands), and IT support. Aravind supplied a list of this equipment, its make and model, and its costs. The production and disposal impacts of these items need to be allocated across product life spans. The life span of computers and their electrical components is highly variable and dependent upon disk size, RAM availability, the condition of their hardware and systems, malware, and initial quality, as well as cultural and community factors. Here, we assume a replacement rate or product lifetime of 5 years for computers [25,26], 4 years for printers and scanners [27], 5.7 years for routers and switches [28], 9.5 years for computer screens [29], and 10 years for cables and miscellaneous equipment. Disposal impacts were estimated based on the weight of the electronic product, according to the manufacturer's product specifications. We assumed all electronic equipment was sent to e-waste recycling and all consumables were sent to a sanitary landfill.

Inventory: Energy Consumption

To estimate the energy consumed while using the EMR, we directly measured the kWh consumed by desktop computers and screens using a watt meter, and we used the manufacturer's product specifications to estimate the power consumption of remaining capital goods such as laptops or tablets, printers, scanners, and routers. We assumed cables and miscellaneous equipment consumed no additional wattage. As a reflection of the average work week and power-down policies at Aravind, we assumed that computers and digital screens are on 10 hours per day and turned off 14 hours per day, 7 days a week. This will likely overestimate their energy consumption, as Aravind typically sees fewer patients on Sundays. These assumptions were validated by Aravind's staff, and variations to these assumptions were tested in our sensitivity analyses. Routers and switches are rarely powered off at the end of a workday, so their assumed usage is 24 hours per day. For the printers, we assume they are in "printing" mode for 15 minutes each day and in "ready or on" the rest of the 24-hour day. As noted above, the energy required to cool this equipment was not included in this study. We assumed all power was drawn from the local electric grid, using the Indian national average mix of power generation sources.

For data storage, Aravind purchases 500 GB annually from the Google Cloud Platform. The energy consumption of cloud-based data storage is highly variable. In addition, Google Cloud Platform claims to be carbon neutral using carbon offsets [30]. For the purposes of this study, we estimated the power consumption of a typical hard disk using manufacturer specifications and multiplied this wattage (8.5 W) over a year. This will likely overestimate the energy consumption and emissions associated with data storage, particularly for medical records, which are not being constantly modified as other cloud-based data may be.

Life Cycle Inventory and Impact Assessment Methods

All LCI was created using SimaPro (version 9.3.0.2; Pré Consultants) [31] and the Ecoinvent (version 3.8; Ecoinvent) database [32], one of the most comprehensive LCA databases available, using an allocation and cut-off by classification approach. Two exceptions are the production of pens and pencils. A full list of LCI and chosen unit processes can be found in Table S1 in Multimedia Appendix 1 [31]. The LCIA step was conducted using the US Environmental Protection Agency's Tool for Reduction and Assessment of Chemicals and Other Environmental Impacts (TRACI; 2.1 version 1.06/US 2008; US EPA) [33]. Here, we report units of GHG emissions in kg carbon dioxide equivalents (CO_2e). Other emission categories, such as acidification, eutrophication, and air pollution, can be found in Multimedia Appendix 1.

Sensitivity Analyses and Monte Carlo Assessment

To test our assumptions and account for variations in practices outside of Aravind, we conducted multiple best- and worst-case sensitivity analyses. For the EMR, we analyzed the impact of shorter and longer life spans for capital equipment, more or less energy efficient capital equipment, and the use of solar power instead of the Indian electric grid (energy estimates from previous literature and from Ecoinvent [32]). Table S2 in Multimedia Appendix 1 contains a detailed list of sensitivity assumptions. In addition, we constructed a Monte Carlo assessment (MCA) through SimaPro software to estimate the uncertainty in our models. The unit processes we select to represent the LCI stage are representative of industry averages rather than our specific use case. For example, the unit process for computer manufacturing selected from Ecoinvent represents the manufacturing and shipping of computers globally. The manufacturer used by Aravind could be emitting more or less than the industry average. The MCA allows us to account for that uncertainty. Using 1000 runs and a 95% CI, we assessed the differences in emissions categories between Aravind's paper record-keeping system and their EMR system, with the Indian grid and with solar power. The MCA, however, excludes impacts from the production of pens and pencils (the data for

XSL•FO
RenderX

which came from non-Ecoinvent sources) and disposal pathways. The detailed results of the MCA can be found in Multimedia Appendix 1.

Ethical Considerations

This study was deemed "non-human subjects research" by Aravind Eye Care System and thus did not require ethical review board approval. No human participants were used in this study, and any data reporting the number of patients served by this system were provided, anonymized, and aggregated by the study location.

Results

The EMR system was found to emit substantially more GHGs than the paper medical record system, shown in Figure 2 and Table S1 in Multimedia Appendix 2. A majority of the EMR system's emissions (90%, or 175,800 kg CO_2e per year) are from the electricity used to run the system. The production of computers accounts for 5% of the EMR system's GHG emissions (10,300 kg CO_2e per year), and toner production accounts for another 1% (2100 kg CO_2e per year). Emissions for the paper medical record system are nearly 100% from the production of printed paper (20,700 kg CO_2e per year). Per patient, both systems generate a small quantity of emissions: 0.361 kg CO_2e for the EMR system and 0.037kg CO_2e for the paper system. Results per patient and results for other emissions impact categories can be found in Multimedia Appendix 1.

Figure 2. Annual greenhouse gas (GHG) emissions from Aravind's paper medical record-keeping system and their electronic medical record (EMR) system, with Indian electric grid and with solar power (renewable energy).

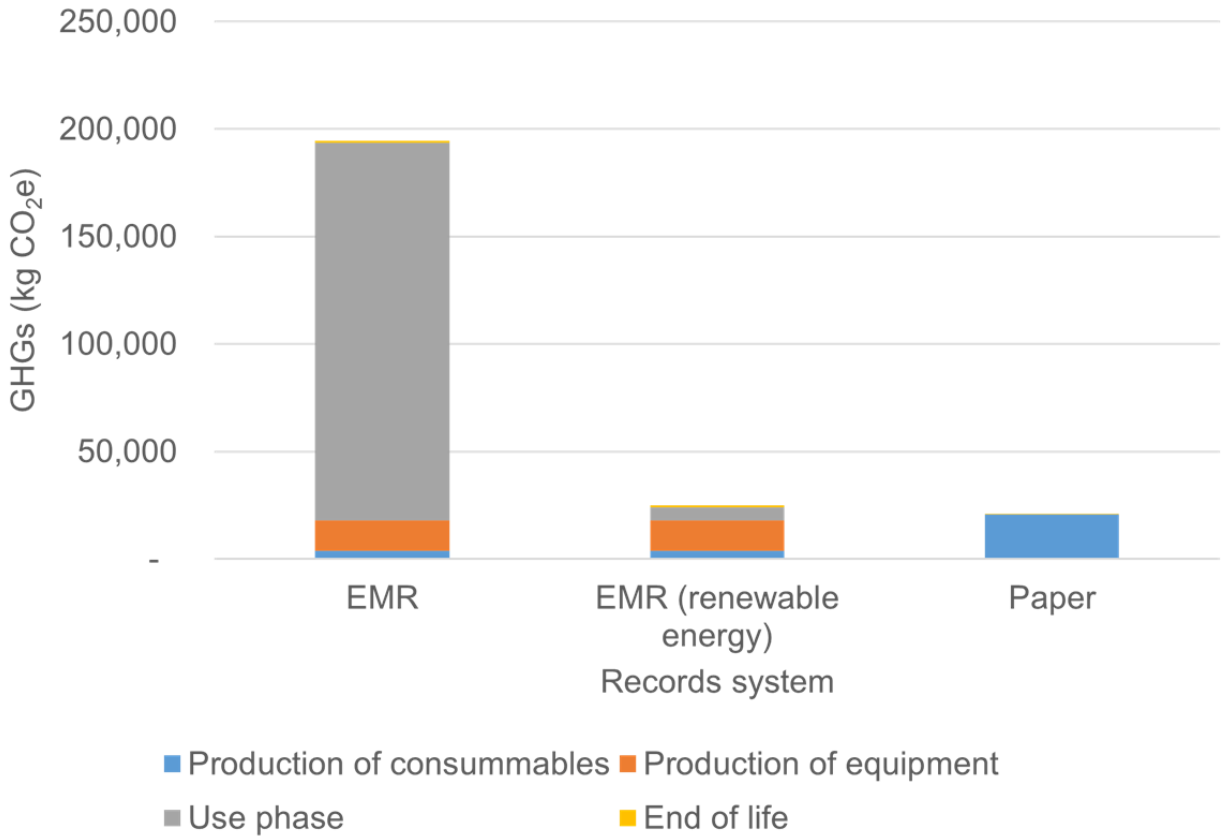

The sensitivity analyses illustrate extreme variability in GHG emissions and the potential for reducing the EMR system's emissions to a level more comparable to the paper medical record system. Renewable electricity sources seem to be the largest factor, with solar power reducing the EMR system's emissions by 87%, from 195,000 kg CO_2e to 24,900 kg CO_2e (Figure 3). More energy-efficient equipment could reduce emissions by 55% (87,100 kg CO_2e), while less energy-efficient equipment could increase emissions by 60% (312,000 kg CO_2e). Aravind is likely already using capital equipment for an optimal life span. Increasing life expectancy (from 10 to 20 years for cables or from 5 to 10 years for computers; Table S3 in Multimedia Appendix 1) could reduce emissions by 4% (187,000 kg CO_2e), but increasing the frequency of replacement to levels more common in countries like the United States (from 10 to 5 years for cables and from 5 to 1 year for computers) could increase the EMR emissions by 25% (243,000 kg CO_2e).

XSL•FO
RenderX

Figure 3. Effects of model and input variability on life cycle greenhouse gas emissions from Aravind's electronic medical record (EMR) system.

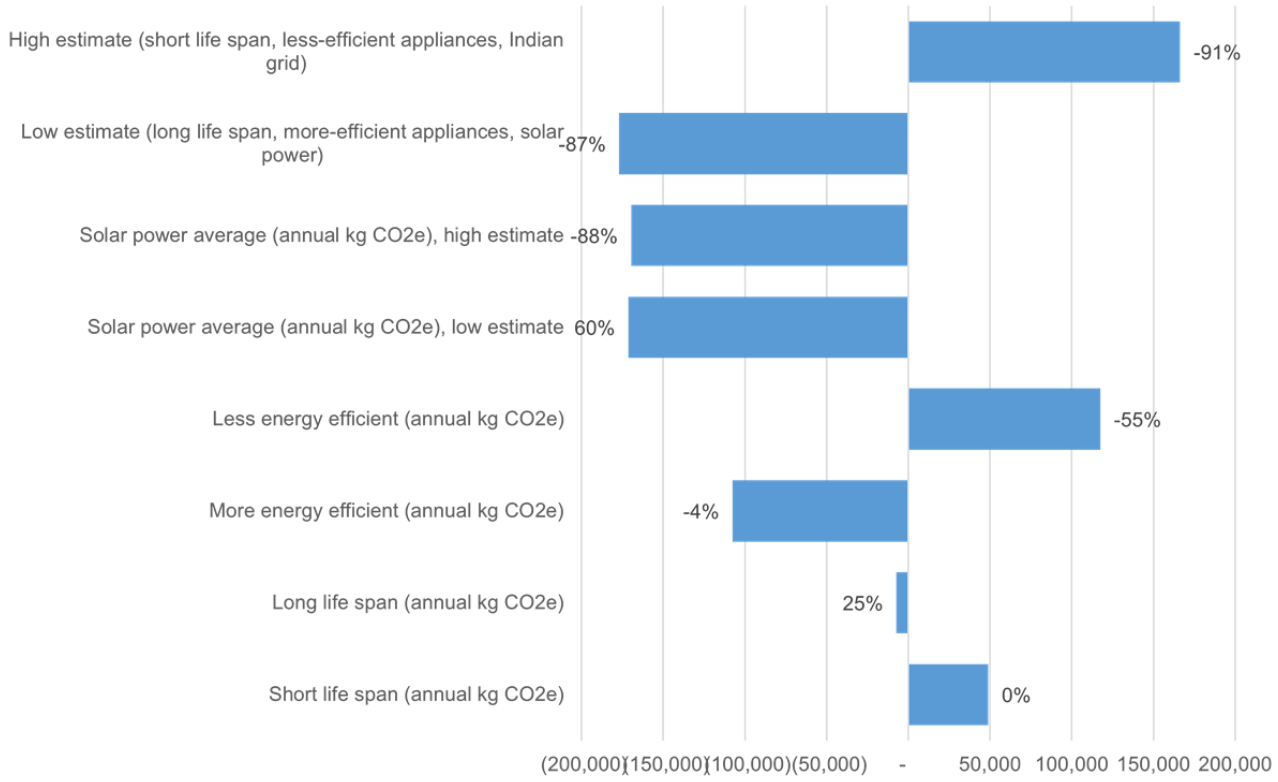

When assuming the EMR system is powered by solar panels (a renewable energy source), the emissions are more comparable to those of the paper record-keeping system, though the paper record-keeping system still outperforms a renewably powered EMR system (Figure 4). However, the share of emissions shifts when solar power is used, with the electricity required to power the EMR system representing only 25% (6170/24,900 kg CO_2e) of GHG emissions, rather than 90% 175,796/194,538 kg CO_2e) (Table S2 in Multimedia Appendix 2). The production of computers becomes the largest source of emissions, at 41%, or 10,300 kg CO_2e per year (0.02 kg CO_2e per patient). Production of capital equipment results in 57% (14,220/24,900 kg CO_2e) of emissions in this scenario, and production of consumables results in 15% (3670/24,900 kg CO_2e) of GHG emissions.

Figure 4. Greater detail of annual Greenhouse Gas (GHG) emissions from Aravind's paper medical record–keeping system and their electronic medical record (EMR) system with renewable energy (solar power).

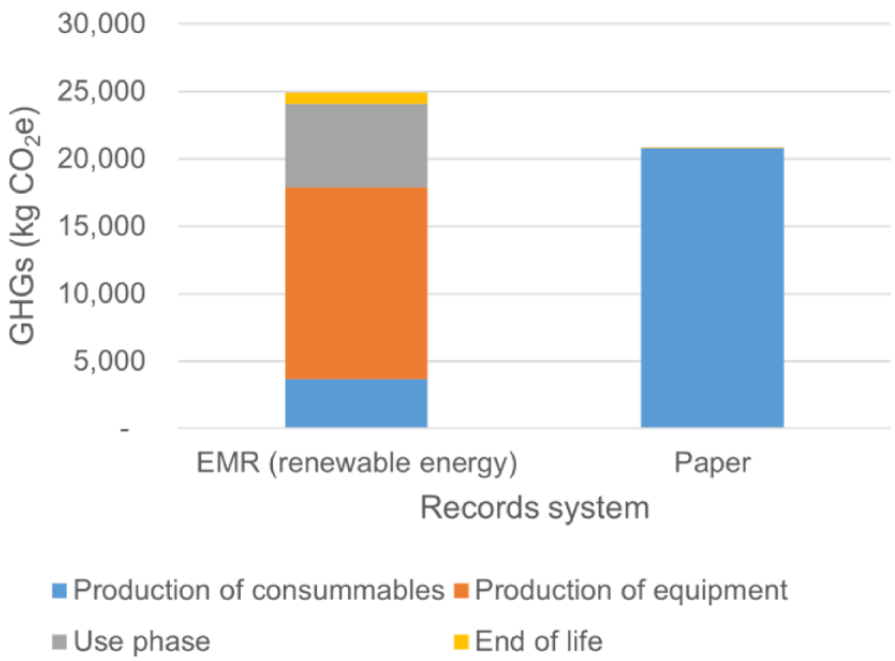

XSL•FO
RenderX

The MCA assesses model uncertainty and shows that the paper system is preferable to the EMR system on an Indian grid in nearly all environmental impact categories except carcinogenic and noncarcinogenic impacts, where they perform similarly. With solar power, the EMR system's impacts overlap more with the paper system's impacts. EMR is comparable to paper record-keeping for carcinogenic and noncarcinogenic emissions categories, where the distribution of EMR impacts is less than paper in about 50% (solar-powered EMR was less than paper record-keeping in 590/1000 runs and 515/1000, respectively) of MCA runs. For air-related categories such as fossil fuel depletion, respiratory effects, global warming, smog formation, and ozone depletion, EMR emissions are lower than paper for about 10% (0/1000, 42/1000, 7/1000, 12/1000, and 362/1000, respectively) of MCA runs (Table S5, Figure S6 in Multimedia Appendix 1). However, for water-related categories such as acidification, eutrophication, and ecotoxicity, the paper record-keeping system outperforms the solar-powered EMR system in nearly 100% (821/1000, 942/1000, and 1000/1000) of MCA runs (95% CI).

Discussion

Changing the Footprint of the EMR

Assuming the use of the Indian electric grid, Aravind's EMR system generates an estimated 195,000 kg CO_2e per year, or 0.361 kg CO_2e per patient visit. This far exceeds the GHGs emitted from their paper record-keeping system, with an estimated 20,800 kg CO_2e per year or 0.037 kg CO_2e per patient. This is equivalent to the use of 42 passenger vehicles over the course of a year for the EMR system and the use of 4.5 passenger vehicles in the case of the paper record-keeping system.

However, sensitivity analyses show that the effect of electricity sources is a major factor in determining which record-keeping system emits fewer GHGs. Indeed, this study overestimates Aravind's emissions, as they are using solar panels for about 22% of their electricity needs. Total reliance on renewable energy sources such as solar or wind power will reduce the emissions from the EMR system to a level comparable to paper record-keeping (approximately 24,900 kg CO_2e). Therefore, beyond the use of solar panels, decarbonizing electric grids could massively reduce impact while also benefiting all other industries reliant on the grid. Of note, proper sourcing of solar photovoltaic systems and other renewable technologies should be a priority, given the labor violations currently in the market [34].

Study Limitations

This study was conducted 1 year after Aravind implemented their EMR system; however, some residual and unnecessary practices likely remained from the paper-record-keeping system. At the time of this study, Aravind still used a mix of paper and electronic materials. Switching entirely away from paper may result in increased emissions from computer equipment or electricity, but it could reduce the remaining impacts from paper and toner production. A longer-term follow-up study may help assess environmental impacts after the EMR system has been fully implemented [35]. Examples of fuller use of the EMR system include barcode scanners for noting sterilization, devices that allow patients to sign consent forms digitally, or increased training for staff to increase their comfort levels with the EMR system.

This study was conducted at a single institution, and the EMR system set up here may be unique to this setting. We have included sensitivity analyses in part to account for variation in practice, but this study's results may not be translatable to other settings. In addition, some elements of both systems were not included in this study. The paper record-keeping system did use some computers and printing, which we did not account for. We have also made assumptions around the life spans of EMR technologies, EMR energy use, and the number of consumables used in each system, which can change the emissions profiles of both systems.

The Unmodeled Environmental Impacts of EMRs

Regardless of electricity source, implementing EMRs, particularly in emerging and expanding health systems in LMICs, has a variety of benefits that are important to consider and have not been captured in this LCA. EMRs improve patient overall care delivery outcomes through enhanced care coordination, enabled data sharing, and decreased clinical errors. Providing the right care to the right patient at the right time ensures emissions from medical practice are not unnecessary [36]. One clinical error that also causes significant waste and unnecessary emissions, ordering duplicative lab tests, has been significantly eliminated in health systems after the adoption of EMRs [37]. Simultaneously, EMRs can increase efficiencies within hospitals and reduce costs in health care systems [7,8]. Aravind was able to reclaim hospital space previously allocated to record storage, facilitating more patient care. In addition, 50% of the staff needed for medical record keeping were reassigned to other critical tasks.

There are other environmental impacts of the transition to EMRs that are not captured by this LCA, a primary one being facilitating telemedicine through enabling the digital transfer of information between patient and doctor. Previous evidence in the US context suggests that transitioning away from paper record-keeping can reduce deforestation from paper production and decrease transportation to hospitals due to increased telemedicine visits or email correspondences [38]. This supports a sizable body of research from high-income countries that points to telemedicine reducing the carbon footprint of appointments, primarily by reducing transport-associated emissions, although the exact carbon footprint savings can vary widely depending on the context [39,40]. Supporting telemedicine has a multitude of other benefits, particularly for LMICs, as the expansion of telemedicine improves access for many vulnerable and geographically remote patients. Since implementing an EMR system, Aravind has created a network of clinics that videoconference with physicians in the main hospital, enabling greater access and triage for patients while maximizing the time and reach of clinicians. Likewise, it has reduced travel and travel-related environmental impacts for providing necessary care.

XSL•FO
RenderX

Conclusion

EMRs have the potential to transform how health care is delivered in LMICs. However, EMRs have an environmental footprint that should be considered, given the devastating health consequences of climate change. When implementing an EMR system, health care systems should use energy-efficient technologies wherever possible and develop processes and maintenance protocols to increase the life span of these products. Moreover, investing in strategies to decarbonize electric grids is essential to ensuring that health care expansion is conducted more sustainably. This could be done locally by individual hospitals through the installation of solar or wind power or more strategically at a national level, which would have greater influence on other industries as well. Improving the environmental performance of EMRs is an important aspect of making healthcare safe and accessible across the world.

Acknowledgments

The authors would like to thank Dr David Chang and the American Society of Cataract and Refractive Surgery (ASCRS) organization for their contributions to the article processing fees.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental information (tables and figures).
[DOCX File , 489 KB - jmir_v26i1e42140_app1.docx]

Multimedia Appendix 2
Supplementary tables.
[DOCX File , 30 KB - jmir_v26i1e42140_app2.docx]

References

1. Giacalone A, Marin L, Febbi M, Franchi T, Tovani-Palone MR. eHealth, telehealth, and telemedicine in the management of the COVID-19 pandemic and beyond: lessons learned and future perspectives. World J Clin Cases 2022 Mar 16;10(8):2363-2368 [FREE Full text] [doi: 10.12998/wjcc.v10.i8.2363] [Medline: 35434056]
2. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;25(Suppl 1):S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
3. Hoffer-Hawlik MA, Moran AE, Burka D, Kaur P, Cai J, Frieden TR, et al. Leveraging telemedicine for chronic disease management in low- and middle-income countries during Covid-19. Glob Heart 2020;15(1):63 [FREE Full text] [doi: 10.5334/gh.852] [Medline: 33150128]
4. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
5. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
6. Boonstra A, Versluis A, Vos JFJ. Implementing electronic health records in hospitals: a systematic literature review. BMC Health Serv Res 2014;14(1):370 [FREE Full text] [doi: 10.1186/1472-6963-14-370] [Medline: 25190184]
7. Adane K, Gizachew M, Kendie S. The role of medical data in efficient patient care delivery: a review. Risk Manag Healthc Policy 2019;12:67-73 [FREE Full text] [doi: 10.2147/RMHP.S179259] [Medline: 31114410]
8. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
9. Hodgson T, Burton-Jones A, Donovan R, Sullivan C. The role of electronic medical records in reducing unwarranted clinical variation in acute health care: systematic review. JMIR Med Inform 2021;9(11):e30432 [FREE Full text] [doi: 10.2196/30432] [Medline: 34787585]
10. Li R, Niu Y, Scott SR, Zhou C, Lan L, Liang Z, et al. Using electronic medical record data for research in a Healthcare Information and Management Systems Society (HIMSS) analytics Electronic Medical Record Adoption Model (EMRAM) stage 7 hospital in Beijing: cross-sectional study. JMIR Med Inform 2021;9(8):e24405 [FREE Full text] [doi: 10.2196/24405] [Medline: 34342589]

XSL•FO
RenderX

11. Seino Y, Sato N, Idei M, Nomura T. The reduction in medical errors on implementing an intensive care information system in a setting where a hospital electronic medical record system is already in use: retrospective analysis. JMIR Perioper Med 2022;5(1):e39782 [FREE Full text] [doi: 10.2196/39782] [Medline: 35964333]
12. Karliner J, Slotterback S, Boyd R, Ashby B, Steele K, Wang J. Health care's climate footprint: the health sector contribution and opportunities for action. Eur J Public Health 2019;30(Supplement_5):ckaa165.843 [FREE Full text] [doi: 10.1093/eurpub/ckaa165.843]
13. Parasrampuria S, Henry J. Hospitals' use of electronic health records data, 2015-2017. ONC Data Brief 2019;46:1-13 [FREE Full text] [doi: 10.1201/9781351057790-2]
14. Rasheed FN, Baddley J, Prabhakaran P, De Barros EF, Reddy KS, Vianna NA, et al. Decarbonising healthcare in low and middle income countries: potential pathways to net zero emissions. BMJ 2021;375:n1284 [FREE Full text] [doi: 10.1136/bmj.n1284] [Medline: 34753746]
15. Haripriya A, Chang DF, Ravindran RD. Endophthalmitis reduction with intracameral moxifloxacin prophylaxis: analysis of 600 000 surgeries. Ophthalmology 2017;124(6):768-775 [FREE Full text] [doi: 10.1016/j.ophtha.2017.01.026] [Medline: 28214101]
16. Le HG, Ehrlich JR, Venkatesh R, Srinivasan A, Kolli A, Haripriya A, et al. A sustainable model for delivering high-quality, efficient cataract surgery in Southern India. Health Aff (Millwood) 2016;35(10):1783-1790 [FREE Full text] [doi: 10.1377/hlthaff.2016.0562] [Medline: 27702949]
17. Venkatesh R, van Landingham SW, Khodifad AM, Haripriya A, Thiel CL, Ramulu P, et al. Carbon footprint and cost-effectiveness of cataract surgery. Curr Opin Ophthalmol 2016;27(1):82-88 [FREE Full text] [doi: 10.1097/ICU.0000000000000228] [Medline: 26569528]
18. Thiel CL, Schehlein E, Ravilla T, Ravindran RD, Robin AL, Saeedi OJ, et al. Cataract surgery and environmental sustainability: waste and lifecycle assessment of phacoemulsification at a private healthcare facility. J Cataract Refract Surg 2017;43(11):1391-1398 [FREE Full text] [doi: 10.1016/j.jcrs.2017.08.017] [Medline: 29223227]
19. Drew J, Christie SD, Tyedmers P, Smith-Forrester J, Rainham D. Operating in a climate crisis: a state-of-the-science review of life cycle assessment within surgical and anesthetic care. Environ Health Perspect 2021;129(7):76001 [FREE Full text] [doi: 10.1289/EHP8666] [Medline: 34251875]
20. Sherman JD, MacNeill A, Thiel C. Reducing pollution from the health care industry. JAMA 2019;322(11):1043-1044. [doi: 10.1001/jama.2019.10823] [Medline: 31373600]
21. ISO 14040:2006: Environmental management: Life cycle assessment: Principles and framework. International Organization for Standardization. 2006. URL: https://www.iso.org/standard/37456.html [accessed 2023-12-08]
22. Robbins K. Wood pencil carbon life cycle. Prezi. 2014. URL: https://prezi.com/rt0ti5r55vvd/carbon-life-cycle-of-a-wood-pencil/ [accessed 2023-12-08]
23. It's just a pen, right? EthicalStationery. 2019. URL: https://ethicalstationery.com/its-just-a-pen-right/ [accessed 2023-12-08]
24. Life cycle analysis of wooden pencil. portfolium.com. URL: https://portfolium.com/entry/life-cycle-analysis-of-wooden-pencil [accessed 2023-12-08]
25. Yu J, Williams E, Ju M, Yang Y. Forecasting global generation of obsolete personal computers. Environ Sci Technol 2010;44(9):3232-3237. [doi: 10.1021/es903350q] [Medline: 20302363]
26. Babbitt CW, Kahhat R, Williams E, Babbitt GA. Evolution of product lifespan and implications for environmental assessment and management: a case study of personal computers in higher education. Environ Sci Technol 2009;43(13):5106-5112 [FREE Full text] [doi: 10.1021/es803568p] [Medline: 19673314]
27. Papiewski J. What is the life span of a laser printer? ItStillWorks. 2019. URL: https://yourbusiness.azcentral.com/life-span-laser-printer-11859.html [accessed 2023-12-08]
28. How much is the life of a WiFi router device? Quora. 2023. URL: https://www.quora.com/How-much-is-the-life-of-a-WiFi-router-device [accessed 2023-12-08]
29. Fakhredin F, Huisman J. Analyzing end of life LCD TV WEEE flows in Europe. 2013 Presented at: Proceedings of EcoDesign 2013 International Symposium; 2013; eju Island, South Korea.
30. Corio A. Five years of 100% renewable energy–and a look ahead to a 24/7 carbon-free future. Google Cloud. 2022. URL: https://cloud.google.com/blog/topics/sustainability/5-years-of-100-percent-renewable-energy [accessed 2023-12-08]
31. LCA software for informed changemakers. SimaPro. 2022. URL: https://simapro.com/ [accessed 2023-12-13]
32. Frischknecht R, Jungbluth N, Althaus HJ, Doka G, Dones R, Heck T, et al. The ecoinvent database: overview and methodological framework (7 pp). Int J Life Cycle Assessment 2004;10(1):3-9 [FREE Full text] [doi: 10.1065/lca2004.10.181.1]
33. Bare J. TRACI 2.0: the tool for the reduction and assessment of chemical and other environmental impacts 2.0. Clean Techn Environ Policy 2011;13(5):687-696 [FREE Full text] [doi: 10.1007/s10098-010-0338-9]
34. Cockayne J, Huerta E, Burcu O. The energy of freedom'? Solar Energy, Modern Slavery and the Just Transition. 2022. URL: https://www.thebritishacademy.ac.uk/documents/4338/Just-transitions-energy-freedom1.pdf [accessed 2023-12-08]
35. Garcia G, Crenner C. Comparing international experiences with electronic health records among emergency medicine physicians in the United States and Norway: semistructured interview study. JMIR Hum Factors 2022;9(1):e28762 [FREE Full text] [doi: 10.2196/28762] [Medline: 34994702]

XSL•FO
RenderX

36. Sherman JD, Thiel C, MacNeill A, Eckelman MJ, Dubrow R, Hopf H, et al. The green print: advancement of environmental sustainability in healthcare. Resour Conserv Recycl 2020;161:104882 [FREE Full text] [doi: 10.1016/j.resconrec.2020.104882]
37. Kerwin TC, Leighton H, Buch K, Avezbadalov A, Kianfar H. The effect of adoption of an electronic health record on duplicate testing. Cardiol Res Pract 2016;2016:1950191 [FREE Full text] [doi: 10.1155/2016/1950191] [Medline: 27088033]
38. Turley M, Porter C, Garrido T, Gerwig K, Young S, Radler L, et al. Use of electronic health records can improve the health care industry's environmental footprint. Health Aff (Millwood) 2011;30(5):938-946 [FREE Full text] [doi: 10.1377/hlthaff.2010.1215] [Medline: 21555478]
39. Purohit A, Smith J, Hibble A. Does telemedicine reduce the carbon footprint of healthcare? A systematic review. Future Healthc J 2021;8(1):e85-e91 [FREE Full text] [doi: 10.7861/fhj.2020-0080] [Medline: 33791483]
40. Donald N, Irukulla S. Greenhouse gas emission savings in relation to telemedicine and associated patient benefits: a systematic review. Telemed J E Health 2022;28(11):1555-1563. [doi: 10.1089/tmj.2022.0047] [Medline: 35446668]

Abbreviations

CO2e: carbon dioxide equivalents
EMR: electronic medical record
GHG: greenhouse gas
ISO: International Organization for Standardization
LCA: life cycle assessment
LCI: life cycle inventory
LCIA: life cycle impact assessment
LMICs: low- and middle-income countryies
MCA: Monte Carlo assessment
TRACI: Tool for Reduction and Assessment of Chemicals and Other Environmental Impacts

Edited by A Mavragani; submitted 13.09.22; peer-reviewed by M Jones, P Han; comments to author 16.02.23; revised version received 22.03.23; accepted 19.04.23; published 06.02.24.

Please cite as:
Kwon C, Essayei L, Spencer M, Etheridge T, Venkatesh R, Vengadesan N, Thiel CL
The Environmental Impacts of Electronic Medical Records Versus Paper Records at a Large Eye Hospital in India: Life Cycle Assessment Study
J Med Internet Res 2024;26:e42140
URL: https://www.jmir.org/2024/1/e42140
doi:10.2196/42140
PMID:38319701

©Cordelia Kwon, Lernik Essayei, Michael Spencer, Tom Etheridge, Rengaraj Venkatesh, Natrajan Vengadesan, Cassandra L Thiel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing and Implementation of a Digitalized Intersectoral Discharge Management System and Its Effect on Readmissions: Mixed Methods Approach

Christoph Strumann[1], PhD; Lisa Pfau[1], MD; Laila Wahle[2], MBA; Raphael Schreiber[1], MD; Jost Steinhäuser[1], Prof Dr Med, MD

[1]Institute of Family Medicine, University Medical Centre Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

[2]Lacanja GmbH Health Innovation Port, Hamburg, Germany

Corresponding Author:
Christoph Strumann, PhD
Institute of Family Medicine
University Medical Centre Schleswig-Holstein, Campus Lübeck
Ratzeburger Allee 160
Lübeck, 23538
Germany
Phone: 49 451 3101 8005
Email: c.strumann@uni-luebeck.de

Abstract

Background: Digital transformation offers new opportunities to improve the exchange of information between different health care providers, including inpatient, outpatient and care facilities. As information is especially at risk of being lost when a patient is discharged from a hospital, digital transformation offers great opportunities to improve intersectoral discharge management. However, most strategies for improvement have focused on structures within the hospital.

Objective: This study aims to evaluate the implementation of a digitalized discharge management system, the project "Optimizing instersectoral discharge management" (SEKMA, derived from the German Sektorübergreifende Optimierung des Entlassmanagements), and its impact on the readmission rate.

Methods: A mixed methods design was used to evaluate the implementation of a digitalized discharge management system and its impact on the readmission rate. After the implementation, the congruence between the planned (logic model) and the actual intervention was evaluated using a fidelity analysis. Finally, bivariate and multivariate analyses were used to evaluate the effectiveness of the implementation on the readmission rate. For this purpose, a difference-in-difference approach was adopted based on routine data of hospital admissions between April 2019 and August 2019 and between April 2022 and August 2022. The department of vascular surgery served as the intervention group, in which the optimized discharge management was implemented in April 2022. The departments of internal medicine and cardiology formed the control group.

Results: Overall, 26 interviews were conducted, and we explored 21 determinants, which can be categorized into 3 groups: "optimization potential," "barriers," and "enablers." On the basis of these results, 19 strategies were developed to address the determinants, including a lack of networking among health care providers, digital information transmission, and user-unfriendliness. On the basis of these strategies, which were prioritized by 11 hospital physicians, a logic model was formulated. Of the 19 strategies, 7 (37%; eg, electronic discharge letter, providing mobile devices to the hospital's social service, and generating individual medication plans in the format of the national medication plan) have been implemented in SEKMA. A survey on the fidelity of the application of the implemented strategies showed that 3 of these strategies were not yet widely applied. No significant effect of SEKMA on readmissions was observed in the routine data of 14,854 hospital admissions (P=.20).

Conclusions: This study demonstrates the potential of optimizing intersectoral collaboration for patient care. Although a significant effect of SEKMA on readmissions has not yet been observed, creating a digital ecosystem that connects different health care providers seems to be a promising approach to ensure secure and fast networking of the sectors. The described intersectoral optimization of discharge management provides a structured template for the implementation of a similar local digital care networking infrastructure in other care regions in Germany and other countries with a similarly fragmented health care system.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47133) doi:10.2196/47133

KEYWORDS

digitalization; intersectoral; discharge management; readmission; mixed methods design

Introduction

Background

Digital patient process systems offer several advantages over analog systems. On the one hand, this can lead to more systematic, targeted use of resources, and on the other hand, easier communication and transmission of data can enable better coordination of the various cooperating partners [1]. Patient records are becoming increasingly digitalized, with some countries being prototypes in this area, such as Latvia, Denmark, and Spain [2].

In Germany, there have been several governmental attempts to shape different elements of health care digitalization. A recent example is the Hospital Future Act (Krankenhauszukunftsgesetz) from 2020. It was designed to support digitalization in hospitals by promoting the technical equipment of hospitals through state-funded investments. The investments are expected to improve process organization, documentation, and communication (internal, sectoral, and intersectoral) [3]. The results suggest that the Hospital Future Act, together with the COVID-19 pandemic, led to an increase in the digital maturity of hospitals and, thus, reduced the digitalization backlog [4]. Another approach to promote health care digitalization is the introduction of an electronic health record (EHR) within a secure telematics infrastructure. The EHR should not only simplify rapid communication within and across different health care institutions but also enable further eHealth applications, for example, electronic prescriptions [5]. However, the introduction of EHRs as well as other reforms promoting health care digitalization have been accompanied with strong resistance underpinned by arguments of data protection and security as well as by technical problems. Especially in the outpatient sector, the latter has resulted in a perceived disproportionate administrative effort without adequate financial compensation for the care providers such as private practices [6]. As a result, Germany lags behind other industrialized countries in the digitalization of the health care system [7,8].

An EHR could make treatment pathways more transparent and improve communication between different health care providers, including inpatient, outpatient and care facilities [9]. The exchange of information is particularly susceptible if a patient is discharged from hospital. With regard to the strongly pronounced sectoral separation in Germany [10,11], information loss is particularly high between inpatient and outpatient care. Moreover, owing to the accelerated tendency toward shortening the length of stay of patients in the inpatient sector as a result of the introduction of the diagnosis-related group–based reimbursement system [12], hospitals no longer provide care and treatment until full recovery [13]. Instead, parts of the treatment and recovery process are moved to the posthospital setting [14]. Similar developments have been observed after introducing the diagnosis-related group–based reimbursement system in other countries, for example, the United States [15-17]. Shortened length of stay and ineffectively designed transitions are associated with adverse events, higher risks of readmission, and higher costs [18-21]. Up to 1 (18%) in 5 patients are readmitted to the hospital within 30 days of discharge [22,23]. Individualized discharge management can reduce the number of readmissions of older patients with a health problem [24], leading to potential cost savings for the health care system [25]. To date, many strategies to improve discharge management have focused on structures within the hospital. However, to ensure a holistic and continuous treatment, the cooperation between different health care providers from the inpatient and outpatient sectors as well as care facilities should also be considered.

As there is still no EHR accessible to all caregivers in Germany, experience with digitalized health information systems has been gathered only in model projects, which are intended to provide insights into possible barriers and enablers for a successful implementation [5,26-30].

Objectives

This study aims to explore the determinants of a digitalized discharge management system, to implement such a system within 1 area, and to evaluate its impact on the readmission rate.

Methods

The evaluation was done within the project “Optimizing intersectoral discharge management” (SEKMA, derived from the German Sektorübergreifende Optimierung des Entlassmanagements).

Study Design

A mixed methods design was chosen to evaluate SEKMA. Owing to the complexity of the intervention, the evaluation was based on the framework of developing tailored interventions [31]. This approach allows a detailed description and analysis of the components of the intervention that contributed to its effectiveness or ineffectiveness. For this purpose, this framework distinguishes between a development and an application phase. In the first step, barriers and enabling factors for a successful implementation of a digitalized discharge management system such as SEKMA were explored using qualitative research methods, that is, interviews. Second, strategies were developed for addressing these determinants. Third, these strategies were prioritized using a (quantitative) questionnaire, and a logic model was formulated to describe the logical linkages among the resources and activities needed to achieve the results. After the implementation (application phase), the congruence between the planned intervention (logic model) and the implemented intervention was evaluated. In this step, the fidelity of the use of the different strategies in the routine was examined [32]. Finally, the effectiveness of the implementation on the readmission rate (outcome) was evaluated based on routine data of hospital admissions.

XSL•FO
RenderX

Setting

The digitalized discharge management system was implemented at a medium-sized hospital (approximately 350 beds) in the northern German federal state Schleswig-Holstein in the Metropolitan area of Hamburg, the second-largest city in Germany. Before the intervention, the internal and external exchange of information was typically performed by phone, fax, and email. As the network between the various caregivers was rather weak, communication occurred only on request, tying up resources and causing delays in the transfer of information.

SEKMA aimed to develop and implement a digitalized, intersectoral discharge management system that considers the patient's entire treatment pathway, from hospital admission to possible admission to a care facility, and the follow-up treatment by general practitioners (GPs). All information relevant to ongoing (postinpatient) treatment and care should be available quickly and easily to all care providers involved. This includes providers from the inpatient and outpatient sectors as well as care facilities. For this purpose, an ecosystem of hospital and postinpatient care facilities has been implemented within a digital infrastructure based on a standardized and harmonized IT system for data exchange [33]. The workflow of the digitalized, intersectoral discharge management can be described as follows:

1. The hospital coordinates and organizes follow-up care in a timely manner based on the patient's agreement with the hospital's discharge management.
2. A discharge plan for medication, follow-up care, and rehabilitation is created and all professionals in the hospital are involved. This includes admission staff, medical service, nursing service, social service, and the patient information system.
3. In cooperation with the nursing staff and social service, the patient is informed and advised about care options and structures that correspond to their illness. The contents are prepared digitally.
4. The patient is discharged from the hospital and transitions to outpatient, rehabilitative, or nursing care. All documents necessary for discharge and further treatment are available digitally and can be transmitted directly to the relevant sectors.
5. If a patient contacts a primary care physician for outpatient follow-up treatment, the patient's digital discharge documentation is already in the system of the private practice.
6. In case of a query or deterioration of health status, the primary care physician can contact the hospital and previously treating physicians directly.
7. If there is a readmission, the hospital can digitally access documentation on posttreatment care and procedures, as well as the medical history, at any time and continue treatment directly. The same applies to nursing and rehabilitation facilities.

The information transfer across the distinct health care provider is organized via KIM (Kommunikation im Medizinwesen) embedded in the telematics infrastructure. All organizations involved in the project have a KIM connection. Using KIM, participants can transmit documents in a secure and encrypted manner [34]. Overall, all communication processes have been digitalized compared with before the intervention. Since April 2022, optimized discharge management has been implemented in the department of vascular surgery.

Individual Interviews

Enabling factors and barriers leading toward successful digital discharge management were identified through individual interviews with physicians, medical assistants, social workers and nurses at the hospital, GPs, and staff from nursing homes and care services. This was performed using the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines for qualitative studies (Multimedia Appendix 1 provides details of the COREQ guidelines [35]). Originally, a combination of interviews and focus groups was planned. Owing to the COVID-19 pandemic, focus groups had to be abandoned.

The hospital, along with collaborating partners such as physician networks and nursing homes, conducted participant recruitment for interviews through face-to-face interactions, telephone calls, and emails. Previously developed partially standardized interview guidelines were used and pilot tested (Multimedia Appendix 2). The interviews were conducted by telephone by a medical student (LP) between April 30, 2020, and October 9, 2020, at the workplace of the interviewees. A theoretical saturation effect in the statements made during the interviews resulted in the final number of interviewees.

The individual interviews were conducted in a protected setting and subsequently pseudonymized, thus providing the opportunity to explore the personal opinions of the interviewees beyond any possible social group pressures. The interviews were recorded using a digital dictaphone and were transcribed orthographically. The material was subsequently analyzed using structured content analysis according to Mayring [36]. The development of the categories was initially based on the questions (deductive) listed in the partially standardized interview guideline (Multimedia Appendix 2). In addition, categories were extracted from the text (inductive). Five persons were involved in the development of the category scheme (LP: medical student [female researcher], JS: GP and experienced health service researcher including qualitative research [male researcher], CS: health economist with some experience in qualitative research [male researcher], a legal project advisor [female researcher], and a physiotherapist [female researcher]; all of them except LP were employed at the Institute of Family Medicine at the University of Lübeck at the time of the analysis). After individual coding, a coding scheme was discussed in a consensus meeting. The final coding scheme was applied to the interview material.

Development and Evaluation of Strategies

On the basis of the described processes for treating the patients, the optimization potential, and the determinants from the evaluated individual interviews as well as the workshop with clinicians and physicians in private practice, strategies for the implementation of optimized discharge management were developed. These strategies were developed in such a way that they addressed the determinants identified and were, thus, conducive to a successful implementation.

XSL•FO
RenderX

During a project meeting on February 3, 2022, employees of the Institute of Family Medicine at the University of Lübeck and the chief and senior physicians of the involved hospital discussed these results. Subsequently, the hospital's chief or senior physicians were invited to evaluate each identified strategy according to its relevance and feasibility using a 6-point Likert scale (very high, rather high, high, rather low, low, and very low) to avoid the central tendency bias.

The resulting list of the ranked strategies formed the logic model. This model was finally compared with the list of strategies implemented in the project.

Routine Data Analysis

The focus of the evaluation of the optimized discharge management was the reduction of (unnecessary) readmissions. With the help of the evaluation of the routine admission data of the involved hospital, the effect of optimized discharge management on rehospitalization was analyzed.

Routine Data and Study Design

The hospital extracted routine data from its internal patient information system. The extracted data were provided by the hospital in an anonymized form. For each inpatient case, the data consisted of information on the date of admission and discharge, the reason for admission and discharge, diagnoses and conducted medical procedures, demographic information of the patients, and the department or departments where the patients had been treated.

Within the framework of a longitudinal study design, a pre- and postcomparison was performed. The intervention group was the department of vascular surgery, in which the optimized discharge management was implemented since April 2022. A case was assigned to the intervention group if the patient was admitted to or discharged from the department of vascular surgery. The outbreak of COVID-19 during the sample period might have affected the readmissions of the entire hospital. To minimize the risk of bias owing to the pandemic on the intervention effect, in addition to the pre-post comparison of the department of vascular surgery, a control group comparison was applied to enrich the empirical strategy. To ensure that the patients in the intervention group were as similar as possible to those in the control group, the departments of internal medicine (medical clinic) and cardiology formed the control group.

Statistical Analysis

The effect of the implementation was estimated using the difference-in-difference (DiD) approach. The sample covers the period from 2019 to August 2022. To counteract the possible COVID-19 pandemic bias, patients admitted between January 2020 and March 2022 were not considered in the analysis. To avoid any seasonal influences on the results, we restricted the preintervention period such that it covered exactly the period after the implementation, that is, from April to August. Therefore, the baseline period (T_0) consisted of April 1, 2019, to August 31, 2019, whereas the intervention period (T_1) started from April 1, 2022.

In addition to the bivariate analysis, a multivariate logistic regression model was applied. By including control variables, differences between patients from the intervention and control group were minimized. In the first step, risk factors for rehospitalization were determined by estimating separate bivariate logistic regression models. The identified risk factors served as control variables in the multivariate DiD regression analysis. A P value <.05 was considered statistically significant. Statistical analyses were performed with Stata (version 15; StataCorp LLC).

Ethical Considerations

The study was approved by the ethics committee of the University of Lübeck before recruitment commenced on December 11, 2019 (approval number 19-387). This study was conducted in accordance with the Declaration of Helsinki.

All participants provided verbal and written informed consent for their participation in the interviews and surveys. The participants were informed that they could withdraw their consent at any time. No identifiable information was recorded to ensure the confidentiality of the participants. No compensation was paid for participation.

For the analysis of routine hospital data, only anonymized data were transferred to the evaluating institution. Owing to the anonymization of the data, no additional informed consent was required to perform the routine data analysis in accordance with German law, ethical standards, and the Declaration of Helsinki. No data requiring informed consent will be presented in the routine data analysis. The ethics committee of the University of Lübeck waived the requirement for informed consent owing to the retrospective nature of this study.

Results

Interviews

Sample

A total of 26 interviews were conducted. These consisted of 14 employees of the hospital (3 doctors, 4 nurses, 4 social workers, and 3 administrative staff), 9 employees from nursing homes or mobile nursing services, and 3 GPs. The average age of the participants was 42.4 (SD 8.9; range 25-65) years, and the proportion of female participants was 54% (14/26). The average interview duration was 33 minutes and 11 seconds. An overview of the characteristics of the interview participants is provided in Table S1 in Multimedia Appendix 3.

Categories

A total of 21 determinants were explored with various subcategories for the introduction of successful digitalized discharge management. These could be divided into 3 categories: "optimization potential," "barriers," and "enablers." The aspects mentioned for optimizing the discharge process covered all areas from admission to follow-up and included inter- and intrasectoral transmission of information (Textbox 1).

XSL•FO
RenderX

Textbox 1. Potential for optimization of discharge management.

Category and subcategories

- Transmission of information
  - Preliminary discharge letter before discharge
  - Final discharge letter at the time of discharge
  - Digital transmission (mail, chat, and video call)
  - Platform for information exchange
  - Standardized information
  - Increased readiness to communicate
- Admission
  - Information exchange at admission
  - Consent to discharge management
- Preparation for discharge
  - Awareness of the existence of discharge management in the hospital
  - Timely completion of the discharge process
  - Continuous preparation for (unplanned) discharge
  - Improvement of patient communication
  - Faster approvals by health insurances
- Discharge
  - Discharge in the morning of the working day
  - Material transfer, issuing of prescriptions and incapacity certificate
  - Nursing services accompany discharge from hospital
  - Increase in the availability of patient transport
- Aftercare
  - Visits to general practitioner after discharge
  - More aftercare places
- Process standardization (standard operating procedure)
  - Training on discharge
  - Digital checklist
  - Standardized processes
  - Clarified responsibilities
- Intersectoral optimization
  - Knowledge of the performance and processes at other facilities
- Structural
  - Evaluation of criticism or review
  - Supervision
  - Ethics committee

In the German health care system, the discharge letter is at the center of information transmission between the inpatient and outpatient sectors. Participants saw a need for improvement in the early, or at least timely, delivery of this letter. In the best case, information would already be transmitted during the hospital stay to the follow-up service providers such as private practices or care facilities:

XSL•FO
RenderX

> *To have all the information and data, everything before the patient arrives here. That would be the absolute dream. [...] You can just admit the person better[...] if you just have preliminary information.* [P03]

Digital transmission of data was also perceived as beneficial; the participants could imagine using conventional media such as email or video calls as well as via a platform provided specifically for this purpose:

> *If you could even find some other common platform where information can be exchanged.* [P01]

Furthermore, the potential for optimization was seen in the standardization of the information. The information to be communicated should be transmitted through a central entity, and at the same time, selected contacts who can be reached on demand and who can provide information about the patients would be beneficial:

> *Yes, standard, standard, standard. So, that you try to agree on what information I need and then it has to appear—in a structured form, so in principle already like my patient information.* [P10]

Some participants also noted that, in principle, a greater willingness to communicate between the individual players would improve the transmission of information.

Participants noted that for a seamless discharge, information about the patient should already be available at the time of admission to the hospital:

> *Discharge or discharge planning and a good discharge process starts at admission. [...] The important thing is not to think about discharge on the day of discharge, but already on admission.* [P25]

Improved patient communication was also considered important by interviewees:

> *And that is certainly a wish that I would have that the patients in the hospital are also informed about what they actually have, what has happened and what the next steps are.* [P01]

An optimal discharge should ideally take place in the morning on a working day, and the handing over of medication and required materials should be regulated. This is considered to be the case by nursing homes and outpatient care services as well as by hospital staff:

> *From 9 or 10 a.m onwards, the number of patients in the emergency room increases and drops again from 8 p.m onwards. And during this peak time, there are few beds available in the hospital. Afterwards, however, when we are closed, the hospital finally loses cases and at night we have more free capacity again. And that is a mismatch between demand and capacity which can be improved.* [P20]

Textbox 2 shows the barriers and enabling factors for intersectoral collaboration in the context of optimizing discharge management. In addition to the technical aspects and subjective reasons, there were concerns about data protection and fear that a change in the discharge process would require more time:

> *Time pressure is always an issue, both in the hospital and in outpatient care. We just often don't have the time for some processes that we would all consider useful.* [P01]

XSL•FO
RenderX

Textbox 2. Determinants for intersectoral cooperation.

Barriers

- Privacy
  - Data transmission security
  - Legal uncertainties
- Fear of contact
  - Leaving known structures and processes
  - Lack of electronic data processing experience
- Lack of time
  - Higher time consumption
  - Lack of personnel
- Subjective reasons
  - Limitation of one's own competence
  - Unclear communication processes
  - No perceived benefit
  - Low appreciation for discharge management
  - No priority of discharge management
  - No consequences for noncompliance
- Technical aspects
  - User-unfriendly system
  - Electronic data processing errors
  - Interface problems
  - Outdated technical equipment
- Change processes
  - Lack of education or communication
  - Lack of networking among health care providers

Enablers

- Structures
  - Clear responsibilities, instructions, contact persons, or responsibilities
  - Surveillance
  - Introduction or training of new processes
  - No overload and enough time
  - Regular exchange for networking
- Recognizable advantage
  - Time saving
  - Workload reduction
  - Improved exchange of information
  - Feedback loops
- Introduction of change management
  - Priority in the management

XSL•FO
RenderX

- Communicating the advantages
- Involvement of employees

In contrast, a possible reduction in workload owing to digitalized processes was seen as conducive:

> *Digitalization must not be an end in itself, in my opinion, but it must really mean an advantage for the processes, increase safety, increase communication, but it must not be a question of just because it is digital, that it is better in every case and is then associated with the fact that medical or nursing working time is lost or additionally created.* [P23]

For the changeover to be successful, the communication of the advantages associated with optimized discharge management was emphasized above all as part of change management.

Strategies

On the basis of the surveyed processes, the optimization potential, and the determinants from the evaluated individual interviews as well as the workshop with clinicians and physicians in private practice, 19 strategies for the implementation of optimized discharge management were developed. To rank these strategies, chief physicians of the hospital were invited to rate their relevance and feasibility.

A total of 11 physicians participated in the survey to evaluate the strategies (Table 1). The strategies of always sending the discharge letter to the GP, equipping the hospital's social service with mobile devices (eg, laptops and tablets), generating individual medication plans in the format of the national medication plan, and exclusively using the federal medication plan received the highest ratings. In contrast, the introduction of a chat function used exclusively by physicians for direct exchange between hospital and office-based physicians received the lowest rating.

XSL•FO
RenderX

Table 1. Evaluation of the strategies (N=11).

Strategy	Sum[a], mean (SD)	Relevance[a], mean (SD)	Feasability[a], mean (SD)	Implemented intervention
Always send the discharge letter to the GP[b]	11.1 (1.4)	5.7 (0.7)	5.4 (0.8)	Yes
Providing mobile devices to the hospital's social services department (eg, laptop and tablet)	10.9 (1.5)	5.6 (0.7)	5.3 (0.9)	Yes
Generating individual medication plans in the format of the national medication plan	10.8 (1.3)	5.6 (0.7)	5.2 (0.8)	Yes
Exclusive use of the national medication plan	10.6 (1.8)	5.5 (0.7)	5.1 (1.3)	No[c]
Development of digital checklists for discharge processes to be integrated into clinic software	9.9 (1.6)	5.3 (0.9)	4.6 (1.3)	No[c]
Standardized consent process for discharge management at the time of admission	9.8 (3.1)	5.0 (1.6)	4.8 (1.5)	Yes
Sending or receiving an electronic discharge letter (eg, compatible with the electronic patient record)	9.7 (2.0)	5.2 (1.1)	4.5 (1.1)	No[c]
Always send the discharge letter to the patient as a message	9.6 (1.8)	4.3 (1.3)	5.3 (1.0)	No
Organization of necessary (postdischarge) follow-up treatments (eg, by a medical assistant employed at the hospital)	9.5 (2.6)	5.1 (1.4)	4.4 (1.4)	No
Preventing multiple information collection through automated exchange within the different clinic software	9.2 (1.8)	5.7 (0.9)	3.5 (1.5)	No[c]
Conduct training for all people involved in the process to raise awareness of the relevance of discharge management	9.0 (2.1)	4.8 (1.3)	4.2 (1.2)	No[c]
Upon discharge, information is sent to the GP about which medications on the medication plan have changed, if any	9.0 (1.7)	5.2 (0.6)	3.8 (1.2)	No
Ensuring that patients have had at least 1 complete physical examination before discharge	9.0 (2.6)	5.0 (1.1)	4.0 (1.7)	No
Medication plan is available to the GP on the morning of the day of discharge (eg, sent electronically in advance)	8.9 (2.0)	5.0 (0.9)	3.9 (1.3)	No[c]
Hotline used exclusively by physicians for direct exchange between hospital and GPs	8.8 (2.3)	5.2 (0.9)	3.6 (1.6)	Yes
Reasons for medication changes in the medication plan, if \applicable, are communicated to the GP at discharge	8.5 (2.2)	4.9 (1.0)	3.6 (1.3)	No
Continuous digital maintenance of a potential discharge letter (eg, in preparation for an unplanned discharge)	8.1 (2.7)	4.3 (1.3)	3.8 (1.5)	Yes
Establishment of a quality circle for discharge management	7.9 (2.6)	4.0 (1.5)	3.9 (1.4)	No
Introduction of a chat function used exclusively by physicians for direct exchange between the hospital and GPs	7.0 (2.6)	3.9 (1.5)	3.1 (1.2)	No[c]

[a]Mean over participants (6=very high, 5=rather high, 4=high, 3=rather low, 2=low, and 1=very low).

[b]GP: general practitioner.

[c]Not part of the intervention but planned for the future by the hospital.

On the basis of these ratings, the hospital staff discussed which of these strategies were already being implemented or planned for implementation in the near future. Of the 19 strategies, 6 (31%) were assessed as already implemented, 7 (37%) were assessed as planned, and 6 (31%) were assessed as not feasible to implement in the project. The 7 strategies rated highest in the development of the logic model (planned implementation) have been implemented or will be implemented in the near future as part of SEKMA.

To summarize, by April 2022, at the department of vascular surgery, (1) discharge letters were continuously updated digitally, (2) they were always sent, (3) they were sent electronically to the GP (via the infrastructure of KIM), (4) the hospital social service was equipped with mobile devices, (5) individual medication plans were in the format of the national

XSL•FO
RenderX

medication plan, (6) the discharge management consent process at admission was standardized, and (7) a hotline for direct communication between hospital physicians and primary care physicians was implemented. The information transfer via the discharge letter was oriented by the standard of medical information objects (MIOs) eArztbrief. The development of this standard was initiated in 2022 by the National Association of Statutory Health Insurance Physicians and the German Hospital Association. It defines a standard for the electronic hospital discharge letter within the EHR ensuring the transition of relevant information from inpatient to subsequent care in a structured and secure manner [37]. The MIO eArztbrief was not yet ready during the project; however, the current status of the MIO was incorporated into the letter as much as possible.

Fidelity Analysis

After the implementation of the optimized discharge management into the routine in the department of vascular surgery as well as at the external partners in April 2022, the stakeholders participating in the project were asked in a fidelity analysis in September 2022 to what extent the identified strategies were implemented in practice. The survey showed that many of these strategies were not yet widely applied.

A total of 14 individuals responded to the survey (Table S2 in Multimedia Appendix 2). Of the 14 individuals, 11 (79%) were employed at the hospital and 1 (7%) each at an outpatient nursing service, nursing home, and private practice. Employees from social services and medical assistants did not participate in this survey. Of those surveyed, >30% (4/13) stated that they were satisfied with the implementation of the change in discharge management.

There are differences in the fidelity of use among the strategies implemented (Table S3 in Multimedia Appendix 3). Although sending or receiving an electronic discharge letter was always or sometimes used in their routine by only a quarter of respondents, approximately 85% (11/13) of the respondents indicated that medication plans from the hospital were always in the format of the federal medication plan at discharge.

Routine Data Analysis

Readmissions

In total, 12,407 patients were admitted to the hospital as inpatients during the study period (from April 2019 to August 2019 and from April 2022 to August 2022), corresponding to 14,854 cases treated. The internal medicine department (medical clinic) treated most of the cases (4175/14,854, 28.11%). Cases treated in the interventional group (vascular surgery) accounted for 5.11% (759/14,854) of all inpatient cases. Overall, 8.73% (994/11,386) of the patients were readmitted after 30 days. In terms of treated cases, the readmission rate was 9.07% (1222/13,477). The rates increased to 17.1% (1542/9016) for patients and 18.85% (1975/10,478) for cases when considering a longer time horizon for the readmission (90 days). Readmission rates were generally higher in the intervention group (80/705, 11.3%) at 30 days and 28.8% (161/560 at 90 days) than in the hospital as a whole and the control group. Table S4 in Multimedia Appendix 3 provides the number of admitted patients and cases treated as well as the readmissions after 30, 60, and 90 days for the total hospital cases and the departments involved.

Risk Factors

Risk factors for readmission were identified to take the differences between patients from different departments into account for the evaluation of the project's implementation effect.

Older patients, as well as cases with a length of stay of >6 days, had a significantly higher risk of readmission. Similarly, discharge time influenced the readmission risk: patients discharged during the night (9 PM to 5 AM) had a higher risk of readmission. Similarly, there were significant differences in readmissions between cases with different ICD-10 (*International Statistical Classification of Diseases, Tenth Revision*) chapters of principal and secondary diagnoses (Table S5 in Multimedia Appendix 3).

Intervention Effect

Table 2 shows the implementation effects on the readmission rate after 30, 60, and 90 days (DiD) of the bivariate analysis. In the intervention group, the 30-day readmission rate increased by 2.33 percentage points from 10.4% (45/431) to 12.8% (35/274) after SEKMA was implemented. For the 60- and 90-day readmission rate, the increase was even higher (60 days: 2.25 and 90 days: 3.94). These increases have been smaller in the control group. Therefore, a reduction effect of the intervention on the readmission rate (ie, a negative DiD estimate) cannot be observed. Concentrating the analysis on patients aged ≥65 years revealed similar results (Table S6 in Multimedia Appendix 3). As a robustness check, the preintervention period was extended to include admissions between 2011 and 2019. These results confirm the previous findings.

XSL•FO
RenderX

Table 2. Bivariate analysis of the intervention effect[a].

	Intervention group: department of vascular surgery	Control group: departments of internal medicine and cardiology	DiD[b] estimator: Δ (*P* value)
Readmission 30 days: T_1 until August 3, 2022			
Baseline (T_0)[c], n/N (%)	45/431 (10.4)	354/3206 (11.04)	N/A[d]
Intervention (T_1)[e], n/N (%)	35/274 (12.8)	245/2344 (10.45)	N/A
Difference T_0 and T_1(*P* value)	2.33 (.35)	–0.59 (.48)	2.92 (.27)
Readmission 60 days: T_1 until July 4, 2022			
Baseline (T_0)[c], n/N (%)	97/431 (22.5)	549/3206 (17.12)	N/A
Intervention (T_1)[e], n/N (%)	51/206 (24.8)	302/1761 (17.15)	N/A
Difference T_0 and T_1 (*P* value)	2.25 (.53)	0.03 (.98)	2.23 (.56)
Readmission 90 days: T_1 until June 4, 2022			
Baseline (T_0)[c], n/N (%)	120/431 (27.8)	707/3206 (22.05)	N/A
Intervention (T_1)[e], n/N (%)	41/129 (31.8)	257/1186 (21.67)	N/A
Difference T_0 and T_1 (*P* value)	3.94 (.40)	–0.38 (.78)	4.32 (.37)

[a]Only admissions between April 2019 and August 2019 and between April 2022 and August 2022 were considered.

[b]DiD: difference-in-difference, Δ: Difference between the readmission rates of the intervention and the control group at T_0 and T_1, respectively.

[c]Intervention period (T_1): from April 1, 2022.

[d]N/A: not applicable.

[e]Baseline period (T_0): April 1, 2019, to August 31, 2019.

The results of the multivariate logistic regression model (Table 3) confirm the results of the bivariate analysis that there were higher readmission rates in the intervention group and that there was no significant effect of the optimized discharge management on readmissions in the available data. This result was also confirmed for patients aged >65 years (Table S7 in Multimedia Appendix 3). Furthermore, the insignificance of the effect of the implementation of SEKMA on readmission rates was also confirmed in a pre-post comparison estimated by a multivariate logistic regression based on vascular surgery cases only (Table S8 in Multimedia Appendix 3). Finally, the estimated effects remained very similar if the preintervention period began in 2011 and ended at the end of 2019.

XSL•FO
RenderX

Table 3. Multivariate logistic regression[a].

Variable[b]	Readmissions after		
	30 days: T_1[c] until August 3, 2022	60 days: T_1 until July 4, 2022	90 days: T_1 until June 4, 2022
DiD[d] effect, odds ratio (95% CI)	1.39 (0.84-2.31)	1.16 (0.76-1.77)	1.27 (0.80-2.02)
Intervention period, odds ratio (95% CI)	0.96 (0.81-1.15)	1.03 (0.88-1.21)	0.99 (0.84-1.17)
Intervention group, odds ratio (95% CI)	0.90 (0.63-1.29)	1.32 (1.01-1.73)	1.29 (1.00-1.66)
Social demographics , odds ratio (95% CI)			
Female	0.93 (0.79-1.09)	0.91 (0.79-1.05)	0.90 (0.78-1.03)
Aged >65 years	1.44 (1.17-1.76)	1.41 (1.19-1.67)	1.48 (1.25-1.74)
Length of stay (reference: <3 days) , odds ratio (95% CI)			
3-5	1.51 (1.16-1.95)	1.45 (1.17-1.80)	1.45 (1.18-1.78)
6-9	2.12 (1.63-2.76)	2.14 (1.72-2.68)	2.04 (1.65-2.53)
>10	2.49 (1.88-3.29)	2.23 (1.76-2.82)	2.25 (1.79-2.83)
Transfer within the hospital, odds ratio (95% CI)			
Transfer between departments	0.96 (0.1-1.18)	1.02 (0.85-1.21)	1.06 (0.89-1.25)
Intensive care	0.50 (0.31-0.82)	0.69 (0.47-1.00)	0.75 (0.53-1.08)
Discharge time (reference: 6 AM to 12 AM), odds ratio (95% CI)			
1 PM to 5 PM	1.17 (0.99-1.39)	1.18 (1.02-1.37)	1.12 (0.97-1.30)
6 PM to 8 PM	1.13 (0.84-1.51)	1.24 (0.96-1.6)	1.23 (0.96-1.57)
9 PM to 5 AM	0.44 (0.22-0.88)	0.45 (0.25-0.81)	0.43 (0.24-0.77)
Observations, N	6224	5578	4945

[a]Only admissions between April 2019 and August 2019 and between April 2022 and August 2022 were considered.

[b]In addition to the variables listed here, the International Statistical Classification of Diseases, Tenth Revision chapters of the principal and secondary diagnoses were also included as control variables.

[c]Intervention period (T_1): from April 1, 2022.

[d]DiD: difference-in-difference.

Discussion

This study aims to explore the barriers and enablers of a digitalized discharge management system, to implement such a system using a logic model developed from these determinants, and to evaluate its impact on the readmission rate.

Determinants and Implementation Strategies

The importance of the transmission of information for improved discharge management is also highlighted in the high rating of the strategies regarding the discharge letter, that is, developing an electronic discharge letter, continuously entering information into the letter, and always sending it to the GP. The discharge letter is the standard communication tool between inpatient and ambulatory care and found to be a source for deficits in information transfer [38]. In particular, delay and incompleteness of medication-related information endanger patients' safety [39,40], leading to an increased risk of hospital readmission [41]. As shown for a sample of 20 Dutch hospitals, discharge letters vary in quality depending on patient and admission characteristics [42]. A standardized discharge letter can reduce transcription time and improve medical communication between physicians [43]. In addition, GPs prefer that discharge letters be written in a clear, concise, and understandable manner [44]. An electronic discharge letter generated from a computer-based document not only avoids transcription errors and lacks standardization but also ensures timely delivery [45]. In Germany, the discharge letter played a central role in approaches to creating a standard for intersectoral information exchange. For example, the VHitG (derived from the German "Verband der Hersteller von IT-Lösungen im Gesundheitswesen") initiative "Intersectoral Communication" developed an implementation to facilitate the exchange of discharge letters between sectors, which is integrated into the existing IT system [46]. Another example is the recent approach by the National Association of Statutory Health Insurance Physicians and the German Hospital Association to create a standard for the electronic hospital discharge letter within the EHR [37].

To improve the standardization of the transmitted medication information, the use of the format of the nationwide medication plan was considered an important strategy in this study. In Germany, several projects have shown that physicians, pharmacists, and patients realize the benefits and accept the nationwide medication plan [47-49]. It can serve for the health care providers as a promising tool to improve the interdisciplinary and multiprofessional collaboration, especially

XSL•FO
RenderX

as a digital solution that can realize its full potential [50]. Similar results have been reported in other countries [51,52]. In this study, participants suggested transferring medication-related information electronically and always in the format of the national medication plan. In the participating hospital, this strategy has been implemented during the project. For older patients in particular, shared medication records have the potential to reduce hospital readmissions [51].

Concerns about technical and temporal integrability were identified as an important barrier to the implementation of optimized discharge management. This includes an expected higher time consumption for the introduction of digitalized processes, a general fear of contact (owing to leaving known structures and a lack of electronic data processing experience), and further technical aspects (as a user-unfriendly system, electronic data processing errors, and interface problems). Similar barriers were identified in related eHealth projects [53-57]. Although the digitalization of processes was expected, in general, to be associated with time advantages, many of those involved associate the introduction with additional work effort. To overcome these concerns, successful implementation requires streamlining, simplifying, and redesigning the existing health care practices as a first step [58]. The strategy of introducing a physician-only hotline and a chat function for direct communication between the hospital and GPs could be seen as a simplification of communication instead of relying solely on the legally required discharge letter.

Effect on Readmissions

A possible explanation for the low level of fidelity as well as the insignificant effect of SEKMA on readmissions could be the relatively short application period of half a year (from April 2022 to September 2022). Complex implementations such as those elaborated in SEKMA may require a longer time before they are applied in daily routines. Another reason for the insignificant effect on readmissions could be the rather good baseline level of the outcome in national comparison. Although other studies in Germany showed readmission rates, for example, of 18.1% (30 days) to 35.4% (90 days) for older patients (aged >65 years) [22], these rates were substantially lower for the patients in this study, that is, 11.8% (30 days) to 23.6% (90 days).

Limitations

Our study had several limitations. First, the restrictions that existed owing to the COVID-19 pandemic might have affected the effectiveness of the implementation. All stakeholders involved in SEKMA faced a high workload owing to the pandemic as well as the requirements and measures resulting from the pandemic. However, the study results show that even under the special circumstances of the pandemic, it was possible to develop and implement an intersectoral optimization of discharge management. The infrastructure for the intersectoral care of patients created by the project has great potential to increase the quality of care, even if this could not yet be demonstrated with regard to readmissions. Future research should analyze the routine hospital data over the next 5 years.

Although the study included all relevant health care providers and considered the entire patient care pathway, the number of respondents from some professions may be rather small. For example, only 3 GPs were interviewed. However, the theoretical saturation effect in the statements made during the interviews suggests that this number is sufficient to identify the optimization potential as well as determinants.

Conclusions

Creating a digital ecosystem that connects different health care providers seems to be a promising approach to ensure secure and fast networking of the sectors and to promote rapid information exchange between the sectors. The described intersectoral optimization of discharge management provides a structured template for the implementation of a similar local digital care networking infrastructure in other care regions in Germany and other countries with a similarly fragmented health care system.

Acknowledgments

This study was financially supported by the Ministry of Justice and Health (Ministerium für Justiz und Gesundheit), Schleswig-Holstein. This study was conducted independently.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to the votum of the Ethics Committee of the University of Lübeck.

Authors' Contributions

CS contributed to conceptualization, formal analysis, investigation, methodology, and validation; prepared the original draft; and reviewed and edited the manuscript. LP participated in methodology, conducted and analyzed the interviews, and reviewed and edited the draft. LW participated in conceptualizing the digital discharge system and reviewed and edited the draft. RS was involved in conceptualization and reviewing and editing the draft. JS contributed to conceptualization, investigation, methodology, and validation and reviewed and edited the draft. All authors have read and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

LW was a hospital manager with a focus on digitalization at the hospital under study during the time of the project, Sektorübergreifende Optimierung des Entlassmanagements (SEKMA). LW is the founder of the company Lacanja GmbH Health Innovation Port, Hamburg, Germany, and is a member of several committees, including the expert group of the Gematik IOP (Interop) Council. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 26 KB - jmir_v26i1e47133_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File , 22 KB - jmir_v26i1e47133_app2.docx]

Multimedia Appendix 3
Supplemental tables with results of supplemental analyses.
[DOCX File , 50 KB - jmir_v26i1e47133_app3.docx]

References

1. Imison C, Castle-Clarke S, Watson R, Edwards N. Delivering the benefits of digital health care. Nuffield Trust. 2016. URL: https://www.nuffieldtrust.org.uk/sites/default/files/2017-01/delivering-the-benefits-of-digital-technology-web-final.pdf [accessed 2024-03-11]
2. Bonomi S. The electronic health record: a comparison of some European countries. In: Ricciardi F, Harfouche A, editors. Information and Communication Technologies in Organizations and Society. Cham, Switzerland: Springer; 2016.
3. Jorzig A. Digitalisierung von krankenhäusern und krankenhauszukunftsgesetz. Gynäkologie 2022 Oct 11;55:880-884. [doi: 10.1007/s00129-022-05019-z]
4. Burmann A, Fischer B, Brinkkötter N, Meister S. Managing directors' perspectives on digital maturity in German hospitals-a multi-point online-based survey study. Int J Environ Res Public Health 2022 Aug 06;19(15):9709 [FREE Full text] [doi: 10.3390/ijerph19159709] [Medline: 35955066]
5. Graf A, Fehring L, Henningsen M, Zinner M. Going digital in Germany: an exploration of physicians' attitudes towards the introduction of electronic prescriptions - a mixed methods approach. Int J Med Inform 2023 Jun;174:105063. [doi: 10.1016/j.ijmedinf.2023.105063] [Medline: 37028259]
6. Nohl-Deryk P, Brinkmann JK, Gerlach FM, Schreyögg J, Achelrod D. [barriers to digitalisation of healthcare in Germany: a survey of experts]. Gesundheitswesen 2018 Nov;80(11):939-945. [doi: 10.1055/s-0043-121010] [Medline: 29301149]
7. Caumanns J. [For discussion: the state of digitization of the German healthcare system]. Z Evid Fortbild Qual Gesundhwes 2019 Jun;143:22-29. [doi: 10.1016/j.zefq.2019.04.002] [Medline: 31204267]
8. Lupiáñez-Villanueva F, Devaux A, Valverde-Albacete J, Folkvord F, Fauli C, Altenhofer M, et al. Benchmarking deployment of eHealth among general practitioners. Publications Office of the European Union. 2018. URL: https://www.rand.org/pubs/external_publications/EP67863.html [accessed 2024-03-11]
9. Melby L, Brattheim BJ, Hellesø R. Patients in transition--improving hospital-home care collaboration through electronic messaging: providers' perspectives. J Clin Nurs 2015 Dec;24(23-24):3389-3399. [doi: 10.1111/jocn.12991] [Medline: 26374139]
10. Blümel M, Spranger A, Achstetter K, Maresso A, Busse R. Germany: health system review. Health Syst Transit 2020 Dec;22(6):1-272 [FREE Full text] [Medline: 34232120]
11. Büyükdurmus T, Kopetsch T, Schmitz H, Tauchmann H. On the interdependence of ambulatory and hospital care in the German health system. Health Econ Rev 2017 Dec;7(1):2 [FREE Full text] [doi: 10.1186/s13561-016-0132-4] [Medline: 28097610]
12. Koné I, Maria Zimmermann B, Nordström K, Simone Elger B, Wangmo T. A scoping review of empirical evidence on the impacts of the DRG introduction in Germany and Switzerland. Int J Health Plann Manage 2019 Jan;34(1):56-70. [doi: 10.1002/hpm.2669] [Medline: 30426573]
13. von Eiff W, Schüring S, Greitemann B, Karoff M. [REDIA--impacts of DRG introduction in the acute sector on medical rehabilitation]. Rehabilitation (Stuttg) 2011 Aug;50(4):214-221. [doi: 10.1055/s-0031-1275720] [Medline: 21800267]
14. Braun B. Auswirkungen der DRGs auf Versorgungsqualität und Arbeitsbedingungen im Krankenhaus. In: Manzei A, Schmiede R, editors. 20 Jahre Wettbewerb im Gesundheitswesen. Wiesbaden, Germany: Springer VS; 2014.
15. Sager MA, Easterling DV, Kindig DA, Anderson OW. Changes in the location of death after passage of medicare's prospective payment system. A national study. N Engl J Med 1989 Feb 16;320(7):433-439. [doi: 10.1056/NEJM198902163200705] [Medline: 2643774]

XSL•FO
RenderX

16. Rogers WH, Draper D, Kahn KL. Quality of care before and after implementation of the DRG-based prospective payment system: a summary of effects. JAMA 1990 Oct 17;264(15):1989-1994. [doi: 10.1001/jama.1990.03450150089037]
17. Kosecoff J, Kahn KL, Rogers WH. Prospective payment system and impairment at discharge: the `quicker-and-sicker' story revisited. JAMA 1990 Oct 17;264(15):1980-1983. [doi: 10.1001/jama.1990.03450150080035]
18. Freund T, Mahler C, Erler A, Gensichen J, Ose D, Szecsenyi J, et al. Identification of patients likely to benefit from care management programs. Am J Manag Care 2011 May;17(5):345-352 [FREE Full text] [Medline: 21718082]
19. Shulan M, Gao K, Moore CD. Predicting 30-day all-cause hospital readmissions. Health Care Manag Sci 2013 Jun;16(2):167-175. [doi: 10.1007/s10729-013-9220-8] [Medline: 23355120]
20. Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW. Adverse drug events occurring following hospital discharge. J Gen Intern Med 2005 Apr;20(4):317-323 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.30390.x] [Medline: 15857487]
21. Solet DJ, Norvell JM, Rutan GH, Frankel RM. Lost in translation: challenges and opportunities in physician-to-physician communication during patient handoffs. Acad Med 2005 Dec;80(12):1094-1099. [doi: 10.1097/00001888-200512000-00005] [Medline: 16306279]
22. Ruff C, Gerharz A, Groll A, Stoll F, Wirbka L, Haefeli WE, et al. Disease-dependent variations in the timing and causes of readmissions in Germany: a claims data analysis for six different conditions. PLoS One 2021 Apr 26;16(4):e0250298 [FREE Full text] [doi: 10.1371/journal.pone.0250298] [Medline: 33901203]
23. Zabawa C, Cottenet J, Zeller M, Mercier G, Rodwin VG, Cottin Y, et al. Thirty-day rehospitalizations among elderly patients with acute myocardial infarction: impact of postdischarge ambulatory care. Medicine (Baltimore) 2018 Jun;97(24):e11085 [FREE Full text] [doi: 10.1097/MD.0000000000011085] [Medline: 29901621]
24. Gonçalves-Bradley DC, Lannin NA, Clemson L, Cameron ID, Shepperd S. Discharge planning from hospital. Cochrane Database Syst Rev 2022 Feb 24;2(2):CD000313 [FREE Full text] [doi: 10.1002/14651858.CD000313.pub6] [Medline: 35199849]
25. Kripalani S, Chen G, Ciampa P, Theobald C, Cao A, McBride M, et al. A transition care coordinator model reduces hospital readmissions and costs. Contemp Clin Trials 2019 Jun;81:55-61 [FREE Full text] [doi: 10.1016/j.cct.2019.04.014] [Medline: 31029692]
26. Schönemann-Gieck P, Evers A. Klinikentlassungen älterer Patienten mit sozialem Interventionsbedarf: Möglichkeiten und Grenzen Kommunalen Handelns am Beispiel des Wiesbadener Gesundheitsnetzes "GeReNet.Wi". Pflege & Gesellschaft. 2018. URL: https://tinyurl.com/yc56yfr2 [accessed 2024-03-11]
27. Held LA, Wewetzer L, Steinhäuser J. Determinants of the implementation of an artificial intelligence-supported device for the screening of diabetic retinopathy in primary care - a qualitative study. Health Informatics J 2022;28(3):14604582221112816 [FREE Full text] [doi: 10.1177/14604582221112816] [Medline: 35921547]
28. Dünnebeil S, Sunyaev A, Blohm I, Leimeister JM, Krcmar H. Determinants of physicians' technology acceptance for e-health in ambulatory care. Int J Med Inform 2012 Nov;81(11):746-760. [doi: 10.1016/j.ijmedinf.2012.02.002] [Medline: 22397989]
29. Hennemann S, Beutel ME, Zwerenz R. Drivers and barriers to acceptance of web-based aftercare of patients in inpatient routine care: a cross-sectional survey. J Med Internet Res 2016 Dec 23;18(12):e337 [FREE Full text] [doi: 10.2196/jmir.6003] [Medline: 28011445]
30. Poss-Doering R, Kunz A, Pohlmann S, Hofmann H, Kiel M, Winkler EC, et al. Utilizing a prototype patient-controlled electronic health record in Germany: qualitative analysis of user-reported perceptions and perspectives. JMIR Form Res 2018 Aug 03;2(2):e10411 [FREE Full text] [doi: 10.2196/10411] [Medline: 30684411]
31. Jäger C, Szecsenyi J, Freund T, Reichel JK, Kuhlmey C, Wensing M, et al. [Developing a tailored intervention: implementing recommendations for polypharmacy in multimorbid patients (PomP)]. Z Evid Fortbild Qual Gesundhwes 2014;108(5-6):270-277. [doi: 10.1016/j.zefq.2013.08.014] [Medline: 25066345]
32. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implement Sci 2007 Nov 30;2:40 [FREE Full text] [doi: 10.1186/1748-5908-2-40] [Medline: 18053122]
33. Sektorübergreifende optimierung des entlassmanagements. SEKMA. URL: https://www.sekma.de/ [accessed 2023-02-07]
34. KIM - Der standard für sicheren E-Mail und datenaustausch im gesundheitswesen. Gematik. URL: https://fachportal.gematik.de/fileadmin/Fachportal/Veranstaltungen/gematik_KIM_DMEA_2021.pdf [accessed 2023-11-01]
35. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
36. Mayring P. Qualitative inhaltsanalyse. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in der Psychologie. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften; 2010.
37. Informationsobjekt "Krankenhaus-Entlassbrief": letter of intent bekräftigt zusammenarbeit zwischen KBV und DKG. Deutsche Krankenhausgesellschaft. 2021 Apr 15. URL: https://www.dkgev.de/dkg/presse/details/gemeinsame-pressemitteilung/ [accessed 2023-10-30]
38. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007 Feb 28;297(8):831-841. [doi: 10.1001/jama.297.8.831] [Medline: 17327525]

XSL•FO
RenderX

39. Freyer J, Kasprick L, Sultzer R, Schiek S, Bertsche T. A dual intervention in geriatric patients to prevent drug-related problems and improve discharge management. Int J Clin Pharm 2018 Oct;40(5):1189-1198. [doi: 10.1007/s11096-018-0643-7] [Medline: 30051223]
40. Schwarz CM, Hoffmann M, Schwarz P, Kamolz LP, Brunner G, Sendlhofer G. A systematic literature review and narrative synthesis on the risks of medical discharge letters for patients' safety. BMC Health Serv Res 2019 Mar 12;19(1):158 [FREE Full text] [doi: 10.1186/s12913-019-3989-1] [Medline: 30866908]
41. Pesko MF, Gerber LM, Peng TR, Press MJ. Home health care: nurse-physician communication, patient severity, and hospital readmission. Health Serv Res 2018 Apr;53(2):1008-1024 [FREE Full text] [doi: 10.1111/1475-6773.12667] [Medline: 28217974]
42. Langelaan M, Baines RJ, de Bruijne MC, Wagner C. Association of admission and patient characteristics with quality of discharge letters: posthoc analysis of a retrospective study. BMC Health Serv Res 2017 Mar 21;17(1):225 [FREE Full text] [doi: 10.1186/s12913-017-2149-8] [Medline: 28327139]
43. Doyle MA, Malcolm JC, Liu D, Maranger J, Ooi TC, Keely E. Using a structured discharge letter template to improve communication during the transition from a specialized outpatient diabetes clinic to a primary care physician. Can J Diabetes 2015 Dec;39(6):457-466. [doi: 10.1016/j.jcjd.2015.06.009] [Medline: 26454683]
44. Weetman K, Spencer R, Dale J, Scott E, Schnurr S. What makes a "successful" or "unsuccessful" discharge letter? Hospital clinician and general practitioner assessments of the quality of discharge letters. BMC Health Serv Res 2021 Apr 15;21(1):349 [FREE Full text] [doi: 10.1186/s12913-021-06345-z] [Medline: 33858383]
45. Schiele F, Lemesle G, Angoulvant D, Krempf M, Kownator S, Cheggour S, et al. Proposal for a standardized discharge letter after hospital stay for acute myocardial infarction. Eur Heart J Acute Cardiovasc Care 2020 Oct;9(7):788-801. [doi: 10.1177/2048872619844444] [Medline: 30990337]
46. Klar R, Pelikan E. Telemedicine in Germany. In: Kramme R, Hoffmann KP, Pozos RS, editors. Springer Handbook of Medical Technology. Berlin, Heidelberg: Springer; 2011.
47. Mueller MA, Opitz R, Grandt D, Lehr T. The federal standard medication plan in practice: an observational cross-sectional study on prevalence and quality. Res Social Adm Pharm 2020 Oct;16(10):1370-1378. [doi: 10.1016/j.sapharm.2020.01.013] [Medline: 32061549]
48. Ulmer I, Mildner C, Krämer I. PS-029 Feasibility of utilisation and patient satisfaction with a nationwide standardised electronic medication plan. Eur J Hospital Pharmacy 2017;24:A239-A240. [doi: 10.1136/ejhpharm-2017-000640.535]
49. Botermann L, Krueger K, Eickhoff C, Kloft C, Schulz M. Patients' handling of a standardized medication plan: a pilot study and method development. Patient Prefer Adherence 2016 Apr 22;10:621-630. [doi: 10.2147/PPA.S96431] [Medline: 27175062]
50. Dormann H, Maas R, Eickhoff C, Müller U, Schulz M, Brell D, et al. [Standardized national medication plan : the pilot projects MetropolMediplan 2016, model region Erfurt, and PRIMA]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Sep 25;61(9):1093-1102. [doi: 10.1007/s00103-018-2789-9] [Medline: 30046864]
51. Payen A, Godard-Sebillotte C, Sourial N, Soula J, Verloop D, Defebvre MM, et al. The impact of including a medication review in an integrated care pathway: a pilot study. Br J Clin Pharmacol 2023 Mar;89(3):1036-1045. [doi: 10.1111/bcp.15543] [Medline: 36164674]
52. Bugnon B, Geissbuhler A, Bischoff T, Bonnabry P, von Plessen C. Improving primary care medication processes by using shared electronic medication plans in Switzerland: lessons learned from a participatory action research study. JMIR Form Res 2021 Jan 07;5(1):e22319. [doi: 10.2196/22319] [Medline: 33410753]
53. Eden KB, Totten AM, Kassakian SZ, Gorman PN, McDonagh MS, Devine B, et al. Barriers and facilitators to exchanging health information: a systematic review. Int J Med Inform 2016 Apr;88:44-51 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.01.004] [Medline: 26878761]
54. Pfeuffer N, Beyer A, Penndorf P, Leiz M, Radicke F, Hoffmann W, et al. Evaluation of a health information exchange system for geriatric health care in rural areas: development and technical acceptance study. JMIR Hum Factors 2022 Sep 15;9(3):e34568 [FREE Full text] [doi: 10.2196/34568] [Medline: 36107474]
55. Anderson JG. Social, ethical and legal barriers to e-health. Int J Med Inform 2007;76(5-6):480-483. [doi: 10.1016/j.ijmedinf.2006.09.016] [Medline: 17064955]
56. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]
57. Jimma BL, Enyew DB. Barriers to the acceptance of electronic medical records from the perspective of physicians and nurses:a scoping review. Inform Med Unlocked 2022;31:100991. [doi: 10.1016/j.imu.2022.100991]
58. Pohlmann S, Kunz A, Ose D, Winkler EC, Brandner A, Poss-Doering R, et al. Digitalizing health services by implementing a personal electronic health record in Germany: qualitative analysis of fundamental prerequisites from the perspective of selected experts. J Med Internet Res 2020 Jan 29;22(1):e15102 [FREE Full text] [doi: 10.2196/15102] [Medline: 32012060]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DiD: difference-in-difference
EHR: electronic health record
GP: general practitioner
ICD-10: International Statistical Classification of Diseases, Tenth Revision
KIM: Kommunikation im Medizinwesen
MIO: medical information object
SEKMA: Sektorübergreifende Optimierung des Entlassmanagements

Edited by T Leung; submitted 09.03.23; peer-reviewed by P Nohl-Deryk, S Meister; comments to author 21.04.23; revised version received 13.06.23; accepted 31.01.24; published 26.03.24.

Please cite as:
Strumann C, Pfau L, Wahle L, Schreiber R, Steinhäuser J
Designing and Implementation of a Digitalized Intersectoral Discharge Management System and Its Effect on Readmissions: Mixed Methods Approach
J Med Internet Res 2024;26:e47133
URL: https://www.jmir.org/2024/1/e47133
doi:10.2196/47133
PMID:38530343

©Christoph Strumann, Lisa Pfau, Laila Wahle, Raphael Schreiber, Jost Steinhäuser. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sociotechnical Cross-Country Analysis of Contextual Factors That Impact Patients' Access to Electronic Health Records in 4 European Countries: Framework Evaluation Study

Jonas Moll[1], PhD; Isabella Scandurra[1†], PhD; Annika Bärkås[2,3], PhD; Charlotte Blease[2,4], PhD; Maria Hägglund[2,3], PhD; Iiris Hörhammer[5], DSC; Bridget Kane[2,6], PhD; Eli Kristiansen[7], ME; Peeter Ross[8,9], MD, PhD; Rose-Mharie Åhlfeldt[10], PhD; Gunnar O Klein[1], MD, PhD

[1]Centre for Empirical Research on Information systems, School of Business, Örebro University, Örebro, Sweden
[2]Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
[3]Medtech Science & Innovation Centre, Uppsala University Hospital, Uppsala, Sweden
[4]Digital Psychiatry, Department of Psychiatry Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States
[5]Department of Computer Science, Aalto University, Espoo, Finland
[6]Karlstad University Business School, Karlstad, Sweden
[7]Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[8]E-Medicine Centre, Department of Health Technologies, Tallinn University of Technology, Tallinn, Estonia
[9]Research Department, East Tallinn Central Hospital, Tallinn, Estonia
[10]School of Informatics, University of Skövde, Skövde, Sweden
[†]deceased

Corresponding Author:
Jonas Moll, PhD
Centre for Empirical Research on Information systems
School of Business
Örebro University
Nova building, 4th floor
Fakultetsgatan 1
Örebro, 70182
Sweden
Phone: 46 0703174383
Email: jonas.moll@oru.se

Abstract

Background: The NORDeHEALTH project studies patient-accessible electronic health records (PAEHRs) in Estonia, Finland, Norway, and Sweden. Such country comparisons require an analysis of the sociotechnical context of these services. Although sociotechnical analyses of PAEHR services have been carried out in the past, a framework specifically tailored to in-depth cross-country analysis has not been developed.

Objective: This study aims to develop and evaluate a method for a sociotechnical analysis of PAEHRs that advances a framework for sociotechnical analysis of eHealth solutions first presented by Sittig and Singh. This first article in a series presents the development of the method and a cross-country comparison of the contextual factors that enable PAEHR access and use.

Methods: The dimensions of the framework for sociotechnical analysis were thoroughly discussed and extended in a series of workshops with international stakeholders, all being eHealth researchers focusing on PAEHRs. All countries were represented in the working group to make sure that important national perspectives were covered. A spreadsheet with relevant questions related to the studied services and the various dimensions of the sociotechnical framework was constructed and distributed to the 4 participating countries, and the project participants researched various national sources to provide the relevant data for the comparisons in the 10 sociotechnical dimensions.

Results: In total, 3 dimensions were added to the methodology of Sittig and Singh to separate clinical content from features and functions of PAEHRs and demonstrate basic characteristics of the different countries regarding national and regional steering of health care and information and communications technology developments. The final framework contained the following

XSL•FO
RenderX

dimensions: metadata; hardware and software computing infrastructure; features and functions; clinical content shared with patients; human-computer interface; people; workflow and communication; the health care organization's internal policies, procedures, and culture; national rules, regulations, and incentives; system measurement and monitoring; and health care system context. The dimensions added during the study mostly concerned background information needed for cross-country comparisons in particular. Several similarities were identified among the compared countries, especially regarding hardware and software computing infrastructure. All countries had, for example, one national access point, and patients are provided a PAEHR automatically. Most of the differences could be identified in the *health care system context* dimension. One important difference concerned the governing of information and communications technology development, where different levels (state, region, and municipality) were responsible in different countries.

Conclusions: This is the first large-scale international sociotechnical analysis of services for patients to access their electronic health records; this study compared services in Estonia, Finland, Norway, and Sweden. A methodology for such an analysis was developed and is presented to enable comparison studies in other national contexts to enable future implementations and evaluations of PAEHRs.

(J Med Internet Res 2024;26:e55752) doi:10.2196/55752

KEYWORDS

electronic health record; EHR; health data; national survey; web-based medical record; web-based record access; patient access; patient-accessible electronic health record; patient portal; sociotechnical analysis

Introduction

Background

Patients' web-based access to their electronic health records (EHRs) is increasingly being implemented internationally [1-3], and a growing body of literature indicates strong benefits to patients, including understanding their care plans better [4], feeling more in control of their care [4,5], being better informed about their medication [6], improved communication with and trust in their clinicians [5,7], and improved patient safety [8].

Patient-accessible EHRs (PAEHRs) are web-based services providing patients with secure access to view and sometimes edit or comment on their EHRs made available by their health care providers [9]. A PAEHR is directly linked to the EHRs, which are shared patient records entered and maintained by health care service providers and contain historical data about a patient [10]. Medical and health data and information in the EHR are created and managed by authorized providers in a digital format capable of being shared with other providers across more than one health care organization [11]. In addition, a PAEHR may include access to the services supporting a person's access to health care services (eg, e-booking and e-consultation) and to evidence-based tools (eg, patient guidelines, educational materials, and reimbursement information). It is important to distinguish between PAEHRs and personal health records—the latter being a health record that the patients themselves control and maintain to track their own health, for example, on paper; in Microsoft Excel sheets; or, predominantly, in health care apps. Personal health records are external applications that can potentially be linked with EHRs to enable sharing [10].

In this study, we focused on national services in the 4 countries of Estonia, Finland, Norway, and Sweden and excluded web-based services designed to provide access to the services of only 1 health care provider. Among the studied countries, Estonia was the first to offer citizens web-based access to their EHR through the service Digilugu that was launched nationwide in 2008 [12]. The Finnish counterpart, Omakanta, was launched nationwide in 2010 with only limited functionality and reached full functionality in 2015 after a step-by-step adoption of functions [13]. In Sweden, the national PAEHR service Journalen was launched in the Region Uppsala in 2012 and has been accessible to all citizens since 2018 [14]. In Norway, the national PAEHR service Helsenorge was first launched in 1 of 4 regions in 2015, and as of 2023, citizens in 3 out of 4 regions can use the service to access their PAEHRs [15].

Despite the reported benefits of PAEHRs and patients' web-based record access, implementation is often slow and challenging, and the complexity of health care systems and technical infrastructure leads to great diversity in, for example, the information to which patients are given access and when they can access it across regions and health care settings. Similarly, patients' adoption and use of PAEHRs also varies across contexts. To understand why these differences exist and better adapt the design and implementation of PAEHRs to a specific context, there is a need for a more fine-grained understanding of the social and technical underpinnings of this innovation.

Sociotechnical Systems

As described by Baxter and Sommerville [16], the problems that arise when designing and implementing complex IT systems are not just technical, engineering problems. These systems are developed and operated by people working in organizations that inevitably have different, often conflicting goals and views on the role and design of the system. The IT system is part of a broader "sociotechnical" system, and to understand success factors and barriers to implementing web-based record access and identify best practices and guidelines, there is a need to approach these eHealth services as complex sociotechnical systems.

Sittig and Singh [17] have proposed a multidimensional sociotechnical framework in which any health IT innovation, intervention, application, or device implemented within a complex adaptive health care system can be studied. The

XSL•FO
RenderX

sociotechnical framework by Sittig and Singh [17] identifies eight dimensions of sociotechnical systems in health care that need to be considered in both development and evaluation: (1) hardware and software computing infrastructure; (2) clinical content; (3) human-computer interface; (4) people; (5) workflow and communication; (6) internal organizational policies, procedures, and culture; (7) external rules, regulations, and pressures; and (8) system measurement and monitoring [17]. The framework breaks down components of the technology to enable researchers to identify specific problems with implementation. It also includes monitoring processes and government structures that need to be in place for the system to achieve its goals. The interrelatedness of the components makes the framework pertinent when eHealth technologies and users are at the core of the investigation.

In 2017, Hägglund and Scandurra [18] began analyzing the Swedish PAEHR system from a sociotechnical perspective using the framework by Sittig and Singh [17], and their results laid the foundation for continued work within the NORDeHEALTH research project [19] involving partners from Estonia, Finland, Norway, and Sweden. These countries were deemed appropriate for the development and first test of the cross-country sociotechnical analysis method derived in this study as they have comparable government and health care structures. In addition, they have all reached maturity in the use of their respective PAEHR systems—these countries are all among the early adopters of this specific type of eHealth service for patients. In a 2021 workshop, a first draft of an extended version of the framework by Sittig and Singh [17] as a sociotechnical template for analyzing PAEHR implementations was introduced and discussed [20]. However, the original framework by Sittig and Singh [17] needed to be more specifically adapted to PAEHRs and their contexts to enable useful cross-country comparisons, a process that will be described further in this paper.

The full results of the sociotechnical analysis of the 4 countries involved in the NORDeHEALTH project are being published in an article series; it is intended for each publication to offer results related to specific and related framework dimensions and, thereby, focus on different themes. This paper is intended to offer a detailed description of the overarching research method as well as a sociotechnical analysis of dimensions providing the context for other articles of the series.

Aims and Research Questions

The aims of this study were to (1) develop a detailed collaborative method suitable for cross-country sociotechnical analyses of PAEHRs and (2) use the data collection to compare contextual factors that enable PAEHR access and use.

Although the proposed work process, including the production of an extensive sociotechnical analysis template, is described in the context of comparing PAEHRs in 4 specific countries, it can be adjusted to cover other eHealth services for patients across other nations. The comparison and discussion regarding the national contexts will provide a robust foundation and understanding of other parts of the sociotechnical analysis, such as included features and national and regional incentives for the use and promotion of PAEHRs. The following research questions (RQs) guided the work presented in this paper:

- RQ 1: how can a sociotechnical framework for health IT be adapted and a collaborative method be developed that is suitable for cross-country sociotechnical analyses of PAEHR services?
- RQ 2: how can the method be used to compare the PAEHRs in Estonia, Finland, Norway, and Sweden?

Methods

Research Method

This study has 2 main parts inspired by different phases in the design science methodology [21]. In the first part, a collaborative method, including a data collection template, for sociotechnical cross-country analysis of PAEHRs was derived: (1) problem identification and motivation; that is, based on earlier experiences with the framework by Sittig and Singh [17] and the fact that the framework is not adapted for comparisons across contexts, the following problem was defined, how can we conduct fruitful comparative analysis among PAEHRs in different regions and health care settings? (2) objectives of a solution; that is, due to the identified need to perform a complete sociotechnical comparison among health care contexts, the main objective of the new artifact (being a method for collecting data, including a refined framework and a data collection instrument) is to enable a comparison that would consider the sociotechnical contexts of PAEHRs; and (3) design and development; that is, a dedicated collaborative method was developed with a focus on adjusting the framework for cross-country comparisons and deriving a data collection template enabling the collection of all relevant data from the specific countries involved.

The second part of our study (in this paper referred to as *data collection*) consists of the collection and comparison of the sociotechnical characteristics of the 4 countries participating in the NORDeHEALTH project: Estonia, Finland, Norway, and Sweden. This can be seen as case studies of the more general questions regarding the availability and use of the PAEHR in a country. However, it is also an evaluation of the method as an artifact in the design science paradigm: (1) demonstration—the data collection template derived from the first part of the study was used to conduct a cross-country comparison among 4 Nordic countries—and (2) evaluation—finally, the results from the comparison were analyzed to evaluate whether the approach provides an understanding that explains differences among countries in the adoption of PAEHRs by regions and providers and patients.

In the remaining sections, the different points of the process are elaborated on, with emphasis on the design and development, demonstration, and evaluation steps. The work was led by a core analysis team (JM, IS, GK, and AB) of health informatics researchers.

Development of the Collaborative Method

The development of the data collection instrument based on the sociotechnical framework proposed by Sittig and Singh [17] started before the NORDeHEALTH project had begun.

XSL•FO
RenderX

In 2017, a first sociotechnical analysis of the Swedish PAEHR service Journalen was conducted with the aim of increasing the understanding of factors that influence the design, implementation, adoption, and use of the service [18]. However, the analysis did not go into the details of each dimension of the framework but, rather, highlighted some overall challenges. The results of this early analysis were used as input to an international workshop held in 2021, for which a more detailed template specifically adapted to PAEHRs was developed (Textbox 1). In addition to including specific PAEHR-related questions to each dimension, a new dimension, *Features and functions*, was added. Some other dimensions were renamed. During the workshop, international experts provided feedback on the template that was incorporated into further refinements, leading to the data collection form presented in this study.

Textbox 1. The sociotechnical framework dimensions included in an early version of the data collection template following the work by Sittig and Singh.

Dimension and description

- Hardware and software computing infrastructure: focuses only on the hardware and software required to run the applications.
- Features and functions: important features and functions in the patient-accessible electronic health record (PAEHR) service or in related services. As there is not yet a strict definition of what functionality is included in a PAEHR or not, we included functions that may be considered external to a PAEHR in some contexts.
- Clinical content shared with patients: includes everything on the data-information-knowledge continuum that is stored in the PAEHR service and made accessible to patients.
- Human-computer interface: focuses on the usability of the PAEHR service.
- People: represents the humans involved in all aspects of the implementation and use of the eHealth application and how they experience the use.
- Workflow and communication: focuses on collaboration and communication among different users and assessing how well the eHealth application supports the current clinical workflow.
- The health care organization's internal policies, procedures, and culture: affects every other dimension in this model as it includes any internal IT policy documents and managerial procedures that may influence the implementation and use of eHealth.
- National rules, regulations, and incentives: focuses on external forces that facilitate or place constraints on the design, development, implementation, use, and evaluation of eHealth in the respective clinical settings.
- System measurement and monitoring: focuses on the need for an effective system measurement and monitoring program to identify the availability of features and functions and how they are used as well as expected outcomes and unintended consequences of the PAEHR service.

The resulting Microsoft Excel template was then used as a basis for developing a complete data collection form within the scope of the NORDeHEALTH project, including questions, response options, and comment sections for all the dimensions in the sociotechnical analysis framework used. The first version of the form was developed by the core analysis team. The sociotechnical dimensions were not changed by the team, but the added "Features and functions" dimension was retained, and the questions used in the Microsoft Excel template were clarified in many cases. The work process followed a weekly workshop format in which different framework dimensions were in focus and that lasted between February 2021 and April 2021. The draft template took form within a shared Google document, enabling collaborators in the participating countries to continuously offer feedback on the ongoing work. This format ensured that the items in the data collection form were relevant in the context of all involved countries. Finally, an analysis dimension called *metadata* was added during this iterative process, whereby information about the data collection itself—including the name of the system, the name of the researcher responsible for data collection, and information sources—could be noted.

After the first complete draft had been developed, a digital workshop was held in early May 2021 in which the core analysis team from Sweden (n=4) as well as representatives from Estonia (n=1), Finland (n=1), and Norway (n=2) participated. All participants are coauthors of this paper and are health informatics researchers with several years of experience following the implementation and subsequent use of the PAEHR system in their respective countries. The focus of the workshop was to discuss how to interpret the framework dimensions and uppermost to elaborate on and resolve some of the question formulations that had elicited many comments from other project partners in the shared document. After the 1-hour workshop, the template was refined, after which input was again sought from other project partners through email. After the last questions had been resolved, in October 2021, data collection could start based on the finalized template.

Data Collection

Data collection was undertaken in 2 different phases. The first, longer phase started in October 2021 and ended in November 2021. One representative from each country, who also took part in the workshop where the data collection template was discussed, was assigned to be the main responsible data collector from their country, and communication among these representatives occurred continuously during the data collection period. Project participants from each country filled out a copy of the data collection form (Multimedia Appendix 1) for their main PAEHR systems and with the systems shown in Table 1 in focus.

XSL•FO
RenderX

Table 1. The studied national patient-accessible electronic health record (EHR) systems for patients' web-based EHR access.

Country	System
Sweden	1177.se and Journalen [22]
Estonia	Digilugu [23]
Finland	Omakanta [24]
Norway	Helsenorge [25]

Some ambiguities were identified during the data collection, and these were handled through communication among the responsible researchers. When needed, the data collection master file was updated in accordance with agreed-upon solutions while making sure that everyone always used the question formulations in the latest version. After completion, all main responsible researchers from each country shared their filled-out forms in a shared folder for everyone to access.

After a preliminary walk-through of the collected data performed by the core analysis team, a need for further clarification was identified. In some cases, a few questions remained unanswered, and in other cases, the level of detail varied, requiring follow-up questions. When questions that the researchers from each country needed to elaborate on had been compiled, an additional and final data collection took place in December 2021. This shorter phase of data collection took place through email exchange between the leader of the core analysis team (JM) and the researchers responsible for data collection in each country. Multimedia Appendix 2 summarizes the sources of information that were used during the data collection in each country.

As a last step of data collection for this particular paper, a project representative from each involved country was tasked with collecting and summarizing information to enable a broader overview of national contexts than the developed data collection form could provide. In this final stage, information related to government structure, overall health care system, digital care organization, and steering of health information and communications technology (ICT) developments was gathered from each country. Information was gathered from national statistics, agency web pages, and local contacts within each health care system. Data collection for this study ended in November 2022.

Analysis

When data collection was completed, the core analysis team began the analysis work. The first step was to copy all answers from the completed data collection forms into a shared analysis document. In this document, each dimension had its own sheet to simplify comparisons within each dimension. In each sheet, the questions from the data collection form were added as rows, and each PAEHR system was added as a column to create a matrix where each column included the answers related to a specific PAEHR system. Each answer, possibly in combination with an additional comment, was added to the corresponding cells in the matrix. Some examples are shown in the *Results* section. The content of the sheets was then compared across columns (countries) for all questions (rows) to identify similarities as well as aspects that are unique to the specific countries involved. In this step, representatives from Estonia, Finland, Norway, and Sweden were once again invited to discuss the identified similarities and differences.

Ethical Considerations

The study presented in this paper is part of the larger NORDeHEALTH project, which received ethics approval from the appropriate national ethical bodies. This particular study did not involve human participants or sensitive personal data and only focused on contextual and technical details of PAEHRs; there were no specific ethical requirements that needed to be addressed.

Results

Final Data Collection Instrument

Overview

The final data collection form inspired by the sociotechnical dimensions in the framework developed by Sittig and Singh [17], including all updates made during the initial data collection round, can be found in Multimedia Appendix 1. Table 2 describes the overall content and structure of the form. The data collection form is the end result of the method development and is intended to enable sociotechnical comparisons of eHealth services for patients across countries, with a specific aim to compare PAEHR systems. It is important to note here that the form derived for this study is specifically aimed at comparison of PAEHR systems. For other types of eHealth services for patients, such as self-tracking applications, a different form may have to be derived through the process suggested in the *Methods* section.

XSL•FO
RenderX

Table 2. The sociotechnical framework dimensions of the final data collection form (following the work by Sittig and Singh [17])[a].

Framework dimension	General description	Questions
Metadata	Information related to the data collection itself, such as the organizations and persons that are responsible for the data collection in the other dimensions, as well as the geographical and organizational context of the providers of PAEHRs[b] in the country	1.1-1.6
Hardware and software computing infrastructure	The hardware and software required to run the applications; this also includes issues regarding information security, access control, and standards used	2.1-2.8
Features and functions	Important features and functions in the PAEHR service or in related services; some functions that may be considered external to a PAEHR in some contexts have been included, such as appointment booking, requesting renewal of prescriptions, and access to logs	3.1-3.14
Clinical content shared with patients	An inventory of which parts of a professional record are shared with the patients; examples are medications, laboratory test results of various kinds, images, and text notes	4.1-4.21
Human-computer interface	The "Human-computer interface" dimension captures information on studies conducted on the usability of the PAEHR service; it includes both the variables that have been measured and the results of measures from different stakeholders' perspectives	5.1-5.10
People	"People" includes an overview of various characteristics of the population in the country, including language groups, educational levels, and internet use; it also includes user demographics of the PAEHR system	6.1-6.6
Workflow and communication	This dimension captures information focusing on collaboration and communication between health care professionals and patients and the explicit role of the PAEHR if identified and promoted in the country, possibly for distinct patient groups	7.1-7.11
The health care organization's internal policies, procedures, and culture	An inventory of internal IT policy documents and managerial procedures that may influence the implementation and use of the PAEHR	8.1-8.6
National rules, regulations, and incentives	An inventory of national regulations that may facilitate or place constraints on the design, development, and implementation of the PAEHR	9.1-9.9
System measurement and monitoring	Collection of information on existing system measurement and monitoring programs to identify the use of the PAEHR system as well as individual functions; this also includes unintended consequences of the PAEHR service and other feedback to the national systems	10.1-10.5
Health care system context	The general context information about the health care system in the specific country where data are collected is recorded in this dimension; information about governance structure and EU[c] membership as well as primary care organization and financing of health care is included in this dimension, and it is especially important if one performs comparisons across several countries	11.1-11.4

[a]The questions refer to the numbering in Multimedia Appendix 1.

[b]PAEHR: patient-accessible electronic health record.

[c]EU: European Union.

During the work process, some of the dimensions by Sittig and Singh [17] were adjusted, and some dimensions were added in response to identified needs that were not fulfilled by the original framework. The following dimensions were either updated or added (Table 2):

1. Metadata: this dimension was added during the work process to enable collection of information related to the data collection process itself. It stores general information that is important for keeping track, especially in large-scale projects.
2. Features and functions: we decided to use the existing framework dimension Clinical content (after renaming it—see the following item on this list) to describe the information to which patients or citizens are given access through the PAEHR (eg, laboratory test results, medications, and clinical notes). However, there is also variation in the functions that PAEHRs provide (eg, whether patients can comment or fill out forms). Considering the importance of these types of differences, we determined that a new dimension was warranted. Thus, this dimension ensured that the focus was not only on the clinical information that patients have access to in the PAEHR but also on the other important functions or features that they can use.
3. Clinical content shared with patients: in the original framework, this dimension was called "Clinical content." The dimension was renamed to make it more PAEHR specific.
4. The health care organization's internal policies, procedures, and culture: in the original framework, this dimension was more generally about policies, procedures, and culture. The decision was made to add "Health care organizations" as a specification of this dimension. This made it better suited for PAEHR analysis.
5. National rules, regulations, and incentives: in the original framework, this dimension was named "External rules, regulations, and pressures." The redefined dimension puts the focus more on the national context in relation to rules and regulations, again to make it more suitable for PAEHR analysis.

XSL•FO
RenderX

6. Health care system context: this dimension was also added during the work process to understand the basics of the countries involved in the comparison. This understanding is needed when analyzing the data in the other dimensions as the national context, including health care system financing and steering of health ICT system development, has an effect on all the other dimensions. It is especially important to gather general data about the national context when comparing across several countries. This dimension was added to the final data collection form after the information had been collected in each country, and hence, it was added in response to a need that the data collection form used did not fulfill.

Identified Clusters of Dimensions

As the collected material is extensive, the dimensions were clustered to enable more manageable partitions of the data set. Dimensions that had similar focuses were grouped together. Textbox 2 presents the 4 resulting clusters. The *Contextual factorsenabling PAEHR access and use* cluster gathers dimensions focusing on the user base as well as technical and governmental prerequisites for PAEHR access and use. The added *Metadata* dimension is also included in this cluster as it is a prerequisite for the type of comparative studies that we have conducted, and it also includes contextual information. The *Features and content* cluster includes dimensions focusing on what is actually offered to patients in the PAEHR, that is, the content (eg, clinical notes, test results, and images) and functions (eg, secure messaging and prescription renewal) that are provided to patients. The *Evaluations of human-computer interaction and use* cluster includes the dimensions focusing on how one measures and evaluates PAEHR use, as well as the results of such evaluations. Both internal service provider evaluations and external evaluations are included to provide a broad coverage. Finally, the *National and local policies, regulations for use, promotion, workflow, and communication* cluster includes dimensions focusing on laws and regulations; the focus in this cluster is also more on health care professionals and their relationship to PAEHRs than on the patients.

Textbox 2. Clustering of the sociotechnical framework dimensions used in the reporting of the findings.

Contextual factors enabling patient-accessible electronic health record access and use

- Metadata
- Hardware and software computing infrastructure
- People
- Health care system context

Features and content

- Features and functions
- Clinical content shared with patients

Evaluations of human-computer interaction and use

- Human-computer interface
- System measurement and monitoring

National and local policies, regulations for use, promotion, workflow, and communication

- Workflow and communication
- The health care organization's internal policies, procedures, and culture
- National rules, regulations, and incentives

In the article series about the results of the sociotechnical analysis, each article focuses on different clusters from Textbox 2. This paper has the *Contextual factors enabling PAEHR access and use* cluster in focus; hence, the dimensions belonging to this cluster were considered in detail, and the results of the data collection in relation to those dimensions are presented in the following sections.

Results From the Sociotechnical Comparison of the "Contextual Factors Enabling PAEHR Access and Use" Cluster

In this section, the results gathered by means of the derived sociotechnical data collection form, with a focus on contextual factors, are presented.

Metadata

The metadata dimension, which was not part of the original sociotechnical framework presented by Sittig and Singh [17], includes basic information related to the data collection process itself and the system in question. Collected information about the PAEHR systems is presented in Table 3. Differences can be observed regarding the type of provider responsible for sharing EHR services in each country. In Sweden, the responsible provider is a publicly owned company, Inera AB, whereas the responsible providers in the other countries are institutions or ministries. In Sweden, Finland, and Estonia, the PAEHR service is national; however, in Norway, the studied service, Helsenorge, is only used for PAEHRs in 3 of 4 regions. All countries only provide 1 national PAEHR service (Table

XSL•FO
RenderX

3); however, local PAEHRs can exist in parallel with the national PAEHR. In all studied countries, health information from different sources is collected in a single PAEHR.

Table 3. Data collected for the items of the Metadata dimension by country[a].

	Sweden	Norway	Finland	Estonia
Name of the national PAEHR[b] service	1177 Journalen[c]	Helsenorge	Omakanta	Digilugu
Responsible provider	The Swedish eHealth organization Inera AB	The Norwegian Directorate of eHealth and Norwegian Health Network	The Finnish Social Insurance Institution (Kela)	The Estonian Ministry of Social Affairs (and the Health and Welfare Information Systems Centre and the National Institute for Health Development)
Geographic area	National	3 of 4 health regions in Norway	National	National
Number of national PAEHRs that 1 patient can have	1	1	1	1

[a]The items shown in the table correspond to questions 1.1 to 1.4 in the data collection form (Multimedia Appendix 1). Question 1.6 regarding information sources is answered in Multimedia Appendix 2.

[b]PAEHR: patient-accessible electronic health record.

[c]A total of 2 different Swedish systems (1177 and Journalen) were analyzed. As 1177 includes Journalen (which presents most of the PAEHR information) as well as some other related features, the results from the 2 systems are merged here. In cases in which, for example, the setup of Journalen differs from that of the rest of 1177, the differences will be highlighted.

People and Demographics

This dimension was more complex than the others covered in this paper as demographic data were not available from the same time intervals and age intervals in the different countries. In addition, some of the countries did not have user group statistics of the PAEHR available. Hence, a complete comparison across all countries cannot be made, and as a consequence, results will only be summarized at an overall level in this section.

Sweden has the highest number of inhabitants (10.2 million) of the countries compared, and Estonia has the smallest (1.3 million). Norway and Finland have 5.4 and 5.5 million inhabitants, respectively. Thus, there are large differences in population size among the countries. In addition, there are large differences regarding the proportion of immigrants in the countries, with the lowest proportion (9%) in Finland and the highest proportion (31% non-Estonians) in Estonia. In Norway and Sweden, the proportion is approximately 20%. Data on internet use among the populations of the 4 countries were available from 2019 in Finland and 2020 in Estonia, Norway, and Sweden. The data showed that a high proportion of the populations were internet users, with the highest number in Norway (98%) and the lowest in Estonia (89%). Data from all countries showed a general trend of increasing internet use. Estonia and Norway lack statistics on the use of PAEHRs in different user groups (such as age, gender, and profession). In Finland, statistics on PAEHR use in different age groups are collected, with the highest number of users being aged 36 to 50 years (94%) and 18 to 35 years (93%). In Sweden, data are collected for the PAEHR service Journalen when it comes to age intervals and gender. The service is used most frequently by individuals aged 20 to 29 years and 30 to 39 years, and slightly more female (53%) than male (47%) individuals use it. Educational levels are comparable across countries, with the vast majority of the population reaching upper secondary education. The highest proportion with at least 3 years of higher education can be found in Sweden (37%), and the lowest proportion can be found in Finland and Estonia (23%). No statistics from any of the countries could be found regarding the proportion of the population that prefers an interpreter in their contacts with health care.

Hardware and Software Computing Infrastructure

This dimension includes several important contextual factors that are presented in Table 4. Among the compared countries, 2 different ways of storing data are represented. In Norway, Finland, and Sweden, data are stored in local EHRs and then extracted and presented in the PAEHRs at runtime. For Digilugu in Estonia, centralized storage is used. In most of the studied countries, data are provided to the PAEHR from both private and public providers and from both primary and secondary care. The exception is Helsenorge in Norway, where only secondary care provides data to the PAEHR. Although all health care personnel in Norway are obliged to provide data to the EHR, it is decided on a regional level whether these data should be electronically accessible in the PAEHR. There are clear similarities among the compared countries when it comes to access points, enrollment, and authentication—all countries use 1 national access point, and each patient of a connected provider automatically receives a PAEHR. Authentication in all countries is made possible through a national electronic ID. Web browsers as well as mobile-adapted web browsers can be used by patients in all studied countries to access the PAEHR. In Norway and Sweden, there is also the possibility of using apps for iOS or Android. There are also some differences regarding how data are provided to the PAEHR. In the case of 1177, Omakanta, and Helsenorge, data are automatically linked to the source EHRs at runtime when a patient logs in to the system. In Digilugu, on the other hand, data are either automatically

XSL•FO
RenderX

uploaded from source EHRs to a central server or manually uploaded from the EHR by a health care professional. Hence, in Estonia but not in the other studied countries, it is possible for a health care professional to add content to the PAEHR specifically.

Table 4. Data collected for the items of the Hardware and software computing infrastructure dimension by country[a].

	Sweden (Vårdguiden 1177)	Norway (Helsenorge)	Finland (Omakanta)	Estonia (Digilugu)
Centralized or distributed data storage?	Data stored in local EHRs[b]	Data stored in local EHRs	Centralized storage	Centralized storage
Who provides data to the PAEHR[c]?	Private and public providers and primary and secondary care	Public providers and secondary care	Private and public providers and primary and secondary care	Private and public providers and primary and secondary care
One national access point per patient portal to the PAEHR?	Yes	Yes	Yes	Yes
How is enrollment done?	Each patient of a provider automatically receives a PAEHR	Each patient of a provider automatically receives a PAEHR	Each patient of a provider automatically receives a PAEHR	Each patient of a provider automatically receives a PAEHR
How are users authenticated?	Patients use a national electronic ID of some type	Patients use a national electronic ID of some type	Patients use a national electronic ID of some type	Patients use a national electronic ID of some type
What technical platform can the patients use for access?	Web browser, mobile-adapted web browser, app for iOS, and app for Android	Web browser, mobile-adapted web browser, app for iOS, and app for Android	Web browser and mobile-adapted web browser	Web browser and mobile-adapted web browser
How are data provided to the PAEHR?	Automatically linked to source EHRs at runtime	Automatically linked to source EHRs at runtime	Automatically linked to source EHRs at runtime	Automatically uploaded from source EHRs to a central server and manually uploaded from EHRs by a health professional
Are international standards (eg, FHIR[d], openEHR, and ISO[e] 13606) used in the interface between local EHRs and the PAEHR?	A national architecture for information services exists that defines information content that is not expressed in any international standard. However, many standards are used to build the integration. International terminologies are used for some aspects, such as *ICD-10*[f] for diagnosis and ATC[g] for class of medicinal product.	IHE[h] XDS[i] and IHE plus national information structure standards	National profiles of information content expressed in various HL7[j] syntaxes—(HL7 version 3: CDA R2[k] and HL7 FHIR) plus a number of modern web-based service standards for exchange and information security	National profile of old-type HL7 standards, such as HL7 version 3 and HL7 CDA R2 combined with international terminologies such as *ICD-10*, ATC, and LOINC[l]

[a]The questions correspond to questions 2.1 to 2.8 in the data collection template (Multimedia Appendix 1, where all response options can also be found).
[b]EHR: electronic health record.
[c]PAEHR: patient-accessible EHR.
[d]FHIR: Fast Healthcare Interoperability Resources.
[e]ISO: International Organization for Standardization.
[f]ICD-10: International Classification of Diseases, 10th Revision.
[g]ATC: The Anatomical Therapeutic Chemical Classification System.
[h]IHE: Integrating the Healthcare Enterprise.
[i]XDS: Cross Enterprise Document Sharing.
[j]HL7: Health Level Seven.
[k]CDA R2: Clinical Document Architecture Release 2.
[l]LOINC: Logical Observation Identifiers Names and Codes.

Health Care System Contexts

During the analysis of the collected data, it became clear that it is necessary to provide basic descriptions of the health care system contexts for readers who are unfamiliar with them. Therefore, a summary of the most essential information for understanding the PAEHR context is provided in Textbox 3. The full data are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Textbox 3. Data collected for the items of the "Health care system context" dimension by country. The aspects considered correspond to questions 11.1 to 11.4 in the data collection template (Multimedia Appendix 1). For these items, there were only free-text responses.

Governance structure

- Finland
  - A parliamentary republic, which is a member of the European Union (EU)
  - There are 3 levels of government: state, regions (19), and municipalities (310)
- Estonia
  - A parliamentary republic, which is a member of the EU
  - There are 3 levels of government: state, counties (15), and municipalities (79)
- Norway
  - A constitutional monarchy, which is not a member of the EU (but a member of the European Economic Area)
  - There are 3 levels of government: state, counties (11), and municipalities (356)
- Sweden
  - A constitutional monarchy, which is a member of the EU
  - There are 3 levels of government: state, regions (21), and municipalities (296)

General health care system financing

- Finland
  - Municipalities are responsible for providing health care that is financed through local tax.
  - Every resident in the country is entitled to health care services from the tax-funded system (funds from the 3 levels of government and a small private sector). Municipal authority hospitals provide specialist care, and privately owned hospitals supplement with, for example, day surgery.
- Estonia
  - The health care is organized in 3 levels—primary or family care, specialist care, and nursing care
  - Payroll tax covers 78% of all expenditures. The Estonian Health Insurance Fund (EHIF) is the sole provider of universal health coverage. The EHIF covers approximately 95% of the population; the rest is covered by the Ministry of Social Affairs. The ministry also has a main responsibility.
  - All health care providers are independent entities; family physicians are local entrepreneurs in private companies.
- Norway
  - The health care is organized in 2 levels—primary care (municipalities) and specialist care (the state). Specialist care is divided into 4 health regions and provides specialist care (mainly hospital based) and ambulance service. The state financing is channeled through the health regions.
  - The tax-funded system covers health care for all citizens.
- Sweden
  - The health care is organized on 3 levels—the state, regions, and municipalities. The regions with their own elected parliaments decide on regional tax, which provides the major financing for health care. National tax funds also constitute large parts of regions' and municipalities' health care budgets. Home care is funded through local municipality taxes.
  - The tax-funded system covers everyone, including recent immigrants. It is free for inpatient care, and outpatient care has a low cost. Most of the care is operated by regions, but especially primary care and some specialist services are performed by private companies under contracts with the regions. The private share varies a lot among regions, with the largest private providership in the capital region where also a small private part exists independent of the public system that is financed mainly through optional insurance.

Primary care

- Finland
  - These are municipality-arranged services at municipal health centers.
  - They include population health monitoring and the promotion of well-being and health as well as prevention, diagnostic services, and treatment.

XSL•FO
RenderX

- Estonia
  - Every EHIF-insured individual (and every Estonian resident) is assigned to a personal general practitioner (GP), who is the first level of contact. The insured individual can choose the GP. Approximately 70% of GPs are in solo practices. Practice lists cover the entire population. There are no copayment fees for primary care services in the EHIF package.
  - The GP is the main point of contact for health benefits; they are expected to manage most patient pathways and to refer patients to specialist care or long-term nursing care and rehabilitation.
  - Provides >50% of ambulatory care visits.
- Norway
  - Services are arranged by the municipalities based on the GP scheme. All citizens have the right to register with a GP of their choice. There are approximately 5000 GPs in total.
  - The GPs are the first line of contact; they coordinate care, are gatekeepers for welfare goods, and manage referrals to specialist care.
  - Municipalities also offer, for example, emergency care, home care, and rehabilitation services.
- Sweden
  - Services are arranged by the regions. Each citizen is connected to a primary care team (often physicians, nurses, psychologists, and physiotherapists), which is the first line of contact. There are approximately 900 primary care centers, between 10% and 50% private depending on the region. Primary care is never provided by single physicians. Care centers are relatively large, with 2 to 10 physicians and a total staff of 20 to 80 people. Organization differs somewhat among regions.
  - Primary care handles many issues directly but may also refer to specialist inpatient or outpatient clinics. They also work with health promotion and preventive care (eg, vaccination) and maternity and child care and are the direct clinical contact of the municipalities' home health care.

Steering of health information and communications technology (ICT) development

- Finland
  - The national solution (Omakanta) is provided by the Social Insurance Institution (Kela). Private and public providers often also provide local portals with no connection to Omakanta.
  - Decisions on regional health ICT systems are often made by municipalities or hospital districts. Health Village (portal implemented by all university hospitals) and Omaolo (municipality collaborative) are national initiatives.
- Estonia
  - There are 2 main domains: Central databases, services, and applications: the National Health Information System (central database and services governed by the Ministry of Social Affairs) and the database of the EHIF are the major components. Databases and applications of health care stakeholders: electronic medical records (EMRs) and health information systems (HISs) of care facilities, as well as applications provided and maintained through private companies, are the major components. EMRs and HISs need to comply with central systems and legal regulations on, for example, data sharing.
- Norway
  - The Directorate of eHealth was established in 2016 and is responsible for implementing national policy and steering and coordinating eHealth initiatives with stakeholders. They are generally responsible for the work in national eHealth programs, including the national portal Helsenorge. Since 2020, the Norwegian Health Network has had the responsibility to develop and operate the portal. It is the health regions' decision whether to use the national offered health portal for patient-accessible electronic health records (PAEHRs) in their region.
- Sweden
  - Major decisions are made by the 21 autonomous regions; hence, there is not much national coordination. However, collaborations on common procurement of electronic health record (EHR) systems among regions, as well as among secondary care, primary care, and municipalities in some larger regions, have happened recently (Cambio COSMIC is procured by 17 of 21 regions).
  - Inera (subsidiary of the Swedish Association of Local Authorities and Regions) supports regions with ICT development and interoperability issues. Inera manages the 1177 patient portal, where Journalen resides.
  - The eHealth authority (under the Ministry of Health and Social Affairs) manages electronic prescriptions of medicinal products. They hold the prescription database and link together all pharmacies.

XSL•FO
RenderX

Discussion

Aims and Motivation

The study presented in this paper had two main aims: (1) to develop a detailed collaborative method suitable for cross-country sociotechnical analyses of eHealth services for patients' access to their records and (2) to illustrate the data collection method by comparing results regarding contextual factors enabling PAEHR access and use in 4 European countries. Starting from a sociotechnical analysis framework developed by Sittig and Singh [17] with an intention to produce a detailed data collection template for cross-country sociotechnical comparisons, a collaboration method was developed and implemented. The resulting data collection instrument was then used in a case study where a cross-country analysis was conducted based on collected data from Estonia, Finland, Norway, and Sweden.

This sociotechnical analysis is valuable because, while innovation in PAEHRs is important, there must be a concomitant, ongoing focus on how these tools are integrated into health systems. Digital innovations can fragment care or risk not being used, and examining the social and technical factors pertaining to PAEHRs in different settings is crucial. This will help us understand potential nonadoption; abandonment; and the effects of using such tools in health care on a wide range of stakeholders, including patients and clinicians.

Method Development

Developing a method that involved collaboration among knowledgeable representatives from all countries involved was a necessity for arriving at the desired result—a data collection instrument that enabled a detailed cross-country sociotechnical system comparison. Earlier research based on more limited sociotechnical comparisons has shown that PAEHR systems differ greatly across countries, mostly due to differences in health care policy, underlying care structures, and health ICT initiatives [3]. To be able to develop a data collection instrument that both covers important aspects of the investigated countries and systems and is relevant in the respective health care system contexts, it was deemed of high importance to involve experts from each country.

The NORDeHEALTH project participants were not experts on every aspect covered, but they had established connections with relevant health authorities and development companies that could provide necessary information through documents or in-person communication. These contacts were necessary for the success of the study, and hence, the continuous contact with these external national agencies and companies was a vital part of the developed collaboration method. On the basis of the experience with the method development, we argue that these external connections need to be established before doing similar research in the future.

The sociotechnical analysis framework by Sittig and Singh [17] on which this work was based was general and theoretical in the sense that dimensions but not any specific data collection items were presented. The framework was also aimed toward sociotechnical health systems in general and not PAEHR systems or even eHealth services for patients in particular. Hence, a big part of the method was devoted to developing a data collection instrument that would work in this specific context—a cross-country analysis of PAEHR systems in the 4 involved countries. An inevitable consequence of this is that the data collection instrument developed in this study cannot be used to enable sociotechnical analysis of other kinds of health care systems. This being said, a similar method can be used to develop other data collection instruments that include items for the different framework dimensions that are better suited to other kinds of health care systems.

The framework by Sittig and Singh [17] has indeed been adapted for sociotechnical analysis of health care technology systems but not for cross-country comparisons. As a consequence, an important part of this work was to amend it for the purpose of comparing health care systems across countries. In total, 2 dimensions were added to the framework for this purpose—metadata and health care system context. It is beneficial to include the *Metadata* dimension in any kind of sociotechnical analysis, but we argue that it is especially important to include this dimension when comparing data from different countries with their own systems, responsible data collectors, and information sources.

The collection of data for the *Health care system context* dimension occurred after the data collection that was based on the developed data collection instrument. After data collection, this dimension was added as a last step of the iterative development of the method and the data collection instrument in line with the design science research approach followed in this study. As earlier research has shown that PAEHR systems vary across countries due to contextual factors [3], this dimension is of high importance to consider in a cross-country comparison of the kind carried out in this study. The national setup of primary and specialist care, as well as general digital care infrastructure and steering of health ICT development, can affect the results of all other framework dimensions. Hence, we argue that this dimension is necessary to include in these kinds of studies.

Country Comparison

The sociotechnical cross-country comparisons regarding contextual factors brought to light several differences and similarities among the involved countries. *Hardware and software computing infrastructure* was the dimension in which most of the similarities could be found. However, it is noteworthy that the use of standards differs considerably across countries, which could potentially give rise to interoperability issues in case, for example, the Nordic countries were to move toward a joint service. It is also of interest to note that physicians in Estonia manually upload some content to the PAEHR, whereas Sweden and Norway only use the automatic link to EHRs at runtime. In Finland, content is also uploaded to the centralized PAEHR but by technical or administrative staff rather than clinicians. This means that health professionals in Estonia can be said to be active users of the PAEHR services. There are some risks associated with being an active user. Aside from the fact that the workload of health care professionals

XSL•FO
RenderX

increases, there is also a risk that the personal opinions of health care professionals may affect their willingness to share information and, consequently, the patients' potential to access their information. Overall, the level of engagement of health care professionals affects whether patients will be able to access some information. In Norway and in some regions in Sweden, patients can only see notes that are signed by a health care professional. Some regions in these countries also have a default delay in publishing notes to patients regardless of whether they are signed. This makes it possible to hide information from patients for some time, for example, by not signing a note.

When it comes to user demographic information, similarities could also be found, but it was difficult to compare data due to differences among the statistical information that the countries provide and how these data are presented. In these kinds of analyses, it is important to be able to compare user statistics on how different groups use the PAEHR services, and this comparison was not possible as the statistics covered different user groups in Sweden and Finland. Estonia and Norway did not (at the time of data collection) collect statistics on individual PAEHR use. The difference in available national statistics on demographic use clearly shows the need to collect user statistics preferably based on comparable user groups and age intervals. This would aid future comparisons as well as evaluations of PAEHR services within countries.

In the *Health care system context* dimension, several differences could be found regarding health care system setup and financing. While all involved countries have 3 levels of government (state, regions or counties, and municipalities), there are differences in which levels provide most of the funding and are responsible for different levels of care. While the regions are responsible for most health care services in Sweden, the responsibility is more divided in the other countries. Differences are also found regarding steering of health ICT development. While, in most countries, the national PAEHR solution is managed by government-owned companies or insurance institutions, there are differences in other areas of health ICT.

Strengths and Limitations

To our knowledge, this kind of detailed comparison among PAEHR systems in different countries is novel. While Essén et al [3] compared PAEHR systems in 10 different countries, including Estonia, Finland, Norway, and Sweden, their comparison focused on the effect of differences in regulations on a few aspects of PAEHRs. Our study expands this research by using a complete sociotechnical framework as a base, enabling more detailed comparisons and a more in-depth analysis of similarities and differences across countries. This kind of comparison is important if we want to understand the effects of contextual factors on the realization of PAEHR systems and will also make it possible for systems to learn from each other. This learning takes place not only when analyzing the results from the data collection but also when carrying out the developed collaboration method.

This study used a specific sociotechnical framework as a base, and even though the framework by Sittig and Singh [17] is tailored toward health systems, it risks constraining the focus to certain dimensions. There are other sociotechnical frameworks that could have been used instead. This being said, the involved researchers did add and change some dimensions to better fit PAEHR systems in particular and the cross-country comparison. Hence, even though a specific framework was used as an important base, it did not completely constrain the focus.

The number of countries involved—and, hence, the number of compared systems—was fairly limited, especially considering the fact that there are several clear similarities among the Nordic countries when it comes to national contexts. Comparisons involving more diverse countries should be conducted when validating the method.

Future Work

This paper presented the results of a sociotechnical cross-country comparison among PAEHR systems in Estonia, Finland, Norway, and Sweden related to contextual factors. The dimensions that were in focus in this paper cover 4/11 of the total dimensions in the data collection instrument. Thus, there are more in-depth comparisons being made based on the data collected in this study. In future work, more dimensions will be included, and the results will be published in subsequent papers.

In future research, it would also be beneficial to increase the number of countries to be compared. The next step in this direction could be to extend the analysis conducted by Essén et al [3] by using the framework by Sittig and Singh [17] and the collaboration method presented in this paper. This would not only include a comparison of PAEHR systems from more countries with larger differences in national contexts, but it would also validate the method in a more complex study. These kinds of more expanded comparisons could also use other frameworks as their base. It is also important to note here that digital forms were not developed in this study as the number of systems to be included in the data collection in the 4 involved countries was fairly limited and, hence, the statistical analysis made possible in many digital tools would not add any value. Instead, the data collection was prepared as a Word template to be filled out digitally. In future research including more countries and systems, it would be well worth introducing digital forms and statistical analyses.

The study presented in this paper only focused on PAEHR systems. Even though these systems are in increasing focus in research nowadays, there are many more types of eHealth services for patients as well as other types of health care systems. In future research, we would like to see this collaboration method used for developing data collection instruments related to other kinds of health care systems, such as apps for self-tracking or systems for video visits.

Future research could also usefully focus in greater depth on particular sociotechnical dimensions of PAEHRs in different countries. For example, regarding human-computer interfaces, investigators might explore the highly specific design features of portals, which may augment psychological dispositions to access health information, or (possibly more importantly) offer feedback to providers.

XSL•FO
RenderX

Conclusions

In this work, we aimed to develop and test a method that would enable cross-country sociotechnical analysis of PAEHRs in 4 countries where PAEHR use has reached maturity: Estonia, Finland, Norway, and Sweden. The main artifact produced during this work was a sociotechnical data collection template that was based on a sociotechnical framework for studying eHealth services derived by Sittig and Singh [17]. The data collection template not only considers the dimensions in the framework by Sittig and Singh [17] tailored for single-system analysis but also includes parts deemed necessary for cross-country comparisons of several systems. Close collaboration during the process among researchers from all involved countries ensured relevance in all settings.

The data collection template was tested through data collection in the 4 countries and a sociotechnical analysis. Results indicated several important similarities and differences among the 4 countries, clarifying that the process followed and the data collection template enabled an in-depth cross-country sociotechnical analysis of PAEHR systems. This paper presents the first part of the results from the sociotechnical analysis—similarities and differences among contextual factors—and companion articles will delve deeper into the remaining parts.

Acknowledgments

This work was supported by NordForsk through the funding to the Nordic eHealth for Patients: Benchmarking and Developing for the Future (NORDeHEALTH project 100477); by the Strategic Research Council at the Academy of Finland (grants 352501 and 352503); and by the Swedish Research Council for Health, Working Life, and Welfare through the funding to Beyond Implementation of eHealth (project 2020-01229).

Data Availability

The data sets generated and analyzed during this study are not publicly available due to the fact that additional publications based on this data set are being produced, but are available from the corresponding author on reasonable request.

Conflicts of Interest

CB is the Associate Editor of *JMIR Mental Health*.

Multimedia Appendix 1
Final data collection instrument.
[DOC File , 134 KB - jmir_v26i1e55752_app1.doc]

Multimedia Appendix 2
Information sources.
[DOC File , 29 KB - jmir_v26i1e55752_app2.doc]

Multimedia Appendix 3
Descriptions of health care system contexts.
[DOC File , 52 KB - jmir_v26i1e55752_app3.doc]

References

1. Cijvat CD, Cornet R, Hägglund M. Factors influencing development and implementation of patients' access to electronic health records-a comparative study of Sweden and the Netherlands. Front Public Health 2021 Jun 8;9:621210 [FREE Full text] [doi: 10.3389/fpubh.2021.621210] [Medline: 34169054]
2. Hägglund M, DesRoches C, Petersen C, Scandurra I. Patients' access to health records. BMJ 2019 Oct 02;367:l5725. [doi: 10.1136/bmj.l5725] [Medline: 31578180]
3. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018 Mar;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
4. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019 May 06;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
5. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 01;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
6. DesRoches CM, Bell SK, Dong Z, Elmore J, Fernandez L, Fitzgerald P, et al. Patients managing medications and reading their visit notes: a survey of OpenNotes participants. Ann Intern Med 2019 Jul 02;171(1):69-71. [doi: 10.7326/M18-3197] [Medline: 31132794]

XSL•FO
RenderX

7. Bell SK, Folcarelli P, Fossa A, Gerard M, Harper M, Leveille S, et al. Tackling ambulatory safety risks through patient engagement: what 10,000 patients and families say about safety-related knowledge, behaviors, and attitudes after reading visit notes. J Patient Saf 2021 Dec 01;17(8):e791-e799. [doi: 10.1097/PTS.0000000000000494] [Medline: 29781979]
8. Bell SK, Mejilla R, Anselmo M, Darer JD, Elmore JG, Leveille S, et al. When doctors share visit notes with patients: a study of patient and doctor perceptions of documentation errors, safety opportunities and the patient-doctor relationship. BMJ Qual Saf 2017 Apr;26(4):262-270 [FREE Full text] [doi: 10.1136/bmjqs-2015-004697] [Medline: 27193032]
9. Zanaboni P, Kummervold PE, Sørensen T, Johansen MA. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020 Feb 07;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
10. World Health Organization. Digital health platform handbook: building a digital information infrastructure (infostructure) for health. 2020. URL: https://apps.who.int/iris/handle/10665/337449 [accessed 2023-12-22]
11. What is an electronic health record (EHR)? HealthIT.gov. URL: https://www.healthit.gov/faq/what-electronic-health-record-ehr [accessed 2023-12-22]
12. Metsallik J, Ross P, Draheim D, Piho G. Ten years of the e-health system in Estonia. In: Proceedings of the 3rd International Workshop on (Meta)Modelling for Healthcare Systems. 2018 Presented at: MMHS 2018; June 13-15, 2018; Bergen, Norway.
13. Kujala S, Hörhammer I, Väyrynen A, Holmroos M, Nättiaho-Rönnholm M, Hägglund M, et al. Patients' experiences of web-based access to electronic health records in Finland: cross-sectional survey. J Med Internet Res 2022 Jun 06;24(6):e37438 [FREE Full text] [doi: 10.2196/37438] [Medline: 35666563]
14. Blease C, Salmi L, Rexhepi H, Hägglund M, DesRoches CM. Patients, clinicians and open notes: information blocking as a case of epistemic injustice. J Med Ethics 2021 May 14;48(10):785-793 [FREE Full text] [doi: 10.1136/medethics-2021-107275] [Medline: 33990427]
15. Om pasientjournal. Helsenorge. URL: https://www.helsenorge.no/pasientjournal/om/ [accessed 2023-09-26]
16. Baxter G, Sommerville I. Socio-technical systems: From design methods to systems engineering. Interact Comput 2011 Jan;23(1):4-17 [FREE Full text] [doi: 10.1016/j.intcom.2010.07.003]
17. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010 Oct;19 Suppl 3(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
18. Hägglund M, Scandurra I. A socio-technical analysis of patient accessible electronic health records. Stud Health Technol Inform 2017;244:3-7 [FREE Full text] [Medline: 29039366]
19. Hägglund M. Nordic countries lead new initiative on patient access to EHRs. BMJ Opinion. 2021 May 18. URL: https://blogs.bmj.com/bmj/2021/05/18/maria-hagglund-nordic-countries-lead-new-initiative-on-patient-access-to-ehrs/ [accessed 2023-12-22]
20. Hägglund M, Blease C, Johansen MA, Kujala S, Scandurra I. Exploration of a socio-technical analysis template for patient accessible electronic health records. In: Proceedings of the 2020 Special Topic Conference of the European Federation for Medical Informatics. 2020 Presented at: EFMI STC 2020; November 26-27, 2020; Virtual Event URL: https://tinyurl.com/349djn9n
21. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manage Inf Syst 2014 Dec 08;24(3):45-77. [doi: 10.2753/MIS0742-1222240302]
22. 1177 Vårdguiden. URL: https://www.1177.se/ [accessed 2024-07-02]
23. Terviseportaal. URL: https://www.terviseportaal.ee/ [accessed 2024-07-02]
24. OmaKanta. Kanta. URL: https://www.kanta.fi/omakanta [accessed 2024-07-02]
25. Helsenorge. URL: https://www.helsenorge.no/ [accessed 2024-07-02]

Abbreviations

EHR: electronic health record
ICT: information and communications technology
PAEHR: patient-accessible electronic health record
RQ: research question

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.12.23; peer-reviewed by A Mahmud, P Pohl; comments to author 29.02.24; revised version received 05.04.24; accepted 08.04.24; published 26.08.24.

Please cite as:
Moll J, Scandurra I, Bärkås A, Blease C, Hägglund M, Hörhammer I, Kane B, Kristiansen E, Ross P, Åhlfeldt RM, Klein GO Sociotechnical Cross-Country Analysis of Contextual Factors That Impact Patients' Access to Electronic Health Records in 4 European Countries: Framework Evaluation Study
J Med Internet Res 2024;26:e55752
URL: https://www.jmir.org/2024/1/e55752
doi:10.2196/55752
PMID:

©Jonas Moll, Isabella Scandurra, Annika Bärkås, Charlotte Blease, Maria Hägglund, Iiris Hörhammer, Bridget Kane, Eli Kristiansen, Peeter Ross, Rose-Mharie Åhlfeldt, Gunnar O Klein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Building and Sustaining Public Trust in Health Data Sharing for Musculoskeletal Research: Semistructured Interview and Focus Group Study

Zainab K Yusuf[1,2], BSc, MSc, PhD; William G Dixon[1,3], MBBS, PhD; Charlotte Sharp[1,2,4], BSc, MBBS, PhD; Louise Cook[1], BA; Søren Holm[5], PhD; Caroline Sanders[2,6], MSc, PhD

[1]Centre for Epidemiology Versus Arthritis, University of Manchester, Manchester, United Kingdom
[2]NIHR School for Primary Care Research, University of Manchester, Manchester, United Kingdom
[3]Manchester Biomedical Research Centre, National Institute for Health and Care Research, Manchester, United Kingdom
[4]The Kellgren Centre for Rheumatology, Manchester Royal Infirmary, Manchester, United Kingdom
[5]Centre for Social Ethics and Policy, University of Manchester, Manchester, United Kingdom
[6]Centre for Primary Care and Health Services Research, NIHR Greater Manchester Patient Safety Research Collaboration, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Caroline Sanders, MSc, PhD
Centre for Primary Care and Health Services Research
NIHR Greater Manchester Patient Safety Research Collaboration
University of Manchester
Williamson Building
Oxford Road
Manchester, M13 9PL
United Kingdom
Phone: 44 01612757619
Email: caroline.sanders@manchester.ac.uk

Abstract

Background: Although many people are supportive of their deidentified health care data being used for research, concerns about privacy, safety, and security of health care data remain. There is low awareness about how data are used for research and related governance. Transparency about how health data are used for research is crucial for building public trust. One proposed solution is to ensure that affected communities are notified, particularly marginalized communities where there has previously been a lack of engagement and mistrust.

Objective: This study aims to explore patient and public perspectives on the use of deidentified data from electronic health records for musculoskeletal research and to explore ways to build and sustain public trust in health data sharing for a research program (known as “the Data Jigsaw”) piloting new ways of using and analyzing electronic health data. Views and perspectives about how best to engage with local communities informed the development of a public notification campaign about the research.

Methods: Qualitative methods data were generated from 20 semistructured interviews and 8 focus groups, comprising 48 participants in total with musculoskeletal conditions or symptoms, including 3 carers. A presentation about the use of health data for research and examples from the specific research projects within the program were used to trigger discussion. We worked in partnership with a patient and public involvement group throughout the research and cofacilitated wider community engagement.

Results: Respondents were supportive of their health care data being shared for research purposes, but there was low awareness about how electronic health records are used for research. Security and governance concerns about data sharing were noted, including collaborations with external companies and accessing social care records. Project examples from the Data Jigsaw program were viewed positively after respondents knew more about how their data were being used to improve patient care. A range of different methods to build and sustain trust were deemed necessary by participants. Information was requested about: data management; individuals with access to the data (including any collaboration with external companies); the National Health Service’s national data opt-out; and research outcomes. It was considered important to enable in-person dialogue with affected communities in addition to other forms of information.

XSL•FO
RenderX

Conclusions: The findings have emphasized the need for transparency and awareness about health data sharing for research, and the value of tailoring this to reflect current and local research where residents might feel more invested in the focus of research and the use of local records. Thus, the provision for targeted information within affected communities with accessible messages and community-based dialogue could help to build and sustain public trust. These findings can also be extrapolated to other conditions beyond musculoskeletal conditions, making the findings relevant to a much wider community.

(J Med Internet Res 2024;26:e53024) doi:10.2196/53024

KEYWORDS

data sharing; public trust; musculoskeletal; marginalized communities; underserved communities

Introduction

Background

Electronic health records contain information documented during consultations between patients and health care professionals, such as diagnoses, test results, and treatments provided. While primarily used for direct care, these records can be deidentified to support research and planning without explicit individual consent [1], supporting the development of new treatments and services [2-6]. In the United Kingdom, anonymized data from National Health Service (NHS) patients are made available for research with relevant governance approvals. Individuals may opt out of data sharing for research by completing the national opt-out process [7]. Before the COVID-19 pandemic, research suggested that most patients were positive and altruistic regarding the confidential reuse of health data [8], and this has continued to be reflected in post–COVID-19 publications [9]. Greater public confidence has been reported in health care organizations (such as the NHS in the United Kingdom), educational institutions (such as universities), or charitable sectors for reusing health data for research rather than in commercial companies [8-11].

There have been ongoing concerns about the low levels of awareness and understanding about data sharing for research, especially in people from underserved and marginalized communities who are less likely to have been included in discussion and research on this topic. Following previous high-profile debates and scandals regarding health data sharing [12-14], there is a need for greater understanding among the research community of how best to build and sustain public trust, including with diverse communities. Issues such as low awareness, consent preferences, and data security remain key areas of consideration when trying to establish public trust [15-21]. Recent reports have highlighted key variations in levels of trust according to factors such as age, with older people less likely to trust data sharing [22]. In addition, there has been lower trust in data sharing among British ethnic minorities [23,24], leading to a recognition that there is more to be done to enable seldom-heard voices to be a part of public dialogue and governance associated with data sharing [25]. This will be crucial for safeguarding and building trust across diverse communities, especially where people may associate data sharing with personal risks and negative consequences. Such negative consequences may include concern about increased risk of discrimination if sensitive health information is shared, or insurance premiums rising if private companies are provided access to an individual's data [26-28]. Varied approaches might be needed to tackle these different types of concerns. The link between low levels of awareness and low trust emphasizes the need for designing better ways of communicating and building public trust in affected communities.

Data Security and Public Trust

With the introduction of the integrated care systems, which join up NHS organizations, local authorities, and other third sector bodies to deliver integrated health and care services [29], there has been a move toward ensuring data are managed securely [30]. The recent review by Goldacre et al [31] recommended moving toward a model of data sharing based on the Secure Data Environment (SDE). The SDE enables researchers and analysts to work within a secure computing environment. SDEs are being implemented across several of the United Kingdom's health data sharing initiatives, including Health Data Research UK [32], Our Future Health [33], and Genomics England [34]. The Goldacre review stipulates that SDEs may help build public trust through "provable, credible steps" to protect patient privacy, and "by being transparent with everyone about everything that is done" with their data [31]. Protecting patient data is, arguably, a moral imperative and of prime importance when trying to build public trust. There is a need for members of the public to learn about how their personal health records are being used for research and the associated benefits and managed risks, so that they can better understand how their data are kept secure [35].

Purpose and Research Questions

Transparency and public trust are acknowledged as fundamental to enabling and sustaining health data sharing for service improvement and research [31]. Without consent from the public, potential gains from health research would be put at risk. Previous studies have explored ways to improve transparency and establish public trust in the use of health records for research, focusing on generic and conceptual discussions on data sharing [15,18-21]. Beyond a few specific examples, such as in cancer research [36], there has been less research focused on enabling dialogue about health data sharing for specific disease areas. There is also a lack of research focused on local communities which might include perspectives of those whose data are being used for specific research relevant to their health or health conditions. Therefore, it is important to engage in dialogue with affected communities (local and specific to the disease area) to build and sustain public trust, particularly in underserved and marginalized communities. Understanding the perspectives of people from diverse contexts can help tailor appropriate communication materials and methods.

XSL•FO
RenderX

This Study

This study fills previous gaps in focusing on the specific disease area of musculoskeletal research using health records, and in adopting an inclusive approach to include underserved and marginalized communities to understand and build public trust for this research. The overall aim of this qualitative study was to explore patient perspectives on health data sharing for musculoskeletal research and to improve communications for building public trust. There were 2 main research questions: (1) What are the attitudes and views of patients and carers toward health data sharing for musculoskeletal research and (2) How can we improve transparency and enable dialogue with local communities about health data sharing for musculoskeletal research to build and sustain public trust?

Methods

Patient and Public Involvement and Engagement Group

The research team formed a patient and public involvement and engagement (PPIE) group comprising patients with musculoskeletal conditions and leads from relevant community and voluntary sector organizations (n=5). The aim of working with this group was to enable public involvement in all stages of the qualitative research and the coproduction of communications materials to inform patients and public of research being conducted using local health care records. The latter was part of a wider research program known as "the Data Jigsaw" [37]. Authors ZKY and C Sanders met with the PPIE group quarterly to discuss the different aspects of the study, such as the design of participant information sheets, interview and focus group topic guides, and communication materials. They also explained the use of health data to the PPIE group and worked together with them to discuss and refine emerging findings from the interviews and focus groups. A detailed paper focusing on the distinct PPIE aspects will be reported separately.

Recruitment and Methods

We developed a semistructured interview and focus group topic guide based on themes within previous literature (such as awareness of data sharing, trust, views about potential benefits, and any concerns) and discussions with our PPIE group to explore participants' views on electronic health data sharing. The topic guide was tested out and refined through multiple meetings with our PPIE group before data collection. The guide started with a brief introduction, followed by open questions about what health data sharing meant to participants, and any prior understanding of health data sharing that they had before joining the study. Initial discussions were based upon their original views and awareness before we presented further information, including a short film to prompt discussion regarding understanding and views about data governance, and project-specific examples delivered as short films by the researchers involved. The latter prompted discussions regarding potential benefits and concerns associated with health data sharing for the musculoskeletal research being undertaken within the local data integration pilot, known as the Data Jigsaw [37]. Discussions were connected to specific research questions being addressed through the database research in the wider Data Jigsaw program (eg, developing an algorithm to detect axial spondyloarthritis, safety and effectiveness of analgesics, and linking health and social care records) [37]. Findings from the initial interviews and focus groups were used to inform co-design of a communications campaign regarding the Data Jigsaw program. Four follow-up focus groups were then conducted to capture perspectives on the communication materials along with discussing general views regarding health data sharing for research (copy of the topic guide used before and after the development of the communications campaign in Multimedia Appendix 1).

Offering participants the choice between taking part in interviews and focus groups allowed people to choose how they would prefer to participate, and provided the study with a combination of different types of data. The interviews provided in-depth patient perspectives. Focus groups enabled open discussion between several participants [38]; generating dialogue about the lack of awareness of health data sharing, security concerns, and privacy of data; as well as ways to improve transparency and awareness of health data sharing in local communities.

We recruited participants, who were not previously known to the researchers, from Greater Manchester, United Kingdom. Participants were recruited using social media platforms and local advertisements in clinical and community settings. All participants met the eligibility criteria if they were >18 years old, able to provide informed consent, had either been medically diagnosed with a musculoskeletal condition or had experience of symptoms (such as chronic pain and joint symptoms), or were carers for people fulfilling those definitions. We used a purposive sample based on whether participants self-identified with the eligibility criteria, and to maximize diversity according to age, gender, ethnicity, and socioeconomic background. We aimed to continue recruitment until identified themes were recurring through iterative data analysis, indicating that data saturation was achieved.

Overall, 8 focus groups were conducted (n=4, 50% before and n=4, 50% after the development of the communications campaign), consisting of 3 to 8 participants each. Focus groups were led by 2 members of the team (C Sanders and ZKY) with extensive expertise in qualitative research. C Sanders is a professor of medical sociology and was the lead facilitator for focus groups. ZKY is an experienced postdoctoral researcher who cofacilitated and took notes throughout the focus groups. Discussions lasted 60 to 90 minutes. We conducted a total of 20 semistructured interviews before the development of the communications campaign, each lasting 30 to 90 minutes. The focus groups and interviews were audio-recorded with consent. Six of the 8 (75%) focus groups were web-based and conducted via Zoom (Zoom Video Communications). The remaining focus groups took place in a local community space in Salford, Greater Manchester. Overall, 19 (95%) interviews were conducted via Zoom (by ZKY) and 1 (5%) interview took place at a local community center (conducted by C Sanders; for completed COREQ [Consolidated Criteria for Reporting Qualitative Research] form [39] for reporting qualitative interviews and focus groups for Multimedia Appendix 2).

XSL•FO
RenderX

Community Engagement

To enable wider engagement with affected communities, particularly with members of underserved groups, informal discussions took place at local community spaces in Salford, Greater Manchester. This was part of our PPIE strategy to include a diverse range of public and patients. The informal discussions supported wider public dialogue regarding the Data Jigsaw research program [37] and discussed the qualitative research and development of communications for a public notification campaign. These discussions included local members of the public who may not have had the opportunity to express their opinions about health data sharing previously and included people experiencing poverty and homelessness (via a community-based drop-in center), as well as members of the D/deaf community. With the help of the PPIE advisory group, authors ZKY and C Sanders arranged to attend a drop-in center for people at risk of homelessness on 2 occasions to generate interest. Approximately 71 people took part in the informal discussions. Two informal discussions were arranged and took place at the community-based drop-in center with approximately 32 people. Further, ZKY liaised with community gatekeepers to arrange informal discussions with the D/deaf community (15/71) and with volunteers who were part of a community allotment group (24/71).

Informal discussions took place at the same time as the formal focus groups and interviews were being conducted with research participants. The informal discussions were not recorded and no personal data were collected. The ethics committee approved plans for public involvement and wider community engagement. However, formal recorded consent was not required as part of the ethical governance of the study because attendees at informal discussions were not research participants. An introduction to the research topic and the purpose of the discussion was provided at the start of the discussion. It was made clear that the discussion was informal and that we would not be recording any identifiable information and would not be audio-recording the discussions. Similar to the research focus groups, open questions were asked about views and understanding of using health records for research. Brief notes were created during and following the discussions but without any personal identifiers. The notes were in bullet point form to identify prevalent issues. While the informal discussions were not part of the qualitative analysis, we have referred to these where there was resonance with analytic themes. This serves to reflect how public involvement and engagement were embedded throughout (for the completed Guidance for Reporting Involvement of Patients and the Public [GRIPP2] checklist short form [40] refer to Multimedia Appendix 3).

Ethical Considerations

The study received ethics approval from the NHS London Bloomsbury Research Ethics Committee (21/NW/0354). Participants were informed of the research aims and interests of the study topic using a detailed Patient Information Sheet at least 24 hours in advance of taking part, to enable them to consider participation and address any questions to the research team. The information specified issues that would be discussed during interviews and focus groups, as well as arrangements for audio-recording and use of anonymized data from the transcripts. Participants were informed that they could withdraw from the study at any time (although none did so) and that they could raise any questions or complaints should they wish to. Participants who agreed to be interviewed were asked to provide informed consent before participation. People participating on the web provided verbal consent recorded in advance of the interview or focus group. People participating in a face-to-face interview or focus group provided written consent immediately before the event. At the start of interviews or participation in the focus group, participants were reminded of the purpose of the discussion, and an outline of what would be covered was presented. Assurance of confidentiality and arrangements for the use of anonymized data was provided. Ground rules were outlined at the start of the focus group to support confidentiality and inclusive discussions. We also assured participants that we could pause or stop discussions when requested and give support, should anyone need this. Participants were provided with a £20 (US $26.5) gift shop voucher, bank transfer, or cash, to thank them for their time.

Data Analysis

Interviews and focus groups were transcribed verbatim by an external supplier and the transcripts were checked by ZKY. Notes were made during the interviews and focus groups and were also used to inform analysis. The data were analyzed thematically, drawing on some techniques of a grounded theory approach, including open coding and constant comparison so that transcripts were continuously compared for the iterative development of key themes [41]. Both C Sanders and ZKY independently read transcripts. Primary open coding on all data was completed by ZKY using NVivo (version 12; Lumivero) to support the coding and organization of the data. C Sanders completed independent coding of 2 initial focus groups to compare and discuss terminology for coding. C Sanders provided reflections and comments on descriptive accounts of coded data from interviews and focus groups throughout. C Sanders and ZKY also met regularly throughout the research to discuss the development and refinement of codes and themes from the data. As noted above, discussions within the interviews and focus groups began with open questions to elicit initial awareness and understanding of health data sharing, before progressing to further discussions focusing on examples of research using health records within the Data Jigsaw program and views about communication needs. This approach inevitably had some bearing on the development of coding, final themes, and elements of both inductive and deductive analysis. Transcripts were not returned to participants for comments. Codes and themes were shared and discussed periodically at wider team meetings, including with the PPIE advisory group. Participants were provided with a summary of the findings at the end of the analysis period and invited to contact the research team with any comments or questions.

Results

Participants

Between June 2022 and September 2023, 48 participants took part in the study. Participants were recruited via Twitter (n=10,

XSL•FO
RenderX

21%), Facebook (n=3, 6%), PPIE health research networks (n=11, 23%), local advertisements (n=5, 10%), local community groups (n=9, 19%), and word of mouth (n=1, 2%). Remaining participants (n=9, 19%) contacted the researcher by email after seeing an advert via one of the previous routes (source of advert not specified).

The participant sample included men (21/48, 44%) and women (27/48, 56%) of various ages, ethnic, and socioeconomic backgrounds (Table 1).

Table 1. Demographic characteristics of research participants (N=48).

Characteristics	Participants, n (%)
Diagnosis of a musculoskeletal condition	
Yes	44 (92)
No	4 (8)
Musculoskeletal symptoms	
Chronic pain	37 (77)
Inflamed joints	36 (75)
Fatigue	33 (69)
Loss of mobility and dexterity	36 (75)
Stiffness	42 (88)
Other musculoskeletal symptoms	12 (25)
Age (y)	
18-25	4 (8)
26-35	5 (10)
36-45	10 (21)
46-55	5 (10)
56-65	4 (8)
66-75	12 (25)
>75	8 (17)
Gender	
Men	21 (44)
Women	27 (56)
Ethnicity	
Asian or Asian British	1 (2)
Black African, Black British, or Caribbean	7 (15)
Mixed or multiple ethnic groups	8 (17)
White	32 (67)
Employment	
Employed full-time	14 (29)
Employed part-time	8 (17)
Self-employed	2 (4)
Not employed, looking for work	1 (2)
Not employed, not looking for work	0 (0)
Volunteer (unpaid work)	5 (10)
Student	0 (0)
Retired	18 (38)

XSL•FO
RenderX

Findings

The findings are presented in three overarching themes: (1) awareness and motivation for supporting data sharing for musculoskeletal research, (2) issues regarding security and sharing of data for musculoskeletal research beyond the NHS, and (3) transparency and communication needs for building trust in data sharing for musculoskeletal research. The development of 3 broad themes reflected a number of subthemes (Textbox 1), elaborated upon in the following sections.

Textbox 1. Summary of main themes and subthemes.

Theme 1: awareness of and motivation for supporting data sharing for musculoskeletal research

- Understanding of data sharing for research
- Understanding of security and consent for data sharing
- Expectation that data are shared to improve patient care
- Perceived public awareness of data sharing for research
- Limited knowledge of the ability to opt-out
- Altruism and data sharing for the greater good
- Improves patient and safety and benefit
- Patient empowerment

Theme 2: issues of security and sharing data for musculoskeletal research beyond the National Health Service (NHS)

- Sharing between NHS and social care
- Stigma and sensitivity of data
- Data breaches
- The role of private companies

Theme 3: transparency and communication needs for building trust in data sharing for musculoskeletal research

- Need for simple information
- Need for positive messages and examples
- Need for transparency about security
- Enabling communication of findings from research using health records

Awareness of and Motivation for Supporting Data Sharing for Musculoskeletal Research

When asked about their views on data sharing, individuals reported a perceived lack of awareness about data sharing for research, not necessarily knowing what it entailed, for example:

> *I don't think people know about it as much...I didn't know...that much about it before I got more involved.* [Interview 14, participant # F11]

and

> *[T]hey don't know it's happening...* [Interview 2, participant # M9]

Focus group participants reported a lack of public notifications that raise awareness about data sharing initiatives (with the exception of Cancer Research UK [36]):

> *[T]here's nothing...I've never seen anything on the walls, or anything like that...the only research I know is cancer research. That's the only one I know of...there's nothing on anything else, as far as I know.* [Focus group 3, participant # F41]

Others agreed, noting a marked lack of awareness. Cancer research may be particularly salient to the public because of the national advertising campaigns by Cancer Research UK [36]. Beyond specific examples like cancer, public notification about health data sharing tends to be generic, possibly reducing public interest. The comparison to other conditions, such as cancer, exemplifies the lack of awareness of data sharing initiatives for musculoskeletal or other disease-specific research, thus reinforcing the need for tailored communications about disease-specific research. Participants also emphasized the lack of trust in data sharing initiatives due to a lack of awareness (Textbox 2):

> *I don't think the intricacies of it are well-known. I think people are still fearful of data sharing and it falling into the wrong hands...so they're quite nervous about it still.* [Interview 3, participant # F10]

XSL•FO
RenderX

Textbox 2. Quotations representing awareness and motivations for supporting data sharing for musculoskeletal research.

- "I think there are a lot of people who don't understand it and are very suspicious of it." [Interview 5, participant # M12]
- "I think it is a lot of unknown you don't know who it is going to, where it is going to, what is going..." [Focus group 2, participant # F5]
- "I think if it was explained fully enough, as to why it was being used...look these are the benefits we're using it for, this is what the potential could be. I think that would allay a lot of my fears, if I'm honest." [Interview 10, participant # F17]
- "I don't think people know that all the data's being collected and you have to opt-out." [Interview 12, participant # F23]
- "...my personal understanding would just simply have been my personal data, that would be shared but with my consent, where it's needed to be shared." [Interview 10, participant #F17]
- "it [data sharing] guides the researchers, it also helps innovation, invention and new methods of treatment can only be achieved with medical data." [Focus group 4, participant # M25]
- "...the benefits would be that actually it's going to improve healthcare, it's going to improve diagnosis rates, it's going to improve what offers are out there for support." [Interview 10, participant # F17]
- "I'm happy for anyone to have my blood, it doesn't matter who they are... it's a very nice feeling... when people know the research they're doing is helping other people, and in the way it's helping other people and how things are being developed, I think that's the positive side." [Interview 16, participant # F31]

Before being informed about the Data Jigsaw program, few participants had awareness of the possibility that members of the public could opt out of having any of their personal health data shared by registering this choice via the NHS "opt-out" system. One participant who had previously worked for the NHS suggested that the opt-out had been kept a secret stating that, "I was there...for 20-odd years, I never knew that there was an opt-out...that's a secret...I'm certainly not happy with that, that the opt-out is not transparent...I think it would be nice to have this option, I can't understand why it's hidden..." (Interview 13, participant # F24).

The discussions around opting out seemed to reflect some distrust of the NHS, making people feel disempowered about their choice to share data. There were also misunderstandings of opt-in and opt-out models of consent, suggesting there is a lack of awareness about the current opt-out model used by the NHS for research purposes with one of the participants stating that, "I know that there is an option in there where you can opt-in to share all of your medical records to be visible for research..." (Focus group 2, participant # F7).

It was recommended that the NHS opt-out be made more transparent, allowing the public the opportunity to opt-out if they wish to do so:

> *I think they (the public) need to be made aware that it happens, that it's not freely available to pull out of it. It needs to be transparent I think; for people to be able to say no I'm not happy with that, don't use my details et cetera.* [Interview 13, participant # F24]

In addition, there were some misunderstandings about how deidentified data can be used without individual consent for research purposes. Many presumed that individual consent was required stating, "I had always assumed, and I may be completely wrong that [data sharing for research] would require specific consent from the patient..." (Focus group 1, participant # F3).

This issue further highlights the lack of awareness regarding the opt-out consent model.

Although there was a general lack of awareness, many perceived data sharing for research positively, after being informed of the Data Jigsaw research program (Textbox 2). Data sharing for research was considered an opportunity to improve patient care. Participants voluntarily shared background experiences which prompted discussions about data sharing and the specific research being undertaken within the Data Jigsaw program. Some examples related to a project in the data pilot about the safety and effectiveness of analgesics, suggesting the clinical questions in the data pilot were of interest to the local population stating that "some of the medication...like non-steroidal anti-inflammatory drugs, I would like to hear more about them or how relevant it is to me, if it is good to me, you guys (researchers) will help me and tell me what more I need to do about it as well" (Interview 11, #M18).

Another respondent suggested medication safety was of paramount importance for patient benefit:

> *I personally am sensitive to opioids...a lot of the girls I've been speaking to on the [fibromyalgia blog]...not just in Salford, but they've been saying throughout hospitals in Manchester...a lot of the hospitals are taking people off the opioids because they think they're being overused. They need to look into what type of ones are being given and how they are working, what are the benefits, what are not the benefits, how they work with...individual people.* [Interview 6, participant # F13]

For some, data sharing was compared with blood and organ donation and seemed to be viewed as a gift exchange that could benefit others. Thus, they felt that the benefits of health data sharing should be communicated more widely to establish trust. Examples, such as advertisements about organ donation were mentioned as models of building public trust:

> *[T]hey've got to promote the positive place of what comes out of this research and instances of how it supports people and how they can contribute to give people more confidence in the fact that it is being put to good use and that others will benefit as well as you. There's an advert...on TV...and...it lists all the*

XSL•FO
RenderX

> *people (who benefit). Someone who got their eyes, somebody who got their heart, kidneys, tissue, so it's the wider picture of what seems a simple instance of this is what can happen.* [Interview 15, participant # F30]

While many participants seemed motivated by the idea that musculoskeletal research could help them directly because of their personal experiences of musculoskeletal problems, this also highlighted altruistic motivations (ie, sharing data for the greater good and helping others). The above also indicates that learning from other campaigns may be beneficial and suggests that the impact of communication may be improved if they include motivation for supporting data sharing in this context.

A lack of awareness about the opt-out was also evident in the informal discussions with underserved communities, many of whom were not aware of this. One member had been notified of the NHS opt-out and chose to opt out of their data being used for research, noting their lack of trust in NHS organizations to keep data secure.

Issues of Security and Sharing Data for Musculoskeletal Research Beyond the NHS

Privacy concerns remained a prevalent issue throughout the discussions. Some participants mentioned concerns regarding stigmatization stating that "there is so much fear about people being really scared of being judged and the stigma that comes along with if you've got a disability or if you've got a long-term health condition." (Interview 12, participant # F23; see also Textbox 3).

Textbox 3. Quotations representing issues of security and sharing data for musculoskeletal research beyond the National Health Service.

- "It can't be emphasized enough how it [patient data] should be extremely secretive." [Focus group 2, participant #M6]
- "I think if I had something that I was very mindful that I didn't want other people knowing about me, I would be quite dubious..." [Interview 12, participant # F23]
- "[They] are a data management company...making algorithms and then feeding that back into the NHS...It's not for their own benefits of...well, we've found this out and we'll keep it to ourselves. It's actually being fed back into the NHS to actually help the public...yeah, on that aspect...I would be inclined to go forth and let them have my data." [Interview 3, participant # F10]
- "...people are really concerned...thinking about data generally how often it is sold for commercial use so I think that is the level of reassurance I certainly would want...that it suddenly wasn't going to find itself with someone else who perhaps doesn't follow the same ethical practices as the NHS..." [Focus group 1, participant # F3]
- "I think that would be for the good of everybody...and obviously they [external company] would be checked for security." [Interview 2, participant # M9]

Discussions about the possibility of linking health and social care records for research indicated concerns about researchers accessing social care records, finding it intrusive stating that "...you feel like sometimes they pry a bit" (Interview 6, participant # F13). When discussing this issue in the context of a specific exemplar project, such as linking health and social care records to understand the impact of musculoskeletal problems [37], participants seemed positive:

> *[A] lot of stuff gets missed. There needs to be a better understanding. I think it is a need-to-know basis and if they need to know then it should be...there definitely needs to be a link between the social care and...rheumatology because again it's getting back to this whole thing of, if they haven't got the information they don't know how to help you.* [Interview 6, participant # F13]

This suggests that participants were open to the idea of linking health and social care records once they became aware of the potential for patient benefit. Using a project-specific example also allowed the participants to consider whether it would be accessed on a "need to know" basis, establishing a basis for trust by providing transparent information about the benefits. This again strengthens the case for tailored communications about disease-specific research (rather than unspecified research) being conducted, for which there is then positive support.

Perspectives on sharing beyond the NHS to enable earlier diagnosis of musculoskeletal conditions, such as axial spondyloarthritis were shared. There were varying degrees of trust in commercial companies accessing patient data (Textbox 3):

> *I'm not sure about that...it can go a different way, can't it? It's all about the money, isn't it, then...?* [Interview 14, participant # M29]

However, when discussing this issue in the context of a specific exemplar project, such as developing an algorithm for enabling earlier diagnosis [37], there were no objections from participants to data sharing for an external company in partnership with the university conducting the research. The partnership was viewed positively, on the condition their data were kept secure with participants stating "I don't mind...as long as they keep...within the parameters of their research" (Interview 14, participant # M29).

Others drew on their struggles to attain a timely diagnosis and viewed the partnership positively, suggesting the Data Jigsaw research program was relevant to participants. This may have contributed to their positive views regarding collaborations with commercial companies. This may have also been influenced by the fact that the commercial company was working in partnership with trusted institutions, such as the NHS and an educational institution (Textbox 3):

> *I think that's a great idea [to work with a private company to develop an algorithm to improve diagnosis]...in the early 80s...I started with a pain in*

XSL•FO
RenderX

> *my back...I was thrown from pillar to post, I was given pain killers...I had x-rays and they said I had this spondylarthritis...that went on for several years...it was the early 2000s when I had an actual body scan of my back...to be told that it was too late, they couldn't do anything. So, with everything you've just said to me, that could have been avoided...* [Interview 13, participant # F24]

Individuals seemed to trust the NHS to keep their data secure, however, they noted the importance of improved transparency regarding external collaborations, to build public trust. It was recommended that a private company's credentials be made transparent, suggesting trust can be established if the company is perceived as reputable and has a positive track record of working for patient benefit, the latter being one of the reported motivations for participants to share data:

> *I'd want to know about the private company and their background. How legit are they? We do trust the NHS. Private companies are popping up all over the place, it could be a private company from abroad wanting that data. So, you'd be like oh, put the brakes on. What's happening here? Why's a private company wanting that? What's their past? What have they done in the past for research and patients? You'd need to be able to backup who they actually were. They'd have to have credentials.* [Interview 3, participant # F10]

The concerns around privacy and discrimination were also apparent in informal community discussions. Such discussions highlighted a lack of trust in commercial companies and concerns that sensitive information may be shared with others and potentially used against them; for example, in relation to judgments about their parenting capacity. However, they were also supportive of the specific projects described as part of the Jigsaw program and related these to potential benefits relevant to their own musculoskeletal problems.

Transparency and Communication Needs for Building Trust in Data Sharing for Musculoskeletal Research

Participants expressed a need for different ways to raise awareness and build trust within affected communities about health data sharing for musculoskeletal research. Transparency about data security and governance was considered vital to building trust (Textbox 4). A key message to relay to the public was that their data was being used for research and that it was kept safe with one of the participants stating that "The message you need to bang home is that it is only going to be for research, and it is safe" (Focus group 1, participant # F3).

Textbox 4. Quotations representing transparency and communication needs for building trust in data sharing for musculoskeletal research.

- "[W]hat you want to do is reassure people that notwithstanding that they have got your information there really is no problem to you as a member of the public." [Focus group 1, participant # F3]
- "I think it is good for people to hear how their data is being used, how it is being stored and all of those procedures..." [Focus group 2, participant # F5]
- "[F]or me it would be where is my information going, who is going to use it and how are they going to use it." [Focus group 2, participant # F5]
- "[I]f you are delivering that message more generally it will need to be a little more digestible about how you are managing the data...although you are identified by your NHS number it still isn't really a threat to you personally, there is not going to be a great leakage of your personal data so I think that is something that you could probably work on." [Focus group 1, participant # F3]
- "[P]eople who are hesitant about, I don't want to share the information, I don't want my information used...it may be reassuring to them to see that this is secure, it's been deidentified, that if you choose you could opt out but here's the kind of way it's being used and how much good it can do." [Focus group 6, participant # F45]
- "[I]t's better to have the findings and sell the positive message of we have done this and this is what you have told us and we are thinking about this." [Interview 16, participant # F31]

Although the majority expressed a view to raise awareness about data security protocols, a few individuals suggested that it is quite complex to understand and might cause confusion, suggesting it is important to consider the layers of information required to engage the public without added complexity (Textbox 4) with statements such as "it just needs to be short and snappy but not overburdened with—I think one of the terms used was locked box and whilst I know what a locked box is, I...wonder is that terminology describing a process that is used internally that actually somebody outside of that might not understand" (Focus group 2, participant # F7).

In addition, to establish trust, it was considered crucial to promote the collaborative relationship between the educational institution involved in the Data Jigsaw research program [35] and the NHS, reflecting the public's social investment in support of the NHS for the greater good. The short films used in this study demonstrated a strong link between the two. One of the participants stated that "with this group, a lot of the researchers are either still practicing clinicians or there is a strong link between the educational institution and the NHS trust there is some kind of positive relationship that isn't there for financial gain, it is there for the advancement of scientific knowledge for research helping the development of new treatments" (Focus group 2, participant # F7).

When asked if they would like information about any research projects using their health data, participants were overwhelmingly positive. One respondent suggested producing a summary report that could be circulated and to demonstrate this, used an example of donating to charity where donors are updated about their contributions. They said that "if you donate to Oxfam you get a letter each year saying your goods have been sold for this amount of money and it just makes you think

XSL•FO
RenderX

that some use has come out of you donating your time if you like to the research...” (Focus group 1, participant # F3).

This suggests that receiving information about how patient data has been used could build and sustain public trust and echoes the patient’s willingness to be acknowledged as an active agent in data sharing initiatives [42].

Research participants highlighted the importance of accessible communications as well as the importance of actively engaging with communities:

> *[W]e’re being challenged, rightly so, by communities saying, you’re not getting information out to us, it’s a one-way process, it’s not a discussion...it’s really about valuing those community links we’ve got...You’ve got to make sure...that the information that’s going out is what you want the information to say...* [Interview 12, participant #F23]

Others reiterated that “community service” is imperative to raising awareness (Interviews 17 and 18, participants #M32 and #M33).

During discussions with members of underserved communities, the issue of digital exclusion was raised. People highlighted the perceived lack of engagement and suggested researchers commit to ongoing dialogue with local communities by regularly visiting community spaces in-person and speaking directly to those who may be impacted by health data sharing and valuing community links.

This highlights the importance of digital inequalities and the need to consider digital inclusion as part of transparency and communication needs to ensure information is accessible and includes a balance of audio-visual communications and in-person discussions.

Discussion

Principal Findings

This study used examples from a local data integration pilot for arthritis research (known as the Data Jigsaw) [37] to explore participants’ views about health data sharing for musculoskeletal research. It explored ways to improve transparency and enable dialogue with local communities about health data sharing for research, to inform the design and delivery of better ways of communicating and building public trust. Three key themes characterized participants’ views: awareness and motivations for supporting data sharing for musculoskeletal research, issues of security and sharing data beyond the NHS for musculoskeletal research, and transparency and communication needs for building trust in data sharing for musculoskeletal research. The discussion is centered on the 3 key themes below.

Awareness and Motivation for Supporting Data Sharing for Musculoskeletal Research

In keeping with previous studies [4,6,8], a general lack of awareness about data sharing for research was evident among participants. The responses also resonate with previous research suggesting patients with musculoskeletal problems adopt an altruistic attitude toward helping others and a sense of social responsibility to share data for the greater good, along with personal motivations to share data for self-benefit [4]. However, previous research and communication materials to inform the public about health data sharing has often focused at a more general level rather than specific conditions within local contexts [9]. After being informed of specific examples of research within the Data Jigsaw program, people expressed positive views about sharing health data for musculoskeletal research. Participants often related discussions about data sharing in this context to their own experiences of musculoskeletal conditions and potential benefits for themselves as well as others. This highlights the potential value of tailoring specific information for a local and condition-specific population to enable meaningful engagement and greater awareness and support for specific research.

In line with previous studies [16,17], there were misunderstandings of opt-in and opt-out models of consent due to limited awareness. This meant that few participants were aware of the national data opt-out as has been found in previous studies [43], with participants suggesting it was hidden from them. This suggests education about the national data opt-out is required and that it is important to notify patients about the opportunity to opt out if they wish to do so.

Respondents believed sharing specific examples of the potential for patient benefit from health data sharing was important to gain public trust. They provided examples of mass advertising campaigns as exemplar models of establishing public trust. Learning from other campaigns may be considered beneficial and suggests that communication and trust may be improved if such campaigns reflect the public’s motivation for supporting data sharing, such as reflecting on the public’s sense of social responsibility to help others. Currently, there are minimal opportunities for members of the public to learn more about how their health data are used for research [43], a finding that our research supports. Promoting the benefits of data sharing, as well as exploring a range of ways to disseminate the research outcomes and the national opt-out may help garner further support. Examples might include developing a range of resources to interact with the public, such as posters and leaflets; or using social media or websites; and speaking to local community members to increase awareness. Organizations like Understanding Patient Data [44] have done this well, developing various resources about health data sharing for research. Dissemination of the research outcomes via public events may also attract attention and help establish public trust. It is recommended that wider advertising both to inform the public about the national opt-out, the patient benefits of health data sharing, and a thank-you for taking part may help build public trust.

Issues of Security and Sharing Data Beyond the NHS for Musculoskeletal Research

Similar to previous studies [4,6,9,43], respondents held positive views toward the NHS and educational institutions collaborating to improve patient care. There were issues of data security when sharing data beyond the NHS, particularly when accessing sensitive information that could potentially be used against them in wider aspects of life, including social care records and the

XSL•FO
RenderX

involvement of private companies. Conversely, respondents held positive views when discussing data sharing in the context of specific exemplar projects in the Data Jigsaw program. Their comments suggest the project-specific exemplars may have influenced participants' views toward data sharing for research once they became aware of the potential for patient benefit.

The principles of data minimization (ie, accessing data by trained individuals and not sharing data more widely), were also highlighted, suggesting transparency about who has access to the data can help increase the acceptability of the reuse of health data for research and establish public trust [43]. This aligns with the SDE model [35] and has been applied in the Data Jigsaw program [37]. The findings in this study reflected positive attitudes toward collaborations with external companies to improve patient care, such as improving diagnosis. This relates back to participants' reported motivations for data sharing, as it seemed participants viewed external collaborations positively if the company had an altruistic motive to benefit patient care. This again draws on the public's motivations and sense of social responsibility [4] to share data for personal and community benefit and suggests views may be influenced depending on the partnerships commercial companies have with trusted institutions, such as the NHS. These findings emphasize the potential benefits of providing specific information regarding collaborations with external companies, as well as the company's ethos of working for patient benefit to garner further support. In addition, including details of any potential collaborations with trusted institutions, such as the NHS, and describing who has access to data may mitigate the public's security concerns (ie, whether an external company has access to patient data). Researchers can also provide a data access report or data receipt, providing clarity about who has accessed their data.

Although public notification might be considered one way of garnering positive support, it is important to consider that campaigns rely on resources and funding. Mass advertising campaigns can be well-funded and effective, however, those with limited resources may suffer. It may be necessary to explore how ethical and governance processes can be made robust and transparent to ensure that the public is aware about of how their data are used without relying on mass advertising to improve awareness.

Transparency and Communication Needs for Building Trust in Data Sharing for Musculoskeletal Research

The findings suggest participants valued transparency about data security, the national data opt-out, and research outcomes. Various ways about how best to communicate with the wider public about data sharing in local communities were suggested, such as developing simple and layered communications with accessible messages. Participants also gave positive feedback on audio-visual resources and summaries used to describe examples from the Data Jigsaw program. The results of this study have been used to inform a public notification campaign, the development and outcome of which will be reported separately.

The findings highlighted a lack of awareness and misunderstanding regarding security and governance of the use of personal health data for research and identified a need to fill a gap in resources to explain SDEs. In line with the Goldacre review's recommendations to be transparent about how data are used [31], participant views informed the development of a short film to explain SDEs in a creative and understandable way for use in a public notification campaign for the Jigsaw program [45]. Raising awareness about the national data opt-out was also considered and implemented via the study website [37].

Our study maximized the inclusion of diverse communities, both through diversity among research participants and public involvement and engagement. This is important as literature has pointed to lower trust in data sharing among British ethnic minorities [23,24] and the need for greater inclusion of seldom heard voices to be part of public dialogue and governance associated with data sharing [25]. Feedback from diverse participants indicated that ongoing dialogue and engagement by reporting research outcomes through various methods, such as visiting community spaces in-person can be implemented to improve transparency and build and sustain public trust in local communities. Previous research has noted that fostering engagement and open dialogue with those who may be directly affected by data sharing may be considered challenging for researchers, but this is important for enabling support for data sharing and for sustaining public trust [42]. It also places emphasis on the importance of reciprocity and coproduction, enabling participants to be active agents in the patient-researcher relationship [42].

Strengths and Limitations

This study further explores issues raised in previous research within general population contexts [4,6,8,9,43] with the benefit of having a specific clinical context (ie, musculoskeletal research). This adds to previous work to raise awareness and collaboration on data sharing with patients affected by specific conditions, such as cancer [36] or genetic conditions [46]. The specific clinical focus with clearly specified research questions is considered a strength in being able to study perspectives and information needs directly with affected communities.

A key strength has been our extensive and inclusive approach to ensure public involvement and engagement has been embedded throughout the research, including a small coproduction working group and our wider public engagement using informal discussions in community-based settings. Speaking to people informally and in-person enabled wider engagement with the public, particularly those without access to information technology, such as members of underserved communities. The value of an inclusive and informal approach enabled relaxed discussion and communication about data sharing within busy community spaces where people already meet informally, enhancing the breadth of perspectives included in this research and providing an example of how other research studies might include communities who have previously been labeled as being "hard-to-reach" [24].

Conducting focus groups and interviews on the web via Zoom presented unique challenges. While most participants opted to use their video cameras, a small number used the camera minimally. For example, some members introduced themselves

XSL•FO
RenderX

at the start with the camera on but then turned the camera off for periods of the discussion. In such cases, this seemed to alter the group dynamics and brought some challenges for enabling dialogue and discussion. This required active management by researchers to ensure all members felt included and maintain the momentum of discussion as recognized in previous research [47].

Conclusions

Views of participants in this research were generally positive about the sharing of health and social care data for musculoskeletal research purposes but there were also low levels of awareness regarding processes and governance of health data sharing. There were associated views about the need for transparency about data security and how patient data would be used to inform patient benefit, reflecting similarities in wider literature [4,6,15-18]. This study used project-specific examples from a local data integration pilot for arthritis research (known as the Data Jigsaw) and found that participants with musculoskeletal conditions (or at risk of such conditions) were supportive of examples discussed when details were explained, the purpose of analysis, the partnerships entailed, how data would be kept secure, and specific potential benefits for people with musculoskeletal conditions. The findings have important implications for research policy and practice. Building and sustaining public trust has been identified as crucial for enabling health data sharing for research. However, few studies have included a focus on understanding and building public trust [9]. This paper adds novel insights by focusing on how to build public trust based on the perspectives of people with relevant musculoskeletal problems within a local area where health records were being used for musculoskeletal research. The combination of public involvement combined with qualitative research was found to be extremely valuable for a more inclusive approach [48,49] similar to other studies in the digital health field. In addition, the wider community-based informal discussions enabled wider inclusion of “seldom heard” voices considered vital for building public trust in health data sharing. Our findings indicate that transparency, and tailoring communications to reflect current and local research using health records, can help to raise awareness and build public trust and support in health data sharing for musculoskeletal research. As done for the Jigsaw program [37], we recommend that researchers using health data records should include a direct weblink to the NHS opt-out on communications, such as a poster or leaflet or a dedicated website to the research itself. The project also highlights the value of engagement within community settings, in addition to audio-visual materials for enabling inclusive dialogue that may contribute to building public trust in data sharing. It is recommended that researchers using data from shared health care records include adequate resource patient and public engagement in their project plans from the onset of the research.

Acknowledgments

The authors wish to thank members of our public and patient involvement and engagement group for ongoing input and advice throughout the project and members of community organizations who supported the work. The authors also thank the patients and members of the public who participated in the research.

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in education, welfare, and justice. It also funds student programs that provide opportunities for young people to develop skills in quantitative and scientific methods. The Nuffield Foundation is the founder and cofunder of the Nuffield Council on Bioethics, the Ada Lovelace Institute, and the Nuffield Family Justice Observatory. Although the Foundation has funded this project (grant OBF/43995), the views expressed are those of the authors and not necessarily of the Foundation [50].

The work was also supported by the Centre for Epidemiology Versus Arthritis (grant 21755).

This research was further supported by the National Institute for Health and Care (National Institute for Health and Care Research, NIHR) Manchester Biomedical Research Centre (grant NIHR203308). C Sanders is part-funded by the NIHR Greater Manchester Patient Safety Research Collaboration (grant NIHR204295). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

All the authors contributed to the conception and design of the study. ZKY and C Sanders were involved in preparing research materials, data collection, and data analysis. WGD and C Sanders supervised the study. All authors supported data interpretation. ZKY wrote the first draft of the manuscript, and all authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

WGD has received consultancy fees from Google, unrelated to this work.

Multimedia Appendix 1
Topic guide for focus groups and interviews.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 4770 KB - jmir_v26i1e53024_app1.pdf]

Multimedia Appendix 2
Completed COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 423 KB - jmir_v26i1e53024_app2.pdf]

Multimedia Appendix 3
Completed Guidance for Reporting Involvement of Patients and the Public (GRIPP2) checklist.
[PDF File (Adobe PDF File), 77 KB - jmir_v26i1e53024_app3.pdf]

References

1. The Caldicott Principles. United Kingdom Government. 2020 Dec 8. URL: https://www.gov.uk/government/publications/the-caldicott-principles [accessed 2023-09-11]
2. Uslu AM, Stausberg J. Value of the electronic patient record: an analysis of the literature. J Biomed Inform 2008 Aug;41(4):675-682 [FREE Full text] [doi: 10.1016/j.jbi.2008.02.001] [Medline: 18359277]
3. Stevenson JE, Nilsson G. Nurses' perceptions of an electronic patient record from a patient safety perspective: a qualitative study. J Adv Nurs 2012 Mar;68(3):667-676. [doi: 10.1111/j.1365-2648.2011.05786.x] [Medline: 21781148]
4. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: a qualitative study. J Med Internet Res 2016 Apr 15;18(4):e66 [FREE Full text] [doi: 10.2196/jmir.5011] [Medline: 27083521]
5. Data saves lives: reshaping health and social care with data. Department of Health & Social Care, United Kingdom Government. 2022 Jun 15. URL: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data [accessed 2023-06-14]
6. Stockdale J, Cassell J, Ford E. "Giving something back": a systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. Wellcome Open Res 2019 Jan 17;3:6 [FREE Full text] [doi: 10.12688/wellcomeopenres.13531.2] [Medline: 30854470]
7. Opt out of sharing your health records. National Health Service. URL: https://www.nhs.uk/using-the-nhs/about-the-nhs/opt-out-of-sharing-your-health-records/ [accessed 2023-09-11]
8. Ghafur S, Van Dael J, Leis M, Darzi A, Sheikh A. Public perceptions on data sharing: key insights from the UK and the USA. Lancet Digit Health 2020 Sep;2(9):e444-e446 [FREE Full text] [doi: 10.1016/S2589-7500(20)30161-8] [Medline: 32838250]
9. Baines R, Stevens S, Austin D, Anil K, Bradwell H, Cooper L, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 05;26:e50421 [FREE Full text] [doi: 10.2196/50421] [Medline: 38441944]
10. Alghamdi NS, Alghamdi SM. The role of digital technology in curbing COVID-19. Int J Environ Res Public Health 2022 Jul 07;19(14):8287 [FREE Full text] [doi: 10.3390/ijerph19148287] [Medline: 35886139]
11. Pujolar G, Oliver-Anglès A, Vargas I, Vázquez ML. Changes in access to health services during the COVID-19 pandemic: a scoping review. Int J Environ Res Public Health 2022 Feb 03;19(3):1749 [FREE Full text] [doi: 10.3390/ijerph19031749] [Medline: 35162772]
12. van Staa TP, Goldacre B, Buchan I, Smeeth L. Big health data: the need to earn public trust. BMJ 2016 Jul 14;354:i3636. [doi: 10.1136/bmj.i3636] [Medline: 27418128]
13. Sheather J, Brannan S. Patient confidentiality in a time of care.data. BMJ 2013 Nov 27;347:f7042. [doi: 10.1136/bmj.f7042] [Medline: 24284340]
14. Hays R, Daker-White G. The care.data consensus? A qualitative analysis of opinions expressed on Twitter. BMC Public Health 2015 Sep 02;15:838 [FREE Full text] [doi: 10.1186/s12889-015-2180-9] [Medline: 26329489]
15. Kass NE, Natowicz MR, Hull SC, Faden RR, Plantinga L, Gostin LO, et al. The use of medical records in research: what do patients want? J Law Med Ethics 2003;31(3):429-433 [FREE Full text] [doi: 10.1111/j.1748-720x.2003.tb00105.x] [Medline: 14626550]
16. Kassam I, Ilkina D, Kemp J, Roble H, Carter-Langford A, Shen N. Patient perspectives and preferences for consent in the digital health context: state-of-the-art literature review. J Med Internet Res 2023 Feb 10;25:e42507 [FREE Full text] [doi: 10.2196/42507] [Medline: 36763409]
17. Taylor MJ, Taylor N. Health research access to personal confidential data in England and Wales: assessing any gap in public attitude between preferable and acceptable models of consent. Life Sci Soc Policy 2014 Dec;10:15 [FREE Full text] [doi: 10.1186/s40504-014-0015-6] [Medline: 26085451]
18. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73 [FREE Full text] [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]

XSL•FO
RenderX

19. Cheung S. Disambiguating the benefits and risks from public health data in the digital economy. Big Data Soc 2020 Jun 18;7(1). [doi: 10.1177/2053951720933924]
20. Horn R, Kerasidou A. Sharing whilst caring: solidarity and public trust in a data-driven healthcare system. BMC Med Ethics 2020 Nov 03;21(1):110 [FREE Full text] [doi: 10.1186/s12910-020-00553-8] [Medline: 33143692]
21. Gille F, Smith S, Mays N. Evidence-based guiding principles to build public trust in personal data use in health systems. Digit Health 2022 Jul 17;8:20552076221111947 [FREE Full text] [doi: 10.1177/20552076221111947] [Medline: 35874863]
22. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit Health 2021 Apr;3(4):e260-e265 [FREE Full text] [doi: 10.1016/S2589-7500(20)30317-4] [Medline: 33678589]
23. Black data matters: how missing data undermines equitable societies. The Ada Lovelace Institute. 2020 Sep 9. URL: https://tinyurl.com/2ydhzjs2 [accessed 2023-06-20]
24. Waterman LZ, Pillay M, Katona C. Sharing health data for immigration control affects marginalised communities. BMJ 2021 Apr 26;373:n1042. [doi: 10.1136/bmj.n1042] [Medline: 33903126]
25. Castell S, Robinson L, Ashford H. Future data-driven technologies and the implications for use of patient data: dialogue with public, patients and healthcare professionals. Ipsos MORI. 2018 Nov. URL: https://acmedsci.ac.uk/file-download/6616969 [accessed 2023-05-12]
26. Mulrine S, Blell M, Murtagh M. Beyond trust: amplifying unheard voices on concerns about harm resulting from health data-sharing. Med Access Point Care 2021 Oct 01;5:23992026211048421 [FREE Full text] [doi: 10.1177/23992026211048421] [Medline: 36204496]
27. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
28. Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, NY: St. Martin's Publishing Group; 2018.
29. Introducing integrated care systems. House of Commons. 2023 Feb 8. URL: https://committees.parliament.uk/publications/33872/documents/185310/default/ [accessed 2023-09-11]
30. Building an integrated care system intelligence function. National Health Service England. 2023 Feb 23. URL: https://www.england.nhs.uk/long-read/building-an-ics-intelligence-function/ [accessed 2023-09-11]
31. Better, broader, safer: using health data for research and analysis. United Kingdom Government. 2022 Apr 7. URL: https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis [accessed 2023-06-10]
32. Varma S, Hubbard T, Seymour D, Brassington N, Madden S. Building trusted research environments - principles and best practices; towards TRE ecosystems. UK Health Data Research Alliance. 2021 Dec 8. URL: https://zenodo.org/records/5767586#.YoYSk6jMJPY [accessed 2023-05-10]
33. Tender opportunity: public cloud infrastructure, trusted research environment and billing partner. Our Future Health. 2021 Dec 17. URL: https://ourfuturehealth.org.uk/news/tender-opportunity-public-cloud-infrastructure-trusted-research-environment-and-billing-partner/ [accessed 2023-05-10]
34. The research environment. Genomics England. URL: https://www.genomicsengland.co.uk/research/research-environment#:~:text=The%20Research%20Environment,clinical%2C%20phenotypic%20and%20longitudinal%20dataAccessed [accessed 2023-05-10]
35. Secure data environment for NHS health and social care data - policy guidelines. Department of Health & Social Care, United Kingdom Government. 2022 Dec 23. URL: https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines [accessed 2023-09-10]
36. Donate to beat cancer. Cancer Research UK. URL: https://donate.cancerresearchuk.org/donate?gclid=EAIaIQobChMIxu3i6oiqgQMVxdLtCh0ABQhnEAAYASAAEgJl-fD_BwE&gclsrc=aw.ds [accessed 2023-09-13]
37. Your health data can improve arthritis care. Data Jigsaw. URL: https://thedatajigsaw.co.uk/ [accessed 2023-07-20]
38. Kitzinger J, Barbour R. Developing Focus Group Research: Politics, Theory and Practice. Thousand Oaks, CA: SAGE Publications; 1998.
39. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
40. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017 Aug 02;358:j3453 [FREE Full text] [doi: 10.1136/bmj.j3453] [Medline: 28768629]
41. Strauss AL. Qualitative Analysis for Social Scientists. Cambridge, UK: Cambridge University Press; 1987.
42. Courbier S, Dimond R, Bros-Facer V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection - quantitative survey and recommendations. Orphanet J Rare Dis 2019 Jul 12;14(1):175 [FREE Full text] [doi: 10.1186/s13023-019-1123-4] [Medline: 31300010]

XSL•FO
RenderX

43. Atkin C, Crosby B, Dunn K, Price G, Marston E, Crawford C, et al. Perceptions of anonymised data use and awareness of the NHS data opt-out amongst patients, carers and healthcare staff. Res Involv Engagem 2021 Jun 14;7(1):40 [FREE Full text] [doi: 10.1186/s40900-021-00281-2] [Medline: 34127076]
44. Understanding Patient Data homepage. Understanding Patient Data. URL: https://understandingpatientdata.org.uk/ [accessed 2023-07-12]
45. The data jigsaw: secure data environments. YouTube. URL: https://tinyurl.com/DataJigsawSDE [accessed 2023-07-20]
46. Genetic Alliance homepage. Genetic Alliance. URL: https://geneticalliance.org/resources-and-information [accessed 2024-03-26]
47. Archibald MM, Ambagtsheer RC, Casey MG, Lawless M. Using Zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. Int J Qual Methods 2019 Sep 11;18. [doi: 10.1177/1609406919874596]
48. Rai T, Morton K, Roman C, Doogue R, Rice C, Williams M, et al. Optimizing a digital intervention for managing blood pressure in stroke patients using a diverse sample: integrating the person-based approach and patient and public involvement. Health Expect 2021 Apr;24(2):327-340 [FREE Full text] [doi: 10.1111/hex.13173] [Medline: 33316120]
49. Morgan H, Thomson G, Crossland N, Dykes F, Hoddinott P, 'BIBS' study team. Combining PPI with qualitative research to engage 'harder-to-reach' populations: service user groups as co-applicants on a platform study for a trial. Res Involv Engagem 2016 Mar 24;2:7 [FREE Full text] [doi: 10.1186/s40900-016-0023-1] [Medline: 29062508]
50. Nuffield Foundation homepage. Nuffield Foundation. URL: https://www.nuffieldfoundation.org/ [accessed 2024-09-23]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
GRIPP2: Guidance for Reporting Involvement of Patients and the Public
NHS: National Health Service
PPIE: patient and public involvement and engagement
SDE: Secure Data Environment

Edited by K Williams; submitted 22.09.23; peer-reviewed by C Marten, G Humphreys, C Laranjeira; comments to author 26.01.24; revised version received 26.06.24; accepted 09.07.24; published 15.10.24.

Please cite as:
Yusuf ZK, Dixon WG, Sharp C, Cook L, Holm S, Sanders C
Building and Sustaining Public Trust in Health Data Sharing for Musculoskeletal Research: Semistructured Interview and Focus Group Study
J Med Internet Res 2024;26:e53024
URL: https://www.jmir.org/2024/1/e53024
doi:10.2196/53024
PMID:

©Zainab K Yusuf, William G Dixon, Charlotte Sharp, Louise Cook, Søren Holm, Caroline Sanders. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Health Record Protection Beyond the Health Insurance Portability and Accountability Act: Mixed Methods Study

Hemang Subramanian[1*], BE, EMBA, MS, PhD; Arijit Sengupta[1*], BTECH, MS, PhD; Yilin Xu[1*], BS, MS

Florida International University, Miami, FL, United States

[*]all authors contributed equally

Corresponding Author:
Hemang Subramanian, BE, EMBA, MS, PhD
Florida International University
11200 SW 8th Street
Miami, FL, 33199-2156
United States
Phone: 1 3053488446
Email: hsubrama@fiu.edu

Abstract

Background: The security and privacy of health care information are crucial for maintaining the societal value of health care as a public good. However, governance over electronic health care data has proven inefficient, despite robust enforcement efforts. Both federal (HIPAA [Health Insurance Portability and Accountability Act]) and state regulations, along with the ombudsman rule, have not effectively reduced the frequency or impact of data breaches in the US health care system. While legal frameworks have bolstered data security, recent years have seen a concerning increase in breach incidents. This paper investigates common breach types and proposes best practices derived from the data as potential solutions.

Objective: The primary aim of this study is to analyze health care and hospital breach data, comparing it against HIPAA compliance levels across states (spatial analysis) and the impact of the Omnibus Rule over time (temporal analysis). The goal is to establish guidelines for best practices in handling sensitive information within hospitals and clinical environments.

Methods: The study used data from the Department of Health and Human Services on reported breaches, assessing the severity and impact of each breach type. We then analyzed secondary data to examine whether HIPAA's storage and retention rule amendments have influenced security and privacy incidents across all 50 states. Finally, we conducted a qualitative analysis of textual data from vulnerability and breach reports to identify actionable best practices for health care settings.

Results: Our findings indicate that hacking or IT incidents have the most significant impact on the number of individuals affected, highlighting this as a primary breach category. The overall difference-in-differences trend reveals no significant reduction in breach rates (P=.50), despite state-level regulations exceeding HIPAA requirements and the introduction of the ombudsman rule. This persistence in breach trends implies that even strengthened protections and additional guidelines have not effectively curbed the rising number of affected individuals. Through qualitative analysis, we identified 15 unique values and associated best practices from industry standards.

Conclusions: Combining quantitative and qualitative insights, we propose the "SecureSphere framework" to enhance data security in health care institutions. This framework presents key security values structured in concentric circles: core values at the center and peripheral values around them. The core values include employee management, policy, procedures, and IT management. Peripheral values encompass the remaining security attributes that support these core elements. This structured approach provides a comprehensive security strategy for protecting patient health information and is designed to help health care organizations develop sustainable practices for data security.

(J Med Internet Res 2024;26:e59674) doi:10.2196/59674

KEYWORDS

security; privacy; security breach; breach report; health care; health care infrastructure; regulatory; law enforcement; Omnibus Rule; qualitative analysis; AI-generated data; artificial intelligence; difference-in-differences; best practice; data privacy; safe practice

XSL•FO
RenderX

Introduction

Overview

The HIPAA (Health Insurance Portability and Accountability Act) [1] was enacted by the United States Congress aiming to develop a set of guidelines by which personally identifiable information maintained by health care providers and insurers is protected from fraud, theft, and unauthorized access. At the most general level, HIPAA prevents health care providers from disclosing such information to anyone other than the patient herself and their health providers without explicit consent from the patient. Since its enactment in 1996, it has been implemented in all states of the United States and is widely accepted by both patients and medical providers [2]. HIPAA is now a globally accepted public health law that governs the protection and security of patient health information. Many countries globally also implement various aspects of HIPAA compliance as a measure of security and protection of personal health record (PHR) data [3].

Presently, HIPAA is enforced, and failures to adhere to the act may result in civil and criminal penalties. While the enactment of HIPAA has certainly increased the privacy of patient information, the attempts at subverting the protection and illegally accessing private information have also increased multiple fold. In this paper, we analyze a publicly available dataset related to data breaches in health care systems with both quantitative and qualitative methods to decide how effective HIPAA has been and what current health care systems may need to do to protect individuals from such breaches.

In HIPAA, Privacy and Security Rules work together to protect the privacy and security of individuals' PHR information. The Privacy Rule focuses on limiting how PHR can be used and disclosed, while the Security Rule focuses on protecting the confidentiality, integrity, and availability of electronic protected health information through technical and administrative safeguards. Both rules are enforced by the Department of Health and Human Services (HHS) and aim to protect individuals' health information from unauthorized access, use, and disclosure. Table 1 compares the HIPAA Privacy Rule with the HIPAA Security Rule with respect to the most common attributes that concern health care data protection.

Table 1. Comparing the HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule and HIPAA Security Rule across various features of the rule.

Features	HIPAA Privacy Rule	HIPAA Security Rule
Purpose	Protects the privacy of an individual's health information	Protects the security of ePHI[a]
Covered entities	Health plans, health care clearinghouses, and health care providers	Covered entities and their business associates
Privacy standards	Limits how covered entities can use and disclose PHI[b]	Ensures confidentiality, integrity, and availability of ePHI
Required measures	Covered entities must have policies and procedures in place to protect PHI	Covered entities must implement administrative, physical, and technical safeguards to protect ePHI
Access controls	Covered entities must limit access to PHI to authorized individuals	Covered entities must implement technical policies and procedures to control access to ePHI
Data encryption	Required only for ePHI transmitted over an open network	Required for all ePHI, both in transit and for storage of data
Breach notification	Covered entities must report breaches of PHI to affected individuals and the HHS[c]	Covered entities must notify affected individuals, HHS, and the media (for large breaches) of any breach of unsecured ePHI
Enforcement	Enforced by the HHS OCR[d]	Enforced by the HHS OCR and the HHS Office of the Inspector General

[a]ePHI: electronic protected health information.

[b]PHI: protected health information.

[c]HHS: Department of Health and Human Services.

[d]OCR: Office for Civil Rights.

Despite stringent HIPAA rules affecting both the privacy and security of PHR data and their implementation, we observe an increase in attack surface and loss of PHR across US hospital systems. In this paper, using health vulnerability data from hospital systems over 15 years, we study the following research question: What are the best practices to protect the security and privacy of PHR data in hospital information systems in addition to HIPAA, state-specific security guidelines, and Omnibus regulations?

To answer the above research question, we use an extensive dataset of health data breaches from January 2009 to January 2023 published by the US HHS [4]. First, we analyze different health data system vulnerabilities by finding the correlation between the number of users affected and the various parameters and types of vulnerabilities. We further look at the different types of vulnerabilities, create clusters of states based on the extent of HIPAA adoption, and analyze vulnerabilities between these clusters. We then introduce the Omnibus Rule as an exogenous shock to see the effect of this rule on the breaches. Finally, we create a cluster of the types of vulnerabilities based on a qualitative analysis of the breach information description to identify the security "values" that are important to protect

XSL•FO
RenderX

health care systems, which should provide guidance to future protection systems.

In summary, the aim of this paper is to analyze the data from existing health care security breaches and cross-compare the information against the extent of HIPAA adoption across the states (spatial) as well as the effect of the Omnibus Rule (temporal) to provide key health care security guidelines.

Background and Literature Review

Much of the health care privacy and security literature falls in the realm of PHRs, specifically electronic health records (EHRs). Unlike other security-intensive information such as banking and taxation, health records have more transparency and must be viewed by and shared with multiple health personnel. Therefore, privacy and security are important drivers in adopting health data, particularly PHRs. A 2008 Markle Foundation survey [5] of 1580 adults reported privacy concerns related to the misuse of personal data by marketers (77%), employers (56%), and insurers (53%). While the report states that American people overwhelmingly believe that EHRs can improve their health, most participants were concerned about privacy factors. The main areas of concern were identity theft and limitation of employment opportunities due to specific health conditions and hence preferred to control who could access their health information.

Other studies also concur with the above findings. A study by Keselman et al [6] specifically mentions that security in PHR access must be configured to support privacy and security without violating HIPAA requirements. In another focus group study by Kerns et al [7], participants strongly opposed commercially developed PHRs and cited issues with privacy, security, and accuracy as a deterrent to adoption. Several studies identify potential security concerns related to data sharing, tampering, improper use, and illegal tracking as well as manipulation by third parties [8,9]. Despite the security concerns, new patients mostly prefer to engage electronically for their PHRs [6,7] so enhancing the security of the systems and databases that are used for the storage and transmission of such data has become a priority.

Despite the potential concerns among users about having highly personal health information in shared electronic infrastructure, having massive amounts of information of health information allows health providers to better identify and manage high-risk and high-cost patients, thereby benefiting both the health providers as well as patients [10]. This has been made using clinical analytics—big data techniques to reduce the costs of health care by identifying high-cost patients, readmissions, triage, decompensation, adverse events, and treatment optimization for diseases affecting multiple organ systems. Such techniques require careful aggregation and data integrity techniques to ensure that all associations are privacy-preserving and not subject to inappropriate tampering [11].

While the HIPAA regulations do provide the possibility of comparative effectiveness studies, they also introduce barriers including inconsistent institutional review board policies and complicated and costly procedures to obtain the consent of patients for release of their information. This necessitates the development of a new policy framework that will allow and encourage the use of health information in all forms—fully identifiable, partially anonymized, and deidentified [12].

Development of a new framework around providing security to health information will require the reduction in ambiguity and confusion surrounding the concept of “health security,” particularly when combining information from diverse global sources. To counteract this, the global public health community will need to develop a common understanding of the security protocols that involve stakeholders in developing countries, industrialized countries, the humanitarian community, as well as military organizations [13].

Multiple techniques have been proposed in the literature for securing and preserving the anonymity and privacy of EHR clients’ symmetric and asymmetric keys for deidentification [14]. Access control, encryption, and auditing tools have also been proposed and developed [15]. Health care organizations are also developing privacy architecture using a combination of enterprise data warehouse and a software intelligence and analytics layer to counteract these threats. One such system developed by Houston Methodist Hospital [16] uses methods like specialized derecognizable proof of information, limited information access, and security controls in the hidden specialized stages to safeguard patient protection. Many other current efforts also use different forms of anonymization and encryption to preserve security and privacy [17]. The most prominent other features that are essential in such privacy preservation include system and application access control, compliance with security requirements, interoperability, integration and sharing, consent and choice mechanisms, policies and regulation, applicability and scalability, and cryptography techniques [18].

Prior research by Peddicord et al [12] similarly identifies barriers to research initiatives imposed by HIPAA and provides a policy-based framework as a collaborative mechanism for creating a process of managed consent and development of a research safe harbor to protect the privacy of health information. Similar governance frameworks have also been reviewed in general public and government sector implementations [19]. Having partnerships between the public and private sectors in such governance is important to ensure such reforms are successful in any information and communication technology solutions [20].

Other comparable techniques used in the literature for maintaining integrity in big data in EHR include blockchain-based solutions [11,21]. Privacy issues surrounding health data have become increasingly prevalent in recent years, particularly with the rise of EHRs and other digital health technologies.

The sensitive nature of health data makes it particularly vulnerable to privacy breaches, which can have serious consequences for individuals, health care providers, and health care systems. Recent studies have found that blockchain-based EHRs can give patients greater control over their data while ensuring their information is kept secure and confidential. Similarly, blockchains can be used to create a secure and private

XSL•FO
RenderX

platform for sharing medical data among health care providers without compromising patient privacy [19].

HIPAA Privacy Rule and Security Rule

To protect health data across multiple sources and stakeholders, various safeguards have been researched in the literature. These methods include physical safeguards such as physical locks and other access control methods for computers and other equipment; technical safeguards such as encryption, firewalls, strong passwords, antivirus and antimalware systems, mobile agents, and dual authentication schemes; and administrative safeguards such as risk assessments, personal device use policies, and deidentification of samples [22].

Finally, there is literature on the agreement of multiple states or stakeholders toward the development of a collaboration agreement, and harmonizing the state privacy laws and measures would be a vital component of the implementation of these [23]—our work does provide an analytical view of how the implementation of HIPAA across state barriers makes a difference in the effect of data breaches, and our analysis suggests that states with higher protection level than HIPAA are typically far better in terms of the effect any data breaches may have.

Creation of the Omnibus HIPAA Rule

On January 17, 2013, the HHS released the long-awaited "Omnibus Rule." The Omnibus Rule implemented most of the privacy and security provisions of the Health Information Technology for Economic and Clinical Health Act. The rule made the following changes to the security and privacy of health data providers and health systems: (1) extended the reach and limits of HIPAA with respect to obligations to business associates and their subcontractors, (2) modified the breach notification standard and expanded patient rights to access and restrict disclosure of protected health information (PHI), (3) imposed new rules governing the uses and disclosures of PHI, (4) clarified enforcement approaches, and (5) addressed obligations under the Genetic Information Nondiscrimination Act of 2008.

Of these changes implemented by the HHS, rules 1-4 impact the security and privacy of user data stored on health care information systems and are expected to significantly improve it. This significant change enabled several covered entities (such as hospitals, partners, and clinics) to analyze gaps in their security practice and perform the following services: (1) revising privacy and security policies and procedures; (2) revising breach notification policies, procedures, and breach response plans; (3) amending notices of privacy practices (and making sure the revised notices are properly posted and distributed); (4) training the workforce and promoting more ongoing awareness about privacy and security policies; (5) revising business associate contract templates and process of amending or renegotiating each one within the framework of the Omnibus Rule; (6) determining whether any forms, such as requests for access, should be updated or created; (7) continuing—or making an increased effort—to take advantage of the safe harbor provision by encrypting PHI according to HHS' guidance; and (8) making sure an updated risk analysis is in place and reflects vulnerabilities addressed in HHS guidance, such as mobile devices.

Methods

Ethical Considerations

Ethics approval is not required for this study, as human participants were not involved.

Study Design

First, we used data from HHS breaches and analyzed the severity of breach types and the impact of those breaches. Second, using secondary data and publicly available information, we tried to analyze whether data storage and retention changes to rules of HIPAA played any role in reducing security and privacy violations across the 50 US states. Finally, we used the textual data present in the vulnerability or breach data to analyze and recommend best practices for health care clinics and hospitals by doing a detailed qualitative analysis. We believe that hospital systems can alleviate the major concerns facing their health care systems overall.

The primary dataset used in this study was HHS breach report data from 2009 until January 2023 [4]. Using this dataset, the different types of data breaches were analyzed by analyzing their statistics. We further developed a model that studies the extent of the effect of a particular type of data breach and the size of such an effect on the user. We supplemented these data with data about 50 US states from the Report on Adoption of HIPAA from HealthInfoLaw.org. We analyzed 2 major events in our sample. First, we used a difference-in-differences analysis to examine the presence of state-specific health data storage rules that are more stringent than the current HIPAA regulations, in order to determine how states with only HIPAA compared to those with stricter rules. Second, we analyzed the impact of the Omnibus Rule change, in relation to state-level HIPAA stringency, on the overall number of customers. Further, based on our analysis of breach descriptions, we determined the severity of breach categories to determine the key security "values" that guide future security practices. The sequence of analyses performed in this paper is summarized in the research design flow diagram of Figure 1.

XSL•FO
RenderX

Figure 1. Research design flow. AI: artificial intelligence; HHS: Department of Health and Human Services; HIPAA: Health Insurance Portability and Accountability Act.

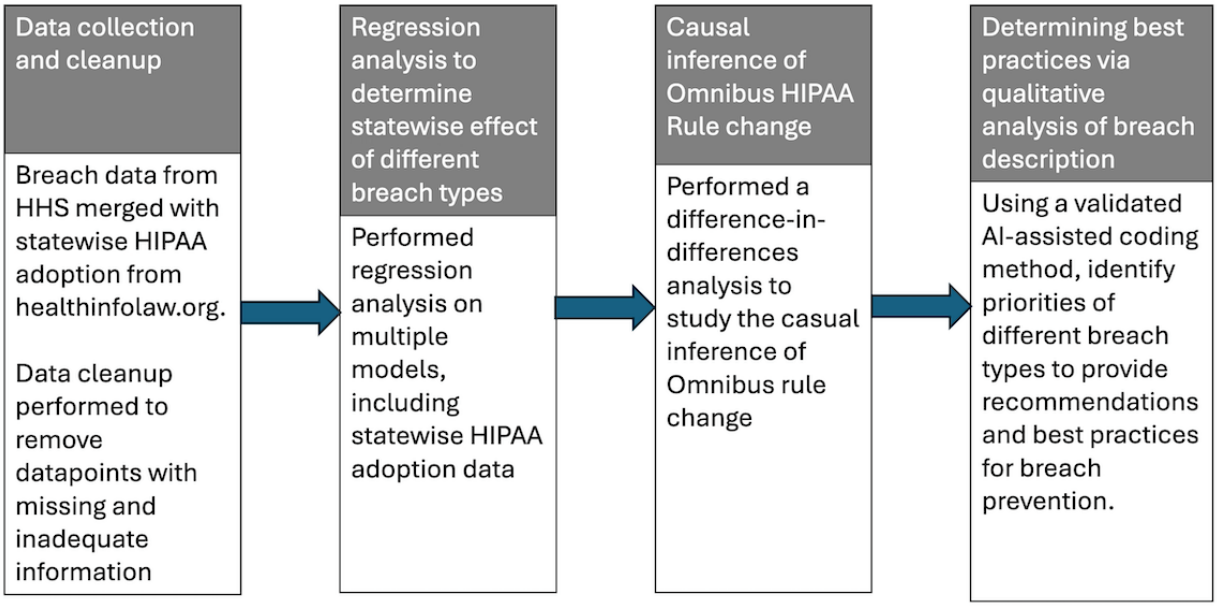

The final step was to perform a ground-up qualitative analysis of the breach descriptions provided in the dataset to classify distinct topics within the dataset. Each code in this analysis represents a portion of data that are tagged as relevant to one of several predetermined categories. Each category represents a specific aspect of data breach reports that helps organize the data to make it easier to understand patterns, correlations, and trends. The resulting categorization of the codes thus represents research themes that determine the key security "measures" that can lead to best practices for identifying and preventing future breaches.

With over 3000 data points in the dataset, we faced a potential research dilemma—performing a qualitative analysis on such a large dataset could potentially be time- and resource-intensive. To automate this process, we used ATLAS.ti, a well-known research coding tool that uses machine learning algorithms to analyze qualitative data and identify patterns and themes. Based on the information provided by ATLAS.ti, the system internally used ChatGPT 3.5 developed by OpenAI as the model to conduct fully automated inductive coding. The motivation to apply the technique instead of manual coding is that this process of semiautomated coding can save time and potentially improve the consistency of the codes since the dataset is too large to code manually. However, to ensure that the codes generated by the system are on par with human coding, one of the researchers manually coded a small subset (n=50) of the breach descriptions using NVivo (version 14; Lumivero), followed by performing the automated coding on the same dataset by ATLAS.ti. An interrater reliability analysis was performed on the resulting codes. The Cohen κ score for this analysis was 0.88, indicating a reliable coding performance from ATLAS.ti. We posited, then, that using ATLAS.ti for the qualitative analysis would be sufficient for the entire dataset. A complete comparison analysis is included in Multimedia Appendix 1.

To be more specific, the codes were created by ATLAS.ti using large language models from the GPT family with the foundation of vast amounts of different texts and additional training by the researchers. During the procedure, the system split the text from the dataset into smaller chunks and sent these data to the GPT models for repetition analysis. The results were then combined algorithmically to generate an optimal mix of codes including various topics while avoiding too many redundant codes.

Serving as the application of large language models, GPT is sensitive to the surrounding text as the context to have a window of attention that ensures that the size of the data does not overflow the model. Thus, the contexts generated are not greater than 100 characters long and are broken up at natural paragraph boundaries. During the coding process, the paragraphs that appear more than 1 time were excluded from the data as inputs. Next, the large language model separated out the context-specific codes and assigned them to a category based on "context," "sentiment," "meaning," and such similar features, as documented in the GPT-4.

Results

Summary of Data

Table 2 depicts the breach types that are common along with the type of incident shown. From this visualization, we can see that hacking or IT incidents are the types of breaches that have the highest impact in terms of the proportion of breaches. When considering the number of individuals affected, we see an even higher impact of hacking and IT-related incidents, which indicates hacking to be a continual issue in health care data privacy.

XSL•FO
RenderX

Table 2. Incidents and individuals by types of breaches.

Type of breach	Incidents (n=5158), n (%)	Individuals affected (n=382,533,378), n (%)
Hacking or IT incident	2465 (47.7)	316,771,372 (82.8)
Unauthorized access or disclosure	1268 (24.5)	24,074,273 (6.2)
Theft	999 (19.3)	26,876,573 (7)
Loss	220 (4.2)	9,449,646 (2.4)
Improper disposal	110 (2.1)	2,093,033 (0.5)
Other	84 (1.6)	1,188,272 (0.3)
Unknown	12 (0.2)	2,080,209 (0.5)

Is HIPAA Implementation Alone Sufficient to Reduce Data Breaches?

We proposed a model with statewise fixed effects to analyze the role of the type of incidents and the location of the incident on the economic value of the affected party. We then used the number of affected individuals as the dependent variable indicating the economic effect of the breaches. Further, we used the health classification of states as either being HIPAA compliant or having local laws that are more stringent than HIPAA, which affect the number of individuals affected by the breach. The following equation specifies the model:

$$ln(Individuals\ affected) = \alpha + \beta_{1\text{-}5}\ Type\ of\ breach\ indicator + \beta_{6\text{-}9}\ C.E.\ type + \beta_{10\text{-}60}\ State\ indicator + \beta_{61}\ StHIPAA$$

Table 3 contains 3 primary models, model 1, model 2, and model 3, which incrementally add variables (Multimedia Appendix 2). Model 1 is the baseline model. Model 2 adds additional covariates of health care provider, health care clearinghouse, and health plan. Model 3 includes the indicator stronger than HIPAA variable, which indicates whether the corresponding state-enabled rules are stronger than HIPAA.

XSL•FO
RenderX

Table 3. Regression models display the effect of the health care breach type on individuals affected[a].

Dependent variable=ln(Individuals affected)	Model 1		Model 2		Model 3	
	β	*P* value	β	*P* value	β	*P* value
Hacking or IT incident	Included	—[b]	Included	—	Included	—
Improper disposal	–1.162[c]	<.001	–1.176[c]	<.001	–1.176[c]	<.001
Loss	–1.342[c]	<.001	–1.339[c]	<.001	–1.339[c]	<.001
Other	–1.014[c]	<.001	–1.085[c]	<.001	–1.085[c]	<.001
Theft	–1.168[c]	<.001	–1.218[c]	<.001	–1.218[c]	<.001
Unauthorized access or disclosure	–1.336[c]	<.001	–1.362[c]	<.001	–1.362[c]	<.001
Unknown	–1.249[d]	.02	–1.291[d]	.02	–1.291[d]	.02
Business associate	Included	—	Included	—	Included	—
Health plan	—	—	–0.382[c]	<.001	–0.382[c]	<.001
Health care clearing house	—	—	–0.393	.49	–0.393	.49
Health care provider	—	—	–0.411[c]	<.001	–0.411[c]	<.001
Stagewise fixed effect	Included	—	Included	—	Included	—
StHIPAA[e]	—	—		—	1.271[d]	.03
Constant	9.144[c]	<.001	8.743[c]	<.001	8.743[c]	<.001
Observations	5158	—	5140	—	5140	—
R^2	0.12	—	0.14	—	0.14	—

[a]Breach report is present in Multimedia Appendix 2.

[b]Not applicable.

[c]*P*<.001.

[d]*P*<.05.

[e]StHIPAA: stronger than Health Insurance Portability and Accountability Act.

From these results, we see that each of the separate types of vulnerability (ie, hacking or IT incident: β=9.144, improper disposal: β=–1.139, loss: β=–1.085, other: β=–1.218, theft: β=–1.362, and unauthorized access or disclosure: β=–1.291) significantly affect *ln(Individuals affected)*. From these, we see that the hacking or IT incident affects the number of individuals affected the highest, followed by loss of data, improper disposal, other types of hacking incidents, and theft of data. Overall, while different types of vulnerabilities affect and influence data uniquely, it is essential to observe that hacking or IT incident, which is the baseline category, has the highest effect on the number of individuals affected.

Similarly, the states that have laws that are significantly more stringent than HIPAA seem to be affected more by these different types of vulnerabilities (ie, stronger than HIPAA: β=1.271). Our results indicate that the higher the number of hacking and related incidents, the higher the chances for individuals being affected, and as a result, the most we need to protect against those types of attacks.

Effect of the Omnibus Rule Amendment to HIPAA on Breaches Across US States

The 2013 Omnibus Rule offers an opportunity to examine how the changes implemented on January 17, 2013—captured by a variable where time=0 represents the pre hoc period and time=1 represents the post hoc period—affect the number of individuals impacted in both the treated states (ie, states with regulations stronger than HIPAA) and states adhering solely to HIPAA standards. Building upon models 1, 2, and 3, we use a difference-in-differences model to estimate how the "Omnibus" Rule affected *ln(Individuals affected)* in both the treatment and control groups. Our difference-in-differences econometric estimation is described in Table 4.

XSL•FO
RenderX

Table 4. The result of the difference-in-differences estimation.

	Model 1	*P* value[a]
Hacking or IT incident	Included	—[b]
Improper disposal	–1.170[c]	<.001
Loss	–1.330[c]	<.001
Other	–1.074[c]	<.001
Theft	–1.205[c]	<.001
Unauthorized access or disclosure	–1.361[c]	<.001
Unknown	–1.267[d]	.02
Business associate	Included	—
Health plan	–0.382[c]	<.001
Health care clearinghouse	–0.388	.50
Health care provider	–0.412[c]	<.001
Treated time (difference-in-differences effect)	0.0828	.50
Constant	8.741[c]	<.001
Observations	5140	—
R^2	0.14	—

[a]Two-tailed *t* test.

[b]Not applicable.

From our analysis, we observe that the overall difference-in-differences trend is not significant (coefficient=0.0828; *P*=.50). This demonstrates that despite laws being stronger than HIPAA in several states and the Omnibus Rule being issued, there are no guarantees that there will be a decrease overall, showing that despite additional guidelines and strengthening of protections, the number of individuals affected and breach trends continue to increase overall. Therefore, we can conclude that although HIPAA, state-specific stringent security rules, and Omnibus Rules provide guidelines, they are not sufficient on their own to prevent data breaches. Separate hospital-specific and breach-specific guidelines must also be followed.

Qualitative Analysis Results

We performed a qualitative coding procedure on the description of each of the 4000+ data points using the artificial intelligence (AI)–assisted coding process. This method produces a substantial number of codes and likely more than what is perceived as useful data. To solve the problem, we used embeddings, a feature of language models for all the words or phrases that correspond to a point within multidimensional space. Using this technique, initially, similar words or phrases were clustered together. The system then repeatedly combined codes nearest to each other into a collection as clusters until an expected number of collections are achieved. For enhanced rigor, we developed and analyzed clusters of 9, 10, 11, 12, and 15 separate collections and manually examined the unique level 2 codes belonging to each collection for similarity and overlap. Two authors (HS and YX) separately analyzed the collections of codes and found that the 15 chosen categories had the lowest overlap and similarity among the choices above. We finalized 15 chosen categories, which are displayed in Table 5.

XSL•FO
RenderX

Table 5. Final set of categories with their weights from the coding analysis.

Type of breach	Weight (n=8517), n (%)
Data breach	2533 (30)
Security	1350 (16)
Privacy	1292 (15)
Protected health information	726 (9)
Communication	632 (7)
Regulatory compliance	355 (4)
Security breach	355 (4)
Risk management	305 (4)
Data privacy	234 (3)
Employee management	192 (2)
Access controls	177 (2)
IT	136 (2)
Policies	121 (1)
Procedures	71 (1)
Electronic communication	38 (0)

Table 5 shows the distribution of the generated codes in 15 categories representing the cluster of similar codes obtained from the analysis. From these results, we observed that the combination of data breach, security, and privacy accounts for more than 60% of the breach reports' description, validating our original quantitative results.

Figure 2 shows screen captures illustrating the process of category generation as embeddings in the ATLAS.ti user interface. Specifically, the larger neon green circles represent the categories, while the smaller green dots connected to them represent the relevant codes with similar characteristics.

XSL•FO
RenderX

Figure 2. (A) Generated codes and clusters and (B) the list of codes and categories generated.

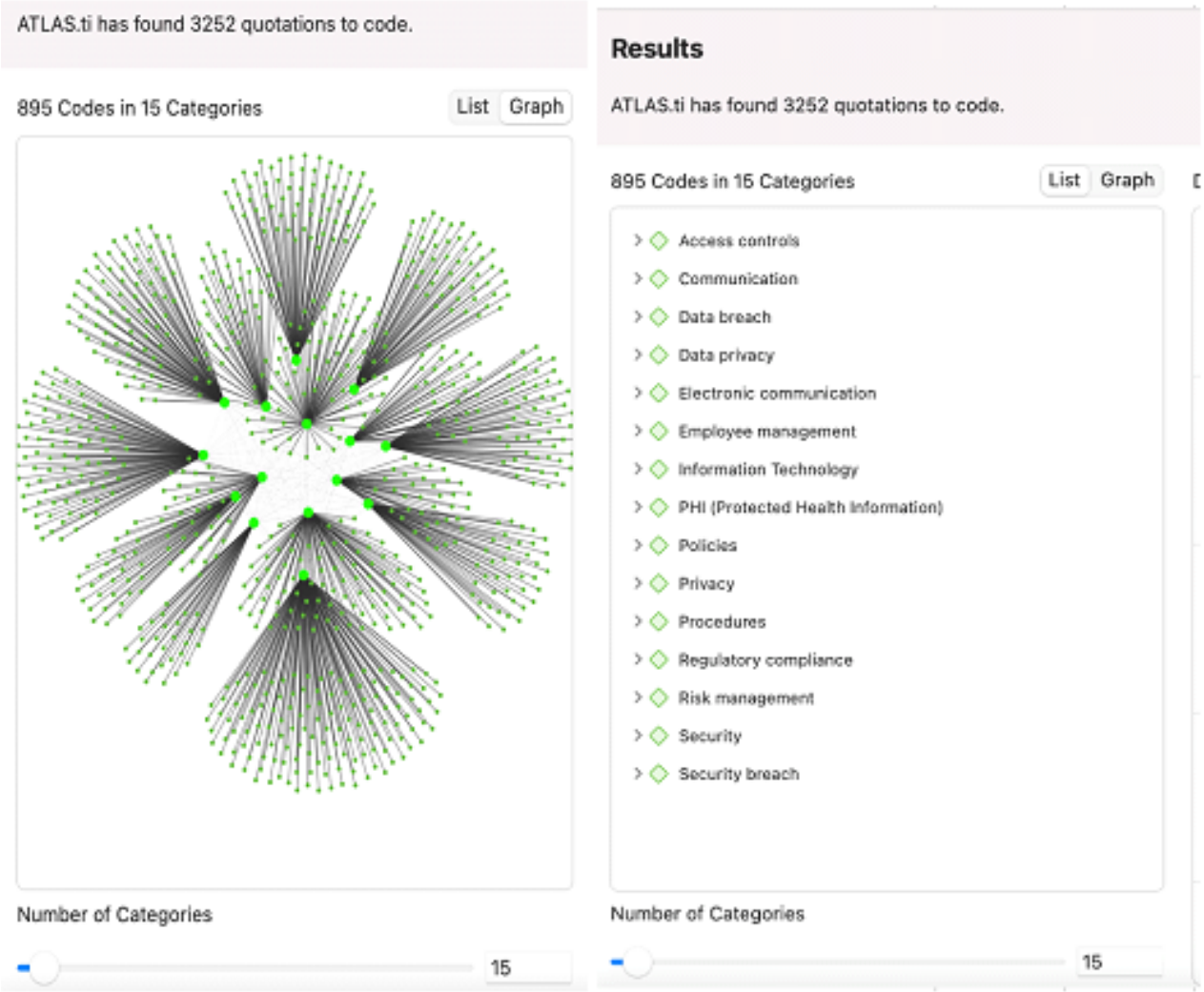

We thus see 895 different codes created within the set of 15 categories. These categories correspond to the 15 underlying values based on the description of the breach report in the dataset: access controls, communication, data breach, data privacy, electronic communication, employee management, IT, PHI, policies, privacy, procedures, regulatory compliance, risk management, security, and security breach. By mapping these values onto their respective best practices as recommendations, we listed the definition of each core security-breach value as well as how organizations use these recommendations as exemplars in Multimedia Appendix 3 [24-85].

Based on the AI-generated summary from ATLAS.ti from the dataset in the health care industry, we find that the health care industry has faced numerous breaches of PHI, affecting a range of organizations such as hospitals, dental offices, insurance companies, and medical clinics. As a result, these breaches have led to unauthorized access, theft, and disclosure of sensitive personal and medical information through various means, including stolen laptops and physical documents, email errors, employee misconduct, and improper disposal of documents. In response to these incidents, investigations have been conducted by the Office for Civil Rights (OCR), and fines were imposed. Corrective actions have included enhancing data security measures, implementing encryption, conducting risk analyses, and providing breach notifications and credit monitoring to affected individuals. These breaches underscore the critical importance of safeguarding patient data, complying with HIPAA regulations, and continuously strengthening security protocols in the health care sector to prevent future incidents.

Discussion

Overview

Based on our quantitative and qualitative analyses of the breach data, we now discuss a further theoretical contribution of our paper that we call SecureSphere, a structured framework for health care security, based on the security values derived earlier.

The SecureSphere Framework for Health Care Data System Security

The final step in our research is to classify the most important values that prevent large value data breaches overall and present the SecureSphere framework (Figure 3). The SecureSphere framework recommends a specific scope of action for the key security “values” and generic actions for preventing and sustaining data system security at hospitals and other health institutions that store, retrieve, and manage PHI.

XSL•FO
RenderX

Figure 3. SecureSphere framework for Hospital Data Breach Protection. The inner circle represents the core values that guide and represent 80% of the breaches. The outer circle represents the peripheral values that form key business practices that influence the inner circle. PHI: protected health information.

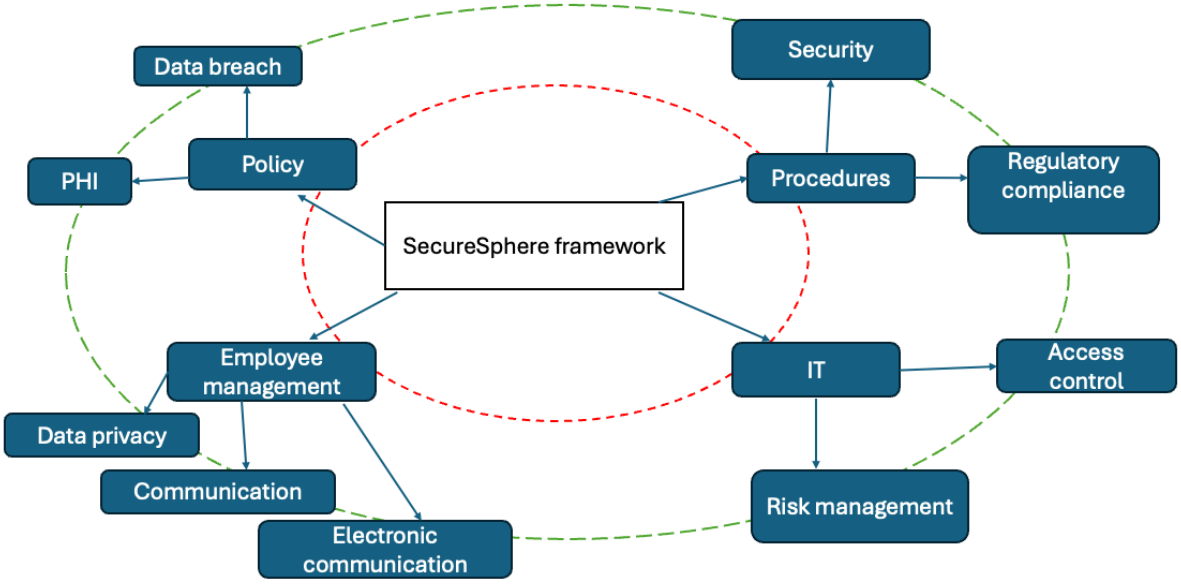

We represent values in concentric circles. The innermost circle represents the core values that include employee management, policy, procedures, and IT. The outer circle represents peripheral values associated with each core value. Thus, the framework demonstrates how data breach and PHI influence policy, security, and regulatory compliance influence procedures within the hospital. Then, we have access control and risk management that influence IT procedures within a company. Finally, we have email communication, general communication, and data handling (privacy), which influence employee management.

While the core focal values are essential focus areas of health care companies, the outer values form the essential components within these focus areas. The SecureSphere framework identifies key security values and categorizes them into concentric circles to emphasize their influence on hospital data system security. The best practices by value category are summarized in Textboxes 1-3.

XSL•FO
RenderX

Textbox 1. Core focal values (inner circle).

Employee management

- Training and awareness: Regular training programs to educate employees on security protocols and best practices.
- Access control: Implementing strict access controls to ensure that only authorized personnel can access sensitive data.
- Monitoring and auditing: Continuous monitoring of employee activities and regular audits to detect and prevent unauthorized actions.

Policy

- Comprehensive security policies: Developing clear and comprehensive security policies that outline acceptable use, data protection, and incident response.
- Policy enforcement: Ensuring strict enforcement of security policies through regular checks and penalties for noncompliance.
- Policy review and update: Regularly reviewing and updating policies to adapt to new threats and regulatory changes.

Procedures

- Standard operating procedures (SOPs): Creating detailed SOPs for handling data, responding to incidents, and managing IT infrastructure.
- Incident response procedures: Establishing clear procedures for responding to security incidents, including containment, investigation, and remediation.
- Regular audits: Conducting regular audits to ensure adherence to established procedures and identify areas for improvement.

IT

- IT infrastructure security: Implementing robust security measures for IT infrastructure, including firewalls, intrusion detection systems, and antivirus software.
- System hardening: Regularly updating and patching systems to protect against vulnerabilities.
- Incident response: Establishing an incident response team to quickly address and mitigate security breaches.

Textbox 2. Peripheral values (outer circle).

Data breach and protected health information (PHI)

- Data encryption: Encrypting sensitive data to protect it from unauthorized access during transmission and storage.
- Data minimization: Collecting only the minimum necessary data to reduce the risk of breaches.
- PHI handling: Ensuring that PHI is handled in compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act).

Regulatory compliance

- Compliance monitoring: Regularly monitoring compliance with regulations such as HIPAA, general data protection regulation, and others.
- Compliance training: Providing ongoing training to employees on regulatory requirements and compliance practices.
- Reporting and documentation: Maintaining thorough documentation and reporting mechanisms to demonstrate compliance.

Security

- Layered security approach: Implementing multiple layers of security controls to protect data at various levels.
- Risk assessments: Conducting regular risk assessments to identify and mitigate potential security threats.
- Security best practices: Adopting industry best practices for securing networks, systems, and data.

Procedures (influenced by policy, security, and regulatory compliance)

- Regular procedure updates: Continuously updating procedures to reflect changes in policy, security requirements, and regulatory standards.
- Procedure training: Ensuring all employees are trained on the latest procedures and understand their roles in maintaining security.

XSL•FO
RenderX

Textbox 3. Additional influences.

Access control and risk management (influencing IT procedures)

- Role-based access control (RBAC): Implementing RBAC to ensure users have access only to the information necessary for their role.
- Risk management strategies: Developing and implementing strategies to identify, assess, and mitigate risks.

Email communication, general communication, and data handling (influencing employee management)

- Secure email practices: Using encryption and secure email gateways to protect communication.
- Communication policies: Establishing clear policies for secure communication, both within and outside the organization.
- Data handling best practices: Implementing best practices for data handling, including secure storage, transmission, and disposal.

By following these best practices, hospitals can effectively enhance their data system security, minimize the risk of breaches, and ensure compliance with relevant regulations.

Conclusions

This paper uses the HIPAA privacy and security rules and analyzes various types of breaches in the US health care system. Though HIPAA rules are stringent, they are not sufficient to prevent some of the basic types of vulnerabilities that influence and affect the major players overall. Our comprehensive analysis of the HIPAA privacy and security rules in conjunction with a detailed study of various types of breaches within the US health care system reveals a complex landscape. Despite the robust framework provided by HIPAA and subsequent amendments like the Omnibus Rule, breaches continue to occur with alarming frequency and impact. The persistence of these breaches underscores the necessity for continuous reassessment and enhancement of security protocols.

Our empirical analysis highlights that while regulatory frameworks form a critical backbone for data protection, they must be dynamically updated and rigorously enforced to adapt to evolving technological landscapes and sophisticated cyber threats. The addition of stringent state laws appears to correlate with better outcomes in some cases, suggesting that a multilayered approach to regulation may be beneficial.

Moreover, our findings advocate for the importance of fostering a culture of security within health care organizations that transcends compliance. Investing in cutting-edge technologies, training staff on a continuous basis, and taking a proactive stance on security are imperative.

The results of our detailed qualitative analysis of all text data provide support for 15 core values related to security breach management. These categories include access controls, communication, data breach, data privacy, electronic communication, employee management, IT, PHI, policies, privacy, procedures, regulatory compliance, risk management, security, and security breach. For each category, specific definitions, best practice recommendations, and exemplary organizations are identified. For instance, data breach involves unintentional disclosures, with Mayo Clinic cited for best practices like ongoing employee training and encryption measures. Regulatory compliance emphasizes aligning with laws, showcased by Deloitte's ethics programs and Ernst & Young's compliance management. IT highlights reducing uncertainty in organizational processes, with Apple noted for developing AI tools. This structured approach aims to enhance understanding and application of security measures in response to the frequent and varied breaches affecting health care entities.

Finally, it is evident that to truly safeguard patient data, collaboration across various stakeholders—including health care providers, patients, technology firms, and policy makers—is crucial. By integrating insights from this extensive analysis and categorizing values into the inner core and peripheral values, we present the SecureSphere framework for the health care industry, which can better protect itself against future breaches and ensure the integrity and privacy of sensitive health information. We highlight key issues with the framework that enables health care executives to make key decisions to enable better health care IT system security.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details about human–artificial intelligence intercoder reliability with subsample of the dataset.
[DOCX File , 189 KB - jmir_v26i1e59674_app1.docx]

Multimedia Appendix 2
Breach report.
[XLS File (Microsoft Excel File), 5161 KB - jmir_v26i1e59674_app2.xls]

Multimedia Appendix 3

XSL•FO
RenderX

Values for information security practices with exemplars.
[DOCX File , 25 KB - jmir_v26i1e59674_app3.docx]

References

1. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Centers for Disease Control and Prevention, Public Health Law. 2024. URL: https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html?CDC_AAref_Val=https://www.cdc.gov/phlp/publications/topic/hipaa.html [accessed 2024-10-10]
2. Slutsman J, Kass N, McGready J, Wynia M. Health information, the HIPAA privacy rule, and health care: what do physicians think? Health Aff (Millwood) 2005;24(3):832-842. [doi: 10.1377/hlthaff.24.3.832] [Medline: 15886179]
3. Mbonihankuye S, Nkunzimana A, Ndagijimana A. Healthcare data security technology: HIPAA compliance. Wirel Commun Mob Comput 2019;2019:1-7 [FREE Full text] [doi: 10.1155/2019/1927495]
4. Breach portal: notice to the secretary of HHS breach of unsecured protected health information. US Department of Health and Human Services, Office of Civil Rights. URL: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf [accessed 2024-10-10]
5. Americans overwhelmingly believe electronic personal health records could improve their health. Markle Foundation. 2008. URL: https://www.markle.org/publications/401-americans-overwhelmingly-believe-electronic-personal-health-records-could-improve-t/ [accessed 2024-10-10]
6. Keselman A, Slaughter L, Smith CA, Kim H, Divita G, Browne A, et al. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. AMIA Annu Symp Proc 2007;2007:399-403 [FREE Full text] [Medline: 18693866]
7. Kerns JW, Krist AH, Longo DR, Kuzel AJ, Woolf SH. How patients want to engage with their personal health record: a qualitative study. BMJ Open 2013;3(7):e002931 [FREE Full text] [doi: 10.1136/bmjopen-2013-002931] [Medline: 23901027]
8. Basil NN, Ambe S, Ekhator C, Fonkem E. Health records database and inherent security concerns: a review of the literature. Cureus 2022;14(10):e30168 [FREE Full text] [doi: 10.7759/cureus.30168] [Medline: 36397924]
9. Kruse CS, Frederick B, Jacobson T, Monticone DK. Cybersecurity in healthcare: a systematic review of modern threats and trends. Technol Health Care 2017;25(1):1-10. [doi: 10.3233/THC-161263] [Medline: 27689562]
10. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33(7):1123-1131. [doi: 10.1377/hlthaff.2014.0041] [Medline: 25006137]
11. Kask M, Piho G, Ross P. Systematic literature review of methods for maintaining data integrity. In: Advances in Model and Data Engineering in the Digitalization Era MEDI 2021. Cham, Switzerland: Springer; 2021.
12. Peddicord D, Waldo AB, Boutin M, Grande T, Gutierrez L. A proposal to protect privacy of health information while accelerating comparative effectiveness research. Health Aff (Millwood) 2010;29(11):2082-2090. [doi: 10.1377/hlthaff.2010.0635] [Medline: 21041752]
13. Aldis W. Health security as a public health concept: a critical analysis. Health Policy Plan 2008;23(6):369-375. [doi: 10.1093/heapol/czn030] [Medline: 18689437]
14. Fernández-Alemán JL, Señor IC, Lozoya PAO, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
15. Jahan S, Chowdhury M, Islam R, Gao J. Security and privacy protection for eHealth data. In: Future Network Systems and Security. 2018 Presented at: 4th International Conference on Future Network Systems and Security, FNSS 2018; July 9-11, 2018; Montparnasse, France p. 197-205 URL: https://researchoutput.csu.edu.au/en/publications/security-and-privacy-protection-for-ehealth-data [doi: 10.1007/978-3-319-94421-0_16]
16. Vaishnav R, Panditi MDD, Dhiman V, Aarthy CCJ, Kumari YS, Mohiddin MK. Data security in healthcare management analysis and future prospects. Mater Today Proc 2022;51:2202-2206. [doi: 10.1016/j.matpr.2021.11.280]
17. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data 2018;5(1):1-18. [doi: 10.1186/s40537-017-0110-7]
18. Rezaeibagha F, Win KT, Susilo W. A systematic literature review on security and privacy of electronic health record systems: technical perspectives. Health Inf Manag 2015;44(3):23-38. [doi: 10.1177/183335831504400304] [Medline: 26464299]
19. Tan E, Mahula S, Crompvoets J. Blockchain governance in the public sector: a conceptual framework for public management. Gov Inf Q 2022;39(1):101625 [FREE Full text] [doi: 10.1016/j.giq.2021.101625]
20. Cordella A, Bonina CM. A public value perspective for ICT enabled public sector reforms: a theoretical reflection. Gov Inf Q 2012;29(4):512-520 [FREE Full text] [doi: 10.1016/j.giq.2012.03.004]
21. Mayer AH, da Costa CA, Righi RDR. Electronic health records in a blockchain: a systematic review. Health Informatics J 2020;26(2):1273-1288. [doi: 10.1177/1460458219866350] [Medline: 31566472]
22. Kruse CS, Smith B, Vanderlinden H, Nealand A. Security techniques for the electronic health records. J Med Syst 2017;41(8):127 [FREE Full text] [doi: 10.1007/s10916-017-0778-4] [Medline: 28733949]

XSL•FO
RenderX

23. Dimitropoulos L, Rizk S. A state-based approach to privacy and security for interoperable health information exchange. Health Aff (Millwood) 2009;28(2):428-434. [doi: 10.1377/hlthaff.28.2.428] [Medline: 19275999]
24. Cheng L, Liu F, Yao DD. Enterprise data breach: causes, challenges, prevention, and future directions. WIREs Data Min Knowl Discov 2017;7(5):e1211. [doi: 10.1002/widm.1211]
25. Avis E. Mayo Clinic's focus on staff needs, training and efficiency pays off. Health Facilities Management. 2022. URL: https://www.hfmmagazine.com/articles/4601-mayo-clinics-focus-on-staff-needs-training-and-efficiency-pays-off [accessed 2024-10-10]
26. Loughlin S. In contracts with device vendors, Mayo Clinic emphasizes security. Biomed Instrum Technol 2016;50(1):53-55. [doi: 10.2345/0899-8205-50.1.53] [Medline: 26829140]
27. Privacy policy. Mayo Clinic. 2024. URL: https://www.mayoclinic.org/about-this-site/privacy-policy [accessed 2024-10-10]
28. Rothchild E. What is security? Daedalus 1995;124(3):53-98.
29. We stop threats hiding in plain sight. Symantec. 2024. URL: https://sep.securitycloud.symantec.com/v2/landing [accessed 2024-10-10]
30. Design zone for security. Cisco. 2024. URL: https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-security/index.html#:~:text=The%20Cisco%20Design%20Zone%20for,defense%2C%20and%20other%20security%20architectures [accessed 2024-10-10]
31. Infosec. 2024. URL: https://www.infosecinstitute.com/iq/welcome/ [accessed 2024-10-10]
32. Solove DJ. Understanding Privacy. Cambridge, MA: Harvard University Press; 2010:272.
33. Compliance guidance. US Department of Health and Human Services Office of Inspector General. 2023. URL: https://oig.hhs.gov/compliance/compliance-guidance/ [accessed 2024-10-10]
34. Data privacy day at Apple: improving transparency and empowering users. Apple Newsroom. 2021. URL: https://www.apple.com/bh/newsroom/2021/01/data-privacy-day-at-apple-improving-transparency-and-empowering-users/ [accessed 2024-10-10]
35. Martinez J. Facebook unveils privacy education feature for new users. The Hill. 2012. URL: https://thehill.com/policy/technology/133497-facebook-unveils-privacy-education-feature-for-new-users/ [accessed 2024-10-10]
36. How Google anonymizes data. Google. URL: https://policies.google.com/technologies/anonymization?hl=en-US [accessed 2024-10-10]
37. Isola S, Al Khalili Y. Protected Health Information. Treasure Island, FL: StatPearls Publishing; 2024.
38. Security and privacy policies. Epic. 2023. URL: https://www.epic.com/privacypolicies/ [accessed 2024-10-10]
39. McGee MK. Cleveland Clinic: assessing risks. ISMG Network. 2013. URL: https://www.healthcareinfosecurity.com/interviews/cleveland-clinic-assessing-risks-i-1820 [accessed 2024-10-10]
40. Notice of privacy practices. US Department of Health and Human Services. 2023. URL: https://www.hhs.gov/hipaa/for-individuals/notice-privacy-practices/index.html [accessed 2024-10-10]
41. Luhmann N. What is communication? Commun Theory 1992;2(3):251-259 [FREE Full text] [doi: 10.1111/j.1468-2885.1992.tb00042.x]
42. Share without insecurity. Signal. URL: https://signal.org/#:~:text=State%2Dof%2Dthe%2Dart,%2C%20every%20call%2C%20every%20time [accessed 2024-10-10]
43. How to give your team meetings a status update. Atlassian. 2016. URL: https://www.atlassian.com/blog/teamwork/give-team-meetings-status-update [accessed 2024-10-10]
44. Learning more effective communication techniques for your communication style. LinkedIn Learning Blog. 2023. URL: https://www.linkedin.com/business/learning/blog/effective-communication-techniques-for-your-communication-style [accessed 2024-10-10]
45. Kharbili ME, Ma Q, Kelsen P, Pulvermueller E. CoReL: policy-based and model-driven regulatory compliance management. : IEEE; 2011 Presented at: 2011 IEEE 15th International Enterprise Distributed Object Computing Conference; August 29-September 2, 2011; Helsinki, Finland p. 247-256. [doi: 10.1109/edoc.2011.23]
46. Ethics and compliance programs. Deloitte. URL: https://www2.deloitte.com/br/en/pages/risk/solutions/programas-de-etica-compliance.html [accessed 2024-10-10]
47. EY regulatory compliance manager. EY. 2024. URL: https://www.ey.com/en_gl/services/consulting/regulatory-compliance-manager [accessed 2024-10-10]
48. Internal audit, risk and compliance services. KPMG. 2024. URL: https://kpmg.com/dp/en/home/services/advisory/risk-consulting/internal-audit-risk-compliance-services.html#:~:text=KPMG's%20Internal%20audit%2C%20Risk%20%26%20compliance,supply%20chain%20to%20augment%20and [accessed 2024-10-10]
49. Hovav A, Andoh-Baidoo F, Dhillion G. Classification of security breaches and their impact on the market value of firms. 2007 Presented at: Proceedings of the 6th Annual Security Conference; April 11-12, 2007; Las Vegas, NV URL: https://www.researchgate.net/profile/Anat-Hovav/publication/267790166_Classification_of_Security_Breaches_and_Their_Impact_on_the_Market_Value_of_Firms/links/597637b8458515e26d25b3ef/Classification-of-Security-Breaches-and-Their-Impact-on-the-Market-Value-of-Firms.pdf

XSL•FO
RenderX

50. Cyber incident response planning. IBM. 2021. URL: https://www.ibm.com/docs/en/ftmswsfz300?topic=program-cyber-incident-response-planning [accessed 2024-10-10]
51. Maude MMJ. A post breach analysis: Okta support unit. BeyondTrust. 2024. URL: https://www.beyondtrust.com/webinars/a-post-breach-analysis-okta-support-unit [accessed 2024-10-10]
52. Global data breach resolution. Experian. 2023. URL: https://www.experian.com/data-breach/solutions/global-data-breach-resolution [accessed 2024-10-10]
53. Power M. The risk management of everything. J Risk Fin 2004;5(3):58-65. [doi: 10.1108/eb023001]
54. Enterprise Risk Management Initiative. NC State University. 2024. URL: https://erm.ncsu.edu/ [accessed 2024-10-10]
55. Neelis M. A crisis communications plan for data breaches. PUSHKINPR Authentic Communication. URL: https://www.pushkinpr.com/blog/crisis-communications-plan-data-breaches/ [accessed 2024-10-10]
56. Vimercati SDCD, Foresti S, Livraga G, Samarati P. Data privacy: definitions and techniques. Int J Unc Fuzz Knowl Based Syst 2012;20(06):793-817. [doi: 10.1142/s0218488512400247]
57. Privacy and data management overview. Microsoft. 2024. URL: https://learn.microsoft.com/en-us/compliance/assurance/assurance-privacy [accessed 2024-10-10]
58. Privacy governance. Apple. URL: https://www.apple.com/legal/privacy/en-ww/governance/ [accessed 2024-10-10]
59. Creating trust through transparency. Google Privacy. URL: https://cloud.google.com/transparency [accessed 2024-10-10]
60. Blount Y. Employee management and service provision: a conceptual framework. Inf Technol People 2011;24(2):134-157. [doi: 10.1108/09593841111137331]
61. Who we are shapes the work we do. Google Belonging. URL: https://about.google/belonging/at-work/ [accessed 2024-10-10]
62. Training services—life is for learning, let's turn it into success. Siemens. URL: https://www.siemens.com/global/en/products/services/digital-enterprise-services/training-services.html [accessed 2024-10-10]
63. HR lessons from the Adobe performance management overhaul. Performyard. 2023. URL: https://www.performyard.com/articles/adobe-performance-management [accessed 2024-10-10]
64. The small business guide to internal communication. Salesforce. 2024. URL: https://www.salesforce.com/resources/articles/small-business-guide-internal-communication/ [accessed 2024-10-10]
65. Bhayani T. Crafting culture: insights from the Patagonia employee handbook for your business. AirMason. 2024. URL: https://blog.airmason.com/patagonia-employee-handbook-pdf/#:~:text=Patagonia%20champions%20employee%20empowerment%20and,special%20benefits%20for%20breastfeeding%20mothers [accessed 2024-10-10]
66. Sandhu RS, Samarati P. Access control: principle and practice. IEEE Commun Mag 1994;32(9):40-48. [doi: 10.1109/35.312842]
67. Microsoft Entra multifactor authentication (MFA). Microsoft. URL: https://www.microsoft.com/en-us/security/business/identity-access/microsoft-entra-mfa-multi-factor-authentication [accessed 2024-10-10]
68. Control access to AWS resources using policies. Amazon Web Services. URL: https://docs.aws.amazon.com/IAM/latest/UserGuide/access_controlling.html [accessed 2024-10-10]
69. What is role-based access control (RBAC)? Okta. 2024. URL: https://www.okta.com/identity-101/what-is-role-based-access-control-rbac/#:~:text=Role%2Dbased%20access%20control%20(RBAC)%20systems%20assign%20access%20and,different%20roles%20have%20different%20rights [accessed 2024-10-10]
70. Dewett T, Jones GR. The role of information technology in the organization: a review, model, and assessment. J Manage 2001;27(3):313-346. [doi: 10.1016/S0149-2063(01)00094-0]
71. Introducing Apple's on-device and server foundation models. Apple Machine Learning Research. 2024. URL: https://machinelearning.apple.com/research/introducing-apple-foundation-models [accessed 2024-10-10]
72. Best practices implementing zero trust with Palo Alto Networks. paloaltoTECHDOCS. 2024. URL: https://docs.paloaltonetworks.com/best-practices/zero-trust-best-practices/zero-trust-best-practices/the-five-step-methodology/step-1-asset-discovery-and-prioritization [accessed 2024-10-10]
73. AWS plans to invest 2.26 trillion yen into its Japanese cloud infrastructure by 2027. Amazon Web Services. 2024. URL: https://press.aboutamazon.com/aws/2024/1/aws-plans-to-invest-2-26-trillion-yen-into-its-japanese-cloud-infrastructure-by-2027 [accessed 2024-10-10]
74. von Solms R, von Solms B. From policies to culture. Comput Secur 2004;23(4):275-279 [FREE Full text] [doi: 10.1016/j.cose.2004.01.013]
75. Providing a safe and trusted experience for everyone. GooglePlay. URL: https://play.google/developer-content-policy/ [accessed 2024-10-10]
76. Modifying a policy. IBM. 2023. URL: https://www.ibm.com/docs/en/data-risk-manager/2.0.6?topic=policies-modifying-policy [accessed 2024-10-10]
77. Fenton E. Salesforce subprocessors: understanding key vendors and their roles.: visualping; 2024. URL: https://visualping.io/blog/salesforce-subprocessors [accessed 2024-10-10]
78. Bandor MS. Process and procedure definition: a primer. Software Engineering Institute. 2007. URL: https://insights.sei.cmu.edu/documents/3263/2007_017_001_23937.pdf [accessed 2024-10-10]

XSL•FO
RenderX

79. Updating system software by using Enterprise Class services. IBM. 2024. URL: https://www.ibm.com/docs/en/flashsystem-9x00/8.4.x?topic=uss-updating-system-software-by-using-enterprise-class-services [accessed 2024-10-10]
80. Our global impact—sustainability/ESG. 3M. 2024. URL: https://www.3m.com/3M/en_US/sustainability-us/governance/innovation-management/ [accessed 2024-10-10]
81. Procedures. Google developer documentation style guide. 2024. URL: https://developers.google.com/style/procedures [accessed 2024-10-10]
82. van der Kam WJ, Moorman PW, Koppejan-Mulder MJ. Effects of electronic communication in general practice. Int J Med Inform 2000;60(1):59-70. [doi: 10.1016/s1386-5056(00)00096-4] [Medline: 10974641]
83. Slack communication: enhancing collaboration and efficiency in the workplace. Cerkl. 2024. URL: https://cerkl.com/blog/slack-communication/ [accessed 2024-10-10]
84. Safeguard individual privacy with cloud services from Microsoft. Microsoft. URL: https://www.microsoft.com/en-us/trust-center/privacy/gdpr-overview#GDPR [accessed 2024-10-10]
85. The best ways to secure communication channels in the enterprise environment. Cybersecurity Insiders. URL: https://www.cybersecurity-insiders.com/the-best-ways-to-secure-communication-channels-in-the-enterprise-environment/ [accessed 2024-10-10]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
HHS: Department of Health and Human Services
HIPAA: Health Insurance Portability and Accountability Act
PHI: protected health information
PHR: personal health record

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 18.04.24; peer-reviewed by S Velichety, L Zhou; comments to author 28.05.24; revised version received 29.07.24; accepted 15.09.24; published 06.11.24.

Please cite as:
Subramanian H, Sengupta A, Xu Y
Patient Health Record Protection Beyond the Health Insurance Portability and Accountability Act: Mixed Methods Study
J Med Internet Res 2024;26:e59674
URL: https://www.jmir.org/2024/1/e59674
doi:10.2196/59674
PMID:

©Hemang Subramanian, Arijit Sengupta, Yilin Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Impact of Patient Access to Electronic Health Records on Health Care Engagement: Systematic Review

Dalia Alomar[1], MSc; Maryam Almashmoum[2], BSc, MSc; Iliada Eleftheriou[1], MSc, PhD; Pauline Whelan[1], MSc, PhD; John Ainsworth[1], MSc, PhD

[1]Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science System, The University of Manchester, Manchester, United Kingdom

[2]Nuclear Medicine Department, Faisal Sultan Bin Eissa, Kuwait Cancer Control Center, Kuwait City, Kuwait

Corresponding Author:
Dalia Alomar, MSc
Division of Informatics, Imaging and Data Sciences
School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science System
The University of Manchester
Vaughan House, Portsmouth St
Manchester, M13 9GB
United Kingdom
Phone: 44 7917941877
Email: dalia.alomar@postgrad.manchester.ac.uk

Abstract

Background: Health information technologies, including electronic health records (EHRs), have revolutionized health care delivery. These technologies promise to enhance the efficiency and quality of care through improved patient health information management. Despite the transformative potential of EHRs, the extent to which patient access contributes to increased engagement with health care services within different clinical setting remains a distinct and underexplored facet.

Objective: This systematic review aims to investigate the impact of patient access to EHRs on health care engagement. Specifically, we seek to determine whether providing patients with access to their EHRs contributes to improved engagement with health care services.

Methods: A comprehensive systematic review search was conducted across various international databases, including Ovid MEDLINE, Embase, PsycINFO, and CINAHL, to identify relevant studies published from January 1, 2010, to November 15, 2023. The search on these databases was conducted using a combination of keywords and Medical Subject Heading terms related to *patient access to electronic health records*, *patient engagement*, and *health care services*. Studies were included if they assessed the impact of patient access to EHRs on health care engagement and provided evidence (quantitative or qualitative) for that. The guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement were followed for study selection, data extraction, and quality assessment. The included studies were assessed for quality using the Mixed Methods Appraisal Tool, and the results were reported using a narrative synthesis.

Results: The initial search from the databases yielded 1737 studies, to which, after scanning their reference lists, we added 10 studies. Of these 1747 studies, 18 (1.03%) met the inclusion criteria for the final review. The synthesized evidence from these studies revealed a positive relationship between patient access to EHRs and health care engagement, addressing 6 categories of health care engagement dimensions and outcomes, including treatment adherence and self-management, patient involvement and empowerment, health care communication and relationship, patient satisfaction and health outcomes, use of health care resources, and usability concerns and barriers.

Conclusions: The findings suggested a positive association between patient access to EHRs and health care engagement. The implications of these findings for health care providers, policy makers, and patients should be considered, highlighting the potential benefits and challenges associated with implementing and promoting patient access to EHRs. Further research directions have been proposed to deepen our understanding of this dynamic relationship.

(J Med Internet Res 2024;26:e56473) doi:10.2196/56473

XSL•FO
RenderX

KEYWORDS

electronic health records; personal health record; health care engagement; empowerment; patient experience; patient satisfaction; health care services; systematic review

Introduction

Background

Health information technologies (HITs) have played a pivotal role in reshaping the landscape of patient care in the 21st century, offering promises of enhanced efficiency and improved quality through advanced health information management practices [1]. Electronic health records (EHRs), a pivotal component of HITs, have revolutionized the management of patient information, promising to elevate the efficiency and quality of health care delivery through streamlined data access and use [2-4].

In addition, governments across the globe are championing the cause of patient-centered care, recognizing it as a fundamental tenet in the pursuit of health care excellence. Therefore, they are increasingly recognizing the pivotal role of EHRs in reshaping health care dynamics, highlighting their central role in the quest for patient empowerment and improved health care outcomes [5]. In 2015, a basic EHR system was adopted by 75% of US hospitals to enhance communication between patients and their health care providers (HCPs), reduce health care costs, and increase patients' health care engagement [6]. The prioritization of citizen access to health records is a central focus of the MyHealthEData initiative by the US government. This is supported by initiatives by private entities such as Apple and Google, allowing patients to store their records on mobile devices. The 21st Century Cures Act enacted in the United States in December 2016 also includes crucial provisions that have the potential to substantially influence the availability, usability, and accessibility of health information for patients [7]. In the United Kingdom, primary care EHR providers, through TPP's SystmOnline, Egton Medical Information Systems Patient Access, and In Practice Systems' Patient Services, are actively providing patients with secure access to their health records. This access is further facilitated through the recently launched National Health Service app [5].

The significance of patient engagement in health care decision-making processes is now more pronounced than ever, with governments acknowledging its impact on both individual well-being and the overall quality of health care systems, enhancing patients' right to be active participants in decision-making processes related to their health [8,9]. As nations strive to build health care systems that are responsive to individual needs, patient engagement, achieved by patients' access to their EHRs, emerges as a linchpin, fostering a sense of ownership and partnership between patients and HCPs [10].

The Evolving Role of EHRs

EHRs, also called electronic medical records (EMRs), are digital versions of patients' comprehensive health information, including medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory test results. EHRs are designed to provide a centralized, electronic repository of health data that can be easily accessed and shared by authorized HCPs across different health care settings. The primary goal of EHRs is to improve the efficiency, quality, and safety of patient care regimen by facilitating accurate and timely access to pertinent health information, promoting care coordination among HCPs, and enhancing patient engagement in their health care journey [10].

Traditionally, EHRs were primarily viewed as tools for HCPs to record and retrieve patient information, with functionalities geared toward billing, documentation, and reimbursement purposes, thus streamlining the completion of various patient care tasks. The paradigm, however, has shifted as patients are now being granted direct access to their health records. This shift introduces a novel dynamic, empowering individuals to actively participate in their health care journey. Patient access to EHRs holds the promise of fostering a more collaborative and informed patient-HCP relationship, potentially influencing health care engagement in ways that extend beyond the traditional provider-centric model [11].

Patient access to EHRs empowers them with real-time and comprehensive health information, enabling informed decision-making aligned with personal preferences. This access facilitates seamless care continuity, offering a holistic view of one's health care journey and contributing to personalized and coordinated care as well as adherence to appointments and treatment plans [12,13]. In addition, EHRs act as educational tools, enhancing health literacy and promoting proactive health management [14]. By fostering shared decision-making, patients become active contributors to their care, ensuring a patient-centered approach. Furthermore, EHR access strengthens patient advocacy, allowing individuals to actively engage with HCPs, seek second opinions, control their possible health risks, and contribute to discussions about their treatment options [15,16]. This redefined perspective on health care engagement involves providing patients access to EHRs, enhancing patient-centered care experiences, and fostering dynamic and collaborative relationships between patients and the health care system [17,18].

Despite the recognized benefits of EHRs and the increasing focus on patient-centered care, the specific impact of allowing patients access to their EHRs on improving their engagement with health care services remains a distinct and underexplored area. Therefore, this study aimed to systematically review the literature describing, examining, and evaluating the impact of patient access to EHRs on their health care engagement within different clinical settings (inpatient, outpatient, and emergency). Through this focused lens of patient-centered care, we aim to investigate the impact of patient access to EHRs on health care engagement, thereby contributing to the ongoing efforts to optimize health care quality and outcomes.

XSL•FO
RenderX

Methods

Research Question

The systematic review provided in this study was guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [19]. Multimedia Appendix 1 shows PRISMA checklist used in this study. The research question was created based on the population, intervention, comparator, and outcome (PICO) framework [20], as shown in Textbox 1.

Textbox 1. Population, intervention, control, and outcome framework for the study strategy.

Population

- The population involves individuals who have access to electronic health records (EHRs). It involves patients across various demographic groups and health care systems or with different medical conditions.

Intervention

- The intervention is patient access to EHRs.

Comparator

- The comparator involves comparing individuals with access to EHRs to those without such access.

Outcome

- The outcome is the improvement in engagement dimensions with health care services.

Accordingly, the research question of this review was defined as follows: Does patient access to EHRs improve engagement with health care services?

Search Strategy and Sources of Information

The search strategy involved constructing specific terms related to "patient access," "EHRs," and "patient health care engagement" to ensure precision in the search results. The terms were expanded using Medical Subject Heading terms, synonyms, acronyms, abbreviations, and alternative spellings. These terms were combined using the Boolean operators AND and OR to optimize the search. Therefore, this phase involved identifying a comprehensive collection of studies relevant to the research question. This procedure encompassed *identification of the search keywords* and *definition of the search scope* [20].

During the *identification of the search keywords* phase, we established specific terms related to "patient access," "EHRs," and "patient health care engagement" to ensure precision in the search results. Following the approach recommended by Kitchenham et al [21], we broke down the research question into distinct units, treating each as a separate research unit. This involved encompassing Medical Subject Heading terms derived from the chosen databases, synonyms, acronyms, abbreviations, and alternative spellings, which were then amalgamated using the Boolean operators AND and OR to enhance the search strategy by locating the relevant studies. During the *definition of the search scope* phase, the source studies were acquired from the selected electronic databases (Ovid MEDLINE, Ovid Embase, PsycINFO, and CINAHL) through searches conducted up to November 15, 2023, using the formulated research keywords. The search strategies for the selected databases are presented in Multimedia Appendix 2.

Eligibility Criteria

The eligibility criteria for inclusion in this study were carefully defined to ensure relevance and reliability in addressing the research question. Textbox 2 outlines the criteria for inclusion and exclusion in this study. Studies were included if they met the criteria, while those that did not meet the criteria were excluded.

XSL•FO
RenderX

Textbox 2. Eligibility criteria (inclusion and exclusion criteria).

Inclusion criteria

- Article type: primary research studies including randomized controlled trials and observational studies (cohort and case control)
- Population: diverse patient populations across various demographics, medical conditions, and health care settings (inpatient, outpatient, and emergency)
- Intervention: examines the impact of patient access to electronic health records (EHRs) or electronic medical records (EMRs) or personal health record (PHR), as a main intervention, on health care engagement
- Focus: patient experiences and attitudes about the accessibility and use of EHRs (the demand side)
- Study design: quantitative, qualitative, and mixed methods designs
- Outcome measures: evaluates multiple dimensions of health care engagement (eg, involvement in health care decisions, adherence to treatment plans, communication, use of health care services, and experiences with patient-centered care)
- Publication type: peer-reviewed studies
- Publication date: studies published within the last 14 years (2010-2023)
- Language: studies published in English
- Availability: full text available or can be retrieved

Exclusion criteria

- Article type: secondary research (eg, reviews and meta-analyses) or other nonprimary research types (eg, editorials)
- Population: studies that lack suitable comparators or control groups essential for evaluating the impact of patient access to EHRs on health care engagement
- Intervention: studies that do not address patient access to EHRs, EMRs or PHR as a main intervention
- Focus: practitioner experiences and attitudes about EHR use (the supply side)
- Study design: studies that do not use quantitative, qualitative, or mixed methods designs
- Outcome measures: studies that do not evaluate the relevant dimensions of health care engagement as the main outcome
- Publication type: non–peer-reviewed publications (eg, editorials, book chapters, opinion pieces, commentaries, and conference abstracts)
- Publication date: studies published outside the 2010-2023 time frame
- Language: studies published in languages other than English
- Availability: studies with unavailable full text or those that cannot be retrieved

Selection Process

All studies identified through the database searches were exported to web-based EndNote (Clarivate Analytics) for screening and removing any duplicates. The selection process followed a systematic approach to identify and include relevant studies. It was conducted in accordance with the PRISMA 2020 statement to ensure transparency and reproducibility. First, a comprehensive search was conducted using 4 selected electronic databases, including MEDLINE, CINAHL, PsycINFO, and Embase. The search used the predefined keywords and eligibility criteria. Then, duplicates were eliminated, which arise due to the presence of certain studies in multiple databases. Titles and abstracts of the remaining studies were screened independently by 2 authors (DA and MA) to assess their relevance to the research question and eligibility criteria. Accordingly, all studies that did not meet these criteria were excluded, while studies passing this initial screening underwent a thorough full-text review by the 2 authors. These 2 authors assessed the entire text of the articles to confirm their eligibility and relevance. Nonrelevant studies were removed. Any discrepancies in study selection were resolved through discussion and consensus among the review team members.

Data Collection Process

Following the final selection of eligible studies by the 2 authors, data were systematically extracted by the primary author (DA). The extracted data encompassed various aspects, including study characteristics (eg, author, publication year, country, aim, intervention specifics, study design, study setting, patient demographics, and sample size), study outcome measures, key findings pertaining to health care engagement and the systematic review's research question, and quality of the study (based on the evidence strength). The accuracy of the extracted data was verified by the research team.

Risk of Bias and Quality Assessment of the Included Studies

After the final selection of the included studies, the risk of bias was assessed using the Mixed Methods Appraisal Tool (MMAT) [22]. It is used for assessment in reviews encompassing quantitative (randomized controlled trial [RCT], nonrandomized, or descriptive), qualitative, and mixed methods studies. The MMAT advises against assigning a singular score based on the evaluation [22]. Following a previous study [23], we used the κ statistic to evaluate the quality of each study, supporting our

XSL•FO
RenderX

final decisions based on the predetermined inclusion and exclusion criteria. On the basis of the fulfillment of criteria, the studies were categorized as high, medium, or low quality. A study attained a high-quality classification if it met all 5 MMAT criteria, a medium quality if 3 or 4 criteria were met (ie, meeting some criteria), and a low-quality classification if only 1 or 2 criteria were met (ie, meeting minimum criteria) [23].

Data Synthesis

Data extracted from the included studies were synthesized and analyzed using a narrative synthesis. A narrative synthesis is fitting and applicable in this study because it incorporates qualitative, quantitative, and mixed methods findings [24]. Studies with comparable outcomes, concerning dimensions of patient health care engagement, were pooled for summarization. The findings were presented by categorizing the results of different studies based on common outcomes of the patient health care engagement. This involved determining the number and percentage of studies associated with each outcome and leveraging these findings to draw conclusions. Given the variability observed in EHRs, health care settings, and patient populations, a meta-analysis was deemed inappropriate. Data analysis was conducted by one author (DA), and any discrepancies were resolved through discussions with the other author (MA). Through the analysis, we identified categories within patient health care engagement, as shown in Table 1.

XSL•FO
RenderX

Table 1. Main categories and dimensions of patient health care engagement outcomes (N=18).

Categories and dimensions	Studies, n (%)	Reference
Treatment adherence and self-management		
Adherence to prescribed treatments and medications plans	7 (39)	• Zarcadoolas et al [25] • Wang et al [26] • Nazi et al [27] • Suija et al [28] • Wass et al [29] • Spratt et al [30] • van der Vaart et al [31]
Self-monitoring or tracking disease and health parameters over time	4 (22)	• Klein et al [32] • Ryu et al [33] • Haggstrom et al [34] • Hanna et al [35]
Self-management of diseases by using self-management tools (eg, mobile apps or PHRs[a])	2 (11)	• Fuller et al [36] • Hanna et al [35]
Patient involvement and empowerment		
Improved accessibility to or usability of health care information	6 (34)	• Nazi et al [27] • Suija et al [28] • van der Vaart et al [31] • Haggstrom et al [34] • Hanna et al [35] • Moll et al [37]
More involvement in treatment and active participation in disease and medication management decisions (eg, patient-generated data and refilling prescriptions)	4 (22)	• Zarcadoolas et al [25] • Suija et al [28] • Wass et al [29] • van der Vaart et al [31]
Enhanced patient agency and responsibility for their care	2 (11)	• Wass et al [29] • Hanna et al [35]
Confidence in managing one's health	1 (5)	• Wolff et al [38]
More informed (health literacy) and better understanding of their condition (disease and treatment)	8 (44)	• Klein et al [32] • Nazi et al [27] • Suija et al [28] • Wass et al [29] • van der Vaart et al [31] • Wolff et al [38] • Moll et al [37] • Mák et al [39]
Health care communication and relationship		
Better prepared for office visits	1 (5)	• Wolff et al [38]
Improved frequency and quality of communication between patients and HCPs[b] regarding disease monitoring	10 (56)	• Klein et al [32] • Zarcadoolas et al [25] • Nazi et al [27] • van der Vaart et al [31] • Fuller et al [36] • Wolff et al [38] • Hanna et al [35] • Moll et al [37] • Wagner et al [40] • Ibrahim et al [41]
Discussion and collaboration between patients and HCPs about patient treatment plans	4 (22)	• Klein et al [32] • Nazi et al [27] • Spratt et al [30] • Wolff et al [38]

XSL•FO
RenderX

Categories and dimensions	Studies, n (%)	Reference
Trust in HCPs and improved patient-HCP relationship	4 (22)	• Klein et al [32] • Nazi et al [27] • van der Vaart et al [31] • Fuller et al [36]
Shared decision-making	4 (22)	• Wang et al [26] • Ryu et al [33] • Haggstrom et al [34] • Hanna et al [35]
Patient satisfaction and health outcomes		
Perceived effectiveness of treatments	4 (22)	• Klein et al [32] • Wang et al [26] • Ryu et al [33] • Wagner et al [40]
Improvement in health conditions and quality of health care and life	4 (22)	• Wang et al [26] • Nazi et al [27] • van der Vaart et al [31] • Hanna et al [35]
Functions tailored to personalized preferences	3 (17)	• Zarcadoolas et al [25] • Krist et al [42] • Haggstrom et al [34]
Achievement of health-related goals (eg, health risk assessment)	2 (11)	• Krist et al [42] • Haggstrom et al [34]
Overall satisfaction with health care services	2 (11)	• Mák et al [39] • Ibrahim et al [41]
Use of health care resources		
Frequency of health care visits	1 (5)	• Zarcadoolas et al [25]
Efficient use of health care and information resources (review, correct, and update health information)	1 (5)	• Krist et al [42]
Use of preventive care services	1 (5)	• Krist et al [42]
Usability concerns and barriers		
Privacy and security concerns	2 (11)	• Zarcadoolas et al [25] • Haggstrom et al [34]
Technical barriers in accessing or using health records features	3 (17)	• Zarcadoolas et al [25] • Suija et al [28] • Spratt et al [30]
Barriers in comprehension of health information or medical terms	3 (17)	• Zarcadoolas et al [25] • Wass et al [29] • Suija et al [28]
Medications and follow-up concerns after discharge from the hospital	1 (5)	• Fuller et al [36]

[a]PHR: personal health record.

[b]HCP: health care provider.

Results

Study Selection

The electronic search from the 4 selected databases retrieved a total of 1737 study records, of which 183 (10.54%) were from Ovid MEDLINE, 220 (12.66%) were from Ovid Embase, 18 (1.04%) were from Ovid PsycINFO, and 1316 (75.76%) were from EBSCOhost CINAHL. Following the removal of duplicate records (n=310, 17.74%) and addition of 10 articles based on scanning reference lists of the articles initial searched, 1437 (82.26%) studies remained that underwent assessment during title and abstract screening. Of these 1437 studies, 83 (5.77%)

XSL•FO
RenderX

met the eligibility criteria for full-text screening, and the final number of studies included in the review was 18 (1.25%). The selection process and rationale for choosing these studies are shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement flowchart for the search and selection process of the included studies. EHR: electronic health record; HCP: health care provider.

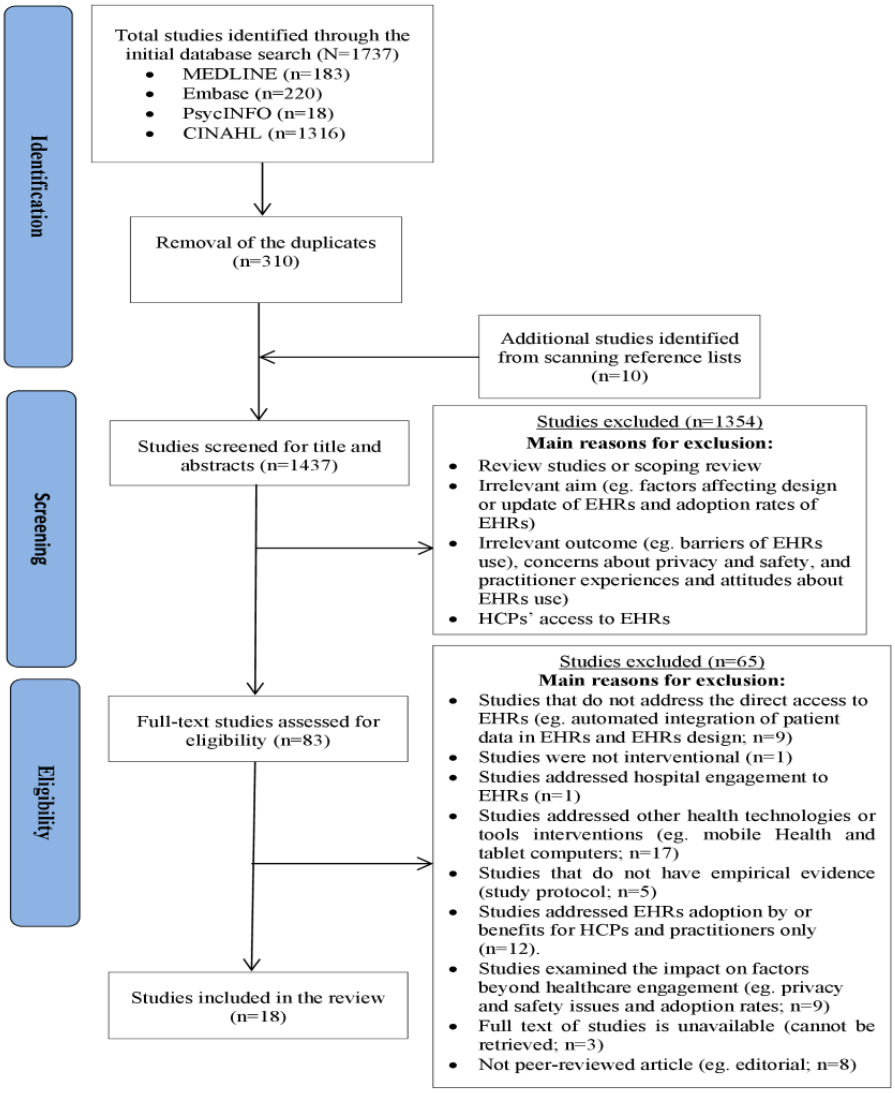

Characteristics of the Included Studies

Multimedia Appendix 3 [25-42] provides a summary of the characteristics of the studies that have been included. All included studied aimed at investigating or describing the impact of patient access to or the implementation of EHRs with their different applications and platforms on patient engagement, including their attitudes, perceptions, experience, empowerment, and satisfaction. The 18 studies included in the review spanned various global regions: 8 (49%) were from the United States, 2 (11%) were from Sweden, 1 (5%) was from Germany, 1 (5%) was from China, 1 (5%) was from Estonia, 1 (5%) was from South Korea, 1 (5%) was from the Netherlands, 1 (5%) was from Australia, 1 (5%) was from Canada, and 1 (5%) was from Malaysia.

Of the 18 studies, qualitative methods (focus groups, interviews, semistructured interviews, and observations) were used in 3 (17%) of studies; quantitative methods (surveys, questionnaires, laboratory test, and mobile phone app recording activity log [eg, MyHealthKeeper]) were used in 10 (56%) of studies; while 5 (27%) studies adopted a mixed approach. Regarding the study designs, the studies used a RCT (4/18, 22%), cross-sectional (2/18, 11%), and pretest-posttest implementation design (2/18, 11%), and the remaining studies used observational, case report, or exploratory study with a retrospective cohort design.

XSL•FO
RenderX

For participants demographics, the studies varied regarding their participants; few studies were focused on both patients and physicians (eg, dermatologists), while most targeted patients only. All included studies shared that the patient sample encompassed adult patients (aged 21-83 years), with sex diversity (male individuals and female individuals), had experience with computers and use the internet, diagnosed in most studies with a specific disease (eg, chronic obstructive pulmonary disease [COPD], diabetes, uncontrolled hypertension, or metabolic syndrome), and they enrolled in or accessed EHRs with various applications and web-based portals or platforms during the study period. The clinical setting of the included studies encompassed various settings, including inpatient, outpatient, and ambulatory clinics.

Within the scope of patient access to EHRs, the 18 included studies revealed a spectrum of interventions related to patient access to or implementation of EHRs either via patient portals (n=5, 28%), a specific mobile app such as MyChart or MyHealthKeeper (n=2, 11%), or other web-based platforms (n=11, 61%). Notably, a subset of studies concentrated on disease-specific applications, such as using EHRs in psoriasis (1/18, 5%) or diabetes care (1/18, 5%). Studies describing patients accessing their medical records either through personal health records (PHRs) tethered to EHRs or EMRs (5/18, 28%) or interactive PHRs (IPHRs; 1/18, 5%) stood out as approaches generating tailored recommendations from EHRs or EMRs, emphasizing personalized health care. Other studies explored unique interventions, such as web-based coaching programs (1/18, 5%), the MyHealtheVet (MHV) Pilot Program (1/18, 5%), and the implementation of a discharge video checklist (1/18, 5%), all aiming to enhance patient engagement through electronic records.

All included studies [25-42] addressed evaluating ≥1 dimensions and measures of health care engagement outcomes because of patient EHRs access, including patient involvement in health care decisions, adherence to treatment plans, communication with HCPs, use of health care services, and experiences with patient-centered care. The identified categories of theses dimensions of health care engagement outcomes are shown in Table 1.

Multimedia Appendix 4 shows the quality of the included studies [25-42]. Just 1 (5%) of the 18 studies was considered as having low quality [28]. Despite its methodological limitations, such as a small sample size and issues with data quality inspection, the study's findings offer qualitative depth that complements quantitative evidence in understanding the complexities of health data use via access to the EHRs. Therefore, this paper needs to be documented in this study. Specifically, Suija et al [28] identified 3 crucial themes for the experiences of patients and primary care physicians with health data, particularly during the COVID-19 pandemic. These themes included access to health records, experiences with using data in health records, and the use of patient-generated data, providing unique insights into the perspectives of both patients and physicians.

Quality (Risk of Bias) of the Included Studies

Multimedia Appendix 4 [25-42] summarizes the outcomes of the quality assessment conducted for the included studies. Of the 18 studies, 14 (78%) were classified as high quality, meeting all 5 MMAT criteria: 1 (5%) qualitative, 8 (45%) quantitative, and 5 (28%) mixed methods studies. In addition, 3 (17%) of the 18 studies were deemed medium quality, fulfilling 3 or 4 of the MMAT criteria: 1 (5%) qualitative and 2 (12%) mixed methods studies. Of the 18 studies, 1 (5%) with a qualitative approach was identified as low quality, meeting only 2 of the MMAT criteria. Given the exploratory nature of this review and its limited research evidence and in accordance with the suggestion of Hong et al [22], we opted not to exclude the study meeting only 2 of the MMAT criteria from the final review.

Synthesis of the Results

Overview

Table 1 shows how patient engagement to health care services is improved through their access to EHRs. We categorized the identified dimensions and outcomes of patient engagement that were affected by patient EHRs access by grouping them according to their shared characteristics, as outlined in prior studies [1-24]. These dimensions were classified into 6 categories based on the nature of engagement: treatment adherence and self-management, patient involvement and empowerment, health care communication and relationship, patient satisfaction and health outcomes, use of health care resources, and usability concerns and barriers.

Categories of the Dimensions of Patient Health Care Engagement Outcomes Affected by EHRs Access

As mentioned in the *Synthesis of the Results* section, the identified dimensions and outcomes of patient engagement that were affected by patient access to EHRs were classified into 6 main categories, as shown in Table 1. The following sections synthesize the categories of those dimensions.

Treatment Adherence and Self-Management

As shown in Table 1, there are 3 identified dimensions of patient engagement in this category. The most cited dimension of them is the adherence to prescribed treatments and medication plans. It is cited in 7 (39%) of the 18 studies [25-31]. Adherence to prescribed treatments and medication plans refers to the extent to which patients follow the recommendations and instructions provided by HCPs regarding their medical treatments, including medication regimens. Adherence is a critical factor in achieving positive health outcomes. It directly influences the effectiveness of treatments, disease management, and overall well-being. Poor adherence can lead to treatment failure, worsening of health conditions, and increased health care costs.

Patient access to EHRs plays a pivotal role in influencing adherence to prescribed treatments. With access to their health information, including medication lists, dosage instructions, and treatment plans, patients found that they could gain a clearer understanding of their prescribed regimens [28,29,31]. This increased awareness empowers patients to take an active role in managing their health. Patient access to their medical records through a PHR connected to an EHR, which included wellness reminders, helped them take action and adhere to their treatment plans [27]. In addition, EHRs can facilitate automated reminders for medication schedules and upcoming appointments, annual

XSL•FO
RenderX

visits, and screenings. These reminders serve as prompts for patients to adhere to their prescribed plans consistently [25]. For evaluation of a web-based coaching program using EHRs for patients with COPD, Wang et al [26] illustrated that this intervention comprises web-based EHRs containing extensive information about preventive measures, treatment options, pulmonary rehabilitation, and variations in the disease. These resources motivated patients to actively participate in healthy behaviors and enhance adherence to prescribed medications, oxygen therapy, and respiratory exercises. Moreover, The Duke PillBox application, a medication management application integrated into the EHRs' patient portal using Substitutable Medical Applications and Reusable Technologies on Fast Healthcare Interoperability Resources technology, is considered beneficial due to enhanced patient interactions, leading to improved adherence as patients follow prescribed medication regimens [30].

In total, 4 (22%) of the 18 studies illustrated that patient access to their EHRs significantly helped in disease and health parameters monitoring and tracking over time [32-35]. Using an EHR enabled patients to monitor their disease progression over time [32]. Personally controlled EHR (PCEHR) enabled patient to review and self-monitor their health status and treatment [35]. MyHealthKeeper, an EHR-tethered PHR application, demonstrated the effectiveness of a health tracker system managed by patients and guided by clinicians [33]. The MHV program, integrated into PHRs, also empowered users to monitor their health by incorporating various types of self-reported health information, including vital signs, laboratory results, tests, and records of food and activity journals [34]. Regarding the active engagement of individuals in managing and controlling aspects of their health related to a specific medical condition, it is reported in 2 (11%) of the 18 studies that patient access to EHRs plays a significant role in diseases self-management by using self-management tools (eg, mobile apps or PHRs) [35,36].

Patient Involvement and Empowerment

Patient involvement and empowerment refer to the active participation, engagement, and empowerment of individuals in their own health care processes and decision-making. Empowerment involves providing individuals with the knowledge, skills, and resources necessary to actively participate in discussions about their health, set personal health goals, and contribute to the decision-making process. Five dimensions of patient engagement are identified in studies for this category.

Of the 18 included studies, 6 (34%) addressed the improved accessibility to health care information using EHRs. PHRs enhanced the value of users' data by ensuring information is more accessible and understandable [34]. Using the MHV program integrated into PHRs, patients engaged with self-instructional materials. This facilitated their ability to find pertinent information through the health education library; input data using the website's Healthelog features; search for information on specific diseases, conditions, or treatments; and engage in discussions with their HCPs. The features most valued by users included accessing portions of their medical record, reviewing prescription history, and checking upcoming appointments [27]. Suija et al [28] demonstrated that both patients and physicians recognized the importance of accessing health records. Patients highlighted the significance of accessing documents such as hospital discharge summaries, consultation responses, and test results. In contrast, physicians emphasized the need for convenient access to past patient health records to effectively prepare comprehensive treatment plans. The hospital-based patient web portal, offering remote access to EMRs, also demonstrated ease of use and high utility, with minimal reported issues [31]. The PCEHR provided patients with direct and immediate access to their health information without the need for intermediaries, such as HCP gatekeepers. This empowered patients to use the information collaboratively with HCPs in making informed choices about their health care [35]. Moreover, Moll et al [37] showed that patient access to EHRs through Journalen, the Swedish national patient portal, facilitated retrieving their test results, immunization history, and visit details, improving their accessibility and usability of their health care information.

There were 4 (22%) out of 18 studies that focused on patient engagement through more involvement in their treatment and active participation in disease and medication management decisions [25,28,29,31]. Patient portals played a crucial role in promoting patient empowerment, particularly in medication management by using features such as making appointments, refilling prescriptions, and contributing patient-generated data [25]. Suija et al [28], using individual semistructured interviews with patients and physicians, reported that patients expressed a willingness to generate crucial data necessary for medical decisions, such as completing questionnaires before appointments to systematize their complaints and save time during consultations. Physicians also acknowledged the value of patient-generated data and considered it valuable when making treatment decisions. The accessibility of EHRs is linked to heightened patient involvement in their treatment. This increased accessibility results in patients being more likely to follow the advice provided by HCPs, have better recall of information, understand discussions during appointments more clearly, and communicate more effectively with HCPs about their situation. As a result, they can actively participate in managing their diseases and medication decisions [29,31].

There were 8 (44%) out of 18 studies that highlighted that patient access to EHRs contributes to increased patient awareness and understanding of their condition, fostering improved health literacy [27-29,31,32,37-39]. Patient access to EHRs enhances comprehension of their medical problems and treatments [27]. This improved understanding extends to a better grasp of their overall care, fostering feelings of being more informed and secure [37,38]. The access to EHRs facilitated the understanding, discussion, and coordination of documented information, making it easier for patients to comprehend what was said during appointments and communicate with HCPs about their situation (diseases and instructions for self-care) [28,29]. Patients also found it intriguing to compare their data with those of others, enabling more informed discussions with HCPs [32]. Among patients who logged into the hospital-based patient web portal with home access to EMR, 37% reported having increased knowledge about their diseases and treatment

XSL•FO
RenderX

[31]. However, in exploring the impact of web-based access to laboratory test results in British Columbia on patient experiences, Mák et al [39] conducted an exploratory study with a retrospective cohort design, revealing that while both service users and the comparison group reported understanding their test results, the rate was slightly lower for service users (75.55% vs 84.69%).

Patient access to EHRs significantly enhanced patient agency and responsibility for their care [29,35]. Web-based access to EHRs made it easier for patients to take greater responsibility for their care [29]. Use of Medenote as a PCEHR promoted patient agency by enhancing their capacity for self-management and making independent, informed choices about their health [35]. Furthermore, according to Wolff et al [38], patients exhibited heightened confidence in managing their health.

Health Care Communication and Relationship

Patient access to EHRs plays a significant role in improving patient engagement through enhancing the physician-patient communication and relationship. There are 5 identified dimensions for this patient engagement outcome. Using EHRs, patients felt better prepared for office visits [38]. Furthermore, more than half of the studies (10/18, 56%) focused on the crucial effect of patient access to EHRs on the frequency and quality of communication between patients and HCPs regarding disease monitoring [25,27,31,32,35-38,40,41]. This access played a crucial role in enhancing their communication with HCPs [25,27,31,35,37]. The frequency of this communication is associated with the regular use of PHRs or EHRs [40,41]. It is positively impacted by patient access to visualized data [32]. Patients perceive that when preparing for discharge from the hospital, the integration of digital health tools with EHRs such as discharge videos, checklists, and postdischarge SMS text messaging fostered communication with their health care team [36]. In addition, the electronic delivery of doctors' visit notes to patients and authorized care partners through OpenNotes (Deaconess Medical Center) contributes to enhanced communication with HCPs [38].

Several studies (4/18, 22%) found that access to EHRs not only enhances communication and relationships between HCPs and patients but also facilitates collaborative discussions about patient treatment plans, particularly when based on visualized data [32]. The discussion of medication reconciliation, including indications, side effects, and barriers to use, becomes vital for improved communication when patients have access to EHRs [27,30]. Wolff et al [38] illustrated that patients and care partners reported better agreement about patient treatment plans and more productive discussions regarding patient care, emphasizing the positive impact of EHR access on collaborative health care discussions.

In total, 22% (4/18) of the studies revealed that shared decision-making is a crucial aspect of HCP-patient communication, which is influenced by patient access to EHRs [26,33-35]. It is reported that the use of a web-based coaching program using EHRs is instrumental in fostering the sharing of patients' medical information among hospital and community nurses, facilitating dynamic management, and enabling follow-up analysis of patients' diseases [26]. MyHealthKeeper, an EHR-tethered PHR application, allows the seamless sharing of life log data between clinicians and patients [33]. Patient access to the MHV program integrated into PHRs revealed preference for information sharing with HCPs among most users [34]. In addition, through access to PCEHR, patients reported their ability to use health information in conjunction with HCPs, empowering them to make informed choices about their health care [35].

In total, 17% (3/18) of the studies found that trust in HCPs and improved patient-provider relationships are crucial dimensions influenced by patient access to EHRs [31,32,36]. It is the result of the improved communication between HCPs and patient. Patient access to EHRs and discussions centered on visualized data enhanced communication and, consequently, improved the relationship between patients and physicians [32]. Furthermore, the trust of patients in HCPs was heightened through patient access to EMRs [31,36].

Patient Satisfaction and Health Outcomes

This category of patient health care engagement encompasses the overall contentment of patients with their health care experiences and the impact of these experiences on their well-being. It examines the extent to which patients are pleased with the health care services they receive and explores the resulting effects on their health status.

There were 4 (22%) out of 18 studies that concluded that patient access to EHRs affects their satisfaction through perceived effectiveness of treatments [26,32,33,40]. Patients reported that using an EHR resulted in an enhancement in the quality of their treatment [32]. When Wang et al [26] investigated the effectiveness of a web-based coaching program using EHRs for improving physical function in patients with COPD across 2 hospitals in China, they indicated that this program could delay the decline in lung function, reduce dyspnea, and enhance physical capacity. Through analyzing data on health outcome measures such as weight change, changes in blood biochemical parameters (cholesterol, triglycerides, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol), Ryu et al [33] revealed that MyHealthKeeper PHR application was found to improve patient clinical profiles, as evidenced by weight loss and lower triglyceride levels [33]. In contrast, Wagner et al [40] conducted an intervention involving a PHR tethered to the patient's EMR for 453 patients, including those with metabolic syndrome, in ambulatory clinics. Their results showed no observed impact of the PHR on blood pressure. However, in the subanalysis of that study, a further examination of patients participating in the intervention, specifically those who identified themselves as active users of the PHR, revealed a decrease of 5.25 points in diastolic blood pressure.

Improvement in health conditions and quality of health care and life was reported in 4 (22%) of the 18 studies [26,27,31,35]. Patient access to EHRs has improved the quality of care and life through various applications and platforms, including a web-based coaching program using EHRs [26], a web portal [31], PCEHRs [35], and MHV (a PHR connected to EHRs). These resources, which include wellness reminders, help patients take action, adhere to their treatment plans, and enhance communication with their HCPs [27].

XSL•FO
RenderX

Functions tailored to personalized preferences are a significant aspect of patient satisfaction influenced by access to EHRs, as reported in 17% (3/18) of the studies [25,34,42]. Accessing EHRs through patient portals positively influenced health outcomes by providing personalized reminders for scheduled appointments, annual visits, and screenings [25]. The IPHR demonstrated its capability to offer personalized recommendations for patients [42]. Usability tests conducted on PHRs provided valuable insights into tailoring functions according to individual preferences [34].

Achievement of health-related goals is a valuable dimension in patient satisfaction that is affected by patient access to EHRs, as reported in 2 (11%) of the 18 studies [34,42]. It concerns about enhancing various aspects of patient daily life, such as physical, mental, and social well-being. Using EMRs, the IPHR effectively involved patients in preventive care by conducting a concise health risk assessment. This assessment collects information that may not be well documented electronically, including health behaviors and psychosocial measures [42]. The MHV program, integrated into PHRs, also enabled patients adapting to posttraumatic stress disorder and healthy eating [34].

In 11% (2/18) of the studies, patient access to EHRs was found to improve the overall satisfaction with health care services [39,41], especially a larger proportion of users of the web-based access to laboratory test results service received their results more promptly, typically within "a few days," in contrast to just over a third of the comparison group. This discrepancy in result delivery times contributed to higher overall satisfaction levels among service users [39].

Use of Health Care Resources

Use of health care resources is considered an important category of patient engagement impacted by patient access to EHRs. It involved the frequency of health care visits via EHRs. According to 6% (1/18) of the studies, accessing EHRs through patient portals resulted in longer doctor's visits, allowing patients more time with their doctors [25]. This contributed to improving the effectiveness of visits with HCPs.

Moreover, this access to EHRs is found to achieve efficient use of health care and information resources in 5% (1/18) of the studies [42]. Krist et al [42] revealed that using EMRs, the IPHR effectively established an open database connection to link with the EMR of the patient's designated clinician, extracting pertinent clinical data. This enables patients to access information such as their medical history, medications, immunizations, and test dates. The platform empowers patients to review, correct, and update their information as needed.

Another dimension of efficient use of health care resources is the use of preventive care services. Krist et al [42] revealed that by using EMRs, the IPHR effectively involved patients in preventive care through enabling patient access to results related to preventive care, such as recommendations for chronic care and managing high cholesterol, achieving a use rate that varied from 1.5% to 28.3% across different health care practices.

Usability Concerns and Barriers

Some of the included studies reported that there were some usability concerns and barriers resulting from using or accessing EHRs. Privacy and security concerns were raised by 11% (2/18) of studies [25,42]. It is found that PHR registration should strike a balance between simplicity and security [34]. Concerns related to the privacy and security of web-based records, such as potential damage or hacking, were expressed by a small number of participants using EHRs via patient portals [25].

Technical barriers in accessing or using health records features are reported in 17% (3/18) of the studies [25,28,30]. Consumers expressed apprehensions about various obstacles to use, including intricate visual layouts and suboptimal usability features [25]. Patient access to documents in health records through the patient portal is deemed intricate [28]. Usability challenges encompass EHR integration, concerns with the EHR-portal communication interface, as well as issues related to browser access and compatibility, affecting both patients and HCPs [30].

Furthermore, barriers related to comprehension of health information or medical terms are reported in 17% (3/18) of the studies [25,28,29]. Consumers expressed concerns about the complexity of language used in EHRs [25]. Challenges such as patients grappling with medical terminology and the necessity for rigorous inspection of medical documents were noted [28]. In a study by Wass et al [29], only 9% of participants reported being worried or upset about the information in the EHRs, with some finding it challenging to comprehend due to the inclusion of medical terms. In addition, concerns regarding medications and follow-up after discharge from the hospital were raised in 5% (1/18) of the studies [36].

Discussion

Principal Findings

This systematic review comprises 18 studies exploring the impact of patient access to EHRs on various aspects of health care engagement. Our findings revealed that most studies reported positive findings for patient health care engagement, including enhanced patient empowerment and involvement, improved adherence to treatment plans and self-management, increased patient confidence in managing their health, promoted health care communication and relationship and trust between HCPs and their patients, effective use of health care resources, and improved patient satisfaction and health outcomes. In addition, some studies raised some usability concerns and barriers for EHRs access.

This comprehensive analysis directly addresses our research question, enhancing our understanding of the role EHRs play in shaping patient engagement and guiding future research and health care practices toward more patient-centered care. However, it is crucial to acknowledge the diversity among the studies, encompassing variations in interventions related to EHRs access, types of applications and platforms used for EHRs access, health care settings, methodologies, and outcome measures. This diversity adds complexity to the interpretation

XSL•FO
RenderX

of findings and underscores the need for a nuanced understanding of the implications of patient access to EHRs.

A broad range of interventions, all related to EHRs access, included access to EHRs via patient portals, web-based coaching programs using EHRs, use of PHRs tethered to EHRs, integration of digital health tools with EHRs (eg, discharge videos, checklists, and SMS text messaging), use of mobile apps (eg, Duke PillBox, MyHealthKeeper, and MHV) for accessing EHRs, implementation of health tracker systems (eg, Misfit), open access to laboratory test results through EHRs, and delivery of physician's visit notes electronically via OpenNotes. Similarly, the included studies encompassed several types of EHRs, including EMRs accessed through a web portal, web-based access to health information through PHRs (eg, via MHV program integrated into PHRs), PCEHRs, Duke PillBox application (integrated into EHRs patient portal), Journalen (personal accessible EHRs), and IPHR. This variety in EHR-related interventions and EHR types collectively represents a diverse array of strategies and platforms aimed at enhancing patient engagement and health care outcomes as well as emphasizes the multifaceted nature of patient engagement in the context of electronic health information systems.

Comparison With the Literature

Our findings in this systematic review indicated that patient access to EHRs has a significant influence on diverse categories and dimensions of patient health care engagement. The six main categories or themes that emerged from the included studies encompassed the following: (1) treatment adherence and self-management, (2) patient involvement and empowerment, (3) health care communication and relationship, (4) patient satisfaction and health outcomes, (5) use of health care resources, and (6) usability concerns and barriers. Our findings are consistent with the Patient Portal Engagement Framework developed by Zhou et al [43], which includes 4 levels of engagement: Inform Patients, Involve Patients, Partner with Patients, and Support Ecology of Care.

However, the 6 categories or themes identified in our review extend beyond the Patient Portal Engagement Framework by providing a more granular view of patient engagement outcomes. This comprehensive set of outcome measures for patient health care engagement, as highlighted in Table 1, offers a broader perspective than what is typically covered in the literature. While some studies may focus on 1 or 2 dimensions of patient engagement, our work integrates multiple dimensions, thereby providing a holistic view of patient engagement outcomes.

Moreover, the 6 categories or themes have not been previously described in the literature as a unified framework, making this a novel contribution of our systematic review. This framework provides a structured approach to understanding the multifaceted nature of patient engagement resulting from EHRs accessibility, encompassing both benefits and challenges. This unique categorization can serve as a valuable contribution for knowledge and practical applications in enhancing patient engagement through EHRs. This comprehensive framework can be discussed as follows.

Patient access to EHRs positively affects treatment adherence and self-management for patients. This access empowers patients by providing information on medications, dosage instructions, and treatment plans [28,29,31]. This heightened awareness leads to active patient participation in health care management [27]. EHRs, especially when integrated into patient portals or PHRs, prove beneficial through wellness reminders, encouraging patients to adhere to treatment plans consistently [25,30].

Our findings confirm those of previous studies, which consistently show that EHRs are powerful tools for self-monitoring and self-management of disease and health parameters over time. Patients using EHRs can longitudinally track disease progression [32]. The PCEHR system allows patients to review and self-monitor their health status and treatment [35]. MyHealthKeeper and MHV also empower users to monitor their health through self-reported information, enhancing engagement [33,34]. In self-management of diseases through tools such as mobile apps or PHRs, patient access to EHRs plays a significant role [35,36], especially for patients with chronic disease (eg, patients with diabetes) [44]. Patients actively engage in managing their health, supported by the functionalities provided by EHRs [35,36]. This is in line with prior studies that indicated that EHRs, when integrated into a patient-centered self-management strategy, have the potential to empower patients for more effective self-management of their diseases and can help address challenges that contribute to unsuccessful treatment outcomes [45]. For example, individuals using the Canadian MyChart PHRs began associating their dietary choices with their laboratory test results, specifically sodium levels [46].

The studies consistently demonstrated that patient access to EHRs enhances the accessibility and usability of health care information. PHRs, such as MHV, were instrumental in providing patients with self-instructional materials, contributing to a better understanding of their health [28]. The hospital-based patient web portal and PCEHR further extended the ease of access to health records, promoting patient engagement [31,35]. Therefore, these findings align with the broader theme of empowering patients through improved access to their health care information.

Patient portals emerged as facilitators of active patient involvement, especially in medication management. Patients expressed a willingness to generate crucial personalized health data, emphasizing the importance of their active participation in medical decisions [25,28,29,31]. The increased accessibility to EHRs was linked to heightened patient involvement, indicating that informed patients are more likely to actively engage in decisions related to their treatment plans.

The studies consistently highlighted that patient access to EHRs contributes to enhanced patient agency and a greater sense of responsibility for their care. Web-based access to EHRs made it easier for patients to take control of their health [29]. For instance, Medenote, as a PCEHR, played a pivotal role in promoting patient agency, allowing them to make independent and informed choices about their health [35]. Therefore, this finding underscores the transformative impact of EHR access

XSL•FO
RenderX

on patients' roles in managing their health care. Wolff et al [38] revealed a crucial aspect of patient empowerment—heightened confidence in managing their health. This aligns with the broader narrative that patient access to EHRs not only provides information but also instills a sense of confidence in patients about managing their health effectively.

The studies consistently demonstrated that patient access to EHRs contributes to increased patient awareness and understanding of their condition. Patients found it intriguing to compare their data with others, fostering more informed discussions with HCPs [32]. However, it is noteworthy that the exploratory study by Mák et al [39] revealed slightly lower rates of understanding test results among service users, indicating potential variations in health literacy outcomes.

Regarding the impact of EHRs access on health care communication and relationship, the findings revealed multifaceted impact. Patients expressed feeling better prepared for office visits when granted access to EHRs. This preparation likely stems from their ability to review and familiarize themselves with their medical records, enabling more informed and productive discussions during appointments [38].

Most of the studies underscored the crucial role of patients' regular access to and use of EHRs in enhancing the frequency and quality of communication between patients and HCPs concerning disease monitoring. Patients perceive digital health tools integrated with EHRs, such as videos, checklists, and SMS text messaging, as valuable contributors to fostering communication with their health care teams [36]. The delivery of visit notes electronically through OpenNotes further contributed to enriched communication [38].

Access to EHRs not only enhances communication but also facilitates collaborative discussions about patient treatment plans. Visualized data play a pivotal role in this dimension, emphasizing the value of data presentation in health care discussions [32]. The collaborative nature of discussions, particularly around medication reconciliation and barriers to use, is highlighted as vital for improved communication [27,30,38]. Shared decision-making emerged as a significant outcome influenced by patient access to EHRs. The use of a web-based coaching program using EHRs contributes to the sharing of patients' medical information among HCPs, promoting dynamic management and follow-up analysis of patients' diseases [26]. The seamless sharing of life log data through EHR-tethered PHR applications supports collaborative decision-making [33]. Patient access to the MHV program, for instance, reveals a preference for information sharing, emphasizing the collaborative nature of health care decisions [34]. In addition, access to PCEHR empowers patients to make informed choices about their health care [35].

The studies underscore that trust in HCPs and the overall patient-HCP relationship are positively influenced by patient access to EHRs [31,32,36]. Improved communication, particularly discussions centered on visualized data, contributes to the enhanced relationship between patients and physicians [32]. The heightened trust in HCPs is also evident through patient access to EMRs [31,36].

Patient satisfaction and health outcomes represent a critical dimension in understanding the impact of patient access to EHRs. Across the 18 studies included in this review, several dimensions emerged, reflecting the diverse effects on patients' satisfaction and overall health. A notable finding suggests that patient access to EHRs contributes to perceived effectiveness in treatments [26,32,33,40]. Patients reported an enhancement in the quality of their treatment, as seen in the positive health outcomes associated with interventions such as web-based coaching programs [26]. However, variations in effectiveness were observed, as seen in the study by Wagner et al [40], where no impact on blood pressure was observed in the overall cohort, but active PHR users showed a significant reduction in diastolic blood pressure.

Similarly, several studies highlighted the improvement in health conditions and the quality of health care and life through patient access to EHRs [26,27,31,35]. Access through different platforms, including web-based coaching programs and patient portals, demonstrated that positive impacts on patient care resulted from treatment adherence and communication with HCPs [27]. However, it is crucial to note that not all interventions demonstrated uniform positive effects. For instance, the study by Wagner et al [40] found no impact of PHRs on blood pressure. Furthermore, some studies emphasized the significance of functions tailored to personalized preferences [25,34,42]. Patient portals positively influenced health outcomes by providing personalized reminders for appointments and screenings [25]. The IPHR showcased its capability to offer personalized recommendations, and usability tests on PHRs provided insights into the importance of tailoring functions to individual preferences [34,42].

Two studies highlighted the achievement of health-related goals through patient access to EHRs [34,42]. The IPHR effectively engaged patients in preventive care, conducting a concise health risk assessment that considers various aspects of patients' daily lives, including physical, mental, and social well-being [42]. The MHV program, integrated into PHRs, also facilitated patients in adapting to specific health-related goals [34]. Furthermore, notably, the prompt delivery of results through web-based access to laboratory test results contributed to higher satisfaction levels among service or EHRs users [39].

One study indicated that accessing EHRs through patient portals positively impacted the effective use of health care resources by extending doctor's visits, which ultimately improved the effectiveness of interactions with HCPs [25]. This suggests that patient access to EHRs may play a role in optimizing the use of health care resources during medical appointments. Furthermore, efficient use of health care and information resources is identified as another dimension influenced by patient access to EHRs.

A study by Krist et al [42] demonstrated that the IPHR, using EMRs, established an open database connection to link with the EMR of the patient's designated clinician. This connection enabled patients to access critical clinical data, including medical history, medications, immunizations, and test dates. The platform empowered patients to actively engage in the review, correction, and updating of their health information as needed.

XSL•FO
RenderX

This emphasizes the potential of patient access to EHRs in achieving efficient use of health care and information resources.

Focusing on the use of preventive care services, Krist et al [42] highlighted that, through the use of EMRs and the IPHR, patients were effectively engaged in preventive care. This involvement enabled patient access to results related to preventive care, with a use rate ranging from 1.5% to 28.3% across different health care practices. This suggests that patient access to EHRs may contribute to increased participation in preventive care services, promoting a proactive approach to health care management.

Notably, given the focus of the included studies on the use of health information resources, there is a lack of studies addressing the effectiveness of health care resource use and adherence to follow-up appointments, addressing this nuanced aspect of patient engagement that requires further exploration.

Importantly, some of the included studies highlighted some usability concerns and barriers related to patient access to and use of EHRs. One study emphasized the need for PHR registration to strike a balance between simplicity, security, and privacy [34]. Participants using EHRs via patient portals expressed concerns about the potential damage or hacking of their web-based records [25]. Technical barriers were also reported in 3 studies [25,28,30]. Consumers expressed apprehensions about obstacles to use, including intricate visual layouts and suboptimal usability features [25]. Patient access to documents in health records through the patient portal was considered intricate [28].

Usability challenges encompassed EHR integration, concerns with the EHR-portal communication interface, and issues related to browser access and compatibility, affecting both patients and HCPs [30]. In addition, barriers in the comprehension of health information and medical terms were reported in several studies [25,28,29]. Consumers expressed concerns about the complexity of language used in EHRs [25]. Challenges included patients grappling with medical terminology and the necessity for rigorous inspection of medical documents [28,29], which hindered their engagement with health care services. One study [39] demonstrated that there was no notable distinction between the groups (comprising those with web-based access to laboratory test results via EHRs and the no access control group) concerning the levels of anxiety reported following the receipt of test results through EHRs.

All these EHR usability–related concerns align with the 8 reasons identified by Valeur et al [47] for participants' hesitance to embrace patient-accessible EHRs. Patients expressed a variety of concerns, including finding PAEHRs unnecessary; preferring oral communication with HCPs; difficulty understanding the medical terminology in the records; dissatisfaction with the records' focus solely on disease; concerns about emotional reactions; perceiving it as an additional burden on their role as a patient; feeling that it was cumbersome, particularly among those lacking digital competence; and expressing skepticism about the overall impact of digital transformation on individual and social life.

As a result, there is a scarcity of research examining the possibility that EHR access could heighten feelings of uncertainty and anxiety when patients encounter clinical information that is unclear, particularly in the context of severe clinical conditions [5]. However, it is noteworthy that such an impact was not observed in the studies included in our analysis.

Interestingly, the diverse array of study designs, settings, and sample demographics observed across the 18 included studies underscores the multifaceted nature of patient engagement research, providing an excellent opportunity to highlight gaps in the existing literature and emphasize the unique contribution of our study. While observational cross-sectional studies dominate the literature, revealing prevalent trends, they inherently lack the capacity to establish causation over time [32,34,37]. By contrast, the presence of RCTs and mixed methods designs showcases the field's commitment to rigorous investigation, yet limitations such as ethical concerns and challenges in real-world applicability persist [26,30,33,34,36,42]. The inclusion of case reports, while offering unique insights, raises questions about generalizability [38].

Regarding the reliability of the positive impact of EHRs on patient engagement that was reported in most of the included studies, the critical evaluation of this evidence reveals a spectrum of strength influenced by study designs, sample sizes, and primary outcome measures. Robust evidence emerged from large-scale RCTs [26,33,40], nonrandomized [31,39,41], or descriptive studies [27,32,37,38] with large number of participants, where patient engagement serves as the primary outcome. These studies offer statistically significant insights, enhancing the reliability of reported outcomes. In contrast, qualitative studies [25,28,35], though rich in depth, exhibit a lower strength of evidence due to smaller sample sizes and exploratory nature. Furthermore, the variability in methodologies and the potential for publication bias, which were acknowledged in some studies [34,35], necessitate a nuanced interpretation.

Despite demographic diversity, with variations in age, sex, and education levels, certain populations remain underexplored, and the need for targeted interventions in EHRs access areas becomes evident [25-42]. Moreover, the diverse clinical settings, from dermatological outpatient clinics to primary care practices and academic medical centers, highlight the variability in contextual factors influencing patient engagement interventions [28,32,34,42]. Identified gaps include the limited representation of specific age groups or cultural backgrounds, a scarcity of longitudinal studies to assess sustained effects, and a dearth of real-world implementation studies. In addressing these gaps, our study contributes novel insights and practical solutions for advancing patient engagement research related to access to patient EHRs. By examining various types, designs, settings, and demographics in our included studies, we offer a more comprehensive understanding of EHR effectiveness for patient health care engagement across diverse health care settings.

It is important to note the "supply side" of patient engagement issues because the success or failure of this engagement depends on the role played by HCPs, HIT teams, and health systems regarding EHRs. The implementation and support of EHRs

XSL•FO
RenderX

present multifaceted challenges for these parties. HCPs contend with an onerous administrative burden, grappling with the time-consuming demands of data entry, documentation requirements, and regulatory compliance. Concurrently, HIT teams navigate technical complexities, encompassing software customization, data security, and interoperability, amidst an ever-evolving landscape of health care technology. Resource constraints within health systems, including financial limitations and staffing shortages, further exacerbate the challenges of investing in EHR infrastructure, providing adequate training, and addressing emerging implementation issues. Compounding these challenges is user resistance, as HCPs and staff may exhibit skepticism toward the benefits of EHRs and apprehension about workflow disruptions. Moreover, achieving seamless interoperability and data exchange between different EHR systems remains challenging due to incompatible standards, data silos, and privacy concerns. Addressing and actively mitigating these challenges necessitates a collaborative effort among health care stakeholders, encompassing policy makers, health care organizations, technology vendors, and clinicians for effective EHR implementation and support, thus enhancing patient health care engagement and delivery.

In addition, administrative burden relative to patient portals and EHRs is different among different health systems. Many studies included in this systematic review are from countries with national health systems. The clinical workflows and administrative burdens in these countries can differ significantly from those in the United States, where most included studies (8/18, 44%) were conducted. In the United States, HCPs often face a high administrative burden due to extensive documentation requirements, data entry, and regulatory compliance. This burden is exacerbated by the complexity of the US health care system, which involves multiple payers and fragmented care delivery.

In contrast, several studies come from countries with national health systems, including Sweden, Canada, Australia, and the Netherlands. These countries often experience more streamlined administrative processes. For example, Sweden's national health system emphasizes integrated care and standardized EHR systems, which can reduce the administrative burden on clinicians. Similarly, Canada's single-payer system and Australia's national health infrastructure support more cohesive HIT integration, easing the documentation load for HCPs.

Overall, while the administrative burden associated with EHRs and patient portals can be substantial in the United States, it may be lower in countries with more integrated and streamlined HIT processes, national-level support, and fewer regulatory complexities. Addressing these differences is crucial for understanding the global implementation and impact of EHRs and patient portals on health care engagement.

In addition, the use of EHRs is influenced by cultural practices and health care resources across diverse health care settings. Cultural norms regarding privacy, patient autonomy, and technology shape patient engagement with EHRs, emphasizing the need for culturally aligned systems. Disparities in health care resources, including technological access, contribute to variations in EHR availability, underscoring the importance of equitable access. In addition, practice cultures influence EHR adoption, with collaboration and continuous learning essential for optimizing outcomes. Addressing these factors can enhance EHR use and health care engagement globally.

Strengths of the Study

This study exhibits several notable strengths that contribute to the robustness and reliability of its findings. First, the comprehensive nature of our systematic review, which included 18 diverse studies addressing diverse EHRS-related applications and platforms and using various methodologies or study types (quantitative, qualitative, and mixed methods), ensures a thorough exploration of patient engagement in the context of EHRs. In addition, the inclusion of different study designs (ranging from observational cross-sectional studies to RCTs) enriches the depth and breadth of our analysis. Furthermore, our study meticulously examined the variations in patient demographics across these diverse studies, providing valuable insights into the generalizability of findings across different populations. The systematic approach used in our review, adhering to established protocols and guidelines (ie, PRISMA) enhances the transparency and validity of the synthesis. In addition, by identifying gaps and variations in the existing literature, our study sets the stage for future research endeavors to address these nuances and contribute to the evolving field of patient engagement with EHRs.

Moreover, our research distinguishes itself by providing a more comprehensive array of outcome measures for patient health care engagement. Unlike the included studies, which predominantly offered 1 or 2 measures, our study adopts a more holistic approach. By encompassing a broader spectrum of outcomes, we aim to capture the multifaceted nature of patient engagement, contributing a nuanced perspective to the existing body of literature. This approach ensures a more thorough understanding of the impact of patient access to EHRs on diverse facets of health care engagement.

Limitations of the Study

It is crucial to acknowledge certain constraints inherent in this systematic review. Publication bias may influence the results, with a potential inclination toward publishing studies with positive outcomes. The diverse methodologies and designs among the included studies introduce variability in evidence quality, challenging uniform analysis. Constraints include the limited availability of studies focusing on specific dimensions of patient engagement, impacting the review's comprehensiveness. Variations in sample sizes, demographics, and clinical settings may also affect the generalizability of findings. In addition, the dynamic nature of health care systems and technological advancements could render some findings time sensitive. Moreover, the use of 4 databases, while relevant to health care publications, may introduce the inadvertent inclusion of unrelated studies. Despite a comprehensive search strategy, PsycINFO yielded a limited number of results (n=18), potentially influencing the diversity of included studies and impacting the overall synthesis. These limitations underscore the importance of interpreting findings with consideration for the scope and potential variations in search outcomes.

XSL•FO
RenderX

Importantly, while the included studies provided valuable insights into the impact of interventions involving patient access to EHRs on patient engagement, it is important to acknowledge a limitation inherent in the design of these interventions. Many of the identified studies encompass multifaceted interventions, incorporating various components alongside patient access to EHRs. The holistic evaluation of these interventions poses a challenge in isolating and quantifying the distinct contribution of EHR access to observed improvements in patient engagement. This limitation implies that the specific impact of patient access to EHRs, when considered in isolation, remains unclear.

Implications for Future Research

Clinical Implications

The findings of this systematic review carry significant clinical implications for health care practitioners and HCPs. Understanding the positive impact of patient engagement with EHRs on various dimensions, including improved adherence to treatment plans, enhanced communication, and increased patient satisfaction, underscores the potential for integrating EHRs as a crucial component of patient care. Clinicians can leverage these insights to encourage and facilitate patient access to EHRs, promoting active involvement in their health care journey. Efforts to enhance user-friendly interfaces and address privacy concerns can further optimize the clinical utility of EHRs. Moreover, the identified gaps in the literature suggest avenues for targeted interventions to address specific dimensions of patient engagement that may currently be underexplored in clinical practice.

The clinical implications may extend to specific patient populations and health care settings, such as patients with long-term conditions and chronic disease in primary care in the United Kingdom, emphasizing the substantial potential for continuity of care and enhanced outcomes through EHR accessibility. Furthermore, patient use of EHRs can benefit the health care system by contributing to alleviating the economic impact of chronic diseases by efficiently managing population health within the community. Therefore, EHRs presents an optimal platform for proactive engagement for both HCPs and patients. In addition, in universal health coverage systems, patient access to primary care EHRs contributes to a more holistic approach to health care. It is crucial to recognize that EHR accessibility requires careful consideration to avoid potential harms related to information exchange, emphasizing the need for a robust electronic bioinformatic initiative scheme.

Research Implications

The systematic review reveals several research implications that can guide future investigations in the realm of patient engagement with EHRs. The identified gaps in the literature, such as the limited focus on certain dimensions of patient engagement and variations in study methodologies, highlight areas where further research is warranted. Future studies should aim for more standardized approaches to enable meaningful comparisons across interventions and settings. In addition, the potential publication bias and time-sensitive nature of findings emphasize the need for continuous updates and assessments to capture the evolving landscape of health care technology. Researchers are encouraged to delve deeper into specific dimensions, populations, and health care settings that may not have been extensively studied. Advancements in technology and changes in health care delivery models should be taken into account to ensure the relevance and applicability of future research in enhancing patient engagement with EHRs.

Conclusions

This systematic review provides a comprehensive synthesis of existing literature on patient engagement with EHRs. The findings highlight the positive impact of EHR accessibility on various dimensions, including increased adherence to treatment plans, promoted involvement and empowerment, improved communication, and enhanced patient satisfaction. While acknowledging the limitations and variations in study methodologies, this review underscores the importance of EHRs as a valuable tool in promoting patient engagement and active participation in health care. The identified gaps in the literature present opportunities for future research to explore specific dimensions of patient engagement and optimize the clinical utility of EHRs. Overall, this study contributes to the evolving understanding of the role of EHRs in patient health care engagement, emphasizing the need for continued exploration and refinement in this critical area of health care delivery.

Acknowledgments

The Medical Subject Heading terms with the Boolean operators were developed with the help of a Teaching and Learning Librarian from the University of Manchester, Timothy Calvert. This study was completed as part of a PhD study at the University of Manchester. JA is funded by the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre. The views expressed in this manuscript are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health.

Conflicts of Interest

PW is a director and shareholder of a for-profit digital mental health company, CareLoop Health Ltd. The remaining authors declare no conflict of interest.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 29 KB - jmir_v26i1e56473_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Results of the search strategies used in Ovid (MEDLINE), Ovid (Embase), Ovid (PsycINFO), and EBSCOhost (CINAHL).
[DOC File , 54 KB - jmir_v26i1e56473_app2.doc]

Multimedia Appendix 3
The characteristics of the selected studies.
[DOC File , 131 KB - jmir_v26i1e56473_app3.doc]

Multimedia Appendix 4
Quality assessment of the included studies using the Mixed Methods Appraisal Tool.
[DOC File , 97 KB - jmir_v26i1e56473_app4.doc]

References

1. Smith EA, Lapinski J, Lichty-Hess J, Pier K. Using health information technology and data to improve chronic disease outcomes in federally qualified health centers in Maryland. Prev Chronic Dis 2016;13:160445 [FREE Full text] [doi: 10.5888/pcd13.160445] [Medline: 28033091]
2. Assiri G. The impact of patient access to their electronic health record on medication management safety: a narrative review. Saudi Pharm J 2022 Mar;30(3):185-194 [FREE Full text] [doi: 10.1016/j.jsps.2022.01.001] [Medline: 35498224]
3. Varroud-Vial M. Improving diabetes management with electronic medical records. Diabetes Metab 2011 Dec;37 Suppl 4:S48-S52. [doi: 10.1016/S1262-3636(11)70965-X] [Medline: 22208710]
4. Jung EY, Park DK, Kang HW, Lim YS. Personalized health management services based on personal health record (PHR). Stud Health Technol Inform 2013;192:956. [doi: 10.3233/978-1-61499-289-9-956] [Medline: 23920730]
5. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care 2021 Jun 02;46(2):192-204. [doi: 10.1080/17538157.2021.1879810] [Medline: 33840342]
6. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. Health Aff (Millwood) 2015 Dec;34(12):2174-2180. [doi: 10.1377/hlthaff.2015.0992] [Medline: 26561387]
7. Lye CT, Forman HP, Daniel JG, Krumholz HM. The 21st Century Cures Act and electronic health records one year later: will patients see the benefits? J Am Med Inform Assoc 2018 Sep 01;25(9):1218-1220 [FREE Full text] [doi: 10.1093/jamia/ocy065] [Medline: 30184156]
8. Davis S, Roudsari A, Courtney KL. Shared decision making via personal health record technology for routine use of diabetic youth: a study protocol. Stud Health Technol Inform 2017;235:63-67. [doi: 10.3233/978-1-61499-753-5-63] [Medline: 28423756]
9. Davis S, Roudsari A, Raworth R, Courtney KL, MacKay L. Shared decision-making using personal health record technology: a scoping review at the crossroads. J Am Med Inform Assoc 2017 Jul 01;24(4):857-866. [doi: 10.1093/jamia/ocw172] [Medline: 28158573]
10. Roehrs A, da Costa CA, Righi RD, de Oliveira KS. Personal health records: a systematic literature review. J Med Internet Res 2017 Jan 06;19(1):e13 [FREE Full text] [doi: 10.2196/jmir.5876] [Medline: 28062391]
11. Shimoga SV, Lu YZ. Role of provider encouragement on patient engagement via online portals. J Am Med Inform Assoc 2019 Oct 01;26(10):968-976. [doi: 10.1093/jamia/ocz026] [Medline: 30925585]
12. Tai-Seale M, Rosen R, Ruo B, Hogarth M, Longhurst CA, Lander L, et al. Implementation of patient engagement tools in electronic health records to enhance patient-centered communication: protocol for feasibility evaluation and preliminary results. JMIR Res Protoc 2021 Aug 26;10(8):e30431. [doi: 10.2196/30431] [Medline: 34435960]
13. Pawelek J, Baca-Motes K, Pandit JA, Berk BB, Ramos E. The power of patient engagement with electronic health records as research participants. JMIR Med Inform 2022 Jul 08;10(7):e39145. [doi: 10.2196/39145] [Medline: 35802410]
14. Noblin AM, Wan TT, Fottler M. The impact of health literacy on a patient's decision to adopt a personal health record. Perspect Health Inf Manag 2012;9(Fall):1-13. [Medline: 23209454]
15. Davis AM, Hanrahan LP, Bokov AF, Schlachter S, Laroche HH, Waitman LR. Patient engagement and attitudes toward using the electronic medical record for medical research: the 2015 greater plains collaborative health and medical research family survey. JMIR Res Protoc 2019 Mar 12;8(3):e11148. [doi: 10.2196/11148] [Medline: 30860485]
16. Talboom-Kamp E, Tossaint-Schoenmakers R, Goedhart A, Versluis A, Kasteleyn M. Patients' attitudes toward an online patient portal for communicating laboratory test results: real-world study using the eHealth impact questionnaire. JMIR Form Res 2020 Mar 04;4(3):e17060 [FREE Full text] [doi: 10.2196/17060] [Medline: 32024632]
17. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019 Apr 11;21(4):e12779. [doi: 10.2196/12779] [Medline: 30973347]

XSL•FO
RenderX

18. Tabche C, Raheem M, Alolaqi A, Rawaf S. Effect of electronic health records on doctor-patient relationship in Arabian gulf countries: a systematic review. Front Digit Health 2023;5:1252227 [FREE Full text] [doi: 10.3389/fdgth.2023.1252227] [Medline: 37877127]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906. [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
20. Petticrew M, Roberts H. Systematic Reviews in the Social Sciences: A Practical Guide. Hoboken, NJ: John Wiley & Sons; 2008.
21. Kitchenham B, Pretorius R, Budgen D, Pearl Brereton O, Turner M, Niazi M, et al. Systematic literature reviews in software engineering – a tertiary study. Inform Softw Technol 2010 Aug;52(8):792-805. [doi: 10.1016/j.infsof.2010.03.006]
22. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inform 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
23. Vusio F, Thompson A, Birchwood M, Clarke L. Experiences and satisfaction of children, young people and their parents with alternative mental health models to inpatient settings: a systematic review. Eur Child Adolesc Psychiatry 2020 Dec;29(12):1621-1633. [doi: 10.1007/s00787-019-01420-7] [Medline: 31637520]
24. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. J Health Serv Res Policy 2005 Jul;10 Suppl 1:6-20. [doi: 10.1258/1355819054308576] [Medline: 16053580]
25. Zarcadoolas C, Vaughon WL, Czaja SJ, Levy J, Rockoff ML. Consumers' perceptions of patient-accessible electronic medical records. J Med Internet Res 2013 Aug 26;15(8):e168. [doi: 10.2196/jmir.2507] [Medline: 23978618]
26. Wang L, He L, Tao Y, Sun L, Zheng H, Zheng Y, et al. Evaluating a web-based coaching program using electronic health records for patients with chronic obstructive pulmonary disease in China: randomized controlled trial. J Med Internet Res 2017 Jul 21;19(7):e264. [doi: 10.2196/jmir.6743] [Medline: 28733270]
27. Nazi KM, Hogan TP, McInnes DK, Woods SS, Graham G. Evaluating patient access to Electronic Health Records: results from a survey of veterans. Med Care 2013 Mar;51(3 Suppl 1):S52-S56. [doi: 10.1097/MLR.0b013e31827808db] [Medline: 23407012]
28. Suija K, Mardo LA, Laidoja R, Nahkur S, Parvelo A, Kalda R. Experiences and expectation with the use of health data: a qualitative interview study in primary care. BMC Prim Care 2022 Jun 23;23(1):159 [FREE Full text] [doi: 10.1186/s12875-022-01764-1] [Medline: 35739479]
29. Wass S, Vimarlund V, Ros A. Exploring patients' perceptions of accessing electronic health records: innovation in healthcare. Health Informatics J 2019 Mar;25(1):203-215. [doi: 10.1177/1460458217704258] [Medline: 28457195]
30. Spratt SE, Ravneberg D, Derstine B, Granger BB. Feasibility of electronic health record integration of a SMART application to facilitate patient-provider communication for medication management. Comput Inform Nurs 2022 Aug 01;40(8):538-546. [doi: 10.1097/CIN.0000000000000891] [Medline: 35234708]
31. van der Vaart R, Drossaert CH, Taal E, Drossaers-Bakker KW, Vonkeman HE, van de Laar MA. Impact of patient-accessible electronic medical records in rheumatology: use, satisfaction and effects on empowerment among patients. BMC Musculoskelet Disord 2014 Mar 26;15:102 [FREE Full text] [doi: 10.1186/1471-2474-15-102] [Medline: 24673997]
32. Klein TM, Augustin M, Kirsten N, Otten M. Attitudes towards using electronic health records of patients with psoriasis and dermatologists: a cross-sectional study. BMC Med Inform Decis Mak 2020 Dec 30;20(1):344. [doi: 10.1186/s12911-020-01302-y] [Medline: 33380329]
33. Ryu B, Kim N, Heo E, Yoo S, Lee K, Hwang H, et al. Impact of an electronic health record-integrated personal health record on patient participation in health care: development and randomized controlled trial of MyHealthKeeper. J Med Internet Res 2017 Dec 07;19(12):e401 [FREE Full text] [doi: 10.2196/jmir.8867] [Medline: 29217503]
34. Haggstrom DA, Saleem JJ, Russ AL, Jones J, Russell SA, Chumbler NR. Lessons learned from usability testing of the VA's personal health record. J Am Med Inform Assoc 2011 Dec;18 Suppl 1(Suppl 1):i13-i17 [FREE Full text] [doi: 10.1136/amiajnl-2010-000082] [Medline: 21984604]
35. Hanna L, Gill SD, Newstead L, Hawkins M, Osborne RH. Patient perspectives on a personally controlled electronic health record used in regional Australia. Health Inf Manag 2017 Jan;46(1):42-48. [doi: 10.1177/1833358316661063] [Medline: 27486184]
36. Fuller TE, Pong DD, Piniella N, Pardo M, Bessa N, Yoon C, et al. Interactive digital health tools to engage patients and caregivers in discharge preparation: implementation study. J Med Internet Res 2020 Apr 28;22(4):e15573 [FREE Full text] [doi: 10.2196/15573] [Medline: 32343248]
37. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 01;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
38. Wolff JL, Darer JD, Berger A, Clarke D, Green JA, Stametz RA, et al. Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. J Am Med Inform Assoc 2017 Apr 01;24(e1):e166-e172 [FREE Full text] [doi: 10.1093/jamia/ocw108] [Medline: 27497795]

XSL•FO
RenderX

39. Mák G, Smith Fowler H, Leaver C, Hagens S, Zelmer J. The effects of web-based patient access to laboratory results in British Columbia: a patient survey on comprehension and anxiety. J Med Internet Res 2015 Aug 04;17(8):e191. [doi: 10.2196/jmir.4350] [Medline: 26242801]
40. Wagner PJ, Dias J, Howard S, Kintziger KW, Hudson MF, Seol YH, et al. Personal health records and hypertension control: a randomized trial. J Am Med Inform Assoc 2012;19(4):626-634. [doi: 10.1136/amiajnl-2011-000349] [Medline: 22234404]
41. Ibrahim AA, Ahmad Zamzuri MI, Ismail R, Ariffin AH, Ismail A, Muhamad Hasani MH, et al. The role of electronic medical records in improving health care quality: a quasi-experimental study. Medicine (Baltimore) 2022 Jul 29;101(30):e29627. [doi: 10.1097/MD.0000000000029627] [Medline: 35905245]
42. Krist AH, Peele E, Woolf SH, Rothemich SF, Loomis JF, Longo DR, et al. Designing a patient-centered personal health record to promote preventive care. BMC Med Inform Decis Mak 2011 Nov 24;11:73. [doi: 10.1186/1472-6947-11-73] [Medline: 22115059]
43. Zhou J, Arriaga RI, Liu H, Huang M. A tale of two perspectives: harvesting system views and user views to understand patient portal engagement. In: Proceedings of the IEEE 10th International Conference on Healthcare Informatics. 2022 Presented at: ICHI 2022; June 11-14, 2022; Rochester, MN p. 373-383. [doi: 10.1109/ichi54592.2022.00059]
44. Mantwill S, Fiordelli M, Ludolph R, Schulz PJ. EMPOWER-support of patient empowerment by an intelligent self-management pathway for patients: study protocol. BMC Med Inform Decis Mak 2015 Mar 20;15:18 [FREE Full text] [doi: 10.1186/s12911-015-0142-x] [Medline: 25890197]
45. Toni E, Pirnejad H, Makhdoomi K, Mivefroshan A, Niazkhani Z. Patient empowerment through a user-centered design of an electronic personal health record: a qualitative study of user requirements in chronic kidney disease. BMC Med Inform Decis Mak 2021 Nov 24;21(1):329 [FREE Full text] [doi: 10.1186/s12911-021-01689-2] [Medline: 34819050]
46. Prashad R. The role of personal health record systems in chronic disease management. Stud Health Technol Inform 2017;234:275-279. [doi: 10.3233/978-1-61499-742-9-275] [Medline: 28186054]
47. Valeur HS, Lie AK, Moen K. Patient rationales against the use of patient-accessible electronic health records: qualitative study. J Med Internet Res 2021 May 28;23(5):e24090. [doi: 10.2196/24090] [Medline: 34047711]

Abbreviations

COPD: chronic obstructive pulmonary disease
EHR: electronic health record
EMR: electronic medical record
HCP: health care provider
HIT: health information technology
IPHR: interactive patient health record
MHV: MyHealtheVet
MMAT: Mixed Methods Appraisal Tool
PCEHR: personally controlled electronic health record
PHR: personal health record
PICO: population, intervention, comparator, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by A Mavragani; submitted 17.01.24; peer-reviewed by T Karhula, K Masters, S Babbott; comments to author 05.04.24; revised version received 15.04.24; accepted 23.08.24; published 20.11.24.

Please cite as:
Alomar D, Almashmoum M, Eleftheriou I, Whelan P, Ainsworth J
The Impact of Patient Access to Electronic Health Records on Health Care Engagement: Systematic Review
J Med Internet Res 2024;26:e56473
URL: https://www.jmir.org/2024/1/e56473
doi:10.2196/56473
PMID:

©Dalia Alomar, Maryam Almashmoum, Iliada Eleftheriou, Pauline Whelan, John Ainsworth. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patients' Use of Electronic Health Records Facilitates Patient-Centered Communication: Findings From the 2017 Health Information National Trends Survey

Suhwoo Ahn[1], PhD; Chul-joo Lee[2], PhD; Inhwan Bae[3], MA

[1]Hubbard School of Journalism and Mass Communication, University of Minnesota, Twin Cities, Minneapolis, MN, United States
[2]Department of Communication, Seoul National University, Gwanak-gu, Seoul, Republic of Korea
[3]Department of Communication, Cornell University, Ithaca, NY, United States

Corresponding Author:
Chul-joo Lee, PhD
Department of Communication
Seoul National University
504 IBK Communication Center
1 Gwanak-ro
Gwanak-gu, Seoul, 08826
Republic of Korea
Phone: 82 2 880 4105
Email: chales96@snu.ac.kr

Abstract

Background: Patient-centered communication refers to interaction between patients and health professionals that considers patients' preferences and empowers patients to contribute to their own care. Research suggests that patient-centered communication promotes patients' satisfaction with care, trust in physicians, and competence in their abilities to manage their health.

Objective: The study aims to explore the role of patients' use of electronic health records (EHRs) in promoting patient-centered communication. Specifically, we investigated how health information efficacy mediates the association of EHR use with patient-centered communication and whether and how the relationship between EHR use and health information efficacy varies according to patients' perceived social support levels.

Methods: We conducted mediation and multigroup analyses using nationally representative data from the Health Information National Trends Survey 5 cycle 1 conducted in the United States (N=3285). Among respondents, we analyzed those who received care from health professionals over the previous year (2823/3285, 85.94%).

Results: EHR use by patients was associated with high levels of health information efficacy (unstandardized coefficient=0.050, SE 0.024; *P*=.04). In turn, health information efficacy was positively related to patient-centered communication (unstandardized coefficient=0.154, SE 0.024; *P*<.001). The indirect pathway from EHR use to patient-centered communication, mediated by health information efficacy, was statistically significant (unstandardized coefficient=0.008, SE 0.004; *P*=.04). Among patients with high social support (2349/2823, 83.21%), EHR use was not significantly associated with health information efficacy (unstandardized coefficient=0.038, SE 0.026; *P*=.15), although health information efficacy was linked to high levels of patient-centered communication (unstandardized coefficient=0.151, SE 0.030; *P*<.001). The indirect relationship in this group was not significant (unstandardized coefficient=0.006, SE 0.004; *P*=.11). However, among those with low social support (474/2823, 16.79%), EHR use was positively associated with health information efficacy (unstandardized coefficient=0.155, SE 0.048; *P*=.001), which in turn relates to high levels of patient-centered communication (unstandardized coefficient=0.137, SE 0.050; *P*=.01). The indirect pathway was also significant (unstandardized coefficient=0.021, SE 0.010; *P*=.03).

Conclusions: Patients who use EHRs may build health information efficacy, which seems to promote communication between patients and health care providers. This indirect pathway was not detected among patients with high social support. However, among those with low social support, EHR use seems to enhance health information efficacy, which may in turn facilitate patient-centered communication. Given the nature of the dataset used, the findings of this study are more relevant to the United States than other contexts.

(J Med Internet Res 2024;26:e50476) doi:10.2196/50476

XSL•FO
RenderX

KEYWORDS

electronic health record; health information efficacy; patient-centered communication; social support; patient-centered care

Introduction

Background

Current health care aims to promote patient-centeredness as a key element of high-quality care [1]. A primary goal of patient-centered care is patient-centered communication [2]. Patient-centered communication refers to interaction between patients and health professionals that considers patients' preferences and empowers patients to contribute to their own care [1,3]. Specifically, patient-centered communication is defined by its functions: fostering relationships between patients and health professionals, exchanging information, responding to patients' emotional reactions, assisting patients in relieving uncertainty, involving patients with medical decision-making, and encouraging patients to manage their health [2-4]. Research suggests that patient-centered communication promotes patients' satisfaction with care [5], trust in physicians [6,7], and competence in their abilities to manage their health [8]. In addition, patient-centered communication plays a crucial role in preventing patients from believing health misinformation [9,10]. Because health misinformation is rampant on social media platforms [10,11] and misinformation is likely to be disseminated more rapidly than corrective information [12], health professionals today use various strategies to counter patients' misinformed beliefs [9].

Emerging evidence indicates that patients' use of electronic health records (EHRs) has the potential to promote patient-centered communication [4,13-17]. An EHR is an electronic version of a patient's personal health information that contains health conditions and medication history [4,18-20]. Health care providers generate health records, and patients can access this information from EHRs [4,18,19]. To our knowledge, however, no studies have investigated the underlying mechanism between EHR use and patient-centered communication.

This study suggests that patients' confidence in their abilities to obtain health information they want, termed health information efficacy [21], may play a role here. Because EHR use empowers patients to control their health status [16], we argue that EHR use is positively linked to health information efficacy, which may in turn be related to patient-centered communication. Moreover, we posit that EHR use is beneficial to patients with little social support. Given that EHRs may compensate for patients' lack of social support, we postulate that the relationship between use of EHRs and health information efficacy might be strong for patients who perceive little social support.

The Development of EHRs

EHRs and electronic medical records (EMRs) are the 2 main computerized systems used to share patient information [19]. Although EHRs and EMRs are sometimes used interchangeably, they are distinct from each other [22]. EMRs were developed to deliver and receive patient information among clinicians within a particular hospital [20,22,23]. Unlike an EMR system, EHRs allow health professionals to share patient information with health care providers in other institutions [19,20,22]. Moreover, EHRs are designed to be used by both health professionals and patients [19,20,22,23]. Thus, patients who use EHRs are able to access their health records, obtain health information, and communicate with health professionals [14,15,19,24].

While EHRs and EMRs are typically controlled by physicians, personal health records (PHRs) are primarily managed by patients themselves, and various entities, such as physicians, patients, and pharmacies, could enter information on PHRs [20,25,26]. That is, PHRs serve as a compilation of individual patients' health information over their lifetime [26]. Although PHRs allow patients to organize their own health information logically, health professionals raise concerns regarding the inaccuracy of health information within PHRs [20]. Therefore, some scholars argue that EHRs need to be combined with PHRs [26,27].

The Relationship Between the Use of EHRs and Patient-Centered Communication

It is not always easy for patients to get useful health information from the media and interpersonal sources. Understanding medical terminologies or jargon that health care professionals use requires basic health literacy and cognitive abilities [28]. Although the traditional media and the internet deliver a wealth of information, patients should spend much time finding proper health information [29]. In addition, not all patients have friends or family members who could provide useful health information [30,31].

However, EHRs enable patients to easily access health information they need [15,17,19,24]. By granting patients access to their personal medical history, including treatment plans, medications, radiology images, allergy information, laboratory results, and immunization dates [32], patients who use EHRs can obtain health information tailored to their needs and circumstances [14,15]. In addition, patients can efficiently communicate with health care professionals through the messaging features within EHR platforms [14,33]. For example, patients are able to send quick messages to health professionals and receive immediate responses through EHRs whenever they have questions about their health status [14,18]. Consequently, patients who use EHRs may perceive that acquiring necessary health information is not as difficult.

Self-efficacy theory posits that people who experience success develop confidence in their capabilities to attain goals [34]. In a health context, patients who acquire useful health information build health information efficacy, which refers to one's confidence in their abilities to get the health information they want [21]. This study proposes that EHR use may enhance health information efficacy. Patients who use EHRs are able to cultivate experiences of successfully obtaining the health information they need by accessing their personal health information, such as laboratory results or medications stored

XSL•FO
RenderX

within EHRs [14]. Through this process, patients may come to realize that it is not as difficult to obtain personal health information from EHRs. Moreover, enhanced confidence may lead patients to believe that acquiring general health information from other information sources, such as the media or interpersonal networks, is also less challenging than previously thought. In other words, continued access to personal health information from EHRs may foster patients' confidence in their abilities to obtain general health information from various information sources beyond EHRs.

Therefore, we advance the following hypothesis: EHR use will be positively related to health information efficacy (hypothesis 1).

Patients' efficacy beliefs about acquiring health information may be applied to situations where patients interact with health professionals. That is, patients with health information efficacy will have the confidence that they can acquire information from health care providers as well. Such confidence may allow patients to communicate efficiently with health care professionals for the following reasons. First, patients with health information efficacy will seek information and prepare a list of well-informed questions before consulting with physicians [35]. Physicians are likely to offer more information to patients who ask many questions than to those with few questions [36,37]. In the process of asking questions and getting answers, patients will communicate actively with their physicians. Second, patients with efficacy beliefs about obtaining health information may learn about their health while interacting with their physicians. Patients who are well-informed are likely to engage in shared decision-making with health professionals [38]. Accordingly, it is expected that patients with health information efficacy have patient-centered communication with health care professionals.

Hence, we posit the following hypothesis: health information efficacy will be positively related to patient-centered communication (hypothesis 2).

Taken together, it is proposed that health information efficacy will mediate the linkage between EHR use and patient-centered communication (hypothesis 3).

The Moderating Role of Social Support in Using EHRs

Social support is defined as individuals' perceptions that they have someone to rely on [39-41]. People feel supported when receiving assistance from significant others, such as family members, friends, or coworkers [40,42]. In a health context, supportive significant others provide patients with health information, such as coping strategies for stress [40] or advice on how to care for infants [43]. Moreover, social support provides patients with a source of efficacy beliefs about obtaining and understanding health information [41].

Research has shown that eHealth, defined as the use of information and communication technologies to improve health care [44-46], has the potential to assist patients in overcoming deficits in social support [47-50]. In this study, we argue that EHRs, a type of eHealth service, may compensate for patients' lack of social support.

EHRs may enable patients with little social support to build efficacy beliefs about obtaining health information. Patients who lack social support are less likely to have supportive others who offer essential health information and emotional support [40,48]. Those patients may receive useful health information through EHR use [15] because EHRs provide patients with health information in an edited form that is easier to understand [51]. In addition, patients can ask questions through EHRs whenever they encounter complicated terms in their medical records [14], which makes them feel comfortable about their care. Naturally, patients who lack social support will cultivate successful experiences of acquiring health information through EHRs. Such mastery experiences may develop patients' efficacy beliefs about getting health information. Thus, among patients with low social support, the linkage between the use of EHRs and health information efficacy may be strong.

On the contrary, the relationship between EHR use and health information efficacy may be weak among individuals with high social support. Because social support increases patients' confidence in obtaining and understanding health information [41], patients with high social support may already have a sufficient level of health information efficacy. Therefore, individuals with high levels of social support may gain limited benefits from EHR use when it comes to building confidence in their ability to acquire health information. That is, a ceiling effect may occur.

Thus, EHR use may narrow the gap in health information efficacy between those with high and low social support. Accordingly, we suggest the following hypothesis: the association of EHR use with health information efficacy will be stronger among those with low social support than those with high social support (hypothesis 4).

Methods

Study Participants

This study analyzed nationally representative survey data from cycle 1 of the fifth Health Information National Trends Survey (HINTS 5 cycle 1) conducted in the United States [52]. The National Cancer Institute in the United States managed the development of survey questions and survey administration [52]. Participants received the first mailing and reminder postcard along with a cover letter [52]. The cover letter states that the survey aims to understand individuals' use of health information. In addition, the cover letter emphasizes the voluntary nature of participation. The survey was conducted through mail among individuals aged ≥18 years in the United States from January 25, 2017, to May 5, 2017 [52]. On the basis of the formula of the American Association for Public Opinion Research response rate 2, the overall response rate was 32.39%. In total, 3285 people completed the survey [52]. Those who had not met health professionals for their own care during the past year, except for emergency room visits, were excluded from the analyses. Ultimately, of the 3285 cases, 2823 (85.94%) were included in the study.

XSL•FO
RenderX

Measures

EHR Use

On a 5-point scale (0=*0* to 4=*≥10 times*), respondents were asked, "How many times did you access your online medical record in the last 12 months?" (mean 0.63, SD 1.03) [53]. Multimedia Appendix 1 contains a copy of the survey, which includes all the questions used in the study.

Health Information Efficacy

Health information efficacy was measured on a 5-point scale (1=*completely confident* to 5=*not confident at all*) by asking participants to rate their confidence levels to obtain health-related advice or information when necessary [54-56]. Responses to the item were reversely coded so that higher scores indicate higher health information efficacy (mean 3.79, SD 0.92).

Patient-Centered Communication

Patient-centered communication was operationalized using 7 items adopted by several studies [3,57]. Respondents reported their experiences of patient-centered communication with health care professionals during the previous 12 months. Respondents were asked how often their health care providers did the following things on a 4-point scale (1=*always* to 4=*never*): (1) provided them with an opportunity to ask any question they had about health, (2) paid attention to their emotions and feelings, (3) engaged them in making decisions regarding their health care, (4) ensured that they understand things they should do to manage their health, (5) clarified things, (6) spent sufficient amount of time with them, and (7) assisted them in coping with uncertainty feelings regarding their health or medical care. Responses to the 7 items were reversely coded. Then, the recoded responses were averaged (Cronbach α=0.93; mean 3.39, SD 0.65).

Social Support

Social support items were adopted from the study by Emanuel et al [58]. Respondents were asked whether they had someone they could rely on for emotional support when they needed it, including talking over problems or assisting them in making difficult decisions. In addition, respondents were asked if they had any family members or friends to discuss their health. Each social support item was assessed on a 2-point scale (0=*no* and 1=*yes*). These 2 items showed a high correlation (r=0.52; P<.001). To investigate the moderating role of social support, we combined the responses from these 2 items into 1 dichotomous indicator. Participants who answered *yes* to both items were considered as those with high social support (2349/2823, 83.21%), while the others were classified as those with low social support (474/2823, 16.79%). Please note that classification into high and low social support groups is based on individuals' self-perceptions regarding social support.

Control Variables

We included 3 types of control variables. First, age (mean 56.99, SD 15.92 years), male or female (1068/2763, 38.65% male), race or ethnicity (2037/2646, 76.98% White), education (1=*<8 years* to 7=*postgraduate*; mean 4.98, SD 1.59), income per year (1=<US $9999 to 9=≥US $200,000; mean 5.54, SD 2.25), employment status (1366/2744, 49.78% employed), marital status (1458/2727, 53.47% married), place of birth (2417/2758, 87.64% born in the United States), and having children (637/2574, 24.75% having children aged <18 years) were included as control variables. Second, we controlled for health-related variables: general health status (1=*poor* to 5=*excellent*; mean 3.35, SD 0.95), cancer history (471/2801, 16.82% ever been diagnosed as having cancer), regular health provider (2165/2787, 77.68% have health professionals that they see most often), and health insurance (2698/2787, 96.81% covered by health insurance). Third, we controlled for health information seeking from any source (2287/2768, 82.62% ever sought health information) because health information seeking from health care professionals or the media could increase health information efficacy or enhance the perceived quality of communication with physicians.

Statistical Analyses

Mediation Analyses

We first examined whether health information efficacy mediates the relationship between EHR use and patient-centered communication. We conducted path analyses using Mplus (version 7.4) developed by Muthén and Muthén [59]. Our model included control variables significantly associated with endogenous variables at bivariate correlation analyses [60]: race or ethnicity, education, income per year, employment, place of birth, general health, regular health provider, health insurance, and health information seeking. The skewness and the kurtosis of the variables in the models indicate that all variables were normally distributed [61]. Unstandardized coefficients were reported following the recommendation of mediation analyses [62].

Following a methodology document made by the Health Information National Trends Survey (HINTS) [63], we used a sample weight and replicate weights. A sample weight corrects for oversampling and generalizes the result to the US population [63]. In addition, replicate weights allow researchers to recalculate SEs, which lowers the risk of committing a type 1 error [63].

Additional Analyses

One could argue that the directionality of the variables in our model may be reversed. That is, patient-centered communication might increase health information efficacy [57], or health information efficacy could promote EHR use. Through the model comparison analyses, we explored whether our hypothesized mediation model fits the data better than its competing models with different causal orders among variables [64,65]. Although a model comparison result does not support causal links among variables, we can compare the performance of each model.

As shown in Figure 1, we developed 5 alternative models. Model 2 assumes the indirect pathway from EHR use to health information efficacy through patient-centered communication. In model 3, EHR use mediates the relationship between patient-centered communication and health information efficacy. Model 4 posits that the relationship between patient-centered communication and EHR use is mediated by health information

XSL•FO
RenderX

efficacy. In model 5, EHR use is a mediator for the path from health information efficacy to patient-centered communication. Model 6 assumes the indirect pathway from health information efficacy to EHR use via patient-centered communication. All models included control variables significantly associated with endogenous variables in the bivariate correlation analyses [60]. In models 1 and 2, where EHR use is an endogenous variable, EHR use was significantly related to race or ethnicity, education, income per year, employment, place of birth, general health, regular health provider, health insurance, and health information seeking at a bivariate level. Thus, these variables were included as control variables. In models 3 and 4, where patient-centered communication is an endogenous variable, patient-centered communication was significantly associated with race or ethnicity, income per year, place of birth, general health, regular health provider, and health insurance at a bivariate level. Therefore, these variables were controlled. In models 5 and 6, where health information efficacy is an endogenous variable, health information efficacy was significantly related to education, income per year, employment, place of birth, general health, and health information seeking at a bivariate level. These covariates were controlled in each model. In addition, we used a sample weight and replicate weights following the HINTS methodology document [63]. Figure 1 provides detailed information about additional models.

Figure 1. Comparison of mediation models. Unstandardized coefficients are shown with SEs and *P* values in parentheses. Paths with solid lines are statistically significant, while those with dotted lines are insignificant. Control variables are not presented in the figure. EHR: electronic health record; PCC: patient-centered communication.

Models	Figures	Coefficients
Model 1	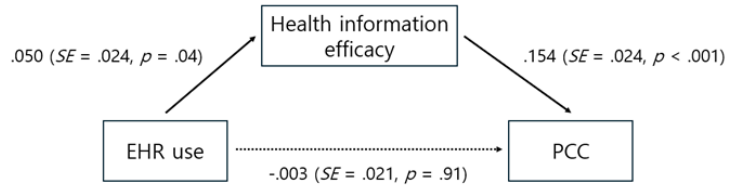	Total path: .005 (*SE* = .021, *p* = .81) Indirect path: .008 (*SE* =.004, *p* = .04) Direct path: -.003 (*SE* =.021, *p* = .91)
Model 2	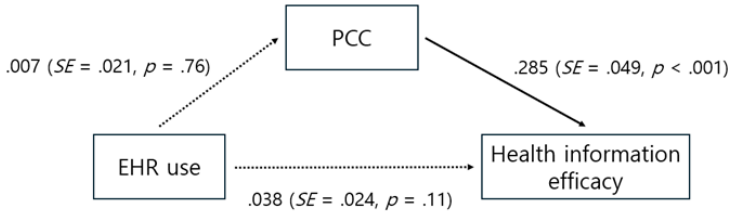	Total path: .040 (*SE* =.024, *p* = .09) Indirect path: .002 (*SE* =.006, *p* = .76) Direct path: .038 (*SE* =.024, *p* = .11)
Model 3	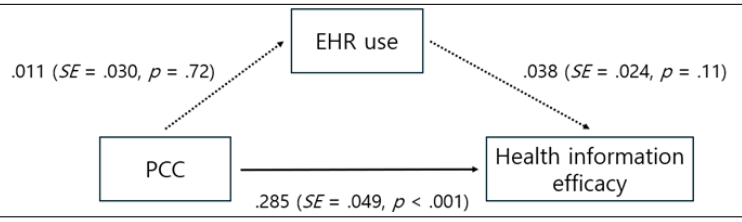	Total path: .286 (*SE* =.049, *p* < .001) Indirect path: .000 (*SE* =.001, *p* = .73) Direct path: .285 (*SE* =.049, *p* < .001)
Model 4	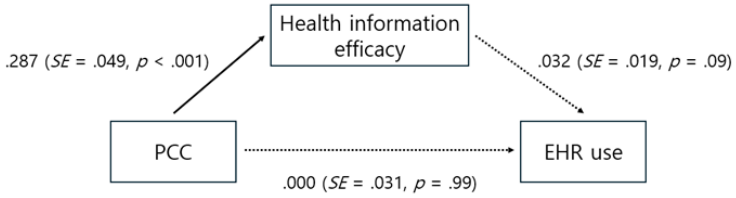	Total path: .010 (*SE* =.030, *p* = .75) Indirect path: .009 (*SE* =.005, *p* = .10) Direct path: .000 (*SE* =.031, *p* = .99)
Model 5	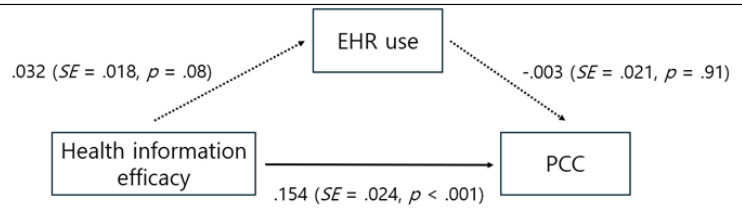	Total path: .154 (*SE* =.023, *p* < .001) Indirect path: .000 (*SE* =.001, *p* = .91) Direct path: .154 (*SE* =.024, *p* < .001)
Model 6	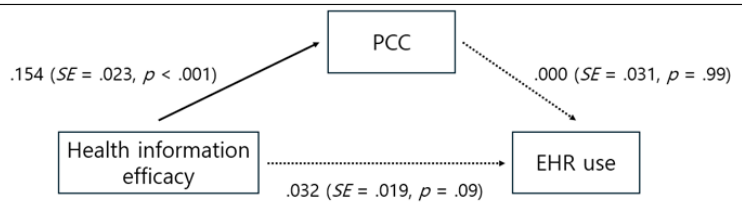	Total path: .032 (*SE* =.018, *p* = .08) Indirect path: .000 (*SE* =.005, *p* = .99) Direct path: .032 (*SE* =.019, *p* = .09)

We used standard criteria for measuring fits to find the best-fitting model: (1) Akaike information criterion, (2) Bayesian information criterion, (3) root mean square error of approximation (RMSEA), and (4) standardized root mean

XSL•FO
RenderX

squared residual (SRMR). Chi-square statistics were not used because Mplus does not provide it with replicate weights. We compared models based on each model's average ranking across all fit statistics. The ranks of the 4 fit statistics of each model were averaged. A model with a high mean fit rank was considered better than its alternatives.

Multigroup Path Analyses

Next, we investigated whether and how social support moderates the linkage between EHR use and health information efficacy. A multigroup path analysis was conducted. We tested the difference of coefficients of the path from EHR use to health information efficacy between those who have high and low levels of social support. If the difference was significant, it was considered that the association of EHR use with health information efficacy differs by social support. We included the same set of control variables: race or ethnicity, education, income per year, employment, place of birth, general health, regular health provider, health insurance, and health information seeking. In addition, a sample weight and replicate weights were applied following the methodology document by the HINTS [63].

Finally, we conducted post hoc power analyses of variables (EHR use, health information efficacy, and patient-centered communication) for the comparability of the groups (high and low social support groups). First, we calculated effect sizes (Cohen *d*) for each variable between high and low social support groups. Then, we used G*Power (version 3.1) software, developed by Faul et al [66], to compute statistical power analyses.

Ethical Considerations

This study analyzed data from the HINTS 5 cycle 1 survey. The Westat Institutional Review Board in the United States reviewed and approved the survey on March 28, 2016 (project number: 6048.14) [67]. Westat is a research firm that conducted the survey under contract with the United States Department of Health and Human Services. The survey received a "not human subjects research" determination from the National Institutes of Health Office of Human Subjects Research on April 25, 2016 (exempt number: 13204) [67]. Participants were invited to take the survey, receiving up to 3 mailings and 1 reminder postcard [63]. Each mail contains a cover letter, questionnaire, and return envelope. The cover letter informed participants that their participation was voluntary and their responses would not be linked to their names or any other information that could identify them or their households in accordance with the Privacy Act [63]. Survey data were stored with restricted access. Prepaid compensation of US $2 was provided to all potential participants to encourage participation [63].

Results

Descriptive Statistics

We presented descriptive statistics of all variables in our analyses in Table 1. Tables 2 and 3 also show bivariate correlations of the variables.

XSL•FO
RenderX

Table 1. Descriptive statistics.

	All respondents (N=2823)	Respondents with high social support[a] (n=2349)	Respondents with low social support[a] (n=474)
EHR[b] use[c], mean (SD)	0.63 (1.03)	0.65 (1.04)	0.52 (0.95)
HIE[d,e], mean (SD)	3.79 (0.92)	3.85 (0.89)	3.47 (1.02)
PCC[f,g], mean (SD)	3.39 (0.65)	3.45 (0.62)	3.12 (0.77)
Age (y), mean (SD)	56.99 (15.92)	56.93 (16.09)	57.10 (14.98)
Gender (male), n (%)	1068 (38.65)	884 (38.24)	171 (40.91)
Social support[a] (high), n (%)	2349 (83.21)	—[h]	—
Race or ethnicity, n (%)			
African American	486 (18.37)	396 (17.83)	83 (21.01)
American Indian or Alaska Native	103 (3.89)	84 (3.78)	19 (4.81)
Hispanic or Latinx	328 (12.78)	265 (12.31)	58 (14.99)
White	2037 (76.98)	1725 (77.67)	289 (73.16)
Education[i], mean (SD)	4.98 (1.59)	5.03 (1.59)	4.77 (1.56)
Income[j] (per year), mean (SD)	5.54 (2.25)	5.68 (2.22)	4.85 (2.26)
Employment status, n (%)			
Employed	1366 (49.78)	1154 (50.17)	197 (47.58)
Retired	900 (32.8)	767 (33.35)	122 (29.47)
Disabled	218 (7.94)	160 (6.96)	54 (13.04)
Homemaker	148 (5.39)	126 (5.48)	21 (5.07)
Unemployed	144 (4.15)	89 (3.87)	25 (6.04)
Marital status, n (%)			
Married	1458 (53.47)	1286 (56.3)	158 (38.44)
Divorced	435 (15.58)	327 (14.32)	90 (21.9)
Single, never been married	416 (15.25)	324 (14.19)	86 (20.92)
Widowed	283 (10.38)	233 (10.2)	47 (11.44)
Born in the United States, n (%)	2417 (87.64)	2030 (88.03)	359 (85.68)
Having children, n (%)	637 (24.75)	534 (24.68)	98 (25.39)
Health status[k], mean (SD)	3.35 (0.95)	3.40 (0.93)	3.08 (1.04)
Cancer history, n (%)	471 (16.82)	403 (17.23)	68 (15.96)
With health provider, n (%)	2165 (77.68)	1841 (79.39)	294 (69.01)
Have health insurance, n (%)	2698 (96.81)	2257 (97.24)	400 (94.12)
HIS[l] (seeking), n (%)	2287 (82.62)	1924 (83.47)	326 (77.99)

[a]Social support was operationalized by 2 items. Those who answered *yes* to both items were considered those with high social support. The others were regarded as those with low social support.

[b]EHR: electronic health record.

[c]EHR use was assessed on a 5-point scale: 0=*0* to 4=*≥10 times*.

[d]HIE: health information efficacy.

[e]HIE was measured on a 5-point scale: 1=*completely confident* to 5=*not confident at all*. Responses were reversely coded.

[f]PCC: patient-centered communication.

[g]Mean of 7 PCC items. Each item was measured on a 4-point scale: 1=*always* to 4=*never*. Responses were reversely coded.

[h]Not available.

[i]Education was assessed on a 7-point scale: 1=*<8 years* to 7=*postgraduate*. Responses were averaged.

XSL•FO
RenderX

[j]Income per year was measured on a 9-point scale: 1=<*US $9999* to 9=≥*US $200,000*. Responses were averaged.

[k]General health status was measured on a 5-point scale: 1=*poor* to 5=*excellent*.

[l]HIS: health information seeking.

XSL•FO
RenderX

Table 2. Bivariate correlations of variables.

	EHR[a]	HIE[b]	PCC[c]	Social support	Age	Male or female	Race or ethnicity	Education
HIE								
r	0.078	__[d]	—	—	—	—	—	—
P value	<.001	—	—	—	—	—	—	—
PCC								
r	0.066	0.242	—	—	—	—	—	—
P value	.01	<.001	—	—	—	—	—	—
Social support								
r	0.030	0.154	0.117	—	—	—	—	—
P value	.23	<.001	<.001	—	—	—	—	—
Age								
r	−0.004	−0.026	0.045	0.098	—	—	—	—
P value	.90	.50	.21	.006	—	—	—	—
Male or female								
r	−0.050	−0.008	0.002	−0.019	0.004	—	—	—
P value	.10	.79	.96	.58	.89	—	—	—
Race or ethnicity								
r	0.066	0.021	0.119	0.069	0.115	0.012	—	—
P value	.006	.49	<.001	.04	<.001	.59	—	—
Education								
r	0.199	0.066	0.063	0.126	−0.058	−0.038	0.166	—
P value	<.001	.03	.06	<.001	.13	.03	<.001	—
Income per year								
r	0.184	0.131	0.093	0.133	−0.010	0.088	0.117	0.361
P value	<.001	<.001	.01	<.001	.79	.01	<.001	<.001
Employment status								
r	0.094	0.089	0.015	0.046	−0.252	0.091	0.029	0.260
P value	<.001	.006	.66	.22	.001	.008	.30	<.001
Marital status								
r	0.130	0.037	0.035	0.093	0.203	0.056	0.140	0.118
P value	<.001	.25	.20	.003	<.001	.001	<.001	<.001
Born in the United States								
r	0.010	0.069	0.136	0.038	0.137	−0.011	0.304	0.007
P value	.70	.01	<.001	.33	<.001	.76	<.001	.84
Having children aged <18 years								
r	0.035	0.011	−0.018	−0.078	−0.219	−0.069	−0.030	0.015
P value	.18	.73	.53	.05	<.001	.02	.22	.67
General health								
r	0.043	0.153	0.147	0.213	−0.137	−0.003	0.082	0.267
P value	.12	<.001	<.001	<.001	.001	.94	.003	<.001
Cancer history								
r	0.030	−0.036	−0.004	0.006	0.241	−0.046	0.028	−0.024

XSL•FO
RenderX

	EHR[a]	HIE[b]	PCC[c]	Social support	Age	Male or female	Race or ethnicity	Education
P value	.19	.09	.82	.79	<.001	.03	.16	.13
Regular health provider								
r	0.192	0.034	0.194	0.063	0.217	−0.040	0.157	0.044
P value	<.001	.25	<.001	.05	<.001	.30	<.001	.13
Health insurance								
r	0.101	0.030	0.192	0.024	0.096	0.007	0.070	0.140
P value	<.001	.56	<.001	.50	.004	.88	.09	<.001
HIS[e]								
r	0.166	0.138	0.035	0.077	−0.010	−0.029	0.109	0.261
P value	<.001	.002	.20	.04	.81	.38	<.001	<.001

[a]EHR: electronic health record.

[b]HIE: health information efficacy.

[c]PCC: patient-centered communication.

[d]Not available.

[e]HIS: health information seeking.

XSL•FO
RenderX

Table 3. Additional bivariate correlations of variables.

	Income per year	Employment status	Marital status	Born in the United States	Having children aged <18 years	General health	Cancer history	Regular health provider	Health insurance
Income per year									
r	__[a]	—	—	—	—	—	—	—	—
P value	—	—	—	—	—	—	—	—	—
Employment status									
r	0.393	—	—	—	—	—	—	—	—
P value	<.001	—	—	—	—	—	—	—	—
Marital status									
r	0.344	0.160	—	—	—	—	—	—	—
P value	<.001	<.001	—	—	—	—	—	—	—
Born in the United States									
r	0.024	−0.079	−0.033	—	—	—	—	—	—
P value	.41	.007	.29	—	—	—	—	—	—
Having children aged <18 years									
r	0.149	0.195	0.210	−0.086	—	—	—	—	—
P value	<.001	<.001	<.001	.02	—	—	—	—	—
General health									
r	0.225	0.200	0.089	−0.003	0.034	—	—	—	—
P value	<.001	<.001	.007	.95	.23	—	—	—	—
Cancer history									
r	−0.004	−0.128	0.055	0.030	−0.077	−0.074	—	—	—
P value	.85	<.001	.27	.08	<.001	.001	—	—	—
Regular health provider									
r	0.060	−0.084	0.131	0.157	−0.075	−0.047	0.095	—	—
P value	.14	.03	<.001	<.001	.04	.17	<.001	—	—
Health insurance									
r	0.075	0.064	0.082	−0.003	−0.055	0.012	0.045	0.217	—
P value	.07	.03	.003	.92	.25	.69	.002	<.001	—
HIS[b]									
r	0.164	0.124	0.099	0.022	−0.017	0.070	0.057	0.043	0.074
P value	<.001	<.001	.007	.44	.60	.04	.001	.20	.12

[a]Not available.

[b]HIS: health information seeking.

Mediation Analyses

We conducted path analyses to investigate whether health information efficacy mediates the relationship between EHR use and patient-centered communication. Our mediation model fit the data well (RMSEA=0.015, 90% CI 0.000-0.031; SRMR=0.006). EHR use was positively associated with health information efficacy (unstandardized coefficient=0.050, SE 0.024; *P*=.04). In turn, health information efficacy was positively linked to patient-centered communication (unstandardized coefficient=0.154, SE 0.024; *P*<.001). The indirect pathway hypothesized in our study was statistically significant: health information efficacy mediated the association of EHR use with patient-centered communication (unstandardized coefficient=0.008, SE 0.004; *P*=.04).

Additional Analyses

Next, we conducted model comparison analyses. We compared our hypothesized mediation model (model 1) with its competing 5 models (models 2-6). Table 4 shows 4 fit statistics (Akaike information criterion, Bayesian information criterion, RMSEA, and SRMR) and the rank of each fit statistic. The average rank

XSL•FO
RenderX

of our hypothesized model (model 1) was 1.5, which is the best among the 6 plausible models.

Table 4. Comparison of the fit measures for the 6 models, with the rank order of fit indicated.

Models	AIC[a]		BIC[b]		RMSEA[c]		SRMR[d]		Rank, mean (SD)
	Values	Rank	Values	Rank	Values	Rank	Values	Rank	
Model 1[e]	68,812.127	1	69,311.553	1	0.015	1	0.006	3	1.5 (0.87)
Model 2[f]	68,813.358	2	69,312.785	2	0.017	6	0.006	3	3.25 (1.64)
Model 3[g]	72,461.655	6	73,044.319	6	0.016	5	0.006	3	5 (1.22)
Model 4[h]	72,461.418	5	73,044.082	5	0.015	1	0.006	3	3.5 (1.66)
Model 5[i]	72,461.026	3	73,043.690	3	0.015	1	0.005	1	2 (1.00)
Model 6[j]	72,461.054	4	73,043.718	4	0.015	1	0.005	1	2.5 (1.50)

[a]AIC: Akaike information criterion.

[b]BIC: Bayesian information criterion.

[c]RMSEA: root mean square error of approximation.

[d]SRMR: standardized root mean squared residual.

[e]Our hypothesized model.

[f]Electronic health record use→patient-centered communication→health information efficacy.

[g]Patient-centered communication→electronic health record use→health information efficacy.

[h]Patient-centered communication→health information efficacy→electronic health record use.

[i]Health information efficacy→electronic health record use→patient-centered communication.

[j]Health information efficacy→patient-centered communication→electronic health record use.

Multigroup Path Analyses

Furthermore, we conducted a multigroup path analysis to explore the moderating role of social support in the relationship between EHR use and health information efficacy. The model with a total sample showed a good fit (RMSEA=0.030, 90% CI 0.016-0.045; SRMR=0.010). In addition, the model fit the data well in each group: high social support group (RMSEA=0.030, 90% CI 0.015-0.046; SRMR=0.009) and low social support group (RMSEA=0.033, 90% CI 0.000-0.074; SRMR=0.013). The path coefficients of the association of EHR use with health information efficacy were significantly different between high and low social support groups (unstandardized coefficient=0.118, SE 0.058; *P*=.04).

Among patients with high social support, EHR use was not significantly related to health information efficacy (unstandardized coefficient=0.038, SE 0.026; *P*=.15). By contrast, health information efficacy was positively associated with patient-centered communication (unstandardized coefficient=0.151, SE 0.030; *P*<.001). The indirect pathway from EHR use to patient-centered communication through health information efficacy was not significant among those who have high levels of social support (unstandardized coefficient=0.006, SE 0.004; *P*=.11).

However, among patients with low social support, EHR use was positively related to health information efficacy (unstandardized coefficient=0.155, SE 0.048; *P*=.001), which in turn was positively associated with patient-centered communication (unstandardized coefficient=0.137, SE 0.050; *P*=.01). Health information efficacy was a significant mediator for the path from EHR use to patient-centered communication among those with low levels of social support (unstandardized coefficient=0.021, SE 0.010; *P*=.03). The relationship among variables with unstandardized coefficients is presented in Figure 2.

Finally, we conducted post hoc power analyses of variables to explore the comparability of high and low social support groups. Results show an effect size (Cohen *d*) of 0.126 for EHR use, 0.408 for health information efficacy, and 0.512 for patient-centered communication. In addition, findings from post hoc power analyses with a 2-tailed *t* test and an α level of .05 show that the power is 66.71% for EHR use, 100% for health information efficacy, and 100% for patient-centered communication.

XSL•FO
RenderX

Figure 2. A multigroup path analysis predicting high and low social support (SSP) groups' patient-centered communication (PCC). Unstandardized coefficients are shown with SEs and *P* values in parentheses. Coefficients for all participants are first reported, followed by coefficients for those with high SSP and low SSP. Paths with solid lines are statistically significant, while those with dotted lines are insignificant. In addition, bold lines indicate significant differences between those with high and low SSP. Control variables are not presented in the figure. The model fits the data well (root mean square error of approximation=0.030, 90% CI 0.016-0.045; standardized root mean squared residual=0.010). EHR: electronic health record.

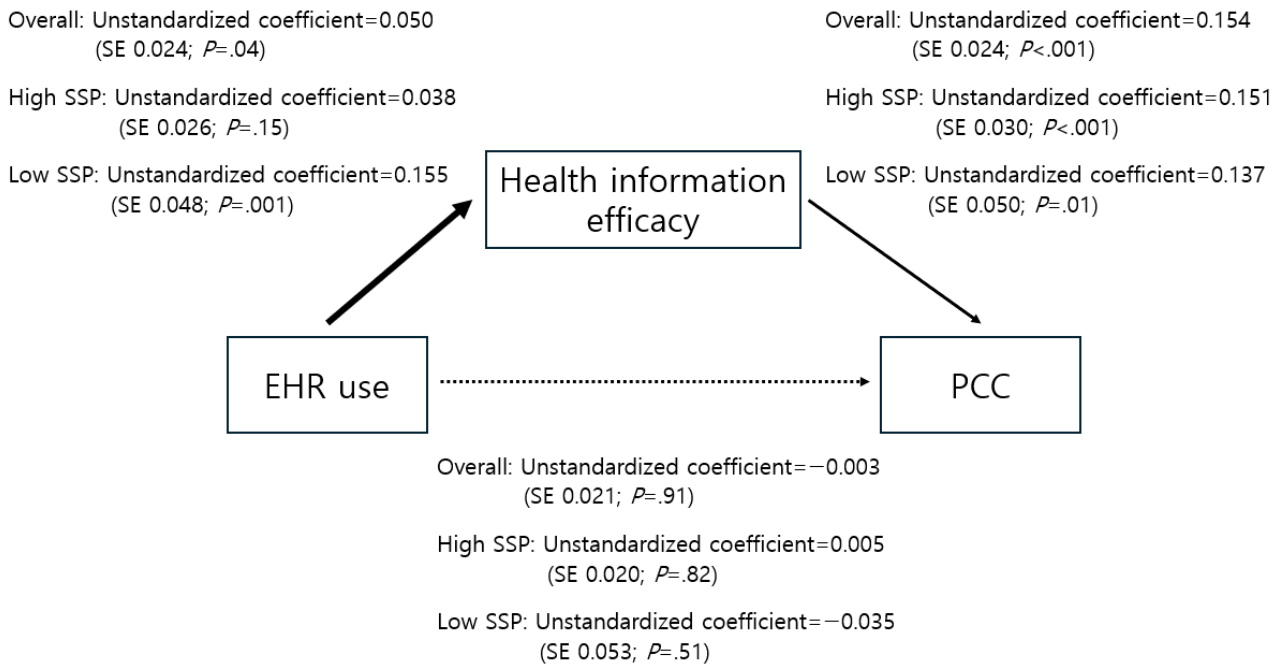

Discussion

Principal Findings

The existing literature pays little attention to the underlying mechanism of how patients' EHR use is beneficial to patient-centered communication. We found that health information efficacy, to some extent, accounts for the linkage between EHR use and patient-centered communication. The results indicate that EHR use seems to increase patients' beliefs that they can obtain necessary health information. In turn, patients with health information efficacy have patient-centered communication with health care professionals, probably because these patients receive the information they desire from physicians while asking questions and become informed of their health, and they thus easily participate in medical decision-making. In addition, EHR use was positively associated with health information efficacy, and we detected such a relationship only among patients who lack social support. Patients who feel they do not receive support from others may meet their needs through the use of EHRs.

Prior research in this field, most of which are qualitative studies, reported mixed findings regarding the relationship between EHR use and patient-centered communication. While EHRs enable patients to communicate with physicians effectively [14,17], there is a possibility that EHR use can interfere with patient-centered communication for certain individuals. One of the challenges associated with EHR systems is maintaining the privacy of patient information [19,68,69]. Concerns about the confidentiality of health records stored in EHRs may discourage patients from sharing their health status with physicians [2,68]. Such concerns might prevent patients from communicating with health professionals. Furthermore, some patients still struggle to understand medical terminologies or interpret test results in EHRs [70]. In addition, health care professionals often experience stress due to the high volume of messages they receive from patients through EHRs [71], which might compromise the quality of communication between physicians and patients. Thus, future studies should adopt a more nuanced approach by considering who uses EHRs in what contexts and what characteristics of EHRs could act as barriers to patient-centered communication. It would be worthwhile to explore further whether the impact of EHR use on patient-centered communication differs according to patients' levels of security concerns or physicians' proficiency in managing eHealth technologies.

Limitations

There are some limitations in this study. First, we could not establish a causal relationship among EHR use, health information efficacy, and patient-centered communication because the survey was cross-sectional. Second, we measured EHR use and health information efficacy using only 1 item. A single-item measure could result in measurement errors, which may reduce the chances of obtaining statistically significant associations. Third, unmeasured factors, such as patients' personality traits or medical mistrust, could affect patients' communication experiences with health professionals [1]. Fourth, this study relies on self-reported perceptions of patient-centered communication. There is the possibility of recall bias in the self-reported measures [72-74]. While self-reported measures of patient-centered communication assess patients' overall evaluations of their interactions with health professionals [3,72,75-77], these measures do not allow researchers to investigate what specific types of patients or health professionals' behaviors influence such assessments of

XSL•FO
RenderX

patient-centered communication [75]. To address these limitations, scholars have suggested the use of observational measures, such as training coders to review video recordings of patients consulting with health professionals [72,75,77]. Future research may benefit from using both self-reported and observational measures of patient-centered communication within the same study. Fifth, categorization into high and low social support groups relies on individuals' self-perceptions of social support. Because perceived social support measures may be influenced by perceptual or judgmental differences among individuals [78], future studies could use other measures of social support, such as received social support that captures specific behaviors or actions provided to individuals [78]. Sixth, previous studies referred to questions asking about patients' web-based medical record use in the HINTS as EHR use [53,79-82]. However, the HINTS items did not specifically ask what type of web-based health record patients accessed, specifically whether patients used EHRs, EMRs, or PHRs, which could limit the generalizability of the results. Finally, the findings of this study are relevant to the United States. Future studies could explore how the relationship between web-based medical record use and patient-centered communication varies across countries.

Implications

Our study has several theoretical implications. First, we empirically investigated the mechanism through which EHR use by patients could improve the quality of communication between health professionals and patients. Drawing from the self-efficacy theory, we explicate how health information efficacy links the relationship between EHR use and patient-centered communication. Moreover, this study suggests that EHRs can compensate for a lack of social support by enhancing beliefs in one's ability to acquire health information. Existing studies show that social capital, resources embedded within one's social network, could increase the use of diverse health information sources [83] and frequent health information seeking [30,31]. This means that the resources that are obtained through social relationships may have differential influences on health behaviors. However, our study suggests that EHR use may help bridge these gaps. Patients who lack social support could derive benefits from EHRs, which might improve the quality of patient-centered communication.

This study has practical implications as well. First, it is important for patients to develop efficacy beliefs about health information acquisition. According to the literature on self-efficacy, individuals with a high sense of efficacy set high goals and commit to them [34]. Patients with health information efficacy might aim to acquire health information from health providers and prepare for meetings with them. Thus, enhancing patients' health information efficacy should be a crucial element in patient education and public health interventions. By boosting efficacy beliefs, health educators and public health professionals may successfully enhance communication between patients and physicians.

Furthermore, EHR use is beneficial for patients, especially those with low social support, in building health information efficacy. However, patients' access to health records on the web remains low, although most hospitals in the United States have already adopted EHRs [84,85]. Similarly, of those who have met health professionals over the past year, 65.3% (1794/2748) of respondents have never had access to EHRs, according to the HINTS data we analyzed. Given that physicians' recommendation for EHR use is the main factor promoting patients' EHR use [53], health care providers should encourage patients to use EHRs.

Conclusions

To understand the pathway from patients' use of EHRs to patient-centered communication, we investigated the mediating role of health information efficacy. We used the data from the HINTS 5 cycle 1, which was a survey with a nationally representative sample, and the statistical weighting makes our results generalizable to the US population aged ≥18 years. Our findings indicate that EHR use by patients appears to promote health information efficacy, which may in turn facilitate patient-centered communication. This indirect relationship was not detected among patients with high social support. However, among patients with low social support, EHR use seems to enhance health information efficacy, which may subsequently promote patient-centered communication. These findings are more relevant to the United States than other contexts. In addition, using a cross-sectional survey prevents us from making strong causal claims about any relationship reported in our study.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF-2021S1A5B8096358) and the Institute of Communication Research at Seoul National University.

Data Availability

The datasets generated during and analyzed during this study are available in the Health Information National Trends Survey repository [52].

Authors' Contributions

The project was administered by SA and CL. The conceptualization of the study was carried out by all authors. All authors were responsible for the validation process. Data analyses and data curation were conducted by SA. The original draft of the manuscript was prepared by SA and IB. All authors reviewed and made edits to the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Copy of the survey including all the questions used in the study.
[DOCX File , 43 KB - jmir_v26i1e50476_app1.docx]

References

1. Epstein RM, Franks P, Fiscella K, Shields CG, Meldrum SC, Kravitz RL, et al. Measuring patient-centered communication in patient-physician consultations: theoretical and practical issues. Soc Sci Med 2005 Oct;61(7):1516-1528 [FREE Full text] [doi: 10.1016/j.socscimed.2005.02.001] [Medline: 16005784]
2. Rathert C, Mittler JN, Banerjee S, McDaniel J. Patient-centered communication in the era of electronic health records: what does the evidence say? Patient Educ Couns 2017 Jan;100(1):50-64 [FREE Full text] [doi: 10.1016/j.pec.2016.07.031] [Medline: 27477917]
3. Trivedi N, Moser RP, Breslau ES, Chou WY. Predictors of patient-centered communication among U.S. adults: analysis of the 2017-2018 Health Information National Trends Survey (HINTS). J Health Commun 2021 Jan 02;26(1):57-64 [FREE Full text] [doi: 10.1080/10810730.2021.1878400] [Medline: 33648425]
4. Butler JM, Gibson B, Lewis L, Reiber G, Kramer H, Rupper R, et al. Patient-centered care and the electronic health record: exploring functionality and gaps. JAMIA Open 2020 Oct;3(3):360-368 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa044] [Medline: 33215071]
5. Brédart A, Bouleuc C, Dolbeault S. Doctor-patient communication and satisfaction with care in oncology. Curr Opin Oncol 2005 Jul;17(4):351-354 [FREE Full text] [doi: 10.1097/01.cco.0000167734.26454.30] [Medline: 15933466]
6. Saha S, Beach MC. The impact of patient-centered communication on patients' decision making and evaluations of physicians: a randomized study using video vignettes. Patient Educ Couns 2011 Sep;84(3):386-392 [FREE Full text] [doi: 10.1016/j.pec.2011.04.023] [Medline: 21600723]
7. Jiang S, Wu Z, Zhang X, Ji Y, Xu J, Liu P, et al. How does patient-centered communication influence patient trust?: the roles of patient participation and patient preference. Patient Educ Couns 2024 May;122:108161 [FREE Full text] [doi: 10.1016/j.pec.2024.108161] [Medline: 38308973]
8. Liu PL, Zhang L, Ma X, Zhao X. Communication matters: the role of patient-centered communication in improving old adults' health competence and health outcomes. Health Commun 2024 Apr;39(2):363-375 [FREE Full text] [doi: 10.1080/10410236.2023.2166209] [Medline: 36628509]
9. Bylund CL, Mullis MD, Alpert J, Markham MJ, Onega T, Fisher CL, et al. Clinician communication with patients about cancer misinformation: a qualitative study. JCO Oncol Pract 2023 Mar;19(3):e389-e396. [doi: 10.1200/OP.22.00526] [Medline: 36626708]
10. Chou WY, Tiner J, Senft N. Emerging challenges in advanced cancer care: opportunities for enhancing patient-centered communication. In: Steel JL, Carr BI, editors. Psychological Aspects of Cancer: A Guide to Emotional and Psychological Consequences of Cancer, Their Causes, and Their Management. Cham, Switzerland: Springer; Jan 03, 2022.
11. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
12. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 09;359(6380):1146-1151 [FREE Full text] [doi: 10.1126/science.aap9559] [Medline: 29590045]
13. Benjamins J, Haveman-Nies A, Gunnink M, Goudkuil A, de Vet E. How the use of a patient-accessible health record contributes to patient-centered care: scoping review. J Med Internet Res 2021 Jan 11;23(1):e17655 [FREE Full text] [doi: 10.2196/17655] [Medline: 33427683]
14. Rief JJ, Hamm ME, Zickmund SL, Nikolajski C, Lesky D, Hess R, et al. Using health information technology to foster engagement: patients' experiences with an active patient health record. Health Commun 2017 Mar;32(3):310-319 [FREE Full text] [doi: 10.1080/10410236.2016.1138378] [Medline: 27223684]
15. Snyder CF, Wu AW, Miller RS, Jensen RE, Bantug ET, Wolff AC. The role of informatics in promoting patient-centered care. Cancer J 2011;17(4):211-218 [FREE Full text] [doi: 10.1097/PPO.0b013e318225ff89] [Medline: 21799327]
16. Urowitz S, Smith K, Alkazaz N, Apatu E, Quartey NK, Wiljer D. Patient accessible electronic health records for the chronically ill: a review of the literature. J Hosp Adm 2012 Sep 09;1(2):64 [FREE Full text] [doi: 10.5430/jha.v1n2p64]
17. Ndabu T, Agrawal L, Sharman R. The role of access type and age group in the breadth of use of patient portals: observational study. J Med Internet Res 2022 Dec 27;24(12):e41972 [FREE Full text] [doi: 10.2196/41972] [Medline: 36574284]
18. McGinn CA, Grenier S, Duplantie J, Shaw N, Sicotte C, Mathieu L, et al. Comparison of user groups' perspectives of barriers and facilitators to implementing electronic health records: a systematic review. BMC Med 2011 Apr 28;9:46 [FREE Full text] [doi: 10.1186/1741-7015-9-46] [Medline: 21524315]

XSL•FO
RenderX

19. McMullen PC, Howie WO, Philipsen N, Bryant VC, Setlow PD, Calhoun M, et al. Electronic medical records and electronic health records: overview for nurse practitioners. J Nurse Pract 2014 Oct;10(9):660-665 [FREE Full text] [doi: 10.1016/j.nurpra.2014.07.013]
20. Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: a more personalized healthcare and public health policy. Health Policy Technol 2017 Mar;6(1):20-25 [FREE Full text] [doi: 10.1016/j.hlpt.2016.08.002]
21. Basu A, Dutta MJ. The relationship between health information seeking and community participation: the roles of health information orientation and efficacy. Health Commun 2008;23(1):70-79 [FREE Full text] [doi: 10.1080/10410230701807121] [Medline: 18443994]
22. Garrett P, Seidman J. EMR vs EHR – what is the difference? Health IT Buzz. 2011 Jan 4. URL: https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference [accessed 2024-10-21]
23. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304 [FREE Full text] [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
24. van der Vaart R, Drossaert CH, Taal E, Drossaers-Bakker KW, Vonkeman HE, van de Laar MA. Impact of patient-accessible electronic medical records in rheumatology: use, satisfaction and effects on empowerment among patients. BMC Musculoskelet Disord 2014 Mar 26;15:102 [FREE Full text] [doi: 10.1186/1471-2474-15-102] [Medline: 24673997]
25. Electronic vs. personal health record. Harvard T.H. Chan. 2008. URL: https://www.hsph.harvard.edu/news/magazine/fall08ehrpersonalvshealth/ [accessed 2024-10-21]
26. Ruhi U, Chugh R. Utility, value, and benefits of contemporary personal health records: integrative review and conceptual synthesis. J Med Internet Res 2021 Apr 29;23(4):e26877 [FREE Full text] [doi: 10.2196/26877] [Medline: 33866308]
27. Agarwal R, Kochhar TA. A review of PHR, EMR and EHR integration: personalized healthcare and public health. Int J Inform Commun Comput Technol 2017;5(2):301-307 [FREE Full text]
28. Pieterse AH, Jager NA, Smets EM, Henselmans I. Lay understanding of common medical terminology in oncology. Psychooncology 2013 May;22(5):1186-1191 [FREE Full text] [doi: 10.1002/pon.3096] [Medline: 22573405]
29. Ishikawa Y, Kondo N, Kawachi I, Viswanath K. Are socioeconomic disparities in health behavior mediated by differential media use? Test of the communication inequality theory. Patient Educ Couns 2016 Nov;99(11):1803-1807 [FREE Full text] [doi: 10.1016/j.pec.2016.05.018] [Medline: 27349600]
30. Ahn S, Lee CJ, Ko Y. Network social capital and health information acquisition. Patient Educ Couns 2022 Sep;105(9):2923-2933 [FREE Full text] [doi: 10.1016/j.pec.2022.05.007] [Medline: 35637049]
31. Song L, Chang TY. Do resources of network members help in help seeking? Social capital and health information search. Soc Netw 2012 Oct;34(4):658-669 [FREE Full text] [doi: 10.1016/j.socnet.2012.08.002]
32. What are electronic health records (EHRs)? HealthIT.gov. URL: https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/what-are-electronic-health-records-ehrs [accessed 2024-10-21]
33. Kuo A, Dang S. Secure messaging in electronic health records and its impact on diabetes clinical outcomes: a systematic review. Telemed J E Health 2016 Sep;22(9):769-777 [FREE Full text] [doi: 10.1089/tmj.2015.0207] [Medline: 27027337]
34. Bandura A. Self-Efficacy: The Exercise of Control. New York, NY: W.H. Freeman and Company; 1997.
35. Bass SB, Ruzek SB, Gordon TF, Fleisher L, McKeown-Conn N, Moore D. Relationship of internet health information use with patient behavior and self-efficacy: experiences of newly diagnosed cancer patients who contact the National Cancer Institute's Cancer Information Service. J Health Commun 2006 Mar;11(2):219-236. [doi: 10.1080/10810730500526794] [Medline: 16537289]
36. Street RLJ. Information-giving in medical consultations: the influence of patients' communicative styles and personal characteristics. Soc Sci Med 1991;32(5):541-548. [doi: 10.1016/0277-9536(91)90288-n] [Medline: 2017721]
37. Street RLJ, Gordon HS, Ward MM, Krupat E, Kravitz RL. Patient participation in medical consultations: why some patients are more involved than others. Med Care 2005 Oct;43(10):960-969. [doi: 10.1097/01.mlr.0000178172.40344.70] [Medline: 16166865]
38. Lee CJ, Gray SW, Lewis N. Internet use leads cancer patients to be active health care consumers. Patient Educ Couns 2010 Dec;81 Suppl(1):S63-S69. [doi: 10.1016/j.pec.2010.09.004] [Medline: 20889279]
39. Cobb S. Social support as a moderator of life stress. Psychosom Med 1976;38(5):300-314 [FREE Full text] [doi: 10.1097/00006842-197609000-00003]
40. Thoits PA. Social support as coping assistance. J Consult Clin Psychol 1986 Aug;54(4):416-423. [doi: 10.1037//0022-006x.54.4.416] [Medline: 3745593]
41. Guan M, Han JY, Shah DV, Gustafson DH. Exploring the role of social support in promoting patient participation in health care among women with breast cancer. Health Commun 2021 Nov;36(13):1581-1589. [doi: 10.1080/10410236.2020.1773704] [Medline: 32500731]
42. Arora NK, Finney Rutten LJ, Gustafson DH, Moser R, Hawkins RP. Perceived helpfulness and impact of social support provided by family, friends, and health care providers to women newly diagnosed with breast cancer. Psychooncology 2007 May;16(5):474-486. [doi: 10.1002/pon.1084] [Medline: 16986172]
43. Leahy-Warren P, McCarthy G, Corcoran P. First-time mothers: social support, maternal parental self-efficacy and postnatal depression. J Clin Nurs 2012 Feb;21(3-4):388-397. [doi: 10.1111/j.1365-2702.2011.03701.x] [Medline: 21435059]

XSL•FO
RenderX

44. Borrelli B, Ritterband LM. Special issue on eHealth and mHealth: challenges and future directions for assessment, treatment, and dissemination. Health Psychol 2015 Dec;34S:1205-1208. [doi: 10.1037/hea0000323] [Medline: 26651461]
45. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20. [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
46. Matusitz J, Breen GM, Wan TT. The use of eHealth services in US nursing homes as an improvement of healthcare delivery to residents. Aging Health 2013 Jan 30;9(1):25-33 [FREE Full text] [doi: 10.2217/ahe.12.75]
47. Kim H, Faw M, Michaelides A. Mobile but connected: harnessing the power of self-efficacy and group support for weight loss success through mHealth intervention. J Health Commun 2017 May;22(5):395-402. [doi: 10.1080/10810730.2017.1296510] [Medline: 28339324]
48. Rains SA, Keating DM. The social dimension of blogging about health: health blogging, social support, and well-being. Commun Monogr 2011 Dec;78(4):511-534 [FREE Full text] [doi: 10.1080/03637751.2011.618142]
49. Rains SA, Meng J. Social enhancement and compensation in online social support among cancer patients: the role of social network properties. Health Commun 2022 Apr;37(4):490-497. [doi: 10.1080/10410236.2020.1853327] [Medline: 33272063]
50. Han JY, Kim E, Lee YI, Shah DV, Gustafson DH. A longitudinal investigation of empathic exchanges in online cancer support groups: message reception and expression effects on patients' psychosocial health outcomes. J Health Commun 2019;24(6):615-623. [doi: 10.1080/10810730.2019.1644401] [Medline: 31340721]
51. Earnest MA, Ross SE, Wittevrongel L, Moore LA, Lin CT. Use of a patient-accessible electronic medical record in a practice for congestive heart failure: patient and physician experiences. J Am Med Inform Assoc 2004;11(5):410-417. [doi: 10.1197/jamia.M1479] [Medline: 15187074]
52. What is HINTS? National Cancer Institute: Health Information National Trends Survey. URL: https://hints.cancer.gov/ [accessed 2024-10-21]
53. Strekalova YA. Electronic health record use among cancer patients: insights from the Health Information National Trends Survey. Health Informatics J 2019 Mar;25(1):83-90. [doi: 10.1177/1460458217704246] [Medline: 28434278]
54. Zhao X, Cai X. The role of risk, efficacy, and anxiety in smokers' cancer information seeking. Health Commun 2009 Apr;24(3):259-269. [doi: 10.1080/10410230902805932] [Medline: 19415558]
55. Patel V, Beckjord E, Moser RP, Hughes P, Hesse BW. The role of health care experience and consumer information efficacy in shaping privacy and security perceptions of medical records: national consumer survey results. JMIR Med Inform 2015 Apr 02;3(2):e14. [doi: 10.2196/medinform.3238] [Medline: 25843686]
56. Link E, Baumann E, Kreps GL, Czerwinski F, Rosset M, Suhr R. Expanding the health information national trends survey research program internationally to examine global health communication trends: comparing health information seeking behaviors in the U.S. and Germany. J Health Commun 2022 Aug 03;27(8):545-554. [doi: 10.1080/10810730.2022.2134522] [Medline: 36250315]
57. Yu L, Zheng F, Xiong J, Wu X. Relationship of patient-centered communication and cancer risk information avoidance: a social cognitive perspective. Patient Educ Couns 2021 Sep;104(9):2371-2377. [doi: 10.1016/j.pec.2021.02.004] [Medline: 33583647]
58. Emanuel AS, Kiviniemi MT, Howell JL, Hay JL, Waters EA, Orom H, et al. Avoiding cancer risk information. Soc Sci Med 2015 Dec;147:113-120. [doi: 10.1016/j.socscimed.2015.10.058] [Medline: 26560410]
59. Muthén MK, Muthén B. Mplus Version 8 User's Guide. Los Angeles, CA: Muthén & Muthén; 2017.
60. Little TD. Longitudinal Structural Equation Modeling. New York, NY: Guilford Publications; 2013.
61. Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol Method 1996;1(1):16-29 [FREE Full text] [doi: 10.1037//1082-989x.1.1.16]
62. Hayes AF. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. Commun Monogr 2009 Dec;76(4):408-420 [FREE Full text] [doi: 10.1080/03637750903310360]
63. Methodology reports. National Cancer Institute: Health Information National Trends Survey. URL: https://hints.cancer.gov/data/methodology-reports.aspx [accessed 2024-10-21]
64. Eveland WPJ, Hayes AF, Shah DV, Kwak N. Understanding the relationship between communication and political knowledge: a model comparison approach using panel data. Polit Commun 2005 Oct;22(4):423-446 [FREE Full text] [doi: 10.1080/10584600500311345]
65. Slater MD, Hayes AF, Ford VL. Examining the moderating and mediating roles of news exposure and attention on adolescent judgments of alcohol-related risks. Commun Res 2007 Aug 01;34(4):355-381 [FREE Full text] [doi: 10.1177/0093650207302783]
66. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191 [FREE Full text] [doi: 10.3758/BF03193146]
67. Institutional Review Board (IRB) approvals for the Health Information National Trends Survey (HINTS). National Cancer Institute: Health Information National Trends Survey. URL: https://hints.cancer.gov/about-hints/institutional-review-board.aspx [accessed 2024-10-21]
68. Stablein T, Loud KJ, DiCapua C, Anthony DL. The catch to confidentiality: the use of electronic health records in adolescent health care. J Adolesc Health 2018 May;62(5):577-582. [doi: 10.1016/j.jadohealth.2017.11.296] [Medline: 29422435]
69. Shenoy A, Appel JM. Safeguarding confidentiality in electronic health records. Camb Q Healthc Ethics 2017 Apr;26(2):337-341. [doi: 10.1017/S0963180116000931] [Medline: 28361730]

XSL•FO
RenderX

70. Freise L, Neves AL, Flott K, Harrison P, Kelly J, Darzi A, et al. Assessment of patients' ability to review electronic health record information to identify potential errors: cross-sectional web-based survey. JMIR Form Res 2021 Feb 26;5(2):e19074. [doi: 10.2196/19074] [Medline: 33635277]
71. Lieu TA, Altschuler A, Weiner JZ, East JA, Moeller MF, Prausnitz S, et al. Primary care physicians' experiences with and strategies for managing electronic messages. JAMA Netw Open 2019 Dec 02;2(12):e1918287. [doi: 10.1001/jamanetworkopen.2019.18287] [Medline: 31880798]
72. Blanch-Hartigan D, Chawla N, Moser RP, Finney Rutten LJ, Hesse BW, Arora NK. Trends in cancer survivors' experience of patient-centered communication: results from the Health Information National Trends Survey (HINTS). J Cancer Surviv 2016 Dec;10(6):1067-1077. [doi: 10.1007/s11764-016-0550-7] [Medline: 27193357]
73. Zaidi M, Amante DJ, Anderson E, Ito Fukunaga M, Faro JM, Frisard C, et al. Association between patient portal use and perceived patient-centered communication among adults with cancer: cross-sectional survey study. JMIR Cancer 2022 Aug 09;8(3):e34745 [FREE Full text] [doi: 10.2196/34745] [Medline: 35943789]
74. Nichols HM, Dababnah S, Berger Z, Long C, Sacco P. Can you hear me now? Effects of patient-centered communication with young adults aged 26 to 39. J Patient Exp 2021;8:23743735211033116 [FREE Full text] [doi: 10.1177/23743735211033116] [Medline: 34604509]
75. McCormack LA, Treiman K, Rupert D, Williams-Piehota P, Nadler E, Arora NK, et al. Measuring patient-centered communication in cancer care: a literature review and the development of a systematic approach. Soc Sci Med 2011 Apr;72(7):1085-1095. [doi: 10.1016/j.socscimed.2011.01.020] [Medline: 21376443]
76. Moser RP, Trivedi N, Murray A, Jensen RE, Willis G, Blake KD. Patient-Centered Communication (PCC) scale: psychometric analysis and validation of a health survey measure. PLoS ONE 2022 Dec 30;17(12):e0279725 [FREE Full text] [doi: 10.1371/journal.pone.0279725]
77. Street RLJ. Analyzing communication in medical consultations. Do behavioral measures correspond to patients' perceptions? Medical Care 1992;30(11):976-988 [FREE Full text] [doi: 10.1097/00005650-199211000-00002] [Medline: 1434961]
78. Haber MG, Cohen JL, Lucas T, Baltes BB. The relationship between self-reported received and perceived social support: a meta-analytic review. Am J Community Psychol 2007 Mar;39(1-2):133-144 [FREE Full text] [doi: 10.1007/s10464-007-9100-9] [Medline: 17308966]
79. Anthony DL, Campos-Castillo C. A looming digital divide? Group differences in the perceived importance of electronic health records. Inf Commun Soc 2015 Feb 03;18(7):832-846 [FREE Full text] [doi: 10.1080/1369118X.2015.1006657]
80. Campos-Castillo C, Anthony DL. The double-edged sword of electronic health records: implications for patient disclosure. J Am Med Inform Assoc 2015 Apr;22(e1):e130-e140 [FREE Full text] [doi: 10.1136/amiajnl-2014-002804] [Medline: 25059953]
81. Calixte R, Islam S, Osakwe ZT, Rivera A, Camacho-Rivera M. Pattern of use of electronic health record (EHR) among the chronically ill: a health information national trend survey (HINTS) analysis. Int J Environ Res Public Health 2021 Jul 07;18(14):7254 [FREE Full text] [doi: 10.3390/ijerph18147254] [Medline: 34299705]
82. Mahmood A, Mosalpuria K, Wyant DK, Bhuyan SS. Association between having a regular health provider and access to services linked to electronic health records. Hosp Top 2019 Jan;97(1):1-10. [doi: 10.1080/00185868.2018.1551102] [Medline: 30596471]
83. Chen W, Lee KH, Straubhaar JD, Spence J. Getting a second opinion: social capital, digital inequalities, and health information repertoires. J Assoc Inf Sci Technol 2014 Apr 30;65(12):2552-2563 [FREE Full text] [doi: 10.1002/asi.23130]
84. Parasrampuria S, Henry J. Hospitals' use of electronic health records data, 2015-2017. The Office of the National Coordinator for Health Information Technology. 2019. URL: https://www.healthit.gov/sites/default/files/page/2019-04/AHAEHRUseDataBrief.pdf [accessed 2024-10-21]
85. Johnson C, Richwine C, Patel V. Individuals' access and use of patient portals and smartphone health apps, 2020. Office of the National Coordinator for Health Information Technology. 2021. URL: https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2020 [accessed 2024-10-21]

Abbreviations

EHR: electronic health record
EMR: electronic medical record
HINTS: Health Information National Trends Survey
PHR: personal health record
RMSEA: root mean square error of approximation
SRMR: standardized root mean squared residual

XSL•FO
RenderX

Edited by A Coristine; submitted 11.07.23; peer-reviewed by L Freise, H Teza; comments to author 01.12.23; revised version received 22.02.24; accepted 25.09.24; published 25.11.24.

Please cite as:
Ahn S, Lee CJ, Bae I
Patients' Use of Electronic Health Records Facilitates Patient-Centered Communication: Findings From the 2017 Health Information National Trends Survey
J Med Internet Res 2024;26:e50476
URL: https://www.jmir.org/2024/1/e50476
doi:10.2196/50476
PMID:

©Suhwoo Ahn, Chul-joo Lee, Inhwan Bae. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association Between Patient Portal Engagement and Weight Loss Outcomes in Patients After Bariatric Surgery: Longitudinal Observational Study Using Electronic Health Records

Xinmeng Zhang[1], BS; Kaidi Kang[2], MS; Chao Yan[3], PhD; Yubo Feng[1], MS; Simon Vandekar[2], PhD; Danxia Yu[4], PhD; S Trent Rosenbloom[3,5], MD; Jason Samuels[6], MD; Gitanjali Srivastava[5,6,7,8], MD; Brandon Williams[6,8], MD; Vance L Albaugh[9], MD, PhD; Wayne J English[6,8], MD; Charles R Flynn[6,8], PhD; You Chen[1,3], PhD

[1]Department of Computer Science, Vanderbilt University, Nashville, TN, United States
[2]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[4]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[5]Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, TN, United States
[6]Department of Surgery, Vanderbilt University School of Medicine, Nashville, TN, United States
[7]Division of Diabetes, Endocrinology & Metabolism, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, United States
[8]Vanderbilt Weight Loss Center, Vanderbilt University Medical Center, Nashville, TN, United States
[9]Metamor Institute, Pennington Biomedical Research Center, Baton Rouge, LA, United States

Corresponding Author:
You Chen, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave.
Nashville, TN, 37203
United States
Phone: 1 615 343 1939
Email: you.chen@vumc.org

Abstract

Background: Bariatric surgery is an effective intervention for obesity, but comprehensive postoperative self-management is essential for optimal outcomes. While patient portals are generally seen as beneficial in engaging patients in health management, the link between their use and post–bariatric surgery weight loss remains unclear.

Objective: This study aimed to investigate the association between patient portal engagement and postoperative BMI reduction among patients after bariatric surgery.

Methods: This retrospective longitudinal study included patients who underwent Roux-en-Y gastric bypass or sleeve gastrectomy at Vanderbilt University Medical Center between January 2018 and March 2021. Patient portal engagement was measured during 4 stages: early (within 3 months after surgery), early midterm (3-6 months), late midterm (6-9 months), and late (9-12 months). Using generalized estimating equations, we estimated the associations between patients' portal engagements at these stages and the percentage of BMI reduction (%BMIR) at 3, 6, and 12 months after surgery. Covariates included duration since surgery, patient's age at the time of surgery, sex, race and ethnicity, type of bariatric surgery, severity of comorbid conditions, and socioeconomic disadvantage.

Results: The study included 1415 patients, predominantly female (n=1145, 80.9%), with a racial composition of 76.9% (n=1088) White and 19.9% (n=282) Black. Overall, 805 (56.9%) patients underwent Roux-en-Y gastric bypass and 610 (43.1%) underwent sleeve gastrectomy. By 1 year after surgery, the median %BMIR was 31.5% (IQR 25.2%-36.8%), and the median number of active days on the patient portal was 54 (IQR 33-80). Early portal engagement was significantly associated with %BMIR at various postoperative times. Specifically, each additional 10 days of early portal engagement was associated with a 0.37% (95% CI –0.55% to –0.18%; $P<.001$) lower expected %BMIR at 3 months, a 1.11% (95% CI 0.82%-1.41%; $P<.001$) higher expected %BMIR at 6 months, and a 0.78% (95% CI 0.25%-1.31%; $P=.004$) higher expected %BMIR at 12 months. Furthermore, early midterm portal engagement was associated with a 0.36% (95% CI –0.69 to –0.03; $P=.03$) lower expected %BMIR at 6 months,

XSL•FO
RenderX

but it was not significant at 12 months (*P*=.88). Late midterm and late portal engagement were not significantly associated with %BMIR at 12 months (*P*=.27 and *P*=.12, respectively). Furthermore, early engagement in various portal functions, such as messaging and accessing medical records, was significantly associated with a lower %BMIR at 3 months and a higher %BMIR at both 6 and 12 months (all *P*<.05).

Conclusions: Higher patient portal engagement within 3 months after surgery—suggestive of stronger adherence to postoperative instructions and improved communication with care teams—is associated with less favorable weight loss immediately after surgery but enhanced postoperative weight loss outcomes at 6 and 12 months. However, the limitations of retrospective data-driven studies highlight the need for future intervention-based studies to validate these associations and establish causality.

(J Med Internet Res 2024;26:e56573) doi:10.2196/56573

KEYWORDS

patient portal; patient engagement; bariatric surgery; retrospective; longitudinal study; data-driven; weight loss; obesity; electronic health records; postoperative care

Introduction

Obesity continues to be a significant health concern in the United States, with approximately 42% and 9.2% of adults classified as having obesity (BMI≥30 kg/m^2) and severe obesity (BMI≥40 kg/m^2), respectively [1]. Research has consistently demonstrated that individuals with obesity face an elevated risk for a range of health complications, including, but not limited to, cardiovascular disease, diabetes, respiratory problems, and metabolic syndrome [2-6]. Bariatric surgery remains the most effective intervention for individuals who struggle to achieve sustainable weight loss solely through diet and exercise, whether through its effects of restricting food intake or modifying the digestive system [7,8].

Comprehensive postoperative care following hospital discharge plays a pivotal role in achieving optimal outcomes and ensuring success for patients after bariatric surgery [9]. Effective health tracking, patient self-management, and efficient collaboration between patients and their care teams are of utmost importance in working toward shared goals, including sustained weight loss, improved health conditions, and enhanced overall quality of life. Such postoperative needs have been increasingly supported by both web-based and mobile app–based patient portals [10], which typically provide patients access to their own electronic health information and serve as platforms for timely communication and regular health information exchange with their care teams. Growing evidence underscores the efficacy of portal usage in delivering continuous support and guidance, monitoring key indicators related to primary interventions, enhancing medication adherence, and fostering self-management in promoting behavioral and lifestyle modifications [11-16]. Although the multifaceted benefits of patient portals in health care settings are well documented, the specific impact of portal use on BMI changes in patients after bariatric surgery remains underexplored. Understanding the time-varying association between portal engagement at various critical stages (ie, early, midterm, and late) and the percentage of BMI reduction (%BMIR) is crucial, as it can provide valuable insights into the optimal timing and types of portal interactions that most effectively support postsurgery weight management. By identifying the most influential timing for portal engagement, health care providers can tailor their digital interventions to maximize their impact, thus enhancing patient adherence, self-management, and ultimately, health outcomes.

In this study, we investigated the association between patient portal engagement at various stages—early (within 3 months after surgery), early midterm (3 to 6 months), late midterm (6 to 9 months), and late (9 to 12 months)—and the %BMIR at 3, 6, and 12 months after surgery. We hypothesized that portal engagement at these different stages has varying associations with %BMIR at various postoperative times. Should the hypothesis test results reveal significant differences in the association across these stages, such findings would prompt a re-evaluation of current postoperative care practices.

Methods

Ethical Considerations

This study was approved by the institutional review boards (IRBs) at Vanderbilt University Medical Center (VUMC) under IRB#221459. A full waiver of written informed consent from patients was granted by the IRB because this study is retrospective with minimal risks to patients and the study data are deidentified. This study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline.

Study Settings

We conducted a single-site, retrospective, longitudinal study at VUMC, a large academic medical center in Nashville, Tennessee, that provides primary and specialty referral care to patients from across the Southeastern United States. This study included patients who underwent Roux-en-Y gastric bypass (RYGB) or sleeve gastrectomy (SG) operations, the most commonly performed operations [17], between January 1, 2018, and March 1, 2021, inclusive, as part of their participation in the VUMC Surgical Weight Loss Program. The specific end date reflected the most current data update within the available dataset. The VUMC Surgical Weight Loss Program is accredited by the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) [18]. VUMC deployed its patient portal, My Health at Vanderbilt (MHAV), in 2004, and migrated to Epic's MyChart platform in late 2017 as part of a wider electronic health record (EHR) system update across VUMC. Like most patient portals, MHAV allows patients

XSL•FO
RenderX

to access their own electronic health information, make appointments, manage medications, and interact with their care providers through a secure messaging system [19]. MHAV currently has over 1 million users and is accessed over 30 million times annually.

Data

The cohort was obtained from the VUMC bariatric surgery Quality, Efficacy, and Safety registry, a VUMC-specific database used exclusively for internal research and quality enhancement. We considered initial RYGB or SG operations performed at VUMC and excluded any subsequent surgical revisions due to their inherent complexity. We identified the date of each patient's initial bariatric surgery (day 0) as the index event. Following the surgery, patients were scheduled for clinic follow-ups at 3, 6, and 12 months after surgery. We collected BMI data from the Quality, Efficacy, and Safety registry and EHR. The observations with missing weights at the follow-up visits at 3, 6, or 12 months after surgery and the patients who lacked weight records immediately before surgery (baseline) were excluded.

Patient demographic and clinical information were extracted from the Epic EHR system, which was deployed at VUMC in 2017. Demographic information includes age at surgery, sex (female or male), and self-reported race and ethnicity (Black, White, or other races or ethnicities) from the EHR. We grouped American Indian or Alaska Native, Asian, Asian Indian, Chinese, Cuban, Filipino, Guamanian or Chamorro, Hispanic or Latino, Japanese, Korean, Mexican, Mexican American or Chicano, Native Hawaiian, other Asian, other Pacific Islander, Puerto Rican, Samoan, Vietnamese, and none of the above into "other races or ethnicities" to avoid unstable estimates due to their small cohort sizes. Past medical conditions of each patient, encompassing a period of 10 years up to the index event, were extracted to calculate the Charlson Comorbidity Index (CCI) [20,21]. To account for socioeconomic status, we used the Area Deprivation Index (ADI) [22], a widely adopted measure indicating the neighborhood's socioeconomic disadvantage level. In our analysis, we determined patients' ADI by their respective 5-digit zip codes. All patients did not have missing data on the type of surgery, the time of surgery, sex, age, and zip code. We excluded patients who had missing records of race and ethnicity. Study cohort inclusion and exclusion criteria were detailed in Figure S1 in Multimedia Appendix 1.

We extracted patient portal engagement history from MHAV event logs, which record every action taken by system users through provided interfaces [23-28]. These included unique patient identifiers, event time stamps, and event types. For each patient, we focused on the time frame starting from day 5 after surgery—typically regarded as the commencement of the postdischarge phase—up to 12 months after surgery. We categorized event types into 7 portal functions based on previous literature, which are messaging (support communication between health care providers and patients), visits (facilitate appointment management), my record (provide lists of allergies, immunizations, medical history, medications, pharmacy, preventive care, test results, and vitals), medical tools (allow patients to view document and add devices), billing (support account payment and insurance management), resources (provide patient education materials), and others (include additional functions not specified above, such as "Send proxy invite") [23]. The full list of event types associated with these functions is reported in Table S1 in Multimedia Appendix 1.

Measure Definition

There is no universally agreed-upon measure of patient portal engagement. In this study, we quantified engagement by measuring the number of days in which a patient performed any action within MHAV after logging in, as recorded in the event logs of MHAV. Days that solely consisted of log-in and log-out events were not counted. As such, we use "active days of portal use" interchangeably to refer to this measurement. We tracked the number of active portal usage days at various postoperative stages to evaluate the effects of portal engagement at different stages over time. Specifically, we categorized the engagement as follows: "early engagement" from day 5 to 3 months after surgery, "early midterm engagement" from 3 to 6 months, "late midterm engagement" from 6 to 9 months, and "late engagement" from 9 to 12 months.

Postoperative weight loss outcome is defined as the %BMIR at month *m* compared with the baseline BMI (kg/m^2):

We selected %BMIR because it accounts for differences in baseline body weight and provides standardized comparisons across diverse patient populations. Compared with other weight loss measures, such as percentage of weight loss and percentage of excess weight loss, %BMIR provides a more sensitive and consistent indicator of surgery-induced weight loss [29,30]. The primary goal of this study is to examine the time-varying association between patients' portal engagements at different postoperative stages and %BMIR.

Statistical Analysis

The data analysis was performed between February 2023 and November 2023. A descriptive analysis was performed to summarize the demographic and clinical characteristics of the bariatric surgery cohort. We used generalized estimating equations [31] with an identity link function and an autoregressive correlation structure (ie, AR-1) for statistical modeling. The use of the AR-1 correlation structure allows for the consideration of the time-dependent nature of within-patient correlation in %BMIR, ensuring the closer sequential measurements are more strongly correlated. The independent and dependent variables are the number of active days of portal use at different stages (winsorized at 95th percentile to limit extreme values and reduce the influence of outliers) and the %BMIR, respectively. All independent variables were included in the same model. We included the following covariates in all the models to control possible confounding: the duration since surgery (in months), the patient's age at the time of surgery (in years), sex (female or male), race and ethnicity (Black, White, or other races or ethnicities), type of bariatric surgery (RYGB or SG), CCI score, and ADI. The covariates were chosen based on previous literature, which was detailed in the conceptual framework (Figure S2 in Multimedia Appendix 1). Furthermore,

XSL•FO
RenderX

to model the time-varying effects of patient portal engagement amount at different postoperative stages on the %BMIR, we included the interaction terms between portal engagement at different stages and the duration since surgery (in months). In addition to the multivariate analyses, the univariate generalized estimating equation models were also fitted to explore the effects of the early, early midterm, late midterm, and late engagements on the outcome, separately. There were no missing data for the covariates. All analyses used complete cases for %BMIR at the available follow-up time points.

We centered our analyses around the investigation of the association between patient portal engagement at various stages (ie, early, early midterm, late midterm, and late portal engagement) and the %BMIR at various postoperative times (ie, 3, 6, and 12 months after surgery). Our primary hypothesis is that portal engagement at these distinct stages has different associations with %BMIR at these respective time points. We also conducted a secondary analysis to test the relative influence of specific portal functions. As part of this, we restricted our analysis to the active days of each portal function separately. A sensitivity analysis was performed by adding BMI immediately before surgery (BMI at baseline) to the primary model to account for its potential mediation effect. In addition, considering the overlap of our study period with the COVID-19 pandemic, during which MHAV usage spiked as a portion of health care moved to virtual delivery [32], we conducted an additional sensitivity analysis by dividing the patient cohort based on whether their surgeries occurred before March 1, 2020. We then determined if the associations between portal engagement and %BMIR were consistent before and during the pandemic. Considering that older patients may experience different levels of success with digital tools compared with younger adults [33], we conducted a sensitivity analysis to assess the influence of age on the relationship between portal engagement at various postoperative stages and weight loss outcomes. This analysis included interaction terms between stages of portal engagement and patient age.

All statistical analyses were conducted using R software (version 4.3.1; R Core Team). A *P* value less than .05 was considered statistically significant. In addition to *P* values, effect size estimates were also reported using a robust effect size index (RESI) along with their 95% CIs [34,35]. The RESI is equal to 0.5 Cohen's *d* under some assumptions [36], so Cohen's *d*–suggested interpretations for effect size are as follows: none to small (RESI=0-0.1), small to medium (RESI=0.1-0.25), and medium to large (RESI=0.25-0.4). The direction of the RESI estimates indicates the association direction.

Results

Descriptive Analysis

The cohort of those who underwent bariatric surgery comprised 1415 patients, with 3377 observations within 1 year of follow-up visits (the study cohort flow diagram is illustrated in Figure S1 in Multimedia Appendix 1). In the included cohort, there were no missing data in the covariates. There were 805 (56.9%) patients who underwent RYGB and 610 (43.1%) who underwent SG. There were 1145 (80.9%) female patients (Table 1). The mean age as of the surgery date was 44.5 (SD 11.4) years, with a median baseline BMI of 45.5 (IQR 41.4-51.4) kg/m^2. After surgery, 95.1% (n=1345), 76.7% (n=1086), and 66.9% (n=946) of patients had a follow-up at the 3, 6, and 12 months, respectively. The median %BMIR at the 3, 6, and 12 months was 15.8% (IQR 13.7%-18.1%), 24.4% (IQR 20.4%-27.9%), and 31.5% (IQR 25.2%-36.8%), respectively. Over 98% of patients who underwent bariatric surgery used MHAV at least once after surgery. Patients had a median engagement of 23 (IQR 15-33), 35 (IQR 21-50), and 54 (IQR 33-80) active days within the first 3, 6, and 12 months after surgery, respectively.

XSL•FO
RenderX

Table 1. Characteristics of study cohort.

Characteristics	Overall (N=1415)	Portal active days Q1[a,b] (n=349)	Portal active days Q2[a,c] (n=352)	Portal active days Q3[a,d] (n=354)	Portal active days Q4[a,e] (n=360)
BMI records missing, n (%)					
3-month follow-up	70 (4.9)	19 (5.4)	14 (4)	17 (4.8)	20 (5.6)
6-month follow-up	329 (23.3)	126 (36.1)	92 (26.1)	74 (20.9)	37 (10.3)
12-month follow-up	469 (33.1)	201 (57.6)	131 (37.2)	87 (24.6)	50 (13.9)
Sex, n (%)					
Female	1145 (80.9)	266 (76.2)	285 (81)	305 (86.2)	289 (80.3)
Male	270 (19.1)	83 (23.8)	67 (19)	49 (13.8)	71 (19.7)
Race and ethnicity, n (%)					
Black	282 (19.9)	91 (26.1)	66 (18.8)	61 (17.2)	64 (17.8)
White	1088 (76.9)	246 (70.5)	278 (79)	285 (80.5)	279 (77.5)
Other races or ethnicities	45 (3.2)	12 (3.4)	8 (2.3)	8 (2.3)	17 (4.7)
Operation, n (%)					
RYGB[f]	805 (56.9)	187 (53.6)	193 (54.8)	214 (60.5)	211 (58.6)
SG[g]	610 (43.1)	162 (46.4)	159 (45.2)	140 (39.5)	149 (41.4)
Age at baseline (years), mean (SD)	44.5 (11.4)	45.9 (11.6)	43.6 (11.5)	43.7 (10.7)	44.8 (11.7)
Area Deprivation Index, mean (SD)	56.3 (19.1)	59.4 (18.8)	56.5 (18.9)	57.0 (18.9)	52.3 (19.2)
Charlson Comorbidity Index, mean (SD)	1.23 (1.59)	1.13 (1.43)	0.963 (1.41)	1.21 (1.60)	1.62 (1.79)
Baseline BMI (kg/m^2), median (IQR)	45.5 (41.4-51.4)	45.5 (41.1-51.0)	44.8 (41.3-50.7)	45.7 (41.4-51.1)	45.6 (41.6-52.6)
%BMIR, median (IQR)					
3 months	15.8 (13.7-18.1)	15.6 (13.4-17.8)	15.8 (13.8-18.3)	15.9 (13.6-18.1)	16.0 (14.0-18.1)
6 months	24.4 (20.4-27.9)	23.1 (19.8-27.4)	25.1 (21.1-28.4)	24.1 (20.4-27.7)	24.4 (20.8-28.2)
12 months	31.5 (25.2-36.8)	30.8 (24.7-35.6)	32.7 (26.8-37.4)	31.4 (25.2-36.1)	31.5 (25.1-37.3)
Portal active days[h], median (IQR)					
Early engagement	23.0 (15.0-33.0)	10.0 (4.0-15.0)	20.0 (16.0-25.0)	27.0 (22.0-32.0)	38.0 (32.0-45.0)
Early midterm engagement	11.0 (5.0-18.0)	3.0 (1.0-6.0)	8.0 (5.0-11.0)	14.0 (11.0-17.0)	25.0 (19.0-34.0)
Late midterm engagement	10.0 (4.0-17.0)	2.0 (0-5.0)	7.0 (4.75-10.0)	12.0 (9.0-16.0)	22.0 (17.0-30.0)
Late engagement	8.0 (3.0-15.0)	1.0 (0-3.0)	5.0 (3.0-9.0)	11.0 (7.0-15.0)	21.0 (14.0-31.0)
Function active days at 12 months[a], median (IQR)					
Billing	16.0 (7.0-27.0)	5.0 (1.0-10.0)	13.0 (7.0-20.0)	21.5 (14.0-30.0)	32.0 (21.0-45.0)
Medical tools	26.0 (4.0-61.0)	2.0 (0-9.0)	28.0 (4.0-42.0)	48.0 (9.25-62.0)	90.0 (42.5-108)
Messaging	32.0 (20.0-50.0)	12.0 (6.0-17.0)	27.0 (22.8-31.0)	40.0 (33.0-48.0)	61.0 (50.0-77.0)
My record	52.0 (32.0-78.0)	19.0 (9.0-26.0)	41.0 (37.0-47.0)	63.0 (57.0-70.0)	102 (89.0-124.0)
Resources	0 (0-10.0)	0 (0-0)	0 (0-2.0)	0 (0-17.8)	9.0 (0-38.0)
Visit	49.0 (27.0-75.0)	16.0 (8.0-24.0)	38.0 (33.0-45.0)	61.0 (55.0-68.8)	97.0 (85.0-122)
Others	9.0 (4.0-16.0)	2.0 (1.0-5.0)	7.0 (4.0-9.25)	12.0 (8.0-16.0)	21.0 (13.0-32.3)

[a]Patient characteristics were stratified by quartiles of portal active days at 12 months. The overall median of portal active days at 12 months was 54 (IQR 33-80). Quartile 1 included patients with fewer than 33 portal active days, quartile 2 included patients with 33 to 54 portal active days, quartile 3 included patients with 54 to 80 portal active days, and quartile 4 included those with more than 80 portal active days.

[b]Q1: quartile 1.

[c]Q2: quartile 2.

[d]Q3: quartile 3.

XSL•FO
RenderX

[e]Q4: quartile 4.

[f]RYGB: Roux-en-Y gastric bypass.

[g]SG: sleeve gastrectomy.

[h]The portal active days were not winsorized.

Multivariate Analysis

Our results revealed significant associations between patient early portal engagement and %BMIR: an additional 10-day increment in early portal engagement was associated with a 0.37% (95% CI –0.55% to –0.18%; *P*<.001) lower expected %BMIR at the 3-month follow-up (Table 2). At 6-month and 12-month follow-ups, an additional 10-day increment in early portal engagement was associated with a 1.11% (95% CI 0.82%-1.41%; *P*<.001) and 0.78% (95% CI 0.25%-1.31%; *P*=.004) higher expected %BMIR, respectively (Table 2). Early midterm portal engagement was associated with a 0.36% (95% CI –0.69% to –0.03%; *P*=.03) lower expected %BMIR at the 6-month follow-up. Its association with the %BMIR at the 12-month follow-up was not significant (β=–.05, 95% CI –0.70 to 0.61; *P*=.88). Late midterm portal engagement (β=–.36, 95% CI –1.01 to 0.28; *P*=.27) or late portal engagement (β=–.42, 95% CI –0.96 to 0.12; *P*=.12) did not show significant associations with the %BMIR at 12 months. Figure 1 visualizes the relationship between patient portal engagement at different time stages and expected %BMIR over follow-up time. Univariate analyses showed coherent average marginal effects of portal engagements at different time stages over follow-up time, respectively (Table S2 in Multimedia Appendix 1).

Table 2. Model results of the relationship between portal engagement and percentage of BMI reduction over time following bariatric surgery. All other independent variables were factored in as covariates.

Factor	Regression coefficient (95% CI)	*P* value	RESI[a] (95% CI)
Intercept	16.98 (15.41 to 18.55)	<.001	0.56 (0.50 to 0.64)
Early engagement at 3-month follow-up (per 10-day)[b]	–0.37 (–0.55 to –0.18)	<.001	–0.10 (–0.15 to –0.05)
Early engagement at 6-month follow-up (per 10-day)[b]	1.11 (0.82 to 1.41)	<.001	0.20 (0.14 to 0.26)
Early engagement at 12-month follow-up (per 10-day)[b]	0.78 (0.25 to 1.31)	.004	0.08 (0.02 to 0.13)
Early midterm engagement at 6-month follow-up (per 10-day)[b]	–0.36 (–0.69 to –0.03)	.03	–0.06 (–0.11 to –0.00)
Early midterm engagement at 12-month follow-up (per 10-day)[b]	–0.05 (–0.70 to 0.61)	.88	–0.00 (–0.06 to 0.05)
Late midterm engagement at 12-month follow-up (per 10-day)[b]	–0.36 (–1.01 to 0.28)	.27	–0.03 (–0.08 to 0.02)
Late engagement at 12-month follow-up (per 10-day)[b]	–0.42 (–0.96 to 0.12)	.12	–0.04 (–0.09 to 0.01)
Age (per 5-years)	–0.28 (–0.39 to –0.16)	<.001	–0.12 (–0.18 to –0.07)
Male[c]	0.19 (–0.43 to 0.81)	.55	0.02 (–0.04 to 0.06)
Black[d]	–2.64 (–3.28 to –1.99)	<.001	–0.21 (–0.26 to –0.16)
Other races or ethnicities[d]	–1.55 (–3.24 to 0.14)	.07	–0.05 (–0.11 to 0.00)
ADI[e]	0.00 (–0.01 to 0.02)	.59	0.01 (–0.04 to 0.07)
CCI[f]	–0.30 (–0.47 to –0.13)	<.001	–0.09 (–0.15 to –0.04)
SG[g,h]	–2.97 (–3.48 to –2.47)	<.001	–0.31 (–0.36 to –0.25)
Months post operation	1.46 (1.35 to 1.58)	<.001	0.66 (0.59 to 0.73)

[a]RESI: robust effect size index.

[b]The variables are interaction terms between portal engagement at different stages and the duration since surgery.

[c]The reference group for sex is female.

[d]The reference group for race and ethnicity is White.

[e]ADI: Area Deprivation Index.

[f]CCI: Charlson Comorbidity Index.

[g]SG: sleeve gastrectomy.

[h]The reference group for operation type is Roux-en-Y gastric bypass.

XSL•FO
RenderX

Figure 1. The expected percentage of BMI reduction over patient portal engagement at different stages by follow-up time, after fixing other covariates at the reference levels (ie, age set at 45 years, female sex, race and ethnicity of White, Charlson Comorbidity Index of 0, and Area Deprivation Index of 0). (A) Early portal engagement, (B) early midterm portal engagement, (C) late midterm portal engagement, and (D) late portal engagement.

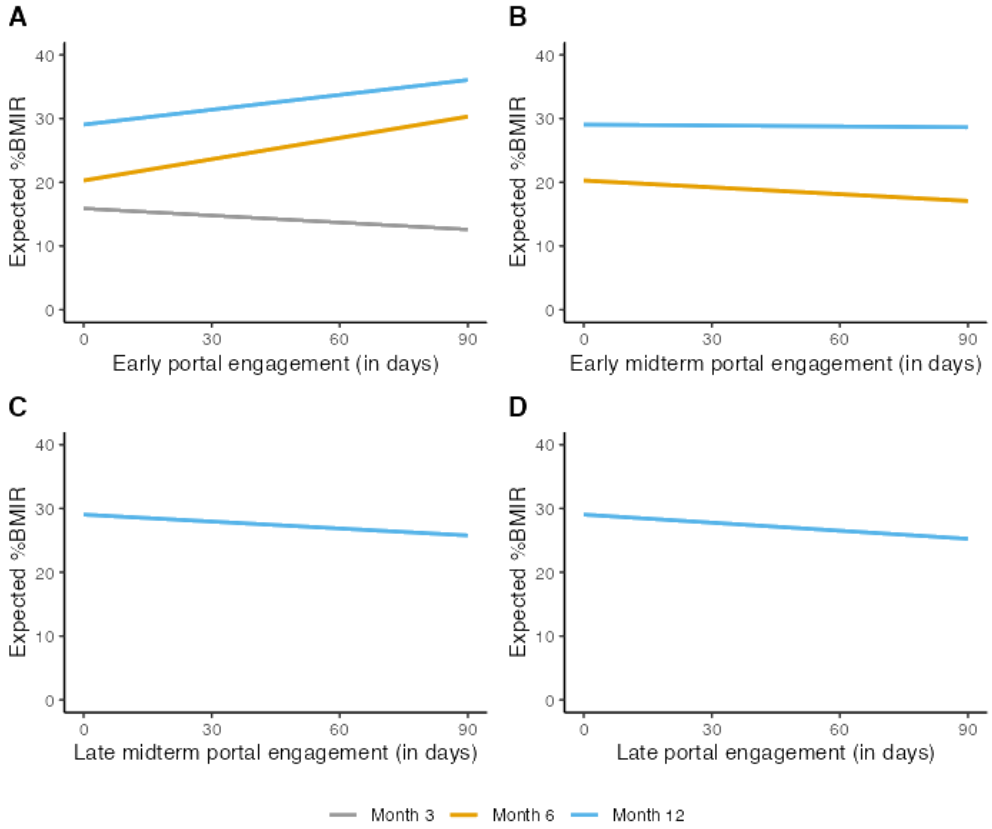

After controlling for other factors, a significantly lower expected %BMIR was observed in Black patients (β=–2.64, 95% CI –3.28 to –1.99; *P*<.001) compared with White patients (Table 2). Such difference was not significant for patients who were self-identified as other races or ethnicities (β=–1.55, 95% CI –3.24 to 0.14; *P*=.07). In terms of age at operation, a 5-year increase was found to be significantly associated with a 0.28% lower expected %BMIR (β=–.28, 95% CI –0.39 to –0.16; *P*<.001). When evaluating comorbidities, our findings indicated that patients with a higher CCI score were associated with a significantly lower expected %BMIR (β=–.30, 95% CI –0.47 to –0.13; *P*<.001). In the context of surgery type, our results indicated the expected %BMIR among the patients who underwent SG was 2.97% lower than those who received RYGB (β=–2.97, 95% CI –3.48 to –2.47; *P*<.001). In addition, patients from areas with different ADI did not show significant differences in expected %BMIR (β=.00, 95% CI –0.01 to 0.02; *P*=.59).

Regarding specific portal functions, “My Record” and “Visit” functions, which enable patients to access their medications and test results, and manage their appointments, respectively, were the most frequently used (Table 1). Our analysis found that the higher engagement in almost every portal function was associated with a significantly lower expected %BMIR at 3-month follow-up and a significantly higher expected %BMIR at 6- and 12-month follow-up (except the “Resources” function; Table 3). Specifically, “Resources,” “Others,” and “Billing” functions showed the strongest associations at 3 months (β=–2.63, 95% CI –5.07 to –0.20, *P*=.03 for “Resources”; β=–1.24, 95% CI –1.84 to –0.64, *P*<.001 for “Others”; β=–1.02, 95% CI –1.37 to –0.66, *P*<.001 for “Billing”). “Others,” “Billing,” and “Messaging” functions showed the strongest positive association at 6 and 12 months (at 6 months: β=3.28, 95% CI 2.35-4.21, *P*<.001 for “Others”; β=1.77, 95% CI 1.24-2.30, *P*<.001 for “Billing”; β=1.38, 95% CI 1.02-1.75, *P*<.001 for “Messaging”; at 12 months: β=1.97, 95% CI 0.32-3.62, *P*=.01 for “Others”; β=1.11, 95% CI 0.21-2.02, *P*=.01 for “Billing”; β=.78, 95% CI 0.10-1.46, *P*=.02 for “Messaging”).

XSL•FO
RenderX

Table 3. Model results of the estimated associations between specific portal functions early engagement and the percentage of BMI reduction by follow-up time (ie, 3, 6, and 12 months).

Portal function	3 months		6 months		12 months	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Billing	–1.02 (–1.37 to –0.66)	<.001	1.77 (1.24 to 2.30)	<.001	1.11 (0.21 to 2.02)	.01
Medical tools	–0.39 (–0.55 to –0.23)	<.001	0.89 (0.57 to 1.20)	<.001	0.59 (0.10 to 1.08)	.01
Messaging	–0.35 (–0.60 to –0.10)	.005	1.38 (1.02 to 1.75)	<.001	0.78 (0.10 to 1.46)	.02
My record	–0.38 (–0.57 to –0.18)	<.001	1.14 (0.85 to 1.44)	<.001	0.76 (0.22 to 1.30)	.006
Resources	–2.63 (–5.07 to –0.20)	.03	2.23 (–2.36 to 6.82)	.34	2.36 (–5.69 to 10.41)	.56
Visits	–0.30 (–0.48 to –0.12)	.001	1.06 (0.77 to 1.35)	<.001	0.59 (0.03 to 1.14)	.03
Others	–1.24 (–1.84 to –0.64)	<.001	3.28 (2.35 to 4.21)	<.001	1.97 (0.32 to 3.62)	.01

The sensitivity analysis (Figure 1 and Table S3 in Multimedia Appendix 1) indicated that the significant association between early portal engagement and %BMIR at 3, 6, and 12 months persisted after adjusting for baseline BMI (β=–.36, 95% CI –0.55 to –0.18, *P*<.001 at 3 months; β=1.11, 95% CI 0.82-1.31, *P*<.001 at 6 months; and β=.78, 95% CI 0.24-1.31, *P*=.004 at 12 months). Portal engagement increased significantly during the COVID-19 pandemic, with median engagement rising from 44.0 (IQR 26.0-68.8) active days before the pandemic to 71 (IQR 50-97) days amid it at 12 months after surgery (Table S4 in Multimedia Appendix 1). The analysis confirmed that the positive association between portal engagement and %BMIR was consistent across the prepandemic and pandemic periods (Tables S5 and S6 in Multimedia Appendix 1). In the sensitivity analysis of incorporating the interaction terms between portal engagement at different stages and age, there was no sufficient statistical evidence that the associations between portal engagements and weight loss outcome would vary by age (Table S7 in Multimedia Appendix 1).

Discussion

Principal Findings

Bariatric surgery necessitates continued postoperative care, and previous studies have suggested that patient portals can enhance health awareness and therapy adherence [11,37,38]. Echoing findings from these studies, our retrospective longitudinal study reveals complex relationships between portal engagement and weight loss outcomes following bariatric surgery. Specifically, engagement within the first 3 months (early portal engagement) was associated with poorer weight loss outcomes at the 3-month follow-up but associated with better outcomes at 6 and 12 months. Similarly, engagement between 3 and 6 months (early midterm portal engagement) was linked to less favorable outcomes at the 6-month follow-up. These patterns suggest that simultaneous portal usage during the same stage as the follow-up may reflect greater patient needs or complications, potentially leading to increased interactions with health care providers and resources. Over time, this heightened engagement is likely to support more effective health management, thereby improving long-term outcomes. This insight underscores the dynamic and crucial role of patient portals in patient care, emphasizing the importance of maintained engagement for long-term success in weight management after bariatric surgery.

Further investigation into the specific functions of the portal revealed that the “Messaging” functions have one of the strongest positive associations with %BMIR at 6-month follow-up, suggesting these functions could be particularly beneficial to postoperative management. Notably, by the 6-month follow-up, patients engaging with the “Messaging” functions for every 100 active days saw an additional reduction in BMI by 13.8%. These findings align with a previous study, which indicates that portal may be a source of context-based educational materials for self-management [12]. While “Billing” and “other functions” showed positive associations with %BMIR, they frequently co-occurred with other functions. In particular, “Billing” and “other functions” occurred together with “Visit” functions in 92% and 90% active days, respectively, suggesting that their statistical significance may be an artifact of metric design rather than evidence of clinically meaningful postoperative patient engagement through the portal.

Comparison With Previous Work

Previous studies have highlighted the potential of digital health engagement in obesity treatment. For example, 1 study found that individuals with access to a mobile weight loss intervention tool achieved significantly greater weight loss at 6 months after surgery compared with those without access to the tool [39]. Another randomized controlled study demonstrated that patients who received a digital intervention before bariatric surgery engaged in more physical activity at an 8-week postoperative follow-up compared with those receiving usual care, although the intervention did not significantly affect BMI [40]. In addition, a survey study has shown that online forums support patients by providing information and fostering motivation [41]. These findings are consistent with the results of our study, However, previous studies focused on different types of digital health engagement. Our study focused on the patient portal, which is integrated with EHRs and an inevitable method to deliver health service. Furthermore, we are the first longitudinal study to investigate the influence of engagement at different time stages on weight loss outcomes. Suggesting that digital tools could serve as a valuable complement to standard perioperative and postoperative care in bariatric surgery.

XSL•FO
RenderX

Similar to previous bariatric surgery studies, loss to follow-ups remains a significant challenge [40-43]. In our study, follow-up rates were 95.1%, 76.7%, and 66.9% at 3, 6, and 12 months, respectively. Potential reasons for this loss included patients with fewer symptoms opting out of later follow-up appointments at the surgery center, unsuccessful weight loss (<50% excess weight loss), changes in their residence and insurance status, and so on [42,43]. This decline in follow-up rates from 6 months to 12 months may reduce the power to capture the association between portal engagement and BMI reduction and may cause a potential selection bias at 12 months. Since the reasons for patients' loss to follow-up visits were unknown, it may potentially cause an overestimation or underestimation of the benefits of patient portal engagement.

In addition to patient portal engagement, we observed that younger patients tended to achieve better postoperative outcomes (Table 2). This could be attributed to faster metabolic rates in younger individuals [44]. In addition, patient's digital access and technical skills may also vary by age [33]. However, age cannot be used as a proxy for these since age's relationship with the weight loss outcome is complex. Our findings align with previous studies that reported racial differences in weight loss after bariatric surgery, noting that self-identified Black patients experienced less weight loss compared with their White counterparts; however, the underlying causative reasons remain unclear [45]. We also found that patients with multiple comorbidities exhibited less weight loss, highlighting the need for tailored postoperative care for patients with more complex health conditions. Our analysis revealed that patients who underwent RYGB achieved more weight loss than those who had SG, consistent with existing literature [8,46]. The duration since surgery was positively associated with BMI reduction. It highlighted the importance of considering the timeline in post–bariatric surgery outcomes analysis. While these observations were significant, they were not the primary focus of this work and warrant further investigation.

Clinical Significance

Our research provides in-depth, longitudinal analysis of patient portal use in bariatric surgery, highlighting its benefits in the management of postoperative weight loss resources and the communication between clinicians and patients. However, more needs to be done to familiarize both patients and health care professionals with portal usage. Encouragement from health care providers and user-friendly designs from vendors are essential [47-49]. Nonetheless, the increasing use of portals may lead to clinician burnout [50,51], underscoring the need for strategies that balance efficient digital interaction management with health care team well-being. Achieving this equilibrium is crucial for maximizing patient satisfaction and the effective use of digital health tools.

Future Directions

Building on the insights from our data-driven study, the next steps should focus on hypothesis-driven research to delve deeper into the nuances of patient portal engagement and its impact on postoperative weight management in patients after bariatric surgery. Recommended future studies include conducting randomized controlled trials to test the effectiveness of targeted portal interventions at different postoperative stages, exploring the reasons behind patient engagement through qualitative methods, and incorporating demographic variables like age to personalize portal use. In addition, longitudinal randomized controlled trials could rigorously test the causality between portal use and weight loss, while expanding research to diverse populations could validate the generalizability of the findings. Optimizing portal features based on the most impactful functions and evaluating collaborative care models that integrate real-time, provider-patient interactions through portals would also be beneficial. These steps would refine understanding and enhance the efficacy of patient portals in supporting long-term weight management success post bariatric surgery.

Strengths and Limitations

Several limitations in this study are worth acknowledging. First, we would like to clarify that our research was not to establish causality. Our study was a data-driven, retrospective study designed to measure the "association" between postoperative patient portal usage and %BMIR at 3, 6, and 12 months after surgery. The insights gleaned from our findings are intended to inform the design of future clinical trials. Second, this study was conducted at a single academic medical center. The generalizability of the findings needs to be confirmed in other large surgical weight loss programs. Third, we acknowledged the need for better data collection during presurgery screening and postoperative follow-up visits. While we adjusted for a range of common factors associated with bariatric surgery outcomes, there could be confounding effects from other factors not within our data access, such as individual-level socioeconomic status, lifestyle factors such as diet and physical activity levels, health literacy, digital literacy, digital access, reasons for loss to follow-up visits, and engagement with other health monitoring apps or other forms of interventions. Due to the limitation of our retrospective study, these variables had low coverage in our cohort. Fourth, we observed a data imbalance in the proportion of Black patients between the surgery groups. Although we adjusted for the surgery type and race and ethnicity as covariates, the imbalance may influence the observed association between race and ethnicity and %BMIR. To address this potential concern, one approach would be to explore the association within each surgery type separately. Finally, the direct effects of specific clinic-oriented activities conducted through the portal—such as medication refill requests or appointment scheduling—on improved weight loss outcomes have yet to be determined. In this study, these actions were not investigated in detail due to the smaller sample sizes they represent, which could result in a lack of statistical power and unstable estimates. Moving forward, our research aims to identify and scrutinize these pivotal actions more closely and assess their associations with post–bariatric surgery outcomes.

Conclusions

In this retrospective study, we analyzed longitudinal data from the VUMC bariatric surgery registry alongside patient portal use records. Our findings indicate that higher early portal engagement was associated with less favorable weight loss immediately after surgery but associated with improved outcomes at 6 to 12 months. The portal engagement metric and

XSL•FO
RenderX

longitudinal analysis framework developed in this study hold promise for application in other chronic conditions requiring long-term management. Future research should aim to include additional postoperative factors such as dietary intake and comorbidity improvements to provide a more comprehensive understanding of the long-term outcomes following bariatric surgery.

Acknowledgments

This research was supported, in part, by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award R01DK126721. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

XZ, CY, KK, and YC contributed to study conception and design. XZ and YC accessed and verified the underlying data. XZ, KK, CY, YF, and YC performed data analysis and interpretation. KK, XZ, and SV contributed to statistical design. XZ and YC drafted the manuscript. YC obtained funding and supervised this study. All authors contributed to critical revision of the manuscript for important intellectual content.

Conflicts of Interest

GS receives consulting fees from Novo Nordisk, Rhythm Pharmaceuticals, and Eli Lilly. GS is on the Speaker's Bureau for Novo Nordisk and Rhythm Pharmaceutical and receives industry support for clinical trials from Eli Lilly.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 387 KB - jmir_v26i1e56573_app1.docx]

References

1. Prevalence of obesity and severe obesity among adults: United States, 2017–2018. Centers for Disease Control and Prevention. 2020 Feb. URL: https://www.cdc.gov/nchs/products/databriefs/db360.htm [accessed 2023-12-27]
2. Managing overweight and obesity in adults: systematic evidence review from the Obesity Expert Panel. National Heart, Lung, and Blood Institute. 2013 Nov. URL: https://www.nhlbi.nih.gov/health-topics/managing-overweight-obesity-in-adults [accessed 2013-11-01]
3. Li TY, Rana JS, Manson JE, Willett WC, Stampfer MJ, Colditz GA, et al. Obesity as compared with physical activity in predicting risk of coronary heart disease in women. Circulation 2006;113(4):499-506 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.105.574087] [Medline: 16449729]
4. Hsu CY, McCulloch CE, Iribarren C, Darbinian J, Go AS. Body mass index and risk for end-stage renal disease. Ann Intern Med 2006;144(1):21-28. [doi: 10.7326/0003-4819-144-1-200601030-00006] [Medline: 16389251]
5. Aminian A, Tu C, Milinovich A, Wolski KE, Kattan MW, Nissen SE. Association of weight loss achieved through metabolic surgery with risk and severity of COVID-19 infection. JAMA Surg 2022;157(3):221-230 [FREE Full text] [doi: 10.1001/jamasurg.2021.6496] [Medline: 34964827]
6. Lavie CJ, Milani RV, Ventura HO. Obesity and cardiovascular disease: risk factor, paradox, and impact of weight loss. J Am Coll Cardiol 2009;53(21):1925-1932 [FREE Full text] [doi: 10.1016/j.jacc.2008.12.068] [Medline: 19460605]
7. Wolfe BM, Kvach E, Eckel RH. Treatment of obesity: weight loss and bariatric surgery. Circ Res 2016;118(11):1844-1855 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.307591] [Medline: 27230645]
8. Maciejewski ML, Arterburn DE, van Scoyoc L, Smith VA, Yancy WS, Weidenbacher HJ, et al. Bariatric surgery and long-term durability of weight loss. JAMA Surg 2016;151(11):1046-1055 [FREE Full text] [doi: 10.1001/jamasurg.2016.2317] [Medline: 27579793]
9. Madura JA, Dibaise JK. Quick fix or long-term cure? Pros and cons of bariatric surgery. F1000 Med Rep 2012;4:19 [FREE Full text] [doi: 10.3410/M4-19] [Medline: 23091563]
10. Robinson JR, Davis SE, Cronin RM, Jackson GP. Use of a patient portal during hospital admissions to surgical services. AMIA Annu Symp Proc 2017;2016:1967-1976 [FREE Full text] [Medline: 28269956]
11. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
12. Johnson KB, Ibrahim SA, Rosenbloom ST. Ensuring equitable access to patient portals-closing the "Techquity" gap. JAMA Health Forum 2023;4(11):e233406 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.3406] [Medline: 37948065]
13. Antonio MG, Petrovskaya O, Lau F. The State of evidence in patient portals: umbrella review. J Med Internet Res 2020;22(11):e23851 [FREE Full text] [doi: 10.2196/23851] [Medline: 33174851]

XSL•FO
RenderX

14. Han HR, Gleason KT, Sun CA, Miller HN, Kang SJ, Chow S, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019;6(4):e15038 [FREE Full text] [doi: 10.2196/15038] [Medline: 31855187]
15. Lee J, Kim JGB, Jin M, Ahn K, Kim B, Kim S, et al. Beneficial effects of two types of personal health record services connected with electronic medical records within the hospital setting. Comput Inform Nurs 2017;35(11):574-581. [doi: 10.1097/CIN.0000000000000362] [Medline: 28548973]
16. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care 2021;46(2):192-204. [doi: 10.1080/17538157.2021.1879810] [Medline: 33840342]
17. Kizy S, Jahansouz C, Wirth K, Ikramuddin S, Leslie D. Bariatric surgery: a perspective for primary care. Diabetes Spectr 2017;30(4):265-276 [FREE Full text] [doi: 10.2337/ds17-0034] [Medline: 29151717]
18. Metabolic and bariatric surgery accreditation and quality improvement program. American College of Surgeons. URL: https://tinyurl.com/yrrw8f7w [accessed 2023-12-27]
19. Steitz BD, Wong JIS, Cobb JG, Carlson B, Smith G, Rosenbloom ST. Policies and procedures governing patient portal use at an academic medical center. JAMIA Open 2019;2(4):479-488 [FREE Full text] [doi: 10.1093/jamiaopen/ooz039] [Medline: 32025645]
20. Gasparini A. Comorbidity: an R package for computing comorbidity scores. JOSS 2018;3(23):648. [doi: 10.21105/joss.00648]
21. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol 2004;57(12):1288-1294. [doi: 10.1016/j.jclinepi.2004.03.012] [Medline: 15617955]
22. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood atlas. N Engl J Med 2018;378(26):2456-2458 [FREE Full text] [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
23. Di Tosto G, McAlearney AS, Fareed N, Huerta TR. Metrics for outpatient portal use based on log file analysis: algorithm development. J Med Internet Res 2020;22(6):e16849 [FREE Full text] [doi: 10.2196/16849] [Medline: 32530435]
24. Yan C, Zhang X, Yang Y, Kang K, Were MC, Embí P, et al. Differences in health professionals' engagement with electronic health records based on inpatient race and ethnicity. JAMA Netw Open 2023;6(10):e2336383 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36383] [Medline: 37812421]
25. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: a new goldmine for health services research? J Biomed Inform 2020;101:103343 [FREE Full text] [doi: 10.1016/j.jbi.2019.103343] [Medline: 31821887]
26. Chen B, Alrifai W, Gao C, Jones B, Novak L, Lorenzi N, et al. Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning. J Am Med Inform Assoc 2021;28(6):1168-1177 [FREE Full text] [doi: 10.1093/jamia/ocaa338] [Medline: 33576432]
27. Zhang X, Yan C, Malin BA, Patel MB, Chen Y. Predicting next-day discharge via electronic health record access logs. J Am Med Inform Assoc 2021;28(12):2670-2680 [FREE Full text] [doi: 10.1093/jamia/ocab211] [Medline: 34592753]
28. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. J Am Med Inform Assoc 2020;27(3):480-490 [FREE Full text] [doi: 10.1093/jamia/ocz196] [Medline: 31750912]
29. Courcoulas AP, Christian NJ, Belle SH, Berk PD, Flum DR, Garcia L, et al. Weight change and health outcomes at 3 years after bariatric surgery among individuals with severe obesity. JAMA 2013;310(22):2416-2425 [FREE Full text] [doi: 10.1001/jama.2013.280928] [Medline: 24189773]
30. Belle SH, Berk PD, Courcoulas AP, Engel S, Flum DR, Gourash W, et al. Reporting weight change: standardized reporting accounting for baseline weight. Surg Obes Relat Dis 2013;9(5):782-789 [FREE Full text] [doi: 10.1016/j.soard.2012.11.006] [Medline: 23337770]
31. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22. [doi: 10.2307/2336267]
32. Richwine C. Progress and persistent disparities in patient access to electronic health information. JAMA Health Forum 2023;4(11):e233883 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.3883] [Medline: 37948063]
33. Finkelstein R, Wu Y, Brennan-Ing M. Older adults' experiences with using information and communication technology and tech support services in New York City: findings and recommendations for post-pandemic digital pedagogy for older adults. Front Psychol 2023;14:1129512 [FREE Full text] [doi: 10.3389/fpsyg.2023.1129512] [Medline: 37138998]
34. Kang K, Jones MT, Armstrong K, Avery S, McHugo M, Heckers S, et al. Accurate confidence and bayesian interval estimation for non-centrality parameters and effect size indices. Psychometrika 2023;88(1):253-273 [FREE Full text] [doi: 10.1007/s11336-022-09899-x] [Medline: 36725775]
35. Kang K, Seidlitz J, Bethlehem RAI, Xiong J, Jones MT, Mehta K, et al. Study design features increase replicability in brain-wide association studies. Nature 2024 [FREE Full text] [doi: 10.1038/s41586-024-08260-9]
36. Vandekar S, Tao R, Blume J. A robust effect size index. Psychometrika 2020;85(1):232-246 [FREE Full text] [doi: 10.1007/s11336-020-09698-2] [Medline: 32232646]
37. Singh P, Jonnalagadda P, Morgan E, Fareed N. Outpatient portal use in prenatal care: differential use by race, risk, and area social determinants of health. J Am Med Inform Assoc 2022;29(2):364-371 [FREE Full text] [doi: 10.1093/jamia/ocab242] [Medline: 34741505]

XSL•FO
RenderX

38. Lyles CR, Sarkar U, Ralston JD, Adler N, Schillinger D, Moffet HH, et al. Patient-provider communication and trust in relation to use of an online patient portal among diabetes patients: the diabetes and aging study. J Am Med Inform Assoc 2013;20(6):1128-1131 [FREE Full text] [doi: 10.1136/amiajnl-2012-001567] [Medline: 23676243]
39. Markkanen JO, Oikarinen N, Savolainen MJ, Merikallio H, Nyman V, Salminen V, et al. Mobile health behaviour change support system as independent treatment tool for obesity: a randomized controlled trial. Int J Obes (Lond) 2024;48(3):376-383 [FREE Full text] [doi: 10.1038/s41366-023-01426-x] [Medline: 38062218]
40. Sysko R, Michaelides A, Costello K, Herron DM, Hildebrandt T. An initial test of the efficacy of a digital health intervention for bariatric surgery candidates. Obes Surg 2022;32(11):3641-3649 [FREE Full text] [doi: 10.1007/s11695-022-06258-8] [Medline: 36074201]
41. Robinson A, Husband AK, Slight RD, Slight SP. Digital support for patients undergoing bariatric surgery: narrative review of the roles and challenges of online forums. JMIR Perioper Med 2020;3(2):e17230 [FREE Full text] [doi: 10.2196/17230] [Medline: 33393926]
42. Higa K, Ho T, Tercero F, Yunus T, Boone KB. Laparoscopic Roux-en-Y gastric bypass: 10-year follow-up. Surg Obes Relat Dis 2011;7(4):516-525. [doi: 10.1016/j.soard.2010.10.019] [Medline: 21333610]
43. Vidal P, Ramón JM, Goday A, Parri A, Crous X, Trillo L, et al. Lack of adherence to follow-up visits after bariatric surgery: reasons and outcome. Obes Surg 2014;24(2):179-183. [doi: 10.1007/s11695-013-1094-9] [Medline: 24101091]
44. Woźniewska P, Diemieszczyk I, Groth D, Szczerbiński L, Choromańska B, Błachnio-Zabielska A, et al. The influence of patient's age on metabolic and bariatric results of laparoscopic sleeve gastrectomy in 2-year observation. BMC Surg 2020;20(1):323 [FREE Full text] [doi: 10.1186/s12893-020-00989-0] [Medline: 33298009]
45. Admiraal WM, Celik F, Gerdes V, Dallal RM, Hoekstra JB, Holleman F. Ethnic differences in weight loss and diabetes remission after bariatric surgery: a meta-analysis. Diabetes Care 2012;35(9):1951-1958 [FREE Full text] [doi: 10.2337/dc12-0260] [Medline: 22923683]
46. Grönroos S, Helmiö M, Juuti A, Tiusanen R, Hurme S, Löyttyniemi E, et al. Effect of laparoscopic sleeve gastrectomy vs Roux-en-Y gastric bypass on weight loss and quality of life at 7 years in patients with morbid obesity: the SLEEVEPASS randomized clinical trial. JAMA Surg 2021;156(2):137-146 [FREE Full text] [doi: 10.1001/jamasurg.2020.5666] [Medline: 33295955]
47. Sada A, Harmsen WS, McKenzie TJ, Kellogg TA, Ghanem O, Kendrick ML, et al. Telemedicine video follow-up for bariatric surgery patients: what do patients prefer? A prepandemic pilot randomized clinical trial. Am Surg 2023;89(6):2806-2809 [FREE Full text] [doi: 10.1177/00031348211054547] [Medline: 34775810]
48. An Q, Kelley MM, Hanners A, Yen PY. Sustainable development for mobile health apps using the human-centered design process. JMIR Form Res 2023;7:e45694 [FREE Full text] [doi: 10.2196/45694] [Medline: 37624639]
49. Gleason KT, Wu MMJ, Wec A, Powell DS, Zhang T, Wolff JL. Patient portal use among older adults with dementia diagnosis. JAMA Intern Med 2023;183(8):881-883 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1568] [Medline: 37358848]
50. Stillman M. Death by Patient Portal. JAMA 2023;330(3):223-224. [doi: 10.1001/jama.2023.11629] [Medline: 37389857]
51. Budd J. Burnout related to electronic health record use in primary care. J Prim Care Community Health 2023;14:21501319231166921 [FREE Full text] [doi: 10.1177/21501319231166921] [Medline: 37073905]

Abbreviations

%BMIR: percentage of BMI reduction
ADI: Area Deprivation Index
CCI: Charlson Comorbidity Index
EHR: electronic health record
IRB: institutional review board
MBSAQIB: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program
MHAV: My Health at Vanderbilt
RESI: robust effect size index
RYGB: Roux-en-Y gastric bypass
SG: sleeve gastrectomy
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VUMC: Vanderbilt University Medical Center

XSL•FO
RenderX

Edited by N Cahill; submitted 19.01.24; peer-reviewed by C Li, A Rule; comments to author 29.02.24; revised version received 25.04.24; accepted 30.10.24; published 09.12.24.

Please cite as:
Zhang X, Kang K, Yan C, Feng Y, Vandekar S, Yu D, Rosenbloom ST, Samuels J, Srivastava G, Williams B, Albaugh VL, English WJ, Flynn CR, Chen Y
Association Between Patient Portal Engagement and Weight Loss Outcomes in Patients After Bariatric Surgery: Longitudinal Observational Study Using Electronic Health Records
J Med Internet Res 2024;26:e56573
URL: https://www.jmir.org/2024/1/e56573
doi:10.2196/56573
PMID:

©Xinmeng Zhang, Kaidi Kang, Chao Yan, Yubo Feng, Simon Vandekar, Danxia Yu, S Trent Rosenbloom, Jason Samuels, Gitanjali Srivastava, Brandon Williams, Vance L Albaugh, Wayne J English, Charles R Flynn, You Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Insights and Trends in Open Note Access: Retrospective Observational Study

Randeep Singh Badwal[1], MD; Paul Cavo[2], BS; Mandip Panesar[1,2], MD, MS

[1]Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, United States

[2]Erie County Medical Center, Buffalo, NY, United States

Corresponding Author:
Mandip Panesar, MD, MS
Erie County Medical Center
462 Grider Street
Buffalo, NY, 14215
United States
Phone: 1 716 898 1417
Email: MPanesar@ecmc.edu

Abstract

Background: As of 2021, at least 4 out of every 5 hospitals offered patients access to clinical notes via a web-based patient portal, a number that is expected to grow because of the 21st Century Cures Act. There is limited data on how open note use may have evolved over time or which types of clinical interactions were viewed most in the outpatient setting.

Objective: This study aims to analyze trends in outpatient open note access over time; characterize usage in terms of age, sex, and clinical interaction type; and assess the method of access to help uncover areas of improvement in patient engagement and identify further areas of research.

Methods: A retrospective observational study was conducted at Erie County Medical Center from November 1, 2021, to December 31, 2022, to coincide with the time that open notes went live. Outpatient note access and account logs were downloaded from the portal and combined into a single dataset consisting of 18,384 note accesses by 4615 users, with column headings of the patient index, sex, age, note title that was accessed, clinical interaction type, time stamp of note creation, time stamp of access, and method of access (web vs mobile). A separate table was created with sex data for all 35,273 portal accounts. Microsoft Excel and Microsoft Power Query were used to combine and analyze the data.

Results: During the study period, 4615 portal users viewed 12,150 documents for a total of 18,384 times, averaging 2.6 notes per patient viewed 4 times. Only 13.1% (4615/35,273) of all portal inpatient and outpatient registrants viewed their outpatient notes. There was a female predominance in those who viewed notes (2926/4615, 63.4%; $P<.001$), while 56.8% (20,047/35,273) of all portal registrants were female. Users in their 30s and 50s accessed more notes than other age groups. The ratio of mobile-to-web access of notes tended to decrease as a function of increasing age, which was not observed in those aged ≥90 years. Notes regarding COVID-19 assessments were the most accessed among all clinical interactions (4725/12,150, 38.9%). Overall, the number of users accessing notes reached a maximum of 1968 before declining to 1027 by the end of the study period.

Conclusions: Open note access was largely dominated by COVID-19 assessments, and the number of users viewing their notes has declined over time as the pandemic subsided. Furthermore, female patients and those aged in their 30s as well as 50s viewed more notes than other groups. Finally, the percentage of notes viewed via a mobile device tended to decrease as a function of increasing age, showing that web-based access of open notes is an important modality for older patients.

(J Med Internet Res 2024;26:e55982) doi:10.2196/55982

KEYWORDS

open note trends; open notes; open note access; open note use, open note sex; open note specialty; clinical note views; patient portal; patients; trends; hospitals; engagement; retrospective observational study; outpatient; assessment; older patients; adults; pandemic; COVID-19

XSL•FO
RenderX

Introduction

Medical notes are comprehensive narratives generated by health care providers during encounters with patients. They form an important backbone of the medical record since they summarize the problem, diagnosis, and treatment plan for various ailments that a patient may experience throughout their lifetime. As of 2021, a total of 78% of office-based physicians and 96% of nonfederal acute care hospitals have implemented a certified electronic health record (EHR) into practice [1], which affords the ability to document these notes digitally. Although previously it was at the discretion of health care institutions to share medical notes with patients, the 21st Century Cures Act's Final Rule [2] implemented in April 2021 mandates that all electronic health information, including notes, be made accessible to patients. This concept, referred to as "open notes" [3], is thought to foster greater transparency between a patient and their health care team, and has been shown to make patients feel more in control of their care [4] and increase medication adherence [5].

As of 2021, more than 4 out of 5 hospitals offered patients access to clinical notes via a web-based patient portal [6], a number that is expected to rise because of legislation. However, despite its widespread availability, literature published specifically on open note access remains relatively scarce. Sangal et al [7] showed that patients who visit the emergency department were less likely to read notes if they were under the age of 18 years; male; Black non-Hispanic, Hispanic, or Latino; admitted; Spanish speaking; or on public insurance. Blok et al [8] showed that patients who viewed their outpatient notes were younger; were mostly White; were mostly female; had better income status; had slightly higher rates of primary care visits; and were higher users of secure messaging than nonusers. However, how open note access may have evolved over time or which type of clinical interactions were accessed the most are worthwhile questions to explore.

For this study, we decided to focus on the outpatient rather than the inpatient setting due to differences in motivation for accessing notes in each setting. For example, there is a greater emphasis on self-care in the outpatient setting, which may motivate patients to obtain a deeper understanding of their treatment plan via a medical note, as opposed to the hospital where health care interactions are continuous. Due to the potential for such motivations to influence the number of patient interactions with the portal, we concentrated on outpatient notes rather than looking at all notes as a conglomerate. In particular, we sought to study trends in access over time; characterize access in terms of age, sex, and type of clinical interaction; and assess the method of access to help uncover areas of improvement in patient engagement as well as identify further areas of research.

Methods

Setting

The study was conducted at Erie County Medical Center, a level 1 regional trauma and teaching center with 573 inpatient beds and 24 practice locations in Buffalo, New York. There were approximately 154,501 visits in 2023. At registration, all patients are given the option of signing up for the Follow My Health Patient Portal, which provides the ability to request appointments and medication refills; message a provider; record health data such as blood pressure; view medical history and results; and, as of November 2021, view open notes.

Study Design

A retrospective observational study was performed using patient portal audit data. The study period was 14 months (from November 1, 2021, to December 31, 2022).

The data were retrieved from Follow My Health, a system that administers the portal to patients who have their data stored within the EHR. On the administrative side, audit logs in .CSV format could be downloaded and thereby analyzed to inform trends about portal use. We were particularly interested in the "Health Record Viewed" audit file, which records the access of clinical documents by patient username, and the "Connected Patients" file, which lists the demographic information for each username in the portal. The portal contained a total of 35,273 accounts as of January 12, 2023.

Since we were particularly interested in the outpatient setting, the "Health Record Viewed" file was filtered by outpatient documents at the time of download, and a list of unique note titles was curated. Some titles included in this initial dataset did not correspond to notes authored by health care professionals, so we compared a list of note titles present prior to November 2021 to those that appeared after open notes went live. This led to a final list of titles that was verified with the IT team to consist only of outpatient clinical notes. The process allowed us to identify 18,385 outpatient notes that were accessed by patients and authored by health care workers, which included physicians, nurses, advanced practice providers, dieticians, and social workers. The note titles were subsequently grouped and labeled with a clinical interaction type. For example, titles such as "Internal Medicine Follow Up" and "Patient Portal Family Med Note" were grouped under "FM, IM, or primary care." Notable absences were psychiatry notes that were documented in the inpatient EHR, as well as obstetrics and gynecology, which operate as an external entity to the institution.

During the process of labeling, we compared time stamps of document creation and discovered that a particular note may be uploaded as multiple distinct notes in the patient portal. For example, there were 1643 infectious disease (ID) notes that had identical creation time stamps with COVID-19 notes, and a sample of these notes showed that they were identical. The ID notes were relabeled as COVID-19. In addition, there were instances of duplicate notes with similar titles and other notes forming a subsection of larger notes. Although not exhaustive, 2 examples of these included "patient portal COVID progress note" and "COVID progress note," or "annual wellness visit" and "patient portal health maintenance note." There were 179 pairs, 2 triplets, and 1 quadruplet sets of notes that were accessed by users (1829 in total including the COVID-19 or ID notes mentioned earlier). Since a patient may access 1 version of a note while overlooking another, or they may engage with different versions on distinct dates, we chose to retain duplicate

XSL•FO
RenderX

notes to capture the complete range of patient interactions with the portal.

Ultimately, the data were merged within Microsoft Power Query to obtain a single dataset for analysis, consisting of username, patient first and last names, sex, date of birth, note title that was accessed, type of clinical interaction, time stamp of note creation, time stamp of access, and method of access (web vs mobile). Demographic information of each username within the portal was carefully reviewed against the EHR to fill in missing information, uncover multiple usernames for a single patient who may have registered more than once, or identify multiple usernames for patients who had name changes but were indeed a single entity. This was done to obtain an accurate account of unique users who accessed their notes. Only 1 patient who accessed a single note (bariatrics) was deleted from the dataset because they could not be uniquely identified from the EHR, resulting in 18,384 total entries. Date of birth values were converted into age in years and all patient identifiers were removed to obtain a final dataset. A separate, single-column table was derived from the "Connected Patients" file that only listed the sex of each patient connected to the portal. Pivot tables and charts within Microsoft Excel were used to analyze the data. A chi-square test was used to compare the proportion of female patients versus male patients accessing their notes.

Ethical Considerations

The study was approved by the institutional review board (IRB) at the University at Buffalo (IRB ID STUDY00006926) and the Research Department at Erie County Medical Center. As part of the approval, the IRB granted a full waiver of individual authorization to use protected health information for research purposes in accordance with the Health Insurance Portability and Accountability Act (HIPAA). All data were deidentified, securely stored, analyzed, and managed in compliance with regulations. There was no compensation.

Results

From November 1, 2021, to December 31, 2022, a total of 13.1% (4615/35,273) of all portal registrants viewed outpatient notes, with the denominator consisting of both inpatient and outpatient registrants. The total number of notes accessed was 18,384, out of which 12,150 were unique. This equates to each user reading 2.6 unique notes a total of 4 times. More female patients viewed notes than male patients (2926/4615, 63.4%; $P<.001$). Of all portal accounts, 56.8% (20,047/35,273) were female. In terms of age, more people in their 30s and 50s viewed notes compared to other groups (Figure 1). Approximately 82.7% (15,196/18,384) of all notes were accessed via a mobile device, a trend that tended to decrease with progressing age but was not observed in those aged 90 years and older (Figure 2).The distribution of open note accesses by clinical interaction type is shown in Table 1. The "COVID-19" category had the most instances of notes being accessed. This was followed by "family medicine, internal medicine, or primary care" with the remainder being clinical interactions from other specialties. There were 565 documents that were uncategorized since multiple specialties shared a single note title in these circumstances.

Figure 3 shows the number of unique users accessing open notes over each quarter-year, where a sharp increase was observed when open notes went live followed by a gradual decline of 47.8% (–941/1968) from the maximum. Figure 4 excludes COVID-19 assessments from the dataset. Here, a more gradual increase is observed, followed by a slight decline of 7.4% (–74/1000) from the maximum by the end of the study period.

Figure 1. Age distribution of open note users. A bimodal peak is observed in the age range of 30s and 50s.

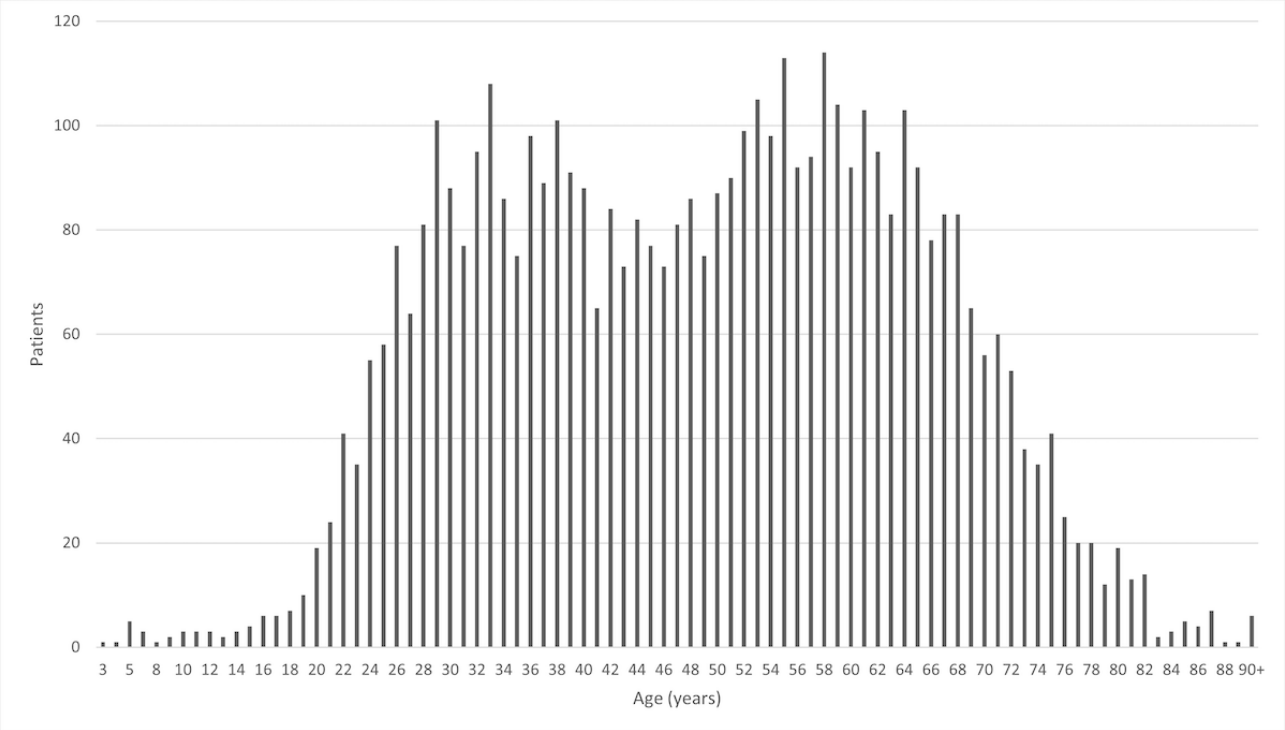

XSL•FO
RenderX

Figure 2. Ratio of mobile-to-web access per age group. A gradual decline in mobile use relative to web use is seen with progressing age, although this trend did not extend to those aged ≥90 years.

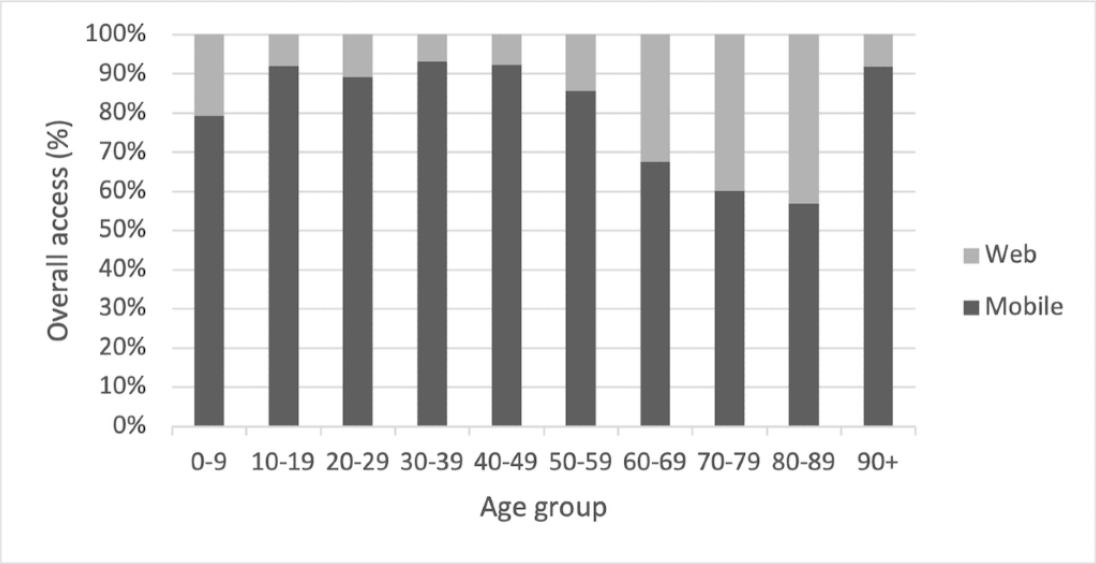

XSL•FO
RenderX

Table 1. Notes accessed under each clinical interaction type.

Clinical interaction type	Total documents, n	Unique documents, n
COVID-19	7476	4725
FM[a], IM[b], or primary care	3043	2213
Head and neck	1029	593
Hematology and oncology	912	654
Bariatrics	861	674
Gastroenterology	810	430
Cardiology	664	437
Infectious disease	653	474
Unknown	565	343
Rheumatology	475	300
Dental oncology	364	266
Pulmonology	339	209
Telemedicine	281	200
Nutrition	186	165
Occupational medicine	138	90
Orthopedics	119	70
Physiatry	94	63
Urology	73	47
Marijuana clinic	71	53
Oncology	49	30
Neurology	41	26
Pain	40	28
Cardiothoracic surgery	34	12
General surgery	34	23
Accident	31	23
Neurosurgery	1	1
Social worker	1	1
Total	18,384	12,150

[a]FM: family medicine.

[b]IM: internal medicine.

XSL•FO
RenderX

Figure 3. Number of patients viewing open notes each quarter (COVID-19 notes included). Q: quarter.

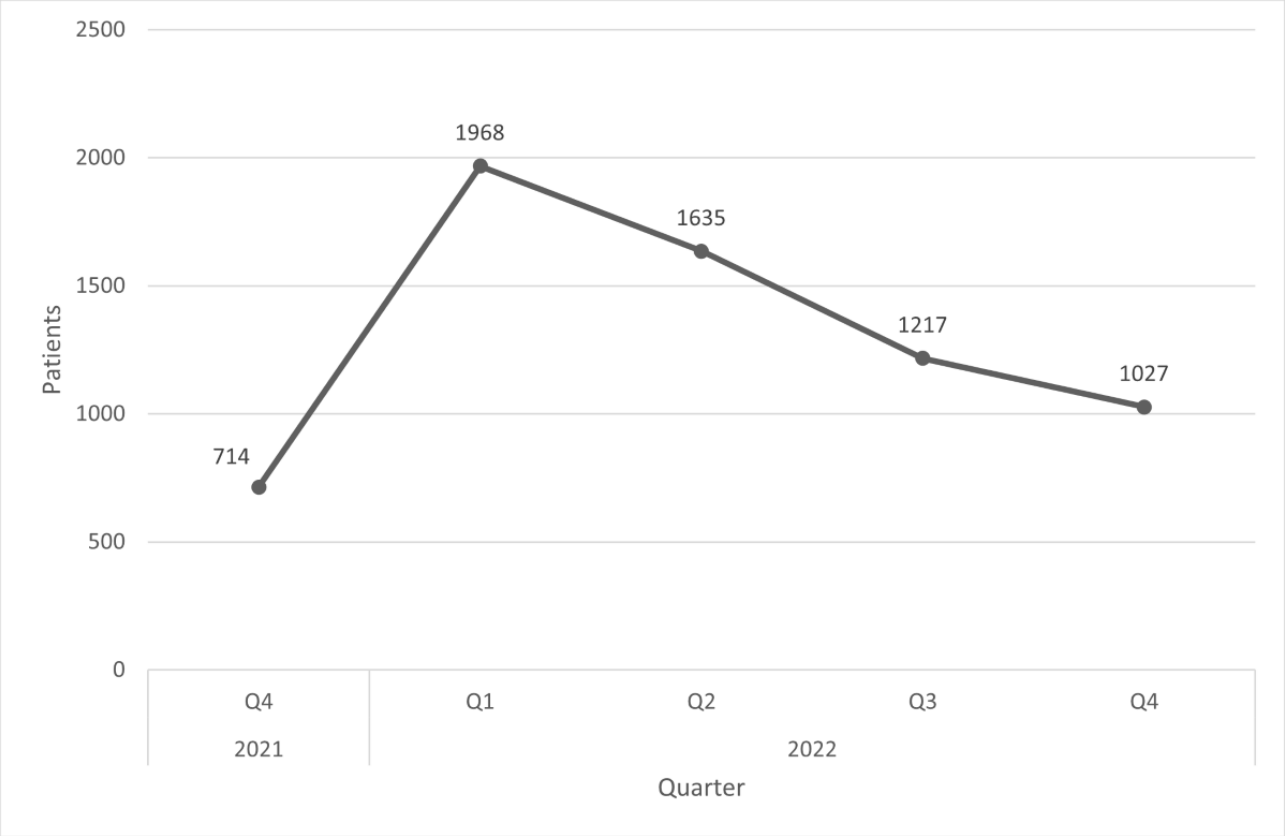

Figure 4. Number of patients viewing open notes each quarter (COVID-19 notes excluded). Q: quarter.

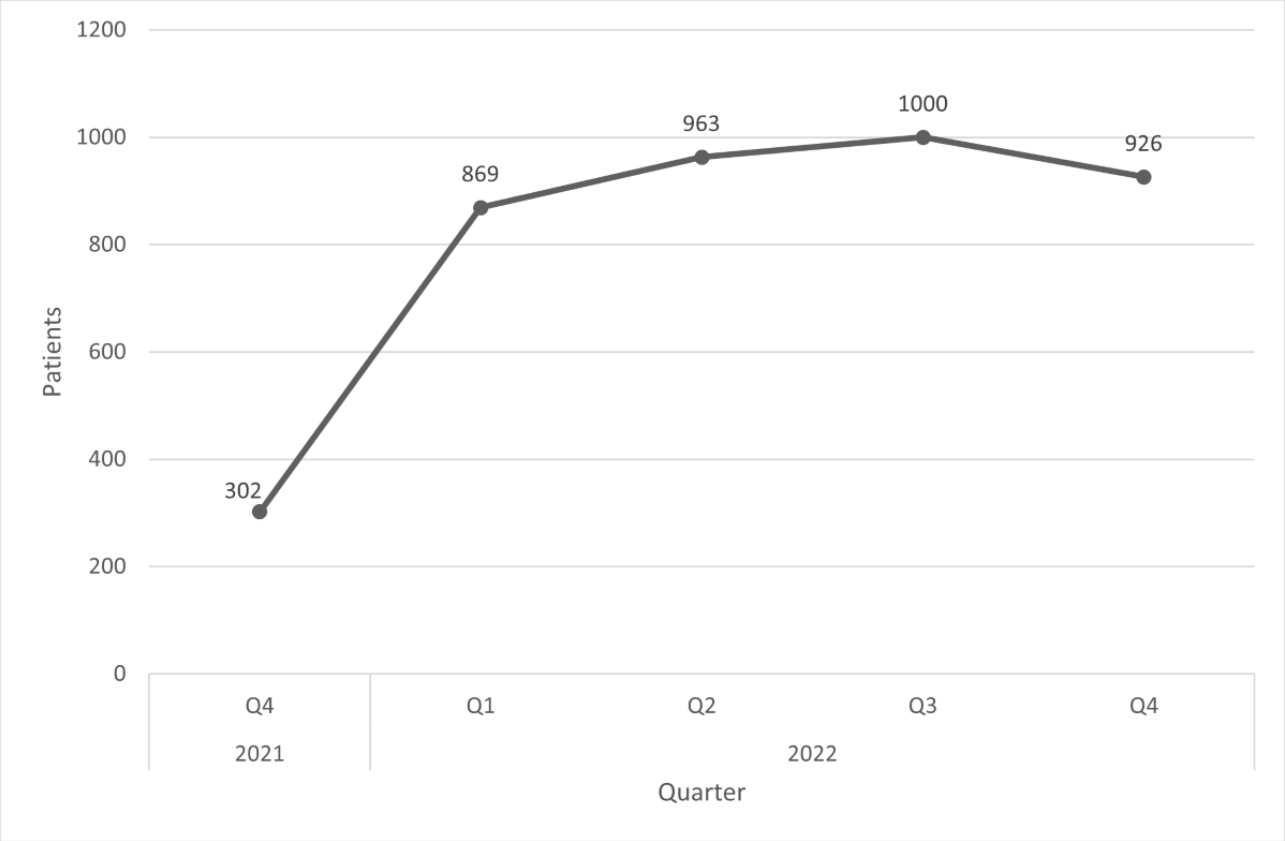

XSL•FO
RenderX

Discussion

Principal Findings

The primary purpose of the study was to quantify outpatient open note access and characterize its users at our institution. With 35,273 accounts in the portal, our study highlighted that 13.1% (4615/35,273) of portal registrants viewed their outpatient notes. This is in line with previous studies that found user rates of 13.7% [7] and 16.2% [8]. Some studies have also noted a female bias with respect to patient portal usage [9-12] and note access specifically [7,8], and our data showed that open note access is no different, with 63.4% (2926/4615) of patients being female. However, this number may be more reflective of the entire patient population connected to the portal (female: 20,047/35,273, 56.8%) rather than any sex-specific attributes of open note access. Interestingly, the bias was not as apparent in those aged 0-19 and ≥80 years, who may be managed by a proxy [11], but this was not delineated further in our study. It would be useful to explore whether there were sex disparities in the total number of clinic visits among users who accessed notes to gain further insight into this trend.

This study also uncovered a bimodal distribution in the age of patients that viewed open notes (30s and 50s). This finding is not unique in the literature, and it has been suggested that younger individuals may be more comfortable using the functions of the portal while older individuals are likely to be more medically complex [10] and have more notes to access. Furthermore, the percentage of notes accessed via a mobile device tended to decrease as a function of increasing age, indicating that web-based access of portal function is still a vital part of the way older patients interact with their data. However, this trend was not observed in those aged ≥90 years, who may have younger proxies managing their portal access [11]. Since this subset only consisted of 6 individuals in our study, it is imperative that further research be conducted with individuals within this age group to further give credence to this hypothesis of the possibility of proxy access.

When open notes went live, a sharp increase in the number of users accessing notes was observed, implying a high level of initial interest. The increase in the beginning may also be the result of more patients creating portal accounts and pursuing registration data would have created a useful point of comparison. However, as time evolved, the number of users accessing notes steadily declined by 47.8% (–941/1968). A likely explanation could be the content of the notes themselves: approximately 38.9% (4725/12,150) represented COVID-19 assessments and the falling levels may be due to fewer visits as the pandemic subsided. To further outline this trend, Figure 4 was plotted without COVID-19 assessments. Here, the number of users accessing open notes more gradually increased, reaching a peak in the third quarter of 2022, and ending with a slight decline of 7.4% (–74/1000) from the maximum. This suggests that outside of the pandemic context, user saturation may have been achieved among those interested in viewing open notes. The slight decrease at the end may be indicative of fading interest in open notes, although more data would be needed to establish whether this trend continued after the study period. We hypothesize that it will likely level off to a point below the maximum but still be appreciable, similar to what was observed for other digital tools such as telemedicine [13,14]. Future studies should compare open note users to all active users of the portal and examine factors that would increase user engagement.

Limitations

This study has several limitations. First, although 12,150 unique documents were accessed in the portal, we identified 1829 (15.1%) duplicates that we retained due to reasons mentioned in the methods. Since we relied on exact time stamp data to identify duplicates, system-related delays between the upload of identical documents to the portal may underestimate this value. In addition, the “clinical interaction type” data are also predicated on the assumption that all providers used the correct note title to document the encounter. Although we expect any deviations from this assumption to be relatively small, the contents of Table 1 may not be a true reflection of reality. This research led to a quality improvement project to improve the upload of documents so that a single note in the EHR corresponded to a single viewable link in the portal with a title reflective of the context. Second, although our analysis reported a 13.1% (4615/35,273) rate of open note use among all portal registrants, the number may be misleadingly low because the denominator included both outpatient and inpatient registrants. The 2 categories are not mutually exclusive but do not completely overlap either. Separating the outpatient from inpatient accounts would have required significantly complex queries with the use of more audit files, which we did not attempt to pursue in this study. Third, the portal likely contained dormant accounts from patients who registered but did not actively use it, and the definition of an “active user” could have multiple meanings [15], which was not taken into consideration during our study. Using a standard definition of active user would have allowed us to better compare open note use to other portal functions such as medication refills or provider messaging. Fourth, we did not pursue account registration data versus time nor the number of clinic visits by open note users, which was beyond the scope of the objectives. Obtaining this information would have allowed a greater understanding of the trends in open note use, and future studies are planned to explore this further. Finally, there may be errors with demographic data when patients register for the portal, such as inputting the wrong birthdate, which we did not quantify, although we expect this number to be small.

Conclusions

Despite the limitations, our study demonstrates that open note access was largely dominated by viewing COVID-19 assessments, and the number of users accessing their notes has decreased over time as the pandemic receded. Furthermore, female patients and those aged in their 30s as well as 50s viewed more notes than other groups, and web-based access of open notes remains an important modality for older patients.

XSL•FO
RenderX

Acknowledgments

The authors thank Peter L Elkin and the Department of Biomedical Informatics at the University at Buffalo for supporting this research and its publication. The authors also thank Erie County Medical Center for granting access to the data and facilitating its analysis.

Conflicts of Interest

None declared.

References

1. National trends in hospital and physician adoption of electronic health records. HealthIT.gov. URL: https://tinyurl.com/5e29sr9p [accessed 2024-11-16]
2. ONC's Cures Act Final Rule. HealthIT.gov. 2022. URL: https://www.healthit.gov/topic/oncs-cures-act-final-rule [accessed 2023-12-13]
3. U.S. federal rule mandates open notes. OpenNotes. URL: https://www.opennotes.org/onc-federal-rule/ [accessed 2023-12-13]
4. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
5. DesRoches CM, Bell SK, Dong Z, Elmore J, Fernandez L, Fitzgerald P, et al. Patients managing medications and reading their visit notes: a survey of OpenNotes participants. Ann Intern Med 2019;171(1):69-71. [doi: 10.7326/M18-3197] [Medline: 31132794]
6. Hospital capabilities to enable patient electronic access to health information, 2021. HealthIT.gov. 2022. URL: https://www.healthit.gov/data/data-briefs/hospital-capabilities-enable-patient-electronic-access-health-information-2021 [accessed 2023-07-28]
7. Sangal RB, Powers E, Rothenberg C, Ndumele C, Ulrich A, Hsiao A, et al. Disparities in accessing and reading open notes in the emergency department upon implementation of the 21st Century CURES Act. Ann Emerg Med 2021;78(5):593-598. [doi: 10.1016/j.annemergmed.2021.06.014] [Medline: 34353651]
8. Blok AC, Amante DJ, Hogan TP, Sadasivam RS, Shimada SL, Woods S, et al. Impact of patient access to online VA notes on healthcare utilization and clinician documentation: a retrospective cohort study. J Gen Intern Med 2021;36(3):592-599 [FREE Full text] [doi: 10.1007/s11606-020-06304-0] [Medline: 33443693]
9. Liu SK, Osborn AE, Bell S, Mecchella JN, Hort S, Batsis JA. Patient characteristics and utilization of an online patient portal in a rural academic general internal medicine practice. BMC Med Inform Decis Mak 2022;22(1):42 [FREE Full text] [doi: 10.1186/s12911-022-01778-w] [Medline: 35172805]
10. Tsai R, Bell EJ, Woo H, Baldwin K, Pfeffer MA. How patients use a patient portal: an institutional case study of demographics and usage patterns. Appl Clin Inform 2019;10(1):96-102 [FREE Full text] [doi: 10.1055/s-0038-1677528] [Medline: 30727003]
11. Oest SER, Hightower M, Krasowski MD. Activation and utilization of an electronic health record patient portal at an academic medical center-impact of patient demographics and geographic location. Acad Pathol 2018;5:2374289518797573 [FREE Full text] [doi: 10.1177/2374289518797573] [Medline: 30302394]
12. Casacchia NJ, Rosenthal GE, O'Connell NS, Bundy R, Witek L, Wells BJ, et al. Characteristics of adult primary care patients who use the patient portal: a cross-sectional analysis. Appl Clin Inform 2022;13(5):1053-1062 [FREE Full text] [doi: 10.1055/a-1951-3153] [Medline: 36167336]
13. Pullyblank K, Krupa N, Scribani M, Chapman A, Kern M, Brunner W. Trends in telehealth use among a cohort of rural patients during the COVID-19 pandemic. Digit Health 2023;9:20552076231203803 [FREE Full text] [doi: 10.1177/20552076231203803] [Medline: 37799503]
14. Hung M, Ocampo M, Raymond B, Mohajeri A, Lipsky MS. Telemedicine among adults living in America during the COVID-19 pandemic. Int J Environ Res Public Health 2023;20(9):5680 [FREE Full text] [doi: 10.3390/ijerph20095680] [Medline: 37174198]
15. Beal LL, Kolman JM, Jones SL, Khleif A, Menser T. Quantifying patient portal use: systematic review of utilization metrics. J Med Internet Res 2021;23(2):e23493 [FREE Full text] [doi: 10.2196/23493] [Medline: 33629962]

Abbreviations

EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
ID: infectious disease
IRB: institutional review board

XSL•FO
RenderX

Edited by G Tsafnat; submitted 02.01.24; peer-reviewed by M Avdagovska, J Sorra; comments to author 12.03.24; revised version received 04.08.24; accepted 20.10.24; published 17.12.24.

Please cite as:
Badwal RS, Cavo P, Panesar M
Insights and Trends in Open Note Access: Retrospective Observational Study
J Med Internet Res 2024;26:e55982
URL: https://www.jmir.org/2024/1/e55982
doi:10.2196/55982
PMID:

©Randeep Singh Badwal, Paul Cavo, Mandip Panesar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Psychometric Evaluation of a Tablet-Based Tool to Detect Mild Cognitive Impairment in Older Adults: Mixed Methods Study

Josephine McMurray[1,2*], MBA, PhD; AnneMarie Levy[1*], MSc, PhD; Wei Pang[1,3*], BTM; Paul Holyoke[4], PhD

[1]Lazaridis School of Business & Economics, Wilfrid Laurier University, Brantford, ON, Canada

[2]Health Studies, Faculty of Human and Social Sciences, Wilfrid Laurier University, Brantford, ON, Canada

[3]Biomedical Informatics & Data Science, Yale University, New Haven, CT, United States

[4]SE Research Centre, Markham, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Josephine McMurray, MBA, PhD
Lazaridis School of Business & Economics
Wilfrid Laurier University
73 George St
Brantford, ON, N3T3Y3
Canada
Phone: 1 548 889 4492
Email: jmcmurray@wlu.ca

Abstract

Background: With the rapid aging of the global population, the prevalence of mild cognitive impairment (MCI) and dementia is anticipated to surge worldwide. MCI serves as an intermediary stage between normal aging and dementia, necessitating more sensitive and effective screening tools for early identification and intervention. The BrainFx SCREEN is a novel digital tool designed to assess cognitive impairment. This study evaluated its efficacy as a screening tool for MCI in primary care settings, particularly in the context of an aging population and the growing integration of digital health solutions.

Objective: The primary objective was to assess the validity, reliability, and applicability of the BrainFx SCREEN (hereafter, the SCREEN) for MCI screening in a primary care context. We conducted an exploratory study comparing the SCREEN with an established screening tool, the Quick Mild Cognitive Impairment (Qmci) screen.

Methods: A concurrent mixed methods, prospective study using a quasi-experimental design was conducted with 147 participants from 5 primary care Family Health Teams (FHTs; characterized by multidisciplinary practice and capitated funding) across southwestern Ontario, Canada. Participants included health care practitioners, patients, and FHT administrative executives. Individuals aged ≥55 years with no history of MCI or diagnosis of dementia rostered in a participating FHT were eligible to participate. Participants were screened using both the SCREEN and Qmci. The study also incorporated the Geriatric Anxiety Scale–10 to assess general anxiety levels at each cognitive screening. The SCREEN's scoring was compared against that of the Qmci and the clinical judgment of health care professionals. Statistical analyses included sensitivity, specificity, internal consistency, and test-retest reliability assessments.

Results: The study found that the SCREEN's longer administration time and complex scoring algorithm, which is proprietary and unavailable for independent analysis, presented challenges. Its internal consistency, indicated by a Cronbach α of 0.63, was below the acceptable threshold. The test-retest reliability also showed limitations, with moderate intraclass correlation coefficient (0.54) and inadequate κ (0.15) values. Sensitivity and specificity were consistent (63.25% and 74.07%, respectively) between cross-tabulation and discrepant analysis. In addition, the study faced limitations due to its demographic skew (96/147, 65.3% female, well-educated participants), the absence of a comprehensive *gold standard* for MCI diagnosis, and financial constraints limiting the inclusion of confirmatory neuropsychological testing.

Conclusions: The SCREEN, in its current form, does not meet the necessary criteria for an optimal MCI screening tool in primary care settings, primarily due to its longer administration time and lower reliability. As the number of digital health technologies increases and evolves, further testing and refinement of tools such as the SCREEN are essential to ensure their efficacy and reliability in real-world clinical settings. This study advocates for continued research in this rapidly advancing field to better serve the aging population.

XSL•FO
RenderX

International Registered Report Identifier (IRRID): RR2-10.2196/25520

(J Med Internet Res 2024;26:e56883) doi:10.2196/56883

KEYWORDS

cognitive dysfunction; dementia neuropsychological tests; evaluation study; technology; aged; mobile phone

Introduction

Background

Mild cognitive impairment (MCI) is a syndrome characterized by a slight but noticeable and measurable deterioration in cognitive abilities, predominantly memory and thinking skills, that is greater than expected for an individual's age and educational level [1,2]. The functional impairments associated with MCI are subtle and often impair instrumental activities of daily living (ADL). Instrumental ADL include everyday tasks such as managing finances, cooking, shopping, or taking regularly prescribed medications and are considered more complex than ADL such as bathing, dressing, and toileting [3,4]. In cases in which memory impairment is the primary indicator of the disease, MCI is classified as amnesic MCI and when significant impairment of non–memory-related cognitive domains such as visual-spatial or executive functioning is dominant, MCI is classified as nonamnesic [5].

Cognitive decline, more so than cancer and cardiovascular disease, poses a substantial threat to an individual's ability to live independently or at home with family caregivers [6]. The Centers for Disease Control and Prevention reports that 1 in 8 adults aged ≥60 years experiences memory loss and confusion, with 35% reporting functional difficulties with basic ADL [7]. The American Academy of Neurology estimates that the prevalence of MCI ranges from 13.4% to 42% in people aged ≥65 years [8], and a 2023 meta-analysis that included 233 studies and 676,974 participants aged ≥50 years estimated that the overall global prevalence of MCI is 19.7% [9]. Once diagnosed, the prognosis for MCI is variable, whereby the impairment may be reversible; the rate of decline may plateau; or it may progressively worsen and, in some cases, may be a prodromal stage to dementia [10-12]. While estimates vary based on sample (community vs clinical), annual rates of conversion from MCI to dementia range from 5% to 24% [11,12], and those who present with multiple domains of cognitive impairment are at higher risk of conversion [5].

The risk of developing MCI rises with age, and while there are no drug treatments for MCI, nonpharmacologic interventions may improve cognitive function, alleviate the burden on caregivers, and potentially delay institutionalization should MCI progress to dementia [13]. To overcome the challenges of early diagnosis, which currently depends on self-detection, family observation, or health care provider (HCP) recognition of symptoms, screening high-risk groups for MCI or dementia is suggested as a solution [13]. However, the Canadian Task Force on Preventive Health Care recommends against screening adults aged ≥65 years due to a lack of meaningful evidence from randomized controlled trials and the high false-positive rate [14-16]. The main objective of a screening test is to reduce morbidity or mortality in at-risk populations through early detection and intervention, with the anticipated benefits outweighing potential harms. Using brief screening tools in primary care might improve MCI case detection, allowing patients and families to address reversible causes, make lifestyle changes, and access disease-modifying treatments [17].

There is no agreement among experts as to which tests or groups of tests are most predictive of MCI [16], and the gold standard approach uses a combination of positive results from neuropsychological assessments, laboratory tests, and neuroimaging to infer a diagnosis [8,18]. The clinical heterogeneity of MCI complicates its diagnosis because it influences not only memory and thinking abilities but also mood, behavior, emotional regulation, and sensorimotor abilities, and patients may present with any combination of symptoms with varying rates of onset and decline [4,8]. For this reason, a collaborative approach between general practitioners and specialists (eg, geriatricians and neurologists) is often required to be confident in the diagnosis of MCI [8,19,20].

In Canada, diagnosis often begins with screening for cognitive impairment followed by referral for additional testing; this process takes, on average, 5 months [20]. The *current usual practice* screening tools for MCI are the Mini-Mental State Examination (MMSE) [21,22] and the Montreal Cognitive Assessment (MoCA) 8.1 [3]. Both are paper-and-pencil screens administered in 10 to 15 minutes, scored out of 30, and validated as MCI screening tools across diverse clinical samples [23,24]. Universally, the MMSE is most often used to screen for MCI [20,25] and consists of 20 items that measure orientation, immediate and delayed recall, attention and calculation, visual-spatial skills, verbal fluency, and writing. The MoCA 8.1 was developed to improve on the MMSE's ability to detect early signs of MCI, placing greater emphasis on evaluating executive function as well as language, memory, visual-spatial skills, abstraction, attention, concentration, and orientation across 30 items [24,26]. Scores of <24 on the MMSE or ≤25 on the MoCA 8.1 signal probable MCI [21,27]. Lower cutoff scores for both screens have been recommended to address evidence that they lack specificity to detect mild and early cases of MCI [4,28-31]. The clinical efficacy of both screens for tracking change in cognition over time is limited as they are also subject to practice effects with repeated administration [32].

Novel screening tools, including the Quick Mild Cognitive Impairment (Qmci) screen, have been developed with the goal of improving the accuracy of detecting MCI [33,34]. The Qmci is a sensitive and specific tool that differentiates normal cognition from MCI and dementia and is more accurate at differentiating MCI from controls than either the MoCA 8.1 (Qmci area under the curve=0.97 vs MoCA 8.1 area under the curve=0.92) [25,35] or the Short MMSE [33,36]. It also demonstrates high test-retest reliability (intraclass correlation

XSL•FO
RenderX

coefficient [ICC]=0.88) [37] and is clinically useful as a rapid screen for MCI as the Qmci mean is 4.5 (SD 1.3) minutes versus 9.5 (SD 2.8) minutes for the MoCA 8.1 [25].

The COVID-19 pandemic and the necessary shift to virtual health care accelerated the use of digital assessment tools, including MCI screening tools such as the electronic MoCA 8.1 [38,39], and the increased use and adoption of technology (eg, smartphones and tablets) by older adults suggests that a lack of proficiency with technology may not be a barrier to the use of such assessment tools [40,41]. BrainFx is a for-profit firm that creates proprietary software designed to assess cognition and changes in neurofunction that may be caused by neurodegenerative diseases (eg, MCI or dementia), stroke, concussions, or mental illness using ecologically relevant tasks (eg, prioritizing daily schedules and route finding on a map) [42]. Their assessments are administered via a tablet and stylus. The BrainFx 360 performance assessment (referred to hereafter as the 360) is a 90-minute digitally administered test that was designed to assess cognitive, physical, and psychosocial areas of neurofunction across 26 cognitive domains using 49 tasks that are timed and scored [42]. The BrainFx SCREEN (referred to hereafter as the SCREEN) is a short digital version of the 360 that includes 7 of the cognitive domains included in the 360, is estimated to take approximately 10 to 15 minutes to complete, and was designed to screen for early detection of cognitive impairment [43,44]. Upon completion of any BrainFx assessment, the results of the 360 or SCREEN are added to the BrainFx Living Brain Bank (LBB), which is an electronic database that stores all completed 360 and SCREEN assessments and is maintained by BrainFx. An electronic report is generated by BrainFx comparing an individual's results to those of others collected and stored in the LBB. Normative data from the LBB are used to evaluate and compare an individual's results.

The 360 has been used in clinical settings to assess neurofunction among youth [45] and anecdotally in other rehabilitation settings (T Milner, personal communication, May 2018). To date, research on the 360 indicates that it has been validated in healthy young adults (mean age 22.9, SD 2.4 years) and that the overall test-retest reliability of the tool is high (ICC=0.85) [42]. However, only 2 of the 7 tasks selected to be included in the SCREEN produced reliability coefficients of >0.70 (visual-spatial and problem-solving abilities) [42]. Jones et al [43] explored the acceptability and perceived usability of the SCREEN with a small sample (N=21) of Canadian Armed Forces veterans living with posttraumatic stress disorder. A structural equation model based on the Unified Theory of Acceptance and Use of Technology suggested that behavioral intent to use the SCREEN was predicted by facilitating conditions such as guidance during the test and appropriate resources to complete the test [43]. However, the validity, reliability, and sensitivity of the SCREEN for detecting cognitive impairment have not been tested.

Objectives

McMurray et al [44] designed a protocol to assess the validity, reliability, and sensitivity of the SCREEN for detecting early signs of MCI in asymptomatic adults aged ≥55 years in a primary care setting (5 Family Health Teams [FHTs]). The protocol also used a series of semistructured interviews and surveys guided by the fit between individuals, task, technology, and environment framework [46], a health-specific model derived from the Task-Technology Fit model by Goodhue and Thompson [47], to explore the SCREEN's acceptability and use by HCPs and patients in primary care settings (manuscript in preparation). This study is a psychometric evaluation of the SCREEN's validity, reliability, and sensitivity for detecting MCI in asymptomatic adults aged ≥55 years in primary care settings.

Methods

Study Location, Design, and Data Collection

This was a concurrent, mixed methods, prospective study using a quasi-experimental design. Participants were recruited from 5 primary care FHTs (characterized by multidisciplinary practice and capitated funding) across southwestern Ontario, Canada. FHTs that used a registered occupational therapist on staff were eligible to participate in the study, and participating FHTs received a nominal compensatory payment for the time the HCPs spent in training; collecting data for the study; administering the SCREEN, Qmci, and Geriatric Anxiety Scale–10 (GAS-10); and communicating with the research team. A multipronged recruitment approach was used [44]. A designated occupational therapist at each location was provided with training and equipment to recruit participants, administer assessment tools, and submit collected data to the research team.

The research protocol describing the methods of both the quantitative and qualitative arms of the study is published elsewhere [44].

Ethical Considerations

This study was approved by the Wilfrid Laurier University Research Ethics Board (ORE 5820) and was reviewed and approved by each FHT. Participants (HCPs, patients, and administrative executives) read and signed an information and informed consent package in advance of taking part in the study. We complied with recommendations for obtaining informed consent and conducting qualitative interviews with persons with dementia when recruiting patients who may be affected by neurocognitive diseases [48-50]. In addition, at the end of each SCREEN assessment, patients were required to provide their consent (electronic signature) to contribute their anonymized scores to the database of SCREEN results maintained by BrainFx. Upon enrolling in the study, participants were assigned a unique identification number that was used in place of their name on all study documentation to anonymize the data and preserve their confidentiality. A master list matching participant names with their unique identification number was stored in a password-protected file by the administering HCP and principal investigator on the research team. The FHTs received a nominal compensatory payment to account for their HCPs' time spent administering the SCREEN, collecting data for the study, and communicating with the research team. However, the individual HCPs who volunteered to participate and the patient participants were not financially compensated for taking part in the study.

XSL•FO
RenderX

Participants

Patients who were rostered with the FHT, were aged ≥55 years, and had no history of MCI or dementia diagnoses to better capture the population at risk of early signs of cognitive impairment were eligible to participate [51,52]. It was necessary for the participants to be rostered with the FHTs to ensure that the HCPs could access their electronic medical record to confirm eligibility and record the testing sessions and results and to ensure that there was a responsible physician for referral if indicated. As the SCREEN is administered using a tablet, participants had to be able to read and think in English and discern color, have adequate hearing and vision to interact with the administering HCP, read 12-point font on the tablet, and have adequate hand and arm function to manipulate and hold the tablet. The exclusion criteria used in the study included colorblindness and any disability that might impair the individual's ability to hold and interact with the tablet. Prospective participants were also excluded based on a diagnosis of conditions that may result in MCI or dementia-like symptoms, including major depression that required hospitalization, psychiatric disorders (eg, schizophrenia and bipolar disorder), psychopathology, epilepsy, substance use disorders, or sleep apnea (without the use of a continuous positive airway pressure machine) [52]. Patients were required to complete a minimum of 2 screening sessions spaced 3 months apart to participate in the study and, depending on when they enrolled to participate, could complete a maximum of 4 screening sessions over a year.

Data Collection Instruments

GAS-10 Instrument

A standardized protocol was used to collect demographic data, randomly administer the SCREEN and the Qmci (a validated screening tool for MCI), and administer the GAS-10 immediately before and after the completion of the first MCI screen at each visit [44]. This was to assess participants' general anxiety as it related to screening for cognitive impairment at the time of the assessment, any change in subjective ratings after completion of the first MCI screen, and change in anxiety between appointments. The GAS-10 is a 10-item, self-report screen for anxiety in older adults [53] developed for rapid screening of anxiety in clinical settings (the GAS-10 is the short form of the full 30-item Geriatric Anxiety Scale [GAS]) [54]. While 3 subscales are identified, the GAS is reported to be a unidimensional scale that assesses general anxiety [55,56]. Validation of the GAS-10 suggests that it is optimal for assessing average to moderate levels of anxiety in older adults, with subscale scores that are highly and positively correlated with the GAS and high internal consistency [53]. Participants were asked to use a 4-point Likert scale (0=*not at all*, 1=*sometimes*, 2=*most of the time*, and 3=*all of the time*) to rate how often they had experienced each symptom over the previous week, including on the day the test was administered [54]. The GAS-10 has a maximum score of 30, with higher scores indicating higher levels of anxiety [53,54,57].

The SCREEN

HCPs completed the required training to become certified BrainFx SCREEN administrators before the start of the study. To this end, HCPs completed a web-based training program (developed and administered through the BrainFx website) that included 3 self-directed training modules. For the purpose of the study, they also participated in 1 half-day in-person training session conducted by a certified BrainFx administrator (T Milner, BrainFx chief executive officer) at one of the participating FHT locations. The SCREEN (version 0.5; beta) was administered on a tablet (ASUS ZenPad 10.1" IPS WXGA display, 1920 × 1200, powered by a quad-core 1.5 GHz, 64-bit MediaTek MTK 8163A processor with 2 GB RAM and 16-GB storage). The tablet came with a tablet stand for optional use and a dedicated stylus that is recommended for completion of a subset of questions. At the start of the study, HCPs were provided with identical tablets preloaded with the SCREEN software for use in the study. The 7 tasks on the SCREEN are summarized in Table 1 and were taken directly from the 360 based on a clustering and regression analysis of LBB records in 2016 (N=188) [58]. A detailed description of the study and SCREEN administration procedures was published by McMurray et al [44].

XSL•FO
RenderX

Table 1. Summary of the 7 SCREEN tasks and the length of time allotted to complete them.

Task	Description	Time to complete
Abstract reasoning	20 everyday items are displayed, and the patient touches the item on the screen and slides each item, one at a time, into 1 of 5 categories to which they best belong while timed.	90 seconds
Constructive ability	2 rounds of a photo are displayed, and it is broken into 9 pieces. The patient touches the pieces and slides each piece into a grid to reassemble the picture while timed.	90 seconds
Prioritizing	5 everyday activities or tasks are presented, the patient is told what time of day it is (eg, 7 PM), and they touch the screen and slide each item to prioritize the order in which the activities or tasks should be completed.	60 seconds
Numerical problem-solving	10 math questions, with 1- or 2-digit answers, are presented for patient response using a numerical pad (+, –, ×, and /) while timed.	90 seconds
Visual-spatial ability	2 rounds of the patient selecting (by touch) into which shape a word fits best while timed.	30 seconds
Divided attention	The patient watches a pot on the stove about to boil over (denoted by boiling water and a red signal) and must touch the pot and move it to the sink to dump out the water while also touching the screen to match as many objects as they can within the kitchen scene.	90 seconds
Route finding	In round 1, the patient traces the most efficient route between 2 locations while timed. In round 2, the patient traces the most efficient route between 2 locations but is instructed to make 2 stops on the way while timed.	90 seconds

An *activity score* is generated for each of the 7 tasks on the SCREEN. It is computed based on a combination of the accuracy of the participant's response and the processing speed (time in seconds) that it takes to complete the task. The relative contribution of accuracy and processing speed to the final activity score for each task is proprietary to BrainFx and unknown to the research team. The participant's activity score is compared to the mean activity score for the same task at the time of testing in the LBB. The mean activity score from the LBB may be based on the global reference population (ie, all available SCREEN results in the LBB), or the administering HCP may select a specific reference population by filtering according to factors including but not limited to age, sex, or diagnosis. If the participant's activity score is >1 SD below the LBB activity score mean for that task, it is labeled as an *area of challenge*. Each of the 7 tasks on the SCREEN are evaluated independently of each other, producing a report with 7 activity scores showing the participant's score, the LBB mean score, and the SD. The report also provides an overall performance and processing speed score. The overall performance score is an average of all 7 activity scores; however, the way in which the overall processing speed score is generated remains proprietary to BrainFx and unknown to the research team. Both the overall performance and processing speed scores are similarly evaluated against the LBB and identified as an area of challenge using the criteria described previously. For the purpose of this study, participants' mean activity scores on the SCREEN were compared to the results of people aged ≥55 years in the LBB.

The Qmci

The Qmci evaluated 6 cognitive domains: orientation (10 points), registration (5 points), clock drawing (15 points), delayed recall (20 points), verbal fluency (20 points), and logical memory (30 points) [59]. Administering HCPs scored the text manually, with each subtest's points contributing to the overall score out of 100 points, and the cutoff score to distinguish normal cognition from MCI was ≤67/100 [60]. Cutoffs to account for age and education have been validated and are recommended as the Qmci is sensitive to these factors [60]. A 2019 meta-analysis of the diagnostic accuracy of MCI screening tools reported that the sensitivity and specificity of the Qmci for distinguishing MCI from normal cognition is similar to usual standard-of-care tools (eg, the MoCA, Addenbrooke Cognitive Examination–Revised, Consortium to Establish a Registry for Alzheimer's Disease battery total score, and Sunderland Clock Drawing Test) [61]. The Qmci has also been translated into >15 different languages and has undergone psychometric evaluation across a subset of these languages. While not as broadly adopted as the MoCA 8.1 in Canada, its psychometric properties, administration time, and availability for use suggested that the Qmci was an optimal assessment tool for MCI screening in FHT settings during the study.

Psychometric Evaluation

Overview

To date, the only published psychometric evaluation of any BrainFx tool is by Searles et al [42] in *Athletic Training & Sports Health Care*; it assessed the test-retest reliability of the 360 in 15 healthy adults between the ages of 20 and 25 years. This study evaluated the psychometric properties of the SCREEN and included a statistical analysis of the tool's internal consistency, construct validity, test-retest reliability, and sensitivity and specificity. McMurray et al [44] provide a detailed description of the data collection procedures for administration of the SCREEN and Qmci completed by participants at each visit.

Validity Testing

Face validity was outside the scope of this study but was implied, and assumptions are reported in the Results section. Construct validity, whether the 7 activities that make up the SCREEN were representative of MCI, was assessed through comparison with a substantive body of literature in the domain and through principal component analysis using varimax rotation. Criterion validity measures how closely the SCREEN results corresponded to the results of the Qmci (used here as an

XSL•FO
RenderX

"imperfect gold standard" for identifying MCI in older adults) [62]. A BrainFx representative hypothesized that the ecological validity of the SCREEN questions (ie, using tasks that reflect real-world activities to detect early signs of cognitive impairment) [63] makes it a more sensitive tool than other screens (T Milner, personal communication, May 2018) and allows HCPs to equate activity scores on the SCREEN with real-world functional abilities. Criterion validity was explored first using cross-tabulations to calculate the sensitivity and specificity of the SCREEN compared to those of the Qmci. Conventional screens such as the Qmci are scored by taking the sum of correct responses on the screen and a cutoff score derived from normative data to distinguish normal cognition from MCI. The SCREEN used a different method of scoring whereby each of the 7 tasks was scored and evaluated independently of each other and there were no recommended guidelines for distinguishing normal cognition from MCI based on the aggregate areas of challenge identified by the SCREEN. Therefore, to compare the sensitivity and specificity of the SCREEN against those of the Qmci, the results of both screens were coded into a binary format as 1=healthy and 2=unhealthy, where *healthy* denoted no areas of challenge identified through the SCREEN and a Qmci score of ≥67. Conversely, *unhealthy* denoted one or more areas of challenge identified through the SCREEN and a Qmci score of <67.

Criterion validity was further explored using discrepant analysis via a resolver test [44]. Following the administration of the SCREEN and Qmci, screen results were evaluated by the administering HCP. HCPs were instructed to refer the participant for follow-up with their primary care physician if the Qmci result was <67 regardless of whether any areas of challenge were identified on the SCREEN. However, HCPs could use their clinical judgment to refer a participant for physician follow-up based on the results of the SCREEN or the Qmci, and all the referral decisions were charted on the participant's electronic medical record following each visit and screening. In discrepant analysis, the results of the *imperfect gold standard* [64], as was the role of the Qmci in this study, were compared with the SCREEN results. A resolver test (classified as whether the HCP referred the patient to a physician for follow-up based on their performance on the SCREEN and the Qmci) was used on discordant results [64,65] to determine sensitivity and specificity. To this end, a new variable, *Referral to a Physician for Cognitive Impairment*, was coded as the *true status* (1=no referral; 2=referral was made) and compared to the Qmci as the *imperfect gold standard* (1=healthy; 2=unhealthy).

Reliability Testing

The reliability of a screening instrument is its ability to consistently measure an attribute and how well its component measures fit together conceptually. Internal consistency identifies whether the items in a multi-item scale are measuring the same underlying construct; the internal consistency of the SCREEN was assessed using the Cronbach α. Test-retest reliability refers to the ability of a measurement instrument to reproduce results over ≥2 occasions (assuming the underlying conditions have not changed) and was assessed using paired *t* tests (2-tailed), ICC, and the κ coefficient. In this study, participants completed both the SCREEN and the Qmci in the same sitting in a random sequence on at least 2 different occasions spaced 3 months apart (administration procedures are described elsewhere) [44]. In some instances, the screens were administered to the same participant on 4 separate occasions spaced 3 months apart each, and this provided up to 3 separate opportunities to conduct test-retest reliability analyses and investigate the effects of repeated practice. There are no clear guidelines on the optimal time between tests [66,67]; however, Streiner and Kottner [68] and Streiner [69] recommend longer periods between tests (eg, at least 10-14 days) to avoid recall bias, and greater practice effects have been experienced with shorter test-retest intervals [32].

Statistics

Analysis of the quantitative data was completed using Stata (version 17.0; StataCorp). Assumptions of normality were not violated, so parametric tests were used. Collected data were reported using frequencies and percentages and compared using the chi-square or Fisher exact test as necessary. Continuous data were analyzed for central tendency and variability; categoric data were presented as proportions. Normality was tested using the Shapiro-Wilk test, and nonparametric data were tested using the Mann-Whitney *U* test. A *P* value of .05 was considered statistically significant, with 95% CIs provided where appropriate. We powered the exploratory analysis to validate the SCREEN using an estimated effect size of 12%—understanding that Canadian prevalence rates of MCI were not available [1]—and determined that the study required at least 162 participants. For test-retest reliability, using 90% power and a 5% type-I error rate, a minimum of 67 test results was required.

The time taken for participants to complete the SCREEN was recorded by the HCPs at the time of testing; there were 6 missing HCP records of time to complete the SCREEN. For these 6 cases of missing data, we imputed the mean time to complete the SCREEN by all participants who were tested by that HCP and used this to populate the missing cells [70]. There were 3 cases of missing data related to the SCREEN reports. More specifically, the SCREEN report generated by BrainFx did not include 1 or 2 data points each for the route finding, divided attention, and prioritizing tasks. The clinical notes provided by the HCP at the time of SCREEN administration did not indicate that the participant had not completed those questions, and it was not possible to determine the root cause of the missing data in report generation according to BrainFx (M Milner, personal communication, July 7, 2020). For continuous variables in analyses such as exploratory factor analysis, Cronbach α, and *t* test, missing values were imputed using the mean. However, for the coded healthy and unhealthy categorical variables, values were not imputed.

Results

Data collection began in January 2019 and was to conclude on May 31, 2020. However, the emergence of the global COVID-19 pandemic resulted in the FHTs and Wilfrid Laurier University prohibiting all in-person research starting on March 16, 2020.

XSL•FO
RenderX

Participant Demographics

A total of 154 participants were recruited for the study, and 20 (13%) withdrew following their first visit to the FHT. The data of 65% (13/20) of the participants who withdrew were included in the final analysis, and the data of the remaining 35% (7/20) were removed, either due to their explicit request (3/7, 43%) or because technical issues at the time of testing rendered their data unusable (4/7, 57%). These technical issues were related to software issues (eg, any instance in which the patient or HCP interacted with the SCREEN software and followed the instructions provided, the software did not work as expected [ie, objects did not move where they were dragged or tapping on objects failed to highlight the object], and the question could not be completed). After attrition, a total of 147 individuals aged ≥55 years with no previous diagnosis of MCI or dementia participated in the study (Table 2). Of the 147 participants, 71 (48.3%) took part in only 1 round of screening on visit 1 (due to COVID-19 restrictions imposed on in-person research that prevented a second visit). The remaining 51.7% (76/147) of the participants took part in ≥2 rounds of screening across multiple visits (76/147, 51.7% participated in 2 rounds; 22/147, 15% participated in 3 rounds; and 13/147, 8.8% participated in 4 rounds of screening).

Table 2. Study participant demographics (highest level of education attained and age; N=147).

Characteristics	Total, n (%)	Female participants (n=96, 65.3%), n (%)	Male participants (n=51, 34.7%), n (%)
Education			
Lower than high school	6 (4.1)	5 (3.4)	1 (0.7)
High school	45 (30.6)	30 (20.4)	15 (10.2)
College diploma or certificate	39 (26.5)	27 (18.4)	12 (8.2)
University degree	34 (23.1)	22 (15)	12 (8.2)
Postgraduate degree	23 (15.6)	12 (8.2)	11 (7.5)
Age (years)			
55-59	14 (9.5)	10 (6.8)	4 (2.7)
60-64	17 (11.6)	12 (8.2)	5 (3.4)
65-69	35 (23.8)	25 (17)	10 (6.8)
70-74	33 (22.4)	23 (15.6)	10 (6.8)
75-79	25 (17)	12 (8.2)	13 (8.8)
80-84	14 (9.5)	9 (6.1)	5 (3.4)
85-89	9 (6.1)	5 (3.4)	4 (2.7)
≥90	0 (0)	0 (0)	0 (0)

The sample population was 65.3% (96/147) female (mean 70.2, SD 7.9 years) and 34.7% (51/147) male (mean 72.5, SD 8.1 years), with age ranging from 55 to 88 years; 65.3% (96/147) achieved the equivalent of or higher than a college diploma or certificate (Table 2); and 32.7% (48/147) self-reported living with one or more chronic medical conditions (Table 3). At the time of screening, 73.5% (108/147) of participants were also taking medications with side effects that may include impairments to memory and thinking abilities [71-75]; therefore, medication use was accounted for in a subset of the analyses. Finally, 84.4% (124/147) of participants self-reported regularly using technology (eg, smartphone, laptop, or tablet) with high proficiency. A random sequence generator was used to determine the order for administering the MCI screens; the SCREEN was administered first 51.9% (134/258) of the time.

XSL•FO
RenderX

Table 3. Self-reported status of participants' health with respect to preexisting conditions and chronic disease types (N=147).

Participant health status	Participants, n (%)
Preexisting conditions	
None	99 (67.3)
Single condition	37 (25.2)
Comorbidity	7 (4.8)
Multimorbidity	4 (2.7)
Chronic disease types reported	
Cardiovascular diseases	7 (4.8)
Chronic respiratory diseases	3 (2)
Diabetes	5 (3.4)
Musculoskeletal disorders	10 (6.8)
Neurological conditions	0 (0)
Cancer	6 (4.1)
Mental illnesses	19 (12.9)
Other (eg, autoimmune or chronic pain)	11 (7.5)

Validity Testing

Construct Validity

Construct validity was assessed through a review of relevant peer-reviewed literature that compared constructs included in the SCREEN with those identified in the literature as 2 of the most sensitive tools for MCI screening: the MoCA 8.1 [76] and the Qmci [25]. Memory, language, and verbal skills are assessed in the MoCA and Qmci but are absent from the SCREEN. Tests of verbal fluency and logical memory have been shown to be particularly sensitive to early cognitive changes [77,78] but are similarly absent from the SCREEN.

Exploratory factor analysis was performed to examine the SCREEN's ability to reliably measure risk of MCI. The Kaiser-Meyer-Olkin measure yielded a value of 0.79, exceeding the commonly accepted threshold of 0.70, indicating that the sample was adequate for factor analysis. The Bartlett test of sphericity returned a chi-square value of χ^2_{21}=167.1 (P<.001), confirming the presence of correlations among variables suitable for factor analysis. A principal component analysis revealed 2 components with eigenvalues of >1, cumulatively accounting for 52.12% of the variance, with the first factor alone explaining 37.8%. After the varimax rotation, the 2 factors exhibited distinct patterns of loadings, with the visual-spatial ability factor loading predominantly on the second factor. The SCREEN tasks, except for visual-spatial ability, loaded substantially on the factors (>0.5), suggesting that the SCREEN possesses good convergent validity for assessing the risk of MCI.

Criterion Validity

The coding of SCREEN scores into a binary *healthy* and *unhealthy* outcome standardized the dependent variable to allow for criterion testing. Criterion validity was assessed using cross-tabulations and the analysis of confusion matrices and provided insights into the sensitivity and specificity of the SCREEN when compared to the Qmci. Of the 144 cases considered, 20 (13.9%) were true negatives, and 74 (51.4%) were true positives. The SCREEN's sensitivity, which reflects its capacity to accurately identify healthy individuals (true positives), was 63.25% (74 correct identifications/117 actual positives). The specificity of the test, indicating its ability to accurately identify unhealthy individuals (true negatives), was 74.07% (20 correct identifications/27 actual negatives). Then, sensitivity and specificity were derived using discrepant analysis and a resolver test previously described (whether the HCP referred the participant to a physician following the screens). The results were identical, the estimate of the SCREEN sensitivity was 63.3% (74/117), and the estimate of the specificity was 74% (20/27).

Reliability Testing

Internal Reliability

A Cronbach α=0.70 is acceptable, and at least 0.90 is required for clinical instruments [79]. The estimate of internal consistency for the SCREEN (N=147) was Cronbach α=0.63.

Test-Retest Reliability

Test-retest reliability analyses were conducted using ICC for the SCREEN activity scores and the κ coefficient for the healthy and unhealthy classifications. Guidelines for interpretation of the ICC suggest that anything <0.5 indicates poor reliability and anything between 0.5 and 0.75 suggests moderate reliability [80]; the ICC for the SCREEN activity scores was 0.54. With respect to the κ coefficient, a κ value of <0.2 is considered to have no level of agreement, a κ value of 0.21 to 0.39 is considered minimal, a κ value of 0.4 to 0.59 is considered weak agreement, and anything >0.8 suggests strong to almost perfect agreement [81]. The κ coefficient for healthy and unhealthy classifications was 0.15.

XSL•FO
RenderX

Analysis of the Factors Impacting Healthy and Unhealthy Results

The Spearman rank correlation was used to assess the relationships between participants' overall activity score on the SCREEN and their total time to complete the SCREEN; age, sex, and self-reported levels of education; technology use; medication use; amount of sleep; and level of anxiety (as measured using the GAS-10) at the time of SCREEN administration. Lower overall activity scores were moderately correlated with being older (r_{s142}=–0.57; *P*<.001) and increased total time to complete the SCREEN (r_{s142}=0.49; *P*<.001). There was also a moderate inverse relationship between overall activity score and total time to compete the SCREEN (r_{s142}=–0.67; *P*<.001) whereby better performance was associated with quicker task completion. There were weak positive associations between overall activity score and increased technology use (r_{s142}=0.34; *P*<.001) and higher level of education (r_{s142}=0.21; *P*=.01).

A logistic regression model was used to predict the SCREEN result using data from 144 observations. The model's predictors explain approximately 21.33% of the variance in the outcome variable. The likelihood ratio test indicates that the model provides a significantly better fit to the data than a model without predictors (*P*<.001).

The SCREEN outcome variable (*healthy* vs *unhealthy*) was associated with the predictor variables sex and total time to complete the SCREEN. More specifically, female participants were more likely to obtain *healthy* SCREEN outcomes (*P*=.007; 95% CI 0.32-2.05). For all participants, the longer it took to complete the SCREEN, the less likely they were to achieve a *healthy* SCREEN outcome (*P*=.002; 95% CI –0.33 to –0.07). Age (*P*=.25; 95% CI –0.09 to 0.02), medication use (*P*=.96; 95% CI –0.9 to 0.94), technology use (*P*=.44; 95% CI –0.28 to 0.65), level of education (*P*=.14; 95% CI –0.09 to 0.64), level of anxiety (*P*=.26; 95% CI –1.13 to 0.3), and hours of sleep (*P*=.08; 95% CI –0.06 to 0.93) were not significant.

Impact of Practice Effects

The SCREEN was administered approximately 3 months apart, and separate, paired-sample *t* tests were performed to compare SCREEN outcomes between visits 1 and 2 (76/147, 51.7%; Table 4), visits 2 and 3 (22/147, 15%), and visits 3 and 4 (13/147, 8.8%). Declining visits were partially attributable to the early shutdown of data collection due to the COVID-19 pandemic, and therefore, comparisons between visits 2 and 3 or visits 3 and 4 were not reported. Compared to participants' SCREEN performance on visit 1, their overall mean activity score and overall processing time improved on their second administration of the SCREEN (score: t_{75}=–2.86 and *P*=.005; processing time: t_{75}=–2.98 and *P*=.004). Even though the 7 task-specific activity scores on the SCREEN also increased between visits 1 and 2, these improvements were not significant, indicating that the difference in overall activity scores was cumulative and not attributable to a specific task (Table 4).

Table 4. Change in participants' SCREEN scores between visits 1 and 2.

SCREEN task	Visit 1 (n=76), mean (SD)	Visit 2 (n=76), mean (SD)	*t* test (*df*)	*P* value
Abstract reasoning	73.51 (1.61)	75.41 (1.87)	–1.04 (75)	.30
Constructive ability	44.15 (2.79)	46.76 (2.58)	–0.91 (75)	.37
Visual-spatial ability	82.22 (2.04)	85.70 (2.0)	–1.19 (75)	.24
Numerical problem-solving	77.82 (1.64)	80.12 (1.71)	–1.27 (75)	.21
Route finding	57.75 (2.64)	56.52 (2.58)	–0.39 (75)	.70
Prioritizing	57.11 (2.1)	55.57 (2.48)	0.55 (75)	.59
Divided attention	56.44 (2.15)	60.04 (2.28)	–1.77 (75)	.08

Discussion

Principal Findings

Our study aimed to evaluate the effectiveness and reliability of the BrainFx SCREEN in detecting MCI in primary care settings. The research took place during the COVID-19 pandemic, which influenced the study's execution and timeline. Despite these challenges, the findings offer valuable insights into cognitive impairment screening.

Brief MCI screening tools help time-strapped primary care physicians determine whether referral for a definitive battery of more time-consuming and expensive tests is warranted. These tools must optimize and balance the need for time efficiency while also being psychometrically valid and easily administered [82]. The importance of brevity is determined by a number of factors, including the clinical setting. Screens that can be completed in approximately ≤5 minutes [13] are recommended for faster-paced clinical settings (eg, emergency rooms and preoperative screens), whereas those that can be completed in 5 to 10 minutes or less are better suited to primary care settings [82-84]. Identifying affordable, psychometrically tested screening tests for MCI that integrate into clinical workflows and are easy to consistently administer and complete may help with the following:

1. Initiating appropriate diagnostic tests for signs and symptoms at an earlier stage
2. Normalizing and destigmatizing cognitive testing for older adults
3. Expediting referrals
4. Allowing for timely access to programs and services that can support aging in place or delay institutionalization
5. Reducing risk

XSL•FO
RenderX

6. Improving the psychosocial well-being of patients and their care partners by increasing access to information and resources that aid with future planning and decision-making [85,86]

Various cognitive tests are commonly used for detecting MCI. These include the Addenbrook Cognitive Examination–Revised, Consortium to Establish a Registry for Alzheimer's Disease, Sunderland Clock Drawing Test, Informant Questionnaire on Cognitive Decline in the Elderly, Memory Alternation Test, MMSE, MoCA 8.1, and Qmci [61,87]. The Addenbrook Cognitive Examination–Revised, Consortium to Establish a Registry for Alzheimer's Disease, MoCA 8.1, Qmci, and Memory Alternation Test are reported to have similar diagnostic accuracy [61,88]. The HCPs participating in this study reported using the MoCA 8.1 as their primary screening tool for MCI along with other assessments such as the MMSE and Trail Making Test parts A and B.

Recent research highlights the growing use of digital tools [51,89,90], mobile technology [91,92], virtual reality [93,94], and artificial intelligence [95] to improve early identification of MCI. Demeyere et al [51] developed the tablet-based, 10-item Oxford Cognitive Screen–Plus to detect slight changes in cognitive impairment across 5 domains of cognition (memory, attention, number, praxis, and language), which has been validated among neurologically healthy older adults. Statsenko et al [96] have explored improvement of the predictive capabilities of tests using artificial intelligence. Similarly, there is an emerging focus on the use of machine learning techniques to detect dementia leveraging routinely collected clinical data [97,98]. This progression signifies a shift toward more technologically advanced, efficient, and potentially more accurate diagnostic approaches in the detection of MCI.

Whatever the modality, screening tools should be quick to administer, demonstrate consistent results over time and between different evaluators, cover all major cognitive areas, and be straightforward to both administer and interpret [99]. However, highly sensitive tests such as those suggested for screening carry a significant risk of false-positive diagnoses [15]. Given the high potential for harm of false positives, it is important to validate the psychometric properties of screening tests across different populations and understand how factors such as age and education can influence the results [99].

Our study did not assess the face validity of the SCREEN, but participating occupational therapists were comfortable with the test regimen. Nonetheless, the research team noted the absence of verbal fluency and memory tests in the SCREEN, both of which McDonnell et al [100] identified as being more sensitive to the more commonly seen amnesic MCI. Two of the most sensitive tools for MCI screening, the MoCA 8.1 [76] and Qmci [25], assess memory, language, and verbal skills, and tests of verbal fluency and logical memory have been shown to be particularly sensitive to early cognitive changes [77,78].

The constructs included in the SCREEN (Table 1) were selected based on a single non–peer-reviewed study [58] using the 360 and traumatic brain injury data (N=188) that identified the constructs as predictive of brain injury. The absence of tasks that measure verbal fluency or logical memory in the SCREEN appears to weaken claims of construct validity. The principal component analysis of the SCREEN assessment identified 2 components accounting for 52.12% of the total variance. The first component was strongly associated with abstract reasoning, constructive ability, and divided attention, whereas the second was primarily influenced by visual-spatial abilities. This indicates that constructs related to perception, attention, and memory are central to the SCREEN scores.

The SCREEN's binary outcome (healthy or unhealthy) created by the research team was based on comparisons with the Qmci. However, the method of identifying *areas of challenge* in the SCREEN by comparing the individual's mean score on each of the 7 tasks with the mean scores of a global or filtered cohort in the LBB introduces potential biases or errors. These could arise from a surge in additions to the LBB from patients with specific characteristics, self-selection of participants, poorly trained SCREEN administrators, inclusion of nonstandard test results, underuse of appropriate filters, and underreporting of clinical conditions or factors such as socioeconomic status that impact performance in standardized cognitive tests.

The proprietary method of analyzing and reporting SCREEN results complicates traditional sensitivity and specificity measurement. Our testing indicated a sensitivity of 63.25% and specificity of 74.07% for identifying healthy (those without MCI) and unhealthy (those with MCI) individuals. The SCREEN's Cronbach α=.63, slightly below the threshold for clinical instruments, and reliability scores that were lower than the ideal standards suggest a higher-than-acceptable level of random measurement error in its constructs. The lower reliability may also stem from an inadequate sample size or a limited number of scale items.

The SCREEN's results are less favorable compared to those of other digital MCI screening tools that similarly enable evaluation of specific cognitive domains but also provide validated, norm-referenced cutoff scores and methods for cumulative scoring in clinical settings (Oxford Cognitive Screen–Plus) [51] or of validated MCI screening tools used in primary care (eg, MoCA 8.1, Qmci, and MMSE) [51,87]. The SCREEN's unique scoring algorithm and the dynamic denominator in data analysis necessitate caution in comparing these results to those of other tools with fixed scoring algorithms and known sensitivities [101,102]. We found the SCREEN to have lower-than-expected internal reliability, suggesting significant random measurement error. Test-retest reliability was weak for the healthy or unhealthy outcome but stronger for overall activity scores between tests. The variability in identifying areas of challenge could relate to technological difficulties or variability from comparisons with a growing database of test results.

Potential reasons for older adults' poorer scores on timed tests include the impact of sensorimotor decline on touch screen sensation and reaction time [38,103], anxiety related to taking a computer-enabled test [104-106], or the anticipated consequences of a negative outcome [107]. However, these effects were unlikely to have influenced the results of this study. Practice effects were observed [29,108], but the SCREEN's novelty suggests that familiarity is not gained through prepreparation or word of mouth as this sample was self-selected

XSL•FO
RenderX

and not randomized. Future research might also explore the impact of digital literacy and cultural differences in the interpretation of software constructs or icons on MCI screening in a randomized, older adult sample.

Limitations

This study had methodological limitations that warrant attention. The small sample size and the demographic distribution of the 147 participants aged ≥55 years, with most (96/147, 65.3%) being female and well educated, limits the generalizability of the findings to different populations. The study's design, aiming to explore the sensitivity of the SCREEN for early detection of MCI, necessitated the exclusion of individuals with a previous diagnosis of MCI or dementia. This exclusion criterion might have impacted the study's ability to thoroughly assess the SCREEN's effectiveness in a more varied clinical context. The requirement for participants to read and comprehend English introduced another limitation to our study. This criterion potentially limited the SCREEN tool's applicability across diverse linguistic backgrounds as individuals with language-based impairments or those not proficient in English may face challenges in completing the assessment [51]. Such limitations could impact the generalizability of our findings to non–English-speaking populations or to those with language impairments, underscoring the need for further research to evaluate the SCREEN tool's effectiveness in broader clinical and linguistic contexts.

Financial constraints played a role in limiting the study's scope. Due to funding limitations, it was not possible to include specialist assessments and a battery of neuropsychiatric tests generally considered the *gold standard* to confirm or rule out an MCI diagnosis. Therefore, the study relied on differential verification through 2 imperfect reference standards: a comparison with the Qmci (the tool with the highest published sensitivity to MCI in 2019, when the study was designed) and the clinical judgment of the administering HCP, particularly in decisions regarding referrals for further clinical assessment. Furthermore, while an economic feasibility assessment was considered, the research team determined that it should follow, not precede, an evaluation of the SCREEN's validity and reliability.

The proprietary nature of the algorithm used for scoring the SCREEN posed another challenge. Without access to this algorithm, the research team had to use a novel comparative statistical approach, coding patient results into a binary variable: healthy (SCREEN=no areas of challenge OR Qmci≥67 out of 100) or unhealthy (SCREEN=one or more areas of challenge OR Qmci<67 out of 100). This may have introduced a higher level of error into our statistical analysis. Furthermore, the process for determining areas of challenge on the SCREEN involves comparing a participant's result to the existing SCREEN results in the LBB at the time of testing. By the end of this study, the LBB contained 632 SCREEN results for adults aged ≥55 years, with this study contributing 258 of those results. The remaining 366 original SCREEN results, 64% of which were completed by individuals who self-identified as having a preexisting diagnosis or conditions associated with cognitive impairment (eg, traumatic brain injury, concussion, or stroke), could have led to an overestimation of the means and SDs of the study participants' results at the outset of the study.

Unlike other cognitive screening tools, the SCREEN allows for filtering of results to compare different patient cohorts in the LBB using criteria such as age and education. However, at this stage of the LBB's development, using such filters can significantly reduce the reliability of the results due to a smaller comparator population (ie, the denominator used to calculate the mean and SD). This, in turn, affects the significance of the results. Moreover, the constantly changing LBB data set makes it challenging to meaningfully compare an individual's results over time as the evolving denominator affects the accuracy and relevance of these comparisons. Finally, the significant improvement in SCREEN scores between the first and second visits suggests the presence of practice effects, which could have influenced the reliability and validity of the findings.

Conclusions

In a primary care setting, where MCI screening tools are essential and recommended for those with concerns [85], certain criteria are paramount: time efficiency, ease of administration, and robust psychometric properties [82]. Our analysis of the BrainFx SCREEN suggests that, despite its innovative approach and digital delivery, it currently falls short in meeting these criteria. The SCREEN's comparatively longer administration time and lower-than-expected reliability scores suggest that it may not be the most effective tool for MCI screening of older adults in a primary care setting at this time.

It is important to note that, in the wake of the COVID-19 pandemic, and with an aging population living and aging by design or necessity in a community setting, there is growing interest in digital solutions, including web-based applications and platforms to both collect digital biomarkers and deliver cognitive training and other interventions [109,110]. However, new normative standards are required when adapting cognitive tests to digital formats [92] as the change in medium can significantly impact test performance and results interpretation. Therefore, we recommend caution when interpreting our study results and encourage continued research and refinement of tools such as the SCREEN. This ongoing process will ensure that current and future MCI screening tools are effective, reliable, and relevant in meeting the needs of our aging population, particularly in primary care settings where early detection and intervention are key.

Acknowledgments

The researchers gratefully acknowledge the Ontario Centres of Excellence Health Technologies Fund for their financial support of this study; the executive directors and clinical leads in each of the Family Health Team study locations; the participants and

XSL•FO
RenderX

their friends and families who took part in the study; and research assistants Sharmin Sharker, Kelly Zhu, and Muhammad Umair for their contributions to data management and statistical analysis.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JM contributed to the conceptualization, methodology, validation, formal analysis, data curation, writing—original draft, writing—review and editing, visualization, supervision, and funding acquisition. AML contributed to the conceptualization, methodology, validation, investigation, formal analysis, data curation, writing—original draft, writing—review and editing, visualization, and project administration. WP contributed to the validation, formal analysis, data curation, writing—original draft, writing—review and editing, and visualization. Finally, PH contributed to conceptualization, methodology, writing—review and editing, supervision, and funding acquisition.

Conflicts of Interest

None declared.

References

1. Casagrande M, Marselli G, Agostini F, Forte G, Favieri F, Guarino A. The complex burden of determining prevalence rates of mild cognitive impairment: a systematic review. Front Psychiatry 2022;13:960648 [FREE Full text] [doi: 10.3389/fpsyt.2022.960648] [Medline: 36213927]
2. Petersen RC, Caracciolo B, Brayne C, Gauthier S, Jelic V, Fratiglioni L. Mild cognitive impairment: a concept in evolution. J Intern Med 2014 Mar;275(3):214-228 [FREE Full text] [doi: 10.1111/joim.12190] [Medline: 24605806]
3. Knopman DS, Petersen RC. Mild cognitive impairment and mild dementia: a clinical perspective. Mayo Clin Proc 2014 Oct;89(10):1452-1459 [FREE Full text] [doi: 10.1016/j.mayocp.2014.06.019] [Medline: 25282431]
4. Anderson ND. State of the science on mild cognitive impairment (MCI). CNS Spectr 2019 Feb;24(1):78-87. [doi: 10.1017/S1092852918001347] [Medline: 30651152]
5. Tangalos EG, Petersen RC. Mild cognitive impairment in geriatrics. Clin Geriatr Med 2018 Nov;34(4):563-589. [doi: 10.1016/j.cger.2018.06.005] [Medline: 30336988]
6. Ng R, Maxwell C, Yates E, Nylen K, Antflick J, Jette N, et al. Brain disorders in Ontario: prevalence, incidence and costs from health administrative data. Institute for Clinical Evaluative Sciences. 2015. URL: https://www.ices.on.ca/publications/research-reports/brain-disorders-in-ontario-prevalence-incidence-and-costs-from-health-administrative-data/ [accessed 2024-04-01]
7. Centers for Disease ControlPrevention (CDC). Self-reported increased confusion or memory loss and associated functional difficulties among adults aged ≥ 60 years - 21 states, 2011. MMWR Morb Mortal Wkly Rep 2013 May 10;62(18):347-350 [FREE Full text] [Medline: 23657108]
8. Petersen RC, Lopez O, Armstrong MJ, Getchius TS, Ganguli M, Gloss D, et al. Practice guideline update summary: mild cognitive impairment: report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. Neurology 2018 Jan 16;90(3):126-135 [FREE Full text] [doi: 10.1212/WNL.0000000000004826] [Medline: 29282327]
9. Song WX, Wu WW, Zhao YY, Xu HL, Chen GC, Jin SY, et al. Evidence from a meta-analysis and systematic review reveals the global prevalence of mild cognitive impairment. Front Aging Neurosci 2023;15:1227112 [FREE Full text] [doi: 10.3389/fnagi.2023.1227112] [Medline: 37965493]
10. Chen Y, Denny KG, Harvey D, Farias ST, Mungas D, DeCarli C, et al. Progression from normal cognition to mild cognitive impairment in a diverse clinic-based and community-based elderly cohort. Alzheimers Dement 2017 Apr;13(4):399-405 [FREE Full text] [doi: 10.1016/j.jalz.2016.07.151] [Medline: 27590706]
11. Langa KM, Levine DA. The diagnosis and management of mild cognitive impairment: a clinical review. JAMA 2014 Dec 17;312(23):2551-2561 [FREE Full text] [doi: 10.1001/jama.2014.13806] [Medline: 25514304]
12. Zhang Y, Natale G, Clouston S. Incidence of mild cognitive impairment, conversion to probable dementia, and mortality. Am J Alzheimers Dis Other Demen 2021;36:15333175211012235 [FREE Full text] [doi: 10.1177/15333175211012235] [Medline: 34032119]
13. Prince M, Bryce R, Ferri CP. World Alzheimer report 2011: the benefits of early diagnosis and intervention. Alzheimer's Disease International. 2011. URL: https://www.alzint.org/u/WorldAlzheimerReport2011.pdf [accessed 2024-04-01]
14. Patnode CD, Perdue LA, Rossom RC, Rushkin MC, Redmond N, Thomas RG, et al. Screening for cognitive impairment in older adults: updated evidence report and systematic review for the US preventive services task force. JAMA 2020 Feb 25;323(8):764-785 [FREE Full text] [doi: 10.1001/jama.2019.22258] [Medline: 32096857]

XSL•FO
RenderX

15. Canadian Task Force on Preventive Health Care, Pottie K, Rahal R, Jaramillo A, Birtwhistle R, Thombs BD, et al. Recommendations on screening for cognitive impairment in older adults. CMAJ 2016 Jan 05;188(1):37-46 [FREE Full text] [doi: 10.1503/cmaj.141165] [Medline: 26622001]
16. Tahami Monfared AA, Phan NT, Pearson I, Mauskopf J, Cho M, Zhang Q, et al. A systematic review of clinical practice guidelines for Alzheimer's disease and strategies for future advancements. Neurol Ther 2023 Aug;12(4):1257-1284 [FREE Full text] [doi: 10.1007/s40120-023-00504-6] [Medline: 37261607]
17. Mattke S, Jun H, Chen E, Liu Y, Becker A, Wallick C. Expected and diagnosed rates of mild cognitive impairment and dementia in the U.S. medicare population: observational analysis. Alzheimers Res Ther 2023 Jul 22;15(1):128 [FREE Full text] [doi: 10.1186/s13195-023-01272-z] [Medline: 37481563]
18. Manly JJ, Tang MX, Schupf N, Stern Y, Vonsattel JP, Mayeux R. Frequency and course of mild cognitive impairment in a multiethnic community. Ann Neurol 2008 Apr;63(4):494-506 [FREE Full text] [doi: 10.1002/ana.21326] [Medline: 18300306]
19. Black CM, Ambegaonkar BM, Pike J, Jones E, Husbands J, Khandker RK. The diagnostic pathway from cognitive impairment to dementia in Japan: quantification using real-world data. Alzheimer Dis Assoc Disord 2019;33(4):346-353. [doi: 10.1097/WAD.0000000000000322] [Medline: 31135457]
20. Ritchie CW, Black CM, Khandker RK, Wood R, Jones E, Hu X, et al. Quantifying the diagnostic pathway for patients with cognitive impairment: real-world data from seven European and north American countries. J Alzheimers Dis 2018;62(1):457-466. [doi: 10.3233/JAD-170864] [Medline: 29439347]
21. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
22. Tsoi KK, Chan JY, Hirai HW, Wong SY, Kwok TC. Cognitive tests to detect dementia: a systematic review and meta-analysis. JAMA Intern Med 2015 Sep;175(9):1450-1458. [doi: 10.1001/jamainternmed.2015.2152] [Medline: 26052687]
23. Lopez MN, Charter RA, Mostafavi B, Nibut LP, Smith WE. Psychometric properties of the Folstein mini-mental state examination. Assessment 2005 Jun;12(2):137-144. [doi: 10.1177/1073191105275412] [Medline: 15914716]
24. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
25. O'Caoimh R, Timmons S, Molloy DW. Screening for mild cognitive impairment: comparison of "MCI specific" screening instruments. J Alzheimers Dis 2016;51(2):619-629 [FREE Full text] [doi: 10.3233/JAD-150881] [Medline: 26890758]
26. Trzepacz PT, Hochstetler H, Wang S, Walker B, Saykin AJ, Alzheimer’s Disease Neuroimaging Initiative. Relationship between the Montreal cognitive assessment and mini-mental state examination for assessment of mild cognitive impairment in older adults. BMC Geriatr 2015 Sep 07;15:107 [FREE Full text] [doi: 10.1186/s12877-015-0103-3] [Medline: 26346644]
27. Nasreddine ZS, Phillips N, Chertkow H. Normative data for the Montreal Cognitive Assessment (MoCA) in a population-based sample. Neurology 2012 Mar 06;78(10):765-766. [doi: 10.1212/01.wnl.0000413072.54070.a3] [Medline: 22391608]
28. Monroe T, Carter M. Using the Folstein Mini Mental State Exam (MMSE) to explore methodological issues in cognitive aging research. Eur J Ageing 2012 Sep;9(3):265-274 [FREE Full text] [doi: 10.1007/s10433-012-0234-8] [Medline: 28804426]
29. Damian AM, Jacobson SA, Hentz JG, Belden CM, Shill HA, Sabbagh MN, et al. The Montreal cognitive assessment and the mini-mental state examination as screening instruments for cognitive impairment: item analyses and threshold scores. Dement Geriatr Cogn Disord 2011;31(2):126-131. [doi: 10.1159/000323867] [Medline: 21282950]
30. Kaufer DI, Williams CS, Braaten AJ, Gill K, Zimmerman S, Sloane PD. Cognitive screening for dementia and mild cognitive impairment in assisted living: comparison of 3 tests. J Am Med Dir Assoc 2008 Oct;9(8):586-593. [doi: 10.1016/j.jamda.2008.05.006] [Medline: 19083293]
31. Gagnon C, Saillant K, Olmand M, Gayda M, Nigam A, Bouabdallaoui N, et al. Performances on the Montreal cognitive assessment along the cardiovascular disease continuum. Arch Clin Neuropsychol 2022 Jan 17;37(1):117-124 [FREE Full text] [doi: 10.1093/arclin/acab029] [Medline: 33960374]
32. Cooley SA, Heaps JM, Bolzenius JD, Salminen LE, Baker LM, Scott SE, et al. Longitudinal change in performance on the Montreal cognitive assessment in older adults. Clin Neuropsychol 2015;29(6):824-835 [FREE Full text] [doi: 10.1080/13854046.2015.1087596] [Medline: 26373627]
33. O'Caoimh R, Gao Y, McGlade C, Healy L, Gallagher P, Timmons S, et al. Comparison of the quick mild cognitive impairment (Qmci) screen and the SMMSE in screening for mild cognitive impairment. Age Ageing 2012 Sep;41(5):624-629 [FREE Full text] [doi: 10.1093/ageing/afs059] [Medline: 22610464]
34. O'Caoimh R, Molloy DW. Comparing the diagnostic accuracy of two cognitive screening instruments in different dementia subtypes and clinical depression. Diagnostics (Basel) 2019 Aug 08;9(3):93 [FREE Full text] [doi: 10.3390/diagnostics9030093] [Medline: 31398811]
35. Clarnette R, O'Caoimh R, Antony DN, Svendrovski A, Molloy DW. Comparison of the Quick Mild Cognitive Impairment (Qmci) screen to the Montreal Cognitive Assessment (MoCA) in an Australian geriatrics clinic. Int J Geriatr Psychiatry 2017 Jun;32(6):643-649. [doi: 10.1002/gps.4505] [Medline: 27427212]

XSL•FO
RenderX

36. Glynn K, Coen R, Lawlor BA. Is the Quick Mild Cognitive Impairment screen (QMCI) more accurate at detecting mild cognitive impairment than existing short cognitive screening tests? A systematic review of the current literature. Int J Geriatr Psychiatry 2019 Dec;34(12):1739-1746. [doi: 10.1002/gps.5201] [Medline: 31418473]
37. Lee MT, Chang WY, Jang Y. Psychometric and diagnostic properties of the Taiwan version of the quick mild cognitive impairment screen. PLoS One 2018;13(12):e0207851 [FREE Full text] [doi: 10.1371/journal.pone.0207851] [Medline: 30507928]
38. Wallace SE, Donoso Brown EV, Simpson RC, D'Acunto K, Kranjec A, Rodgers M, et al. A comparison of electronic and paper versions of the Montreal cognitive assessment. Alzheimer Dis Assoc Disord 2019;33(3):272-278. [doi: 10.1097/WAD.0000000000000333] [Medline: 31335458]
39. Gagnon C, Olmand M, Dupuy EG, Besnier F, Vincent T, Grégoire CA, et al. Videoconference version of the Montreal cognitive assessment: normative data for Quebec-French people aged 50 years and older. Aging Clin Exp Res 2022 Jul;34(7):1627-1633 [FREE Full text] [doi: 10.1007/s40520-022-02092-1] [Medline: 35178685]
40. Friemel TN. The digital divide has grown old: determinants of a digital divide among seniors. New Media & Society 2014 Jun 12;18(2):313-331. [doi: 10.1177/1461444814538648]
41. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
42. Searles C, Farnsworth JL, Jubenville C, Kang M, Ragan B. Test–retest reliability of the BrainFx 360® performance assessment. Athl Train Sports Health Care 2019 Jul;11(4):183-191. [doi: 10.3928/01913913-20181005-01]
43. Jones C, Miguel-Cruz A, Brémault-Phillips S. Technology acceptance and usability of the BrainFx SCREEN in Canadian military members and veterans with posttraumatic stress disorder and mild traumatic brain injury: mixed methods UTAUT study. JMIR Rehabil Assist Technol 2021 May 13;8(2):e26078 [FREE Full text] [doi: 10.2196/26078] [Medline: 33983125]
44. McMurray J, Levy A, Holyoke P. Psychometric evaluation and workflow integration study of a tablet-based tool to detect mild cognitive impairment in older adults: protocol for a mixed methods study. JMIR Res Protoc 2021 May 21;10(5):e25520 [FREE Full text] [doi: 10.2196/25520] [Medline: 34018966]
45. Wilansky P, Eklund JM, Milner T, Kreindler D, Cheung A, Kovacs T, et al. Cognitive behavior therapy for anxious and depressed youth: improving homework adherence through mobile technology. JMIR Res Protoc 2016 Nov 10;5(4):e209 [FREE Full text] [doi: 10.2196/resprot.5841] [Medline: 27833071]
46. Ammenwerth E, Iller C, Mahler C. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Med Inform Decis Mak 2006 Jan 09;6:3 [FREE Full text] [doi: 10.1186/1472-6947-6-3] [Medline: 16401336]
47. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Q 1995 Jun;19(2):213-236. [doi: 10.2307/249689]
48. Beuscher L, Grando VT. Challenges in conducting qualitative research with individuals with dementia. Res Gerontol Nurs 2009 Jan;2(1):6-11 [FREE Full text] [doi: 10.3928/19404921-20090101-04] [Medline: 20077988]
49. Howe E. Informed consent, participation in research, and the Alzheimer's patient. Innov Clin Neurosci 2012 May;9(5-6):47-51 [FREE Full text] [Medline: 22808449]
50. Thorogood A, Mäki-Petäjä-Leinonen A, Brodaty H, Dalpé G, Gastmans C, Gauthier S, Global Alliance for GenomicsHealth, AgeingDementia Task Team. Consent recommendations for research and international data sharing involving persons with dementia. Alzheimers Dement 2018 Oct;14(10):1334-1343 [FREE Full text] [doi: 10.1016/j.jalz.2018.05.011] [Medline: 30293575]
51. Demeyere N, Haupt M, Webb SS, Strobel L, Milosevich ET, Moore MJ, et al. Introducing the tablet-based Oxford Cognitive Screen-Plus (OCS-Plus) as an assessment tool for subtle cognitive impairments. Sci Rep 2021 Apr 12;11(1):8000 [FREE Full text] [doi: 10.1038/s41598-021-87287-8] [Medline: 33846501]
52. Nasreddine ZS, Patel BB. Validation of Montreal cognitive assessment, MoCA, alternate French versions. Can J Neurol Sci 2016 Sep;43(5):665-671. [doi: 10.1017/cjn.2016.273] [Medline: 27670209]
53. Mueller AE, Segal DL, Gavett B, Marty MA, Yochim B, June A, et al. Geriatric anxiety scale: item response theory analysis, differential item functioning, and creation of a ten-item short form (GAS-10). Int Psychogeriatr 2015 Jul;27(7):1099-1111. [doi: 10.1017/S1041610214000210] [Medline: 24576589]
54. Segal DL, June A, Payne M, Coolidge FL, Yochim B. Development and initial validation of a self-report assessment tool for anxiety among older adults: the Geriatric Anxiety Scale. J Anxiety Disord 2010 Oct;24(7):709-714. [doi: 10.1016/j.janxdis.2010.05.002] [Medline: 20558032]
55. Balsamo M, Cataldi F, Carlucci L, Fairfield B. Assessment of anxiety in older adults: a review of self-report measures. Clin Interv Aging 2018;13:573-593 [FREE Full text] [doi: 10.2147/CIA.S114100] [Medline: 29670342]
56. Gatti A, Gottschling J, Brugnera A, Adorni R, Zarbo C, Compare A, et al. An investigation of the psychometric properties of the Geriatric Anxiety Scale (GAS) in an Italian sample of community-dwelling older adults. Aging Ment Health 2018 Sep;22(9):1170-1178. [doi: 10.1080/13607863.2017.1347141] [Medline: 28675312]
57. Yochim BP, Mueller AE, June A, Segal DL. Psychometric properties of the Geriatric Anxiety Scale: comparison to the beck anxiety inventory and geriatric anxiety inventory. Clin Gerontol 2010 Dec 06;34(1):21-33. [doi: 10.1080/07317115.2011.524600]

XSL•FO
RenderX

58. Recent concussion (< 6 months ago) analysis result. Daisy Intelligence. 2016. URL: https://www.daisyintelligence.com/retail-solutions/ [accessed 2024-04-01]
59. Malloy DW, O'Caoimh R. The Quick Guide: Scoring and Administration Instructions for The Quick Mild Cognitive Impairment (Qmci) Screen. Waterford, Ireland: Newgrange Press; 2017.
60. O'Caoimh R, Gao Y, Svendovski A, Gallagher P, Eustace J, Molloy DW. Comparing approaches to optimize cut-off scores for short cognitive screening instruments in mild cognitive impairment and dementia. J Alzheimers Dis 2017;57(1):123-133 [FREE Full text] [doi: 10.3233/JAD-161204] [Medline: 28222528]
61. Breton A, Casey D, Arnaoutoglou NA. Cognitive tests for the detection of mild cognitive impairment (MCI), the prodromal stage of dementia: meta-analysis of diagnostic accuracy studies. Int J Geriatr Psychiatry 2019 Feb;34(2):233-242. [doi: 10.1002/gps.5016] [Medline: 30370616]
62. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. PLoS One 2019;14(10):e0223832 [FREE Full text] [doi: 10.1371/journal.pone.0223832] [Medline: 31603953]
63. Espinosa A, Alegret M, Boada M, Vinyes G, Valero S, Martínez-Lage P, et al. Ecological assessment of executive functions in mild cognitive impairment and mild Alzheimer's disease. J Int Neuropsychol Soc 2009 Sep;15(5):751-757 [FREE Full text] [doi: 10.1017/S135561770999035X] [Medline: 19570310]
64. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat Med 2001 Jul 15;20(13):1987-2001. [doi: 10.1002/sim.819] [Medline: 11427955]
65. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. Epidemiology 2005 Sep;16(5):604-612. [doi: 10.1097/01.ede.0000173042.07579.17] [Medline: 16135935]
66. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. J Clin Epidemiol 2003 Aug;56(8):730-735. [doi: 10.1016/s0895-4356(03)00084-2] [Medline: 12954464]
67. Paiva CE, Barroso EM, Carneseca EC, de Pádua Souza C, Dos Santos FT, Mendoza López RV, et al. A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. BMC Med Res Methodol 2014 Jan 21;14:8 [FREE Full text] [doi: 10.1186/1471-2288-14-8] [Medline: 24447633]
68. Streiner DL, Kottner J. Recommendations for reporting the results of studies of instrument and scale development and testing. J Adv Nurs 2014 Sep;70(9):1970-1979. [doi: 10.1111/jan.12402] [Medline: 24684713]
69. Streiner DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry 1993 Mar;38(2):140-148. [doi: 10.1177/070674379303800214] [Medline: 8467441]
70. Peyre H, Leplège A, Coste J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. Qual Life Res 2011 Mar;20(2):287-300. [doi: 10.1007/s11136-010-9740-3] [Medline: 20882358]
71. Nevado-Holgado AJ, Kim CH, Winchester L, Gallacher J, Lovestone S. Commonly prescribed drugs associate with cognitive function: a cross-sectional study in UK Biobank. BMJ Open 2016 Nov 30;6(11):e012177 [FREE Full text] [doi: 10.1136/bmjopen-2016-012177] [Medline: 27903560]
72. Moore AR, O'Keeffe ST. Drug-induced cognitive impairment in the elderly. Drugs Aging 1999 Jul;15(1):15-28. [doi: 10.2165/00002512-199915010-00002] [Medline: 10459729]
73. Rogers J, Wiese BS, Rabheru K. The older brain on drugs: substances that may cause cognitive impairment. Geriatr Aging 2008;11(5):284-289 [FREE Full text]
74. Marvanova M. Drug-induced cognitive impairment: effect of cardiovascular agents. Ment Health Clin 2016 Jul;6(4):201-206 [FREE Full text] [doi: 10.9740/mhc.2016.07.201] [Medline: 29955471]
75. Espeland MA, Rapp SR, Manson JE, Goveas JS, Shumaker SA, Hayden KM, WHIMSYWHIMS-ECHO Study Groups. Long-term effects on cognitive trajectories of postmenopausal hormone therapy in two age groups. J Gerontol A Biol Sci Med Sci 2017 Jun 01;72(6):838-845 [FREE Full text] [doi: 10.1093/gerona/glw156] [Medline: 27506836]
76. Luis CA, Keegan AP, Mullan M. Cross validation of the Montreal cognitive assessment in community dwelling older adults residing in the Southeastern US. Int J Geriatr Psychiatry 2009 Feb;24(2):197-201. [doi: 10.1002/gps.2101] [Medline: 18850670]
77. Cunje A, Molloy DW, Standish TI, Lewis DL. Alternate forms of logical memory and verbal fluency tasks for repeated testing in early cognitive changes. Int Psychogeriatr 2007 Feb;19(1):65-75. [doi: 10.1017/S1041610206003425] [Medline: 16684396]
78. Molloy DW, Standish TI, Lewis DL. Screening for mild cognitive impairment: comparing the SMMSE and the ABCS. Can J Psychiatry 2005 Jan;50(1):52-58. [doi: 10.1177/070674370505000110] [Medline: 15754666]
79. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 4th edition. Oxford, UK: Oxford University Press; 2008.
80. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]

XSL•FO
RenderX

81. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
82. Zhuang L, Yang Y, Gao J. Cognitive assessment tools for mild cognitive impairment screening. J Neurol 2021 May;268(5):1615-1622. [doi: 10.1007/s00415-019-09506-7] [Medline: 31414193]
83. Zhang J, Wang L, Deng X, Fei G, Jin L, Pan X, et al. Five-minute cognitive test as a new quick screening of cognitive impairment in the elderly. Aging Dis 2019 Dec;10(6):1258-1269 [FREE Full text] [doi: 10.14336/AD.2019.0115] [Medline: 31788337]
84. Feldman HH, Jacova C, Robillard A, Garcia A, Chow T, Borrie M, et al. Diagnosis and treatment of dementia: 2. Diagnosis. CMAJ 2008 Mar 25;178(7):825-836 [FREE Full text] [doi: 10.1503/cmaj.070798] [Medline: 18362376]
85. Sabbagh MN, Boada M, Borson S, Chilukuri M, Dubois B, Ingram J, et al. Early detection of mild cognitive impairment (MCI) in primary care. J Prev Alzheimers Dis 2020;7(3):165-170. [doi: 10.14283/jpad.2020.21] [Medline: 32463069]
86. Milne A. Dementia screening and early diagnosis: the case for and against. Health Risk Soc 2010 Mar 05;12(1):65-76. [doi: 10.1080/13698570903509497]
87. Screening tools to identify adults with cognitive impairment associated with dementia: diagnostic accuracy. Canadian Agency for Drugs and Technologies in Health. 2014. URL: https://www.cadth.ca/sites/default/files/pdf/htis/nov-2014/RB0752%20Cognitive%20Assessments%20for%20Dementia%20Final.pdf [accessed 2024-04-01]
88. Chehrehnegar N, Nejati V, Shati M, Rashedi V, Lotfi M, Adelirad F, et al. Early detection of cognitive disturbances in mild cognitive impairment: a systematic review of observational studies. Psychogeriatrics 2020 Mar;20(2):212-228. [doi: 10.1111/psyg.12484] [Medline: 31808989]
89. Chan JY, Yau ST, Kwok TC, Tsoi KK. Diagnostic performance of digital cognitive tests for the identification of MCI and dementia: a systematic review. Ageing Res Rev 2021 Dec;72:101506. [doi: 10.1016/j.arr.2021.101506] [Medline: 34744026]
90. Cubillos C, Rienzo A. Digital cognitive assessment tests for older adults: systematic literature review. JMIR Ment Health 2023 Dec 08;10:e47487 [FREE Full text] [doi: 10.2196/47487] [Medline: 38064247]
91. Chen R, Foschini L, Kourtis L, Signorini A, Jankovic F, Pugh M, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019 Presented at: KDD '19; August 4-8, 2019; Anchorage, AK p. 2145 URL: https://dl.acm.org/doi/10.1145/3292500.3330690 [doi: 10.1145/3292500.3330690]
92. Koo BM, Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. Innov Aging 2019 Jan;3(1):igy038 [FREE Full text] [doi: 10.1093/geroni/igy038] [Medline: 30619948]
93. Zygouris S, Ntovas K, Giakoumis D, Votis K, Doumpoulakis S, Segkouli S, et al. A preliminary study on the feasibility of using a virtual reality cognitive training application for remote detection of mild cognitive impairment. J Alzheimers Dis 2017;56(2):619-627. [doi: 10.3233/JAD-160518] [Medline: 28035922]
94. Liu Q, Song H, Yan M, Ding Y, Wang Y, Chen L, et al. Virtual reality technology in the detection of mild cognitive impairment: a systematic review and meta-analysis. Ageing Res Rev 2023 Jun;87:101889. [doi: 10.1016/j.arr.2023.101889] [Medline: 36806377]
95. Fayemiwo MA, Olowookere TA, Olaniyan OO, Ojewumi TO, Oyetade IS, Freeman S, et al. Immediate word recall in cognitive assessment can predict dementia using machine learning techniques. Alzheimers Res Ther 2023 Jun 15;15(1):111 [FREE Full text] [doi: 10.1186/s13195-023-01250-5] [Medline: 37322550]
96. Statsenko Y, Meribout S, Habuza T, Almansoori TM, van Gorkom KN, Gelovani JG, et al. Patterns of structure-function association in normal aging and in Alzheimer's disease: screening for mild cognitive impairment and dementia with ML regression and classification models. Front Aging Neurosci 2022;14:943566 [FREE Full text] [doi: 10.3389/fnagi.2022.943566] [Medline: 36910862]
97. Roebuck-Spencer TM, Glen T, Puente AE, Denney RL, Ruff RM, Hostetter G, et al. Cognitive screening tests versus comprehensive neuropsychological test batteries: a national academy of neuropsychology education paper†. Arch Clin Neuropsychol 2017 Jun 01;32(4):491-498. [doi: 10.1093/arclin/acx021] [Medline: 28334244]
98. Jammeh EA, Carroll CB, Pearson SW, Escudero J, Anastasiou A, Zhao P, et al. Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. BJGP Open 2018 Jul;2(2):bjgpopen18X101589 [FREE Full text] [doi: 10.3399/bjgpopen18X101589] [Medline: 30564722]
99. Riello M, Rusconi E, Treccani B. The role of brief global cognitive tests and neuropsychological expertise in the detection and differential diagnosis of dementia. Front Aging Neurosci 2021;13:648310 [FREE Full text] [doi: 10.3389/fnagi.2021.648310] [Medline: 34177551]
100. McDonnell M, Dill L, Panos S, Amano S, Brown W, Giurgius S, et al. Verbal fluency as a screening tool for mild cognitive impairment. Int Psychogeriatr 2020 Sep;32(9):1055-1062 [FREE Full text] [doi: 10.1017/S1041610219000644] [Medline: 31258101]
101. Wojtowicz A, Larner AJ. Diagnostic test accuracy of cognitive screeners in older people. Prog Neurol Psychiatry 2017 Mar 20;21(1):17-21. [doi: 10.1002/pnp.457]
102. Larner AJ. Cognitive screening instruments for the diagnosis of mild cognitive impairment. Prog Neurol Psychiatry 2016 Apr 07;20(2):21-26. [doi: 10.1002/PNP.421]

XSL•FO
RenderX

103. Heintz BD, Keenan KG. Spiral tracing on a touchscreen is influenced by age, hand, implement, and friction. PLoS One 2018;13(2):e0191309 [FREE Full text] [doi: 10.1371/journal.pone.0191309] [Medline: 29389940]
104. Laguna K, Babcock RL. Computer anxiety in young and older adults: implications for human-computer interactions in older populations. Comput Human Behav 1997 Aug;13(3):317-326. [doi: 10.1016/s0747-5632(97)00012-5]
105. Wild KV, Mattek NC, Maxwell SA, Dodge HH, Jimison HB, Kaye JA. Computer-related self-efficacy and anxiety in older adults with and without mild cognitive impairment. Alzheimers Dement 2012 Nov;8(6):544-552 [FREE Full text] [doi: 10.1016/j.jalz.2011.12.008] [Medline: 23102124]
106. Wiechmann D, Ryan AM. Reactions to computerized testing in selection contexts. Int J Sel Assess 2003 Jul 30;11(2-3):215-229. [doi: 10.1111/1468-2389.00245]
107. Gass CS, Curiel RE. Test anxiety in relation to measures of cognitive and intellectual functioning. Arch Clin Neuropsychol 2011 Aug;26(5):396-404. [doi: 10.1093/arclin/acr034] [Medline: 21636602]
108. Barbic D, Kim B, Salehmohamed Q, Kemplin K, Carpenter CR, Barbic SP. Diagnostic accuracy of the Ottawa 3DY and short blessed test to detect cognitive dysfunction in geriatric patients presenting to the emergency department. BMJ Open 2018 Mar 16;8(3):e019652 [FREE Full text] [doi: 10.1136/bmjopen-2017-019652] [Medline: 29549205]
109. Owens AP, Ballard C, Beigi M, Kalafatis C, Brooker H, Lavelle G, et al. Implementing remote memory clinics to enhance clinical care during and after COVID-19. Front Psychiatry 2020;11:579934 [FREE Full text] [doi: 10.3389/fpsyt.2020.579934] [Medline: 33061927]
110. Geddes MR, O'Connell ME, Fisk JD, Gauthier S, Camicioli R, Ismail Z, Alzheimer Society of Canada Task Force on Dementia Care Best Practices for COVID‐19. Remote cognitive and behavioral assessment: report of the Alzheimer Society of Canada task force on dementia care best practices for COVID-19. Alzheimers Dement (Amst) 2020;12(1):e12111 [FREE Full text] [doi: 10.1002/dad2.12111] [Medline: 32999916]

Abbreviations

ADL: activities of daily living
FHT: Family Health Team
GAS: Geriatric Anxiety Scale
GAS-10: Geriatric Anxiety Scale–10
HCP: health care provider
ICC: intraclass correlation coefficient
LBB: Living Brain Bank
MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
Qmci: Quick Mild Cognitive Impairment

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 29.01.24; peer-reviewed by J Gao, MJ Moore; comments to author 20.02.24; revised version received 05.03.24; accepted 19.03.24; published 19.04.24.

Please cite as:
McMurray J, Levy A, Pang W, Holyoke P
Psychometric Evaluation of a Tablet-Based Tool to Detect Mild Cognitive Impairment in Older Adults: Mixed Methods Study
J Med Internet Res 2024;26:e56883
URL: https://www.jmir.org/2024/1/e56883
doi:10.2196/56883
PMID:38640480

©Josephine McMurray, AnneMarie Levy, Wei Pang, Paul Holyoke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Food Frequency Questionnaire Assessing Adherence to the Norwegian Food–Based Dietary Guidelines and Other National Lifestyle Recommendations: Instrument Validation Study

Hege Berg Henriksen[1], PhD; Markus Dines Knudsen[1,2,3], PhD; Anette Hjartåker[1], PhD; Rune Blomhoff[1,4], PhD; Monica Hauger Carlsen[1], PhD

[1]Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[2]Section for Colorectal Cancer Screening, Cancer Registry of Norway, Oslo, Norway
[3]Department of Transplantation Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital, Oslo, Norway
[4]Department of Clinical Service, Division of Cancer Medicine, Oslo University Hospital, Oslo, Norway

Corresponding Author:
Hege Berg Henriksen, PhD
Department of Nutrition
Institute of Basic Medical Sciences
University of Oslo
Sognsvannsveien 9
PO Box 1046, Blindern
Oslo, 0372
Norway
Phone: 47 99459673
Email: h.b.henriksen@medisin.uio.no

Abstract

Background: Valid assessment tools are needed when investigating adherence to national dietary and lifestyle guidelines.

Objective: The relative validity of the new digital food frequency questionnaire, the DIGIKOST-FFQ, against 7-day weighed food records and activity sensors was investigated.

Methods: In total, 77 participants were included in the validation study and completed the DIGIKOST-FFQ and the weighed food record, and of these, 56 (73%) also used the activity sensors. The DIGIKOST-FFQ estimates the intake of foods according to the Norwegian food–based dietary guidelines (FBDGs) in addition to lifestyle factors.

Results: At the group level, the DIGIKOST-FFQ showed good validity in estimating intakes according to the Norwegian FBDG. The median differences were small and well below portion sizes for all foods except "water" (median difference 230 g/day). The DIGIKOST-FFQ was able to rank individual intakes for all foods (r=0.2-0.7). However, ranking estimates of vegetable intakes should be interpreted with caution. Between 69% and 88% of the participants were classified into the same or adjacent quartile for foods and between 71% and 82% for different activity intensities. The Bland-Altman plots showed acceptable agreements between DIGIKOST-FFQ and the reference methods. The absolute amount of time in "moderate to vigorous intensity" was underestimated with the DIGIKOST-FFQ. However, estimated time in "moderate to vigorous intensity," "vigorous intensity," and "sedentary time" showed acceptable correlations and good agreement between the methods. The DIGIKOST-FFQ was able to identify adherence to the Norwegian FBDG and physical activity recommendations.

Conclusions: The DIGIKOST-FFQ gave valid estimates of dietary intakes and was able to identify individuals with different degrees of adherence to the Norwegian FBDG and physical activity recommendations. Moderate physical activity was underreported, water was overreported, and vegetables showed poor correlation, which are important to consider when interpreting the data. Good agreement was observed between the methods in estimating dietary intakes and time in "moderate to vigorous physical activity," "sedentary time," and "sleep."

(J Med Internet Res 2024;26:e53442) doi:10.2196/53442

XSL•FO
RenderX

KEYWORDS

digital food frequency questionnaire; lifestyle assessment tool; relative validity; physical activity; Norwegian food–based dietary guidelines; Norway; Norwegian; food; foods; diet; dietary; lifestyle; assessment; digital questionnaire; investigation; food intake; dietary intake; dietary intakes; observation; monitoring; youths; adolescent; adolescents; teen; teens; teenager; teenagers; cross-sectional study

Introduction

Monitoring dietary intake and physical activity with valid and feasible methods in a population is of great importance, as there is a growing body of evidence on the strong association between lifestyle and risk of several chronic diseases [1-5].

Even though self-administered web-based 24-hour dietary recalls and food are increasingly used in epidemiological studies [6,7], the most common type of dietary assessment tool in observational and intervention studies is still the food frequency questionnaire (FFQ) [8,9]. Generally, the FFQs assess the frequency of food intakes and portion size of the food consumed, are typically self-administrated, and vary with regard to which food items or nutrients are included, portion sizes, frequency categories, length of the questionnaire, and duration of the registration period (ie, weeks, months, or years) [6,10-12]. The FFQs may be used to estimate habitual dietary intakes over a longer period of time and to rank individuals according to dietary intakes and may generate data on dietary intakes on a group level. Thus, FFQs are valuable in assessing the degree of adherence to dietary recommendations in a population [8,9,13-15].

As with all dietary assessment methods, FFQs are prone to reporting errors, particularly systematic errors, often related to memory, knowledge of portion sizes, and understanding of the questions [11,12,16,17]. This can result in over- or underestimation of dietary intakes [18,19], which may attenuate the observed association between diet and health outcomes [20,21]. Systematic error may be reduced with clear instruction and the use of relevant examples, such as pictures of portion sizes and informative text.

The emerging use of digital applications has resulted in modifications of traditional printed assessment tools into web-based tools in both research and the clinic [22,23]. Web-based dietary assessment tools have demonstrated good validity and give comparable data with printed versions [6,23-25].

We have developed a short digital semiquantitative food and lifestyle frequency questionnaire (the DIGIKOST-FFQ) based on a validated printed questionnaire called NORDIET-FFQ [26] developed at the Department of Nutrition, University of Oslo, Norway. The DIGIKOST-FFQ is designed to assess adherence to the Norwegian food–based dietary guidelines (FBDGs) and national lifestyle guidelines [3,27] and has been revised according to the results from a qualitative evaluation study (ie, focus group interviews and usability testing) [27].

Generally, the DIGIKOST-FFQ delivers similar data on the main dietary and physical activity recommendations as the NORDIET-FFQ. However, some adjustments have been made based on the results from the validation study of the NORDIET-FFQ (Multimedia Appendix 1 [3,26-38]). In addition, the DIGIKOST-FFQ is based on a digital platform constituting several technical functions to improve the feasibility and the understandability of the questions [27].

Therefore, investigating the relative validity of this new digital assessment tool is of great importance to current and future studies applying the assessment method for dietary data collection. Thus, in this project, we aimed to investigate the relative validity of the DIGIKOST-FFQ in assessing dietary intake, physical activity, and adherence to the Norwegian FBDG in an adult and healthy population in Norway, with weighed food record (WR) and activity sensor as the reference methods.

Methods

Study Design and Participants

Participants were recruited between April and September 2021 and randomly extracted by the National Registry of Norway and by advertisements on Facebook. Adult (18 years and older) men and women living in the Oslo region in Norway were eligible for the study. The study design was cross-sectional. All participants completed the DIGIKOST-FFQ, and after 1-2 months, they filled in a 7-day WR and used an activity sensor (SenseWear Armband Mini [SWA]; BodyMedia). In addition to the WR, the participants received a digital scale and an instruction on how to weigh and record all foods and beverages consumed during a period of 7 consecutive days. Due to the COVID-19 pandemic, the instructions were given in a video meeting (Zoom; Zoom Video Communications) by the responsible researchers in the study. Then, the participants were invited to meet outside the study center at the Department of Nutrition, University of Oslo at a pickup point to receive the WR, a digital scale, SWA, and the prepaid postal envelopes for returning the equipment after use.

During the WR registration period, the participants also wore the SWA to record all physical activity, sedentary time, and sleep. By the end of the registration period, the participants returned the completed WRs and the SWA to the study center by postal mail. At the end of the study, all participants were offered to voluntarily fill out the DIGIKOST-FFQ once more in order to receive an individual feedback report benchmarking their dietary intakes and physical activity against the Norwegian FBDG. In addition, they could be randomly selected to win a gift certificate of NOK 500 (US $56.91). The digital scale was also given to the participants as a gift.

DIGIKOST

The first version of the DIGIKOST-FFQ underwent evaluation in focus group interviews and usability testing [27]. The final version of the DIGIKOST-FFQ is described in Multimedia Appendix 1. In brief, the DIGIKOST is a digital diet and lifestyle assessment tool designed to assess dietary intake and

XSL•FO
RenderX

other lifestyle factors and evaluate these according to the Norwegian FBDG. DIGIKOST also has a report function, the DIGIKOST report, which is an individual feedback report on respondents' adherence to the Norwegian FBDG, with specific and personalized advice on how to fulfill the recommendations. The DIGIKOST-FFQ is based on a software platform called Nettskjema, developed and administered by the University Information Technology Center at the University of Oslo, Norway [28]. The main login option in DIGIKOST is the ID-port (e-ID used by the Norwegian Agency for Public Management and eGovernment [Difi]) [29,30]. The DIGIKOST-FFQ takes approximately 20 minutes to complete [27] and includes 103 food and lifestyle items, of which 78 questions about food items (grams per day), 7 questions about physical activity (minutes per week), sedentary time, and sleep (hours per day), 8 questions about tobacco use, and 10 questions about body weight and demographic data. The main food groups cover foods rich in fiber (ie, fruits and berries, vegetables, and whole grain products), fish, dairy products, meat, oils, margarine, and beverages (Table S1 in Multimedia Appendix 1). The responses from the DIGIKOST-FFQ are directly transferred into a secure server called services for sensitive data, and the crude variables are automatically transformed by unique algorithms into food groups, activities, and lifestyle indices according to the national recommendations [26]. The DIGIKOST data set also contains 2 indices, the Norwegian diet index and the Norwegian lifestyle index [32] (Multimedia Appendix 1). The Norwegian diet index consists of 12 components corresponding to the Norwegian FBDG with a 3-level scoring approach including 3 categories representing low, intermediate, and high adherence, giving a composite diet index ranging in scores from 0 (lowest adherence) to 20 (highest adherence) points. The Norwegian lifestyle index consists of 5 components (ie, diet, physical activity, normal weight, alcoholic drinks, and tobacco use) with a 3-level equal scoring approach, and a composite lifestyle index ranging from 0 to 5 points.

Seven-Day WR

Dietary data were retrieved from the WR and manually coded and imported into the food and nutrient calculation system, KBS (version 7.3, 2018, AE-10 database), developed at the Department of Nutrition, University of Oslo [39].

Physical Activity Sensor

We used the activity sensors called SWA to objectively measure different activity intensities. The SWA was placed on the nondominant arm at the upper part around the triceps muscle. The participants were instructed to use the SWA all day and night and to only remove it during water-based activities like showers or swimming. All activity data generated from the SWA were exported to a computer, and all calculations on activities were conducted in the SenseWear Professional software (version 7.0; BodyMedia Inc).

Sample Size

An adequate number of participants in validation studies are shown to be 80-100 [10,11,40,41]. This ensured statistical power of 80% to detect differences of 1 portion of "fruit" or "vegetable" per day (1 portion=100 g) between test and reference methods, assuming an SD of 1.6 portions (or 160 g) and a significance level of 5% [42,43]. Moreover, in order to detect a Pearson correlation coefficient of 0.5 or higher between the DIGIKOST-FFQ and the reference methods, a sample size of 38 men and 38 women is required to achieve a significance level of 5% and a power of 90% [44].

Statistical Analyses

All analyses were performed using SPSS software (version 29; IBM Corp). Normal distribution for continuous variables was checked by visual inspection of histograms and quantile-quantile plots. Dietary and activity variables were presented in median and IQR. Comparisons between the methods were presented as median differences, and categorical variables were presented in frequencies and percentages. Paired 2-tailed *t* tests were used to compare differences in normal distributed variables, and Wilcoxon signed rank test in nonnormal distributed variables. *P* values <.05 were considered statistically significant. Ranking of individual dietary intakes and activity levels were tested by Spearman ρ for nonparametric variables and the 95% CI for the correlation estimate. A correlation below 0.3 was defined as poor, whereas a correlation between 0.3 and 0.49 was fair, and above 0.5 was satisfactory, according to Hankin et al [45].

Cross-classification of individual intakes between the methods was estimated by ranking participants' intakes, dividing them into quartiles, and calculating how many were classified in the same and same plus adjacent, misclassified with 2 quartiles, or classified in the opposite quartile (grossly misclassified, 3 quartiles apart).

Bland-Altman plots were used to explore and visualize bias, such as under- or overreporting (mean differences), limits of agreement (mean difference and 1.96 SD), and the presence of outliers in the data. Sensitivity and specificity were calculated to evaluate the participants' adherence to the Norwegian FBDGs identified by both methods. The sensitivity was defined as the proportion of participants who were classified as not fulfilling the recommendations both by the DIGIKOST-FFQ and the WR or SWA divided by the number of participants not fulfilling the recommendations according to the WR or SWA only. Specificity was defined as the proportion of participants who were classified as fulfilling the recommendations both by the DIGIKOST-FFQ and WR or SWA divided by the number of participants fulfilling the recommendations according to the WR or SWA only. Values above 60% were defined as good for both measures. Adherence to the Norwegian diet index and physical activity recommendations were examined on a continuous scale. For all the different diet and physical activity recommendations, participants were assigned points of adherence, defined as high (1 point), intermediate (0.5 point), and low (0 point) to each specific recommendation.

Ethical Considerations

This study was carried out in accordance with the Helsinki Declaration. The Norwegian Data Protection Services (Sikt, project registration 277679) has approved the DIGIKOST protocol and the informed consent. The participant signed the informed consent electronically prior to participating in the study. They were also informed about the ability to opt out of

XSL•FO
RenderX

the study as well as information regarding the safe storage of all personal data. All data were analyzed in the safe and approved server, and with restricted access, called services for sensitive data at the University of Oslo [28]. Only tables and figures without individual sensitive data were exported. No data were collected about the invited participants who did not participate.

Results

Participant Characteristics

In total, 77 (male: n=16 and female: n=61) participants, corresponding to 6% (n=77) of the 1249 who were invited and 59% (n=77) of the 131 who consented, participated in this study. All of these 77 participants fulfilled the inclusion criterion by the completion of both the DIGIKOST-FFQ and the WR. Of these, 56 (73%) also used the SWA (Figure 1 and Table 1). The mean age was 45 (SD 14.6) years, and the mean BMI was 24.5 (SD 3.9) kg/m^2. Most of the participants were highly educated (college or higher: n=70, 91%) and currently employed (n=53, 69%). Total daily energy intake estimated from the WR was on mean 8057 (SD 2157) kJ, and energy expenditure measured from the SWA was on mean 10,365 (SD 1827) kJ/day (Table 1).

Figure 1. Flow of participants. FFQ: food frequency questionnaire; SWA: SenseWear Armband Mini; WR: weighed food record.

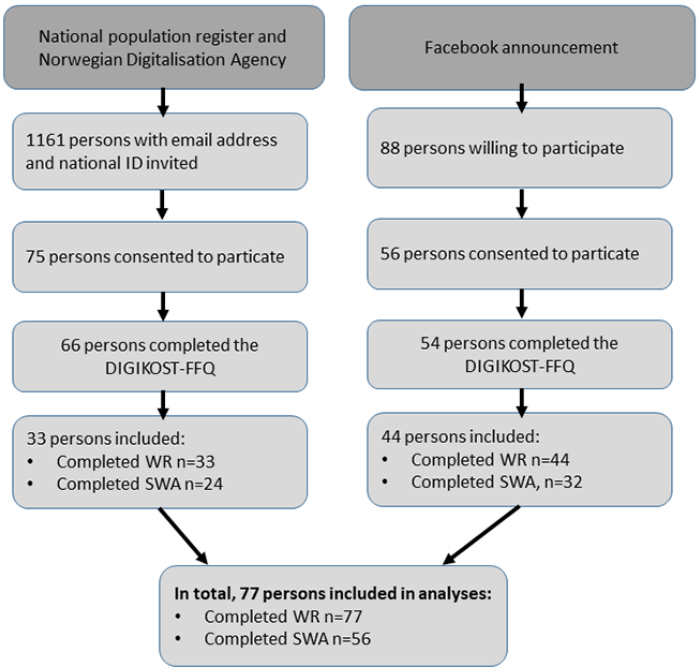

XSL•FO
RenderX

Table 1. Participant's characteristics, anthropometrics, and demography (N=77).

	Values
General characteristics	
Gender, n (%)	
Men	16 (21)
Women	61 (79)
Age (years), mean (SD)	45.2 (14.6)
Energy intake per day (kJ), mean (SD)	8057 (2157)
Days between completion of tools	
Mean (SD)	19.4 (16.8)
Range	1-94
Activity level per day (n=56), mean (SD)	
Total energy expenditure (kJ)	10,365 (1827)
Active energy expenditure (kJ)	4884 (1680)
Steps	8868 (3780)
Anthropometry, mean (SD)	
Height (cm)	170.4 (8.1)
Weight (kg)	71.4 (12.8)
BMI (kg/m^2)	24.5 (3.9)
Education, n (%)	
Upper secondary school	7 (9)
College or university	70 (91)
Ethnicity, n (%)	
Asia	1 (1)
Europe	74 (96)
Other country	2 (3)
Work status, n (%)	
In work or employed	53 (69)
Student	9 (13)
Retired	6 (8)
Sick leave and unemployed	8 (10)

Dietary Intakes From the DIGIKOST-FFQ and WR

Overview

Estimates of median dietary intakes from DIGIKOST-FFQ and WR, median differences, correlation coefficients, and cross-classifications are presented in Table 2. All estimated intakes of food groups from the DIGIKOST-FFQ were significantly different from the WR estimated intakes, except for processed meat, sugar-rich beverages, and dietary supplements.

XSL•FO
RenderX

Table 2. Estimated intake of food groups from DIGIKOST-FFQ[a] and WR[b] for all participants (N=77).[c]

Food group	DIGIKOST-FFQ (g/day), median (P10-P90)	WR (g/day), median (P10-P90)	Difference (g/day), median (P10-P90)	*P*value[d]	Spearman ρ (95% CI)	Correct quartile (%)	Correct+adjacent quartile (%)	Misclassification 2 quartiles (%)	Gross misclassification (%)
Fruits and berries	189 (81-431)	157 (41-495)	31 (–107 to 203)	.02	0.64 (0.48-0.76)	42	87	12	1
Berries	29 (0-93)	14 (0-90)	6.5 (–41 to –86)	.02	0.35 (0.13-0.54)	42	71	21	8
Vegetables	212 (107-414)	150 (53-304)	45 (–120 to 244)	<.001	0.24 (0.0-0.45)	30	69	21	10
Legumes	14.0 (0-57)	0 (0-61)	1.6 (–33 to 37)	.03	0.45 (0.25-0.62)	30	84	16	0
Unsalted nuts	9 (0-30)	0 (0-18)	2.3 (–6 to 26)	<.001	0.48 (0.28-0.64)	50	74	13	13
Whole grains	77 (34-155)	39 (11-84)	43 (–10 to 94)	<.001	0.51 (0.32-0.67)	47	81	17	3
Fish	60 (18-146)	43 (0-120)	22 (–27 to 84)	<.001	0.52 (0.33-0.67)	41	80	17	3
Fatty fish	28 (10-83)	28 (0-79)	10 (–23 to 45)	.002	0.61 (0.44-0.74)	43	82	18	0
Low-fat dairy products	70 (0-351)	150 (1-370)	–37 (–228 to 137)	<.001	0.65 (0.49-0.77)	47	86	12	3
High-fat dairy products	14 (3-131)	31 (6-72)	–9.4 (–51 to 70)	<.001	0.33 (0.11-0.52	35	73	20	8
Red meat	35 (0-105)	54 (0-145)	–13 (–77 to 35)	.002	0.58 (0.40-0.72)	48	84	13	3
Processed meat	29 (0-96)	38 (0-101)	–2.5 (–60 to 40)	.13	0.54 (0.35-0.69)	51	82	13	5
Water	1100 (300-1737)	757 (94-1968)	230 (–737 to 1020)	.01	0.43 (0.22-0.59)	42	79	17	4
Sugar-rich beverages	0.0 (0-69)	28 (0-95)	0.0 (–60 to 57)	.14	0.30 (0.08-0.50)	58	82	18	0
Ethanol	6 (0-20)	8 (0-29)	–0.8 (–17 to 5)	.002	0.71 (0.57-0.81)	54	88	9	3
Sugar- and fat-rich foods	34 (0-114)	64 (18-148)	–27 (–113 to 29)	<.001	0.43 (0.22-0.60)	40	79	14	7

[a]FFQ: food frequency questionnaire.

[b]WR: weighed food record.

[c]For dietary supplements, n=40 (52%) and n=14 (18%) for DIGIKOST-FFQ and WR, respectively, and Fischer exact test was performed (*P*<.39).

[d]Wilcoxon signed rank test.

The median differences were however generally small and below a regular portion size for the different food groups. The ranking of individual intakes was fair and satisfactory for all food groups and poor for "vegetables." Based on the Bland-Altman plot evaluations, the agreement between the 2 methods was good, with most of the differences within the 95% limit of agreement for each food group and evenly distributed above and below the mean difference (Figures 2 and 3). In total, 69% or more of the participants were classified into the same or adjacent quartile for all food groups, and gross misclassification ranged from 0 for legumes, fatty fish, and alcohol to 13% for unsalted nuts (Table 2). The sensitivity and specificity of the DIGIKOST-FFQ are presented in Table 3. The sensitivity for different food groups ranged from 32% for red meat to 90% for low-fat dairy products. For specificity, the food groups ranged from 40% for unsalted nuts to 100% for sugar- and fat-rich foods. Most of the participants achieved intermediate adherence to the diet score recommendation with both methods (Table 4) [32]. More detailed results are presented in the sections below.

XSL•FO
RenderX

Figure 2. Bland-Altman plots depicting the mean differences (DIGIKOST–WR) for intake of food groups in grams per day: (A) fruits and berries including juice (LoA: –265 to 323), (B) vegetables (LoA: –201 to 329), (C) unsalted nuts (LoA: –18 to 30), (D) whole grains (LoA: –48 to 133), (E) fish (LoA: –68 to 117), (F) fatty fish (LoA: –48 to 71), (G) red meat (LoA: –118 to 94), and (H) processed meat (–96 to 83). The solid line represents the mean, and the dashed lines represent the LoA equal to 1.96 SDs of the observations. FFQ: food frequency questionnaire; LoA: limits of agreement; WR: weighed food dairy.

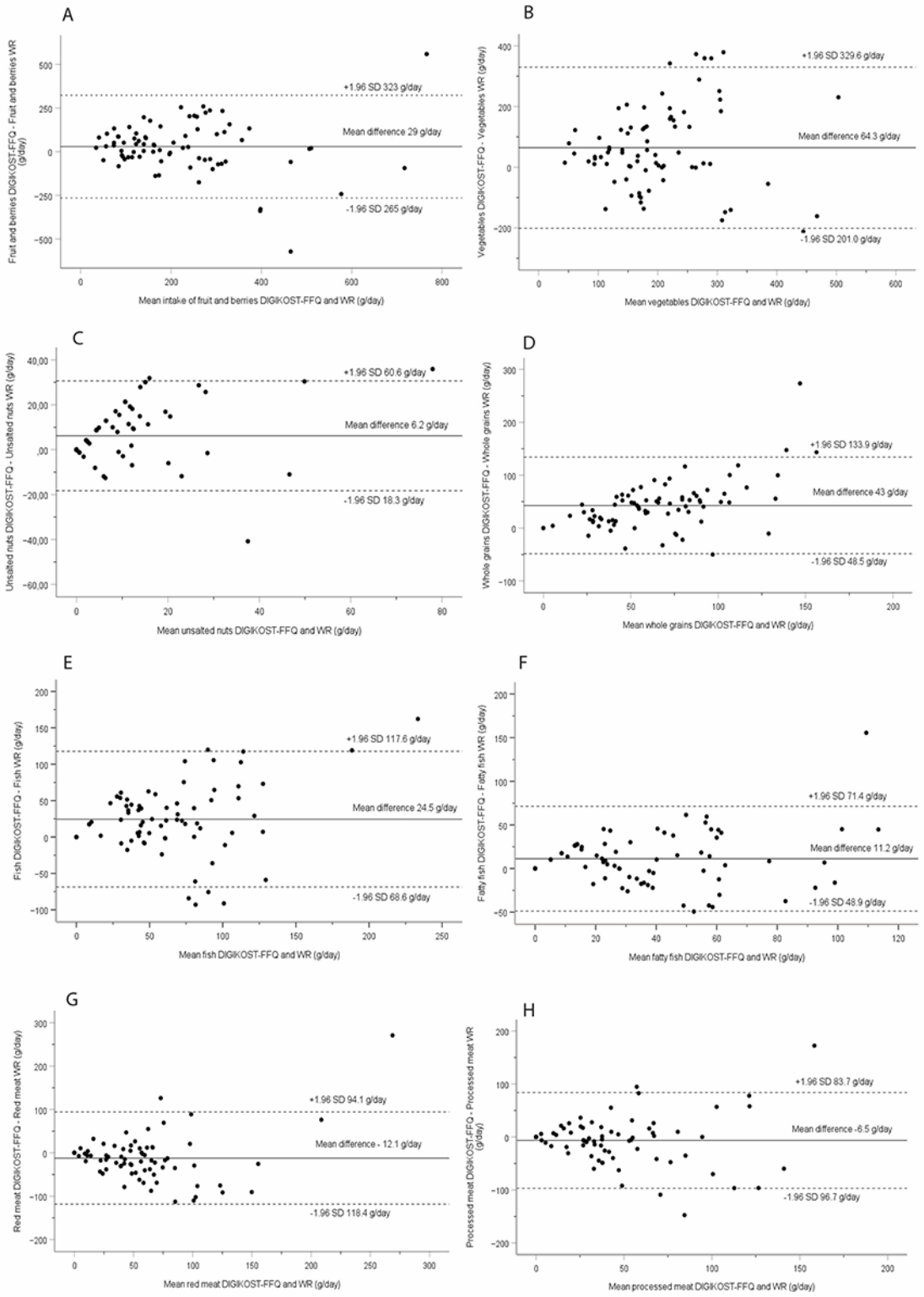

XSL•FO
RenderX

Figure 3. Bland-Altman plots depicting the mean differences (DIGIKOST–WR) for intake of food groups in grams per day: (A) low-fat dairy products (LoA: –468 to 399), (B) high-fat dairy products (LoA: –124 to 134), (C) berries (LoA: –90 to 111), (D) legumes (LoA: –61 to 72), (E) water (LoA: –1228 to 1572), (F) ethanol (LoA: –23 to 15), (G) sugar- and fat-rich foods (LoA: –163 to 96), and (H) sugar-rich beverages (LoA: –136 to 131). The solid line represents the mean, and the dashed lines represent the LoA equal to 1.96 SDs of the observations. FFQ: food frequency questionnaire; LoA: limits of agreement; WR: weighed food dairy.

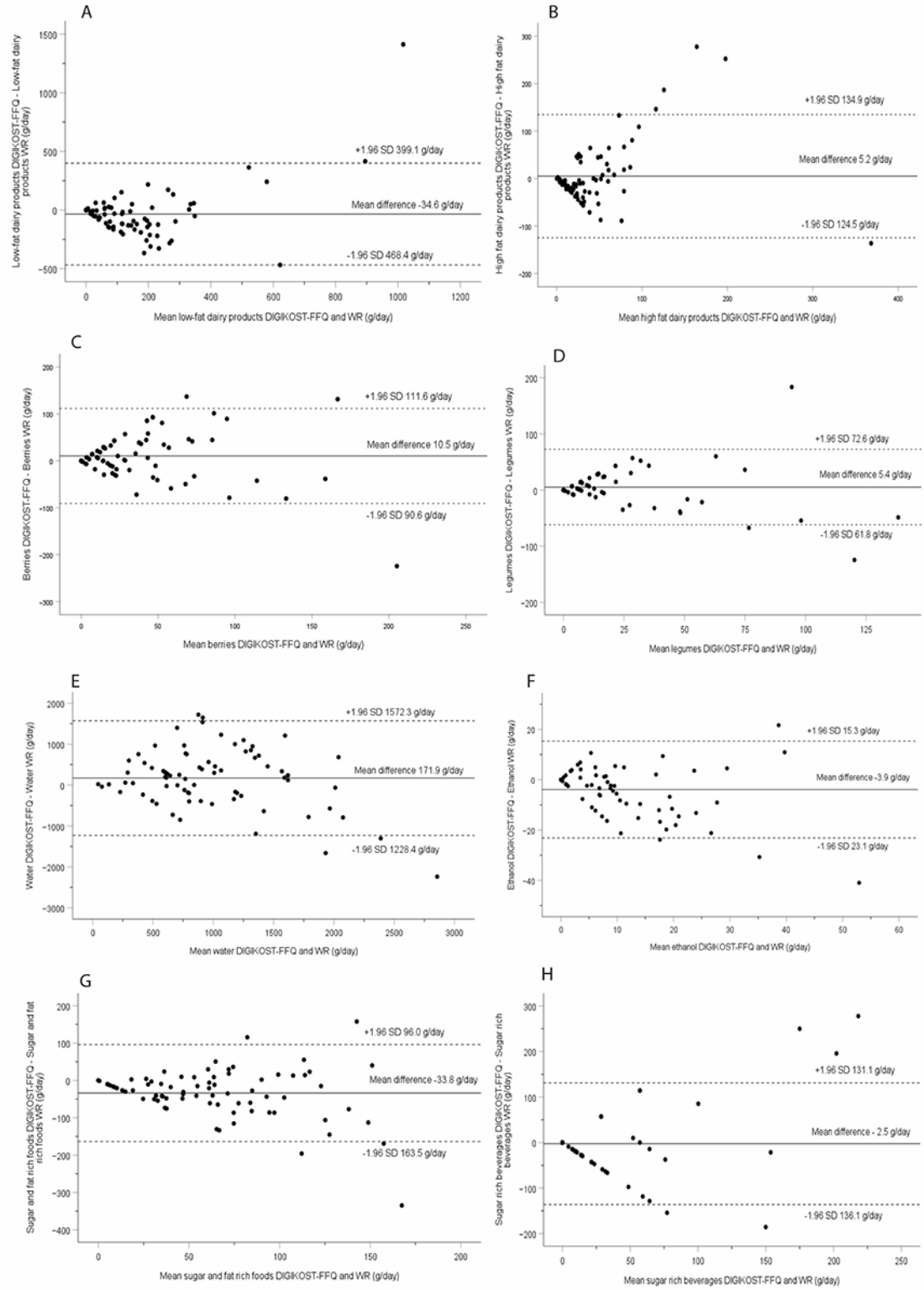

XSL•FO
RenderX

Table 3. Proportion of individuals with low adherence (sensitivity) and high adherence (specificity) measured by both assessment methods.

Variables	Sensitivity (low adherence) (%)	Specificity (high adherence) (%)
Fruits and berries	80	70
Vegetables	70	82
Unsalted nuts	82	40
Whole grains	41	85
Fish (lean and fat)	33	87
Red meat (processed and unprocessed)	32	88
Processed meat (white and red)	77	56
Sugar- and fat-rich foods	73	100
Drinks with added sugar	33	87
Low-fat dairy products	90	55
Margarines, butter, and oils	62	66
Ethanol	80	65
Dietary supplements	64	49
MVPA[a] score[b]	100	65.5

[a]MVPA: moderate to vigorous physical activity.
[b]Fischer exact test ($P<.4$).

Table 4. Adherence to composite diet score and MVPA[a] score (components in the Norwegian lifestyle index [32]).

Component score and assessment tool	Adherence (%)		
	0[b]	0.5[c]	1[d]
MVPA score			
DIGIKOST-FFQ[e]	16.1	19.6	64.3
SWA[f]	28.6	0.0	71.4
Diet score			
DIGIKOST-FFQ	18.2	66.2	15.6
WR[g]	10.4	64.9	24.7

[a]MVPA: moderate to vigorous physical activity.
[b]No adherence.
[c]Intermediate adherence.
[d]High adherence.
[e]FFQ: food frequency questionnaire.
[f]SWA: SenseWear Armband Mini.
[g]WR: weighed food record.

Fruits, Berries, Vegetables, Whole Grains, and Unsalted Nuts

The DIGIKOST-FFQ slightly overestimated intakes of "fruits and berries" with a median difference of 31 g/day (Table 2). The correlation was satisfactory (r=0.64), and 87% of the individual intakes were classified in the correct or adjacent quartile, and only 1% was grossly misclassified. For intakes up to 200 g/day, the limits of agreement were 200 g above and below the mean difference, which increased with higher intakes (Figure 2A). "Vegetables" were overestimated with 45 g/day and with a similar distribution of limits of agreement as for "fruits and berries" (Figure 2B). In total, 69% of the individual intakes of "vegetables" were classified in the correct or adjacent quartile, and 10% were grossly misclassified. The ranking of individual intakes of "vegetables" was poor (r=0.24). The median difference for intakes of "whole grains" was 43 g/day, the correlation was satisfactory (r=0.51), and 81% of the participants were classified in the same or adjacent quartile, whereas 1% was grossly misclassified. Limits of agreements ranged from –48 to 133 g/day, and the plot indicated a trend for overestimation with increased intakes (Figure 2D). "Unsalted nuts" were estimated with a median difference of 2.3 g/day and with a borderline satisfactory correlation (r=0.48). In total, 74%

XSL•FO
RenderX

of the participants were classified in the same or adjacent quartile, and 13% were grossly misclassified (Table 2). The distribution of the differences increased with higher intakes, and the limits of agreement were wide (Figure 2C).

Fish, Dairy Products, and Meat

Median differences for "fish" and "fatty fish" intakes were 22 and 10 g/day, respectively (Table 2). Ranking of individual intakes was satisfactory with correlations of r=0.52 and r=0.61 for "fish" and "fatty fish," respectively. Limits of agreements ranged from –68 to 117 g/day, and the distribution of the differences increased with higher mean intakes up to about 100 g/day of "fish." For "fatty fish," the agreement was good with evenly distributed differences above and below the mean difference and narrow limits of agreement (–49 to 71 g/day; Figure 2E and F). Classification of individual intakes was good for both "fish" and "fatty fish" with low gross misclassification (Table 2).

The DIGIKOST-FFQ showed a minor underestimation of dairy products on the group level, acceptable ranking (r=0.65), low gross misclassification of individual intakes, and good agreement between the methods (Table 2 and Figure 3A and B). The DIGIKOST-FFQ was able to estimate intakes of "red meat" and "processed meat" on the group level and to rank individual intakes of both food groups (Table 2 and Figure 2G and H). More than 82% of the participants were classified in the same or adjacent quartile, and few were grossly misclassified.

Sugar- and Fat-Rich Foods and Beverages, Other Beverages, and Dietary Supplements

There were good agreements between the methods in estimating intakes of "sugar- and fat-rich foods" and "sugar-rich beverages" on the group level, fair ranking of individual intakes, and good quartile classifications (Table 2 and Figure 3G and H). Intakes of "water" were overestimated by the DIGIKOST-FFQ, and the ranking of individual intake was fair (Table 2 and Figure 3E). The DIGIKOST-FFQ was able to estimate the intake of "ethanol" both on group and individual levels (Table 2). More than 79% of the participants were classified in the same or adjacent quartile for intakes of "sugar- and fat-rich foods," "sugar-rich beverages," "water," and "ethanol" (Table 2).

Physical Activity, Sleep, and Sedentary Time

The DIGIKOST-FFQ underestimated time in moderate intensity (median –468 min/week), which increased with higher amounts of time in moderate physical intensity (Table 5 and Figure 4A). The correlation between the methods was borderline fair (r=0.29; Table 4). In total, 73% of the participants were classified in the correct or adjacent quartile, and 5% were grossly misclassified. On the group level, vigorous intensity was overestimated by 18 min/week, and the individual differences increased with increased time in "vigorous physical activity" (Table 5 and Figure 4B). There was a fair correlation between the methods for vigorous intensity (P=.40), and no gross misclassification. The DIGIKOST-FFQ was able to rank individual time in "moderate to vigorous intensity" (P=.64), and 79% were classified in the correct or adjacent quartile, and few were grossly misclassified (Table 5). On the group level, time in "moderate to vigorous intensity" was underestimated by DIGIKOST-FFQ, which also increased with higher amounts of time in this intensity (Table 5 and Figure 4C). The underreporting of "moderate to vigorous intensity" was particularly seen among individuals with BMI <25 kg/m^2 (Figure 4D).

Table 5. Estimated physical activity, sedentary time, and sleep from DIGIKOST-FFQ[a] and SWA[b] for all participants (n=56).

Physical activity intensity	DIGIKOST-FFQ, median (P10-P90)	SWA, median (P10-P90)	Difference, median (P10-P90)	*P*value[c]	Spearman ρ (95% CI)	Correct quartile	Correct+adjacent quartile (%)	Misclassification 2 quartiles (%)	Grossly misclassification (%)
MPA[d] (min/week)	123 (33-428)	753 (247-1492)	–468 (–1333 to –156)	<.001	0.29 (0.02 to 0.50)	30	73	21	5
VPA[e] (min/week)	62 (0-226)	21 (0-178)	18 (–42 to 123)	<.001	0.64 (0.45 to 0.78)	43	82	18	0
MVPA[f] (min/week)	236 (48-507)	754 (258-1534)	–453 (–1424 to 123)	<.001	0.40 (0.14 to 0.60)	34	79	16	5
Sedentary time (not sleep; h/day)	10 (5-14)	11 (8-13)	–0.63 (–6 to 3)	.06	0.27 (–0.001 to 0.5)	41	75	16	9
Sleep (h/day)	7 (7-8)	7 (6-8)	0.21 (–1.3 to 1.7)	.03	0.19 (0.08 to 0.43)	36	71	20	9

[a]FFQ: food frequency questionnaire.

[b]SWA: SenseWear Armband Mini.

[c]Wilcoxon signed rank test.

[d]MPA: moderate physical activity.

[e]VPA: vigorous physical activity.

[f]MVPA: moderate to vigorous physical activity.

XSL•FO
RenderX

Figure 4. Bland-Altman plots depicting the mean differences (DIGIKOST–activity sensor [SWA]) for activities in minutes per week: (A) MPA (LoA: –1423 to 227), (B) VPA (LoA: –118 to 183), (C) MVPA (LoA: –1424 to 293), (D) BMI, (E) ST (LoA: –7 to 6), and (F) sleep (–1.7 to 1.9). The solid line represents the mean, and the dashed lines represent the LoA equal to 1.96 SDs of the observations. FFQ: food frequency questionnaire; LOA: limits of agreement; MPA: moderate physical activity; MVPA: moderate to vigorous physical activity; ST: sedentary time; SWA: SenseWear Armband Mini; VPA: vigorous physical activity.

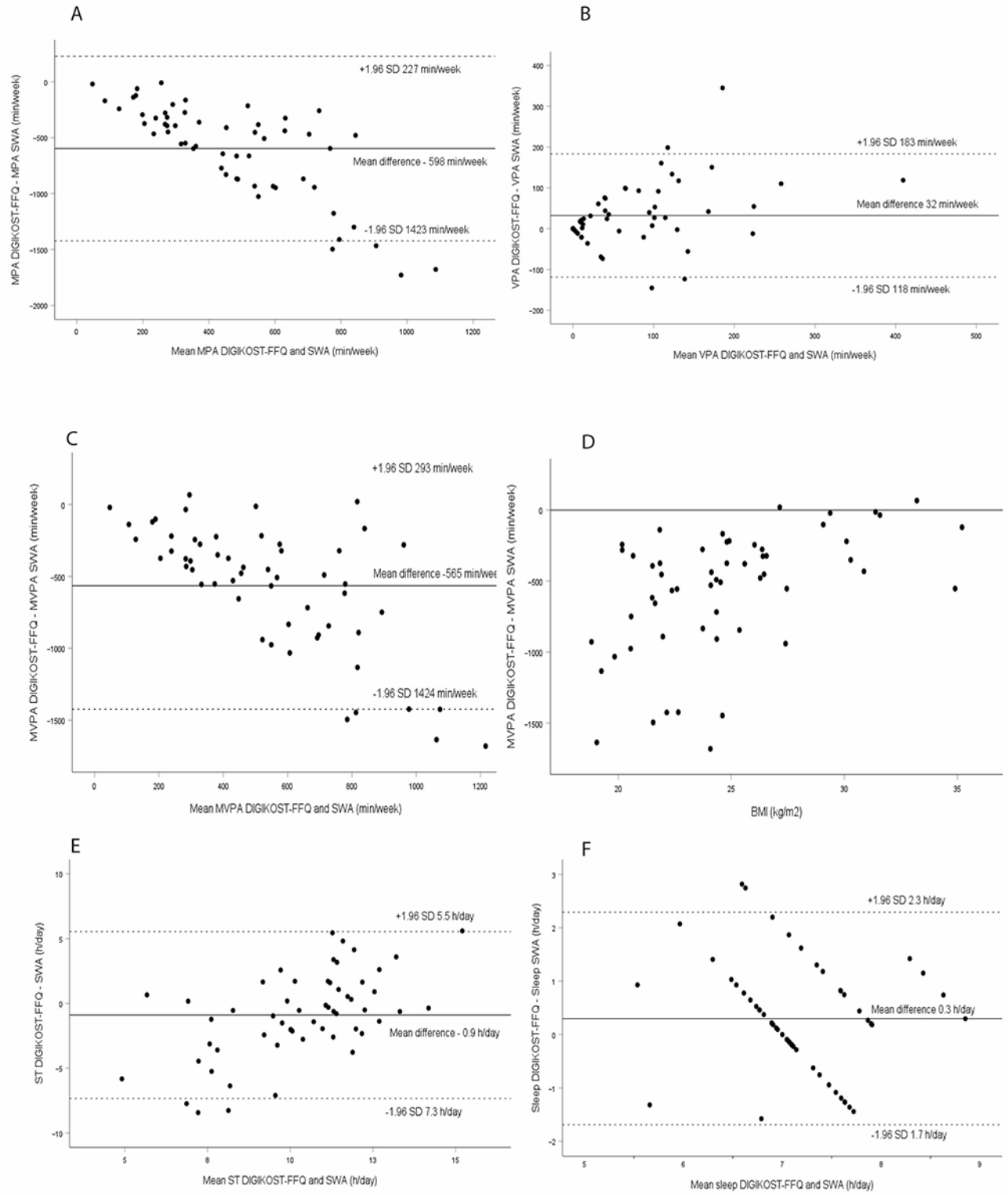

The DIGIKOST-FFQ was able to estimate time in "sedentary" activities on the group level, with a trend for overreporting with higher amounts of time in this activity, and the correlation between the methods was poor (r=0.27; Table 5 and Figure 4E). A high number of participants (75%) were classified in the correct or adjacent quartile, and 9% were grossly misclassified.

Median time in "sleep" was estimated to be 7 hours per day with both methods, with a trend for underreporting by the DIGIKOST-FFQ with increased time in "sleep" (Table 5 and Figure 4F). The correlation was poor (r=0.19), but the classification of individual time in quartiles was good (71%; Table 5).

XSL•FO
RenderX

Adherence to Dietary Guidelines and Physical Activity Recommendations

The sensitivity was above 60% for all dietary guidelines except for "whole grains," "fish," "red meat," and "drinks with added sugar" (Table 3). The specificity was above 60% for all dietary guidelines except for "unsalted nuts," "processed meat," "low-fat dairy products," and "dietary supplements" (Table 3). The total mean diet score of the Norwegian diet index (ranging from 0 to 20 points) was estimated to be 11.2 (SD 3.2) and 10.4 (SD 3.1) points with the DIGIKOST-FFQ and WR, respectively (Multimedia Appendix 2) [32]. Adherences to each component in the Norwegian diet index measured by DIGIKOST-FFQ, WR, and SWA are shown in Tables 4 and 6. Intermediate adherences were the most frequent for the composite "diet score," "fruits and berries," "vegetables," and "margarine, butter, and oils" estimated with both DIGIKOST-FFQ and WR. Low adherences were shown by both methods for "unsalted nuts," "processed meat," and "sugar- and fat-rich foods," whereas high adherences were shown for "fish," "sugar-rich drinks," and "dietary supplements." Approximately, evenly distributed adherences (ie, low, intermediate, and high) with both methods were shown for "red meat" (Table 6). High adherence to the recommendation of physical activity was estimated as the most frequent by both methods (Table 4).

XSL•FO
RenderX

Table 6. Adherence to the Norwegian diet index [32].

Food component and dietary assessment tool	Adherence (%)		
	0[a]	0.5[b]	1[c]
Fruits and berries			
DIGIKOST-FFQ[d]	24.7	40.3	35.1
WR[e]	37.7	32.5	29.9
Vegetables			
DIGIKOST-FFQ	16.9	37.7	45.5
WR	41.6	44.2	14.3
Unsalted nuts			
DIGIKOST-FFQ	63.6	16.9	19.5
WR	81.8	11.7	6.5
Whole grains			
DIGIKOST-FFQ	11.7	24.7	63.6
WR	45.5	37.7	16.9
Fish			
DIGIKOST-FFQ	13.0	10.4	76.6
WR	28.6	22.1	49.4
Red meat			
DIGIKOST-FFQ	19.5	29.9	50.6
WR	36.4	32.5	31.2
Processed meat			
DIGIKOST-FFQ	66.2	13.0	20.8
WR	67.5	10.4	22.1
Margarine or oils			
DIGIKOST-FFQ	13.0	37.7	49.4
WR	26.0	35.1	37.7
Dairy			
DIGIKOST-FFQ	39.0	23.4	37.7
WR	24.7	14.3	61
Foods high in sugar and fat			
DIGIKOST-FFQ	64.9	10.4	24.7
WR	89.6	7.8	2.6
Sugar-sweetened beverages			
DIGIKOST-FFQ	20.8	0	79.2
WR	3.09	6.5	54.5
Dietary supplements			
DIGIKOST-FFQ	53.2	__[f]	46.8
WR	18.2	—	81.8

[a]No adherence.
[b]Intermediate adherence.
[c]High adherence.
[d]FFQ: food frequency questionnaire.
[e]WR: weighed food record.

XSL•FO
RenderX

[f]Not available.

Discussion

Principal Findings

We investigated the relative validity of the new digital assessment tool, DIGIKOST-FFQ, with 7-day WR and activity sensors as reference methods. Generally, the DIGIKOST-FFQ was able to rank individual intakes of foods and time spent in physical activities adequately and also able to evaluate dietary intakes and adherences to the Norwegian FBDG and physical activity recommendation.

The DIGIKOST-FFQ is based on the paper form NORDIET-FFQ, and improvements were made to enhance the quality of the DIGIKOST-FFQ estimates. For example, in the DIGIKOST-FFQ, questions about fruits and berries were presented as single food items as opposed to aggregated questions in the paper FFQ, which underreported the fruit, berry, and vegetable intakes [26]. In this validation study, improved correlations and agreement between the test and reference methods for estimated fruit, berry, and vegetable intakes were observed. Moreover, WR has been documented to underreport energy and food intake [46,47]. Thus, the overreporting of fruit and vegetable intakes in the DIGIKOST-FFQ along with the improved correlations and agreement between the methods in this validation study indicate a higher accuracy of the intake estimates from the DIGIKOST-FFQ for these food groups as compared to the NORDIET-FFQ. In particular, the DIGIKOST-FFQ showed a borderline fair correlation in ranking individual intakes of vegetables. However, the median difference between the methods was well below 1 portion of vegetable (ie, 100 g), the agreement between the methods was good, and the DIGIKOST-FFQ was able to identify adherence to the recommendation of vegetables.

Few participants reported intakes of unsalted nuts, and low amounts were reported for those with intakes. This may explain the high share grossly misclassification of nut intakes in this study.

Adding single-item questions regarding whole grain products along with automatic counting of slices of bread consumed per week to help the responder to report intakes of bread in the DIGIKOST-FFQ (Multimedia Appendix 1) may have successfully contributed to the improvement of the reported intakes of whole grains. In addition, legumes were added as a new item in the DIGIKOST-FFQ, and the good agreement and correlation observed for intakes of legumes contributed to the overall good ability of DIGIKOST-FFQ to identify individuals with different adherences to the recommendations for all fiber-rich foods.

Reported intakes of fish, meat, dairy products, sugar- and fat-rich foods, sugar-rich beverages, ethanol, water, and dietary supplements were good with the DIGIKOST-FFQ. We hypothesize that the overreporting of water, corresponding to 1 glass of water per day, might be explained by the warmer climate in June 2021, during which most of the participants completed the DIGIKOST-FFQ, as compared to the colder climate in August 2021 when they completed the WR, thus resulting in higher intakes of water in June.

Underreporting of physical activity by questionnaires is a common issue when compared to objective activity sensors, which record all activities within 24 hours a day, whereas questionnaires rely only on a few questions supported by others [48,49]. In particular, sex and BMI have been shown to alter the association between self-reported and objectively measured physical activity [49]. Additionally, questionnaires are subjected to memory and conceptualization of different degree of intensities, such as the difference between moderate and vigorous intensities. In this study, there was a trend for normal-weighed individuals to underreport moderate physical activity more than overweighed or obese individuals. Overall, on the group level, the DIGIKOST-FFQ underreported moderate to vigorous physical activity by 450 minutes per week when compared to SWA. In particular, low and moderate intensities of physical activity, such as transport to work (eg, biking), gardening, and housework, have been challenging to report in questionnaires [50,51], which was also observed in this study. Moreover, the aim of the DIGIKOST tool is to estimate adherence to the recommendations and to give advice in the individual report in order to spend more time in physical activity. Therefore, ranking individual time in different activities and classifying the individuals according to adherence are important abilities of the assessment tool.

Adding illustrations of different intensities with images and text for the participants to better differentiate between the performance of physical activity during the day and week may increase the quality of the estimated time in moderate physical activity. Generally, the results from this study showed that the DIGIKOST-FFQ was able to rank individual dietary intakes and identify adherences to most of the dietary guidelines and physical activity recommendations. This conforms to previous studies showing the ability of FFQs to classify individuals according to different intakes of foods and time in activities [8,9].

Strengths and Limitations

We included 77 participants in the study, which is an adequate number in validation studies. The announcement by email lists from the National Registry of Norway was accompanied by an advertisement in Facebook. Recruitment through Facebook has shown to be the most successful strategy. Generally, the use of social media for recruitment in health research has shown to be effective in hard-to-reach populations [52,53]. In this study, most of the participants were women with higher education than the general Norwegian population [54]. This might compromise the external validity of this study.

In this study, the DIGIKOST-FFQ (test method) was administered prior to the reference methods in order to avoid the learning effect from the reference method [11]. Moreover, the DIGIKOST-FFQ asks for dietary intakes and activities the last 2 months, whereas the reference methods (WR and SWA) prospectively recorded dietary intake and activities for 7 consecutive days and completed within 1-2 months after

XSL•FO
RenderX

fulfilling the DIGIKOST-FFQ. The average range of days between completion of DIGIKOST-FFQ and reference methods was 19 (SD 16.8) days. The expectancy of meaningful changes in dietary intakes or activities during this period is low.

The measurement errors usually associated with printed questionnaires have also been documented in digital questionnaires such as underreporting and recall errors [6,10,12,26,31,41,55,56]. Closed questions, predefined categories of frequency and amounts, and foods not included in the questionnaire may result in misclassification and underreporting. In DIGIKOST-FFQ, underreporting was documented for some food groups and activities, but there were few gross misclassifications, which indicates the good ability to identify individuals with different adherence to the recommendations. Ideally, in validation studies, the error in the reference method should be independent from the test method [16]. This study used a self-reported instrument, that is, the WR, as the reference method against the DIGIKOST-FFQ. Since both methods are self-reported, the errors may not be independent, thus inflating the validity correlation coefficient and an optimistic evaluation of the validity of the FFQ. Even though the WR is commonly used as a reference method in the validation of FFQs [11,12] and does not rely on the memory of food intakes, no interpretation of portion size is needed due to the weighing. Moreover, the WR provides data on foods that are really consumed, which are not provided by using biomarkers as a reference method. The DIGIKOST-FFQ does not cover a whole diet, since the aim of the tool is to estimate adherence to the recommendations, thus biomarkers for energy intake or nutrients are not used in this study. In the future, revisions of the DIGIKOST tool developments may be conducted to enable estimations of energy and food intakes; however, that is not the design and aim of this version.

Conclusions

The DIGIKOST-FFQ was able to assess dietary intake and physical activity and identify individuals with different adherences to the Norwegian FBDG and physical activity recommendation. However, moderate physical activity was underreported, water was overreported, and vegetables showed a low correlation between the methods, which is important to take into account when interpreting the data in future studies. The new digital questionnaire contributes with valuable data for research projects focusing on improved lifestyle, prevention of diseases, and increased quality of life and survival, as well as a screening tool in the clinic.

Acknowledgments

The authors would like to acknowledge project coordinator Nina Nordberg Gjerrestad and the participants in the study. The authors would also like to thank Professor Sveinung Berntsen for generating the physical activity data from the SenseWear Armband. The Throne Holst Foundation for Nutrition Research and Aktieselskabet Freia Chocolade Fabriks Medisinske Fond funded this study.

Authors' Contributions

HBH had the main responsibility for writing the paper. HBH, MDK, MHC, AH, and RB contributed to the conception and the design of the study and drafting of the paper. MDK and HBH contributed to the acquisition of data. HBH, MDK, and MHC contributed to analyses. All authors contributed to the interpretation of the data and the writing and approval of the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
DIGIKOST: a digital application for diet and lifestyle assessment and benchmarking against national guidelines.
[PDF File (Adobe PDF File), 10042 KB - jmir_v26i1e53442_app1.pdf]

Multimedia Appendix 2
Adherence to total diet score and each single component measured with both assessment tools.
[DOCX File , 15 KB - jmir_v26i1e53442_app2.docx]

References

1. Diet, nutrition and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation. World Health Organization. 2003. URL: https://www.who.int/publications/i/item/924120916X [accessed 2024-03-27]
2. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. World Cancer Research Fund/American Institute for Cancer Research. 2007. URL: https://www3.paho.org/hq/dmdocuments/2011/nutrition-AICR-WCR-food-physical-activ.pdf [accessed 2024-03-27]
3. Kostråd for å Fremme Folkehelsen og Forebygge Kroniske Sykdommer Metodologi og Vitenskapelig Kunnskapsgrunnlag: Nasjonalt Råd for Ernæring 2011. Oslo: Helsedirektoratet; 2011:353.

XSL•FO
RenderX

4. Diet, nutrition, physical activity and cancer: a global prespective. World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. 2018. URL: https://www.wcrf.org/wp-content/uploads/2021/02/Summary-of-Third-Expert-Report-2018.pdf [accessed 2024-03-27]
5. Scientific report of the 2020 Dietary Guidelines Advisory Committee: Advisory report to the Secretary of Agriculture and the Secretary of Health and Human Services. Dietary Guidelines Advisory Committee. 2020. URL: https://www.dietaryguidelines.gov/sites/default/files/2020-07/ScientificReport_of_the_2020DietaryGuidelinesAdvisoryCommittee_first-print.pdf [accessed 2024-04-17]
6. Cade JE. Measuring diet in the 21st century: use of new technologies. Proc Nutr Soc 2017;76(3):276-282 [FREE Full text] [doi: 10.1017/S0029665116002883] [Medline: 27976605]
7. Conrad J, Nöthlings U. Innovative approaches to estimate individual usual dietary intake in large-scale epidemiological studies. Proc Nutr Soc 2017;76(3):213-219 [FREE Full text] [doi: 10.1017/S0029665116003025] [Medline: 28162110]
8. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009 [FREE Full text] [doi: 10.4178/epih/e2014009] [Medline: 25078382]
9. Veierød MB, Thelle D, Hjartåker A, Laake P. Epidemiologiske og Kliniske Forskningsmetoder. Oslo: Gyldendal Akademisk; 2007:33-44.
10. Cade J, Thompson R, Burley V, Warm D. Development, validation and utilisation of food-frequency questionnaires—a review. Public Health Nutr 2002;5(4):567-587 [FREE Full text] [doi: 10.1079/PHN2001318] [Medline: 12186666]
11. Margetts BM, Nelson M. Design Concepts in Nutritional Epidemiology. New York: Oxford University Press; 1997.
12. Willett W. Nutritional Epidemiology. Oxford: Oxford University Press; 2013.
13. Araujo MC, Yokoo EM, Pereira RA. Validation and calibration of a semiquantitative food frequency questionnaire designed for adolescents. J Am Diet Assoc 2010;110(8):1170-1177 [FREE Full text] [doi: 10.1016/j.jada.2010.05.008] [Medline: 20656092]
14. Dehghan M, del Cerro S, Zhang X, Cuneo JM, Linetzky B, Diaz R, et al. Validation of a semi-quantitative food frequency questionnaire for Argentinean adults. PLoS One 2012;7(5):e37958 [FREE Full text] [doi: 10.1371/journal.pone.0037958] [Medline: 22662256]
15. Tucker KL, Smith CE, Lai CQ, Ordovas JM. Quantifying diet for nutrigenomic studies. Annu Rev Nutr 2013;33:349-371 [FREE Full text] [doi: 10.1146/annurev-nutr-072610-145203] [Medline: 23642200]
16. Kirkpatrick SI, Baranowski T, Subar AF, Tooze JA, Frongillo EA. Best practices for conducting and interpreting studies to validate self-report dietary assessment methods. J Acad Nutr Diet 2019;119(11):1801-1816 [FREE Full text] [doi: 10.1016/j.jand.2019.06.010] [Medline: 31521583]
17. Park Y, Dodd KW, Kipnis V, Thompson FE, Potischman N, Schoeller DA, et al. Comparison of self-reported dietary intakes from the automated self-administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. Am J Clin Nutr 2018;107(1):80-93 [FREE Full text] [doi: 10.1093/ajcn/nqx002] [Medline: 29381789]
18. Kirkpatrick SI, Troiano RP, Barrett B, Cunningham C, Subar AF, Park Y, et al. Measurement error affecting web- and paper-based dietary assessment instruments: insights from the multi-cohort eating and activity study for understanding reporting error. Am J Epidemiol 2022;191(6):1125-1139 [FREE Full text] [doi: 10.1093/aje/kwac026] [Medline: 35136928]
19. Kipnis V, Midthune D, Freedman LS, Bingham S, Schatzkin A, Subar A, et al. Empirical evidence of correlated biases in dietary assessment instruments and its implications. Am J Epidemiol 2001;153(4):394-403 [FREE Full text] [doi: 10.1093/aje/153.4.394] [Medline: 11207158]
20. Adamson AJ, Collerton J, Davies K, Foster E, Jagger C, Stamp E, et al. Nutrition in advanced age: dietary assessment in the Newcastle 85+ study. Eur J Clin Nutr 2009;63(Suppl 1):S6-S18 [FREE Full text] [doi: 10.1038/ejcn.2008.60] [Medline: 19190647]
21. Liu L, Wang PP, Roebothan B, Ryan A, Tucker CS, Colbourne J, et al. Assessing the validity of a self-administered food-frequency questionnaire (FFQ) in the adult population of Newfoundland and Labrador, Canada. Nutr J 2013;12:49 [FREE Full text] [doi: 10.1186/1475-2891-12-49] [Medline: 23590645]
22. Falomir Z, Arregui M, Madueño F, Corella D, Coltell Ó. Automation of food questionnaires in medical studies: a state-of-the-art review and future prospects. Comput Biol Med 2012;42(10):964-974 [FREE Full text] [doi: 10.1016/j.compbiomed.2012.07.008] [Medline: 22898338]
23. Forster H, Fallaize R, Gallagher C, O'Donovan CB, Woolhead C, Walsh MC, et al. Online dietary intake estimation: the Food4Me food frequency questionnaire. J Med Internet Res 2014;16(6):e150 [FREE Full text] [doi: 10.2196/jmir.3105] [Medline: 24911957]
24. Carlsen MH, Andersen LF, Hjartåker A. Reproducibility and feasibility of an online self-administered food frequency questionnaire for use among adult Norwegians. Food Nutr Res 2021;65:7561 [FREE Full text] [doi: 10.29219/fnr.v65.7561] [Medline: 34908922]
25. Kristal AR, Kolar AS, Fisher JL, Plascak JJ, Stumbo PJ, Weiss R, et al. Evaluation of web-based, self-administered, graphical food frequency questionnaire. J Acad Nutr Diet 2014;114(4):613-621 [FREE Full text] [doi: 10.1016/j.jand.2013.11.017] [Medline: 24462267]

XSL•FO
RenderX

26. Henriksen HB, Carlsen MH, Paur I, Berntsen S, Bøhn SK, Skjetne AJ, et al. Relative validity of a short food frequency questionnaire assessing adherence to the Norwegian dietary guidelines among colorectal cancer patients. Food Nutr Res 2018;62:1306 [FREE Full text] [doi: 10.29219/fnr.v62.1306] [Medline: 29545734]
27. Henriksen HB, Knudsen MD, Carlsen MH, Hjartåker A, Blomhoff R. A short digital food frequency questionnaire (DIGIKOST-FFQ) assessing dietary intake and other lifestyle factors among Norwegians: qualitative evaluation with focus group interviews and usability testing. JMIR Form Res 2022;6(11):e35933 [FREE Full text] [doi: 10.2196/35933] [Medline: 36346647]
28. Usability testing. University Center for Information Technology (USIT). 2017. URL: https://www.usit.uio.no/om/organisasjon/bnt/web/ux/blogg/2017/brukertest.html [accessed 2024-03-27]
29. Norwegian Digitalisation Agency. 2023. URL: https://www.digdir.no/digdir/about-norwegian-digitalisation-agency/887 [accessed 2024-03-27]
30. Difi. 2023. URL: https://eid.difi.no/nb/id-porten [accessed 2024-03-27]
31. Henriksen HB, Berntsen S, Paur I, Zucknick M, Skjetne AJ, Bøhn SK, et al. Validation of two short questionnaires assessing physical activity in colorectal cancer patients. BMC Sports Sci Med Rehabil 2018;10:8-20 [FREE Full text] [doi: 10.1186/s13102-018-0096-2] [Medline: 29854408]
32. Henriksen HB, Berg HB, Andersen LF, Weedon-Fekjær H, Blomhoff R. Development of the Norwegian diet index and the Norwegian lifestyle index and evaluation in a national survey. Food Nutr Res 2023;67:-11 [FREE Full text] [doi: 10.29219/fnr.v67.9217] [Medline: 37808205]
33. Nettskjema. 2023. URL: https://nettskjema.no/ [accessed 2024-04-15]
34. Dalane J, Bergvatn TAM, Kielland E, Carlsen MH. Mål, vekt og porsjonsstørrelser for matvarer [Weights, measures and portion sizes for foods]. Universitetet i Oslo Helsedirektoratet. 2015. URL: https://tinyurl.com/4mjn3vj2 [accessed 2024-04-15]
35. Norwegian Directorate of Health. 2021. URL: https://www.helsenorge.no/kosthold-og-ernaring/sma-grep-for-et-sunt-kosthold/dagens-maltider/#tallerkenmodellen [accessed 2024-04-15]
36. Norwegian Institute of Public Health. 2021. URL: https://www.fhi.no/en/ [accessed 2024-04-15]
37. World Health Organization (WHO), Europe. URL: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi?source=post_page [accessed 2024-04-15]
38. Web-form. University of Oslo. 2021. URL: https://www.uio.no/english/services/it/adm-services/nettskjema/ [accessed 2024-04-15]
39. Rimestad AH, Løken EB, Nordbotten A. Den norske matvaretabellen og beregningsdatabasen ved institutt for ernæringsforskning. Nor J Epidemiol 2009;10(1):2000 [FREE Full text] [doi: 10.5324/nje.v10i1.509]
40. Willett WC. Diet and cancer: an evolving picture. JAMA 2005;293(2):233-234 [FREE Full text] [doi: 10.1001/jama.293.2.233] [Medline: 15644551]
41. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 1985;122(1):51-65 [FREE Full text] [doi: 10.1093/oxfordjournals.aje.a114086] [Medline: 4014201]
42. Glasson C, Chapman K, James E. Fruit and vegetables should be targeted separately in health promotion programmes: differences in consumption levels, barriers, knowledge and stages of readiness for change. Public Health Nutr 2011;14(4):694-701 [FREE Full text] [doi: 10.1017/S1368980010001643] [Medline: 20576194]
43. Totland TH, Melnæs BK, Lundberg-Hallén N, Helland-Kigen KM, Lund-Blix NA, Myhre JB, et al. Norkost 3: en Landsomfattende Kostholdsundersøkelse Blant Menn og Kvinner i Norge i Alderen 18-70 år, 2010-11. Oslo: Helsedirektoratet; 2012.
44. Bland M. An Introduction to Medical Statistics. Oxford: Oxford University Press; 2000.
45. Hankin JH, Wilkens LR, Kolonel LN, Yoshizawa CN. Validation of a quantitative diet history method in Hawaii. Am J Epidemiol 1991;133(6):616-628 [FREE Full text] [doi: 10.1093/oxfordjournals.aje.a115934] [Medline: 2006649]
46. Garden L, Clark H, Whybrow S, Stubbs RJ. Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific? Eur J Clin Nutr 2018;72(7):1026-1034 [FREE Full text] [doi: 10.1038/s41430-018-0199-6] [Medline: 29789710]
47. Stubbs RJ, O'Reilly LM, Whybrow S, Fuller Z, Johnstone AM, Livingstone MBE, et al. Measuring the difference between actual and reported food intakes in the context of energy balance under laboratory conditions. Br J Nutr 2014;111(11):2032-2043 [FREE Full text] [doi: 10.1017/S0007114514000154] [Medline: 24635904]
48. Warren JM, Ekelund U, Besson H, Mezzani A, Geladas N, Vanhees L, Experts Panel. Assessment of physical activity—a review of methodologies with reference to epidemiological research: a report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. Eur J Cardiovasc Prev Rehabil 2010;17(2):127-139 [FREE Full text] [doi: 10.1097/HJR.0b013e32832ed875] [Medline: 20215971]
49. Quinlan C, Rattray B, Pryor D, Northey JM, Anstey KJ, Butterworth P, et al. The accuracy of self-reported physical activity questionnaires varies with sex and body mass index. PLoS One 2021;16(8):e0256008 [FREE Full text] [doi: 10.1371/journal.pone.0256008] [Medline: 34379676]

XSL•FO
RenderX

50. Cleland JA. The qualitative orientation in medical education research. Korean J Med Educ 2017;29(2):61-71 [FREE Full text] [doi: 10.3946/kjme.2017.53] [Medline: 28597869]
51. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56-70 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
52. Darko EM, Kleib M, Olson J. Social media use for research participant recruitment: integrative literature review. J Med Internet Res 2022;24(8):e38015-e38037 [FREE Full text] [doi: 10.2196/38015] [Medline: 35925655]
53. Gelinas L, Pierce R, Winkler S, Cohen IG, Lynch HF, Bierer BE. Using social media as a research recruitment tool: ethical issues and recommendations. Am J Bioeth 2017;17(3):3-14 [FREE Full text] [doi: 10.1080/15265161.2016.1276644] [Medline: 28207365]
54. Educational attainment of the population. Statistics Norway. 2020. URL: https://www.ssb.no/en/utdanning/utdanningsniva/statistikk/befolkningens-utdanningsniva [accessed 2024-02-05]
55. Carlsen MH, Lillegaard ITL, Karlsen A, Blomhoff R, Drevon CA, Andersen LF. Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records. Nutr J 2010 Sep 15;9:37-46 [FREE Full text] [doi: 10.1186/1475-2891-9-37] [Medline: 20843361]
56. McClung HL, Ptomey LT, Shook RP, Aggarwal A, Gorczyca AM, Sazonov ES, et al. Dietary intake and physical activity assessment: current tools, techniques, and technologies for use in adult populations. Am J Prev Med 2018 Oct 01;55(4):e93-e104 [FREE Full text] [doi: 10.1016/j.amepre.2018.06.011] [Medline: 30241622]

Abbreviations

FBDG: food-based dietary guidelines
FFQ: food frequency questionnaire
SWA: SenseWear Armband Mini
WR: weighed food record

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 07.10.23; peer-reviewed by D Brassard; comments to author 19.12.23; revised version received 12.02.24; accepted 09.03.24; published 30.04.24.

Please cite as:
Henriksen HB, Knudsen MD, Hjartåker A, Blomhoff R, Carlsen MH
Digital Food Frequency Questionnaire Assessing Adherence to the Norwegian Food–Based Dietary Guidelines and Other National Lifestyle Recommendations: Instrument Validation Study
J Med Internet Res 2024;26:e53442
URL: https://www.jmir.org/2024/1/e53442
doi:10.2196/53442
PMID:38687986

©Hege Berg Henriksen, Markus Dines Knudsen, Anette Hjartåker, Rune Blomhoff, Monica Hauger Carlsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Instruments for Measuring Psychological Dimensions in Human-Robot Interaction: Systematic Review of Psychometric Properties

Roberto Vagnetti[1], PhD; Nicola Camp[1], MSc; Matthew Story[2], PhD; Khaoula Ait-Belaid[3], PhD; Suvobrata Mitra[4], PhD; Massimiliano Zecca[3], PhD; Alessandro Di Nuovo[2], PhD; Daniele Magistro[1], PhD

[1]Department of Sport Science, School of Science and Technology, Nottingham Trent University, Nottingham, United Kingdom

[2]Department of Computing & Advanced Wellbeing Research Centre, Sheffield Hallam University, Sheffield, United Kingdom

[3]Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, United Kingdom

[4]Department of Psychology, Nottingham Trent University, Nottingham, United Kingdom

Corresponding Author:
Daniele Magistro, PhD
Department of Sport Science
School of Science and Technology
Nottingham Trent University
College Drive
Clifton
Nottingham, NG11 8NS
United Kingdom
Phone: 44 115 848 3522
Email: daniele.magistro@ntu.ac.uk

Abstract

Background: Numerous user-related psychological dimensions can significantly influence the dynamics between humans and robots. For developers and researchers, it is crucial to have a comprehensive understanding of the psychometric properties of the available instruments used to assess these dimensions as they indicate the reliability and validity of the assessment.

Objective: This study aims to provide a systematic review of the instruments available for assessing the psychological aspects of the relationship between people and social and domestic robots, offering a summary of their psychometric properties and the quality of the evidence.

Methods: A systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines across different databases: Scopus, PubMed, and IEEE Xplore. The search strategy encompassed studies meeting the following inclusion criteria: (1) the instrument could assess psychological dimensions related to social and domestic robots, including attitudes, beliefs, opinions, feelings, and perceptions; (2) the study focused on validating the instrument; (3) the study evaluated the psychometric properties of the instrument; (4) the study underwent peer review; and (5) the study was in English. Studies focusing on industrial robots, rescue robots, or robotic arms or those primarily concerned with technology validation or measuring anthropomorphism were excluded. Independent reviewers extracted instrument properties and the methodological quality of their evidence following the Consensus-Based Standards for the Selection of Health Measurement Instruments guidelines.

Results: From 3828 identified records, the search strategy yielded 34 (0.89%) articles that validated and examined the psychometric properties of 27 instruments designed to assess individuals' psychological dimensions in relation to social and domestic robots. These instruments encompass a broad spectrum of psychological dimensions. While most studies predominantly focused on structural validity (24/27, 89%) and internal consistency (26/27, 96%), consideration of other psychometric properties was frequently inconsistent or absent. No instrument evaluated measurement error and responsiveness despite their significance in the clinical context. Most of the instruments (17/27, 63%) were targeted at both adults and older adults (aged ≥18 years). There was a limited number of instruments specifically designed for children, older adults, and health care contexts.

Conclusions: Given the strong interest in assessing psychological dimensions in the human-robot relationship, there is a need to develop new instruments using more rigorous methodologies and consider a broader range of psychometric properties. This is essential to ensure the creation of reliable and valid measures for assessing people's psychological dimensions regarding social

XSL•FO
RenderX

and domestic robots. Among its limitations, this review included instruments applicable to both social and domestic robots while excluding those for other specific types of robots (eg, industrial robots).

(J Med Internet Res 2024;26:e55597) doi:10.2196/55597

KEYWORDS

psychometric; human-robot interaction; psychological dimensions; robot; assessment; systematic review

Introduction

Background

There is a growing interest in the field of social robotics when it comes to creating robots that can cater to people's needs. This is evidenced by the increasing number of publications covering various aspects of robotics [1]. This interest stems from the desire to develop robots that can engage in social interaction with humans, serving as collaborators, companions, tutors, and partners in various applications. Applications of social and domestic robots cover widespread areas; for instance, they have been proposed for educational purposes [2], for mental health and well-being [3], to support older adults in their homes [4-6], or to support different clinical populations such as people with autism spectrum disorder [7] or dementia [8].

While many studies have explored users' opinions and requirements to design and develop this technology to meet their needs in a participatory design framework [9-13], a major challenge for the success of social robots is the fact that their mere presence in everyday life does not automatically increase their chances of being accepted or users' willingness to interact with them [14]. Thus, understanding the perspectives and preferences of people regarding robots represents a crucial point for their development and acceptance [15-17]. How a robot is perceived plays a major role in the human-robot relationship [18]. Existing literature has identified several factors linked to individuals' predispositions toward robots and how robots are used [19]. Peoples' robot acceptance is influenced by attitudes and intentions to use robots [20]. According to the Unified Theory of Acceptance and Use of Technology [21], factors in the intention to use robots include attitude, perceived usefulness, perceived ease of use, enjoyment, trust, and anxiety. However, many other psychological dimensions have been investigated within the human-robot relations, such as beliefs [22], adaptability, control, companionship, sociability [19], attractiveness [23], social presence [24], intentionality [25], and expectations [26]. Thus, numerous user-related psychological dimensions can significantly influence the dynamics of the human-robot relationship. Systematic reviews focusing on different dimensions related to human-robot interactions with social robots reportedly indicate that most of the assessments are made using self-report measurements, raising concerns about their suitability [27-29].

For developers and researchers, it is crucial to have a comprehensive understanding of the psychometric properties of the available instruments. To make a reasoned decision regarding the use of instruments in research, it is crucial to possess an understanding of instrument properties and make comparisons between them [30]. Indeed, psychometric properties encompass attributes of an instrument that serve as indicators of its reliability and validity [31]. They help ascertain whether the measure accurately assesses what it is meant to assess and consistently gauges the intended dimension. In this context, systematic reviews of instrument psychometric properties can assist practitioners and researchers in choosing the most suitable measurement instrument tailored to their specific needs [32]. These reviews are valuable because they consider both the instrument psychometric properties and the methodological quality of the studies conducted to assess them [33]. This knowledge is essential for making informed decisions and effectively evaluating the performance and impact of robots in various applications.

Objectives

The aim of this research was to conduct a systematic review of the instruments documented in the literature for assessing individuals' psychological dimensions in relation to social and domestic robots, such as attitudes, beliefs, perceptions, opinions, and emotions. In this review, *instrument* refers to a specific tool used for data collection and measurement, such as questionnaires, scales, and interviews. This review assessed both the instrument psychometric properties and the quality of evidence linked to each property with a view to (1) provide practitioners and researchers with a comprehensive guide to the available instruments and their psychometric properties, enabling them to make informed choices based on their specific requirements; and (2) establish indications for the future development and validation of such instruments.

Methods

Search Strategy and Eligibility Criteria

A systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [34]. The search was conducted from June 2023 to July 2023 in the following computerized databases: Scopus, PubMed, and IEEE Xplore. This was done as searching at least 2 databases is recommended for the best coverage of the topic and to decrease chances of inappropriate conclusions [35]. The search strategy aimed to find literature related to the validation and assessment of the psychometric properties of instruments designed to evaluate individuals' psychological dimensions in relation to robots. To accomplish this, the following index terms—"robot*," "social," "home," "domestic," "questionnaire," "survey," "assessment," "measure*," "psychom*," "valid," and "reliab*"—were used. The search strategy is provided in Multimedia Appendix 1.

The search strategy aimed to incorporate instruments suitable for use with social and domestic robots. Therefore, those specifically designed for other types of robots were excluded. The search strategy targeted studies meeting the following

XSL•FO
RenderX

inclusion criteria: (1) the instrument could assess psychological dimensions related to social and domestic robots, including attitudes, beliefs, opinions, feelings, and perceptions; (2) the study focused on validating the instrument; (3) the study evaluated the psychometric properties of the instrument; (4) the study underwent peer review; and (5) the study was in English. Given the focus of our research, studies centered on industrial robots, rescue robots, or robotic arms or those primarily validating technology or measuring anthropomorphism were excluded. In total, 3 members of the research group independently assessed the eligibility of the articles after establishing the criteria with the research team. Initially, the titles and abstracts of the articles resulting from the search were screened based on the established criteria. Those that passed the screening were then evaluated through full-text reading. At the end of each step (title and abstract screening and full-text screening), interrater agreement among the 3 reviewers was evaluated, indicating good agreement (Fleiss κ=0.83 and 0.92, respectively), and any disagreements were resolved through discussion. The systematic review and protocol were not registered with any relevant database.

Data Extraction

From the included studies, the following data were extracted for each instrument: the name of the scale; references identified during the systematic review process; the total number of items; a description of the type of items; a description of the construct measured; a description of the subscales, if any; the number of items in each subscale; the administration of the instrument; the target population of the instrument; and the characteristics of the population used to validate the instrument, including nationality, sample sizes, and age (mean, SD, and range). A total of 2 reviewers extracted these data independently, and any disagreements were resolved through consensus with a third reviewer.

Assessment of the Instrument Properties and Methodological Quality

The Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines were adapted to evaluate instrument properties and the methodological quality of the evidence obtained from the identified studies [36-38]. The instrument properties defined and considered by the COSMIN guidelines include content validity (which assesses item relevance, comprehensiveness, and comprehensibility), structural validity, internal consistency, cross-cultural validity, measurement invariance, reliability, measurement error, criterion validity, construct validity, and responsiveness.

The overall rating of each psychometric property per instrument could be sufficient (+), insufficient (−), indeterminate (?), or inconsistent (±) depending on the scores obtained across all the studies for that given measure. While "sufficient" or "insufficient" clearly indicate whether the criteria were met, the procedure considered studies as "indeterminate" when they addressed the relevant aspect but failed to provide sufficient information to determine whether the criteria had been met [36,38]. Furthermore, the "inconsistent" category encompassed a combination of both "sufficient" and "insufficient" results. We chose not to resolve inconsistent results but, instead, to provide this process to individuals interested in using the reviewed instruments, considering this study a valuable summary of the instrument properties available to date. In the context of content validity evaluations, "insufficient" was assigned to each subcategory (relevance, comprehensiveness, and comprehensibility) when these aspects were not evaluated during the development or validation of the measure. In terms of hypothesis testing for construct validity, and considering the multiple dimensions assessed using the identified measures, each study was independently evaluated based on the following suggested generic hypothesis [36]: when the instruments measure related but dissimilar constructs, correlations should fall within the range of 0.30 to 0.50, and when they measure similar constructs, the correlations should be ≥0.50. We considered group differences when hypotheses were clearly stated, supported by the literature, and used specifically to assess instrument properties. An important aspect to consider is that, due to the various constructs that the instrument could assess and the relatively recent development of these scales, it is challenging to establish a reliable gold standard. Consequently, this study did not assess criterion validity among the indicators used to validate the instruments.

The methodological quality of each instrument property in each study was assessed as "high," "moderate," "low," or "very low" following the COSMIN guidelines [36-38]. Subsequently, the overall quality of the body of evidence for each psychometric property could be downgraded based on 3 factors: risk of bias, inconsistency of findings (less relevant for content validity), and imprecision (low sample sizes). It is worth noting that "indirectness" was not evaluated because the review lacked a defined target population.

In total, 2 independent raters extracted data from each record included in the research process and assessed the risk of bias for each psychometric property in each study. Any disagreements that arose were resolved through consensus with a third reviewer.

Results

Overview

The search strategy resulted in a total of 3828 articles. After the removal of 14.52% (556/3828) of duplicates, a further 83.49% (3196/3828) of the articles were excluded during title and abstract screening. Then, of the remaining 76 articles, 42 (55%) were excluded during the full-text evaluation for not meeting the research criteria.

The overall search strategy resulted in the inclusion of a total of 34 articles evaluating 27 measures aimed at assessing people's attitudes toward social robots. A summary of the research process is provided in Figure 1. The information extracted for each measure is reported in Table 1, whereas ratings and quality of evidence are reported in Table 2. In the following sections, we discuss the results per instrument, grouping them according to the included population.

XSL•FO
RenderX

Figure 1. Flowchart of the search and study selection.

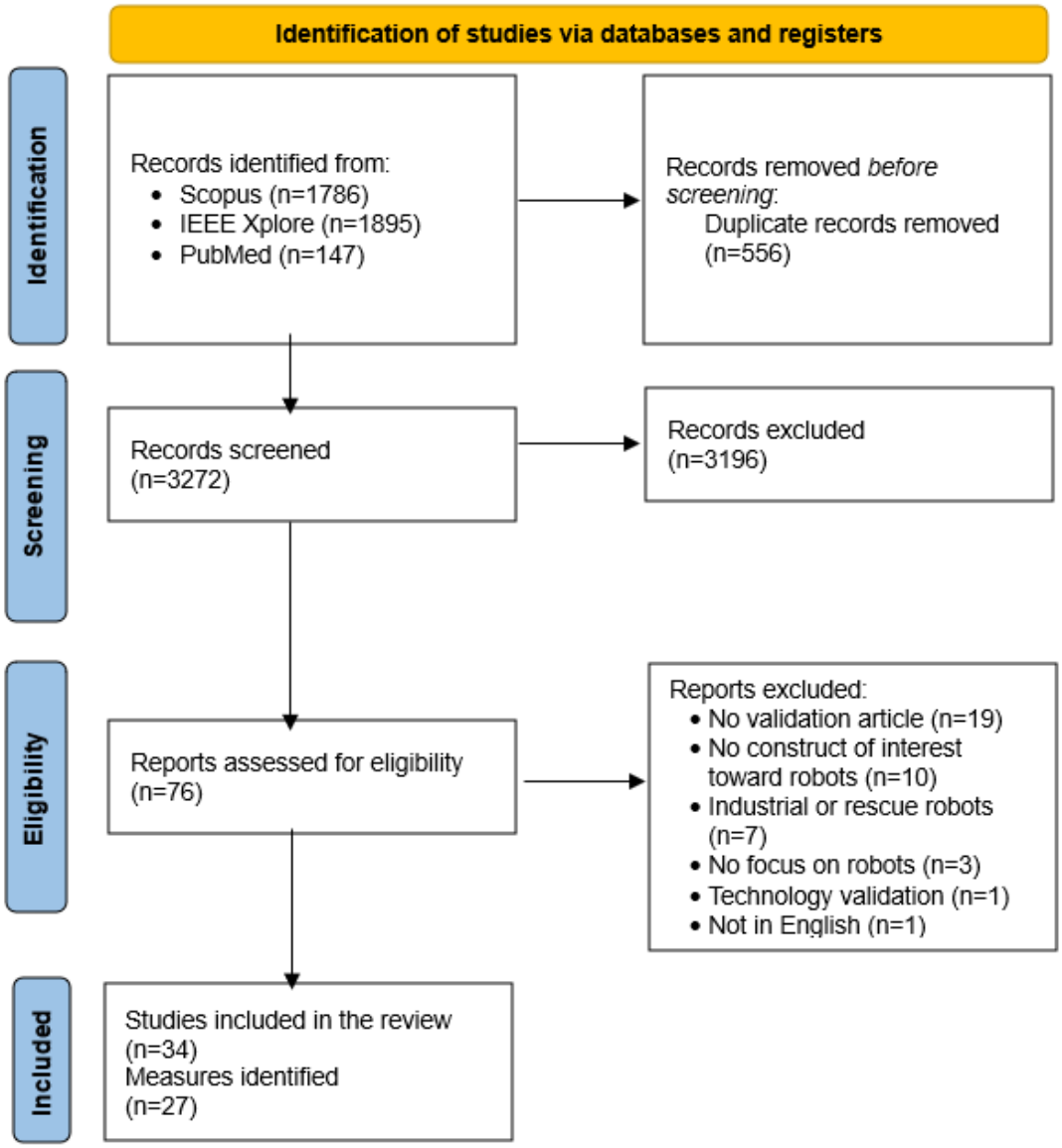

XSL•FO
RenderX

Table 1. Summary of the data extracted for the instruments identified through the systematic review.

Instrument name	Study	Total items	Type of items	Construct or constructs	Subscales (number of items)	Type of administration	Target population	Validation population nationality and number of participants	Age of the validation population (y)
COIRS[a]	Robert and van den Bergh [39]	12	Scale from 0 to 4	Assess openness to robot interaction	Intrinsic interest in interacting with a robot (3); openness to socioemotional interactions with a robot (5); openness to utilitarian interaction with a robot (4)	Self-report	Children	American	Range 8-11
Robot gratification questionnaire[b]	de Jong et al [40]	26	5-point Likert scale	Assess gratification (sought and obtained)	Hedonic gratification—sought (3); hedonic gratification—obtained (3); informative gratification—sought (3); informative gratification—obtained (3); social gratification—obtained (3); social gratification—sought (3); experiential gratification—sought (4); experiential gratification—obtained (4)	Observational	Children	Dutch—24 children (15 male; 9 female)	Mean 9.31 (SD 1.15; range 7-11)
Robot social presence scale[b]	Chen et al [41]	17	5-point Likert scale	Assess robot social presence	Perceived presence (4); interaction behavior perception (4); interactive expression and information understanding (4); perceived emotional interdependence (4); attention allocation (4); emotional understanding and expressiveness (4)	Self-report	General population	Chinese—494 (174 male; 320 female)	Range 18-60
REI[c]	Bevilacqua et al [42]	41	5-point Likert scale	Assess the acceptability of SARs[d] in the older adult population	Perceived robot personality (11); human-robot interaction (10); perceived benefit (6); easiness of use (6); perceived usefulness (7)	Self-report	Older adults	Not reported—21 (13 male; 8 female)	Mean 62.9 (SD 3.9)
UNRAQ[e]	Tobis et al [43,44]	34	5-point Likert scale	Assess older persons' needs and requirements regarding the properties and functions of a robot	Interaction with the robot and technical issues (10); assistive role of the robot (13); social aspects of using the robot (6); ethical issues (5)	Self-report	General population	Not reported—720 (179 male; 541 female)	Mean 52.0 (SD 37.0; range 19-91)
A revision of the TSES-R[f]	Zhang et al [45]	20	5-point Likert scale	Parental expectations regarding robots for health care	Capabilities dimension (5); fictional view dimension (5); social or emotional dimension (4); playful distraction dimension (3); assistive role dimension (3)	Self-report	Parents	Not reported	Not reported

XSL•FO
RenderX

Instrument name	Study	Total items	Type of items	Construct or constructs	Subscales (number of items)	Type of administration	Target population	Validation population nationality and number of participants	Age of the validation population (y)
GAToRS[g]	Koverola et al [46]	20	7-point Likert scale	People's attitudes toward robots	Personal-level positive attitude (5); personal-level negative attitude (5); societal-level positive attitude (5); societal-level negative attitude (5)	Self-report	General population	Finnish—477 (192 male; 283 female)	Mean 40.23 (SD 13.51); all participants were aged >18 years
SPRS[h]	Mandl et al [47]	18	Semantic differential scale	Measure social perception of robots	Anthropomorphism (6); morality or sociability (6); activity or cooperation (4)	Self-report	General population	German, Austrian, Swiss, and from other countries—1032 (538 male; 480 female; 14 nonbinary)	Mean 33.83 (SD 12.66)
SSRIT[i]	Chi et al [48]	50	5-point Likert scale	Assess trust in interactions with AI[j] social robots in service delivery	Familiarity (4); robot use self-efficacy (5); social influence (4); technology attachment (3); trust stance in technology (3); anthropomorphism (7); robot performance (9); effort expectancy (4); perceived service risk (5); robot service fit (3); facilitating robot use condition (3)	Self-report	Consumers	Ethnicities were reported—sample 1: 452 (38.9% male; 60.8% female; 0.2% other); sample 2: 362 (38.7% male; 61% female; 0.3% other)	Range >18->65
ATTUSR-C[k]	Chen et al [49]	15	5-point Likert scale	Assess ATTUSR-C questionnaire for use with Taiwanese health personnel	Unidimensional	Self-report	Health personnel	Taiwanese—95 (95% female)	Mean 44.5 (SD 11.9; range 25-63)
Intentional Acceptance of Social Robots	de Jong et al [50]	4	5-point bar scale	Assess children's intention to use social robots	Unidimensional	Self-report	Children	Dutch—87 (39 male; 48 female)	Mean 9.17 (SD 0.85; range 7-11)
ERAS[l]	Sisman et al [51]	17	5-point Likert scale	Measure the attitudes of school students toward the use of humanoid robots in educational settings	Engagement (5); enjoyment (4); anxiety (4); intention (4)	Self-report	Children	Not reported—232 (128 male; 104 female)	Range 10-13
SE-HRI[m]	der Pütten and Bock [52]	18	6-point Likert scale	Measure people's perceived self-efficacy in dealing with robots	Unidimensional	Self-report	General population	German—450 (288 female; 4 unknown); American—209 (104 male; 105 female)	German: mean 25.15 (SD 6.66; range 18-59); American: mean 26.48 (SD 9.11; range 16-69)

XSL•FO
RenderX

Instrument name	Study	Total items	Type of items	Construct or constructs	Subscales (number of items)	Type of administration	Target population	Validation population nationality and number of participants	Age of the validation population (y)
SE-HRI short version	der Pütten and Bock [52]	10	6-point Likert scale	Measure people's perceived self-efficacy in dealing with robots	Unidimensional	Self-report	General population	English-speaking and German-speaking sample—196 (101 male; 95 female)	Mean 36.91 (SD 13.97; range 18-82)
MCRS[n]	Nomura et al [53]	21	7-point Likert scale	Measure moral concern for robots	Basic moral concern (12); concern for psychological harm (9)	Self-report	General population	Japanese—group 1: 121 (66 male; 55 female); group 2: 200 (100 male; 100 female)	Mean 20.1 (SD 1.6) for group 1; range 20s-60s for group 2
RoSAS[o]	Carpinella et al [54] and Oliveira et al [55]	18 (note: 10 for Portuguese version)	9-point Likert scale	Measure social perception of robots	Warmth (6); competence (6); discomfort (6)	Self-report	General population	Not reported—210 (105 male; 104 female; 1 not identified); Portuguese—185 (45% male; 55% female)	Not reported—Portuguese sample: mean 23.40 (SD 5.21; range 18-35)
HEXACO-60 for HRI[b,p]	Siri et al [56]	60	5-point Likert scale	Evaluate how people perceive the personality traits of robots	Empathy, altruism, or sociability; integrity; dependability; self-confidence	Self-report	General population	Italian—133 (not reported)	Mean 34.46 (SD 14.17; range 19-65)
Sense of safety and security for robots in elder care[b]	Akalin et al [57]	12	Semantic difference scales	Measure sense of safety and security for robots in eldercare	Sense of safety (6); sense of security (6)	Self-report	General population	Not reported—100 (47 male; 53 female)	Mean 35.48 (SD 10.58; range 14-62)
PERNOD[q]	Kamide et al [58]	33	7-point Likert scale	Evaluate humanoid robots	Familiarity (12); utility (7); motion (4); controllability (5); toughness (5)	Self-report	General population	Japanese—380 university students (140 male; 239 female; 1 unknown)	Mean 20.31 (SD 2.89)
Multidimensional Robot Attitude Scale	Ninomiya et al [59]	49	7-point Likert scale	Assess attitudes toward domestic robots	Familiarity (6); interest (7); negative attitude (5); self-efficacy (4); appearance (7); utility (5); cost (3); variety (3)	Self-report	Adults	Japanese—175 (77.8% male); Taiwanese—130 (46.9% male); Chinese—40.5% male	Japanese: mean 22.3 (SD 1.9); Chinese: mean 23.6 (SD 1.6); Taiwanese: mean 24.2 (SD 5.0)
RUSH[r]	Turja et al [60]	6	5-point Likert scale	Measure robot use self-efficacy in health care work	Unidimensional	Self-report	Health care workers	Finnish—3 samples: 200 home care workers (93.5% female), 1889 nurses (89.8% female),and 1554 nurses and physiotherapists (95% female)	Sample 1: mean 43.2 (SD 11.8; range 19-65); sample 2: mean 45.5 (SD 12.1; range 17-68); sample 3: mean 47.5 (SD 10.4; range 19-70)

XSL•FO
RenderX

Instrument name	Study	Total items	Type of items	Construct or constructs	Subscales (number of items)	Type of administration	Target population	Validation population nationality and number of participants	Age of the validation population (y)
RAS[s]	Nomura et al [61] and Cai et al [62]	11	6-point Likert scale	Measuring the anxiety that prevents individuals from interacting with robots that have functions of communication in daily life	Anxiety toward communication capability of robots (3); anxiety toward behavioral characteristics of robots (4); anxiety toward discourse with robots (4)	Self-report	General population	Japanese: 400 university students (197 male; 199 female; 4 unknown); Chinese: sample 1 composed of 305 adults (138 male; 167 female) and sample 2 composed of 740 adults (319 male; 421 female)	Japanese: mean 21.4; Chinese: sample 1 range 18-≥60 and sample 2 range 18-60
RERS[t]	Nomura and Kanda [63]	18	7-point Likert scale	Measure people's expectations for rapport	Expectation as a conversation partner (11); expectation for togetherness (7)	Self-report	General population	Not reported—2s0 university students (not reported)	Not reported
Child-robot relationship formation[b]	Straten et al [64]	13	5-point bar scales	Assess child-robot relationship formation	Closeness (5); trust (4); perceived social support (5)	Self-reported	Children in middle childhood	Dutch—87 children (39 male; 48 female)	Mean 9.17 (SD 0.85; range 7-11)
Almere model	Heerink et al [21] and He et al [65]	41 (30 for the Mandarin version)	5-point Likert scale	Acceptance of assistive social agents by older adults	Anxiety (4); attitude toward the assistive social agent (3); facilitating conditions (2); intention to use (3); perceived adaptiveness (3); perceived enjoyment (5); perceived ease of use (5); perceived sociability (4); perceived usefulness (3); social influence (2); social presence (5); trust (2)	Self-reported	Older adults	Dutch—experiment 1: 40 older adults (18 male; 22 female), experiment 2: 88 participants (28 male; 60 female), experiment 3: 30 older adults (8 male; 22 female), and experiment 4: 30 older adults (16 male; 14 female); Chinese—317 (55.5% female)	Dutch—experiment 1: range 65-89, experiment 2: NA[u], experiment 3: range 65-94, and experiment 4: range 65-89; Chinese—mean 70.3 (SD 7.5)
Frankenstein Syndrome Questionnaire	Nomura et al [66,67]	30	7-point Likert scale	Measure acceptance of humanoid robots, including expectations and anxieties regarding this technology in the public	General anxiety toward humanoid robots (13); apprehension toward social risks of humanoid robots (5); trustworthiness for developers of humanoid robots (4); expectation for humanoid robots in daily life (5)	Self-reported	General population	Japanese (the questionnaire is also available in English)—1000 (500 male; 500 female)	Range 20s-60s

XSL•FO
RenderX

Instrument name	Study	Total items	Type of items	Construct or constructs	Subscales (number of items)	Type of administration	Target population	Validation population nationality and number of participants	Age of the validation population (y)
NARS[v]	Nomura et al [68], Piçarra et al [69], Pochwatko et al [70], Syrdal et al [71], and Xia and LeTendre [72]	14 (the Portuguese and Polish versions have 12 items, and the English version has 11 items)	5-point Likert scale (7-point Likert scale for the Portuguese and Polish versions)	Measure humans' negative attitudes toward robots	Negative attitudes toward situations of interaction with robots (6); negative attitudes toward the social influence of robots (5); negative attitudes toward emotions in interaction with robots (3; note: the Portuguese and Polish versions have 2 dimensions, NARHT[w] and NATIR[x], and the English version has 3 dimensions measuring different constructs)	Self-reported	General population	Japanese—240 university students (146 male; 92 female; 2 unknown); American—54 undergraduate students (13 male; 41 female); Portuguese—4 studies with a total sample of 997 (401 male; 598 female; 3 not reported); Polish—213 (80 male; 91 female; 42 not reported); English—28 university students and staff (14 male; 14 female)	Japanese—mean 22.0 (SD not reported); American—range 18-25; Portuguese—range 18-71; Polish—mean 29.36 (SD 10.15); English—range 18-55

[a]COIRS: Children's Openness to Interacting With a Robot Scale.
[b]The name of the instrument was not provided in the original article.
[c]REI: Robot-Era Inventory.
[d]SAR: socially assistive robot.
[e]UNRAQ: Users' Needs, Requirements, and Abilities Questionnaire.
[f]TSES-R: Technology-Specific Expectation Scale–R.
[g]GAToRS: General Attitudes Toward Robots Scale.
[h]SPRS: Social Perception of Robots Scale.
[i]SSRIT: Social Service Robot Interaction Trust.
[j]AI: artificial intelligence.
[k]ATTUSR-C: Chinese version of the Attitudes Toward The Use of Social Robots.
[l]ERAS: Educational Robot Attitude Scale.
[m]SE-HRI: Self-Efficacy in Human-Robot Interaction Scale.
[n]MCRS: Moral Concern for Robots Scale.
[o]RoSAS: Robotic Social Attributes Scale.
[p]HRI: human-robot interaction.
[q]PERNOD: Perception to Humanoid scale.
[r]RUSH: Robot Use Self-Efficacy in Healthcare Work.
[s]RAS: Robot Anxiety Scale.
[t]RERS: Rapport-Expectation With a Robot Scale.
[u]Not available.
[v]NARS: Negative Attitudes Toward Robots Scale.
[w]NARHT: negative attitudes toward robots with human traits.
[x]NATIR: negative attitudes toward interactions with robots.

XSL•FO
RenderX

Table 2. Results obtained from the overall rating (OR) quality of evidence (QoE)[a].

Instrument	Content validity								Structural validity		Internal consistency		Cross-cultural validity		Measurement invariance		Reliability		Construct validity	
	Overall		Relevance		Comprehensiveness		Comprehensibility		OR	QoE	OR	QoE	OR	QoE	OR	QoE	OR	QoE	OR	QoE
	OR	QoE	OR	QoE	OR	QoE	OR	QoE												
COIRS[b]	±[c]	Low	–[d]	Low	+[e]	Low	+	Low	+	High	+	High	—[f]	—	—	—	—	—	?[g]	Low
Robot gratification questionnaire[h]	±	Low	+	Low	−	Low	−	Low	?	Low	±	Low	—	—	—	—	—	—	±	Very low
Robot social presence scale[h]	+	Moderate	+	Moderate	+	Moderate	+	Moderate	+	Moderate	+	High	—	—	—	—	—	—	—	—
REI[i]	−	Very low	±	Low	−	Very low	−	Very low	?	Very low	−	Very low	—	—	—	—	—	—	?	Very low
UNRAQ[j]	−	Very low	−	Very low	−	Very low	−	Very low	—	—	±	High	—	—	—	—	+	Moderate	—	—
TSES-R[k]	−	Very low	−	Very low	−	Very low	−	Very low	?	Very low	+	Moderate	—	—	—	—	—	—	—	—
GAToRS[l]	±	Very low	+	Low	−	Very low	−	Very low	+	High	+	High	—	—	—	—	—	—	±	High
SPRS[m]	−	Very low	−	Low	−	Very low	−	Very low	−	High	±	High	—	—	—	—	—	—	±	High
SSRIT[n]	+	Moderate	+	Moderate	+	Moderate	+	Low	+	High	+	High	—	—	—	—	—	—	+	High
ATTUSR-C[o]	±	Very low	+	Very low	−	Very low	+	Very low	?	Moderate	+	Moderate	—	—	—	—	—	—	—	—
Intentional Acceptance of Social Robots	−	Very low	−	Very low	−	Very low	−	Low	+	Moderate	+	Moderate	—	—	+	Low	—	—	±	Very low
ERAS[p]	−	Moderate	−	Moderate	−	Moderate	+	Moderate	?	High	+	High	—	—	—	—	—	—	—	—
SE-HRI[q]	±	Very low	+	Very low	−	Very low	+	Very low	+	High	+	High	—	—	—	—	—	—	+	High
SE-HRI (short version)	—	—	—	—	—	—	—	—	—	—	+	High	—	—	—	—	—	—	+	High
MCRS[r]	−	Very low	−	Very low	−	Very low	−	Very low	?	Moderate	+	High	—	—	—	—	—	—	?	Very low
RoSAS[s]	−	Very low	−	Very low	−	Very low	+	Low	+	High	+	High	—	—	—	—	±	Moderate	±	Moderate
HEXACO-60 for HRI[d,t]	−	Very low	±	Very low	−	Very low	−	Very low	?	Very low	—	—	—	—	—	—	—	—	—	—
Sense of safety and security for robots in elder care[h]	−	Very low	−	Very low	−	Very low	−	Very low	?	Very low	+	Moderate	—	—	—	—	—	—	—	—
PERNOD[u]	−	Very low	−	Low	−	Very low	−	Very low	?	High	+	High	—	—	—	—	—	—	—	—

XSL•FO
RenderX

Instrument	Content validity								Structural validity		Internal consistency		Cross-cultural validity		Measurement invariance		Reliability		Construct validity	
	Overall		Relevance		Comprehensiveness		Comprehensibility		OR	QoE	OR	QoE	OR	QoE	OR	QoE	OR	QoE	OR	QoE
	OR	QoE	OR	QoE	OR	QoE	OR	QoE												
Multidimensional Robot Attitude Scale	−	Very low	−	Very low	−	Low	−	Very low	—	—	±	High	—	—	—	—	—	—	—	—
RUSH[v]	−	Very low	−	Very low	−	Very low	−	Very low	?	High	+	High	—	—	—	—	—	—	−	Very low
RAS[w]	+	Moderate	+	Moderate	+	Moderate	?	Moderate	+	High	+	High	—	—	—	—	—	—	±	High
RERS[x]	−	Very low	±	Very low	−	Very low	−	Very low	?	Very low	+	Moderate	—	—	—	—	—	—	?	Low
Child-robot relationship formation[h]	±	Very low	±	Very low	−	Very low	+	Low	+	Moderate	+	Moderate	—	—	—	—	—	—	±	Very low
Almere model	+	Moderate	+	Moderate	+	Moderate	?	Moderate	+	High	+	High	—	—	—	—	+	High	?	Low
Frankenstein Syndrome Questionnaire	−	Moderate	−	Moderate	−	Moderate	−	Moderate	?	Moderate	+	High	−	Very low	—	—	—	—	—	—
NARS[y]	−	Low	−	Low	−	Low	−	Very low	−	High	+	High	−	Very low	—	—	±	Low	±	High

[a]Measurement error and responsiveness are absent from the table because no article assessed these properties, and criterion validity is not reported in accordance with the explanation given in the *Methods* section.

[b]COIRS: Children's Openness to Interacting With a Robot Scale.

[c]Inconsistent rating.

[d]Insufficient rating.

[e]Sufficient rating.

[f]Blank cells represent psychometric properties that were not evaluated for that instrument.

[g]Indeterminate rating.

[h]The name of the instrument was not reported in the original article.

[i]REI: Robot-Era Inventory.

[j]UNRAQ: Users' Needs, Requirements, and Abilities Questionnaire.

[k]TSES-R: Technology-Specific Expectation Scale–R.

[l]GAToRS: General Attitudes Toward Robots Scale.

[m]SPRS: Social Perception of Robots Scale.

[n]SSRIT: Social Service Robot Interaction Trust.

[o]ATTUSR-C: Chinese version of the Attitudes Toward the Use of Social Robots.

[p]ERAS: Educational Robot Attitude Scale.

[q]SE-HRI: Self-Efficacy in Human-Robot Interaction Scale.

[r]MCRS: Moral Concern for Robots Scale.

[s]RoSAS: Robotic Social Attributes Scale.

[t]HRI: human-robot interaction.

[u]PERNOD: Perception to Humanoid scale.

[v]RUSH: Robot Use Self-Efficacy in Healthcare Work.

[w]RAS: Robot Anxiety Scale.

[x]RERS: Rapport-Expectation With a Robot Scale.

[y]NARS: Negative Attitudes Toward Robots Scale.

XSL•FO
RenderX

Instruments to Assess Children's Psychological Dimensions Toward Robots

The *Children's Openness to Interacting With a Robot Scale* [39] measures openness to new experiences and psychological boundaries related to robot interactions. The scale was developed through focus groups with parents, teachers, and researchers and underwent cognitive pretesting with colleagues and researchers. During the validation, exploratory factor analysis (EFA) revealed a 3D structure with good internal consistency (Cronbach α ranging from 0.72 to 0.78) and sufficient structural validity (root mean square error of approximation [RMSEA]=0.07; comparative fit index [CFI]=0.93; root mean square residual=0.07) for the 3 dimensions. Construct validity was assessed by correlating the average Children's Openness to Interacting With a Robot Scale score with those of other scales. However, correlations were not performed with the subscales, making the construct validity for each subscale unclear. A comparison by age and gender found no significant differences, although the purpose of the comparison was not reported.

The questionnaire developed by de Jong et al [40] aimed to assess children's uses of and gratifications regarding social robots based on the literature on children's media gratifications. After a brief interaction with a social robot, 88 Dutch children were interviewed. Through coding of their responses to an open-ended question, categories of gratifications were identified, and a questionnaire was developed to measure 4 types of gratification. The items were derived from previous questionnaires and children's answers. The gratification types were subsequently categorized into sought and obtained, although the theoretical rationale for this choice was not provided. During the validation, 2 subscales did not reach sufficient internal consistency. The EFA results did not provide information about the goodness of the 8D solution. Some of the subscales did not provide sufficient evidence for the expected hypothesis tested by the authors (Pearson correlations ranging from 0.12 to 0.78).

The *Intentional Acceptance of Social Robots* [50] is a unidimensional instrument developed to assess children's intentional acceptance of social robots, defined as children's intention to use a social robot repeatedly or for a long period in their daily life. The researchers reviewed existing measures and focused on the scale by Heerink et al [21]. They adjusted and refined items by referencing specific activities and adapting the language for children, through discussions, and with suggestions from primary school teachers. The items were also translated into Dutch. Pilot-testing with 4 children led to further adjustments. The confirmatory factor analysis (CFA) revealed a good fit of the data (N=87, χ^2_2=3.6, *P*=.16; CFI=0.97; standardized root mean square residual=0.04). Measurement invariance was assessed between boys and girls, showing sufficient results. Internal consistency showed sufficient results for the overall sample (range 0.72-0.85). According to hypothesis testing, the scale showed enough correlation with the enjoyment measure (*r*=0.49) but low correlation with other measures (ie, social presence [*r*=0.24] and social anxiety [*r*=−0.20]).

The *Educational Robot Attitude Scale* [51] was developed through a process involving the creation of an item pool based on existing literature. The scale was reviewed by experts for content and face validity. A pilot test with 20 schoolchildren was conducted to assess item comprehension. The scale showed a 4D solution according to EFA (fitting indexes were not reported). The reliability of the scale was satisfactory (Cronbach α ranged from 0.81 to 0.85).

Straten et al [64] developed a measure to assess child-robot relationships using 3 self-report scales of closeness, trust, and perceived social support in which constructs were derived from theories of interpersonal relationships. The researchers developed the scales by reviewing existing measures and refining item content translated into Dutch. Comprehensibility was assessed through teachers and pilot studies. The measure's validation demonstrated a good model fit based on CFA results (N=87, χ^2_{62}=62.3, *P*=.47; CFI=0.999; standardized root mean square error [SRMR]=0.052). Hypothesis testing with concurrently measured variables, which were significantly shortened, yielded mixed results.

Instruments to Assess Psychological Dimensions of Adults (Aged ≥18 Years) Toward Robots

Chen et al [41] proposed a 6D questionnaire to assess robots' social presence. Researchers retrieved papers related to social presence and identified questions for a human-robot interaction scale divided into theoretical dimensions following expert evaluation and translation. A total of 3 experts in artificial intelligence, psychology, and sociology respectively assessed the proposed definition and model, tested face validity, and reviewed content and discriminant validity for each dimension of the scale. Then, 5 respondents experienced in using social robots were invited for structured interviews to assess the clarity, precision, repetition, conflict, and understandability of the questionnaire. Validation results indicated good fit from EFA (chi-square–to–df ratio=2.160; RMSEA=0.048; Tucker-Lewis index [TLI]=0.928; normed fit index=0.939; Adjusted Goodness of Fit Index [AGFI]=0.926; SRMR=0.052; CFI=0.966; goodness-of-fit index [GFI]=0.950), and the Cronbach α values were of >0.70.

The *Users' Needs, Requirements, and Abilities Questionnaire* [43,44] was developed through a process that involved a literature review and collaboration with the ENRICHME project partners. It is an instrument that can be used to collect data on the use of social robots in the care of older people. The validation sample consisted of 720 older adult participants, 125 of whom repeated the assessment 2 weeks apart. Evaluation of psychometric properties indicated good Cronbach α values for each dimension (all >0.70) and test-retest reliability for each subscale measured using the intraclass correlation coefficient (range 0.81-0.93).

The *General Attitudes Toward Robots Scale* [46] was developed to assess attitude as a predisposition to respond favorably or unfavorably to objects in the world and makes a distinction between personal and societal levels of attitudes toward robots, differentiating them between positive and negative. In the pilot study, the authors only reported that the measure was developed

XSL•FO
RenderX

partly based on other instruments. The 4D factor was considered adequate as it fell between the suggested factors of different evaluation methods. Only 2 of the dimensions had a Cronbach α value of >0.70. The authors developed a revised version of the questionnaire by conducting a pilot study, extracting items from other instruments, and collecting new items through open questions posted in science fiction forums and robotics-oriented Facebook groups. The authors further refined these items through various EFAs. The final version of the questionnaire consisted of 20 items along with 4 criterion items. A CFA indicated good fit (χ^2_{164}=430.0, *P*<.001; CFI=0.91; TLI=0.896; RMSEA=0.058, 90% CI 0.052-0.064; SRMR=0.057), and the Cronbach α values for each subscale were of >0.70. Correlations with the Negative Attitudes Toward Robots Scale (NARS) indicated mixed results (range 0.2-0.8); however, the authors did not report a specific hypothesis and, given that the General Attitudes Toward Robots Scale measures attitude toward robots, we would have expected strong correlations of >0.50.

The *Social Perception of Robots Scale* [47] was developed as a short scale for measuring social perceptions of robots that comprises sociability, competence, morality, and anthropomorphism that can be applied to different robots in diverse research settings. Although a definition for each scale was provided, the authors did not describe a theoretical background for the social perception dimension and for its subcomponents. The authors composed items based on 3 different instruments to address the 3 main dimensions of social perception. The EFA results indicated a 3D factor (anthropomorphism, morality or sociability, and activity or cooperation), and a subsequent CFA did not indicate good fit (χ^2_{115}=508.1, *P*<.001; RMSEA=0.101; CFI=0.796; TLI=0.759; SRMR=0.096). Regarding internal consistency, the third dimension resulted in a low index (Cronbach α=0.64), whereas the first and second dimensions had sufficient indexes (Cronbach α=0.82 and 0.85, respectively). Regarding hypothesis testing, only some of the expected correlations were confirmed (*r* range 0.08-0.96), indicating mixed results.

The *Self-Efficacy in Human-Robot Interaction Scale* [52] was developed to create a German and an English version of a valid and reliable instrument for measuring people's perceived self-efficacy in dealing with robots. The first version of the Self-Efficacy in Human-Robot Interaction Scale consisted of items that were either adapted from different questionnaires or theoretically generated. An EFA indicated a 2D solution (namely, self-efficacy and loss of control), which showed good internal consistency (Cronbach α=0.945 and 0.864, respectively). A CFA was conducted with the German version of the measure and a different sample; however, it did not reach sufficient structural validity (chi-square–to–df ratio of 5.21 and poor values for the other fit indexes: RMSEA=0.097; CFI=0.84; SRMR=0.055), and a subsequent analysis with reduced items indicated a 1-factor solution and a good model fit (chi-square–to–df ratio=2.98; RMSEA=0.066; CFI=0.95; SRMR=0.029). This result was replicated for the English version. The comprehensibility of the German version was assessed with 6 older adults. Hypothesis testing performed with correlations indicated sufficient values (*r*>0.30); however, we should note that, with only 1 scale, a general self-efficacy measure was close to this value (*r*=0.271 for the German sample and *r*=0.298 for the English sample). The authors also proposed a short version based on results from the EFA. A CFA indicated good fit of the short version for both the German and English samples. In addition, hypothesis testing indicated correlations of >0.30; however, this was true also for 1 scale that the authors used as a discriminant measure.

Nomura et al [53] developed the *Moral Concern for Robots Scale*. A definition or a theoretical background of moral concern was not clearly provided. The Moral Concern for Robots Scale was obtained by adopting items from existing questionnaires. In addition, they created items based on human moral treatment and scenes of possible robot abuse. Through a questionnaire-based survey, the collected data were analyzed using factor analysis, resulting in a 2-factor structure. No fitting statistics were reported. Each dimension indicated good internal consistency (Cronbach α=0.912 and 0.876). Most of the correlations conducted by the authors for construct validity were of <0.30, and 2 dimensions indicated high correlations with the developed measure, namely, "Mental state" and "Social partner"; however, they were not assessed using validated measures, so these results could not properly be considered as evidence of construct validity.

The *HEXACO-60 for Robots* [56] is based on the HEXACO model of personality and proposes that individuals are characterized by 6 domains. The authors adapted the items of the HEXACO-60 original questionnaire [73] addressing "a robot" as the subject of each original item. Even though the construct was clearly described and had a theoretical background, a representative population was not involved in the elicitation of relevant items; thus, relevance was considered indeterminate. The authors performed an EFA that indicated a 4D solution; fitting statistics were not reported.

Akalin et al [57] developed a scale to evaluate the sense of safety and security of robots for older adult care. The authors developed the items after videos of different types of robot interactions were shown to participants; 3 items were based on the Godspeed Questionnaire Series [74]. Definitions of safety and security used to construct items were not clearly reported. The authors calculated the Cronbach α for the 2 dimensions and for each video scenario presented to the participants. All the Cronbach α values reported were of >0.70, indicating good internal consistency of the scales. Factor analysis was performed to identify the most important item associated with the 2 dimensions.

The *Robotic Social Attributes Scale* (RoSAS) developed by Carpinella et al [54] assesses warmth, competence, and discomfort perceived in robots. While the first 2 dimensions were drawn from social psychology, they lacked a clear definition, making it challenging to assess the content of items related to these dimensions. The development of this scale involved 4 studies. In the first study, an EFA was conducted on the Goodspeed Questionnaire Series [74], resulting in 3 factors reflecting anthropomorphism, perceived intelligence, and likeability. In the second study, participants were presented with the Godspeed items, a list of attributes from the stereotype

XSL•FO
RenderX

content model, and the Bem Sex-Role Inventory [75,76]. Participants indicated whether each item was associated with robots. EFA reduced the number of items and suggested 3 dimensions: warmth (Cronbach α=0.91), competence (Cronbach α=0.84), and discomfort (Cronbach α=0.82). The third study trialed the developed RoSAS, presenting participants with familiar and unfamiliar animals and human linguistic categories to demonstrate that the dimension of "discomfort" emerges when individuals are evaluating robots. In the fourth study, the questionnaire was validated by comparing different types of robots to assess whether participants' perceptions varied based on the scale. However, references to support the hypotheses were not provided. In a separate study, Oliveira et al [55] performed a Portuguese translation of the RoSAS and assessed the comprehension of its items. A CFA suggested that the 3 dimensions were a good solution (CFI=0.98; RMSEA=0.05; SRMR=0.06), leading to a reduction in the number of items. Correlations with other measures for construct validity and reliability assessments yielded conflicting results.

The *Rapport-Expectation With a Robot Scale* [63] was designed to measure people's expectations regarding rapport with robots. To create this scale, students watched science fiction movie clips featuring robots and were asked about their feelings toward interacting with robots, distinguishing between fictional and real robots. Items were developed based on participant responses and from previous research. Subsequently, an EFA was conducted with a small sample, revealing a 2D solution (Cronbach α=0.919 and 0.848). Unfortunately, no fit indexes were reported. To assess construct validity, the same participants were used, with the assumption that there would be variations in their responses based on the different video clips they had viewed. Differences in scores were indeed found, but it is difficult to interpret these results as there was no provided evidence to support the formulated hypothesis. Subsequently, an experimental task was carried out to evaluate predictive validity. However, the results were inconsistent as only 1 of the 2 hypotheses was confirmed.

The *Robot Anxiety Scale* (RAS) [61] was developed to measure anxiety that inhibits people from interacting with robots. The items for this scale were generated through a pilot survey, and content validity was assessed. A subsequent EFA revealed a 3D factor solution. Following this, a CFA indicated a good fit (GFI=0.949; AGFI=0.917; RMSEA=0.066) for each scale (Cronbach α=0.840, 0.844, and 0.796). Construct validity was evaluated by comparing the RAS with 2 other anxiety measures, all showing correlations of <0.30. Cai et al [62] translated the scale from Japanese to Chinese and assessed its comprehensibility and item content validity. Their study included a CFA that confirmed the RAS's structural validity (chi square–to–df ratio=3.26; SRMR=0.02; CFI=0.99; GFI=0.96; TLI=0.98; RMSEA=0.06), and correlations for construct validity indicated good construct validity (absolute values of *r* ranged from 0.42 to 0.81). Overall, the correlations between the 2 studies yielded mixed results.

The *Frankenstein Syndrome Questionnaire* (FSQ) developed by Nomura et al [67] is a questionnaire to gauge people's acceptance of humanoid robots. To develop this questionnaire, a survey was conducted to gather opinions, attitudes, and feelings regarding humanoid robots from students in both Japan and the United Kingdom. A group of experts later reviewed the extracted items for content validity. The questionnaire was administered on the web, and a factor analysis revealed a 4D solution (Cronbach α range 0.693-0.909). GFIs were not reported. In a subsequent study, the cross-cultural validity of the FSQ was examined [66], revealing differences in responses between Japanese and UK populations.

Nomura et al [68] developed the NARS to assess the predispositions in behavior or reactions toward robots. They initially gathered opinions through a pilot survey extracting 13 sentences and obtained an additional 20 sentences from 2 other measures. The content validity was confirmed through expert discussions. During the validation, an EFA revealed a 4-factor structure, and a CFA indicated a good fit (GFI=0.900; AGFI=0.856; RMSEA=0.080), with Cronbach α coefficients ranging from 0.648 to 0.782. Construct validity was assessed using Pearson correlation with a measure of anxiety, but all coefficients were of <0.30. Test-retest reliability, assessed using Pearson correlation, showed mixed results: 2 subscales had good reliability (*r*=0.706 and *r*=0.740), but the "Negative attitudes toward emotions in interaction with robots" subscale did not (*r*=0.538). Syrdal et al [71] assessed the NARS in the English population after translating it. They removed 3 items, although the Cronbach α was not reported for each subscale. They conducted a principal component analysis to assess item loadings on each dimension. Construct validity was assessed using 12 personality traits [77], yielding mixed results. A Portuguese validation of the measure was conducted by Piçarra et al [69], resulting in a 2D solution and a good model fit (CFI=0.93; TLI=0.90; RMSEA=0.065). Each subscale showed good internal consistency (Cronbach α=0.73 and 0.75), although construct validity was not evaluated using other standardized measures. The Polish version of the measure, as conducted by Pochwatko et al [70], resulted in a 2D solution with 2 items removed. Both subscales exhibited good internal consistency (Cronbach α=0.84 and 0.79), but the study did not provide sufficient information to assess construct validity. Xia and LeTendre [72] conducted a cross-cultural validation of the questionnaire, recruiting American and international-background students. A CFA confirmed the 3-factor structure (CFI=0.93; TLI=0.91; RMSEA=0.08; SRMR=0.08), and internal consistency was also confirmed (Cronbach α ranged from 0.773 to 0.818). The study revealed differences between the 2 groups of students. It is important to note that the structural validity of the NARS yielded conflicting results, with some studies suggesting a 3D solution whereas others proposed a 2D solution.

The *Perception to Humanoid* scale developed by Kamide et al [58] was designed to assess people's perspectives when evaluating humanoid robots. University students were required to describe their impressions after viewing a video recording of a humanoid robot. The responses were categorized into groups and adapted into items. An EFA indicated a 5D solution, with each dimension demonstrating good internal reliability (Cronbach α ranging from 0.79 to 0.86). However, no CFA or GFIs were reported.

To assess attitudes toward domestic robots, Ninomiya et al [59] developed the *Multidimensional Robot Attitude Scale*. The

XSL•FO
RenderX

authors did not provide a specific definition of "attitude," and they generated scale items based on descriptions provided by study participants. EFA was conducted, revealing a 12D structure. Subsequently, 2 to 7 items were selected for each factor based on their loadings with the aim of ensuring sufficient differentiation among them. The Cronbach α values for most dimensions exceeded 0.70 except for the value for the "control" dimension, which was 0.643.

The *Social Service Robot Interaction Trust* [48] assesses consumers' trust in interactions with artificial intelligence social robots. The scale's items were generated through a literature review process and interviews, subsequently evaluated through a focus group. An EFA revealed an 11-factor solution. The Cronbach α values ranged from 0.82 to 0.94, and a CFA indicated a good model fit (RMSEA=0.03; CFI=0.96; TLI=0.96; SRMR=0.05). Concurrent validity was assessed by comparing the Social Service Robot Interaction Trust with the Interpersonal Trust Scale [78] and the Technology Artifact Scale [79], revealing high correlations (r=0.78 and r=0.84, respectively).

Instruments to Assess Older Adults' Psychological Dimensions Toward Robots

The *Robot-Era Inventory* [42] was designed to measure older adults' acceptance of social robots across 5 dimensions based on the Robot-Era Model proposed by the authors. The inventory items were derived from existing scales found in the literature. A preliminary validation of the questionnaire was conducted. The internal consistency analysis yielded mixed results, with 2 of the proposed subscales showing insufficient Cronbach α values (0.67 and 0.69). Construct validity was assessed by examining the correlations between the Robot-Era Inventory and the Unified Theory of Acceptance and Use of Technology, although a clear hypothesis was not reported. The associations between the dimensions showed mixed results.

The *Almere model* was developed to test the acceptance of assistive social agents by older adults [21]. The questionnaire items were adapted from the UTAUT to fit the context of assistive robot and screen agent technology, specifically addressing older adult users in a care home environment. Additional constructs were considered, and items were adapted from questionnaires in the literature. A path analysis was used to test hypothesized relationships among the dimensions. In an experiment comparing responses to a robot in more social versus less social conditions, differences were found on 4 subscales. The study showed sufficient internal consistency of the instrument. He et al [65] translated the Almere Technology Acceptance Questionnaire into Mandarin Chinese and evaluated its psychometric properties among older adults in China. They performed a content analysis with 6 experts. EFA followed by CFA revealed a 9D solution (chi square–to–df ratio=2.006; RMSEA=0.069; root mean-square residual=0.059; GFI=0.816; incremental fit index=0.913; TLI=0.896; CFI=0.912). Cronbach α coefficients indicated mixed results, ranging from 0.664 to 0.891, indicating varied internal consistency across dimensions. The test-retest reliability coefficient was satisfactory, with an overall value of 0.980 and domain-specific values ranging from 0.918 to 0.986 [65].

Instruments to Assess Psychological Dimensions of Health Care Professionals Toward Robots

The *Robot Use Self-Efficacy in Healthcare Work* [60] is a measure developed and validated with health care workers to assess their self-efficacy in using robots in their work. There was no reported information regarding the items' development. The validation of the measure indicated sufficient internal consistency (Cronbach α=0.90); only factor loadings of the factor analysis performed were reported. Regarding correlations performed for construct validity, results obtained for the 6-item version of the measure were not reported. Instead, the authors performed a correlation analysis for the short version of the measure (3-item version). Most correlations with other measures were insufficient (r<0.30) except for one (r=0.33), which was measured with only 1 item, and its validation was not reported.

The Chinese version of the *Attitudes Toward the Use of Social Robots* (ATTUSR-C) [49] questionnaire is a modified and translated version of the questionnaire proposed by Costescu and David [80]. Although the original version provides a clear definition of "attitude," the study was not aimed at validating the questionnaire, and there was no evidence of concept elicitation or literature search in item generation. In this version, a panel of 5 expert academic nursing professors assessed the content validity of the ATTUSR-C questionnaire, rating item clarity and appropriateness. Items with an item content validity index of <70% were eliminated. In addition, 10 clinical instructors assessed the instrument for face validity by evaluating the clarity of each questionnaire item. This process indicated sufficient evidence for relevance and comprehensibility of the items; however, the professionals were not asked about the comprehensiveness of the items, and thus, it was evaluated as insufficient. During validation, the EFA interpretation led to a 1D solution, with no reported fit indexes. The Cronbach α was sufficient (0.84).

Among the instruments designed for health care, we can highlight the *Technology-Specific Expectation Scale–R* [45], which was developed to assess parents' expectations in health-related robot interactions. The scale consists of items adapted from the work of Alves-Oliveira et al [81], and additional items were created by the authors, organized into 3 dimensions. Principal component analysis was used to determine the item loadings for each dimension; however, fit indexes were not reported. Each subscale demonstrated good internal consistency (Cronbach α=0.869, 0.839, and 0.800). Details regarding the sample used for this analysis were not provided.

Discussion

Principal Findings

The use of social robots has generated substantial research interest, and it is unsurprising that numerous studies have explored the variables that influence the human-robot relationship. This exploration is essential for understanding people's attitudes toward these emerging technological tools. This review aimed to provide both practitioners and researchers with an up-to-date framework of psychometrically validated instruments for assessing the psychological dimensions relevant

XSL•FO
RenderX

to the interaction with social and domestic robots. This systematic literature review identified a total of 27 validated measures across 34 articles. These findings suggest a growing interest in psychological constructs related to understanding human-robot relationships, indicating their increasing importance and relevance. Indeed, as detailed in Table 1, the dimensions assessed using the validated scales encompass different constructs.

Although it indicates validated instruments to assess different dimensions, this review also highlights important limitations in terms of psychometric properties. To enhance the quality and accuracy of the available instruments, these limitations should be considered in future development or revisions of instruments for assessing people's psychological dimensions related to robots. Most of the studies (24/27, 89%) primarily concentrated on assessing the structural validity (12/24, 50% of evidence) and internal consistency (26/27, 96%, of which 18/26, 69% had a high quality of evidence) of the instruments. Construct validity was considered for 63% (17/27) of the instruments (7/17, 41% with a high quality of evidence). Cross-cultural validity was evaluated for only 7% (2/27) of the instruments (both of which exhibited low quality of evidence), and measurement invariance was evaluated for only 4% (1/27) of the instruments (low quality of evidence). Notably, the measurement error and responsiveness aspects were disregarded across all the instruments. Content validity was identified in most of the studies (26/27, 96%); however, none of them exhibited an overall high quality of evidence. Moreover, there was a noticeable scarcity of tools specifically tailored for children, older adults, and health care contexts. This highlights the necessity for the development and validation of instruments encompassing a more comprehensive range of psychometric properties. Such an approach is vital for the advancement of this growing area of research, ensuring that assessments are not only thorough but also tailored to the unique characteristics and needs of diverse populations and contexts.

Regarding content validity, many studies (17/26, 65%) inadequately assessed this property, often demonstrating very low methodological quality and neglecting aspects such as item relevance, comprehensiveness, and comprehensibility. This result aligns with previous findings from other reviews, which have shown that studies often offer unclear definitions of constructs or fail to provide any definition at all [28,29]. Thus, given the interest in validating instruments, the importance of considering this aspect should be further stressed. Content validity refers to how the content of the scale adequately reflects the construct the instrument is intended to measure [82], and it is considered to be the most important measurement property [37]. Relevance and comprehensiveness refer to how well the items are aligned with the construct of interest, ensuring that all key aspects of the construct are thoroughly evaluated [38]. Comprehensibility considers how well items are interpreted, which can have an impact on the quality of responses and measurement accuracy [83]. Therefore, careful consideration of the construct's definition and its theoretical basis should be taken during the development of the instrument to enhance methodological rigor and improve the quality of the assessment.

Measurement error and responsiveness were not addressed in any of the studies identified. Measurement error indicates the amount of error, systematic and random, that could not be attributed to a true change in the construct measured [82]. It could be assessed through minimally important change, which indicates whether a change in the measurement is considered important [84], or the smallest detectable change, which indicates whether the change in score is of sufficient magnitude to have a low probability of being a random error [85]. Responsiveness indicates how the instrument could detect change over time in the measured construct [86], which is considered to reflect longitudinal validity [87]. However, even if they are important, we should also note that these 2 properties place a strong emphasis in the clinical context [87,88]. Thus, we suggest considering these 2 properties with caution and within the context and aim for which the instrument is used.

Most studies (24/27, 89%) primarily focused on structural validity, typically through EFA or CFA, as well as internal consistency as measured using the Cronbach α. However, it is important to note that several studies (12/24, 50%) had "indeterminate" findings on structural validity and did not report goodness-of-fit statistics for their models or provide sufficient information to assess the appropriateness of their structural models. This result expands upon what Naneva et al [29] previously reported. Thus, the results suggest improvement of the structural assessment of the instruments. It could be suggested to report the goodness of fit in exploratory analysis [89] and further conduct confirmatory analyses to evaluate model fit in relation to this psychometric property. Indeed, while EFA is an exploratory approach to determine the appropriate number of factors, CFA requires a strong empirical foundation and is typically used in later phases on empirical and theoretical grounds [90].

Regarding cross-cultural validity, only 7% (2/27) of the instruments, the NARS [72] and the FSQ [66], assessed this aspect. However, in both cases, there was insufficient evidence to demonstrate cross-cultural validity, and the quality of the methodology was low. Despite the challenges associated with considering this property, cross-cultural validity offers valuable instruments for diverse cultures [91]. Its importance is evident from the multinational studies conducted on the topic of human-robot interactions [92,93] and is also highlighted by the diverse nationalities of the validation samples included in this review. This aspect should be given further consideration in the context of psychological measurements for human-robot relations, with particular attention to the methods used.

Similarly, measurement invariance was examined in only 3% (1/34) of the studies [50]. This is particularly concerning as it would indicate that differences between groups evaluated using most of these instruments could be due to group-specific characteristics rather than to true differences in the dimensions assessed by them [94]. Thus, they should be interpreted with some caution.

Reliability, which refers to the proportion of the overall variance in the measure that can be attributed to true differences between individuals [36], or, in other words, how the variability observed between individuals is not influenced by errors [95], was largely overlooked. Only 15% (4/27) of the instruments provided evidence for the assessment of this property: the Users' Needs,

XSL•FO
RenderX

Requirements, and Abilities Questionnaire; the RoSAS; the Almere model; and the NARS [44,55,65,68]. Consequently, many of the identified instruments did not demonstrate reliability, which represents a significant limitation of most available instruments.

When assessing construct validity, a significant proportion of studies (14/17, 82%) used correlations with other instruments. Nevertheless, these correlations often yielded inconsistent results. A problem faced in this evaluation was that many studies (7/14, 50%) did not establish hypotheses regarding expected correlations beforehand. In the validation of these measures, it is recommended to formulate valuable and clear hypotheses that address the construct under investigation.

Most of the instruments (17/27, 63%) reviewed had a target population of young adults to older adults; however, we should note that they did not consider measurement invariance due to age-related differential item functioning, and thus, it could not be established whether certain items could favor individuals from different age groups with different backgrounds or due to specific response formats [96]. In this regard, this aspect should be considered when developing these instruments. Only 7% (2/27) of the instruments, the Robot-Era Inventory [42] and the Almere model [21], were designed specifically for the older adult population. However, they exhibited limited psychometric properties, indicating the need to develop instruments for this specific demographic group.

Only 11% (3/27) of the instruments considered the clinical context. The Robot Use Self-Efficacy in Healthcare Work [60] and ATTUSR-C [49] were designed for health care professionals, whereas the Technology-Specific Expectation Scale–R focused on parents' expectations [45]. However, these 3 instruments only demonstrated sufficient internal consistency, indicating that there is still a need to develop psychometrically valid and reliable instruments in the health care context. This is particularly important given the literature's emphasis on the use of social robots in health care settings [97].

This review indicated that only 19% (5/27) of the instruments in the literature were validated for children. Most of them (3/5, 60%) only demonstrated sufficient structural validity and internal consistency, suggesting that the available measures to assess the psychological dimensions of children toward robots have important limitations. Consequently, there is a need for the development of improved instruments for children.

Finally, it is worth noting that most of the reviewed studies (26/27, 96%) did not effectively use item response theory. While there is some debate regarding the best approach [98], given the conditions of the validation study [99], authors should also take this framework into consideration.

Despite this review indicating a strong interest in developing instruments to assess the psychological facets of the human-robot relationship, it also highlights that only some psychometric properties are systematically considered, whereas other important ones tend to be overlooked. Psychometric properties indicate whether the instrument used is a valid and reliable form to assess the dimension of interest [31]. Poorly or incompletely validated instruments have limited use for specific conditions, populations, and countries [100]. Limitations in these properties may raise concerns regarding the accuracy of reported outcomes in research and in making informed decisions [101].

A significant limitation of the available instruments is the absence of consideration or clear description of context for the robot's use during development and validation. A precise delineation of the use context is an integral aspect of instrument development and content validity evaluation [38], which serves to indicate the relevance of the developed items composing the instrument. This is important as there is preliminary evidence suggesting that the context in which the robot is presented or used may impact the components of human-robot interaction [29]. This information is critical for practitioners and clinicians as it indicates the appropriate use of these instruments for specific purposes and clinical populations.

This suggests a need to develop instruments with a broader range of psychometric properties through studies with higher methodological quality of evidence. The analysis suggests that, in addition to the commonly assessed psychometric properties, particular attention should be paid to content validity, cross-cultural validity, reliability, measurement invariance, and construct validity through rigorous methodologies. Particular attention should be paid to targeted groups and the potential application of the instruments in different contexts. Measurement error and responsiveness remain important properties, and their assessment should be guided by the rationale of the developed instrument. Researchers should consider that these properties have significant weight in a clinical context.

While this study yielded intriguing results, it is important to acknowledge its limitations. The review and analysis in this study primarily focused on questionnaires suitable for assessing social and domestic robots. However, it is crucial to note that questionnaires tailored for other specific types of robots, such as industrial robots, exist and warrant evaluation to offer valuable insights into those domains as well. In addition, there are alternative measures in the existing literature for assessing psychological constructs in the context of human-robot interaction. In this regard, it is worth mentioning the Godspeed Questionnaire Series [74]. The search strategy in this review focused on studies dedicated to the validation of instruments, considering eligible those that addressed this aspect as one of their primary objectives. In the context of the relatively novel field of instrument development for human-robot interactions, it was not feasible to identify gold standards for assessing criterion validity in this review. Nevertheless, it is worth emphasizing that the findings of this study may contribute to the identification of gold standards for other instruments in the future. Indeed, there is significant variability in their use, with some scales being rarely used whereas others are more commonly used (eg, the Almere model) in the literature. The cause of this variability cannot be definitively determined. Notably, the NARS, the Almere model, and the RAS are the oldest scales identified in the systematic literature, potentially contributing to their continued use. This raises the possibility that certain important constructs may be systematically overlooked or the psychometric properties of the instruments

XSL•FO
RenderX

might be disregarded. The primary objective of this review was to offer a comprehensive overview of the instruments available to measure various dimensions and conduct a critical selection based on the currently available psychometric properties of these instruments. As another aspect to consider, individuals may have distinct preferences regarding various physical characteristics of robots, and these preferences are likely influenced by personal factors. However, the extent to which appearance can impact or enhance the human-robot relationship remains a topic that requires more comprehensive exploration. Indeed, determining the ideal form that an agent, such as a robot, should take is particularly challenging [102].

Conclusions

Numerous psychometrically validated instruments exist for assessing various psychological constructs within the realm of human-robot relationships applicable to both social and domestic robots. This review aimed to provide a comprehensive overview of these instruments, offering insights into their psychometric properties. While there is a notable interest in developing and validating such instruments, this review also puts forth guidelines and considerations for both the creation of new instruments and the review of existing ones. This review indicates the necessity to develop and validate new instruments for human-robot interactions encompassing more methodologically rigorous approaches and a broader spectrum of psychometric properties. Researchers should carefully consider the targeted populations and the context of use during development.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council and the National Institute for Health and Care Research (grant EP/W031809/1; Intelligent Multimodal Assessment and Coaching Through Identification of Vulnerabilities in Older People [IMACTIVE]). The funding sponsors had no role in the design of the study; the collection, analyses, or interpretation of the data; the writing of the paper; or the decision to publish the results.

Data Availability

Data sharing is not applicable to this paper as no data sets were generated or analyzed during this study.

Authors' Contributions

RV, MZ, ADN, and DM contributed to conceptualization. Methodology was designed by RV, SM, and DM. Formal analysis was conducted by RV, NC, and DM. Investigation was carried out by RV, NC, MS, KAB, and DM. Resources were provided by MZ, ADN, and DM. Data curation was managed by RV and NC. The original draft of the manuscript was written by RV, NC, MS, and KAB. RV contributed to visualization. Supervision was provided by SM, MZ, ADN, and DM. Project administration was handled by SM and DM. Funding acquisition was secured by SM, MZ, ADN, and DM. All the authors contributed to writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for each database.
[DOCX File , 13 KB - jmir_v26i1e55597_app1.docx]

Multimedia Appendix 2
PRISMA checklist.
[PDF File (Adobe PDF File), 126 KB - jmir_v26i1e55597_app2.pdf]

References

1. Mejia C, Kajikawa Y. Bibliometric analysis of social robotics research: identifying research trends and knowledgebase. Appl Sci 2017 Dec 18;7(12):1316. [doi: 10.3390/app7121316]
2. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F. Social robots for education: a review. Sci Robot 2018 Aug 15;3(21):eaat5954. [doi: 10.1126/scirobotics.aat5954] [Medline: 33141719]
3. Scoglio AA, Reilly ED, Gorman JA, Drebing CE. Use of social robots in mental health and well-being research: systematic review. J Med Internet Res 2019 Jul 24;21(7):e13322 [FREE Full text] [doi: 10.2196/13322] [Medline: 31342908]
4. Camp N, Lewis M, Hunter K, Johnston J, Zecca M, Di Nuovo A, et al. Technology used to recognize activities of daily living in community-dwelling older adults. Int J Environ Res Public Health 2020 Dec 28;18(1):163 [FREE Full text] [doi: 10.3390/ijerph18010163] [Medline: 33379319]

XSL•FO
RenderX

5. Di Nuovo A, Broz F, Wang N, Belpaeme T, Cangelosi A, Jones R, et al. The multi-modal interface of Robot-Era multi-robot services tailored for the elderly. Intel Serv Robotics 2017 Sep 2;11(1):109-126. [doi: 10.1007/s11370-017-0237-6]
6. Robinson H, MacDonald B, Broadbent E. The role of healthcare robots for older people at home: a review. Int J Soc Robotics 2014 Jul 3;6(4):575-591. [doi: 10.1007/s12369-014-0242-2]
7. Di Nuovo A, Conti D, Trubia G, Buono S, Di Nuovo SF. Brief review of robotics in low-functioning autism therapy. In: Proceedings of the Second Symposium on Psychology-Based Technologies. 2020 Presented at: Second Symposium on Psychology-Based Technologies; September 28-29, 2020; Naples, Italy.
8. Yu C, Sommerlad A, Sakure L, Livingston G. Socially assistive robots for people with dementia: systematic review and meta-analysis of feasibility, acceptability and the effect on cognition, neuropsychiatric symptoms and quality of life. Ageing Res Rev 2022 Jun;78:101633 [FREE Full text] [doi: 10.1016/j.arr.2022.101633] [Medline: 35462001]
9. Camp N, Di Nuovo A, Hunter K, Johnston J, Zecca M, Lewis M, et al. Perceptions of socially assistive robots among community-dwelling older adults. In: Proceedings of the 14th International Conference on Social Robotics. 2022 Presented at: ICSR 2022; December 13-16, 2022; Florence, Italy. [doi: 10.1007/978-3-031-24670-8_48]
10. Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Gathering expert opinions for social robots' ethical, legal, and societal concerns: findings from four international workshops. Int J Soc Robotics 2019 Nov 20;12(2):441-458. [doi: 10.1007/s12369-019-00605-z]
11. Ramírez-Duque AA, Aycardi LF, Villa A, Munera M, Bastos T, Belpaeme T, et al. Collaborative and inclusive process with the autism community: a case study in Colombia about social robot design. Int J Soc Robotics 2020 Feb 07;13(2):153-167. [doi: 10.1007/s12369-020-00627-y]
12. Vagnetti R, Camp N, Story M, Ait-Belaid K, Bamforth J, Zecca M, et al. Robot companions and sensors for better living: defining needs to empower low socio-economic older adults at home. In: Proceedings of the 15th International Conference on Social Robotics. 2023 Presented at: ICSR 2023; December 3-7, 2023; Doha, Qatar. [doi: 10.1007/978-981-99-8715-3_31]
13. Winkle K, Senft E, Lemaignan S. LEADOR: a method for end-to-end participatory design of autonomous social robots. Front Robot AI 2021;8:704119 [FREE Full text] [doi: 10.3389/frobt.2021.704119] [Medline: 34926589]
14. Bartneck C, Nomura T, Kanda T, Suzuki T, Kennsuke K. A cross-cultural study on attitudes towards robots. In: Proceedings of the International Conference on Human-Computer Interaction. 2005 Presented at: HCI 2005; June 20-23, 2005; Las Vegas, NV. [doi: 10.13140/RG.2.2.35929.11367]
15. Cavallo F, Esposito R, Limosani R, Manzi A, Bevilacqua R, Felici E, et al. Robotic services acceptance in smart environments with older adults: user satisfaction and acceptability study. J Med Internet Res 2018 Sep 21;20(9):e264 [FREE Full text] [doi: 10.2196/jmir.9460] [Medline: 30249588]
16. Conti D, Di Nuovo S, Buono S, Di Nuovo A. Robots in education and care of children with developmental disabilities: a study on acceptance by experienced and future professionals. Int J Soc Robotics 2016 Jun 1;9(1):51-62. [doi: 10.1007/s12369-016-0359-6]
17. de Graaf MM, Ben Allouch S, van Dijk JA. Why would I use this in my home? A model of domestic social robot acceptance. Hum Comput Interact 2017 Jul 21;34(2):115-173. [doi: 10.1080/07370024.2017.1312406]
18. Caruana N, Moffat R, Miguel-Blanco A, Cross ES. Perceptions of intelligence and sentience shape children's interactions with robot reading companions. Sci Rep 2023 May 05;13(1):7341 [FREE Full text] [doi: 10.1038/s41598-023-32104-7] [Medline: 37147422]
19. de Graaf MM, Ben Allouch S. Exploring influencing variables for the acceptance of social robots. Robot Auton Syst 2013 Dec;61(12):1476-1486. [doi: 10.1016/j.robot.2013.07.007]
20. Shin DH, Choo H. Modeling the acceptance of socially interactive robotics. Interact Stud 2011 Dec 6;12(3):430-460. [doi: 10.1075/is.12.3.04shi]
21. Heerink M, Kröse B, Evers V, Wielinga B. Assessing acceptance of assistive social agent technology by older adults: the Almere model. Int J Soc Robotics 2010 Sep 4;2(4):361-375. [doi: 10.1007/s12369-010-0068-5]
22. Chatzoglou PD, Lazaraki V, Apostolidis SD, Gasteratos AC. Factors affecting acceptance of social robots among prospective users. Int J Soc Robotics 2023 Jul 06. [doi: 10.1007/s12369-023-01024-x]
23. Wilson RK, Eckel CC. Judging a book by its cover: beauty and expectations in the trust game. Polit Res Q 2016 Jul 02;59(2):189-202. [doi: 10.1177/106591290605900202]
24. Chen YC, Yeh SL, Lin W, Yueh HP, Fu LC. The effects of social presence and familiarity on children-robot interactions. Sensors (Basel) 2023 Apr 24;23(9):4231 [FREE Full text] [doi: 10.3390/s23094231] [Medline: 37177434]
25. Ziemke T. Understanding social robots: attribution of intentional agency to artificial and biological bodies. Artif Life 2023 Aug 01;29(3):351-366 [FREE Full text] [doi: 10.1162/artl_a_00404] [Medline: 36943757]
26. Dou X, Wu CF, Wang X, Niu J. User expectations of social robots in different applications: an online user study. In: Proceedings of the HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence. 2020 Presented at: HCII 2020; July 19-24, 2020; Copenhagen, Denmark. [doi: 10.1007/978-3-030-60117-1_5]
27. David D, Thérouanne P, Milhabet I. The acceptability of social robots: a scoping review of the recent literature. Comput Hum Behav 2022 Dec;137:107419. [doi: 10.1016/j.chb.2022.107419]
28. Krägeloh CU, Bharatharaj J, Sasthan Kutty SK, Nirmala PR, Huang L. Questionnaires to measure acceptability of social robots: a critical review. Robotics 2019 Oct 21;8(4):88. [doi: 10.3390/robotics8040088]

XSL•FO
RenderX

29. Naneva S, Sarda Gou M, Webb TL, Prescott TJ. A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. Int J Soc Robotics 2020 Jun 03;12(6):1179-1201. [doi: 10.1007/s12369-020-00659-4]
30. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford, UK: Oxford University Press; 2015.
31. Ginty AT. Psychometric properties. In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine. New York, NY: Springer; 2013.
32. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res 2009 Apr;18(3):313-333 [FREE Full text] [doi: 10.1007/s11136-009-9451-9] [Medline: 19238586]
33. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012 May;21(4):651-657 [FREE Full text] [doi: 10.1007/s11136-011-9960-1] [Medline: 21732199]
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
35. Ewald H, Klerings I, Wagner G, Heise TL, Stratil JM, Lhachimi SK, et al. Searching two or more databases decreased the risk of missing relevant studies: a metaresearch study. J Clin Epidemiol 2022 Sep;149:154-164 [FREE Full text] [doi: 10.1016/j.jclinepi.2022.05.022] [Medline: 35654269]
36. Mokkink LB, Prinsen CA, de Vet HC, Terwee CB. COSMIN methodology for conducting systematic reviews of patient-reported outcome measures (PROMs). In: Maggino F, editor. Encyclopedia of Quality of Life and Well-Being Research. Cham, Switzerland: Springer; 2023.
37. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018 May;27(5):1147-1157 [FREE Full text] [doi: 10.1007/s11136-018-1798-3] [Medline: 29435801]
38. Terwee CB, Prinsen CA, Chiarotto A, de Vet HC, Bouter LM, Alonso J, et al. COSMIN methodology for assessing the content validity of PROMs: user manual. COSMIN. 2018. URL: https://cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf [accessed 2023-10-12]
39. Robert D, van den Bergh V. Children's Openness to Interacting with a Robot Scale (COIRS). In: Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication. 2014 Presented at: ROMAN 2014; August 25-29, 2014; Edinburgh, UK. [doi: 10.1109/roman.2014.6926372]
40. de Jong C, Kühne R, Peter J, Straten CL, Barco A. What do children want from a social robot? Toward gratifications measures for child-robot interaction. In: Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 2019 Presented at: RO-MAN 2019; October 14-18, 2019; New Delhi, India. [doi: 10.1109/ro-man46459.2019.8956319]
41. Chen N, Liu X, Zhai Y, Hu X. Development and validation of a robot social presence measurement dimension scale. Sci Rep 2023 Feb 19;13(1):2911 [FREE Full text] [doi: 10.1038/s41598-023-28817-4] [Medline: 36807328]
42. Bevilacqua R, Di Rosa M, Riccardi GR, Pelliccioni G, Lattanzio F, Felici E, et al. Design and development of a scale for evaluating the acceptance of social robotics for older people: the robot era inventory. Front Neurorobot 2022 Jul 07;16:883106 [FREE Full text] [doi: 10.3389/fnbot.2022.883106] [Medline: 35874107]
43. Tobis S, Cylkowska-Nowak M, Wieczorowska-Tobis K, Pawlaczyk M, Suwalska A. Occupational therapy students' perceptions of the role of robots in the care for older people living in the community. Occup Ther Int 2017 Feb 07;2017:9592405 [FREE Full text] [doi: 10.1155/2017/9592405] [Medline: 29097983]
44. Tobis S, Neumann-Podczaska A, Kropinska S, Suwalska A. UNRAQ-A questionnaire for the use of a social robot in care for older persons. A multi-stakeholder study and psychometric properties. Int J Environ Res Public Health 2021 Jun 07;18(11):6157 [FREE Full text] [doi: 10.3390/ijerph18116157] [Medline: 34200294]
45. Zhang F, Broz F, Ferrari O, Barakova E. TSES-R: an extended scale for measuring parental expectations toward robots for children in healthcare. In: Proceedings of the Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. 2023 Presented at: HRI '23; March 13-16, 2023; Stockholm, Sweden. [doi: 10.1145/3568294.3580084]
46. Koverola M, Kunnari A, Sundvall J, Laakasuo M. General attitudes towards robots scale (GAToRS): a new instrument for social surveys. Int J of Soc Robot 2022 Jun 15;14:1559-1581. [doi: 10.1007/s12369-022-00880-3]
47. Mandl S, Bretschneider M, Asbrock F, Meyer B, Strobel A. The social perception of robots scale (SPRS): developing and testing a scale for successful interaction between humans and robots. In: Proceedings of the 23rd Working Conference on Virtual Enterprises. 2022 Presented at: PRO-VE 2022; September 19-21, 2022; Lisbon, Portugal. [doi: 10.1007/978-3-031-14844-6_26]
48. Chi OH, Jia S, Li Y, Gursoy D. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. Comput Hum Behav 2021 May;118:106700. [doi: 10.1016/j.chb.2021.106700]
49. Chen SC, Jones C, Moyle W. Health professional and workers attitudes towards the use of social robots for older adults in long-term care. Int J Soc Robot 2019 Dec 07;12:1135-1147. [doi: 10.1007/s12369-019-00613-z]

XSL•FO
RenderX

50. de Jong C, Kühne R, Peter J, van Straten CL, Barco A. Intentional acceptance of social robots: development and validation of a self-report measure for children. Int J Hum Comput Stud 2020 Jul;139:102426. [doi: 10.1016/j.ijhcs.2020.102426]
51. Sisman B, Gunay D, Kucuk S. Development and validation of an educational robot attitude scale (ERAS) for secondary school students. Interact Learn Environ 2018 May 13;27(3):377-388. [doi: 10.1080/10494820.2018.1474234]
52. Der Pütten AR, Bock N. Development and validation of the self-efficacy in human-robot-interaction scale (SE-HRI). ACM Trans Hum Robot Interact 2018 Dec 05;7(3):1-30. [doi: 10.1145/3139352]
53. Nomura T, Kanda T, Yamada S. Measurement of moral concern for robots. In: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction. 2019 Presented at: HRI '19; March 11-14, 2019; Daegu, Korea (South). [doi: 10.1109/hri.2019.8673095]
54. Carpinella CM, Wyman AB, Perez MA, Stroessner SJ. The robotic social attributes scale (RoSAS): development and validation. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. 2017 Presented at: HRI '17; March 6-9, 2017; Vienna, Austria. [doi: 10.1145/2909824.3020208]
55. Oliveira R, Arriaga P, Stroessner SJ, Paiva A. Preliminary validation of the European Portuguese version of the robotic social attributes scale (RoSAS). Human Behav Emerg Technol 2021 Dec;3(5):750-758. [doi: 10.1002/hbe2.311]
56. Siri G, Marchesi S, Wykowska A, Chiorri C. The personality of a robot. An adaptation of the HEXACO – 60 as a tool for HRI. In: Proceedings of the 13th International Conference Social Robotics. 2021 Presented at: ICSR 2021; November 10-13, 2021; Singapore, Singapore. [doi: 10.1007/978-3-030-90525-5_62]
57. Akalin N, Kiselev A, Kristoffersson A, Loutfi A. An evaluation tool of the effect of robots in eldercare on the sense of safety and security. In: Proceedings of the 9th International Conference Social Robotics. 2017 Presented at: ICSR 2017; November 22-24, 2017; Tsukuba, Japan. [doi: 10.1007/978-3-319-70022-9_62]
58. Kamide H, Takubo T, Ohara K, Mae Y, Arai T. Impressions of humanoids: the development of a measure for evaluating a humanoid. Int J Soc Robot 2013 May 17;6:33-44. [doi: 10.1007/s12369-013-0187-x]
59. Ninomiya T, Fujita A, Suzuki D, Umemuro H. Development of the multi-dimensional robot attitude scale: constructs of people's attitudes towards domestic robots. In: Proceedings of the 7th International Conference Social Robotics. 2015 Presented at: ICSR 2015; October 26-30, 2015; Paris, France. [doi: 10.1007/978-3-319-25554-5_48]
60. Turja T, Rantanen T, Oksanen A. Robot use self-efficacy in healthcare work (RUSH): development and validation of a new measure. AI Soc 2017 Jul 29;34:137-143. [doi: 10.1007/s00146-017-0751-2]
61. Nomura T, Suzuki T, Kanda T, Kato K. Measurement of anxiety toward robots. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication. 2006 Presented at: ROMAN 2006; September 6-8, 2006; Hatfield, UK. [doi: 10.1109/roman.2006.314462]
62. Cai J, Sun Y, Niu C, Qi W, Fu X. Validity and reliability of the Chinese version of robot anxiety scale in Chinese adults. Int J Hum Comput Interact 2023 Mar 14:1-10. [doi: 10.1080/10447318.2023.2188535]
63. Nomura T, Kanda T. Rapport–expectation with a robot scale. Int J Soc Robot 2015 Mar 13;8:21-30. [doi: 10.1007/s12369-015-0293-z]
64. Straten CL, Kühne R, Peter J, de Jong C, Barco A. Closeness, trust, and perceived social support in child-robot relationship formation. Interact Stud 2020 Jan 24;21(1):57-84. [doi: 10.1075/is.18052.str]
65. He Y, Liu Q, He Q, Li L. Translation, adaptation, and psychometric testing of the Almere technology acceptance questionnaire (ATAQ) among older adults in China. Clin Interv Aging 2022 Sep 12;17:1353-1364 [FREE Full text] [doi: 10.2147/CIA.S367781] [Medline: 36117570]
66. Nomura T, Syrdal DS, Dautenhahn K. Differences on social acceptance of humanoid robots between Japan and the UK. In: Proceedings of the 4th International Symposium on New Frontiers in Human-Robot Interaction. 2015 Presented at: HRI 2015; April 21-22, 2015; Canterbury, UK.
67. Nomura T, Sugimoto K, Syrdal DS, Dautenhahn K. Social acceptance of humanoid robots in Japan: a survey for development of the frankenstein syndorome questionnaire. In: Proceedings of the 12th IEEE-RAS International Conference on Humanoid Robots. 2012 Presented at: HUMANOIDS 2012; November 29-December 1, 2012; Osaka, Japan. [doi: 10.1109/humanoids.2012.6651527]
68. Nomura T, Suzuki T, Kanda T, Kato K. Measurement of negative attitudes toward robots. Interact Stud 2006 Nov 15;7(3):437-454. [doi: 10.1075/is.7.3.14nom]
69. Piçarra N, Giger JC, Pochwatko G, Gonçalves G. Validation of the Portuguese version of the negative attitudes towards robots scale. Eur Rev Appl Psychol 2015 Mar;65(2):93-104. [doi: 10.1016/j.erap.2014.11.002]
70. Pochwatko G, Giger JC, Różańska-Walczuk M, Świdrak J, Kukiełka K, Możaryn J, et al. Polish version of the negative attitude toward robots scale (NARS-PL). J Autom Mob Robot Intell Syst 2015 Jul 27;9(3):65-72. [doi: 10.14313/JAMRIS_2-2015/25]
71. Syrdal DS, Dautenhahn K, Koay KL, Walters ML. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. In: Proceedings of the Artificial Intelligence and the Simulation of Behaviour (AISB). 2009 Presented at: AISB'09; April 6-9, 2009; Edinburgh, UK.
72. Xia Y, LeTendre G. Robots for future classrooms: a cross-cultural validation study of "negative attitudes toward robots scale" in the U.S. Context. Int J Soc Robot 2020 Jun 27;13:703-714. [doi: 10.1007/s12369-020-00669-2]

XSL•FO
RenderX

73. Ashton MC, Lee K. The HEXACO-60: a short measure of the major dimensions of personality. J Pers Assess 2009 Jul;91(4):340-345. [doi: 10.1080/00223890902935878] [Medline: 20017063]
74. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 2008 Nov 20;1:71-81. [doi: 10.1007/s12369-008-0001-3]
75. Bem SL. The measurement of psychological androgyny. J Consult Clin Psychol 1974;42(2):155-162. [doi: 10.1037/h0036215]
76. Fiske ST, Cuddy AJ, Glick P. Universal dimensions of social cognition: warmth and competence. Trends Cogn Sci 2007 Feb;11(2):77-83. [doi: 10.1016/j.tics.2006.11.005] [Medline: 17188552]
77. Eysenck HJ, Eysenck M. Personality and Individual Differences: A Natural Science Approach. New York, NY: Springer; 1985.
78. McKnight DH, Choudhury V, Kacmar C. Developing and validating trust measures for e-commerce: an integrative typology. Inf Syst Res 2002 Sep 01;13(3):334-359. [doi: 10.1287/isre.13.3.334.81]
79. Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: an investigation of its components and measures. ACM Trans Manage Inf Syst 2011 Jul 01;2(2):1-25. [doi: 10.1145/1985347.1985353]
80. Costescu CA, David DO. Attitudes toward using social robots in psychotherapy. Transylvanian J Psychol 2014;15(1):3 (forthcoming) [FREE Full text]
81. Alves-Oliveira P, Ribeiro T, Petisca S, di Tullio E, Melo FS, Paiva A. An empathic robotic tutor for school classrooms: considering expectation and satisfaction of children as end-users. In: Proceedings of the 7th International Conference Social Robotics. 2015 Presented at: ICSR 2015; October 26-30, 2015; Paris, France. [doi: 10.1007/978-3-319-25554-5_3]
82. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010 Jul;63(7):737-745. [doi: 10.1016/j.jclinepi.2010.02.006] [Medline: 20494804]
83. Lenzner T. Effects of survey question comprehensibility on response quality. Field Methods 2012 Sep 07;24(4):409-428. [doi: 10.1177/1525822X12448166]
84. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 2007 Feb;16(1):131-142 [FREE Full text] [doi: 10.1007/s11136-006-9109-9] [Medline: 17033901]
85. Polit DF, Yang FM. Measurement and the Measurement of Change: A Primer for the Health Professions. Alphen aan den Rijn, The Netherlands: Wolters Kluwer; 2016.
86. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989;42(5):403-408. [doi: 10.1016/0895-4356(89)90128-5] [Medline: 2659745]
87. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. BMC Med Res Methodol 2011 Nov 18;11:152; author reply 152 [FREE Full text] [doi: 10.1186/1471-2288-11-152] [Medline: 22099330]
88. Terwee CB, Peipert JD, Chapman R, Lai JS, Terluin B, Cella D, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 2021 Oct;30(10):2729-2754 [FREE Full text] [doi: 10.1007/s11136-021-02925-y] [Medline: 34247326]
89. Fabrigar LR, Wegener DT. Exploratory Factor Analysis. Oxford, UK: Oxford University Press; 2011.
90. Hoyle RH. Handbook of Structural Equation Modeling. New York, NY: Guilford Publications; 2012.
91. Huang WY, Wong SH. Cross-cultural validation. In: Michalos AC, editor. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht, The Netherlands: Springer; 2014.
92. Conti D, Cattani A, Di Nuovo S, Di Nuovo A. Are future psychologists willing to accept and use a humanoid robot in their practice? Italian and English students' perspective. Front Psychol 2019 Sep 18;10:2138 [FREE Full text] [doi: 10.3389/fpsyg.2019.02138] [Medline: 31620061]
93. Papadopoulos I, Wright S, Koulouglioti C, Ali S, Lazzarino R, Martín-García Á, et al. Socially assistive robots in health and social care: acceptance and cultural factors. Results from an exploratory international online survey. Jpn J Nurs Sci 2023 Apr;20(2):e12523. [doi: 10.1111/jjns.12523] [Medline: 36732396]
94. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Med Care 2006 Nov;44(11 Suppl 3):S78-S94 [FREE Full text] [doi: 10.1097/01.mlr.0000245454.12228.8f] [Medline: 17060839]
95. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006 Oct;59(10):1033-1039. [doi: 10.1016/j.jclinepi.2005.10.015] [Medline: 16980142]
96. Ownby RL, Waldrop-Valverde D. Differential item functioning related to age in the reading subtest of the test of functional health literacy in adults. J Aging Res 2013;2013:654589 [FREE Full text] [doi: 10.1155/2013/654589] [Medline: 24089638]
97. González-González CS, Violant-Holz V, Gil-Iranzo RM. Social robots in hospitals: a systematic review. Appl Sci 2021 Jun 27;11(13):5976. [doi: 10.3390/app11135976]
98. Sébille V, Hardouin JB, Le Néel T, Kubis G, Boyer F, Guillemin F, et al. Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients--a simulation study. BMC Med Res Methodol 2010 Mar 25;10:24 [FREE Full text] [doi: 10.1186/1471-2288-10-24] [Medline: 20338031]

XSL•FO
RenderX

99. Jabrayilov R, Emons WH, Sijtsma K. Comparison of classical test theory and item response theory in individual change assessment. Appl Psychol Meas 2016 Nov;40(8):559-572 [FREE Full text] [doi: 10.1177/0146621616664046] [Medline: 29881070]
100. O'Connor A, McGarr O, Cantillon P, McCurtin A, Clifford A. Clinical performance assessment tools in physiotherapy practice education: a systematic review. Physiotherapy 2018 Mar;104(1):46-53. [doi: 10.1016/j.physio.2017.01.005] [Medline: 28844474]
101. Friberg JC. Considerations for test selection: how do validity and reliability impact diagnostic decisions? Child Lang Teach Ther 2010 Apr 08;26(1):77-92. [doi: 10.1177/0265659009349972]
102. Seaborn K, Sekiguchi T, Tokunaga S, Miyake NP, Otake-Matsuura M. Voice over body? Older adults' reactions to robot and voice assistant facilitators of group conversation. Int J Soc Robot 2023;15(2):143-163 [FREE Full text] [doi: 10.1007/s12369-022-00925-7] [Medline: 36406778]

Abbreviations

AGFI: Adjusted Goodness of Fit Index
ATTUSR-C: Chinese version of the Attitudes Toward the Use of Social Robots
CFA: confirmatory factor analysis
CFI: comparative fit index
COSMIN: Consensus-Based Standards for the Selection of Health Measurement Instruments
EFA: exploratory factor analysis
FSQ: Frankenstein Syndrome Questionnaire
GFI: goodness-of-fit index
NARS: Negative Attitudes Toward Robots Scale
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAS: Robot Anxiety Scale
RMSEA: root mean square error of approximation
RoSAS: Robotic Social Attributes Scale
SRMR: standardized root mean square error
TLI: Tucker-Lewis index

Edited by S Ma; submitted 18.12.23; peer-reviewed by E Otaka, S Tokunaga; comments to author 28.02.24; revised version received 03.04.24; accepted 25.04.24; published 05.06.24.

Please cite as:
Vagnetti R, Camp N, Story M, Ait-Belaid K, Mitra S, Zecca M, Di Nuovo A, Magistro D
Instruments for Measuring Psychological Dimensions in Human-Robot Interaction: Systematic Review of Psychometric Properties
J Med Internet Res 2024;26:e55597
URL: https://www.jmir.org/2024/1/e55597
doi:10.2196/55597
PMID:38682783

©Roberto Vagnetti, Nicola Camp, Matthew Story, Khaoula Ait-Belaid, Suvobrata Mitra, Massimiliano Zecca, Alessandro Di Nuovo, Daniele Magistro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cerebral Microbleeds in Different Brain Regions and Their Associations With the Digital Clock-Drawing Test: Secondary Analysis of the Framingham Heart Study

Samia C Akhter-Khan[1,2*], MSc; Qiushan Tao[2,3*], MSc, MD; Ting Fang Alvin Ang[2,4,5,6*], MD; Cody Karjadi[2,4], MSc; Indira Swetha Itchapurapu[3], MPH; David J Libon[7,8], PhD; Michael Alosco[9,10], PhD; Jesse Mez[2,9,10], MSc, MD; Wei Qiao Qiu[2,3,10,11], MD, PhD; Rhoda Au[2,4,5,6,9,10], PhD

[1]Department of Global Health & Social Medicine, King's College London, London, United Kingdom
[2]Framingham Heart Study, Boston University School of Medicine, Boston, MA, United States
[3]Pharmacology, Physiology & Biophysics, Boston University School of Medicine, Boston, MA, United States
[4]Department of Anatomy & Neurobiology, Boston University School of Medicine, Boston, MA, United States
[5]Department of Epidemiology, Boston University School of Public Health, Boston, MA, United States
[6]Slone Epidemiology Center, Boston University School of Medicine, Boston, MA, United States
[7]Department of Geriatrics and Gerontology, Rowan University, Glassboro, NJ, United States
[8]Department of Psychology, New Jersey Institute for Successful Aging, School of Osteopathic Medicine, Rowan University, Glassboro, NJ, United States
[9]Department of Neurology, Boston University School of Medicine, Boston, MA, United States
[10]Alzheimer's Disease and Chronic Traumatic Encephalopathy Centers, Boston University, Boston, MA, United States
[11]Department of Psychiatry, Boston University School of Medicine, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Wei Qiao Qiu, MD, PhD
Framingham Heart Study
Boston University School of Medicine
72 East Concord Street
R-623D
Boston, MA, 02118
United States
Phone: 1 6176384336
Email: wqiu67@bu.edu

Abstract

Background: Cerebral microbleeds (CMB) increase the risk for Alzheimer disease. Current neuroimaging methods that are used to detect CMB are costly and not always accessible.

Objective: This study aimed to explore whether the digital clock-drawing test (DCT) may provide a behavioral indicator of CMB.

Methods: In this study, we analyzed data from participants in the Framingham Heart Study offspring cohort who underwent both brain magnetic resonance imaging scans (Siemens 1.5T, Siemens Healthcare Private Limited; T2*-GRE weighted sequences) for CMB diagnosis and the DCT as a predictor. Additionally, paper-based clock-drawing tests were also collected during the DCT. Individuals with a history of dementia or stroke were excluded. Robust multivariable linear regression models were used to examine the association between DCT facet scores with CMB prevalence, adjusting for relevant covariates. Receiver operating characteristic (ROC) curve analyses were used to evaluate DCT facet scores as predictors of CMB prevalence. Sensitivity analyses were conducted by further including participants with stroke and dementia.

Results: The study sample consisted of 1020 (n=585, 57.35% female) individuals aged 45 years and older (mean 72, SD 7.9 years). Among them, 64 (6.27%) participants exhibited CMB, comprising 46 with lobar-only, 11 with deep-only, and 7 with mixed (lobar+deep) CMB. Individuals with CMB tended to be older and had a higher prevalence of mild cognitive impairment and higher white matter hyperintensities compared to those without CMB (P<.05). While CMB were not associated with the

XSL•FO
RenderX

paper-based clock-drawing test, participants with CMB had a lower overall DCT score (CMB: mean 68, SD 23 vs non-CMB: mean 76, SD 20; *P*=.009) in the univariate comparison. In the robust multiple regression model adjusted for covariates, deep CMB were significantly associated with lower scores on the drawing efficiency (β=–0.65, 95% CI –1.15 to –0.15; *P*=.01) and simple motor (β=–0.86, 95% CI –1.43 to –0.30; *P*=.003) domains of the command DCT. In the ROC curve analysis, DCT facets discriminated between no CMB and the CMB subtypes. The area under the ROC curve was 0.76 (95% CI 0.69-0.83) for lobar CMB, 0.88 (95% CI 0.78-0.98) for deep CMB, and 0.98 (95% CI 0.96-1.00) for mixed CMB, where the area under the ROC curve value nearing 1 indicated an accurate model.

Conclusions: The study indicates a significant association between CMB, especially deep and mixed types, and reduced performance in drawing efficiency and motor skills as assessed by the DCT. This highlights the potential of the DCT for early detection of CMB and their subtypes, providing a reliable alternative for cognitive assessment and making it a valuable tool for primary care screening before neuroimaging referral.

(J Med Internet Res 2024;26:e45780) doi:10.2196/45780

KEYWORDS

cerebral microbleeds; CMB; digital clock-drawing test; DCT; Alzheimer disease; dementia; early screening; Boston Process Approach; cerebral microbleed; neuroimaging; cerebrovascular diseases; aging; MRI; magnetic resonance imaging; clock-drawing test; cognitive function

Introduction

It is widely shown that cerebrovascular diseases increase the risk of Alzheimer disease (AD) and other dementia [1,2]. Cerebral microbleeds (CMB) are one of such cerebrovascular abnormalities, defined as small chronic brain hemorrhages, likely caused by structural abnormalities of the small vessels of the brain [3]. The prevalence of CMB is estimated to be as high as 34% in people with ischemic stroke and 60% in people with nontraumatic intracerebral hemorrhage [4]. CMB have also been associated with cognitive impairment and increased risk for AD development across multiple studies [5,6]. CMB can be divided into 2 subclasses based on their location in the brain, that is, lobar and deep CMB. A recent meta-analysis reported a 75% increased risk of dementia with deep or mixed CMB [7].

In light of population aging demographics, these figures are concerning, and solutions for early detection of emergent disease and AD risk factors at a preclinical stage to prevent the disease's development are urgently needed [8]. Neuroimaging methods, such as brain magnetic resonance imaging (MRI) and computerized axial tomography scans, are valuable tools to detect cerebrovascular pathology. Currently, lobar and deep CMB can only be identified by brain MRI. However, these imaging tools are costly and, in most cases, not accessible in rural areas and low-income contexts. There are over 50 million people estimated to live with dementia and AD worldwide, with the highest increases in lower- and middle-income countries [9]. The worldwide costs of dementia are estimated to amount to over US $1 trillion [10]. To drive down the costs and promote early detection of early brain pathology for AD risk, including CMB, which have subtle clinical symptoms, one promising approach is to explore new, inexpensive technologies such as digital neuropsychological assessments coupled with machine learning analytics to detect and screen for cerebrovascular diseases in clinic before applying neuroimaging tools and neuropsychological assessments [11].

The clock-drawing test (CDT; traditional, paper based) is an easily applicable cognitive test, and prior research has documented that patients with dementia with MRI evidence of vascular disease make more clock-drawing errors than other groups [12-14]. However, the traditional analog, manually scored CDT is limited by its low sensitivity and specificity for different brain diseases, especially at an early stage or where there is subtle brain pathology. Moreover, analog scoring systems are only able to generate a limited number of clock-drawing metrics [15]. Developing efficient tools to measure cognitive change and brain health, especially in the preclinical stage of AD, is necessary. Recently, the CDT has been transformed through the use of a digital ballpoint pen, replacing a conventional ink pen, and coupled with a dot pattern that provides raw, time-stamped data that capture the full performance sequence (digital clock-drawing test; DCT), thus, generating large, detailed data on cognition that cannot be derived using the traditional CDT [11]. Recently, the Framingham Heart Study (FHS) analyzed DCTs from 1833 participants using machine learning to detect cognitive nuances. The DCT showed superiority over existing methods such as the Mini Mental Status Examination in detecting early cognitive impairment and characterizing individuals along the AD trajectory [16]. In addition, the DCT has evidenced its diagnostic value for early screening for mild cognitive impairment (MCI) and AD [17-19]. Further, the DCT was also associated with biomarkers relevant to cognitive impairment and AD, collected by positron emission tomography imaging, for example, amyloid and τ pathology [20].

Given the important role of CMB in the development of AD, the value of operationalizing putative negative effects on behavior caused by CMB using less costly and more accessible diagnostic tools is evident. Using data from the FHS, we predicted that the DCT would be useful in detecting CMB in older adults without dementia. We also examined whether the DCT could detect CMB in different brain regions, for example, lobar and deep CMB.

XSL•FO
RenderX

Methods

Study Sample

FHS is a multigeneration, community-based, prospective cohort study in Framingham, Massachusetts. The FHS offspring cohort (Generation 2) has been longitudinally examined in 9 core examinations, with examinations occurring on average every 4 years between 1971 and 2014. Details about this cohort have been previously described elsewhere [21]. These participants also had serial neuropsychological and MRI scans on average every 5 years between 1999 and 2019 [22]. For this study, FHS Generation 2 participants who were over 45 years old and had brain MRI CMB data (2000-2009) and DCT assessments (2011-2018; n=1072) were included. For primary analyses, participants were excluded if they had dementia (n=23) or had a history of stroke (n=29).

Clock-Drawing Measures: the CDT Versus DCT

The manual for administering and scoring the CDT has been previously described [23]. The CDT score using this analog (hand-scored) method ranged from 0 (no abnormality) to 3 (severe impairment). The DCT was obtained using the digital pen technology from Anoto, and the time-stamped features were processed by Linus Health Inc. Similar to the original paper-pencil CDT, the DCT contains a command and a copy condition [20]. In brief, the DCT contains multiple objective measurements that were derived from approximately 5000 digital clock drawings using machine learning algorithms to precisely evaluate nuances in performance beyond successful task completion [24]. Variables from the command and copy versions were combined from the machine learning calculations into an overall command or copy score ranging from 0 to 100. A similar technique was used for the domain-specific subscores measuring drawing efficiency, information processing, simple motor, and spatial reasoning (see previous publications [20,25] and Table S1 in Multimedia Appendix 1 for details).

Brain MRI

The FHS MRI protocol and CMB diagnosis criteria have been previously described [26]. Briefly, participants were imaged by a Siemens 1.5T MRI, using a 3D T1-weighted coronal spoiled gradient-recalled echo sequence. All images were transferred to and processed by the University of California Davis Medical Center without knowledge of clinical information. The MRI scans were conducted between 2000 and 2009 with gradient recalled echo T2-weighted sequences, allowing for the detection of CMB. Using recently published guidelines [27], CMB were defined as rounded or ovoid hypointense lesions on a T2*-GRE weighted sequence, measuring <10 mm in diameter and surrounded by brain parenchyma over at least half the circumference of the lesion. The presence, number, and location of CMB were determined. Reliability measures for CMB readings have previously been described [28,29]. In line with previous studies [29], CMB location in the brain was classified into subgroups based on assumed pathophysiology (cerebral amyloid angiopathy and hypertensive vasculopathy), and it was classified into 2 locations: deep and lobar, with each participant potentially having 1 or more CMB. All participants were grouped into 4 subgroups based on assumed pathophysiology (cerebral amyloid angiopathy and hypertensive vasculopathy), which included the no CMB (control) group and groups with lobar-only, deep-only, and mixed (deep+lobar) CMB. White matter hyperintensities (WMHs) were segmented with fluid-attenuated inversion recovery and gray matter with T1-weighted images by semiautomated procedures, as previously described [30]. WMHs were adjusted for head size (by dividing WMHs by the total cerebral volume).

Statistical Analysis

Baseline characteristics of study participants were evaluated for the total sample and for CMB status. To compare the CMB status groups, 2-tailed *t* tests were applied for continuous variables, and χ^2 tests or Fisher exact tests were applied for categorical variables. To facilitate standardized comparisons between different scales in the models adjusted for covariates, the digital clock variables were transformed to *z* scores (mean 0, SD 1), for example, the overall DCT score (percentage range 0%-100%) was rescaled to a *z* score after logit transformation; all other DCT scores were also rescaled to *z* scores to standardize the variables that have different value ranges. Robust multivariable linear regression models were applied to assess whether the DCT scores (outcomes variables) were significantly different between different CMB subgroups and no CMB (control) group. All models were adjusted for age, sex, education, WMHs, and the time difference between brain MRI and DCT. A sensitivity analysis was conducted by including subjects with dementia (n=23), or a history of stroke (n=29) that were excluded from the main models. To test whether the DCT facet scores could distinguish different CMB subgroups from no CMB (control group), we also calculated the area under the curve (AUC) of a receiver operating characteristic (ROC) curve analysis based on multinomial classification models that were adjusted for main confounding factors such as age, sex, time between examinations, and MCI. The AUC, ranging from 0.5 to 1, is a key metric for evaluating a classifier's ability to distinguish between positive and negative outcomes. A value nearing 1 indicates a highly accurate model, while an AUC of 0.5 suggests performance equivalent to random chance, indicating no predictive power. The results were shown as beta estimates (β) with 95% CIs. Statistical significance was indicated by a *P* value <.05 (2-tailed tests). All statistical analyses were performed using R (version 4.2.1; R Foundation for Statistical Computing).

In our data analysis, we used several R packages and functions. Specifically, we used the *rlm* function from the *MASS* library for robust linear model fitting [31], the *lmrob* function from the *robustbase* library for MM-type estimator calculation in linear regression [32], and the *sandwich* library for robust SE estimation in nonlinear models [33]. Additionally, the *pROC* library facilitated ROC curve analysis and AUC value calculation [34].

Ethical Considerations

The study was conducted per the Declaration of Helsinki, and ethical approval was provided by Boston University's Institutional Review Board (H-40620). All participants provided informed consent, and data was de-identified.

XSL•FO
RenderX

Results

The 1020 participants from the FHS Generation 2 sample were on average aged 72 (SD 7.9) years, and 57.35% (n=585) of them were female (Table 1). Among them, 64 (6.27%) participants had at least 1 CMB. Participants with CMB were more likely to be older (P<.001), have MCI (P=.02), and have greater WMHs (P<.001). There were no differences between the participants with and without CMB for the traditional, analog-scored CDT, in either the command or copy condition. By contrast, participants with CMB showed a significantly lower overall combined command or copy DCT score (P=.01) than those without CMB (Table 1).

Table 1 shows that participants with CMB showed worse performance on several command and copy DCT domains. Specifically, participants with CMB scored lower on the command spatial reasoning (P=.03), copy drawing efficiency (P=.03), and information processing (P=.01) subscales. We further divided participants with CMB into those who had CMB exclusively in lobar regions (46/64, 72%), those who had CMB only in deep regions (11/64, 17%), and those who had mixed lobar+deep CMB (7/64, 11%) and used the DCT domain score to examine CMB (Table 2). After adjusting for covariates, there were no statistically significant differences in the overall DCT or domain scores between participants with any CMB (P>.05), compared to those without CMB. However, when analyzing the subscales, participants with deep-only CMB had lower scores on the command simple motor subscale than the reference group (β=–0.86, 95% CI –1.43 to –0.30; P=.003). Additionally, participants with mixed (lobar+deep) CMB had lower scores on the command spatial reasoning subscale than the reference group (β=–1.70, 95% CI –2.29 to –1.11; P<.001). In contrast, participants with lobar CMB did not show associations with any domain score under both command and copy conditions of the DCT. The results were similar when including participants with dementia or stroke in the analysis (Table S2 in Multimedia Appendix 1).

Robust multivariate linear regression models were applied with DCT scores (overall score and domains) and CMB subtypes (lobar-only, deep-only, and any deep). The reference group were participants without CMB (n=956). The overall DCT score (percentage) was rescaled to a z score (mean 0, SD 1) after logit transformation and other DCT scores were rescaled to z scores (mean 0, SD 1) after logit transformation. Participants with dementia or stroke were excluded from the analysis (see sensitivity analysis with inclusion in Table S2 in Multimedia Appendix 1). All models were adjusted for age, sex, education, MCI, WMHs, and the time difference between brain MRI and the DCT. The results are shown as standardized β coefficients (β) with 95% CI. P values for statistical significance are indicated.

Next, we used all facets to explore the diagnostic potential of the DCT for CMB, with the subtypes of CMB as the predicted outcomes (lobar, deep, or mixed). In addition to including all facets in the analysis, we investigated which specific facet score had a relationship with the 3 CMB subtypes (ie, lobar-only, deep-only, and mixed; Tables S2 and S3 in Multimedia Appendix 1). Some features from the command condition were mainly associated with deep CMB, whereas the features from the copy condition were more likely to be associated with lobar CMB (Figure 1). For example, in the command condition, the features including vertical spatial placement, horizontal spatial placement, clock face circularity, component placement, max speed, initiation speed, average speed, ink length, and drawing size were associated with deep CMB but not with lobar CMB. Whereas oscillatory motion was lower among participants with lobar CMB, this feature was higher among participants with deep CMB in the command condition. In the copy condition, the features long latency count, stroke count conformity, and noise were positively associated with lobar-only CMB, but not with deep-only CMB. The mixed (lobar+deep) CMB was significantly associated with lower oscillatory motion, lower horizontal spatial placement, higher clock face circularity, and higher component placement scores in the command condition and higher noise score in the copy condition (Figure 1 and Tables S3 and S4 in Multimedia Appendix 1).

In the ROC curve analysis, the results indicated that DCT facet scores alone demonstrated discrimination, as evidenced by the AUCs between individuals without CMB and those with lobar CMB (AUC 0.74, 95% CI 0.66-0.82), deep CMB (AUC 0.80, 95% CI 0.63-0.98), and mixed CMB (AUC 0.89, 95% CI 0.68-1.00; Figure 2A). After adjusting for sex and age, the AUC values were further improved. The AUC (95% CI) values for individuals without CMB and those with lobar CMB, deep CMB, or mixed CMB were 0.77 (0.71-0.84), 0.85 (0.73-0.98), and 0.97 (0.95-0.99), respectively (Figure 2B). Further adjustment for the time difference between brain MRI and DCT measurement and the prevalence of MCI at the brain MRI scan yielded similar AUC values for individuals without CMB and those with lobar CMB (0.76, 95% CI 0.69-0.83), deep CMB (0.88, 95% CI 0.78-0.98), or mixed CMB (0.98, 95% CI 0.96-1.00; Figure 2C). Notably, the highest AUC was observed for the mixed CMB group, followed by a modest increase for the deep CMB group, and the lobar CMB group when compared to the no CMB group, highlighting the enhanced discriminatory ability of DCT facets in identifying different CMB subtypes after adjusting for relevant covariates.

XSL•FO
RenderX

Table 1. Characteristics of the study sample stratified by brain MRI[a] CMB[b] status.

Characteristics	Overall (N=1020)	CMB status (no=0 and yes=1)		
		No (n=956)	Yes (n=64)	*P* values[c]
Age (years), mean (SD)	72 (7.9)	71 (7.8)	76 (8.2)	*<.001* [d]
Female, n (%)	585 (57)	551 (58)	34 (53)	.78
Education, n (%)				.26
High school or less	246 (24)	224 (23)	22 (34)	
Some college	316 (31)	295 (31)	21 (33)	
College or higher	458 (45)	437 (46)	21 (33)	
Mild cognitive impairment, n (%)	43 (4)	36 (4)	7 (11)	*.02*
White matter hyperintensities, mean (SD)	0.0 (1.0)	–0.04 (1.0)	0.49 (1.0)	*<.001*
Traditional clock-drawing test[e]				
Command clock, median (IQR)	0 (0-3)	0 (0-2)	0 (0-3)	.36
Copy clock, median (IQR)	0 (0-2)	0 (0-1)	0 (0-2)	.30
DCT[f], mean (SD)				
Overall DCT score	76 (20)	76 (20)	68 (23)	*.009*
Composite score, mean (SD)				
Domains of command clock				
Drawing efficiency	62 (10)	62 (10)	60 (12)	.41
Simple motor	63 (8.1)	63 (8.0)	63 (9.0)	.89
Information processing	60 (9.9)	60 (9.9)	59 (8.8)	.83
Spatial reasoning	65 (17)	65 (17)	59 (20)	*.03*
Domains of copy clock				
Drawing efficiency	61 (8.4)	61 (8.3)	59 (9.2)	*.03*
Simple motor	61 (6.7)	61 (6.7)	60 (7.4)	.80
Information processing	61 (11)	61 (11)	57 (11)	*.008*
Spatial reasoning	67 (18)	68 (17)	62 (21)	.06

[a]MRI: magnetic resonance imaging.

[b]CMB: cerebral microbleed.

[c]*t* tests (*df*=1018) were applied for continuous variable comparisons between 2 groups (cerebral microbleed versus no cerebral microbleed). χ^2 (female *df*=1, education *df*=2) tests were used for categorical variables, while Fisher exact tests were used in cases of low frequency. *P* values for statistical significance are shown.

[d]Significant *P* values are italicized.

[e]Due to the skewed distributions, we performed a Kruskal-Wallis rank sum test for the traditional clock-drawing test.

[f]DCT: digital clock-drawing test.

XSL•FO
RenderX

Table 2. The association between cerebral microbleed subtypes and the domains of the DCT[a].

DCT	Lobar-only (n=46)		Deep-only (n=11)		Mixed (lobar+deep; n=7)	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Overall DCT score	0.05 (–0.19 to 0.29)	.67	–0.36 (–0.84 to 0.12)	.15	–0.29 (–0.90 to 0.32)	.35
Command domains						
Drawing efficiency	0.20 (–0.05 to 0.45)	.12	–0.65 (–1.15 to –0.15)	*.01* [b]	–0.28 (–0.90 to 0.35)	.39
Simple motor	0.25 (–0.03 to 0.53)	.08	–0.86 (–1.43 to –0.30)	*.003*	0.52 (–0.19 to 1.23)	.15
Information processing	0.04 (–0.22 to 0.31)	.75	–0.13 (–0.66 to 0.39)	.62	–0.29 (–0.95 to 0.37)	.39
Spatial reasoning	–0.06 (–0.30 to 0.18)	.60	–0.12 (–0.59 to 0.35)	.60	–1.70 (–2.29 to –1.11)	*<.001*
Copy domains						
Drawing efficiency	–0.18 (–0.45 to 0.09)	.19	–0.05 (–0.57 to 0.47)	.86	0.05 (–0.60 to 0.71)	.88
Simple motor	0.08 (–0.20 to 0.36)	.57	–0.18 (–0.73 to 0.38)	.54	0.48 (–0.22 to 1.18)	.18
Information processing	–0.22 (–0.49 to 0.05)	.11	–0.14 (–0.67 to 0.40)	.62	–0.09 (–0.76 to 0.58)	.79
Spatial reasoning	–0.01 (–0.29 to 0.26)	.92	–0.22 (–0.77 to 0.32)	.43	–0.16 (–0.85 to 0.52)	.64

[a]DCT: digital clock-drawing test.

[b]Significant *P* values are italicized.

Figure 1. The association between facets of the DCT and the CMB subtypes. Robust multivariate linear regression models were applied with DCT scores (facets for the command and copy condition) and CMB subtypes (lobar CMB, deep CMB, and mixed [lobar+deep] CMB). The reference group were participants without CMB (n=956). The overall DCT score (percentage) was rescaled to a z score (mean 0, SD 1) after logit transformation and other DCT scores were rescaled to z scores (mean 0, SD 1) after log10 transformation. All models were adjusted for age, sex, education, WMHs, MCI, and the time difference between brain MRI and the DCT. Only facets with statistical significance for any type of CMB are shown (as indicated by the red color, with *P*<.05). The results are shown as standardized β coefficients (β) with 95% CI. CMB: cerebral microbleed; DCT: digital clock-drawing test; MCI: mild cognitive impairment; WMH: white matter hyperintensity. A: Command condition; B: Copy condition.

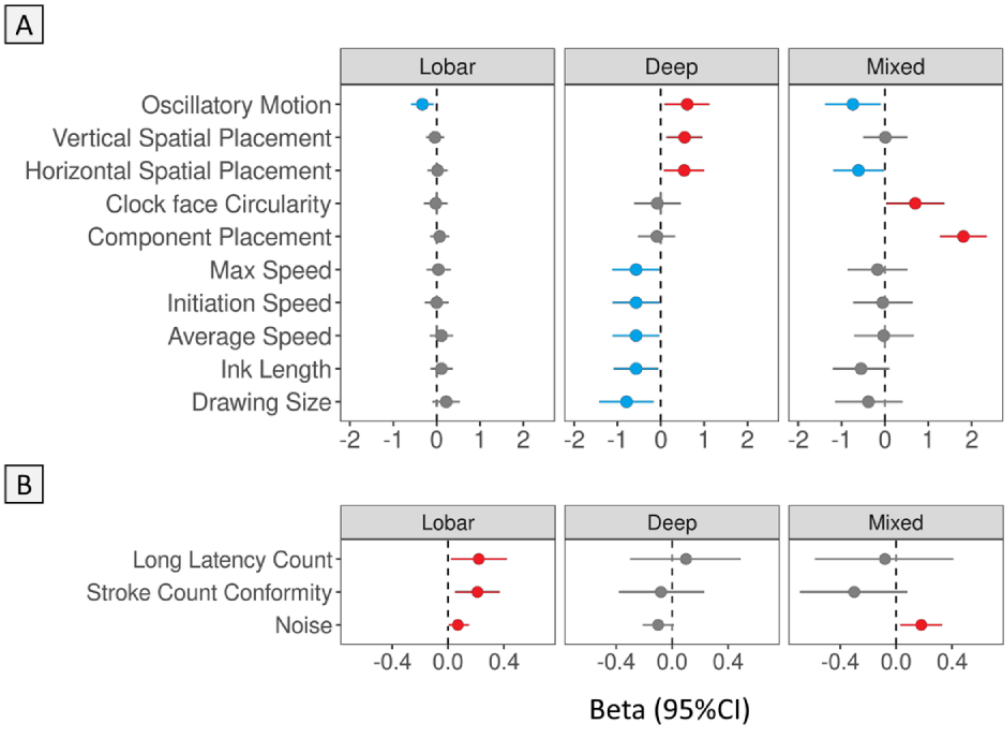

XSL•FO
RenderX

Figure 2. Multiclass classification of CMB evaluation with ROC curves using (A) all DCT facet scores, (B) all DCT facet scores+sex+age, and (C) all DCT facet scores+sex+age+time+MCI, where time is the time (years) difference between the brain MRI date and DCT date. AUC: area under curve; CMB: cerebral microbleed; DCT: digital clock drawing test; MCI: mild cognitive impairment; MRI: magnetic resonance imaging; ROC: receiver operating characteristic.

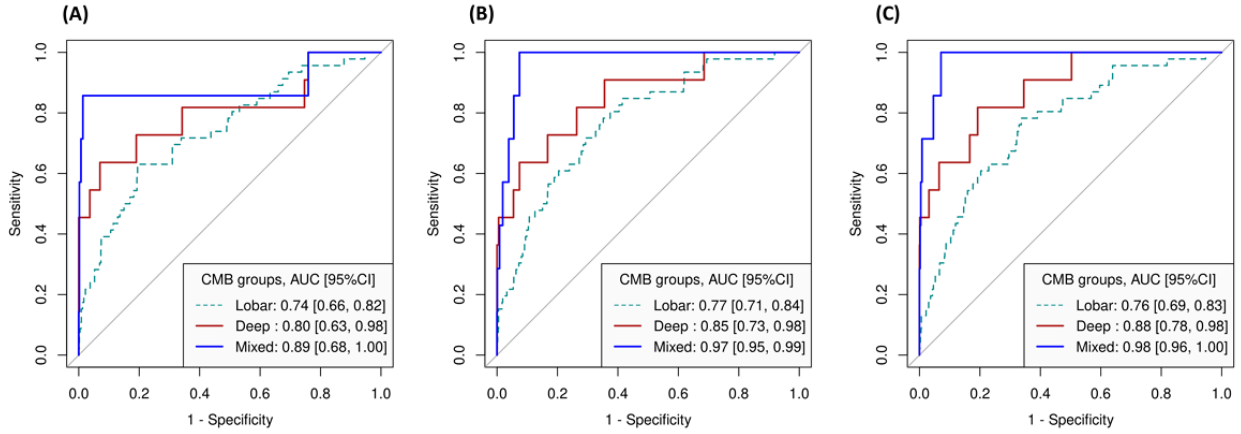

Discussion

Overview

This cross-sectional study in the FHS reveals the potential of the DCT as a cost-effective and objective screening tool for detecting CMB in different brain regions. Unlike the traditional CDT, the DCT offers detailed insights into cognitive function and demonstrates significant associations with CMB, particularly deep and mixed subtypes. The study highlights the limitations of traditional cognitive tests in detecting subtle brain abnormalities such as CMB and underscores the DCT's value in early prediction of dementia risk.

Principal Findings

To our knowledge, our study is the first to evidence that the DCT reveals detailed, nuanced, and hidden information and can serve as a potentially useful screening tool to detect the presence of CMB in different brain regions. As CMB is a risk factor for AD [28] but is accompanied by no or very subtle clinical symptoms, it is costly for clinicians to use neuroimaging to detect CMB. In comparison, the DCT is more cost-effective and can be completed by patients within a couple of minutes with minimal assistance. Another advantage of the DCT is that it objectively captures detailed brain functions through automated digital collection and analysis, while the manual traditional counterpart depends on the subjective assessment by trained testers. Using the DCT may be useful for in future clinical practice for early screening and detection of CMB, triggering interventions that can delay the progress of the disease or prevent AD onset [35].

Comparison to Prior Work

Whereas the traditional CDT has low sensitivity and specificity to screen or diagnose participants with CMB, the DCT scores were significantly associated with CMB in our study, especially deep and mixed CMB, including across 3 different measurement levels (ie, overall score, domains, and facets of the DCT). Previous studies have reported inconsistent associations between CMB and the traditional CDT. For example, whereas 1 study found that CMB were a risk factor for low performance on the CDT [36], 2 others did not find a relationship [37,38]. Another study that did not exclude participants with dementia illustrated that lobar, but not deep, CMB were associated with CDT [39]. It is possible that CDT's crude measurements contain mixtures of different cognitive functions that cannot be broken down into more detailed measurements such as facet scores of DCT and thus may not be able to clearly detect fine brain abnormalities like CMB.

Strengths and Limitations

There were several strengths in our study. First, in our study, the diagnostic discriminability using the DCT facet scores was able to differentiate those with and without CMB, especially the mixed subtype, independent from MCI. Second, the DCT is simple, can be self-administered, and could serve as a screening test before administering costly neuroimaging tests. Third, our study found that participants with lobar CMB and deep CMB had different DCT performance patterns. Whereas the command condition was more strongly associated with deep or mixed CMB, the copy condition may be more associated with lobar CMB. Since deep but not lobar CMB have been identified as risk factors for dementia development in previous studies [28], the DCT may be a valuable tool for the early prediction of dementia risk. By capturing multiple facets of cognitive function as well as their fine-grained interrelations, the DCT affords substantially more sensitive analyses compared to typical measures of domain-specific cognitive functions that are observed in isolation (or aggregated in sum scores that overlook fine-grained interrelations).

This study had several limitations. First, the sample size of deep and mixed CMB cases is relatively small, potentially limiting the generalizability of findings and statistical power to detect associations. Future studies should aim for larger and more diverse cohorts. Second, the associative study design restricts the ability to establish causal relationships between variables, despite efforts to control for confounding factors. Future research could benefit from longitudinal or interventional approaches to explore causality further. Lastly, the lack of

XSL•FO
RenderX

diversity in the FHS cohort may limit the applicability of findings to broader populations. Future studies should strive to include more diverse cohorts to enhance generalizability and reduce potential biases.

Future Directions

Large cohorts with multiethnicities should be used to confirm that the DCT and other digital tools detect CMB and similar pathologies in different brain regions and can serve as a cost-effective screening tool to better identify people at risk earlier in the preclinical stages of the disease [11]. More importantly, user-oriented assessment devices such as the DCT may promote objectivity and equity within public health research, especially in underserved populations.

Acknowledgments

We thank Lilly Buck (MSc) and Leon Li (PhD) for their helpful feedback on an earlier draft of this paper. This work was supported by the National Heart, Lung, and Blood Institute contract (N01-HC-25195) and by grants from the National Institute of Neurological Disorders and Stroke (NS-17950) and the National Institute on Aging (AG-08122, 016495, 068753: RA, and AG-09899: WQQ).

Data Availability

The data sets generated or analyzed during this study are available through the Framingham Heart Study central database. For detailed information on data access, please visit the Framingham Heart Study Brain Aging Program (FHS-BAP) Data Access Portal [40].

Authors' Contributions

SCA-K and QT contributed equally to drafting this paper, study design, and data analysis. TFAA and WQQ contributed to the study design and substantial revision of this paper. RA contributed to the initiation of digital data collection in the Framingham Heart Study, data quality control, data sharing, and review and revision of this paper. CK managed digital data used in this study. ISI, DJL, MA, and JM contributed to the review and revision of this paper.

Conflicts of Interest

RA is a scientific advisor to Signant Health, Novo Nordisk, and Davos Alzheimer's Collaborative. JM receives grants from the NIH (National Institutes of Health) and DOD (U.S. Department of Defense). All other authors have no conflict of interest to declare.

Multimedia Appendix 1
The descriptions of each subdomain of digital clock-drawing test (DCT) scores.
[DOCX File , 33 KB - jmir_v26i1e45780_app1.docx]

References

1. de Bruijn RFAG, Ikram MA. Cardiovascular risk factors and future risk of Alzheimer's disease. BMC Med 2014;12(1):130 [FREE Full text] [doi: 10.1186/s12916-014-0130-5] [Medline: 25385322]
2. Savva GM, Stephan BCM, Alzheimer's Society Vascular Dementia Systematic Review Group. Epidemiological studies of the effect of stroke on incident dementia: a systematic review. Stroke 2010;41(1):e41-e46. [doi: 10.1161/STROKEAHA.109.559880] [Medline: 19910553]
3. Martinez-Ramirez S, Greenberg SM, Viswanathan A. Cerebral microbleeds: overview and implications in cognitive impairment. Alzheimers Res Ther 2014;6(3):33 [FREE Full text] [doi: 10.1186/alzrt263] [Medline: 24987468]
4. Cordonnier C, Al-Shahi Salman R, Wardlaw J. Spontaneous brain microbleeds: systematic review, subgroup analyses and standards for study design and reporting. Brain 2007;130(Pt 8):1988-2003. [doi: 10.1093/brain/awl387] [Medline: 17322562]
5. Wu R, Feng C, Zhao Y, Jin A, Fang M, Liu X. A meta-analysis of association between cerebral microbleeds and cognitive impairment. Med Sci Monit 2014;20:2189-2198 [FREE Full text] [doi: 10.12659/MSM.891004] [Medline: 25377475]
6. Akoudad S, Wolters FJ, Viswanathan A, de Bruijn RF, van der Lugt A, Hofman A, et al. Association of cerebral microbleeds with cognitive decline and dementia. JAMA Neurol 2016;73(8):934-943 [FREE Full text] [doi: 10.1001/jamaneurol.2016.1017] [Medline: 27271785]
7. Wang M, Hu H, Wang Z, Ou Y, Qu Y, Ma Y, et al. Association of cerebral microbleeds with risks of cognitive impairment and dementia: A systematic review and meta-analysis of prospective studies. Brain Disorders 2021;2:100010. [doi: 10.1016/j.dscb.2021.100010]
8. Rasmussen J, Langerman H. Alzheimer's disease—why we need early diagnosis. Degener Neurol Neuromuscul Dis 2019;9:123-130 [FREE Full text] [doi: 10.2147/DNND.S228939] [Medline: 31920420]

XSL•FO
RenderX

9. Guerchet M, Prince M, Prina M. Numbers of people with dementia worldwide: an update to the estimates in the World Alzheimer Report 2015. Alzheimer's Disease International. UK: Alzheimer's Disease International; 2015. URL: https://www.alzint.org/resource/numbers-of-people-with-dementia-worldwide/ [accessed 2024-05-29]
10. Wimo A, Guerchet M, Ali G, Wu Y, Prina AM, Winblad B, et al. The worldwide costs of dementia 2015 and comparisons with 2010. Alzheimer's Dementia 2017;13(1):1-7 [FREE Full text] [doi: 10.1016/j.jalz.2016.07.150] [Medline: 27583652]
11. Au R, Ritchie M, Hardy S, Ang TFA, Lin H. Aging well: using precision to drive down costs and increase health quality. Adv Geriatr Med Res 2019;1:e190003 [FREE Full text] [doi: 10.20900/agmr20190003] [Medline: 31342014]
12. Cosentino S, Jefferson A, Chute DL, Kaplan E, Libon DJ. Clock drawing errors in dementia: neuropsychological and neuroanatomical considerations. Cogn Behav Neurol 2004;17(2):74-84. [doi: 10.1097/01.wnn.0000119564.08162.46] [Medline: 15453515]
13. Libon DJ, Swenson RA, Barnoski EJ, Sands LP. Clock drawing as an assessment tool for dementia. Arch Clin Neuropsychol 1993;8(5):405-415. [Medline: 14589710]
14. Libon DJ, Malamut BL, Swenson R, Sands LP, Cloud BS. Further analyses of clock drawings among demented and nondemented older subjects. Arch Clin Neuropsychol 1996;11(3):193-205. [Medline: 14588923]
15. Libon DJ, Swenson R, Lamar M, Price CC, Baliga G, Pascual-Leone A, et al. The Boston Process Approach and digital neuropsychological assessment: past research and future directions. J Alzheimers Dis 2022;87(4):1419-1432. [doi: 10.3233/JAD-220096] [Medline: 35466941]
16. Souillard-Mandar W, Penney D, Schaible B, Pascual-Leone A, Au R, Davis R. DCTclock: clinically-interpretable and automated artificial intelligence analysis of drawing behavior for capturing cognition. Front Digit Health 2021;3:750661 [FREE Full text] [doi: 10.3389/fdgth.2021.750661] [Medline: 34723243]
17. Müller S, Herde L, Preische O, Zeller A, Heymann P, Robens S, et al. Diagnostic value of digital clock drawing test in comparison with CERAD neuropsychological battery total score for discrimination of patients in the early course of Alzheimer's disease from healthy individuals. Sci Rep 2019;9(1):3543 [FREE Full text] [doi: 10.1038/s41598-019-40010-0] [Medline: 30837580]
18. Souillard-Mandar W, Davis R, Rudin C, Au R, Libon DJ, Swenson R, et al. Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. Mach Learn 2016;102(3):393-441 [FREE Full text] [doi: 10.1007/s10994-015-5529-5] [Medline: 27057085]
19. Binaco R, Calzaretto N, Epifano J, McGuire S, Umer M, Emrani S, et al. Machine learning analysis of digital clock drawing test performance for differential classification of mild cognitive impairment subtypes versus Alzheimer's disease. J Int Neuropsychol Soc 2020;26(7):690-700. [doi: 10.1017/S1355617720000144] [Medline: 32200771]
20. Rentz DM, Papp KV, Mayblyum DV, Sanchez JS, Klein H, Souillard-Mandar W, et al. Association of digital clock drawing with PET amyloid and tau pathology in normal older adults. Neurology 2021;96(14):e1844-e1854 [FREE Full text] [doi: 10.1212/WNL.0000000000011697] [Medline: 33589537]
21. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. Nat Rev Cardiol 2019;16(11):687-698. [doi: 10.1038/s41569-019-0202-5] [Medline: 31065045]
22. Au R, Piers RJ, Devine S. How technology is reshaping cognitive assessment: lessons from the Framingham Heart Study. Neuropsychology 2017;31(8):846-861 [FREE Full text] [doi: 10.1037/neu0000411] [Medline: 29376667]
23. Nyborn JA, Himali JJ, Beiser AS, Devine SA, Du Y, Kaplan E, et al. The Framingham Heart Study clock drawing performance: normative data from the offspring cohort. Exp Aging Res 2013;39(1):80-108 [FREE Full text] [doi: 10.1080/0361073X.2013.741996] [Medline: 23316738]
24. Davis R, Libon DJ, Au R, Pitman D, Penney DL. THink: inferring cognitive status from subtle behaviors. Proc AAAI Conf Artif Intell 2014;2014:2898-2905 [FREE Full text] [Medline: 27066307]
25. Matusz EF, Price CC, Lamar M, Swenson R, Au R, Emrani S, et al. Dissociating statistically determined normal cognitive abilities and mild cognitive impairment subtypes with DCTclock. J Int Neuropsychol Soc 2023;29(2):148-158. [doi: 10.1017/S1355617722000091] [Medline: 35188095]
26. DeCarli C, Massaro J, Harvey D, Hald J, Tullberg M, Au R, et al. Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. Neurobiol Aging 2005;26(4):491-510. [doi: 10.1016/j.neurobiolaging.2004.05.004] [Medline: 15653178]
27. Greenberg SM, Vernooij MW, Cordonnier C, Viswanathan A, Al-Shahi Salman R, Warach S, et al. Cerebral microbleeds: a guide to detection and interpretation. Lancet Neurol 2009;8(2):165-174 [FREE Full text] [doi: 10.1016/S1474-4422(09)70013-4] [Medline: 19161908]
28. Romero JR, Beiser A, Himali JJ, Shoamanesh A, DeCarli C, Seshadri S. Cerebral microbleeds and risk of incident dementia: the Framingham Heart Study. Neurobiol Aging 2017;54:94-99 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2017.02.018] [Medline: 28347929]
29. Romero JR, Preis SR, Beiser A, DeCarli C, Viswanathan A, Martinez-Ramirez S, et al. Risk factors, stroke prevention treatments, and prevalence of cerebral microbleeds in the Framingham Heart Study. Stroke 2014;45(5):1492-1494 [FREE Full text] [doi: 10.1161/STROKEAHA.114.004130] [Medline: 24713533]

XSL•FO
RenderX

30. Maillard P, Seshadri S, Beiser A, Himali JJ, Au R, Fletcher E, et al. Effects of systolic blood pressure on white-matter integrity in young adults in the Framingham Heart Study: a cross-sectional study. Lancet Neurol 2012;11(12):1039-1047 [FREE Full text] [doi: 10.1016/S1474-4422(12)70241-7] [Medline: 23122892]
31. Venables W, Ripley B. Modern Applied Statistics with S.. New York City: Springer; 2002.
32. Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Soft 2009;32(3):1-47. [doi: 10.18637/jss.v032.i03]
33. Zeileis A. Object-oriented computation of sandwich estimators. J Stat Soft 2006;16(9):1-16. [doi: 10.18637/jss.v016.i09]
34. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf 2011;12:77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
35. Chojdak-Łukasiewicz J, Dziadkowiak E, Zimny A, Paradowski B. Cerebral small vessel disease: a review. Adv Clin Exp Med 2021;30(3):349-356 [FREE Full text] [doi: 10.17219/acem/131216] [Medline: 33768739]
36. Fang M, Feng C, Xu Y, Hua T, Jin A, Liu X. Microbleeds and silent brain infarctions are differently associated with cognitive dysfunction in patients with advanced periventricular leukoaraiosis. Int J Med Sci 2013;10(10):1307-1313 [FREE Full text] [doi: 10.7150/ijms.6430] [Medline: 23983590]
37. Büyükşerbetçi G, Saka E, Oğuz KK, Göçmen R, Arsava EM, Topçuoğlu MA. Cognitive dysfunction in relation to topography and burden of cerebral microbleeds. Noro Psikiyatr Ars 2018;55(1):84-90 [FREE Full text] [Medline: 30042647]
38. Poliakova T, Levin O, Arablinskiy A, Vasenina E, Zerr I. Cerebral microbleeds in early Alzheimer's disease. J Neurol 2016;263(10):1961-1968. [doi: 10.1007/s00415-016-8220-2] [Medline: 27389080]
39. Chung C, Chou K, Chen W, Liu L, Lee W, Chen L, et al. Strictly lobar cerebral microbleeds are associated with cognitive impairment. Stroke 2016;47(10):2497-2502. [doi: 10.1161/STROKEAHA.116.014166] [Medline: 27625380]
40. Chobanian & Avedisian School of Medicine. Access ou data. URL: https://www.bumc.bu.edu/fhs-bap/for-researchers/access-our-data/ [accessed 2024-05-29]

Abbreviations

AD: Alzheimer disease
AUC: area under the curve
CDT: clock-drawing test
CMB: cerebral microbleed
DCT: digital clock-drawing test
FHS: Framingham Heart Study
MCI: mild cognitive impairment
MRI: magnetic resonance imaging
ROC: receiver operating characteristic
WMH: white matter hyperintensity

Edited by T Leung; submitted 17.01.23; peer-reviewed by K Johnson, F Abdulla; comments to author 24.01.24; revised version received 18.03.24; accepted 31.03.24; published 29.07.24.

Please cite as:
Akhter-Khan SC, Tao Q, Ang TFA, Karjadi C, Itchapurapu IS, Libon DJ, Alosco M, Mez J, Qiu WQ, Au R
Cerebral Microbleeds in Different Brain Regions and Their Associations With the Digital Clock-Drawing Test: Secondary Analysis of the Framingham Heart Study
J Med Internet Res 2024;26:e45780
URL: https://www.jmir.org/2024/1/e45780
doi:10.2196/45780
PMID:39073857

©Samia C Akhter-Khan, Qiushan Tao, Ting Fang Alvin Ang, Cody Karjadi, Indira Swetha Itchapurapu, David J Libon, Michael Alosco, Jesse Mez, Wei Qiao Qiu, Rhoda Au. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Multidimensional Approach for Evaluating Reality in Social Media: Mixed Methods Study

HyunYi Cho[1], PhD; Wenbo Li[2], PhD; Rachel Lopez[1], MA

[1]The Ohio State University, Columbus, OH, United States

[2]Stony Brook University, Stony Brook, NY, United States

Corresponding Author:
HyunYi Cho, PhD
The Ohio State University
154 N Oval Mall
Columbus, OH, 43210
United States
Phone: 1 6142923400
Fax: 1 614 2922055
Email: cho.919@osu.edu

Abstract

Background: Misinformation is a threat to public health. The effective countering of misinformation may require moving beyond the binary classification of fake versus fact to capture the range of schemas that users employ to evaluate social media content. A more comprehensive understanding of user evaluation schemas is necessary.

Objective: The goal of this research was to advance the current understanding of user evaluations of social media information and to develop and validate a measurement instrument for assessing social media realism.

Methods: This research involved a sequence of 2 studies. First, we used qualitative focus groups (n=48). Second, building on the first study, we surveyed a national sample of social media users (n=442). The focus group data were analyzed using the constant comparison approach. The survey data were analyzed using confirmatory factor analyses and ordinary least squares regression.

Results: The findings showed that social media reality evaluation involves 5 dimensions: falsity, naturality, authenticity, resonance, and social assurance. These dimensions were differentially mapped onto patterns of social media use. Authenticity was strongly associated with the existing global measure of social media realism ($P<.001$). Naturality, or the willingness to accept artificiality and engineered aspects of social media representations, was linked to hedonic enjoyment ($P<.001$). Resonance predicted reflective thinking ($P<.001$), while social assurance was strongly related to addictive use ($P<.001$). Falsity, the general belief that much of what is on social media is not real, showed a positive association with both frequency ($P<.001$) and engagement with ($P=.003$) social media. These results provide preliminary validity data for a social media reality measure that encompasses multiple evaluation schemas for social media content.

Conclusions: The identification of divergent schemas expands the current focus beyond fake versus fact, while the goals, contexts, and outcomes of social media use associated with these schemas can guide future digital media literacy efforts. Specifically, the social media reality measure can be used to develop tailored digital media literacy interventions for addressing diverse public health issues.

(J Med Internet Res 2024;26:e52058) doi:10.2196/52058

KEYWORDS

fake; fact; misinformation; reality; social media; scale; measure; instrument; user-centric; tailoring; digital media literacy

Introduction

Background

The surgeon general of the United States declared that misinformation is a serious threat to public health and urged all Americans to help deter its spread [1]. Research has found that misinformation is widespread in diverse domains of public health including, but not limited to, drugs and smoking, communicable and noncommunicable diseases, eating disorders, treatments and medical interventions, and mental health [2,3].

XSL•FO
RenderX

Given its ubiquity, addressing misinformation requires whole-of-society efforts [1]. At the same time, the surgeon general's advisory calls on researchers and educators to take specific actions. Researchers are encouraged to develop a user-centric understanding of how different individuals interpret and interact with information on social media. On the basis of this understanding, educators are asked to offer programs that assist diverse individuals to become more discerning consumers of digital media content.

Overall, there seems to be a paradoxical reliance on, and distrust in, social media. Research suggests that users generally doubt the accuracy of social media information [4,5]; for example, a recent survey found that 62% of participants from 7 countries, including the United States, reported seeing misinformation on social media [4]. Another study found that across a variety of digital media contexts, participants consistently expected other users to lie about their appearance [5]. Concurrently, 72% of Americans use social media platforms and visit them at least once a day [6].

The extensive social media use and simultaneous concern about misinformation suggest that the motivations for, and gratifications from, social media use may go beyond obtaining factual information. Previous research in mass media has shown that user evaluations of realism use multiple dimensions [7,8], that is, users use divergent schemas or guiding frameworks comprising a collection and structure of information when interpreting media messages. Similarly, studies suggest that when evaluating social media information, users may use more than just a dichotomy of fake versus fact [9-11], and different evaluative schemas may be used for different goals and contexts [12,13]. However, despite this understanding, a comprehensive examination of these dimensions has yet to be undertaken.

Objectives

The goal of this research was to advance the current knowledge about user evaluations of social media content and to develop and initially validate a measurement instrument for assessing social media realism. Given the pervasiveness of misinformation [1-3], our goal was to offer a general measure that can be used in diverse public health domains. These efforts aim to inform and enhance public health misinformation–countering efforts using digital media literacy. Toward this end, 2 studies were conducted. First, qualitative focus group research examined salient themes in social media realism judgments among users. Second, building on the identified themes and existing literature, a quantitative survey study investigated the schemas' connections with patterns of social media use.

Conceptual Foundation

User judgments of media's correspondence with reality are commonly referred to as perceived realism. This construct has been found to be one of the primary determinants of media effects on various public health behaviors and outcomes. One of these domains is violence prevention. Research indicates that when viewers perceive harmful media such as violent content to be highly realistic, the effects are amplified [14,15]. By contrast, melanoma prevention interventions using a media literacy approach found that correcting perceived media realism related to tanned looks was central to the intervention efficacy evidenced at a 6-month follow-up [16,17]. Anti–substance abuse public service announcements were more persuasive when viewers judged that the realism of the messages was high rather than low [7].

Perceived realism has also been found to be important in determining social media effects pertaining to health behaviors; for instance, exposing individuals to a juxtaposition of real and idealized Instagram images of the same woman reduced body dissatisfaction compared to exposure to either the real image or ideal image alone [18]. Perceived realism moderated the effects of exposure to social media e-cigarette messages among teenagers. When perceived realism was low, the effect of the exposure on the provaping attitude was mitigated [19].

Moreover, research suggests that user judgments of social media information may be multifaceted. One of the facets is authenticity, which refers to the congruency between digital and offline selves and behaviors [20]. Authenticity has been a focus of studies on social media realism effects on health behaviors. Cho et al [19] investigated teenagers' consideration of how user-created content on social media reflected the content creators' true selves using items such as "People's posts portray how they really live their lives." Similarly, Tiggemann and Anderberg [18] assessed young women's perceptions using items such as "The models in the advertisements looked like they would look like in person." Importantly, these studies found that a low perception of authenticity reduced the harmful effects of social media on e-cigarette use and eating disorders [18,19]. Recognizing this significance, Jenkins et al [21] called for more public health research assessing the predictors of the perceptions of authenticity concerning nutrition messages on social media.

Authenticity may not be the only dimension users consider when evaluating social media content. Studies suggest that users may rely on social cues to judge the proximity of social media posts to reality. In addition to the core content of the message, the metrics generated by others may provide a bandwagon heuristic for assessing social media messages [22]. Popularity cues such as others' support and approval of a message may provide a rule of thumb for evaluating the credibility of digital media [23]; for instance, the effects of both pro- and antivaping social media messages were increased by strong rather than weak popularity cues [24].

We suspect that another criterion people use to evaluate social media messages may be resonance. A significant body of research has documented that, on social media, users tend to prefer information that is consonant, rather than dissonant, with their existing beliefs. Communication on social media has often been described as occurring in "echo chambers" where users' attitudes are reinforced through repeated interactions with like-minded others [25]. Related mechanisms include selective exposure [26] and confirmation bias [27], whereby people seek information consistent with their preexisting perspectives [28] and process information in a way that affirms their prior values [29]. This biased process suggests that users may give more credibility and evaluate more positively the social media messages that resonate with their predispositions. Exposure to

XSL•FO
RenderX

opposing positions can facilitate polarization through heightened social identities on social media [30].

Study 1: Qualitative Focus Groups

Overview

Existing research suggests that users may be skeptical of misinformation on social media, but they may also consider other factors, such as authenticity, when evaluating the content, rather than relying on a binary distinction of fake versus fact. However, there is still a limited understanding of potential additional factors. To address this gap in knowledge, we used a qualitative method, focus groups, to identify these factors.

Methods

Design and Participants

For our focus group research, we recruited 48 social media users from a large public university in the United States. We chose young adults because they use a wider range of social media platforms more frequently than older adults [6], providing us with a deeper understanding of user experience [31]. The participants were aged between 18 and 34 (mean 20.3, SD 3.3) years. Of the 48 participants, 31 (65%) were female. None of the participants chose to drop out from the study, and no cases were excluded from analysis. Of the 48 participants, 13 (27%) were Asian, 5 (10%) were Black, and 30 (63%) were White. They reported using the following social media platforms: Instagram (41/48, 85%), YouTube (31/48, 65%), TikTok (29/48, 60%), Facebook (24/48, 50%), Twitter (subsequently rebranded X; 22/48, 46%), and Reddit (14/48, 29%).

Ethical Considerations

This study was determined to be exempt by the institutional review board of the Ohio State University (2020E1164). It was deemed that this study, which involved interview procedures, would not reasonably place the participants at risk of harm. Before the focus group participation, individuals were presented with a consent form, informing them of the purpose of the research. Only those who consented participated in the focus groups. Compensation was given in the form of extra credit at a rate of 0.25 credits per 15 minutes of participation. All analyses involved deidentified data.

Procedure

Due to the ongoing COVID-19 pandemic, focus groups were conducted over Zoom (Zoom Video Communications, Inc) using audio and video functions. The focus group topic guide was developed by the researchers, with the overall interview procedure being guided by the work of Krueger and Casey [32]. Each session included 4 to 10 participants and lasted an average of 80 minutes. No parties other than the researchers and participants were present during the sessions. Upon finishing the seventh focus group, the researchers determined that both data saturation and theoretical saturation had been reached; increasing instances of redundancy were found in the data, and successive groups provided diminishing contributions to the development of themes and categories [33]. Thus, data collection ceased after 7 focus group sessions.

The sessions used a semistructured approach, beginning with the identification of unrealistic content. Participants were asked to recall their experiences with, and reactions to, unrealistic content, followed by a detailed discussion on realistic content in later parts of the session. While all 7 focus groups followed this general flow and organization of topics, the specific questions used for each focus group were adjusted to probe the different themes and issues that arose during prior sessions. Given the far-reaching impact of social media on both mental and physical health, the participants were encouraged throughout to consider a variety of content and platforms. At the conclusion of each session, the moderator provided a verbal summary of the discussion, inviting participants to provide feedback on the accuracy of the summary and to offer any additional comments. All sessions were audio-recorded.

Data Analysis

Immediately after each focus group, the researchers met to debrief on their notes, considering observed trends, unique concepts, and suggestions for future discussions. Inductive, deductive, and iterative approaches were used. The deductive approach was based on existing research reviewed in the Conceptual Foundation section above [7,8,19,20,23]. Concurrently, because novel themes emerged, coding took an inductive approach: these themes were iteratively reviewed and refined to identify the salient schemas for evaluating social media information and the contexts of this evaluation.

Transcription was carried out manually, with the assistance of Zoom-produced transcripts, and cross-validated with the audio recordings. The transcripts for the 7 sessions comprised 76,704 words in total. Prominent themes across the focus group sessions were identified using the constant comparison approach [34]. NVivo software (Lumivero) was used to code responses by speaker (ie, as a form of case) and to attach participant information such as race, sex and focus group session number as case attributes. The data were chunked into smaller units, and new patterns were compared across successive focus groups using an emergent-systematic design. Specific quotes were coded to each theme in NVivo, allowing for the easy consolidation and creation of the schemas of interest.

As all authors (faculty and graduate students) study the effects of media on health outcomes, we used our knowledge about research findings in this area to inform the interpretation. All researchers, comprising female and male individuals, were involved in the hosting of the focus groups and the theme development process. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist is provided in Multimedia Appendix 1.

Results

Overview

Overall, the focus group research revealed how social media users judge the realism of content. Multiple approaches were identified: falsity, authenticity, naturality, resonance, and social assurance. Textbox 1 presents conceptual definitions of these schemas.

XSL•FO
RenderX

Textbox 1. Schemas for evaluating social media content.

Schema and definition

- Falsity: general sense of distrust or doubt about social media content
- Naturality: willingness to accept artificiality and engineered aspects of social media representations
- Authenticity: the extent to which one's social media posts align with one's inner thoughts and feelings
- Resonance: felt relatedness between a social media post and the user's self
- Social assurance: reliance on metrics associated with social media posts

Falsity

There was a general sentiment that much of what is on social media is not real, or at least some of it cannot be trusted. A pervasive sense of disillusion and doubt was evident throughout the focus group sessions. Many participants assumed that most, if not all, of what is on social media is not real or at least does not present "the whole picture." This ranged from fake news (usually fake unless coming from a credible platform with citations and verification) to accounts with few followers, "weird" usernames, low-quality pictures in profiles and posts, no mutual followers, and repetitive posts. The motivation for these behaviors was considered to be for money, fame, popularity, and a need for attention. These specific visible cues and concrete motivations associated with the widespread falsity of information on social media were noted by participants consistently across the 7 focus group sessions.

In addition to these more surface-level determinants, participants' comments revealed another layer. They perceived that there may be a "fake ideology" on social media that can easily create a "false perception" of somebody:

> *Everyone just has this invisible hand pushing them to do something and not do something.* [P8; male; group 2]
>
> *There's a lot of like performative activism. I'm not sure if that's the right word, but people just post things because they want to make themselves look better, not because they actually believe in whatever.... So, it's like really hard to tell sometimes on social media versus real life.* [P10; female; group 2]

Regarding a more subtle level of fakeness, participants showed a certain level of resignation, tolerance, and acceptance of this aspect. They said that they tend to ignore it or to not pay attention:

> *I try not to read too much into other people's posts.... I know I put a disclaimer on my Facebook that says that I'm showing more mountain tops than the valleys that I also have in my life, like everyone else.* [P14; female; group 2]

Others were concerned that the lack of realism on social media can have "toxic" effects:

> *I think that it is generally unimportant for people to be realistic on social media, and it has given us a false perception of the way people are in our communities, in our society.* [P8; male; group 2]

Authenticity

Participants expressed concerns about the degree to which the content posted by others is a true representation of the self. They cared about the extent to which an individual's posts reflected their internal thoughts and values. Content that demonstrated authenticity was perceived as more real. Whereas 1 facet of falsity involved conforming to social norms and "invisible hands," authenticity was seen as asserting one's uniqueness:

> *Posts that I think are genuine [are ones] about trying to not be like everybody else.* [P41; female; group 6]
>
> *Personally, I like real content like...if you're being authentic then you, I think you can build like a friendship or bond with that person because you're being vulnerable with each other. Versus like if you're just fake all the time like you're posting Photoshopped images like you're in it to look good. You're not, you're in it for the popularity points I think instead of the friendship or the realness.* [P38; female; group 6]

Participants provided examples of authentic content. Self-disclosure was widely regarded as a signifier of real content. It did not need to focus on negative experiences, although some found that events "like arguments and stuff like that...are more likely to be real than people being nice to each other... [since they are] harder to fake" (P29; female; group 5). Another participant said that realism was conveyed through the disclosure of one's "deep emotions or secret struggles...or personal issues" because "[it] makes you feel closer to that person, or like you know them a little bit better as a person" (P31; female; group 5), and yet another participant added that "it's like really getting more insight of them personally instead of [just seeing] their outside positive side" (P34; female; group 5). Disclosure provided a bridge between an individual's inner thoughts and outer expressions, allowing them to be consistent to their true self. This evaluation applied to public figures as well:

> *People always put their best foot forward on social media, usually.... So, I think I get an indicator that something is real when an influencer takes a minute and says, "Here's a life update: I actually haven't been the happiest." I tend to believe that those things are real, just because people don't often open up like that on social media.* [P31; female; group 5]

XSL•FO
RenderX

Naturality

Although participants generally wanted others' presentation of the self and events to mimic external reality and its flaws, they made allowance for some degree of latitude and deviation. Participants indicated a certain willingness to accept some of the artificial and engineered aspects of self-presentation and storytelling that digital media affordances allow.

At a basic level, the practice of editing visual content such as photos or videos had implications for perceptions of naturality. Participants seemed to embrace a tolerable level of artificiality, as they distinguished between "heavy," "extreme," and "body-altering" editing from "enhancements." The former were generally regarded as unreal, treated with disapproval or even light ridicule. Multiple participants also indicated an ease with identifying these forms of heavy editing, that "it's really obvious" (P27; male; group 4). By contrast, administering enhancements such as color corrections and "basic" filters was regarded as not only acceptable but also a predominant practice on the web. Content was still considered naturally real as long as edits were unobtrusive and refrained from altering the substance of the photos, such as erasing portions of the background or morphing body parts.

As a participant described, these types of edits are meant "to change the quality [of the image] instead of changing the actual person" [in the post].... Everyone does that" (P21; female; group 4). Filters serve to "make [photos] pop out and bring [them] to life a little more," allowing audiences to "perceive what [the posters are] seeing, what [they are] trying to display" (P5; male; group 1). In other words, rather than strictly adhering to factual realism by refraining from manipulating the content, users might instead provide naturality through enhancing edits. These changes intentionally overemphasize features of the post to compensate for the experiential gap between the poster and viewer. Thus, enhancing edits demonstrate naturality by maintaining a balance between fidelity to reality and (unobtrusive) manipulation.

Resonance

A sense of resonance was another lens through which participants evaluated realism. The degree to which the social media representation reverberated with one's own experience was important:

> *I feel like we base what's real or not...on what our own experiences are. So, if we've experienced something similar...through ourselves or other people, we're more likely to see it [as real]. So, if no one that we know closely has ever done X, Y, or Z, then we're less likely to believe that another person outside of our circle has ever done that.* [P29; female; group 5]

Another participant explained that relatable content felt more real:

> *[L]ike when people like sharing more personal events on their Instagram or YouTube, I feel like people are able to relate to it more, and I think it makes it more believable, the message they're trying to put out.* [P1; male; group 1]

In this context, participants tended to perceive content from peers as more realistic because these sources were considered more relatable. Peers were seen as relatable because they shared similar traits and life experiences with the participants; for instance, a participant described how he tended to "compare [himself] way more to people [his] own age and people in college than [to] celebrities" (P27; male; group 4). In this case, the participant's identity as a young adult and a student served as a standard for comparison to other social media users.

The benefit and value of resonance were described by another participant:

> *So, when people see others talking about what's going on in their lives or something that they relate to, and seeing in a context that it's okay to feel this way. It's okay to be this way. It's okay to look this way. It's okay to be like, who you are in this facet. The more exposure they have to that type of like positivity and acceptance, the more widespread it is, so, I feel like realness helps others come to terms with their own selves.* [P20; female; group 4]

Social Assurance

Another factor in determining the realism of social media content was the degree of social validation associated with it. Online activity data, which serve as a signifier that others have endorsed the content, influenced the evaluation of social media content. The visibility of others' engagement behaviors may indicate an existing consensus regarding the realism of the content. Preexisting engagement with a post can influence users' perceptions of the content's validity by suggesting how real others have already determined it to be.

Participants reported that seeing the interactions of others influenced their perceptions of realism; for instance, a participant described how engagement metrics caused her to reconsider her initial evaluation of a post because after seeing "all of [her close friends and family] posting positive comments and stuff like that, it [made her] think... [she] should like this post [too] because they believe it" (P10; female; group 2). This participant also noted how a *lack* of likes or comments on a post sowed discomfort and doubt about the realism of the post;

> *If I do see a post that I really like, and nobody has commented on it yet, I feel weird. I don't want to be the first one [to interact with it].* [P10; female; group 2]

Engagement metrics such as likes, comments, and shares influenced users' assessments and advocacy, or lack thereof, of the realism of certain content, biasing subsequent evaluations from other users.

Discussion

Overview

The 5 main themes that emerged from the focus group sessions suggest that users manage divergent schemas to assess the relationship between social media and reality. The binary distinction of fake versus fact may not appropriately capture these nuances. While participants expressed that they were

XSL•FO
RenderX

distrustful of social media content, this schema coexisted with others, which allowed for deviations, different standards, and less rigorous applications. Altogether, these schemas suggest that users are willing to negotiate the boundaries between social media and reality based on the context and motivations.

Of note, these findings build upon and extend previous research on several key areas, including the effects of mass media on health behaviors [7,14,15] and authenticity [18-20] or social media metrics [22-24] pertaining to health-relevant decisions and actions. The results illuminate the distinct nature of judgments regarding social media realism compared to mass media realism. The findings underscore that the evaluation of social media realism encompasses dimensions that go beyond the consideration of authenticity or social media metrics. This study advances our understanding of how individuals perceive digital information, suggesting multiple pathways through which social media information can impact health behaviors.

Dimensions

Falsity may reflect an overarching skepticism toward social media content. This concern ranged from the more visible factors, including fake accounts using weird usernames without a profile photo, to the more subtle issues, including people not being honest on social media. This latter concern was expressed using terms such as “fake ideology” and “invisible hands” on social media that drive people to be less than forthright.

The remaining 4 dimensions could be grouped into 2 categories. The first category includes authenticity and naturality, which concern the closeness between true others and their social media representations. Authenticity may be more about the inner thoughts and feelings of others, whereas naturality focuses on their external features and characteristics. The second category includes social themes, including resonance and social assurance. Resonance is based on a deeper level of understanding of others and the self, whereas social assurance relies on surface-level cues created by others.

Naturality refers to the tolerable degree of departure from the true forms and shapes of entities in reality. It is a willingness to accept a certain level of artificiality, straddling a line between the original and genuine and the made-up, manufactured forms and shapes in our world. These findings echo prior research reporting that “no-makeup” movements and calls for “natural beauty” are actually positively associated with cosmetic sales [35]. Users seem to place a premium on the appearance of naturality, even when it is a product of planned and contrived efforts. This openness to compromise seems to be stronger when social media use is motivated by enjoyment rather than surveillance. Some participants even noted that realism does not matter on these occasions; for example, although some participants described TikTok as fake, others defended it as a platform where they can be silly and have a good laugh.

Authenticity refers to the extent to which a person’s social media representations align with their true self. It entails evaluating the consistency between the 2 [15]. Authenticity requires the willingness to be genuine and distinguish oneself from others, while falsity often arises from conforming to what is trendy, popular, and acceptable on social media. Users expect authentic self-representations to be consistent and stable over time and not be influenced by the social norms of superficiality and positivity.

Resonance is the self-validation that one derives from others’ content and posts. It occurs when others’ posts reverberate with one’s own experiences, beliefs, and attitudes. Whereas authenticity refers to the alignment between the true and social media–constructed self, resonance is the similarity of self and others on social media. Users find the feelings of connection and relatability meaningful and rewarding because they provide opportunities for reflection and validation of the self and their thoughts and experiences. Therefore, resonance refers to the relatedness between a post and the user’s self that is felt internally.

By contrast, social assurance is a heuristic that users use to quickly judge the veracity and social value of a message. Unlike the assessment of authenticity and resonance, which may require a relatively closer examination of social media content, social assurance relies on more readily visible metrics associated with a post, including likes. Knowledge of other users’ engagement with a post may serve as cues for the credibility [23] as well as general acceptance of the content [22]. As a contrast to mass media, these findings highlight the value of shared experiences on social media.

In summary, the widespread concern about falsity coexists with the willingness to accept manufactured social media representations by others and rely on others’ judgments rather than one’s own. While users may be suspicious about others’ posts and motives, they still value authenticity and report benefiting from resonance experiences. Moreover, the focus group discussion suggests that these schemas may be related to social media use motivations, gratifications, and the contexts that different platforms differentially afford.

Limitations

There are some limitations to this study. The participants were college students, who may not represent the broader population of social media users. This sample was chosen because young adults use social media more frequently and on a wider range of platforms than older adults. As samples of experienced users allow researchers to gain a more in-depth insight into user perspectives [25], these participants may appropriately capture the purpose of this focus group research. Future research should include a more heterogeneous sample. Due to the focus of this paper, we could not include all contextual information that emerged during the focus group sessions. Finally, focus group data involve the communication of ideas within a social setting. Participants may express their thoughts differently within a social setting compared to a private or anonymous setting.

Study 2: Quantitative Survey

Overview

The focus group research provided valuable insights into the diverse approaches that people use to navigate the array of information on social media. On the basis of these findings, we developed measurement items capturing users’ schemas for evaluating social media content. The items were then

XSL•FO
RenderX

administered to a general sample of social media users to initially assess their content validity and construct validity.

Hypotheses

Overview

The results of our focus group research showed that social media reality judgments may involve 5 dimensions: falsity, naturality, authenticity, resonance, and social assurance. On this basis, we formulated the following hypothesis:

- H1—Perceived social media realism is a multidimensional construct comprising falsity, naturality, authenticity, resonance, and social assurance.

To test the distinctiveness of the dimensions, we formulated hypotheses predicting the relationship between each dimension and ≥1 external variables. Informed by the insights gained from our focus groups and literature review, we expected that the schemas would be differentially associated with features of social media use.

Falsity

We considered that falsity would be distinct from the other dimensions in that it is a general sense of distrust about what is presented on social media. Despite this skepticism, the participants in our study, as well as those in a global study on information literacy [4], continued to use social media. However, the frequency and level of engagement may differ, depending on the strength of the falsity beliefs. Some participants noted that they do not "read too much into" others' posts, while other participants reported closing their accounts on certain platforms. For these reasons, we hypothesized that falsity will be negatively associated with the frequency of use and engagement with social media. We also expect that stronger falsity beliefs would lead to greater information seeking to verify the social media representations against established facts. Information seeking is a behavior to reduce uncertainty [36], which could be associated with suspicion about the veracity of the stimuli at hand. Accordingly, we formulated the following hypotheses:

- H2a—Falsity will be negatively associated with use frequency and engagement with social media.
- H2b—Falsity will be positively associated with greater information seeking.

Authenticity

We thought that the anticipated multiple dimensions should be differentially associated with the currently available global measure of perceived social media realism [18,19]. This existing global measure, as discussed earlier, seems to primarily capture authenticity, which is one of the most frequently studied aspects of social media realism judgments. Therefore, we anticipated that the extant global measure will be more strongly associated with authenticity than with the other dimensions. We anticipated that the beliefs about authenticity could be linked to eudaimonic gratification from social media use. Research has found that one's psychological well-being is associated with the sense that one can manage and maintain one's authentic self [37]. Building on these findings, we conjectured that the desire for, and rewards from, encountering and interacting with authentic others could lead to a meaningful social media use experience, thereby facilitating eudaimonic gratification. Accordingly, we developed the following hypotheses:

- H3a—Authenticity will be more strongly associated with the existing global measure of realism than the other dimensions.
- H3b—Authenticity will be positively associated with eudaimonic gratification.

Naturality

Naturality, the willingness to embrace artificiality in social media representations, may be related to other types of gratifications or motivations. We hypothesized that the reliance on the naturality schema may be associated with hedonic enjoyment gratification and the motivation to escape from the mundane and routine world through social media use. During our focus group sessions, some participants commented that realism does not really matter when they want to have fun being on social media. This pursuit of happiness through media use has been conceptualized as hedonic enjoyment [38]. Escape is the desire to get away from the daily rituals of work and life [39]. In this state, people may prefer new, novel, and exotic stimuli to the familiar [40]. Accordingly, the following hypotheses were constructed:

- H4a—Naturality will be positively associated with hedonic gratification.
- H4b—Naturality will be positively associated with escape motivation.

Resonance

As users manage networks that influence the kinds of information they will consume, network characteristics may be related to the schema they use to evaluate social media messages. The preference for ideas that affirm and reinforce existing ones—resonance—may be associated with a homophilous social media network. Homophily refers to the degree of similarity of node attributes, reflecting the socialization process and social attitudes [41]. The propensity to value consonant social media information may be linked to how individuals use the information [27]. Specifically, confirmation bias was found to be positively associated with the time spent processing information and the degree of cognitive reflection on the information [28]. These findings suggest that resonance may predict greater reflective thinking when exposed to self-affirming information. Accordingly, we formulated the following hypotheses:

- H5a—Resonance will be positively associated with homophilous social networks.
- H5b—Resonance will be positively associated with greater reflective thinking.

Social Assurance

Social assurance, the dependence on external metrics rather than the internal message content to form judgments, may serve a surveillance goal when using social media, that is, when a person's primary purpose is to find out what is going on in society and what is popular and trendy, external metrics can be useful. However, this dependency on mental shortcuts of social

XSL•FO
RenderX

assurance is unlikely to be positively associated with either information-seeking or reflective-thinking behaviors. The following hypotheses were developed:

- H6a—Social assurance will be positively associated with surveillance motivation.
- H6b—Social assurance will not be associated with information seeking or reflective thinking.

Research Question

In addition, we wondered whether these 5 schemas of realism judgments might be differentially related to the addictive use of social media [42,43]. Prior research found that perceived social media realism is a factor predicting social media addiction [44]. This finding suggests that correcting social media realism may mitigate social media addiction. However, the study used a global measure of realism, and it is unknown which specific dimension may have a stronger association. As little prior research provided us a basis to formulate a hypothesis, we proposed the following research question (RQ) to examine the possible associations:

- What are the relationships between the 5 dimensions and social media addiction?

Methods

Design and Participants

We recruited social media–using adults through the CloudResearch web-based panel (Prime Research Solutions LLC). To ensure a diverse sample, we used quota sampling to ensure that our sample comprised 50% young adults (aged 18-29 y) and 50% adults (aged ≥30 y). This decision was based on research indicating that younger adults are more likely to use social media than older adults [6]. The sample size was 442 participants, after excluding those who failed 1 of 4 attention checks (n=28). The participants' ages ranged from 18 to 96 (mean 41.8, SD 20.7; median 29.0) years. Of the 442 participants, 219 (49.5%) were female. They reported using the following social media platforms: Facebook (392/442, 88.7%), Instagram (290/442, 65.6%), Reddit (180/442, 40.7%), TikTok (234/442, 52.9%), and Twitter (213/442, 48.2%).

Ethical Considerations

This study was determined to be exempt by the institutional review board of the Ohio State University (2022E0763). It was deemed that this study, which involved survey procedures, would not reasonably place the participants at risk of harm. Before the survey participation, individuals were presented with a consent form. Only those who consented participated in the survey. The web-based panel company CloudResearch provided the compensation to participants, which was up to US $3. All analyses involved deidentified data.

Measures

Measures were given on a scale ranging from 1=*strongly disagree* to 5=*strongly agree*, unless noted otherwise. On the basis of the findings from the focus groups and existing research, items were developed to measure the 5 aspects of social media reality judgments: falsity, naturality, authenticity, resonance, and social assurance. Multimedia Appendix 2 presents the full list of items, along with scale means, SDs, and reliability.

Social media use frequency was assessed for 5 platforms, including Facebook and Twitter. For each platform, participants were asked how often they used it, and they rated their use on a 5-point scale: 1=*I don't use this*, 2=*seldom*, 3=*sometimes*, 4=*often*, and 5=*very often* (mean 2.94, SD 1.09; Cronbach α=0.72). Social media engagement was measured by asking participants how often they engaged in liking, commenting, replying to others' comments, and sharing. The scale ranged from 1=*never* to 5=*very often* (mean 3.44, SD 0.91; Cronbach α=0.84).

With regard to information seeking, following the stem "after seeing others' social media posts," participants were given items that included how often they "look up additional information to know more about the topic." The scale ranged from 1=*never* to 4=*often* (mean 2.77, SD 0.74; Cronbach α=0.85).

Items for the global measure of social media realism included "What's on social media reflects real life for the most part" (mean 3.01, SD 0.95; Cronbach α=0.90) [18].

Gratifications included eudaimonic and hedonic enjoyment. Items were adapted from the study by Oliver and Raney [38]. Eudaimonic gratification items included social media "provide a greater understanding of life" (mean 3.41, SD 0.89; Cronbach α=0.86). Hedonic enjoyment items included "I have fun while on social media" (mean 3.79, SD 0.78; Cronbach α=0.85).

Motivations included escape and surveillance. Escape motivation was measured with the scale developed by Rubin and Perse [39]. Following the stem "I use social media to," items included "forget about work and other things" (mean 3.41, SD 0.99; Cronbach α=0.90). Surveillance motivation was assessed using the social media use motivation scale developed and validated by Cho et al [19]. Respondents indicated the degree to which they use social media to "keep up to date with what is trending" (mean 3.48, SD 0.87; Cronbach α=0.84).

Network homophily [45] was assessed by presenting the stem "most people in my social media connections" and following it with items that included "have thoughts that are similar to mine" and "express attitudes similar to mine" (mean 3.49, SD 0.80; Cronbach α=0.83).

Regarding reflective thinking, participants were asked, "When seeing other people's social media posts, how often do you think about the following?" Items included "what may be the motivation for the post," and the scale ranged from 1=*never* to 4=*often* (mean 2.82, SD 0.67; Cronbach α=0.84). These items were based on the study by Chen [46].

Addictive social media use was measured with 6 items taken from the study by Caplan [42]. The items represented compulsive use, problematic use, and withdrawal dimensions. As these items loaded onto the same factor (loadings: 0.75-0.85), we combined them into a single factor (mean 2.53, SD 0.99; Cronbach α=0.88).

XSL•FO
RenderX

Data Analysis

H1 was tested with confirmatory factor analyses (CFA) using the R package *lavaan*. H2 to H6 and the RQ were tested using a series of ordinary least squares multiple regressions. In each model, the dependent variable was regressed onto all dimensions, rather than only the one predicted to be associated, for a more rigorous evaluation. No issue of multicollinearity was found in the data, with variance inflation factors <4, tolerance values >0.2, and condition indices <30. All analyses were conducted using R (R Foundation for Statistical Computing).

Results

Factor Structure

H1 hypothesized that the items would comprise a 5-factor structure. Participants' responses to the items were included for the CFAs. First, a first-order single-factor model was estimated. Second, a first-order oblique 5-factor model was estimated, where the 5 factors were allowed to be correlated with each other. Third, we estimated a first-order 3-factor model comprising (1) falsity, (2) naturality and authenticity, and (3) resonance and social assurance. Finally, a second-order single-factor model was estimated, where the second-order factor loaded on the 5 first-order factors, which were not allowed to be correlated.

Three indices evaluated the fit of the CFA models: the comparative fit index (desirably >0.90) [47], the root mean square error of approximation (desirably <0.08) [48], and the difference of Bayesian information criterion across the models [49]. Model comparison was performed using sequential likelihood ratio tests. Table 1 shows the model fit indices. These results show that the first-order oblique 5-factor model yielded the best fit to the data. The model comparison demonstrated that the first-order 5-factor model performed significantly better than the first-order single-factor model ($\Delta\chi^2_{10}$ =2556.1; P<.001), the first-order 3-factor model ($\Delta\chi^2_{7}$ =1326.6; P<.001), and the second-order single-factor model ($\Delta\chi^2_{5}$ =51.0; P<.001). The best of these 4 models is therefore the first-order 5-factor model. This model also has the lowest information criteria of all models (Akaike information criterion=27,890, Bayesian information criterion=28,168), which suggests that it is the best fitting model out of the 4. These findings support H1. Table 2 summarizes the results, where the numbers are standardized regression coefficients. Figure 1 presents the factor structure.

Table 1. Confirmatory factor analysis results.

Model	Chi-square (*df*)	CFI[a]	RMSEA[b]
First-order single-factor	3519.9 (377)	0.606	0.137
First-order 3-factor	2290.4 (374)	0.760	0.108
First-order 5-factor	963.8 (367)	0.925	0.061
Second-order single-factor	1014.8 (372)	0.919	0.063

[a]CFI: comparative fit index.

[b]RMSEA: root mean square error of approximation.

XSL•FO
RenderX

Table 2. Ordinary least squares regression results of testing H2 to H6 and the research question.

	Falsity	Naturality	Authenticity	Resonance	Social assurance
SM[a] use frequency	0.199[b]	0.257[b]	0.065	0.095	0.230[b]
SM engage	0.128[c]	0.213[b]	0.163[c]	0.208[b]	0.064
Info seek	0.197[b]	0.138[c]	0.082	0.248[b]	0.054
Global	0.004	0.012	0.467[b]	0.142[b]	0.267[b]
Eudaimonic	0.114[c]	0.138[c]	0.175[b]	0.303[b]	0.085
Hedonic	0.023	0.321[b]	0.188[b]	0.312[b]	–0.091
Escape	0.219[b]	0.228[b]	–0.024	0.119[d]	0.280[b]
Homophily	0.102[d]	0.047	0.242[b]	0.201[b]	0.150[d]
Reflect think	0.105[d]	–0.009	0.135[d]	0.266[b]	0.107
Surveillance	0.049	0.155[b]	0.076	0.277[b]	0.193[b]
Addictive use	0.062	0.102[d]	0.101[d]	0.041	0.469[b]

[a]SM: social media.
[b]$P \leq .001$.
[c]$P < .01$.
[d]$P < .05$.

XSL•FO
RenderX

Figure 1. Factor structure.

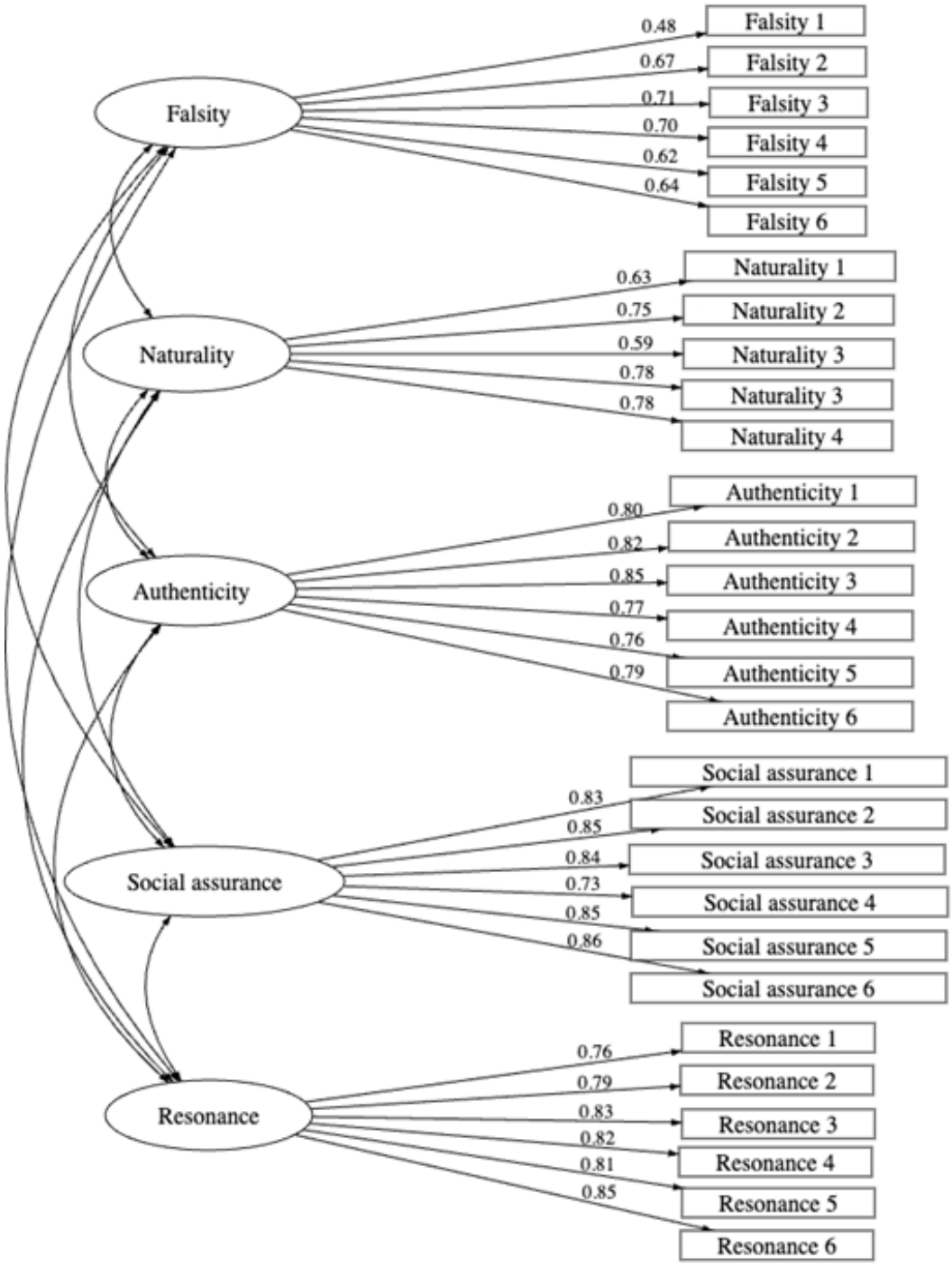

Relationships With External Variables

Zero-order correlations between the variables are provided in Multimedia Appendix 3. Table 2 presents the results of testing H2 to H6 and the RQ.

H2a predicted that falsity would be negatively associated with social media use (frequency and engagement), and H2b predicted that falsity would be positively associated with information seeking. H2a was rejected because falsity was positively linked to social media use frequency and engagement. On the other hand, falsity positively predicted information seeking. Other dimensions, including resonance, predicted information seeking. H2b was partially supported.

H3a expected that authenticity would be positively associated with the existing global measure of social media realism, and H3b expected that authenticity would be positively associated with eudaimonic enjoyment. Supporting H3a, authenticity was substantially related to the global realism measure, with other dimensions showing weaker associations. Authenticity positively predicted eudaimonic enjoyment, but so did resonance. These results provide partial support for H3b.

H4a hypothesized that naturality would be positively associated with hedonic enjoyment, and H4b hypothesized that naturality would be positively associated with escape motivation. Naturality was indeed positively associated with hedonic enjoyment and escape motivation. Hedonic enjoyment was also

XSL•FO
RenderX

predicted by other dimensions such as resonance. Escape motivation was related to other dimensions, including social assurance and falsity. H4a and H4b were partially supported.

H5a anticipated that resonance would be positively associated with social media network homophily, and H5b anticipated that resonance would be positively associated with reflective thinking. Resonance was positively associated with social network homophily. However, other dimensions, including authenticity and social assurance, were also linked to network homophily. H5a was partially supported. Supporting H5b, resonance showed a stronger positive association with reflective thinking than any other dimensions.

H6a hypothesized that social assurance would be positively associated with surveillance gratification, and H6b hypothesized that social assurance would not be associated with information seeking or reflective thinking. Social assurance was positively associated with surveillance gratification; other dimensions, including resonance, were also linked to surveillance. These findings partially support H6a. Consistent with H6b, social assurance showed no association with information seeking or reflective thinking.

The RQ asked whether the 5 dimensions would be differentially associated with addictive social media use. Social assurance was more strongly associated with addictive social media use than the other dimensions. Authenticity also showed a positive association. Falsity, naturality, and resonance were unrelated to addictive social media use.

Discussion

Summary

This survey study provides initial quantitative evidence that social media reality judgments comprise 5 dimensions—falsity, naturality, authenticity, resonance, and social assurance—and that these dimensions are differentially associated with the patterns of social media use.

Dimensions

Falsity was the strongest belief among the dimensions, as shown in Table 2. The strength of this schema observed in this study may reflect the finding of a recent global study on information literacy [4] that the majority of social media users are aware of the problem of misinformation. However, this belief did not deter people from using, and engaging with, social media. An alternative explanation may be that frequent social media users are more likely to encounter misinformation. Future research should examine this potential using a longitudinal design. By contrast, falsity did motivate users to seek information and engage in reflective thinking. The falsity schema may function as a general self-protective mechanism against potential harm on social media. The fact that falsity was related to escape motivation suggests that less purposeful and more passive use of social media [50] may present users with less-than-credible information.

Naturality is a unique dimension that emerged in this study, representing openness to artificiality in social media presentations. This schema was linked to hedonic enjoyment and escape motivation for social media use. Naturality may be particularly relevant when people use social media for entertainment and diversion purposes because they may be more willing to embrace the use of filters, digital editing and alterations, and preplanned and prechoreographed videos. As hedonic gratifications are central to social media use [51], it is important for future research to understand the mechanisms and effects of naturality. Furthermore, the schema of naturality can be useful for research on visual misinformation and visual information processing, with concerns growing over the increasing prevalence of deep fakes and cheap fakes [52].

Authenticity, as hypothesized, was more strongly associated with the current global measure of social media realism than the other dimensions. This finding highlights the necessity for multidimensional measures, such as the ones proposed in this study, for a more comprehensive understanding of user judgments of social media realism. Concurrently, the results show that authenticity plays important and multiple roles in social media use. It was positively linked to eudaimonic and hedonic enjoyment, but the belief that people's social media posts represent their true selves also predicted addictive social media use.

Resonance is another distinctive dimension identified in this study. It was a stronger predictor than the other dimensions of reflective thinking and information seeking after social media use. This finding corroborates the confirmation bias theory, which argues that people tend to search for, and focus on, information that supports their preexisting beliefs. Perhaps as a result, resonance facilitated social media engagement, but it was unrelated to the frequency of use. In light of these findings, future research should investigate how users' social media networks influence resonance. Network homophily was associated with resonance in this study. Future research should expand on this result to identify other network features.

Social assurance, a heuristic-based judgment, was unrelated to reflective thinking or information seeking. It was positively associated with surveillance motivation for keeping up with what is trendy. Social assurance was a stronger predictor than the other dimensions of addictive social media use, suggesting that it may have a less-than-healthy role in social media use. While unrelated to eudaimonic or hedonic enjoyment, social assurance was positively associated with escape motivation. This schema was positively linked to the frequency of social media use but not to engagement with social media. These results suggest that people depend on social assurance when they engage in passive use of social media.

The results offer further perspectives on the distinctiveness of the dimensions, as well as their connectedness. Although resonance and social assurance share a social-orientation focus, the results show that they function starkly differently. Resonance promoted reflective thinking and information seeking, while social assurance was unrelated to either. The former fostered eudaimonic and hedonic enjoyment, but the latter did not. Social assurance was strongly linked to addictive use, but resonance was not. Resonance predicted engagement, while social assurance was related to the frequency of use. Resonance is an internally sensed connection between a post and the user, whereas social assurance is a dependence on external cues.

XSL•FO
RenderX

Authenticity and naturality are both judgments about others, but the former focuses on internal thoughts and feelings, while the latter focuses on external features, of others. Authenticity was strongly associated with the existing global measure of social media realism, while naturality was not. Considering the important role of naturality evidenced in this study, more research is needed to examine this dimension and its effects on digital media uses and outcomes.

In addition, our findings suggest that resonance can be an important dimension influencing the effects of social media use on health-related outcomes. Beyond our predictions, resonance was linked to the gratifications and motivations related to eudaimonic well-being, hedonic pleasure, and surveillance. These associations demonstrate the importance of the sense of connection with others in social media use and its impact on how individuals evaluate digital information. Consistent with our expectation, resonance predicted reflective thinking more strongly than the other dimensions. This finding suggests the central role of resonance in influencing how individuals may engage with health information on social media. Collectively, these findings highlight the need for future research to closely examine the role of resonance in determining the evaluation of health information on social media. Resonance is likely based on one's prior experiences and beliefs, which can bias subsequent judgments and actions [53].

Limitations

This study has limitations. As it used a cross-sectional design, making a causal inference is difficult. Although the predictions were based on theory, findings should be interpreted keeping this limitation in mind. Future research could apply the measures to diverse platforms and to more specialized populations to examine how these differences may shape the schemas. The affordances of a given platform may promote the use of certain schemas rather than others. There may be differences in schema use across social media use experiences.

The results of this survey study provide a better understanding of social media reality judgments. The results further illustrate that the different schemas may coexist in users for different goals, gratifications, contexts, and outcomes. This understanding could be useful for future efforts for digital media literacy education to prevent the harmful effects of various forms of misinformation, including visual misinformation.

General Discussion

Principal Findings

Across the 2 studies, this research identified prominent user schemas for evaluating social media reality and investigated the schemas' association with social media use behaviors. These efforts aimed to develop and preliminarily validate a social media reality measure, an instrument designed to capture multifaceted user schemas used in assessing social media messages. Given the pervasiveness of misinformation, our goal was to provide a general measure that can be used in diverse domains of public health. The areas of application can include, but are not limited to, a wide range of public health issues such as drugs and smoking, communicable and noncommunicable diseases, diet and eating disorders, treatments and medical interventions, and mental health [2,3,7,9-11,14-19,21,24,42-44].

The development of these measures opens up possibilities for future research on social media and public health. The multifaceted user schemas suggest a new avenue for enhancing current efforts to counter misinformation through tailored digital media literacy initiatives. The 5 dimensions suggest that research may need to move beyond the current binary distinction of fake versus fact to obtain a more dynamic understanding of the user interpretations of social media information. On the basis of this understanding, user-centric tailored digital media literacy interventions can be developed.

Through formative research, interventions can specifically identify whether users have low or high levels of these schemas; for example, for users with high levels of beliefs on certain dimensions (eg, naturality, resonance, and social assurance), tailored interventions can seek to reduce the beliefs by providing facts, statistics, or anecdotal evidence. Conversely, for users with low levels of beliefs on dimensions such as falsity, a tailored intervention can focus on content that provides new information to increase these beliefs. For users with the desirable levels of beliefs (eg, low levels of beliefs on social assurance), interventions can focus on maintaining or reinforcing the beliefs. The optimal level of beliefs for each dimension will depend on the specific public health issue and context, necessitating a user-centric approach. The scales developed in this study can usefully assist these efforts.

In this approach, educators can also encourage users to critically examine and discern the various facets of a social media post and understand how the facets impact their judgments and actions; for instance, a funny TikTok video may be deemed strongly authentic, although not highly natural. A glamorous Instagram image may be perceived as natural yet lacking in authenticity. By challenging users to more critically examine their beliefs, educators can help users recognize how their existing beliefs and values shape their evaluations and improve their ability to differentiate social media information from reality.

By using these scales, researchers can also examine how users deal with diverse messages within the contexts of differential motivations and gratifications. The findings of this study (eg, users being more willing to embrace artificiality in social media presentations when motivated by escapism) can be used in educational efforts so that users can better understand how their evaluation is context dependent.

The findings suggest that the multiple functions served by each schema in various contexts and goals should be acknowledged. The results and the social media reality measures of this study can assist educators in assessing users' existing schemas and the context within which the schemas operate; for example, although perceived authenticity may provide enjoyment, it was weakly linked to the addictive use of social media. Therefore, users can be prompted to question their assumptions about authenticity. Likewise, users who hold beliefs about naturality may be obtaining hedonic enjoyment, but they may also be vulnerable to the growing presence of visual misinformation on social media, including deep fakes and cheap fakes.

XSL•FO
RenderX

The diametrically different functions of the 2 socially focused schemas should also be addressed. The use of the resonance schema can be rewarding, but its stronger association with reflective thinking and information seeking compared to falsity suggests that it may reinforce confirmation bias. Digital media literacy efforts can inform users of this tendency and encourage openness to different opinions. Users who rely on social assurance should critically examine the content they quickly accept based on metrics generated by others.

In addition, research could use the scales to examine the role of the schemas in the mechanisms of social media effects because different schemas, when primed, may activate different types of information processing and produce differential outcomes.

Although the initial form of the social media reality measure would benefit from further validation, it holds potential for future applications. To improve on our research, future research could use more diverse samples and use a longitudinal design that allows causal inference of the realism effects.

The results of this research advance our existing knowledge. They demonstrate that judgments of social media realism differ significantly from those of mass media realism [7,8]. The findings also indicate that the dimensions are broader than the current focus on authenticity [18-20] or social media metrics [23], encompassing aspects such as falsity, naturality, and resonance. Furthermore, the results illuminate how these different dimensions are differentially connected to health-relevant perceptions and practices, including information seeking and reflective thinking, social media network structures, and social media addiction. These insights suggest avenues for modification and correction. By unpacking the complexity of social media content evaluation, research can better develop tailored interventions and strategies that improve digital media literacy and foster healthier web-based engagement.

Conclusions

The schemas identified in this research contribute to a more nuanced understanding of how users interact with, and interpret, social media. We hope that these studies will stimulate further research toward a deeper and more dynamic understanding of user evaluations of social media information. Understanding where users stand regarding these schemas can be a basis for developing personalized digital media literacy education tailored to each individual's distinctive set of schemas and for proactively shielding users from the harmful influence of misinformation on social media.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 437 KB - jmir_v26i1e52058_app1.pdf]

Multimedia Appendix 2
Dimensions and items.
[DOCX File , 17 KB - jmir_v26i1e52058_app2.docx]

Multimedia Appendix 3
Bivariate correlations between variables.
[DOCX File , 19 KB - jmir_v26i1e52058_app3.docx]

References

1. Confronting health misinformation: the U.S. surgeon general's advisory on building a health information environment. U.S. Department of Health and Human Services. 2021. URL: https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf [accessed 2024-07-15] [WebCite Cache ID NBK572169]
2. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
3. Yeung AW, Tosevska A, Klager E, Eibensteiner F, Tsagkaris C, Parvanov ED, et al. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. J Med Internet Res 2022 Jan 25;24(1):e28152 [FREE Full text] [doi: 10.2196/28152] [Medline: 34951864]
4. A global study on information literacy: understanding generational behaviors and concerns around false and misleading information online. Poynter Institute for Media Studies, MediaWise, YouGov Inc., Google. 2022 Aug. URL: https://www.poynter.org/wp-content/uploads/2022/08/A-Global-Study-on-Information-Literacy-1.pdf [accessed 2024-07-14]
5. Drouin M, Miller D, Wehle SM, Hernandez E. Why do people lie online? "Because everyone lies on the internet". Comput Hum Behav 2016 Nov;64:134-142. [doi: 10.1016/j.chb.2016.06.052]
6. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 07. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-07-14]

XSL•FO
RenderX

7. Cho H, Shen L, Wilson K. Perceived realism: dimensions and roles in narrative persuasion. Commun Res 2012 Jul 12;41(6):828-851. [doi: 10.1177/0093650212450585]
8. Hall A. Reading realism: audiences' evaluations of the reality of media texts. J Commun 2003 Dec;53(4):624-641. [doi: 10.1111/j.1460-2466.2003.tb02914.x]
9. Livingston J, Holland E, Fardouly J. Exposing digital posing: the effect of social media self-disclaimer captions on women's body dissatisfaction, mood, and impressions of the user. Body Image 2020 Mar;32:150-154. [doi: 10.1016/j.bodyim.2019.12.006] [Medline: 32014705]
10. Bury B, Tiggemann M, Slater A. Disclaimer labels on fashion magazine advertisements: does timing of digital alteration information matter? Eat Behav 2017 Apr;25:18-22 [FREE Full text] [doi: 10.1016/j.eatbeh.2016.08.010] [Medline: 27591965]
11. Naderer B, Peter C, Karsay K. This picture does not portray reality: developing and testing a disclaimer for digitally enhanced pictures on social media appropriate for Austrian tweens and teens. J Child Media 2021 Jun 23;16(2):149-167. [doi: 10.1080/17482798.2021.1938619]
12. Egelhofer JL, Boyer M, Lecheler S, Aaldering L. Populist attitudes and politicians' disinformation accusations: effects on perceptions of media and politicians. J Commun 2022 Dec;72(6):619-632. [doi: 10.1093/joc/jqac031]
13. Ferrari E. Fake accounts, real activism: political faking and user-generated satire as activist intervention. New Media Soc 2017 Sep 21;20(6):2208-2223. [doi: 10.1177/1461444817731918]
14. Barlett CP, Rodeheffer C. Effects of realism on extended violent and nonviolent video game play on aggressive thoughts, feelings, and physiological arousal. Aggress Behav 2009;35(3):213-224. [doi: 10.1002/ab.20279] [Medline: 19280624]
15. Huesmann LR. The impact of electronic media violence: scientific theory and research. J Adolesc Health 2007 Dec;41(6 Suppl 1):S6-13 [FREE Full text] [doi: 10.1016/j.jadohealth.2007.09.005] [Medline: 18047947]
16. Cho H, Song CC, Adams D. Efficacy and mediators of a web-based media literacy intervention for indoor tanning prevention. J Health Commun 2020 Feb 01;25(2):105-114 [FREE Full text] [doi: 10.1080/10810730.2020.1712500] [Medline: 31984873]
17. Cho H, Yu B, Cannon J, Zhu YM. Efficacy of a media literacy intervention for indoor tanning prevention. J Health Commun 2018;23(7):643-651 [FREE Full text] [doi: 10.1080/10810730.2018.1500659] [Medline: 30058942]
18. Tiggemann M, Anderberg I. Social media is not real: the effect of 'Instagram vs reality' images on women's social comparison and body image. New Media Soc 2019 Nov 16;22(12):2183-2199. [doi: 10.1177/1461444819888720]
19. Cho H, Li W, Shen L, Cannon J. Mechanisms of social media effects on attitudes toward e-cigarette use: motivations, mediators, and moderators in a national survey of adolescents. J Med Internet Res 2019 Jun 27;21(6):e14303 [FREE Full text] [doi: 10.2196/14303] [Medline: 31250830]
20. Schwarz KC, Williams JP. Introduction to the social construction of identity and authenticity. In: Studies on the Social Construction of Identity and Authenticity. Milton Park, UK: Routledge; 2020.
21. Jenkins EL, Ilicic J, Barklamb AM, McCaffrey TA. Assessing the credibility and authenticity of social media content for applications in health communication: scoping review. J Med Internet Res 2020 Jul 23;22(7):e17296 [FREE Full text] [doi: 10.2196/17296] [Medline: 32706675]
22. Waddell TF, Sundar SS. Bandwagon effects in social television: how audience metrics related to size and opinion affect the enjoyment of digital media. Comput Hum Behav 2020 Jun;107:106270. [doi: 10.1016/j.chb.2020.106270]
23. Metzger MJ, Flanagin AJ, Medders RB. Social and heuristic approaches to credibility evaluation online. J Commun 2010 Sep;60(3):413-439. [doi: 10.1111/j.1460-2466.2010.01488.x]
24. Cho H, Shen L, Peng L. Examining and extending the influence of presumed influence hypothesis in social media. Media Psychol 2021;24(3):413-435 [FREE Full text] [doi: 10.1080/15213269.2020.1729812] [Medline: 34421390]
25. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of misinformation online. Proc Natl Acad Sci U S A 2016 Jan 19;113(3):554-559 [FREE Full text] [doi: 10.1073/pnas.1517441113] [Medline: 26729863]
26. Hazard PD, Klapper JT. The effects of mass communication. Am Q 1962;14(1):103-104. [doi: 10.2307/2710239]
27. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 1998 Jun;2(2):175-220. [doi: 10.1037//1089-2680.2.2.175]
28. Knobloch-Westerwick S, Mothes C, Polavin N. Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. Commun Res 2017 Jul 18;47(1):104-124. [doi: 10.1177/0093650217719596]
29. Munro GD, Stansbury JA. The dark side of self-affirmation: confirmation bias and illusory correlation in response to threatening information. Pers Soc Psychol Bull 2009 Sep 02;35(9):1143-1153. [doi: 10.1177/0146167209337163] [Medline: 19491331]
30. Asker D, Dinas E. Thinking fast and furious: emotional intensity and opinion polarization in online media. Public Opin Q 2019 Sep 09;83(3):487-509. [doi: 10.1093/poq/nfz042]
31. Aspers P, Corte U. What is qualitative in qualitative research. Qual Sociol 2019;42(2):139-160 [FREE Full text] [doi: 10.1007/s11133-019-9413-7] [Medline: 31105362]
32. Krueger RA, Casey MA. Focus Groups: A Practical Guide for Applied Research. Thousand Oaks, CA: SAGE Publications; 2014.
33. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]

XSL•FO
RenderX

34. Strauss A, Corbin J. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, Second Edition. Thousand Oaks, CA: SAGE Publications, Inc; 1998.
35. Smith RK, Yazdani E, Wang P, Soleymani S, Ton LA. The cost of looking natural: why the no-makeup movement may fail to discourage cosmetic use. J Acad Mark Sci 2021 Aug 10;50(2):324-337. [doi: 10.1007/s11747-021-00801-2]
36. Afifi WA, Weiner JL. Toward a theory of motivated information management. Commun Theory 2006 Jan 10;14(2):167-190. [doi: 10.1111/j.1468-2885.2004.tb00310.x]
37. Ryff CD. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. J Pers Soc Psychol 1989;57(6):1069-1081. [doi: 10.1037//0022-3514.57.6.1069]
38. Oliver MB, Raney AA. Entertainment as pleasurable and meaningful: identifying hedonic and eudaimonic motivations for entertainment consumption. J Commun 2011 Oct 03;61(5):984-1004. [doi: 10.1111/j.1460-2466.2011.01585.x]
39. Rubin AM, Perse EM. Audience activity and television news gratifications. Commun Res 1987;14(1):58-84. [doi: 10.1177/009365087014001004]
40. Farquhar LK, Meeds R. Types of fantasy sports users and their motivations. J Comput Mediat Commun 2007 Jul 01;12(4):1208-1228. [doi: 10.1111/j.1083-6101.2007.00370.x]
41. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. Annu Rev Sociol 2001 Aug;27(1):415-444. [doi: 10.1146/annurev.soc.27.1.415]
42. Caplan SE. Problematic internet use and psychosocial well-being: development of a theory-based cognitive–behavioral measurement instrument. Comput Hum Behav 2002 Sep;18(5):553-575. [doi: 10.1016/S0747-5632(02)00004-3]
43. Dogan H, Norman H, Alrobai A, Jiang N, Nordin N, Adnan A. A web-based intervention for social media addiction disorder management in higher education: quantitative survey study. J Med Internet Res 2019 Oct 02;21(10):e14834 [FREE Full text] [doi: 10.2196/14834] [Medline: 31579018]
44. Li W, Cho H. Modifiable factors associated with social media addiction: gratifications sought, social media realism, and social network characteristics. Commun Rep 2022 Sep 15;36(2):69-81. [doi: 10.1080/08934215.2022.2121976]
45. Ma L, Lee CS, Goh DH. Understanding news sharing in social media: an explanation from the diffusion of innovations theory. Online Inf Rev 2014;38(5):598-615. [doi: 10.1108/OIR-10-2013-0239]
46. Chen YC. The effectiveness of different approaches to media literacy in modifying adolescents' responses to alcohol. J Health Commun 2013 Jun;18(6):723-739. [doi: 10.1080/10810730.2012.757387] [Medline: 23496333]
47. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
48. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociol Methods Res 1992;21(2):230-258. [doi: 10.1177/0049124192021002005]
49. Raftery AE. Bayesian model selection in social research. Sociol Methodol 1995;25(1995):111-163. [doi: 10.2307/271063]
50. Escobar-Viera CG, Shensa A, Bowman ND, Sidani JE, Knight J, James AE, et al. Passive and active social media use and depressive symptoms among United States adults. Cyberpsychol Behav Soc Netw 2018 Jul;21(7):437-443. [doi: 10.1089/cyber.2017.0668] [Medline: 29995530]
51. de Oliveira Santini F, Ladeira WJ, Pinto DC, Herter MM, Sampaio CH, Babin BJ. Customer engagement in social media: a framework and meta-analysis. J Acad Mark Sci 2020 May 27;48:1211-1228. [doi: 10.1007/s11747-020-00731-5]
52. Paris B, Donovan J. Deepfakes and cheap fakes: the manipulation of audio and visual evidence. Data & Society. 2019 Sep 18. URL: https://datasociety.net/library/deepfakes-and-cheap-fakes/ [accessed 2024-07-15]
53. Cho H, Cannon J, Lopez R, Li W. Social media literacy: a conceptual framework. New Media Soc 2024 Feb;26(2):941-960 [FREE Full text] [doi: 10.1177/14614448211068530] [Medline: 38328341]

Abbreviations

CFA: confirmatory factor analyses
COREQ: Consolidated Criteria for Reporting Qualitative Research
RQ: research question

Edited by T Leung, G Eysenbach; submitted 21.08.23; peer-reviewed by A Faux-Nightingale, M Zhou; comments to author 06.09.23; revised version received 26.09.23; accepted 01.07.24; published 06.08.24.

Please cite as:
Cho H, Li W, Lopez R
A Multidimensional Approach for Evaluating Reality in Social Media: Mixed Methods Study
J Med Internet Res 2024;26:e52058
URL: https://www.jmir.org/2024/1/e52058
doi:10.2196/52058
PMID:39106092

XSL•FO
RenderX

©HyunYi Cho, Wenbo Li, Rachel Lopez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

True Mitotic Count Prediction in Gastrointestinal Stromal Tumors: Bayesian Network Model and PROMETheus (Preoperative Mitosis Estimator Tool) Application Development

Salvatore Lorenzo Renne[1,2*], MD; Manuela Cammelli[1*], MD; Ilaria Santori[1], MD; Marta Tassan-Mangina[1], MD; Laura Samà[3], MD; Laura Ruspi[3], MD; Federico Sicoli[3], MD; Piergiuseppe Colombo[1,2], MD; Luigi Maria Terracciano[1,2], MD; Vittorio Quagliuolo[3], MD; Ferdinando Carlo Maria Cananzi[1,3], MD

[1]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

[2]Pathology Department, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital, Rozzano, Milan, Italy

[3]Sarcoma, Melanoma and Rare Tumors Surgery Unit, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital, Rozzano, Milan, Italy

[*]these authors contributed equally

Corresponding Author:
Salvatore Lorenzo Renne, MD
Department of Biomedical Sciences
Humanitas University
Via Rita Levi Montalcini 4
Pieve Emanuele, Milan, 20072
Italy
Phone: 39 0282247743
Fax: 39 02 82244791
Email: salvatore.renne@hunimed.eu

Abstract

Background: Gastrointestinal stromal tumors (GISTs) present a complex clinical landscape, where precise preoperative risk assessment plays a pivotal role in guiding therapeutic decisions. Conventional methods for evaluating mitotic count, such as biopsy-based assessments, encounter challenges stemming from tumor heterogeneity and sampling biases, thereby underscoring the urgent need for innovative approaches to enhance prognostic accuracy.

Objective: The primary objective of this study was to develop a robust and reliable computational tool, PROMETheus (Preoperative Mitosis Estimator Tool), aimed at refining patient stratification through the precise estimation of mitotic count in GISTs.

Methods: Using advanced Bayesian network methodologies, we constructed a directed acyclic graph (DAG) integrating pertinent clinicopathological variables essential for accurate mitotic count prediction on the surgical specimen. Key parameters identified and incorporated into the model encompassed tumor size, location, mitotic count from biopsy specimens, surface area evaluated during biopsy, and tumor response to therapy, when applicable. Rigorous testing procedures, including prior predictive simulations, validation utilizing synthetic data sets were employed. Finally, the model was trained on a comprehensive cohort of real-world GIST cases (n=80), drawn from the repository of the Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital, with a total of 160 cases analyzed.

Results: Our computational model exhibited excellent diagnostic performance on synthetic data. Different model architecture were selected based on lower deviance and robust out-of-sample predictive capabilities. Posterior predictive checks (retrodiction) further corroborated the model's accuracy. Subsequently, PROMETheus was developed. This is an intuitive tool that dynamically computes predicted mitotic count and risk assessment on surgical specimens based on tumor-specific attributes, including size, location, surface area, and biopsy-derived mitotic count, using posterior probabilities derived from the model.

Conclusions: The deployment of PROMETheus represents a potential advancement in preoperative risk stratification for GISTs, offering clinicians a precise and reliable means to anticipate mitotic counts on surgical specimens and a solid base to stratify patients for clinical studies. By facilitating tailored therapeutic strategies, this innovative tool is poised to revolutionize clinical

XSL•FO
RenderX

decision-making paradigms, ultimately translating into improved patient outcomes and enhanced prognostic precision in the management of GISTs.

(J Med Internet Res 2024;26:e50023) doi:10.2196/50023

KEYWORDS

GIST mitosis; risk classification; mHealth; mobile health; neoadjuvant therapy; patient stratification; Gastrointestinal Stroma; preoperative risk

Introduction

Gastrointestinal stromal tumor (GIST) is the most common sarcoma type [1]. The majority harbors activating mutation in *KIT* or *PDGFRA* [2-5]. Even if these mutations represent early events in carcinogenesis—being shared by clinically irrelevant and very aggressive GIST [6,7]—they are the molecular basis for the very active tyrosine-kinase-inhibitor (TKI) therapy [8-10]. TKIs have revolutionized GIST treatment and have been used in all the disease stages since their introduction [11-14]. Notably, they can be administered as neoadjuvant treatment for patients with high-risk disease or for reducing the extent of surgery in GIST in a peculiar location (ie, rectum and duodenum) [14-17]. While neoadjuvant therapy with Imatinib is beneficial for patients with high-risk disease [18] and for those who may not achieve R0 resection or can undergo less mutilating, function-sparing surgery if there is a volumetric reduction [17], TKIs present certain challenges. One major issue is that they may impair accurate postoperative risk assessment, as tumor response to therapy can prevent correct risk evaluation [5]. Several risk assessments in GIST have been developed to identify size, site, and mitotic count as important features [19-21]. Indeed, mitotic count on the surgical specimen after TKI therapy can be greatly modified—especially in the case of tumor response. Therefore, in these patients, the subsequent management will be guided by a risk assessment computed with the mitotic count from the biopsy, and this can lead to mistreatment.

On the other hand, during initial patient management, the correct identification of high-risk patients can also fail: a very small amount of tissue on biopsy is required to make the GIST diagnosis since good immunohistochemical markers (rather specific and sensitive) exist [22-25]. Therefore, whereas size and site can be accurately assessed by imaging, mitotic count on biopsy can face several limitations, some purely biological (such as tumor heterogeneity) and others more physical (ie, the size of the specimen available for counting, which is a classic example of sampling bias) [26-30]. Thus, it may happen to incorrectly classify the risk of a GIST preoperatively, and this might lead to surprises after the mitotic count on the surgical specimen is performed, often due to an underestimation of the mitotic count [31-33].

These limitations underscore the critical need for innovative approaches to refine preoperative risk stratification in GIST, aiming to mitigate the risks of misclassification and subsequent therapeutic mismanagement. The discrepancies between preoperative risk assessments and postoperative findings underscore the imperative for precision tools that can dynamically estimate mitotic count on surgical specimens, enhancing the accuracy of patient stratification and treatment planning.

In line with this imperative, we aim to develop an advanced computational tool, termed PROMETheus (Preoperative Mitosis Estimator Tool), designed to predict mitotic count on surgical specimens. By leveraging state-of-the-art Bayesian modeling techniques and integrating comprehensive clinicopathological variables, PROMETheus seeks to address the limitations of current risk assessment methodologies, offering clinicians a reliable means to anticipate postoperative mitotic counts and refine preoperative treatment strategies effectively.

Methods

Modeling Strategy

Bayesian Network and Workflow

As a modeling strategy, we used the Bayesian network with the aim of predicting the mitotic count on the surgical specimen. Of note, we use the term “Bayesian network” to indicate the model’s graphical representation and the collection of functions necessary to use it for statistical learning. This loose definition is often used in practice; however, it is broader than the one defined by the term’s inventor, Judea Pearl. In his 2009 book *Causality*, he identifies the directed acyclic graphs (DAGs) with the term Bayesian networks: “Directed graphs, especially DAGs, have been used to represent causal or temporal relationships…and came to be known as Bayesian networks, a term coined…to emphasize three aspects: (1) the subjective nature of the input information; (2) the reliance on Bayes’ conditioning as the basis for updating information; and (3) the distinction between causal and evidential modes of reasoning [34]. In this paper, the meaning of “Bayesian network” is closer to the one of structural causal models (SCMs) [35]. Briefly, we designed an SCM of the variables, created a mock data set, wrote a probabilistic program, validated it on the data simulation, fit the model to the data, and compared multiple models with different structures; these procedures are often collectively referred as the Bayesian workflow [36-39].

Causal Modeling

The graphical representation of the Bayesian network was done with DAGs: the variables were represented by nodes and the conditional dependencies through directed edges. Based on the graphical representation, we built an SCM, which we used for data simulation and as the model for the fits.

Probabilistic Programming

We wrote and fit the models using R software (version 4.1.2; R Foundation for Statistical Computing) and Stan (version

XSL•FO
RenderX

2.21.0; Stan Development Team) [40-43]. Stan is a probabilistic programming language that runs a No U-Turn sampler, an extension to Hamiltonian Monte Carlo (HMC) sampling, which is itself a form of Markov Chain Monte Carlo (MCMC) [44]. To promote regularization and reduce overfitting, we used a multilevel-hierarchical modeling strategy [39,45,46]. Distributions (likelihood and priors) were chosen with maximum entropy criteria [39]. To understand priors' implications, we run prior predictive simulations [37,38,46]. To minimize divergent transitions, we reparametrized the models with a noncentered equivalent form when appropriated [39,47]. To ensure a good representation of the sample space, we visually inspected the chains with trace plots and trankplots [48]. We then monitored the chains with postmodeling diagnostics such as the number of effective samples and the Gelman-Rubin convergence diagnostic .

[45,48]. Of note, . is defined in the cited references is different from the classic definition by Gelman and Rubin (1992). All models' fits were plotted against the fitted data to ensure a good representation of the outcome space (posterior predictive check) [39,49,50]. Compatibility intervals (CIs) were calculated as the highest posterior density interval (HDPI) [51].

Model Selection

For the aim of pure prediction—as in our case—the best model can be selected based on information theory by estimating model performance using the widely applicable information criterion (WAIC, a generalization of Akaike information criteria) and Pareto Smooth Important Sampling Leave-One-Out Cross-Validation (PSIS-LOO-CV) criteria [52-54]. We checked that the 2 statistics gave the same results, leading us to trust their results [39,55,56]. We then selected the model with the lowest deviance in out-of-sample performance.

Forecasting

The posterior probability density of the coefficients was then used to create an application that, given the chosen variables, computed the posterior probability distribution in the outcome space (ie, the mitotic count on the specimen). Moreover, we programmed the application to calculate the risk class from this computed posterior distribution of the mitotic count.

Study Population

Data Set

The cases came from a prospectively maintained database including all the patients who underwent surgery for primary sarcoma in the Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital (Milan, Italy). This database comprises extended clinical and pathological information and contains 233 GISTs operated from January 2000 to March 2022.

Inclusion and Exclusion Criteria

We included patients who had preoperative diagnostic biopsy and underwent surgical resection, with informed consent for research and available histologic material from both the biopsy and the surgical specimen. The histology was reviewed by a sarcoma pathologist (author SLR), and cases with diagnosis other than GIST were excluded.

Pathology

Microscope Calibration

We calibrated the microscope with a stage micrometer slide and calculated the number of high-power fields (HPFs) needed to reach the size of 5 mm^2. In line with the published guidelines, the number of HPFs to evaluate was 23.5 [30,57-59].

Mitosis

We defined mitosis as basophilic, dark, hairy material representing the chromosomes. Mitosis was counted when the chromosomes were either clotted (as at the beginning of metaphase), in a plane (as in metaphase and anaphase), or in separate clots (as in telophase), as previously described [60].

Biopsy Measurement

We measured the surface available under the microscope counting HPFs filled by neoplastic specimens up to 5 mm^2 for very small biopsies we approximated the surface as a fraction of a field of view.

Tumor Response

Some of the series cases underwent preoperative therapy. To use these cases without polluting the estimate for mitotic count coefficient, we also recorded the response to treatment; this had a different meaning from a classical pathological response and was defined as follows: if a mitotically active area was identified on the surgical specimen (regardless of the size) and the mitotic count in this area was equal to or more than the biopsy count, the tumor was classified as "no response." Therefore, this mitotic count on the biopsy was used in the model as if the case did not undergo preoperative therapy. Conversely, if the count on the surgical specimen was less than the count on the biopsy, the tumor was classified as "response" and the count was used to estimate a different coefficient that we did not use for prediction (see model description in the Results section and custom code for greater detail).

Ethical Considerations

All patients signed an institutional written informed consent to research. The Independent Ethics Committee of IRCCS Humanitas Clinical Research reviewed the following project, giving approval on February 21, 2023 (145/23). Data were deidentified before analysis.

Results

Causal Modeling

We designed a causal model to identify the covariates to be included in the model for estimating the mitotic count on the surgical specimen. We assumed that the true mitotic count is the one counted on the surgical specimen (M^S) The more tumor cell replicates, the bigger the tumor is (D). We also identified the anatomical location (L) as a cause of tumor dimension (D), in the sense that in certain locations the symptoms would appear earlier, thus influencing the measured size at diagnosis.

XSL•FO
RenderX

Moreover, location directly causes the total amount of tissue available for evaluation at biopsy (ie, the measured biopsy surface, S); some sites are more difficult to reach than others. Finally, mitotic count on the biopsy (M^B) reflects the mitotic count on the surgical specimen (M^S) and due to tumor heterogeneity and sampling bias, the measurement on the biopsy also depends on the surface examined (S; Figure 1 and Multimedia Appendix 1 for an extended version).

Figure 1. To determine the model covariates, we utilized a directed acyclic graph (DAG). In this causal framework, several factors influence the mitotic count on the surgical specimen. First, the dimension of the tumor impacts both the mitotic count on the biopsy and the surgical specimen. Larger tumors tend to exhibit higher mitotic activity. Second, the location of the tumor plays a crucial role in its growth pattern. For instance, gastric neoplasms often have more space to expand, leading to the development of symptoms with larger masses. Additionally, tumor location influences the accessibility of the biopsy site, as some sites are inherently more accessible than others. Finally, the amount of surface area available on the biopsy directly impacts the accuracy of the mitotic count estimation. A larger surface area allows for a more representative sampling of the tumor. This simplified causal model elucidates the relationship between various factors affecting GIST mitotic count. Please refer to the main text and supplementary materials for further details. In the model notation,Drepresents the dimension of the tumor, L denotes the location, M^B signifies the mitotic count on the biopsy, M^S indicates the mitotic count on the surgical specimen, and S represents the surface area of the biopsy.

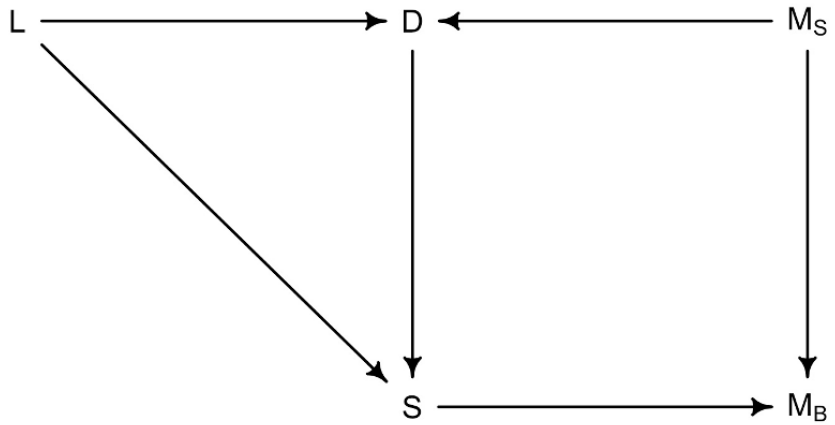

Probabilistic Modeling

Our inferential target was the mitotic count on the surgical specimen (M^S); as the name suggests, it is a count variable; therefore, we chose a Poisson distribution to model it, as in Equation (1). Poisson distributions have just 1 parameter, λ. It is the expected value and the expected variance of the count variable, and it is the parameter used for the generalized linear model. It needs to be positive, and a common link function is to exponentiate the model. Given each patient i, we estimated a coefficient for the tumor dimension (β) and for surface of the biopsy (γ) for each location (L) using a multilevel-hierarchical model for both of them ($\beta_{[L]}D_i$ and $\gamma_{[L]}S_i$ respectively); the parameters δ and [illegible] were alternatively switched on and off by the presence of response to therapy R, as defined in the methods section; finally, we set an intercept a, as in Equation (2). To justify the prior choice, we used prior predictive simulation.

Prior Predictive Simulation

The majority of GISTs are clinically irrelevant, with a mitotic count less than 5/5 mm^2. A few of them can have a higher mitotic count, even to a greater order of magnitude; however, it is biologically implausible to expect many cases with a mitotic count greater than 50. Using this field-specific knowledge, we chose normal distributions for model coefficients, as seen in Equations (3), (5), (4), (7), (8), (10), and (11), and exponential distributions for scalar coefficients, as seen in Equations (6) and (9). Through serial simulations, we narrowed the numerical values. For a graphical representation of part of the coefficients, see Multimedia Appendix 2). The results of the prior predictive simulation show that most of the simulated λ have values that conform to the field-specific knowledge (Figure 2).

XSL•FO
RenderX

Figure 2. Prior predictive simulation of the λ parameter (ie, what the model anticipates before encountering the data). This plot displays 80 simulated cases derived from the priors. The majority of the expected values indicate a very low mitotic count, aligning with real-world expectations. However, the model is not startled by higher mitotic counts, even though it anticipates encountering them in a minority of cases without prior training.

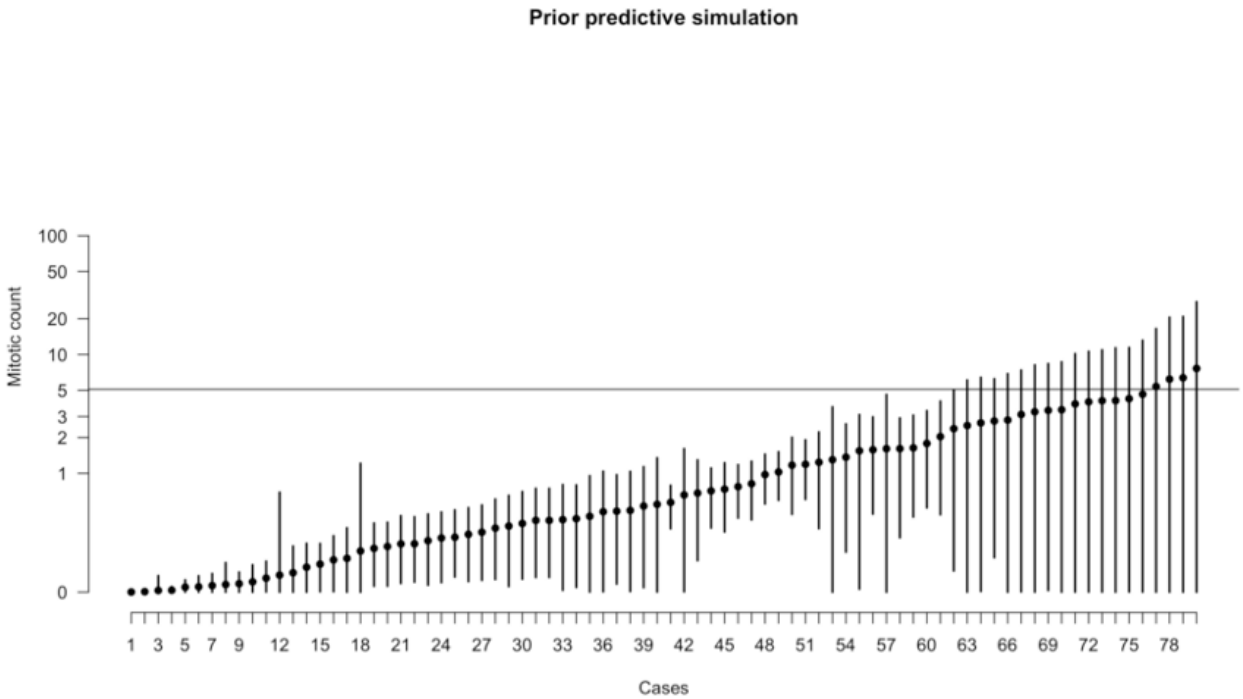

Fitting the Mock Data Set

To test the model performance, we fitted it on a simulated data set of 100 cases. The custom code to produce the mock data set is available on the cited repository. The fit with a centered parameterization resulted in 2% divergent transitions; therefore, we rewrote the model in a noncentered form. The fit diagnostics were satisfactory: ⊠. was obtained for all the parameters, and the energy from the Hamiltonian had a Gaussian outlook and the trankplots of the log-probability showed a satisfactory convergence of the chains (Multimedia Appendix 3). All the parameters had a satisfactory number of effective samples and a good outlook of the trankplots (Multimedia Appendix 4). The posterior probability density for each coefficient is depicted in Multimedia Appendix 5. To check the model fitness, we compared the inferred λ by the model to the true λ used for data simulation. This procedure revealed that the model regularized the values within each modeled site and was able to recover a value closer to λ even if provided with lower values (Figure 3).

XSL•FO
RenderX

Figure 3. Posterior predictive simulation. The upper panel shows the imputed λ parameter for each sample against the ground truth; consider the logarithmic scale on the y-axis. The lower panel shows the same imputed λ parameter against the fitted data. On average, the mitotic count on either the surgical specimen or the biopsy is lower than the expected value. This is due to the asymmetry of the Poisson distribution. The size of the X is proportional to the biopsy surface, and the size of the empty dots is proportional to the tumor size. The color of the background corresponds to the tumor site. NAC: N-acetylcysteine.

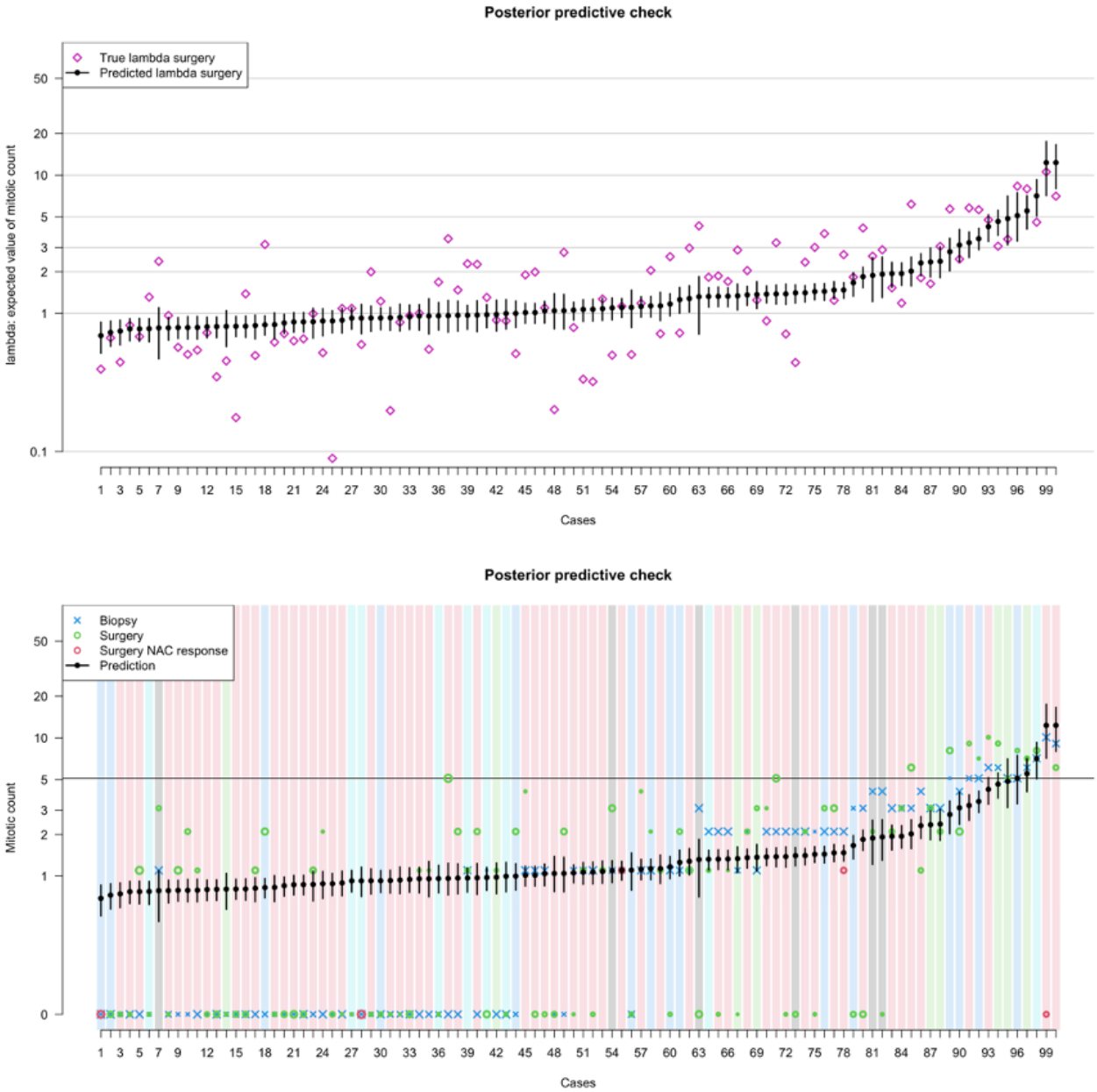

Fitting the Real Data Set

We then fitted the model to the real data. We were able to fit the data with the centered version of the model. Similarly to the simulated dataset, the fit's diagnostics were satisfactory: $\hat{R}$. was obtained for all the parameters. The energy from the Hamiltonian had a Gaussian outlook and the trankplots of the log-probability showed a satisfactory convergence of the chains (Multimedia Appendix 6). Moreover, all the parameters had a satisfactory number of effective samples and a good outlook of the trankplots (Multimedia Appendix 7); the posterior probability density for the model coefficient is shown in Multimedia Appendix 8 and Table 1. The site (L) was an index variable with values from 1 to 4, representing coefficients for the colon-rectum, duodenum, small intestine, and stomach, respectively. Of note, the model posterior distribution for the biopsy count parameter (δ) is 1, consistent with our expectations. As in the data simulation, to see what the model learned, we moved to form the parameter space to the outcome space and simulated from the whole posterior distribution a λ parameter—that tells us the expected value of the Poisson distribution—for each case and plotted it against the data fitted. This posterior predictive simulation is depicted in Figure 4.

XSL•FO
RenderX

Table 1. Coefficients.

Parameter	Mean (SD)	5.5%	94.5%	N eff	
$\beta_{L[1]}$	3.81 (1.91)	1.40	7.10	4396	1
$\beta_{L[2]}$	0.12 (0.15)	–0.12	0.36	8407	1
$\beta_{L[3]}$	0.30 (0.16)	0.03	0.55	5850	1
$\beta_{L[4]}$	0.40 (0.04)	0.35	0.46	9032	1
$\gamma_{L[1]}$	0.38 (0.44)	–0.25	1.13	6905	1
$\gamma_{L[2]}$	–0.09 (0.15)	–0.33	0.15	8769	1
$\gamma_{L[3]}$	–1.22 (0.34)	–1.77	–0.70	5884	1
$\gamma_{L[4]}$	–0.17 (0.06)	–0.26	-0.08	9371	1
α	1.75 (0.05)	1.67	1.83	10178	1
	–0.85 (0.20)	–1.17	–0.53	8664	1
	0.16 (0.05)	0.08	0.23	7978	1
δ	1.01 (0.08)	0.89	1.13	8031	1
	–0.90 (0.20)	–1.22	–0.57	9447	1
σ_γ	0.89 (0.36)	0.44	1.55	6856	1
σ_β	2.07 (0.74)	1.10	3.40	5274	1

Figure 4. Posterior predictive simulation on the true data set. The size of the X is proportional to the biopsy surface, the size of the empty dots is proportional to the tumor size; the color of the background corresponds to the tumor site.

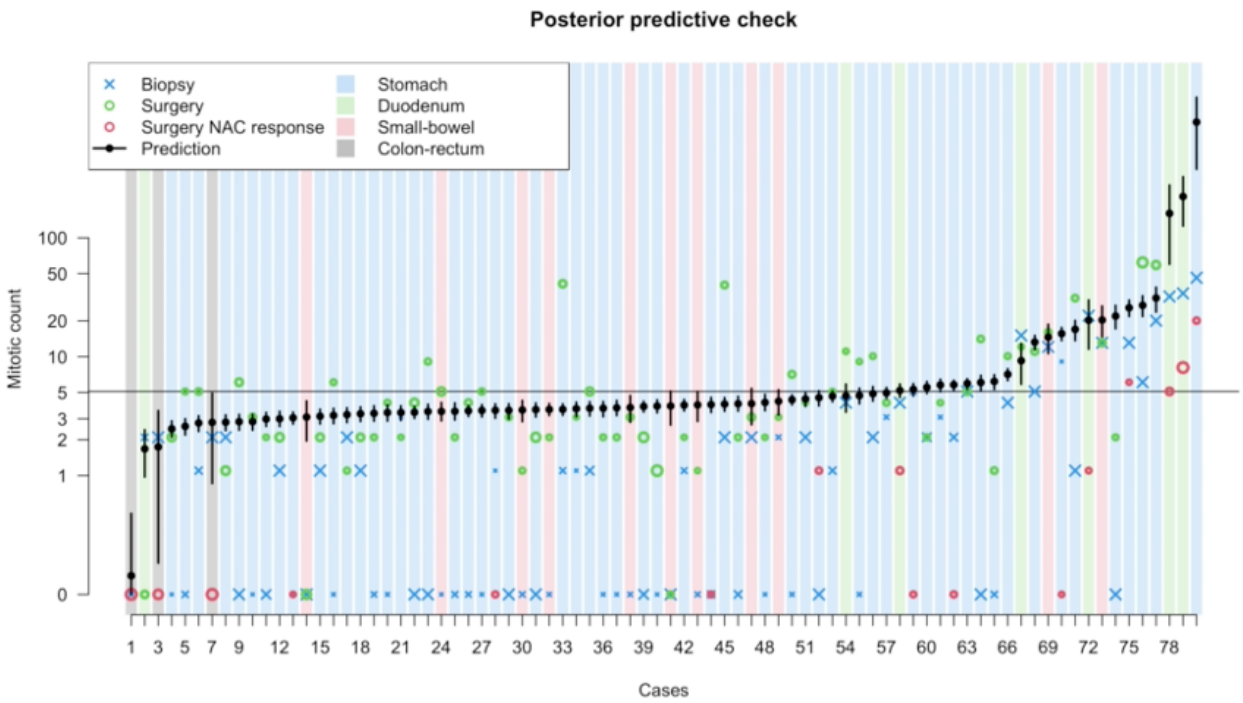

Model Selection

In general, adding parameters in the multilevel modeling reduces overfitting and improves the out-of-sample performance, whereas adding parameters without a hierarchical structure can reduce deviance within the sample, but results in lower out-of-sample performance (ie, increases overfitting). We therefore computed the deviance using WAIC and PSIS-LOO-CVC on our model (which had a multilevel hierarchical structure for the tumor size (β_L) and biopsy surface

XSL•FO
RenderX

(γ_L) parameters. We also evaluated alternative models: one incorporating mitotic count on the biopsy with a hierarchical structure, found in Equation 12 in Multimedia Appendix 9; one with only tumor size with a hierarchical structure, as in Equation 13 in Multimedia Appendix 9; one with only tumor size without accounting for biopsy surface, as in Equation (14) in Multimedia Appendix 9; and one without a hierarchical structure without tumor size, as in Equation (15) in Multimedia Appendix 9. The model with a hierarchical structure for tumor size and biopsy surface parameters, as in Equation (2), had the lowest out-of-sample deviance (Figure 5). Therefore, we chose the model with a hierarchical structure for the tumor size and biopsy surface parameters to develop the application.

Figure 5. Model selection. The vertical line indicates the mean deviance of the reference model (the one with the lowest out of sample deviance). The filled dots are values within the sample. The empty dots represent the mean out-of-sample deviance with the bar indicating the 89% CI; the triangle is the contrast between the model and the reference model.

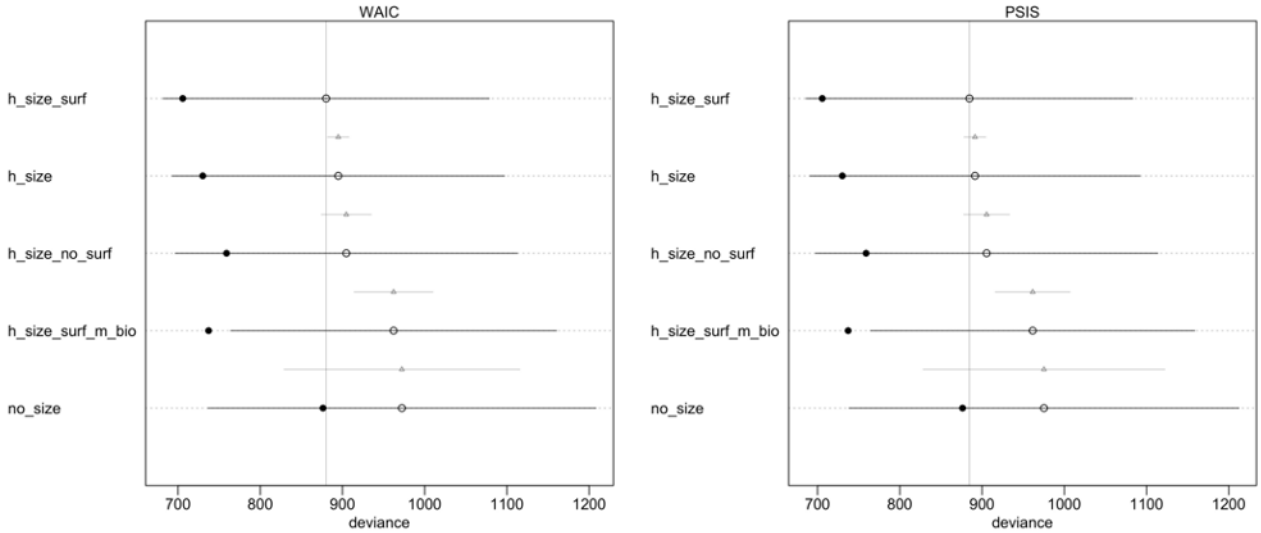

Application Development

Using the posterior probabilities from the model, we developed PROMETheus, a web-based application [61]. The user interface has an input panel to insert the tumor location, tumor size, the mitotic count on the biopsy, and the available surface on the biopsy. Then, the application dynamically computes the risk class according to Miettinen and Lasota. Moreover, using the inputted data and the full posterior distribution, the application computes the expected mitotic count on the surgical specimen (indicating the most probable rendered counts) and the predicted risk class for the new posterior distribution of M^S provided (Figure 6).

XSL•FO
RenderX

Figure 6. The PROMETheus (Preoperative Mitosis Estimator Tool) app. The image shows a screenshot for a gastrointestinal stromal tumor (GIST) measuring 72 mm with a biopsy surface of 14.3 high-power fields (HPFs; about 3 mm^2). With these biological characteristics, the risk class is low. However, the model shows that the mitotic count is likely to be underestimated, with the most probable count on the surgical specimen predicted to be between 5 to 8 mitoses in 5 mm^2. Given this predicted mitotic count distribution, it is much more probable for the risk class to be high (more than 60%).

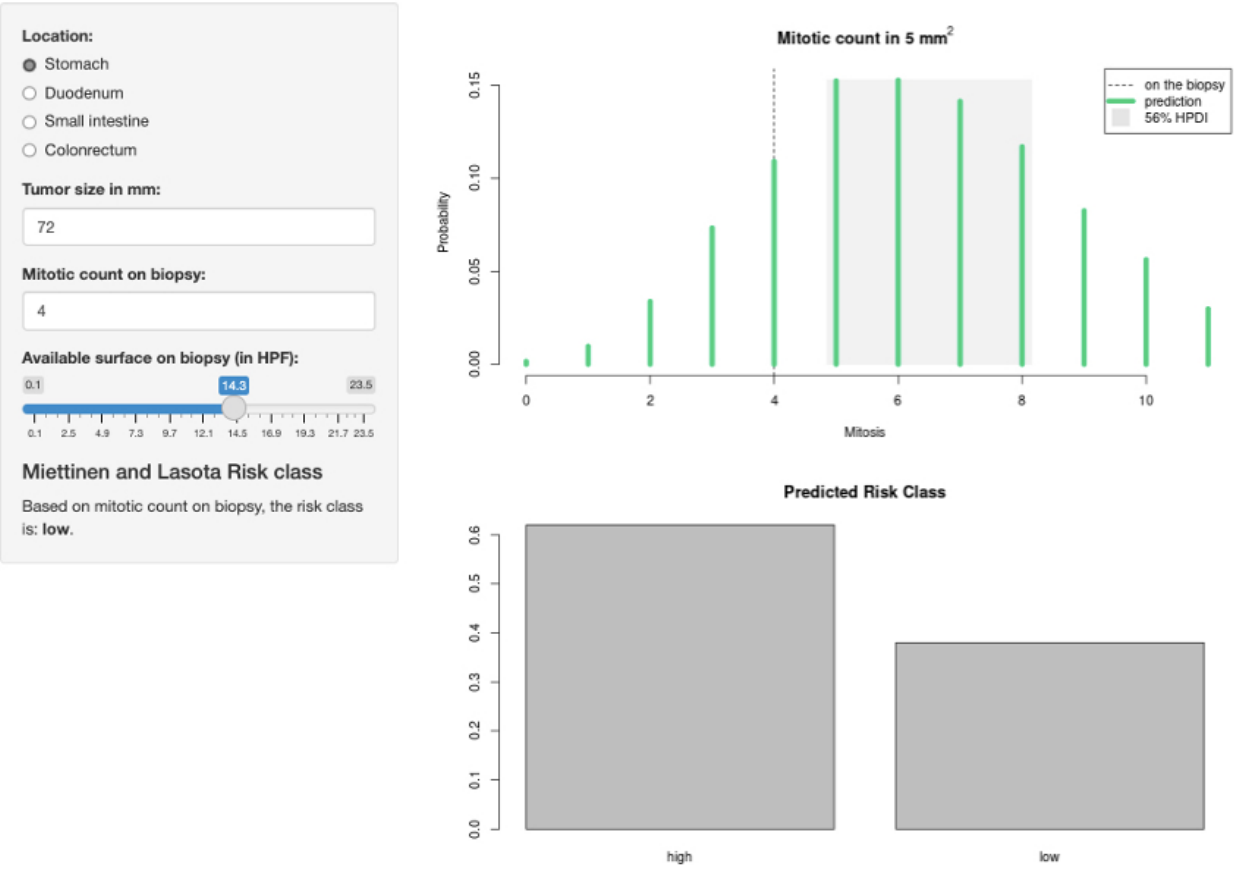

Development of the Preoperative Classification for GIST

To facilitate stage-adapted treatment planning and enhance data comparison across institutions, we developed a preoperative classification system for GIST based on the results of our computational model, PROMETheus. This classification system aims to provide a standardized definition of preoperative classification for GISTs, which can be used to guide therapeutic decisions and improve clinical trial designs (Table 2). The development of this classification included the probabilistic output of PROMETheus risk classification, important surgical parameters (such as site and size), and respectability.

Table 2. Proposed preoperative classification for GIST[a].

Clinical type	GIST's resectability	Features
0	Resectable	• Gastric, mini tumor (<2 cm), PROMETheusb low risk
1	Resectable	• Gastric, small tumor (2-10 cm), PROMETheus low risk
2	Resectable locally advanced	• Large tumor (gastric: >10 cm; nongastric: >5 cm) • Nongastric site • Elevated risk of intraoperative rupture • Need to mutilating or complex multivisceral resection • PROMETheus high risk
3	Resectable metastatic	• Resectable synchronous metastatic disease
4	Unresectable	• Unresectable nonmetastatic disease

[a]GIST: gastrointestinal stromal tumor.

[b]PROMETheus: Preoperative Mitosis Estimator Tool.

XSL•FO
RenderX

Discussion

Principal Findings

Our study aimed to develop an innovative computational tool, PROMETheus, to accurately predict mitotic count on surgical specimens of GISTs, thereby addressing the challenges associated with preoperative risk assessment and treatment planning. The primary objective was to bridge the gap between clinical judgment and computed risk class, empowering clinicians to make informed decisions in complex scenarios. By using Bayesian networks and rigorous covariate selection methodologies, we succeeded in achieving this objective, providing clinicians with a novel approach to preoperative risk stratification in GISTs.

Our findings represent a potential advancement in the field of oncology, particularly in the context of sarcomas, where precise risk assessment is paramount for optimal treatment outcomes. By accurately predicting the mitotic count on surgical specimens, PROMETheus offers clinicians a tool to navigate the complexities of GIST management, enabling tailored treatment strategies based on individual patient characteristics. This approach aligns with the evolving paradigm of precision medicine, where treatment decisions are increasingly guided by molecular and pathological insights.

Mitotic count serves as a vital indicator of biological aggressiveness in oncology, playing a significant role in the grading systems of various tumors. With chemotherapy often administered to patients with high-grade tumors, mitotic count becomes a de facto predictive biomarker, providing valuable insights into treatment response. In the realm of GISTs, alongside size and site, the number of mitoses holds the utmost importance in current risk classifications [19-21]. However, unlike size and site, which can be easily assessed via imaging or endoscopy in the preoperative setting, accurately determining mitotic count poses challenges. Factors such as tumor heterogeneity and limited specimen size for counting (a classic example of sampling bias) hinder precise estimation during preoperative biopsy.

The introduction of effective therapies like Imatinib, which targets the molecular alterations driving GIST, has led to a clinical tendency to use this treatment preoperatively, even in cases where the risk of disease progression is not high. However, this approach has several drawbacks: (1) variable response: not all patients exhibit a decrease in tumor size, which means the extent of the surgery may remain unchanged; (2) side effects: Imatinib is not without side effects, highlighting the need for a tool to identify patients at genuine risk of metastasis who would benefit from neoadjuvant treatment; (3) risk classification after treatment: for patients treated with neoadjuvant Imatinib, it becomes challenging to perform risk classification on the surgical specimen, leaving uncertainty about the subsequent adjuvant therapy.

Integrating PROMETheus into clinical practice could have far-reaching implications for patient care. Not only does it enable more precise risk stratification in GISTs, but it could also facilitate stage-adapted treatment planning, ensuring that patients receive the most appropriate interventions based on their individual risk profiles. The preoperative classification system shown in Table 2 offers a common language for patient follow-up and treatment planning. It is a revision of previous work by our group [26], refined based on the insights gained from our study. This classification could lay the foundation for improved data collection and comparison across institutions, fostering collaboration and advancing research efforts in the field of GIST management.

Comparison to Prior Work

Importantly, our study diverges from previous approaches that primarily focused on comparing biopsy and surgical specimens to evaluate the reliability of biopsy-based risk assessments. Instead, we recognized the inherent limitations of preoperative biopsy in predicting tumor grade, especially in the context of TKI therapy–induced tumor response. By focusing on predicting the mitotic count directly, our approach transcends these limitations, providing clinicians with a more accurate and reliable tool for preoperative risk assessment.

There are many tools for risk stratification in oncological practice, including in the sarcoma field, mainly based on nomograms, which are usually used to predict overall survival and the risk of metastasis [62]. While these tools have found utility in reanalyzing previous clinical studies and selecting patients for future trials [63], they are not tailored for GISTs. Our innovative approach stands out in 2 crucial ways. First, our Bayesian methodology empowers clinicians by providing the full posterior probability, capturing the inherent uncertainty often overlooked by traditional frequentist approaches that focus on central estimates. Second, unlike prognostic prediction tools, our method solely aims to forecast the mitotic count on the surgical specimen, a distinctive objective.

Strengths and Limitations

Given the widespread use of mitotic count as a grading parameter, our approach holds potential for scalability across diverse clinical settings, such as breast cancer, solitary fibrous tumors, and soft tissue sarcoma [30,64]. The integration of our posterior probability with other tools could enhance their effectiveness. However, our choice of prediction methods was largely constrained by the data set's size, as machine learning methods like deep learning demand substantial computational resources and larger sample sizes [65], which are often limited in rare diseases. Bayesian networks offer an advantageous alternative with principled variable selection, interpretability, and no minimum sample size requirement, making them a fitting choice for our study.

Despite our study's promising findings, several limitations warrant consideration. First, our use of a complex multilevel hierarchical model, meticulously designed according to the DAG, introduces potential challenges in model efficacy and implementation. However, we addressed this complexity by subjecting the model to rigorous testing with simulated data, ensuring its robustness and reliability.

Second, the performance evaluation of PROMETheus is based on data from a relatively limited patient population, which may restrict the generalizability of our findings. However, unlike

XSL•FO
RenderX

conventional approaches that rely solely on in-sample performance metrics for the model selection, we used PSIS-LOO-CV to identify the most effective model structure. Nonetheless, further validation on a larger and more diverse population is essential before considering the clinical deployment of PROMETheus, ensuring its efficacy and applicability across different clinical settings and patient demographics.

Third, the development and deployment of PROMETheus as a web-based application are at an early stage. Although we have made the tool publicly available as an open-source resource to encourage validation and application by other researchers and clinicians, the current version may lack certain functionalities and user-friendly features (such as the inclusion of risk classifications other than Miettinen and Lasota). Future work should focus on enhancing the application's interface, usability, and integration with clinical workflows, as well as providing comprehensive user training and support to facilitate its adoption in clinical practice.

Future Directions

We have developed a cutting-edge application that could revolutionize the estimation of mitotic count on surgical specimens, not only providing accurate quantification but also addressing the uncertainty that clinicians face when encountering challenging cases. By bridging the gap between clinical judgment and computed risk class based on available parameters, our application empowers clinicians to make informed decisions in complex scenarios. While previous studies have primarily focused on comparing biopsy and surgical specimens to evaluate the reliability of the former as the gold standard, we diverge from this approach. Our study transcends the limitations of preoperative biopsy in predicting tumor grade, shedding light on the underestimation of aggressiveness often associated with this method [66-69]. Embracing the widely accepted practice of causal modeling for covariate selection in epidemiological studies [70], we harness the power of Bayesian networks to pioneer a novel tool capable of predicting the mitotic count—a breakthrough in the existing literature.

In conclusion, our study represents a significant step forward in the development of precision tools for oncological risk assessment. By accurately predicting the mitotic count on surgical specimens in GISTs, PROMETheus enables clinicians to make informed treatment decisions, ultimately improving patient outcomes and advancing the field of sarcoma management. Moving forward, continued research and validation efforts will be essential to further refine and optimize the utility of PROMETheus in clinical practice, ultimately realizing its full potential in guiding personalized treatment strategies for patients with GISTs and other sarcomas.

Acknowledgments

We acknowledge the reSeARChOMA group of the Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital. This work was partially supported by Ricerca Corrente funding from the Italian Ministry of Health to IRCCS Humanitas Research Hospital.

Data Availability

Deidentified patient data are available upon motivated request on the public repository Zenodo [71]. Custom code for the analysis is available at the PROMETheus (Preoperative Mitosis Estimator Tool) GitHub repository [72].

Conflicts of Interest

None declared.

Multimedia Appendix 1

A more elaborated directed acyclic graph (DAG). During optimization cycles of the modeling, we simplified both the DAG and the probabilistic program used. Indeed, our first modeling attempts included several latent variables (circled) as the true mitotic rate (M) and the biologic aggressiveness (U), and modeled in detail the disease presentation (including symptoms among the variables), the therapy indication and its response. Although this DAG is more accurate, its complexity impairs the parameter estimation efficiency. Moreover, within the domain of the Bayesian network, a different approach using a system of equations that translate the structural causal model (SCM) in the code for the fitting might have been used; however, the choice of a more simplified approach (ie, using the M^s as inferential target instead of as a variable) removes the need to refit the model to impute the value.

[PNG File , 79 KB - jmir_v26i1e50023_app1.png]

Multimedia Appendix 2

Prior predictive simulation for the model coefficients. The coefficients for tumor size and surface (β and γ, respectively) show a wider (t) distribution compared to the biopsy count (δ or if a response to neoadjuvant therapy, N-acetylcysteine [NAC], is present). GIST: gastrointestinal stromal tumor.

[PNG File , 133 KB - jmir_v26i1e50023_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3
Fit's diagnostics on the simulated data set. The upper panel plots the number of effective samples against the R ^; the vertical gray line represents the total number of samples (half of the 8000 iterations: the first 4000 are discarded, multiplied for the 4 chains). The second panel shows the energy of the Hamiltonian, nicely following along a normal distribution (background blue line). The third panel shows a rank histogram plot (trankplot) for the log probability of the model. Each of the 4 chains alternates in dominating the ranking; this indicates that the chains mixed and converged.
[PNG File , 131 KB - jmir_v26i1e50023_app3.png]

Multimedia Appendix 4
Rank histogram plot (trankplot) for the all parameters. All the parameters have a nice mixing of the chains.
[PNG File , 440 KB - jmir_v26i1e50023_app4.png]

Multimedia Appendix 5
Model coefficients from fitting the simulated data set.
[PNG File , 156 KB - jmir_v26i1e50023_app5.png]

Multimedia Appendix 6
Fit's diagnostics on the real gastrointestinal stromal tumor (GIST) cohort. The upper panel plots the number of effective samples against the R ^; the vertical gray line represents the total number of samples (half of the 4000 iterations multiplied for the 4 chains; the first 2000 are discarded). The second panel shows the energy of the Hamiltonian, nicely following along a normal distibution (background blue line). The trankplot for the log-probability of the model shows a nice mixing of the chains.
[PNG File , 144 KB - jmir_v26i1e50023_app6.png]

Multimedia Appendix 7
Rank histogram plot (trankplot) for all parameters. All parameters have a nice mixing of the chains.
[PNG File , 473 KB - jmir_v26i1e50023_app7.png]

Multimedia Appendix 8
Model coefficients from fitting the simulated data set.
[PNG File , 126 KB - jmir_v26i1e50023_app8.png]

Multimedia Appendix 9
Supplementary materials.
[DOCX File , 18 KB - jmir_v26i1e50023_app9.docx]

References

1. de Pinieux G, Karanian M, Le Loarer F, Le Guellec S, Chabaud S, Terrier P, NetSarc/RePPS/ResSosFrench Sarcoma Group- Groupe d'Etude des Tumeurs Osseuses (GSF-GETO) networks. Nationwide incidence of sarcomas and connective tissue tumors of intermediate malignancy over four years using an expert pathology review network. PLoS One 2021;16(2):e0246958 [FREE Full text] [doi: 10.1371/journal.pone.0246958] [Medline: 33630918]
2. Hirota S, Isozaki K, Moriyama Y, Hashimoto K, Nishida T, Ishiguro S, et al. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. Science 1998 Jan 23;279(5350):577-580. [doi: 10.1126/science.279.5350.577] [Medline: 9438854]
3. Joensuu H, Hohenberger P, Corless CL. Gastrointestinal stromal tumour. Lancet 2013 Sep 14;382(9896):973-983 [FREE Full text] [doi: 10.1016/S0140-6736(13)60106-3] [Medline: 23623056]
4. Rubin BP, Heinrich MC. Genotyping and immunohistochemistry of gastrointestinal stromal tumors: An update. Semin Diagn Pathol 2015 Sep;32(5):392-399 [FREE Full text] [doi: 10.1053/j.semdp.2015.02.017] [Medline: 25766843]
5. von Mehren M, Joensuu H. Gastrointestinal stromal tumors. J Clin Oncol 2018 Jan 10;36(2):136-143 [FREE Full text] [doi: 10.1200/JCO.2017.74.9705] [Medline: 29220298]
6. Rossi S, Gasparotto D, Toffolatti L, Pastrello C, Gallina G, Marzotto A, et al. Molecular and clinicopathologic characterization of gastrointestinal stromal tumors (GISTs) of small size. Am J Surg Pathol 2010 Oct;34(10):1480-1491 [FREE Full text] [doi: 10.1097/PAS.0b013e3181ef7431] [Medline: 20861712]
7. Heinrich MC, Patterson J, Beadling C, Wang Y, Debiec-Rychter M, Dewaele B, et al. Genomic aberrations in cell cycle genes predict progression of -mutant gastrointestinal stromal tumors (GISTs). Clin Sarcoma Res 2019 Mar 5;9(1):3 [FREE Full text] [doi: 10.1186/s13569-019-0112-7] [Medline: 30867899]

XSL•FO
RenderX

8. Joensuu H, Roberts PJ, Sarlomo-Rikala M, Andersson LC, Tervahartiala P, Tuveson D, et al. Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. N Engl J Med 2001 Apr 05;344(14):1052-1056. [doi: 10.1056/NEJM200104053441404] [Medline: 11287975]
9. Demetri GD, von Mehren M, Blanke CD, Van den Abbeele AD, Eisenberg B, Roberts PJ, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. N Engl J Med 2002 Aug 15;347(7):472-480. [doi: 10.1056/NEJMoa020461] [Medline: 12181401]
10. Verweij J, Casali PG, Zalcberg J, LeCesne A, Reichardt P, Blay J, et al. Progression-free survival in gastrointestinal stromal tumours with high-dose imatinib: randomised trial. Lancet 2004;364(9440):1127-1134 [FREE Full text] [doi: 10.1016/S0140-6736(04)17098-0] [Medline: 15451219]
11. Somerville MA. A postmodern moral tale: the ethics of research relationships. Nat Rev Drug Discov 2002 Apr;1(4):316-320. [doi: 10.1038/nrd773] [Medline: 12120283]
12. Heinrich MC. Imatinib treatment of metastatic GIST: don't stop (believing). Lancet Oncol 2010 Oct;11(10):910-911 [FREE Full text] [doi: 10.1016/S1470-2045(10)70225-4] [Medline: 20864404]
13. Cohen P, Cross D, Jänne PA. Kinase drug discovery 20 years after imatinib: progress and future directions. Nat Rev Drug Discov 2021 Jul;20(7):551-569 [FREE Full text] [doi: 10.1038/s41573-021-00195-4] [Medline: 34002056]
14. Casali PG, Blay JY, Abecassis N, Bajpai J, Bauer S, Biagini R, ESMO Guidelines Committee, EURACANGENTURIS. Electronic address: clinicalguidelines@esmo.org. Gastrointestinal stromal tumours: ESMO-EURACAN-GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2022 Jan;33(1):20-33 [FREE Full text] [doi: 10.1016/j.annonc.2021.09.005] [Medline: 34560242]
15. Eisenberg BL, Harris J, Blanke CD, Demetri GD, Heinrich MC, Watson JC, et al. Phase II trial of neoadjuvant/adjuvant imatinib mesylate (IM) for advanced primary and metastatic/recurrent operable gastrointestinal stromal tumor (GIST): early results of RTOG 0132/ACRIN 6665. J Surg Oncol 2009 Jan 01;99(1):42-47 [FREE Full text] [doi: 10.1002/jso.21160] [Medline: 18942073]
16. Wang D, Zhang Q, Blanke CD, Demetri GD, Heinrich MC, Watson JC, et al. Phase II trial of neoadjuvant/adjuvant imatinib mesylate for advanced primary and metastatic/recurrent operable gastrointestinal stromal tumors: long-term follow-up results of Radiation Therapy Oncology Group 0132. Ann Surg Oncol 2012 Apr;19(4):1074-1080 [FREE Full text] [doi: 10.1245/s10434-011-2190-5] [Medline: 22203182]
17. Rutkowski P, Gronchi A, Hohenberger P, Bonvalot S, Schöffski P, Bauer S, et al. Neoadjuvant imatinib in locally advanced gastrointestinal stromal tumors (GIST): the EORTC STBSG experience. Ann Surg Oncol 2013 Sep;20(9):2937-2943 [FREE Full text] [doi: 10.1245/s10434-013-3013-7] [Medline: 23760587]
18. Casali P, Le Cesne A, Velasco A, Kotasek D, Rutkowski P, Hohenberger P, et al. Final analysis of the randomized trial on imatinib as an adjuvant in localized gastrointestinal stromal tumors (GIST) from the EORTC Soft Tissue and Bone Sarcoma Group (STBSG), the Australasian Gastro-Intestinal Trials Group (AGITG), UNICANCER, French Sarcoma Group (FSG), Italian Sarcoma Group (ISG), and Spanish Group for Research on Sarcomas (GEIS). Ann Oncol 2021 Apr;32(4):533-541 [FREE Full text] [doi: 10.1016/j.annonc.2021.01.004] [Medline: 33482247]
19. Fletcher CDM, Berman JJ, Corless C, Gorstein F, Lasota J, Longley BJ, et al. Diagnosis of gastrointestinal stromal tumors: A consensus approach. Hum Pathol 2002 May;33(5):459-465. [doi: 10.1053/hupa.2002.123545] [Medline: 12094370]
20. Miettinen M, Lasota J. Gastrointestinal stromal tumors: pathology and prognosis at different sites. Semin Diagn Pathol 2006 May;23(2):70-83. [doi: 10.1053/j.semdp.2006.09.001] [Medline: 17193820]
21. Joensuu H. Risk stratification of patients diagnosed with gastrointestinal stromal tumor. Hum Pathol 2008 Oct;39(10):1411-1419 [FREE Full text] [doi: 10.1016/j.humpath.2008.06.025] [Medline: 18774375]
22. Hornick JL, Fletcher CD. Immunohistochemical staining for KIT (CD117) in soft tissue sarcomas is very limited in distribution. Am J Clin Pathol 2002 Feb 01;117(2):188-193. [doi: 10.1309/lx9u-f7p0-uwdh-8y6r]
23. West RB, Corless CL, Chen X, Rubin BP, Subramanian S, Montgomery K, et al. The novel marker, DOG1, is expressed ubiquitously in gastrointestinal stromal tumors irrespective of KIT or PDGFRA mutation status. Am J Pathol 2004 Jul;165(1):107-113. [doi: 10.1016/s0002-9440(10)63279-8]
24. Espinosa I, Lee C, Kim M, Rouse B, Subramanian S, Montgomery K, et al. A novel monoclonal antibody against DOG1 is a sensitive and specific marker for gastrointestinal stromal tumors. Am J Surg Pathol 2008 Feb;32(2):210-218. [doi: 10.1097/PAS.0b013e3181238cec] [Medline: 18223323]
25. Miettinen M, Wang Z, Lasota J. DOG1 antibody in the differential diagnosis of gastrointestinal stromal tumors: a study of 1840 cases. Am J Surg Pathol 2009 Sep;33(9):1401-1408 [FREE Full text] [doi: 10.1097/PAS.0b013e3181a90e1a] [Medline: 19606013]
26. Cananzi FCM, Judson I, Lorenzi B, Benson C, Mudan S. Multidisciplinary care of gastrointestinal stromal tumour: a review and a proposal for a pre-treatment classification. Eur J Surg Oncol 2013 Nov;39(11):1171-1178 [FREE Full text] [doi: 10.1016/j.ejso.2013.08.030] [Medline: 24063969]
27. Grazzini G, Guerri S, Cozzi D, Danti G, Gasperoni S, Pradella S, et al. Gastrointestinal stromal tumors: relationship between preoperative CT features and pathologic risk stratification. Tumori 2021 Dec;107(6):556-563 [FREE Full text] [doi: 10.1177/0300891621996447] [Medline: 33620027]

XSL•FO
RenderX

28. Polkowski M, Gerke W, Jarosz D, Nasierowska-Guttmejer A, Rutkowski P, Nowecki ZI, et al. Diagnostic yield and safety of endoscopic ultrasound-guided trucut [corrected] biopsy in patients with gastric submucosal tumors: a prospective study. Endoscopy 2009 Apr;41(4):329-334 [FREE Full text] [doi: 10.1055/s-0029-1214447] [Medline: 19340737]
29. Cananzi FCM, Ruspi L, Sama` L, Renne SL, Sicoli F, Quagliuolo V. The gist of surgical margins in GIST: a narrative review. Laparosc Surg 2022 Jan;6:4-4. [doi: 10.21037/ls-20-139]
30. Renne SL. How to measure your microscope's HPF. A critical guide for residents. Pathologica 2023 Dec;115(6):302-307. [doi: 10.32074/1591-951x-900]
31. Larghi A, Fuccio L, Chiarello G, Attili F, Vanella G, Paliani GB, et al. Fine-needle tissue acquisition from subepithelial lesions using a forward-viewing linear echoendoscope. Endoscopy 2014 Jan;46(1):39-45 [FREE Full text] [doi: 10.1055/s-0033-1344895] [Medline: 24218311]
32. Kobara H, Mori H, Rafiq K, Fujihara S, Nishiyama N, Chiyo T, et al. Analysis of the amount of tissue sample necessary for mitotic count and Ki-67 index in gastrointestinal stromal tumor sampling. Oncol Rep 2015 Jan;33(1):215-222 [FREE Full text] [doi: 10.3892/or.2014.3608] [Medline: 25405369]
33. Christensen AW, Goldberg AF. Mitotic count of fine needle aspiration material of gastrointestinal stromal tumours of the stomach underestimates actual mitotic count. Cytopathology 2022 Jan 28;33(1):100-106. [doi: 10.1111/cyt.13050] [Medline: 34390589]
34. Pearl J. Causality: Models, Reasoning and Inference. 2nd edition. Cambridge, UK: Cambridge University Press; 2009.
35. Peters J, Janzing D, Schölkopf B. Elements of Causal Inference Foundations and Learning Algorithms. Cambridge, MA: MIT Press; 2017:978.
36. Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. Validating Bayesian inference algorithms with simulation-based calibration. arXiv Preprint posted online on October 21, 2020 [FREE Full text]
37. Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. J R Stat Soc A 2019 Jan 15;182(2):389-402. [doi: 10.1111/rssa.12378]
38. Gelman A, Vehtari A, Simpson D, Margossian C, Carpenter B, Yao Y, et al. Bayesian workflow. arXiv Preprint posted online on November 3, 2020 [FREE Full text]
39. McElreath R. Statistical Rethinking: A Bayesian Course With Examples in R and Stan. Boca Raton, FL: CRC press; Dec 21, 2015.
40. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw 2017;76:a [FREE Full text] [doi: 10.18637/jss.v076.i01] [Medline: 36568334]
41. Stan Development Team. RStan: The R interface to Stan. Stan. URL: https://mc-stan.org/ [accessed 2024-10-01]
42. Stan Development Team. The Stan Core Library. Stan. 2019. URL: http://mc-stan.org/ [accessed 2024-10-01]
43. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing URL: https://www.R-project.org/ [accessed 2024-10-01]
44. Hoffman M, Gelman A. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. arXiv Preprint posted online on November 18, 2011 [FREE Full text]
45. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. Bayesian Data Analysis. Boca Raton, FL: CRC press; 2013.
46. Gelman A, Vehtari A. What are the most important statistical ideas of the past 50 years? J Am Stat Assoc 2021 Jul 08;116(536):2087-2097. [doi: 10.1080/01621459.2021.1938081]
47. Betancourt M. Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. arXiv Preprint posted online on April 3, 2016 [FREE Full text]
48. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P. Rank-normalization, folding, and localization: an improved Rˆ for assessing convergence of MCMC (with discussion). Bayesian Anal 2021 Jun 1;16(2). [doi: 10.1214/20-ba1221]
49. Gabry J. Graphical posterior predictive checks using the bayesplot package. Stan. URL: https://mc-stan.org/bayesplot/articles/graphical-ppcs.html [accessed 2024-10-01]
50. Stan DT. Reparameterization. Stan. URL: https://mc-stan.org/docs/stan-users-guide/ppcs.html [accessed 2024-10-01]
51. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. R Journal. 2006. URL: https://journal.r-project.org/archive/ [accessed 2024-10-01]
52. Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 1973;60(2):255-265. [doi: 10.1093/biomet/60.2.255]
53. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 2010;11:3594 [FREE Full text]
54. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 2016 Aug 30;27(5):1413-1432. [doi: 10.1007/s11222-016-9696-4]
55. Watanabe S. Mathematical theory of Bayesian statistics. Watanabe. URL: http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/mt_bs.html [accessed 2024-10-01]
56. Watanabe S. Mathematical Theory of Bayesian Statistics. Boca Raton, FL: CRC press; 2018:9781482238082.
57. Walsh S, MacDuff E, Wong N, Novelli M. G103 dataset for histopathological reporting of gastrointestinal stromal tumours. The Royal College of Pathologists. URL: https://www.rcpath.org/uploads/assets/6741f729-adf6-4ee6-b06144967e2d734a/g103-Dataset-for-histopathological-reporting-of-gastrointestinal-stromal-tumours.pdf [accessed 2024-10-01]

XSL•FO
RenderX

58. Laurini J. Protocol for the examination of resection specimens from patients with gastrointestinal stromal tumor (GIST). College of American Pathologists. URL: https://documents.cap.org/protocols/Stomach.GIST_4.2.0.0.REL_CAPCP.pdf [accessed 2024-10-01]
59. Laurini J. Protocol for the examination of biopsy specimens from patients with gastrointestinal stromal tumor (GIST). College of American Pathologists. URL: https://documents.cap.org/protocols/Stomach.GIST.Bx_4.2.0.0.REL_CAPCP.pdf [accessed 2024-10-01]
60. Jannink I, Risberg B, Van Diest PJ, Baak JP. Heterogeneity of mitotic activity in breast cancer. Histopathology 1996 Nov;29(5):421-428 [FREE Full text] [doi: 10.1046/j.1365-2559.1996.d01-509.x] [Medline: 8951486]
61. PROMETheus. URL: https://slrenne.shinyapps.io/PROMETheus/ [accessed 2024-10-02]
62. Callegaro D, Miceli R, Bonvalot S, Ferguson P, Strauss DC, Levy A, et al. Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis. Lancet Oncol 2016 May;17(5):671-680. [doi: 10.1016/S1470-2045(16)00010-3] [Medline: 27068860]
63. Pasquali S, Palmerini E, Quagliuolo V, Martin-Broto J, Lopez-Pousa A, Grignani G, et al. Neoadjuvant chemotherapy in high-risk soft tissue sarcomas: A Sarculator-based risk stratification analysis of the ISG-STS 1001 randomized trial. Cancer 2022 Jan 01;128(1):85-93 [FREE Full text] [doi: 10.1002/cncr.33895] [Medline: 34643947]
64. Cree IA, Tan PH, Travis WD, Wesseling P, Yagi Y, White VA, et al. Counting mitoses: SI(ze) matters!. Mod Pathol 2021 Sep;34(9):1651-1657 [FREE Full text] [doi: 10.1038/s41379-021-00825-7] [Medline: 34079071]
65. Thompson N, Greenewald K, Lee K, Manso G. The computational limits of deep learning. arXiv Preprint posted online on July 27, 2022 [FREE Full text] [doi: 10.21428/bf6fb269.1f033948]
66. Destounis S, Seifert P, Somerville P, Murphy P, Morgan R, Arieno A, et al. Underestimation of papillary breast lesions by core biopsy: correlation to surgical excision. Breast Cancer 2014 Mar 5;21(2):128-134. [doi: 10.1007/s12282-012-0361-2] [Medline: 22477267]
67. Yıldızlı, Üntan, Demirci D. What is the consistency between the results of needle biopsy and prostatectomy specimen pathology results? A pilot study. Turk J Med Sci 2021 Jun 28;51(3):1360-1364 [FREE Full text] [doi: 10.3906/sag-2009-73] [Medline: 33535735]
68. Lago V, Martín B, Ballesteros E, Cárdenas-Rebollo JM, Minig L. Tumor grade correlation between preoperative biopsy and final surgical specimen in endometrial cancer: the use of different diagnostic methods and analysis of associated factors. Int J Gynecol Cancer 2018 Sep 01;28(7):1258-1263. [doi: 10.1097/igc.0000000000001304]
69. Dik E, Ipenburg N, Adriaansens S, Kessler P, Es RV, Willems S. Poor correlation of histologic parameters between biopsy and resection specimen in early stage oral squamous cell carcinoma. Am J Clin Pathol 2015 Sep;144(4):666. [doi: 10.1309/ajcpfivhhh7q3blx]
70. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013 Feb 15;177(4):292-298 [FREE Full text] [doi: 10.1093/aje/kws412] [Medline: 23371353]
71. Zenodo. URL: https://zenodo.org/communities/humanitasirccs/ [accessed 2024-10-02]
72. PROMETheus repository. URL: https://www.github.com/slrenne/PROMETheus [accessed 2022-10-02]

Abbreviations

CI: compatibility interval
DAG: directed acyclic graph
GIST: gastrointestinal stromal tumor
HDPI: highest posterior density interval
HMC: Hamiltonian Monte Carlo
HPF: high-power field
IRCCS: Istituto di Ricovero e Cura a Carattere Scientifico
MCMC: Markov Chain Monte Carlo
PROMETheus: Preoperative Mitosis Estimator Tool
PSIS-LOO-CV: Pareto Smoothed Importance Sampling Leave-One-Out Cross-Validation Criteria
SCM: structural causal model
TKI: tyrosine-kinase inhibitor
WAIC: widely applicable information criterion

XSL•FO
RenderX

Edited by T Leung; submitted 16.06.23; peer-reviewed by F Pennoni; comments to author 02.05.24; revised version received 14.05.24; accepted 21.07.24; published 22.10.24.

Please cite as:
Renne SL, Cammelli M, Santori I, Tassan-Mangina M, Samà L, Ruspi L, Sicoli F, Colombo P, Terracciano LM, Quagliuolo V, Cananzi FCM
True Mitotic Count Prediction in Gastrointestinal Stromal Tumors: Bayesian Network Model and PROMETheus (Preoperative Mitosis Estimator Tool) Application Development
J Med Internet Res 2024;26:e50023
URL: https://www.jmir.org/2024/1/e50023
doi:10.2196/50023
PMID:

©Salvatore Lorenzo Renne, Manuela Cammelli, Ilaria Santori, Marta Tassan-Mangina, Laura Samà, Laura Ruspi, Federico Sicoli, Piergiuseppe Colombo, Luigi Maria Terracciano, Vittorio Quagliuolo, Ferdinando Carlo Maria Cananzi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Web-Based Health Information Seeking by People Living With Multiple Sclerosis: Qualitative Investigation of the Multiple Sclerosis Online Course

William Bevens[1,2], PhD; Rebekah Davenport[3], MCP; Sandra Neate[1], MBBS; Maggie Yu[1], MPsy, MPH; Pia Jelinek[1], MBBS; George Alexander Jelinek[1], MD; Jeanette Reece[1], MPH, PhD

[1]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Carlton, Australia
[2]Institute for Clinical and Translational Science, University of California, Irvine, Irvine, CA, United States
[3]Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, Australia

Corresponding Author:
William Bevens, PhD
Centre for Epidemiology and Biostatistics
Melbourne School of Population and Global Health
The University of Melbourne
335 Bouverie Street
Carlton, 3053
Australia
Phone: 61 3 8344 2173
Email: wbevens@hs.uci.edu

Abstract

Background: Digital technologies have afforded people living with multiple sclerosis (MS) access to telehealth consultations, diagnostic tools, and monitoring. Although health care professionals remain the most trusted source of information, the internet has emerged as a valuable resource for providing MS-related information, particularly during the COVID-19 pandemic. Notably, people living with MS are increasingly seeking educational content for a range of topics related to the self-management of MS; however, web-based information seeking remains largely underevaluated. To address this gap and ensure that web-based health-related information is accessible and engaging, this study used qualitative methods to analyze the reflections from participants of web-based educational programs for people living with MS.

Objective: This study aimed to explore the motivations, behaviors, and expectations of web-based health information seeking for people living with MS.

Methods: We conducted semistructured interviews for 38 people living with MS 1 month after they completed the novel MS Online Course, which provided information on modifiable lifestyle-related risk factors for people living with MS. Of the 38 participants, 22 (58%) completed the intervention course and 16 (42%) completed the standard care course. Inductive thematic analysis was used within a qualitative paradigm, and 2 authors coded each interview separately and arrived at themes with consensus.

Results: We identified 2 themes: motivation to learn and MS information on the web. The diagnosis of MS was described as a pivotal moment for precipitating web-based information seeking. People living with MS sought lifestyle-related information to facilitate self-management and increase control of their MS. Although social media sites and MS websites were considered useful for providing both support and information, discretion was needed to critically appraise information. Recognizable institutions were frequently accessed because of their trustworthiness.

Conclusions: This study provided novel insights into the motivations of people living with MS for seeking web-based health information. Furthermore, their preferences for the content and format of the web-based information accessed and their experiences and reactions to this information were explored. These findings may guide educators, researchers, and clinicians involved in MS care to optimize the engagement and processing of web-based health information seeking by people living with MS.

(J Med Internet Res 2024;26:e53372) doi:10.2196/53372

XSL•FO
RenderX

KEYWORDS

information-seeking behavior; self-management; lifestyle; digital health

Introduction

Background

Multiple sclerosis (MS) is a chronic neurodegenerative disease that affects almost 3 million people worldwide [1]. Despite the advent of a range of disease-modifying medications that can modestly affect disease progression, MS is associated with a large disease burden because of relatively early onset and the accumulation of disability with no cure available. Self-management approaches, in which individuals are encouraged to become active participants in the management of their disease [2,3], are highly relevant to people living with MS. MS-related knowledge and self-management skills are integral in assisting people living with MS to manage their disease [4]. Knowledge of and self-management skills for MS are mediated by how people living with MS access and engage with relevant health information sources [5]; however, this remains an underexplored area of MS care.

It is well understood that people living with MS seek information related to MS very early on in the disease course, often occurring before their first neurology appointment [6]. Much of this information is sourced on the web [6-8], whereby people living with MS report not only empowerment but also skepticism toward information found on websites [9,10] and social media [11]. Neurologists and other health care professionals (HCPs) can assist in critically appraising sourced information [12]; however, people living with MS often seek information independently of formal health care environments [13]. In the context of an increasingly complex digital world, structured educational programs from trusted providers may provide important mechanisms for health information delivery for people living with MS. Several groups have developed web-based programs to deliver educational content to people living with MS, with the goal of increasing MS-related knowledge [14,15], and have demonstrated early success [16]. What remains unexplored is how the self-directed web-based health information–seeking behavior of people living with MS relates to their engagement with educational intervention programs delivered via web-based platforms.

This qualitative study formed a part of a large randomized controlled trial (RCT) comparing the effects of 2 web-based educational lifestyle programs on medium- and long-term health outcomes in people living with MS. This study presents a qualitative analysis of semistructured interviews of participants conducted 1 month after the completion of the MS Online Course (MSOC) RCT [17].

Objectives

This qualitative analysis specifically aimed to investigate health information seeking in a subset of people living with MS interested in lifestyle behavior modification (ie, those who selected to actively engage in a lifestyle education program via enrollment in the MSOC RCT). Notably, this study also sought to explore the perceptions of people living with MS encountering health information on the web and how they navigate different types of health information across various sources.

Methods

Ethical Considerations

This ancillary study was approved by the University of Melbourne Human Research Ethics subcommittee on November 2, 2021 (ID: 1851781.2). This protocol for the RCT was prospectively registered on November 25, 2021, and was approved by the Australian and New Zealand Clinical Trials Registry (ACTRN: ACTRN12621001605886). Written informed consent was obtained from all participants before inclusion in the RCT, with additional verbal consent obtained from all participants in this study.

MSOC RCT Study Design

The MSOC RCT study design and course content have been comprehensively described previously [17]. We developed 2 courses in conjunction with people living with MS: the standard care course (SCC) based on the information from MS websites and the intervention course (IC) based on the Overcoming MS program. The content and format of the Overcoming MS program have been described comprehensively elsewhere [18]; however, briefly, it is a lifestyle management program recommending a plant-based wholefood diet, exercise and stress reduction activities, sunlight exposure, and vitamin D consumption. The 7 modules of the first round of the MSOC were delivered across 6 weeks from August to September 2022.

We recruited participants for the MSOC RCT web-based across many Facebook (Meta Platforms, Inc) groups and MS websites and forums worldwide, as previously described [17]. The consenting participants completed a screening questionnaire on the study website to confirm eligibility: a self-reported diagnosis of MS by a physician and aged at least 18 years. A baseline survey consisting of questions on sociodemographic information, lifestyle habits, and health status was included in the first module of the course. This baseline survey was voluntary for participants to complete, and consequently, only the participants who completed this baseline survey were included in the MSOC RCT for subsequent data analysis.

Qualitative Recruitment

We collected qualitative data 1 month after course completion. The possibility of an invitation to participate in an interview in the poststudy period was outlined in the RCT consent form and plain language statement. All participants who completed the first round of the MSOC and baseline and postcourse evaluation surveys were emailed an invitation to participate in a semistructured interview along with a plain language statement and consent form specific for the qualitative interviews that outlined the time commitment and process.

The Interviewers

The interviews were conducted by SN, JR, RD, and PJ, all with clinical and qualitative research expertise. We had access to the

XSL•FO
RenderX

course materials and were not blinded to participants' course allocation, which is a necessary step in understanding participants' views on the course content. Single blinding of the participants was maintained.

The Interviews

The question guide for the 1-month postcourse interview was developed by WB, RD, SN, PJ, and JR (provided in Multimedia Appendix 1 [19-23]).

We conducted interviews by videoconference using the Zoom (Zoom Video Communications, Inc) software licensed by the University of Melbourne or via WhatsApp. The first 3 interviews were conducted with a lead interviewer and 2 observers, with the permission of the interviewee, to assess the interview schedule and facilitate consistency between the interviewers. Following the first 3 interviews, the interview schedule was minimally revised, based on our observations, to reduce repetition and improve the clarity of the questions. Further interviews were conducted with 1 or 2 interviewers depending on our availability. The interviews were digitally recorded and transcribed verbatim using the voice recognition software [24].

The interviewer manually edited the transcripts while listening to the recordings to ensure accuracy. Edits to transcripts, including the removal of data, were made only for the following reasons:

1. Grammatical or spelling errors produced by the software during transcription
2. Removal of personal, sensitive, or other information that was deemed irrelevant to the overarching aims of the study

Transcripts were imported into NVivo (Lumivero) software for data management. In accordance with best practice data management, NVivo document coding development and theme evolution were maintained within the NVivo outputs [25].

Qualitative Data Analysis

The primary analytical approach to this data set was an inductive, reflexive thematic analysis (RTA), which allows for epistemological flexibility when determining theories from which to interpret data [26]. The strengths of RTA allow for researcher autonomy on what type of theme to identify (eg, latent or semantic) and what type of analysis to apply (inductive vs theoretical) [27]. Despite its flexibility, RTA still requires a systematic approach to analysis, which is categorized into 6 distinct phases [26]. WB and RD performed data immersion by reviewing all the transcripts and checking against the recordings to ensure transcript accuracy. Initial code generation was then completed separately using the NVivo software by WB and RD. In phase 3, WB and RD initially divided the data corpus between the IC and SCC arms and independently generated codes and themes from 1 arm at a time. The analysis team (WB, RD, SN, and JR) met to examine these codes, and it was determined at this stage that, owing to the diverse results, different broad themes were incongruent for simultaneous analysis. On the basis of this decision, the data corpus was divided into 4 separate data sets (Figure 1). The data presented in this study were themes falling under the "information-seeking" data set (represented as a black line) that WB and RD independently identified. Once this new data set was determined, WB and RD performed data immersion, familiarization, and refinement of themes under the "information-seeking" data set. We reviewed and renamed themes before the final themes were determined. Notably, the gray line themes identified are not discussed in this study.

XSL•FO
RenderX

Figure 1. Process of the development of codes, themes, and subthemes. MS: multiple sclerosis.

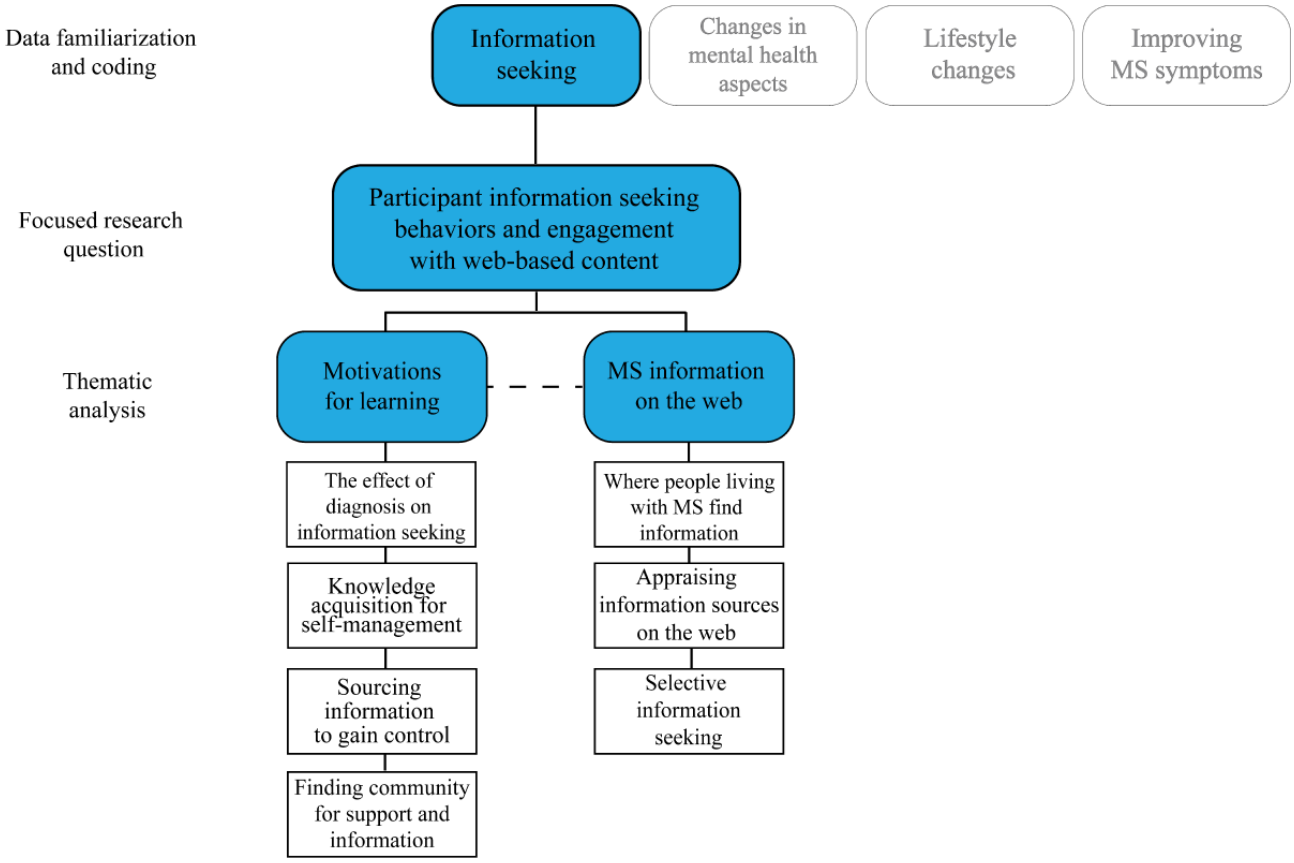

Results

Participant Characteristics

A total of 54 participants were contacted and invited to participate in a semistructured interview. A total of 22 (79%) of the 28 participants in the IC arm and 16 (62%) of the 26 participants in the SCC arm agreed to be interviewed. Participants were predominantly from the United Kingdom, the United States, and Australia or New Zealand. The mean age of the participants in the SCC and IC was 50.7 (SD 15.6) years and 48.1 (SD 12.3) years, respectively, and almost 89% (34/38) of the participants were female (Table 1). Participant characteristics were similar across the IC and SCC; however, there was a higher proportion of people with relapsing-remitting MS and men from the IC compared with the SCC.

XSL•FO
RenderX

Table 1. Participant sociodemographic and clinical characteristics (n=38).

Characteristics	SCC[a] (n=16)	IC[b] (n=22)
Sex, n (%)		
Male	1 (6)	3 (14)
Female	15 (94)	19 (86)
Country of residence, n (%)		
Australia or NZ[c]	2 (13)	5 (23)
United States or Canada	5 (31)	5 (23)
United Kingdom	4 (25)	3 (14)
Other	5 (31)	9 (41)
MS[d] phenotype, n (%)		
RRMS[e]	9 (56)	18 (82)
PPMS[f] or SPMS[g]	4 (25)	3 (14)
Other	3 (19)	1 (55)
Age (y), mean (SD)	50.7 (15.6)	48.1 (12.2)
MS duration from diagnosis (y), mean (SD)	6.9 (5.7)	5.4 (4.3)

[a]SCC: standard care course.
[b]IC: intervention course.
[c]NZ: New Zealand.
[d]MS: multiple sclerosis.
[e]RRMS: relapsing-remitting multiple sclerosis.
[f]PPMS: primary progressive multiple sclerosis.
[g]SPMS: secondary progressive multiple sclerosis.

Themes

Overview

We identified 2 themes from the data: (1) “motivations for learning” described the motivations behind why participants seek web-based information in general and their reasons for engaging with the MSOC specifically, and (2) “MS information on the web” explored the relationships between participants and the web-based health information with which they engage.

Verbatim quotes are used to illustrate the themes. The themes identified from the “information-seeking” data set were similar across both study arms. Therefore, both study arms are reported together; however, example quotations from each study arm are provided where possible. The quotes are followed by the study arm (eg, IC or SCC) and participant number.

Motivations for Learning

The theme “motivations for learning” explored participants’ reasons for enrolling in and engaging with the MSOC specifically and why they sought MS-related information on the web in general. Furthermore, the impact of the diagnosis on the information-seeking behavior of participants was investigated, particularly relating to how these behaviors informed the desire of people living with MS for content related to self-management.

Effect of Diagnosis on Information Seeking

The participants described receiving an MS diagnosis as a critical event in their lives, which precipitated information-seeking behavior for health-related knowledge. The participants described difficulties in accessing timely, comprehensible, and relevant information at diagnosis, and the MSOC was considered positive compared with self-managed research. Furthermore, the participants mentioned gaining information as quickly as possible following diagnosis while filtering out unnecessary information from the “information overload” as follows:

> *I feel like [the online course] would’ve been perfect when I was first diagnosed. When I was just desperately searching for anything, for any information about [MS].* [SCC7]

The initial impact of the diagnosis resulted in a desire to learn as much as possible as fast as possible, and the participants reported readily accessing many different web-based resources:

> *I was newly diagnosed and was in a phase where I was trying to learn as much as I could about MS as fast as I possibly can.* [IC4]

> *I completely stopped everything in my life and turned my life in a completely different direction with my diagnosis. I jumped onto all kinds of bandwagons, let’s just put it that way.* [SCC10]

XSL•FO
RenderX

The impact of the diagnosis affected some people's ability to process the volume and type of information available, leading to a need for the controlled delivery of information that could be taken at their own pace:

> *When you first get MS, you do get some information and it's sort of just thrown at you.* [IC2]
>
> *When I was first diagnosed, it wasn't that it was confrontational, it was just that it was too much.* [SCC4]

In addition to the information provided at diagnosis, some participants were advised by their HCPs to seek further information via the internet. For some participants, the timing of their diagnosis was relevant. For those who received their diagnosis during the pandemic, there was little alternative than to seek web-based information. In these instances, HCPs recommended reputable websites for information seeking:

> *When I was first diagnosed I was told "look, you know that there's lots of misinformation out there and there's lots of anecdotal experiential information out there." [My neurologist] told me to stick to the recommended sources, the MS Trust and the MS Society. And so I did read all that and they have some good information and they run some good webinars.* [IC19]
>
> *I was diagnosed during lockdown and the advice given to me was, "you're just gonna have to go look up stuff on the internet." They told me to go to certain sites: MS Trust, MS Society. They're the big ones in the UK.* [SCC16]

For some participants, the avoidance of information appeared to be functional within the early stages of adjusting to an MS diagnosis:

> *There was avoidance at first, probably denial, you know, the grieving stages. But that did not last long cause then I had to decide what medicine I was going to go on. So then I just read everything.* [IC22]

Many participants were selective in the information they sought following their diagnosis, with some preferentially seeking realistic and positively framed information. Some participants saw the MSOC as a source of positively presented information as follows:

> *When you get diagnosed, there can be a lot of negative information and obviously just in your own life, you can go through a lot of hard changes after being diagnosed. But the course was something that was really positive.* [SCC6]

Knowledge Acquisition for Self-Management

The participants sought knowledge to better understand their MS, with information on lifestyle modifications acting as a potential tool to facilitate self-management skills. For most participants, discovering novel information was the primary motivator for engaging with the web-based courses, and they were open to new ideas:

> *I wanted to hear something new since - this disease - everybody has an opinion on it.* [IC17]
>
> *I wanted to just learn more about this disease...Is there anything new that I can learn from it?* [SCC8]

Many participants reported that discovering new and specific information that was not available in previously explored web-based spaces was a principal reason for engaging with the MSOC, that is, participants felt that learning certain information about lifestyle behaviors would help them manage their MS and potentially improve future health outcomes:

> *Most of the information was available through online publications...in terms of literature, a lot of information is available already. So, I was after the in-between sort of information, things like amounts, for example, how much to eat, in terms of quantity, and in terms of what not to do.* [IC5]

For some participants in the SCC, the level of lifestyle-related information was not as complex or detailed as they had hoped:

> *I felt like some of the information was reinforcing information that I had already heard and sometimes I was hoping that maybe it went a little more in depth.* [SCC7]

Furthermore, this knowledge had to be easily comprehensible, especially for those with cognitive impairment, where the accessibility of information within the MSOC enhanced knowledge acquisition:

> *I think what the course really did for me was it underpinned what I had read. For me because I have cognitive issues, I need to really simplify things.* [IC19]

Moreover, the participants commonly reported that the importance of undertaking the course was to consolidate their existing knowledge base as opposed to learning new information. Others used the course to "fill in the gaps," having previously undertaken their own research:

> *I thought that it's very positive reinforcement because I've done my research, I am doing this course and it's telling me the same things.* [SCC4]

Sourcing Information to Gain Control

The subtheme "sourcing information to gain control" focused on the individuals' intentions to seek information, exemplified by their enrollment into the MSOC with the objective of enhancing their perceived control over their MS. The participants expressed the belief that acquiring a deeper understanding of MS and lifestyle modifications could substantially enhance their capacity to effectively manage and exert control over the disease:

> *I'm just trying to cover all bases. Trying to take some control of a condition that I don't really have much control of.* [IC19]
>
> *I think I was just looking for a bit more information so I felt more informed...just so I felt more in control.* [SCC16]

XSL•FO
RenderX

The participants had a common sentiment of not wanting to be passive in their experience of living with MS:

> *Since I got diagnosed with MS I've learned a lot and I keep learning. I don't wanna be one of those people that just sits back and does nothing, just accepts that the doctors said I got this horrible disease and then don't look any further.* [IC9]

The knowledge gained led to an enhanced sense of capacity to manage MS:

> *[MS is] the disease without a cure and you've got no control over it. So it [knowledge] gives you that sense of some control at least.* [SCC9]

Finding Community for Support and Information

Some participants experienced isolation and a lack of face-to-face communication and other support. Some participants with limited resources or a smaller population in their country of origin sought web-based connections or engagement. The forum enabled a certain degree of interaction and support:

> *Because I don't know other people in Taiwan with MS, I could see other people on the internet and I could relate to them, and they could be someone like me and I could see they were doing something for their health.* [SCC3]

> *Cause I was alone. I have nobody to talk to about my diagnosis. I didn't know anyone with it...It was awful. I didn't find any groups on the internet or on Facebook, and there was no Instagram at that time. I was kind of isolated.* [SCC8]

The participants felt supported by hearing others' experiences of MS on the community forum:

> *I wanted to explore the learning experiences of other people with MS.* [IC20]

> *A bit of refreshing my understanding on what MS is and then also a bit more moral support.* [SCC5]

Some participants described feeling comforted knowing that others were participating even if they did not communicate with them:

> *I commented maybe once or twice in the forum. It was mostly seeing that there's other people doing the course...* [IC12]

MS Information on the Web

Overview

The theme "MS information on the web" encompassed a range of subthemes that explored how and where people living with MS accessed and engaged with web-based information and the value they attributed to different spaces and their unique roles. People living with MS seek information from a variety of sources on the web; however, not all sources are considered equal in quality. Furthermore, this theme explored challenges associated with accessing MS-related information on the web, including conflicting opinions on disease management approaches and negative presentations of MS. Finally, the role that HCPs play in facilitating or supporting web-based information dispersal is discussed.

Where People Living With MS Find Information

The most commonly reported source of frequently accessed information on the web was via Facebook groups. Many participants used Facebook to gather more information about MS and understand others' experiences of living with MS:

> *I joined all these MS groups on Facebook too because I'm just hoping to get the science.* [IC4]

> *Facebook groups, even in Ukraine, are where participants discuss how they change their life. They share their knowledge. They share their experience of different diets, Wahls and the others.* [SCC8]

In addition, Facebook was a useful tool for people living with MS to connect to others where no other meeting place existed:

> *There is a group on Facebook that is called "live with MS"...And maybe a few years after I was diagnosed, I was thinking about it, cause I was alone.* [SCC8]

> *I facilitate a Facebook group for Atlantic Canada because we're not as big as Melbourne. We have a number of Atlantic Canadian provinces in one group. And because of the pandemic, things have gone mostly online, especially with the MS crowd.* [SCC4]

MS society websites were, in general, described as being useful, particularly at the point of diagnosis. In addition, MS societies provided an initial point of contact for people living with MS to access web-based information, where advice on where to seek further information was available.

Some participants used the Overcoming MS charity website and the related Overcoming MS Facebook group with mixed opinions:

> *I've been on the [OMS] website and MS Facebook groups... where there's so much support as well... The OMS Facebook group, that would be most likely where they would get similar support. Even though it is not always a professor that answers; it's really peer support.* [IC16]

> *I was also on the OMS support group on Facebook... actually I got off the Facebook group in the end. I thought "it's not doing anything for me."* [IC11]

The Overcoming MS website was cited by some participants in both the intervention and standard care arms as a useful source of information, particularly as it related to mindfulness or meditation:

> *Often I'll do 15 minutes of just sitting and meditating in the morning and in the evening. And then when I lay down in the middle of the day, if I do, then I'll do a mindfulness thing off the OMS website.* [IC11]

> *I've done some OMS webinars. And again, they haven't been fully as informative as I would like but I am still doing them.* [SCC10]

XSL•FO
RenderX

Appraising Web-Based Information Sources

The participants in the MSOC consistently described the value of scientific evidence and trust in the sources of information when considering the nature of the web-based information they engaged with. Generally, participants appeared to have a high standard for evidence, and they were vigilant of unsubstantiated information, relying mainly on the reputation of the web-based source to formulate discretionary approaches to seeking web-based information:

> *I don't believe anything which isn't scientifically proven. That's just my analytical mind.* [IC8]
>
> *I really need information that is not "somebody said something somewhere," but some scientifically proven information; checked information.* [SCC8]

The method through which the participants critically appraised sources of information was primarily through name recognition and referrals from trusted institutions.

> *You guys were in my mind "you're a reputable institution doing meaningful work", and I had heard about the study through like that society and through my neuroscience center.* [SCC7]
>
> *I saw the Melbourne University so I thought it's like "alright, I'll do that."* [IC8]

Similarly, the participants were especially conscious of misinformation on the web, whether from perceived experts or from other people living with MS:

> *There's a lot of hoax things on the internet on how to cure, how to manage, how to do this and that.* [IC8]
>
> *I've joined a few sort of forum type things to find a lot of information from there. I tend to apply a sort of 80 / 20 rule to a lot of these things. It's like 20% of it's probably rubbish and 80% is okay.* [SCC16]
>
> *I have looked up things in Google, which has brought me to a blog or something, and I don't care for blogs. I think a lot of the information is negative and the only things I really investigate on my own are the studies or research.* [SC11]

Selective Information Seeking

In addition to appraising web-based information sources, many people living with MS were discriminatory in the content of MS-related information sought on the web, choosing to avoid web-based information or describing a hesitancy toward information seeking, whereas others were keen to learn as much about MS as they could. Notably, many participants described a shift from avoiding information early in their disease to actively seeking it later with time:

> *I decided I don't want to read anything about anyone's progress getting worse. That's why I was seeking out positive stories.* [IC12]

Conversely, many people living with MS were not particularly concerned with encountering negatively valanced information on the web:

> *I've wanted to get as much information, knowledge about my MS as I can. I'm one of those people.* [IC9]
>
> *There are a lot of things that will hit you like a ton of bricks whether you try and avoid it or not...* [SCC6]

Discussion

Principal Findings

People living with MS frequently seek MS-related education information, with the knowledge acquired playing an important role in facilitating self-management of the disease [2,3]. Using thematic analysis to understand how a cohort of people living with MS seek and engage with information related to MS across web-based platforms, this study provides an increased understanding of the information sources and needs of the people living with MS. Drawing from a larger data corpus within the context of the MSOC RCT, we identified how people living with MS seek MS-related information and the perceived barriers they face in accessing and integrating information. Two overarching themes emerged from the data: (1) "motivations for learning," which explored individuals' reasons for seeking and engaging with web-based information in general and the MSOC course specifically and (2) "MS information on the web," pertaining to where the people living with MS seek web-based information and their experiences of sourcing this information.

Comparison With Prior Work

Motivations for Learning

Understanding an individual's motivation to access the MSOC and other web-based information is crucial for identifying personal attributes that may precipitate information seeking and influence both the sources and content of web-based learning. Although there has been a substantial qualitative investigation into the informational needs of people living with MS at diagnosis [10,28-34], including a meta-synthesis of qualitative studies looking at diagnosis more broadly [35], the personal behavior and experiences of people living with MS seeking web-based information, especially at diagnosis, have previously been underinvestigated. Notably, this study identified the time of diagnosis as a crucial time point that precipitated information seeking.

Despite concerns around the disclosure of an MS diagnosis [36], respondents described being encouraged to seek web-based information by their HCPs in addition to being independently motivated to seek information. Subsequently, people living with MS were often overloaded by a great deal of information, which at times led to confusion and jumping "on all sorts of bandwagons." The participants in this study considered this initial self-guided "informational rush" as an intense experience and not necessarily always productive, possibly increasing the likelihood of engaging with misinformation, particularly for newly diagnosed people living with MS [13,31]. The negativity of information found at diagnosis was previously described in a similar study by Solari et al [29].

Although the experience of seeking information was confronting for many participants in the early phase of diagnosis, the process of information seeking itself appeared to provide support as individuals' access to information may have facilitated a sense of illness coherence (ie, understanding of their MS). This is

XSL•FO
RenderX

congruent with prior quantitative findings suggesting that a greater sense of illness coherence may assist people living with MS in managing their emotions [37], which in turn may facilitate seeking MS-related knowledge in a more measured manner that may result in people living with MS accessing more relevant knowledge. Relatedly, many participants reported that participating in the MSOC was rewarding and identified the positive way in which information was presented as the reason. This presents an important consideration for designers of educational content targeted at people living with MS, whereby balancing negativity and accuracy continues to be critical for engaging people living with MS [38].

The importance of MS-related information seeking following diagnosis was further emphasized through respondents' reports that information presented at diagnosis within the health care setting was often too much or too little. Consistent with previous reports [29,30,32], respondents were generally dissatisfied with the information coupled with an emphasis on self-directed information seeking and learning that they received at the point of diagnosis. The task of self-directed information seeking proved challenging for some participants, which may have been a result of the overwhelming amount of information on the web and having to discern its trustworthiness [6,9]. This emphasizes the importance of personalized care at the time of diagnosis during this confronting period [39], including supplying information to patients guided by their readiness and desire for such information. Furthermore, our findings indicate that providing guidance to people living with MS on what type of information to source on the web and where to source information is an important priority for HCPs, particularly around the time of diagnosis. This may be of particular relevance in the current climate, with respondents reporting greater reliance on web-based sources at diagnosis following the COVID-19 pandemic, which has previously been described as less desirable by people living with MS [29].

The desire for enhanced capability in self-managing MS served as an important motivator for information seeking. In the process of simply seeking and gaining new knowledge, the participants hoped to have a sense of greater capacity to manage their MS. Personal control perceptions, referring to an individual's sense of empowerment regarding the effectiveness of personal coping behaviors [39], have been reliably linked to motivation for self-management in chronic disease populations [40], lower levels of depression in people living with MS [41], and better quality of life. Information seeking for this purpose was perceived as a protective action compared with a passive or fatalistic approach to disease management.

MS Information on the Web

The second major theme identified was the sourcing of MS-related health information on the web by the study participants and their relationships with these sources. Facebook was the most commonly reported source of web-based information through official institutional MS Facebook pages or Facebook community groups. Facebook provided not only direct educational content for people but also a sense of peer support where asynchronous communication could take place. Furthermore, Facebook was useful for creating new communities in areas where they previously did not exist, particularly in countries with smaller populations or a low prevalence of MS.

Seeking information through communication with other patients is a common theme identified among chronic disease populations [42], and it represents a motivating factor for people to continue to engage with or develop these web-based communities [43,44]. Furthermore, information seeking through MS Society Facebook pages or charitable organizations such as MS Societies and the Overcoming MS website was also common among the participants. MS Society websites were recognized as an initial point for health information, particularly around the time of diagnosis, whereas the Overcoming MS website was often sought by people living with MS seeking information specifically related to lifestyle modification. Overall, our findings suggest that people living with MS engage with the Facebook community, organization groups, or MS Society websites for general information seeking, information clarification, and community support. Through these web-based sources, people living with MS are able to search for filtered information according to their health needs, such as lifestyle modification information.

As reported previously, the role of trust was of great importance to participants when encountering information on the web [45]; it was emphasized that the respondents required the information presented to be scientifically validated. Although some respondents reported having some scientific skills (eg, identifying studies and interpreting research findings), which facilitated the critical appraisal of information on the web, most respondents relied on the name recognition of institutions. Similar to other studies [6,9,10], misinformation was still a considerable concern for respondents who were cautious of material from unfamiliar sources, particularly information presented as testimony or hearsay from others. Building on the results from prior studies, this study adds context to challenging web-based misinformation, whereby trust in institutions can play a key role in engaging people living with MS with appropriate health content.

People living with MS described encountering negative health-related information on the web frequently. There were 2 common responses to encountering negative information: avoiding or accepting. The most common reason for avoiding information was seeing portrayals of disability. Other studies have similarly described how people living with MS struggle with encountering depictions of disability on the web, particularly close to the diagnosis [28,38]. According to some participants, the immediacy of addressing management options and health concerns played a role in overcoming their initial avoidance of confronting information and ushered them toward broad information seeking. Another subset of participants found it best to accept the presence of negative information from the outset, as this was impossible to avoid when seeking MS-related health information.

Strengths and Limitations

This study has strengths in consolidating the common reasons for seeking web-based MS-related information and preferences for information types and capturing individuals' experiences with this information. However, this study has some limitations.

XSL•FO
RenderX

First, this qualitative study population was drawn from a larger RCT in which people living with MS self-selected their involvement with the study through voluntary participation. This may limit the generalizability to the broader MS population as participants who seek and choose to enroll in web-based programs for MS may possess higher levels of patient activation compared with the general population. Furthermore, the participants completing the course, baseline study, and postevaluation survey and consenting to postcourse interviewing may represent an even more select and motivated subpopulation. Second, the views expressed are those of the interviewed participants and do not represent all potential views. Some participant characteristics are underrepresented in this cohort, such as the number of men and individuals with more progressive MS phenotypes, which may further limit generalizability.

Conclusions

This study of a subset of participants in the MSOC RCT explored information seeking and reasons for seeking this information and engaging with the MSOC and other web-based information. Notably, this study identified previously underexamined barriers to people living with MS seeking information, including the timing of diagnosis and the validity of information available on the web. These factors have the potential to influence individuals' experiences of engaging with health information on the web. These findings should assist HCPs, researchers, and policy makers in determining optimal ways to provide reliable educational support to people living with MS, especially at critical points such as MS diagnosis. Ultimately, as people living with MS continue to engage with increasingly prevalent web-based resources related to their health, a greater understanding of the information they seek is critical for ensuring that they gain the maximum benefit of information in terms of content and potential impact.

Acknowledgments

The authors would like to thank all the participants in this study for their generosity in both time and spirit. Without their contribution, this study would not have been possible. Furthermore, the authors acknowledge the role of the industry partner who assisted in the development of the Multiple Sclerosis Online Course, JMA Creative.

Data Availability

The data sets used for quantitative analysis and anonymized transcripts used for qualitative analysis are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed equally to the critical revisions of the paper, and all authors accept the final responsibility for the decision to submit for publication. WB, SN, GAJ, and JR were responsible for the development of the semistructured question prompts. RD, SN, PJ, JR, and MY were responsible for conducting the semistructured interviews, and WB was responsible for the transcription. WB and RD were responsible for condensing and organizing data. WB and RD were primarily responsible for identifying themes and conducting analysis, with JR and SN involved in the iterative refinement process. WB and RD were responsible for drafting the original manuscript. All authors contributed to redrafting, revising, and submitting the final manuscript.

Conflicts of Interest

GAJ and SN previously received remuneration for facilitating lifestyle modification workshops. GAJ receives royalties from his book "Overcoming Multiple Sclerosis: The 7 Step Recovery Program."

Multimedia Appendix 1
Interview schedule of questions for participants.
[DOCX File , 36 KB - jmir_v26i1e53372_app1.docx]

References

1. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: insights from the Atlas of MS, third edition. Mult Scler 2020 Dec 11;26(14):1816-1821 [FREE Full text] [doi: 10.1177/1352458520970841] [Medline: 33174475]
2. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004 Apr;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
3. Lorig K, Ritter PL, Villa FJ, Armas J. Community-based peer-led diabetes self-management: a randomized trial. Diabetes Educ 2009 Apr 30;35(4):641-651. [doi: 10.1177/0145721709335006] [Medline: 19407333]
4. Kidd T, Carey N, Mold F, Westwood S, Miklaucich M, Konstantara E, et al. A systematic review of the effectiveness of self-management interventions in people with multiple sclerosis at improving depression, anxiety and quality of life. PLoS One 2017;12(10):e0185931 [FREE Full text] [doi: 10.1371/journal.pone.0185931] [Medline: 29020113]

XSL•FO
RenderX

5. Köpke S, Solari A, Rahn A, Khan F, Heesen C, Giordano A. Information provision for people with multiple sclerosis. Cochrane Database Syst Rev 2018 Oct 14;10(10):CD008757 [FREE Full text] [doi: 10.1002/14651858.CD008757.pub3] [Medline: 30317542]
6. Hay MC, Strathmann C, Lieber E, Wick K, Giesser B. Why patients go online: multiple sclerosis, the internet, and physician-patient communication. Neurologist 2008 Nov;14(6):374-381. [doi: 10.1097/NRL.0b013e31817709bb] [Medline: 19008743]
7. Bishop M, Frain MP, Espinosa CT, Stenhoff DM. Sources of information about multiple sclerosis: information seeking and personal, demographic, and MS variables. J Vocat Rehabil 2009;31(2):107-117. [doi: 10.3233/JVR-2009-479]
8. Higueras Y, Salas E, Meca-Lallana V, Carrascal Rueda P, Rodríguez De la Fuente O, Cabello-Moruno R, et al. Information-seeking strategies of people with multiple sclerosis in Spain: the INFOSEEK-MS study. Patient Prefer Adherence 2022 Jan 11;16:51-60 [FREE Full text] [doi: 10.2147/PPA.S344690] [Medline: 35046641]
9. Synnot AJ, Hill SJ, Garner KA, Summers MP, Filippini G, Osborne RH, et al. Online health information seeking: how people with multiple sclerosis find, assess and integrate treatment information to manage their health. Health Expect 2016 Jun;19(3):727-737 [FREE Full text] [doi: 10.1111/hex.12253] [Medline: 25165024]
10. Marrie RA, Salter AR, Tyry T, Fox RJ, Cutter GR. Preferred sources of health information in persons with multiple sclerosis: degree of trust and information sought. J Med Internet Res 2013 Mar 17;15(4):e67 [FREE Full text] [doi: 10.2196/jmir.2466] [Medline: 23635393]
11. Colombo C, Mosconi P, Confalonieri P, Baroni I, Traversa S, Hill SJ, et al. Web search behavior and information needs of people with multiple sclerosis: focus group study and analysis of online postings. Interact J Med Res 2014 Jul 24;3(3):e12 [FREE Full text] [doi: 10.2196/ijmr.3034] [Medline: 25093374]
12. Langhorne AL, Thomas P, Kolaczkowski L. Untangling the web of e-Health: multiple sclerosis patients' perceptions of online health information, information literacy, and the impact on treatment decision making. In: Proceedings of the 5th International Conference, OCSC 2013, Held as Part of HCI International 2013. 2013 Presented at: 5th International Conference, OCSC 2013, Held as Part of HCI International 2013; July 21-26, 2013; Las Vegas, NV. [doi: 10.1007/978-3-642-39371-6_35]
13. Lavorgna L, Lanzillo R, Brescia Morra V, Abbadessa G, Tedeschi G, Bonavita S. Social media and multiple sclerosis in the posttruth age. Interact J Med Res 2017 Sep 27;6(2):e18 [FREE Full text] [doi: 10.2196/ijmr.7879] [Medline: 28954717]
14. Claflin SB, Gates R, Maher M, Taylor BV. Building a successful massive open online course about multiple sclerosis: a process description. J Med Internet Res 2020 Jul 29;22(7):e16687 [FREE Full text] [doi: 10.2196/16687] [Medline: 32723726]
15. Russell RD, Black LJ, Purdue J, Daly A, Begley A. A collaborative approach to designing an online nutrition education program for people with multiple sclerosis. Disabil Rehabil 2023 Mar 12:1-10. [doi: 10.1080/09638288.2023.2186499] [Medline: 36908015]
16. Bessing B, van der Mei I, Taylor BV, Honan CA, Blizzard L, Claflin SB. Evaluating the impact of the understanding multiple sclerosis online course on participant MS knowledge, health literacy, resilience, self-efficacy, quality of life, and MS symptom severity. Mult Scler Relat Disord 2022 Apr;60:103717. [doi: 10.1016/j.msard.2022.103717] [Medline: 35259682]
17. Reece JC, Yu M, Bevens W, Simpson-Yap S, Jelinek G, Jelinek P, et al. Study protocol for an online lifestyle modification education course for people living with multiple sclerosis: the multiple sclerosis online course (MSOC). BMC Neurol 2023 Jun 29;23(1):249 [FREE Full text] [doi: 10.1186/s12883-023-03298-0] [Medline: 37386385]
18. Bevens W, Weiland TJ, Gray K, Neate SL, Nag N, Simpson-Yap S, et al. The feasibility of a web-based educational lifestyle program for people with multiple sclerosis: a randomized controlled trial. Front Public Health 2022 Apr 27;10:852214 [FREE Full text] [doi: 10.3389/fpubh.2022.852214] [Medline: 35570898]
19. Headspace home page. Headspace National Youth Mental Health Foundation. URL: http://headspace.org.au [accessed 2024-01-21]
20. Home - distress centres of Toronto. Distress Centres. URL: https://www.torontodistresscentre.com/ [accessed 2024-01-29]
21. Lifeline Aotearoa home page. Lifeline Aotearoa. URL: https://www.lifeline.org.nz/ [accessed 2024-01-29]
22. Samaritans home page. Samaritans. URL: https://www.samaritans.org/ [accessed 2024-01-29]
23. The lifeline and 988. 988 Suicide & Crisis Lifeline. URL: https://988lifeline.org/current-events/the-lifeline-and-988/ [accessed 2024-01-19]
24. Audio to text automatic transcription service and app. Temi. URL: https://www.temi.com/ [accessed 2024-01-30]
25. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 02;16(1). [doi: 10.1177/1609406917733847]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Campbell KA, Orr E, Durepos P, Nguyen L, Li L, Whitmore C, et al. Reflexive thematic analysis for applied qualitative health research. Qual Report 2021 Jun 20;26(6):2011-2028. [doi: 10.46743/2160-3715/2021.5010]
28. Ghafari S, Fallahi-Khoshknab M, Nourozi K, Mohammadi E. Patients' experiences of adapting to multiple sclerosis: a qualitative study. Contemp Nurse 2015 Jul 27;50(1):36-49. [doi: 10.1080/10376178.2015.1010252] [Medline: 26213256]

XSL•FO
RenderX

29. Solari A, Acquarone N, Pucci E, Martinelli V, Marrosu MG, Trojano M, et al. Communicating the diagnosis of multiple sclerosis - a qualitative study. Mult Scler 2007 Jul 09;13(6):763-769. [doi: 10.1177/1352458506074689] [Medline: 17613605]
30. Fallahi-Khoshknab M, Ghafari S, Nourozi K, Mohammadi E. Confronting the diagnosis of multiple sclerosis: a qualitative study of patient experiences. J Nurs Res 2014 Dec;22(4):275-282 [FREE Full text] [doi: 10.1097/jnr.0000000000000058] [Medline: 25386871]
31. Koopman W, Schweitzer A. The journey to multiple sclerosis: a qualitative study. J Neurosci Nurs 1999 Feb;31(1):17-26. [doi: 10.1097/01376517-199902000-00003] [Medline: 10207829]
32. Heesen C, Schäffler N, Kasper J, Mühlhauser I, Köpke S. Suspected multiple sclerosis - what to do? Evaluation of a patient information leaflet. Mult Scler 2009 Sep 22;15(9):1103-1112. [doi: 10.1177/1352458509106508] [Medline: 19625332]
33. Edwards RG, Barlow JH, Turner AP. Experiences of diagnosis and treatment among people with multiple sclerosis. J Eval Clin Pract 2008 Jun 26;14(3):460-464. [doi: 10.1111/j.1365-2753.2007.00902.x] [Medline: 18373567]
34. Russell RD, Black LJ, Sherriff JL, Begley A. Dietary responses to a multiple sclerosis diagnosis: a qualitative study. Eur J Clin Nutr 2019 Apr 25;73(4):601-608. [doi: 10.1038/s41430-018-0252-5] [Medline: 29941913]
35. Topcu G, Mhizha-Murira JR, Griffiths H, Bale C, Drummond A, Fitzsimmons D, et al. Experiences of receiving a diagnosis of multiple sclerosis: a meta-synthesis of qualitative studies. Disabil Rehabil 2023 Mar 07;45(5):772-783 [FREE Full text] [doi: 10.1080/09638288.2022.2046187] [Medline: 35254195]
36. Scalfari A, Neuhaus A, Daumer M, Ebers GC, Muraro PA. Age and disability accumulation in multiple sclerosis. Neurology 2011 Sep 27;77(13):1246-1252 [FREE Full text] [doi: 10.1212/WNL.0b013e318230a17d] [Medline: 21917763]
37. Calandri E, Graziano F, Borghi M, Bonino S. Coping strategies and adjustment to multiple sclerosis among recently diagnosed patients: the mediating role of sense of coherence. Clin Rehabil 2017 Oct;31(10):1386-1395. [doi: 10.1177/0269215517695374] [Medline: 28933614]
38. Bevens W, Reece J, Jelinek PL, Weiland TJ, Nag N, Simpson-Yap S, et al. The feasibility of an online educational lifestyle program for people with multiple sclerosis: a qualitative analysis of participant semi-structured interviews. Digit Health 2022 Sep 04;8:20552076221123713 [FREE Full text] [doi: 10.1177/20552076221123713] [Medline: 36081754]
39. Leventhal H, Phillips LA, Burns E. The common-sense model of self-regulation (CSM): a dynamic framework for understanding illness self-management. J Behav Med 2016 Dec;39(6):935-946. [doi: 10.1007/s10865-016-9782-2] [Medline: 27515801]
40. Breland JY, Wong JJ, McAndrew LM. Are common sense model constructs and self-efficacy simultaneously correlated with self-management behaviors and health outcomes: a systematic review. Health Psychol Open 2020 Jan 23;7(1):2055102919898846 [FREE Full text] [doi: 10.1177/2055102919898846] [Medline: 32030192]
41. Vaughan R, Morrison L, Miller E. The illness representations of multiple sclerosis and their relations to outcome. Br J Health Psychol 2003 Sep;8(Pt 3):287-301. [doi: 10.1348/135910703322370860] [Medline: 14606974]
42. Kanter E, Bevan JL, Dorros SM. The use of online support groups to seek information about chronic illness: applying the theory of motivated information management. Commun Q 2019;67(1):100-121. [doi: 10.1080/01463373.2018.1539406]
43. Steadman J, Pretorius C. The impact of an online Facebook support group for people with multiple sclerosis on non-active users. Afr J Disabil 2014 Nov 21;3(1):132 [FREE Full text] [doi: 10.4102/ajod.v3i1.132] [Medline: 28730005]
44. Gerritzen EV, Lee AR, McDermott O, Coulson N, Orrell M. Online peer support for people with multiple sclerosis: a narrative synthesis systematic review. Int J MS Care 2022;24(6):252-259 [FREE Full text] [doi: 10.7224/1537-2073.2022-040] [Medline: 36545647]
45. Neate SL, Bevens W, Jelinek PL, Gray KM, Weiland TJ, Nag N, et al. A multiple sclerosis lifestyle behavior online course: qualitative analysis of participants' motivations, expectations and experiences. Front Public Health 2022 Dec 7;10:1022185 [FREE Full text] [doi: 10.3389/fpubh.2022.1022185] [Medline: 36568793]

Abbreviations

ACTRN: Australian and New Zealand Clinical Trials Registry
HCP: health care professional
IC: intervention course
MS: multiple sclerosis
MSOC: Multiple Sclerosis Online Course
RCT: randomized controlled trial
RTA: reflexive thematic analysis
SCC: standard care course

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, T Leung; submitted 06.10.23; peer-reviewed by B Crocker, T Titcomb; comments to author 30.12.23; revised version received 04.01.24; accepted 12.01.24; published 09.02.24.

Please cite as:
Bevens W, Davenport R, Neate S, Yu M, Jelinek P, Jelinek GA, Reece J
Web-Based Health Information Seeking by People Living With Multiple Sclerosis: Qualitative Investigation of the Multiple Sclerosis Online Course
J Med Internet Res 2024;26:e53372
URL: https://www.jmir.org/2024/1/e53372
doi:10.2196/53372
PMID:38335016

©William Bevens, Rebekah Davenport, Sandra Neate, Maggie Yu, Pia Jelinek, George Alexander Jelinek, Jeanette Reece. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Dental Students' Satisfaction With Web-Based Learning During the Initial Phase of the COVID-19 Pandemic: Mixed Methods Study

Minjung Lee[1], PhD; So Youn An[2], PhD, DDS; Jungjoon Ihm[3], PhD

[1]Yale School of Nursing, Yale University, New Haven, CT, United States

[2]College of Dentistry, Wonkwang University, Iksan, Republic of Korea

[3]Dental Research Institute, School of Dentistry, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Jungjoon Ihm, PhD
Dental Research Institute, School of Dentistry
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 08826
Republic of Korea
Phone: 82 8802343 ext 2343
Email: ijj127@snu.ac.kr

Abstract

Background: The COVID-19 pandemic has precipitated an accelerated shift in education, moving from traditional learning to web-based learning. This transition introduced a notable transactional distance (TD) between the instructors and learners. Although disease control and staff and students' safety are the top priorities during a pandemic, the successful delivery of education is equally crucial. However, the ramifications of this swift transition are particularly critical in the context of dental education. Dental education is inherently practice oriented, necessitating hands-on training and manual skills development, which poses unique challenges to distance learning approaches.

Objective: This study aims to examine dental students' web-based learning satisfaction and experience of TD, investigate the predictors of web-based learning satisfaction, and explore the perceptions of students about the advantages and disadvantages of web-based learning.

Methods: This study explored the factors associated with web-based learning satisfaction using TD theory during the transition to web-based education. Psychological factors that could influence satisfaction were adapted from the health belief model. We conducted a cross-sectional web-based survey of 345 dental students from 2 institutions in South Korea who were enrolled in the spring semester of 2020. Data were collected between July 8 and September 14, 2020. Qualitative analysis was used to examine responses to open-ended questions concerning perceptions of web-based learning.

Results: A multivariate hierarchical linear regression model was used to analyze the effects of student characteristics, TD, and psychological factors (ie, perceived risk of infection and efficacy belief of web-based learning) on web-based learning satisfaction. The average score for web-based learning satisfaction was 3.62 (SD 0.84), just above the midpoint of the possible range (1-5). Self-regulated learning (β=0.08; P=.046), learner-instructor interaction (β=0.08; P=.03), and learner-content interaction (β=0.64; P<.001) were associated with higher levels of satisfaction. Moreover, a significant association was revealed between high efficacy beliefs in web-based learning (β=0.20; P<.001) and satisfaction. Although the learning structure (synchronous vs asynchronous) did not exhibit a significant association with satisfaction, the qualitative analysis results revealed that each structure possesses distinct strengths and weaknesses. The students in synchronous learning (79/345, 22.9%) recognized heightened autonomy in the "learning environment" (19/79, 24%); however, technical issues (28/79, 35%) and reduced concentration (15/79, 19%) were identified as downsides. Conversely, the students in asynchronous settings (266/345, 77.1%) emphasized unlimited access to learning content (74/266, 27.8%) and the flexibility of "learning in preferred time" (69/266, 25.9%). Nevertheless, challenges, such as self-management difficulties (66/266, 24.8%) and limited interactions (55/266, 20.7%), were evident.

Conclusions: The findings suggest that efforts to minimize TD, facilitating self-regulated learning and interaction among students and instructors, are critical for achieving web-based learning satisfaction. Moreover, establishing a common understanding among students regarding the necessity and efficacy of web-based learning during epidemics could enhance their satisfaction.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50278) doi:10.2196/50278

KEYWORDS

web-based learning; student satisfaction; transactional distance theory; health profession education; COVID-19

Introduction

Background

The battle to combat COVID-19 has resulted in an unprecedented public health crisis globally. At the early stage of the pandemic, with heightened concerns owing to the scarcity of clinical measures and the emergence of precautionary behaviors, social distancing and surveillance were considered essential for public safety [1]. Education was no exception; educational institutions were forced to close to prevent virus transmission by eliminating or limiting the contact between students and instructors. These changes transformed classroom education into web-based learning via recorded or live lectures and web-based platforms. Moreover, this change has affected >900 million students globally, including undergraduates in preclinical, clinical, and postgraduate medical education [2-4]. In South Korea, the commencement of the new semester in March 2020 was delayed because of the necessary preparations to accommodate an unforeseen shift to web-based education.

Of particular concern is the field of dentistry, which, given its focus on oral practice, faced unique challenges during the COVID-19 pandemic. The nature of the virus, which is primarily transmitted through respiratory means, exacerbated the situation. Consequently, numerous dental practical classes had to be temporarily postponed to prioritize the safety of both student dentists and patients [5-7]. This adjustment in educational practices has led to increased vigilance within health care professional education in South Korea, with a heightened focus on infection control measures in clinical settings. The necessity of these adaptations underscores the resilience and adaptability of the academic community in responding to the challenges posed by the global health crisis.

The transition from traditional learning to web-based learning driven by the COVID-19 pandemic has affected learning in various fields [8-11]. However, the impact of this transition can be especially crucial for dental education. Dental education is a practice-oriented course of study, which cannot depend on distance learning but requires manual skills training and continuous contact with patients [12-14]. Studies on dental students' experiences with and perceptions of web-based learning during the COVID-19 pandemic have reported that dental students found web-based learning favorable owing to the effectiveness, safety, and additional time to connect with family [5,12,15]. In contrast, dissatisfaction because of less interaction with other students and instructors [16]; lower quality teaching compared with traditional face-to-face teaching [17]; and high levels of anxiety, depression, and stress [18] were also revealed. Nonetheless, the studies examining the outcomes of web-based learning because of the pandemic and its predictors among dental students remain limited.

The sudden transition from traditional learning to web-based learning driven by the COVID-19 pandemic increased transactional distance (TD). The TD theory defines TD as "the psychological or communicative space that separates the instructor from the learner in the transaction between them, occurring in the structured or planned learning situation" [19]. According to this theory, 3 sets of variables define the pedagogical aspects of education: structure, learner autonomy, and dialogue. Structure is defined as constructs to be learned and measures an "educational program's responsiveness to learners' individual needs" [20]. Autonomy is a learner characteristic that refers to self-management during interactions with teachers within a designed structure [21]. Finally, dialogue denotes the interaction between instructors and learners during the implementation of a structured program [22]. This interaction is further classified into learner-content, learner-instructor, and learner-learner interactions. In this study, we adopted Moore's theory as a theoretical framework to examine TD in a web-based learning environment and its relationship with learning outcomes.

Furthermore, to deepen our understanding and contribution to studies on web-based learning, we adopted psychological factors from health behavior models, such as the perceived risk of COVID-19 and efficacy belief in web-based learning, to prevent the spread of the virus. The health belief model [23] and the extended parallel process model [24] are among the most famous health behavior models, which propose that perceived risk and efficacy beliefs are key contributors in determining individuals' willingness to make behavioral changes. Multiple studies have reported the robust role of these psychological factors in behavioral responses to reduce the health risk induced by the pandemic [1,25,26]. Although previous studies have revealed a favorable attitude toward the safety and effectiveness of web-based learning [5], perceived risk and efficacy beliefs can relate to the acceptance of the transition from traditional learning to web-based learning and its outcomes.

Objectives

Disease control and ensuring the safety of students and faculty are the top priorities during any public health crisis. However, the quality and outcomes of education should not be compromised. Furthermore, web-based learning has received substantial attention in the field of dentistry and medicine, and its implementation can offer long-term advantages for dental education beyond the challenges posed by the pandemic. Hence, this study focused on identifying the factors influencing learning outcomes in web-based education, specifically focusing on learning satisfaction among dental students. The objectives of this study are (1) to examine dental students' web-based learning satisfaction and experience with TD, (2) to investigate the predictors of web-based learning satisfaction, and (3) to explore the perceptions of students regarding the advantages and disadvantages of web-based learning. This study presents several implications for the development and implementation of web-based learning. We believe that the results of this study will provide evidence and implications for guiding the direction

XSL•FO
RenderX

of web-based education in the field of medicine and dentistry in the post–COVID-19 era. Furthermore, it will contribute to the development of proactive measures in anticipation of potential public health crises in the future.

Methods

Study Design

A cross-sectional web-based survey was designed and conducted during the early stages of the COVID-19 pandemic, inviting students enrolled in the spring 2020 semester to participate. The sample size was calculated using G*Power (latest version 3.1.9.7; Heinrich-Heine-Universität Düsseldorf), which indicated that 315 participants were required to detect small effect sizes (Cohen f^2=0.08) in linear multiple regression models with up to 10 predictor variables, given a 5% type I error rate and 95% power to calculate the sample size [27].

When this study was conducted, the South Korean government requested that the public minimize face-to-face interactions and isolate themselves at home. Most classes were conducted on the web; potential respondents were invited electronically via email and web-based school bulletin boards. An invitation to participate in the study, including a brief introduction to the background and objective of the study, voluntary nature of participation, and declarations of confidentiality and anonymity, was sent via email to dental students of 2 participating dental schools (N=400). The students who enrolled in the spring 2020 semester at Seoul National University and Wonkwang University in South Korea were invited to participate in this study. These 2 schools differ in several aspects: one is a national institution, whereas the other is a private institution; and one is located in the capital city of Seoul, whereas the other is located in a regional area. In addition, they varied in their approach to web-based classes, with one school using a synchronous format and the other using an asynchronous format. These differences make these 2 schools well suited for gathering a more heterogeneous study sample.

Data Collection

A survey questionnaire was developed to (1) evaluate dental students' TD, perceptions, and satisfaction with web-based learning and (2) explore students' attitudes toward web-based learning driven by the COVID-19 pandemic. An anonymous web-based questionnaire was used to evaluate dental students' satisfaction with web-based learning. Specifically, it investigated the effects of dialogue on web-based self-regulated learning, perceived risk of COVID-19, and efficacy beliefs about safety and support during web-based learning. Furthermore, the survey included an open-ended question exploring students' perceived advantages and disadvantages of web-based learning. The survey was conducted using Google Forms. Data were collected from July 8 to September 14, 2020. A total of 345 responses were collected, corresponding to an 86.3% response rate.

Measurements

Satisfaction With Web-Based Learning

Satisfaction with web-based learning was measured using the standardized web-based Online Course Satisfaction Scale (OCSS) [28], which was modified to fit the context of the study. The OCSS presented 7 items on the general contentment level of the students with learning experiences related to instructors and course design. Examples include "I am satisfied with the instructional style" and "I am satisfied with the use of the online discussion forum." An item on overall satisfaction was included (ie, "Overall, I am satisfied with this course"). Items were rated on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). The Cronbach α value was 0.91.

Student Interaction (Learner-Learner, Learner-Instructor, and Learner-Content)

The 3 subscales of interaction (ie, learner-learner, learner-instructor, and learner-content) were adapted from an instrument formulated by Kuo et al [29] to examine student interaction and satisfaction in a blended learning environment. The items were revised to fit the web-based environment of this study. Learner-learner interaction consisted of 8 items that measured the level of communication with fellow students (eg, "I shared my thoughts or ideas about the lectures and their application with other students during this class"). It also included 1 item on overall evaluation (ie, "Overall, I had numerous interactions related to the course content with fellow students"). Learner-instructor interaction consisted of 6 items that measured the level of communication between the learner and instructor and 1 item on overall evaluation (ie, "I asked the instructor my questions through different electronic means, such as email, discussion board, and instant messaging tools"). Finally, learner-content interaction consisted of 4 items (eg, "Online course materials helped me to better understand the class content"). All items in the subscales were measured using a 5-point Likert-type scale, ranging from 1 (strongly disagree) to 5 (strongly agree). The Cronbach α values were 0.92, 0.79, and 0.89, respectively.

Metacognitive Self-Regulation

The metacognitive self-regulation subscale of the Motivated Strategies for Learning Questionnaire assessed self-regulated learning and the extent to which students use planning, monitoring, and regulating strategies during learning [30]. A few examples include "When I become confused about something I'm reading for this class, I go back and try to figure it out," and "I ask myself questions to ensure I understand the material I have been studying in this class." Items were rated on a 7-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). The Cronbach α value was 0.87.

Perceived Risk and Efficacy Belief

Regarding psychological factors, we examined the perceived risk of COVID-19 infection (2 items) comprising the perceived susceptibility (individual beliefs about the possibility of infection during face-to-face learning) and the perceived severity of infection [31]. The items were "What do you think is the possibility that you will contract COVID-19 during face-to-face learning?" and "How severe will the COVID-19 infection be?" Items were rated on a 5-point Likert-type scale (1=very low, 3=neither low nor high, and 5=very high). The Cronbach α value was 0.66. The items such as "To what extent do you think online learning is an effective means for reducing the risk of

XSL•FO
RenderX

COVID-19 infection?" measured the efficacy beliefs. The items were rated on a 5-point Likert-type scale ranging from 1 (not at all) to 5 (extremely).

Furthermore, we investigated whether the students received the necessary support for web-based learning using 1 of the following resources: lecturer, teaching assistant, classmates, web-based support, or none. To conduct the regression analyses, the responses were converted into binary answers (none=0 and otherwise=1). In addition, the survey included 2 open-ended questions: "What do you think are the advantages of learning online?" and "What do you think are the disadvantages of learning online?"

Ethical Considerations

This study was approved by the institutional review board of the School of Dentistry, Wonkwang University, Republic of Korea (number WKIRB-202104-SB-021). All methods were performed in accordance with the relevant guidelines and regulations. All participants were informed of their involvement in this study and gave their written consent through Google Forms, agreeing to the anonymous use of their data for publication. The data were collected and analyzed anonymously. No compensation was provided to the study participants.

Statistical Analyses

Statistical analyses were performed using R software (version 3.5.1; R Foundation for Statistical Computing). The results of the quantitative variables were reported as mean, SD, or frequency (%). A multivariate hierarchical linear regression analysis examined the effects of sociodemographic factors, academic factors, structure, dialogue, self-regulated learning, psychological factors, and web-based learning support on web-based learning satisfaction.

From a qualitative perspective, we conducted content analysis of the responses to the open-ended item. The students' responses from the 2 dental schools were analyzed separately, given the differences in the web-based lecture structure (synchronous vs asynchronous). All responses to the open-ended item were analyzed inductively with open codes and, subsequently, focus coded using the grounded theory approach [32].

From a qualitative perspective, thematic analysis was used for the written responses on the advantages and disadvantages of web-based learning because of the pandemic [33]. This process aimed to find the main patterns of meaning that occurred across the data and a course of data familiarization, generating codes for features of the data, organizing related codes into main themes, and iteratively reviewing and refining themes to fit the data. The data were analyzed using NVivo software (version 12; QSR International). The first author ML developed a coding frame during the manual coding of responses and identified themes that arose from these codes. For credibility, The authors JI and SYA assessed 50 randomly selected responses and identified their corresponding themes [34]. We reviewed the differences in thematic associations in this subsample and the categorization of codes into themes to reach a consensus.

Results

Descriptive Analyses

We collected data from 345 students (male students: 213/345, 61.7%; female students: 132/345, 38.3%), with a mean age of 24.08 (SD 2.68) years (Table 1). The response rate was 86%. In South Korea, dental education operates through 2 parallel tracks: a dental college system comprising a 3-year predental course followed by a 4-year Doctor of Dental Surgery (DDS) degree program and a professional graduate-entry school system that offers a 4-year DDS degree program. The students pursuing the professional graduate-entry system are required to hold a bachelor's degree before admission, whereas those pursuing the dental college system typically enter with a high school diploma. The students from both systems integrate their studies to achieve the qualification of a dentist with a DDS degree [35,36]. Among the participating schools, 1 school operated simultaneously as a dental college and a professional graduate-entry school system, and the other school operated only as a dental college with a predental program. Among the participants, 84.6% (292/345) of them were from *dental colleges* and 15.4% (53/345) of them were from *professional graduate-entry schools*. The average academic duration was 4.18 (SD 1.54) years. Of the 345 students, 266 (77.1%) and 79 (22.9%) took asynchronous and synchronous web-based classes, respectively.

XSL•FO
RenderX

Table 1. Descriptive statistics of the respondents (N=345).

Characteristics	Values
Sex, n (%)	
Male	213 (61.7)
Female	132 (38.3)
Age (years), mean (SD)	24.08 (2.68)
Academic track, n (%)	
Dental college	292 (84.6)
Professional graduate-entry school	53 (15.4)
Academic period (years)	
Value, mean (SD)	4.18 (1.54)
≤3, n (%)	107 (31)
>3, n (%)	238 (69)

Table 2 summarizes each scale's average score and reliability information based on the study sample. The mean OCSS score of the entire sample was 3.62 (SD 0.84), which is close to the midpoint (3) of the possible range (1-5). Each subscale for dialogue had an average score close to the midpoint of the corresponding scale, except for learner-content interaction (mean 3.55, SD 0.86), which exhibited a mean higher than the midpoint. The Cronbach α values were high for each subscale (learner-learner: Cronbach α=0.92 and learner-instructor: Cronbach α=0.79, and learner-content: Cronbach α=0.89). For self-regulated learning, the average score was 4.37 (SD 0.89), with a high Cronbach α value of 0.91. The average score for the perceived risk of COVID-19 infection as a psychological factor was 3.46 (SD 0.99).

Table 2. Descriptive statistics of structure, dialogue, self-regulated learning, and psychological factors related to web-based learning during the COVID-19 pandemic (N=345).

Variables	Values
Structure, n (%)	
Synchronous	79 (22.9)
Asynchronous	266 (77.1)
Dialogue (5-point Likert scale), mean (SD)	
Learner-learner interaction	3.00 (0.86)
Learner-instructor interaction	3.00 (0.52)
Learner-content Interaction	3.55 (0.86)
Self-regulated learning (7-point Likert scale), mean (SD)	4.37 (0.89)
Psychological factors (5-point Likert scale), mean (SD)	
Perceived risk of COVID-19 infection	3.46 (0.99)
Efficacy beliefs of web-based learning on safety	3.99 (1.08)
Web-based learning satisfaction (5-point scale) mean (SD)	3.62 (0.84)
Web-based learning support, n (%)	
Instructor and teaching assistant	57 (16.5)
Classmates	223 (64.6)
Web-based support	42 (12.2)
No support	23 (6.6)

The perceived risk of contracting COVID-19 during face-to-face learning was measured using a 5-point Likert-type scale. The average perceived susceptibility was higher than "neutral" (score=3; mean 3.04, SD 1.20). The results indicated that 11.3% (39/345) and 26.4% (91/345) of the students reported the perceived chance of infection as "very high" (score=5) and "high" (score=4), respectively. Many students (105/345, 30.4%) stated that their chance of infection was "half and half." The average perceived severity score was higher than the perceived susceptibility score, close to "high" (score=4; mean 3.88, SD 1.12). However, 40% (138/345) and 33% (114/345) of the students reported perceived severity as "very high" (score=5)

XSL•FO
RenderX

and "high" (score=4), respectively. Alternatively, the average score for efficacy belief about safety was close to "high" (score=4; mean 3.99, SD 1.08). The majority (138/345, 40%) reported that web-based learning effectively reduced the risk of COVID-19 ("extremely"; score=5).

Furthermore, we investigated whether the students received support during web-based learning. The majority (223/345, 64.6%) reported that they received help from classmates, followed by instructors and teaching assistants (57/345, 16.5%) and web-based support through various platforms, such as email, discussion boards, and instant SMS text messaging tools (42/345, 12.2%). However, of the 345 students, 23 (6.6%) reported receiving no support.

Factors Influencing Students' Web-Based Learning Satisfaction

Hierarchical linear regression models tested the association between the factors influencing web-based learning satisfaction (ie, web-based course structure, dialogue, self-regulated learning, and psychological factors) and web-based learning support (Table 3). Students' sex, self-regulated learning, and learner-content interaction (model 1) accounted for 60.6% of the variance. Psychological factors such as the perceived risk of COVID-19 and the efficacy belief of web-based learning in preventing COVID-19 (model 2) explained an additional 3.5% of the variance. Conversely, web-based learning support, as the predictor variable (model 3), explained an additional 1% of the variance in web-based learning satisfaction. In summary, the effect of learner-content interaction (β=0.64; P<.001) was significant and the highest, followed by efficacy beliefs of the safety of web-based learning (β=0.20; P<.001) and support during web-based learning (β=0.10; P<.001). Furthermore, the positive effects of self-regulated learning (β=0.08; P=.046) and learner-instructor interaction (β=0.08; P=.03) were statistically significant. The effect of the structure of web-based learning was not significant (P=.52).

Table 3. Multivariate hierarchical linear regression analysis of factors of satisfaction with web-based learning.

Variables[a]	Model 1[b]			Model 2[c]			Model 3[d]		
	Unstandardized coefficient, B	Standardized coefficient, β	*P* value	Unstandardized coefficient, B	Standardized coefficient, β	*P* value	Unstandardized coefficient, B	Standardized coefficient, β	*P* value
Sex (male: 1 and female: 2)	–0.141	–0.082	.02	–0.121	–0.070	.03	–0.118	–0.069	.03
Academic period	0.038	0.069	.06	0.036	0.066	.053	0.031	0.056	.10
Academic track (DC[e]: 1 and professional: 2)	–0.128	–0.055	.32	–0.133	–0.057	.28	–0.154	–0.066	.20
Structure (synchronous: 1 and asynchronous: 2)	–0.062	–0.031	.58	–0.058	–0.029	.59	–0.068	–0.034	.52
Self-regulated learning	0.078	0.083	.047	0.078	0.083	.04	0.076	0.080	.046
Learner-learner interaction	0.010	0.011	.78	0.023	0.023	.53	0.009	0.009	.80
Learner-instructor interaction	0.091	0.078	.06	0.097	0.084	.04	0.098	0.084	.03
Learner-content interaction	0.689	0.703	<.001	0.618	0.630	<.001	0.622	0.635	<.001
Perceived risk of COVID-19	N/A[f]	N/A	N/A	–0.038	–0.045	.18	–0.040	–0.047	.16
Efficacy beliefs of web-based learning on safety	N/A	N/A	N/A	0.161	0.207	<.001	0.154	0.197	<.001
Web-based learning support (none: 0 and otherwise: 1)	N/A	N/A	N/A	N/A	N/A	N/A	0.346	0.103	.001

[a]Nominal variable consists of two options (eg, male and female), and the numerals represent the values assigned to each option.

[b]Model 1: R^2=0.64; ΔR^2=0.606; P≤.001.

[c]Model 2: R^2=0.67; ΔR^2=0.035; P≤.001.

[d]Model 3: R^2=0.68; ΔR^2=0.010; P=.001.

[e]DC: dental college.

[f]N/A: not applicable.

XSL•FO
RenderX

Perceived Advantages and Disadvantages of Web-Based Learning

Analysis of responses to the 2 open-ended questions, "What do you think are the advantages of learning online?" and "What do you think are the disadvantages of learning online?" exhibited several themes that frequently recurred (Tables 4 and 5).

Table 4. Perceived advantages and disadvantages of synchronous web-based learning (n=79).

Theme and perceived advantages and disadvantages	Values, n (%)	Description	Excerpts from written responses
Advantages			
Efficient use of time	27 (34)	Reduction in travel time and time saving	• "I like it because I don't have to spend time traveling to take classes." • "I like being able to use the time before and after class efficiently."
The autonomy to choose one's preferred learning environment	19 (24)	Attend classes from a preferred place (eg, home, coffee shop, or anywhere there is Wi-Fi)	• "Taking classes in my own comfortable environment, optimized to learn the most, is the best thing of online learning." • "I can take classes regardless of location."
Protective measure during the pandemic	14 (18)	Reduced risk of COVID-19, maintaining social distancing, not a face-to-face learning, and no need to wear facial masks	• "It's a way to escape from the risk of coronavirus infection." • "I feel safe because I don't have to go to a crowded environment." • "I think it is absolutely necessary to prevent infection in the pandemic era."
Heightened concentration	7 (9)	Easier to concentrate and lecture structure is more focused	• "As I take classes alone, I seem to be able to concentrate and listen more easily." • "Classes are often held with lecture materials on display, allowing you to concentrate more on the lecture materials while listening to the lecture."
Disadvantages			
Technical challenges and disruptions	28 (35)	Usability problems (eg, internet cut off during class and electronic devices needed)	• "There are times when the instructor's microphone, speaker, etc. do not work properly." • "If the internet was cut off, lectures would be interrupted, and attendance checks were often not carried out properly." • "If you happen to miss a class due to an internet connection, etc., it is often inconvenient because you cannot view the recording due to the nature of real-time lectures."
Reduced concentration	15 (19)	Requires more effort to concentrate; Easier to get distracted	• "It is easier to not concentrate or to sneak away and do other things during class." • "The duration of concentration does not last long."
Limited interaction	10 (13)	Limited interaction with classmates and lecturers, hard to speak up, and feeling disconnected	• "It is difficult to actively communicate and interact with professors and students." • "It seems like student participation in class has dropped significantly."

XSL•FO
RenderX

Table 5. Perceived advantages and disadvantages of asynchronous web-based learning (n=266).

Theme and perceived advantages and disadvantages	Values, n (%)	Description	Excerpts from written responses
Advantages			
Unlimited access to lecture content	74 (27.8)	Repeated learning is possible and reviewing any particularly difficult part of the lecture is possible	• "If there's something you don't know, you can rewatch it at any time. It was very helpful to understood some parts." • "It's good because you can repeat the class and have time to pause and think when necessary."
The autonomy to choose one's preferred learning time	69 (25.9)	Attending to classes when one is prepared and when one can focus the most and the option to progress through the class at one's own pace	• "Because I can study when I am ready, I have deeper concentration and understanding." • "I have the advantage of being able to pause and listen if I lack concentration." • "Learning online helps me learn at my own pace and in my own way, which helps boost my motivation."
Protective measures during the pandemic	48 (18)	Reduced risk of COVID-19, maintaining social distancing, not a face-to-face learning, and no need to wear facial masks	• "Protecting the safety of students and instructors is of utmost importance in this time." • "The spread of coronavirus can be prevented in college-related populations."
Disadvantages			
Difficulties with self-regulated learning	66 (24.8)	Difficulty in attending classes consistently and one must manage their own learning well	• "Too much freedom making me lazy." • "Pushing away from studying and anxiety right before exams."
Limited interaction	55 (20.7)	No real-time feedback, difficulty in asking questions, lack of direct communication, and feeling disconnected	• "The downside is that it is difficult to immediately resolve questions when they arise. It can be only resolved through email or other communication methods." • "Lack of interaction between professors and students makes it hard to form a relationship." • "Unable to go to school or meet classmates."
Increased class assignments	27 (10.2)	Instructor assigns more assignments to students for learning management	• "Since attendance could not be confirmed non-face-to-face, a lot of homework was given to prove attendance." • "Sometimes I felt burdened because there were too many assignments in one day."

Regarding the advantages, both students in synchronous and asynchronous learning consistently identified heightened autonomy as a positive aspect of web-based learning. However, there were nuanced differences between the 2 settings. The students in synchronous settings emphasized autonomy in the "learning environment" (19/79, 24%), whereas those in asynchronous settings stressed the flexibility of "learning in preferred time" (69/266, 25.9%):

> *Taking classes in my own comfortable environment, optimized to learn the most, is the best thing of online learning.* [Student #18, synchronous setting]
>
> *Learning online helps me learn at my own pace and in my own way, which helps boost my motivation.* [Student #23, asynchronous setting]

Moreover, a shared perception among students in both synchronous (14/79, 18%) and asynchronous (48/266, 18%) settings was that web-based learning served as a protective measure to reduce the risk of infection for themselves and their families:

> *I feel safe because I don't have to go to a crowded environment.* [Student #56, synchronous setting]
>
> *Protecting the safety of students and instructors is of utmost importance in this time.* [Student #10, asynchronous setting]

However, beyond these advantages, challenges in self-management of learning were also evident. The students in synchronous settings mentioned the effort required to concentrate during class (15/79, 19%). In contrast, the students in asynchronous settings highlighted the advantage of learning at their preferred times (69/266, 25.9%):

> *The duration of concentration does not last long.* [Student #6, synchronous setting]
>
> *Because I can study when I am ready, I have deeper concentration and understanding.* [Student #110, asynchronous setting]

Technical challenges were more pronounced in the synchronous setting (28/79, 35%), with issues such as malfunctioning

XSL•FO
RenderX

microphones and speakers. In contrast, limited interaction emerged as a more common response among students in asynchronous settings (55/266, 20.7%), expressing difficulty in forming relationships because of a lack of interaction between professors and students:

> *There are times when the instructor's microphone, speaker, etc. do not work properly.* [Student #123, synchronous setting]
>
> *Lack of interaction between professors and students makes it hard to form a relationship.* [Student #203, asynchronous setting]

Discussion

Principal Findings

The study's findings revealed student satisfaction and experiences with web-based learning, particularly regarding TD, during the abrupt shift from traditional learning to web-based education in the early stages of the COVID-19 pandemic. Given the determinants of TD, higher self-regulated learning, learner-instructor interaction, and learner-content interaction were related to higher satisfaction levels. The learning structure (synchronous vs asynchronous) did not reveal a significant association with satisfaction. Nevertheless, the qualitative data analysis results of this study clearly illustrated the advantages and disadvantages of web-based classes in both synchronous and asynchronous structures. Notably, the high efficacy belief scores that web-based learning is an effective means of reducing the risk of COVID-19 were related to increased satisfaction levels with web-based learning.

Several findings from the results are worth noting. Learner-content interaction was the strongest predictor of student satisfaction, which is consistent with the results of previous studies that were conducted before the pandemic [29,37]. Learner-content interaction refers to a 1-way process of reflecting on the course content or course material, and it occurs when learners talk or think about the knowledge and information comprising the course experience [29,38]. Learner-content interaction is an essential cognitive process for organizing and reflecting on new knowledge by integrating prior knowledge [22,38]. Students spend most of their time on the required reading or assignments; thus, they should be supported and receive timely and sufficient learning materials and content.

Although numerous studies identified 2-way communication, such as "learner-learner" [39,40] and "learner-instructor" interactions [39,41-43], as the most important contributors to satisfaction with web-based learning, learner-learner interaction had a weak relationship with satisfaction in this study. A possible explanation for this finding may lie in the specific course objectives and instructional goal orientation. Learner-learner interaction is relevant to group discussions, group projects, and idea sharing, which occur during collaborative activities [40]. In addition, limited opportunities for interaction between students and instructors during web-based learning represented a significant barrier to web-based learning during the first year of the pandemic [44-46]. This study was conducted in the first semester of implementing web-based classes because of the COVID-19 pandemic; therefore, the limitation of interaction may also be applicable in this case.

Regarding learner-instructor interaction, the students' desires or expectations regarding interaction with instructors may be low, leading to less impact on satisfaction with web-based learning. Students' expectations are closely related to their values and preferences, which directly influence their perception of learning experiences. From the perspective of fulfillment theory, satisfaction is achieved by obtaining an agent's desired state of affairs [47]. Even before the pandemic, students in South Korea typically refrained from being active in class by asking questions because of the psychological stress and cultural norms that inhibited active participation [48]. The educational environments in South Korea are not conducive to training students for questioning and discussion. In contrast, lectures with examination-oriented evaluations bring a competitive atmosphere to the class and only require students to listen to instructors. Therefore, the interactions between instructors and students are frequently low in South Korea [49], and there is a tendency of students to prefer web-based chats or calls rather than traditional forms of communication [50].

This study emphasizes the significance of efficacy beliefs, specifically individuals' perceptions of the effectiveness of web-based learning in mitigating the risk of COVID-19, as a significant predictor of satisfaction. Additionally, one-fifth of the students surveyed indicated that the primary advantage of web-based education was its efficient and effective protection against COVID-19. Given that students perceived the likelihood of contracting COVID-19 to be on par with equal probability and perceived the severity of the virus to be high, efficacy beliefs likely played a crucial role in their willingness to adopt web-based learning. Numerous studies have tested the effect of efficacy beliefs on prevention behaviors for COVID-19 [1,25,26], cancer [51], driving [52], and the risk of using chemical products [53]. Therefore, although web-based classes were not the students' preferred choice, firmer efficacy beliefs could lead to high levels of satisfaction with web-based learning. Similarly, previous studies have shown that psychological safety and perceived usefulness in web-based medical education are crucial to student satisfaction with web-based learning [54-56]. Therefore, this study's results suggest that schools must try to identify a shared understanding and consensus with faculty, staff, students, and parents about the need for web-based learning to ensure their safety against public health emergencies such as COVID-19 to increase satisfaction. In other words, to manage the increasing anxiety among school members and to help reach an understanding among them, providing concrete grounds for decisions through open and sufficient communication is necessary for the school leadership [57].

The results of this study provide implications for education and future research. First, understanding the differences between synchronous and asynchronous learning when selecting a learning structure for web-based learning is essential. A distinction was noted between the perceptions of students regarding the advantages and disadvantages of web-based learning structures. Both students from synchronous and asynchronous learning cited autonomy of learning as an

XSL•FO
RenderX

advantage of web-based learning. However, it is noteworthy that although students from synchronous learning predominantly mentioned the autonomy of location, students from asynchronous learning primarily highlighted the advantage of time management. In addition, synchronous teaching is preferred when peer interactions are required and critical thinking skills are taught at the novice level, whereas asynchronous instruction can facilitate autonomy among students [58,59]. Hybrid learning, which uses synchronous and asynchronous elements, may be effective for teaching integrated content or clinical scenarios [60]. Moreover, it may facilitate learner-instructor and learner-learner interactions and self-regulated learning [61]. Selecting between synchronous and asynchronous structures is critical for educators and students [20,62].

Second, communication models should be developed to facilitate 2-way interaction in web-based learning. Scholars have introduced several communication models for instructors and students in distance learning environments [63,64]. However, these models focus on the examination and content analysis of interactions instead of facilitating or educating instructors. Encouraging 2-way communication (eg, requiring students to keep their "cameras on" to foster appropriate videoconferencing behavior) [65] and using appropriate communication tools [66] can encourage students to contact instructors when they require further assistance. Moreover, efforts to strengthen students' self-regulated learning are important. This study demonstrated that self-regulated learning influences affective outcomes, such as satisfaction and academic achievement or performance [29,67]. Web-based learning's flexibility requires students to use additional self-regulatory skills [68], which can be taught before starting an web-based course or by providing support for skill development within the course itself.

This study has several limitations. First, we did not investigate other forms of interaction in web-based learning, which were suggested as additional factors to be incorporated into Moore's model [19], such as learner-interface [69], learner-task [70], and learner-tool [71] interactions. Additional studies examining the influence of these interactions are required. Second, despite being conducted at >1 institution, the study could not investigate the potential predictors of satisfaction at the institutional level, such as tuition. Factors related to the institutional level may contribute to student satisfaction, which warrants further study.

Several studies have examined objective learning outcomes during the pandemic, such as grade point averages and test scores [8,72]. Subjective outcomes such as student satisfaction are equally important and of increasing interest [9-11]. Student satisfaction reflects the extent to which the learner positively perceives learning experiences and is an essential indicator of program-related and student-related outcomes, such as increased persistence and commitment to the program [73-75]. However, future studies investigating both subjective and objective learning outcomes could provide valuable implications for enhancing web-based learning.

Conclusions

A call has been issued for dental educators to share their stories, practices, and experiences during the COVID-19 pandemic. The results of this study indicate that designing a learning structure to effectively achieve specific course objectives and facilitate student interactions is essential. Interactions between students and content were the most predictable factors for satisfaction. Therefore, support should be provided to ensure timely and sufficient learning materials and content. Furthermore, support is needed to strengthen the self-regulated learning of students. The results of this study can inform decision-making related to future policies, practices, and research. Moreover, understanding students' perceptions of and satisfaction with web-based learning will aid policy makers and university authorities in planning and managing dental education to prepare for future crises.

Data Availability

The data sets generated during this study are not publicly available owing to the limitations of ethics approval involving student data and anonymity but are available from the corresponding author on reasonable request.

Authors' Contributions

All authors have reviewed and approved the manuscript. ML conducted data analysis and drafted the paper as the first author. JI and ML participated in the conceptualization and methodology of the study. JI and SYA contributed to data collection and data discussions. JI conducted the manuscript review and editing as the corresponding author.

Conflicts of Interest

None declared.

References

1. Lee M, You M. Psychological and behavioral responses in South Korea during the early stages of coronavirus disease 2019 (COVID-19). Int J Environ Res Public Health 2020 Apr 25;17(9):2977 [FREE Full text] [doi: 10.3390/ijerph17092977] [Medline: 32344809]
2. Gordon M, Patricio M, Horne L, Muston A, Alston SR, Pammi M, et al. Developments in medical education in response to the COVID-19 pandemic: a rapid BEME systematic review: BEME Guide No. 63. Med Teach 2020 Aug 26;42(11):1202-1215. [doi: 10.1080/0142159x.2020.1807484]

XSL•FO
RenderX

3. Liu X, Zhou J, Chen L, Yang Y, Tan J. Impact of COVID-19 epidemic on live online dental continuing education. Eur J Dent Educ 2020 Nov 30;24(4):786-789 [FREE Full text] [doi: 10.1111/eje.12569] [Medline: 32648989]
4. Kabir H, Hasan MK, Mitra DK. E-learning readiness and perceived stress among the university students of Bangladesh during COVID-19: a countrywide cross-sectional study. Ann Med 2021 Dec 10;53(1):2305-2314 [FREE Full text] [doi: 10.1080/07853890.2021.2009908] [Medline: 34889699]
5. Herr L, Jih MK, Shin J, Chae YK, Lee HS, Choi SC, et al. The perspective of undergraduate dental students on web-based learning in pediatric dentistry during the COVID-19 pandemic: a Korean multicenter cross-sectional survey. BMC Med Educ 2021 Sep 25;21(1):505 [FREE Full text] [doi: 10.1186/s12909-021-02928-w] [Medline: 34563185]
6. Si J. Medical students' self-directed learning skills during online learning amid the COVID-19 pandemic in a Korean medical school. Korean J Med Educ 2022 Jun;34(2):145-154 [FREE Full text] [doi: 10.3946/kjme.2022.226] [Medline: 35676881]
7. Lim J, Ko H, Park J, Ihm J. Effect of active learning and online discussions on the academic performances of dental students. BMC Med Educ 2022 Apr 25;22:312. [doi: 10.1186/s12909-022-03377-9]
8. Prigoff J, Hunter M, Nowygrod R. Medical student assessment in the time of COVID-19. J Surg Educ 2021 Mar;78(2):370-374 [FREE Full text] [doi: 10.1016/j.jsurg.2020.07.040] [Medline: 32819868]
9. Coughlan J, Timuş D, Crnic T, Srdoč D, Halton C, Dragan IF. Impact of COVID-19 on dental education in Europe: the students' perspective. Eur J Dent Educ 2022 Aug 15;26(3):599-607. [doi: 10.1111/eje.12736] [Medline: 34882932]
10. Dutta S, Ambwani S, Lal H, Ram K, Mishra G, Kumar T, et al. The satisfaction level of undergraduate medical and nursing students regarding distant preclinical and clinical teaching amidst COVID-19 across India. Adv Med Educ Pract 2021 Feb;Volume 12:113-122. [doi: 10.2147/amep.s290142]
11. Fatani TH. Student satisfaction with videoconferencing teaching quality during the COVID-19 pandemic. BMC Med Educ 2020 Oct 31;20(1):396 [FREE Full text] [doi: 10.1186/s12909-020-02310-2] [Medline: 33129295]
12. Schlenz MA, Wöstmann B, Krämer N, Schulz-Weidner N. Update of students' and lecturers' perspectives on online learning in dental education after a five-semester experience due to the SARS-CoV-2 (COVID-19) pandemic: insights for future curriculum reform. BMC Med Educ 2023 Aug 08;23(1):556 [FREE Full text] [doi: 10.1186/s12909-023-04544-2] [Medline: 37553629]
13. Hattar S, AlHadidi A, Sawair FA, Alraheam IA, El-Ma'aita A, Wahab FK. Impact of COVID-19 pandemic on dental education: online experience and practice expectations among dental students at the University of Jordan. BMC Med Educ 2021 Mar 08;21(1):151 [FREE Full text] [doi: 10.1186/s12909-021-02584-0] [Medline: 33685451]
14. Goetz K, Wenz HJ, Hertrampf K. Certainty in uncertain times: dental education during the COVID-19 pandemic-a qualitative study. Int J Environ Res Public Health 2023 Feb 10;20(4):3090 [FREE Full text] [doi: 10.3390/ijerph20043090] [Medline: 36833785]
15. Maragha T, Dempster L, Shuler C, Lee V, Mendes V, von Bergmann H. Exploring students' perspectives from two Canadian dental schools toward online learning experiences. J Dent Educ 2023 Jul 27;87(7):1047-1056. [doi: 10.1002/jdd.13207] [Medline: 36973937]
16. Di Giacomo P, Di Paolo C. COVID-19 and dental distance-based education: students' perceptions in an Italian university. BMC Med Educ 2021 Aug 02;21(1):414 [FREE Full text] [doi: 10.1186/s12909-021-02840-3] [Medline: 34340662]
17. Otuyemi OD, Jebose OJ. Experience and perceptions of nigerian undergraduate dental students to virtual learning during the COVID-19 pandemic. Niger J Clin Pract 2023 May;26(5):566-573 [FREE Full text] [doi: 10.4103/njcp.njcp_265_22] [Medline: 37357471]
18. Braz-José C, Morais Caldas I, de Azevedo Á, Pereira M. Stress, anxiety and depression in dental students: impact of severe acute respiratory syndrome-coronavirus 2 pandemic. Eur J Dent Educ 2023 Aug 20;27(3):700-706 [FREE Full text] [doi: 10.1111/eje.12858] [Medline: 36176054]
19. Moore MG. Theory of transactional distance. In: Keegan D, editor. Theoretical Principles of Distance Education. Milton Park, UK: Routledge; 1997.
20. Roach VA, Attardi SM. Twelve tips for applying Moore's Theory of Transactional Distance to optimize online teaching. Med Teach 2021 Apr 21;44(8):859-865. [doi: 10.1080/0142159x.2021.1913279]
21. Shearer RL, Park E. The theory of transactional distance. In: Jung I, editor. Open and Distance Education Theory Revisited: Implications for the Digital Era. Singapore, Singapore: Springer Nature; 2019.
22. Moore MG, Kearsley G. Distance Education: A Systems View of Online Learning. Boston, MA: Cengage Learning; 2011.
23. Janz NK, Becker MH. The Health Belief Model: a decade later. Health Educ Q 1984 Sep 04;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
24. Witte K. Putting the fear back into fear appeals: the extended parallel process model. Commun Monogr 1992 Dec;59(4):329-349. [doi: 10.1080/03637759209376276]
25. Lee M, Kang BA, You M. Knowledge, attitudes, and practices (KAP) toward COVID-19: a cross-sectional study in South Korea. BMC Public Health 2021 Feb 05;21(1):295 [FREE Full text] [doi: 10.1186/s12889-021-10285-y] [Medline: 33546644]
26. Lee M, You M. Effects of COVID-19 emergency alert text messages on practicing preventive behaviors: cross-sectional web-based survey in South Korea. J Med Internet Res 2021 Feb 25;23(2):e24165 [FREE Full text] [doi: 10.2196/24165] [Medline: 33544691]
27. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]

XSL•FO
RenderX

28. Wei HC, Chou C. Online learning performance and satisfaction: do perceptions and readiness matter? Distance Educ 2020 Feb 16;41(1):48-69. [doi: 10.1080/01587919.2020.1724768]
29. Kuo YC, Walker AE, Schroder KE, Belland BR. Interaction, internet self-efficacy, and self-regulated learning as predictors of student satisfaction in online education courses. Internet High Educ 2014 Jan;20:35-50. [doi: 10.1016/j.iheduc.2013.10.001]
30. Pintrich PR, Smith DA, Garcia T, Mckeachie WJ. Reliability and predictive validity of the motivated strategies for learning questionnaire (Mslq). Educ Psychol Meas 2016 Sep 07;53(3):801-813. [doi: 10.1177/0013164493053003024]
31. Rimal RN, Juon HS. Use of the risk perception attitude framework for promoting breast cancer prevention. J Appl Soc Pyschol 2010 Feb 16;40(2):287-310. [doi: 10.1111/j.1559-1816.2009.00574.x]
32. Charmaz K, Belgrave LL. Qualitative interviewing and grounded theory analysis. In: Gubrium JF, Holstein JA, Marvasti AB, McKinney KD, editors. The SAGE Handbook of Interview Research: The Complexity of the Craft. Thousand Oaks, CA: SAGE Publications, Inc; 2012.
33. Krippendorff K. The Content Analysis Reader. Thousand Oaks, CA: SAGE Publications, Inc; 2009.
34. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018 Dec 05;24(1):120-124 [FREE Full text] [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]
35. Lee M, Song Y, You M, Park SY, Ihm J. Dentists' attitudes toward patient-centered care and its predictors: a cross-sectional study in South Korea. BMC Oral Health 2023 Feb 06;23(1):75 [FREE Full text] [doi: 10.1186/s12903-023-02791-9] [Medline: 36747198]
36. Lee M, Ihm J. Empathy and attitude toward communication skill learning as a predictor of patient-centered attitude: a cross-sectional study of dental students in Korea. BMC Med Educ 2021 Apr 21;21(1):225 [FREE Full text] [doi: 10.1186/s12909-021-02674-z] [Medline: 33882935]
37. Prabhu M NB, Bolar K, Mallya J, Roy P, Payini V, K T. Determinants of hospitality students' perceived learning during COVID 19 pandemic: role of interactions and self-efficacy. J Hosp Leis Sport Tour Educ 2022 Jun;30:100335 [FREE Full text] [doi: 10.1016/j.jhlste.2021.100335] [Medline: 34512192]
38. Moore MG. Editorial: three types of interaction. Am J Distance Educ 1989;3(2):1-7 [FREE Full text]
39. Sher A. Assessing the relationship of technology-mediated interaction to student learning in web-based e-learning environment. In: Proceedings of the 2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT). 2009 Presented at: ICADIWT; August 04-06, 2008; Ostrava, Czech Republic URL: https://ieeexplore.ieee.org/document/4664403 [doi: 10.1109/icadiwt.2008.4664403]
40. Jung I, Choi S, Lim C, Leem J. Effects of different types of interaction on learning achievement, satisfaction and participation in web-based instruction. Innov Educ Teach Int 2002 Jan;39(2):153-162. [doi: 10.1080/14703290252934603]
41. Battalio J. Interaction online: a reevaluation. Q Rev Distance Educ 2007;8(4):339-352.
42. Bolliger DU. Key factors for determining student satisfaction in online courses. Int J E Learn 2004;3(1):61-67 [FREE Full text]
43. Chan EH, Chan VH, Roed J, Chen JY. Observed interactions, challenges, and opportunities in student-led, web-based near-peer teaching for medical students: interview study among peer learners and peer teachers. JMIR Med Educ 2023 May 15;9:e40716 [FREE Full text] [doi: 10.2196/40716] [Medline: 37184931]
44. Alpert JB, Young MG, Lala SV, McGuinness G. Medical student engagement and educational value of a remote clinical radiology learning environment: creation of virtual read-out sessions in response to the COVID-19 pandemic. Acad Radiol 2021 Jan;28(1):112-118. [doi: 10.1016/j.acra.2020.09.011] [Medline: 33268209]
45. Bastos RA, Carvalho DR, Brandão CF, Bergamasco EC, Sandars J, Cecilio-Fernandes D. Solutions, enablers and barriers to online learning in clinical medical education during the first year of the COVID-19 pandemic: a rapid review. Med Teach 2021 Oct 05;44(2):187-195. [doi: 10.1080/0142159x.2021.1973979]
46. Papan C, Schmitt M, Becker SL. Teaching medical microbiology with a web-based course during the COVID-19 pandemic: retrospective before-and-after study. JMIR Med Educ 2023 Feb 27;9:e39680 [FREE Full text] [doi: 10.2196/39680] [Medline: 36848212]
47. Murphy MC. The simple desire - fulfillment theory. Nous 2002 Dec 17;33(2):247-272. [doi: 10.1111/0029-4624.00153]
48. Kim JM, Pak SJ. The role of cognitive strategic questioning in the changes of students' conceptions about heat and temperature. J Korean Assoc Sci Educ 1995;15(3):363-370.
49. Jung M, Kim JH, Kim J, So HJ. Mark-on: encouraging student questions in class. In: Proceedings of HCI Korea. 2016 Presented at: HCIK '16; January 27-29, 2016; Jeongseon, Republic of Korea. [doi: 10.17210/hcik.2016.01.79]
50. Lee K, Fanguy M, Lu XS, Bligh B. Student learning during COVID-19: it was not as bad as we feared. Distance Educ 2021 Feb 10;42(1):164-172. [doi: 10.1080/01587919.2020.1869529]
51. Sullivan HW, Burke Beckjord E, Finney Rutten LJ, Hesse BW. Nutrition-related cancer prevention cognitions and behavioral intentions: testing the risk perception attitude framework. Health Educ Behav 2008 Dec 30;35(6):866-879 [FREE Full text] [doi: 10.1177/1090198108326164] [Medline: 19011220]
52. Dillow MR, Walsh AG, Spellman F, Quirk M. Testing the risk perception attitude framework in the context of texting while driving. Commun Res Reports 2015 Mar 26;32(2):132-142. [doi: 10.1080/08824096.2015.1016144]

XSL•FO
RenderX

53. Lee M, You M. Safety behaviors to reduce risk of using chemical household products: an application of the risk perception attitude framework. Int J Environ Res Public Health 2020 Feb 27;17(5):1528 [FREE Full text] [doi: 10.3390/ijerph17051528] [Medline: 32120915]
54. McLeod E, Gupta S. The role of psychological safety in enhancing student engagement in online synchronous learning. Med Teach 2021 Jun 14;44(3):336-337. [doi: 10.1080/0142159x.2021.1937588]
55. Rajeh MT, Abduljabbar FH, Alqahtani SM, Waly FJ, Alnaami I, Aljurayyan A, et al. Students' satisfaction and continued intention toward e-learning: a theory-based study. Med Educ Online 2021 Dec 02;26(1):1961348 [FREE Full text] [doi: 10.1080/10872981.2021.1961348] [Medline: 34338161]
56. Nishimura Y, Ochi K, Tokumasu K, Obika M, Hagiya H, Kataoka H, et al. Impact of the COVID-19 pandemic on the psychological distress of medical students in Japan: cross-sectional survey study. J Med Internet Res 2021 Feb 18;23(2):e25232 [FREE Full text] [doi: 10.2196/25232] [Medline: 33556033]
57. Park SW, Jang HW, Choe YH, Lee KS, Ahn YC, Chung MJ, et al. Avoiding student infection during a Middle East respiratory syndrome (MERS) outbreak: a single medical school experience. Korean J Med Educ 2016 Jun;28(2):209-217 [FREE Full text] [doi: 10.3946/kjme.2016.30] [Medline: 27240893]
58. Iyer P, Aziz K, Ojcius DM. Impact of COVID-19 on dental education in the United States. J Dent Educ 2020 Jun 27;84(6):718-722. [doi: 10.1002/jdd.12163] [Medline: 32342516]
59. Stuart J, O'Donnell AW, Scott R, O'Donnell K, Lund R, Barber B. Asynchronous and synchronous remote teaching and academic outcomes during COVID-19. Distance Educ 2022 Jun 20;43(3):408-425. [doi: 10.1080/01587919.2022.2088477]
60. Bennardo F, Buffone C, Fortunato L, Giudice A. COVID-19 is a challenge for dental education-A commentary. Eur J Dent Educ 2020 Nov 28;24(4):822-824 [FREE Full text] [doi: 10.1111/eje.12555] [Medline: 32542796]
61. Ballouk R, Mansour V, Dalziel B, McDonald J, Hegazi I. Medical students' self-regulation of learning in a blended learning environment: a systematic scoping review. Med Educ Online 2022 Dec 28;27(1):2029336 [FREE Full text] [doi: 10.1080/10872981.2022.2029336] [Medline: 35086439]
62. Ogunbiyi MO, Obiri-Darko E. Medical students' corner: barriers to communication during the COVID-19 pandemic. JMIR Med Educ 2020 Nov 27;6(2):e24989 [FREE Full text] [doi: 10.2196/24989] [Medline: 33197232]
63. Oliver R, McLoughlin C. An investigation of the nature and form of interactions in Live Interactive Television. Educational Resources Information Center (Eric). 1996. URL: https://files.eric.ed.gov/fulltext/ED396738.pdf [accessed 2024-02-16]
64. Offir B, Barth I, Lev Y, Shteinbok A. Teacher–student interactions and learning outcomes in a distance learning environment. Internet High Educ 2003 Jan;6(1):65-75. [doi: 10.1016/s1096-7516(02)00162-8]
65. Leung HT, Bruce H, Korszun A. To see or not to see: should medical educators require students to turn on cameras in online teaching? Med Teach 2021 Jan 21;43(9):1099. [doi: 10.1080/0142159x.2021.1873258]
66. Malik AS, Malik RH. Twelve tips for conducting team-based learning session online in synchronous setting. Med Teach 2021 Apr 09;44(5):486-493. [doi: 10.1080/0142159x.2021.1910642]
67. Mcmanus TF. Individualizing instruction in a web-based hypermedia learning environment: nonlinearity, advance organizers, and self-regulated learners. J Interact Learn Res 2000;11(2):219-251 [FREE Full text]
68. Artino AR. Online military training: using a social cognitive view of motivation and self-regulation to understand students' satisfaction, perceived learning, and choice. Q Rev Distance Educ 2007;8(3):191.
69. Gunawardena CN, Lowe CA, Anderson T. Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. J Educ Comput Res 1998 Feb 01;17(4):397-431. [doi: 10.2190/7mqv-x9uj-c7q3-nrag]
70. Herrington J, Reeves TC, Oliver R. Authentic tasks online: a synergy among learner, task, and technology. Distance Educ 2006 Aug;27(2):233-247. [doi: 10.1080/01587910600789639]
71. Hirumi A. A framework for analyzing and designing e-learning interactions. In: Juwah C, editor. Interactions in Online Education: Implications for Theory and Practice. Milton Park, UK: Routledge; 2007:46-71.
72. Inoue N, Aldosari M, Park SE, Ohyama H. The impact of COVID-19 pandemic on student performance and self-evaluation in preclinical operative dentistry. Eur J Dent Educ 2022 May 18;26(2):377-383 [FREE Full text] [doi: 10.1111/eje.12713] [Medline: 34520609]
73. Yukselturk E, Yildirim Z. Investigation of interaction, online support, course structure and flexibility as the contributing factors to students' satisfaction in an online certificate program. Educ Technol Soc 2008;11(4):51-65.
74. Gray JA, DiLoreto M. The effects of student engagement, student satisfaction, and perceived learning in online learning environments. Int J Leadersh Educ Prep 2016;11(1) [FREE Full text]
75. Shen D, Cho MH, Tsai CL, Marra R. Unpacking online learning experiences: online learning self-efficacy and learning satisfaction. Internet High Educ 2013 Oct;19:10-17. [doi: 10.1016/j.iheduc.2013.04.001]

Abbreviations

DDS: Doctor of Dental Surgery
OCSS: Online Course Satisfaction Scale
TD: transactional distance

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 25.06.23; peer-reviewed by SH Jang, J Mira; comments to author 02.10.23; revised version received 07.11.23; accepted 04.02.24; published 08.03.24.

Please cite as:
Lee M, An SY, Ihm J
Dental Students' Satisfaction With Web-Based Learning During the Initial Phase of the COVID-19 Pandemic: Mixed Methods Study
J Med Internet Res 2024;26:e50278
URL: https://www.jmir.org/2024/1/e50278
doi:10.2196/50278
PMID:38457808

©Minjung Lee, So Youn An, Jungjoon Ihm. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Utility of Large Language Models for Health Care Professionals and Patients in Navigating Hematopoietic Stem Cell Transplantation: Comparison of the Performance of ChatGPT-3.5, ChatGPT-4, and Bard

Elisabetta Xue[1], MD; Dara Bracken-Clarke[1], MD; Giovanni Maria Iannantuono[2], MD; Hyoyoung Choo-Wosoba[3], PhD; James L Gulley[1], MD; Charalampos S Floudas[1], MD

[1]Center for Immuno-Oncology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States
[2]Genitourinary Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States
[3]Biostatistics and Data Management Section, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States

Corresponding Author:
Elisabetta Xue, MD
Center for Immuno-Oncology
Center for Cancer Research, National Cancer Institute
National Institutes of Health
9000 Rockville Pike
Building 10, B2L312
Bethesda, MD, 20892
United States
Phone: 1 2403518904
Email: elisabetta.xue@nih.gov

Abstract

Background: Artificial intelligence is increasingly being applied to many workflows. Large language models (LLMs) are publicly accessible platforms trained to understand, interact with, and produce human-readable text; their ability to deliver relevant and reliable information is also of particular interest for the health care providers and the patients. Hematopoietic stem cell transplantation (HSCT) is a complex medical field requiring extensive knowledge, background, and training to practice successfully and can be challenging for the nonspecialist audience to comprehend.

Objective: We aimed to test the applicability of 3 prominent LLMs, namely ChatGPT-3.5 (OpenAI), ChatGPT-4 (OpenAI), and Bard (Google AI), in guiding nonspecialist health care professionals and advising patients seeking information regarding HSCT.

Methods: We submitted 72 open-ended HSCT–related questions of variable difficulty to the LLMs and rated their responses based on consistency—defined as replicability of the response—response veracity, language comprehensibility, specificity to the topic, and the presence of hallucinations. We then rechallenged the 2 best performing chatbots by resubmitting the most difficult questions and prompting to respond as if communicating with either a health care professional or a patient and to provide verifiable sources of information. Responses were then rerated with the additional criterion of language appropriateness, defined as language adaptation for the intended audience.

Results: ChatGPT-4 outperformed both ChatGPT-3.5 and Bard in terms of response consistency (66/72, 92%; 54/72, 75%; and 63/69, 91%, respectively; *P*=.007), response veracity (58/66, 88%; 40/54, 74%; and 16/63, 25%, respectively; *P*<.001), and specificity to the topic (60/66, 91%; 43/54, 80%; and 27/63, 43%, respectively; *P*<.001). Both ChatGPT-4 and ChatGPT-3.5 outperformed Bard in terms of language comprehensibility (64/66, 97%; 53/54, 98%; and 52/63, 83%, respectively; *P*=.002). All displayed episodes of hallucinations. ChatGPT-3.5 and ChatGPT-4 were then rechallenged with a prompt to adapt their language to the audience and to provide source of information, and responses were rated. ChatGPT-3.5 showed better ability to adapt its language to nonmedical audience than ChatGPT-4 (17/21, 81% and 10/22, 46%, respectively; *P*=.03); however, both failed to consistently provide correct and up-to-date information resources, reporting either out-of-date materials, incorrect URLs, or unfocused references, making their output not verifiable by the reader.

XSL•FO
RenderX

Conclusions: In conclusion, despite LLMs' potential capability in confronting challenging medical topics such as HSCT, the presence of mistakes and lack of clear references make them not yet appropriate for routine, unsupervised clinical use, or patient counseling. Implementation of LLMs' ability to access and to reference current and updated websites and research papers, as well as development of LLMs trained in specialized domain knowledge data sets, may offer potential solutions for their future clinical application.

(J Med Internet Res 2024;26:e54758) doi:10.2196/54758

KEYWORDS

hematopoietic stem cell transplant; large language models; chatbot; chatbots; stem cell; large language model; artificial intelligence; AI; medical information; hematopoietic; HSCT; ChatGPT

Introduction

The applications of large language model (LLM)–based chatbots, artificial intelligence tools trained to understand, interact with, and produce human-readable text, are garnering increasing interest in many fields. In medicine, LLMs are successfully passing board examinations [1-3] and show potential in information retrieval and finer conceptual application [4]. LLMs are accessible to health care professionals and patients; therefore, their ability to deliver complex medical information is of particular interest; so far, several applications have been explored, including patients' education [5], patient-trial matching [6], administrative tasks, and training purposes [7].

Hematopoietic stem cell transplantation (HSCT) is a complex medical field requiring extensive knowledge, background, and training to practice successfully and can be challenging for the nonspecialist audience to comprehend. Here, we evaluated the performance of different chatbots in answering HSCT-related questions and assessed their reliability and verifiability, with the aim to determine which LLM can best assist the nonspecialists, including nonhematology medical professionals but also patients and caregivers, in navigating this field.

Methods

We compared the applicability to HSCT of 3 LLM chatbots, ChatGPT-3.5 (OpenAI), ChatGPT-4 (OpenAI), and Bard (Google AI) that were prominent and widely available at the time of the study design (July 2023), by assessing their responses to HSCT-related questions. ChatGPT-3.5 and ChatGPT-4 share a similar architecture, but the former is a free, easily accessible platform, whereas the second is subscription only, was released more recently, and advertised for having better performance; therefore, we included both to examine the difference in their performance. We selected four HSCT-related topics: (1) drugs (mechanisms of action and toxicities), (2) transplant indications and conditioning platforms, (3) infectious, and (4) noninfectious complications. For each topic, we generated 18 open-ended questions, with 3 levels of difficulty ranging from "easy," testing superficial factual knowledge (eg, drug toxicities), to "difficult," testing complex clinical scenarios (eg, inferring causative drug from a toxicity and guiding subsequent patient-tailored management; see Materials and Methods section in Multimedia Appendix 1 for the complete list of submitted questions). All questions were submitted between July 14 and 18, 2023, responses were referenced from the 7th edition European Bone Marrow Transplantation Handbook [8], Lexicomp [9], and BeTheMatch [10] reviewed on July 13, 2023. Each question was submitted 3 times consecutively, without providing feedback to the chatbot: if the chatbot declined to answer any of the 3 submissions, no further evaluation was conducted for that question. If it responded all 3 times, we evaluated the responses for consistency, defined as the ability to convey the same information at each submission. If the 3 responses were consistent, we scored them together for (1) veracity, defined as correctness of the information, (2) language comprehensibility, defined as clarity of the output, and (3) specificity, defined as focus on the question: each variable was rated from "1" (low performance) to "3" (best performance); analyses were then conducted comparing ratings "1 and 2" versus "3." Inconsistent answers were not analyzed further. In addition, answers that scored "2" or "3" in veracity were evaluated for completeness to assess for lack of relevant information. Finally, we assessed for *hallucinations*, defined as nonsensical, fabricated information [11], among incorrect (rated as "1" in veracity) and inconsistent answers (Figure 1). For each step, 2 physicians (EX and DBC) independently graded the answers and reached consensus for any discrepancy; interrater reliability was evaluated through Cohen κ statistic. Average answer word count was also calculated.

Subsequently, after identifying the 2 overall best performing LLMs, we aimed to determine their utility as an information resource for nonspecialist audiences: we rechallenged them by resubmitting the "difficult" questions, prompting the chatbots to respond as if communicating with either a health care professional or a patient and to provide clear reference sources and concise responses. Answers were then regraded, as described above, and additionally assessed for language appropriateness, defined as language adaptation for the intended audience, to evaluate the LLMs' ability to convey the same information using either simple or more technical terminology.

Fisher exact, Wilcoxon rank sum, and Kruskal-Wallis tests were used for categorical and continuous variables, respectively. All the statistical analyses were exploratory and performed using R (version 4.3.1; R Foundation Statistical Computing).

This study was exempt from ethical review since no human subjects were involved and 45 CFR part 46 did not apply.

XSL•FO
RenderX

Figure 1. Rating system. Each question was submitted 3 times and rated according to the rating system reported. AI: artificial intelligence.

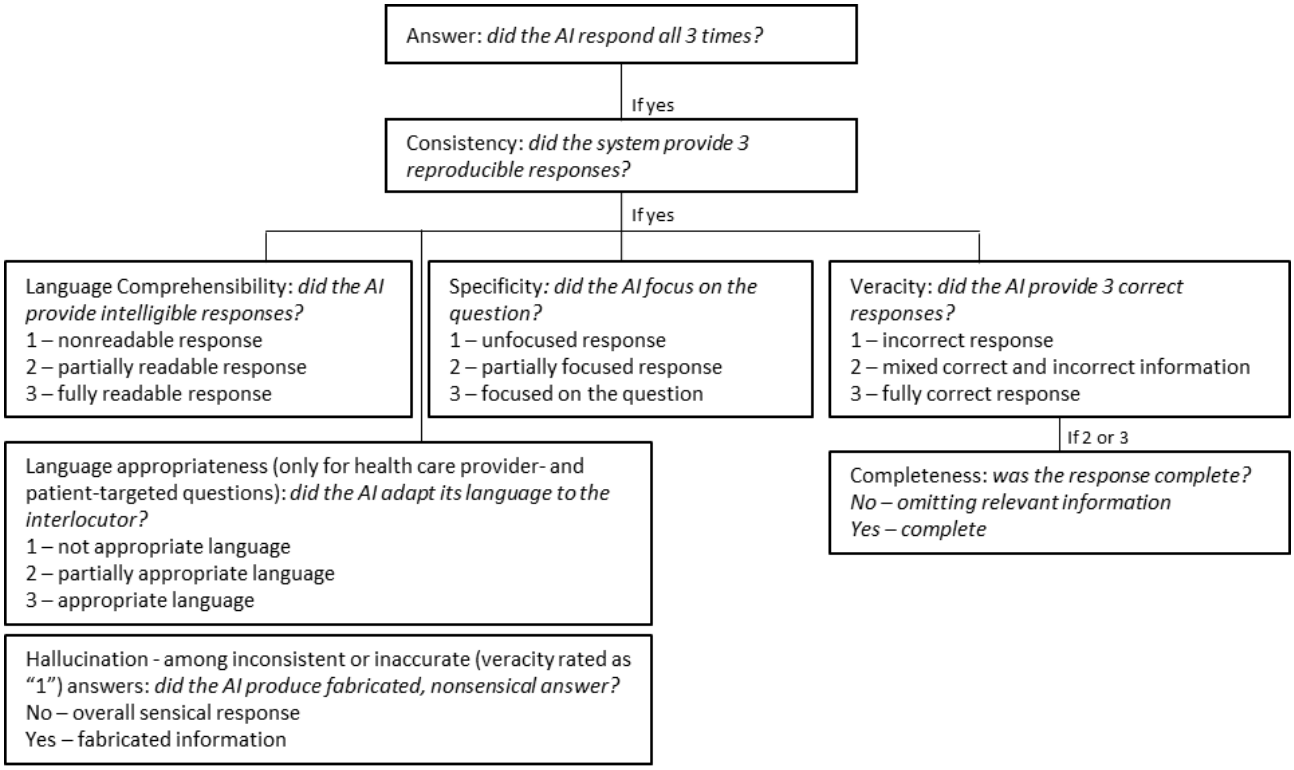

Results

Detailed questions and responses are included in Multimedia Appendix 1. Cohen κ ranged between moderate to near perfect agreement (Table S1 in Multimedia Appendix 1). Figure 2 and Table S2 in Multimedia Appendix 1 display LLMs' performances; ChatGPT-3.5 and ChatGPT-4 responded to all questions, whereas Bard did not answer 3 of 72 (4%; *P*=.12, reporting "I'm not able to help with that, as I'm only a language model") questions. ChatGPT-4 had the highest rate of consistent responses (66/72, 92%) followed by Bard (63/69, 91%) and ChatGPT-3.5 (54/72, 75%; *P*=.007).

Consistent responses were evaluated for veracity, language comprehensibility, and specificity. ChatGPT-4 performed best in terms of veracity, with 58 of 66 (88%) of answers considered correct (rated as "3"), followed by ChatGPT-3.5 (40/54, 74%) and Bard (16/63, 25%; *P*<.001). For instance, when asked why a female patient who received an allogeneic HSCT from a male donor develops leukemia with 46XY karyotype, only ChatGPT-4 and Bard recognized this as donor-derived leukemia, whereas ChatGPT-3.5 wrongly suggested relapse of patient's original disease. Bard had the highest rate of incorrect responses (rated as "1," 21/63, 33%), especially among "moderate" and "difficult" questions. The LLMs also proved ineffective at calculating well-established risk scores (eg, Hematopoietic Cell Transplantation–specific Comorbidity Index). With respect to completeness, ChatGPT-3.5 and ChatGPT-4 answers were deemed complete in >80% of evaluable cases compared to approximately 60% for Bard answers (*P*<.001).

For language comprehensibility, ChatGPT-3.5 and ChatGPT-4 performance was equivalent, with only <4% of the answers rated less than "3" (ie, less than fully comprehensible) compared to 18% (11/63) of Bard answers (*P*=.002), especially among "moderate" and "difficult" questions (data not shown). Specifically, Bard exhibited a trend of repetitive language, yielding a less clear output.

Regarding specificity, ChatGPT-4 responses were rated as "3" (ie, very focused on the topic) in >90% of cases, followed by ChatGPT-3.5 (43/54, 80%) and Bard (27/63, 43%). For instance, when asked which drugs should be administered before antithymocyte globulin, ChatGPT-3.5 and ChatGPT-4 correctly listed the premedication and its purpose, whereas Bard listed all the premedication's side effects. Bard provided more specific answers to the "easy" questions, whereas ChatGPT-3.5 and ChatGPT-4 performed similarly across the difficulty levels (data not shown).

All exhibited episodes of hallucinations, with ChatGPT-3.5, ChatGPT-4, and Bard showing at least 1 hallucinated answer in 7 of 24 (29%), 3 of 7 (43%), and 14 of 27 (52%) of evaluable cases, respectively. ChatGPT-4 provided shorter answers, with a median of 213 (IQR 191-261) words per answer, followed by ChatGPT-3.5 with 247 (IQR 212-307) words, and Bard with 303 (IQR 260-384) words (*P*<.001).

Due to their overall better performance, ChatGPT-4 and ChatGPT-3.5 were selected for the assessment of audience-tailored information delivery. In the "health care professional-targeted" version, with respect to specificity, ChatGPT-3.5 more frequently yielded unfocused (ie, rated as "1" or "2") answers compared to ChatGPT-4 (6/18, 33% and 0/21, 0% respectively); no noticeable differences were seen in other parameters, including language appropriateness (Table S3 in Multimedia Appendix 1). ChatGPT-3.5 did not provide verifiable information sources despite the prompt requiring them to do so, reporting "The information I provide is based on

XSL•FO
RenderX

the knowledge I was trained on until September 2021"; ChatGPT-4 referenced scientific literature in 10 of 21 (48%) evaluable cases, frequently with relevant but out-of-date papers, or with inaccurate authorship, title, or Digital Object Identifier.

In the "patient-targeted" version (Table S4 in Multimedia Appendix 1), ChatGPT-4 yielded a higher rate of correct responses compared to ChatGPT-3.5, with 19 of 22 (86%) and 10 of 21 (48%) cases rated as "3" in veracity, respectively; however, ChatGPT-4 showed excessively technical language, with only 10 of 22 (46%) rated as "3" in language appropriateness compared to 17 of 21 (81%) for ChatGPT-3.5. No differences were seen in the other parameters. Both failed to return information sources but provided website resources targeted for patients in 62% (13/21) and 95% (21/22) of the cases, respectively. Notably, ChatGPT-3.5 returned several broken URLs, likely corresponding to no longer existing pages, while ChatGPT-4 tended to provide overly general websites for very specific queries (eg, American Cancer Society web page [12] for information on sinusoidal obstruction syndrome). Both consistently acknowledged the potential for case-to-case variability and recommended referring to the medical team for any case-specific questions.

Figure 2. ChatGPT-3.5's, ChatGPT-4's, and Bard's performance. Seventy-two questions were submitted to each LLM. (A) Number of questions answered 3 times, (B) number of answers consistent with each other, (C) comprehensibility, (D) specificity, and (E) veracity. Veracity, language comprehensibility, and specificity were evaluated among consistent responses, and were rated as "1" (poor performance), "2" (mediocre performance), or "3" (best performance), and (F) completeness was evaluated among responses with veracity rated as "2" or "3." LLM: large language model.

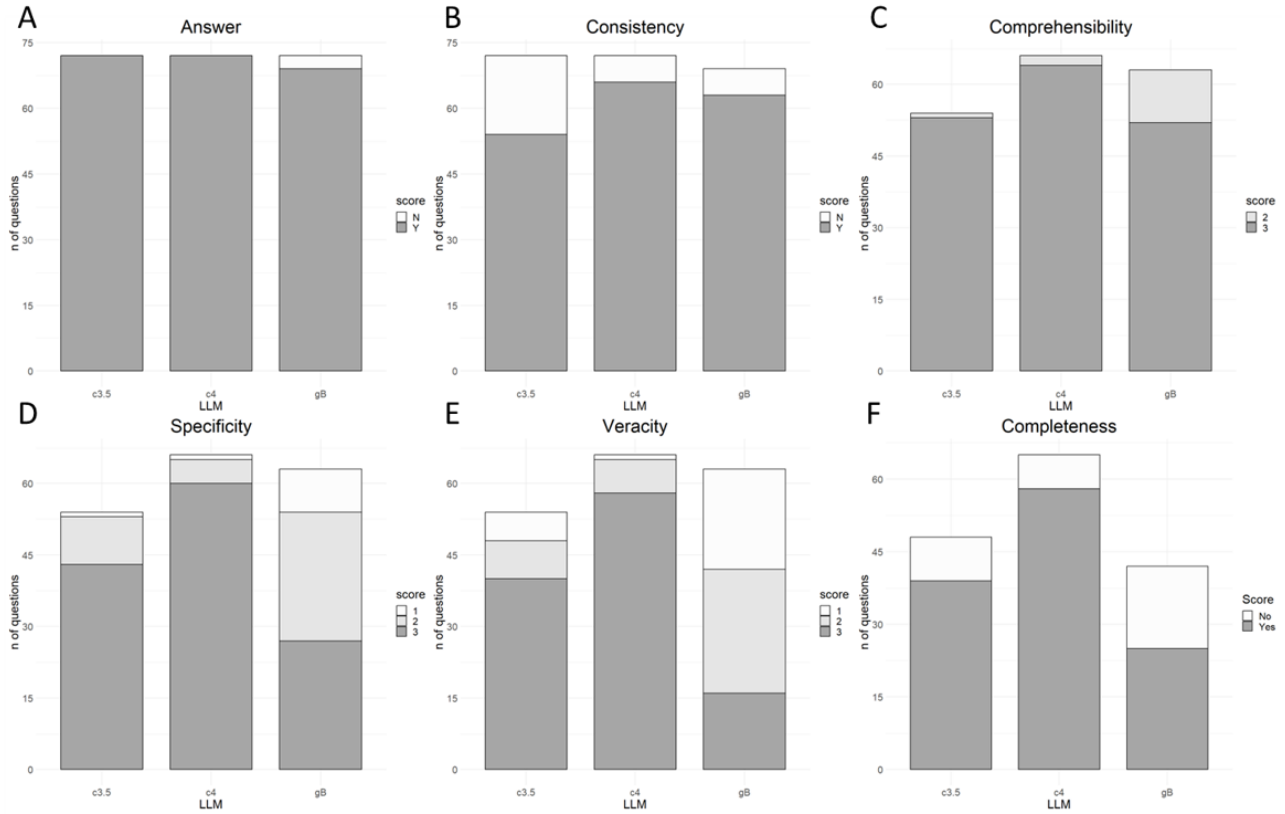

Discussion

The emergence of LLMs has expanded the accessibility of medical information to the general public [5,13]; however, their reliability remains of concern [14]. In our study, all 3 LLMs correctly answered most of the "easy" questions, but only ChatGPT-3.5 and ChatGPT-4 successfully addressed more complex scenarios, and both outperformed Bard in producing comprehensible and specific responses. However, all exhibited episodes of hallucinations; thus, the potential for mistakes in diagnosis and recommendations remains a major obstacle to their routine unsupervised use.

When testing LLMs as support learning tools for laypeople, ChatGPT-3.5 adopted a friendly tone and, interestingly, exhibited some degree of emotional support (eg, "I understand your concern" and "take care"), showing a greater ability in adjusting its language to the audience. Adapting language to the general community and avoiding technical jargon would be optimal tools for making complex information accessible to patients and caregivers. In our opinion, LLMs cannot replace effective patient-doctor communication but rather may potentially supplement it, eventually reducing the risk of misinformation from nonscientific websites and sources. However, in our experience, current LLMs failed to consistently provide specific and updated web-based references, likely due to ChatGPT's then lack of real-time access to the internet, thus making their output frequently unverifiable by the reader.

From a physician's perspective, LLMs cannot replace conferences or scientific literature but may effectively support personal learning. Unfortunately, limited access to current web-based data, errors in reporting peer-reviewed material, and

XSL•FO
RenderX

failure to provide valid references severely compromise this application [15].

Our study has limitations, including the submission of each question 3 consecutive times, without opening a new chat session each time, potentially urging the chatbot to provide a different answer at each submission. Furthermore, we subjectively selected 3 among the most popular available LLMs, 2 of which are developed by the same company; as more are becoming available, our observation might not apply to other LLMs. Finally, this is a rapidly changing field: since the completion of our analysis, ChatGPT has gained access to real time internet data, and Bard was updated into Google Gemini, and thus might yield a different output if tested today.

In conclusion, our evaluation suggests that, given the higher rate of correct and focused responses provided, at the time of this analysis, ChatGPT-3.5 and ChatGPT-4 are not yet appropriate for routine, unsupervised clinical use for both the general population and health care providers, or patient counseling. Their use at present should only be considered under expert supervision or for research purposes. Nevertheless, because of the rapid progression and the clear potential of LLMs to revolutionize workflows in medicine, including specialized fields, we need to engage proactively with this technology. Implementation of LLMs' ability to access and to reference current and updated websites and research papers, as well as the development of LLMs trained in specialized domain knowledge data sets, may offer potential solutions for their future clinical application.

Authors' Contributions

EX, DBC, GMI, and CSF designed the study; EX, GMI, and HCW performed the analysis and prepared the figure; EX, DBC, and CSF wrote the paper; JLG and CSF provided supervision. All authors accepted the final draft of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 450 KB - jmir_v26i1e54758_app1.docx]

References

1. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023 Nov 01;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
2. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
4. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
5. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 2023 Sep;143(9):779-782 [FREE Full text] [doi: 10.1080/00016489.2023.2254809] [Medline: 37694729]
6. Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, et al. Matching patients to clinical trials with large language models. AarXiv Preprint posted online on April 20, 2024. [Medline: 37576126]
7. Tripathi S, Sukumaran R, Cook TS. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. J Am Med Inform Assoc 2024 Jan 25:ocad258. [doi: 10.1093/jamia/ocad258] [Medline: 38273739]
8. The manual of choice for doctors and practitioners involved in hematopoietic cell transplantation (HCT) and cellular therapies. EBMT Handbook. URL: https://www.ebmt.org/education/ebmt-handbook [accessed 2023-07-13]
9. UptoDate® Lexidrug™. URL: https://online.lexi.com/lco/action/home [accessed 2023-07-13]
10. Unlocking the paths to possibility. 2024 NMDP Galas: New York, Los Angeles, Minneapolis. www.BeTheMatch.org. (n.d.). URL: https://bethematch.org/ [accessed 2023-07-13]
11. Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023;15(4):e37432 [FREE Full text] [doi: 10.7759/cureus.37432] [Medline: 37182055]
12. Every cancer. Every life. American Cancer Society. URL: https://www.cancer.org/ [accessed 2023-07-13]

XSL•FO
RenderX

13. Campbell DJ, Estephan LE, Sina EM, Mastrolonardo EV, Alapati R, Amin DR, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid 2024;34(3):371-377. [doi: 10.1089/thy.2023.0491] [Medline: 38010917]
14. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol 2023;13:1268915 [FREE Full text] [doi: 10.3389/fonc.2023.1268915] [Medline: 37731643]
15. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus 2023;15(5):e39238 [FREE Full text] [doi: 10.7759/cureus.39238] [Medline: 37337480]

Abbreviations

HSCT: hematopoietic stem cell transplantation
LLM: large language model

Edited by G Tsafnat; submitted 21.11.23; peer-reviewed by M Chatzimina, E Hermes-DeSantis, T Ma, R Yang, K Fitzner; comments to author 16.03.24; revised version received 22.03.24; accepted 22.03.24; published 17.05.24.

Please cite as:
Xue E, Bracken-Clarke D, Iannantuono GM, Choo-Wosoba H, Gulley JL, Floudas CS
Utility of Large Language Models for Health Care Professionals and Patients in Navigating Hematopoietic Stem Cell Transplantation: Comparison of the Performance of ChatGPT-3.5, ChatGPT-4, and Bard
J Med Internet Res 2024;26:e54758
URL: https://www.jmir.org/2024/1/e54758
doi:10.2196/54758
PMID:38758582

©Elisabetta Xue, Dara Bracken-Clarke, Giovanni Maria Iannantuono, Hyoyoung Choo-Wosoba, James L Gulley, Charalampos S Floudas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial

Wenyi Gan[1*], PhD; Jianfeng Ouyang[2*], PhD; Hua Li[3*], PhD; Zhaowen Xue[1*], PhD; Yiming Zhang[1*], PhD; Qiu Dong[1], PhD; Jiadong Huang[4], MD; Xiaofei Zheng[1], PhD; Yiyi Zhang[1], PhD

[1]The First Clinical Medical College of Jinan University, The First Affiliated Hospital of Jinan University, Guangzhou, China

[2]Department of Joint Surgery and Sports Medicine, Zhuhai People's Hospital (Zhuhai Hospital Affiliated With Jinan University), Zhuhai, Guangdong, China

[3]Department of Orthopaedics, Beijing Jishuitan Hospital, Beijing, China

[4]Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yiyi Zhang, PhD
The First Clinical Medical College of Jinan University, The First Affiliated Hospital of Jinan University
No. 613, Huangpu Avenue West
Tianhe District
Guangzhou, 510630
China
Phone: 86 130 76855735
Fax: 86 020 38688563
Email: yiyizjun@126.com

Abstract

Background: ChatGPT is a natural language processing model developed by OpenAI, which can be iteratively updated and optimized to accommodate the changing and complex requirements of human verbal communication.

Objective: The study aimed to evaluate ChatGPT's accuracy in answering orthopedics-related multiple-choice questions (MCQs) and assess its short-term effects as a learning aid through a randomized controlled trial. In addition, long-term effects on student performance in other subjects were measured using final examination results.

Methods: We first evaluated ChatGPT's accuracy in answering MCQs pertaining to orthopedics across various question formats. Then, 129 undergraduate medical students participated in a randomized controlled study in which the ChatGPT group used ChatGPT as a learning tool, while the control group was prohibited from using artificial intelligence software to support learning. Following a 2-week intervention, the 2 groups' understanding of orthopedics was assessed by an orthopedics test, and variations in the 2 groups' performance in other disciplines were noted through a follow-up at the end of the semester.

Results: ChatGPT-4.0 answered 1051 orthopedics-related MCQs with a 70.60% (742/1051) accuracy rate, including 71.8% (237/330) accuracy for A1 MCQs, 73.7% (330/448) accuracy for A2 MCQs, 70.2% (92/131) accuracy for A3/4 MCQs, and 58.5% (83/142) accuracy for case analysis MCQs. As of April 7, 2023, a total of 129 individuals participated in the experiment. However, 19 individuals withdrew from the experiment at various phases; thus, as of July 1, 2023, a total of 110 individuals accomplished the trial and completed all follow-up work. After we intervened in the learning style of the students in the short term, the ChatGPT group answered more questions correctly than the control group (ChatGPT group: mean 141.20, SD 26.68; control group: mean 130.80, SD 25.56; P=.04) in the orthopedics test, particularly on A1 (ChatGPT group: mean 46.57, SD 8.52; control group: mean 42.18, SD 9.43; P=.01), A2 (ChatGPT group: mean 60.59, SD 10.58; control group: mean 56.66, SD 9.91; P=.047), and A3/4 MCQs (ChatGPT group: mean 19.57, SD 5.48; control group: mean 16.46, SD 4.58; P=.002). At the end of the semester, we found that the ChatGPT group performed better on final examinations in surgery (ChatGPT group: mean 76.54, SD 9.79; control group: mean 72.54, SD 8.11; P=.02) and obstetrics and gynecology (ChatGPT group: mean 75.98, SD 8.94; control group: mean 72.54, SD 8.66; P=.04) than the control group.

Conclusions: ChatGPT answers orthopedics-related MCQs accurately, and students using it excel in both short-term and long-term assessments. Our findings strongly support ChatGPT's integration into medical education, enhancing contemporary instructional methods.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry Chictr2300071774; https://www.chictr.org.cn/hvshowproject.html ?id=225740&v=1.0

(J Med Internet Res 2024;26:e57037) doi:10.2196/57037

KEYWORDS

ChatGPT; medical education; orthopedics; artificial intelligence; large language model; natural language processing; randomized controlled trial; learning aid

Introduction

ChatGPT, a natural language processing model developed by OpenAI, is based on a sophisticated machine learning algorithm that can be iteratively updated and optimized to accommodate the changing and complex requirements of human verbal communication [1-3]. ChatGPT-4.0 is significantly superior to ChatGPT-3.5 in terms of language comprehension, context comprehension, generation speed, and interpretability [4]. It can be applied to a variety of natural language processing duties and provides individuals with more precise, efficient, and intelligent natural language processing services [5,6]. Numerous researchers have reported that ChatGPT can achieve satisfactory results on multidisciplinary medical practitioner examinations in a variety of countries as well as provide detailed evidence-based explanations when responding to input clinical scenarios [7-10]. These studies have investigated the function of ChatGPT in the vast field of medicine and demonstrated the feasibility of investigating the application of ChatGPT in medical education.

Due to the vast medical knowledge system, diverse content, and lengthy learning cycle, it is difficult to accomplish a quality breakthrough in medical education [11,12]. Both general practitioners and specialists must keep relearning medical knowledge to avoid forgetting some minor but essential knowledge points during clinical practice [13,14]. The internet has become a common learning resource for physicians and medical students due to its convenience. However, the internet's general search results are vast and complex, necessitating the use of very specific search terms so that users can find the answers they seek [15]. Designing a specific learning application based on the network can increase the output of knowledge points, but it may not be able to respond promptly to the personalized user query input [16]. By identifying keywords in queries and analyzing their relevance, search engines provide users with a variety of search results; however, users must frequently determine the authenticity and veracity of the answers [17,18]. It has been indicated that while ChatGPT and professional forum answers exhibit comparable levels of logic and accuracy in their specific responses, ChatGPT demonstrates significantly greater empathy than the answers provided by verified individuals on forums, despite the fact that ChatGPT relies on big data [8,19]. Therefore, ChatGPT's response is more akin to a web-based answer retrieval, with the veracity of the retrieved answer being filtered based on the user's personalized query and the expression being intensively processed prior to outputting the result.

Although many academics believe that it is a double-edged sword to use ChatGPT as an auxiliary instrument in medical education [20-22], the development and promotion of ChatGPT will undoubtedly fuel the innovation of medical education [23,24]. The major reason why academics are worried about ChatGPT's inadequacies in medical education is that its potent text creation feature brings the risk of making students too reliant on it as a writing tool [25,26]. However, through interactive learning and immediate feedback, ChatGPT as a virtual teaching assistant can increase learning efficiency [23]. Artificial intelligence (AI)–assisted learning breaks the mode of 1-way input of theoretical knowledge in the previous stage of basic medical knowledge education, and its ability to generate different clinical scenarios according to different diseases can assist medical students in constructing a bridge between professional medical theory and clinical practice [27-29].

Therefore, we conducted a prospective randomized controlled trial concerning the application of ChatGPT in learning orthopedics. We first validated ChatGPT-4.0's accuracy in answering multiple-choice questions (MCQs) related to orthopedics and then used it as a learning aid intervention and conducted short- and long-term follow-up. Through the results of the randomized controlled trial and short-term follow-up, the ultimate objective was to determine whether ChatGPT can be used as an effective learning tool for undergraduate medical students.

Methods

Study Design

A parallel-design randomized controlled trial was used for this investigation. First, the accuracy of ChatGPT-4.0's responses to orthopedics-related MCQs was examined. In addition, for a group experiment, 129 third-year medical students from Jinan University's Medical College were recruited. They were divided into 2 groups at random, namely, the control group and the ChatGPT-4.0–assisted learning group (ChatGPT group). The internet-using students in the control group were not permitted to use any OpenAI-related software or programs, whereas those in the ChatGPT group used only ChatGPT-4.0 as the learning tool. Only after completing the orthopedics course, the orthopedics exercises, the review of the fundamental concepts in orthopedics, using the internet or ChatGPT, the orthopedic examination, and the final examination for the semester's teaching task did the participants finish the experiment. The detailed process of experimental arrangement is shown in Figure 1. Both the CONSORT (Consolidated Standards of Reporting Trials) checklist (Multimedia Appendix 1) and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth; version 1.6.1) checklist (Multimedia Appendix 2) were used for this trial [30,31].

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) 2010 flow diagram showing the randomized controlled trial process. MCQs: multiple-choice questions.

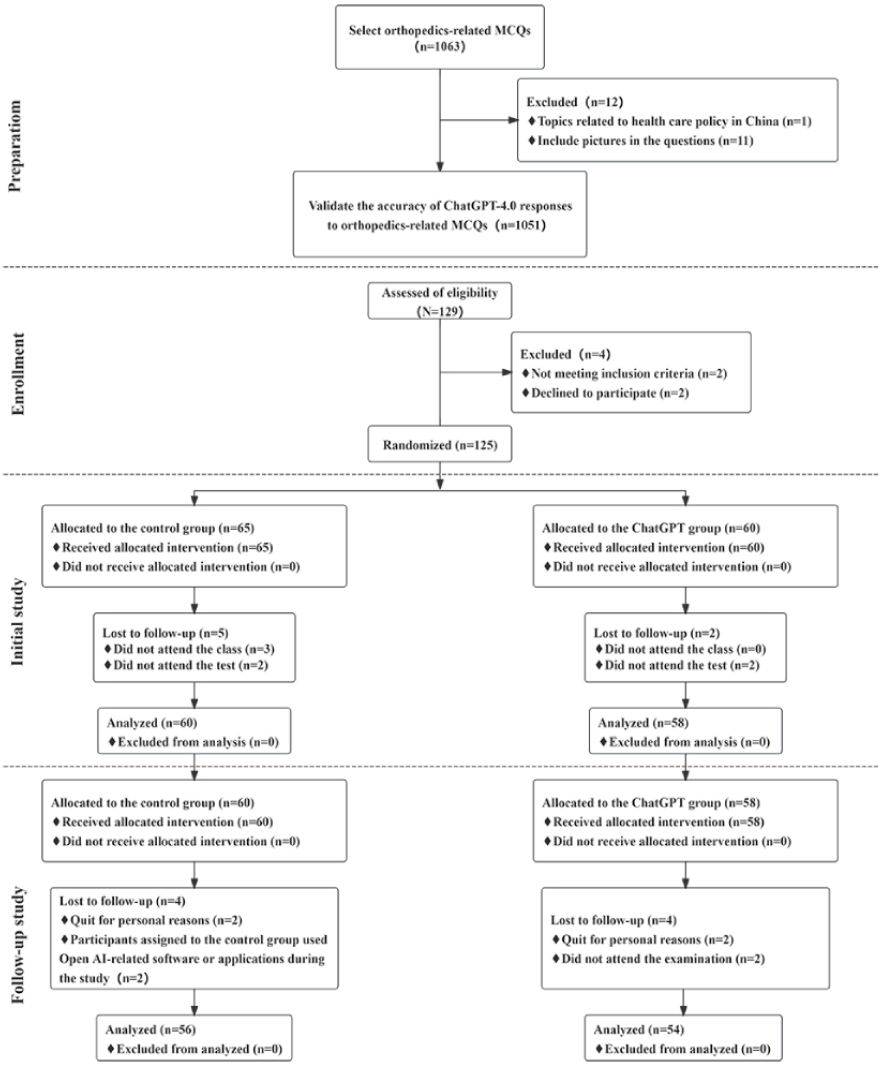

Participants

In China's undergraduate medical education curriculum, third-year students have already completed coursework in foundational subjects such as anatomy, histology, embryology, and physiology, and are presently immersed in acquiring theoretical knowledge pertaining to clinical disciplines. Consequently, to investigate the efficacy of ChatGPT-4.0 in aiding undergraduate medical students' learning of concepts in orthopedics, we recruited third-year undergraduate medical students from Jinan University.

We used a convenience sampling approach to recruit as many participants as possible from the target population. Recruitment was conducted through collective promotional efforts during class meetings in each individual class. This approach allowed us to efficiently disseminate information about the study and encourage widespread participation among the third-year undergraduate medical students.

Inclusion Criteria

Participants were included if they had completed courses in human anatomy, physiology, biochemistry, pathology, pathophysiology, and diagnostics.

Exclusion Criteria

The exclusion criteria were as follows: failing final examinations for 6 courses in human anatomy, physiology, biochemistry, pathology, pathophysiology, and diagnostics that were completed before; switching majors, stopping classes, or dropping out during the current academic year; refusing to join or leaving in the middle for private reasons; failing to finish the multiple-choice exercises or the orthopedics course; using OpenAI-related software or apps in the control group; not completing the orthopedics multiple-choice examination; and missing subjects on the semester's final test.

Preparation

The "National Medical Electronic Schoolbag" (People's Military Medical Press) is the multiple-choice practice software for the standardized training theory examination of China's medical industry (Figure 2B). It is the first digital teaching reform project

XSL•FO
RenderX

in medical higher education funded by China. After screening by 2 orthopedics specialists (WYG and ZWX), we removed 12 questions from the 1075 orthopedics-related MCQs (Multimedia Appendix 3) that were downloaded from the "National Medical Electronic Schoolbag" software (1 question was related to medical policies with Chinese characteristics, and 11 inquiries were disqualified because they contained images). A1, A2, A3/A4, and case analysis questions are the different categories of inquiries. Type A1 questions are primarily about the fundamental knowledge of orthopedics. A2 questions have a brief medical history as the topic stem. Type A3/A4 questions describe a simple patient-centered clinical situation (eg, questions 169 and 170 from the first set of questions in Multimedia Appendix 3), and case analysis questions describe a clinical situation focused on a single patient or family (eg, questions 208-217 from the first set of questions in Multimedia Appendix 3). Chinese was used as the text input language for ChatGPT in this investigation. Finally, 1051 orthopedics-related MCQs were sequentially entered into ChatGPT-4.0 (Figure 2A), and ChatGPT-4.0's responses were recorded.

Figure 2. Scenarios for performing the educational clinical trial. (A) ChatGPT interface for answering orthopedics-related multiple-choice questions; (B) the "National Medical Electronic Schoolbag" is the multiple-choice practice software (English schematic diagram); (C) the participants in the ChatGPT group using ChatGPT to assist in learning orthopedics knowledge; and (D) the participants in 2 groups doing the orthopedics examination.

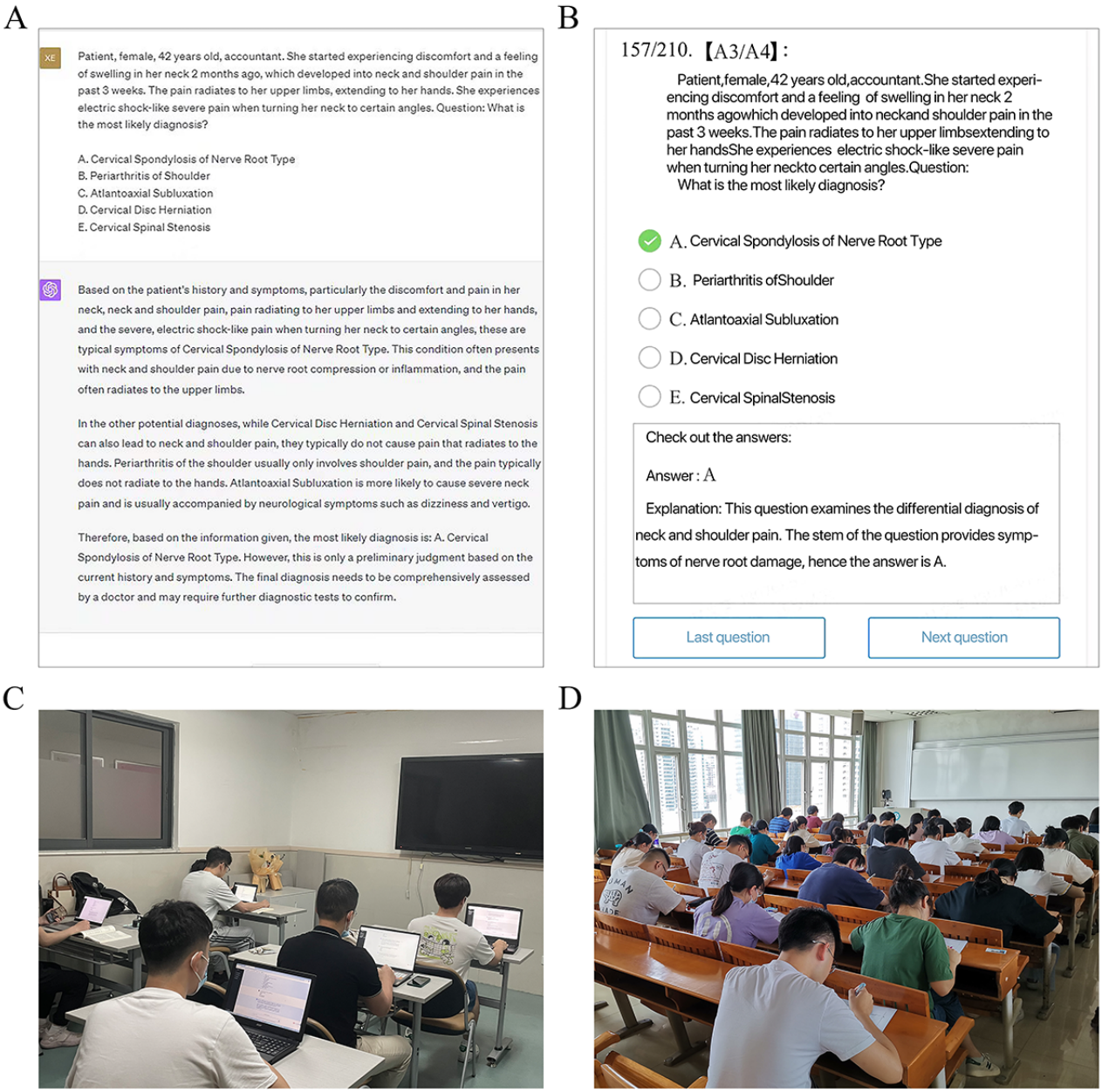

XSL•FO
RenderX

Intervention

Initial Study

Participants were enrolled in a 1-week orthopedics course. Upon completion of the course, they took a multiple-choice test on orthopedics-related topics, and each participant's accuracy rate was recorded. The control group and the ChatGPT group were subsequently formed out of the participants. In the final step, participants in various groups used the internet and ChatGPT (Figure 2C) to evaluate the omissions both in the course and the exercises during the week following the completion of the orthopedic course.

Follow-Up Study

For orthopedics, there was a 7-day review period. The ChatGPT group's students were required to use ChatGPT-4.0 to search up knowledge points to help their learning. Other web-based search engines and message boards were not allowed to be used by the students in this group to search for information. The students in the control group were required to use the internet to support their study by scouring forums and search engines for relevant information. However, they were not allowed to use any software connected to OpenAI. After a week of study using the internet or the ChatGPT, the subjects completed an orthopedics multiple-choice examination. There were 66 questions of type A1, 88 questions of type A2, 29 questions of type A3/A4, and 31 questions requiring case analysis. Participants in diverse groups provided the correct response and total score for each kind of question, which were recorded.

Following the orthopedics examination, the final examination review period was scheduled in accordance with the semester learning plan. We did not obstruct the participants' learning to complete their reviews at this time. The results of this semester's internal medicine, surgery, pediatrics, obstetrics and gynecology, and infectious diseases final examinations (Figure 2D) were systematically gathered via the academic affairs office with the participants' informed consent. Multimedia Appendix 4 contrasts all interventions between the experimental group and the control group.

Outcomes

In the preparation phase, 2 orthopedics specialists (WYG and ZWX) curated and inputted 1075 orthopedics-related MCQs into ChatGPT-4.0, with the aim of evaluating the accuracy of ChatGPT-4.0 in answering these orthopedics MCQs, while also recording its performance across different question types. We had logged answers from "National Medical Electronic Schoolbag" MCQs and ChatGPT-4.0 responses, uploading the data as Multimedia Appendix 5.

Next, after students completed a 1-week prestudy course in orthopedics and used different methods to review the knowledge for 1 week, we assigned ChatGPT-4.0 as an auxiliary learning review tool to students in the ChatGPT-4.0 group, while students in the control group were not allowed to use large language models (LLMs) for auxiliary review. After the review, we tested all students with the same test questions, enabling us to gauge the quality of short-term revision using different learning auxiliary tools. We recorded the performance of the different student groups in the orthopedics MCQ test, which served as the basis for the short-term impact analysis of the learning intervention.

Subsequently, at the end of the current semester, we accessed and recorded the clinical subject examination scores of the different student groups through the academic system. During the period from the end of our intervention to the end of the final examination, our follow-up content consisted of recording the use of LLMs by the ChatGPT group and the control group. We did not carry out any additional interventions during this time. The scores in the subject examinations at the end of the semester served as the basis for long-term impact analysis of the changes in students' learning methods after using ChatGPT-4.0. In this study, the accuracy of ChatGPT-4.0, short-term impact analysis, and long-term impact analysis are all primary outcomes.

Blinding

To eradicate subjective bias in the grading process, the collector was unaware of the classification of the participants when collecting the results from the orthopedics-related multiple-choice exercises and examinations. Other information relied on data from the school's pedagogical administration system, and the personnel who collected the data did not know the group information of the experiment participants.

Randomization

After completing the orthopedics-related multiple-choice exercises, the participants were randomly assigned to different groups using the sealed envelope method to minimize systemic bias. To ensure balanced group sizes, a clinician not involved in the program prepared sealed envelopes containing group assignment information. The participants then selected envelopes to determine their group allocation without knowing the group information beforehand. This approach served as a blocking method to achieve balanced group sizes while reducing artificial bias.

Statistics

For statistical analysis, SPSS software (version 26.0; IBM Corp) was used. The chi-square test was used to analyze gender differences between different groups. The Kolmogorov-Smirnov test was used to determine whether the data exhibited a normal distribution. If the data did not conform to a normal distribution, the Mann-Whitney U test was used for analysis. In addition, the data were processed according to the Levene test, and it was determined that the variance was homogeneous; accordingly, an independent-samples 2-tailed t test was conducted. When the P value was <.05, the difference was considered statistically significant. GraphPad Prism 8 was used to create bar charts. The results of continuous variables were displayed as follows: mean difference between the experimental group and the control group (mean, SD of the difference, P value), whereas the accuracy rate of ChatGPT was displayed as: correct number/total number.

Ethical Considerations

This study was approved by the First Affiliated Hospital of Jinan University's Ethics Committee (KY-2023-171), and it

XSL•FO
RenderX

was also registered in the Chinese Clinical Trials Registry (Chictr2300071774). To ensure confidentiality of the participants, access to the original experimental data necessitates a valid request that is sent to the email address of the corresponding author. Prior to manuscript submission, the research team was required to submit the content of the uploaded materials to the Science and Technology Department of the First Affiliated Hospital of Jinan University for review. This step ensured that no personal information of the participants was disclosed inadvertently. Concurrently, the department was also able to verify the implementation of the participants' rewards. In accordance with the Declaration of Helsinki, the participants' written informed consent was obtained before any information about them was acquired.

Results

Overview

We began collectively recruiting the Jinan University third-year undergraduate class of medical students on April 1, 2023, and finished the recruitment process on April 7, 2023. We finished the "initial study" and short-term follow-up through the course of the next 2 weeks. The long-term follow-up work was then conducted by gathering the final examination results of the participating students after the semester ended in June 2023. From April 1, 2023, to April 7, 2023, a total of 129 eligible candidates from the Medical College of Jinan University were assessed for participation. During the recruitment phase, 4 individuals dropped out, 7 more during the initial study, and 8 during the follow-up, resulting in a total of 110 participants who completed the study, with 56 in the control group and 54 in the ChatGPT group (Figure 1). As part of our follow-up study, we carried out telephone interviews to assess the extent of ChatGPT usage among the participants in both the control and experimental groups. The interviews revealed that, prior to the final examination, only 2 individuals from the control group had engaged with LLMs. In contrast, every participant in the experimental group had used different types of LLMs to varying degrees.

All of the participants mentioned their ages and grade point averages as of the school year to the researchers after completing an informed consent form for this study. We analyzed age (mean –0.02, SD 0.14 years; *P*=.89), sex (*P*=.44), grade point average (mean 0.10, SD 0.11; *P*=.38), and orthopedic practice accuracy rate (mean 0.12, SD 1.34; *P*=.93) across the 2 groups and found no significant differences after omitting those who were lost to follow-up (Table 1).

Table 1. Baseline characteristics of the final participants.

Characteristics	Control group (n=56)	ChatGPT group (n=54)	Total (n=110)	*P* value
Age (years), mean (SD), range	22.46 (0.79), 21-24	22.44 (0.69), 21-24	22.45 (0.74), 21-24	.89
Male sex, n (%)	28 (50)	23 (43)	51:59	.44
Grade point average, mean (SD), range	3.23 (0.51), 2.5-4.3	3.23 (0.68), 2.4-4.4	3.28 (0.60), 2.4-4.4	.38
Orthopedic MCQs[a] practice accuracy rate, mean (SD), range	56.08 (7.34), 42.94-70.91	56.19 (6.74), 41.44-73.92	56.13 (7.02), 41.44-73.92	.93

[a]MCQ: multiple-choice question.

The Accuracy of ChatGPT-4.0 Responses to Orthopedics-Related MCQs

A total of 330 A1 MCQs, 448 A2 MCQs, 131 A3/A4 MCQs, and 142 case analysis MCQs made up to a total of 1051 questions (Figure 3). ChatGPT-4.0 answered the 1051 orthopedics-related MCQs with a satisfactory accuracy rate of 70.60% (742/1051; Figure 3E and 3F). Specifically, the accuracy was 71.8% (237/330) for all A1 MCQs (Figure 3A), 73.7% (330/448) for all A2 MCQs (Figure 3B), and 70.2% (92/131) for all A3/A4 MCQs (Figure 3C), whereas it was only 58.5% (83/142) for all case analysis MCQs (Figure 3D). Among the case analysis MCQs, the accuracy of questions with multiple correct answers was as low as 33% (13/39) (Figure 3D). We found that ChatGPT explained each answer option and enumerated the associated knowledge points (Figure 2A).

XSL•FO
RenderX

Figure 3. ChatGPT's performance in answering multiple-choice questions related to orthopedics. (A-D) Accuracy of various types of questions answered by ChatGPT; (E) overall accuracy of questions answered by ChatGPT; and (F) the percentage of each question type in the overall questions. MCQs: multiple-choice questions.

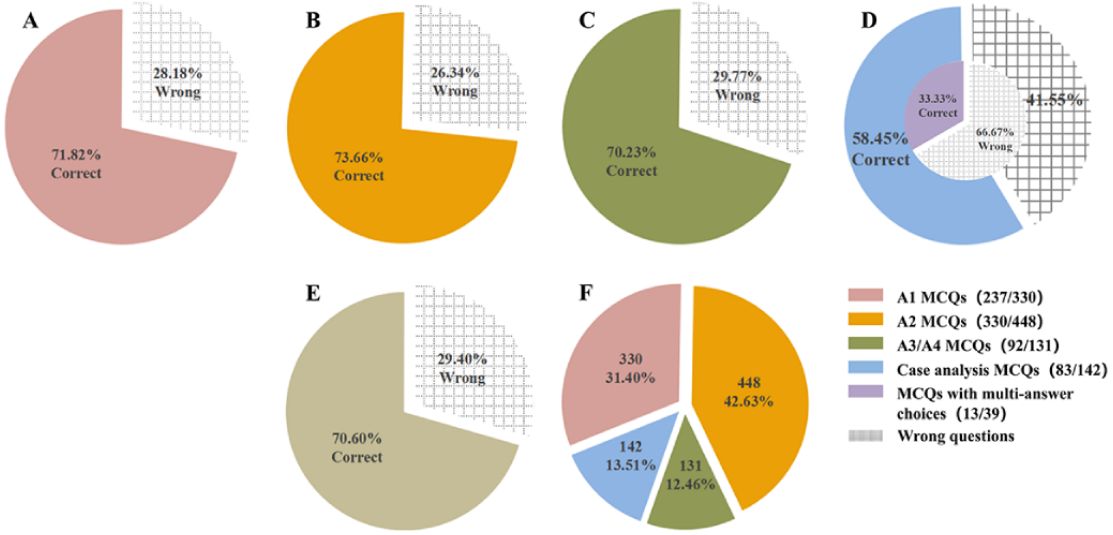

The Accuracy of the 2 Groups of Participants in the Orthopedics Examination

In the orthopedics examination, we were surprised to find out that the participants in the ChatGPT group were typically able to respond to 138.46 (SD 26.97) questions correctly, whereas those in the control group were able to respond to only 130.80 (SD 25.56) questions correctly on average, a significant decline in accuracy (mean 10.40, SD 4.98; *P*=.04; Figure 4B). As shown in Figure 4A, correct answers to A1 (mean 4.40, SD 1.75; *P*=.01), A2 (mean 3.93, SD 1.95; *P*=.047), and A3/4 (mean 3.11, SD 0.96; *P*=.002) among them demonstrated that the ChatGPT group performed significantly better than the control group. Despite the fact that the ChatGPT group showed a lower accuracy in case analysis questions than the control group (mean –1.04, SD 0.72; *P*=.16), the difference was not statistically significant (Figure 4A).

Figure 4. Short-term orthopedics test results and long-term final subjects examination results. (A) Accuracy of various types of orthopedics examination questions; (B) overall accuracy of the orthopedics examination; (C) the 2 groups' final test results for the semester; and (D) the 2 groups' final test total results for the semester. ChatGPT group: ChatGPT-4.0–assisted learning group; MCQs: multiple-choice questions; ns: not significant. **P* value less than .05; ***P* value less than .01.

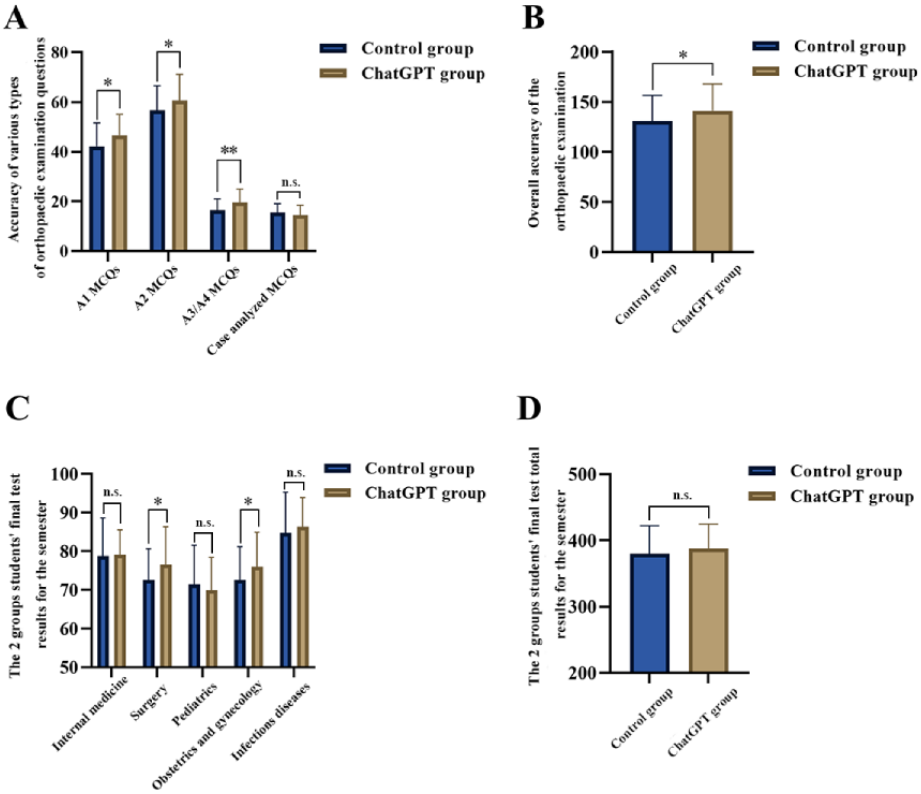

XSL•FO
RenderX

The Score of the Final Examination for the Semester's Teaching Task

The statistical findings and the participants' scores of the final examination for the semester are shown in Figures 4C and 4D. We were pleasantly surprised to find that the ChatGPT-assisted participants scored higher on the final examinations in surgery (mean 4.00, SD 1.72; *P*=.02), internal medicine (mean 0.23, SD 1.58; *P*=.88), obstetrics and gynecology (mean 3.45, SD 1.68; *P*=.04), and infectious diseases (mean 1.60, SD 1.75; *P*=.36) than the control group, although the difference was statistically significant only for surgery and obstetrics and gynecology (Figure 4C). In the pediatric examination, the ChatGPT group scored worse than the control group (mean –1.58, SD 1.79; *P*=.38), but the difference was not statistically significant (Figure 4C). The mean total final examination scores of the ChatGPT group were also higher than those of the control group (mean 7.71, SD 7.53; *P*=.31), but the difference was not statistically significant (Figure 4D).

Discussion

Principal Findings

This is the first prospective clinical trial using ChatGPT as an intervention in the instruction of medical undergraduates. We discovered that ChatGPT has a high rate of accuracy when responding to orthopedics-pertinent MCQs, and that it explains each answer option and lists the corresponding knowledge points. After successfully enrolling the final 110 participants in a clinical study, the students using ChatGPT-4.0 were able to have a better comprehension of orthopedics-related knowledge points and performed better on end-of-semester examinations in other disciplines. These findings provide a firm foundation for medical students to use ChatGPT as an auxiliary learning aid in order to improve learning efficacy and promote the use of ChatGPT in medical education.

The purpose of undergraduate medical education is to assist students in developing a strong theoretical foundation in medicine and fundamental clinical abilities. The MCQs test has grown to be one of the most popular tools for objectively evaluating medical theoretical knowledge [32,33]. The aim of our research was to determine whether undergraduate medical students who use ChatGPT as a learning tool in orthopedics can learn more effectively. It is the first prospective randomized controlled trial that used ChatGPT as the main intervention. Through this research, we discovered that using ChatGPT as a learning tool might increase the effectiveness of acquiring information relevant to orthopedics and lead to improved performance on the MCQs test. The short-term follow-up revealed that the students who used ChatGPT as an additional tool for learning also performed better on assessments for the majority of the subjects at the end of the semester, indicating that ChatGPT users might show willing to alter their learning strategies and include ChatGPT as one of their daily learning tools.

We fed ChatGPT-4.0 with the orthopedics-pertinent practice questions and found that it had a respectable accurate answer rate for all of them (A1 [237/330, 71.8%], A2 [330/448, 73.7%], A3/A4 [92/131, 70.2%], and case analysis [83/142, 58.5%]). While doing so, we saw that ChatGPT provided not just knowledge points but also explanations for each answer choice. ChatGPT was also performed at a level of 54.96% accuracy on the Plastic Surgery Inservice Training Examination, which consisted of 242 questions [7]. Because correct rates of 58.5% and 54.96% are not acceptable in clinical work, we believe that ChatGPT is better suited as an auxiliary learning tool for the basic teaching of medical knowledge than for the auxiliary clinical diagnosis and treatment, relying on the powerful and rapid ability to collect and collate data and the ability to answer questions instantly. In addition to textual information such as the patient's complaint, current history, and prior history, the clinician's physical examination and imaging must be incorporated into the clinical diagnosis and treatment [34,35]. Although some academics enter the results of physical examinations and imaging examinations into ChatGPT as text for big data analysis [7], the quality of such input content is inconsistent, and systematic errors cannot be ruled out.

The vast and intricate medical knowledge system is the primary source of medical education's complexity. In addition to cramming for examinations and studying for examinations, medical students must devote significant effort to accumulating and organizing information and making causal links between different domains of knowledge [36,37]. Medical students often reread the medical material many times to ensure a thorough grasp of the material. Moreover, it is also very challenging to realize the transformation of medical knowledge from theory to practice [38,39] because the process of clinical practice requires the interaction between doctors and patients, that is, the immediate output and immediate feedback between them. With the evolution of teaching concepts and learning aids, medical education model has undergone significant transformations over time [40,41]. The traditional education model consists of a 1-way transmission from professors to students [42]. Mass Open Online Course, a nonfixed multidirection input model, has emerged progressively with the advent of the internet [43]. To enhance students' capacity for self-exploration, education sector constructs problem-based learning and promotes interaction during the learning process [44,45]. Simultaneously, virtual reality–assisted instruction is being developed to enhance medical students' perceptual comprehension of medical knowledge through the interaction and immediate feedback of programmed audiovisual operations [46]. Through big data analysis and data aggregation, ChatGPT interacts with users via text and provides immediate feedback [5,6]. As LLMs technology continues to evolve and improve, its application prospects in medical education will become even more extensive. Educators should actively embrace this technological innovation [29]. These advanced AI tools can provide personalized learning support, helping students better understand complex medical concepts such as anatomy [47], physiology [48], and biochemistry [48]. Through interactive dialogues with LLMs, students can ask questions, obtain detailed explanations and examples, and deepen their understanding of abstract concepts [49,50]. Moreover, LLMs can generate engaging learning materials, such as clinical case studies, question-and-answer exercises, and knowledge summaries, to enhance student participation and learning motivation [49,51].

XSL•FO
RenderX

This immersive learning experience can stimulate students' curiosity and encourage them to actively explore knowledge in the biomedical field [51]. Hence, as a supplementary learning aid for medical students, ChatGPT might be able to realize nonprocedural multidirectional interaction and continuously enhance students' knowledge systems [29] via immediate feedback.

As an auxiliary instrument for medical education, ChatGPT is currently viewed by many academics as a 2-way street [52]. Notably, ChatGPT as a supplementary aid can assist users in collecting the correct target information more efficiently, thereby enhancing work and learning efficiency [5]. However, many academics believe that ChatGPT is merely a plagiarist that lacks initiative and can only collect information, which has become a slang term for slothful people [53-55]. GPTZero (accessed on June 6, 2023), software created by Edward Tian, a student at Princeton University, for statistical analysis of whether text has been generated by AI, has gone a long way toward alleviating the public's anti-AI sentiment. Instead of emphasizing excessively the "writing shortcuts" provided by ChatGPT's text-generation function, users should take advantage of its big data rapid retrieval summary and immediate feedback to improve their study and work efficiency. As society, science, and technology advance, it is inevitable that students will take the initiative to adapt and combine their own learning mode with the updates and iteration of learning media and supplementary learning tools [56]. This general trend cannot be reversed. Therefore, society should hold higher standards for teaching methods, teaching content, and student assessment in this irreversible trend.

Limitations

First, although MCQs are often used to gauge medical students' fundamental theoretical knowledge in many nations, China may have a distinct custom for framing questions and placing a different priority on knowledge points than other nations. Second, this study focuses more on orthopedic intervention than on multidisciplinary intervention, which might result in some limitations in the research findings. Third, during the preparation stage of this study, while assessing ChatGPT-4.0's proficiency in answering MCQs pertaining to orthopedics, we made the decision to omit MCQs that included visual elements. This choice might have, to some degree, constrained the breadth of our investigation and consequently fell short of delivering an all-encompassing appraisal of ChatGPT-4.0's aptitude in tackling a wide spectrum of medical inquiries. This study was a single-center randomized controlled trial; thus, more prospective multicenter and interdisciplinary investigations are required to fully examine ChatGPT's potential as a teaching tool for medical education.

Conclusions

ChatGPT has a high rate of accuracy in answering orthopedics-related MCQs, and it explains each answer choice and lists the corresponding knowledge points. Compared with the students assigned to the control group, those in the ChatGPT group had a better understanding of the knowledge points related to orthopedics after a short intervention period and performed better on the end-of-semester examinations in some other disciplines. These findings provided a solid foundation for medical students to use ChatGPT as an auxiliary learning aid to enhance learning efficacy and help promote the use of ChatGPT in medical education.

Acknowledgments

We would like to thank all the Jinan University medical students who took part in this research and short-term follow-up. The Academic Affairs Office of Jinan University, the Basic Medical College of Jinan University, and the First Clinical Medical College of Jinan University are thanked for their assistance with the planning and compilation of data for this study. We are grateful to Dr Guorong She, who assisted us in completing the randomization process for the experiment. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Disclaimer

All authors affirm that neither ChatGPT nor any AI-assisted software was used in the text creation of this article. The screenshot example used to illustrate learning with ChatGPT unequivocally demonstrates that the interface originates from ChatGPT. ChatGPT was used exclusively as an experimental intervention and not as a paper-writing instrument in this study.

Authors' Contributions

WG led the conceptualization, investigation, visualization, and writing (original draft, review, and editing). ZX initiated the data curation, formal analysis, and writing (original draft). Yiming Z and HL initiated the data curation and formal analysis. QD, and JH led the assistance with finishing the study and providing orthopedics instructions. JO, XZ, and Yiyi Z led the conceptualization, writing (review and editing), and study supervision. All authors concur that neither ChatGPT nor any other OpenAI-related software or application was used during the writing of the paper's text. All authors read and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) 2010 Checklist.
[DOC File , 197 KB - jmir_v26i1e57037_app1.doc]

Multimedia Appendix 2
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth; version 1.6.1) checklist.
[PDF File (Adobe PDF File), 2914 KB - jmir_v26i1e57037_app2.pdf]

Multimedia Appendix 3
Multiple-choice questions.
[DOCX File , 176 KB - jmir_v26i1e57037_app3.docx]

Multimedia Appendix 4
Illustration showing different parts of the training were undertaken by each group.
[DOCX File , 13 KB - jmir_v26i1e57037_app4.docx]

Multimedia Appendix 5
The reference answers and the responses from ChatGPT-4.0.
[XLSX File (Microsoft Excel File), 37 KB - jmir_v26i1e57037_app5.xlsx]

References

1. Baumgarten KM, Gerlach D, Galatz LM, Teefey SA, Middleton WD, Ditsios K, et al. Cigarette smoking increases the risk for rotator cuff tears. Clin Orthop Relat Res 2010;468(6):1534-1541 [FREE Full text] [doi: 10.1007/s11999-009-0781-2] [Medline: 19283436]
2. Agathokleous E, Saitanis CJ, Fang C, Yu Z. Use of ChatGPT: what does it mean for biology and environmental science? Sci Total Environ 2023;888:164154. [doi: 10.1016/j.scitotenv.2023.164154] [Medline: 37201835]
3. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
4. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus 2023;15(4):e37023 [FREE Full text] [doi: 10.7759/cureus.37023] [Medline: 37143631]
5. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
6. Cahan P, Treutlein B. A conversation with ChatGPT on the role of computational systems biology in stem cell research. Stem Cell Reports 2023;18(1):1-2 [FREE Full text] [doi: 10.1016/j.stemcr.2022.12.009] [Medline: 36630899]
7. Gupta R, Herzog I, Park JB, Weisberger J, Firouzbakht P, Ocon V, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023;43(12):NP1078-NP1082. [doi: 10.1093/asj/sjad128] [Medline: 37128784]
8. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024;52(6):1542-1545 [FREE Full text] [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
10. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the situational judgement Test-A professional dilemmas-based examination for doctors in the United Kingdom. JMIR Med Educ 2023;9:e48978 [FREE Full text] [doi: 10.2196/48978] [Medline: 37548997]
11. Malau-Aduli BS, Lee AY, Cooling N, Catchpole M, Jose M, Turner R. Retention of knowledge and perceived relevance of basic sciences in an integrated case-based learning (CBL) curriculum. BMC Med Educ 2013;13:139 [FREE Full text] [doi: 10.1186/1472-6920-13-139] [Medline: 24099045]
12. Atchley TJ, Vukic B, Vukic M, Walters BC. Review of cerebrospinal fluid physiology and dynamics: a call for medical education reform. Neurosurgery 2022;91(1):1-7. [doi: 10.1227/neu.0000000000002000] [Medline: 35522666]

XSL•FO
RenderX

13. Hilty DM, Turvey C, Hwang T. Lifelong learning for clinical practice: how to leverage technology for telebehavioral health care and digital continuing medical education. Curr Psychiatry Rep 2018;20(3):15. [doi: 10.1007/s11920-018-0878-y] [Medline: 29527637]
14. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79(10 Suppl):S70-S81. [doi: 10.1097/00001888-200410001-00022] [Medline: 15383395]
15. Ullrich PF, Vaccaro AR. Patient education on the internet: opportunities and pitfalls. Spine (Phila Pa 1976) 2002;27(7):E185-E188. [doi: 10.1097/00007632-200204010-00019] [Medline: 11923675]
16. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. JMIR Form Res 2024;8:e51346 [FREE Full text] [doi: 10.2196/51346] [Medline: 38717811]
17. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020;22(8):e16504 [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
18. Rouleau G, Gagnon M, Côté J, Payne-Gagnon J, Hudson E, Dubois C, et al. Effects of e-learning in a continuing education context on nursing care: systematic review of systematic qualitative, quantitative, and mixed-studies reviews. J Med Internet Res 2019;21(10):e15118 [FREE Full text] [doi: 10.2196/15118] [Medline: 31579016]
19. Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X. Quality and dependability of ChatGPT and DingXiangYuan forums for remote orthopedic consultations: comparative analysis. J Med Internet Res 2024;26:e50882 [FREE Full text] [doi: 10.2196/50882] [Medline: 38483451]
20. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
21. Gordijn B, Have HT. ChatGPT: evolution or revolution? Med Health Care Philos 2023;26(1):1-2. [doi: 10.1007/s11019-023-10136-0] [Medline: 36656495]
22. NA. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613(7945):612. [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]
23. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
24. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? J Med Syst 2023;47(1):61. [doi: 10.1007/s10916-023-01957-w] [Medline: 37160568]
25. The Lancet Digital health. ChatGPT: friend or foe? Lancet Digit Health 2023;5(3):e102 [FREE Full text] [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
26. Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. Sci Rep 2023;13(1):4569 [FREE Full text] [doi: 10.1038/s41598-023-31341-0] [Medline: 37024502]
27. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
28. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023;98(8):867-868. [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
29. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
30. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, CONSORT. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg 2012;10(1):28-55 [FREE Full text] [doi: 10.1016/j.ijsu.2011.10.001] [Medline: 22036893]
31. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
32. DeSantis M, McKean TA. Efficient validation of teaching and learning using multiple-choice exams. Adv Physiol Educ 2003;27(1-4):3-14 [FREE Full text] [doi: 10.1152/advan.00016.2001] [Medline: 12594068]
33. Gupta V, Williams ER, Wadhwa R. Multiple-choice tests: A-Z in best writing practices. Psychiatr Clin North Am 2021;44(2):249-261. [doi: 10.1016/j.psc.2021.03.008] [Medline: 34049647]
34. Thorborg K, Reiman MP, Weir A, Kemp JL, Serner A, Mosler AB, et al. Clinical examination, diagnostic imaging, and testing of athletes with groin pain: an evidence-based approach to effective management. J Orthop Sports Phys Ther 2018;48(4):239-249. [doi: 10.2519/jospt.2018.7850] [Medline: 29510653]
35. Chou R, Qaseem A, Snow V, Casey D, Cross JT, Shekelle P, Clinical Efficacy Assessment Subcommittee of the American College of Physicians, American College of Physicians, American Pain Society Low Back Pain Guidelines Panel. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med 2007;147(7):478-491 [FREE Full text] [doi: 10.7326/0003-4819-147-7-200710020-00006] [Medline: 17909209]
36. Maniar KP, Arva N, Blanco LZ, Mao Q, Morency EG, Rodriguez R, et al. Accreditation council for graduate medical education self-study for pathology: one institution's experience and lessons learned. Arch Pathol Lab Med 2019;143(10):1271-1277 [FREE Full text] [doi: 10.5858/arpa.2018-0467-RA] [Medline: 31017451]

XSL•FO
RenderX

37. McAdams CD, McNally MM. Continuing medical education and lifelong learning. Surg Clin North Am 2021;101(4):703-715. [doi: 10.1016/j.suc.2021.05.015] [Medline: 34242611]
38. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2009;14(4):595-621. [doi: 10.1007/s10459-007-9090-2] [Medline: 18034364]
39. Slotnick HB. How doctors learn: physicians' self-directed learning episodes. Acad Med 1999;74(10):1106-1117. [doi: 10.1097/00001888-199910000-00014] [Medline: 10536633]
40. Wong G, Greenhalgh T, Pawson R. Internet-based medical education: a realist review of what works, for whom and in what circumstances. BMC Med Educ 2010;10:12 [FREE Full text] [doi: 10.1186/1472-6920-10-12] [Medline: 20122253]
41. Venkatesan M, Mohan H, Ryan JR, Schürch CM, Nolan GP, Frakes DH, et al. Virtual and augmented reality for biomedical applications. Cell Rep Med 2021;2(7):100348 [FREE Full text] [doi: 10.1016/j.xcrm.2021.100348] [Medline: 34337564]
42. Sattar MU, Palaniappan S, Lokman A, Hassan A, Shah N, Riaz Z. Effects of Virtual Reality training on medical students' learning motivation and competency. Pak J Med Sci 2019;35(3):852-857 [FREE Full text] [doi: 10.12669/pjms.35.3.44] [Medline: 31258607]
43. Milbourn B, Black MH, Afsharnejad B, Snyman Z, Baker-Young E, Thompson C, et al. The "Talk-to-Me" MOOC intervention for suicide prevention and mental health education among tertiary students: protocol of a multi-site cross-over randomised controlled trial. Contemp Clin Trials 2022;112:106645. [doi: 10.1016/j.cct.2021.106645] [Medline: 34861409]
44. Al-Azri H, Ratnapalan S. Problem-based learning in continuing medical education: review of randomized controlled trials. Can Fam Physician 2014;60(2):157-165 [FREE Full text] [Medline: 24522680]
45. Jin J, Bridges SM. Educational technologies in problem-based learning in health sciences education: a systematic review. J Med Internet Res 2014;16(12):e251 [FREE Full text] [doi: 10.2196/jmir.3240] [Medline: 25498126]
46. Gan W, Mok T, Chen J, She G, Zha Z, Wang H, et al. Researching the application of virtual reality in medical education: one-year follow-up of a randomized trial. BMC Med Educ 2023;23(1):3 [FREE Full text] [doi: 10.1186/s12909-022-03992-6] [Medline: 36597093]
47. Temsah M, Alhuzaimi AN, Almansour M, Aljamaan F, Alhasan K, Batarfi MA, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst 2024;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
48. Luke WANV, Seow Chong L, Ban KH, Wong AH, Zhi Xiong C, Shuh Shing L, et al. Is ChatGPT 'ready' to be a learning tool for medical undergraduates and will it perform equally in different subjects? Comparative study of ChatGPT performance in tutorial and case-based learning questions in physiology and biochemistry. Med Teach 2024:1-7. [doi: 10.1080/0142159X.2024.2308779] [Medline: 38295769]
49. Sauder M, Tritsch T, Rajput V, Schwartz G, Shoja MM. Exploring generative artificial intelligence–assisted medical education: assessing case-based learning for medical students. Cureus 2024;16(1):e51961 [FREE Full text] [doi: 10.7759/cureus.51961] [Medline: 38333501]
50. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience 2024;27(5):109713 [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
51. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023;9:e50945 [FREE Full text] [doi: 10.2196/50945] [Medline: 37578830]
52. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
53. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport 2023;40(2):615-622 [FREE Full text] [doi: 10.5114/biolsport.2023.125623] [Medline: 37077800]
54. Graham F. Daily briefing: will ChatGPT kill the essay assignment? Nature 2022. [doi: 10.1038/d41586-022-04437-2] [Medline: 36517680]
55. Stokel-Walker C. AI bot ChatGPT writes smart essays—should professors worry? Nature 2022. [doi: 10.1038/d41586-022-04397-7] [Medline: 36494443]
56. Xu J, Lio A, Dhaliwal H, Andrei S, Balakrishnan S, Nagani U, et al. Psychological interventions of virtual gamification within academic intrinsic motivation: a systematic review. J Affect Disord 2021;293:444-465. [doi: 10.1016/j.jad.2021.06.070] [Medline: 34252688]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth
LLM: large language model
MCQ: multiple-choice question

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 02.02.24; peer-reviewed by Z Liao, F Tang; comments to author 22.05.24; revised version received 10.06.24; accepted 27.06.24; published 20.08.24.

Please cite as:
Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, Huang J, Zheng X, Zhang Y
Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial
J Med Internet Res 2024;26:e57037
URL: https://www.jmir.org/2024/1/e57037
doi:10.2196/57037
PMID:

©Wenyi Gan, Jianfeng Ouyang, Hua Li, Zhaowen Xue, Yiming Zhang, Qiu Dong, Jiadong Huang, Xiaofei Zheng, Yiyi Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Blended Learning Course on the Diagnostics of Mental Disorders: Multicenter Cluster Randomized Noninferiority Trial

Gabriel Bonnin[1,2*], MSc; Svea Kröber[1,2*], PhD; Silvia Schneider[1,2], PhD; Jürgen Margraf[1,2], PhD; Verena Pflug[1,2], PhD; Alexander L Gerlach[3], PhD; Timo Slotta[3], PhD; Hanna Christiansen[2,4], PhD; Björn Albrecht[2,4], PhD; Mira-Lynn Chavanon[2,4], PhD; Gerrit Hirschfeld[5], PhD; Tina In-Albon[6], PhD; Meinald T Thielsch[7], PhD; Ruth von Brachel[1,2], PhD

[1]Mental Health Research and Treatment Center (FBZ), Faculty of Psychology, Ruhr University Bochum, Bochum, Germany
[2]German Center for Mental Health (DZPG), Bochum-Marburg, Germany
[3]Clinical Psychology and Psychotherapy, Department of Psychology, University of Cologne, Cologne, Germany
[4]Clinical Child and Adolescent Psychology, Department of Psychology, Philipps University Marburg, Marburg, Germany
[5]Faculty of Business, University of Applied Sciences Bielefeld, Bielefeld, Germany
[6]Clinical Child and Adolescent Psychology and Psychotherapy, Department of Psychology, University of Mannheim, Mannheim, Germany
[7]Work and Environmental Psychology, Department of Psychology, University of Wuppertal, Wuppertal, Germany
[*]these authors contributed equally

Corresponding Author:
Gabriel Bonnin, MSc
Mental Health Research and Treatment Center (FBZ)
Faculty of Psychology
Ruhr University Bochum
Massenbergstraße 9-13
Bochum, 44787
Germany
Phone: 49 (0)234 32 217
Email: gabriel.bonnin@rub.de

Abstract

Background: Clinical diagnoses determine if and how therapists treat their patients. As misdiagnoses can have severe adverse effects, disseminating evidence-based diagnostic skills into clinical practice is highly important.

Objective: This study aimed to develop and evaluate a blended learning course in a multicenter cluster randomized controlled trial.

Methods: Undergraduate psychology students (N=350) enrolled in 18 university courses at 3 universities. The courses were randomly assigned to blended learning or traditional synchronous teaching. The primary outcome was the participants' performances in a clinical diagnostic interview after the courses. The secondary outcomes were diagnostic knowledge and participants' reactions to the courses. All outcomes were analyzed on the individual participant level using noninferiority testing.

Results: Compared with the synchronous course (74.6% pass rate), participation in the blended learning course (89% pass rate) increased the likelihood of successfully passing the behavioral test (odds ratio 2.77, 95% CI 1.55-5.13), indicating not only noninferiority but superiority of the blended learning course. Furthermore, superiority of the blended learning over the synchronous course could be found regarding diagnostic knowledge (β=.13, 95% CI 0.01-0.26), course clarity (β=.40, 95% CI 0.27-0.53), course structure (β=.18, 95% CI 0.04-0.32), and informativeness (β=.19, 95% CI 0.06-0.32).

Conclusions: Blended learning can help to improve the diagnostic skills and knowledge of (future) clinicians and thus make an important contribution to improving mental health care.

Trial Registration: ClinicalTrials.gov NCT05294094; https://clinicaltrials.gov/study/NCT05294094

(J Med Internet Res 2024;26:e54176) doi:10.2196/54176

XSL•FO
RenderX

KEYWORDS

diagnosis; structured clinical interviews; blended learning; dissemination; therapist training; clinical interview; clinical diagnosis; clinical practice; psychology students; diagnostic test; health personnel; mental health services; mental health

Introduction

The reliable and valid diagnosis of mental disorders is associated with a more favorable therapeutic course and outcome [1,2]. However, although structured clinical interviews are acknowledged as the gold standard for diagnosing mental disorders [3-5], clinicians often rely on unstructured, experience-based explorations of symptoms [6,7]. As a result, both the under- and overdiagnosis of mental disorders are common [8-10], leading to undertreatment [11] or inappropriate or unnecessary psychotherapy or medication [12-14]. In view of the high rates of misdiagnoses, there is an urgent need to improve diagnostics of mental disorders by disseminating evidence-based assessment procedures into clinical practice. While there is increasing awareness of the importance of disseminating evidence-based treatment [15,16], the foundation of successful treatment, namely evidence-based diagnostics, has not been sufficiently addressed in dissemination research [17,18]. Therefore, the aim of this study was to develop and evaluate a blended learning course to disseminate evidence-based diagnostics of mental disorders.

Learning environments can be classified into 4 groups based on their modality (offline vs online) and synchronicity (synchronous vs asynchronous). Traditional classroom teaching combines synchronous with offline teaching, while webinars are an example of synchronous online teaching. Asynchronous teaching can be realized online through learning management systems or offline with the use of printed learning materials. Recent meta-analytical evidence indicates that there are no significant differences in learning outcomes between synchronous offline and online learning [19,20], as well as asynchronous online instruction [20]. In addition, there are various combinations of these learning environments, including blended learning, which combines (synchronous) offline and (asynchronous) online learning [21]. This approach addresses the limitations of both traditional offline and online learning, such as reduced student engagement and inconvenient time and space requirements [22]. By combining the best of both approaches, blended learning provides a more effective learning experience. Recent meta-analytical evidence suggests that blended learning is consistently more effective than either synchronous online or offline learning alone and is better accepted than traditional offline teaching [22-24].

In addition, a blended learning approach may be especially appropriate for disseminating evidence-based diagnostics, as it addresses challenges commonly associated with its training and dissemination in a face-to-face setting [25]. First, time and costs that are required for training in evidence-based methods act as important barriers for attendance [26]. Web-based or blended training methods can overcome this barrier by allowing clinical knowledge and skills to be trained in a time- and cost-efficient, easily accessible, flexible, and highly standardized way [27-29]. Second, viewing case studies of and practicing diagnostic situations are essential for acquiring diagnostic skills. As videos of diagnostic situations with simulated patients can be included in the asynchronous online part, whereas practicing in role plays can occur during the synchronous face-to-face part of the course, a blended learning approach allows learning practical diagnostic skills in a flexible and time-saving way. Third, clinicians underestimate patient acceptance of structured interviews and seem to have various preconceptions against their use [6]. These preconceptions can be reduced by addressing them explicitly and by intensifying training in the implementation of structured interviews [18,26,30]. By conveying content in a highly accessible and standardized way, blended learning courses can contribute to the intensification and standardization of training in structured diagnostic interviews and hereby reduce prejudices against their use.

Although the use and acceptance of online teaching methods increased globally during the COVID-19 pandemic [31,32], until today, only a few studies evaluated blended learning in randomized controlled trials [33-36]. Despite the high relevance of disseminating evidence-based diagnostics into clinical practice, to our knowledge, blended learning for teaching diagnostic skills was not yet evaluated at all. We aimed to fill this gap by conducting a cluster randomized controlled trial at 3 German universities and comparing a blended learning course with regular face-to-face teaching in a noninferiority analysis at the individual participant level. As there is evidence that the impact of training on more experienced practitioners does not last over time [37,38], and it is considered that such training may be more effective for those at the beginning of their clinical careers [39], we targeted a relatively inexperienced sample of preprofessionals, specifically undergraduate psychology students.

We hypothesize that students' interviewing skills, knowledge acquisition, and reactions in a novel blended learning course will be noninferior to those in traditional synchronous courses.

Methods

Study Design

The study was a multicenter cluster randomized controlled trial, comparing 2 university teaching formats: a blended learning course and a traditional synchronous course. A cluster randomization of courses was chosen because individual randomization of participants was not feasible given the constraints of the existing university setting. Clusters were 18 courses in clinical diagnostics at the 3 cooperating universities. Courses were randomly assigned to 1 of the 2 teaching conditions, stratified by study site. Participants could choose between courses in the online registration systems of the respective universities. To minimize any selection bias, course information available to participants (eg, content and instructor) was held constant in both conditions. Importantly, participants had no information about whether the teaching condition was synchronous or blended. Since the study was conducted at 3 different universities with different numbers of students and

XSL•FO
RenderX

teachers, the number and size of courses at each center varied. There were 3 assessments—before the start of the courses (t1), before the last course session (t2), and after the last course session (t3).

While both teachers and participants were aware of the teaching condition to which they were assigned, key aspects of the assessment process were conducted under blinded conditions. Specifically, the actors portraying patients in the diagnostic skills test were unaware of the participants' assigned teaching conditions to ensure impartiality in their interactions. In addition, the outcome assessors responsible for scoring the videotaped simulated diagnostic scenarios were also blinded to the participants' group assignments, ensuring that the assessment of diagnostic skills was not influenced by knowledge of the training method.

The trial is registered on ClinicalTrials.gov (NCT05294094). Due to governmental protective restrictions in Germany related to the COVID-19 pandemic, synchronous sessions that were originally planned as face-to-face classes had to be conducted as online webinars. This deviation from the registered study protocol is addressed in the *Discussion* section of this paper, where the potential consequences and limitations are analyzed in detail.

In accordance with the journal guidelines for reporting randomized controlled trials of eHealth interventions, the CONSORT-EHEALTH checklist is included in Multimedia Appendix 1.

Participants

A total of 3 universities took part in the study (Ruhr University Bochum, University of Cologne, and Philipps University Marburg). The eligibility criterion for clusters was an undergraduate course on clinical diagnostics at cooperating universities. Eligibility criteria for individual participants were (1) age >18 years, (2) undergraduate psychology students at a cooperating university, and (3) willingness to give informed consent online.

Participation in a course on the diagnostics of mental disorders was mandatory in the curriculum of the undergraduate psychology program at all cooperating universities. In total, 18 courses were offered, 10 of which focused on the diagnostics of mental disorders in adulthood and 8 of which focused on the diagnostics of mental disorders in childhood and adolescence. Participants were recruited over the course of 2 semesters between April 2021 and February 2022. During this period, the courses were attended by 400 students. Participation was possible for all students at each of the 3 measurement time points separately.

Procedure

Overview

Before the start of the course, written informed consent was obtained. Study participation was voluntary and compensated with a test participant certificate (mandatory part of the study program) and a shopping voucher (€10-20 [US $10.84-21.68], value depending on the scope of study participation). While the synchronous online classes were held weekly from the beginning, the blended learning course started 6 weeks into the current semester for organizational reasons. The asynchronous online part of the blended learning course was accessible to all students through the Moodle learning platform.

Experimental Condition: Blended Learning Course

Overview

The blended learning course followed a flipped classroom model, in which asynchronous online lessons focused on content delivery and synchronous online sessions were used to apply and deepen clinical skills under the guidance of an instructor [40]. The course consisted of 8 asynchronous online lessons and 3 synchronous online sessions and was designed considering the current knowledge regarding the conditions under which blended learning is effective (eg, including case studies, interactive elements with personalized feedback, or collaborative activities during synchronous sessions [32,41]) and well accepted by students (eg, user-friendly and functional design [42]).

A detailed overview of the course content is illustrated in Table 1. Access to the asynchronous online course can be provided by the corresponding author (GB) on request.

XSL•FO
RenderX

Table 1. Structure and content of the blended learning course.

Part and lesson	Adulthood	Childhood and adolescence
Part I: Asynchronous lessons		
Lessons 1-3: Diagnostic fundamentals and evidence-based assessment	1. Introduction to classificatory diagnostics, diagnostic approaches, and classification systems 2. The diagnostic process, standardized clinical assessment, and biasing influences on the diagnostic process 3. Structure, conduction, and development of the (Kinder-)DIPS-OA[a]	1-3. Same as adulthood
Lessons 4-7: Diagnostic criteria and conduction of the (Kinder-)DIPS-OA for specific disorders	4. Panic disorder, agoraphobia, social anxiety disorder, and generalized anxiety disorder 5. Bipolar disorders, major depression, persistent depressive disorder, and OCD[c] 6. PTSD[d], somatic symptoms disorder, and illness anxiety disorder 7. Anorexia nervosa, bulimia nervosa, and alcohol use disorder	4. ADHD[b], oppositional defiant disorder, and conduct disorder 5. Separation anxiety disorder, specific phobia, and social anxiety disorder 6. Generalized anxiety disorder, selective mutism, and major depression 7. PTSD, OCD, and anorexia nervosa
Lesson 8: Evaluation of the (Kinder-)DIPS-OA	8. Evaluation of the (Kinder-)DIPS-OA, feedback of a diagnosis, difficult situations conducting the (Kinder-)DIPS-OA, and acceptance and psychometric properties of the (Kinder-)DIPS-OA	8. Same as adulthood
Part II: Synchronous Sessions		
Lessons 9-10	9. Apply skills and conduct the (Kinder-)DIPS-OA as the interviewer and as a patient with fellow students. Get direct feedback from peers and teacher. 10. Other nonspecified content was based on the students' questions and interests (eg, questions regarding the diagnostic criteria, the diagnostic process, and how to conduct the [Kinder-]DIPS).	9-10. Same as adulthood

[a](Kinder-)DIPS-OA: Diagnostic Interview for Mental Disorders (in Children and Adolescents) – Open Access.

[b]ADHD: attention-deficit/hyperactivity disorder.

[c]OCD: obsessive compulsive disorder.

[d]PTSD: posttraumatic stress disorder.

Blended Learning Course—Asynchronous Lessons

In total, 2 separate versions of the asynchronous online course were developed—a version with a focus on the diagnostics of mental disorders in childhood and adolescence and a version with a focus on the diagnostics of mental disorders in adulthood. Both versions were parallel in content, except for age-specific diagnostic procedures and some of the disorders presented, as they typically occur at different developmental stages (Table 1). Furthermore, both versions focused on teaching the conduction of a semistructured diagnostic interview, the Diagnostic Interview for Mental Disorders – Open Access 1.2 (DIPS-OA1.2) [43], and the Diagnostic Interview for Mental Disorders in Children and Adolescents – Open Access (Kinder-DIPS-OA) [44].

Content, usability, and design of the online courses were formatively evaluated during development by students and research associates from the Ruhr University Bochum and the University of Koblenz-Landau.

Each lesson included an introduction and conclusion sequence, a downloadable handout, and final evaluation questions. In the disorder-specific lessons (4-7), video-based case studies (played by actors) were presented to illustrate the conduction of a structured interview and to allow participants to test and apply their acquired knowledge through interactive elements (eg, multiple choice questions, automatic feedback, and matching tasks). Participants were able to navigate through the lessons and subchapters independently; however, working through the course content in sequential order was recommended. A tutorial video was provided, explaining how to navigate through the course, as well as how to use the various interactive course elements. In addition, the course included a forum where participants could ask questions about the course content, which were answered by the first 2 study authors (GB and SK).

Blended Learning Course—Synchronous Sessions

Following the asynchronous online course, participants of the experimental condition took part in 3 weekly synchronous online sessions (90 min each). In these sessions, they could discuss questions about the asynchronous course content with a lecturer and apply their skills in role plays with the other participants.

Control Condition: Synchronous University Course

The synchronous university course took place in attendance and consisted of 11 weekly online sessions (90 min each), representing the usual teaching of clinical diagnostic knowledge and skills at the 3 cooperating universities. The teachers were instructed to work through mandatory content, which was based on the asynchronous online course to ensure comparability between the 2 conditions. Before the start of the course, a training session was held for the teachers. In addition, course material was provided in the form of Microsoft PowerPoint slides. In addition to the mandatory content, teachers were

XSL•FO
RenderX

allowed to provide additional information relevant to the field of clinical diagnostics.

Measures and Assessments

Primary Outcome: Practical Diagnostic Skills

The primary outcome was the students' performance in a simulated structured diagnostic interview. At t2, course participants individually conducted a 15-minute section of a structured clinical interview (Diagnostic Interview for Mental Disorders [in Children and Adolescents] - Open Access; [Kinder-]DIPS-OA) with patients played by previously trained actors through video chat. All actors were blinded to the assigned teaching condition. Patient roles were based on 1 out of 3 case vignettes distributed evenly across courses, each for a different disorder (generalized anxiety disorder, obsessive-compulsive disorder, or major depression; "Section A" in Multimedia Appendix 2). Each case vignette included instructions to the actors to simulate difficult interview situations (eg, "Miss the point with your answer to this question."; "Section A" in Multimedia Appendix 2). The interviews were videotaped and then rated by 4 blinded and independent evaluators using a coding scheme ("Section B" in Multimedia Appendix 2), which assessed 2 facets of interview performance—formal interviewing skills (10 items; eg, "The interviewer asks relevant additional questions beyond the interview guide to assess the presence of the diagnostic criteria.") and interpersonal interviewing skills (9 items; eg, "The interviewer uses non-verbal and paraverbal interviewing techniques."). Both dimensions were assessed on scales ranging from 0 to 100. To succeed in adequately conducting the structured interview, participants had to score at least 50% correct on both scales. The cutoff of 50% is commonly used in the German education system.

All the outcome assessors had a master's degree in psychology, were certified and experienced conducting the (Kinder-)DIPS-OA, and received at least 2 years of postgraduate cognitive behavioral therapy training. Interrater reliability for each item was calculated based on 40 jointly coded interviews, with Fleiss κ ranging between fair (0.34) and almost perfect (0.96) agreement between outcome assessors [45].

Secondary Outcomes

Diagnostic Knowledge

Two parallel 15-item versions of a test of basic clinical diagnostic knowledge were created, which participants answered at t1 and t3 (refer to "Section C" in Multimedia Appendix 2 for example items). The format of the items varied (single choice, multiple select, and multiple-true-false) and the items were previously piloted with laypersons (30 undergraduate students in their first semester) and experts (44 therapists in postgraduate training). Items were selected based on item-scale correlation and discrimination between these 2 groups. In addition, at t1, the self-reported diagnostic knowledge was assessed on an 11-point Likert-type scale ("How knowledgeable are you in the area of 'clinical diagnostics'?"; 0= "I don't know anything about it", 10= "I am very knowledgeable in this area").

Participants' Reactions

Participants' reactions to the courses and the estimated patient acceptance of structured interviews were evaluated at t3 by means of an online questionnaire, which consisted of 32 selected items (Table 2) from several instruments [6,46-50]. There were 8 additional items only administered in the blended learning condition. Unless otherwise described, a 7-point Likert-type scale was used for the items, ranging from 1="strongly disagree" to 7="strongly agree", with higher scores indicating a better outcome.

XSL•FO
RenderX

Table 2. Overview of the items assessing participants' reactions.

Questionnaire and subscale	Items, n	Example	Cronbach α
MFE-Sr[a] [46]			
Intent to recommend	1	"I would recommend this course to other students."	__[b]
Experience of overload	3	"The content of this course was too difficult for me."[c]	0.77
Subjective learning success	1	"I learned a lot in this course."	—
Web-CLIC[d] [46]			
Clarity	3	"The contents of the course are clearly presented"	0.83
Likeability	3	"The course arouses my interest"	0.93
Informativeness	3	"The information is of high quality"	0.91
Credibility	3	"I can trust the information in the course"	0.95
Short scale for academic course evaluation [47]			
Course structure	3	"The course was clearly structured."	0.73
UMUX-Lite[e] [48]			
Usability	2	"This system is easy to use"[f]	0.82-0.83
VisAWI-S[g] [49]			
Visual Aesthetics	4	"The layout is professional"[f]	0.76
Items designed by the study authors			
Visual Aesthetics	1	"DiSkO is designed to be visually appealing"[f]	—
Credibility	1	"I completely trusted the content in DiSkO"[f]	—
Overall impression	1	"Overall: I give the course an overall grade of …"[c,h]	—
Acceptance of structured interviews questionnaire [6]			
Global satisfaction rating	1	"Please indicate on the accompanying scale how satisfied you think patients are or would be with structured diagnostic interviews in general."[i]	—
Mental effort and emotional reaction to structured interviews	10	"After a structured interview, patients feel more confused than before."[j]	—

[a]MFE-Sr: Münster Questionnaire for the Evaluation of Seminars – revised.
[b]Not applicable
[c]Lower scores indicate a better outcome.
[d]Web-CLIC: Website-Clarity, Likeability, Informativeness, and Credibility.
[e]UMUX-Lite: Usability Metric for User Experience – Lite.
[f]These items were only administered in the blended learning condition.
[g]VisAWI-S: Visual Aesthetics of Websites Inventory – Short.
[h]This item used the German grading system ranging from 1 (excellent) to 6 (insufficient).
[i]Visual analog scale ranging from 0 (not at all satisfied) to 100 (completely satisfied).
[j]4-point Likert scale ranging from 0 (disagree) to 3 (completely agree).

Statistical Analyses

All outcomes were evaluated at the individual participant level using noninferiority analyses. To assess noninferiority between the teaching conditions, 95% CIs were calculated. Although our hypothesis regarding the noninferiority of the blended learning course is inherently 1-sided, the use of a 2-sided 95% CI is standard practice [51] and recommended by the guidelines of the European Medicines Agency and the US Food and Drug Administration [52,53]. This approach allows for the interpretation of noninferiority if the entire 95% CI lies above the prespecified noninferiority margin.

A logistic regression was conducted, predicting the primary outcome (passing the performance test) based on the predictor's teaching condition (blended learning vs synchronous), study site (center 1 vs center 2 vs center 3), course focus (adulthood vs childhood and adolescence), study year, self-reported diagnostic knowledge, and the score in the knowledge test at t1. To account for the effects of cluster randomization, the

XSL•FO
RenderX

course variable was included as a random effect (random-intercept). Odds ratios (ORs) were calculated by unconditional maximum likelihood estimation and 95% CI using normal approximation. Based on experience with the traditional synchronous course format in diagnostic teaching, a passing rate of 85% was assumed for the synchronous course. As the passing rate after blended learning should be at least as good as that in traditional face-to-face instruction due to the positive effects of blended learning on learning outcomes [24], a 90% passing rate was assumed in the blended learning course. To test for noninferiority, the assumed passing rates and noninferiority margin of 5% were transferred to ORs [54,55]:

Accordingly, a power analysis for noninferiority trials with dichotomous data revealed for expected success rates of 85% and 90%, respectively; noninferiority margin of 5%; α=.05; and power of 80% a required sample size of 135 per treatment group [56].

All secondary outcome measures were tested using multiple linear regression models with the following predictors: teaching condition, study site, course focus, study year, self-reported diagnostic knowledge, and the score in the knowledge test at t1. Noninferiority of the blended learning course was assumed, when the lower bound of the CI of the predictor teaching condition was larger than β=–.10, corresponding to a small negative effect.

To test for systematic differences between teaching conditions at baseline, *t* tests and Fisher exact tests were conducted. Furthermore, we tested for differences in assigned teaching condition and practical diagnostic test performance between completers (participation in t1, t2, and t3) and noncompleters using chi-square tests.

All available data were analyzed for each statistical test performed. All analyses were run in R (R Core Team) [57]. The anonymized dataset [58] and R code [59] are available online.

Ethical Considerations

The authors declare that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. This study protocol was reviewed and approved by the local ethics committee of the faculty of psychology at the Ruhr University Bochum (2021/686). Written informed consent was obtained from participants to participate in the study.

Results

Participant Characteristics at Baseline

Figure 1 shows the distribution of participants among the courses and the sample sizes at the measurement time points. A total of 350 participants took part in at least 1 of the 3 measurement time points, 203 of whom participated in all of them. Demographic data were missing from 17 participants because they did not complete the online survey that was part of the behavioral test. Participants (n=333) had a mean age of 23.6 (SD 4.52) years and the majority identified as women (279/333, 83.8%). The average study year was 2.82 (SD 0.93).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart of students enrolled in the courses and participating in the study.

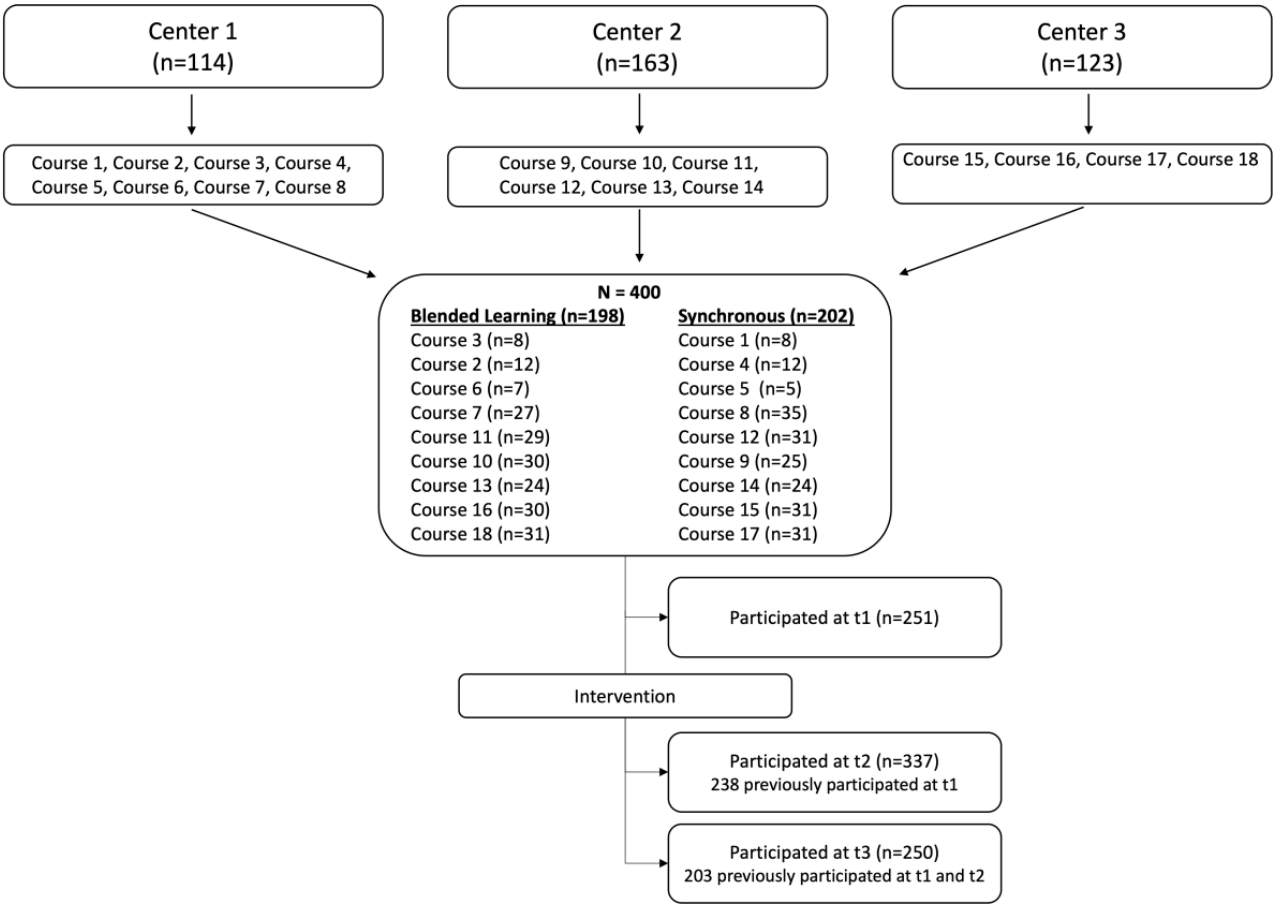

The *t* test and Fisher exact test revealed no significant differences between teaching conditions at baseline (all *P*>.05; Table 3). Chi-square tests showed no significant differences between completers and noncompleters regarding teaching condition (χ^2_1=0.79, *P*=.37) and behavioral performance (χ^2_1=0.82, *P*=.37).

Table 3. Participants' baseline demographic characteristics and diagnostic knowledge.

Characteristic	Participants, n	Blended learning	Synchronous	*t* test (*df*)	*P* value
Age (years), mean (SD)	333	24.1 (4.20)	23.2 (4.76)	1.81 (331)	.07
Study year, mean (SD)	333	2.88 (0.93)	2.76 (0.94)	1.09 (331)	.28
Diagnostic knowledge, mean (SD)					
Test score (t1)	251	9.44 (1.71)	9.36 (1.79)	0.37 (249)	.71
Self-rating	333	3.46 (2.07)	3.51 (2.18)	–0.22 (331)	.82
Gender, n (%)	350				.05
Missing		11 (6.47)	7 (3.89)	—[a]	
Female		127 (74.71)	152 (84.44)	—	
Male		31 (18.23)	18(10.0)	—	
Diverse		1 (0.59)	3 (1.67)	—	

[a]Not applicable.

Primary Outcome—Practical Diagnostic Skills

Overall, participants showed high levels of interpersonal (blended learning: mean 74.8, SD 16.2; synchronous: mean 70.7, SD 18.1) and formal skills (blended learning: mean 86.1, SD 14.4; synchronous: mean 82.8, SD 16.6). The passing rate was 89% in the blended learning condition and 74.6% in the synchronous condition, corresponding to an OR of 2.77 (95% CI 1.52-5.03).

We furthermore tested whether this finding still held up when several covariates were considered. For this, we fitted a logistic mixed model (adjusted model; n=320) including the course variable as random effect to take cluster randomization into account (Table 4; for all model coefficients, refer to "Section

XSL•FO
RenderX

D" in Multimedia Appendix 2). As the model resulted in convergence errors with all predictors, the knowledge test score at t1 and the study center had to be excluded. More complex models accounting for the nesting of courses within universities were attempted to be fitted but resulted in convergence errors. Intraclass correlation is not reported as the results of the generalized linear mixed model did not contain the residual variance required to calculate the intraclass correlation.

Table 4. Odds ratio (OR) and 95% CI for the unadjusted and adjusted model.

Predictors	Unadjusted model, OR (95% CI)	Adjusted model, OR (95% CI)
Teaching condition	2.77 (1.55-5.13)	3.20 (1.56-6.71)
Center 2	—[a]	—
Center 3	—	—
Course focus	—	0.42 (0.17-0.96)
Study year	—	1.36 (0.82-2.42)
Self-reported knowledge	—	1.06 (0.89-1.25)
Knowledge test (t1)	—	—
Random effects		
σ^2	—	3.29
N_{course}	—	18
Observations	337	320
Tjur D	0.035	0.214

[a]Not applicable.

Secondary Outcomes

Overview

To describe the magnitude of the differences in secondary outcomes between the groups, multiple linear regression models were calculated (Table 5). For the complete covariate-adjusted models, refer to "Section E" in Multimedia Appendix 2.

XSL•FO
RenderX

Table 5. Means, SDs, and β-coefficients with CIs for all secondary outcomes.

Secondary outcome	Range	Blended learning (n=117), mean (SD)	Synchronous (n=132), mean (SD)	Teaching condition, β (95% CI)
Diagnostic knowledge				
Knowledge test (t3)	0-15	12.0 (1.65)	11.4 (1.68)	.13 (0.01 to 0.26)
Participants' reactions				
Intent to recommend	0-7	6.21 (1.06)	6.09 (1.13)	.09 (–0.05 to 0.22)
Experience of overload[a]	0-7	2.70 (1.04)	2.40 (1.04)	.20 (0.07 to 0.34)
Subjective learning success	0-7	5.89 (1.02)	5.80 (1.06)	.04 (–0.096 to 0.18)
Clarity	0-7	6.18 (0.74)	5.48 (0.91)	.40 (0.27 to 0.53)
Likeability	0-7	6.09 (0.96)	5.86 (1.21)	.09 (–0.04 to 0.23)
Informativeness	0-7	6.38 (0.68)	6.19 (0.72)	.19 (0.06 to 0.33)
Credibility	0-7	6.46 (0.64)	6.45 (0.62)	.08 (–0.05 to 0.22)
Course structure	0-7	6.30 (0.72)	5.98 (0.92)	.18 (0.04 to 0.32)
Overall impression	1-6	1.56 (0.81)	1.72 (0.75)	–.12 (–0.26 to 0.01)
Visual Aesthetics	0-7	5.87 (0.91)	__[b]	—
Usability	0-100	86.3 (14.1)	—	—
Acceptance				
Global rating	0-100	77.7 (14.2)	76.8 (16.2)	.04 (–0.095 to 0.18)
"More confused"[a]	0-3	0.26 (0.48)	0.42 (0.58)	–.11 (–0.25 to 0.03)
"questioned out"[a]	0-3	1.18 (0.74)	1.20 (0.85)	.05 (–0.08 to 0.19)
"too many questions"[a]	0-3	1.11 (0.81)	1.06 (0.76)	.03 (–0.11 to 0.16)
"exhausting"[a]	0-3	1.15 (0.71)	1.07 (0.77)	.06 (–0.08 to 0.20)
"taken seriously"	0-3	2.33 (0.88)	2.33 (0.84)	–.01 (–0.15 to 0.13)
"positive relationship"	0-3	1.95 (0.80)	2.04 (0.75)	–.10 (–0.24 to 0.04)
"not report everything"[a]	0-3	1.33 (0.81)	1.32 (0.86)	.04 (–0.10 to 0.18)
"better understanding"	0-3	1.56 (0.76)	1.50 (0.80)	.07 (–0.07 to 0.21)
"enough detail"	0-3	2.30 (0.75)	2.36 (0.72)	–.05 (–0.19 to 0.09)
"helpful"	0-3	2.12 (0.74)	2.14 (0.72)	–.03 (–0.17 to 0.12)

[a]Lower scores indicate better outcome. For negatively-keyed items, noninferiority of the blended learning course can be assumed if the upper bound of the CI is greater than 0.10.

[b]Not applicable.

Diagnostic Knowledge

Participation in the blended learning course increased the knowledge score at t3 (β=.13, 95% CI 0.01-0.26). Thus, noninferiority and superiority of the blended learning course regarding diagnostic knowledge at t3 can be assumed.

Participants' Reactions

Noninferiority of the blended learning course regarding participants' reactions to the courses could be observed in most measures collected. Only with regard to the experience of overload the blended learning course was inferior to the synchronous course, with lower scores indicating a more favorable outcome (β=.20, 95% CI 0.07-0.34). Furthermore, the superiority of the blended learning over the synchronous course could be found in the following subscales: clarity (β=.40, 95% CI 0.27-0.53), course structure (β=.18, 95% CI 0.04-0.32), and informativeness (β=.19, 95% CI 0.06-0.32).

Regarding the estimated patient acceptance of structured interviews, noninferiority of the blended learning course was observed for the global acceptance rating (β=.04, 95% CI –0.095 to 0.18) and the items "After a structured interview, patients feel more confused than before" (β=–0.11, 95% CI –0.25 to 0.03) and "Patients have the feeling that they understand themselves and their problems better, after a structured interview" (β=.07, 95% CI –0.07 to 0.21). For the other items, student's estimation did not differ between the blended learning and the synchronous courses.

XSL•FO
RenderX

Discussion

Principal Findings

The aim of the present study was to establish whether a blended learning course with a reduced personal contact time results in comparable clinical diagnostic skills as a traditional synchronous online course.

The results of this study are in line with and extend the existing literature on blended learning [24,35,36]. First, noninferiority and superiority of the blended learning course over the synchronous course could be found for the primary outcome measure—the performance in a simulated structured diagnostic interview. Second, noninferiority and superiority were also observed for the diagnostic knowledge test score at t3 and several reaction measures, such as clarity, informativeness, and structure of the course. Furthermore, noninferiority could be observed regarding the intention to recommend the course to other students, subjective learning success, likeability, credibility, overall impression of the course, and 3 items of the estimated patient acceptance of structured clinical interviews. Third, inferiority of the blended learning compared with the synchronous course was found for the participants' experience of overload.

Despite the described differences, participants in both courses showed high levels of interpersonal and formal skills, good diagnostic knowledge, positive reactions to the courses, and high-estimated patient acceptance. While therapists were found to underestimate patient acceptance of structured interviews [6], estimated patient acceptance ratings in this study correspond more closely to patients' actual acceptance ratings [60], indicating that participants of the present study estimated patient acceptance more accurately than did therapists in the aforementioned study.

Limitations and Strengths

The study has some limitations that should be mentioned. First, the blended learning course was presented as a block, meaning that participants had 3 weeks of time to work through the asynchronous online content followed by 3 weekly synchronous online sessions. In contrast, the synchronous online course consisted of 11 weekly sessions. The fact that participants only had 3 weeks for the online content, which was equivalent to 8 sessions of 90 minutes each, might be considered a disadvantage for the blended learning course. This might explain why participants in the blended learning course reported higher levels of overload than those in the synchronous online course. To reduce the experience of overload and possibly even further enhance students' performance, the blended learning course should be provided on a continuous basis in the future, ensuring continuous student activity [61]. Second, adherence in the control condition was not assessed. Although teachers were informed about the mandatory content in a training session and received course material before the start of the course, it remains unclear whether all mandatory content was in fact taught in the control condition. In contrast, the asynchronous online component of the blended learning course was meticulously developed. It was ensured that the blended learning course contained all the necessary content for the diagnostic skills and knowledge tests. In addition, formative evaluations were conducted to assess the content, usability, and design of the course. Based on the feedback received, the course was subsequently revised. Therefore, the blended course was designed and implemented with more effort than the control condition, which followed a "teaching as usual" rationale. Third, we did not evaluate how the participants used the courses and materials for preparation, repetition, and reflection of the lectures. The convenient repetition offered by the asynchronous materials in the blended-learning course may have been particularly beneficial for the practical skills and knowledge tests. Fourth, all synchronous sessions in both the experimental and control conditions had to be conducted online due to governmental restrictions associated with the COVID-19 pandemic. As a result, the experimental condition had to be adapted to a hybrid combination of synchronous and asynchronous online teaching, although it was initially intended to be a combination of synchronous face-to-face teaching and asynchronous online teaching. Thus, it could be argued that our experimental condition does not strictly qualify as blended learning since it did not include face-to-face teaching. To evaluate the potential implications of this adaption, it is important to consider several differences between synchronous online and offline teaching. While both provide the benefit of immediate educational support and feedback [62], online teaching has logistical, instructional, and financial benefits over offline teaching [63]. However, online learning's logistical flexibility also has the potential downside of causing social isolation for the learner [64]. In addition, the integration of technology in educational settings inevitably increases the likelihood of technical issues, which can decrease satisfaction and participation [64,65]. Recent meta-analyses suggest that there are no significant differences in learning success between online and offline teaching [19,20]. Furthermore, online teaching received significantly higher satisfaction ratings compared with offline teaching [19]. It is important to note that our blended learning course primarily consisted of asynchronous online sessions. It can be assumed that the impact of 3 synchronous sessions, regardless of whether they are taught online or offline, is relatively small. However, it remains unclear how the blended learning course would compare with a traditional face-to-face course.

Besides these limitations, the study also has some notable strengths. First, as a multicenter cluster randomized controlled trial, it makes an important contribution to the scarce evidence on the efficacy of blended learning in general [33-35] and, more specifically, for teaching evidence-based diagnostics. Second, the design of the evaluation study was developed very carefully. For instance, the outcome measures were assessed with reliability and validity in mind, the case vignettes for students' performance tests included very precise instructions, and actors were trained beforehand to ensure a high standardization. In addition, the items of the knowledge test were piloted on laypersons and therapists. Third, as undergraduate psychology students from 3 German universities attending a mandatory seminar of the diagnostics of mental disorders were invited to participate, a large sample of 337 participants could be included in the analysis of the primary outcome and 203 participants took part at all 3 measurement time points. Fourth, as the study was

XSL•FO
RenderX

conducted in an ongoing university setting, a high external validity and generalizability of the study results can be assumed.

Clinical Implications and Future Research

As the results indicate that the blended learning course can be used to teach evidence-based diagnostics, we aim to disseminate the blended learning course open access throughout Germany—at universities (undergraduate and graduate courses), at institutions of tertiary education, and among practicing psychotherapists. In order to facilitate the adoption of the blended learning course [66], a technical infrastructure was chosen which is available free of charge and provides ongoing technical support. In addition, an interesting question for future research is whether structured interviews are in fact used more frequently after attending the blended learning course. Increasing the use of structured interviews in clinical practice is an important goal as therapists appeared to use structured interviews only with 14.8% (55/370) of their patients [6]. Until today, research on therapist training is limited, especially when it comes to web-based training [67]. Therefore, to extend the promising findings of this study, future research should also focus on the development and evaluation of further blended learning courses to improve evidence-based practice in clinical psychology in general.

Conclusion

In conclusion, this study suggests that a blended learning course, compared with a synchronous online course in a cluster randomized controlled trial, can be used to efficiently teach evidence-based diagnostics. The results indicate that the blended learning approach was more effective than synchronous online teaching in acquiring practical diagnostic skills and diagnostic knowledge, and that it was well received by the students. The blended learning course can therefore help to improve the skills and knowledge of (future) clinicians in a time- and cost-efficient way and thus make an important contribution to improving the diagnostics of mental disorders and the mental health care situation in the long term.

Acknowledgments

The authors would like to thank Julia Glombiewski and Saskia Scholten for their help with the formative evaluation of the course material. This work was funded by the German Federal Ministry of Education and Research (grant 16DHB3010).

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.x).
[PDF File (Adobe PDF File), 1043 KB - jmir_v26i1e54176_app1.pdf]

Multimedia Appendix 2
Case vignette and instructions for the actors, coding scheme, knowledge test, logistic regression coefficients for the primary outcome, and multiple linear regression coefficients for secondary outcomes.
[DOCX File , 100 KB - jmir_v26i1e54176_app2.docx]

References

1. Jensen-Doss A, Weisz JR. Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. J Consult Clin Psychol 2008;76(5):711-722. [doi: 10.1037/0022-006X.76.5.711] [Medline: 18837589]
2. Pogge DL, Wayland-Smith D, Zaccario M, Borgaro S, Stokes J, Harvey PD. Diagnosis of manic episodes in adolescent inpatients: structured diagnostic procedures compared to clinical chart diagnoses. Psychiatry Res 2001;101(1):47-54. [doi: 10.1016/s0165-1781(00)00248-1] [Medline: 11223119]
3. Ehlert U. Eine Psychotherapie ist immer nur so gut wie ihre Diagnostik. Psychotherapy is only as good as its diagnostics. Verhaltenstherapie 2007;17(2):81-82. [doi: 10.1159/000103156]
4. Merten EC, Schneider S. Clinical interviews with children and adolescents. In: Hofmann SG, editor. Clinical Psychology: A Global Perspective. Hoboken, NJ: Wiley; 2017.
5. Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. Int J Methods Psychiatr Res 2009;18(3):169-184 [FREE Full text] [doi: 10.1002/mpr.289] [Medline: 19701924]
6. Bruchmüller K, Margraf J, Suppiger A, Schneider S. Popular or unpopular? Therapists' use of structured interviews and their estimation of patient acceptance. Behav Ther 2011;42(4):634-643. [doi: 10.1016/j.beth.2011.02.003] [Medline: 22035992]
7. Jensen-Doss A, Youngstrom EA, Youngstrom JK, Feeny NC, Findling RL. Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. J Consult Clin Psychol 2014;82(6):1151-1162 [FREE Full text] [doi: 10.1037/a0036657] [Medline: 24773574]

XSL•FO
RenderX

8. Merten EC, Cwik JC, Margraf J, Schneider S. Overdiagnosis of mental disorders in children and adolescents (in developed countries). Child Adolesc Psychiatry Ment Health 2017;11:5 [FREE Full text] [doi: 10.1186/s13034-016-0140-5] [Medline: 28105068]
9. Mojtabai R. Clinician-identified depression in community settings: concordance with structured-interview diagnoses. Psychother Psychosom 2013;82(3):161-169. [doi: 10.1159/000345968] [Medline: 23548817]
10. Ruggero CJ, Zimmerman M, Chelminski I, Young D. Borderline personality disorder and the misdiagnosis of bipolar disorder. J Psychiatr Res 2010;44(6):405-408 [FREE Full text] [doi: 10.1016/j.jpsychires.2009.09.011] [Medline: 19889426]
11. Vermani M, Marcus M, Katzman MA. Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study. Prim Care Companion CNS Disord 2011;13(2):PCC.10m01013 [FREE Full text] [doi: 10.4088/PCC.10m01013] [Medline: 21977354]
12. Berardi D, Menchetti M, Cevenini N, Scaini S, Versari M, de Ronchi D. Increased recognition of depression in primary care. Comparison between primary-care physician and ICD-10 diagnosis of depression. Psychother Psychosom 2005;74(4):225-230. [doi: 10.1159/000085146] [Medline: 15947512]
13. Bruchmüller K, Margraf J, Schneider S. Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. J Consult Clin Psychol 2012;80(1):128-138. [doi: 10.1037/a0026582] [Medline: 22201328]
14. Margraf J, Schneider S. From neuroleptics to neuroscience and from Pavlov to psychotherapy: more than just the "emperor's new treatments" for mental illnesses? EMBO Mol Med 2016;8(10):1115-1117 [FREE Full text] [doi: 10.15252/emmm.201606650] [Medline: 27621275]
15. McMain S, Newman MG, Segal ZV, DeRubeis RJ. Cognitive behavioral therapy: current status and future research directions. Psychother Res 2015;25(3):321-329. [doi: 10.1080/10503307.2014.1002440] [Medline: 25689506]
16. Weisz JR, Ng MY, Bearman SK. Odd couple? Reenvisioning the relation between science and practice in the dissemination-implementation Era. Clinical Psychological Science 2013;2(1):58-74. [doi: 10.1177/2167702613501307]
17. Hunsley J, Mash EJ. Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. Psychol Assess 2005;17(3):251-255. [doi: 10.1037/1040-3590.17.3.251] [Medline: 16262451]
18. Jensen-Doss A, Hawley KM. Understanding barriers to evidence-based assessment: clinician attitudes toward standardized assessment tools. J Clin Child Adolesc Psychol 2010;39(6):885-896 [FREE Full text] [doi: 10.1080/15374416.2010.517169] [Medline: 21058134]
19. He L, Yang N, Xu L, Ping F, Li W, Sun Q, et al. Synchronous distance education vs traditional education for health science students: a systematic review and meta-analysis. Med Educ 2021;55(3):293-308. [doi: 10.1111/medu.14364] [Medline: 32881047]
20. Ebner C, Gegenfurtner A. Learning and satisfaction in webinar, online, and face-to-face instruction: a meta-analysis. Front. Educ 2019;4:92. [doi: 10.3389/feduc.2019.00092]
21. Laster S, Otte G, Picciano A, Sorg S. Redefining blended learning. 2005 Presented at: 2005 Sloan-C Workshop on Blended Learning; April 18, 2005; Chicago, IL.
22. Yu Z, Xu W, Sukjairungwattana P. Meta-analyses of differences in blended and traditional learning outcomes and students' attitudes. Front Psychol 2022;13:926947 [FREE Full text] [doi: 10.3389/fpsyg.2022.926947] [Medline: 36186290]
23. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020;22(8):e16504 [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
24. Schneider M, Preckel F. Variables associated with achievement in higher education: a systematic review of meta-analyses. Psychol Bull 2017;143(6):565-600. [doi: 10.1037/bul0000098] [Medline: 28333495]
25. Shafran R, Clark DM, Fairburn CG, Arntz A, Barlow DH, Ehlers A, et al. Mind the gap: improving the dissemination of CBT. Behav Res Ther 2009;47(11):902-909. [doi: 10.1016/j.brat.2009.07.003] [Medline: 19664756]
26. Stewart RE, Chambless DL, Baron J. Theoretical and practical barriers to practitioners' willingness to seek training in empirically supported treatments. J Clin Psychol 2012;68(1):8-23 [FREE Full text] [doi: 10.1002/jclp.20832] [Medline: 21901749]
27. Fairburn CG, Cooper Z. Therapist competence, therapy quality, and therapist training. Behav Res Ther 2011;49(6-7):373-378 [FREE Full text] [doi: 10.1016/j.brat.2011.03.005] [Medline: 21492829]
28. Jackson CB, Quetsch LB, Brabson LA, Herschell AD. Web-based training methods for behavioral health providers: a systematic review. Adm Policy Ment Health 2018;45(4):587-610 [FREE Full text] [doi: 10.1007/s10488-018-0847-0] [Medline: 29352459]
29. Khanna MS, Kendall PC. Bringing technology to training: web-based therapist training to promote the development of competent cognitive-behavioral therapists. Cognitive and Behavioral Practice 2015;22(3):291-301. [doi: 10.1016/j.cbpra.2015.02.002]
30. Seehagen S, Pflug V, Schneider S. Psychotherapy and science - harmony or dissonance? [Article in German]. Z Kinder Jugendpsychiatr Psychother 2012;40(5):301-306. [doi: 10.1024/1422-4917/a000186] [Medline: 22869223]
31. Mali D, Lim H. How do students perceive face-to-face/blended learning as a result of the COVID-19 pandemic? Int J Manag Educ 2021;19(3):100552. [doi: 10.1016/j.ijme.2021.100552]

XSL•FO
RenderX

32. Singh J, Steele K, Singh L. Combining the best of online and face-to-face learning: hybrid and blended learning approach for COVID-19, post vaccine, and post-pandemic world. J Educ Technol Syst 2021;50(2):1-32. [doi: 10.1177/00472395211047865]
33. Ilic D, Nordin RB, Glasziou P, Tilson JK, Villanueva E. A randomised controlled trial of a blended learning education intervention for teaching evidence-based medicine. BMC Med Educ 2015;15:39 [FREE Full text] [doi: 10.1186/s12909-015-0321-6] [Medline: 25884717]
34. Liu Q, Peng W, Zhang F, Hu R, Li Y, Yan W. The effectiveness of blended learning in health professions: systematic review and meta-analysis. J Med Internet Res 2016;18(1):e2 [FREE Full text] [doi: 10.2196/jmir.4807] [Medline: 26729058]
35. Lozano-Lozano M, Fernández-Lao C, Cantarero-Villanueva I, Noguerol I, Álvarez-Salvago F, Cruz-Fernández M, et al. A blended learning system to improve motivation, mood state, and satisfaction in undergraduate students: randomized controlled trial. J Med Internet Res 2020;22(5):e17101 [FREE Full text] [doi: 10.2196/17101] [Medline: 32441655]
36. Ma L, Lee CS. Evaluating the effectiveness of blended learning using the ARCS model. Computer Assisted Learning 2021;37(5):1397-1408. [doi: 10.1111/jcal.12579]
37. Beidas RS, Kendall PC. Training therapists in evidence-based practice: a critical review of studies from a systems-contextual perspective. Clin Psychol (New York) 2010;17(1):1-30. [doi: 10.1111/j.1468-2850.2009.01187.x] [Medline: 20877441]
38. Chu BC, Crocco ST, Arnold CC, Brown R, Southam-Gerow MA, Weisz JR. Sustained implementation of cognitive-behavioral therapy for youth anxiety and depression: long-term effects of structured training and consultation on therapist practice in the field. Prof Psychol Res Pr 2015;46(1):70-79 [FREE Full text] [doi: 10.1037/a0038000] [Medline: 26366037]
39. McCarty RJ, Cooke DL, Lazaroe LM, Guzick AG, Guastello AD, Budd SM, et al. The effects of an exposure therapy training program for pre-professionals in an intensive exposure-based summer camp. Cogn Behav Ther 2022;15:e5. [doi: 10.1017/s1754470x22000010]
40. Karabulut‐Ilgu A, Jaramillo Cherrez N, Jahren CT. A systematic review of research on the flipped learning method in engineering education. Brit J Educational Tech 2017;49(3):398-411. [doi: 10.1111/bjet.12548]
41. van der Kleij FM, Feskens RCW, Eggen TJHM. Effects of feedback in a computer-based learning environment on students' learning outcomes. Rev Educ Res 2015;85(4):475-511. [doi: 10.3102/0034654314564881]
42. Diep AN, Zhu C, Struyven K, Blieck Y. Who or what contributes to student satisfaction in different blended learning modalities? Brit J Educational Tech 2016;48(2):473-489. [doi: 10.1111/bjet.12431]
43. Margraf J, Cwik JC, von Brachel R, Suppiger A, Schneider S. DIPS Open Access 1.2: Diagnostic Interview for Mental Disorders. Bochum, Germany: Mental Health Research and Treament Center, Ruhr-Universität Bochum; 2021.
44. Schneider S, Pflug V, In-Albon T, Margraf J. Kinder-DIPS Open Access: Diagnostisches Interview bei psychischen Störungen im Kindes- und Jugendalter. Bochum, Germany: Forschungs- und Behandlungszentrum für psychische Gesundheit, Ruhr-Universität Bochum; 2017.
45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]
46. Hirschfeld G, Thielsch M. In: Glöckner-Rist A, editor. Münsteraner Fragebogen zur Evaluation von Seminaren - revidiert (MFE-Sr). Bonn, Germany: Zusammenstellung sozialwissenschaftlicher Items und Skalen ZIS Version 140; 2010.
47. Thielsch MT, Hirschfeld G. Facets of website content. Human–Computer Interaction 2018;34(4):279-327. [doi: 10.1080/07370024.2017.1421954]
48. Zumbach J, Spinath B, Schahn J, Friedrich M, Kögel M. Entwicklung einer Kurzskala zur Lehrevaluation. Development of a short scale for teaching evaluation. In: Psychologiedidaktik und Evaluation VI. Psychological Didactics and Evaluation VI. Göttingen, Germany: V & R Unipress; 2007:317-325.
49. Lewis JR, Utesch BS, Maher DE. UMUX-LITE - When there's no time for the SUS. 2013 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 27, 2013; Paris, France. [doi: 10.1145/2470654.2481287]
50. Moshagen M, Thielsch M. A short version of the visual aesthetics of websites inventory. Behav Inf Technol 2013;32(12):1305-1311. [doi: 10.1080/0144929x.2012.694910]
51. Cuzick J, Sasieni P. Interpreting the results of noninferiority trials-a review. Br J Cancer 2022;127(10):1755-1759 [FREE Full text] [doi: 10.1038/s41416-022-01937-w] [Medline: 36104512]
52. Non-inferiority clinical trials to establish effectiveness: guidance for industry. U.S. Food and Drug Administration. 2016. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/non-inferiority-clinical-trials [accessed 2024-08-13]
53. Choice of a non-inferiority margin - scientific guideline. European Medicines Agency. 2005. URL: https://www.ema.europa.eu/en/choice-non-inferiority-margin-scientific-guideline [accessed 2024-08-13]
54. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Stat Med 2003;22(2):169-186. [doi: 10.1002/sim.1425] [Medline: 12520555]
55. Rief W, Hofmann SG. Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example. Psychol Med 2018;48(8):1392-1394. [doi: 10.1017/S0033291718000247] [Medline: 29439745]
56. Laster LL, Johnson MF, Kotler ML. Non-inferiority trials: the 'at least as good as' criterion with dichotomous data. Stat Med 2006;25(7):1115-1130. [doi: 10.1002/sim.2476] [Medline: 16381070]

XSL•FO
RenderX

57. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
58. Bonnin G, Kröber S, Schneider S, Margraf J, Pflug V, Gerlach A, et al. Dataset for: Blended learning for diagnostic skills: a multicenter cluster randomized non-inferiority trial. PsychArchives. 2023. URL: https://psycharchives.org/en/item/5cecec0f-b732-4012-a1e3-661a3e3c5182 [accessed 2024-11-19]
59. Bonnin G, Kröber S, Schneider S, Margraf J, Pflug V, Gerlach AL, et al. Code for: Blended learning for diagnostic skills: a multicenter cluster randomized non-inferiority trial. PsychArchives 2023:1. [doi: 10.23668/psycharchives.12494]
60. Suppiger A, In-Albon T, Hendriksen S, Hermann E, Margraf J, Schneider S. Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. Behav Ther 2009;40(3):272-279. [doi: 10.1016/j.beth.2008.07.002] [Medline: 19647528]
61. van Leeuwen A, Bos N, van Ravenswaaij H, van Oostenrijk J. The role of temporal patterns in students' behavior for predicting course performance: a comparison of two blended learning courses. Brit J Educational Tech 2018;50(2):921-933. [doi: 10.1111/bjet.12616]
62. Chen NS, Ko HC, Kinshuk *, Lin T. A model for synchronous learning using the Internet. Innov Educ Teach Int 2006;42(2):181-194. [doi: 10.1080/14703290500062599]
63. Alnabelsi T, Al-Hussaini A, Owens D. Comparison of traditional face-to-face teaching with synchronous e-learning in otolaryngology emergencies teaching to medical undergraduates: a randomised controlled trial. Eur Arch Otorhinolaryngol 2015;272(3):759-763. [doi: 10.1007/s00405-014-3326-6] [Medline: 25308244]
64. Cook DA. Web-based learning: pros, cons and controversies. Clin Med (Lond) 2007;7(1):37-42 [FREE Full text] [doi: 10.7861/clinmedicine.7-1-37] [Medline: 17348573]
65. Cook DA, Dupras DM, Thompson WG, Pankratz VS. Web-based learning in residents' continuity clinics: a randomized, controlled trial. Acad Med 2005;80(1):90-97. [doi: 10.1097/00001888-200501000-00022] [Medline: 15618102]
66. Porter WW, Graham CR. Institutional drivers and barriers to faculty adoption of blended learning in higher education. Brit J Educational Tech 2015;47(4):748-762. [doi: 10.1111/bjet.12269]
67. Cooper Z, Bailey-Straebler S, Morgan KE, O'Connor ME, Caddy C, Hamadi L, et al. Using the internet to train therapists: randomized comparison of two scalable methods. J Med Internet Res 2017;19(10):e355 [FREE Full text] [doi: 10.2196/jmir.8336] [Medline: 29046265]

Abbreviations

DIPS-OA1.2: Diagnostic Interview for Mental Disorders – Open Access 1.2
(Kinder-)DIPS-OA: Diagnostic Interview for Mental Disorders (in Children and Adolescents) – Open Access
OR: odds ratio

Edited by T de Azevedo Cardoso; submitted 31.10.23; peer-reviewed by A Ramaprasad, B Pratumvinit, A Segev; comments to author 28.02.24; revised version received 05.04.24; accepted 23.09.24; published 27.11.24.

Please cite as:
Bonnin G, Kröber S, Schneider S, Margraf J, Pflug V, Gerlach AL, Slotta T, Christiansen H, Albrecht B, Chavanon ML, Hirschfeld G, In-Albon T, Thielsch MT, von Brachel R
A Blended Learning Course on the Diagnostics of Mental Disorders: Multicenter Cluster Randomized Noninferiority Trial
J Med Internet Res 2024;26:e54176
URL: https://www.jmir.org/2024/1/e54176
doi:10.2196/54176
PMID:

©Gabriel Bonnin, Svea Kröber, Silvia Schneider, Jürgen Margraf, Verena Pflug, Alexander L Gerlach, Timo Slotta, Hanna Christiansen, Björn Albrecht, Mira-Lynn Chavanon, Gerrit Hirschfeld, Tina In-Albon, Meinald T Thielsch, Ruth von Brachel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Virtual Simulations Versus Mannequins and Real Persons in Medical and Nursing Education: Meta-Analysis and Trial Sequential Analysis of Randomized Controlled Trials

Nan Jiang[1*], MD; Yuelun Zhang[2*], PhD; Siyu Liang[3], MD; Xiaohong Lyu[4], MD; Shi Chen[3], MD; Xiaoming Huang[5], MD; Hui Pan[3], MD

[1]4 + 4 Medical Doctor Program, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Medical Research Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[3]Department of Endocrinology, Key Laboratory of Endocrinology of National Health Commission, Translation Medicine Centre, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China

[4]Department of Breast Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[5]Department of Family Medicine & Division of General Internal Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Shi Chen, MD
Department of Endocrinology
Key Laboratory of Endocrinology of National Health Commission
Translation Medicine Centre, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences
No.1 Shuaifuyuan,Wangfujing, Dongcheng District
Beijing, 100730
China
Phone: 86 13683136205
Email: cspumch@163.com

Abstract

Background: Virtual simulation (VS) is a developing education approach with the recreation of reality using digital technology. The teaching effectiveness of VSs compared to mannequins and real persons (RPs) has never been investigated in medical and nursing education.

Objective: This study aims to compare VSs and mannequins or RPs in improving the following clinical competencies: knowledge, procedural skills, clinical reasoning, and communication skills.

Methods: Following Cochrane methodology, a meta-analysis was conducted on the effectiveness of VSs in pre- and postregistration medical or nursing participants. The Cochrane Library, PubMed, Embase, and Educational Resource Information Centre databases were searched to identify English-written randomized controlled trials up to August 2024. Two authors independently selected studies, extracted data, and assessed the risk of bias. All pooled estimates were based on random-effects models and assessed by trial sequential analyses. Leave-one-out, subgroup, and univariate meta-regression analyses were performed to explore sources of heterogeneity.

Results: A total of 27 studies with 1480 participants were included. Overall, there were no significant differences between VSs and mannequins or RPs in improving knowledge (standard mean difference [SMD]=0.08; 95% CI –0.30 to 0.47; I^2=67%; P=.002), procedural skills (SMD=–0.12; 95% CI –0.47 to 0.23; I^2=75%; P<.001), clinical reasoning (SMD=0.29; 95% CI –0.26 to 0.85; I^2=88%; P<.001), and communication skills (SMD=–0.02; 95% CI: –0.62 to 0.58; I^2=86%; P<.001). Trial sequential analysis for clinical reasoning indicated an insufficient sample size for a definitive judgment. For procedural skills, subgroup analyses showed that VSs were less effective among nursing participants (SMD=–0.55; 95% CI –1.07 to –0.03; I^2=69%; P=.04). Univariate meta-regression detected a positive effect of publication year (β=.09; P=.02) on communication skill scores.

XSL•FO
RenderX

Conclusions: Given favorable cost-utility plus high flexibility regarding time and space, VSs are viable alternatives to traditional face-to-face learning modalities. The comparative effectiveness of VSs deserves to be followed up with the emergence of new technology. In addition, further investigation of VSs with different design features will provide novel insights to drive education reform.

Trial Registration: PROSPERO CRD42023466622; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=466622

(J Med Internet Res 2024;26:e56195) doi:10.2196/56195

KEYWORDS

artificial intelligence; clinical virtual simulation; medical education; meta-analysis; nursing education; virtual patient; virtual reality

Introduction

The ultimate goal of health profession education is to promote the transfer of theoretical knowledge into clinical practice. Studies in cognitive psychology suggest that recall and application of information are best in learning environments similar to workplace [1]. Clinical simulation, a technique that replicates real experiences with guided experiences, has thereby gained popularity in the last 2 decades and is currently the main tool for training health care professionals [2]. There are five main categories of simulation: (1) low-tech simulators (ie, models or mannequins); (2) standardized patients (SPs); (3) screen-based computer programs; (4) high-fidelity computer simulators integrated with visual, audio, and touch cues; and (5) computer-driven, full-length mannequins with realistic anatomy and physiology [3]. With the help of simulation, health care professionals refine the knowledge, skills, and attitudes needed to deliver quality patient care while patients are protected from unnecessary risk [3]. In medical and nursing education, simulation-based teaching has demonstrated superior effectiveness compared to didactic teaching [4,5]; moreover, it reduces anxiety and increases the confidence of students entering practice [6,7]. While simulation-based education may bring downstream benefits of reducing medical errors and improving patient care, the current level of evidence is still low due to a lack of high-quality studies with patient-centered outcomes [8-12].

Driven by COVID-19 social distancing and modern technological advancement, there has been a marked shift toward learning on digital platforms or software. Virtual simulation (VS) refers to the 2D or 3D recreation of reality by digital technology, where students are able to interact with digital scenes, instruments, and characters [13,14]. Compared to conventional simulations (eg, mannequins and SPs), VSs can be undertaken flexibly with no limit of time and space [15]; they can provide a more realistic experience with the aid of artificial intelligence (AI) and virtual reality (VR) when certain pathological findings cannot be readily expressed [16-18]. For educators, VSs can potentially save costs due to reduced instructor time, manpower, and space resources [19]. One recent analysis found that VR simulations required 22% less time and 40% lower cost than traditional high-fidelity simulations to achieve the same learning outcomes [20]. Documentation and evaluation of student performance can also be automated with digital technologies [19].

Despite these promising advantages, the comparative effectiveness of training health professionals using mannequins or real persons (RPs; eg, SPs, role-play, actual patients) and VSs remains uncertain. Previous meta-analyses have shown that compared to traditional education (eg, lectures, reading exercises, in-class group discussions, mannequins, SPs), digital education is at least as effective in training medical students' communication skills [15]; virtual patients (VPs) can more effectively improve skills and can be at least as effective in enhancing knowledge [21]; VR can more effectively improve not only skills but also knowledge [22]. The comparison with an extensive category of "traditional education" is insufficient to determine the value and viability of implementing VS as substitutes for mannequins and RPs, so specific recommendations for educational reform cannot be made. Thus, this study aims to investigate the teaching efficacy of VSs versus mannequins and RPs in a fine-grain spectrum of clinical competencies, including knowledge, procedural skills, clinical reasoning, and communication skills.

Methods

Reporting Guidelines

This study follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines [23]. Multimedia Appendix 1 provides a PRISMA checklist of this meta-analysis. The protocol was registered in PROSPERO (CRD42023466622).

Eligibility Criteria

This study included randomized controlled trials (RCTs) from inception to August 2024. Inclusion and exclusion criteria are listed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Year: From inception to August 2024
- Study design: Randomized controlled trials
- Language: English
- Pre- and post-registration medical/nursing students and staff
- Intervention: Virtual simulations
- Comparison: Mannequins or real persons
- Outcome: Objective posttest ratings for knowledge, procedural skills, clinical reasoning, or communication skills
- Data availability: Yes

Exclusion criteria

- Intervention: Blended simulations where virtual simulations and mannequins or real persons were both applied
- Outcome: Self-evaluations

The four outcomes of interest were defined as follows: (1) knowledge: remembering and understanding basic concepts, measured by question-based theoretical tests [24]; (2) procedural skills: following a series of steps to accomplish certain tasks, demonstrated by simulation exams [24]; (3) clinical reasoning: evaluating and reacting to clinical situations, demonstrated by simulation examinations [24]; and (4) communication skills: Interacting with patients or colleagues verbally and nonverbally, demonstrated by simulation examinations [24].

Search Strategy

A thorough search was carried out in the Cochrane Library, PubMed, Embase, and Educational Resource Information Centre databases by specifying the desired population, intervention, and comparison. Search terms were selected judiciously, with specific terms customized for each database (Multimedia Appendix 2). Additionally, we searched Google and reference lists of the selected studies to retrieve other relevant publications.

Search records were imported into the EndNote library (version X9; Clarivate). After eliminating duplicates, the remaining studies underwent eligibility screening by 2 independent reviewers (NJ and SYL) according to predefined inclusion and exclusion criteria (Textbox 1). The initial screening process involved the assessment of titles and abstracts for relevance. Subsequently, full-text screening was performed, and the rationale for exclusion was documented in the PRISMA flowchart. In cases of discrepancies between the 2 reviewers, a third reviewer (SC) was consulted to reach a consensus. Study authors were contacted for crucial missing information. Finally, all researchers agreed on the final literature to be included in the analysis.

Data Extraction

Two reviewers (NJ and SYL) independently extracted data in a structured form in Microsoft Excel, including author, publication date, country, participants, sample size, type of intervention, type of comparison, and outcomes. If a study had multiple posttest measurements, only the first measurement was recorded due to concern about the learning effect. If participants were rated by both SPs and independent evaluators, the ratings of the latter were preferred. If an outcome was assessed by multiple rating scales, the scale with the highest intraclass correlation coefficient was chosen. If an outcome was reported as multiple items instead of a total score, the primary item was selected, and if that was impossible, the mean score of all items was calculated. If a study had both mannequin and RP as control groups, each pairwise comparison was included separately while splitting the shared intervention group approximately evenly among the comparisons. The outcomes were recorded as mean values and SDs. For studies that reported outcomes as median values and IQRs or as mean values and CIs, these measures were converted to mean values and SDs using methods described in previous literature [25,26].

Risk of Bias Assessment

Two reviewers (NJ and SYL) independently assessed the risk of bias of the included studies using the Cochrane tool [25]. The following domains were considered: random sequence generation, allocation sequence concealment, blinding of participants or personnel, blinding of outcome assessment, complete outcome data, selective reporting, and other sources of bias. Discrepancies were resolved by discussion with a third author. Studies were not excluded from data extraction or analysis based on bias assessment scores.

Statistical Analysis

Since data were measured using different tools, the mean differences were recalculated into standardized mean differences (SMDs). The results are displayed in forest plots, with pooled SMDs computed using random-effects meta-analysis models. We interpreted the effect size as small (SMD=0.2), moderate (SMD=0.5), or large (SMD=0.8) [25]. Publication bias was assessed by the Egger regression asymmetry test.

Cochran Q (chi-square test) was used to evaluate statistical heterogeneity, with a statistical significance threshold at $P<.10$. I^2 statistic was adopted to quantify the degree of heterogeneity, where I^2 was categorized as unimportant (0%-40%), moderate

XSL•FO
RenderX

(30%-60%), substantial (50%-90%), and considerable (75%-100%) [25]. To investigate possible sources of heterogeneity, the leave-one-out method was first applied. Next, for each outcome, we conducted subgroup analyses by discipline (medicine; nursing), level (undergraduate; nonundergraduate [graduate students or clinical staff]), and comparison (mannequin; RP) to see if I^2<50% in both groups [27]. Univariate random-effects meta-regression analyses were then performed to investigate whether heterogeneity between trials could be attributed to year of publication, age of participants, discipline, level, and comparison. Multivariate meta-regression was not performed due to the limited number of studies and the risk of overfitting.

Trial sequential analysis (TSA) was used to evaluate the strength of evidence and adjusted for potential errors. The analysis set specific values for type 1 and 2 errors (5% and 20%) and used these values to calculate trial sequential monitoring boundaries, futility boundaries, and the required information size (RIS) [28]. The mean difference to generate RIS was set to detect a mean difference of 2.0 between intervention and comparison. The variance was estimated by pooling all included trials. Heterogeneity was adjusted based on the estimated ratio between the variance in the random-effects model and the variance in the fixed-effects model. Finally, a graphical evaluation was used to determine if the cumulative *Z* curve met defined thresholds.

All data analyses were performed using package meta of R software (version 4.2.3; R Foundation for Statistical Computing) [29] and TSA software (version 0.9.5.10 beta; Copenhagen Trial Unit). Two-sided *P* values<.05 were considered to be statistically significant.

Results

Included Studies

After the selection process, a total of 27 studies with 1480 participants were included (Figure 1). In terms of discipline, participants in 14 studies were from the field of medicine [16,18,30-41]; participants in 11 studies were from the field of nursing [42-52]; and the remaining 2 studies enrolled mixed participants from medicine and nursing [53,54]. In terms of level, participants in 14 studies were undergraduate students [16,31,35,39,42-46,50-54]; participants in 8 studies were graduate students or clinical staff [18,32-34,36-38,40,41,49]; and the remaining 2 studies enrolled mixed undergraduate and nonundergraduate participants [30,48]. In terms of comparison, 13 studies compared VSs and mannequins [33-36,38,40,44-51]; 10 studies compared VSs and RPs [16,18,30,31,37,39,41-43,52-54]; and the remaining study included mannequins and RPs as comparison groups simultaneously [32]. Detailed information and extracted data for each study are summarized in Multimedia Appendix 3 [16,18,30-54].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) flow diagram.

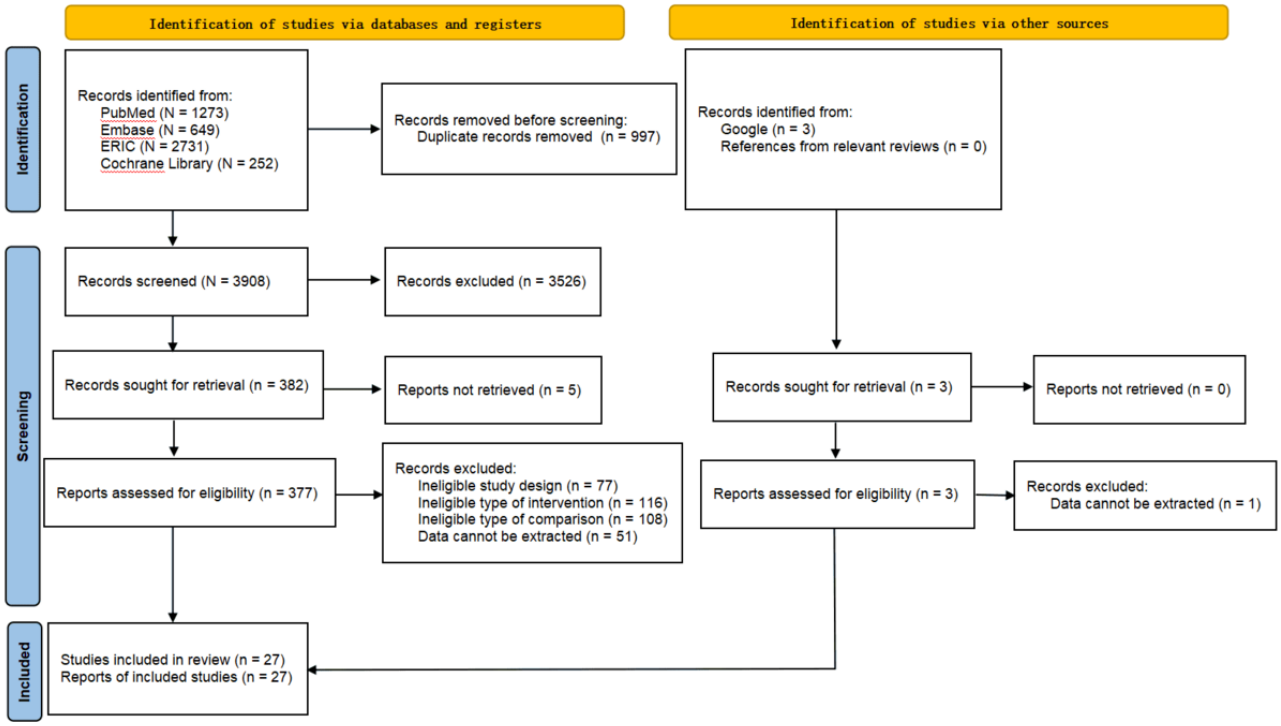

Risk of Bias

Following Cochrane guidance [25], the risk of bias was assessed for all the included studies. The results are summarized in Figure 2 [16,18,30-54]. Out of 27 studies, 15 studies described an adequate random sequence generation method; 11 studies did not provide a clear description; and the remaining 1 study allocated participants to intervention and comparison arms based on the order in which they volunteered. Allocation concealment was not explicitly mentioned in most studies except 2. Blinding participants was impossible to avoid in this type of research but should not raise a major concern. Four studies had high-risk performance bias since researchers were not blinded and might give biased ratings; 12 studies were low-risk due to blinding of

XSL•FO
RenderX

personnel; the others did not provide a clear description. The risk of detection bias was rated as unclear in most studies except 4, which clearly stated that statistical analysts were blinded. Most studies had low-risk attrition bias because no participants were lost, while participant dropout in 6 studies had an unclear impact on outcome assessment. Due to the absence of protocols, most studies had unclear reporting bias; 7 studies with protocols provided were rated as low risk. Given the difficulty of evaluating other biases (eg, volunteer bias), the risk was judged to be unclear for all studies.

Figure 2. Risk of bias summary [16,18,30-54]. +: low risk of bias; –: high risk of bias; ?: unclear risk of bias.

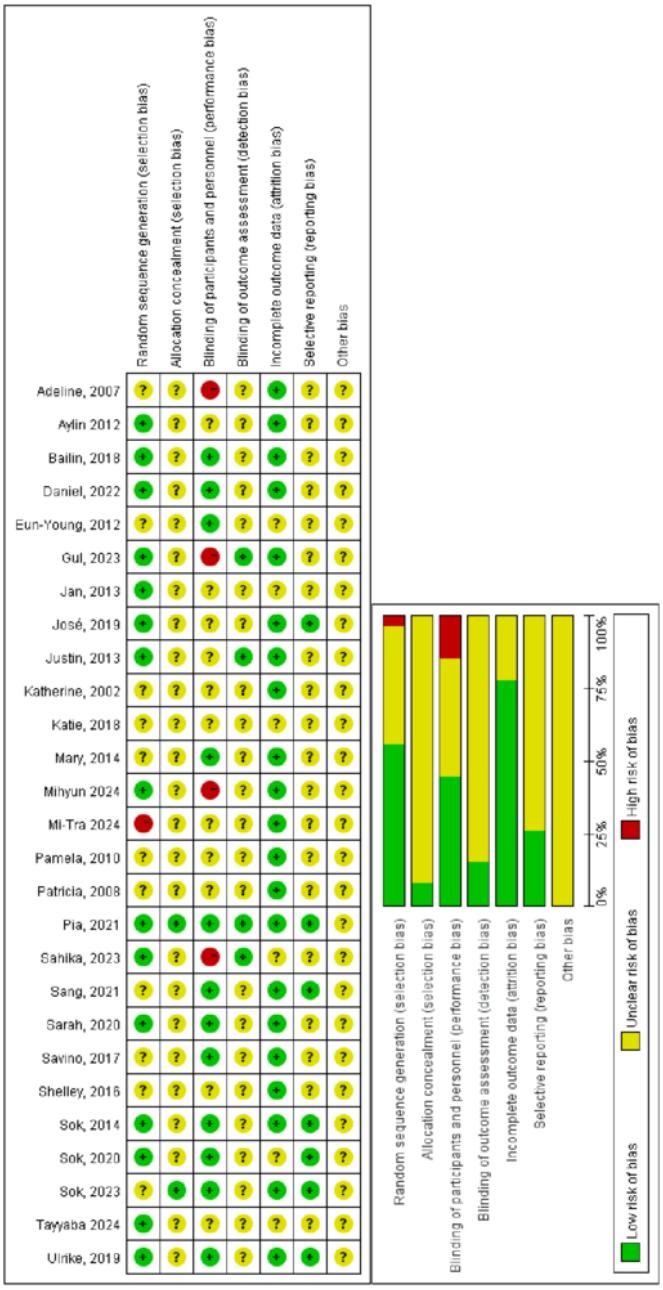

Effects of Interventions

Knowledge

A total of 8 studies assessed the outcome of knowledge [32,35,37,41,44-46,50]. The Egger test showed no statistically significant publication bias (*P*=.73). In Figure 3 [32,35,37,41,44-46,50], the pooled effect size did not reflect a significant difference between VSs and mannequins or RPs on the knowledge outcome (SMD=0.08; 95% CI –0.30 to 0.47; I^2=67%; *P*=.002). TSA showed that the "inner wedge" area had been reached, indicating strong evidence that further studies would hardly be able to change the statistically insignificant result (Figure S1 in Multimedia Appendix 4). This result was consistent according to the leave-one-out analysis (Figure S2 in Multimedia Appendix 4 [32,35,37,41,44-46,50]).

XSL•FO
RenderX

Figure 3. Forest plot of virtual simulations versus mannequins or real persons for knowledge [32,35,37,41,44-46,50].

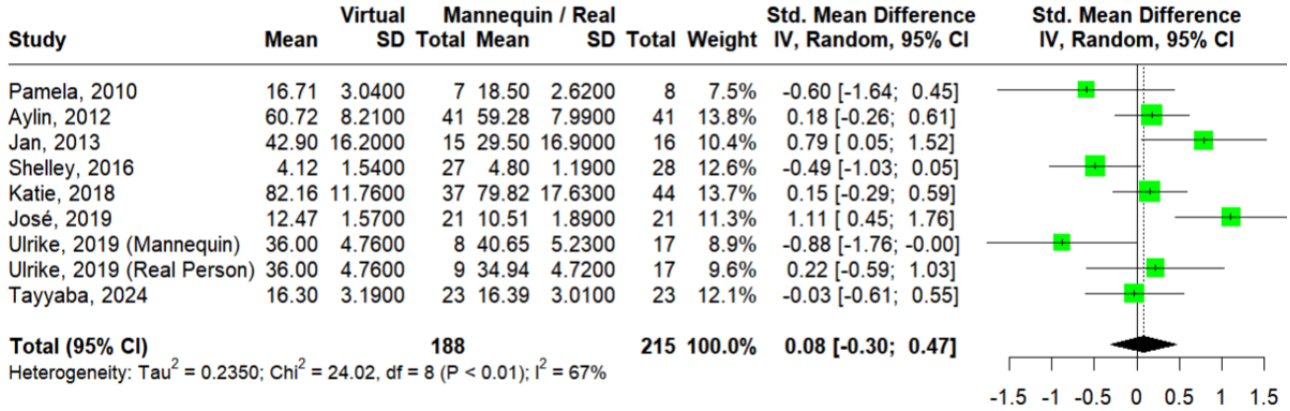

For subgroup analysis by discipline (Figure S3a in Multimedia Appendix 4 [32,35,37,41,44-46,50]), 4 studies [32,35,37,41] and 4 studies [44-46,50] enrolled participants from medicine and nursing, respectively. Neither of these groups showed a significant difference between VSs and mannequins or RPs (SMD=–0.05; 95% CI –0.62 to 0.51; I^2=59%; P=.04 for medicine and SMD=0.22; 95% CI –0.39 to 0.82; I^2=78%; P<.001 for nursing), with no significant subgroup difference at P=.53.

For subgroup analysis by level (Figure S3b in Multimedia Appendix 4 [32,35,37,41,44-46,50]), 5 studies [35,44-46,50] and 3 studies [32,37,41] enrolled undergraduate and nonundergraduate participants, respectively. Neither of these groups showed a significant difference between VSs and mannequins or RPs (SMD=0.31; 95% CI –0.21 to 0.84; I^2=75%; P<.001 for undergraduate and SMD=–0.25; 95% CI –0.71 to 0.22; I^2=29%; P=.24 for nonundergraduate), with no significant subgroup difference at P=.12.

For subgroup analysis by comparison (Figure S3c in Multimedia Appendix 4 [32,35,37,41,44-46,50]), 2 studies [37,41] and 5 studies [35,44-46,50] included RP and mannequin arms, respectively; Weber et al [32] compared web-based simulators to both mannequins and RPs. Neither mannequins nor RPs was significantly different from VSs (SMD=0.15; 95% CI –0.40 to 0.71; I^2=77%; P<.001 for mannequin and SMD=–0.06; 95% CI –0.48 to 0.37; I^2=0%; P=.48 for RP), with no significant subgroup difference at P=.56.

Univariate meta-regression revealed that year of publication (β=.01; P=.92), age of participants (β=−.05; P=.29), nursing discipline (β=.27; P=.52), undergraduate level (β=.58; P=.14), and real person comparison (β=−.25; P=.58) had no effects on knowledge scores (Table S1 in Multimedia Appendix 4).

Procedural Skills

A total of 10 studies assessed the outcome of procedural skills [16,18,32-34,36,37,42,49,51]. The Egger test showed no statistically significant publication bias (P=.40). In Figure 4 [16,18,32-34,36,37,42,49,51], the pooled effect size did not reflect a significant difference between VSs and mannequins or RPs on the outcome of procedural skills (SMD=–0.12; 95% CI –0.47 to 0.23; I^2=75%; P<.001). TSA showed that the RIS had been reached, implying a conclusive statistically insignificant result (Figure S4 in Multimedia Appendix 4). This result was consistent according to the leave-one-out analysis (Figure S5 in Multimedia Appendix 4 [16,18,32-34,36,37,42,49,51]).

Figure 4. Forest plot of virtual simulations versus mannequins or real persons for procedural skills [16,18,32-34,36,37,42,49,51].

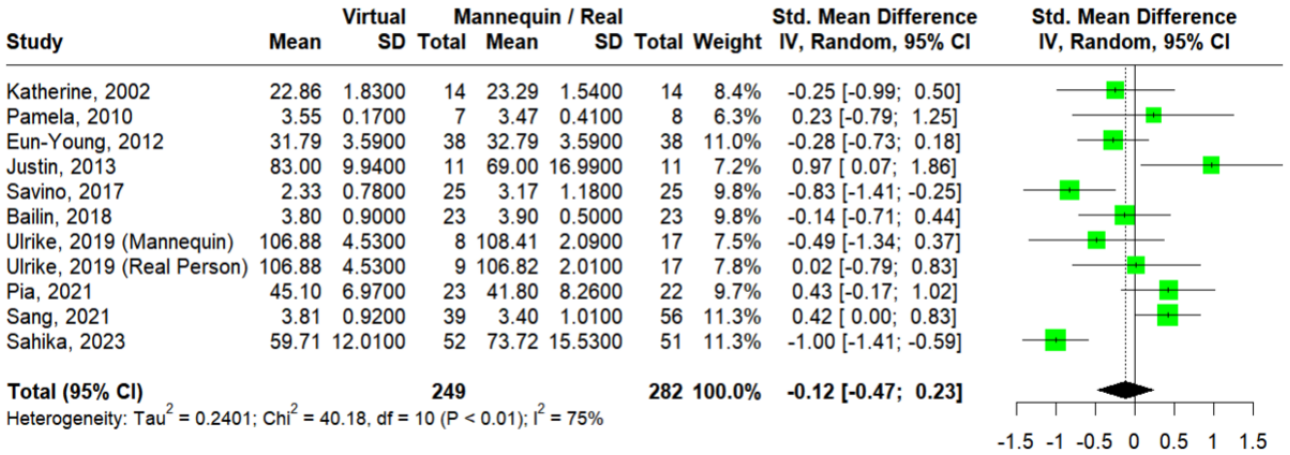

For subgroup analysis by discipline (Figure S6a in Multimedia Appendix 4 [16,18,32-34,37,42,49,51]), 7 studies [16,18,32-34,36,37] and 3 studies [42,49,51] enrolled participants from medicine and nursing respectively. Only the nursing group showed a significant difference between VSs and mannequins or RPs (SMD=0.06; 95% CI –0.33 to 0.46; I^2=64%;

XSL•FO
RenderX

P<.001 for medicine and SMD=–0.55; 95% CI –1.07 to –0.03, I^2=69%; *P*=.04 for nursing), with no significant subgroup difference at *P*=.07.

For subgroup analysis by level (Figure S6b in Multimedia Appendix 4 [16,18,32-34,36,37,42,49,51]), 3 studies [16,42,51] and 7 studies [18,32-34,36,37,49] enrolled undergraduate and nonundergraduate participants respectively. Neither of these groups showed a significant difference between VSs and mannequins or RPs (SMD=–0.29; 95% CI –1.10 to 0.52; I^2=91%; *P*<.001 for undergraduate and SMD=–0.04; 95% CI –0.43 to 0.35; I^2=56%; *P*=.03 for nonundergraduate), with no significant subgroup difference at *P*=.59.

For subgroup analysis by comparison (Figure S6c in Multimedia Appendix 4 [16,18,32-34,36,37,42,49,51]), 5 studies [33,34,36,49,51] and 4 studies [16,18,37,42] included mannequin and RP arms, respectively; Weber et al [32] compared web-based simulators to both mannequins and RPs. Neither mannequins nor RPs was significantly different from VSs (SMD=–0.22; 95% CI –0.63 to 0.19; I^2=56%; *P*=.04 for mannequins and SMD=–0.00; 95% CI –0.60 to 0.59; I^2=86%; *P*<.001 for RPs), with no significant subgroup difference at *P*=.57.

Univariate meta-regression revealed that year of publication (β=–.01; *P*=.71), age of participants (β=.05; *P*=.33), nursing discipline (β=–.60; *P*=.08), undergraduate level (β=–.25; *P*=.52), and real person comparison (β=.19; *P*=.61) had no effects on procedural skill scores (Table S1 in Multimedia Appendix 4).

Clinical Reasoning

A total of 9 studies assessed the outcome of clinical reasoning [38,40,43,45,47,48,50,52,53]. The Egger test showed no statistically significant publication bias (*P*=.88). In Figure 5 [38,40,43,45,47,48,50,52,53], the pooled effect size did not reflect a significant difference between VSs and mannequins or RPs on the outcome of procedural skills (SMD=0.29; 95% CI –0.26 to 0.85; I^2=88%; *P*<.01). TSA showed that the RIS had not been reached, implying further studies would be needed to verify the statistically insignificant result (Figure S7 in Multimedia Appendix 4). This result was consistent according to the leave-one-out analysis (Figure S8 in Multimedia Appendix 4).

Figure 5. Forest plot of virtual simulations versus mannequins or real persons for clinical reasoning [38,40,43,45,47,48,50,52,53].

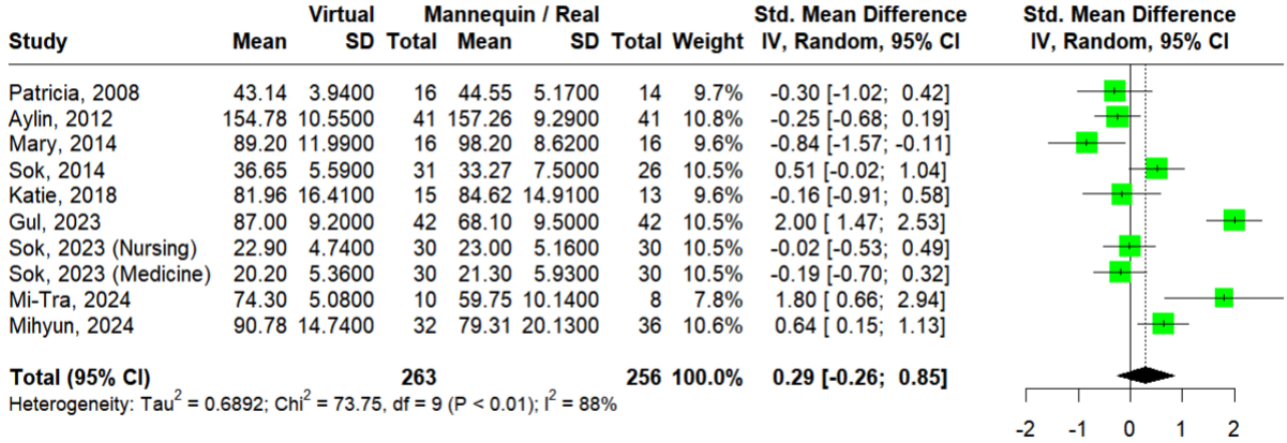

For subgroup analysis by discipline (Figure S9a in Multimedia Appendix 4 [38,40,43,45,47,48,50,52,53]), 2 studies [38,40] and 6 studies [43,45,47,48,50,52] enrolled participants from medicine and nursing respectively. Liaw et al [53] enrolled participants from both fields. Neither of these groups showed a significant difference between VSs and mannequins or RPs (SMD=0.35; 95% CI –0.89 to 1.60; I^2=81%; *P*<.001 for medicine and SMD=0.28; 95% CI –0.39 to 0.95; I^2=90%; *P*<.001 for nursing), with no significant subgroup difference at *P*=.92.

For subgroup analysis by level (Figure S9b in Multimedia Appendix 4 [38,40,43,45,47,48,50,52,53]), 6 studies [43,45,47,50,52,53] and 2 studies [38,40] enrolled undergraduate and nonundergraduate participants, respectively; the study by Johnson et al [48] was excluded due to a mixed enrollment. Neither of these groups showed a significant difference between VSs and mannequins or RPs (SMD=0.36; 95% CI –0.23 to 0.96; I^2=89%; *P*<.001 for undergraduate and SMD=0.70; 95% CI –1.35 to 2.75; I^2=89%; *P*<.001 for nonundergraduate), with no significant subgroup difference at *P*=.76.

For subgroup analysis by comparison (Figure S9c in Multimedia Appendix 4 [38,40,43,45,47,48,50,52,53]), 6 studies [38,40,45,47,48,50] and 3 studies [43,52,53] included mannequin and RP arms, respectively. Neither mannequins nor RPs was significantly different from VSs (SMD=0.06; 95% CI –0.57 to 0.68; I^2=75%; *P*<.001 for mannequins and SMD=0.60; 95% CI –0.37 to 1.58; I^2=93%; *P*<.001 for RPs), with no significant subgroup difference at *P*=.35.

Univariate meta-regression revealed that year of publication (β=.08, *P*=.06), age of participants (β=–.02; *P*=.81), nursing discipline (β=–.05; *P*=.94), undergraduate level (β=–.27; *P*=.72), and real person comparison (β=.53; *P*=.35) had no effects on clinical reasoning scores (Table S1 in Multimedia Appendix 4).

Communication Skills

A total of 5 studies assessed the outcome of communication skills [30,31,39,52,54]. The Egger test showed no statistically significant publication bias (*P*=.55). In Figure 6 [30,31,39,52,54], the pooled effect size did not reflect a

XSL•FO
RenderX

significant difference between VSs and mannequins or RPs on the outcome of procedural skills (SMD=–0.02; 95% CI –0.62 to 0.58; I^2=86%; P<.01). TSA showed that the "inner wedge" area had been reached, indicating strong evidence that further studies would hardly be able to change the statistically insignificant result (Figure S10 in Multimedia Appendix 4). This result was consistent according to the leave-one-out analysis (Figure S11 in Multimedia Appendix 4 [30,31,39,52,54]).

Figure 6. Forest plot of virtual simulations versus mannequins or real persons for communication skills [30,31,39,52,54].

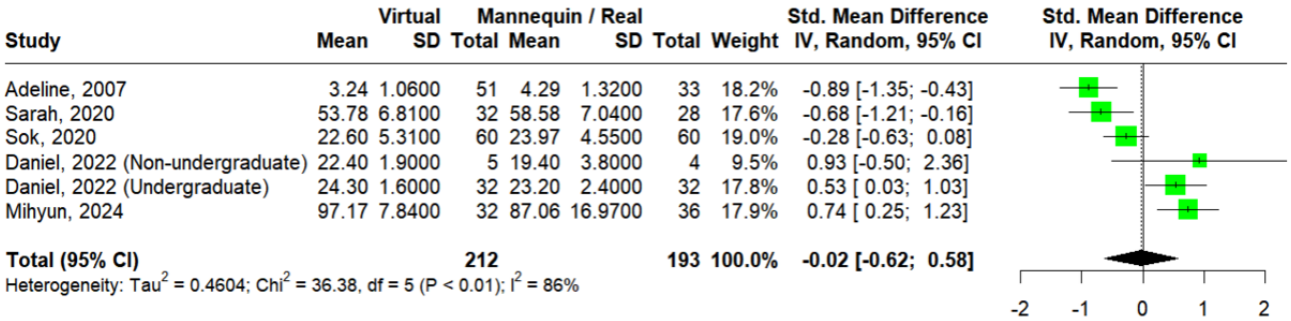

Since all 5 studies compared VSs to RPs, only subgroup analyses by discipline and level could be conducted. For subgroup analysis by discipline (Figure S12a in Multimedia Appendix 4 [30,31,39,52,54]), 3 studies [30,31,39] and 1 study [52] enrolled participants from medicine and nursing, respectively; and the study by Liaw et al [54] was excluded due to a mixed enrollment. Only the nursing group showed a significant difference between VSs and mannequins or RPs (SMD=–0.14; 95% CI –0.97 to 0.69; I^2=86%; P<.001 for medicine and SMD=0.74; 95% CI 0.25-1.23 for nursing), with no significant subgroup difference at P=.07.

For subgroup analysis by level (Figure S12b in Multimedia Appendix 4 [30,31,39,52,54]), 4 studies enrolled undergraduate participants [31,39,52,54] and the study by Sapkaroski et al [30] enrolled both undergraduate and nonundergraduate participants. Neither of these groups showed a significant difference between VSs and mannequins or RPs (SMD=–0.12; 95% CI –0.75 to 0.51; I^2=88%; P<.001 for undergraduate and SMD=0.93; 95% CI –0.50 to 2.36 for nonundergraduate), with no significant subgroup difference at P=.19.

Univariate meta-regression revealed that year of publication (β=.09; P=.02) had a positive effect on communication skill scores, whereas age of participants (β=–.07; P=.84), nursing discipline (β=.88; P=.33), undergraduate level (β=–1.05; P=.32), and real person comparison (β=.19; P=.61) had no effects (Table S1 in Multimedia Appendix 4). Since no studies compared VSs to mannequins, the effect of real person comparison could not be evaluated.

Discussion

Principal Findings

The aim of this meta-analysis was to evaluate the effectiveness of VSs in comparison with mannequins and RPs in medical and nursing education. We found that VSs and conventional simulations were not statistically different in teaching knowledge, procedural skills, clinical reasoning, and communication skills. Among nursing participants, VSs were more effective in improving procedural skills. For the training of communication skills, VSs tended to be increasingly effective with time.

This meta-analysis followed the original protocol of PROSPERO (CRD42023466622). TSA and meta-regression were added to further evaluate the accumulative evidence and identify potential sources of heterogeneity. We believe that these analyses do not negatively affect the reliability of our results; instead, they offer novel insights and strengthen the existing content. Here, we state the aforementioned deviations to ensure our study's transparency.

Comparison With Prior Work

There was no significant difference in knowledge gained between VSs and mannequins or RPs. This result aligns with a previous meta-analysis comparing 2D VSs with traditional education methods [21], while other analyses suggested that 3D VSs were more effective [22,55-57]. In our meta-analysis, almost all studies adopted 3D VSs, but the comparison was specifically narrowed down to conventional simulations. Our pooled effect size tended to favor VSs. Indeed, it can be beneficial to use VSs for knowledge delivery. First, digital platforms or software provide repeatable, information-rich feedback for learning [58]. Second, immersive technology (ie, devices that provide a sense of realism and immersion in the computer-generated world through sensory stimuli) such as VR promotes students' interest with enhanced satisfaction, self-efficacy, and engagement [59]. Although VSs are viable alternatives to mannequins and RPs, undesirable usability may hinder the learning process. For instance, Cobbett and Snelgrove-Clarke [46] reported in their study that nearly half of the students expressed dislike for the VS due to technological issues such as the "online program was slow" and "didn't know where to find things." Consequently, the development of VSs should consider features that enhance ease of use and users' level of satisfaction (eg, step-by-step instructions, user-friendly interfaces) so that students allocate their time appropriately on learning instead of wrestling with the platform or software.

There was no significant difference in improving procedural skills between VSs and mannequins or RPs, while our pooled effect size tended to favor mannequins or RPs. Subgroup analyses showed that VSs were less effective among nursing

XSL•FO
RenderX

participants. These results were inconsistent with previous meta-analyses comparing VSs with traditional education methods [21,22,57]. Such discrepancies may be due to different definitions of comparison. While VSs are advantageous for practicing and mastering skills relative to didactic teaching, this is not necessarily true if they are compared with mannequins and RPs. A gap exists between digital environments and real circumstances, and the abstraction of procedures may not be transferable to reality [60]. Notwithstanding, VSs are cost-effective alternatives that enable learners to repeat the training freely without worrying about time, space, and patient harm [61]. VSs may be designed to achieve better skill acquisition. First, the addition of haptic technology, for example, offers more realistic hands-on experiences [36]. Second, too much immersion may not be desirable as it can lead to cognitive overload and hamper procedure learning [42].

There was no significant difference in fostering clinical reasoning between VSs and mannequins or RPs. Again, this result does not conform to previous meta-analyses which compared VSs to no intervention or traditional education methods [21,62]. Our pooled effect size tended to favor VSs. It has been suggested that VSs, with a large number of customizable clinical scenarios, are highly suitable for improving clinical assessment and decision-making [63]. In particular, the following elements may be beneficial: a training duration of more than 30 minutes, varied clinical scenarios, nonimmersive 2D digital environments, and postscenario feedback [62]. Similar to procedural skills, nonimmersive VSs appear to be more effective than their immersive counterparts. According to the cognitive load theory. 3D digital environments can increase cognitive load as students get distracted by irrelevant stimuli [62]. More data are still required to reach a conclusive result regarding the comparative effectiveness of VSs in improving clinical reasoning.

There was no significant difference in enhancing communication skills between VSs and mannequins or RPs. This is in line with the findings of a previous meta-analysis comparing digital education to traditional learning [15]. Our pooled effect size tended to favor mannequins or RPs. It is noteworthy, however, that more recent studies tended to favor VSs. A possible explanation is that advances in technology can increasingly overcome the long-standing lack of realism and means for interaction in VSs [30]. Greater immersion can elicit a greater sense of “being there” for users [64]. This psychological experience is critical for communication training, to make learners feel as if they are having face-to-face conversations with digital avatars so that higher levels of empathy can be induced. Moreover, VSs have fewer limits on time or place, eliminate the costs of SP training, and offer a variety of clinical scenarios available for student communication training. These properties render VSs with great potential as novel methods for teaching communication skills.

Strengths and Limitations

To the best of our knowledge, this is the first meta-analysis comparing the effectiveness of VSs and conventional simulations. Strengths of this study include pragmatic research questions, well-defined outcome measures, comprehensive searches of the most up-to-date RCTs (ie, the gold standard for evaluating the effectiveness of interventions), and critical analyses of the evidence. Our results indicate that VSs are viable alternatives to face-to-face learning modalities. For institutions, time and money saved can be invested in other research and teaching projects. For students, the removal of time, space, and location limitations leads to greater freedom of learning and wider popularization of education.

On the other hand, limitations must be acknowledged when interpreting the results of this study. First, researchers in several studies were not blinded and might introduce bias when assessing participants. Second, certain subgroup or meta-regression analyses were not feasible or included only one entry due to the limited number of studies. Further analyses may be possible when more RCTs are published. Third, the heterogeneity of the included studies could not be fully explained by leave-one-out, subgroup, or meta-regression analyses. This lack of explanation may be attributed to other unspecified sources of methodological and subject heterogeneity.

Conclusions

Overall, this meta-analysis did not find significant differences between VSs and mannequins or RPs in improving knowledge, procedural skills, clinical reasoning, and communication skills. For procedural skill acquisition, VSs were less effective than mannequins or RPs among nursing participants. For communication skill development, VSs exhibited increasing potency with more advanced technology over time. Given other prominent advantages of VSs (eg, cost-effectiveness, flexibility regarding time or space), it is worth considering these tools as alternatives to traditional education methods. As technology continues to evolve, the comparative effectiveness of VSs will need to be reevaluated and updated. In particular, it is anticipated to witness the integration of emerging large language models (eg, ChatGPT) in VSs and see if they can revolutionize educational practices. We also recommend future research to investigate the effectiveness of VSs across various design configurations.

Acknowledgments

This study was funded by the Peking Union Medical College Hospital Teaching Programme (X361011).

XSL•FO
RenderX

Authors' Contributions

HP, SC, and XH contributed equally to the work. NJ, YZ, SC, and XH contributed to the conception and design of the study. NJ performed search strategies. SL and XL selected articles, extracted data, and assessed the risk of bias. NJ and YZ analyzed data. NJ prepared the original draft. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 29 KB - jmir_v26i1e56195_app1.docx]

Multimedia Appendix 2
Search strategies.
[DOCX File , 14 KB - jmir_v26i1e56195_app2.docx]

Multimedia Appendix 3
Characteristics of selected 27 randomized controlled trials.
[DOCX File , 30 KB - jmir_v26i1e56195_app3.docx]

Multimedia Appendix 4
Trial sequential, leave-one-out, subgroup and meta-regression analyses.
[DOCX File , 4611 KB - jmir_v26i1e56195_app4.docx]

References

1. Khan K, Pattison T, Sherwood M. Simulation in medical education. Med Teach 2011;33(1):1-3. [doi: 10.3109/0142159X.2010.519412] [Medline: 21182376]
2. Herrera-Aliaga E, Estrada LD. Trends and innovations of simulation for twenty first century medical education. Front Public Health 2022;10:619769 [FREE Full text] [doi: 10.3389/fpubh.2022.619769] [Medline: 35309206]
3. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. Acad Med 2003;78(8):783-788. [doi: 10.1097/00001888-200308000-00006] [Medline: 12915366]
4. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011;86(6):706-711 [FREE Full text] [doi: 10.1097/ACM.0b013e318217e119] [Medline: 21512370]
5. Shin S, Park JH, Kim JH. Effectiveness of patient simulation in nursing education: meta-analysis. Nurse Educ Today 2015;35(1):176-182. [doi: 10.1016/j.nedt.2014.09.009] [Medline: 25459172]
6. Oliveira Silva G, Oliveira FSE, Coelho ASG, Cavalcante AMRZ, Vieira FVM, Fonseca LMM, et al. Effect of simulation on stress, anxiety, and self-confidence in nursing students: systematic review with meta-analysis and meta-regression. Int J Nurs Stud 2022;133:104282. [doi: 10.1016/j.ijnurstu.2022.104282] [Medline: 35679634]
7. Watmough S, Box H, Bennett N, Stewart A, Farrell M. Unexpected medical undergraduate simulation training (UMUST): Can unexpected medical simulation scenarios help prepare medical students for the transition to foundation year doctor? BMC Med Educ 2016;16:110 [FREE Full text] [doi: 10.1186/s12909-016-0629-x] [Medline: 27079898]
8. Nielsen RP, Nikolajsen L, Paltved C, Aagaard R. Effect of simulation-based team training in airway management: a systematic review. Anaesthesia 2021;76(10):1404-1415 [FREE Full text] [doi: 10.1111/anae.15375] [Medline: 33497486]
9. Luzzi A, Hellwinkel J, O'Connor M, Crutchfield C, Lynch TS. The efficacy of arthroscopic simulation training on clinical ability: a systematic review. Arthroscopy 2021;37(3):1000-1007.e1. [doi: 10.1016/j.arthro.2020.09.018] [Medline: 33220467]
10. James HK, Chapman AW, Pattison GTR, Griffin DR, Fisher JD. Systematic review of the current status of cadaveric simulation for surgical training. Br J Surg 2019;106(13):1726-1734 [FREE Full text] [doi: 10.1002/bjs.11325] [Medline: 31573088]
11. Sarfati L, Ranchon F, Vantard N, Schwiertz V, Larbre V, Parat S, et al. Human-simulation–based learning to prevent medication error: a systematic review. J Eval Clin Pract 2019;25(1):11-20. [doi: 10.1111/jep.12883] [Medline: 29383867]
12. Zendejas B, Brydges R, Wang AT, Cook DA. Patient outcomes in simulation-based medical education: a systematic review. J Gen Intern Med 2013;28(8):1078-1089 [FREE Full text] [doi: 10.1007/s11606-012-2264-5] [Medline: 23595919]
13. Turner S, Harder N, Vigier D, Cooper A, Pinel K, Mitchell K. Lessons from implementing virtual simulations: a multi-program evaluation. Clin Simul Nurs 2023;74:57-64. [doi: 10.1016/j.ecns.2022.09.001]
14. Lioce L, Anderson M, Diaz D, Robertson J, Chang T, Downing D, et al. Healthcare Simulation Dictionary. 2nd Edition. Rockville, MD: Agency for Healthcare Research and Quality; 2020.

XSL•FO
RenderX

15. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(8):e12967 [FREE Full text] [doi: 10.2196/12967] [Medline: 31456579]
16. Han SG, Kim YD, Kong TY, Cho J. Virtual reality–based neurological examination teaching tool (VRNET) versus standardized patient in teaching neurological examinations for the medical students: a randomized, single-blind study. BMC Med Educ 2021;21(1):493 [FREE Full text] [doi: 10.1186/s12909-021-02920-4] [Medline: 34526004]
17. Dante A, Marcotullio A, Masotta V, Caponnetto V, La CC, Bertocchi L, et al. From high-fidelity patient simulators to robotics and artificial intelligence: a discussion paper on new challenges to enhance learning in nursing education. 2020 Presented at: 10th International Conference on Methodologies and Intelligent Systems for Technology Enhanced Learning; June 17-19, 2020; L´Aquila, Italy. [doi: 10.1007/978-3-030-52287-2_11]
18. Pietersen PI, Jørgensen R, Graumann O, Konge L, Skaarup SH, Lawaetz Schultz HH, et al. Training thoracic ultrasound skills: a randomized controlled trial of simulation-based training versus training on healthy volunteers. Respiration 2021;100(1):34-43. [doi: 10.1159/000509298] [Medline: 33454705]
19. Choules AP. The use of elearning in medical education: a review of the current situation. Postgrad Med J 2007;83(978):212-216 [FREE Full text] [doi: 10.1136/pgmj.2006.054189] [Medline: 17403945]
20. Bumbach MD, Culross BA, Datta SK. Assessing the financial sustainability of high-fidelity and virtual reality simulation for nursing education: a retrospective case analysis. Comput Inform Nurs 2022;40(9):615-623. [doi: 10.1097/CIN.0000000000000916] [Medline: 35524349]
21. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
22. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(1):e12959 [FREE Full text] [doi: 10.2196/12959] [Medline: 30668519]
23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
24. Sim JH, Abdul Aziz YF, Mansor A, Vijayananthan A, Foong CC, Vadivelu J. Students' performance in the different clinical skills assessed in OSCE: What does it reveal? Med Educ Online 2015;20:26185 [FREE Full text] [doi: 10.3402/meo.v20.26185] [Medline: 25697602]
25. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.4. Cochrane. 2023. URL: https://training.cochrane.org/handbook [accessed 2023-08-22]
26. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 2014;14:135 [FREE Full text] [doi: 10.1186/1471-2288-14-135] [Medline: 25524443]
27. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
28. Wetterslev J, Jakobsen JC, Gluud C. Trial sequential analysis in systematic reviews with meta-analysis. BMC Med Res Methodol 2017;17(1):39 [FREE Full text] [doi: 10.1186/s12874-017-0315-7] [Medline: 28264661]
29. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019;22(4):153-160 [FREE Full text] [doi: 10.1136/ebmental-2019-300117] [Medline: 31563865]
30. Sapkaroski D, Mundy M, Dimmock MR. Immersive virtual reality simulated learning environment versus role-play for empathic clinical communication training. J Med Radiat Sci 2022;69(1):56-65 [FREE Full text] [doi: 10.1002/jmrs.555] [Medline: 34706398]
31. O'Rourke SR, Branford KR, Brooks TL, Ives LT, Nagendran A, Compton SN. The emotional and behavioral impact of delivering bad news to virtual versus real standardized patients: a pilot study. Teach Learn Med 2020;32(2):139-149. [doi: 10.1080/10401334.2019.1652180] [Medline: 31437006]
32. Weber U, Zapletal B, Base E, Hambrusch M, Ristl R, Mora B. Resident performance in basic perioperative transesophageal echocardiography: comparing 3 teaching methods in a randomized controlled trial. Medicine 2019;98(36):e17072 [FREE Full text] [doi: 10.1097/MD.0000000000017072] [Medline: 31490407]
33. Jiang B, Ju H, Zhao Y, Yao L, Feng Y. Comparison of the efficacy and efficiency of the use of virtual reality simulation with high-fidelity mannequins for simulation-based training of fiberoptic bronchoscope manipulation. Simul Healthc 2018;13(2):83-87. [doi: 10.1097/SIH.0000000000000299] [Medline: 29621098]
34. Spadaro S, Karbing DS, Fogagnolo A, Ragazzi R, Mojoli F, Astolfi L, et al. Simulation training for residents focused on mechanical ventilation: a randomized trial using mannequin-based versus computer-based simulation. Simul Healthc 2017;12(6):349-355 [FREE Full text] [doi: 10.1097/SIH.0000000000000249] [Medline: 28825930]
35. Ahlqvist JB, Nilsson TA, Hedman LR, Desser TS, Dev P, Johansson M, et al. A randomized controlled trial on 2 simulation-based training methods in radiology: effects on radiologic technology student skill in assessing image quality. Simul Healthc 2013;8(6):382-387. [doi: 10.1097/SIH.0b013e3182a60a48] [Medline: 24096919]

XSL•FO
RenderX

36. LeBlanc J, Hutchison C, Hu Y, Donnon T. A comparison of orthopaedic resident performance on surgical fixation of an ulnar fracture using virtual reality and synthetic models. J Bone Joint Surg Am 2013;95(9):e60, S1-e60, S5. [doi: 10.2106/JBJS.K.01284] [Medline: 23636199]
37. Andreatta PB, Maslowski E, Petty S, Shim W, Marsh M, Hall T, et al. Virtual reality triage training provides a viable solution for disaster-preparedness. Acad Emerg Med 2010;17(8):870-876 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00728.x] [Medline: 20670325]
38. Youngblood P, Harter PM, Srivastava S, Moffett S, Heinrichs WL, Dev P. Design, development, and evaluation of an online virtual emergency department for training trauma teams. Simul Healthc 2008;3(3):146-153. [doi: 10.1097/SIH.0b013e31817bedf7] [Medline: 19088658]
39. Deladisma AM, Cohen M, Stevens A, Wagner P, Lok B, Bernard T, et al. Do medical students respond empathetically to a virtual patient? Am J Surg 2007;193(6):756-760. [doi: 10.1016/j.amjsurg.2007.01.021] [Medline: 17512291]
40. Tran MT, Ahmad M, Patel K, Argyriou O, Davies A, Shalhoub J. Comparing virtual reality and simulation to teach the assessment and management of acute surgical scenarios: a pilot study. Health Sci Rep 2024;7(7):e2245 [FREE Full text] [doi: 10.1002/hsr2.2245] [Medline: 38983682]
41. Malik TG, Mahboob U, Khan RA, Alam R. Virtual patients versus standardized patients for improving clinical reasoning skills in ophthalmology residents. A randomized controlled trial. BMC Med Educ 2024;24(1):429 [FREE Full text] [doi: 10.1186/s12909-024-05241-4] [Medline: 38649884]
42. Simsek-Cetinkaya S, Cakir SK. Evaluation of the effectiveness of artificial intelligence assisted interactive screen-based simulation in breast self-examination: an innovative approach in nursing students. Nurse Educ Today 2023;127:105857. [doi: 10.1016/j.nedt.2023.105857] [Medline: 37253303]
43. Sahin Karaduman G, Basak T. Is virtual patient simulation superior to human patient simulation: a randomized controlled study. Comput Inform Nurs 2023;41(6):467-476. [doi: 10.1097/CIN.0000000000000957] [Medline: 36633879]
44. Padilha JM, Machado PP, Ribeiro A, Ramos J, Costa P. Clinical virtual simulation in nursing education: randomized controlled trial. J Med Internet Res 2019;21(3):e11529 [FREE Full text] [doi: 10.2196/11529] [Medline: 30882355]
45. Haerling KA. Cost-utility analysis of virtual and mannequin-based simulation. Simul Healthc 2018;13(1):33-40. [doi: 10.1097/SIH.0000000000000280] [Medline: 29373382]
46. Cobbett S, Snelgrove-Clarke E. Virtual versus face-to-face clinical simulation in relation to student knowledge, anxiety, and self-confidence in maternal-newborn nursing: a randomized controlled trial. Nurse Educ Today 2016;45:179-184. [doi: 10.1016/j.nedt.2016.08.004] [Medline: 27537670]
47. Liaw SY, Chan SWC, Chen FG, Hooi SC, Siau C. Comparison of virtual patient simulation with mannequin-based simulation for improving clinical performances in assessing and managing clinical deterioration: randomized controlled trial. J Med Internet Res 2014;16(9):e214 [FREE Full text] [doi: 10.2196/jmir.3322] [Medline: 25230684]
48. Johnson MP, Hickey KT, Scopa-Goldman J, Andrews T, Boerem P, Covec M, et al. Manikin versus web-based simulation for advanced practice nursing students. Clin Simul Nurs 2014;10(6):e317-e323. [doi: 10.1016/j.ecns.2014.02.004]
49. Chang KKP, Chung JWY, Wong TKS. Learning intravenous cannulation: a comparison of the conventional method and the cathSim intravenous training system. J Clin Nurs 2002;11(1):73-78. [doi: 10.1046/j.1365-2702.2002.00561.x] [Medline: 11845758]
50. Durmaz A, Dicle A, Cakan E, Cakir S. Effect of screen-based computer simulation on knowledge and skill in nursing students' learning of preoperative and postoperative care management: a randomized controlled study. Comput Inform Nurs 2012;30(4):196-203. [doi: 10.1097/NCN.0b013e3182419134] [Medline: 22228217]
51. Jung EY, Park DK, Lee YH, Jo HS, Lim YS, Park RW. Evaluation of practical exercises using an intravenous simulator incorporating virtual reality and haptics device technologies. Nurse Educ Today 2012;32(4):458-463. [doi: 10.1016/j.nedt.2011.05.012] [Medline: 21664014]
52. Lee M, Kim SK, Go Y, Jeong H, Lee Y. Positioning virtual reality as means of clinical experience in mental health nursing education: a quasi-experimental study. Appl Nurs Res 2024;77:151800. [doi: 10.1016/j.apnr.2024.151800] [Medline: 38796255]
53. Liaw SY, Sutini, Chua WL, Tan JZ, Levett-Jones T, Ashokka B, et al. Desktop virtual reality versus face-to-face simulation for team-training on stress levels and performance in clinical deterioration: a randomised controlled trial. J Gen Intern Med 2023;38(1):67-73 [FREE Full text] [doi: 10.1007/s11606-022-07557-7] [Medline: 35501626]
54. Liaw SY, Ooi SW, Rusli KDB, Lau TC, Tam WWS, Chua WL. Nurse-physician communication team training in virtual reality versus live simulations: randomized controlled trial on team communication and teamwork attitudes. J Med Internet Res 2020;22(4):e17279 [FREE Full text] [doi: 10.2196/17279] [Medline: 32267235]
55. Chen FQ, Leng YF, Ge JF, Wang DW, Li C, Chen B, et al. Effectiveness of virtual reality in nursing education: meta-analysis. J Med Internet Res 2020;22(9):e18290 [FREE Full text] [doi: 10.2196/18290] [Medline: 32930664]
56. Woon APN, Mok WQ, Chieng YJS, Zhang HM, Ramos P, Mustadi HB, et al. Effectiveness of virtual reality training in improving knowledge among nursing students: a systematic review, meta-analysis and meta-regression. Nurse Educ Today 2021;98:104655. [doi: 10.1016/j.nedt.2020.104655] [Medline: 33303246]
57. Liu K, Zhang W, Li W, Wang T, Zheng Y. Effectiveness of virtual reality in nursing education: a systematic review and meta-analysis. BMC Med Educ 2023;23(1):710 [FREE Full text] [doi: 10.1186/s12909-023-04662-x] [Medline: 37770884]

XSL•FO
RenderX

58. Kiegaldie D, Shaw L. Virtual reality simulation for nursing education: effectiveness and feasibility. BMC Nurs 2023;22(1):488 [FREE Full text] [doi: 10.1186/s12912-023-01639-5] [Medline: 38114940]
59. Ryan GV, Callaghan S, Rafferty A, Higgins MF, Mangina E, McAuliffe F. Learning outcomes of immersive technologies in health care student education: systematic review of the literature. J Med Internet Res 2022;24(2):e30082 [FREE Full text] [doi: 10.2196/30082] [Medline: 35103607]
60. Gu Y, Zou Z, Chen X. The effects of vSIM for Nursing™ as a teaching strategy on fundamentals of nursing education in undergraduates. Clin Simul Nurs 2017;13(4):194-197. [doi: 10.1016/j.ecns.2017.01.005]
61. Plotzky C, Loessl B, Kuhnert B, Friedrich N, Kugler C, König P, et al. My hands are running away—learning a complex nursing skill via virtual reality simulation: a randomised mixed methods study. BMC Nurs 2023;22(1):222 [FREE Full text] [doi: 10.1186/s12912-023-01384-9] [Medline: 37370124]
62. Sim JJM, Rusli KDB, Seah B, Levett-Jones T, Lau Y, Liaw SY. Virtual simulation to enhance clinical reasoning in nursing: a systematic review and meta-analysis. Clin Simul Nurs 2022;69:26-39 [FREE Full text] [doi: 10.1016/j.ecns.2022.05.006] [Medline: 35754937]
63. LaManna JB, Guido-Sanz F, Anderson M, Chase SK, Weiss JA, Blackwell CW. Teaching diagnostic reasoning to advanced practice nurses: positives and negatives. Clin Simul Nurs 2019;26:24-31. [doi: 10.1016/j.ecns.2018.10.006]
64. Cummings JJ, Bailenson JN. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychol 2015;19(2):272-309. [doi: 10.1080/15213269.2015.1015740]

Abbreviations

AI: artificial intelligence
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses
RCT: randomized controlled trial
RIS: required information size
RP: real person
SMD: standardized mean difference
SP: standardized patient
TSA: trial sequential analysis
VP: virtual patient
VR: virtual reality
VS: virtual simulation

Edited by A Mavragani; submitted 09.01.24; peer-reviewed by J Zhang, YG Abdildin; comments to author 03.05.24; revised version received 23.06.24; accepted 16.09.24; published 05.12.24.

Please cite as:
Jiang N, Zhang Y, Liang S, Lyu X, Chen S, Huang X, Pan H
Effectiveness of Virtual Simulations Versus Mannequins and Real Persons in Medical and Nursing Education: Meta-Analysis and Trial Sequential Analysis of Randomized Controlled Trials
J Med Internet Res 2024;26:e56195
URL: https://www.jmir.org/2024/1/e56195
doi:10.2196/56195
PMID:

©Nan Jiang, Yuelun Zhang, Siyu Liang, Xiaohong Lyu, Shi Chen, Xiaoming Huang, Hui Pan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An e-Learning Intervention for Professionals to Promote Family-Centered Cancer Care When a Significant Caregiver for Children Is at End of Life: Mixed Methods Evaluation Study

Cherith Jane Semple[1], PhD; Carla O'Neill[2], PhD; Sarah Sheehan[2], MSc; Tanya McCance[3], PhD; Amanda Drury[4], PhD; Jeffrey R Hanna[1], PhD

[1]South Eastern Health & Social Care Trust, Ulster University, Belfast, United Kingdom
[2]University College Dublin, Dublin, Ireland
[3]Ulster University, Belfast, United Kingdom
[4]Dublin City University, Dublin, Ireland

Corresponding Author:
Cherith Jane Semple, PhD
South Eastern Health & Social Care Trust
Ulster University
York Street
Belfast, BT15 1ED
United Kingdom
Phone: 44 2870123456
Email: c.semple@ulster.ac.uk

Abstract

Background: Families are often unsure how best to prepare dependent children for the death of a significant caregiver with a poor cancer prognosis and seek guidance and support from health care teams. Health and social care professionals (hereafter referred to as *professionals*) often lack educational opportunities to gain the desired knowledge, skills, and confidence to provide family-centered supportive cancer care. e-Learning has positively impacted access and reach, improving educational opportunities in health care.

Objective: We aimed to evaluate the acceptability, usability, and effectiveness of an evidence-based, theory-driven e-learning intervention to equip and promote professionals' self-efficacy to deliver family-centered supportive cancer care when a significant caregiving member for dependent children is at the end of life.

Methods: Guided by the "person-based approach," a mixed methods outcome evaluation was used. To determine the effect on self-efficacy, participants completed a validated pretest and posttest 12-item self-efficacy survey. The use of one-on-one, remote semistructured interviews and single-item questions determined the usability by professionals of the e-learning intervention and the acceptability of perceived learning in clinical practice. To generate enhanced insights, quantitative and qualitative data were integrated through a 4-stage, modified pillar integration process.

Results: Overall, 158 participants completed the pretest survey for the e-learning resource, with 99 (62.7%) completing the posttest survey. Semistructured interviews were conducted with 12 professionals at least 1 month after the intervention. Findings highlighted a statistically significant improvement in posttest self-efficacy (99/158, 62.7%; $P<.001$). Usability of the e-learning intervention was positive, with participants reporting that it was clear and organized (mean 4.84, SD 0.373), the layout was appealing (mean 4.71, SD 0.539), the language was easy to understand (mean 4.71, SD 0.407), and graphics and media were purposeful (mean 4.76, SD 0.495) and engaging (mean 4.67, SD 0.703). Determining acceptability, participants considered that the intervention would positively impact practice (mean 4.60, SD 0.589) and increase knowledge (mean 4.56, SD 0.677), with appropriate practical examples to support learning (mean 4.58, SD 0.610). Following engagement with the e-learning intervention, professionals reported preparedness to deliver supportive adult-professional end-of-life cancer care, when an adult with significant caregiving responsibilities is dying. Findings demonstrated transferable learning to additional contexts, such as other close adult-child relational bonds (grandparents) and to life-limiting conditions.

Conclusions: The systematic and iterative person-based approach optimized the acceptability of a novel e-learning intervention, having the potential to promote family-centered supportive end-of-life cancer care. This accessible e-learning intervention makes an important contribution to the recognized global gap of educational interventions in this field. Equipping professionals with

XSL•FO
RenderX

family-centered supportive end-of-life care improves self-efficacy and preparedness to engage in challenging conversations, with the potential to promote better outcomes for affected adults and children and mediate adverse outcomes for adults and children before and after bereavement.

(J Med Internet Res 2024;26:e65619) doi:10.2196/65619

KEYWORDS

e-learning; intervention evaluation; mixed methods; end-of-life care; educational intervention; professionals; self-efficacy; cancer; family-centered care; cancer care; person-based approach; qualitative; evidence-based; parent; adverse outcome; eHealth

Introduction

A recent systematic review indicated that a substantial proportion (14%-25%) of people diagnosed with cancer are parents or have significant caregiving responsibilities for dependent children (aged <18 years) [1]. Having cancer alongside significant caregiving responsibilities for children presents unique challenges, contributing significantly to higher levels of depression and anxiety [2]. Strikingly, 1 in 5 dependent children who have a parent with cancer experience the death of their parent from cancer [3].

Many families are often unsure how best to prepare children for the end-of-life experience of a significant caregiver who has a poor cancer prognosis [4]. Often, adults in this situation not only experience higher levels of psychological distress but also are at a greater risk of family dysfunction [3]. In recent years, there has been an increased focus on communication challenges within families and the provision of supportive programs [5,6]. Nonetheless, the needs of families frequently remain unmet and efforts to address the psychosocial needs of families when a significant adult is at the end of life remain in its infancy [5]. Consequently, many children are unprepared for this expectant and traumatic experience, predisposing them to adverse outcomes before and after bereavement [3,7]. Studies have shown that a child who is unprepared for the death of a significant adult in childhood is associated with an increased risk of mental health problems and threats to emotional well-being, such as anxiety, depression, and a perceived lack of control over what happens in one's life [7,8]. It is also correlated with increased criminality rates, risk-taking behaviors, behavioral problems, and poor educational attainment [3,7].

Health and social care professionals (hereafter referred to as *professionals*) are well-placed to support families as they prepare children for the death of a significant adult caregiver to cancer [9-11]. Despite this, professionals often avoid significant end-of-life conversations with adults regarding their caregiving responsibilities for children for fear of making a challenging and fraught situation worse [12,13]. Professionals have repeatedly highlighted a need for training to equip themselves with the skills, knowledge, and confidence to provide meaningful family-centered cancer care at the end of life [9,12,13]. Despite this, a systematic review conducted by our team identified a paucity of available educational interventions for professionals (n=2) and an imminent need for robust educational interventions to be developed [5].

Given the avid need for training, our team designed and evaluated a face-to-face, evidence-based, and theory-driven educational intervention for professionals to support adult-child end-of-life communication and management [14]. A mixed methods evaluation demonstrated statistically significant (P<.001) improvements in professionals' perceived self-efficacy after completing the face-to-face educational intervention, with in-depth interviews after the intervention identifying that clinicians had gained new tools and strategies on how to start supportive conversations with adults regarding the children [14]. To progress family-centered end-of-life cancer conversations in routine practice at a global level, there is an inherent need to promote the accessibility and availability of the provision of such education [14,15].

There is evidence to suggest that greater flexibility and access to digital education for professionals is transforming the delivery of health care education [16]. It also positively impacts promoting person-centered outcomes, with patients and carers reporting greater satisfaction, empowerment, and inclusion in their physical and psychological health needs [17-19]. It has been suggested that the *success* of e-learning interventions in promoting positive behavior change is associated with the interventions being cocreated with end users. This ensures that the content, features, and navigational aspects of the resource are relevant, applicable, motivating, enjoyable, and informative [20-22]. These principles align with the person-based approach to developing e-learning interventions [23].

Guided by the person-based approach, the research team sought to adapt the face-to-face educational intervention [14] to a stand-alone, self-directed e-learning intervention for professionals. To support this process, key behavioral issues, needs, and challenges that the e-learning intervention must address were outlined [23]. This consisted of the development of a logic model that was underpinned by the family resilience theory [24] and social cognitive theory [25]. In addition, guiding principles were developed, comprising design objectives and distinctive features of the e-learning intervention that were necessary to achieve the project objectives and logic model [26]. The research team collaborated with an expert group and a team of learning technologists to refine the content of the face-to-face educational intervention [14] and identified appropriate and relevant components for the e-learning intervention (the content and logic model has been published in detail elsewhere) [26]. The e-learning intervention prototype was developed using the authoring tool RISE360 (Articulate). Alongside appropriate visuals and icons, interactive elements were incorporated to the e-learning intervention prototype such as flip-card activities, expandable icons, and reflective activities [26].

XSL•FO
RenderX

The e-learning intervention prototype was optimized with end users to ensure it was appropriate, relevant, and applicable [27]. An iterative approach involving 3 cycles of data collection and analysis was used to enact changes to the e-learning intervention with professionals and patient and public involvement representatives [26]. In line with the person-based approach, there is a need to determine the effectiveness, usability, and acceptability of the e-learning intervention [16,27,28].

Using the person-based approach, the aim of this study is to evaluate the acceptability, usability, effectiveness, and outcomes of an e-learning intervention developed to equip professionals to deliver family-centered supportive cancer care when an adult with significant caregiving responsibilities for children is at the end of life with cancer. The objectives of the study are as follows:

1. Explore professionals' acceptability and usability of engaging with an e-learning intervention on how best to support adults at the end of life with cancer who have significant caregiving responsibilities for dependent children.
2. Determine the effectiveness of an e-learning intervention on professionals' perceived self-efficacy to provide family-centered end-of-life cancer care in routine practice.
3. Explore professionals' experience and perceived learning from the e-learning intervention in clinical practice.

Methods

Overview

To guide the evaluation process, this project focuses on levels 1 and 2 (reactions and learning) of the 4 levels of evaluation in Kirkpatrick's model [29]. Using a mixed methods approach, quantitative pretest and posttest surveys were completed before and after participants completed the e-learning intervention. Qualitative interviews were conducted at least 4 weeks after completion of the e-learning intervention to further explore users' experience of and perceived impact of the e-learning intervention on professionals' practice. The main purpose of adopting a mixed methods design was to enable both complementarity and expansion of data, with both quantitative and qualitative components given equal weighting for this applied research study of a complex issue.

Setting

Between October 2023 and June 2024, individuals were invited through a study advertisement to evaluate the e-learning intervention if they were a health and social care professional involved in the care of families impacted by cancer at the end of life. There was no limitation placed on geographic location, health or social care context, or health care provider. Participants were informed about the resource evaluation through social media posts on X (formerly Twitter), hospice organizations, flyers that were distributed at relevant national and international cancer and palliative care conferences and events and the research team's professional networks.

Sampling

A minimum sample of 34 was required (power calculation in G*power based on repeated measures 2-tailed *t* test using a power of 0.8 and medium effect size of 0.05, with a Cronbach α value of 0.05). A planned sample size of 50 was selected to allow for missing data at follow up [30].

Eligibility criteria included registered health or social care professionals involved in end-of-life care of families impacted by cancer, who consented to the study with an ability to understand and converse in English. Trainee health and social care professionals were excluded from participation.

Procedure

The study advertisement indicated that individuals could access the e-learning intervention by registering for a free account on the All-Ireland Institute of Hospice and Palliative Care's Palliative Hub Learning Platform. Eligible participants could then access the participant information sheet and consent form that were embedded in a Qualtrics survey on a separate URL link. The participant information sheet provided individuals with explanations as to why the study was being conducted, what participation would include, the risks involved, and whom to contact if they had any questions regarding the study. Once interested and willing professionals completed the consent form, the pretest survey was automatically loaded within the Qualtrics link on the learning platform page. Professionals could only access the pretest survey if they completed the consent form. An honor-based system was in place where participants could access the e-learning intervention without completing a pretest survey.

Data Collection

Pretest Survey

Participants completed the Self-Efficacy Questionnaire (SE-12) for measuring clinical communication skills on a Likert scale ranging from 0 (very uncertain) to 10 (very certain) for 12 questions [31]. The maximum score was 120, which reflects high self-efficacy [31]. Of note, the framing of questions in the SE-12 was applied to conversations with adults at the end of life with cancer regarding the children. Multimedia Appendix 1 presents the modified version of the SE-12 scale used in this study. Appropriate demographic questions were included to capture details about the convenience sample. Participants were asked to create a unique code to match each of their pretest and posttest surveys with their first initial and date of birth in the format "xddmmyy."

Posttest Survey

The posttest survey included the same version of the SE-12 scale as the pretest survey [31]. Alongside this, participants completed single-item questions to evaluate the usefulness and relevance of the e-learning interventions on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The same demographic questions were included in the posttest survey to capture details about individuals who may have only completed the posttest survey.

XSL•FO
RenderX

Qualitative Interviews

A total of 29 email addresses were provided by participants who completed the posttest survey. These individuals were emailed at least 4 weeks after completing the posttest survey and were provided with a participant information sheet and consent form from the third author (SS). A follow-up invitation was sent to individuals 2 weeks after the initial invitation to aid participation.

Semistructured interviews were conducted between February and June 2024. A topic guide was developed by the research team, guided by Kirkpatrick's model of evaluation [29] and other relevant literature (see Textbox 1 for a sample of topics). Interviews were completed when no further categories emerged from the data. Interviews were conducted on Zoom, were audio-recorded, and lasted between 12.5 and 35.47 (mean average 19.96) minutes. Interviews were conducted by the third author (SS) who had no prior relationships with the participants. Recruitment was discontinued following the completion of 12 interviews as no new main themes were being identified.

Textbox 1. Sample of topics explored during the qualitative interviews.

- Explore professional background, location, and clinical experience.
- Explore professional's motivations for completing the e-learning intervention.
- Explore the professional's experience of the e-learning intervention.
- Explore the professional's perceived learning from the e-learning intervention.
- Explore the professional's perceived impact of the e-learning intervention on practice.
- Explore the professional's perception of training needs.
- Anything else that is relevant.

Data Analysis

Overview

The integration of the quantitative and qualitative data was achieved through a modified pillar integration process [32-34]. The process facilitates integration through 4 stages: listing, matching, checking, and pillar building [32]. This process allows meta-inferences about the data to be generated to enhance insights gained relating to the impact of the e-learning intervention [34].

Quantitative Analysis

Close-ended questions from the pretest and posttest surveys were analyzed using SPSS (version 29; IBM Corp). Descriptive statistics were used to analyze demographic and single-item questions. A 1-way ANOVA was used to analyze the scores of self-efficacy in the pretest and posttest surveys.

Qualitative Analysis

Open-ended questions from the posttest questionnaire and semistructured interviews were analyzed using reflexive thematic analysis [35-37]. Initial coding, generation of initial themes, development and review of themes, and the definition and naming of themes were conducted. Transcripts were transcribed verbatim, and these transcripts were verified by SS. Then, SS coded the data by marking similar words and phrases within the transcripts. This process was managed using NVIVO (version 1.7). Codes were reviewed by CON. Following this, both SS and CON identified where the codes merged as themes. Themes were discussed and refined through critical dialogue with all coauthors.

Integration of Quantitative and Qualitative Data

Data from the quantitative and qualitative data were arranged in 2 columns. A description of the interpretation of these findings were listed alongside each data item (ie, listing). Relational quantitative and qualitative data were then aligned with each other in the joint display (ie, matching). These matches were then reviewed and refined independently by all research team members to verify that the matched data represent appropriately matched points of integration (ie, checking). Differences were resolved through critical dialogue as a team. The insights generated from the process were conceptualized as pillars that formed the central column of the joint display (ie, pillar building themes column).

Ethical Considerations

Participants were informed that all information shared would be confidential. Unique codes were used to match the pretest and posttest survey data. Transcripts were anonymized using unique codes. Data were stored on the University College Dublin OneDrive with password protection and are accessible only by the research team. Data protection procedures are in place to destroy all data in accordance with the General Data Protection Regulation and the Data Protection Act 2018. Ethics approval was obtained from Ulster University (FCNUR-23-002) and University College Dublin (LS-22-65).

Results

Overview

Overall, 234 individuals registered to access the course. A total of 158 (67.5%) participants completed the pretest survey, and of those completing the pretest survey 99 (62.7%) participants completed the posttest survey. Of the 158 participants who completed the pretest survey, most were registered nurses (n=100, 63.3%) who were residing in the United Kingdom (n=88, 55.7%) or the Republic of Ireland (n=45, 28.5%). Of the 99 participants who completed the posttest survey, most were

XSL•FO
RenderX

registered nurses (n=65, 66%), residing in the United Kingdom (n=59, 60%) or the Republic of Ireland (n=22, 22%).

A total of 12 qualitative interviews were conducted between 4 and 17 weeks after the intervention (m Avg 7.4 weeks). Of these 12 participants, 2 (17%) were medical professionals, 9 (75%) were registered nurses, and 1 (8%) was a health care assistant. These professionals were predominately residing in the United Kingdom (8/12, 67%), and had 22.92 mean average years of clinical experience, with 10.92 mean average years specifically working in cancer care. Sample characteristics are reported in Table 1.

Pillars were generated from an integration of the analysis of the open and closed responses to the pre and posttest surveys and qualitative interviews. A joint display (Multimedia Appendix 2) presents the integrated findings of the quantitative and qualitative data. The integrated analysis generated two pillars: (1) experience of engaging with the e-learning intervention and (2) professionals feeling equipped to have family-centered supportive end-of-life conversations.

XSL•FO
RenderX

Table 1. Sample characteristics for all participants (pretest survey, posttest survey, and qualitative interviews).

Characteristics	Participants Values, n (%)
Pretest survey (n=158)	
Professional role	
Registered nurse	100 (63.3)
Medical doctor	17 (11.8)
Allied health professional	15 (9.5)
Social worker	8 (5.1)
Counselor	3 (1.9)
Support worker (before or after bereavement)	7 (4.4)
Health care assistant	6 (3.8)
Student nurse	1 (0.6)
Psychotherapist	1 (0.6)
Location of participants	
Albania	8 (5.1)
Australia	1 (0.6)
Croatia	1 (0.6)
Cyprus	1 (0.6)
France	1 (0.6)
India	1 (0.6)
Ireland	45 (28.5)
Malawi	1 (0.6)
Malta	1 (0.6)
South Africa	1 (0.6)
Spain	6 (3.8)
Thailand	1 (0.6)
Turkey	2 (1.3)
United Kingdom	88 (55.7)
Years of professional experience	
0-4	15 (9.5)
5-9	23 (14.6)
10-14	38 (29.7)
15-19	15 (11.7)
20-24	21 (16.4)
25-29	12 (7.6)
>30	30 (19)
Missing	4 (2.5)
Years of experience working in cancer care	
0-4	43 (27.2)
5-9	36 (22.8)
10-14	27 (17.1)
15-19	14 (8.9)
20-24	18 (28.4)
25-29	7 (4.4)

XSL•FO
RenderX

Characteristics	Participants Values, n (%)
>30	9 (5.7)
Missing	4 (2.5)
Previous training	
Yes	20 (12.7)
Advanced communication skills	6 (3.8)
Face-to-face training from team	6 (3.8)
Charity organization training	6 (3.8)
Postgraduate studies	2 (1.3)
No	135 (85.4)
Missing	3 (1.9)
Children (aged <18 years) of age at home	
Yes	71 (44.9)
No	83 (52.5)
Missing	4 (2.5)
Posttest survey (n=99)	
Professional role	
Registered nurse	65 (65.7)
Medical doctor	9 (9.1)
Allied health professional	6 (6.1)
Social worker	8 (8.1)
Counselor	3 (3)
Support worker (before or after bereavement)	5 (5.1)
Health care assistant	1 (1)
Missing	2 (2)
Location of participants	
Albania	8 (8.1)
Australia	1 (1)
Croatia	1 (1)
France	1 (1)
Republic of Ireland	22 (22.2)
Malawi	1 (1)
South Africa	1 (1)
Spain	2 (2)
Turkey	1 (1)
United Kingdom	59 (59.6)
Missing	2 (2)
Years of professional experience	
0-4	9 (9.1)
5-9	16 (16.2)
10-14	22 (22.2)
15-19	6 (6.1)
20-24	17 (17.2)
25-29	7 (7.1)

XSL•FO
RenderX

Characteristics	Participants Values, n (%)
>30	16 (16.2)
Missing	6 (6.1)
Years of experience working in cancer care	
0-4	24 (24.2)
5-9	23 (23.2)
10-14	19 (19.2)
15-19	8 (8.1)
20-24	12 (12.1)
25-29	4 (4)
>30	3 (3)
Missing	6 (6.1)
Previous training	
Yes	14 (14.1)
No	80 (80.8)
Missing	5 (5.1)
Children <18 years of age at home	
Yes	41 (41.1)
No	53 (53.5)
Missing	5 (5.1)
Qualitative interviews (n=12)	
Professional role	
Registered nurse	9 (75)
Palliative care consultant	1 (8.3)
Palliative care coordinator	1 (8.3)
Health care assistant	1 (8.3)
Location of participants	
Australia	1 (8.3)
Republic of Ireland	2 (16.7)
Spain	1 (8.3)
United Kingdom	8 (66.7)
Years of professional experience	
0-4	2 (16.7)
5-9	0 (0)
10-14	1 (8.3)
15-19	0 (0)
20-24	4 (33.3)
25-29	0 (0)
>30	5 (41.7)

Pillar 1: Experience of Engaging With the e-Learning Intervention

In the posttest survey, participants (97/99, 98%) rated the graphics and media integrated into the resource as being purposeful (mean 4.76, SD 0.495). Similar findings were identified within the open-text posttest responses and interview data, with most participants highlighting the positive and helpful nature of the learning from the videos (Q.1.1.1 and Q.1.1.2; Multimedia Appendix 2):

XSL•FO
RenderX

> *These short bite size videos like those, they're excellent, you're fully engaged right the whole way through and you think oh yes I remember that, that was really good. So that, and also the, I won't call them actors but the people who participated in the video, you know they were very, very good, spoke very well, spoke very short, concise, gave short concise information which was very beneficial to the actual course.* [Interview 02, registered nurse]

Some professionals felt there was a need for additional educational video resources roleplaying scenarios of how best to navigate conversations with adults who are emotionally resistant or reluctant to tell their children the reality of the poor cancer prognosis (Q.1.1.3 and Q.1.1.4).

In the posttest survey, participants (97/99, 98%; Multimedia Appendix 2) positively rated the usability of the e-learning intervention, reporting it as easy to use (mean 4.74, SD 0.485; Multimedia Appendix 2), clear, and organized (mean 4.84, SD 0.373; Multimedia Appendix 2), with an appealing layout (mean 4.71, SD 0.539; Multimedia Appendix 2) and with integrated, easy-to-understand language (mean 4.71, SD 0.407; Multimedia Appendix 2). Qualitative data complemented these findings, with professionals positively reporting on the flexibility of being able to complete the e-learning intervention in their own time (Q.2.2.2; Multimedia Appendix 2). Other aspects that were positively reported in the qualitative interviews included the navigational ease of the e-learning intervention (Q.1.2.1, Q.1.2.4, and Q.1.2.5; Multimedia Appendix 2) and the engaging and varying nature of the learning tools such as the interactive elements, videos, and written pieces (Q.1.2.6; Multimedia Appendix 2):

> *Yeah, like that I thought every segment was well laid out. I think it had a very logical and methodical and the story was well told, you know with that mixture of whether you were clicking on pictures or whether you were watching the videos.* [Interview 09, registered nurse]

While most professionals highlighted a preference for completing this education as e-learning in the posttest survey, a fifth of the participants considered a face-to-face delivery of the training would have been more appropriate (22/97, 23%; Multimedia Appendix 2). Further explored in the qualitative interviews, participants reported that a blended approach would have enabled an opportunity for further experiential learning in a training environment to aid learning. Participants highlighted the importance of having the opportunity to rehearse difficult family-centered cancer care end-of-life conversations with peers "in a safe space" before integrating within their clinical practice (Q.1.2.3; Multimedia Appendix 2). A participant stated, "...would love a follow up face-to-face, to practice the skill set myself and gain more insight into this topic" (an open-ended response to posttest, registered nurse).

Pillar 2: Professionals Feeling Equipped to Have Family-Centered Supportive End-of-Life Conversations

Overwhelmingly, participants consistently stated that they had an increased awareness of the importance of engaging in these end-of-life conversations with families (Q.2.1.1, Q.2.1.2, and Q.2.1.3; Multimedia Appendix 2). Alongside this, participants reported that they gained knowledge on how and when to have end-of-life conversations after completing the e-learning intervention, with participants highly rating the single-item "increased knowledge" (97/99, 98%; mean 4.56, SD 0.677; Multimedia Appendix 2). Highlighted in the qualitative data, professionals reflected on acquiring new strategies and tools from the "talking, telling, sharing framework" (Q.2.5.1; Multimedia Appendix 2) on how to open conversations with adults regarding the children, and the appropriate language to provide adults with on how best to share the poor prognosis with the children (Q.2.6.1 and Q.2.6.2; Multimedia Appendix 2):

> *I thought it was very clear, because again it was something I wasn't really aware of. Certainly the markers that explained you know the why, where, when, that type of thing. I found that all really useful, because although I won't necessarily always remember it straight off, if I'm in that situation, I have the tools here I can go back to and remember.* [Interview 02, registered nurse]

Participants considered the e-learning intervention relevant to their clinical practice (97/99, 98%; mean 4.68, SD 0.550) and were confident that it would positively impact their provision of family-centered cancer care (97/99, 98%; mean 4.60, SD 0.589; Multimedia Appendix 2). Within the qualitative data, professionals highlighted the positive impact that the e-learning intervention would have on their ability to engage in these conversations during their clinical work, with participants describing feeling more equipped to navigate these conversations and better support families (Q.2.2.1, Q.2.2.2, Q.2.2.3, and Q.2.2.4; Multimedia Appendix 2):

> *It's made me less fearful, because I didn't think that was in my role. I didn't think it was my job, but it is, and now I am very confident to actually do it.* [Interview 12, registered nurse]

In the posttest survey, 96% (93/97) of participants "agreed" or "strongly agreed" that they would recommend the e-learning intervention to a colleague working in cancer or end-of-life care (Multimedia Appendix 2). This converged with the qualitative findings, with participants highlighting that they had recommended the e-learning intervention to colleagues, clinical managers, and senior staff members as they believed it was relevant and beneficial to their clinical workloads (Q.2.7.1 and Q.2.7.2; Multimedia Appendix 2). Of note, 39.9% (63/158) of the participants in the pretest survey indicated their motivations for completing the e-learning intervention were due to a recommendation from a colleague:

> *I went to talk with my supervisor. I believe she could share it with the team you know, or with another team.*

XSL•FO
RenderX

> *Because I think they would benefit from taking this course.* [Interview 08, registered nurse]

In the qualitative data, professionals reflected on their learnings and identified its relevance to different family situations, such as grandparents (Q.2.4.3 and Q.2.4.4; Multimedia Appendix 2), other life-limiting conditions, and cultural contexts outside of the one the intervention was developed for (Q.2.4.1 and Q.2.4.2; Multimedia Appendix 2). Some participants considered the e-learning intervention to be especially suitable for nonspecialist professionals (Q.2.4.5; Multimedia Appendix 2):

> *We come into contact with older people, who have adult children and I would be asked about grandchildren. So it would maybe, for me even to being a bit more tuned in if someone has got grandchildren.* [Interview 02, registered nurse]

As measured by the modified SE-12 scale for measuring clinical communication skills, it was identified that participant's perceptions of self-efficacy to successfully communicate with adults in these situations increased significantly from before (mean 71.25, SE 2.76) to after completion of the e-learning intervention (mean 103.03, SE 1.34; $F_{1,98}$=188.059; P<.001; Multimedia Appendix 2). Those with previous training reported higher scores of self-efficacy in the pretraining survey, compared with those without training. This difference was statistically significant. However, this gap closed in the posttest survey, where both groups were similar (Multimedia Appendix 2).

In total, 99% (96/97) of participants indicated in the posttest survey that they would be "likely" or "very likely" to apply the learning to their clinical role (Multimedia Appendix 2). Further explored in the qualitative findings, some participants provided examples of when they had used their learning in practice since completing the e-learning intervention. This included professionals encouraging adults that it was in the best interest of the children to tell them the reality of the situation, and providing advice and guidance to adults on how to start the conversation with the children (Q.2.3.3; Multimedia Appendix 2):

> *And because I had watched that video, I got the confidence to say to him (adult) "you know what, research does actually prove you are better being honest and open." And he said to me "do you think so?." I said "yes, honestly trust me'." And he came back into me on the Monday, and he said "I need to speak to you." I went oh god it went badly wrong. And he said to me "thank you." I said "for what?" He said "for giving me that advice, because those children and my daughter and everybody was relieved, and we've had the most wonderful weekend.* [Interview 12, registered nurse]

Discussion

Principal Findings

This novel e-learning intervention that is evidence based, theory driven, and free to access provides a much needed and desired educational intervention for professionals to promote end-of-life family-centered cancer conversations [9,38]. The rigorous development of this e-learning intervention was guided by the person-based approach digital intervention development framework [23]. This enabled the incorporation of formative research findings by the team on the support needs of families at the end of life [4,9,39-43], integration of established theories [24,25] with systematic and iterative feedback from end users, and patient and public involvement integrated at all stages of the design and optimization process [26]. Importantly, following engagement with the e-learning intervention, professionals reported an improvement in self-efficacy and enhanced knowledge on the support needs of families at the end of life and felt equipped to have family-centered end-of-life conversations with this population in clinical practice.

The retention rate for the evaluation of this e-learning intervention (99/158, 62.7%) could be considered as positive compared to other nonmandatory, professional e-learning interventions for health and social care professionals [44,45]. Acknowledging the likelihood of a self-motivated population, this study's high retention rate could otherwise be due to incorporating a range of interactive learning tools, accounting for different learning preferences during the course design. Learning tools comprised flip-card activities, expandable icons, introductory and role-play videos with transcripts, and reflective activities. The role of the learning technologists proved integral [46]; contributing essential skills in learning theory, curriculum development, and course design adapted for the web-based environment [47]. Consequently, pertinent content and design features were integrated for visual and aural preferences (ie, educational video resources) and logical preferences (ie, a step-by-step communication framework [42] and content presented chronologically relative to the end-of-life trajectory) [48]. In addition, intrapersonal preferences were accounted for with reflective activities to give adult learners the opportunity to construct their own knowledge [49]. Other factors that may help explain successful retention rates include the relatively short duration of the e-learning intervention (approximately 40 minutes), avoiding cognitive overloading [50], alongside the careful and respectful use of diverse vector illustrations to moderate for content being considered as lacking cultural relevance [26].

Some professionals stated their motivations for participating in the study was a notable caseload increase of adults at the end of life with significant caregiving responsibilities for children. Within the literature, professionals repeatedly report a desire to provide evidence-based care on how best to support adults to prepare their dependent children for this end of life experience [9,51,52]. As cancer rates increases globally [53], professionals will progressively encounter adult patients with parenting responsibilities who have incurable cancer forming a part of their caseload. It is important for professionals to have readily available educational opportunities to improve self-efficacy and equip them (ie, professionals) with knowledge and tools to engage and progress important family-centered end-of-life cancer conversations [9-11]. This will promote protective opportunities for children to be involved in the end-of-life experience [40]; with educational, social, and psychological benefits, before and after bereavement [3,54]. The advantage of this positively evaluated intervention being available in an

XSL•FO
RenderX

e-learning format, in addition to a face-to-face format [14], is the ability to expand, upscale, and disseminate its reach, enhancing global equity [55].

While the posttest survey highlighted a statistically significant increase in scores of self-efficacy, the qualitative data highlighted that professionals would not "shy away from these important conversations" in clinical practice. This is an important finding, in view of the discomfort confronted by professionals caring for families with dependent children at the end of life [9,10]. Consequently, professionals often avoid these challenging conversations; evoked with a sense of helplessness and fear of worsening an emotionally fraught situation [56,57]. However, for most professionals in this study, it was unclear how, or if, skill acquisition to provide family-centered supportive end-of-life care could be inferred from the e-learning intervention. Despite using the most widely cited evaluation framework for educational interventions [29], there remain challenges to determine a causal link between an educational intervention and clinical outcomes. Determining skill acquisition and behavioral application is a well-recognized challenge when evaluating communication skills training [58] and is furthermore acknowledged by our team [14,46]. Of the 4 levels of evaluation within the Kirkpatrick framework (1: reaction; 2: learning; 3: behavior; and 4: results), levels 3 and 4 prove more challenging to report upon, with Kirkpatrick [29] providing limited guidance on how to demonstrate if learning has been applied in practice. Beyond this mixed methods evaluation, there remains a requirement to determine the significance of the training on skill acquisition when professionals are communicating with adults experiencing the end of life, concerning their children [14]. It is therefore considered necessary to evaluate familial impact through the lens of adults and children, to explore if the provision of educational training for professionals has an impact on the care and support received at the end of life.

Of note, the findings highlighted the applicability of the e-learning intervention beyond the intended target audience of adults with cancer who have significant caregiving responsibilities for children; with transferability to those with other close adult-child relational bonds (such as grandparents) and to other life-limiting conditions. Although participants from 14 countries completed the e-learning intervention, it is unclear how applicable and relevant the educational intervention would be for professionals and families from non-Western contexts. However, there are core principles that are important irrespective of cultural context [58]. This includes the following: (1) exploring adults' beliefs, attitudes, and readiness to share the poor prognosis with the children; (2) explaining to adults the benefits of telling the children about the poor prognosis; and (3) children's developmental understanding of illness and death [59,60]. There is the potential for adaptation of this e-learning intervention to specific cultural contexts, to further meet the current global gap of family-centered end-of-life educational interventions for professionals [14]. Despite reported barriers, such as scarce technology and internet access, low digital literacy, and poor infrastructure in low-income and middle-income countries [61], there are emerging data that the world's highest low-income countries are progressing rapidly with regard to implementing e-learning technology [62], closing the digital health care educational divide.

Most professionals highlighted the beneficial learning obtained from viewing, and for some reviewing, the videos within the e-learning intervention that role-played parent-professional communication across the end-of-life trajectory. However, to enable further application of the learning and to problem-solve challenging parent-professional conversations, several participants indicated the desire for experiential and participatory learning in a training environment. Role-play has been advocated as an effective means for professionals to gain confidence in communication skills [63]. Some communication skills training programs offer blended learning with actors or "standardized" patients to help learners gain experience in simulated clinical environments [64]. Nonetheless, the training of the actors or "standardized" patients is costly and time-consuming [65]. Alongside this, individuals who are actors for standardized patient-professional role-play may present biases and experience feelings of anxiety, fatigue, and physical discomfort, albeit with limited long-term negative consequences [66]. With technological advances, there is an opportunity to use artificial intelligence and machine learning to generate natural language within the context of simulated virtual patients, enabling interaction with professional learners for the development of communication skills [67]. The evidence collated from a recent scoping review rated the use of artificial intelligence and machine learning in communication skills training as viable and highly promising [68]. Using digital learning enables the creation of various scenarios that could be customized to specific real-life situations [69], such as adult-professional end-of-life communication. Professionals have the potential to benefit from practicing simulated patient encounters that are tailored to the challenges they confront in routine clinical practice and gain real-time feedback.

Study Strengths and Limitations

At project commencement, the research team considered collecting engagement and preference data analytics with a desire to gain an understanding of patterns, trends, and interactions with the e-learning intervention, such as the amount of time spent on each section, where individuals dropped off, and if users returned to sections. Taking into consideration the data analytics process with the learning technologists, the content would have required development as "microlearning blocks" on Rise360. This would incur additional development costs, necessitating 5 "microlearning blocks," alongside the requirement of learners to "scroll and click" at multiple points to progress through sections within the e-learning intervention. Following user consultation, this concept was not pursued to reduce navigational burden and user fatigue. Furthermore, the team took cognizance of the rigorous and rich learning obtained from the iterative user-testing process undertaken in adapting and optimizing the e-learning resource with end users, which provided data on engagement and preferences [26]. In addition, the planned qualitative element of this mixed methods outcome evaluation enabled the collection of usability and acceptability data. Consequently, the e-learning intervention was developed as one sharable content object reference model (SCORM) package.

XSL•FO
RenderX

The e-learning intervention was hosted on the All-Ireland Institute of Hospice and Palliative Care educational hub, being a leading organization with national and international influence and promoting excellence in palliative care. Our team established a collaborative agreement for hosting the e-learning intervention from study inception (ie, grant writing), which proved valuable in promoting the professional integrity of the educational intervention, mitigating platform hosting costs, and enabling longevity of the intervention beyond the study period. A potential drawback of having the e-learning intervention hosted and embedded on an organizational learning platform is professionals who search on Google or Bing (which accounts for 93% of all internet searches [70]) for educational interventions are less likely to identify embedded courses on an initial search, which could impact the reach of the intervention. Platform hosting costs, reach, and infrastructure should be important considerations for funders and researchers of e-learning interventions at project commencement.

Conclusions

Engagement with this coproduced, theory-driven, and evidence-based e-learning educational intervention demonstrates positive improvements in professionals' self-efficacy and preparedness to engage in supportive parent-professional end-of-life cancer care conversations, when an adult with significant caregiving responsibilities is dying. Having the educational intervention in a freely available e-learning format has the potential to improve reach at a national and international level, bridging the recognized global gap of educational interventions in this field. Findings indicated applicability of the e-learning intervention beyond the intended target audience of adults with cancer who have significant caregiving responsibilities for children, with transferability to those with other close adult-child relational bonds and to other life-limiting conditions. Equipping professionals on family-centered supportive care has the potential to promote a better end-of-life experience for the family and mediate adverse outcomes before and after bereavement.

Acknowledgments

The authors express their sincere thanks to all participants, steering group members, bereaved parents, and children who willingly gave their time to be involved in and support this project. They extend their thanks to the All-Ireland Institute of Hospice and Palliative Care for hosting the e-learning intervention and their continued support to the family-centered cancer care program of work.

The authors would like to acknowledge and thank the Higher Education Authority North-South Research Programme, who generously funded this project.

Data Availability

The dataset that supports the study findings is available at the University College Dublin Repository and available upon reasonable request from the first (CJS) and second (CON) authors (principal investigators of the study). The data are not publicly available due to ethical restrictions, as detailed in the ethics committee approval from Ulster University (FCNUR-23-002) and University College Dublin (LS-22-65).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Modified Self-Efficacy Questionnaire tool.
[DOCX File , 17 KB - jmir_v26i1e65619_app1.docx]

Multimedia Appendix 2
FCCCEoLEd e-learning Evaluation Pillar Integration Process with joint display of quantitative (pretest and posttest surveys) and qualitative data.
[DOCX File , 25 KB - jmir_v26i1e65619_app2.docx]

References

1. Inhestern L, Bultmann JC, Johannsen LM, Beierlein V, Möller B, Romer G, et al. Estimates of prevalence rates of cancer patients with children and well-being in affected children: a systematic review on population-based findings. Front Psychiatry 2021 Nov 25;12:765314 [FREE Full text] [doi: 10.3389/fpsyt.2021.765314] [Medline: 34899425]
2. Johannsen L, Frerichs W, Inhestern L, Bergelt C. Exploring the perspectives of cancer patients parenting minor children: a qualitative study on family-centered cancer care experiences. Patient Educ Couns 2023 Dec;117:107989. [doi: 10.1016/j.pec.2023.107989] [Medline: 37812964]
3. Ellis J, Dowrick C, Lloyd-Williams M. The long-term impact of early parental death: lessons from a narrative study. J R Soc Med 2013 Feb;106(2):57-67 [FREE Full text] [doi: 10.1177/0141076812472623] [Medline: 23392851]

XSL•FO
RenderX

4. Hanna JR, McCaughan E, Semple CJ. Challenges and support needs of parents and children when a parent is at end of life: a systematic review. Palliat Med 2019 Sep;33(8):1017-1044 [FREE Full text] [doi: 10.1177/0269216319857622] [Medline: 31244381]
5. Sheehan S, Hanna JR, Drury A, McCance T, Semple CJ, O'Neill C. A systematic review of educational interventions to equip health and social care professionals to promote end-of-life supportive care when a parent with dependent children is dying with cancer. Semin Oncol Nurs 2023 Oct;39(5):151474 [FREE Full text] [doi: 10.1016/j.soncn.2023.151474] [Medline: 37481410]
6. Milbury K, Ann-Yi S, Jones M, Li Y, Whisenant M, Yousuf S, et al. Patients with advanced cancer and their spouses parenting minor children: the role of the relationship context in parenting concerns. Psychooncology 2024 Mar 27;33(3):e6310. [doi: 10.1002/pon.6310] [Medline: 38411282]
7. Høeg BL, Christensen J, Banko L, Frederiksen K, Appel CW, Dalton SO, et al. Psychotropic medication among children who experience parental death to cancer. Eur Child Adolesc Psychiatry 2023 Jan;32(1):155-165 [FREE Full text] [doi: 10.1007/s00787-021-01846-y] [Medline: 34302529]
8. Syse A, Aas GB, Loge JH. Children and young adults with parents with cancer: a population-based study. Clin Epidemiol 2012;4:41-52 [FREE Full text] [doi: 10.2147/CLEP.S28984] [Medline: 22442635]
9. Hanna JR, McCaughan E, Beck ER, Semple CJ. Providing care to parents dying from cancer with dependent children: health and social care professionals' experience. Psychooncology 2021 Mar;30(3):331-339 [FREE Full text] [doi: 10.1002/pon.5581] [Medline: 33091180]
10. Franklin P, Arber A, Ream E. Priorities for enhancing nurses' and social workers' competence and confidence in helping families support dependent children through parental death. A classic-Delphi survey. BMC Palliat Care 2024 May 17;23(1):122 [FREE Full text] [doi: 10.1186/s12904-024-01452-0] [Medline: 38760809]
11. Dalton LJ, McNiven A, Hanna JR, Rapa E. Exploring healthcare professionals' beliefs, experiences and opinions of family-centred conversations when a parent has a serious illness: a qualitative study. PLoS One 2022;17(11):e0278124 [FREE Full text] [doi: 10.1371/journal.pone.0278124] [Medline: 36441706]
12. Hanna JR, Rapa E, Miller M, Turner M, Dalton LJ. Conversations about children when an important adult is at end of life: an audit. Am J Hosp Palliat Care 2022 Jul;39(7):806-811 [FREE Full text] [doi: 10.1177/10499091211046241] [Medline: 34538106]
13. Franklin P, Arber A, Reed L, Ream E. Health and social care professionals' experiences of supporting parents and their dependent children during, and following, the death of a parent: a qualitative review and thematic synthesis. Palliat Med 2019 Jan;33(1):49-65 [FREE Full text] [doi: 10.1177/0269216318803494] [Medline: 30371147]
14. Hanna JR, Semple CJ. Mixed-methods evaluation of a face-to-face educational intervention for health and social care professionals to deliver family-centred cancer supportive care when a parent with dependent children is at end of life. Psychooncology 2024 Jul;33(7):e6374 [FREE Full text] [doi: 10.1002/pon.6374] [Medline: 38977423]
15. Dalton L, Rapa E, Stein A. Protecting the psychological health of children through effective communication about COVID-19. Lancet Child Adolesc Health 2020 May;4(5):346-347 [FREE Full text] [doi: 10.1016/s2352-4642(20)30097-3]
16. Meinert E, Eerens J, Banks C, Maloney S, Rivers G, Ilic D, et al. Exploring the cost of eLearning in health professions education: scoping review. JMIR Med Educ 2021 Mar 11;7(1):e13681 [FREE Full text] [doi: 10.2196/13681] [Medline: 33704073]
17. Bingham SL, Semple CJ, Flannagan C, Dunwoody L. Enhancing healthcare professional-led sexual support in cancer care: acceptability and usability of an eLearning resource and its impact on attitudes towards providing sexual support. Psychooncology 2022 Sep;31(9):1555-1563 [FREE Full text] [doi: 10.1002/pon.5993] [Medline: 35781720]
18. Kimura R, Matsunaga M, Barroga E, Hayashi N. Asynchronous e-learning with technology-enabled and enhanced training for continuing education of nurses: a scoping review. BMC Med Educ 2023 Jul 13;23(1):505 [FREE Full text] [doi: 10.1186/s12909-023-04477-w] [Medline: 37442970]
19. Orr F, Kelly M, Virdun C, Power T, Phillips A, Gray J. The development and evaluation of an integrated virtual patient case study and related online resources for person-centred nursing practice. Nurse Educ Pract 2021 Feb;51:102981 [FREE Full text] [doi: 10.1016/j.nepr.2021.102981] [Medline: 33610969]
20. Darley A, Carroll Á. Conducting co-design with older people in a digital setting: methodological reflections and recommendations. Int J Integr Care 2022;22(4):18 [FREE Full text] [doi: 10.5334/ijic.6546] [Medline: 36561633]
21. Yardley S, Williams H, Bowie P, Edwards A, Noble S, Donaldson L, et al. Which human factors design issues are influencing system performance in out-of-hours community palliative care? Integration of realist approaches with an established systems analysis framework to develop mid-range programme theory. BMJ Open 2022 Jan 03;12(1):e048045 [FREE Full text] [doi: 10.1136/bmjopen-2020-048045] [Medline: 34980606]
22. Brooks CF, Lund S, Kryl D, Lloyd Jones S, Myall M. "We all see things through a different lens based on our life experiences": co-production of a web-based implementation toolkit with stakeholders across the health and social care system. Front Health Serv 2024;4:1356961 [FREE Full text] [doi: 10.3389/frhs.2024.1356961] [Medline: 38812599]
23. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]

XSL•FO
RenderX

24. Walsh F. Family resilience: a developmental systems framework. Eur J Dev Psychol 2016 Mar 02;13(3):313-324 [FREE Full text] [doi: 10.1080/17405629.2016.1154035]
25. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26 [FREE Full text] [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
26. O'Neill C, Hanna JR, Sheehan S, McCance T, Drury A, Semple CJ. Adapting and testing an eLearning resource for professionals to support families when a significant caregiver for children is dying with cancer. BMC Palliat Care 2024 Nov 21;23(1):268 [FREE Full text] [doi: 10.1186/s12904-024-01601-5] [Medline: 39574104]
27. Yardley L, Ainsworth B, Arden-Close E, Muller I. The person-based approach to enhancing the acceptability and feasibility of interventions. Pilot Feasibility Stud 2015;1:37 [FREE Full text] [doi: 10.1186/s40814-015-0033-z] [Medline: 27965815]
28. O'Connor SR, Flannagan C, Parahoo K, Steele M, Thompson S, Jain S, et al. Efficacy, use, and acceptability of a web-based self-management intervention designed to maximize sexual well-being in men living with prostate cancer: single-arm experimental study. J Med Internet Res 2021 Jul 26;23(7):e21502 [FREE Full text] [doi: 10.2196/21502] [Medline: 34309580]
29. Kirkpatrick DL. The four levels of evaluation. In: Brown SM, Seidner CJ, editors. Evaluating Corporate Training: Models and Issues. Dordrecht, The Netherlands: Springer; 1998:95-112.
30. Kang H. Sample size determination and power analysis using the G*Power software. J Educ Eval Health Prof 2021;18:17 [FREE Full text] [doi: 10.3352/jeehp.2021.18.17] [Medline: 34325496]
31. Axboe MK, Christensen KS, Kofoed PE, Ammentorp J. Development and validation of a self-efficacy questionnaire (SE-12) measuring the clinical communication skills of health care professionals. BMC Med Educ 2016 Oct 18;16(1):272 [FREE Full text] [doi: 10.1186/s12909-016-0798-7] [Medline: 27756291]
32. Johnson RE, Grove AL, Clarke A. Pillar integration process: a joint display technique to integrate data in mixed methods research. J Mixed Methods Res 2017 Dec 01;13(3):301-320 [FREE Full text] [doi: 10.1177/1558689817743108]
33. Drury A, Goss J, Afolabi J, McHugh G, O'Leary N, Brady AM. A mixed methods evaluation of a pilot multidisciplinary breathlessness support service. Eval Rev 2023 Apr 04;47(5):820-870 [FREE Full text] [doi: 10.1177/0193841x231162402]
34. Sheehan S, Dowling M, Shewbridge A, Diez de Los Rios la Serna C, Bağçivan G, Kotronoulas G, et al. Digital Education in Healthcare: A Mixed-Methods Evaluation of the Advanced Breast Cancer for Nurses (ABC4Nurses) eLearning Programme. Nurse Education Today 2024. [doi: 10.1016/j.nedt.2024.106533]
35. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597 [FREE Full text] [doi: 10.1080/2159676x.2019.1628806]
36. Braun V, Clarke V. Supporting best practice in reflexive thematic analysis reporting in Palliative Medicine: a review of published research and introduction to the Reflexive Thematic Analysis Reporting Guidelines (RTARG). Palliat Med 2024 Jun;38(6):608-616 [FREE Full text] [doi: 10.1177/02692163241234800] [Medline: 38469804]
37. Braun V, Clarke V, Hayfield N, Davey L, Jenkinson E. Doing reflexive thematic analysis. In: McBeath A, Bager-Charleson S, editors. Supporting Research in Counselling and Psychotherapy: Qualitative, Quantitative, and Mixed Methods Research. Cham, Switzerland: Springer; 2023:19-38.
38. Cockle-Hearne J, Reed E, Todd J, Ream E. The dying parent and dependent children: a nationwide survey of hospice and community palliative care support services. BMJ Support Palliat Care 2022 Nov;12(e5):e696-e704 [FREE Full text] [doi: 10.1136/bmjspcare-2019-001947] [Medline: 32152037]
39. Hanna JR, McCaughan E, Semple CJ. Immediate bereavement experiences when a parent of dependent children has died of cancer: funeral directors' perspectives. Death Stud 2022;46(4):969-978 [FREE Full text] [doi: 10.1080/07481187.2020.1793433] [Medline: 32678718]
40. McCaughan E, Semple CJ, Hanna JR. 'Don't forget the children': a qualitative study when a parent is at end of life from cancer. Support Care Cancer 2021 Dec;29(12):7695-7702 [FREE Full text] [doi: 10.1007/s00520-021-06341-3] [Medline: 34143326]
41. Semple CJ, McCaughan E, Beck ER, Hanna JR. 'Living in parallel worlds' - bereaved parents' experience of family life when a parent with dependent children is at end of life from cancer: a qualitative study. Palliat Med 2021 May;35(5):933-942 [FREE Full text] [doi: 10.1177/02692163211001719] [Medline: 33765868]
42. Semple CJ, McCaughan E, Smith R, Hanna JR. Parent's with incurable cancer: 'nuts and bolts' of how professionals can support parents to communicate with their dependent children. Patient Educ Couns 2022 Mar;105(3):775-780 [FREE Full text] [doi: 10.1016/j.pec.2021.06.032] [Medline: 34294491]
43. Hanna JR, Semple CJ. 'I didn't know what was in front of me'-bereaved parents' experience of adapting to life when a co-parent of dependent children has died with cancer. Psychooncology 2022 Oct;31(10):1651-1659 [FREE Full text] [doi: 10.1002/pon.6010] [Medline: 35949162]
44. Degerfält J, Sjöstedt S, Fransson P, Kjellén E, Werner MU. E-learning programs in oncology: a nationwide experience from 2005 to 2014. BMC Res Notes 2017 Jan 13;10(1):39 [FREE Full text] [doi: 10.1186/s13104-017-2372-8] [Medline: 28086959]
45. Gillan C, Papadakos J, Brual J, Harnett N, Hogan A, Milne E, et al. Impact of high-fidelity e-learning on knowledge acquisition and satisfaction in radiation oncology trainees. Curr Oncol 2018 Dec 01;25(6):e533-e538 [FREE Full text] [doi: 10.3747/co.25.4090] [Medline: 30607120]

XSL•FO
RenderX

46. Dowling M, Shewbridge A, Ryan C, Clancy C, Meade E, Sheehan S, et al. Development and implementation of an online education program on advanced breast cancer for European cancer nurses: ABC4Nurses project: a brief report. J Cancer Educ 2023 Oct;38(5):1662-1666 [FREE Full text] [doi: 10.1007/s13187-023-02319-3] [Medline: 37336799]
47. Park SY, Kim JH. Instructional design and educational satisfaction for virtual environment simulation in undergraduate nursing education: the mediating effect of learning immersion. BMC Med Educ 2022 Sep 12;22(1):673 [FREE Full text] [doi: 10.1186/s12909-022-03728-6] [Medline: 36096798]
48. Rasheed F, Wahid A. Learning style detection in e-learning systems using machine learning techniques. Expert Syst Appl 2021 Jul;174:114774 [FREE Full text] [doi: 10.1016/j.eswa.2021.114774]
49. Yilmaz FG, Keser H. The impact of reflective thinking activities in e-learning: a critical review of the empirical research. Comput Educ 2016 Apr;95:163-173 [FREE Full text] [doi: 10.1016/j.compedu.2016.01.006]
50. Bawa P. Retention in online courses: exploring issues and solutions—a literature review. Sage Open 2016 Jan 05;6(1) [FREE Full text] [doi: 10.1177/2158244015621777]
51. Wray A, Seymour J, Greenley S, Boland JW. Parental terminal cancer and dependent children: a systematic review. BMJ Support Palliat Care 2022 Jan 28:bmjspcare-2021-003094 [FREE Full text] [doi: 10.1136/bmjspcare-2021-003094] [Medline: 35091436]
52. Dieperink KB, Semple C. Parental cancer in young families - a unique opportunity for family-centered care. Semin Oncol Nurs 2023 Oct;39(5):151485 [FREE Full text] [doi: 10.1016/j.soncn.2023.151485] [Medline: 37604761]
53. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
54. Birgisdóttir D, Grenklo TB, Kreicbergs U, Steineck G, Fürst CJ, Kristensson J. Family cohesion predicts long-term health and well-being after losing a parent to cancer as a teenager: a nationwide population-based study. PLoS One 2023;18(4):e0283327 [FREE Full text] [doi: 10.1371/journal.pone.0283327] [Medline: 37043474]
55. McNutt M. Higher education for all. Science 2022 Nov 11;378(6620):579 [FREE Full text] [doi: 10.1126/science.adf6263] [Medline: 36356153]
56. Wickramasinghe LM, Yeo ZZ, Chong PH, Johnston B. Communicating with young children who have a parent dying of a life-limiting illness: a qualitative systematic review of the experiences and impact on healthcare, social and spiritual care professionals. BMC Palliat Care 2022 Jul 12;21(1):125 [FREE Full text] [doi: 10.1186/s12904-022-01007-1] [Medline: 35820910]
57. Fearnley R, Boland JW. Communication and support from health-care professionals to families, with dependent children, following the diagnosis of parental life-limiting illness: a systematic review. Palliat Med 2017 Mar;31(3):212-222 [FREE Full text] [doi: 10.1177/0269216316655736] [Medline: 27383635]
58. Mata AN, de Azevedo KP, Braga LP, de Medeiros GC, de Oliveira Segundo VH, Bezerra IN, et al. Training in communication skills for self-efficacy of health professionals: a systematic review. Hum Resour Health 2021 Mar 06;19(1):30 [FREE Full text] [doi: 10.1186/s12960-021-00574-3] [Medline: 33676515]
59. Rapa E, Hanna JR, Pollard T, Santos-Paulo S, Gogay Y, Ambler J, et al. Exploring the experiences of healthcare professionals in South Africa and Uganda around communicating with children about life-threatening conditions: a workshop-based qualitative study to inform the adaptation of communication frameworks for use in these settings. BMJ Open 2023 Jan 27;13(1):e064741 [FREE Full text] [doi: 10.1136/bmjopen-2022-064741] [Medline: 36707115]
60. Chandratre S, Soman A. End of life care practices for Hindu patients during COVID-19. J Palliat Care 2022 Jan;37(1):3-7 [FREE Full text] [doi: 10.1177/08258597211036243] [Medline: 34397284]
61. Reis FJ, Fernandes LG, Saragiotto BT. Telehealth in low- and middle-income countries: bridging the gap or exposing health disparities? Health Policy Technol 2021 Dec;10(4):100577 [FREE Full text] [doi: 10.1016/j.hlpt.2021.100577] [Medline: 34745854]
62. Global diffusion of eHealth: making universal health coverage achievable. World Health Organization. URL: https://www.afro.who.int/publications/global-diffusion-ehealth-making-universal-health-coverage-achievable [accessed 2024-07-20]
63. Gartmeier M, Bauer J, Fischer MR, Hoppe-Seyler T, Karsten G, Kiessling C, et al. Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. Instr Sci 2015 Apr 7;43(4):443-462 [FREE Full text] [doi: 10.1007/s11251-014-9341-6]
64. Cappi V, Artioli G, Ninfa E, Ferrari S, Guarnieri MC, Martucci G, et al. The use of blended learning to improve health professionals' communication skills: a literature review. Acta Biomed 2019 Mar 28;90(4-S):17-24 [FREE Full text] [doi: 10.23750/abm.v90i4-S.8330] [Medline: 30977745]
65. Stuij SM, Drossaert CH, Labrie NH, Hulsman RL, Kersten MJ, van Dulmen S, et al. Developing a digital training tool to support oncologists in the skill of information-provision: a user centred approach. BMC Med Educ 2020 May 01;20(1):135 [FREE Full text] [doi: 10.1186/s12909-020-1985-0] [Medline: 32357886]
66. Plaksin J, Nicholson J, Kundrod S, Zabar S, Kalet A, Altshuler L. The benefits and risks of being a standardized patient: a narrative review of the literature. Patient 2016 Feb;9(1):15-25 [FREE Full text] [doi: 10.1007/s40271-015-0127-y] [Medline: 26002043]

XSL•FO
RenderX

67. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. J Med Internet Res 2019 Oct 29;21(10):e14658 [FREE Full text] [doi: 10.2196/14658] [Medline: 31663857]
68. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25:e43311 [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]
69. Anbro SJ, Szarko AJ, Houmanfar RA, Maraccini AM, Crosswell LH, Harris FC, et al. Using virtual simulations to assess situational awareness and communication in medical and nursing education: a technical feasibility study. J Organ Behav Manage 2020 Apr 02;40(1-2):129-139 [FREE Full text] [doi: 10.1080/01608061.2020.1746474]
70. Theuring J, Barnard K. Bing vs Google: search engine comparison. Impression. 2024 Aug 7. URL: https://www.impressiondigital.com/blog/bing-differ-google/ [accessed 2024-11-22]

Abbreviations

SCORM: sharable content object reference model
SE-12: self-efficacy questionnaire

Edited by A Mavragani; submitted 20.08.24; peer-reviewed by D Sheehan, Y Chu; comments to author 30.09.24; revised version received 01.10.24; accepted 05.11.24; published 10.12.24.

Please cite as:
Semple CJ, O'Neill C, Sheehan S, McCance T, Drury A, Hanna JR
An e-Learning Intervention for Professionals to Promote Family-Centered Cancer Care When a Significant Caregiver for Children Is at End of Life: Mixed Methods Evaluation Study
J Med Internet Res 2024;26:e65619
URL: https://www.jmir.org/2024/1/e65619
doi:10.2196/65619
PMID:

©Cherith Jane Semple, Carla O'Neill, Sarah Sheehan, Tanya McCance, Amanda Drury, Jeffrey R Hanna. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of a Rice-Farming Simulation Video Game on Nature Relatedness, Nutritional Status, and Psychological State in Urban-Dwelling Adults During the COVID-19 Pandemic: Randomized Waitlist Controlled Trial

Seulki Lee[1*], MA; Chisung Yuh[2*], MS, DDS; Yu-Bin Shin[1], PhD; Heon-Jeong Lee[1], MD, PhD; Young-Mee Lee[2], MD, PhD; Jungsil Lee[3], MD, MPH; Chul-Hyun Cho[1,4], MD, PhD

[1]Department of Psychiatry, Korea University College of Medicine, Seoul, Republic of Korea
[2]Department of Medical Education, Korea University College of Medicine, Seoul, Republic of Korea
[3]Department of Environmental Medicine, College of Medicine, Ewha Womans University, Seoul, Republic of Korea
[4]Department of Biomedical Informatics, Korea University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Chul-Hyun Cho, MD, PhD
Department of Psychiatry, Korea University College of Medicine
73 Goryeodae-ro Seongbuk-gu
Seoul, 02841
Republic of Korea
Phone: 82 0221993827
Email: david0203@gmail.com

Abstract

Background: During the COVID-19 pandemic, urban inhabitants faced significant challenges in maintaining connections with nature, adhering to nutritional guidelines, and managing mental well-being.

Objective: Recognizing the urgent need for innovative approaches, this study was designed to explore the potential benefits of a specific digital intervention, the rice-farming simulation game *Sakuna: Of Rice and Ruin*, for nature relatedness, nutritional behaviors, and psychological well-being.

Methods: A total of 66 adults without any prior major psychiatric disorders residing in an urban area were recruited for the study. They were randomly assigned to 2 groups through block randomization: the immediate intervention group (IIG; 34/66, 52%) and the waitlist group (32/66, 48%). Participants in the IIG were instructed to play the game for at least 4 days per week for 3 weeks, with each session lasting from 30 minutes to 3 hours. Assessments were performed at baseline, week 1, and week 3. The Nature Relatedness Scale (NR) and Nutrition Quotient Scale were used to evaluate nature relatedness and nutritional state, respectively. Furthermore, psychological state was assessed using the World Health Organization Quality of Life–Brief Version (WHOQOL-BREF), Brief Fear of Negative Evaluation Scale, Social Avoidance and Distress Scale, Toronto Alexithymia Scale, State-Trait Anxiety Inventory, Center for Epidemiologic Studies Depression Scale Revised, and Korean Resilience Quotient.

Results: This study's results revealed significant time interactions between the IIG and waitlist group for both the total NR score (P=.001) and the score of the *self* subdomain of NR (P<.001), indicating an impact of the game on nature relatedness. No group×time interactions were found for the total Nutrition Quotient Scale and subdomain scores, although both groups showed increases from baseline. For psychological state, a significant group×time interaction was observed in the total WHOQOL-BREF score (P=.049), suggesting an impact of the game on quality of life. The psychological (P=.01), social (P=.003), and environmental (P=.04) subdomains of the WHOQOL-BREF showed only a significant time effect. Other psychological scales did not display any significant changes (all P>.05).

Conclusions: Our findings suggest that the rice-farming game intervention might have positive effects on nature relatedness, nature-friendly dietary behaviors, quality of life, anxiety, depression, interpersonal relationships, and resilience among urban adults during the COVID-19 pandemic. The impact of pronature games in confined urban environments provides valuable evidence of how digital technologies can be used to enhance urban residents' affinity for nature and psychological well-being. This understanding can be extended in the future to other digital platforms, such as metaverses.

XSL•FO
RenderX

Trial Registration: Clinical Research Information Service (CRIS) KCT0007657; http://tinyurl.com/yck7zxp7

(J Med Internet Res 2024;26:e51596) doi:10.2196/51596

KEYWORDS

video game; digital intervention; nature relatedness; nutritional status; psychological state; COVID-19; urban-dwelling adults

Introduction

Background

Urbanization and industrialization have reshaped societies all over the world, with a substantial fraction of populations now residing in city centers [1]. Urban lifestyles, although convenient, may divorce individuals from nature and promote behaviors that can negatively impact both physical and mental health [1,2]. Recognizing the importance of nature for human well-being, it is important for research to be conducted using an integrated approach to examine the effects of urban living on health promotion behaviors, nature relatedness, nutritional status, and psychological states [3-5].

Urban living can engender lifestyle habits that impede connection with nature and hinder balanced nutrition [6]. Many urban dwellers lack access to green spaces, leading to a significant detachment from nature and possibly unhealthy eating habits. The convenience and ubiquity of fast food in urban areas, coupled with limited access to fresh produce, may lead to diets that are high in fat, sugar, and calories. Combined with the sedentary habits required by city living, the lack of connection with nature and poor nutrition escalate the risks of obesity, diabetes, and other health issues [6-8].

The stressors of urban living, social isolation, and alienation from nature can have severe psychological consequences. Restricted access to green spaces contributes to feelings of estrangement from nature, exacerbating stress, anxiety, and depression and undermining the overall well-being. Indeed, studies have shown that exposure to nature can enhance feelings of well-being and mitigate stress [9,10].

The COVID-19 pandemic and the ensuing social distancing policies amplified urban dwellers' physical and social isolation, increased negative dietary patterns, decreased outdoor physical activities, and escalated psychological distress [11-13]. A rapid rise in obesity and metabolic diseases were associated with such changes in dietary patterns in previously healthy young adults in South Korea during the COVID-19 pandemic [14]. Moreover, mental health, which is highly correlated with obesity and metabolic health, became a serious issue during the pandemic for similar reasons, such as individual isolation measures in closed spaces and detachment from the surrounding community and nature [14]. Because future pandemics are a real possibility, urban dwellers' susceptibility necessitates the development of strategies to promote their physical and psychological well-being [15,16]. These efforts are not only important for pandemic preparedness but also vital interventions in the modern, urban-centric society.

The video game industry, which saw an expansion during the era of social distancing, presents an opportunity to address urban dwellers' disconnection from nature and associated health issues [17]. The World Health Organization has suggested that video games may have positive effects for alleviating stress and social isolation [18]. Recently, games and game-like software have been deployed in clinical contexts for health promotion and disease treatment, as well as to promote psychological well-being, highlighting their potential as digital therapeutics [19].

Gamification has become increasingly prevalent in health interventions, capitalizing on the persuasive capabilities of technology to effectuate positive behavior changes [20]. By implementing game mechanics, including points and levels, in nongame contexts, this strategy increases participation and encourages behavior modification [21]. A substantial proportion of the examined studies yielded inconclusive results, indicating that although gamification may be beneficial, its effects are not universally favorable and may differ depending on the context and quality of implementation. Although the literature frequently discusses short-term engagement, the results regarding long-term impact and intrinsic motivation are inconsistent, necessitating further investigation into the ways in which extrinsic motivators can evolve into sustained intrinsic motivation [21].

Sakuna: Of Rice and Ruin, a rice-farming action role-playing game, provides a unique blend of entertainment and cultural education [22]. Players assume the role of Sakuna, a harvest goddess who has been banished to an island where she must battle demons and cultivate rice. The game introduces players to the intricacies of traditional rice farming and ties these agricultural practices directly to the character's progression. We believe that this immersive experience could enhance players' connection with nature, even in a game-based setting, and provide psychological benefits through exposure to slow and methodical tasks.

As mentioned earlier, healthier food choices, such as traditional homemade rice-based meals in South Korea, could lead to a decrease in obesity and metabolic and mental health issues [14]. Rice is the staple food for more than half of the world's population, and its cultivation plays an important role in Asian culture. In South Korea and Japan, rice is deeply associated with cultural identity, memories of family, and the community [23]. Healthy homemade meals are mostly perceived to be rice-based meals, which tend to be more balanced, higher in nutrients, and lower in calories than Western foods; several studies have been published on the association of rice-centered Japanese diet with the quality of life, sleep, impulsivity, depressive symptoms, and biochemical changes [24,25].

This Study

Therefore, we hypothesized that engaging in this rice-farming game would lead to a positive perception of nature and a rice-based balanced diet, which could subsequently affect actual food choices. Improved nature relatedness could positively

XSL•FO
RenderX

influence psychological well-being and mood, which, in turn, could also help in choosing a healthier diet in real life. This study aims to explore the potential impacts of playing *Sakuna: Of Rice and Ruin* on nature relatedness, nutritional status, and psychological well-being among urban dwellers during the COVID-19 pandemic, especially during the period of social distancing. We investigated the game's ability to provide a meaningful and accessible connection with nature and improve nutritional behavior and its potential influence on players' mental health, focusing on the game's realism and immersive content.

Methods

Participants

Our study used a recruitment strategy involving web-based bulletin boards of social network services and gaming communities in South Korea. The web-based recruitment process included posting institutional review board–approved announcements on our laboratory's official website and sharing the link to this page on YouTube (Google LLC), Naver (Naver Corp), Twitch (Amazon, Inc), and 8 other game-related web-based platforms. From those who applied, we included only young adults aged 19 to 40 years with a minimum residency period of 6 months in the urban area at the time of study enrollment. The exclusion criteria included individuals diagnosed with major psychiatric disorders, game addiction, mental retardation, traumatic brain injury, or epilepsy or those who had used the game *Sakuna: Of Rice and Ruin* within the last 6 months.

Once potential participants were identified, they were sent a brief video explaining research consent requirements via email. Subsequently, an electronic signature of the web-based informed consent form was obtained from all participants. Of the 71 individuals screened, 68 (96%) agreed to participate and provided their informed consent. We used block randomization to assign participants equally to the immediate intervention group (IIG) and waitlist group (WLG), each of which initially included 34 members. However, 2 (6%) out of 34 individuals from the WLG could not be contacted before the start of the study. Therefore, a final total of 32 participants were in the WLG, resulting in 66 participants overall.

The study required participants to engage in the *Sakuna: Of Rice and Ruin* game under specific conditions. Participants were asked to use the game's PC version if possible, although the use of the Nintendo Switch (Nintendo Co, Ltd) and PlayStation (Sony Interactive Entertainment) versions was also permitted. The game had a total running time of 35 hours, and participants were expected to play the game for at least 4 days per week for 3 weeks, with each session lasting from 30 minutes to 3 hours in length. We set a maximum daily playtime of 5 hours to prevent potential game addiction, with a strongly recommended 10-minute break time. An additional week was granted to any participant who was unable to complete the game within the initial 3 weeks.

Participants were required to purchase the game independently for the purpose of the study, thus ensuring that each held individual copyright. The researchers did not provide the game to the participants. In addition, the study design involved real-time remote monitoring of gaming activities. The research team observed participants' in-game progress through instant messaging platforms, where participants were instructed to share snapshots of the game's start, pause, and finish screens as well as any achievements obtained, to track their engagement with the game.

For the IIG, the gaming intervention took place in April 2021; the first assessment occurred before the game-playing phase, and subsequent evaluations occurred at the end of weeks 1 and 3. The WLG continued their usual daily activities for the first 2 weeks, starting the game only after this period. They also underwent assessments before the gaming phase and at the end of weeks 1 and 3. All the assessments were conducted via a web-based survey site. Figure 1 presents the schematic research flow with evaluations and procedures.

XSL•FO
RenderX

Figure 1. Schematic research flow. Comparison of the waitlist group and immediate intervention group (white), and game intervention in the entire sample (gray).

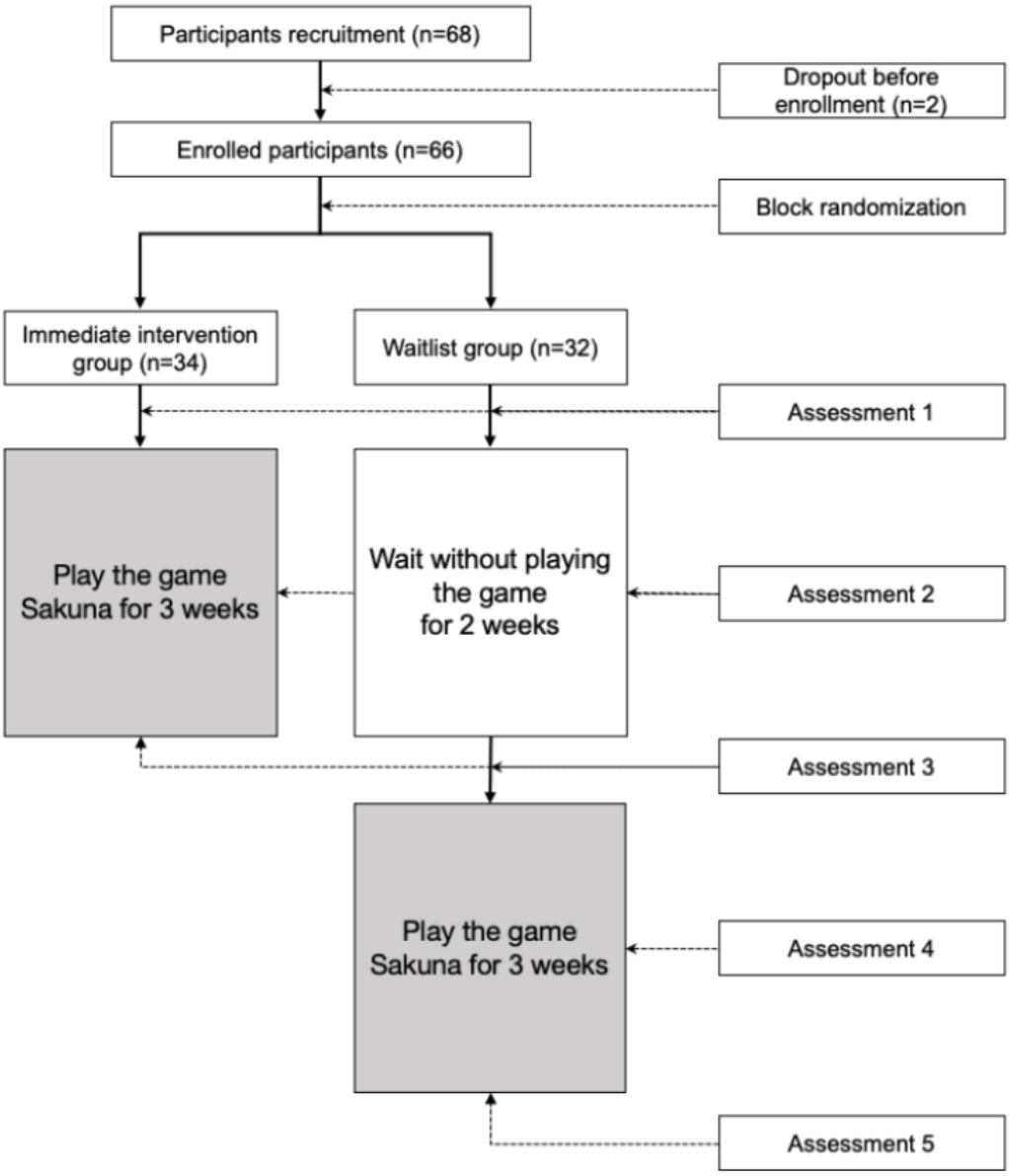

A Rice-Farming Video Game Digital Intervention (Sakuna: Of Rice and Ruin)

Overview

Sakuna: Of Rice and Ruin (Marvelous Inc) is an action role-playing simulation game that provides a distinctive blend of agriculture and action gameplay, thus offering a unique gaming experience. The game is divided into 2 primary sections: an agricultural simulator, in which players must cultivate rice, and an action segment, in which players must combat demonic creatures using farming tools (Figure 2). The agrarian component spans several in-game seasons and involves multiple stages. These multiple stages range from tilling the soil and transplanting rice to adding fertilizer to managing water levels to weeding to eventually harvesting and threshing rice (Multimedia Appendices 1 and 2).

Each successful rice harvest permanently enhances the character's abilities. The quality of the harvest can be improved through meticulous crop management. Players who complete each task access new skills that expedite and enhance the cultivation process. The action element of the game features combat against demonic animals, for which rice harvests bolster players' strength. Set in nature and modeled around real-world farming practices, it is reasonable to assume that the game may be influenced by the user's connection with nature, nutritional status, and psychosocial state.

In addition to rice farming, the game emphasizes the importance of traditional family dinners and a well-balanced diet. Sakuna, with her human friends, such as family, makes her own healthy daily dinner from the available natural ingredients (gathered, hunted, harvested, or stored), which are relatively low in calories but rich in nutrients. The available sources for the menu include various grains, 10 types of meat, 11 types of seafood and freshwater fish, 18 types of vegetables and fruits, and 15 types

XSL•FO
RenderX

of natural sauce and plants. Sharing meals with family and having dinner conversations in the virtual world could make the player feel at *home* in the game. When Sakuna becomes picky about food, her friends encourage her to try a balanced meal. The player should consider a realistic and sustainable meal choice, as there are expiration periods for the natural food ingredients; expired foods cannot be consumed and should be turned into fertilizers. Furthermore, the menu choice is directly linked to Sakuna's power and ability to explore the next morning. Eating a well-balanced meal increases Sakuna's natural healing potential; missing a meal or major nutrients reduces the healing potential to zero, making it difficult to further explore or play an action. A summary of the game components is presented in Multimedia Appendix 3.

The game and its associated elements may contribute to the dietary effects outlined in Textbox 1.

Dietary outcomes might also be shaped by the profound psychological impacts of the game. The major hypothesis of this study was that as the participants were immersed in a virtual green environment with natural background music, their nature relatedness would improve, which could lead to reduced stress, improved mood and resilience, and enhanced mindfulness. These positive psychological effects could subsequently improve the participants' dietary choices, such as by bringing about a reduction in stress eating (thus reducing the consumption of fast foods) and promoting mindful eating. In addition, increased nature relatedness might lead participants to feel a connection with food sources and to place greater value on them, with shifts toward natural food preferences and proenvironmental behaviors, such as sustainable food choices. However, these could not be fully determined by this study alone; further research is necessary to determine a clearer association between game elements and their effects.

Using *Sakuna: Of Rice and Ruin*, we evaluated our study participants' affinity to nature, mental health, and nutritional eating patterns. In particular, we focused on their enhanced awareness of rice cultivation and rural areas, their mental health, and their health behaviors in a context that required the application of their knowledge of rice farming.

Figure 2. Key scenes from the Sakuna: Of Rice and Ruin game (images used with permission from Edelweiss, the video game developers).

XSL•FO
RenderX

Textbox 1. Dietary effects of the game.

- Increased appreciation for a well-balanced diet (with grains, plants, and good protein sources): with an emphasis on balanced nutrition contributing to the character's power and natural healing potential, participants may develop a positive perception of well-balanced homemade food and wholesome nutrition with minimal processing. Experiencing the process of rice farming may deepen players' understanding and appreciation of the labor and resources required to produce rice and whole grains. In the postgame interview, some of the participants responded that after experiencing the milling process, they realized how precious the grains were and even

  purchased

  the specific types of whole grains in their real life.
- Shift toward natural or homemade food and reduced fast food consumption: the simulation of traditional home cooking and eating natural foods in the game could influence the player to shift toward natural food selection or homemade meals rather than processed or fast foods. Some participants mentioned that after playing this game, they felt more fun in cooking and tried home cooking more frequently, which reduced the consumption of outside foods. A participant said, "I tried cooking at home the traditional dish I had in Sakuna because it looked so good."
- Sociocultural reinforcement of healthy food choices: Through daily interaction with nonplayer characters (friends at family dinners and the community), participants are reinforced of the importance of well-balanced diets and traditional Japanese-style homemade meals. Although the sociocultural effect significantly influences Sakuna's dietary practice in the virtual world, it could also influence participants' dietary practices in real life.
- Sustainable dietary behavior: Owing to the expiration periods of natural ingredients and foods, participants should make a realistic choice of ingredients, considering both their freshness and nutritional values. Repeatedly experiencing the process of creating natural fertilizers from remnant foods can influence participants into making their dietary behaviors sustainable.

Assessment of Nature Relatedness and Nutritional Status

The Nature Relatedness Scale (NR) and Nutrition Quotient Scale (NQ) were used to assess nature relatedness and nutritional status, respectively. The NR measures and evaluates an individual's connection with nature, their worldview with respect to nature, and their physical familiarity with nature across 3 subscales: *self*, *perspective*, and *experience* [26]. The reliability coefficients for the NR in the baseline and end evaluations were 0.693 and 0.569, respectively. The NQ measures and evaluates the quality and nutritional value of individuals' meals. It assesses several aspects of a person's dietary habits across 4 subscales: *balance*, *diversity*, *moderation*, and *dietary behavior* [27]. Balance is the frequency of intake of fruits, eggs, dairy products, beans or bean products, fish, and breakfast. Diversity indicates the level of deviated food habits and the frequency of vegetable and water intake. Moderation is the frequency of having sweet and greasy bread, processed beverages, instant noodles, sugar-sweetened beverages, late-night snacks, and street foods. Dietary behavior includes checking nutrition labels, performing exercises for more than 30 minutes, and washing hands before meals.

Assessment of Psychological State

We used multiple scales and inventories to evaluate the psychological states of participants (Textbox 2).

XSL•FO
RenderX

Textbox 2. Instruments used to assess the psychological states of participants.

The World Health Organization's Quality of Life–Brief Version (WHOQOL-BREF)

- The overall quality of life and well-being were assessed across 4 subscales: physical health, psychological health, social relations, and environment [28]. The reliability coefficients for the WHOQOL-BREF scale in the baseline and end evaluations were 0.895 and 0.925, respectively.

The Brief Fear of Negative Evaluation Scale (BFNE)

- This scale measures the fear of being negatively judged by others, where higher scores indicate greater fear [29]. The reliability coefficients for the BFNE scale in the baseline and end evaluations were 0.725 and 0.659, respectively.

The Social Avoidance and Distress Scale (SADS)

- This self-report tool assesses social anxiety and the tendency to avoid uncomfortable social situations [30]. The reliability coefficients for the SADS in the baseline and end evaluations were 0.940 and 0.958, respectively.

The Toronto Alexithymia Scale (TAS)

- This self-report measure assesses alexithymia, a personality trait characterized by difficulties in recognizing and expressing emotions [31]. The reliability coefficients for the TAS in the baseline and end evaluations were 0.940 and 0.884, respectively.

The State-Trait Anxiety Inventory (STAI)

- This inventory measures state and trait anxiety in both research and clinical practice contexts [32]. The reliability coefficients for the STAI–State in the baseline and end evaluations were 0.933 and 0.950, respectively. The reliability coefficients for the STAI–Trait in the baseline and end evaluations were 0.938 and 0.950, respectively.

The Center for Epidemiologic Studies Depression Scale Revised (CESD-R)

- This scale comprises 20 questions about various symptoms of depression as they may have appeared over the previous week [33]. The reliability coefficients for the CESD-R in the baseline and end evaluations were 0.915 and 0.912, respectively.

The Korean Resilience Quotient (KRQ)

- This measure assesses resilience across 3 areas: self-regulation ability, interpersonal relationships, and positivity [34]. The reliability coefficients for the KRQ scale in the baseline and end evaluations were 0.946 and 0.959, respectively.

The General Self-Efficacy Scale (GSE)

- This scale measures individuals' beliefs in their own abilities and assesses their confidence across various problem situations [35]. The reliability coefficients for the GSE in the baseline and end evaluations were 0.888 and 0.918, respectively.

The Spiritual Well-Being Scale (SWBS)

- This self-report measure assesses spiritual well-being, which refers to individuals' sense of purpose, connection to something greater than themselves, and inner peace [36]. The reliability coefficients for the SWBS in the baseline and end evaluations were 0.895 and 0.905, respectively.

Statistical Analyses

Data analysis was performed using SPSS (version 24, IMB Corp), incorporating a range of statistical tests, such as 2-tailed *t* tests, chi-square tests, and repeated-measures ANOVA. Repeated-measures ANOVA was specifically used to compare 2 groups of repeated observations captured at baseline, after 1 week of gameplay, and after 3 weeks of gameplay in the IIG and at the corresponding time points in the WLG. Given that the digital game intervention was eventually implemented in both groups, the 2 groups were amalgamated into a single sample (sample of the actual same gameplay at each time point in the IIG arm and after the wait in the WLG arm) to analyze intervention-induced changes. A *P* value of less than .05 was established as the threshold for statistical significance in all analyses. As a total of 10 psychological scales were applied to the analysis, multiple correction was performed. Bonferroni correction was performed for 7 scales (World Health Organization Quality of Life–Brief Version [WHOQOL-BREF], Brief Fear of Negative Evaluation Scale, Social Avoidance and Distress Scale, Toronto Alexithymia Scale, State-Trait Anxiety Inventory [STAI]–State, STAI–Trait, and Center for Epidemiologic Studies Depression Scale Revised) that were measured twice during the study period and for 3 scales (Korean Resilience Quotient, General Self-Efficacy Scale, and Spiritual Well-Being Scale) that were measured thrice during the study period.

Ethical Considerations

This study underwent a thorough ethical review and was approved by the Chungnam National University Sejong Hospital Institutional Review Board (reference 2021-01-012) and the Clinical Research Information Service (CRIS; reference KCT0007657). Written informed consent was obtained from all parents of the participating children, and informed consent was obtained from all participating children. The study data were anonymized and unidentified. Participants received compensation of KRW ₩70,000 (US $50) for their involvement in the study.

XSL•FO
RenderX

Results

Overview

Table 1 presents the characteristics of all 66 participants. There were no significant differences between the IIG (n=34) and the WLG (n=32) in terms of age, sex, socioeconomic status, residence type, or marital status (all *P*>.05).

Table 1. Comparison of participant characteristics.

Characteristics	Total (n=66)	Groups		Statistics	*P* values
		IIG[a] (n=34)	WLG[b] (n=32)		
Age (y), mean (SD)	27 (5.07)	28 (5.04)	27 (5.10)	t_{64}=0.7	.98
Age group (y), n (%)				χ^2_1=0.4	.53
19-29	45 (68)	22 (65)	23 (72)		
30-40	21 (32)	12 (35)	9 (28)		
Sex, n (%)				χ^2_1=0.01	.98
Male	56 (85)	29 (85)	27 (84)		
Female	10 (15)	5 (15)	5 (16)		
Socioeconomic status, n (%)				χ^2_3=6.6	.09
Wealthy	26 (39)	12 (35)	14 (44)		
Moderate	26 (39)	18 (53)	8 (25)		
Temporary	11 (17)	3 (9)	8 (25)		
Poor	3 (5)	1 (3)	2 (6)		
Residence type, n (%)				χ^2_2=1.4	.50
Living alone	15 (23)	9 (27)	6 (19)		
Living with another person	10 (15)	6 (18)	4 (12)		
Living with ≥2 persons	40 (61)	18 (55)	22 (69)		
Marital status, n (%)				χ^2_1=0.7	.79
Unmarried	57 (86)	29 (85)	28 (87)		
Married	9 (14)	5 (15)	4 (13)		

[a]IIG: immediate intervention group.

[b]WLG: waitlist group.

Impacts of the Game on Nature Relatedness and Nutritional Status in the IIG and WLG

Analysis of the impact of the game on nature relatedness and nutritional status (as shown in Table 2 and Multimedia Appendix 4) revealed significant group×time interactions for both the total NR score ($F_{1,60}$=13.418; *P*=.001) and the *self* subscale of the NR ($F_{1,62}$=16.977; *P*<.001). The total NQ score ($F_{2,116}$=6.209; *P*=.003) and moderation ($F_{2,118}$=3.954; *P*=.02) and dietary behavior ($F_{2,124}$=12.619; *P*<.001) subscale scores generally increased from baseline in both groups, but no group×time interactions were found.

XSL•FO
RenderX

Table 2. Comparison of the impact of the game on nature relatedness and nutritional status between the immediate invention group (IIG) and waitlist group (WLG).

Measures and group	Baseline, mean (SD)	1 week after starting the game or 1 week after waiting for the game, mean (SD)	3 weeks after starting the game or 2 weeks after waiting for the game, mean (SD)	Group×time interaction, *F* test (*df*)	*P* values	Time effect, *F* test (*df*)	*P* values
NR[a]							
Total NR				13.418 (1, 60)	.001	10.053 (1, 60)	.002
IIG	56.59 (6.95)	N/A[b]	61.22 (6.00)				
WLG	58.07 (7.25)	N/A	57.73 (7.31)				
Self				16.977 (1, 62)	<.001	18.185 (1, 62)	<.001
IIG	24.74 (6.36)	N/A	28.62 (5.51)				
WLG	26.07 (5.15)	N/A	26.13 (5.48)				
Perspective				0.297 (1, 63)	.59	0.051 (1, 63)	.82
IIG	15.45 (3.77)	N/A	15.55 (3.36)				
WLG	16.56 (3.00)	N/A	16.34 (3.13)				
Experience				2.453 (1, 63)	.12	0.682 (1, 63)	.41
IIG	15.79 (2.47)	N/A	16.39 (2.05)				
WLG	15.56 (2.82)	N/A	15.38 (2.37)				
NQ[c]							
Total NQ				1.461 (2, 116)	.24	82.038 (2, 116)	.003
IIG	48.00 (11.59)	49.78 (10.62)	51.46 (11.50)				
WLG	50.15 (10.65)	51.07 (8.50)	51.36 (9.14)				
Balance				0.486 (2, 126)	.62	1.798 (2, 126)	.17
IIG	28.64 (15.53)	28.85 (14.24)	29.81 (15.43)				
WLG	31.92 (14.92)	30.68 (13.24)	33.81 (17.77)				
Diversity				1.319 (2, 126)	.27	1.670 (2, 126)	.19
IIG	46.75 (19.36)	50.58 (18.88)	51.29 (16.86)				
WLG	49.60 (15.97)	50.57 (13.58)	49.45 (13.58)				
Moderation				0.950 (2, 118)	.39	3.954 (2, 118)	.02
IIG	69.65 (12.18)	69.80 (12.35)	73.54 (12.97)				
WLG	68.31 (11.93)	69.06 (11.28)	70.01 (13.69)				
Dietary behavior				1.733 (2, 124)	.18	12.619 (2, 124)	<.001
IIG	40.91 (17.46)	45.68 (13.84)	47.76 (16.60)				
WLG	44.91 (19.78)	48.24 (17.91)	47.91 (19.52)				

[a]NR: Nature Relatedness Scale.

[b]N/A: not applicable.

[c]NQ: Nutrition Quotient Scale.

Impact of the Game on Psychological State in the IIG and WLG

Analysis of the impact of the game on psychological state with Bonferroni correction (Table 3 and Multimedia Appendix 4) showed a significant group×time interaction concerning the total WHOQOL-BREF score ($F_{1,62}$=7.652; *P*=.049). Scores for the psychological ($F_{1,64}$=6.530; *P*=.01), social ($F_{1,63}$=9.677; *P*=.003), and environmental ($F_{1,63}$=4.250; *P*=.04) subscales showed a significant time effect at week 3 compared with baseline, but no significant group×time interaction was observed. No significant time effects or group×time interactions were found for the other psychological scales (all *P*>.05).

XSL•FO
RenderX

Table 3. Comparison of the impact of the game on psychological state between the immediate intervention group (IIG) and waitlist group (WLG).

Measures and group	Baseline, mean (SD)	1 week after starting the game or 1 week after waiting for the game, mean (SD)	3 weeks after starting the game or 2 weeks after waiting for the game, mean (SD)	Group×time interaction, *F* test (*df*)	*P* values	Time effect, *F* test (*df*)	*P* values
WHOQOL-BREF[a]							
Total score				7.652 (1, 62)	.049[c]	8.862 (1, 62)	.02[c]
IIG	87.79 (11.84)	N/A[b]	93.06 (13.08)				
WLG	89.65 (12.90)	N/A	89.84 (12.30)				
Physical				3.649 (1, 63)	.06	1.433 (1, 63)	.24
IIG	24.29 (3.98)	N/A	25.56 (4.51)				
WLG	24.65 (5.05)	N/A	24.35 (4.37)				
Psychological				3.192 (1, 64)	.08	6.530 (1, 64)	.01
IIG	19.44 (3.82)	N/A	20.68 (4.25)				
WLG	19.56 (4.19)	N/A	19.78 (3.78)				
Social				3.140 (1, 63)	.08	9.677 (1, 63)	.003
IIG	9.68 (1.97)	N/A	10.62 (1.81)				
WLG	9.61 (2.19)	N/A	9.87 (2.16)				
Environment				3.877 (1, 63)	.05	4.250 (1, 63)	.04
IIG	27.55 (3.65)	N/A	28.91 (3.75)				
WLG	28.88 (4.16)	N/A	28.91 (3.97)				
BFNE[d]				1.065 (1, 62)	.14[c]	0.630 (1, 62)	.01[c]
IIG	27.91 (7.89)	N/A	27.76 (6.74)				
WLG	29.32 (4.92)	N/A	30.48 (5.27)				
SADS[e]				0.730 (1, 58)	.77[c]	4.173 (1, 58)	.32[c]
IIG	84.13 (19.72)	N/A	81.35 (20.48)				
WLG	86.21 (15.39)	N/A	85.07 (16.60)				
TAS[f]				0.065 (1, 61)	.60[c]	1.423 (1, 61)	.67[c]
IIG	46.50 (9.63)	N/A	45.16 (11.01)				
WLG	45.68 (10.34)	N/A	44.81 (9.25)				
STAI-S[g]				0.427 (1, 62)	.61[c]	0.054 (1, 62)	0.71[c]
IIG	39.00 (8.75)	N/A	38.41 (9.85)				
WLG	40.84 (10.94)	N/A	41.13 (9.99)				
STAI-T[h]				2.514 (1, 59)	.83[c]	5.299 (1, 59)	.18[c]
IIG	41.80 (10.16)	N/A	39.00 (11.18)				
WLG	42.03 (10.19)	N/A	41.52 (10.93)				
CESD[i]				0.370 (1, 61)	.82[c]	5.617 (1, 61)	.15[c]

XSL•FO
RenderX

Measures and group	Baseline, mean (SD)	1 week after starting the game or 1 week after waiting for the game, mean (SD)	3 weeks after starting the game or 2 weeks after waiting for the game, mean (SD)	Group×time interaction, *F* test (*df*)	*P* values	Time effect, *F* test (*df*)	*P* values
IIG	9.69 (9.88)	N/A	6.91 (6.99)				
WLG	10.55 (9.01)	N/A	8.90 (9.89)				
KRQ[j]				3.804 (2, 106)	.17[k]	0.381 (2, 106)	.62[k]
IIG	178.58 (27.56)	184.50 (28.23)	183.50 (31.97)				
WLG	184.00 (23.52)	182.90 (23.65)	183.00 (24.61)				
GES[l]				0.231 (2, 110)	.38[k]	0.251 (2, 110)	.45[k]
IIG	80.10 (13.61)	79.03 (20.34)	80.48 (19.03)				
WLG	82.39 (13.64)	81.68 (13.39)	81.46 (13.41)				
SWBS[m]				0.307 (2, 114)	.21[k]	0.432 (2, 114)	.95[k]
IIG	67.04 (16.00)	67.39 (18.19)	68.46 (18.14)				
WLG	64.45 (16.46)	65.29 (15.14)	64.84 (14.24)				

[a]WHOQOL-BREF: World Health Organization's Quality of Life–Brief Version.

[b]N/A: not applicable.

[c]Bonferroni multiple corrections (*P* value×7) for a total of 7 psychological scales (World Health Organization's Quality of Life–BREF, Brief Fear of Negative Evaluation Scale, Social Avoidance and Distress Scale, Toronto Alexithymia Scale, State-Trait Anxiety Inventory–State, Sate-Trait Anxiety Inventory–Trait, and Center for Epidemiologic Studies Depression Scale).

[d]BFNE: Brief Fear of Negative Evaluation Scale.

[e]SADS: Social Avoidance and Distress Scale.

[f]TAS: Toronto Alexithymia Scale.

[g]STAI-S: State-Trait Anxiety Inventory–State.

[h]STAI-T: State-Trait Anxiety Inventory–Trait.

[i]CESD: Center for Epidemiologic Studies Depression Scale.

[j]KRQ: Korean Resilience Quotient.

[k]Bonferroni multiple corrections (*P* value×3) for a total of 3 psychological scales (Korean Resilience Quotient, General Self-Efficacy Scale, and Spiritual Well-Being Scale).

[l]GES: General Self-Efficacy Scale.

[m]SWBS: Spiritual Well-Being Scale.

Impact of the Game on Nature Relatedness, Nutrition Status, and Psychological State Across the Entire Sample

As both IIG and WLG ultimately played the game, we performed additional analyses on the changes in each scale from before to after the game for the entire sample of these 2 groups combined. Multimedia Appendices 5 and 6 present the overall changes in nature relatedness, nutritional status, and psychological state for all participants before and after gameplay. There were statistically significant increases in the total NR score (t_{62}=−4.150; *P*<.001) and total NQ score ($F_{2,114}$=4.775; *P*=.01) after the 3-week game period. The total WHOQOL-BREF score also significantly increased after gameplay (t_{63}=−4.027; *P*<.007). The STAI–Trait score (t_{59}=3.765; *P*<.007) statistically significantly decreased following gameplay compared with baseline. No statistically significant differences were found for any of the other psychological scales (all *P*>.05).

Discussion

Principal Findings

Our study explored the impacts of a nature-friendly, rice-farming action video game on urban dwellers during the social distancing period of the COVID-19 pandemic, specifically examining changes in nature relatedness, nutritional behavior, and psychological state [37] The digital game intervention used in

XSL•FO
RenderX

this study, similar to the concept of drug repositioning or the discovery of therapeutic factors from existing substances, sought to repurpose a recreational activity to take advantage of its therapeutic benefits [38]. This approach is gaining importance in the age of digital therapies and the emerging metaverse, where synergistic interactions between virtual and real-world experiences can potentially enhance mental health, promote healthy nutrition, and foster a deeper connection with nature [39,40].

As hypothesized, the results revealed significant changes after 3 weeks of gameplay in the total NR score and the *self* subscale of the NR in the IIG compared with the WLG. This indicates that the game intervention bolstered participants' connection with nature, especially by influencing their personal relationships. For urban dwellers, who are increasingly disconnected from nature, this indicates a promising strategy for nurturing an affinity with the natural environment [5]. The enhancement in nature relatedness could have clinical applications, particularly in the context of eco-friendly therapeutic approaches, which are emerging as effective components of mental health treatment [41]. By fostering a connection with nature, such game interventions may provide a complementary tool for treating affective disorders and stress-related conditions. They can also be incorporated into mindfulness-based therapies to augment the therapeutic experience, especially for those with limited access to natural settings [42]. These implications suggest a valuable role for nature-centric games in the broader spectrum of mental health care approaches.

These findings are particularly relevant in light of the ongoing COVID-19 pandemic, which has necessitated social distancing and led to an increase in indoor activities, such as video gaming [17]. Video games, accessible through various platforms, including computers, consoles, and mobile devices, have been recommended by the World Health Organization as a means of alleviating stress [43,44]. In this context, our results suggest that video games have the potential to enhance nature connectivity, which is encouraging. This opens up possibilities for the development of the so-called serious games designed with therapeutic objectives [43].

The study also found significant changes in the total WHOQOL-BREF scores in the IIG compared with the WLG, indicating potential positive impacts on psychological state. These findings suggest that nature-friendly video games have an impact on the quality of life in general, in addition to improving specific psychological and nutritional aspects [38]. However, it was not possible to conclusively attribute these changes to specific factors in the video game, such as the joy derived from the gameplay or the cathartic impact of the narrative of the game.

The observed improvements in the WHOQOL-BREF scores may have therapeutic implications for individuals with diminished quality of life due to chronic conditions or individuals undergoing rehabilitation [45]. Aspects of the game that enhance psychological well-being could be particularly beneficial in clinical settings, offering a noninvasive means to support traditional therapies. Such games could be explored as a method to deliver positive psychological interventions, potentially aiding in the management of conditions such as chronic pain or fatigue, where patient engagement in enjoyable activities has been shown to contribute to the overall well-being.

For additional analysis, we investigated the pregame and postgame results for all participants. The findings showed significant changes in the total NR score, scores of the *self* and *experience* subscales of the NR, the total NQ score, and score of the dietary behavior subscale. In psychological terms, not only did the total WHOQOL-BREF score and scores of all of its subscales show significant changes, but the Social Avoidance and Distress Scale, Toronto Alexithymia Scale, STAI, and Center for Epidemiologic Studies Depression Scale Revised scores also showed significant changes. Although the pre-post comparison of all participants was limited to the same group without comparison, the broad positive changes in nature relatedness, nutritional status, and psychological states observed when the sample size was increased can be considered a basis for extrapolating the therapeutic and educational effects of a game intervention with targeted content and features. If future studies can ensure sufficient group-specific sample sizes, we can expect these findings to be further emphasized. In addition, the observation of dietary behavior changes through a nature-friendly serious game in this study raises the possibility of using such games as part of digital therapies for improving and managing dietary habits among urban dwellers.

Our study used the commercial game *Sakuna: Of Rice and Ruin*, which was launched by the South Korean branch of Nintendo in November 2020. The participants were randomized into the IIG and WLG conditions to evaluate the effects of the video game intervention. From the protective effects of nature on mental health [46], it can be assumed that an increased connection with nature through gaming may have some association with improved mental health and quality of life. Although our study showed significant results in terms of the NR, a case-control study examining the impact of specific game elements on nature connectivity could offer more precise insights. Furthermore, well-designed studies are necessary to ascertain which game factors facilitated the observed changes in the NR.

Previous studies have presented mixed results, suggesting that future research should prioritize understanding the long-term effects of gamification and the elements contributing to its efficacy [21]. It is crucial to identify the *active ingredients* that drive sustained behavioral changes and how they interact with individual differences, including personality traits and demographics [47]. Moreover, comprehensive longitudinal studies are essential to investigate gamification's long-term impact [20]. Although gamification shows promise in enhancing health and educational outcomes, its long-term effectiveness and core components require further exploration.

Strengths and Limitations

This study has several strengths. First, it uses a commercial game for therapeutic and educational purposes. Using *Sakuna: Of Rice and Ruin*, we offered a real-world context for the exploration of the potential impact of video games on nature connectivity, dietary behavior, and psychological states. Similar

XSL•FO
RenderX

to exploring the development of new drugs using natural plants or existing drugs, exploring therapeutic or educational elements in existing commercial games will be necessary in the coming digital age. This study has significant strengths in that regard. Second, the study design, which incorporated the random allocation of participants into either the IIG or the WLG, helped reduce the risk of selection bias, thereby improving the reliability of the findings. Third, the comprehensive evaluation of a broad range of outcomes, including changes in nature relatedness, nutritional status, quality of life, and various psychological measures, provided a multifaceted perspective on the potential benefits of video game interventions.

This study has some limitations. First, although the period in which we conducted our study was unique, with the COVID-19 pandemic spreading in South Korea and social distancing being mandated as a national policy, this study primarily targeted the general public, and the participants were not a patient group. Most psychological scales used in this study were designed for psychological and pathological conditions; because we targeted a general population, and not a clinical group, the baseline values of the psychological scale were not high. This posed limitations for the evaluation of the game's psychological improvement effects compared with a patient group. Future studies would benefit from examining the effects on patient groups with psychological or physical issues or high-risk groups. Second, this study did not clearly identify the specific elements of the video game that contributed to the observed behavioral and emotional changes. Further research is needed to determine which elements of the game are the most influential in producing significant effects. Finally, the sample size in this study was relatively small to determine statistical significance. To increase the reliability and statistical power of the findings, it is necessary to conduct another study with a larger sample size. Along these lines, it is also necessary to confirm the evidence produced through additional independent studies to ensure reproducibility.

Conclusion

Our study provides empirical evidence to suggest that nature-based video games, specifically *Sakuna: Of Rice and Ruin*, may enhance nature relatedness and positively influence nutritional and psychological states in an urban population. In light of our findings and the growing popularity of video games, particularly in the context of an increased tendency to stay at home owing to the COVID-19 pandemic, serious consideration should be given to the integration of digital games into strategies for promoting healthy behaviors and mental well-being. This is especially relevant as we navigate the challenges of our increasingly urbanized and digitized society. In conclusion, our study suggests that digital games, specifically those with a nature-based theme, can have a positive impact on users' connection with nature, nutritional habits, and psychological health. Further studies are needed to corroborate these findings, identify the various influences of specific game elements, and determine the long-term effectiveness of these games as a form of digital therapy. Furthermore, it is necessary to optimize these specific game tools and integrate them into a real-world ecosystem to maximize their therapeutic potential.

Acknowledgments

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and Information and Communications Technology (MSIT), Government of Korea (NRF-2020R1C1C1007463 and NRF-2021R1A5A8032895); Information and Communications Technology and Future Planning for Convergent Research in the Development Program for R&D Convergence Over Science and Technology Liberal Arts (NRF-2022M3C1B6080866); and an Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT; RS-2023-00224823).

Data Availability

The data supporting the findings of this study are available upon request from the corresponding author. The data are not publicly available because of privacy concerns.

Authors' Contributions

CHC and JL are cocorresponding authors of this paper and are responsible for data and materials, manuscript submission, peer review, publication process, authorship details, and ethics committee approval. JL and CHC conceived and designed the study. SL, CY, YBS, JL, and CHC performed the analyses. SL, CY, JL, and CHC wrote the first draft of the manuscript. SL, CY, YBS, JL, and CHC collected the data. SL, CY, HJL, YML, JL, and CHC edited all versions of the manuscript. All authors were involved in interpreting the results and read, commented on, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Scenes captured from the game Sakuna: Of Rice and Ruin (images used with permission from Edelweiss, the video game developers).
[DOCX File , 6310 KB - jmir_v26i1e51596_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Summary of game characteristics.
[DOCX File , 16 KB - jmir_v26i1e51596_app2.docx]

Multimedia Appendix 3
Summary of game components.
[DOCX File , 16 KB - jmir_v26i1e51596_app3.docx]

Multimedia Appendix 4
Significant findings regarding the impact of the game in the immediate invention group and waitlist group.
[DOCX File , 36 KB - jmir_v26i1e51596_app4.docx]

Multimedia Appendix 5
Impact of the computer game on nature relatedness, nutritional status, and psychological state in the entire sample.
[DOCX File , 21 KB - jmir_v26i1e51596_app5.docx]

Multimedia Appendix 6
Significant findings regarding the impact of the game in the entire sample.
[DOCX File , 48 KB - jmir_v26i1e51596_app6.docx]

Multimedia Appendix 7
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 15077 KB - jmir_v26i1e51596_app7.pdf]

References

1. Harpham T. Urbanization and mental health in developing countries: a research role for social scientists, public health professionals and social psychiatrists. Soc Sci Med 1994 Jul;39(2):233-245. [doi: 10.1016/0277-9536(94)90332-8] [Medline: 8066502]
2. Moore M, Gould P, Keary BS. Global urbanization and impact on health. Int J Hyg Environ Health 2003 Aug;206(4-5):269-278. [doi: 10.1078/1438-4639-00223] [Medline: 12971682]
3. Dean JH, Shanahan DF, Bush R, Gaston KJ, Lin BB, Barber E, et al. Is nature relatedness associated with better mental and physical health? Int J Environ Res Public Health 2018 Jun 29;15(7):1371 [FREE Full text] [doi: 10.3390/ijerph15071371] [Medline: 29966307]
4. Walsh R. Lifestyle and mental health. Am Psychol 2011 Oct;66(7):579-592. [doi: 10.1037/a0021769] [Medline: 21244124]
5. Martyn P, Brymer E. The relationship between nature relatedness and anxiety. J Health Psychol 2016 Jul 04;21(7):1436-1445. [doi: 10.1177/1359105314555169] [Medline: 25370570]
6. Popkin BM. Urbanization, lifestyle changes and the nutrition transition. World Development 1999 Nov;27(11):1905-1916. [doi: 10.1016/s0305-750x(99)00094-7]
7. Yach D, Stuckler D, Brownell KD. Epidemiologic and economic consequences of the global epidemics of obesity and diabetes. Nat Med 2006 Jan 01;12(1):62-66. [doi: 10.1038/nm0106-62] [Medline: 16397571]
8. Basu S, Stuckler D, McKee M, Galea G. Nutritional determinants of worldwide diabetes: an econometric study of food markets and diabetes prevalence in 173 countries. Public Health Nutr 2012 Jun 13;16(1):179-186. [doi: 10.1017/s1368980012002881]
9. Ulrich RS. Visual landscapes and psychological well‐being. Landscape Res 2007 Feb 24;4(1):17-23. [doi: 10.1080/01426397908705892]
10. Groenewegen PP, van den Berg AE, de Vries S, Verheij RA. Vitamin G: effects of green space on health, well-being, and social safety. BMC Public Health 2006 Jun 07;6(1):149 [FREE Full text] [doi: 10.1186/1471-2458-6-149] [Medline: 16759375]
11. Hwang TJ, Rabheru K, Peisah C, Reichman W, Ikeda M. Loneliness and social isolation during the COVID-19 pandemic. Int Psychogeriatr 2020 May 26;32(10):1217-1220. [doi: 10.1017/s1041610220000988]
12. Williams SN, Armitage CJ, Tampe T, Dienes K. Public perceptions and experiences of social distancing and social isolation during the COVID-19 pandemic: a UK-based focus group study. BMJ Open 2020 Jul 20;10(7):e039334 [FREE Full text] [doi: 10.1136/bmjopen-2020-039334] [Medline: 32690752]
13. Hoffart A, Johnson SU, Ebrahimi OV. Loneliness and social distancing during the COVID-19 pandemic: risk factors and associations with psychopathology. Front Psychiatry 2020 Nov 20;11:589127 [FREE Full text] [doi: 10.3389/fpsyt.2020.589127] [Medline: 33329136]

XSL•FO
RenderX

14. Lim S, Sohn M. How to cope with emerging viral diseases: lessons from South Korea's strategy for COVID-19, and collateral damage to cardiometabolic health. Lancet Reg Health West Pac 2023 Jan;30:100581 [FREE Full text] [doi: 10.1016/j.lanwpc.2022.100581] [Medline: 36093123]
15. Michie S, West R. Sustained behavior change is key to preventing and tackling future pandemics. Nat Med 2021 May 10;27(5):749-752. [doi: 10.1038/s41591-021-01345-2] [Medline: 33972794]
16. The Lancet Respiratory Medicine. Future pandemics: failing to prepare means preparing to fail. Lancet Respir Med 2022 Mar;10(3):221-222. [doi: 10.1016/s2213-2600(22)00056-x]
17. Barr M, Copeland-Stewart A. Playing video games during the COVID-19 pandemic and effects on players' well-being. Games Culture 2021 May 06;17(1):122-139. [doi: 10.1177/15554120211017036]
18. Online games encourage players to stay mentally and physically healthy at home. World Health Organization. 2021 Nov 16. URL: https://www.who.int/news/item/16-11-2021-online-games-encourage-players-to-stay-mentally-and-physically-healthy-at-home [accessed 2023-06-23]
19. Ellis LA, Lee MD, Ijaz K, Smith J, Braithwaite J, Yin K. COVID-19 as 'game changer' for the physical activity and mental well-being of augmented reality game players during the pandemic: mixed methods survey study. J Med Internet Res 2020 Dec 22;22(12):e25117 [FREE Full text] [doi: 10.2196/25117] [Medline: 33284781]
20. Cugelman B. Gamification: what it is and why it matters to digital health behavior change developers. JMIR Serious Games 2013 Dec 12;1(1):e3 [FREE Full text] [doi: 10.2196/games.3139] [Medline: 25658754]
21. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
22. Sakuna: Of Rice and Ruin. Wikipedia. URL: https://en.wikipedia.org/wiki/Sakuna:_Of_Rice_and_Ruin [accessed 2023-06-23]
23. Ohnuki-Tierney E. Rice as Self: Japanese Identities through Time. Princeton, NJ: Princeton University Press; 1993.
24. Koga M, Toyomaki A, Kiso Y, Kusumi I. Impact of a rice-centered diet on the quality of sleep in association with reduced oxidative stress: a randomized, open, parallel-group clinical trial. Nutrients 2020 Sep 24;12(10):2926 [FREE Full text] [doi: 10.3390/nu12102926] [Medline: 32987839]
25. Kim HS, Jung SJ, Mun EG, Kim MS, Cho SM, Cha YS. Effects of a rice-based diet in Korean adolescents who habitually skip breakfast: a randomized, parallel group clinical trial. Nutrients 2021 Mar 05;13(3):853 [FREE Full text] [doi: 10.3390/nu13030853] [Medline: 33807703]
26. Steg L, Perlaviciute G, van der Werff E, Lurvink J. The significance of hedonic values for environmentally relevant attitudes, preferences, and actions. Environ Behav 2012 Sep 10;46(2):163-192. [doi: 10.1177/0013916512454730]
27. Kang MH, Lee JS, Kim HY, Kwon S, Choi YS, Chung HR, et al. Selecting items of a food behavior checklist for the development of nutrition quotient (NQ) for children. Korean J Nutr 2012;45(4):372. [doi: 10.4163/kjn.2012.45.4.372]
28. The WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. The WHOQOL Group. Psychol Med 1998 May 01;28(3):551-558. [doi: 10.1017/s0033291798006667] [Medline: 9626712]
29. Carleton RN, McCreary DR, Norton PJ, Asmundson GJ. Brief Fear of Negative Evaluation Scale-revised. Depress Anxiety 2006;23(5):297-303. [doi: 10.1002/da.20142] [Medline: 16688736]
30. Turner SM, McCanna M, Beidel DC. Validity of the social avoidance and distress and fear of negative evaluation scales. Behav Res Ther 1987;25(2):113-115. [doi: 10.1016/0005-7967(87)90081-7] [Medline: 3593163]
31. Bagby RM, Parker JD, Taylor GJ. The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. J Psychosomatic Res 1994 Jan;38(1):23-32. [doi: 10.1016/0022-3999(94)90005-1]
32. Spielberger CD, Gonzalez-Reigosa F, Martinez-Urrutia A, Natalicio LF, Natalicio DS. The State-Trait Anxiety Inventory. Revista Interamericana De Psicología 2017;5(3 & 4). [doi: 10.30849/rip/ijp.v5i3 & 4.620]
33. Van Dam NT, Earleywine M. Validation of the Center for Epidemiologic Studies Depression Scale--Revised (CESD-R): pragmatic depression assessment in the general population. Psychiatry Res 2011 Mar 30;186(1):128-132. [doi: 10.1016/j.psychres.2010.08.018] [Medline: 20843557]
34. Shin WY, Kim MG, Kim JH. Developing measures of resilience for korean adolescents and testing cross, convergent and discriminant validity. Stud Korean Youth 2009;20:105-131.
35. Tipton RM, Worthington EL. The measurement of generalized self-efficacy: a study of construct validity. J Pers Assess 1984 Oct;48(5):545-548. [doi: 10.1207/s15327752jpa4805_14] [Medline: 16367514]
36. Paloutzian RF, Ellison CW. Manual for the spiritual well-being scale. Westmont College. 1991. URL: https://westmont.edu/sites/default/files/users/user401/SWBS%20Manual%202.0_0.pdf [accessed 2023-12-19]
37. Grabowska-Chenczke O, Wajchman-Świtalska S, Woźniak M. Psychological well-being and nature relatedness. Forests 2022 Jul 02;13(7):1048. [doi: 10.3390/f13071048]
38. Ferrari M, McIlwaine SV, Reynolds JA, Archie S, Boydell K, Lal S, et al. Digital game interventions for youth mental health services (Gaming My Way to Recovery): protocol for a scoping review. JMIR Res Protoc 2020 Jun 24;9(6):e13834 [FREE Full text] [doi: 10.2196/13834] [Medline: 32579117]
39. Qiu CS, Majeed A, Khan S, Watson M. Transforming health through the metaverse. J R Soc Med 2022 Dec;115(12):484-486 [FREE Full text] [doi: 10.1177/01410768221144763] [Medline: 36480946]

XSL•FO
RenderX

40. Wiederhold BK. Metaverse games: game changer for healthcare? Cyberpsychol Behav Soc Netw 2022 May 01;25(5):267-269. [doi: 10.1089/cyber.2022.29246.editorial] [Medline: 35549346]
41. Chaudhury P, Banerjee D. "Recovering with nature": a review of ecotherapy and implications for the COVID-19 pandemic. Front Public Health 2020 Dec 10;8:604440 [FREE Full text] [doi: 10.3389/fpubh.2020.604440] [Medline: 33363096]
42. Luykx JJ, Voetterl HT. Toward ecologically sustainable mental health care-a call for action from within psychiatry. JAMA Psychiatry 2022 Jun 01;79(6):524-525. [doi: 10.1001/jamapsychiatry.2022.0594] [Medline: 35416980]
43. Abd-Alrazaq A, Alajlani M, Alhuwail D, Schneider J, Akhu-Zaheya L, Ahmed A, et al. The effectiveness of serious games in alleviating anxiety: systematic review and meta-analysis. JMIR Serious Games 2022 Feb 14;10(1):e29137 [FREE Full text] [doi: 10.2196/29137] [Medline: 35156932]
44. Can video games make us healthier? World Health Organization. 2021 Dec 8. URL: https://www.who.int/europe/news/item/08-12-2021-can-video-games-make-us-healthier [accessed 2023-06-23]
45. Richardson CR, Franklin B, Moy ML, Jackson EA. Advances in rehabilitation for chronic diseases: improving health outcomes and function. BMJ 2019 Jun 17;365:l2191. [doi: 10.1136/bmj.l2191] [Medline: 31208954]
46. Liu H, Ren H, Remme RP, Nong H, Sui C. The effect of urban nature exposure on mental health—a case study of Guangzhou. J Clean Prod 2021 Jul;304:127100 [FREE Full text] [doi: 10.1016/j.jclepro.2021.127100]
47. Smiderle R, Rigo SJ, Marques LB, Peçanha de Miranda Coelho JA, Jaques PA. The impact of gamification on students' learning, engagement and behavior based on their personality traits. Smart Learn Environ 2020 Jan 09;7(1). [doi: 10.1186/s40561-019-0098-x]

Abbreviations

CRIS: Clinical Research Information Service
IIG: immediate intervention group
NQ: Nutrition Quotient Scale
NR: Nature Relatedness Scale
STAI: State-Trait Anxiety Inventory
WHOQOL-BREF: World Health Organization Quality of Life–Brief Version
WLG: waitlist group

Edited by G Eysenbach; submitted 07.08.23; peer-reviewed by K Lee, T Baranowski; comments to author 31.08.23; revised version received 23.11.23; accepted 30.11.23; published 22.01.24.

Please cite as:
Lee S, Yuh C, Shin YB, Lee HJ, Lee YM, Lee J, Cho CH
Effects of a Rice-Farming Simulation Video Game on Nature Relatedness, Nutritional Status, and Psychological State in Urban-Dwelling Adults During the COVID-19 Pandemic: Randomized Waitlist Controlled Trial
J Med Internet Res 2024;26:e51596
URL: https://www.jmir.org/2024/1/e51596
doi:10.2196/51596
PMID:38252464

©Seulki Lee, Chisung Yuh, Yu-Bin Shin, Heon-Jeong Lee, Young-Mee Lee, Jungsil Lee, Chul-Hyun Cho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of Digital Dance on Brain Imagery, Cognition, and Health: Randomized Controlled Trial

Heng-Hsin Tung[1,2,3], PhD; Chen-Yuan Kuo[4], PhD; Pei-Lin Lee[3], PhD; Chih-Wen Chang[3], MPH; Kun-Hsien Chou[4,5], PhD; Ching-Po Lin[4,5,6], PhD; Liang-Kung Chen[3,7,8], PhD

[1]Department of Nursing, National Yang Ming Chiao Tung University, Taipei, Taiwan
[2]Tungs' Taichung MetroHarbor Hospital, Taichung, Taiwan
[3]Center for Healthy Longevity and Aging Sciences, National Yang Ming Chiao Tung University, Taipei, Taiwan
[4]Institute of Neuroscience, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan
[6]Department of Education and Research, Taipei City Hospital, Taipei, Taiwan
[7]Center for Geriatrics and Gerontology, Taipei Veterans General Hospital, Taipei, Taiwan
[8]Taipei Municipal Gan-Dau Hospital, Taipei, Taiwan

Corresponding Author:
Liang-Kung Chen, PhD
Center for Geriatrics and Gerontology
Taipei Veterans General Hospital
No.201, Sec. 2, Shipai Rd
Beitou District
Taipei, 11217
Taiwan
Phone: 886 2 2875 7808
Fax: 886 2 2875 7858
Email: lkchen2@vghtpe.gov.tw

Abstract

Background: Multidomain interventions have demonstrable benefits for promoting healthy aging, but self-empowerment strategies to sustain long-term gains remain elusive.

Objective: This study evaluated the effects of digital somatosensory dance game participation on brain imagery changes as primary outcomes and other physical and mental health measures as secondary outcomes related to healthy aging.

Methods: Between August 31, 2020, and June 27, 2021, this randomized controlled trial recruited 60 eligible participants older than 55 years with no recent engagement in digital dance games. A computer-generated randomization sequence was used to allocate participants 1:1, without stratification, to an intervention group (n=30) who underwent digital somatosensory dance game training or a control group (n=30). An anonymized code masked the intervention allocations from the investigators, and individuals who assigned the interventions were not involved in analyzing the study data. The intervention entailed two 30-minute dance game sessions per week for 6 months, and the control group received healthy aging education. Primary outcomes were brain imagery changes. All variables were measured at baseline and the 6-month follow-up, and intervention effects were estimated using *t* tests with intention-to-treat analyses.

Results: Compared with the control group, intervention participants had significantly different brain imagery in the gray matter volume (GMV) of the left putamen (estimate 0.016, 95% CI 0.008 to 0.024; $P<.001$), GMV of the left pallidum (estimate 0.02, 95% CI 0.006 to 0.034; $P=.004$), and fractional amplitude of low frequency fluctuations of the left pallidum (estimate 0.262, 95% CI 0.084 to 0.439; $P=.004$). Additionally, the intervention group had different imagery in the cerebellum VI GMV (estimate 0.011, 95% CI 0.003 to 0.02; $P=.01$). The intervention group also had improved total Montreal Cognitive Assessment scores (estimate 1.2, 95% CI 0.27 to −2.13; $P<.01$), quality of life (estimate 7.08, 95% CI 2.35 to 11.82; $P=.004$), and time spent sitting on weekdays (estimate −1.96, 95% CI −3.33 to −0.60; $P=.005$). Furthermore, dance performance was significantly associated with cognitive performance ($P=.003$), health status ($P=.14$), resilience ($P=.007$), and demoralization ($P<.001$).

XSL•FO
RenderX

Conclusions: Digital somatosensory dance game participation for 6 months was associated with brain imagery changes in multiple regions involving somatosensory, motor, visual, and attention functions, which were consistent with phenotypic improvements associated with healthy aging.

Trial Registration: ClinicalTrials.gov NCT05411042; https://clinicaltrials.gov/study/NCT05411042

(J Med Internet Res 2024;26:e57694) doi:10.2196/57694

KEYWORDS

digital; somatosensory dance; somatosensory game; cognitive performance; physical function; resilience; demoralization; quality of life; dance; dancer; dancing; movement; sport; sports; cognitive; cognition; brain; neuroscience; image; imagery; imaging; RCT; randomized; controlled trial; controlled trials; somatosensory; gerontology; geriatric; geriatrics; older adult; older adults; elder; elderly; older person; older people; ageing; aging; aged; game; games; gaming

Introduction

The process of aging manifests complex interplays between declining organ function, diminished physiological reserve, impaired functional performance, and the development of age-related health conditions. In 2015, the World Health Organization (WHO) published the "World Report on Aging and Health," which proposed a function-centric, integrated approach to reforming the health and social care systems [1]; this paradigm shift is intended to enable health care systems to respond to the challenges of global population aging. To encourage the implementation of healthy aging principles internationally, the WHO launched its action plan "Decade of Healthy Aging" in 2020 [2]. The WHO defines "healthy aging" as "the process of developing and maintaining the functional abilities that enables wellbeing in older age" [2], which pursues healthy longevity by minimizing unhealthy years over the life course. A pragmatic, participatory, multidomain intervention for older community dwellers substantially improved their physical function, cognitive performance, nutritional status, and quality of life, even among people with multimorbidity [3,4]. With the success of identifying at-risk populations and validating the efficacy of intervention programs, limited research to date has affirmed consistent long-term gains in both physical and mental components [4]; however, the beneficial effects of interventions weaken as the frequency of supervised activities tapers off, possibly because of a lack of self-empowerment, especially during the COVID-19 pandemic [4].

"Gamification," which incorporates pleasure and fun into health promotion activities, may contribute to sustaining beneficial intervention effects [5,6]. Digital somatosensory games are designed to engage participants in an interactive virtual environment that connects body movements with music and videos to achieve predefined goals [7]. Previous studies have reported that a gaming dance exergame is considered an enjoyable way for older adults to keep exercising. Playing digital somatosensory games not only includes physical activities but also involves sensory and cognitive coordination [8]. Hence, digital somatosensory games have been used in health care, rehabilitation, athletics, and other activities [9]. For example, Nintendo Wii Fit Plus games reportedly improve postural control and functional gait for frail older players [10], and nursing home residents who played digital somatosensory games were happier and more willing to exercise regularly [7]. Prior studies have also suggested that the use of digital games could promote structural and functional alterations in specific brain regions, such as the hippocampus, prefrontal cortex, and parietal lobe, leading to improvements in cognitive performance [11-13]. Additionally, other studies have shown that digital somatosensory games can substantially improve social, sensory, and emotional functions; physical activity levels; and disease prevention [8].

Theoretically, digital somatosensory dancing games represent suitable interventions to promote healthy aging in older adults because they integrate multitask activities that require extensive sensory coordination and are pleasurable [14]. However, no studies have evaluated the impact of digital somatosensory dance games with an explicit focus on brain signatures and healthy aging. Hence, this randomized controlled trial (RCT) aimed to study hypothetical benefits of digital somatosensory dance games on structural and functional brain imaging findings, cognitive performance, physical health, and mental health biomarkers.

Methods

Participants

This RCT enrolled community residents aged ≥55 years in Taipei, Taiwan, from August 31, 2020, to June 27, 2021, and followed them for 6 months; however, the trial was interrupted for approximately 3 months due to the COVID-19 pandemic (Figure 1). Participant inclusion criteria were as follows: (1) age ≥55 years, (2) ≥6 years of formal education, (3) spoke Mandarin and Taiwanese, (4) engagement in digital dance games <3 times per year during the past 3 years and not within the past 3 months, (5) ability to understand the research procedures and adhere to the prescribed study activities and assessments, and (6) provision of written informed consent. Participants meeting any of the following criteria were excluded from the study: (1) inability to communicate effectively with study staff, (2) undergoing active chemotherapy for malignancy, (3) life expectancy <12 months, and (4) presence of contraindications for magnetic resonance imaging (MRI) such as ferromagnetic foreign bodies or metal implants.

Study Design and Participant Recruitment

This study was a single-blind, parallel-group RCT. Eligible participants were randomly allocated 1:1, without stratification, to either the intervention or control group according to a computer-generated random number sequence. Participants

XSL•FO
RenderX

were not explicitly informed about their group assignment, and there was no contact between study investigators and participants during the study period. The intervention allocation was masked from investigators by anonymized coding, and the individuals who assigned interventions were not involved in assessing outcomes and analyzing data. The investigators also avoided any potential interactions between participants in either group. The participants were recruited from the neighborhood community, and flyers were posted in neighborhood health centers and local district service offices. The content of the flyer included the research purpose, health examination results, and compensation information.

Staff Training

All researchers completed education workshops before the study began using a written instruction manual to ensure standard implementation of study interventions.

Ethical Considerations

Human Subject Ethics Review and Approvals

This study was approved by the institutional review board of National Yang Ming Chiao Tung University (YM108123F).

Informed Consent

Informed consent was obtained, and the right of participants to withdraw from the program was explained.

Privacy and Confidentiality

The electronic data for analysis were deidentified, and the questionnaires with the participants' signatures were stored in a locked place to safeguard participant information.

Compensation Details

The participants received breakfast following their morning fasting blood work and a gift card worth NT$100 (US $3.30) as compensation during both the baseline and posttest phases.

Procedures

Baseline Measurements

Before the first dance game session, participants in both study arms underwent 3-T brain MRI scans and completed questionnaires on physical activity, resilience, demoralization, and quality of life; baseline measurements included physical performance, body composition, and blood biochemistry biomarkers.

Digital Somatosensory Dance Game Intervention

The intervention group played a digital somatosensory dance game (DANZ BASE; International Games System Company) at a community health center; these sessions, each lasting ≥30 minutes (including a 5-7–minute warm-up and cooldown), took place twice weekly for 6 months.

In each digital somatosensory dance game session, participants danced to 3 songs categorized as 1, 2, or 3 stars according to the increasing complexity of the dance movements. Each of these songs also had 4 levels of difficulty (1: easy; 2: ordinary; 3: hard; 4: master), which were used in successive intervention periods. DANZ-BASE has standard levels of interactive dance movements for each song, and after completing each song, participants are scored based on the accuracy with which their dance movements followed those of the on-screen virtual performers. An integrated digital camera evaluates the dance movements of performers at 4 predefined levels (miss, nice, great, and perfect) based on the timing and positioning of participants compared with the images on the screen (Tables S1-S3 and Figure S1 in Multimedia Appendix 1).

The scheduled 6-month program was modified to mitigate withdrawals from the study, especially during the COVID-19 pandemic, which could have potentially lessened observable intervention effects. Notably, the intervention had to be interrupted at difficulty level 4 for 1 month due to the pandemic, and the remaining sessions were recommenced at level 3 or level 2.

Control group participants underwent the same assessments and examinations at baseline and at the end of the study as the intervention group, including the questionnaire, physical examination, laboratory tests, and MRI (Table S1 in Multimedia Appendix 1); however, during the 6-month intervention period, they only received healthy aging education fact sheets and monthly telephone calls to monitor their daily activities.

We tried to enhance adherence in several ways, including assigning participants a dedicated research assistant who contacted them monthly by telephone to determine if they had any difficulties with adhering to the program and to help resolve any issues as needed.

Outcomes and Measurements

Primary Outcome: Neuroimaging

The primary outcome was brain imagery changes as shown on neuroimaging.

All participants underwent 2 MRI brain scans. The first scan used a Siemens 3T MAGNETOM Trio scanner; the second MRI scan was done with an upgraded Siemens MAGNETOM Prisma^{fit} system, which includes enhanced gradient performance and upgraded software. Both Siemens scanner systems are widely used in neuroimaging studies. Although MRI scanner upgrades may introduce system effects in acquired images, we did not change the main static MRI magnet and sought to minimize image acquisition effects. Whole-brain T1-weighted anatomical data from both MRI scan sessions were acquired using a 3-dimensional magnetization-prepared rapid-acquisition gradient-echo (3D-MPRAGE) sequence without in-plane interpolation and interslice gap with the following parameters: TR (repetition time)/TE (echo time)/TI (inversion time)=3500/3.5/1100 ms; flip angle=7°; field of view=256 × 256 mm^2; matrix size=256 × 256; voxel size=1.0 × 1.0 × 1.0 mm^3; number of excitations=1; number of sagittal slices=192. Whole-brain resting-state functional MRIs were acquired using an axial T2*-weighted gradient-echo echo planar imaging sequence with the following imaging parameters: TR/TE=2500/27 ms; flip angle=77°, field of view=220 × 220 mm^2; image matrix=64 × 64; slice thickness=3.4 mm; 200 continuous image volumes; 43 interleaved partitions without an interpartition gap. An experienced researcher visually

XSL•FO
RenderX

assessed all anatomical scans for image quality to exclude images with brain trauma, tumors, hemorrhagic or infarct lesions, or severe motion artifacts.

To investigate potential brain signature alterations associated with digital dance game training, we estimated the individual voxel-wise brain maps, including gray matter volume (GMV), regional homogeneity (ReHo), fractional amplitude of low frequency fluctuations (fALFF), and global brain connectivity (GBC). During image preprocessing of structural data, we used the enhanced Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra voxel-based morphometry (DARTEL-VBM) procedure to generate individual GMV maps. Image processing was done using Statistical Parametric Mapping (SPM12, version 7487; Wellcome Institute of Neurology) software using MATLAB (R2020a, Mathworks). Details of the enhanced DARTEL-VBM framework are described in our previous studies [15-17]. Briefly, image processing of T1-weighted images entails (1) bias-field correction, (2) 3 tissue-type segmentations (gray matter, white matter, and cerebrospinal fluid), (3) spatial normalization to a study template in the Montreal Neurological Institute (MNI) space, (4) tissue modulation for the GMV map, and (5) spatial smoothing with a 6-mm Gaussian kernel. Finally, we estimated the smoothed GMV voxel-wise map, which represents the brain structure.

During image preprocessing of resting-state functional MRI (rsfMRI) data, all rsfMRI image processing was performed using a combination of the Analysis of Functional Neuroimaging (AFNI, version 20.3.03) software package, the Functional Magnetic Resonance Imaging of the Brain Software Library (FSL, version 5.0.10), and SPM12. The preprocessing of the individual rsfMRI data sets was carried out according to the methods outlined in our previous study [18], and the pipeline included the following steps: (1) removal of the first 10 volumes for steady‐state magnetization; (2) motion correction for functional volume alignment; (3) slice timing correction to correct the temporal misalignment during image acquisition across slices; (4) brain extraction to remove nonbrain tissue; (5) despiking and detrending to remove the spike artifact and polynomial trend; (6) linear regression to remove nuisance signals including Friston 24-motion parameters, white matter, and cerebrospinal fluid signals; (7) spatial normalization and resampling to the standard MNI space with a 2-mm isotropic voxel size. Subsequently, we calculated 3 functional measurements through appropriate analyses, including ReHo, fALFF, and GBC. These measurements represent aspects of spontaneous brain activity and brain architecture.

The calculation of ReHo involved assessing regional synchronization using the Kendall coefficient of concordance (KCC) for a voxel's time series in conjunction with its nearest 26 neighboring voxels [19]. Prior to ReHo computation, bandpass temporal filtering (0.01-0.08 Hz) was applied to mitigate low-frequency drift, high-frequency noise, and nonneural signals. Individual ReHo maps were generated by computing the KCC for each voxel and subsequently normalizing them by subtracting the mean value of the entire brain and dividing by the standard deviation, thereby producing standardized ReHo maps. Finally, these resultant maps were subjected to spatial smoothing using a 6-mm full-width at half-maximum (FWHM) Gaussian kernel, resulting in individual, smoothed ReHo voxel-wise maps.

In the process of measuring fALFF, the preprocessed functional data were smoothed with a 6‐mm FWHM Gaussian kernel [20]. Subsequently, the time series of each voxel was converted to the frequency domain, and the power spectrum was computed using fast Fourier transform. Voxel-wise fALFF maps were then generated by determining the fraction of power within the specific low-frequency band (0.01-0.08 Hz) relative to the power spectrum amplitudes across the entire frequency range (0-0.2 Hz). Following this calculation, the fALFF maps were normalized by subtracting the average value and dividing by the standard deviation, resulting in individualized, normalized fALFF voxel-wise maps.

Before estimating the GBC measurements, the preprocessed images underwent a bandpass temporal filter and spatial smoothing using a 6-mm FWHM Gaussian kernel [21]. The images were then resampled to 3-mm voxel size, and a gray matter mask was applied to reduce computational load. Subsequently, the individual GBC voxel-wise maps were generated by computing the correlation between each gray matter voxel and all other gray matter voxels. These correlations were then transformed into Fisher *z* scores and averaged.

Furthermore, the whole brain GMV, ReHo, fALFF, and GBC voxel-wise maps were derived by averaging values across various brain regions defined in a composite brain atlas. This atlas comprised 110 cortical regions sourced from the Harvard-Oxford template [22] and 27 cerebellar regions from a spatially unbiased infratentorial template [17]. Subsequently, these averaged brain signatures for each participant were used in subsequent group-level statistical analyses.

Secondary Outcomes

Secondary outcomes included cognitive performance, physical health (physical activity, body composition, and physical performance), mental health (resilience, demoralization, and quality of life), and related biomarkers (hematology and urinalysis).

Cognitive performance was evaluated using the Montreal Cognitive Assessment (MoCA) [23]; the MoCA contains 8 subdomains: (1) orientation, (2) short-term memory and delayed recall, (3) executive function and visuospatial ability, (4) language skills, (5) abstract thinking, (6) verbal fluency, (7) attention, and (8) clock drawing test. The MoCA score ranges from 0 to 30, with ≥26 considered normal [24].

The International Physical Activity Questionnaire (IPAQ) [25] was used to assess physical activity, which includes working, household chores, commuting, and exercise. Greater scores indicate greater physical activity. IPAQ has been used in many countries with a content validity index of 0.994 and a test-retest reliability of 0.67 for the short-form version [25,26].

Resilience was measured using the Brief Resilience Scale (BRS), which is a short, self-administered questionnaire that assesses respondents' abilities to recover from stress and adversity [27]. The BRS has 6 questions to elicit responses on a 5-point Likert-type scale, ranging from 1 (strongly disagree) to 5

XSL•FO
RenderX

(strongly agree). Responses to negatively phrased questions are reverse scored, and the mean score of all 6 items constitutes the overall resilience status, with higher scores indicating better resilience [27].

The Demoralization Scale-Mandarin Version that was used to measure demoralization comprises 24 items, each rated on a 5-point Likert scale, ranging from 0 (never) to 4 (all the time) [28,29]; a cut-off score of ≥30 is considered to affirm demoralization [28]. The German version sets cut-offs of 30 to indicate moderate demoralization and 36 to signify a high level of demoralization in patients with advanced cancer [30].

Quality of life was evaluated using the EQ-5D questionnaire, which assesses 5 domains: mobility, self-care, pain, usual activities, and psychological status [31]. Responses are categorized as grade 1 (no problem), 2 (moderate problem), or 3 (severe problem) [31]. The EQ-5D also includes a visual analogue scale from 0 to 100, with a higher score indicating better general health status [32].

Other Variables

Other study data included demographic characteristics, health behaviors (including tobacco or alcohol use), physical performance, and self-reported physician-diagnosed medical conditions.

Statistical Analysis

Based on a previous study of somatosensory games with older adults [33], we used a functional connectivity gaming comparison before and after data collection. By inputting the samples of the 2 groups (45 and 42) and the *t* value (–4.00) into a Cohen *d* calculation, the resulting Cohen *d* was –0.8582. This value was entered into G-power with a *t* test. The difference between the 2 independent means (2 groups), with a 2-tailed effect size of –0.8582 and power of 0.9, resulted in a total sample size of 60 (30 in each group).

Numerical variables are presented as the mean (SD). Discrete variables are presented as the number (percentage). Student *t* tests were used to compare the changes between pre- and postintervention, demonstrating effects of the intervention on the primary and secondary outcomes. Intention-to-treat analysis was used with the Student *t* test and mean values to manage missing data under the original RCT design. In the intervention group, associations between the study intervention and secondary outcomes were estimated using a stepwise model in regression analysis. Dance intervention performance was measured using the completion and correction rates. In the primary outcomes, we first conducted statistical analysis on the averaged GMV, then explored the statistical potential changes of functional signatures (fALFF, ReHo, and GBC) in significant structural brain regions.

A 2-sided, uncorrected *P* value <.05 and 95% CIs not spanning the null hypothesis value were considered statistically significant. All analyses were performed using SPSS Windows, version 25.0 (IBM Corp). In addition, the threshold of significance was set at *P*<.05 corrected by the false discovery rate (FDR) for multiple comparisons. Finally, to further investigate the relationship between regional GMV alterations and cognitive performance in the intervention group, exploratory partial Pearson correlation analyses were then used to examine the associations between regional GMV differences and cognitive performance difference (total MoCA score and the 8 subdomain scores), with age, sex, and total intracranial volume serving as nuisance variables. The significance threshold for the partial Pearson correlation analyses was set at an uncorrected *P*<.05.

Results

Participants

From August 31, 2020, to June 27, 2021, 94 potential participants were screened for eligibility, 60 of whom met the inclusion criteria and were enrolled and randomized (Figure 1). Table 1 shows the baseline characteristics of each study arm. Both groups had a mean age of approximately 68 years (intervention: 68.3, SD 5.1 years; control: 68.2, SD 5.5 years), and women represented 97% (29/30) of the intervention group and 60% (18/30) of the control group.

XSL•FO
RenderX

Figure 1. Study flowchart. MRI: magnetic resonance imaging.

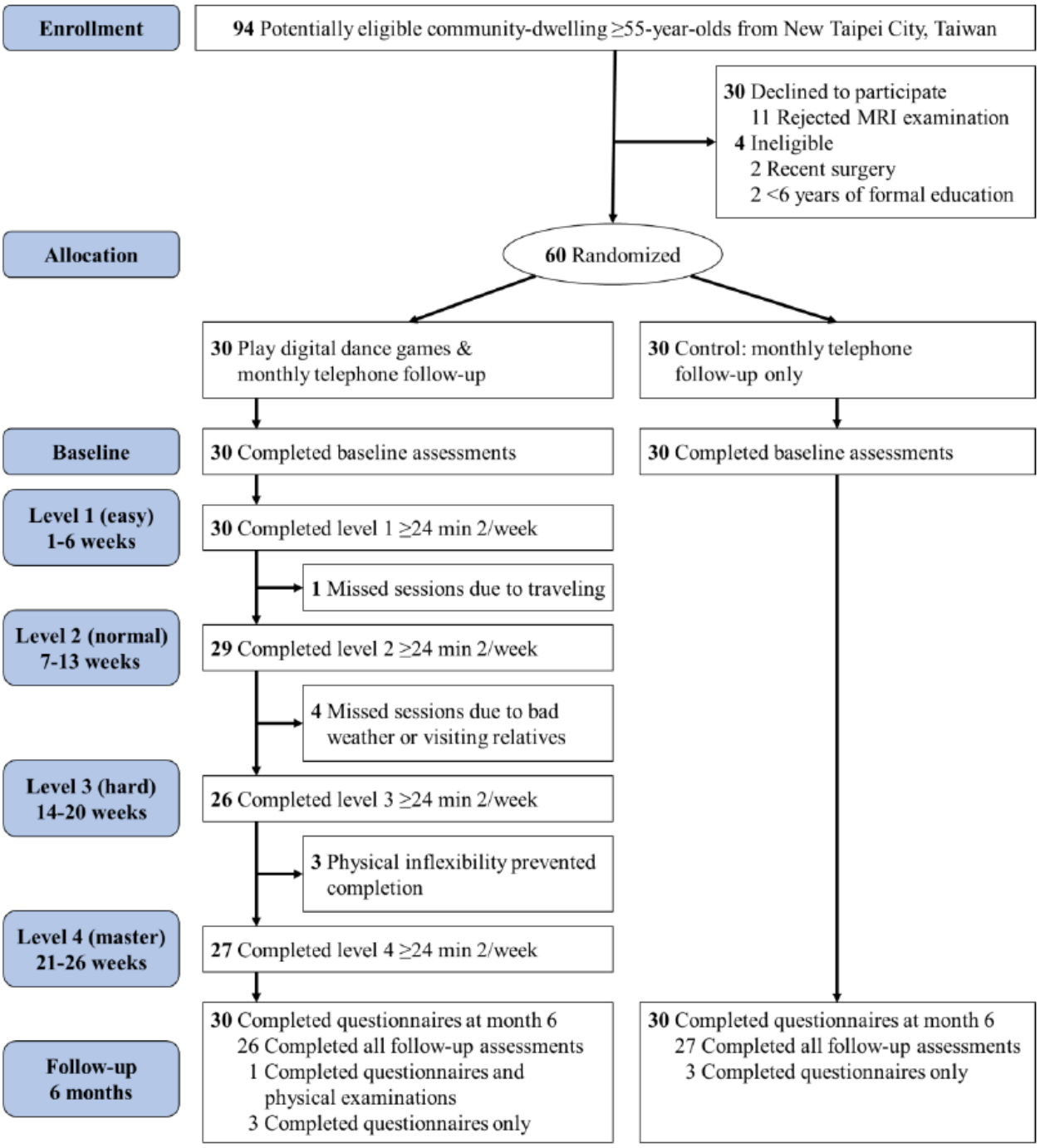

XSL•FO
RenderX

Table 1. Baseline participant characteristics.

Characteristics	Intervention group (n=30)	Control group (n=30)
Age (years), mean (SD)	68.3 (5.1)	68.2 (5.5)
Sex, n (%)		
Female	29 (97)	18 (60)
Male	1 (3)	12 (40)
Education (years), mean (SD)	12.2 (3.0)	14.1 (3.6)
Employment status, n (%)		
Full-time	2 (7)	4 (13)
Part-time	6 (20)	1 (3)
Homemaker	2 (7)	2 (7)
Retired	20 (66)	23 (77)
Alcohol consumption, n (%)		
Lifelong abstainer	9 (30)	6 (20)
Former drinker	18 (60)	10 (33)
Current drinker	3 (10)	14 (47)
Tobacco smoking, n (%)		
Never	29 (97)	23 (77)
Former smoker	1 (3)	7 (23)
Current smoker	0	0
Immunization status, n (%)		
Influenza	10 (33)	17 (57)
Pneumococcus	10 (33)	14 (47)
Varicella zoster	2 (7)	3 (10)
COVID-19	26 (87)	27 (90)
Medication use, n (%)		
Pharmaceutical agents	16 (53)	21 (70)
Health supplements	22 (73)	19 (63)

The completion rate in the intervention group was influenced by the difficulty level of dance movements and the participation rate. The easiest level (level 1) was used for the first 6 weeks, with a completion rate of 100%; however, due to the COVID-19 pandemic, participation and completion rates later fell below 100%. Table S1 in Multimedia Appendix 1 presents the details and the rationale for the 4 participants who failed to complete the scheduled intervention.

Primary Outcome

During this study, intervention participants had significantly different brain imagery of the left putamen GMV (estimate 0.016, 95% CI 0.008 to 0.024; $P<.001$, FDR-corrected $P=.14$), left pallidum GMV (estimate 0.02, 95% CI 0.006 to 0.034; $P=.004$, FDR-corrected $P=.27$), left pallidum fALFF (estimate 0.262, 95% CI 0.084 to 0.439; $P=.004$, FDR-corrected $P=.14$), and cerebellum VI GMV (estimate 0.011, 95% CI 0.003 to 0.02; $P=.01$, FDR-corrected $P=.38$; Table 2). Additionally, the intervention participants had a significant change in the GMV and GBC of the left putamen, GMV and fALFF of the left pallidus, and GMV and ReHo of the center cerebellum VI (Table 2). Figure 2 shows that, compared with the control group, the intervention group exhibited a significant increase in the slope of the left putamen GMV and left pallidus GMV and a notable decrease in the downward slope of the center cerebellum VI GMV. Furthermore, there was a significant difference in the diagonal trajectory of the left pallidus fALFF between the intervention and control groups. Figure S2 exhibits the significant relationships between the change in the left pallidum GMV and total MoCA score ($r=-0.393$, $P=.04$) and abstract thinking subdomain score of the MoCA ($r=-0.531$, $P=.004$).

XSL•FO
RenderX

Table 2. Changes in brain regions of interest between the intervention and control groups.

Variables	Intervention (n=30)				Control (n=30)				Intervention effect[a]		FDR[b]-corrected *P* value
	Baseline (T0), mean (SD)	6 months (T1), mean (SD)	T1-T0	*P* value	Baseline (T0), mean (SD)	6 months (T1), mean (SD)	T1-T0	*P* value	Estimate (95% CI)	*P* value	
Left putamen											
GMV[c]	0.50 (0.03)	0.52 (0.04)	0.01 (0.01)	<.001	0.51 (0.04)	0.51 (0.04)	0.00 (0.02)	.74	0.016 (0.008 to 0.024)	<.001	.14
ReHo[d]	–0.24 (0.19)	–0.26 (0.25)	–0.02 (0.20)	.56	–0.12 (0.28)	–0.22 (0.24)	–0.10 (0.23)	.02	0.079 (–0.028 to 0.185)	.15	.67
fALFF[e]	–0.01 (0.26)	0.04 (0.36)	0.05 (0.29)	.39	0.17 (0.39)	0.12 (0.35)	–0.05 (0.33)	.39	0.100 (–0.056 to 0.255)	.21	.65
GBC[f]	0.31 (0.02)	0.32 (0.02)	0.01 (0.02)	.03	0.32 (0.03)	0.32 (0.02)	0.00 (0.02)	.81	0.006 (–0.003 to 0.016)	.18	.59
Left pallidum											
GMV	0.28 (0.05)	0.34 (0.06)	0.06 (0.03)	<.001	0.28 (0.06)	0.32 (0.06)	0.04 (0.03)	<.001	0.020 (0.006 to 0.034)	.004	.27
ReHo	–0.59 (0.20)	–0.58 (0.24)	0.00 (0.21)	.96	–0.48 (0.24)	–0.57 (0.23)	–0.09 (0.25)	.06	0.092 (–0.023 to 0.208)	.12	.67
fALFF	–0.52 (0.23)	–0.39 (0.30)	0.13 (0.30)	.03	–0.34 (0.34)	–0.48 (0.32)	–0.13 (0.41)	.08	0.262 (0.084 to 0.439)	.004	.14
GBC	0.16 (0.01)	0.16 (0.01)	0 (0.01)	.23	0.15 (0.01)	0.15 (0.01)	0.00 (0.01)	.74	0.002 (–0.003 to 0.007)	.44	.73
Center cerebellum VI											
GMV	0.49 (0.05)	0.48 (0.05)	–0.01 (0.01)	<.001	0.48 (0.06)	0.46 (0.05)	–0.02 (0.02)	<.001	0.011 (0.003 to 0.02)	.01	.38
ReHo	–0.31 (0.25)	–0.20 (0.26)	0.10 (0.19)	.005	–0.11 (0.28)	0.11 (0.33)	0.21 (0.29)	<.001	–0.110 (–0.232 to 0.012)	.08	.56
fALFF	–0.32 (0.42)	–0.26 (0.31)	0.06 (0.35)	.37	–0.20 (0.32)	–0.11 (0.34)	0.09 (0.43)	.24	–0.035 (–0.229 to 0.160)	.73	.91
GBC	0.37 (0.01)	0.38 (0.01)	0.01 (0.02)	.14	0.38 (0.02)	0.38 (0.03)	0.01 (0.03)	.08	–0.003 (–0.014 to 0.008)	.54	.76

[a]Intervention effect computed as the between-group difference in mean change from baseline to month 6 (change in intervention group minus change in control group).

[b]FDR: false discovery rate.

[c]GMV: gray matter volume.

[d]ReHo: regional homogeneity.

[e]fALFF: fractional amplitude of low-frequency fluctuations.

[f]GBC: global brain connection.

XSL•FO
RenderX

Figure 2. Comparison between the intervention and control groups of the slopes of the (A) left putamen gray matter volume (GMV), (B) left pallidum GMV, (C) center cerebellum VI GMV, and (D) left pallidum fractional amplitude of low frequency fluctuations (fALFF).

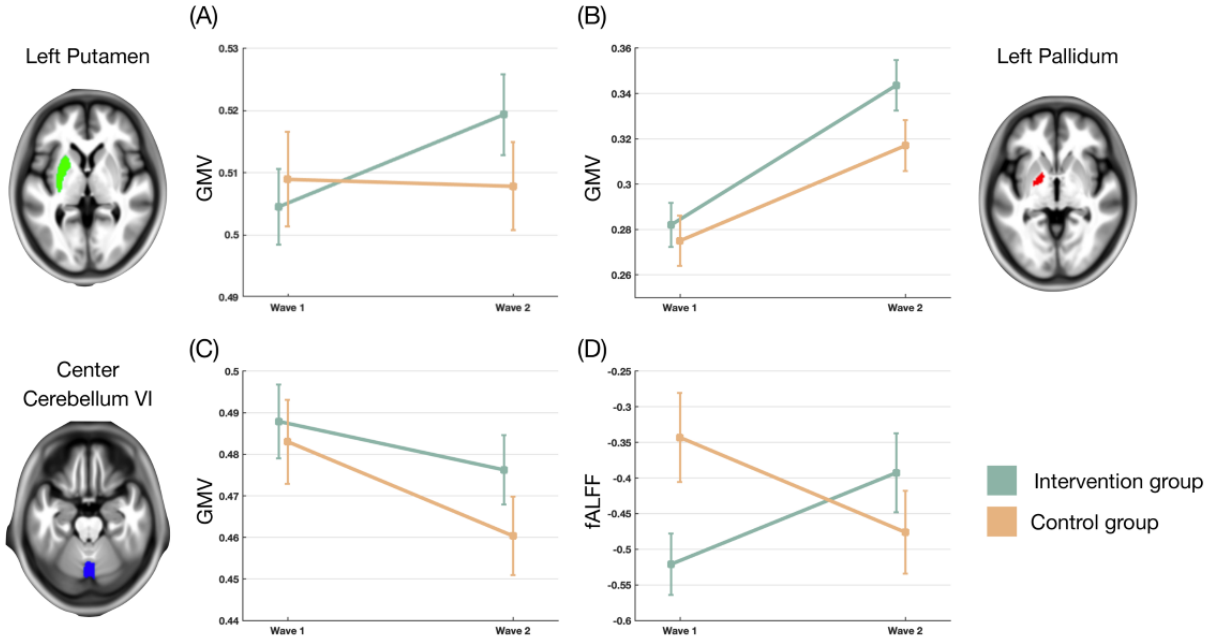

Secondary Outcomes

Table 3 also shows the changes in cognitive performance, physical activity, resilience, demoralization, and quality of life, together with estimates of the intervention effects. There were significant postintervention between-group differences in total MoCA (estimate 1.2, 95% CI 0.27 to −2.13; $P<.01$, FDR-corrected $P=.08$), time spent sitting on weekdays (estimate −1.96, 95% CI −3.33 to −0.60; $P=.005$, FDR-corrected $P=.053$), and quality of life (estimate 7.08, 95% CI 2.35 to 11.82; $P=.04$, FDR-corrected $P=.053$). The MoCA abstract thinking subdomain score significantly improved from baseline to 6 months in the intervention group but not in the control group; the difference between the groups was not statistically different. For body composition and physical performance, body weight decreased in the intervention group. In addition, body muscle mass, muscle mass of both legs, skeletal muscle mass, and calories expended per day decreased in both groups. For biomarkers, mean corpuscular volume (estimate −1.20, 95% CI −2.21 to −0.20; $P=.02$, FDR-corrected $P=.11$) and thyroid-stimulating hormone (estimate −0.69, 95% CI −1.11 to −0.28; $P=.002$, FDR-corrected $P=.053$) were different between the groups.

XSL•FO
RenderX

Table 3. Comparisons of changes in secondary outcomes between the intervention and control groups.

Outcomes	Intervention group (n=30)			Control group (n=30)			Intervention effect[a]		Corrected *P* value[b]
	Baseline, mean (SD)	6 months, mean (SD)	*P* value	Baseline, mean (SD)	6 months, mean (SD)	*P* value	Estimate (95% CI)	*P* value	
Cognitive performance (MoCA[c])									
Total score	26.03 (2.66)	27.42 (2.68)	.001	26.63 (2.22)	26.81 (2.69)	.50	1.20 (0.27 to 2.13)	.01	.08
Verbal fluency (naming)	2.65 (0.58)	2.80 (0.53)	.06	2.96 (0.18)	3.00 (0)	.27	0.12 (–0.06 to 0.29)	.18	.65
Abstract thinking	1.42 (0.54)	1.73 (0.50)	.003	1.63 (0.53)	1.85 (0.34)	.005	0.09 (–0.15 to 0.32)	.47	.79
Physical activity (IPAQ[d])									
Time spent sitting on weekdays (hours)	2.92 (2.25)	1.37 (2.18)	.002	2.44 (3.17)	2.85 (3.18)	.44	–1.96 (–3.33 to –0.60)	.005	.053
MET[e] (kcal per week)	2419.26 (1763.94)	2393.41 (1368.57)	.90	1814.36 (1006.39)	1885.38 (1326.06)	.68	–96.87 (–620.58 to 426.84)	.71	.88
5 times sit to stand test (second)	7.03 (1.71)	7.14 (1.53)	.68	8.91 (2.69)	9.65 (2.68)	.18	–0.63 (–1.84 to 0.59)	.31	.70
Resilience (BRS[f] score)	4.05 (0.45)	4.16 (0.51)	.09	3.80 (0.63)	3.89 (0.51)	.36	0.03 (–0.20 to 0.26)	.80	.89
Demoralization (total DS-MV[g] score)	16.65 (10.05)	17.57 (10.51)	.50	21.74 (12.23)	21.55 (9.99)	.89	1.11 (–2.63 to 4.85)	.56	.79
Quality of life (EQ-5D)									
Visual analog score	76.73 (9.82)	81.50 (10.09)	.02	78.06 (8.56)	75.74 (8.80)	.11	7.08 (2.35 to 11.82)	.004	.053
Body composition and physical performance									
Body weight (kg)	58.72 (7.84)	57.76 (7.71)	.007	63.73 (8.46)	63.06 (8.79)	.09	–0.30 (–1.31 to 0.72)	.56	.79
Body mass index (kg/m^2)	24.35 (3.35)	23.97 (3.29)	.005	24.09 (2.57)	23.92 (2.66)	.23	–0.20 (–0.57 to 0.17)	.29	.70
Body fat mass (kg)	19.28 (5.24)	18.87 (5.15)	.23	18.07 (5.32)	17.98 (5.09)	.85	–0.32 (–1.49 to 0.86)	.59	.79
Body fat (%)	32.40 (5.62)	32.20 (5.40)	.58	28.29 (6.88)	28.38 (6.21)	.87	–0.30 (–1.68 to 1.07)	.66	.85
Body muscle mass (kg)	21.22 (2.75)	20.93 (2.61)	.004	25.28 (4.86)	24.62 (4.25)	.12	0.38 (–0.47 to 1.23)	.38	.79
Right leg muscle mass (kg)	5.88 (0.83)	5.76 (0.83)	.008	7.13 (1.35)	6.90 (1.34)	<.001	0.05 (–0.09 to 0.19)	.50	.79
Left leg muscle mass (kg)	5.83 (0.78)	5.72 (0.78)	.005	7.06 (1.27)	6.87 (1.27)	<.001	0.09 (–0.02 to 0.19)	.12	.49
Handgrip strength (kg)	25.54 (4.77)	25.38 (4.81)	.64	28.11 (6.37)	29.09 (7.19)	.13	–1.14 (–2.57 to 0.29)	.12	.49
Knee height (cm)	45.75 (2.14)	45.50 (2.22)	.19	48.89 (2.10)	47.98 (1.97)	<.001	0.66 (0.14 to 1.19)	.01	.08
Corrected vision of right eye	0.56 (0.21)	0.59 (0.22)	.26	0.66 (0.26)	0.73 (0.24)	.03	–0.04 (–0.11 to 0.03)	.29	.70
Skeletal muscle index (kg/m2)	6.39 (0.76)	6.31 (0.78)	.02	7.00 (1.13)	6.90 (1.11)	.008	0.01 (–0.08 to 0.10)	.78	.89

XSL•FO
RenderX

Outcomes	Intervention group (n=30)			Control group (n=30)			Intervention effect[a]		Corrected *P* value[b]
	Baseline, mean (SD)	6 months, mean (SD)	*P* value	Baseline, mean (SD)	6 months, mean (SD)	*P* value	Estimate (95% CI)	*P* value	
Phase angle (rad)	4.96 (0.42)	4.86 (0.47)	.04	5.11 (0.45)	4.97 (0.43)	.01	0.05 (–0.09 to 0.19)	.50	.79
Bone mass index (gm/cm2)	2.33 (0.25)	2.28 (0.23)	<.001	2.65 (0.42)	3.30 (3.22)	.30	0.01 (–0.08 to 0.10)	.78	.89
Body moisture content (kg)	28.95 (3.37)	28.54 (3.19)	.001	33.53 (5.42)	33.11 (5.23)	.03	0.42 (–12.58 to 13.41)	.95	.98
Burn calories per day (kcal)	1221.88 (98.53)	1210.19 (93.60)	.002	1356.04 (158.98)	1343.93 (153.20)	.04	0.01 (–0.42 to 0.46)	.95	.98
Right foot resistance (Ohm)	201.05 (26.71)	206.99 (29.32)	.04	187.60 (33.30)	197.30 (33.18)	<.001	–3.75 (–10.55 to 3.04)	.27	.70
Left foot resistance (Ohm)	204.70 (25.89)	210.85 (29.56)	.02	189.50 (30.43)	198.17 (32.23)	<.001	–2.52 (–8.85 to 3.81)	.43	.79
Hematology and urinalysis									
Mean corpuscular volume (fL)	94.10 (3.64)	93.20 (3.64)	.005	93.63 (6.23)	93.93 (6.36)	.47	–1.20 (–2.21 to –0.20)	.02	.11
Alanine aminotransferase (U/L)	30.50 (27.65)	26.31 (26.24)	.004	26.89 (17.48)	22.70 (8.11)	.09	–0.01 (–5.51 to 5.50)	≥.99	≥.99
Homocysteine (umol/L)	15.18 (3.23)	12.52 (2.67)	<.001	16.65 (7.41)	14.66 (4.77)	.02	–0.68 (–2.70 to 1.34)	.50	.79
Albumin to creatinine ratio	10.77 (10.32)	12.15 (17.60)	.62	24.66 (84.97)	28.95 (107.48)	.31	–2.90 (–12.93 to 7.11)	.56	.79
High-density lipoprotein (mg/dL)	60.80 (10.61)	56.50 (9.40)	.003	59.44 (11.06)	57.40 (10.81)	.15	–2.27 (–6.08 to 1.54)	.24	.70
Thyroid-stimulating hormone (uIU/mL)	2.42 (1.28)	2.05 (1.08)	.004	1.87 (1.15)	2.19 (1.25)	.07	–0.69 (–1.11 to -0.28)	.002	.053

[a]Intervention effect computed as the between-group difference in mean change from baseline to month 6 (change in intervention group minus change in control group).

[b]Corrected for the false discovery rate.

[c]MoCA: Montreal Cognitive Assessment.

[d]IPAQ, International Physical Activity Questionnaire.

[e]MET: metabolic equivalent.

[f]BRS: Brief Resilience Scale.

[g]DS-MV: Demoralization Scale Mandarin Version.

Hierarchical regression analyses (Table 4) showed that dancing performance (completion rate of the intervention for >24 minutes per session) was independently associated with cognitive performance (β=0.647, *P*=.21) and self-rated health status on the IPAQ (β=1.159, *P*=.008). Moreover, the correct rate for the assigned dance at a difficulty level of 3 was related with resilience (β=–5.629, *P*=.007) and demoralization (β=−0.251, *P*<.001).

XSL•FO
RenderX

Table 4. Associations between dance intervention performance and secondary outcomes (n=30).

Outcomes	Unstandardized B (SE)	Standardized β	*t* test (*df*)	*P* value
Increased cognitive function (MoCA[a])				
Completion rate of interventions for >24 minutes per session	0.647 (0.264)	0.394	2.453 (4)	.02
Correct rate for assigned dance with level 2 difficulty	0.937 (0.356)	0.420	2.633 (4)	.01
Correct rate for assigned dance with level 3 difficulty	0.955 (0.295)	0.417	3.237 (4)	.003
Increased self-rated daily health status (EQ-VAS[b])				
Completion rate of interventions for >24 minutes per session	–1.159 (0.399)	–0.455	–2.905 (4)	.008
Correct rate for assigned dance with level 3 difficulty	–1.177 (0.446)	–0.331	–2.638 (4)	.01
Affecting increased resilience (BRS[c])				
Correct rate for assigned dance with level 3 difficulty	–5.629 (1.903)	–0.415	–2.958 (4)	.007
Affecting decreased demoralization level (DS-MV[d])				
Correct rate for assigned dance with level 1 difficulty	–0.382 (0.124)	–0.622	–3.081 (4)	.005
Correct rate for assigned dance with level 2 difficulty	–0.370 (0.105)	–0.628	–3.529 (4)	.002
Correct rate for assigned dance with level 3 difficulty	–0.351 (0.087)	–0.579	–4.030 (4)	<.001

[a]MoCA: Montreal Cognitive Assessment.

[b]VAS: visual analogue scale.

[c]BRS: Brief Resilience Scale.

[d]DS-MV: Demoralization Scale Mandarin Version.

Discussion

Principal Findings

To the best of our knowledge, this is the first RCT to substantiate beneficial effects of playing digital somatosensory dance games on measures of healthy aging, especially concerning neuroimaging findings. The results support the hypothesis that participating in digital somatosensory dance games not only results in favorable changes in brain structures and function in the putamen, pallidum, and cerebellum but also promotes cognitive performance, daily physical activity, healthier body composition, and quality of life.

The unique strength of this study was the robust primary outcome of brain signatures, which have not been previously reported. The clinical benefits of exergames or somatosensory games in previous studies were shown among small samples in case-control designs with relatively short intervention periods, questionnaires and interview-based assessments for outcomes, and neither biomarker measurements nor brain signatures. Our study showed evident clinical benefits of digital somatosensory dance game participation on cognitive performance, and favorable brain signature changes strongly support the biological effects of the intervention, despite the intervention activities being interrupted by the COVID-19 pandemic.

Comparison With Prior Work

Digital somatosensory dance game training entails visual observation and physical practice to acquire motor and coordination skills; this involves the integration of multiple cognitive domains, including somatosensory, motor, visual, attention, executive function, and reward processing. Previous studies have shown that these cognitive functions involve multiple brain regions, including the medial prefrontal, cingulate, premotor, parietal, and occipitotemporal cortices [34]. Our results show that digital somatosensory dance game training significantly increases structural and functional characteristics in the left putamen GMV, left pallidum GMV, center cerebellum VI GMV, and left pallidum fALFF. Pertinently, these brain regions are linked to higher-order cognitive functions, such as somatosensory, attention, episodic memory, and motor imagery [35].

Most video games include rewarding effects to increase the motivation of players to continue to participate. Such rewards may substantially alter the reward system, either functionally or structurally, in brain regions that include the striatum, pallidum, and cingulate cortex [36]. Consistent with reported reward-related effects, we also observed GMV increases in the putamen and pallidum among intervention participants; notably, we detected significant changes in the cerebellum VI, which controls motor functions and is also involved in visual working memory, verbal fluency, creativity, and cognition-related activities [37,38]. Manifest brain structure changes not only support phenotypic improvements in physical and cognitive functions but are also relevant to higher-order cognitive function. Our previous research showed that the cerebellum-limbic neurocircuit plays significant roles in the development of physiocognitive decline syndrome, which is a key phenotype associated with unhealthy aging [7]. The results of this RCT and another recent study [39] support considering the cerebellum-limbic connection as a biosignature of healthy aging. Besides benefiting physical activities and physical performance, playing the digital somatosensory dance game for 6 months significantly preserved brain regions implicated in reward processing.

XSL•FO
RenderX

Although digital dance games were originally developed for youth entertainment, they can be adapted to improve the physical and mental health of older adults [5]. The favorable effects of participating in digital somatosensory dance games support expanding their applications to promoting healthy aging rather than merely entertaining youngsters. People playing such games must perform complex and rhythmic dance moves that entail extensive coordination of visual and auditory functions and active body movements. Adcock et al [5] evaluated an exergame that incorporated tai chi and dance exercises, showing positive associations between physical activities and functioning in a small sample. Others have demonstrated the feasibility of unsupervised sensor-based exercise (step mat training and Microsoft Kinect) at home for older adults, which significantly reduced the likelihood of falls and improved cognitive function [40]. Previous research has shown that playing somatosensory games may contribute to improving cognitive function; however, which domains of cognitive function are affected has not been firmly established [6,7].

In the aging process, domain-specific declines in cognitive function suggest different trajectories of aging, with verbal, language, and executive function domains being both the earliest to decline over time and potentially reversible to promote healthy aging [41]. Improvements in physical and cognitive performance and in brain structural images associated with participating in digital somatosensory dance games are consistent with the aforementioned studies. Importantly, there was a statistically significant relationship between the rate of completing somatosensory dance game levels and improved cognitive function.

Although resilience and demoralization were not significantly improved after the intervention, dance performance was associated with resilience and demoralization. Notably, psychosocial health is an important aspect of healthy aging, and previous studies have identified demoralization and resilience as factors that are associated with good quality of life [42,43]. A lack of a statistically significant intervention effect on resilience and demoralization may be explained by ceiling effects; participants had a good quality of life, good resilience, and low demoralization levels. Nevertheless, the pleasure derived from playing the digital somatosensory dance game is important for both mental well-being and for adherence to and participation in health promotion activities. Furthermore, game-playing also provides alternatives to augment conventional multidomain intervention programs that increase the diversity of activities and ensure their effectiveness. Motor-cognitive dual-task training has been shown to improve the physical and cognitive performance of people with various neurodegenerative disorders [44,45], and digital somatosensory dance games involve a multitask format that incorporates training of motor, coordination, rhythm, reaction, memory, executive function, and other cognitive domains. In addition, the social participation and pleasure of playing games with others also lessen the risks of loneliness, social isolation, and depressive symptoms [3].

Our observations of reduced body weight in the intervention participants were concordant with those of past studies [46]. Moreover, time spent sitting on weekdays diminished after the intervention compared with baseline levels, consistent with previous findings that somatosensory games increase physical activity [7,47]. Even with the impact of the COVID-19 pandemic, participants changed their lifestyles and daily activities through regular participation in digital somatosensory dance games, which implies potential long-term benefits of playing digital somatosensory dance games in promoting healthy aging.

Associations between higher homocysteine levels and cognitive impairment have been repeatedly demonstrated [48,49]. In our study, intervention group participants had comparatively lower serum homocysteine levels, which may be relevant to important theories of neuroinflammation and DNA methylation in the development of Alzheimer disease [50]. Nevertheless, our study showed some muscle mass loss, which is inconsistent with the reduced serum levels of homocysteine in other intervention studies [51]. Similar to other studies showing the adverse effects of the COVID-19 pandemic on health promotion intervention activities [52], the reduction in muscle mass and calories expended per day in both groups may be explained by the interruption of interventions and restricted daily activity related to pandemic control measures.

Limitations

Despite its merits, this RCT had several limitations, particularly the impact of the COVID-19 pandemic, which restricted and interrupted participation in the intervention program due to disease control measures. This may have consequently attenuated beneficial effects of the intervention. In addition, the pandemic also adversely affected psychosocial well-being, which may have counteracted possible intervention effects on resilience and demoralization. Notably, quality of life was different between groups, which supports the effect of gamification. The predominance of female participants, reflecting that women tend to participate in recreational activities more than men, may also have led to sex-specific effects on the study results [3-5]. Given these limitations, the evident favorable effects of digital somatosensory dance game participation on cognitive performance and brain signatures are very encouraging for healthy aging. Although digital somatosensory dance game participation benefits older adults, it is important to generalize to availability within community settings and to further evaluate the effectiveness and sustainability of such interventions.

Conclusions

Participation in digital somatosensory dance games for 6 months not only facilitated structural and functional alterations in various brain regions associated with somatosensory, motor, visual, and attentional functions but also improved physical and mental health. Our findings also suggest that distinctive alterations in the reward system and cerebellum play a significant role in improving motor function, skilled movements, and motivation. Gamifying interventions may provide a potentially sustainable, effective approach to promoting healthy aging and may be considered an effective component of community-based, multidomain, healthy aging programs.

XSL•FO
RenderX

Acknowledgments

David Neil PhD, of Dr. Word Ltd, Taiwan, provided editorial services supported by funding from Taipei Veterans General Hospital.

This study was funded by the International Games System Company, New Taipei City, Taiwan, and National Health Research Institutes, Taiwan (NHRI-11A1-CG-CO-01-2225-1).

Authors' Contributions

HHT conceptualized the study, curated the data, performed the formal analysis, designed the methodology, performed project administration, validated the data, created data visualizations, and wrote the original manuscript draft. CWC curated the data, performed the investigation, and performed project administration. CYK and PLL analyzed, validated, and visualized the data. KHC and CPL validated and visualized the data. LKC conceptualized the study, designed the methodology, validated the data, reviewed and edited the manuscript draft, and gave final approval for the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Further details of the dance intervention.
[DOCX File , 632 KB - jmir_v26i1e57694_app1.docx]

Multimedia Appendix 2
CONSORT checklist.
[PDF File (Adobe PDF File), 33 KB - jmir_v26i1e57694_app2.pdf]

References

1. World report on ageing and health. World Health Organization. 2015 Sep 29. URL: https://www.who.int/publications-detail-redirect/9789241565042 [accessed 2024-07-11]
2. WHO's work on the UN Decade of Healthy Ageing (2021-2030). World Health Organization. URL: https://www.who.int/initiatives/decade-of-healthy-ageing [accessed 2024-07-11]
3. Chen L, Hwang A, Lee W, Peng L, Lin M, Neil DL, Taiwan Health Promotion Intervention Study for Elders research group. Efficacy of multidomain interventions to improve physical frailty, depression and cognition: data from cluster-randomized controlled trials. J Cachexia Sarcopenia Muscle 2020 Jun;11(3):650-662. [doi: 10.1002/jcsm.12534] [Medline: 32134208]
4. Lee W, Peng L, Lin C, Chen R, Lin S, Loh C, et al. Effects of incorporating multidomain interventions into integrated primary care on quality of life: a randomised controlled trial. The Lancet Healthy Longevity 2021 Nov;2(11):e712-e723. [doi: 10.1016/s2666-7568(21)00248-8]
5. Adcock M, Sonder F, Schättin A, Gennaro F, de Bruin ED. A usability study of a multicomponent video game-based training for older adults. Eur Rev Aging Phys Act 2020;17:3 [FREE Full text] [doi: 10.1186/s11556-019-0233-2] [Medline: 31938075]
6. Liu Y, Liu JC, Lin M, Chen W. Participation of senior citizens in somatosensory games: a correlation between the willingness to exercise and happiness. J Ambient Intell Human Comput 2020 Apr 13:1. [doi: 10.1007/s12652-020-01918-y]
7. Yu C, Rau PP, Liu X. Development and preliminary usability evaluation of a somatosensory square dance system for older Chinese persons: mixed methods study. JMIR Serious Games 2020 May 28;8(2):e16000 [FREE Full text] [doi: 10.2196/16000] [Medline: 32463376]
8. Yoong SQ, Wu VX, Jiang Y. Experiences of older adults participating in dance exergames: A systematic review and meta-synthesis. Int J Nurs Stud 2024 Apr;152:104696 [FREE Full text] [doi: 10.1016/j.ijnurstu.2024.104696] [Medline: 38301305]
9. Wiemeyer J, Kliem A. Serious games in prevention and rehabilitation—a new panacea for elderly people? Eur Rev Aging Phys Act 2011 Dec 08;9(1):41-50. [doi: 10.1007/s11556-011-0093-x]
10. Gomes GCV, Simões MS, Lin SM, Bacha JMR, Viveiro LAP, Varise EM, et al. Feasibility, safety, acceptability, and functional outcomes of playing Nintendo Wii Fit Plus for frail older adults: A randomized feasibility clinical trial. Maturitas 2018 Dec;118:20-28. [doi: 10.1016/j.maturitas.2018.10.002] [Medline: 30415751]
11. West GL, Zendel BR, Konishi K, Benady-Chorney J, Bohbot VD, Peretz I, et al. Playing Super Mario 64 increases hippocampal grey matter in older adults. PLoS One 2017 Dec 6;12(12):e0187779 [FREE Full text] [doi: 10.1371/journal.pone.0187779] [Medline: 29211727]

XSL•FO
RenderX

12. Kral TRA, Stodola DE, Birn RM, Mumford JA, Solis E, Flook L, et al. Neural correlates of video game empathy training in adolescents: a randomized trial. NPJ Sci Learn 2018 Aug 7;3(1):13 [FREE Full text] [doi: 10.1038/s41539-018-0029-6] [Medline: 30631474]
13. West GL, Konishi K, Diarra M, Benady-Chorney J, Drisdelle BL, Dahmani L, et al. Impact of video games on plasticity of the hippocampus. Mol Psychiatry 2018 Jul 8;23(7):1566-1574. [doi: 10.1038/mp.2017.155] [Medline: 28785110]
14. Trinkler I, Chéhère P, Salgues J, Monin M, Tezenas du Montcel S, Khani S, et al. Contemporary dance practice improves motor function and body representation in Huntington's disease: a pilot study. JHD 2019 Feb 13;8(1):97-110. [doi: 10.3233/jhd-180315]
15. Chou K, Lee P, Peng L, Lee WJ, Wang PN, Chen LK, et al. Classification differentiates clinical and neuroanatomic features of cerebral small vessel disease. Brain Commun 2021;3(2):fcab107 [FREE Full text] [doi: 10.1093/braincomms/fcab107] [Medline: 34131645]
16. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006 Jul 01;31(3):968-980. [doi: 10.1016/j.neuroimage.2006.01.021] [Medline: 16530430]
17. Diedrichsen J. A spatially unbiased atlas template of the human cerebellum. Neuroimage 2006 Oct 15;33(1):127-138. [doi: 10.1016/j.neuroimage.2006.05.056] [Medline: 16904911]
18. Chou K, Kuo C, Liang C, Lee P, Tsai C, Tsai C, et al. Shared patterns of brain functional connectivity for the comorbidity between migraine and insomnia. Biomedicines 2021 Oct 09;9(10):1420 [FREE Full text] [doi: 10.3390/biomedicines9101420] [Medline: 34680538]
19. Zang Y, Jiang T, Lu Y, He Y, Tian L. Regional homogeneity approach to fMRI data analysis. Neuroimage 2004 May;22(1):394-400. [doi: 10.1016/j.neuroimage.2003.12.030] [Medline: 15110032]
20. Zou Q, Zhu C, Yang Y, Zuo X, Long X, Cao Q, et al. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. J Neurosci Methods 2008 Jul 15;172(1):137-141 [FREE Full text] [doi: 10.1016/j.jneumeth.2008.04.012] [Medline: 18501969]
21. Cole MW, Pathak S, Schneider W. Identifying the brain's most globally connected regions. Neuroimage 2010 Feb 15;49(4):3132-3148. [doi: 10.1016/j.neuroimage.2009.11.001] [Medline: 19909818]
22. Atlases. FSL. URL: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases [accessed 2024-07-11]
23. Coen RF, Robertson DA, Kenny RA, King-Kallimanis BL. Strengths and limitations of the MoCA for assessing cognitive functioning: findings from a large representative sample of Irish older adults. J Geriatr Psychiatry Neurol 2016 Jan 06;29(1):18-24. [doi: 10.1177/0891988715598236] [Medline: 26251108]
24. Kim H, Yu K, Lee B, Kim B, Kang Y. Validity of the Montreal Cognitive Assessment (MoCA) index scores: a comparison with the cognitive domain scores of the Seoul Neuropsychological Screening Battery (SNSB). Dement Neurocogn Disord 2021 Jul;20(3):28-37 [FREE Full text] [doi: 10.12779/dnd.2021.20.3.28] [Medline: 34354755]
25. Liou YM, Jwo CJC, Yao KG, Chiang L, Huang L. Selection of appropriate Chinese terms to represent intensity and types of physical activity terms for use in the Taiwan version of IPAQ. Journal of Nursing Research 2008;16(4):252-263. [doi: 10.1097/01.JNR.0000387313.20386.0a]
26. Zwolinsky S, McKenna J, Pringle A, Widdop P, Griffiths C. Physical activity assessment for public health: efficacious use of the single-item measure. Public Health 2015 Dec;129(12):1630-1636. [doi: 10.1016/j.puhe.2015.07.015] [Medline: 26296848]
27. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
28. Kissane DW, Wein S, Love A, Lee XQ, Kee PL, Clarke DM. The Demoralization Scale: a report of its development and preliminary validation. J Palliat Care 2004;20(4):269-276. [Medline: 15690829]
29. Hung HC, Chen HW, Chang YF, Yang YC, Liu CL, Hsieh RK, et al. Evaluation of the reliability and validity of the Mandarin version of demoralization scale for cancer patients. Journal of Internal Medicine of Taiwan 2010;21(6):427-435. [doi: 10.6314/JIMT.2010.21(6).06]
30. Mehnert A, Vehling S, Höcker A, Lehmann C, Koch U. Demoralization and depression in patients with advanced cancer: validation of the German version of the demoralization scale. J Pain Symptom Manage 2011 Nov;42(5):768-776 [FREE Full text] [doi: 10.1016/j.jpainsymman.2011.02.013] [Medline: 21592722]
31. Lin H, Li C, Lin FJ, Chang J, Gau C, Luo N, et al. Valuation of the EQ-5D-5L in Taiwan. PLoS One 2018 Dec 26;13(12):e0209344 [FREE Full text] [doi: 10.1371/journal.pone.0209344] [Medline: 30586400]
32. Schrag A, Selai C, Jahanshahi M, Quinn NP. The EQ-5D--a generic quality of life measure-is a useful instrument to measure quality of life in patients with Parkinson's disease. J Neurol Neurosurg Psychiatry 2000 Jul;69(1):67-73 [FREE Full text] [doi: 10.1136/jnnp.69.1.67] [Medline: 10864606]
33. Wu L, Xu J, Song K, Zhu L, Zhou N, Xu L, et al. Emotional bias modification weakens game-related compulsivity and reshapes frontostriatal pathways. Brain 2022 Dec 19;145(12):4210-4221. [doi: 10.1093/brain/awac267] [Medline: 35861265]
34. Cross ES, Kraemer DJ, Hamilton AFDC, Kelley WM, Grafton ST. Sensitivity of the action observation network to physical and observational learning. Cereb Cortex 2009 Feb 30;19(2):315-326 [FREE Full text] [doi: 10.1093/cercor/bhn083] [Medline: 18515297]

XSL•FO
RenderX

35. Cavanna A, Trimble M. The precuneus: a review of its functional anatomy and behavioural correlates. Brain 2006 Mar;129(Pt 3):564-583. [doi: 10.1093/brain/awl004] [Medline: 16399806]
36. Palaus M, Marron EM, Viejo-Sobera R, Redolar-Ripoll D. Neural basis of video gaming: a systematic review. Front Hum Neurosci 2017 May 22;11:248 [FREE Full text] [doi: 10.3389/fnhum.2017.00248] [Medline: 28588464]
37. Schmahmann JD. The cerebellum and cognition. Neurosci Lett 2019 Jan 01;688:62-75. [doi: 10.1016/j.neulet.2018.07.005] [Medline: 29997061]
38. Gao Z, Liu X, Zhang D, Liu M, Hao N. The indispensable role of the cerebellum in visual divergent thinking. Sci Rep 2020 Oct 06;10(1):16552 [FREE Full text] [doi: 10.1038/s41598-020-73679-9] [Medline: 33024190]
39. Jung SJ, Vlasov K, D'Ambra AF, Parigi A, Baya M, Frez EP, et al. Novel cerebello-amygdala connections provide missing link between cerebellum and limbic system. Front Syst Neurosci 2022 May 13;16:879634 [FREE Full text] [doi: 10.3389/fnsys.2022.879634] [Medline: 35645738]
40. Gschwind YJ, Schoene D, Lord SR, Ejupi A, Valenzuela T, Aal K, et al. The effect of sensor-based exercise at home on functional performance associated with fall risk in older people - a comparison of two exergame interventions. Eur Rev Aging Phys Act 2015 Nov 30;12(1):11 [FREE Full text] [doi: 10.1186/s11556-015-0156-5] [Medline: 26865875]
41. Wu Y, Liu L, Chen W, Lee W, Peng L, Wang P, et al. Cognitive function in individuals with physical frailty but without dementia or cognitive complaints: results from the I-Lan Longitudinal Aging Study. J Am Med Dir Assoc 2015 Oct 01;16(10):899.e9-899.16. [doi: 10.1016/j.jamda.2015.07.013] [Medline: 26321467]
42. Chang T, Huang P, Hsu C, Yen T. Demoralization in oral cancer inpatients and its association with spiritual needs, quality of life, and suicidal ideation: a cross-sectional study. Health Qual Life Outcomes 2022 Apr 02;20(1):60 [FREE Full text] [doi: 10.1186/s12955-022-01962-6] [Medline: 35366908]
43. Koivunen K, Portegijs E, Sillanpää E, Eronen J, Kokko K, Rantanen T. Maintenance of high quality of life as an indicator of resilience during COVID-19 social distancing among community-dwelling older adults in Finland. Qual Life Res 2022 Mar 27;31(3):713-722 [FREE Full text] [doi: 10.1007/s11136-021-03002-0] [Medline: 34570331]
44. Kalyani H, Sullivan K, Moyle G, Brauer S, Jeffrey ER, Roeder L, et al. Effects of dance on gait, cognition, and dual-tasking in Parkinson's disease: a systematic review and meta-analysis. JPD 2019 May 23;9(2):335-349. [doi: 10.3233/jpd-181516]
45. De Freitas TB, Leite PHW, Doná F, Pompeu JE, Swarowsky A, Torriani-Pasin C. The effects of dual task gait and balance training in Parkinson's disease: a systematic review. Physiother Theory Pract 2020 Oct 03;36(10):1088-1096. [doi: 10.1080/09593985.2018.1551455] [Medline: 30501424]
46. Wu H, Tu J, Hsu C, Tsao TH. Effects of low-impact dance on blood biochemistry, bone mineral density, the joint range of motion of lower extremities, knee extension torque, and fall in females. J Aging Phys Act 2016 Jan;24(1):1-7. [doi: 10.1123/japa.2014-0088] [Medline: 25642949]
47. Liu L, Chou K, Hsu CH, Peng L, Lee W, Chen W, et al. Cerebellar-limbic neurocircuit is the novel biosignature of physio-cognitive decline syndrome. Aging (Albany NY) 2020 Nov 25;12(24):25319-25336 [FREE Full text] [doi: 10.18632/aging.104135] [Medline: 33234736]
48. Li R, Weng H, Pan Y, Meng X, Liao X, Wang M, et al. Relationship between homocysteine levels and post-stroke cognitive impairment in female and male population: from a prospective multicenter study. J Transl Int Med 2021 Dec 01;9(4):264-272 [FREE Full text] [doi: 10.2478/jtim-2021-0035] [Medline: 35136725]
49. Tawfik A, Elsherbiny NM, Zaidi Y, Rajpurohit P. Homocysteine and age-related central nervous system diseases: role of inflammation. Int J Mol Sci 2021 Jun 10;22(12):6259 [FREE Full text] [doi: 10.3390/ijms22126259] [Medline: 34200792]
50. Qazi TJ, Quan Z, Mir A, Qing H. Epigenetics in Alzheimer's disease: perspective of DNA methylation. Mol Neurobiol 2018 Feb 14;55(2):1026-1044. [doi: 10.1007/s12035-016-0357-6] [Medline: 28092081]
51. Vidoni M, Pettee Gabriel K, Luo S, Simonsick EM, Day RS. Relationship between homocysteine and muscle strength decline: the Baltimore Longitudinal Study of Aging. J Gerontol A Biol Sci Med Sci 2018 Mar 14;73(4):546-551 [FREE Full text] [doi: 10.1093/gerona/glx161] [Medline: 28958086]
52. Hansen P, Main C, Hartling L. Dance intervention affects social connections and body appreciation among older adults in the long term despite COVID-19 social isolation: a mixed methods pilot study. Front Psychol 2021 Feb 25;12:635938 [FREE Full text] [doi: 10.3389/fpsyg.2021.635938] [Medline: 33716905]

Abbreviations

3D-MPRAGE: 3-dimensional magnetization-prepared rapid-acquisition gradient-echo
AFNI: Analysis of Functional Neuroimaging
BRS: Brief Resilience Scale
DARTEL-VBM: Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra voxel-based morphometry
fALFF: fractional amplitude of low frequency fluctuations
FDR: false discovery rate
FSL: Functional Magnetic Resonance Imaging of the Brain Software Library
FWHM: full-width at half-maximum

XSL•FO
RenderX

GBC: global brain connectivity
GMV: gray matter volume
IPAQ: International Physical Activity Questionnaire
KCC: Kendall coefficient of concordance
MNI: Montreal Neurological Institute
MoCA: Montreal Cognitive Assessment
MRI: magnetic resonance imaging
RCT: randomized controlled trial
ReHo: regional homogeneity
rsfMRI: resting-state functional MRI
SPM: Statistical Parametric Mapping
TE: echo time
TI: inversion time
TR: repetition time
WHO: World Health Organization

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 24.02.24; peer-reviewed by J Verghese; comments to author 22.03.24; revised version received 13.05.24; accepted 05.07.24; published 30.07.24.

Please cite as:
Tung HH, Kuo CY, Lee PL, Chang CW, Chou KH, Lin CP, Chen LK
Efficacy of Digital Dance on Brain Imagery, Cognition, and Health: Randomized Controlled Trial
J Med Internet Res 2024;26:e57694
URL: https://www.jmir.org/2024/1/e57694
doi:10.2196/57694
PMID:39078687

©Heng-Hsin Tung, Chen-Yuan Kuo, Pei-Lin Lee, Chih-Wen Chang, Kun-Hsien Chou, Ching-Po Lin, Liang-Kung Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of a Gamified Intervention on Daily Steps in Real-Life Conditions: Retrospective Analysis of 4800 Individuals

Alexandre Mazéas[1,2,3], PhD; Cyril Forestier[4], PhD; Guillaume Harel[3], MSc; Martine Duclos[2,5], MD, PhD; Aïna Chalabaev[1], PhD

[1]Laboratoire Sport et Environnement Social (SENS), Université Grenoble Alpes, Grenoble, France
[2]National Research Institute for Agriculture, Food and Environment (INRAE), Clermont-Ferrand, France
[3]Kiplin, Nantes, France
[4]Laboratoire Motricité, Interactions, Performance (MIP - UR4334), Nantes Université, Nantes, France
[5]Department of Sport Medicine and Functional Exploration, University Hospital Clermont-Ferrand, Hospital G. Montpied, Clermont-Ferrand, France

Corresponding Author:
Alexandre Mazéas, PhD
Laboratoire Sport et Environnement Social (SENS)
Université Grenoble Alpes
1741 Rue de la Piscine
Grenoble, 38000
France
Phone: 33 476635081
Email: alexandre.mazeas@univ-grenoble-alpes.fr

Abstract

Background: Digital interventions integrating gamification features hold promise to promote daily steps. However, results regarding the effectiveness of this type of intervention are heterogeneous and not yet confirmed in real-life contexts.

Objective: This study aims to examine the effectiveness of a gamified intervention and its potential moderators in a large sample using real-world data. Specifically, we tested (1) whether a gamified intervention enhanced daily steps during the intervention and follow-up periods compared to baseline, (2) whether this enhancement was higher in participants in the intervention than in nonparticipants, and (3) what participant characteristics or intervention parameters moderated the effect of the program.

Methods: Data from 4819 individuals who registered for a mobile health Kiplin program between 2019 and 2022 were retrospectively analyzed. In this intervention, participants could take part in one or several games in which their daily step count was tracked, allowing individuals to play with their overall activity. Nonparticipants were people who registered for the program but did not take part in the intervention and were considered as a control group. Daily step counts were measured via accelerometers embedded in either commercial wearables or smartphones of the participants. Exposure to the intervention, the intervention content, and participants' characteristics were included in multilevel models to test the study objectives.

Results: Participants in the intervention group demonstrated a significantly greater increase in mean daily steps from baseline than nonparticipants ($P<.001$). However, intervention effectiveness depended on participants' initial physical activity. The daily steps of participants with <7500 baseline daily steps significantly improved from baseline both during the Kiplin intervention (+3291 daily steps) and the follow-up period (+945 daily steps), whereas participants with a higher baseline had no improvement or significant decreases in daily steps after the intervention. Age ($P<.001$) and exposure ($P<.001$) positively moderated the intervention effect.

Conclusions: In real-world settings and among a large sample, the Kiplin intervention was significantly effective in increasing the daily steps of participants from baseline during intervention and follow-up periods compared to nonparticipants. Interestingly, responses to the intervention differed based on participants' initial steps, with the existence of a plateau effect. Drawing on the insights of self-determination theory, we can assume that the effect of gamification could depend of the initial motivation and activity of participants.

(J Med Internet Res 2024;26:e47116) doi:10.2196/47116

XSL•FO
RenderX

KEYWORDS

behavior change; daily steps; exercise; gamification; intervention; mHealth; mobile health; mobile phone; physical activity; real world data; retrospective; self-determination theory; step; steps

Introduction

Background

Physically inactive individuals are at higher risk of developing noncommunicable diseases—such as cardiovascular diseases, cancers, type 2 diabetes mellitus, or obesity—and mental health issues than those who are most active [1]. However, one-third of the world's population is insufficiently active [2,3], and the trend is downward, with adults performing on average 1000 fewer steps per day than 2 decades ago [4]. In addition, it has recently been reported that the global population step count did not return to prepandemic levels in the 2 years following the onset of the COVID-19 outbreak [5]. The number of steps per day is a simple and convenient measure of physical activity (PA). Recent research suggests that an increase in the daily step count is associated with a progressively lower risk of all-cause mortality. Walking an additional 1000 steps per day can help reduce the risk of all-cause mortality [6]. For adults aged ≥60 years, this reduction in mortality rates is observed with up to approximately 6000 to 8000 steps per day, whereas for adults aged <60 years, the threshold is approximately 8000 to 10,000 steps per day [7]. However, sustaining this increase over time is crucial to achieve tangible health benefits [8]. Despite the efficacy of current programs in eliciting initial changes in individuals' PA, they often struggle with inducing long-term behavioral shifts [9]. In this context, there is an urgent need to sustainably increase the number of daily steps of individuals in primary, secondary, and tertiary prevention.

Digital behavior change interventions are promising avenues to promote daily steps. Smartphones and digital tools, ubiquitous in our daily lives, offer several advantages, including their widespread availability, relatively low cost, and ability to access content quickly from anywhere [10-12]. Moreover, these technologies can collect real-time data in natural contexts (ie, daily step counts can be measured via accelerometers embedded in either commercial wearables such as Fitbit or smartphones) and present them in quantified formats, providing opportunities for exploration and reflection. This facilitates the implementation of powerful behavior change techniques such as goal setting and self-monitoring, potentially influencing behaviors [11]. However, there are concerns about the ability of digital programs to engage participants once the novelty wears off or to be effective on any type of audience regardless of their age, sociodemographic characteristics, or health status. In this context, gamification strategies introduce an exciting road map for addressing these challenges.

Gamification refers to the use of game elements in nongame contexts [13] and allows for the transformation of a routine activity into a more engaging one. Self-determination theory (SDT) [14] is a commonly used theoretical framework for understanding the motivational impact of gamification on behavior. SDT suggests the existence of different types of motivation that can be pictured on a continuum ranging from lack of motivation to completely autonomous motivation in which the behavior comes from the individual's will. By contrast, controlled motivation will lead the individual to practice for the consequences that the activity can bring and not for the activity itself. SDT holds that people will be more likely to perform the behavior in the long term when their motivation is autonomous rather than controlled. Thus, autonomous forms of motivation represent more sustainable drivers of engagement and are an important predictor of the long-term maintenance of physical practice [15,16]. Autonomous motivation occurs when people perform an activity for their own satisfaction, inherent interest, and enjoyment. Moreover, 3 basic psychological needs are presumed to achieve self-determination: the need for autonomy (ie, need to feel responsible for one's own actions), competence (ie, need to feel effective in one's interactions with the environment), and relatedness (ie, need to feel connected to other people).

In addition to providing fun and playful experiences to users, gamification can effectively address basic psychological needs [17]. First, gamification strategies such as point scoring, badges, levels, and competitions serve to sustain the need for competence by offering feedback on users' behaviors. Second, customizable game environments or user choices can support autonomy. Finally, features such as leaderboards, team structures, groups, or communication functions can foster a sense of relatedness. From this perspective, a gamified intervention would feed the autonomous motivation of participants and would be more correlated with the long-term adherence to PA. However, from another perspective, several criticisms have been leveled at gamification, including the fact that these mechanisms are reward oriented and that, still in line with SDT, the use of external rewards can reduce autonomous motivation [18,19].

A recent meta-analysis [20] revealed that digital gamified interventions lasting on average 12 weeks improved daily steps by 1600 steps on average. Importantly, the results showed that gamified interventions (1) appear more effective than digital nongamified interventions, (2) seem appropriate for any type of user regardless of their age or health status, and (3) lead to a persistent PA improvement after follow-up periods lasting on average 14 weeks with a very small to small effect size. As a result, gamified interventions are emerging as interesting behavior change tools to tackle the physical inactivity pandemic. However, these findings obtained from randomized controlled trials do not always reflect what happens in real-life settings [21]. In addition, the effect sizes reported in this meta-analysis were heterogeneous, and the authors found high between-study heterogeneity (eg, I^2=82%).

If this heterogeneity can be explained by differences in study quality or diversity of designs in the included studies, the behavior change intervention ontology proposed by Michie et al [22] argues that heterogeneity in behavioral interventions could also be explained by different variables such as

XSL•FO
RenderX

intervention characteristics (eg, content and delivery), the context (eg, characteristics of the population targeted, such as demographics, and setting, such as the policy environment or physical location), exposure of participants to the program (eg, engagement and reach), and the mechanisms of action (the processes through which interventions influence the target behavior). Considering these variables within gamification contexts could provide a useful means to better understand the conditions under which interventions are successful. Furthermore, based on SDT, we can envisage that gamification techniques will not have the same impact on all users depending on their initial motivation and the way they perceive games.

This study investigated these questions based on a retrospective analysis of real-world data collected from a large sample of adult participants who were proposed a mobile health gamified intervention developed by the company Kiplin in France from 2019 to 2022. In this intervention, participants could take part in one or several collective games in which their daily step count was tracked, allowing individuals to play with their overall activity. In addition to offering the possibility of direct intervention on people's activity habits in a natural context, the capacity of this mobile app to collect a large amount of objective real-world data in real time can be useful for understanding the processes and outcomes of behavioral health interventions [23]. More specifically, these data can help make explicit when, where, for whom, and in what state for the participant the intervention will produce the expected effect, notably owing to continuous data collection over time. The within-person evolution in daily steps obtained via the app combined with between-person individual factors and intervention parameters is of great interest in this perspective.

Objectives

Thus, the objectives of this study were to analyze the data collected to (1) examine within-individual evolutions of daily steps before, during, and after the intervention; (2) test the effectiveness of a gamified program in real-life conditions on daily steps among participants versus nonparticipants; and (3) explore the variables that could explain heterogeneity in responses to the intervention. On the basis of previous results on gamification [20], we first hypothesized that daily steps would increase during and after the gamified program compared to baseline (hypothesis 1). Second, we hypothesized that this improvement will be greater for participants than for nonparticipants (ie, participants who registered on the app but did not complete any games; hypothesis 2). Finally, we expected that the intervention's characteristics (ie, type and number of games), the context within which the intervention was performed (ie, population and setting), and the exposure to the intervention (ie, engagement of participants with the app) will moderate the intervention effect (hypothesis 3).

Methods

Study Design and Participants

This study retrospectively analyzed data from adult participants who had registered for a Kiplin program and had given consent for their data to be collected. To be included, participants must be aged ≥18 years; have registered on the app between January 1, 2019, and January 2, 2022; and logged daily steps (measured via their smartphone or an activity monitor) on a time frame of at least 90 days with <20% of missing daily observations. Of the 134,040 individuals who registered on the Kiplin app during this time span, 4819 (3.6%) met the eligibility criteria. Figure 1 shows the study flowchart.

Figure 1. Study flowchart and screening of the Kiplin database.

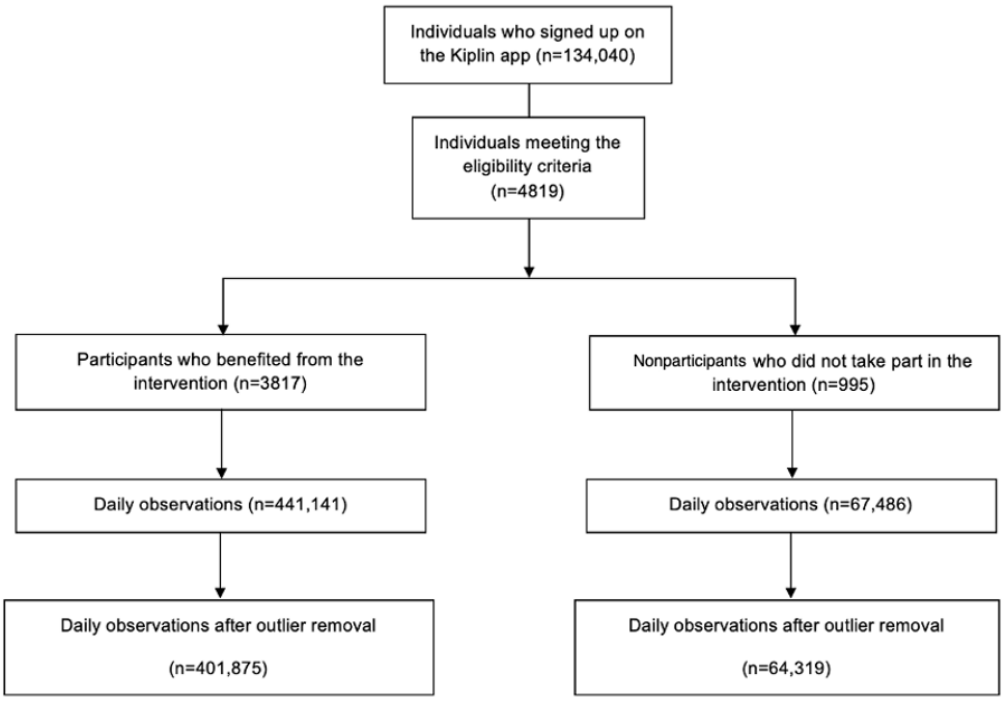

XSL•FO
RenderX

Nonwear days were defined as days with <1000 steps and considered as missing observations—as previous research has suggested that daily step values of <1000 may not represent full data capture [24,25]. Days before the first day of the first game were considered as *baseline* (median 14, SD 42.9 days), the period between the first day of the first game and the last day of the last game was considered as the *intervention period* (median 19, SD 31.2 days), and the days after the last day of the last game were considered as the *follow-up* (median 90, SD 22.8 days). We restricted the follow-up periods to 90 days after the intervention (ie, 3 months).

Participants could receive the Kiplin intervention (1) in the context of their work (ie, primary prevention with employees), (2) in an older adult program (ie, primary prevention with volunteer retirees), or (3) as part of their chronic disease care (ie, patients mainly treated for obesity or cancer). In all the aforementioned conditions, the program was paid not by the participant but by their employer or health care center.

Some participants registered for the program and created an account but did not take part in the intervention (ie, did not complete any games). These individuals were considered *nonparticipants* and were used as a control group (as proposed in previous research [26]). Similarly, the baseline period for these nonparticipants corresponded to the days before the date in which they were supposed to start the intervention period.

Ethical Considerations

This study was approved by the local ethics committee (IRB00013412; CHU de Clermont-Ferrand institutional review board 1; institutional review board number 2022-CF063) with compliance with the French policy of individual data protection.

The Kiplin Intervention

The Kiplin intervention proposes time-efficient collective games accessible through an Android or iOS app. In all games, participants' daily step counts are converted into points, allowing for progression within the games. The Kiplin app retrieves participants' daily step counts by integrating with the application programming interfaces (APIs) of the apps used by participants to track their activity (such as Apple Health for iPhone users, Google Health for Android users, and Garmin Health). In this way, participants could connect a wearable if they already owned one. In addition, participants had access to a visual tool to monitor their daily and weekly step counts and to a chat for communication with other participants. Depending on the program, participants were offered one or several games lasting approximately 14 days each. If several games were proposed, these games followed each other in an interval of <60 days.

Participants could take part in 4 different games with no option for selection. In The Adventure, the objective was to reach step goals collectively to progress toward a final destination. Players could track their progress on a map, with checkpoints representing distances between different cities of a digital world tour (Figure 2A). In The Mission, participants engaged in PA and collective challenges to unlock clues and attempt to solve missions (Figure 2B). In The Board Game, participants took on the role of forest rangers tasked with extinguishing fires. Achieving step goals allowed for progress on the board and advancement to higher levels, ultimately aiming to extinguish all fires and save forest residents (Figure 2C). Finally, in The Challenge, players aimed to achieve the highest number of steps and complete challenges to earn trophies for their team. Team and individual rankings were available (Figure 2D).

These games included a multitude of gamification mechanisms such as points, trophies, leaderboards, a chat, challenges, and narratives—mechanics that are closely linked to proven behavior change techniques [27]. Table 1 gives an overview of the gamification strategies included in the Kiplin games following the taxonomy proposed by Schmidt-Kraepelin et al [28] and the associated behavior change techniques. While the games share common characteristics (eg, collective gameplay and in-game challenges), it is important to note that The Adventure and The Challenge emphasize competition more than the others.

Figure 2. Screenshots of the Kiplin games. (A) The Adventure. (B) The Mission. (C) The Board Game. (D) The Challenge.

XSL•FO
RenderX

Table 1. Analysis of the game mechanics and content of the Kiplin app and associated behavior change techniques following the taxonomies by Schmidt-Kraepelin et al [28] and Michie et al [27].

Dimension	Characteristic	Gamification technique	Associated behavior change techniques
Gamification concept-to-user communication	Mediated	The gamification concept communicates with the user through an avatar named Pilot Kiplin (ie, a real Kiplin team member animating the app who takes the persona of a funny mascot). Pilot Kiplin launches in-game challenges, announces results, and delivers internal messages aimed at motivating participants. These messages include tips to plan and implement PA[a] in daily life and information on the benefits of walking on health.	• Instruction on how to perform a behavior (4.1) • Information about health consequences (5.1)
User identity	Static self-selected identity	Participants are only able to select a nickname and personalize the name of their team.	—[b]
Rewards	Internal	Rewards that participants can earn through the app are solely internal and virtual, such as points and trophies.	• Feedback on behavior (2.2) • Self-monitoring of behavior (2.3) • Cue signaling reward (7.2) • Behavioral practice and rehearsal (8.1) • Nonspecific reward (10.3)
Competition	Direct and indirect	Participants can compete directly via in-game challenges (ie, at specific intervals, a competitive challenge commences between teams; to secure victory and earn a trophy, one team must accumulate more points or steps than their rival team within the allotted time frame) or indirectly via the point system and leaderboards. These challenges are announced in advance, encouraging players to plan their activities so as to be active on the day of the battle.	• Action planning (1.4) • Feedback on behavior (2.2) • Social comparison (6.2)
Target group	Patients and healthy individuals	The app can be offered to patients in clinical settings as well as in preventive initiatives among employees or older adults.	—
Collaboration	Cooperative	All games operate on a collaborative basis: participants are organized into teams and must collectively complete challenges to advance or win. They can use the chat feature to communicate, support each other, and exchange ideas with their teammates.	• Social support (unspecified, 3.1) • Social support (emotional, 3.3)
Goal setting	Externally set	Goals, cutoffs, and the number of steps required to earn a trophy are predetermined by the app developer or the health care professional.	• Review behavior goals (1.5) • Discrepancy between current behavior and goal (1.6)
Narrative	Episodical	Narratives are unlocked as participants reach new milestones, such as unlocking a new clue in The Mission, reaching a higher board level in The Board Game, or reaching a new city in The Adventure.	—
Reinforcement	Positive	The app focuses solely on highlighting current and future successes, aiming to avoid stigmatizing users based on their initial inactivity or health status.	—
Persuasive intent	Behavior change	The app aims to enhance individuals' daily step counts through game mechanics.	—
Level of integration	Inherent	To participate in the games, individuals must engage in PA. Progression within the game is contingent upon the steps performed; without activity, players cannot advance.	—
User advancement	Presentation only	The progression is presented through the clues collected in The Mission to the progression to different cities in The Adventure and to new board levels in The Board Game. This structure encourages participants to engage in and repat the target behavior.	• Graded tasks (8.7) • Reward approximation (14.4)
Other tools of the app	—	Self-monitoring tool and notifications	• Self-monitoring (2.2) • Prompt and cues (7.1)

[a]PA: physical activity.

[b]No correspondence with the Behavior Change Technique Taxonomy.

XSL•FO
RenderX

Variables

The variables of interest were selected based on the behavior change intervention ontology by Michie et al [22] and included (1) the longitudinal evolution of daily steps, (2) the exposure of each participant to the intervention, (3) the intervention parameters, and (4) the context (participants' characteristics and setting) as these variables are likely to influence the intervention effect. Table 2 specifies the measures of interest and their operationalization.

Table 2. Operationalization of the variables.

Outcome	Operationalization
Primary outcome—target behavior (dependent variable)	
Daily step count	PA[a] was assessed via the daily step count, measured using the smartphone or activity monitor of the participant. The daily step count is a trusted proxy for PA [29]. During onboarding, participants were asked to connect to their tracking device (eg, Apple Health, Google Fit, Fitbit, or Garmin) for synchronization of their step count data. In this way, the daily step count of the participants was automatically synchronized on the Kiplin app, and the app could retrieve the daily step count for the previous 15 days.
Intervention (content and delivery) and mechanisms of action	
Type of game	Participants could play 4 types of games (ie, The Challenge, The Adventure, The Board Game, and The Mission).
Exposure	
Compliance ratio	The engagement of participants with the app was computed as the compliance ratio representing the number of days with a log-in during the game period divided by the duration of the game periods. This variable allows for measuring the frequency of the engagement with the service [30].
Number of games played	The total number of games played during the intervention period
Context (population and setting)	
Self-reported age and gender	Filled out by participants when they registered on the app
Population	Employees, older adults, or patients (treated for obesity or cancer)
Confounding factors	
Season	The season (winter, spring, summer, or autumn) when the step data were logged was controlled for as the season can influence PA [31].
Type of device	The type of device used to assess daily step count (ie, Android or iOS smartphones or Garmin, Withings, Polar, Fitbit, or TomTom wearables) was controlled for as smartphone apps and wearable devices differ in accuracy and precision [32].
Lockdown	The study period was characterized by the COVID-19 pandemic. In France, 3 lockdowns were implemented to mitigate the spread of COVID-19: in spring 2020 from March 17 to May 11, in fall 2020 from October 30 to December 15, and in spring 2021 from April 3 to May 3. During these periods, French citizens were required to remain at home with exceptions for essential activities such as going to work, shopping for necessities, health purposes, and engaging in individual PA near their residence. Failure to provide documentation justifying outdoor movement during inspections could result in fines. As these periods had a strong influence on the PA of individuals [33], we controlled for the lockdown periods in our analyses.

[a]PA: physical activity.

Statistical Analyses

We calculated the step count increase by subtracting the baseline average daily step count from the average daily step count during the intervention or follow-up periods for each participant and then computed the relative change (in percentage).

Mixed-effects models were used to (1) analyze within-person evolution across time (ie, changes in daily steps throughout the baseline, intervention, and follow-up periods) and across participants and nonparticipants and (2) examine the associations among intervention parameters, exposure to the intervention, participants' characteristics and settings, and daily step evolution. This statistical approach controls for the nested structure of the data (ie, multiple observations nested within participants); does not require an equal number of observations from all participants [34]; and separates between-person from within-person variance, providing unbiased estimates of the parameters [35,36].

First, an unconditional model (ie, with no predictor) was estimated for each variable to calculate intraclass correlation coefficients and estimate the amount of variance at the between- and within-individual levels, which allowed us to determine whether conducting multilevel models was relevant or not. Then, a model that allowed for random slope over time (ie, model with random intercept and random slope) was compared to the null model (ie, with only random intercept) using an ANOVA to evaluate whether the less parsimonious model explained a significantly higher proportion of the variance of the outcome than the unconditional model [37,38]. Third, between-level predictors and confounding variables were added to another model (model 1; the equation for the model was as follows: Y_{ij}

XSL•FO
RenderX

$= [\beta_0 + \gamma_{0i} + \theta_{0j}] + [\beta_1 + \theta_{1j}]\ time_j + \beta_2\ phase_j + \beta_3\ age_j + \beta_4\ sex_j + \beta_5\ population_j + \beta_6\ season_j + \beta_7\ captor_j + \beta_8$ baseline $PA_j + \beta_9\ lockdown_j + \beta_{10}\ condition_j \times phase_j + \varepsilon_{ij}$, where β_0 to β_{10} are the fixed-effects coefficients, θ_{0j} and θ_{1j} are the random effect for participant *j* (1 random intercept and 1 random slope), γ_{0i} is the random effect for time *i* [random intercept], and ε_{ij} is the error term) and compared to the previous models. Finally, intervention characteristics, as well as their interactions with the phases (ie, baseline, intervention, or follow-up) of the study, were added in a final model excluding nonparticipants (model 2; the equation for the model was as follows: $Y_{ij} = [\beta_0 + \gamma_{0i} + \theta_{0j}] + [\beta_1 + \theta_{1j}]\ time_j + \beta_2\ phase_j + \beta_3\ age_j \times phase_j + \beta_4\ sex_j + \beta_5\ population_j \times phase_j + \beta_6\ season_j + \beta_7\ captor_j \times phase_j + \beta_8$ baseline $PA_j \times phase_j + \beta_9\ lockdown_j + \beta_{10}$ compliance $ratio_j \times phase_j + \beta_{11}$ number of games $played_j \times phase_j + \beta_{12}$ type of $game_j + \varepsilon_{ij}$, where β_0 to β_{12} are the fixed-effects coefficients, θ_{0j} and θ_{1j} are the random effect for participant *j* (1 random intercept and 1 random slope), γ_{0i} is the random effect for time *i* [random intercept], and ε_{ij} is the error term). Model fit was assessed via the Bayesian information criterion and –2 log-likelihood [39]. All models were performed using the *lmerTest* package in the R software (R Foundation for Statistical Computing) [40]. An estimate of the effect size was reported using the marginal and conditional pseudo-R^2. When the interaction terms turned significant, contrast analyses were computed using the *emmeans* package [41]. The models' reliability (estimated using residual analyses) and outlier detection were performed using the *Performance* package [42]. In addition to subtracting nonwear days (defined previously), we removed outliers via the *check_outliers* function [42] that checks for influential observations via several distance and clustering methods (ie, *Z* scores, IQR, and equal-tailed interval). Sensitivity analyses were conducted using all data (including data before outlier imputation) and are available in Multimedia Appendix 1.

The data and code for the statistical analyses used in this study are available on the Open Science Framework [43].

Results

Descriptive Results

Descriptive results are presented in Table 3. The final sample included 4819 adults (mean age 42.7, SD 11.5 y; 2823/4819, 58.58% women). Participants wore an activity monitor measuring their daily step count for an average of 113 (SD 58.01; range 90-686) days. A total of 34,922 daily step observations were missing (ie, daily data missing or considered as a nonwear day), which is equivalent to 6.4% of missing data for the full data set.

We tested for statistical differences in sociodemographic variables and baseline daily steps between participants and nonparticipants using 2-tailed *t* tests and chi-square tests. Results revealed significant differences for age ($t_{82,500}$=–6.9149; P<.001), gender (χ^2_2=4028.3; P<.001), and baseline daily steps ($t_{22,721}$=–19.75; P<.001). However, in large samples, *P* values may drop below the α level despite effect sizes that are not practically meaningful [44]. Therefore, we mainly examined the magnitude of the effect sizes of these differences and observed very small to small effects (d=–0.03 for age, d=–0.17 for baseline daily steps, and w=0.09 for gender). According to Magnusson [45], the interpretation of these effect sizes suggests that, for age and baseline daily steps, approximately 98.8% and 93.2% of individuals in both groups overlapped, respectively. In addition, there is approximately a 50.8% and 54.8% chance that a randomly selected individual from the nonparticipant group would have a higher score than a randomly selected individual from the participant group. Therefore, we considered that the differences were minor between the 2 groups. Finally, these variables were controlled in our mixed-effects models as they were included as fixed effects.

XSL•FO
RenderX

Table 3. Descriptive statistics.

	Participants (n=3817)	Nonparticipants (n=995)
Sociodemographic characteristics		
Age (y), mean (SD)	43.2 (11.08)	41.0 (12.81)
Female sex, n (%)	2313 (62.6)	510 (53.26)
Employees, n (%)	3526 (92.38)	978 (98.29)
Patients, n (%)	194 (5.16)	17 (2.09)
Older adults, n (%)	97 (2.54)	__[a]
Exposure, mean (SD)		
Compliance ratio	0.84 (0.23)	0 (0)
Games played	1.28 (0.9)	0 (0)
In-game days	22.06 (16.24)	0 (0)
Observations in each type of game, n (%)		
The Adventure	21,316 (32.73)	—
The Board Game	4093 (6.28)	—
The Challenge	32,801 (50.37)	—
The Mission	6915 (10.62)	—
Type of device used, n (%)		
Android smartphone	1076 (28.19)	286 (28.74)
iOS smartphone	810 (21.22)	533 (53.57)
Fitbit	750 (19.65)	52 (5.23)
Garmin	1071 (28.06)	109 (10.95)
Polar	5 (0.08)	—
TomTom	3 (0.08)	—
Withings	90 (2.36)	9 (0.9)
Observations in each season, n (%)		
Winter	110,517 (23.87)	17,451 (24.4)
Spring	94,961 (20.51)	21,162 (29.6)
Summer	129,039 (27.87)	8804 (12.31)
Fall	138,429 (29.9)	24,086 (33.67)
Observations in each lockdown, n (%)		
First lockdown (spring 2020)	10,872 (2.35)	925 (1.29)
Second lockdown (fall 2020)	32,298 (6.89)	4110 (5.75)
Third lockdown (spring 2021)	23,435 (5.06)	1757 (2.46)

[a]Not applicable.

Hypothesis 1: Is the Gamified Program Effective to Promote PA?

During the intervention period, participants increased their daily steps by 2619 steps per day on average (+55.6%) compared to the baseline period and by 317 steps per day on average during the follow-up period (+13.8%) compared to the baseline. In comparison, the daily step count of the control group remained more or less stable throughout the same time frame, with a mean increase of 151 daily steps compared to baseline (+7.5%).

Overall, contrast analyses of the model for the intervention participants (model 2; Table S1 in Multimedia Appendix 1) revealed a negative effect of the intervention on the daily step count during the intervention phase compared to baseline activity (b=–0.09, 95% CI –0.14 to –0.05; P<.001) and no significant effect (b=0.01, 95% CI –0.05 to 0.06; P=.79) during the follow-up periods compared to baseline. However, the patterns were different when participants were stratified by baseline PA. Participants with lower baseline daily steps (<5000 steps per day or 5001-7500 steps per day) showed a significant increase in their daily steps during the intervention (b=0.25,

XSL•FO
RenderX

95% CI 0.22-0.28; *P*<.001) and follow-up (*b*=0.12, 95% CI 0.09-0.15; *P*<.001) periods both compared to the baseline. Participants with initial values between 7501 and 10,000 steps did not have a significant increase in their daily steps during the intervention (*b*=0.00, 95% CI –0.05 to 0.05; *P*=.99) or during the follow-up period (*b*=–0.01, 95% CI –0.04 to 0.02; *P*=.44) compared to baseline. Participants who performed >10,000 baseline steps had significant deteriorations during the intervention (*b*=–0.13, 95% CI –0.19 to –0.08; *P*<.001) and follow-up (*b*=–0.06, 95% CI –0.10 to –0.03; *P*<.001) periods. These trends are depicted in Figure 3 and Table 4. Results were similar in sensitivity analyses that used data without outlier imputation except for participants with initial daily step counts between 7501 and 10,000, who showed significant improvements during and after the intervention (Tables S2 and S3 in Multimedia Appendix 1).

In parallel, contrast analyses comparing the effectiveness of the Kiplin intervention on participants who used smartphones to collect their daily steps in comparison to participants who used a wearable showed a significantly greater effect among smartphone users during both the intervention phase (*b*=0.09, 95% CI 0.07-0.11; *P*<.001) and the follow-up period (*b*=0.04, 95% CI 0.01-0.06; *P*=.001). These results are illustrated in Figure S1 in Multimedia Appendix 1.

Figure 3. Changes in daily steps throughout the study phases for participants who received the Kiplin program stratified by baseline physical activity.

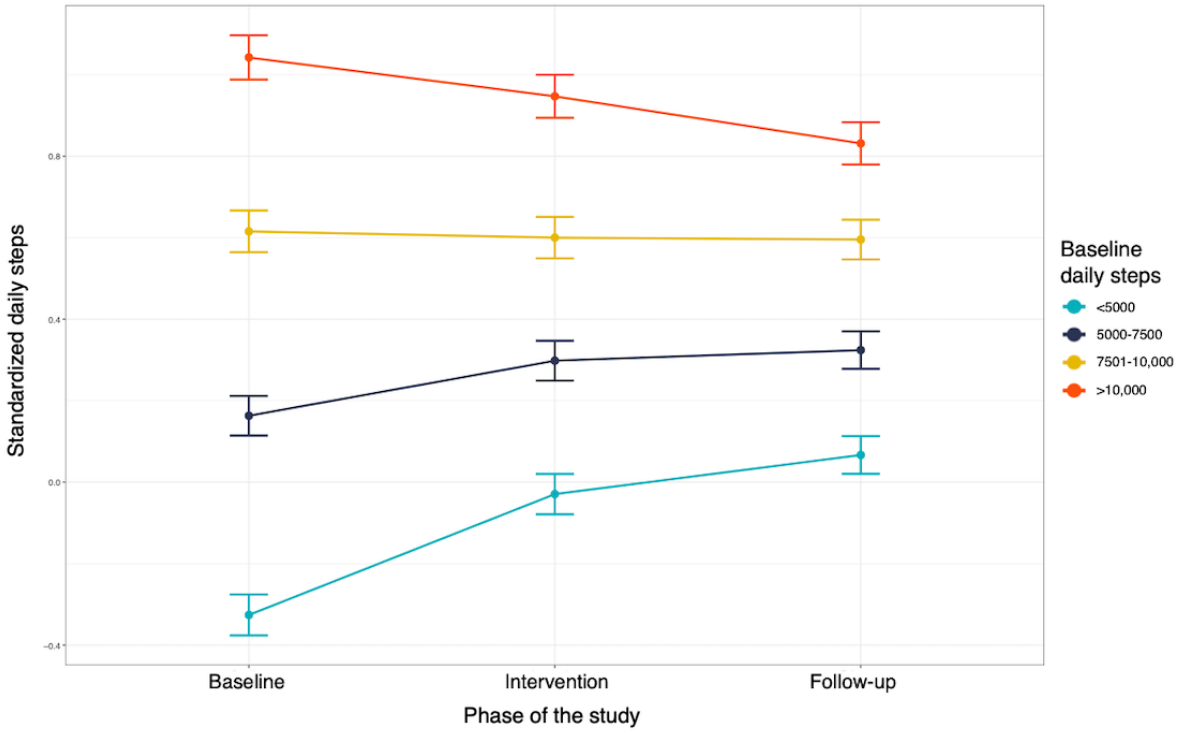

Table 4. Description of the mean daily step count during the baseline, intervention, and follow-up periods; changes; and relative changes from baseline according to participants' baseline daily step count.

	Participants with <5000 steps	Participants with 5000-7500 steps	Participants with 7501-10,000 steps	Participants with >10,000 steps
Baseline daily step count, mean (SD)	3671 (902.73)	6096 (747.49)	8818 (824.84)	10,111 (1789.2)
Intervention daily step count, mean (SD)	7490 (3804.69)	8855 (3786.71)	10,301 (3627.73)	11,388 (3518.07)
Follow-up daily step count, mean (SD)	5119 (2062.40)	6534 (1889.99)	7971 (2074.43)	9424 (2390.33)
Change from baseline during the intervention	+3820	+2762	+2187	+1309
Change from baseline during follow-up	+1459	+431	–156	–697
Relative change during intervention (%)	+118.8	+47.2	+28.8	+16.9
Relative change during follow-up (%)	+49.5	+8.2	–1	–4.3

XSL•FO
RenderX

Hypothesis 2: Is the Intervention Effect Greater for Participants Than for Nonparticipants?

In model 1 (Table S1 in Multimedia Appendix 1), participants who received the Kiplin intervention had a significantly greater increase in mean daily steps between baseline and the intervention period compared with nonparticipants (*b*=0.54, 95% CI 0.52-0.58; *P*<.001). The results were similar in sensitivity analyses (Table S3 in Multimedia Appendix 1). The comparison of the means, changes, and relative changes from baseline for participants and nonparticipants are available in Table S4 of Multimedia Appendix 1.

Hypothesis 3: What Are the Moderators of the Intervention Effect?

The model 2 estimates are shown in Table S1 in Multimedia Appendix 1. The variables under consideration explained 39% of the variance in daily steps. In this model, we tested the hypothesized interactions to investigate predictors associated with the efficiency of the intervention (Table S5 in Multimedia Appendix 1). Contrast analyses were conducted on significant interactions and revealed that the age (*b*=0.05; *P*<.001) and compliance ratio (*b*=0.37; *P*<.001) were positively associated with the change in daily steps between baseline and the intervention period. Specifically, the older the age, the more regularly the individuals played and the more effective the intervention was. On the other hand, the number of games played by participants was negatively associated with this change (*b*=–0.02; *P*=.02). In other words, the longer the intervention and the higher the number of games, the less effective the intervention. For categorical outcomes, contrast analyses revealed differences in the intervention effect among the different populations (Figure 4). Compared to employees, patients treated for cancer (*b*=–0.18; *P*<.001) and older adults (*b*=–0.19; *P*<.001) showed a significantly weaker effect of the intervention in comparison to baseline PA. There was no significant difference between employees and patients treated for obesity (*b*=–0.07; *P*=.13). All the results of these analyses are available in Multimedia Appendix 1.

Finally, model 2 estimates revealed that participants were significantly more active in The Adventure and The Challenge compared to The Board Game and The Mission (Table S1 in Multimedia Appendix 1).

Figure 4. Changes in daily steps throughout the study phases for the different populations who received the Kiplin program.

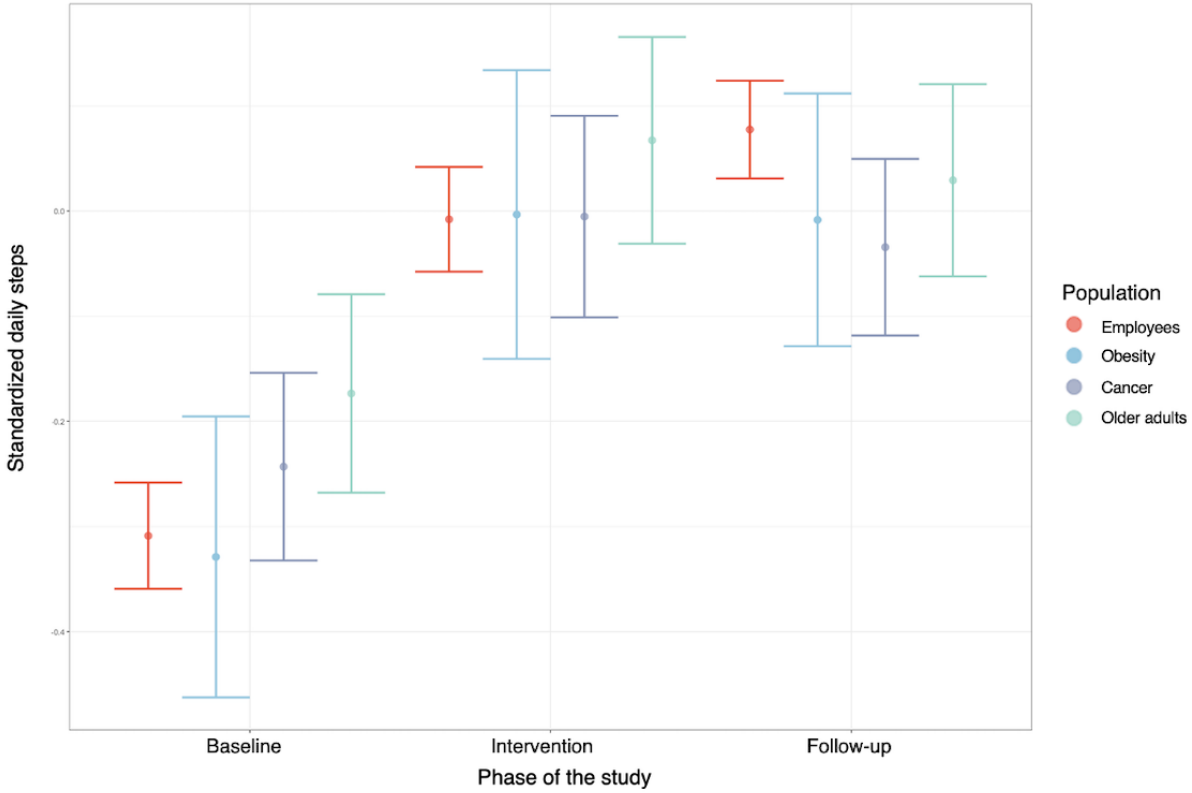

Discussion

Principal Findings

This study demonstrated a significant increase in daily steps among participants engaging with the Kiplin intervention compared to nonparticipants over the same period. Interestingly, the intervention effect varied according to the baseline daily step count of individuals. Participants with lower baseline steps (<7500 steps per day) significantly improved their PA during both the intervention (between +34% and +76%) and follow-up (between +10% and +33%) periods, whereas participants with >7500 steps had no significant change or significant decreases.

These results suggest that a gamified program is more efficient for inactive individuals compared to active ones, with the existence of a plateau effect. They also support recent findings [20,46] and the ability of gamified interventions to improve daily steps both during and after the end of the program and in real-life settings [47]—at least for the more inactive individuals.

XSL•FO
RenderX

This efficacy is noteworthy given the challenges faced by current behavioral interventions in promoting PA in the long haul [9].

SDT offers a valuable framework for elucidating the disparate outcomes observed among initially active and inactive participants. Gamification strategies could enhance the autonomous motivation of inactive participants, as suggested by a previous study [48], whereas the use of rewards on already motivated people could undermine this motivation. Known as the overjustification effect [18], this phenomenon suggests that, if people receive rewards for doing an activity that they used to enjoy, they are likely to discount the internal reason and, thus, become less intrinsically motivated than before receiving the rewards. This could explain why the same intervention had positive effects on inactive participants, who performed more daily steps after the end of the intervention (ie, during follow-up periods), compared to its effects on already active ones, who observed significant decreases after the intervention compared to their baseline daily steps.

Moreover, results indicating that the intervention was more effective among users who used their smartphones to track their step counts through the Kiplin app compared to those who already owned and used a wearable device—and were significantly more active at baseline—further reinforce this argument. Individuals who already possess an activity monitor are likely motivated to monitor their daily steps, potentially diminishing the additional impact of gamification rewards. Consequently, the introduction of gamification may have less influence or even produce counterproductive effects on their behavior, particularly when compared to those who solely rely on their smartphones for activity tracking in the context of the intervention.

The results of this study also stressed that older age may not be incompatible with gamified interventions. Indeed, intervention effectiveness was moderated by the age of the individual, and gamification was more efficient among older individuals compared to younger ones. These findings are in line with those of a previous study [49] that reported higher use of gamification features among older users. The authors postulated that older adults pay generally more attention to their health and, thus, have a stronger intention to engage in a health program. From another perspective, and in light of the gamification strategies embedded in the Kiplin intervention, these results could also be explained by the fact that these strategies are accessible—inspired by traditional board game rules and mechanics widely known in the general population—and, thus, may be more attractive for older populations. Previous research has suggested that the most engaging game mechanics may diverge between youths and other populations [50], and we can expect that younger populations may prefer more complex game mechanics and need more novelty during the intervention to stay interested in the service.

Regarding the effects of the gamified intervention according to the characteristics of the population, a stronger effect was found for programs among employees and patients treated for obesity. While these results warrant caution due to the variability observed in patients or older adult participants, these findings suggest that gamified interventions are suitable for both primary and tertiary prevention, as suggested by previous work [20].

Practical Implications

The findings of this study also offer valuable insights that could help improve future intervention design. First, exposure to the content is essential for the gamified intervention to be effective. It is interesting, as gamification has often been assimilated into a self-fulfilling process permitting automatic engagement of participants. These results are consistent with previous findings demonstrating that higher use of gamification features was associated with greater intervention effectiveness [49,51]. If gamification can ultimately increase program engagement, developers need first to design their apps to be as attractive as possible and optimize retention.

Second, the results revealed that the total number of games played was negatively associated with the intervention effect, suggesting that shorter interventions could be more beneficial for behavior change. These results are in line with those of previous research [20,52] suggesting that digital interventions of <3 months tend to yield greater benefits. It also suggests a “dose-response” relationship in an inverted U shape, with an optimal “middle” to find. Nevertheless, it is important to consider that Kiplin programs incorporating multiple games are built in such a way as to administer several doses at regular intervals. Therefore, periods without games were considered in the intervention phases and could explain why, overall, the shorter games were more efficient. More refined analyses of the intervention effect over time will be necessary in the future.

Third, the daily step count of participants was significantly higher in The Adventure and The Challenge. These 2 games are characterized by their competitive nature, placing a stronger emphasis on leaderboards than the other 2 games, which are more centered on collaboration. In this vein, Patel et al [53] observed that the competitive version of their gamified intervention outperformed the collaborative and supportive arms. Moreover, various studies have highlighted that leaderboards are a particularly successful gamification mechanic [49,54].

Strengths and Limitations

This study has several strengths, including its large sample size, the intensive objective PA measurement in real-life conditions through daily steps, and the longer baseline and follow-up duration compared with most trials on gamification that typically incorporate measurement bursts dispersed across time [20]. However, several limitations should be considered. First, this study was observational and not a randomized controlled trial. Thus, we cannot establish the causality of the intervention’s effect on outcome improvement. The nonparticipants are not a true control group. If they did not receive the intervention, it may be because they were unable to join or for underlying motivational reasons that could impact their PA. Second, intervention lengths differed between participants. Third, although mixed-effects models are useful for describing trends in PA behavior change over time, they are limited in their capacity to assess precise fluctuation patterns of nonstationary behavior, such as daily step counts [55] across time. Future

XSL•FO
RenderX

longitudinal studies could benefit from using time-series analyses to more accurately describe these patterns of change. Finally, the compliance ratio used in this study as a proxy for engagement tends to oversimplify the exposure of participants to the service. Complementary measures of behavioral engagement (eg, using the number of log-ins, time spent per log-in, and the number of components accessed) and affective engagement (eg, emotions and pleasure) should be considered to draw the longitudinal impact of the engagement of the participants on the intervention effect.

Conclusions

In this study, we conducted a comprehensive analysis of real-world data from >4800 individuals, suggesting the impact of a gamified intervention in real-life settings. Our findings indicate that the Kiplin intervention led to a significantly greater increase in mean daily steps from baseline among users than among nonparticipants. Interestingly, responses to the intervention were significantly different as a function of individuals' initial daily step counts. Participants with <7500 baseline daily steps had significant improvements during both the intervention and follow-up periods with +3291 daily steps during the program and +945 after the intervention on average, whereas the intervention had no effect on participants with initial values of >7500. Therefore, the motivational effect of gamification could depend on the initial PA and motivational profile of the participants. This result can also be interpreted in light of our observation that participants who already owned a wearable and, thus, were likely already motivated to engage in PA exhibited significantly lower effects compared to less experienced participants who used their smartphones to track their step counts. This study also revealed that the age of participants and their engagement with the app were positively and significantly associated with the intervention effect, whereas the number of games played was negatively associated with it.

Overall, the results of this study suggest that gamification holds promise in promoting the daily steps of inactive populations, with demonstrated short- and medium-term effects. Importantly, this study represents a pioneering effort as one of the first to examine the longitudinal effect of a gamified program outside the context of a trial using intensive real-world data. As such, the findings are quite generalizable to similar settings and reaffirm the value of gamification in both primary and tertiary prevention efforts across a diverse range of age groups.

Acknowledgments

The work of AM is supported by the French National Association of Research and Technology (Cifre PhD thesis grant) and by the company Kiplin. The funders had no input in the design of the study and no influence on the interpretation or publication of the study results.

Data Availability

The anonymized data used in this study and the R code are available on the Open Science Framework [43].

Authors' Contributions

AM conceptualized the study. AM and GH performed data curation. Investigation was led by AM, while AM and CF contributed to methodology and statistics. AM conducted the formal analysis using the R software. AM was responsible for writing the original draft. All authors contributed to writing (review and editing).

Conflicts of Interest

AC, CF, and MD declare that they have no competing interests. GH is employed by Kiplin. The results of this study could be beneficial to Kiplin from a marketing point of view. The Kiplin company had no input in the design of the study and no influence on the interpretation or publication of the study results.

Multimedia Appendix 1
Supplementary Figure S1 and Table S1-S5.
[PDF File (Adobe PDF File), 300 KB - jmir_v26i1e47116_app1.pdf]

References

1. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet 2012 Jul 21;380(9838):219-229 [FREE Full text] [doi: 10.1016/S0140-6736(12)61031-9] [Medline: 22818936]
2. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018 Oct;6(10):e1077-e1086 [FREE Full text] [doi: 10.1016/S2214-109X(18)30357-7] [Medline: 30193830]
3. Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U, et al. Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet 2012 Jul 21;380(9838):247-257. [doi: 10.1016/S0140-6736(12)60646-1] [Medline: 22818937]

XSL•FO
RenderX

4. Conger SA, Toth LP, Cretsinger C, Raustorp A, Mitáš J, Inoue S, et al. Time trends in physical activity using wearable devices: a systematic review and meta-analysis of studies from 1995 to 2017. Med Sci Sports Exerc 2022 Feb 01;54(2):288-298. [doi: 10.1249/MSS.0000000000002794] [Medline: 34559725]
5. Tison GH, Barrios J, Avram R, Kuhar P, Bostjancic B, Marcus GM, et al. Worldwide physical activity trends since COVID-19 onset. Lancet Glob Health 2022 Oct;10(10):e1381-e1382. [doi: 10.1016/s2214-109x(22)00361-8]
6. Hall KS, Hyde ET, Bassett DR, Carlson SA, Carnethon MR, Ekelund U, et al. Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. Int J Behav Nutr Phys Act 2020 Jun 20;17(1):78 [FREE Full text] [doi: 10.1186/s12966-020-00978-9] [Medline: 32563261]
7. Paluch AE, Bajpai S, Bassett DR, Carnethon MR, Ekelund U, Evenson KR, Steps for Health Collaborative. Daily steps and all-cause mortality: a meta-analysis of 15 international cohorts. Lancet Public Health 2022 Mar;7(3):e219-e228 [FREE Full text] [doi: 10.1016/S2468-2667(21)00302-9] [Medline: 35247352]
8. Global status report on physical activity. World Health Organization. 2022. URL: https://www.who.int/teams/health-promotion/physical-activity/global-status-report-on-physical-activity-2022 [accessed 2024-04-29]
9. Dunton GF, Rothman AJ, Leventhal AM, Intille SS. How intensive longitudinal data can stimulate advances in health behavior maintenance theories and interventions. Transl Behav Med 2021 Feb 11;11(1):281-286 [FREE Full text] [doi: 10.1093/tbm/ibz165] [Medline: 31731290]
10. Arigo D, Jake-Schoffman DE, Wolin K, Beckjord E, Hekler EB, Pagoto SL. The history and future of digital health in the field of behavioral medicine. J Behav Med 2019 Feb;42(1):67-83. [doi: 10.1007/s10865-018-9966-z] [Medline: 30825090]
11. Walsh JC, Groarke JM. Integrating behavioral science with mobile (mHealth) technology to optimize health behavior change interventions. Eur Psychol 2019 Jan;24(1):38-48. [doi: 10.1027/1016-9040/a000351]
12. Zhao J, Freeman B, Li M. Can mobile phone apps influence people's health behavior change? An evidence review. J Med Internet Res 2016 Oct 31;18(11):e287 [FREE Full text] [doi: 10.2196/jmir.5692] [Medline: 27806926]
13. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining "gamification". In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments. 2011 Presented at: MindTrek '11; September 28-30, 2011; Tampere, Finland p. 9-15 URL: https://dl.acm.org/doi/10.1145/2181037.2181040 [doi: 10.1145/2181037.2181040]
14. Deci EL, Ryan RM. Handbook of Self-Determination Research. Rochester, NY: University of Rochester Press; 2002.
15. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/S15327965pli1104_01]
16. Teixeira PJ, Carraça EV, Markland D, Silva MN, Ryan RM. Exercise, physical activity, and self-determination theory: a systematic review. Int J Behav Nutr Phys Act 2012;9:78 [FREE Full text] [doi: 10.1186/1479-5868-9-78] [Medline: 22726453]
17. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
18. Ryan RM, Deci EL. Cognitive evaluation theory, part I. The effects of rewards, feedback, and other external events on intrinsic motivation. In: Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY: Guilford Press; 2017:123-127.
19. Zuckerman O, Gal-Oz A. Deconstructing gamification: evaluating the effectiveness of continuous measurement, virtual rewards, and social comparison for promoting physical activity. Pers Ubiquit Comput 2014 Jul 5;18(7):1705-1719. [doi: 10.1007/S00779-014-0783-2]
20. Mazeas A, Duclos M, Pereira B, Chalabaev A. Evaluating the effectiveness of gamification on physical activity: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Jan 04;24(1):e26779 [FREE Full text] [doi: 10.2196/26779] [Medline: 34982715]
21. Vincent JL. We should abandon randomized controlled trials in the intensive care unit. Crit Care Med 2010 Oct;38(10 Suppl):S534-S538. [doi: 10.1097/CCM.0b013e3181f208ac] [Medline: 21164394]
22. Michie S, Thomas J, Johnston M, Aonghusa PM, Shawe-Taylor J, Kelly MP, et al. The human behaviour-change project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. Implement Sci 2017 Oct 18;12(1):121 [FREE Full text] [doi: 10.1186/s13012-017-0641-5] [Medline: 29047393]
23. Hekler EB, Michie S, Pavel M, Rivera DE, Collins LM, Jimison HB, et al. Advancing models and theories for digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):825-832. [doi: 10.1016/j.amepre.2016.06.013] [Medline: 27745682]
24. Bassett DR, Wyatt HR, Thompson H, Peters JC, Hill JO. Pedometer-measured physical activity and health behaviors in U.S. adults. Med Sci Sports Exerc 2010 Oct;42(10):1819-1825 [FREE Full text] [doi: 10.1249/MSS.0b013e3181dc2e54] [Medline: 20305579]
25. Kang M, Rowe DA, Barreira TV, Robinson TS, Mahar MT. Individual information-centered approach for handling physical activity missing data. Res Q Exerc Sport 2009 Jun;80(2):131-137. [doi: 10.1080/02701367.2009.10599546] [Medline: 19650377]

XSL•FO
RenderX

26. Wang G, Yang M, Hong M, Krauss J, Bailey JF. Clinical outcomes one year after a digital musculoskeletal (MSK) program: an observational, longitudinal study with nonparticipant comparison group. BMC Musculoskelet Disord 2022 Mar 11;23(1):237 [FREE Full text] [doi: 10.1186/s12891-022-05188-x] [Medline: 35277161]
27. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
28. Schmidt-Kraepelin M, Thiebes S, Tran MC, Sunyaev A. What's in the game? Developing a taxonomy of gamification concepts for health apps. In: Proceedings of the 51st Hawaii International Conference on System Sciences. 2018 Presented at: HICSS '18; January 3-6, 2018; Waikoloa Village, HI p. 1217-1226 URL: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/93fa9001-cd42-4379-b4d7-607c2d223d40/content [doi: 10.24251/hicss.2018.150]
29. Saint-Maurice PF, Troiano RP, Bassett DR, Graubard BI, Carlson SA, Shiroma EJ, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160 [FREE Full text] [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
30. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
31. Turrisi TB, Bittel KM, West AB, Hojjatinia S, Hojjatinia S, Mama SK, et al. Seasons, weather, and device-measured movement behaviors: a scoping review from 2006 to 2020. Int J Behav Nutr Phys Act 2021 Feb 04;18(1):24 [FREE Full text] [doi: 10.1186/s12966-021-01091-1] [Medline: 33541375]
32. Piccinini F, Martinelli G, Carbonaro A. Accuracy of mobile applications versus wearable devices in long-term step measurements. Sensors (Basel) 2020 Nov 05;20(21):6293 [FREE Full text] [doi: 10.3390/s20216293] [Medline: 33167361]
33. Genin PM, Lambert C, Larras B, Pereira B, Toussaint J, Baker JS, et al. How did the COVID-19 confinement period affect our physical activity level and sedentary behaviors? Methodology and first results from the French national ONAPS survey. J Phys Act Health 2021 Mar 01;18(3):296-303. [doi: 10.1123/jpah.2020-0449] [Medline: 33581686]
34. Sonnentag S. Dynamics of well-being. Annu Rev Organ Psychol Organ Behav 2015 Apr 10;2(1):261-293. [doi: 10.1146/annurev-orgpsych-032414-111347]
35. Boisgontier MP, Cheval B. The anova to mixed model transition. Neurosci Biobehav Rev 2016 Sep;68:1004-1005. [doi: 10.1016/j.neubiorev.2016.05.034] [Medline: 27241200]
36. Judd CM, Westfall J, Kenny DA. Experiments with more than one random factor: designs, analytic models, and statistical power. Annu Rev Psychol 2017 Jan 03;68:601-625. [doi: 10.1146/annurev-psych-122414-033702] [Medline: 27687116]
37. Bliese P. Multilevel modeling in R (2.7): a brief introduction to R, the multilevel package and the NLME package. Cran R. URL: https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf [accessed 2024-04-29]
38. Finch WH, Bolin JE. Multilevel Modeling Using Mplus. Boca Raton, FL: Chapman & Hall; 2017.
39. Bollen KA, Harden JJ, Ray S, Zavisca J. BIC and alternative Bayesian information criteria in the selection of structural equation models. Struct Equ Modeling 2014 Jan 31;21(1):1-19 [FREE Full text] [doi: 10.1080/10705511.2014.856691] [Medline: 31360054]
40. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26. [doi: 10.18637/JSS.V082.I13]
41. Lenth RV, Bolker B, Buerkner P, Giné-Vázquez I, Herve M, Jung M, et al. Estimated marginal means, aka least-squares means. Cran R. 2018. URL: https://cran.r-project.org/web/packages/emmeans/emmeans.pdf [accessed 2024-04-29]
42. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. performance: an R package for assessment, comparison and testing of statistical models. J Open Source Softw 2021 Apr;6(60):3139. [doi: 10.21105/joss.03139]
43. Mazéas A, Forestier C, Harel G, Duclos M, Chalabaev A. The impact of a gamified intervention on physical activity in real-life conditions: a retrospective analysis of 4800 individuals. medRxiv Preprint posted online November 20,2022 [FREE Full text] [doi: 10.1101/2022.11.18.22282458]
44. Lin M, Lucas HC, Shmueli G. Research commentary—too big to fail: large samples and the p-value problem. Inf Syst Res 2013 Dec;24(4):906-917. [doi: 10.1287/isre.2013.0480]
45. Magnusson K. A causal inference perspective on therapist effects. PsyArXiv Preprints posted online June 9, 2023 [FREE Full text] [doi: 10.31234/osf.io/f7mvz]
46. Patel MS, Small DS, Harrison JD, Hilbert V, Fortunato MP, Oon AL, et al. Effect of behaviorally designed gamification with social incentives on lifestyle modification among adults with uncontrolled diabetes: a randomized clinical trial. JAMA Netw Open 2021 May 03;4(5):e2110255 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10255] [Medline: 34028550]
47. de Buisonjé DR, Brosig F, Breeman LD, Bloom EL, Reijnders T, Janssen VR, et al. Put your money where your feet are: the real-world effects of StepBet gamified deposit contracts for physical activity. Internet Interv 2023 Mar;31:100610 [FREE Full text] [doi: 10.1016/j.invent.2023.100610] [Medline: 36873308]
48. Höchsmann C, Infanger D, Klenk C, Königstein K, Walz SP, Schmidt-Trucksäss A. Effectiveness of a behavior change technique-based smartphone game to improve intrinsic motivation and physical activity adherence in patients with type 2 diabetes: randomized controlled trial. JMIR Serious Games 2019 Feb 13;7(1):e11444 [FREE Full text] [doi: 10.2196/11444] [Medline: 30758293]

XSL•FO
RenderX

49. Maher CA, Olds T, Vandelanotte C, Plotnikoff R, Edney SM, Ryan JC, et al. Gamification in a physical activity app: what gamification features are being used, by whom, and does it make a difference? Games Health J 2022 Jun;11(3):193-199. [doi: 10.1089/g4h.2021.0207] [Medline: 35501981]
50. Schwarz AF, Huertas-Delgado FJ, Cardon G, DeSmet A. Design features associated with user engagement in digital games for healthy lifestyle promotion in youth: a systematic review of qualitative and quantitative studies. Games Health J 2020 Jun;9(3):150-163. [doi: 10.1089/g4h.2019.0058] [Medline: 31923363]
51. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PLoS One 2017;12(3):e0173403 [FREE Full text] [doi: 10.1371/journal.pone.0173403] [Medline: 28362821]
52. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019 Mar 19;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
53. Patel MS, Small DS, Harrison JD, Fortunato MP, Oon AL, Rareshide CA, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the united states: the step up randomized clinical trial. JAMA Intern Med 2019 Sep 09:1-9 [FREE Full text] [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]
54. De Croon R, Geuens J, Verbert K, Vanden Abeele V. A systematic review of the effect of gamification on adherence across disciplines. In: Proceedings of the 3rd International Conference, HCI-Games 2021, Held as Part of the 23rd HCI International Conference. 2021 Presented at: HCII '21; July 24-29, 2021; Virtual Event p. 168-184 URL: https://dl.acm.org/doi/abs/10.1007/978-3-030-77277-2_14 [doi: 10.1007/978-3-030-77277-2_14]
55. Heino MT, Knittle K, Noone C, Hasselman F, Hankonen N. Studying behaviour change mechanisms under complexity. Behav Sci (Basel) 2021 May 14;11(5):77 [FREE Full text] [doi: 10.3390/bs11050077] [Medline: 34068961]

Abbreviations

API: application programming interface
PA: physical activity
SDT: self-determination theory

Edited by G Greco; submitted 25.04.23; peer-reviewed by S Payne, A Bucher; comments to author 26.01.24; revised version received 10.03.24; accepted 01.05.24; published 12.08.24.

Please cite as:
Mazéas A, Forestier C, Harel G, Duclos M, Chalabaev A
The Impact of a Gamified Intervention on Daily Steps in Real-Life Conditions: Retrospective Analysis of 4800 Individuals
J Med Internet Res 2024;26:e47116
URL: https://www.jmir.org/2024/1/e47116
doi:10.2196/47116
PMID:39133533

©Alexandre Mazéas, Cyril Forestier, Guillaume Harel, Martine Duclos, Aïna Chalabaev. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

COVID-19–Related Social Isolation, Self-Control, and Internet Gaming Disorder Among Chinese University Students: Cross-Sectional Survey

Yufang Guo[1], PhD; Fangyan Yue[1], BS; Xiangyu Lu[2], MS; Fengye Sun[1], BS; Meixing Pan[1], MS; Yannan Jia[3], MS

[1]School of Nursing and Rehabilitation, Shandong University, Jinan, China

[2]Hebei Provincial Mental Health Center, Baoding, China

[3]Committee of The Communist Youth League, Shandong University, Jinan, China

Corresponding Author:
Yufang Guo, PhD
School of Nursing and Rehabilitation
Shandong University
44 Wen Hua Xi Road
Jinan, 250012
China
Phone: 86 15269163352
Email: cdguoyufang@163.com

Abstract

Background: Internet gaming disorder among university students has become a great concern for university counsellors worldwide since the COVID-19 pandemic. The factors influencing the development of internet gaming disorder in students during the COVID-19 pandemic could be different from those before the pandemic.

Objective: This study aims to explore the associations among social isolation, self-control, and internet gaming disorder in Chinese university students and to examine whether self-control mediates the positive effects of social isolation on internet gaming disorder.

Methods: A cross-sectional survey was employed to collect data from university students in Shandong province of China from April to September 2022. The Isolation subscale of the Self-Compassion Scale, Self-Control Scale, and Internet Gaming Disorder Scale were used to assess the social isolation, self-control, and internet gaming disorder among university students, respectively. Models 4 and 5 of PROCESS software were used to analyze the mediating role of self-control and the moderating role of gender on the association between social isolation and internet gaming disorder.

Results: A total of 479 students were recruited from 6 universities located in 3 different regions of Shandong, China. Students had low levels of internet gaming disorder and moderate levels of social isolation and self-control, with mean scores of 8.94 (SD 9.06), 12.04 (SD 3.53), and 57.15 (SD 8.44), respectively. Social isolation was positively correlated with internet gaming disorder (r=0.217; P<.001), and self-control was negatively correlated with social isolation (r=–0.355; P<.001) and internet gaming disorder (r=–0.260; P<.001). Self-control played a mediating role in the association between social isolation and internet gaming disorder (β=–.185, 95% CI –.295 to –.087). The effects of social isolation on internet gaming disorder among female students were lower than those among male students.

Conclusions: Self-control was a mediator in the association between social isolation and internet gaming disorder. Moreover, gender played a moderating role in the association between social isolation and internet gaming disorder. This study highlights the need to alleviate the development of internet gaming disorder among students during a pandemic, especially that of male students. Effective interventions that lessen social isolation and promote self-control should be developed.

(J Med Internet Res 2024;26:e52978) doi:10.2196/52978

KEYWORDS

COVID-19 pandemic; internet gaming disorder; self-control; social isolation; university students; game; gaming; games; addict; addictive; addiction; addictions; university; universities; college; colleges; postsecondary; higher education; student; students;

XSL•FO
RenderX

China; Chinese; isolation; isolated; self-compassion; mental health; association; associations; correlation; causal; correlated; correlations

Introduction

The COVID-19 pandemic was declared a global health emergency by the World Health Organization in March 2020, which significantly impacted educational training institutions worldwide [1]. Previous studies have shown that because of educational interruptions, remote learning, less physical activity, financial concerns, social isolation, lack of belongingness, and the rapid rise in the number of COVID-19 infections, university students were one of the most vulnerable groups to experience anxiety, depression, frustration, loneliness, substance misuse, and addiction [2-4]. To alleviate the negative impacts of COVID-19 on mental and physical health, university students choose internet gaming as the main source of entertainment [5]. According to the research report on the internet usage of minors in China in 2023 released by China Internet Network Information Center, 522 million Chinese citizens often played internet games, and university students are one of the main groups who often play internet games [6,7].

Internet gaming disorder (IGD), classified as an addictive disorder by the *Diagnostic and Statistical Manual of Mental Disorders, fifth edition* (*DSM-5*) and the *eleventh revision of the International Classification of Diseases* (*ICD-11*), is defined as a pattern of continuous or repeated gaming behavior, manifested through 5 of the 9 diagnosis criteria in the *DSM-5* (withdrawal, loss of control, loss of interest, preoccupation or obsession, tolerance, deceiving, continued overuse, functional impairment, and escape of negative feelings) or diagnosis in the *ICD-11* (impaired control of the game, increasing priority for gaming, continuous or upgraded gaming despite negative outcomes) [8,9]. Son et al [10] found that 36.4% of the university students played internet games 3-4 times per week, 19.5% played internet games daily, and 48.1% played an internet game for over 2 hours. Vahidi et al [11] found that 48.5% of the university students played internet games, and among them, 4.3% were diagnosed with IGD. Siste et al [12] found that the prevalence rate of IGD was 2.03% among university students, and students with IGD played games 19 hours per week. Previous evidence shows that IGD leads to maladjustments in students' daily life, such as headaches, attention-deficit/hyperactivity disorder, alexithymia, increased aggression, poor psychological well-being, and a decline in study achievement [13-15]. IGD also inclines students to flee their homes and engage in substance abuse, self-harm, and even suicidality when it becomes severe [16,17]. Therefore, exploring the influencing factors for IGD in university students during the COVID-19 pandemic is crucial for university counsellors and students themselves to take effective interventions to treat and relieve symptoms of IGD and prevent IGD in future pandemics.

Several qualitative and quantitative studies have indicated that social isolation plays a key role in the emergence of IGD, especially as seen during the COVID-19 pandemic [18,19]. Social isolation is an objective measure reflecting an individual's absence or limitation of social contact or interactions with friends, family, neighbors, and colleagues [20]. Social isolation causes profound changes in social relationships, which are linked with loneliness, anxiety, depression, and exhaustion. It is possible that individuals who feel social isolation are more likely to surf online to address their loneliness [21,22]. Moreno et al [23] reported that because of poor quality and infrequent social interactions, the prevalence of IGD among young adults was 4.43% during the COVID-19 pandemic. Gaming addiction is referred to as a consequence of dealing with COVID-19–related psychological distress and social isolation [24]. According to previous studies, social isolation has damaging effects on the IGD of university students influenced by personality characteristics [25,26]. However, the critical factors and impact pathway between social isolation and IGD remain unclear and need further investigation.

Self-control refers to individuals' ability to regulate their thoughts, feelings, and actions, when permanently valued purposes collide with momentarily more satisfying goals [27]. As a facet of personality traits, self-control can involve engaging in desired behaviors and deserting undesired ones despite challenges to doing so. Studies suggest that good self-control benefits students' task performance, positive emotions, psychological well-being, and health-promoting behaviors [28,29]. Studies have documented that self-control has a negative correlation with IGD, and poor self-control may have increased the likelihood of students developing IGD during the COVID-19 pandemic [30]. Because of unsupportive living environments, controlling relationships, and social withdrawal during the pandemic, students' self-control was undermined, which resulted in increased IGD.

Questions regarding the correlation between social isolation and self-control and the impact of self-control on the relationship between social isolation and IGD remain largely unanswered. The limited research on this topic has reported that social isolation during the COVID-19 pandemic significantly increased the probability of internet gaming, while students with high self-control are less likely to fall into internet gaming dependence [31]. Moreover, some studies have found that self-control as a coping method is greatly associated with a variety of social isolation–related negative experiences such as anxiety, frustration, and loneliness [32,33]. According to the self-determination theory, social isolation can reduce students' self-motivation and decision-making abilities, while maladaptive motivations (eg, introjected regulation, amotivation, strong extrinsic motivations) are positively associated with low self-control [34,35]. Low self-control was verified as an important predictor for serious IGD [36]. Therefore, social isolation may be negatively correlated with self-control, and self-control may play a mediating role in the relationship between social isolation and IGD. Students who experienced severe social isolation had poor self-control and were consequently addicted to internet gaming.

Based on these findings, this study aims to explore the associations among social isolation, self-control, and IGD and to examine the mediating effect of self-control on the

XSL•FO
RenderX

relationship between social isolation and IGD in university students during the COVID-19 pandemic and related lockdown procedures in mainland China.

Methods

Study Participants

A cross-sectional study design was employed. University students who were Chinese and familiar with communication software (eg, WeChat, QQ) were invited to participate in this study. The surveys were conducted between April and September 2022. Convenience sampling was used to recruit university student participants from Jinan, Zibo, and Binzhou in Shandong, China. Jinan, the provincial capital of Shandong, has 23 universities. Zibo and Binzhou have 2 and 1 universities, respectively. The number of university students were 18,000-70,000. A total of 504 students completed the web-based survey. After excluding 25 questionnaires with incomplete data, 479 questionnaires were valid, with a response rate of 95.04% (479/504).

Ethics Approval

The study survey was approved by the institutional review board of the School of Nursing and Rehabilitation, Shandong University (2021-R-043). All procedures of this study followed the guidelines and requirements of the institutional review board. An anonymous web-based survey was administered, and students who submitted the questionnaires were considered as consenting to participate in the study. Students received ¥5 (US $0.69) as compensation for their time and effort. The data were collected and stored in a computer owned by the research team, and only the researchers of this study knew the password for the data file. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (Shandong University, China) and with the Helsinki Declaration of 1975, as revised in 2000.

Measurements

Self-Compassion Scale

The Isolation subscale of the Self-Compassion Scale [37] is a 4-item measure of social isolation. Each item contains a negative adjective description (ie, alone, inadequate, down, and struggling) to express feelings of isolation. A 5-point Likert scale from 1 (never) to 5 (always) was used to describe participants' experience of social isolation. The Chinese version of the Self-Compassion Scale has been shown to have good validity and reliability among university students [38]. In this study, the α reliability rate for the Isolation subscale of the Self-Compassion Scale was 0.780.

Self-Control Scale

The Self-Control Scale, developed by Tangney et al [39], is a 36-item scale and consists of 5 dimensions: self-discipline (11 items), impulse control (10 items), healthy habits (7 items), work or study performance (4 items), and dependability (4 items). A 5-point Likert scale (1=completely out of line, 5=fully compliant) is used to score the items. The Chinese version, translated and revised by Tan and Guo [40], is a 19-item scale with 5 dimensions: impulse control (6 items), healthy habits (3 items), resisting temptation (4 items), work ethics (3 items), and abstaining from entertainment (3 items). The Chinese version also had good validity and reliability among university students, with the confirmatory factor analysis being statistically significant (χ^2=1.5, root mean square error of approximation=0.050, goodness of fit index=0.91, incremental fit index=0.93, non–normed fit index=0.91, comparative fit index=0.93), and the α reliability rates for the scale and its dimension were between 0.606 and 0.862. In this study, the α reliability rates for the Self-Control Scale and its dimensions were 0.868, 0.811, 0.640, 0.648, 0.575, and 0.512, respectively.

IGD Scale

The IGD Scale [41] is a 9-item short scale that is scored via a 5-point Likert scale with 0 representing never and 5 representing always. The IGD Scale has been translated into Chinese, Arabic, and Turkish with good validity and reliability. The Chinese version of the IGD Scale has been widely used among young adults [42,43]. In this study, the α reliability rate for the scale was 0.888.

Procedure

This study was conducted from April to September 2022. During this 6-month period, there was 1 outbreak of the COVID-19 pandemic in Shandong, which led to lockdown in the universities. Therefore, since students were required to maintain social distance, refrain from going outdoors, and study at home or in their dormitory, several web-based and offline methods (eg, recruitment advertisements, messages, calls) were used to recruit student participants for this study.

The Wen-Juan-Xing platform [44] was used for data collection. We created a web-based survey through Wen-Juan-Xing for business to ensure its security and functionality. The survey consisted of 2 components: an explanation and the questionnaire with 5 pages. The web-based survey link and its related QR code were sent to university counselors from 6 universities, who distributed the link and QR code to their students. Student participants were required to complete the sociodemographic questionnaire, Isolation subscale of the Self-Compassion Scale, Self-Control Scale, and IGD Scale, in that order. Each IP address can submit only 1 questionnaire. Only when the technical completeness check for the questionnaires was passed, the web-based questionnaire could be successfully submitted.

Statistical Analysis

Harman single-factor analysis [45] was used to explore the common-method variance in this study. Descriptive analysis (mean, SD, frequency, percentage, and 95% CIs) was conducted to report the measured variables. Bivariate analysis was used to examine the differences in IGD by sociodemographic characteristics. Pearson correlation coefficients were used to test the correlations among social isolation, self-control, and IGD. Model 4 for PROCESS (Andrew F Hayes) was employed to test the mediating role of self-control on the association between social isolation and IGD. Model 5 for PROCESS was employed to explore the moderating role of gender on the association between social isolation and IGD. Bootstrap samples were set as 5000, and a 95% CI was used for the mediation analysis.

XSL•FO
RenderX

Results

Common-Method Variance Analysis

To avoid self-report measure–related common-method variance, we used anonymity and separating dimensions of the questionnaires in the data collection process [45]. Furthermore, Harman single-factor analysis showed that only 25.1% of the variance can be explained by the first factor, which was less than 40%. Thus, no common-method variance was found in this study.

Descriptive Sociodemographic Characteristics and Associations Between Sociodemographic Characteristics and IGD of the Student Participants

Table 1 presents the descriptive sociodemographic characteristics of the student participants. The average age of the students was 19.97 (SD 1.81) years; 72.9% (349/479) were females, 63.5% (304/479) were not an only child, 49.3% (236/479) were freshmen, and 48.2% (231/479) were majoring in medicine. The 2-sided *t* test and analysis of variance showed that gender and major were significant influencing factors for students' IGD (each *P*<.05).

Table 1. Descriptive sociodemographic characteristics of the student participants and the distribution of internet gaming disorders (N=479).

Characteristics	Values, n (%)	IGD[a], mean (SD)	*t* test/*F* test *(df)*	*P* value
Gender			7.824 (3)	<.001
Male	130 (27.1)	14.15 (9.13)		
Female	349 (72.9)	6.99 (8.24)		
Only child			1.847 (3)	.07
Yes	175 (36.5)	9.95 (9.33)		
No	304 (63.5)	8.35 (8.86)		
Grade			1.334 (3, 475)	.26
Freshman year	236 (49.3)	9.31 (8.93)		
Sophomore year	78 (16.3)	8.67 (9.79)		
Junior year	80 (16.7)	9.83 (8.35)		
Senior year	85 (17.7)	7.31 (9.30)		
Major			5.863 (3, 475)	.001
Liberal art	98 (20.5)	8.32 (9.29)		
Science and Engineering	118 (24.6)	11.91 (9.50)		
Medicine	231 (48.2)	7.81 (8.54)		
Others	32 (6.7)	8.03 (8.31)		

[a]IGD: internet gaming disorder.

Descriptive Social Isolation, Self-Control, and IGD and the Correlations Among the Measured Variables of Student Participants

Table 2 shows the descriptive social isolation, self-control, and IGD and the correlations among these variables of students. The total score for social isolation was 12.04 (SD 3.53), which indicated that students experienced moderate levels of social isolation during the COVID-19 pandemic. The total score for self-control was 57.15 (SD 8.44), which indicated that students had moderate levels of self-control. The total score for IGD was 8.94 (SD 9.06), which indicated that students had low levels of IGD. The Pearson correlation analysis showed that social isolation was positively correlated with IGD (r=0.217; P<.001), while social isolation and IGD were negatively correlated with self-control ($r_{\text{social isolation}}$=–0.355, r_{IGD}=–0.260; each P<.001).

XSL•FO
RenderX

Table 2. Descriptive statistics of the measured variables and the correlations among the variables (N=479).

Variables	Mean (SD)	Social isolation	Self-control	Internet gaming disorder
Social isolation	12.04 (3.53)	__[a]	—	—
r	—	—	—	—
P value	—	—	—	—
Self-control	57.15 (8.44)	—	—	—
r	—	–0.355	—	—
P value	—	<.001	—	—
Internet gaming disorder	8.94 (9.06)	—	—	—
r	—	0.217	–0.260	—
P value	—	<.001	<.001	—

[a]Not applicable.

Mediation Analysis

Model 4 of PROCESS was employed to analyze the mediating role of self-control in the association between social isolation and IGD. Gender and major were included in the model as covariate variables. Table 3 shows the results of the mediation analysis.

Table 4 presents the effects of social isolation on IGD, which is mediated by self-control. Social isolation had a positive direct effect on IGD (β=.526; *t*=3.033; *P*<.001). When self-control as a mediator was included in the model, the positive direct effect of social isolation on IGD was also significant (β=.341; *t*=3.033; *P*=.003); moreover, social isolation had a negative direct effect on self-control (β=–.844; *t*=–8.233; *P*<.001), and self-control had a negative direct effect on IGD (β=–.219; *t*=–4.659; *P*<.001). The results showed that self-control was a significant mediator in the relationship between social isolation and IGD.

Table 3. Mediation model of self-control on the association between social isolation and internet gaming disorder (N=479).

Outcome variables, predictive variables	β	SE	*t* test	*P* value	*F* test *(df)*	*P* value	R^2
Internet gaming disorder					32.137 (4, 474)	<.001	0.163
Social isolation	.526	.205	4.896	<.001			
Gender	–6.995	–.344	–8.209	<.001			
Major	–.524	–.051	–1.225	.22			
Self-control					23.231 (4, 474)	<.001	0.128
Social isolation	–.844	.103	–8.233	<.001			
Gender	–.087	.813	–0.106	.92			
Major	–.447	.408	–1.095	.27			
Internet gaming disorder					30.582 (4, 474)	<.001	0.205
Social isolation	.341	.112	3.033	.003			
Self-control	–.219	.047	–4.659	<.001			
Gender	–6.976	.834	–8.364	<.001			
Major	–.426	.419	–1.016	.31			

XSL•FO
RenderX

Table 4. Effects of social isolation on internet gaming disorder mediated by self-control (N=479).

Effect	β (95% Boot CI)	Boot SE	Effect-total effect ratio (%)[a]
Direct effect of X[b] on Y[c]	.341 (.120 to .562)	.112	64.83
Indirect effect: X→M1[d]→Y	–.185 (–0.295 to 0.087)	–.052	35.17

[a]Social isolation was not only directly associated with internet gaming disorder but also indirectly associated with internet gaming disorder through the path of self-control. The direct effect was 2 times stronger than the indirect effect, with effect-total effect ratios of 64.83% and 35.17%, respectively.

[b]X: social isolation.

[c]Y: internet gaming disorder.

[d]M: self-control.

Moderation Analysis

Model 5 of PROCESS was conducted to examine whether gender mitigates the association between social isolation and IGD. Table 5 presents the moderation analysis results. The interaction variable (social isolation × gender) was negatively correlated with IGD (β=–.496; *t*=–2.003; *P*=.046), indicating that gender was a moderator in the association between social isolation and IGD.

For further analysis of the moderation of gender, a simple slope analysis was conducted. The moderation analysis of gender in the association between social isolation and IGD is depicted in Multimedia Appendix 1. The results showed that for male students, social isolation was positively correlated with IGD (β=.906; *t*=3.891; *P*<.001); for female students, social isolation was also positively correlated with IGD (β=.410; *t*=3.407; *P*=.001), but the effect size was lower than that of male students. Our results indicate that gender moderated the association between social isolation and IGD.

Table 5. Moderating effects of gender in the association between social isolation and internet gaming disorder, the outcome variable (N=479).

Outcome variable, predictive variables	β	SE	*t* test	*P* value	*F* test (*df*)	R^2
Internet gaming disorder					25.423 (5, 473)	0.212
Self-control	–.219	.047	–4.673	<.001		
Major	–.447	.418	–1.068	.28		
Social isolation	.359	.112	3.198	.001		
Gender	–6.915	.832	–8.312	<.001		
Social isolation × gender	–.496	.248	–2.003	.046		

Discussion

Social isolation was a crucial negative feeling experienced by university students during the COVID-19 pandemic. Since students are at a higher risk of social isolation, they develop IGD. Our study reports the associations among social isolation, self-control (a personality characteristic), and IGD and examines the mediating effect of self-control on the association between social isolation and IGD.

In our study, social isolation was found to be positively correlated with IGD, which is in line with that reported previously. Yang et al [46] suggested that because of protracted periods of social isolation during the pandemic, crisis-induced distress and technology-based activities increased greatly, thereby leading to or intensifying IGD among individuals. Social isolation is believed to disrupt students' adaptive routines and learning habits and increase their loneliness, boredom, anxiety, frustration, and depression [47]. Due to isolation-related disruption and negative emotions, students tend to socialize on the internet platform to alleviate their stress and cater to their psychological needs. Hence, students are more likely to overuse internet games and ultimately experience IGD.

Our study shows a negative correlation between social isolation and self-control. This important finding is the first evidence, to the best of our knowledge, revealing the relationship between social isolation and self-control. Previous studies have shown that social isolation indicates individuals' objective deficiency in interactions and support from others and the wider community, and the high prevalence of social isolation during the pandemic affected university students' economic and social resources, coping ability, and mental health status [48]. Social isolation has been identified as a significant risk factor for emotional dysregulation and negative emotions [49,50]. Students with high levels of negative emotions tend to lose self-control in the face of difficulties and setbacks. Li et al [51] found that highly negative emotions consume individuals' cognitive resources and impair their cognitive activities, thus decreasing their self-control capacity.

In this study, a negative correlation was found between self-control and IGD. This result corroborates those of Xiang et al [52] and Cudo et al [53], who suggested that self-control serves as a protective factor for mitigating IGD among students. Previous studies have reported that self-control represents the reasoned procedure and ability to manage impulses deliberatively [54], and students with high self-control may find it easy to withstand the temptation of internet games and to

XSL•FO
RenderX

cease playing [55]. Moreover, some studies have indicated that high self-control could reduce one's aggression and sensation-seeking attitudes, and aggressive behaviors have been related to problematic internet gaming and IGD [56].

The most interesting finding of our study was that self-control was a significant mediating factor in the association between social isolation and IGD. This important finding provides evidence for understanding the generation process of IGD among university students during a global pandemic and developing effective interventions for alleviating IGD. Because of lockdown requirements during the pandemic, university students experienced high levels of social isolation, which impacted students' self-control [57]. Lower self-control increases the possibility of internet game playing, which eventually leads to IGD [30]. Therefore, university students become an extremely susceptible group for developing IGD. Targeted interventions to alleviate IGD should include training to promote the ability of self-control, increase alternative resources for human interactions, and remove sources of gaming temptations to reduce intentions and behaviors of playing internet games.

In this study, another interesting finding was that gender played a moderating role in the association between social isolation and IGD. Compared with male students, female students who experienced social isolation tended to have lesser IGD, indicating that male students are more vulnerable to confronting the challenges of social isolation and are more likely to play internet games to cope with isolation-related stress. Therefore, more attention should be given to male students when implementing interventions for reducing IGD.

Several limitations of this study should be considered before employing its findings. First, a cross-sectional study design was used that may hardly examine the causal relationships among the measured variables. A longitudinal or cohort study should be used instead to explore the causal effects of social isolation and self-control on IGD. Second, limited sociodemographic characteristics and personality traits (self-control) were investigated. Other personality traits (eg, self-efficacy, self-esteem, resilience, psychological capital) and family characteristics (eg, economic status, parents' job, relationship between parents and children) should be considered in future studies. Third, convenience sampling was used to recruit student participants, which may raise some respondent bias. Random sampling is suggested to reduce these biases. Fourth, our study used the Isolation subscale of the Self-Compassion Scale to measure social isolation. Future studies are suggested to employ more complex and reliable instruments such as the Lubben Social Network Scale-6 [58] and the Orth-Gomer Measure of Social Networks Support [59]. Our study has several notable contributions to the literature on IGD in university students in mainland China. This study provides evidence of the associations among social isolation, self-control, and IGD in university students during the COVID-19 pandemic in mainland China. In particular, the negative association between social isolation and self-control was first verified. Moreover, the significant mediating effect of self-control in the relationship between social isolation and IGD was proved. Additionally, we established evidence of the impact of gender on the relationship between social isolation and IGD. In conclusion, our study emphasizes on the challenges of social isolation and IGD among university students during a pandemic and recommend that effective interventions and prevention measures be implemented with the least delay possible.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Moderation analysis of gender in the association between social isolation and internet game addiction.
[DOCX File , 18 KB - jmir_v26i1e52978_app1.docx]

References

1. Haddad JM, Macenski C, Mosier-Mills A, Hibara A, Kester K, Schneider M, et al. The impact of social media on college mental health during the COVID-19 pandemic: a multinational review of the existing literature. Curr Psychiatry Rep 2021 Oct 06;23(11):70 [FREE Full text] [doi: 10.1007/s11920-021-01288-y] [Medline: 34613542]
2. Rainford L, Zanardo M, Buissink C, Decoster R, Hennessy W, Knapp K, et al. The impact of COVID-19 upon student radiographers and clinical training. Radiography (Lond) 2021 May;27(2):464-474 [FREE Full text] [doi: 10.1016/j.radi.2020.10.015] [Medline: 33223416]
3. Anderson AR, Knee E, Anderson KR, Ramos WD. Campus recreational participation and COVID-19: impact on college student health and well-being. J Am Coll Health 2024;72(6):1820-1827. [doi: 10.1080/07448481.2022.2093116] [Medline: 35816740]
4. Browning MHEM, Larson LR, Sharaievska I, Rigolon A, McAnirlin O, Mullenbach L, et al. Psychological impacts from COVID-19 among university students: Risk factors across seven states in the United States. PLoS One 2021;16(1):e0245327 [FREE Full text] [doi: 10.1371/journal.pone.0245327] [Medline: 33411812]
5. Wang Y, Liu M, Nogueira OCBL. Prevalence and risk factors of internet gaming disorder under the COVID-19 pandemic among university students in Macao. SAGE Open Nurs 2023;9:23779608231158158 [FREE Full text] [doi: 10.1177/23779608231158158] [Medline: 36865753]

XSL•FO
RenderX

6. Report on the internet usage of minors in China. Doc88.com. URL: https://www.doc88.com/p-13047264158693.html [accessed 2023-06-25]
7. Wang Y, Liu B, Zhang L, Zhang P. Anxiety, depression, and stress are associated with internet gaming disorder during COVID-19: fear of missing out as a mediator. Front Psychiatry 2022;13:827519 [FREE Full text] [doi: 10.3389/fpsyt.2022.827519] [Medline: 35222126]
8. American PA. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Arlington, VA, USA: American Psychiatric Publishing; 2023.
9. The ICD-11 classification of mental and behavioral disorders: diagnostic criteria for research. World Health Organization. 2018. URL: https://icd.who.int/en [accessed 2024-09-04]
10. Son S, Oh J, Jeon B. Correlational study between online friendship network and internet game disorder among university students. Brain Behav 2021 Nov;11(11):e2392 [FREE Full text] [doi: 10.1002/brb3.2392] [Medline: 34661980]
11. Vahidi M, Zamanzadeh V, Musavi S, Roshangar F, Janani R. Gaming disorder among students of Tabriz University of Medical Sciences: The frequency and related factors. Med J Islam Repub Iran 2021;35:98 [FREE Full text] [doi: 10.47176/mjiri.35.98] [Medline: 34956944]
12. Siste K, Hanafi E, Sen LT, Wahjoepramono POP, Kurniawan A, Yudistiro R. Potential correlates of internet gaming disorder among Indonesian medical students: cross-sectional study. J Med Internet Res 2021 Apr 19;23(4):e25468 [FREE Full text] [doi: 10.2196/25468] [Medline: 33871379]
13. Ayraler Taner H, Akın Sarı B, Kulaçoğlu KC, Mayalık A, Mecit H, Özekicioğlu K, et al. Internet gaming disorder's correspondence with attention deficit hyperactivity disorder and game/in-game purchases among university students. Entertainment Computing 2022 May;42:100486. [doi: 10.1016/j.entcom.2022.100486]
14. T'ng ST, Ho KH, Sim DE, Yu CH, Wong PY. The mediating effect of internet gaming disorder's symptoms on loneliness and aggression among undergraduate students and working adults in Malaysia. Psych J 2020 Feb;9(1):96-107. [doi: 10.1002/pchj.320] [Medline: 31797557]
15. Evren C, Evren B, Dalbudak E, Topcu M, Kutlu N, Elhai JD. Relationship between internet gaming disorder symptoms with attention deficit hyperactivity disorder and alexithymia symptoms among university students. ADDICTA: The Turkish Journal on Addictions 2020 Apr 07;7(1):1-9. [doi: 10.5152/addicta.2020.19076]
16. Burkauskas J, Griskova-Bulanova I, Đorić A, Balhara YPS, Sidharth A, Ransing R, et al. Association of internet gaming disorder symptoms with anxiety and depressive symptoms and substance use: an international cross-sectional study. Middle East Curr Psychiatry 2022 Feb 14;29(1):14. [doi: 10.1186/s43045-022-00180-6]
17. Cho H, Hwang Y. The effect of gaming disorder on suicidality: focusing on the mediating effect of social support and self-esteem. Korean J Food & Health Converg. 2022. URL: https://koreascience.or.kr/article/JAKO202210951905144.pdf [accessed 2024-09-04]
18. King DL, Kaptsis D, Delfabbro PH, Gradisar M. Craving for internet games? Withdrawal symptoms from an 84-h abstinence from Massively Multiplayer Online gaming. Computers in Human Behavior 2016 Sep;62:488-494. [doi: 10.1016/j.chb.2016.04.020]
19. Tavormina MGM, Tavormina R. Video games and COVID-19: How do lockdown and addiction interact? Psychiatr Danub 2021 Sep;33(Suppl 9):152-157 [FREE Full text] [Medline: 34559795]
20. De JGJ, Van TT, Dykstra P. Loneliness and social isolation. In: The Cambridge Handbook of Personal Relationships. Cambridge, UK: Cambridge University Press; 2006.
21. Elhai JD, Gallinari EF, Rozgonjuk D, Yang H. Depression, anxiety and fear of missing out as correlates of social, non-social and problematic smartphone use. Addict Behav 2020 Jun;105:106335. [doi: 10.1016/j.addbeh.2020.106335] [Medline: 32062337]
22. Liu C, Liu Z, Yuan G. Cyberbullying victimization and problematic internet use among Chinese adolescents: Longitudinal mediation through mindfulness and depression. J Health Psychol 2021 Dec;26(14):2822-2831. [doi: 10.1177/1359105320934158] [Medline: 32567374]
23. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719 [FREE Full text] [doi: 10.2196/27719] [Medline: 34081596]
24. Han TS, Cho H, Sung D, Park M. A systematic review of the impact of COVID-19 on the game addiction of children and adolescents. Front Psychiatry 2022;13:976601 [FREE Full text] [doi: 10.3389/fpsyt.2022.976601] [Medline: 36061296]
25. Ali AM, Al-Amer R, Atout M, Ali TS, Mansour AMH, Khatatbeh H, et al. The Nine-Item Internet Gaming Disorder Scale (IGDS9-SF): its psychometric properties among Sri Lankan students and measurement invariance across Sri Lanka, Turkey, Australia, and the USA. Healthcare (Basel) 2022 Mar 07;10(3):490 [FREE Full text] [doi: 10.3390/healthcare10030490] [Medline: 35326968]
26. Shirasaka T, Tateno M, Tayama M, Tsuneta M, Kimura H, Saito T. Survey of the relationship between internet addiction and social withdrawal (HIKIKOMORI) in Japan. Nihon Arukoru Yakubutsu Igakkai Zasshi 2016 Oct;51(5):275-282. [Medline: 30462389]
27. Duckworth AL, Taxer JL, Eskreis-Winkler L, Galla BM, Gross JJ. Self-control and academic achievement. Annu Rev Psychol 2019 Jan 04;70:373-399. [doi: 10.1146/annurev-psych-010418-103230] [Medline: 30609915]

XSL•FO
RenderX

28. Ozturk Eyimaya A, Sezer TA, Tezel A. Self - control and problem - solving skills of undergraduate nursing program students, and an analysis of influential factors. Perspect Psychiatric Care 2021 May 20;58(2):464-470. [doi: 10.1111/ppc.12864]
29. Li S, Nie Y. Spirituality and cyberaggression: mediating and moderating effect of self-control and school climate. Int J Environ Res Public Health 2023 Feb 08;20(4):2973 [FREE Full text] [doi: 10.3390/ijerph20042973] [Medline: 36833669]
30. Yu Y, Fong VWI, Ng JH, Wang Z, Tian X, Lau JTF. The associations between loneliness, hopelessness, and self-control and internet gaming disorder among university students who were men who have sex with men: cross-sectional mediation study. J Med Internet Res 2023 Jan 17;25:e43532 [FREE Full text] [doi: 10.2196/43532] [Medline: 36649059]
31. Liu Y, Liu J, Wang W. The effect of future time perspective on internet dependence among college students in pandemic time: a moderated mediation analysis. PRBM 2023 Feb;16:509-520. [doi: 10.2147/prbm.s396622]
32. Hamama R, Ronen T, Feigin R. Self-control, anxiety, and loneliness in siblings of children with cancer. Social Work in Health Care 2000 Aug 14;31(1):63-83. [doi: 10.1300/j010v31n01_05]
33. Barrón-Martínez JB, Salvador-Cruz J. Social abilities in young Mexicans with Down syndrome during the COVID-19 pandemic. Int J Dev Disabil 2023;69(6):888-895 [FREE Full text] [doi: 10.1080/20473869.2022.2038527] [Medline: 37885837]
34. Ryan RM, Deci EL, Bouffard L. Self-determination theory. Basic psychological needs in motivation, development and wellness. Revue québécoise de psychologie 2017;38(3):231. [doi: 10.7202/1041847ar]
35. Converse BA, Juarez L, Hennecke M. Self-control and the reasons behind our goals. J Pers Soc Psychol 2019 May;116(5):860-883. [doi: 10.1037/pspp0000188] [Medline: 30070566]
36. Mills DJ, Allen JJ. Self-determination theory, internet gaming disorder, and the mediating role of self-control. Computers in Human Behavior 2020 Apr;105:106209. [doi: 10.1016/j.chb.2019.106209]
37. Neff KD. The development and validation of a scale to measure self-compassion. Self and Identity 2003 Jul;2(3):223-250. [doi: 10.1080/15298860309027]
38. Chen J, Yan L, Zhou L. Reliability and validity of Chinese version of self-compassion scale. Chinese J Clini Psychol. 2011. URL: https://psycnet.apa.org/record/2012-05439-005 [accessed 2024-09-04]
39. Tangney JP, Baumeister RF, Boone AL. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. J Pers 2004 Apr;72(2):271-324. [doi: 10.1111/j.0022-3506.2004.00263.x] [Medline: 15016066]
40. Tan S, Guo Y. Revision of self-control scale for Chinese college students. Chinese J Clinil Psychol. 2008. URL: https://psycnet.apa.org/record/2008-19347-007 [accessed 2024-09-04]
41. Lemmens JS, Valkenburg PM, Gentile DA. The Internet Gaming Disorder Scale. Psychol Assess 2015 Jun;27(2):567-582. [doi: 10.1037/pas0000062] [Medline: 25558970]
42. Lei W, Liu K, Zeng Z, Liang X, Huang C, Gong K, et al. The psychometric properties of the Chinese version internet gaming disorder scale. Addict Behav 2020 Jul;106:106392. [doi: 10.1016/j.addbeh.2020.106392] [Medline: 32193072]
43. Wu AMS, Lai MHC, Zhang M, Yogo M, Yu SM, Mao S, et al. Effects of psychological distress and coping resources on internet gaming disorder: comparison between Chinese and Japanese University Students. IJERPH 2022 Mar 03;19(5):2951. [doi: 10.3390/ijerph19052951]
44. Wen-Juan-Xing. Changsha Ranxing Information Technology. URL: https://www.wjx.cn [accessed 2022-04-01]
45. Chang S, van Witteloostuijn A, Eden L. From the Editors: Common method variance in international business research. J Int Bus Stud 2010 Feb 8;41(2):178-184. [doi: 10.1057/jibs.2009.88]
46. Yang X, Yip BHK, Lee EKP, Zhang D, Wong SYS. The relationship between technology use and problem technology use and potential psychosocial mechanisms: population-based telephone survey in community adults during COVID-19. Front Psychol 2021;12:696271 [FREE Full text] [doi: 10.3389/fpsyg.2021.696271] [Medline: 34434146]
47. King DL, Achab S, Higuchi S, Bowden-Jones H, Müller KW, Billieux J, et al. Gaming disorder and the COVID-19 pandemic: Treatment demand and service delivery challenges. J Behav Addict 2022 Apr 12;11(2):243-248 [FREE Full text] [doi: 10.1556/2006.2022.00011] [Medline: 35413005]
48. Leigh-Hunt N, Bagguley D, Bash K, Turner V, Turnbull S, Valtorta N, et al. An overview of systematic reviews on the public health consequences of social isolation and loneliness. Public Health 2017 Nov;152:157-171. [doi: 10.1016/j.puhe.2017.07.035] [Medline: 28915435]
49. Liu B, Wu X, Shi L, Li H, Wu D, Lai X, et al. Correlations of social isolation and anxiety and depression symptoms among patients with breast cancer of Heilongjiang province in China: The mediating role of social support. Nurs Open 2021 Jul;8(4):1981-1989 [FREE Full text] [doi: 10.1002/nop2.876] [Medline: 33939294]
50. Wolters NE, Mobach L, Wuthrich VM, Vonk P, Van der Heijde CM, Wiers RW, et al. Emotional and social loneliness and their unique links with social isolation, depression and anxiety. J Affect Disord 2023 May 15;329:207-217 [FREE Full text] [doi: 10.1016/j.jad.2023.02.096] [Medline: 36842647]
51. Li X, Li W, Liu M, Xiao W, Zhou H. How does shyness affect Chinese college students' tendency to mobile phone addiction? testing the mediating roles of social anxiety and self-control. Front Public Health 2022;10:902425 [FREE Full text] [doi: 10.3389/fpubh.2022.902425] [Medline: 35910898]

XSL•FO
RenderX

52. Xiang G, Gan X, Jin X, Zhang Y, Zhu C. Developmental assets, self-control and internet gaming disorder in adolescence: testing a moderated mediation model in a longitudinal study. Front Public Health 2022;10:808264 [FREE Full text] [doi: 10.3389/fpubh.2022.808264] [Medline: 35186844]
53. Cudo A, Wojtasiński M, Tużnik P, Fudali-Czyż A, Griffiths MD. The relationship between depressive symptoms, loneliness, self-control, and gaming disorder among Polish male and female gamers: the indirect effects of gaming motives. Int J Environ Res Public Health 2022 Aug 22;19(16):10438 [FREE Full text] [doi: 10.3390/ijerph191610438] [Medline: 36012072]
54. Finkenauer C, Engels RCME, Baumeister RF. Parenting behaviour and adolescent behavioural and emotional problems: The role of self-control. International Journal of Behavioral Development 2016 Jun 30;29(1):58-69. [doi: 10.1080/01650250444000333]
55. Yu Y, Mo PK, Zhang J, Li J, Lau JT. Impulsivity, self-control, interpersonal influences, and maladaptive cognitions as factors of internet gaming disorder among adolescents in China: cross-sectional mediation study. J Med Internet Res 2021 Oct 27;23(10):e26810 [FREE Full text] [doi: 10.2196/26810] [Medline: 34704960]
56. Jeong H, Yim HW, Lee S, Lee HK, Potenza MN, Jo S, et al. Low self-control and aggression exert serial mediation between inattention/hyperactivity problems and severity of internet gaming disorder features longitudinally among adolescents. J Behav Addict 2020 Jun;9(2):401-409 [FREE Full text] [doi: 10.1556/2006.2020.00039] [Medline: 32634112]
57. Loades ME, Chatburn E, Higson-Sweeney N, Reynolds S, Shafran R, Brigden A, et al. Rapid systematic review: the impact of social isolation and loneliness on the mental health of children and adolescents in the context of COVID-19. J Am Acad Child Adolesc Psychiatry 2020 Nov;59(11):1218-1239.e3 [FREE Full text] [doi: 10.1016/j.jaac.2020.05.009] [Medline: 32504808]
58. Lubben J, Blozik E, Gillmann G, Iliffe S, von Renteln Kruse W, Beck JC, et al. Performance of an abbreviated version of the Lubben Social Network Scale among three European community-dwelling older adult populations. Gerontologist 2006 Aug;46(4):503-513 [FREE Full text] [doi: 10.1093/geront/46.4.503] [Medline: 16921004]
59. Undén AL, Orth-Gomér K. Development of a social support instrument for use in population surveys. Soc Sci Med 1989;29(12):1387-1392. [doi: 10.1016/0277-9536(89)90240-2] [Medline: 2629121]

Abbreviations

***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, fifth edition*
***ICD-11*:** *eleventh revision of the International Classification of Diseases*
IGD: internet gaming disorder

Edited by T de Azevedo Cardoso; submitted 21.09.23; peer-reviewed by A AL-Asadi, D Abdel-Hady; comments to author 19.06.24; revised version received 24.06.24; accepted 07.08.24; published 10.09.24.

Please cite as:
Guo Y, Yue F, Lu X, Sun F, Pan M, Jia Y
COVID-19–Related Social Isolation, Self-Control, and Internet Gaming Disorder Among Chinese University Students: Cross-Sectional Survey
J Med Internet Res 2024;26:e52978
URL: https://www.jmir.org/2024/1/e52978
doi:10.2196/52978
PMID:

©Yufang Guo, Fangyan Yue, Xiangyu Lu, Fengye Sun, Meixing Pan, Yannan Jia. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Virtual Reality Motor-Cognitive Training for Older People With Cognitive Frailty: Multicentered Randomized Controlled Trial

Rick Yiu Cho Kwan[1*], PhD; Justina Liu[2,3*], PhD; Olive Suk Kan Sin[4], MSW, MSc; Kenneth N K Fong[5], PhD; Jing Qin[2], PhD; Joe Chi Yin Wong[6], BEng; Claudia Lai[2], PhD

[1]School of Nursing, Tung Wah College, Hong Kong, China (Hong Kong)
[2]School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[3]Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[4]Board of Director, Pok Oi Hospital, Hong Kong, China (Hong Kong)
[5]Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[6]Sengital Limited, Hong Kong, China (Hong Kong)
[*]these authors contributed equally

Corresponding Author:
Rick Yiu Cho Kwan, PhD
School of Nursing
Tung Wah College
31 Wylie Road, Homantin
Hong Kong, NA
China (Hong Kong)
Phone: 852 34686813
Fax: 852 27821566
Email: rickkwan@twc.edu.hk

Abstract

Background: Cognitive frailty refers to a clinical syndrome in which physical frailty and mild cognitive impairment coexist. Motor-cognitive training and virtual reality (VR) have been used to launch various therapeutic modalities to promote health in older people. The literature advocates that motor-cognitive training and VR are effective in promoting the cognitive and physical function of older people. However, the effects on older people with cognitive frailty are unclear.

Objective: This study examined the effects of VR motor-cognitive training (VRMCT) on global cognitive function, physical frailty, walking speed, visual short-term memory, inhibition of cognitive interference, and executive function in older people with cognitive frailty.

Methods: This study used a multicentered, assessor-blinded, 2-parallel-group randomized controlled trial design. Participants were recruited face-to-face in 8 older adult community centers. Eligible participants were aged ≥60 years, were community dwelling, lived with cognitive frailty, had no dementia, and were not mobility restricted. In the intervention group, participants received VRMCT led by interventionists with 16 one-hour training sessions delivered twice per week for 8 weeks. In the control group, participants received the usual care provided by the older adult community centers that the investigators did not interfere with. The primary outcome was global cognitive function. The secondary outcomes included physical frailty, walking speed, verbal short-term memory, inhibition of cognitive interference, and executive function. Data were collected at baseline (T0) and the week after the intervention (T1). Generalized estimating equations were used to examine the group, time, and interaction (time × group) effects on the outcomes.

Results: In total, 293 eligible participants enrolled in the study. The mean age of the participants was 74.5 (SD 6.8) years. Most participants were female (229/293, 78.2%), had completed primary education (152/293, 52.1%), were married (167/293, 57.2%), lived with friends (127/293, 43.3%), and had no VR experience (232/293, 79.5%). In the intervention group, 81.6% (119/146) of participants attended >80% (13/16, 81%) of the total number of sessions. A negligible number of participants experienced VR sickness symptoms (1/146, 0.7% to 5/146, 3%). VRMCT was effective in promoting global cognitive function (interaction effect: P=.03), marginally promoting executive function (interaction effect: P=.07), and reducing frailty (interaction effect: P=.03). The effects were not statistically significant on other outcomes.

XSL•FO
RenderX

Conclusions: VRMCT is effective in promoting cognitive functions and reducing physical frailty and is well tolerated and accepted by older people with cognitive frailty, as evidenced by its high attendance rate and negligible VR sickness symptoms. Further studies should examine the efficacy of the intervention components (eg, VR vs non-VR or dual task vs single task) on health outcomes, the effect of using technology on intervention adherence, and the long-term effects of the intervention on older people with cognitive frailty at the level of daily living.

Trial Registration: ClinicalTrials.gov NCT04730817; https://clinicaltrials.gov/study/NCT04730817

(J Med Internet Res 2024;26:e57809) doi:10.2196/57809

KEYWORDS

virtual reality; motor-cognitive training; cognitive frailty; gamification

Introduction

Background

Cognitive frailty (also known as potentially reversible cognitive frailty) refers to a clinical syndrome in which physical frailty (including physical prefrailty or physical frailty) and mild cognitive impairment (MCI) coexist [1,2]. The prevalence in older people is increasing, as shown in a systematic review that showed that the pooled estimates of cognitive frailty prevalence in community-dwelling older populations increased from 6% (2012-2017) to 11% (2018-2020) [3]. Another systematic review showed that cognitive frailty is commonly seen in older people aged ≥60 years and that older age and a lower level of engagement in activities are associated with a higher risk of cognitive frailty [4]. It was found to be a significant predictor of adverse health outcomes, including all-cause mortality and dementia. Cognitive frailty was found to be a better predictor of adverse health outcomes (eg, dementia and depression) than physical frailty alone [5,6]. Recent evidence has shown that effective interventions enhancing cognitive and physical functions in older people with cognitive frailty are mostly multidimensional but commonly include the component of physical and cognitive training [2,7,8].

Motor-cognitive training (also known as dual-task training) refers to training in which both physical and cognitive training are performed at the same time (ie, thinking about how to solve a problem while moving) [9]. The Guided Plasticity Facilitation framework proposed that the combination of physical and cognitive activities has positive synergistic effects that exceed the positive effects of either cognitive or physical training alone because the facilitation effects of physical training and the guidance effects of cognitive training trigger a neurophysiological mechanism to promote neuroplasticity, as evidenced by the enhanced release of neurotrophic factors (eg, brain-derived neurotrophic factor) [10]. In the literature, evidence has shown that motor-cognitive training is effective in treating many neurological diseases (eg, chronic stroke and Parkinson disease) to improve gait performances and risk of falls [11,12], but its effects on cognitive function are not clear in people with Parkinson disease [13]. Recently, motor-cognitive training has been used to treat people with MCI, and there is preliminary evidence suggesting that it might improve cognitive function in addition to physical function [14-16].

Virtual reality (VR) refers to the use of a computer system that aims to immerse users in a virtual environment by replacing visual and aural environments to achieve a sense of presence so that users perceive themselves as being part of this virtual environment [17]. VR is often used as a platform to launch various therapeutic modalities in older people to promote various aspects of their health, including cognitive, physical, and mental health, with promising effects when compared with traditional training methods (eg, conducted by a therapist in real time) [18-20]. People with MCI commonly display impaired attentional processing and working memory capacity [21]. Physical activity adherence is proven to be poor in older people with cognitive impairment [22]. VR is an ideal platform to launch therapies because it allows for a more controllable intervention environment through simulation and it attracts participants' attention, resulting in more interaction with the training contributing to better training results [23].

Objectives

The literature advocates that the elements of motor-cognitive training and VR are effective in promoting the cognitive and physical functions of older people with various chronic conditions (eg, stroke and MCI) [11,14,24]. However, their effects on older people with cognitive frailty are unclear. A pilot study was conducted by our team in Hong Kong that examined the effects of a VR motor-cognitive training (VRMCT) system (ie, the same training system tested in this study) on 17 older people with cognitive frailty [25]. The preliminary evidence (ie, the within-group effect of 9 participants) of the pilot study supported that the VRMCT was effective in promoting cognitive and physical functions and reducing physical frailty and the intervention was feasible to be implemented in people with cognitive frailty [25]. Nevertheless, there are no full-powered randomized controlled trials to provide proof of its effects. Therefore, the objectives of this study were to examine the effects of VRMCT on global cognitive function, physical frailty, walking speed, visual short-term memory, inhibition of cognitive interference, and executive function in older people with cognitive frailty.

Methods

Trial Design

This study used a multicentered, assessor-blinded, 2-parallel-group, 1:1-allocation-ratio randomized controlled trial design. The CONSORT (Consolidated Standards of Reporting Trials) 2010 guidelines were followed to report on this trial [26]. This trial was registered at ClinicalTrials.gov

XSL•FO
RenderX

with the last updated version on August 27, 2021 (ClinicalTrials.gov ID: NCT04730817).

Participants

Overview

Participant recruitment was conducted in 8 older adult community centers (ie, District Elderly Community Centres), which provide support services at the district level in Hong Kong to enable older people to remain in the community and lead a healthy, respectful, and dignified life [27]. The research team produced a set of intervention promotional materials (eg, flyers, videos, newsletters, and presentation slides). The staff members of the older adult community centers used promotional materials with the institutional affiliation's name displayed to invite interested and potentially eligible center members to enroll in this study. Upon the referral by the staff members of the older adult community centers and with the approval of the potential participants, the research team members conducted eligibility assessments and, subsequently, the demographic and outcome assessment for eligible participants at the older adult community centers. All eligible older people were invited to participate in this study.

Inclusion Criteria

The inclusion criteria were as follows:

1. The participants were aged ≥60 years.
2. The participants were community dwelling, defined as not living in a long-term care facility in the previous 12 months.
3. The participants had cognitive frailty, defined as the coexistence of MCI and physical frailty without concurrent dementia [2,28]. MCI is measured through a Montreal Cognitive Assessment (MoCA) score of <26/30 [29]. Physical frailty is measured as a Fried frailty phenotype (FFP) score of >0 [30]. The measurement of each phenotypic criterion followed the Fried method using local normative data [31]. Exclusion of possible dementia was measured through a MoCA score of <19/30 [32].

Exclusion Criteria

The exclusion criteria were as follows:

1. There was a confirmed diagnosis of dementia, defined as the documented diagnosis on the participants' medical record.
2. The participants had restricted mobility, defined as a Modified Functional Ambulation Classification score of <7 [33].

Interventions

Design

There were 2 parallel groups in this study. The intervention group received VRMCT, whereas the control group received usual care.

Intervention Group: VRMCT

The intervention was to deliver physical and cognitive training simultaneously on the VR platform. As shown in Figure 1, the VRMCT system comprises 3 parts: hardware, software and VR headset, and computer. The hardware consists of a motion sensor, a microcontroller unit with a Bluetooth Low Energy feature, and a power-controlled integrated circuit. The hardware also consists of a bicycle, which is an underdesk ergometer with adjustable cycling resistance (ie, DeskCycle 2). The motion sensor is attached to the paddle of the bicycle to capture the cycling motion of the participants. Once the sensor captures motion data, it communicates with the microcontroller unit. The motion sensor is specifically designed for seamless integration with the system. The second part of the VRMCT system is the software and VR headset, which consists of a VR headset with related hardware, including a head-mounted VR display with a pair of headphones and wireless handheld controllers (ie, HTC Vive Focus Plus). The VR games are installed in the HTC Vive Focus Plus headset. Using the communication protocol, the motion sensor is connected wirelessly to the VR headset. As a result, the participants can control their movement in the virtual environment by cycling on the motion sensor–connected bicycle. The third part of the VRMCT system is the computer, which stores users' information and records the data logged by the system during the process of training. Using a similar communication protocol, the computer is connected wirelessly to the software and VR headset to record the data.

XSL•FO
RenderX

Figure 1. The design of the virtual reality (VR) motor-cognitive training system.

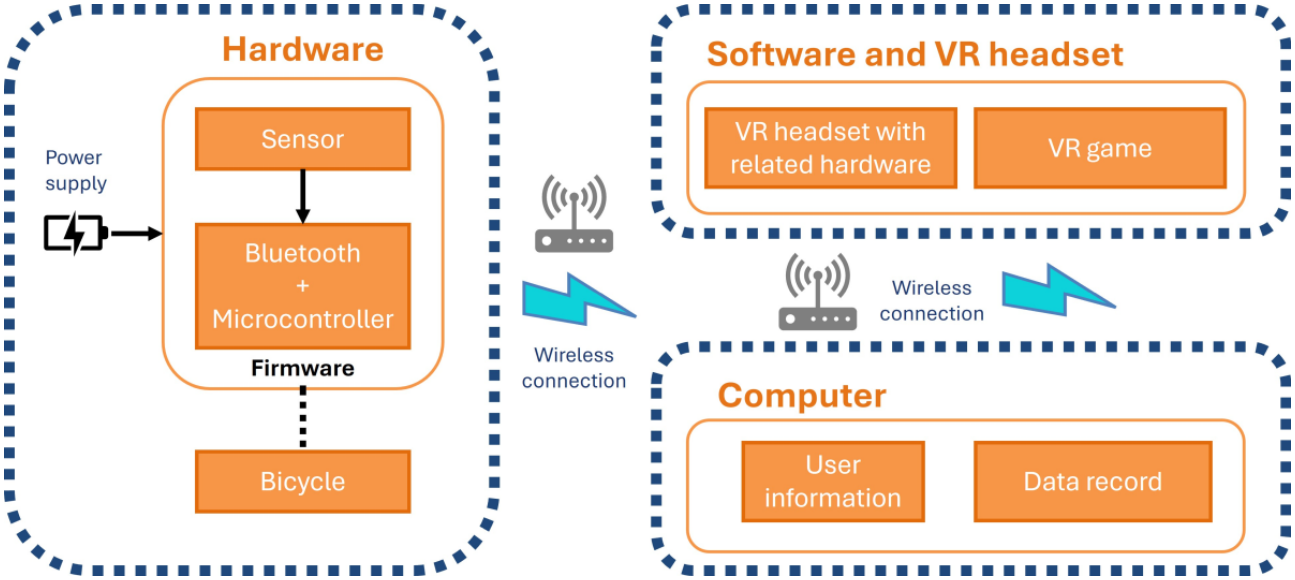

The VR game was developed by an interprofessional team with the needed expertise, including nurses (RYCK, JL, and CL), an occupational therapist (KNKF), a social worker (OSKS), and computer engineers (JQ and JCYW) aiming to develop a prototype effective at promoting the health of older people with cognitive frailty and technically usable and accepted by them. As shown in Table 1, the VR game contains 8 different themes: orientation, finding a bus stop, reporting lost items, finding a supermarket, grocery shopping, cooking, finding a travel hot spot, and bird watching. They mimic and gamify the problems faced and activities undertaken by older people regularly. In week 1, orientation sessions were provided to the participants aiming to teach them to master all the commands needed in the training (eg, cycling the ergometer to control the movement and speed in the virtual world and using the handheld controller to select and unselect items). In the remaining weeks, participants were instructed to navigate the virtual world by cycling on the motion sensor–connected ergometer and solve the gamified problems by exercising their cognitive capabilities in the virtual world. For example, in week 2, participants were expected to navigate the virtual world to find a designated bus stop to take a bus to a supermarket to continue the game. Participants were expected to memorize and recognize the visual cues in the virtual environment (eg, traffic lights and building names) to find the designated bus stop; while they were doing so, they would be exercising their visuospatial abilities, as shown in Multimedia Appendix 1. As another example, in week 5, participants were expected to shop for food in the supermarket. Participants were expected to exercise their memory to recall the names of the food items needed for cooking, pay attention to find the image of the food items needed, and do the calculation to settle the payment, as shown in Multimedia Appendix 2. The illustration of the setup of the intervention on the participants is shown in Figure 2.

Figure 2. The setup of the virtual reality motor-cognitive training system.

XSL•FO
RenderX

Table 1. Themes of the virtual reality motor-cognitive training system.

Week	Themes	Purposes	Cognitive functions required
1	Orientation	Learn all the commands in the training	__[a]
2	Finding a bus stop	Find a bus stop on a given route in a city	• Attention • Visuospatial function
3	Reporting lost items	Report a lost item found on the street to police	• Problem-solving • Visuospatial function
4	Finding a supermarket	Find a particular supermarket in the city	• Attention • Visuospatial function
5	Grocery shopping	Shop in a supermarket for a list of food items and calculate the price of the selected items	• Memory • Attention • Calculation
6	Cooking	Flip eggs at a specific time interval	• Mental processing speed
7	Finding a travel hot spot	Find a travel hot spot in a park	• Attention • Visuospatial function
8	Bird watching	Identify the birds in a park that were shown at the beginning of the game	• Attention • Memory

[a]Not applicable.

The VR motor-cognitive intervention was provided in 8 older adult community centers in Hong Kong. The duration of the whole intervention was 8 consecutive weeks, including two 1-hour sessions each week. The actual use of VR devices was limited to 30 minutes per session to minimize possible cyber fatigue or cybersickness because a longer exposure time increases the risk of VR sickness [34]. The duration and frequency were designed according to the pilot study, which was shown to be sufficient to yield clinical outcomes on global cognitive function and physical frailty [25]. Within the same week, the theme of the 2 training sessions was the same, but the levels of difficulty of the 2 sessions in terms of cognitive challenge increased. Participants attended the easier one first, followed by the harder one. The total number of training sessions was 16, and each training session lasted for 1 hour, including briefing, setup, VRMCT (30 minutes), and posttraining observation. In the briefing, participants were primed with the training objectives of the session (eg, to find a bus stop) and reminded about the essential controls needed. In the setup, all items of the VRMCT system (ie, motion sensor–connected ergometer, VR headset, and the computer) were wirelessly connected via Bluetooth. The resistance level of the ergometer was then set, with a goal of progressive increase to the maximum tolerable level according to the participants' previous training experience (eg, knee pain, muscle ache, and fatigue) and motivation to increase the training dose (ie, the resistance level of the ergometer). After that, VR headset fitting was conducted on the participants aiming to make sure that the position of the VR headset allowed for the best-quality 3D vision experienced by the participants. In the VRMCT, the participants followed the computer-generated and interventionist's instructions to complete the training tasks. Each session was expected to be completed within 20 to 30 minutes according to participants' cognitive performance. The session ended if the training lapsed for 30 minutes. Finally, immediately after the training, participants were placed in a safe sitting position with the VR headset removed and observed for at least 10 minutes. This was because potential adverse effects were commonly observed in older people (eg, VR sickness, dizziness, and eyestrain) during the period [35]. Participants were only allowed to leave if they had no adverse symptoms.

During the intervention, the participants wore the head-mounted VR display and were immersed in the virtual world of the self-developed VR game with the voice-over and written instructions in the VR game and the verbal instructions of the accompanying on-site interventionists. The interventionists were undergraduate nursing students who were trained by following an implementation protocol (eg, setup of the system, technical problem-solving, common adverse effect identification, and provision of cues to participants in given situations) to guide participants through the training sessions. The intervention was delivered by the interventionists at a 1:1 interventionist-to-participant ratio. The interventionists delivered the interventions and accompanied the participants throughout the whole intervention period. All interventionists completed a standardized training, including a 1-hour training and returned demonstration. The training contents included the setup of the equipment, communication with older people with cognitive frailty, equipment fitting on participants, training objectives of each session, instructions provided to participants, troubleshooting, observable adverse effects, and emergency management. To ensure intervention fidelity, a research team member (author RYCK) regularly and randomly site visited the performance of the interventionists' implementation. A checklist mapped with the intervention implementation protocol was used to guide the observation of the interventionists. When deviations from the intervention implementation protocol were observed, remedial training was provided to the interventionists by the

XSL•FO
RenderX

research team member (RYCK) at the site. When technical problems were observed or reported, they were fixed by the technical team (JCYW).

To ensure adherence to the intervention, the participants were reminded by either the interventionists or the older adult community center staff members to attend the training 1 day before the training. All participants who completed the training for >80% of the sessions (ie, 13/16, 81%) were presented with a certificate of completion in a mini ceremony of completion held in the older adult community center. Some scheduled sessions could not be attended by participants because of some unforeseeable reasons (eg, illnesses) and were rescheduled to ensure that 16 sessions were provided to all participants. To ensure an adequate intervention dose, up to 2 makeup sessions were arranged within 2 weeks after the completion of the scheduled period of 8 consecutive weeks for participants who missed some sessions.

Participants accessed the intervention at the designated older adult community centers. The intervention was delivered by the interventionists who were authorized to operate the VRMCT system. Only authorized interventionists could deliver the intervention and operate the system. The systems were owned by the research team. All eligible participants in the study received the intervention for free.

Apart from the intervention provided by the research team, participants led their normal lives and were not restricted from enrolling in or engaging in any activities (eg, physical or social activities) provided by the community centers or organizations outside the community centers.

Waitlist Control Group: Usual Care

Participants in this group received usual care. The research team did not provide any interventions to the participants in this group during the 8-week intervention period. Similarly to the participants in the intervention group, participants in the control group were also not restricted from enrolling on or engaging in any activities (eg, physical or social activities) provided by the community centers or organizations outside the community centers. However, a waitlist control method was adopted, and an 8-week and 16-session VRMCT was provided to all participants in this group after the completion of follow-up data collection (ie, 8 weeks after receiving usual care from the baseline).

Outcomes

Overview

Demographic data were collected at baseline (ie, T0 at week 0) that included age, number of chronic illnesses according to the Charlson Comorbidity Index [36], BMI, gender, educational level, marital status, living status, and VR experience. Outcome variables included global cognitive function (primary), physical frailty, walking speed, visual short-term memory, inhibition of cognitive interference, and executive function. These were measured at both baseline (ie, T0 at week 0) and the week immediately after the completion of the intervention (ie, T1 at week 9).

Global Cognitive Function

Global cognitive function was measured using the MoCA, which comprises 30 dichotomous items [32]. One point is assigned to one correct answer. The total score ranges from 0 to 30. A higher score indicates a higher level of cognitive function. The MoCA has a strong correlation with the Mini-Mental State Examination (r=0.90) and Saint Louis University Mental Status Examination (r=0.83) [37], as well as a good criterion validity in detecting MCI [29,32].

Physical Frailty

Physical frailty was measured using the FFP, which assesses physical frailty based on 5 components: weight loss, exhaustion, low physical activity, slow gait, and weakness [30]. One point is assigned to the presence of one component, and the FFP score ranges from 0 to 5. A higher score indicates a higher level of frailty.

Walking Speed

Walking speed was measured using the Timed Up and Go test, which assesses the total time needed for a participant to stand up from a chair, walk a 3-meter distance, walk back to the chair, and sit down. A longer time needed indicates a slower walking speed [38].

Verbal Short-Term Memory

Verbal short-term memory was measured using the digit span test [39]. The digit span test is conducted through forward and backward digit recall and scored using the longest forward and backward digit string recalled without error. A higher score indicates a higher level of verbal short-term memory. The digit span test was validated to classify people with different levels of memory impairment or disorder [40,41]. Verbal short-term memory was assessed as an outcome because verbal short-term memory training was integrated as part of the VRMCT intervention.

Inhibition of Cognitive Interference

Inhibition of cognitive interference was measured using the Stroop Color and Word Test. This popular test assesses a person's ability to inhibit cognitive interference when processing a specific stimulus while a second stimulus is simultaneously introduced. This outcome was assessed because participants were also trained to respond to 2 stimuli (cognitive and motor) as part of the VRMCT intervention. The number of correct answers produced in a fixed time (ie, 30 seconds) were counted separately in the color test, word test, and color-word test. The inhibition score is calculated using the formula I = color-word test – ([word test + color test] / 2) [42]. A higher score indicates a higher level of inhibition of cognitive interference.

Executive Function

Executive function was measured using the Trail Making Test (TMT). Time is counted for the completion of both the TMT part A and TMT part B. The TMT score is calculated by dividing the TMT part B by the TMT part A [43]. A higher score indicates a lower level of executive function. This outcome was assessed because participants were also trained to exercise their executive function as part of the VRMCT intervention.

XSL•FO
RenderX

Ancillary Variables

Ancillary variables included VR sickness, unstructured untoward effects on participants, and unplanned events in the training sessions. These were measured during the intervention period every time after the completion of each session of the training.

VR sickness was measured using the Virtual Reality Sickness Questionnaire (VRSQ) [44]. The VRSQ consists of 9 simulator sickness symptoms: general discomfort, fatigue, eyestrain, difficulty focusing, headache, fullness of head, blurred vision, dizziness, and vertigo. The severity of each symptom is rated using a 4-point Likert scale (ie, 0=*never*; 3=*very*). The total score was the summation of all item scores and was then converted to a percentage score. A higher score indicates a higher severity of VR sickness. The VRSQ was validated as a reliable tool (Cronbach α=0.85-0.89) [44].

Open-ended questions were posed to all participants (ie, "Do you have any uncomfortable symptoms apart from the symptoms covered in the VRSQ?") and interventionists ("Did you observe any unusual symptoms on the participants or conditions throughout the intervention period?"). The questions were asked through a questionnaire launched on a web-based platform (ie, Qualtrics [Qualtrics International Inc]).

Sample Size

We adopted a previous power analysis using the web-based software GLIMMPSE and used a general linear mixed model [45]. We set the level of significance at 0.05, the power at 0.9, the number of repeated measures at 2 (ie, T0 and T1), the number of groups at 2 (ie, the intervention and control groups), and the allocation ratio between the 2 groups at 1:1. To estimate the effects, we referred to the interaction (ie, group × time) effect on global cognitive function (ie, primary outcome) observed in the pilot study with a highly identical design [25]. The estimated sample size was 220. We assumed a dropout rate of 6%, as observed in the pilot study [25]. The total sample size was expected to be at least 234 participants, with 117 in each group.

Randomization

A permuted block randomization with a block size of 8 in an intervention-to-control ratio of 1:1 was used. A list of random numbers of either 1 or 2 (ie, 1=intervention; 2=control) of each block was generated independently by author RYCK using the RAND formula in Microsoft Excel (Microsoft Corp). After each round of eligibility screening, a list of unique participant codes was assigned to each eligible participant by the participant recruitment team. A research assistant assigned all eligible participants to either the intervention or control group by matching the participant codes with the list of random group numbers. To ensure concealment, the research assistant managing the random group allocation did not participate in any part of recruitment.

Blinding

All outcome assessors were blinded to the group labels. We also endeavored to blind the participants to the group label at baseline assessment through an arrangement in which the random group allocation was only conducted after the completion of the baseline assessment. However, it was not possible to blind the interventionists or participants given the nature of the study.

Statistical Methods

SPSS (version 26; IBM Corp) was used to conduct the data analysis. The demographic and outcome variables at continuous levels of measurement were first checked for normality, which is assumed when the absolute skewness is <2 and the absolute kurtosis is <7 [46]. Continuous variables with normality assumed were described using mean and SD or otherwise using median and IQR. Variables at categorical levels of measurement were described using frequency and percentage. To test the effects of the intervention over the 6 outcomes, 6 independent generalized estimating equations were used. The independent variables included group (ie, intervention vs control), time (ie, T0 vs T1), and group × time. The dependent variables were the 6 outcomes of the study (ie, cognitive function, physical frailty, walking speed, inhibition of cognitive interference, executive function, and walking speed). Missing values were estimated using the generalized estimating equations [47]. The level of significance was set at 0.05. Intention-to-treat analysis was adopted.

Ethical Considerations

All participants signed the written informed consent form. Before signing the written consent form, the participants were well informed about the expected benefits of the intervention, their rights (eg, right of withdrawal), the possibility of being assigned to either intervention or control, and possible risks. All data are anonymized to ensure confidentiality. Because all participants received the interventions (waitlist control was used), no compensation was provided to participants for their participation in the research. The ethical clearance for research involving human subjects for this study was provided by the institutional review board of the Hong Kong Polytechnic University (reference HSEAR202100115001-01).

Results

Participant Flow

This study was conducted from January 2021 to April 2023. As shown in Figure 3, a total of 561 participants were assessed for eligibility. In total, 47.8% (268/561) of the participants were excluded because they did not fulfill the eligibility criteria. No eligible participants declined to participate. A total of 293 eligible participants were randomly allocated to the 2 groups: 146 (49.8%) were allocated to the intervention group, and 147 (50.2%) were allocated to the control group. In the intervention group, 4.8% (7/146) of the participants did not receive the allocated intervention, defined as attending 0 training sessions. A total of 81.6% (119/146) of the participants attended ≥81% (13/16) of the expected sessions, 5.9% of the participants attended 50% (8/16) to 75% (12/16) of the expected sessions, and 12.5% of the participants attended 0% to 44% (7/16) of the expected sessions. In the control group, all 147 participants received usual care. After the completion of the intervention, in the intervention group, 14.4% (21/146) of the participants were lost to follow-up, and 23% (7/21) of these participants

XSL•FO
RenderX

discontinued the intervention, defined as attending 0 training sessions. In the control group, 7.5% (11/147) of the participants were lost to follow-up.

Figure 3. Participant flowchart.

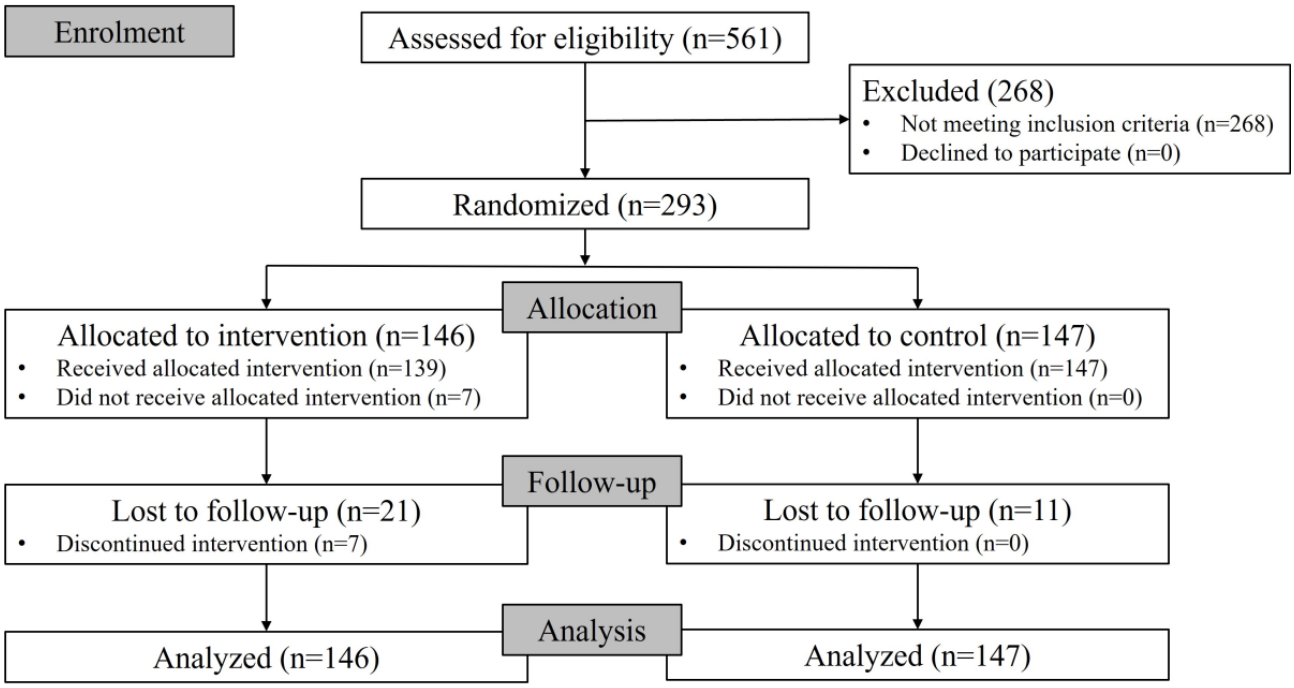

Baseline Data

As shown in Table 2, the mean age of the participants was 74.5 (SD 6.8) years, and most of the participants had no chronic illnesses (175/293, 59.7%). The mean BMI of the participants was 24.3 (SD 3.6) kg/m^2. Most of the participants were female (229/293, 78.2%), had completed primary education (152/293, 52.1%), were married (167/293, 57.2%), lived with friends (127/293, 43.3%), and had no VR experience (232/293, 79.5%). The mean MoCA score was 21.3 (SD 2.4), the mean FFP score was 1.6 (SD 0.8), the median Timed Up and Go test score was 10.9 (IQR 3.6) seconds, the mean digit span forward test score was 7.1 (SD 1.2), the mean digit span backward test score was 3.8 (SD 1.3), the mean Stroop Color and Word Test score was –21.4 (SD 9.1), and the mean TMT score was 1.7 (SD 0.7).

XSL•FO
RenderX

Table 2. Demographic and outcome variables at baseline (N=293).

Variables	All	Intervention (n=146)	Control (n=147)
Demographic variables			
Age (y), mean (SD)	74.5 (6.8)	75.2 (7.1)	73.9 (6.6)
Number of CIs[a], n (%)			
0	175 (59.7)	87 (59.6)	88 (59.9)
1-3	113 (38.6)	57 (39.0)	56 (38.1)
4-6	5 (1.7)	2 (1.4)	3 (2)
BMI (kg/m^2), mean (SD)	24.3 (3.6)	24.4 (3.7)	24.2 (3.5)
Gender, n (%)			
Men	64 (21.8)	36 (24.7)	28 (19)
Women	229 (78.2)	110 (75.3)	119 (81)
Educational level, n (%)			
Tertiary	20 (6.8)	12 (8.2)	8 (5.4)
Secondary	98 (33.4)	44 (30.1)	54 (36.7)
Primary	152 (51.8)	81 (55.5)	71 (48.3)
No education	22 (7.5)	8 (5.5)	14 (9.5)
Marital status, n (%)			
Unmarried	12 (4.1)	5 (3.4)	7 (4.8)
Married	167 (57.2)	90 (62.1)	77 (52.4)
Separated	2 (0.7)	0 (0)	2 (1.4)
Divorced	17 (5.8)	10 (6.9)	7 (4.8)
Widowed	9.4 (32.2)	40 (27.6)	54 (36.7)
Living status, n (%)			
Living alone	101 (34.5)	49 (33.6)	52 (35.4)
Living with family	42 (14.3)	23 (15.8)	19 (12.9)
Living with friends	127 (43.3)	60 (40.1)	67 (45.6)
Living with others	22 (7.4)	14 (9.6)	8 (5.4)
VR[b] experience, n (%)			
Yes	60 (20.5)	18 (12.4)	42 (28.6)
No	232 (79.5)	127 (87.6)	105 (71.4)
Outcome variables			
MoCA[c] score, mean (SD)	21.3 (2.4)	21.1 (2.42)	21.5 (2.4)
FFP[d] score, mean (SD)	1.6 (0.8)	1.6 (0.8)	1.7 (0.8)
TUG[e] score (seconds), median (IQR)	10.9 (3.6)	10.7 (3.1)	11.5 (4.1)
VSTM[f], mean (SD)			
DSF[g] score	7.1 (1.2)	7.1 (1.3)	7.1 (1.1)
DSB[h] score	3.8 (1.3)	3.8 (1.3)	3.8 (1.3)
SCWT[i] score, mean (SD)	–21.4 (9.1)	–22.1 (9.4)	–20.8 (8.7)
TMT[j] score, mean (SD)	1.7 (0.7)	1.7 (0.1)	1.6 (0.1)

[a]CI: chronic illness.
[b]VR: virtual reality.

XSL•FO
RenderX

[c]MoCA: Montreal Cognitive Assessment.
[d]FFP: Fried frailty phenotype.
[e]TUG: Timed Up and Go test.
[f]VSTM: visual short-term memory.
[g]DSF: digit span forward test.
[h]DSB: digit span backward test.
[i]SCWT: Stroop Color and Word Test.
[j]TMT: Trail Making Test.

Numbers Analyzed

As shown in Table 3, the number of participants with missing data on all variables in the intervention group, including both demographic and outcome variables, ranged from 0% to 28.8% (42/146) at baseline (T0) and from 14.4% (21/146) to 30.8% (45/146) at follow-up (T1). The number of participants with missing data for all variables in the control group ranged from 0% to 22.4% (33/147) at baseline (T0) and from 7.5% (11/147) to 27.9% (41/147) at follow-up (T1). The data for all participants who entered the randomized group allocation were analyzed (ie, n=146 in the intervention group and n=147 in the control group).

Table 3. The number of participants with missing data.

Variables	Intervention (n=146), n (%)		Control (n=147), n (%)	
	T0[a]	T1[b]	T0	T1
Demographic variables				
Age (y)	2 (1.4)	—[c]	0 (0)	—
Number of CIs[d]	1 (0.7)	—	0 (0)	—
BMI	1 (0.7)	—	1 (0.7)	—
Gender	0 (0)	—	0 (0)	—
Educational level	1 (0.7)	—	0 (0)	—
Marital status	1 (0.7)	—	0 (0)	—
Living status	1 (0.7)	—	1 (0.7)	—
VR[e] experience	1 (0.7)	—	0 (0)	—
Outcome variables				
MoCA[f] score	1 (0.7)	21 (14.4)	0 (0)	11 (7.5)
FFP[g] score	1 (0.7)	21 (14.4)	0 (0)	11 (7.5)
TUG[h] score	1 (0.7)	21 (14.4)	0 (0)	14 (9.5)
DSF[i] score	1 (0.7)	21 (14.4)	1 (0.7)	13 (8.8)
DSB[j] score	1 (0.7)	21 (14.4)	1 (0.7)	14 (9.5)
SCWT[k] score	42 (28.8)	45 (30.8)	33 (22.4)	41 (27.9)
TMT[l] score	1 (0.7)	21 (14.4)	1 (0.7)	12 (8.2)

[a]T0: baseline.
[b]T1: follow-up.
[c]Demographic data were only collected at baseline.
[d]CI: chronic illness.
[e]VR: virtual reality.
[f]MoCA: Montreal Cognitive Assessment.
[g]FFP: Fried frailty phenotype.
[h]TUG: Timed Up and Go test.
[i]DSF: digit span forward test.
[j]DSB: digit span backward test.
[k]SCWT: Stroop Color and Word Test.
[l]TMT: Trail Making Test.

XSL•FO
RenderX

Outcomes and Estimation

Cognitive function improved from T0 (mean MoCA score 21.12, SE 0.204) to T1 (mean MoCA score 22.10, SE 0.273) and reached the level of significance (within-group $P<.001$) in the intervention group but not in the control group (within-group $P=.33$). The extent of improvement in the intervention group was also significantly larger than that in the control group (group × time $P=.03$), as shown in Table 4.

XSL•FO
RenderX

Table 4. Outcome estimates.

Outcome and group	Values, mean (SE)		MD[a] (95% CI)	*P* value			
	T0[b]	T1[c]		Within group	Between groups	Time	Group × time
Global cognitive function (MoCA[d])							
Intervention	21.12 (0.204)	22.10 (0.273)	0.97 (1.38 to 0.56)	<.001[e]	.89	<.001[e]	.03[e]
Control	21.29 (0.200)	21.54 (0.299)	0.25 (0.73 to –0.23)	.33	__[f]	—	—
Physical frailty (FFP[g])							
Intervention	1.60 (0.060)	1.29 (0.068)	–0.31 (–0.17 to –0.46)	<.001[e]	.006[e]	.003[e]	.01[e]
Control	1.69 (0.064)	1.66 (0.093)	–0.03 (0.15 to –0.21)	.74	—	—	—
Walking speed (TUG[h])							
Intervention	11.25 (0.311)	10.77 (0.242)	–0.48 (0.06 to –1.03)	.08	.02[e]	.78	.13
Control	12.10 (0.454)	12.80 (0.829)	0.71 (2.13 to –0.72)	.33	—	—	—
VSTM[i]: DSF[j]							
Intervention	7.13 (0.107)	7.43 (0.112)	0.30 (0.51 to 0.08)	.006[e]	.44	.001[e]	.46
Control	7.08 (0.096)	7.27 (0.108)	0.19 (0.38 to –0.01)	.04[e]	—	—	—
VSTM: DSB[k]							
Intervention	3.88 (0.104)	3.92 (0.149)	0.04 (0.34 to –0.27)	.82	.80	.48	.74
Control	3.82 (0.109)	3.92 (0.115)	0.10 (0.33 to –0.13)	.40	—	—	—
ICI[l] (SCWT[m])							
Intervention	–22.19 (0.897)	–21.33 (0.953)	0.86 (2.59 to –0.88)	.33	.34	.78	.31
Control	–20.45 (0.814)	–20.94 (0.981)	–0.49 (1.41 to –2.39)	.62	—	—	—
Executive function (TMT[n])							
Intervention	1.68 (0.054)	1.51 (0.048)	–0.17 (–0.04 to –0.29)	.01[e]	.47	.12	.07
Control	1.63 (0.056)	1.65 (0.063)	0.01 (0.16 to –0.13)	.85	—	—	—

[a]MD: mean difference.
[b]T0: baseline.
[c]T1: follow-up.
[d]MoCA: Montreal Cognitive Assessment.
[e]$P<.05$.
[f]Not applicable.
[g]FFP: Fried frailty phenotype.
[h]TUG: Timed Up and Go test.
[i]VSTM: visual short-term memory.
[j]DSF: digit span forward test.
[k]DSB: digit span backward test.

XSL•FO
RenderX

[l]ICI: inhibition of cognitive interference.

[m]SCWT: Stroop Color and Word Test.

[n]TMT: Trail Making Test.

Physical frailty decreased from T0 (mean FFP score 1.60, SE 0.060) to T1 (mean FFP score 1.29, SE 0.068) and reached the level of significance (within-group $P<.001$) in the intervention group but not in the control group (within-group $P=.74$), and the extent of reduction in the intervention group was also significantly larger than that in the control group (group × time $P=.01$), as shown in Table 4.

Executive function improved from T0 (mean TMT score 1.68, SE 0.054) to T1 (mean TMT score 1.51, SE 0.048) and reached the level of significance (within-group $P=.01$) in the intervention group but not in the control group (within-group $P=.85$). The extent of improvement in the intervention group was not but almost significantly larger than that in the control group (group × time $P=.07$), as shown in Table 4.

Walking speed, visual short-term memory, and inhibition of cognitive inference did not show more improvement in the intervention group compared to the control group, as shown in Table 4.

Harms

In all the VRMCT sessions (n=4874) conducted on participants in both intervention and waitlist control groups (including some preintervention trial sessions on some participants), the VR sickness symptoms reported by participants using the VRSQ ranged from 0.7% on symptoms of difficulty focusing and headache to 3% on the symptom of vertigo. Among all the reported VR sickness symptoms, the most reported frequency was “less than half” (ie, 82%-93%). Regarding the most reported VR sickness symptom (ie, vertigo), only participants in 1% of the total sessions reported to have experienced it “more than half” during the VR training. No participants reported severe untoward effects requiring medical treatment after the completion of the training. However, 0.5% of the total training sessions were reported to have technical issues related to the training system.

Discussion

Principal Findings

This is one of the very few studies that provides direct evidence using a randomized controlled trial that a VRMCT intervention using a gamified and light-intensity approach is effective in promoting global cognitive function, marginally promoting executive function, and reducing frailty specifically in the at-risk population of older people with cognitive frailty, who also adhered well to the intervention.

Previous studies have shown that motor-cognitive training has positive effects on cognitive function in cognitively healthy older people [48], on executive function in older people [49], and on dual-task walking ability in older people with MCI [50]. It also brings cognitive benefits probably because it yields moderations at the neuronal level (eg, the dopaminergic system showing decreases in mean diffusivity in the frontal and subcortical brain areas in functional magnetic resonance imaging) [51], although the improvement may not be directly due to the moderation of the Aβ metabolism [52]. The findings of this study also agree with previous findings that motor-cognitive training promotes cognitive function and was also effective in a specific group of older people with cognitive frailty [25,49]. This study recommends the use of VRMCT to promote cognitive function and reduce physical frailty for older people with cognitive frailty. However, this study cannot conclude that the improvement in clinical outcomes was caused by or correlated with neuroplastic changes, as proposed in the literature [10]. Future studies should also examine whether this intervention can also yield beneficial moderation at the neuronal level.

The implementation methods and dose are also important. Motor-cognitive training is a general term without a clear gauge of the portion of time spent on simultaneous training throughout the training period. The conventional and effective simultaneous motor-cognitive training protocols usually use 100% motor-cognitive training time throughout the training period [53]. However, recently, some types of gamified motor-cognitive training have adopted a more amusing approach with a lower portion of simultaneous training time and have also demonstrated effectiveness in improving executive function and dual-task performance [24]. The intensity of the training in terms of training hours per session and training period is usually high (eg, 60-minute sessions for 12 weeks), which may not be easily tolerated by the vulnerable group of older people with cognitive frailty [24,48]. This intervention protocol adopted a gamified and light-intensity approach (ie, 30 minutes [actual VRMCT time] per session twice per week for 8 weeks) and also showed effectiveness in promoting cognitive function.

This piece of evidence opens up the possibility of a less serious and less intensive training model implementable in the community to promote the cognitive function of older people. This study recommends that gamified and light-intensity training interventions should be promoted to community-dwelling older people with cognitive frailty in community centers for cognitive promotion purposes. This intervention protocol also adopted training content similar to activities of daily living using VR technology, which aimed at lowering the learning barrier and improving the ecological validity of the training to be translated to improving the daily activity of the participants. We believe that performing these functional tasks with internalized real-time feedback through VR may have a more significant impact on various executive functions, leading to a significantly improved global cognitive function in the experimental group. In addition, the enjoyment and attractiveness of the VR characteristics likely increased motivation and contributed to more extensive training effects on executive function in the experimental group compared to the control group.

Recent evidence has shown that physical frailty prevalence is reduced in older people’s cognitive frailty after the completion of motor-cognitive training [54]. However, the available evidence shows that spontaneous revision to normality from

XSL•FO
RenderX

frailty is possible [55]. This study used a more rigorous approach of a randomized controlled trial that showed that the improvement in frailty was significantly greater in the intervention group compared with the control group and where the risk of spontaneous revision to normality was comparable between groups. It provided greater confidence that motor-cognitive training is effective in reducing physical frailty. Although plenty of evidence shows that motor-cognitive training is more effective in promoting physical functions (eg, gait speed, single-task walking function, and balance) than single-task training [11,53], the findings of this study still could not prove whether motor-cognitive training is more effective than single-task training in reducing physical frailty. Future studies should examine whether motor-cognitive training is more effective than single-task training in reducing physical frailty.

The adherence rate of the intervention in this study is also appealing, with 81.6% of participants attending >81% (13/16) of the total number of sessions. A systematic review showed that the average adherence rate to community-based group exercise interventions for older people is only 69.1% [56]. This finding also agrees with the literature on the fact that technology-enhanced (ie, using VR) exercise programs can improve adherence to exercise interventions in older people [57]. Future studies should also compare which technology-enhanced interventions may lead to a better adherence rate.

Despite the rigorous design of this study with favorable outcomes, there are several limitations. First, there were 14.4% (21/146) and 7.5% (11/147) of participants lost to follow-up in the intervention and control groups, respectively. Although the research team executed many measures to reduce the risk of loss to follow-up (eg, sending reminders and providing certificates of completion to those who completed the T1 data collection), it was not avoidable because some centers were closed sporadically in response to the unpredictably varying severity of COVID-19 during the study period. Data collection could not be carried on because some centers were closed for a prolonged time because of COVID-19. Fortunately, the amount of loss to follow-up was only 10.6%, which is <20%, so it is unlikely to pose a serious threat [58]. Second, there was a relatively large group of participants who refused to participate in the Stroop tests (ie, 33/147, 22.4% to 45/146, 30.8%), as shown in Table 3, because they either did not understand the instructions or could not finish the Stroop test within the scheduled period of assessment. Some participants spent too long completing other tests. The Stroop test was arranged as the last task so that it could not be completed by some participants. Therefore, the insignificant results could be caused by unintentional selection bias. Third, this study only evaluated the effects of the intervention on short-term outcomes. Future studies should also evaluate its longer-term effects after the improvement in cognitive function, such as social participation and life-space mobility [31]. Fourth, although a VR element was added to this intervention, because of the trial design (ie, the control group received only usual care), the efficacy of the VR element contributing to favorable clinical outcomes cannot be concluded in this study. Future studies should examine whether VR-integrated motor-cognitive training is more effective than non–VR-integrated motor-cognitive training. Fifth, the generalization of the findings depends on the access to the training prototype. To date, VRMCT is still a prototype, and its effects cannot benefit the target population immediately. Commercialization of the prototype with cost control measures should be implemented to ensure that the effects of VRMCT can be accessible and affordable to benefit the population in an ethical manner, such as regarding equity. Sixth, the contents of VRMCT were all culturally adapted to Hong Kong Chinese culture (eg, language, the virtual environment, and the activities familiar to older people). This culture-specific prototype limits the generalizability and the direct use of VRMCT. Cultural modification is needed to generalize the effects of VRMCT to other cultures and geographic locations. Seventh, only short-term effects of VRMCT (ie, the week immediately after the completion of the intervention) were evaluated. Longer-term effects should be examined in future studies. Eighth, the clinical significance of our results may be minimal given the MoCA increase of only 0.93 points. Therefore, the interpretation of the findings in this study should be approached with caution. In future research, we will explore the impact of the observed MoCA score improvement in the context of daily clinical practice and cognitive function. Specifically, we will assess whether this degree of improvement translates into meaningful benefits for patients, such as enhanced cognitive abilities or improved quality of life. Finally, an unexpected finding showed that the forward digital span scores significantly improved from T0 to T1 in both the intervention and control groups, as shown in Table 4. It might have been caused by the practice effect. Evidence has shown that digit span tests and many commonly used attentional tests do not always have optimal test-retest stability because the practice effect often causes retest scores to be higher [59,60]. Future studies can adopt a composite measure of several tasks or use alternate forms of the measures to yield a more stable result [59,60].

Conclusions

To conclude, VRMCT is effective in promoting cognitive functions and reducing physical frailty and is also well tolerated and accepted by older people with cognitive frailty, as evidenced by its high attendance rate and negligible VR sickness symptoms. Further studies should examine the efficacy of the intervention components (eg, VR vs non-VR and dual vs single task) on health outcomes, the effect of using technology on intervention adherence, and the long-term effects of the intervention on older people with cognitive frailty at the level of daily living.

XSL•FO
RenderX

Acknowledgments

This project was funded by the Innovation and Technology Fund for Better Living, Innovation, Technology, and Industry Bureau, Hong Kong (project reference ITB/FBL/4015/19/P), and a matching fund provided by the School of Nursing, the Hong Kong Polytechnic University.

Authors' Contributions

All authors contributed to the development of the study and the intervention. OSKS was responsible for the coordination of data collection at the older adult community centers. RYCK was responsible for the data collection, intervention implementation monitoring, and research team training. RYCK drafted the manuscript. All authors read, critically reviewed, and revised the manuscript. All authors reviewed and agreed on the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video of the week 2 theme: finding a bus stop.
[MP4 File (MP4 Video), 28567 KB - jmir_v26i1e57809_app1.mp4]

Multimedia Appendix 2
Video of the week 5 theme: grocery shopping.
[MP4 File (MP4 Video), 65348 KB - jmir_v26i1e57809_app2.mp4]

Multimedia Appendix 3
The CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 67 KB - jmir_v26i1e57809_app3.pdf]

References

1. Ruan Q, Yu Z, Chen M, Bao Z, Li J, He W. Cognitive frailty, a novel target for the prevention of elderly dependency. Ageing Res Rev 2015 Mar;20:1-10. [doi: 10.1016/j.arr.2014.12.004] [Medline: 25555677]
2. Sugimoto T, Arai H, Sakurai T. An update on cognitive frailty: its definition, impact, associated factors and underlying mechanisms, and interventions. Geriatr Gerontol Int 2022 Feb;22(2):99-109 [FREE Full text] [doi: 10.1111/ggi.14322] [Medline: 34882939]
3. Qiu Y, Li G, Wang X, Zheng L, Wang C, Wang C, et al. Prevalence of cognitive frailty among community-dwelling older adults: a systematic review and meta-analysis. Int J Nurs Stud 2022 Jan;125:104112 [FREE Full text] [doi: 10.1016/j.ijnurstu.2021.104112] [Medline: 34758429]
4. Zhang T, Ren Y, Shen P, Jiang S, Yang Y, Wang Y, et al. Prevalence and associated risk factors of cognitive frailty: a systematic review and meta-analysis. Front Aging Neurosci 2022 Jan 28;13:755926 [FREE Full text] [doi: 10.3389/fnagi.2021.755926] [Medline: 35153714]
5. Bu Z, Huang A, Xue M, Li Q, Bai Y, Xu G. Cognitive frailty as a predictor of adverse outcomes among older adults: a systematic review and meta-analysis. Brain Behav 2021 Jan;11(1):e01926 [FREE Full text] [doi: 10.1002/brb3.1926] [Medline: 33159430]
6. Kwan RY, Leung AY, Yee A, Lau LT, Xu XY, Dai DL. Cognitive frailty and its association with nutrition and depression in community-dwelling older people. J Nutr Health Aging 2019;23(10):943-948 [FREE Full text] [doi: 10.1007/s12603-019-1258-y] [Medline: 31781723]
7. Tam AC, Chan AW, Cheung DS, Ho LY, Tang AS, Christensen M, et al. The effects of interventions to enhance cognitive and physical functions in older people with cognitive frailty: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2022 Aug 24;19(1):19 [FREE Full text] [doi: 10.1186/s11556-022-00299-9] [Medline: 36002799]
8. Kwan RY, Lee D, Lee PH, Tse M, Cheung DS, Thiamwong L, et al. Effects of an mHealth brisk walking intervention on increasing physical activity in older people with cognitive frailty: pilot randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e16596 [FREE Full text] [doi: 10.2196/16596] [Medline: 32735218]
9. Lauenroth A, Ioannidis AE, Teichmann B. Influence of combined physical and cognitive training on cognition: a systematic review. BMC Geriatr 2016 Jul 18;16:141 [FREE Full text] [doi: 10.1186/s12877-016-0315-1] [Medline: 27431673]
10. Herold F, Hamacher D, Schega L, Müller NG. Thinking while moving or moving while thinking - concepts of motor-cognitive training for cognitive performance enhancement. Front Aging Neurosci 2018 Aug 6;10:228 [FREE Full text] [doi: 10.3389/fnagi.2018.00228] [Medline: 30127732]

XSL•FO
RenderX

11. Huber SK, Knols RH, Arnet P, de Bruin ED. Motor-cognitive intervention concepts can improve gait in chronic stroke, but their effect on cognitive functions is unclear: a systematic review with meta-analyses. Neurosci Biobehav Rev 2022 Jan;132:818-837 [FREE Full text] [doi: 10.1016/j.neubiorev.2021.11.013] [Medline: 34815131]
12. Johansson H, Folkerts AK, Hammarström I, Kalbe E, Leavy B. Effects of motor-cognitive training on dual-task performance in people with Parkinson's disease: a systematic review and meta-analysis. J Neurol 2023 Jun;270(6):2890-2907 [FREE Full text] [doi: 10.1007/s00415-023-11610-8] [Medline: 36820916]
13. Xiao Y, Yang T, Shang H. The impact of motor-cognitive dual-task training on physical and cognitive functions in Parkinson's disease. Brain Sci 2023 Mar 03;13(3):437 [FREE Full text] [doi: 10.3390/brainsci13030437] [Medline: 36979247]
14. Jung J, Kim YS, Lee S. The motor-cognitive training on cognition and physical performance in the older adults with mild cognition impairment : a literature review. Phys Ther Rehabil Sci 2021 Dec 30;10:493-502. [doi: 10.14474/ptrs.2021.10.4.493]
15. Sale A, Noale M, Cintoli S, Tognoni G, Braschi C, Berardi N, et al. Long-term beneficial impact of the randomised trial 'Train the Brain', a motor/cognitive intervention in mild cognitive impairment people: effects at the 14-month follow-up. Age Ageing 2023 May 01;52(5):afad067. [doi: 10.1093/ageing/afad067] [Medline: 37167616]
16. Tao M, Liu H, Cheng J, Yu C, Zhao L. Motor-cognitive interventions may effectively improve cognitive function in older adults with mild cognitive impairment: a randomized controlled trial. Behav Sci (Basel) 2023 Sep 04;13(9):737 [FREE Full text] [doi: 10.3390/bs13090737] [Medline: 37754015]
17. Snoswell AJ, Snoswell CL. Immersive virtual reality in health care: systematic review of technology and disease states. JMIR Biomed Eng 2019 Sep 26;4(1):e15025. [doi: 10.2196/15025]
18. Dermody G, Whitehead L, Wilson G, Glass C. The role of virtual reality in improving health outcomes for community-dwelling older adults: systematic review. J Med Internet Res 2020 Jun 01;22(6):e17331 [FREE Full text] [doi: 10.2196/17331] [Medline: 32478662]
19. Skurla MD, Rahman AT, Salcone S, Mathias L, Shah B, Forester BP, et al. Virtual reality and mental health in older adults: a systematic review. Int Psychogeriatr 2022 Feb;34(2):143-155. [doi: 10.1017/S104161022100017X] [Medline: 33757619]
20. Kwan RY, Ng F, Lam LC, Yung RC, Sin OS, Chan S. The effects of therapeutic virtual reality experience to promote mental well-being in older people living with physical disabilities in long-term care facilities. Trials 2023 Aug 26;24(1):558 [FREE Full text] [doi: 10.1186/s13063-023-07592-7] [Medline: 37633916]
21. Saunders NL, Summers MJ. Attention and working memory deficits in mild cognitive impairment. J Clin Exp Neuropsychol 2010 Apr;32(4):350-357. [doi: 10.1080/13803390903042379] [Medline: 19787522]
22. van der Wardt V, Hancox J, Gondek D, Logan P, Das Nair R, Pollock K, et al. Adherence support strategies for exercise interventions in people with mild cognitive impairment and dementia: a systematic review. Prev Med Rep 2017 May 18;7:38-45 [FREE Full text] [doi: 10.1016/j.pmedr.2017.05.007] [Medline: 28593121]
23. Yu D, Li X, Lai FH. The effect of virtual reality on executive function in older adults with mild cognitive impairment: a systematic review and meta-analysis. Aging Ment Health 2023 Apr;27(4):663-673. [doi: 10.1080/13607863.2022.2076202] [Medline: 35635486]
24. Wang RY, Huang YC, Zhou JH, Cheng SJ, Yang YR. Effects of exergame-based dual-task training on executive function and dual-task performance in community-dwelling older people: a randomized-controlled trial. Games Health J 2021 Oct;10(5):347-354. [doi: 10.1089/g4h.2021.0057] [Medline: 34491113]
25. Kwan RY, Liu JY, Fong KN, Qin J, Leung PK, Sin OS, et al. Feasibility and effects of virtual reality motor-cognitive training in community-dwelling older people with cognitive frailty: pilot randomized controlled trial. JMIR Serious Games 2021 Aug 06;9(3):e28400 [FREE Full text] [doi: 10.2196/28400] [Medline: 34383662]
26. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
27. District elderly community centre. Social Welfare Department, Government of the Hong Kong Special Administrative Region of the People's Republic of China. URL: https://www.swd.gov.hk/en/pubsvc/elderly/cat_commsupp/elderly_centres/ [accessed 2023-09-11]
28. Kelaiditi E, Cesari M, Canevelli M, van Kan GA, Ousset PJ, Gillette-Guyonnet S, et al. Cognitive frailty: rational and definition from an (I.A.N.A./I.A.G.G.) international consensus group. J Nutr Health Aging 2013 Sep;17(9):726-734 [FREE Full text] [doi: 10.1007/s12603-013-0367-2] [Medline: 24154642]
29. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
30. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar 01;56(3):M146-M157. [doi: 10.1093/gerona/56.3.M146]
31. Kwan RY, Cheung DS, Lo SK, Ho LY, Katigbak C, Chao YY, et al. Frailty and its association with the Mediterranean diet, life-space, and social participation in community-dwelling older people. Geriatr Nurs 2019;40(3):320-326. [doi: 10.1016/j.gerinurse.2018.12.011] [Medline: 30777380]
32. Yeung PY, Wong LL, Chan CC, Leung JL, Yung CY. A validation study of the Hong Kong version of Montreal Cognitive Assessment (HK-MoCA) in Chinese older adults in Hong Kong. Hong Kong Med J 2014 Dec 15;20(6):504-510 [FREE Full text] [doi: 10.12809/hkmj144219] [Medline: 25125421]

XSL•FO
RenderX

33. Rosanna Chau M, Chan SP, Wong YW, Lau MY. Reliability and validity of the Modified Functional Ambulation Classification in patients with hip fracture. Hong Kong Physiother J 2013 Jun;31(1):41-44 [FREE Full text] [doi: 10.1016/j.hkpj.2013.01.041]
34. Saredakis D, Szpak A, Birckhead B, Keage HA, Rizzo A, Loetscher T. Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. Front Hum Neurosci 2020;14:96 [FREE Full text] [doi: 10.3389/fnhum.2020.00096] [Medline: 32300295]
35. Høeg ER, Povlsen TM, Bruun-Pedersen JR, Lange B, Nilsson NC, Haugaard KB, et al. System immersion in virtual reality-based rehabilitation of motor function in older adults: a systematic review and meta-analysis. Front Virtual Real 2021 Apr 12;2. [doi: 10.3389/frvir.2021.647993]
36. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383 [FREE Full text] [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
37. Stewart S, O'Riley A, Edelstein B, Gould C. A preliminary comparison of three cognitive screening instruments in long term care: the MMSE, SLUMS, and MoCA. Clin Gerontol 2012 Jan;35(1):57-75. [doi: 10.1080/07317115.2011.626515]
38. Podsiadlo D, Richardson S. The timed "Up and Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991 Feb;39(2):142-148 [FREE Full text] [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]
39. Ryan JJ, Lopez SJ. Wechsler Adult Intelligence Scale-III. In: Dorfman WI, Hersen M, editors. Understanding Psychological Assessment. Boston, MA: Springer; 2001.
40. Tam CW, Lam LC. Neuropsychological profiles of elderly Chinese people with amnestic and multiple-domain mild cognitive impairment. Hong Kong J Psychiatry 2006;16:21-26 [FREE Full text]
41. Schroeder RW, Twumasi-Ankrah P, Baade LE, Marshall PS. Reliable Digit Span: a systematic review and cross-validation study. Assessment 2012 Mar 06;19(1):21-30. [doi: 10.1177/1073191111428764] [Medline: 22156721]
42. Scarpina F, Tagini S. The stroop color and word test. Front Psychol 2017 Apr 12;8:557 [FREE Full text] [doi: 10.3389/fpsyg.2017.00557] [Medline: 28446889]
43. Arbuthnott K, Frank J. Trail making test, part B as a measure of executive control: validation using a set-switching paradigm. J Clin Exp Neuropsychol 2010 Aug 09;22(4):518-528. [doi: 10.1076/1380-3395(200008)22:4;1-0;ft518]
44. Kim HK, Park J, Choi Y, Choe M. Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl Ergon 2018 May;69:66-73 [FREE Full text] [doi: 10.1016/j.apergo.2017.12.016] [Medline: 29477332]
45. Kreidler SM, Muller KE, Grunwald GK, Ringham BM, Coker-Dukowitz Z, Sakhadeo UR, et al. GLIMMPSE: online power computation for linear models with and without a baseline covariate. J Stat Softw 2013 Sep;54(10):i10 [FREE Full text] [doi: 10.18637/jss.v054.i10] [Medline: 24403868]
46. Kim HY. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restor Dent Endod 2013 Feb;38(1):52-54 [FREE Full text] [doi: 10.5395/rde.2013.38.1.52] [Medline: 23495371]
47. Salazar A, Ojeda B, Dueñas M, Fernández F, Failde I. Simple generalized estimating equations (GEEs) and weighted generalized estimating equations (WGEEs) in longitudinal studies with dropouts: guidelines and implementation in R. Stat Med 2016 Aug 30;35(19):3424-3448 [FREE Full text] [doi: 10.1002/sim.6947] [Medline: 27059703]
48. Gallou-Guyot M, Mandigout S, Bherer L, Perrochon A. Effects of exergames and cognitive-motor dual-task training on cognitive, physical and dual-task functions in cognitively healthy older adults: an overview. Ageing Res Rev 2020 Nov;63:101135 [FREE Full text] [doi: 10.1016/j.arr.2020.101135] [Medline: 32768594]
49. Wollesen B, Wildbredt A, van Schooten KS, Lim ML, Delbaere K. The effects of cognitive-motor training interventions on executive functions in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020 Jul 02;17(1):9 [FREE Full text] [doi: 10.1186/s11556-020-00240-y] [Medline: 32636957]
50. Kuo HT, Yeh NC, Yang YR, Hsu WC, Liao YY, Wang RY. Effects of different dual task training on dual task walking and responding brain activation in older adults with mild cognitive impairment. Sci Rep 2022 May 19;12(1):8490 [FREE Full text] [doi: 10.1038/s41598-022-11489-x] [Medline: 35589771]
51. Takeuchi H, Magistro D, Kotozaki Y, Motoki K, Nejad KK, Nouchi R, et al. Effects of simultaneously performed dual-task training with aerobic exercise and working memory training on cognitive functions and neural systems in the elderly. Neural Plasticity 2020 Jun 05;2020:1-17. [doi: 10.1155/2020/3859824]
52. Yokoyama H, Okazaki K, Imai D, Yamashina Y, Takeda R, Naghavi N, et al. The effect of cognitive-motor dual-task training on cognitive function and plasma amyloid β peptide 42/40 ratio in healthy elderly persons: a randomized controlled trial. BMC Geriatr 2015 May 28;15(1):60 [FREE Full text] [doi: 10.1186/s12877-015-0058-4] [Medline: 26018225]
53. He Y, Yang L, Zhou J, Yao L, Pang MY. Dual-task training effects on motor and cognitive functional abilities in individuals with stroke: a systematic review. Clin Rehabil 2018 Jul 23;32(7):865-877. [doi: 10.1177/0269215518758482] [Medline: 29473480]
54. Merchant RA, Chan YH, Hui RJ, Tsoi CT, Kwek SC, Tan WM, et al. Motoric cognitive risk syndrome, physio-cognitive decline syndrome, cognitive frailty and reversibility with dual-task exercise. Exp Gerontol 2021 Jul 15;150:111362 [FREE Full text] [doi: 10.1016/j.exger.2021.111362] [Medline: 33887381]

XSL•FO
RenderX

55. Canevelli M, Bruno G, Remiddi F, Vico C, Lacorte E, Vanacore N, et al. Spontaneous reversion of clinical conditions measuring the risk profile of the individual: from frailty to mild cognitive impairment. Front Med (Lausanne) 2017 Oct 30;4:184 [FREE Full text] [doi: 10.3389/fmed.2017.00184] [Medline: 29164119]
56. Farrance C, Tsofliou F, Clark C. Adherence to community based group exercise interventions for older people: a mixed-methods systematic review. Prev Med 2016 Jun;87:155-166 [FREE Full text] [doi: 10.1016/j.ypmed.2016.02.037] [Medline: 26921655]
57. Valenzuela T, Okubo Y, Woodbury A, Lord SR, Delbaere K. Adherence to technology-based exercise programs in older adults: a systematic review. J Geriatr Phys Ther 2018;41(1):49-61. [doi: 10.1519/JPT.0000000000000095] [Medline: 27362526]
58. Dettori JR. Loss to follow-up. Evid Based Spine Care J 2011 Feb 10;2(1):7-10 [FREE Full text] [doi: 10.1055/s-0030-1267080] [Medline: 22956930]
59. Waters GS, Caplan D. The reliability and stability of verbal working memory measures. Behav Res Methods Instrum Comput 2003 Nov;35(4):550-564. [doi: 10.3758/bf03195534] [Medline: 14748499]
60. Lemay S, Bédard MA, Rouleau I, Tremblay PL. Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. Clin Neuropsychol 2004 May;18(2):284-302. [doi: 10.1080/13854040490501718] [Medline: 15587675]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FFP: Fried frailty phenotype
MCI: mild cognitive impairment
MoCA: Montreal Cognitive Assessment
TMT: Trail Making Test
VR: virtual reality
VRMCT: virtual reality motor-cognitive training
VRSQ: Virtual Reality Sickness Questionnaire

Edited by B Puladi; submitted 29.02.24; peer-reviewed by D Levine, A Bezabih, H Kerminen; comments to author 11.04.24; revised version received 05.06.24; accepted 14.07.24; published 11.09.24.

Please cite as:
Kwan RYC, Liu J, Sin OSK, Fong KNK, Qin J, Wong JCY, Lai C
Effects of Virtual Reality Motor-Cognitive Training for Older People With Cognitive Frailty: Multicentered Randomized Controlled Trial
J Med Internet Res 2024;26:e57809
URL: https://www.jmir.org/2024/1/e57809
doi:10.2196/57809
PMID:

©Rick Yiu Cho Kwan, Justina Liu, Olive Suk Kan Sin, Kenneth N K Fong, Jing Qin, Joe Chi Yin Wong, Claudia Lai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Game-Based Assessment of Peripheral Neuropathy Combining Sensor-Equipped Insoles, Video Games, and AI: Proof-of-Concept Study

Antao Ming[1], PhD; Vera Clemens[1], MSc; Elisabeth Lorek[1]; Janina Wall[1]; Ahmad Alhajjar[1]; Imke Galazky[2], MD; Anne-Katrin Baum[2]; Yang Li[3], PhD; Meng Li[4], PhD; Sebastian Stober[5], PhD; Nils David Mertens[1], MD; Peter Rene Mertens[1], MD

[1]University Clinic for Nephrology and Hypertension, Diabetology and Endocrinology, Otto von Guericke University Magdeburg, Magdeburg, Germany
[2]University Clinic for Neurology, Otto von Guericke University Magdeburg, Magdeburg, Germany
[3]Pure-systems, Magdeburg, Germany
[4]Affective Neuroscience Lab, Friedrich Schiller University Jena, Jena, Germany
[5]Artificial Intelligence Lab, Otto von Guericke University Magdeburg, Magdeburg, Germany

Corresponding Author:
Peter Rene Mertens, MD
University Clinic for Nephrology and Hypertension, Diabetology and Endocrinology
Otto von Guericke University Magdeburg
Leipziger Straße 44
Magdeburg, 39120
Germany
Phone: 49 391 6713236
Fax: 49 391 6715440
Email: peter.mertens@med.ovgu.de

Abstract

Background: Detecting peripheral neuropathy (PNP) is crucial in preventing complications such as foot ulceration. Clinical examinations for PNP are infrequently provided to patients at high risk due to restrictions on facilities, care providers, or time. A gamified health assessment approach combining wearable sensors holds the potential to address these challenges and provide individuals with instantaneous feedback on their health status.

Objective: We aimed to develop and evaluate an application that assesses PNP through video games controlled by pressure sensor–equipped insoles.

Methods: In the proof-of-concept exploratory cohort study, a complete game-based framework that allowed the study participant to play 4 video games solely by modulating plantar pressure values was established in an outpatient clinic setting. Foot plantar pressures were measured by the sensor-equipped insole and transferred via Bluetooth to an Android tablet for game control in real time. Game results and sensor data were delivered to the study server for visualization and analysis. Each session lasted about 15 minutes. In total, 299 patients with diabetes mellitus and 30 with metabolic syndrome were tested using the game application. Patients' game performance was initially assessed by hypothesis-driven key capabilities that consisted of reaction time, sensation, skillfulness, balance, endurance, and muscle strength. Subsequently, specific game features were extracted from gaming data sets and compared with nerve conduction study findings, neuropathy symptoms, or disability scores. Multiple machine learning algorithms were applied to 70% (n=122) of acquired data to train predictive models for PNP, while the remaining data were held out for final model evaluation.

Results: Overall, clinically evident PNP was present in 247 of 329 (75.1%) participants, with 88 (26.7%) individuals showing asymmetric nerve deficits. In a subcohort (n=37) undergoing nerve conduction study as the gold standard, sensory and motor nerve conduction velocities and nerve amplitudes in lower extremities significantly correlated with 79 game features ($|R|>0.4$, highest R value +0.65; $P<.001$; adjusted $R^2=0.36$). Within another subcohort (n=173) with normal cognition and matched covariates (age, sex, BMI, etc), hypothesis-driven key capabilities and specific game features were significantly correlated with the presence of PNP. Predictive models using selected game features achieved 76.1% (left) and 81.7% (right foot) accuracy for PNP detection.

XSL•FO
RenderX

Multiclass models yielded an area under the receiver operating characteristic curve of 0.76 (left foot) and 0.72 (right foot) for assessing nerve damage patterns (small, large, or mixed nerve fiber damage).

Conclusions: The game-based application presents a promising avenue for PNP screening and classification. Evaluation in expanded cohorts may iteratively optimize artificial intelligence model efficacy. The integration of engaging motivational elements and automated data interpretation will support acceptance as a telemedical application.

(J Med Internet Res 2024;26:e52323) doi:10.2196/52323

KEYWORDS

diabetes mellitus; metabolic syndrome; peripheral neuropathy; sensor-equipped insoles; video games; machine learning; feature extraction

Introduction

For primary care clinicians, it is a challenge to diagnose peripheral neuropathy (PNP) due to its diverse forms and presentations. Especially, in patients with diabetes mellitus or metabolic syndrome (MetS), the onset is insidious, and most affected individuals are not aware of distal, often symmetric PNP [1,2]. Further common causes consist of excessive and chronic alcohol use, nerve injury, toxin exposure, nutritional deficiencies, chemotherapy, and genetic disorders [3].

According to the International Diabetes Federation, more than 435 million people worldwide are diagnosed with diabetes, which is expected to rise to 693 million in 2045 [4]. A total of 25% to 50% of patients with diabetes are impacted by PNP, and the severity varies according to age, duration of diabetes, and level of glucose control [3,5,6]. PNP significantly impacts the mobility of patients and is associated with disturbed gait and coordination, increased likelihood of falls, higher incidence rates of diabetic foot syndrome, and frail mental health [1]. Peripheral nerves encompass large nerve fibers (Aα- and Aβ-fibers; vibratory and pressure sensation) and small nerve fibers (unmyelinated C- and Aδ-fibers; pain and warm or cold sensation) [7,8]. Besides diminished sensation, PNP may also cause involuntary symptoms of discomfort and pain. Complications such as tissue damage, infections, and ultimately minor and major foot ulcerations are estimated to affect 19% to 34% of individuals with diabetes during their lifetime, especially with delayed diagnosis of PNP or inadequate preventive measures [9]. Notably, every fifth moderate to severe diabetes-related infection will prompt lower extremity amputation [9,10]. On the other hand, 4 of 5 amputations may be prevented by adequate podiatry care [11].

Due to their invasive nature, nerve conduction studies (NCSs) and skin biopsies are not applicable as screening tools [12]. According to the position statement published already 15 years ago by the American Diabetes Association, “assessment should include a careful history and either temperature or pinprick sensation [...] and vibration sensation [...]” [5,11]. National Institute for Health and Care Excellence [13] recommends vibration perception testing using a 128-Hz tuning fork together with a 10-g (Semmes-Weinstein) monofilament for the screening of PNP. However, neuropathy testing is infrequently provided to patients at high risk, possibly due to potential barriers such as the unavailability of health care (restrictions on facilities, care providers, or time), patient-related factors (absence of pain sensation in feet, lack of understanding of diabetes, and its possible complication), and differences in health care system [14]. The implementation of clinical testing for PNP by monofilament and vibration sensation tests is still not optimal. First, numbers of testing are as low as 35% in high-risk cohorts [15]. Second, these tests identify diabetic peripheral neuropathy at a late, irreversible stage [16,17]. An abnormal monofilament test is associated with a high 3-year relative risk of foot ulceration or lower limb amputation, reaching 15% (95% CI 9.0-26.0) [18]. Although small fibers undergo early and progressive injury in diabetes, their assessment is not included in annual diabetic foot screening programs [19,20]. Some innovations as point-of-care devices have been brought forward; however, they are not established as routines, such as DPN Check (large fibers), NeuroQuick, NeuroPad, Corneal Confocal Microscopy, and Sudoscan (small fibers) [1].

All these aspects urge for the optimization of PNP diagnosis, possibly even prevention and management. Up to 50% of affected individuals remain asymptomatic, while the remainder develop numbness, tingling, pain, or weakness [5,21]. Symptoms commonly originate distally (ie, at the tips of the toes) and spread proximally with a symmetric distribution. Unilateral or asymmetric forms of PNP (eg, diabetic radiculoplexus neuropathy) have also been reported [22]. Further investigation of PNP screening techniques needs to address these challenges.

Recently, using gamification in health assessments combined with wearable sensors has garnered great attention [23,24]. The integration of game-like elements can enhance user engagement, leading to heightened motivation and increased adherence to health goals. Moreover, the continuous data acquisition capabilities of wearables facilitate real-time health monitoring, providing users with instantaneous feedback and enabling more precise trend analysis [25,26]. This synergistic approach holds the potential to instigate positive behavioral changes, ensuring that individuals are not only informed about their health status but are also incentivized to maintain or improve it. Therefore, our proof-of-concept study was designed to identify patients at risk for PNP in a playful approach and further enable timely intervention to prevent complications [27].

The primary objectives of this study were to develop and evaluate a playful assessment device for peripheral nerve function through video-based games that are coordinated through the feet with pressure sensor–equipped insoles.

Specifically, we (1) designed a system with real-time transfer of pressure recordings to a tablet application for gaming control and a central database for analysis. (2) Video-based games were

XSL•FO
RenderX

designed with varying difficulties. (3) Feature extraction methodologies were set up to explore data analysis from a pilot study. Herein, the game-based data set was correlated with NCS and clinical findings. Two ways were followed: predefined hypothesis-driven parameter analyses such as reaction time, sensation, skillfulness, endurance, balance, and muscle strength as well as feature extraction. (4) Optimization of the data analyses was performed by training and validation data sets to yield classification models.

Methods

Ethical Considerations

The study was conducted following approval of the study protocol by the local ethical committee of the Otto von Guericke University Magdeburg (approval 28/17; March 13, 2017). All study participants provided written informed consent upon a detailed explanation of the study protocol, test procedure, and data policy. No compensation was provided for participants except for travel costs. Information regarding study participant identities was removed. The pseudonymized study data were used for analyses. The figures provided in this work are deidentified.

Study Design and Recruitment Procedures

In this proof-of-concept exploratory cohort study, the game platform development and clinical study were carried out at the University Clinic for Nephrology and Hypertension, Diabetes and Endocrinology at the Otto von Guericke University Magdeburg, Germany.

Patients were recruited from July 2020 to January 2023 from the university outpatient clinic, which is specialized in care for comorbidities in patients with diabetes and MetS. A questionnaire about past medical history, diabetes mellitus history (duration, type, treatment, sensory disturbances, complaints, and movement restrictions in daily life), autonomic diabetic neuropathy (dizziness, heart rate arrhythmia, urination disorders, and sweating function), diabetes-associated comorbidities, and daily activities (sports, handedness, and dominant foot) was performed. A physical examination with bedside neurological and sensory nerve testing was performed by a study physician with a standardized evaluation to identify symptoms and signs. The Montreal Cognitive Assessment (MoCA) or Mini Mental State Examination (MMSE) was performed to determine and quantify the cognitive capabilities of the participants.

Patients with diabetes mellitus or MetS aged between 18 and 85 years were eligible for the study, except for those with the following conditions: macroangiopathy of the lower extremities (ankle-brachial index <0.9 or ≥1.3 in either leg), physical deformities (amputations and foot and leg deformities requiring orthopedic shoe fitting), manifest neuropathic foot ulceration, visual disorders including visual acuity of less than 0.8 (except for corrected myopia and hyperopia), muscular diseases or motor diseases, myocardial infarction <12 weeks, heart failure (New York Heart Association III or IV), transient ischemic attack, stroke, tremor, lack of ability to give consent for any reason, and lack of ability to use a mobile phone. Patients with painful diabetic neuropathy, clinical severe autonomic neuropathy (eg, chronic diarrhea), and other chronic pain conditions (necessitating painkillers) that may impair the ability to perform study-related tasks or vestibular disorders (with signs of vertigo) were not evaluated in the study. We opted to exclude these patients, as our study was intended as a proof-of-concept study to generate a basis for predictive scores of PNP.

A study physician meticulously detailed the study protocol, test procedures, and data policy to the patients deemed eligible. Subsequently, these patients were enrolled in the study upon the provision of written informed consent. Compensation was not paid to the study participants except for travel costs.

PNP Assessment

PNP screening was performed by a study physician with a standardized evaluation to identify symptoms and signs. The neuropathy symptom score (NSS) was used to summarize the type, localization, severity, and exacerbation of neuropathic symptoms on a scale ranging from 0 to 10, with more points indicating more severe symptoms (categorization: 3-4=mild, 5-6=moderate, and 7-10=severe symptoms) [28]. The neuropathic signs were classified with the most widely used and accepted neuropathy disability score (NDS, 0-10 points) that included examination of vibratory perception (using a 128-Hz tuning fork, measurement dorsal on big toe joint, 0=normal and 1=reduced or absent), temperature sensation (using a Tip Therm; Tip Therm GmbH, measurement on the dorsum of the foot, 0=normal and 1=reduced or absent), pain sensation (using pinprick, measurement on the dorsum of the foot, 0=normal and 1=reduced or absent), and ankle reflex (0=normal and 1=reduced or absent) [29]. According to the NDS score, the severity of neuropathy was graded as follows: 3-5=mild, 6-8=moderate, and 9-10=severe. In addition, a monofilament test was performed on 3 test sites on the dorsum of the feet (using 10-g Semmes-Weinstein monofilaments, 0=no insensate sites and 1=with ≥1 insensate sites) [30].

According to the aforementioned physical examinations, we defined PNP as 2 or more positive findings related to neuropathic symptoms or signs (vibration perception, temperature sensation, pain sensation, ankle reflex, and pressure sensation evaluated by the monofilament testing) per foot side [31]. Our definition of PNP is equal to “probable” neuropathy introduced in the Toronto Consensus Guidelines [32]. Based on the assumption that asymmetric (unilateral) PNP may be observed in patients with diabetes [22] and the impact on game outcomes is uncertain, we analyzed data from left and right feet separately. Asymmetric PNP was defined as at least 1 asymmetric finding in the clinical examination. Small fiber neuropathy (SFN) was diagnosed when pain sensation (pinprick) or temperature sensation was reduced or absent [7]. Large fiber neuropathy (LFN) was diagnosed with abnormal results of vibration perception, monofilament testing, or Achilles tendon reflexes [8].

For a subcohort of participants, neurophysiological testing using NCS [33] was carried out by board-certified neurologists in the University Clinic for Neurology, Otto von Guericke University Magdeburg. Nerve conduction velocities (NCVs) and amplitudes

XSL•FO
RenderX

were quantified for sensory (sural) and motor (tibial) nerves in the lower extremities.

Study Cohorts

Overall, 329 patients were analyzed in the study (Figure 1). In subcohort 1, we recruited 37 patients aged 50 to 85 years and diagnosed with diabetes mellitus. They were of normal cognitive status (MoCA>25) who completed the video games and underwent NCS as outlined [33]. In subcohort 2, we analyzed 173 patients with normal cognitive status (MoCA>25 or MMSE>24) aged 55 to 85 years with complete clinical examination and video game data sets. The age, sex, weight, BMI, and diabetes type were matched between patients with PNP and those without (w/o) PNP in subcohort 2.

Figure 1. Overview of study cohorts. A total of 145 patients were excluded from the subgroup analyses for not meeting the predefined inclusion criteria outlined in the Venn diagram, e.g., cognitive dysfunction. NCS: nerve conduction study.

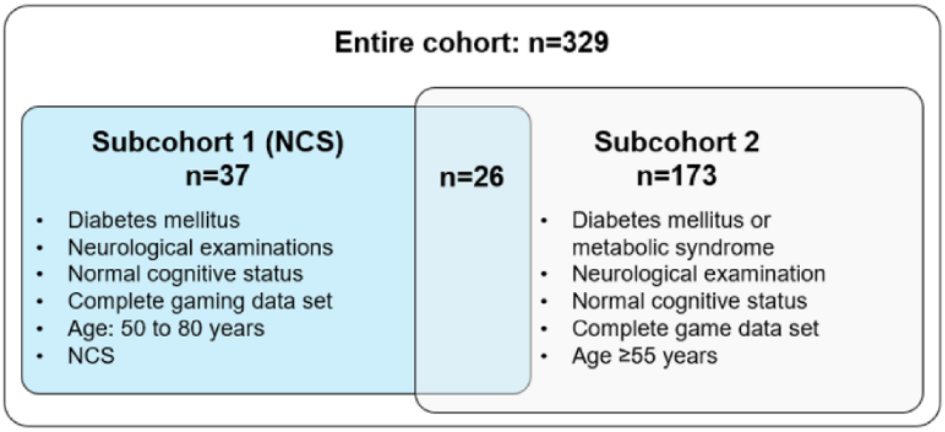

Standardization of Test Procedure

Game sessions were performed with each study participant receiving a size-matching pair of shoes harboring sensor-equipped insoles (ActiSense System, IEE S.A; Figure 2) [34]. The insoles encompassed an electronic control unit and 8 pressure sensors (5.57-cm^2 high-dynamic HD002 force–sensing resistors) that were integrated into the heel, lateral arch, metatarsal 1, 3, and 5, hallux, and toe region of the foot (Figure S1 in Multimedia Appendix 1). The sensors allowed for pressure detection with a sensitivity of 3.4 mbar in the range of 250 mbar to 7 bar with a sampling rate up to 500 Hz, data synchronization (between insoles and smart devices), automatic detection of foot side, internal storage of 16 GB, and up to 10-hour energy supply. The sensors were embedded in foil and did not protrude. The insoles were integrated into the footwear with a support layer of ethylene-vinyl acetate-30 material and a protective layer of sponge.

During the gaming session (Figure 3), sensor data were recorded at 200 Hz and transferred in real time to the application via Bluetooth (5.0) for smooth steering of games. During the test, the participants were seated on a chair without armrests in front of a table on which an Android tablet (Samsung Galaxy Tab A T580) was positioned that connected to the insoles. The participants initially calibrated the insoles through 8 standardized steps [34]. The measured pressure thresholds were used to normalize the absolute pressure values to the range of 0 to 1. Subsequently, participants familiarized themselves with the setup and control techniques of the video games through standardized tutorials.

Each gaming session consisted of 4 games: Apple Catch (AC), Balloon Flying (BF), Cross-Pressure (CP), and Island Jump (IJ), which were played sequentially (Figure 2). An introduction video for the gaming application is provided as Multimedia Appendix 2. Each session lasted about 15 minutes. The acquired insole sensor data and game outcomes were transferred in a complete data package to a remote server for data visualization and analyses. A spider chart provided feedback on the game performance to participants after the session (key capabilities: reaction time, sensation, skillfulness, muscle strength, balance, and endurance).

XSL•FO
RenderX

Figure 2. Game-based assessment of peripheral neuropathy. The sensor-equipped gaming insoles harbored 8 embedded pressure sensors in distinct areas of the plantar pedis. The insole served as a steering unit and was connected to a control unit for real-time data transmission via Bluetooth to the game application, which is run on a tablet. The setup allowed the participant to play games solely by modulating plantar pressure values. Each gaming session included 8 calibration steps and 4 games that are Balloon Flying, Apple Catch, Cross-Pressure, and Island Jump. Subsequently, the data were uploaded to the IQ trial server as a data package. The IQ trial server is a central database set up to collect and archive all data sets for further exploration. The participants received a short summary of their performance with a spider chart, scores, levels, and capabilities (beginner, average, advanced, and expert). Further data visualization for the physician and data analyses were realized to indicate the performance of participants in the different games in comparison to maximum achievement levels.

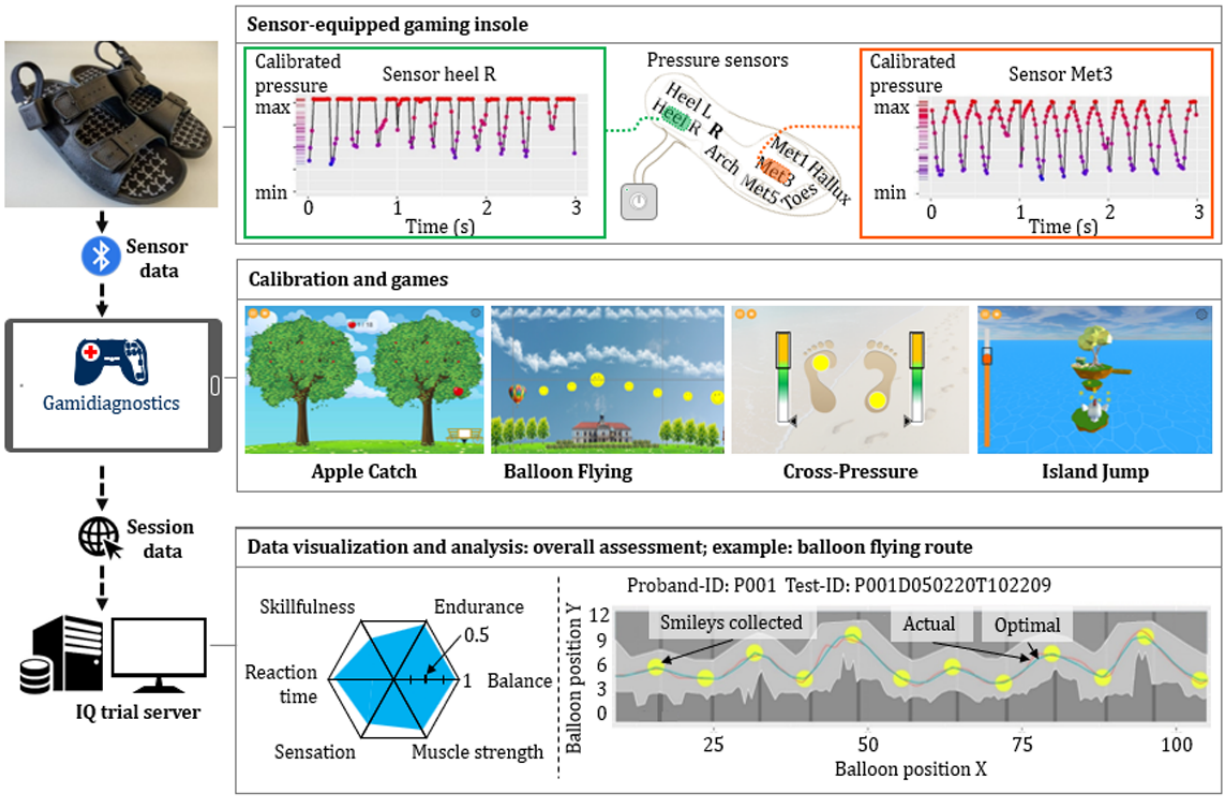

Figure 3. Visualization of the gaming session setup (Balloon Flying). The study participant is instructed to wear a pair of size-matching slippers with pressure sensor–equipped insoles and is seated on an armless chair with hands resting on the table (left image). The insoles serve as control units for the video-based gaming application that is run on a tablet (here the Balloon Flying game image is shown in the top right). The task is to collect as many smileys as possible by modulating the plantar pressure (bottom right image).

XSL•FO
RenderX

Game Design and Challenges

Detailed game design and challenges have been published in the previous work [34]. Briefly, common elements of games were combined in the application, which were playful challenges in a digital setting. The steering control unit was implemented through pressure sensor–equipped insoles. The game design and challenges herein addressed the following requirements: (1) a low complexity setup allowed a quick entry into all games and an easy understanding of control functions through pressure measuring insoles. (2) Motivational elements encouraged completion of tasks and endurance over 15 minutes. (3) Standardized calibration steps and tutorials before each game allowed initial steering attempts to familiarize with the games. Tutorials were repeated on demand. (4) Standardized data acquisition processed with time stamps link sensor data over the course of games, even in the event of failed efforts (maximum allowance of 3 failed efforts per game). Comparison of data sets through time frames was maintained. (5) Definition of distinct challenges in each game provided information on movement control of both feet and legs with variables affected by muscle strength, sensation, balance, and coordination. These variables were tested for each foot separately and both feet in concert. (6) Immediate feedback to the participants on gaming results and overall performance was provided.

To normalize pressure values according to weight and maximum pressure applied by the participants, 8 calibration steps were performed before the commencement of the games [34]. These steps included the application of minimal and maximal pressure at all positions. Furthermore, the participants were instructed to stand up for 5 seconds and keep balance on each foot alone for 5 seconds. The steering unit of the insoles was programmed to apply normalized values in the games, which is minimal-maximal normalization with the transformation of all values to the range of 0 to 1.

Special care was taken in the design of the tutorials for the games. These provided standardized instructions on how to proceed with the games, allowed for some early steps in handling the insoles (examples of pressure application and guiding with insoles), motivational elements (scoring system with smileys), and the possibility to repeat instructions. The remote IQ trial server stored all sensor data together with the pseudonymized data from each participant (eg, personal data, medical history, and clinical examination results) and performed sensor data interpretation (Figure 2). Moreover, it visualized the data and provided feedback to the study participants.

In the AC game, the player was situated in an autumn harvest scenario and tried to catch as many apples as possible with a carriage that was controlled by the plantar pressure application in both forefeet (sensors Met1/3/5) [34]. The apples grew in size and fell from the tree one by one at equal time intervals. The target pressure had to be adapted for the guidance of the carriage below the apple. Additionally, the player must maintain the appropriate pressure until the apple falls into the carriage; otherwise, the carriage moves out of the ideal area. Following each task, the carriage was automatically reset to the middle line. In total, 11 distinct parameters per task were defined that are the basis for player performance assessment. The length of the AC game was standardized and encompassed a total of 14 tasks (apples). Furthermore, the tasks were combined (task combination [TC]) and assigned to different groups that were of relevance for feature extraction.

In the BF game, the player guided a balloon over a skyline (Figure 3) [34]. The flying height was adjusted by the applied pressure detected at the forefoot of the right or left insole, respectively. The balloon approached the ground if no pressure was applied. The ideal flying route was paved by 12 smileys that also suited the game's scoring system. To collect the maximum number of smileys, the player must maneuver the balloon through the skyline and preclude collisions with obstacles, such as clouds, buildings, and trees. In the case of a collision with an obstacle and the absence of corrective measures in 5 seconds, a restart was automatically initiated. The Balloon Flying parcours consisted of 12 distinct tasks, indicating 12 tasks. For each task, distinct parameters were defined to calculate the players' performance. These included (1) the overall number of collected smileys, (2) collision frequency, (3) minimum distance of balloon to smiley and match with optimal flight position, and (4) pressure-related parameters (normalized pressure, pressure difference, pressure gradient, and pressure-time integration). For feature extraction, 4 TCs were furthermore defined for each foot side (task combination for the left foot [TCL] or task combination for the right foot [TCR] 1-4) to assess the performance by low, intermediate, and high obstacles.

In the CP game, the players were instructed to apply pressure on different foot areas (forefoot or heel) with differing target pressure levels (low or high) [34]. Low pressure was indicated by green color, high by yellow color. The actual pressure was visualized by black arrows in a pressure bar to the left and right of the displayed feet. To achieve optimal scores, the player must readily adjust the applied pressure on the corresponding plantar foot areas and maintain the correct pressure level for at least 4.5 seconds. A smiley and checkmark confirmed the accomplishment of the task. If the insole detected no valid action within 25 seconds, the game proceeded to the next task. In total, 16 tasks corresponding to 16 combinations of foot areas and ideal pressure levels were designed in the game. Distinct parameters per task were defined to calculate the players' performance. These parameters included the anticipation time, time outside the optimal pressure zone, relaxation time, normalized pressure, pressure difference, pressure gradient, and pressure-time integration of the left or right foot. The CP game consisted of 16 tasks with target pressure levels (low vs high) and different foot areas (left or right foot, forefoot, or heel). For feature extraction, TCs were furthermore defined for each foot side (TCL or TCR 1-6) to assess the performance by different pressure levels (low vs high) and foot areas (left or right foot, forefoot, or heel).

In the IJ game, the player steered a digital bird with jump movements from island to island in an ocean until a final destiny harboring its home was reached [34]. The player adjusted the jumping distance by modulating the plantar pressure in his forefeet. The jump direction, left or right, was adjusted by the relative pressure distribution below the right and left forefoot. To achieve the optimal score, the player must adjust according

XSL•FO
RenderX

to the predefined pressure values and release the pressure at once when this has been reached. If the optimal pressure was not maintained within narrow limits, the bird jumped into the water. This initiated a game restart, and the player faced the same challenge again. There was a total of 16 islands (tasks). The players' performance was calculated by the following parameters: attempt count, deviation from ideal pressure, anticipation time, execution time, mean pressure of execution phase, normalized pressure, pressure difference, pressure gradient, and pressure-time integration. For feature extraction, 3 TCs were furthermore defined for each foot side (TCL or TCR 1-3) to evaluate the performance by different pressure levels (low, middle, or high) and foot side (left, right, or both).

Game Feature Extraction

Feature extraction of distinct parameters from defined tasks and TCs for the 4 games was set up (Figure S2 in Multimedia Appendix 1) [34]. Game parameters calculated from each game task were considered as primary features, such as reaction time in the first task of the AC game. Apart from that, the concept of TC was introduced, that is, a set of game tasks with similar specifications or macromeasurements were combined. For example, TCL1 covered all tasks that rely on the left foot for controlling the carriage movement to catch apples. The corresponding TCR1 consisted of all tasks that rely on the right foot. The sum, mean, and SD of primary features over game tasks of TCs were treated as secondary features. For example, in the AC game, the reaction time of TCL1 was a secondary feature that was computed from the average reaction time over tasks with the left foot. Overall, 2622 distinctive parameters reflexing players' performance in the entirety and among similar tasks were extracted for each foot side per data set (per participant).

Statistical Analysis

Descriptive statistics are presented in proportions and frequencies for categorical variables. For continuous variables, mean (SD) and median (IQR) are used for the description of normally and nonnormally distributed data, respectively. For comparison between groups, chi-square tests were performed on categorical variables. Either 2-tailed *t* tests or Mann-Whitney *U* tests were selected depending on data distribution. In addition, the Kruskal-Wallis *H* test or 1-way ANOVA was used for comparing multiple groups. Two-sided *P* values below .05 were considered as statistically significant, with the Holm-Bonferroni method adjusting *P* values in pairwise tests [35]. Associations between variables were determined through correlation tests and linear regression using Pearson or Spearman correlation based on data distribution. Missing NCS data were excluded during the correlation analysis. Covariates between the groups were balanced using optimal full matching, which was based on propensity scores [36]. The analysis focused on estimating the average treatment effect specifically for those who received the treatment.

Development of Predictive Models for PNP

Binary classification models were derived to identify feet with PNP within individuals diagnosed with diabetes or MetS using game parameters extracted from acquired data sets. In the preliminary stage of feature selection, game parameters exhibiting minor discrepancies across the comprehensive data set were excluded. A correlation coefficient threshold of 0.75 was set to filter out redundant game parameters with high intercorrelations [31]. For subsequent feature selection, the data set was bifurcated into test and training data sets at a 3:7 ratio. This stratified random sampling ensured class distribution conservation. Further feature selection was performed on the training data set.

Each feature was initially assessed for its correlation with the observed class via univariate statistical techniques (eg, Wilcoxon tests or 2-tailed *t* tests). Features exhibiting interclass differences underwent further model-based variable importance evaluation supported by the R caret library [37]. Multiple models were harnessed to determine feature importance, including random forest, lasso and elastic-net regularized generalized linear models (GLMNETs), support vector machines, k-nearest neighbor, neural networks, and gradient boosting machine. Some of the methods (eg, random forest) intrinsically provide variable importance metrics. For classifiers lacking this intrinsic capability, the importance was inferred via individual receiver operating characteristic curve analyses.

Subsequently, models with different subsets of the top 20 important features (ie, variances of numbers or orders of features) were tested. The aforementioned classifiers were used during modeling. The area under the receiver operating characteristic curve (AUC-ROC) was selected as the performance metric to compare the models trained with different feature combinations. A 5-fold 10 repeats cross-validation was used in training to avoid overfitting and derive a more accurate estimate of the model performance. This statistical method repeatedly divided the training data set into 5 subsets with approximately equal size 3 times. Each subset contained the same proportion of labels as the complete data set. In total, 4 of the 5 subsets were used in the model training, while the remaining subset was used for validation. The average AUC-ROC of cross-validation was considered in the grid search of parameter combination that improves the model performance the most. Ultimately, the obtained models were further applied to the hold-out testing data set to evaluate the models' predictive performance. Classification performance was evaluated by calculating AUC-ROC, balanced accuracy, sensitivity, and specificity, as compared with a reference standard of clinical examination.

Besides, multiple classification models were trained to predict small, large, or mixed nerve fiber damage of the feet in patients with diabetes or MetS. All data sets were used for modeling, and the 5-fold 10 repeats cross-validation was applied due to limited sample size and imbalanced class distribution. The multiclass AUC-ROC was computed to evaluate the model performance [38]. Data processing, statistical computing, and machine learning algorithms were performed using R programming language (version 4.4.1; R Foundation for Statistical Computing) and related open-source libraries as well as Zstats (version 1.0; Hangzhou Yunxiang Statistical Technology Co, Ltd).

XSL•FO
RenderX

Results

Characteristics of Study Participants

Overall, 299 patients with diabetes and 30 individuals with MetS were enrolled into the study and completed the game sessions. Only data sampled from the first game attempt were analyzed. Patient characteristics and clinical findings are summarized in Table 1. Moderate or severe neuropathy symptoms and deficits were observed in 201 (61.1%) and 124 (37.7%) patients, respectively. In total, 10% (n=33) of enrolled participants scored <4 in NSS and >6 in NDS, indicating the presence of an asymptomatic neuropathy (Table S1 in Multimedia Appendix 1). Clinically evident PNP was identified in 247 participants (left foot: n=236 and right foot: n=234). A total of 88 (26.7%) participants were diagnosed with asymmetric affections of PNP (Table S2 in Multimedia Appendix 1).

Table 1. Demographic and clinical profiles of study participants.

Cohort	Overall (N=329)	NCS[a] subcohort 1 (n=37)	Subcohort 2 (n=173)
Sex, n (%)			
Male	204 (62)	16 (43.2)	107 (61.8)
Female	125 (38)	21 (56.8)	66 (38.2)
Age (years), median (IQR)	67.0 (59.0-72.0)	65.0 (61.0-70.0)	68.0 (64.0-72.0)
Weight (kg), median (IQR)	86.0 (77.0-98.0)	85.0 (78.0-92.0)	86.0 (78.0-96.0)
BMI (kg/m^2), median (IQR)	28.7 (25.5-32.5)	28.7 (26.2-32.2)	29.1 (26.1-32.3)
Type of diabetes, n (%)			
No	30 (9.1)	0 (0)	29 (16.8)
Type 1	75 (22.8)	7 (18.9)	29 (16.8)
Type 2	224 (68.1)	30 (81.1)	115 (66.4)
Duration of diabetes (years), median (IQR)	13.0 (6.0-23.5)	13.0 (8.0-25.0)	12.0 (7.0-24.0)
Neuropathy symptom score, n (%)			
Normal (0-2)	106 (32.2)	9 (24.3)	57 (32.9)
Mild (3-4)	22 (6.7)	5 (13.5)	9 (5.2)
Moderate (5-6)	70 (21.3)	7 (18.9)	35 (20.2)
Severe (7-10)	131 (39.8)	16 (43.2)	72 (41.6)
Neuropathy disability score, n (%)			
Normal (0-2)	114 (34.7)	11 (29.7)	65 (37.6)
Mild (3-5)	91 (27.7)	16 (43.2)	52 (30.1)
Moderate (6-8)	103 (31.3)	9 (24.3)	44 (25.4)
Severe (9-10)	21 (6.4)	1 (2.7)	12 (6.9)
Clinically evident PNP[b,c], n (%)			
Left foot	236 (71.7)	24 (64.9)	118 (68.2)
Right foot	234 (71.1)	26 (70.3)	120 (69.4)
Overall	247 (75.1)	27 (73)	126 (72.8)
Asymmetrical PNP[d]	88 (26.7)	5 (13.5)	45 (26)

[a]NCS: nerve conduction study.

[b]PNP: peripheral neuropathy.

[c]PNP was defined as clinical findings of 2 or more neuropathic symptoms, signs, or reflex abnormalities per foot side.

[d]Asymmetric (unilateral) PNP was defined as at least 1 asymmetric finding in the clinical examination.

Game Features Analysis in Subcohort 1

NCS were performed in 37 patients (subcohort 1, aged ≥50 years, Table 1), all diagnosed with diabetes mellitus and normal cognitive status (MoCA>25). This subcohort had similar distributions in age, weight, BMI, and neuropathic scores as the entire cohort (Table S3 in Multimedia Appendix 1). The sex distribution is lightly skewed toward female patients in this subcohort. Patients' game performance was significantly correlated to the sensory and motor NCV and amplitude in the lower extremities. Overall, 277 independent game features were significantly correlated with NCV or nerve amplitude, 79 of

XSL•FO
RenderX

these presented moderate correlations (|*R*|>0.4). Eight representative features are depicted in Figure 4. The strongest correlation was observed between the motor nerve conduction amplitude of the right foot and the maximal execution time in task 8 of the IJ game (*R*=0.65; *P*<.001; adjusted R^2=0.36; last plot in Figure 4).

Figure 4. Performance in nerve conduction study and correlation with game-based findings in subcohort 1. The equation and adjusted R2 were calculated from simple linear regression models. Pearson and Spearman correlation were used for the analysis of normally and nonnormally distributed data, respectively. In total, 5 of 37 patients had incomplete nerve conduction study data. TCL: task combination for the left foot; TCR: task combination for the right foot.

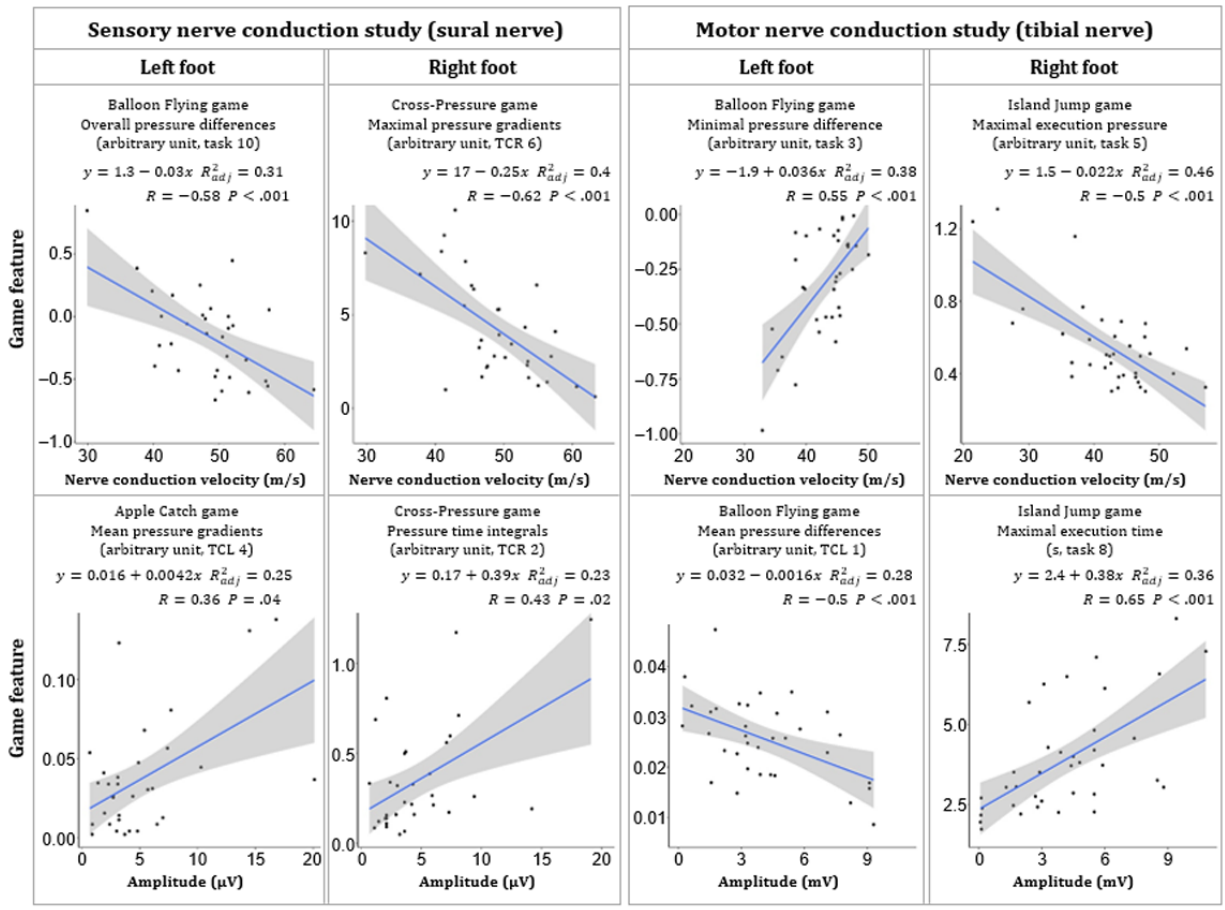

BF and AC games contributed more features that were correlated with the NCV and amplitude of the left foot. However, more features extracted from BF and IJ games were associated with the NCV and amplitude of the right foot. In total, 164 and 113 independent game features were significantly correlated with motor and sensory nerve status, respectively. Overall, 146 independent game features were found to be associated with NCV, and 131 were correlated to the amplitude of nerves of the foot. Only 5 patients had incomplete NCS data (Table S4 in Multimedia Appendix 1).

Game Features Analysis in Subcohort 2

Of all participants (N=329), a subcohort 2 (n=173) with normal cognitive state (MoCA>25) aged 55 years and older was generated (Table 1 and Table S5 in Multimedia Appendix 1). No significant differences in age, sex, weight, BMI, and type of diabetes were observed between patients w/o clinical PNP and with clinical PNP. Individuals with neuropathy had a significantly longer duration of diabetes and more severe neuropathic symptoms. Clinically evidence for PNP was present in 118 left and 120 right feet. For all these feet, at least 1 insensate site in the 10-g monofilament testing was detected. PNP was identified in both feet in 112 patients. No PNP was present in 47 patients, 26% (n=45) of patients had unilateral PNP.

According to the hypothesis that critical skills related to nerve functions are required for successful game performance, 6 hypothesis-driven key capabilities were defined: reaction time (understanding; immediate response to tasks), sensation (fine-tuning of pressure application in subtasks), skillfulness (overall achievements in each game), muscle strength (achievements with high-pressure application), balance (pressure distribution left vs right foot), and endurance (steadiness of pressure application in tasks). The consideration on hypothesis-driven key capabilities and scoring of video games is presented in Figure 5.

XSL•FO
RenderX

Figure 5. Considerations on key capabilities and scoring of video games (hypothesis-driven). Six hypothesis-driven key capabilities were identified to quantify the player's game performance: reaction time (speed and accuracy of immediate task responses), sensation (precision in pressure application during subtasks), skillfulness (overall performance across games), muscle strength (performance in tasks requiring high-pressure application), balance (pressure distribution between left and right foot), and endurance (consistency of pressure application over time). Each key capability score is derived from various underlying game parameters and normalized to a scale from 0 to 1, with 1 representing perfect performance. The actual scores are visualized in a spider diagram.

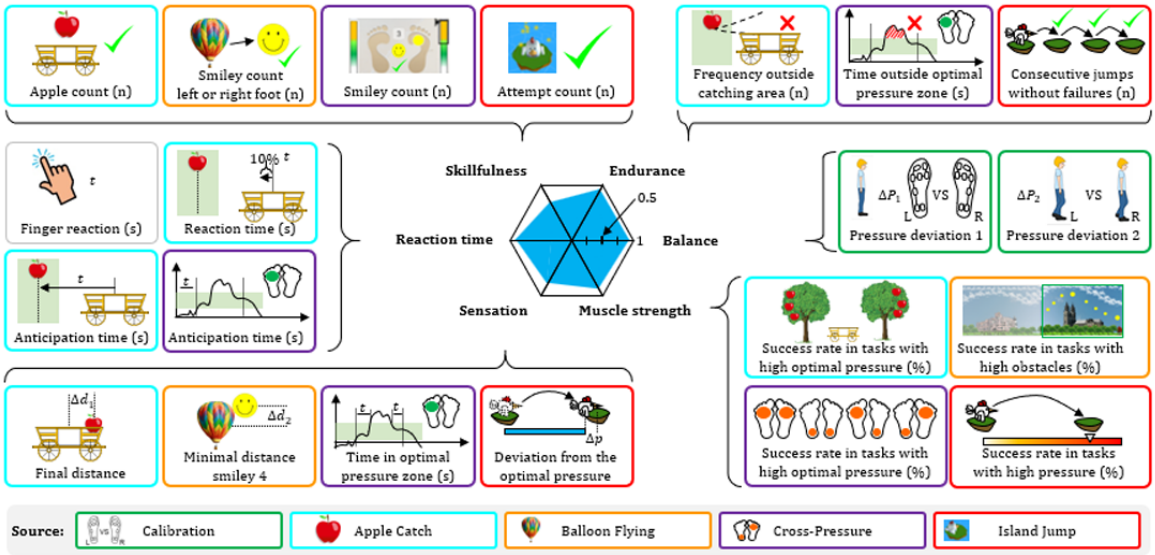

All key capabilities revealed significant differences between patients w/o PNP (n=47) and with symmetrical PNP (n=112; Figure 6). Exemplary game performance differences between 2 participants are visualized for all games (w/o PNP: blue color; with PNP: red color; Figure 7). Of 2622 extracted game features, 50 (left side) and 56 (right side) independent game features revealed significant differences between the "w/o PNP" and "with PNP" groups. Representative examples of these game features were compared and visualized for each game (Figure 8).

Figure 6. Presentation of hypothesis-driven key capabilities between participants without (w/o) and with (w) symmetrical peripheral neuropathy (PNP). Data outliers were excluded in boxplots. Significance levels: **P*=.01-.05, ***P*=.001-.01.

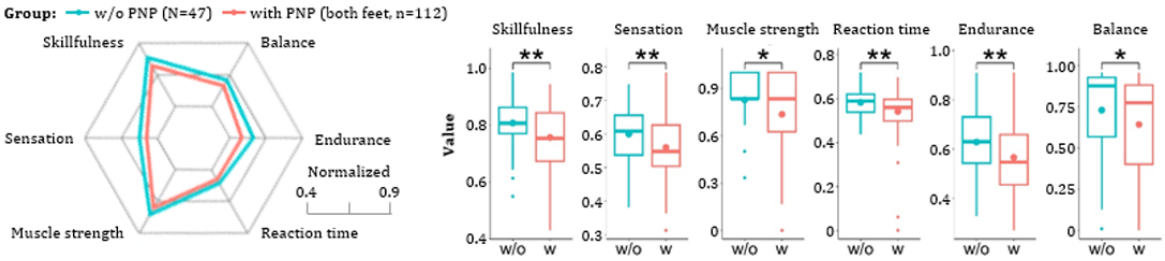

XSL•FO
RenderX

Figure 7. Visualization of game performance of 2 study participants (blue: healthy individual without [w/o] PNP and red: patient with symmetrical PNP). PNP: peripheral neuropathy.

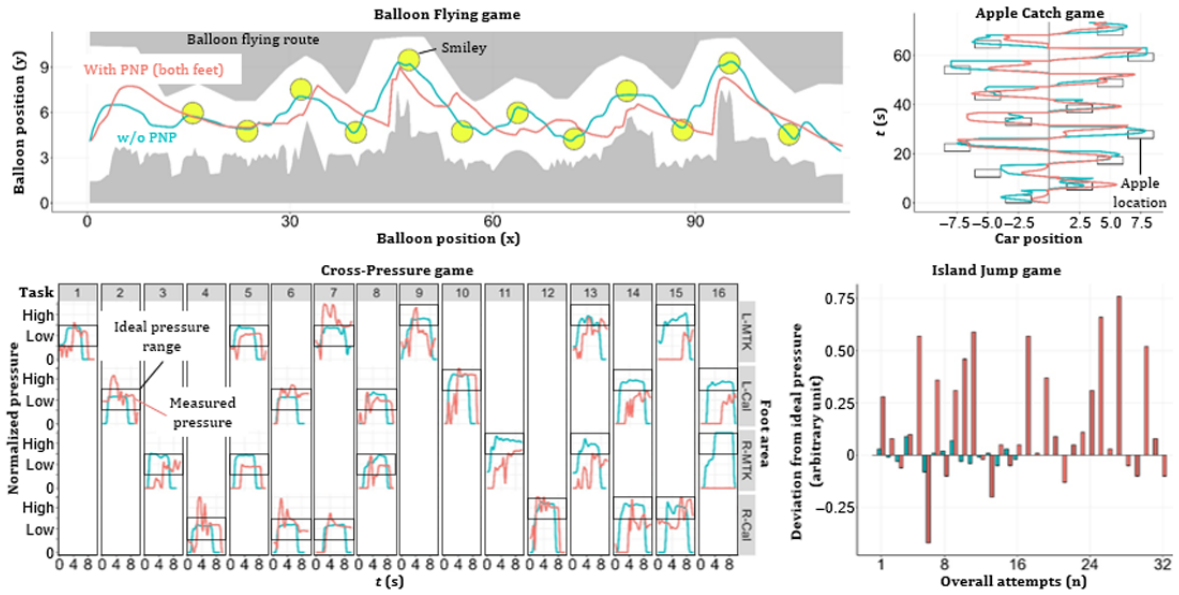

Figure 8. Representative game features correlated with the presence of PNP. Data outliers were excluded in boxplots. AC: Apple Catch; BF: Balloon Flying; CP: Cross-Pressure; IJ: Island Jump; PNP: peripheral neuropathy; w/o: without. Significance levels: **P*=.01-.05, ***P*=.001-.01.

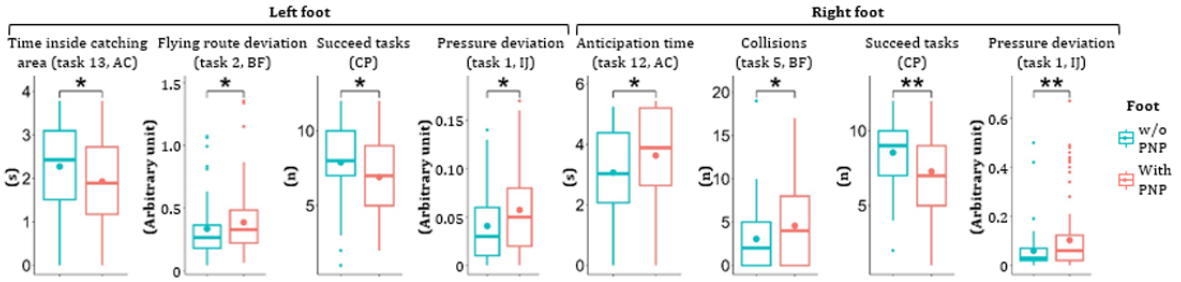

XSL•FO
RenderX

Figure 9. Independent game features that correlated significantly with clinical phenotypes of PNP. Data outliers were excluded in boxplots. AC: Apple Catch; BF: Balloon Flying; CP: Cross-Pressure; IJ: Island Jump; LFN: large fiber neuropathy; PNP: peripheral neuropathy; SFN: small fiber neuropathy; TCL: task combination for the left foot; TCR: task combination for the right foot. Significance levels: ns: *P*>.05, **P*=.01-.05, ***P*=.001-.01, ****P*=.0001-0.001.

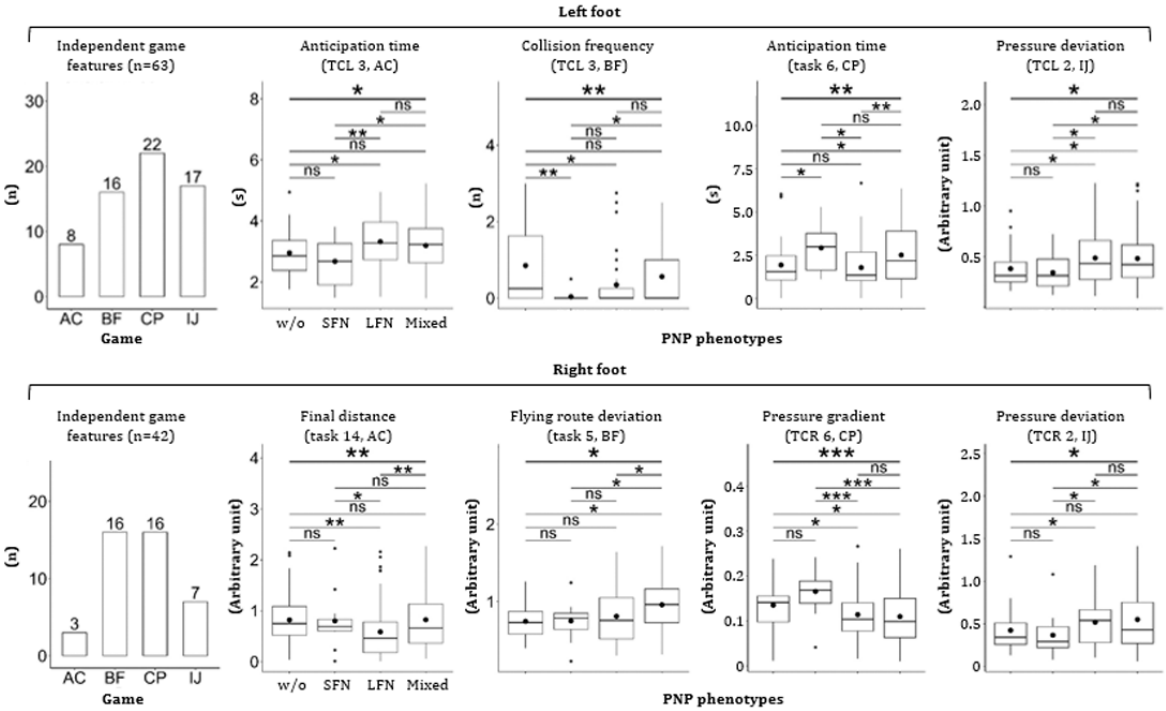

PNP Predictive Models

With the extracted game features from 173 patients in subcohort 2, predictive models for PNP were trained using the 5-fold 10 repeats cross-validation. In total, 70% (n=122) of game data sets were used for model training, and 30% (n=51) for testing. In the hold-out testing data set, the obtained GLMNET identified PNP in the left foot with an adjusted accuracy of 76.1% (sensitivity: 77.1%; specificity: 75%; and AUC-ROC: 0.72). For determining PNP in the right foot, another GLMNET model achieved an accuracy of 81.7% (sensitivity: 83.3%; specificity: 80%; and AUC-ROC: 0.75). An overview of the model's predictive performance is presented in Tables 2 and 3.

Table 2. Predictive performance of peripheral neuropathy (PNP) models on the held-out testing data set (n=51).

	Left foot	Right foot
Model	GLMNET[a]	GLMNET
Tuning parameters		
α	.1	.1
λ	0.000431594129792057	0.00209857537479028
Accuracy (%)	76.1	81.7
Sensitivity (%)	77.1	83.3
Specificity (%)	75	80
AUC-ROC[b]	0.72	0.75

[a]GLMNET: lasso and elastic-net regularized generalized linear model.

[b]AUC-ROC: area under the receiver operating characteristic curve.

XSL•FO
RenderX

Table 3. Confusion matrixes of peripheral neuropathy (PNP) predictive models on the held-out testing data set (n=51).

Foot	Predicted with PNP	Predicted without PNP
Left foot		
With PNP	27	8
Without PNP	4	12
Right foot		
With PNP	30	6
Without PNP	3	12

In addition, multiclassification models were established to differentiate feet w/o PNP versus SFN, LFN, and mixed nerve fiber damages (Table 4). Game features were extracted separately for the left and right foot from the game data set. The representative game features with significant differences between the 4 groups are depicted (Figure 9). The trained gradient boosting machine model yielded a multiclass AUC-ROC of 0.76 and 0.72 when identifying the nerve fiber damage pattern for the left and right foot, respectively (Figure 10). Both models performed better in distinguishing between SFN and LFN. Game features selected by the learning models are provided in Figures S3-S6 in Multimedia Appendix 1.

Figure 10. Visualization of multiclass AUC-ROC curves for the PNP subclassification (small or large or mixed nerve fiber dysfunctions). AUC-ROC: area under the receiver operating characteristic curve; LFN: large fiber neuropathy; PNP: peripheral neuropathy; SFN: small fiber neuropathy.

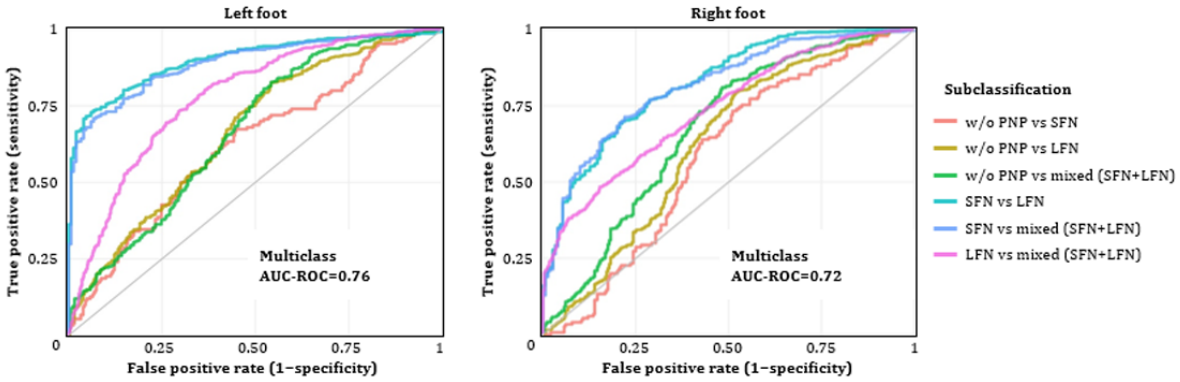

Table 4. Clinical phenotypes of PNPa according to clinical examinationsb.

PNP phenotypes	Without	SFN[c]	LFN[d]	Mixed
SFN (Aδ- or C-fiber deficits, 0-2 points)				
Temperature sensitivity (0=normal and 1=reduced or absent)	0	1-2	0	1-2
Pain sensitivity (pin-prick; 0=normal and 1=reduced or absent)	0	1-2	0	1-2
LFN (Aα- or Aβ-fiber deficits, 0-3 points)				
Ankle reflexes (0=normal and 1=reduced or absent)	0	0	1-3	1-3
Vibration perception (128-Hz tuning fork; 0=normal and 1=reduced or absent)	0	0	1-3	1-3
10-g monofilament (3 test sites; 0=no insensate sites and 1=1-3 insensate sites)	0	0	1-3	1-3

[a]PNP: peripheral neuropathy.

[b]Left feet (overall n=173): without: n=31, SFN: n=15, LFN: n=62, and mixed: n=65; and right feet (overall n=173): without: n=28, SFN: n=15, LFN: n=64, and mixed: n=66.

[c]SFN: small fiber neuropathy.

[d]LFN: large fiber neuropathy.

Discussion

Principal Findings

This proof-of-concept study tested the feasibility and accuracy of a game-based setup to detect PNP in patients with diabetes mellitus or MetS. The games were designed to detect key capabilities related to nerve functions with result interpretations based on hypothesis-driven assumptions as well as algorithms developed by artificial intelligence (AI) methodologies. The findings from NCS data suggested that a standardized gaming

XSL•FO
RenderX

session of 15-minute length is sufficient to acquire an assessment of peripheral nerve function. The extracted game features were correlated with both NCV and amplitude of the major nerves of the foot, suggesting a possible combination of axonal degeneration and segmental demyelination in the study cohort [39]. The detected correlations of game features with NCS findings yielded maximum *R* values of 0.65. These are of similar magnitude as data sets using NDS alone (R=–0.69; P<.001) or with the application of the Toronto Clinical Neuropathy Score (R=–0.58; P<.001) [40].

Learning models trained with gaming data sets from subcohort 2 allowed for the discrimination between feet with PNP and healthy feet with accuracies of 76.1% and 81.7% for each foot side, respectively. The definition of PNP is based on the widely used diagnostic criteria (a combination of NSS, NDS, and monofilament testing) but implemented separately for each foot side [31]. Clinically evident PNP was identified in 75.1% (n=247) of study participants, suggesting a very high risk of developing foot complications in our cohort [41]. This unusual high prevalence of PNP identified in our study cohort is likely due to a patient selection bias. For the specialized outpatient clinic, referral is mostly due to inadequate blood sugar control and comorbidities with diabetes. The patients were mostly at an advanced age with a long duration of diabetes [42]. Notably, the proposed diagnostic criteria for "probable" PNP fitted into the concept of the Toronto Consensus Guidelines [32], which was applied to both feet separately following clinical examination.

In addition, the chosen diagnostic criteria according to clinical examination allowed to discriminate between nerve damage patterns of PNP. The obtained multiclassification models predicted the nerve damage pattern (SFN, LFN, or mixed fiber neuropathy) with AUC-ROC of 0.76 and 0.72 for the left and right foot, respectively. Importantly, gaming data from patients with cognitive impairment (MoCA≤25 or MMSE≤24) were excluded from analyses and AI modeling [43].

Asymptomatic neuropathy was observed in 10% (n=33) of the study participants. This underscores the importance of developing easy-to-use screening tools to avoid or delay the progression of PNP. In addition, more than one-fourth of participants were diagnosed with asymmetric neuropathy, which is rarely reported in other studies [44]. Therefore, we performed analyses and modeling separately for each foot to circumvent the methodological errors caused by asymmetric neuropathy. The results are in line with our concern and yielded different results for the left and right foot.

Strengths and Limitations

The presented game-based approach combines several aspects inherent in screening tools recommended by the American Diabetes Association and guidelines from the Canadian Diabetes Association [5,32]. The procedure may be performed in a highly standardized manner without guidance by health care professionals, given that the tutorial is self-explanatory. The barrier to completion in this study is low, even for those with little or no gaming experience. The immediate response of the study participants upon completion of the games was overwhelmingly positive. More than 90% (n=297) wished to repeat the sessions. However, the results of the second course were not included in this study to rule out learning effects.

However, the number of patients enrolled in the study was still too low to build more refined AI models. The trained learning models using gaming data sets from the age-matched subcohort 2 could achieve binary classification for PNP with moderate accuracy. The obtained multiple classification models performed differently in identifying various clinical phenotypes of PNP. The AUC-ROC for distinguishing between SFN and LFN was obviously higher than differentiating SFN from the "w/o PNP group." However, only a limited number of patients with pure small fiber deficits were enrolled. Small fiber deficits are believed to be precursors of large fiber impairment in PNP [19]. It remains unclear whether game-based diagnostics (ie, incorporation of game-playing elements into the diagnosis of neuropathy) will be similarly sensitive in detecting isolated Aδ- or C-fiber neuropathy. Therefore, more patients in the early stages of diabetic polyneuropathy need to be tested in subsequent studies. Notably, with early diagnosis, preventive measures may be taken, such as weight control or normalization of hyperglycemia [45].

A systematic analysis of the influences of age and cognitive impairment on the execution of the games will also be of interest. The results may set the stage for exergaming approaches and tests of learning capabilities, for example, through repetition of games [46]. A learning effect and changes in the outcome of test findings with repeated performance may not be excluded. Repeated game data sets from the same study participants (n>100) will facilitate the exploration of the influence of the learning effect and enable the testing of the robustness of the study findings.

The order of gaming tasks was also fixed in our setup, as we wished to standardize it as much as possible. Ramifications by introducing randomized sequences of gaming tasks would be of interest in follow-up studies. In addition, the handedness may confound the findings, which may not be unraveled by our analyses. The data sets for left-handedness were too scarce to bring forward a meaningful assessment. A specific cohort of left-handed individuals exceeding 50 is required to refine the findings and explore the impact of left versus right-handedness on the proposed game-based assessment. Reproducibility of findings by others will be key before generalization may be advocated.

Future Directions

In future prospective studies, we envision further cooperation with clinicians in the fields of endocrinology and neurology to implement telemedical platforms for regular PNP assessments. Besides, given the high potential of virtual reality in medical applications, we wish to use such technologies to monitor patients' eye movements and emotional responses in our gaming environment and explore their correlations with game performance [47].

Conclusions

Our proof-of-concept study with a game-based tool to identify patients with PNP was successful and held the potential to ease PNP screening. The findings revealed accuracies in classifying

XSL•FO
RenderX

PNP and nerve damage patterns. Playful motivational elements and automatic data interpretation may be implemented in future telemedical setups. Future testing of larger cohorts will iteratively improve the performance of the feature extraction procedure and promote AI models.

Acknowledgments

The authors would like to thank C Piehler, J Hoetzsch, A Petrow, and I Siddiquee for their support in the performance of the study, such as software development, recruitment process, and test organization. The authors thank T Meyer (IEE S.A.) for fruitful interactions in the establishment of sensor-equipped insoles as well as V Herbst for advice on insole design. This project was funded by the Ministry of Science, Economics, and Digitalization of the State of Saxony-Anhalt and the European Fund for Regional Development under the Autonomy in Old Age Program (ZS/2016/05/78615 and ZS/2018/12/95325) as well as the Chinese Scholarship Council (201508120093). The primary sponsor of this study is the Medical Faculty of the Otto von Guericke University Magdeburg. The sponsors played no role in the design, execution, or evaluation of the study data.

Data Availability

All data generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AM, VC, and PRM conceptualized and designed the study. AM managed application development. AM, VC, AA, and PRM designed the interventions. AM, EL, JW, AA, SS, YL, ML, NDM, and PRM contributed to the acquisition, analysis, or interpretation of data. IG and AKB performed nerve conduction studies. AM, NDM, and PRM drafted the manuscript. AM, SS, NDM, and PRM critically revised the manuscript for important intellectual content. AM and SS did the statistical analysis. PRM was the initiator of the game platform and supervised the study.

Conflicts of Interest

AM and PRM filed a patent on game-based assessment of peripheral neuropathy. All authors declare no competing interests.

Multimedia Appendix 1
Subcohort analyses and visualization of game features for peripheral neuropathy classification models.
[PDF File (Adobe PDF File), 1709 KB - jmir_v26i1e52323_app1.pdf]

Multimedia Appendix 2
Introduction video on the game-based assessment application for peripheral neuropathy.
[MP4 File (MP4 Video), 186960 KB - jmir_v26i1e52323_app2.mp4]

References

1. Selvarajah D, Kar D, Khunti K, Davies MJ, Scott AR, Walker J, et al. Diabetic peripheral neuropathy: advances in diagnosis and strategies for screening and early intervention. Lancet Diabetes Endocrinol 2019;7(12):938-948 [FREE Full text] [doi: 10.1016/S2213-8587(19)30081-6] [Medline: 31624024]
2. Doughty CT, Seyedsadjadi R. Approach to peripheral neuropathy for the primary care clinician. Am J Med 2018;131(9):1010-1016. [doi: 10.1016/j.amjmed.2017.12.042] [Medline: 29408540]
3. Callaghan BC, Price RS, Feldman EL. Distal symmetric polyneuropathy: a review. JAMA 2015;314(20):2172-2181 [FREE Full text] [doi: 10.1001/jama.2015.13611] [Medline: 26599185]
4. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018;138:271-281. [doi: 10.1016/j.diabres.2018.02.023] [Medline: 29496507]
5. Pop-Busui R, Boulton AJM, Feldman EL, Bril V, Freeman R, Malik RA, et al. Diabetic neuropathy: a position statement by the American Diabetes Association. Diabetes Care 2017;40(1):136-154 [FREE Full text] [doi: 10.2337/dc16-2042] [Medline: 27999003]
6. Gwathmey KG, Pearson KT. Diagnosis and management of sensory polyneuropathy. BMJ 2019;365:l1108. [doi: 10.1136/bmj.l1108] [Medline: 31068323]
7. Hoeijmakers JG, Faber CG, Lauria G, Merkies IS, Waxman SG. Small-fibre neuropathies—advances in diagnosis, pathophysiology and management. Nat Rev Neurol 2012;8(7):369-379. [doi: 10.1038/nrneurol.2012.97] [Medline: 22641108]
8. Kanji JN, Anglin RES, Hunt DL, Panju A. Does this patient with diabetes have large-fiber peripheral neuropathy? JAMA 2010;303(15):1526-1532. [doi: 10.1001/jama.2010.428] [Medline: 20407062]

XSL•FO
RenderX

9. Armstrong DG, Boulton AJM, Bus SA. Diabetic foot ulcers and their recurrence. N Engl J Med 2017;376(24):2367-2375. [doi: 10.1056/NEJMra1615439] [Medline: 28614678]
10. Richard JL, Lavigne JP, Got I, Hartemann A, Malgrange D, Tsirtsikolou D, et al. Management of patients hospitalized for diabetic foot infection: results of the French OPIDIA study. Diabetes Metab 2011;37(3):208-215. [doi: 10.1016/j.diabet.2010.10.003] [Medline: 21169044]
11. Singh N, Armstrong DG, Lipsky BA. Preventing foot ulcers in patients with diabetes. JAMA 2005;293(2):217-228. [doi: 10.1001/jama.293.2.217] [Medline: 15644549]
12. Pinheiro DS, Manzano GM, Nóbrega JAM. Reproducibility in nerve conduction studies and F-wave analysis. Clin Neurophysiol 2008;119(9):2070-2073. [doi: 10.1016/j.clinph.2008.05.006] [Medline: 18621579]
13. Diabetic foot problems: prevention and management: NICE guideline. National Institute for Health and Care Excellence. 2019. URL: https://www.nice.org.uk/guidance/ng19 [accessed 2024-08-28]
14. van Houtum WH. Barriers to implementing foot care. Diabetes Metab Res Rev 2012;28(S1):112-115. [doi: 10.1002/dmrr.2238]
15. Lavery LA, Hunt NA, Lafontaine J, Baxter CL, Ndip A, Boulton AJM. Diabetic foot prevention: a neglected opportunity in high-risk patients. Diabetes Care 2010;33(7):1460-1462 [FREE Full text] [doi: 10.2337/dc10-0310] [Medline: 20424223]
16. Tan LS. The clinical use of the 10g monofilament and its limitations: a review. Diabetes Res Clin Pract 2010;90(1):1-7. [doi: 10.1016/j.diabres.2010.06.021] [Medline: 20655123]
17. Young MJ, Breddy JL, Veves A, Boulton AJ. The prediction of diabetic neuropathic foot ulceration using vibration perception thresholds. A prospective study. Diabetes Care 1994;17(6):557-560. [doi: 10.2337/diacare.17.6.557] [Medline: 8082524]
18. Rith-Najarian SJ, Stolusky T, Gohdes DM. Identifying diabetic patients at high risk for lower-extremity amputation in a primary health care setting. A prospective evaluation of simple screening criteria. Diabetes Care 1992;15(10):1386-1389. [doi: 10.2337/diacare.15.10.1386] [Medline: 1425105]
19. Breiner A, Lovblom LE, Perkins BA, Bril V. Does the prevailing hypothesis that small-fiber dysfunction precedes large-fiber dysfunction apply to type 1 diabetic patients? Diabetes Care 2014 May;37(5):1418-1424. [doi: 10.2337/dc13-2005] [Medline: 24574353]
20. Burgess J, Frank B, Marshall A, Khalil RS, Ponirakis G, Petropoulos IN, et al. Early detection of diabetic peripheral neuropathy: a focus on small nerve fibres. Diagnostics (Basel) 2021;11(2):165 [FREE Full text] [doi: 10.3390/diagnostics11020165] [Medline: 33498918]
21. Lipsky BA, Berendt AR, Cornia PB, Pile JC, Peters EJG, Armstrong DG, et al. 2012 Infectious Diseases Society of America Clinical Practice Guideline for the diagnosis and treatment of diabetic foot infections. J Am Podiatr Med Assoc 2013;103(1):2-7. [doi: 10.7547/1030002] [Medline: 23328846]
22. Izenberg A, Perkins B, Bril V. Diabetic neuropathies. Semin Neurol 2015;35(4):424-430. [doi: 10.1055/s-0035-1558972] [Medline: 26502765]
23. van der Kolk NM, de Vries NM, Kessels RPC, Joosten H, Zwinderman AH, Post B, et al. Effectiveness of home-based and remotely supervised aerobic exercise in Parkinson's disease: a double-blind, randomised controlled trial. Lancet Neurol 2019;18(11):998-1008. [doi: 10.1016/S1474-4422(19)30285-6] [Medline: 31521532]
24. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
25. Burgon T, Casebeer L, Aasen H, Valdenor C, Tamondong-Lachica D, de Belen E, et al. Measuring and improving evidence-based patient care using a web-based gamified approach in primary care (QualityIQ): randomized controlled trial. J Med Internet Res 2021;23(12):e31042 [FREE Full text] [doi: 10.2196/31042] [Medline: 34941547]
26. Hamaya R, Fukuda H, Takebayashi M, Mori M, Matsushima R, Nakano K, et al. Effects of an mHealth app (Kencom) with integrated functions for healthy lifestyles on physical activity levels and cardiovascular risk biomarkers: observational study of 12,602 users. J Med Internet Res 2021;23(4):e21622 [FREE Full text] [doi: 10.2196/21622] [Medline: 33900203]
27. Watson JC, Dyck PJB. Peripheral neuropathy: a practical approach to diagnosis and symptom management. Mayo Clin Proc 2015;90(7):940-951. [doi: 10.1016/j.mayocp.2015.05.004] [Medline: 26141332]
28. Dyck PJ. Detection, characterization, and staging of polyneuropathy: assessed in diabetics. Muscle Nerve 1988;11(1):21-32. [doi: 10.1002/mus.880110106] [Medline: 3277049]
29. Weintrob N, Amitay I, Lilos P, Shalitin S, Lazar L, Josefsberg Z. Bedside neuropathy disability score compared to quantitative sensory testing for measurement of diabetic neuropathy in children, adolescents, and young adults with type 1 diabetes. J Diabetes Complications 2007;21(1):13-19. [doi: 10.1016/j.jdiacomp.2005.11.002] [Medline: 17189869]
30. Olmos PR, Cataland S, O'Dorisio TM, Casey CA, Smead WL, Simon SR. The Semmes-Weinstein monofilament as a potential predictor of foot ulceration in patients with noninsulin-dependent diabetes. Am J Med Sci 1995;309(2):76-82. [doi: 10.1097/00000441-199502000-00004] [Medline: 7847445]
31. Martin CL, Albers JW, Pop-Busui R, DCCT/EDIC Research Group. Neuropathy and related findings in the diabetes control and complications trial/epidemiology of diabetes interventions and complications study. Diabetes Care 2014;37(1):31-38 [FREE Full text] [doi: 10.2337/dc13-2114] [Medline: 24356595]

XSL•FO
RenderX

32. Tesfaye S, Boulton AJM, Dyck PJ, Freeman R, Horowitz M, Kempler P, et al. Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments. Diabetes Care 2010;33(10):2285-2293 [FREE Full text] [doi: 10.2337/dc10-1303] [Medline: 20876709]
33. Baum AK. Neurophysiologie in der Praxis. German: Kohlhammer; 2023.
34. Ming A, Lorek E, Wall J, Schubert T, Ebert N, Galatzky I, et al. Unveiling peripheral neuropathy and cognitive dysfunction in diabetes: an observational and proof-of-concept study with video games and sensor-equipped insoles. Front Endocrinol (Lausanne) 2024;15:1310152 [FREE Full text] [doi: 10.3389/fendo.2024.1310152] [Medline: 38495786]
35. Holm S. A simple sequentially rejective multiple test procedure. Scand J Statist 1979;6(2):65-70 [FREE Full text]
36. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Soft 2011;14(42):1-28. [doi: 10.18637/jss.v042.i08]
37. Kuhn M. Building predictive models in R using the caret package. J Stat Soft 2008;28(5):1-26. [doi: 10.18637/jss.v028.i05]
38. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 2001;45(2):171-186. [doi: 10.1023/A:1010920819831]
39. Morgenstern J, Groener JB, Jende JME, Kurz FT, Strom A, Göpfert J, et al. Neuron-specific biomarkers predict hypo- and hyperalgesia in individuals with diabetic peripheral neuropathy. Diabetologia 2021;64(12):2843-2855 [FREE Full text] [doi: 10.1007/s00125-021-05557-6] [Medline: 34480211]
40. Abraham A, Barnett C, Katzberg HD, Lovblom LE, Perkins BA, Bril V. Toronto Clinical Neuropathy Score is valid for a wide spectrum of polyneuropathies. Eur J Neurol 2018;25(3):484-490. [doi: 10.1111/ene.13533] [Medline: 29194856]
41. Boyko EJ, Zelnick LR, Braffett BH, Pop-Busui R, Cowie CC, Lorenzi GM, et al. Risk of foot ulcer and lower-extremity amputation among participants in the diabetes control and complications trial/epidemiology of diabetes interventions and complications study. Diabetes Care 2022;45(2):357-364 [FREE Full text] [doi: 10.2337/dc21-1816] [Medline: 35007329]
42. Charles M, Ejskjaer N, Witte DR, Borch-Johnsen K, Lauritzen T, Sandbaek A. Prevalence of neuropathy and peripheral arterial disease and the impact of treatment in people with screen-detected type 2 diabetes: the ADDITION-Denmark study. Diabetes Care 2011;34(10):2244-2249 [FREE Full text] [doi: 10.2337/dc11-0903] [Medline: 21816977]
43. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
44. Sloan G, Selvarajah D, Tesfaye S. Pathogenesis, diagnosis and clinical management of diabetic sensorimotor peripheral neuropathy. Nat Rev Endocrinol 2021;17(7):400-420. [doi: 10.1038/s41574-021-00496-z] [Medline: 34050323]
45. Davies M, Aroda V, Collins B, Gabbay R, Green J, Maruthur N, et al. Management of hyperglycemia in type 2 diabetes, 2022. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2022;45(11):2753-2786. [doi: 10.2337/dci22-0034]
46. Ferraz DD, Trippo KV, Duarte GP, Neto MG, Bernardes Santos KO, Filho JO. The effects of functional training, bicycle exercise, and exergaming on walking capacity of elderly patients with Parkinson disease: a pilot randomized controlled single-blinded trial. Arch Phys Med Rehabil 2018;99(5):826-833. [doi: 10.1016/j.apmr.2017.12.014] [Medline: 29337023]
47. Sun P, Zhao Y, Men J, Ma ZR, Jiang HZ, Liu CY, et al. Application of virtual and augmented reality technology in hip surgery: systematic review. J Med Internet Res 2023;25:e37599 [FREE Full text] [doi: 10.2196/37599] [Medline: 36651587]

Abbreviations

AC: Apple Catch
AI: artificial intelligence
AUC-ROC: area under the receiver operating characteristic curve
BF: Balloon Flying
CP: Cross-Pressure
GLMNET: lasso and elastic-net regularized generalized linear model
IJ: Island Jump
LFN: large fiber neuropathy
MetS: metabolic syndrome
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
NCS: nerve conduction study
NCV: nerve conduction velocity
NDS: neuropathy disability score
NSS: neuropathy symptom score
PNP: peripheral neuropathy
SFN: small fiber neuropathy
TC: task combination
TCL: task combination for the left foot

XSL•FO
RenderX

TCR: task combination for the right foot

Edited by A Mavragani; submitted 04.09.23; peer-reviewed by K Blondon, S Munusamy, A Hidki; comments to author 06.04.24; revised version received 31.05.24; accepted 24.06.24; published 01.10.24.

Please cite as:
Ming A, Clemens V, Lorek E, Wall J, Alhajjar A, Galazky I, Baum AK, Li Y, Li M, Stober S, Mertens ND, Mertens PR
Game-Based Assessment of Peripheral Neuropathy Combining Sensor-Equipped Insoles, Video Games, and AI: Proof-of-Concept Study
J Med Internet Res 2024;26:e52323
URL: https://www.jmir.org/2024/1/e52323
doi:10.2196/52323
PMID:

©Antao Ming, Vera Clemens, Elisabeth Lorek, Janina Wall, Ahmad Alhajjar, Imke Galazky, Anne-Katrin Baum, Yang Li, Meng Li, Sebastian Stober, Nils David Mertens, Peter Rene Mertens. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Interaction Effects Between Low Self-Control and Meaning in Life on Internet Gaming Disorder Symptoms and Functioning in Chinese Adolescents: Cross-Sectional Latent Moderated Structural Equation Modeling Study

Ted CT Fong[1,2], PhD; Kunjie Cui[3], PhD; Paul SF Yip[3,4,5], PhD

[1]Research Hub of Population Studies, The University of Hong Kong, Pokfulam, Hong Kong, China (Hong Kong)

[2]Centre on Behavioral Health, The University of Hong Kong, Pokfulam, Hong Kong, China (Hong Kong)

[3]Research Institute of Social Development, Southwestern University of Finance and Economics, Chengdu, China

[4]The HKJC Centre for Suicide Research and Prevention, The University of Hong Kong, Pokfulam, Hong Kong, China (Hong Kong)

[5]Department of Social Work and Social Administration, The University of Hong Kong, Pokfulam, Hong Kong, China (Hong Kong)

Corresponding Author:
Kunjie Cui, PhD
Research Institute of Social Development
Southwestern University of Finance and Economics
No. 555 Liutai Road
Wenjiang District
Chengdu, 611100
China
Phone: 86 17512828565
Email: cuikunjie@swufe.edu.cn

Abstract

Background: Internet gaming disorder (IGD) is an emerging behavioral addiction with mental health implications among adolescents. Low self-control is an established risk factor of IGD. Few studies have, however, examined the moderating role of meaning in life (MIL) on the relationships between low self-control and IGD symptoms and functioning.

Objective: This study aimed to examine the effects of low self-control and MIL and their interaction effects on IGD symptoms and family and school functioning in a structural equation model.

Methods: A sample of 2064 adolescents (967, 46.9% male; mean age 14.6 years) was recruited by multistage cluster random sampling from 5 middle schools in Sichuan, China, in 2022. The participants completed a self-report questionnaire with validated measures on low self-control, presence of MIL, search for MIL, IGD symptoms, school commitment, and family functioning. Construct validity, measurement invariance, and structural invariance of the measures were evaluated by confirmatory factor analysis across sex. Structural equation modeling was conducted to examine the indirect effects of low self-control and MIL on family and school functioning through IGD symptoms. Latent moderated structural equation modeling was performed to examine the interaction effects between low self-control and MIL on IGD symptoms, school commitment, and family functioning.

Results: All scales showed satisfactory model fit and scalar measurement invariance by sex. Males showed significantly greater IGD symptoms and lower levels of self-control (Cohen d=0.25-1.20, P<.001) than females. IGD symptoms were significantly and positively associated with impulsivity (β=.20, P=.01), temper (β=.25, P<.001), and search for meaning (β=.11, P=.048) and significantly and negatively associated with presence of meaning (β=–.21, P<.001). Presence of MIL and impulsivity showed a significant and negative interaction effect (β=–.11, SE .05; P=.03) on IGD symptoms. The positive effect of impulsivity on IGD symptoms was stronger among adolescents with low presence of MIL than those with high presence of MIL. Temper showed significant and positive interaction effects with presence of MIL (β=.08, SE .04; P=.03) and search for MIL (β=.08, SE .04; P=.04) on family functioning. The negative effects of temper on family functioning were stronger among adolescents with low levels of MIL than among those with high levels of MIL.

Conclusions: This study provides the first findings on the interaction effects between low self-control and presence of MIL and search for MIL on IGD symptoms and functioning among a large sample of adolescents in rural China. The results have implications for targeted interventions to help male adolescents with lower self-control and presence of meaning.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59490) doi:10.2196/59490

KEYWORDS

Chinese; impulsivity; interaction effects; internet gaming disorder; latent moderation; meaning in life; self-control; temper

Introduction

Technological advances in the past decade have contributed to increased gaming behaviors among younger generations. Internet gaming disorder (IGD) is a psychiatric condition involving persistent internet gaming behaviors that lead to significant distress [1]. IGD refers to the frequent occurrence of at least 5 of the 9 diagnostic criteria of *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) such as preoccupation (obsession with gaming), withdrawal symptoms, tolerance (need to play internet games for longer times or with greater intensity), displacement (neglect of work, study, or social life due to gaming), and deception (tendency to hide gaming time or negative consequences) [2]. Though a previous study [3] found only a minimal prevalence (0.3%-1%) of gamers qualified for IGD, this phenomenon has been aggravated by the social isolation brought on by the COVID-19 pandemic [4]. A recent meta-analysis estimated the pooled prevalence of IGD to be 8.8% among adolescents [5].

Adolescence is a crucial transition period with greater vulnerability to addictive behaviors and adolescents exhibit higher tendencies toward sensation-seeking and risk-taking behaviors [6]. Apart from existing risk factors of IGD such as loneliness, hopelessness, being bullied, and low self-esteem [5,7], low self-control has been positively linked with maladaptive gaming motivations and IGD symptoms in adolescents [8-10]. Impulsivity and temper have shown significant associations with greater IGD symptoms in adolescent samples [11,12]. IGD has been associated with poor family functioning and school achievement in adolescents [13,14]. Recent studies have linked IGD symptoms with parental factors and family conflicts [15,16].

Adolescents may play internet games for motivations such as socialization, risk-seeking, achievement, immersion, or escape from reality [17,18]. The self-determination theory posits that adolescents seek to fulfil 3 basic psychological needs: autonomy, competence, and relatedness [19]. In the gaming context, adolescents could make decisions and control their virtual characters to acquire a sense of autonomy in the context of "virtual reality." They could gain a sense of competence by overcoming in-game challenges and tasks and achieve a sense of relatedness by connecting with other gamers. A previous study [8] found that failure to meet these psychological needs could undermine individuals' self-control in regulating behaviors and impulses, which in turn contributes to greater IGD symptoms.

Meaning in life (MIL) measures the coherence and purpose of one's life [20] and is operationalized in 2 dimensions: presence of MIL and search for MIL [21]. Though a recent study [22] found that MIL mediated the relationship between basic psychological needs and IGD, the authors only used the total MIL score and did not differentiate between the 2 dimensions. Lower presence of MIL has been consistently associated with greater IGD symptoms in adolescents [23-25]. In comparison, search for MIL has shown positive or nonsignificant associations with IGD symptoms [24,25]. A previous study found that the identity formation process during adolescence could contribute to internet gaming behaviors in the search for identity [26]. Excessive internet gaming could make this exploration maladaptive and lead to negative player identity [27]. The paradoxical roles of the 2 MIL factors on IGD symptoms warrant further investigation.

Both self-control and the presence of MIL could be important psychological resources for resilience and coping. Despite the negative association between search for MIL and subjective well-being [28], the presence of MIL showed positive moderating effects on happiness when individuals were searching for MIL. Schnell and Krampe [29] found that the presence of MIL could buffer the adverse effect of low self-control on mental distress. Another study [30] found that presence of MIL and search for MIL attenuated and amplified the effects of bullying victimization on IGD, respectively, in Chinese male adolescents. However, no existing studies have investigated the interaction effects between MIL and low self-control on IGD symptoms and functioning. Such an investigation would provide insights into how adolescents' presence of MIL and search for MIL may moderate the effects of low self-control on IGD symptoms and functioning.

This study had 3 objectives. Figure 1 depicts the conceptual model and hypothesized relationships of the present study. In the figure, positive and negative associations are hypothesized and depicted in orange solid arrows and blue dashed arrows, respectively. The green dotted arrows denote the potential moderating role of meaning in life on the relationships between low self-control and outcome variables. First, we aimed to examine the relationships among low self-control, MIL, IGD symptoms, and family and school functioning in a structural equation model (SEM). Second, we investigated the indirect effects of low self-control and MIL on adolescents' family and school functioning through IGD symptoms. Third, we aimed to examine the interactive effects between MIL and low self-control on IGD symptoms and family and school functioning of adolescents. The study proposed several hypotheses. In hypothesis 1, low self-control would be positively associated with IGD symptoms and negatively with family and school functioning. In hypothesis 2, the presence of MIL would be negatively associated with IGD symptoms. In hypothesis 3, search for MIL would show positive or nonsignificant associations with IGD symptoms. In hypothesis 4, presence of MIL and search for MIL would have interaction effects with low self-control on IGD symptoms and family and school functioning.

XSL•FO
RenderX

Figure 1. Conceptual model and hypothesized relationships among the study variables in the present study. Positive and negative associations are hypothesized and depicted in solid orange arrows and dash blue arrows, respectively. The dotted green arrows denote the potential moderating role of meaning in life on the relationships between low self-control and outcome variables.

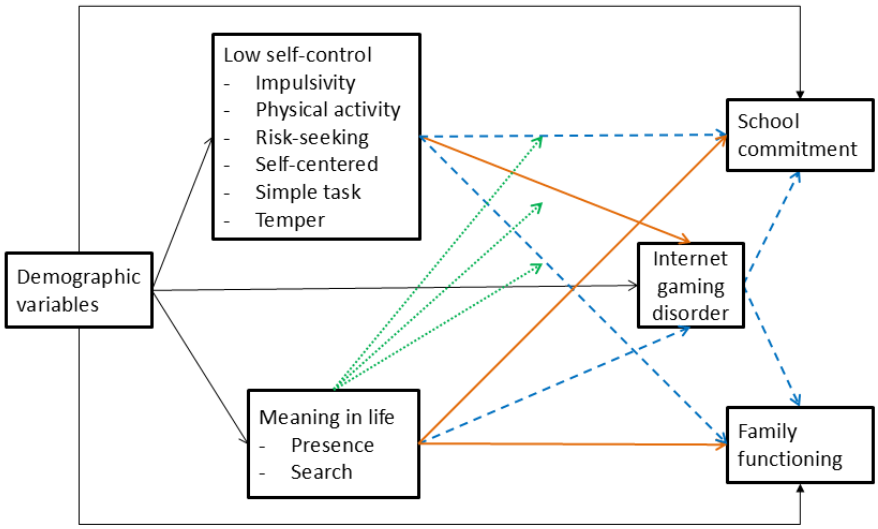

Methods

Study Design and Procedure

This study was cross-sectional in nature and adopted a multistage cluster random sampling method. In the first stage, 5 middle schools were randomly sampled in Liangshan Yi Autonomous prefecture of Sichuan province in China. The Liangshan prefecture is an impoverished region known with its high ethnic diversity in rural China [31]. Adolescents in this region are at risks of developing IGD given the shortage of qualified teachers, limited collaboration between schools and families, and low effectiveness of family education [10]. Next, grade 7-8 classes were randomly selected from each school between June and July 2022. A total of 2378 students were invited from 38 classes through invitation letters in the 5 middle schools. Inclusion criteria of the study were age 11 to 17 years, students studying in grade 7 or 8, and ability to read simplified Chinese. Exclusion criteria of the study were presence of substantial visual and psychiatric impairment that interfered with completion of the questionnaire and inability to obtain written informed consent from the parent or guardian of the students. The eligibility of the students to participate was screened by a research assistant with reference to school records. The overall response rate of the survey was 86.8% and 2064 participants voluntarily elected to join the survey.

Measures

In this study, the self-report questionnaire included the following measures on low self-control, MIL, IGD symptoms, school commitment, and family functioning. The questionnaire also asked the following sociodemographic characteristics of the participants, namely, sex, age, ethnicity, urban or rural registration, and whether they were left-behind children. The questionnaire was administered in a paper-and-pencil format and completed in 20-30 minutes in classroom settings without interference from the school teachers. Participants were reminded to avoid discussion during the survey completion and encouraged to answer the questionnaire honestly. The questionnaire was field tested in May 2022 with minor revisions based on comments from the adolescents.

Low self-control was measured by the Grasmick Low Self-Control Scale [32]. This 23-item scale assessed 6 trait types of low self-control: impulsivity (3 items), physical activity (4 items), risk-seeking (4 items), self-centeredness (4 items), simple task (4 items), and temper (4 items). The items are scored on a 4-point scale ranging from 1=“strongly disagree” to 4=“strongly agree.” Higher scores indicated lower levels of self-control.

MIL was assessed by the 6-item Meaning in Life Questionnaire–short form [21] on 2 dimensions of MIL: presence of MIL (3 items) and search for MIL (3 items). Items are scored on a 7-point scale ranging from 1=“absolutely untrue” to 7=“absolutely true.” Higher scores indicated higher levels of presence of and search for MIL.

The 9-item IGD Scale–short form [33] was used to assess the severity of IGD symptoms according to the *DSM-5* diagnostic criteria over the past year. The 9 items are scored on a 5-point scale ranging from 1=“Never” to 5=“Very often” and added up for the total IGD score (theoretical range=9-45). Higher scores indicated greater levels of IGD symptoms.

School commitment was assessed by 4 self-constructed items on a 5-point scale from 1=“strongly disagree” to 5=“strongly agree.” Example of items included “Getting good grades is important to me.” Family functioning was measured by 5 self-constructed items on a 5-point scale ranging from 1=“absolutely untrue” to 5=“absolutely true.” Example of items included “When problems arise, we are able to work together to solve them and rely on each other.” Higher scores indicated greater levels of commitment to school and family functioning.

Composite reliability of the latent factors was evaluated using McDonald omega (ω) with values ≥.70 indicating good reliability. All of the measures showed good levels of reliability (ω=.71-.94).

XSL•FO
RenderX

Sample Size Planning

Power analysis was performed through Monte Carlo simulation with 1000 replications to determine the required sample size for SEM [34]. In the SEM, we specified a latent predictor with 4 observed indicators, a latent mediator with 3 observed indicators, and a latent outcome variable with 5 observed indicators. The factor loadings (λ) of the latent factors were fixed at 0.7. The acceptable α error level was set at $P<.05$. Effect size of the standardized path coefficients was denoted by 0.14 (small effect) and 0.39 (moderate effect) [35]. The sample size of male (n=967) and female (n=1097) subsamples showed adequate statistical power for detecting statistical significance for direct effects (91.2%-97.3%) and indirect effects (83.5%-89.6%) with small effect sizes.

Data Analyses

Construct Validity of Measures

This study conducted data analysis in 4 steps. The first step examined the factorial validity of the measures by confirmatory factor analysis (CFA). The 47 items on low self-control, MIL, IGD symptoms, school commitment, and family functioning were analyzed in separate CFA models and a combined 11-factor CFA model in Mplus (version 8.6, Muthén and Muthén) [36]. Preliminary item analysis found substantial floor effects for the low self-control and IGD items and substantial ceiling effects for the school commitment and family functioning items. Given the skewed distributions of the items, these items were treated as ordinal items using the robust weighted least square estimator [37]. Model fit was evaluated through the following criteria on the fit indices [38]: comparative fit index (CFI) ≥.95, Tucker-Lewis index (TLI) ≥.95, root-mean-square error of approximation (RMSEA) ≤.06, and standardized root-mean-square residuals (SRMR) ≤.06. Statistical significance was set at $P<.05$ in the present study.

Measurement and Structural Invariance Across Sex

The second step evaluated the measurement invariance of the scales across sex. Multiple group CFA estimated the configural invariance model with different item loadings and thresholds across sex. Scalar invariance model fixed all item loadings and thresholds to be equal across sex [39]. Model comparison was performed using the chi-square difference test and change in model fit indices, with ΔCFI >–.01 and ΔRMSEA and ΔSRMR <.01 supporting the more constrained model [40]. Structural invariance tests compared the means of latent factors across sex on a standardized metric. Cohen *d* denoted the effect sizes with cutoff scores of 0.2, 0.5, and 0.8 for small, moderate, and large sizes, respectively [41]. Missing data were minimal and were handled using full information maximum likelihood under the missing-at-random assumption [42].

Structural Equation Model

Third, SEM was conducted to examine the direct and indirect effects from low self-control and MIL to school commitment and family functioning through IGD symptoms. In the SEM, the 6 low self-control factors and 2 MIL factors were the latent predictors; the IGD factor was the latent mediator; and the school commitment and family functioning factors were the latent outcomes. The model included sex, age, ethnic minority, urban registration, and left-behind children as control variables. Indirect effects were estimated using 5000 bootstrap draws and considered statistically significant if their 95% bootstrapped CI excluded 0. R^2 values denoted the proportion of explained variance for the dependent variables. Sensitivity analysis was performed by conducting multigroup SEM across sex to examine potential sex differences.

Interaction Effects Between MIL and Low Self-Control

In the fourth step, we conducted latent interaction analysis by adding interaction terms between the latent MIL factors and low self-control factors to the SEM separately. The latent interaction SEM was estimated using the XWITH option with numerical integration in Mplus, which allowed us to examine whether MIL factors moderated the relationships between low self-control and IGD symptoms and latent outcomes. Standardized coefficients for the latent interaction terms were interpreted in terms of direction and statistical significance. Simple slopes analysis was conducted to probe the interaction effects by testing the conditional effects of low-self-control on IGD symptoms and functioning at mean, low (mean –1.5SD), and high (mean + 1.5SD) levels of MIL.

Ethical Considerations

The study received ethical approval from the Ethics Committee of the Southwestern University of Finance and Economics (approval number 2422CSH068). The study objectives and procedures were clearly explained to the participants by the research assistant in each classroom. Participation was completely voluntary. Written informed consent was obtained from the parents or legal guardians of the adolescents. Participants had the right to opt out of the survey and not answer specific questions without any adverse outcomes. All collected data was anonymized to preserve confidentiality of personal information. No financial compensation was provided to the study participants. All procedures contributing to this work comply with the Helsinki Declaration of 1975, as revised in 2008.

Results

Sample Characteristics

As shown in Table S1 in Multimedia Appendix 1, a total of 1097 out of 2064 participants (53.1%) were females, and the average age was 14.6 (SD 1.10) years. Most of the participants were ethnic minorities (1931/2064, 93.6%) and had rural registration (1929/2064, 93.5%). Over 1 quarter (592/2064, 28.7%) of the sample were left-behind children in their hometown. Out of 2064, a total of 1281 (62.1%) participants were regular gamers who played internet games for at least 3 hours per week. Overall, the sample reported moderate levels of self-control, presence of MIL, and search for MIL, moderately high levels of school commitment and family functioning, and low levels of IGD symptoms (Multimedia Appendix 1). The sample prevalence of IGD was 2% (36/2064). The IGD prevalence was significantly higher (χ^2_1=14.1, $P<.001$) in males (prevalence=3%; 28/967) than females (prevalence=1%; 8/1097).

XSL•FO
RenderX

Construct Validity of Measures

All measurement scales showed acceptable model fit to the data in CFA with CFI>.95, TLI≥.95, RMSEA≈.06, and SRMR<.06 (Multimedia Appendix 2). Substantial factor loadings were found for low self-control (λ=0.53-0.85), MIL (λ=0.68-0.87), IGD (λ=0.70-0.88), school commitment (λ=0.45-0.77), and family functioning (λ=0.83-0.93).

Measurement and Structural Invariance Across Sex

Both configural and scalar invariance models provided adequate model fits across sex. Despite significant chi-square difference test between the 2 models ($\Delta\chi^2_{155}$=357.6, *P*<.001), the scalar invariance model did not show deterioration in fit indices. As Table 1 shows, males showed significantly higher latent means in physical activity, risk-seeking, self-centeredness, and IGD symptoms (*d*=0.25-1.20, *P*<.001) and significantly lower latent means in school commitment (*d*=–0.42, *P*<.001) than females. The latent mean differences in other factors were not significant (*d* =–0.08-0.14, *P*=.12-.86).

Table 1. Latent mean difference of males compared with females in the study variables in the scalar invariance model (N=2064).

Variables	Cohen $d^{a,b}$	SE	*P* value
Low self-control			
Impulsivity	0.03	0.054	.53
Physical activity	0.33	0.056	<.001
Risk-seeking	0.34	0.064	<.001
Self-centeredness	0.25	0.062	<.001
Simple task	0.14	0.057	.06
Temper	0.01	0.053	.86
Presence of meaning	0.01	0.046	.77
Search for meaning	–0.07	0.048	.17
Internet gaming disorder symptoms	1.20	0.209	<.001
School commitment	–0.42	0.058	<.001
Family functioning	–0.08	0.051	.12

[a]All latent means were fixed at zeros in the female subgroup for identification purpose.

[b]Cohen *d*: SD for males over females.

Structural Equation Model

The SEM with covariates provided an adequate fit to the whole sample. Multimedia Appendix 3 shows the standardized effects of demographic variables on the latent factors in the model. Figure 2 depicts the standardized regression paths from low self-control and MIL to school commitment and family functioning through IGD symptoms. For simplicity of presentation, the figure does not display the 47 observed items of the latent variables, demographic variables, and their associated pathways. Positive and negative paths are represented by continuous orange lines and dash blue lines, respectively. Significant paths involving IGD symptoms are highlighted in bold. Residual correlation between school commitment and family functioning is shown in black dashed arrows. IGD symptoms were significantly and positively associated with impulsivity (β=.20, *P*=.01), temper (β=.25, *P*<.001), and search for meaning (β=.11, *P*=.048), and significantly and negatively associated with presence of meaning (β=–.21, *P*<.001). IGD symptoms were significantly and negatively associated with school commitment (β=–.34, *P*<.001) and family functioning (β=–.08, *P*=.03). The model explained 37.1%, 41.1%, and 33.7% of the variances of IGD symptoms, school commitment, and family functioning, respectively.

As Table 2 shows, out of the 6 low self-control factors, impulsivity and temper showed significant and negative indirect effects on school commitment and family functioning through IGD symptoms. The presence of MIL had significant and positive indirect effects on school commitment and family functioning through IGD symptoms. Search for MIL showed a significant and negative indirect effect on school commitment through IGD symptoms. Multigroup SEM found comparable indirect effects from impulsivity, temper, and presence of MIL to school commitment and family functioning through IGD symptoms across sex. The negative indirect effect from search for meaning to school commitment through IGD symptoms was significant in males (αβ=–.059, 95% CI –0.115 to –0.008) but not in females (αβ=–.012, 95% CI –0.074 to 0.046).

XSL•FO
RenderX

Figure 2. Standardized path coefficients and standard errors (in parenthesis) from low self-control factors and meaning in life factors to school commitment and family functioning through internet gaming disorder symptoms in the structural equation model. *P<.05; **P<.01.

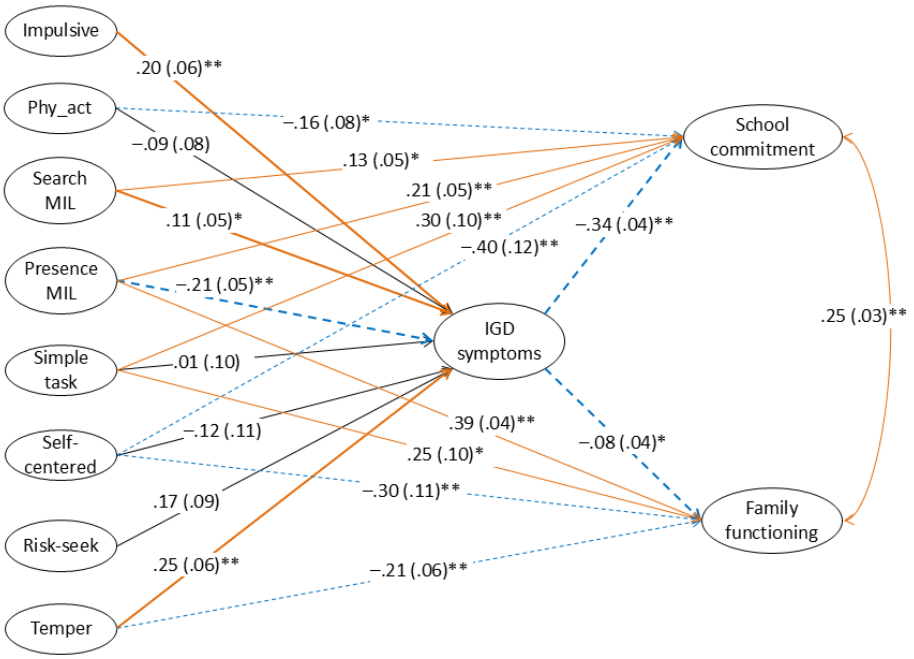

Table 2. Standardized indirect effects from low self-control and meaning in life to school commitment and family functioning through internet gaming disorder symptoms in the structural equation model (N=2064).

Predictor variable	School commitment	Family functioning
	Estimate (95% CI)	Estimate (95% CI)
Impulsivity	–0.066 (–0.119 to –0.016)	–0.016 (–0.041 to –0.001)
Physical activity	0.029 (–0.018 to 0.083)	0.007 (–0.005 to 0.025)
Risk-seeking	–0.058 (–0.132 to 0.003)	–0.014 (–0.038 to 0.000)
Self-centered	0.039 (–0.038 to 0.119)	0.010 (–0.011 to 0.034)
Simple task	–0.004 (–0.073 to 0.049)	–0.001 (–0.021 to 0.012)
Temper	–0.086 (–0.136 to –0.044)	–0.021 (–0.046 to –0.004)
Presence of meaning in life	0.071 (0.036 to 0.107)	0.018 (0.002 to 0.035)
Search for meaning in life	–0.037 (–0.075 to –0.006)	–0.009 (–0.023 to 0.000)

Interaction Effects Between MIL and Low Self-Control

Latent interaction SEM was conducted to examine 4 interaction effects between MIL factors and impulsivity and temper. Presence of MIL and impulsivity showed a significant and negative interaction effect (β=–.11, SE 0.05; P=.03) on IGD symptoms. As Figure 3 shows, the positive effect of impulsivity on IGD symptoms was stronger (dash orange line) among adolescents with low presence of MIL and was much weaker (solid blue line) among those with high presence of MIL. The effects of other 3 interaction terms on IGD symptoms were not significant (β=–.06-.01, SE 0.05-0.06; P=.17-.92). Both impulsivity and temper showed positive but nonsignificant interaction effects with presence of MIL and search for MIL on school commitment (β=.04-.08, SE 0.05; P=.09-.44). Presence of MIL and search for MIL showed significant and positive interaction effects (β=.08, SE 0.04; P=.03-.045) with temper on family functioning. As Figure 4 shows, the negative effects of temper on family functioning were stronger (dash orange lines) among adolescents with low levels of MIL and were much weaker (solid blue lines) among those with high levels of MIL.

XSL•FO
RenderX

Figure 3. Effects of impulsivity on internet gaming disorder symptoms conditional on presence of meaning in life. For presence of meaning in life, high levels=mean+1.5SD and low levels=mean−1.5SD.

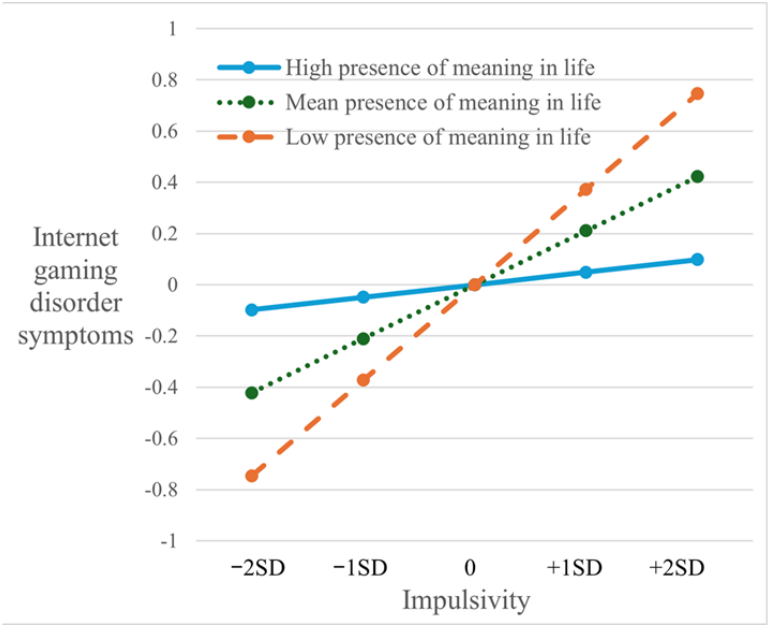

Figure 4. Effects of temper on family functioning conditional on presence of meaning in life and search for meaning in life. For presence of meaning in life and search for meaning in life, high levels=mean+1.5SD and low levels=mean−1.5SD.

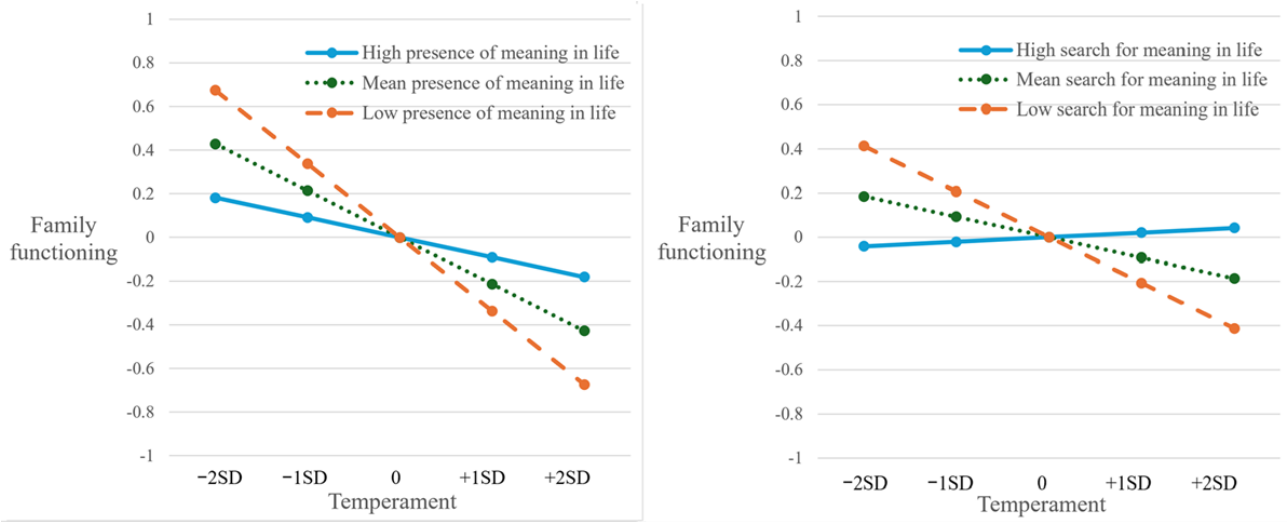

Discussion

Principal Findings

This study evaluated the effects of low self-control and MIL on IGD symptoms and family and school functioning in a large sample of Chinese adolescents. Two low self-control factors (impulsivity and temper) were significantly associated with greater IGD symptoms, and 5 low self-control factors (except simple task) were directly or indirectly associated with worse school commitment and family functioning. These provided empirical support for hypothesis 1. IGD symptoms were found to have significant and negative associations with presence of MIL and significant and positive associations with search for MIL. These results are consistent with hypothesis 2 and hypothesis 3. Consistent with hypothesis 4, MIL factors showed interaction effects with impulsivity and temper on IGD symptoms and family functioning.

Comparison With Previous Work

Consistent with previous literature [22-24,43], adolescents with greater presence of MIL showed lower IGD symptoms and better family functioning and school commitment. Presence of MIL has found a significant and negative indirect effect on suicidal ideation through psychological distress in Hong Kong young adults [44]. The indirect effects from low self-control and MIL to school commitment were stronger than those for family functioning. This appears to suggest greater relevance of IGD in the school context than the family context. Furthermore, this study found that presence of MIL could play a buffering role on the adverse effect of impulsivity on IGD symptoms among adolescents. This is consistent with previous findings on the moderating role of presence of MIL on drug use, suicide risk, and internet addiction [45-47].

In line with recent findings [25], search for MIL showed a positive effect on IGD symptoms after controlling for presence

XSL•FO
RenderX

of MIL. In contrast to the positive direct effect from search for MIL to school commitment, the negative indirect effect through IGD symptoms implies that adolescents could search for MIL by internet gaming and maladaptive gaming behaviors could lead to decrement in school commitment. Nevertheless, search for MIL was found to positively modulate the negative effect of temper on family functioning. Overall, the present results suggest that search for MIL plays a complex role in relation to adolescents' functioning through their IGD symptoms.

In our study, the sex differences in the IGD symptoms and self-control agree with the literature [8,48-50]. In a qualitative study [51], gaming was a meaningful activity for problem gamers and the gaming experience offered a sense of meaning to them. Our multigroup SEM results suggest a more salient role for search for MIL in predicting IGD symptoms and school commitment in males. A previous study [30] found that presence of MIL and search for MIL attenuated and amplified the effects of bullying victimization on IGD, respectively, in male adolescents but not in female adolescents. This discrepancy highlights greater service needs to foster a sense of meaning among male adolescents.

IGD is a prominent public health issue in Asian countries such as China and South Korea [52]. The present sample displayed substantially lower levels of IGD symptoms (Cohen d=0.87, P<.01) than that reported in a recent study among young people in Hong Kong [49]. Adolescents in China are only allowed to play internet games for at most 1.5 hours on weekdays and 3 hours on weekends and holidays [53]. It is plausible that the government restriction on internet gaming contributed to substantially lower severity of IGD in adolescents in China than in Hong Kong. However, presence of floor effects for low self-control and IGD and ceiling effects for functioning indicated potential lack of sensitivity of the measures in capture individual differences. Underestimation of the true variability could cast doubts on the validity of the present findings.

Implications for Future Work

In Liangshan, many parents migrated to urban areas for work opportunities and better livelihood, resulting in many left-behind children. These adolescents showed lower levels of presence of meaning, family functioning, and self-control in the present sample. Previous studies have found inadequate parental supervision and greater loneliness among left-behind children in rural China [54,55]. Interestingly, ethnic minorities showed fewer IGD symptoms and higher school commitment than the Han counterparts. The high proportion of ethnic minorities denotes a sense of cultural similarity and better social support, both of which could contribute to lower IGD risks [56]. From a practical perspective, positive psychology interventions such as character strength intervention [57] and life skills development programs [58] could be implemented in schools to facilitate the adolescents' search for identity and MIL to promote holistic development during adolescence. A previous study found that awareness of medicalization of IGD was linked with reduced gaming time, fewer maladaptive cognitions, and greater help-seeking intention in Chinese adolescents [59]. School-based prevention programs should aim to increase the adolescents' awareness of the IGD symptoms and their adverse effects in a nonstigmatizing manner.

Besides, recent studies have suggested the use of mindfulness-based intervention [60] and self-control intervention [61] to reduce aggression and promote self-control in adolescents. For instance, a recent study [62] has examined the effectiveness of relapse prevention therapy on parent-child communication, parent-child relationships, and parenting needs for youths with IGD. A recent study identifies digital phenotypes (faster stroke acceleration, reduced word spacing, decreased deletion behavior, and longer horizontal strokes) by student tablet data for early detection of IGD among 168 Korean adolescents [63]. Large-scale studies are recommended to evaluate the effectiveness of digital phenotyping as an early detection method of IGD for adolescents. Further research should explore the associations between preferences toward gaming devices and genres and gamification user types [64]. The results could inform the development of targeted treatment programs toward different types of gamers in different gaming devices. Future studies could examine the inter-relationships between IGD and other risk factors such as social withdrawal, cybersex experiences, deviant behaviors at a symptom level through network analysis [65-67].

Limitations

This study used the full latent variable approach to account for the measurement errors of the item indicators, which mitigates the measurement bias and provides more reliable estimates of the regression paths. However, the present study has a few limitations. First, the cross-sectional design did not permit inference of directional relationships between MIL, self-control, and IGD symptoms. Reciprocal effects from IGD to self-control and MIL could not be ruled out. Panel studies are needed to elucidate their temporal relationships. Second, the present study focused on the effects of low self-control and MIL on IGD symptoms and functioning. Omission of important confounding variables could lead to omitted variable bias in the present results. Further studies should examine the effects of factors such as maladaptive cognitions, peer influences, social withdrawal, resilience on IGD [68-70]. Third, the adolescents were recruited from 5 middle schools and were mostly ethnic minorities. There are self-selection and response biases associated with cluster sampling method. Cautions are warranted in generalizing the present results to the general population of adolescents in China.

Fourth, the present results were based solely on self-report data from the adolescents and were subject to common method bias. It was plausible that participants with IGD would under-report their symptoms, leading to an underestimation of the IGD severity in the sample. Future studies should incorporate additional sources of information from parents or teachers to provide a more accurate evaluation of IGD symptoms. Fifth, recent studies have indicated potential overlap between IGD and smartphone addiction and social media addiction [24,71,72]. A recent study found that gaming devices (computer or mobile phone) contributed to differences in gaming time, gaming motivations, and IGD symptoms [73]. It is plausible that IGD and smartphone addiction are simply manifestations of the same

XSL•FO
RenderX

underlying disorder as a result of confounding effect of gaming device. Further studies should clarify the comorbid relationships and the nature of digital addictions through network analysis to ensure precise use of diagnostic categories.

Conclusions

This study provided the first findings on the interaction effects between low self-control and presence of MIL and search for MIL on IGD symptoms and functioning among a large sample of adolescents in rural China. Presence of MIL and search for MIL could mitigate the adverse effects of impulsivity on IGD symptoms and temper on family functioning, respectively. The results have implications for targeted interventions to help male adolescents with lower self-control and presence of meaning.

Acknowledgments

The authors would like to thank all adolescents for their participation in the study and the school principals and teachers for their help in the study coordination and data collection. This study was supported by the National Social Science Fund of China (grant 22CSH068), a ChiangKiang visiting scholar program, and a General Research Fund grant (GRF/HKU 17608522) by the Research Grants Council of Hong Kong. The sponsors had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Data Availability

The raw data and Mplus syntax analyzed in this study are available from the first author upon reasonable request.

Authors' Contributions

TCTF conceptualized the current study, conducted the formal analysis, visualized and interpreted the results, and wrote the original draft. KC was responsible for conceptualization, data curation, project administration, funding acquisition, and reviewed and edited the manuscript. PSFY contributed to conceptualization, methodology, validation, funding acquisition, supervision, and reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographics and descriptive statistics of the sample.
[DOCX File , 15 KB - jmir_v26i1e59490_app1.docx]

Multimedia Appendix 2
Fit indices of confirmatory factor analysis models of the measurement scales, measurement invariance models across gender, and structural equation modeling models in the whole sample and across gender subgroups.
[DOCX File , 19 KB - jmir_v26i1e59490_app2.docx]

Multimedia Appendix 3
Standardized effects of demographic variables on the latent variables in the structural equation model.
[DOCX File , 16 KB - jmir_v26i1e59490_app3.docx]

References

1. Yao YW, Potenza MN, Zhang JT. Internet gaming disorder within the DSM-5 framework and with an eye toward ICD-11. Am J Psychiatry 2017;174(5):486-487. [doi: 10.1176/appi.ajp.2017.16121346] [Medline: 28457151]
2. Poon LYJ, Tsang HWH, Chan TYJ, Man SWT, Ng LY, Wong YLE, et al. Psychometric properties of the internet gaming disorder scale-short-form (IGDS9-SF): systematic review. J Med Internet Res 2021;23(10):e26821 [FREE Full text] [doi: 10.2196/26821] [Medline: 34661543]
3. Przybylski AK, Weinstein N, Murayama K. Internet gaming disorder: investigating the clinical relevance of a new phenomenon. Am J Psychiatry 2017;174(3):230-236 [FREE Full text] [doi: 10.1176/appi.ajp.2016.16020224] [Medline: 27809571]
4. Teng Z, Pontes HM, Nie Q, Griffiths MD, Guo C. Depression and anxiety symptoms associated with internet gaming disorder before and during the COVID-19 pandemic: a longitudinal study. J Behav Addict 2021;10(1):169-180 [FREE Full text] [doi: 10.1556/2006.2021.00016] [Medline: 33704085]

XSL•FO
RenderX

5. Gao YX, Wang JY, Dong GH. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
6. Balogh KN, Mayes LC, Potenza MN. Risk-taking and decision-making in youth: relationships to addiction vulnerability. J Behav Addict 2013;2(1). [doi: 10.1556/JBA.2.2013.1.1] [Medline: 24294500]
7. Yu Y, Fong VWI, Ng JHY, Wang Z, Tian X, Lau JTF. The associations between loneliness, hopelessness, and self-control and internet gaming disorder among university students who were men who have sex with men: cross-sectional mediation study. J Med Internet Res 2023;25:e43532 [FREE Full text] [doi: 10.2196/43532] [Medline: 36649059]
8. Mills DJ, Allen JJ. Self-determination theory, internet gaming disorder, and the mediating role of self-control. Computers in Human Behavior 2020;105:106209. [doi: 10.1016/j.chb.2019.106209]
9. Jo SJ, Yim HW, Jeong H, Lee HK. Moderating effects of depressive symptoms and self-control trait on the association between problematic internet gaming and time spent. Cyberpsychol Behav Soc Netw 2022;25(4):237-244. [doi: 10.1089/cyber.2021.0221] [Medline: 35363564]
10. Quancai L, Meng C, Kunjie C. Social control and self-control: factors linking exposure to domestic violence and adolescents' internet gaming addiction. Front Psychiatry 2023;14:1245563 [FREE Full text] [doi: 10.3389/fpsyt.2023.1245563] [Medline: 37680450]
11. Yu Y, Mo PKH, Zhang J, Li J, Lau JTF. Impulsivity, self-control, interpersonal influences, and maladaptive cognitions as factors of internet gaming disorder among adolescents in China: cross-sectional mediation study. J Med Internet Res 2021;23(10):e26810 [FREE Full text] [doi: 10.2196/26810] [Medline: 34704960]
12. Siste K, Hanafi E, Sen LT, Wahjoepramono POP, Kurniawan A, Yudistiro R. Correction: potential correlates of internet gaming disorder among indonesian medical students: cross-sectional study. J Med Internet Res 2021;23(4):e29790 [FREE Full text] [doi: 10.2196/29790] [Medline: 33882024]
13. Wartberg L, Kriston L, Thomasius R. Internet gaming disorder and problematic social media use in a representative sample of German adolescents: prevalence estimates, comorbid depressive symptoms and related psychosocial aspects. Computers in Human Behavior 2020;103:31-36. [doi: 10.1016/j.chb.2019.09.014]
14. Bonnaire C, Phan O. Relationships between parental attitudes, family functioning and internet gaming disorder in adolescents attending school. Psychiatry Res 2017;255:104-110. [doi: 10.1016/j.psychres.2017.05.030] [Medline: 28535475]
15. She R, Zhang Y, Yang X. Parental factors associated with internet gaming disorder among first-year high school students: longitudinal study. JMIR Serious Games 2022;10(4):e33806 [FREE Full text] [doi: 10.2196/33806] [Medline: 36346660]
16. Shin YB, Kim JJ, Kim H, Kim SJ, Eom H, Jung YH, et al. Managing game-related conflict with parents of young adults with internet gaming disorder: development and feasibility study of a virtual reality app. JMIR Serious Games 2021;9(1):e22494 [FREE Full text] [doi: 10.2196/22494] [Medline: 33459603]
17. Wang HY, Cheng C. The associations between gaming motivation and internet gaming disorder: systematic review and meta-analysis. JMIR Ment Health 2022;9(2):e23700 [FREE Full text] [doi: 10.2196/23700] [Medline: 35175204]
18. Marques LM, Uchida PM, Aguiar FO, Kadri G, Santos RIM, Barbosa SP. Escaping through virtual gaming-what is the association with emotional, social, and mental health? a systematic review. Front Psychiatry 2023;14:1257685 [FREE Full text] [doi: 10.3389/fpsyt.2023.1257685] [Medline: 38025467]
19. Ryan RM, Deci EL. Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices, and future directions. Contemporary Educational Psychology 2020;61:101860. [doi: 10.1016/j.cedpsych.2020.101860]
20. King LA, Hicks JA. The science of meaning in life. Annu Rev Psychol 2021;72:561-584. [doi: 10.1146/annurev-psych-072420-122921] [Medline: 32898466]
21. Steger MF, Frazier P, Oishi S, Kaler M. The meaning in life questionnaire: assessing the presence of and search for meaning in life. Journal of Counseling Psychology 2006;53(1):80-93. [doi: 10.1037/0022-0167.53.1.80]
22. Kaya A, Türk N, Batmaz H, Griffiths MD. Online gaming addiction and basic psychological needs among adolescents: the mediating roles of meaning in life and responsibility. Int J Ment Health Addict 2023;22:2413-2437 [FREE Full text] [doi: 10.1007/s11469-022-00994-9] [Medline: 36643385]
23. Zhang MX, Wang X, Yu SM, Wu AMS. Purpose in life, social support, and internet gaming disorder among Chinese university students: a 1-year follow-up study. Addict Behav 2019;99:106070. [doi: 10.1016/j.addbeh.2019.106070] [Medline: 31430620]
24. Li Y, Mu W, Xie X, Kwok SYCL. Network analysis of internet gaming disorder, problematic social media use, problematic smartphone use, psychological distress, and meaning in life among adolescents. DIGITAL HEALTH 2023;9:205520762311580. [doi: 10.1177/20552076231158036]
25. Fraser R, Slattery J, Yakovenko I. Escaping through video games: using your avatar to find meaning in life. Computers in Human Behavior 2023;144:107756. [doi: 10.1016/j.chb.2023.107756]
26. Monacis L, de Palo V, Griffiths MD, Sinatra M. Exploring individual differences in online addictions: the role of identity and attachment. Int J Ment Health Addict 2017;15(4):853-868 [FREE Full text] [doi: 10.1007/s11469-017-9768-5] [Medline: 28798553]

XSL•FO
RenderX

27. Kokkini V, Tseliou E, Abakoumkin G, Bozatzis N. “Immersed in World of Warcraft”: a discursive study of identity management talk about excessive online gaming. Journal of Language and Social Psychology 2022;41(5):590-612. [doi: 10.1177/0261927x211067820]
28. Cohen K, Cairns D. Is searching for meaning in life associated with reduced subjective well-being? confirmation and possible moderators. J Happiness Stud 2011;13(2):313-331. [doi: 10.1007/s10902-011-9265-7]
29. Schnell T, Krampe H. Meaning in life and self-control buffer stress in times of COVID-19: moderating and mediating effects with regard to mental distress. Front Psychiatry 2020;11:582352 [FREE Full text] [doi: 10.3389/fpsyt.2020.582352] [Medline: 33173525]
30. Zhao H, Li X, Zhou J, Nie Q, Zhou J. The relationship between bullying victimization and online game addiction among Chinese early adolescents: the potential role of meaning in life and gender differences. Children and Youth Services Review 2020;116:105261. [doi: 10.1016/j.childyouth.2020.105261]
31. Cao M, Xu D, Xie F, Liu E, Liu S. The influence factors analysis of households' poverty vulnerability in southwest ethnic areas of China based on the hierarchical linear model:a case study of Liangshan Yi autonomous prefecture. Applied Geography 2016;66:144-152. [doi: 10.1016/j.apgeog.2015.11.016]
32. Arneklev BJ, Grasmick HG, Tittle CR, Bursik RJ. Low self-control and imprudent behavior. J Quant Criminol 1993;9(3):225-247. [doi: 10.1007/bf01064461]
33. Pontes HM, Griffiths MD. Measuring DSM-5 internet gaming disorder: development and validation of a short psychometric scale. Computers in Human Behavior 2015;45:137-143. [doi: 10.1016/j.chb.2014.12.006]
34. Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. Educ Psychol Meas 2013;76(6):913-934 [FREE Full text] [doi: 10.1177/0013164413495237] [Medline: 25705052]
35. Miočević M, O'Rourke HP, MacKinnon DP, Brown HC. Statistical properties of four effect-size measures for mediation models. Behav Res Methods 2018;50(1):285-301 [FREE Full text] [doi: 10.3758/s13428-017-0870-1] [Medline: 28342072]
36. Muthén LK, Muthén BO. Mplus statistical analysis with latent variables: User's guide. Los Angeles: Muthe´n & Muthe´n; 2017. URL: https://www.statmodel.com/ugexcerpts.shtml
37. Rhemtulla M, Brosseau-Liard P, Savalei V. When can categorical variables be treated as continuous? a comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods 2012;17(3):354-373. [doi: 10.1037/a0029315] [Medline: 22799625]
38. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999;6(1):1-55. [doi: 10.1080/10705519909540118]
39. van de Schoot R, Lugtig P, Hox J. A checklist for testing measurement invariance. European Journal of Developmental Psychology 2012;9(4):486-492. [doi: 10.1080/17405629.2012.686740]
40. Chen FF. Sensitivity of goodness of fit Indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal 2007;14(3):464-504. [doi: 10.1080/10705510701301834]
41. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988.
42. Little RJ, Rubin DB. Statistical Analysis with Missing Data, Third edition. Hoboken, New Jersey: John Wiley & Sons; 2019.
43. Wu AMS, Lei LLM, Ku L. Psychological needs, purpose in life, and problem video game playing among Chinese young adults. Int J Psychol 2013;48(4):583-590. [doi: 10.1080/00207594.2012.658057] [Medline: 22506646]
44. Fong TCT, Ho RTH, Yip PSF. Psychometric properties of the patient health questionnaire-4 among Hong Kong young adults in 2021: associations with meaning in life and suicidal ideation. Front Psychiatry 2023;14:1138755 [FREE Full text] [doi: 10.3389/fpsyt.2023.1138755] [Medline: 36970263]
45. Krause N, Pargament KI, Ironson G, Hill P. Religious involvement, financial strain, and poly-drug use: exploring the moderating role of meaning in life. Subst Use Misuse 2017;52(3):286-293. [doi: 10.1080/10826084.2016.1225096] [Medline: 27759484]
46. Wang Z, Wang X, Peng Y, Liu C, He J. Recalled childhood maltreatment and suicide risk in chinese college students: the mediating role of psychache and the moderating role of meaning in life. J Adult Dev 2022;30(1):156-165. [doi: 10.1007/s10804-022-09422-7]
47. Zhang Y, Mei S, Li L, Chai J, Li J, Du H. The relationship between impulsivity and internet addiction in Chinese college students: a moderated mediation analysis of meaning in life and self-esteem. PLoS One 2015;10(7):e0131597 [FREE Full text] [doi: 10.1371/journal.pone.0131597] [Medline: 26171860]
48. Shulman EP, Harden KP, Chein JM, Steinberg L. Sex differences in the developmental trajectories of impulse control and sensation-seeking from early adolescence to early adulthood. J Youth Adolesc 2015;44(1):1-17. [doi: 10.1007/s10964-014-0116-9] [Medline: 24682958]
49. Fong TCT, Cheng Q, Pai CY, Kwan I, Wong C, Cheung SH, et al. Uncovering sample heterogeneity in gaming and social withdrawal behaviors in adolescent and young adult gamers in Hong Kong. Soc Sci Med 2023;321:115774. [doi: 10.1016/j.socscimed.2023.115774] [Medline: 36796169]

XSL•FO
RenderX

50. Su W, Han X, Yu H, Wu Y, Potenza MN. Do men become addicted to internet gaming and women to social media? A meta-analysis examining gender-related differences in specific internet addiction. Computers in Human Behavior 2020;113:106480. [doi: 10.1016/j.chb.2020.106480]
51. Shi J, Renwick R, Turner NE, Kirsh B. Understanding the lives of problem gamers: the meaning, purpose, and influences of video gaming. Computers in Human Behavior 2019;97:291-303. [doi: 10.1016/j.chb.2019.03.023]
52. Liao Z, Chen X, Huang Q, Shen H. Prevalence of gaming disorder in East Asia: a comprehensive meta-analysis. J Behav Addict 2022;11(3):727-738 [FREE Full text] [doi: 10.1556/2006.2022.00050] [Medline: 35932469]
53. Zhan JD, Chan HC. Government regulation of online game addiction. CAIS 2012;30(1):13. [doi: 10.17705/1cais.03013]
54. Jin F, Liu Z, Liu Y, Yao C, Cheng Y. Health status of left-behind children and parenting behaviors of caregivers in poor rural areas - 6 provinces, China, 2018. China CDC Wkly 2021;3(3):54-57 [FREE Full text] [doi: 10.46234/ccdcw2021.017] [Medline: 34594956]
55. Tang D, Choi WI, Deng L, Bian Y, Hu H. Health status of children left behind in rural areas of Sichuan province of China: a cross-sectional study. BMC Int Health Hum Rights 2019;19(1):4 [FREE Full text] [doi: 10.1186/s12914-019-0191-9] [Medline: 30691456]
56. Jian W. Reflections on the difficulties of educational development in minority areas of Western China. Chinese Education & Society 2017;50(1):4-11. [doi: 10.1080/10611932.2016.1262180]
57. Schnitker SA, Balkaya-Ince M, Houltberg B. On the importance of attending to meaning, contextualization, and collaboration for character intervention development for adolescents. The Journal of Positive Psychology 2024;19(5):753-757. [doi: 10.1080/17439760.2024.2325460]
58. To SM. Loneliness, the search for meaning, and the psychological well-being of economically disadvantaged Chinese adolescents living in Hong Kong: Implications for life skills development programs. Child Youth Serv Rev 2016;71:52-60. [doi: 10.1016/j.childyouth.2016.10.037]
59. Yu Y, Li JB, Lau JTF. Awareness and potential impacts of the medicalization of internet gaming disorder: cross-sectional survey among adolescents in China. J Med Internet Res 2021;23(2):e22393 [FREE Full text] [doi: 10.2196/22393] [Medline: 33625362]
60. Zhang A, Zhang Q. How could mindfulness-based intervention reduce aggression in adolescent? mindfulness, emotion dysregulation and self-control as mediators. Curr Psychol 2021;42(6):4483-4497. [doi: 10.1007/s12144-021-01778-5]
61. Gabbiadini A, Riva P, Andrighetto L, Volpato C, Bushman BJ. Preliminary evidence of the effectiveness of a brief self‐control intervention on reducing the short‐term harmful consequences of violent video games on adolescents. J Applied Social Pyschol 2021;52(4):246-258. [doi: 10.1111/jasp.12854]
62. Kapetanovic S, Gurdal S, Einarsson I, Werner M, André F, Håkansson A, et al. Relapse prevention therapy for problem gaming or internet gaming disorder in swedish child and youth psychiatric clinics: protocol for a randomized controlled trial. JMIR Res Protoc 2023;12:e44318 [FREE Full text] [doi: 10.2196/44318] [Medline: 36602846]
63. Cho K, Kim M, Cho Y, Hur JW, Kim DH, Park S, et al. Digital phenotypes for early detection of internet gaming disorder in adolescent students: explorative data-driven study. JMIR Ment Health 2024;11:e50259 [FREE Full text] [doi: 10.2196/50259] [Medline: 38683658]
64. Fong TCT, Lu S, Ye X, Junus A, Yip PSF. Validation of the 12-item gamification user types hexad scale in 1,610 Hong Kong young gamers. Int J Ment Health Addiction 2024. [doi: 10.1007/s11469-024-01367-0]
65. Fong TCT, Junus A, Wen M, Yip PSF. Comorbidity among symptoms of internet gaming disorder, social withdrawal, and depression in 3430 young people in Hong Kong: a network analysis. J Affect Disord 2024;359:319-326. [doi: 10.1016/j.jad.2024.05.091] [Medline: 38777272]
66. Cui K, Fong TCT, Yip PSF. Latent heterogeneity of deviant behaviors and associated factors among ethnic minority adolescents: a latent class analysis. Child Adolesc Psychiatry Ment Health 2024;18(1):93 [FREE Full text] [doi: 10.1186/s13034-024-00771-7] [Medline: 39085944]
67. Fong TCT, Cheung DYT, Choi EPH, Fong DYT, Ho RTH, Ip P, et al. Latent heterogeneity of online sexual experiences and associations with sexual risk behaviors and behavioral health outcomes in Chinese young adults: cross-sectional study. JMIR Public Health Surveill 2024;10:e50020 [FREE Full text] [doi: 10.2196/50020] [Medline: 38277190]
68. Ji Y, Yin MXC, Zhang AY, Wong DFK. Risk and protective factors of internet gaming disorder among Chinese people: a meta-analysis. Aust N Z J Psychiatry 2022;56(4):332-346. [doi: 10.1177/00048674211025703] [Medline: 34250835]
69. Fong TCT, Yip PSF. Prevalence of hikikomori and associations with suicidal ideation, suicide stigma, and help-seeking among 2,022 young adults in Hong Kong. Int J Soc Psychiatry 2023;69(7):1768-1780. [doi: 10.1177/00207640231174376]
70. Lin S, Chen X, Tan L, Liao Z, Li Y, Tang Y, et al. Psychometric properties of the metacognitions about online gaming scale in the Chinese population and its relationship with internet gaming disorder: cross-sectional study. JMIR Serious Games 2024;12:e45985 [FREE Full text] [doi: 10.2196/45985] [Medline: 38648634]
71. Kim H, Choi IY, Kim DJ. Excessive smartphone use and self-esteem among adults with internet gaming disorder: quantitative survey study. JMIR Mhealth Uhealth 2020;8(9):e18505 [FREE Full text] [doi: 10.2196/18505] [Medline: 32990637]
72. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022;8(1):e27719 [FREE Full text] [doi: 10.2196/27719] [Medline: 34081596]

XSL•FO
RenderX

73. Li YF, Tang Y, Huang SC, Tan LX, Huang QP, Chen XX, et al. Role of gaming devices associated with internet gaming disorder in China: cross-sectional study. JMIR Serious Games 2023;11:e40130 [FREE Full text] [doi: 10.2196/40130] [Medline: 36645700]

Abbreviations

CFA: confirmatory factor analysis
CFI: comparative fit index
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
IGD: Internet gaming disorder
MIL: meaning in life
RMSEA: root-mean-square error of approximation
SEM: structural equation modeling
SRMR: standardized root-mean-square residuals
TLI: Tucker-Lewis index

Edited by A Schwartz; submitted 13.04.24; peer-reviewed by X Wang, A Rizvi; comments to author 05.07.24; revised version received 24.07.24; accepted 17.09.24; published 04.11.24.

Please cite as:
Fong TCT, Cui K, Yip PSF
Interaction Effects Between Low Self-Control and Meaning in Life on Internet Gaming Disorder Symptoms and Functioning in Chinese Adolescents: Cross-Sectional Latent Moderated Structural Equation Modeling Study
J Med Internet Res 2024;26:e59490
URL: https://www.jmir.org/2024/1/e59490
doi:10.2196/59490
PMID:

©Ted CT Fong, Kunjie Cui, Paul SF Yip. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Facilitating Thought Progression to Reduce Depressive Symptoms: Randomized Controlled Trial

Shai-Lee Yatziv[1], PhD; Paola Pedrelli[2,3], PhD; Shira Baror[1], PhD; Sydney Ann DeCaro[2], MA; Noam Shachar[1], MA; Bar Sofer[1], MA; Sunday Hull[4], BA; Joshua Curtiss[2], PhD; Moshe Bar[1], PhD

[1]The Leslie and Susan Gonda Brain Science Center, Ramat-Gan, Israel

[2]Depression Clinical and Research Program, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States

[3]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

[4]Old Dominion University, Norfolk, VA, United States

Corresponding Author:
Moshe Bar, PhD
The Leslie and Susan Gonda Brain Science Center
Bar- Ilan University Building number 901
Ramat-Gan, 5290002
Israel
Phone: 972 35317795
Email: moshe.bar@biu.ac.il

Abstract

Background: The constant rise in the prevalence of major depressive disorder calls for new, effective, and accessible interventions that can rapidly and effectively reach a wide range of audiences. Recent developments in the digital health domain suggest that dedicated online platforms may potentially address this gap. Focusing on targeting ruminative thought, a major symptomatic hallmark of depression, in this study we hypothesized that delivering a digital health–based intervention designed to systematically facilitate thought progression would substantially alleviate depression.

Objective: The study aims to investigate the efficacy of a novel digital intervention on the reduction of depressive symptoms. This intervention was designed as an easy-to-use gamified app specifically aimed to facilitate thought progression through intense practicing of associative, semantically broad, fast, and creative thought patterns.

Methods: A randomized clinical trial was conducted, comparing changes in depression symptoms between participants who used the app in the intervention group (n=74) and waitlist control group (n=27) over the course of 8 weeks. All participants filled out a battery of clinical questionnaires to assess the severity of depression at baseline and 4 and 8 weeks after starting the study. These primarily included the Montgomery-Åsberg Depression Rating Scale (MADRS) and the Patient Health Questionnaire-9 as well as the Positive Affect Negative Affect Scale-Negative Affect Score, Ruminative Response Scale, and Symptoms of Depression Questionnaire. Additional questionnaires were implemented to assess anxiety, positive affect, anhedonia, and quality of life.

Results: The results indicate that across multiple clinical measurements, participants in the intervention group who played the gamified app showed greater and faster improvement in depressive symptoms compared with their waitlist control counterparts. The difference between the groups in MADRS improvement was –7.01 points (95% CI –10.72 to –3.29; $P<.001$; Cohen $d=0.67$). Furthermore, the difference in improvement between groups persisted up to 4 weeks posttrial (MADRS differences at week 12: $F_{49,2}=6.62$; $P=.003$; $\eta p^2=0.21$). At the end of the trial, participants who played the app showed high interest in continuing using the app.

Conclusions: The results demonstrate that a gamified app designed to facilitate thought progression is associated with improvement in depressive symptoms. Given its innovative and accessibility features, this gamified method aiming to facilitate thought progression may successfully complement traditional treatments for depression in the future, providing a safe and impactful way to enhance the lives of individuals experiencing depression and anxiety.

Trial Registration: ClinicalTrials.gov NCT05685758; https://clinicaltrials.gov/study/NCT05685758

(J Med Internet Res 2024;26:e56201) doi:10.2196/56201

XSL•FO
RenderX

KEYWORDS

depression; cognitive neuroscience; facilitating thought progression; FTP; mobile phone; digital health; gamification; depression symptoms; randomized controlled trial; RCT; app; depressive disorder; web-based platforms; effectiveness

Introduction

Background

Depressive and anxiety disorders have been on a constant rise as major global disease burdens since 1990 and were among the 10 most common noncommunicable diseases in 2019, even among children as young as 10 years [1]. Major depressive disorder (MDD), which is highly comorbid with anxiety, is characterized by persistent feelings of sadness, loss of interest or pleasure (ie, anhedonia), changes in appetite or weight, disturbances in sleep patterns, fatigue, feelings of worthlessness or guilt, impaired cognitive and social function, and recurrent thoughts of death or suicide [2], altogether making it a highly complex disease.

The global COVID-19 pandemic contributed significantly to the sharp increase in the already-high MDD and anxiety conditions; in 2020, the global number of mental health disorder cases rose significantly, with an additional 76.2 million cases of MDD and 53.2 million cases of anxiety disorder [3]. The COVID-19 pandemic–associated lockdowns and social distancing have led patients to seek and successfully engage with remote psychotherapeutic options, such as online sessions with their therapists, or digital health tools, such as mobile phone apps, for the diagnosis and treatment of depression and anxiety [4]. A recent meta-analysis has demonstrated great efficacy for digital mental health interventions [5], despite their limited face-to-face component and the constant technical support they require. In that manner, digital health has emerged as a promising avenue for addressing mental health challenges in 3 critical aspects: accessibility, retention, and innovation.

Accessibility

Digital health has the clear benefit of allowing existing treatments a wider reach. For example, telehealth and internet-based approaches that translate existing treatments to online platforms have already been shown to be feasible and as effective as in-person care [5-9]. These remote evidence-based interventions use mobile phone apps as an additional platform for delivering care, encompassing clear advantages given their extensive accessibility, immediacy, low cost, and the ability for patients to use these tools at their convenience without having to wait for mental health professionals to be available [10,11]. In particular, cognitive digital health interventions have attracted considerable attention because of their potential for targeting cognitive processes associated with depression more directly [5,9,12-14]. Digital mental health therapies can significantly improve well-being, potentially overcoming stigma-related barriers to seeking treatment and reaching a much wider population, all while enhancing existing treatments [15].

Retention

Digital health platforms can amplify retention in favor of mental health causes by harnessing and customizing online gaming features. A review of studies on digital health apps has found that gamification, the inclusion of game elements within nongame contexts, increased motivation and engagement and resulted in improved outcomes [16]. Such interventions leverage the power of technology to deliver not only accessible and cost-effective support but also its power to sustain long-term motivation and engagement for diverse mental health problems [17-19]. Recent studies have found that online gaming platforms incorporating cognitive or behavioral techniques can help alleviate depression [20-22], demonstrating how gaming can complement existing approaches and potentially improve treatment outcomes for various mental health issues [23-25].

Innovation

The success rate of traditional treatments for MDD, including psychotherapy and pharmacotherapy, remains unsatisfactory: approximately one-third of individuals with MDD do not respond to standard treatments, and high rates of relapse and residual symptoms further highlight the limitations of currently available therapies [26-32]. Furthermore, a large portion (40%-50%) of the population experiencing depression does not even seek treatment [33]. Consequently, there is a clear need for new and innovative solutions, which digital health solutions are well positioned to provide. A central problem with existing treatment options is their focus on symptoms rather than the underlying causes of depression [34,35]. Therefore, although creating gamified online versions of existing treatments may improve accessibility, cost, and engagement, these developments have yet to solve the limited effectiveness of current therapeutic approaches. In other words, beyond successfully implementing remote versions of old treatments, there is an urgent need for remote and gamified solutions that implement innovative interventions and improve treatment outcomes, making life easier for those coping with depression.

Facilitating Thought Progression

This study evaluated a new app-based intervention developed on the basis of a cognitive neuroscience framework for mood and depression, which we termed *facilitating thought progression* (FTP) [36]. FTP is based on research showing that specific thought patterns are directly and reciprocally connected with mood [37-41]. For example, narrow, slow, and repetitive thinking patterns have been associated with a more negative mood, whereas broad, rapid, and expanding thinking patterns are associated with a more positive mood. The FTP approach postulates that persistent rumination not only dampens mood but over time also leads to structural changes in critical brain areas, such as the hippocampus and the prefrontal cortex [42-44]. Given the consistent findings linking depression and anxiety with ruminative thought patterns (repetitive, circular thinking, sluggish thinking, and narrow semantic representation), an FTP-based intervention would focus on reducing ruminative thoughts by broadening associative scope, reducing mental inhibition, and increasing cognitive flexibility.

In addition, it has been demonstrated that people with depression show a reduced ability for globalization—they exhibit a

XSL•FO
RenderX

localized, detail-oriented perspective and are less able to develop a global one. This means that they are limited in their ability to think about things in global terms or in their holistic semantic representation [45,46]. Given the bidirectional relationship between cognitive thought process and mood [35,39,40,47,48], individuals with depression may benefit from cognitive training to facilitate thought patterns that are more broadly associative, more global, and more rapid.

Consistently, our group developed a farm-like mobile phone app game called MoodVille with the aim of providing this form of cognitive training. The app included 5 minigames, and participants needed to complete levels in each of them to earn collectible rewards for their farm development. The therapeutic games were collectively aimed to facilitate thought progression and specifically to enhance cognitive flexibility, associative breadth, creativity, rapid thinking, and globalization. An overview of the app and examples of the levels for each therapeutic game are provided in Figure S1 in Multimedia Appendix 1.

The study aimed to evaluate the efficacy of this FTP-based therapeutic intervention in improving symptoms among patients with MDD.

Methods

Participants

The study was a single-blind, randomized controlled trial (RCT) to test the efficacy of a newly developed FTP-based mobile phone app in alleviating symptoms of depression in individuals with MDD. Clinicians evaluating the participants were blinded to their trial condition. Overall, 117 participants aged between 20 and 50 (mean age 32.43, SD 8.5) years were found eligible for the study. Of the 117 participants, 13 (11.1%) from the intervention group and 3 (2.5%) from the waitlist control group (WLC) discontinued either due to technical and administrative reasons (n=5, 4.2%), personal reasons (n=1, 0.8%), or failure to meet minimum game playing requirements (n=10, 8.5%). Therefore, the intervention group consisted of 74 participants, while the WLC group consisted of 27 participants.

Study Procedure

The study procedures were approved by the Mass General Brigham Institutional Review Board. The study was preregistered (2022P002100). Participants were recruited between October 17, 2022, and May 31, 2023, from online channels, including Craigslist, Facebook (Meta), Instagram (Meta), Reddit, ClinicalTrials.gov, and the recruitment website for research studies conducted in the Mass General Brigham [49]. Interested individuals were first screened over the phone by the study staff. Eligible participants were scheduled to complete a visit with clinical study personnel, where they completed the consent procedure and were evaluated for inclusion and exclusion criteria. The inclusion criteria were meeting the criteria for current MDD per the Mini-International Neuropsychiatric Interview [50] and having mild to moderate depression per the Montgomery-Åsberg Depression Rating Scale (MADRS; MADRS score ≥15 and <35) [51]. The minimum score of 15 was chosen because a cutoff score of 15 to 16 has been shown in 1 study to be the most sensitive for MDD [52], as it falls between mild and moderate severity, suggesting that it is the ideal score to indicate the presence of MDDs. The upper score of 35 was chosen because it indicates the presence of severe depression, which was not identified by the authors as a treatment population and raised safety concerns about participation in a placebo-controlled trial in which the treatment intervention had not yet been validated. Exclusion criteria included the following: meeting the criteria for schizoaffective disorder, bipolar I and II disorder, current posttraumatic stress disorder, panic disorder, obsessive-compulsive disorder that significantly impairs their functioning or lasts >1 hour per day, or personality disorder as long as their MDD diagnosis was primary; having treatment-resistant depression; planning to change treatment regimen (therapy or psychotropic medications) during the 8-week study period; having changed treatment regime (therapy or psychotropic medication) in the 6 weeks before the study initiation; using a computer, internet, or smartphone software–based app for mental health or depression treatment in the 6 weeks before the study; and presence of neurodegenerative diseases or uncorrected visual or dominant-hand motor deficits. Eligible participants were then randomly assigned (3:1) to either the intervention group or the WLC group by a research coordinator. To control for sex at birth and gender-diverse status, blocks of randomized number sequences were used at randomization.

Following randomization, participants in the intervention group began playing the FTP-based mobile app, while WLC participants were informed that they would gain access to the app after an 8-week waiting period. Consistently, study staff contacted the WLC participants at their 8-week mark and offered them a link to download the FTP-based app. During the 8-week study period, participants from both groups were administered web-based weekly questionnaires on depression [53] (Patient Health Questionnaire-9 [PHQ-9]), anxiety [54] (General Anxiety Disorder-7), rumination [55] (Ruminative Response Scale), anhedonia [56] (Snaith-Hamilton Pleasure Scale [SHAPS]), and mood [57] (Positive and Negative Affect Schedule); a monthly quality of life questionnaire [58] (World Health Organization Quality of Life Brief Version); and the Symptoms of Depression Questionnaire [59]. At the end of the study, the intervention group also received a user feedback questionnaire regarding the app itself. All web-based questionnaires were administered and stored via REDCap (Research Electronic Data Capture; Vanderbilt University), a secure web app for surveys and databases. MDD symptoms were evaluated monthly using the MADRS diagnostic questionnaire. Mini-International Neuropsychiatric Interview and MADRS assessments were conducted by clinicians with Master's, Doctor of Philosophy, or Doctor of Medicine-level qualifications who were blind to group assignment and were extensively trained to administer the assessments.

Intervention

Intervention arm participants were given a link to download the FTP-based MoodVille app free of charge and were instructed to play the app for at least 4 days per week and 15 minutes per day for 8 weeks. Participants who did not meet the play

XSL•FO
RenderX

requirement for >2 weeks over the course of the 8-week trial (ie, missing >25% of minimum play requirements) were discontinued, as per the study's protocol and consent form. Study staff regularly monitored each participant's play rate and alerted participants who were close to not meeting their targets. In total, 47 (54%) of the 87 participants in the intervention group were receiving either psychotherapy (n=5, 11%), medication (n=27, 57%), or a combination of both (n=15, 32%). All participants were instructed not to change their treatment regimens during the 8-week study.

The intervention was a farm-like mobile phone app game called MoodVille, and it included 5 minigames, each designed to train 1 of the following cognitive facets: associative breadth, cognitive flexibility, thinking speed, creativity, and global attention (Figure S1 in Multimedia Appendix 1).

Control Group

The WLC allowed participants to continue their regular MDD treatment protocols, if such existed, and were instructed not to change them for the duration of the 8-week waiting period. A total of 19 participants in the WLC group were receiving either psychotherapy (n=3, 16%), medication (n=8, 42%), or a combination of both (n=8, 42%). Study coordinators contacted the WLC participants at their 8-week mark and offered them a link to download the FTP-based app.

Primary and Secondary Outcomes

The primary outcomes of the clinical trial were changes in the clinical questionnaires. Questionnaires were filled out each week, and changes were compared between the intervention and the control groups. The secondary outcomes of the clinical trial were postintervention effects, measured through data collected at week 12 of the trial from participants from both groups who extended their engagement. Additional secondary outcomes were evaluated through participants' ratings of the app, contributing to the assessment of the intervention's efficacy.

Data Analytic Strategy

The primary analysis of interest in this study was the effect of the FTP-based app on MDD severity. Our hypothesis was that clinical scores would be significantly improved for the intervention group compared with WLC. Several analyses were conducted to test this hypothesis. First, a 2-way mixed repeated measures ANOVA was run to assess differences between the 2 groups in their clinical scores, which were defined as the outcome variables. In this study, the baseline and fourth and eighth weeks of the clinical trial were used as the time points of reference for the within-subject independent variable, and group allocation was used as the between-subject independent variable. To address sample size differences between the 2 groups, additional randomization analysis was conducted: the 2-way repeated measures ANOVA was performed in multiple iterations (n=1000), in which random data subsampling of participants from the intervention group was performed so that the sample size in each iteration was equal between the intervention and WLC groups. The proportion of significant results out of the number of iterations is reported in the Results section. Second, 1-tailed *t* test analyses tested whether the outcome variable of change in clinical scores from baseline was significantly enhanced for the intervention group compared with the WLC group after 4 weeks as well as after 8 weeks (group and weeks operationalized as independent variables). Third, the clinical benefit of the intervention was assessed by computing the standardized mean difference (SMD) for each assessment tool using Cohen effect size. Finally, to assess the rate of clinical improvement over time, a mixed linear regression analysis was run, estimating the change in each clinical questionnaire's score over the study's 8 weeks for both groups. A linear model was built for each questionnaire, in which weeks (1-8) and group (intervention and WLC) were implemented as a fixed effect and participants were implemented as the random effect.

Secondary analyses focused on postintervention data. As participants could continue using the app for an additional 4 weeks, these analyses assessed the long-lasting effect of app use on clinical assessment. In this study, the outcome variable of postintervention clinical data was compared between groups at week 12, while dividing the intervention group into those maintaining high engagement (HE) postintervention and those who maintained low engagement (LE) postintervention (engagement as the independent variable). In addition, analyses focused on participants' metaexperimental feedback, contributing to assessing the intervention's efficacy. Finally, exploratory analyses were aimed at examining the intervention's effectiveness specifically on female players, as the intersection between their prevalence in depression diagnoses and their prevalence in adopting play-like apps positions them to gain maximal benefit from the intervention. This is particularly important given the growing interest in the intersection of mental health and gaming. A recent literary review has shed light on the potential positive impact of video games specifically on women in terms of enhancing cognitive, social, and physical abilities [60]. The latter results are reported in Multimedia Appendix 1.

Sample Size and Power

A sample size of 100 (75 intervention and 25 WLC) was calculated using G*Power (version 3.1; Heinrich-Heine-Universität Düsseldorf), under the assumption of power of 0.80, for which a 2-tailed *t* test would be able to detect a large effect size (0.8) and a 1-tailed *t* test would be able to detect a moderate effect size (0.57). Aiming to detect a moderate effect size was chosen based on recent findings in online depression treatment research showing a 0.6 effect size in PHQ-9 improvement in internet-based cognitive behavioral therapy (CBT) interventions [61], and corresponding to a recent meta-analysis revealing a mean effect size of 0.67 across internet-based studies [62]. The effect size of 0.57 was chosen to detect a slightly smaller effect within the same expected moderate range.

Ethical Considerations

The study was ethically approved by the institutional review board of Mass General Hospital (approved protocol number 2022P002100), where the study was conducted. All participants provided informed consent before enrollment and were informed that they could opt out at any time. Data files, including participants' demographic information (eg, age and gender),

XSL•FO
RenderX

clinical measures collected by the clinical team as well as self-administered, and app use data, were kept separate from any identifying information, and all analyses were conducted using deidentified data. No deviations were made from the registered protocol. Participants were reimbursed via monthly payment card deposits or e-checks for the following components: baseline visit (first month only): US $25, visits with a study clinician (once every 4 weeks): US $10, and weekly questionnaires: US $5. Participants in the intervention group received an additional US $35 in the week they completed at least 75% of daily available game levels for 4 different games on at least 4 days of that week. Participants not meeting this weekly use and those in the WLC received US $9 per week. Data collection ended on July 26, 2023.

Results

Participants

Overview

A total of 348 eligible individuals underwent phone screening, and 163 (46.8%) of them met the inclusion criteria for completing a clinical MDD assessment. Of the 163 individuals, 117 (71.7%) were found to be eligible for the study and underwent randomization; of the 117 individuals, 87 (74.4%) were assigned to the intervention group and 30 (25.6%) to the WLC group. The mean age of the participants enrolled in the study was 32.43 (SD 8.5; range 20-50) years, 77.8% (91/117) were women, and 61.5% (72/117) were White. The study was conducted at Mass General Hospital. The primary diagnosis for all 117 participants was MDD, and the mean MADRS score at baseline was 26.97 (SD 4.65). There were no statistically significant differences between the 2 groups on any demographic or clinical variables. During the course of the study, participants were discontinued if they failed to reach the minimum app use requirements for >2 weeks, missing 25% of minimal engagement requirements, as well as if they failed to complete their monthly clinical assessments. Discontinued participants were removed from all analyses. By the end of the study, 13 (14.9%) of the 87 participants from the intervention group and 3 (10%) of the 30 participants from the WLC ended up either discontinuing due to technical and administrative reasons (n=5, 4.2%), personal reasons (n=1, 0.8%), or failure to meet minimum app playing requirements (n=10, 8.5%). Therefore, the intervention group consisted of 73.3% (74/101) participants, while the WLC group consisted 26.7% (27/101) participants (Figure 1). The demographic and clinical characteristics of the 86.3% (101/117) of the participants from both groups who successfully completed the study were not statistically different at baseline from the original groups and are described in Table 1 and Table S1 in Multimedia Appendix 1.

In line with the minimum app playing requirements, which required participants to engage in the app for at least 15 minutes a day for 4 days a week, the mean weekly minutes played was 60.32 (SE 2), showing on-par compliance and little variability across participants.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) table. MADRS: Montgomery-Åsberg Depression Rating Scale; MDD: major depressive disorder; MINI: Mini-International Neuropsychiatric Interview; OCD: obsessive-compulsive disorder; PTSD: posttraumatic stress disorder; SSN: social security number; WLC: waitlist control.

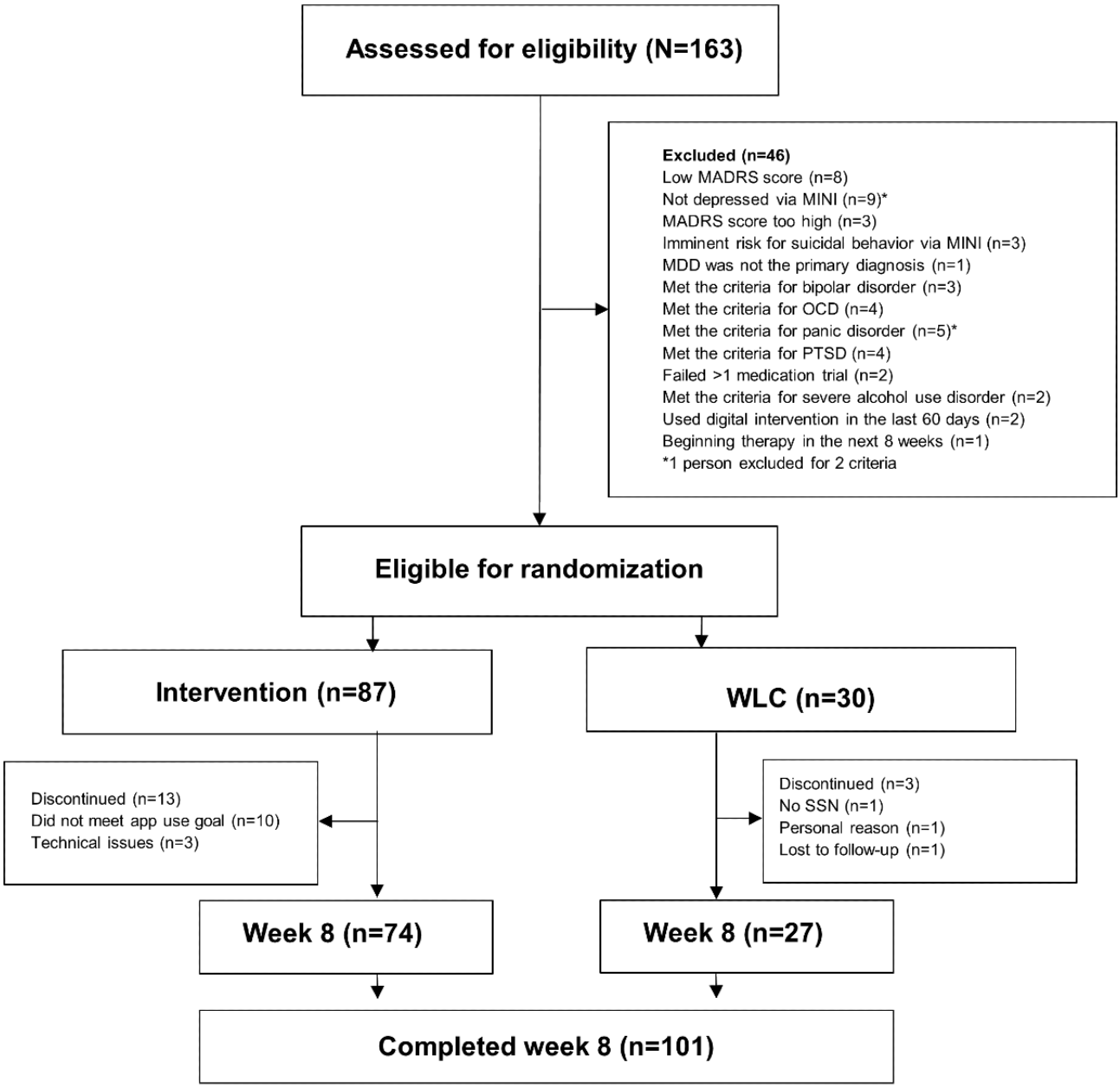

XSL•FO
RenderX

Table 1. Demographic and clinical characteristics at baseline.

	Intervention (n=74)	WLC[a] (n=27)	*P* value[b]
Demographic characteristics			
Age (y), mean (SD)	33.22 (8.6)	32.44 (8.53)	.69
Female, n (%)	57 (77)	21 (78)	.83
Ethnicity, n (%)			.56
American Indian or Alaska Native	2 (2)	0 (0)	
Asian	12 (16)	8 (30)	
Black or African American	8 (11)	2 (7)	
Native Hawaiian or other Pacific Islander	0 (0)	0 (0)	
White	48 (65)	16 (59)	
Prefer not to answer	4 (5)	1 (4)	
University level education, n (%)	58 (78)	18 (67)	.23
Married or domestic partners, n (%)	26 (35)	8 (30)	.6
Employment status, n (%)			.25
Employed	52 (70)	16 (60)	
Unemployed	13 (18)	4 (15)	
Student	9 (12)	7 (26)	
Annual household income (US $), n (%)			.18
≤19,999	5 (7)	3 (11)	
20,000-39,999	13 (18)	9 (33)	
40,000-59,999	13 (18)	3 (11)	
60,000-79,999	7 (9)	5 (18)	
80,000-99,999	11 (15)	1 (4)	
≥100,000	19 (26)	6 (22)	
Prefer not to answer	6 (8)	0 (0)	
English as the primary language, n (%)	71 (96)	23 (85)	.06
Clinical characteristics			
MDD[c] as the primary diagnosis, n (%)	74 (100)	27 (100)	__[d]
Secondary diagnosis, n (%)			.40
Suicide behavior disorder	3 (4)	1 (4)	
Panic disorder	5 (7)	2 (7)	
Social anxiety disorder	12 (16)	6 (22)	
Obsessive-compulsive disorder	0 (0)	2 (7)	
Alcohol use disorder	2 (3)	0 (0)	
GAD[e]	21 (28)	11 (41)	
MADRS[f] at baseline, mean (SD)	27.53 (4.71)	26.78 (4.13)	.47
PHQ-9[g] at baseline, mean (SD)	15.05 (4.59)	15.70 (3.48)	.51
GAD-7 at baseline, mean (SD)	9.27 (5.71)	10.52 (5.73)	.33
RRS[h] at baseline, mean (SD)	55.12 (11.63)	58.96 (7.99)	.07
SDQ[i] at baseline, mean (SD)	141.3 (18.26)	146.7 (14.68)	.17
PANAS-NA[j] at baseline, mean (SD)	23.72 (7.63)	25.93 (7.69)	.19

XSL•FO
RenderX

	Intervention (n=74)	WLC[a] (n=27)	*P* value[b]
PANAS-PA[k] at baseline, mean (SD)	17.78 (6.23)	20.15 (8.11)	.12
WHO-QoL-BREF[l] at baseline, mean (SD)	77.86 (10.05)	76.48 (9.63)	.54
SHAPS[m] at baseline, mean (SD)	22.93 (6.31)	22.48 (6.30)	.75
Receiving psychotherapy, MDD medication, or both n (%)	42 (56.8)	16 (59.3)	.82

[a]WLC: waitlist control group.

[b]*P* values indicate the significance level and are derived from either the student *t* test or chi-square test by comparing treatment conditions on binary variables. Data from the originally randomized 117 participants did not differ statistically and are presented in Table S1 in Multimedia Appendix 1.

[c]MDD: major depressive disorder.

[d]Not available.

[e]GAD: General Anxiety Disorder.

[f]MADRS: Montgomery-Åsberg Depression Rating Scale.

[g]PHQ-9: Patient Health Questionnaire-9.

[h]RRS: Ruminative Response Scale.

[i]SDQ: Symptoms of Depression Questionnaire.

[j]PANAS-NA: Positive Affect Negative Affect Scale-Negative Affect Score.

[k]PANAS-PA: Positive Affect Negative Affect Scale-Positive Affect Score.

[l]WHO-QoL-BREF: World Health Organization Quality of Life Brief Version.

[m]SHAPS: Snaith-Hamilton Pleasure Scale.

Playing the FTP-Based Game Leads to a Significant Reduction in Negative Clinical Scores Compared With Control Group

To assess the effectiveness of the intervention compared with the WLC, we conducted a mixed repeated measures ANOVA, separately for each questionnaire, implementing groups as the between-subject variable and time points (ie, baseline, week 4, and week 8) as the within-subject variable. MADRS analysis shows a significant interaction between groups and time points ($F_{97,2}$=7.79; *P*<.001), revealing that while there is no difference in MADRS scores between the groups at baseline (intervention: mean 27.52, SD 0.54 and WLC: mean 26.77, SD 0.79; t_{99}=0.77; *P*=.44), the intervention group shows significantly lower MADRS scores in the fourth and eighth weeks of the experiment compared with the WLC group (week 4 intervention: mean 18.64, SD 0.94 and WLC: mean 22.37, SD 1.54; t_{99}=–2.05; *P*=.04 and week 8 intervention: mean 15.7, SD 1.07, and WLC: mean 21.96, SD 1.75; t_{99}=–3.04; *P*=.004). To account for group size differences, an additional permutation test of significance was carried out. For each questionnaire, the intervention group's participants' data were subsampled 1000 times to fit the sampling size of the WLC group, and the statistical significance was recalculated. This allowed us to conservatively estimate the proportion of permutations showing a significant interaction effect between groups and time points. In the MADRS analysis, this randomized permutation procedure resulted in a significant interaction in 90% of permutations. These results suggest that the FTP-based intervention effectively reduced clinical symptoms to a greater degree than the WLC. Figure 2 depicts the elaborated results.

Follow-up analyses examined the difference between the groups in their measured clinical change from baseline after 4 weeks as well as after 8 weeks into the trial. The mean change from baseline in the MADRS score after 4 weeks was –8.87 (SE 0.95) in the intervention group and –4.4 (SE 1.4) in the WLC group (difference: –4.47, 95% CI –7.89 to –1.04; *P*=.01) and after 8 weeks was –11.82 (SE 1.12) in the intervention group and –4.81 (SE 1.47) in the WLC group (difference: –7.01, 95% CI –10.72 to –3.29; *P*<.001), indicating a significant difference between the groups in improvement in both time points. Of note, there was no significant difference in the MADRS score at baseline between the groups (difference: 0.75; *P*=.44). Together, these findings suggest the rate of improvement is significantly faster in the intervention group (refer to Figure 3 and Table S2 in Multimedia Appendix 1 for detailed results).

Further evaluation of the clinical benefit of the intervention was done by computing the SMD for each assessment tool. This calculation was based on Cohen effect size estimating the difference between the groups at the end of the study (week 8), in line with past studies evaluating the SMD of MDD improvement in mental health digital therapy apps [9]. The results of this analysis support the intervention's effectiveness over WLC in all assessment questionnaires, with the exception of the SHAPS questionnaire (Figure 4 [9]). The mean effect size across all questionnaires was –0.51, with larger effect sizes of –0.67 in the MADRS and –0.92 in the PHQ-9 questionnaires.

XSL•FO
RenderX

Figure 2. Greater improvement in clinical symptoms following facilitating thought progression–based intervention compared with waitlist control group (WLC). Clinical scores at the first, fourth, and eighth weeks in the intervention and WLC groups. The interaction results of a mixed-effects 3-way repeated measures ANOVA are reported. The proportion of significance in the randomized permutation procedure is indicated in parenthesis. (A) Montgomery-Åsberg Depression Rating Scale (MADRS): F97,2=7.79; *P*<.001 (90% permutation significance). (B) Patient Health Questionnaire-9 (PHQ-9): F95,2=7.17; *P*<.001 (88% permutation significance). (C) General Anxiety Disorder-7 (GAD-7): F95,2=1.75; *P*=.17 (10% permutation significance). (D) Rumination Response Scale (RRS): F97,2=5.67; *P*=.004 (68% permutation significance). (E) Symptoms of Depression Questionnaire (SDQ): F97,2=4.26; *P*=.01 (57% permutation significance). (F) Positive Affect Negative Affect Scale-Negative Affect Score (PANAS-NA): F95,2=1.51; *P*=.22 (6% permutation significance). (G) Positive Affect Negative Affect Scale-Positive Affect Score (PANAS-PA): F95,2=5.75; *P*=.003 (73% permutation significance). (H) World Health Organization Quality of Life Brief Version (WHO-QoL-BREF): F95,2=4.26; *P*=.01 (29% permutation significance). (I) Snaith-Hamilton Pleasure Scale (SHAPS): F97,2=1.08; *P*=.34 (1% permutation significance).

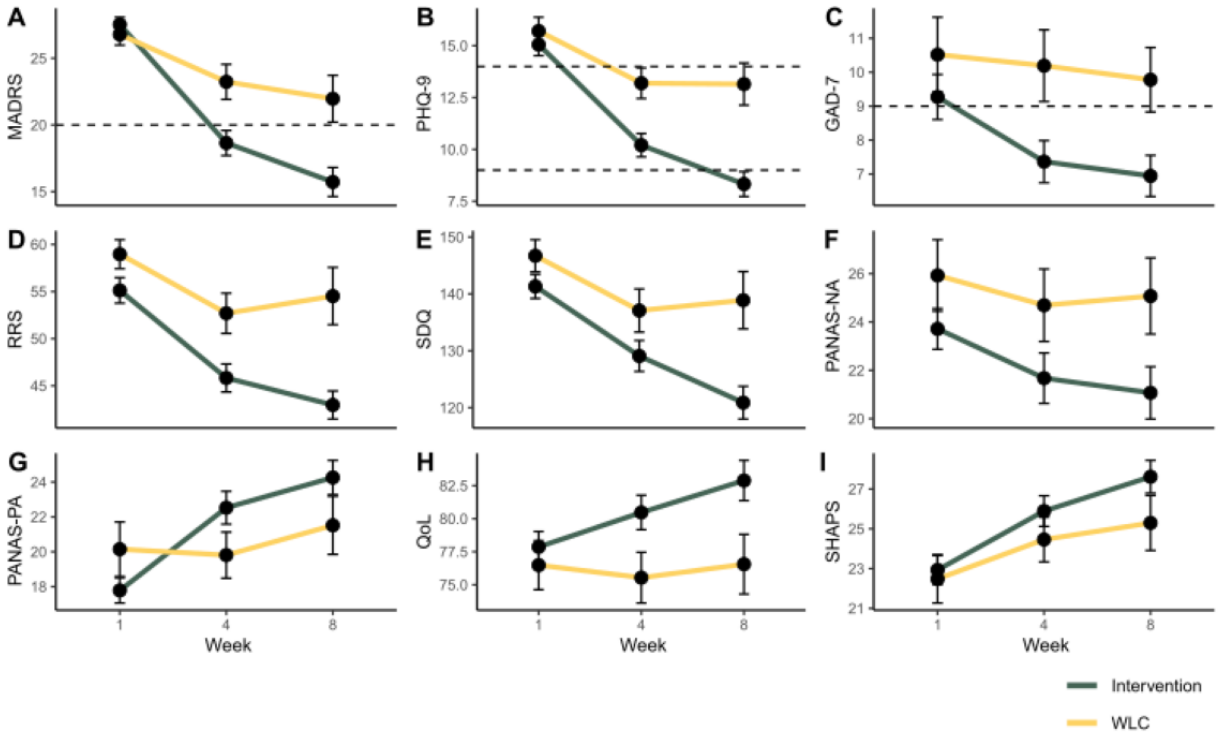

XSL•FO
RenderX

Figure 3. Enhanced improvement in clinical symptoms following the facilitating thought progression (FTP)–based intervention compared with a waitlist control group (WLC). The figure depicts the percentage change from baseline after 4 weeks (A) as well as after 8 weeks (B) of enrolling in the clinical trial in either the FTP-based intervention or the WLC groups. Asterisks denote significant differences in the percentage change from baseline between the intervention and the WLC groups. *P≤.01, **P≤.05, ***P≤.001. Error bars denote SEM. Results are in a normalized format. GAD-7: General Anxiety Disorder-7; MADRS: Montgomery-Åsberg Depression Rating Scale; PANAS-NA: Positive Affect Negative Affect Scale-Negative Affect Score; PANAS-PA: Positive Affect Negative Affect Scale-Positive Affect Score; PHQ-9: Patient Health Questionnaire-9; RRS: Rumination Response Scale; SHAPS: Snaith-Hamilton Pleasure Scale; SDQ: Symptoms of Depression Questionnaire; WHO-QoL-BREF: World Health Organization Quality of Life Brief Version.

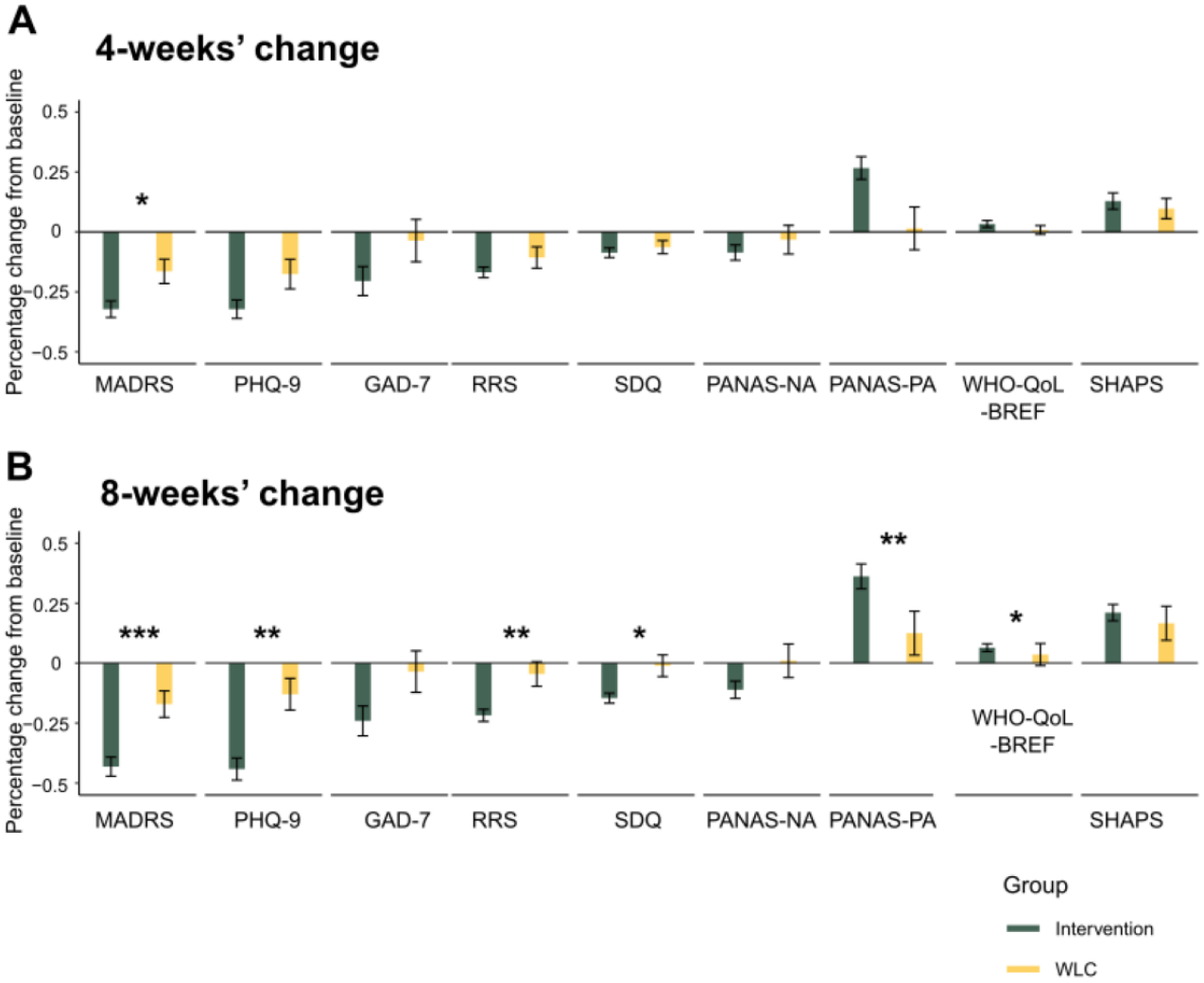

XSL•FO
RenderX

Figure 4. The effect of the facilitating thought progression (FTP)–based app intervention on depression. Dots denote Cohen d effect size. Negative values provide evidence in support of the intervention's efficiency, while positive values provide evidence in support of waitlist efficiency. Horizontal lines denote CIs. Mean effect sizes for each questionnaire are as follows: Montgomery-Åsberg Depression Rating Scale (MADRS)=−0.67, Patient Health Questionnaire-9 (PHQ-9)=−0.92, General Anxiety Disorder-7 (GAD-7)=−0.54, Rumination Response Scale (RRS)=−0.84, Symptoms of Depression Questionnaire (SDQ)=−0.71, Positive Affect Negative Affect Scale-Negative Affect Score (PANAS-NA)=−0.44, Positive Affect Negative Affect Scale- Positive Affect Score (PANAS-PA)=−0.31, World Health Organization Quality of Life Brief Version (WHO-QoL-BREF)=−0.49, Snaith-Hamilton Pleasure Scale (SHAPS)=0.32. The green triangle denotes the mean effect size across all questionnaires. The pink triangle denotes the mean effect size across all studies reported by Lee et al [9].

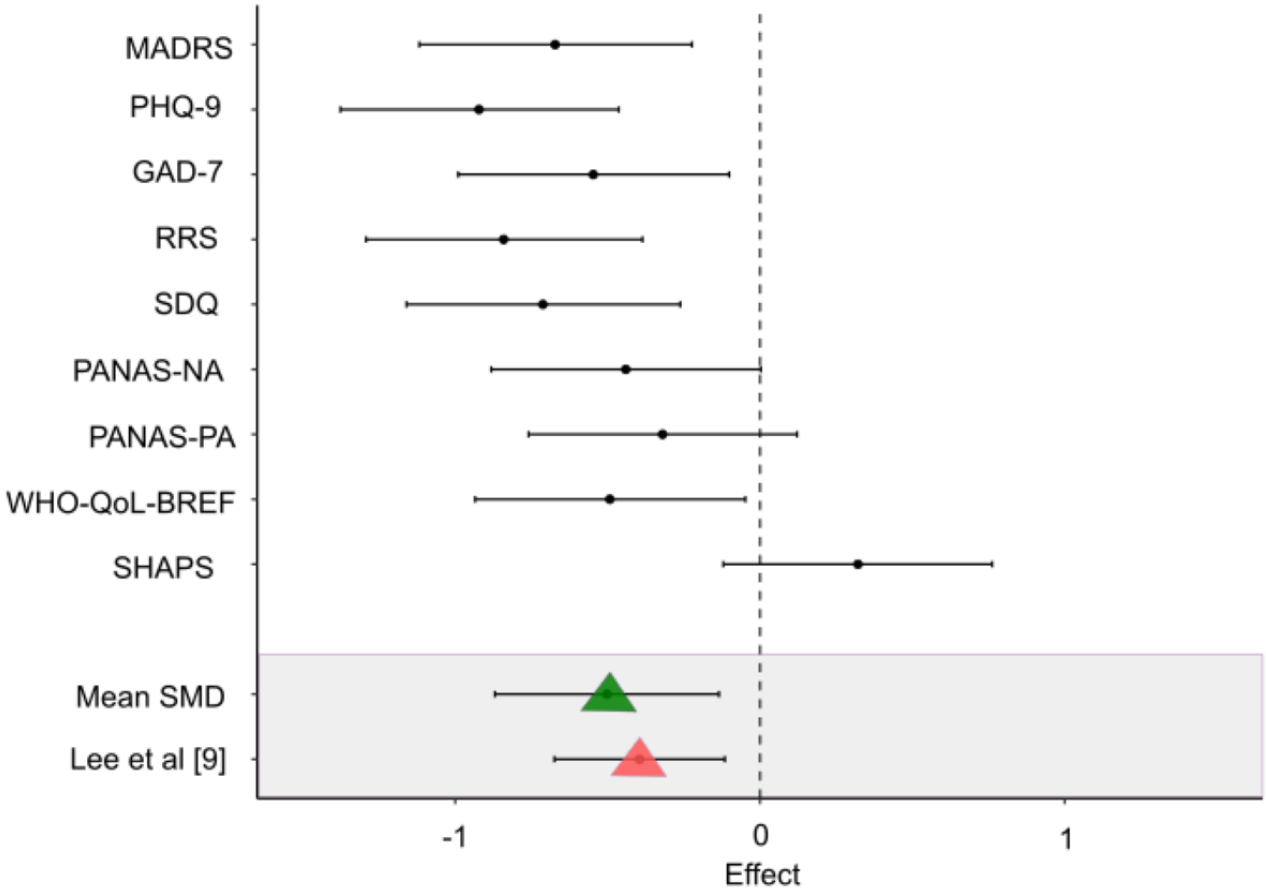

Playing an FTP-Based Game Leads to a Significantly Faster Reduction in Negative Clinical Scores Compared With Control Group

To assess the rate of week-by-week clinical improvement, mixed-effects linear models were composed for each clinical questionnaire. These models included all weeks as well as group as fixed effects and participants as a random effect and were found statistically significant for all 9 questionnaires. A significant group effect was found for 6 (66%) out of 9 questionnaires, suggesting that the rates of clinical improvement across these assessments were faster for playing participants compared with WLC (Table 2).

XSL•FO
RenderX

Table 2. Rate of improvement in clinical symptoms following facilitating thought progression–based intervention compared with waitlist control group[a].

Assessment	β (SE; 95% CI)	*t* test (*df*)	*P* value	Effect favors intervention significantly	Model *P* values
MADRS[b]					<.001
Time	–1.24 (0.1; –1.45 to –1.03)	–11.74 (298,5)	<.001	__[c]	
Group	3.07 (1.33; 0.45 to 5.7)	2.3 (298,5)	.02	Yes	
PHQ-9[d]					<.001
Time	–.64 (0.04; –0.72 to –0.56)	–16.2 (298,5)	<.001	—	
Group	3.13 (0.85; 1.45 to 4.81)	3.65 (298,5)	<.001	Yes	
GAD-7[e]					<.001
Time	–.29 (0.03; –0.36 to –0.21)	–7.69 (298,5)	<.001	—	
Group	2.07 (1.03; 0.03 to 4.1)	1.99 (298,5)	.04	Yes	
RRS[f]					<.001
Time	–1.32 (0.9; –1.5 to –1.14)	–14.55 (298,5)	<.001	—	
Group	8.03 (2.41; 3.29 to –12.76)	3.32 (298,5)	<.001	Yes	
SDQ[g]					<.001
Time	–2.13 (0.28; –2.7 to –1.56)	–7.35 (298,5)	<.001	—	
Group	10.41 (3.88; 2.76 to 18.05)	2.68 (298,5)	.007	Yes	
PANAS-NA[h]					<.001
Time	–.35 (0.05; –0.51 to –0.27)	–6.27 (298,5)	<.001	—	
Group	3.46 (1.64; 0.24 to 6.68)	2.11 (298,5)	.03	Yes	
PANAS-PA[i]					<.001
Time	.56 (0.06; 0.44 to 0.68)	9.09 (298,5)	<.001	—	
Group	–.99 (1.4; –3.74 to 1.75)	–0.71 (298,5)	.47	No	
WHO-QoL-BREF[j]					<.001
Time	.46 (0.11; 0.23 to 0.68)	4.03 (298,5)	<.001	—	
Group	–4.08 (2.21; –8.44 to 0.27)	–1.84 (298,5)	.06	No	
SHAPS[k]					<.001
Time	.46 (0.05; 0.36 to 0.57)	8.61 (298,5)	<.001	—	
Group	–1.03 (1.27; –3.53 to 1.46)	–0.81 (298,5)	.41	No	

[a]Mixed-effects general linear models were conducted for each clinical measurement, assessing the rate of clinical symptom change as a function of time (weeks into the clinical trial) and group (playing participants vs waitlist control group). Participants were included in the models as a random variable. Each model's parameters are reported for the different time and group effects. All significant group effects favor the intervention group. Each model's *P* values are false discovery rate–corrected for multiple comparisons.

[b]MADRS: Montgomery-Åsberg Depression Rating Scale.

[c]Not available.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]GAD-7: General Anxiety Disorder-7.

[f]RRS: Ruminative Response Scale.

[g]SDQ: Symptoms of Depression Questionnaire.

[h]PANAS-NA: Positive Affect Negative Affect Scale-Negative Affect Score.

[i]PANAS-PA: Positive Affect Negative Affect Scale-Positive Affect Score.

[j]WHO-QoL-BR EF: World Health Organization Quality of Life Brief Version.

[k]SHAPS: Snaith-Hamilton Pleasure Scale.

XSL•FO
RenderX

Secondary Outcomes

Posttrial Follow-Up in Week 12

While the study required 8 weeks of engagement, participants could choose to continue for an additional 4 weeks. During this additional phase, participants were compensated for completing the assessment but not for engaging with the app. In total, 90% (58/64) of the participants who were offered the option chose to continue and completed 12 weeks of app use. Though the intervention group contained 73.3% (74/101) of participants who reached the end of week 8, only participants who were recruited before May 2023 were offered to continue playing for another 4 weeks, due to time limitations. In total, 58% (43/74) of the participants from the intervention group and 56% (15/27) of the participants from the WLC group completed a MADRS assessment and filled out self-assessment questionnaires at week 12. In total, 16% (12/74) of the participants from the intervention group maintained the high playing regiment of at least 4 days a week (henceforth HE group), and 42% (31/74) of the participants kept playing but did not maintain regiment-level engagement (ie, missed at least 1 week of playing). These players used the app for 8 (SD 7) minutes throughout the postintervention period (henceforth LE group). In total, 15 (56%) participants from the WLC group underwent MADRS assessment at week 12, out of which 6 (22%) chose to also play the game. These participants played for 24 (SD 10) minutes throughout the 4 (range 12-37) weeks, notably below the minimum 60 minutes per week engagement criteria. The remaining 22% (6/27) of the participants did not play the game at all. Analysis shows that although similar in their baseline MADRS score, participants in the HE group differed from the LE and WLC groups in their clinical scores at week 8, such that the HE group exhibited the lowest MADRS scores at week 8, compared to their LE and WLC group counterparts ($F_{49,2}$=7.14; P=.002; colored lines in Figure 5A). This difference between the groups was preserved at week 12 ($F_{49,2}$=6.62; P=.003; colored bars in Figure 5A), as each group maintained an equivalent level to the scores obtained at week 8, suggesting a possible long-lasting effect for the LE group. Similar results were found for the PHQ-9 (week 8: $F_{52,2}$=8.48; P<.001 and week 12: $F_{52,2}$=8.57; P<.001; Figure 5B). These results point to the intervention's possible long-lasting effect as well as suggest that participants for whom the intervention was most effective voluntarily continued to play the app.

Figure 5. Group differences 1 month after the intervention. (A) Montgomery-Åsberg Depression Rating Scale (MADRS) scores of participants at week 12. (B) Patient Health Questionnaire-9 (PHQ-9) scores of participants at week 12. Bars depict the week-12 scores of high-engagement participants (HE; dark-green bar), low-engagement participants (LE; light-green bar), and waitlist control group (WLC) participants (yellow bar). Horizontal lines depict individual groups' mean scores at week 8 for reference.

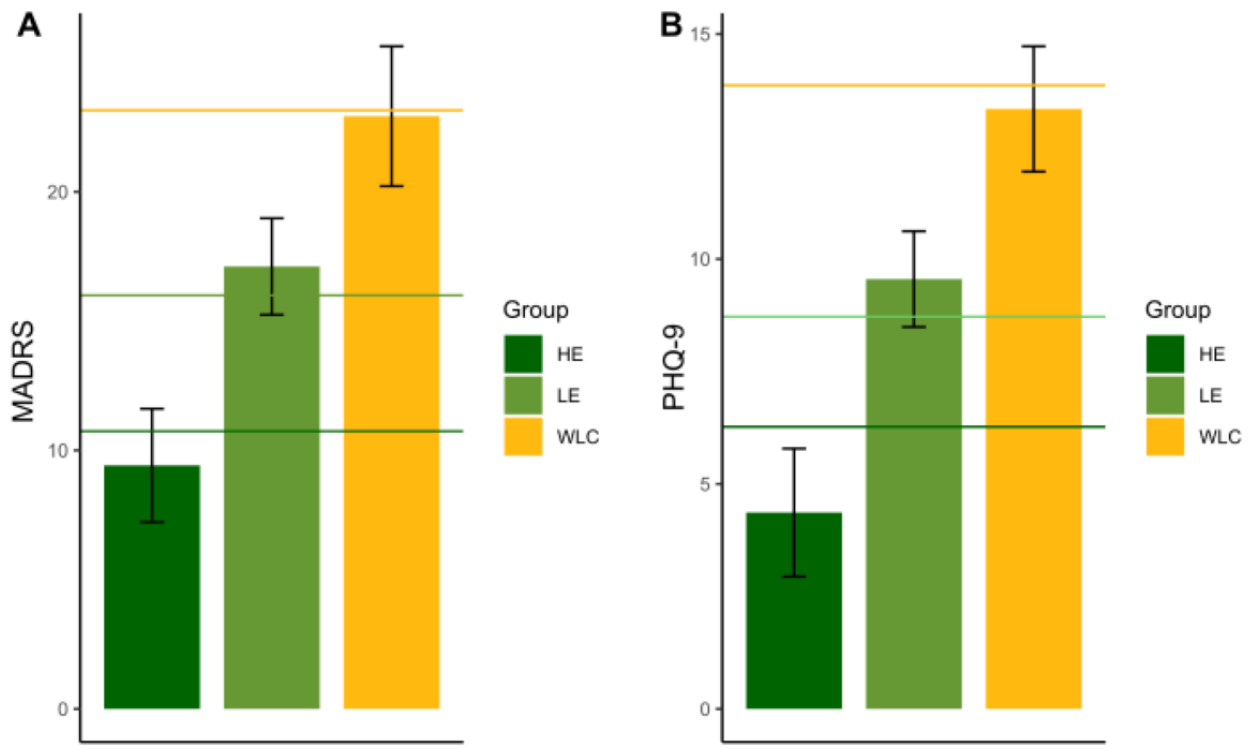

Subjective Engagement and Benefits

With respect to participant feedback, 47% (35/74) of the participants reported that they saw themselves using the app at least once a week if the app was to be made publicly available. In addition, a high percentage of the participants in the intervention group reported either "agree" or "strongly agree" to the following statements: "Use of this app is likely to decrease low mood and sadness" (34/74, 46%), "The app was easy to use" (70/74, 94%), "Playing this app was fun" (55/74, 74%), as well as "This app is likely to increase awareness of the importance of addressing low mood and sadness" (32/74, 43%) and "This app is likely to increase intentions or motivation to address low mood and sadness" (38/74, 51%), suggesting high

XSL•FO
RenderX

subjective perceived benefits. In total, 53% (40/74) of the participants gave the app either 4 or 5 ratings on a 1 to 5 scale, 1 being very negative and 5 being very positive views of the app. The mean score was 3.59, and the median score was 4 (IQR 3-4).

Discussion

Principal Findings

This RCT aimed to test the efficacy of a digital health intervention, which was designed to facilitate thought progression, on alleviating symptoms of depression over the course of 8 weeks. This intervention stems from the theoretical grounds showing that flexible, fast, broad, creative, and global thinking patterns are bidirectionally tied with mood. On the one hand, they are dampened among people experiencing negative mood disorders, but on the other hand, practicing and instigating these patterns may improve negative mood [38].

The primary outcomes in this trial show that the rate of clinical improvement in the intervention group was faster, and its overall magnitude was significantly larger compared with that of the WLC. This was true both for several main gold standard measurements for depression, most importantly MADRS and PHQ-9, as well as for General Anxiety Disorder-7, which assessed the improvement in anxiety symptoms. The effect size of the results was on par with previous results found in other existing apps for depression. Secondary outcomes showed that the positive clinical effect persisted for at least 1 month after the trial and that the intervention was perceived as fun, engaging, and beneficial. Overall, these findings demonstrate the efficacy of an FTP-based app in the long-term reduction of depressive symptoms.

Of note, other existing treatments have focused on the existence of maladaptive thought in depression. For example, CBT focuses on identifying negative, irrational, or distorted thoughts and encourages patients to assess their validity and replace them with more beneficial alternatives [63,64]. However, while the CBT intervention aims at teaching strategies to change the content of thought [65,66], the FTP-based approach focuses on changing the structure and pattern of thoughts by having patients engage in cognitive exercises that are independent from thought content. Therefore, the FTP approach is agnostic to the content of the thought or to the level to which one's thought content corresponds with realistic truth and rather focuses on the structure of thought patterns and how they progress. Therefore, it is likely that an intervention based on the FTP framework could be a complement to CBT.

Beyond its relation to CBT, this study aligns well with the accumulating literature showing sustained beneficial effects of online digital interventions in mental health [11-14]. These found that mobile apps may have the greatest impact on people with mild to moderate depression, which is the audience targeted in this study. These studies also show that video games facilitate cognitive control and multitasking, accompanied by attention-related neural changes, which suggest that digital interventions may lead not only to externally measured improvement but to changes in the underlying mechanisms as well. While the effect of the FTP-based intervention on neural processes remains to be tested, the significant findings in this study corroborate this literature by showing significant improvement in depressive and anxiety symptoms following consistent cognitive training that aims to sustain healthy thinking patterns.

Theoretical Implications and Future Directions

The main results indicated that participants in the intervention group showed faster and greater clinical improvement compared with WLC. This improvement was further sustained for 4 weeks after the critical period, and overall high compliance was assessed, supporting the efficacy of the intervention and its potential to engage other clinical populations in need. This is the first study showing that an app based on the idea of the FTP framework is associated with improvement of depression. Therefore, altering thought structure may be as important as changing the content of thought. The findings support the notion that practicing faster, broader, and more associative thought patterns may alleviate depression. Future studies may examine whether applying this intervention in parallel to other psychotherapeutic interventions, such as CBT, as suggested earlier, or even classical psychotherapy, boosts their anticipated effect. It is possible that engaging in CBT-like treatments, which focus on rerouting the content of negative thought, conjointly with engaging with the FTP-based digital intervention, which focuses on the structural (eg, breadth) and dynamic-related (eg, pace) aspects of thought, described and tested in this study, may show promising results in targeting depression.

The intervention discussed in this study provides proof of concept for the feasibility of an FTP-based digital intervention to clinically aid populations with depression. Future studies are required to understand how each specific component within the FTP approach impacts depression. For example, fostering broad associative thoughts may have a specific positive effect on rumination reduction, while fast thinking may instate a needed sense of progress. The combination between these elements may also lead to a unique effect on depressive symptoms. Given that this trial evaluated engagement of all 5 therapeutic "minigames," studies that will further dissociate the FTP approach to its individual therapeutic elements may provide insight into whether different elements can be tailored to different patients with depression for maximal clinical improvement.

Limitations

This study has several limitations that would require future work to mitigate. First, the experimental design does not allow us to discern whether increased app engagement accelerates the reduction in depressive symptoms. The study required all participants in the intervention group to engage in the FTP-based app for a minimum of 15 minutes daily for 8 weeks, and most participants requested to continue playing for 4 weeks more. However, little variability was observed in the overall duration of engagement, leaving the question open as to whether more extensive engagement would provide additional benefits. Future studies in which participants are free to play as much or as little as they like would help fine-tune the parameters of optimal

XSL•FO
RenderX

engagement (eg, daily duration and number of levels solved) that will maximize efficacy in alleviating depression symptoms.

Second, another possible limitation is the cohort's low ethnic and socioeconomic diversity. Future research should involve better diversifying strategies to improve the ability to conclude efficacy for the broader MDD population.

Third, the study lacks a sham app control group to control for possible placebo effects. For example, while the reported significant changes in MADRS scores in both the intervention and the WLC groups may have resulted from the placebo effects related to being assessed by professional clinicians, the boosted improvement in MADRS in the intervention group may have resulted from an additional placebo effect caused by mere app engagement. A sham app would have controlled for this potential placebo effect, making the study a true double-blind design. Considering this limitation, follow-up examinations evaluated the possibility that our findings were driven by placebo effects, revealing several key insights into why a mere placebo account is unlikely. First, improvement was not observed on all assessment tools, as SHAPS scores did not improve for either group. Second, in an exploratory analysis, it was found that for women, the intervention was more effective as age increased (Multimedia Appendix 1 [67-71]). Together, these findings provide additional support for the specific efficacy of the app for most clinical measurements, though not all, and suggest that its effectiveness may vary more significantly among certain populations. Third, a concern for a placebo effect is generally attributed, in major part, to the possible worsening of symptoms among the WLC, which may eventually account for misattributed beneficial intervention results. In this study, however, the WLC showcased improvement in symptoms, yet it was outperformed by the intervention group, which further supports the app's effectiveness. Finally, while the RCT did not involve a sham app, the SMD score of all 9 clinical measurements, and specifically the SMD of MADRS, fell well on the high end of effective measures compared with other digital therapeutic RCTs. This analysis suggests that the intervention method developed and studied in this trial is comparable in its magnitude of depressive symptoms reduction to other recently developed online treatments, making a pure placebo account highly unlikely.

Conclusions

Within the global context of the rising need for innovative solutions that effectively alleviate depression and anxiety, a digital intervention that was designed to facilitate thought progression was found to effectively reduce depressive symptoms across multiple measurements in this study. It was also found to be appealing to patients with MDD, who showed interest in further engagement. These findings support the implementation of FTP-based interventions in future therapeutic efforts or along with other forms of therapeutic intervention, serving the effective facilitation of widely accessible mental health solutions. The findings further bolster the promising potential of digital health solutions to bridge the increasing gap in access to effective treatments for depression and anxiety.

Acknowledgments

The study was funded by Hedonia Ltd.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request. The up-to-date app version can be accessed via the MoodBloom app, which is available on the App Store or Google Play Store.

Authors' Contributions

MB and SLY designed the clinical trial. PP, SH, and JC oversaw participant recruitment and acquired clinical data at Mass General Hospital. SLY, SB, NS, and BS conducted data analyses. SLY, SB, and MB wrote the initial drafts of the paper. All authors approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Illustration of the gamified web-based app used by participants in the intervention group. This game includes 5 minigames aimed to facilitate participant's thought progression.
[DOCX File , 722 KB - jmir_v26i1e56201_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1072 KB - jmir_v26i1e56201_app2.pdf]

References

XSL•FO
RenderX

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Association; 2013.
3. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
4. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 06;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
5. Bae H, Shin H, Ji HG, Kwon JS, Kim H, Hur JW. App-based interventions for moderate to severe depression: a systematic review and meta-analysis. JAMA Netw Open 2023 Nov 01;6(11):e2344120 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.44120] [Medline: 37983028]
6. Khanna MS, Kendall PC. Computer-assisted cognitive behavioral therapy for child anxiety: results of a randomized clinical trial. J Consult Clin Psychol 2010 Oct;78(5):737-745. [doi: 10.1037/a0019739] [Medline: 20873909]
7. Mohr DC, Vella L, Hart S, Heckman T, Simon G. The effect of telephone-administered psychotherapy on symptoms of depression and attrition: a meta-analysis. Clin Psychol (New York) 2008;15(3):243-253 [FREE Full text] [doi: 10.1111/j.1468-2850.2008.00134.x] [Medline: 21369344]
8. Muller I, Yardley L. Telephone-delivered cognitive behavioural therapy: a systematic review and meta-analysis. J Telemed Telecare 2011 Feb 28;17(4):177-184. [doi: 10.1258/jtt.2010.100709] [Medline: 21357672]
9. Lee S, Oh JW, Park KM, Lee S, Lee E. Digital cognitive behavioral therapy for insomnia on depression and anxiety: a systematic review and meta-analysis. NPJ Digit Med 2023 Mar 25;6(1):52 [FREE Full text] [doi: 10.1038/s41746-023-00800-3] [Medline: 36966184]
10. Anguera JA, Jordan JT, Castaneda D, Gazzaley A, Areán PA. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. BMJ Innov 2016 Jan 04;2(1):14-21 [FREE Full text] [doi: 10.1136/bmjinnov-2015-000098] [Medline: 27019745]
11. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]
12. Anguera JA, Boccanfuso J, Rintoul JL, Al-Hashimi O, Faraji F, Janowich J, et al. Video game training enhances cognitive control in older adults. Nature 2013 Sep 05;501(7465):97-101 [FREE Full text] [doi: 10.1038/nature12486] [Medline: 24005416]
13. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016 Dec 20;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
14. Morimoto SS, Altizer RA, Gunning FM, Hu W, Liu J, Cote SE, et al. Targeting cognitive control deficits with neuroplasticity-based computerized cognitive remediation in patients with geriatric major depression: a randomized, double-blind, controlled trial. Am J Geriatr Psychiatry 2020 Sep;28(9):971-980. [doi: 10.1016/j.jagp.2020.05.023] [Medline: 32591170]
15. Conner KO, Copeland VC, Grote NK, Koeske G, Rosen D, Reynolds CF, et al. Mental health treatment seeking among older adults with depression: the impact of stigma and race. Am J Geriatr Psychiatry 2010 Jun;18(6):531-543. [doi: 10.1097/jgp.0b013e3181cc0366]
16. Alahäivälä T, Oinas-Kukkonen H. Understanding persuasion contexts in health gamification: a systematic analysis of gamified health behavior change support systems literature. Int J Med Inform 2016 Dec;96:62-70. [doi: 10.1016/j.ijmedinf.2016.02.006] [Medline: 26944611]
17. Eichenberg C, Schott M. Serious games for psychotherapy: a systematic review. Games Health J 2017 Jun;6(3):127-135. [doi: 10.1089/g4h.2016.0068] [Medline: 28628385]
18. Fitzgerald M, Ratcliffe G. Serious games, gamification, and serious mental illness: a scoping review. Psychiatr Serv 2020 Feb 01;71(2):170-183. [doi: 10.1176/appi.ps.201800567] [Medline: 31640521]
19. Merry SN, Stasiak K, Shepherd M, Frampton C, Fleming T, Lucassen MF. The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: randomised controlled non-inferiority trial. BMJ 2012 Apr 18;344(apr18 3):e2598 [FREE Full text] [doi: 10.1136/bmj.e2598] [Medline: 22517917]
20. Kim Y, Hong S, Choi M. Effects of serious games on depression in older adults: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Sep 06;24(9):e37753 [FREE Full text] [doi: 10.2196/37753] [Medline: 36066964]
21. Ruiz M, Moreno M, Girela-Serrano B, Díaz-Oliván I, Muñoz LJ, González-Garrido C, et al. Winning the game against depression: a systematic review of video games for the treatment of depressive disorders. Curr Psychiatry Rep 2022 Jan 03;24(1):23-35 [FREE Full text] [doi: 10.1007/s11920-022-01314-7] [Medline: 35113313]

XSL•FO
RenderX

22. Townsend C, Humpston C, Rogers J, Goodyear V, Lavis A, Michail M. The effectiveness of gaming interventions for depression and anxiety in young people: systematic review and meta-analysis. BJPsych Open 2022 Jan 07;8(1):e25 [FREE Full text] [doi: 10.1192/bjo.2021.1078] [Medline: 34991767]
23. Kowal M, Conroy E, Ramsbottom N, Smithies T, Toth A, Campbell M. Gaming your mental health: a narrative review on mitigating symptoms of depression and anxiety using commercial video games. JMIR Serious Games 2021 Jun 16;9(2):e26575 [FREE Full text] [doi: 10.2196/26575] [Medline: 34132648]
24. Pine R, Fleming T, McCallum S, Sutcliffe K. The effects of casual videogames on anxiety, depression, stress, and low mood: a systematic review. Games Health J 2020 Aug;9(4):255-264. [doi: 10.1089/g4h.2019.0132] [Medline: 32053021]
25. Villani D, Carissoli C, Triberti S, Marchetti A, Gilli G, Riva G. Videogames for emotion regulation: a systematic review. Games Health J 2018 Apr;7(2):85-99 [FREE Full text] [doi: 10.1089/g4h.2017.0108] [Medline: 29424555]
26. Blackburn TP. Depressive disorders: treatment failures and poor prognosis over the last 50 years. Pharmacol Res Perspect 2019 Jun 03;7(3):e00472 [FREE Full text] [doi: 10.1002/prp2.472] [Medline: 31065377]
27. Boschloo L, Hieronymus F, Lisinski A, Cuijpers P, Eriksson E. The complex clinical response to selective serotonin reuptake inhibitors in depression: a network perspective. Transl Psychiatry 2023 Jan 21;13(1):19 [FREE Full text] [doi: 10.1038/s41398-022-02285-2] [Medline: 36681669]
28. Cuijpers P, Pineda BS, Quero S, Karyotaki E, Struijs SY, Figueroa CA, et al. Psychological interventions to prevent the onset of depressive disorders: a meta-analysis of randomized controlled trials. Clin Psychol Rev 2021 Feb;83:101955 [FREE Full text] [doi: 10.1016/j.cpr.2020.101955] [Medline: 33333441]
29. Leichsenring F, Steinert C, Rabung S, Ioannidis JP. The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. World Psychiatry 2022 Feb;21(1):133-145 [FREE Full text] [doi: 10.1002/wps.20941] [Medline: 35015359]
30. Casacalenda N, Perry JC, Looper K. Remission in major depressive disorder: a comparison of pharmacotherapy, psychotherapy, and control conditions. Am J Psychiatry 2002 Aug;159(8):1354-1360. [doi: 10.1176/appi.ajp.159.8.1354] [Medline: 12153828]
31. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry 2006 Jan;163(1):28-40. [doi: 10.1176/appi.ajp.163.1.28] [Medline: 16390886]
32. Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry 2006 Nov;163(11):1905-1917. [doi: 10.1176/ajp.2006.163.11.1905] [Medline: 17074942]
33. Puyat JH, Kazanjian A, Goldner EM, Wong H. How often do individuals with major depression receive minimally adequate treatment? A population-based, data linkage study. Can J Psychiatry 2016 Mar 24;61(7):394-404. [doi: 10.1177/0706743716640288]
34. Greer TL, Kurian BT, Trivedi MH. Defining and measuring functional recovery from depression. CNS Drugs 2010 Apr;24(4):267-284. [doi: 10.2165/11530230-000000000-00000] [Medline: 20297853]
35. Insel TR, Wang PS. The STAR*D trial: revealing the need for better treatments. Psychiatr Serv 2009 Nov;60(11):1466-1467. [doi: 10.1176/ps.2009.60.11.1466] [Medline: 19880463]
36. Bar M. From objects to unified minds. Curr Dir Psychol Sci 2021 Mar 10;30(2):129-137. [doi: 10.1177/0963721420984403]
37. Bar M. A cognitive neuroscience hypothesis of mood and depression. Trends Cogn Sci 2009 Nov;13(11):456-463 [FREE Full text] [doi: 10.1016/j.tics.2009.08.009] [Medline: 19819753]
38. Herz N, Baror S, Bar M. Overarching states of mind. Trends Cogn Sci 2020 Mar;24(3):184-199. [doi: 10.1016/j.tics.2019.12.015] [Medline: 32059121]
39. Akbari Chermahini S, Hommel B. More creative through positive mood? Not everyone!. Front Hum Neurosci 2012;6:319 [FREE Full text] [doi: 10.3389/fnhum.2012.00319] [Medline: 23189050]
40. Baas M, De Dreu CK, Nijstad BA. A meta-analysis of 25 years of mood-creativity research: hedonic tone, activation, or regulatory focus? Psychol Bull 2008 Nov;134(6):779-806. [doi: 10.1037/a0012815] [Medline: 18954157]
41. Pronin E, Jacobs E. Thought speed, mood, and the experience of mental motion. Perspect Psychol Sci 2008 Nov 01;3(6):461-485. [doi: 10.1111/j.1745-6924.2008.00091.x] [Medline: 26158973]
42. Kühn S, Vanderhasselt MA, De Raedt R, Gallinat J. Why ruminators won't stop: the structural and resting state correlates of rumination and its relation to depression. J Affect Disord 2012 Dec 10;141(2-3):352-360. [doi: 10.1016/j.jad.2012.03.024] [Medline: 22497878]
43. Zuo N, Fang J, Lv X, Zhou Y, Hong Y, Li T, et al. White matter abnormalities in major depression: a tract-based spatial statistics and rumination study. PLoS One 2012 May 30;7(5):e37561 [FREE Full text] [doi: 10.1371/journal.pone.0037561] [Medline: 22666366]
44. Harel EV, Tennyson RL, Fava M, Bar M. Linking major depression and the neural substrates of associative processing. Cogn Affect Behav Neurosci 2016 Dec 23;16(6):1017-1026 [FREE Full text] [doi: 10.3758/s13415-016-0449-9] [Medline: 27553369]
45. de Fockert JW, Cooper A. Higher levels of depression are associated with reduced global bias in visual processing. Cogn Emot 2014 Apr 25;28(3):541-549. [doi: 10.1080/02699931.2013.839939] [Medline: 24067089]

XSL•FO
RenderX

46. Gasper K. Do you see what I see? Affect and visual information processing. Cognit Emotion 2004 Apr;18(3):405-421. [doi: 10.1080/02699930341000068]
47. Fredrickson BL, Branigan C. Positive emotions broaden the scope of attention and thought-action repertoires. Cogn Emot 2005 May 01;19(3):313-332 [FREE Full text] [doi: 10.1080/02699930441000238] [Medline: 21852891]
48. Mason MF, Bar M. The effect of mental progression on mood. J Exp Psychol Gen 2012 May;141(2):217-221 [FREE Full text] [doi: 10.1037/a0025035] [Medline: 21823806]
49. Rally with Mass General Brigham. Mass General Brigham. URL: https://rally.massgeneralbrigham.org/ [accessed 2024-10-25]
50. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33;quiz 34. [Medline: 9881538]
51. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979 Apr 01;134(4):382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
52. Ntini I, Vadlin S, Olofsdotter S, Ramklint M, Nilsson KW, Engström I, et al. The Montgomery and Åsberg depression rating scale - self-assessment for use in adolescents: an evaluation of psychometric and diagnostic accuracy. Nord J Psychiatry 2020 Aug 03;74(6):415-422. [doi: 10.1080/08039488.2020.1733077] [Medline: 32125211]
53. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
54. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
55. Treynor W, Gonzalez R, Nolen-Hoeksema S. Rumination reconsidered: a psychometric analysis. Cognit Therapy Res 2003:247-259. [doi: 10.1023/A:1023910315561]
56. Snaith RP, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. Br J Psychiatry 1995 Jul;167(1):99-103. [doi: 10.1192/bjp.167.1.99] [Medline: 7551619]
57. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 1988;54(6):1063-1070. [doi: 10.1037/0022-3514.54.6.1063]
58. Development of the World Health Organization WHOQOL-BREF quality of life assessment. The WHOQOL Group. Psychol Med 1998 May;28(3):551-558. [doi: 10.1017/s0033291798006667] [Medline: 9626712]
59. Pedrelli P, Blais MA, Alpert JE, Shelton RC, Walker RS, Fava M. Reliability and validity of the Symptoms of Depression Questionnaire (SDQ). CNS Spectr 2014 Oct 02;19(6):535-546. [doi: 10.1017/s1092852914000406]
60. Lopez-Fernandez O, Williams AJ, Griffiths MD, Kuss DJ. Female gaming, gaming addiction, and the role of women within gaming culture: a narrative literature review. Front Psychiatry 2019;10:454 [FREE Full text] [doi: 10.3389/fpsyt.2019.00454] [Medline: 31354536]
61. Johansson P, Westas M, Andersson G, Alehagen U, Broström A, Jaarsma T, et al. An internet-based cognitive behavioral therapy program adapted to patients with cardiovascular disease and depression: randomized controlled trial. JMIR Ment Health 2019 Oct 03;6(10):e14648 [FREE Full text] [doi: 10.2196/14648] [Medline: 31584000]
62. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
63. Wright JH, Beck AT. Cognitive therapy of depression: theory and practice. Hosp Community Psychiatry 1983 Dec;34(12):1119-1127. [doi: 10.1176/ps.34.12.1119] [Medline: 6642458]
64. McGinn LK. Cognitive behavioral therapy of depression: theory, treatment, and empirical status. Am J Psychother 2000 Apr;54(2):257-262 [FREE Full text] [doi: 10.1176/appi.psychotherapy.2000.54.2.257] [Medline: 10928248]
65. Martinengo L, Stona AC, Griva K, Dazzan P, Pariante CM, von Wangenheim F, et al. Self-guided cognitive behavioral therapy apps for depression: systematic assessment of features, functionality, and congruence with evidence. J Med Internet Res 2021 Jul 30;23(7):e27619 [FREE Full text] [doi: 10.2196/27619] [Medline: 34328431]
66. Gericke F, Ebert DD, Breet E, Auerbach RP, Bantjes J. A qualitative study of university students' experience of Internet‐based CBT for depression. Counsel Psychother Res 2021 Aug 17;21(4):792-804. [doi: 10.1002/capr.12465]
67. Albert PR. Why is depression more prevalent in women? J Psychiatry Neurosci 2015 Jul 1;40(4):219-221 [FREE Full text] [doi: 10.1503/jpn.150205] [Medline: 26107348]
68. Ford DE, Erlinger TP. Depression and C-reactive protein in US adults: data from the Third National Health and Nutrition Examination Survey. Arch Intern Med 2004 May 10;164(9):1010-1014. [doi: 10.1001/archinte.164.9.1010] [Medline: 15136311]
69. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. Psychol Bull 2017 Aug;143(8):783-822 [FREE Full text] [doi: 10.1037/bul0000102] [Medline: 28447828]
70. Doherty DT, Kartalova‐O'Doherty Y. Gender and self‐reported mental health problems: predictors of help seeking from a general practitioner. Br J Health Psychol 2011 Jan 07;15(1):213-228. [doi: 10.1348/135910709x457423]
71. Rogler LH, Cortes DE. Help-seeking pathways: a unifying concept in mental health care. Am J Psychiatry 1993 Apr;150(4):554-561. [doi: 10.1176/ajp.150.4.554] [Medline: 8465869]

XSL•FO
RenderX

Abbreviations

CBT: cognitive behavioral therapy
FTP: facilitating thought progression
HE: high engagement
LE: low engagement
MADRS: Montgomery-Åsberg Depression Rating Scale
MDD: major depressive disorder
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
SHAPS: Snaith-Hamilton Pleasure Scale
SMD: standardized mean difference
WLC: waitlist control group

Edited by T de Azevedo Cardoso; submitted 10.01.24; peer-reviewed by K Byrne, C de Miquel; comments to author 09.05.24; revised version received 09.06.24; accepted 30.09.24; published 07.11.24.

Please cite as:
Yatziv SL, Pedrelli P, Baror S, DeCaro SA, Shachar N, Sofer B, Hull S, Curtiss J, Bar M
Facilitating Thought Progression to Reduce Depressive Symptoms: Randomized Controlled Trial
J Med Internet Res 2024;26:e56201
URL: https://www.jmir.org/2024/1/e56201
doi:10.2196/56201
PMID:

©Shai-Lee Yatziv, Paola Pedrelli, Shira Baror, Sydney Ann DeCaro, Noam Shachar, Bar Sofer, Sunday Hull, Joshua Curtiss, Moshe Bar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Open-World Games' Affordance of Cognitive Escapism, Relaxation, and Mental Well-Being Among Postgraduate Students: Mixed Methods Study

Ailin Anto[1*], BSc; Arunima Basu[1*], BSc; Rania Selim[1*], BSc; Thomas Foscht[2*], PhD; Andreas Benedikt Eisingerich[3*], BSc, MPhil, PhD

[1]Faculty of Medicine, Imperial College London, London, United Kingdom

[2]Department of Marketing, The University of Graz, Styria, Austria

[3]Imperial College Business School, Imperial College London, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Ailin Anto, BSc
Faculty of Medicine
Imperial College London
Exhibition Rd, South Kensington
London, SW7 2AZ
United Kingdom
Phone: 44 020 7589 5111
Email: ailin.anto18@imperial.ac.uk

Abstract

Background: Open-world games, characterized by their expansive and interactive environments, may offer unique cognitive escapism opportunities, potentially leading to relaxation and enhanced well-being. These games, such as "The Legend of Zelda: Breath of the Wild" and "The Legend of Zelda: Tears of the Kingdom," allow players to experience a sense of freedom and autonomy, which can reduce stress and improve mental health. While previous research has examined the general impact of video games on mental well-being, specific studies on the effects of open-world games among postgraduate students are limited.

Objective: This study aims to investigate the relationships between cognitive escapism provided by open-world games and their effects on relaxation and well-being. The goal was to understand how the immersive nature of these games contributes to stress reduction and overall mental health improvement among postgraduate students.

Methods: A mixed methods approach was used, which involved in-depth exploratory qualitative interviews and a survey of 609 players of popular open-world games. Quantitative data were collected using standardized questionnaires to measure open-world games' affordance of cognitive escapism, relaxation, and well-being. Qualitative data were obtained through 32 in-depth interviews that explored players' experiences and perceptions of cognitive escapism, relaxation, and mental well-being.

Results: Qualitative data (n=32; n=15, 47% female; n=16, 50% male; n=1, 3% preferred not to disclose gender; mean age 23.19, SD 2.19 y) revealed that cognitive escapism through immersive game worlds allowed players to temporarily disconnect from real-world stressors, resulting in enhanced mood and psychological well-being. Players indicated that the nonlinear gameplay and freedom to explore interactive environments provided a sense of relaxation and mental rejuvenation. Quantitative analysis (N=609) showed a substantial mediating role of relaxation in the relationship between cognitive escapism offered by open-world games and well-being. Specifically, cognitive escapism had a significant positive effect on players' relaxation (β=.15; SE 0.04; P<.001; 95% CI 0.0695-0.2331), which in turn had a significant and positive effect on players' well-being scores (β=.12; SE 0.04; P=.002; 95% CI 0.0445-0.2032).

Conclusions: The study demonstrates that open-world games offer substantial benefits for cognitive escapism, significantly improving relaxation and well-being among postgraduate students. The immersive and autonomous nature of these games is crucial in reducing stress and enhancing mental health. Future research may investigate the long-term effects of regular engagement with open-world games and explore their potential therapeutic applications for managing stress and anxiety.

(J Med Internet Res 2024;26:e63760) doi:10.2196/63760

XSL•FO
RenderX

KEYWORDS

open-world games; cognitive escapism; relaxation; mental well-being; students; video games; stress; freedom to explore

Introduction

Overview

Open-world games with their expansive, immersive environments and nonlinear gameplay have become a significant genre in the video game industry. Titles such as "The Legend of Zelda: Breath of the Wild" (Nintendo) or "The Legend of Zelda: Tears of the Kingdom" (Nintendo) provide players with computer-simulated environment to explore, fostering a sense of freedom, autonomy, and discovery. In this study, we explore the impact of open-world games on postgraduate students' well-being by offering cognitive escapism and relaxation.

Today, an increasing number of young people report high levels of stress, anxiety, and sadness [1]. Prior research has demonstrated that video games can serve as effective tools for stress relief [2-4]. In particular, in this study, we posit that open-world games with their expansive environments and opportunities for leisurely exploration may create a sense of escapism and relaxation. Previous work found that casual video game play may significantly reduce stress and improve mood [5], suggesting potential benefits for players of open-world games, which often offer similarly engaging yet nonpressuring experiences. Moreover, prior research reviewed the cognitive benefits of action video games and found improvements in attention, spatial skills, and problem-solving abilities [6]. These cognitive benefits are likely applicable to open-world games due to their intricate gameplay and diverse challenges. For instance, open-world games often require players to solve complex puzzles, strategize, and make decisions that can enhance cognitive functions.

Cognitive escapism refers to the use of media to divert attention from real-world stressors and engage in alternative, often more gratifying, cognitive activities. Open-world games, with their detailed and expansive environments, may provide ideal platforms for this form of escapism. They allow players to immerse themselves in different realities, which can be mentally stimulating and refreshing. For instance, previous work on the psychological benefits of immersive virtual environments highlights the potential for such environments to offer significant cognitive relief [7].

In recent years, there have been some exploration of escapism in emerging contexts, including the metaverse [8-10] and recreational activities, such as gaming, web-based streaming, or running [4,11]. This study extends the work of previous research that has primarily focused on general gaming or specific genres, such as action or casual games [12-19]. By specifically focusing on open-world games, new insights can be gained about the nuanced ways these games facilitate cognitive escapism.

Specifically, the expansive nature of open-world games allows players to engage in non–goal-oriented activities, such as exploration, crafting, or simply enjoying the game's scenery, which can foster relaxation. Unlike the high-intensity demands of many action games, open-world games often allow a more leisurely pace, contributing to relaxation and stress relief. Prior studies have shown that casual gaming can reduce stress. However, the open-world genre has been less frequently studied in this context. A focused study could, thus, provide robust evidence of how these games' specific features, such as vast, interactive worlds and varied activity options, uniquely contribute to relaxation.

Despite the potential benefits of gaming, there is a risk of addiction associated with prolonged gaming [20]. Open-world games with their immersive nature may also lead to excessive play. The World Health Organization has recognized "gaming disorder" as a behavioral addiction, highlighting the need for moderation and awareness [21]. Furthermore, extended periods of gameplay can negatively affect physical health, contributing to a sedentary lifestyle. A prior study indicated that high levels of sedentary behavior are associated with increased risks of cardiovascular diseases and other health issues [22]. Players of open-world games, which often engage in extended play sessions, may balance their gaming with physical activity.

Thus, open-world games hold significant potential for enhancing well-being through stress reduction, cognitive skill development, and social interaction. However, the risks of addiction and physical health impacts must be carefully managed. Additional research is needed to explore the potential benefits of open-world games while addressing their potential harms. In this study, we aimed to explore the effects of open-world games on postgraduate students' mental well-being. Students frequently operate in environments of stress and pressure [23,24]. Thus, we focus on examining the impact of open-world games in postgraduate students' lives.

In brief, this study investigates the relationships between cognitive escapism provided by open-world games and their effects on relaxation and well-being. The goal was to understand how the immersive nature of these games contributes to stress reduction and overall mental health improvement among postgraduate students.

In the following sections, we provide background information on the study and examine the exploratory qualitative studies (study 1, n=17; study 3, n=15) and the quantitative cross-sectional survey (study 2, N=609), along with their findings. We conclude with a discussion of the research implications, limitations, and avenues for future research.

Background on Open-World Video Games

"Open-world" refers to games that provide a large, freely explorable environment, where players can move and interact with the game world with minimal restrictions. Here are some characteristics that make a game, such as Nintendo's "Zelda: Breath of the Wild," an open-world game. First, a key characteristic of open-world games is the freedom to explore; that is, players can roam the world at their own pace and choose their path and activities. Indeed, Shigeru Miyamoto shared that his main inspiration when creating the Zelda games series was

XSL•FO
RenderX

to capture the essence of a young kid exploring outdoors, just as he did in the countryside of his hometown throughout his childhood. Moreover, open-world games often allow for nonlinear gameplay. Hence, the game does not force players to follow a set sequence of events or missions. Furthermore, they offer large, detailed worlds to explore. Specifically, the games may feature an expansive world filled with diverse landscapes, towns, and hidden secrets. For example, see Figure 1, which depicts Zelda's protagonist, Link, in "The Legend of Zelda: Breath of the Wild"(Nintendo). Thus, by inviting exploration and discovery, open-world games provide a unique and engaging experience that contrasts with more competitive or goal-oriented gameplay. This approach allows players to immerse themselves in the game world, fostering a sense of wonder and curiosity.

Open-world games provide a unique gaming experience by creating vast, immersive environments that players can explore and discover at their own pace. This focus on exploration and discovery is distinct from traditional game objectives, such as competition or beating the game. More specifically, open-world games often feature large, meticulously designed environments filled with diverse landscapes, cities, villages, and natural wonders. These expansive settings encourage players to wander and investigate, providing a sense of freedom and autonomy. The detailed worlds often contain hidden secrets, lore, and environmental storytelling that reward curious players. Furthermore, unlike linear games with a set path, open-world games offer nonlinear gameplay. Players are not bound by a strict sequence of events and can choose their own path. This freedom allows them to engage with the game world on their own terms, whether they prefer to follow the main story, complete side quests, or simply explore.

The self-directed playstyle of open-world games promotes a deeper connection with the game world, with a primary focus on exploration. In contrast, competitive games, such as "Fortnite," are structured around set objectives and a defined path. For instance, in the case of Fortnite, the primary goal is to be the last player or team standing, which drives a clear, competitive progression. Fortnite's gameplay is fast paced and intense, designed to keep players constantly engaged and on their toes. The competitive nature drives a high level of excitement and urgency. Open-world games, in contrast, often emphasize player-driven experiences over predefined goals. This allows players to set their own objectives at their own pace and preference, whether it is building a new settlement, taming wild creatures, or mapping out uncharted territories.

In addition, open-world games offer a dynamic environment where the world responds to player actions and the time of day, with changes in weather, nonplayer characters' behavior, and more. In addition to "Zelda: Breath of the Wild or Zelda: Tears of the Kingdom," there are several popular video games that fall into the "open-world" category. Here are some notable examples across different genres and platforms: "The Elder Scrolls V: Skyrim" is a fantasy role-playing game that offers a vast world full of quests, dungeons, and a deep lore. "Red Dead Redemption" is a Western-themed action-adventure game with a highly immersive and realistic open world. "Minecraft" is a sandbox game that allows players to explore, build, and survive in a blocky, procedurally generated world. "Ghost of Tsushima" is an action-adventure game set in feudal Japan, known for its beautiful landscapes and samurai combat. These games, to name but a few, are known for their expansive, immersive worlds that offer players a high degree of freedom and numerous activities to engage in beyond the main storyline. Next, we discuss the potential effects of open-world games on players' cognitive escapism, relaxation, and mental well-being.

Figure 1. The open world of The Legend of Zelda: Breath of the Wild.

XSL•FO
RenderX

Open-World Video Games and Mental Well-Being

Well-being may encompass a wide range of psychological states, including happiness, stress levels, and overall mental health [25]. Open-world games, by providing opportunities for escapism and relaxation, can positively influence these aspects of well-being. The autonomy and control players experience in these games can also enhance feelings of competence and satisfaction. This study on well-being in relation to open-world games builds on existing literature that examines the psychological impacts of video games [26,27]. Previous research on the motivational aspects of video game engagement suggests that games fulfilling basic psychological needs can significantly enhance well-being [19]. Thus, a study on open-world games may further elucidate how the freedom and expansive environments in these games meet these psychological needs more effectively than other genres.

Open-world games may provide a unique form of cognitive escapism by immersing players in expansive, interactive environments that divert their attention from real-world stressors. This immersive experience may lead to significant relaxation and improvements in overall well-being. Cognitive escapism refers to the process of diverting one's thoughts away from real-life concerns to immerse in an engaging activity. Open-world games facilitate this through their richly detailed and varied worlds, compelling narratives, and freedom of exploration (Figures 2 and 3). The immersive nature of these games allows players to temporarily set aside their worries and become absorbed in an alternate reality. The process of cognitive escapism in open-world games can lead to profound relaxation.

According to a prior study, engaging in video games can help alleviate stress by providing a temporary escape from daily pressures, allowing the mind to rest and recuperate [12]. The nonlinear and often serene environments in open-world games, such as the tranquil landscapes in "The Legend of Zelda: Breath of the Wild," enhance this effect by allowing players to explore and interact at their own pace without the immediate pressures of goal-oriented tasks.

Furthermore, prior work indicates that casual gaming may significantly reduce stress and improve mood [9]. The study highlights that engagement in low-pressure, enjoyable activities can lead to relaxation. Open-world games, with their extensive and varied environments, offer similar benefits by allowing players to choose their activities and engage in relaxing, noncompetitive gameplay. Cognitive escapism in open-world games may also provide emotional relief by allowing players to experience positive emotions and alleviate negative ones. Previous work found that immersive virtual environments could significantly improve mood and reduce anxiety [11], indicating that similar benefits could be derived from open-world game environments.

Well-being is a multifaceted construct that includes emotional, psychological, and social dimensions [25]. Cognitive escapism through open-world games may enhance well-being by providing mental stimulation, emotional relief, and opportunities for social interaction. The immersive experiences in open-world games can contribute to psychological well-being by fulfilling basic psychological needs, such as autonomy, competence, and relatedness, as described by self-determination theory [13]. The autonomy offered by open-world games allows players to make choices and control their in-game actions, which can lead to increased feelings of competence and satisfaction.

To explore the difference that playing an open-world game makes to students' daily lives, we first conducted an exploratory study with qualitative in-depth interviews. The findings of the qualitative study informed the survey study and the second qualitative study. In the following sections, we discuss these studies.

XSL•FO
RenderX

Figure 2. Link in The Legend of Zelda: Breath of the Wild, with parts of the Hyrule region in the background.

Figure 3. Link in The Legend of Zelda: Breath of the Wild, with parts of the Gerudo Highlands in the background.

Methods

Study 1

Qualitative Study Design

To generate novel insights from data gathered in a natural setting, we first conducted semistructured, cross-sectional interviews using open-ended, predetermined questions [28]. We report the findings of qualitative study 1 in accordance with the Standards for Reporting Qualitative Research guidelines [29]. This gave us a chance to explore and consider various personal views and opinions [30] and, in so doing, follow the spirit of theory in use approach [31]. Participants were invited to share their thoughts about open-world gaming and how it affects their daily lives. Specifically, during the interview, we asked respondents to describe how open-world gaming affects their daily lives.

As researchers with prior experience in qualitative studies, we were aware of the importance of in-depth data collection to

XSL•FO
RenderX

formulate and derive new theories. We ensured that questions were open ended. Moreover, to minimize any potential bias, 2 trained research assistants (RAs) conducted the interviews, and any discrepancies were discussed. In addition to this, the 2 RAs were trained not to guide the participants and their responses in any shape or form. The team of RAs and the entire research team was diverse in gender, age, and ethnicity; this ensured increased comfortability with participants during interviews and subsequent discussions.

Ethical Considerations

Ethics approval for the study was given by Imperial College London at Departmental level, Analytics, Marketing, and Operations. The study was conducted in accordance with the principles of the Declaration of Helsinki. Informed consent was obtained, and participants had the right to withdraw from the study at any timepoint. Study data were anonymized and kept confidential within the research team. Participants were compensated US $30 for their time.

Participants and Procedure

We used a convenience sampling method and recruited study participants via posters and handing out flyers to students on a university campus. Eligibility criteria for taking part in the study was that they were full-time students and they had some personal experience playing an open-world game before. Over the course of 5 weeks (beginning of June 2024), 17 (n=9, 53% female; n=7, 41% male; n=1, 6% preferred not to disclose gender; mean age 22.88, SD 2.15 y) exploratory in-depth interviews were conducted with respondents, who were full-time students (student ID check) and have played an open-world game before (the trained RA asked some basic screening questions, including which specific open-world game or games they had play before and the basic content [story] of the games).

We followed established procedures for inductive qualitative data analysis when examining our qualitative data. Specifically, following constant comparison techniques in prior work [28], we collected and analyzed the interview data in tandem and compared the newly collected data against the existing insights. With this approach, we first aimed to group conceptually linked data and reduce them to a set of meaningful concepts [28]. First, we analyzed the data to develop a holistic understanding of respondents' evaluations of open-world gaming and key elements of it that impact their lives. Furthermore, we used open coding and used descriptive phrases to name these codes to develop insights about key motivators and process explanations regarding respondents' evaluation of open-world games and key elements thereof that may influence their mental well-being [32]. Finally, by integrating relevant literature and conducting joint discussions, we finalized the elements of our conceptual framework to reflect the key elements of open-game worlds that affect players' well-being [28].

Interviews were conducted in the alumni lounge and a café of a university and lasted between 20 and 95 minutes. Interviewees were reassured of the strict confidentiality of their responses and were told that their participation would inform academic research. At the end of each interview, we gave interviewees a chance to ask any potential question that they had.

Qualitative Data Analysis

First, each RA familiarized themselves with the data collected by transcribing the interviews and reading through the transcripts. Second, the authors then analyzed the data to develop a holistic understanding of the qualitative data. We used open coding and used descriptive phrases to name these codes to develop insights about the potential effects of open-world games on users. Interview data coding was conducted in duplicate by the authors, and any discrepancies were discussed and resolved. We used the constant comparison technique and compared the newly collected interview data against the existing insights as we conducted more interviews. In doing so, our first goal was to group conceptually linked data to reduce them to a set of meaningful concepts. As a third step, we finalized the elements of our conceptual framework to reflect the key effects of open-world games on users' well-being. This was done by grouping codes together under common ideas to derive overarching themes. Fourth, the themes were then revisited and reviewed by the researchers and compared with all the data collected to ensure that all key concepts had been captured by the key themes. Fifth and finally, findings were presented through a report that highlighted the key identified effects of open-world games with a description and exemplary quotes.

Study 2

Survey Study Design

We used convenience sampling and a cross-sectional survey as part of study 2. Eligibility for this study required participants to be full-time university students with personal experience playing an open-world game. The study was conducted in June 2024. For this study, we followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines to report the observational, cross-sectional studies (Multimedia Appendix 1) [33].

Ethical Considerations

Ethics approval for study 2 was received from Imperial College London at Departmental level, Analytics, Marketing, and Operations. Data were anonymized and kept confidential within the research team. Participants provided informed consent and had the right to withdraw from the study at any time.

Participants and Procedure

We recruited study participants via posters and handing out flyers to students on a university campus during June 2024. Potential participants were informed that the study would take no longer than 20 minutes and that they would receive US $8 as a thank you for their participation. A total of 630 full-time university postgraduate students (ID check), who played an open-world game (2 trained RAs asked brief screening questions before the study, eg, which open-world game they had played and brief content or story of that game), were recruited to take part in this study over the course of 3 weeks. In total, 21 respondents did not complete the entire survey, and thus, the effective number of responses is 609.

In case participants played >1 open-world game, study participants were invited to think of the game that they played the most in terms of hours spent per week when answering the

XSL•FO
RenderX

survey questions. This study was conducted in a lecture theater of a university and was supervised by 2 trained RAs. As part of the study, participants were invited to complete a brief paper-and-pencil questionnaire about their mental well-being. The questions were adapted from a previously published psychological well-being scale [34], which included items such as "I feel there is meaning to my present and past life," "I have a sense of continued self-development and growth," and "I have a positive attitude toward myself and life in general." Participants also answered questions about the extent to which playing an open-world game provided relaxation, adapted from an established emotional relaxation scale [32]. These included statements such as "Playing an open-world game makes me feel content," "When I play an open-world game, I feel relaxed," and "Playing an open-world game makes me feel stressed" (reverse-coded). In addition, questions about cognitive escapism were included, such as "Playing an open-world game helps me escape from my everyday problems," "Open-world games help me forget the troubles and anxiety in my life," and "Playing open-world games makes me forget my worries." Responses were measured on a 7-point Likert scale (1=strongly disagree and 7=strongly agree). No demographic data (gender, age, etc) were collected as part of this study. At the end of the study, each participant was thanked and received US $8 as compensation for their participation. In addition, participants had a chance to ask any questions that they may have.

Survey Data Analysis

Before testing the effects of open-world games' affordance of cognitive escapism on relaxation and mental well-being, we evaluated the factor structure and the reliability scores for the measurement items in the study. Specifically, we followed best practices for conducting confirmatory factor analysis, and the results showed that all the measurement items significantly loaded on their intended factors (P<.001) with weak cross-loadings (<0.20) [35,36]. The average variance extracted were higher than the recommended threshold of 0.50, indicating the convergent validity among the items of each scale [36,37]. In support of discriminant validity of the constructs, the average variance extracted was greater than the squared correlations between constructs and the other variables [38]. Table 1 lists the specific measurement reliabilities and item factor loadings.

As part of this study, we also conducted a mediation analysis with the PROCESS macro [39] (model 4; bootstrapped samples=5000; 95% CI) to examine the effects of open-world games' affordance of cognitive escapism (independent variable) on players' relaxation (mediator) and, consequently, on players' mental well-being (dependent variable).

Table 1. Measurement items' factor loadings and reliabilities.

Construct and measurement items	Factor loadings
Mental well-being (α=.92; CR[a] =0.94; AVE[b] =0.85)	
I feel there is meaning to my present and past life	0.93
I have a sense of continued self-development and growth	0.92
I have a positive attitude toward myself and life in general	0.91
Relaxation (α=.93; CR=0.95; AVE=0.87)	
Playing an open-world game makes me feel content	0.94
When I play an open-world game I feel relaxed	0.93
Playing an open-world games makes me feel stressed (reverse)	0.93
Cognitive escapism (α=.91; CR=0.94; AVE=0.84)	
Playing an open-world game helps me escape from my everyday problems	0.92
Open-world games help me forget the trouble and anxiety in my life	0.92
Playing open-world games makes me forget my worries	0.91

[a]CR: composite reliability.

[b]AVE: average variance extracted.

Study 3

Qualitative Study Design

Following our first qualitative and survey studies, we conducted a second qualitative study to further explore the specific elements of open-world games that may facilitate cognitive escapism. Specifically, we followed the same procedure as in qualitative study 1 (conducting semistructured, cross-sectional interviews using open-ended, predetermined questions and convenience sampling); we report the findings of qualitative study 3 in accordance with the Standards for Reporting Qualitative Research guidelines [29]. As part of this study, we conducted 15 semistructured in-depth interviews. As in study 1, we paid special attention to the importance of in-depth data collection to formulate and derive new theories. More specifically, and as in study 1, we ensured that questions were open ended. Furthermore, in an effort to minimize any potential bias, 2 trained RAs conducted the interviews, and any discrepancies were discussed. Moreover, as in study 1, the 2 RAs were trained not to guide the participants and their responses in any shape or form. The team of RAs and the entire research team was diverse in gender, age, and ethnicity; this ensured increased comfortability with participants during interviews and subsequent discussions.

XSL•FO
RenderX

Ethical Considerations

Ethics approval for the study was granted by the Department of Analytics, Marketing, and Operations at Imperial College London. All data were anonymized and kept confidential within the research team. Participants provided informed consent and had the right to withdraw from the study at any time.

Participants and Procedure

As in our qualitative study 1, we used a convenience sampling method and recruited study 3 participants using posters and handing out flyers to students on a university campus. The 2 eligibility criteria for taking part in this study were that participants were full-time students and that they had some personal experience playing an open-world game before. Over the course of 4 weeks (June 2024), 15 (n=6, 40% female; n=9, 60% male; mean age 23.53, SD 2.26 y) exploratory in-depth interviews were conducted with respondents, who had their student ID checked and had played an open-world game before (a trained RA asked some basic screening questions, including which specific open-world games they had play before and the basic content [story] of the games).

During the interviews, participants were invited to share their thoughts about the specific elements of open-world games that influenced their lives ("Which elements of open-world games influence how you feel in your daily life?" "How do they do so?" "Are there certain aspects of open-world games that you find most impactful in your daily life in terms of how they make you feel?" "Why do you enjoy playing open-world games, if at all?" "Are there any other elements of open-world games that you find noteworthy in terms of how they make you feel?" "Can you please share how these manage to do so?").

As in study 1, interviews were again conducted in the alumni lounge and a campus café of a university and lasted between 75 and 135 minutes. Interviewees were informed about the confidentiality of their responses and were told that their participation would inform academic research. Each interviewee was thanked and received US $30 as compensation for their participation. Interviewees also had the chance to ask any questions they potentially had at the end of the study.

Qualitative Data Analysis

Following the same steps as in study 1, first each RA familiarized themselves with the data collected by transcribing the interviews and reading through the transcripts carefully. As a second step, the authors analyzed all the qualitative data to develop a holistic understanding of it. Specifically, we used open coding and descriptive phrases to name these codes to develop insights about the potential effects of open-world games on users. Interview data coding was conducted in duplicate by the authors, and any discrepancies were discussed and resolved. We again used the constant comparison technique and compared the newly collected interview data against the existing insights as we conducted more interviews to group conceptually linked data to reduce them to a set of meaningful concepts. Third, we finalized the key elements identified by grouping codes together under common ideas to derive overarching themes. Fourth, the themes were then revisited and reviewed by the researchers and cross-checked and compared with all the data collected to ensure that all key concepts had been captured by the key themes. Finally and fifth, we presented findings through a report that highlighted the key identified effects of open-world games with a brief description and exemplary quotes.

Results

Study 1: Results

The interview findings showed that open-world games' affordance of cognitive escapism played a key role in their effect on players' lives. Table 2 lists the representative quotes, which illustrate the extent to which open-world games were seen as providing cognitive escapism and players' appreciation of cognitive escapism when playing. One interviewee noted the following:

> *Finally I'm feeling inner-peace. Exploring...wandering around...it allows me to forget my daily worries. It is liberating in a way. I don't have to think about things that would otherwise upset me.* [Respondent #8]

Furthermore, the interview findings demonstrated the key role played by relaxation offered by open-world gaming. Respondents noted that reduced stress and enhanced relaxation were some of the key benefits of playing open-world games (Table 2). For instance, a respondent said the following:

> *There is this constant background noise. It stresses me out. It is exhausting and I get no respite. When I am playing I am less stressed. I breathe more easily. It calms me down. This is a true form of relaxation or being at peace with myself and the world moment.* [Respondent #3]

In addition to this, respondents shared their points about the extent to which open-world games impacted their mental well-being. One of the respondents, for instance, shared the following:

> *There are a million of things in my life that need taking care of. It never ends. [Game name] is something I look forward to. No matter how bad the day is, I know I can grow, explore, and keep going in [Game name]. Without it I would feel like something important is missing in my life. To keep me healthy and sane, oh, yes, spending time playing [Game name] is probably the way at the moment in my life.* [Respondent #5]

XSL•FO
RenderX

Table 2. Representative quotes of respondents regarding the key effects of open-world games.

Open-world games' impact	Representative quotes
Cognitive escapism	• "I need something that allows me to switch off at times. Constantly thinking about things drives me mad. Playing [Game name] allows me to switch off and not worry." [Respondent #2] • "It's just always there. It never stops. Worrying about grades. Worrying about what others think. What's going on in the world. It is exhausting to keep thinking about this all the time. I need some rest from all of this and this is what [Game name] gives me. It gives me space to not have to think about my daily worries all the time." [Respondent #5] • -"[Game name] is my form of mediation. There is no other time during the day when I feel at ease, no concerns, no distractions of thinking about various things. Yes, it really is the closest to mediation I can think of. It's like a blessing and a refuge from stress." [Respondent #6] • "Exploring at my own pace, discovering new things, trying out new things...when I spend time playing [Game name] I feel absolutely at ease. [Game name] makes me feel less worried and less stressed out by all the things." [Respondent #9] • "What Nintendo does really well and I say Nintendo in particular is giving players a sense of peace and inner calm. Their Zelda: Breath of the Wild or Zelda: Tears of the Kingdom games are so positive. They make me feel I can escape the negativity in today's world and be positive again. I can think of something else other than the madness that characterises social media and life in general these days." [Respondent #10] • "[Game name] is my home away from home. It is where I want to be. [Game name] gives me respite from all the stress. I need a break. I just need a break sometimes. [Game name] is the break I need." [Respondent #13] • "One could say it is a form of relief of stress and a way not to have to constantly fight the daily battle with issues. That's what it is for me. [Game name] allows me to breathe easy, to just give it a rest and not overthink things all the time. That inner voice inside my head is finally silent. [Game name] gives me a chance to be calm inside my head." [Respondent #14] • "As a student I face constant pressure from peers, the programme, social media conversations, etc. [Game name] is the only place I have in my life that gives me peace from all these stressors. When I play [Game name] I worry less about these kind of things." [Respondent #15] • "[Game name] gives me a chance to lose myself in it. [Game name] is a break from all the things going on in my life right now." [Respondent #16]
Relaxation	• "It's almost instant or after a short while of playing [Game name]. I start to feel at ease. I am less tensed. I am definitely less stressed out and worried about life." [Respondent #1] • "[Game name] is incredible. I feel more relaxed and rejuvenated after playing than after getting a massage. My body and mind are fully at ease. All the tension is gone." [Respondent #3] • "Primarily it is a soothing experience for me. [Game name] makes me feel at ease and less stressed out. It is relaxing for me." [Respondent #4] • "I can feel the difference it makes. Playing [Game name] makes me feel a lot calmer. My mind doesn't feel so tense. And I can sense the effects on my body too. Even my muscles overall feel less tensed as a result." [Respondent #5] • "My muscles feel relief and so does my mind. [Game name] gives me relaxation. It relaxes me. I feel less stressed out and tense." [Respondent #7] • "I find it super relaxing. Everybody needs a Nintendo Switch. I recommend it wholeheartedly. Probably that's the best investment you can make with your money today. Play Zelda on the Nintendo Switch and you can thank me later." [Respondent #8] • "It's both. Both my body and mind feel relaxed. It's definitely a relaxing experience." [Respondent 10] • "[Game name] calms me down. Playing [Game name] calms me emotionally. It soothes me. I feel at ease and some form of inner-peace. While it may sound a bit crazy, it's real; [Game name] gives me inner-peace." [Respondent #11] • "[Game name] soothes my tired and exhausted mind." [Respondent #12] • "It's relaxing and nothing is quite like it. I don't have found anything that comes even close in terms of giving me a chance to breathe and be less tense, to feel loose and relaxed." [Respondent #14] • -"[Game name] alleviates my tension. It's just soothing and relaxing to play. Very relaxing." [Respondent #17]

XSL•FO
RenderX

Open-world games' impact	Representative quotes
Mental well-being	• "[Game name] has made me a far more positive, satisfied person." [Respondent #1] • "Playing [Game name] has given me a new perspective in life. It's not just about worries and anxieties. I can explore. I have a sense of achievement, of growth, of doing more with my life. [Game name] has had a very positive impact on my life. I feel a lot more content with what I have and who I am." [Respondent #2] • "[Game name] gives me a sense of fulfilment in my life. A sense of accomplishment and as a result I feel more satisfied and happy." [Respondent #5] • "Before [Game name] I often felt a loss of purpose and I felt there was just emptiness. [Game name] gave me a new perspective on life in general. There are still so many things I need to deal with as part of what's going on but at least [Game name] gives me a mission, a quest, something different. To some others it may be trivial and meaningless but to me it holds deep meaning. I feel [Game name] gives me the strength I need to keep going in spite of everything." [Respondent #8] • "When I play [Game name] I feel a sense of direction a certain calling. [Game name] gives me a chance to explore and keep growing as a person. It is deeply satisfying. Overall it has a tremendous impact on my health and well-being on a day-to-day basis." [Respondent #9] • "Too many gaming companies go down the route of trying to lure people in and then rip them off. It's like a casino. Stay away from such games. They are like social media, where your brain is reduced to clickbait and constant dopamine rushes that you dearly pay for. I love Nintendo for being so different and unique in the industry. Their open-world games are truly offering an immersive, positive sense of exploration and adventure. It reminds me of the times when I was a kid, exploring the countryside by myself. That was the best time of my life hands down. When I play Zelda: Breath of the Wild I feel like that kid again. There is nothing quite like it. I deeply enjoy and treasure that feeling." [Respondent #10] • "In an interesting way [Game name] makes me actually more at ease and accepting of my situation and who I am. I am happy with what I have got. I am happy with who I am and that's also thanks to [Game name]." [Respondent #11] • "It's a rejuvenating experience. It clears my mind and I feel happier, more positive, less doom and gloom. There is so much negativity everywhere it is depressing. [Game name] allows me to think of other, more positive things in life. I feel less depressed and lonely. [Game name] gives me happiness." [Respondent #13] • "My mental health and well-being has improved significantly thanks to [Game name]. I feel a lot happier now than before." [Respondent #14] • "Life is depressing if it is only about solving problems. There will always be challenges and this and that to cope with. But life can't only be about that. The freedom to explore, to grow, to learn, to see new places and do new things perhaps someone has not tried before, it's just refreshing and that is exactly what [Game name] gives me." [Respondent #16] • "At one point I started to question the meaning of it all. Why even bother? What's the meaning of all this? Of life? Is my life different from that of an ant? Does it make a difference to the universe? [Game name] just gives me a different perspective. I don't have to know the answers to these questions. It's OK. Life is OK. [Game name] gives me a sense of meaning, growth, development, exploration, wonder, appreciation, gratitude, relief, peace, beauty of the natural things around me. [Game name] has opened my mind to new things in my life or perhaps there were there but I could not see and appreciate them deeply. [Game name] has changed my life for the better." [Respondent #17]

Study 2: Mediation Analyses Results

The results for the mediation analyses showed that the direct effect of cognitive escapism on mental well-being was not statistically significant because 95% CIs included 0 (P=.43; Table 3). Notably, we found evidence for an indirect path of cognitive escapism through relaxation (standardized β=0.02; bootstrapped SE 0.009; 95% CI 0.0042-0.0390; Table 4). Specifically, the detailed results indicated that cognitive escapism afforded by open-world gaming had a significant positive impact on relaxation by the players (standardized β=.15; SE 0.04; P<.001; 95% CI 0.0695-0.2331; $F_{1,607}$=13.21; P<.001, R^2=0.02), which in turn had a significant and positive impact on players' mental well-being (standardized β=.12; SE 0.04; P<.002; 95% CI 0.0445-0.2032, $F_{2,606}$=5.48; P=.004, R^2=0.02). These results suggest that relaxation significantly mediates the effects of cognitive escapism on mental well-being (Table 4), supported by ANOVA results (Table 5).

Table 3. Mediation results.

	Relaxation			Mental well-being		
	Standardized coefficient (95% CI)	Coefficient (SE)	P value	Standardized coefficient (95% CI)	Coefficient (SE)	P value
Constant	—[a] (3.11 to 3.84)	3.47 (0.19)	<.001	— (2.85 to 3.76)	3.30 (0.23)	<.001
Cognitive escapism	0.15 (0.07 to 0.23)	0.15 (0.04)	<.001	0.03 (–0.05 to 0.12)	0.03 (0.04)	.43
Relaxation	—	—	—	0.12 (0.05 to 0.20)	0.12 (0.04)	.002

[a]Not applicable.

XSL•FO
RenderX

Table 4. Total and direct effects of cognitive escapism on mental well-being.

	Effect (SE; 95% CI)	*P* value
Total effect	0.05 (0.04; –0.03 to 0.13)	.22
Direct effect	0.03 (0.04; –0.05 to 0.12)	.43

Table 5. Indirect and standardized indirect effects of cognitive escapism on mental well-being.

	Effect (bootstrap 95% CI)	Bootstrap SE
Indirect effects	0.02 (0.01-0.04)	0.01
Standardized indirect effects	0.02 (0.01-0.04)	0.01

Study 3: Results

The findings of our second qualitative study showed which specific elements of open-world games participants noted as being effective in facilitating cognitive escapism, namely, (1) offering a sense of exploration, (2) giving players a chance to develop and experience mastery and skill, (3) offering a sense of positivity, and (4) providing a sense of purpose and meaning in life. One respondent noted the following:

> *It's the freedom to roam around, to keep searching, to take the road less travelled...that makes me want to come back to [Game name]. It clears my mind and allows me to think freely, to entertain thoughts I'd otherwise not enjoy having. It is that liberty to go out there and explore the map, which then makes me feel free and lighter.* [Respondent #1]

Another respondent shared the following:

> *To be able to put in the hours but also see a reward, to see the character grow, do more, be more...it's so satisfying. It's empowering really and helps me free up my headspace and shut up that inner voice that tells me "no, you can't do this or can't do that."* [Respondent #6]

Furthermore, as part of our interviews, 1 respondent indicated the following:

> *After a hard and difficult day I feel exhausted and empty. And The Legend of Zelda games are so positive. No matter how much trial and tribulation Link and Zelda have to go through, there is a positive side to it. An encouraging message. A path that opens up or is worth searching for and support to keep going. The positivity of the games such as Zelda go a long way. It truly helps me cope with the stress and challenges in my life. It gives me a chance to breathe easy if only for a while but also later in the day I look forward to [playing Zelda] and this positivity carries over and gets me through the day.* [Respondent #2]

In addition to this, a respondent shared the following:

> *Life sometimes feels meaningless. I mean why bother if our lives really don't matter much to the universe than the life of an ant? Sure, some people think themselves more important than ants but to the cosmos really they may be the same. Sometimes I struggle because of this and playing [Game name] gives me a reason and something to look forward to, goals to reach, things to explore and conquer, to keep learning, and I enjoy the process, the journey. It may sound crazy but it is meaningful to me this journey. Let's say I start of as Link in The Legend of Zelda: Breath of the Wild in the shrine of resurrection...I wake up...walk out of the shrine and see this vast world, free to explore, free to try, free to fail, free to keep going, growing...it actually is quite moving. As in life, you start of butt-naked. Well, they give you basic clothes but no armour, no shield, no bow and arrow...you gain all these by exploring the world on offer along the way. During the day and night I may be engaged in tasks that...who knows if they are meaningful or not but when I spend time in Hyrule I feel a sense of meaning. The game gives me purpose in life. It is a beautiful reason for me to go through some of the other things in life. It grounds me. It anchors me. Yes, it gives me meaning in life.* [Respondent #3]

Additional representative quotes for these specific elements of open-world games noted by the respondents are listed in Tables 6-9.

XSL•FO
RenderX

Table 6. Representative quotes for the theme of "offering a sense of exploration."

Open-world game elements and description	Representative quotes
Offering a sense of exploration	
Nonlinear gameplay allowing exploration and interaction at own pace	• "It's a very soothing experience to be able to enjoy and explore [Game name] and different paths within the game at my own pace." [Respondent #1] • "Sometimes it is super tedious to be asked or forced to follow a certain, very linear and predictable play in a game. It becomes a chore not a game anymore. I feel like I have to do certain things and simply follow instructions. That's not what I want. Once I feel I feel boxed in with all these rules, it just loses its appeal. Being able to make my own decisions and draw my own path and make decisions on the fly as to what to explore next is priceless. This is where the exploration element comes in for me that gives me flow and losing myself in the exploration and journey itself." [Respondent #5] • "The moment I begin to feel it is all about pressing the right buttons at the right time I am out. Count me out. This is stress for me and quite the opposite that I need and seek in a game. I feel like a rat in a laboratory then asked to press various buttons at the right time. I don't want to be that rat. Let me decide, give me the freedom to make a decision what step to take next in the game. I may regret it but at least I chose." [Respondent #8] • "You never know what's around the corner. That's why it is so magical. There may always be a surprise knocking on your door. Maybe you did not expect it but it surprised you and it put a smile on your face. I love those moments. That's why life is magic. [Game name] does this really well. One moment you are fighting a mystical creature, and the next moment you are shield-surfing in the snowy mountains. Perfection. I feel like the time stops in those moments and my mind and body are rushed with feelings of complete calm." [Respondent #10]
Immersive, detailed environments that offer a sense of freedom	• "There is this immense open world that is out there to be explored. It is inviting me to keep looking, to take different turns and enjoy the twists in doing so. Each region has its own characteristic and I feel like I am on this grand adventure, conquering different parts of the map that I enjoy and find so liberating in many ways." [Respondent #3] • "From the majestic and vast, vast Hebra Mountains to the Wetlands of Lanayru and the scenic, leafy fields of the Akkala Highlands, all the way to the punishing heat of Death Mountain, and the sandy beaches of the Necluda Sea, and the extreme temperatures of the Gerudo Desert, and not to mention of course the Hyrule Field, and the Great Plateau, where all the story starts...I could go on, there is so much beauty, so much to freely explore. WOW! It truly is a world out there. It is special and I feel I want to be a part of it all. It is a part of me. That is what bliss feels like for me the freedom to explore these beautiful regions." [Respondent #6] • "One can easily spend weeks and weeks exploring the Hebra Mountains alone. C'mon that place is wild. It's crazy. It's so beautiful and rich in depth and full of adventure and danger lurking everywhere. Get your shield and get boarding on snow! You even get the world-champion tutoring you, when you find her. It's just nuts. I completely forget about time when I play. When I want some quality me time this is it." [Respondent #11] • "There was this ad with someone on the bus driving home after work. This hit hard. I was like that's me on the bus. I got [Game name] and I am grateful I did. Actually do you know that our planet Earth is beautiful? You wouldn't believe it reading all the negative stuff on social media and so called news out there. It is difficult not to get depressed by it all. So here I was and I started to play, not exactly knowing what to expect, and it was such a thing of magic. There was adventure and a sense of exploration around each corner...it blew my mind and even now after months and months it makes me feel warm and fuzzy inside thinking about it. The world is so detailed and varied and just incredible what may explore in it. It is brilliantly done. Breath taking and so calming too with a chance to do something without the constant worry about stress and being anxious about this and that in my life. Wonderful." [Respondent #15]

XSL•FO
RenderX

Open-world game elements and description	Representative quotes
Vast and varied game worlds provide a chance for personal journey and growth	• "Often I catch myself thinking I am the on the journey here. It's me doing the exploring and growing." [Respondent #4] • "Through all this trial and tribulation, the frustrations of tackling some of the shrines that are really a pain in the neck to figure out and to find I have grown as a person. Patience. Self-Discipline. Kindness. Friendship. Dedication to the craft...yes, I have grown as a warrior not just Link." [Respondent #8] • "[Game name] is fabulous. Full of adventure and challenges and so many places and crafts to explore. Each place has hidden secrets. To unlock them one needs patience and a lot of grit. Getting lost in Hebra Mountains, c'mon, they are just so vast. It is wild. I learned a lot about myself when playing [Game name]. You think it is all about moving around Link but it's not. It is about you as a player, as an explorer, of the captain of your own life and the adventures to unlock along the way." [Respondent #10] • "Did you know that Bushido means 'the way of the warrior'? Courage and bravery but also politeness and kindness as well as a sense of justice and faithfulness are things I have considered when playing [Game name]. It has become a part of me and Link's journey in many ways is my journey." [Respondent #12]

Table 7. Representative quotes for the theme of "offering a sense of mystery and skill."

Open-world game elements and description	Representative quotes
Offering a sense of mastery and skill	
Enhances feelings of satisfaction and competence by enabling players to overcome various challenges and struggles	• "Finally I managed to get the master sword. What a moment. I waited for this moment so long! That's the feeling you live for. Total bliss and total feeling of relief and peace." [Respondent #3] • "I often struggle with the fact that I seem to hit a wall. How to proceed in life? What to do next? It all seems so random and chaotic. Everyone is negative and deeply pessimistic, selfish, tribal, and self-obsessed in this world right now. It's a welcomed escape for me to...well, learn new things in [Game name], see Link gain new skills, I learn new skills. I thought I would never manage to beat the four divine beasts but finally I did it. The first time I found a Korok seed, solved one of those puzzles...I will never forget those moments. They are staying with me." [Respondent #6] • "Mezza Lo Shrine was impossible to complete. It took me days and days and I couldn't get it done. That buck always escaped. And you know what? I hit the wrong button! Hahaha, instead of the L button I kept spamming the R button. It was hilarious. What a hoot. It really brought home the message of never giving up. And in the end I did it. I completed that damn shrine. Hahaha. Heart in the mouth stuff. I felt I could manage it and do other things too. If I can manage the Mezza Lo Shrine I can overcome other challenges." [Respondent #11] • "Never look back too much unless it's for gratefulness and learning that's what [Game name] taught me. Nunquam retro as the old Latin saying goes. [Game name] threw challenges at me I never thought I'd manage ever. They were impossible. At least that's what they looked like to me at the time. Funny how I managed in the end. I learned so much. The frustrations were huge at times. You know how it feels you try and try and try and it simply won't work? Yes, that's exactly how I felt. But then I did it. I was able to solve the riddles, unlock the paths, it was amazing." [Respondent #13]
Enables players in their daily lives by offering a feeling of being capable and able to shape one's environment or destiny	• "What Link has shown me is that I can be more than a one trick pony. I still have got the moves and can put them to use in dangerous situations, facing obstacles and enemies. There is no reason to believe we cannot change anything. This enabled me in my life too. To know and to think one isn't done yet. Life is not over until it's over." [Respondent #9] • "I used to spend four to six hours on social media every day. It got to me in the end. It made me feel lonely, sad, anxious, depressed and deeply pessimistic about everyone and everything. It was a rabbit hole and I couldn't get out. Utterly powerless and a sense of being at a loss that's what it felt like day in and day out. [Game name] changed my life. It's crazy. I felt peace and inner calm. I could play, nobody else, I played and that's it. After months into the game I thought to myself 'well, life isn't too bad'. I can do this." [Respondent #11] • "I very much felt empowered by this. If I can conquer the different terrains of [Game name] I should be able to handle the other things here in life." [Respondent #14] • "It's not this static thing. Absolutely it's not the case that there is no response. There is. As the protagonist you can change the route of the story by the actions you take. This made me think about my own life and what I do and what I don't do and how one step opens up other things and so on." [Respondent #15]

XSL•FO
RenderX

Table 8. Representative quotes for the theme of "offering a sense of positivity."

Open-world game elements and description	Representative quotes
Offering a sense of positivity	
Engaging and positive narratives provide encouraging emotional experiences	• "Switching on the TV or even worse watching anything on the internet is saying hello go gloom and doom. The end is neigh. Pack your packs the world as we know it already ended and that's what the internet is all about 24/7. Negativity rules in anti-social media. The sad part is how addictive it all is scrolling down reading the shocking, sad, outrageous things. It messes with your head and leaves you ultimately sucked dry and empty. I tell my friends 'everyone needs a Nintendo Switch'. The positivity of [Game name] is what makes it for me. The story is so positive and keeps you positive." [Respondent #2] • "Well, it's not really novel exactly the story...when you think about it, it's that ancient good versus evil. But that's OK. That story never gets old. It's part of human existence. It's what we are all about individually and as a species. We all struggle with it and it's a part of us and who we are. So that's all OK. Don't be worried about it too much. Any story can be told in multiple ways. It can be sad and leave you anxious. It can be encouraging and make you feel stronger and more positive. Beating Calamity Ganon may be the goal but it's the story and journey along the way that makes you more positive. Also it's true genius that they allow you to fight Calamity Ganon over and over again without having to start a new game. That's absolutely brilliant. I love the positive ending. It's such a positive and encouraging story." [Respondent #5] • "I want more positivity in my life. [Game name] gives me that, which makes me feel better and at ease, less stressed out and worried about all sorts of things going on." [Respondent #14]
Gives players a sense of belonging and makes them feel as though they have a home away from home	• "I was travelling and I didn't bring my console. So I couldn't play [Game name] for a few weeks. And it's like, I missed it. The moment I saw Hyrule Fields, I am like 'I am back. I am home'. It was such a warm feeling." [Respondent #1] • "[Game name] is this place I belong. I feel I belong there and it's part of me." [Respondent #10] • "[Game name] is so familiar to me, yet offering always something new. It's so comforting in one way and giving adventure and exploration also." [Respondent #12] • "You can always go back to it. I know it is there for me when I need it. In many ways it is like a home, where you can enjoy, be yourself, not worry about the madness of the world. It's home." [Respondent #14]

XSL•FO
RenderX

Table 9. Representative quotes for the theme of "offering a sense of purpose and meaning."

Open-world game elements and description	Representative quotes
Offering a sense of purpose and meaning	
Provides opportunities for introspection and self-reflection as part of gameplay	• "Life can be so dull and always the same thing, the same old, the same loneliness, the same negative things, and so [Game name] is this break I need from all that. It gives me a chance to think about other things and...then I think about things that probably otherwise I wouldn't have thought about at all." [Respondent #3] • "What does it say about me as a player and what does it mean for me if I don't manage this I would ask myself quite often when playing [Game name]. But most often I wouldn't think about anything in particular and that was the beautiful part. It's quite rare not thinking about a zillion things at a time. Then I thought about my life and what I do and then I thought about what I want to explore. It allowed me to explore my own situation more deeply, which is rare." [Respondent #7] • "It can all appear as to have to meaning. My friends always say that it doesn't matter. When the lights are switched off then that's that. End of story. There could be more. What if there is more? In [Game name] the story does not end, why should my life end?" [Respondent #11] • "Perhaps that's what it's all about? Maybe we are already in a game, in a simulation? Elon Musk thinks so. Why not? It could be real. It's real. [Game name] is a story and I play it and I think how my story is going and how it will end maybe." [Respondent #13]
Allows players to transcend the worries and stressors of daily life	• "WOW! It's the small things that kill me. You know, that email or that comment or this picture or that someone's words...In the bigger scheme of things they are all, well, do they matter? In [Game name] I feel there are more important things than these petty small stuff to sweat. There are more to life and to see and enjoy." [Respondent #4] • "There is so much pressure from university. My parents are stressed out and I am stressed out. I don't know how to cope. [Game name] finally allows me to forget about these worries and move on not being constantly bombarded with stressful things." [Respondent #9] • "I get worried so much and it's hard to stop worrying. So [Game name] helps me breathe easy and not have to be worried all the time non-stop. I can think of other things." [Respondent #13] • "Hit the reset button I tell myself and it is not easy to do. It is easy to say but it is not easy to do. Everyone needs a little help I think. Everyone needs a bit of something else besides stress and worries and problems to be solved all the time. [Game name] is my reset button." [Respondent #15]

Discussion

Principal Findings

By examining how open-world games afford cognitive escapism, relaxation, and mental well-being, this study significantly contributes to the understanding of video games' complex psychological impacts. Specifically, this research may inform game design, promoting features that maximize these benefits, and guide mental health professionals in using open-world games as therapeutic tools.

In brief, the findings of this study show that the cognitive escapism afforded by open-world games contributes significantly to relaxation and overall well-being among postgraduate students. By providing immersive environments that allow mental diversion, emotional relief, and meaning, these games can serve as valuable tools for enhancing psychological and emotional health. We invite future research to continue to explore these relationships to better understand how to maximize the therapeutic potential of open-world games while mitigating any potential negative impacts.

Taken together, this study on open-world games and their potential effects on cognitive escapism, relaxation, and well-being fills a significant gap in the current literature. By offering a detailed exploration of open-world games' unique game characteristics, the findings of this study provide valuable insights into how digital environments may positively impact mental health and well-being, paving the way for more targeted and effective uses of video games in therapeutic settings.

Comparison to Prior Work

Previous studies [9,15] demonstrated that casual gaming could reduce stress and improve mood, suggesting that video games can serve as effective tools for relaxation. This study builds on these findings by focusing specifically on open-world games and confirming that the freedom and autonomy provided by these games contribute to relaxation and mental rejuvenation.

While previous research has generally addressed the benefits of video games [19], this study focuses on the unique aspects of open-world games, confirming their significant positive impact on mental health in the context of postgraduate students. Future research should continue to explore these relationships, particularly the long-term effects and therapeutic potential of open-world games for managing stress and anxiety.

Limitations and Future Research

This study has several limitations, which also offer promising avenues for future work. For instance, while in-depth interviews were conducted and open-world game players were surveyed as part of this research, the reliance on self-reported data is a limitation that is also an opportunity for future work. While self-reported data are valuable, they may be complemented by physiological or observational measures to provide a more comprehensive assessment. Future research that uses

XSL•FO
RenderX

physiological measures to examine the effects of open-world gaming on mental well-being is richly deserving. Furthermore, arguably this study explored more of the immediate or medium-term effects of gameplay but was not able to measure possible long-term effects. Medium-term stress reduction and relaxation may not necessarily translate into sustained mental health benefits in the long term. Longitudinal studies are needed to assess how regular engagement with open-world games influences mental health over time. In addition, the relatively small R^2 value in our quantitative analyses (study 2) may be due to potential confounding variables that could be considered to further strengthen the power of the analysis. For example, prior research has highlighted the anxiety and stress that social media and university studies may cause [40-43] as well as the positive effects of gamification [44-48], hope and well-being benefits [49,50], and the role of gaming in enhancing critical skills and knowledge, which contribute to individuals' connections and well-being in life [47,51-57]. Future research examining open-world games' influence on people's lives and potential negative effects as well as their ability to contribute to overall happiness and life satisfaction in the long term is richly deserving.

Summary

This study investigates the impact of open-world games on cognitive escapism, relaxation, and well-being among postgraduate students. Open-world games, known for their expansive, immersive environments and player autonomy, offer unique opportunities for players to escape from real-world stressors. The study used a mixed methods approach, conducting 32 in-depth exploratory interviews and surveying 609 postgraduate student players to assess the effects of these games. The findings of this study show that players reported significant cognitive escapism, with the ability to divert their thoughts from real-life concerns to engaging in-game activities. Enhanced well-being was observed, with players reporting increased relaxation, improved mood, and greater psychological satisfaction.

Conclusions

This study indicates that open-world games significantly contribute to cognitive escapism, relaxation, and overall well-being. Specifically, these games' immersive environments and player autonomy play a crucial role in reducing stress and enhancing mental health. Open-world games could be used as therapeutic tools for stress and anxiety management, offering a cost-effective and accessible method to improve mental health. Developers should consider incorporating features that promote relaxation and cognitive escapism to enhance the well-being of players. The finding that open-world games may enhance people's well-being through enhanced escapism and relaxation is not trivial, given the growing evidence that other forms of entertainment, such as traditional social media, contribute to adolescent anxiety and depression [58-60]. We invite future research to build on the findings of this study and examine the role of open-world games in people's lives further.

Acknowledgments

The authors would like to thank all the study participants for their time and insights. The authors used OpenAI's ChatGPT o1 [61] for spellchecking and grammar and not the contents of the manuscript.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the writing—review and editing, conceptualization and methodology, and data collection and analyses.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for cross-sectional studies.
[DOCX File , 30 KB - jmir_v26i1e63760_app1.docx]

References

1. Insights: youth mental health and wellbeing over time. headspace: National Youth Mental Health Foundation. 2020. URL: https://headspace.org.au/assets/Uploads/Insights-youth-mental-health-and-wellbeing-over-time-headspace-National-Youth-Mental-Health-Survey-2020.pdf [accessed 2024-12-01]
2. Huang TL, Yeh JR, Liao GY, Cheng TC, Chang YC, Teng CI. How does escapism foster game experience and game use? Decis Support Syst 2024 Jun;181:114207. [doi: 10.1016/j.dss.2024.114207]
3. Marques LM, Uchida PM, Aguiar FO, Kadri G, Santos RI, Barbosa SP. Escaping through virtual gaming-what is the association with emotional, social, and mental health? A systematic review. Front Psychiatry 2023 Nov 8;14:1257685 [FREE Full text] [doi: 10.3389/fpsyt.2023.1257685] [Medline: 38025467]

XSL•FO
RenderX

4. Stenseng F, Steinsholt IB, Hygen BW, Kraft P. Running to get "lost"? Two types of escapism in recreational running and their relations to exercise dependence and subjective well-being. Front Psychol 2022 Jan 25;13:1035196 [FREE Full text] [doi: 10.3389/fpsyg.2022.1035196] [Medline: 36760907]
5. Russoniello C, O'Brien K, Parks JM. The effectiveness of casual video games in improving mood and decreasing stress. J Cyber Ther Rehabil 2009;2(1):53-66.
6. Bediou B, Adams DM, Mayer RE, Tipton E, Green CS, Bavelier D. Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. Psychol Bull 2018 Jan;144(1):77-110. [doi: 10.1037/bul0000130] [Medline: 29172564]
7. Villani D, Repetto C, Cipresso P, Riva G. May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. Interact Comput 2012 Jul;24(4):265-272. [doi: 10.1016/j.intcom.2012.04.008]
8. Dwivedi YK, Hughes L, Wang Y, Alalwan AA, Ahn SJ, Balakrishnan J, et al. Metaverse marketing: how the metaverse will shape the future of consumer research and practice. Psychol Mark 2022 Dec 16;40(4):750-776. [doi: 10.1002/mar.21767]
9. Mogaji E, Wirtz J, Belk RW, Dwivedi YK. Immersive time (ImT): conceptualizing time spent in the metaverse. Int J Inf Manag 2023 Oct;72:102659. [doi: 10.1016/j.ijinfomgt.2023.102659]
10. Pal D, Arpnikanondt C. The sweet escape to metaverse: exploring escapism, anxiety, and virtual place attachment. Comput Hum Behav 2024 Jan;150:107998. [doi: 10.1016/j.chb.2023.107998]
11. Stenseng F, Falch-Madsen J, Hygen BW. Are there two types of escapism? Exploring a dualistic model of escapism in digital gaming and online streaming. Psychol Pop Media 2021 Jul;10(3):319-329. [doi: 10.1037/ppm0000339]
12. Kuo A, Lutz R, Hiler J. Brave new World of Warcraft: a conceptual framework for active escapism. J Consum Mark 2016 Nov 14;33(7):498-506. [doi: 10.1108/JCM-04-2016-1775]
13. Ryan RM, Rigby CS, Przybylski A. The motivational pull of video games: a self-determination theory approach. Motiv Emot 2006 Nov 29;30:344-360. [doi: 10.1007/s11031-006-9051-8]
14. Shen C, Williams D. Unpacking time online: connecting internet and massively multiplayer online game use with psychosocial well-being. Commun Res 2010 Nov 30;38(1):123-149. [doi: 10.1177/0093650210377196]
15. Granic I, Lobel A, Engels RC. The benefits of playing video games. Am Psychol 2014 Jan;69(1):66-78. [doi: 10.1037/a0034857] [Medline: 24295515]
16. Hussain Z, Griffiths MD. The attitudes, feelings, and experiences of online gamers: a qualitative analysis. Cyberpsychol Behav 2009 Dec;12(6):747-753. [doi: 10.1089/cpb.2009.0059] [Medline: 19788376]
17. Seaborn K, Fels DI. Gamification in theory and action: a survey. Int J Hum Comput Stud 2015 Feb;74:14-31. [doi: 10.1016/j.ijhcs.2014.09.006]
18. Cole H, Griffiths MD. Social interactions in massively multiplayer online role-playing gamers. Cyberpsychol Behav 2007 Aug;10(4):575-583. [doi: 10.1089/cpb.2007.9988] [Medline: 17711367]
19. Przybylski AK, Rigby CS, Ryan RM. A motivational model of video game engagement. Rev Gen Psychol 2010 Jun 01;14(2):154-166. [doi: 10.1037/a0019440]
20. Brand M, Rumpf HJ, Demetrovics Z, King DL, Potenza MN, Wegmann E. Gaming disorder is a disorder due to addictive behaviors: evidence from behavioral and neuroscientific studies addressing cue reactivity and craving, executive functions, and decision-making. Curr Addict Rep 2019 May 30;6(3):296-302. [doi: 10.1007/s40429-019-00258-y]
21. International classification of diseases 11th revision. World Health Organization. URL: https://icd.who.int/ [accessed 2024-06-22]
22. Vandelanotte C, Sugiyama T, Gardiner P, Owen N. Associations of leisure-time internet and computer use with overweight and obesity, physical activity and sedentary behaviors: cross-sectional study. J Med Internet Res 2009 Jul 27;11(3):e28 [FREE Full text] [doi: 10.2196/jmir.1084] [Medline: 19666455]
23. Pereira S, Reay K, Bottell J, Walker L, Dzikiti C. University student mental health survey 2018: a large scale study into the prevalence of student mental illness within UK universities. The Insight Network and Dig-In. 2019 Mar. URL: https://uploads-ssl.webflow.com/561110743bc7e45e78292140/5c7d4b5d314d163fecdc3706_Mental%20Health%20Report%202018.pdf [accessed 2024-06-22]
24. Ribeiro Í, Pereira R, Freire IV, de Oliveira BG, Casotti CA, Boery EN. Stress and quality of life among university students: a systematic literature review. Health Prof Educ 2018;4(2):70-77. [doi: 10.1016/j.hpe.2017.03.002]
25. Ruggeri K, Garcia-Garzon E, Maguire Á, Matz S, Huppert FA. Well-being is more than happiness and life satisfaction: a multidimensional analysis of 21 countries. Health Qual Life Outcomes 2020 Jun 19;18(1):192 [FREE Full text] [doi: 10.1186/s12955-020-01423-y] [Medline: 32560725]
26. Alsaad F, Binkhamis L, Alsalman A, Alabdulqader N, Alamer M, Abualait T, et al. Impact of action video gaming behavior on attention, anxiety, and sleep among university students. Psychol Res Behav Manag 2022 Jan;Volume 15:151-160. [doi: 10.2147/prbm.s347694]
27. Fazeli S, Zeidi IM, Lin CY, Namdar P, Griffiths MD, Ahorsu DK, et al. Corrigendum to "depression, anxiety, and stress mediate the associations between internet gaming disorder, insomnia, and quality of life during the COVID-19 outbreak" [Addict. Behav. Rep. 12 (2020) 100307]. Addict Behav Rep 2020 Dec;12:100322 [FREE Full text] [doi: 10.1016/j.abrep.2020.100322] [Medline: 33364329]

XSL•FO
RenderX

28. Glaser BG, Strauss AL. The Discovery of Grounded Theory: Strategies for Qualitative Research. New York, NY: Aldine Publishing Company; 1967.
29. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
30. Edwards R, Holland J. What is Qualitative Interviewing?. London, UK: Bloomsbury Publishing; 2013.
31. Zeithaml VA, Jaworski BJ, Kohli AK, Tuli KR, Ulaga W, Zaltman G. A theories-in-use approach to building marketing theory. J Mark 2019 Nov 14;84(1):32-51. [doi: 10.1177/0022242919888477]
32. Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research: notes on the Gioia methodology. Organ Res Methods 2012 Jul 24;16(1):15-31. [doi: 10.1177/1094428112452151]
33. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
34. Ryff CD. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. J Pers Soc Psychol 1989;57(6):1069-1081. [doi: 10.1037//0022-3514.57.6.1069]
35. Foa EB, Kozak MJ. Emotional processing of fear: exposure to corrective information. Psychol Bull 1986;99(1):20-35. [doi: 10.1037//0033-2909.99.1.20]
36. Anderson JC, Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach. Psychol Bull 1988;103(3):411-423. [doi: 10.1037//0033-2909.103.3.411]
37. Bagozzi RP, Yi Y. On the evaluation of structural equation models. J Acad Mark Sci 1988 Mar;16:74-94. [doi: 10.1007/bf02723327]
38. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.2307/3151312]
39. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: Guilford Press; 2017.
40. Hartnup B, Dong L, Eisingerich AB. How an environment of stress and social risk shapes student engagement with social media as potential digital learning platforms: qualitative study. JMIR Med Educ 2018 Jul 13;4(2):e10069 [FREE Full text] [doi: 10.2196/10069] [Medline: 30006324]
41. Anto A, Asif RO, Basu A, Kanapathipillai D, Salam H, Selim R, et al. Exploring the impact of social media on anxiety among university students in the United Kingdom: qualitative study. JMIR Form Res 2023 Jun 16;7:e43037 [FREE Full text] [doi: 10.2196/43037] [Medline: 37327030]
42. Anto A, Asif RO, Basu A, Kanapathipillai D, Salam H, Selim R, et al. Exploring the facilitators and barriers to addressing social media's impact on anxiety within primary care: a qualitative study. BJGP Open 2024 Jul 25;8(2):BJGPO.2023.0190 [FREE Full text] [doi: 10.3399/BJGPO.2023.0190] [Medline: 38272495]
43. Eisingerich AB, Chun HH, Liu Y, Jia H, Bell SJ. Why recommend a brand face‐to‐face but not on Facebook? How word‐of‐mouth on online social sites differs from traditional word‐of‐mouth. J Consum Psychol 2014 May 29;25(1):120-128. [doi: 10.1016/j.jcps.2014.05.004]
44. Eisingerich AB, Marchand A, Fritze MP, Dong L. Hook vs. hope: how to enhance customer engagement through gamification. Int J Res Mark 2019 Jun;36(2):200-215. [doi: 10.1016/j.ijresmar.2019.02.003]
45. El-Hilly AA, Iqbal SS, Ahmed M, Sherwani Y, Muntasir M, Siddiqui S, et al. Game on? Smoking cessation through the gamification of mHealth: a longitudinal qualitative study. JMIR Serious Games 2016 Oct 24;4(2):e18 [FREE Full text] [doi: 10.2196/games.5678] [Medline: 27777216]
46. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018 Jun 15;6(6):e10024 [FREE Full text] [doi: 10.2196/10024] [Medline: 29907557]
47. Eisingerich AB, Kretschmer T. In e-commerce, more is more. Harvard Business Review. 2008 Mar. URL: https://hbr.org/2008/03/in-e-commerce-more-is-more [accessed 2024-12-01]
48. Tudor-Sfetea C, Rabee R, Najim M, Amin N, Chadha M, Jain M, et al. Evaluation of two mobile health apps in the context of smoking cessation: qualitative study of cognitive behavioral therapy (CBT) versus non-CBT-based digital solutions. JMIR Mhealth Uhealth 2018 Apr 18;6(4):e98 [FREE Full text] [doi: 10.2196/mhealth.9405] [Medline: 29669708]
49. Lin YT, MacInnis DJ, Eisingerich AB. Strong anxiety boosts new product adoption when hope is also strong. J Mark 2020 Jul 10;84(5):60-78. [doi: 10.1177/0022242920934495]
50. Eisingerich AB, MacInnis DJ, Whan Park C. Do CSR efforts that focus on helping the environment influence brand purchase more than other forms of CSR? J Bus Res 2023 Nov;168:114240. [doi: 10.1016/j.jbusres.2023.114240]
51. Lin YT, Foscht T, Eisingerich AB. A time for heroes? Conceptualization, development and validation of the brand hero scale. Eur J Mark 2023 May 15;57(13):1-26. [doi: 10.1108/ejm-09-2021-0700]
52. Bell SJ, Auh S, Eisingerich AB. Unraveling the customer education paradox: when, and how, should firms educate their customers? J Serv Res 2017 Feb 15;20(3):306-321 [FREE Full text] [doi: 10.1177/1094670517691847] [Medline: 22973420]
53. Sun X, Foscht T, Eisingerich AB. Does educating customers create positive word of mouth? J Retail Consum Serv 2021 Sep;62:102638. [doi: 10.1016/j.jretconser.2021.102638]

XSL•FO
RenderX

54. Foscht T, Lin Y, Eisingerich AB. Blinds up or down? The influence of transparency, future orientation, and CSR on sustainable and responsible behavior. Eur J Mark 2018 Feb 02;52(3/4):476-498. [doi: 10.1108/EJM-10-2016-0576]
55. Eisingerich AB, MacInnis DJ, Whan Park C. A theoretical framework exploring three foundational benefits of brand attachment. In: King C, Murillo E, editors. A Research Agenda for Brand Management in a New Era of Consumerism. Cheltenham, UK: Edward Elgar Publishing; 2023:161-175.
56. Fritze MP, Marchand A, Eisingerich AB, Benkenstein M. Access-based services as substitutes for material possessions: the role of psychological ownership. J Serv Res 2020 Mar 13;23(3):368-385. [doi: 10.1177/1094670520907691]
57. Fritze MP, Eisingerich AB, Benkenstein M. Digital transformation and possession attachment: examining the endowment effect for consumers' relationships with hedonic and utilitarian digital service technologies. Electron Commer Res 2018 Jun 11;19(2):311-337 [FREE Full text] [doi: 10.1007/s10660-018-9309-8] [Medline: 26811110]
58. Haidth J. The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness. London, UK: Penguin Books; 2024.
59. Varma P, Junge M, Meaklim H, Jackson ML. Younger people are more vulnerable to stress, anxiety and depression during COVID-19 pandemic: a global cross-sectional survey. Prog Neuropsychopharmacol Biol Psychiatry 2021 Jul 13;109:110236 [FREE Full text] [doi: 10.1016/j.pnpbp.2020.110236] [Medline: 33373680]
60. Shannon H, Bush K, Villeneuve PJ, Hellemans KG, Guimond S. Problematic social media use in adolescents and young adults: systematic review and meta-analysis. JMIR Ment Health 2022 Apr 14;9(4):e33450 [FREE Full text] [doi: 10.2196/33450] [Medline: 35436240]
61. ChatGPT o1. OpenAI. URL: https://openai.com/o1/ [accessed 2024-12-03]

Abbreviations

RA: research assistant
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by T de Azevedo Cardoso; submitted 28.06.24; peer-reviewed by YT Lin; comments to author 01.08.24; revised version received 05.08.24; accepted 08.11.24; published 17.12.24.

Please cite as:
Anto A, Basu A, Selim R, Foscht T, Eisingerich AB
Open-World Games' Affordance of Cognitive Escapism, Relaxation, and Mental Well-Being Among Postgraduate Students: Mixed Methods Study
J Med Internet Res 2024;26:e63760
URL: https://www.jmir.org/2024/1/e63760
doi:10.2196/63760
PMID:

©Ailin Anto, Arunima Basu, Rania Selim, Thomas Foscht, Andreas Benedikt Eisingerich. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors That Influence Young Adults' Preferences for Virtual Reality Exergames in a Weight Control Setting: Qualitative Study

Yanya Chen[1,2*], MSN; Lina Guan[1*], BSN; Weixin Wu[3*], BSPH; Liang Ye[2], MSPH; Yan He[2], MSN; Xiaofen Zheng[4], MS; Sicun Li[2], MS; Bingsheng Guan[5*], PhD; Wai-kit Ming[2*], PhD

[1]School of Nursing, Jinan University, Guangzhou, China

[2]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[4]School of Medicine, Jinan University, Guangzhou, China

[5]The First Affiliated Hospital of Jinan University, Guangzhou, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Wai-kit Ming, PhD
Department of Infectious Diseases and Public Health
Jockey Club College of Veterinary Medicine and Life Sciences
City University of Hong Kong
To Yuen Building
31 To Yuen Street
Hong Kong
China (Hong Kong)
Phone: 852 56117116
Email: wkming2@cityu.edu.hk

Abstract

Background: Obesity could compromise people's health and elevate the risk of numerous severe chronic conditions and premature mortality. Young adults are at high risk of adopting unhealthy lifestyles related to overweight and obesity, as they are at a phase marked by several significant life milestones that have been linked to weight gain. They gain weight rapidly and excess adiposity mostly accrues, compared with middle-aged and older adults when weight stabilizes or even decreases. Virtual reality exergames have the potential to increase physical activity in people's daily lives. However, the factors that influence young adults' preference for using virtual reality exergames for weight control remain unclear.

Objective: The objective of this study is to identify, characterize, and explain the factors influencing young adults' preference for weight control using virtual reality exergames.

Methods: This qualitative study used semistructured interviews. In total, 4 focus group interviews were conducted with 23 young adults aged between 18 and 25 years. The qualitative data were analyzed using the Colaizzi phenomenological methodology.

Results: In total, 3 major factors were found to influence young adults' preference for virtual reality exergames in weight control settings: individual factors, social or environmental factors, and expectations of virtual reality exergames. Individual factors included experience with previous weight control methods, previous experience with virtual reality, psychological status, attitudes toward personal BMI, preference for exercise type, and acceptance of virtual reality exergames. Social or environmental factors included social definitions of beauty, weather or public health events, and knowledge of virtual reality provided. Expectations of virtual reality exergames included cost of the device, the fun of virtual reality exergames, supervision, modality of virtual reality exergames, feedback after exercise, convenience to use, and weight loss effect.

Conclusions: Young adults take various factors into account when deciding whether to use virtual reality exergames for weight control. These factors can inform the development and further refinement of devices, guides, and policies related to virtual reality exergames for controlling weight.

(J Med Internet Res 2024;26:e58422) doi:10.2196/58422

XSL•FO
RenderX

KEYWORDS

virtual reality; games; weight control; preferences; young adults; qualitative research

Introduction

Obesity could compromise people's health and elevate the risk of numerous severe chronic conditions and premature mortality [1]. An estimated 5 million people died in 2019 from noncommunicable diseases caused by higher-than-optimal BMI, including diabetes, cancer, cardiovascular disease, digestive disorders, neurological disorders, and chronic respiratory diseases [2,3]. Besides serious threats to people's health, it was predicted that the global costs of overweight and obesity would reach US$ 3 trillion per year by 2030 and will exceed US$ 18 trillion by 2060 if no action is taken [4]. Over the last 2 decades, obesity prevalence has doubled worldwide, with 43% (2.5 billion) of adults aged 18 years and older being overweight and 16% (890 million) being obese in 2022 [5]. Especially, young adults (aged 18-25 years) are at high risk of adopting unhealthy lifestyles leading to overweight and obesity. They are at a phase marked by several significant life milestones that have been linked to weight gain. These milestones include embarking on a college education, entering into marriage, and initiating the journey of starting a family [6]. At the young adult stage, adults gain weight rapidly and excess adiposity mostly accrues, compared with middle-aged and older adults when weight stabilizes or even decreases [7]. According to the report of the World Obesity Federation, the decrease in physical activity and a lack of exercise have contributed to a global rise in overweight or obesity [8,9]. In 2016, the prevalence of insufficient physical activity increased steadily and reached 28.5%, and over a quarter of adults (27.5%) were insufficiently physically active [10]. Global physical activity targets in 2025 (reducing insufficient physical activity by 10%) will not be achieved if current trends continue [10]. In order to increase population levels of physical activity, measures and policies must be urgently prioritized and scaled up.

The term exergames (a portmanteau of "exercise" and "games") referred to technology-based physical activities where people could move their entire bodies for the purpose of playing [11]. As an emerging field in technology, virtual reality had attracted attention for its ability to enhance participation [12]. It allowed users to gain real-world experiences by interacting with it in a similar way to what they would experience in the real world [13]. Although the definition of exergames was different from that of virtual reality, there was an overlap between them for health [14]. That was to say, some exergames could consist of virtual reality content, while some did not [15]. These years have seen the popularity, acceptance, affordability, and accessibility of some commercial virtual reality exergames, including Nintendo Switch, Microsoft Xbox Kinect, and Wii sports [16]. Virtual reality exergames could create a presence where the users could be fully immersed in the virtual environment. Compared with traditional aerobic exercise, the adherence to playing virtual reality exergames was reported to increase by 30% in terms of attendance at physical activity, emphasizing the potential of virtual reality exergames to motivate individuals with overweight to participate in positive activities more regularly [17]. Besides promoting light- to moderate-intensity physical activity, virtual reality exergames could even be used as an effective exercise strategy for meeting physical recommendations [18]. Currently, virtual reality exergames were reported to help Parkinson's disease patients to increase their balance skills and individuals with chronic neck pain to improve their depression and pain interference [19,20]. They have also been used for overweight and obese people, demonstrating the effect of decreasing BMI among overweight middle-aged women and increasing energy expenditure in children and adolescents with overweight or obesity [12,21]. However, to our knowledge, systematic factors that influence young adults' preferences for virtual reality exergames in a weight-control setting remain unclear. To develop and further refine the devices, guides, and policies for weight control–related virtual reality exergames, we need to better understand the factors that influence young adults' preferences. Hence, this study aimed to identify, characterize, and explain the factors influencing young adults' preference for virtual reality exergames in weight-control settings through a qualitative study.

Methods

Participants and Study Design

This qualitative study was conducted on young adults from a community in Guangzhou (Canton), making sure that their environmental factors and sociocultural backgrounds are relatively consistent. A convenience sample of young adults was used due to the small number of participants available. The inclusion criteria were shown as follows: (1) the participants were required to be within the age range of 18-25 years, (2) not having a self-reported diagnosed physical or mental disease, and (3) voluntary participation. Recruitment began with posting the advertisement on the WeChat Official Account Platform. Then, interested potential participants get in touch with us by scanning the QR code on the advertisement to add our principal investigator on WeChat. Invitation letters were sent to potential participants who provided contact details and met the inclusion criteria, explaining the purpose of the study and asking if they would like to participate. Finally, 23 young adults (ie, 12 male and 11 female) fulfilled the inclusion criteria and took part in the study. Participants received a certain amount of honoraria (50 RMB [US $6.85] per individual) after the interview.

Compared with one-on-one interviews, focus groups are premised on the assumption that group interactions facilitate the expression of interviewees' viewpoints, as participants tend to feel safer and less apprehensive when discussing challenging experiences [22]. Hence, since focus group interviews can provide a more in-depth understanding of one's feelings or views regarding a particular topic, they were adopted to explore the factors that influenced young adults' preferences in this study. Based on previous literature [22], the number of participants in focus group interviews was typically 5-6 participants.

XSL•FO
RenderX

Ethical Considerations

Ethical approval was obtained from the institutional review board of City University of Hong Kong (reference number: HU-STA-00001075). Participation was strictly voluntary. Interviews were conducted only after participant consent had been obtained. The purpose of the research, the methodology, and the procedures involved were explained to the participants. Without any setbacks, participants could withdraw at any time. Additionally, we protected the participants' privacy with strict confidentiality measures. The information of participants was anonymized during transcription and only accessible to authorized members of the research team.

Data Collection

Data were collected in September and October 2023. Individuals interested in participating in the study were required to contact the principal investigator to arrange for data collection. Qualitative data collection was conducted in a quiet meeting room at a university conveniently located near the participants' community. Participants sat comfortably in a circle so that facial expressions and body posture could be used to convey information. Each participant was informed that they could withdraw at any time. Focus group interviews were conducted using the Mandarin language in a semistructured interview format and recorded (the guiding questions are listed in Table 1). The interview comprised 2 distinct sections. During the initial section, participants explained their perceptions of previous weight control methods and virtual reality. In the subsequent section, the researchers offered a concise summary of the present situation of virtual reality exergames in the realm of weight management, ensuring that all participants attained a fundamental and comprehensive understanding. Respondents could then articulate their expectations and ideas about the application of virtual reality in weight control programs through conversations. As the data had reached saturation, we decided not to recruit other participants after interviewing 23 participants. Each interview lasted approximately 30 minutes and was video recorded and transcribed by a professional transcription service.

Table 1. Focus group outline.

Part	Number	Question
1	1	Please share your feelings about previous weight control methods.
1	2	Have you experienced something related to virtual reality?
2	1	If virtual reality exergames could be used to control weight, would you like to try it?
2	2	If virtual reality exergames could be used to control weight, what factors would you deem important?
2	3	Do you have any suggestions on when virtual reality exergames will be used to control weight?
2	4	Is there any additional information you wish to include in the interview content?

Data Analysis

The Colaizzi phenomenological methodology and NVivo (version 14; Lumivero) software were adopted to analyze the qualitative data. The Colaizzi analytical approach encompassed seven distinct steps: (1) understanding the content and significance of each transcript, (2) identifying noteworthy or key from the narrative, (3) deriving or building significance from serious statements, (4) dividing explicit meanings into thematic clusters, (5) providing a detailed depiction of the phenomenon, (6) identifying the potential framework for this phenomenon, and (7) offering exhaustive descriptions to the participants. The findings of the paper were presented in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) to guarantee transparency and reproducibility of our methods.

Results

Overview

This study involved 23 participants (ie, 12 male and 11 female) aged 18-25 years (Table 2). The focus group interviews were conducted with 4 separate groups of participants. In total, three main themes emerged: (1) individual factors, (2) social or environmental factors, and (3) expectations of virtual reality exergames. These main themes were further categorized into 16 subthemes (Table 3).

XSL•FO
RenderX

Table 2. Characteristics of included young adults (N=23).

Characteristics	Values
Age (years), mean (SD)	21.04 (1.988)
BMI (kg/m^2), mean (SD)	25.97 (4.244)
Gender, n (%)	
Male	12 (52.2)
Female	11 (47.8)
Setting, n (%)	
Rural	12 (52.2)
Urban	11 (47.8)

Table 3. Themes and subthemes.

Theme	Subtheme
Individual factors	• Experience with previous weight control methods • Previous experience with virtual reality • Psychological status • Attitudes toward personal BMI • Preference of exercise type • Acceptance of virtual reality exergames
Social or environmental factors	• Social definition of beauty • Weather or public health event • Knowledge of virtual reality provided
Expectations of virtual reality exergames	• Cost of the device • Fun of virtual reality exergames • Supervision • Modality of virtual reality exergames • Feedback after exercise • Convenience to use • Weight loss effect

Individual Factors

Experience With Previous Weight Control Methods

Previous weight control methods included diet, exercise, physical therapy, and traditional Chinese medicine. Some young adults were confused as they could not find a suitable weight loss method. Several participants had difficulty keeping the weight off. Exercise as a method of weight control was tedious, and it was difficult to participate in others' “sports teams suddenly” sometimes.

> *I feel like every sport has a threshold. Picking up a new sport, such as badminton and basketball... can lead to difficulty integrating myself into other experienced players.*
>
> *Dancing to web-based videos can be dull, but to control my weight, I would endure the boredom. If possible, I would choose more interesting ways of exercising.*

Previous Experience With Virtual Reality

Approximately 50% of interviewees had experienced virtual reality before, and most participants preferred to use virtual reality exergames to control weight since they regarded it as interesting, immersive, and engaging. After experiencing it, a minority of participants considered virtual reality exergames uncomfortable and dizzy, so they lost interest.

> *I experienced virtual reality in science museums, viewing snowy mountain scenery from a high altitude with virtual reality glasses. I was engrossed in the novelty. Perhaps it would be more appealing to control weight via VR than outdoor sports.*
>
> *Now it's (virtual reality) quite realistic. I played a similar roller coaster simulation game before. It simulates the perspective of you sitting on a roller coaster. Although you don't feel the actual weightlessness, when it rushes down, it still feels quite realistic…very willing to try (virtual reality exergames)*

XSL•FO
RenderX

Psychological Status

Some young adults were perceived as introverted and unwilling to go out. Therefore, they were partial to exercise at home, adopting virtual reality exergames. Some young adults with large BMI, low self-esteem, and sensitivity favored controlling weight at home through virtual reality exergames, avoiding encountering acquaintances in exercise procedures.

> *Some people like me simultaneously possess a social butterfly identity online and a socially awkward persona in real life…we feel comfortable approaching strangers in video games and becoming acquainted virtually. However, if we encounter someone on the nearby treadmill in a gym, it is impossible for us to walk up and start a conversation. I prefer online workout buddies as long as our relationship serves to motivate exercising.*

> *I'm an 'i-person' (introverted person). I just don't like going out. Virtual reality exergames feel more inclusive for people like us.*

Attitudes Toward Personal BMI

Some young adults believed their weight was acceptable or did not care about others' comments about their figures. Accordingly, there was no intense aspiration to use virtual reality exergames to control weight.

> *I have a 'let it be' mentality…Regardless of my size, this is my body… I do not have a strong motivation for weight intervention except it affects my health…whether through virtual reality exergames or outdoor activities.*

> *I guess others might have negative opinions about me, but I don't care. If there are virtual reality exergames, I'll use them; if not, it's no big deal.*

Preference of Exercise Type

Currently, virtual reality exergames offer a wide range of sports options, encompassing football, badminton, fencing, and dancing. However, the type of physical activity could influence young adults' usage and adherence to virtual reality exergames for weight management. For instance, female-identified persons preferred dance-based games within the virtual reality exercise realm while disinterested in excessively aggressive game genres. On the other hand, male-identified persons leaned toward ball games or fencing as their preferred forms of virtual reality exergames. Nevertheless, both genders expressed reluctance toward using virtual reality glasses for exercising through virtual reality exergames.

> *The duration of applying virtual reality exergames can last for one month or year… depends on if we are entertained by the type of game and exercise.*

Acceptance of Virtual Reality Exergames

More than half of the young adults were willing to adopt virtual reality exergames for weight control, although a few participants did not agree that it had good results due to poor virtual reality experience and other reasons. Participants with a positive attitude believed that virtual reality exergames could motivate people to exercise more consistently by offering a variety of exercise options and fun experiences. They also found that these games were challenging enough to leave them feeling fatigued after working out. This supportive attitude would provide a basis for young adults to use virtual reality exergames for weight control.

> *In fact, the greatest challenge about weight control is persistence. As long as the game is well designed, engaging and immersive, surely it will be a boost for those people having trouble sticking to their workouts, so why not opt for such a form of exercise?*

Social or Environmental Factors

Social Definition of Beauty

Almost all interviewees agreed that society defined beauty as being slender. This popular aesthetic trend became a driving force for some people to exercise, making them willing to try a new exercising way that was easier to stick to and had fun.

> *Clothes sizes are also designed smaller and smaller these days, guiding people to standardize beauty with slenderness. In order to look more attractive (slim), I have to exercise, and the virtual reality exergames can be an option for me.*

> *Usually, because we study in Guangdong (Canton), there are many local students here, and indeed, the girls from Guangdong are generally petite and very slim. Standing next to them, I can't help but feel a bit envious... I still want to try various weight control methods as long as they are effective.*

Weather or Public Health Event

Many interviewees said that when they encountered weather conditions such as rain, cold winters, and public health events such as the COVID-19 pandemic, which prevented people from going out to exercise, they would choose the virtual reality exergames to exercise at home.

> *During COVID-19, I was required to quarantine at a hotel for 14 days and at home for another 7 days, and because of the lack of exercise and the greasy meals, I gained weight very quickly, but I was not allowed to get out of the house, and at that time I was thinking how nice it would be to have a set of consoles such as the Nintendo, that would be good to both kill time and control my weight.*

Knowledge of Virtual Reality Provided

About half of the interviewees were unclear about the definition of virtual reality, not to mention how exercise was done through virtual reality. In total, 4 interviewees said that they usually did not have access to knowledge related to virtual reality and did not understand the development of virtual reality, so they hesitated to use the virtual reality exergames for weight control.

> *I don't know much about new technology like virtual reality, ... also not interested in virtual reality exergames.*

XSL•FO
RenderX

For new technologies like this (virtual reality), I feel that there hasn't been a widespread way to access them in the past. Currently, most of the promotion focuses on exercise, and short videos make it quite impactful. So, it seems that diet and exercise are still the main methods.

Expectations of Virtual Reality Exergames

Cost of the Device

For young adults, their income was not high; most of them were students and still received living expenses from their parents. Some interviewees knew that some interesting virtual games could be used for exercise and weight control. However, they could not afford expensive virtual gaming equipment. Therefore, they said that they would only favor exercising in this way if the equipment cost could be a bit more affordable.

As a junior in college, I get all my living expenses from my parents. Therefore, I don't want to buy expensive gaming equipment, even though it might help me lose weight. I can just run a few laps on the track field by myself.... I would only choose to use it if (the gaming device) was cheaper.

Fun of Virtual Reality Exergames

The fun of virtual reality exergames was one of the important factors influencing young adults' choice of weight control measures. Most participants highlighted that the gratifying and engaging nature of these games would play a role in maintaining their engagement. However, the fun aspect tends to diminish if players are restricted to a single game type for too long, particularly when it ceases to offer new challenges.

In team sports such as volleyball and bowling... a lack of strong interaction or a failure to meet expectations could lead to infrequent participation.

Supervision

For the majority of interviewees, supervision provided by virtual reality exergames was essential in managing weight. They were more likely to frequently use these games under supervision, but in its absence, their commitment tends to wane.

Given specific daily check-in tasks, I'm likely to adhere to them; however, in the absence of such directives, I tend not to engage in virtual reality exercises at home.

If there's no check-in or supervision, you just won't use it.

Modality of Virtual Reality Exergames

Most interviewees revealed dislike for exercising alone, emphasizing their commitment to exercise increases with the presence of family or friends. This sentiment held true for managing weight through virtual reality exergames, where they valued the companionship of family and friends.

Nintendo offers games like badminton, which initially seemed uninteresting when I played alone…… But during the Chinese New Year festivities, engaging in competitive play with friends transformed the experience, leading me to maintain regular exercise throughout the winter break surprisingly.

Feedback After Exercise

Many participants hoped that the virtual reality exergames system could provide some feedback, encompassing win-loss records, heart rate, blood pressure, etc. Being aware of the positive impact of their exercise or favorable game results, a portion of these individuals might experience a sense of success.

Receiving feedback is undeniably better than not getting any... Witnessing my heart rate rise in the workout summary assures me that I'm actually exercising.

This (virtual reality exergames) needs to provide feedback after exercise for me to use it. Otherwise, without any kind of reference point, I wouldn't know how much or how well I'm exercising.

Convenience to Use

In total, 3 participants mentioned that the user-friendliness and portability of the gaming system influenced their inclination to use virtual reality exergames.

I lean towards more compact models, which allow me to easily take the equipment along even when I'm traveling for a few days.

I might start by trying it out... Also, the convenience of the equipment is important; otherwise, it would be difficult for me to stick with using it(virtual reality exergames) to control my weight.

Weight Loss Effect

In total, 7 interviewees noted that success in weight loss impacts their choice to use virtual reality exergames for weight management. This is especially significant for those with a higher BMI, for whom the effectiveness in controlling weight is a major factor.

If weight loss is ineffective, the VR game is mere entertainment, lacking its promised weight reduction functionality. My initial purpose for buying the game was to sustain a regular exercise routine. It feels like a waste of money if it fails to give significant results.

Discussion

Principal Findings

This qualitative study identified, characterized, and explained the factors influencing young adults' preference for weight control using virtual reality exergames. Through interviewing 23 young adults, it was concluded that the respondents were mainly affected by individual factors, social or environmental factors, and expectations of virtual reality exercise games.

The effect of weight loss was one of the important factors affecting whether or not young adults choose virtual reality exergames for weight control. As Comeras-Chueca et al [23] reported, different types of virtual reality exergames produced different weight loss effects and were no worse than traditional exercise. Graves et al [24] examined the energy expended while

XSL•FO
RenderX

playing 3 distinct Wii Sports games and demonstrated that boxing resulted in the highest expenditure of energy (4.05 kcal/min), surpassed by tennis (3.05 kcal/min) and bowling (2.75 kcal/min). Similar results also revealed that the energy expended during Dance Dance Revolution and Wii Sports boxing was comparable to that of walking at a moderate intensity (5.7 km/h) [25]. Furthermore, virtual reality exergames could be an effective way for obesity management among overweight middle-aged women, as the results of the research showed that the mean BMI significantly decreased from 26.04 to 24.65 kg/m^2 after 8 weeks of the virtual reality exergames program [12]. Meanwhile, these virtual reality exergames aim to enhance the fun and enjoyment of physical activities by providing a virtual environment that is tailored to the preferences of young adults, especially for those with obesity and low motor skills [23]. These make those young adults feel more agile, more skilled, and more powerful, so that these young adults will have more desire and more interest in practicing other sports [23]. Therefore, the virtual reality exergames caught the attention of young adults who are reluctant to engage in traditional physical activities. That is to say, virtual reality exergames change the attitude of young adults from rejection to habit, helping them get a more active lifestyle at the same time.

The results of this study showed that most young adults preferred to use virtual reality exergames in the company of friends or family. Young adults consumed more energy when playing virtual reality exergames with a peer [26]. Second, with the engagement of friends and family members, young adults' adherence to exercise would increase. In the virtual world, the companionship of friends or family could boost individuals' perceived competency, satisfaction, and self-belief in executing group tasks [27]. The social interaction might provide feedback and recognition, resulting in a sense of satisfaction when their accomplishments are noted and appreciated by others [28]. When playing virtual reality exergames socially with peers, it was encouraged to engage in more intense and sustained physical activity throughout each gaming session, particularly in the posterior regions [29]. It was speculated that team cohesion enhanced adherence to an exercise regimen when individuals engage in physical activities as part of a group [30]. Third, social support and self-esteem increased when people worked out in teams, which served as motivation and enabled sustained physical activities [31].

The cost of the device influenced young adults' preference for weight control using virtual reality exergames. Virtual reality exergames did not reach all overweight and obese adolescents due to their high cost; the software and hardware were expensive [32]. For example, the cost of an average head-mounted virtual reality was still relatively high (average CAD $ 585 [US$ 460]) [33], which limited the use of virtual reality exergames for weight loss in a subset of the population. Especially 18- to 25-year-old young adults who are studying or just entering work do not have too much income. However, with the growing interest in virtual reality exergames, young adults spend more and more on them. Statistics reveal that in 2017, US consumers spent a total of 29.1 billion dollars on video games, marking an increase of 10 billion dollars compared to the previous 5-year period [34]. In addition, with technology development, virtual reality exergames development costs may be getting lower and lower. The cost will not be the main factor that influences young adults' preferences in the future.

This study had one strength of being a detailed qualitative analysis with a limited number of participants, enabling a deeper understanding of their perceptions. Through the in-depth understanding of young adults' perspective, the results can benefit the development and further refinement of devices, guides, and policies related to virtual reality exergames for controlling weight. An additional strength of this study was using focus groups as a methodology. The direct interaction between the participants in one group enabled participants to ask the key questions and explore a variety of opinions. This study also had some limitations. First, this study may lack representativeness of the general population due to the non-random selection of participants and potential failure to capture the population's diverse characteristics. Future research should address this limitation by using larger sample sizes and mixed methods, such as surveys and focus groups. Another limitation of this study was that it was conducted in a single community and might not translate to other areas. This should be overcome by seeking variation among participants and paying attention to more general factors that can be applied to other areas in future studies.

Conclusions

The results indicated that young adults consider complex factors when considering whether to use virtual reality exergames for weight control. These included individual factors such as the experience of previous weight control methods, previous experience with virtual reality, psychological status, etc, social or environmental factors such as the social definition of beauty, weather or public health events, etc, and expectations of virtual reality exergames. The results of this study can inform the development and further refinement of devices, guides, and policies related to virtual reality exergames for controlling weight.

Acknowledgments

This research was partially supported by SIRG-CityU Strategic Interdisciplinary Research (grant 7020093); the Basic and Applied Basic Research Project of Guangzhou Basic Research Program (2023A04J1918); the scientific research project of the Chinese medicine bureau in Guangdong province (20241064); and the University Youth Innovative Talent Project of Guangdong Province (2023KQNCX006).

XSL•FO
RenderX

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CYY, GLN, and WWX made equal contributions as the co–first authors of this manuscript. MWK and GBS, as the corresponding authors, made equal contributions to this manuscript. MWK, GBS, and CYY led the study concept and design. GLN, WWX, and ZXF performed data collection. MWK, GBS, CYY, YL, and HY performed data analysis. CYY, GLN, WWX, and LSC prepared the manuscript. The final draft was revised and approved by all authors.

Conflicts of Interest

None declared.

References

1. Johnston JD, Massey AP, Devaneaux CA. Innovation in weight loss programs: a 3-dimensional virtual-world approach. J Med Internet Res 2012;14(5):e120 [FREE Full text] [doi: 10.2196/jmir.2254] [Medline: 22995535]
2. Fu R, Zhang Y, Yu K, Mao D, Su H. Bariatric surgery alleviates depression in obese patients: a systematic review and meta-analysis. Obes Res Clin Pract 2022;16(1):10-16. [doi: 10.1016/j.orcp.2021.11.002] [Medline: 34802982]
3. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet 2020;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
4. Okunogbe A, Nugent R, Spencer G, Powis J, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for 161 countries. BMJ Glob Health 2022;7(9):e009773 [FREE Full text] [doi: 10.1136/bmjgh-2022-009773] [Medline: 36130777]
5. World Health Organization. Obesity and Overweight. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight[accessed [accessed 2024-07-04]
6. Liu E, Chang SH. Self-esteem and weight status of young adults: findings from a pilot study. J Educ Health Promot 2022;11:263 [FREE Full text] [doi: 10.4103/jehp.jehp_1195_21] [Medline: 36325221]
7. Chen C, Ye Y, Zhang Y, Pan XF, Pan A. Weight change across adulthood in relation to all cause and cause specific mortality: prospective cohort study. Br Med J 2019;367:l5584 [FREE Full text] [doi: 10.1136/bmj.l5584] [Medline: 31619383]
8. World obesity atlas 2023. World Obesity Federation. URL: https://data.worldobesity.org/publications/?cat=19(accessed [accessed 2024-03-20]
9. Huang J, Chen Y, Wang X, Wang C, Yang J, Guan B. Change in adipokines and gastrointestinal hormones after bariatric surgery: a meta-analysis. Obes Surg 2023;33(3):789-806. [doi: 10.1007/s11695-022-06444-8] [Medline: 36607567]
10. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018;6(10):e1077-e1086 [FREE Full text] [doi: 10.1016/S2214-109X(18)30357-7] [Medline: 30193830]
11. Abd-Alrazaq A, Alhuwail D, Al-Jafar E, Ahmed A, Shuweihdi F, Reagu SM, et al. The effectiveness of serious games in improving memory among older adults with cognitive impairment: systematic review and meta-analysis. JMIR Serious Games 2022;10(3):e35202 [FREE Full text] [doi: 10.2196/35202] [Medline: 35943792]
12. Seo EY, Kim YS, Lee YJ, Hur MH. Virtual reality exercise program effects on body mass index, depression, exercise fun and exercise immersion in overweight middle-aged women: a randomized controlled trial. Int J Environ Res Public Health 2023;20(2):900 [FREE Full text] [doi: 10.3390/ijerph20020900] [Medline: 36673656]
13. García-Bravo S, Cano-de-la-Cuerda R, Domínguez-Paniagua J, Campuzano-Ruiz R, Barreñada-Copete E, López-Navas MJ, et al. Effects of virtual reality on cardiac rehabilitation programs for ischemic heart disease: a randomized pilot clinical trial. Int J Environ Res Public Health 2020;17(22):8472 [FREE Full text] [doi: 10.3390/ijerph17228472] [Medline: 33207670]
14. Bond S, Laddu DR, Ozemek C, Lavie CJ, Arena R. Exergaming and virtual reality for health: implications for cardiac rehabilitation. Curr Probl Cardiol 2021;46(3):100472. [doi: 10.1016/j.cpcardiol.2019.100472] [Medline: 31606141]
15. Ko J, Jang SW, Lee HT, Yun HK, Kim YS. Effects of virtual reality and non–virtual reality exercises on the exercise capacity and concentration of users in a ski exergame: comparative study. JMIR Serious Games 2020;8(4):e16693. [doi: 10.2196/16693] [Medline: 33112240]
16. Yen HY, Chiu HL. Virtual reality exergames for improving older adults' cognition and depression: a systematic review and meta-analysis of randomized control trials. J Am Med Dir Assoc 2021;22(5):995-1002. [doi: 10.1016/j.jamda.2021.03.009] [Medline: 33812843]
17. Warburton DER, Bredin SSD, Horita LTL, Zbogar D, Scott JM, Esch BTA, et al. The health benefits of interactive video game exercise. Appl Physiol Nutr Metab 2007;32(4):655-663. [doi: 10.1139/H07-038] [Medline: 17622279]
18. Peng W, Lin JH, Crouse J. Is playing exergames really exercising? A meta-analysis of energy expenditure in active video games. Cyberpsychol Behav Soc Netw 2011;14(11):681-688. [doi: 10.1089/cyber.2010.0578] [Medline: 21668370]

XSL•FO
RenderX

19. Maranesi E, Casoni E, Baldoni R, Barboni I, Rinaldi N, Tramontana B, et al. The effect of non-immersive virtual reality exergames versus traditional physiotherapy in parkinson's disease older patients: preliminary results from a randomized-controlled trial. Int J Environ Res Public Health 2022;19(22):14818 [FREE Full text] [doi: 10.3390/ijerph192214818] [Medline: 36429537]
20. Glavare M, Stålnacke BM, Häger CK, Löfgren M. Virtual reality exercises in an interdisciplinary rehabilitation programme for persons with chronic neck pain: a feasibility study. J Rehabil Med Clin Commun 2021;4:1000067 [FREE Full text] [doi: 10.2340/20030711-1000067] [Medline: 34527201]
21. Comeras-Chueca C, Villalba-Heredia L, Pérez-Llera M, Lozano-Berges G, Marín-Puyalto J, Vicente-Rodríguez G, et al. Assessment of active video games' energy expenditure in children with overweight and obesity and differences by gender. Int J Environ Res Public Health 2020;17(18):6714 [FREE Full text] [doi: 10.3390/ijerph17186714] [Medline: 32942663]
22. Chen Y, Li Z, Zou X, Yang Q, Guan B, Ming WK. Nursing students' perceptions of using virtual patients to prepare for clinical practice. J Nurs Educ 2022;61(7):398-402. [doi: 10.3928/01484834-20220613-04] [Medline: 35858137]
23. Comeras-Chueca C, Villalba-Heredia L, Perez-Lasierra JL, Marín-Puyalto J, Lozano-Berges G, Matute-Llorente A, et al. Active video games improve muscular fitness and motor skills in children with overweight or obesity. Int J Environ Res Public Health 2022;19(5):2642 [FREE Full text] [doi: 10.3390/ijerph19052642] [Medline: 35270330]
24. Graves LEF, Ridgers ND, Stratton G. The contribution of upper limb and total body movement to adolescents' energy expenditure whilst playing Nintendo Wii. Eur J Appl Physiol 2008;104(4):617-623. [doi: 10.1007/s00421-008-0813-8] [Medline: 18607619]
25. Graf DL, Pratt LV, Hester CN, Short KR. Playing active video games increases energy expenditure in children. Pediatrics 2009;124(2):534-540 [FREE Full text] [doi: 10.1542/peds.2008-2851] [Medline: 19596737]
26. Staiano AE, Calvert SL. Wii tennis play for low-income African American adolescents' energy expenditure. Cyberpsychology (Brno) 2011;5(1):4 [FREE Full text] [Medline: 24058381]
27. Staiano AE, Abraham AA, Calvert SL. Motivating effects of cooperative exergame play for overweight and obese adolescents. J Diabetes Sci Technol 2012;6(4):812-819 [FREE Full text] [doi: 10.1177/193229681200600412] [Medline: 22920807]
28. Marker AM, Staiano AE. Better together: outcomes of cooperation versus competition in social exergaming. Games Health J 2015;4(1):25-30 [FREE Full text] [doi: 10.1089/g4h.2014.0066] [Medline: 26181677]
29. Peng W, Crouse J. Playing in parallel: the effects of multiplayer modes in active video game on motivation and physical exertion. Cyberpsychol Behav Soc Netw 2013;16(6):423-427. [doi: 10.1089/cyber.2012.0384] [Medline: 23509986]
30. Carron AV, Bray SR, Eys MA. Team cohesion and team success in sport. J Sports Sci 2002;20(2):119-126. [doi: 10.1080/026404102317200828] [Medline: 11811568]
31. Lieberman DA, Chamberlin B, Medina E, Franklin BA, Sanner BM, Vafiadis DK, Power of Play: Innovations in Getting Active Summit Planning Committee. The power of play: innovations in getting active summit 2011: a science panel proceedings report from the American Heart Association. Circulation 2011;123(21):2507-2516. [doi: 10.1161/CIR.0b013e318219661d] [Medline: 21518980]
32. Riva G, Serino S. Virtual reality in the assessment, understanding and treatment of mental health disorders. J Clin Med 2020;9(11):3434 [FREE Full text] [doi: 10.3390/jcm9113434] [Medline: 33114623]
33. Farič N, Smith L, Hon A, Potts HWW, Newby K, Steptoe A, et al. A virtual reality exergame to engage adolescents in physical activity: mixed methods study describing the formative intervention development process. J Med Internet Res 2021;23(2):e18161 [FREE Full text] [doi: 10.2196/18161] [Medline: 33538697]
34. Kracht CL, Joseph ED, Staiano AE. Video games, obesity, and children. Curr Obes Rep 2020;9(1):1-14 [FREE Full text] [doi: 10.1007/s13679-020-00368-z] [Medline: 32077041]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by T de Azevedo Cardoso; submitted 18.03.24; peer-reviewed by DAA Abshire, KM Leung; comments to author 06.06.24; revised version received 29.07.24; accepted 06.11.24; published 30.12.24.

Please cite as:
Chen Y, Guan L, Wu W, Ye L, He Y, Zheng X, Li S, Guan B, Ming WK
Factors That Influence Young Adults' Preferences for Virtual Reality Exergames in a Weight Control Setting: Qualitative Study
J Med Internet Res 2024;26:e58422
URL: https://www.jmir.org/2024/1/e58422
doi:10.2196/58422
PMID:

XSL•FO
RenderX

©Yanya Chen, Lina Guan, Weixin Wu, Liang Ye, Yan He, Xiaofen Zheng, Sicun Li, Bingsheng Guan, Wai-kit Ming. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identifying the Risk Factors of Allergic Rhinitis Based on Zhihu Comment Data Using a Topic-Enhanced Word-Embedding Model: Mixed Method Study and Cluster Analysis

Dongxiao Gu[1], PhD; Qin Wang[1], MD; Yidong Chai[1], PhD; Xuejie Yang[1], PhD; Wang Zhao[1], PhD; Min Li[1], MD; Oleg Zolotarev[2], PhD; Zhengfei Xu[1], MD; Gongrang Zhang[1], PhD

[1]School of Management, Hefei University of Technology, Hefei, China

[2]Russian New University, Moscow, Russian Federation

Corresponding Author:
Dongxiao Gu, PhD
School of Management, Hefei University of Technology
193 Tunxi Road
Hefei, 230009
China
Phone: 86 13866167367
Email: gudongxiao@hfut.edu.cn

Abstract

Background: Allergic rhinitis (AR) is a chronic disease, and several risk factors predispose individuals to the condition in their daily lives, including exposure to allergens and inhalation irritants. Analyzing the potential risk factors that can trigger AR can provide reference material for individuals to use to reduce its occurrence in their daily lives. Nowadays, social media is a part of daily life, with an increasing number of people using at least 1 platform regularly. Social media enables users to share experiences among large groups of people who share the same interests and experience the same afflictions. Notably, these channels promote the ability to share health information.

Objective: This study aims to construct an intelligent method (TopicS-ClusterREV) for identifying the risk factors of AR based on these social media comments. The main questions were as follows: How many comments contained AR risk factor information? How many categories can these risk factors be summarized into? How do these risk factors trigger AR?

Methods: This study crawled all the data from May 2012 to May 2022 under the topic of *allergic rhinitis* on Zhihu, obtaining a total of 9628 posts and 33,747 comments. We improved the Skip-gram model to train topic-enhanced word vector representations (TopicS) and then vectorized annotated text items for training the risk factor classifier. Furthermore, cluster analysis enabled a closer look into the opinions expressed in the category, namely gaining insight into how risk factors trigger AR.

Results: Our classifier identified more comments containing risk factors than the other classification models, with an accuracy rate of 96.1% and a recall rate of 96.3%. In general, we clustered texts containing risk factors into 28 categories, with season, region, and mites being the most common risk factors. We gained insight into the risk factors expressed in each category; for example, seasonal changes and increased temperature differences between day and night can disrupt the body's immune system and lead to the development of allergies.

Conclusions: Our approach can handle the amount of data and extract risk factors effectively. Moreover, the summary of risk factors can serve as a reference for individuals to reduce AR in their daily lives. The experimental data also provide a potential pathway that triggers AR. This finding can guide the development of management plans and interventions for AR.

(J Med Internet Res 2024;26:e48324) doi:10.2196/48324

KEYWORDS

social media platforms; disease risk factor identification; chronic disease management; topic-enhanced word embedding; text mining

XSL•FO
RenderX

Introduction

Background

Over the past few decades, the prevalence of chronic diseases has increased significantly, becoming a global public health concern. The World Health Organization has listed allergic diseases as one of the disease types that require priority research and prevention in the 21st century [1]. As a common chronic disease, allergic rhinitis (AR) is a multifactorial disease that is induced by environmental conditions or certain genes [2]. AR not only has a significant impact on individuals' sleep, social life, and work attendance but also triggers comorbidities such as conjunctivitis, atopic dermatitis, and asthma [3]. Large-scale flow survey data showed that AR currently affects several people in China alone [4] and with an estimated prevalence between 15% and 20% worldwide [5]. The direct and indirect costs associated with the management of AR are also a significant burden on society. For instance, the total cost of AR in Sweden, with a population of 9.5 million, was estimated at €1.3 (US $1.41) billion annually [6]. These unexpectedly high costs could be related to the high prevalence of disease, in combination with the previously often underestimated indirect costs that arise from reduced work efficiency and absenteeism and the potential costs associated with treating AR comorbidities [6].

Currently, there is no cure for AR, and individuals need to avoid the disease risk factors such as exposure to allergens and inhalation irritants [7] during the long self-management process. Therefore, identifying AR risk factors can provide a reference for patients to help reduce the condition in their daily lives [8].

A plethora of studies have been proposed to identify AR risk factors. These studies recruited participants with symptoms of AR and control participants without AR symptoms from a specific age group or a particular geographical area. These studies collected demographic information, lifestyle habits, family history, comorbidities, and residential areas through questionnaires. Subsequently, they used correlation methods to explore the relationship between these data and AR, aiming to identify the risk factors for AR within the specified age group or geographical area [9]. However, these studies have 2 limitations. First, these studies specifically target certain age groups or geographical areas, and questionnaires can only gather data on specific pieces of information. Owing to the constraints of questionnaire surveys, it is challenging to identify potential risk factors that may be present in individuals' daily lives. As a result, the risk factors identified through survey-based studies have a limited scope and are incomplete. As such, they provide limited insights for a broader patient population. Second, the survey-based approach demands a commitment to long-term investigation and a substantial effort to collect representative responses [10]. In contrast, collecting information from social media platforms can cover large geographical areas at a comparatively low cost [10]. Social media platforms allow users to share experiences and opinions on various topics [11,12], including personal health issues [13]. Over time, highly unstructured and implicit knowledge has been generated in communities where users frequently participate [14,15], which can provide daily health records that are difficult to obtain from traditional questionnaire surveys. Therefore, social media can become a potential source of information for identifying risk factors for diseases such as AR [16].

Text-mining techniques are an effective tool for using voluminous social media data [17]. Some studies have combined social media data analysis to obtain knowledge about disease risk factors [18,19]. However, the abovementioned studies on disease risk factors used only shallow text features such as the number of social media text items and word cooccurrences, which are not conducive to identifying disease risk factors in the context of colloquial and diverse user expressions [20]. In this study, we designed a text-processing framework to automatically identify risk factors from social media data [21]. We used social media comments to construct a natural language processing–based AR risk factor identification method, aiming to tackle the problems of omission and low accuracy in traditional disease-related information identification methods that rely solely on shallow text features such as word frequency.

To be more specific, we developed an AR risk factor identification method that integrates pretrained word embeddings with text convolutional neural networks (CNNs). The Word2vec algorithm has proven to be superior in text vector representation [20]. This is a prediction-based approach that predicts the neighboring words that are most likely to appear within a window size around a center word in a corpus, resulting in high-dimensional vector representations that capture semantic aggregation. As social media users may mention related topics, such as symptoms and treatments, when describing risk factors in their comments, we used a local context window to achieve better semantic aggregation of AR risk factors, a method that has been demonstrated to be effective for such aggregation. In addition, using the Skip-gram model to train word pairs enables the incorporation of word thematic information, thus improving attention to risk factor phrases. The convolutional network can convolve the text in the word vector dimension and extract critical information through the max-pooling layer operation. In addition, this study used a clustering method with review mechanisms to concentrate on a large amount of text that contains risk factors within the observable range, thereby ensuring the usefulness of the content obtained through text mining.

Our main contributions were as follows:

1. First, this study proposed a framework (TopicS-ClusterREV) based on natural language processing for identifying the risk factors of AR. We used pretrained word embeddings and text convolutional networks to process social media text. Our model can identify more risk factors from social media comments with high accuracy and recall. To the best of our knowledge, this is the first study to use natural language processing techniques to identify risk factors for AR in social media comments.
2. Second, this study proposes a topic-enhanced word-embedding model. TopicS enhances the thematic information of words by adding a task that predicts the theme to which the center word belongs. This generates high-dimensional word vector representations with semantic aggregation and theme enhancement. We trained 2 types

XSL•FO
RenderX

of word vectors using both the Skip-gram and TopicS models and separately input them into each risk factor classifier. The results showed that TopicS outperformed the baseline on the text classification task, demonstrating the effectiveness of our topic-enhanced word-embedding model.

3. Finally, we introduced automatic and manual review mechanisms to improve the single-pass algorithm, which allowed us to effectively identify and focus on a large amount of text that contains risk factors within the observable range. We ultimately identified 28 categories of risk factors including the common risk factors that lead to most individuals developing symptoms and previously overlooked risk factors that were not within the scope of previous research.

Identification of AR Risk Factors Through Surveys

AR has become a major global issue with a substantial increase in its prevalence in recent years. In Europe, the prevalence of AR among Danish adults progressively increased from 19% to 32% over the past 3 decades [22]. Understanding the risk factors, such as genetic, environmental, and lifestyle factors, helps in the management of AR, thus motivating many studies to focus on identifying potential risk factors. These studies are summarized in Table 1. From Table 1, we observed that the previous studies were based on survey methods, including cross-sectional surveys, cohort studies, and case-control studies.

Table 1. Summary of the literature related to risk factors for allergic rhinitis (AR)[a].

Study, year	Method	Risk factors
Chiang et al [23], 2016	Case control	Exposure to sulfur dioxide
Kurganskiy et al [24], 2021	Cross-sectional	Grass and tree pollen
Lee et al [25], 2021	Cross-sectional	The widespread use of industrial chemicals
Paciência et al [26], 2020	Cross-sectional	Indoor decoration materials containing volatile organic chemicals
Saulyte et al [27], 2014	Case control	Active smoking
Kong et al [28], 2021	Cohort	Stress
Han et al [29], 2016	Cross-sectional	Obesity
Kanazawa et al [30], 2018	Cross-sectional	TYRO3 gene
Alm et al [31], 2014	Cross-sectional	Using antibiotics in the first week after birth

[a]We searched for the literature related to AR risk factors and presented 9 papers from the past decade to showcase the methods and the identified risk factors.

These studies typically recruited participants with symptoms of AR and control participants without AR symptoms from a specific age group or a particular geographical area, collected demographic information through questionnaires, and then conducted correlation analysis, such as logistic regression, to explore the relationship between those metadata and AR [32]. For instance, Gao et al [9] conducted a cross-sectional survey to investigate the prevalence and risk factors of adult self-reported AR in the plain lands and hilly areas of Shenmu City in China and analyzed the differences between regions. The content of the web-based questionnaire included demographic factors, smoking status, the comorbidities of other allergic disorders, family history of allergies, and place of residence. The unconditional logistic regression analysis was used to screen for factors influencing AR. Finally, they found that the prevalence of AR existed in regional differences. Genetic and environmental factors were the important risk factors associated with AR. However, these studies have 2 limitations. First, these studies specifically targeted certain age groups or geographical areas, and questionnaires can only gather data on specific pieces of information. Owing to the constraints of questionnaire surveys, it is challenging to identify potential risk factors that may be present in individuals' daily lives. As a result, the risk factors identified through survey-based studies have limited scope and are incomplete and they may provide limited insights for a broader patient population. Second, the survey-based approach demands a commitment to long-term investigation and a massive effort to collect representative responses [10].

Identification of Disease Risk Factors From Social Media Through Text Mining

Social media sites provide a convenient way for users to continuously update their day-to-day activities, which allows large groups of people to create and share information, opinions, and experiences about health conditions through web-based discussion [11]. Hence, social media can be considered a new data source to assess population health. As shown in Table 2, some studies have combined text-mining techniques to classify and summarize voluminous social media data to obtain knowledge about chronic disease risk factors. Zhang and Ram [33] extracted behavioral features from Twitter posts of asthma users using keywords from an existing knowledge base. Griffis et al [34] collected 25,000 tweets containing and not containing diabetes, identified 5000 common words, used logistic regression to determine which common words were high-frequency expressions of diabetes, and finally grouped these high-frequency words using latent Dirichlet allocation to obtain the risk factors for diabetes. Schäfer et al [35] used syntactic analysis to identify portions of risk factors occurring before or after causal terms, grouped these portions using latent Dirichlet allocation, and obtained the risk factors for gastric discomfort. Pradeepa et al [19] performed clustering on stroke-related tweets using the Probability Neural Network,

XSL•FO
RenderX

used the Apriori algorithm to identify frequent word sets related to risk, and thus identified risk factors for stroke [19]. In addition to the aforementioned approaches that use shallow text features such as keywords, frequent word sets, high-frequency words, and syntactic features for disease risk factor identification, other studies [36-38] trained risk factor classifiers using machine learning methods such as Naive Bayes, Maximum Entropy Model, and Naive Bayes Classifier–Term Frequency Inverse Document Frequency. These classifiers predict the presence of risk factors in text based on discrete vector representations such as bag-of-words and n-gram.

Table 2. Summary of the literature related to disease risk factors based on social media data[a].

Study, year	Social media platforms	Data	Methods	Features	Diseases	Identified risk factors
Zhang and Ram [33], 2020	Twitter	Posts	Semisupervised learning	Knowledge base	Asthma	Behavioral attributes
Griffis et al [34], 2020	Twitter	Posts	LDA[b]	Word frequency	Diabetes	BMI, waist, drugs, alcohol, and obesity
Schäfer et al [35], 2020	Doctissimo, Aufeminin	Answers	LDA	Syntactic	Gastrointestinal discomfort	Food and psychological factors
Pradeepa et al [19], 2020	Twitter	Posts	A priori	Frequent word set	Stroke	Lifestyle, family history, heart disease
Alswedani et al [39]	Twitter	Posts	Keywords list	Word frequency	Psychological health	Social and economic factors, individual factors, diseases and disorders
Chung et al [40]	Telegram (app)	Time	MLP[c]	Meta data	Respiratory diseases	Pollution
Neisani Samani et al [41]	Twitter	Posts	Content analysis methods	Word frequency	Oropharyngeal cancer	Drinking, smoking

[a]We searched for studies related to identifying disease risk factors based on social media data. We found 7 papers from the past decade, highlighting the social media platforms, data, methods, features, diseases, and risk factors involved in research.

[b]LDA: latent Dirichlet allocation.

[c]MLP: multilayer perceptron.

The current methods for identifying disease risk factors on social media fall into 2 categories: shallow text feature methods and discrete word vector representations. Shallow text feature techniques often fail to capture important risk factors resulting in low accuracy, whereas discrete word vector approaches struggle to keep up with the dynamic vocabulary of social media text, missing new words, and trending expressions, thus inadequately representing the information conveyed.

Word Embedding and Text Classification Based on Deep Learning

Natural language processing technology promotes text analysis based on social media comments [39]; this technology can learn the deeper semantic features of the comment text and the features that are consistent with the current context, according to different training corpus, to input a better text vector representation for downstream classification tasks. Some researchers have used large-scale pretrained language models [40], global matrix decomposition [41], and local context windows [42] for text vector representation. Local context windows are more suitable for semantically aggregating AR risk factors [43]. Skip-gram and Continuous Bag-of-Words Model (CBOW) are prediction-based methods that learn the semantic representation of a center word by predicting the most likely neighboring words within a window size in a corpus. When users narrate risk factors in their comments, they may also mention symptoms, treatments, and other topics. These global contexts may dilute the key features of the risk factors expression. CBOW averages the context words to predict the target word and tends to predict high-frequency words in the corpus. In contrast, Skip-gram gives each word a chance to be a center word, making it better at predicting rare words compared with CBOW [44]. Therefore, in situations where social media users express a wide variety of ideas, the Skip-gram model can yield satisfactory outcomes. Moreover, the Skip-gram approach uses word pair training, which facilitates the incorporation of topic information into words [45], resulting in the generation of high-dimensional word vectors that feature semantic aggregation and topic enhancement. Therefore, we selected Skip-gram as the word-embedding model for our study.

Text classification has evolved to deep learning models, mainly including CNN-based models [46], recurrent neural network (RNN)–based models [47], and transformer models [48]. For the CNN algorithm, convolutional networks can convolve text on the word vector dimensions and extract key information through pooling layer operations. Consequently, this algorithm is capable of using essential data for classification tasks. Therefore, we used TextCNN for classifier training and evaluated the performance of RNN and transformer models on this task.

XSL•FO
RenderX

Methods

Framework

The framework used in this study consisted of 3 parts as shown in Figure 1. The first part was data collection and processing, aimed at obtaining a clean data set. The second part was risk factor identification, which included the proposed TopicS method and training of a risk factor classifier. The implementation steps were as follows: (1) semiautomatically constructing a risk factor topic dictionary, (2) generating high-dimensional word vectors enhanced by TopicS-generated topics, and (3) vectorizing annotated text and training a risk factor classifier. The third part is text clustering and keyword extraction, which uses the ClusterREV method to cluster the identified risk factors and extract keywords from every category.

Figure 1. Allergic rhinitis (AR) risk factor identification method based on the topic-enhanced word-embedding model (TopicS-ClusterREV). The figure shows the research framework of our study. The framework consists of 3 parts. The first part is data collection and processing aimed at obtaining a clean data set. The second part is risk factor identification, which includes the proposed TopicS method and training of a risk factor classifier. The third part is text clustering and keyword extraction, which uses the ClusterREV method to cluster identified risk factors and extract keywords from every category.

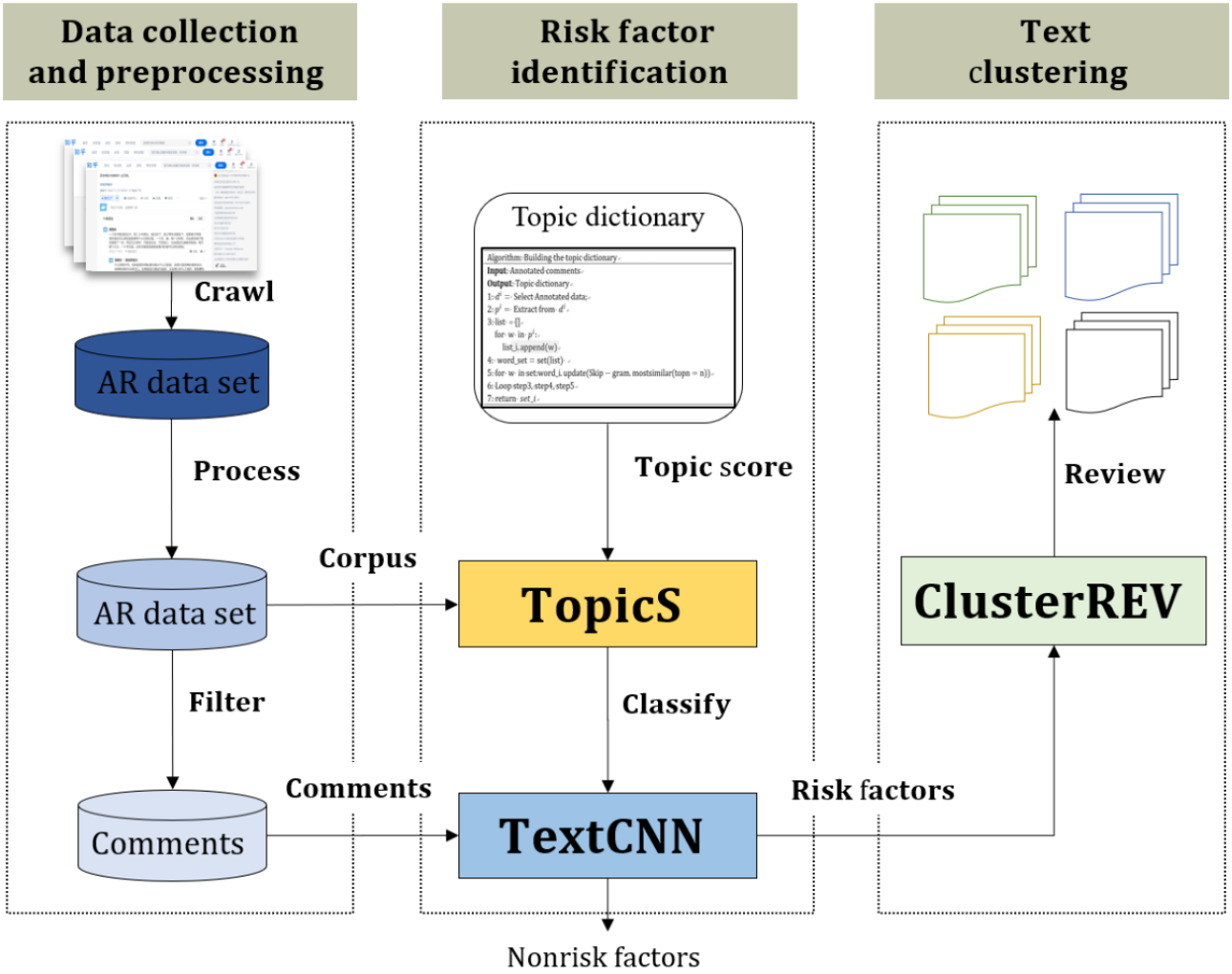

Data Set

Zhihu is a Chinese social media platform where people discuss topics in an web-based forum format. In May 2022, the Zhihu subcommunity *allergic rhinitis* had 1.04 million discussions. The posts on this social media platform allow other users to comment [49], and people can explain their situations to provide support or seek help effectively. Therefore, these comments provide a rich source of data for investigating the risk factors reported by different users [50]. In this study, we trained domain-specific word representations based on experimental data. A relatively domain-specific input corpus [51] is better at extracting meaningful semantic relations than a generic pretrained language model [52]. We crawled all the data from May 2012 to May 2022 under the topic *allergic rhinitis* on Zhihu, obtaining a total of 9628 posts and 33,747 comments, including the post ID, comment ID, and post and comment content.

In this study, we preprocessed the data through regularization, stop word removal, and word separation. First, we removed special symbols, such as URLs and emoticons, in the comments through regularization and stop word removal to reduce the interference of noise with the text analysis task. Then, we compiled a dictionary of 169 specialized terms, including types of AR, medications, and comorbidities, to reduce the probability

XSL•FO
RenderX

of incorrect word segmentation. After word separation, we obtained a lexicon of 68,863 words and ranked the words according to the number of occurrences. We found that the top 10,000 words accounted for 94.83% of the total words, suggesting that many words recurred and a relatively simple word vector could effectively train the model [53]. This further confirms the efficacy of our decision to use Skip-gram as the foundational model.

We observed ultrashort comment noise in the comments (eg, “Thank you!”). It is important to note that these ultrashort comments do not include any personal medical information. The ultrashort comments were filtered, resulting in 33,039 valid comments. This operation can effectively minimize the impact of noise on downstream text classification tasks. Table S1 in Multimedia Appendix 1 presents the examples of valid comments.

Annotation

The data must be labeled before supervised learning and then trained end to end. If a comment directly mentions an allergen or indicates a condition that leads to the appearance or worsening of symptoms, the comment will be labeled as 1, indicating the presence of risk factors, as shown in Figure 2.

Figure 2. Examples of short text annotation. The figure shows examples of data labeled as 1 including the text and label. In the figure, phrases with a blue background indicate those that were specifically noted during manual annotation, and the presence of these marker phrases often suggests potential risk factors in the sentence. The yellow background highlights the risk factors in the text.

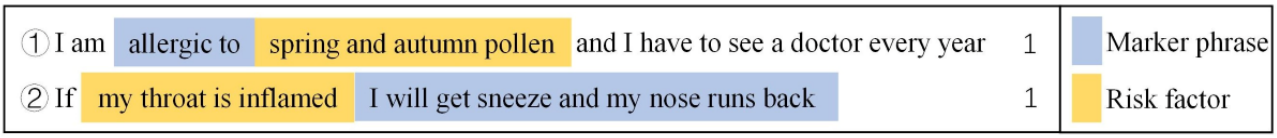

We randomly chose 2030 comments from the 33,039 comments, and 3 researchers labeled each comment as containing or not containing risk factors. To ensure high interannotator consistency, all 3 researchers annotated all 2030 comments. In cases with uncertainty in labeling, the 3 researchers discussed and arrived at a final label. After annotating and eliminating comments with religiously controversial content, 2000 labeled comments remained, consisting of 996 comments containing risk factors and 1004 comments not containing risk factors. The data set was divided into a 90% training set and a 10% test set. The 90% training set was further divided into 10 subsets, with 9 subsets used for training and the remaining subset used for validation, performing 10-fold cross-validation.

Topic Dictionary Construction

We used a combination of manual labeling and similarity calculation to identify keywords related to risk factors. Subsequently, we constructed a table of topic words using a semiautomated approach. The process of constructing the dictionary is depicted in Textbox 1 and is as follows: (1) label 400 randomly selected comments as described in the *Annotation* section, thereby obtaining 198 comments with risk factors; (2) extract risk factor phrases from annotated comments; (3) obtain risk factors topic word list; (4) remove duplicate word list, and the words in the current topic are used as seed words, *word_set*; (5) use Skip-gram to find the top similar words to expand the topic words; (6) repeat steps 3 through 5 to expand the topic word; and (7) finally, obtain the topic words for the risk factor. A large weight was assigned to the risk factor theme words. Table S2 in Multimedia Appendix 1 shows examples of the risk factor topic dictionary.

Textbox 1. The algorithm for building the topic dictionary. This textbox outlines the algorithm process for building the topic dictionary with explanations for each step provided in the text.

```
Input: annotated comments
Output: topic dictionary
1. d^i= Select Annotated data;
2. p^i= Extract from d^i
3. list =
for w in p^i:
list_i.append(w)
4. word_set=set(list)
5. for w in set: word_i.update(Skip-gram.mostsimilar(topn=n))
6. Loop step3, step4, step5
```


XSL•FO
RenderX

Ethical Considerations

As the use of text data from social media involves user privacy, this study adopted the following steps for deidentification: (1) We removed user account information and retained only anonymous comment information. (2) We used regular expressions to match and delete URLs and email addresses in the comments. (3) During the annotation process, annotators received only text that did not involve personal information. To evaluate the quality of deidentification, we randomly selected 500 text items for manual inspection and did not find any instances containing personal identity information. Our data are sourced from public discussions on Zhihu, a social media platform that can be accessed without registration. We followed strict ethical research protocols similar to the guidelines by Eysenbach and Till [54]. In addition, to protect the anonymity of participants, we have implemented measures including the removal of user information and avoiding verbatim quotations to prevent identification through search engines, protecting the privacy and security of personal data. It should be mentioned that our study was focused on the post level; we do not anticipate any negative ethical impact from our analysis.

Topic-Enhanced Word Embedding

TopicS performed 2 tasks during training, as shown in Figure 3. The first task was to predict the neighboring words within the window of the central word. The second task was to predict the topic of the central word; the topic dictionary used for this purpose is described in the *Topic Dictionary Construction* section.

Figure 3. Topic-enhanced word-embedding model (TopicS). The figure illustrates the vector changes within the TopicS model. The rectangular boxes in both the input and output represent one-hot vectors. Within the input, the dark blue circles signify the center words, representing a value of 1, whereas the light blue circles denote other words in the training text, with a value of 0. For the output's task 1, the dark blue circles depict context words surrounding the center word, signifying a value of 1, whereas the light blue circles represent noncontext words with a value of 0. The various colored circles in the output's task 2 indicate the topics to which the center word belongs. If it pertains to the risk factor topic, it is marked by a dark blue circle, symbolizing a value of 1, whereas circles of other colors represent a value of 0.

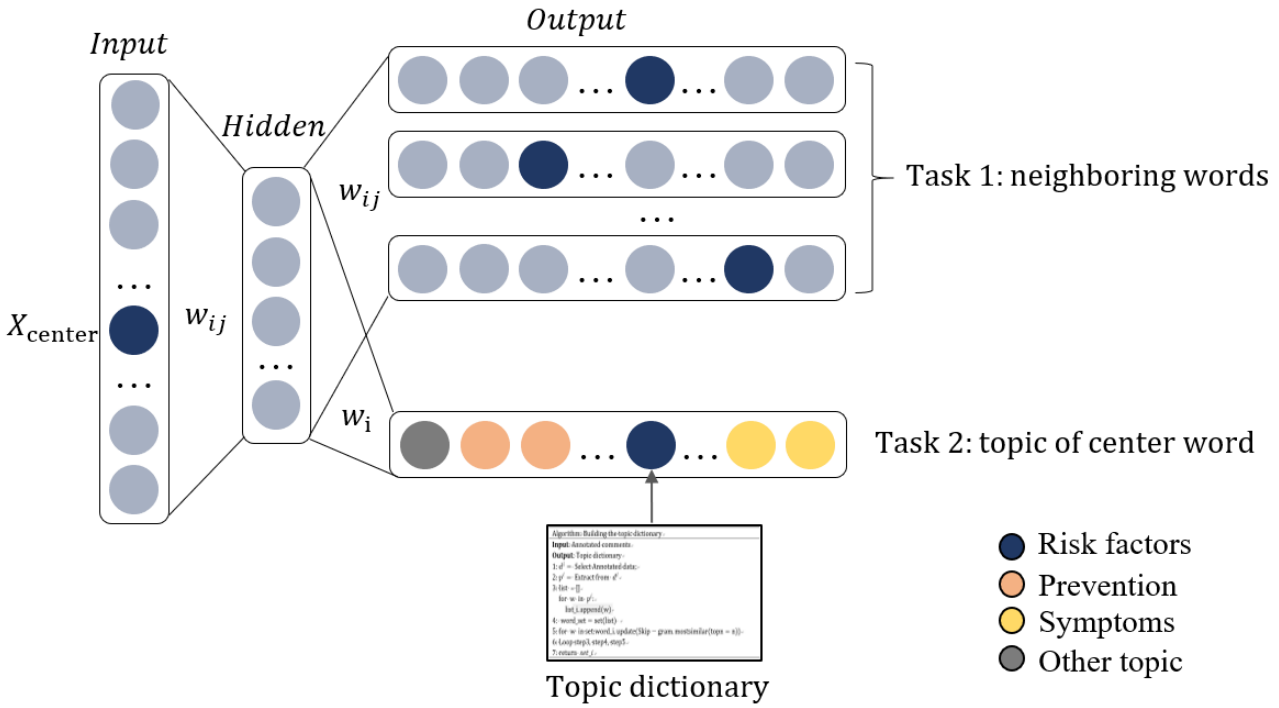

The specific formula calculations for the loss function design, parameter updates, and error backpropagation of TopicS are explained subsequently.

First, we defined the loss function. For each word in the corpus, we used it as the central word for a sliding operation with a window size of c; let S be the training sequence $(w_1,w_2,...,w_T)$, whereas w_i denotes the ith word in the sequence. The subscript T represents the total number of unique words in the corpus. In addition to predicting the contextual word of the central word, we must also predict the topic score of the central word. Therefore, the loss function comprised 2 parts: L_{cont} and L_{topic}, and the overall loss was denoted by L_s. Our training objective was to minimize the loss function:

The initial word vector was represented as one-hot vector. The central word was denoted as w_{center}, the surrounding words of the central word were denoted as $w_{(O,cont(i))}$, and the central word topic information was denoted as $w_{(O,topic)}$. The weight parameter λ was used to balance the loss of L_{cont} and L_{topic}. After sliding the window to browse the entire corpus to average all loss window losses, error backpropagation was performed to update the parameters. The input matrix was represented as the central word vector. The actual loss function can be expressed as

XSL•FO
RenderX

Second, we introduced the rules for updating parameters. The parameters were updated through error backpropagation

Here, y_{true} represents the true value of the contextual words in the window, and y_{pred} represents the predicted value.

Finally, we can update the word representation.

Text Classification

In this study, we chose TextCNN as the classification model. In the risk factor identification task, some key semantic information is more important, and TextCNN can efficiently use the key information for classification with minimal cost consumption. We represented the manually annotated text as a vector matrix using high-dimensional word vector representations trained by the TopicS model, which aggregates local contextual and topic information and uses it as input for the TextCNN model. Then, the TextCNN algorithm leverages convolutional kernels of different sizes to extract multiple n-gram text features and uses convolutional operations in a fixed window to combine word representations to capture local information. Our input word vector combined the topic information of words, and the most important features in the convolution operation can be extracted using the maximum pooling operation as shown in Figure 4.

Figure 4. Framework of the classification model with different word embedding. This figure illustrates the TextCNN modeling process for text vectorization using both the skip-gram and TopicS techniques. In the example sentence, “spring” and “pollen” are highlighted as risk factors. These words are represented by blue squares in TopicS, suggesting that TopicS incorporates topic information, unlike the skip-gram method. These thematic data are subsequently integrated into the convolution, max-pooling, and softmax procedures to enhance the model’s classification capabilities.

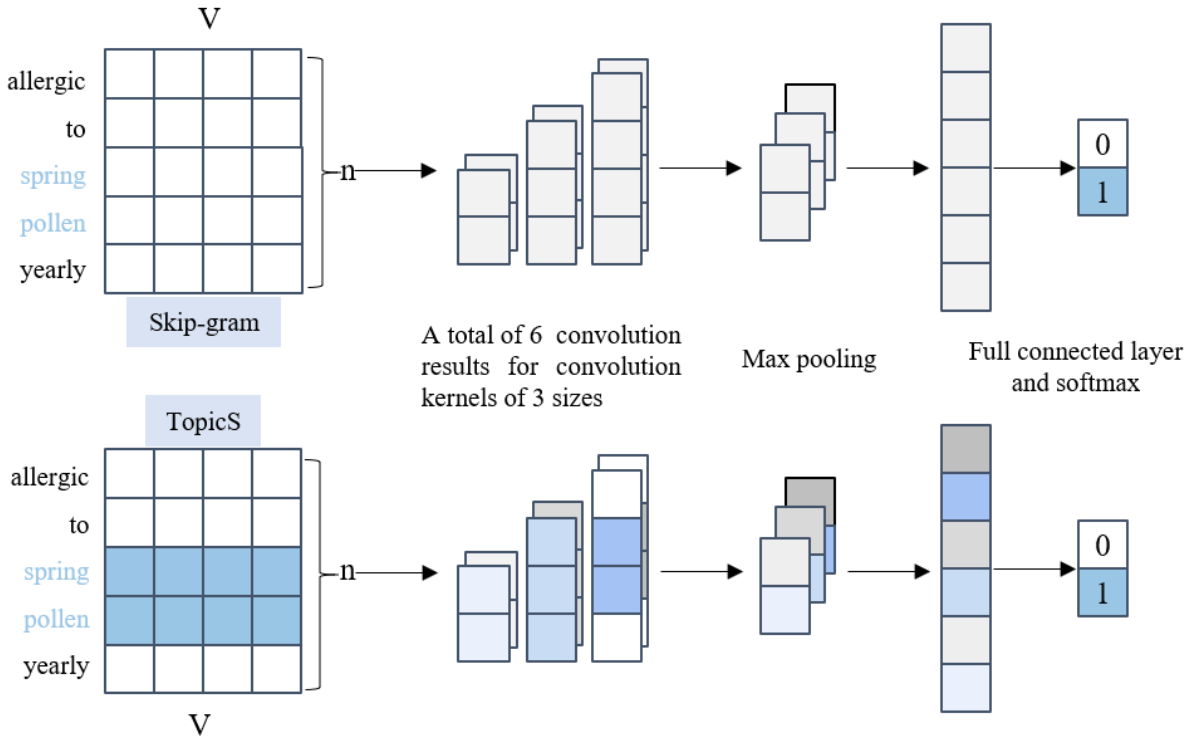

Clustering With a Review Mechanism

The clustering task is to group similar risk factors. In this study, a large amount of text containing risk factors was clustered into a manually observable number of categories, making it easier to comprehend their content. This study enhances the single-pass algorithm and integrates it with a manual review to cluster the risk factors identified in the text classification, ensuring the validity of the clustering results. The main concept of single-pass clustering [55] is to match informational text items based on their similarity values without the need to determine the number of clusters in advance. This makes it suitable for clustering tasks with an unknown number of clusters. However, traditional single-pass clustering uses only one-loop traversal, which may result in previously entered text items completing the traversal earlier. This can cause their similarity to the previous topics to be slightly lower than the threshold and lead to them being recreated as new categories, ultimately affecting the clustering effect.

As shown in Figure 5, we improved the single-pass algorithm by retraversing the categories that were clustered separately after all the text items had been traversed to handle any missed text. After the automated clustering was completed, we conducted a manual review to ensure the reliability of the clustering.

XSL•FO
RenderX

Figure 5. Cluster method with review mechanisms (ClusterREV). This figure depicts the process of the ClusterREV algorithm. The rectangular boxes represent category state transitions. The circles below the rectangles indicate the texts awaiting clustering. The algorithm assesses the distance between the current text and existing categories, classifying the text based on the minimal distance and a set threshold. Once all texts have been clustered, texts within a solitary category undergo automatic review. Finally, we manually reviewed the clustering results.

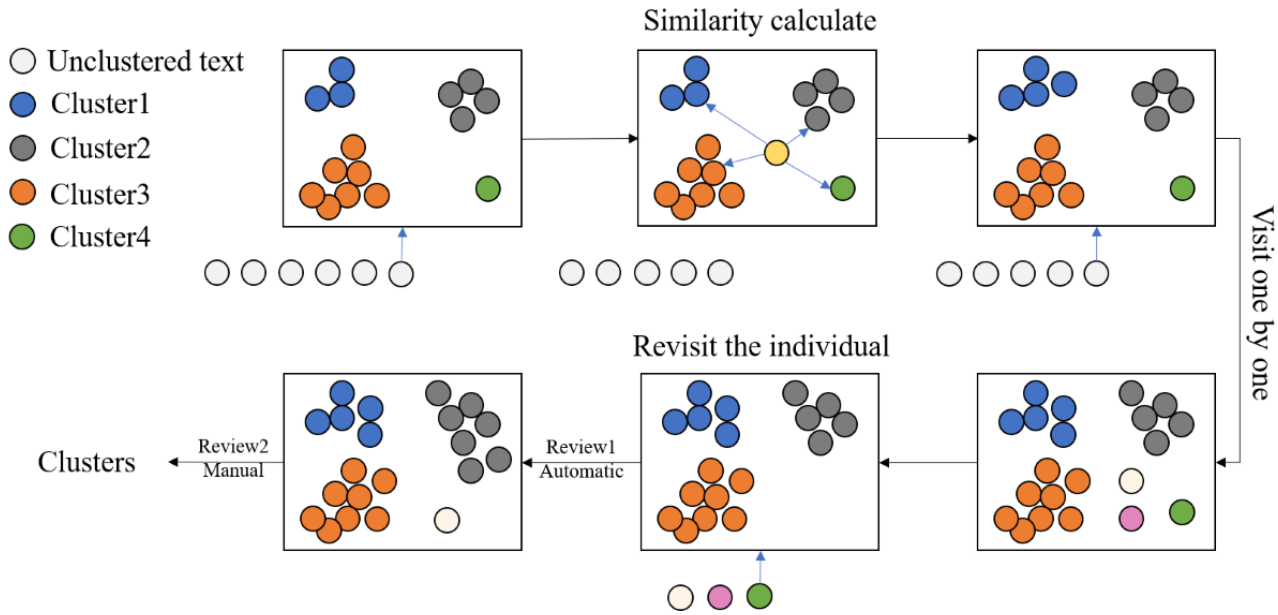

Moreover, this study uses a keyword cloud visualization of category content to quickly understand the themes and characteristics of each cluster and compare the differences between different clusters. TextRank [56] was selected to extract category keywords, which considers only the voting scores of words in a single document; common words that frequently appear in a single document easily obtain high scores [57]. We treated each category as a single document for keyword extraction. As risk factors appear more frequently in categories, TextRank can effectively extract risk factors and surrounding words, preserving category content information as much as possible and reflecting the true content of the risk factors.

Results

Overview

In this section, we present the performance of the classifier and the findings based on the categorization of all the comments in the clean data set using the classifier. Our approach involved visualizing the clustering results of the risk factors to comprehend the primary elements of these factors. We also explored the pathogenic mechanisms associated with these risk factors.

Classifier Performance

We used standard text-mining evaluation metrics such as accuracy, precision, recall, and F_1-score to evaluate the performance. Precision assesses how many risk factors the model identifies correctly, and recall measures how many risk factors the model can identify on its test set. As we aimed to identify as many AR risk factors as possible to provide comprehensive references for individuals, recall was more important than precision in our study.

We set 7-word embedding dimensions ranging from 100 to 400. Table 3 displays the classification results of the TextCNN classification model with the 7 dimensions of Skip-gram and TopicS word vectors. In addition, TextRNN and transformer models were evaluated with the 7-word embedding dimensions of TopicS or Skip-gram, as shown in Tables S3 and S4 in Multimedia Appendix 1; the classification models performed better when the word-embedding dimension was 100 or 150, as shown in Table 4, which includes the results with best-performing dimensions. This study conducted word representation learning on a domain-specific input corpus, where low dimensionality was found to be sufficient to represent the features of the corpus [58]. Moreover, TopicS not only improved precision but also significantly increased recall for all 3 models, as shown in Table 4.

XSL•FO
RenderX

Table 3. Word-embedding dimension parameters with TextCNN.

Evaluation metrics and model	100	150	200	250	300	350	400
Accuracy (%)							
Skip-gram	93.75	94.85	94.00	93.15	93.80	93.49	93.40
TopicS[a]	95.10[b]	96.10	94.80	94.45	94.95	95.25	95.40
Precision (%)							
Skip-gram	91.92	95.32	94.07	93.16	93.33	93.28	93.09
TopicS	96.32	95.95	95.56	94.11	95.20	95.87	96.53
Recall (%)							
Skip-gram	94.00	94.50	94.00	93.30	94.40	93.90	93.90
TopicS	95.90	96.30	94.00	94.90	94.70	94.60	94.20
F_1-score (%)							
Skip-gram	93.90	94.88	93.95	93.17	93.84	93.48	93.44
TopicS	95.09	96.10	94.77	94.48	94.94	95.22	95.34

[a]TopicS represents the topic-enhanced word-embedding model proposed in this paper.

[b]Italicization represents that the metrics of TopicS are better than Skip-gram for each metric.

Table 4. Accuracy, precision, recall, and F_1-score of Skip-gram and TopicS with different classification models.

Model (Embed_size[a])	Accuracy (%)	Precision (%)	Recall (%)	F_1-score (%)	Time (s)
TextCNN					
Skip-gram (150 dimens)	94.85	95.32	94.50	94.88	40.30
TopicS (150 dimens)	96.10[b]	95.94	96.30	96.10	35.23
TextRNN					
Skip-gram (150 dimens)	94.85	95.32	94.50	94.88	40.30
TopicS (150 dimens)	96.10[b]	95.94	96.30	96.10	35.23
Transformer					
Skip-gram (100 dimens)	85.45	85.06	78.80	81.13	55.16
TopicS (150 dimens)	90.70	90.90	90.60	90.68	49.32

[a]Embed_size represents the word-embedding size.

[b]Italicization represents that the metrics of TopicS are better than Skip-gram for each model.

Table 4 shows that TextCNN has the highest accuracy and recall rate among the 3 classification models. The highest accuracy achieved by our classification model was 0.9594, which used a 150-dimension word-embedding representation obtained from TopicS. In other words, TextCNN can detect more risk factors and minimize the loss of risk factors resulting from classification errors. The CNN model can extract key information similar to n-grams in sentences. The combination of TopicS and TextCNN can enhance topic information and achieve an aggregation effect. Our implementation process was the simplest and consumed the least resources. Our model examined 30,372 comments and identified 5221 comments containing risk factors.

Risk Factor Clustering Results

We clustered the text items obtained from the text classification into 28 categories and extracted keywords from each category to better understand the content. Table 5 shows the top 5 categories and their corresponding keywords. The complete list can be found in Table S5 in Multimedia Appendix 1. We used category 1 as an example to explain the category formation process and demonstrate the validity of the qualitative results. As shown in Table 4, we labeled category 1 as *Season* based on the analysis of keyword weights and relative comments. The comments related to this category focused on seasonally induced AR, with factors such as changes in the weather during seasonal transitions and colder temperatures during winter, which can exacerbate symptoms. We also counted the number of text items in each category and found that seasonal, regional, mites, and weather changes were common risk factors for most patients. In addition, patients' unhealthy lifestyle habits were also important risk factors widely present in research investigations. Furthermore, most patients reported experiencing symptoms at specific times (eg, "morning"), but researchers have paid little

XSL•FO
RenderX

attention to the timing of symptom occurrence (which we refer to as time points).

Table 5. Category keyword distribution and visualization.

Category	Top 10 words (weight)	Word cloud	Number of text items
Season	Summer (0.031), winter (0.031), season (0.02), season change (0.014), spring (0.013), autumn (0.013), seasonal (0.013), nose (0.011), air conditioning (0.01), month (0.007)	Multimedia Appendix 2	852
Region	Beijing (0.035), Shenzhen (0.019), air (0.010), Wuhan (0.010), Guangdong (0.01), city (0.009), Shanghai (0.009), dust mites (0.009), nose (0.009), university (0.009)	Multimedia Appendix 3	644
Mites	Dust mites (0.111), mites (0.045), dust (0.02), allergy (0.012), pollen (0.008), allergens (0.007), effect (0.006), child (0.005), nose (0.005), cold air (0.005)	Multimedia Appendix 4	608
Weather	Cold air (0.071), weather (0.023), temperature (0.021), nose (0.02), winter (0.009), changes (0.008), air (0.008), alternate (0.008), summer (0.007), air conditioning (0.007)	Multimedia Appendix 5	538
Other diseases	Cold and flu (0.095), conjunctivitis (0.019), urticaria (0.016), nose (0.01), asthma (0.009), cough (0.008), eczema (0.008), winter (0.004), eyes (0.004), physique (0.004)	Multimedia Appendix 6	372

The Possible Pathway of Several Risk Factors Triggers AR

We referred to the relevant literature on the risk factors associated with AR to confirm whether the extracted risk factors were consistent with the general medical consensus. Our findings are novel compared with those in the literature [59]. Previous survey-based studies have explored only the correlation between risk factors and AR, whereas our experimental data provide insight into the potential pathogenesis of reported risk factors. The following section provides a theoretical discussion of potential pathways for several risk factors that trigger AR:

- *Season*: (1) seasonal risk factors are manifested in pollen allergens. Tree allergens such as elm and cypress pollen are prevalent in early spring, followed by ash, pine, and birch pollen in late spring. In summer, grasses, artemisia, and flowering plants grow vigorously owing to increased rainfall, leading to increased pollen spread from these plants. In autumn, weeds account for the largest proportion of pollen allergens. (2) Different climatic conditions in different seasons contribute to the development of allergies. For example, in early spring, frequent cold and high-pressure air activity in East Asia causes intense atmospheric circulation, resulting in alternating hot and cold temperatures that impair the immune regulatory function of the human body, leading to increased allergy attacks. In autumn, changeable weather, large temperature differences, and sunlight and UV radiation can stimulate allergic reactions in people with weak lungs or those who are prone to AR. In addition, seasonal changes and increasing temperature differences between day and night can disrupt the human immune system.
- *Poor habits*: major keywords for this topic were “smoking,” “staying up late,” and “resistance.” (1) Habits such as staying up late, lack of exercise, smoking, and alcohol abuse can weaken immunity and resistance. Gangl et al [60] found that smoking can reduce the integrity and barrier function of respiratory epithelial cells, thereby making smokers more susceptible to allergens. (2) An irregular diet can damage the spleen and stomach, which is also a key factor in the development of AR. (3) The frequent use of air conditioning in summer can cause nasal mucosa irritation owing to temperature fluctuations. Long-term exposure to adverse stimuli can cause dryness of the nasal cavity and weaken the resistance of the mucosal epithelium, which may lead to AR.
- *Allergens*: we grouped clusters that included mites, plants, food, animals, and mold as allergens. (1) The findings of this study suggest that dust mites are the primary allergen, and exposure to a certain concentration of indoor dust mites can lead to AR. The ideal humidity level for dust mite growth is between 75% and 80%, and dust mites tend to thrive during spring and autumn and in warm and humid environments. Studies have shown that a large number of dust mites may be attached to uncleaned air conditioning filters, confirming that air conditioning is an important route of transmission for household dust mites [61]. (2) Allergenic pollen species are closely related to regions and seasons, and some regions now provide pollen concentration and allergy index broadcasts based on meteorological conditions, which is highly convenient for individuals experiencing allergy. (3) Food allergens such as milk, eggs, wheat, soybeans, and peanuts can also trigger AR. (4) Apart from dust mites, other perennial indoor allergens include animal dander, cockroach excrement, and molds.
- *Outdoor environment*: this topic had “dust,” “air quality,” “trust,” and “allergen” as high scoring words. (1) Various substances present in the outdoor environment can trigger AR. Industrialization has increased the content of aromatic hydrocarbon particles, ethanol, and formaldehyde in diesel exhaust, which can damage the mucous membrane and serve as a strong stimulus for AR attacks. (2) Air pollution can affect the distribution of allergens such as mold and pollen. In hazy weather, allergens tend to stay in the air longer, increasing the chance and duration of contact with the human body and leading to AR. (3) High winds can raise dust, pollen, mites, bacteria, and other allergenic factors, increasing their concentration in the air and making it easier to trigger AR.
- *Time points*: patients with AR are more likely to experience symptoms during 2 specific time points, morning and evening. Schenkel et al [62] assessed the severity of 4 nasal symptoms (sneezing, blockage, nasal runny nose, and nasal itch) at different times of the day, revealing that morning

XSL•FO
RenderX

and evening symptoms were the most severe. This may be because of the circadian rhythm, pollen concentration, or personal behavior exacerbating the symptoms. In the evening, when the wind subsides, pollen settles closer to the ground and can be inhaled more easily. In addition, although humans rest at night in a horizontal position, nasal ventilation may be more difficult, leading to more severe symptoms. In the morning, low temperatures can cause congestion and swelling of the nasal mucosa because of the temperature difference between the environment and the body. This cluster had words such as “evening,” “get up early,” and “nose” as highly rated words.

This theoretical discussion regarding the potential pathway of risk factors that trigger AR can guide the development of detailed AR intervention measures. For example, patients with AR can pay attention to pollen concentration and temperature changes and adjust their outings and clothing accordingly based on the characteristics of the season; they can set the air conditioner to turn on or off based on their waking time to reduce the inhalation of cold air when waking up. Furthermore, they can adjust their sleeping position to reduce the frequency of nighttime symptoms.

Discussion

Principal Findings

This study aimed to identify the risk factors for AR based on social media comments. To do so, a data set of comments related to AR was collected, processed, and analyzed. The data set covered a consecutive period from May 2012 to May 2022. Overall, this analysis provided new insights into three main questions: (1) How many comments contained AR risk factor information? (2) How many categories can these risk factors be summarized into? (3) How do these risk factors trigger AR?

In assessing the identification of AR risk factors, we found that TopicS enhanced both precision and recall. TextCNN outperformed other models, achieving an accuracy of 0.9594 with a 150-dimension TopicS embedding. Analyzing 30,372 comments, our model pinpointed 5221 comments with risk factors. Categorizing the text items led to 28 distinct categories, with seasonal factors, regional variations, mites, weather changes, and unhealthy lifestyle habits emerging as common risks.

Furthermore, our research into AR risk factors revealed how risk factors trigger AR and uncovered the frequently reported, but underresearched, risk factors by affected individuals. Seasonal changes, especially during spring and autumn, increase exposure to pollen allergens, with varying climatic conditions affecting the development of allergies. Poor habits, such as smoking, irregular sleep, and frequent use of air conditioning, compromise immunity and heighten AR susceptibility. Dust mites, influenced by humidity, stand out as a primary allergen, with food items and indoor factors, such as animal dander, also triggering AR. Industrial pollutants and outdoor environmental factors amplify AR risk. Notably, AR symptoms intensify during mornings and evenings, which is likely influenced by circadian rhythms and environmental factors.

Limitations and Future Work

This study has some limitations. Our study was based on the self-reported nature of social media data, and the lack of more detailed information from the study participants was a concern. Our statistics showed that seasonal factors, regional variations, mites, weather changes, and unhealthy lifestyle habits emerge as common risk factors, which is consistent with the findings of other studies based on surveys. Although social media may lack in-depth patient information, it provides an effective method of collecting breadth of data. Social media data can be gathered 24 hours a day and are an extremely efficient way to rapidly update new knowledge into the risk factor knowledge base. In the future, our framework can be expanded in 2 ways. First, the framework can track the development trends and changes in AR risk factors by leveraging real-time internet data sets. Second, the framework can be generalized and extended to detect patterns, trends, and risk factors for other chronic diseases such as type 2 diabetes.

Conclusions

In this model improvement study, we proposed a topic-enhanced word-embedding model to improve the accuracy and recall of the text classification, namely to uncover less common or other types of risk factors based on social media data that have not been previously reported. The risk factors identified in this study can be a helpful reference for people with AR to reduce the development of the disease in their daily lives. This study establishes a knowledge base of potential risk factors for individuals who may not be aware of the factors that could trigger their symptoms. Patients can compare their lifestyle habits and medical history to identify their risk factors, which could help reduce the frequency of episodes and prevent the decline in their quality of life caused by blindly avoiding potential triggers. Our findings demonstrate the practicality and feasibility of using social media data for investigating disease knowledge. These findings may provide guidance for the development of management plans and interventions for AR.

Acknowledgments

The data set collection and analysis of this research were partially supported by the National Natural Science Foundation of China (grants 72131006, 72071063, and 72271082); Anhui Provincial Key Research and Development Plan Project (grant 2022i01020003); and the Fundamental Research Funds for the Central Universities (grant JS2023ZSPY0063).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

DG conceptualized and investigated the study. QW drafted the methodology, performed the software analysis, and prepared the original draft. YC reviewed and edited the draft. XY completed the investigation. WZ drafted the methodology and supervised the study. ML supervised the study. ZX conceptualized the study. GZ and ZO supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of social media text, topic dictionary examples, word-embedding dimension parameters with TextRNN, word-embedding dimension parameters with transformer, and social media category distribution and visualization.
[DOCX File , 693 KB - jmir_v26i1e48324_app1.docx]

Multimedia Appendix 2
Word cloud 1.
[PNG File , 119 KB - jmir_v26i1e48324_app2.png]

Multimedia Appendix 3
Word cloud 2.
[PNG File , 120 KB - jmir_v26i1e48324_app3.png]

Multimedia Appendix 4
Word cloud 3.
[PNG File , 122 KB - jmir_v26i1e48324_app4.png]

Multimedia Appendix 5
Word cloud 4.
[PNG File , 121 KB - jmir_v26i1e48324_app5.png]

Multimedia Appendix 6
Word cloud 5.
[PNG File , 111 KB - jmir_v26i1e48324_app6.png]

References

1. Pawankar R, Baena-Cagnani CE, Bousquet J, Walter Canonica G, Cruz AA, Kaliner MA, et al. State of world allergy report 2008: allergy and chronic respiratory diseases. World Allergy Org J 2008;1:S4-17. [doi: 10.1186/1939-4551-1-S1-S4]
2. Krishna MT, Mahesh PA, Vedanthan PK, Mehta V, Moitra S, Christopher DJ. The burden of allergic diseases in the Indian subcontinent: barriers and challenges. Lancet Glob Health 2020 Apr;8(4):e478-e479 [FREE Full text] [doi: 10.1016/S2214-109X(20)30061-9] [Medline: 32199115]
3. Greiner AN, Hellings PW, Rotiroti G, Scadding GK. Allergic rhinitis. Lancet 2011 Dec 17;378(9809):2112-2122. [doi: 10.1016/S0140-6736(11)60130-X] [Medline: 21783242]
4. Wang XD, Zheng M, Lou HF, Wang CS, Zhang Y, Bo MY, et al. An increased prevalence of self-reported allergic rhinitis in major Chinese cities from 2005 to 2011. Allergy 2016 Aug 13;71(8):1170-1180 [FREE Full text] [doi: 10.1111/all.12874] [Medline: 26948849]
5. Price D, Smith P, Hellings P, Papadopoulos N, Fokkens W, Muraro A, et al. Current controversies and challenges in allergic rhinitis management. Expert Rev Clin Immunol 2015 Aug 29;11(11):1205-1217. [doi: 10.1586/1744666x.2015.1081814]
6. Cardell LO, Olsson P, Andersson M, Welin KO, Svensson J, Tennvall GR, et al. TOTALL: high cost of allergic rhinitis-a national Swedish population-based questionnaire study. NPJ Prim Care Respir Med 2016 Feb 04;26(1):15082 [FREE Full text] [doi: 10.1038/npjpcrm.2015.82] [Medline: 26845513]
7. Terreehorst I, Hak E, Oosting AJ, Tempels-Pavlica Z, de Monchy JG, Bruijnzeel-Koomen CA, et al. Evaluation of impermeable covers for bedding in patients with allergic rhinitis. N Engl J Med 2003 Jul 17;349(3):237-246 [FREE Full text] [doi: 10.1056/NEJMoa023171] [Medline: 12867607]
8. Zhang Y, Zhang L. Increasing prevalence of allergic rhinitis in China. Allergy Asthma Immunol Res 2019 Mar;11(2):156-169 [FREE Full text] [doi: 10.4168/aair.2019.11.2.156] [Medline: 30661309]
9. Gao H, Niu Y, Wang Q, Shan G, Ma C, Wang H, et al. Analysis of prevalence and risk factors of adult self-reported allergic rhinitis and asthma in plain lands and hilly areas of Shenmu City, China. Front Public Health 2021 Jan 4;9:749388 [FREE Full text] [doi: 10.3389/fpubh.2021.749388] [Medline: 35059372]

XSL•FO
RenderX

10. Li L, Zhou J, Ma Z, Bensi MT, Hall MA, Baecher GB. Dynamic assessment of the COVID-19 vaccine acceptance leveraging social media data. J Biomed Inform 2022 May;129:104054 [FREE Full text] [doi: 10.1016/j.jbi.2022.104054] [Medline: 35331966]
11. Castillo A, Benitez J, Llorens J, Luo X. Social media-driven customer engagement and movie performance: theory and empirical evidence. Decis Support Syst 2021 Jun;145:113516. [doi: 10.1016/j.dss.2021.113516]
12. Stieglitz S, Dang-Xuan L. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. J Manag Inf Syst 2014 Dec 08;29(4):217-248. [doi: 10.2753/MIS0742-1222290408]
13. Kumar A, Srinivasan K, Wen-Huang C, Zomaya AY. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Inf Process Manag 2020 Jan;57(1):102141. [doi: 10.1016/j.ipm.2019.102141]
14. Liu X, Wang GA, Fan W, Zhang Z. Finding useful solutions in online knowledge communities: a theory-driven design and multilevel analysis. Inf Syst Res 2020 Sep;31(3):731-752. [doi: 10.1287/isre.2019.0911]
15. Lindelöf G, Aledavood T, Keller B. Dynamics of the negative discourse toward COVID-19 vaccines: topic modeling study and an annotated data set of Twitter posts. J Med Internet Res 2023 Apr 12;25:e41319 [FREE Full text] [doi: 10.2196/41319] [Medline: 36877804]
16. Zhu J, Li Z, Zhang X, Zhang Z, Hu B. Public attitudes toward anxiety disorder on Sina Weibo: content analysis. J Med Internet Res 2023 Apr 04;25:e45777 [FREE Full text] [doi: 10.2196/45777] [Medline: 37014691]
17. Shin D, He S, Lee GM, Whinston AB, Cetintas S, Lee KC. Enhancing social media analysis with visual data analytics: a deep learning approach. MIS Q 2020 Dec 1;44(4):1459-1492. [doi: 10.25300/MISQ/2020/14870]
18. Paul M, Dredze M. You are what you tweet: analyzing Twitter for public health. Proc Int AAAI Conf Web Social Media 2021 Aug 03;5(1):265-272. [doi: 10.1609/icwsm.v5i1.14137]
19. Pradeepa S, Manjula KR, Vimal S, Khan MS, Chilamkurti N, Luhach AK. DRFS: detecting risk factor of stroke disease from social media using machine learning techniques. Neural Process Lett 2020 Jun 09;55(4):3843-3861. [doi: 10.1007/s11063-020-10279-8]
20. Xie J, Liu X, Zeng DD, Fang X. Understanding medication nonadherence from social media: a sentiment-enriched deep learning approach. MIS Q 2022 Feb 25;46(1):341-372. [doi: 10.25300/MISQ/2022/15336]
21. Navale V, McAuliffe M. The integration of a canonical workflow framework with an informatics system for disease area research. Data Intell 2022;4(2):186-195. [doi: 10.1162/dint_a_00125]
22. Zhang Y, Lan F, Zhang L. Advances and highlights in allergic rhinitis. Allergy 2021 Nov 17;76(11):3383-3389. [doi: 10.1111/all.15044] [Medline: 34379805]
23. Chiang TY, Yuan TH, Shie RH, Chen CF, Chan CC. Increased incidence of allergic rhinitis, bronchitis and asthma, in children living near a petrochemical complex with SO pollution. Environ Int 2016 Nov;96:1-7. [doi: 10.1016/j.envint.2016.08.009] [Medline: 27585759]
24. Kurganskiy A, Creer S, de Vere N, Griffith GW, Osborne NJ, Wheeler BW, PollerGEN Consortium, et al. Predicting the severity of the grass pollen season and the effect of climate change in Northwest Europe. Sci Adv 2021 Mar 26;7(13):eabd7658 [FREE Full text] [doi: 10.1126/sciadv.abd7658] [Medline: 33771862]
25. Lee JY, Lee J, Huh DA, Moon KW. Association between environmental exposure to phthalates and allergic disorders in Korean children: Korean National Environmental Health Survey (KoNEHS) 2015-2017. Int J Hyg Environ Health 2021 Sep;238:113857. [doi: 10.1016/j.ijheh.2021.113857] [Medline: 34644676]
26. Paciência I, Cavaleiro Rufo J, Silva D, Mendes F, Farraia M, Delgado L, et al. Effects of indoor endocrine-disrupting chemicals on childhood rhinitis. J Investig Allergol Clin Immunol 2020 Jun 18;30(3):195-197 [FREE Full text] [doi: 10.18176/jiaci.0471] [Medline: 31833476]
27. Saulyte J, Regueira C, Montes-Martínez A, Khudyakov P, Takkouche B. Active or passive exposure to tobacco smoking and allergic rhinitis, allergic dermatitis, and food allergy in adults and children: a systematic review and meta-analysis. PLoS Med 2014 Mar 11;11(3):e1001611 [FREE Full text] [doi: 10.1371/journal.pmed.1001611] [Medline: 24618794]
28. Kong IG, Rhee CS, Lee JW, Yim H, Kim MJ, Choi Y, et al. Association between perceived stress and rhinitis-related quality of life: a multicenter, cross-sectional study. J Clin Med 2021 Aug 19;10(16):3680 [FREE Full text] [doi: 10.3390/jcm10163680] [Medline: 34441978]
29. Han YY, Forno E, Gogna M, Celedón JC. Obesity and rhinitis in a nationwide study of children and adults in the United States. J Allergy Clin Immunol 2016 May;137(5):1460-1465 [FREE Full text] [doi: 10.1016/j.jaci.2015.12.1307] [Medline: 26883461]
30. Kanazawa J, Masuko H, Yatagai Y, Sakamoto T, Yamada H, Kitazawa H, et al. Association analyses of eQTLs of the TYRO3 gene and allergic diseases in Japanese populations. Allergol Int 2019 Jan;68(1):77-81 [FREE Full text] [doi: 10.1016/j.alit.2018.07.004] [Medline: 30082152]
31. Alm B, Goksör E, Pettersson R, Möllborg P, Erdes L, Loid P, et al. Antibiotics in the first week of life is a risk factor for allergic rhinitis at school age. Pediatr Allergy Immunol 2014 Aug 09;25(5):468-472 [FREE Full text] [doi: 10.1111/pai.12244] [Medline: 24912441]
32. Ho CL, Wu WF. Risk factor analysis of allergic rhinitis in 6-8 year-old children in Taipei. PLoS One 2021 Apr 2;16(4):e0249572 [FREE Full text] [doi: 10.1371/journal.pone.0249572] [Medline: 33798255]

XSL•FO
RenderX

33. Zhang W, Ram S. A comprehensive analysis of triggers and risk factors for asthma based on machine learning and large heterogeneous data sources. MIS Q 2020 Jan 01;44(1):305-349. [doi: 10.25300/misq/2020/15106]
34. Griffis H, Asch DA, Schwartz HA, Ungar L, Buttenheim AM, Barg FK, et al. Using social media to track geographic variability in language about diabetes: analysis of diabetes-related tweets across the United States. JMIR Diabetes 2020 Jan 26;5(1):e14431 [FREE Full text] [doi: 10.2196/14431] [Medline: 32044757]
35. Schäfer F, Faviez C, Voillot P, Foulquié P, Najm M, Jeanne JF, et al. Mapping and modeling of discussions related to gastrointestinal discomfort in French-speaking online forums: results of a 15-year retrospective infodemiology study. J Med Internet Res 2020 Nov 03;22(11):e17247 [FREE Full text] [doi: 10.2196/17247] [Medline: 33141087]
36. Oyebode O, Orji R. Detecting factors responsible for diabetes prevalence in Nigeria using social media and machine learning. In: Proceedings of the 15th International Conference on Network and Service Management (CNSM). 2019 Presented at: 15th International Conference on Network and Service Management (CNSM); October 21-25, 2019; Halifax, NS. [doi: 10.23919/cnsm46954.2019.9012679]
37. Ramsingh J, Bhuvaneswari V. A big data framework to analyze risk factors of diabetes outbreak in indian population using a map reduce algorithm. In: Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS). 2018 Presented at: Second International Conference on Intelligent Computing and Control Systems (ICICCS); June 14-15, 2018; Madurai, India. [doi: 10.1109/iccons.2018.8663143]
38. Ramsingh J, Bhuvaneswari V. An integrated multi-node Hadoop framework to predict high-risk factors of diabetes mellitus using a multilevel MapReduce based fuzzy classifier (MMR-FC) and modified DBSCAN algorithm. Appl Soft Comput 2021 Sep;108:107423. [doi: 10.1016/j.asoc.2021.107423]
39. Alswedani S, Mehmood R, Katib I, Altowaijri SM. Psychological health and drugs: data-driven discovery of causes, treatments, effects, and abuses. Toxics 2023 Mar 20;11(3):102681. [doi: 10.3390/toxics11030287] [Medline: 36977052]
40. Chung JE, Mustapha IZ, Li J, Gu X. Discourse about human papillomavirus (HPV)-associated oropharyngeal cancer (OPC) on Twitter: lessons for public health education about OPC and dental care. Public Health Pract (Oxf) 2022 Jun;3:100239 [FREE Full text] [doi: 10.1016/j.puhip.2022.100239] [Medline: 36101754]
41. Neisani Samani Z, Karimi M, Alesheikh A. Environmental and infrastructural effects on respiratory disease exacerbation: a LBSN and ANN-based spatio-temporal modelling. Environ Monit Assess 2020 Jan 04;192(2):90. [doi: 10.1007/s10661-019-7987-x] [Medline: 31902018]
42. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv. Preprint posted online October 16, 2013 [FREE Full text] [doi: 10.5555/2999792.2999959]
43. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); June 22-27, 2014; Baltimore, MD. [doi: 10.3115/v1/p14-1146]
44. Yilmaz S, Toklu S. A deep learning analysis on question classification task using Word2vec representations. Neural Comput Appl 2020 Jan 21;32(7):2909-2928. [doi: 10.1007/s00521-020-04725-w]
45. Shi B, Fu Z, Bing L, Lam W. Learning domain-sensitive and sentiment-aware word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1232]
46. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1181]
47. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv. Preprint posted online September 8, 2014 [FREE Full text] [doi: 10.48550/ARXIV.1409.2329]
48. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online June 12, 2017 .
49. Zhang K, Moe W. Measuring brand favorability using large-scale social media data. Inf Syst Res 2021 Dec;32(4):1128-1139. [doi: 10.1287/isre.2021.1030]
50. Abbasi A, Li J, Adjeroh D, Abate M, Zheng W. Don't mention it? Analyzing user-generated content signals for early adverse event warnings. Inf Syst Res 2019 Sep;30(3):1007-1028. [doi: 10.1287/isre.2019.0847]
51. Guo Q, Chen W, Wan H. AOL4PS: a large-scale data set for personalized search. Data Intell 2021;3(4):548-567. [doi: 10.1162/dint_a_00104]
52. Khatua A, Khatua A, Cambria E. A tale of two epidemics: contextual Word2Vec for classifying Twitter streams during outbreaks. Inf Process Manag 2019 Jan;56(1):247-257. [doi: 10.1016/j.ipm.2018.10.010]
53. Gu D, Li M, Yang X, Gu Y, Zhao Y, Liang C, et al. An analysis of cognitive change in online mental health communities: a textual data analysis based on post replies of support seekers. Inf Process Manag 2023 Mar;60(2):103192. [doi: 10.1016/j.ipm.2022.103192]
54. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001 Nov 10;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]

XSL•FO
RenderX

55. Shen D, Yang Q, Sun JT, Chen Z. Thread detection in dynamic text message streams. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006 Presented at: SIGIR '06; August 6-11, 2006; Seattle, WA. [doi: 10.1145/1148170.1148180]
56. Mihalcea R, Tarau P. TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004 Presented at: 2004 Conference on Empirical Methods in Natural Language Processing; July 25-26, 2004; Barcelona, Spain URL: https://aclanthology.org/W04-3252.pdf
57. Long S, Yan L. Rank-IDF: a statistical and network based feature words selection in big data text analysis. In: Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence. 2020 Presented at: ICMAI '20; April 10-13, 2020; Chengdu, China. [doi: 10.1145/3395260.3395291]
58. Wright AP, Jones CM, Chau DH, Matthew Gladden R, Sumner SA. Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media. J Biomed Inform 2021 Jul;119:103824 [FREE Full text] [doi: 10.1016/j.jbi.2021.103824] [Medline: 34048933]
59. Zhou J, Zhang Q, Zhou S, Li X, Zhang X. Unintended emotional effects of online health communities: a text mining-supported empirical study. MIS Q 2023 Mar 01;47(1):195-226. [doi: 10.25300/misq/2022/17018]
60. Gangl K, Reininger R, Bernhard D, Campana R, Pree I, Reisinger J, et al. Cigarette smoke facilitates allergen penetration across respiratory epithelium. Allergy 2009 Mar 23;64(3):398-405. [doi: 10.1111/j.1398-9995.2008.01861.x] [Medline: 19120070]
61. Liu Z, Bai Y, Ji K, Liu X, Cai C, Yu H, et al. Detection of dermatophagoides farinae in the dust of air conditioning filters. Int Arch Allergy Immunol 2007 Aug 10;144(1):85-90. [doi: 10.1159/000102619] [Medline: 17505140]
62. Schenkel EJ. Effect of desloratadine on the control of morning symptoms in patients with seasonal and perennial allergic rhinitis. Allergy Asthma Proc 2006 Nov 01;27(6):465-472. [doi: 10.2500/aap.2006.27.2936] [Medline: 17176780]

Abbreviations

AR: allergic rhinitis
CBOW: Continuous Bag-of-Words Model
CNN: convolutional neural network
RNN: recurrent neural network

Edited by A Mavragani; submitted 19.04.23; peer-reviewed by X Liu, Y Cao; comments to author 12.10.23; revised version received 30.10.23; accepted 03.01.24; published 22.02.24.

Please cite as:
Gu D, Wang Q, Chai Y, Yang X, Zhao W, Li M, Zolotarev O, Xu Z, Zhang G
Identifying the Risk Factors of Allergic Rhinitis Based on Zhihu Comment Data Using a Topic-Enhanced Word-Embedding Model: Mixed Method Study and Cluster Analysis
J Med Internet Res 2024;26:e48324
URL: https://www.jmir.org/2024/1/e48324
doi:10.2196/48324
PMID:38386404

©Dongxiao Gu, Qin Wang, Yidong Chai, Xuejie Yang, Wang Zhao, Min Li, Oleg Zolotarev, Zhengfei Xu, Gongrang Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

What Adolescents Say in Text Messages to Motivate Peer Networks to Access Health Care and Sexually Transmitted Infection Testing: Qualitative Thematic Analysis

Marguerita Lightfoot[1,2], PhD; Chadwick Campbell[2,3], PhD; Allysha C Maragh-Bass[4,5], PhD; Joi Jackson-Morgan[6], MPH; Kelly Taylor[2,7], PhD

[1]School of Public Health, Oregon Health & Science University – Portland State University, Portland, OR, United States
[2]Division of Prevention Science, University of California, San Francisco, San Francisco, CA, United States
[3]Herbert Wertheim School of Public Health & Human Longevity Science, University of California, San Diego, San Diego, CA, United States
[4]Behavioral, Epidemiological, Clinical Sciences Division, FHI 360, Durham, NC, United States
[5]Duke Global Health Institute, Duke University, Durham, NC, United States
[6]3rd Street Youth Center and Clinic, San Francisco, CA, United States
[7]Institute for Global Health Sciences, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Marguerita Lightfoot, PhD
School of Public Health
Oregon Health & Science University – Portland State University
1810 SW 5th Ave., Suite 510
Portland, OR, 97201
United States
Phone: 1 503 494 4345
Email: lightfom@ohsu.edu

Abstract

Background: While rates of HIV and sexually transmitted infections (STIs) are extremely high among adolescents and young adults in the United States, rates of HIV and STI testing remain low. Given the ubiquity of mobile phones and the saliency of peers for youths, text messaging strategies may successfully promote HIV or STI testing among youths.

Objective: This study aimed to understand the types of messages youths believe were motivating and persuasive when asked to text friends to encourage them to seek HIV or STI testing services at a neighborhood clinic.

Methods: We implemented an adolescent peer-based text messaging intervention to encourage clinic attendance and increase STI and HIV testing among youths (n=100) at an adolescent clinic in San Francisco, California. Participants were asked to send a text message to 5 friends they believed were sexually active to encourage their friends to visit the clinic and receive STI or HIV screening. Thematic analysis was used to analyze the content of the text messages sent and received during the clinic visit. Member checking and consensus coding were used to ensure interrater reliability and significance of themes.

Results: We identified four themes in the messages sent by participants: (1) calls to action to encourage peers to get tested, (2) personalized messages with sender-specific information, (3) clinic information such as location and hours, and (4) self-disclosure of personal clinic experience. We found that nearly all text messages included some combination of 2 or more of these broad themes. We also found that youths were inclined to send messages they created themselves, as opposed to sending the same message to each peer, which they tailored to each individual to whom they were sent. Many (40/100, 40%) received an immediate response to their message, and most participants reported receiving at least 1 positive response, while a few reported that they had received at least 1 negative response. There were some differences in responses depending on the type of message sent.

Conclusions: Given the high rates of STI and HIV and low rates of testing among adolescents, peer-driven text messaging interventions to encourage accessing care may be successful at reaching this population. This study suggests that youths are willing to text message their friends, and there are clear types of messages they develop and use. Future research should use these methods with a large, more diverse sample of youths and young adults for long-term evaluation of care seeking and care retention outcomes to make progress in reducing HIV and STI among adolescents and young adults.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e44861) doi:10.2196/44861

KEYWORDS

adolescents; clinics; HIV/STI testing; intervention; mobile health; peer; screening; sexually active; STI; text messaging; young adult

Introduction

Adolescents are overrepresented among those diagnosed with sexually transmitted infections (STIs). More than half of all STIs in the United States are among adolescents aged between 15 and 24 years [1]. More specifically, adolescents aged between 15 and 24 years comprise 62.6% of all chlamydia infections, and there were increases in rates of gonorrhea, syphilis, and chlamydia incidence among young men and women in this age group between 2016 and 2017 [1]. However, STI testing rates among adolescents are low. Only 27.8% of sexually experienced females and 9.6% of sexually experienced male candidates, aged between 15 and 25 years, reported being tested for STIs in the previous 12 months. Importantly, there are gender differences in testing rates. Among those with 4 or more lifetime partners, 40.8% of females reported receiving an STI test in the previous 12 months, compared to only 3.1% of males [2]. Numerous barriers to STI testing have been identified, including concerns about confidentiality, embarrassment, limits to access (eg, clinic hours and payment), and low-risk perception [1,2].

In 2018, adolescents aged between 13 and 24 years accounted for 21% of new HIV diagnoses in the United States and were the least likely to be virally suppressed of all groups living with HIV [3]. Among those aged between 13 and 19 years, the highest rate of new diagnoses was among African American individuals. Among those aged between 20 and 24 years, those identifying as African American, multiracial, and Hispanic had the first, second, and third highest new diagnosis rates, respectively. These rates are alarming, and they are between 2 and 7 times the rate of new diagnoses for all adults and adolescents in the United States [1]. Despite the high percentage of HIV diagnoses among teens and young adults, similar to STI testing rates, HIV testing rates are low. Among high school students who have ever had sex, the estimated proportion of having ever received an HIV test is between 13% and 22%, and only 33% of those aged between 18 and 24 years have ever tested for HIV [4]. It is estimated that only 40% of those aged between 13 and 29 years who live with HIV are aware of their status [3]. Many barriers exist related to adolescents and young adults getting tested and knowing their status. Some adolescents reported low perceptions of risk, concerns about confidentiality, difficulty remembering to go and get tested, and very few cues to action for HIV testing, such as being offered a test by a medical provider [5]. Profoundly apparent in these numbers, strategies that encourage adolescents and young adults to get tested for STIs and HIV are desperately needed.

A model that uses peer-driven technology-based communication methods to encourage testing among youths may be effective in decreasing the barriers reported in this population. Recent data suggest that, across race and ethnicity, gender, and socioeconomic status, as many as 95% of teens in the United States have access to a smartphone [6]. Text message–based interventions have shown promise in affecting positive changes in health behavior for a number of health issues, including diabetes, weight loss, medication adherence, and smoking cessation [7]. With widespread phone access among American adolescents and with the average youth sending more than 30 messages a day, mobile phone–based health interventions offer an opportunity to reach many adolescents and young adults to improve health outcomes [8]. Importantly, youths have responded positively to receiving SMS text messages as a part of health promotion activities. Youths (aged between 16 and 20 years) who identified as African American and heterosexual expressed openness to receiving HIV prevention and sexual health messages by text and expressed a desire for a text intervention to include the opportunity to build community and share information with each other [9]. A study of adolescents aged between 15 and 20 years described text messaging as convenient for health promotion because they are “always on their phones” [10] and a text-based peer education intervention in which youths could text sexual health questions to a peer health educator was found to be feasible and acceptable [8]. However, in each of these interventions, text messages were sent by clinic or research staff, and the messages themselves were standardized.

Peer strategies are based on the premise that members of a social network know and trust each other and often share similar health behaviors [11]. Building on the widespread usage of text messaging and the important role of peer influence, we implemented a peer-delivered texting intervention to increase attendance at a youth health clinic in San Francisco. This word-of-mouth communication strategy [12] is a practical application of diffusion theory [13]. Studies indicate this communication strategy is highly persuasive [14]. The implementation of the peer-based text messaging strategy showed promise for both increasing the number of youths accessing health services and finding youths engaging in sexual risk behaviors and most in need of sexual health screening and services [15]. This study will examine the SMS text messages that adolescents and young adults sent in order to explicate the types of messages they believed would be persuasive among their friends. Understanding what young people find motivating and persuasive can inform other peer-based models of influence and intervention.

Methods

Study Design

We analyzed the content of messages sent and received during the initial clinic visit.

Study Population

A total of 100 participants were recruited from an adolescent medicine clinic in San Francisco, California, which served youths aged between 12 and 24 years. We approached

XSL•FO
RenderX

consecutive adolescent clinic patients to send text messages to 5 peers they believed were sexually active and lived in the clinic's service area. Of the 153 patients approached by research staff, 18% (28/153) refused, 16% (25/153) did not have a phone with them, and 65% (100/153) provided informed consent and sent text messages to at least 5 peers.

Participants were told that the goal of their messages was to encourage their friends to visit the clinic and receive STI screening. Participants developed their own messages and were provided with an SMS text messaging guide that reviewed considerations for developing messages (eg, let your friend know you care about them, example messages). The SMS text message and any responses from friends while the participant was still in the clinic were recorded by research staff.

Data Analysis

We analyzed text messages using thematic analysis, as described by Braun and Clarke [16,17]. The second author conducted the initial analysis, including reading all text messages, generating initial codes, and identifying emergent themes. After completing the initial analysis, the first author reviewed the codes and initial themes. Together, the first and second authors discussed the initial analysis, merged related themes, identified the most salient themes, and eliminated those that lacked enough supporting data. Decision trails were maintained throughout the coding and analysis process.

Ethical Considerations

The study was approved by the University of California, San Francisco, institutional review board (#12-08516).

Results

Participants

A total of 430 new patients were seen at the clinic during data collection, of whom 78.6% (338/430) were female. The new patients were 61.2% (263/430) African American, 17.4% (75/430) Hispanic, had a mean age of 19.6 (SD 2.6 years, range 13-25), and 86.9% (374/430) were primarily identified as heterosexual. The demographics of participants sending messages mirrored those of the overall clinic.

All participants sent 5 messages, with 18% (18/100) sending more than the required 5. In the brief posttexting interview, most participants reported no concerns, with 55% (55/100) sending the text message, 18% (18/100) expressing worry about their friends' response, and 27% (27/100) expressing concern that their friend would not heed their advice (eg, friend would not call the clinic or a friend would not listen). In the 1-week postmessaging interview, no youth reported adverse events or negative outcomes from sending the message.

Thematic Findings

We identified 4 overarching themes that emerged from participants' text messages. Nearly all text messages (101/108, 93.5%) included some combination of 2 or more of these themes. The four themes were (1) calls to action, (2) personalized messages, (3) clinic information, and (4) self-disclosure of services received at the clinic or positive experiences and perceptions about clinic services and staff.

Calls to Action

The most common characteristic of text messages (94/108, 87% of all messages) were calls to action (CTA) that were intended to encourage friends to get tested and know their status, or to get some other service (eg, birth control, a mom's group, or job services). Some CTAs were more direct than others. For example, some texts directly called for their friend to come to the clinic to receive a service (eg, "…you should check it out and get tested for stds…" (participant 40). Others were less direct, which we coined passive CTA, and included messages that were more general about the importance of staying on top of one's health (eg, "Its a good idea to come to 3rd st youth clinic and get checked out" (participant 70).

Personalized Messages

Although clinic staff provided participants with an SMS text message guide including prewritten text messages, most SMS text messages (87/108, 80.6%) were written by participants. Most of these SMS text messages simply informed the recipient of general clinic information and encouraged clinic attendance, while a small number of these messages (n=10) were more personal in nature. For example, some referred to something specific about themselves in the message (eg, "…remember I was a health educator…" participant 102) or referenced something specific to the text recipient ("wen u cum bck frm Texas, [I'm] here at the clinic on 3rd, u shuld stp bi" participant 5). Others started with "hey. This is [name]."

Clinic Information

More than two-thirds (74/108, 68.5%) of SMS text messages included clinic-specific information. Some included basic information about the clinic, such as the address or other location marker (eg, "…It's near the new wing stop on 3rd…" participant 58), phone number, or hours of operation. Others included specific information about services offered at the clinic (eg, "…They offer free std testing…" participant 78).

Self-Disclosure and Positive Clinic Experiences

In more than half (60/108, 55.5%) of SMS text messages, adolescents disclosed something about their own clinic participation (34/108, 31.5%) or positive experiences about the clinic (26/108, 24.1%). These included general statements about having been to the clinic, such as "…theres a good clinic ive been going for years…" (participant 48). Others included more specific information on why they attended the clinic: "…I went and got tested. And I know my status…" (participant 26). In addition to disclosing their own clinic attendance, some participants included positive information about their own experience at the clinic (eg, "I just got checked out at the 3rd st clinic, and they were really nice…" (participant 41). Others did not disclose anything about their personal experience but shared positive perceptions of the clinic: "…they will not tell your business to no one so go check them out…" (participant 103). Within this theme, 41.6% (25/60) included both self-disclosure and positive experiences, while most (35/60, 58.3%) disclosed their clinic attendance only or did not disclose

XSL•FO
RenderX

receiving any services but shared positive perceptions about the clinic.

Responses Received from Peers

A total of 40% (40/100) participants reported that they received responses from the friends they texted. Among those reporting that they received a response, most 57.5% (23/40) reported receiving at least 1 positive response, while 10% (4/40) reported that they had received at least 1 negative response. There were some differences in responses depending on the type of message sent. Of those using personalized messages, 32% (28/87) received any response, and of those responses, 60.7% (17/28) were positive. Approximately a quarter (15/100, 15%) of the staff provided SMS text messages as is without making any changes. Of those, 53.3% (8/15) received some type of response, and 37.5% (3/8) of those responses were positive. While a greater proportion of the staff providing SMS text messages received a response compared to the personalized messages, a smaller proportion of them received positive responses. Similarly, of those text messages that used direct calls to action, 32.1% (17/53) received any response, with 70.6% (12/17) of those being positive. Of those using passive calls to action, 41.5% (17/41) received a response, but only 41.2% (7/17) of those were positive responses.

Discussion

Principal Results

The overarching purpose of this study was to examine what youths view as motivating and persuasive when participating in peer-driven text messaging approaches to promoting sexual health service access. Consistent with the word-of-mouth communication strategy [12], we found that youths viewed tailoring the content as a facilitator of a response from their intended recipient, as they chose not to use the few predesigned example messages found in the provided guidance. Tailoring messages is a health communication best practice previously used in SMS-based counseling interventions with youths [18-20]. Over 40% (40/100) of participants reported that their peers replied to their message. There were few negative text responses and no negative consequences or adverse events reported by participants in follow-up interviews. This suggests that the messages sent by participants may also be acceptable to the recipients of the message.

We categorized the content of the messages into four themes and noted that most messages used at least two of the four themes: (1) calls to action, (2) personalized messages, (3) clinic information, and (4) self-disclosure of services received or positive clinic experiences and perceptions. Calls to action, where youths encouraged their peers to access HIV or STI testing and other health services, were highly prevalent in the messages that participants crafted. Peer influence has been widely studied in a number of maladaptive behaviors [21]. However, direct peer endorsement of services has been demonstrated to promote health-seeking behaviors of a wide variety, including HIV prevention [22,23], physical activity [24], and most recently, COVID-19 vaccination uptake. Future research should explore the use of peer-designed and peer-delivered text interventions to promote different health-seeking behaviors related to sexual health, including web-based information seeking and the engagement of others in decision-making related to sexual behaviors.

Personalized messages included informing peers about clinic information as well as whether the individual sending the message had accessed the clinic. More than two-thirds of messaging also included specific information about the clinic (eg, location, hours of operation, and services). Informing peers about available services, hours of operation, and location suggests that peers view this information as important for facilitating access. This may be because it facilitated transportation seeking, on which many youths would be dependent [25,26].

Finally, self-disclosure about visiting the clinic and reporting positive experiences may also serve to encourage clinic attendance through reporting patient satisfaction [27]. Self-disclosure may also serve to normalize and destigmatize sexual health service use, which has important implications for future research since stigma is a major barrier to sexual health service use among adolescents and young adults [28,29]. Given that youths often share many similarities with their peers, future research should evaluate the impact of self-reported positive care experiences on health care access and health care stigma experiences among youths.

Limitations

Several limitations exist with this study's findings. While we did record the immediate response from message recipients, we were unable to examine if all recipients read the messages or to directly ask recipients their reactions. This study was also conducted within the initial phase of a text message intervention deployed in an adolescent medicine clinic serving youths aged between 12 and 24 years. Therefore, we cannot ascertain the direct relationship between messages and actual testing rates, HIV or STI diagnose rates within this study's analyses. Further, the patient population included in this study and the larger intervention are likely insured with some existing access to care and may not represent youths and young adults experiencing profound marginalization and the greatest HIV or STI inequities. Lastly, our findings are supported by existing literature but may not have generalizability to all youths, youths in the region where the research was conducted, or the youths included in the larger study.

Conclusions

The disproportionately high STI and HIV incidence among adolescents and young adults underscores the need for identifying approaches to interrupt the spread of STIs and HIV and prevent new infections, particularly among minority youths. Adolescents rely on text messaging as a primary method of communication, which has implications for leveraging peer influence for health promotion, especially among youths of color who experience the highest HIV and STI incidence. Future interventions should incorporate guidance for youths as they craft messages for sharing with peers. Self-disclosure of care seeking was a common method used by youths, which may be a powerful approach for destigmatizing HIV and STI testing and should be expanded on in future research and trainings with

XSL•FO
RenderX

youths as they craft messages. The use of low-cost, easily accessible peer-driven text-based interventions to promote sexual health care service usage among adolescents and young adults has great potential for reducing inequities and improving health outcomes.

Acknowledgments

The authors would like to thank the youths who participated in this study and the staff at the 3rd Street Youth Center and Clinic. This work was supported by the National Institute of Drug Abuse (1R34DA031806-01A1) to the first author. The National Institute on Drug Abuse had no role in the design of the study, in the collection, analysis, and interpretation of the data, or in writing the manuscript.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ML obtained funding for and supervised the study. ML and JJM conceptualized, designed, and implemented the study. ML and CC analyzed and interpreted the data. ML and CC prepared the manuscript and were major contributors to its writing. ACMB and KT were major contributors to reviewing and editing the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Centers for Disease Control Prevention. Diagnoses of HIV infection among adolescents and young adults in the United States and 6 dependent areas, 2012–2017. HIV Surveill Suppl Rep 2018;24(5):1-47 [FREE Full text]
2. Cuffe KM, Newton-Levinson A, Gift TL, McFarlane M, Leichliter JS. Sexually transmitted infection testing among adolescents and young adults in the United States. J Adolesc Health 2016;58(5):512-519 [FREE Full text] [doi: 10.1016/j.jadohealth.2016.01.002] [Medline: 26987687]
3. Zanoni BC, Mayer KH. The adolescent and young adult HIV cascade of care in the United States: exaggerated health disparities. AIDS Patient Care STDS 2014;28(3):128-135 [FREE Full text] [doi: 10.1089/apc.2013.0345] [Medline: 24601734]
4. Kann L, McManus T, Harris WA, Shanklin SL, Flint KH, Queen B, et al. Youth risk behavior surveillance - United States, 2017. MMWR Surveill Summ 2018;67(8):1-114 [FREE Full text] [doi: 10.15585/mmwr.ss6708a1] [Medline: 29902162]
5. Schnall R, Rojas M, Travers J. Understanding HIV testing behaviors of minority adolescents: a health behavior model analysis. J Assoc Nurses AIDS Care 2015;26(3):246-258 [FREE Full text] [doi: 10.1016/j.jana.2014.08.005] [Medline: 25283353]
6. Anderson M, Jiang J. Teens, social media and technology 2018. Pew Research Center. 2018. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2022-06-01]
7. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015;36:393-415 [FREE Full text] [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
8. O'Malley TL, Horowitz KR, Garth J, Mair C, Burke JG. A technology-based peer education intervention: results from a sexual health textline feasibility study. Am J Sex Educ 2017;12(4):383-394 [FREE Full text] [doi: 10.1080/15546128.2017.1372831]
9. Wright E, Fortune T, Juzang I, Bull S. Text messaging for HIV prevention with young Black men: formative research and campaign development. AIDS Care 2011;23(5):534-541 [FREE Full text] [doi: 10.1080/09540121.2010.524190] [Medline: 21287416]
10. Perry RCW, Kayekjian KC, Braun RA, Cantu M, Sheoran B, Chung PJ. Adolescents' perspectives on the use of a text messaging service for preventive sexual health promotion. J Adolesc Health 2012;51(3):220-225 [FREE Full text] [doi: 10.1016/j.jadohealth.2011.11.012] [Medline: 22921131]
11. Social Network Strategy (SNS) for HIV testing recruitment. Effective interventions. Centers for Disease Control Prevention. URL: https://cdc.gov/hiv/effective-interventions/diagnose/social-network-strategy?Sort=Title%3A%3Aasc [accessed 2024-01-18]
12. Hennig-Thurau T, Walsh G, Walsh G. Electronic word-of-mouth: motives for and consequences of reading customer articulations on the internet. Int J Electron Commer 2014;8(2):51-74 [FREE Full text] [doi: 10.1080/10864415.2003.11044293]
13. Rogers EM, Singhal A, Quinlan MM. Diffusion of Innovations. New York: Routledge; 2014.

XSL•FO
RenderX

14. Bansal HS, Voyer PA. Word-of-mouth processes within a services purchase decision context. J Serv Res 2016;3(2):166-177 [FREE Full text] [doi: 10.1177/109467050032005]
15. Lightfoot M, Jackson-Morgan J, Pollack L, Bennett A. Acceptability and feasibility of peer-to-peer text messaging among adolescents to increase clinic visits and sexually transmitted infection testing: interrupted times-series analysis. JMIR Form Res 2022;6(6):e32416 [FREE Full text] [doi: 10.2196/32416] [Medline: 35686737]
16. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
17. Braun V, Clarke V. What can "thematic analysis" offer health and wellbeing researchers? Int J Qual Stud Health Well-being 2014;9:26152 [FREE Full text] [doi: 10.3402/qhw.v9.26152] [Medline: 25326092]
18. Cortese J, Lustria MLA. Can tailoring increase elaboration of health messages delivered via an adaptive educational site on adolescent sexual health and decision making? J Am Soc Inf Sci Tec 2012;63(8):1567-1580. [doi: 10.1002/asi.22700]
19. Castro RP, Haug S, Debelak R, Jakob R, Kowatsch T, Schaub MP. Engagement with a mobile phone-based life skills intervention for adolescents and its association with participant characteristics and outcomes: tree-based analysis. J Med Internet Res 2022;24(1):e28638 [FREE Full text] [doi: 10.2196/28638] [Medline: 35044309]
20. Schwebel FJ, Larimer ME. Text message reminders as an adjunct to a substance use intervention for adolescents and young adults: pilot feasibility and acceptability findings. Digit Health 2020;6:2055207620965052 [FREE Full text] [doi: 10.1177/2055207620965052] [Medline: 33110614]
21. Henneberger AK, Mushonga DR, Preston AM. Peer influence and adolescent substance use: a systematic review of dynamic social network research. Adolescent Res Rev 2020;6(1):57-73. [doi: 10.1007/s40894-019-00130-0]
22. He J, Wang Y, Du Z, Liao J, He N, Hao Y. Peer education for HIV prevention among high-risk groups: a systematic review and meta-analysis. BMC Infect Dis 2020;20(1):338 [FREE Full text] [doi: 10.1186/s12879-020-05003-9] [Medline: 32398032]
23. Sodik MA. Analysis of improved attitude of youth in HIV/AIDS prevention through the provision of health education with peer education. 2nd Jt Int Conf 2018;2(2):495-502 [FREE Full text]
24. Haidar A, Ranjit N, Archer N, Hoelscher DM. Parental and peer social support is associated with healthier physical activity behaviors in adolescents: a cross-sectional analysis of Texas School Physical Activity and Nutrition (TX SPAN) data. BMC Public Health 2019;19(1):640 [FREE Full text] [doi: 10.1186/s12889-019-7001-0] [Medline: 31132999]
25. Doll M, Fortenberry JD, Roseland D, McAuliff K, Wilson CM, Boyer CB. Linking HIV-negative youth to prevention services in 12 U.S. cities: barriers and facilitators to implementing the HIV prevention continuum. J Adolesc Health 2018;62(4):424-433. [doi: 10.1016/j.jadohealth.2017.09.009] [Medline: 29224988]
26. Philbin MM, Tanner AE, Duval A, Ellen J, Kapogiannis B, Fortenberry JD. Linking HIV-positive adolescents to care in 15 different clinics across the United States: creating solutions to address structural barriers for linkage to care. AIDS Care 2014;26(1):12-19 [FREE Full text] [doi: 10.1080/09540121.2013.808730] [Medline: 23777542]
27. Wilson KS, Beima-Sofie KM, Moraa H, Wagner AD, Mugo C, Mutiti PM, et al. "At our age, we would like to do things the way we want:" a qualitative study of adolescent HIV testing services in Kenya. AIDS 2017;31(Suppl 3):S213-S220 [FREE Full text] [doi: 10.1097/QAD.0000000000001513] [Medline: 28665879]
28. Tang W, Ritchwood TD, Wu D, Ong JJ, Wei C, Iwelunmor J, et al. Crowdsourcing to improve HIV and sexual health outcomes: a scoping review. Curr HIV/AIDS Rep 2019;16(4):270-278 [FREE Full text] [doi: 10.1007/s11904-019-00448-3] [Medline: 31155691]
29. Ibitoye M, Lappen H, Freeman R, Jordan AE, Aronson ID. Technology-based interventions to increase point-of-care HIV testing and linkage to care among youth in the US: a systematic review. AIDS Behav 2021;25(6):1829-1838 [FREE Full text] [doi: 10.1007/s10461-020-03112-9] [Medline: 33231846]

Abbreviations

CTA: call to action
STI: sexually transmitted infection

Edited by T Leung; submitted 06.12.22; peer-reviewed by R Chandrasekaran, R Raeside; comments to author 20.03.23; revised version received 06.07.23; accepted 11.01.24; published 28.02.24.

Please cite as:
Lightfoot M, Campbell C, Maragh-Bass AC, Jackson-Morgan J, Taylor K
What Adolescents Say in Text Messages to Motivate Peer Networks to Access Health Care and Sexually Transmitted Infection Testing: Qualitative Thematic Analysis
J Med Internet Res 2024;26:e44861
URL: https://www.jmir.org/2024/1/e44861
doi:10.2196/44861
PMID:38416541

XSL•FO
RenderX

©Marguerita Lightfoot, Chadwick Campbell, Allysha C Maragh-Bass, Joi Jackson-Morgan, Kelly Taylor. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Types of Social Support Exchanged by Survivors of Pediatric Stroke and Their Families in an Online Peer Support Community: Qualitative Thematic Analysis

William J A Wright[1], MB, BChir, MA; Charlotte Howdle[1]; Neil S Coulson[2], PhD; Anna De Simoni[1,3], PhD

[1]School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

[2]Medical School, Nottingham City Hospital, University of Nottingham, Nottingham, United Kingdom

[3]Centre for Primary Care, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Anna De Simoni, PhD
Centre for Primary Care
Wolfson Institute of Population Health
Queen Mary University of London
58 Turner Street
London, E1 2AB
United Kingdom
Phone: 44 882 2520
Fax: 44 882 2520
Email: a.desimoni@qmul.ac.uk

Abstract

Background: Pediatric stroke is relatively rare and underresearched, and there is little awareness of its occurrence in wider society. There is a paucity of literature on the effectiveness of interventions to improve rehabilitation and the services available to survivors. Access to online health communities through the internet may be a means of support for patients with pediatric stroke and their families during recovery; however, little research has been done in this area.

Objective: This study aims to identify the types of social support provided by an online peer support group to survivors of pediatric stroke and their families.

Methods: This was a qualitative thematic analysis of posts from a pediatric stroke population on a UK online stroke community active between 2004 and 2011. The population was split into 2 groups based on whether stroke survivors were aged ≤18 years or aged >18 years at the time of posting. The posts were read by 2 authors who used the adapted Social Support Behavior Code to analyze the types of social support exchanged.

Results: A total of 52 participants who experienced a pediatric stroke were identified, who posted a total of 425 messages to the community. About 41 survivors were aged ≤18 years at the time of posting and were written about by others (31/35 were mothers), while 11 were aged >18 years and were writing about themselves. Survivors and their families joined together in discussion threads. Support was offered and received by all participants, regardless of age. Of all 425 posts, 193 (45.4%) contained at least 1 instance of social support. All 5 types of social support were identified: informational, emotional, network, esteem support, and tangible aid. Informational and emotional support were most commonly exchanged. Emotional support was offered more often than informational support among participants aged ≤18 years at the time of posting; this finding was reversed in the group aged >18 years. Network support and esteem support were less commonly exchanged. Notably, the access subcategory of network support was not exchanged with the community. Tangible aid was the least commonly offered type of support. The exchanged social support provided insight into rehabilitation interventions and the unmet needs of pediatric stroke survivors.

Conclusions: We found evidence of engagement of childhood stroke survivors and their families in an online stroke community, with peer support being exchanged between both long- and short-term survivors of pediatric stroke. Engagement of long-term survivors of pediatric stroke through the online community was key, as they were able to offer informational support from lived experience. Further interventional research is needed to assess health and rehabilitation outcomes from engagement with online support groups. Research is also needed to ensure safe, nurturing online communities.

(J Med Internet Res 2024;26:e49440) doi:10.2196/49440

XSL•FO
RenderX

KEYWORDS

child; internet-based intervention; online health communities; peer support; qualitative analysis; rehabilitation; self-help group; self-help; social support; stroke; support groups; thematic analysis

Introduction

More than 400 children are diagnosed with stroke in the United Kingdom each year [1]. Despite this sizeable number, diagnosis is often a shock to parents because there is little awareness that strokes can affect children [2]. Recovery from pediatric stroke is a long process, with impairments lasting decades after the event [3]. Novel deficits may present many years after the stroke itself, and existing issues evolve and become more numerous as the survivor grows, with as many as three-quarters of families experiencing at least 1 unmet need after stroke [4]. Pediatric stroke is underresearched [5], and as a result, little is known and available to survivors and their families to support them during recovery [2,4,6,7]. Finally, during the recovery process, survivors and their families feel lonely and isolated from other members of their family, previous friends, peers, and other survivors [3]. Peer support groups may be well placed to alleviate these barriers by providing a space where users can share information, advice, and support each other.

A survey of internet access in the United Kingdom conducted in 2020 revealed that the internet is available in the homes of 93% of the population [8]. This provides an opportunity for individuals to connect online and has led to the formation of online peer-support groups for health care conditions. The advantages of having peer support on an online platform have been outlined by a number of studies [9-12]. These include transcending geographic or temporal boundaries to communication, allowing users to interact regardless of where they live and their time commitments. In addition, posts are written anonymously and asynchronously, allowing participants to speak frankly about their experience without fear of being recognized and, furthermore, spend time composing their posts rather than rushing into a response.

A systematic review found that the roles of online peer support groups for long term conditions consist of a shared social identity, learning from the experiences of others, fostering personal growth, and supporting others [13]. Emerging quantitative data shows that online social support groups may have a therapeutic benefit for some users [14], and a systematic review of social networking site interventions for health care conditions found that they had a positive effect on health-related behaviors [15]. A quantitative study on the use of an online peer support group for women with breast cancer found that the group membership had a moderate effect on reducing depression scores, perceived stress, and cancer-related trauma [16].

The role of in-person peer support groups for adult stroke survivors has been evaluated. A systematic review found they offered a platform for shared experience, social comparison, vicarious learning, and mutual gain [17]. Qualitative reviews additionally found that peer support groups for adult stroke were a place where people shared knowledge [18], felt they belonged [18,19], found a purpose in mentoring each other [18,19], and formed and maintained friendships outside the group setting [18,19]. A study looking at the uses of an online forums for stroke survivors found that users shared a story, requested information and/or support, and provided information and/or support [20]. Adult stroke survivors and their families received information and support on an individual basis from an online forum with 95% of user intentions being met [20].

There is a growing literature evaluating the types of peer support offered in a variety of patient groups with chronic conditions using the Social Support Behavior Code (SSBC) framework and the adapted SSBC [10,21,22]. The SSBC measures 5 types of social support: informational support, emotional support, esteem support (expressing respect and confidence in others), network support (sharing a feeling of belonging to the group), and tangible assistance (providing or offering help). Studies using the SSBC to look at the types of support offered on online forums found the same broad pattern. Emotional and informational support were most commonly exchanged. Esteem and network support were less common, and tangible assistance was rare [9,10,22-24]. Online health communities allow larger degree of personal disclosure, possibly due to anonymity when posting [10].

Exploring the support given by peers in an online stroke community could provide insight about unmet needs with regard to rehabilitation services. Indeed, the “Stroke in Childhood Clinical Guidelines,” published in May 2017, requested research that aimed to identify “rehabilitation interventions in line with the emerging evidence of motor, social, behavioral, and communication sequelae following stroke” [5]. This study aims to analyze whether online peer support was exchanged among survivors of pediatric stroke and what type of peer support was exchanged. A secondary aim is to examine whether time (number of years) since the pediatric stroke was related to engagement and the type of support offered.

Methods

Design

A qualitative analysis using the adapted SSBC framework [10] on posts by a pediatric stroke population in an online community.

Setting

The analysis used posts from the 2004-2011 Talkstroke online forum, a UK-based moderated online forum hosted by the Stroke Association website. The forum was set up as part of the charity website to facilitate communication between stroke survivors and caregivers. In total, the archives included 22,173 posts, written by 2583 unique usernames. A total of 58 participants who had a pediatric stroke were identified from the characterization of individual community users, as reported in a previous study [20]. We further excluded 2 users when analysis revealed their age at stroke was 18 years or older and 4 users because their age at the time of posting was unknown. A sample of 52 users remained, who produced a total of 425 posts that

XSL•FO
RenderX

were collected in an Excel (Microsoft Corporation) spreadsheet. Characteristics of stroke survivors, including demographics, employment, education, stroke type, initial impairments as well as impairments at the time of posting, support needs, and independence, were retrieved from within the posts in a previous study [20].

Data Analysis

WJAW read through all posts to become familiar with the data and patient narratives. Considering the potentially different perspectives of survivors who wrote as adults on their experience of rehabilitation from pediatric stroke and survivors who were written about by their parents, posts were split into 2 categories: whether the pediatric stroke survivor was older than 18 years or aged 18 years or younger at the time of posting. This assessed whether the time since the stroke was related to the amount and type of social support given. Rather than excluding posts that were not relevant (ie, did not include instances of social support), all posts were taken into account for the analysis to give an idea of how prevalent social support exchange was within the discussion threads. WJAW analyzed all posts, and CH and ADS independently coded a random 10% of posts. Posts were analyzed individually, not in the context of other posts in the forum thread. Deductive thematic analysis was applied, as described by Braun and Clarke [25]. The adapted SSBC (Textbox 1) was used as a framework to identify the types of social support provided through engagement with the online forums [10], and results have been reported according to the 5 types of social support: informational, emotional, esteem, network support, and tangible assistance. Coding was discussed with CH and ADS until agreement was reached.

Textbox 1. Social Support Behavior Code, from Coulson and Greenwood [10].

Support types and definitions

- Informational support
  - Advice: provides ideas or suggestions for action.
  - Referral: refers the recipient to other sources of information or help.
  - Situational appraisal: helps reassess or redefine the situation being faced by the recipient.
  - Teaching: offers detailed information, facts, or news.
- Emotional support
  - Relationship: conveys the importance of closeness.
  - Physical affection: offers physical contact, such as hugs and kisses.
  - Confidentiality: keeps the recipient's problem in confidence.
  - Sympathy: sorrow or regret for the situation faced by the recipient.
  - Understanding or empathy: expressions of understanding of the situation or discloses similar experience in a way that conveys understanding.
  - Encouragement: provides the recipient with hope and confidence.
  - Prayer: offers prayer for the recipient.
- Esteem support
  - Compliment: says positive things about the recipient.
  - Validation: provides agreement with the views of the recipient.
  - Relief of blame: alleviates any feelings of guilt the recipient has about the situation.
- Network support
  - Access: provides the recipient with access to new people.
  - Presence: offers to be there.
  - Companions: reminds the recipient that there are others who share similar experiences and are available.
- Tangible aid
  - Loan: lend money to the recipient.
  - Direct task: offers to do a direct task.
  - Indirect task: offers to take over a task from the recipient while they are stressed.
  - Active participation: offers to join the recipient in an activity.
  - Willingness: offers or expressions of willingness to help.

XSL•FO
RenderX

Patient and Public Involvement

A survivor of pediatric stroke aged 22 years (aged 0 years at the time of stroke and not a member of an online stroke community) was contacted after the initial analysis was completed and read the first draft of the results, providing insightful comments that helped to finalize the analysis and informed the discussion.

Ethical Considerations

The Stroke Association provided access to the archived forums and gave permission for the data to be used for this research purpose. The data from Talkstroke were stored and accessed through the University of Cambridge Clinical School Secure Data Hosting Service. Users of the forums had previously agreed that their data would become public upon registration within the forums, and there is consensus that internet data that are freely and publicly accessible can be used for research without needing ethics committee approval [20]. In order to protect the identity and intellectual property of forums participants, direct quotes have not been used, despite this being normal practice in qualitative research. Summative descriptions of quotes will instead be used throughout the paper. De Simoni et al [20] report a detailed description of the ethics linked to the research in the Talkstroke archives.

Results

Participants' Characteristics

A total of 41 survivors were aged ≤18 years at the time of participation, contributing a total of 273 posts; 11 survivors were aged >18 years and contributed 152 posts. Most survivors in the group aged ≤18 years took part in the community less than 1 year after their stroke, with the majority of content contributed indirectly through third-party users (31/35 were mothers, 89%). Content from the group aged >18 years was reported firsthand by adult survivors of pediatric stroke (Table 1). The time between stroke and participation in the online community for both groups ranged from 2 weeks to 46 years. Further information about our population can be found in Howdle et al [3].

XSL•FO
RenderX

Table 1. Characteristics of the online stroke community participants as identified in the posts.

Characteristics	Frequency
Participants overall	
Total number of participants, n	52
Age of survivor at time of stroke (years), median (range)	6 (0-17)
Posts	
Total number of posts, n	425
Number of posts per participant, median (range)	3 (1-56)
Identity of person posting, n	
Stroke survivor	17
Mother	31
Other (aunt, family friend, cousin, mother, and father)	5
Sex of stroke survivor, n	
Male	24
Female	26
Not stated	2
Time since stroke (years), n	
0-1	20
1-2	5
>2	27
Not stated	0
Participants aged ≤18 years at time of posting	
Total number of participants, n	41
Age of survivor at time of stroke (years), median (range)	4 (0-7)
Posts	
Total number of posts, n	273
Number of posts per participant, median (range)	3 (1-56)
Identity of person posting, n	
Stroke survivor	6
Mother	31
Other (aunt, family friend, cousin, mother, and father)	4
Sex of stroke survivor, n	
Male	19
Female	20
Not stated	2
Time since stroke (years), n	
0-1	20
1-2	5
>2	16
Not stated	0
Participants aged >18 years at time of posting	
Total number of participants, n	11
Age of survivor at time of stroke (years), median (range)	13 (0-17)
Posts	

XSL•FO
RenderX

Characteristics	Frequency
Total number of posts, n	152
Number of posts per participant, median (range)	13 (1-33)
Identity of person posting, n	
Stroke survivor	11
Mother	0
Other (aunt, family friend, cousin, mother, and father)	0
Sex of stroke survivor, n	
Male	5
Female	6
Not stated	0
Time since stroke (years), n	
0-1	0
1-2	0
>2	11
Not stated	0
Participants who stated they were...	
Holding a driving license	2
Not holding a driving license	2
At university or some university	4
In full-time employment	1
In part-time employment	1

Types of Support Requested by Participants

Posts written by participants at the start of a discussion thread followed a similar structure: an introduction followed by an account of their stroke and recovery up until the time of writing. Participants would then request help or ask for general or specific advice. Within the group aged ≤18 years, nearly half of the posts involved sharing their experience of pediatric stroke.

One parent wrote that they were not sure where to start and explained that their child was born with a congenital heart defect. They described the location of (stroke) damage in the child's brain. They explained that the progress the child has made has amazed everyone, describing the regaining of functions. The user then requested general help and advice as it was all new (p1, stroke aged 0, 0 years poststroke, mother).

Some posts in the group aged >18 years described their experience with pediatric stroke. In general, rather than sharing their whole story, they shared parts that were relevant to the topic of discussion within a thread. Their aim when participating typically seemed to be to give hope and motivation to other survivors.

A survivor aged older than 18 years wrote, first, to say they felt sorry for the original poster's loss. They then explained their own story about how their childhood stroke was missed. Since then, they have had many tests in the hope of finding a cause; however, the cause of the stroke was still unknown. They went on to complete their General Certificate of Secondary Education and A-level examinations and were currently studying at university. They finished by writing that strokes do happen in young people and about the need for more societal awareness. They then told the survivor that they too struggled to find anyone to help support their rehabilitation, however, reassuring them they were not alone and to not hesitate to ask any questions (p43, stroke aged 13, 7 years poststroke, survivor).

Reaching out for help occurred in posts by those aged ≤18 years and >18 years. These posts tended to involve asking for advice on dealing with a specific type of symptom.

A survivor aged ≤18 years wrote that her headaches were seriously affecting her everyday life. She followed the doctor's advice to rest, drink water, and take painkillers, but she found she was resting all the time, and it was affecting her sleep. She apologized for "moaning" and then asked for any ideas (p47, stroke aged 15, 1 year poststroke, survivor).

A user aged >18 years wrote that after 30 years, their eyes were "dizzy." They then subsequently asked if anyone had medication or other advice to stop this (p15, stroke aged 11, 30 years poststroke).

SSBC Analysis

About half of the posts contained instances of social support. It was common for posts to contain more than one form of support, for example, a user relating their situation to the recipient's to show understanding and then providing advice based on their experience of a similar situation.

XSL•FO
RenderX

Informational Support

All 4 subcategories of informational support were represented (advice, referral, situation appraisal, and teaching). Comparing the 2 age categories, informational support was proportionally offered more by the group aged >18 years than the group aged ≤18 years. Advice was offered on a range of topics. In the group aged ≤18 years, these included therapy types, disability allowance, travel, and not giving up and carving out time as a caregiver. Survivors aged >18 years also gave advice on therapy types and financial aid, but additionally discussed driving and how to be a good caregiver.

One user aged ≤18 years told another participant to look into claiming disability living allowance for children aged 16 years, saying it could save them 17% of the cost of alterations to their home (p6, stroke aged 0, 14 years poststroke, mother).

One user aged >18 years advised a caregiver to take their child out of their comfort zone and make them realize that there is more to life than sitting in a chair. They advised that this would motivate the survivor and be lots of fun (p52, stroke aged 17, 21 years poststroke, survivor).

In the referral subcategory, participants recommended research groups and organizations to each other for support and to aid financial claims.

One user wrote that one good organization for their child's one-sided weakness was "Hemihelp" (p3, stroke aged 0, 11 years poststroke, mother).

Situation appraisal was common among participants in the group aged ≤18 years. This occurred to help others understand that pediatric strokes occurred more commonly than widely perceived, that the recovery process was different in children compared to adults, and to normalize the grief that parents were feeling for their "lost child."

A user wrote not to be discouraged about hearing recovery stories of adult survivors of stroke, and that recovery for a child was different, as the plasticity of the brain is a childhood phenomenon (p12, stroke aged 1, 0 years poststroke, mother).

There were several instances of users teaching each other in both groups. Topics involved included how specific medications worked and their dosage, the meaning of medical tests, appointments, and the different types of strokes the disability support available and how to access it, and how to deal with the side effects of stroke.

One user aged >18 years advised another to start a diary to write appointments and other important things in. The user said that this piece of information was invaluable, as it would keep their mind on the stroke and not worrying about forgetting things (p52, stroke aged 17, 21 years poststroke, survivor).

A user explained what a magnetic resonance angiography scan is, saying that it is basically a magnetic resonance imaging scan, but that then they add a dye to show how well the blood is flowing through veins and arteries (p22, stroke aged 5, 4 years poststroke, mother).

Emotional Support

Emotional support was offered proportionally more by participants aged ≤18 years than >18 years. Support was offered to survivors of stroke in both age categories and to caregivers who wrote on behalf of the survivors. About 5 of the 7 subcategories were seen relationship, physical affection, confidentiality, sympathy, understanding or empathy, and encouragement.

Sympathy was the most displayed subcategory. Messages expressed sorrow for participants' stroke events, bad news since the stroke, bad experiences, ongoing symptoms, and the struggle to find a cause.

One user wrote that she was sorry to hear about another user's stroke (p4, stroke aged 0, 0 years poststroke, mother).

Understanding and empathy were displayed in the context of loneliness, fighting the urge to give up, acknowledging similar experiences, symptoms experienced, and difficulty finding travel insurance.

One user wrote they knew what another user meant when saying they felt alone as they had that most of their life. (p36, stroke aged 0, 35 years poststroke, survivor).

A user empathized with another user, writing that they too understood the challenges of dealing with a young child (with stroke) as a single parent (p12, stroke aged 1, 0 years poststroke, mother).

Messages of encouragement were present, often in the form of remarks like "good luck," "chin up," or "it gets easier over time." Most encouragement occurred in the context of helping people start their recovery journey, inspiring people not to give up, or wishing good fortune when attending health care institutions.

A health care worker who previously had a stroke as a child shared her story in the hope it would inspire and encourage people that life does not have to end following a stroke and that with determination you can get on and find other ways of doing things. They finished by stating not to let the stroke win (p39, stroke aged 8, 23 years poststroke, survivor).

The relationship subcategory describes the closeness felt between members of the online community. Users told others they did the best thing by joining the community and expressed joy in finding people who were going through similar experiences.

One user wrote how sorry they were to hear another person's story but that it was lovely to hear from them as it was difficult to find people in a similar situation. They wanted to be kept up to date about how the other survivors were getting on (p2, stroke aged 0, 2 years poststroke, mother).

Physical affection was displayed through affectionate language and symbols. The most common symbol used was an *x* to send kisses.

One user wrote that they hoped another's child was well and that their heart truly went to their family, then signed off with their name (p11, stroke aged 1, 1 year poststroke, mother).

XSL•FO
RenderX

Another user wished someone a happy new year, followed by 3 kisses (p49, stroke aged 17, 2 years poststroke, survivor).

Network Support

Around 2 out of 3 subcategories of network support were identified within the data: presence and companions. The most common incidences of presence, a member offering to be there for another member, were related to 2 users finding themselves in a similar position. This prompted them to stay in touch, especially when somebody was offering advice, to see if it worked. Adult survivors of childhood stroke also offered to keep in touch with younger participants in the long term, to keep answering questions related to their recovery.

A survivor aged >18 years wrote that a user could contact them with any questions and they would do their best to answer them (p49, stroke aged 17, 2 years poststroke, survivor).

A mother of a survivor wrote to another to stay in touch; a few mothers had found each other, and it did help (p12, stroke aged 1, 0 years poststroke, mother).

The community offered a lot in the way of companionship. Users reminded each other that everyone on the site had been through similar experiences and were all there to support each other. Users were often complimentary about others in the community.

One user stated that if another felt down, then they could just talk to those on the site, as everyone was there for each other (p36, stroke aged 0, 35 years poststroke, survivor).

A user prompted another to keep posting because the people on the site were brilliant (p26, stroke aged 9, 0 years poststroke, mother).

Esteem Support

All subcategories of esteem support were present: compliments, validation, and relief of blame.

Compliments were exchanged to thank people for advice and recovery ideas, for sharing their stories, to motivate caregivers, and to congratulate people for joining the community.

One user wrote to another that they admired their confidence and how they wanted to thank them for always being on the site with advice and support (p26, stroke aged 9, 0 years poststroke, mother).

Another user complimented another by saying they thought the rehabilitation measures they had in place were a brilliant idea (p22, stroke aged 5, 4 years poststroke, mother).

Validation was given when survivors had been through similar experiences. Notable themes within validation were missed or delayed diagnoses, poor awareness of pediatric stroke, isolation, feeling tired, and wanting live chatrooms.

One user agreed with another about the difficulty of convincing doctors to do investigations for stroke. They said their child had all the signs of stroke, but they only obtained a computed tomography scan by pushing the consultant hard, who then agreed to get a radiologist out of bed (p12, stroke aged 1, 0 years poststroke, mother).

Another user agreed that previous friends found it difficult to comprehend how difficult recovery was for stroke survivors (p28, stroke aged 9, 2 years poststroke, parents).

There was a single account of relief of blame, reassuring a survivor that they were not a burden and that everyone needs support at some point (p12, stroke aged 1, 0 years poststroke, mother).

Tangible Aid

Tangible aid was the least common type of social support offered. Loans, direct tasks, indirect tasks, and active participation were not identified within the data set. However, members displayed a willingness to help wherever possible.

One user wrote that they would be more than happy to help if required (p49, 17 years, 4 years poststroke, survivor).

Patient and Public Involvement

The analysis was read by a survivor of multiple childhood strokes while aged <1 year, who is now studying at university. She commented that she and her family had never had any experience with peer support, whether in person or online; support had only been provided by medical professionals. However, she thought peer support would have been useful for her and her family. She especially valued the teaching and situational appraisal subcategories in informational support, emotional support, and the validation subcategory in esteem support. Some extracts of her comments are reported in Textbox 2.

XSL•FO
RenderX

Textbox 2. Patient and Public Involvement representative's specific comments on the study results.

Regarding situational appraisal, the survivor said:

- "I know my family were told this a lot while we were waiting to see if my stroke at <0 years would affect me in my first decade. It was emphasised to be aware of the difference between childhood and adult stroke and not to compare the speed of recovery."

Regarding teaching, the survivor said:

- "I am sure this support would have been invaluable to my family if these forums existed when I was born. Instead informational support was given by doctors only."

Regarding emotional, the survivor said:

- "Again I'm sure this emotional support would have been of great help especially in the uncertain times following a stroke, especially since my family were unaware and unable to connect with any of other families who had experienced pediatric stroke. Additionally, emotional support would have been a great help during the early years when there was uncertainty about the future implications of the stroke as it's apparently hard to judge the extent of the effects, they wait to see if the child walks, talks or can communicate and learn at school."

Validation with regards to medical professionals being less experienced with diagnosing pediatric strokes:

- "I'm sure my family can sympathise with this, it took one very persistent doctor/nurse who was adamant I was having a stroke to investigate further."

Discussion

Overview

This study provides qualitative evidence that online peer support, facilitated by the online community, played an important role in meeting the rehabilitation needs of patients with pediatric stroke. Most posts analyzed displayed social support. Discussion threads engaged a range of people: those who were recently recovering from a pediatric stroke, their caregivers and families, and adult survivors of pediatric stroke. Users appreciated finding others they could share their lived experience of stroke with and relate to similar incidences, particularly regarding the difficulty in getting a diagnosis and rehabilitation. They additionally discussed both the lack of specific local support and available resources and support services. All 5 categories of social support in the SSBC were evidenced; emotional support was the most common, followed by informational support. The group aged ≤18 years provided more emotional support than informational support; however, the group aged >18 years provided more informational support than emotional support. This difference may arise as members of the group aged >18 years engage with the community in order to provide informational support based on their own lived experience of decades of rehabilitation. There were instances of network support and esteem support, and tangible support was exchanged the least. There were no instances of subcategory access support. It was noticeable that survivors in the group aged >18 years displayed empathy less often than survivors in the group aged ≤18 years.

There are 3 strengths to this study. First, the online community facilitated communication and discussion between participants who recently had a stroke and users who had a stroke many years before. This brought unprecedented insight into the long-term lived experiences of pediatric stroke recovery and survivors' unmet needs. Second, the methodological design of analyzing posts on a forums meant that messages were viewed and assessed directly by the researcher. This is a nontraditional method that helps to avoid potential problems such as retrospective self-reports, recall bias, and researchers' bias. Third, the population that uses the forums might include people who do not partake in traditional research studies, allowing the needs of an underrepresented patient population to be studied [11].

Limitations of the study include the analysis of a single UK-based online peer-support community dated 2004-2011, decreasing the generalizability of the study. Additionally, although messages imply that users appreciated support from others, the impact of the support given was hard to measure; performing interviews with participants would have been more successful in measuring the effect of comments on users. Moreover, the study population may not be representative of survivors who do not have access to a computer, do not know how to use the online forums, and therefore may not be able to participate in the online community, resulting in a potential patient population being missed from this study [16]. Further to this, participants who were adult survivors needed to be able to read and type, abilities sometimes affected by stroke and potentially acting as barriers to participation. This study was not set to explore the drawbacks of using online forums as support for health conditions, as posts were analyzed individually and not in the context of the thread to which they belonged. This meant we could not analyze how a post affected other users taking part in that thread. Nevertheless, a previous qualitative study of the same stroke online community found that users would promptly counter inappropriate medical information or health behavior [20].

The ability to access support is a form of social medical capital, defined as the "advantages that any user (patient or caregiver) can gain from participation in the social networks provided by online health communities" [26]. This confirms findings from similar studies across a variety of chronic conditions [9,10,23].

Only half of the posts analyzed in the study contained instances of social support. This was much lower than previous studies, which found rates between 83.8% and 98.9% [9,24]. This could be for a number of reasons: there is a paucity of literature for

XSL•FO
RenderX

survivors of pediatric stroke to access, resulting in families and survivors having many questions during their recovery. It is possible that participants were asking more questions than they were able to answer. Also, the forum was for survivors of both adult and pediatric stroke, and it is possible that the adult stroke survivors provided support, which was not included in this study sample. Similar to other studies, informational and emotional support were the 2 most commonly exchanged categories of support; network and esteem support were the next most common, with tangible aid being the least common form of support received [9,10,22-24]. This may be reflective of the actual online platform as the means for providing support, where the geographical limitations of the group mean that the ability to perform tasks to help others is limited.

Informational support involved participants giving each other practical and recovery advice. In addition, participants taught each other strategies for dealing with things referred them to resources and helped them appraise situations. Emotional support was commonly offered; participants exhibited affection for each other, and the forums appeared to provide a safe space to share experiences where users could sympathize, empathize, or encourage each other. Network support was present on the forums; there were many messages that expressed a sense of comradery, reiterating that members of the site were there to support each other. This could be particularly useful in the context of pediatric stroke, where the incidence is relatively low and the chances of meeting another survivor who lives close by are low. Esteem support was provided to thank people for their contribution to the site as well as validate people's recovery processes.

Notably, there were no examples of the access subcategory. Access is defined as providing "the recipient with access to new people." This is contrary to other studies [9,10,22,23]. This may be another illustration of the low incidence of pediatric stroke, and so not many survivors and families are well known. In addition, medical and public awareness of pediatric stroke in society is low [2,3,7] and as a result, there are very few people and sources of literature available for survivors. There were no examples of confidentiality, a finding found in other similar studies on online forums; this was as expected as posts were made on a public platform [9,10].

A new finding from this study is that posts made by users whose child was aged ≤18 years at the time of posting offered emotional support more commonly than informational support. However, survivors who were aged >18 years at the time of posting gave more informational support than emotional support. This finding can be interpreted in the context of the theory of optimal matching [27], which hypothesizes that specific types of social support may be beneficial in aiding specific types of stress. According to this theory, the controllability of a situation plays an important role in determining what kind of social support will be most beneficial to the individual. Informational support is a type of action-facilitating support that fosters behaviors designed to mitigate a stressor. Emotional support is a nurturing support that helps individuals cope with the emotional consequences of a stressor [22]. It proposes that individuals with controllable problems should benefit most from informational support because they can use this information, advice, and guidance to help them deal with the cause of their difficulties. However, those with uncontrollable problems should benefit more from emotional support because this will help them cope with unpleasant emotions and the stressful negative effects of being in an uncontrollable situation. In the context of pediatric stroke, survivors who are closer to their stroke event may perceive to have less control over events in their life and so offer emotional support. In contrast, survivors aged >18 years may view their stroke event as more controllable and therefore may have perceived that providing informational support from decades of lived experience was more useful. In addition, socioemotional selectivity theory [28] suggests that the degree to which an illness is relatively more chronic or acute could influence the support seeker's time perspective, whereby chronic illness facilitates problem-focused coping that favors action-facilitating types of support [22]. Pediatric stroke is an illness with both an acute and chronic phase. As predicted by the social-emotional selectivity theory, emotional support is more prevalent in the group aged ≤18 years coping with acute illness, whereas informational support is more common in the group aged >18 years.

Conclusions

This study brings qualitative evidence that a nation-wide online peer support group was beneficial in drawing support to survivors of pediatric stroke and their families. The community also represented an opportunity for adult survivors of pediatric stroke to validate their experiences many decades post stroke, share information gained through their rehabilitation journeys, provide insight about unmet needs with regard to rehabilitation services, and provide hope to families with more acute stroke incidences. The study additionally highlights that recovery from stroke is a long process, with adult survivors reaching out for advice. More research is needed, in particular interventional research studies, to evaluate the effectiveness of online peer support groups for survivors of pediatric stroke, as well as research to ensure online communities are safe and nurtured.

Acknowledgments

The authors are grateful to the Stroke Association for sharing the archive file of Talkstroke Online with us. We thank our patient and public involvement representative for her comments on the manuscript. ADS was partly funded by a National Institute for Health and Care Research (NIHR) Programme Grant for Applied Research (NIHR202037). Views are those of the authors and not necessarily those of the National Health Service, NIHR, or the Department of Health and Social Care.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. What is childhood stroke? Stroke Association. URL: https://www.stroke.org.uk/childhood-stroke/about-childhood-stroke [accessed 2012-03-23]
2. Soufi S, Chabrier S, Bertoletti L, Laporte S, Darteyre S. Lived experience of having a child with stroke: a qualitative study. Eur J Paediatr Neurol 2017;21(3):542-548. [doi: 10.1016/j.ejpn.2017.01.007] [Medline: 28185801]
3. Howdle C, Wright W, Mant J, De Simoni A. JMIR 2023 submitted to JMIR ms#49409. NA 2023.
4. Gordon AL, Nguyen L, Panton A, Mallick AA, Ganesan V, Wraige E, et al. Self-reported needs after pediatric stroke. Eur J Paediatr Neurol 2018;22(5):791-796. [doi: 10.1016/j.ejpn.2018.06.003] [Medline: 29960841]
5. Stroke in childhood. Clinical guidelines for diagnosis, management and rehabilitation. Royal College of Paediatrics and Child Health. 2017. URL: https://www.rcpch.ac.uk/sites/default/files/2021-02/Stroke%20guideline%2008.04.19%20updated%202021.pdf [accessed 2023-12-03]
6. Denham AMJ, Wynne O, Baker AL, Spratt NJ, Turner A, Magin P, et al. "This is our life now. Our new normal": a qualitative study of the unmet needs of carers of stroke survivors. PLoS One 2019;14(5):e0216682 [FREE Full text] [doi: 10.1371/journal.pone.0216682] [Medline: 31067286]
7. McKevitt C, Topor M, Panton A, Mallick AA, Ganesan V, Wraige E, et al. Seeking normality: parents' experiences of childhood stroke. Child Care Health Dev 2019;45(1):89-95. [doi: 10.1111/cch.12622] [Medline: 30255632]
8. A review of Ofcom's research on digital exclusion among adults in the UK. Chapter 4. How many people in the UK are digitally excluded? Ofcom. Digital Exclusion. 2022. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/234364/digital-exclusion-review-2022.pdf [accessed 2023-04-23]
9. Coulson NS, Buchanan H, Aubeeluck A. Social support in cyberspace: a content analysis of communication within a huntington's disease online support group. Patient Educ Couns 2007;68(2):173-178 [FREE Full text] [doi: 10.1016/j.pec.2007.06.002] [Medline: 17629440]
10. Coulson NS, Greenwood N. Families affected by childhood cancer: an analysis of the provision of social support within online support groups. Child Care Health Dev 2012;38(6):870-877. [doi: 10.1111/j.1365-2214.2011.01316.x] [Medline: 21916931]
11. De Simoni A, Shanks A, Mant J, Skelton JR. Making sense of patients' internet forums: a systematic method using discourse analysis. Br J Gen Pract 2014;64(620):e178-e180 [FREE Full text] [doi: 10.3399/bjgp14X677671] [Medline: 24567657]
12. Jamison J, Sutton S, Mant J, De Simoni A. Online stroke forum as source of data for qualitative research: insights from a comparison with patients' interviews. BMJ Open 2018;8(3):e020133 [FREE Full text] [doi: 10.1136/bmjopen-2017-020133] [Medline: 29602848]
13. Shilling V, Morris C, Thompson-Coon J, Ukoumunne O, Rogers M, Logan S. Peer support for parents of children with chronic disabling conditions: a systematic review of quantitative and qualitative studies. Dev Med Child Neurol 2013;55(7):602-609 [FREE Full text] [doi: 10.1111/dmcn.12091] [Medline: 23421818]
14. Coulson NS, Bullock E, Rodham K. Exploring the therapeutic affordances of self-harm online support communities: an online survey of members. JMIR Ment Health 2017;4(4):e44 [FREE Full text] [doi: 10.2196/mental.8084] [Medline: 29030324]
15. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015;22(1):243-256 [FREE Full text] [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
16. Winzelberg AJ, Classen C, Alpers GW, Roberts H, Koopman C, Adams RE, et al. Evaluation of an internet support group for women with primary breast cancer. Cancer 2003;97(5):1164-1173 [FREE Full text] [doi: 10.1002/cncr.11174] [Medline: 12599221]
17. Clark E, MacCrosain A, Ward NS, Jones F. The key features and role of peer support within group self-management interventions for stroke? A systematic review. Disabil Rehabil 2020;42(3):307-316. [doi: 10.1080/09638288.2018.1498544] [Medline: 30325686]
18. Wijekoon S, Wilson W, Gowan N, Ferreira L, Phadke C, Udler E, et al. Experiences of occupational performance in survivors of stroke attending peer support groups. Can J Occup Ther 2020;87(3):173-181. [doi: 10.1177/0008417420905707] [Medline: 32115988]
19. Christensen ER, Golden SL, Gesell SB. Perceived benefits of peer support groups for stroke survivors and caregivers in Rural North Carolina. N C Med J 2019;80(3):143-148 [FREE Full text] [doi: 10.18043/ncm.80.3.143] [Medline: 31072940]
20. De Simoni A, Shanks A, Balasooriya-Smeekens C, Mant J. Stroke survivors and their families receive information and support on an individual basis from an online forum: descriptive analysis of a population of 2348 patients and qualitative study of a sample of participants. BMJ Open 2016;6(4):e010501 [FREE Full text] [doi: 10.1136/bmjopen-2015-010501] [Medline: 27053271]

XSL•FO
RenderX

21. Cutrona CE, Suhr JA. Controllability of stressful events and satisfaction with spouse support behaviours. Commun Res 1992;19:154-174 [FREE Full text] [doi: 10.1177/009365092019002002]
22. Rains SA, Peterson EB, Wright KB. Communicating social support in computer-mediated contexts: a meta-analytic review of content analyses examining support messages shared online among individuals coping with illness. Commun Monogr 2015;82(4):403-430. [doi: 10.1080/03637751.2015.1019530]
23. Mo PKH, Coulson NS. Exploring the communication of social support within virtual communities: a content analysis of messages posted to an online HIV/AIDS support group. Cyberpsychol Behav 2008;11(3):371-374. [doi: 10.1089/cpb.2007.0118] [Medline: 18537512]
24. Smedley R, Coulson N, Gavin J, Rodham K, Watts L. Online social support for complex regional pain syndrome: a content analysis of support exchanges within a newly launched discussion forum. Computers in Human Behavior 2015;51:53-63 [FREE Full text] [doi: 10.1016/j.chb.2015.04.040]
25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Panzarasa P, Griffiths CJ, Sastry N, De Simoni A. Social medical capital: how patients and caregivers can benefit from online social interactions. J Med Internet Res 2020;22(7):e16337 [FREE Full text] [doi: 10.2196/16337] [Medline: 32720910]
27. Cutrona CE, Russell DW. Type of social support and specific stress: toward a theory of optimal matching. In: Sarason BR, Sarason IG, Pierce GR, editors. Social Support: Aninteractional View. Oxford: John Wiley & Sons; 1992:319-366.
28. Carstensen LL. Evidence for a life-span theory of socioemotional selectivity. Curr Dir Psychol Sci 1995 Oct;4(5):151-156. [doi: 10.1111/1467-8721.ep11512261]

Abbreviations

NIHR: National Institute for Health and Care Research
SSBC: Social Support Behavior Code

Edited by T Leung; submitted 29.05.23; peer-reviewed by S Atanasova, D Riggins; comments to author 19.09.23; revised version received 07.11.23; accepted 29.01.24; published 15.03.24.

Please cite as:
Wright WJA, Howdle C, Coulson NS, De Simoni A
Exploring the Types of Social Support Exchanged by Survivors of Pediatric Stroke and Their Families in an Online Peer Support Community: Qualitative Thematic Analysis
J Med Internet Res 2024;26:e49440
URL: https://www.jmir.org/2024/1/e49440
doi:10.2196/49440
PMID:38488858

©William J A Wright, Charlotte Howdle, Neil S Coulson, Anna De Simoni. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Influencing Recovery From Pediatric Stroke Based on Discussions From a UK-Based Online Stroke Community: Qualitative Thematic Study

Charlotte Howdle[1]; William James Alexander Wright[1], BChir, MB, MA; Jonathan Mant[2], MBBS, MA, MSc, MD; Anna De Simoni[2,3], MBBS, PhD, MRCGP

[1]School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

[2]Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

[3]Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Anna De Simoni, MBBS, PhD, MRCGP
Wolfson Institute of Population Health
Queen Mary University of London
58 Turner Street
Centre for Primary Care
London, E1 2AB
United Kingdom
Phone: 44 207882 ext 252
Email: a.desimoni@qmul.ac.uk

Abstract

Background: The incidence of stroke in children is low, and pediatric stroke rehabilitation services are less developed than adult ones. Survivors of pediatric stroke have a long poststroke life expectancy and therefore have the potential to experience impairments from their stroke for many years. However, there are relatively few studies characterizing these impairments and what factors facilitate or counteract recovery.

Objective: This study aims to characterize the main barriers to and facilitators of recovery from pediatric stroke. A secondary aim was to explore whether these factors last into adulthood, whether they change, or if new factors impacting recovery emerge in adulthood.

Methods: We performed a qualitative thematic analysis based on posts from a population of participants from a UK-based online stroke community, active between 2004 and 2011. The analysis focused on users who talked about their experiences with pediatric stroke, as identified by a previous study. The posts were read by 3 authors, and factors influencing recovery from pediatric stroke were mapped into 4 areas: medical, physical, emotional, and social. Factors influencing recovery were divided into short-term and long-term factors.

Results: There were 425 posts relating to 52 survivors of pediatric stroke. Some survivors of stroke posted for themselves, while others were talked about by a third party (mostly parents; 31/35, 89% mothers). In total, 79% (41/52) of survivors of stroke were aged ≤18 years and 21% (11/52) were aged >18 years at the time of posting. Medical factors included comorbidities as a barrier to recovery. Medical interventions, such as speech and language therapy and physiotherapy, were also deemed useful. Exercise, particularly swimming, was deemed a facilitator. Among physical factors, fatigue and chronic pain could persist decades after a stroke, with both reported as a barrier to feeling fully recovered. Tiredness could worsen existing stroke-related impairments. Other long-standing impairments were memory loss, confusion, and dizziness. Among emotional factors, fear and uncertainty were short-term barriers, while positivity was a major facilitator in both short- and long-term recovery. Anxiety, grief, and behavioral problems hindered recovery. The social barriers were loneliness, exclusion, and hidden disabilities not being acknowledged by third parties. A good support network and third-party support facilitated recovery. Educational services were important in reintegrating survivors into society. Participants reported that worrying about losing financial support, such as disability allowances, and difficulties in obtaining travel insurance and driving licenses impacted recovery.

Conclusions: The lived experience of survivors of pediatric stroke includes long-term hidden disabilities and barriers to rehabilitation. These are present in different settings, such as health care, schools, workplaces, and driving centers. Greater awareness of these issues by relevant professional groups may help ameliorate them.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e49409) doi:10.2196/49409

KEYWORDS

child; stroke; rehabilitation; barriers; facilitators; internet-based intervention; self-help group; thematic analysis

Introduction

A pediatric stroke is classified as either perinatal (occurring when the child is 20 weeks in gestation to 28 days after birth) or childhood (from 29 days after birth to 18 years). There are >400 cases of pediatric stroke in the United Kingdom annually [1]. Pediatric stroke can be a debilitating disease, leaving survivors and their families coping with persisting issues during their recovery. These problems evolve and become more numerous as the survivor grows up, with as many as 75% of families of a child with stroke having at least 1 unmet need [2].

There is limited literature on factors that impact recovery. Survivors of pediatric stroke often develop comorbidities. Studies have found that survivors have an increased incidence of the following: epilepsy [3-5], attention-deficit/hyperactivity disorder [3,5], depression [5], anxiety [4,5], and autism [6,7], compared with their nonstroke counterparts. Fatigue [8] and behavioral problems [4,9] also hinder recovery. Mental health support [10] and follow-up neuropsychological assessments [10,11] positively influence outcomes. Qualitative studies with parents revealed that many disabilities experienced by stroke survivors are *hidden* [9], resulting in them not receiving the support they need.

There are also emotional barriers that impact recovery. Pediatric stroke negatively affects the family of a survivor, with reports of poor parental mental health [4,11], guilt [4,12,13], and uncertainty about the future [12]. Having good parental welfare is important—parents are the primary caregivers following their child's stroke, and their well-being impacts their child's recovery. Good family functioning [4] also facilitates recovery from childhood stroke. Parents reported that treating the child as "normal" aided recovery [14]. A supportive social network, from close family and friends [14] to the wider society [13], is important for both parents and children to facilitate recovery.

Certain social factors are known to impact recovery. Qualitative reviews asking about patients' experiences of health care showed that after discharge, parents felt abandoned by professionals [9,12]. Health care services may not be flexible, they may be difficult to access, and parents may not know where they can access therapy for their child [12]. When health care staff could be approached, parents may not have known what questions to ask [15]. Clear communication with parents by medical professionals about the causes of the stroke and events occurring around the stroke has been identified as an important issue [4]. Patients have appreciated positivity from clinicians [14], close and ongoing medical support [14,15], involvement with goal-setting approaches [16,17], and continuity of care. There is limited medical awareness or literature to support parents during their child's recovery from stroke [2,9,12]. Support from charities and both in-person and online community groups partly addresses this issue [4,13,15]. Accessing disability aid also appears to facilitate recovery from childhood stroke [15].

A growing body of literature supports the potential of online health communities to provide opportunities for individuals to share their personal experiences and learn from others with similar conditions [18,19]. These communities also serve as a reliable and novel source of information about patients' unmet needs [20,21], complementing more traditional research methods [22,23].

There are an increasing number of studies that explore the barriers to and facilitators of recovery from pediatric stroke, considering that these factors may be numerous and long lasting as children have a long poststroke life expectancy. However, there are relatively few studies that characterize the long-term impact of stroke on children; currently, the longest time frame studied is 5 years after stroke [24]. In May 2017, the Royal College of Paediatrics and Child Health in the United Kingdom published *The Stroke in Childhood Clinical Guidelines*, which had a set of "research recommendations" that included "reviewing the complications that children experience" following pediatric stroke to assess the "rehabilitation needs of pediatric stroke patients" and evaluate the "long-term outcomes" of survivors of pediatric stroke [25]. This study aims to address these recommendations by characterizing what survivors of pediatric stroke report as the main barriers and facilitators to their recovery in a UK-based online stroke community. As a secondary aim, the study also explores whether these factors last into adulthood or not, whether they change, or if new factors impacting recovery emerge in adulthood.

Methods

Study Design

We conducted a qualitative thematic analysis on posts from a pediatric stroke population within a UK-based online stroke community.

Setting

The analysis used the archived TalkStroke online community, a UK-based, moderated online community hosted on the Stroke Association website from 2004 to 2011. In total, the TalkStroke archive contains 22,173 posts written by 2583 unique usernames. A previous study by De Simoni et al [18] identified 58 usernames that posted about experiences of pediatric stroke, contributing to a total of 469 posts. We excluded some users: 2 after further analysis revealed that their age at stroke was >18 years and a further 4 users because their age at the time of posting was unknown. A sample of 52 users remained. The characteristics of the survivors of pediatric stroke, including demographics, employment, education, stroke type, initial impairments as well as impairments at the time of posting, support needs, and independence, were retrieved from the data set of a previous study [18].

XSL•FO
RenderX

Data Analysis

A deductive approach was used to develop predetermined themes before the start of the analysis. Factors (barriers or facilitators) were split into 4 themes based on recommendations by the Clinical Guidelines for Stroke in Childhood [25]: medical, physical, emotional, and social.

All relevant posts were collected in an Excel (Microsoft Inc) spreadsheet. WJAW and CH read all the posts to become familiar with the data. Themes were further split into subcategories using a data-driven approach by applying thematic analysis, as described by Braun and Clarke [26]. Coding was discussed until agreement was reached and a final coding framework was agreed upon. Each individual post was considered on its own, outside the context of the thread to which the posts belonged.

To consider the potentially different perspectives of participants over time, the posts were split into 2 categories: whether the survivor of pediatric stroke was aged ≤18 years or >18 years at the time of posting. This provided an assessment of whether the factors impacting recovery were short-term or long-term factors. We defined "long-term factors" as those affecting participants into adulthood and "short-term factors" as those impacting patients from the time immediately following a stroke until they reach 18 years of age. Sometimes, third parties (ie, parents or a member of the wider family) wrote on behalf of the survivor of pediatric stroke.

In this analysis, the term recovery describes the improvement of any aspect of stroke-related impairment. Rehabilitation is defined in concordance with the definition of the World Health Organization: "interventions designed to optimize functioning and reduce disability in individuals with health conditions in interaction with their environment" [27].

Patient and Public Involvement

A 22-year-old survivor of pediatric stroke (when aged 0 years) was contacted after the analysis was completed. The patient and public involvement (PPI) participant was an acquaintance of a coauthor's friend and was approached informally. We obtained written consent for her contribution to the work to be included in the paper.

The initial results were written and sent to the PPI participant, who read and provided insightful comments that highlighted important subcategories and informed our *Discussion* section.

Ethical Considerations

The Stroke Association collected the data from the archived TalkStroke forum and provided the data set, together with their permission, for analysis for research purposes. The data are stored through the University of Cambridge Clinical School's Secure Data Hosting Service, with reference S0126—Stroke Needs & Exp. This analysis was assessed by the University of Cambridge institutional review board and exempted from ethics approval, provided that permission to use the data was granted by the Stroke Association. At the time of registering with the community, users agreed that their data would be public. However, to protect the identity and intellectual property of participants, this analysis does not use direct quotes; instead, quotes are paraphrased. A more detailed account of the ethics involved with research on the TalkStroke archives is described in the study by De Simoni et al [18].

Results

Participants' Characteristics

In total, 79% (41/52) of users were aged ≤18 years at the time of participation, contributing a total of 273 posts; 21% (11/52) of users were aged >18 years and contributed 152 posts. The majority of data from the ≤18-year-old group were collected through third-party users (35/41, 85%). Of these, most were written by the mother of the survivor (31/35, 89%). Data from the >18-year-old group were reported firsthand by adult survivors of pediatric stroke (Table 1).

Among the 11 participants who were aged >18 years at the time of posting, 2 (18%) held a driving license, 1 (9%) was considering applying for one, and 1 (9%) stated they do not drive. In total, 4 (36%) participants were university graduates or attending university, 1 (9%) was in part-time employment, and 1 (9%) was in full-time employment.

XSL•FO
RenderX

Table 1. Characteristics of the TalkStroke online community participants as identified in the posts.

Sample characteristics	All participants (n=52)	Participants aged ≤18 y (n=41)[a]	Participants aged >18 y (n=11)[a]
Age (y) at stroke (if survivor of stroke)			
Mean	7.3	6	12
Median (range)	6 (0-17)	4 (0-17)	13 (0-17)
Total number of posts, n	425	273	152
Number of posts per participant			
Mean	8.2	6.7	13.8
Median (range)	3 (1-56)	3 (1-56)	13 (1-33)
Identity of the person posting, n (%)			
Survivor of stroke	17 (33)	6 (15)	11 (100)
Mother	31 (60)	31 (76)	0 (0)
Other (aunt, family friend, or cousin)	4 (8)	4 (10)	0 (0)
Sex of the survivor of stroke, n (%)			
Male	24 (46)	19 (46)	5 (45)
Female	26 (50)	20 (49)	6 (55)
Not stated	2 (4)	2 (5)	0 (0)
Time since stroke (y), n (%)			
0<x[b]<1	20 (38)	20 (49)	0 (0)
1≤x<2	5 (0)	5 (12)	0 (0)
>2	27 (52)	16 (39)	11 (100)
Not stated	0 (0)	0 (0)	0 (0)
Participants who stated they were, n (%)			
Holding a driving license	2 (18)	N/A[c]	2 (18)
Not holding a driving license	2 (18)	N/A	2 (18)
At university or some university	4 (36)	N/A	4 (36)
In full-time employment	1 (9)	N/A	1 (9)
In part-time employment	1 (9)	N/A	1 (9)

[a]Age at time of participation in the online stroke community.

[b]x: time or age between the 2 values.

[c]N/A: not applicable.

Causes of Stroke and Resulting Impairments

Of the 52 survivors, 13 (25%) were survivors of right-sided strokes, 22 (42%) were survivors of left-sided strokes, 2 (4%) were survivors of stroke affecting both sides, and 15 (29%) were not reported. The causes of stroke were reported for a small number of participants: 4 (8%) were after surgery (3/4, 75% was due to cardiac operations and 1/4, 25% was unknown); 6 (11%) were after infection (1/6, 17% was meningitis; 3/6, 50% was chicken pox; 1/6, 17% was herpes encephalopathy; and 1/6, 17% was maternal shingles during gestation); 2 (4%) were ischemic; 3 (6%) were hemorrhagic; 1 (2%) was linked to acute lymphoblastic leukemia treatment; 2 (4%) were dissections; 1 (2%) was linked to a brain tumor; 1 (2%) was linked to arterial stenosis; 1 (2%) was linked to a septal defect; 1 (2%) was linked to an oral contraceptive pill; and 30 (58%) were not reported.

The initial impairments after stroke included the following: hemiplegia, hemiparesis, poststroke epilepsy, visual disturbances, tiredness, increased emotions, dystonia, dysarthria, facial droop, memory impairments, headaches, and no impairments. The difference between initial impairments and residual impairments at the time of posting varied greatly for users. The time between stroke and participation in the community ranged from 2 weeks to 46 years.

Themes

The subcategories of the 4 themes are presented in Table 2, which shows whether the factors are short term, mostly reported

XSL•FO
RenderX

by parents discussing their children with stroke, or long term, as reported by adult survivors of pediatric strokes.

Table 2. Themes and subcategories of factors impacting pediatric stroke.

Theme	Barrier	Facilitator	Short term	Long term
Medical factors				
Comorbidities	✓		✓	✓
Medical interventions		✓	✓	
Physical factors				
Fatigue	✓		✓	✓
Pain	✓		✓	✓
Neurological sequelae	✓		✓	✓
Emotional factors				
Positivity		✓	✓	✓
Grief and bereavement	✓		✓	✓
Fear	✓		✓	
Anxiety	✓		✓	✓
Behavioral problems	✓		✓	✓
Social factors				
Support network		✓	✓	✓
Loneliness	✓		✓	✓
Exclusion or bullying	✓		✓	
Hidden disabilities	✓			✓
Education	✓	✓	✓	
Driving	✓		✓	
Travel	✓		✓	
Third-party support	✓		✓	✓
Financial support	✓	✓	✓	✓

Medical Factors

Comorbidities

Epilepsy and depression were most commonly mentioned. Parents found it tough to cope with the additional diagnosis of epilepsy alongside pediatric stroke, with some reporting seeing their children getting more ill rather than better:

> *A parent writes they were completely shattered as their child had already suffered so much.* [Mother; participant 30; ≤18-year-old group; stroke at the age of 11 years; 2 years since the stroke]
>
> *One survivor wrote their depression sets them in a really low state of mind, where they cannot control their emotion.* [Survivor; participant 47; ≤18-year-old group; stroke at the age of 15 years; 1 year since the stroke]
>
> *Another, 35 years after their stroke, queried if it was possible to still feel depressed due to the stroke.* [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]

Medical Interventions

The survivors credited physiotherapy and speech therapy with helping them regain functionality and speech, respectively. The only downside mentioned was long waiting times to access services:

> *One parent advised others to get on top of physio immediately, saying that she went privately as waiting lists were long.* [Mother; participant 12; ≤18-year-old group; stroke at the age of 1 year; 11 years after the stroke]
>
> *One adult survivor writes that having physiotherapy sessions helped her walk.* [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]

Exercise was an activity recommended by medical practitioners that aided recovery. Parents and survivors mentioned the importance of repeatedly doing both general exercise and muscle-specific exercises to prevent muscle wasting, build muscle strength, help induce sleep, and increase fitness. Comments from the >18-year-old group also supported the utility of exercise, stressing the importance of regular daily

XSL•FO
RenderX

activity. Swimming was written about consistently; it was recommended by parents as a way to get their children's limbs to move, and it was noted that gains in swimming ability became an indicator of recovery and a source of excitement for the survivors and their families:

> *One individual advised parents to make a child do their exercises every day.* [Survivor; participant 52; >18-year-old group; stroke at the age of 17 years; 21 years after the stroke]
>
> *One parent wrote that swimming is a great way of exercising and moving their child's limbs without trauma.* [Mother; participant 11; ≤18-year-old group; stroke at the age of 1 year; 1 year after the stroke]
>
> *Another wrote their child had 1 to 1 swimming lessons and could nearly do breaststroke again.* [Mother; participant 25; ≤18-year-old group; stroke at the age of 8 years; 0 years after the stroke]

Other medical interventions that the participants mentioned were the use of Botox for tightness in muscles. Although it was often useful, it did not always help. In addition, the participants aged >18 years endorsed the use of SaeboFlex [28] for regaining functionality, tai chi for mental health, and quinine for muscle cramps and spasms.

Physical Factors

Fatigue

Fatigue was the most commonly reported physical barrier to recovery, with some parents writing that it was their main concern. When sleep duration was reported, it was at least 12 hours each night. The users told each other that tiredness was often caused by the large effort that the survivors were putting into their recovery. Fatigue was also a long-term factor:

> *A parent advised another that tiredness as a result of stroke is normal.* [Mother; participant 26; ≤18-year-old group; stroke at the age of 9 years; 0 years after the stroke]
>
> *One user reported that there was underlying tiredness throughout middle age, when they were working full time. In another post, they wrote that having a stroke resulted in less stamina, tolerance and energy in the context of noisy, busy backgrounds.* [Survivor; participant 45; >18-year-old group; stroke at the age of 15 years; 46 years after the stroke]

When a survivor became tired, this caused a worsening in disability. This affected speech, movement, and coordination:

> *A parent reported that when their child became tired on holiday, they were sad to see that their leg and arms were dragging.* [Mother; participant 12; ≤18-year-old group; stroke at the age of 1 year; 0 years after the stroke]

Pain

Pain was another commonly cited barrier preventing people from feeling fully recovered. Sites of pain were varied, including back pain, headaches, and limb muscle cramps. Pain emerged as a long-term consequence of a pediatric stroke. When the users discussed their pain with medical professionals, they were told that it was a result of their stroke and that it would ease with time. Being cold exacerbated chronic pain both in the short and long term:

> *One user reported that their headaches were seriously affecting their everyday life.* [Survivor; participant 47; ≤18-year-old group; stroke at the age of 15 years; 1 year after the stroke]
>
> *A user still has pain in their joints 35 years poststroke.* [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]
>
> *A parent wrote that their child's affected limbs were more painful in the cold weather.* [Mother; participant 32; >18-year-old group; stroke at the age of 13 years; 1 year after the stroke]

Neurological Sequelae

Poststroke memory loss, confusion, and dizziness were the barriers to recovery both in the short and long term:

> *A parent described her child's memory loss as causing as much trouble as the arm or foot.* [Mother; participant 32; ≤18-year-old group; stroke at the age of 13 years; 1 year after the stroke]
>
> *An individual 3 years poststroke commented that his eyes are dizzy.* [Survivor; participant 41; >18-year-old group; stroke at the age of 11 years; 30 years after the stroke]

Emotional Factors

Positivity

Adult survivors of pediatric stroke stressed that positivity was an important factor in ensuring a successful recovery. It was mentioned less often by parents and younger survivors of pediatric stroke. However, humor was brought up by both groups as an important facilitator in coping with the consequences of stroke. Many possible fun activities were exchanged on the site. There was a large proportion of music-related activities, for example, singing songs with hand puppets:

> *One survivor stresses staying positive is the key to a successful recovery.* [Survivor; participant 49; >18-year-old group; stroke at the age of 17 years; 2 years after the stroke]
>
> *One parent writes their family tries to laugh themselves through bad times, rather than cry again.* [Mother; participant 32; ≤18-year-old group; stroke at the age of 13 years; 1 year after the stroke]
>
> *One parent wrote that music is a great mental stimulator as well as being fun.* [Mother; participant 12; ≤18-year-old group; stroke at the age of 1 year; 0 years after the stroke]

Grief and Bereavement

Following a pediatric stroke, the survivors wrote that they viewed themselves as different people and grieved for the person they once were. The families of survivors also commented that

XSL•FO
RenderX

they experienced similar emotions and that the stroke had a long-term effect on the whole family:

> *One parent commented they have a different child now.* [Mother; participant 31; ≤18-year-old group; stroke at the age of 11 years; 5 years after the stroke]
>
> *Another survivor queried whether anyone on the forum felt that stroke was like a loss in the family and caused a grieving process.* [Survivor; participant 51; >18-year-old group; stroke at the age of 17 years; 4 years after the stroke]
>
> *One parent commented that they don't think any of their family will ever be the same as before the stroke.* [Mother; participant 32; ≤18-year-old group; stroke at the age of 13 years; 1 year after the stroke]

Fear

Fear of uncertainty and the unknown was commonly cited as a barrier to recovery. Parents wrote that the fear of another stroke event, how their child will grow up and fit into society, and having no explanation of the cause of the stroke were particularly difficult issues. In contrast, adult survivors encouraged users to accept the stroke and not let fear stand in the way of recovery:

> *A parent wrote that it was hard to comprehend the unknown future.* [Mother; participant 11; ≤18-year-old group; stroke at the age of 1 year; 1 year after the stroke]
>
> *One adult survivor wrote that acceptance of stroke is the first stage of the healing process and survivors must move on and get on with their lives.* [Survivor; participant 46; >18-year-old group; stroke at the age of 15 years; 28 years after the stroke]

Anxiety

Both parents and survivors described the scenarios that caused upset by reminding them of the trauma surrounding the stroke event:

> *One parent writes that her daughter was admitted to hospital and it brings back too many memories which makes their calm slip a little inside.* [Mother, participant 12; ≤18-year-old group; stroke at the age of 1 year; 0 years after the stroke]
>
> *An adult survivor training to be a health care professional writes that having a placement on a stroke unit is difficult to cope with.* [Survivor; participant 51; >18-year-old group; stroke at the age of 17 years; 4 years after the stroke]

Behavioral Problems

Behavioral problems were commonly mentioned. Parents reported mood swings with negative emotions, such as getting upset, anger, and frustration. There were multiple posts trying to rationalize why these changes occur, possible reasons being parental spoiling of the child either before or after the stroke and typical teenage mood swings. Adult survivors of pediatric stroke mentioned some lasting behavioral problems:

> *A parent wrote that their child gets upset with how she feels and that her brain seems to tell her things that she can't cope with.* [Mother; participant 30; ≤18-year-old group; stroke at the age of 11 years; 2 years after the stroke]
>
> *One survivor wrote that he takes himself to bed when he gets really grumpy.* [Survivor; participant 52; >18-year-old group; stroke at the age of 17 years; 21 years after the stroke]
>
> *Another survivor queries whether it is possible to have mood swings and feel low 35 years after a stroke as she is struggling.* [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]

Social Factors

Support Network

Parents stressed the importance of supporting their children throughout the stroke recovery process. Support was appreciated by the survivors of stroke, who reiterated the importance of love and support in motivating them during recovery:

One parent wrote *it is completely down to the family to get their child through recovery.* [Mother; participant 12; ≤18-year-old group; stroke at the age of 1 year; 0 years since the stroke]

> *An adult survivor of pediatric stroke wrote that they fought every inch of the way with the love and support of their parents.* [Survivor; participant 39; >18-year-old group; stroke at the age of 8 years; 23 years since the stroke]
>
> *A user mentioned their friend by name and thanked them for getting them over the mental side of things.* [Survivor; participant 42; >18-year-old group; stroke at the age of 13 years; 7 years after the stroke]

Loneliness

Loneliness emerged in many contexts as both a short-term and a long-term barrier to recovery from stroke. Loneliness made the survivors feel different from those around them, which negatively impacted their well-being. The survivors felt isolated from fellow survivors of stroke, friends, family, and peers:

> *A male survivor wrote he struggled to find any help and support with his rehabilitation.* [Survivor; participant 43; >18-year-old group; stroke at the age of 13 years; 7 years after the stroke]
>
> *A female survivor commented she never talks about her stroke to her friends because she doesn't want to be judged.* [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]
>
> *The parent of a survivor writes that her child's class look at him differently.* [Mother; participant 20; ≤18-year-old group; stroke at the age of 3 years; 2 years after the stroke]

Loneliness had long-lasting effects on survivors of stroke:

XSL•FO
RenderX

One member of the >18 group reported they felt alone most of their life. They then go on to tell a user they have done the best thing joining the site. [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]

Exclusion or Bullying

There were reports of bullying and exclusion behaviors among children at school. Furthermore, parents reported that social restrictions, such as not allowing the survivors to ride a bicycle or go out with friends, were a barrier to recovery:

A survivor's mother wrote that her daughter is taunted at school by her own friends for having a stroke. [Mother; participant 26; ≤18-year-old group; stroke at the age of 9 years; 0 years after the stroke]

One parent wrote that the restrictions placed on her son prompted him to tell them that he may as well be dead. [Mother; participant 22; ≤18-year-old group; stroke at the age of 5 years; 4 years after the stroke]

Hidden Disabilities

Being doubted about their degree of disability was reported by several adult survivors of pediatric stroke. They described the medical staff and the public as not understanding the hidden effects of stroke and assuming that if there are no visible impairments, they are recovered. This discredited the survivor's struggle for recovery and made them feel upset:

A survivor wrote how she told a midwife about her stroke and the midwife's response was that because they hadn't noticed at first glance, it couldn't have affected the survivor that badly. The survivor commented how angry that made her feel. [Survivor; participant 36; >18-year-old group; stroke at the age of 0 years; 35 years after the stroke]

A survivor reported people discrediting their tiredness, accusing them of being selfish or lazy. [Survivor; participant 51; >18-year-old group; stroke at the age of 17 years; 4 years after the stroke]

Education

There were mixed reviews about the education services accessed; however, interaction with the education sector was highlighted as important in getting the survivor of pediatric stroke back into school and therefore integrating back into society:

One parent writes that her child's new school is not dealing with her needs very well, and they have had to go in on several occasions. [Mother; participant 32; >18-year-old group; stroke at the age of 13 years; 1 year after the stroke]

Driving

There was concern from parents of survivors of stroke as to whether their children could drive. There was much reassurance and advice given by other parents and survivors in the >18-year-old group:

One parent writes that her child had his stroke at 11 and was then 16 and wondered if he would be able to drive. [Mother; participant 31; ≤18-year-old group; stroke at the age of 11 years; 5 years after the stroke]

One survivor commented he overcame medical predictions and had been driving since the age of 21. [Survivor; participant 46; >18-year-old group; stroke at the age of 15 years; 28 years after the stroke]

Travel

Parents reported that obtaining travel insurance for their child was difficult. Many companies did not provide insurance at all; they only insured per trip and not for longer periods as parents wanted, and the price of insurance was high, deterring the families from traveling:

One parent commented that they found it mind-blowing that they must pay so much and no one else will cover them. [Survivor; participant 52; >18-year-old group; stroke at the age of 17 years; 21 years after the stroke]

Third-Party Support

The members often recommended external resources to each other. These were most often helplines or information on websites. Organizations that were mentioned were the Stroke Association [1], HemiHelp [29], Maypole project [30], Different Strokes [31], Mobilise [32], Disability Information and Advice Line [33], and Headway [34].

Financial Support

Financial support for disability was available for survivors of pediatric stroke and was identified as helpful. However, there was discussion among adult survivors of pediatric stroke who expressed fear of having financial support taken away from them:

One survivor wrote that they had read an article about how people with mental health issues are more likely to fail the assessment test, as they have multiple symptoms which vary and she was worried that this would disadvantage stroke survivors also. [Survivor; participant 45; >18-year-old group; stroke at the age of 15 years; 46 years after the stroke]

PPI Feedback

The analysis was read by a survivor of multiple childhood strokes while aged <1 year who is now studying at a university. She reported strongly agreeing with uncertainty about the future, the benefits of swimming, the hidden disabilities because of stroke, and the lack of public awareness of pediatric stroke. Some quotes are reported in the subsequent section to illustrate the PPI feedback:

- Regarding swimming:

  I was also told it was important for me to swim to help prevent any damage as I developed so swam from 6 months on.

- Uncertainty about the future:

XSL•FO
RenderX

This strongly resonates with me and my family. For my parents it was a huge unknown as the doctors couldn't tell them if I would be able to walk, speech or do well in school. They also couldn't tell if I was going to have another one, but if I didn't have another by 12 I should be clear from not having another.

- Hidden disabilities:

I would agree hidden effects of a stroke are not spoken about. In the main all I was warned of was physical disabilities and those which would be major hindrances in my life. By this I mean walking, talking, learning and doing physical exercise/sport well. Once it was clear at ~7 or 8 year old I was able to conduct those tasks no further check ups were taken and no-one checked for any other less major/obvious effects. As for the public, I would strongly agree many have no idea of the effects of strokes, even I didn't realise tiredness was an effect until I read this paper! I have an absolutely terrible memory and I do think if this is very likely down to my stroke and this I would say is a hidden disability. I'm almost certain my stroke affected the memory side of my brain so it would make some sense. However, I assume few members of the public would understand this and no medical staff checked or mentioned smaller effects other than physical movement/speech/learning.

Discussion

Principal Findings

Medical, physical, emotional, and social themes were identified as impacting the recovery from pediatric stroke. Exercise, swimming, speech therapy, and physiotherapy were the medical factors that facilitated recovery, whereas having a comorbidity hindered recovery. A novel finding of this study is that "hidden" physical impairments, such as fatigue, pain, memory loss, confusion, and dizziness, affected survivors of pediatric stroke in the long term, with the lack of awareness of these impairments by the general public and by professionals in health care, school, and workplace settings also hindering recovery. Grief, fear of the future, restrictions on the child's life, and behavioral problems were the emotional factors that slowed recovery, whereas positivity and a good support network aided recovery. Isolation, hidden disabilities, triggering events, finding travel insurance, and fear of not being able to drive were the social factors hindering recovery. Third-party support and financial aid were facilitators of recovery, albeit there was a fear of losing financial aid as survivors of pediatric stroke aged.

Identifying the factors that impact recovery from pediatric stroke is important, as there is the potential for survivors to live for many decades after the stroke, and therefore, they may live many years with disabling factors. Our study found that fatigue persists many decades after stroke, which expands on current literature that fatigue is an identified barrier up to 5 years after the stroke [8]. In addition, tiredness was reported to exacerbate stroke-related disabilities. This relationship has been reported in adult brain injury [35], but to the best of our knowledge, this is the first time it has been reported in pediatric stroke. In addition, this study is the first to report chronic pain as a barrier to recovery in pediatric stroke, despite chronic pain being well known because of adult stroke [36].

Rehabilitation therapies are recommended following pediatric stroke; however, there is little evidence that describes survivors' experiences of these services. This study was the first to report that survivors of pediatric stroke found that both muscle-specific exercise and general exercise helped in recovery. The Royal College of Physicians national guidelines for adult stroke recommend muscle-specific exercises to improve functionality and found general exercise to help aerobic fitness, gait, and prevent regression of cognitive and functional gains after stroke as well as have positive psychological effects [37]. To our knowledge, no such research has established this for pediatric stroke. A new finding from this study was that swimming facilitated recovery. This has not been researched in relation to pediatric stroke, but a feasibility study on an introductory performance-focused swimming intervention for adult cerebral palsy found that swimming helped fatigue, physical function, and mental health [38]. Speech and language therapy is commonly used during rehabilitation, despite a lack of evidence [25]. This study provides qualitative evidence that participants found this intervention useful and recommended it to other users. Similarly, this is the first qualitative study to explore patients' experiences of physiotherapy.

Our study found that stroke-related social restriction on a child's life has a negative emotional effect. This is an important finding, as survivors of pediatric brain injuries, including stroke, commonly have lower levels of community activity and peer social play at school [39]. Another finding from this study was that isolation and exclusion were barriers to recovery. Denham et al [15] found that following a pediatric stroke, families feel abandoned by friends or their intimate support network. Our study further characterized this dimension as "isolation" from other survivors, friends, the wider family, and society. Bullying has also been found to be a barrier in previous studies [40,41]. Uncertainty about the future has also been reported [12]. This study found this is a short-term barrier, as participants in the >18-year-old group did not share this concern, instead putting the stroke behind them and being more positive about the future. Financial support is accessible after recovery from stroke. However, a new finding from this study was that there was a long-term fear of losing financial support as disability allowance was reassessed. Driving is an important indicator of independence and recovery after an adult stroke [42]. This study found that the parents and the survivors were concerned about whether they could drive after a pediatric stroke, and this concern was often expressed many years after the stroke. This was only a short-term concern, as the users were reassured by adult survivors of pediatric stroke who had driving licenses and could explain the process of learning to drive with a disability. Difficulty obtaining travel insurance has been reported so far only in the literature on adult stroke [43].

This study found that both parents and children experienced anxiety-triggering events that reminded them of the happenings of their original stroke. Although a diagnosis cannot be reached from these comments, Lehman et al [44] found increased symptoms of posttraumatic stress disorder in children and their

XSL•FO
RenderX

parents, suggesting that the memory of stroke has long-lasting emotional effects on the family. Our findings support previous evidence that developing comorbidities [3-7], behavioral problems [4,9], and the bereavement process caused by the stroke [45] were factors hindering recovery. A good support network [13] and third-party support and information aided recovery [4,13,15], and the users often recommended useful resources to each other.

Strengths and Limitations

The strength of this study lies in the source of data, which is a UK-wide online community with several survivors of pediatric stroke at varying stages in their lives. First, the community facilitated the discussion between participants who had a stroke recently and users who had a stroke many years ago. This allowed unprecedented insight into longer-term factors inhibiting recovery. Second, the discussions were initiated by the participants and continued in an asynchronous way, with no time, geographical, or behavioral constraints on communication. This dynamic cannot be replicated in traditional research approaches, for example, in interviews. Finally, the population that uses the forum might include people who do not partake in traditional research studies, thereby including perspectives from an underrepresented patient population [23]. The limitations of this study are that the users may not mention all the factors that affect their recovery; therefore, our analysis may not be comprehensive. First, this may be because the users do not raise everything pertinent to the research question, the users were only active over a limited period of their recovery, or the forum was moderated, so some posts may have been removed or affected by the moderation process. Second, the authenticity of posts could not be determined. Third, the data set is individuals aged >10 years, and factors affecting recovery may have changed in the time between the posts and this analysis. PPI feedback was limited to a single individual. Finally, there is a nonactive population that reads but does not compose messages. These users may be more numerous than the registered users themselves; one study [46] reported a 26:1 ratio of lurkers for every author of a message in an online forum. This population cannot be quantified or classified.

Conclusions

This study identified novel findings regarding factors affecting recovery from pediatric stroke. Raising awareness about the lived experience of survivors of pediatric stroke and the type and impact of long-term impairments is needed in health care settings, schools, workplaces, driving centers, and travel agencies so that appropriate support and information can be provided. Recovery from a stroke is an evolving process that lasts decades. Although this study has highlighted some long-term factors impacting recovery, more research needs to be performed to further establish these as well as design interventions to alleviate these barriers. This will result in effective, long-term support for survivors. In particular, fatigue, pain, and loneliness were the physical problems that were present many decades after the stroke, and there are no effective interventions reported yet.

Acknowledgments

The authors are grateful to the Stroke Association for sharing the archived file of TalkStroke on the web with them. The authors thank their patient and public involvement representative for comments on the manuscript.

ADS was partly funded by a National Institute for Health and Care Research (NIHR) Programme Grant for Applied Research (NIHR202037). The views presented in this paper are those of the authors and not necessarily those of the National Health Service, NIHR, or the Department of Health and Social Care.

Conflicts of Interest

None declared.

References

1. What is childhood stroke? Stroke Association. URL: https://www.stroke.org.uk/childhood-stroke/about-childhood-stroke [accessed 2023-03-12]
2. Gordon AL, Nguyen L, Panton A, Mallick AA, Ganesan V, Wraige E, et al. Self-reported needs after pediatric stroke. Eur J Paediatr Neurol 2018 Sep;22(5):791-796. [doi: 10.1016/j.ejpn.2018.06.003] [Medline: 29960841]
3. Greenham M, Gordon A, Anderson V, Mackay MT. Outcome in childhood stroke. Stroke 2016 Apr;47(4):1159-1164. [doi: 10.1161/strokeaha.115.011622]
4. Dunbar M, Kirton A. Perinatal stroke. Semin Pediatr Neurol 2019 Dec;32:100767. [doi: 10.1016/j.spen.2019.08.003] [Medline: 31813521]
5. Rivella C, Zanetti A, Bertamino M, Primavera L, Moretti P, Viterbori P. Emotional and social functioning after stroke in childhood: a systematic review. Disabil Rehabil 2023 Dec 17;45(25):4175-4189. [doi: 10.1080/09638288.2022.2144490] [Medline: 36384380]
6. Hamner T, Shih E, Ichord R, Krivitzky L. Children with perinatal stroke are at increased risk for autism spectrum disorder: prevalence and co-occurring conditions within a clinically followed sample. Clin Neuropsychol 2022 Jul 24;36(5):981-992. [doi: 10.1080/13854046.2021.1955150] [Medline: 34308766]

XSL•FO
RenderX

7. Sundelin H, Söderling J, Bang P, Bolk J. Risk of autism after pediatric ischemic stroke. Neurology 2022 May 10;98(19):e1953-e1963. [doi: 10.1212/wnl.0000000000200253]
8. Wrightson JG, Zewdie E, Kuo HC, Millet GY, Kirton A. Fatigue in children with perinatal stroke: clinical and neurophysiological associations. Dev Med Child Neurol 2020 Feb 20;62(2):234-240 [FREE Full text] [doi: 10.1111/dmcn.14273] [Medline: 31222717]
9. Soufi S, Chabrier S, Bertoletti L, Laporte S, Darteyre S. Lived experience of having a child with stroke: a qualitative study. Eur J Paediatr Neurol 2017 May;21(3):542-548. [doi: 10.1016/j.ejpn.2017.01.007] [Medline: 28185801]
10. Amlie-Lefond C. Evaluation and acute management of ischemic stroke in infants and children. Continuum (Minneap Minn) 2018 Feb;24(1, Child Neurology):150-170. [doi: 10.1212/CON.0000000000000559] [Medline: 29432241]
11. Mrakotsky C, Williams TS, Shapiro KA, Westmacott R. Rehabilitation in pediatric stroke: cognition and behavior. Semin Pediatr Neurol 2022 Dec;44:100998. [doi: 10.1016/j.spen.2022.100998] [Medline: 36456041]
12. McKevitt C, Topor M, Panton A, Mallick AA, Ganesan V, Wraige E, et al. Seeking normality: parents' experiences of childhood stroke. Child Care Health Dev 2019 Jan 15;45(1):89-95. [doi: 10.1111/cch.12622] [Medline: 30255632]
13. Kirton A, deVeber G. Paediatric stroke: pressing issues and promising directions. Lancet Neurol 2015 Jan;14(1):92-102. [doi: 10.1016/s1474-4422(14)70227-3]
14. Williams TS, McDonald KP, Roberts SD, Westmacott R, Dlamini N, Tam EW. Understanding early childhood resilience following neonatal brain injury from parents' perspectives using a mixed-method design. J Int Neuropsychol Soc 2019 May 3;25(04):390-402. [doi: 10.1017/s1355617719000079]
15. Denham AM, Wynne O, Baker AL, Spratt NJ, Turner A, Magin P, et al. "This is our life now. Our new normal": a qualitative study of the unmet needs of carers of stroke survivors. PLoS One 2019 May 8;14(5):e0216682 [FREE Full text] [doi: 10.1371/journal.pone.0216682] [Medline: 31067286]
16. Rosewilliam S, Roskell CA, Pandyan AD. A systematic review and synthesis of the quantitative and qualitative evidence behind patient-centred goal setting in stroke rehabilitation. Clin Rehabil 2011 Jun 25;25(6):501-514. [doi: 10.1177/0269215510394467] [Medline: 21441308]
17. Salinas J, Sprinkhuizen SM, Ackerson T, Bernhardt J, Davie C, George MG, et al. An international standard set of patient-centered outcome measures after stroke. Stroke 2016 Jan;47(1):180-186 [FREE Full text] [doi: 10.1161/STROKEAHA.115.010898] [Medline: 26604251]
18. De Simoni A, Shanks A, Balasooriya-Smeekens C, Mant J. Stroke survivors and their families receive information and support on an individual basis from an online forum: descriptive analysis of a population of 2348 patients and qualitative study of a sample of participants. BMJ Open 2016 Apr 06;6(4):e010501 [FREE Full text] [doi: 10.1136/bmjopen-2015-010501] [Medline: 27053271]
19. Panzarasa P, Griffiths CJ, Sastry N, De Simoni A. Social medical capital: how patients and caregivers can benefit from online social interactions. J Med Internet Res 2020 Jul 28;22(7):e16337 [FREE Full text] [doi: 10.2196/16337] [Medline: 32720910]
20. Balasooriya-Smeekens C, Bateman A, Mant J, De Simoni A. Barriers and facilitators to staying in work after stroke: insight from an online forum. BMJ Open 2016 Apr 06;6(4):e009974 [FREE Full text] [doi: 10.1136/bmjopen-2015-009974] [Medline: 27053267]
21. Jamison J, Sutton S, Mant J, De Simoni A. Barriers and facilitators to adherence to secondary stroke prevention medications after stroke: analysis of survivors and caregivers views from an online stroke forum. BMJ Open 2017 Jul 16;7(7):e016814 [FREE Full text] [doi: 10.1136/bmjopen-2017-016814] [Medline: 28713074]
22. Jamison J, Sutton S, Mant J, De Simoni A. Online stroke forum as source of data for qualitative research: insights from a comparison with patients' interviews. BMJ Open 2018 Mar 30;8(3):e020133 [FREE Full text] [doi: 10.1136/bmjopen-2017-020133] [Medline: 29602848]
23. De Simoni A, Shanks A, Mant J, Skelton JR. Making sense of patients' internet forums: a systematic method using discourse analysis. Br J Gen Pract 2014 Feb 24;64(620):e178-e180. [doi: 10.3399/bjgp14x677671]
24. Greenham M, Anderson V, Cooper A, Hearps S, Ditchfield M, Coleman L, et al. Early predictors of psychosocial functioning 5 years after paediatric stroke. Dev Med Child Neurol 2017 Oct 17;59(10):1034-1041 [FREE Full text] [doi: 10.1111/dmcn.13519] [Medline: 28815654]
25. Stroke in childhood: clinical guideline for diagnosis, management and rehabilitation. Royal College of Paediatrics and Child Health. 2017 May. URL: https://www.rcpch.ac.uk/sites/default/files/2021-02/Stroke%20guideline%2008.04.19%20updated%202021.pdf [accessed 2023-03-12]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Rehabiliatation. World Health Organization. 2023 Jan 30. URL: https://www.who.int/news-room/fact-sheets/detail/rehabilitation [accessed 2023-03-12]
28. Saebo UK homepage. Saebo UK. URL: https://uk.saebo.com/ [accessed 2023-03-19]
29. HemiHelp – a history. Contact. URL: https://contact.org.uk/help-for-families/information-advice-services/hemihelp/hemiplegia-support/ [accessed 2023-03-12]
30. The Maypole project homepage. The Maypole Project. URL: https://www.themaypoleproject.co.uk/ [accessed 2023-03-12]

XSL•FO
RenderX

31. Different Strokes homepage. Different Strokes. URL: https://differentstrokes.co.uk/ [accessed 2023-03-12]
32. Mobilise homepage. Mobilise. URL: https://www.mobiliseonline.co.uk/ [accessed 2023-03-12]
33. Welcome to DIAL UK. Disability Information and Advice Line. URL: https://www.dialuk.info/ [accessed 2023-03-12]
34. Headway - the brain injury association homepage. Headway - The Brain Injury Asscoiation. URL: https://www.headway.org.uk/ [accessed 2023-03-12]
35. Ezekiel L, Field L, Collett J, Dawes H, Boulton M. Experiences of fatigue in daily life of people with acquired brain injury: a qualitative study. Disabil Rehabil 2021 Oct 04;43(20):2866-2874. [doi: 10.1080/09638288.2020.1720318] [Medline: 32017641]
36. Harrison RA, Field TS. Post stroke pain: identification, assessment, and therapy. Cerebrovasc Dis 2015 Mar 5;39(3-4):190-201 [FREE Full text] [doi: 10.1159/000375397] [Medline: 25766121]
37. Stroke guidelines 2016. Royal College of Physicians. 2016. URL: https://www.rcplondon.ac.uk/guidelines-policy/stroke-guidelines-2016 [accessed 2023-02-12]
38. Dutia IM. The effect of performance-focused swimming training on clinical outcomes in young people with cerebral palsy who have high support needs. School of Human Movement and Nutrition Sciences, The University of Queensland. 2020. URL: https://espace.library.uq.edu.au/view/UQ:8951f19 [accessed 2024-04-01]
39. van Tol E, Gorter JW, DeMatteo C, Meester-Delver A. Participation outcomes for children with acquired brain injury: a narrative review. Brain Inj 2011 Sep 30;25(13-14):1279-1287. [doi: 10.3109/02699052.2011.613089] [Medline: 21961572]
40. Kirton A, deVeber G. Life after perinatal stroke. Stroke 2013 Nov;44(11):3265-3271. [doi: 10.1161/strokeaha.113.000739]
41. Sporns PB, Fullerton HJ, Lee S, Kirton A, Wildgruber M. Current treatment for childhood arterial ischaemic stroke. Lancet Child Adolesc Health 2021 Nov;5(11):825-836. [doi: 10.1016/s2352-4642(21)00167-x]
42. Devos H, Hawley CA, Conn AM, Marshall SC, Akinwuntan AE. Driving after stroke. In: Platz T, editor. Clinical Pathways in Stroke Rehabilitation. Cham, Switzerland: Springer; 2021.
43. Hodson T, Gustafsson L, Cornwell P. The lived experience of supporting people with mild stroke. Scand J Occup Ther 2020 Apr 02;27(3):184-193. [doi: 10.1080/11038128.2019.1633401] [Medline: 31264497]
44. Lehman LL, Maletsky K, Beaute J, Rakesh K, Kapur K, Rivkin MJ, et al. Prevalence of symptoms of anxiety, depression, and post-traumatic stress disorder in parents and children following pediatric stroke. J Child Neurol 2020 Jun;35(7):472-479. [doi: 10.1177/0883073820909617] [Medline: 32202201]
45. Yehene E, Brezner A, Ben-Valid S, Golan S, Bar-Nadav O, Landa J. Factors associated with parental grief reaction following pediatric acquired brain injury. Neuropsychol Rehabil 2021 Jan 26;31(1):105-128. [doi: 10.1080/09602011.2019.1668280] [Medline: 31556807]
46. Burri M, Baujard V, Etter JF. A qualitative analysis of an internet discussion forum for recent ex-smokers. Nicotine Tob Res 2006 Dec;8 Suppl 1(1):S13-S19. [doi: 10.1080/14622200601042513] [Medline: 17491166]

Abbreviations

PPI: patient and public involvement

Edited by A Mavragani; submitted 29.05.23; peer-reviewed by G Engvall; comments to author 15.01.24; revised version received 04.02.24; accepted 08.03.24; published 16.04.24.

Please cite as:
Howdle C, Wright WJA, Mant J, De Simoni A
Factors Influencing Recovery From Pediatric Stroke Based on Discussions From a UK-Based Online Stroke Community: Qualitative Thematic Study
J Med Internet Res 2024;26:e49409
URL: https://www.jmir.org/2024/1/e49409
doi:10.2196/49409
PMID:38625726

©Charlotte Howdle, William James Alexander Wright, Jonathan Mant, Anna De Simoni. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of Prosocial Behaviors on e-Consultations in a Web-Based Health Care Community: Panel Data Analysis

Xiaoxiao Liu[1,2], PhD; Huijing Guo[3], PhD; Le Wang[4], PhD; Mingye Hu[5], PhD; Yichan Wei[1], BBM; Fei Liu[6], PhD; Xifu Wang[7], MCM

[1]School of Management, Xi'an Jiaotong University, Xi'an, China
[2]China Institute of Hospital Development and Reform, Xi'an Jiaotong University, Xi'an, China
[3]School of Economics and Management, China University of Mining and Technology, Xuzhou, China
[4]College of Business, City University of Hong Kong, Hong Kong, China (Hong Kong)
[5]School of Economics and Management, Xi'an University of Technology, Xi'an, China
[6]School of Management, Harbin Engineering University, Harbin, China
[7]Healthcare Simulation Center, Guangzhou First People's Hospital, Guangzhou, China

Corresponding Author:
Xifu Wang, MCM
Healthcare Simulation Center
Guangzhou First People's Hospital
1 Pan Fu Road
Yuexiu District
Guangzhou, 510180
China
Phone: 86 13560055951
Email: wangxifu.simulation@gmail.com

Abstract

Background: Patients using web-based health care communities for e-consultation services have the option to choose their service providers from an extensive digital market. To stand out in this crowded field, doctors in web-based health care communities often engage in prosocial behaviors, such as proactive and reactive actions, to attract more users. However, the effect of these behaviors on the volume of e-consultations remains unclear and warrants further exploration.

Objective: This study investigates the impact of various prosocial behaviors on doctors' e-consultation volume in web-based health care communities and the moderating effects of doctors' digital and offline reputations.

Methods: A panel data set containing information on 2880 doctors over a 22-month period was obtained from one of the largest web-based health care communities in China. Data analysis was conducted using a 2-way fixed effects model with robust clustered SEs. A series of robustness checks were also performed, including alternative measurements of independent variables and estimation methods.

Results: Results indicated that both types of doctors' prosocial behaviors, namely, proactive and reactive actions, positively impacted their e-consultation volume. In terms of the moderating effects of external reputation, doctors' offline professional titles were found to negatively moderate the relationship between their proactive behaviors and their e-consultation volume. However, these titles did not significantly affect the relationship between doctors' reactive behaviors and their e-consultation volume (P=.45). Additionally, doctors' digital recommendations from patients negatively moderated both the relationship between doctors' proactive behaviors and e-consultation volume and the relationship between doctors' reactive behaviors and e-consultation volume.

Conclusions: Drawing upon functional motives theory and social exchange theory, this study categorizes doctors' prosocial behaviors into proactive and reactive actions. It provides empirical evidence that prosocial behaviors can lead to an increase in e-consultation volume. This study also illuminates the moderating roles doctors' digital and offline reputations play in the relationships between prosocial behaviors and e-consultation volume.

(J Med Internet Res 2024;26:e52646) doi:10.2196/52646

XSL•FO
RenderX

KEYWORDS

prosocial behaviors; proactive behaviors; reactive behaviors; reputations; e-consultation volume; live streaming

Introduction

Background

e-Consultations, offered through web-based health care communities [1], are increasingly becoming vital complements to traditional hospital services [2-4]. In hospital consultations, patients can only passively accept treatment [5] from a limited pool of medical resources within a geographical radius. However, when engaging with web-based health care communities, patients can search for primary care solutions [6] from an extensive digital market in a relatively short time [7]. Given that the diagnostic accuracy of e-consultations matches that of hospital consultations [8-10], e-consultations are becoming increasingly attractive to patients [3,11].

Doctors are also showing a growing interest in e-consultations, motivated by economic and social benefits. First, doctors can achieve economic gains by participating in e-consultations [7,12]. Web-based consultation platforms facilitate an efficient reputation system, enabling patients to easily provide feedback about doctors. Consequently, doctors can use e-consultation to strengthen their relationship with patients [13,14] and foster positive word-of-mouth [15]. More e-consultations can benefit doctors by retaining current patients, attracting new ones, and boosting in-person hospital visits [16,17]. Second, doctors could also receive social returns from engaging in e-consultation [7]. Active participation in e-consultations allows doctors to demonstrate their skills, attitude, and experience, aiding in accumulating professional capital [7], building their reputation [18], and increasing their social influence [19]. Given these tangible and intangible benefits, it is essential for doctors to diligently provide the desired e-consultations and make additional efforts to highlight their service attributes to stand out [6,20,21]. This involves engaging in prosocial behaviors in web-based health care communities, which is the primary research focus of this study.

Prior studies have examined the effects of prosocial behaviors on financial outcomes, such as actions reflecting social responsibility in the workplace [22]. In the health care sector, previous research has explored doctors' prosocial behaviors within traditional, offline medical services. Doctors, working in established medical institutes and serving patients with limited choices of clinical service providers, often aim for self-satisfaction and patient satisfaction with their offline prosocial behaviors. For example, research indicates that doctors may act prosocially to regulate their self-oriented feelings [23] and foster a caring and understanding attitude toward patients [24,25]. Additionally, doctors who demonstrate more empathy and care can elicit positive emotions in patients and improve the doctor-patient relationship [26,27].

Compared to the offline context, doctors' prosocial behaviors in a digital context may differ in 2 aspects. First, the internet allows patients to choose from a broader, more diverse range of doctors without the constraints of time and space [7]. However, the uncertainty inherent in the digital environment creates a more pronounced information asymmetry between patients and doctors [28], consequently making it more challenging for patients to establish trust. Therefore, doctors' prosocial behaviors are crucial in building their self-image, establishing patients' trust, and assisting patients in identifying suitable doctors [29,30]. Second, unlike offline environments, web-based medical platforms offer a range of functions, including asynchronous activities such as publishing articles, as well as real-time interactional actions such as answering questions during live streams. This array of functions facilitates the adoption of more diverse prosocial behaviors by doctors.

Although these differences underscore the importance of studying doctors' prosocial behavior, there has been limited research focusing on the impact of such behaviors in the digital context. One previous study has scrutinized the impact of prosocial behaviors, such as answering patients' questions freely, on patient engagement within web-based health care communities [31]. An aspect that requires further exploration is how doctors' motivations and patients' involvement vary in doctors' helping behaviors. Consequently, studies on web-based health care communities should differentiate between diverse prosocial actions to understand their effects on doctors' web-based service outcomes. This study aims to contribute new knowledge regarding the full breadth of doctors' prosocial behaviors.

Unlike the previous study that exclusively investigated doctors' asynchronous behaviors in web-based health care communities [31], this study also explores the role of synchronous reactive actions in achieving optimal doctors' e-consultation volume. Recently, web-based health care communities have developed and released live-streaming functions to assist doctors in providing voluntary interactions with patients. The effect of doctors' engagement in medical live streaming on e-consultation services remains unexplored. While these behaviors could demonstrate doctors' ethical traits and ability to fulfill an e-consultation workflow, a potential trade-off with e-consultations may exist when doctors engage in prosocial behaviors.

In summary, this study examines the effects of doctors' proactive and reactive prosocial behaviors, considering their digital and offline reputations as potential moderating factors. First, drawing from functional motives theory (FMT), we explore the impact of doctors' web-based proactive actions on their e-consultation volume. Proactive behaviors are actions in which individuals exceed their assigned work, focusing on long-term goals to prevent future problems [32,33]. According to FMT, these behaviors reflect helping actions that satisfy personal needs [34], driven by self-focused motivations [35], such as impression management and the realization of self-worth goals. For example, knowledge-based proactive behaviors, such as disseminating expertise to preempt future issues, are self-initiated and not reactions to immediate requests [36]. This study categorizes doctors' sharing of professional articles as a form of proactive behavior that creates a professional image for

XSL•FO
RenderX

their patient audience. This is because these actions aim to assist patients with future health concerns rather than directly responding to patients' immediate needs.

Second, this study explores the role of doctors' reactive prosocial behaviors in increasing e-consultations, guided by social exchange theory (SET). Unlike proactive behaviors, reactive behaviors are characterized by instances of individuals engaging in helping activities [35], typically in response to others' needs [34]. SET posits that individuals incurring additional social costs in relationships may anticipate reciprocal value [37,38]. Reactive prosocial behaviors, per SET, are initiated by the motivation to satisfy others' desires, leading to the development of cooperative social values. In our context, medical live streams facilitate real-time, synchronized interactions, enabling patients to ask questions and doctors to provide immediate responses. Patients' health questions during these streams indicate their immediate needs. Thus, a higher frequency of live streams within a certain period suggests doctors are increasingly responding to patients' needs during that time. Therefore, this study uses the number of medical live-streaming sessions conducted by doctors as a measure for their synchronous reactive behaviors.

Finally, considering that doctors' reputations play a crucial role in their workflow on web-based health care communities [39,40], we test the moderating roles of digital and offline reputation—measured by doctors' offline professional titles and patients' recommendations in the digital context, respectively—on the main effects.

Objective

Based on previous studies and practices within web-based health care communities, we aim to extend the literature by testing the impact of 2 types of web-based prosocial behaviors by doctors: proactive and synchronous reactive actions on e-consultation volume. We then explore the moderating roles of doctors' offline and digital reputations on these main effects.

Research Framework and Hypothesis Development

Overview

We have developed a research framework, shown in Figure 1, to identify effective prosocial strategies used by doctors within web-based health care communities to achieve a preferred e-consultation volume from the supply side.

Figure 1. Research framework.

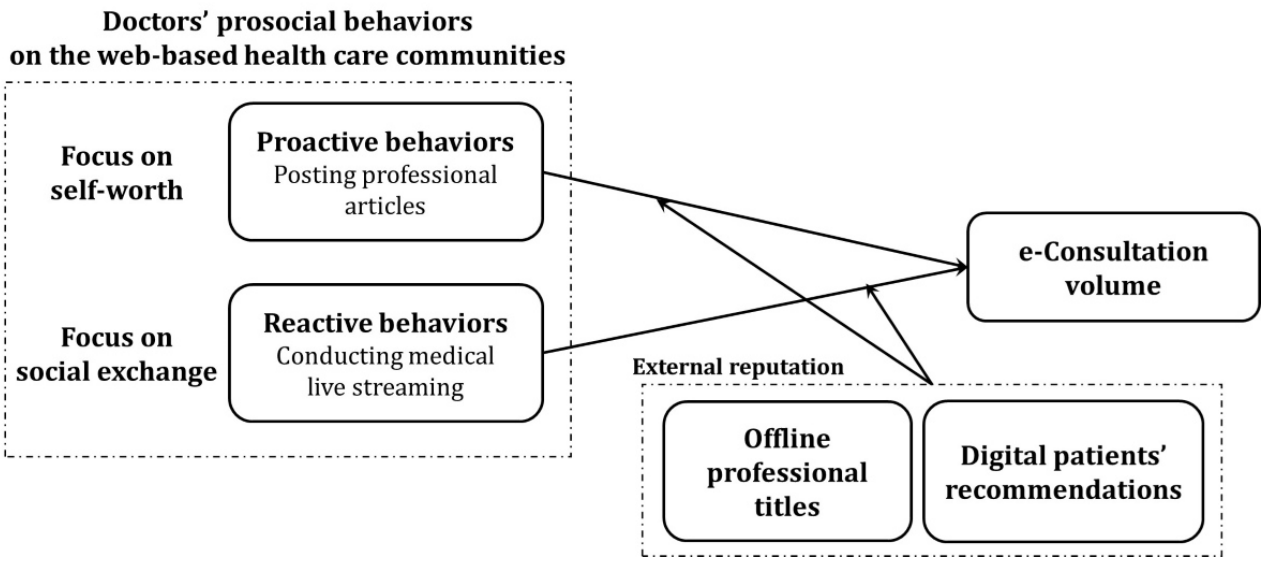

Primarily, we explore the relationships between doctors' prosocial behaviors and e-consultation volume, drawing on FMT and SET. These theories are widely adopted for measuring and classifying the outcomes of prosocial behaviors from 2 fundamental perspectives based on human nature [34]. While doctors' offline prosocial behaviors may help satisfy patients [24,25], who are already service acceptors, the outcomes of doctors' web-based prosocial behaviors still need careful distinction. It is essential to clearly differentiate between various types of doctors' prosocial behaviors to identify their nature. In this study, following the leads of FMT and SET, we test 2 kinds of prosocial behavior: proactive (posting professional articles to achieve self-worth) and reactive (conducting medical live streaming to create cooperative social values).

Subsequently, we examine how doctors' external reputation moderates the impacts of doctors' proactive and reactive prosocial behaviors. This examination is conducted from the perspectives of reducing uncertainty and building trust, respectively.

Doctors' Proactive Behaviors and e-Consultation Volume

FMT places emphasis on the primary motivations behind individuals' behaviors, adopting an atheoretical stance [41]. Through the exploratory process, previous studies have provided examples to identify the functional motivations behind prosocial behaviors [42], such as expressing important personal values. In web-based health care communities, doctors have the opportunity to demonstrate personal traits through proactive behaviors. According to FMT, these proactive behaviors stem from the actors' active efforts to satisfy their own needs and achieve self-worth [34,35].

XSL•FO
RenderX

Doctors might post professional articles, such as clinical notes and scientific papers, on web-based health care communities to help patient readers handle future health problems. These proactive prosocial behaviors are primarily driven by a desire to showcase personal medical competence, a crucial characteristic of a professional image [43], in medical consultations. By posting professional articles, doctors can display their medical knowledge, care delivery capability, and service quality, thereby enhancing their professional image. We hypothesize that this effort will lead to an increase in the e-consultation volume. Therefore, we propose the following hypothesis:

- Hypothesis 1: The posting of professional articles by doctors positively impacts their e-consultation volume on web-based health care communities.

Doctors' Reactive Behaviors and e-Consultation Volume

Considering the social environment in the working context, SET suggests that reactive prosocial behaviors stem from responding to others' needs [34]. Engaging in such behaviors can foster positive perceptions among the audience and build cooperative social values [44] through reactive social exchange. People with a high orientation toward cooperative social values act to maximize mutual interests [45], a trait highly valued in the medical field.

We use medical live streaming as a measure of doctors' reactive behaviors on web-based health care communities. Volunteering to provide interactional live streaming, a typical reactive behavior that may generate cooperative social value, gives the patient audience the impression that the doctors will prioritize demand-side interests during e-consultation services. Additionally, engaging in medical live streaming allows doctors to present themselves as authentic and recognized experts. This enhances their social presence [46], potentially leading to increased service use [47] and greater popularity [48]. Consequently, patients are more likely to perceive doctors who participate in medical live streaming as trustworthy for consultations. Given that e-consultations are closely related to the health conditions of the demand side, a credible doctor is likely to attract more e-consultations. Therefore, we propose the following hypothesis:

- Hypothesis 2: The conduct of medical live streaming by doctors positively impacts their e-consultation volume on web-based health care communities.

Moderating Roles of Offline and Digital Reputation

As doctors' proactive and reactive behaviors potentially affect their consultation performance, based on 2 distinct theoretical foundations of human nature, there exists a discrepancy in how doctors' reputations influence the relationship between various prosocial behaviors and e-consultation.

We formulate hypotheses regarding the moderating effects within the context of digital health care, by taking into account the inherent information asymmetry and the significance of establishing patient trust. Specifically, our hypotheses explore the influence of reputation on the relationship between doctors' proactive behaviors and e-consultation volume, with a focus on reducing uncertainty. Additionally, we examine how reputation moderates the impact of doctors' reactive behaviors, emphasizing the perspective of trust building.

First, in the marketing literature, service providers' reputations, which can reduce information asymmetry and purchase uncertainty [49], are key factors influencing purchasing behavior and sales performance in the digital context [50-52]. Similarly, for doctors, reputations are related to the experiences and beliefs of other stakeholders [53]. As health care services are credence goods [54]—whose quality patients cannot discern even after experiencing the services—and given the nature of web-based platforms (eg, the absence of face-to-face meetings), there is a significant information asymmetry [51]. This increases patients' uncertainty regarding the quality of doctors. Consequently, doctors' reputations play crucial roles in patients' decision-making processes [18,39]. We use doctors' professional titles and patients' recommendations on web-based health care communities to measure doctors' offline and digital reputations.

Proactive behaviors by low-reputation doctors can create deeper professional impressions [34,35] to reduce uncertainty in e-consultations than high-reputation doctors, who are less uncertain in medical services. Then, doctors' reputations—measured by offline professional titles and digital patients' recommendations on web-based health care communities—will negatively moderate the relationship between proactive behavior and e-consultation volume. Thus, we propose the following hypotheses:

- Hypothesis 3a: Doctors' offline professional titles negatively moderate the relationship between the posting of professional articles and e-consultation volume on web-based health care communities.
- Hypothesis 3b: Doctors' digital recommendations from patients negatively moderate the relationship between the posting of professional articles and e-consultation volume on web-based health care communities.

Second, one of the central elements of SET is the concept of trust between actors in the exchange process [55-58]. In the context of digital health, patient's trust in doctors is important to establish in order to refine the doctor-patient relationship. Doctors' reputations can reflect their personality traits [39] and promote trust from patients [53]. Conducting medical live streaming, a form of reactive prosocial behavior, includes doctors' cooperative social value orientations that are preferred in e-consultations. For low-reputation doctors, such as those with relatively junior professional titles and few digital patient recommendations, conducting medical live streaming will build patients' confidence in e-consultations to a greater extent than doctors with high reputations, who are usually already highly trusted. Then, offline and digital reputation may negatively moderate the relationship between engaging in medical live streaming and e-consultation volume. Thus, we propose the final hypotheses:

- Hypothesis 4a: Doctors' offline professional titles will negatively moderate the relationship between conducting medical live streaming and e-consultation volume on web-based health care communities.

XSL•FO
RenderX

- Hypothesis 4b: Doctors' digital recommendations from patients will negatively moderate the relationship between conducting medical live streaming and e-consultation volume on web-based health care communities.

Methods

Research Context and Data Collection

Our research context is one of the largest web-based health care communities in China. This platform, established in 2006, offers e-consultation services to patients. As of July 2023, it boasts over 260,000 active doctors from 10,000 hospitals nationwide and has provided web-based medical services to 79 million patients.

The platform allows doctors to create home pages where they can display relevant information such as offline professional titles, experiences shared by other patients, and personal introductions. Patients can select doctors for e-consultation by browsing this information. Besides e-consultation, doctors can engage in prosocial behavior primarily focused on knowledge sharing. This includes posting professional articles in various formats (text, voice, and short videos) and conducting medical live streams for real-time interaction with patients.

We collected data over a 22-month period, from January 2021 to October 2022, focusing on common diseases such as diabetes, depression, infertility, skin diseases, and gynecological diseases. To ensure that our findings are generalizable to a typical and active doctor on the platform, we included doctors who had posted at least 1 article and conducted at least 1 live stream before the end of the study period in our analysis [59-61]. Our sample consists of 2880 doctors and includes the following information for each doctor: professional title, patient recommendations, records of experiences shared by the doctor's patients, records of professional articles posted, records of live streams conducted, and records of the doctor's e-consultations.

Variable Operationalization

Overview

Our unit of analysis is each doctor. We investigate how doctors' prosocial behaviors, including proactive behaviors (posting professional articles) and reactive behaviors (conducting medical live streams), influence their e-consultation volume.

Dependent Variable

Our dependent variable is the doctors' e-consultation volume, denoted as $Consultation_{it}$, which is measured by the number of e-consultations of doctor i in month t.

Independent Variables

Our independent variables are doctors' proactive behaviors and reactive behaviors. Doctors' proactive behavior is operationalized as the posting of professional articles. Specifically, we denote proactive behavior as $Articles_{it}$, which is measured by the number of professional articles posted by doctor i in month t. Doctors' reactive behavior is operationalized as medical live streaming. This variable is denoted as $LiveStreaming_{it}$, which is calculated as the number of medical live streams conducted by doctor i in month t.

Moderating Variables

We are also interested in how doctors' external reputation, including their offline professional titles and digital recommendations from patients, influences the relationship between prosocial behaviors and e-consultation volume. A doctor's offline professional title is denoted as $Title_i$, which is a dummy variable indicating whether doctor i is a chief doctor ($Title_i$=1 indicates the doctor is a chief doctor, and $Title_i$=0 indicates the doctor has a lower-ranked title). Digital recommendations are captured by $Recommendations_i$, which is the digital recommendation level of doctor i as calculated by the platform based on the recommendations provided by their past patients.

Control Variables

We incorporated several control variables to account for factors that may influence patient's choices of doctors in the digital context. The shared experiences of patients regarding a doctor's treatment [39], as well as the number of patients who have previously consulted with the doctors [17,62], can indicate the doctor's overall popularity. This, in turn, may affect patient choice. Therefore, we controlled for (1) the total number of patients who consulted with doctor i in the digital context before month t ($TotalPatients_{it}$) and (2) the total number of patient-shared experiences about offline treatment by doctor i before month t ($TotalExperiences_{it}$). Furthermore, doctors' past behaviors, including article publishing and live streaming, can influence their current practices in posting articles and conducting live streams. Simultaneously, these factors may also act as signals affecting patients' judgments and selection of doctors [12]. To account for these influences, we also controlled for (1) the total number of articles posted by doctor i before month t ($TotalArticles_{it}$) and (2) the total number of medical live streams conducted by doctor i before month t ($TotalLiveStreaming_{it}$).

To control for both observed and unobserved doctor-specific factors that do not change over time, individual-fixed effects were added. Additionally, time-fixed effects were introduced into our analysis to account for both observed and unobserved factors that vary over time but remain constant across doctors. Table 1 shows the variables and their definitions.

XSL•FO
RenderX

Table 1. Variable definition.

Variables	Definition
Dependent variable	
$Consultation_{it}$	The number of e-consultations of doctor *i* in month *t*.
Independent variables	
$Articles_{it}$	The number of professional articles posted by doctor *i* in month *t*.
$LiveStreaming_{it}$	The number of medical live streams conducted by doctor *i* in month *t*.
Moderating variables	
$Title_i$	A dummy variable indicating whether doctor *i* is a chief doctor (takes the value 1 for a chief doctor and 0 for doctors with lower-ranked titles).
$Recommendations_i$	The digital recommendation level of doctor *i* by other patients.
Control variables	
$TotalPatients_{it}$	The total number of patients who consulted doctor *i* in the digital context before month *t*.
$TotalExperiences_{it}$	The total number of patient-shared experiences about offline treatment by doctor *i* before month *t*.
$TotalArticles_{it}$	The total number of articles posted by doctor *i* before month *t*.
$TotalLiveStreaming_{it}$	The total number of medical live streams doctor *i* conducted before month *t*.

Estimation Model

To estimate the direct impact of doctors' proactive behaviors and reactive behaviors on their e-consultation volume, the following 2-way fixed effects regression model was used:

$$Consultation_{it} = \beta_0 + \beta_1 Articles_{it} + \beta_2 LiveStreaming_{it} + \beta_3 TotalPatients_{it} + \beta_4 TotalExperiences_{it} + \beta_5 TotalArticles_{it} + \beta_6 TotalLiveStreaming_{it} + \alpha_i + \delta_t + \mu_{it} \quad \textbf{(1)}$$

where *i* denotes doctor, *t* denotes month, α_i is doctor-fixed effects, δ_t is month-fixed effects, $Consultation_{it}$ is the number of e-consultations of doctor *i* in month *t*, $Articles_{it}$ is the number of professional articles posted by doctor *i* in month *t*, $LiveStreaming_{it}$ is the number of medical live streams conducted by doctor *i* in month *t*, $TotalPatients_{it}$ is the total number of patients who consulted doctor *i* in the digital context before month *t*, $TotalExperiences_{it}$ is the total number of patient-shared experiences about offline treatment by doctor *i* before month *t*, $TotalArticles_{it}$ is the total number of articles posted by doctor *i* before month *t*, $TotalLiveStreaming_{it}$ is the total number of medical live streams doctor *i* conducted before month *t*, β is the coefficient, and μ_{it} is the error term. We took the log transformation for our continuous variables in the model to reduce the skewness of the variables [63].

Next, the moderating effects of doctors' offline professional titles and digital recommendations by patients were investigated based on the following specification:

$$Consultation_{it} = \beta_0 + \beta_1 Articles_{it} + \beta_2 LiveStreaming_{it} + \beta_3 Articles_{it} \times Title_i + \beta_4 LiveStreaming_{it} \times Title_i + \beta_5 Articles_{it} \times Recommendation_i + \beta_6 LiveStreaming_{it} \times Recommendation_i + \beta_7 TotalPatients_{it} + \beta_8 TotalExperiences_{it} + \beta_9 TotalArticles_{it} + \beta_{10} TotalLiveStreaming_{it} + \alpha_i + \delta_t + \mu_{it} \quad \textbf{(2)}$$

where $Title_i$ indicates whether doctor *i* is a chief doctor ($Title_i$=1 indicates the doctor is a chief doctor, and $Title_i$=0 indicates the doctor has a lower-ranked title). $Recommendations_i$ is the digital recommendation level of doctor *i* by other patients.

Ethical Considerations

This study used secondary publicly available data obtained from a website and did not involve the collection of original data pertaining to human participants. As such, there is no evidence of unethical behavior in the study. Consequently, ethics approval by an ethics committee or institutional review board was not deemed necessary.

Results

Overview

In this section, we present our empirical results. The descriptive statistics are shown in Table 2, and the correlation matrix is shown in Table 3.

XSL•FO
RenderX

Table 2. Descriptive statistics.

Variable	Mean (SD)	Minimum-maximum
Consultation	17.529 (41.172)	0-1148
Articles	1.478 (8.385)	0-658
LiveStreaming	0.196 (0.928)	0-39
Title	0.413 (0.492)	0-1
Recommendations	3.849 (0.518)	1.5-5
TotalPatients	1469.475 (3237.306)	0-72,559
TotalExperiences	209.578 (363.915)	0-4626
TotalArticles	59.552 (203.213)	0-5795
TotalLiveStreaming	2.313 (7.171)	0-163

Table 3. Correlation matrix.

Variable	*Consultation*	*Articles*	*LiveStreaming*	*Title*	*Recommendations*	*TotalPatients*	*TotalExperiences*	*TotalArticles*	*TotalLiveStreaming*
Consultation									
r	1	0.177	0.146	0.132	0.561	0.629	0.628	0.355	0.192
P value	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Articles									
r	0.177	1	0.39	0.018	0.156	0.097	0.117	0.353	0.233
P value	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001
LiveStreaming									
r	0.146	0.390	1	0.006	0.072	0.066	0.071	0.176	0.348
P value	<.001	<.001	—	.14	<.001	<.001	<.001	<.001	<.001
Title									
r	0.132	0.018	0.006	1	0.203	0.238	0.248	0.174	0.014
P value	<.001	<.001	.14	—	<.001	<.001	<.001	<.001	<.001
Recommendations									
r	0.561	0.156	0.072	0.203	1	0.42	0.648	0.304	0.091
P value	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001
TotalPatients									
r	0.629	0.097	0.066	0.238	0.420	1	0.774	0.569	0.245
P value	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001
TotalExperiences									
r	0.628	0.117	0.071	0.248	0.648	0.774	1	0.513	0.208
P value	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001
TotalArticles									
r	0.355	0.353	0.176	0.174	0.304	0.569	0.513	1	0.415
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001
TotalLiveStreaming									
r	0.192	0.233	0.348	0.014	0.091	0.245	0.208	0.415	1
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

XSL•FO
RenderX

Empirical Results

Results for Direct Effects

The analysis was conducted progressively. We first estimated the equation without control variables (model 1) and then added control variables in model 2. The estimated results are shown in Table 4. From the results, we can see that the coefficient of *Articles* is significant and positive in model 2 (β=.093; *P*<.001), indicating that doctors' proactive behaviors (ie, posting professional articles) can help them obtain more e-consultations. Thus, hypothesis 1 is supported. Regarding doctors' engagement in medical live streaming, the results show that the coefficient of *LiveStreaming* is significantly positive (β=.214; *P*<.001), which suggests that doctors' reactive behaviors (ie, conducting medical live streaming) can increase their e-consultation volume. This supports hypothesis 2.

Table 4. The effects of doctors' proactive behaviors and reactive behaviors on their e-consultation volume[a].

Variables	e-Consultation volume: *Consultation*			
	Model 1[b]		Model 2[c]	
	β (SE)	*P* value	β (SE)	*P* value
Articles	.096 (0.008)	<.001	.093 (0.008)	<.001
LiveStreaming	.235 (0.016)	<.001	.214 (0.016)	<.001
TotalPatients	N/A[d]	N/A	.195 (0.027)	<.001
TotalExperiences	N/A	N/A	.256 (0.037)	<.001
TotalArticles	N/A	N/A	.073 (0.014)	<.001
TotalLiveStreaming	N/A	N/A	–.003 (0.015)	.81
Constant	1.704 (0.003)	<.001	–.617 (0.120)	<.001

[a]All models include doctor-fixed effects and month-fixed effects; robust SEs clustered by doctors are reported; the number of doctors is 2880, and the number of observations is 63,360.

[b]R^2=0.843; $F_{2,2879}$=175.98; *P*<.001.

[c]R^2=0.851; $F_{6,2879}$=119.72; *P*<.001.

[d]N/A: not applicable.

Results for Moderating Effects

The results for moderating effects are shown in Table 5. In model 1, interaction terms were initially introduced between *Title* and *Articles*, as well as between *Title* and *LiveStreaming*, to estimate the moderating effect of doctors' offline professional titles. The interaction terms were then added between *Recommendations* and *Articles*, as well as between *Recommendations* and *LiveStreaming*, to estimate the moderating effect of doctors' digital recommendations in model 2. Finally, a full model was estimated by incorporating all interaction terms. We find that the results are consistent across all models. Wald tests and likelihood ratio were used to compare the fit among nested models [64,65], and the results show that the inclusion of moderating variables significantly enhances the model's fit.

Regarding the moderating effect of doctors' offline professional titles, we find that the coefficient of *Articles*×*Title* in model 1 of Table 5 is significantly negative (β=–.058; *P*<.001), which supports hypothesis 3a that doctors' offline professional titles have a negative moderating effect on the relationship between doctors' proactive behaviors and e-consultation volume. However, the coefficient of *LiveStreaming*×*Title* is insignificant (β=–.024; *P*=.45), which suggests that doctors' offline professional titles have no moderating effect on the relationship between doctors' reactive behaviors and e-consultation volume. Thus, hypothesis 4a is not supported.

XSL•FO
RenderX

Table 5. The moderating effects of doctors' offline professional titles and digital patient recommendations[a].

Variables	e-Consultation volume: *Consultation*					
	Model 1[b]		Model 2[c]		Model 3[d]	
	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value
Articles	.117 (0.011)	<.001	.313 (0.057)	<.001	.303 (0.057)	<.001
LiveStreaming	.223 (0.022)	<.001	.603 (0.114)	<.001	.601 (0.112)	<.001
Articles×Title	–.058 (0.016)	<.001	N/A[e]	N/A	–.047 (0.016)	.003
LiveStreaming×Title	–.024 (0.031)	.45	N/A	N/A	.007 (0.031)	.83
Articles×Recommendations	N/A	N/A	–.055 (0.014)	<.001	–.048 (0.014)	<.001
LiveStreaming×Recommendations	N/A	N/A	–.100 (0.028)	<.001	–.100 (0.027)	<.001
TotalPatients	.193 (0.027)	<.001	.193 (0.027)	<.001	.193 (0.027)	<.001
TotalExperiences	.259 (0.037)	<.001	.262 (0.037)	<.001	.263 (0.037)	<.001
TotalArticles	.071 (0.014)	<.001	.072 (0.014)	<.001	.071 (0.014)	<.001
TotalLiveStreaming	–.002 (0.015)	.87	–.002 (0.015)	.88	–.002 (0.015)	.91
Constant	–.614 (0.121)	<.001	–.628 (0.121)	<.001	–.625 (0.121)	<.001

[a]All models include doctor-fixed effects and month-fixed effects; robust SEs clustered by doctors are reported; the number of doctors is 2880, and the number of observations is 63,360.

[b]R^2=0.851; $F_{8,2879}$=89.98; *P*<.001; Wald test: *P*<.001; likelihood ratio: *P*<.001.

[c]R^2=0.851; $F_{8,2879}$=89.13; *P*<.001; Wald test: *P*<.001; likelihood ratio: *P*<.001.

[d]R^2=0.852; $F_{10,2879}$=71.44; *P*<.001; Wald test: *P*<.001; likelihood ratio: *P*<.001.

[e]N/A: not applicable.

For the moderating effect of digital patient recommendations, we find that both of the coefficients of *Articles ×Recommendations* and *LiveStreaming×Recommendations* are negative and significant (β=–.055; *P*<.001 and β=–.100; *P*<.001, respectively, in model 2 of Table 5). This indicates that digital recommendations from patients have negative moderating effects on the relationship between doctors' proactive behaviors and e-consultation volume as well as on the relationship between doctors' reactive behaviors and e-consultation volume; this finding supports hypotheses 3b and 4b.

Robustness Check

First, additional analysis was performed to check whether our findings are robust to different measures of doctors' reactive behaviors. In the main analysis, we used the number of medical live streams to construct doctors' reactive behaviors. In the robustness check, doctors' reactive behaviors were measured using the following measures: (1) the length of time spent in medical live streaming ($LSDuration_{it}$), which is calculated as the total duration of all medical live streams conducted by doctor *i* in month *t*; and (2) the number of doctor-patient interactions in the medical live streams ($LSInteractions_{it}$), which is calculated as the total number of interactions between doctor *i* and patients in medical live streams in month *t*. This measure is likely to more effectively capture the reactive element of the behavior. The estimated results are shown in Table 6, and we can see that the results are consistent with the main results.

Second, in the above analysis, the total number of articles posted by the doctors was used to measure doctors' proactive behaviors. As doctors can post articles that are either their own original work or reposts from others, we further used the number of original articles ($OriArticles_{it}$) to measure doctors' proactive prosocial behaviors. Specifically, the number of articles was replaced with the number of original articles posted by doctor *i* in month *t* ($OriArticles_{it}$). Models 1 and 2 in Table 7 show the results. We can see that using this alternative measure of proactive behavior does not materially change the results.

Third, as our dependent variable takes nonnegative values, negative binomial regression was further used to re-estimate our models. We find that the results (models 3 and 4 in Table 7) are similar to the main results.

Fourth, to further enhance the robustness and validity of our findings, article quality was used as a measure of doctors' proactive behaviors. This approach is based on the premise that article quality more accurately reflects the effort and time invested by doctors in content creation. Specifically, we assessed article quality based on either the length of each article or the number of likes it received and then re-estimated our model. As indicated in Table 8, the results remain consistent with our main findings, thereby further reinforcing the validity of our conclusions.

XSL•FO
RenderX

Table 6. Robustness checks: alternative measure of reactive behaviors[a].

Variables	e-Consultation volume: *Consultation*							
	Use *LSDuration* to measure reactive behaviors				Use *LSInteractions* to measure reactive behaviors			
	Model 1[b]		Model 2[c]		Model 3[d]		Model 4[e]	
	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value
Articles	.095 (0.008)	<.001	.304 (0.058)	<.001	.103 (0.008)	<.001	.351 (0.059)	<.001
LSDuration	.023 (0.002)	<.001	.071 (0.011)	<.001	N/A[f]	N/A	N/A	N/A
LSInteractions	N/A	N/A	N/A	N/A	.041 (0.004)	<.001	.114 (0.035)	.001
Articles×Title	N/A	N/A	–.045 (0.016)	.004	N/A	N/A	–.049 (0.016)	.002
LSDuration×Title	N/A	N/A	–.000 (0.003)	.96	N/A	N/A	N/A	N/A
LSInteractions×Title	N/A	N/A	N/A	N/A	N/A	N/A	.003 (0.009)	.69
Articles×Recommendations	N/A	N/A	–.048 (0.014)	.001	N/A	N/A	–.057 (0.014)	<.001
LSDuration×Recommendations	N/A	N/A	–.012 (0.003)	<.001	N/A	N/A	N/A	N/A
LSInteractions×Recommendations	N/A	N/A	N/A	N/A	N/A	N/A	–.019 (0.008)	.02
TotalPatients	.194 (0.027)	<.001	.192 (0.027)	<.001	.194 (0.027)	<.001	.191 (0.027)	<.001
TotalExperiences	.256 (0.037)	<.001	.263 (0.037)	<.001	.258 (0.038)	<.001	.266 (0.038)	<.001
TotalArticles	.073 (0.014)	<.001	.071 (0.014)	<.001	.076 (0.014)	<.001	.074 (0.014)	<.001
TotalLiveStreaming	.005 (0.015)	.71	.008 (0.015)	.60	–.005 (0.015)	.73	–.004 (0.015)	.80
Constant	–.617 (0.120)	<.001	–.626 (0.121)	<.001	–.621 (0.120)	<.001	–.628 (0.122)	<.001

[a]All models include doctor-fixed effects and month-fixed effects; robust SEs clustered by doctors are reported; the number of doctors is 2880, and the number of observations is 63,360.

[b]R^2=0.851; $F_{6,2879}$=131.71; P<.001.

[c]R^2=0.851; $F_{10,2879}$=79.29; P<.001; Wald test: P<.001; likelihood ratio: P<.001.

[d]R^2=0.850; $F_{6,2879}$=112.52; P<.001.

[e]R^2=0.851; $F_{10,2879}$=68.43; P<.001; Wald test: P<.001; likelihood ratio: P<.001.

[f]N/A: not applicable.

XSL•FO
RenderX

Table 7. Robustness checks: alternative measure of proactive behaviors and negative binomial regression[a].

Variables	e-Consultation volume: *Consultation*							
	Use *OriArticles* to measure proactive behaviors				Negative binomial regression			
	Model 1[b]		Model 2[c]		Model 3[d]		Model 4[e]	
	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value
Articles	N/A[f]	N/A	N/A	N/A	.040 (0.008)	<.001	.277 (0.058)	<.001
OriArticles	.093 (0.008)	<.001	.290 (0.057)	<.001	N/A	N/A	N/A	N/A
LiveStreaming	.215 (0.016)	<.001	.626 (0.111)	<.001	.166 (0.013)	<.001	.741 (0.150)	<.001
Articles×Title	N/A	N/A	N/A	N/A	N/A	N/A	–.040 (0.016)	.010
OriArticles×Title	N/A	N/A	–.051 (0.016)	.001	N/A	N/A	N/A	N/A
LiveStreaming×Title	N/A	N/A	.008 (0.031)	.80	N/A	N/A	–.026 (0.031)	.40
Articles×Recommendations	N/A	N/A	N/A	N/A	N/A	N/A	–.050 (0.013)	<.001
OriArticles×Recommendations	N/A	N/A	–.044 (0.014)	.001	N/A	N/A	N/A	N/A
LiveStreaming×Recommendations	N/A	N/A	–.106 (0.027)	<.001	N/A	N/A	–.133 (0.036)	<.001
TotalPatients	.195 (0.027)	<.001	.192 (0.027)	<.001	.067 (0.023)	.003	.064 (0.023)	.006
TotalExperiences	.256 (0.037)	<.001	.263 (0.037)	<.001	.401 (0.026)	<.001	.412 (0.026)	<.001
TotalArticles	.073 (0.014)	<.001	.070 (0.014)	<.001	–.011 (0.018)	.55	–.015 (0.015)	.32
TotalLiveStreaming	–.003 (0.015)	.84	–.001 (0.015)	.94	.020 (0.016)	.21	.020 (0.019)	.31
Constant	–.614 (0.120)	<.001	–.621 (0.121)	<.001	–1.174 (0.076)	<.001	–1.196 (0.091)	<.001

[a]All models include doctor-fixed effects and month-fixed effects; robust SEs clustered by doctors are reported in models 1 and 2; bootstrap SEs in models 3 and 4.

[b]R^2=0.851; $F_{6,2879}$=118.99; P<.001.

[c]R^2=0.852; $F_{10,2879}$=71.04; P<.001; Wald test: P<.001; likelihood ratio: P<.001.

[d]Log likelihood=–150,015.36.

[e]Log likelihood=–149,888.24.

[f]N/A: not applicable.

XSL•FO
RenderX

Table 8. Robustness checks: using article quality to measure proactive behavior[a].

Variables	e-Consultation volume: *Consultation*							
	Using the number of likes received by each article to measure article quality				Using the length of each article to measure article quality			
	Model 1[b]		Model 2[c]		Model 1[d]		Model 2[e]	
	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value	β (SE)	*P* value
ArticleQuality	.084 (0.007)	<.001	.539 (0.060)	<.001	.029 (0.002)	<.001	.117 (0.016)	<.001
LiveStreaming	.265 (0.016)	<.001	.770 (0.121)	<.001	.241 (0.016)	<.001	.642 (0.117)	<.001
ArticleQuality×Title	N/A[f]	N/A	–.059 (0.013)	<.001	N/A	N/A	–.016 (0.004)	<.001
LiveStreaming×Title	N/A	N/A	–.021 (0.032)	.51	N/A	N/A	–.006 (0.032)	.85
ArticleQuality×Recommendations	N/A	N/A	–.099 (0.014)	<.001	N/A	N/A	–.020 (0.004)	<.001
LiveStreaming×Recommendations	N/A	N/A	–.127 (0.029)	<.001	N/A	N/A	–.103 (0.028)	<.001
TotalPatients	.194 (0.027)	<.001	.195 (0.027)	<.001	.194 (0.027)	<.001	.192 (0.027)	<.001
TotalExperiences	.256 (0.037)	<.001	.259 (0.037)	<.001	.257 (0.037)	<.001	.263 (0.037)	<.001
TotalArticles	.077 (0.014)	<.001	.077 (0.014)	<.001	.075 (0.014)	<.001	.074 (0.014)	<.001
TotalLiveStreaming	.004 (0.015)	.80	.006 (0.015)	.66	.002 (0.014)	.88	.004 (0.015)	.77
Constant	–.622 (0.120)	<.001	–.639 (0.121)	<.001	–.627 (0.120)	<.001	–.638 (0.121)	<.001

[a]All models include doctor-fixed effects and month-fixed effects; robust SEs clustered by doctors are reported; the number of doctors is 2880, and the number of observations is 63,360.

[b]R^2=0.851; $F_{6,2879}$=127.75; P<.001.

[c]R^2=0.851; $F_{10,2879}$=89.97; P<.001; Wald test: P<.001; likelihood ratio: P<.001.

[d]R^2=0.851; $F_{6,2879}$=133.39; P<.001.

[e]R^2=0.852; $F_{10,2879}$=84.94; P<.001; Wald test: P<.001; likelihood ratio: P<.001.

[f]N/A: not applicable.

Discussion

Analysis of Results

Web-based medical platforms offer a variety of functions to support doctors' engagement in different types of prosocial behaviors. However, few studies have investigated the effects of these behaviors. Drawing on FMT and SET, this study categorized doctors' prosocial practices in web-based health care communities into proactive and reactive actions and examined their effects on e-consultation volume. Briefly, prosocial behaviors positively impact on e-consultation, and a doctor's digital and offline reputation moderates the relationship between prosocial behavior and e-consultation, albeit with some nuances.

First, we expanded upon existing literature on proactive prosocial behaviors, concluding that these actions can help doctors create professional images [43] in the medical consultation context. Our panel data analysis reveals that doctors' posting of professional articles, which contribute to their professional image in the digital context, attracts more e-consultations. This finding aligns with the prior study [31], which observed that a health professional's previous asynchronous prosocial behavior positively influences their future economic performance.

Second, drawing from SET, we analyzed the impact of synchronous reactive prosocial behaviors, a less explored area in prior literature. Our findings confirm that engaging in medical live streaming, a form of reactive prosocial behavior, leads to higher e-consultation volumes. Interestingly, we found that the positive impact of conducting a live stream exceeds that of posting an article.

Third, we expanded our research by testing the moderating roles of digital and offline reputations, measured by doctors' offline professional titles and patients' recommendations on web-based health care communities. We found that digital reputations significantly moderate the relationships between both types of prosocial behaviors and e-consultation volume. Specifically, doctors who post professional articles or conduct medical live streams attract more e-consultations when they have fewer patient recommendations compared to those with higher recommendations. Regarding offline professional titles, our results indicate a significant moderating effect on the relationship between proactive prosocial behaviors and e-consultation volume. Notably, junior doctors should focus more on posting articles in web-based health care communities to compensate for limitations associated with their titles [66]. However, the moderating effect of offline titles on the impact of reactive prosocial behaviors was found to be insignificant. We attribute this to the unique dynamics of trust conversion in Chinese health care settings. As doctors' offline titles are granted by medical institutions, these titles could enhance patients' trust in doctors only if there is a conversion of trust from the organization to the individual doctor, which represents different

XSL•FO
RenderX

types of trust [67]. Consequently, doctors with the same offline titles from different hospitals may be perceived differently. For example, a senior doctor from a 3-A hospital is usually seen as highly professional in their clinical field, while a doctor with the same title in a 1-A hospital might typically handle primary diseases. Due to this trust conversion phenomenon, patients may not uniformly trust doctors from different hospitals with the same offline titles, leading to the insignificant moderating effect of offline titles on the impact of reactive prosocial behaviors.

In summary, this study underscores the importance of prosocial behaviors and reputation in shaping doctors' e-consultation volumes on web-based health care communities, offering valuable insights for health care professionals aiming to increase their consultation outreach.

Implications

This study makes several theoretical implications. First, this study contributes to web-based health care community literature by offering a nuanced understanding of how doctors' prosocial behaviors enhance e-consultation volume. While a limited number of studies have examined the effects of doctors' freely provided behaviors in the digital context [31], the specific impact of different types of prosocial behaviors on e-consultation volume remains largely unexplored. This study addresses this knowledge gap by theoretically categorizing doctors' prosocial behaviors in web-based health care communities into proactive and reactive types and exploring their impacts on e-consultations.

Second, this study enriches web-based health care communities and live streaming literature by validating the role of medical live streaming in web-based health services. Prior research on live streaming has mainly concentrated on e-commerce [68], web-based gaming [69], and web-based learning [70]. Our study extends this research to the health care context, highlighting the importance of live streaming on web-based health care platforms. Specifically, this study delves into how doctors' synchronous, reactive volunteer interactions via live streaming influence patient decision-making.

Finally, this study advances FMT and SET by highlighting the importance of context in theory development and providing guidance for context-specific theorizing on web-based health platforms. It also sheds light on how the impact of different prosocial behaviors on e-consultation volume varies depending on a doctor's offline and digital reputations. Notably, this study validates that proactive behaviors work more effectively in promoting e-consultations for doctors with lower titles or fewer digital recommendations, while reactive behaviors are more effective for doctors with fewer digital recommendations.

This study offers several practical implications for doctors and platform managers. First, the beneficial effects of prosocial behaviors suggest that doctors should adapt their engagement activities when participating in web-based health care platforms. Nowadays, an increasing number of doctors are joining web-based health care communities and focusing on e-consultations, attracted by the economic and social benefits. Based on our results, posting professional articles can help doctors establish a professional image, potentially leading to more e-consultations. Additionally, conducting medical live streams can bolster e-consultations by fostering cooperative social value for doctors and enhancing their credibility among patient audiences. Therefore, doctors may prefer engaging in both proactive and reactive prosocial activities in web-based health care communities to attract more patients to their e-consultation services.

Second, the boundary conditions of the effects of prosocial behaviors imply that doctors should strategically leverage the beneficial effect of proactive and reactive behaviors according to their offline and digital reputations. Doctors with fewer digital recommendations should focus more on prosocial behavior to attract patients to e-consultations. Meanwhile, doctors with lower titles should devote their efforts to proactive behaviors to demonstrate their capability in fulfilling the e-consultations, thereby reducing information asymmetry between patients and themselves.

Third, our findings offer implications for web-based health care platform managers in designing effective functions. An increasing number of platforms are launching various features to better serve doctors and patients, meeting the needs of both groups more effectively. Our empirical findings suggest that doctors' proactive and reactive prosocial behaviors, such as posting professional articles and conducting medical live streams, can help them establish professional image and enhance patient trust, leading to improved performance. Importantly, these behaviors also benefit patients by enhancing their health knowledge and literacy. Thus, platform managers could introduce functions (eg, article posting, live streaming, and doctor-driven communities) to encourage more prosocial behaviors by doctors. Additionally, platform managers might consider incorporating guidelines or incentive mechanisms for prosocial behaviors into their platforms. For example, it is recommended that platforms collect and analyze doctors' proactive and reactive prosocial behaviors and guide them on how to effectively use these functions and engage in different types of activities.

Limitations

Despite its contributions, this study also presents several limitations that future research should consider. First, various classifications of prosocial behavior are available; for instance, Richaud et al [71] classified such behavior as altruistic, compliant, emotional, public, anonymous, or dire actions. Given the intricacy of web-based medical services, future studies would benefit from further exploring the roles of these other types of prosocial behavior exhibited by doctors on web-based health care communities. Second, our research model was constructed primarily from the doctor's perspective and thus did not investigate the influence of doctors' prosocial behaviors on patients' satisfaction and well-being. Future research should delve into these relationships to obtain a more comprehensive understanding of the impacts of doctors' prosocial behaviors. Finally, this study focused only on the quantity of medical live-streaming sessions, overlooking the quality aspect, which could be a crucial factor influencing e-consultation volume. Future research will concentrate on exploring this aspect.

XSL•FO
RenderX

Conclusions

Building upon prior studies on doctors' prosocial behaviors on web-based health care communities, this study further delineates doctors' beneficial actions into proactive and synchronous reactive behaviors. This distinction is based on the divergence in doctors' motives for engaging and patients' levels of involvement. Drawing from FMT and SET, this study offers insights that could aid doctors in increasing their e-consultation volume by adopting these beneficial behaviors. Concurrently, this research augments our understanding of the roles a doctor's reputation plays in the relationships between various prosocial behaviors—specifically, proactive and reactive actions—and their e-consultation volume. This study may inspire doctors with comparatively lower offline professional titles and digital popularity to achieve their desired e-consultation volume.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72001170 and 72102179), the Fundamental Research Funds for the Central Universities (SK2024028), the Ministry of Education in China Project of Humanities and Social Sciences (21XJC630003), the China Postdoctoral Science Foundation (2022T150515, 2023M742818, and 2020M673432), the National Natural Science Foundation of China (72004042), and the Heilongjiang Natural Science Foundation (YQ2023G003), and the grants from City University of Hong Kong (projects 7005959, 7006152, and 7200725).

Conflicts of Interest

None declared.

References

1. Qi M, Cui J, Li X, Han Y. Influence of e-consultation on the intention of first-visit patients to select medical services: results of a scenario survey. J Med Internet Res 2023;25:e40993 [FREE Full text] [doi: 10.2196/40993] [Medline: 37115615]
2. Cao B, Huang W, Chao N, Yang G, Luo N. Patient activeness during online medical consultation in China: multilevel analysis. J Med Internet Res 2022;24(5):e35557 [FREE Full text] [doi: 10.2196/35557] [Medline: 35622403]
3. Jiang S. Talk to your doctors online: an internet-based intervention in China. Health Commun 2021;36(4):405-411 [FREE Full text] [doi: 10.1080/10410236.2019.1692493] [Medline: 31736365]
4. Jiang J, Cameron AF, Yang M. Analysis of massive online medical consultation service data to understand physicians' economic return: observational data mining study. JMIR Med Inform 2020;8(2):e16765 [FREE Full text] [doi: 10.2196/16765] [Medline: 32069213]
5. Street RL, Millay B. Analyzing patient participation in medical encounters. Health Commun 2001;13(1):61-73 [FREE Full text] [doi: 10.1207/S15327027HC1301_06] [Medline: 11370924]
6. Ren D, Ma B. Effectiveness of interactive tools in online health care communities: social exchange theory perspective. J Med Internet Res 2021;23(3):e21892 [FREE Full text] [doi: 10.2196/21892] [Medline: 33709940]
7. Guo S, Guo X, Fang Y, Vogel D. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. J Manag Inf Syst 2017;34(2):487-519 [FREE Full text] [doi: 10.1080/07421222.2017.1334480]
8. Tande AJ, Berbari EF, Ramar P, Ponamgi SP, Sharma U, Philpot L, et al. Association of a remotely offered infectious diseases eConsult service with improved clinical outcomes. Open Forum Infect Dis 2020;7(1):ofaa003 [FREE Full text] [doi: 10.1093/ofid/ofaa003] [Medline: 31988969]
9. Richardson BR, Truter P, Blumke R, Russell TG. Physiotherapy assessment and diagnosis of musculoskeletal disorders of the knee via telerehabilitation. J Telemed Telecare 2017;23(1):88-95. [doi: 10.1177/1357633X15627237] [Medline: 26985005]
10. Castaneda PR, Duffy B, Andraska EA, Stevens J, Reschke K, Osborne N, et al. Outcomes and safety of electronic consult use in vascular surgery. J Vasc Surg 2020;71(5):1726-1732 [FREE Full text] [doi: 10.1016/j.jvs.2019.08.231] [Medline: 31611112]
11. Jiang S. The relationship between face-to-face and online patient-provider communication: examining the moderating roles of patient trust and patient satisfaction. Health Commun 2020;35(3):341-349 [FREE Full text] [doi: 10.1080/10410236.2018.1563030] [Medline: 30821506]
12. Li J, Tang J, Jiang L, Yen DC, Liu X. Economic success of physicians in the online consultation market: a signaling theory perspective. Int J Electron Commer 2019;23(2):244-271 [FREE Full text] [doi: 10.1080/10864415.2018.1564552]
13. Guo S, Guo X, Zhang X, Vogel D. Doctor–patient relationship strength's impact in an online healthcare community. Inf Technol Dev 2017;24(2):279-300 [FREE Full text] [doi: 10.1080/02681102.2017.1283287]
14. Audrain-Pontevia AF, Menvielle L. Do online health communities enhance patient-physician relationship? an assessment of the impact of social support and patient empowerment. Health Serv Manage Res 2018;31(3):154-162 [FREE Full text] [doi: 10.1177/0951484817748462] [Medline: 29280679]
15. Yang Y, Zhang X, Lee PK. Improving the effectiveness of online healthcare platforms: an empirical study with multi-period patient-doctor consultation data. Int J Prod Econ 2019;207:70-80 [FREE Full text] [doi: 10.1016/j.ijpe.2018.11.009]

XSL•FO
RenderX

16. Fan W, Zhou Q, Qiu L, Kumar S. Should doctors open online consultation services? An empirical investigation of their impact on offline appointments. Inf Syst Res 2023;34(2):629-651 [FREE Full text] [doi: 10.1287/isre.2022.1145]
17. Liu QB, Liu X, Guo X. The effects of participating in a physician-driven online health community in managing chronic disease: evidence from two natural experiments. MIS Q 2020;44(1):391-419 [FREE Full text] [doi: 10.25300/misq/2020/15102]
18. Yan Z, Wang T, Chen Y, Zhang H. Knowledge sharing in online health communities: a social exchange theory perspective. Inf Manag 2016;53(5):643-653 [FREE Full text] [doi: 10.1016/j.im.2016.02.001]
19. Luo P, Chen K, Wu C, Li Y. Exploring the social influence of multichannel access in an online health community. J Assoc Inf Sci Technol 2018;69(1):98-109 [FREE Full text] [doi: 10.1002/asi.23928]
20. Chai S, Bagchi-Sen S, Morrell C, Rao HR, Upadhyaya SJ. Internet and online information privacy: an exploratory study of preteens and early teens. IEEE Trans Profess Commun 2009;52(2):167-182 [FREE Full text] [doi: 10.1109/tpc.2009.2017985]
21. Ahearne M, Rapp A, Hughes DE, Jindal R. Managing sales force product perceptions and control systems in the success of new product introductions. J Mark Res 2010;47(4):764-776 [FREE Full text] [doi: 10.1509/jmkr.47.4.764]
22. Awaysheh A, Heron RA, Perry T, Wilson JI. On the relation between corporate social responsibility and financial performance. Strateg Manag J 2020;41(6):965-987 [FREE Full text] [doi: 10.1002/smj.3122]
23. Coll MP, Grégoire M, Eugène F, Jackson PL. Neural correlates of prosocial behavior towards persons in pain in healthcare providers. Biol Psychol 2017;128:1-10 [FREE Full text] [doi: 10.1016/j.biopsycho.2017.06.005] [Medline: 28669784]
24. Kapıkıran NA. Sources of ethnocultural empathy: personality, intergroup relations, affects. Curr Psychol 2021;42(14):11510-11528 [FREE Full text] [doi: 10.1007/s12144-021-02286-2]
25. Yin Y, Wang Y. Is empathy associated with more prosocial behaviour? a meta - analysis. Asian J Soc Psychol 2023;26(1):3-22 [FREE Full text] [doi: 10.1111/ajsp.12537]
26. Finset A. "I am worried, Doctor!" emotions in the doctor-patient relationship. Patient Educ Couns 2012;88(3):359-363 [FREE Full text] [doi: 10.1016/j.pec.2012.06.022] [Medline: 22819270]
27. Larson EB, Yao X. Clinical empathy as emotional labor in the patient-physician relationship. JAMA 2005;293(9):1100-1106 [FREE Full text] [doi: 10.1001/jama.293.9.1100] [Medline: 15741532]
28. Bloom G, Standing H, Lloyd R. Markets, information asymmetry and health care: towards new social contracts. Soc Sci Med 2008;66(10):2076-2087 [FREE Full text] [doi: 10.1016/j.socscimed.2008.01.034] [Medline: 18316147]
29. Guo F, Zou B, Guo J, Shi Y, Bo Q, Shi L. What determines academic entrepreneurship success? A social identity perspective. Int Entrep Manag J 2019;15(3):929-952 [FREE Full text] [doi: 10.1007/s11365-019-00569-6]
30. Yang H, Zhang X. Investigating the effect of paid and free feedback about physicians' telemedicine services on patients' and physicians' behaviors: panel data analysis. J Med Internet Res 2019;21(3):e12156 [FREE Full text] [doi: 10.2196/12156] [Medline: 30900997]
31. Yan Z, Kuang L, Qiu L. Prosocial behaviors and economic performance: evidence from an online mental healthcare platform. Prod Oper Manag 2022;31(10):3859-3876 [FREE Full text] [doi: 10.1111/poms.13792]
32. Frese M, Fay D. 4. Personal initiative: an active performance concept for work in the 21st century. Res Organ Behav 2001;23:133-187 [FREE Full text] [doi: 10.1016/s0191-3085(01)23005-6]
33. Frese M, Kring W, Soose A, Zempel J. Personal initiative at work: differences between East and West Germany. Acad Manage J 1996;39(1):37-63 [FREE Full text] [doi: 10.5465/256630]
34. Spitzmuller M, Van Dyne L. Proactive and reactive helping: contrasting the positive consequences of different forms of helping. J Organ Behavior 2013;34(4):560-580 [FREE Full text] [doi: 10.1002/job.1848]
35. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52(1):1-26 [FREE Full text] [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
36. Mittal S, Sengupta A, Agrawal NM, Gupta S. How prosocial is proactive: developing and validating a scale and process model of knowledge-based proactive helping. J Manag Organ 2018;26(4):625-650 [FREE Full text] [doi: 10.1017/jmo.2017.80]
37. Walter A, Ritter T, Gemünden HG. Value creation in buyer–seller relationships: theoretical considerations and empirical results from a supplier's perspective. Ind Mark Manag 2001;30(4):365-377 [FREE Full text] [doi: 10.1016/s0019-8501(01)00156-0]
38. Eggert A, Ulaga W, Schultz F. Value creation in the relationship life cycle: a quasi-longitudinal analysis. Ind Mark Manag 2006;35(1):20-27 [FREE Full text] [doi: 10.1016/j.indmarman.2005.07.003]
39. Deng Z, Hong Z, Zhang W, Evans R, Chen Y. The effect of online effort and reputation of physicians on patients' choice: 3-wave data analysis of China's good doctor website. J Med Internet Res 2019;21(3):e10170 [FREE Full text] [doi: 10.2196/10170] [Medline: 30848726]
40. Kurihara H, Maeno T, Maeno T. Importance of physicians' attire: factors influencing the impression it makes on patients, a cross-sectional study. Asia Pac Fam Med 2014;13(1):2 [FREE Full text] [doi: 10.1186/1447-056X-13-2] [Medline: 24397871]
41. Cooper ML, Shapiro CM, Powers AM. Motivations for sex and risky sexual behavior among adolescents and young adults: a functional perspective. J Pers Soc Psychol 1998;75(6):1528-1558. [doi: 10.1037//0022-3514.75.6.1528] [Medline: 9914665]

XSL•FO
RenderX

42. Clary EG, Snyder M, Ridge RD, Copeland J, Stukas AA, Haugen J, et al. Understanding and assessing the motivations of volunteers: a functional approach. J Pers Soc Psychol 1998;74(6):1516-1530. [doi: 10.1037//0022-3514.74.6.1516] [Medline: 9654757]
43. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287(2):226-235 [FREE Full text] [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
44. Zhang L, Qiu Y, Zhang N, Li S. How difficult doctor patient relationships impair physicians' work engagement: the roles of prosocial motivation and problem-solving pondering. Psychol Rep 2020;123(3):885-902 [FREE Full text] [doi: 10.1177/0033294119826887] [Medline: 30732545]
45. Dabholkar PA, Johnston WJ, Cathey AS. The dynamics of long-term business-to-business exchange relationships. J Acad Mark Sci 1994;22(2):130-145 [FREE Full text] [doi: 10.1177/0092070394222003]
46. Ochs M, Mestre D, de Montcheuil G, Pergandi JM, Saubesty J, Lombardo E, et al. Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. J Multimodal User Interfaces 2019;13(1):41-51 [FREE Full text] [doi: 10.1007/s12193-018-0289-8]
47. Schroeder T, Seaman K, Nguyen A, Gewald H, Georgiou A. Enablers and inhibitors to the adoption of mHealth apps by patients—a qualitative analysis of German doctors' perspectives. Patient Educ Couns 2023;114:107865 [FREE Full text] [doi: 10.1016/j.pec.2023.107865] [Medline: 37356116]
48. Liu X, Hu M, Xiao BS, Shao J. Is my doctor around me? investigating the impact of doctors' presence on patients' review behaviors on an online health platform. J Assoc Inf Sci Technol 2022;73(9):1279-1296 [FREE Full text] [doi: 10.1002/asi.24632]
49. Chen Y, Xie J. Online consumer review: word-of-mouth as a new element of marketing communication mix. Manage Sci 2008;54(3):477-491 [FREE Full text] [doi: 10.1287/mnsc.1070.0810]
50. Ye Q, Li Y, Kiang M, Wu W. The impact of seller reputation on the performance of online sales: evidence from TaoBao Buy-It-Now (BIN) data. SIGMIS Database 2009;40(1):12-19 [FREE Full text] [doi: 10.1145/1496930.1496935]
51. Liu X, Guo X, Wu H, Wu T. The impact of individual and organizational reputation on physicians' appointments online. Int J Electron Commer 2016;20(4):551-577 [FREE Full text] [doi: 10.1080/10864415.2016.1171977]
52. Dewan S, Hsu V. Adverse selection in electronic markets: evidence from online stamp auctions. J Ind Econ 2004;52(4):497-516. [doi: 10.1111/j.0022-1821.2004.00237.x]
53. Torres E, Vasquez-Parraga AZ, Barra C. The path of patient loyalty and the role of doctor reputation. Health Mark Q 2009;26(3):183-197 [FREE Full text] [doi: 10.1080/07359680903263565] [Medline: 19813122]
54. Gottschalk F, Mimra W, Waibel C. Health services as credence goods: a field experiment. Econ J 2020;130(629):1346-1383 [FREE Full text] [doi: 10.1093/ej/ueaa024]
55. Nunkoo R, Ramkissoon H. Power, trust, social exchange and community support. Ann Tour Res 2012;39(2):997-1023 [FREE Full text] [doi: 10.1016/j.annals.2011.11.017]
56. Blau PM. Exchange and Power in Social Life. New Brunswick: Transaction Publishers; 1964.
57. Molm LD, Takahashi N, Peterson G. Risk and trust in social exchange: an experimental test of a classical proposition. Am J Sociol 2000;105(5):1396-1427 [FREE Full text] [doi: 10.1086/210434]
58. Lioukas CS, Reuer JJ. Isolating trust outcomes from exchange relationships: social exchange and learning benefits of prior ties in alliances. Acad Manage J 2015;58(6):1826-1847 [FREE Full text] [doi: 10.5465/amj.2011.0934]
59. Malgonde OS, Saldanha TJ, Mithas S. Resilience in the open source software community: how pandemic and unemployment shocks influence contributions to others' and one's own projects. MIS Q 2023;47(1):361-390 [FREE Full text] [doi: 10.25300/misq/2022/17256]
60. Mayya R, Ye S, Viswanathan S, Agarwal R. Who forgoes screening in online markets and why? Evidence from Airbnb. MIS Q 2021;45(4):1745-1776 [FREE Full text] [doi: 10.25300/misq/2021/15335]
61. Moqri M, Mei X, Qiu L, Bandyopadhyay S. Effect of "following" on contributions to open source communities. J Manag Inf Syst 2018;35(4):1188-1217 [FREE Full text] [doi: 10.1080/07421222.2018.1523605]
62. Chen Q, Jin J, Zhang T, Yan X. The effects of log-in behaviors and web reviews on patient consultation in online health communities: longitudinal study. J Med Internet Res 2021;23(6):e25367 [FREE Full text] [doi: 10.2196/25367] [Medline: 34081008]
63. Mo J, Sarkar S, Menon S. Competing tasks and task quality: an empirical study of crowdsourcing contests. MIS Q 2021;45(4):1921-1948 [FREE Full text] [doi: 10.25300/misq/2021/14639]
64. Gourieroux C, Holly A, Monfort A. Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. Econometrica 1982;50(1):63. [doi: 10.2307/1912529]
65. Peng CH, Yin D, Zhang H. More than words in medical question-and-answer sites: a content-context congruence perspective. Inf Syst Res 2020;31(3):913-928 [FREE Full text] [doi: 10.1287/isre.2020.0923]
66. Guo L, Jin B, Yao C, Yang H, Huang D, Wang F. Which doctor to trust: a recommender system for identifying the right doctors. J Med Internet Res 2016;18(7):e186 [FREE Full text] [doi: 10.2196/jmir.6015] [Medline: 27390219]
67. Zheng S, Hui SF, Yang Z. Hospital trust or doctor trust? A fuzzy analysis of trust in the health care setting. J Bus Res 2017;78:217-225 [FREE Full text] [doi: 10.1016/j.jbusres.2016.12.017]

XSL•FO
RenderX

68. Mao Z, Du Z, Yuan R, Miao Q. Short-term or long-term cooperation between retailer and MCN? new launched products sales strategies in live streaming e-commerce. J Retail Consum Serv 2022;67:102996 [FREE Full text] [doi: 10.1016/j.jretconser.2022.102996]
69. Hilvert-Bruce Z, Neill JT, Sjöblom M, Hamari J. Social motivations of live-streaming viewer engagement on Twitch. Comput Hum Behav 2018;84:58-67 [FREE Full text] [doi: 10.1016/j.chb.2018.02.013]
70. Tang YM, Chen PC, Law KMY, Wu CH, Lau YY, Guan J, et al. Comparative analysis of student's live online learning readiness during the coronavirus (COVID-19) pandemic in the higher education sector. Comput Educ 2021;168:104211 [FREE Full text] [doi: 10.1016/j.compedu.2021.104211] [Medline: 33879955]
71. Richaud MC, Mesurado B, Cortada AK. Analysis of dimensions of prosocial behavior in an Argentinean sample of children. Psychol Rep 2012;111(3):687-696 [FREE Full text] [doi: 10.2466/10.11.17.PR0.111.6.687-696] [Medline: 23402038]

Abbreviations

FMT: functional motives theory
SET: social exchange theory

Edited by G Eysenbach; submitted 11.09.23; peer-reviewed by P Luo, Y Zhu, C Fu; comments to author 05.10.23; revised version received 30.12.23; accepted 09.03.24; published 25.04.24.

Please cite as:
Liu X, Guo H, Wang L, Hu M, Wei Y, Liu F, Wang X
Effect of Prosocial Behaviors on e-Consultations in a Web-Based Health Care Community: Panel Data Analysis
J Med Internet Res 2024;26:e52646
URL: https://www.jmir.org/2024/1/e52646
doi:10.2196/52646
PMID:38663006

©Xiaoxiao Liu, Huijing Guo, Le Wang, Mingye Hu, Yichan Wei, Fei Liu, Xifu Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Diverse Mentoring Connections Across Institutional Boundaries in the Biomedical Sciences: Innovative Graph Database Analysis

Toufeeq Ahmed Syed[1], MS, PhD; Erika L Thompson[2], PhD; Zainab Latif[3], MCS; Jay Johnson[1], HS; Damaris Javier[4], MA; Katie Stinson[1], MLIS; Gabrielle Saleh[4], BS; Jamboor K Vishwanatha[4], PhD

[1]University of Texas Health Science Center at Houston, Houston, TX, United States
[2]University of Texas Health Science Center at San Antonio, San Antonio, TX, United States
[3]Northwestern University, Evanston, IL, United States
[4]University of North Texas Health Science Center at Fort Worth, Fort Worth, TX, United States

Corresponding Author:
Toufeeq Ahmed Syed, MS, PhD
University of Texas Health Science Center at Houston
7000 Fannin Street, Suite 600
Houston, TX, 77030
United States
Phone: 1 713 500 3591
Email: toufeeq.a.syed@uth.tmc.edu

Abstract

Background: With an overarching goal of increasing diversity and inclusion in biomedical sciences, the National Research Mentoring Network (NRMN) developed a web-based national mentoring platform (MyNRMN) that seeks to connect mentors and mentees to support the persistence of underrepresented minorities in the biomedical sciences. As of May 15, 2024, the MyNRMN platform, which provides mentoring, networking, and professional development tools, has facilitated more than 12,100 unique mentoring connections between faculty, students, and researchers in the biomedical domain.

Objective: This study aimed to examine the large-scale mentoring connections facilitated by our web-based platform between students (mentees) and faculty (mentors) across institutional and geographic boundaries. Using an innovative graph database, we analyzed diverse mentoring connections between mentors and mentees across demographic characteristics in the biomedical sciences.

Methods: Through the MyNRMN platform, we observed profile data and analyzed mentoring connections made between students and faculty across institutional boundaries by race, ethnicity, gender, institution type, and educational attainment between July 1, 2016, and May 31, 2021.

Results: In total, there were 15,024 connections with 2222 mentees and 1652 mentors across 1625 institutions contributing data. Female mentees participated in the highest number of connections (3996/6108, 65%), whereas female mentors participated in 58% (5206/8916) of the connections. Black mentees made up 38% (2297/6108) of the connections, whereas White mentors participated in 56% (5036/8916) of the connections. Mentees were predominately from institutions classified as Research 1 (R1; doctoral universities—very high research activity) and historically Black colleges and universities (556/2222, 25% and 307/2222, 14%, respectively), whereas 31% (504/1652) of mentors were from R1 institutions.

Conclusions: To date, the utility of mentoring connections across institutions throughout the United States and how mentors and mentees are connected is unknown. This study examined these connections and the diversity of these connections using an extensive web-based mentoring network.

(J Med Internet Res 2024;26:e47560) doi:10.2196/47560

KEYWORDS

online platform; mentorship; diversity; network analysis; graph database; online communities

XSL•FO
RenderX

Introduction

Web-based networks have the power and capacity to connect individuals unlike anything we have previously experienced in our society. This connection capacity is especially important for those traditionally underrepresented in the biomedical sciences. Black or African American and Hispanic or Latinx individuals use social media at higher rates than their White counterparts, which may aid their ability to seek mentoring beyond their local networks [1]. This is especially important in the realm of health care as diversity in the biomedical workforce is essential for addressing health disparities and other public health needs [2-4]. The National Institutes of Health (NIH) has invested in diversity-focused initiatives to promote representation from underrepresented groups in higher levels of the biomedical workforce [3]. One of these investments is in the National Research Mentoring Network (NRMN), which aims to promote diversity among undergraduate, graduate, postdoctoral, and workforce career stages through mentorship and mentoring tools [5,6].

NRMN provides resources (eg, culturally responsive mentoring and networking) remotely through a portal, NRMNet, for the biomedical, educational, and workforce pipeline [5]. A key component of NRMNet is the MyNRMN platform, which allows people from all education and career stages to remotely mentor or receive mentorship throughout the United States and territories [7]. Mentoring is an integral factor in supporting longevity and persistence in the biomedical science fields and training, as well as for professional development and developing a science identity [8-11]. Moreover, mentoring early and during impressionable stages in a person's education and career pathways may have a cumulative effect for downstream biomedical career success [12]. By engaging with MyNRMN, users gain access to a network of mentors and mentees promoting education and career advancement for individuals who may not have access to these resources locally.

Historically, persons from underrepresented minority groups are at a disadvantage when accessing and receiving mentoring compared with nonminority groups [13]. Mentorship from persons who identify with challenges specific to underrepresented minorities is a valuable asset for professional development and training [14]. Off-site mentorship and self-reflection are key aspects to influential and beneficial mentorships, which often are not accessible within limited networks [15]. Furthermore, mentoring or coaching networks, including peer mentors, provide different forms of support for training and development [12,15,16].

MyNRMN provides an accessible platform for mentorship with the goal of providing meaningful connections for persistence in the biomedical sciences. The MyNRMN platform is unique because it crosses state lines and institutional boundaries for mentorship, which is vital for underrepresented minorities who may not have access to mentoring networks or expansive social capital at their own institution. The development of MyNRMN, as discussed by Ahmed et al [7], describes the importance of incorporating social capital and social networks while building and creating this platform. The connections built within MyNRMN increase an individual's social capital, specifically in the biomedical sciences, enabling them to move further in their education and career through the "informational, emotional, and instrumental resources and supports" [7]. By observing and understanding the connections within MyNRMN, we can observe how an individual's social capital increases through this network. Future studies will discuss the effects of increased social capital within MyNRMN.

This study aimed to examine the large-scale mentoring connections, facilitated by MyNRMN, between students (mentees) and faculty (mentors) across institutional and geographic boundaries on a large, national mentoring network for biomedical sciences. We assessed the diversity of connections on the platform by gender, race, ethnicity, educational attainment, and minority-serving institution (MSI)–Carnegie classification. We hypothesized that building this web-based platform would create a space for diverse mentoring networks that expand beyond an individual's immediate proximity and personal identity.

Methods

MyNRMN Platform

Details of the MyNRMN platform capabilities have been described in previous publications [5,7]. Briefly, the platform is a remote mentoring, networking, and career development space where individuals can enroll, create a profile, and sign up to be a mentor or mentee within the network. Additionally, the platform provides a space for "connections" (ie, people linking accounts across the platform) and to form mentoring relationships. In typical academic settings, mentoring is limited to a particular department or institution, but with MyNRMN, a mentee or mentor can connect with a plethora of like-minded biomedical science professionals and students across the nation (8473 mentors and 15,852 mentees as of May 15, 2024), increasing their social capital beyond their immediate location. This includes connections across types of institutions and professional or education levels. For example, a mentee from a 2-year community college in Idaho can connect with a mentor from an institution classified by Carnegie as a Research 1 (R1; doctoral universities—very high research activity) on the East Coast, or a mentee from a 4-year program in Atlanta can connect with somebody from Washington state in a PhD program. In this fashion, MyNRMN can facilitate interactions for those seeking professional or career support, or advice on majors, job opportunities, research areas, graduate school applications, and so on.

Graph Database Infrastructure

MyNRMN is built on a robust graph database called Neo4j, which facilitates creating a rich social graph that mimics real-world social and mentoring connections. Each node can be a mentee, mentor, member, institution, career, location, group, course, mentoring program, or cohort. Additionally, each connection (the link between 2 nodes) is categorized and labeled. For example, a link between a mentor and a mentee is labeled as a mentor-mentee connection and a link between a mentee and a mentee is labeled as a mentee-mentee connection. Figure 1 explains the different types of nodes and the labels of links

XSL•FO
RenderX

between the nodes. We have built our platform and graph database to work together, so that every time a mentee connects to a mentor, a mentor connects to another mentor, a mentee connects to another mentee, or a member joins a group, the system creates an entry in our database and captures that in the graph database as a connection by these 2 nodes.

Figure 1. MyNRMN Neo4j graph database structure.

Measures

In the MyNRMN platform, we capture the profile information, which includes demographic fields (eg, gender, race, ethnicity, institution, whether they are first in the family to go to college, and education level). When a member joins MyNRMN, the system syncs the member's profile to our Neo4j graph database, creating a node for the person, location, school, and career, and creates the links between these nodes. Each node includes properties such as system ID and name. For example, the member node includes the member's profile information, for example, system ID, full name, email, alternate email, education, degrees, careers, interest, and has changed school in past 6 months. Likewise, each activity in MyNRMN is also recorded

XSL•FO
RenderX

in Neo4j. Through this process, when a member joins a group, enrolls into a course, or joins a cohort, a link is created between the member node and the node of the group, the course, or the cohort the member has joined.

For member connections, the system only records accepted connections; pending and rejected connection requests are not posted to our graph database. However, connections can be made in a multitude of ways using different engagement features [7]. For example, a mentee can start a search or seek a mentor using the Find a Mentor tool. The mentee will start with a few keywords and search for different mentors. Once they find a mentor they are interested in connecting with and seeking advice from, they will send a request for connection. The mentor will receive a connection request via email and will have the choice to accept or reject that request. If the request is accepted, the system will create a Connected_to link between the mentee and the mentor nodes in Neo4j. If the request is of a formal mentoring type, a Mentored_by link will be created between the mentee and the mentor nodes. The link can also include properties just like nodes; for example, the Mentored_by link includes information such as mentorship type (one-on-one mentoring or program mentoring), program name, and mentoring path (postbaccalaureate, undergraduate, graduate, postdoctoral, and junior faculty).

If the connection is not accepted, it will stay pending in our platform and will not be added to our graph database. Hence, our graph database maintains the active connections between mentors and mentees. For this paper, we are reporting the network and connection analysis between mentors and mentees for active, accepted connections. To describe our data set, we have extracted 3 different types of data sets analyzing mentee-mentor connections, mentee-mentee connections, and mentor-mentor connections (the latter 2 are considered peer mentoring). For each of these analyses, we have included only the accepted connections on our platform.

For each variable, such as gender, race, ethnicity, education level, and MSI-Carnegie classification, we counted the number of accepted connections based on the role of the requester (member who requested the connection) and the role of the receiver (member who accepted the connection) as follows: mentee-mentor, mentee-mentee, and mentor-mentor. To obtain the number of connections between the categories defined above, we retrieved the profile information of the requester and the receiver of the connection request. For gender, race, ethnicity, and education level, we used the information provided by the member in their profile. The connections by MSI-Carnegie institutions were determined by the institution stated in the member profile. These were then classified into one of the following categories: Asian American and Native American Pacific Islander–serving institutions (AANAPISIs), historically Black colleges and universities (HBCUs), Hispanic-serving institutions (HSIs), tribally controlled colleges and universities (TCCUs), Carnegie doctoral university R1, and Carnegie doctoral university R2 (doctoral universities—high research activity). We have defined the MSIs based on the US Department of Education (2020) Eligibility Matrix Report. The Carnegie classification of institutions was based on The Carnegie Classification of Institutions of Higher Education [17].

Sample

To examine the mentoring connections occurring on MyNRMN, we captured data from July 1, 2016, until May 31, 2021. These data show the MyNRMN connections between mentees and mentors from different genders, races, ethnicities, education levels, and MSI-Carnegie institutions across the country, based on the user's profile data. During this time frame there were 2261 mentees and 1583 mentors contributing connection data.

Data Analysis

This study presents the descriptive counts and frequencies for connections accepted across the MyNRMN platform. These connections are stratified by demographics and institution type. Moreover, we estimated the average connections by demographic and institution characteristics. Alluvial graphs were developed to show the percentage of accepted connections between platform users.

Ethical Considerations

Research reported in this publication was supported by the National Institute of General Medical Sciences of the NIH under award numbers U54GM119023 and U24GM132217. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. All user data collected is protected under the North Texas Regional Institutional Review Board, reference number 2015-0720, and stored securely.

Results

Characteristics

During the July 1, 2016, to May 31, 2021, time frame, there were 2222 mentees and 1652 mentors contributing connection data. Tables 1 and 2 describe the number of connections and average connections per user during the study period stratified by gender, race, ethnicity, and institution type. In total, there were 15,024 connections during the study period. Among the connections collected in the aforementioned window, mentors and mentees came from 1625 institutions. The institutions with the highest representation included the partner institutions Tuskegee University (151 connections), Savannah State University (98 connections), University of North Texas Health Science Center (97 connections), and University of Wisconsin-Madison (93 connections). A majority of connections occurred between different institutions (n=4603 connections).

XSL•FO
RenderX

Table 1. Mentor and mentee connections.

Demographic	Mentees, n[a]	Connections by mentees, n[b]	Connections per mentee, mean (SD)	Mentors, n[c]	Connections by mentors, n[d]	Connections per mentor, mean (SD)
Total	2222	6108	2.75 (5.39)	1652	8916	5.54 (16.52)
Gender						
Female	1309	3996	3.05 (6.11)	895	5206	5.82 (19.73)
Male	499	1345	2.7 (5.11)	624	3217	5.16 (12.42)
Other	9	27	3 (3.84)	4	9	2.25 (1.5)
Race						
White	638	1590	2.49 (4.4)	902	5036	5.58 (19.63)
Black	727	2297	3.16 (4.97)	277	1592	5.75 (10.92)
Asian	281	603	2.15 (3.05)	178	902	5.07 (17.83)
Native American	52	95	1.83 (1.54)	16	51	3.19 (3.15)
Hawaiian-Pacific Islander	11	14	1.27 (0.65)	6	9	1.5 (0.84)
Two or more races	20	121	6.05 (15.88)	19	115	6.05 (5.25)
Other	124	314	2.53 (3.65)	49	341	6.96 (10.99)
Ethnicity						
Hispanic or Latino	349	943	2.7 (4.5)	172	908	5.28 (8.71)
Not Latino	1208	3681	3.05 (6.38)	1169	6210	5.31 (13)
Other	85	237	2.79 (4.62)	51	268	5.25 (8.66)
Education						
Undergraduate	563	1703	3.02 (5.87)	33	239	7.24 (16.07)
Graduate	470	1440	3.06 (5.27)	77	868	11.27 (54.55)
Postdoctoral	292	716	2.45 (3.08)	205	885	4.32 (6.52)
Other (currently working/faculty/not in school or formal program)	437	1378	3.15 (7.82)	1221	6574	5.38 (12.89)
Institution						
AANAPISI[e]	48	155	3.23 (5.3)	25	151	6.04 (7.97)
HBCU[f]	307	1072	3.49 (5.87)	121	684	5.65 (6.54)
HSI[g]	284	538	1.89 (1.89)	105	560	5.33 (12.76)

[a]Number of mentees with at least 1 active connection.

[b]Number of active connections requested or accepted by mentees.

[c]Number of mentors with at least 1 active connection.

[d]Number of active connections requested or accepted by mentors.

[e]AANAPISI: Asian American and Native American Pacific Islander–serving institution.

[f]HBCU: historically Black colleges and universities.

[g]HSI: Hispanic-serving institution.

XSL•FO
RenderX

Table 2. Connection numbers by type of connection.

	Mentee-mentor connections, n[a]	Mentee-mentor connections, mean (SD)	Mentor-mentee connections, n[b]	Mentor-mentee connections, mean (SD)	Mentee-mentee connections, n[c]	Mentee-mentee connections, mean (SD)	Mentor-mentor connections, n[d]	Mentor-mentor connections, mean (SD)
Total	3388	2.17 (4.25)	930	2.93 (9.44)	895	2.05 (2.79)	2299	4.35 (14)
Gender								
Female	2125	2.37 (4.47)	500	2.92 (11.14)	636	2.42 (3.3)	1441	5.06 (17.08)
Male	769	2.43 (5.44)	357	3.16 (7.84)	150	1.58 (1.37)	719	3.84 (10.27)
Other	9	1.13 (0.35)	0	0 (0.0)	2	1 (0.0)	0	0 (0.0)
Race								
White	841	2.02 (2.68)	572	3.4 (12.77)	226	2.09 (3.45)	1295	4.89 (17.39)
Black	1166	2.24 (4.03)	161	2.6 (2.46)	405	2.29 (2.49)	367	3.67 (6.1)
Asian	392	1.95 (2.87)	79	3.16 (3.65)	69	1.44 (0.85)	295	5.67 (18.47)
Native American	36	1.29 (0.53)	9	2.25 (1.5)	5	1 (0.0)	16	2 (2.07)
Hawaiian-Pacific Islander	10	1.25 (0.71)	1	1 (0.0)	0	0 (0.0)	4	1.33 (0.58)
Two or more races	96	6.4 (17.13)	11	3.67 (2.08)	5	1 (0.0)	4	1.33 (0.58)
Other	190	1.98 (2.64)	27	2.08 (1.71)	36	2.12 (2.18)	105	5 (8.14)
Ethnicity								
Hispanic or Latino	530	2.29 (4.73)	61	1.97 (1.68)	109	1.68 (1.15)	153	3.4 (4.14)
Not Latino	1927	2.41 (4.86)	581	2.7 (5.9)	583	2.39 (3.44)	1487	4.14 (9.11)
Other	146	2.56 (3.7)	21	2.1 (2.02)	23	2.09 (2.7)	82	4.82 (6.54)
Education								
Undergraduate	916	2.25 (5.58)	118	8.43 (21.02)	323	2.15 (2.18)	442	3.86 (3.89)
Graduate	890	2.51 (3.88)	193	9.19 (31.17)	192	2.31 (3.69)	1499	12.63 (45.62)
Postdoctoral	369	2.32 (2.92)	62	1.77 (2.14)	68	1.33 (1.07)	189	3.32 (3.01)
Other (currently working/faculty/not in school or formal program)	648	2.58 (5.58)	495	2.31 (21.02)	184	2.75 (4.23)	81	4.12 (8.94)
Institution								
AANAPISI[e]	106	3.79 (6.56)	14	1.56 (1.01)	9	1.8 (0.84)	29	2.9 (3.51)
HBCU[f]	463	2.19 (5.18)	105	2.69 (2.76)	246	2.37 (2.24)	143	2.8 (3.21)
HSI[g]	311	1.47 (1.3)	27	2.08 (1.89)	60	1.46 (0.74)	106	3.21 (4.28)

[a]Number of connections requested by mentees and accepted by mentors.

[b]Number of connections requested by mentors and accepted by mentees.

[c]Number of connections requested by mentees and accepted by other mentees.

[d]Number of connections requested by mentors and accepted by other mentors.

[e]AANAPISI: Asian American and Native American Pacific Islander–serving institution.

[f]HBCU: historically Black colleges and universities.

[g]HSI: Hispanic-serving institution.

XSL•FO
RenderX

Gender

Mentee-Mentor Accepted Connections by Gender

Of the accepted connections between mentees and mentors (n=3645 connections), female mentees made up 72% (2624/3645) of the accepted connections, male mentees made up 27% (999/3645), and those who identified as other gender made up 0.6% (22/3645); see Figure 2. Alternatively, female mentors made up 59% (2136/3645) of the accepted connections, male mentors made up 41% (1503/3645), and those who identified as other made up 0.2% (6/3645) of the accepted connections. Of the accepted connections, female mentees connected with female mentors approximately 61% (1604/2624) of the time, and they connected with male mentors approximately 39% (1016/2624) of the time. Male mentees were almost evenly split in their connection percentages between male and female mentors. Of the accepted connections, female mentors connected with female mentees approximately 75% (1604/2136) of the time, whereas male mentors connected with female mentees approximately 68% (1016/1503) of the time.

Figure 2. Connections by gender: mentee-mentor.

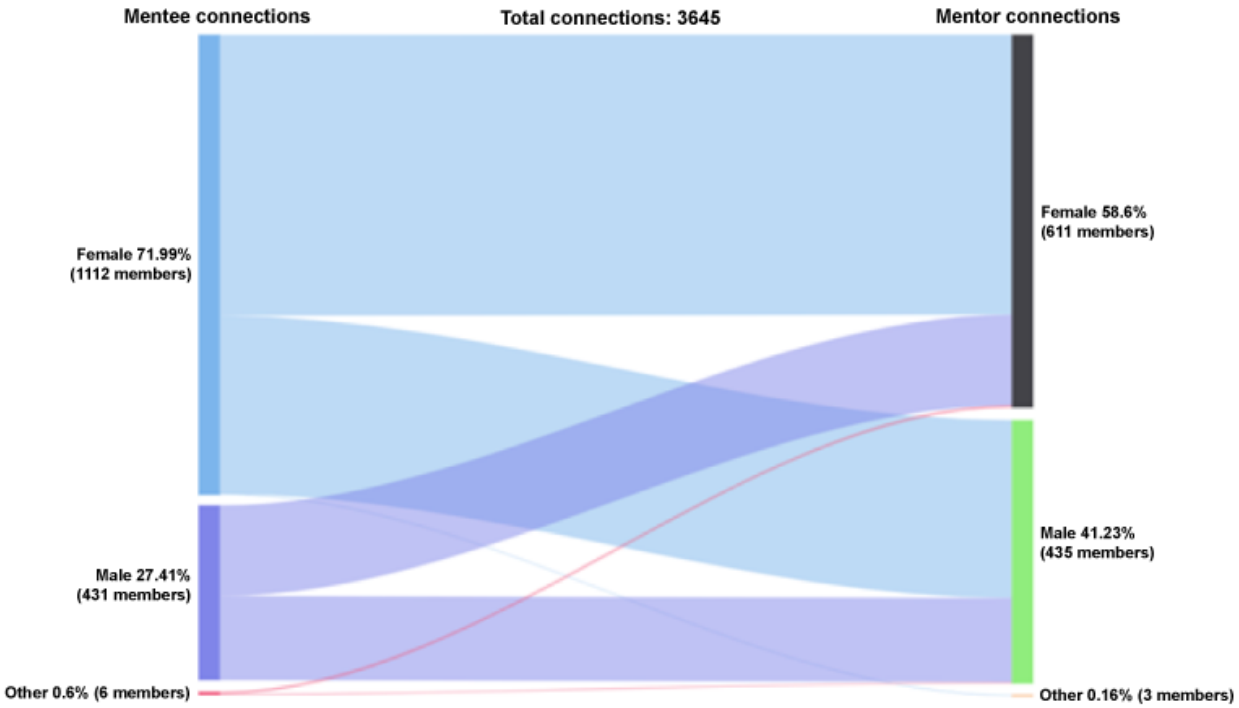

Mentee-Mentee Accepted Connections by Gender

Female mentees represented 80% (641/805) of the initiated and accepted mentee-mentee connections (n=805 connections), whereas male mentees made up 20% (162/805) and those who identify as other gender represented 0.25% (2/805); see Figure 3. Among the mentees who received and accepted peer mentee requests, 75% (604/805) were female, 24% (195/805) were male, and 0.75% (6/805) identified as other gender. Female mentees who initiated the mentee-mentee connection reached out to female mentees 75% (483/641) of the time, male mentees 24% (153/641) of the time, and those who identified as other 0.8% (5/641) of the time. Alternatively, male mentees who initiated the mentee-mentee connection reached out to female mentees 73% (119/162) of the time, male mentees 26% (42/162) of the time, and those who identified as other 0.6% (1/162) of the time. The mentees who identified as other connected only with female mentees.

XSL•FO
RenderX

Figure 3. Connections by gender: mentee-mentee.

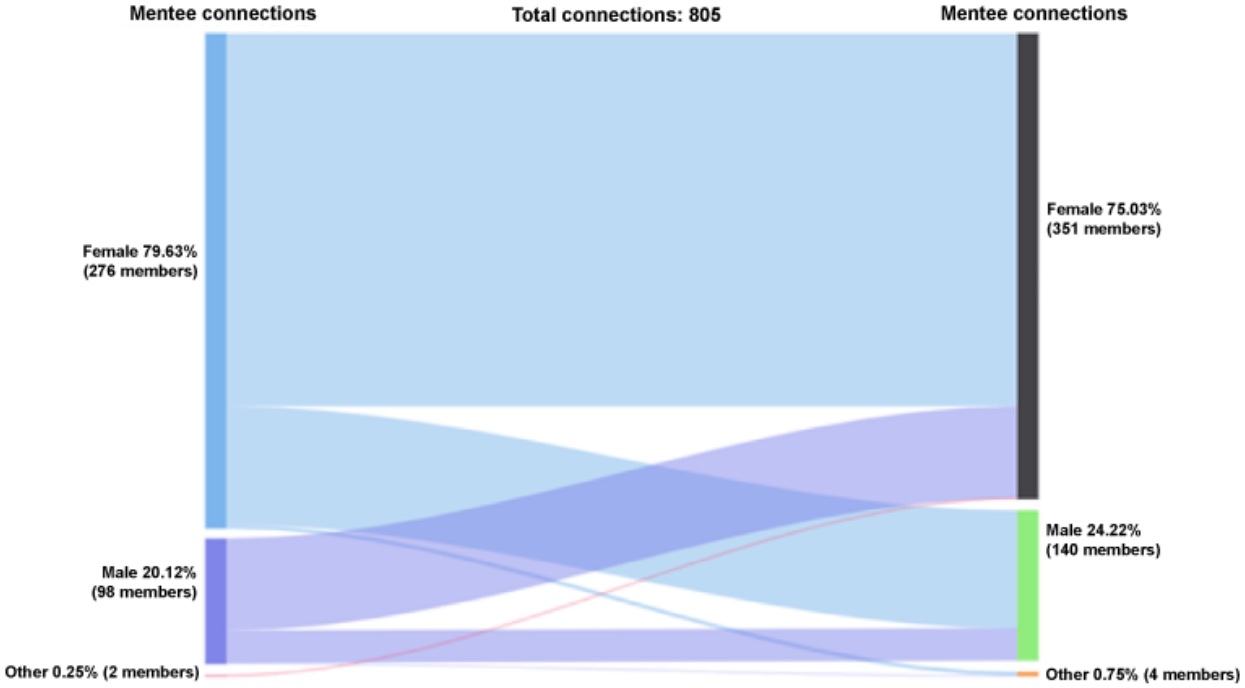

Mentor-Mentor Accepted Connections by Gender

Mentors can also connect with one another. Of the mentors who initiated the accepted mentoring connections (n=1985 connections), 69% (1367/1985) were female and 31% (618/1985) were male (Figure 4). Among the mentors who received and accepted peer mentor requests, 62% (1237/1985) were female and 38% (748/1985) were male. Among the female-initiated mentor-mentor connections, 64% (869/1367) were female-female and 36% (498/1367) were female-male connections. Among the male-initiated mentor-mentor connections, 60% (368/618) were male-female and 40% (250/618) were male-male connections.

Figure 4. Connections by gender: mentor-mentor.

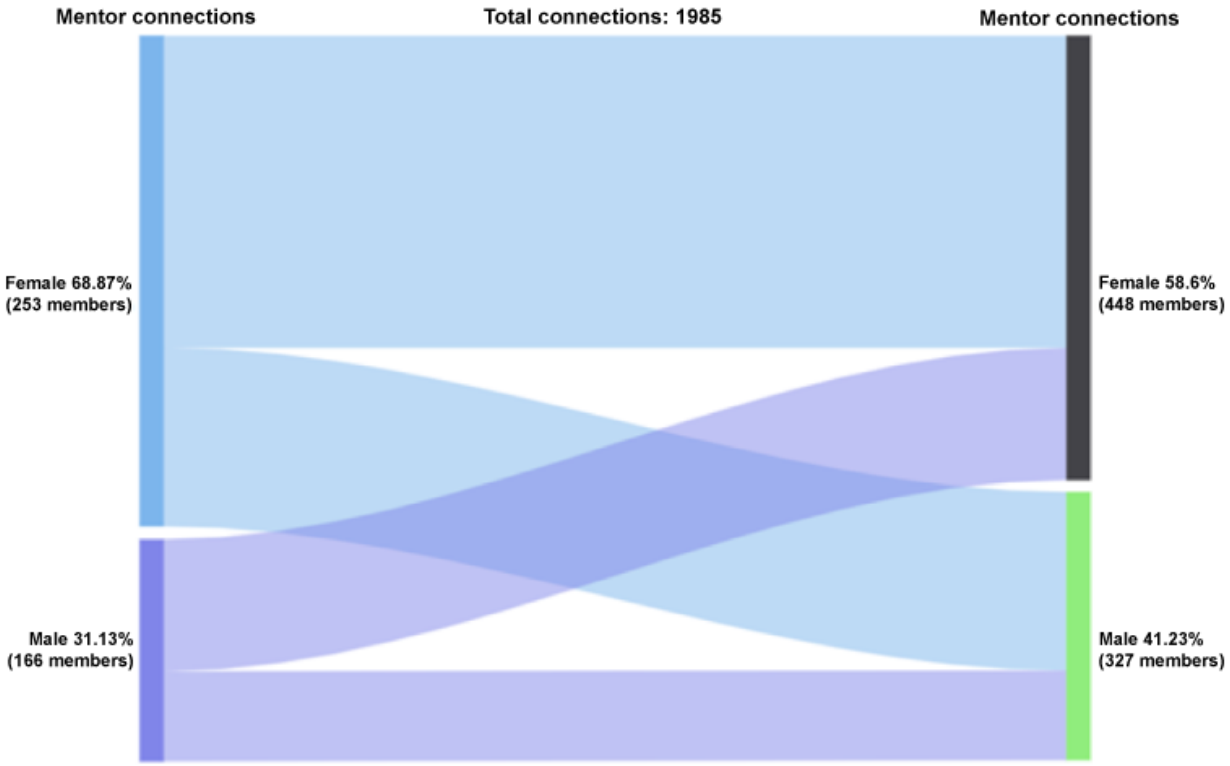

XSL•FO
RenderX

Race

Mentee-Mentor Accepted Connections by Race

Of the accepted connections between mentees and mentors (n=3302 connections), mentee connections were 42% (1384/3302) Black, 33% (1100/3302) White, 13% (424/3302) Asian, 3% (91/3302) Native American, 3% (85/3302) 2 or more races, 0.3% (10/3302) Hawaiian-Pacific Islander, and 6% (208/3302) other (Figure 5). Alternatively, 63% (2,076/3302) of accepted mentor connections were White mentors, followed by 22% (718/3302) Black, 10% (338/3302) Asian, 2% (68/3302) 2 or more races, 0.4% (12/3302) Native American, and 3% (88/3302) other.

Figure 5. Connections by race: mentee-mentor.

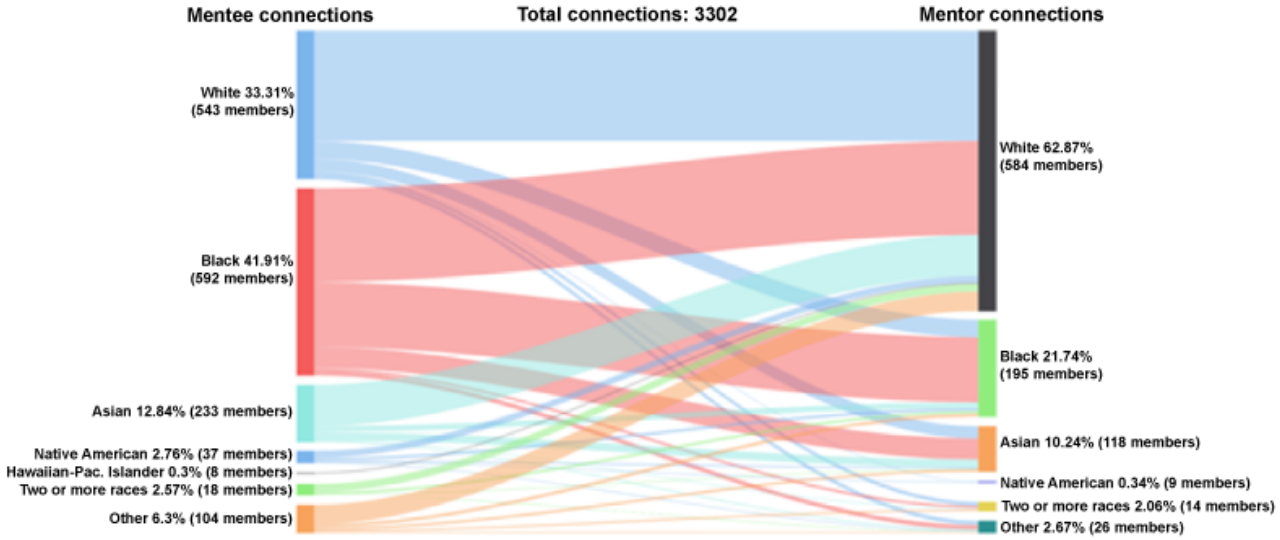

Of the 1100 accepted connections among White mentees, the mentors were 74% (817/1100) White, 12% (131/1100) Black, 8% (90/1100) Asian, 0.3% (3/1100) Native American, 3% (28/1100) 2 or more races, and 3% (31/1100) other. Of the 1384 accepted connections among Black mentees, the mentors were 50% (694/1384) White, 35% (481/1384) Black, 11% (155/1384) Asian, 0.1% (1/1384) Native American, 0.1% (1/1384) Hawaiian-Pacific Islander, 1% (19/1384) 2 or more races, and 3% (33/1384) other. Among the 424 accepted connections among Asian mentees, the mentors were 72% (306/424) White, 10% (43/424) Black, 14% (59/424) Asian, 0.2% (1/424) Native American, 1% (6/424) 2 or more races, and 2% (9/424) other. Among the 91 accepted connections among Native American mentees, the mentors were 55% (50/91) White, 21% (19/91) Black, 10% (9/91) Asian, 7% (6/91) Native American, 0% (0/91) Hawaiian-Pacific Islander, 1% (1/91) 2 or more races, and 7% (6/91) other. Among the 85 connections for mentees identifying as 2 or more races, mentors were 68% (58/85) White, 22% (19/85) Black, 7% (6/85) Asian, and 2% (2/85) other. Among the 10 connections for Hawaiian-Pacific Islander mentees, mentors were 70% (7/10) White, 10% (1/10) Black, 10% (1/10) Hawaiian-Pacific Islander, and 10% (1/10) 2 or more races. Finally, among the 208 connections from mentees identifying as other, mentors were 69% (144/208) White, 12% (24/208) Black, 9% (19/208) Asian, 0.5% (1/208) Native American, 6% (13/208) 2 or more races, and 4% (7/208) other.

Of the 2076 connections accepted by White mentors, mentees were 39% (817/2076) White, 33% (694/2076) Black, 15% (306/2076) Asian, 2% (50/2076) Native American, 0.3% (7/2076) Hawaiian-Pacific Islander, 3% (58/2076) 2 or more races, and 7% (144/2076) other. Of the 718 connections accepted by Black mentors, mentees were 18% (131/718) White, 67% (481/718) Black, 6% (43/718) Asian, 3% (19/718) Native American, 0.1% (1/718) Hawaiian-Pacific Islander, 3% (19/418) 2 or more races, and 3% (24/718) other. Among the 338 connections accepted by Asian mentors, mentees were majority Black 46% (155/338), followed by White (90/338, 27%), Asian (59/338, 17%), other race (19/338, 6%), Native American (9/338, 3%), and 2 or more races (6/338, 2%). Of the 88 connections accepted by mentors who identified as other race, approximately one-third were with Black and White mentees (33/88, 39% and 31/88, 35%, respectively). The remaining accepted connections comprised Asian (9/88, 10%), other race (7/88, 8%), Native American (6/88, 7%), and mentees of 2 or more races (2/88, 2%). Among the 68 connections accepted by mentors of 2 or more races, a majority were with White mentees (28/68, 41%), followed by Black (19/68, 28%), other (13/68, 19%), Asian (6/68, 9%), Native American (1/68, 1%), and Hawaiian-Pacific Islander (1/68, 1%). Only 12 connections were accepted by Native American mentors, and half were with Native American mentees. Similarly, there were only 2 connections accepted by Hawaiian-Pacific Islander mentors.

Mentee-Mentee Accepted Connections by Race

Black mentees represented 56% (422/749) of the initiated and accepted mentee-mentee connections (n=749 connections), followed by 28% (213/749) White, 8% (63/749) Asian, 2% (15/749) Native American, 4% (32/749) other, and 0.5% (4/749) 2 or more races (Figure 6). Among the mentees who received and accepted peer mentee requests, 50% (375/749) were Black, 34% (254/749) were White, 7% (53/749) were Asian, 2% (13/749) were Native American, 0.3% (2/749) were Hawaiian-Pacific Islander, 1% (7/749) were 2 or more races, and 6% (45/749) identified as other race.

XSL•FO
RenderX

Figure 6. Connections by race: mentee-mentee.

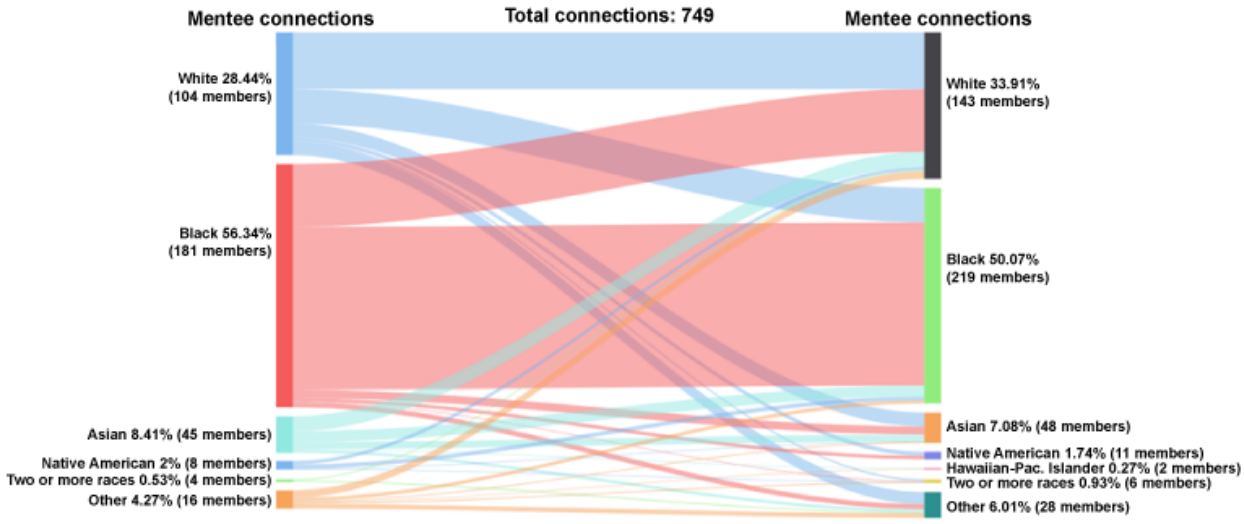

Black mentees initiated the most connections compared with other racial groups. Black mentees who initiated peer mentee-mentee connections (n=422 connections) connected with mentees who were Black (283/422, 67%), White (109/422, 26%), and less than 3% of Asian (14/422), Native American (6/422), other (9/422), and 2 or more races (1/422). White mentees who initiated the mentee-mentee connections (n=213 connections) reached out to mentees who were White (98/213, 46%), Black (60/213, 28%), Asian (24/213, 11%), Native American (7/213, 3%), other race (21/213, 10%), and 2 or more races (3/213, 1%). Asian mentees initiated 63 connections. These connections were directed toward mentees who were White (26/63, 41%), Black (19/63, 30%), Asian (12/63, 19%), other race (5/63, 8%), and Hawaiian-Pacific Islander (1/63, 2%). Mentees of other racial backgrounds initiated 32 connections primarily to White (13/32, 41%), other (8/32, 25%), and Black (19%, 6/32) mentees. The remaining connections were with Asian mentees (2/32, 6%), mentees of 2 or more races (2/32, 6%), and Hawaiian-Pacific Islander (1/32, 3%) mentees. A total of 15 mentee-mentee connections were initiated among Native American mentees; of these connections, 7/15 (47%) were with Black mentees, 6/15 (40%) to White mentees, 1/15 to Asian mentees, and 1/15 to mentees of 2 or more races. Only 4 connections were initiated by mentees of 2 or more races.

Mentor-Mentor Accepted Connections by Race

Of the mentors who initiated the accepted mentoring connections (n=1804 connections), 64% (1162/1804) were White, 16% (292/1804) were Black, 14% (250/1804) were Asian, 5% (83/1804) were other race, and less than 1% were Native American (12/1804), Hawaiian-Pacific Islander (3/1804), and 2 or more races (2/1804); see Figure 7. Among the mentors who received and accepted peer mentor requests, 68% (1219/1804) were White, 17% (315/1804) were Black, 10% (178/1804) were Asian, 3% (58/1804) were other, 1% (21/1804) were 2 or more races, and 0.7% (13/1804) were Native American.

Figure 7. Connections by race: mentor-mentor.

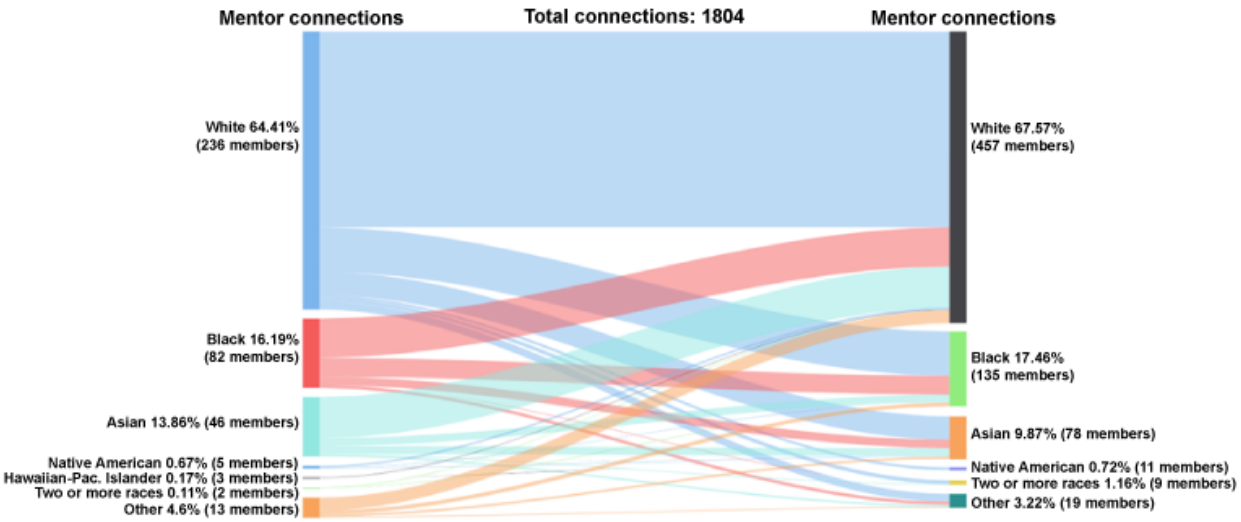

XSL•FO
RenderX

White mentors had the most initiated connections compared with other racial groups. White mentors who initiated peer mentor-mentor connections (n=1162 connections) connected with White (818/1162, 70%), Black (186/1162, 16%), and Asian (98/1162, 8%) mentors. The remaining connections comprised less than 5% (Native American, 2 or more races, and other combined total 60/1162) of the initiated connections. Black mentors who initiated mentor-mentor connections (n=292 connections) reached out to White (164/292, 56%), Black (79/292, 27%), Asian (35/292, 12%), and other race (13/292, 5%) mentors. Asian mentors initiated 250 connections to White (172/250, 69%), Black (32/292, 13%), and Asian (33/250, 13%) mentors. The remaining connections comprised less than 5% (Native American, 2 or more races, and other combined total 13/250) of initiated connections. Mentors of other racial backgrounds initiated 83 connections primarily to White (54/83, 95%), Black (16/83, 19%), and Asian (9/83, 11%) mentors. The remaining 5% (4/83) of connections were initiated by mentors of other racial backgrounds. Twelve mentor-mentor connections were initiated among Native American mentors; of these connections, 67% (8/12) were with White mentors. Only 5 total connections were initiated by mentors from Hawaiian-Pacific Islander and more than 1 racial background.

Ethnicity

Mentee-Mentor Accepted Connections by Ethnicity

Of the accepted connections between mentees and mentors (n=2936 connections), Hispanic or Latinx mentees made up 22% (645/2936) of the accepted connections, non-Hispanic or Latinx mentees made up 72% (2121/2936), and those who identified as other were 6% (170/2936); see Figure 8. Non-Hispanic or Latinx mentors were a majority of the accepted connections (2389/2936, 81%), whereas Hispanic or Latinx mentors were 15% (429/2936) of the accepted connections. Of the accepted connections with Hispanic or Latinx mentees (n=645 connections), 64% (411/645) were with non-Hispanic or Latinx mentors and 30% (196/645) were with Hispanic or Latinx mentors. Of the accepted connections with Hispanic or Latinx mentors (n=429), 46% (196/429) were with Hispanic or Latinx mentees, 44% (189/429) with non-Hispanic or Latinx mentees, and 10% (44/429) with other ethnic backgrounds. Of the accepted connections with non-Hispanic or Latinx mentees (n = 2121), 88% (1862/2121) were with non-Hispanic or Latinx mentors, 9% (189/2121) with Hispanic or Latinx mentors, and 3% (70/2121) with other mentors. Of the accepted connections with non-Hispanic or Latinx mentors (n=2389), most were with non-Hispanic or Latinx mentees (1862/2389, 78%), followed by Hispanic or Latinx mentees (411/2389, 17%) and other mentees (116/2389, 5%). Accepted connections with mentees of other ethnicities occurred 170 times, and accepted connections with mentors of other ethnicities occurred 118 times.

Figure 8. Connections by ethnicity: mentee-mentor.

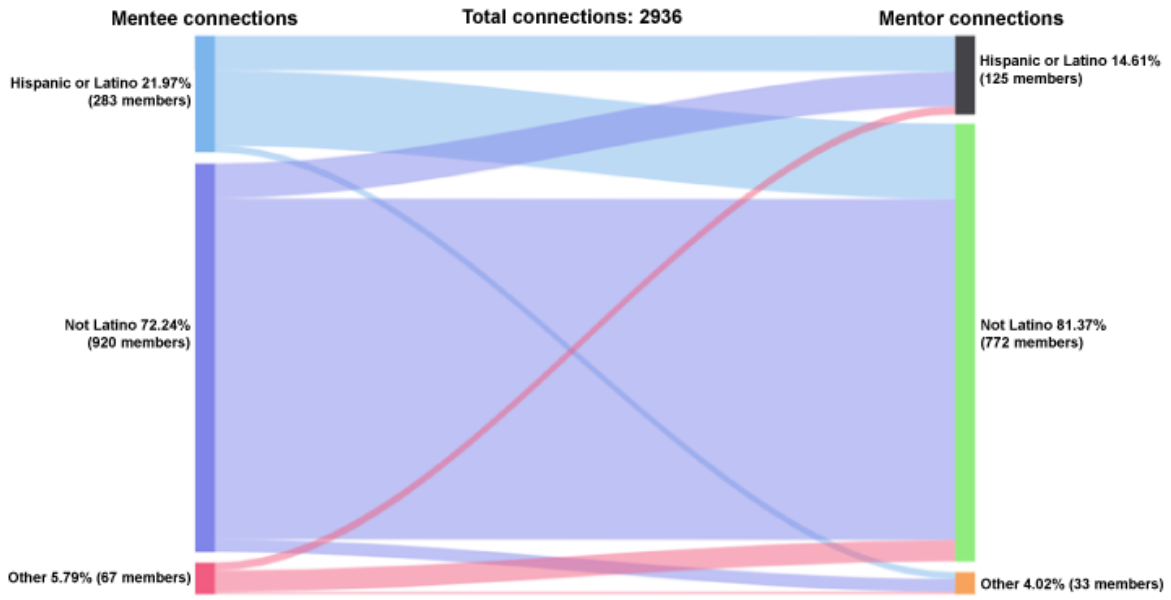

Mentee-Mentee Accepted Connections by Ethnicity

Non-Hispanic or Latinx mentees represent 81% (565/700) of the initiated and accepted mentee-mentee connections, while Hispanic or Latinx mentees were 16% (111/700) of those connections (n=700 connections; Figure 9). Among the mentees who received and accepted peer mentee requests, 75% (528/700) were non-Hispanic or Latinx, 20% (142/700) were Hispanic or Latinx, and 4% (30/700) identified as other. Hispanic or Latinx mentees who initiated mentee-mentee connections (n=111 connections) reached out to Hispanic or Latinx (62/111, 56%), non-Hispanic or Latinx (44/111, 40%), and other (5/111, 5%) mentees. Non-Hispanic or Latinx mentees who initiated mentee-mentee connections (n=565) connected with non-Hispanic or Latinx (469/565, 83%), Hispanic or Latinx (75/565, 13%), and other (21/565, 4%) mentees. A total of 24 connections occurred for people of other ethnic backgrounds; one-third were with non-Hispanic or Latinx mentees.

XSL•FO
RenderX

Figure 9. Connections by ethnicity: mentee-mentee.

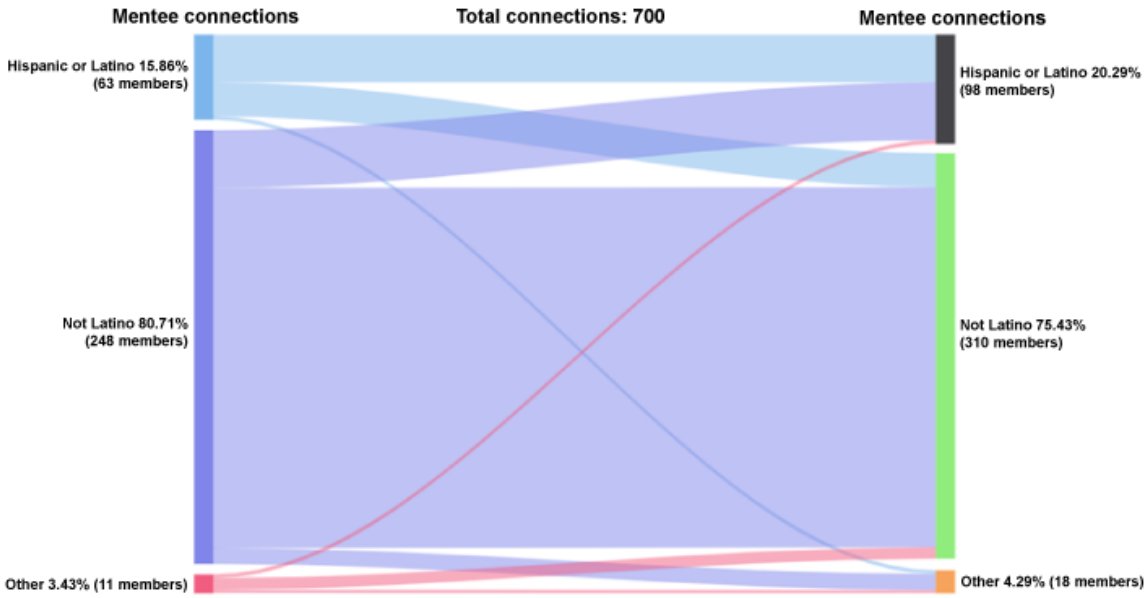

Mentor-Mentor Accepted Connections by Ethnicity

Mentors can connect with each other, and of the mentors who initiated and accepted mentoring connections (n=1441 connections), 88% (1262/1441) were non-Hispanic or Latinx, 8% (112/1441) were Hispanic or Latinx, and 5% (67/1441) were other ethnicity (Figure 10). Among the mentors who received and accepted peer mentor requests, 86% (1235/1441) were non-Hispanic or Latinx, 11% (162/1441) were Hispanic or Latinx, and 3% (44/1441) were other ethnicity. Hispanic or Latinx mentors who initiated mentor-mentor connections (n=112 connections) reached out to non-Hispanic or Latinx (75/112, 67%), Hispanic or Latinx (33/112, 29%), and other (4/112, 4%) mentees. Non-Hispanic or Latinx mentors who initiated mentor-mentor connections (n=1262) connected with non-Hispanic or Latinx (1104/1262, 87%), Hispanic or Latinx (119/1262, 9%), and other (39/1262, 3%) mentors. A total of 67 connections occurred for people of other ethnic backgrounds; 84% (56/67) were with non-Hispanic or Latinx mentors.

Figure 10. Connections by ethnicity: mentor-mentor.

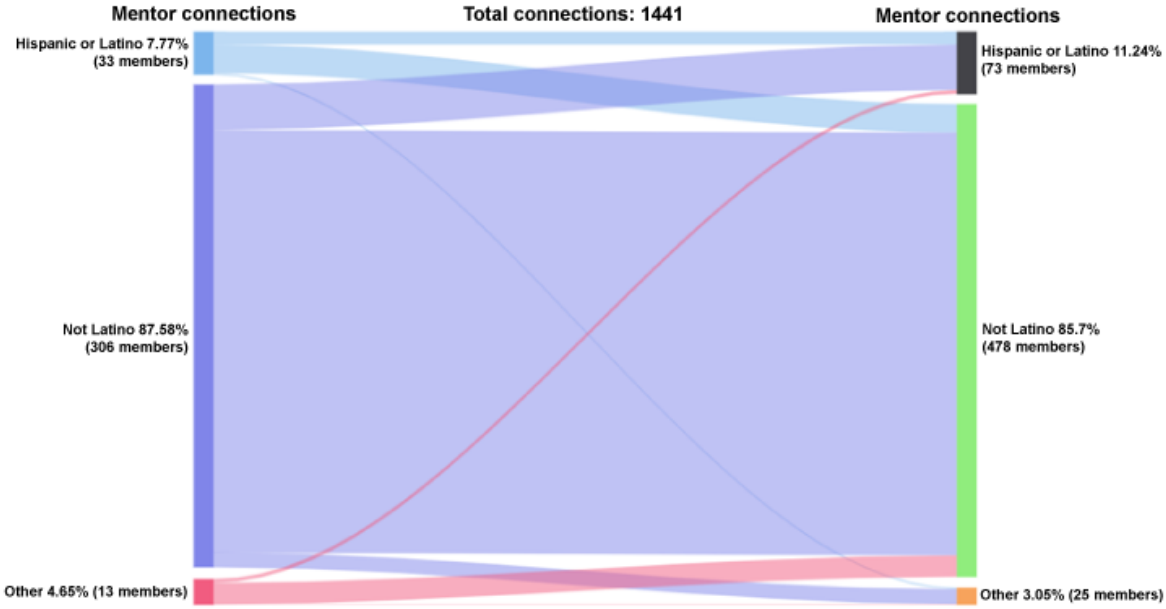

XSL•FO
RenderX

MSI-Carnegie Institutions

Mentee-Mentor Accepted Connections by the MSI-Carnegie Classification

Of the accepted connections between mentees and mentors with institution information (n=1542), mentees' connections represented the following institutions: R1 (660/1542, 42%), HBCU (384/1543, 25%), HSI (262/1542, 17%), R2 (141/1542, 9%), AANAPISI (65/1542, 4%), and TCCU (30/1542, 2%); see Figure 11. Following a similar pattern, 54% (837/1542) of accepted mentor connections were at R1 universities, followed by 19% (290/1542) HBCUs, 16% (245/1542) HSIs, 7% (113/1542) R2 universities, 4% (56/1542) AANAPISIs, and less than 1% (1/1542) TCCUs (see Textbox 1 for a list of definitions for the MSI-Carnegie classification).

Figure 11. Connections by MSI Classification: mentee-mentor. AANAPISI: Asian American and Native American Pacific Islander–serving institution; HBCU: historically Black colleges and universities; HSI: Hispanic-serving institution; MSI: minority-serving institution; TCCU: tribally controlled colleges and universities.

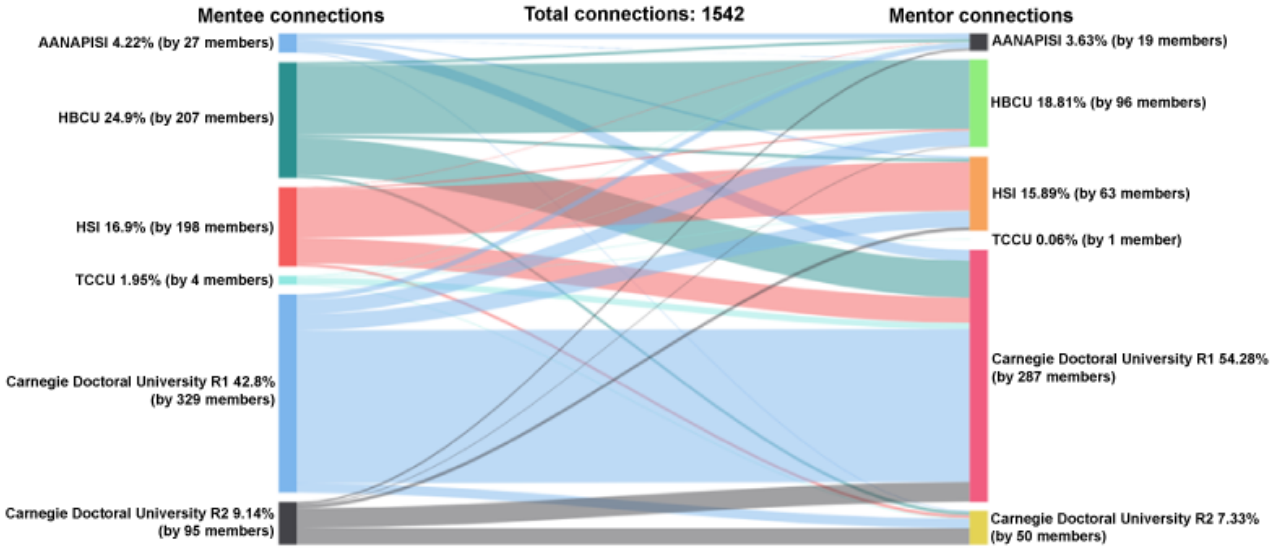

Of the 660 accepted connections from mentees at R1 universities, the mentors represented universities that were also R1 (50/660, 77%), followed by HSI (53/660, 8%), HBCU (49/660, 7%), R2 (32/660, 5%), and AANAPISI (18/660, 3%). Among the 384 accepted connections among mentees at HBCUs, a majority of mentors were from HBCUs (230/384, 60%) or R1 universities (122/384, 32%). The remaining mentors came from other university designation types for less than 3%. There were 262 accepted connections for mentees from HSIs. Most of the connections were with other HSIs (161/262, 61%) and R1 universities (84/262, 32%), followed by R2 universities (10/262, 4%) and HBCUs (6/262, 2%). Of the 141 accepted connections for mentees at R2 universities, almost half connected with R1 mentors (66/141, 47%) and over a third with R2 mentors (54/141, 38%). The remaining mentors came from HSIs (11/141, 8%), AANAPISIs (7/141, 5%), and HBCUs (3/141, 2%). A total of 65 connections were accepted for mentees from AANAPISIs, who connected with mentors from R1 universities (36/65, 55%), AANAPISIs (18/65, 28%), HSIs (7/65, 11%), R2 universities (3/65, 5%), and HBCUs (1/65, 2%). Only 30 connections were accepted for mentees from TCCUs; 70% (21/30) of those connections were with R1 university mentors.

Textbox 1. Minority-serving institution–Carnegie classification list.

- AANAPISI: Asian American and Native American Pacific Islander serving institution
- HBCU: historically black colleges and universities
- HIS: Hispanic-serving institution
- TCCU: tribal college or university
- Carnegie doctoral university R1: doctoral universities—very high research activity
- Carnegie doctoral university R2: doctoral universities—high research activity

Of the 837 connections accepted by R1 university mentors, mentees were from institutions designated as R1 (508/837, 61%), HBCU (122/837, 15%), HSI (84/837, 10%), R2 (66/837, 8%), AANAPISI (36/837, 4%), and TCCU (21/837, 3%). A total of 290 connections were accepted by HBCU mentors, with mentees from HBCUs (230/290, 79%), R1 universities (49/290, 17%), HSIs (6/290, 2%) and R2 universities (3/290, 1%). Of the 245 connections accepted by HSI mentors, mentees represented HSIs (161/245, 66%), R1 (53/245, 22%) and R2 (11/245, 4%) universities, HBCUs (11/245, 4%), AANAPISIs (7/245, 3%), and TCCUs (2/245, 1%). A total of 113 connections were accepted by mentors at R2 universities who

XSL•FO
RenderX

connected with mentees from R2 universities (54/113, 48%), R1 universities (32/113, 28%), HBCUs (11/113, 10%), HSIs (10/113, 9%), AANAPISIs (3/113, 3%), and TCCUs (3/113, 3%). Of the 56 connections accepted by mentors at AANAPISIs, one-third were with mentees from AANAPISIs and R1 universities (18/56 each). The remaining came from HBCUs (10/56, 18%), R2 universities (7/56, 13%), TCCUs (2/56, 4%), and HSIs (1/56, 2%). Only 1 connection was accepted by a mentor at a TCCU.

Mentee-Mentee Accepted Connections by the MSI-Carnegie Classification

HBCU mentees represent 50% (236/472) of initiated and accepted mentee-mentee connections (n=472 connections), followed by 28% (131/472) R1 mentees, 11% (53/472) HSI mentees, 8% (37/472) R2 mentees, 2% (9/472) AANAPISI mentees, and 1% (6/472) TCCU mentees (Figure 12). Among the mentees who received and accepted peer mentee requests, 42% (196/472) were from HBCUs, 37% (174/472) were from R1 universities, 11% (53/472) were from HSIs, 8% (39/472) were from R2 universities, 2% (9/472) were from AANAPISIs, and less than 1% (1/472) were from TCCUs.

Figure 12. Connections by MSI Classification: mentee-mentee. AANAPISI: Asian American and Native American Pacific Islander–serving institution; HBCU: historically Black colleges and universities; HSI: Hispanic-serving institution; MSI: minority-serving institution; TCCU: tribally controlled colleges and universities.

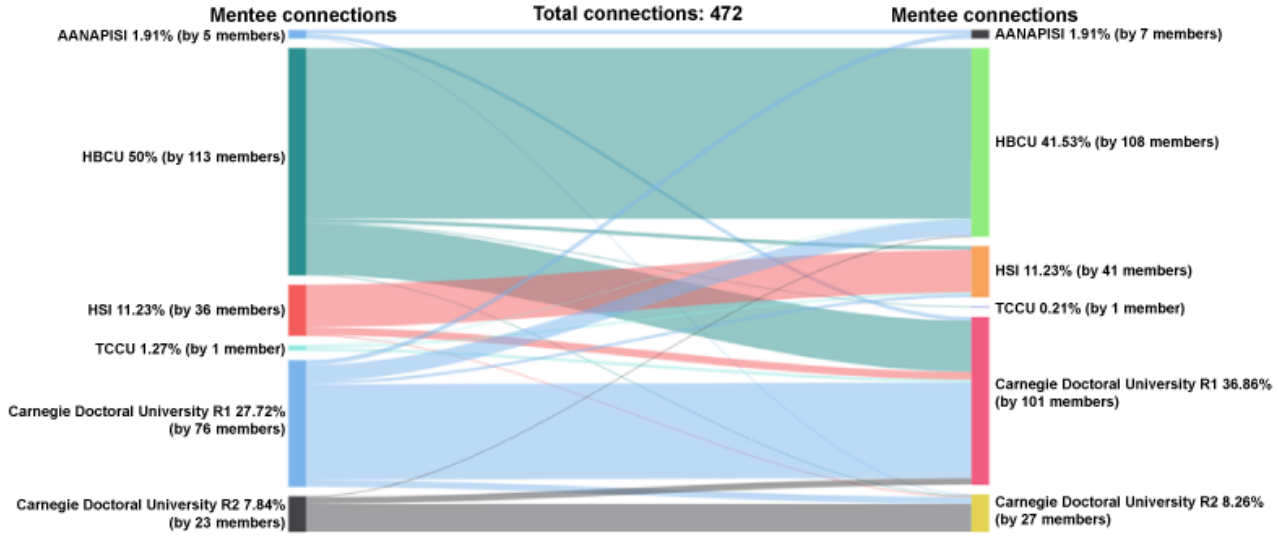

HBCU mentees who initiated peer mentee-mentee connections (n=236 connections) connected with mentees from predominantly HBCUs (177/236, 75%), followed by R1 universities (53/236, 22%) and only 2% (4/236) HSIs. R1 mentees who initiated peer mentee-mentee connections (n=131 connections) connected with mentees from R1 universities (99/131, 76%), HBCUs (17/131, 13%), R2 universities (7/131, 5%), AANAPISIs (5/131, 4%), and HSIs (3/131, 2%). HSI mentees who initiated peer mentee-mentee connections (n=53 connections) connected with mentees from HSIs (44/53, 83%), R1 universities (8/53, 15%), and R2 universities (1/53, 2%). R2 mentees who initiated peer mentee-mentee connections (n=37) connected with R2 universities (29/37, 78%), R1 universities (7/37, 19%), and HBCUs (1/37, 3%). TCCUs and AANAPISIs had fewer than 10 connections each.

Mentor-Mentor Accepted Connections by MSI Designation

Of the mentors who initiated and accepted mentoring connections (n=851 connections), 66% (565/851) were R1 connections followed by 13% (110/851) HBCUs, 8% (66/851) R2 universities, 7% (61/851) HSIs, 6% (48/851) AANAPISIs, and less than 1% (1/851) TCCUs (Figure 13). Among the mentors who received and accepted peer mentor requests, 65% (549/851) were from R1 institutions, 13% (107/851) from HBCU institutions, 10% (85/851) from HSIs, 8% (66/851) from R2 institutions, 5% (43/851) from AANAPISIs, and less than 1% (1/851) from TCCU institutions.

XSL•FO
RenderX

Figure 13. Connections by MSI Classification: mentor-mentor. AANAPISI: Asian American and Native American Pacific Islander–serving institution; HBCU: historically Black colleges and universities; HSI: Hispanic-serving institution; MSI: minority-serving institution; TCCU: tribally controlled colleges and universities.

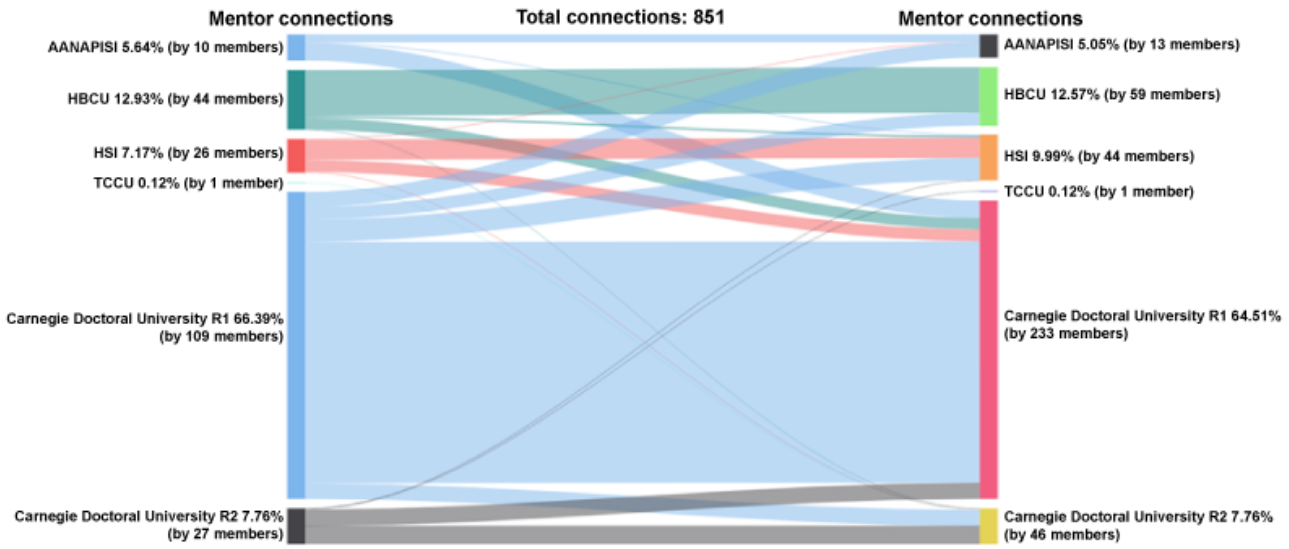

Among the 565 connections initiated by mentors at R1 universities, connections were with mentors at other R1 universities (444/565, 79%), followed by HSIs (41/565, 7%), AANAPISIs (28/565, 5%), R2 universities (29/565, 5%), and HBCUs (23/565, 4%). Among the connections sent by mentors at HBCUs (110 connections), 76% (84/110) were with mentors at other HBCUs, 19% (21/110) to R1 institutions, HSIs (4/110, 4%), and R2 universities (1/110, 1%). R2 mentors connected (n=66 connections) with predominantly R2 (34/66, 52%) or R1 (30/66, 45%) institutions, and only 1 connection each for HSIs and TCCUs. Mentors from HSI locations connected (n=61 connection) with mostly mentors from HSIs (37/61, 61%), followed by R1 universities (22/61, 36%) and only 1 connection each for AANAPISIs and R2 universities. Mentors from AANAPISIs connected (n=48 connections) with R1 mentors (32/48, 67%) and AANAPISI mentors (14/48, 29%). Only 2 connections were made with HSI mentors. Only 1 mentor connection was initiated from a mentor at a TCCU.

Educational Attainment

Mentee-Mentor Accepted Connections by Education Level

Of the accepted connections between mentees and mentors with educational attainment data (n=3578), mentees have the following distribution for education level: undergraduate (1074/3578, 30%), graduate (1076/3578, 30%), postdoctoral fellow (553/3578, 15%), and other (ie, currently working, faculty, etc; 875/3578, 24%); see Figure 14. In contrast, most mentor-accepted connections were with mentors working in the field (2667/3578, 75%), followed by postdocs (369/3578, 10%), undergraduate students (304/3578, 9%), and graduate students (238/3578, 7%).

XSL•FO
RenderX

Figure 14. Connections by educational attainment: mentee-mentor.

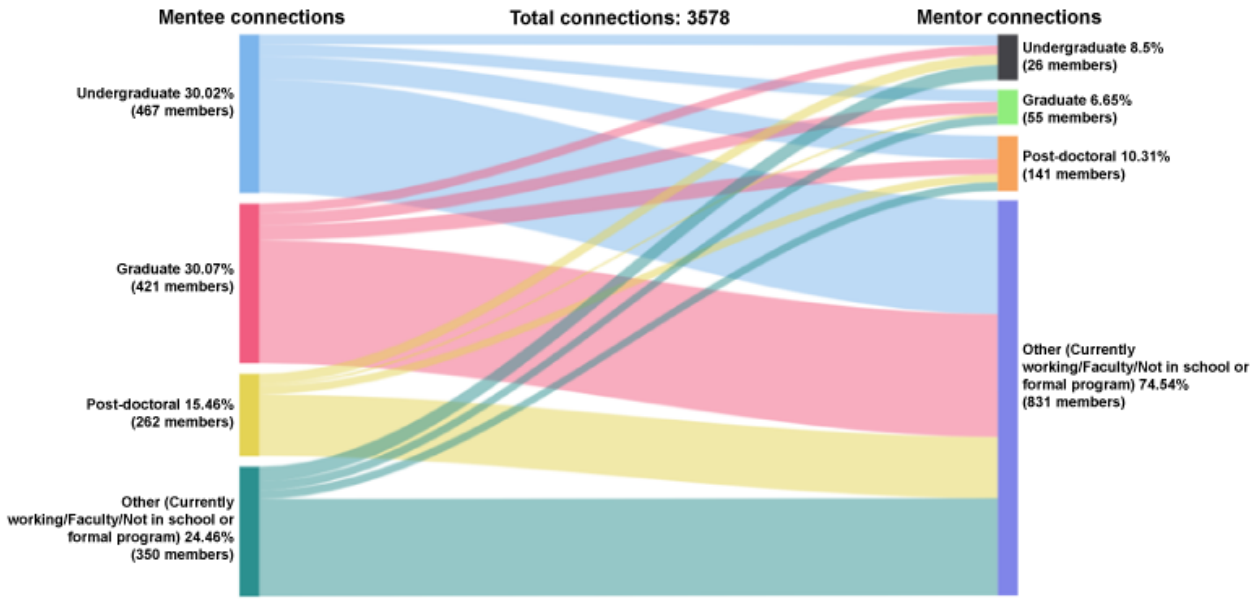

Of the 1074 accepted connections from undergraduate mentees, the mentors were currently working (767/1074, 71%), postdocs (155/1074, 14%), graduate students (81/1074, 8%), and other undergraduate students (71/1074, 7%). Among the 1076 accepted connections from graduate mentees, the mentors were currently working (829/1076, 77%), postdocs (102/1076, 10%), graduate students (81/1076, 8%), and undergraduate students (64/1076, 6%). Fewer mentee connections occurred from people currently working (n=875). They connected with people who were also currently working (657/875, 75%), undergraduate students (100/875, 11%), graduate students (60/875, 7%), and postdocs (58/875, 7%). There were 553 connections from postdocs connecting to people currently working (414/553, 75%), undergraduate students (69/553, 12%), other postdocs (54/553, 10%), and graduate students (16/553, 3%).

Of the 2667 connections accepted by mentors currently working, mentees were graduate students (829/2667, 31%), undergraduate students (767/2667, 29%), others currently working (657/2667, 25%), and postdocs (414/2667, 16%). A similar number of connections were accepted by postdoctoral (n=369 connections) and undergraduate (n=304) mentors. Connections with postdoctoral mentors came from undergraduate (155/369, 42%), graduate school (102/369, 28%), currently working (58/369, 16%), and other postdoctoral (54/369, 15%) mentees. Connections with undergraduate mentors came from working (100/304, 33%), postdoctoral or undergraduate (71/304, 23%), and graduate student (64/304, 21%) mentees. A total of 238 connections were accepted by graduate student mentors from undergraduate (81/238, 34%) and graduate students (81/238, 34%), people currently working (60/238, 25%), and postdocs (16/238, 7%).

Mentee-Mentee Accepted Connections by Educational Attainment

Undergraduate students represent 44% (347/783) of initiated and accepted mentee-mentee connections (n=738 connections), followed by graduate students (198/783, 25%), people currently working (162/783, 21%), and postdocs (76/783, 10%); see Figure 15. Among the mentees who received and accepted peer mentee requests, 26% (282/783) were undergraduate students, 33% (262/783) were people working in the field, 18% (140/783) were graduate students, and 13% (99/783) were postdocs.

XSL•FO
RenderX

Figure 15. Connections by educational attainment: mentee-mentee.

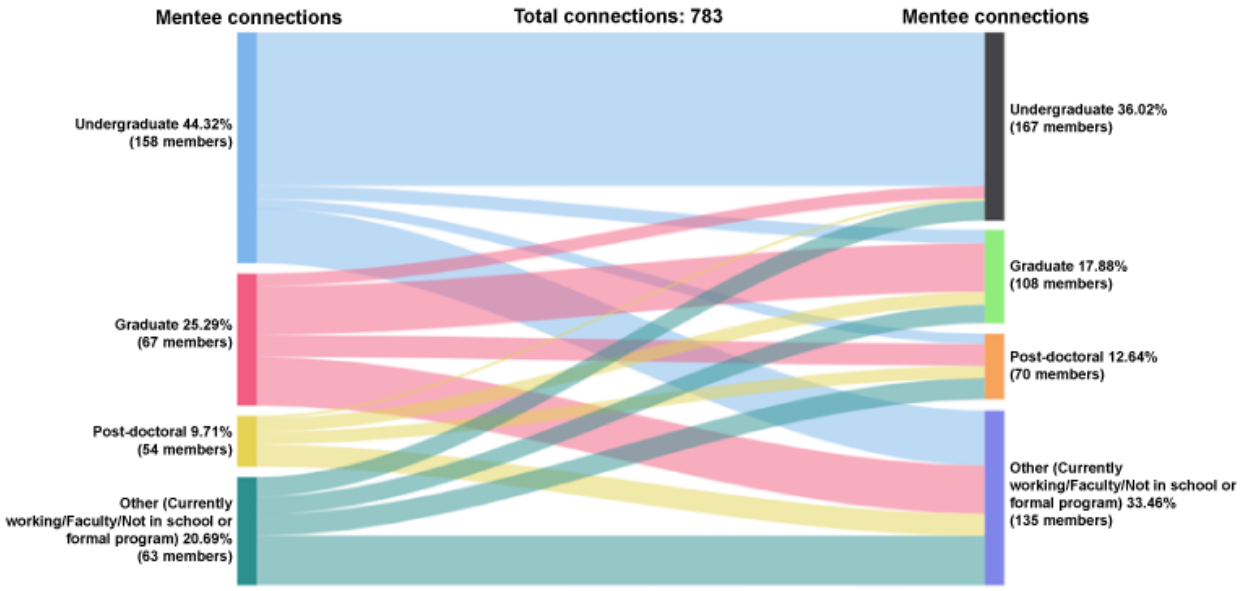

Undergraduate mentees who initiated peer mentee-mentee connections (n=347 connections) connected with undergraduate mentees (230/347, 66%), people working (82/347, 24%), graduate mentees (20/347, 6%), and postdocs (15/347, 4%). Among the 198 connections with graduate mentees, 37% (73/198) were with mentees who were also graduate students, people currently working (73/198, 37%), postdocs (33/198, 17%), and undergraduates (19/198, 6%). Currently working mentees had 162 peer connections with mainly other people working (74/162, 46%), followed by postdocs (32/162, 20%), undergraduate mentees (29/162, 18%), and graduate mentees (27/162, 17%). Postdoc mentees initiated 76 connections with people currently working (33/76, 43%), graduate students (20/76, 26%), other postdocs (19/76, 25%), and undergraduate students (4/76, 5%).

Mentor-Mentor Accepted Connections by Educational Attainment

Of the mentors who initiated and accepted mentoring connections (n=2037 connections), 66% (1348/2037) were with people currently working, followed by undergraduates (323/2037, 16%), graduate students (198/2037, 10%), and postdocs (168/2037, 8%); see Figure 16. Among mentors who received and accepted peer mentor requests, a majority were from people currently working (1688/2037, 83%) followed by postdocs (202/2037, 10%), graduate students (113/2037, 6%), and undergraduate students (34/2037, 2%).

Figure 16. Connections by educational attainment: mentor-mentor.

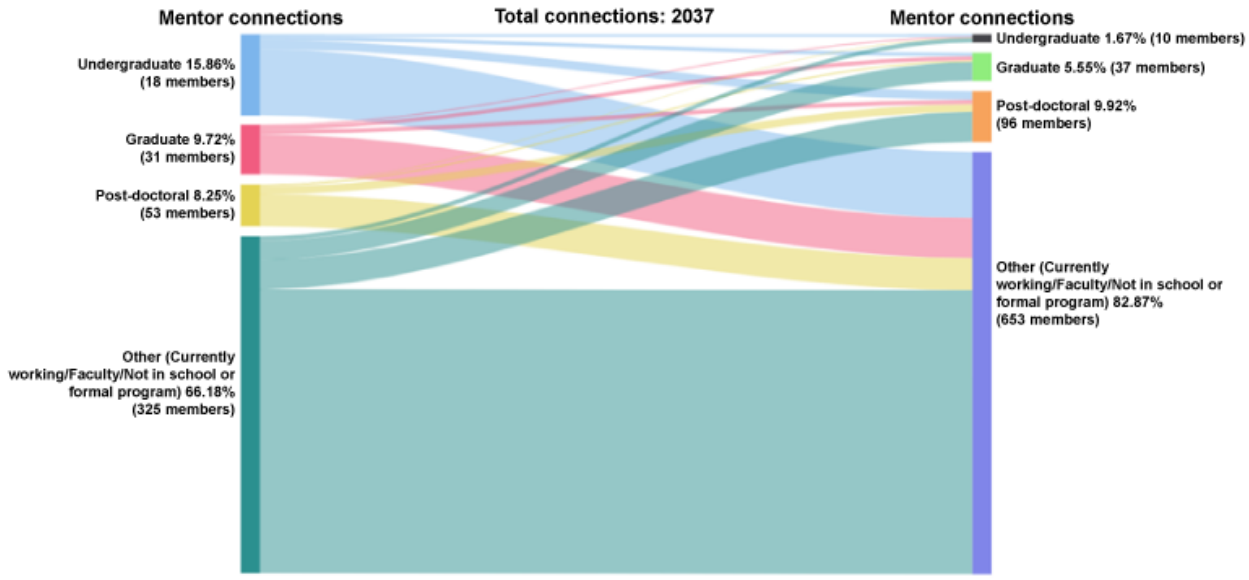

XSL•FO
RenderX

Among the 1348 connections initiated by currently working mentors, connections were with mentors currently working (1136/1348, 84%), postdocs (119/1348, 9%), graduate students (75/1348, 6%), and undergraduate students (18/1348, 1%). Undergraduate mentors initiated 323 connections with mentors currently working (264/323, 82%), postdocs (35/323, 11%), graduate students (15/323, 5%), and undergraduates (9/323, 3%). Within connections from graduate student mentors (198 connections), most were also to people currently working (160/198, 81%), followed by postdocs (17/198, 9%), graduate students (16/198, 8%), and undergraduates (5/198, 3%). Among the connections sent by postdoc mentors (168 connections), most were with either people currently working (128/168, 76%) or other postdocs (31/168, 18%).

Discussion

Principal Findings

Mentoring is a vital aspect of training to create the next generation of a diverse biomedical research workforce. The MyNRMN platform provides a powerful tool to enable remote mentoring across institutions throughout the United States and territories. Before this study, the utility of the mentoring connections and mentoring networks across institutions and how mentors and mentees connected was unknown. This study aimed to examine the large-scale mentoring connections facilitated by our web-based platform between students (mentees) and faculty (mentors) across institutional and geographic boundaries by gender, race, ethnicity, institution type, and educational attainment.

We developed the MyNRMN platform to increase access to diverse mentoring for mentees across institutions and geographic boundaries across the United States and territories. Access to a diverse and more extensive mentoring network grows the social capital of the mentees. It was hypothesized that by providing a remote platform accessible to a large swath of biomedical students, as well as individuals in the workforce, mentoring connections would occur [7]. Our previous work described the intentional recruitment strategies across the United States, with a specific emphasis on minority-serving institutions and conferences that reach diverse audiences [5]. The results of these successful strategies provide evidence demonstrating that the hypothesis is correct, as observed by the substantial number of connections between heterogeneous individuals. As such, the MyNRMN platform is addressing the NIH's initiative to provide mentoring support to underrepresented students and scientists [2,18]. By addressing this initiative, we can now observe the growth of individuals' social capital, which is crucial to their persistence and advancement in the biomedical sciences [7].

As evident from the data in this study, the MyNRMN platform provides a diverse cohort of mentees access to diverse mentors across the nation. Our platform's goal is to increase diversity of the biomedical workforce, and our mentees are predominantly Black compared with other racial groups, representing a key demographic of interest by the NIH [19]. Other racial and ethnic groups also have strong representation on the MyNRMN platform, and there are opportunities to further expand in some subgroups, such as Hispanic or Latinx students. Previous research found that Latina women in the biomedical sciences experience isolation and a low sense of belonging in their undergraduate programs [20]; thus remote mentoring and connections may bridge the gap for less inclusive environments in biomedical science programs. Additionally, we found that a majority of mentees and mentors on this platform identified as female. In fact, there was a gender skew in the proportion of requested and accepted connections to female mentors. This finding may represent a mentoring burden that women, especially women of color, face in academia and training [21]. Future work could assess the MyNRMN mentors' perceived workload and burden in these mentoring roles to determine how to ease any encumbrance through the platform. It would also be imperative to assess the variability of workloads across demographics and identities, such as race and ethnicity, and gender, respectively.

Another benefit of the MyNRMN platform is the connectivity with mentors at other institutions. This ability to connect beyond institutional boundaries is crucial for Black or African American and Hispanic or Latinx individuals as it provides support and builds a community of mentors enabling persistence, providing role models, and increasing social capital in spaces that were previously untapped or underresourced [1]. There were over 4500 connections between mentors and mentees at different institutions, including significant crossover among HBCUs, MSIs, and R1 and R2 institutions. These connections are contingent on a bridge due to location, resources, and physical distance. An added benefit of these connections is that some institutions may not have senior faculty members with the bandwidth for mentoring or even a faculty member within a mentee's specific discipline. Additionally, remote mentoring can help foster conversations that may not occur in an in-person environment as demonstrated by mentoring during the COVID-19 pandemic [22]. Furthermore, having a mentor outside of a person's immediate proximity can provide valuable insight on professional development [23] that cannot be obtained within the organizational culture of a home institution. MyNRMN provides a solution to enable these cross-institutional and long-distance collaborations. Future research should evaluate the ability to recruit and sustain these cross-institutional partnerships on the web-based platform.

The data reported represent connections as the unit of observation rather than persons; thus, some persons may have more connections than others and be more heavily represented in the data. We provided the average number of connections in Table 1 to demonstrate these potential differences by demographic. Additionally, data on demographic characteristics are passively collected based on profile forms on the network, and thus, missing data are an area of concern. As of now, we do not actively collect many elements of diversity, including LGBTQIA+ identity, which is an important facet of the lived experience [24]. Thus, the findings from this paper should be considered in the context of potential limitations. In the future, we plan to adjust our data profile fields to be representative of sexual and gender underrepresented groups. Moreover, we did not present data on nonaccepted connections, meaning a user reached out to another user to connect, but the connection was

XSL•FO
RenderX

not accepted. It should be noted that all mentor and mentee participation on the platform is voluntary, and competing demands may result in a nonacceptance. We have also observed that connection acceptances are cyclical with the ebbs and flows of the academic semesters when workloads may shift. A mentor's participation is also conditional based on their capacity to engage with additional mentees outside of any requirements or obligations for their position. Future research will explore user experiences with the platform to inform adjustments to meet end-user needs.

Conclusions

Access to mentors is crucial for career advancement and increasing our nation's current and future biomedical workforce. We developed a web-based national platform to connect mentees and mentors across institutions and geographic boundaries toward this goal. The MyNRMN platform is achieving this goal by facilitating mentoring connections and developing diverse mentoring networks for diverse mentees. By analyzing the organic evolution of mentoring connections throughout MyNRMN, we can observe the value of facilitating and nurturing these connections. In this study, we examined large-scale mentoring connections and the diversity of these connections and addressed a gap in our understanding of how mentees and mentors connect across institutions throughout the United States. We observed that a web-based remote space for mentors and mentees to connect and build their network can enable diverse connections between mentors and mentees.

Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award numbers U54GM119023 and U24GM132217. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors' Contributions

This study was conceptualized by TAS, ELT, ZL, DJ, KS, and JKV. Analysis was conducted by TAS and ZL. The MyNRMN platform was developed by TAS, ZL, and JJ. The original draft of the manuscript was written by TAS, ELT, ZL, DJ, KS, and GS. Review, editing, and revisions were completed by TAS, ELT, ZL, JJ, DJ, KS, GS, and JKV.

Conflicts of Interest

None declared.

References

1. Montgomery BL. Building and sustaining diverse functioning networks using social media and digital platforms to improve diversity and inclusivity. Front Digit Humanit 2018;5(2018):1-11 [FREE Full text] [doi: 10.3389/fdigh.2018.00022]
2. Valantine HA, Collins FS. National Institutes of Health addresses the science of diversity. Proc Natl Acad Sci U S A 2015 Oct 06;112(40):12240-12242 [FREE Full text] [doi: 10.1073/pnas.1515612112] [Medline: 26392553]
3. Valantine HA, Lund PK, Gammie AE. From the NIH: a systems approach to increasing the diversity of the biomedical research workforce. CBE Life Sci Educ 2016;15(3):1-5 [FREE Full text] [doi: 10.1187/cbe.16-03-0138] [Medline: 27587850]
4. Swartz TH, Palermo AGS, Masur SK, Aberg JA. The science and value of diversity: closing the gaps in our understanding of inclusion and diversity. J Infect Dis 2019 Aug 20;220(220 Suppl 2):S33-S41 [FREE Full text] [doi: 10.1093/infdis/jiz174] [Medline: 31430380]
5. Javier D, Stinson K, Zavala M, Ahmed T, Vishwanatha JK. NRMNet: building a national resource for mentorship, networking and professional development to enhance diversity. Ethn Dis 2021;31(3):469-480 [FREE Full text] [doi: 10.18865/ed.31.3.469] [Medline: 34295135]
6. Sorkness CA, Pfund C, Ofili EO, Okuyemi KS, Vishwanatha JK, NRMN team, Duplicate Of Weber-Main, et al. A new approach to mentoring for research careers: the National Research Mentoring Network. BMC Proc 2017;11(Suppl 12):171-200 [FREE Full text] [doi: 10.1186/s12919-017-0083-8] [Medline: 29375663]
7. Ahmed T, Johnson J, Latif Z, Kennedy N, Javier D, Stinson K, et al. MyNRMN: A national mentoring and networking platform to enhance connectivity and diversity in the biomedical sciences. FASEB Bioadv 2021 Jul;3(7):497-509 [FREE Full text] [doi: 10.1096/fba.2020-00102] [Medline: 34258519]
8. Estrada M, Hernandez PR, Schultz PW. A longitudinal study of how quality mentorship and research experience integrate underrepresented minorities into STEM careers. CBE Life Sci Educ 2018;17(1):1-13 [FREE Full text] [doi: 10.1187/cbe.17-04-0066] [Medline: 29351912]
9. Haggard DL, Dougherty TW, Turban DB, Wilbanks JE. Who is a mentor? A review of evolving definitions and implications for research. Journal of Management 2010 Oct 28;37(1):280-304 [FREE Full text] [doi: 10.1177/0149206310386227]
10. Packard BWL. Successful STEM Mentoring Initiatives for Underrepresented Students: A Research-Based Guide for Faculty and Administrators. Sterling, VA: Stylus Publishing; 2015.

XSL•FO
RenderX

11. Atkins K, Dougan BM, Dromgold-Sermen MS, Potter H, Sathy V, Panter AT. "Looking at Myself in the Future": how mentoring shapes scientific identity for STEM students from underrepresented groups. Int J STEM Educ 2020;7(1):1-15 [FREE Full text] [doi: 10.1186/s40594-020-00242-3] [Medline: 32850287]
12. McGee R, Saran S, Krulwich TA. Diversity in the biomedical research workforce: developing talent. Mt Sinai J Med 2012;79(3):397-411 [FREE Full text] [doi: 10.1002/msj.21310] [Medline: 22678863]
13. Thomas KM, Willis LA, Davis J. Mentoring minority graduate students: issues and strategies for institutions, faculty, and students. Equal Oppor Int 2007;26(3):178-192. [doi: 10.1108/02610150710735471]
14. Zambrana RE, Ray R, Espino MM, Castro C, Douthirt CB, Eliason J. "Don't leave us behind": the importance of mentoring for underrepresented minority faculty. Am Educ Res J 2015;52(1):40-72 [FREE Full text] [doi: 10.3102/0002831214563063]
15. Montgomery BL. Mapping a mentoring roadmap and developing a supportive network for strategic career advancement. SAGE Open 2017 Jun 03;7(2):1-13 [FREE Full text] [doi: 10.1177/2158244017710288]
16. Deanna R, Merkle BG, Chun KP, Navarro-Rosenblatt D, Baxter I, Oleas N, et al. Community voices: the importance of diverse networks in academic mentoring. Nat Commun 2022;13(1):1-7 [FREE Full text] [doi: 10.1038/s41467-022-28667-0] [Medline: 35338138]
17. The Carnegie Classification of Institutions of Higher Education. Indiana University Center for Postsecondary Research. 2021. URL: https://tinyurl.com/2ufbwc5e [accessed 2024-04-30]
18. Enhancing the diversity of the NIH-funded workforce. National Institute of General Medical Sciences. 2022. URL: https://www.nigms.nih.gov/training/dpc [accessed 2024-04-30]
19. Notice of NIH's interest in diversity. National Institutes of Health. 2019. URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-031.html [accessed 2024-04-30]
20. Rodriguez SL, Blaney JM. "We're the unicorns in STEM": understanding how academic and social experiences influence sense of belonging for Latina undergraduate students. J Divers High Educ 2021;14(3):441-455 [FREE Full text] [doi: 10.1037/dhe0000176]
21. Casad BJ, Franks JE, Garasky CE, Kittleman MM, Roesler AC, Hall DY, et al. Gender inequality in academia: problems and solutions for women faculty in STEM. J Neurosci Res 2021;99(1):13-23. [doi: 10.1002/jnr.24631] [Medline: 33103281]
22. Speer JE, Lyon M, Johnson J. Gains and losses in virtual mentorship: a descriptive case study of undergraduate mentees and graduate mentors in STEM research during the COVID-19 pandemic. CBE Life Sci Educ 2021;20(2):1-9 [FREE Full text] [doi: 10.1187/cbe.20-06-0128] [Medline: 33734867]
23. Seal DW, Smith NG, Sun CJ, Thompson EL. Reflections on a mentoring partnership journey. Health Behav Res 2019;2(4):1-5 [FREE Full text] [doi: 10.4148/2572-1836.1060]
24. Cech EA, Waidzunas TJ. Systemic inequalities for LGBTQ professionals in STEM. Sci Adv 2021;7(3):1-9 [FREE Full text] [doi: 10.1126/sciadv.abe0933] [Medline: 33523910]

Abbreviations

AANAPISI: Asian American and Native American Pacific Islander–serving institution
HBCU: historically Black colleges and universities
HSI: Hispanic-serving institution
MSI: minority-serving institution
NIH: National Institutes of Health
NRMN: National Research Mentoring Network
TCCU: tribally controlled colleges and universities

Edited by K Williams; submitted 24.03.23; peer-reviewed by N Mungoli, D Carvalho, E Krug; comments to author 05.10.23; revised version received 30.11.23; accepted 09.03.24; published 17.06.24.

Please cite as:
Syed TA, Thompson EL, Latif Z, Johnson J, Javier D, Stinson K, Saleh G, Vishwanatha JK
Diverse Mentoring Connections Across Institutional Boundaries in the Biomedical Sciences: Innovative Graph Database Analysis
J Med Internet Res 2024;26:e47560
URL: https://www.jmir.org/2024/1/e47560
doi:10.2196/47560
PMID:38885013

©Toufeeq Ahmed Syed, Erika L Thompson, Zainab Latif, Jay Johnson, Damaris Javier, Katie Stinson, Gabrielle Saleh, Jamboor K Vishwanatha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association Between Prosuicide Website Searches Through Google and Suicide Death in the United States From 2010 to 2021: Lagged Time-Series Analysis

Nora Clancy Kelsall[1], MS; Catherine Gimbrone[1], MPH; Mark Olfson[2], MD, MPH; Madelyn S Gould[1,2], MD, MPH; Jeffrey Shaman[3,4], PhD; Katherine Keyes[1], PhD, MPH

[1]Department of Epidemiology, Columbia University, New York, NY, United States
[2]Department of Psychiatry, Columbia University, New York, NY, United States
[3]Department of Environmental Health Sciences, Columbia University, New York, NY, United States
[4]Columbia Climate School, Columbia University, New York, NY, United States

Corresponding Author:
Nora Clancy Kelsall, MS
Department of Epidemiology
Columbia University
722 West 168th Street, Room 733
New York, NY, 10032
United States
Phone: 1 2023600113
Email: nk3067@cumc.columbia.edu

Abstract

Background: The rate of suicide death has been increasing, making understanding risk factors of growing importance. While exposure to explicit suicide-related media, such as description of means in news reports or sensationalized fictional portrayal, is known to increase population suicide rates, it is not known whether prosuicide website forums, which often promote or facilitate information about fatal suicide means, are related to change in suicide deaths overall or by specific means.

Objective: This study aimed to estimate the association of the frequency of Google searches of known prosuicide web forums and content with death by suicide over time in the United States, by age, sex, and means of death.

Methods: National monthly Google search data for names of common prosuicide websites between January 2010 and December 2021 were extracted from Google Health Trends API (application programming interface). Suicide deaths were identified using the CDC (Centers for Disease Control and Prevention) National Vital Statistics System (NVSS), and 3 primary means of death were identified (poisoning, suffocation, and firearm). Distributed lag nonlinear models (DLNMs) were then used to estimate the lagged association between the number of Google searches on suicide mortality, stratified by age, sex, and means, and adjusted for month. Sensitivity analyses, including using autoregressive integrated moving average (ARIMA) modeling approaches, were also conducted.

Results: Months in the United States in which search rates for prosuicide websites increased had more documented deaths by intentional poisoning and suffocation among both adolescents and adults. For example, the risk of poisoning suicide among youth and young adults (age 10-24 years) was 1.79 (95% CI 1.06-3.03) times higher in months with 22 searches per 10 million as compared to 0 searches. The risk of poisoning suicide among adults aged 25-64 was 1.10 (95% CI 1.03-1.16) times higher 1 month after searches reached 9 per 10 million compared with 0 searches. We also observed that increased search rates were associated with fewer youth suicide deaths by firearms with a 3-month time lag for adolescents. These models were robust to sensitivity tests.

Conclusions: Although more analysis is needed, the findings are suggestive of an association between increased prosuicide website access and increased suicide deaths, specifically deaths by poisoning and suffocation. These findings emphasize the need to further investigate sites containing potentially dangerous information and their associations with deaths by suicide, as they may affect vulnerable individuals.

(J Med Internet Res 2024;26:e53404) doi:10.2196/53404

XSL•FO
RenderX

KEYWORDS

pro-suicide forum; suicide; google search; social media; online forum; internet search; death; United States; suicide death; forum; analysis; association; poisoning; suffocation

Introduction

Death by suicide in the United States has increased from 10.7 per 100,000 in 2001 to 14.1 per 100,000 in 2021 [1], accelerating its public health burden. Rates of death by suicide have increased across all ages, with the greatest increases among men and those aged 10-24 years [1] While known individual-level determinants of suicide death include past nonfatal suicide attempts, depressive disorders, stressful life events, and lack of social support [2], understanding potential novel risk factors is necessary for addressing the increasing rates of suicidality.

There is evidence that exposure to content that is explicit regarding suicide, most notably after inappropriate reporting of high-profile suicide deaths, such as description of specific means, increases population rates of suicide [3,4]. It is less well understood whether novel sources of information about suicide, such as prosuicide websites with discussion forums, are related to recent changes in suicide death rates. While the majority of suicide-related websites provide resources such as helplines and encouragement to seek mental health care [5,6], not all websites share this approach. Prosuicide websites are those that promote or facilitate suicide, including advocating for an individual's "right" to end their own life, creating communities of individuals who desire to end their own lives, as well as sharing explicit information about means of killing oneself. [7,8]. While prosuicide websites are relatively few in number [5,6] and accessibility differs by browser [9,10], search engine [11], and search terms [11,12], they can easily be reached across the world [11,13,14], and many suicide forums and related explicit information are accessible to the public without age or other restriction, and thus, are available to suicidal and otherwise vulnerable individuals [15-17].

Prosuicide website access may affect suicide risk through several different mechanisms. While some individuals find comfort or social support in the communities created on the forums, for others these sites could contribute to escalating plans to end their lives [16,18-20]. Access to the sites might also promote suicidal behavior and encourage individuals with psychiatric disorders to seriously consider or attempt suicide by normalizing suicide death or making it a more realistic option. The Interpersonal Theory of Suicide suggests that the repeated exposure to painful or fearful events may provide an increased acquired capability for suicide by reducing fear surrounding suicide and violence, and thus increase risk to an individual [21]. This construct may become pertinent when individuals engage in prolonged interactions with prosuicide websites, experiencing recurrent exposure to instances of self-harm by fellow users, alongside the dissemination of explicit details regarding methods of suicide. Users often share information about lethal means of death, which may increase the probability that an attempt will be fatal, in addition to providing supportive messages about chosen means. However, attempts through means such as firearms, which are highly lethal, are often discouraged due to the potentials for long-term disability if an attempt does not result in death, and for the greater likelihood that an attempt will be aborted when firearms are involved [22].

Previous findings are mixed on the impact of prosuicide websites. One study found an increase in suicidal ideation in those who had used the internet for suicide or mental health related reasons, including searching for suicide methods [23], whereas another found no significant association with suicidal ideation [19]. However, there remains limited information on the population level effects of these websites. In addition, some literature on the understanding of the effect of these websites may be outdated [7,8] and have limited use due to the rapidly changing landscape of contemporary internet use.

Google trend data on suicide related terms have previously been shown to be associated with the suicide rate globally [24,25], yet previous studies have not included specific terms for prosuicide websites, a potentially important addition. The present study aimed to estimate the association of the frequency of Google searches of known prosuicide web forums and content with death by suicide over time in the United States, by age as well as means of suicide death.

Methods

Google Searches

National monthly Google search data for names of common prosuicide websites and related terms between January 2010 and December 2021 were extracted from Google Health Trends API (application programming interface) [26] (previously the Google Flu Trends API). We queried combinations of search terms including names of prosuicide websites and related terms (exact search terms available upon request). Queries randomly select 10%-15% of total Google searches for the given month and returned the probability of a search for the given terms per 10 million searches. Geographies below the national level (eg, state or designated market area) were largely suppressed for many of the study months due to insufficient search numbers; thus, we focus on national search numbers. We requested 10 data sets and averaged the results to estimate monthly search levels.

Deaths by Suicide

We used restricted data from the CDC National Vital Statistics System (NVSS) to identify monthly US national suicide deaths. Suicide deaths were defined as those that listed X60-X84 (*ICD-10* [*International and Statistical Classification of Diseases, Tenth Revision*]), intentional self-harm, U03, intentional self-harm resulting from terrorism, and Y87.0 (*ICD-10*), sequelae of intentional self-harm, as an underlying cause of death. Deaths were further categorized using all listed causes of death by means including: firearm (X72-X74), suffocation (X70), and poisoning (X60-X69). These means categories encompassed 92.4% of the total national suicides between 2010 and 2021; other means were not sufficiently prevalent for analysis. We also identified deaths that involved opioid overdose

XSL•FO
RenderX

as a subcategory of poisoning for sensitivity analyses. Suicides involving opioids were identified as any suicide as defined above, which also listed any cause of death as T40.0-T40.4 or T40.6. We aggregated counts of death by suicide for 2 age groups: 10-24 and 25-64 years by sex and by means category. We excluded the oldest age group of 65 years and older due to potentially differing patterns of internet use. Yearly national population data by age and sex were collected from the national census bureau.

Statistical Analysis

For the primary statistical analysis, we used distributed lag nonlinear models (DLNMs) [27] to estimate the lagged association between the number of Google searches on suicide mortality. DLNMs describe the association between an exposure and effect as well as a lag and effect. Monthly Google search data were used as the explanatory time series, and suicide deaths, stratified by age, sex, and means were analyzed as dependent time series. User demographics are not available for Google data. From the NVSS data, deaths were stratified by age, sex, and means, due to robust evidence that temporal and seasonal patterns of death by suicide vary by age [28,29], sex [30], as well as means [31,32]. Age was stratified into 10-24 and 25-64 years to determine differences among youth and young adults versus adult populations. These stratifications were used to evaluate potential differential associations with suicide deaths in these populations. Therefore, we investigated each dependent time series separately as we aimed to explore how searches may vary in association with different patterns of deaths. Peaks (see Figure 1) related to events such as high-profile news articles on the topic were removed from the explanatory time series.

Figure 1. Monthly search volume of a comprehensive set of terms encompassing prosuicide websites and related terms between January 2010 and December 2021. Noted peaks were removed prior to analysis. Note prosuicide resource release refers to initial publication of a website or book with prosuicide material.

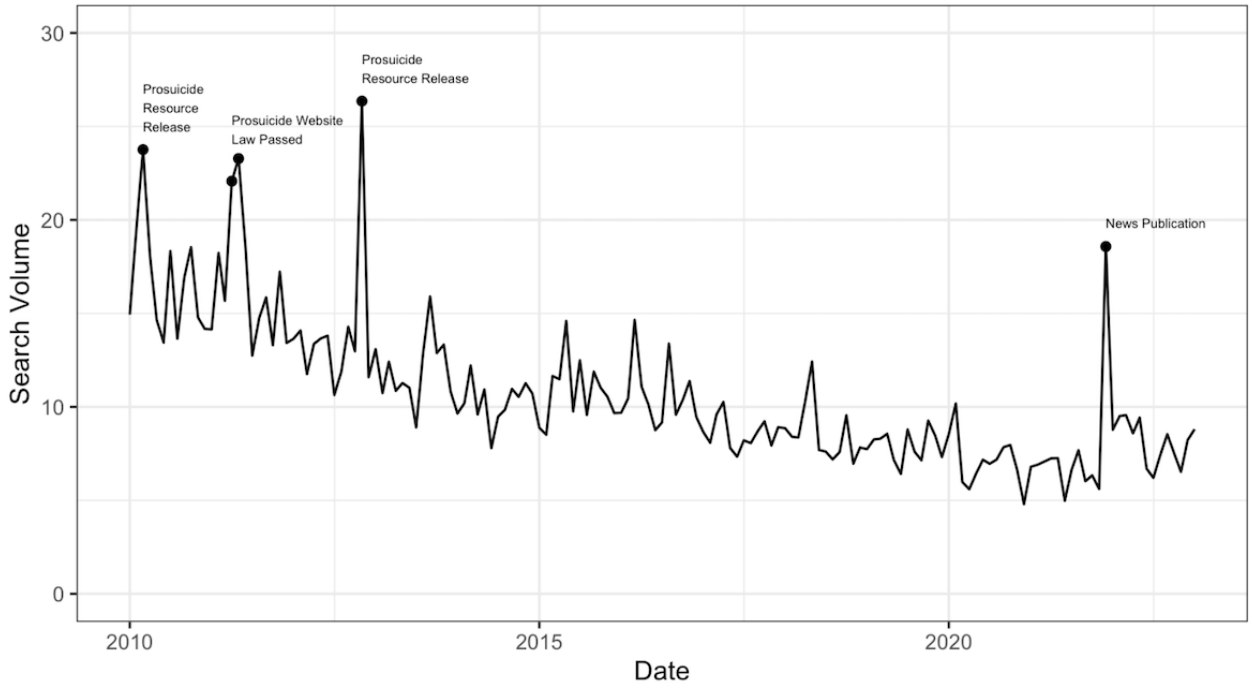

DLNMs use a cross-basis function, which is a 2D matrix that describes the exposure-lag-response relationship using known transformations [33]. To define the cross-basis function of the Google searches, we used 3 months as the maximum lag based on previous literature examining the association between Google searches and suicide deaths [23,24,34]. We then tested linear, polynomial, and multiple degree of freedom splines for the lag-response and exposure-response. The best fitting model was selected based on the Akaike information criterion of each model [35]. In our final models, we used a linear model of the lag-response relationship and a spline with 3 degrees of freedom to model the exposure-response relationship.

We then used the cross-basis function of searches in a generalized linear Poisson model controlling for month of the year (to account for seasonality) using a spline with 4 degrees of freedom, as well as an offset term for population. We used 0 searches as a reference point. Analyses were stratified by age at death, sex, and cause of death. Results with 95% CIs that did not include 1, as computed using a normal approximation of the estimators, were considered statistically significant. Results only discuss statistically significant results.

Sensitivity Analyses

To test whether results were robust to alternative model specifications, we also used (seasonal) autoregressive integrated moving average [(S)ARIMA] models to evaluate the correlation between the explanatory and dependent time series following the steps described in Tran et al 2017. We first fit a (S)ARIMA model to the explanatory time series. We auto-selected this model using the *forecast* package in R, which considers the possibility of a seasonal component. This model was evaluated using plots of the standardized residuals, autocorrelation functions of the residuals, and Q-Q plots of the standardized residuals. The residuals from this model were considered the

XSL•FO
RenderX

prewhitened explanatory time series. The same model was then fit to the dependent time series, and the residuals from this model were considered filtered dependent time series. The correlation between the prewhitened explanatory time series and the filtered dependent time series was then evaluated using a cross-correlation function. A positive correlation at a negative lag would indicate that an increase in searches occurred prior to an increase in deaths, and vice versa for a negative correlation. Like the DLNM models, we used 3 months as our maximum lag and stratified by age, sex, and cause of death.

To estimate whether results from the DLNM models were robust to the reporting period, we conducted sensitivity analyses including shorter time periods. Two time periods were selected that were an equal number of months, from January 2010 to December 2015, and January 2016 to December 2021. The same process for the distributed lag models was followed separately for each of these time periods. In addition, given that opioid deaths have significantly increased in the United States in the past 10 years [36], we wanted to evaluate whether the results may be spuriously produced due to increasing trends in suicide deaths that involved opioids. Thus, to evaluate a potential differential association between searches and opioid deaths, lagged associations between suicide deaths from only opioid poisoning, and only nonopioid poisoning, were also evaluated.

We also repeated the DLNM analyses using Google search terms that were related to suicide but not specific to prosuicide websites. The search terms that were combined and used for this analysis included “suicide -squad,” “suicides,” and “suicidality,” which were drawn from Tran et al 2017 [24]. The minus sign in “suicide -squad” indicates removing the searches with the word squad, in order to avoid references to the movie. We considered other search terms that are reliably associated with suicide deaths, especially those about specific methods. Given that such terms may direct individuals to suicide website forums, we chose more general terms. Analysis for these terms followed the same DLNM procedure as the primary search terms as described above.

Ethical Considerations

The New York State Institute Institutional Review Board determined that this research was exempt from human participants review (protocol AAAS5604).

Results

Monthly Google search data were collected for the time period between January 2010 and December 2021 (Figure 1). There were 4 distinct peaks during this period, during the months of March 2010, April-May 2011, November 2012, and December 2021. These peaks corresponded to releases of prosuicide resources, passing of relevant laws, and a news article, and were removed from the time series for analysis.

Deaths by suicide varied by age group, and rates of death by suicide were higher in 2021 than in 2010 for all age groups. Individuals aged 10-24 years had the lowest rates of death by suicide (Figure 2, Table 1).

Figure 2. Monthly rates of death by suicide per 100,000 between January 2010 and December 2021 for age groups 10-24 and 25-64 years.

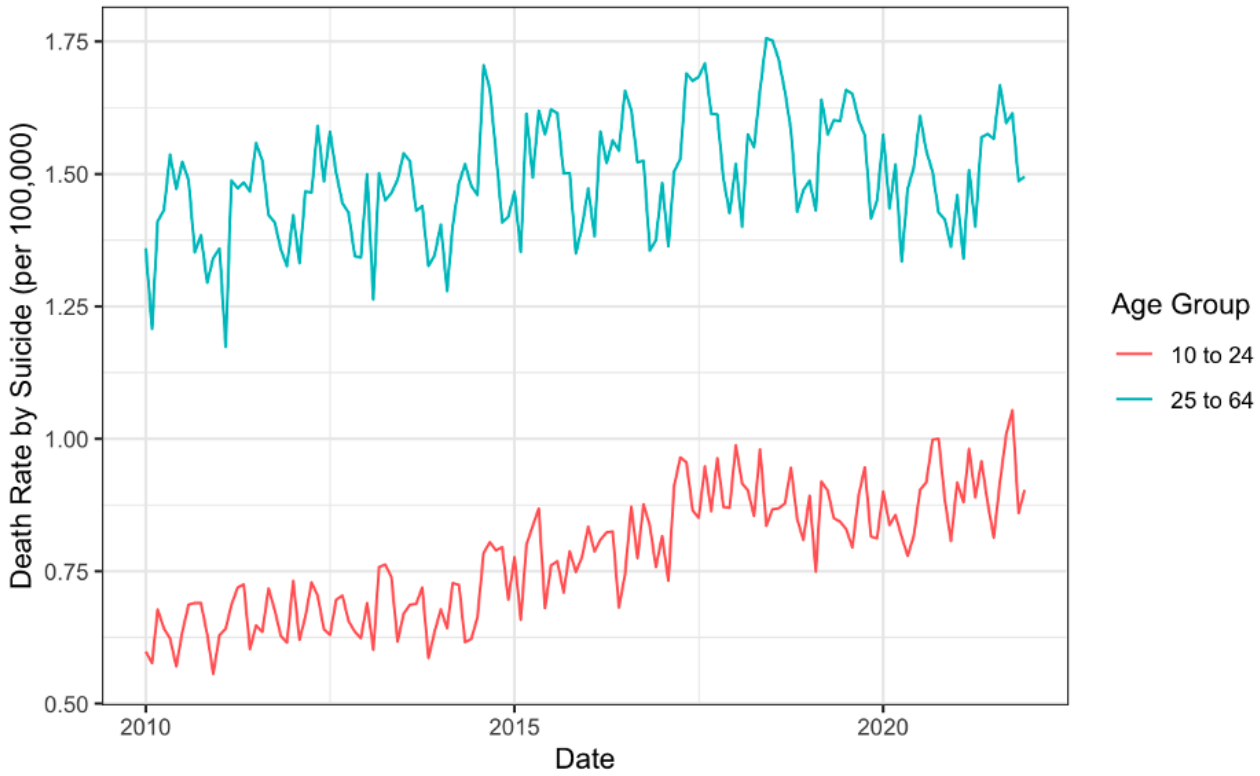

XSL•FO
RenderX

Table 1. Demographics and means of deaths by suicide across age between 2010 and 2021.

	10-24 (N=71,961), n (%)	25-64 (N=362,344), n (%)
Sex		
Male	57,053 (79.3)	277,968 (76.7)
Female	14,908 (20.7)	84,376 (23.3)
Means[a]		
Firearm	33,516 (46.6)	169,661 (46.8)
Suffocation	27,301 (37.9)	102,951 (28.4)
Poisoning	6397 (8.9)	66,476 (18.3)

[a]Means are not mutually exclusive.

DLNM Models

Youth Aged 10-24 Years

Figure 3 shows a plot of the association between search rates and deaths among those aged 10-24 years by sex and method at lags of 0 months (concurrent), 1 month, 2 months, and 3 months; areas in red are those with a positive association (increases in searches were associated with higher suicide rates), and areas in blue are those with a negative association (increases in searches were associated with lower suicide rates). A positive association at a lag of 1 month means that an increase in searches was associated with an increase in deaths by suicide in the following month. Associations with the total deaths among youth and young adults aged 10-24 years as well as deaths by sex showed no significant associations (Figures 3A-3C).

Figure 3. Log of relative risks of an increase from 0 searches for (A) total suicides for those aged 10-24 years, (B) male suicides among (10-24 years), (C) female suicides (10-24 years), (D) poisoning suicides (10-24 years), (E) suffocation suicides (10-24 years), and (F) firearm suicides (10-24 years). Boxed areas are statistically significant.

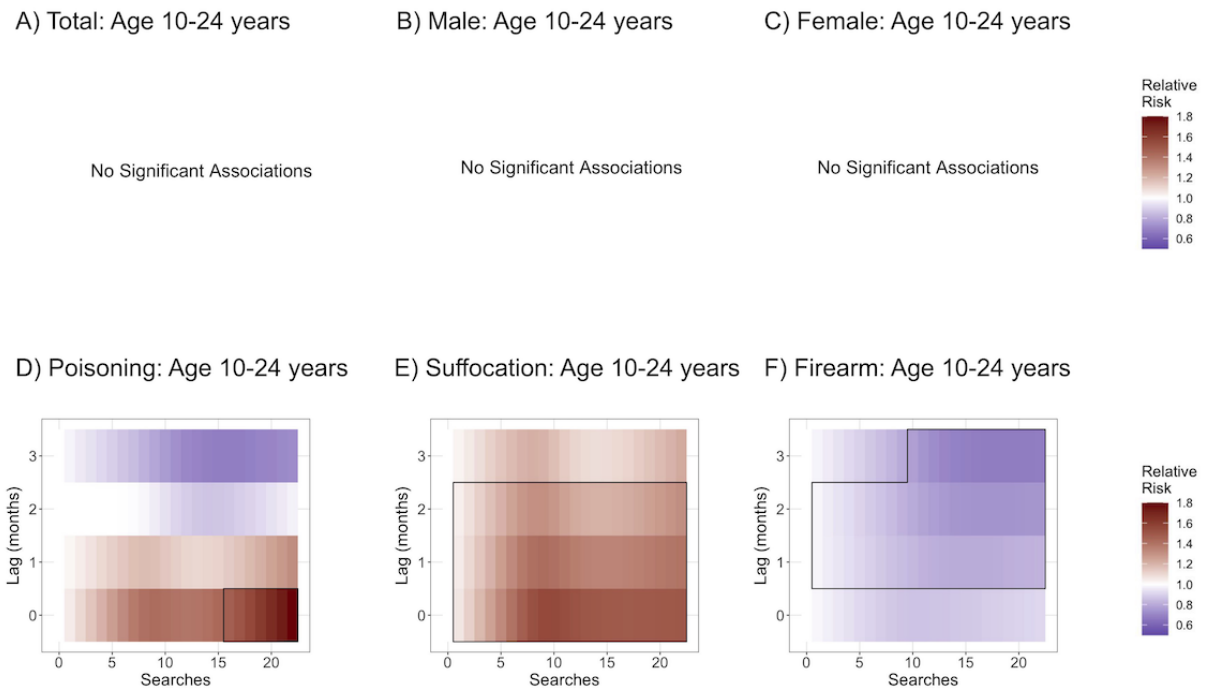

Associations with youth intentional poisoning deaths showed a positive association at lag 0 with a maximum relative risk of 1.79 (95% CI 1.06-3.03) at the maximum search volume of 22 searches per 10 million as compared to no searches, indicating that the risk of suicide was 1.79 times higher in months with 22 searches per 10 million as compared to 0 searches. There were no significant association at lags 1, 2, or 3 (Figure 3D).

For deaths involving suffocation, there was a significant positive association at lags 0, 1, and 2 months with maximum relative risks as compared to no searches of 1.55 (95% CI 1.23-1.96) at a search volume of 9 per 10 million, 1.43 (95% CI 1.26-1.61) at a search volume of 9 per 10 million, and 1.31 (95% CI 1.09-1.59) at a search volume of 8 per 10 million, respectively, with the strongest association at lag 0 (Figure 3E).

XSL•FO
RenderX

Associations with firearm deaths showed a significant negative association at lags 1, 2, and 3 months with minimum relative risks as compared to no searches for lags 1, 2, and 3 months of 0.80 (95% CI 0.71-0.91) at a search volume of 14 per 10 million, 0.74 (95% CI 0.64-0.85) at a search volume of 15 per 10 million, and 0.67 (95% CI 0.50-0.90) at a search volume of 17 per 10 million, respectively. The reverse association was strongest at a lag of 3 months (Figure 3F).

Adults Aged 25-64 years

The association between search rates and deaths by suicide for adults aged 25-64 years old shown in Figure 4 differed from associations found for youth. Search association with total adult deaths was statistically significant at lags 1 and 2 months, with a maximum relative risk as compared to no searches of 1.10 (95% CI 1.03-1.16) at a search volume of 9 per 10 million and 1.09 (95% CI 1.02-1.16) at a search volume of 8 per 10 million, respectively, indicating that the strongest signal for adults aged 25-64 years was for increased search rates in a given month associated with increased suicide deaths the following month and 2 months later.

Figure 4. Log of relative risks of an increase from 0 searches for (A) total suicides for those aged 25-64 years, (B) male suicides (25-64 years), (C) female suicides (25-64 years), (D) poisoning suicides (25-64 years), (E) suffocation suicides (25-64 years), and (F) firearm suicides (25-64 years). Boxed areas are statistically significant (*P*<.05).

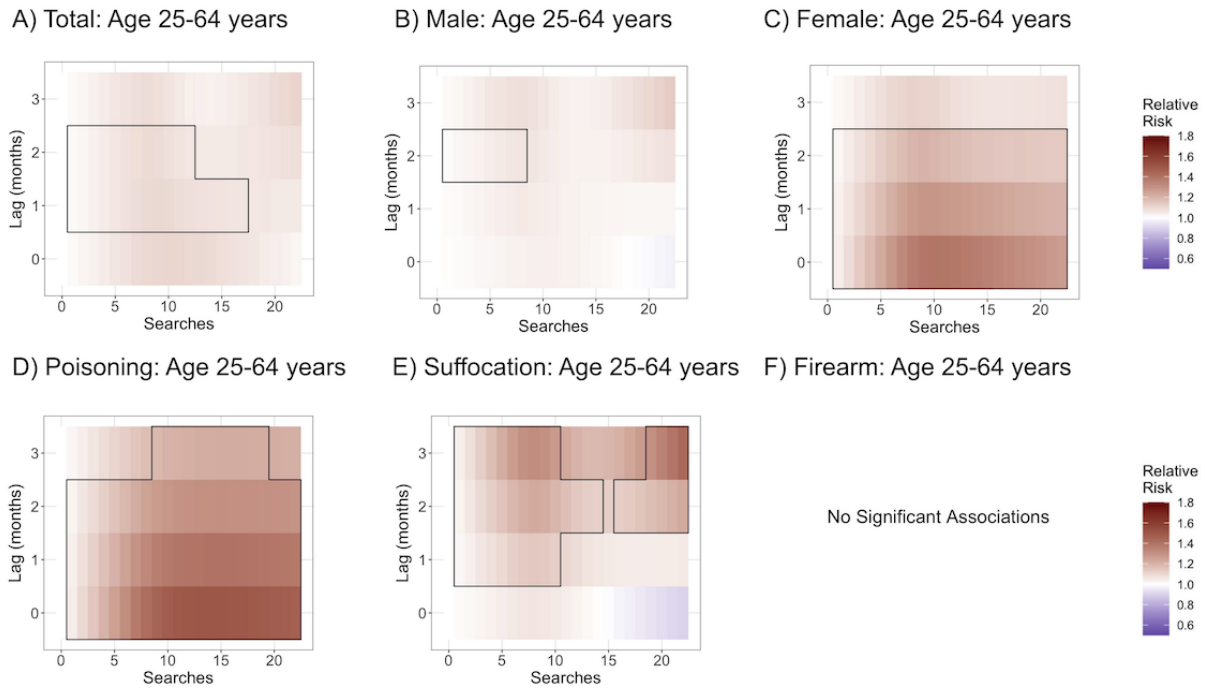

There were few significant associations with male deaths, with significant associations only at lag 2, with a maximum relative risk as compared to no searches of 1.06 (95% CI 1.00-1.11) at a search volume of 8 per 10 million; however, there was a significant association with female deaths at lags 0, 1, and 2 months with maximum relative risks as compared to no searches of 1.40 (95% CI 1.19-1.63) at a search volume of 10 per 10 million, 1.29 (95% CI 1.19-1.40) at a search volume of 10 per 10 million, and 1.20 (95% CI 1.10-1.31) at a search volume of 9 per 10 million, respectively. The strongest association was at lag 0.

Associations with intentional poisoning deaths were positive at all lags with maximum relative risks as compared to no searches of 1.50 (95% CI 1.29-1.75) at a search volume of 14 per 10 million, 1.40 (95% CI 1.30-1.51) at a search volume of 15 per 10 million, 1.31 (95% CI 1.21-1.43) at a search volume of 15 per 10 million, and 1.23 (95% CI 1.03-1.45) at a search volume of 16 per 10 million for lags 0, 1, 2, and 3, respectively. The strongest association was at lag 0.

Associations with deaths by suffocation were similarly positive at all lags; however, associations at lag 0 did not reach significance. Maximum relative risks as compared to no searches were 1.14 (95% CI 1.03-1.28), 1.23 (95% CI 1.09-1.38), and 1.44 (95% CI 1.05-1.97) for lags 1, 2, and 3 months, respectively, all at a search volume of 8 per 10 million.

Associations with firearm deaths were not significant at any lag.

Sensitivity Analyses

SARIMA model results were distinct between youth and adults. For youth, similar to the DLNM analysis, SARIMA analysis showed a consistent peak at 0 lag for total youth deaths by suicide, including among men and women, and for suffocation. However, unlike the DLNM models, there was a peak at 0 lag in firearm deaths. Deaths by poisoning were the only group without a statistically significant positive association with search rates at lag 0; however, estimates show a positive trend. All groups of youth deaths except females also had a significant negative association at either lag 2 or lag 3 (Figure S1 in

XSL•FO
RenderX

Multimedia Appendix 1). Unlike DLNM models, when investigating associations for adults aged 25-64 years, there was no evidence of an association for any group at any lag (Figure S2 in Multimedia Appendix 1).

The patterns of association in the 2 subperiods were mostly consistent with the results for the entire record (Figure S3 in Multimedia Appendix 1). The time period between 2016 and 2021 had stronger associations between searches and means-specific deaths for both youth and adults. However, the association with total deaths was stronger in the earlier time period for youths.

Opioid-related suicide deaths accounted for (21,441/85,399, 25.1%) of intentional poisoning deaths between 2010 and 2021, and included the use of synthetic opioids, heroin, methadone, codeine, and morphine. The pattern of the relationship with search rates for prosuicide websites was similar between poisoning suicides due to opioids and those that were not opioid-related; however, for youth, the magnitude of the association between search rates and deaths was only significant for opioid-related deaths as compared to those that were not opioid-related (Figure S4 in Multimedia Appendix 1).

Repeating the DLNM analyses using Google search terms that were not specific to prosuicide websites identified no significant associations between search rates and deaths by suicide for female youth and adults, deaths by poisoning for youth and adults, or deaths by suffocation for adults. In addition, the significant associations with deaths were largely negative, except for the association between searches and deaths by suffocation for youth aged 10-24 years (Figure S5 in Multimedia Appendix 1).

Discussion

During months when searches of prosuicide websites increased, there were more deaths by intentional poisoning and suffocation, among both youth and adults. However, for youth, we also observed that increased searches were associated with fewer suicide deaths by firearm at lags of 1, 2, and 3 months. These findings are suggestive of an association between searches and suicide deaths, and warrant further exploration. There are multiple potential explanations for this association, both causal and noncausal. The findings could be the evidence of a causal relationship with these sites influencing individuals to take their own life. On the other hand, increases in searches for such sites could be a marker of an increase in population suicidality, although results of sensitivity analyses for other suicide-related terms suggest otherwise.

We observed distinctly different patterns among adolescents and young adults in associations with means of death. Among these age groups, while there were positive concurrent associations between searches and deaths by suffocation and poisoning, there were negative time-lagged associations with suicide deaths by firearm. Messaging common on prosuicide websites may explain this pattern. While there has been little academic study of the content of suicide websites, our review of the posts consistently noted that deaths by firearm were often discouraged and firearm self-injury was discussed as having a high risk of long-lasting physical disability (as has been noted by others [37]). In contrast, poisoning and suffocation were discussed more often and were “encouraged.” Associations may also be higher with deaths by poisoning and suffocation due to more complex procedures required for death by these means; thus, increased access to specific information may have a greater impact on lethality.

Within specific means, the associations differed by age group. There was a positive association at all lags for all age groups with deaths by suffocation; however, for youth, associations were of greater magnitude. In contrast, associations between searches and death by poisoning were of greater magnitude for adults than youth. These patterns may be due to differing availability of means for youth, as suffocation is the most frequent nonfirearm means used by youth presumably due to greater ease of access to these means [38-40]. Poisoning, which tends to involve more complex procedures and difficult to obtain materials, is less common among youth [41].

Sensitivity analyses showed that there was some variation in the association of prosuicide site search rates with suicide by time period. Most associations were closer to the null between 2010 and 2015 as compared to between 2016 and 2021. This may be due to a change in the material shared on these sites over time, or different sites being created or shut down. Data were too sparse to conduct COVID-19–specific sensitivity analyses between 2020 and 2021; however, this is a question that requires further investigation.

Sensitivity analyses also showed that for youth, the associations were stronger for opioid-related intentional poisoning deaths compared to nonopioid-related intentional poisoning deaths; however, there was no corresponding difference for adults. While speculative, it should be noted that deaths from opioids in the United States have increased rapidly during this study period, including deaths designated as suicides among youth [36,42]. Individuals dependent on opioids have higher suicide rates than the general population, and thus suicidal individuals with access to opioids, especially highly lethal synthetic opioids, might have been more vulnerable to prosuicide messages. In addition, as adults are more likely to use multiple medications than youth [43], the effect of searches on adult opioid overdoses may be lower due to the availability of other medications.

Results also show that the patterns of association differed in models based on general suicide search terms from prosuicide website specific search terms. While more research is needed to fully investigate these findings, significant associations between searches of general suicide-related terms and deaths by suicide were largely negative, indicating a potential protective effect of seeking general suicide information using search engines, which could be due to the majority of web-based sources providing recovery and support resources [6], and showing that the current findings are indicative of a distinct effect of prosuicide sites. This also suggests that the association between searches and deaths by suicide is not solely a marker of an increase in population level suicidality due to differing patterns in associations.

For youth, the ARIMA models showed similar results for most groups as the DLNM models. They showed a positive

XSL•FO
RenderX

association between searches and deaths by suicide at 0 lag, with a negative association at a 2- or 3-month lag. They varied most from the DLNM models when estimating association with deaths by firearms; DLNM models showed only negative associations, while ARIMA models showed a pattern similar to other deaths by suicide. In addition, ARIMA models showed no association for adults, whereas DLNM models showed significant associations for most groups. These differences in results may be due to DLNM allowing for fitting of a nonlinear model to the associations, as well as differing methods for controlling for long-term variability.

Given the dangerous information they provide, efforts to limit accessibility to prosuicide websites and information have already taken place in several countries. The original domain name of one of the most well-known prosuicide websites was blocked or restricted in Australia, Italy, and Germany [44,45]. In the United States, following a high-profile news article of this website and associated deaths, there were legal actions proposed including criminal actions toward the owner of the website. Tech companies such as Google and Bing were also urged to limit accessibility [45]. Companies including Google and Meta have already taken some precautionary measures surrounding availability of such information, such as automatically providing information on suicide prevention resources based on search terms and allowing users to flag posts about self-harm or suicide [46,47]. However, these measures have limited efficacy [48]. Meanwhile, this website and other similar forums continue to be openly accessed with no effective age barriers and give explicit instructions on how to die by suicide [11,13,14]. It is well understood that while mentioning suicide in general and inquiring about suicidal ideation does not increase risk of suicide, discussion of specific suicide means and sensationalizing suicide deaths can increase the risk [3,4].

Our results suggest that there is a population level association between search rates of prosuicide websites and deaths by suicide. However, there are several limitations to this study. Google search volume does not indicate the population conducting these searches, limiting the interpretation of age or sex-specific associations. In addition, by using Google search volume to estimate access to these websites, we likely underestimated the true website seeking behavior, as there are other ways of accessing these sites that are not captured in Google searches, including via other search engines and direct links from other web pages. Further, Google search volume data were suppressed at the state level; thus, we were not able to account for state-level differences in potential covariates such as access to lethal means. While we were able to attend to some potential confounders such as increased lethality of the opioid supply, and included a seasonal term in the model to account for seasonal patterns, other potential confounders may be considered in future studies, including factors varying over time such as economic factors and media exposure. This study also used fatal suicides as a measure of suicidality; however, prosuicide websites may be associated with chosen means and subsequent fatality of an attempt rather than rates of attempts, which should be investigated in further research.

Our results emphasize potentially pernicious population-level effects of prosuicide websites on suicide risk among youth and adults. Although preliminary, the findings suggest that searches of these websites are associated with deaths by suicide among youth and adults and underscore the need to regulate or close these sites.

Acknowledgments

The study was supported by a National Institutes of Health grant (R01 MH 121410).

Data Availability

Data used for these analyses are not publicly available in order to limit dissemination of names of prosuicide websites, but are available from the corresponding author on reasonable request.

Authors' Contributions

NCK worked on conceptualization, data curation, formal analysis, investigation, methodology, visualization and writing of the original draft, review, and editing. CG worked on the conceptualization, methodology, review, and editing. MO worked on methodology as well as draft review and editing. MSG worked on conceptualization as well as draft review and editing. JS worked on methodology as well as draft review and editing. KK worked on conceptualization, funding, supervision, investigation, methodology, as well as draft review and editing.

Conflicts of Interest

JS and Columbia University disclose partial ownership of SK Analytics.

Multimedia Appendix 1
Additional figures.
[DOCX File , 3707 KB - jmir_v26i1e53404_app1.docx]

References

XSL•FO
RenderX

1. Garnett MF, Curtin SC. Suicide mortality in the United States, 2001-2021. NCHS Data Brief 2023(464):1-8 [FREE Full text] [Medline: 37093258]
2. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
3. Niederkrotenthaler T, Kirchner S, Till B, Sinyor M, Tran US, Pirkis J, et al. Systematic review and meta-analyses of suicidal outcomes following fictional portrayals of suicide and suicide attempt in entertainment media. EClinicalMedicine 2021;36:100922 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100922] [Medline: 34308310]
4. Niederkrotenthaler T, Braun M, Pirkis J, Till B, Stack S, Sinyor M, et al. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. BMJ 2020;368:m575 [FREE Full text] [doi: 10.1136/bmj.m575] [Medline: 32188637]
5. Kemp CG, Collings SC. Hyperlinked suicide: assessing the prominence and accessibility of suicide websites. Crisis 2011;32(3):143-151. [doi: 10.1027/0227-5910/a000068] [Medline: 21616763]
6. Thornton L, Handley T, Kay-Lambkin F, Baker A. Is a person thinking about suicide likely to find help on the internet? an evaluation of google search results. Suicide Life Threat Behav 2017;47(1):48-53. [doi: 10.1111/sltb.12261] [Medline: 27111845]
7. Westerlund M. The production of pro-suicide content on the internet: a counter-discourse activity. New Media & Society 2011;14(5):764-780. [doi: 10.1177/1461444811425221]
8. Durkee T, Hadlaczky G, Westerlund M, Carli V. Internet pathways in suicidality: a review of the evidence. Int J Environ Res Public Health 2011;8(10):3938-3952 [FREE Full text] [doi: 10.3390/ijerph8103938] [Medline: 22073021]
9. Mörch CM, Côté LP, Corthésy-Blondin L, Plourde-Léveillé L, Dargis L, Mishara BL. The Darknet and suicide. J Affect Disord 2018;241:127-132. [doi: 10.1016/j.jad.2018.08.028] [Medline: 30118946]
10. Biddle L, Derges J, Mars B, Heron J, Donovan JL, Potokar J, et al. Suicide and the internet: changes in the accessibility of suicide-related information between 2007 and 2014. J Affect Disord 2016;190:370-375. [doi: 10.1016/j.jad.2015.10.028] [Medline: 26546772]
11. Chen YY, Hung GC, Cheng Q, Tsai CW, Wu KCC. Searching for suicide-related information on Chinese websites. Psychiatry Res 2017;258:506-510. [doi: 10.1016/j.psychres.2017.08.087] [Medline: 28886904]
12. Huang CY, Tsai CW, Chi YC, Wu KC, Chen YY. Changes in accessibility of suicide-related information on websites in Taiwan during 2016 and 2019. J Formos Med Assoc 2022;121(1 Pt 2):335-341 [FREE Full text] [doi: 10.1016/j.jfma.2021.05.005] [Medline: 34024719]
13. Till B, Niederkrotenthaler T. Surfing for suicide methods and help: content analysis of websites retrieved with search engines in Austria and the United States. J Clin Psychiatry 2014;75(8):886-892. [doi: 10.4088/JCP.13m08861] [Medline: 25099284]
14. Eriksen CU, Konradsen F, Vildekilde T. Searching online for methods of suicide. Crisis 2021;42(2):136-143. [doi: 10.1027/0227-5910/a000701] [Medline: 32672520]
15. Voros V, Fekete S, Szabo Z, Torma E, Nagy A, Fekete J, et al. High prevalence of suicide-related internet use among patients with depressive disorders - a cross-sectional study with psychiatric in-patients. Psychiatry Res 2022;317:114815. [doi: 10.1016/j.psychres.2022.114815] [Medline: 36063749]
16. Harris KM, McLean JP, Sheffield J. Examining suicide-risk individuals who go online for suicide-related purposes. Arch Suicide Res 2009;13(3):264-276. [doi: 10.1080/13811110903044419] [Medline: 19591000]
17. Minkkinen J, Oksanen A, Kaakinen M, Keipi T, Räsänen P. Victimization and exposure to pro-self-harm and pro-suicide websites: a cross-national study. Suicide Life Threat Behav 2017;47(1):14-26. [doi: 10.1111/sltb.12258] [Medline: 27094352]
18. Marchant A, Hawton K, Stewart A, Montgomery P, Singaravelu V, Lloyd K, et al. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: the good, the bad and the unknown. PLoS One 2017;12(8):e0181722 [FREE Full text] [doi: 10.1371/journal.pone.0181722] [Medline: 28813437]
19. Mok K, Jorm AF, Pirkis J. The perceived impact of suicide-related internet use: a survey of young Australians who have gone online for suicide-related reasons. Digit Health 2016;2:2055207616629862 [FREE Full text] [doi: 10.1177/2055207616629862] [Medline: 29942550]
20. Niezen R. Internet suicide: communities of affirmation and the lethality of communication. Transcult Psychiatry 2013;50(2):303-322. [doi: 10.1177/1363461512473733] [Medline: 23315147]
21. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. Psychol Rev 2010;117(2):575-600 [FREE Full text] [doi: 10.1037/a0018697] [Medline: 20438238]
22. Marzano L, Katsampa D, Mackenzie JM, Kruger I, El-Gharbawi N, Ffolkes-St-Helene D, et al. Patterns and motivations for method choices in suicidal thoughts and behaviour: qualitative content analysis of a large online survey. BJPsych Open 2021;7(2):e60 [FREE Full text] [doi: 10.1192/bjo.2021.15] [Medline: 33622447]
23. Sueki H, Yonemoto N, Takeshima T, Inagaki M. The impact of suicidality-related internet use: a prospective large cohort study with young and middle-aged internet users. PLoS One 2014;9(4):e94841 [FREE Full text] [doi: 10.1371/journal.pone.0094841] [Medline: 24740115]

XSL•FO
RenderX

24. Tran US, Andel R, Niederkrotenthaler T, Till B, Ajdacic-Gross V, Voracek M. Low validity of google Trends for behavioral forecasting of national suicide rates. PLoS One 2017;12(8):e0183149 [FREE Full text] [doi: 10.1371/journal.pone.0183149] [Medline: 28813490]
25. Yang AC, Tsai SJ, Huang NE, Peng CK. Association of internet search trends with suicide death in Taipei city, Taiwan, 2004-2009. J Affect Disord 2011;132(1-2):179-184. [doi: 10.1016/j.jad.2011.01.019] [Medline: 21371755]
26. Neumann K, Mason SM, Farkas K, Santaularia NJ, Ahern J, Riddell CA. Harnessing google health trends data for epidemiologic research. Am J Epidemiol 2023;192(3):430-437 [FREE Full text] [doi: 10.1093/aje/kwac171] [Medline: 36193858]
27. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. Stat Med 2010;29(21):2224-2234 [FREE Full text] [doi: 10.1002/sim.3940] [Medline: 20812303]
28. McCleary R, Chew KS, Hellsten JJ, Flynn-Bransford M. Age- and sex-specific cycles in United States suicides, 1973 to 1985. Am J Public Health 1991;81(11):1494-1497. [doi: 10.2105/ajph.81.11.1494] [Medline: 1951813]
29. Yu J, Yang D, Kim Y, Hashizume M, Gasparrini A, Armstrong B, et al. Seasonality of suicide: a multi-country multi-community observational study. Epidemiol Psychiatr Sci 2020;29:e163 [FREE Full text] [doi: 10.1017/S2045796020000748] [Medline: 32829741]
30. Mergl R, Havers I, Althaus D, Rihmer Z, Schmidtke A, Lehfeld H, et al. Seasonality of suicide attempts: association with gender. Eur Arch Psychiatry Clin Neurosci 2010;260(5):393-400. [doi: 10.1007/s00406-009-0086-1] [Medline: 19915990]
31. Maes M, Cosyns P, Meltzer HY, De Meyer F, Peeters D. Seasonality in violent suicide but not in nonviolent suicide or homicide. Am J Psychiatry 1993;150(9):1380-1385. [doi: 10.1176/ajp.150.9.1380] [Medline: 8352350]
32. Lester D, Frank ML. Sex differences in the seasonal distribution of suicides. Br J Psychiatry 1988;153:115-117. [doi: 10.1192/bjp.153.1.115] [Medline: 3224232]
33. Gasparrini A, Scheipl F, Armstrong B, Kenward MG. A penalized framework for distributed lag non-linear models. Biometrics 2017;73(3):938-948 [FREE Full text] [doi: 10.1111/biom.12645] [Medline: 28134978]
34. Sueki H. Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: data from 2004 to 2009 in Japan. Psychiatry Clin Neurosci 2011;65(4):392-394 [FREE Full text] [doi: 10.1111/j.1440-1819.2011.02216.x] [Medline: 21569178]
35. Gasparrini A. Modeling exposure-lag-response associations with distributed lag non-linear models. Stat Med 2014;33(5):881-899 [FREE Full text] [doi: 10.1002/sim.5963] [Medline: 24027094]
36. Spencer MR, Miniño AM, Warner M. Drug overdose deaths in the United States, 2001-2021. NCHS Data Brief 2022(457):1-8 [FREE Full text] [Medline: 36598401]
37. Martin CW. How not to kill yourself : a portrait of the suicidal mind. New York: Pantheon Books; 2023:431.
38. Glenn CR, Kleiman EM, Kellerman J, Pollak O, Cha CB, Esposito EC, et al. Annual research review: a meta-analytic review of worldwide suicide rates in adolescents. J Child Psychol Psychiatry 2020;61(3):294-308. [doi: 10.1111/jcpp.13106] [Medline: 31373003]
39. Sarchiapone M, Mandelli L, Iosue M, Andrisano C, Roy A. Controlling access to suicide means. Int J Environ Res Public Health 2011;8(12):4550-4562 [FREE Full text] [doi: 10.3390/ijerph8124550] [Medline: 22408588]
40. Kosky RJ, Dundas P. Death by hanging: implications for prevention of an important method of youth suicide. Aust N Z J Psychiatry 2000;34(5):836-841. [doi: 10.1080/j.1440-1614.2000.00807.x] [Medline: 11037371]
41. Lynn E, Cousins G, Lyons S, Bennett KE. Comparing characteristics of suicide to non-suicide drug poisoning deaths, by sex, in Ireland. J Affect Disord 2022;306:80-89 [FREE Full text] [doi: 10.1016/j.jad.2022.03.030] [Medline: 35306124]
42. Friedman J, Godvin M, Shover CL, Gone JP, Hansen H, Schriger DL. Trends in drug overdose deaths among US adolescents, January 2010 to June 2021. JAMA 2022;327(14):1398-1400 [FREE Full text] [doi: 10.1001/jama.2022.2847] [Medline: 35412573]
43. Miller TR, Swedler DI, Lawrence BA, Ali B, Rockett IRH, Carlson NN, et al. Incidence and lethality of suicidal overdoses by drug class. JAMA Netw Open 2020;3(3):e200607 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0607] [Medline: 32202643]
44. Scarr L. Cops ban pro-suicide site after WA youth's death. The West Austrailian. URL: https://thewest.com.au/politics/federal-politics/cops-ban-pro-suicide-site-after-wa-youths-death-ng-b881210157z [accessed 2019-05-24]
45. Trahan C. Trahan leads bipartisan call for action to stop "sanctioned suicide" website. URL: https://trahan.house.gov/news/documentsingle.aspx?DocumentID=2336 [accessed 2021-12-21]
46. Cohen N. 'Suicide' query prompts Google to offer hotline. URL: https://www.nytimes.com/2010/04/05/technology/05google.html?ref=todayspaper [accessed 2010-04-04]
47. Kottasova I. Instagram launches suicide prevention tool. URL: https://money.cnn.com/2016/10/20/technology/instagram-suicide-prevention-tools/ [accessed 2016-10-20]
48. Cheng Q, Yom-Tov E. Do search engine helpline notices aid in preventing suicide? analysis of archival data. J Med Internet Res 2019;21(3):e12235 [FREE Full text] [doi: 10.2196/12235] [Medline: 30912753]

XSL•FO
RenderX

Abbreviations

API: application programming interface
ARIMA: autoregressive integrated moving average
DLNM: distributed lag nonlinear model
ICD-10: International and Statistical Classification of Diseases, Tenth Revision
NVSS: National Vital Statistics System
(S)ARIMA: (seasoned) autoregressive integrated moving average

Edited by T de Azevedo Cardoso; submitted 05.10.23; peer-reviewed by B Lawrence, M Sinyor; comments to author 12.12.23; revised version received 22.12.23; accepted 11.04.24; published 26.07.24.

Please cite as:
Kelsall NC, Gimbrone C, Olfson M, Gould MS, Shaman J, Keyes K
Association Between Prosuicide Website Searches Through Google and Suicide Death in the United States From 2010 to 2021: Lagged Time-Series Analysis
J Med Internet Res 2024;26:e53404
URL: https://www.jmir.org/2024/1/e53404
doi:10.2196/53404
PMID:

©Nora Clancy Kelsall, Catherine Gimbrone, Mark Olfson, Madelyn S Gould, Jeffrey Shaman, Katherine Keyes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Relationship Between Instagram Use and Self-Criticism, Self-Compassion, and Body Dissatisfaction in the Spanish Population: Observational Study

Andrea Varaona[1], MD; Miguel Angel Alvarez-Mon[1,2,3,4], MD, PhD; Irene Serrano-Garcia[5], BM; Marina Díaz-Marsá[5], MD, PhD; Jeffrey C L Looi[6,7], MBBS, MD, DMedSc; Rosa M Molina-Ruiz[5,8], MD, PhD

[1]Department of Medicine and Medical Specialities, Faculty of Medicine and Health Sciences, University of Alcala, Alcalá de Henares, Spain
[2]CIBERSAM-ISCIII (Biomedical Research Networking Centre in Mental Health), Madrid, Spain
[3]Ramón y Cajal Institute of Sanitary Research, Madrid, Spain
[4]Department of Psychiatry and Mental Health, Hospital Universitario Infanta Leonor, Madrid, Spain
[5]Department of Psychiatry and Mental Health, Hospital Clínico San Carlos, Madrid, Spain
[6]Academic Unit of Psychiatry and Addiction Medicine, The Australian National University Medical School, Canberra, Australia
[7]Consortium of Australian Academic Psychiatrists for Policy, Research and Analysis, Canberra, Australia
[8]Fundación para la Investigación Biomédica del Hospital Clínico San Carlos, Madrid, Spain

Corresponding Author:
Miguel Angel Alvarez-Mon, MD, PhD
Department of Medicine and Medical Specialities
Faculty of Medicine and Health Sciences
University of Alcala
C/19, Carretera Madrid-Barcelona, Km 33,600
Alcalá de Henares, 28805
Spain
Phone: 34 918854503
Email: maalvarezdemon@icloud.com

Abstract

Background: The widespread use of online social networks, particularly among the younger demographic, has catalyzed a growing interest in exploring their influence on users' psychological well-being. Instagram (Meta), a visually oriented platform, has garnered significant attention. Prior research has consistently indicated that Instagram usage correlates with heightened levels of perfectionism, body dissatisfaction, and diminished self-esteem. Perfectionism is closely linked to self-criticism, which entails an intense self-scrutiny and is often associated with various psychopathologies. Conversely, self-compassion has been linked to reduced levels of perfectionism and stress, while fostering greater positive affect and overall life satisfaction.

Objective: This study investigates the relationship between Instagram usage (time of use and content exposure) and users' levels of self-compassion, self-criticism, and body dissatisfaction.

Methods: This study comprised 1051 adult participants aged between 18 and 50 years, either native to Spain or residing in the country for at least a decade. Each participant completed a tailored questionnaire on Instagram usage, along with abbreviated versions of the Self-Compassion Scale, the Body Shape Questionnaire, and the Depressive Experiences Questionnaire, spanning from January 23 to February 25, 2022.

Results: A positive correlation was observed between daily Instagram usage and self-criticism scores. Participants of all age groups who spent over 3 hours per day on Instagram exhibited higher self-criticism scores than users who spent less than 1 hour or between 1 and 3 hours per day. Contrary to previous findings, no significant relationship was detected between Instagram usage time and levels of self-compassion or body dissatisfaction. Furthermore, content centered around physical appearance exhibited a positive correlation with self-criticism and body dissatisfaction scores. Among younger participants (aged 18-35 years), those who primarily viewed beauty or fashion content reported higher self-criticism scores than those consuming science-related content. However, this association was not significant for participants aged 35-50 years. Conversely, individuals who predominantly engaged with sports or fitness or family or friends content exhibited higher levels of body dissatisfaction than

XSL•FO
RenderX

those focusing on science-related content. No significant associations were observed between self-compassion scores and daily Instagram usage or most-viewed content categories.

Conclusions: The findings of this study underscore the considerable impact of Instagram usage on self-criticism and body dissatisfaction—2 variables known to influence users' psychological well-being and be associated with various symptoms and psychological disorders.

(J Med Internet Res 2024;26:e51957) doi:10.2196/51957

KEYWORDS

Instagram; self-compassion; self-esteem; self-criticism; self-worth; body dissatisfaction; dissatisfaction; satisfaction; appearance; psychological; social media; body; mental health; mental wellbeing; Spain; Spanish; Hispanic; depression; depressive; usage; correlation; association

Introduction

The advent of social media has revolutionized how we interact and connect with others. In 2021, there were approximately 3.78 billion active users worldwide [1]. In 2022, it is estimated that Facebook (Meta) had 2.91 billion active users, followed by YouTube (Google LLC) with 2.56 billion, WhatsApp (Meta) and Instagram with 2 billion active users, and TikTok (ByteDance) with 1 billion users [2]. Social media serves as a way to stay connected with others, whether for work, educational, or social purposes [3]. The use of social networks by young adults is considered a normative experience and a new digital environment that impacts their psychological and social development [4].

According to the literature, the most visually oriented social media platforms have a greater psychological impact on users. Specifically, previous studies have examined the impact of using Instagram on self-esteem, body image satisfaction, social comparison, perfectionism, and other variables relevant to psychological well-being [5-12]. College women who used Facebook and Instagram reported a greater number of appearance-related thoughts than those who used other less visual platforms [5]. The results from a different study showed that the use of Instagram was closely related to dysmorphic concerns in men and women [7]. Moreover, in other studies, it has been observed that individuals who use Instagram for more hours per day tend to exhibit lower levels of self-esteem and psychological well-being, as well as greater dissatisfaction with their body image [10,12].

However, the association between Instagram usage and self-compassion and self-criticism has not been thoroughly studied yet. In a recent study, it was reported that users with higher levels of self-compassion used Instagram less frequently [13]. Self-criticism, besides hurting people's moods, is also considered a vulnerability factor for the development of depression and other mental disorders, such as social anxiety, eating disorders, posttraumatic stress disorder, personality disorders, and psychotic symptoms [14-18]. The implication of self-criticism in a wide range of psychopathologies has led us to consider it as a transdiagnostic process [16,17]. Furthermore, self-criticism has been associated with self-injuries and suicidal thoughts [19-22]. On the other hand, self-compassion has been found to have multiple psychological benefits [13,23,24]. Therefore, the objective of this study was to explore the potential associations between daily Instagram usage time and the most viewed content with levels of self-compassion, self-criticism, and body dissatisfaction among users.

Methods

Study Design and Participants

This study uses a cross-sectional observational design, wherein participants were recruited anonymously and voluntarily. The survey was administered using Google Forms and commenced with an introductory section outlining this study's objectives, along with obtaining informed consent regarding anonymity and voluntary participation. Participants were then prompted to provide responses regarding sociodemographic information, Instagram usage patterns, and other pertinent measures relevant to this study's aims.

The recruitment process involved promoting the survey through one of the authors' Instagram profiles dedicated to mental health, which boasts over 100,000 followers. Additionally, some followers shared the survey on their own Instagram stories, extending this study's reach to a broader audience, whereas other authors shared the survey via email and WhatsApp. Data collection took place over the course of 1 month, from January 23 to February 25, 2022.

The inclusion criteria were (1) being aged between 18 and 40 years, (2) regular usage of Instagram, and (3) Spanish nationality or living in Spain for at least 10 years. Participants with a current or past diagnosis of eating disorders were excluded from this study.

Assessment Instruments and Variables Considered

We used the following measurement instruments: (1) our own researcher-designed questionnaire collecting sociodemographic data and Instagram use variables, (2) the Self-Compassion Scale (SCS); (3) the Depressive Experiences Questionnaire (DEQ); (4) the Body Shape Questionnaire (BSQ). In the Instagram questionnaire (available in Multimedia Appendix 1), aspects such as daily usage time on Instagram, the number of followers, the number of accounts they followed, the most viewed content (science, gastronomy, traveling, family or friends, beauty or fashion, sports, fitness, or lifestyle, humor, news, or others), and the number of stories and posts each participant published daily were also assessed.

Regarding the SCS [25], we used the short version translated into Spanish and validated by Garcia-Campayo et al [26]. This

XSL•FO
RenderX

self-applied scale consists of 12 items that assess self-compassion in general (total score) and 3 specific components of self-compassion: common humanity, mindfulness, and self-compassion. Although the construct of self-compassion is defined from these 3 components, the scale is divided into 6 subscales representing the positive and negative aspects of each of these 3 components: self-compassion, self-judgment, common humanity, isolation, mindfulness, and overidentification. Participants are asked to respond to the items according to how they usually act toward themselves in difficult times using a Likert-type scale ranging from 1 ("almost never") to 5 ("almost always"). In this short Spanish version, Cronbach α is 0.85, indicating good internal consistency. As for the correlation between the items of the 26-item and 12-item versions, high correlations were observed between the different subscales: self-compassion (r=0.89), self-judgment (r=0.90), common humanity (r=0.81), isolation (r=0.81), mindfulness (r=0.83), and overidentification (r=0.90). Regarding the correlation between the total score of both versions, a high correlation was also observed (r≥0.92). Higher scores on this scale are considered to indicate higher levels of self-compassion [27].

Furthermore, we used the DEQ [28]. The short version was translated into Spanish and validated [29]. This questionnaire is composed of 32 items that measure 3 factors: dependence, self-criticism, and relationship. In this version, the relationship factor is grouped within the dependence factor. Although both refer to the phenomenon of separation, the dependency factor concerns a more primitive level of development and refers to feelings of helplessness, hopelessness, and fear of separation, while the relationship factor refers to feelings of loss and loneliness in response to the breakup of a particular relationship, this last factor being a more mature way of coping after the loss of a significant other. The self-criticism factor refers to a self-oriented variable and includes items regarding feelings of guilt, emptiness, hopelessness, dissatisfaction with oneself, failure to meet one's own expectations, and the tendency to be critical about oneself. The questions are to be answered following a Likert-type scale ranging from 1 to 7, where 1 represents "strongly disagree" and 7 indicates "strongly agree." Scores are calculated independently for each subscale since dependence and self-criticism are differentially associated with different types of depressive symptoms and interpersonal behaviors [30]. The minimum and maximum score on each subscale is 1 and 7, respectively. Although there is no cutoff point in the interpretation of this scale, higher scores indicate higher levels of self-criticism or dependence. In this version of the questionnaire, Cronbach α indicated good internal consistency for the 3 factors of dependence (α=0.82), self-criticism (α=0.85), and relationship (α=0.71).

Concerning the BSQ [31] we used the short version (BSQ-14) validated by Dowson and Henderson [32], which consists of 14 items measuring body dissatisfaction in the last 2 weeks using a 6-point Likert scale, where 1 is equivalent to "never" and 6 is to "always." The total score is obtained by adding the score of each item so that it ranges from 14 to 84. Although there is no established cutoff point for body image dissatisfaction, higher scores indicate greater dissatisfaction or concern about body shape [33]. The Cronbach α of this version is 0.93, showing good internal reliability. In a study in which 8 different versions of this scale were compared, this was 1 of the 3 versions that showed the most favorable results in comparison to the original 34-item scale [34]. In a study with Mexican university women, an adequate fit (comparative fit index=0.95; normed fit index=0.92; goodness-of-fit index=0.86) and high internal consistency (α=0.96) were obtained [35]. In addition, it correlated positively with the Eating Attitudes Test-26, which measures the presence of eating disorder symptoms [36].

Ethical Considerations

This study received approval (MPGS_2021_012) from the University of Comillas Research Ethics Committee and was compliant with the research ethics principles of the Declaration of Helsinki. An informed consent was obtained from every participant after acknowledging the nature and objectives of this study. This study's data are anonymous and voluntary, as no compensation was offered to the participants.

Statistical Analysis

Data analysis was conducted using SPSS Statistics (version 26; IBM Corp). For descriptive analysis of the data, qualitative variables are presented with a distribution of frequencies, and quantitative variables are presented using mean and SD. The continuous nonnormally distributed variables were summarized by the median (IQR 25th percentile to 75th percentile). To compare the groups, the ANOVA test for scores obtained on the scales and quantitative normally distributed variables and the nonparametric Kruskal-Wallis test for nonnormally distributed variables were used. Bonferroni test for multiple comparisons was applied to determine which means differed from each other. In the case of qualitative variables, comparison was evaluated using the chi-square test, or by the Fisher exact test in case more than 25% of the expected values were less than five. These analyses were performed to compare the 3 different Instagram usage times groups (less than 1 hour, between 1 and 3 hours, and more than 3 hours) and to compare scales of the most viewed content on Instagram. Subsequently, linear regression models were performed to explore the association between each of the dependent variables (SCS, self-criticism and dependency subscale of the DEQ, and the BSQ) and the time spent using Instagram daily and the type of content most viewed on this platform. The variables associated with the 3 different Instagram usage times groups (those variables which, in the univariate analyses, showed a level of statistical significance of P<.05) were considered to adjust the linear regression models. In the adjusted models, it was found that the effect of the Instagram usage times groups on the scales did not change by more than 15%. Therefore, the main model is considered without adjustment. Further, linear regression models were carried out with the total sample and by age groups (aged 18-35 and 35-50 years). For all the aforementioned tests, a significance value of 5% was accepted. Missing data due to nonresponse were not considered in the analyses. Therefore, for each variable, the analysis was of complete cases.

XSL•FO
RenderX

Results

Sociodemographic and Instagram Usage Characteristics of Participants

The final sample consisted of 1051 participants, from which 90.2% (948/1051) were women, 9% (95/1051) were men, 0.005% (5/1051) did not specify their gender, and 0.002% (2/1051) would rather not tell; mean age 30.58, SD 8.9 years). The sociodemographic data of our sample is presented in Table 1.

Descriptive data on participants' Instagram usage is presented in Table 2. The vast majority of the sample (648/1051, 61.7%) use Instagram between 1 and 3 hours per day, compared to 25.2% (265/1051) and 13.1% (138/1051) of the participants who use it less than 1 hour per day and for more than 3 hours per day, respectively. A decrease in daily Instagram usage time was observed as age increased. Thus, 16% (63/394) of the group aged between 18 and 25 years use Instagram for more than 3 hours a day, while 15.1% (54/357) and 7% (21/300) of the participants in the groups of those aged 25-35 and 35-50 years, respectively, use Instagram for more than 3 hours a day.

Table 1. Sociodemographic variables[a].

Variables	Total (N=1051)	Women (n=948)	Men (n=95)
Age (years), mean (SD)	30.58 (8.9)	30.77 (8.9)	24.23 (9.1)
Age groups (years), n (%)			
18-25	394 (37.5)	344 (36.3)	44 (46.3)
25-35	357 (34)	326 (34.4)	30 (31.6)
35-50	300 (28.5)	278 (29.3)	21 (22.1)
Educational level, n (%)			
Elementary education	1 (0.1)	1 (0.01)	0 (0)
Mandatory high school education	40 (3.8)	35 (3.7)	4 (4.2)
High school	162 (15.4)	141 (14.9)	18 (18.9)
Professional study	187 (17.8)	163 (17.2)	22 (23.2)
College degree	357 (34)	331 (34.9)	25 (26.3)
Master's degree	284 (27)	260 (27.4)	23 (24.2)
Doctorate degree	20 (1.9)	17 (1.8)	3 (3.2)
Work, n (%)			
Yes	688 (66.5)	627 (66.1)	58 (61.1)
No	341 (32.4)	301 (31.8)	35 (36.8)
Other	22 (2.1)	20 (2.1)	2 (2.1)
Marital status, n (%)			
Single	371 (35.3)	310 (32.7)	55 (57.9)
In a relationship	435 (41.4)	410 (43.2)	23 (24.2)
Married	204 (19.4)	190 (20)	14 (14.7)
Separated or divorced	35 (3.3)	32 (3.4)	3 (3.2)
Other	6 (0.6)	6 (0.6)	0 (0)
Children (yes), n (%)	221 (21)	206 (21.7)	15 (15.8)
Physical activity (yes)	616 (58.6)	550 (58)	64 (67.4)
Substance use (yes)	546 (52)	480 (50.6)	64 (67.4)

[a]Sociodemographic characteristics of the total sample of participants and by sex. The variable "children (yes)" refers to participants having one or more children, "physical activity (yes)" refers to participants who practice some type of physical activity, and "substance use (yes)" refers to participants who consume some type of substance (alcohol, tobacco, or other).

XSL•FO
RenderX

Table 2. Descriptive data on Instagram usage of the total sample and by age group.

Variables	Total (N=1051)	Aged 18-25 years (n=394)	Aged 25-35 years (n=357)	Aged 35-50 years (n=300)
Daily usage time[a] (hours), n (%)				
<1	265 (25.2)	90 (23)	88 (24.6)	87 (29)
1-3	648 (61.7)	241 (61.2)	215 (60.2)	192 (64)
>3	138 (13.1)	63 (16)	54 (15.1)	21 (7)
Followers[b], n (%)				
<100 followers	275 (26.2)	36 (9.1)	86 (24.1)	153 (51)
101-500 followers	530 (50.4)	205 (52)	207 (58)	118 (39.3)
>500 followers	233 (22.2)	149 (37.8)	61 (17.1)	23 (7.7)
Users followed[b], n (%)				
<100	102 (9.7)	8 (2)	29 (8.1)	65 (21.7)
101-500	577 (54.9)	203 (51.5)	205 (57.4)	169 (56.3)
>500	359 (34.2)	181 (45.9)	119 (33.3)	59 (19.7)
Most viewed content[c], n (%)				
Sports, fitness, or lifestyle	96 (12.7)	33 (8.4)	35 (9.8)	28 (9.3)
Beauty or fashion	134 (12.7)	45 (11.4)	51 (14.3)	38 (12.7)
Family or friends	259 (24.6)	133 (33.8)	88 (24.6)	38 (12.7)
Science	245 (23.3)	43 (10.9)	74 (20.7)	128 (42.7)
Gastronomy	54 (5.1)	22 (5.6)	15 (4.2)	17 (5.7)
Humor	146 (13.9)	75 (19)	51 (14.3)	20 (6.7)
Travel	33 (3.1)	15 (3.8)	12 (3.4)	6 (2)
News	17 (1.6)	3 (0.8)	6 (1.7)	8 (2.7)
Other	67 (6.4)	25 (6.3)	25 (7)	17 (5.7)
Time using Instagram[d] (years), n (%)				
<4	197 (18.7)	35 (8.9)	58 (16.2)	104 (34.7)
5-9	352 (33.5)	193 (49)	104 (29.1)	55 (18.3)
>9	159 (15.1)	71 (18)	67 (18.8)	21 (7)

[a]Daily usage time: expressed on several minutes or hours on app usage time settings.

[b]Followers and users you follow: expressed by directly observing Instagram account statistics.

[c]Most viewed content: participants expressed their 3 most observed content, in order from highest to lowest consumption.

[d]Time using Instagram: expressed by observing the year they created their Instagram account.

Correlation Between the Most Observed Type of Content and Psychological Variables

The total sample obtained an average of 2.75 (SD 0.41) on the SCS, an average of 3.61 (SD 1.76) on the self-criticism subscale of the DEQ (DEQ-A), 3.73 (SD 1.58) on the dependence subscale of the DEQ (DEQ-D) and, finally, an average of 47.69 (SD 20.10) on the BSQ-14. There are significant statistical differences between the 3 age groups in the scores obtained in both subscales of the DEQ (P<.001), in BSQ-14 (P<.001), and in the SCS scale (P=.005).

The Bonferroni test indicated that in the 3 scales used in our investigation, there were significant differences between the scores of the youngest group (aged 18-25 years) and the oldest group (35-50 years), but not between the scores of the age groups of 18-25 years and 25-35 years. Due to this, the younger and middle-aged groups were combined to create a single group compromising participants aged 18-35 years. The results displayed in Table 3 refer to 2 age groups, one including participants aged 18-35 years and the other including those aged 35-50 years. The t test (2-tailed) for independent samples showed that there are significant statistical differences in the scores obtained in the 3 scales between the aged 18-35 years group and the aged 35-50 years group.

XSL•FO
RenderX

Table 3. Descriptive data for the total sample and scales mean comparisons between different age groups.

Dependent variables	Total	Aged 18-35 years	Aged 35-50 years	*P* value[a]
SCS[b], mean (SD)	2.75 (0.41)	2.73 (0.41)	2.81 (0.41)	.002
DEQ-A[c], mean (SD)	3.61 (1.76)	3.80 (1.74)	3.13 (1.7)	<.001
DEQ-D[d], mean (SD)	3.73 (1.58)	3.82 (1.63)	3.49 (1.4)	.002
BSQ[e], mean (SD)	47.69 (20.1)	49.23 (20.43)	43.83 (18.73)	<.001

[a]*P*<.05.

[b]SCS: Self-Compassion Scale.

[c]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[d]DEQ-D: dependency subscale of the Depressive Experiences Questionnaire.

[e]BSQ: Body Shape Questionnaire.

Relation Between Time Using Instagram and Self-Criticism

Mean scores and SDs of the scales as a result of time spent using Instagram, as well as the *P* value after performing ANOVA analysis are presented in Table 4. The results revealed significant statistical differences across the 3 different usage times groups in the mean scores obtained in the DEQ-A and DEQ-D (*P*<.001). However, no significant statistical differences were found in the mean scores obtained in the SCS (*P*=.12) or the BSQ-14 (*P*=.43) among the 3 different times of use.

Table 4. Scores of the scales and mean comparison according to daily Instagram time of use.

Dependent variables	Less than 1 hour	Between 1 and 3 hours	More than 3 hours	*P* value
SCS[a], mean (SD)	2.79 (0.39)	2.76 (0.41)	2.69 (0.42)	.12
DEQ-A[b], mean (SD)	3.30 (1.69)	3.63 (1.72)	4.12 (1.94)	<.001
DEQ-D[c], mean (SD)	3.53 (1.55)	3.71 (1.55)	4.21 (1.64)	<.001
BSQ-14[d], mean (SD)	46.38 (19.75)	48 (20.05)	48.75 (21.03)	.43

[a]SCS: Self-Compassion Scale.

[b]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[c]DEQ-D: dependency subscale of the Depressive Experiences Questionnaire.

[d]BSQ-14: Body Shape Questionnaire.

Type of Content Viewed Is Related to Body Dissatisfaction

Descriptive data of the scales according to the type of content most viewed on Instagram and the results of the 1-factor ANOVA analysis are presented in Table 5. In such analysis, the 5 types of content most viewed by the participants (family or friends, science, humor, beauty or fashion, and sport, fitness, or lifestyle) were analyzed. Statistically significant differences were found only in body dissatisfaction (*P*=.02) according to the type of most viewed content. However, no statistically significant differences were observed in self-compassion, self-criticism, and dependence when taking the most observed content type into account (*P*=.36, *P*=.16, and *P*=.07, respectively). Bonferroni test results indicated that there are statistically significant differences in the mean scores of body dissatisfaction between those groups whose most observed content was “science” and “humor” (*P*=.007).

Results of the linear regression model conducted to explore the relationships between daily Instagram use and self-compassion, self-criticism, and body dissatisfaction are presented in Table 6. To begin with, no statistically significant relationship was found between using Instagram between 1 and 3 hours per day or for more than 3 hours per day and SCS. As for self-criticism (DEQ-A), a statistically significant association was observed between this variable and using Instagram between 1 and 3 hours a day (*P*=.01) and for more than 3 hours a day (*P*<.001). Participants who use Instagram between 1 and 3 hours a day scored on average 0.325 (SD 0.13) points higher on this subscale than those who use Instagram for less than 1 hour a day. On the other hand, those who use Instagram for more than 3 hours a day scored on average 0.817 (SD 0.18) points higher on this subscale than those who use it for less than 1 hour a day. As for dependence (DEQ-D), a statistically significant relationship was observed between this variable and using Instagram for more than 3 hours a day (*P*<.001), but not between using Instagram between 1 and 3 hours a day (*P*=.11). Compared to those who use Instagram less than one hour a day, those who use this social network for more than 3 hours a day scored on average 0.680 (SD 0.16) points higher on this subscale. Finally, no statistically significant relationship was found between consuming Instagram between 1 and 3 hours a day or for more than 3 hours a day on body dissatisfaction.

XSL•FO
RenderX

Table 5. Descriptive data and mean comparison on the self-compassion, self-criticism, and body dissatisfaction scales and the types of content most frequently observed by the participants.

DV[a]	Sport, fitness, or lifestyle	Beauty or fashion	Family or friends	Science	Humor	*P* value
SCS[b]	2.76 (0.39)	2.73 (0.37)	2.75 (0.42)	2.80 (0.42)	2.70 (0.39)	.36
DEQ-A[c]	3.47 (1.7)	3.85 (1.77)	3.56 (1.69)	3.48 (1.72)	3.76 (1.76)	.16
DEQ-D[d]	3.49 (1.66)	3.92 (1.58)	3.81 (1.54)	3.57 (1.49)	3.74 (1.6)	.07
BSQ-14[e]	50.06 (21.19)	47.38 (18.7)	48.12 (20.2)	43.96 (19.49)	51.82 (21.37)	.02

[a]DV: dependent variables.

[b]SCS: Self-Compassion Scale.

[c]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[d]DEQ-D: dependency subscale of the Depressive Experiences Questionnaire.

[e]BSQ-14: Body Shape Questionnaire.

Table 6. Unadjusted linear regression model and *P* value for the association between psychological variables and daily Instagram use in the overall sample (N=1051)[a].

	SCS[b]		DEQ-A[c]		DEQ-D[d]		BSQ-14[e]	
	β[f]	*P* value	β	*P* value	β	*P* value	β	*P* value
Between 1 and 3 hours	–0.11	.71	0.325	.01	0.184	.11	1.624	.27
More than 3 hours	–0.83	.05	0.817	<.001	0.680	<.001	2.376	.26

[a]Reference category in a linear regression model to which other times of usage were compared to less than 1 hour.

[b]SCS: Self-Compassion Scale.

[c]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[d]DEQ-D: dependence subscale of the Depressive Experiences Questionnaire.

[e]BSQ-14: Body Shape Questionnaire.

[f]β: regression coefficient.

Age Differences Between Daily Instagram Use and Psychological Variables

Results of the linear regression model conducted to explore the relationships between daily Instagram usage time and self-compassion, self-criticism, dependence, and body dissatisfaction by age group are shown in Table 7. No statistically significant relationships were found between using Instagram between 1 and 3 hours per day and for more than 3 hours per day and self-compassion and body dissatisfaction scores were not found in either of the 2 age groups. However, in both groups, a statistically significant relationship was observed between using Instagram for more than 3 hours a day and self-criticism. In the aged 18-35 years group, those using Instagram for more than 3 hours a day scored on average 0.615 (SD 0.21) points higher on the self-criticism subscale than those using Instagram for less than 1 hour daily (*P*=.003). In the aged 35- to 50-year group, people who use Instagram for more than 3 hours a day scored on average 1.111 (SD 0.41) points higher on this subscale than people who use Instagram for less than 1 hour per day (*P*=.007). Regarding the dependence variable, a statistically significant relationship was observed between consuming Instagram for more than 3 hours a day in both age groups, but not between using Instagram 1-3 hours a day and this variable. In this sense, individuals aged between 18 and 35 years who consume Instagram for more than 3 hours a day scored on average 0.571 (SD 0.19) points higher in dependency than those who use this social network for less than 1 hour a day (*P*=.003). Regarding the 35- to 50-year age group, individuals who use Instagram for more than 3 hours a day scored on average 0.911 (SD 0.34) points higher in dependence than those who use Instagram for less than 1 hour (*P*=.007).

XSL•FO
RenderX

Table 7. Unadjusted linear regression estimate and *P* value for the association between psychological variables and daily Instagram use stratified by age[a].

	SCS[b]		DEQ-A[c]		DEQ-D[d]		BSQ-14[e]	
	β[f]	*P* value	β	*P* value	β	*P* value	β	*P* value
Aged 18-35 years (n=751)								
Between 1 and 3 hours	–0.019	.59	0.277	.07	0.165	.25	0.969	.59
More than 3 hours	–0.075	.12	0.615	.003	0.571	.003	0.035	.99
Aged 35-50 years (n=300)								
Between 1 and 3 hours	0.015	.78	0.362	.097	0.193	.28	2.435	.32
More than 3 hours	–0.062	.54	1.111	.007	0.911	.007	7.443	.10

[a]Reference category in a linear regression model to which other times of usage were compared to less than 1 hour.

[b]SCS: Self-Compassion Scale.

[c]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[d]DEQ-D: dependence subscale of the Depressive Experiences Questionnaire.

[e]BSQ-14: Body Shape Questionnaire.

[f]β: regression coefficient.

Psychological Variables Affected by the Most Observed Type of Content

Table 8 provides the results of the linear regression model that analyzed the relationship between the type of most viewed content on Instagram and self-compassion, self-criticism, dependence, and body dissatisfaction. Regarding the self-compassion variable, the linear regression equation was statistically significant only for the "humor" content type (*P*=.02). Participants who mainly watch "humor" content scored on average 0.100 (SD 0.04) points lower in self-compassion than those who watch mostly science. Regarding self-criticism, no statistically significant relationship was found between this variable and the type of most watched content on Instagram. Regarding dependence, only a statistically significant relationship was found between this variable and watching mostly "beauty or fashion" content (*P*=.04). Compared to individuals whose most viewed content on Instagram is "science," individuals who view mainly "beauty or fashion" content scored on average 0.355 (SD 0.17) points higher in dependence. Finally, a statistically significant relationship was observed between body dissatisfaction and observing mainly "sport, fitness, or lifestyle" content (*P*=.01), "family or friends" (*P*=.02), and "humor" (*P*<.001). Individuals who mainly view "sports, fitness, or lifestyle" content scored on average 6.099 (SD 2.41) points higher in body dissatisfaction than those whose most viewed content is "science." On the other hand, individuals whose most viewed content is "family or friends" and "humor" scored on average 4.156 (SD 1.78) and 7.859 (SD 2.10) higher in body dissatisfaction, respectively, than those who mainly watch "science" content. However, no statistically significant association was found between watching mostly "beauty or fashion" content and body dissatisfaction.

Table 8. Unadjusted linear regression model and *P* value for the association between psychological variables and the most observed type of content in the overall sample[a].

	SCS[b]		DEQ-A[c]		DEQ-D[d]		BSQ-14[e]	
Most viewed content	β[f]	*P* value	β	*P* value	β	*P* value	β	*P* value
Sports, fitness, or lifestyle	–0.041	.40	–0.002	.99	–0.085	.66	6.099	.01
Beauty or fashion	–0.068	.12	0.366	.05	0.355	.04	3.417	.11
Family or friends	–0.045	.22	0.078	.62	0.240	.09	4.156	.02
Humor	–0.100	.02	0.268	.16	0.170	.30	7.859	<.001

[a]Reference category in a linear regression model to which other types of content were compared to science.

[b]SCS: Self-Compassion Scale.

[c]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[d]DEQ-D: dependence subscale of the Depressive Experiences Questionnaire.

[e]BSQ-14: Body Shape Questionnaire.

[f]β: regression coefficient.

XSL•FO
RenderX

Impact of the Most Viewed Type of Content on Younger Individuals

At last, Table 9 displays results from the linear regression model that analyzed the relationship between the type of most viewed content on Instagram and self-compassion, self-criticism, dependence, and body dissatisfaction by age group. No statistically significant relationship was found in any of the age groups between the type of most viewed content on Instagram and self-compassion. In regards to the aged 18-35 years group, a statistically significant relationship was observed between self-criticism and watching mostly "beauty or fashion" content (*P*=.03). Individuals in this age group who watch mainly "beauty or fashion" content scored on average 0.530 (SD 0.24) points higher in self-criticism than those who mainly watch "science" content. On the other hand, a statistically significant relationship was observed between dependence and mainly watching "beauty or fashion" content (*P*=.02). Compared to individuals whose most watched content is "science," individuals who mainly watch "beauty or fashion" content scored on average 0.508 (SD 0.22) points higher in dependence. As for body dissatisfaction, a statistically significant relationship was observed only between this variable and observing mainly "humor" content (*P*=.01). Individuals aged 18-35 years who consume mainly "humor" content scored on average 6.46 (SD 2.62) points higher in body dissatisfaction than those who consume mainly "science" content. In the aged 35-50 years group, no statistically significant relationships were observed between the type of content most viewed on Instagram and the studied variables.

Table 9. Unadjusted linear regression model and *P* value for the association between psychological variables and the most observed type of content stratified by age[a].

	SCS[b]		DEQ-A[c]		DEQ-D[d]		BSQ-14[e]	
	β[f]	*P* value	β	*P* value	β	*P* value	β	*P* value
Aged 18-35 years (n=751)								
Sports, fitness, or lifestyle	–0.013	.84	0.059	.82	0.015	.95	5.212	.09
Beauty or fashion	–0.070	.21	0.530	.03	0.508	.02	3.044	.28
Family or friends	–0.005	.91	–0.017	.93	0.273	.14	2.865	.22
Humor	–0.071	.17	0.096	.67	0.207	.32	6.460	.01
Aged 35-50 years (n=300)								
Sports, fitness, or lifestyle	–0.049	.56	–0.453	.20	–0.351	.23	5.156	.19
Beauty or fashion	0.005	.95	–0.370	.24	–0.055	.83	1.064	.76
Family or friends	–0.136	.08	–0.346	.27	–0.025	.93	1.643	.64
Humor	–0.051	.60	0.214	.60	–0.146	.67	5.656	.21

[a]Reference category in a linear regression model to which other types of content were compared to science.

[b]SCS: Self-Compassion Scale.

[c]DEQ-A: self-criticism subscale of the Depressive Experiences Questionnaire.

[d]DEQ-D: dependence subscale of the Depressive Experiences Questionnaire.

[e]BSQ-14: Body Shape Questionnaire.

[f]β: regression coefficient.

Discussion

Principal Findings

We investigated the impact of Spanish people's Instagram use on psychological well-being as assessed through the domains of self-compassion, self-criticism, and body dissatisfaction. The innovations in our study include an assessment of the type of content and the total time of use, together with the impact on these psychological domains. According to our findings, greater Instagram daily time of use is associated with higher self-criticism scores in participants of all ages. In regards to the most viewed type of content on this social network, observing primarily sports or fitness and family or friends content predicted higher body dissatisfaction scores in all the samples. At last, observing predominantly beauty or fashion content predicted higher self-criticism scores only among the youngest participants.

There was a positive association between Instagram usage (we refer to duration of use as usage hereinafter) and self-criticism. Users who spent more time on Instagram, particularly those spending over 3 hours per day on the platform, exhibited higher self-criticism scores. When we analyzed self-criticism to the time spent on Instagram within different age groups, we found a positive and significant association between consuming Instagram for more than 3 hours a day in both age groups, but no association with using the social network for 1-3 hours a day. These findings align with previous research exploring the link between Instagram usage time and other variables related to self-criticism. In a recent study, greater Instagram usage time was associated with higher levels of concern and yearning for perfection [8]. Additionally, problematic use of Facebook and Instagram (characterized by a strong motivation to access the social network and loss of control over usage time impacting psychological well-being) has been linked to negative

XSL•FO
RenderX

self-oriented perfectionism and socially prescriptive perfectionism [37].

However, we did not find a statistically significant relationship between Instagram usage time and self-compassion in either of the 2 age groups. In contrast to our findings, a different study found that higher Instagram use was associated with lower levels of self-compassion and poorer psychological well-being [13]. In that same study, individuals with higher levels of self-compassion tended to spend less time using Instagram.

No association was found between body dissatisfaction and daily Instagram usage. These unexpected findings contrast with previous studies that indicated greater daily usage of Instagram was associated with a greater tendency to think about one's appearance, body dissatisfaction, and lower self-esteem and psychological well-being [7,10,12,38]. The results from a recent study showed that the relationship between Instagram use and body image concerns is mediated by comparisons related to physical appearance [7]. Women who engage in appearance-related comparisons on Instagram and Facebook are more vulnerable to body dissatisfaction, particularly regarding their face, skin, and hair [39,40]. Social interactions on Facebook, such as looking at other profiles, liking, and leaving comments, have also been associated with higher levels of body dissatisfaction, but this variable was not associated with overall Facebook exposure [41]. These findings suggest that user body dissatisfaction may be influenced not just by the duration of Instagram use, but also by the nature of social comparisons and interactions regarding physical appearance. It is possible that duration of use does not fully explain the variation in users' body dissatisfaction, and other variables may be more relevant in explaining the relationship between Instagram usage and body dissatisfaction.

We found age differences for the impact of the type of content viewed. For those aged 18-35 years, a positive association was observed between self-criticism and predominantly viewing "beauty or fashion" content. Compared to those aged 18-35 years who primarily observe "science" content, those who mainly watch "beauty or fashion" content had higher self-criticism scores. It has been found that Instagram users regularly post pictures about their physical appearance that often feature filters or have been previously edited [7], promoting beauty standards and stereotypes that are difficult to achieve for most people [5,42,43]. Viewing images based on stereotyped beauty and idealized body forms increases the likelihood that users of this social network will engage in upward social comparison, leading to feelings of inferiority [42,44]. Those who predominantly observe "beauty or fashion" content may have a greater tendency to be more self-critical and experience greater feelings of inferiority, as they perceive the people depicted in such images to be more attractive than themselves.

However, in those aged 35-50 years, we found no correlation between predominantly viewing "beauty or fashion" content and self-criticism. It is plausible that participants aged 35-50 years may be less likely to compare themselves to the Instagram "beauty or fashion" images and might then have more realistic expectations about their self-image, which could decrease the likelihood of experiencing feelings of inferiority and the development of a negative view of themselves. Older women are less affected by beauty and thinness stereotypes promoted by society and, therefore, have higher self-esteem and psychological well-being [45-48]. Moreover, older people set more realistic expectations about themselves and are less concerned about their social status [49,50].

There was a significant association between body dissatisfaction and predominantly observing "sport, fitness, or lifestyle," "family or friends," and "humor" content. These findings are consistent with previous research where observing "fitness" images has been linked to lower body satisfaction and a greater desire to achieve thinness [40,51,52]. The internalization of beauty standards and the tendency to compare one's own appearance with that of women in fitness images likely mediates the development of body image concerns [40]. A recent study found that comparing oneself to idealized bodies perceived as more attractive led to greater weight and appearance dissatisfaction and lower confidence [53]. In this study, women who were more perfectionistic about their appearance experienced greater weight and appearance dissatisfaction after comparing themselves to Instagram models, compared to women who did not worry about having an imperfect appearance. Furthermore, the positive relationship between predominantly observing "family or friends" content and body dissatisfaction may also be explained by the phenomenon of upward social comparison. Family members and friends may use Instagram's filters and image editing strategies, presenting a version of themselves that does not correspond to their true self, but rather an attempt to achieve beauty stereotypes frequently promoted on this platform. Therefore, observing this type of content may lead users to perceive their physical appearance as less attractive than that of their relatives or friends.

However, there was no significant association between the most viewed content type and self-compassion, except "humor" content, which had a negative association with self-compassion. This contrasts with previous research finding women and men who viewed physical appearance-related content (images of muscular and toned bodies) reported lower levels of self-compassion compared to those who viewed neutral images about architecture [51,52]. The lack of our predicted positive relationship between predominantly observing appearance-related content and self-compassion may be explained by the emergence of health campaigns in fashion and social media, such as the "body positive" movement, that challenges dominant beauty and idealized body stereotypes as well as promoting acceptance and appreciation of all body types and self-compassion [52,54,55]. The lack of a statistically significant relationship between predominantly consuming beauty or fashion content and body dissatisfaction may be elucidated by the potential positive influence of the aforementioned campaigns. Furthermore, participants in our study may perceive beauty or fashion content as emphasizing characteristics related to their face, makeup, hair, or attire, in contrast to the sports, fitness, or lifestyle category, which may be perceived as centered on body shape and health. The instrument used to measure body dissatisfaction in our study specifically targeted aspects related to body shape, rather than encompassing characteristics of other anatomical regions, such

XSL•FO
RenderX

as the face. This specificity in measurement focus could contribute to the observed absence of a significant relationship between body dissatisfaction and beauty or fashion content.

We found higher Instagram usage time is associated with higher levels of self-criticism and that viewing physical appearance-related content on Instagram, such as beauty or fashion content and fitness, sports, or lifestyle content is associated with higher levels of self-criticism in the youngest participants and body dissatisfaction in the general sample, respectively. Although our study did not find an association between Instagram use and self-compassion, perhaps interventions that target the development of a self-compassionate attitude may contribute to a more positive vision that users have of themselves compared to others. Self-compassion may have a protective function against stressful situations in both adults and adolescents [23]. A nonjudgmental, accepting attitude toward personal characteristics and body image might help decrease the impact of Instagram-related social comparisons and self-criticism, and promote greater psychological well-being.

Related research on general social media use, such as the viewing of news during disasters, has demonstrated negative impacts on psychological well-being associated with duration of use. For example, social media use of more than or equal to 2 hours per day was associated with posttraumatic stress disorder and depressive symptoms in a longitudinal study of social unrest in Hong Kong from 2009 to 2019 [56]. Similarly, a systematic review of social networking site use found that differing engagement styles might explain the development of anxiety or depression, as we found for Instagram users in our study [57]. However, this broader systematic review did not identify a particular usage time that predicted effects. There is related literature on problematic social media usage, such as the use of Facebook, where users display addictive-type behavior that is often associated with the development of anxiety and depression [58].

Limitations

However, our study has some limitations that may affect the interpretation of our results. First, due to the cross-sectional nature of this study, causality cannot be attributed to the observed associations. Second, we have not conducted a clinical evaluation or a comprehensive psychological characterization of the participants beyond the use of self-administered questionnaires and scales. Conducting such evaluations would have added value to this study, considering that certain personality traits have been associated with a predisposition to internalize ideals propagated through media [59]. Third, we did not control for the absence of depression or anxiety diagnoses among our participants, which could have potentially influenced our results. Higher scores of self-criticism and body dissatisfaction may be indicative of symptoms related to these disorders rather than a direct consequence of Instagram use. Fourth, the majority of participants in our study were women, which limits our ability to assess the influence of gender on the results. We also could not evaluate the impact of culture since all participants were of Spanish nationality. Therefore, we do not know if our results are generalizable to individuals from other cultures.

Conclusions

Increased Instagram usage, especially for more than 3 hours per day, may hurt individuals' self-perception and beliefs concerning their physical appearance, and, through increased self-criticism reduce psychological well-being. Perhaps it is time to develop evidence-based guidelines on how to adopt a balanced approach to using social media, particularly visually oriented platforms such as Instagram. Further longitudinal studies to examine the effects of usage of Instagram and other social media platforms on body image and self-perception are needed to inform the development of more detailed evidence-based usage guidelines. In the interim, based on our research, perhaps Instagram users focusing on physical appearance content should limit their use to less than three hours a day to avoid negative effects on psychological well-being from social comparison.

Furthermore, these guidelines should take into account that the psychological impact of social media usage varies according to age, with younger individuals being more susceptible. Future longitudinal studies that examine the relationship between the type of Instagram usage and other social media platforms according to gender and culture are needed to provide a more exhaustive comprehension of these platforms and elaborate more specific interventions and guidelines on social media usage.

Acknowledgments

We extend our gratitude to one of the authors whose expansive Instagram outreach facilitated the broadening of our study's audience. We also thank the reviewers for their insightful feedback that helped enhance the quality of our paper.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire on Instagram usage.
[PDF File (Adobe PDF File), 133 KB - jmir_v26i1e51957_app1.pdf]

XSL•FO
RenderX

References

1. Statista. Number of global social network users? 2025. 2024. URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [accessed 2024-06-12]
2. Statista. Global social networks ranked by number of users 2022. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [accessed 2024-06-12]
3. Wong A, Ho S, Olusanya O, Antonini MV, Lyness D. The use of social media and online communications in times of pandemic Covid-19. J Intensive Care Soc 2021;22(3):255-260 [FREE Full text] [doi: 10.1177/1751143720966280] [Medline: 34422109]
4. Pertegal-Vega MA, Oliva-Delgado A, Rodríguez-Meirinhos A. Revisión sistemática del panorama de la investigación sobre redes sociales: taxonomía sobre experiencias de uso. Comunicar 2019;60(27):81-91 [FREE Full text] [doi: 10.3916/c60-2019-08]
5. Engeln R, Loach R, Imundo MN, Zola A. Compared to Facebook, Instagram use causes more appearance comparison and lower body satisfaction in college women. Body Image 2020;34:38-45. [doi: 10.1016/j.bodyim.2020.04.007] [Medline: 32505866]
6. Hanley S, Watt SE, Coventry W. Taking a break: the effect of taking a vacation from Facebook and Instagram on subjective well-being. PLoS One 2019;14(6):e0217743 [FREE Full text] [doi: 10.1371/journal.pone.0217743] [Medline: 31170206]
7. Senín-Calderón C, Perona-Garcelán S, Rodríguez-Testal JF. The dark side of Instagram: predictor model of dysmorphic concerns. Int J Clin Health Psychol 2020;20(3):253-261 [FREE Full text] [doi: 10.1016/j.ijchp.2020.06.005] [Medline: 32994798]
8. Simon PD, Cu SMO, De Jesus KEM, Go NTS, Lim KTF, Say CLC. Worried about being imperfect? The mediating effect of physical appearance perfectionism between Instagram addiction and body esteem. Pers Individ Differ 2022;186:111346 [FREE Full text] [doi: 10.1016/j.paid.2021.111346]
9. Skowronski M, Busching R, Krahé B. Links between exposure to sexualized Instagram images and body image concerns in girls and boys. J Media Psychol 2021;34(1):55-62 [FREE Full text] [doi: 10.1027/1864-1105/a000296]
10. Steinsbekk S, Wichstrøm L, Stenseng F, Nesi J, Hygen BW, Skalická V. The impact of social media use on appearance self-esteem from childhood to adolescence—a 3-wave community study. Comput Hum Behav 2021;114:106528 [FREE Full text] [doi: 10.1016/j.chb.2020.106528]
11. Tiggemann M, Anderberg I. Social media is not real: the effect of 'Instagram vs reality' images on women's social comparison and body image. New Media Soc 2020;22(12):2183-2199 [FREE Full text] [doi: 10.1177/1461444819888720]
12. Tiggemann M, Hayden S, Brown Z, Veldhuis J. The effect of Instagram "likes" on women's social comparison and body dissatisfaction. Body Image 2018;26:90-97. [doi: 10.1016/j.bodyim.2018.07.002] [Medline: 30036748]
13. Keyte R, Mullis L, Egan H, Hussain M, Cook A, Mantzios M. Self-compassion and Instagram use is explained by the relation to anxiety, depression, and stress. J Technol Behav Sci 2021;6(2):436-441 [FREE Full text] [doi: 10.1007/s41347-020-00186-z]
14. Sturman E, Mongrain M. Self-criticism and major depression: an evolutionary perspective. Br J Clin Psychol 2005;44(Pt 4):505-519. [doi: 10.1348/014466505X35722] [Medline: 16368030]
15. Sherry SB, Richards JE, Sherry DL, Stewart SH. Self-critical perfectionism is a vulnerability factor for depression but not anxiety: a 12-month, 3-wave longitudinal study. J Res Pers 2014;52:1-5. [doi: 10.1016/j.jrp.2014.05.004]
16. McIntyre R, Smith P, Rimes K. The role of self-criticism in common mental health difficulties in students: a systematic review of prospective studies. Ment Health Prev 2018;10:13-27 [FREE Full text] [doi: 10.1016/j.mhp.2018.02.003]
17. Werner AM, Tibubos AN, Rohrmann S, Reiss N. The clinical trait self-criticism and its relation to psychopathology: a systematic review—update. J Affect Disord 2019;246:530-547. [doi: 10.1016/j.jad.2018.12.069] [Medline: 30599378]
18. Ferreira C, Pinto-Gouveia J, Duarte C. Self-criticism, perfectionism and eating disorders: the effect of depression and body dissatisfaction. Int J Psychol Psychol Ther 2014;14(3):409-420 [FREE Full text] [doi: 10.1016/j.eatbeh.2013.01.005] [Medline: 23557822]
19. Carpenter R, Alloway TP. Exploring working memory, self-criticism, and rumination as factors related to self-harm. Psychol Rep 2023;126(4):1642-1660. [doi: 10.1177/00332941221074258] [Medline: 35084275]
20. Hooley JM, Ho DT, Slater J, Lockshin A. Pain perception and nonsuicidal self-injury: a laboratory investigation. Pers Disord: Theory, Res, Treat 2010;1(3):170-179. [doi: 10.1037/a0020106] [Medline: 22448633]
21. Xavier A, Gouveia JP, Cunha M. Non-suicidal self-injury in adolescence: the role of shame, self-criticism and fear of self-compassion. Child Youth Care Forum 2016;45(4):571-586 [FREE Full text] [doi: 10.1007/s10566-016-9346-1]
22. Campos RC, Holden RR, Baleizão C, Caçador B, Fragata AS. Self-criticism, neediness, and distress in the prediction of suicide ideation: results from cross-sectional and longitudinal studies. J Psychol 2018;152(4):237-255. [doi: 10.1080/00223980.2018.1446895] [Medline: 29667539]
23. Bluth K, Roberson PNE, Gaylord SA, Faurot KR, Grewen KM, Arzon S, et al. Does self-compassion protect adolescents from stress? J Child Fam Stud 2016;25(4):1098-1109 [FREE Full text] [doi: 10.1007/s10826-015-0307-3] [Medline: 26997856]
24. Neff K. Self-compassion: an alternative conceptualization of a healthy attitude toward oneself. Self Identity 2003;2(2):85-101 [FREE Full text] [doi: 10.1080/15298860309032]

XSL•FO
RenderX

25. Neff KD. The development and validation of a scale to measure self-compassion. Self Identity 2003;2(3):223-250 [FREE Full text] [doi: 10.1080/15298860309027]
26. Garcia-Campayo J, Navarro-Gil M, Andrés E, Montero-Marin J, López-Artal L, Demarzo MMP. Validation of the Spanish versions of the long (26 items) and short (12 items) forms of the Self-Compassion Scale (SCS). Health Qual Life Outcomes 2014;12:4 [FREE Full text] [doi: 10.1186/1477-7525-12-4] [Medline: 24410742]
27. Neff KD. The Self-Compassion Scale is a valid and theoretically coherent measure of self-compassion. Mindfulness 2016;7(1):264-274 [FREE Full text] [doi: 10.1007/s12671-015-0479-3]
28. Blatt SJ, D'Afflitti JP, Quinlan DM. Depressive experiences questionnaire. APA PsycTests 1976. [doi: 10.1037/t02165-000]
29. Linares L, Herrero-Fernández D, Iriarte L, Estévez A. Psychometric properties from a Spanish validation of Depressive Experiences Questionnaire (DEQ). Accion Psicol 2017;14(1):137-150. [doi: 10.5944/ap.14.1.17989]
30. Blatt SJ. Experiences of depression: theoretical, clinical, and research perspectives. Washington, D.C., United States: American Psychological Association; 2004.
31. Cooper PJ, Taylor MJ, Cooper Z, Fairburn CG. The development and validation of the Body Shape Questionnaire. Int J Eating Disord 1987;6(4):485-494 [FREE Full text] [doi: 10.1002/1098-108x(198707)6:4<485::aid-eat2260060405>3.0.co;2-o]
32. Dowson J, Henderson L. The validity of a short version of the Body Shape Questionnaire. Psychiatry Res 2001;102(3):263-271. [doi: 10.1016/s0165-1781(01)00254-2] [Medline: 11440777]
33. Imperatori C, Panno A, Giacchini M, Massullo C, Carbone GA, Clerici M, et al. Electroencephalographic correlates of body shape concerns: an eLORETA functional connectivity study. Cogn Neurodyn 2020;14(5):723-729 [FREE Full text] [doi: 10.1007/s11571-020-09618-1] [Medline: 33014184]
34. Pook M, Tuschen-Caffier B, Brähler E. Evaluation and comparison of different versions of the Body Shape Questionnaire. Psychiatry Res 2008;158(1):67-73. [doi: 10.1016/j.psychres.2006.08.002] [Medline: 18037499]
35. Franco-Paredes K, Bautista-Díaz ML, Díaz-Reséndiz FJ, Arredondo-Urtíz EM. Propiedades psicométricas del Body Shape Questionnaire (BSQ-14) en mujeres universitarias mexicanas. Evaluar 2018;18(1):1-11 [FREE Full text] [doi: 10.35670/1667-4545.v18.n1.19765]
36. Garner DM, Olmsted MP, Bohr Y, Garfinkel PE. The Eating Attitudes Test: psychometric features and clinical correlates. Psychol Med 1982;12(4):871-878. [doi: 10.1017/s0033291700049163] [Medline: 6961471]
37. Harren N, Walburg V, Chabrol H. Studying social media burnout and problematic social media use: the implication of perfectionism and metacognitions. Comput Hum Behav Rep 2021;4:100117 [FREE Full text] [doi: 10.1016/j.chbr.2021.100117]
38. Alfonso-Fuertes I, Alvarez-Mon MA, Sanchez Del Hoyo R, Ortega MA, Alvarez-Mon M, Molina-Ruiz RM. Time spent on Instagram and body image, self-esteem, and physical comparison among young adults in Spain: observational study. JMIR Form Res 2023;7:e42207 [FREE Full text] [doi: 10.2196/42207] [Medline: 37027197]
39. Baker N, Ferszt G, Breines JG. A qualitative study exploring female college students' instagram use and body image. Cyberpsychol Behav Soc Netw 2019;22(4):277-282. [doi: 10.1089/cyber.2018.0420] [Medline: 30855190]
40. Fardouly J, Diedrichs PC, Vartanian LR, Halliwell E. Social comparisons on social media: the impact of Facebook on young women's body image concerns and mood. Body Image 2015;13:38-45. [doi: 10.1016/j.bodyim.2014.12.002] [Medline: 25615425]
41. Kim JW, Chock TM. Body image 2.0: associations between social grooming on Facebook and body image concerns. Comput Hum Behav 2015;48:331-339. [doi: 10.1016/j.chb.2015.01.009]
42. Gilbert P. Evolution and social anxiety. the role of attraction, social competition, and social hierarchies. Psychiatr Clin North Am 2001;24(4):723-751. [doi: 10.1016/s0193-953x(05)70260-4] [Medline: 11723630]
43. Jiang S, Ngien A. The effects of Instagram use, social comparison, and self-esteem on social anxiety: a survey study in Singapore. Soc Media + Soc 2020;6(2):1-10 [FREE Full text] [doi: 10.1177/2056305120912488]
44. Thompson R, Zuroff D. The levels of self-criticism scale: comparative self-criticism and internalized self-criticism. Pers Individ Differ 2004;36(2):419-430 [FREE Full text] [doi: 10.1016/s0191-8869(03)00106-5]
45. Esnaola I, Rodríguez A, Goñi A. Body dissatisfaction and perceived sociocultural pressures: gender and age differences. Salud Ment 2010;33(1):21-29 [FREE Full text]
46. Tiggemann M. Body image across the adult life span: stability and change. Body Image 2004;1(1):29-41. [doi: 10.1016/S1740-1445(03)00002-0] [Medline: 18089139]
47. Tiggemann M. Body-size dissatisfaction: individual differences in age and gender, and relationship with self-esteem. Pers Individ Differ 1992;13:39-43 [FREE Full text] [doi: 10.1016/0191-8869(92)90215-b]
48. Webster J, Tiggemann M. The relationship between women's body satisfaction and self-image across the life span: the role of cognitive control. J Genet Psychol 2003;164(2):241-252. [doi: 10.1080/00221320309597980] [Medline: 12856818]
49. Blatt SJ. A fundamental polarity in psychoanalysis: implications for personality development, psychopathology, and the therapeutic process. Psychoanalytic Inquiry 2007;26(4):494-520 [FREE Full text] [doi: 10.1080/07351690701310581]
50. Blatt SJ, Luyten P. A structural-developmental psychodynamic approach to psychopathology: two polarities of experience across the life span. Dev Psychopathol 2009;21(3):793-814. [doi: 10.1017/S0954579409000431] [Medline: 19583884]

XSL•FO
RenderX

51. Barron AM, Krumrei-Mancuso EJ, Harriger JA. The effects of fitspiration and self-compassion Instagram posts on body image and self-compassion in men and women. Body Image 2021;37:14-27. [doi: 10.1016/j.bodyim.2021.01.003] [Medline: 33556914]
52. Slater A, Varsani N, Diedrichs PC. #fitspo or #loveyourself? The impact of fitspiration and self-compassion Instagram images on women's body image, self-compassion, and mood. Body Image 2017;22:87-96. [doi: 10.1016/j.bodyim.2017.06.004] [Medline: 28689104]
53. McComb SE, Mills JS. Young women's body image following upwards comparison to Instagram models: the role of physical appearance perfectionism and cognitive emotion regulation. Body Image 2021;38:49-62 [FREE Full text] [doi: 10.1016/j.bodyim.2021.03.012] [Medline: 33798801]
54. Brown Z, Tiggemann M. A picture is worth a thousand words: the effect of viewing celebrity Instagram images with disclaimer and body positive captions on women's body image. Body Image 2020;33:190-198. [doi: 10.1016/j.bodyim.2020.03.003] [Medline: 32289571]
55. Tiggemann M, Anderberg I, Brown Z. #Loveyourbody: the effect of body positive Instagram captions on women's body image. Body Image 2020;33:129-136. [doi: 10.1016/j.bodyim.2020.02.015] [Medline: 32151992]
56. Ni MY, Yao XI, Leung KSM, Yau C, Leung CMC, Lun P, et al. Depression and post-traumatic stress during major social unrest in Hong Kong: a 10-year prospective cohort study. Lancet 2020;395(10220):273-284. [doi: 10.1016/S0140-6736(19)33160-5] [Medline: 31928765]
57. Seabrook EM, Kern ML, Rickard NS. Social networking sites, depression, and anxiety: a systematic review. JMIR Ment Health 2016;3(4):e50 [FREE Full text] [doi: 10.2196/mental.5842] [Medline: 27881357]
58. Marino C, Gini G, Vieno A, Spada MM. The associations between problematic Facebook use, psychological distress and well-being among adolescents and young adults: a systematic review and meta-analysis. J Affect Disord 2018;226:274-281. [doi: 10.1016/j.jad.2017.10.007] [Medline: 29024900]
59. Retallack H, Ringrose J, Lawrence E. “Fuck your body image”: teen girls’ Twitter and Instagram feminism in and around school. In: Coffey J, Budgeon S, Cahill H, editors. Learning Bodies. Perspectives on Children and Young People, vol 2. Singapore: Springer; 2016.

Abbreviations

BSQ-14: Body Shape Questionnaire
DEQ: Depressive Experiences Questionnaire
SCS: Self-Compassion Scale

Edited by A Mavragani; submitted 17.08.23; peer-reviewed by V Astha, F Kong; comments to author 30.11.23; revised version received 15.01.24; accepted 23.03.24; published 01.08.24.

Please cite as:
Varaona A, Alvarez-Mon MA, Serrano-Garcia I, Díaz-Marsá M, Looi JCL, Molina-Ruiz RM
Exploring the Relationship Between Instagram Use and Self-Criticism, Self-Compassion, and Body Dissatisfaction in the Spanish Population: Observational Study
J Med Internet Res 2024;26:e51957
URL: https://www.jmir.org/2024/1/e51957
doi:10.2196/51957
PMID:39088263

©Andrea Varaona, Miguel Angel Alvarez-Mon, Irene Serrano-Garcia, Marina Díaz-Marsá, Jeffrey C L Looi, Rosa M Molina-Ruiz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Topic Modeling to Understand Patients' and Caregivers' Perspectives About Left Ventricular Assist Device: Thematic Analysis

Semyon Melnikov[1], PhD; Stav Klein[2], PhD; Moni Shahar[2], PhD; David Guy[3], MSc

[1]Nursing Department, Steyer School of Health Professions, Faculty of Medical & Health Sciences, Tel Aviv University, Tel Aviv, Israel
[2]The Tel Aviv University Center for AI and Data Science, Tel Aviv University, Tel Aviv, Israel
[3]Steyer School of Health Professions, Faculty of Medical & Health Sciences, Tel Aviv University, Tel Aviv, Israel

Corresponding Author:
Semyon Melnikov, PhD
Nursing Department, Steyer School of Health Professions
Faculty of Medical & Health Sciences
Tel Aviv University
Haim Levanon 55
Tel Aviv, 6997801
Israel
Phone: 972 36405456
Email: melniko@tauex.tau.ac.il

Abstract

Background: Heart failure (HF) is a significant global clinical and public health challenge, impacting 64.3 million individuals worldwide. To address the scarcity of donor organs, left ventricular assist device (LVAD) implantation has become a crucial intervention for managing end-stage HF, serving as a bridge to heart transplantation or as a destination therapy. Web-based health forums, such as MyLVAD.com, play a vital role as trusted sources of information for individuals with HF symptoms and their caregivers.

Objective: We aim to uncover the latent topics within the posts shared by users on the MyLVAD.com website.

Methods: Using the latent Dirichlet allocation algorithm and a visualization tool, our objective was to uncover latent topics within the posts shared on the MyLVAD.com website. Through the application of topic modeling techniques, we analyzed 459 posts authored by recipients of LVAD and their family members from 2015 to 2023.

Results: This study unveiled 5 prominent themes of concern among patients with LVAD and their family members. These themes included family support (39.5% weight value), encompassing subthemes such as family caregiving roles and emotional or practical support; clothing (23.9% weight value), with subthemes related to comfort, normalcy, and functionality; infection (18.2% weight value), covering driveline infections, prevention, and care; power (12% weight value), involving challenges associated with power dependency; and self-care maintenance, monitoring, and management (6.3% weight value), which included subthemes such as blood tests, monitoring, alarms, and device management.

Conclusions: These findings contribute to a better understanding of the experiences and needs of patients implanted with LVAD, providing valuable insights for health care professionals to offer tailored support and care. By using latent Dirichlet allocation to analyze posts from the MyLVAD.com forum, this study sheds light on key topics discussed by users, facilitating improved patient care and enhanced patient-provider communication.

(J Med Internet Res 2024;26:e50009) doi:10.2196/50009

KEYWORDS

left ventricular assist device; LVAD; topic modeling; health care forum; heart disease; cardiovascular condition; medical devices; devices for heart; latent Dirichlet allocation; cardiovascular; device; visualization tool; tool; heart; caregiver; monitoring; management; care; users; communication; heart failure

XSL•FO
RenderX

Introduction

Background

Heart failure (HF) is a serious clinical and public health problem affecting 64.3 million people worldwide [1]. With the limited availability of donor organs, management of end-stage HF has increasingly incorporated the implantation of a left ventricular assist device (LVAD) as a bridge to heart transplantation or as the destination therapy [2]. Web-based health forums have emerged as crucial and highly regarded sources of information for individuals experiencing HF or symptoms [3]. Additionally, these forums serve as valuable platforms for bridging knowledge gaps that may arise between medical consultations. They provide a space where individuals can interact with others who have similar experiences, fostering an environment that promotes enhanced understanding and mutual support [3]. Professionals concerning LVAD established MyLVAD.com, a dedicated website focused on the patient experience and care of individuals implanted with LVADs or those who are considering obtaining one [4]. The forum serves as a platform for patients with LVAD, caregivers, family members, and medical professionals to connect, share experiences, ask questions, and provide support to one another [4].

Prior Work

In a previous study conducted by Naik et al [5], computational modeling was used to analyze discussions from web-based support groups on MyLVAD.com. The researchers used a Structural Topic Model technique to identify major topics of concern among user groups and to uncover the coping strategies used by them. The data set analyzed in this study consisted of posts from users on the MyLVAD.com forum between 2011 and 2016. Nine major topics were detected: (1) supportive and appreciative posts; (2) emotional adjustment after LVAD; (3) carrying, wearing, and handling LVAD; (4) LVADs during travel; (5) journal papers and nursing positions; (6) worried wives; (7) medical complications with LVAD; (8) food; and (9) lists of nursing notes and medical information [5]. It is important to note that the study by Naik et al [5] did not provide specific details regarding the weight of each topic or the words reservoir used to identify the topics. Additionally, the study's findings were limited to posts collected within a specific time period.

In a subsequent study conducted by Austin et al [6], an analysis was conducted on posts, comments, and titles extracted from MyLVAD.com between April 2017 and October 2019. Using computational sentiment analysis, the study found that positive sentiment dominated the 20 most commonly used words, which indicated that, for the majority of patients and their families, LVADs and the associated experiences were viewed in a positive light [6]. Despite the positive sentiment associated with the majority of the most common words from the forum, the most common negative word in the sentiment analysis was "infection" (208 mentions), compared to other complications such as "stroke" (29 mentions), "bleeding" (30 mentions), and "thrombosis" or "clot" (32 mentions) [6].

Natural language processing (NLP) is a branch of artificial intelligence that focuses on analyzing and extracting insights from textual data [7]. Topic or text analysis is a method used to analyze large volumes of unstructured data, aiming to uncover the primary topics within each text and provide meaningful insights [8]. NLP facilitates topic analysis by automating the process through the application of various algorithms, primarily categorized into 2 main approaches: supervised and unsupervised [9]. A supervised algorithm involves the use of predetermined output attributes in addition to the input attributes, whereas an unsupervised algorithm discovers inherent groupings within unlabeled data and subsequently assigns labels to each data value [8]. The main objective of this study was to use an unsupervised algorithm and latent Dirichlet allocation (LDA), a widely used statistical modeling technique in NLP, to uncover the latent topics within the posts shared by users on the MyLVAD.com website.

Methods

Data Collection

For this study, data were collected from the forum section of the website MyLVAD.com using an automated extraction method. A custom Java script was developed to systematically gather information from the forum, including thread content. The script collected all links without applying any inclusion or exclusion criteria, accessed each link, and parsed the HTML to extract the main posts, comments, and titles. The collected data were then saved in a text file and underwent preprocessing and analysis. The data set used in this study comprised 459 threads that were analyzed. Figure 1 shows the data collection and analysis process. This study did not provide compensation to participants, as it was conducted using publicly available data from a web-based forum. The research used existing data, and there was no direct interaction with individuals or collection of private information.

It is important to highlight that our extraction methods strictly followed the terms of service and policies set by the MyLVAD.com website. We were also fully cognizant of the potential risks and concerns associated with using data from web-based forums and social media platforms, such as the potential for reidentification or unintended disclosure of sensitive personal information. No personal or identifying information of forum participants was used or disclosed in this study.

XSL•FO
RenderX

Figure 1. The process of data collection and analysis. LDA: latent Dirichlet allocation.

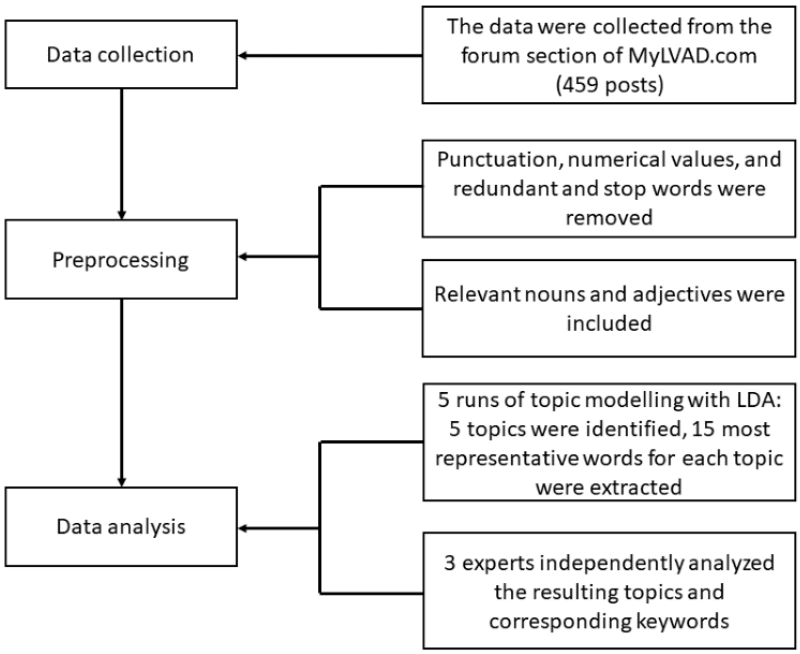

Preprocessing

The initial stage of data preprocessing aimed to enhance the quality of the data set and remove any potentially confounding elements that could impede subsequent analytical processes. Data processing included HTML formatting removal. Then, various forum-specific terms such as “author,” “comments,” “date,” “views,” and the names of days were eliminated from the data set, alongside special characters and numerical values deemed extraneous to this study’s objectives. Next, only nouns and adjectives were included using the NLP library spaCy (Explosion) [10]. Additionally, for each remaining word, its lemma (ie, its base or canonical form) was extracted using spaCy. This filtering process aimed to eliminate extraneous words and reduce the dimensionality of the data while preserving the semantic content of the text. The resulting data set comprised a more parsimonious and standardized set of features.

Before applying the LDA model to the preprocessed data, an analysis of term frequency was conducted to identify any words that may significantly skew the results of the topic modeling. In particular, frequent nouns and adjectives were examined to determine whether they would affect the topic distribution of the remaining words in the corpus. Common nouns and adjectives such as “doctor,” “patient,” and “hospital” were excluded since they were considered semantically vague and potentially detrimental to topic modeling. Additionally, words that appeared frequently throughout the data set, such as “husband,” “battery,” and “equipment,” were also flagged for exclusion due to their nondiscriminative nature concerning topic identification. The resulting data set was thus refined to include only the most salient and discriminative words, providing a more accurate representation of the underlying topics present in the corpus. The LDA analysis, without the exclusion of any words, as well as the standard list of stop words used, appears in Multimedia Appendix 1.

Topic Modeling With LDA

LDA is a statistical modeling technique used in NLP to uncover the underlying topics present in a large corpus of textual data [7]. LDA assumes that each document in the corpus is a mixture of topics and that each topic is a distribution over words. In other words, LDA posits that there are hidden or “latent” variables, namely the topics that generate the observed words in a document. The goal of LDA is to infer the topic proportions for each document and the word probabilities for each topic by analyzing the co-occurrence patterns of words across the corpus. The LDA algorithm works by iteratively estimating the topic proportions for each document and the word probabilities for each topic until convergence. During each iteration, the algorithm randomly assigns each word in the corpus to a topic and re-estimates the topic proportions and word probabilities based on the new assignment. This process continues until the model reaches a stable state where the topic proportions and word probabilities converge [7]. The topics are then extracted from the LDA model by analyzing the estimated parameters of the model, namely the topic-word probabilities and the document-topic probabilities. The topic-word probabilities are the probabilities of each word given a particular topic. They are computed during the training phase of the LDA model and represent the relative frequency of each word in a particular topic. The document-topic probabilities, on the other hand, represent the distribution of topics in each document. These probabilities are also computed during the training phase of the LDA model and are used to determine which topics are most relevant to each document in the corpus.

To obtain robust and reliable results, the LDA model was executed on the preprocessed data for 5 runs (times). Each run was designed to identify 5 topics, and the top 15 most representative words were extracted for each topic (a complete list of the hyper-parameters used in the model, which are critical for interpreting the LDA outputs, is provided in Multimedia Appendix 2). The most representative words linked with each

XSL•FO
RenderX

topic were identified based on the topic-word probabilities. Specifically, words that appeared in at least 2 of the 5 model runs were selected as the top words for each topic. This approach aimed to avoid spurious results and minimize the impact of model-specific hyper-parameters on the selected topic keywords. By selecting the most commonly occurring words across the different model runs, the resulting set of top words is considered more stable and representative of the underlying topics present in the corpus.

To validate the topics and their respective top words identified through the LDA model, 3 health care professionals (2 registered nurses with a Master of Arts and 1 author, a registered nurse with a Doctor of Philosophy) with expertise in LVAD were consulted. Further, 3 experts independently reviewed and analyzed the resulting topics and corresponding keywords. After a thorough examination, they reached an agreement regarding the validity and coherence of the identified topics, as well as their respective representative words. The consensus among the health care professionals added an important layer of verification to the LDA model results, providing further support for the accuracy and reliability of the identified topics and their associated keywords. Each topic had an assigned weight value, which was calculated by dividing the number of words in the collection assigned to the topic by the total number of words.

Ethical Considerations

The Tel Aviv University's ethics committee approved this study (#0006133). The ethics committee approval covers secondary analysis without additional consent.

Results

Overview

We identified 459 posts written by the unique users of MyLVAD.com forums from March 2015 to February 2023. A total of 945 users were identified with unique usernames. It is possible, however, that some individuals might have multiple usernames, although we lack the means to ascertain such instances. The top 5% of the most frequent users included 46 individuals. Among them, the user with the fewest posts had 14 posts, while the user with the most posts had 143 posts. It is worth mentioning that the user who ranked second in terms of post count was labeled "anonymous," indicating those users who opted not to select a name. Further, 5 themes were identified using LDA of the posts of users of MyLVAD.com: (1) family support; (2) clothing; (3) infection; (4) power; and (5) self-care maintenance, monitoring, and management. These 5 themes with 15 keywords each are listed in Table 1. Figure 2 displays the visualizations we generated for the topics. The figure illustrates the relative importance of topics through the sizes of the circles, while the distances between the topics indicate their similarity to each other [11].

Table 1. The top 5 themes of concern of patients with left ventricular assist device and their family members in MyLVAD.com forum discussions.

Theme number	Themes	Keywords	Weight (%)
1	Family support	family, good, recovery, bad, able, transplant, well, infection, many, failure, new, home, post, blood, change	39.5
2	Clothing	bag, belt, clothing, comfortable, controller, pocket, vest, shirt, new, good, great, side, size, small, shoulder	23.9
3	Infection	infection, dressing, skin, change, rash, kit, tape, sterile, water, shower, bandage, area, adhesive, anchor, driveline	18.2
4	Power	bag, charger, power, extra, sure, new, good, generator, line, transplant, unit, pump, shower, water, home	12
5	Self-care maintenance, monitoring, and management	blood, controller, good, transplant, alarm, device, flow, pressure, fluid, exercise, infection, medical, rehab, dad, pump	6.3

XSL•FO
RenderX

Figure 2. Latent Dirichlet allocation visualization of topic 2.

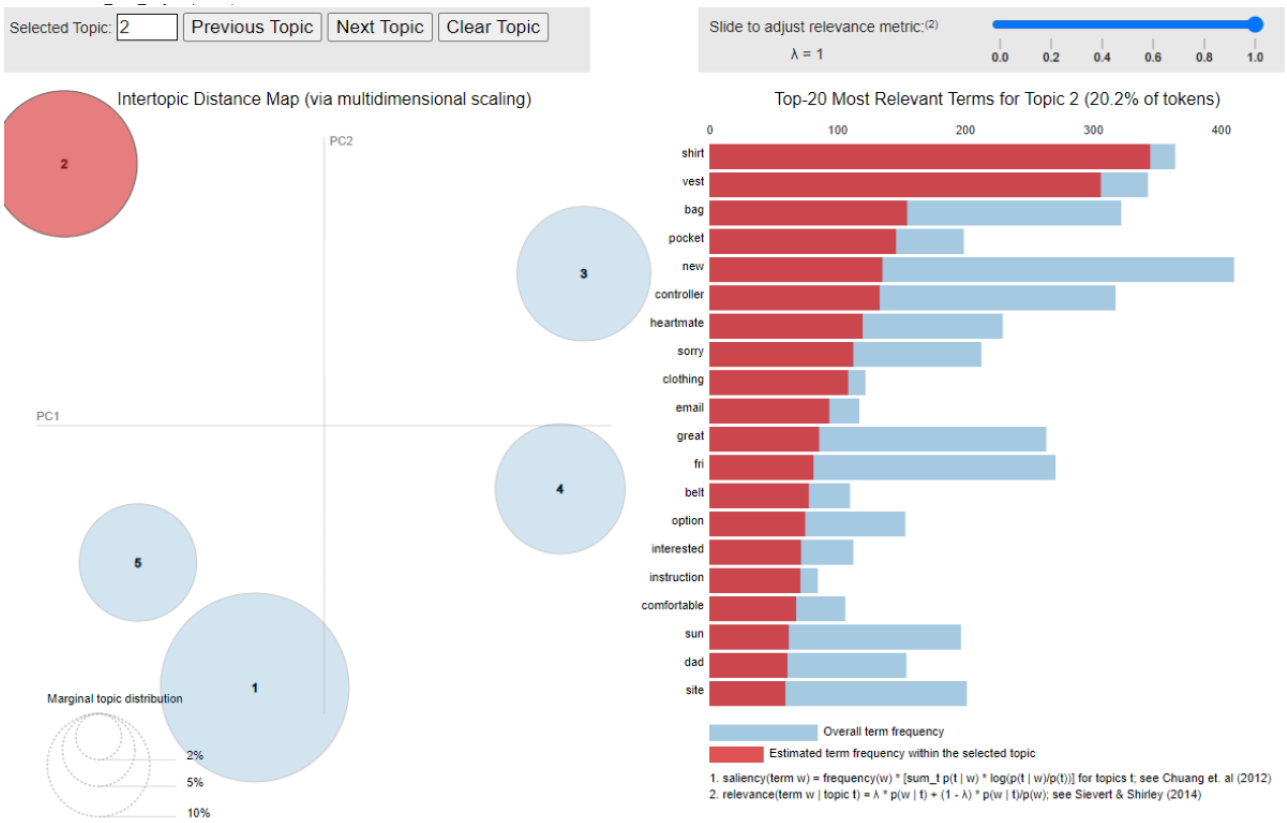

Family Support

The first theme was family or partner support. The weight value of this theme was 39.5%. The impact of family support was frequently mentioned in the context of post-LVAD implantation recovery. The importance of family support was mentioned both by family members as well by the patients themselves. Attributing patients' survival to the fact that their partners or family members were by their side was frequently used. Family caregiving role: the partner (usually the wife) was portrayed as a caregiver in many posts. The partner took care of the person with LVAD, ensuring their well-being, adjusting medications, changing dressings, and maintaining precautions. The theme of caregiving and the partners' dedication to this role was prominent. Emotional and practical support: partners not only took care of the individual's physical needs but also provided emotional reassurance, helping patients with LVADs concentrate on becoming better. The partner's support was seen as crucial for the successful outcome of the individual with the LVAD. Feeling like a burden: patients with LVADs expressed feelings of being useless and a burden to their partner. Family involvement: posts mentioned the involvement of the individual's children as care providers. While the focus was primarily on the partner's role, the presence of family members in the caregiving process was acknowledged.

Clothing

Clothing was another major theme. The weight value of this theme was 23.9%. Some users mentioned various types of vests and shirts designed specifically for individuals with LVADs. These garments were discussed in terms of their functionality, comfort, and ability to accommodate the LVAD's batteries and controller. Comfort and normalcy: posts emphasized the importance of finding clothing options that provided comfort and a sense of normalcy for individuals with LVADs. Wearing specific shirts, vests, or undergarments was mentioned as a way to feel more like a "normal person" despite the medical condition. Pockets and storage: users frequently mentioned the significance of pockets in LVAD clothing. Pockets are discussed in terms of their size, suitability for carrying batteries, and the convenience they offer in distributing the weight of the LVAD. Cost and bargain shopping: users provided information about where to find LVAD clothing at affordable prices, including discounted items or at specific stores. The mention of purchasing vests at marked-down prices and repurposing items from thrift stores highlighted a cost-conscious approach to acquiring suitable clothing.

Infection

Infection was another theme we identified. The weight value of this theme was 18.2%. Driveline infections: numerous posts mentioned driveline infections as a significant concern and fear for patients with LVADs. Prevention and care: multiple posts emphasized the importance of proper care, hygiene, and proactive measures to prevent infection, such as regular dressing changes, cleanliness around the driveline, and following hospital instructions for driveline management and cable integrity. Treatment and management: the text discussed the challenges of treating infections in patients with LVADs and the use of antibiotics, blood tests, debridement, and surgeries to address infection-related issues.

XSL•FO
RenderX

Power

The fourth theme was power. The weight value of this theme was 12%. The main subthemes were as follows. Dependence on power: the posts discussed the reliance on power for various aspects of the individual's medical device, such as the LVAD. They emphasized the importance of maintaining a power supply and switching to battery power when necessary. Power-related challenges: multiple posts mentioned various challenges related to power, such as power failures, the need to bring personal equipment for power connection, potential risks of power switches, and the need to adapt to different power supplies when traveling. External support and coordination: numerous posts mentioned the involvement of external entities, such as LVAD coordinators, electric companies, and the fire department, in managing power-related issues. Showering precautions: there was an emphasis on taking precautions during showering to prevent water from coming into direct contact with the equipment. The posts discussed various experiences and approaches to showering while having a LVAD, including the use of protective bags, adhesive patches, and hand-held shower wands to ensure the safety of the equipment and to prevent water from reaching the device.

Self-Care Maintenance, Monitoring, and Management

The fifth theme was self-care maintenance, monitoring, and management. The weight value of this theme was 6.3%. The following topics were identified. Blood tests and monitoring: numerous posts mentioned the importance of regular international normalized ratio blood assays, to assess the effectiveness of blood thinners. Alarms and device management: many posts mentioned the management of low flow alarms, fixes and adjustments to address alarm issues, the impact on daily life and quality of life, the importance of controller upgrades and team communication, and concerns about clots, internationally normalized ratio levels, and battery time. Physical exercises: several posts discussed finding suitable exercises and engaging in physical activities while considering the medical limitations associated with HF and LVAD. Infection prevention: numerous posts highlighted the risk, consequences, treatment, and prevention of driveline infections in patients with LVAD. Rehabilitation therapy: many posts revolved around the process of recovery and rehabilitation after a LVAD implantation, with a focus on physical rehabilitation, exercise, regaining strength, and the positive outcomes associated with rehabilitation therapy.

Discussion

Principal Findings

The principal findings of this innovative study, conducted through an exploration of patients' concerns in a dedicated LVAD forum, encompassed five major themes: (1) family support plays a crucial role in the recovery and success of patients with LVAD; (2) clothing designed for users of LVADs is important for comfort and a sense of normalcy; (3) infection prevention and management, particularly driveline infections, are significant concerns for patients with LVAD, requiring proactive measures and treatment; (4) power management is essential, including maintaining power supply, dealing with power-related challenges, and involving external support for coordination; and (5) self-care maintenance, monitoring, and management involve regular blood tests, device monitoring, physical exercises, infection prevention, and rehabilitation therapy.

These findings thus advance the understanding of patients with LVAD and caregiver experiences. Our study achieved this by integrating a nuanced exploration of web-based forum discussions with innovative analytical methods. In addition to this new approach, this study also enriched the body of knowledge on LVAD-related patient concerns and needs for support in several key areas.

Insight into familial support dynamics: the theme of "family support" emerged as the most prevalent among the discussions on the MyLVAD.com forums. Previously, using Structural Topic Model analysis, Naik et al [5], identified the "supportive and appreciative posts" topic among 9 major topics in web-based support groups on MyLVAD.com. Naik et al [5] primarily highlighted major topics of concern but did not examine the specifics of familial or caregiver support. Our study thus went a step further and provided an updated contextual analysis of how family support specifically influences the recovery and daily management of patients with LVAD. This study also shed light on the critical role of emotional and practical support from family members, extending the understanding of the social dimensions of care for patients with LVAD.

A previous study by Slade et al [12] identified the cluster "partner and family support" among 9 distinct clusters encompassing the experiences of 18 recipients of LVAD from a national heart transplant center in the United Kingdom. Our study expands on this by exploring a larger and more diverse data set from the MyLVAD.com forum, offering a broader perspective on how familial support is perceived and discussed among a wider community of patients with LVAD and caregivers. The contribution of family or caregiver is expressed not only in the gratitude of the patients as has been found in this study but also by the clinical impact the family or caregiver provides. Another previous study explored the relationships between caregiver status (caregivers that retained or resigned their roles, or patients implanted with a LVAD but without an assigned caregiver) and post-LVAD implant outcomes. They found that the 30-day postimplant readmission rate was 3-fold higher for patients whose caregivers quit [13]. Lastly, another study found that following LVAD implantation, the risk of death was over 3 times more likely among individuals who lived alone compared to those who did not [14]. These findings thus support our study's results regarding the significance and appreciation expressed by patients with LVAD toward their family members in general, and their partners in particular.

LVAD-related equipment and clothing needs: similar to this study, one of the topics identified in the MyLVAD.com users by Naik et al [5] was "carrying, wearing and handling LVAD." Also, similar to this study, this outcome encompassed various topics such as equipment and batteries, dressings and drivelines, and the challenges of finding appropriate clothing to wear with the LVAD equipment, as found among 18 recipients of LVAD from a national heart transplant center in the United Kingdom [12]. However, the current research goes beyond only

XSL•FO
RenderX

identifying the significance of clothing and equipment for patients with LVAD, as noted in such prior studies. Namely, this study explored the specific challenges and innovative solutions devised by patients to manage their LVAD equipment within their daily lives. The clothing theme not only revolved around the difficulty of finding suitable clothing to accommodate the LVAD equipment but also highlighted the patients' desire to maintain a sense of normalcy despite their medical condition. Similarly, previous studies demonstrated that patients with LVAD adjusted their clothing not only to suit their physical needs but also to blend in with the general public [15]. It seems that despite being a lifesaving medical device worn around the clock, the user experience offered by the design of the wearable system and components of LVADs falls short of optimal. Hopefully, modern developments leading to the creation of biocompatible, lightweight, wireless LVAD devices [16] could alleviate the challenges associated with equipment and clothing. Meanwhile, clinicians can apply the current findings to offer practical advice on handling LVAD equipment in everyday life.

Another significant theme identified was "power." This theme encompassed the concerns expressed by patients with LVAD regarding the maintenance of a reliable power supply for their LVADs, as well as the challenges they faced related to power when traveling and showering. Similarly, "LVADs during travel" was one of the major topics identified in posts from users on MyLVAD.com [5]. The current findings are also consistent with the concerns reported by patient-caregiver dyads with LVADs during cardiac rehabilitation, as highlighted in a study by Rapelli et al [17]. In that study, fear of power cuts and the need to maintain a stable power supply were among the worries expressed by patients and their caregivers. Similarly, Slade et al [12] also identified LVAD challenges as a cluster, which included patients' concerns about equipment during shopping and the risk of getting the equipment wet. While previous research has touched upon power as a concern for patients with LVAD, this study provides a more detailed exploration of subthemes within the power category. It delves into specific aspects such as dependence on power, challenges related to power failures and traveling, the need for external support and coordination, the role of external entities such as LVAD coordinators, electric companies, and fire departments, as well as practical strategies for showering precautions. These findings collectively indicate that patients with LVADs attach great importance to a reliable power supply of their LVADs, as it represents a highly significant component of their daily lives. Based on the challenges and needs highlighted in this study, clinicians could provide ongoing self-management training for patients with LVAD to help them gain independence in maintaining a reliable power supply. Recent and potential technological advancements aimed at addressing power supply concerns for LVADs, including innovations in battery technology, wireless charging, or more energy-efficient devices [18], would help mitigate the challenges faced by patients during travel, showering, and daily activities.

"Infection" surfaced as another key theme. This finding aligns with a previous study by Austin et al [6], who found that the term "infection" was mentioned disproportionately more frequently in the MyLVAD.com forum compared to other complications. While the risk of driveline infections is a known concern, this study provides a more nuanced understanding by detailing patients' fears, preventive measures, and the complexities involved in maintaining driveline hygiene. Currently, there is a lack of specialized training for managing cardiovascular infections in general and LVAD-associated infections in particular, despite their complexity and prevalence [19]. El-Dalati et al [19] advocated for creating programs to train health care professionals in infectious diseases, cardiology, and addiction medicine, ensuring expert, coordinated care for cardiovascular infections patients. Introducing appropriate training could potentially lower infection rates and reduce the complications from infections among patients with LVADs. At the same time, clinicians can use the insights from this study to improve patient education programs, focusing on correct driveline maintenance, identifying signs of infection, and understanding when to seek medical assistance, thereby highlighting the critical role of preventing infections.

Another significant theme identified was "self-care maintenance, monitoring, and management." The concept of self-care is crucial in the management of chronic illnesses, as emphasized by Riegel et al [20] who defined self-care as a process that involves maintaining health through health-promoting practices and managing illness. This process comprises 3 components: self-care maintenance, self-care monitoring, and self-care management. For patients with LVADs, self-care maintenance involves behaviors aimed at maintaining physical and emotional stability [21]. These behaviors encompass operating the LVAD system, caring for percutaneous leads, and engaging in lifestyle activities such as maintaining hygiene, adhering to medications, participating in physical exercise, and following an appropriate diet. Self-care monitoring involves the ongoing monitoring of the LVAD device, symptoms and signs of HF, potential side effects and complications, as well as psychological distress. Self-care management refers to the patient's response to symptoms and signs, which includes handling alarms, managing wound care, adjusting medications, diet, and rest, as well as seeking appropriate support from the LVAD and HF health care team [21]. Indeed, the discussions and interactions in MyLVAD.com forums show that patients and their families engage in conversations about self-care, asking questions and offering insights on various aspects of self-care management. These findings support previous results from Naik et al [5] who found "carrying, wearing, and handling LVAD," "medical complications with LVAD," and "food" among the major topics of posts from users on the MyLVAD.com forum. Clinicians can thus use MyLVAD.com forum discussions to develop personalized self-care plans addressing patients' specific needs and lifestyles. Lastly, the benefits of such forums are not limited to users of LVADs. Prior research has shown that social media platforms offer social support and practical means for enabling self-care and emotional assistance to individuals living with various other medical conditions, such as those with diabetes [22].

Limitations and Future Research

This study has several limitations. This study relied on web scraping data only from the forum section of MyLVAD.com.

XSL•FO
RenderX

This introduces a potential bias, as the data collected only represents the discussions and experiences shared on this particular website. Another limitation is a sample size consisting of 459 threads. While this may be a substantial amount of data, the sample size is relatively small considering the diversity and complexity of experiences within the patients with LVAD population. A larger sample size would provide a more comprehensive understanding of the topics and themes discussed by patients with LVAD and caregivers. An additional limitation might be data generalizability. This study focuses on the forum discussions from MyLVAD.com, which may not represent the broader population of patients with LVAD and caregivers. Accordingly, the findings of this study may not be generalizable to other LVAD populations or different cultural contexts.

Future research aims to conduct a longitudinal study to track the experiences of patients with LVAD and their caregivers over an extended period of time. This would provide insights into the long-term challenges, coping strategies, and outcomes associated with LVAD implantation, in addition to a comparative analysis comparing the experiences and outcomes of patients with LVAD who participate in web-based support forums such as MyLVAD.com with those who do not. This could help determine the effectiveness and benefits of web-based support networks for patients with LVAD and identify areas for improvement.

Conclusion

This comprehensive study, using LDA to analyze discussions from the MyLVAD.com forum, unveiled 5 pivotal themes central to the experiences of patients with LVAD: family support, clothing, infection, power, and self-care maintenance, monitoring, and management. Each theme represents a critical aspect of the patient's journey and underscores the multifaceted nature of living with a LVAD. Recognizing these themes' significance not only enhances the understanding but also directs their practical implications for improving patient care and support systems. The prominence of family support as the most significant theme, accounting for 39.5% of the thematic weight, underscores the indispensable role of caregivers in the postimplantation recovery phase. This finding suggests a need for health care systems to implement structured support and education programs for both patients and their caregivers, acknowledging the integral role of the family in the care continuum. The discussions around clothing and power supply challenges highlight practical aspects of daily living with a LVAD, pointing toward a need for innovations in device design and patient education that cater to these needs. The detailed exploration of infection prevention and management strategies reflects ongoing patient concerns despite existing medical guidance, indicating a gap in effective communication or the applicability of current advice. The theme of self-care maintenance, monitoring, and management reinforces the critical role of patient engagement in managing their condition. This finding suggests further investigation into the barriers and facilitators to effective self-care practices among patients with LVAD, potentially leading to the development of more intuitive and supportive self-care tools and resources.

In conclusion, this study contributes to a richer understanding of the lived experiences of patients with LVAD and the complex interplay of clinical, practical, and emotional factors that define their care journey. By illuminating these key themes, our findings provide a foundation for enhancing patient care, guiding future research, and informing the design of more responsive health care interventions for patients with LVAD and their families.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

SM, SK, MS, and DG conceived and designed this study. SK and MS collected and analyzed the data. SM, SK, MS, and DG interpreted the results. SM wrote the initial draft of this paper and SK, MS, and DG edited this paper. All authors approved the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The latent Dirichlet allocation analysis, without the exclusion of any words, and list of stopwords.
[DOCX File , 357 KB - jmir_v26i1e50009_app1.docx]

Multimedia Appendix 2
Latent Dirichlet allocation configuration.
[DOCX File , 14 KB - jmir_v26i1e50009_app2.docx]

References

1. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. Eur J Heart Fail 2020 Aug;22(8):1342-1356 [FREE Full text] [doi: 10.1002/ejhf.1858] [Medline: 32483830]

XSL•FO
RenderX

2. Thiha S, Zaidi ARZ, Robert CA, Abbas MK, Malik BH. A rising hope of an artificial heart: left ventricular assisted device—outcome, convenience, and quality of life. Cureus 2019 Sep 10;11(9):e5617 [FREE Full text] [doi: 10.7759/cureus.5617] [Medline: 31696010]
3. Farnood A, Johnston B, Mair FS. An analysis of the diagnostic accuracy and peer-to-peer health information provided on online health forums for heart failure. J Adv Nurs 2022 Jan;78(1):187-200 [FREE Full text] [doi: 10.1111/jan.15009] [Medline: 34369604]
4. Boyce S, Christensen D. MyLVAD.com. 2023. URL: https://www.mylvad.com/ [accessed 2023-06-02]
5. Naik AR, Lohmueller LC, Rose CP, Antaki JF. Improving decision support tools for LVAD recipients via computational analyses of online support group discussions. Journal of Cardiac Failure 2017 Aug;23(8):S107-S108. [doi: 10.1016/j.cardfail.2017.07.315]
6. Austin MA, Saxena A, O'Malley TJ, Maynes EJ, Moncure H, Ott N, et al. Computational sentiment analysis of an online left ventricular assist device support forum: positivity predominates. Ann Cardiothorac Surg 2021;10(3):375-382 [FREE Full text] [doi: 10.21037/acs-2020-cfmcs-fs-11] [Medline: 34159118]
7. Chowdhary K. Natural language processing. In: Fundamentals of Artificial Intelligence. New Delhi, India: Springer; 2020:603-649.
8. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar;28(1):e100262 [FREE Full text] [doi: 10.1136/bmjhci-2020-100262] [Medline: 33653690]
9. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf A. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Berry MW, Mohamed A, Yap BW, editors. Supervised and Unsupervised Learning for Data Science. Cham, United States: Springer International Publishing; 2020:3-21.
10. Jugran S, Kumar A, Tyagi B, Anand V. Extractive automatic text summarization using spaCy in Python & NLP. 2021 Presented at: International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE); March 4-5, 2021; Greater Noida, India p. 582-585. [doi: 10.1109/icacite51222.2021.9404712]
11. Chuang J, Ramage D, Manning C, Heer J. Interpretation and trust: designing model-driven visualizations for text analysis. USA: ACM; 2012 Presented at: SIGCHI Conference on Human Factors in Computing Systems; May 12, 2012; New York, NY p. 443-452. [doi: 10.1145/2207676.2207738]
12. Slade AL, O'Hara ME, Quinn D, Marley L, Griffith S, Calvert M, et al. Living with a left ventricular assist device: Capturing recipients experiences using group concept mapping software. PLoS One 2022;17(9):e0273108 [FREE Full text] [doi: 10.1371/journal.pone.0273108] [Medline: 36129928]
13. Koeckert M, Vining P, Reyentovich A, Katz SD, DeAnda A, Philipson S, et al. Caregiver status and outcomes after durable left ventricular assist device implantation. Heart Lung 2017;46(2):74-78. [doi: 10.1016/j.hrtlng.2016.12.001] [Medline: 28057337]
14. Bruce CR, Minard CG, Wilhelms LA, Abraham M, Amione-Guerra J, Pham L, et al. Caregivers of patients with left ventricular assist devices: possible impacts on patients' mortality and interagency registry for mechanically assisted circulatory support-defined morbidity events. Circ Cardiovasc Qual Outcomes 2017 Jan;10(1):e002879. [doi: 10.1161/CIRCOUTCOMES.116.002879] [Medline: 28073849]
15. Krimminger DM, Sledge JA. A qualitative study of life with a left ventricular assist device as a bridge to transplant: a new normal. Intensive Crit Care Nurs 2022;71:103230. [doi: 10.1016/j.iccn.2022.103230] [Medline: 35397981]
16. Tu J, Xu L, Li F, Dong N. Developments and challenges in durable ventricular assist device technology: a comprehensive review with a focus on advancements in China. J Cardiovasc Dev Dis 2024 Jan 18;11(1):29 [FREE Full text] [doi: 10.3390/jcdd11010029] [Medline: 38248899]
17. Rapelli G, Giusti EM, Donato S, Parise M, Pagani AF, Pietrabissa G, et al. "The heart in a bag": the lived experience of patient-caregiver dyads with left ventricular assist device during cardiac rehabilitation. Front Psychol 2023;14:1116739 [FREE Full text] [doi: 10.3389/fpsyg.2023.1116739] [Medline: 37089738]
18. Deng J, Sun X, Peng H. Power supplies for cardiovascular implantable electronic devices. EcoMat 2023 Mar 25;5(6):e12343. [doi: 10.1002/eom2.12343]
19. El-Dalati S, Paras ML, Strnad L, Harris AW, Gurley JC, Sekela M, et al. In plain sight: the need for a dedicated cardiovascular infectious disease subspecialty. JACC Adv 2024 Jan;3(1):100748 [FREE Full text] [doi: 10.1016/j.jacadv.2023.100748] [Medline: 38939821]
20. Riegel B, Jaarsma T, Strömberg A. A middle-range theory of self-care of chronic illness. ANS Adv Nurs Sci 2012;35(3):194-204. [doi: 10.1097/ANS.0b013e318261b1ba] [Medline: 22739426]
21. Kato N, Jaarsma T, Ben Gal T. Learning self-care after left ventricular assist device implantation. Curr Heart Fail Rep 2014 Sep;11(3):290-298. [doi: 10.1007/s11897-014-0201-0] [Medline: 24831883]
22. Elnaggar A, Ta Park V, Lee SJ, Bender M, Siegmund LA, Park LG. Patients' use of social media for diabetes self-care: systematic review. J Med Internet Res 2020 Apr 24;22(4):e14209 [FREE Full text] [doi: 10.2196/14209] [Medline: 32329745]

XSL•FO
RenderX

Abbreviations

HF: heart failure
LDA: latent Dirichlet allocation
LVAD: left ventricular assist device
NLP: natural language processing

Edited by T de Azevedo Cardoso; submitted 16.06.23; peer-reviewed by W Ceron, A Sarraju; comments to author 15.02.24; revised version received 31.03.24; accepted 22.05.24; published 13.08.24.

Please cite as:
Melnikov S, Klein S, Shahar M, Guy D
Using Topic Modeling to Understand Patients' and Caregivers' Perspectives About Left Ventricular Assist Device: Thematic Analysis
J Med Internet Res 2024;26:e50009
URL: https://www.jmir.org/2024/1/e50009
doi:10.2196/50009
PMID:

©Semyon Melnikov, Stav Klein, Moni Shahar, David Guy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

How Does an Online Mental Health Community on Twitter Empower Diverse Population Levels and Groups? A Qualitative Analysis of #BipolarClub

Horeya AbouWarda[1], MSc; Mateusz Dolata[1], PhD; Gerhard Schwabe[1], PhD

Department of Informatics, Faculty of Business, Economics and Informatics, University of Zurich, Zurich, Switzerland

Corresponding Author:
Horeya AbouWarda, MSc
Department of Informatics
Faculty of Business, Economics and Informatics
University of Zurich
Binzmuehlestrasse 14
Zurich, 8050
Switzerland
Phone: 41 44 635 75 83
Fax: 41 44 635 68 09
Email: abouwarda@ifi.uzh.ch

Abstract

Background: Social media, including online health communities (OHCs), are widely used among both healthy people and those with health conditions. Platforms like Twitter (recently renamed X) have become powerful tools for online mental health communities (OMHCs), enabling users to exchange information, express feelings, and socialize. Recognized as empowering processes, these activities could empower mental health consumers, their families and friends, and society. However, it remains unclear how OMHCs empower diverse population levels and groups.

Objective: This study aimed to develop an understanding of how empowerment processes are conducted within OMHCs on Twitter by identifying members who shape these communities, detecting the types of empowerment processes aligned with the population levels and groups outlined in Strategy 1 of the Integrated People-Centred Health Services (IPCHS) framework by the World Health Organization (WHO), and clarifying members' involvement tendencies in these processes.

Methods: We conducted our analysis on a Twitter OMHC called #bipolarclub. We captured 2068 original tweets using its hashtag #bipolarclub between December 19, 2022, and January 15, 2023. After screening, 547 eligible tweets by 182 authors were analyzed. Using qualitative content analysis, community members were classified by examining the 182 authors' Twitter profiles, and empowerment processes were identified by analyzing the 547 tweets and categorized according to the WHO's Strategy 1. Members' tendencies of involvement were examined through their contributions to the identified processes.

Results: The analysis of #bipolarclub community members unveiled 5 main classifications among the 182 members, with the majority classified as individual members (n=138, 75.8%), followed by health care–related members (n=39, 21.4%). All members declared that they experience mental health conditions, including mental health and general practitioner members, who used the community as consumers and peers rather than for professional services. The analysis of 547 tweets for empowerment processes revealed 3 categories: individual-level processes (6 processes and 2 subprocesses), informal carer processes (1 process for families and 1 process for friends), and society-level processes (1 process and 2 subprocesses). The analysis also demonstrated distinct involvement tendencies among members, influenced by their identities, with individual members engaging in self-expression and family awareness support and health care–related members supporting societal awareness.

Conclusions: The examination of the #bipolarclub community highlights the capability of Twitter-based OMHCs to empower mental health consumers (including those from underserved and marginalized populations), their families and friends, and society, aligning with the WHO's empowerment agenda. This underscores the potential benefits of leveraging Twitter for such objectives. This pioneering study is the very first to analyze how a single OMHC can empower diverse populations, offering various health care stakeholders valuable guidance and aiding them in developing consumer-oriented empowerment programs using such OMHCs. We also propose a structured framework that classifies empowerment processes in OMHCs, inspired by the WHO's Strategy 1 (IPCHS framework).

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55965) doi:10.2196/55965

KEYWORDS

social media; Twitter; online mental health community; OMHC; empowerment processes; diverse population levels and groups; World Health Organization; WHO; Integrated People-Centred Health Services; IPCHS framework (Strategy 1)

Introduction

The Role of Social Media and Online Mental Health Communities in Consumer Empowerment

In this era of digital interconnectedness, the pervasive use of social media [1], including online health communities (OHCs), has become ubiquitous among both healthy people and those dealing with health conditions [2]. Platforms such as Twitter (recently renamed X) [3,4], Facebook [5], YouTube [6], and Reddit [7] have emerged as powerful tools for the formation and use of online mental health communities (OMHCs), in which users can exchange health-related information, advice, personal experiences, and social support; engage socially; express personal feelings; raise awareness about mental health and influence public conceptions; or merely observe the interactions of others [2,4,5,8-10]. Participating in these "empowerment processes" may lead to significant empowerment outcomes for users of OMHCs [11,12]. This empowerment potential of OMHCs could support various health consumer types, whether they are directly or indirectly engaged with health care systems [13]. It extends beyond mental health consumers (individuals coping with mental health conditions) to include their supporters and informal carers (eg, family and friends) and the broader public. The potential empowerment outcomes for mental health consumers engaging in OMHCs include better health self-management and navigation of health care systems as well as more self-esteem, self-efficacy, and involvement in decision-making during consultations with professionals, which could lead to better health outcomes and quality of life [5,9,14,15]. These online communities could also benefit supporters and informal carers and society at large by enhancing mental health literacy and awareness [16]. However, despite extensive research investigating and acknowledging the pivotal role of social media and OMHCs in fostering consumer empowerment, it has remained unclear how OMHCs empower diverse population levels and groups.

The Global Consumer Empowerment Movement and the Use of Social Media and OMHCs

Consumer empowerment implies providing people with the educational and supportive resources and opportunities they need in order to acquire the knowledge and skills necessary to have active roles in decisions and actions that impact their health [13,17]. Furthermore, the collaboration between health care stakeholders (eg, organizations and professionals) and health consumers is integral to consumer empowerment in health care systems [13]. This collaboration with consumers, commonly known as "coproduction" and "cocreation" of knowledge generation, decision-making, health services, and healthy environments, plays a crucial role in addressing their specific needs and elevating the overall effectiveness of health care [13,18].

Interest in consumer empowerment has fueled a global health care movement in which individuals, their families, informal carers, and communities are supported to actively participate in decisions and actions that influence their health [13,17,19]. Various health care stakeholders have adopted this movement worldwide, including health systems, authorities, institutions, professionals, and researchers, as well as nongovernmental health organizations, such as the World Health Organization (WHO) [17,19]. Consumer empowerment has become a global choice, as it could not only improve consumers' health outcomes but could also address the challenges faced by health care systems (eg, rising health care costs, chronic disease burdens, and staff shortages) and contribute to their effectiveness and sustainability [17].

To steer the consumer empowerment movement, the WHO has provided comprehensive guidelines, exemplified by the Integrated People-Centred Health Services (IPCHS) framework, particularly its Strategy 1 (empowering and engaging people and communities), which outlines four key approaches: (1) empowering and engaging individuals and families, (2) empowering and engaging communities, (3) empowering and engaging informal carers, and (4) reaching the underserved and marginalized [13]. In addition, specifically in mental health, the WHO's Comprehensive Mental Health Action Plan 2013-2030 has embraced consumer empowerment principles [20], emphasizing the importance of empowering and involving mental health consumers, their families, informal carers, and communities in mental health care. These principles are further endorsed in other WHO statements, such as the World Mental Health Report [21] and the Guidance on Community Mental Health Services [22]. In addition to these principles, in the action plan and the other 2 statements, the WHO has encouraged related organizations and professionals to increase the use of social media and digital support groups (eg, online communities) as integral components of digital mental health solutions.

Building on the IPCHS framework's principles, including its Strategy 1, which suggests the integral role of OMHC users as participants in health systems and aims to provide responsive health services that better address people's needs both within and beyond the health sector [13], OMHCs could serve as valuable environments for supporting Strategy 1. These online communities could facilitate consumer value cocreation in health care [18,23], in which health consumers from diverse population levels and groups and multiple health care stakeholders (eg, health organizations, professionals, and OMHC moderators) [5,8,10,16,24] could engage in a collaborative process of sharing knowledge, experiences, and insights. These activities could not only benefit consumers within OMHCs but could also enhance the overall quality and responsiveness of online and offline health care services (eg, the development of consumer-oriented empowerment programs in OMHCs and improvements to conventional mental health services). Thus,

XSL•FO
RenderX

OMHCs could support the objectives of the IPCHS framework and its Strategy 1, extending beyond traditional mental health service paradigms to promote a more inclusive and accessible consumer empowerment approach. Furthermore, although the WHO's encouragement to use these platforms may primarily seek to harness such potential advantages, it may also indicate an attempt to fill gaps in mental health services and resources [25] that may not be adequately or readily available elsewhere. This could be due to the challenges within health care systems, such as insufficient funding for alternative mental health solutions. Despite these potential benefits of OMHCs, they are not immune to the drawbacks of online environments. Some studies have reported risks associated with the use of OMHCs, such as the potential of developing unrealistic expectations or confusion about one's condition due to untrustworthy information from others' experiences or advice [9] and the possibility of adverse events arising from exposure to triggering content or negative interactions [5]. Nonetheless, despite these risks, the vital role of OMHCs in fostering consumer empowerment cannot be overlooked, given their potential to improve access to mental health resources and address existing service gaps.

Although several health care stakeholders, such as mental health organizations and health professionals [26,27], use OMHCs on social media for consumer empowerment activities, their endeavors cannot be fully realized without a nuanced understanding of how these communities empower diverse population levels and groups. In other words, it is crucial to identify the members who use and form OMHCs, the various types of empowerment processes in such communities, and how their members contribute to these processes. These insights can help various health care stakeholders in leveraging the full potential of OMHCs, enabling the design of tailored supportive resources aligned with specific needs and the development of consumer-oriented mental health programs and services. In addition, a recent study has called for a better understanding of OMHCs on social media regarding how these communities are formed, who their members are, and how they work [5], which underscores the necessity for comprehensive insights to fully leverage their potential and contribute valuable knowledge to the ongoing research in this field.

Existing Research on Consumer Empowerment Processes in OMHCs and OHCs

Consumer empowerment processes are activities that allow people to be capable, active, and engaged in controlling actions or decisions that impact their health [11,12]. Research in the field of OMHCs and empowerment activities has primarily focused on the empowerment of a specific population level and/or group, including mental health consumers at the individual level [5,15] and those from racial and ethnic minority groups [8]. Some studies were influenced by conceptual frameworks for personal mental health recovery, such as the POETIC (Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identity, Connectedness) and the CHIME (Connectedness, Hope and Optimism, Identity, Meaning and Purpose, Empowerment) frameworks [28]. Other studies have concentrated on informal carers and supporters, notably family and friends (pivotal in the well-being of individuals with mental health conditions) [24,29], along with societal-level considerations [16]. These studies indicate that OMHCs on social media involve empowerment processes aligned with the population levels and groups outlined in Strategy 1 of the WHO's IPCHS framework. However, to the best of our knowledge, existing research has not yet examined these processes in the context of Strategy 1 within OMHCs. In response to the WHO's call for enhanced use of these online platforms in mental health, gaining deeper insights into empowerment processes aligned with Strategy 1 in OMHCs is crucial to guide effective use. In addition, investigating OMHCs as a digital mental health solution while considering diverse populations and adhering to global standards is of great research interest [30], as it promises to address mental health disparities and enhance global access to care.

Furthermore, current research on empowerment activities has focused mainly on social media–based OHCs addressing various health conditions [31] or web-based OHCs focusing on non–mental health conditions [32,33] rather than specifically investigating social media–based OMHCs. Given the acknowledged differences in the nature of social media–based and web-based OHCs [34] and the variability in support needs across different circumstances [35], extending findings from OHCs not focused on social media and mental health to social media–based OMHCs may not be entirely appropriate. It is also imperative to focus on OHCs dedicated to mental health, given the recent rapid global increase in mental health conditions [36]. This surge in the number of individuals with mental health conditions has put additional strain on health care systems, which may lead, for instance, to longer intervals between consultations. Meanwhile, these individuals may already encounter substantial barriers to using mental health services. For example, recent studies focused on bipolar disorder have reported such barriers faced by those affected, including the delay in obtaining the right diagnosis, limited local availability of psychiatrists, and a lack of insurance coverage [25,37]. Given these challenges, mental health consumers may increasingly turn to OMHCs on popular social media platforms to seek support and connect with peers or professionals, drawing on their familiarity with these platforms from other contexts. Therefore, it is crucial that we urgently examine this specific context.

Twitter as a Mental Health Solution for Consumer Empowerment

Twitter is one of the most popular social media platforms used in health care practices [34,38]. It offers a distinctive outlet for its users to connect and engage in empowering activities over a popular, publicly accessible, and user-friendly information distribution platform, facilitating the widespread dissemination of information to a wider audience without constraints tied to any account or community following [39,40]. Communities on Twitter are often created bottom-up around a hashtag. Hashtags (a # sign followed by one or more words relating to a specific topic) are usually used to label tweets. These hashtags allow several users to view and contribute to the discussed topic, formulating communities of users talking about common topics, such as #depressionsucks [41] and #schizophrenia [27]. Several studies have reported that mental health–related hashtags on

XSL•FO
RenderX

Twitter are valuable solutions for various empowering activities, such as discussing symptoms, personal experiences, life challenges, and mental health care policies; raising awareness; combating stigma; and exchanging different kinds of support [8,16,27,41,42]. Other studies have documented empowerment outcomes for those who use Twitter for mental health discussions [4,43]. Most of these studies have directly and indirectly acknowledged Twitter's features as conducive to creating a mental health support ecosystem capable of reaching a wide audience, highlighting its potential to support global empowerment initiatives for various populations. However, despite extensive research on Twitter and mental health, the specific processes through which OMHCs on Twitter could empower diverse population levels and groups in line with global empowerment objectives, such as Strategy 1 of the WHO's IPCHS framework, have not yet been elucidated.

Purpose of This Study

This study aimed to describe and explain how OMHCs on Twitter empower diverse population levels and groups. Therefore, we formulated three subquestions to address how empowerment processes are conducted within an OMHC on Twitter called #bipolarclub:

1. Who are the members that use and form the OMHC, and how do they participate in it?
2. What are the types of empowerment processes conducted within the OMHC, in accordance with the following population levels and groups outlined in Strategy 1 of the WHO's IPCHS framework?
   - Mental health consumers
   - Underserved and marginalized mental health consumers
   - Their families and informal carers
   - The broader community (society)
3. What are the tendencies of the OMHC's members in contributing to and being involved in these empowerment processes?

Methods

Overview

To achieve our research goals, we conducted a qualitative exploratory study on the #bipolarclub community. This method offers a nuanced understanding and classification of behaviors, surpassing the scope of statistical analyses. Specifically, we opted for qualitative content analysis because it is a widely recognized and validated method used in previous research to systematically code and categorize communication content on social media platforms [41,44].

Our choice of the #bipolarclub community as the subject of our study is deliberate and aligned with our research context; moreover, this community is active, interactive, substantial, heterogeneous, and rich in its activities, meeting the recommended criteria outlined by Kozinets [45] and Salzmann-Erikson and Eriksson [46] for exploratory studies of online communities. Furthermore, the #bipolarclub community is not merely a congregation around a specific hashtag. It is a well-established entity with a dedicated Twitter account managed by the community's moderators, who oversee its activities. This characteristic sets it apart as a stable and enduring community rather than one formed around a transient trend.

Data Collection

We extracted tweets originating from the #bipolarclub hashtag in the Bipolar Club community on Twitter. As evident by its website [47], #bipolarclub is an OMHC that was established in March 2021 by mental health consumers. Despite its name, which suggests a focus on bipolar disorder and a primary aim of supporting those dealing with it, a nuanced perspective is presented on the community's Twitter account. A prominently pinned tweet clarifies that the community extends beyond a specific disorder or particular persons, welcoming anyone interested in mental health issues generally [48].

A total of 2068 tweets were collected for a period of 4 weeks from December 19, 2022, to January 15, 2023. We used the Twitter search application programming interface (API; Twitter Inc; now X Corp) [49] and the Postman API development environment (Postman Inc) to gather data end points from Twitter that included the #bipolarclub hashtag. The collected data for each tweet contained its content and the Twitter profile of its author (user's profile information). From a total of 2068 captured tweets, we excluded some based on the following criteria: (1) the type of tweet was not an original tweet or a reply tweet (retweets and quote tweets, n=1105, 53.4%); (2) the text was not in English (other languages, n=243, 11.8%); and (3) the tweet had no text content or contained irrelevant information (visual content, spam, and irrelevant tweets, n=173, 8.4%). Although tweets should include text content, we also considered embedded media in the analysis, such as images, graphics interchange formats (GIFs), videos, emojis, and links to external websites. After eliminating irrelevant tweets, our final data set consisted of 547 (26.5%) tweets written by 182 authors, as shown at the top of Figure 1.

XSL•FO
RenderX

Figure 1. A flowchart of the data analysis process. IPCHS: Integrated People-Centred Health Services; WHO: World Health Organization.

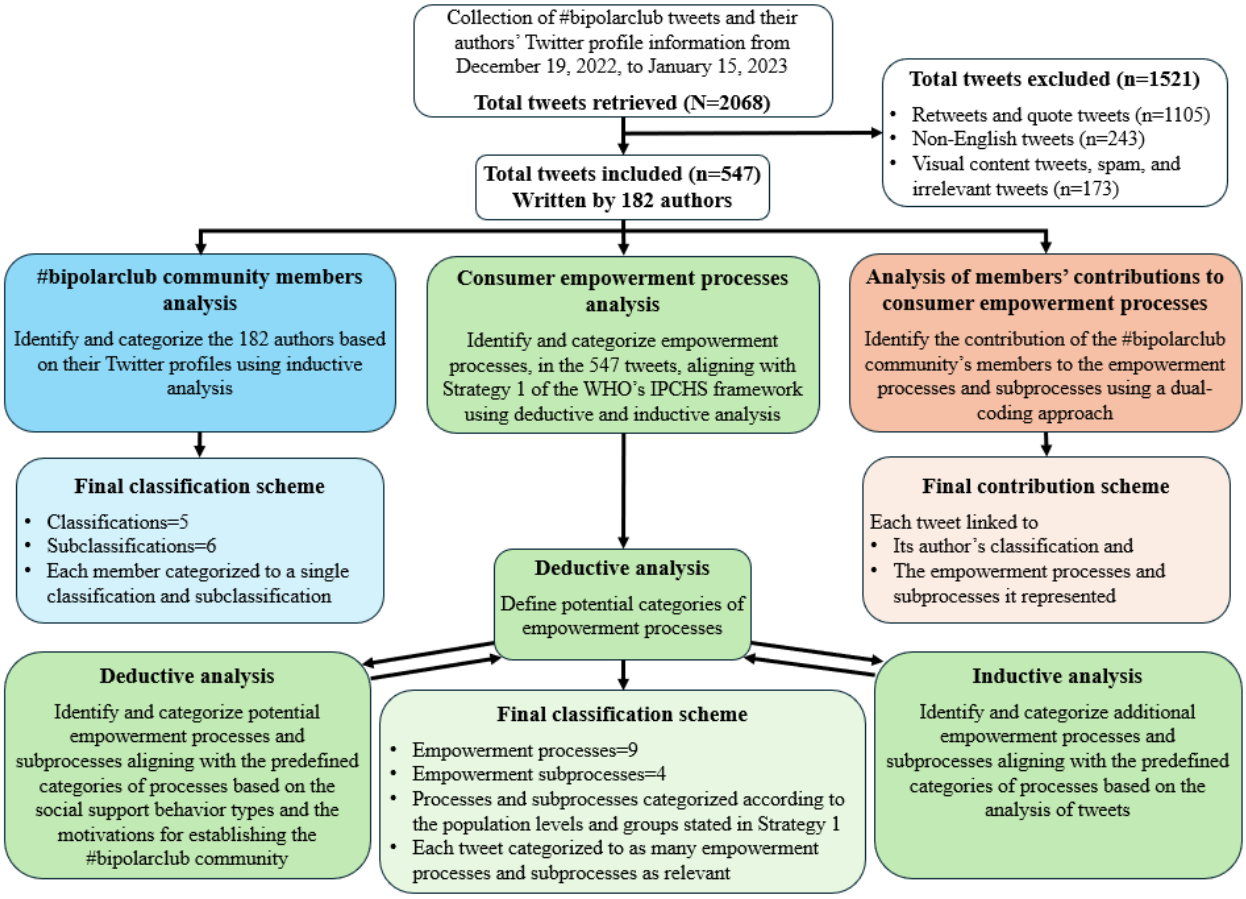

Ethical Considerations

Twitter, the platform that hosts the #bipolarclub community, operates as a public platform, allowing its content to be publicly available for research purposes [16]. Notably, only tweets from users with public profiles could be extracted; thus, private profiles are protected by default from being involved in research studies. Given the public nature of the tweets used for the analysis, our study meets the criteria for waiving informed consent [50,51]. Although we met these criteria, we deemed it ethically necessary to inform the #bipolarclub community's crew about our study intentions beforehand. The first author (HA) communicated with the moderator via private messages from a dedicated Twitter account for the study to the community's account and clearly identified the researcher's identity and the study purpose. Subsequently, the moderator discussed our intentions with other crew members and supported our study through a private message on the established communication channel.

Throughout the study, we also adhered to the ethical guidelines of the Association of Internet Researchers [52] to protect Twitter users' privacy. After the analysis, any information that may have led to the identification of users (eg, @usernames and names) was replaced with pseudonyms in the reported tweets, except for those belonging to institutions that provide mental health care services. The reported tweets were also paraphrased to maintain anonymity, and precautions were taken to ensure that no identifiable information could be retrieved through the Twitter search function or search engines. We ensured that these paraphrased tweets reflect the original tweets' content. This study's research protocol and procedures were reviewed and approved by the Human Subjects Committee of the Faculty of Business, Economics and Informatics at the University of Zurich (OEC IRB # 2022-093).

Data Analysis

Overview

After the data collection, all tweets were imported into a secure password-protected Excel spreadsheet (Microsoft Corp) for subsequent qualitative analysis. To facilitate this analysis, we used NVivo software (version 20; QSR International) to conduct qualitative content analysis [53]. The initial analysis was spearheaded by the first author (HA), who developed a preliminary coding scheme to classify #bipolarclub community members and to thematically analyze the empowerment processes conveyed in the tweets. To ensure a systematic approach, hierarchical structures were generated, including classifications and subclassifications for members analysis, along with descriptive categories for empowerment processes analysis. These structures facilitated a standardized comparison among members and across empowerment processes, and they were independently reviewed by the second and third authors (MD and GS) and discussed again, leading to refinement by the entire research team. The 3 researchers who oversaw this analysis have extensive expertise in qualitative research methodologies.

XSL•FO
RenderX

OMHC Members Analysis

To identify #bipolarclub community members, we applied an inductive approach to categorize them [53]. We used the user information of the 182 authors of the tweets to inspect their Twitter profiles using the Twitter search function to classify their entities. The categorization was based on cues from their personal descriptions (bios) and the tweets posted on their profiles [54], including pinned tweets as well as tweets posted during the 4-week study period (with or without the hashtag #bipolarclub). However, we only considered their tweets that included the hashtag #bipolarclub to detect their engagement within the community. Each member was assigned to the most pertinent category.

Consumer Empowerment Processes Analysis

To detect the empowerment processes aligning with the population levels and groups outlined in Strategy 1 of the WHO's IPCHS framework in the #bipolarclub community, we analyzed the 547 #bipolarclub tweets using a combined deductive and inductive analysis [53]. Initially, we used 3 frameworks: Strategy 1, social support behavior types, and the motivations of mental health consumers for establishing the #bipolarclub community. This deductive phase laid the groundwork for a categorization structure that encompasses potential empowerment processes.

Empowerment process categories were defined based on Strategy 1 [13], and prior research considered this strategy in mHealth [55], aligning with the population levels and groups stated in the 4 substrategies of Strategy 1: individuals and families (Strategy 1.1), communities (Strategy 1.2), informal carers (Strategy 1.3), and underserved and marginalized individuals (Strategy 1.4). We then used the other 2 frameworks, social support behavior types and motivating factors that led to the community's establishment, to identify potential empowerment processes and subprocesses aligning with the predefined process categories. The inclusion of social support behavior types stemmed from their prevalence as empowerment processes within OHCs [31]. We also incorporated the motivations that drove mental health consumers to establish the #bipolarclub community because these drivers encapsulate the support needs that these consumers aim to address through this online community, which is a key facet of their empowerment. The social support behavior types were derived from the social support behavior code [32,56,57]. This code encompasses five social support types: (1) informational support, (2) tangible support, (3) esteem support, (4) network support, and (5) emotional support. The motivations for establishing the #bipolarclub community on Twitter were obtained from its website [47]; there were five drives: (1) connecting and making friends, (2) learning from one another, (3) free self-expression, (4) telling personal stories, and (5) fighting stigma. Duplications and overlaps among the 5 social support types and the 5 motivations for establishing the community were addressed, refining the considered empowerment processes for the deductive analysis.

Through an inductive analysis, we then identified additional empowerment processes that derived directly from the data, supplemented by insights from existing empirical and theoretical research. The identified empowerment processes and subprocesses, derived from both deductive and inductive analyses, were systematically organized based on the predefined categories of empowerment processes that align with the population levels and groups outlined in Strategy 1. Each tweet was coded into as many empowerment processes and subprocesses as were relevant, acknowledging that a single tweet may contain multiple statements reflecting various empowerment aspects. This comprehensive coding approach ensured that our analysis appropriately accounted for the multifaceted nature of tweets and the diversity of their content.

Examination of OMHC Members' Contributions to Consumer Empowerment Processes

To identify the tendencies of contribution and involvement of #bipolarclub community members in the empowerment processes and subprocesses within the community, we applied a dual-coding approach. We coded each tweet with (1) its author's classification code and (2) the empowerment process and subprocess codes it represented. This method enabled us to link the 547 tweets to both their authors' classifications within the community and the specific empowerment processes and subprocesses to which they contributed. Figure 1 demonstrates the data analysis process in our study to address its 3 research objectives.

Results

OMHC Member Analysis

A total of 547 #bipolarclub tweets were written by 182 authors from 13 countries: the United States (n=49, 26.9%), the United Kingdom (n=26, 14.3%), Canada (n=6, 3.3%), Australia (n=3, 1.6%), Norway (n=3, 1.6%), South Africa (n=2, 1.1%), Belgium (n=1, 0.5%), India (n=1, 0.5%), Kenya (n=1, 0.5%), Malaysia (n=1, 0.5%), Russia (n=1, 0.5%), Saudi Arabia (n=1, 0.5%), and Sweden (n=1, 0.5%). However, some users either did not have an available location (n=47, 25.8%) or mentioned a fictional location (n=39, 21.4%).

We categorized the 182 members into 5 main classifications and 6 associated subclassifications. Of the 182 members, the majority were individual members (138/182, 75.8%), followed by health care–related members (39/182, 21.4%), crew and moderator members (3/182, 1.6%), an organizational member (1/182, 0.5%), and the community's Twitter account (1/182, 0.5%). All of these members, including health care–related members (academics and practitioners), the organizational member (the founder and chairman of the peer support foundation), and those managing the community's account, were mental health consumers experiencing mental health conditions. This identification was based on statements within their tweets or their Twitter bios, in which they openly shared their experiences of mental health conditions. Furthermore, a key feature of their bios was that many members provided clear insights into their situations as mental health consumers, such as "trying to deal with bipolar," "mental health advocate," and "Twitter as a venting space." In addition, several members used usernames and names featuring terms relating to "bipolar" to explicitly express their mental health condition. Classifications

XSL•FO
RenderX

and subclassifications of members and their descriptions and proportions of representing the #bipolarclub community are presented in Table 1. Moreover, Multimedia Appendix 1 shows sample paraphrased personal descriptions (bios) and posted tweets on the Twitter profiles of members, offering further clarity regarding their classifications and subclassifications.

Table 1. Classifications and subclassifications of #bipolarclub community members (N=182) and their proportions of representing the community.

Classification and subclassifications	Description	Proportion of representing the community, n (%)
Individual member	A member who is an individual participant and does not have a health care–related affiliation	138 (75.8)
Health care–related member	A member who is an individual participant and has a volunteering, professional, or academic health care–related affiliation	39 (21.4)
Mental health advocate	A member who is an active mental health advocate that supports promoting mental health care	33 (84.6)
Mental health advocate and academic	A member who is an active mental health advocate that supports promoting mental health care and has an academic affiliation in the mental health care field (eg, researcher, graduate, and student)	2 (5.1)
Mental health practitioner	A member who has a professional affiliation in the mental health care field (mental health care service provider)	2 (5.1)
Mental health advocate and general practitioner	A member who is an active mental health advocate that supports promoting mental health care and has a professional affiliation in the health care field (general health care service provider)	1 (2.6)
Mental health academic	A member who has an academic affiliation in the mental health care field (eg, researcher, graduate, and student)	1 (2.6)
Crew and moderator member	A member who has established and/or run the online community	3 (1.6)
Organizational member	A member that is an organizational participant and has a health care–related affiliation	1 (0.5)
Peer support foundation	A member that provides mental health peer support services as a nonprofit organization	1 (100)
Community's account	A digital member represents the online community's entity on Twitter, which is run by its moderators	1 (0.5)

The individual members (138/182, 75.8%) who represented the majority of members in the #bipolarclub community were individuals from the general public with no health care–related affiliations. The tweets posted by these individual members primarily revolved around their personal lives and thoughts. The second largest category of members was composed of 21.4% (39/182) of persons with health care–related affiliations, including volunteering (mental health advocates); professional (mental health or general practitioners); or academic (researchers, graduates, or students) affiliations. In this category, members were classified into 5 subcategories. The largest subcategory included mental health advocates (33/39, 85%), who mainly tweeted content that reflected their active advocacy activities for promoting mental health care, such as volunteering for mental health organizations, providing motivational and awareness talks, supporting those who are struggling, and fighting stereotypes and stigma of mental health. It is also worth noting that both practitioner types (mental health and general practitioners) in this community tweeted as mental health consumers and peers, not for professional services, despite being identifiable as professionals in the health care–related classification. Their tweets suggested that they were primarily seeking support from the community. However, they were posting informative tweets to help as peers. This shows how the #bipolarclub community as an OMHC stimulated health care providers to engage as mental health consumers, seeking support rather than using it for professional health care service purposes.

The third category involved crew and moderator members, who represented 1.6% (3/182) of the online community. These members had established the community and/or were running its activities. These activities included retweeting supportive tweets, replying to members, monitoring discussions, posting announcements, contributing with helpful resources, and moderating audio conversations held on Twitter Spaces [58] through the community's account, which help make the online community a positive outlet for its members [5]. The next classification was formed of an organizational member (1 peer support foundation), representing 0.5% (1/182) of the community. This foundation was a nonprofit organization that provides peer support services. It was tweeting to share its activities, including peer support groups provided by its team members, and to introduce new team members to the community. The last classification consisted of the Twitter account of the #bipolarclub community, with 0.5% (1/182) representing the community's members. We considered the community's account as a member in this categorization because it serves as a digital representative of the online community, tweeting and interacting with its members.

XSL•FO
RenderX

Consumer Empowerment Processes Analysis and Examination of OMHC Members' Contributions

Through the analysis of 547 #bipolarclub tweets, we identified 9 empowerment processes and 4 corresponding subprocesses within the #bipolarclub community. We structured these processes according to the population levels and groups outlined in the WHO's Strategy 1 (IPCHS framework), along with the substrategy corresponding to each population level and group, as shown in Figure 2.

As illustrated in Figure 2, we positioned mental health consumers at the center, with the underserved and marginalized individuals as a subgroup, encircled by 2 tiers: their supporters and informal carers (primarily their families, followed by their friends) in the inner tier and the broader society in the outer tier. This figure suggests that the #bipolaclub community could foster a supportive and empowering environment, with its influence extending beyond individual well-being to a broader societal impact. As indicated by the 3-dotted boxes in Figure 2, we classified the 9 identified empowerment processes and their 4 associated subprocesses into 3 categories: individual-level processes, informal carer processes, and society-level processes. Individual-level processes support mental health consumers and address the needs of the underserved and marginalized individuals. Informal carer processes extend support to family members and friends, while society-level processes aim to benefit society at large. Our examination revealed that the identified empowerment processes within the #bipolarclub community addressed all the 4 substrategies of Strategy 1. Table 2 provides detailed descriptions, frequencies of tweets, and the contribution proportions of the community members for the identified empowerment processes and subprocesses.

Figure 2. A structured framework of empowerment processes and subprocesses within the #bipolarclub community formulated according to the population levels and groups stated in Strategy 1 of the Integrated People-Centred Health Services framework by the World Health Organization and incorporated the 4 substrategies of Strategy 1 to correspond to each population level and group.

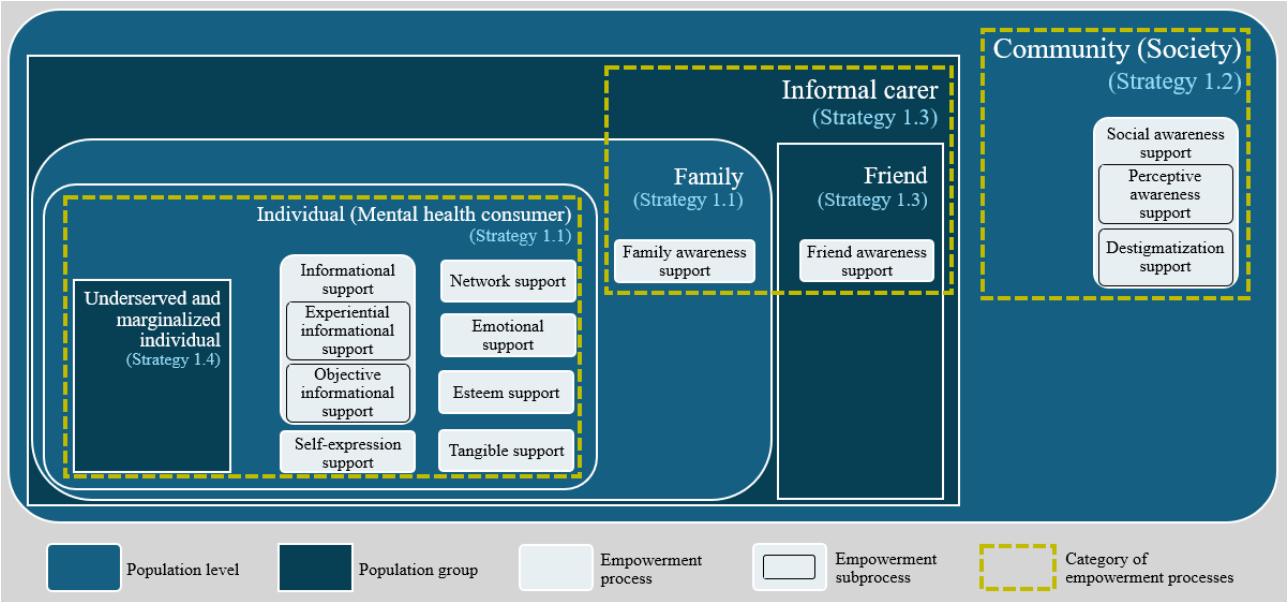

XSL•FO
RenderX

Table 2. Empowerment processes and subprocesses detected from #bipolarclub tweets (n=547) aligned with the population levels and groups outlined in Strategy 1 of the Integrated People-Centred Health Services framework by the World Health Organization, a description and the frequency of tweets for each process, and the contribution proportions of the community members.

Population level and group, substrategy of Strategy 1[a], and empowerment process and subprocess	Description	Sample paraphrased tweet	Tweet frequency, n (%)	Contribution proportions of community members, n (%)
Individuals (mental health consumers), including the underserved and marginalized individuals				
Strategy 1.1				
Informational support	Provision of health-related information about mental health conditions and coping strategies, including both objective and fact-based information as well as experiential insights derived from personal experiences	__[b]	519 (94.9)	• IM[c]: 331 (63.8) • HM[d]: 141 (27.2) • CM[e]: 28 (5.4) • CA[f]: 17 (3.3) • OM[g]: 2 (0.4)
Experiential informational support	Provision of information derived from personal experiences, offering firsthand insights into navigating everyday life with mental health conditions, as well as self-management and coping mechanisms, including medical, therapeutic, and health care system experiences	"I bought myself a new diary... It helped me track my moods and feelings..."	468 (90.2)	• IM: 307 (65.6) • HM: 134 (28.6) • CM: 27 (5.8)
Objective informational support	Provision of information that is impartial and fact based, including mental health–related education materials, advice, and referrals	"...who suffers from bipolar 1 and vestibular migraines, I found this research really interesting..."	113 (21.8)	• IM: 53 (46.9) • HM: 37 (32.7) • CA: 17 (15) • CM: 4 (3.5) • OM: 2 (1.8)
Self-expression support	Self-disclosing personal feelings, thoughts, daily life experiences, and challenges of living with mental health conditions, as well as self-motivational expressions	"I've never had a life. Only surviving..."	414 (75.7)	• IM: 285 (68.8) • HM: 106 (25.6) • CM: 23 (5.6)
Network support	Communicating with affiliation to the online community as well as provision of offers to gain access to its members	"...We #bipolarclub, could all collectively help to end your depression..."	221 (40.4)	• IM: 130 (58.8) • HM: 60 (27.1) • CA: 17 (7.7) • CM: 12 (5.4) • OM: 2 (0.9)
Emotional support	Provision of care, love, encouragement, and understanding expressions	"Thank you so much lovely people. Really thanks a lot. I should be dead! But I'm still alive because of you..."	126 (23)	• IM: 77 (61.1) • HM: 44 (34.9) • CM: 5 (4)
Esteem support	Provision of affirmation in ability and compliment expressions, as well as expressions of agreement on a situation and alleviating a sense of guilt about a situation	"It's great for you, me, and the community that you're open about how you're managing your disease. I appreciate the content you're sharing..."	52 (9.5)	• IM: 33 (63.5) • HM: 18 (34.6) • CM: 1 (1.9)
Tangible support	Provision of offers to help and to join activities or events that are needed to cope with the challenges of mental health conditions, including online and offline peer support groups	"Hello #bipolarclub! Come participate in our Twitter Space gathering today! We'll be discussing communication skills..."	32 (5.9)	• HM: 21 (65.6) • IM: 8 (25) • CA: 2 (6.2) • OM: 1 (3.1)
Underserved and marginalized individuals (mental health consumers)				
Strategy 1.4	Provision of diverse forms of individual-level support to address the specific needs of individuals from underserved and marginalized populations, including children and individuals from racial and ethnic minority groups (Black population)	"...@ThinkTenacity...join free Black mental health support group by Black therapists who understand us..."	21 (3.8)	• HM: 11 (52.4) • IM: 9 (42.9) • CM: 1 (4.8)

XSL•FO
RenderX

Population level and group, substrategy of Strategy 1[a], and empowerment process and subprocess	Description	Sample paraphrased tweet	Tweet frequency, n (%)	Contribution proportions of community members, n (%)
Informal carers, including family members and friends				
Strategy 1.1				
Family awareness support	Provision of information that pertains to both mental health– and family-related aspects, including personal experiences with families	"The past couple of days, I was all right, but today I woke up feeling terrible and can't bring myself to get out of bed. I hope my daughter can understand..."	39 (7.1)	• IM: 27 (69.2) • HM: 12 (30.8)
Strategy 1.3				
Friend awareness support	Provision of information that pertains to both mental health– and friend-related aspects, including personal experiences with friends	"...losing friends from having bipolar...it's hard...no one can understand..."	14 (2.6)	• IM: 9 (64.3) • HM: 3 (21.4) • CM: 1 (7.1) • CA: 1 (7.1)
Community (society)				
Strategy 1.2				
Social awareness support	Provision of information that pertains to both mental health– and societal-related aspects, addressing social misconceptions and stigmatization of mental health conditions and offering perspectives on the actual experiences and realities associated with mental health conditions	—	77 (14.1)	• HM: 36 (46.8) • IM: 33 (42.9) • CM: 5 (6.5) • CA: 3 (3.9)
Perceptive awareness support	Provision of information that addresses social misconceptions surrounding mental health conditions, including clarifications of these misconceptions, insights into the reality of mental health conditions, and personal experiences in society	"...the most thing I hope people to know about bipolar disorder...that it is not the same thing as mood swings..."	65 (84.4)	• HM: 29 (44.6) • IM: 28 (43.1) • CM: 5 (7.7) • CA: 3 (4.6)
Destigmatization support	Provision of information that addresses the destigmatization of mental health conditions in society, including antistigma expressions and personal experiences with mental health stigma	"...I discuss bipolar disorder because I want to destigmatize it... Stigma is a terrifying thing..."	18 (23.4)	• HM: 12 (66.7) • IM: 6 (33.3)

[a]The 4 substrategies of Strategy 1: Strategy 1.1 (empowering and engaging individuals and families), Strategy 1.2 (empowering and engaging communities), Strategy 1.3 (empowering and engaging informal carers), and Strategy 1.4 (reaching the underserved and marginalized).

[b]Not available.

[c]IM: individual member.

[d]HM: health care–related member.

[e]CM: crew and moderator member.

[f]CA: the community's Twitter account.

[g]OM: organizational member.

Individual-Level Empowerment Processes: Supporting Individuals (Mental Health Consumers), Including the Underserved and Marginalized Individuals

Our analysis of empowerment processes that sought to support individuals with mental health conditions in the #bipolarclub community revealed 6 distinct types of processes and 2 associated subprocesses. This individual-level category of processes could benefit a mental health consumer to be empowered. These processes also addressed the specific needs of the underserved and marginalized individuals. Among the 547 eligible tweets, the 6 identified empowerment processes in the #bipolarclub community included informational support (n=519, 94.9%), self-expression support (n=414, 75.7%), network support (n=221, 40.4%), emotional support (n=126, 23%), esteem support (n=52, 9.5%), and tangible support (n=32, 5.9%). Furthermore, 3.8% (n=21) of the tweets in these processes focused on addressing the specific needs of individuals from underserved and marginalized populations. Table 2 illustrates the results of this analysis, descriptions of processes, samples of tweets, and the contribution proportions of the community members.

XSL•FO
RenderX

The most prevalent individual-level empowerment process in the #bipolarclub community was informational support, constituting 94.9% (519/547) of the tweets. In the informational support process, 2 subprocesses emerged, each defined by the type of mental health–related information exchanged through tweets. Experiential insights from personal experiences shaped the experiential informational support subprocess (468/519, 90.2%), while objective and fact-based information formed the objective informational support subprocess (113/519, 21.8%). Tweets in both subprocesses addressed various aspects of mental health conditions and coping strategies. Experiential informational support (468/519, 90.2%) was the predominant subprocess. In this empowerment subprocess, #bipolarclub community members exchanged personal experiences relating to their daily encounters with mental health conditions, as well as self-management and coping strategies, including medical and therapeutic experiences as well as interactions with the health care system:

> *...have a space for venting...it helps...*
>
> *...so happy...first...hospitalization...I feel better...*
>
> *I have been misdiagnosed over 10 times...*
>
> *I've a Zoom appointment with my psychologist...it's better to be in person, but he only visits my area monthly...*
>
> *...if you need support contact @BipolarUK @IntlBipolar...it's helpful...*
>
> *...my insurance doesn't cover my psych meds $1600 a month...*

In the other empowerment subprocess, objective informational support (113/519, 21.8%), #bipolarclub community members exchanged tweets involving mental health–related education content, advice, and referrals:

> *... This article discusses a type of delusion called pseudocyesis, which refers to the false belief of being pregnant...*
>
> *Take a cozy day off (or three) to recharge...*
>
> *... Listening to "This Is Bipolar" podcast...*

The second individual-level empowerment process was self-expression support. It constituted 75.7% (414/547) of the tweets. Tweets involved in this process show that #bipolarclub community members used it as an outlet for venting and sharing their emotions, thoughts, and challenges of living with mental health conditions, as well as self-motivational expressions:

> *I am in an absolutely fantastic mood today!...*
>
> *I feel like I'm losing my fight with my disorder. I feel like my mind is taken a walk off the map...*
>
> *This week, I'm actively searching for a job. Just last week, I was filling out disability papers. It feels like every week is a guessing game...*
>
> *I will be okay, it's not the first time, just like the previous time...*

Network support was the third empowerment process in the mental health consumer category, comprising 40.4% (221/547) of the tweets. Members of the #bipolarclub community in this process tweeted expressions of belonging to the online community and exchanged offers to contact one another:

> *Fellow #bipolarclub...*
>
> *If you're up for a chat, please message me. 11:53 PM. I would like a chat with a Peer about right now...*

The fourth individual-level empowerment process was emotional support, representing 23% (126/547) of the tweets. This process consisted of tweets among #bipolarclub community members expressing their care, love, encouragement, and understanding to one another:

> *... To all my #bipolarclub friends and family! Wishing you joyous holidays and hope everyone enjoyed a peaceful day!...*
>
> *... Thanks #bipolarclub for making this year easier to handle...*
>
> *... I wish you a wonderful day. If you're feeling a bit low, just remember you woke up and faced life today. That's a victory!...*
>
> *...holidays...I know this time of year can be challenging...*

The fifth empowerment process, esteem support, constituted 9.5% (52/547) of the tweets. Tweets in this process revealed that #bipolarclub community members were supporting one another through expressions of positive affirmations regarding capabilities, compliments, agreement on a situation, and alleviation of any feelings of guilt about a situation:

> *Living with bipolar doesn't mean you're broken; it means you are strong and brave for battling your mind every single day...*
>
> *Yes. I've experienced... It's exhausting...*
>
> *Completely agree. I'm unable to get help... My general practitioner referrals have been denied twice, and I don't have the funds for private help...*
>
> *... If you're facing depression, know that you're not alone, millions of people worldwide understand what you're going through...*

The least prominent empowerment process identified in the individual-level category was tangible support, comprising 5.9% (32/547) of the tweets. These tweets demonstrate that the #bipolarclub community was used by its members to offer help to one another and announce activities or events crucial for coping with the challenges of mental health conditions, such as participation in peer support groups:

> *...I'm here if needed...*
>
> *Hello #bipolarclub! Join our Twitter Space this Sunday as we delve into a discussion about establishing healthy boundaries in your relationships...*

Our analysis also revealed that 3.8% (21/547) of the tweets in the identified empowerment processes at the individual level addressed the specific needs of individuals from underserved and marginalized populations. This formed a distinct subcategory of individual-level empowerment in the #bipolarclub community, as illustrated in Figure 2. The content

XSL•FO
RenderX

of tweets in this subcategory was particularly relevant to 2 groups, children and individuals from racial and ethnic minority groups (Black population):

> *The ACEs (adverse childhood experiences) score is important... Watch this video about childhood trauma!...*
>
> *...free access to Black therapists...a group support via zoom...meet real Black therapists... Date of event...*

As shown in Table 2, #bipolarclub community members from all classifications contributed to the individual-level empowerment processes, with proportions reflecting their representation ratios and classifications within the community. Notably, among the 6 identified empowerment processes, health care–related members were the predominant contributors of tangible support content, in contrast to the other 5 processes with the largest proportion of posts by individual members.

Furthermore, health care–related members made their most substantial contribution to the empowerment process of tangible support, with 66% (21/32) of the tweets. By contrast, both individual members and crew and moderator members demonstrated their highest contribution ratios in self-expression support, with 68.8% (285/414) and 5.6% (23/414) of the tweets, respectively. The community's account had its most significant contribution rate in the network support process (17/221, 7.7%), whereas the organizational member had it in the tangible support process (1/32, 3%).

Informal Carer Empowerment Processes: Supporting Informal Carers, Including Family Members and Friends

The examination of empowerment processes dedicated to supporting informal carers in the #bipolarclub community unveiled 2 processes, one tailored to family members and another to friends. We identified these 2 processes as family awareness support and friend awareness support. Of the 547 tweets, family awareness support comprised 7.1% (n=39) of the tweets, and friend awareness support included 2.6% (n=14) of the tweets. Table 2 shows descriptions of these processes, tweet samples, and the contribution proportions of the community members.

The most prominent empowerment process identified in the informal carer category was family awareness support, constituting 7.1% (39/547) of the tweets. Tweets in this process were composed of information that could support the engagement of families in mental health care, raise their awareness, and allow them to understand their crucial role in the well-being of individuals in their families with mental health conditions. The content of these tweets was related to both mental health conditions and family-related aspects. It involved mental health education materials relevant to families as well as personal experiences shared by #bipolarclub community members. These experiences covered interactions with families in general and specific family members, including parents, children, siblings, and partners:

> *... Some reading from the Bipolar Disorder Survival Guide "What You and Your Family Need to Know"...*
>
> *It really hurts being the bipolar child who never gets invited to family gatherings, again...*
>
> *I find it not easy to discuss my condition with others...especially...my parents...for fear of being blamed; that's why I can't be positive, there're lots of people have rougher lives than you, so on...*
>
> *My brother led me to a really dark place that I haven't been in for a very long time, he doesn't realize how much his words hurt all the time...*
>
> *I was feeling scattered mentally yesterday, and I was terrible to my partner. I'm so relieved today, and he seems to have forgiven me...*

In the informal carer category, the other empowerment process that we identified was friend awareness support (14/547, 2.6%). This process included tweets containing valuable information that could help involve friends in mental health care, enhance their awareness, and help them recognize their significant role in supporting and understanding their friends with mental health conditions. The content in these tweets was relevant to both mental health conditions and friend-related perspectives. It mainly consisted of personal experiences shared by #bipolarclub community members in relation to their friends:

> *My friend gifted me this wonderful book...it is really helpful during challenging days of depression…*
>
> *...losing friends from having bipolar...it's hard...no one can understand...*

From the overall contributions of #bipolarclub community members to the 2 empowerment processes of the informal carer category, individual members contributed the most in both processes, family awareness support and friend awareness support. However, the contribution ratio of individual members in family awareness support (27/39, 69%) was higher than in friend awareness support (9/14, 64%). Furthermore, individual members and health care–related members were the only contributors to family awareness support. In addition, health care–related members made their strongest contribution in this category of processes in family awareness support (12/39, 31%). In contrast, the organizational member did not contribute to any process in the informal carer category, while crew and moderator members (1/14, 7%) and the community's account (1/14, 7%) contributed solely to the friend awareness support process at the same percentage.

Society-Level Empowerment Processes: Supporting the Community (Society)

Our analysis of empowerment processes in the #bipolarclub community, aimed at supporting society, indicated that this online community involved 1 overarching process and 2 associated subprocesses. We identified the overarching process as social awareness support and the associated subprocesses as perceptive awareness support and destigmatization support. The findings revealed that of the 547 eligible tweets, social awareness support constituted 14.1% (n=77), wherein perceptive awareness support and destigmatization support accounted for 84% (65/77) and 23% (18/77), respectively. Table 2 describes the processes, samples of tweets, and the contribution proportions of the community members in each process.

XSL•FO
RenderX

The empowerment process of social awareness support (77/547, 14.1%) and its subprocesses, perceptive awareness support (65/77, 84%) and destigmatization support (18/77, 23%), emerged from tweets involving information that could inform and educate the public about mental health conditions and raise their awareness, as well as support to counter stigmatizing attitudes and engage society in mental health care. The provided information pertained to both mental health conditions and aspects associated with society. In addition, notably, tweets in the society-level empowerment processes often featured relevant hashtags such as #MentalHealthAwareness, #BipolarAwareness, #RaiseAwareness, #DepressionIsReal, #MentalHealthMatters, and #BreakTheStigma. This demonstrated that #bipolarclub community members were using their collective voice to reach the public.

The most prevalent subprocess in the social awareness support process was perceptive awareness support (65/77, 84%). This subprocess was composed of tweets containing explanations of misunderstandings regarding mental health conditions, glimpses into the actuality of mental health conditions, and lived experiences of mental health conditions in society:

> *... Stop saying it's "their" FAULT to mental patients! The meds disrupt metabolism!...*
>
> *I've been physically fit with a six-pack but still battled major depression and panic attacks...looking good on the outside doesn't necessarily mean you're mentally "healthy" on the inside...*
>
> *I don't like when people say they're "so bipolar today" just because they can't make up their minds...dear, it doesn't work that way...*
>
> *I don't want anyone passing judgment on me during my episodes... Mental health challenges come in various ways. May we all learn to support each other with love, understanding, and compassion...*
>
> *It seems I have got tardive dyskinesia from my antipsychotics...so embarrassed my twitching is pointed out by my coworker...*

The other subprocess, destigmatization support (18/77, 23%), included expressions aimed at destigmatizing mental health conditions as well as personal experiences with mental health stigma:

> *It isn't acceptable to use disrespectful, stigmatizing words describing any situation...bipolar is a real mental condition...*
>
> *...I really want to appear with my real identity to #bipolarclub. But I am scared of the potential of someone from real life finding me. I wouldn't be able to survive the shame & humiliation again...*

While the major contributors in the empowerment processes of the individual-level category and the informal carer category were individual members, health care–related members were the major contributors in the society-level category. Health care–related members contributed to the social awareness support process by 47% (36/77) of the tweets, while individual members participated by 43% (33/77). Furthermore, health care–related members and individual members were the only contributors to the subprocess destigmatization support. In alignment with the informal carer category, the organizational member did not contribute to the society-level category. Crew and moderator members and the community's account participated in the social awareness support process by 6% (5/77) and 4% (3/77) of the tweets, respectively.

Discussion

Principal Findings

Overview

In this study, we have investigated an OMHC on Twitter called #bipolarclub by conducting a qualitative content analysis of tweets containing this hashtag circulated between December 19, 2022, and January 15, 2023. Through this analysis, we provided profound insights into three key aspects: (1) the members shaping the online community; (2) the various types of empowerment processes in it, aligned with the population levels and groups outlined in Strategy 1 of the WHO's IPCHS framework; and (3) the contributions made by its members to these empowerment processes, elucidating their tendencies of involvement.

Overall, we have demonstrated that OMHCs such as #biplolarclub involve health professional members who have dual roles as both professionals and individuals coping with mental health conditions, which highlights the valuable insights and expertise they contribute to the community. We have also revealed that the #bipolarclub community includes empowerment processes catering to all the population levels and groups outlined in Strategy 1 of the WHO's IPCHS framework. The online community contains processes for mental health consumers, including those from the underserved and marginalized populations, their informal carers (families and friends), and society at large. This finding indicates that an OMHC on Twitter holds promise for empowering diverse populations and supporting global empowerment objectives. Furthermore, our analysis has unveiled distinct contribution tendencies among the members to the empowerment processes in the #bipolarclub community. These tendencies showed diverse patterns of involvement in these processes. In the following sections, we thoroughly discuss our findings and provide suggestions on how these findings can be leveraged to promote consumer empowerment in OMHCs and mental health care.

OMHC Members

The #bipolarclub community is formed of 5 primary types of members: individual member, health care–related member, crew and moderator member, organizational member, and the community's account. This categorization is generally aligned with a previous study that broadly discussed the identities of those who use OMHCs for peer support and the role of moderators in managing the communities [5]. However, our categorization provides deeper insights into the members who form OMHCs on Twitter. All #bipolarclub community members identified with a mental health condition, including practitioner members with health care–related affiliations. However, they used the community as consumers and peers, not for professional

XSL•FO
RenderX

service purposes. This finding indicates that the #bipolarclub community involves a unique type of peers that could be identified as "professional peers." These peers share similar features with those noted in previous studies as formal peers [59] and peer specialists [60]. This peer type in the community provides health-related information based not only on a similar personal experience but also on a professional one. They also serve as linkages with the health system [60].

Furthermore, these peers add value to the #bipolarclub community, as they can bring reliability and credibility to the health-related information provided therein. Compared with a previous study on Reddit, which reported that mental health professionals joined OMHCs solely to offer expert assistance [26], our study shows that professionals may also seek support for themselves as individuals affected by mental health conditions. Thus, we have introduced the concept of "professional peers" as a novel member category in OMHCs on social media that has not been previously documented.

Consumer Empowerment Processes

The identified empowerment processes in the #bipolarclub community imply that an OMHC on Twitter has the potential to achieve the empowerment objectives of Strategy 1 of the WHO's IPCHS framework, along with all its substrategies. This includes empowering mental health consumers as well as addressing the specific needs of the underserved and marginalized individuals, their informal carers (families and friends), and society. The #bipolarclub community comprises 3 categories of empowerment processes, each supporting a specific population level and/or group: the individual-level category (mental health consumers including those from underserved and marginalized populations), the informal carer category (families and friends), and the society-level category.

Regarding the individual-level empowerment processes, the #bipolarclub community involved informational support and its subprocesses, experiential informational support and objective informational support, self-expression support, network support, emotional support, esteem support, and tangible support. Although previous studies indicated that informational and emotional support are the predominant social support types in OHCs [57,61] and are the 2 most common types sought by mental health consumers in online communities [5], our results indicate that the most exchanged social support types in the #bipolarclub community were informational and network support. This finding may be attributed to the networking mechanisms unique to Twitter, such as #hashtags, @username mentions, retweeting, and the following function [40], enabling #bipolarclub community members to engage in more interconnected communication compared with those connected in OHCs on other social media platforms [62]. A prior study that analyzed Twitter hashtags relating to bipolar disorder highlighted emotional support as the predominant social support type [63]. By considering a specific community and the types of members involved therein, we point toward network support as a key aspect. In addition, in a recent study that delved into empowerment processes, "finding recognition" was identified as a process in OMHCs [15]. This process aligns with the network support process in our study, emphasizing the idea that being connected in OMHCs with fellow community members who share similar illnesses and potentially have comparable life experiences can alleviate feelings of loneliness.

Our analysis also unveiled that the experiential informational support subprocess constitutes the majority of its main empowerment process informational support in the #bipolarclub community. This finding suggests that members could primarily use the community to learn from one another their personal experiences of dealing with mental health challenges, which aligns with one of the 5 motivations for establishing the #bipolarclub community (learning from one another) [47]. Although these firsthand experiences can help them in their management and coping strategies [61], they may also lead them to overlook professional guidance. In addition, #bipolarclub community members actively shared information and provided feedback on their experiences with the health care system. This finding indicates that the community could serve a dual function, helping its members to navigate the health care system more effectively [9], as well as offering valuable feedback to health care providers, thereby providing opportunities to enhance health care services [11]. This is particularly relevant in the #bipolarclub community, in which some members have dual roles as both mental health consumers and practitioners.

Our analysis also demonstrated that self-expression support emerged as a frequently practiced empowerment process among #bipolarclub community members. This implies that self-expression may represent a significant need among members, which could be met in the OMHC. Our finding supports previous assertions that mental health consumers commonly use OMHCs on social media [5,9] and Twitter [4] to self-express and vent, reaffirming the notion that OMHCs offer a conducive environment for expressing one's true self owing to the anonymity afforded by social media [9]. The findings also showed that #bipolarclub community members were expressing their current emotions, thoughts, and challenges instantly, indicating that the real-time nature of Twitter could be a valuable tool for immediate release and a sense of catharsis for those struggling with mental health conditions.

Although the #bipolarclub community generally serves underserved and marginalized individuals (mental health consumers) [64], the content of its tweets addresses the needs of individuals who belong to 2 additional underserved and marginalized populations: children and racial and ethnic minority groups (Black population) [13,55]. The fact that Twitter can support consumers with dual underserved and marginalized status toward empowerment is not novel. An earlier study highlighted the creation of the hashtag #YouGoodMan on Twitter, which is specifically tailored for Black men to share their experiences with mental health conditions, exchange support, and navigate challenges stemming from cultural and social factors in the Black community [8].

Regarding the informal carer empowerment processes, the existence of family awareness support and friend awareness support processes in the #bipolarclub community signifies its potential to empower family members and friends of mental health consumers. The dissemination of mental health–related

XSL•FO
RenderX

information linked to families and friends could serve as a valuable resource for enhancing their understanding of the challenges faced by those they are caring for [65,66] and their significant role in their well-being, which could facilitate their active engagement in mental health care practices. In addition, the public nature of Twitter allows users to access a diverse range of perspectives and information. With these findings, we underscore the pivotal role of firsthand information in OMHCs for informal carers, which is sourced directly from individuals who have experienced the situation themselves rather than from other informal carers. This aspect of firsthand information has been overlooked in the existing literature on informal carer empowerment in OMHCs.

Regarding the society-level empowerment processes, the presence of social awareness support and its subprocesses, perceptive awareness support and destigmatization support, highlights the capability of the #bipolarclub community to empower society. The community's potential role as a facilitator of societal education is apparent through clarifications about social misconceptions regarding mental health conditions, personal experiences in the broader societal context, and antistigma posts. This content could enhance societal understanding and awareness of mental health conditions and could foster societal engagement in mental health care. The destigmatization support subprocess in the #bipolarclub community, aligned with the community's core objective of fighting stigma [47], could address the stigmatizing challenges encountered by those dealing with mental health conditions in their daily lives in society [67]. Using the #bipolarclub community to combat stigma resonates with previous studies on Twitter [4,68], emphasizing Twitter's key role in supporting mental health antistigma efforts. In addition, most tweets in the #bipolarclub community centered on raising awareness rather than directly addressing stigmatization. This finding indicates that the community could be adopting a proactive approach to building understanding and empathy with a more inclusive discourse around mental health issues.

Contributions of OMHC Members to Consumer Empowerment Processes

When examining members' contributions to the 3 categories of empowerment processes in the #bipolarclub community, distinct patterns emerged regarding their involvement tendencies, reflecting their identities. Health care–related members were inclined to support societal awareness, mirroring their role as influencers and educators. In contrast, individual members were actively engaged in self-expression and significantly contributed to family awareness, reflecting the importance of personal expression and familial support in this group. In addition, the peer support foundation (organizational member) focused solely on supporting mental health consumers through tangible support, aligning with the core mission of such foundations in providing practical assistance to individuals in need. Akin to individual members, crew and moderator members used the #bipolarclub community as an outlet for personal expression. This tendency conforms with their primary role as individuals coping with mental health conditions, preceding their roles as crew and moderator members. The community's account stood out for its active role in connecting with members and providing network support, aligning with its identity as a facilitator of community cohesion. These nuanced tendencies underscore the collaborative efforts of different member types, each bringing unique perspectives to the community's empowerment processes. Notably, our findings differ from those of a study that examined tweets during Mental Health Awareness Week [16], which found similar contribution tendencies to the discourse among diverse users. However, our analysis focused on an OMHC's tweets, which are not based on a trendy mental health–related hashtag, and we used different criteria for analyzing community members and tweets. Despite differing research goals, these findings suggest that Twitter users' contribution tendencies in engaging in mental health discussions may vary depending on the context of use.

Implications

Our study has extended the research on the role of OMHCs in empowering people in significant ways. The key findings are as follows: (1) we identified the diverse members within an OMHC, clarifying their tendencies for engaging in empowerment processes; (2) while health care practitioners use OMHCs to support consumers, we revealed their multifaceted engagement to fulfill various roles simultaneously, acting as mental health consumers, peers, and providers of health information, sometimes disclosing their professional identities; and (3) we proved that an OMHC not only could facilitate empowerment processes for mental health consumers but could also extend their impacts to individuals from underserved and marginalized populations, informal carers (family and friends), and society at large, aligning with the WHO's empowerment agenda (Strategy 1 of the IPCHS framework). In addition, we have proposed a structured framework for classifying the empowerment processes within OMHCs based on Strategy 1 of the WHO's IPCHS framework, which embraces individual-level, informal carer, and society-level processes.

The study findings hold significant implications for various health care stakeholders, such as national and international health care organizations, health care professionals, and OMHC moderators. For instance, the WHO and other health care institutions can benefit by learning about the potential of OMHCs for empowering not only mental health consumers but also their families and friends as well as society. Thus, they could consider integrating these communities into standards, acknowledge their important roles in guidelines, and effectively support their development as an element in holistic approaches to mental health care. Our findings can also guide health care organizations and professionals to tailor their interventions and outreach strategies using these communities, ensuring more effective and targeted approaches for diverse population levels and groups to foster empowerment in mental health. Furthermore, moderators of OMHCs can leverage insights from our findings to provide dedicated support for diverse empowerment processes. We speculate that they could explore various uses for initiatives, such as motivating mental health consumers to tell their stories (eg, "tell-your-story" week) to express themselves and raise awareness among others.

XSL•FO
RenderX

Study Strengths

This study had several strengths. First, our examination of the #bipolarclub community forms part of a larger study to investigate the concept of consumer empowerment in the community following the netnography methodology that focuses on studying online communities' behaviors [69]. Therefore, the findings of our analysis were based on in-depth immersion in the community. Second, by investigating empowerment processes supporting various population levels and groups, we offer a comprehensive and nuanced view of the empowerment concept within the community. Third, examining empowerment processes aligned with Strategy 1 of the WHO's IPCHS framework has yielded profound insights into the types of processes that adhere to global standards within OMHCs. Often overlooked in prior studies, this aspect contributes a valuable perspective on the empowerment processes within OMHCs to the existing body of research. Thus, our proposed structured framework of empowerment processes' classifications within OMHCs, based on Strategy 1, can serve as a novel foundation for future investigations of various OMHCs.

Study Limitations and Future Research

A possible limitation of this study is the classification of #bipolarclub community members based on their personal bios and the tweets posted on their profiles. This content may not necessarily represent their real-life identity and activities, potentially leading to inaccuracies. However, our passive data collection approach provides real and bias-free insights into how empowerment processes are conducted in the online community [70]. In addition, this approach precluded the examination of a member type in online communities, known as passive members and "lurkers" [71], who may be using the community. Further research could build on our study and explore this member type.

Another limitation of using a passive investigation is the challenge of demonstrating that the #bipolarclub community includes members beyond those who directly experience mental health conditions, such as family members, friends, and the public. However, since our analysis did not include passive members, it is important to acknowledge that merely lurking constitutes a form of engagement in empowerment processes. This implies that these populations still have considerable potential for being involved in the online community. In addition, Twitter's open-access nature allows not only #bipolarclub community members but also other Twitter users to engage in these processes without restrictions, such as account following [39] or registration requirements such as private groups on Facebook. Furthermore, tweets using the #bipolarclub hashtag were often blended with other hashtags, expanding the community's reach to a wider audience of Twitter users outside the scope of the community's members [40], advertently and inadvertently engaging them in the empowerment processes. Future research endeavors could leverage our findings by examining the passive involvement of various population levels and groups within the #bipolarclub community.

We have also primarily focused on one aspect of the consumer empowerment concept, the processes (empowering activities), without considering its other aspect, the outcomes (states of being empowered), which may have limited our findings to provide a comprehensive understanding of the empowerment phenomenon in the online community. However, our study provides in-depth insights into these processes. Future research could investigate the empowerment outcomes in alignment with the empowerment processes we have identified.

Conclusions

Our analysis of members and empowerment processes in the #bipolarclub community highlights the capability of Twitter-based OMHCs to empower mental health consumers, including those from underserved and marginalized populations, along with their families and friends as well as society. Our study demonstrates the ability of a Twitter-based OMHC to facilitate empowerment processes for diverse population levels and groups aligning with the WHO's empowerment agenda (Strategy 1 of the IPCHS framework), highlighting the potential advantages of using Twitter for such empowerment objectives. These findings also acknowledge the relevance of Twitter-based OMHCs in advancing global empowerment goals. As the use of OMHCs and Twitter continues to rapidly grow, exploring their potential holds promise for informing various health care stakeholders. This is particularly relevant for health care organizations, professionals, and OMHC moderators, as these insights could pave the way for developing consumer-oriented services and empowerment programs for different population levels and groups.

Acknowledgments

HA, a doctoral researcher, has been awarded a Swiss Government Excellence Scholarship from the Federal Commission for Scholarships for Foreign Students (FCS). We would like to acknowledge the FCS for supporting this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample paraphrased personal descriptions (bios) and posted tweets on Twitter profiles of the #bipolarclub community's members.
[PDF File (Adobe PDF File), 123 KB - jmir_v26i1e55965_app1.pdf]

References

XSL•FO
RenderX

1. Otokiti A, Williams KS, Warsame L. Impact of digital divide on the adoption of online patient portals for self-motivated patients. Healthc Inform Res 2020 Jul;26(3):220-228 [FREE Full text] [doi: 10.4258/hir.2020.26.3.220] [Medline: 32819040]
2. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
3. Capoot A. Elon Musk rebrands Twitter to 'X,' replaces iconic bird logo. CNBC. 2023 Jul 24. URL: https://www.cnbc.com/2023/07/24/elon-musk-rebrands-twitter-to-x-replaces-iconic-bird-logo.html [accessed 2024-07-09]
4. Berry N, Lobban F, Belousov M, Emsley R, Nenadic G, Bucci S. #WhyWeTweetMH: understanding why people use Twitter to discuss mental health problems. J Med Internet Res 2017 Apr 05;19(4):e107 [FREE Full text] [doi: 10.2196/jmir.6173] [Medline: 28381392]
5. Rayland A, Andrews J. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. JMIR Ment Health 2023 Feb 28;10:e41855 [FREE Full text] [doi: 10.2196/41855] [Medline: 36853738]
6. Shalaby RA, Agyapong VI. Peer support in mental health: literature review. JMIR Ment Health 2020 Jun 09;7(6):e15572 [FREE Full text] [doi: 10.2196/15572] [Medline: 32357127]
7. Sit M, Elliott SA, Wright KS, Scott SD, Hartling L. Youth mental health help-seeking information needs and experiences: a thematic analysis of Reddit posts. Youth Soc 2024 Jan;56(1):24-41. [doi: 10.1177/0044118x221129642]
8. Francis DB. "Twitter is really therapeutic at times": examination of black men's Twitter conversations following hip-hop artist Kid Cudi's depression disclosure. Health Commun 2021 Apr 06;36(4):448-456. [doi: 10.1080/10410236.2019.1700436] [Medline: 33586529]
9. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
10. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. Comput Human Behav 2018 Jan;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
11. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 09;23(2):e19910 [FREE Full text] [doi: 10.2196/19910] [Medline: 33560233]
12. Fullwood C, Chadwick D, Keep M, Attrill-Smith A, Asbury T, Kirwan G. Lurking towards empowerment: explaining propensity to engage with online health support groups and its association with positive outcomes. Comput Human Behav 2019 Jan;90:131-140. [doi: 10.1016/j.chb.2018.08.037]
13. Framework on integrated, people-centred health services: report by the Secretariat. World Health Organization (WHO). 2016. URL: https://apps.who.int/iris/handle/10665/252698 [accessed 2024-07-09]
14. Prescott J, Rathbone AL, Hanley T. Online mental health communities, self-efficacy and transition to further support. Ment Health Rev J 2020 Sep 28;25(4):329-344. [doi: 10.1108/mhrj-12-2019-0048]
15. Liu J, Wang J. Users' intention to continue using online mental health communities: empowerment theory perspective. Int J Environ Res Public Health 2021 Sep 07;18(18):9427 [FREE Full text] [doi: 10.3390/ijerph18189427] [Medline: 34574361]
16. Makita M, Mas-Bleda A, Morris S, Thelwall M. Mental health discourses on Twitter during Mental Health Awareness Week. Issues Ment Health Nurs 2021;42(5):437-450. [doi: 10.1080/01612840.2020.1814914] [Medline: 32926796]
17. Bosi G, De Santis M, Hervás C, Weinman A, Bottarelli V. Patient empowerment of people living with rare diseases. Its contribution to sustainable and resilient healthcare systems. Ann Ist Super Sanita 2019;55(3):283-291 [FREE Full text] [doi: 10.4415/ANN_19_03_15] [Medline: 31553324]
18. Faghihnasiri K, Alves HM, Soares AM. Value co-creation activities role in patient well-being in online healthcare communities. In: Proceedings of the Selected Papers from XXI International Congress on Public and Nonprofit Marketing. 2022 Presented at: IAPNM '22; July 7-8, 2022; Braga, Portugal p. 149-175 URL: https://link.springer.com/chapter/10.1007/978-3-031-29020-6_8 [doi: 10.1007/978-3-031-29020-6_8]
19. Garattini L, Padula A. Patient empowerment in Europe: is no further research needed? Eur J Health Econ 2018 Jun 08;19(5):637-640. [doi: 10.1007/s10198-018-0965-3] [Medline: 29520521]
20. Comprehensive mental health action plan 2013–2030. World Health Organization (WHO). 2021. URL: https://www.who.int/publications/i/item/9789240031029 [accessed 2024-07-10]
21. World mental health report: transforming mental health for all. World Health Organization (WHO). 2022. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-07-10]
22. Guidance on community mental health services: promoting person-centred and rights-based approaches. World Health Organization (WHO). 2021. URL: https://www.who.int/publications/i/item/9789240025707 [accessed 2024-07-08]
23. Rezaei Aghdam A, Watson J, Cliff C, Miah SJ. Improving the theoretical understanding toward patient-driven health care innovation through online value cocreation: systematic review. J Med Internet Res 2020 Apr 24;22(4):e16324 [FREE Full text] [doi: 10.2196/16324] [Medline: 32329736]
24. Ruiz SL, Stadtlander L. Social media as support for partners of veterans with posttraumatic stress disorder. J Soc Behav Health Sci 2015 May 28;9(1):1-18. [doi: 10.5590/JSBHS.2015.09.1.01]

XSL•FO
RenderX

25. Siegel-Ramsay JE, Sharp SJ, Ulack CJ, Chiang KS, Lanza di Scalea T, O'Hara S, et al. Experiences that matter in bipolar disorder: a qualitative study using the capability, comfort and calm framework. Int J Bipolar Disord 2023 Apr 20;11(1):13 [FREE Full text] [doi: 10.1186/s40345-023-00293-9] [Medline: 37079153]
26. Saha K, Ernala SK, Dutta S, Sharma E, De Choudhury M. Understanding moderation in online mental health communities. In: Proceedings of the 2020 Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference. 2020 Presented at: HCII '20; July 19-24, 2020; Copenhagen, Denmark p. 87-107 URL: https://tinyurl.com/36jvdb5b [doi: 10.1007/978-3-030-49576-3_7]
27. Reavley NJ, Pilkington PD. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. PeerJ 2014 Oct 28;2:e647 [FREE Full text] [doi: 10.7717/peerj.647] [Medline: 25374786]
28. Jagfeld G, Lobban F, Humphreys C, Rayson P, Jones SH. How people with a bipolar disorder diagnosis talk about personal recovery in peer online support forums: corpus framework analysis using the POETIC framework. JMIR Med Inform 2023 Nov 08;11:e46544 [FREE Full text] [doi: 10.2196/46544] [Medline: 37962520]
29. Migliorini C, Barrington N, O'Hanlon B, O'Loughlin G, Harvey C. The help-seeking experiences of family and friends who support young people with mental health issues: a qualitative study. Qual Health Res 2023 Feb 17;33(3):191-203. [doi: 10.1177/10497323221147130] [Medline: 36650446]
30. Torous J, Benson NM, Myrick K, Eysenbach G. Focusing on digital research priorities for advancing the access and quality of mental health. JMIR Ment Health 2023 Apr 24;10:e47898 [FREE Full text] [doi: 10.2196/47898] [Medline: 37093624]
31. Sharma S, Khadka A. Role of empowerment and sense of community on online social health support group. Inform Technol Peopl 2019 Dec 02;32(6):1564-1590. [doi: 10.1108/itp-09-2018-0410]
32. Da Moura Semedo C, Bath PA, Zhang Z. Social support in a diabetes online community: mixed methods content analysis. JMIR Diabetes 2023 Jan 06;8:e41320 [FREE Full text] [doi: 10.2196/41320] [Medline: 36607714]
33. Verberne S, Batenburg A, Sanders R, van Eenbergen M, Das E, Lambooij MS. Analyzing empowerment processes among cancer patients in an online community: a text mining approach. JMIR Cancer 2019 Apr 17;5(1):e9887 [FREE Full text] [doi: 10.2196/cancer.9887] [Medline: 30994468]
34. Ukoha C, Stranieri A. Criteria to measure social media value in health care settings: narrative literature review. J Med Internet Res 2019 Dec 16;21(12):e14684 [FREE Full text] [doi: 10.2196/14684] [Medline: 31841114]
35. Sharma E, De Choudhury M. Mental health support and its relationship to linguistic accommodation in online communities. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC p. 1-13 URL: https://dl.acm.org/doi/10.1145/3173574.3174215 [doi: 10.1145/3173574.3174215]
36. Cerasa A, Gaggioli A, Marino F, Riva G, Pioggia G. The promise of the metaverse in mental health: the new era of MEDverse. Heliyon 2022 Nov;8(11):e11762 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e11762] [Medline: 36458297]
37. de Azevedo Cardoso T, Kochhar S, Torous J, Morton E. Digital tools to facilitate the detection and treatment of bipolar disorder: key developments and future directions. JMIR Ment Health 2024 Apr 01;11:e58631 [FREE Full text] [doi: 10.2196/58631] [Medline: 38557724]
38. Gupta P, Khan A, Kumar A. Social media use by patients in health care: a scoping review. Int J Healthc Manag 2022;15(2):121-131. [doi: 10.1080/20479700.2020.1860563]
39. Ju R, Jia M, Cheng J. Promoting mental health on social media: a content analysis of organizational tweets. Health Commun 2023;38(8):1540-1549. [doi: 10.1080/10410236.2021.2018834] [Medline: 34955059]
40. Bernardi R. How do online communities of patients aggregate on Twitter? An affordance perspective. In: Proceedings of the 37th International Conference on Information Systems. 2016 Presented at: ICIS '16; December 11-14, 2016; Dublin, Ireland p. 2993-3011 URL: https://aisel.aisnet.org/icis2016/ISHealthcare/Presentations/18/
41. Tambling RR, D'Aniello-Heyda C, Hynes KC. Manifestations of depression on social media: a content analysis of Twitter posts. J Technol Behav Sci 2023 Jun 23;9:252-261. [doi: 10.1007/s41347-023-00327-0]
42. Lachmar EM, Wittenborn AK, Bogen KW, McCauley HL. #MyDepressionLooksLike: examining public discourse about depression on Twitter. JMIR Ment Health 2017 Oct 18;4(4):e43 [FREE Full text] [doi: 10.2196/mental.8141] [Medline: 29046270]
43. Brown G, Rathbone AL, Prescott J. Social media use for supporting mental health (SMILE). Ment Health Rev J 2021 Jul 08;26(3):279-297. [doi: 10.1108/mhrj-10-2020-0079]
44. Canary HE, Wellman N, Martinez LS. COVID-19, genetics, and risk: content analysis of Facebook posts early in the coronavirus pandemic. Health Commun 2023;38(8):1654-1665. [doi: 10.1080/10410236.2022.2027639] [Medline: 35067113]
45. Kozinets RV. The field behind the screen: using netnography for marketing research in online communities. J Mark Res 2002 Feb;39(1):61-72. [doi: 10.1509/jmkr.39.1.61.18935]
46. Salzmann-Erikson M, Eriksson H. LiLEDDA: a six-step forum-based netnographic research method for nursing science. Aporia 2012;4(4):6-18 [FREE Full text]
47. About Bipolar Club. Bipolar Club community. Archived at https://web.archive.org/web/20230514175728/https://bcdx.org/about/ on May 14, 2023 URL: https://bcdx.org/about/ [accessed 2023-12-20]
48. @BipolarClubDx Twitter account. Twitter (X). URL: https://twitter.com/BipolarClubDx [accessed 2024-07-09]

XSL•FO
RenderX

49. The search API. Twitter (X) Developer Platform. URL: https://developer.twitter.com/en/docs/api-reference-index [accessed 2024-05-01]
50. Bruckman A. Studying the amateur artist: a perspective on disguising data collected in human subjects research on the internet. Ethics Inf Technol 2002 Sep;4:217-231. [doi: 10.1023/A:1021316409277]
51. Whitehead LC. Methodological and ethical issues in internet-mediated research in the field of health: an integrated review of the literature. Soc Sci Med 2007 Aug;65(4):782-791. [doi: 10.1016/j.socscimed.2007.03.005] [Medline: 17512105]
52. Internet research: ethical guidelines 3.0. Association of Internet Researchers (AoIR). 2019. URL: https://aoir.org/ethics/ [accessed 2024-04-09]
53. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. 4th edition. Thousand Oaks, CA: Sage Publications; 2014.
54. Cheng TY, Liu L, Woo BK. Analyzing Twitter as a platform for Alzheimer-related dementia awareness: thematic analyses of tweets. JMIR Aging 2018 Dec 10;1(2):e11542 [FREE Full text] [doi: 10.2196/11542] [Medline: 31518232]
55. Godinho MA, Jonnagaddala J, Gudi N, Islam R, Narasimhan P, Liaw ST. mHealth for integrated people-centred health services in the Western Pacific: a systematic review. Int J Med Inform 2020 Oct;142:104259. [doi: 10.1016/j.ijmedinf.2020.104259] [Medline: 32858339]
56. Cutrona CE, Suhr JA. Controllability of stressful events and satisfaction with spouse support behaviors. Commun Res 1992 Apr 02;19(2):154-174. [doi: 10.1177/009365092019002002]
57. Chen L, Baird A, Straub D. Fostering participant health knowledge and attitudes: an econometric study of a chronic disease-focused online health community. J Manag Inf Syst 2019;36(1):194-229. [doi: 10.1080/07421222.2018.1550547]
58. About X Spaces. Twitter (X) Help Center. URL: https://help.twitter.com/en/using-x/spaces [accessed 2024-07-05]
59. Charles A, Nixdorf R, Ibrahim N, Meir LG, Mpango RS, Ngakongwa F, et al. Initial training for mental health peer support workers: systematized review and international Delphi consultation. JMIR Ment Health 2021 May 27;8(5):e25528 [FREE Full text] [doi: 10.2196/25528] [Medline: 34042603]
60. Cabassa LJ, Camacho D, Vélez-Grau CM, Stefancic A. Peer-based health interventions for people with serious mental illness: a systematic literature review. J Psychiatr Res 2017 Jan;84:80-89. [doi: 10.1016/j.jpsychires.2016.09.021] [Medline: 27701013]
61. Lin X, Kishore R. Social media-enabled healthcare: a conceptual model of social media affordances, online social support, and health behaviors and outcomes. Technol Forecast Soc Change 2021 May;166:120574. [doi: 10.1016/j.techfore.2021.120574]
62. Phua J, Jin SV, Kim J. Uses and gratifications of social networking sites for bridging and bonding social capital: a comparison of Facebook, Twitter, Instagram, and Snapchat. Comput Human Behav 2017 Jul;72:115-122. [doi: 10.1016/j.chb.2017.02.041]
63. Budenz A, Klassen A, Purtle J, Yom Tov E, Yudell M, Massey P. Mental illness and bipolar disorder on Twitter: implications for stigma and social support. J Ment Health 2020;29(2):191-199. [doi: 10.1080/09638237.2019.1677878] [Medline: 31694433]
64. Barron GC, Laryea-Adjei G, Vike-Freiberga V, Abubakar I, Dakkak H, Devakumar D, Lancet Commission on COVID-19: Task Force on Humanitarian Relief, Social Protection and Vulnerable Groups. Safeguarding people living in vulnerable conditions in the COVID-19 era through universal health coverage and social protection. Lancet Public Health 2022 Jan;7(1):e86-e92 [FREE Full text] [doi: 10.1016/S2468-2667(21)00235-8] [Medline: 34906331]
65. Reifegerste D, Blech S, Dechant P. Understanding information seeking about the health of others: applying the comprehensive model of information seeking to proxy online health information seeking. J Health Commun 2020 Feb 02;25(2):126-135. [doi: 10.1080/10810730.2020.1716280] [Medline: 32009552]
66. Ferrell EL, Russin SE, Hardy RM. Informal caregiving experiences in posttraumatic stress disorder: A content analysis of an online community. J Community Psychol 2019 May;47(4):757-771. [doi: 10.1002/jcop.22151] [Medline: 30592051]
67. Robinson P, Turk D, Jilka S, Cella M. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Soc Psychiatry Psychiatr Epidemiol 2019 Jan;54(1):51-58 [FREE Full text] [doi: 10.1007/s00127-018-1571-5] [Medline: 30069754]
68. Farsi D, Martinez-Menchaca HR, Ahmed M, Farsi N. Social media and health care (Part II): narrative review of social media use by patients. J Med Internet Res 2022 Jan 07;24(1):e30379 [FREE Full text] [doi: 10.2196/30379] [Medline: 34994706]
69. Salzmann-Erikson M, Eriksson H. A mapping review of netnography in nursing. Qual Health Res 2023;33(8-9):701-714 [FREE Full text] [doi: 10.1177/10497323231173794] [Medline: 37192601]
70. Costello L, McDermott ML, Wallace R. Netnography: range of practices, misperceptions, and missed opportunities. Int J Qual Methods 2017;16(1):1609406917700647. [doi: 10.1177/1609406917700647]
71. James TL, Villacis Calderon ED, Bélanger F, Lowry PB. The mediating role of group dynamics in shaping received social support from active and passive use in online health communities. Inf Manag 2022 Apr;59(3):103606. [doi: 10.1016/j.im.2022.103606]

XSL•FO
RenderX

Abbreviations

API: application programming interface
CHIME: Connectedness, Hope and Optimism, Identity, Meaning and Purpose, Empowerment
GIFs: graphics interchange formats
IPCHS: Integrated People-Centred Health Services
OHC: online health community
OMHC: online mental health community
POETIC: Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identity, Connectedness
WHO: World Health Organization

Edited by K Williams; submitted 31.12.23; peer-reviewed by F Lobban, P Marshall; comments to author 09.03.24; revised version received 02.05.24; accepted 10.06.24; published 19.08.24.

Please cite as:
AbouWarda H, Dolata M, Schwabe G
How Does an Online Mental Health Community on Twitter Empower Diverse Population Levels and Groups? A Qualitative Analysis of #BipolarClub
J Med Internet Res 2024;26:e55965
URL: https://www.jmir.org/2024/1/e55965
doi:10.2196/55965
PMID:39158945

©Horeya AbouWarda, Mateusz Dolata, Gerhard Schwabe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining the Popularity, Content, and Intersections With the Substance Abuse and Mental Health Services Administration's Definition of Recovery in a Nonclinical Online Cannabis Cessation Community: Infodemiology Study of Reddit Posts

Elyse J Thulin[1*], PhD; Maureen A Walton[1*], MPH, PhD; Erin E Bonar[1*], PhD; Anne Fernandez[1*], PhD

University of Michigan, Ann Arbor, MI, United States

[*]all authors contributed equally

Corresponding Author:
Elyse J Thulin, PhD
University of Michigan
1109 Geddes Ave
Ann Arbor, MI, 48109
United States
Phone: 1 (734) 764 0478
Email: ethulin@umich.edu

Abstract

Background: Cannabis consumption has increased in recent years, as has cannabis use disorder. While researchers have explored public online community discussions of active cannabis use, less is known about the popularity and content of publicly available online communities intended to support cannabis cessation.

Objective: This study aims to examine the level of engagement and dominant content of an online community for cannabis cessation through 3 specific aims. First, we examine the use of a subreddit cannabis cessation community (r/leaves) over time to evaluate the popularity of this type of resource for individuals who want to stop using cannabis. Second, we examine the content of posts in the community to identify popular topics related to cessation. Third, we compare the thematic findings relative to the 4 domains of recovery defined by the Substance Abuse and Mental Health Services Administration (SAMHSA). By examining these 3 gaps, we take the initial steps toward understanding the experiences being shared online among individuals interested in cannabis cessation and compare them with the principles outlined in the SAMHSA definition of recovery.

Methods: Using the Pushshift application programming interface, we collected the count of posts by year between 2011 and 2021 and the narrative of the 100 posts with the most comments per year in a popular cannabis cessation–focused subreddit (r/leaves). A linear model and a nonlinear model were compared to evaluate change in the number of posts by year. Mixed natural language processing and qualitative analyses were applied to identify top terms, phrases, and themes present in posts over time. Overlap between themes and the 4 SAMHSA domains of recovery (health, purpose, community, and home) were examined.

Results: The number of annual posts in r/leaves increased from 420 in 2011 to 34,841 in 2021 (83-fold increase), with exponential growth since 2018. The term that was the most common across posts was "smoke" (2019 posts). Five major themes were identified, and a narrative arc was represented, from motivations and perceived benefits of cannabis use to the negative consequences of use, strategies to change behaviors, and the positive and negative consequences of change. There was substantial overlap between these 5 themes and 3 of SAMHSA's 4 domains of recovery: health, purpose, and community. However, the domain of home was less commonly identified.

Conclusions: Engagement in this online cannabis support community appears to be increasing. Individuals using this forum discussed several topics, including multiple aspects of recovery defined by the SAMHSA. Online communities, such as this one may, serve as an important pathway for individuals seeking to reduce or cease their consumption of cannabis.

(J Med Internet Res 2024;26:e47357) doi:10.2196/47357

KEYWORDS

cannabis use disorder; online community; cannabis; human-computer interaction; mobile phone

XSL•FO
RenderX

Introduction

Background

The prevalence of cannabis consumption and clinical diagnoses of cannabis use disorder (CUD) has increased in the past 2 decades [1-3]. The intensification in delta-9-tetrahydrocannabinol potency of cannabis products in recent years, expanding types of consumption modalities of high-potency products (eg, vaping, dabs, and edibles), and increased perceived acceptability of cannabis use increase the risk for greater use and dependence on cannabis [4-12]. CUD has serious consequences for individuals, families, and communities, including increasing the number of disability-adjusted life years, indicating lower quality of life [13]. However, compared with other substance use disorders, cannabis use change (eg, cessation and reduction) is under-studied [14].

As noted by the Substance Abuse and Mental Health Services Administration (SAMHSA), recovery is often more than changing one's substance use behaviors [15]. Rather, recovery is a multidimensional process related to health (making choices that support physical health and psychological well-being), community (having relationships that yield support, care, and hope), purpose (engagement with meaningful daily activities, such as work, school, or other participation in society), and home (access to a stable, safe space to live). Within SAMHSA's definition, recovery is achieved through self-directed changes that enable individuals to navigate their lives, improve their health and well-being, and work toward their full potential. While there are a handful of evidence-based clinical psychotherapies for CUD [16-18], approximately 85% of individuals who would qualify as having a diagnosis of CUD never receive clinical treatment [3]. Cost, geographic access, perceived social stigma, internalized stigma, and a lack of knowledge of or low belief in the efficacy of treatments are all barriers to treatment [19,20]. Given the increase in cannabis use and low rates of clinical treatment, we aim to investigate cannabis use behavior changes outside of the clinical context and how these behavior changes may or may not intersect with the domains of SAMHSA's multidimensional definition of recovery.

Multiple Pathways to Recovery

Traditionally, the focus of substance use behavior change within clinical and nonclinical settings (eg, mutual help organizations such as Alcoholics Anonymous) has been the prolonged cessation of all substances [21]. While long-term cessation is beneficial for a proportion of the population who desires to change their substance use, focusing on the singular outcome of cessation of all substances has created barriers for individuals to engage with alternative behavior change pathways (eg, medically assisted treatment and behavioral harm reduction strategies) [22]. To reduce barriers to recovery, in 2013, the SAMHSA expanded its definition of recovery to include multiple pathways and desired outcomes of substance use change and identified the holistic nature of change as it relates to the 4 domains of health, purpose, community, and home [23,24]. Identifying the intersection of cannabis change in nonclinical contexts and the 4 domains of SAMHSA's definition of recovery may yield facilitators and barriers to nonclinical routes of cannabis use behavior change, which could then be used in future health interventions to expand access and relevance of programs within and outside of clinical contexts.

Online Spaces Promoting Cannabis Cessation

There is evidence that those wanting to reduce or cease cannabis use are increasingly turning to publicly available online support forums [25,26]. Unique to online forums is the opportunity for interactions to occur organically and asynchronously within an individual's daily life [25]. Studying online communities related to cannabis use behavior change is relevant for 3 overarching reasons. First, unlike in-person forms of peer support (such as mutual help organizations) or social support provided through formal treatment programs, online communities are not limited by geography. This online access is a particularly salient opportunity for individuals living in rural or underresourced settings who may benefit from online care and support [27]. Second, some online communities provide the opportunity to participate and receive support without disclosing one's personal identifying information, including one's physical appearance. Maintaining pseudonymity may be appealing to individuals who desire social support and advice while simultaneously maintaining anonymity because of the perceived stigma associated with substance use disorders [22]. Finally, online spaces are already in existence, are being used worldwide, and are becoming ever more popular over time [28]. Understanding the content and utility of these spaces could guide future research and inform clinical recommendations for new interventions or additional support [29].

The few studies that have examined online cannabis communities have focused on posts from individuals who are actively using cannabis (as opposed to focusing on cessation or reduction) [30-32]. These studies found that Reddit contains multiple procannabis use communities (called subreddits), which have documented increases in cannabis references across time and discussions about newer forms of delivery (eg, vaping and dabbing). In contrast, one study investigated the desire to cease cannabis use in a subreddit community focused on recovery (r/leaves) [33]. In this study, 38% of individuals who posted to this forum reported symptoms aligned with the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* classification of mild CUD severity, 16% with moderate severity, and 12% with severe severity. Though setting the foundation that this community is composed of individuals who are experiencing dependency on cannabis and want to change, research is needed to investigate the relevancy and value of this online community to those interested in recovery and to provide information on recovery from the perspective of those with lived experience. Such data could expand our understanding of how these communities may be of use to individuals with CUD, which could inform future integration or creation of online support groups as treatment adjuncts.

Objectives

Given the negative consequences associated with CUD, examining the use of such nontraditional self-help–based online forums is important to expand understanding of potential

XSL•FO
RenderX

recovery pathways. In this study, we aim to address 3 gaps in the research. First, we examine the use of a subreddit cannabis cessation community (r/leaves) over time to evaluate the popularity of this type of resource for individuals who want to stop using cannabis. Second, we examine the content of posts in the community to identify popular topics related to cessation. Third, we compare the thematic findings relative to the 4 domains of SAMHSA's definition of recovery. By examining these 3 gaps, we take the initial steps toward understanding the experiences being shared on these sites among individuals interested in cannabis cessation and compare them with the principles outlined in the SAMHSA definition of recovery.

Methods

Data Source

Reddit is an online platform that individuals use to gather information and connect with others [34]. Reddit is composed of many community forums (called subreddits), and the platform as a whole has consistently increased in popularity over the past decade [35,36]. Many subreddit communities allow for posts and comments to be viewed without an account (ie, allowing for general exploration). Finally, individuals who create accounts generally use pseudonyms on their profile, thus avoiding disclosure of personal identifying information (eg, name, location, and physical appearance) [37].

We extracted data from Reddit using the Pushshift Reddit application programming interface (API) from the r/leaves subreddit community [38]. Using this API, we collected the number of posts by year and the text of the 100 posts with the greatest number of comments per year. The r/leaves subreddit thread is described as "a support community to help stop smoking cannabis, marijuana, pot, weed, edibles, or getting high." Though many communities reference cannabis-related terms (eg, cannabis, marijuana, and weed) and multiple communities that are dedicated to cannabis-related topics, r/leaves appears to be the most frequently used community dedicated to cessation of cannabis use [30-33]. This community was started on January 8, 2011, and as of June 2022, it had approximately 240,000 members. Though demographic statistics for this subreddit are not available, from our initial review, we found that all posts are written in English. Broadly, the known statistics of Reddit users indicate that over half of them are from the United States, and the distribution of users skews younger and male [39].

Data Extraction

First, we gathered the top 100 most–commented-on posts for each year within the r/leaves community. We only collected original posts, as opposed to comments, which have their own set of popularity rankings. Given prior research findings that a small number of community members post with high frequency (and thus, posts may not be representative of a wider set of community members) [40], we determined the number of unique users who authored the top posts. We noted an archival issue in Pushshift for the year 2013, which prevented the collection of posts from that year; thus, our data includes the top 100 most–commented-on posts by year for 2011, 2012, and 2014 to 2021 for 1000 posts.

Data Analysis: Top Terms and Phrases

To accomplish the first stated goal (understanding the change in use of this forum across time), we calculated the annual number of original posts by year from 2011 to 2021. We plotted the number of posts by year to visualize the trend of posts across time, compared coefficient significance, and adjusted R^2 values in linear versus quadratic models to assess which had the better fit.

Our second goal was to identify the top themes in the most popular posts (those with the greatest number of comments). To do so, we used mixed computational and qualitative methods [41,42]. Given the importance of input from individuals with lived experience with substance use disorders and recovery, the lack of research on cannabis recovery online forums, and the frequent use of clinical samples in the general field of recovery, we chose to examine the most frequent terms and used an inductive qualitative coding approach to allow the data to guide our findings. Once we had the emergent themes present in this online community, we then examined the intersection of these themes with existing SAMHSA recovery definition domains (health, purpose, community, and home). First, we preprocessed text (ie, case standardization, removing punctuation and numbers, and using the stems of all words) using standard procedures [43]. We removed the most common English words that do not have significant meaning (eg, the and a) using the System for the Mechanical Analysis and Retrieval of Text stopword dictionary [44] through the R "quanteda" package [45]; for phrases, stop words were not removed to increase accuracy and interpretability of multiword findings [46]. Preprocessing of text increases the accuracy and parity of natural language processing techniques and increases the replicability of findings [43]. Data scraping and analyses were conducted in R (version 4.1) [47].

Data Analysis: Emergent Themes in Posts

After identifying the top terms and phrases present in preprocessed posts, we randomly selected 10% (100/965) of posts to review to help contextualize findings; mixing natural language processing and qualitative methods increases the accuracy of findings and the ability to analyze a greater amount of text [42]. In line with thematic content analysis [48], the lead author read the subset of posts, created a codebook from the data, and then coded the data using the established codebook; the results from this coding process were reviewed by coauthors (2 of whom are licensed clinical psychologists and 1 of whom is a community psychologist) for saliency and relevancy. Finally, we compared the emergent themes with the 4 domains (health, purpose, community, and home) of SAMHSA's definition of recovery.

Ethical Considerations

This study was determined as unregulated by the institutional review board at the University of Michigan. As data are publicly available and we did not have study participants, informed consent was not collected and no compensation was made. Though the data contain usernames (which are frequently pseudonyms), to uphold privacy we do not report any of those usernames in our work. To uphold deidentification and follow

XSL•FO
RenderX

Reddit's intention of anonymity, the quotes presented below are paraphrased based on the reviewed data [49]. Finally, generative artificial intelligence was not used in any aspect of this study (data scraping, cleaning, manuscript drafting, etc).

Results

Overview

The number of posts in r/leaves by year increased by 8195% from 420 in 2011 to 34,841 in 2021. In the linear model, for every increase in year, the number of posts increased by 3379.3 (SE 341.3; $P<.001$; Figure 1). The linear model accounted for 91.5% of the variance. In the quadratic model, the linear (β=–1,179,000.0, SE 230,900.0; $P=.001$) and squared coefficients (β=293.1, SE 57.27; $P=.001$) were both significant. The interpretation of these values suggests that the starting point is low (eg, intercept) and increases as the curve opens upward (signified by the positive valence of the quadratic term). The quadratic model accounted for 98% of the variance.

Figure 1. Descriptive count of posts from the Reddit cannabis recovery forum, covering the period from 2011 to 2021.

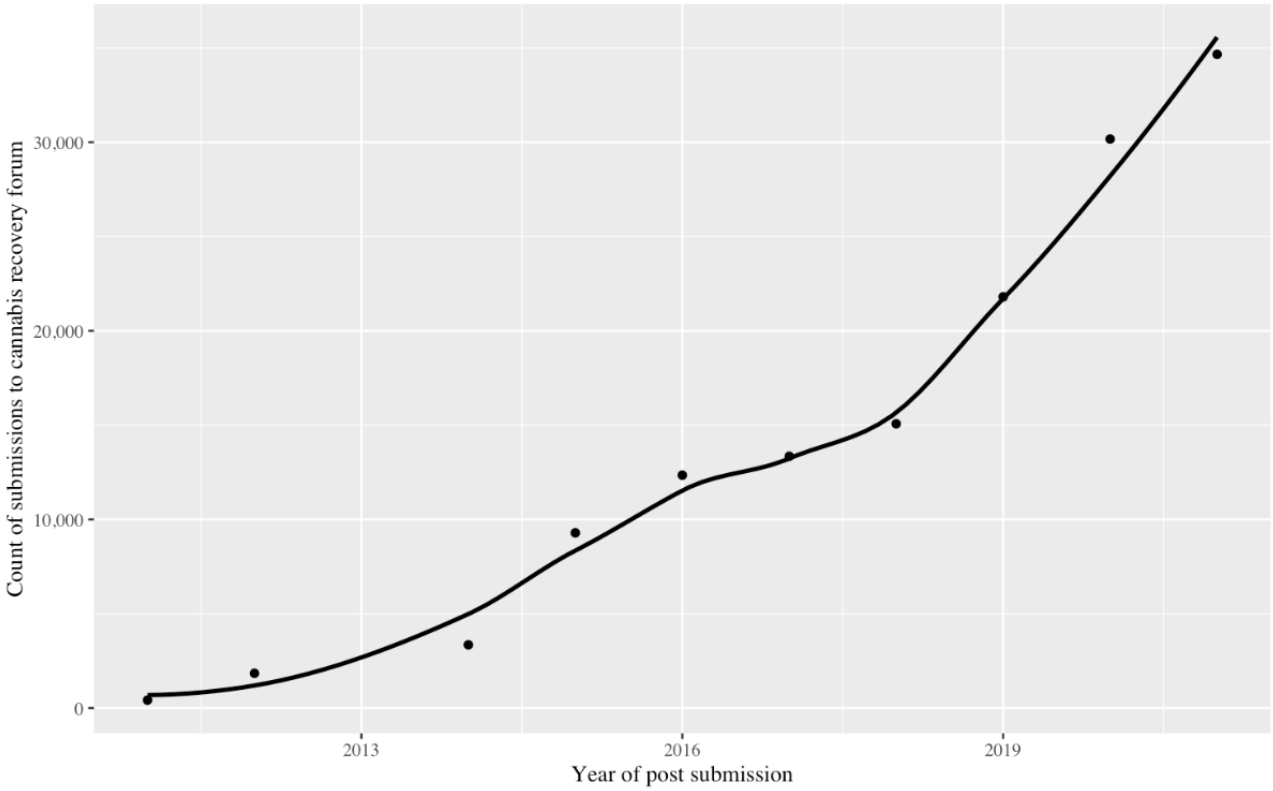

Top Terms and Phrases

Overall, 1000 posts were collected, representing the 100 most–commented-on posts per year across 10 years (excluding 2013, as noted in the Data Extraction section); 965 of the posts included text (as opposed to other media, such as a video link or only a title) and were analyzed in this study. After text preprocessing, there were 7829 unique features (eg, terms). The top 5 terms across all posts were "smoke" (n=2763), "time" (n=1756), "weed" (n=1738), "day" (n=1635), and "feel" (n=1456). A review of terms produced several clustering of topics related to time (eg, day, month, and year), cannabis use (eg, smoke, weed, and high) and perceived identity related to cannabis use (eg, addict and sober), cannabis behavior change (eg, quit, begin, start, and stop), cognitive functioning (eg, thought and sleep) or emotions (eg, happy and hope), mental health (eg, depress and anxieti, which is the root for permutations of "anxiety"), social purpose (eg, job, school, and money), and terms relationships (eg, friend and famili). The top 40 terms are shown in Figure 2.

XSL•FO
RenderX

Figure 2. Descriptive frequency of the most common terms in the top 100 most-liked posts from the Reddit cannabis recovery forum posted from 2011 to 2021.

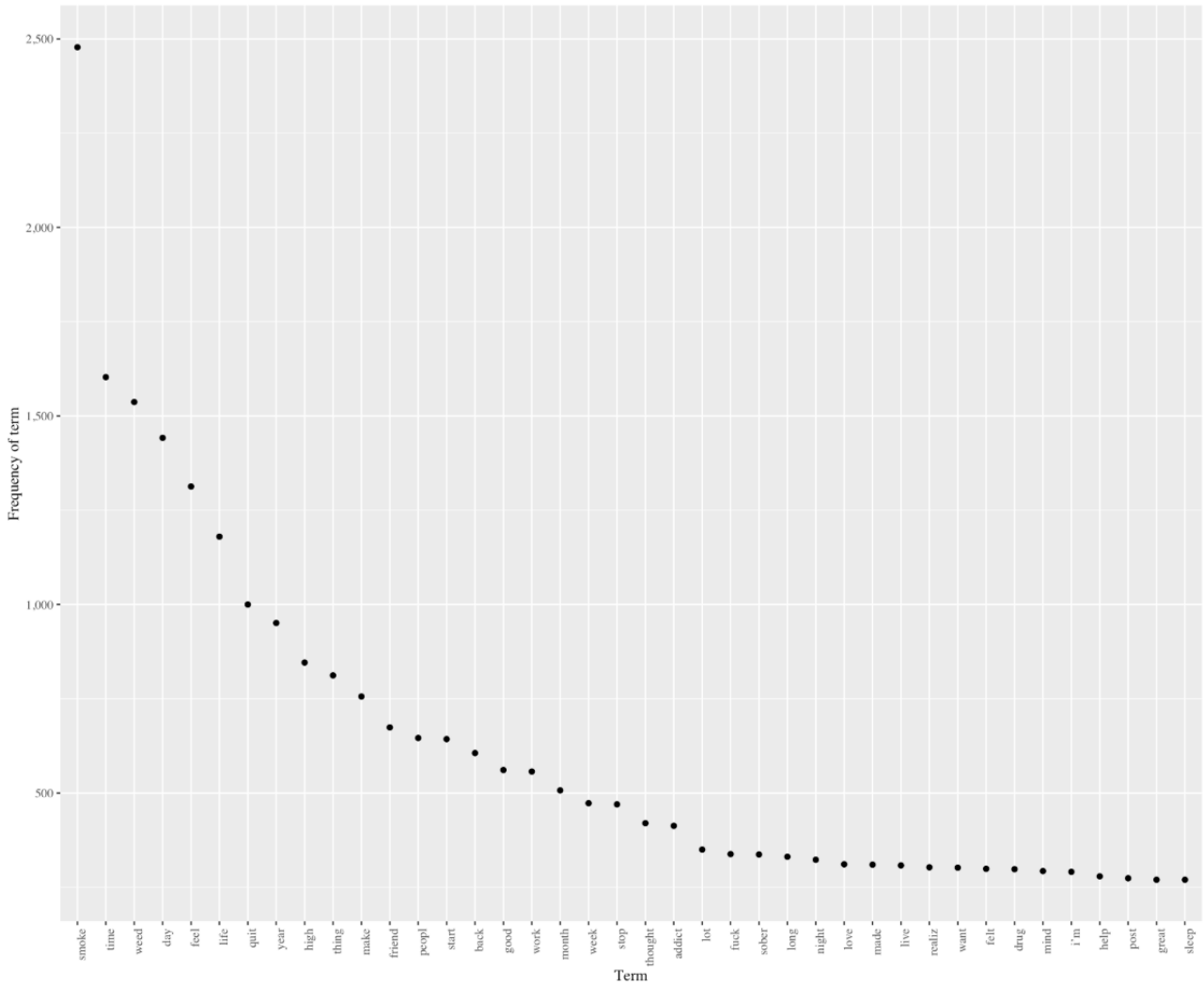

Phrases were defined as strings of 3 to 6 words**,** given the limited interpretability of 2-word strings; notably, stemming at times results in the atypical presentation of words (eg, “decid,” the stem of “decided,” “deciding,” and “decide”). The top phrases were “i want to” (n=287), “a lot of” (n=173), “i need to” (n=140), “when i was” (n=118), “be abl to” (n=117), and “i use to” (n=114). The statements expressing want often referred to outcomes of cessation that the individual hoped to experience, including “I want to feel happy,” “I want to be able to remember my time,” and “I want to find a new job.” Many of the top phrases were I statements, such as “i feel like” (n=93), “i have been” (n=82), “i had a” (n=81), “i had to” (n=78), “i don’t want” (n=71), “i was a” (n=65), “i decid to” (n=63), “i just want” (n=58), and “i don’t know” (n=54). Several of the phrases referenced cannabis use behaviors, such as craving (eg, “want to smoke”; n=67) and desire to cease (eg, “to stop smoke”; n=54). There are also multiple references to behaviors related to cannabis cessation, including “tri to quit” (n=64) and “want to quit” (n=61). The top 50 phrases are presented in Table 1.

XSL•FO
RenderX

Table 1. Top 50 two-to-three-word phrases in cannabis cessation online community posts (N=1000)[a].

Phrases used in posts	Phrase across posts, n
“i want to”	287
“a lot of”	173
“i need to”	140
“when i was”	118
“be abl to”	117
“i use to”	114
“want to be”	113
“in my life”	111
“all the time”	102
“of my life”	99
“i feel like”	93
“just want to”	91
“don’t want to”	82
“i have been”	82
“i had a”	81
“i had to”	78
“the first time”	77
“you want to”	76
“i don’t want”	71
“to get high”	68
“in the morn”	68
“i have a”	67
“want to smoke”	67
“i was a”	65
“for the first”	65
“my life i”	64
“that i was”	64
“the end of”	64
“tri to quit”	64
“i decid to”	63
“you need to”	62
“want to quit”	61
“one of the”	60
“hang out with”	60
“to be a”	59
“it was a”	59
“i just want”	58
“feel like i”	58
“a long time”	57
“want to do”	56
“i just want to”	56
“i don’t want to”	55

XSL•FO
RenderX

Phrases used in posts	Phrase across posts, n
“in the past”	54
“becaus i was”	54
“i don’t know”	54
“to stop smoke”	54
“for a while”	53
“this is a”	53
“you have to”	53
“i was in”	52

[a]Given the sparsity of term and phrase matrices, providing a percentage of the count over the total number of words or phrases is not relevant.

Emergent Themes in Posts

Overview

The results of the thematic content analysis produced five overarching themes: (1) history of cannabis use, (2) consequences of cannabis use, (3) reasons for changing cannabis use, (4) cessation strategies, and (5) consequences of change. In addition to the descriptions provided later in the text, the intersections between themes and SAMHSA domains are presented in Table 2.

Table 2. Intersection of themes present in the top posts posted between 2011 and 2021 in the Reddit Cannabis Cessation Forum and the 4 domains of recovery established by the SAMHSA[a].

Emergent themes in online cessation community posts	SAMHSA’s domains of recovery			
	Health	Community	Purpose	Home
Individual identity and cannabis use	__[b]	“when I started smoking, it was awesome - I loved going to parties and hanging out with friends. Smoking weed felt so good, so I started doing it more and more...”	• “I smoked weed every day for 10 years” • “Looking back, I used to feel like, ‘man, I am a relatively productive stoner’”	—
Consequences of cannabis use	“I experienced psychosis due to drug use, and attempted suicide during a period of psychosis, and yet I still smoked – I used on and off my mental health meds”	“my parents didn’t trust me anymore, because I continuously lied to them about smoking and would occasionally steal from their house when I needed cash”	• “I can’t remember any of those years - those years are just a blurry fog”	“My behaviors when high were negatively impacting my entire family, and my parents were making changes to ensure a healthier environment for my siblings”
Reasons for change	“I hoped the anxiety I felt would be over when I stopped using”	“I felt embarrassed to be high around her because she didn’t use, so I started to moderate my use”	• “I’ve been smoking daily for months. I feel lethargic, and like my life isn’t moving forward”	—
Cessation strategies	“I still get anxious, and when I am, I reach out to friends - we go out to eat, hang out, or play games”	“I asked my partner to help me by hiding my stash”	—	—
Consequences of change	“a year clean, and I am sleeping well”	“...over the past year since I quit, we’ve been able to rebuild a lot of trust and they are helping me get back on track with school”	• “I feel sharper and able to efficiently multitask...I recently was promoted because I started to step up in my role”	—

[a]SAMHSA: Substance Abuse and Mental Health Services Administration.

[b]Not applicable.

History of Cannabis Use

Many posts contained reference to one’s personal history with cannabis use. Reference to personal history included disclosing the age at which the author who posted began using cannabis and the number of months or years the author used cannabis (eg, “I used daily since 16” or “I took my first tok at 13 and was

XSL•FO
RenderX

a daily stoner by 14”). Social networks were also referenced regarding one’s period of active cannabis use; these included positive experiences they had with friends when using cannabis, though these times were often recalled as occurring earlier in one’s history of cannabis use.

Providing a background on the author’s personal experience with cannabis in some posts was stated as evidence against the false social perception that cannabis is not addictive (eg, “call it whatever you want, an addiction, dependence, over use, whatever – as someone who quit 97 days ago after using daily for 19 years, I can say for certain that for some people, having pot in your life means you is taking away your ability to reach your full potential”). For many authors, the reference to time was related to cessation. For some authors, it was the celebration of recent cessation (eg, “it has been five days since I last lit up!”); for others, it was to provide perspective on the experience of not having cannabis in their lives for a prolonged period (eg, “it has been 6 years since I last used the green monster...I’m proud to live a life that is free from the haze of weed”) and to sharing one’s iterative experience with cessation through chronicling experience by period (eg, “month one, no sleep; bizarre dreams if you can drift off; month two, settling in, no longer always thinking about cannabis...month six, new friendships not on the basis of smoking together”).

SAMHSA’s domains of purpose and community were reflected in this theme. Multiple authors discussed the intertwined nature of cannabis use and their daily life activities. During their active use stage, cannabis became a major feature in their daily lives, pulling their attention and resources and informing social networks. The reference to social ties with other people who actively use cannabis maps onto the SAMHSA constructs of community, while feelings of connection intersect with the construct of purpose. Nonetheless, posts suggested that the social connections formed through cannabis use may wane over time, or the cost of prolonged cannabis use on one’s health, decreased feelings of purpose, added burden on other social relationships, and risk toward home may reduce the benefits of the social connections that were built with the use of cannabis. In sum, reference to one’s history was important to many authors, reflecting times in their life that had direct implications for feelings of community connectedness as well as a sense of purpose.

Consequences of Cannabis Use

While there were some perceived benefits to cannabis use, particularly noted when individuals started using, authors noted many negative consequences. Authors most frequently identified consequences as an effect of prolonged, frequent cannabis use. These consequences were related to SAMHSA’s domains of health, purpose, community, and home. Health consequences included cognitive and physical effects. Fuzzy memories, a haziness, or a lack of memories of a given time when high (eg, “I can’t remember any of those years—they are just a blurry fog”) were all examples of cognitive consequences. Cognitive effects extend beyond conscious memories into sleep, in the form of not being able to remember dreams (eg, “while smoking, I never remembered my dreams”). However, in other cases, not remembering was a marker of the severity or habitual nature of cannabis use (eg, “I can’t remember the last holiday where I wasn’t high”). Cognitive effects of cannabis use also impacted the ability to maintain existing or build new social connections, and at times, it resulted in feelings of isolation and loneliness, such as an individual not having friends during their “weed-induced haze,” resulting in them feeling “antisocial” for a period of their life. Weight gain was a physical consequence noted by authors (eg, “weed made me fat—I had no ability to resist snacking”).

While some authors noted a perceived normativity of cannabis consumption during certain phases of life (eg, high school and college), many noted the level of their consumption was greater than others, with negative consequences such as worse grades and poor social interactions at work (eg, “I avoided eye contact with people I didn’t know, and when I’d speak, I’d turn red because I sounded dumb”). Some authors noted ways in which established relationships were deeply strained or even ended because of the author’s use of cannabis (eg, “my parents didn’t trust me anymore, because I continuously lied to them about smoking and would occasionally steal from their house when I needed cash”). This was particularly true for romantic relationships (eg “she got sick of me not being present, just constantly stoned—eventually, she left”). Even in cases in which established relationships were not entirely ended, some authors noted that needing a space to use cannabis limited the types of interactions in which the author and their romantic partner interacted.

Though commented on less, the SAMHSA domain of purpose was also present; some of the authors retrospectively recounted their perceptions that prolonged, frequent use of cannabis resulted in a loss of sense of self, sense of purpose, and feeling like one’s life was stuck and not moving forward (eg, “all my life was going to work then coming home and smoking; I had no inspiration, no intention—weed made me boring and stagnant”). At times, this topic included losing their sense of what they wanted for their lives, and they felt that the ongoing use of cannabis had reduced their understanding of their core values. While the SAMHSA domain of home (having a stable, safe place to live) was also less common, violations of relationship quality or trust, including stealing and lying, at times were associated with having to move out of a given living situation. Overall, posts reflected negative consequences of cannabis use related to all 4 of the SAMHSA recovery domains (health, community, purpose, and home).

Reasons for Change

Throughout posts, authors shared various motivators for changing their use of cannabis. Sometimes this was expressed as a broader sentiment (eg, “I know stopping would be good for my life”). At other times, specific motivations were mentioned, such as a desire to mitigate cognitive consequences. For example, the loss of memory was a motivator for one author who wanted to remember the last bit of time left with a beloved, but sick, pet; another author wanted to be more present and have clearer memories of time with their family. Other authors wanted to stop the “numbing effects” they experienced through frequent, prolonged cannabis use, even if these effects were desired in the early stages of cannabis consumption. Mental health was

XSL•FO
RenderX

also a motivation for change; while some authors acknowledged the use of cannabis when "feeling depressed" or anxious, others noted a desire to change because of the negative effects on their mental health (eg, "I hoped the anxiety I felt would be over when I stopped using [cannabis]"). This included when individuals perceived that cannabis use increased mental health symptoms (eg, "weed makes my depression worse, and my ability to handle dealing with my depression harder"). However, changes in mental health symptoms on reduction or cessation of cannabis use were not always positive (eg, "I grow anxious when I don't smoke"), perhaps reflecting withdrawal symptoms. In addition, even with the knowledge that cannabis may have negative effects on mental health, the negative effects did not always yield immediate behavior change (eg, "I experienced psychosis due to drug use, and attempted suicide during a period of psychosis, and yet I still smoked—I used on and off my mental health meds").

The effects of cannabis on mental health symptoms often interacted with other SAMHSA domains of recovery, including community. For example, some authors observed in retrospect that while cannabis had initially made them feel less anxious in social settings, after a while, it made them more socially anxious and paranoid, resulting in withdrawal from social interactions, including staying home or skipping events. While relationships were at times established or ended because of cannabis use, in some cases, a relationship was a reason for the change (eg, "I felt embarrassed to be high around her because she didn't use, so I started to moderate my use"). Finally, feelings of purpose were at times mentioned as a motivator for change; with less time spent on acquiring, using, and acutely recovering from cannabis use, some individuals noted they would be able to "focus on thriving in life as opposed to just existing." This sentiment also extended to the ability to trust one's internal perceptions and feel less paranoid. Across the various specific reasons for change, most can be described as related to the SAMHSA recovery domains of health and community.

Behavior Change Strategies

While the r/leaves community aims for cannabis cessation, many members recognize that reducing consumption might be a viable option for some people. However, they may not consider it the best choice for the specific author of the post. Choosing cessation as a goal over reduction or moderation at times reflected challenges in maintaining reduction or moderation while continuing to use cannabis. While some individuals in the community reported trying moderation as a step to cessation, moderation often did not help many authors cease their cannabis use, and many of the authors reported stopping "cold turkey."

Many of the strategies for cessation were individually guided and behavioral. For some, this was resisting thoughts (eg, urges) about wanting to resume cannabis use (eg, "I'm learning how to resist thoughts and the want to smoke"); for others, it was learning to handle challenging thoughts via new behavioral and mental coping strategies (eg, "now I write down my thoughts, as it helps me process"). Individuals found helpful strategies and tips through the online subreddit community; for example, after reading posts, one author commented they, "...wrote out a list of reasons to quit and posted it on the wall [in a common area]." This list included reasons why the person quit and reasons why they would not use cannabis again in the future. Another strategy mentioned was to "stay busy" to get through the initial months of cessation and to not return to using cannabis because of boredom or uncertainty of what to do with one's time (and particularly periods when the person frequently used cannabis, such as at night).

Coping skills were also frequently referenced, often in conjunction with one's social relationships. For example, individuals asked for tangible support behaviors from romantic partners ("I asked my partner to help me by hiding my stash"). Friends and family also were at times attributed as forms of support (eg, "I came clean and told my friend I needed support in order to change" or "I'm still working on making changes, but now that I quit, I have the support of my family, which is helping me"). In other cases, individuals noted spending time with friends rather than giving in to triggers to use cannabis (eg, "I still get anxious, and when I am, I reach out to friends—we go out to eat, hang out, or play games"). While social relationships provided a mechanism of support, some individuals noted the need to set boundaries within existing social relationships as a way to sustain cessation (eg, "at my friend's cook out, I avoided my friends who smoked—I was worried I couldn't say no"). Behavioral change strategies were largely related to the SAMHSA recovery domain of health, whereby individuals were finding ways to choose behaviors that helped them achieve their goals of reduction or cessation, and several of the behaviors reflected the role of social support in behavior change, mapping onto the recovery domain of community.

Consequences of Change

There were positive (ie, benefits) and negative consequences of change, touching upon the SAMHSA recovery domains of health, community, and purpose. For health, sleep and dreams were often referenced both as positive and negative consequences. Reflecting on short-term effects, some individuals noted difficulties with sleep (eg, "sleeping isn't happening"). Others found themselves having "vivid," "wild" dreams. However, after some time, other individuals noted that they had returned to better sleep (eg, "a year clean, and I am sleeping well"), which was a benefit of change. Another outcome related to physical fitness, with some individuals noting that they spent more time working out as both a coping strategy (ie, distraction) and as a way to invest in their health and well-being (eg, "I started small, just committed to getting to the gym every day; now I can lie down at night with a body that feels good and tired from a 10 mile bike ride"). Others noted changes to eating habits, either through less snacking because of not being intoxicated or as an intentional way to invest in their behavior change and their health by choosing healthier options for meals.

Cessation was often associated by authors as having effects on their social networks. Sometimes, this was positive because of improving previously strained family relationships (eg, "My behaviors when high were negatively impacting my entire family, and my parents were making changes to ensure a healthier environment for my siblings...over the past year since

XSL•FO
RenderX

I quit, we've been able to rebuild a lot of trust and they are helping me get back on track with school"). Other times, this meant more time to invest in existing social or romantic relationships ("I now spend more intentional time with my partner—I'm not just thinking about when I can go home and get high, but actually pay attention to what we are doing"). Sometimes, new relationships were formed (eg, "I'm dating someone new and they are really awesome and motivated to live a healthy and positive life"). There were also many references to the support that individuals received through the r/leaves subreddit; in some cases, this seemed to help support individuals who were moving away from relationships that were associated with active cannabis (eg, "I gave up my weed friends—I'm really thankful for finding support in this sub[reddit]").

Changes in feelings of purpose because of cannabis behavior change were positive. Within work and school settings, multiple authors noted feeling more productive and having an increased ability to engage with the content or tasks associated with their role because they no longer felt the mental haze of being high (eg, "I feel sharper and able to efficiently multitask...I recently was promoted because I started to step up in my role"). Others felt a better sense of self and direction because of behavior changes (eg, "I'm focusing on defining new goals for my life, one of which I realized was going back to school!"). Other authors noted that they were able to engage in activities that they enjoyed in their life and be more present within those experiences, which added meaning and new insights. Finally, multiple authors noted the importance of helping others; sometimes, this was general advice (eg, "go help others, giving back makes life more meaningful"). Other times, the reflection on helping others was within the r/leaves subreddit community (eg, "I'm grateful for the support I've gotten from this community, and I want to share my story in case it helps others to stop and to feel more fulfilled"). Individuals discussed many outcomes that they attributed to reducing or ceasing cannabis use within the online forum. These outcomes align with the SAMHSA recovery domains, including making choices to promote health, building connections with family and community, and engaging in activities that provide a sense of purpose.

Discussion

Principal Findings

Our findings provide information to help fill 3 important gaps in knowledge on the use of a nonclinical, online cannabis cessation community. First, our findings show that the use of this cannabis recovery forum has increased in recent years, with >80 times the number of posts in 2021 as compared to 2011. This finding supports and expands upon data showing increases in engagement with online mutual help organization cannabis support groups (ie, Marijuana Anonymous) [26]. Second, we identified 5 major themes in the online community: individual history and cannabis use, consequences of cannabis use, motivators for cannabis use change, cessation strategies, and consequences of change. Within these themes, we identified intersections with all 4 of the SAMHSA domains of recovery (health, community, purpose, and home). Our findings taken together highlight the importance of online communities, both in their growing popularity and in their content supporting multidimensional definitions of recovery.

We hypothesized that increases in the community would exist, as evidenced by a higher number of users and posts over time. We found that the rate of increase was not constant but differed based on the period. Specifically, we found that the increase in annual posts steepened around 2018. This temporal trend may be explained by several factors. First, the rise in forum use may be related to the general expansion of Reddit [28]. Second, during this period, several US states legalized cannabis for recreational purposes (eg, Michigan in 2018, Illinois in 2019, and 8 additional states in 2020-2021), which coincided with increases in CUD [50], potentially in part because of the increased risks of dependence posed by higher tetrahydrocannabinol content in legalized cannabis [10,12]. However, it is important to note that not all Reddit users are from the United States and that because of the anonymous nature of community members, it is not possible to determine the demographic composition of the r/leaves community [39]. Regardless, the increased involvement in this community reflects a greater absolute number of individuals seeking support within the online space to quit or reduce cannabis consumption. The increased use supports the notion that online spaces such as r/leaves provide an accessible forum to support those seeking recovery from cannabis.

More Than Just Change: Understanding the Context of Cannabis Use as a Critical Aspect of One's Narrative

The most popular content of the r/leaves forum was the progression from beginning cannabis use through desired changes, strategies for change, and outcomes attributed to change. While consequences of cannabis use, reasons for change, strategies for change, and outcomes related to change may be expected given the r/leaves community mission statement of being "a support community to help stop smoking cannabis, marijuana, pot, weed, edibles, or getting high," the importance of the narrative arc of the circumstances of when someone began using cannabis, including reasons why, and the frequency that individuals listed their duration of use emerged in many posts. For many individuals, cannabis was present for a significant amount of their lives, with authors frequently reflecting on years of use; it appears that sharing the long-standing presence of cannabis in the author's life was important to the author and meaningful in their interaction with the online community. Our findings that narratives were common complemented and expanded other research. Narrative storytelling is common in other nonclinical substance recovery organizations, such as Alcoholics Anonymous, and has been theorized to be a mechanism through which individuals maintain their behavior change [51]. In addition, telling one's story can help reduce forms of stigma and increase help-seeking behavior [52], while narratives with examples of successful substance use change can help inspire those who read the stories to make their own behavior changes [53]. However, unlike groups such as Alcoholics Anonymous, in which individuals may share their stories multiple times throughout their attendance, we found that almost 90% (797/965) of posts within this cannabis

XSL•FO
RenderX

cessation community were written by unique authors. Future work that explores repeated engagement in telling one's own story versus singular sharing may be important to better understand this contrast. In addition, future research could examine if individuals repeatedly share aspects of their story within comments to other posts.

Content of Posts: Intersections With SAMHSA Domains of Recovery

There were many cases of overlap between the 5 themes found in this online nonclinical cannabis cessation community and SAMHSA's 4 domains of recovery (health, purpose, community, and home). For example, references to health in the themes of active cannabis use experiences and history of use, motivations for change, and consequences of change included anxiety and depression. Although there is research into the potential medical uses of cannabis related to mental health [54,55], it is noted that r/leaves focuses on CUD, and the findings herein align with well-documented research from clinical settings on the bidirectional association between greater cannabis use and poorer mental health symptoms. Specifically, multiple studies document the association of greater cannabis use with greater anxiety and depression [56-58], while other researchers have found that cessation of cannabis use increases the likelihood of withdrawal symptoms, which can include acute anxiety and mood changes [14].

While physical health topics such as sleep and dietary choices were sometimes referenced within negative consequences of cannabis use, they were more often talked about within outcomes of change. Sleep was at times negatively impacted by cessation, particularly in the acute period directly after cessation, which is likely a symptom of withdrawal [54]. Specifically, multiple posts referenced a lessening of sleep disturbances over time and the return to perceived normal sleep. In contrast, dietary choices within consequences of change were often referenced as a means of identifying purpose and investing in one's self. While research has shown that cannabis use does not always increase BMI and cannabinoids may increase metabolism and thus decrease BMI [55], identifying food choices as a mechanism of change with the motivation of investing in one's health could be leveraged in both clinical and nonclinical settings as another pathway through which individuals might successfully achieve their intended behavior changes.

Cognitive impairment (sometimes referred to as "fog" or "haze" by authors) was also commonly mentioned in this online community and appeared across multiple themes. Previous research has described the contrasting motivations of using cannabis to both tune out and disconnect from life as a form of relaxation as well as to tune in and focus intensely on a given point of stimuli [56]. Our findings supported this, as authors in the cannabis online community referenced mental haze or fog as the aim of cannabis use to tune out stressors of life within the theme of cannabis use and history of use. However, the cognitive effects of cannabis use were perceived to have negative impacts on social relationships, health, and sense of purpose. Additional work identifying the levels or contexts of cannabis use that cause or increase mental health symptoms and contrasting them to contexts under which cannabis use may be medically relevant are important future directions and could be used to inform future public health campaigns to enhance public awareness and knowledge.

Social relationships were referenced across all 5 themes identified, highlighting their importance. Social relationships are part of SAMHSA's domain of community, which includes "relationships and social networks that provide support, friendship, love, and hope" [15]. Though cannabis use does not always result in severe social consequences, it can negatively impact one's social relationships and ability to engage in prosocial daily activities [22]. In posts, we found that individuals managed their social relationships during cannabis cessation, with, at times, the need to instill relational boundaries that would help to prioritize their cannabis cessation goals. Cessation also appeared to prompt rebuilding relationships that had become strained during one's period of cannabis use, such as with parents or family members, and the opportunity to build new relationships outside the context of using cannabis. The prevalence of social relationships mentioned throughout posts and the inclusion of social relationships within the SAMHSA domain of community stresses the importance of these interpersonal ties when seeking recovery from cannabis use. Although we found a high prevalence of posts related to social relationships, other forms of community building, which have been identified in recovery from addictions to other substances [57], such as engagement with religion or community service, were less common in posts. It may be that individuals who would benefit from community building through religion or service may not use this online community or that these themes may not be as popular.

The SAMHSA recovery domain of home (a stable, safe place to live) appeared less in this online community. Though there was a slight reference to home in regard to negative consequences of use, in that cannabis use was at times stressing one's social relationships with those that they resided with, one hypothesis may be that CUD is less commonly associated with housing insecurity [58]. Alternatively, it may be that those who experience housing insecurity because of cannabis use have less access to the online space because of barriers to access (eg, Wi-Fi or data, computers or smartphones, and electricity to charge phones), or this online space is of much lower priority given other needs. Further exploration of the impact of cannabis on housing is needed outside of this online community to better understand the risk. This research would also help to inform which populations are not served by the type of nonclinical online community examined in this study.

Limitations

Though this study provides important findings on an emerging area of research, it also has limitations. First, we only analyzed the content of the initial post; while providing important information on the types of topics that individuals are seeking to discuss, it does not provide information on the depth and content of discussions that may have occurred in response, which could be a useful direction for future work [59]. Second, though the Pushshift API is a widely used resource [31,60,61], it has some limitations [38], including in relation to missing

XSL•FO
RenderX

data (eg, missing data for the r/leaves subreddit in the year 2013) because of the nature of the archival process; however, in analyses on missingness, researchers found that missingness is not differential by subreddit [62]. In addition, despite the year of missing data, the increase of posts across time within the r/leaves subreddit is significant and thus does not undermine our findings of growth. Finally, selecting the top 100 posts based on the score within Pushshift is a limitation. This is because while the Pushshift API catalogs data at regular intervals, it is inherently an archival tool. However, posts can exist and can be viewed and commented on indefinitely [63]. By creating a static archive, any interactions that occurred between the most recent pull of the cataloging process may not be reflected in the Pushshift data set. While this can create some bias in the Pushshift data set, it is important to contextualize this limitation within the regular life cycle of a post; the first few hours of a post's "life cycle" is when a post typically receives the majority of its views and interaction, after which the post starts to move downwards in the feed, moving to later pages on the subreddit, and thus receives far fewer views or interactions [63,64]. Thus, the catalog process has likely captured most of the activity. In addition, the use of an archived data set increases the replicability of findings, which is critical in building empirical knowledge.

Conclusions

In summary, nonclinical online support communities may be an important resource for individuals who desire to change their cannabis use. We found that the popularity of this forum increased by >8000% from 2011 to 2021. The 5 major themes of discussion represented a narrative arc, covering the proactive engagement in cannabis use, the negative consequences of use, strategies for changing behaviors, and the positive and negative outcomes of those changes. The significant overlap between the themes present in this online community and SAMHSA's 4 domains of recovery supports the fact that this community is engaging in meaningful discussions that overlap with aspects of clinically supported recovery, despite these interactions occurring outside of a clinical setting. Future research should examine the interactions of posts to understand forms of support, information and resources provided, and gaps in the types of content shared as well as to inform opportunities for future online interventions and referrals to online communities for support.

Acknowledgments

The authors would like to acknowledge the contributions of all the individuals who have contributed to this study. The authors appreciate the public availability of data through Pushshift, which is operated by Jason Baumgartner. EJT is funded through a T32-Postdoctoral Training Grant (AA007477-29; principal investigator: Fredric Blow).

Data Availability

Data are not available to share due to Department of Veterans Affairs and ethical retractions; however, annotation instructions and vocabulary are available upon request..

Authors' Contributions

EJT was responsible for the conceptualization, data curation, formal analyses, methodologies, and writing the original draft or reviewing and editing subsequent drafts. MAW was responsible for supervision, review of analyses, and writing review and editing. EEB was responsible for supervision, review of analyses, and writing review and editing. AF was responsible for conceptualization, supervision, analysis review, and writing review and editing.

Conflicts of Interest

None declared.

References

1. Compton WM, Han B, Jones CM, Blanco C. Cannabis use disorders among adults in the United States during a time of increasing use of cannabis. Drug Alcohol Depend 2019 Nov 01;204:107468 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.05.008] [Medline: 31586809]
2. Hasin DS, Saha TD, Kerridge BT, Goldstein RB, Chou SP, Zhang H, et al. Prevalence of marijuana use disorders in the United States between 2001-2002 and 2012-2013. JAMA Psychiatry 2015 Dec;72(12):1235-1242 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.1858] [Medline: 26502112]
3. Key substance use and mental health indicators in the United States: results from the 2020 National Survey on Drug Use and Health (HHS Publication No. PEP21-07-01-003, NSDUH Series H-56). Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. 2021. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt35325/NSDUHFFRPDFWHTMLFiles2020/2020NSDUHFFR1PDFW102121.pdf [accessed 2024-09-03]
4. Cash MC, Cunnane K, Fan C, Romero-Sandoval EA. Mapping cannabis potency in medical and recreational programs in the United States. PLoS One 2020 Mar 26;15(3):e0230167 [FREE Full text] [doi: 10.1371/journal.pone.0230167] [Medline: 32214334]

XSL•FO
RenderX

5. ElSohly MA, Mehmedic Z, Foster S, Gon C, Chandra S, Church JC. Changes in cannabis potency over the last 2 decades (1995-2014): analysis of current data in the United States. Biol Psychiatry 2016 Apr 01;79(7):613-619 [FREE Full text] [doi: 10.1016/j.biopsych.2016.01.004] [Medline: 26903403]
6. ElSohly MA, Chandra S, Radwan M, Majumdar CG, Church JC. A comprehensive review of cannabis potency in the United States in the last decade. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Jun;6(6):603-606. [doi: 10.1016/j.bpsc.2020.12.016] [Medline: 33508497]
7. Jenkins MC, Kelly L, Binger K, Moreno MA. Cyber-ethnography of cannabis marketing on social media. Subst Abuse Treat Prev Policy 2021 Apr 26;16(1):35. [doi: 10.1186/s13011-021-00359-w] [Medline: 33902649]
8. Mahamad S, Wadsworth E, Rynard V, Goodman S, Hammond D. Availability, retail price and potency of legal and illegal cannabis in Canada after recreational cannabis legalisation. Drug Alcohol Rev 2020 May;39(4):337-346. [doi: 10.1111/dar.13069] [Medline: 32291811]
9. Matheson J, Le Foll B. Cannabis legalization and acute harm from high potency cannabis products: a narrative review and recommendations for public health. Front Psychiatry 2020 Sep 23;11:591979 [FREE Full text] [doi: 10.3389/fpsyt.2020.591979] [Medline: 33173527]
10. Sabet K. Marijuana and legalization impacts. Berkeley J Crim Law 2018;23(1):84-98. [doi: 10.15779/Z381V5BD5S]
11. Smart R, Caulkins JP, Kilmer B, Davenport S, Midgette G. Variation in cannabis potency and prices in a newly legal market: evidence from 30 million cannabis sales in Washington state. Addiction 2017 Dec;112(12):2167-2177 [FREE Full text] [doi: 10.1111/add.13886] [Medline: 28556310]
12. Stuyt E. The problem with the current high potency THC marijuana from the perspective of an addiction psychiatrist. Mo Med 2018;115(6):482-486 [FREE Full text] [Medline: 30643324]
13. Degenhardt L, Ferrari AJ, Calabria B, Hall WD, Norman RE, McGrath J, et al. The global epidemiology and contribution of cannabis use and dependence to the global burden of disease: results from the GBD 2010 study. PLoS One 2013 Oct 24;8(10):e76635 [FREE Full text] [doi: 10.1371/journal.pone.0076635] [Medline: 24204649]
14. Bonnet U, Preuss UW. The cannabis withdrawal syndrome: current insights. Subst Abuse Rehabil 2017 Apr 27;8:9-37 [FREE Full text] [doi: 10.2147/SAR.S109576] [Medline: 28490916]
15. SAMHSA working definition of recovery: 10 guiding principles of recovery. Substance Abuse and Mental Health Services Administration. 2012. URL: https://store.samhsa.gov/sites/default/files/pep12-recdef.pdf [accessed 2024-09-03]
16. Budney AJ, Roffman R, Stephens RS, Walker D. Marijuana dependence and its treatment. Addict Sci Clin Pract 2007 Dec;4(1):4-16 [FREE Full text] [doi: 10.1151/ascp07414] [Medline: 18292704]
17. Denis C, Lavie E, Fatséas M, Auriacombe M. Psychotherapeutic interventions for cannabis abuse and/or dependence in outpatient settings. Cochrane Database Syst Rev 2006 Jul 19(3):CD005336. [doi: 10.1002/14651858.CD005336.pub2] [Medline: 16856093]
18. Lozano BE, Stephens RS, Roffman RA. Abstinence and moderate use goals in the treatment of marijuana dependence. Addiction 2006 Nov;101(11):1589-1597. [doi: 10.1111/j.1360-0443.2006.01609.x] [Medline: 17034438]
19. Gates P, Copeland J. Barriers to treatment seeking for cannabis dependence. In: Handbook of Cannabis and Related Pathologies. Cambridge, MA: Academic Press; 2017.
20. Kerridge BT, Mauro PM, Chou SP, Saha TD, Pickering RP, Fan AZ, et al. Predictors of treatment utilization and barriers to treatment utilization among individuals with lifetime cannabis use disorder in the United States. Drug Alcohol Depend 2017 Dec 01;181:223-228 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.09.032] [Medline: 29107786]
21. Hagman BT, Falk D, Litten R, Koob GF. Defining recovery from alcohol use disorder: development of an NIAAA research definition. Am J Psychiatry 2022 Nov 01;179(11):807-813. [doi: 10.1176/appi.ajp.21090963] [Medline: 35410494]
22. Sanger S, Bath PA, Bates J. 'Someone like me': user experiences of the discussion forums of non-12-step alcohol online support groups, June 2019. Addict Behav 2019 Nov;98:106028. [doi: 10.1016/j.addbeh.2019.106028] [Medline: 31302313]
23. Richardson GB, Blount TN, Hanson-Cook BS. Life history theory and recovery from substance use disorder. Rev Gen Psychol 2018 Dec 20;23(2):263-274. [doi: 10.1037/gpr0000173]
24. What individuals in recovery need to know about wellness. Substance Abuse and Mental Health Services Administration. 2016. URL: https://store.samhsa.gov/sites/default/files/sma16-4950.pdf [accessed 2024-09-03]
25. Greiner C, Chatton A, Khazaal Y. Online self-help forums on cannabis: a content assessment. Patient Educ Couns 2017 Oct;100(10):1943-1950. [doi: 10.1016/j.pec.2017.06.001] [Medline: 28602568]
26. Haynes K. High on marijuana anonymous. WebMD. 2022 Mar 16. URL: https://www.webmd.com/mental-health/addiction/features/high-marijuana-anonymous [accessed 2022-06-17]
27. Creedon TB, Schrader KE, O'Brien PL, Lin JR, Carroll CD, Mulvaney-Day N. Rural-nonrural differences in telemedicine use for mental and substance use disorders among medicaid beneficiaries. Psychiatr Serv 2020 Aug 01;71(8):756-764. [doi: 10.1176/appi.ps.201900444] [Medline: 32290806]
28. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 07. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2021-11-16]
29. Allem JP, Majmundar A, Dormanesh A, Donaldson SI. Identifying health-related discussions of cannabis use on Twitter by using a medical dictionary: content analysis of tweets. JMIR Form Res 2022 Feb 25;6(2):e35027 [FREE Full text] [doi: 10.2196/35027] [Medline: 35212637]

XSL•FO
RenderX

30. Benson R, Hu M, Chen AT, Zhu SH, Conway M. Examining cannabis, tobacco, and vaping discourse on Reddit: an exploratory approach using natural language processing. Front Public Health 2022 Jan 05;9:738513 [FREE Full text] [doi: 10.3389/fpubh.2021.738513] [Medline: 35071153]
31. Hu M, Benson R, Chen AT, Zhu SH, Conway M. Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing. Drug Alcohol Depend 2021 Nov 01;228:109016 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.109016] [Medline: 34560332]
32. Meacham MC, Paul MJ, Ramo DE. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. Drug Alcohol Depend 2018 Jul 01;188:364-369 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.03.041] [Medline: 29883950]
33. Sowles SJ, Krauss MJ, Gebremedhn L, Cavazos-Rehg PA. "I feel like I've hit the bottom and have no idea what to do": supportive social networking on reddit for individuals with a desire to quit cannabis use. Subst Abus 2017;38(4):477-482 [FREE Full text] [doi: 10.1080/08897077.2017.1354956] [Medline: 28704167]
34. Tamersoy A, De Choudhury M, Chau DH. Characterizing smoking and drinking abstinence from social media. HT ACM Conf Hypertext Soc Media 2015 Sep;2015:139-148 [FREE Full text] [doi: 10.1145/2700171.2791247] [Medline: 26640831]
35. Kemp S. Reddit now has more active users than Twitter — and is more engaging than porn. The Next Web. 2018 Apr 19. URL: https://thenextweb.com/news/reddit-now-active-users-twitter-engaging-porn [accessed 2022-06-24]
36. Murphy N. Reddit's 2019 year in review. Reddit. 2019 Dec 04. URL: https://www.redditinc.com/blog/reddits-2019-year-in-review/ [accessed 2022-06-24]
37. van der Nagel E. From usernames to profiles: the development of pseudonymity in internet communication. Internet Hist 2017 Oct 27;1(4):312-331. [doi: 10.1080/24701475.2017.1389548]
38. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift Reddit dataset. In: Proceedings of the Fourteenth International AAAI Conference on Web and Social Media. 2020 Presented at: ICWSM 2020; June 8-11, 2020; Atlanta, GA. [doi: 10.1609/icwsm.v14i1.7347]
39. Dean B. Reddit user and growth stats. Backlinko. URL: https://backlinko.com/reddit-users [accessed 2022-06-24]
40. Graham T, Wright S. Discursive equality and everyday talk online: the impact of "superparticipants". J Comput Mediat Commun 2014 Apr 01;19(3):625-642. [doi: 10.1111/jcc4.12016]
41. Chang T, DeJonckheere M, Vydiswaran VG, Li J, Buis LR, Guetterman TC. Accelerating mixed methods research with natural language processing of big text data. J Mix Methods Res 2021 Jun 16;15(3):398-412. [doi: 10.1177/15586898211021196]
42. Guetterman TC, Chang T, DeJonckheere M, Basu T, Scruggs E, Vydiswaran VG. Augmenting qualitative text analysis with natural language processing: methodological study. J Med Internet Res 2018 Jun 29;20(6):e231 [FREE Full text] [doi: 10.2196/jmir.9702] [Medline: 29959110]
43. Denny MJ, Spirling A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit Anal 2018 Mar 19;26(2):168-189. [doi: 10.1017/pan.2017.44]
44. Lewis DD, Yang Y, Rose TG, Li F. RCV1: a new benchmark collection for text categorization research. J Mach Learn Res 2004;5:361-397.
45. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: an R package for the quantitative analysis of textual data. J Open Source Softw 2018 Oct 06;3(30):774. [doi: 10.21105/joss.00774]
46. Hvitfeldt E, Silge J. Supervised Machine Learning for Text Analysis in R. Boca Raton, FL: CRC Press; 2021.
47. R: a language and environment for statistical computing. R Core Team. 2015 Feb 10. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2024-09-03]
48. Smith CP, Atkinson JW, McClelland DC, Veroff J. Motivation and Personality: Handbook of Thematic Content Analysis. Cambridge, UK: Cambridge University Press; 1992.
49. Chen AT, Johnny S, Conway M. Examining stigma relating to substance use and contextual factors in social media discussions. Drug Alcohol Depend Rep 2022 May 05;3:100061 [FREE Full text] [doi: 10.1016/j.dadr.2022.100061] [Medline: 36845987]
50. Cerdá M, Mauro C, Hamilton A, Levy NS, Santaella-Tenorio J, Hasin D, et al. Association between recreational marijuana legalization in the United States and changes in marijuana use and cannabis use disorder from 2008 to 2016. JAMA Psychiatry 2020 Feb 01;77(2):165-171 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.3254] [Medline: 31722000]
51. Lederman LC, Menegatos LM. Sustainable recovery: the self-transformative power of storytelling in alcoholics anonymous. J Groups Addict Recover 2011;6(3):206-227. [doi: 10.1080/1556035x.2011.597195]
52. Nickerson A, Byrow Y, Pajak R, McMahon T, Bryant RA, Christensen H, et al. 'Tell your story': a randomized controlled trial of an online intervention to reduce mental health stigma and increase help-seeking in refugee men with posttraumatic stress. Psychol Med 2019 Apr 11;50(5):781-792. [doi: 10.1017/s0033291719000606]
53. Kim HS, Bigman CA, Leader AE, Lerman C, Cappella JN. Narrative health communication and behavior change: the influence of exemplars in the news on intention to quit smoking. J Commun 2012 Jun 01;62(3):473-492 [FREE Full text] [doi: 10.1111/j.1460-2466.2012.01644.x] [Medline: 22736808]
54. Gates P, Albertella L, Copeland J. Cannabis withdrawal and sleep: a systematic review of human studies. Subst Abus 2016;37(1):255-269. [doi: 10.1080/08897077.2015.1023484] [Medline: 25893849]

XSL•FO
RenderX

55. Clark TM, Jones JM, Hall AG, Tabner SA, Kmiec RL. Theoretical explanation for reduced body mass index and obesity rates in cannabis users. Cannabis Cannabinoid Res 2018 Dec 21;3(1):259-271 [FREE Full text] [doi: 10.1089/can.2018.0045] [Medline: 30671538]
56. Osborne GB, Fogel C. Understanding the motivations for recreational marijuana use among adult Canadians. Subst Use Misuse 2008;43(3-4):539-72; discussion 573-9, 585-7. [doi: 10.1080/10826080701884911] [Medline: 18365950]
57. Walton-Moss B, Ray EM, Woodruff K. Relationship of spirituality or religion to recovery from substance abuse: a systematic review. J Addict Nurs 2013;24(4):217-26; quiz 227-8. [doi: 10.1097/JAN.0000000000000001] [Medline: 24335768]
58. National coalition for the homeless homepage. National Coalition for the Homeless. URL: https://nationalhomeless.org/ [accessed 2023-01-26]
59. Park A, Conway M. Tracking health related discussions on reddit for public health applications. AMIA Annu Symp Proc 2018 Apr 16;2017:1362-1371 [FREE Full text] [Medline: 29854205]
60. Ming Y, Matteson ML, Sun J. Reddit use in the LIS community. J Librariansh Inf Sci 2021 Oct 19;54(4):770-778. [doi: 10.1177/09610006211052084]
61. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying reddit: a systematic overview of disciplines, approaches, methods, and ethics. Soc Media Soc 2021 May 26;7(2):205630512110190. [doi: 10.1177/20563051211019004]
62. Gaffney D, Matias JN. Caveat emptor, computational social science: large-scale missing data in a widely-published Reddit corpus. PLoS One 2018 Jul 6;13(7):e0200162 [FREE Full text] [doi: 10.1371/journal.pone.0200162] [Medline: 29979741]
63. Chandra K. Counting is hard: probabilistic algorithms for view counting at Reddit. In: Proceedings of the QCon San Francisco 2018. 2018 Presented at: QCon SF 2018; November 5-7, 2018; San Francisco, CA.
64. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with Reddit data. Soc Sci Comput Rev 2019 Dec 18;39(5):943-960. [doi: 10.1177/0894439319893305]

Abbreviations

API: application programming interface
CUD: cannabis use disorder
SAMHSA: Substance Abuse and Mental Health Services Administration

Edited by A Mavragani; submitted 16.03.23; peer-reviewed by J Thrul, D Carvalho, M Meacham; comments to author 22.09.23; revised version received 01.01.24; accepted 13.05.24; published 27.09.24.

Please cite as:
Thulin EJ, Walton MA, Bonar EE, Fernandez A
Examining the Popularity, Content, and Intersections With the Substance Abuse and Mental Health Services Administration's Definition of Recovery in a Nonclinical Online Cannabis Cessation Community: Infodemiology Study of Reddit Posts
J Med Internet Res 2024;26:e47357
URL: https://www.jmir.org/2024/1/e47357
doi:10.2196/47357
PMID:39331460

©Elyse J Thulin, Maureen A Walton, Erin E Bonar, Anne Fernandez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding e-Cigarette Addictiveness: Triangulation of Focus Group and Netnographic Data

Marike Andreas[1], MSc; Nadja Grundinger[2], MSc; Nadine Wolber[1], BSc; Daria Szafran[1], MA; Tatiana Görig[3], DrScHum; Ute Mons[4], Dr sc hum; Valerie Lohner[4], PhD; Sabine Vollstädt-Klein[2,5], Dr sc hum; Sven Schneider[1], Dphil

[1]Division of Public Health, Social and Preventive Medicine, Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Department of Addictive Behavior and Addiction Medicine, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[3]Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Erlangen, Germany

[4]Cardiovascular Epidemiology of Aging, Department of Cardiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[5]Mannheim Center for Translational Neurosciences, Medical Faculty of Mannheim, Heidelberg University, Mannheim, Germany

Corresponding Author:
Marike Andreas, MSc
Division of Public Health, Social and Preventive Medicine
Center for Preventive Medicine and Digital Health
Medical Faculty Mannheim, Heidelberg University
Ludolf-Krehl-Straße 13-17
Mannheim, 68167
Germany
Phone: 49 621 383 71814
Email: marikeandreas@gmail.com

Abstract

Background: Numerous studies have shown that e-cigarettes are addictive. For example, we previously showed that users of e-cigarette online forums discuss experiences of addiction in a netnographic analysis. However, it is unclear what makes e-cigarettes addictive apart from nicotine. In a focus group analysis, we recently identified 3 unique features of e-cigarettes that users linked to experiences of addiction: the pleasant taste, unobtrusiveness, and unlimited usability of e-cigarettes.

Objective: This study aimed to validate the previously identified features of e-cigarette addictive potential by triangulating data from the netnographic analysis and focus group discussions.

Methods: Drawing on a netnographic analysis of 3 popular, German-language e-cigarette forums, we studied whether experiences of addiction were linked to specific e-cigarette features. We included 451 threads in the analysis that had been coded for addictive experiences in a previous study by our team. First, we conducted a deductive analysis with preregistered codes to determine whether the features of pleasant taste, unobtrusiveness, and unlimited usability were mentioned in relation to the addictive potential of e-cigarettes in the online forums. Second, an inductive approach was chosen to identify further possible addictive features of e-cigarettes.

Results: Our deductive analysis confirmed that the features highlighted in our previous focus group study (pleasant taste, unobtrusiveness, and unlimited usability) were also frequently discussed in online forums in connection to addictive symptoms. In addition, our inductive analysis identified nicotine dosage as a significant feature linked to addiction. Users reported varying their nicotine doses for different reasons, leading to the identification of four distinct user types based on dosing patterns: (1) high doses for intermittent, (2) high doses for constant use, (3) low doses for constant use, and (4) switching between high and low doses depending on the situation.

Conclusions: Our comprehensive analysis of online forum threads revealed that users' experiences of addiction are linked to 4 specific features unique to e-cigarettes: pleasant taste, unobtrusiveness, unlimited usability, and nicotine dosage. Recognizing these addictive features of e-cigarettes is crucial for designing cessation programs and informing public health policies to reduce the addictiveness of e-cigarettes.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57970) doi:10.2196/57970

KEYWORDS

e-cigarettes; online forums; netnographic analysis; addictive; addiction; smoking cessation; smoker; user; focus group; nicotine; public health; prevalence; smoking behavior

Introduction

Electronic cigarettes, also known as e-cigarettes, are devices powered by batteries. They function by heating a liquid mixture, which usually includes nicotine and flavor, to produce an aerosol that users inhale [1]. e-Cigarettes are often used as a tool to quit smoking, with evidence showing that they are more effective than traditional nicotine replacement products [2]. However, many individuals who successfully quit smoking using these devices tend to continue vaping in the long term [2]. Indeed, research has indicated that e-cigarette users experience addiction, with studies revealing notable levels of nicotine dependence among e-cigarette users [3,4]. Moreover, the uptake of e-cigarettes among young people who have never smoked has risen [5], indicating that e-cigarettes are attractive beyond their use as smoking cessation aids. Thus, understanding the factors driving the attractiveness and addictiveness of e-cigarettes is crucial for improving prevention and cessation strategies.

In a comprehensive netnographic analysis of e-cigarette online forums, we previously examined self-reports of users in online forums using a netnographic approach. Some, but not all, users reported subjective experiences that met the criteria for a tobacco use disorder as classified by the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition* (*DSM-5*) [6]. In a further analysis of focus groups, we identified 3 unique aspects of e-cigarettes that were related to their addictiveness according to e-cigarette users: their pleasant taste, unobtrusiveness, and unlimited usability [7]. The aspect of pleasant taste has frequently been highlighted as the central component of the addictiveness of e-cigarettes. For example, mint flavor has been reported to be more addictive than other flavors by young adult e-cigarette users [8], while fruity flavors are perceived as less of a health risk [9]. The second unique characteristic, unobtrusiveness, describes the possibility of using the e-cigarette almost unnoticed without producing ash or unpleasant smells. This makes e-cigarettes more attractive compared with tobacco products [7]. Furthermore, since the user can control the size and density of vapor clouds, unobtrusiveness encourages using e-cigarettes even in places where it is not allowed to vape [10,11]. According to the results of our focus group study, this feature is associated with a higher frequency of use [7]. Likewise, the third e-cigarette characteristic identified in the focus groups, unlimited usability, was also associated with a high frequency of use and automaticity. Depending on their battery, tank size, and behavior of their user, e-cigarettes can be used for up to 20 hours or 800 puffs and more. Because they do not have to be lit like tobacco cigarettes, users can vape nonstop. The “ready availability” has also been identified by Kim et al [12] as a unique feature of e-cigarettes.

The 3 unique e-cigarette features are depicted in the MAPE (Model of the Addictive Potential of E-cigarettes) model (Figure 1 [7,13]). The development of the MAPE model was based on a qualitative approach in the form of focus group interviews [7]. Using diverse methods of data collection can significantly enhance the depth of understanding regarding the subject under investigation [14]. To validate the MAPE model, we opted for a netnographic approach, in which ethnographic techniques are applied to online social interactions [15]. An advantage of this approach is its capacity to mitigate various sources of bias prevalent in the field, such as selection, groupthink, interviewer, and social desirability effects [16]. In the discussion part of this paper, we will triangulate the findings of the qualitative and netnographic approaches to elaborate on what makes e-cigarettes attractive and addictive. A better understanding of what draws users to these products would support the development of strategies to prevent initiation among nonusers and to promote cessation. Finally, comprehending the addictive nature of e-cigarettes can inform public health policies aimed at reducing their appeal and mitigating long-term health risks.

XSL•FO
RenderX

Figure 1. The model of the addictive potential of e-cigarettes (MAPE) is based on focus group interviews with e-cigarette users.

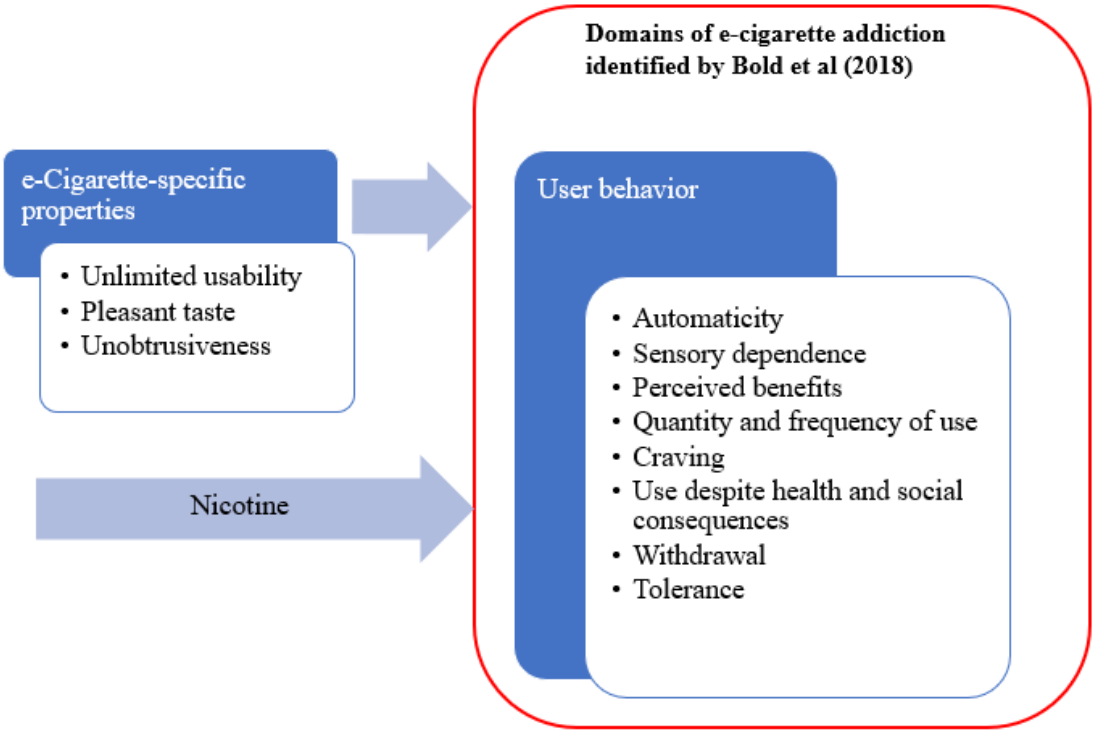

Methods

Study Design and Setting

Since many e-cigarette users discuss their e-cigarette use in online forums [17,18], we chose netnographic analysis as the method for our analysis. Analyzing online discourse allows an understanding of the consumer's perspective without the influence of the social desirability bias to please the researcher [16]. Especially sensitive topics, such as health status or addiction, are often discussed in the anonymous online setting [19], which makes the online setting appropriate for studying the addictive potential of e-cigarettes.

Data Collection

This study is based on results from 2 previous studies from the EVAPE (Evaluation of the Addictive Potential of E-cigarettes) project. In the first study posts in 3 large, German-speaking, e-cigarette online forums were scanned for addiction-specific keywords and subsequently extracted. The keyword search in all forums returned 5337 threads (database lock on April 9, 2021). Of the 5337 threads initially screened in the 3 forums, 451 contained relevant information regarding addiction and were included in the data analysis. Within these posts, self-reported signs of addiction were coded based on the *DSM-5* criteria for substance use disorder [6]. For a full description of data collection within the forums, please see our previous publication [6]. We found that users reported typical symptoms of addiction such as excessive time spent vaping, craving, and continued use despite health issues–related to e-cigarettes. In the second study of the EVAPE project, we conducted focus group discussions to identify e-cigarette–specific features related to addiction symptoms [7]. In 5 focus groups with 14 e-cigarette users overall, we asked users about e-cigarette–specific addiction symptoms. Analysis of the addiction experiences reported in the focus groups was based on the 10 domains of e-cigarette addiction previously identified by Bold et al [13]. We then inductively added codes for the e-cigarette properties that were identified as addictive by users. In this paper, we aimed to triangulate the results of the 2 studies and investigate whether the previously identified self-reported signs of addiction according to the MAPE model are linked to e-cigarette–specific features by the users of e-cigarette online forums (Figure 2 [6,7]).

Figure 2. Overview of the research process to triangulate findings from netnographic analysis of e-cigarette online forums (study I) and focus group discussions with e-cigarette users (study II).

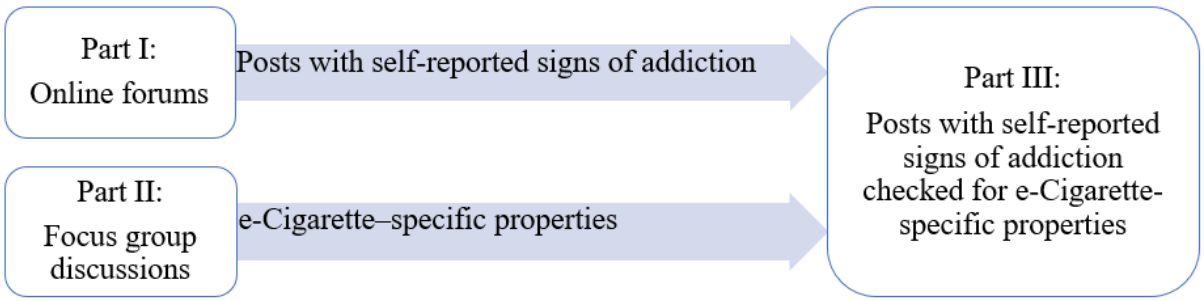

XSL•FO
RenderX

Data Analysis

We used the analysis software MAXQDA (VERBI Software) [20] for content analysis based on the approach by Clarke et al [21]. To evaluate the connection between e-cigarette–specific properties and perceived addiction symptoms, every code-identifying addiction symptom in our previous study [6] was evaluated again to check whether an e-cigarette feature was linked to the experience of addiction. The e-cigarette–specific properties from the MAPE model constituted the deductive coding system (Table 1). A post was only coded when the reported addiction symptom was linked to a specific e-cigarette feature. The deductive codes were registered beforehand on Open Science Framework storage [22]. Two authors (MA and NW) independently double-coded the posts.

In the second step, we used an inductive approach to identify e-cigarette–specific characteristics not covered by the MAPE model, namely, different user types based on the combination of nicotine dosage and frequency of use. Authors MA and NW agreed on the new code and also independently double-coded it. Agreement between the authors was then tested for all codes (deductive and inductive) using MAXQDA's tool for intercoder agreement resulting in a code intersection rate of 87% at the segment level, indicating a substantial level of intercoder agreement. The authors discussed disagreements and, in the few cases of persistent disagreement, a third author was consulted (SS) to reach a consensus. Finally, all quotes were anonymized and forward-translated from German to English.

Table 1. Codebook for the deductive coding process of identifying e-cigarette properties linked to addictiveness in user posts.

e-Cigarette property	User description for coding
Unlimited usability	Users describe that unlimited usability leads to experiences of addiction symptoms. For example, long battery life, large tanks and not having to light up are indicators of the unlimited usability of e-cigarettes.
Pleasant taste	Users describe that the pleasant taste and smell of e-cigarette liquids lead to experiences of addiction symptoms. For example, favorite liquids could be discussed.
Unobtrusiveness	Users describe that unobtrusiveness leads to experiences of addiction symptoms. Features of unobtrusiveness include those e-cigarettes can be used almost everywhere even in places where it is prohibited (stealthing), they do not leave waste (eg, ash), and they do not leave an unpleasant smell.

Ethical Considerations

Our study design was preregistered on the Open Science Framework and detailed in a previous publication to ensure transparency [22,23]. The Medical Ethics Committee of the Medical Faculty Mannheim at the University of Heidelberg approved our methodology for the netnographic and qualitative study (2017-567-N-MA). For the netnographic analysis, the forums' terms and conditions informed users about the public nature of their posts. In addition, to protect user anonymity, we selected forums with a large number of active users and ensured no personal information was documented. This approach aligns with ethical guidelines for online research [24]. Participants in focus groups signed a consent form enabling us to use their anonymized data for research purposes.

Results

Deductive Codes

In the following, the deductive codes based on the MAPE model are listed.

Pleasant Taste

The pleasant taste of e-cigarette liquids was the characteristic discussed most frequently by users in relation to addiction. For example, users wrote about their favorite liquid or which liquids they enjoyed in which situations. Many users highlighted the taste as their favorite aspect of vaping, or "Taste- the holy grail of vapers!" as one user stated. Liquid taste was often mentioned as a reason to vape more, which led to self-reported craving in some users:

> *Sometimes the taste is so good - I just can't get enough. Or I think about what I want to try next: another liquid or a mixture of the current and another one - so I have to use up my tank quickly.*

Especially sweet flavors were mentioned as addictive and were sometimes compared to candy:

> *When I crave a specific sweet, I just choose the matching liquid. For example, when I want to eat chocolate, I vape chocolate liquid.*

Unobtrusiveness

The fact that e-cigarette vapor does not leave a lingering smell and generates no waste, such as ash, was named as a reason for e-cigarette addictiveness. This characteristic was often compared with cigarettes:

> *I feel like I crave vaping constantly. I vape whenever possible. I enjoy vaping because there is no ash or smell in contrast to tobacco smoking.*

The unobtrusiveness of e-cigarettes was also identified as one reason for e-cigarette use at home, at the office, or in the car:

> *When I was still smoking, I did not smoke in my flat because I find that disgusting. Now I vape in my flat and chain vape when I'm sitting in front of the TV.*

This example suggests that the possibility of e-cigarette use at home also led to an increased frequency of vaping. Finally, not only e-cigarette users but also the people surrounding them seem to view vaping as less obtrusive than tobacco second-hand smoke. One user stated:

> *In the office, I vape non-stop because my colleague is not annoyed by that. So most of the time I have my e-cigarette in my mouth.*

XSL•FO
RenderX

Again, this user behavior was often contrasted with previous experiences with tobacco smoking:

> *More and more often I think I vape too much. One disadvantage is that I can vape everywhere: At work, at home, visiting my friends. This was not the case when I was still smoking.*

Some users also admitted to "stealthing", a term used in the e-cigarette community to describe vaping in places where it is forbidden to vape or smoke.

Unlimited Use

e-Cigarette users mentioned the possibility of using their e-cigarettes for a long time as one reason for nonstop vaping and cravings. Unlimited use was often discussed in the context of improving this feature by buying new equipment. One user reported:

> *I was looking for hardware with long battery life because I had to recharge my former e-cigarette twice a day when vaping constantly. I found an e-cigarette model, which I only have to recharge after three days. So, I can vape constantly now.*

The option to use e-cigarettes for a long period was also contrasted with previous cigarette smoking. Another user reported:

> *Imitating my former cigarette smoking is not possible, because there is not "one cigarette" to smoke. So now, I chain vape. And I don´t like that.*

Inductive Codes

One new code was identified during the inductive coding process: the ability to choose nicotine dosage in relation to craving.

Nicotine Dosage

The possibility of choosing liquids with user-specific nicotine doses or mixing their own liquids was a recurring topic in the forums. This feature was related to experiencing addiction and also trying to combat addiction symptoms. We identified 4 user types who choose their nicotine doses according to different rationales. A summary of the 4 user types is provided in Figure 3.

Figure 3. Visualization of 4 different user types identified in the inductive analysis of posts in e-cigarette forums.

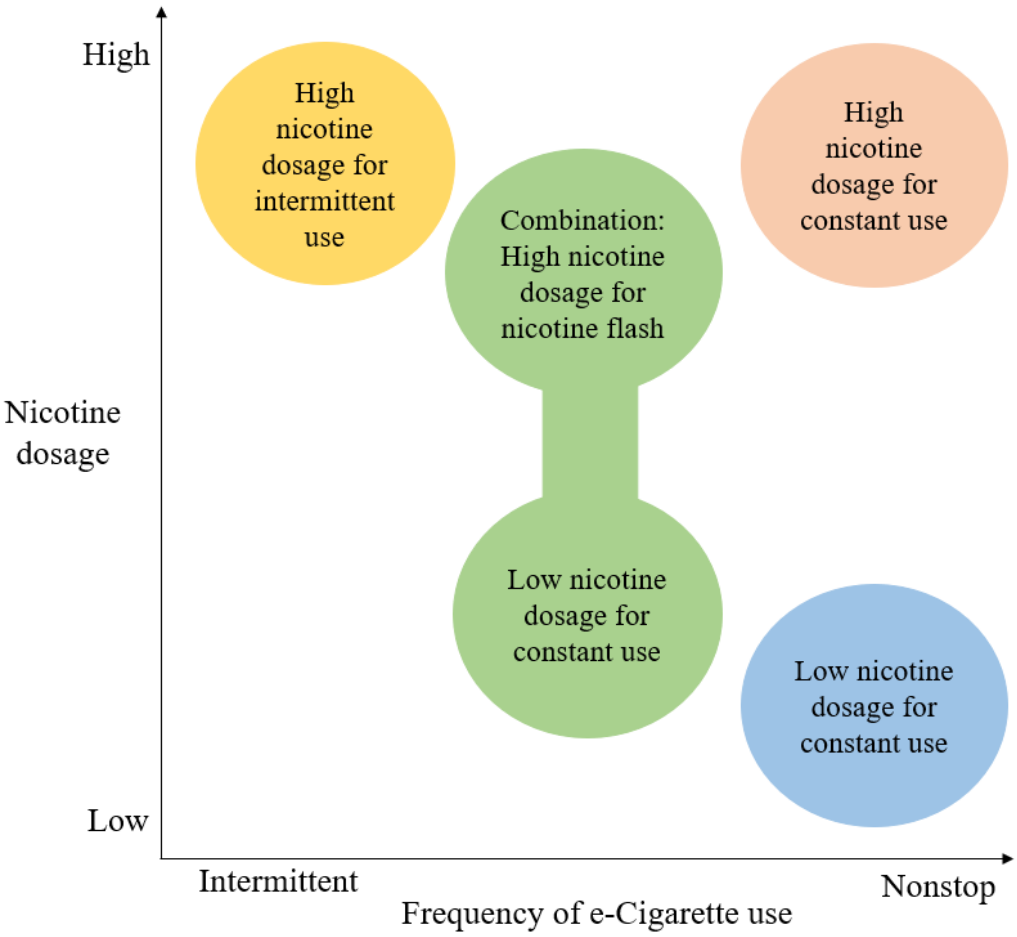

Low Nicotine Dosages for Constant Use

These users reported vaping e-cigarettes with low nicotine doses in order to be able to use their e-cigarettes nonstop:

> *At the beginning, I vaped 6 mg. For me as a chain-vaper, that was too much: I got dizzy. Nowadays 3 mg and everything is fine.*

Typically, the nicotine doses mentioned by these users range from 1 mg to 5 mg. Often, users reported vaping lower doses

XSL•FO
RenderX

to counteract negative health effects or make them more tolerable. For instance, one user posted:

> *At the end of the day I got a headache. I decided to dose down the nicotine in my e-cigarette, so I am able to vape constantly again.*

Consequently, this code was often found in relation to the addiction symptom of e-cigarette use despite negative health effects.

High Nicotine Dosages for Intermittent Use

In contrast to users who use low nicotine doses for constant use, these users choose liquids with high nicotine doses in order to not vape constantly. For instance, one user reported, "I switched from 12 mg to 18 mg for controlling and reducing my chain-vaping." Thus, this user type aims to control their usage behavior through higher doses.

High Nicotine Dose for Constant Use

These e-cigarette users regularly use very high nicotine doses to feel a nicotine flash. They often mentioned having developed a tolerance for lower nicotine doses, as this example demonstrates:

> *I started with 18 mg nicotine. That was not enough for me to get a flash, so I switched to 24 mg. I don't want less nicotine. I have been vaping 24 mg in the last 3 years and I am happy with it.*

Another user reported:

> *I vape 18 mg in the morning with my coffee and as a nicotine boost during the day (…). In the evening on the sofa I then switch to 9 or 12er fruity liquid - in almost continuous vaping mode.*

Therefore, this user group combats tolerance symptoms with higher doses to satisfy their addiction.

High and Low Nicotine Dose in Combination

Some vapers also reported using an e-cigarette with a moderate nicotine dose in combination with a second one with a very high dose: "I have an emergency device with 18 mg. Ready for action." Often, the different e-cigarettes are meant for different situations, as this quote exemplifies:

> *When I know there is no time for vaping, I put 18 mg in my (hardware), 15-18 Watt-4 puffs and I am done. By default, I use 9 mg and 15 Watt for chain-vaping.*

The e-cigarette with higher nicotine concentration is thus used to experience a nicotine flash.

Discussion

Principal Findings

In this study, using data triangulation, we showed that 3 specific characteristics of e-cigarettes identified in the MAPE model were also observed in e-cigarette online forums. The pleasant taste, unlimited usability, and unobtrusiveness of e-cigarettes were discussed by forum members in relation to addictive symptoms. Similarly to the focus groups previously conducted as part of the EVAPE project with e-cigarette users [7], users in the e-cigarette online forums reported that the 3 aspects highlighted in the MAPE model were attractive and thus addictive, therefore, validating the MAPE model in a different setting. Moreover, in contrast to the focus groups, the netnographic context allowed us to access user experiences firsthand without an interviewer bias. In addition, we were able to identify one further unique e-cigarette feature: nicotine dosage. The updated MAPE model is depicted in Figure 4.

Figure 4. The updated model of the addictive potential of e-cigarettes is based on data triangulation.

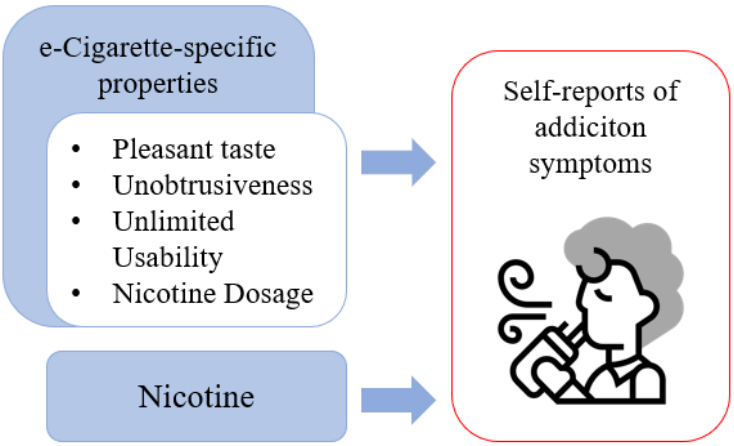

Triangulation

In addition to the 3 distinct features of e-cigarettes (pleasant taste, unobtrusiveness, unlimited use) that we previously identified in a focus group study, the netnographic analysis revealed one additional feature associated with e-cigarette addictiveness. In this study, nicotine dosage emerged as a prominent e-cigarette feature frequently discussed in relation to addiction. The analysis of forum discussions on dosage revealed the existence of 4 distinct patterns among users. Some users reported opting for low nicotine doses in order to use their e-cigarettes continuously, while a second group preferred high doses to reduce the frequency of vaping. In a laboratory study, Dawkins et al [25] compared the user behavior of their participants on 2 different days, being given an e-cigarette with high (24 mg/mL) one day and low (6 mg/mL) nicotine concentration liquids the other day. Vaping the low nicotine concentration, users puffed more frequently, longer, and nearly

XSL•FO
RenderX

doubled the amount of liquid consumed. This is in line with user reports analyzed in our study, showing that low nicotine doses were related to constant use. However, our analysis extends beyond controlled laboratory settings, revealing that vapers actively select low or high nicotine doses depending on the user behavior they prefer. It is possible that there are different motives behind the usage behaviors. The fact that one user group likes to use their e-cigarette nonstop, but higher nicotine doses lead to unpleasant effects, could indicate that these users are motivated by psychological dependence, for example, habituation or sensory aspects (throat hit, stimulation of the respiratory tract), rather than a purely physiological dependence on nicotine. It is already known that nonpharmacological aspects of tobacco cigarette smoking reduce the urge to smoke even more effectively than the direct pharmacological effect of nicotine alone [26,27]. Conversely, people who use high nicotine concentrations but vape less may be driven more by nicotine intake or physiological withdrawal symptoms. Our study indicates that vapers who use high nicotine concentrations, self-titrate to control addiction symptoms, meaning they adjust their nicotine intake either by different use behavior or by adjusting nicotine levels. This behavior helps to effectively manage craving and withdrawal symptoms to reduce constant vaping [28].

Furthermore, we identified 2 groups of forum users, who used liquids with very high nicotine concentrations to experience a nicotine flash. One group exhibited symptoms indicative of nicotine tolerance, leading them to consistently vape liquids with high nicotine doses in order to experience a nicotine flash. This is in line with a retrospective survey study by Browne and Todd [29], who found that e-cigarette users who have been vaping for a longer time used a higher nicotine dosage than short-term users. Longitudinal studies of nicotine dosing changes over time could provide further insights into this pattern. In contrast, another group used liquids with high nicotine concentrations next to lower-dosed liquids to experience a nicotine flash. This was done in specific situations that were perceived as stressful or in which nonstop vaping was not possible. For dual–nicotine format users, Perry et al [30] observed that craving for tobacco cigarettes compared with e-cigarettes increased in stressful situations. Our study indicates that single–nicotine format users experienced cravings in such situations and equip themselves with an “emergency” high nicotine dose device to alleviate their withdrawal symptoms. Earlier studies have shown that many smokers mistakenly believe cigarettes reduce stress when in reality, smoking only alleviates nicotine withdrawal symptoms [31]. Thus, some vapers may perceive high-nicotine liquids as more effective for stress relief because they are more effective at reducing withdrawal symptoms, resulting in a temporary improvement in mood and state of mind, but a long-term increase in stress levels.

The aspect of nicotine dosage and its relation to addiction is not yet well understood in e-cigarette users. A review of 104 studies indicates that increased nicotine concentrations are linked to heightened abuse potential and appeal [32]. While some studies compare puff durations between e-cigarette users [33], no study has yet investigated differences and rationales for self-chosen nicotine concentration. Our analysis implies that e-cigarette users select nicotine doses according to their preferred use behavior. For example, some users use high-nicotine e-cigarettes occasionally for a quick nicotine flash, while others use low-nicotine e-cigarettes more frequently throughout the day. In turn, some users have multiple devices for different situations and choose between them depending on their current nicotine craving or social context. This is consistent with the findings of a previous study of puff topography in natural environments [34]. User reports collected here suggest that the opportunity to select nicotine dosage might thus be a pathway to facilitating nicotine dependency and tolerance. Given the diverse rationales observed in individuals’ choice of nicotine doses and the indications of self-titration found in this study, future research should delve deeper into the relationship between self-chosen nicotine concentration and addiction in e-cigarette users.

Implications for Practice

The updated MAPE model can help inform both preventive and cessation programs for e-cigarette users. Our research demonstrated that the pleasant taste of e-cigarettes is a crucial factor for the enjoyment of vaping and related addiction. Accordingly, banning certain liquid flavors might be an option to reduce the attractiveness of these products. Currently, 6 countries ban e-cigarette liquids with flavors other than tobacco to make vaping less attractive for young people [35]. Notably, in Germany, where our study was conducted, such a ban is not currently in place. Our data suggest that implementing a ban on liquid flavors might mitigate the addictive potential of e-cigarettes, but evidence is currently lacking in reports on the effects of such a ban [36]. Evidence from other countries indicates that a ban on flavored e-cigarettes alone may increase the consumption of tobacco cigarettes or other flavored electronic nicotine delivery systems that are exempt from the restrictions, such as disposable e-cigarettes [37]. In addition, most respondents appeared to continue to use banned flavors despite the restrictions [38]. This illustrates that the effectiveness of such regulations should not be considered independently of other tobacco or nicotine products and their user groups. Since Germany has weak tobacco regulations overall [39], tobacco control should be strengthened on all devices to effectively safeguard public health.

In addition, the ability to use e-cigarettes for an extended duration was associated with increased usage frequency and automatic use, indicative of addiction. For individuals seeking to quit or reduce vaping, physicians could be advised to recommend the use of a puff counter, a device that displays the number of puffs taken within a specific timeframe. In light of our findings, such a device could assist users to regain control over their e-cigarette consumption. The other unique aspects of e-cigarettes (unobtrusiveness and the ability to control nicotine dosage) are more challenging to regulate. These features make it difficult for regulatory measures to effectively address these aspects of e-cigarette use. The different user preferences for nicotine doses found in this study also highlight that this aspect needs to be further investigated to better understand the association between nicotine dosage, duration of use, and the development of addiction.

XSL•FO
RenderX

Limitations

An in-depth discussion of applying a netnographic approach to e-cigarette online forums can be found in a previous paper by Szafran et al [6]. In brief, applying a netnographic approach potentially mitigates biases associated with personal or group interviews, such as interviewer or social desirability biases [40]. Nonetheless, a bias might still remain, as research has shown that forum users selectively disclose information to project a favorable image of themselves to their peers [41]. In addition, our dataset encompasses a specific subgroup of e-cigarette users whose overall positive outlook toward e-cigarettes may have influenced the underreporting of addiction symptoms. In addition, we cannot completely exclude the possibility that comments may have been influenced by tobacco company marketing communications, although this is prohibited by the forum's terms of use. Furthermore, we used terms derived from *DSM-5* criteria for tobacco use disorder to search for user posts in the online forums, thus e-cigarette–specific addiction symptoms such as automaticity may not have been included.

Another limitation of this research approach is that we have no data on user characteristics, such as age or gender, as well as usage characteristics, such as frequency of consumption or total consumption. This also applies to the potential inclusion of posts from dual users, who use both e-cigarettes and tobacco. While we attempted to exclude posts explicitly mentioning dual use, it was impossible to filter out all such content as it often was not reported. Previous research has shown that dual users tend to exhibit higher levels of addiction compared with exclusive e-cigarette users [42], which should be factored into result interpretation. Furthermore, we could not distinguish between users who used nicotine-containing liquids from those who did not, and in cases where nicotine was used, we lack data on nicotine concentration in the liquids used. This was not true for the codes on dosage, however, where users shared specific nicotine doses with each other.

Conclusion

In the MAPE model, we previously summarized the distinctive aspects that make e-cigarettes addictive: their pleasant taste, unobtrusiveness, and unlimited usability. This netnographic analysis enhances the MAPE model by incorporating insights from e-cigarette online forums, revealing the additional dimension of nicotine dosage. The patterns of nicotine dosage, for example, low concentrations for continuous use or high concentrations for intermittent use, reflect the different rationales among users.

These findings have important implications for policy and practice. The role of e-cigarettes' pleasant taste for experiences of addiction highlights the potential role of flavor bans to mitigate addiction. The association between unlimited use and self-reported addiction suggests the potential utility of puff counters in cessation efforts.

In essence, our study contributes a nuanced understanding of what makes e-cigarettes addictive, emphasizing user perspectives on e-cigarettes' unique features. This qualitative exploration also highlighted the need to further explore user rationales for nicotine dosage.

Acknowledgments

The icon in Figure 4 was made by Eucalyp from the Flaticon. We thank the anonymous reviewers for commenting on and improving this paper.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available to ensure user anonymity, but further details may be available from the corresponding author on reasonable request.

Conflicts of Interest

None declared

References

1. WHO report on the global tobacco epidemic 2021: addressing new and emerging products. World Health Organization. 2021 Jul 27. URL: https://www.who.int/publications/i/item/9789240032095 [accessed 2024-09-25]
2. Hartmann-Boyce J, Begh R, Aveyard P. Electronic cigarettes for smoking cessation. BMJ 2022;360(11):j5543. [doi: 10.1136/bmj.j5543] [Medline: 29343486]
3. Lohner V, McNeill A, Schneider S, Vollstädt-Klein S, Andreas M, Szafran D, et al. Understanding perceived addiction to and addictiveness of electronic cigarettes among electronic cigarette users: a cross-sectional analysis of the international tobacco control smoking and vaping (ITC 4CV) England survey. Addiction 2023;118(7):1359-1369 [FREE Full text] [doi: 10.1111/add.16162] [Medline: 36772958]
4. Shiffman S, Sembower MA. Dependence on e-cigarettes and cigarettes in a cross-sectional study of US adults. Addiction 2020;115(10):1924-1931 [FREE Full text] [doi: 10.1111/add.15060] [Medline: 32196810]
5. Jackson SE, Tattan-Birch H, Shahab L, Oldham M, Kale D, Brose L, et al. Who would be affected by a ban on disposable vapes? A population study in Great Britain. Public Health 2024;227:291-298 [FREE Full text] [doi: 10.1016/j.puhe.2023.12.024] [Medline: 38267284]

XSL•FO
RenderX

6. Szafran D, Görig T, Vollstädt-Klein S, Grundinger N, Mons U, Lohner V, et al. Addictive potential of e-cigarettes as reported in e-cigarette online forums: netnographic analysis of subjective experiences. J Med Internet Res 2023;25:e41669 [FREE Full text] [doi: 10.2196/41669] [Medline: 36607713]
7. Andreas M, Grundinger N, Wolber N, Szafran D, Mons U, Vollstädt-Klein S, et al. Subjective experiences of the addictive potential of e-cigarettes: results from focus group discussions. Addiction Research & Theory 2023:1-8. [doi: 10.1080/16066359.2023.2288831]
8. Sargent JD, Stoolmiller M, Dai H, Barrington-Trimis JL, McConnell R, Audrain-McGovern J, et al. First e-cigarette flavor and device type used: associations with vaping persistence, frequency, and dependence in young adults. Nicotine Tob Res 2022;24(3):380-387 [FREE Full text] [doi: 10.1093/ntr/ntab172] [Medline: 34460934]
9. Strombotne K, Buckell J, Sindelar JL. Do JUUL and e-cigarette flavours change risk perceptions of adolescents? Evidence from a national survey. Tob Control 2021;30(2):199-205 [FREE Full text] [doi: 10.1136/tobaccocontrol-2019-055394] [Medline: 32300029]
10. Sanchez S, Kaufman P, Pelletier H, Baskerville B, Feng P, O'Connor S, et al. Is vaping cessation like smoking cessation? A qualitative study exploring the responses of youth and young adults who vape e-cigarettes. Addict Behav 2021;113:106687. [doi: 10.1016/j.addbeh.2020.106687] [Medline: 33045643]
11. Simpson KA, Kechter A, Schiff SJ, Braymiller JL, Yamaguchi N, Ceasar RC, et al. Characterizing symptoms of e-cigarette dependence: a qualitative study of young adults. BMC Public Health 2021;21(1):959 [FREE Full text] [doi: 10.1186/s12889-021-10945-z] [Medline: 34016066]
12. Kim H, Davis AH, Dohack JL, Clark PI. e-Cigarettes use behavior and experience of adults: qualitative research findings to inform e-cigarette use measure development. Nicotine Tob Res 2017;19(2):190-196 [FREE Full text] [doi: 10.1093/ntr/ntw175] [Medline: 27613944]
13. Bold KW, Sussman S, O'Malley SS, Grana R, Foulds J, Fishbein H, et al. Measuring e-cigarette dependence: initial guidance. Addict Behav 2018 Apr;79:213-218 [FREE Full text] [doi: 10.1016/j.addbeh.2017.11.015] [Medline: 29174664]
14. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
15. Kozinets RV. On netnography: initial reflections on consumer research investigations of cyberculture. ACR North American Advances 1998 [FREE Full text] [doi: 10.1093/oso/9780198796978.003.0007]
16. Ulusoy E. Revisiting The Netnography: Implications For Social Marketing Research Concerning Controversial And/Or Sensitive Issues. In: Marketing Dynamism & Sustainability: Things Change, Things Stay the Same. United States: Springer; 2015:422-425.
17. Maloney SF, Soule EK, Palafox S, McFadden K, Guy MC, Eissenberg T, et al. A longitudinal analysis of electronic cigarette forum participation. Addict Behav 2019;91:75-81. [doi: 10.1016/j.addbeh.2018.08.006]
18. Struik L, Yang Y. e-Cigarette cessation: content analysis of a quit vaping community on reddit. J Med Internet Res 2021;23(10):e28303 [FREE Full text] [doi: 10.2196/28303] [Medline: 34694229]
19. Langer R, Beckman SC. Sensitive research topics: netnography revisited. Qualitative Market Research: An International Journal 2005;8(2). [doi: 10.1108/13522750510592454]
20. VERBI Software. MAXQDA. Berlin: VERBI Software; 2022. URL: https://de.linkedin.com/company/verbi-software [accessed 2024-08-30]
21. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Qualitative psychology: a practical guide to research methods. London: SAGE Publications; 2015:222-248.
22. Schneider S, Görig T, Szafran D, Vollstädt-Klein S, Grundinger N, Fischer R. Addiction potential of e-cigarettes: the sociological perspective. Open Science Framework. 2021. URL: https://osf.io/dxgya [accessed 2024-08-30]
23. Vollstädt-Klein S, Grundinger N, Görig T, Szafran D, Althaus A, Mons U, et al. Study protocol: evaluation of the addictive potential of e-cigarettes (EVAPE): neurobiological, sociological, and epidemiological perspectives. BMC Psychol 2021;9(1):181 [FREE Full text] [doi: 10.1186/s40359-021-00682-8] [Medline: 34794514]
24. Roberts LD. Ethical issues in conducting qualitative research in online communities. Qualitative Research in Psychology 2015;12(3):314-325. [doi: 10.1080/14780887.2015.1008909]
25. Dawkins L, Cox S, Goniewicz M, McRobbie H, Kimber C, Doig M, et al. 'Real-world' compensatory behaviour with low nicotine concentration e-liquid: subjective effects and nicotine, acrolein and formaldehyde exposure. Addiction 2018;113(10):1874-1882 [FREE Full text] [doi: 10.1111/add.14271] [Medline: 29882257]
26. Rose JE. Nicotine and nonnicotine factors in cigarette addiction. Psychopharmacology (Berl) 2006;184(3-4):274-285. [doi: 10.1007/s00213-005-0250-x] [Medline: 16362402]
27. Naqvi NH, Bechara A. The airway sensory impact of nicotine contributes to the conditioned reinforcing effects of individual puffs from cigarettes. Pharmacol Biochem Behav 2005;81(4):821-829 [FREE Full text] [doi: 10.1016/j.pbb.2005.06.005] [Medline: 15996724]
28. Dawkins LE, Kimber CF, Doig M, Feyerabend C, Corcoran O. Self-titration by experienced e-cigarette users: blood nicotine delivery and subjective effects. Psychopharmacology (Berl) 2016;233(15-16):2933-2941. [doi: 10.1007/s00213-016-4338-2] [Medline: 27235016]

XSL•FO
RenderX

29. Browne M, Todd DG. Then and now: consumption and dependence in e-cigarette users who formerly smoked cigarettes. Addict Behav 2018;76:113-121. [doi: 10.1016/j.addbeh.2017.07.034] [Medline: 28780356]
30. Perry RN, Girgulis JP, Harrison SL, Barrett SP. Dual users' perceptions of the addictive properties of cigarettes versus E-cigarettes. Addict Behav 2023;136:107483 [FREE Full text] [doi: 10.1016/j.addbeh.2022.107483] [Medline: 36084416]
31. Parrott AC. Does cigarette smoking cause stress? Am Psychol 1999;54(10):817-820. [doi: 10.1037//0003-066x.54.10.817] [Medline: 10540594]
32. Gades MS, Alcheva A, Riegelman AL, Hatsukami DK. The role of nicotine and flavor in the abuse potential and appeal of electronic cigarettes for adult current and former cigarette and electronic cigarette users: a systematic review. Nicotine Tob Res 2022;24(9):1332-1343 [FREE Full text] [doi: 10.1093/ntr/ntac073] [Medline: 35305014]
33. Robinson RJ, Hensel EC, Morabito PN, Roundtree KA. Electronic cigarette topography in the natural environment. PLoS One 2015;10(6):e0129296 [FREE Full text] [doi: 10.1371/journal.pone.0129296] [Medline: 26053075]
34. Lee YO, Morgan-Lopez AA, Nonnemaker JM, Pepper JK, Hensel EC, Robinson RJ. Latent class analysis of e-cigarette use sessions in their natural environments. Nicotine Tob Res 2019;21(10):1408-1413. [doi: 10.1093/ntr/nty164] [Medline: 30107462]
35. Physicians for a Smoke-Free Canada, Restrictions on E-Cigarette Flavours. 2023. URL: https://smoke-free.ca/SUAP/2021/e-cigarette-flavour%20restrictions.pdf [accessed 2024-08-30]
36. Cadham CJ, Liber AC, Sánchez-Romero LM, Issabakhsh M, Warner KE, Meza R, et al. The actual and anticipated effects of restrictions on flavoured electronic nicotine delivery systems: a scoping review. BMC Public Health 2022;22(1):2128 [FREE Full text] [doi: 10.1186/s12889-022-14440-x] [Medline: 36402989]
37. Hammond D, Reid JL, Burkhalter R, Bansal Travers M, Gravely S, Hyland A, et al. e-Cigarette flavors, devices, and brands used by youths before and after partial flavor restrictions in the United States: Canada, England, and the United States, 2017 2020. Am J Public Health 2022;112(7):1014-1024 [FREE Full text] [doi: 10.2105/AJPH.2022.306780] [Medline: 35622007]
38. Yang Y, Lindblom EN, Ward KD, Salloum RG. The impact of flavored e-cigarette bans on e-cigarette use in three US states. medRxiv 2023 [FREE Full text] [doi: 10.1101/2023.05.19.23290249] [Medline: 37292701]
39. Joossens L, Feliu A, Fernandez E. The Tobacco Control Scale 2019 in Europe. Association of European Cancer Leagues Brussels. 2019. URL: https://www.tobaccocontrolscale.org/TCS2019.pdf [accessed 2024-09-25]
40. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull 2007;133(5):859-883. [doi: 10.1037/0033-2909.133.5.859] [Medline: 17723033]
41. Schlosser AE. Self-disclosure versus self-presentation on social media. Curr Opin Psychol 2020;31:1-6. [doi: 10.1016/j.copsyc.2019.06.025] [Medline: 31357096]
42. Rüther T, Wissen F, Linhardt A, Aichert DS, Pogarell O, de Vries H. Electronic cigarettes-attitudes and use in Germany. Nicotine Tob Res 2016;18(5):660-669. [doi: 10.1093/ntr/ntv188] [Medline: 26385930]

Abbreviations

DSM-5*: *Diagnostic and Statistical Manual of Mental Disorders Fifth Edition
EVAPE: Evaluation of the Addictive Potential of E-cigarettes
MAPE: Model of the Addictive Potential of E-cigarettes

Edited by A Mavragani; submitted 05.03.24; peer-reviewed by T Rüther, D Hejlova; comments to author 03.05.24; revised version received 17.05.24; accepted 01.07.24; published 01.10.24.

Please cite as:
Andreas M, Grundinger N, Wolber N, Szafran D, Görig T, Mons U, Lohner V, Vollstädt-Klein S, Schneider S
Understanding e-Cigarette Addictiveness: Triangulation of Focus Group and Netnographic Data
J Med Internet Res 2024;26:e57970
URL: https://www.jmir.org/2024/1/e57970
doi:10.2196/57970
PMID:

©Marike Andreas, Nadja Grundinger, Nadine Wolber, Daria Szafran, Tatiana Görig, Ute Mons, Valerie Lohner, Sabine Vollstädt-Klein, Sven Schneider. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Depression Communities as a Complementary Approach to Improving the Attitudes of Patients With Depression Toward Medication Adherence: Cross-Sectional Survey Study

Runnan Chen[1,2], PhD; Xiaorong Fu[1], PhD; Mochi Liu[3], MIB; Ke Liao[4], BBA; Lifei Bai[1], PhD

[1]Department of Marketing, School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China

[2]Department of Marketing, Business School, University of Edinburgh, Edinburgh, United Kingdom

[3]Education Center for the Master of International Business, School of International Business, Southwestern University of Finance and Economics, Chengdu, China

[4]Department of Tourism Management, School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China

Corresponding Author:
Xiaorong Fu, PhD
Department of Marketing
School of Business Administration
Southwestern University of Finance and Economics
555 Liutai Road
Chengdu, 611130
China
Phone: 86 13981916682
Email: fuxr@swufe.edu.cn

Abstract

Background: Lack of adherence to prescribed medication is common among patients with depression in China, posing serious challenges to the health care system. Online health communities have been found to be effective in enhancing patient compliance. However, empirical evidence supporting this effect in the context of depression treatment is absent, and the influence of online health community content on patients' attitudes toward medication adherence is also underexplored.

Objective: This study aims to explore whether online depression communities (ODCs) can help ameliorate the problem of poor medication taking among patients with depression. Drawing on the stimulus-organism-response and feelings-as-information theories, we established a research model to examine the influence of useful institution-generated content (IGC) and positive user-generated content (UGC) on attitudes toward medication adherence when combined with the mediating role of perceived social support, perceived value of antidepressants, and the moderating role of hopelessness.

Methods: A cross-sectional questionnaire survey method was used in this research. Participants were recruited from various Chinese ODCs, generating data for a main study and 2 robustness checks. Hierarchical multiple regression analyses and bootstrapping analyses were adopted as the primary methods to test the hypotheses.

Results: We received 1515 valid responses in total, contributing to 5 different datasets: model IGC (n=353, 23.3%), model UGC (n=358, 23.63%), model IGC+UGC (n=270, 17.82%), model IGC-B (n=266, 17.56%), and model UGC-B (n=268, 17.69%). Models IGC and UGC were used for the main study. Model IGC+UGC was used for robustness check A. Models IGC-B and UGC-B were used for robustness check B. Useful IGC and positive UGC were proven to have positive impact on the attitudes of patients with depression toward medication adherence through the mediations of perceived social support and perceived value of antidepressants. The findings corroborated the role of hopelessness in weakening or even negating the positive effects of ODC content on the attitudes of patients with depression toward medication adherence.

Conclusions: This study provides the first empirical evidence demonstrating the relationship between ODC content and attitudes toward medication adherence, through which we offer a novel solution to the problem of poor medication adherence among patients with depression in China. Our findings also provide suggestions about how to optimize this new approach—health care practitioners should generate online content that precisely matches the informational needs of patients with depression, and ODC service providers should endeavor to regulate the community atmosphere. Nonetheless, we warn that ODC interventions cannot be used as the only approach to addressing the problem of poor medication taking among patients with severe depressive symptoms.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56166) doi:10.2196/56166

KEYWORDS

online depression communities; attitudes; institution-generated content; user-generated content; perceived social support; antidepressants; hopelessness; cross-sectional study; China; health care system; online health community; depression; medication adherence; social support; health care practitioner; peer support

Introduction

Background

It is estimated that there are currently >95 million people with depression in China [1]. Among this population, one-third have never sought professional help, and only 8.99% have taken medicinal treatments [2]. Moreover, among those taking antidepressants, >70% exhibit poor adherence behaviors, meaning that they either arbitrarily alter prescribed dosages or stop taking prescribed medications altogether [3], which will negatively impact the therapeutic outcomes [4]. Such widespread nonadherence poses a serious challenge to the health care system [5]. However, due to the huge population, uneven distribution, and insufficient medical resources in China [6], the amelioration of this problem is extremely tough and tardy.

Fortunately, recent research on online health communities (OHCs) sheds light on a new approach to confronting such dilemmas. OHCs are online platforms that provide patients and health care seekers with access to health care knowledge, experiential information, and social support [7,8]. They can be found in a variety of virtual spaces, such as chat rooms, forums, and bulletin boards on social media or websites [8,9]. Taking advantage of the internet, OHCs can reach a large population of patients at a much lower cost than offline channels [10]. Grounded on this invaluable strength, a great deal of research interest has been paid to investigate OHCs' potential to address various health care problems. The value of OHCs in enhancing patients' motivation to adhere to prescribed medications has also been corroborated. Yin et al [11] determined that OHC member reciprocity positively affected patients' completion of a 5-year regimen. Wu et al [12] found that informational support from OHCs strengthens patients' trust in health care services and boosts their adherence intentionality. Lu and Zhang [13] and Lu [6] revealed that the physician-patient relationship, along with patients' health information–seeking behaviors, becomes better in OHCs, resulting in patients' increasing willingness to comply with physicians' instructions. These findings ascertain that OHCs can be an effective tool to remediate patients' attitudes toward medication adherence, thereby refining the nonadherence problem.

However, depression has some distinctions compared to general diseases. Patients with depression experience extensive impairments in cognition, well-being, and social life, putting them at high risk of self-harming and even suicide owing to overwhelming agony and desperation [14]. Statistical results have also revealed that a poor medication adherence problem is much more prominent and prevalent among patients with depression [2,3,5]. In line with such uniqueness, online depression communities (ODCs), which are communities exclusive for stakeholders in depression treatment (eg, patients, physicians, and psychological counselors) to discuss depression-related topics, have prosperously developed in cyberspaces and are usually separated from OHCs in academia [15-18]. Therefore, we are concerned that the conclusions of previous OHC research might not necessarily hold true in the depression treatment and ODC context. Additional empirical studies to verify ODCs' influence on the attitudes of patients with depression toward medication adherence are warranted.

Furthermore, previous research on the relationship between OHCs and medication adherence has primarily treated users' participation behaviors or perceptions of the overall OHC environment as determinants of their received benefits, with little emphasis placed on the role of community content. Serving as the touch point between an online community and its members, community content is closely related to how users will be influenced by the community [19]. In particular, Nimrod [20] verified that the influence of ODC content is not homogeneous—content with different features will affect recipients in different ways. This finding suggests the importance of taking the characteristics of the content into consideration when examining ODCs' impact. Therefore, this study intended to explore the value of ODCs in improving the medication adherence attitudes of patients with depression from a community content perspective.

Theoretical Foundation

Stimulus-Organism-Response Theory

ODC content can be viewed as a series of external stimuli to community users, and variation in attitudes toward medication adherence is a potential user response. Therefore, we introduce the stimulus-organism-response (SOR) theory [21] to ascertain the relationship between ODC content and attitude toward medication adherence. Following the environmental psychology perspective, SOR theory states that external stimuli affect human behavior by influencing an individual's internal experience (organism), thereby eliciting a response [21]. Its validity in theorizing the influence of online communications has been corroborated by previous literature. In particular, several studies have used this theory to evaluate the relationship between OHC characteristics and the responses of community members. Their results indicate that users' positive responses are attributable to positive information dissemination within the community, such as emotional support, high-quality content, and an enjoyable atmosphere [7,22,23]. In contrast, negative informational factors, such as information overload and system functionality overload, lead to resistant behavior [24]. On the basis of these findings, we postulate that favorable stimuli evoke affirmative responses, whereas unfavorable stimuli prompt adverse responses. The improvement in medication adherence attitudes is conceivably the result of stimulation of ODC content carrying positive features.

XSL•FO
RenderX

Moreover, SOR theory offers a well-established paradigm to explore the inner mechanism through which patients with depression are influenced by ODC content using a series of psychological activities [25]. On the basis of this theoretical framework, we assume that patients' attitudes toward medication adherence might be improved concurrently with alterations in some subjective factors that closely relate to the adherence to antidepressants, initially through the stimulation of specific ODC content. Previous studies have identified that psychological activities that directly affect patients' willingness to adhere to antidepressant treatment primarily include their perceived support from their social network and their subjective evaluations of the pros and cons of taking antidepressants [26-29]. These mental processes can be boiled down to the concept of perceived social support and perceived value of antidepressants. Both are perception-level factors that can be influenced by information disseminated through online communities [6,30]. Therefore, we predict that they may serve as mediators between the extrinsic stimuli that patients receive from ODCs and their responses.

Feelings-as-Information Theory

Nevertheless, as this study strives to understand how patients with depression, which is a type of mood disorder [14], perceive and respond to external stimuli, it is necessary to take their emotional states into consideration. The SOR theory is valuable for postulating connections between ODC content stimuli and responses in patients with depression. However, it does not explain how mental state as a kind of internal information filter influences individuals' decision patterns. Hence, we incorporate the feelings-as-information theory [31] to cover this deficiency and aid in enhancing the overall integrity of our theoretical framework. The feelings-as-information theory is a critical theory in emotion research. It delves into how subjective feelings, such as emotions, moods, and metacognitive experiences, impact human judgments [31-33], offering a comprehensive model to describe the interplay between emotion and thought. According to this theory, an individual's state of mind affects their processing style when dealing with external information and, thus, impacts their judgments [31,32]. Its effectiveness in interpreting how feelings moderate people's responses to information on social media has also been confirmed by previous literature [34-36]. As the motivation to adhere to antidepressant treatment is a symbol of trying to save oneself from depression, we surmise that hopelessness—a state of mind in which people lose heart and cease to care about life or the future [14]—will play a key role in moderating the positive effect of ODC content on patients' attitudes toward medication adherence.

Research Hypotheses

The Content in ODCs

In essence, all informational materials within an online community are generated by its participants, whereas the information from different content providers may be heterogeneous [18,19]. Typically, the discrepancy in specialized knowledge and expertise can differentiate participants' roles in an online community [19]. Health care providers in particular can educate other users (eg, patients) by sharing professional health information [37,38]. ODCs are technically communities that address health-related issues. Thus, when studying the influence of ODC content, distinguishing between institution-generated content (IGC) and user-generated content (UGC) is essential.

IGC refers to content posted on online communities by specialized agencies, such as governmental agencies, hospitals, or psychology institutes, and personnel associated with those agencies [39]. Researchers have learned that it is common for OHC users to create or disseminate inaccurate or even false information in the communities, which may mislead other platform users [39,40]. IGC is generally considered more credible, authoritative, and trustworthy than content created by lay users [41,42]. When seeking specific information about a condition or disease, people are inclined to go to a reputable source [39]. Nevertheless, studies have shown that, despite the voluminous content in ODCs, individuals with depression still have questions and unmet informational needs [15]. Therefore, we hypothesize that the perceived usefulness of IGC may be a significant factor affecting IGC's influence on content recipients.

UGC is content created by users in the community [18]. Due to the anonymity within ODCs and the fact that all participants experience depression, community members come together in a safe space to communicate their experiences of living with depression and exchange sympathy or encouragement with others [8,40]. While this content is generally not as accurate and reliable as IGC [39,41,42], it provides ODC users with experiential knowledge about depression treatment [16] and emotional support [18], which cannot be accessed through IGC. However, the tone of UGC in ODCs is usually mixed because users come to the sites in a depressed state [16]. In addition to content that offers positive support, ODCs are rife with negative UGC that conveys anguish and despair [16]. Research has shown that individuals who experience depression may be less likely to seek assistance if they perceive the received messages related to depression as negative [17,43]. Therefore, we firmly believe that perceived positivity is a decisive factor in influencing UGC's ability to help patients with depression improve their attitudes toward taking medications.

Overall, we extracted 2 types of content carrying positive features in the ODC setting: useful IGC and positive UGC. In our notion, they can serve as positive extrinsic stimuli for patients with depression, eliciting improved attitudes toward taking medications. Useful IGC provides needed and reliable health information to patients [37,38]. This can enrich their knowledge about depression and antidepressants, through which irrational concerns about the side effects of medications can be alleviated and facts about the purpose and value of medications can be clarified, making patients more inclined to adhere to prescribed medicinal protocols [28,29]. Positive UGC, on the other hand, provides opportunities for community members to communicate with other patients with depression about their experiences of coping with depression [16,40]. Such positive content can alleviate their loneliness during treatment and empowers them to fight depression more actively [9]. Moreover, it helps patients understand their disease and evaluate medical treatments based on the experience of others with the same condition [8,44]. Gradually, as they develop a more

XSL•FO
RenderX

comprehensive view of antidepressant treatment, they may have a more positive attitude toward medicinal treatment and become more willing to take antidepressants [27,45,46]. Consequently, we propose the following:

- The usefulness of IGC has a positive effect on the attitudes of patients with depression toward medication adherence (hypothesis 1a).
- The positivity of UGC has a positive effect on the attitudes of patients with depression toward medication adherence (hypothesis 1b).

The Mediating Role of Perceived Social Support and Perceived Value of Antidepressants

Perceived Social Support

Social support is defined as the exchange of resources between individuals who play the roles of support provider and support recipient with the function of improving the recipient's well-being [47]. Feldhege et al [18] determined that the communications within ODCs can primarily provide 2 types of support to users: informational support and emotional support. In line with this finding, both IGC and UGC can be potential social support resources for ODC members. On the one hand, IGC is a source of authoritative health care information [39], consisting of reliable messages that cover topics such as the disease, treatments, health care services, and coping strategies for depression. They can offer informational support to individuals with depression, helping them gain a better understanding of depression and how to deal with it [15,18]. In contrast, UGC refers to the communications among community users. Social discrimination and lack of public understanding of depression are significant issues in China, causing individuals with depression to be marginalized and ostracized [17]. Through ODCs, patients with depression connect and relate to others in similar situations [40], providing mutual help and mitigating individual loneliness [9]. In this way, UGC generates potential emotional support for ODC participants [40].

Nevertheless, perceived social support is to some extent different from received social support [48]. In addition to availability, perceived social support is also significantly influenced by recipients' satisfaction with the support provided [49]. To positively influence perceived social support, the support provided needs to match people's needs as it is only when individuals feel that their issues are successfully addressed or mitigated that they perceive the benefit of social support [50,51]. Grounded on this rationale, not all ODC content can uplift users' perceived social support. For instance, IGC that is unhelpful to users and UGC that discusses negative topics (eg, self-harm, suicide, or despair) cannot make the recipients feel well supported. Therefore, we argue that the usefulness of IGC and positivity of UGC are important driving factors for perceived social support. Useful IGC fulfills users' specific demands for knowledge about depression and coping strategies, whereas positive UGC soothes users' loneliness and agony, thereby enhancing their perceived social support. Accordingly, we propose the following:

- The usefulness of IGC has a positive effect on the perceived social support of patients with depression (hypothesis 2a).
- The positivity of UGC has a positive effect on the perceived social support of patients with depression (hypothesis 2b).

Patients with depression are frequently pessimistic and unmotivated [14], making them likely to avoid challenges and stop taking prescribed medications [5]. In this case, they might need some additional support to encourage them not to lose hope of recovering. If they perceive strong social support when they are tempted to stop taking their medications, they are reminded that they are not alone and there are others who understand their misery. Thereby, they can be empowered to keep fighting depression [9]. Consistent with this rationale, clinical studies have concluded that patients with depression with strong social support systems generally have better attitudes toward medicinal interventions and are less likely to abandon treatment ahead of schedule [4,26]. Therefore, we propose the following:

- Perceived social support has a positive effect on the attitudes of patients with depression toward medication adherence (hypothesis 3).

Perceived Value of Antidepressants

Perceived value is generally determined by consumers' subjective evaluations of a product's potential benefits and costs [52,53]. As the evaluation process is not objective, perceived value may significantly differ from a product's actual value due to consumers' lack of understanding of the product or excessive concerns about its potential risks [52,54]. Such a phenomenon is typically salient in the setting of antidepressant treatment. Although clinical research has verified that all first-line antidepressants have substantial efficacy in improving depressive symptoms [55] and most side effects recede spontaneously after an initial phase of use [56], a misconception that taking antidepressants is painful and ineffective still widely exists, negatively impacting patients' perceived efficacy and willingness to adhere to antidepressant treatment [27-29,57].

Fortunately, perceived value is a dynamic trait [58]. As consumers gain more exposure to a product, their evaluation of its value tends to become more comprehensive [59]. Irrational concerns about a product's potential risks can also lessen as consumers learn more about it [60]. In line with these findings, clinical research has confirmed that objective knowledge about antidepressants is positively associated with patients' beliefs about their value and willingness to use them [45,46]. Through the IGC channel, patients with depression can conveniently access information about antidepressants and obtain specific details about their features and functions curated by health experts [37,38]. When users perceive that the information presented is believable and useful, the information is stored in their memory and serves to enhance their objective knowledge of the medicines [61]. With more objective knowledge, they better understand the benefits, and their concerns about potential adverse reactions to medications are allayed, both significant factors influencing perceived value. Therefore, we hypothesize the following:

- The usefulness of IGC has a positive effect on patients with depression's perceived value of antidepressants (hypothesis 4a).

XSL•FO
RenderX

UGC has also been found to affect consumer attitudes and value perceptions about products discussed in online communities through peer influence [62]. Peer influence occurs among a group of people who share common characteristics or interests and is actually a mechanism of learning by imitation [63]. Individuals may absorb and internalize information generated by their peers in a conscious or unconscious way, leading to a convergence of group members' opinions and beliefs [63,64]. Relevant studies have also revealed that peers' experiences of a medical treatment will influence patients' judgments on its benefits [44,46,64]. Therefore, we postulate that, when patients are exposed to UGC that relates positive personal experiences of antidepressant treatment, they may internalize the information and develop a more positive perception of antidepressants. Accordingly, we propose the following:

- The positivity of UGC has a positive effect on patients with depression's perceived value of antidepressants (hypothesis 4b).

The relationship between the perceived value of antidepressants and attitudes toward medication adherence is fairly obvious. A high perceived value indicates that consumers believe that using antidepressants can bring more benefits than costs [52], and therefore, they are willing to purchase or use these medications [53]. These patients understand that taking antidepressants can help ease or even alleviate their depression, and this understanding is significantly related to their attitudes toward medication adherence [65]. Hamrin et al [5] also noted that, when patients with depression believe that antidepressants are helpful to them, their intrinsic motivation to comply with treatment is substantially enhanced. Therefore, we suggest the following:

- Perceived value of antidepressants has a positive effect on the attitudes of patients with depression toward medication adherence (hypothesis 5).

Mediation Effects

In line with the SOR theory, an individual's organism, which may be any internal assessment correlated with or aroused by a stimulus, is a key mediator between the external stimuli and the response [25]. The feasibility of applying this theory in the depression treatment context is reinforced by clinical research outcomes. As a pharmacological treatment for a mental disease, the effectiveness of antidepressants is revealed more by patients' self-reporting than by physiological indicators [57]. Whether medical treatments are effective and whether they merit continuation depends largely on patients' subjective evaluations. Consequently, the willingness to accept medicinal interventions has been confirmed to be closely associated with a variety of cognitive factors [3,27-29,65]. Therefore, it is reasonable to theorize that patients' internal experiences play a significant mediation role in our research setting.

Perceived social support is an individual's feeling about whether there is someone able and ready to provide help when they are in need [48]. Perceived value is a subjective evaluation of a product's benefits and costs [52]. They are both subjective perceptions and, thus, meet the criteria to be regarded as organisms [25]. Therefore, we propose that perceived social support and perceived value of antidepressants play mediating roles between ODC content stimuli and the attitudes of patients with depression toward medication adherence. As patients are exposed to useful IGC that enables them to better cope with the challenges of depression and the medication process or positive UGC about others' experiences of fighting depression, perceived social support and perceived value of antidepressants increases, and this in turn boosts their attitudes toward adherence to medicinal treatment. Accordingly, we propose the following hypotheses:

- Perceived social support mediates the effect of the usefulness of IGC on the attitudes of patients with depression toward medication adherence (hypothesis 6a).
- Perceived social support mediates the effect of the positivity of UGC on the attitudes of patients with depression toward medication adherence (hypothesis 6b).
- Perceived value of antidepressants mediates the effect of the usefulness of IGC on the attitudes of patients with depression toward medication adherence (hypothesis 7a).
- Perceived value of antidepressants mediates the effect of the positivity of UGC on the attitudes of patients with depression toward medication adherence (hypothesis 7b).

The Moderating Role of Hopelessness

In line with the feelings-as-information theory, individuals' moods have substantial influence on their processing styles when dealing with external information [31]. In particular, Schwarz et al [66] ascertained that a negative affect can act as an alarm bell, alerting people that they are in a problematic situation. Such information provides cognitive tuning to adapt a person's processing strategies to situational requirements [66]. As having bad moods is an obvious symptom of depression [14], practically all patients with depression satisfy the condition for this effect. However, the negative emotions that patients with depression bear may continue to escalate and become extremely serious. Marchetti [67] noted that people trapped in depression and misery for a long time will develop a sense of hopelessness. With such an emotional state, even when offered the possibility of escape, they choose to remain in the status quo rather than try to alleviate their symptoms [68]. Consistent with this rationale, clinical studies show that poor medication adherence is frequently exhibited by patients with intense negative moods in the treatment of chronic diseases such as HIV or diabetes [69-71]. Hosek et al [71] and DiMatteo et al [72] have suggested that the underlying cause is a feeling of hopelessness that often accompanies severe depressive symptoms, which leads patients to believe that their actions will not bring desired outcomes [67] and extinguishes their motivation to adhere to treatments [72].

Collectively, these findings suggest that a negative emotional state, which can range from mild sadness to intense hopelessness, is closely related to patients' decision patterns in the medication treatment process. Nevertheless, it should be noted that, the current rate of uptake and general adherence to antidepressants in extremely low around the globe [2-5,29]. Skepticism about the medication's purpose can arise due to lack of understanding [27,29]. It is common for those with depression to struggle with sticking to their medication routine, often due

XSL•FO
RenderX

to lack of motivation and positive reinforcement [5]. Even emotional state notwithstanding, they may still have trouble trusting that the medication will work for them. Therefore, we propose that hopelessness, rather than directly determining motivation to adhere to antidepressant treatment, may serve as a significant factor regulating the influence of positive intervention on the medication adherence attitudes of patients with depression. Our hypothesis suggests that the use of positive ODC stimuli can play a crucial role in improving the willingness of patients with depression to adhere to their medication routine. We theorize that hopelessness serves as a key moderator of the relationship between a patient's attitude toward medication adherence and the positive stimuli from ODCs.

More specifically, patients who are hopeful about their lives and futures will be driven to deal with their depression by actively exploring coping strategies [66]. With useful IGC that offers authoritative knowledge about antidepressants and positive UGC that depicts successful recovery experiences, patients with depression may come to understand that antidepressants can help them and be inspired to adhere to treatment. However, when patients reach a certain level of hopelessness, they become overwhelmed by the negative belief that any recovery attempts are useless [67]. With such biased cognition, they remain mired in the status quo and cannot be persuaded to comply with the medicinal treatments even when provided with useful IGC and positive UGC as supports [68,72]. Consequently, we propose the following hypothesis:

- Hopelessness negatively moderates the effects of the usefulness of IGC and the positivity of UGC on the attitudes of patients with depression toward medication adherence such that the effects are stronger when a patient has a low level of hopelessness but weaker or even negated when a patient has a high level of hopelessness (hypothesis 8).

Research Model

Overall, under the guidance of the SOR and feelings-as-information theories, this study established a research model (Figure 1) theorizing the influence of useful IGC and positive UGC on the attitudes of patients with depression toward medication adherence considering the mediating effects of perceived social support and perceived value of antidepressants and the moderating role of hopelessness. This study will provide the first empirical evidence demonstrating whether ODCs can be used as an effective complementary approach to addressing the nonadherence problems among patients with depression, and it will contribute to the preliminary knowledge on its inner mechanism and boundary conditions.

Figure 1. Research model. IGC: institution-generated content; UGC: user-generated content.

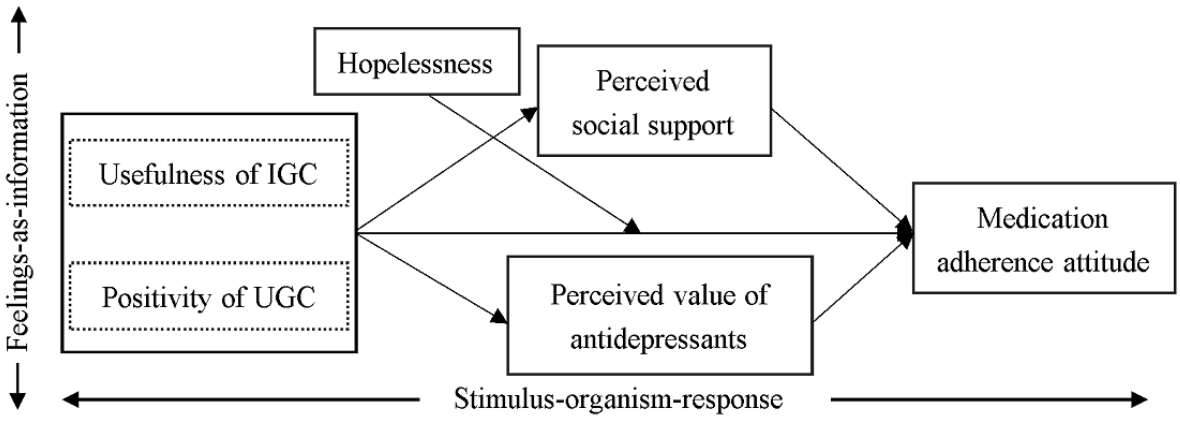

Methods

Study Design

This research aimed to explore the influence of ODC content on the willingness of people with depression to adhere to medicinal treatments. A main study and 2 robustness checks, all of which used a cross-sectional survey method, were designed to test our hypothetical model.

The main study was conducted to examine our research model using data collected from patients with depression who had only read one type of ODC content (ie, IGC [content provided by professional institutions] or UGC [content provided by other patients with depression]). The previous work by Lu [6] revealed that, as the specialization of OHCs develops, some OHCs exclusive for patient-to-physician communications or patient-to-patient communications are being established in China. Under such circumstances, some ODC users have only been exposed to either IGC or UGC. Volunteers who had only read one type of ODC content were less likely to experience the distraction of recalling another type of content, generating more plausible data. Thus, we used the data collected from them (named model IGC and model UGC) for the main study to verify our hypotheses.

Robustness check A, whose data (named model IGC+UGC) were collected from volunteers who had been exposed to both types of content, was implemented to optimize the generalizability of our research findings. This substudy validated whether the effects of useful IGC and positive UGC could function simultaneously and examined whether they would impact patients with depression differently.

XSL•FO
RenderX

Robustness check B, whose data (named model IGC-B and model UGC-B) were also collected from volunteers who had only read one type of ODC content, was conducted to further corroborate our core hypotheses—the effects of ODC content on the attitudes of patients with depression toward medication adherence. In this substudy, participants' ODC use duration and frequency were additionally measured to evaluate how the influence of IGC usefulness and UGC positivity on participants' medication adherence attitudes varied along with their ODC use situations. This study was designed to reduce the likelihood that the identified effects of ODC content on medication adherence attitudes were false-positive results.

Data Collection and Respondent Profile

We used convenience sampling and snowball sampling in our data collection process. Given that users from the same ODC likely have similar perceptions about their community as they are exposed to the same information, we did not rely on a single ODC to obtain data for our research. Instead, we posted research advertisements on several health care websites (eg, "SunForum"), online forums (eg, "Baidutieba"), online depression group chats (eg, "Depression Tree Hole"), and social media platforms (eg, "Xiaohongshu") to recruit participants from various ODCs. Meanwhile, we encouraged participants to share our questionnaire with their peers with depression who were also using ODCs. The sample recruitment advertisement is provided in Multimedia Appendix 1.

Because previous research has found that approximately one-third of individuals with depression in China have never sought professional help and, therefore, have not been officially diagnosed with depression [2], we sought to include this significant cohort in our study population [9]. Both patients with diagnosed depression and self-reported depression were targeted so that our sample reflected the population of individuals with depression in China as a whole, ensuring that our data were not biased and the validity of the study was not impugned.

The general inclusion criteria for this study were that the participants must currently have depression and be users of ODCs. Participants were regarded as experiencing depression if they met one of the following conditions: (1) they had been officially diagnosed with depression by a physician and had not yet recovered (diagnosed depression), (2) they had been suggested to have a high likelihood of being depressed by a self-rating depression scale and believed that they were still in a depressive status (self-reported depression), or (3) they had experienced depression-related symptoms for a rather long time and strongly believed that they might have depression (self-reported depression). The scope of ODCs was defined as "digital communities exclusive for stakeholders in depression treatment (eg, patients, physicians, and psychological counselors) to discuss depression-related topics." The additional criterion for each substudy, which related to the type of ODC content that the participants had read, is provided in the Study Design subsection. Participants were excluded if they reported no depression experience, had already recovered from depression, or were not using ODCs.

To mitigate the limitations of self-reported data, the following steps were taken. First, the survey was anonymous, and we emphasized to the participants that all information they provided to us would be kept strictly confidential and only used for research purposes. This helped reduce the risk of social desirability bias by mitigating participants' privacy concerns and fear of judgment [73]. Second, we repeatedly reminded the participants throughout the questionnaire that the accuracy of their responses was vital to the success of this study and the common well-being of the population of patients with depression. In this way, we sought to encourage participants to allocate more cognitive resources to our study and answer the questions more carefully, honestly, and precisely [74]. Third, we manually scrutinized every response received. Participants whose responses displayed obvious signals of carelessness (eg, answering all questions with the same choice, random answers, and very short response time) were excluded from our samples [75].

The whole data collection fieldwork was conducted in 3 periods. The first period was from June 2022 to September 2022, which collected 469 effective responses for our main study. The second period was from July 2023 to August 2023, generating 173 effective responses for robustness check A. The third period was from April 2024 to June 2024, providing 873 effective responses for our main study, robustness check A, and robustness check B. Overall, 711 effective responses for the main study (n=353, 49.6% for model IGC and n=358, 50.4% for model UGC), 270 effective responses for robustness check A (all for model IGC+UGC), and 534 effective responses for robustness check B (n=266, 49.8% for model IGC-B and n=268, 50.2% for model UGC-B) were collected. Of the overall respondents, 67.1% (1017/1515) were patients with diagnosed depression, and 32.9% (498/1515) were patients with self-reported depression, closely mirroring the statistics provided by Yu et al [2]. Multimedia Appendix 2 provides more detailed sample characteristics.

Ethical Considerations

This study received ethics approval from the School of Business Administration of Southwestern University of Finance and Economics. Informed consent was obtained from all participants and was necessary for submitting the digital questionnaire. The questionnaire was anonymous and did not collect any identifying information, such as name or address. Eligible participants were given CN ¥10 (US $1.40) in cash as compensation.

Measurement Instruments

Overview

The measurements for each construct were adapted from existing research and modified based on the context and the target audience. Specifically, in the preliminary stage, we reviewed related literature to obtain the seminal scales of reference variables. All items that were originally in English were translated into Chinese and cross-culturally adapted. In addition, we interviewed 5 patients with depression to contextualize the items. Next, 2 psychology experts, 2 marketing experts, and 2 medical experts conducted reviews to refine the wording of the instrument items. Their feedback provided the basis for revising

XSL•FO
RenderX

the construct measures and modifying wording and item sequence. Finally, we back translated the Chinese questionnaire into English to ascertain its accuracy, completeness, and consistency compared to the original text. The full measure items are listed in Multimedia Appendix 3. All theoretical variables were measured on 7-point Likert-type scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The descriptive statistics and reliability of each variable are shown in Multimedia Appendix 4.

Usefulness

The perception of IGC usefulness was measured using a 3-item scale developed by Lee and Hong [76].

Positivity

We used a 4-item scale to measure the perception of UGC positivity. The scale was adapted from the friendliness scale by Price and Arnould [77] in accordance with the definition of positive feedback by Fong et al [78].

Perceived Social Support

We measured perceived social support using an 11-item scale adapted from the Multidimensional Scale of Perceived Social Support by Zimet et al [79]. The scale by Zimet et al [79] originally measured 3 dimensions: perceived support from significant other, perceived support from family, and perceived support from friends. This multidimensional measurement style is inapplicable in the ODC context. Thus, we adapted this scale, making it a unidimensional scale measuring the comprehensive perception of support from the overall society.

Perceived Value of Antidepressants

The perceived value of antidepressants was measured using a 5-item scale developed by Sweeney and Soutar [80].

Medication Adherence Attitude

Medication adherence attitude was measured using a 3-item scale adapted from the Adherence Attitude Inventory developed by Lewis and Abell [81].

Hopelessness

We used a 17-item scale adapted from the Hopelessness Scale by Beck et al [82] to measure the current hopelessness intensity of patients with depression.

ODC Use Duration and Frequency

In robustness check B, participants' ODC use duration and frequency were measured. Participants' ODC use duration was measured by calculating the interval between the reported date of their first visit to the ODC and the date when they responded to our survey. Participants' average ODC use frequency was measured using 10 options ranging from "once a month" to "at least once a day." Each option was converted into a visit per day form for statistical analyses; for example, "once a month" was counted as 0.0333 visits per day (1 visit per 30 days), "once a week" was counted as 0.1428 visits per day (1 visit per 7 days), and "at least once a day" was counted as 1 visit per day.

Control Variables

Following previous research [9], we chose participants' mental health condition, gender, age, educational level, and marital status as the basic control variables in our study. Given that collectivism is strongly emphasized in China, especially within the family of origin [17], and that the medication adherence situations of patients with depression are closely associated with their families of origin [3], we decided to include family of origin as a control variable. Furthermore, relevant research suggests that recipients' perception of the credibility of online health information influences its communication effect [39]. To eliminate the interference of this factor, we additionally included the perceived credibility of IGC and UGC as a control variable using a 4-item scale developed by Rimmer and Weaver [83] to measure it.

Assessment of Measurement Validity

Common Method Bias Test

Given that all variables in our study were self-reported by participants, our research might be affected by common method bias [84]. We conducted the single-factor test by Harman [85] to examine the underlying seriousness of common method bias. In the main study, for items in model IGC, 7 factors with an eigenvalue of >1 were extracted, with the first factor accounting for 22.67% of the variance; for the items in model UGC, 6 factors with an eigenvalue of >1 were extracted, with the first factor accounting for 23.56% of the variance. In robustness check A, 9 factors with eigenvalues of >1 were identified, with the first factor accounting for 18.45% of the variance. In robustness check B, 3 factors with an eigenvalue of >1 emerged from model IGC-B, with the first factor explaining 33.01% of the variance; 4 factors with an eigenvalue of >1 emerged from model UGC-B, with the first factor explaining 24% of the variance. As the first factor in each model only accounted for a small proportion of the total variance and several factors emerged, common method bias did not appear to be a serious problem for our research [84].

Convergent Validity and Discriminant Validity

Convergent validity is considered acceptable if the value of the standardized factor loading of each item is >0.5 and the composite reliability (CR) and average variance extracted (AVE) of each variable are >0.7 and >0.5, respectively [86]. Confirmatory factor analysis was conducted, and the results showed that the standardized factor loadings of all variables were >0.5. Corresponding CRs and AVEs computed based on the standardized factor loadings are shown in Multimedia Appendix 5. All CR and AVE values satisfied the required standards, ensuring convergent validity. Discriminant validity differentiates a construct from other constructs in the same model. A classic method to assess discriminant validity is comparing the square root of the AVE and the correlation between 2 constructs [87]. As shown in Multimedia Appendix 5, the square root of the AVE for each variable in our models was greater than its correlations with other variables. Therefore, discriminant validity was confirmed.

XSL•FO
RenderX

Results

Results of the Main Study

Assessment of Model Fit

Before verifying the proposed hypotheses, we needed to confirm whether the hypothesized model fit the data well. While 2 of the latent variables in our models, perceived social support (11 items) and hopelessness (17 items), were measured using a great number of items, a large parameter estimation bias might be incurred if we directly used the original items to construct the models. To avoid this interference, previous research has suggested parceling the scales if they are unidimensional [88]. Both the scales of perceived social support and hopelessness used in this study fulfilled this requirement. Therefore, we parceled them into 3 items each following the factorial algorithm proposed by Rogers and Schmitt [89]. The modified models' fit analyses yielded the following results for model IGC—χ^2/df=2.0, P<.001, comparative fit index (CFI)=0.977, Tucker-Lewis index (TLI)=0.972, and root mean square error of approximation (RMSEA)=0.052—and model UGC—χ^2/df=2.5, P<.001, CFI=0.966, TLI=0.959, and RMSEA=0.065. All fit indexes met the acceptable standards [90], indicating that the hypothesized models fit the data well and it was allowable to conduct the subsequent path analyses.

Hypothesis Testing

Assessment of Direct Effects

We conducted hierarchical multiple regression analyses to examine the hypothesized direct effects in our models. The standardized path coefficient and corresponding significance of each hypothesis are shown in Table 1. As the results indicate, the usefulness of IGC had positive effects on medication adherence attitude (β=.258; P<.001), perceived social support (β=.296; P<.001), and perceived value of antidepressants (β=.316; P<.001); the positivity of UGC also had positive effects on medication adherence attitude (β=.226; P=.001), perceived social support (β=.404; P<.001), and perceived value of antidepressants (β=.251; P<.001). Thus, hypotheses 1, 2, and 4 were supported. Moreover, in both model IGC and model UGC, perceived social support and perceived value of antidepressants were verified to have positive effects on medication adherence attitude (β>0; P<.001). Therefore, hypotheses 3 and 5 were supported as well.

Table 1. Results of direct effect examination.

Hypotheses	β	*P* value
Model IGC[a] (n=353)		
Hypothesis 1a: usefulness→medication adherence attitude	.258	<.001
Hypothesis 2a: usefulness→perceived social support	.296	<.001
Hypothesis 4a: usefulness→perceived value of antidepressants	.316	<.001
Hypothesis 3: perceived social support→medication adherence attitude	.245	<.001
Hypothesis 5: perceived value of antidepressants→medication adherence attitude	.696	<.001
Model UGC[b] (n=358)		
Hypothesis 1b: positivity→medication adherence attitude	.226	.001
Hypothesis 2b: positivity→perceived social support	.404	<.001
Hypothesis 4b: positivity→perceived value of antidepressants	.251	<.001
Hypothesis 3: perceived social support→medication adherence attitude	.229	<.001
Hypothesis 5: perceived value of antidepressants→medication adherence attitude	.677	<.001

[a]IGC: institution-generated content.

[b]UGC: user-generated content.

Assessment of Mediating Effects

Through the above analysis of direct effects, we confirmed that the independent variables in this study had significant relationships with the hypothesized mediating variables and the dependent variable. Therefore, the preconditions for testing the mediating effects were satisfied [91]. In accordance with the suggestion by Zhao et al [92], we used bootstrapping analysis (supported by SPSS version 26.0, PROCESS model 4; IBM Corp) to examine the mediating effects of perceived social support and perceived value of antidepressants. As the results in Table 2 show, all the 95% CIs did not include 0, suggesting that the indirect effects of the usefulness of IGC and positivity of UGC on attitude toward medication adherence through perceived social support or perceived value of antidepressants were statistically significant. Therefore, hypotheses 6 and 7 were supported.

XSL•FO
RenderX

Table 2. Results of mediating effect examination.

Hypotheses	β (SE; 95% CI)
Model IGC[a] (n=353)	
Hypothesis 6a: usefulness→perceived social support→medication adherence attitude	.066 (.0271; .0175-.1243)
Hypothesis 7a: usefulness→perceived value of antidepressants→medication adherence attitude	.249 (.0564; .1403-.3638)
Model UGC[b] (n=358)	
Hypothesis 6b: positivity→perceived social support→medication adherence attitude	.092 (.0433; .0200-.1901)
Hypothesis 7b: positivity→perceived value of antidepressants→medication adherence attitude	.204 (.0808; .0494-.3713)

[a]IGC: institution-generated content.

[b]UGC: user-generated content.

In addition, Table 2 reveals that the mediating role of the perceived value of antidepressants on the relationship between IGC usefulness or UGC positivity and attitude toward medication adherence (β>.20, 95% CI not including 0) seemed to be stronger than that of perceived social support (β<.10, 95% CI not including 0). To assess the significance of this difference in a more modest way, we used the following *Z* statistic:

where β_x are coefficients of the indirect effects and s^2 are the squared SEs of the coefficients [93]. The results show that the indirect effect of IGC usefulness on attitude toward medication adherence via the mediation of perceived social support was significantly weaker than the indirect effect via the mediation of perceived value of antidepressants (*Z* statistic=–2.921∉[–1.96, 1.96]), whereas the difference was not salient for the indirect effects of UGC positivity (*Z* statistic=–1.227∈[–1.96, 1.96]).

Assessment of Moderating Effects

To exclude the possibility that hopelessness might be a direct influencing factor on the attitudes of patients with depression toward medication adherence, we conducted hierarchical multiple regression analyses to examine their relationship. Our conclusion was based on the findings of both model IGC (β=–.023; *P*=.68) and model UGC (β=–.023; *P*=.69). As a result, we can state with certainty that hopelessness is not a significant factor in shaping how patients with depression view medication adherence. To verify the moderating effects of hopelessness, we conducted a bootstrapping analysis (supported by SPSS version 26.0, PROCESS model 1) for the first step. As shown in Table 3, the interaction between usefulness of IGC and hopelessness was negatively related to attitude toward medication adherence (β=–.153, 95% CI –.2398 to –.0671; *P*<.001). Similar results were obtained for the interaction between positivity of UGC and hopelessness related to attitude toward medication adherence (β=–.135, 95% CI –.2246 to –.0446; *P*=.004). Therefore, hypothesis 8 was partially supported.

Table 3. Results of moderating effect examination.

Path and moderator variable	β (95% CI)	*P* value
Model IGC[a] (n=353)		
Usefulness → medication adherence attitude		
Low hopelessness	.493 (.3211 to .6643)	<.001
Medium hopelessness	.322 (.1879 to .4564)	<.001
High hopelessness	.152 (–.0066 to .3098)	.06
Total effect—usefulness × hopelessness	−.153 (–.2398 to –.0671)	<.001
Model UGC[b] (n=358)		
Positivity → medication adherence attitude		
Low hopelessness	.450 (.2465 to .6533)	<.001
Medium hopelessness	.291 (.1249 to .4580)	<.001
High hopelessness	.133 (–.0582 to .3242)	.17
Total effect—positivity × hopelessness	−.135 (–.2246 to –.0446)	.004

[a]IGC: institution-generated content.

[b]UGC: user-generated content.

To better present the nature of the moderating effects, we adopted the procedures by Aiken et al [94] for computing slopes 1 SD above and below the means of hopelessness and the slopes on the means of hopelessness to plot the interactions. As shown

XSL•FO
RenderX

in Figure 2, with the hopelessness level increasing from low to high, the slopes became more and more steady, indicating that the effects of IGC usefulness and UGC positivity on attitude toward medication adherence were declining. Moreover, from Table 3, we can see that, with the conditions of low or medium hopelessness, the positive effects of IGC usefulness and UGC positivity on medication adherence attitude were all significant (P<.001, 95% CI not including 0), whereas with the condition of high hopelessness, the effects became insignificant (P=.06 for model IGC and P=.17 for model UGC, 95% CI including 0). This finding indicates that the significant positive influence of the usefulness of IGC and the positivity of UGC on attitude toward medication adherence can be eliminated by a high level of hopelessness.

Figure 2. The effects of the interaction between the usefulness of institution-generated content (IGC) and hopelessness and the interaction between the positivity of user-generated content (UGC) and hopelessness on medication adherence attitude. M: mean; M-1SD: 1 SD below the mean; M+1SD: 1 SD above the mean.

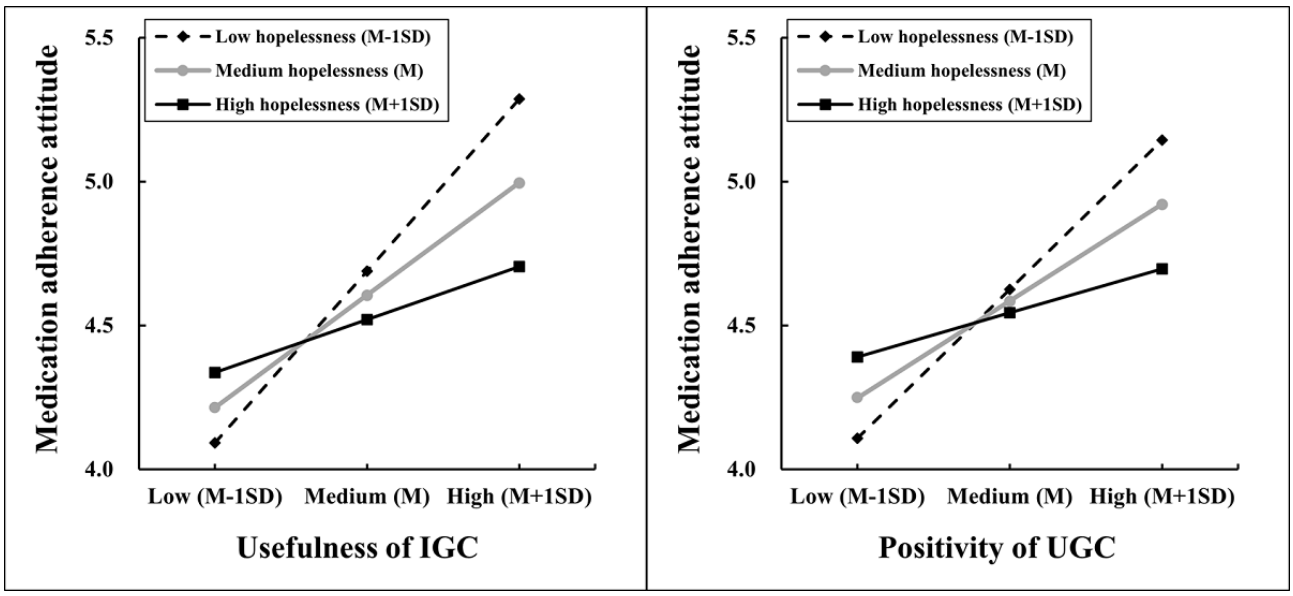

In summary, hopelessness was verified as negatively moderating the effects of the usefulness of IGC and positivity of UGC on attitude toward medication adherence such that the positive effects of IGC usefulness and UGC positivity on attitude toward medication adherence strengthened when patients with depression had a low level of hopelessness but gradually diminished and even disappeared as the hopelessness level continued to increase. Therefore, hypothesis 8 was fully supported.

Results of Robustness Check A

Assessment of Model Fit

We adopted a methodology similar to that of the main study to evaluate the study's soundness in terms of model fit. We used the algorithm by Rogers and Schmitt [89] to parcel the scales of perceived social support and hopelessness into 3 items each. The modified model's fit analysis yielded the following results: χ^2/df=2.4, P<.001, CFI=0.937, TLI=0.925, and RMSEA=0.073. All fit indexes met the acceptable criteria [90], indicating that the hypothesized model fit well enough for follow-up path analyses.

Hypothesis Testing

We conducted hierarchical multiple regression analyses to examine the hypothesized direct effects. In addition, to rigorously compare the differences between the parallel hypotheses, we used the following Z statistic:

where β_x are standardized coefficients and s^2 are the squared SEs of the coefficients [93]. The results are shown in Table 4. All hypotheses were statistically significant (P<.01), ensuring the robustness of our findings and showing that useful IGC and positive UGC had simultaneous positive effects on attitudes toward medication adherence in patients with depression. Perceived social support was confirmed to have a weaker impact on attitude toward medication adherence than perceived value of antidepressants (Z statistic=–4.748∉[–1.96, 1.96]), whereas the differences between other parallel hypotheses were not statistically significant (Z statistic∈[–1.96, 1.96]).

XSL•FO
RenderX

Table 4. Results of direct effect examination of model IGC+UGC (n=270).

Hypotheses and comparison	β (SE)	*P* value	*Z* statistic
Hypothesis 1a: usefulness of IGC[a]→medication adherence attitude	.273 (.082)	<.001	—[b]
Hypothesis 1b: positivity of UGC[c]→medication adherence attitude	.247 (.087)	<.001	—
Comparison: hypothesis 1a and hypothesis 1b	—	—	0.217
Hypothesis 2a: usefulness of IGC→perceived social support	.279 (.078)	<.001	—
Hypothesis 2b: positivity of UGC→perceived social support	.415 (.074)	<.001	—
Comparison: hypothesis 2a and hypothesis 2b	—	—	–1.265
Hypothesis 4a: usefulness of IGC→perceived value of antidepressants	.290 (.064)	<.001	—
Hypothesis 4b: positivity of UGC→perceived value of antidepressants	.225 (.070)	.002	—
Comparison: hypothesis 4a and hypothesis 4b	—	—	0.685
Hypothesis 3: perceived social support→medication adherence attitude	.254 (.067)	<.001	—
Hypothesis 5: perceived value of antidepressants→medication adherence attitude	.681 (.060)	<.001	—
Comparison: hypothesis 3 and hypothesis 5	—	—	–4.748

[a]IGC: institution-generated content.

[b]Not applicable.

[c]UGC: user-generated content.

We used bootstrapping analyses (supported by SPSS version 26.0, PROCESS) to verify the hypothesized mediating and moderating effects. We present the bootstrapping results in Multimedia Appendices 6 and 7. All hypotheses were statistically significant, strengthening the robustness of our research outcomes.

Results of Robustness Check B

Sample Characteristics

The approximate ODC use duration varied from 2 to 795 days among participants in model IGC-B and from 2 to 1268 days among participants in model UGC-B. ODC use frequency reported by participants in both models ranged from once a month (counted as 0.0333 visits per day) to at least once a day (counted as 1 visit per day). The rich diversity of the participants in terms of ODC use situations allowed the following effect examination to be informative. Moreover, as shown in Table 5, the participants recruited for this study covered all categories of ODC use situations. Therefore, we believe that the collected data for robustness check B fit the statistical requirements and study objective well.

XSL•FO
RenderX

Table 5. Participants' online depression community use situations.

Duration[a]	Frequency[b]	Participants, n (%)
Model IGC[c] -B (n=266)		
Low	Low	21 (7.9)
Low	Medium	26 (9.8)
Low	High	10 (3.8)
Medium	Low	31 (11.7)
Medium	Medium	70 (26.3)
Medium	High	24 (9)
High	Low	29 (10.9)
High	Medium	43 (16.2)
High	High	12 (4.5)
Model UGC[d] -B (n=268)		
Low	Low	27 (10.1)
Low	Medium	33 (12.3)
Low	High	15 (5.6)
Medium	Low	29 (10.8)
Medium	Medium	28 (10.4)
Medium	High	13 (4.9)
High	Low	50 (18.7)
High	Medium	52 (19.4)
High	High	21 (7.8)

[a]Low=<1 month; medium=1 to 6 months; high=>6 months.

[b]Low=<1 visit per week; medium=1 to 4 visits per week; high=>4 visits per week.

[c]IGC: institution-generated content.

[d]UGC: user-generated content.

Hypothesis Testing

Hierarchical multiple regression analyses were conducted as the first step for examining the influence of ODC use situations. The standardized path coefficient and corresponding significance are shown in Table 6. As the results show, the direct effects of IGC usefulness and UGC positivity on medication adherence attitude were significantly positive (β>0; *P*<.001), consistent with our hypotheses. The results of model IGC-B and model UGC-B collectively indicate that ODC use duration had no significant relationships with IGC usefulness, UGC positivity, and medication adherence attitude (*P*>.10). In model IGC-B, ODC use frequency showed significantly positive relationships with IGC usefulness and medication adherence attitude (β>0; *P*<.01). In model UGC-B, ODC use frequency showed no significant relationships with UGC positivity and medication adherence attitude (*P*>.10).

XSL•FO
RenderX

Table 6. Results of direct effect examination.

Path	β	*P* value
Model IGC[a] -B (n=266)		
Usefulness→medication adherence attitude	.281	<.001
Duration→usefulness	.038	.41
Duration→medication adherence attitude	.053	.35
Frequency→usefulness	.168	<.001
Frequency→medication adherence attitude	.159	.006
Model UGC[b] -B (n=268)		
Positivity→medication adherence attitude	.266	<.001
Duration→positivity	–.010	.84
Duration→medication adherence attitude	.002	.98
Frequency→positivity	.028	.55
Frequency→medication adherence attitude	.081	.17

[a]IGC: institution-generated content.

[b]UGC: user-generated content.

To further ascertain the role of ODC use situations in the influence of IGC usefulness and UGC positivity on mediation adherence attitude, SPSS (version 26.0, PROCESS model 2) bootstrapping analysis was implemented. As the results in Table 7 show, the effects of IGC usefulness and UGC positivity on medication adherence attitude were insignificant when duration and frequency were both low (*P*>.10, 95% CI including 0), whereas they became stronger with a higher ODC use duration and frequency. When duration and frequency were both high, the effects of IGC usefulness and UGC positivity on medication adherence attitude were strongly significant (*P*<.001, 95% CI not including 0). The interaction between IGC usefulness and ODC use duration and the interaction between UGC positivity and ODC use duration showed no significant relationships with medication adherence attitude (*P*>.10, 95% CI including 0), suggesting that longer ODC use duration could not soundly guarantee a stronger effect of IGC usefulness or UGC positivity. The effects of the interaction between IGC usefulness and ODC use frequency and the interaction between UGC positivity and ODC use frequency on medication adherence attitude were significantly positive (β>0, 95% CI not including 0; *P*<.05). This indicates that higher ODC use frequency could prominently enhance the effects of IGC usefulness and UGC positivity on the attitudes of patients with depression toward medication adherence.

XSL•FO
RenderX

Table 7. Results of interactive effect examination.

Path and duration	Frequency	β (95% CI)	*P* value
Model IGC[a] -B (n=266)[b]			
Usefulness → medication adherence attitude			
Low	Low	.039 (–.2202 to .2989)	.77
Low	Medium	.206 (–.0273 to .4395)	.08
Low	High	.532 (.2371 to .8276)	<.001
Medium	Low	.090 (–.1387 to .3177)	.44
Medium	Medium	.256 (.0556 to .4570)	.01
Medium	High	.583 (.3084 to .8566)	<.001
High	Low	.243 (–.0175 to .5038)	.07
High	Medium	.410 (.1660 to .6538)	.001
High	High	.736 (.4185 to 1.0538)	<.001
Model UGC[c] -B (n=268)[d]			
Positivity → medication adherence attitude			
Low	Low	.121 (–.1115 to .3540)	.31
Low	Medium	.170 (–.0426 to .3830)	.12
Low	High	.426 (.1773 to .6742)	<.001
Medium	Low	.159 (–.0530 to .3714)	.14
Medium	Medium	.208 (.0171 to .3993)	.03
Medium	High	.464 (.2292 to .6983)	<.001
High	Low	.316 (.0179 to .6137)	.04
High	Medium	.365 (.0787 to .6508)	.01
High	High	.620 (.2906 to .9501)	<.001

[a]IGC: institution-generated content.

[b]Total effect—usefulness × duration: β=.0006 (95% CI –.0003 to .0015) and *P*=.17; total effect—usefulness × frequency: β=.7612 (95% CI .3044-1.2180) and *P*=.001.

[c]UGC: user-generated content.

[d]Total effect—positivity × duration: β=.0003 (95% CI –.0002 to .0007) and *P*=.23; total effect—positivity × frequency: β=.4472 (95% CI .0454-.8490) and *P*=.03.

In summary, this robustness check revealed how ODC use situations impacted the value of ODC as a tool to improve the attitudes of patients with depression toward medication adherence. By capturing the nuances between participants with different ODC use situations, the results of this robustness check collectively reduced the possibility that the positive influence of IGC usefulness and UGC positivity on the medication adherence attitudes of patients with depression was spuriously significant. Therefore, the reliability of our findings was further enhanced.

Discussion

Principal Findings and Comparison With Prior Work

In this study, we show that useful IGC and positive UGC in ODCs have positive effects on the attitudes of patients with depression toward medication adherence, with the mediating roles of perceived social support and perceived value of antidepressants and the moderating effect of hopelessness. This study contributes to the theoretical canon in multiple ways.

First, we uncovered new insights about the potential value of ODCs, which extend beyond their direct effects on mental health. ODCs are widely acknowledged as being able to provide a variety of social support to participants, including informational support, emotional support, instrumental support, social network support, and esteem support [8-10,16,18]. Given these advantages of ODCs, researchers generally believe that ODCs are a promising tool for coping with depression. With this grounded assumption, prior work in studies on ODCs has extensively explored ODCs' direct effects on the mental states of individuals experiencing depression. Their findings are rather encouraging. They revealed that ODC engagement can contribute to the improvement of various psychological factors related to depression, such as the sense of loneliness, feeling of being understood, self-efficacy, and well-being [8-10,20,40]. More directly, several studies have found that patients' depressive symptoms decrease over their participation in ODCs

XSL•FO
RenderX

[10,95,96], and the significance of this effect has been confirmed by Goodwin et al [10]. These results offer important insights about ODCs' role in depression treatment, suggesting that ODCs have direct effects on ameliorating patients' mental health and can serve as an informal method to treat depression. However, the potential value of ODCs as an adjunct to formal depression treatments such as medication has not yet been fully appreciated. This niche topic is theoretically important because it indicates the likelihood that ODCs can benefit patients' recovery from depression in diversified ways. Apart from being a direct treatment tool, ODCs may also indirectly impact depression recovery by facilitating patients' appropriate adoption of mental health care services and products. Several scholars have corroborated that participating in OHCs can aid in improving patient compliance in the treatment process [6,11-13], whereas such a benefit in the ODC setting and the field of depression treatment remains unexplored. This study was designed to address this gap, verifying ODCs' positive influence on the perceived value of antidepressants and attitudes toward medication taking among individuals experiencing depression. Our findings provide empirical evidence that the value of ODCs extends beyond influencing the mental status of individuals with depression and serves as an effective complement to formal treatment.

Second, this study introduces the community content perspective into the research on OHCs' value in improving patients' medication adherence. Previous literature on this topic has explored how patients may be affected by the overall idiosyncrasies of OHCs and how their participation patterns impact their ability to benefit from the communities [6,11-13]. However, there is a lack of knowledge about how community content impacts patients' attitudes toward medication adherence. To address this gap, our study borrowed ideas from SOR theory and relevant literature and extracted 2 specific types of ODC content, namely, useful IGC and positive UGC. Our findings reveal that the usefulness of IGC and positivity of UGC do positively relate to patients' willingness to adhere to medications. This offers a fresh perspective on how OHCs can be used as an intervention to address patients' nonadherence problem, thus enriching the existing knowledge base.

Third, this study adds to the understanding of the inner mechanism of OHCs' influence on patients' medication adherence. Wu et al [12] found that health care information in OHCs enhances patients' trust in health care services and, thus, boosts their adherence intentionality. Lu [6] proved that perceived social support plays a significant mediating role in the relationship between patient health information–seeking behaviors and patient compliance. However, little is known about the difference between the mediation effects of the 2 factors: the value judgment of the treatment and the social capital obtained by patients. Our research outcomes ascertain that perceived value of antidepressants plays a stronger mediating effect in the relationship between IGC usefulness and medication adherence attitudes than perceived social support, whereas such a distinction does not appear in the relationship between UGC positivity and attitudes toward medication adherence. This finding demonstrates that the professional health care information within OHCs majorly affects patient compliance by adjusting their perception about the medication's value, whereas patient-to-patient communications impact patients' willingness to follow the treatment through a more mixed mechanism. Thus, this study enriches the knowledge about the inner pathway of the influence of OHCs on patients' medication adherence attitudes.

Fourth, we found that there is a specific limitation to the positive impact of ODCs on patients with depression. Grounded on the feelings-as-information theory, we theorized that the feeling of hopelessness can moderate the relationship between positive external stimuli and medication adherence in patients with depression. Our study discovered that useful IGC and positive UGC are powerful stimuli for enhancing the medication-taking attitudes of patients with depression who remain optimistic. However, when patients lose hope about their lives and the future, they tend to believe that any endeavor to improve their condition is futile, thus undermining the positive effects of ODCs. This finding provides a fundamental boundary condition for the positive effects of ODCs, aiding in clarifying the scenarios in which ODCs can and cannot be adopted as a tool for boosting the adherence of patients with depression to medicinal treatments.

Finally, this study is a call to action for future research in the field of OHCs to emphasize the importance of carefully evaluating the target population's uniqueness. In the treatment process of physiological maladies (eg, HIV and diabetes), the sense of hopelessness is generally believed as a direct reason leading to patients' nonadherence to the medications [71,72], whereas our research outcomes reveal that hopelessness does not directly result in poor medication-taking behaviors among patients with depression. Rather, it serves as a moderator, impacting their processing style for the ODC stimuli and, thus, eliciting different response patterns in terms of medication adherence attitudes. This peculiar phenomenon is conceivably caused by the fact that the current public distrust of the value of antidepressants is prevalent [27,29]. Our findings indicate the significance of recognizing that the same variable can have a vastly differentiated influence mechanism among individuals with different diseases. When theorizing on the impact of OHCs on patients, close scrutiny with regard to the uniqueness of the diseases they experience and the realistic social backgrounds related to the diseases is necessary.

Practical Implications

Our findings have salient practical implications. There is currently a large population of people with depression in China [1], and many of them exhibit poor medication-taking behaviors [2]. The Chinese government has recognized the magnitude of this problem and formulated policies to address it, such as the "Action Plan for Depression Prevention and Treatment" [97], in which the importance of online channels such as ODCs is emphasized. The results of this study confirm that ODCs can be an effective complementary approach to improving the attitudes of patients with depression toward medication adherence and provide meaningful guidelines for practitioners.

One the one hand, governmental bodies, hospitals, or other health-related institutions should ensure that the content they provide is helpful to users. More specifically, they need to

XSL•FO
RenderX

certify the quality of the content they create. Inaccurate information that confuses or misleads users can lead them to take inappropriate actions in the therapeutic process [39]. We urge practitioners to determine what information is genuinely needed by patients with depression and produce high-quality content in response while also purging inaccurate or misleading content from their sites. In contrast, ODC administrators should endeavor to create positive atmospheres within the communities. Our study confirms that ODC content can serve as stimuli for patients and content that presents positive coping strategies can inspire and empower patients to combat their depression. However, negative content such as messages about suicide or self-injurious behaviors [16] is harmful, and administrators should establish guidelines that encourage users to share optimistic content and avoid comments that are offensive or pessimistic. With these efforts, ODCs' positive effects on patients can be strengthened.

Meanwhile, the recommendations we make can also be selectively applied in certain other contexts as depression is a representative mood disorder and often manifests as a comorbidity in patients with other severe diseases [98]. Specifically, the findings of our research can help practitioners better understand how to use OHCs as a tool to enhance treatment adherence for patients with other mood disorders (eg, posttraumatic stress disorder) that have similar symptoms to those of depression or for patients with other illnesses that co-occur with depression or have depression as a comorbidity.

Limitations and Future Research

This study has a few limitations and prospects. Above all, cross-sectional questionnaire survey method was used in this study. This led to several deficiencies. First, we could only measure patients' attitudes toward medication adherence at the time of the investigation. We were unable to capture the dynamic changes in participants' attitudes from a within-subject perspective. As a result, even though we had examined the distinctions between participants with different ODC use situations, we could not fully exclude the potential that the effects of IGC usefulness and UGC positivity on participants' medication adherence attitudes were spuriously significant. Meanwhile, although the attitudes of patients with depression toward medication adherence are known to be significantly associated with their adherence performance [65], we cannot ensure that ODCs' positive effects on medication adherence attitudes are long-lasting and affect actual medication-taking behaviors. Second, as this study used self-reported data to examine the hypothesized model, the reliability of our results might be subject to the quality of the responses that participants provided. Although we made multiple efforts regarding the study design, data collection, and data analyses to avoid this risk, the robustness of our findings still cannot be fully guaranteed. Therefore, we suggest that future researchers conduct a longitudinal study tracing participants' changes in attitudes and objective behaviors to address these remaining concerns.

Moreover, we included patients with both diagnosed depression and self-reported depression in our samples. While this practice helped make our samples more representative of the overall population of individuals with depression, it may have introduced a degree of measurement inaccuracy as people with self-reported depression had no actual experience taking medications and so their answers were based solely on their subjective feelings about antidepressants. Although we set this factor as one of the control variables to reduce the severity of this interference, whether the improvement in their perceived value of antidepressants and medication adherence attitude can be maintained after they start to take the medication remains unexamined. Hence, we encourage future research to explore the deviation between the perceptions of and attitudes toward antidepressants of patients with depression before and after taking them.

Furthermore, we split ODC content into IGC and UGC in this research, but our comparison between their influence on patients with depression was inadequate due to methodological limitations. First, recall bias exists in questionnaire-based surveys and possibly weakened the preciseness of respondents' perceptions of the ODC content. Second, the attributes of IGC and UGC investigated in this research were different—usefulness and positivity, respectively. Consequently, our results cannot rigorously reflect the difference between the influence of IGC and UGC on patients. In our view, the attributes and providers of ODC content are closely linked to the potential for patients to derive substantial benefits from ODC involvement. To attain a deeper and more accurate understanding of the variability of ODC or OHC content, we suggest that future researchers conduct randomized controlled trials to compare the effects of IGC and UGC with the same features on patients.

Finally, the respondents of our survey were all Chinese, and thus, they share a common cultural background. Influenced by Confucian thought for thousands of years, Chinese culture emphasizes that individuals should possess self-control and resilience [17], leading to depression being frequently misunderstood as a kind of sentimentality or sign of weakness in China. As a result, many patients with depression face discrimination and marginalization in society. Under this background, people with depression in China are likely to rely more on ODCs, where the communications are anonymous and emotional support can be accessed [8,40]. Given this cultural influence, the generalizability and external validity of our findings may be limited. Consequently, we recommend that future research re-examine similar models in different cultural contexts.

Conclusions

Grounded on the SOR and feelings-as-information theories, this study proposed a research model to examine the relationship between ODC content and the attitudes of patients with depression toward medication adherence. The results corroborate findings that the attitudes of patients with depression toward medication adherence can be improved as an outcome of the increase in perceived social support and perceived value of antidepressants, initially contributed by the stimulation of positive ODC content (useful IGC and positive UGC). Thereby, this study offers a novel solution to the problem of poor medication adherence among patients with depression in China.

XSL•FO
RenderX

Our findings also provide suggestions for health care practitioners to generate online content that precisely matches the informational needs of patients with depression and recommend that ODC service providers make more efforts to regulate the community atmosphere. Moreover, this study verifies that hopelessness serves as an essential boundary condition for the effect of positive ODC content on the medication adherence attitudes of patients with depression. This finding warns health care practitioners that ODC interventions cannot be used as the only approach to addressing the problem of poor medication taking among patients with severe depressive symptoms.

Acknowledgments

The authors deeply appreciate the editors and reviewers for their valuable and constructive feedback, which contributed to an improved paper. No part of the manuscript generation process involved the use of generative artificial intelligence. This work was supported by the Department of Education of Sichuan Province (College Student Innovation and Entrepreneurship Training Program; grant 20221065100121) and Southwestern University of Finance and Economics.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

This study was conceived and designed by RC and XF. RC, ML, KL, and LB were responsible for conducting the field study and compiling and analyzing the data. RC drafted the paper with critical review and revision from XF. All authors have read and approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample recruitment advertisement.
[DOCX File , 22 KB - jmir_v26i1e56166_app1.docx]

Multimedia Appendix 2
Sample characteristics.
[DOCX File , 22 KB - jmir_v26i1e56166_app2.docx]

Multimedia Appendix 3
Survey questions.
[DOCX File , 31 KB - jmir_v26i1e56166_app3.docx]

Multimedia Appendix 4
Mean, SD, and Cronbach α values for all variables.
[DOCX File , 26 KB - jmir_v26i1e56166_app4.docx]

Multimedia Appendix 5
Convergent validity and discriminant validity.
[DOCX File , 23 KB - jmir_v26i1e56166_app5.docx]

Multimedia Appendix 6
Results of the mediating effect examination of model IGC+UGC (n=270).
[DOCX File , 17 KB - jmir_v26i1e56166_app6.docx]

Multimedia Appendix 7
Results of the moderating effect examination of model IGC+UGC (n=270).
[DOCX File , 17 KB - jmir_v26i1e56166_app7.docx]

References

XSL•FO
RenderX

1. Huang Y, Wang YU, Wang H, Liu Z, Yu X, Yan J, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study. Lancet Psychiatry 2019 Mar;6(3):211-224. [doi: 10.1016/S2215-0366(18)30511-X] [Medline: 30792114]
2. Yu C, Lyu J, Chen Y, Guo Y, Paul S, Bian Z, et al. Epidemiology of major depressive episodes among Chinese adults aged 30-79 years: data from the China Kadoorie Biobank. Zhonghua Liu Xing Bing Xue Za Zhi 2015 Jan;36(1):52-56. [Medline: 25876866]
3. Feng G, Tian J, Zhang L, Zhao F, Wu Y, Qin X. Research progress of medication adherence in depressed patients. J Shanxi Med Univ 2013;44(4):313-316. [doi: 10.3969/J.ISSN.1007-6611.2013.04.020]
4. Banerjee S, Varma RP. Factors affecting non-adherence among patients diagnosed with unipolar depression in a psychiatric department of a tertiary hospital in Kolkata, India. Depress Res Treat 2013;2013:809542-809512. [doi: 10.1155/2013/809542] [Medline: 24381752]
5. Hamrin V, Sinclair VG, Gardner V. Theoretical approaches to enhancing motivation for adherence to antidepressant medications. Arch Psychiatr Nurs 2017 Apr;31(2):223-230. [doi: 10.1016/j.apnu.2016.09.004] [Medline: 28359437]
6. Lu X. The effects of patient health information seeking in online health communities on patient compliance in china: social perspective. J Med Internet Res 2023 Jan 09;25:e38848. [doi: 10.2196/38848] [Medline: 36622741]
7. Pan T, Lu Y. Research on influencing factors of user participation behavior based on SOR model in online health community. Inf Doc Serv 2022;43(2):76-83. [doi: 10.12154/j.qbzlgz.2022.02.009]
8. Sun S, Kang J. Health communication from a social support perspective: an empirical study of patient online communities. J Shanxi Univ 2022;45(01):66-73. [doi: 10.13451/j.cnki.shanxi.univ(phil.soc.).2022.01.008]
9. Nimrod G. Challenging the Internet paradox: online depression communities and well-being. Int J Internet Sci 2013;8(1):30-48 [FREE Full text]
10. Goodwin BC, Ford DE, Hsiung RC, Houston TK, Fogel J, Van Voorhees BW. First, do no harm: referring primary care patients with depression to an internet support group. Telemed J E Health 2018 Jan;24(1):37-44. [doi: 10.1089/tmj.2017.0004] [Medline: 28657881]
11. Yin Z, Warner J, Song L, Hsueh PY, Chen CH, Malin B. Learning hormonal therapy medication adherence from an online breast cancer forum. In: Social Web and Health Research. Cham, Switzerland: Springer; 2019:233-257.
12. Wu JJ, Chen YM, Talley PC, Kuo KM. Does online community participation contribute to medication adherence? An empirical study of patients with chronic diseases. Int J Environ Res Public Health 2021 May 12;18(10):5100. [doi: 10.3390/ijerph18105100] [Medline: 34065820]
13. Lu X, Zhang R. Impact of patient information behaviours in online health communities on patient compliance and the mediating role of patients' perceived empathy. Patient Educ Couns 2021 Jan;104(1):186-193. [doi: 10.1016/j.pec.2020.07.001] [Medline: 32665071]
14. Ratcliffe M. Experiences of Depression: A Study in Phenomenology. Oxford, UK: Oxford University Press; 2015.
15. Barney LJ, Griffiths KM, Banfield MA. Explicit and implicit information needs of people with depression: a qualitative investigation of problems reported on an online depression support forum. BMC Psychiatry 2011 May 18;11(1):88. [doi: 10.1186/1471-244X-11-88] [Medline: 21592335]
16. Nimrod G. From knowledge to hope: online depression communities. Int J Disabil Hum Dev 2012;11(1):23-30. [doi: 10.1515/ijdhd.2012.009]
17. Li G, Zhou X, Lu T, Yang J, Gu N. SunForum: understanding depression in a Chinese online community. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computin. 2016 Presented at: CSCW '16; February 27-March 2, 2016; San Francisco, CA p. 515-526. [doi: 10.1145/2818048.2819994]
18. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020 Mar 15;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
19. Johnson CM. A survey of current research on online communities of practice. Internet High Educ 2001 Jan;4(1):45-60. [doi: 10.1016/S1096-7516(01)00047-1]
20. Nimrod G. Online depression communities: members' interests and perceived benefits. Health Commun 2013 Jul 18;28(5):425-434. [doi: 10.1080/10410236.2012.691068] [Medline: 22809441]
21. Mehrabian A, Russell JA. An Approach to Environmental Psychology. Cambridge, UK: MIT Press; 1974.
22. Zhou P, Zhao S, Ma Y, Liang C, Zhu J. What influences user participation in an online health community? The stimulus-organism-response model perspective. Aslib J Inform Manag 2022 Aug 10;75(2):364-389. [doi: 10.1108/ajim-12-2021-0383]
23. Zhou T. Examining users' knowledge sharing behaviour in online health communities. Data Technol Appl 2019 Oct 18;53(4):442-455. [doi: 10.1108/dta-03-2019-0044]
24. Cao Y, Li J, Qin X, Hu B. Examining the effect of overload on the mHealth application resistance behavior of elderly users: an SOR perspective. Int J Environ Res Public Health 2020 Sep 12;17(18):6658. [doi: 10.3390/ijerph17186658] [Medline: 32932679]
25. Jacoby J. Stimulus‐organism‐response reconsidered: an evolutionary step in modeling (consumer) behavior. J Consum Psychol 2008 Jan 25;12(1):51-57. [doi: 10.1207/s15327663jcp1201_05]

XSL•FO
RenderX

26. Woods A, Solomonov N, Liles B, Guillod A, Kales HC, Sirey JA. Perceived social support and interpersonal functioning as predictors of treatment response among depressed older adults. Am J Geriatr Psychiatry 2021 Aug;29(8):843-852. [doi: 10.1016/j.jagp.2020.12.021] [Medline: 33419660]
27. Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc 2011 Apr;86(4):304-314 [FREE Full text] [doi: 10.4065/mcp.2010.0575] [Medline: 21389250]
28. Jacob SA, Ab Rahman AF, Hassali MA. Attitudes and beliefs of patients with chronic depression toward antidepressants and depression. Neuropsychiatr Dis Treat 2015 May:1339. [doi: 10.2147/ndt.s82563]
29. Mert DG, Turgut NH, Kelleci M, Semiz M. Perspectives on reasons of medication nonadherence in psychiatric patients. Patient Prefer Adherence 2015;9:87-93 [FREE Full text] [doi: 10.2147/PPA.S75013] [Medline: 25609930]
30. Juliana J, Aditi B, Nagoya R, Wisnalmawati W, Nurcholifah I. Tourist visiting interests: the role of social media marketing and perceived value. Int J Data Netw Sci 2022;6(2):469-476. [doi: 10.5267/j.ijdns.2021.12.007]
31. Schwarz N. Feelings as information: informational and motivational functions of affective states. In: Higgins ET, Sorrentino RM, editors. Handbook of Motivation and Cognition: Foundations of Social Behavior, Volume 2. New York, NY: The Guilford Press; 1990:527-561.
32. Keltner D, Ellsworth PC, Edwards K. Beyond simple pessimism: effects of sadness and anger on social perception. J Pers Soc Psychol 1993;64(5):740-752. [doi: 10.1037/0022-3514.64.5.740]
33. Schwarz N, Clore GL. Mood as information: 20 years later. Psychol Inq 2003 Oct;14(3-4):296-303. [doi: 10.1080/1047840X.2003.9682896]
34. Kostyk A, Niculescu M, Leonhardt JM. Less is more: online consumer ratings' format affects purchase intentions and processing. J Consum Behav 2017 Mar 02;16(5):434-441. [doi: 10.1002/cb.1643]
35. Shin S, Chung N, Xiang Z, Koo C. Assessing the impact of textual content concreteness on helpfulness in online travel reviews. J Travel Res 2018 May 03;58(4):579-593. [doi: 10.1177/0047287518768456]
36. Zhou Q, Li B, Scheibenzuber C, Li H. Fake news land? Exploring the impact of social media affordances on user behavioral responses: a mixed-methods research. Comput Hum Behav 2023 Nov;148:107889. [doi: 10.1016/j.chb.2023.107889]
37. Gentles SJ, Lokker C, McKibbon KA. Health information technology to facilitate communication involving health care providers, caregivers, and pediatric patients: a scoping review. J Med Internet Res 2010;12(2):e22 [FREE Full text] [doi: 10.2196/jmir.1390] [Medline: 20562092]
38. Zhou J, Fan T. Understanding the factors influencing patient e-health literacy in online health communities (OHCs): a social cognitive theory perspective. Int J Environ Res Public Health 2019 Jul 10;16(14):2455. [doi: 10.3390/ijerph16142455] [Medline: 31295924]
39. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 02;41(1):433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
40. Powell J, McCarthy N, Eysenbach G. Cross-sectional survey of users of internet depression communities. BMC psychiatry 2003 Dec 10;3(1):1-7. [doi: 10.1186/1471-244x-3-19]
41. Borah P, Xiao X. The importance of 'likes': the interplay of message framing, source, and social endorsement on credibility perceptions of health information on Facebook. J Health Commun 2018 Mar 30;23(4):399-411. [doi: 10.1080/10810730.2018.1455770] [Medline: 29601271]
42. Dutta-Bergman M. Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation. J Med Internet Res 2003 Sep;5(3):e21 [FREE Full text] [doi: 10.2196/jmir.5.3.e21] [Medline: 14517112]
43. Lienemann BA, Siegel JT, Crano WD. Persuading people with depression to seek help: respect the boomerang. Health Commun 2013 Oct;28(7):718-728. [doi: 10.1080/10410236.2012.712091] [Medline: 23066999]
44. Powell J, Clarke A. Information in mental health: qualitative study of mental health service users. Health Expect 2006 Dec 04;9(4):359-365. [doi: 10.1111/j.1369-7625.2006.00403.x] [Medline: 17083562]
45. Sun GC, Hsu MC, Moyle W, Lin MF, Creedy D, Venturato L. Mediating roles of adherence attitude and patient education on antidepressant use in patients with depression. Perspect Psychiatr Care 2011 Jan;47(1):13-22. [doi: 10.1111/j.1744-6163.2010.00257.x] [Medline: 21418069]
46. Sigurdsson E, Olafsdóttir T, Gottfredsson M. Public views on antidepressant treatment: lessons from a national survey. Nord J Psychiatry 2008 Jul 12;62(5):374-378. [doi: 10.1080/08039480801984156] [Medline: 18752102]
47. Shumaker SA, Brownell A. Toward a theory of social support: closing conceptual gaps. J Soc Issues 2010 Apr 14;40(4):11-36. [doi: 10.1111/j.1540-4560.1984.tb01105.x]
48. Haber MG, Cohen JL, Lucas T, Baltes BB. The relationship between self-reported received and perceived social support: a meta-analytic review. Am J Community Psychol 2007 Mar;39(1-2):133-144. [doi: 10.1007/s10464-007-9100-9] [Medline: 17308966]
49. Sarason BR, Sarason IG, Pierce GR. Traditional views of social support and their impact on assessment. In: Sarason BR, Sarason IG, Pierce GR, editors. Social Support: An Interactional View. Oxford, UK: John Wiley & Sons Press; 1990:9-25.
50. Cutrona CE, Russell DW. Type of social support and specific stress: toward a theory of optimal matching. In: Sarason BR, Sarason IG, Pierce GR, editors. Social Support: An Interactional View. Oxford, UK: John Wiley & Sons Press; 1990:319-366.

XSL•FO
RenderX

51. Hobfoll SE, Stokes JP. The process and mechanics of social support. In: Duck S, Hay DF, Hobfoll SE, Ickes W, Montgomery BM, editors. Handbook of Personal Relationships: Theory, Research and Interventions. Oxford, UK: John Wiley & Sons Press; 1988:497-517.
52. Zeithaml VA. Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. J Mark 1988 Jul 01;52(3):2-22. [doi: 10.1177/002224298805200302]
53. Woodruff RB. Customer value: the next source for competitive advantage. J Acad Mark Sci 1997 Mar;25(2):139-153. [doi: 10.1007/bf02894350]
54. Snoj B, Korda AP, Mumel D. The relationships among perceived quality, perceived risk and perceived product value. J Prod Brand Manag 2004;13(3):156-167. [doi: 10.1108/10610420410538050]
55. Yuan Z, Chen Z, Xue M, Zhang J, Leng L. Application of antidepressants in depression: a systematic review and meta-analysis. J Clin Neurosci 2020 Oct;80:169-181. [doi: 10.1016/j.jocn.2020.08.013] [Medline: 33099342]
56. Braund TA, Tillman G, Palmer DM, Gordon E, Rush AJ, Harris AW. Antidepressant side effects and their impact on treatment outcome in people with major depressive disorder: an iSPOT-D report. Transl Psychiatry 2021 Aug 04;11(1):417 [FREE Full text] [doi: 10.1038/s41398-021-01533-1] [Medline: 34349116]
57. Hieronymus F, Lisinski A, Eriksson E, Østergaard SD. Do side effects of antidepressants impact efficacy estimates based on the Hamilton Depression Rating Scale? A pooled patient-level analysis. Transl Psychiatry 2021 Apr 27;11(1):249. [doi: 10.1038/s41398-021-01364-0] [Medline: 33907188]
58. Sánchez-Fernández R, Iniesta-Bonillo M. The concept of perceived value: a systematic review of the research. Mark Theory 2016 Aug 11;7(4):427-451. [doi: 10.1177/1470593107083165]
59. Parasuraman A. Reflections on gaining competitive advantage through customer value. J Acad Mark Sci 1997 Mar;25(2):154-161. [doi: 10.1007/bf02894351]
60. Laroche M, Vinhal Nepomuceno M, Richard MO. How do involvement and product knowledge affect the relationship between intangibility and perceived risk for brands and product categories? J Consum Mark 2010;27(3):197-210. [doi: 10.1108/07363761011038275]
61. Park CW, Mothersbaugh DL, Feick L. Consumer knowledge assessment. J Consum Res 1994 Jun;21(1):71. [doi: 10.1086/209383]
62. Jiao Y, Li Z, Shen Z. A study on the mechanism of new product diffusion under online social network context: based on the process of peer influence. J Ind Eng Eng Manag 2020;34(05):105-113. [doi: 10.13587/j.cnki.jieem.2020.05.011]
63. Hu X, Huang Q, Zhong X, Davison RM, Zhao D. The influence of peer characteristics and technical features of a social shopping website on a consumer's purchase intention. Int J Inf Manage 2016 Dec;36(6):1218-1230. [doi: 10.1016/j.ijinfomgt.2016.08.005]
64. Willis E. Applying the health belief model to medication adherence: the role of online health communities and peer reviews. J Health Commun 2018;23(8):743-750. [doi: 10.1080/10810730.2018.1523260] [Medline: 30280998]
65. Aznar-Lou I, Iglesias-González M, Rubio-Valera M, Peñarrubia-Maria MT, Mendive JM, Murrugarra-Centurión AG, et al. Diagnostic accuracy and treatment approach to depression in primary care: predictive factors. Fam Pract 2019 Jan 25;36(1):3-11. [doi: 10.1093/fampra/cmy098] [Medline: 30423158]
66. Schwarz N, Bless H, Bohner G. Mood and persuasion: affective states influence the processing of persuasive communications. Adv Exp Soc Psychol 1991;24:161-199. [doi: 10.1016/S0065-2601(08)60329-9]
67. Marchetti I. Hopelessness: a network analysis. Cogn Ther Res 2018 Nov 17;43(3):611-619. [doi: 10.1007/s10608-018-9981-y]
68. Maier SF, Seligman ME. Learned helplessness: theory and evidence. J Exp Psychol Gen 1976 Mar;105(1):3-46. [doi: 10.1037/0096-3445.105.1.3]
69. Anastasiades MH, Gupton OA, Fritz A, Caldaza PJ, Stillman MA. Diabetes, depression, and nonadherence: exploring hopelessness as a mediating factor: a preliminary study. Ment Health Fam Med 2016;12:243-249 [FREE Full text]
70. Grenard JL, Munjas BA, Adams JL, Suttorp M, Maglione M, McGlynn EA, et al. Depression and medication adherence in the treatment of chronic diseases in the United States: a meta-analysis. J Gen Intern Med 2011 Oct 1;26(10):1175-1182. [doi: 10.1007/s11606-011-1704-y] [Medline: 21533823]
71. Hosek SG, Harper GW, Domanico R. Predictors of medication adherence among HIV-infected youth. Psychol Health Med 2005 May;10(2):166-179 [FREE Full text] [doi: 10.1080/1354350042000326584] [Medline: 25705113]
72. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000 Jul 24;160(14):2101-2107. [doi: 10.1001/archinte.160.14.2101] [Medline: 10904452]
73. Joinson A. Social desirability, anonymity, and internet-based questionnaires. Behav Res Methods Instrum Comput 1999 Aug;31(3):433-438. [doi: 10.3758/bf03200723] [Medline: 10502866]
74. Biemer PP, Lyberg LE. Introduction to Survey Quality. Oxford, UK: John Wiley & Sons Press; 2003.
75. Goldammer P, Annen H, Stöckli PL, Jonas K. Careless responding in questionnaire measures: detection, impact, and remedies. Leadersh Q 2020 Aug;31(4):101384. [doi: 10.1016/j.leaqua.2020.101384]
76. Lee J, Hong IB. Predicting positive user responses to social media advertising: the roles of emotional appeal, informativeness, and creativity. Int J Inf Manage 2016 Jun;36(3):360-373. [doi: 10.1016/j.ijinfomgt.2016.01.001]

XSL•FO
RenderX

77. Price LL, Arnould EJ. Commercial friendships: service provider–client relationships in context. J Mark 1999 Oct 01;63(4):38-56. [doi: 10.1177/002224299906300405]
78. Fong CJ, Williams KM, Williamson ZH, Lin S, Kim YW, Schallert DL. “Inside out”: appraisals for achievement emotions from constructive, positive, and negative feedback on writing. Motiv Emot 2017 Dec 19;42(2):236-257. [doi: 10.1007/s11031-017-9658-y]
79. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The multidimensional scale of perceived social support. J Pers Assess 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
80. Sweeney JC, Soutar GN. Consumer perceived value: the development of a multiple item scale. J Retail 2001 Jun;77(2):203-220. [doi: 10.1016/s0022-4359(01)00041-0]
81. Lewis SJ, Abell N. Development and evaluation of the adherence attitude inventory. Res Soc Work Pract 2002 Jan 01;12(1):107-123. [doi: 10.1177/104973150201200108]
82. Beck AT, Weissman A, Lester D, Trexler L. The measurement of pessimism: the hopelessness scale. J Consult Clin Psychol 1974 Dec;42(6):861-865. [doi: 10.1037/h0037562] [Medline: 4436473]
83. Rimmer T, Weaver D. Different questions, different answers? Media use and media credibility. Journal Q 1987 Mar 01;64(1):28-44. [doi: 10.1177/107769908706400104]
84. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
85. Harman HH. Modern Factor Analysis. Chicago, IL: University of Chicago Press; 1976.
86. Hair Jr JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. 7th edition. Saddle River, NJ: Prentice Hall Press; 2009.
87. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb 01;18(1):39-50. [doi: 10.1177/002224378101800104]
88. Wu Y, Wen Z. Item parceling strategies in structural equation modelling. Adv Psychol Sci 2011;19(12):1859-1867 [FREE Full text]
89. Rogers WM, Schmitt N. Parameter recovery and model fit using multidimensional composites: a comparison of four empirical parceling algorithms. Multivariate Behav Res 2004 Jul;39(3):379-412. [doi: 10.1207/s15327906mbr3903_1]
90. Mulaik SA, James LR, Van Alstine J, Bennett N, Lind S, Stilwell CD. Evaluation of goodness-of-fit indices for structural equation models. Psychol Bull 1989 May;105(3):430-445. [doi: 10.1037/0033-2909.105.3.430]
91. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986;51(6):1173-1182. [doi: 10.1037/0022-3514.51.6.1173]
92. Zhao X, Lynch JG, Chen Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. J Consum Res 2010 Aug 01;37(2):197-206. [doi: 10.1086/651257]
93. Clogg CC, Petkova E, Haritou A. Statistical methods for comparing regression coefficients between models. Am J Sociol 1995 Mar;100(5):1261-1293. [doi: 10.1086/230638]
94. Aiken LS, West SG, Reno RR. Multiple Regression: Testing and Interpreting Interactions. Thousand Oaks, CA: Sage Publications; Jan 1994.
95. Houston TK, Cooper LA, Ford DE. Internet support groups for depression: a 1-year prospective cohort study. Am J Psychiatry 2002 Dec;159(12):2062-2068. [doi: 10.1176/appi.ajp.159.12.2062] [Medline: 12450957]
96. Lu Y, Pan T, Liu J, Wu J. Does usage of online social media help users with depressed symptoms improve their mental health? Empirical evidence from an online depression community. Front Public Health 2020 Jan 25;8:581088. [doi: 10.3389/fpubh.2020.581088] [Medline: 33569367]
97. Action plan for depression prevention and treatment. China's National Health Commission. URL: http://www.nhc.gov.cn/jkj/s7914/202009/a63d8f82eb53451f97217bef0962b98f.shtml [accessed 2024-01-01]
98. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Primers 2020 Aug 20;6(1):69. [doi: 10.1038/s41572-020-0200-2] [Medline: 32820163]

Abbreviations

AVE: average variance extracted
CFI: comparative fit index
CR: composite reliability
IGC: institution-generated content
ODC: online depression community
OHC: online health community
RMSEA: root mean square error of approximation
SOR: stimulus-organism-response
TLI: Tucker-Lewis index
UGC: user-generated content

XSL•FO
RenderX

Edited by A Coristine; submitted 08.01.24; peer-reviewed by L Xinyi, X Liu; comments to author 12.03.24; revised version received 08.07.24; accepted 09.10.24; published 19.11.24.

Please cite as:
Chen R, Fu X, Liu M, Liao K, Bai L
Online Depression Communities as a Complementary Approach to Improving the Attitudes of Patients With Depression Toward Medication Adherence: Cross-Sectional Survey Study
J Med Internet Res 2024;26:e56166
URL: https://www.jmir.org/2024/1/e56166
doi:10.2196/56166
PMID:

©Runnan Chen, Xiaorong Fu, Mochi Liu, Ke Liao, Lifei Bai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Communal Load Sharing of Miscarriage Experiences: Thematic Analysis of Social Media Community Support

Julia Dubbelman[1*], BSc; Jonelle Ooms[1*], BSc, MSc; Laura Havgry[1*], BSc, MSc; Lianne Simonse[1*], MSc, PhD

Department of Design, Organization, and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands
[*]all authors contributed equally

Corresponding Author:
Lianne Simonse, MSc, PhD
Department of Design, Organization, and Strategy
Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15
Delft, 2628 CE
Netherlands
Phone: 31 616844170
Email: L.W.L.Simonse@tudelft.nl

Abstract

Background: Miscarriage is a common experience, affecting 15% of recognized pregnancies, but societal ignorance and taboos often downplay the mental distress and personal impact following a miscarriage. Emerging stories on social media in which women express their miscarriage grief are breaking such taboos. Research in the area of online health communities is increasingly focused on studying how people share their health experiences on social media. However, a clear understanding on the social support involved in this type of sharing of health experiences is lacking.

Objective: This study explored the use of Instagram in sharing miscarriage experiences, guided by the following research question: How is social community support given to women who share their miscarriage experiences on social media? Considering that social media is increasingly used as a source of social support, in this study, we chose Instagram as the social media platform. The purpose of this research was to create a better understanding of how social media provides support in expressing personal miscarriage experiences and how people engage with such posts.

Methods: This study used a qualitative inductive research method in which a phenomenological strategy and thematic analysis were followed to create a comprehensive understanding of the social community support phenomenon. The dataset was established from a sample of 258 Instagram posts and 736 comments collected over a period of 6 months after initial posts and from 6 different women. These data were categorized and clustered through a thematic analysis.

Results: Three themes were identified: (1) storytelling of emotional turmoil and grief after miscarriage, (2) sharing positivity amidst miscarriage grief, and (3) mentioning personal medical information about miscarriage. Theme 1 represents the emotional experience of women who have had a miscarriage. It encompasses the initial posts that included miscarriage storytelling that express deep grief and mental distress and the emotional impact on both the posters and the commenters. Theme 2 highlights the importance of finding moments of joy and positivity in the midst of mental distress and pain. The posts shared with the online community convey a sense of moving forward and a refusal to let grief become the defining aspect of one's life. Theme 3 focuses on sharing medical and practical advice. This theme includes posts and comments about medications, in vitro fertilization procedures, hospital experiences, and personal physical symptoms.

Conclusions: As an overarching theme for this social support phenomenon, we introduce the term *communal load sharing* to describe the therapeutic role of social media in helping women cope with miscarriage by providing a platform for sharing similar experiences, breaking social taboos, and fostering *load sharing*.

(J Med Internet Res 2024;26:e56680) doi:10.2196/56680

KEYWORDS

miscarriage; miscarriage grief; online health communities; thematic analysis; social support; communal load sharing; peer-to-peer support

XSL•FO
RenderX

Introduction

Social Relevance

Miscarriage is a more common phenomenon than people seem to realize, occurring in 15% of recognized pregnancies [1]. Miscarriage is defined by the European Society of Human Reproduction and Embryology as the loss of pregnancy before 22 weeks of gestation [1]. One in 4 women will experience a miscarriage during their lifetime [2], so why do so few women know how common it is and what effects a miscarriage can have on a person's life? Past research has indicated that societal ignorance and taboos often overshadow the profound effect miscarriage has on a person [3]. Over half of the women who have a miscarriage will experience lasting psychological effects [1,4-6]. Overall, there is an imbalance in research, as the physical effects of a miscarriage are prioritized over other effects, such as social-psychological effects. This study investigates these social-psychological effects, concentrating on postmiscarriage interactions on social media. Exploring the experiences women have with sharing their stories of miscarriage on social media can facilitate an understanding of how social media platforms can be used as a form of social community support [7]. Collecting and analyzing multiple Instagram users' experiences can yield a holistic overview of the miscarriage experience. Through the longitudinal analysis of social media use, this research aims to create a better understanding of the role that social media plays in offering social support following a miscarriage experience. Specifically, this research considers Instagram users who shared their miscarriage experiences. Instagram is utilized as a source of data because it is currently one of the biggest social media platforms with >2 billion monthly users [8]. Against the background of online health communities (OHCs) research, the use of social media is a recognized direction in which OHCs are headed in the provision of social community support [9].

Structure of the Paper

This paper contributes by unpacking important themes for social support on social media OHCs and introduces the overarching term *communal load sharing* to describe the therapeutic role of social media in breaking social taboos and fostering the sharing of similar experiences, ultimately helping women to cope with miscarriage grief.

The subsequent section provides a review of the theoretical background regarding different factors influencing the experience of a miscarriage in relation to social support. This review leads to the research question, followed by the Methods section in which the steps of thematic analysis are described with respect to the sample and analysis of the OHC created by sharing miscarriage experiences on Instagram. This is followed by the Results section, which presents the key findings on identified themes, evidence quotes, and visuals. Finally, the theoretical and practical implications are discussed, including limitations and future research opportunities.

Theoretical Background

Mental Distress After a Miscarriage

While the physical consequences of miscarriages are well known, the psychological impacts are often overlooked. Half of the women experiencing a miscarriage face psychological distress, including anxiety, depression, guilt, and shame [4,10]. Moreover, some women also struggle with feelings of failure or envy of others' children, which can lead to isolation [11]. Furthermore, concerns about future pregnancies, including the fear of experiencing another miscarriage, are commonly reported [4]. In addition, it is important to recognize that the impact of miscarriage leads to an increased use of mental health services and prescription drugs [12]. These psychological consequences may not always be evident in social interactions with health care providers, family, and friends, as they can lack outward physical manifestations [1]. Although miscarriage is often perceived primarily as the woman's experience, it is important to recognize the impact it has on her partner. Despite their own substantial grief, some women report that the support that is offered overlooks their partners [13]. Men tend to grieve less intensely and for shorter periods than women following a miscarriage and have been found to use different coping mechanisms [14]. These differences in response to a miscarriage may contribute to misunderstandings and conflicts in the relationship, and studies have shown that couples have an increased risk of separation following a miscarriage [11,12]. Individuals can often feel unsupported due to differences in outward physical manifestations and reactions to miscarriage [15]. The social sharing of individual experiences might benefit the postmiscarriage coping and recovery process and decrease mental distress. However, a clear understanding of such social sharing and support is currently lacking.

Social Taboo Surrounding Miscarriage

Discussing miscarriages can be a sensitive and complex topic because of the social taboo that surrounds it. This taboo not only makes it difficult for individuals and couples to find support but it also perpetuates harmful myths and misconceptions surrounding miscarriage [3]. A common cultural habit of waiting until the 12th week to announce a pregnancy can further contribute to the negative psychological effects in the event of a miscarriage [16]; if a miscarriage occurs before friends and family are informed, the grieving process may be more difficult. The social stigma also persists in professional settings [17], where women frequently lack recovery time following a miscarriage [12]. Overall, this highlights the importance of breaking down the social stigma and providing adequate social community support for women experiencing miscarriage. More open and supportive communication around this topic and the creation of supportive environments both online and offline might be beneficial. By participating in OHCs through social media, individuals and couples who have experienced a miscarriage might feel less alone and more understood, which ultimately can lead to positive effects on processing the loss [12]. That is why, in this study, we explore the social community support of those who break the social taboo and share their miscarriage experience on social media.

XSL•FO
RenderX

Social Support Following Miscarriage

Prior research has indicated that one of the crucial factors for helping individuals through the experience of a miscarriage is social support [4,6]. This type of support is often offered by the patient's partner, health care professionals, relatives, or friends. Studies have found that a lack of social support leads to isolation from loved ones [1,11]. However, in recent years, there has been a shift in the way people seek health information and emotional support in OHCs [18,19]. Traditionally, OHCs were forums administered by medical foundations or companies, providing a certain level of reputable health information [20]. However, with the rapidly increasing popularity of social media, a form of OHCs started emerging from these platforms as well [21]. In this research, we study social media–based OHCs that have the same function of offering a comparatively safe space for people dealing with miscarriage to share their experiences, find coping strategies, reduce stigma, and share emotional support [22]. As social media is increasingly serving as a source of social community support, it is crucial for health care professionals to recognize its importance in facilitating access to social support for people affected by miscarriage [23]. Moreover, this is also the type of support that health care professionals currently do not provide, as it must be tailored to each individual's needs [5]. Recent studies have found that talking about miscarriage, listening, and providing informative and emotional support through different social networks can improve the quality of care [3,12,24,25]. Two additional reasons for this improvement challenge could be that it is not completely understood how individuals react to miscarriages [5,26] and that health care professionals do not feel educated enough to provide the necessary postmiscarriage support [12]. In this research, we consider social media as an important medium for social community support. On Instagram, there are numerous accounts with posts about the topic of miscarriage. We also recognize that the act of posting about miscarriage on social media is already a form of breaking the stigma surrounding miscarriage. That is why our research explores the social community support for the mental distress that many women go through after a miscarriage.

Given the recognized gap in miscarriage support from friends, family, and health care professionals and the evolving knowledge base on OHCs and the social support offered by community members, we pose the following research question: *How is social community support given to women who share their miscarriage experiences on social media?*

Methods

Qualitative Research Design

Social media, as a form of social community support, is a relatively new area of research. Therefore, we chose to use a qualitative research approach to explore this area and create an in-depth understanding [27]. Within this qualitative research approach, we applied a phenomenological strategy [28,29] to create a comprehensive understanding of the social community support phenomenon. For this purpose, we collected Instagram posts from a selected group of women sharing their miscarriage stories to analyze their perceptions and social support interactions on experiencing a miscarriage on a longitudinal basis. This data collection method enabled us to provide insights and a clear overview of social community engagement in the use of Instagram when sharing stories about miscarriage and how other users engage with the posts. The data, consisting of posts and corresponding comments, were collected within a 6-month time frame following the first post announcing the miscarriage. These posts were not limited to those explicitly addressing the topic of miscarriage; all other posts were also considered relevant for the research, as they contributed to forming a comprehensive overview of an entire miscarriage journey. A 6-month time frame was chosen to narrow down the research dataset, although additional posts about miscarriage were made after this time frame. During data collection, all aspects of engagement were gathered, including posts, comments, and likes. The data were later analyzed using thematic analysis to identify themes and relate each theme to its dynamic over time. We used a thematic analysis to unpack the multifaceted dimensions of social community support for miscarriage experiences.

Instagram as a Data Collection Platform

Due to their growing popularity, social media platforms such as Instagram have become valuable sources of data for researchers seeking to explore various aspects of social science [30]. This approach recognizes the reliability and value of the stories and experiences shared by individuals on social media and highlights the potential for using social media platforms as a source of data for qualitative research. Our research focuses on 2 aspects of using Instagram as a data collection platform: how users share their stories and how other users engage with these posts. Instagram is a popular social networking platform for photo and video sharing that offers a rich and varied multimodal dataset beyond just visual content. This diversity makes it an invaluable source of insights into user interactions, narratives, and community dynamics. Table 1 provides an overview of the data extracted from Instagram in this research and the relevant terminology used by Instagram [31].

XSL•FO
RenderX

Table 1. Multimodal data collected from Instagram.

Name	Definition
Post	A shared picture or other visual content. Either a single picture or a series of up to 10 visuals (referred to as a carrousel)
Post caption	The text below a picture. Has a limit of 2200 characters
Picture comments	Other Instagram users' comments' on the post
Video	A short video of a maximum of 90 seconds
Video caption	The text below a video. Has a limit of 2200 characters
Video comments	The comments on a video
Nested comments	The comments inside another comment
Number of likes on posts and videos	The total amount of likes from other Instagram users at the time of data extraction
Number of comments on posts and videos	The total amount of comments from other Instagram users at the time of data extraction
Number of resends on posts and videos	The total amount of times the post or video is forwarded from one Instagram user to another at the time of data extraction

Sample

The sampling strategy was focused on identifying women who have experienced a miscarriage and shared their stories on Instagram. To build a relevant sample, this research considers miscarriage following pregnancy from both traditional intercourse and in vitro fertilization from heterosexual relationships. The inclusion and exclusion criteria are provided in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Miscarriage after both in vitro fertilization and traditional intercourse
- Accounts with 10,000-100,000 Instagram followers
- Women who have experienced a miscarriage 6 months to 5 years prior
- English language
- Accounts from North American and European women
- Women aged >18 years

Exclusion criteria

- Nonhetero relationships
- Motivational accounts
- Less than 10 posts explicitly mentioning the miscarriage
- More than 75 overall posts
- Inability to give informed consent

The research identified those Instagram users who are so-called "micro-influencers" (10,000-100,000 followers) [32]. These influencers typically nurture more personal and engaging community interactions, essential for sensitive discussions about miscarriage. Their audience's trust and relatability toward them promise more authentic and deeper insights into the subject. In addition, their niche following provides a concentrated yet diverse range of experiences, ensuring a rich and focused dataset for nuanced qualitative analysis. Furthermore, microinfluencers have a tighter-knit social community than larger or smaller accounts [32], which is highly relevant to this research, as its focus is on the use of Instagram as a social community. To begin the sampling, the hashtags #ihadamiscarriage, #miscarriage, and #1in4 were searched on Instagram to identify a list of potential microinfluencers. Instagram profiles that primarily focused on posting motivational content, often featuring quotes, positive affirmations, and generalized encouragement messages, were excluded as they differ distinctively from personal story-driven profiles. Once a list of potential posters had been made, these Instagram users were contacted via direct message on Instagram and provided with an informed consent form with the topic and purpose of the study. Posters were only included in the study when they gave direct permission to use their data. The time frame for post inclusion was chosen to be between 6 months and 5 years ago. This period is long enough to accumulate a substantial body of data for analysis, and the upper limit of 5 years ensures that the collected data are still relevant and reflective of current social media trends and discussions

XSL•FO
RenderX

surrounding miscarriage. Profiles with content in English were selected to minimize translation biases and miscommunication. North American and European women were selected, as their social media use habits are similar [33]. Finally, Instagram profiles with at least 10 posts specifically mentioning the miscarriage were included to have enough robust data to analyze posting on social media about miscarriage. From a potential list of 9 microinfluencers who consented to the use of their data for this research, the number of posts the person made within the 6-month time frame was used as a further exclusion criterion; any influencers with >75 posts in 6 months were excluded. This exclusion led to a more tightly focused analysis that delved deeper into the sampled dataset. This ultimately resulted in a sample of 6 female Instagram posters who discussed their miscarriage experiences, with a total of 258 posts, 736 comments, and 1734 quotations, as shown in Table 2.

Table 2. Sample of collected data.

Poster	Followers[a], n×1000	Posts analyzed[b], n (%)			Comments analyzed[b] (n=736), n (%)	Quotations, (n=1724), n (%)
		Total (n=258)	Videos (n=88)	Pictures (n=170)		
1	19.3	52 (20)	27 (31)	55 (32)	149 (20)	283 (16)
2	65.5	41 (16)	28 (32)	13 (8)	119 (16)	271 (16)
3	27.3	29 (11)	4 (5)	25 (15)	74 (10)	258 (15)
4	22.3	57 (22)	9 (10)	50 (29)	171 (23)	331 (19)
5	22.2	24 (9)	20 (23)	4 (2)	62 (8)	174 (10)
6	12.0	55 (21)	0 (0)	55 (32)	161 (22)	407 (24)

[a]Date of extraction: June 18, 2023.

[b]Date of extraction: between May 15, 2023, and May 31, 2023.

Data Analysis

All data types shown in Table 1 were analyzed over a 6-month time frame to create a comprehensive and holistic overview of how the women share their stories online in interaction with other social media users. The top 3 comments under each post were analyzed, as these comments had the most likes from other users, indicating that these comments were the most popular among the people interacting with the post. In total, this was between 84 and 180 comments per poster. The number of data quotations analyzed was 1724 (Table 2). The first 3 authors were involved in the thematic analysis of the data sample [34]. The data were then structured in a code tree created collaboratively by these 3 researchers. The fourth author provided feedback and suggestions on the draft reporting of the code trees. Throughout this analysis process, discussions took place among all 4 researchers to refine and validate the themes and patterns, enhancing the validity of the findings through triangulation [34]. Throughout the coding process, the emergence of new themes and patterns was continuously tracked to determine if data saturation was achieved. Saturation was considered achieved when no new themes or patterns were identified in the final round of coding, which led to the creation of the final code tree. This gave an overview of the identified themes and the patterns of the presence of these themes during the time span following the first post about miscarriage.

Six Steps of Thematic Analysis

To analyze the phenomenon of social community support for women who share their miscarriage experiences on social media, a thematic analysis approach was applied to determine recurring themes and patterns within the data [28,29]. Both the posts and the comments on the posts were analyzed separately for themes. The six steps highlighted by Braun and Clarke [28] were used: (1) familiarizing with the data, (2) generating initial codes, (3) searching for themes as clustered codes, (4) reviewing the themes, (5) defining and naming the themes, and finally (6) producing the report. The first 2 steps were largely carried out during the data collection described in the Sample section.. This was followed by the initial analysis, which included steps 3 and 4, where the first round of codes was determined (72 codes). Clustered codes (25 codes) were made to include codes that were not used as frequently but which were still important for the research. Following this, steps 5 and 6 were completed to make categories for both the posters and the commenters and then define and name the final themes.

Ethical Considerations

This study has carefully taken ethical considerations into account, given the sensitive nature of the topic. The researchers have ensured that the study only uses stories and data from public Instagram accounts, shared by the individuals themselves. To ensure the privacy of the posters, the research team obtained consent for data use from all analyzed users who have shared their personal miscarriage stories through an Instagram post. The consent included informing the users about the use of their data and the purpose of the study and ensuring that their personal information is not disclosed in any way that could potentially identify them. The data of users engaging in posts (through comments and likes) were also treated with discretion and anonymized for confidentiality to ensure their privacy rights are respected. The project received ethics approval from the Human Research Ethics Committee at Delft University of Technology (approval number 2234), ensuring that the study is conducted in an ethical and responsible manner. The researchers approached this study with sensitivity and respect for the experiences of women who have experienced a

XSL•FO
RenderX

miscarriage, making a conscious effort to use language that is not stigmatizing or insensitive.

Results

Social Community Support for the Multifaceted Experience of Miscarriage

Resulting from the thematic analysis, three core themes emerged: (1) *storytelling of emotional turmoil and grief after miscarriage*, (2) *sharing positivity amidst miscarriage grief*, and (3) *mentioning personal medical information about miscarriage.* These themes collectively represent how diverse experiences and emotions related to miscarriage are shared within the open social media community of Instagram. Figure 1 presents the dual code tree on social community support for miscarriage experiences. The 3 core theme structure results from the data induction of both the women posting their miscarriage experience (18 code clusters into 7 categories) and the online community commenters (6 code clusters into 3 categories).

Figure 1. The dual code tree of social community support for miscarriage experiences.

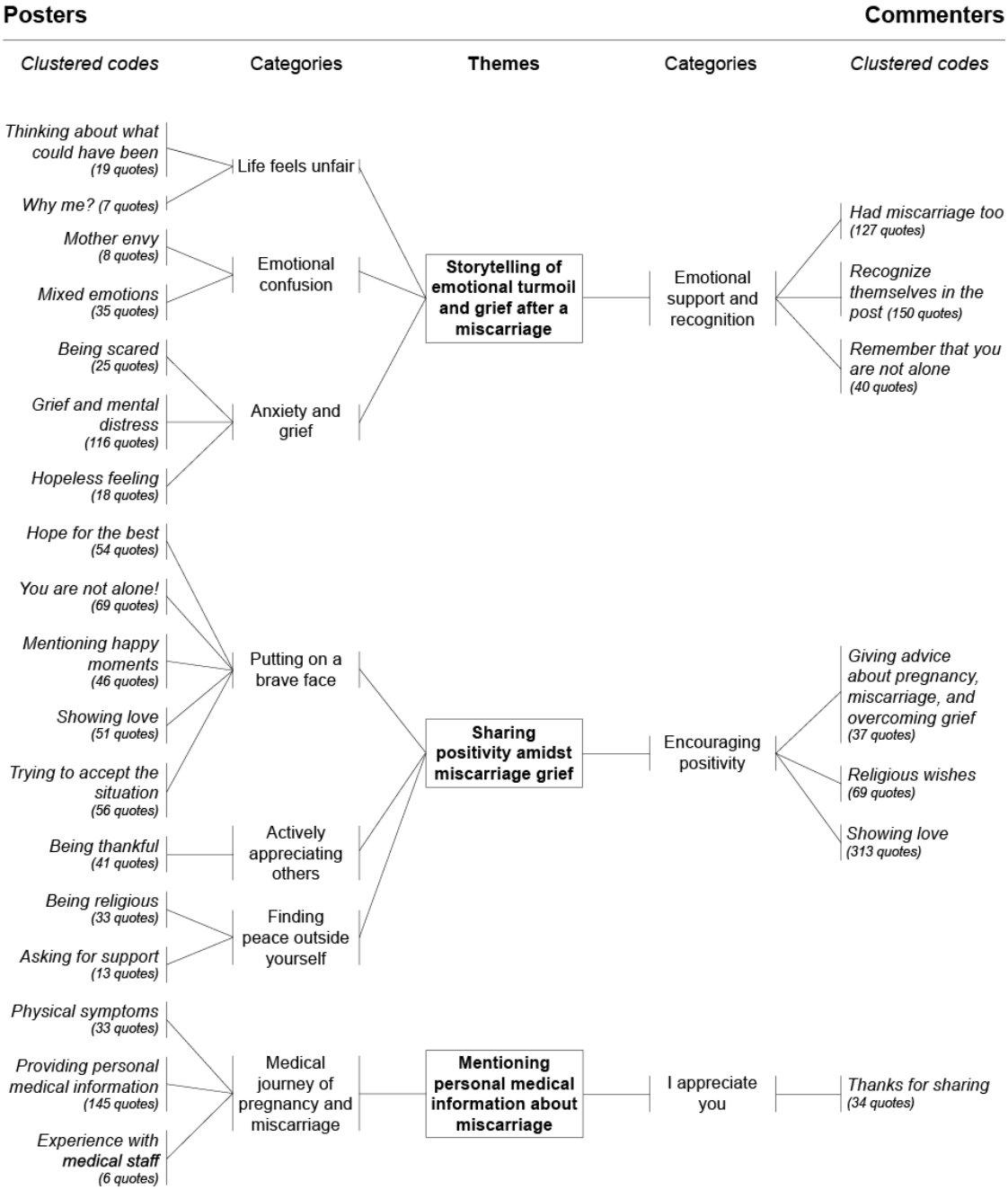

Theme 1, *storytelling of emotional turmoil and grief after miscarriage*, uncovers the deep emotional impact and distress faced by women who have suffered a miscarriage and how, in reaction to their storytelling, commenters provide social-emotional support and recognition.

Theme 2, *sharing positivity amidst miscarriage grief*, emphasizes the resilience and determination of women to find

XSL•FO
RenderX

moments of joy and hope despite the underlying grief and how commenters play a social support role in providing uplifting positive support to peers.

Theme 3, *mentioning personal medical information about miscarriage*, highlights the practical support aspects of experiencing a miscarriage, with the sharing of medical encounters and providing of medical information. The commenters express great appreciation for their openness in sharing such personal details.

While theme 1 is more focused on the internal emotional struggles, theme 2 demonstrates the desire to highlight positive aspects amidst miscarriage grief, and theme 3 takes a less emotional and more practical stance. Overall, these 3 themes illustrate the multifaceted nature of social community support surrounding grief and miscarriage on Instagram and demonstrate how social media is becoming its own support pillar when dealing with grief following miscarriage. The evidence on the core themes and categories is explained in more detail in the next paragraphs.

Storytelling of Emotional Turmoil and Grief After Miscarriage

The first identified theme is *storytelling of emotional turmoil and grief after miscarriage.* This theme represents the emotional experience of women who have had a miscarriage. It encompasses the initial posts of miscarriage storytelling that express deep grief and mental distress and the emotional impact on both the posters and the commenters (Figure 2). The posters express their grief by posting sincere pictures, heartfelt captions, and personal videos; sharing their story; or explaining the situation. The commenters often identify with these posts, and some of them open up about their own loss. Besides this expression of grief from both the posters and commenters, there is a strong element of encouragement that frequently accompanies these posts and comments. Three categories, in particular, were found to stand out among the posters: *life feels unfair*, *emotional confusion*, and *anxiety and grief*. For the commenters, the main category is *emotional support and recognition*.

In the first category of expressions, *life feels unfair*, the posters frequently mentioned how life seemed unfair and they questioned the reason behind their miscarriages (7 quotes), which also related to “feeling hopeless” (18 quotes). One community poster states the following:

> *So while I have reached a point of understanding and divine peace, we will all still grieve what could have been.* [Poster 5]

Figure 2. Evidence example of a posted image on theme 1.

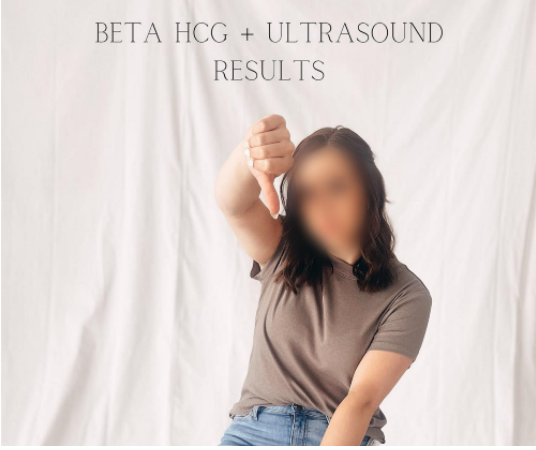

The second category, *emotional confusion* included several instances of mixed emotions (35 quotes). Some women felt envy toward other women who had successful pregnancies (8 quotes). One woman posted about having feelings that were difficult to navigate:

> *It’s been hard to see other pregnant people and be triggered with new announcements because I thought for sure our time would’ve come already.* [Poster 2]

The third category was about feeling *anxiety and grief*, pain, and mental distress. These emotions can be difficult to navigate, which can lead women to feel scared (25 quotes) and frequently, with women explicitly expressing grief in these posts (116 quotes):

> *I had never really felt despair in life before we lost you. There truly isn’t a day that goes by that I don’t think about you and wish you were here. and what makes it even harder is that our journey has brought even more heartbreak since then.* [Poster 2]

Emotional support and recognition was the category of community commenters who recognized themselves in the posted story and shared their recognition (150 quotes) or their own miscarriage story (127 quotes) and offered social-emotional support to the poster by expressing that they were not alone (40 quotes) with reactions such as the following:

> *Hugs mama, please know you are not alone. Thank you for your vulnerability.* [Comment on poster 5]

Sharing Positivity Amidst Miscarriage Grief

This theme focuses on follow-up posts that differ from the initial ones in that they showcase moments of happiness and positivity, despite the underlying grief. These posts shared with the online community convey a sense of moving forward and a refusal to let grief become the defining aspect of one’s life. Each

XSL•FO
RenderX

category—*putting on a brave face*, *actively appreciating others*, and *finding peace outside yourself*—served as a means of giving social community support back to those who might be dealing with a similar experience, thereby sharing the knowledge that even in the face of misery, there can be moments of joy. Within this theme, the main category for commenters is *encouraging positivity*, emphasizing the profound impact of cheerful and positive messages shared with each other, especially during the most challenging times (Figure 3). Commenter messages foster a sense of community support, expressing hope and encouragement for the poster.

The category *putting on a brave face* expresses trying to be upbeat even through the darkest of times. It encompasses various clustered codes such as being hopeful, *hope for the best* (54 quotes), spreading awareness about miscarriages, and making social community support explicit, coded as the expression *you're not alone!* (69 quotes) to everyone in the online community, such as in the following quote:

> *If 2022 didn't turn out to be your year, I just want to say that I see you. You're not alone.* [Poster 4]

In addition to that, *putting on a brave face* involves the *mentioning of happy moments* (46 quotes), *showing love* (51 quotes), and the sharing of *trying to accept the situation* (56 quotes) amidst grief. For an example, see the shared photo shown in Figure 3 and the following quote:

> *I know one day I'll look back at this time in my life and be proud of the woman I've become. I know it isn't all bad. But grief is an interesting thing. It comes and goes, and in many ways stays with you forever. I know it also makes me appreciate aspects of my life more now...It isn't all bad.* [Poster 4]

Figure 3. Evidence example of a comment related to theme 1.

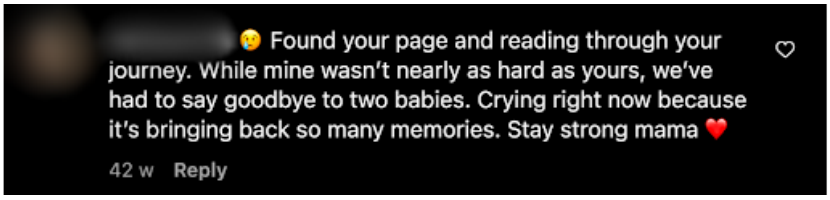

The next category of *being thankful* (41 quotes) concerned gratitude for the community support. Both commenters and posters showed gratitude for their partner in difficult times, and they also expressed their gratitude to the team of health care professionals, their loved ones, and community peers. For example, 1 woman wrote in the caption of a post asking for funding for their surrogate:

> *We are so grateful for the support we have gotten from you all and feel as though we can't fully express how much we love you and appreciate your help and the love you send our way. Someday we will have a happy ending and we appreciate your help to get there.* [Poster 1]

Then another category under this theme concerns *finding peace outside yourself*, which refers to seeking meaning and reaching out for social community support by *asking for support* (13 quotes) and embracing religious beliefs, *being religious* (33 quotes), to provide emotional strength:

> *So when you cry out, in desperation, like I've done countless times, and feel like He isn't opening the doors you want, or isn't moving you in a certain direction, know that BEING STILL is sometimes the hardest, yet most important thing you can do. It is in these moments that we truly develop FAITH. That we give over our will to His and TRUST HIM. It's when we progress beyond measure.* [Poster 5]

The main category for commenters was *encouraging positivity* by being supportive and *giving advice about pregnancy, miscarriage and overcoming grief* (37 quotes) to peers who struggled with the same feelings, with one commentator stating the following:

> *Glad y'all had a nice break together. We just took a break for a vacation as well! I can't recommend it enough.* [Comment on poster 4]

Some commenters expressed their support by offering *religious wishes* (69 quotes), thereby showing solidarity and providing comfort:

> *I feel your pain. I had to pour into something to get me through. I was in a sad, lonely dark place.... no one can understand... until they have lived it. I choose to listen to Christian contemporary music, which helped and I leaned in and listened to many online service through our church app. We had many obstacles knock us down, but kept moving forward with new goals to make our family. We finally had twins! Luck and strength to you.* [Comment on poster 1]

Other commenters expressed empathy by *showing love*, exemplified with heart emojis and quotes such as the following:

> *Words cannot even express how sorry I am for your loss. Your post resonated with me so much. This pain is the hardest. I love you my friend. Sending you and [name of husband] all my love and thoughts during this difficult time.* [Comment on poster 4]

Mentioning Personal Medical Information About Miscarriage

The third theme related to providing practical information and support to peers who have experienced similar situations (Figure 4). It included posts and comments about medications, in vitro fertilization procedures, and hospital experiences. *Physical symptoms* were also shared, and personal advice was given and shared.

XSL•FO
RenderX

Connected under this theme was the category *I appreciate you* from commenters showing gratitude for sharing this information.

Personal medical information providing concerns the category of *medical journey of pregnancy and miscarriage*, which often is not an easy journey. As many posters shared, this involved taking pills, managing diets, making regular visits to a clinic, and various other actions, coded as *sharing personal medical information* (145 quotes). For example, 1 poster wrote the following:

> *So today until trigger I'll be taking Omni trope—a human growth hormone—shot daily. 2.5mg to be exact, along with testosterone gel (1 pump) and DHEA 75mg (daily).* [Poster 4]

Figure 4. Evidence example of a posted reel related to theme 3.

Physical symptoms (33 quotes) cluster codes of posts about sharing physical symptoms and side effects of treatments and medical procedures. This also included posts on appreciating the capabilities of the body and addressing difficulties with trusting their body after the miscarriage:

> *A constellation of battle scars. I once had someone refer to their scars as such and it has stuck with me ever since. I've been saying this for years, I truly believe that polar opposite feelings can co-exist. When I look down at my stomach, I'm torn between feeling really sad and also being proud of all that it's been through.* [Poster 4]

Some also shared their personal encounters with health care professionals, encompassing both positive and negative experiences, coded as *experience with medical staff* (6 quotes):

> *And if I haven't said it enough I really love my clinic and my medical team. My nurse called me today and just made me feel so seen and taken care of. They truly want this for us as well and are fighting with us and I just appreciate them so much. Having my Dr be so personally involved rather than just seeing a tech every time I go in also makes me feel so taken care of. I'm just very grateful for the team we have helping us navigate all of this.* [Poster 4]

Most commenters expressed their appreciation for the poster's courage in openly sharing their story about the sensitive topic of miscarriages, making people who have gone through similar experiences feel less alone, coded as *thanks for sharing* (34 quotes):

> *Thank you for always being so open about it all! It helps feel less alone through it all.* [Comment on poster 4]

Discussion

Principal Findings

This research sought to better understand how women who have experienced miscarriage share their stories on social media and how commenters engage with these posts as a form of social community support. Its in-depth qualitative results offer, besides new insights into the longitudinal experiences of miscarriage grief and breaking down the stigma surrounding miscarriage through the use of social media, a novel contribution to the area of OHCs. In addition to the theoretical understanding of these domains, we identified three overarching themes that address the multifaceted social sharing of grief experiences after a miscarriage on social media: (1) storytelling of emotional turmoil and grief after miscarriage, (2) sharing positivity amidst miscarriage grief, and (3) mentioning personal medical information about miscarriage.

Social Media Community Support

All 3 themes demonstrate how breaking the social taboo by sharing miscarriage experiences on social media can generate social community support in a reciprocal way. Breaking the taboo and stigma surrounding miscarriage is important to create an environment where women feel safe and supported in their healing process. In contrast to the research that indicated that the topic of miscarriage is too sensitive and complex to talk about [3], we found that social media lowers this threshold. Each time a story is shared by uploading a post, a conversation about miscarriage is started. This raises awareness about an often misunderstood topic and helps normalize grief after a

XSL•FO
RenderX

miscarriage. Both posters and commenters are, perhaps unknowingly, challenging societal norms, and their actions contribute to fostering a better environment for those who have experienced this loss. The public sharing of miscarriage grief contributes to breaking down harmful myths and misconceptions about this topic. The 3 identified themes enhance an understanding of how social media can be used as a form of social support. This also relates to the theoretical implications for resolving an overall lack of knowledge on practical aspects of dealing with miscarriage grief [23]. It adds to previous research on social media that indicated that women often search for social forms of support on their own [23,30] but contradicts the research that suggests that the information shared is not valuable [35]. In particular, theme 3, *mentioning personal medical information about miscarriage*, indicates that the opposite is true: social media serves as a medium to educate others. Regarding this theme, the researchers consider it crucial to address the potential risks associated with medical information given by nonprofessionals. The information shared is based on personal experiences and may be inaccurate, as most of the posters and commenters lack the expertise to provide medical advice. This can lead to harmful consequences if others follow such advice without consulting health care professionals. While sharing personal medical information is perceived as helpful, as our findings show, relying solely on this information holds potential risks. Instagram users should seek professional medical help for health concerns.

Multifaceted Miscarriage Grief

The themes *storytelling of emotional turmoil and grief after miscarriage* and *sharing positivity amidst miscarriage grief* are interconnected, as both have to do with emotions related to miscarriage grief. The first largely points to the negative emotional experiences following the miscarriage, while the latter shows how women are trying to cope with their grief and learning to accept the pain by actively focusing on the good things in life.

In this study, it has become evident that women experience profound grief following a miscarriage. The vast majority of the analyzed posts show aspects of an intense grieving process. The main finding of this study, evident in both themes, is the importance of not becoming consumed by grief and sadness. Instead, it highlights the value of actively seeking social support and, alongside sharing emotional struggles, also expressing positivity and gratitude during times of grief. The posters mostly receive uplifting comments and advice, fostering a positive environment of social media support for both the poster and the commenter. This corresponds to a recent study that indicated that social media provides a form of nurturing support, which focuses on consolation more than problem-solving [26].

This supplements the understanding of mental distress caused by a miscarriage. It identifies the multifaceted nature of miscarriage grief and sheds light on an important aspect of social media community support, that is, empathy and consolation from emotional recognition.

Communal Load Sharing

All themes provide new insights into the reciprocal nature of social media community support. Both women who initiate personal storytelling on their miscarriage grief by posting on social media and their peers, the women who interact with these stories, are seeking consolation and empathy through sharing their emotional turmoil; medical information; and, unexpectedly, also their positivity in coping with their mental burden and load within the online community. This *communal load sharing* can be seen as a new form of social support, sought mostly through interactions on social media. Distinct from providing social support, either emotional or information support within OHCs, communal load sharing is a reciprocal sharing of grief experiences, positive moments, and social appreciation. Given the absence of a specific construct to describe this, we propose to use the term *communal load sharing* for this unique form of social support. Communal load sharing stands in contrast to prior studies that report on the indifference and lack of support from friends and family [1] and the consequences of self-isolation [11]. The novel finding in addition to existing research is the prominent outward manifestations of miscarriage experiences. Communal load sharing encapsulates this through social self-expression. The findings of the first theme, *storytelling of emotional turmoil and grief after miscarriage*, correspond with recent studies that indicated that women can have difficulty expressing their emotions following a miscarriage, although this is an integral step in the healing process [23,26]. Posting on social media about the miscarriage is a form of self-expression that can have positive effects on the healing process, and in this study, it was found to be a form of social self-expression in which women collectively cope with their miscarriage grief. This social self-expression and openness confront the often-understated depth of the experience.

In response, the social support from the commenters comes in the form of *emotional recognition support*, with commenters sharing their own stories and reminding the poster that they are not alone in their grief journey. This is in line with an earlier study in which women indicated that 1 of the most important factors that helped them following a miscarriage was hearing other women's stories [4]. In the second theme, *sharing positivity amidst miscarriage grief*, the posters shared positive moments in contrast to the traumatic event of a miscarriage. The commenters appreciated these forms of positivity and reinforced them by expressing their happiness for the poster, wishing them well, and providing religious blessings.

The 2 forms of communal load sharing, *social self-expression* and *emotional recognition support*, also add to the theoretical understanding of OHCs [20,21]. This builds upon corresponding research in which social media has been found to not have the informational reputability and reliability that OHCs formed by medical foundations or companies have [22], although it serves as an increasingly frequently used source of online community support [18] and, as our research indicates, an important form of social community support.

Limitations and Future Research

The researchers acknowledge the potential biases that may arise from using only publicly available data and recognize that it

XSL•FO
RenderX

may not represent the full range of experiences with miscarriage and the perspectives of women who are not active on social media. As is inherent to every qualitative study, we acknowledge the specific limitations of this type of research owing to potential researcher bias when interpreting and analyzing data. This research was purposefully focused on creating a better understanding from a woman's perspective; their partner also plays a large role in the grieving process following a miscarriage, and it would be worthwhile to conduct a study on the partner's point of view in the future.

The thematic analysis was based on a large longitudinal sample of >250 posts and approximately 730 comments; however, the sample was small, comprising 6 Instagram profiles. This may limit the diversity of experiences and perspectives in the context of social support for miscarriages. It is worth noting that the clustered commenter code *showing love* had a significantly high presence. It would be an interesting subject for future research to look into the different ways of showing love. Furthermore, in addition to analyzing the highest engagement comments, additional research could study all types of comments, including low engagement posts; the prioritization of comments; and, if possible, any removed comments. While this research focuses solely on studying posts and comments, future research could extend this analysis to include comments on comments. Moreover, interviews with posters could also lead to a deeper understanding of the topic.

This study sampled data from microinfluencers on Instagram. Microinfluencers often have a more engaged audience in comparison to everyday users or mega-influencers. This engagement can result in more impact of the posts and responses. In addition, the characteristics of microinfluencers may also affect the type of interaction they attract. For further research, it would be interesting to explore the interactions of everyday Instagram users or mega-influencers to research how social community support is given on these accounts. This could provide a broader perspective on the role social media plays in communal load sharing.

Practical Implications

The practical implications of these findings suggest that social media not only helps break the silence around the sensitive topic of miscarriage but also emphasizes the importance of cultivating a supportive online community, which can foster healing and understanding. Our research suggests that social media can be useful for providing social community support for women experiencing miscarriage and how they deal with grief. The main findings from this paper offer practitioners insights into how women navigate and cope with multifaceted miscarriage grief. All 3 themes demonstrate how communal load sharing is integral to social support for dealing with grief following a miscarriage. This encapsulates the interactions between posters and commenters through social self-expressions and emotional recognition support. Understanding women's experiences of miscarriage enables the social support network to provide further targeted assistance. As mentioned previously, health care professionals often do not feel adequately trained to deal with miscarriages [4]; thus, this in-depth analysis of social media support communities can help medical professionals better understand their own role in supporting the patient.

The researchers hope that this paper can serve as inspiration to continue breaking the taboo around the topic of miscarriage and contribute to paving the way for a more empathetic and inclusive society, where the emotional well-being of those affected by miscarriage is acknowledged, valued, and supported.

Conclusions

This study identified three themes that address the social sharing of grief experiences after a miscarriage on social media: (1) storytelling of emotional turmoil and grief after miscarriage, (2) sharing positivity amidst miscarriage grief, and (3) mentioning personal medical information about miscarriage.

As the overarching theme for this social support phenomenon, this study describes *communal load sharing* as a therapeutic role of social media in helping women cope with miscarriage by providing a platform for sharing similar experiences, breaking social taboos, and fostering load sharing.

Acknowledgments

The authors would like to thank the women who posted their stories on Instagram and consented to the use of their data in the study and the reviewers for their constructive feedback.

Authors' Contributions

JD, JO, and LH contributed to this paper equally, in the conceptualization and study design, as well as contributed to the write-up and editing of the manuscript. LS contributed to this paper by providing feedback on the literature review, study design, and results; editing; and reviewing. JD, JO, and LH took on the roles of conceptualization, methodology, data curation (analysis on Atlas.ti), formal analysis (thematic analysis), investigation (code trees and theme identification), project administration, validation (process discussions), visualization, and writing the first draft. JD and JO were also involved in writing, reviewing, and editing the manuscript for publication. LS provided feedback and suggestions on all draft reporting of each section and the code trees, alongside further supervision and editing tasks.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Quenby S, Gallos ID, Dhillon-Smith RK, Podesek M, Stephenson MD, Fisher J, et al. Miscarriage matters: the epidemiological, physical, psychological, and economic costs of early pregnancy loss. Lancet 2021 May 01;397(10285):1658-1667. [doi: 10.1016/S0140-6736(21)00682-6] [Medline: 33915094]
2. Regan L, Rai R. Epidemiology and the medical causes of miscarriage. Baillieres Best Pract Res Clin Obstet Gynaecol 2000 Oct;14(5):839-854. [doi: 10.1053/beog.2000.0123] [Medline: 11023804]
3. The Lancet. Miscarriage: worldwide reform of care is needed. Lancet 2021 May 01;397(10285):1597. [doi: 10.1016/S0140-6736(21)00954-5] [Medline: 33915093]
4. Séjourné N, Callahan S, Chabrol H. Support following miscarriage: what women want. J Reprod Infant Psychol 2010 Nov;28(4):403-411. [doi: 10.1080/02646830903487375]
5. van den Akker OB. The psychological and social consequences of miscarriage. Expert Rev Obstet Gynecol 2014 Jan 10;6(3):295-304. [doi: 10.1586/eog.11.14]
6. Ying LY, Wu LH, Loke AY. Gender differences in experiences with and adjustments to infertility: a literature review. Int J Nurs Stud 2015 Oct;52(10):1640-1652. [doi: 10.1016/j.ijnurstu.2015.05.004] [Medline: 26021885]
7. Simonse L, Albayrak A, Starre S. Patient journey method for integrated service design. Des Health 2019 May 13;3(1):82-97. [doi: 10.1080/24735132.2019.1582741]
8. Dixon SJ. Most used social networks 2024, by number of users. Statista. 2024 Jul 10. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [accessed 2024-11-02]
9. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
10. Athey J, Spielvogel AM. Risk factors and interventions for psychological sequelae in women after miscarriage. Prim Care Update Ob Gyns 2000 Mar 01;7(2):64-69. [doi: 10.1016/s1068-607x(00)00023-8] [Medline: 10725674]
11. Kersting A, Wagner B. Complicated grief after perinatal loss. Dialogues Clin Neurosci 2022 Apr 01;14(2):187-194. [doi: 10.31887/dcns.2012.14.2/akersting]
12. Rellstab S, Bakx P, Garcia-Gomez. Human capital formation: the effect of a miscarriage on mental health, labour market, and family outcomes. Tinbergen Institute. 2022. URL: https://ideas.repec.org/p/tin/wpaper/20220027.html [accessed 2023-04-03]
13. Bellhouse C, Temple-Smith MJ, Bilardi JE. "It's just one of those things people don't seem to talk about..." women's experiences of social support following miscarriage: a qualitative study. BMC Womens Health 2018 Oct 29;18(1):176 [FREE Full text] [doi: 10.1186/s12905-018-0672-3] [Medline: 30373583]
14. Beutel M, Willner H, Deckardt R, Von Rad M, Weiner H. Similarities and differences in couples' grief reactions following a miscarriage: results from a longitudinal study. J Psychosom Res 1996 Mar;40(3):245-253. [doi: 10.1016/0022-3999(95)00520-x] [Medline: 8861120]
15. McLean A, Flynn C. ‘It’s not just a pap-smear’: women speak of their experiences of hospital treatment after miscarriage. Qual Soc Work 2012 Nov 01;12(6):782-798. [doi: 10.1177/1473325012464099]
16. Bute JJ, Brann M, Hernandez R. Exploring societal-level privacy rules for talking about miscarriage. J Soc Pers Relatsh 2017 Sep 26;36(2):379-399. [doi: 10.1177/0265407517731828]
17. Correll SJ, Benard S, Paik I. Getting a job: is there a motherhood penalty? Am J Sociol 2007 Mar;112(5) [FREE Full text] [doi: 10.1086/511799]
18. Prescott J, Mackie L. "You sort of go down a rabbit hole...You're just going to keep on searching": a qualitative study of searching online for pregnancy-related information during pregnancy. J Med Internet Res 2017 Jun 05;19(6):e194 [FREE Full text] [doi: 10.2196/jmir.6302] [Medline: 28583906]
19. Ure C, Cooper-Ryan AM, Condie J, Galpin A. Exploring strategies for using social media to self-manage health care when living with and beyond breast cancer: in-depth qualitative study. J Med Internet Res 2020 May 25;22(5):e16902 [FREE Full text] [doi: 10.2196/16902] [Medline: 32364510]
20. Sanger S, Duffin S, Gough RE, Bath PA. Use of online health forums by people living with breast cancer during the COVID-19 pandemic: thematic analysis. JMIR Cancer 2023 Feb 07;9:e42783 [FREE Full text] [doi: 10.2196/42783] [Medline: 36473015]
21. Wang X, Zhao K, Street N. Analyzing and predicting user participations in online health communities: a social support perspective. J Med Internet Res 2017 Apr 24;19(4):e130 [FREE Full text] [doi: 10.2196/jmir.6834] [Medline: 28438725]
22. Andalibi N, Bowen K. Internet-based information behavior after pregnancy loss: interview study. JMIR Form Res 2022 Mar 02;6(3):e32640 [FREE Full text] [doi: 10.2196/32640] [Medline: 35234656]
23. Mercier RJ, Senter K, Webster R, Henderson Riley A. Instagram users' experiences of miscarriage. Obstet Gynecol 2020 Jan;135(1):166-173. [doi: 10.1097/AOG.0000000000003621] [Medline: 31809440]
24. Jia L, Li W, Liu Y, Wang L. Psychologic sequelae in early pregnancy complications. Int J Womens Health 2023;15:51-57 [FREE Full text] [doi: 10.2147/IJWH.S382677] [Medline: 36643713]
25. Verschoor MA. Management of first trimester miscarriage; new insight in old dilemmas. University of Amsterdam repository PhD thesis. 2017 Mar 28. URL: https://dare.uva.nl/search?identifier=dcc36bfc-fb5f-40ba-a1a5-84ab242474bd [accessed 2023-05-11]

XSL•FO
RenderX

26. Alqassim MY, Kresnye KC, Siek KA, Lee J, Wolters MK. The miscarriage circle of care: towards leveraging online spaces for social support. BMC Womens Health 2022 Jan 29;22(1):23 [FREE Full text] [doi: 10.1186/s12905-022-01597-1] [Medline: 35090452]
27. Creswell JW. The selection of a research approach. In: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications, Inc; 2014.
28. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 02;16(1). [doi: 10.1177/1609406917733847]
30. Felt M. Social media and the social sciences: how researchers employ Big Data analytics. Big Data Soc 2016 Apr 29;3(1). [doi: 10.1177/2053951716645828]
31. Instagram features. Instagram. URL: https://about.instagram.com/features [accessed 2023-05-11]
32. Park J, Lee JM, Xiong VY, Septianto F, Seo Y. David and Goliath: when and why micro-influencers are more persuasive than mega-influencers. J Advert 2021 Oct 26;50(5):584-602. [doi: 10.1080/00913367.2021.1980470]
33. Leading countries based on Instagram audience size as of January 2024. Statista. URL: https://www.statista.com/statistics/578364/countries-with-most-instagram-users/ [accessed 2023-05-11]
34. Ravitch SM, Carl NM. Validity in qualitative research. In: Qualitative Research: Bridging the Conceptual, Theoretical, and Methodological. Thousand Oaks, CA: SAGE Publications; 2015.
35. Chen Y, Sherren K, Smit M, Lee KY. Using social media images as data in social science research. New Media Soc 2021 Aug 18;25(4):849-871. [doi: 10.1177/14614448211038761]

Abbreviations

OHC: online health community

Edited by T de Azevedo Cardoso; submitted 23.01.24; peer-reviewed by S Rajput, A Hassan, C Severinsen; comments to author 13.06.24; revised version received 27.06.24; accepted 27.06.24; published 10.12.24.

Please cite as:
Dubbelman J, Ooms J, Havgry L, Simonse L
Communal Load Sharing of Miscarriage Experiences: Thematic Analysis of Social Media Community Support
J Med Internet Res 2024;26:e56680
URL: https://www.jmir.org/2024/1/e56680
doi:10.2196/56680
PMID:39656509

©Julia Dubbelman, Jonelle Ooms, Laura Havgry, Lianne Simonse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Robust and Interpretable Early Triaging Support System for Patients Hospitalized With COVID-19: Predictive Algorithm Modeling and Interpretation Study

Sangwon Baek[1,2], BA; Yeon joo Jeong[3], MD; Yun-Hyeon Kim[4], MD; Jin Young Kim[5], MD; Jin Hwan Kim[6], MD; Eun Young Kim[7], MD; Jae-Kwang Lim[8], MD; Jungok Kim[9], MD; Zero Kim[1,10], PhD; Kyunga Kim[10,11,12*], PhD; Myung Jin Chung[1,10,12,13*], MD, PhD

[1]Medical AI Research Center, Samsung Medical Center, Seoul, Republic of Korea
[2]Center for Data Science, New York University, New York, NY, United States
[3]Department of Radiology, Research Institute for Convergence of Biomedical Science and Technology, Pusan National University Yangsan Hospital, Yangsan, Republic of Korea
[4]Department of Radiology, Chonnam National University Hospital, Gwangju, Republic of Korea
[5]Department of Radiology, Keimyung University Dongsan Hospital, Daegu, Republic of Korea
[6]Department of Radiology, Chungnam National University Hospital, Daejeon, Republic of Korea
[7]Department of Radiology, Gachon University Gil Medical Center, Incheon, Republic of Korea
[8]Department of Radiology, School of Medicine, Kyungpook National University, Daegu, Republic of Korea
[9]Department of Infectious Diseases, Chungnam National University Sejong Hospital, Sejong, Republic of Korea
[10]Department of Data Convergence & Future Medicine, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[11]Biomedical Statistics Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea
[12]Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea
[13]Department of Radiology, Samsung Medical Center, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Kyunga Kim, PhD
Biomedical Statistics Center, Research Institute for Future Medicine
Samsung Medical Center
81 Irwon-ro, Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 2 3410 6745
Fax: 82 2 3410 2525
Email: kyunga.j.kim@gmail.com

Abstract

Background: Robust and accurate prediction of severity for patients with COVID-19 is crucial for patient triaging decisions. Many proposed models were prone to either high bias risk or low-to-moderate discrimination. Some also suffered from a lack of clinical interpretability and were developed based on early pandemic period data. Hence, there has been a compelling need for advancements in prediction models for better clinical applicability.

Objective: The primary objective of this study was to develop and validate a machine learning–based Robust and Interpretable Early Triaging Support (RIETS) system that predicts severity progression (involving any of the following events: intensive care unit admission, in-hospital death, mechanical ventilation required, or extracorporeal membrane oxygenation required) within 15 days upon hospitalization based on routinely available clinical and laboratory biomarkers.

Methods: We included data from 5945 hospitalized patients with COVID-19 from 19 hospitals in South Korea collected between January 2020 and August 2022. For model development and external validation, the whole data set was partitioned into 2 independent cohorts by stratified random cluster sampling according to hospital type (general and tertiary care) and geographical location (metropolitan and nonmetropolitan). Machine learning models were trained and internally validated through a cross-validation technique on the development cohort. They were externally validated using a bootstrapped sampling technique

XSL•FO
RenderX

on the external validation cohort. The best-performing model was selected primarily based on the area under the receiver operating characteristic curve (AUROC), and its robustness was evaluated using bias risk assessment. For model interpretability, we used Shapley and patient clustering methods.

Results: Our final model, RIETS, was developed based on a deep neural network of 11 clinical and laboratory biomarkers that are readily available within the first day of hospitalization. The features predictive of severity included lactate dehydrogenase, age, absolute lymphocyte count, dyspnea, respiratory rate, diabetes mellitus, c-reactive protein, absolute neutrophil count, platelet count, white blood cell count, and saturation of peripheral oxygen. RIETS demonstrated excellent discrimination (AUROC=0.937; 95% CI 0.935-0.938) with high calibration (integrated calibration index=0.041), satisfied all the criteria of low bias risk in a risk assessment tool, and provided detailed interpretations of model parameters and patient clusters. In addition, RIETS showed potential for transportability across variant periods with its sustainable prediction on Omicron cases (AUROC=0.903, 95% CI 0.897-0.910).

Conclusions: RIETS was developed and validated to assist early triaging by promptly predicting the severity of hospitalized patients with COVID-19. Its high performance with low bias risk ensures considerably reliable prediction. The use of a nationwide multicenter cohort in the model development and validation implicates generalizability. The use of routinely collected features may enable wide adaptability. Interpretations of model parameters and patients can promote clinical applicability. Together, we anticipate that RIETS will facilitate the patient triaging workflow and efficient resource allocation when incorporated into a routine clinical practice.

(J Med Internet Res 2024;26:e52134) doi:10.2196/52134

KEYWORDS

COVID-19; prognosis; prognostic; prognostics; prediction model; early triaging; interpretability; machine learning; predict; prediction; predictive; triage; triaging; emergency; severity; biomarker; biomarkers; SHAP; Shapley; clustering; hospital admission; hospital admissions; hospitalize; hospitalization; hospitalizations; neural network; neural networks; deep learning; Omicron; SARS-CoV-2; coronavirus

Introduction

During the COVID-19 pandemic, the global health care system confronted an urgent threat despite concerted efforts from health care institutions and providers to contain the rapid spread of the disease, which has claimed the lives of 6.97 million people as of October 2023 [1]. The overwhelming influx of patients into hospitals strained medical resources and hindered optimal treatment provision by health care practitioners [2]. This global outbreak may continue to exist with the advent of new SARS-CoV-2 variants due to its tendency to mutate during host adaptation [3]. Recently, a high proportion of population immunity and decreasing fatality rates initiated global movements toward endemic status following the World Health Organization's (WHO) announcement that COVID-19 is no longer a public health emergency of international concern [4,5]. However, COVID-19 cases continue to rise with the emergence of new subvariants, such as SARS-CoV-2 EG.5 and BA.2.86 [6,7]. Therefore, a robust and interpretable early triaging system is necessary to accurately triage patients in preparation for the next pandemic [8].

Many prognostic models for patients with COVID-19 severity and mortality have been proposed, yet most were reported unsuitable for clinical application by several systematic review studies [9-11]. Most models were either at a high or unclear risk of bias (Wynants et al [10]: 305 out of 310 studies, 98.4%; Buttia et al [11]: 312 out of 314 studies, 99.4%) such that their reported discriminative performances were deemed neither reliable nor generalizable [10,11]. These high-risk models were developed with predictors selected based on univariable analysis, failed to deal with model overfitting represented by miscalibration, performed no or limited external validations with sufficient samples, imputed missing data without a clear explanation, or considered a limited number of machine learning (ML) algorithms [10,11]. Although there were some models with a low risk of bias, these models had low to moderate discriminative power, were based on the data from the early pandemic period, and had limited clinical interpretability [9,10]. Therefore, the development of a robust, interpretable, and generalizable model with high discriminative power is required to provide practical benefit in managing the next possible pandemic [12,13].

We aimed to develop and validate an ML-based Robust and Interpretable Early Triaging Support (RIETS) system to predict severity based on routinely collected biomarkers using a nationwide multicenter cohort. In addition, we tried to improve model interpretability through patient clustering and characterization.

Methods

Ethical Considerations

The study protocol was approved and the requirement for informed consent was waived by the institutional review boards (IRBs) of all participating hospitals. In addition, the use and management of a cloud-based data storage platform for the secondary analysis was approved by the IRB of Samsung Medical Center (SMC 2020-09-100-002). All unique identifiers were removed prior to uploading. All data in the study database were assigned a research specific serial number and deidentified to protect the confidentiality of the study patients.

XSL•FO
RenderX

Study Setting and Design

This study was a nationwide, multicenter, retrospective, prognostic study conducted in South Korea. We collected data for adult patients who were confirmed to have COVID-19 via real-time polymerase chain reaction and were hospitalized at 19 main referral hospitals between January 5, 2020, and August 29, 2022 (Methods S1 in Multimedia Appendix 1). Among 9199 hospitalized patients with COVID-19, we excluded 406 patients diagnosed either more than 15 days before or more than 1 day after the hospitalization date, as well as 2848 patients who had missing data in any variable of interest (Figure 1). A total of 5945 patients (5106 nonsevere and 839 severe) remained for the analysis. The 19 hospitals were divided into 4 strata according to hospital type and location: metropolitan area general hospitals, nonmetropolitan area general hospitals, metropolitan area tertiary care hospitals, and nonmetropolitan area tertiary care hospitals. We then used a random cluster sampling method to partition hospitals in each stratum and construct development and validation cohorts. For reporting and bias-risk assessment, we adhered to the following guidelines: Guidelines for Developing Machine Learning Predictive Models in Biomedical Research [14], Transparent Reporting of a multivariable prediction model for Individual Prognosis (TRIPOD; File S1 in Multimedia Appendix 1), and Diagnosis and Prediction Model Risk of Bias Assessment Tool (PROBAST; File S2 in Multimedia Appendix 1) [15].

Figure 1. Patient flowchart depicting the generation of development and validation cohorts among hospitalized patients with COVID-19 (n=5945). Stratified random cluster sampling was applied to segment the cohorts based on hospital type (general vs tertiary) and geographical location (metropolitan vs nonmetropolitan). RT-PCR: real time polymerase chain reaction.

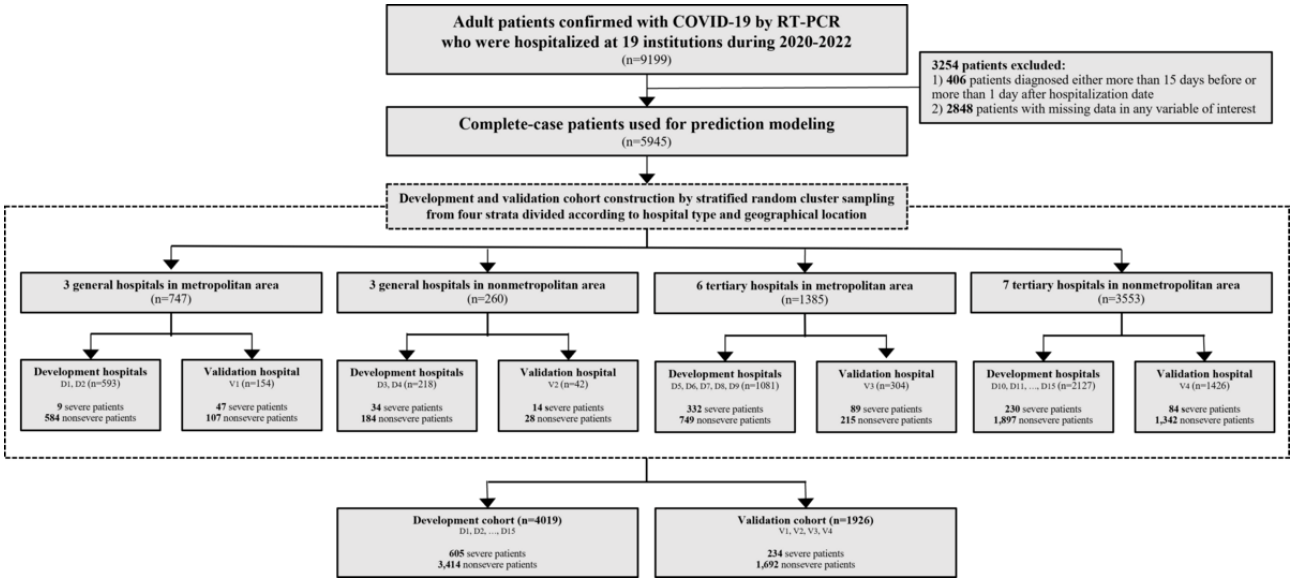

Data Collection

A set of data collection guidelines were predetermined by our clinical experts. We developed a standard data collection form and prepared cloud database storage. Adhering to the set guidelines, researchers affiliated with each participating hospital gathered patient data with 32 features from demographic, clinical, laboratory, and radiological findings within the first day of hospitalization. We specified these features based on previous prognostic models and a literature review describing common biomarkers associated with severe COVID-19 [16]. The final severity status of each patient was determined on day 15 of hospitalization. All data collected in each hospital were deidentified and uploaded onto the cloud database storage. The entire data set underwent a quality assurance process, including typo rectification, outlier handling, and double-checking with the electronic health records in each participating hospital.

Definition of COVID-19 Severity

We declared the COVID-19 severity for patients under one or more of the following conditions during their hospitalization: (1) mechanical ventilation required; (2) extracorporeal membrane oxygenation required; (3) admission to intensive care unit; or (4) patient's death. This criterion aligns closely with severe status (score of 6 or higher) in the WHO Clinical Progression Scale, which is developed by reaching a consensus among a group of international medical experts [17].

Identification of Candidate Feature Subsets

Among the 32 collected features, 27 readily accessible features without missing data remained for prediction modeling (Figure 2A). In order to identify subsets of robust features against feature selection methods, we considered 6 feature engineering methods (FEMs) based on 2 ML algorithms with optimal hyperparameter tuning and 4 feature importance measures: random forest (RF)-based mean decreases in Gini impurity feature importance; RF-based permutation feature importance; RF-based Shapley values; extreme gradient boosting (XGB)-based built-in feature importance; XGB-based permutation feature importance; and XGB-based Shapley values.

XSL•FO
RenderX

Figure 2. Machine learning–based pipeline for developing and validating the prognosis prediction model for COVID-19 severity. AUROC: area under receiver operating characteristic curve; DCA: decision curve analysis; DDRTree: Discriminative dimensionality reduction by learning a tree; DNN: deep neural network; GBM: gradient boosting machine; MLR: multivariable logistic regression; RF: random forest; RF-MDIFI: random forest–based mean decrease in Gini Impurity feature importance method; RF-PFI: random forest–based permutation feature importance method; RF-Shapley: random forest–based Shapley method; ROC: receiver operating characteristic curve; SHAP: Shapley additive explanations; SVM: support vector machine; XGB: extreme gradient boosting; XGB-BFI: extreme gradient boosting–based built-in feature importance method; XGB-PFI: extreme gradient boosting–based permutation feature importance method; XGB-Shapley: extreme gradient boosting–based Shapley method.

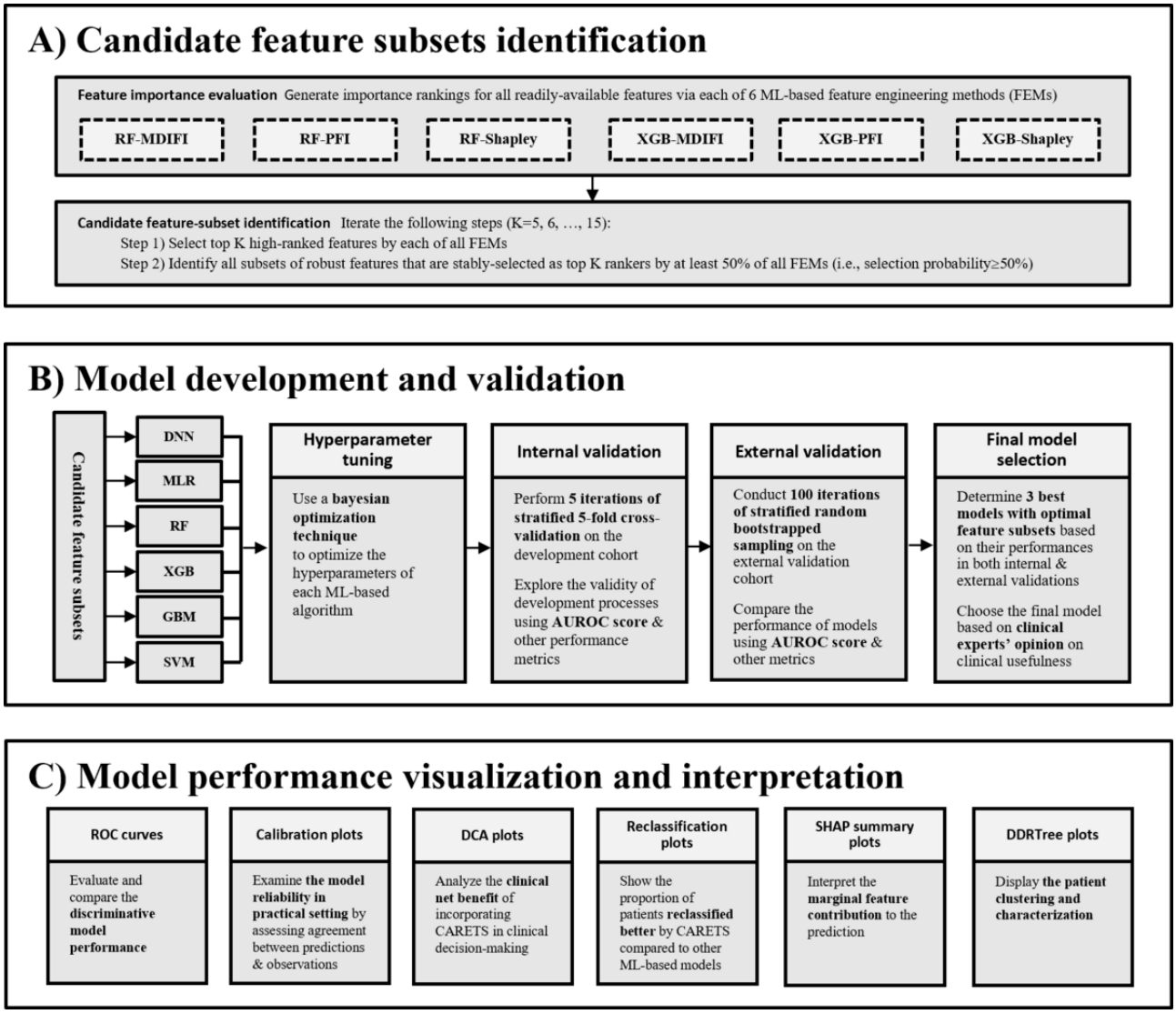

All features were ranked by importance measure by each FEM. We considered various criteria (ie, top K; K=5, 6, …, 15) for high-ranking features, termed high rankers. Based on each criterion K, we generated candidate feature subsets in two steps: (1) we selected high rankers with K highest importance rankings by each FEM, and (2) we identified features that were stably selected as high rankers by at least 50% of all FEMs. This process resulted in candidate subsets of robust features. The set of all 27 features was used as the reference model to identify the performance improvements in subset models during model evaluation.

Model Development and Validation

A total of 60 candidate feature subsets were used for the model development, including 59 identified subsets of robust features and the reference set of all 27 features (Figure S1 in Multimedia Appendix 1). We first fine-tuned the hyperparameters of 6 ML-based algorithms, namely, deep neural network (DNN), multivariable logistic regression, RF, XGB, gradient boosting machine, and support vector machine, by applying Bayesian optimization on the development cohort. Then, we simultaneously developed all possible 360 combinations of 6 ML-based algorithms and those 60 feature subsets and evaluated the performances in both internal and external validations (Figure 2B). The model predictive performance was evaluated using the area under the receiver operating characteristic curve (AUROC) score and other cutoff-based measures, such as sensitivity, specificity, positive and negative predicted values, positive and negative likelihood ratios, and diagnostic odds ratio.

Model development and internal validation were done with the development cohort in the following steps. First, we used 5 iterations of stratified 5-fold cross-validation to explore the internal validity of each combination of feature subsets and ML algorithms as a model development procedure. The procedures

XSL•FO
RenderX

were evaluated by the mean values of performance metrics and their 95% CIs calculated from the repeated cross-validation process (Methods S2 in Multimedia Appendix 1). Second, we used the entire development cohort to construct prediction models based on each of the internally validated development procedures.

External validation was conducted for the prediction models with the validation cohort. Each prediction model was evaluated by the mean values and 95% CIs of performance measures that were calculated from 100 iterations of bootstrapped sampling (Methods S2 in Multimedia Appendix 1). The final prediction model was proposed in three steps: (1) for each ML algorithm, we selected the optimal feature subset that produced the model with best discriminative performance in both internal and external validations; (2) the 3 prediction models with the best predictivity were chosen to compare their usefulness via calibration, reclassification improvement, and decision curve analysis (DCA); (3) the DNN-based final prediction model, RIETS, was proposed by considering its discriminative ability along with clinical applicability [18].

Model Performance Visualization and Feature Interpretation

We used graphical representations to visualize the performance of RIETS contrasted with other ML-based models and provided interpretation for the selected features (Figure 2C; Methods S3 in Multimedia Appendix 1). Receiver operating characteristic curves demonstrated the discriminative model performance. Calibration plots implicated the model's reliability in practical settings by displaying the correlation between predicted and observed risks. DCA plots indicated the net benefit of incorporating the model in clinical decision-making by quantifying the weighted trade-off between true positive and false positive identifications [19]. Reclassification plots displayed the proportion of patients that were reclassified correctly or incorrectly by RIETS compared to other ML-based models. Lastly, the Shapley additive explanations summary plot interpreted the contributions of individual features in RIETS when classifying severe and nonsevere cases [20].

Patient Clustering and Characterization Using Discriminative Dimensionality Reduction

We used discriminative dimensionality reduction via learning a tree (DDRTree) to cluster and characterize patients based on the features in RIETS. DDRTree is a tree-based unsupervised learning technique that reduces multidimensional features into a 2-dimensional space to visualize patients in the form of a tree structure (see Methods S3 in Multimedia Appendix 1 for procedures). This tool is known to capture cluster information with higher accuracy compared to conventional dimensionality reduction methods [21,22].

Each patient in a tree was colored with dark red to indicate high odds for severity and light green to indicate low odds for severity (Figure 3). Then, in Figure 4, dark blue and light green colors were overlaid to represent high and low concentrations of each laboratory marker, respectively. Severity risk of each patient can be identified through Figure 3A, the risk distribution with or without a pre-existing condition can be seen in Figures 3B and 3C, and feature values associated with each patient can be inferred from Figure 4. We integrated these observations to cluster patients into subgroups and characterize each subgroup (subgroup boundaries are shown in Figure 3A).

Figure 3. Patient clustering based on features in RIETS and characterization with dyspnea, DM, age, and severity. (A) DDRTree plot for severity probability. (B) DDRTree plot for patients with dyspnea, DM, and age ≥60 years. (C) DDRTree plot for patients with dyspnea, DM, and age <60 years. Points closer to dark red indicate a high severity probability, while points closer to light green indicate a low severity probability. DDRTree: discriminative dimensionality reduction via learning a tree; DM: diabetes mellitus; LLG: lower left group; LRG: lower right group; MRG: middle right group; RIETS: robust and interpretable early triaging support system; URG: upper right group.

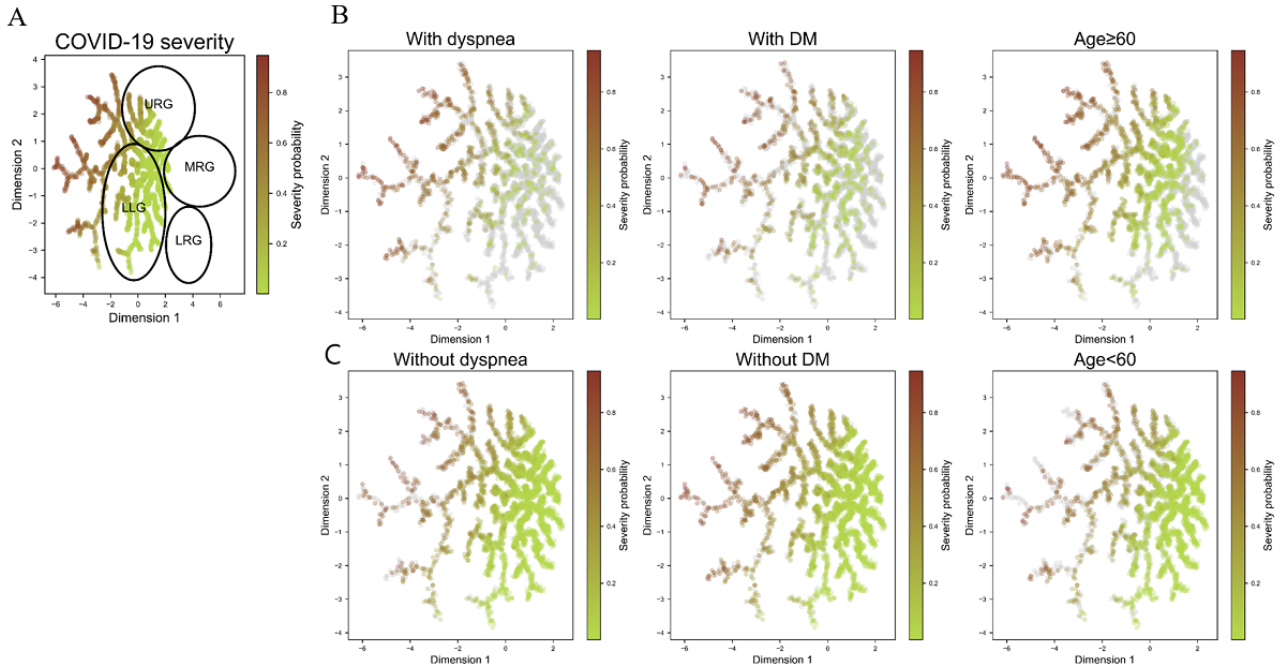

XSL•FO
RenderX

Figure 4. Patient clustering based on features in RIETS and characterization with vital signs and laboratory results. A dark blue color indicates a high concentration and a light green color indicates a low concentration of each corresponding feature. ALC: absolute lymphocyte count; ANC: absolute neutrophil count; CRP: c-reactive protein; DDRTree: discriminative dimensionality reduction via learning a tree; LDH: lactate dehydrogenase; PLT: platelet count; RIETS: Robust and Interpretable Early Triaging Support; RR: respiratory rate; SPO2: saturation of peripheral oxygen; WBC: white blood cell.

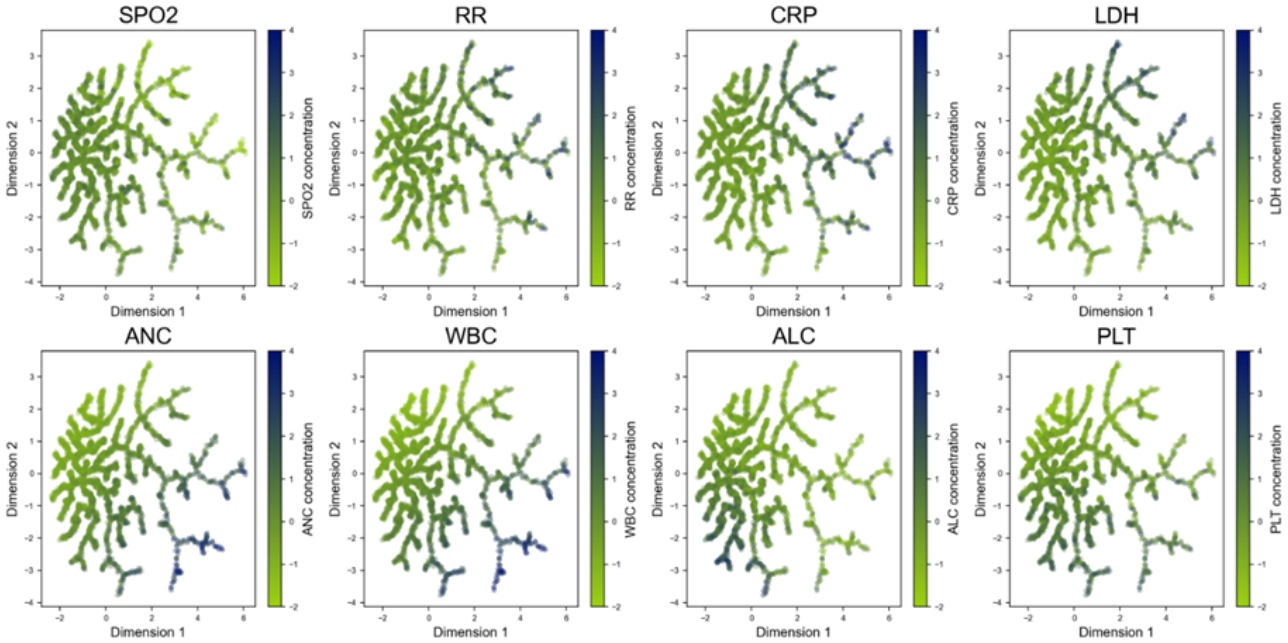

Definition of Variant-Dominant Periods

The predominant circulating variant at the time of hospitalization was identified through viral whole genome sequencing and could differ across nations [23]. According to predominant circulating variants during the pandemic in South Korea, we segmented our study period into 3 variant-dominant periods and constructed the corresponding patient subcohorts: original Alpha-dominant period (January 5, 2020 to May 1, 2021), Delta-dominant period (May 1, 2021 to November 24, 2021), and Omicron-dominant period (November 24, 2021 to August 24, 2022) [23].

Analysis of Model Transportability on Omicron Variant Cases

We developed modifications of RIETS to explore its prediction transportability across different variant-dominant periods. Each modified model was constructed using the variant dominant subcohorts in the development cohort. For instance, the "RIETS-All" model was based on the entire development cohort and the "RIETS-Omicron" model was based on the Omicron-dominant development cohort. We evaluated all possible combinations of modified RIETS and compared their discriminative performances among patients in the external validation cohort. Consequently, we identified the best-performing model, named "RIETS-Ensemble," that integrates the 3 models based on original Alpha-Omicron, Delta-Omicron, and Omicron. Then, the "RIETS-Ensemble" model was contrasted to the "RIETS-All" and "RIETS-Omicron" models to visualize marginal improvements. All developed models were compared using the AUROC as a measure for discriminative performance.

Statistical Analysis

Patient characteristics were summarized as median (IQR) and number (%) for continuous and categorical variables, respectively, and compared between the development and validation cohorts via absolute standardized mean difference (ASMD). The ASMD was calculated using Cohen D and H formulas for continuous and categorical variables, respectively. No considerable difference was identified with an ASMD below 0.2. For cutoff-based performance measures, Youden index was used to find an optimal threshold at which the average of sensitivity and specificity was maximized. The integrated calibration index (ICI), derived from the weighted mean difference between observed and predicted probabilities for the outcome, was used to quantify and assess calibration. ICI was preferred over other calibration metrics (eg, calibration-in-the-large and slope) due to its high stability from capturing the entire range of predicted probabilities during its computation [24]. A 2-sided *P* value below 0.05 was set to declare statistical significance. All statistical analyses were performed using Python (Python Software Foundation, version 3.9).

Results

Patient Characteristics

Among the 5945 hospitalized patients with COVID-19 used for the development and validation of RIETS, 4019 (67.6%) and 1926 (32.4%) were allocated into the development and validation cohorts, respectively (Table 1). The median age was higher in the development cohort than in the validation cohort (mean 60, SD 45-70 years vs mean 55, SD 35-65 years, respectively; ASMD=0.333). The proportion of male patients was similar in both the development (n=2130, 48.8%) and

XSL•FO
RenderX

validation (n=757, 47.9%) cohorts. Hypertension was the most prevalent comorbidity for both the development (n=1622, 37.2%) and validation (n=492, 31.1%) cohorts. While the most frequent symptoms across both cohorts were cough (n=2623, 44.1%) and fever (n=2366, 39.8%), the rankings of observed symptoms were similar in both cohorts. All variables pertaining to vital signs and blood biochemistry showed no considerable difference between cohorts (ASMD<0.2), except for the absolute neutrophil count (ANC; ASMD=0.319).

Table 1. Baseline characteristics in the development and validation cohorts of hospitalized South Korean patients with COVID-19.

	Total cohort (n=5945)	Development cohort (n=4019)	Validation cohort (n=1926)	ASMD[a]
Patient Characteristics				
Age (years), median (IQR)	60 (40-70)	60 (45-70)	50 (35-65)	0.333
Male sex, n (%)	2887 (48.6)	2130 (48.8)	757 (47.9)	0.018
Comorbidities, n (%)				
Hypertension	2114 (35.6)	1622 (37.2)	492 (31.1)	0.127
Diabetes mellitus	1249 (21)	978 (22.4)	271 (17.2)	0.132
Cardiovascular disease	508 (8.5)	389 (8.9)	119 (7.5)	0.050
Cancer	477 (8)	369 (8.5)	108 (6.8)	0.061
Others	2242 (37.7)	1813 (41.5)	429 (27.2)	0.304
Clinical symptoms, n (%)				
Fever	2366 (39.8)	1678 (38.4)	688 (43.5)	0.104
Cough	2623 (44.1)	1852 (42.4)	771 (48.8)	0.128
Sputum	1502 (25.3)	1038 (23.8)	464 (29.4)	0.127
Dyspnea	1318 (22.2)	1042 (23.9)	276 (17.5)	0.159
Myalgia	1388 (23.3)	938 (21.5)	450 (28.5)	0.162
Sore throat	1142 (19.2)	763 (17.5)	379 (24)	0.161
Loss of sensor	330 (5.6)	212 (4.9)	118 (7.5)	0.109
Gastrointestinal symptom	472 (7.9)	295 (6.8)	177 (11.2)	0.157
Vital sign, median (IQR)				
Body temperature (˚C)	36.6 (36.3-37.2)	36.6 (36.3-37.2)	36.5 (36.3-37.2)	0.051
Systolic blood pressure (mmHg)	129 (116-141)	129 (116-140)	129 (117-141)	0.021
Diastolic blood pressure (mmHg)	80 (70-87)	80 (70-86)	80 (72-90)	0.200
Pulse rate (counts)	84 (74-95)	84 (74-95)	86 (76-97)	0.128
Respiratory rate (counts)	20 (18-20)	20 (18-20)	20 (18-20)	0.191
SPO2[b] (%)	97 (96-98)	97 (96-98)	97 (96-98)	0.021
Blood biochemistry, median (IQR)				
White blood cells (10^3/μL)	5.2 (4.1-6.9)	5.3 (4.1-7.1)	5.1 (4.0-6.6)	0.143
Absolute neutrophil count (10^3/μL)	3.5 (2.4-5.4)	3.7 (2.5-5.8)	3.0 (2.1-4.3)	0.319
Absolute lymphocyte count (10^3/μL)	1.2 (0.8-1.7)	1.1 (0.8-1.6)	1.3 (1.0-1.8)	0.105
Platelet count (10^3/μL)	200 (157-248)	198 (154-248)	205 (166-248)	0.068
C-reactive protein (mg/dL)	1.2 (0.3-5.2)	1.5 (0.3-6.0)	0.7 (0.2-2.9)	0.077
Lactate dehydrogenase (U/L)	316 (221-445)	287 (212-428)	369 (289-476)	0.161

[a]ASMD: absolute standardized mean difference.

[b]SPO2: saturation of peripheral oxygen.

The baseline characteristics were also compared between 839 (14.1%) patients with nonsevere COVID-19 and 5106 (85.9%) patients with severe COVID-19 (Table S1 in Multimedia Appendix 1). Patients with severe COVID-19 were older (mean

XSL•FO
RenderX

70, SD 60-75 years vs mean 55, SD 40-70 years; ASMD=0.733). A larger proportion of patients with severe COVID-19 had dyspnea (ASMD=0.902) and diabetes mellitus (DM; ASMD=0.524). Patients with severe COVID-19 were more likely to have an increased respiratory rate (RR; ASMD=0.911) and decreased saturation of peripheral oxygen (SPO2; ASMD=0.705) upon hospital admission. In addition, patients with severe COVID-19 presented with higher ANC (ASMD=0.971), higher lactate dehydrogenase (LDH; ASMD=0.726), and higher white blood cell (WBC) count (ASMD=0.693).

Performance of RIETS

RIETS is a DNN-based final model with the subset of 11 features that demonstrated the highest discriminative power (AUROC=0.937, 95% CI 0.935-0.938; diagnostic odds ratio=46.14, 95% CI 43.40-48.87; specificity=0.867, 95% CI 0.865-0.869; sensitivity=0.869, 95% CI 0.864-0.875) amongst the 6 ML-based models (AUROC=0.862-0.929) (Figure 5A and Table 2). RIETS also exhibited a superior discriminative ability compared to the existing low risk of bias models (AUROC=0.60-0.80; Table S2 in Multimedia Appendix 1).

Figure 5. Performance comparisons of RIETS to other prediction models in the external validation. (A) Receiver operating characteristic curves for displaying the discriminative performances. (B) Calibration plots for showing the practical reliability of risk prediction. (C) Decision curve analysis plots for demonstrating the net clinical utility when deployed in a clinical practice. All dashed lines represent the references. Shaded areas represent the 95% CI bands. AUROC: area under receiver operating characteristic curve; GBM: gradient boosting machine; ICI: integrated calibration index; MLR: multivariable logistic regression; RIETS: Robust and Interpretable Early Triaging System; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting.

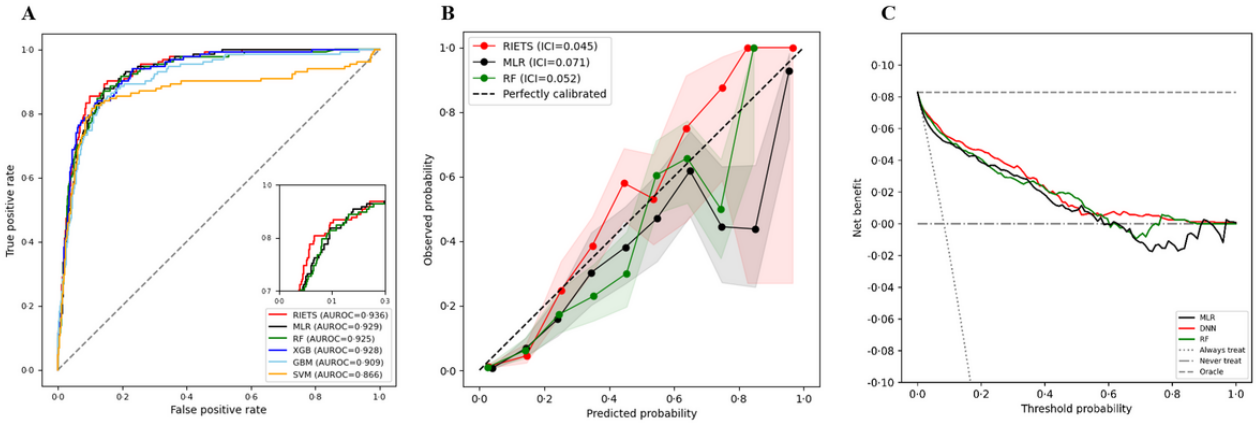

XSL•FO
RenderX

Table 2. Discriminative performance of machine learning–based models based on 11 clinical and laboratory features collected within the first day of hospitalization in the internal and external validation (IV and EV, respectively). The IV and EV results were computed from 5 iterations of stratified 5-fold cross-validations and 100 iterations of bootstrapped sampling, respectively. The Youden index was used to determine the optimal cutoff point.

Model and validation type	Predictive measures (95% CI)								Cutoff
	AUROC[a]	Sensitivity	Specificity	PPV[b]	NPV[c]	LRP[d]	LRN[e]	DOR[f]	
RIETS[g]									0.171
IV	0.891 (0.889-0.892)	0.852 (0.844-0.860)	0.808 (0.796-0.819)	46.58 (45.24-47.93)	96.59 (96.45-96.73)	4.57 (4.30-4.84)	0.18 (0.17-0.19)	25.96 (24.22-27.71)	
EV	0.937 (0.935-0.938)	0.869 (0.864-0.875)	0.867 (0.865-0.869)	37.65 (37.07-38.24)	98.63 (98.57-98.69)	6.56 (6.46-6.66)	0.15 (0.14-0.16)	46.14 (43.40-48.87)	
Multivariable logistic regression									0.164
IV	0.887 (0.886-0.888)	0.847 (0.845-0.849)	0.792 (0.789-0.795)	44.40 (43.99-44.81)	96.40 (96.35-96.45)	4.17 (4.08-4.26)	0.19 (0.19-0.20)	21.93 (21.35-22.52)	
EV	0.929 (0.927-0.930)	0.879 (0.874-0.884)	0.832 (0.830-0.834)	32.61 (32.08-33.14)	98.67 (98.61-98.73)	5.25 (5.18-5.32)	0.15 (0.14-0.15)	38.10 (36.11-40.10)	
Random forest									0.193
IV	0.894 (0.893-0.896)	0.848 (0.831-0.866)	0.807 (0.791-0.823)	46.47 (44.98-47.96)	96.51 (96.17-96.85)	4.55 (4.27-4.82)	0.19 (0.17-0.21)	25.40 (23.75-27.05)	
EV	0.925 (0.923-0.927)	0.863 (0.857-0.869)	0.864 (0.862-0.866)	36.94 (36.30-37.57)	98.56 (98.49-98.62)	6.36 (6.26-6.46)	0.16 (0.15-0.17)	42.35 (40.09-44.60)	
Extreme gradient boosting									0.105
IV	0.878 (0.875-0.880)	0.833 (0.813-0.853)	0.784 (0.772-0.796)	43.34 (42.36-44.33)	96.08 (95.68-96.47)	4.01 (3.82-4.20)	0.21 (0.19-0.23)	19.54 (18.05-21.02)	
EV	0.900 (0.898-0.903)	0.826 (0.820-0.832)	0.836 (0.835-0.838)	31.85 (31.33-32.37)	98.11 (98.04-98.19)	5.07 (5.00-5.14)	0.21 (0.20-0.21)	25.39 (24.21-26.57)	
Gradient boosting machine									0.010
IV	0.879 (0.877-0.882)	0.837 (0.820-0.853)	0.791 (0.775-0.808)	44.66 (42.94-46.37)	96.23 (95.94-96.51)	4.27 (3.95-4.60)	0.20 (0.19-0.22)	21.74 (20.75-22.74)	
EV	0.907 (0.904-0.910)	0.852 (0.847-0.858)	0.841 (0.839-0.843)	33.15 (32.56-33.74)	98.40 (98.33-98.47)	5.38 (5.30-5.47)	0.18 (0.17-0.18)	32.28 (30.58-33.97)	
Support vector machine									0.105
IV	0.834 (0.830-0.837)	0.782 (0.771-0.794)	0.833 (0.826-0.841)	48.15 (47.20-49.09)	95.20 (95.00-95.40)	4.88 (4.69-5.08)	0.26 (0.25-0.27)	19.19 (17.97-20.42)	
EV	0.862 (0.858-0.866)	0.794 (0.787-0.801)	0.899 (0.898-0.901)	42.12 (41.49-42.76)	97.93 (97.85-98.00)	7.92 (7.78-8.05)	0.23 (0.22-0.24)	35.83 (34.22-37.43)	

[a]AUROC: area under receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]LRP: likelihood ratio positive.

[e]LRN: likelihood ratio negative.

[f]DOR: diagnostic odds ratio.

XSL•FO
RenderX

[g]RIETS: Robust and Interpretable Early Triaging System.

In comparison with other ML-based models, RIETS exhibited net reclassification improvement (0.54%-6.14%), especially on nonsevere cases (2.14%-6.14%) (Table S3 and Figure S2 in Multimedia Appendix 1) and had the most stable prediction during cost sensitivity learning [25] (Figure S3 in Multimedia Appendix 1). RIETS also maintained sustainable prediction transportability (AUROC=0.903, 95% CI 0.897-0.910) on the limited number of cases (n=449, 7.6%) in the Omicron-dominant period when an ensemble learning technique was applied (Figure S4, Table S4, and Methods S4 in Multimedia Appendix 1).

Moreover, a PROBAST evaluation indicated that RIETS has a low risk of bias and minimal concerns regarding applicability (Methods S2 in Multimedia Appendix 1). RIETS also attained the best calibration (ICI=0.041) among comparable ML-based models (ICI=0.052-0.071; Figure 5B). Overall, it showed a higher net clinical benefit than the "intervention for none" and "intervention for all" reference strategies in DCA (Figure 5C).

Feature Interpretation

RIETS comprised 11 clinical and laboratory features: LDH, age, absolute lymphocyte counts (ALC), dyspnea, RR, DM, c-reactive protein (CRP), ANC, platelet counts (PLT), WBC, and SPO2. These features were ordered by their contribution to the severity prediction by using Shapley values (Figure 6). LDH was the highest ranked, followed by age, ALC, and dyspnea. In addition, pre-existing conditions (age, dyspnea, and DM) available at the time of admission were generally ranked higher relative to those of laboratory markers (CRP, ANC, PLT, WBC, and SPO2).

Figure 6. Average impact of 11 selected features in RIETS on COVID-19 severity prediction. To compute SHAP values based on the RIETS model, we properly modified KernelExplainer. Specific SHAP values are shown on the right of each feature name in the y-axis labels. ALC: absolute lymphocyte count; ANC: absolute neutrophil count; CRP: c-reactive protein; DM: diabetes mellitus; LDH: lactate dehydrogenase; PLT: platelet count; RIETS: Robust and Interpretable Early Triaging System; RR: respiratory rate; SHAP: Shapley additive explanations; SPO2: saturation of peripheral oxygen; WBC: white blood cell.

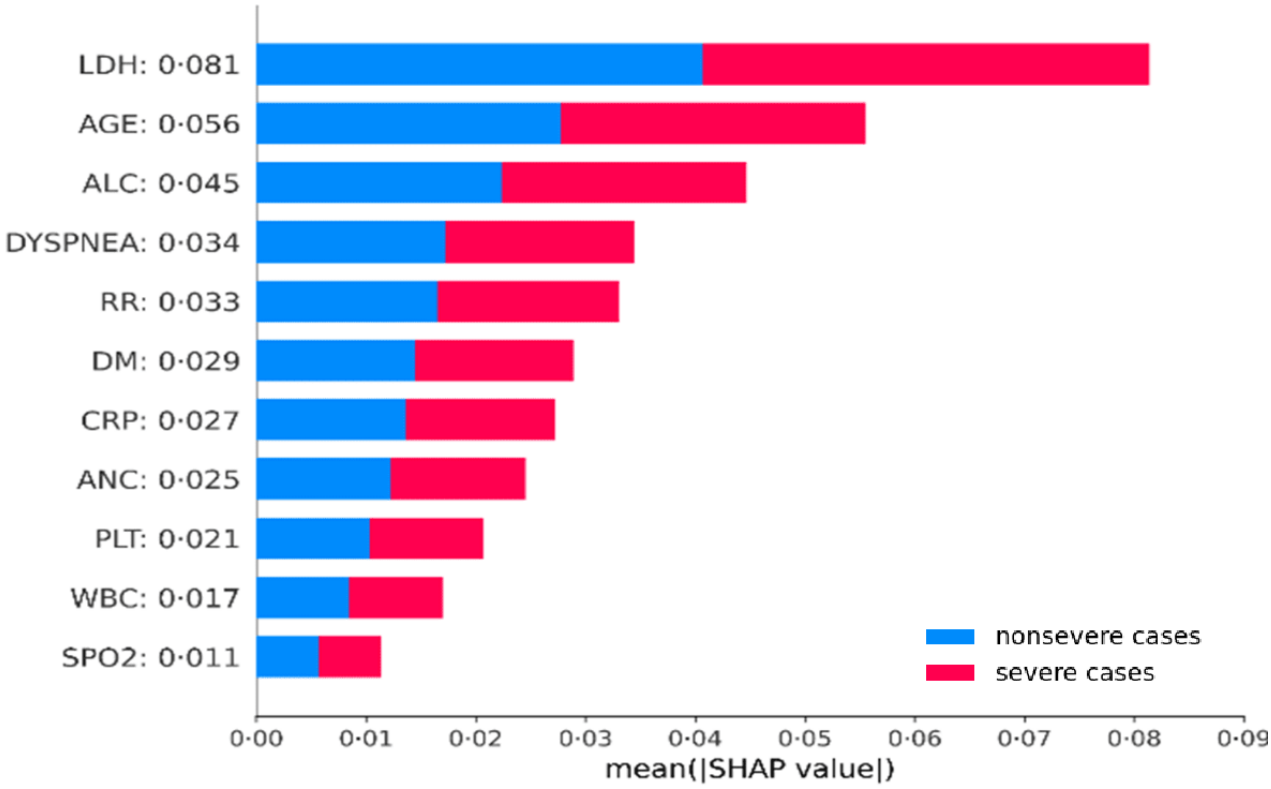

Patient Clustering and Characterization

We identified 4 patient subgroups using DDRTree, a tree-based unsupervised learning technique, based on the features in RIETS: the upper-right group (URG), middle-right group (MRG), lower-right group (LRG), and lower-left group (LLG) (Figure 3A). Among the 4 subgroups, the URG comprised the largest proportion of patients at high risk for severity, followed by the MRG, LRG, and LLG. The majority of patients in the URG and MRG had dyspnea, were older than 60 years, and had elevated RR, CRP, and LDH (Figure 3B). High ANC and WBC were additionally observed in the MRG. Those in the LRG and LLG had elevated ANC, WBC, and PLT. There was a negligible variation in SPO2 or DM across the tree. Moreover, we compared the patient distribution per each variant period (original Alpha-dominant, Delta-dominant, and Omicron-dominant) and found no distinguishable pattern (Figure S5 in Multimedia Appendix 1).

XSL•FO
RenderX

Discussion

Principal Findings

We developed and validated RIETS, an ML-based prognostic model for severity among patients hospitalized with COVID-19, based on a temporally and geographically extensive cohort with heterogeneous feature distributions (Figure S6 in Multimedia Appendix 1). RIETS incorporates 11 promptly and routinely available features upon hospitalization and is intended to assist early patient triaging. RIETS provides risk estimates that indicate the odds for severity progression along with feature and patient interpretation. These outputs can support clinicians in making decisions for appropriate medical measures, such as the administration of antiviral medication, transportation to the intensive care unit, and proactive preparation of medical resources. Although several prognostic models with low risk of bias excel in analogous tasks [26-30], RIETS offers substantial improvements in several aspects owing to its discriminative power and novel interpretability (Table S2 and S5 in Multimedia Appendix 1).

According to PROBAST, RIETS can be regarded as a clinically applicable model with a low risk of bias because of its generalizability and methodologically rigorous procedure. First, RIETS can be generalized across diverse populations because it was developed and validated based on a large data set from a multicenter cohort (19 general and tertiary care hospitals) over the 3-year pandemic period (from January 2020 to August 2022). In contrast, previous prognostic models were either based on a large multicenter cohort during the early pandemic period [28,31,32] or a single center cohort covering a longer pandemic period [29,30,33]. Second, we executed a rigorous modeling procedure to establish RIETS. We exhaustively developed and simultaneously validated all possible combinations of candidate feature subsets and modeling algorithms (Figure S7 in Multimedia Appendix 1). Contrary to our study design, previous prognostic studies relied on a single feature selection approach (clinical consensus, least absolute shrinkage and selection operator regression, recursive feature elimination, and sequential forward selection) [26,28,32,34,35]. Since there is no one-size-fits-all solution in the model fine-tuning process [36], this comprehensive modeling procedure can provide engineering value in attaining optimal prediction with parsimonious feature usage.

RIETS demonstrated superior discriminative performance in contrast to previous prognostic models with a low risk of bias (RIETS: AUROC=0.937, 95% CI 0.935-0.938; previous studies: 95% CI 0.60-0.80) while maintaining comparable calibration (ICI=0.041 vs calibration-in-the-large=0.00; slope=0.96; Table S2 in Multimedia Appendix 1). It has high accuracies both in severe cases (sensitivity=0.869, 95% CI 0.864-0.875) and in nonsevere cases (specificity=0.867, 95% CI 0.865-0.869). This strength can offer considerable benefits in triaging situations because prompt treatment for critically ill patients is facilitated without the resource overutilization on less critical patients [37].

RIETS also can be broadly adaptable across health care systems. Unlike some well-established models based on advanced technology-based, expensive, and time-consuming features [30,32,38], RIETS comprises 11 readily available features obtainable from routine blood tests and patient-reported conditions at admission. Thus, it is interoperable even for health care systems in low- and middle-income countries and may offer significant operational benefits during resource allocation across the global population [12]. In addition, RIETS exhibited sustainable performance on Omicron cases, implicating its potential for transportability across new variant cases with differing virulency [39] and limited case availability (Figure S4 and Table S4 in Multimedia Appendix 1).

Lastly, RIETS offers substantial interpretability that may induce improvements in model reliability and operational workflow. To our knowledge, this is the first attempt in patient clustering and characterization amongst the COVID-19 prognostic models (Figures 3 and 4). Given that bias risks are inevitable in ML systems, the interpretability of RIETS can promote transparent feedback, mitigate those bias risks, and earn trust as a clinical decision support system [40,41]. Moreover, the patient clustering tool in RIETS provides clinicians with useful information for treatment planning and resource preparation. For instance, the graphical representations of patients can enable monitoring of the characteristics of incoming patients and facilitate the identification of representative clusters at the moment. This can be used to plan the customized patient care and to initiate the preemptive preparation of medical resources for those representative patient clusters.

Limitations

This study has some limitations to be addressed. First, the study participants were patients hospitalized with COVID-19 in South Korea from January 2020 to August 2022. Hence, a further study with other ethnic and variant groups is recommended to validate the generalizability of RIETS. Second, the vaccination records were not accounted for during the analysis due to a high missing rate. Although vaccination often decreases the severity [42,43], a recent study showed that some vaccinated patients with certain chief complaints remained at high risk for severity [44]. This finding implicates that the impact of vaccination on severe case discrimination may not be large as long as the distributions of clinical signs remain similar across different variants. Lastly, the information on SARS-CoV-2 variants confirmed by viral whole genome sequencing were not available for each patient. We used variant dominant periods to define variant subcohorts while anticipating some misclassifications.

Conclusions

We developed and validated RIETS, an ML-based COVID-19 severity prediction system, to promote the early triaging of hospitalized patients with COVID-19. RIETS demonstrated high prediction power and considerable reliability with low bias risk. Model development and validation on a nationwide, multicenter cohort implicated its generalizability. The use of routinely collected features for model construction facilitated its adaptability. Visual interpretations of model parameters and patients improved its usability and applicability. When incorporated into routine clinical practice, we anticipate RIETS to have a direct clinical impact for enabling efficient medical resource allocation as well as proactive patient care.

XSL•FO
RenderX

Acknowledgments

The use and management of a cloud-based data storage platform was approved by the IRB of Samsung Medical Center (SMC 2020-09-100-002). This research was supported both by a grant of the Information and Communications Promotion Fund (ICT promotion fund) through the National IT Industry Promotion Agency (NIPA), funded by the Ministry of Science and ICT (MSIT), South Korea; and the Future Medicine 20*30 Project of Samsung Medical Center (SMO1230061). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Authors express appreciation to the patients involved across the 19 participated hospitals and to the healthcare practitioners for their assistance in data collection and extraction. Special thanks are given to Jongyeop Kim, Sungho Jung, Jungmin Choi, and Yena Kim for their roles in assembling, de-identifying, and organizing the raw data.

Data Availability

This study used deidentified data from cloud-based data storage constructed by the Medical AI Center of Samsung Medical Center. Given the multisite origin of these data, access to data will require approval from the clinical sites and participating institutions; requests can be made to the corresponding author. We have deployed RIETS as a web-based open-source application to allow for vast usability from the wider public [45].

Authors' Contributions

SB handled data preprocessing, model development, results interpretation, and manuscript drafting and revision. YJJ, YHK, JYK, JHK, EYK, JKL, and JK provided patient data and revised the manuscript. ZK assisted with administrative support and revised the manuscript. MJC conceived the idea, established a multicenter cloud-data platform, acquired funding, and revised the manuscript. KK designed and supervised the study, interpreted the results, and revised the manuscript. All authors have read and approved the final manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
RIETS for COVID-19 Severity Prediction.
[DOCX File , 1643 KB - jmir_v26i1e52134_app1.docx]

References

1. WHO COVID-19 dashboard. World Health Organization. URL: https://data.who.int/dashboards/covid19/cases?n=c [accessed 2024-01-02]
2. Gilbert A, Ghuysen A. Triage in the time of COVID-19. Lancet Digit Health 2022 Apr;4(4):e210-e211. [doi: 10.1016/s2589-7500(22)00001-2]
3. Aleem A, Samad ABA, Vaqar S. Emerging variants of SARS-CoV-2 and novel therapeutics against coronavirus (COVID-19). In: StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; 2023.
4. Lenharo M. WHO declares end to COVID-19's emergency phase. Nature. 2023 May 05. URL: https://www.nature.com/articles/d41586-023-01559-z [accessed 2024-01-02]
5. Rahimi F, Darvishi M, Bezmin Abadi AT. 'The end' - or is it? Emergence of SARS-CoV-2 EG.5 and BA.2.86 subvariants. Future Virol 2023 Sep;18(13):823 [FREE Full text] [doi: 10.2217/fvl-2023-0150] [Medline: 37736262]
6. Dyer O. Covid-19: infections climb globally as EG.5 variant gains ground. BMJ 2023 Aug 16;382:1900. [doi: 10.1136/bmj.p1900] [Medline: 37586736]
7. Mahase E. Covid-19: new "Pirola" variant BA.2.86 continues to spread in UK and US. BMJ 2023 Sep 13;382:2097. [doi: 10.1136/bmj.p2097] [Medline: 37704230]
8. Hu B, Guo H, Zhou P, Shi Z. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 2021 Mar;19(3):141-154 [FREE Full text] [doi: 10.1038/s41579-020-00459-7] [Medline: 33024307]
9. Gupta RK, Marks M, Samuels TH, Luintel A, Rampling T, Chowdhury H, et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. Eur Respir J 2020 Dec;56(6):2003498 [FREE Full text] [doi: 10.1183/13993003.03498-2020] [Medline: 32978307]
10. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]

XSL•FO
RenderX

11. Buttia C, Llanaj E, Raeisi-Dehkordi H, Kastrati L, Amiri M, Meçani R, et al. Prognostic models in COVID-19 infection that predict severity: a systematic review. Eur J Epidemiol 2023 Apr;38(4):355-372 [FREE Full text] [doi: 10.1007/s10654-023-00973-x] [Medline: 36840867]
12. Alakija A. Leveraging lessons from the COVID-19 pandemic to strengthen low-income and middle-income country preparedness for future global health threats. Lancet Infect Dis 2023 Aug;23(8):e310-e317. [doi: 10.1016/s1473-3099(23)00279-7]
13. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
14. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
15. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51. [doi: 10.7326/m18-1376]
16. Gallo Marin B, Aghagoli G, Lavine K, Yang L, Siff EJ, Chiang SS, et al. Predictors of COVID-19 severity: a literature review. Rev Med Virol 2021 Jan;31(1):1-10 [FREE Full text] [doi: 10.1002/rmv.2146] [Medline: 32845042]
17. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis 2020 Aug;20(8):e192-e197 [FREE Full text] [doi: 10.1016/S1473-3099(20)30483-7] [Medline: 32539990]
18. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol 2016 Mar;71:76-85. [doi: 10.1016/j.jclinepi.2015.10.002] [Medline: 26475568]
19. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016 Jan 25;352:i6 [FREE Full text] [doi: 10.1136/bmj.i6] [Medline: 26810254]
20. Jia R, Dao D, Wang B, Hubis FA, Hynes N, Gurel NM, et al. Towards efficient data valuation based on the Shapley value. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. 2019 Presented at: AISTATS 2019; April 16-18, 2019; Naha, Japan p. 1167-1176.
21. Mao Q, Wang L, Goodison S, Sun Y. Dimensionality reduction via graph structure learning. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD 2015; August 10-13, 2015; Sydney, Australia. [doi: 10.1145/2783258.2783309]
22. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017 Oct;14(10):979-982 [FREE Full text] [doi: 10.1038/nmeth.4402] [Medline: 28825705]
23. Lee D, Kim J, Park AK, Kim D, Kim J, Lim N, et al. Genomic epidemiology of SARS-CoV-2 Omicron variants in the Republic of Korea. Sci Rep 2022 Dec 27;12(1):22414 [FREE Full text] [doi: 10.1038/s41598-022-26803-w] [Medline: 36575217]
24. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019 Sep 20;38(21):4051-4065 [FREE Full text] [doi: 10.1002/sim.8281] [Medline: 31270850]
25. Mienye ID, Sun Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Inform Med Unlocked 2021;25:100690. [doi: 10.1016/j.imu.2021.100690]
26. Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, et al. Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. BMC Med 2021 Jan 21;19(1):23 [FREE Full text] [doi: 10.1186/s12916-020-01893-3] [Medline: 33472631]
27. Knight SR, Gupta RK, Ho A, Pius R, Buchan I, Carson G, et al. Prospective validation of the 4C prognostic models for adults hospitalised with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. Thorax 2022 Jun;77(6):606-615 [FREE Full text] [doi: 10.1136/thoraxjnl-2021-217629] [Medline: 34810237]
28. Goodacre S, Thomas B, Sutton L, Burnsall M, Lee E, Bradburn M, et al. Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: the PRIEST observational cohort study. PLoS One 2021;16(1):e0245840 [FREE Full text] [doi: 10.1371/journal.pone.0245840] [Medline: 33481930]
29. Raman G, Ashraf B, Demir YK, Kershaw CD, Cheruku S, Atis M, et al. Machine learning prediction for COVID-19 disease severity at hospital admission. BMC Med Inform Decis Mak 2023 Mar 07;23(1):46 [FREE Full text] [doi: 10.1186/s12911-023-02132-4] [Medline: 36882829]
30. Alaiad AI, Mugdadi EA, Hmeidi II, Obeidat N, Abualigah L. Predicting the severity of COVID-19 from lung CT images using novel deep learning. J Med Biol Eng 2023;43(2):135-146 [FREE Full text] [doi: 10.1007/s40846-023-00783-2] [Medline: 37077696]
31. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C mortality score. BMJ 2020 Sep 09;370:m3339 [FREE Full text] [doi: 10.1136/bmj.m3339] [Medline: 32907855]

XSL•FO
RenderX

32. Gupta R, Harrison EM, Ho A, Docherty AB, Knight SR, van Smeden M, et al. Development and validation of the ISARIC 4C deterioration model for adults hospitalised with COVID-19: a prospective cohort study. Lancet Respir Med 2021 Apr;9(4):349-359 [FREE Full text] [doi: 10.1016/S2213-2600(20)30559-2] [Medline: 33444539]
33. Nguyen HTT, Le-Quy V, Ho SV, Thomsen JHD, Pontoppidan Stoico M, Tong HV, et al. Outcome prediction model and prognostic biomarkers for COVID-19 patients in Vietnam. ERJ Open Res 2023 Mar;9(2):00481-2022 [FREE Full text] [doi: 10.1183/23120541.00481-2022] [Medline: 37041987]
34. Kim J, Lim H, Ahn J, Lee KH, Lee KS, Koo KC. Optimal triage for COVID-19 patients under limited health care resources with a parsimonious machine learning prediction model and threshold optimization using discrete-event simulation: development study. JMIR Med Inform 2021 Nov 02;9(11):e32726 [FREE Full text] [doi: 10.2196/32726] [Medline: 34609319]
35. Yadaw AS, Li Y, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health 2020 Oct;2(10):e516-e525. [doi: 10.1016/s2589-7500(20)30217-x]
36. Abiodun EO, Alabdulatif A, Abiodun OI, Alawida M, Alabdulatif A, Alkhawaldeh RS. A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Comput Appl 2021;33(22):15091-15118 [FREE Full text] [doi: 10.1007/s00521-021-06406-8] [Medline: 34404964]
37. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019 May 28;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
38. Jiao Z, Choi JW, Halsey K, Tran TML, Hsieh B, Wang D, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. Lancet Digit Health 2021 May;3(5):e286-e294. [doi: 10.1016/s2589-7500(21)00039-x]
39. Esper F, Adhikari TM, Tu ZJ, Cheng YW, El-Haddad K, Farkas DH, et al. Alpha to Omicron: disease severity and clinical outcomes of major SARS-CoV-2 variants. J Infect Dis 2023 Feb 01;227(3):344-352 [FREE Full text] [doi: 10.1093/infdis/jiac411] [Medline: 36214810]
40. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? J Med Ethics 2022 Sep;48(9):581-585 [FREE Full text] [doi: 10.1136/medethics-2020-107102] [Medline: 34006600]
41. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
42. Kim Y, Choe YJ, Kim J, Kim RK, Jang EJ, Lee H, et al. Vaccine effectiveness against severe disease and death for patients with COVID-19 during the Delta-dominant and Omicron-emerging periods: a K-COVE study. J Korean Med Sci 2023 Mar 20;38(11):e87 [FREE Full text] [doi: 10.3346/jkms.2023.38.e87] [Medline: 36942395]
43. Hyams C, Challen R, Marlow R, Nguyen J, Begier E, Southern J, et al. Severity of Omicron (B.1.1.529) and Delta (B.1.617.2) SARS-CoV-2 infection among hospitalised adults: a prospective cohort study in Bristol, United Kingdom. Lancet Reg Health Eur 2023 Feb;25:100556 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100556] [Medline: 36530491]
44. Agrawal U, Bedston S, McCowan C, Oke J, Patterson L, Robertson C, et al. Severe COVID-19 outcomes after full vaccination of primary schedule and initial boosters: pooled analysis of national prospective cohort studies of 30 million individuals in England, Northern Ireland, Scotland, and Wales. Lancet 2022 Oct 15;400(10360):1305-1320 [FREE Full text] [doi: 10.1016/S0140-6736(22)01656-7] [Medline: 36244382]
45. Baek SW. RIETS Web Model. URL: https://riets-web.azurewebsites.net [accessed 2024-01-03]

Abbreviations

ALC: absolute lymphocyte counts
ANC: absolute neutrophil count
ASMD: absolute standardized mean difference
AUROC: area under the receiver operating characteristic curve
CRP: c-reactive protein
DCA: decision curve analysis
DDRTree: discriminative dimensionality reduction via learning a tree
DM: diabetes mellitus
DNN: deep neural network
FEM: feature engineering method
ICI: integrated calibration index
IRB: institutional review board
LDH: lactate dehydrogenase
LLG: lower-left group

XSL•FO
RenderX

LRG: lower-right group
ML: machine learning
MRG: middle-right group
PLT: platelet count
PROBAST: Diagnosis and Prediction Model Risk of Bias Assessment Tool
RF: random forest
RR: respiratory rate
SPO2: saturation of peripheral oxygen
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis
URG: upper-right group
WBC: white blood cell
WHO: World Health Organization
XGB: extreme gradient boosting

Edited by A Mavragani; submitted 24.08.23; peer-reviewed by KF Chen, P Okoro, Y Su; comments to author 18.10.23; revised version received 03.11.23; accepted 25.12.23; published 11.01.24.

Please cite as:
Baek S, Jeong YJ, Kim YH, Kim JY, Kim JH, Kim EY, Lim JK, Kim J, Kim Z, Kim K, Chung MJ
Development and Validation of a Robust and Interpretable Early Triaging Support System for Patients Hospitalized With COVID-19: Predictive Algorithm Modeling and Interpretation Study
J Med Internet Res 2024;26:e52134
URL: https://www.jmir.org/2024/1/e52134
doi:10.2196/52134
PMID:38206673

©Sangwon Baek, Yeon joo Jeong, Yun-Hyeon Kim, Jin Young Kim, Jin Hwan Kim, Eun Young Kim, Jae-Kwang Lim, Jungok Kim, Zero Kim, Kyunga Kim, Myung Jin Chung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of an Interpretable Conformal Predictor to Predict Sepsis Mortality Risk: Retrospective Cohort Study

Meicheng Yang[1], PhD; Hui Chen[2], MD; Wenhan Hu[2], PhD; Massimo Mischi[3], PhD; Caifeng Shan[4,5], PhD; Jianqing Li[1,6*], PhD; Xi Long[3*], PhD; Chengyu Liu[1*], PhD

[1]State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing, China

[2]Department of Critical Care Medicine, Jiangsu Provincial Key Laboratory of Critical Care Medicine, Zhongda Hospital, Southeast University, Nanjing, China

[3]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands

[4]College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China

[5]School of Intelligence Science and Technology, Nanjing University, Nanjing, China

[6]School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

[*]these authors contributed equally

Corresponding Author:
Chengyu Liu, PhD
State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering
Southeast University
35 Jinxianghe Road
Nanjing, 210096
China
Phone: 86 25 83793993
Email: chengyu@seu.edu.cn

Abstract

Background: Early and reliable identification of patients with sepsis who are at high risk of mortality is important to improve clinical outcomes. However, 3 major barriers to artificial intelligence (AI) models, including the lack of interpretability, the difficulty in generalizability, and the risk of automation bias, hinder the widespread adoption of AI models for use in clinical practice.

Objective: This study aimed to develop and validate (internally and externally) a conformal predictor of sepsis mortality risk in patients who are critically ill, leveraging AI-assisted prediction modeling. The proposed approach enables explaining the model output and assessing its confidence level.

Methods: We retrospectively extracted data on adult patients with sepsis from a database collected in a teaching hospital at Beth Israel Deaconess Medical Center for model training and internal validation. A large multicenter critical care database from the Philips eICU Research Institute was used for external validation. A total of 103 clinical features were extracted from the first day after admission. We developed an AI model using gradient-boosting machines to predict the mortality risk of sepsis and used Mondrian conformal prediction to estimate the prediction uncertainty. The Shapley additive explanation method was used to explain the model.

Results: A total of 16,746 (80%) patients from Beth Israel Deaconess Medical Center were used to train the model. When tested on the internal validation population of 4187 (20%) patients, the model achieved an area under the receiver operating characteristic curve of 0.858 (95% CI 0.845-0.871), which was reduced to 0.800 (95% CI 0.789-0.811) when externally validated on 10,362 patients from the Philips eICU database. At a specified confidence level of 90% for the internal validation cohort the percentage of error predictions (n=438) out of all predictions (n=4187) was 10.5%, with 1229 (29.4%) predictions requiring clinician review. In contrast, the AI model without conformal prediction made 1449 (34.6%) errors. When externally validated, more predictions (n=4004, 38.6%) were flagged for clinician review due to interdatabase heterogeneity. Nevertheless, the model still produced significantly lower error rates compared to the point predictions by AI (n=1221, 11.8% vs n=4540, 43.8%). The most important predictors identified in this predictive model were Acute Physiology Score III, age, urine output, vasopressors, and pulmonary infection. Clinically relevant risk factors contributing to a single patient were also examined to show how the risk arose.

XSL•FO
RenderX

Conclusions: By combining model explanation and conformal prediction, AI-based systems can be better translated into medical practice for clinical decision-making.

(J Med Internet Res 2024;26:e50369) doi:10.2196/50369

KEYWORDS

sepsis; critical care; clinical decision-making; mortality prediction; conformal prediction

Introduction

Sepsis is a life-threatening systemic illness resulting from a dysregulated host response to microbial invasion (infection) and is associated with high morbidity and mortality [1,2]. According to the most recent Global Burden of Diseases study, nearly 49 million people experience sepsis each year and approximately 11 million die from sepsis and its complications, accounting for 19.7% of all deaths worldwide [3]. In addition, a Chinese epidemiological study in 2020 showed that the incidence of sepsis in intensive care units (ICUs) was 20.6%, with a mortality rate of 35.5% [4]. Recent evidence suggests that early identification of patients who are critically ill with the potential for acute deterioration is effective in improving clinical outcomes [5]. Therefore, a pragmatic model that could help identify high-risk patients and further improve the prognosis of patients with sepsis is critical.

Recent significant increases in electronic health record data and advancements in artificial intelligence (AI) have led to rapid growth in the development of machine learning algorithms to identify patients with sepsis at high risk of in-hospital mortality [6-8]. However, there are 3 major barriers to the widespread adoption of AI models for their deployment in clinical practice: first, the lack of interpretability of AI models; second, the lack of external validation so that the model generalizability across institutions cannot be guaranteed; and third, the risk of automation bias, where users tend to rely too much on the system output rather than actively seeking information and assessing the model uncertainty.

Although numerous AI methods, especially deep learning, have demonstrated remarkable performance in medicine, surprisingly, few constructed models have been used in clinical practice, with poor interpretability being a major reason [9]. Clinicians need to understand how AI-based algorithms generate their predictions and gain insight into the precise changes in risk induced by certain factors of an individual patient [10,11]. Moreover, the majority of current sepsis mortality prediction models published are built on data from a single hospital or a uniform health care system, where the care processes are standardized or similar [6-8]. It is challenging for AI models to ensure accuracy in different hospital settings [7,12]. Therefore, another challenge in predictive modeling is how to quantify the reliability of the model predictions for new patients, especially when such data are outside the “domain” on which the model was trained [13]. Furthermore, most AI models only provide binary predictions, that is, yes or no, without assessing how reliable a prediction is [14]. However, when using AI models in high-risk environments (such as ICUs), uncertainty quantification is required to avoid unexpected model failures by gaining insight into the confidence of the predictions made by an AI algorithm [15].

To the best of our knowledge, no models have been explicitly developed to estimate the uncertainty of AI-assisted critical illness risk predictions in patients admitted to the ICU using electronic health record data. In this study, we aimed to develop and validate an AI model, called CPMORS (Conformal Predictor for Mortality Risk in Sepsis), to assess the risk of in-hospital sepsis mortality in ICU admissions. To mitigate the impact of insufficient model generalization performance and automation bias on its clinical application, we expected the model to provide confidence measures to monitor predictions and flag uncertain predictions at a customized confidence level for human intervention, as well as to provide interpretable risk factors, in this case, to improve the translation of AI-assisted sepsis prediction systems into medical practice and enable intensivists to use them in clinical decision-making.

Methods

Data Sources

The data used in this retrospective study were obtained from 2 different databases with different clinical information systems: Medical Information Mart for Intensive Care database-IV (MIMIC-IV; version 2.2; MetaVision system) [16] and eICU Collaborative Research Database (eICU-CRD; Philips eICU system) [17]. MIMIC-IV provided critical care data of 73,181 patients admitted to the ICUs at Beth Israel Deaconess Medical Center between 2008 and 2019. eICU-CRD is a large multicenter ICU database including more than 200,000 ICU admissions from 335 units in 208 hospitals across the United States between 2014 and 2015.

Ethical Considerations

The MIMIC-IV and eICU-CRD were publicly available databases and were previously ethically approved by the institutional review boards at Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology in accordance with the tenets of the Declaration of Helsinki. The waiver of the requirement for informed consent was included in the institutional review board approval as all protected health information was deidentified [16,17]. The authors were granted access to the database after completing training in human research and signing a data use agreement in PhysioNet (certification number: 27252652).

Participant Selection

In MIMIC-IV, participants were enrolled based on the sepsis-3 criteria [2], that is, known or suspected infection and a Sequential Organ Failure Assessment (SOFA) score ≥2 points. In eICU-CRD, patients with sepsis were identified according

XSL•FO
RenderX

to the admission diagnosis recorded in the Acute Physiology and Chronic Health Evaluation IV data set [18]. For those with multiple ICU admissions, only the first ICU admission was included. Patients who were discharged from the ICU within 24 hours and were aged <18 years were excluded. Patients with >30% missing individual data were also excluded.

Data Extraction and Preprocessing

Patient data from the first 24 hours after ICU admission were retrieved from the MIMIC-IV and eICU-CRD databases to predict in-hospital mortality risk. The study retrospectively collected the following data: (1) demographic characteristics, including sex, age, BMI, and ethnicity; (2) site of infection, including pulmonary and gastrointestinal infections; (3) comorbidities, including chronic kidney disease, congestive heart failure, chronic pulmonary disease, diabetes, and liver disease; (4) worst reported Glasgow Coma Scale, Acute Physiology Score III (APS III), and SOFA score; (5) vital signs, including heart rate, respiratory rate, systolic blood pressure, mean blood pressure, diastolic blood pressure, temperature, and oxygen saturation; (6) laboratory data, including pH, lactate, bicarbonate, base excess, PaO_2, $PaCO_2$, FiO_2, PaO_2/FiO_2 ratio, hematocrit, hemoglobin, platelets, white blood cells count, albumin, anion gap, blood glucose, blood urea nitrogen, serum calcium, serum creatinine, serum sodium, serum potassium, international normalized ratio, prothrombin time, partial thromboplastin time, alanine transaminase, alkaline phosphatase, aspartate aminotransferase, and total bilirubin; (7) therapeutic management, including the use of vasopressors; and (8) total urine output.

We used one-hot encoding for the representation of categorical variables. For vital signs with multiple measurements during the first day, we included the maximum, minimum, mean, and SD values for analysis. For laboratory values with multiple measurements, we included the maximum and minimum values for analysis. This resulted in a total of 103 features used to train and validate AI models for in-hospital mortality risk prediction (Multimedia Appendix 1). For missing data, the BMI was imputed with the k-nearest neighbors algorithm using the demographic characteristics. For the remaining missing values, if the predictive models could not support the missing data, the mean value from the training data was used to fill the remaining missing values; if the predictive models could support the missing data, no imputation was performed. All numerical features were standardized by removing the mean and scaling to unit variance. To avoid information leakage, the preprocessing operations were derived from the training data and applied to other validation data sets.

Model Development and Validation

The CPMORS model was developed to predict in-hospital mortality risk from sepsis and provide uncertain predictions and risk factors for further possible active management (Figure 1). Patients with sepsis from the MIMIC-IV database were randomly divided into a development set (n=16,746, 80%) for model training and an internal validation set (n=4187, 20%) for testing (Figure 2). Septic ICU admissions derived from the eICU-CRD database were used for external validation. The CPMORS model was constructed using the gradient boosting machines [19] prediction algorithm. Three common machine learning algorithms were also constructed for comparative purposes, including neural decision forest [20], random forest [21], and logistic regression [7]. Conventional scoring systems that have been widely used in clinical practice without machine learning, including APS III and SOFA, were also tested for comparison. Missing data were handled with a mean imputation method for neural decision forest, random forest, and logistic regression, while the gradient boosting machines did not require imputation.

XSL•FO
RenderX

Figure 1. Visualization of the use of CPMORS in AI-assisted sepsis mortality risk prediction. The implication of multiple predictions is that there is insufficient information for the model to discriminate between survival and nonsurvival outcomes, which are flagged for clinician review. Risk factors are the explanations provided by the Shapley values. Risk factors in red are more serious and risk factors in green are in the normal range. AI: artificial intelligence; APS III: Acute Physiology Score III; BUN: blood urea nitrogen; CPMORS: Conformal Predictor for Mortality Risk in Sepsis; GCS: Glasgow Coma Scale; HR: heart rate; MBP: mean blood pressure; SOFA: Sequential Organ Failure Assessment; SpO_2: oxygen saturation; WBC: white blood cell.

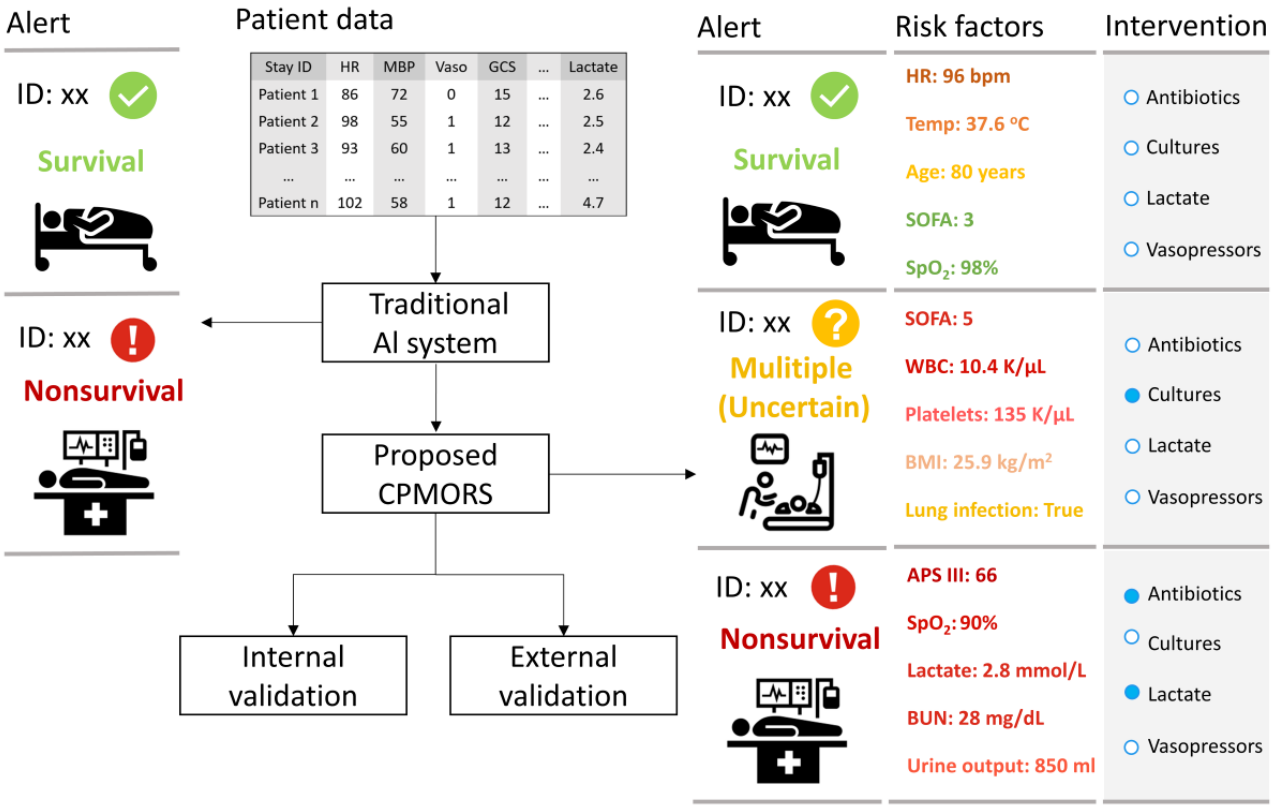

To estimate the uncertainty of the model outputs at a customized or user-specified confidence level that can be set by clinicians, we used the conformal prediction (CP) framework built on top of the prediction algorithm. CP is a user-friendly paradigm for generating statistically rigorous uncertainty sets or intervals for the predictions of unknown samples that differ from the training data [15]. This approach could provide reliable predictions at a user-specified desired error rate (equal to the significance level and 1-confidence). The output of CP is a prediction region, that is, sets of labels, rather than a single value prediction (point prediction) for classification problems. In this study, the possible prediction sets were {survival} or {nonsurvival}, called a single prediction, {survival, nonsurvival}, called multiple predictions, or {null} called the empty set. Central to CP is the use of nonconformity measures to assess how dissimilar a new sample is from the data on which the model was built. In this study, we used a common nonconformity measure, that is, using the predicted probability of an example belonging to a given class to calculate the nonconformity score. The nonconformity was then used to calculate a *P* value for each possible class label when making a prediction using the conformal predictor. The *P* value represented the proportion of observations with more extreme nonconformity scores. Labels were included in the prediction set if the *P* value exceeded a user-specified desired significance level (1-confidence), such as .05 (Table 1). A multiple prediction meant that the prediction was uncertain, and the model could not distinguish between survival and nonsurvival. Empty predictions were examples where the model could not assign any label, typically meaning that the example differed from the data the model was trained on. In this study, a Mondrian CP was specifically implemented to handle classification tasks with unbalanced data [22]. It could work on a class basis to ensure the desired error rate within each class. At a higher confidence level, we got fewer error predictions but more multiple predictions (Table 2). To develop the Mondrian conformal predictor, we further split the development set into a training set (n=13,397, 80%) and a calibration set (n=3349, 20%; Figure 2). The training set was used to train the AI prediction algorithm, while the calibration set was used to construct the conformal predictor and also to tune the model hyperparameters using a Bayesian optimizer [23].

XSL•FO
RenderX

Table 1. Examples of the formation of the prediction set.

Confidence level	Significance level	Higher than the significance level (yes or no)		Prediction set
		P for survival=.37	*P* for nonsurvival=.08	
0.60	.40	No	No	{null}
0.85	.15	Yes	No	{survival}
0.95	.05	Yes	Yes	{survival, nonsurvival}

Table 2. Examples of predictions at 2 different confidence levels.

ID	Confidence level=85%			Confidence level=95%		
	True label[a]	Prediction set[a]	Prediction type[b]	True label[a]	Prediction set[a]	Prediction type[b]
1	0	{0, 1}	Multiple	0	{0, 1}	Multiple
2	1	{1}	Correct	1	{1}	Correct
3	1	{0}	Error	1	{0, 1}	Multiple
4	1	{0, 1}	Multiple	1	{0, 1}	Multiple
5	0	{null}	Empty	0	{1}	Error
6	0	{1}	Error	0	{1}	Error
7	0	{0}	Correct	0	{0}	Correct
8	1	{1}	Correct	1	{1}	Correct
9	1	{0}	Error	1	{0, 1}	Multiple
10	1	{1}	Correct	1	{0, 1}	Multiple

[a]Label 0=survival, label 1=nonsurvival.

[b]Correct and error predictions refer to single predictions.

XSL•FO
RenderX

Figure 2. Workflow for the proposed CPMORS model. CPMORS: Conformal Predictor for Mortality Risk in Sepsis; SHAP: Shapley additive explanation.

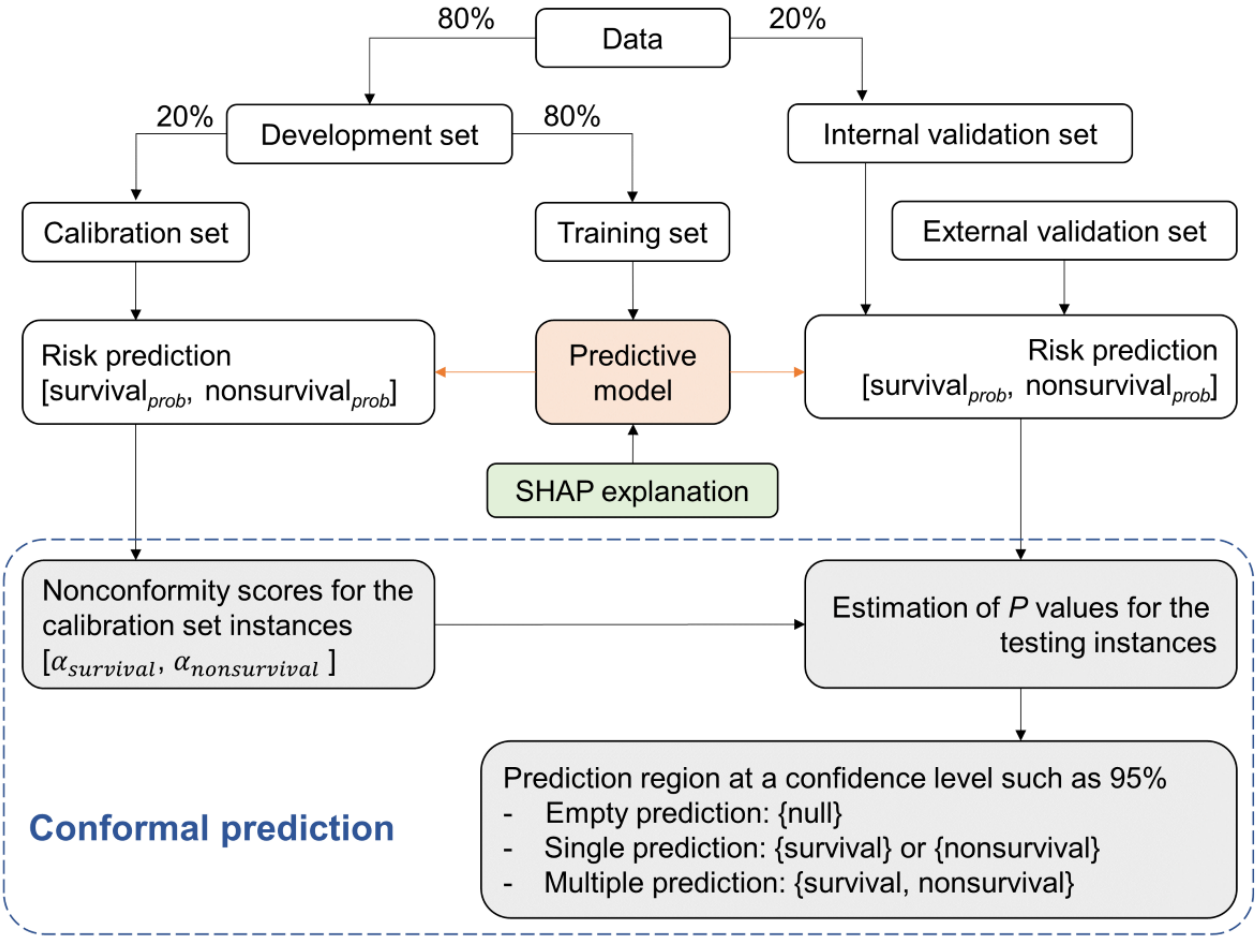

To explain the model, the impact of the features on the risk output was quantified using Shapley additive explanation (SHAP) values [24] to obtain interpretability for the developed model. We provide both global feature importance from the whole population outputs and individual interpretability for a single patient output.

Statistical Analysis

Continuous variables were expressed as median with IQR, and 2 groups were compared using the Wilcoxon rank sum test. Categorical variables were expressed as numbers and percentages and compared using chi-square tests. The discriminative performance of the model for predicting sepsis mortality was assessed using the area under the curve (AUC), and calibration was performed using calibration curves and the Brier score (the lower the better). We calculated the 95% CIs for these metrics using bootstrapping (2000 stratified bootstrap replicates).

The conformal predictor could produce multiple predictions or empty predictions in cases where it could not assign reliable single predictions. Therefore, it was not possible to directly calculate the sensitivity and specificity of the CP. To evaluate the CP framework, we assessed efficiency, defined as the proportion of all predictions that resulted in a single correct prediction, and validity (the error rate), the proportion of all predictions that did not exceed the prespecified significance level [15]. All statistical analysis and calculations were performed using R (version 4.2.2; R Foundation for Statistical Computing) and Python (version 3.8.16; Python Software Foundation).

Results

Patient Description

This study follows the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [25]. A total of 20,933 adult patients from the MIMIC-IV database meeting the sepsis-3 criteria were analyzed, of whom 3457 (16.5%) were nonsurvivors. The external validation cohort from the eICU-CRD included 10,362 patients with sepsis (n=1757, 17% for nonsurvivors). Table 3 describes the baseline characteristics between survivors and nonsurvivors of patients with sepsis admitted to the ICU. Multimedia Appendix 1 shows that 93 out of 103 features were statistically different ($P \leq .05$) between MIMIC-IV and eICU-CRD. In both data sets, compared with patients whose outcome was in-hospital survival, nonsurvivors were older, had a higher BMI, had more pulmonary infections, were more likely to receive vasopressors, had higher APS III and SOFA scores, and had a longer ICU stay.

XSL•FO
RenderX

Table 3. Baseline characteristics of the patients with sepsis following ICU[a] admission.

Variables	MIMIC-IV[b] (n=20,933)			eICU-CRD[c] (n=10,362)		
	Survivors (n=17,476)	Nonsurvivors (n=3457)	*P* value	Survivors (n=8605)	Nonsurvivors (n=1757)	*P* value
Age (years), median (IQR)	67 (56-77)	71 (60-81)	<.001	66 (54-76)	70 (60-80)	<.001
Male, n (%)	10,285 (58.9)	1938 (56.1)	.002	4606 (53.5)	952 (54.2)	.62
BMI (kg/m^2), median (IQR)	27.2 (23.8-31.7)	26.5 (23.1-31.3)	<.001	27.3 (23.1-33.2)	25.9 (22.0-31.1)	<.001
Ethnicity, n (%)			<.001			.97
Asian	508 (2.9)	100 (2.9)		140 (1.6)	27 (1.5)	
Black	1739 (10)	315 (9.1)		868 (10.1)	181 (10.3)	
White	11.972 (68.5)	2134 (61.7)		6699 (77.9)	1361 (77.5)	
Other	3257 (18.6)	908 (26.3)		898 (10.4)	188 (10.7)	
Comorbidities, n (%)						
Chronic kidney disease	2913 (16.7)	673 (19.5)	<.001	1040 (12.1)	277 (15.8)	<.001
Congestive heart failure	5575 (31.9)	1332 (38.5)	<.001	1907 (22.2)	429 (24.4)	.04
Chronic pulmonary disease	4882 (27.9)	1055 (30.5)	.002	2015 (23.4)	415 (23.6)	.86
Liver disease	1209 (6.9)	487 (14.1)	<.001	96 (1.1)	79 (4.5)	<.001
Diabetes	5832 (33.4)	1103 (31.9)	.09	2976 (34.6)	562 (32)	.04
Pulmonary infection, n (%)	6958 (39.8)	1983 (57.4)	<.001	4299 (50)	933 (53.1)	.02
Vasopressors, n (%)	8836 (50.6)	2518 (72.8)	<.001	2450 (28.5)	835 (47.5)	<.001
Urine output (mL), median (IQR)	1640 (1030-2465)	956 (445-1715)	<.001	1320 (689-2223)	700 (250-1351)	<.001
Glasgow Coma Scale (score), median (IQR)	15 (13-15)	15 (12-15)	.02	14 (9-15)	10 (6-14)	<.001
APS III[d] (score), median (IQR)	46 (35-60)	66 (51-84)	<.001	51 (38-67)	71 (54-92)	<.001
SOFA[e] (score), median (IQR)	5 (4-8)	8 (6-11)	<.001	4 (2-7)	7 (4-10)	<.001
Length of ICU stay (hours), median (IQR)	72 (44-138)	109 (58-208)	<.001	68 (43-121)	86 (46-170)	<.001

[a]ICU: intensive care unit.

[b]MIMIC-IV: Medical Information Mart for Intensive Care database-IV.

[c]eICU-CRD: eICU Collaborative Research Database.

[d]APS III: Acute Physiology Score III.

[e]SOFA: Sequential Organ Failure Assessment.

Prediction Performance and Explanation

Figure 3 shows the prediction performance of the developed 6 models in terms of receiver operating characteristic curves. When tested on the internal validation population for MIMIC-IV, the AUC value obtained using the proposed CPMORS model was 0.858 (95% CI 0.845-0.871), which is significantly higher than the other 3 machine learning models ranging from 0.829 (95% CI 0.811-0.844) to 0.853 (95% CI 0.838-0.867) and the 2 scoring systems with 0.708 (95% CI 0.685-0.728) of SOFA and 0.757 (95% CI 0.737-0.774) of APS III. When externally validated on the multicenter database eICU-CRD, the performance of all models deteriorated due to interdatabase heterogeneity. Nevertheless, CPMORS still showed the best performance with AUC 0.800 (95% CI 0.789-0.811) versus other models ranging from 0.693 (95% CI 0.680-0.706) to 0.786 (95% CI 0.775-0.798). In addition, the model showed good calibration (Figure 3C), where the curves were close to the diagonal dash line. The Brier score was 0.101 (95% CI 0.096-0.106) in MIMIC-IV and 0.116 (95% CI 0.113-0.119) in eICU-CRD, indicating its high calibration ability.

XSL•FO
RenderX

Figure 3. Prediction performance and explanation. (A,B) Receiver operating characteristic curves of MMIC-IV and eICU-CRD validation performance of different models in predicting sepsis mortality. (C) Calibration plots of the proposed CPMORS model. (D) Summary of the interpretability of the CPMORS. Beeswax plots show the feature importance across patients for the top 15 features, where each point indicates the feature importance value for 1 patient sample. Where multiple dots fall on the same x position, they are stacked to show density. Features with positive impact values push the risk up, while negative impact values push the risk down. Long tails indicate features that are extremely important for some patients. (E) Examples of the relationship between SHAP value and feature value. (F) Explanation of how the sepsis risk score is output for a nonsurvivor. (G) Explanation of how the sepsis risk score is output for a survivor. APS III: Acute Physiology Score III; aPTT: activated partial thromboplastin time; AUC: area under the curve; BUN: blood urea nitrogen; CPMORS: Conformal Predictor for Mortality Risk in Sepsis; eICU-CRD: eICU Collaborative Research Database; LR: logistic regression; MMIC-IV: Medical Information Mart for Intensive Care database-IV; NDF: neural decision forest; PLT: platelet; RF: random forest; SHAP: Shapley additive explanation; SOFA: Sequential Organ Failure Assessment; SpO_2: oxygen saturation; WBC: white blood cell.

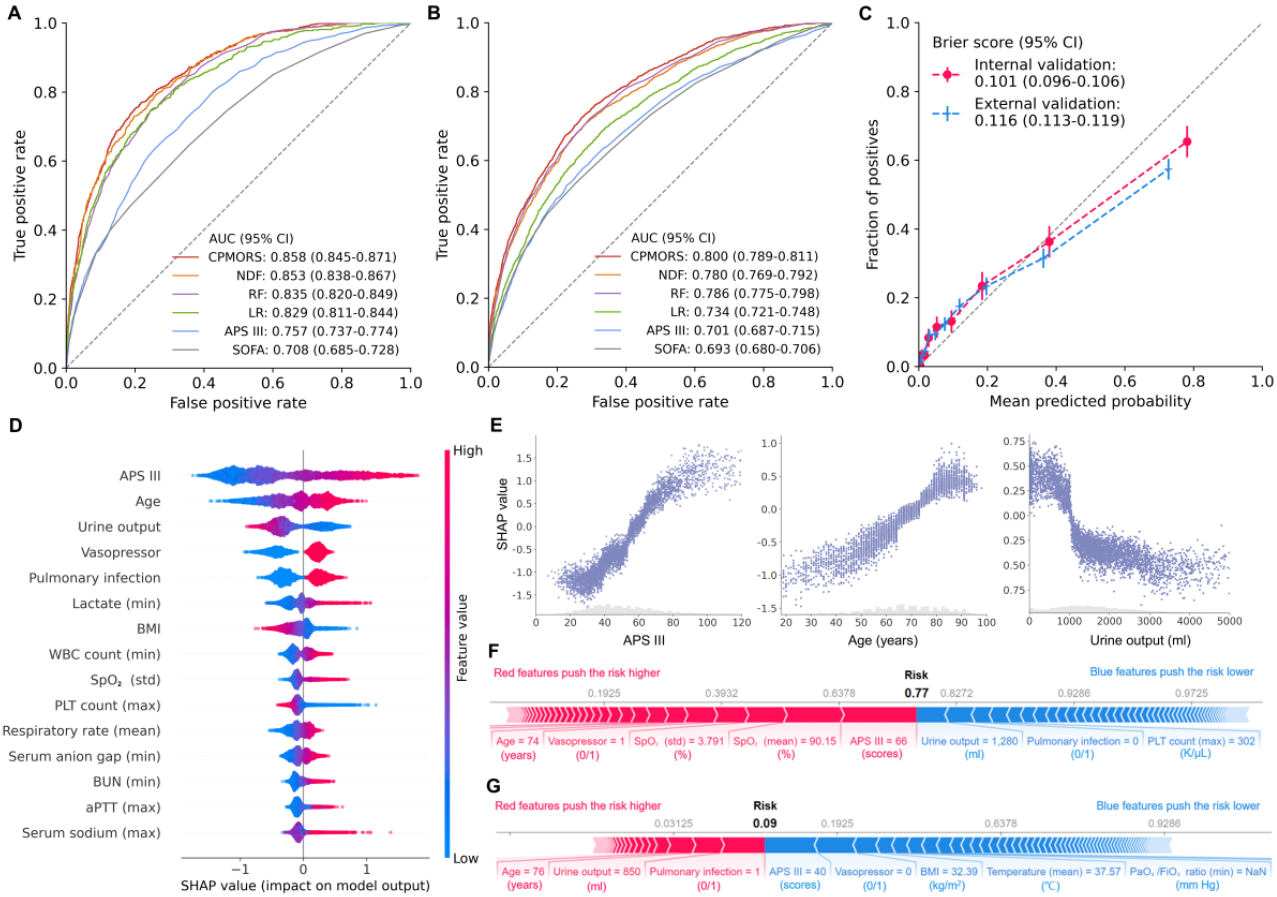

The top 15 clinical features that contributed to the prediction of sepsis mortality risk are summarized in Figure 3D. The interpretable summary of the impact of the features across patients showed that a higher APS III score, older age, oliguria, the use of vasopressor, pulmonary infection, a higher lactate, a lower BMI, a higher white blood cell count, an unstable oxygen saturation, and a lower platelet count were associated with a higher mortality risk, which are consistent with the previous statistical analysis. Figure 3E shows examples of the relationship between SHAP value and feature value. Explanations representing the effects of interpretable sets of extracted features for an individual nonsurvivor and survivor are shown in Figure 3F and G. These effects explained why the model predicted a particular risk, allowing a clinician to plan appropriate interventions.

Results for CP

Calibration curves of the observed prediction error at a significance (1-confidence) level between 0% and 100% showed that CPMORS was well calibrated for the internal MIMIC-IV validation when the populations were from the same hospital system (Figure 4). The calibration curve for external eICU-CRD populations deviated from the ideal diagonal line but was good when the significance level was below 20%. When the conformal predictor was set at a prespecified confidence level of 90% to provide valid predictions, for the MIMIC-IV validation cohort, there were 2520 (efficiency=60.2%) single correct predictions and 1229 (29.4%) multiple predictions (Table 4). The overall validity, the percentage of error predictions (n=438) out of all predictions (n=4187), was 10.5%. In contrast, the AI model without CP made 1449 (34.6%) errors in predicting sepsis mortality risk. When the model was externally validated on a large multicenter database, that is, eICU-CRD, more multiple predictions (n=4004, 38.6%) were flagged for clinician review. In this case, the conformal predictor still produced significantly lower error rates compared to the point predictions by AI (n=1221, 11.8% vs n=4540, 43.8%). No empty predictions were made in either validation cohort.

XSL•FO
RenderX

Figure 4. Plots of conformal prediction for the 2 validation data sets. (A,B) Calibration curves of observed prediction error (the fraction of true labels not included in the prediction region) for different prespecified significance (1-confidence) levels. (C,D) Label distribution plots at different prespecified significance levels, with the model incorporating more multiple predictions at lower significance (higher confidence) levels. The white dotted line indicates the corresponding significance level that produces the highest number of single-label predictions. Single predictions output {survival} or {nonsurvival}, multiple predictions output {survival, nonsurvival}, empty means {null}. eICU-CRD: eICU Collaborative Research Database; MIMIC-IV: Medical Information Mart for Intensive Care database-IV.

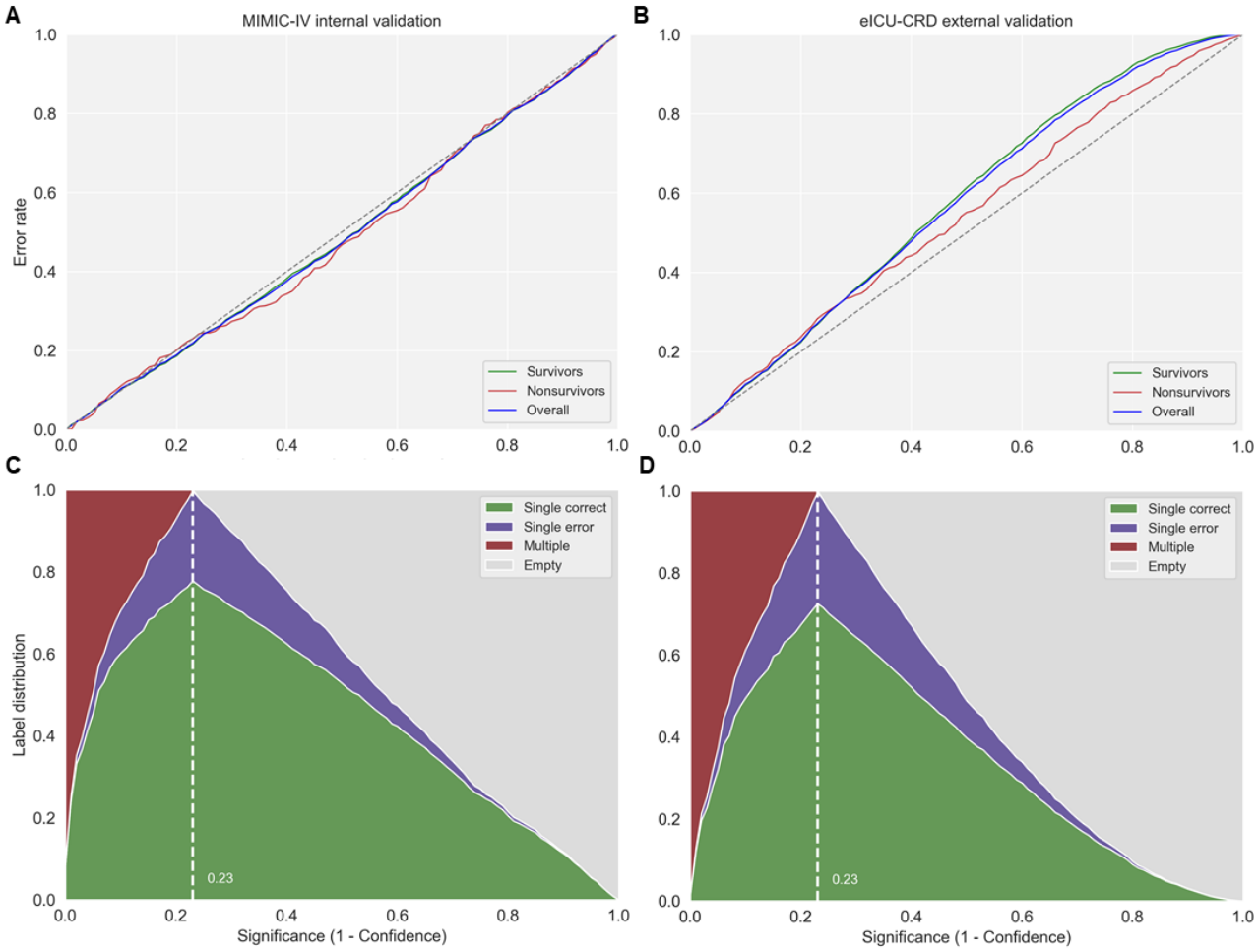

XSL•FO
RenderX

Table 4. Prediction regions on the internal validation and external validation populations.

Confidence level	Prediction regions	MIMIC-IV[a] internal validation			eICU-CRD[b] external validation		
		Survivors (n=3496), n (%)	Nonsurvivors (n=691), n (%)	Overall (n=4187), n (%)	Survivors (n=8605), n (%)	Nonsurvivors (n=1757), n (%)	Overall (n=10,362), n (%)
99.9%							
	Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Single error	43 (1.2)	1 (0.1)	44 (1.1)	81 (0.9)	13 (0.7)	94 (0.9)
	Single correct	868 (24.8)	145 (21)	1013 (24.2)	999 (11.6)	263 (15)	1262 (12.2)
	Multiple	2585 (73.9)	545 (78.9)	3130 (74.8)	7525 (87.4)	1481 (84.3)	9006 (86.9)
95%							
	Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Single error	182 (5.2)	28 (4.1)	210 (5)	461 (5.4)	80 (4.6)	541 (5.2)
	Single correct	1598 (45.7)	306 (44.3)	1904 (45.5)	2704 (31.4)	616 (35.1)	3320 (32)
	Multiple	1716 (49.1)	357 (51.7)	2073 (49.5)	5440 (63.2)	1061 (60.4)	6501 (62.7)
90%							
	Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Single error	361 (10.3)	77 (11.1)	438 (10.5)	997 (11.6)	224 (12.7)	1221 (11.8)
	Single correct	2123 (60.7)	397 (57.5)	2520 (60.2)	4285 (49.8)	852 (48.5)	5137 (49.6)
	Multiple	1012 (28.9)	217 (31.4)	1229 (29.4)	3323 (38.6)	681 (38.8)	4004 (38.6)
85%							
	Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Single error	502 (14.4)	109 (15.8)	611 (14.6)	1454 (16.9)	322 (18.3)	1776 (17.1)
	Single correct	2390 (68.4)	470 (68)	2860 (68.3)	5171 (60.1)	1018 (57.9)	6189 (59.7)
	Multiple	604 (17.3)	112 (16.2)	716 (17.1)	1980 (23)	417 (23.7)	2397 (23.1)
AI[c] point prediction[d]							
	Error	1372 (39.2)	77 (11.1)	1449 (34.6)	4316 (50.2)	224 (12.7)	4540 (43.8)
	Correct	2124 (60.8)	614 (88.9)	2738 (65.4)	4289 (49.8)	1533 (87.3)	5822 (56.2)

[a]MIMIC-IV: Medical Information Mart for Intensive Care database-IV.

[b]eICU-CRD: eICU Collaborative Research Database.

[c]AI: artificial intelligence.

[d]For comparison with the AI model with conformal prediction at a user-specified 90% confidence level, the AI point prediction without conformal prediction was made when setting the error rate in nonsurvivors as 11.1% for MIMIC-IV validation and 12.7% for eICU-CRD validation.

The top 15 features summarized by the SHAP values could not be well differentiated between the multiple predictions in survivors and nonsurvivors (Table 5 and Multimedia Appendix 2). In addition, for those patients who survived, the analysis showed that patients identified for clinician review (those with uncertain predictions) were more likely to develop new acute kidney injury after the first day of ICU admission (n=227, 22.4% vs n=428, 20.2% for MIMIC-IV; n=799, 24% vs n=875, 20.4% for eICU-CRD) and had longer ICU stays with a median of 89 (IQR 54-173) hours versus 55 (IQR 37-99) hours for MIMIC-IV, and 78 (IQR 48-141) hours versus 55 (IQR 40-91) hours for eICU-CRD, compared to the patients who were not flagged for review.

XSL•FO
RenderX

Table 5. Statistical analysis of the top 15 features between the single correct prediction patients and the multiple prediction patients in the MIMIC-IV[a] validation populations.

Variables	Survivors			Nonsurvivors			
	Single correct (n=2123)	Multiple (n=1012)	*P* value	Single correct (n=397)	Multiple (n=217)	*P* value	*P* value[b]
APS III[c] (score), median (IQR)	39 (31-49)	55 (44-66)	<.001	78 (64-94)	58 (47-68)	<.001	.05
Age (years), median (IQR)	64 (54-74)	72 (61-82)	<.001	71 (59-82)	75 (60-81)	.40	.57
Urine output (mL), median (IQR)	1900 (1310-2675)	1350 (808-2065)	<.001	635 (206-1171)	1278 (845-1815)	<.001	.31
Vasopressor, n (%)	959 (45.2)	520 (51.4)	.01	347 (87.4)	134 (61.8)	<.001	.005
Pulmonary infection, n (%)	585 (27.6)	603 (59.6)	<.001	230 (57.9)	127 (58.5)	.89	.77
Lactate (min) (mmol/L), median (IQR)	1.2 (1.0-1.6)	1.4 (1.1-1.9)	<.001	2.2 (1.5-3.4)	1.4 (1.1-1.9)	<.001	.72
BMI (kg/m^2), median (IQR)	27.8 (24.5-32.3)	26.5 (23.0-30.8)	<.001	26.8 (23.5-32.3)	25.9 (23.2-31.4)	.57	.61
WBC[d] count (min) (K/μL), median (IQR)	9.2 (6.6-12.1)	9.9 (6.8-14.0)	<.001	11.7 (7.0-16.4)	10.4 (6.6-14.4)	.06	.50
SpO_2[e] (std) (%), median (IQR)	1.86 (1.36-2.39)	1.93 (1.34-2.55)	.08	2.45 (1.71-3.46)	2.12 (1.43-2.89)	<.001	.01
PLT[f] count (max) (K/μL), median (IQR)	204 (153-269)	213 (147-286)	.11	198 (128-298)	205 (137-279)	.91	.20
Respiratory rate (mean) (bpm), median (IQR)	18 (16-21)	20 (17-22)	<.001	22 (19-25)	20 (17-23)	<.001	.22
Serum anion gap (min) (mmol/L), median (IQR)	12 (10-14)	13 (11-16)	<.001	15 (13-19)	13 (11-16)	<.001	.34
BUN[g] (min) (mg/dL), median (IQR)	16 (12-24)	25 (16-41)	<.001	33 (19-50)	27 (17-42)	.006	.43
aPTT[h] (max) (s), median (IQR)	32.3 (28.3-39.2)	33.8 (29.0-44.5)	<.001	44.9 (33.1-64.3)	34.4 (28.4-47.1)	<.001	.83
Serum sodium (max) (mmol/L), median (IQR)	140 (138-142)	140 (137-144)	.28	140 (136-144)	140 (136-143)	.98	.54

[a]MIMIC-IV: Medical Information Mart for Intensive Care database-IV.

[b]Statistical analysis between survival and nonsurvival multiple predictions.

[c]APS III: Acute Physiology Score III.

[d]WBC: white blood cell.

[e]SpO_2: oxygen saturation.

[f]PLT: platelet.

[g]BUN: blood urea nitrogen.

[h]aPTT: activated partial thromboplastin time.

Discussion

Principal Findings

This study proposed CPMORS, an AI model to provide reliable predictions of mortality risk for patients with sepsis admitted to the ICU. CPMORS was developed using gradient boosting machines with 103 clinical features. Internal and external validation assessed its best predictive performance in terms of discrimination and calibration compared with several other commonly used models. Global feature importance showed that a higher APS III score, older age, oliguria, the use of vasopressor, and pulmonary infection were most associated with a higher mortality risk. A Mondrian CP was built on top of the prediction algorithm to detect uncertain predictions. Compared to traditional AI models that only provide point predictions, our method could provide reliable and uncertain predictions with user-specified confidence, especially when the performance of the external validation could not be guaranteed as good as the internal validation. Interpretable information and estimates of prediction uncertainty enabled CPMORS to provide informative support for clinical decision-making.

Recent guidelines on AI-based predictive models in health care emphasize the importance of AI systems conveying their prediction confidence to users while furnishing accurate predictions and explanations [26]. In previous studies, Zhang et al [27] developed a traditional sepsis mortality risk score system that is interpretable and transparent, but the accuracy of the model is limited, with AUC in the development and validation sets of 0.789 and 0.765. Park et al [6] and Kong et al [7] demonstrated that machine learning can improve the

XSL•FO
RenderX

accuracy of sepsis mortality risk prediction, but there is a lack of strong interpretable analysis. Hu et al [8] performed a detailed and interpretable analysis. However, to date, no studies have estimated the prediction uncertainty of sepsis mortality risk in patients admitted to the ICU. In the case of physicians supported by AI systems, such feedback of prediction uncertainty was invaluable in enabling them to exercise caution and not rely solely on the output of the AI system. This could serve to safeguard patients from the potential hazards of automation bias in AI systems [28]. Various methods for quantifying uncertainty, such as the Gaussian process [29], Bayesian inference [30], deep ensembles [31], and dropout [32], have been implemented in computer vision and natural language processing applications. However, their use in clinical decision-making has been limited due to the accompanying computational cost and modeling complexity. In addition, they did not provide adjustable confidence levels to suit different clinical requirements for predicting critical illness.

In this study, we used the CP framework, which is characterized by its light mathematical nature and offers a potential solution by generating prediction regions. These prediction regions are similar to the confidence intervals used in statistics, except that they are based on individual predictions rather than overall statistics [14]. The use of CP-derived prediction regions could therefore provide reliable and uncertain predictions in clinical settings. In addition, clinicians would not rely entirely on the model's output and make immediate decisions [28]. They still need to review patients' critical information before making decisions. Throughout this process, CPMORS could not only provide explainable risk factors so that clinicians can double-check the reasoning of AI systems and identify problematic elements [9] but could also help to achieve stratified patient management (Figure 1). The single predictions of nonsurvival made by CPMORS could be treated as red flags, meaning that clinicians should pay the most attention. The single survival predictions could be treated as green flags, meaning that the patients are relatively safe and would be monitored regularly. For the multiple predictions, we flag the yellow warnings as such unreliable cases require clinician review; it is clinically very important that AI-based decision support systems also flag uncertainty when they are not certain about the produced output [13]. In this case, CPMORS could produce much fewer false alarms to avoid exposing patients to unnecessary tests and treatments [33]. The process also enables a symbiotic interaction between AI and clinicians, which is the key to AI adoption [34]. The AI system expertly identifies cases where its predictions are highly reliable, allowing clinicians to focus their attention on challenging and less reliable predictions. Therefore, this process would not delay critical decisions. Instead, the proposed AI model could provide explainable information and also help take the pressure off clinicians by stratifying patients. Furthermore, survivors whose predictions were uncertain (multiple predictions) were more likely to develop acute kidney injury and had a longer ICU stay, which proves the validity of their clinician review. In this case, a real-time compatible software pipeline can be developed into clinical decision systems (Figure 1) by fulfilling the requirements stated in de Hond et al [26]. This includes providing an explanation of the model output, enabling end users to visually comprehend the connection between the input data and the predicted output, fostering feedback, and improvement of the predictions.

A systematic review of AI in critical care concluded that it is important to ensure the validity and reliability of predictive models across clinical settings [35]. However, it has been reported that only 6% of studies of AI applications in health care have performed external validation [36], and the sepsis mortality risk prediction model is no exception. Furthermore, previous research on external validation shows a significant decrease when the AI system is applied in different hospital settings or data sets due to several potential threats, such as population heterogeneity, differences in clinical practice, and software diversity [11,37]. Consistent with these studies, the prediction results in this study reported that all AI models deteriorated when externally validated on eICU-CRD, with the AUC of CPMORS decreasing from 0.858 to 0.800. Although transfer learning (TL) methods have been proposed to improve the model generalization ability, additional requirements are needed. For example, additional labeled target data are required for supervised TL (fine-tuning) [38] or many unlabeled data for unsupervised TL (domain adaptation) [39]. Although model performance can be improved, uncertainty estimation is not guaranteed. In this study, the proposed CPMORS did not require target data, could provide confidence measures, and was mathematically proven to be valid, which were important for application across hospital settings. For a given confidence level, the conformal predictor provided a prediction interval within which the true value should lie with a probability of the given confidence, that is, the true class is in the prediction set. Therefore, CPMORS did not aim to directly address the problem of insufficient model generalization performance by simply improving the model's accuracy, as it is impossible for the model to achieve 100% accuracy. CPMORS could help to mitigate the risk of the model generalizability issue from another perspective, by flagging those patients whose predictions are uncertain due to dissimilarity to the training samples and reminding clinicians to check. In addition, CP can also help to detect systematic differences, as shown in Olsson et al [15] for the diagnosis of prostate cancer.

However, we should carefully consider the trade-off between producing single predictions and multiple predictions, as the number of uncertain predictions should be kept limited to avoid overloading clinicians and creating an unmanageable situation. We can see that the proportions of single predictions, empty predictions, and multiple predictions can vary for different levels of significance (1-confidence; Figure 4). To increase confidence in the estimates, a low level of significance must be set, but this would generate an increase in the number of multiple predictions (uncertain predictions). The findings of this study indicate that, when the prespecified confidence level is set to 90%, the CPMORS provides a prediction region around the point prediction that contains the true label with a 90% probability. Our results show that, for the MIMIC-IV and eICU-CRD data sets, CPMORS generated a total of 217 (31.4%) and 681 (38.8%) multiple predictions for nonsurvivors, respectively, and 1012 (28.9%) and 3323 (38.6%) multiple predictions for survivors, respectively (Table 4). The implication of the multiple

XSL•FO
RenderX

predictions in this study is that there is insufficient information for the model to discriminate between survival and nonsurvival outcomes. This is demonstrated by the inability to effectively discriminate the top 15 features between the multiple predictions of survivors and nonsurvivors (Table 5 and Multimedia Appendix 2), which reminds us that new information from patients should be provided to arrive at a single correct prediction.

Limitations

This study has several limitations. First, although we have demonstrated this model can help to mitigate the impact of insufficient model generalization performance and automation bias by providing the uncertainty estimation, we still need further studies to collect weekly or monthly statistics of uncertainty rates so that we can set a balanced confidence level for more efficient patient management. Second, more advanced nonconformal functions should be tested to achieve smaller prediction regions and higher proportions of reliable single predictions. Third, this work was not designed to directly address generalizability and bias issues, although it touched upon them. Future work is needed to solve these 2 issues more thoroughly. Nevertheless, we provided an experimental example to describe how a combination of model interpretability and CP could work to assist clinicians in predicting sepsis mortality in ICU admissions. Conformal predictors are built on top of the underlying prediction algorithm; therefore, the framework can be applied to all prediction algorithms and other predictive tasks in critical illness prediction.

Conclusions

In summary, this study presents the development and validation of an AI model, called CPMORS, for predicting sepsis mortality risk in patients who are critically ill. CPMORS emerges as the most effective model among all the predictive models tested in this study. Importantly, CPMORS offers interpretability, allowing for transparency in the prediction-making process. The internal and external validation procedures demonstrate the ability of CPMORS to make reliable predictions and flag uncertain predictions. These findings suggest that the integration of model explanation and CP can enhance the clinical applicability of AI-assisted systems, thereby facilitating clinical decision-making in the context of sepsis management.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (62171123, 62211530112, and 62071241), the National Key Research and Development Program of China (2023YFC3603600 and 2022YFC2405600), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0088).

Data Availability

The data sets and code generated and analyzed during this study are available on the PhysioNet website [40,41] and GitHub repository [42], respectively.

Authors' Contributions

The study conception and design were developed by MY, XL, JL, and CL. Material preparation, data collection, and analysis were performed by MY, HC, and WH. The predictive model was designed by MY and CL. The first draft of the manuscript was written by MY. MM, CS, and all authors commented on previous versions of the manuscript. JL provided clinical resources. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Statistical analysis of variables between Medical Information Mart for Intensive Care database-IV and eICU Collaborative Research Database.
[DOCX File , 29 KB - jmir_v26i1e50369_app1.docx]

Multimedia Appendix 2
Statistical analysis of the top 15 features between the single correct prediction patients and the multiple prediction patients in the eICU Collaborative Research Database validation populations.
[DOCX File , 23 KB - jmir_v26i1e50369_app2.docx]

References

1. Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, et al. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. Am J Respir Crit Care Med 2016;193(3):259-272 [FREE Full text] [doi: 10.1164/rccm.201504-0781OC] [Medline: 26414292]

XSL•FO
RenderX

2. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
3. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study. Lancet 2020;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]
4. Xie J, Wang H, Kang Y, Zhou L, Liu Z, Qin B, et al. The epidemiology of sepsis in Chinese ICUs: a national cross-sectional survey. Crit Care Med 2020;48(3):e209-e218. [doi: 10.1097/CCM.0000000000004155] [Medline: 31804299]
5. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. N Engl J Med 2020;383(20):1951-1960 [FREE Full text] [doi: 10.1056/NEJMsa2001090] [Medline: 33176085]
6. Park JY, Hsu TC, Hu J, Chen CY, Hsu WT, Lee M, et al. Predicting sepsis mortality in a population-based national database: machine learning approach. J Med Internet Res 2022;24(4):e29982 [FREE Full text] [doi: 10.2196/29982] [Medline: 35416785]
7. Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med Inform Decis Mak 2020;20(1):251 [FREE Full text] [doi: 10.1186/s12911-020-01271-2] [Medline: 33008381]
8. Hu C, Li L, Huang W, Wu T, Xu Q, Liu J, et al. Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. Infect Dis Ther 2022;11(3):1117-1132 [FREE Full text] [doi: 10.1007/s40121-022-00628-6] [Medline: 35399146]
9. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
11. Yang M, Liu C, Wang X, Li Y, Gao H, Liu X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med 2020;48(11):e1091-e1096. [doi: 10.1097/CCM.0000000000004550] [Medline: 32885937]
12. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
13. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say "i don't know". NPJ Digit Med 2021;4(1):134 [FREE Full text] [doi: 10.1038/s41746-021-00504-6] [Medline: 34504260]
14. Alvarsson J, McShane SA, Norinder U, Spjuth O. Predicting with confidence: using conformal prediction in drug discovery. J Pharm Sci 2021;110(1):42-49 [FREE Full text] [doi: 10.1016/j.xphs.2020.09.055] [Medline: 33075380]
15. Olsson H, Kartasalo K, Mulliqi N, Capuccini M, Ruusuvuori P, Samaratunga H, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nat Commun 2022;13(1):7761 [FREE Full text] [doi: 10.1038/s41467-022-34945-8] [Medline: 36522311]
16. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
17. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
18. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34(5):1297-1310. [doi: 10.1097/01.CCM.0000215112.84523.F0] [Medline: 16540951]
19. Chen T, Guestrin CE. XGBoost: A Scalable Tree Boosting System. In: ACM Digital Library.: Association for Computing Machinery; 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794 URL: https://dl.acm.org/doi/proceedings/10.1145/2939672 [doi: 10.1145/2939672.2939785]
20. Kontschieder P, Fiterau M, Criminisi A, Bulò SR. Deep Neural Decision Forests. In: Computer Vision Foundation. 2015 Presented at: IEEE International Conference on Computer Vision (ICCV); December 07-13, 2015; Santiago, Chile URL: https://ieeexplore.ieee.org/document/7410529 [doi: 10.1109/iccv.2015.172]
21. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform 2019;125:55-61. [doi: 10.1016/j.ijmedinf.2019.02.002] [Medline: 30914181]
22. Toccaceli P, Gammerman A. Combination of inductive mondrian conformal predictors. Mach Learn 2018;108(3):489-510. [doi: 10.1007/s10994-018-5754-9]
23. Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. Taking the human out of the loop: a review of Bayesian optimization. Proc IEEE 2016;104(1):148-175 [FREE Full text] [doi: 10.1109/jproc.2015.2494218]
24. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]

XSL•FO
RenderX

25. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
26. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
27. Zhang K, Zhang S, Cui W, Hong Y, Zhang G, Zhang Z. Development and validation of a sepsis mortality risk score for sepsis-3 patients in intensive care unit. Front Med (Lausanne) 2020;7:609769 [FREE Full text] [doi: 10.3389/fmed.2020.609769] [Medline: 33553206]
28. Volovici V, Syn NL, Ercole A, Zhao JJ, Liu N. Steps to avoid overuse and misuse of machine learning in clinical research. Nat Med 2022;28(10):1996-1999 [FREE Full text] [doi: 10.1038/s41591-022-01961-6] [Medline: 36097217]
29. Lederer A, Umlauft J, Hirche S. Uniform error bounds for Gaussian process regression with application to safe control. ArXiv Preprint posted online Dec 19, 2019 [FREE Full text] [doi: 10.48550/arXiv.1906.01376]
30. Ghosh S, Yao J, Doshi-Velez F. Structured variational learning of Bayesian neural networks with horseshoe priors. ArXiv Preprint posted online Jul 31, 2018 [FREE Full text] [doi: 10.48550/arXiv.1806.05975]
31. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. ArXiv Preprint posted online Nov 4, 2017 [FREE Full text] [doi: 10.48550/arXiv.1612.01474]
32. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. PMLR 2016;48:1050-1059 [FREE Full text] [doi: 10.48550/arXiv.1506.02142]
33. Mitka M. Joint commission warns of alarm fatigue: multitude of alarms from monitoring devices problematic. JAMA 2013;309(22):2315-2316. [doi: 10.1001/jama.2013.6032] [Medline: 23757063]
34. Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit Med 2022;5(1):97 [FREE Full text] [doi: 10.1038/s41746-022-00597-7] [Medline: 35864312]
35. Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. Crit Care 2022;26(1):75 [FREE Full text] [doi: 10.1186/s13054-022-03915-3] [Medline: 35337366]
36. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019;20(3):405-410 [FREE Full text] [doi: 10.3348/kjr.2019.0025] [Medline: 30799571]
37. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 2020;11(1):5668 [FREE Full text] [doi: 10.1038/s41467-020-19551-w] [Medline: 33168827]
38. Holder AL, Shashikumar SP, Wardi G, Buchman TG, Nemati S. A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the ICU. Crit Care Med 2021;49(12):e1196-e1205 [FREE Full text] [doi: 10.1097/CCM.0000000000005175] [Medline: 34259450]
39. Zhu Y, Venugopalan J, Zhang Z, Chanani NK, Maher KO, Wang MD. Domain adaptation using convolutional autoencoder and gradient boosting for adverse events prediction in the intensive care unit. Front Artif Intell 2022;5:640926 [FREE Full text] [doi: 10.3389/frai.2022.640926] [Medline: 35481281]
40. MIMIC-IV data. PhysioNet. URL: https://physionet.org/content/mimiciv/2.2/ [accessed 2023-04-28]
41. eICU-CRD Data. MIT Laboratory. URL: http://eicu-crd.mit.edu/ [accessed 2023-04-28]
42. CPMORS (Conformal Predictor for Mortality Risk in Sepsis). GitHub. URL: https://github.com/Meicheng-SEU/CPMORS [accessed 2023-10-15]

Abbreviations

AI: artificial intelligence
APS III: Acute Physiology Score III
AUC: area under the receiver operating characteristic curve
CPMORS: Conformal Predictor for Mortality Risk in Sepsis
CP: conformal prediction
eICU-CRD: eICU Collaborative Research Database
ICU: intensive care unit
MIMIC-IV: Medical Information Mart for Intensive Care database-IV
SOFA: Sequential Organ Failure Assessment
SHAP: Shapley additive explanation
TL: transfer learning

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 29.06.23; peer-reviewed by Z Xu, E Kawamoto; comments to author 12.10.23; revised version received 16.10.23; accepted 24.01.24; published 18.03.24.

Please cite as:
Yang M, Chen H, Hu W, Mischi M, Shan C, Li J, Long X, Liu C
Development and Validation of an Interpretable Conformal Predictor to Predict Sepsis Mortality Risk: Retrospective Cohort Study
J Med Internet Res 2024;26:e50369
URL: https://www.jmir.org/2024/1/e50369
doi:10.2196/50369
PMID:38498038

©Meicheng Yang, Hui Chen, Wenhan Hu, Massimo Mischi, Caifeng Shan, Jianqing Li, Xi Long, Chengyu Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Making Science Computable Using Evidence-Based Medicine on Fast Healthcare Interoperability Resources: Standards Development Project

Andrey Soares[1], PhD; Lisa M Schilling[1], MD, MSPH; Joshua Richardson[2], MS, MLiS, PhD; Bhagvan Kommadi[3,4], MS; Vignesh Subbian[4,5], PhD; Joanne Dehnbostel[4,6], MS, MPH; Khalid Shahin[4,6], BA; Karen A Robinson[4,7], PhD; Muhammad Afzal[4,8], PhD; Harold P Lehmann[4,7], MD, PhD; Ilkka Kunnamo[4,9], MD, PhD; Brian S Alper[4,6], MD, MSPH

[1]Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[2]Center for Informatics, Research Triangle Institute International, Berkeley, CA, United States
[3]Quantica Computacao, Hyderabad, India
[4]Scientific Knowledge Accelerator Foundation, Franklin, NC, United States
[5]College of Public Health, Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, United States
[6]Computable Publishing LLC, Franklin, NC, United States
[7]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States
[8]Department of Computing and Data Science, Birmingham City University, England, United Kingdom
[9]Duodecim Publishing Company Ltd, Helsinki, Finland

Corresponding Author:
Andrey Soares, PhD
Department of Medicine
University of Colorado Anschutz Medical Campus
1890 North Revere Court
Mailstop F443
Aurora, CO, 80045
United States
Phone: 1 3037242825
Email: andrey.soares@cuanschutz.edu

Abstract

Background: Evidence-based medicine (EBM) has the potential to improve health outcomes, but EBM has not been widely integrated into the systems used for research or clinical decision-making. There has not been a scalable and reusable computer-readable standard for distributing research results and synthesized evidence among creators, implementers, and the ultimate users of that evidence. Evidence that is more rapidly updated, synthesized, disseminated, and implemented would improve both the delivery of EBM and evidence-based health care policy.

Objective: This study aimed to introduce the EBM on Fast Healthcare Interoperability Resources (FHIR) project (EBMonFHIR), which is extending the methods and infrastructure of Health Level Seven (HL7) FHIR to provide an interoperability standard for the electronic exchange of health-related scientific knowledge.

Methods: As an ongoing process, the project creates and refines FHIR resources to represent evidence from clinical studies and syntheses of those studies and develops tools to assist with the creation and visualization of FHIR resources.

Results: The EBMonFHIR project created FHIR resources (ie, *ArtifactAssessment*, *Citation*, *Evidence*, *EvidenceReport*, and *EvidenceVariable*) for representing evidence. The COVID-19 Knowledge Accelerator (COKA) project, now Health Evidence Knowledge Accelerator (HEvKA), took this work further and created FHIR resources that express *EvidenceReport*, *Citation*, and *ArtifactAssessment* concepts. The group is (1) continually refining FHIR resources to support the representation of EBM; (2) developing controlled terminology related to EBM (ie, study design, statistic type, statistical model, and risk of bias); and (3) developing tools to facilitate the visualization and data entry of EBM information into FHIR resources, including human-readable interfaces and JSON viewers.

XSL•FO
RenderX

Conclusions: EBMonFHIR resources in conjunction with other FHIR resources can support relaying EBM components in a manner that is interoperable and consumable by downstream tools and health information technology systems to support the users of evidence.

(J Med Internet Res 2024;26:e54265) doi:10.2196/54265

KEYWORDS

evidence-based medicine; FHIR; Fast Healthcare Interoperability Resources; computable evidence; EBMonFHIR; evidence-based medicine on Fast Healthcare Interoperability Resources

Introduction

Background and Significance

Timely and relevant biomedical evidence is essential to provide high-quality health care. Decision makers rely on "synthesized evidence" from systematic reviews (SRs) and clinical practice guidelines (CPGs) to inform clinical care decisions and policies at multiple levels [1-7]. Therefore, actionable evidence is key for optimizing health care delivery and outcomes. Yet, with an estimated 75 clinical trials and 11 SRs published every day [8], efforts to synthesize, disseminate, and implement biomedical evidence throughout the evidence ecosystem to inform decision-making are unsustainable [9].

Evidence-based medicine (EBM) is increasingly recognized as critical in the decision-making process related to patient care and policy development. EBM is "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [10]. The practice of EBM depends on comprehensive and up-to-date synthesized research evidence, which requires continuous updating and reconciliation of new scientific results with previous results. For example, early trial results on prenatal steroids for preterm births were initially inconclusive, but evidence syntheses justified steroidal therapy as the best practice [11-14]. However, EBM practice is laborious, methodologically complex, and often error prone.

Multiple barriers hinder the effective and efficient use of evidence syntheses. First, the time and effort required to synthesize evidence often leads to SRs and CPGs being out of date by the time of their publication [15,16]. Even when synthesized evidence is available in digital form, it is typically in narrative form and accessed via a bibliographical database that requires proactive searching (eg, PubMed or a publisher), an email newsletter (eg, *Journal of the American Medical Association Internal Medicine Newsletter*), or social media notification (eg, X, formerly known as Twitter, or Doximity). These "pull" methods of synthesized evidence are helpful means of dissemination, but they do not facilitate more efficient "push" methods that promote use, implementation, and action at scale. Second, the common mode of evidence dissemination, which is human-readable text, is not standardized in ways that support technical solutions to scalable dissemination and implementation [15-19]. Despite notable past efforts for structuring and disseminating guidance, such as the GuideLine Interchange Format (GLIF) [20,21] and Standards-Based Sharable Active Guideline Environment (SAGE) [22] standards, there are no existing scalable and reusable computer-readable standards for distributing research results and synthesized evidence among creators, implementers, and the ultimate users of that evidence.

The current state is an ecosystem of researchers, informaticians, statisticians, policy makers, epidemiologists, librarians, and other stakeholders in the biomedical research community who synthesize evidence by manually searching for relevant studies, assessing the studies for quality and risk of bias, and compiling the results in labor-intensive ways. Evidence implementers, including CPG and clinical decision support (CDS) creators, whether they are creating or implementing local or third-party tools, must similarly review and verify the evidence before use. The challenges with finding and delivering evidence are amplified by needs due to the COVID-19 pandemic [23-28], where "scientists [have] published well over 100,000 articles about the coronavirus pandemic in 2020" [29] and for which there are currently over 2.5 million publications [30]. Today, these mostly manual, redundant, and disjointed processes seem to be the acknowledged status quo, even though new evidence continues to be generated at a rapid pace. The evidence synthesis ecosystem is, therefore, rife with duplication and uncoordinated efforts [31,32] for identifying, appraising, synthesizing, and disseminating evidence, requiring considerable resources, expertise, and time [33,34]. Evidence that is more rapidly updated, synthesized, disseminated, and implementable would improve both the delivery of EBM and evidence-based health care policy [4,35].

Computable Evidence

Solutions to improve the evidence-to-practice lifecycle [36] should begin with the transformation of study results into "computable evidence" or "knowledge artifacts" that could be consumed by software-based information systems. Key to computable evidence is knowledge representation in machine-interpretable formats that enable findable, accessible, interoperable, and reusable [37] information across the evidence ecosystem. Figure 1 (adapted from the "Digital and Trustworthy Evidence Ecosystem" [38]) depicts this vision where research results at many stages throughout analysis, publication, and synthesis (the EBM on Fast Healthcare Interoperability Resources [EBMonFHIR] area) can be extracted and transformed from numerous disparate sources (such as registry reports, gray literature, preprints, and peer-reviewed literature databases) and stored as machine-readable evidence in an interoperable standard format that could then be used by a wide variety of evidence users and developers from biomedical knowledge bases (ie, evidence repositories) to SRs, CPGs, and CDS systems. Figure 1 also shows where related interoperability solutions [39] applied to CPG and CDS representation (the CPGonFHIR area), such as CDS Connect (a platform with a repository and authoring tool for CDS artifacts) [40] and CDS Hooks (a Health Level Seven [HL7] specification that provides

XSL•FO
RenderX

a way to embed CDS services within the clinician workflow of an electronic health record) [41], can complete the evidence-to-practice implementation within the evidence ecosystem.

Figure 1. Application of Fast Healthcare Interoperability Resources (FHIR) resources to the evidence ecosystem. EBMonFHIR denotes application to results of scientific analysis, synthesis, and publication. CPGonFHIR denotes application to guidance and decision support. CSV, TXT, XLS, PDF, DOC, HTML, JSON, and XML denote file formats. CDMS: clinical data management system; CDS: clinical decision support; CPG: clinical practice guideline; CQL: Clinical Quality Language; EBM: evidence-based medicine; EBMonFHIR: Evidence-Based Medicine on Fast Healthcare Interoperability Resources; EHR: electronic health record; REDCap: Research Electronic Data Capture. Adapted from the "Digital and Trustworthy Evidence Ecosystem" [38].

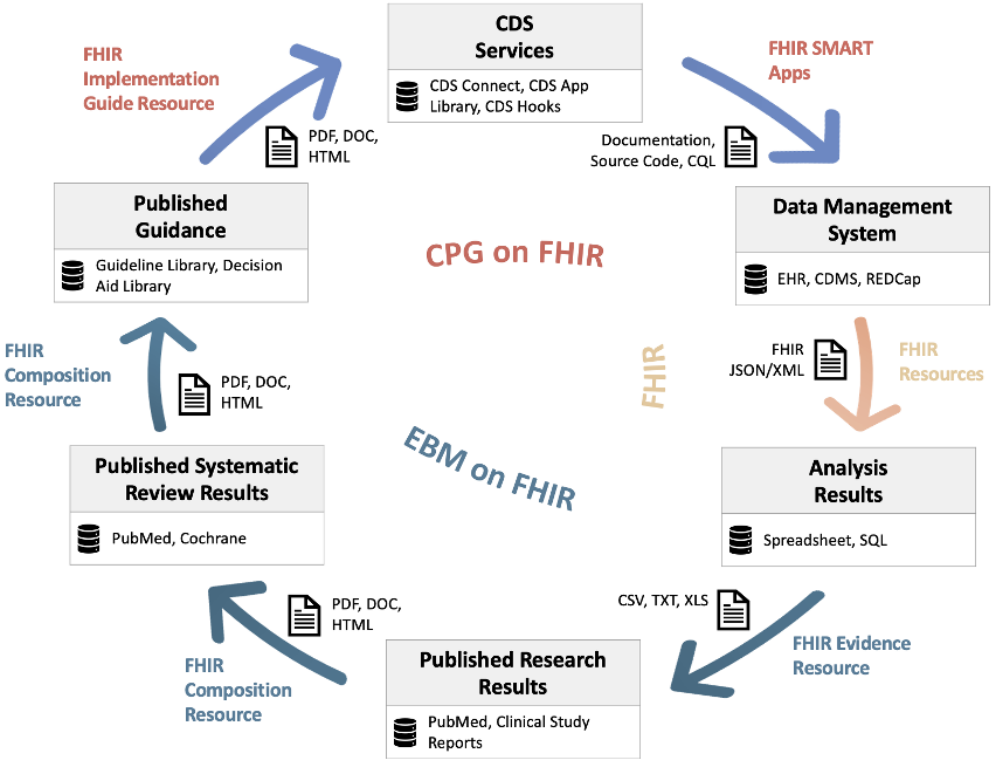

The EBMonFHIR Project

Built upon the success of Fast Healthcare Interoperability Resources (FHIR) [42-44] to promote interoperability and standards for data exchange, we initiated the EBMonFHIR project on May 16, 2018. The aim was to extend the methods and infrastructure of HL7 FHIR to provide an interoperability standard for the electronic exchange of biomedical knowledge from and about clinical research and recommendations [45]. We sought to stand on the shoulders of those who have previously made strides in structuring evidence and standardizing the means to share that evidence (eg, GLIF) by leveraging FHIR's growing popularity and acceptance. The project solicits the participation of known experts and gathers input from broad communities in the evidence ecosystem to determine the data exchange needs for interoperable knowledge assets [45] and defines the FHIR resources related to the domain of clinical research evidence. We have begun showing that FHIR can deliver not just clinical data but also synthesized evidence and knowledge to end users.

Historical Context

The project started with implementers from EBSCO Health, Duodecim Medical Publications Ltd, HarmoniQ, and MAGIC Evidence Ecosystem Foundation and expanded to a multisector project with participation from academia, industry, government, and nonprofit organizations [46].

The COVID-19 pandemic introduced a new impetus in the quest to make biomedical evidence computable and interoperable. In light of the pandemic, where scientists and clinicians urgently needed timely results and evidence, the EBMonFHIR project leaders organized a larger group named the "COVID-19 Knowledge Accelerator (COKA) Initiative" and shifted their focus to the biomedical evidence regarding SARS-CoV-2 [47]. COKA is a nonprofit unincorporated organization with global collaboration to develop and advance interoperability standards for COVID-19 knowledge and to enhance the evidence exchange standards. As of January 2023, the COKA initiative was renamed to Health Evidence Knowledge Accelerator (HEvKA) to better represent the group's now broader scope that is inclusive of clinical, public health, and environmental health domains.

Objectives

In this paper, we introduce the EBMonFHIR project that aims to produce an HL7 FHIR schema to express biomedical evidence as computable evidence—as well as FHIR resource instances, terminologies, and tools—and to promote an effective and efficient evidence ecosystem. We describe the participants involved in this effort, the process to develop the EBMonFHIR standards, and the progress made to represent evidence findings as FHIR resources. We also provide information on initial impacts, limitations, and next steps for stakeholders involved

XSL•FO
RenderX

in standards and controlled terminology development, EBM implementation, and evidence used for both clinical studies and syntheses of those studies. We finally describe how others can get involved in the EBMonFHIR effort by way of the HEvKA initiative.

Methods

HEvKA is advancing EBMonFHIR according to the five aspects of the HL7 standards development process [48], and they are (1) foster consensus, (2) ensure content is fit for purpose, (3) ensure content is implementable, (4) establish an appropriate implementer community, and (5) ensure ongoing maintenance of the standard. Through up to 15 online meetings each week and up to 3 HL7 FHIR Connectathons [49] each year, HEvKA creates and refines FHIR resources to represent evidence from clinical studies and syntheses of those studies and develops tools to assist with the creation and visualization of FHIR resources.

Updates and improvements for any FHIR specification can be developed, proposed, reviewed, improved, voted on, and released within a documented environment in accordance with the American National Standards Institute (ANSI)-sanctioned HL7 ballot process, and ultimately published as part of the official HL7 FHIR specification. [50]

The FHIR resources that HEvKA develops are continuously revised and adapted to reflect the best representation of the knowledge from the community. The HL7 CDS Work Group [51] is responsible for evaluating and approving the additions and changes to these FHIR resources.

To support a culture of transparency and openness about the process of developing the FHIR resources and support tools, HEvKA makes use of numerous digital approaches to document and keep track of its activities and disseminate its progress, including the HL7 Confluence Web Page (content management) [45], Google Drive (content repository) [52], and Microsoft Teams (videoconference) [53]. The content produced is open-source and freely available to the community, with examples published on the Fast Evidence Interoperability Resources (FEvIR) Platform [54]. HEvKA also coordinates input and dissemination across many communities in the evidence ecosystem.

HEvKA has 14 active working groups that meet weekly to address different aspects of the project (ie, Communications; Computable EBM Tools Development; CQL Development; EBM Implementation Guide; Eligibility Criteria; Funding the Ecosystem Infrastructure; GRADE [Grading of Recommendations Assessment, Development and Evaluation] Ontology; Measuring the Rate of Scientific Knowledge Transfer; Ontology Management; Project Management; Risk of Bias Terminology; Setting the Scientific Record on FHIR; Statistic Terminology; and StatisticsOnFHIR). For details, visit the Confluence web page for information about HEvKA [55].

Results

EBM Representation With FHIR Resources

HEvKA first created FHIR resources for representing research results (*Evidence*) and variable definitions (*EvidenceVariable*); and after March 2020, HEvKA further created FHIR resources that express compositions (*EvidenceReport*), citations (*Citation*), and judgments about knowledge (*ArtifactAssessment*). To ease the readability of this article, we will broadly refer to these FHIR resources as the "EBMonFHIR Resources." Table 1 shows the list of EBMonFHIR resources the project developed. From here on, we list FHIR resources, FHIR elements, and data types [56] in *italics*.

An *Evidence* resource [57] provides an expression of the most granular components of evidence. Evidence is often represented by values and parameters of statistical measures (eg, mean, confidence interval, relative risk, and hazard ratio) and expressions of certainty or classifications of these statistical findings. These statistics are about a particular combination of variables from a particular study, so the *Evidence* resource can refer to *EvidenceVariable* resources for definitions of the observed or intended variables. However, to support interoperability across systems, the *Group* resource [58] may be used instead of an *EvidenceVariable* resource [59], especially when referring to a group of people such as a population, sample, or subgroup.

The *Evidence* resource refers to statistical measures and their values with a machine-interpretable expression of a statistic, including the quantity; unit of measure; classification of statistic type; sample size; attribute estimates such as confidence intervals, *P* values, and heterogeneity estimates; and statistical model characteristics (Figure 2). The *statisticType* element has, as of November 2, 2023, a suggested set of 22 possible codes [60], assembled by the HEvKA team, that represent types of statistics (eg, median, relative risk, and incidence rate ratio). HEvKA has drafted 139 terms for statistic types in total (these will eventually replace the Statistic Type Value Set).

Biomedical evidence is comprised of facts and interpretations derived from an analysis of observations of a selective sample. Certainty about any evidence may change due to methodological factors, statistical factors, and contextual factors, and for the relatedness between the sample, the evidence was derived from the population to which the evidence is applied. The *certainty* element provides a machine-interpretable expression of confidence in, or certainty or quality of, the evidence. The *type* subelement can express the aspect of the certainty being rated, using codable concepts from a suggested value set (eg, including the overall certainty, risk of bias, inconsistency, indirectness, imprecision, publication bias, dose-response gradient, plausible confounding, and large effect size [61]).

An *EvidenceVariable* resource [59] provides an expression of a single evidence variable (eg, a single exposure or a single outcome or measured variable).

The *Citation* resource "enables reference to any knowledge artifact for purposes of identification and attribution" [62], including "location, authorship, and contributorship to a journal

XSL•FO
RenderX

article, report, document, resource, or other knowledge artifact" [62]. For instance, the *Citation* resource [63] is a reference to the *Evidence* resource [64], representing the primary outcome from an article reporting the results of a randomized clinical trial [65].

The *EvidenceReport* resource represents a "container for a collection of resources and codable concepts, adapted to support compositions of *Evidence*, *EvidenceVariable*, and *Citation* resources and related concepts" [66]. This resource may bundle knowledge from 1 or multiple studies [66]. The report can be represented in sections of different forms including text, references to codable concepts, or other FHIR resources and sections. The *EvidenceReport* resource is "suited for communicating reports about research and data analysis not specific to individual persons" [66], distinct from the FHIR *Composition* resource commonly used for reports specific to individual persons. However, the *EvidenceReport* resource will be deprecated as the *Composition* resource has been modified to include an EvidenceReport profile to support EBMonFHIR use.

The *ArtifactAssessment* resource "represents one or more assessments of another record or resource" [67]. This resource covers assessments about clinical records, health care provision, and records related to community knowledge (eg, evidence), and may include comments, corrections, classifications, ratings, questions, and responses.

Table 1. List of FHIR[a] resources developed by the EBMonFHIR[b] and HEvKA[c] projects (FHIR version 6.0.0 current build, as of November 2, 2023).

FHIR[a] resource	Description	Reference to URL
ArtifactAssessment	The *ArtifactAssessment* resource provides 1 or more comments, classifiers, or ratings about a resource and supports attribution and rights management metadata for the added content.	[67]
Citation	The *Citation* resource enables reference to any knowledge artifact for purposes of identification and attribution. The *Citation* resource supports existing reference structures and developing publication practices such as versioning, expressing complex contributorship roles, and referencing computable resources.	[62]
Evidence	The *Evidence* resource provides a machine-interpretable expression of an evidence concept including the evidence variables (eg, population, exposures or interventions, comparators, outcomes, measured variables, confounding variables), the statistics, and the certainty of this evidence.	[57]
EvidenceReport	The *EvidenceReport* resource is a specialized container for a collection of resources and codable concepts, adapted to support compositions of *Evidence*, *EvidenceVariable*, and *Citation* resources and related concepts.	[66]
EvidenceVariable	The *EvidenceVariable* resource describes an element that knowledge (*Evidence*) is about.	[59]

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]EBMonFHIR: Evidence-Based Medicine on Fast Healthcare Interoperability Resources.

[c]HEvKA: Health Evidence Knowledge Accelerator.

Figure 2. *Evidence* resource backbone elements are represented with an abbreviated unified modeling language diagram. 0..1 denotes the element is optional and can only have 1 instance. 0..* denotes the element is optional and can have any number of instances. 1..* denotes the element is required and can have any number of instances.

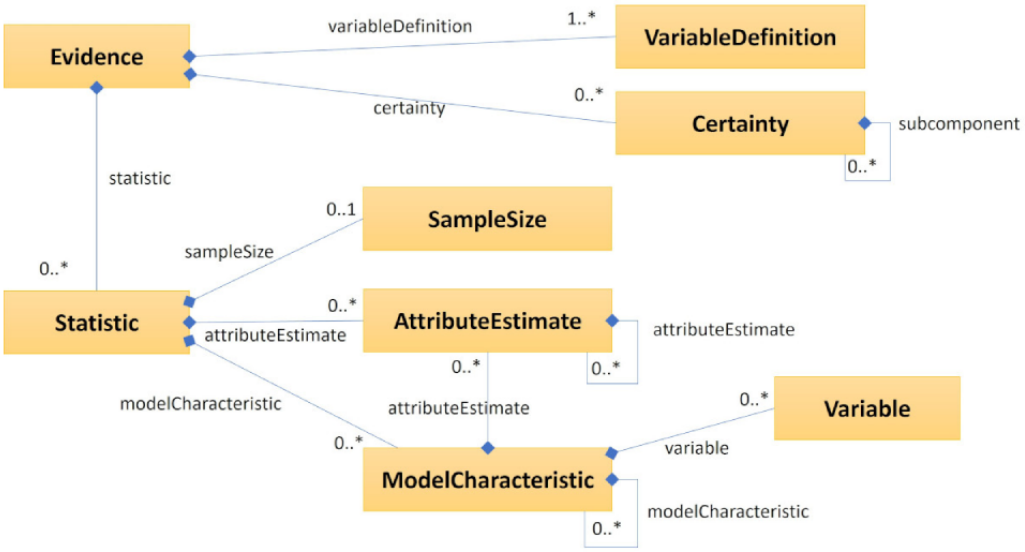

XSL•FO
RenderX

Terminologies

The use of controlled terminologies supports the interoperable representation of multiple concepts. In FHIR, controlled terminologies [68] are represented in code systems and value sets. Reporting a specific term in FHIR uses a *Coding* data type. The *Coding* data type includes a *system* element to identify the terminology system, a *code* element for the precise code for machine use, and a *display* element for human-readable expression. A *CodeableConcept* data type is "a value that is usually supplied by providing a reference to one or more terminologies or ontologies but may also be defined by the provision of text" [56]. A *CodeableConcept* element is composed of a *coding* element (with none, 1, or more instances) and a *text* element (with none or 1 instance).

The *coding* element of a *CodeableConcept* element can use a variety of terminologies, such as Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), RxNORM, and Identification of Medicinal Products (IDMP). FHIR has a noncomprehensive registry of external code systems [68]. New items can be proposed by the community, and implementers can choose to use other code systems not listed in the registry. Figure 3 shows a sample *CodeableConcept* element for the "Disease caused by severe acute respiratory syndrome coronavirus 2."

HEvKA has used, developed, or extended over 40 terminologies for use with *Evidence*, *EvidenceVariable*, *EvidenceReport*, *Citation*, and *ArtifactAssessment* resources. Consolidation of several of these terminologies has resulted in HEvKA leading the effort to create a Scientific Evidence Code System (SEVCO) with nearly 600 terms for study design, risk of bias, and statistics [50]. Multimedia Appendix 1 presents a list of FHIR data elements with reference to controlled terminologies.

Figure 3. Sample *valueCodeableConcept* element showing multiple *coding* elements. In this example, the *coding* element represents the COVID-19 disease with both SNOMED CT and ICD-10 terminologies. ICD-10: International Classification of Diseases, Tenth Revision; SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms.

```
"valueCodeableConcept": {
  "coding": [
  {
     "system": "http://snomed.info/sct",
     "code": "840539006",
     "display": "Disease caused by severe acute respiratory
        syndrome coronavirus 2 (disorder)",
     "userSelected": true
  },
  {
     "system": "http://hl7.org/fhir/sid/icd-10",
     "code": "U07.1",
     "display": "COVID-19, virus identified",
     "userSelected": true
  }
  ],
  "text": "COVID-19"
}
```

Walkthrough: Sample Study Result Represented With EBMonFHIR Resources

HEvKA tested the EBMonFHIR standard by conducting a successful proof-of-concept exercise that used the FHIR *Evidence* resource to represent a critically appraised summary of the primary outcome of a multi-platform randomized controlled trial (RCT) of anticoagulation for hospitalized noncritically ill patients with COVID-19 [69]. The primary outcome was "organ support–free days, evaluated on an ordinal scale that combined in-hospital death and the number of days free of cardiovascular or respiratory organ support up to day 21 among patients who survived to hospital discharge" [69]. The primary result for the overall group was reported as a median adjusted odds ratio of 1.27 (95% credible interval 1.03-1.58), with 939 (80.2%) out of 1171 in the therapeutic dose anticoagulation group and 801 (76.4%) out of 1048 in the usual-care thromboprophylaxis group. An overview of FHIR resources used to represent the appraisal of this outcome is shown in Table 2. See the EBMonFHIR walkthrough [70] of this study with details about how the resources are used to represent the study including samples of JSON code to illustrate the EBM representation with FHIR resources.

XSL•FO
RenderX

Table 2. List of EBMonFHIR[a] resources that represent the result for the primary outcome of the sample study.

Reference	FHIR resource type	Description
[71]	*Citation*	Citation for the article in the NEJM[b]
[63]	*Citation*	Citation for evidence resource with FEvIR[c] Object Identifier (FOI) 7637
[64]	*Evidence*	Summary of 1 unit of evidence (statistical findings for 1 set of variables) from the study
[72]	*Evidence variable*	Exposure in the intervention arm of the study
[73]	*Evidence variable*	Exposure in the comparator arm of the study
[74]	*Evidence variable*	Measured variable for the primary outcome of the study
[75]	*Group*	Group description for the intended population for evidence interpretation
[76]	*Group*	Group description for the observed population in the study

[a]EBMonFHIR: Evidence-Based Medicine on Fast Healthcare Interoperability Resources.

[b]NEJM: New England Journal of Medicine.

[c]FEvIR: Fast Evidence Interoperability Resources.

Discussion

Overview

HEvKA is spearheading transformations and advances of EBMonFHIR on several fronts to support the process of exchanging biomedical evidence in a machine-readable format. As a continuously evolving effort, HEvKA has coordinated efforts and collaborated with industry, academia, government, and nonprofit organizations to develop EBMonFHIR resources and related tools, and has presented progress reports and results of the project to both EBM and informatics communities at several national and international events such as American Medical Informatics Association (AMIA) Annual Symposium, American Medical Informatics Association Informatics Summit, Guidelines International Network (GIN) Conference, Mobilizing Computable Biomedical Knowledge (MCBK), and Cochrane Colloquium.

Status and Maturity of Evidence-Related Standards

The EBMonFHIR standard contains five FHIR resources (ie, *Evidence*, *EvidenceVariable*, *Citation*, *EvidenceReport*, and *ArtifactAssessment* resources) and HEvKA is continually revising and refining the resources with support from members of the EBM community. Currently, the *ArtifactAssessment*, *Citation*, *Evidence*, and *EvidenceVariable* resources are at FHIR Maturity Level 1, "the artifact produces no warnings during the build process and the responsible WG has indicated that they consider the artifact substantially complete and ready for implementation" [77], whereas the *EvidenceReport* resource is at Maturity Level 0 (ie, Draft) and will be deprecated. Dedicated HEvKA working groups are developing functional examples to confirm and demonstrate the use of EBMonFHIR resources and data elements in various use cases. The examples are published on the FEvIR Platform [54]. In addition, following efforts for developing an implementation guide for representing evidence-based CPG recommendations [78], the EBMonFHIR project is producing an EBMonFHIR Implementation Guide [79,80] with 73 profiles of 12 resources to fully represent clinical research and evidence-based guidelines.

The Role of Controlled Terminologies

Communicating evidence in coded and structured forms requires controlled terminologies (or code systems) to uniquely and accurately express essential concepts. HEvKA created a Code System Development protocol to enable standardized terminologies for the exchange of scientific evidence [50]. This protocol is being executed to develop four code sets related to EBM: study design, statistic type, statistical model, and risk of bias.

Relatedly, the SEVCO Expert Working Group (an offshoot of HEvKA) has 39 members from 18 countries as of November 3, 2023. SEVCO has identified 23 commonly used tools and methods for what the code system will support, such as the ROBINS-I tool for risk of bias assessment [81]. As of this writing, there are 602 prospective terms (253 for risk of bias, 76 for study design, and 273 for statistics) to support all recognized commonly used tools and systems. Coordination with the GRADE Working Group, a collaboration to develop a common and transparent approach to grading quality (or certainty) of evidence and strength of recommendations, is underway to support a GRADE Ontology.

Tools for End Users

HEvKA participants have developed tools to facilitate the visualization and data entry of evidence into EBMonFHIR resources, including human-readable interfaces and JSON viewers. The tools are integrated into the FEvIR Platform [54], and include intuitive forms, purposefully created to not require manual JSON coding or working knowledge of FHIR, for viewing and building *Citation*, *Evidence*, *EvidenceVariable*, and *Group* resources; for viewing and building multiple profiles of *Composition* resource (*Guideline*, *Recommendation*, *SummaryOfFindings*); for viewing and building multiple profiles of *ArtifactAssessment* resource (*Classification*, *Rating*, *RiskOfBiasAssessment*, and *RecommendationJustification*); and automated converters to translate data from MEDLINE, RIS, ClinicalTrials.gov, and MAGICapp into the FHIR specification. The viewing of resources created with these tools is open without an account, but an account (at no cost) is required to create content.

XSL•FO
RenderX

This environment may reduce the time spent on SR evidence gathering. The laborious process of searching for articles and screening them by hand could be automated because the search query would be expressed in the "language" of EBMonFHIR and judgments made by previous searchers (such as population classification) can be recorded for reuse. Similarly, automated tools could be used to update evidence within SRs rather than relying on manual updates.

With data available in the "language" of EBMonFHIR, it will become easier to develop tools to analyze and process scientific knowledge. Automated meta-analysis tools could be developed based on the recognition of *Evidence* resources with matching *variableDefinition* element content and processing algorithms mapped to the structured *statistic* content. Tools to accelerate original research will also be developed, such as tools to match clinical trials with potentially eligible patients [82].

The *ArtifactAssessment* resource can also be used to represent the quality or certainty of the evidence itself. The EBMonFHIR Implementation Guide includes a *CertaintyOfEvidence* element profile for this purpose. There is also a *DatasetCitation* element Profile of the *Citation* resource, and an *ArtifactAssessment* resource referencing a *DatasetCitation* element can be used to represent the data quality.

Conclusions

Continuously identifying, synthesizing, and incorporating evidence into care are the key tasks of medical knowledge management, yet they are also prohibitively labor-intensive. HEvKA has taken approaches to standardize and eventually automate these tasks through its efforts around EBMonFHIR, an HL7 standard for making biomedical evidence computable. EBMonFHIR will be used for enabling seamless data flow between published evidence reports, repositories, SR authoring tools, and guideline development tools; automating the searching and matching of evidence with any subgroups of patients; connecting individual patient data with medical knowledge for computerized CDS; and providing individualized effect estimates for different outcomes to facilitate shared decision-making. EBMonFHIR resources in conjunction with other FHIR resources can support relaying EBM components in a manner that is interoperable and consumable by downstream tools and health information systems to support evidence users (eg, creators of biomedical knowledge bases, CPGs, CDS artifacts, and SRs). Anyone may join HEvKA to engage a community of FHIR users and committed volunteers to accelerate the development and implementation of standards for evidence exchange.

Acknowledgments

Health Level Seven (HL7) and Fast Healthcare Interoperability Resources (FHIR) are the registered trademarks of Health Level Seven International and their use of these trademarks does not constitute an endorsement by HL7. Evidence-Based Medicine on FHIR (EBMonFHIR) is an HL7 project. We would like to thank all members of the EBMonFHIR and COVID-19 Knowledge Accelerator (COKA) or Health Evidence Knowledge Accelerator (HEvKA) projects who contributed to the design and development of the FHIR resources to support EBM, and Janice Tufte for reviewing from a lay perspective. We would like to acknowledge the Data Science to Patient Value (D2V) initiative at the University of Colorado School of Medicine. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. This paper was written on behalf of Health Evidence Knowledge Accelerator (HEvKA).

Authors' Contributions

All authors contributed to the conceptualization, writing—review and editing, and approval of the submitted paper and will be accountable for its contents. AS and BA performed supervision. AS, JD, KS, and BSA performed data curation. AS, LMS, BK, VS, JD, KS, MA, and BSA performed investigation and methodology. HPL contributed to methodology. AS, JD, and BSA handled project administration. KS and BSA managed resource and software. JD managed software. BK, JD, and BSA conducted validation. AS, BK, VS, JD, KS, and BSA performed visualization. AS, LMS, JR, BK, JD, MA, and BSA wrote the original draft. Classification based on the CRediT taxonomy.

Conflicts of Interest

BSA owns Computable Publishing LLC and leads the Health Evidence Knowledge Accelerator (HEvKA) and Evidence-Based Medicine on Fast Healthcare Interoperability Resources (EBMonFHIR) projects. JD and KS are employees of Computable Publishing LLC. BSA, KS, JD, MA, BK, HPL, KAR, VS, and IK are board members of the Scientific Knowledge Accelerator Foundation (SKAF). AS, LMS, and JR have no disclosures to report.

Multimedia Appendix 1
List of Fast Healthcare Interoperability Resources (FHIR) data elements with reference to controlled terminologies introduced by the Evidence-Based Medicine on FHIR (EBMonFHIR) project.
[DOCX File , 27 KB - jmir_v26i1e54265_app1.docx]

References

XSL•FO
RenderX

1. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010;11:55 [FREE Full text] [doi: 10.1186/1471-2105-11-55] [Medline: 20102628]
2. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform 2014;51:242-253 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.005] [Medline: 24954015]
3. Becker M, Jaschinski T, Eikermann M, Mathes T, Bühn S, Koppert W, et al. A systematic decision-making process on the need for updating clinical practice guidelines proved to be feasible in a pilot study. J Clin Epidemiol 2018;96:101-109 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.12.011] [Medline: 29289763]
4. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2016;25(Suppl 1):S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]
5. Rosenfeld RM, Shiffman RN. Clinical practice guideline development manual: a quality-driven approach for translating evidence into action. Otolaryngol Head Neck Surg 2009;140(6 Suppl 1):S1-S43 [FREE Full text] [doi: 10.1016/j.otohns.2009.04.015] [Medline: 19464525]
6. Murad MH. Clinical practice guidelines: a primer on development and dissemination. Mayo Clin Proc 2017;92(3):423-433 [FREE Full text] [doi: 10.1016/j.mayocp.2017.01.001] [Medline: 28259229]
7. Therasse P. From clinical trials to evidence-based medicine: how to build the evidence!. Eur J Cancer Suppl 2003;1(6):55-65 [FREE Full text] [doi: 10.1016/s1359-6349(03)90010-3]
8. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7(9):e1000326 [FREE Full text] [doi: 10.1371/journal.pmed.1000326] [Medline: 20877712]
9. Boutron I, Créquit P, Williams H, Meerpohl J, Craig JC, Ravaud P. Future of evidence ecosystem series: 1. introduction evidence synthesis ecosystem needs dramatic change. J Clin Epidemiol 2020;123:135-142 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.01.024] [Medline: 32145367]
10. Sackett DL. Evidence-based medicine. Semin Perinatol 1997;21(1):3-5 [FREE Full text] [doi: 10.1016/s0146-0005(97)80013-4] [Medline: 9190027]
11. Wapner RJ, Waters T. Introduction and historical perspective. Clin Obstet Gynecol 2003;46(1):125-131 [FREE Full text] [doi: 10.1097/00003081-200303000-00017] [Medline: 12686902]
12. Crowley P, Chalmers I, Keirse MJ. The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. Br J Obstet Gynaecol 1990;97(1):11-25 [FREE Full text] [doi: 10.1111/j.1471-0528.1990.tb01711.x] [Medline: 2137711]
13. Crowley PA. Antenatal corticosteroid therapy: a meta-analysis of the randomized trials, 1972 to 1994. Am J Obstet Gynecol 1995;173(1):322-335 [FREE Full text] [doi: 10.1016/0002-9378(95)90222-8] [Medline: 7631713]
14. Briceño-Pérez C, Reyna-Villasmil E, Vigil-De-Gracia P. Antenatal corticosteroid therapy: historical and scientific basis to improve preterm birth management. Eur J Obstet Gynecol Reprod Biol 2019;234:32-37 [FREE Full text] [doi: 10.1016/j.ejogrb.2018.12.025] [Medline: 30639954]
15. Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MP, Grimshaw JM, et al. Validity of the agency for healthcare research and quality clinical practice guidelines: how quickly do guidelines become outdated? JAMA 2001;286(12):1461-1467 [FREE Full text] [doi: 10.1001/jama.286.12.1461] [Medline: 11572738]
16. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007;147(4):224-233 [FREE Full text] [doi: 10.7326/0003-4819-147-4-200708210-00179] [Medline: 17638714]
17. Winters BD, Bharmal A, Wilson RF, Zhang A, Engineer L, Defoe D, et al. Validity of the agency for health care research and quality patient safety indicators and the centers for medicare and medicaid hospital-acquired conditions: a systematic review and meta-analysis. Med Care 2016;54(12):1105-1111 [FREE Full text] [doi: 10.1097/MLR.0000000000000550] [Medline: 27116111]
18. Hulse NC, Rocha RA, Del Fiol G, Bradshaw RL, Hanna TP, Roemer LK. KAT: a flexible XML-based knowledge authoring environment. J Am Med Inform Assoc 2005;12(4):418-430 [FREE Full text] [doi: 10.1197/jamia.M1701] [Medline: 15802477]
19. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schünemann HJ, Living Systematic Review Network. Living systematic reviews: 4. living guideline recommendations. J Clin Epidemiol 2017;91:47-53 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.08.009] [Medline: 28911999]
20. Boxwala AA, Peleg M, Tu S, Ogunyemi O, Zeng QT, Wang D, et al. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform 2004;37(3):147-161 [FREE Full text] [doi: 10.1016/j.jbi.2004.04.002] [Medline: 15196480]
21. Peleg M. Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform 2013;46(4):744-763 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.009] [Medline: 23806274]
22. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, McClay J, et al. The SAGE guideline model: achievements and overview. J Am Med Inform Assoc 2007;14(5):589-598 [FREE Full text] [doi: 10.1197/jamia.M2399] [Medline: 17600098]
23. Tse EG, Klug DM, Todd MH. Open science approaches to COVID-19. F1000Res 2020;9:1043 [FREE Full text] [doi: 10.12688/f1000research.26084.1] [Medline: 33145011]

XSL•FO
RenderX

24. McCartney M. COVID-19: has EBM been replaced by hype-based medicine? Drug Ther Bull 2020;58(7):99-100 [FREE Full text] [doi: 10.1136/dtb.2020.000029] [Medline: 32451323]
25. Romiti GF, Corica B, Cangemi R, Basili S, Raparelli V. Need for innovative and timely synthesis of evidence during COVID-19 outbreak. Eur J Intern Med 2020;77:165-166 [FREE Full text] [doi: 10.1016/j.ejim.2020.06.007] [Medline: 32532662]
26. Atallah Á. COVID-19: better trustworthiness of clinical evidence through clinical trials and systematic reviews. Sao Paulo Med J 2020;138(3):171-173 [FREE Full text] [doi: 10.1590/1516-3180.2019.138311062020] [Medline: 32696833]
27. Bero LA. Producing independent, systematic review evidence: cochrane's response to COVID-19. Am J Public Health 2020;110(7):952-953 [FREE Full text] [doi: 10.2105/AJPH.2020.305734] [Medline: 32407142]
28. Laine C, Taichman DB, Guallar E, Mulrow CD. Keeping up with emerging evidence in (Almost) real time. Ann Intern Med 2020;173(2):153-154 [FREE Full text] [doi: 10.7326/M20-2627] [Medline: 32369539]
29. Else H. How a torrent of COVID science changed research publishing—in seven charts. Nature 2020;588(7839):553 [FREE Full text] [doi: 10.1038/d41586-020-03564-y] [Medline: 33328621]
30. Covid-19 report. Google Data Studio. URL: https://lookerstudio.google.com/reporting/70fdce85-1b53-4605-8698-839b66c8010b/page/JwhJB?feature=opengraph [accessed 2024-06-11]
31. Dunn AG, Bourgeois FT. Is it time for computable evidence synthesis? J Am Med Inform Assoc 2020;27(6):972-975 [FREE Full text] [doi: 10.1093/jamia/ocaa035] [Medline: 32337600]
32. Alper BS, Richardson JE, Lehmann HP, Subbian V. It is time for computable evidence synthesis: the COVID-19 knowledge accelerator initiative. J Am Med Inform Assoc 2020;27(8):1338-1339 [FREE Full text] [doi: 10.1093/jamia/ocaa114] [Medline: 32442263]
33. Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. combining human and machine effort. J Clin Epidemiol 2017;91:31-37 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.08.011] [Medline: 28912003]
34. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4(1):5 [FREE Full text] [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
35. Elliott J, Lawrence R, Minx JC, Oladapo OT, Ravaud P, Jeppesen BT, et al. Decision makers need constantly updated evidence synthesis. Nature 2021;600(7889):383-385 [FREE Full text] [doi: 10.1038/d41586-021-03690-1] [Medline: 34912079]
36. Lang ES, Wyer PC, Haynes RB. Knowledge translation: closing the evidence-to-practice gap. Ann Emerg Med 2007;49(3):355-363 [FREE Full text] [doi: 10.1016/j.annemergmed.2006.08.022] [Medline: 17084943]
37. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
38. Evidence Ecosystem. MAGIC Evidence Ecosystem Foundation. URL: https://www.magicevidence.org/evidence-ecosystem [accessed 2024-06-11]
39. Taber P, Radloff C, Del Fiol G, Staes C, Kawamoto K. New standards for clinical decision support: a survey of the state of implementation. Yearb Med Inform 2021;30(1):159-171 [FREE Full text] [doi: 10.1055/s-0041-1726502] [Medline: 34479387]
40. CDS connect. Agency for Healthcare Research and Quality. URL: https://cds.ahrq.gov/cdsconnect [accessed 2023-01-16]
41. CDS hooks. HL7 and Boston Children's Hospital. URL: https://cds-hooks.hl7.org/ [accessed 2023-01-16]
42. Lehne M, Luijten S, Imbusch PVFG, Thun S. The use of FHIR in digital health—a review of the scientific literature. Stud Health Technol Inform 2019;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
43. Jaffe C, Mandel J, Sartin S, Posnack S, Overhage JM, Tripathi M. W09: FHIR -- implementing the HL7 interoperability platform: a community of implementers for research, patient care, and value-based care. 2019 Presented at: AMIA 2019 Annual Symposium; November 16-20, 2019; Washington DC URL: https://symposium2019.zerista.com/event/member/602305
44. FHIR v4.0.1. HL7® FHIR® Release 4. URL: http://hl7.org/fhir/R4/index.html [accessed 2024-05-17]
45. EBMonFHIR—clinical decision support—confluence. Atlassian Confluence Community License. URL: https://confluence.hl7.org/display/cds/ebmonfhir [accessed 2021-03-28]
46. HL7 searchable project index—FHIR resources for evidence-based medicine knowledge assets (EBMonFHIR). HL7 International. URL: https://www.hl7.org/special/Committees/projman/searchableProjectIndex.cfm?action=edit&ProjectNumber=1422 [accessed 2021-04-05]
47. COVID-19 Knowledge Accelerator (COKA)—clinical decision support—confluence. Atlassian Confluence Community License. URL: https://confluence.hl7.org/pages/viewpage.action?pageId=97468919 [accessed 2021-03-29]
48. Understanding the standards process—HL7—confluence. Atlassian Confluence Community License. URL: https://confluence.hl7.org/display/HL7/Understanding+the+Standards+Process [accessed 2022-08-10]
49. FHIR Connectathons. Atlassian Confluence Community License. URL: https://confluence.hl7.org/display/FHIR/Connectathons [accessed 2023-01-16]

XSL•FO
RenderX

50. Alper BS, Dehnbostel J, Afzal M, Subbian V, Soares A, Kunnamo I, et al. Making science computable: developing code systems for statistics, study design, and risk of bias. J Biomed Inform 2021;115:103685 [FREE Full text] [doi: 10.1016/j.jbi.2021.103685] [Medline: 33486066]
51. Clinical decision support workgroup—confluence. Atlassian Confluence Community License. URL: https://confluence.hl7.org/display/CDS/WorkGroup+Home [accessed 2021-06-29]
52. Personal cloud storage and file sharing platform—Google. Google Drive. URL: https://www.google.com/drive/ [accessed 2023-01-16]
53. Video conferencing, meetings, calling. Microsoft Teams. URL: https://www.microsoft.com/en-us/microsoft-teams/group-chat-software [accessed 2023-01-16]
54. FEvIR platform. FEvIR Platf. URL: https://fevir.net/ [accessed 2024-03-27]
55. HEvKA-Health Evidence Knowledge Accelerator—clinical decision support—confluence. Atlassian Confluence Community License. URL: https://confluence.hl7.org/display/CDS/HEvKA-Health+Evidence+Knowledge+Accelerator [accessed 2023-01-16]
56. Datatypes—FHIR v6.0.0-cibuild. HL7®FHIR® FHIR CI-Build. URL: https://build.fhir.org/datatypes.html [accessed 2023-11-02]
57. Evidence. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/evidence.html [accessed 2023-11-02]
58. Group. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/group.html [accessed 2023-11-02]
59. EvidenceVariable. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/evidencevariable.html [accessed 2023-11-02]
60. Valueset-statistic-type. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/valueset-statistic-type.html [accessed 2023-11-02]
61. Valueset-certainty-type. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/valueset-certainty-type.html [accessed 2023-11-02]
62. Citation. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/citation.html [accessed 2023-11-02]
63. FEvIR citation—citation for FEvIR evidence 7637: critically appraised summary of primary outcome of multi-platform RCT of anticoagulation for non-critically ill COVID-19. FEvIR Platf. URL: https://fevir.net/resources/Citation/7638 [accessed 2023-11-02]
64. FEvIR evidence—critically appraised summary of primary outcome of multi-platform RCT of anticoagulation for non-critically ill patients with COVID-19. FEvIR Platform. URL: https://fevir.net/resources/Evidence/7637 [accessed 2023-11-02]
65. RECOVERY Collaborative Group, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in hospitalized patients with COVID-19. N Engl J Med 2021;384(8):693-704 [FREE Full text] [doi: 10.1056/NEJMoa2021436] [Medline: 32678530]
66. EvidenceReport. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/evidencereport.html [accessed 2023-11-02]
67. ArtifactAssessment. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/artifactassessment.html [accessed 2023-11-02]
68. Terminologies-systems. FHIR v6.0.0-cibuild. URL: https://build.fhir.org/terminologies-systems.html [accessed 2023-11-02]
69. ATTACC Investigators, ACTIV-4a Investigators, REMAP-CAP Investigators, Lawler PR, Goligher EC, Berger JS, et al. Therapeutic anticoagulation with heparin in noncritically ill patients with COVID-19. N Engl J Med 2021;385(9):790-802 [FREE Full text] [doi: 10.1056/NEJMoa2105911] [Medline: 34351721]
70. EBMonFHIR walkthrough. Atlassian Confluence Community License. URL: https://confluence.hl7.org/display/CDS/EBMonFHIR+Walkthrough [accessed 2023-11-03]
71. FEvIR citation—anticoagulation for COVID-19 combined RCTs in NEJM. FEvIR Platf. URL: https://fevir.net/resources/Citation/7636 [accessed 2023-11-02]
72. FEvIR EvidenceVariable—therapeutic-dose anticoagulation with heparin. FEvIR Platf. URL: https://fevir.net/resources/EvidenceVariable/7751 [accessed 2023-11-02]
73. FEvIR EvidenceVariable—usual-care pharmacologic thromboprophylaxis. FEvIR Platf. URL: https://fevir.net/resources/EvidenceVariable/7752 [accessed 2023-11-02]
74. FEvIR EvidenceVariable—organ support-free days. FEvIR Platf. URL: https://fevir.net/resources/EvidenceVariable/7753 [accessed 2023-11-02]
75. FEvIR Group—patients who are hospitalized for COVID-19 and who are not critically ill. FEvIR Platf. URL: https://fevir.net/resources/Group/7749 [accessed 2023-11-02]
76. FEvIR Group—participants in anticoagulation for COVID-19 combined (ATTACC, ACTIV-4a, and REMAP-CAP) RCT (hospitalized, not critically ill). FEvIR Platf. URL: https://fevir.net/resources/Group/7750 [accessed 2023-11-02]
77. FHIR maturity model. Health Level Seven International. URL: https://wiki.hl7.org/FHIR_Maturity_Model [accessed 2021-07-30]
78. Lichtner G, Alper BS, Jurth C, Spies C, Boeker M, Meerpohl JJ, et al. Representation of evidence-based clinical practice guideline recommendations on FHIR. J Biomed Inform 2023;139:104305 [FREE Full text] [doi: 10.1016/j.jbi.2023.104305] [Medline: 36738871]
79. FEvIR project—evidence based medicine implementation guide. FEvIR Platf. URL: https://fevir.net/resources/Project/29736 [accessed 2024-03-27]
80. Evidence based medicine on FHIR implementation guide v1.0.0-ballot. HL7 International. URL: https://build.fhir.org/ig/HL7/ebm/index.html [accessed 2024-03-27]

XSL•FO
RenderX

81. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
82. Alper BS, Dehnbostel J, Shahin K, Ojha N, Khanna G, Tignanelli CJ. Striking a match between FHIR-based patient data and FHIR-based eligibility criteria. Learn Health Syst 2023;7(4):e10368 [FREE Full text] [doi: 10.1002/lrh2.10368] [Medline: 37860063]

Abbreviations

CDS: clinical decision support
COKA: COVID-19 Knowledge Accelerator
CPG: clinical practice guideline
EBM: evidence-based medicine
EBMonFHIR: Evidence-Based Medicine on Fast Healthcare Interoperability Resources
FEvIR: Fast Evidence Interoperability Resources
FHIR: Fast Healthcare Interoperability Resources
GRADE: Grading of Recommendations Assessment, Development and Evaluation
GIN: Guidelines International Network
GLIF: GuideLine Interchange Format
HEvKA: Health Evidence Knowledge Accelerator
HL7: Health Level Seven
IDMP: Identification of Medicinal Products
LOINC: Logical Observation Identifiers Names and Codes
MCBK: Mobilizing Computable Biomedical Knowledge
RCT: randomized controlled trial
SAGE: Standards-Based Sharable Active Guideline Environment
SEVCO: Scientific Evidence Code System
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
SR: systematic review
UML: unified modeling language

Edited by G Tsafnat; submitted 03.11.23; peer-reviewed by S Thun, C Vorisek; comments to author 22.03.24; revised version received 04.04.24; accepted 06.04.24; published 25.06.24.

Please cite as:
Soares A, Schilling LM, Richardson J, Kommadi B, Subbian V, Dehnbostel J, Shahin K, Robinson KA, Afzal M, Lehmann HP, Kunnamo I, Alper BS
Making Science Computable Using Evidence-Based Medicine on Fast Healthcare Interoperability Resources: Standards Development Project
J Med Internet Res 2024;26:e54265
URL: https://www.jmir.org/2024/1/e54265
doi:10.2196/54265
PMID:38916936

©Andrey Soares, Lisa M Schilling, Joshua Richardson, Bhagvan Kommadi, Vignesh Subbian, Joanne Dehnbostel, Khalid Shahin, Karen A Robinson, Muhammad Afzal, Harold P Lehmann, Ilkka Kunnamo, Brian S Alper. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Collaborative Human–Computer Vision Operative Video Analysis Algorithm for Analyzing Surgical Fluency and Surgical Interruptions in Endonasal Endoscopic Pituitary Surgery: Cohort Study

Chia-En Wong[1], MD; Pei-Wen Chen[2,3], MSc; Heng-Jui Hsu[4], MD; Shao-Yang Cheng[2], BS; Chen-Che Fan[2], BS; Yen-Chang Chen[2], BS; Yi-Pei Chiu[2], BS; Jung-Shun Lee[1,3,5*], MSc, MD; Sheng-Fu Liang[2,6*], PhD

[1]Division of Neurosurgery, Department of Surgery, National Cheng Kung University Hospital, Tainan, Taiwan
[2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
[3]Department of Cell Biology and Anatomy, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[4]Department of Otolaryngology-Head and Neck Surgery, National Cheng Kung University Hospital, Tainan, Taiwan
[5]Institute of Basic Medical Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[6]Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan
[*]these authors contributed equally

Corresponding Author:
Sheng-Fu Liang, PhD
Department of Computer Science and Information Engineering
National Cheng Kung University
No 1, University Road
Tainan, 701
Taiwan
Phone: 886 62757575
Email: sfliang@ncku.edu.tw

Abstract

Background: The endonasal endoscopic approach (EEA) is effective for pituitary adenoma resection. However, manual review of operative videos is time-consuming. The application of a computer vision (CV) algorithm could potentially reduce the time required for operative video review and facilitate the training of surgeons to overcome the learning curve of EEA.

Objective: This study aimed to evaluate the performance of a CV-based video analysis system, based on OpenCV algorithm, to detect surgical interruptions and analyze surgical fluency in EEA. The accuracy of the CV-based video analysis was investigated, and the time required for operative video review using CV-based analysis was compared to that of manual review.

Methods: The dominant color of each frame in the EEA video was determined using OpenCV. We developed an algorithm to identify events of surgical interruption if the alterations in the dominant color pixels reached certain thresholds. The thresholds were determined by training the current algorithm using EEA videos. The accuracy of the CV analysis was determined by manual review, and the time spent was reported.

Results: A total of 46 EEA operative videos were analyzed, with 93.6%, 95.1%, and 93.3% accuracies in the training, test 1, and test 2 data sets, respectively. Compared with manual review, CV-based analysis reduced the time required for operative video review by 86% (manual review: 166.8 and CV analysis: 22.6 minutes; $P<.001$). The application of a human-computer collaborative strategy increased the overall accuracy to 98.5%, with a 74% reduction in the review time (manual review: 166.8 and human-CV collaboration: 43.4 minutes; $P<.001$). Analysis of the different surgical phases showed that the sellar phase had the lowest frequency (nasal phase: 14.9, sphenoidal phase: 15.9, and sellar phase: 4.9 interruptions/10 minutes; $P<.001$) and duration (nasal phase: 67.4, sphenoidal phase: 77.9, and sellar phase: 31.1 seconds/10 minutes; $P<.001$) of surgical interruptions. A comparison of the early and late EEA videos showed that increased surgical experience was associated with a decreased number (early: 4.9 and late: 2.9 interruptions/10 minutes; $P=.03$) and duration (early: 41.1 and late: 19.8 seconds/10 minutes; $P=.02$) of surgical interruptions during the sellar phase.

Conclusions: CV-based analysis had a 93% to 98% accuracy in detecting the number, frequency, and duration of surgical interruptions occurring during EEA. Moreover, CV-based analysis reduced the time required to analyze the surgical fluency in

XSL•FO
RenderX

EEA videos compared to manual review. The application of CV can facilitate the training of surgeons to overcome the learning curve of endoscopic skull base surgery.

Trial Registration: ClinicalTrials.gov NCT06156020; https://clinicaltrials.gov/study/NCT06156020

(J Med Internet Res 2024;26:e56127) doi:10.2196/56127

KEYWORDS

algorithm; computer vision; endonasal endoscopic approach; pituitary; transsphenoidal surgery

Introduction

The endonasal endoscopic approach (EEA) is an effective technique for pituitary adenoma (PA) resection [1-3]. The introduction of an operative endoscope into the nasal cavity greatly improves the intraoperative visualization of critical neurovascular structures and facilitates the resection of tumors with suprasellar and cavernous sinus components [1,4]. Compared with microscopic transsphenoidal resections, EEA has been shown to have a higher percentage of complete resection and a lower incidence of neurovascular complications when performed by experienced surgeons [4-6]. Nevertheless, EEA has a long learning curve compared with microscopic approaches. A previous study reported that a 15- to 50-case learning curve is required to improve EEA outcomes [7,8].

To shorten the learning curve for endoscopic surgeries, artificial intelligence (AI) and computer vision (CV) analyses of operative videos provide novel and effective tools to reduce the time required for manual review [9]. Xu et al [10] reported the use of CV in the analysis of thoracoscopic pulmonary lobectomies and demonstrated that bleeding could be detected using CV and was associated with postoperative outcomes. Ryu et al [11] described the use of a CV-based tracking of surgical instruments in robot-assisted laparoscopic surgery to prevent instrument collisions. The use of CV analyses in endoscopic neurosurgery has been reported only recently. Khan et al [12] developed a CV-based workflow analysis to distinguish the nasal, sphenoidal, and sellar phases during EEA and achieved accurate automated recognition of the surgical phases and steps.

Although the system developed by Khan et al [12] could evaluate the time spent in each phase of EEA surgery, however, technical fluency and operative efficiency were not evaluated, which could provide invaluable information for surgeons to refine their operative techniques and improve the learning curve [13,14]. Currently, manual review of EEA operative videos remains time-consuming and ineffective [14]; therefore, an automated CV-based system to analyze the technical fluency and operative efficiency of EEA videos will facilitate the training of surgeons to overcome the learning curve of endoscopic skull base surgery.

To address this knowledge gap, this study aimed to develop a CV-based system for analyzing surgical fluency in EEA videos by detecting the number, frequency, and duration of surgical interruptions that occur during EEA surgeries. Since removal and reintroduction of the endoscope into the nasal cavity require a transition of the surgeons' focus from the video screen to the operative field and frequent redirection of the surgeons' attention was associated with deleterious effects on surgeons' performance [15,16], we aimed to detect the events of removal and reintroduction of the endoscope during EEA to represent surgical interruptions. The time spent on CV-based and manual reviews of the surgical video was compared. Additionally, the learning curve of the EEA and the impact of tubular retractor use were investigated.

Methods

Ethical Considerations

This study was approved by the institutional review board of the National Cheng-Kung University Hospital (approval B-ER-112-452). The requirement for informed consent was waived, as approved by the institutional review board. All patient data, including the operative video, were deidentified. Compensation was not provided to included patients, as approved by the institutional review board. This study was reported in line with the STROCSS (Strengthening the Reporting of Cohort Studies in Surgery) criteria [17] and registered on ClinicalTrials.gov (NCT06156020).

Patients

Patients who underwent EEA for PAs between January 2020 and March 2024 with fully recorded anonymous operative videos were retrospectively reviewed. We excluded patients whose operative videos were incompletely recorded and those with poor-quality operative videos. Patients were excluded if their facial identity could be recognized in the video. All patients underwent standard 2-surgeon EEA techniques performed by an ear, nose, and throat rhinology specialist and a neurosurgeon. All surgeries were performed endoscopically.

Videos from January 2020 to August 2023 were randomly assigned into training and test 1 data sets. Videos from September 2023 to February 2024 were assigned into test 2 data set, which included EEA videos recorded using a new set of operative endoscopes and surgical instruments or performed by another team of surgeons.

Surgical Fluency and Surgical Interruptions

The definitions of surgical interruptions were reviewed and varied according to the types of different surgeries [18]. In this study, surgical interruption was defined on the basis of all of the following criteria: (1) blood contamination of endoscopic vision, (2) removal of the endoscope from the nasal cavity, and (3) cleaning of the endoscope. These criteria indicate cessation of operative maneuvers and were used for both the manual review and development of the CV-based algorithm. The number and duration of the surgical interruptions were also recorded. The frequency and total duration of the surgical interruptions were calculated every 10 minutes.

XSL•FO
RenderX

CV-Based Algorithm

Our algorithm aimed to detect surgical interruptions that necessitated the removal of the endoscope from the nasal cavity for cleaning (Figure 1). All surgical videos were converted to 19201080 pixels. When given a surgical video with red, green, and blue channels, the 8-bit per channel red, green, and blue values of all the pixels in each frame could be obtained using OpenCV software and expressed as a 1920×1080×3 matrix. Under normal illumination, the color signatures of the nasal and sellar mucosa and structures were red-pinkish, the surgical drape was blue-greenish, and the surgical instruments and gauze were silverish-whitish. Using this property, nasal mucosa, surgical drapes, and instruments were classified using *K*-means clustering, a variance-based clustering technique that can be used to classify different objects.

Figure 1. Graphical illustration of the present computer vision–based algorithm to detect a surgical interruption in endonasal endoscopic approach resection of pituitary adenomas. #: number; G–Bx: number of pixels with green-blue dominant color; RGB: red, green, and blue; Rx: number of pixels with red dominant color; SRG-B: silver, red, green and blue; Sx: number of pixels with silver dominant color.

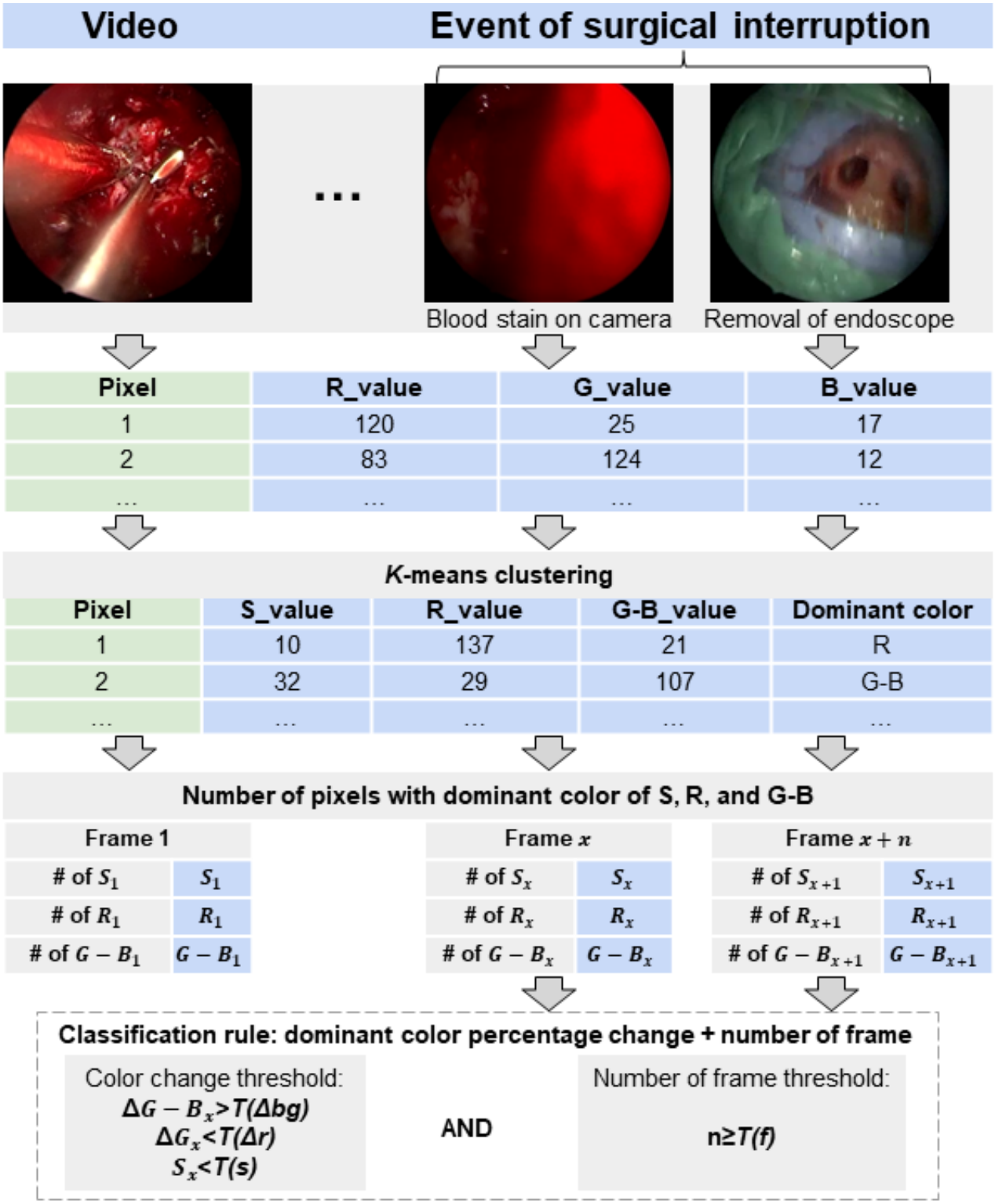

XSL•FO
RenderX

Then, the image color space was converted to the SRB-G space, where S stands for silver and R and B-G stand for the 2-color channels representing the nasal and sellar mucosa and surgical drapes, respectively. The dominant color in terms of SRB-G in each pixel could then be determined, and the number of pixels with the dominant colors S, R, and B-G could be expressed as S, R, and B-G, respectively. The alterations in the dominant color in each frame were compared to those in the previous frames and could further be expressed as ΔS, ΔR, and ΔB-G, respectively. An algorithm was developed to claim surgical interruption events according to the thresholds of S, R, and B-G and ΔS, ΔR, and ΔB-G as follows:

$$\Delta B\text{–}G > T(\Delta bg) \text{ AND } \Delta R < T(\Delta r) \text{ AND } S < T(s)$$

Where $T(\Delta bg)$ and $T(\Delta r)$ are the percentage thresholds of the alterations in the B-G and R values, respectively, indicating that a sufficient increase in the B-G percentage and a simultaneous decrease in the R percentage were required to claim a surgical interruption. Furthermore, $T(s)$ was the threshold of the S percentage, which was used to exclude the impact of moving surgical instruments on alterations in R and B-G values.

Because the event of surgical interruption did not occur in a single frame but rather in consecutive frames, the S, R, and B-G values were compared to the numbers of previous frames, which we further improved the foregoing method by adding a threshold in the number of consecutive frames $T(f)$ compared. An event of surgical interruption should last longer than this threshold; therefore, if an identified event lasted less than $T(f)$ frames, it would not be claimed as a surgical interruption event. The thresholds $T(\Delta bg)$, $T(\Delta r)$, $T(s)$, and $T(\Delta f)$ were determined using the training data set. We found that settings $T(\Delta bg)$ of 35%, $T(\Delta r)$ of 22%, $T(s)$ of 32%, and $T(f)$ of 30% optimized the performance of the foregoing detection rules on the training data set with an overall accuracy of 93.6% (sensitivity=94.7%, specificity=93.4%).

Statistical Analysis

Parametric variables were compared using the 2-tailed Student *t* test. Nonparametric variables were compared using the Mann-Whitney *U* test. Multiple comparisons of parametric variables were performed using ANOVA. Multiple comparisons of paired variables were performed using repeated measures ANOVA with pairwise post hoc comparisons. Statistical tests were conducted using MedCalc (version 19.7.2; MedCalc Software, Ltd).

Results

Operative Videos Included for Analysis

In total, 46 EEA operative videos were analyzed, with 10, 26, and 10 videos used as the training, test 1, and test 2 data sets, respectively. The overall accuracy in the training data set was 93.6% (sensitivity=94.7%, specificity=93.4%). The overall accuracy in test 1 data set was 95.1% (sensitivity=94.5%, specificity=95.3%). The overall accuracy in test 2 data set was 93.3% (sensitivity=93.1%, specificity=93.4%).

The number of surgical interruptions every 10 minutes was similar between the 3 data sets in all surgical phases (ANOVA, nasal phase: training=15.1, test 1=14.5, and test 2=17.0 interruptions/10 minutes; *df*=2; *P*=.65; sphenoidal phase: training=12.4, test 1=14.7, and test 2=12.9 interruptions/10 minutes; *df*=2; *P*=.37; and sellar phase: training=5.6, test 1=4.2, and test 2=3.7 interruptions/10 minutes; *df*=2; *P*=.16; Figure 2A). The duration of surgical interruption was also similar between the data sets (ANOVA, nasal phase: training=65.4, test 1=65.1, and test 2=77.6 seconds/10 minutes; *df*=2; *P*=.70; sphenoidal phase: training=71.8, test 1=81.7, and test 2=78.3 seconds/10 minutes; *df*=2; *P*=.75; and sellar phase: training=39.4, test 1=27.3, and test 2=31.9 seconds/10 minutes; *df*=2; *P*=.15; Figure 2B).

Figure 2. Comparison between the computer vision–based detection results between the training and test data sets. (A) The average number and (B) average total duration of surgical interruptions normalized to every 10 minutes in the nasal, sphenoidal, and sellar phases of the training and test video data sets. ns: not significant.

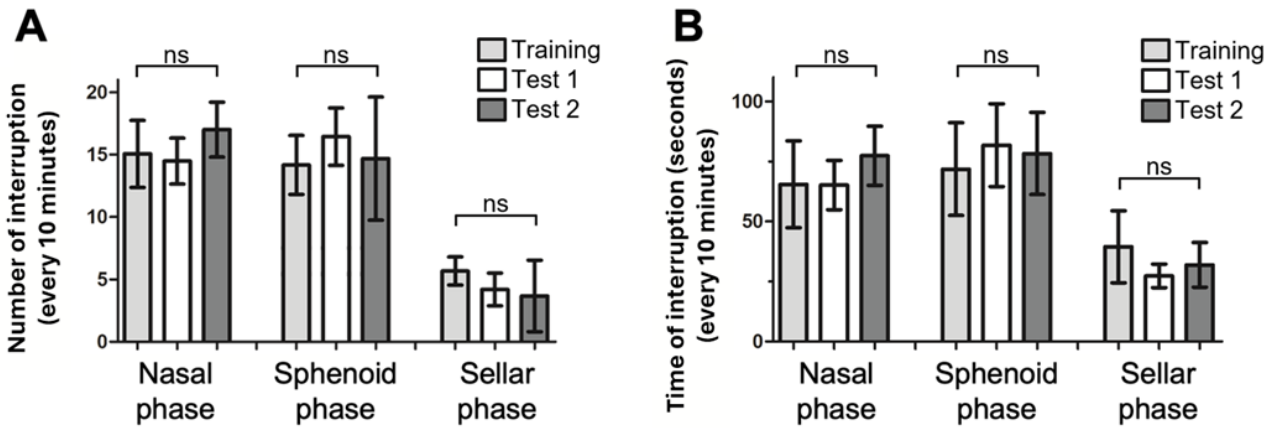

Time Used in CV-Based and Manual Analyses

The time required for manual, CV-based, and human-computer collaborative analyses of EEA operative videos is presented in Figure 3 (repeated measures ANOVA; *df*=27; *P*<.001). The average operative video length was 132.5 (SD 34.6) minutes, and the average time spent for manual video review was 166.8 (SD 43.8) minutes (paired post hoc comparison, surgery: 132.5 and manual review: 166.8 minutes; *P*=.01). The average time required for CV-based automated analysis was 22.6 (SD 7.6)

XSL•FO
RenderX

minutes, which was significantly less than the video length (paired post hoc comparison, CV analysis: 22.6 and surgery: 132.5 minutes; *P*<.001) and the time used in the manual analysis (paired post hoc comparison, CV analysis: 22.6 and manual review: 166.8 minutes; *P*<.001).

Figure 3. The average length of endonasal endoscopic approach videos and the time required for manual, CV-based, and human-computer collaborative video analyses. CV: computer vision. **P*<.05, ***P*<.01, and ****P*<.001.

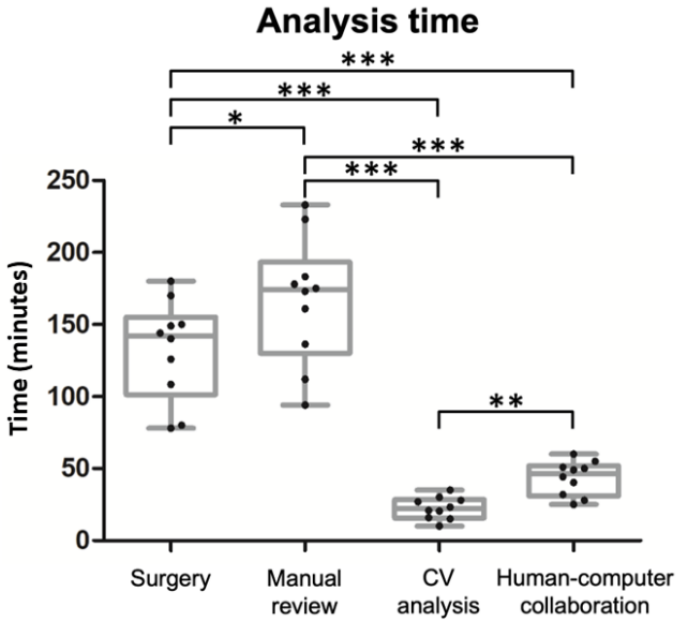

To improve the accuracy of the current CV-based operative video analysis, we used a human-computer collaborative strategy, in which the videos were first analyzed using an automated system, and the identified events of surgical interruptions were subsequently evaluated manually to determine their correctness. In this strategy, the average time spent for analysis was 43.4 (SD 11.8) minutes, which was significantly less than the video length (paired post hoc comparison, human-CV collaboration: 43.4 and surgery: 132.5 minutes; *P*<.001) and the time used in manual analysis (paired post hoc comparison, human-CV collaboration: 43.4 and manual review: 166.8 minutes; *P*<.001) but longer than the time for the CV-based automated analysis (paired post hoc comparison, human-CV collaboration: 43.4 and CV analysis: 22.6 minutes; *P*=.002).

The Learning Curve in the Different Surgical Phases of EEA

We investigated the difference in frequency (ANOVA, nasal phase: 14.9, sphenoidal phase: 15.9, and sellar phase: 4.9 interruptions/10 minutes; *df*=2; *P*<.001) and duration (ANOVA, nasal phase: 67.4, sphenoidal phase: 77.9, and sellar phase: 31.1 seconds/10 minutes; *df*=2; *P*<.001) of surgical interruptions that occurred during the different phases of EEA (Figure 4A and C). The sellar phase had the least number and duration of surgical interruptions every 10 minutes compared with the nasal (paired post hoc comparison, sellar phase: 4.9 and nasal phase: 14.9 interruptions/10 minutes; *P*<.001) and sphenoidal phases (paired post hoc comparison, sellar phase: 4.9 and sphenoidal phase: 15.9 interruptions/10 minutes; *P*<.001). Although the sphenoidal phase had the highest number and longest duration of surgical interruptions, the differences were not significant compared with those in the nasal phase (number of interruptions: paired post hoc comparison, sphenoidal phase: 15.9 and nasal phase: 14.9 interruptions/10 minutes; *P*=.16; duration of interruptions: paired post hoc comparison, sphenoidal phase: 77.9 and nasal phase: 67.4 seconds/10 minutes; *P*=.10).

XSL•FO
RenderX

Figure 4. Surgical interruptions during different phases of EEA and the effect of surgeon experiences. (A) The average number and (C) the average total duration of surgical interruptions normalized to every 10 minutes in the nasal, sphenoidal, and sellar phases of EEA operative videos. Comparisons of (B) the average number and (D) the average total duration of surgical interruptions in EEA operative videos of surgeons with ≤20 procedural experiences and >50 procedural experiences. EEA: endonasal endoscopic approach.

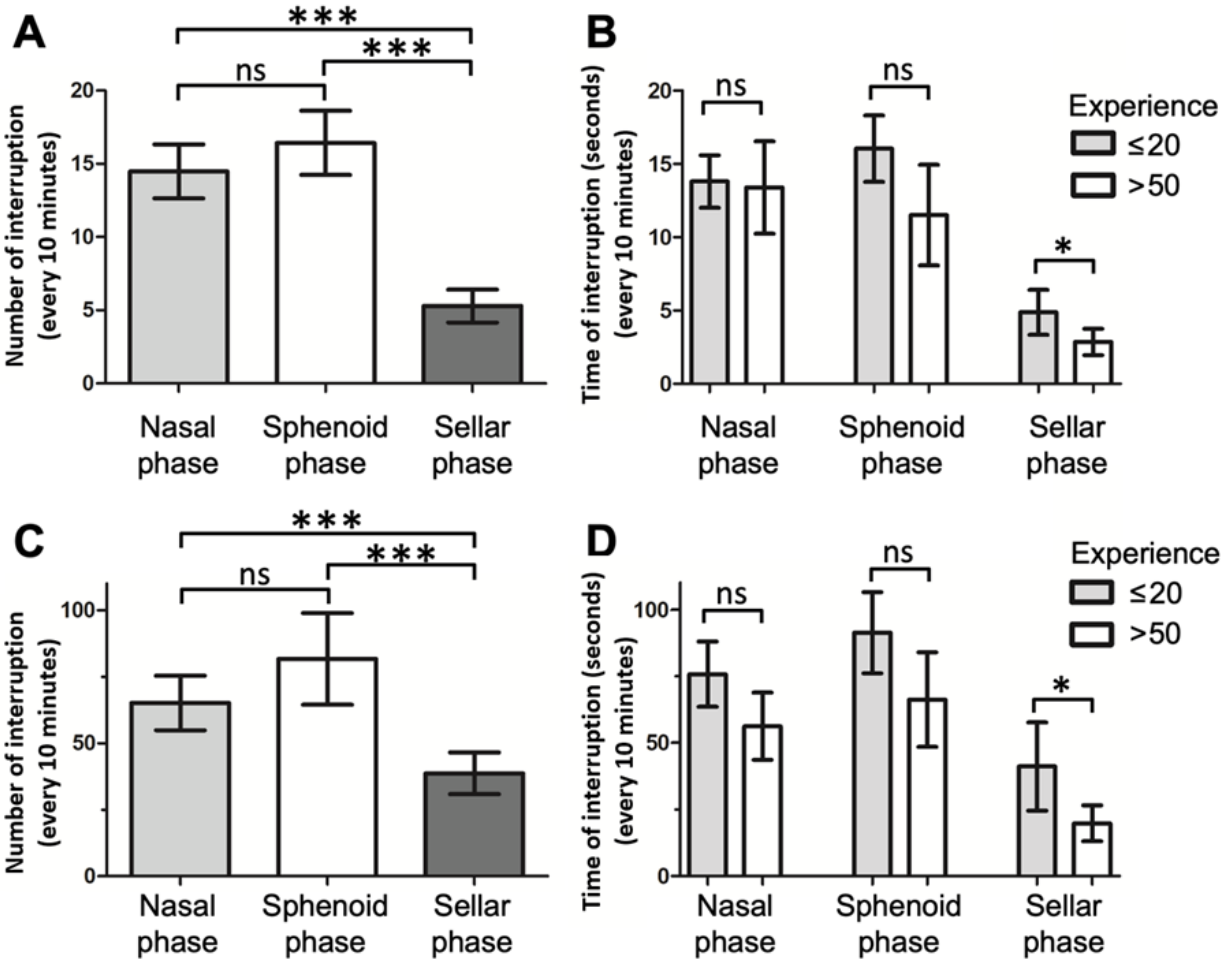

To demonstrate the learning curve of EEA, the analysis results of early (≤20 procedural experiences, n=11) and late operative videos (>50 procedural experiences, n=8) were compared (Figure 4B and D). The results demonstrated a decreased number (2-tailed Student *t* test, early: 4.9 and late: 2.9 interruptions/10 minutes; *P*=.03) and duration (2-tailed Student *t* test, early: 41.1 and late: 19.8 seconds/10 minutes; *P*=.02) of surgical interruptions in the sellar phase in late videos compared with early videos. Additionally, there were mild trends of fewer numbers (2-tailed Student *t* test, early: 16.0 and late: 11.6 interruptions/10 minutes; *P*=.09) and shorter durations (2-tailed Student *t* test, early: 91.2 and late: 69.0 seconds/10 minutes; *P*=.08) of surgical interruptions in the sphenoidal phase of late videos compared with early videos, but the difference was not statistically significant. The use of a tubular retractor in the sellar phase did not affect the operative time (2-tailed Student *t* test, no tube: 73.1 and tube: 87.7 minutes; *P*=.31), number (2-tailed Student *t* test, no tube: 5.1 and tube: 3.7 interruptions/10 minutes; *P*=.10), or duration (2-tailed Student *t* test, no tube: 42.4 and tube: 28.8 seconds/10 minutes; *P*=.11) of surgical interruptions (Figure 5).

XSL•FO
RenderX

Figure 5. The effect of a tubular retractor in the sellar phase of endonasal endoscopic approach (EEA) operative videos. Quantitative comparisons of the (A) operative time, (B) average number, and (C) average total duration of surgical interruptions in the sellar phase of EEA operative videos with or without the use of a tubular retractor.

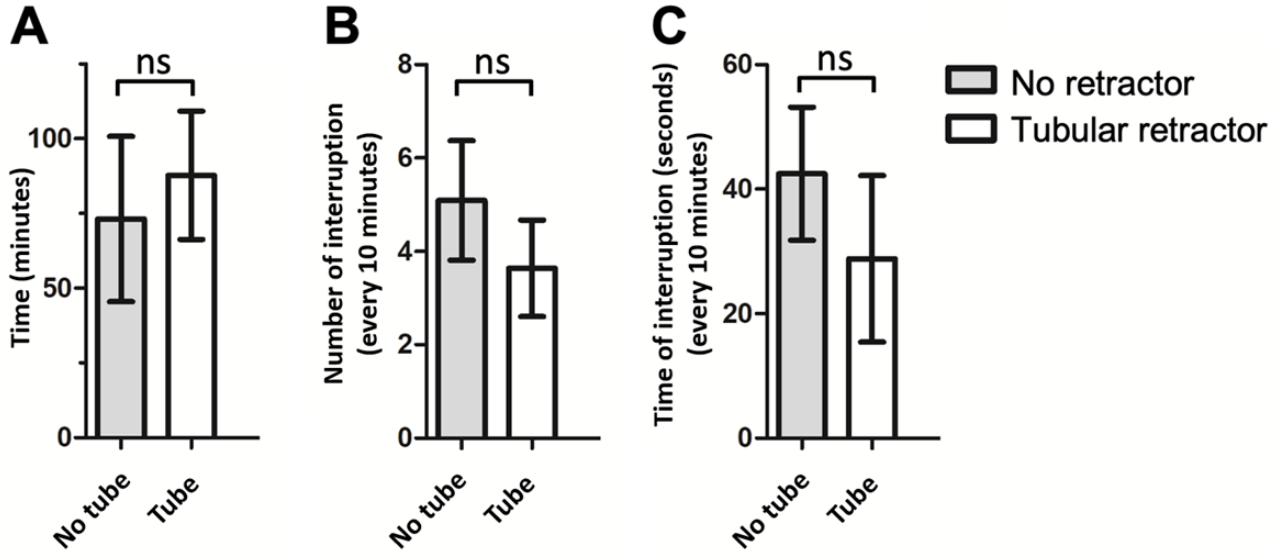

Discussion

Principal Findings

The application of CV can improve the review process of operative videos, thereby facilitating the training of surgeons to overcome the learning curve of endoscopic skull base surgery. In this study, we developed a CV-based video analysis system to detect surgical interruptions and analyze surgical fluency in the EEA with an overall accuracy of 93.6%. The CV-based analysis greatly reduced the time required for operative video reviews. The learning curves of the different phases of EEA surgery were demonstrated in the CV-based analysis.

The algorithm used here was designed to detect events of surgical interruption. Removal and reintroduction of the operative endoscope into the nasal cavity require time and transition of the surgeons' focus from the video screen to the operative field [16]. Frequent redirection of the surgeons' attention during surgery was associated with increased surgeon load on working memory, and intraoperative distraction was associated with a deleterious effect on the surgeons' performance [15,19,20]. Moreover, re-entering the endonasal corridor implies the need for reidentification of critical neurovascular structures, such as the optic nerves and internal carotid artery; thus, frequent endoscope removal and re-entry into the endonasal corridor might increase the risk of misidentification and neurovascular injury [16]. Therefore, the frequency and duration of surgical interruptions are critical intraoperative factors that can affect surgical safety and are used as indicators of surgical fluency in this study. Our results showed that the total amount of time spent in surgical interruption could approach 90-120 seconds every 10 minutes, which was equal to 15% to 20% of the time. A reduction in surgical interruption could potentially reduce both the operative time and the risk of surgical complications [18,21]. Although other parameters, such as instrument collision, were reported as indicators of surgical fluency in thoracoscopic and laparoscopic surgeries [22,23], they were not suitable for the EEA because instrument collision might occur frequently considering the narrow endonasal corridor of the approach.

Comparison to Prior Work

The method used in the present algorithm to detect surgical interruptions was primarily based on the identification of the dominant colors in each pixel [24]. Similar approaches were used by Xu et al [10], who designed a CV-based algorithm to detect bloodstains in thoracoscopic surgeries by identifying red pixels. The use of color-based recognition is a reasonable and effective approach, especially for detecting a surgical interruption in the EEA, because the removal of the endoscope from the nasal cavity creates a considerable color contrast between the reddish-pinkish nasal mucosa and greenish-bluish surgical drapes, which facilitates CV recognition of alterations in the dominant color [25,26]. Moreover, considering the aforementioned color spectra analyzed during the EEA, we transformed the original R-G-B color information into the SRG-B space. Instead of using only 2-color channels, we added an S channel to minimize the interference generated by the silverish surgical instrument and whitish gauze in the video [11]. Combining these approaches together, the present algorithm was able to achieve an overall accuracy of 93.6%. Furthermore, the accuracy of the CV-based method can be further improved using a human-computer collaborative strategy [27,28]. In our approach, videos are first analyzed using an automated system, and the identified surgical interruption events are subsequently evaluated manually to determine their correctness. Although the unidentified (false-negative) surgical interruptions could remain undetected, the false-positive surgical interruptions could be eliminated, and the overall accuracy was increased to 98.5%.

Strengths and Limitations

The main goal of the present CV-based EEA video analysis system is to reduce the time spent reviewing operative videos. Previous applications of CV-based videos in PA surgeries were mainly aimed at recognizing the different phases and determining the surgical workflow [12,29]. Lalys et al [29] used

XSL•FO
RenderX

feature extraction and principal component analysis to classify the microscopic resection of PA into 6 phases, with the best correct classification rate of 82%. Similarly, Khan et al [12] developed an automatic CV-based workflow analysis of endoscopic resection of PAs and successfully classified surgical workflow into 3 phases and 7 steps, with accuracies of 90% and 75%, respectively. However, in these studies, the authors did not report the time required for automated analysis or the time saved compared with manual analysis. Therefore, the efficiency of the system remains to be evaluated. Compared with the classification of different surgical phases, the identification of surgical interruption events throughout surgery is potentially a much more time-consuming task [30]. In manual analysis, a review of the entire length of the operative video is required to recognize all the interrupted events. Moreover, additional time was spent recording the events. Therefore, it is reasonable that our results showed that the time required for manual review was significantly longer than the original length of the video [30]. Therefore, the recognition of surgical interruption is a time-consuming task and is well suited for CV-based analysis to reduce the analysis time. Our results showed that the use of CV-based analysis significantly reduced the time required for analysis by approximately 86%. Although the use of the human-computer collaborative strategy slightly increased the time, there was still a 74% reduction in the time spent for operative video analysis. Hence, these results demonstrate that the primary goal of the CV-based analysis was successfully achieved. The application of the present CV-based system will facilitate the training of surgeons to overcome the learning curve of endoscopic skull base surgery more efficiently.

Moreover, the present CV-based analysis could be further used to characterize the learning curve of the EEA and compare different surgical techniques. A previous study reported that a 15- to 50-case learning curve is required in EEA surgeries [7,8,31]. Previous studies by Lofrese et al [8] and Qureshi et al [31] reported a decreased operative duration with increasing experience. Leach et al [7] reported improvements in the duration of surgery and visual outcomes after approximately 50 procedures. However, conflicting results have been reported regarding operative outcomes, with some studies reporting no correlation between surgeon experience and complication rates [8]. Furthermore, confounding factors such as tumor size could affect the surgical length and patient outcome [32-34]. Therefore, the analysis of operative fluency and the calculation of surgical interruption normalized by unit time provided a more objective and quantitative approach to characterize surgical learning curves [18]. Here, we showed that increased experience was associated with decreased surgical interruptions in the sellar phase of the EEA for PA resection. Since the nasal phase and the initial sphenoidotomy were performed by an expert ear, nose, and throat surgeon who specializes in functional endoscopic sinus surgery, it is reasonable that the experience in the 2-surgeon 3-hand approach has less of an effect on surgical fluency in the nasal and sphenoidal phase [35]. In contrast, our results clearly demonstrated the effect of the learning curve on the duration of surgical interruptions in the sellar phase. Additionally, although the use of tubular retractors did not result in a significant reduction in surgical interruptions, these results demonstrate that CV-based analysis of EEA videos can be applied to compare the impact of different surgical techniques or the use of adjunctive surgical instruments on surgical fluency.

This study had several limitations. First, we included operative videos from a single institution. Nonetheless, we analyzed 2 separate test data sets, which comprised EEA videos recorded using different sets of operative endoscopes and surgical instruments or performed by other teams of surgeons to better demonstrate the generalizability of the present CV algorithm. Still, further investigation with EEA videos from other institutions was required to externally validate the present CV-based analysis method. However, since the present CV-based analysis relied mainly on the identification of the dominant color, the effect of interinstitutional differences is potentially minimized, given the similarity in the colors of the nasal mucosa and surgical drapes. Second, additional intraoperative factors could have also influenced the accuracy of our CV analysis as well as the frequency of surgical interruptions. These factors included hypertrophic nasal mucosa, a deviated nasal septum, intraoperative hemorrhage, and the use of novel surgical instruments, which could result in misidentification and bias in analyzing the EEA learning curve [36-38]. Nonetheless, this CV-based analysis could be used to investigate the effects of these intraoperative factors. Further, we showed that the application of human-computer collaborative strategy could increase the overall accuracy to overcome this limitation. Third, the CV method used in this study was an image processing algorithm based on OpenCV color detection rather than a trained, machine learning AI model. The use of machine learning AI model in surgical analysis could potentially extract more information from the operative videos. However, conventional image processing offered several advantages such as faster inference on computer central processing unit and elimination of the need for training on expensive graphic processing units. Moreover, we were able to achieve a 93.3% to 95.1% overall accuracy in this analysis. Finally, whether the frequency and duration of surgical interruptions are associated with patient outcomes remains unclear. Future studies with comprehensive analyses are warranted to clarify the effects of surgical interruptions and other patient and intraoperative factors on patient outcomes.

Conclusions

This study developed a CV-based system to analyze surgical fluency in EEA videos by detecting the number, frequency, and duration of surgical interruptions that occur during EEA surgeries. The CV-based video analysis system had an overall accuracy of 93.6% and significantly reduced the time required for operative video review. The application of CV can facilitate the training of surgeons to overcome the learning curve of endoscopic skull base surgery.

XSL•FO
RenderX

Acknowledgments

The authors thank Dr Wei-Ting Lee and Dr Shang-Yu Hung for providing operative videos. This work is supported by National Cheng Kung University Higher Education Sprout Project, Ministry of Education, Taiwan.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to institutional board regulations but are available from the corresponding author on reasonable request.

Authors' Contributions

CEW, JSL, and SFL conceived the study. CEW and PWC designed the study methodology. CEW, PWC, HJH, and JSL performed surgery and obtained surgical videos. CEW, SYC, CCF, YCC, and YPC constructed the algorithm. CEW and PWC performed statistical analysis. CEW, JSL, and SFL wrote the original draft of the manuscript. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

References

1. Marigil Sanchez M, Karekezi C, Almeida JP, Kalyvas A, Castro V, Velasquez C, et al. Management of giant pituitary adenomas: role and outcome of the endoscopic endonasal surgical approach. Neurosurg Clin N Am 2019;30(4):433-444. [doi: 10.1016/j.nec.2019.05.004] [Medline: 31471050]
2. Rahimli T, Hidayetov T, Yusifli Z, Memmedzade H, Rajabov T, Aghayev K. Endoscopic endonasal approach to giant pituitary adenomas: surgical outcomes and review of the literature. World Neurosurg 2021;149:e1043-e1055. [doi: 10.1016/j.wneu.2021.01.019] [Medline: 33524611]
3. Wong CE, Wang WH, Lan MY, Lee PH, Huang CC, Su PF, et al. Predicting the need for desmopressin treatment during inpatient and after discharge following endoscopic sellar surgery. Front Neurol 2022;13:843646 [FREE Full text] [doi: 10.3389/fneur.2022.843646] [Medline: 35250842]
4. Prajapati HP, Jain SK, Sinha VD. Endoscopic versus microscopic pituitary adenoma surgery: an institutional experience. Asian J Neurosurg 2018;13(2):217-221 [FREE Full text] [doi: 10.4103/ajns.AJNS_160_16] [Medline: 29682011]
5. Chen J, Liu H, Man S, Liu G, Li Q, Zuo Q, et al. Endoscopic vs. microscopic transsphenoidal surgery for the treatment of pituitary adenoma: a meta-analysis. Front Surg 2021;8:806855 [FREE Full text] [doi: 10.3389/fsurg.2021.806855] [Medline: 35187049]
6. Møller MW, Andersen MS, Glintborg D, Pedersen CB, Halle B, Kristensen BW, et al. Endoscopic vs. microscopic transsphenoidal pituitary surgery: a single centre study. Sci Rep 2020;10(1):21942 [FREE Full text] [doi: 10.1038/s41598-020-78823-z] [Medline: 33318567]
7. Leach P, Abou-Zeid AH, Kearney T, Davis J, Trainer PJ, Gnanalingham KK. Endoscopic transsphenoidal pituitary surgery: evidence of an operative learning curve. Neurosurgery 2010;67(5):1205-1212. [doi: 10.1227/NEU.0b013e3181ef25c5] [Medline: 20871457]
8. Lofrese G, Vigo V, Rigante M, Grieco DL, Maresca M, Anile C, et al. Learning curve of endoscopic pituitary surgery: experience of a neurosurgery/ENT collaboration. J Clin Neurosci 2018;47:299-303. [doi: 10.1016/j.jocn.2017.09.011] [Medline: 28988650]
9. Chadebecq F, Vasconcelos F, Mazomenos E, Stoyanov D. Computer vision in the surgical operating room. Visc Med 2020;36(6):456-462 [FREE Full text] [doi: 10.1159/000511934] [Medline: 33447601]
10. Xu H, Han T, Wang H, Liu S, Hou G, Sun L, et al. Detection of blood stains using computer vision–based algorithms and their association with postoperative outcomes in thoracoscopic lobectomies. Eur J Cardiothorac Surg 2022;62(5):ezac154 [FREE Full text] [doi: 10.1093/ejcts/ezac154] [Medline: 35352106]
11. Ryu J, Choi J, Kim HC. Endoscopic vision–based tracking of multiple surgical instruments during robot-assisted surgery. Artif Organs 2013;37(1):107-112. [doi: 10.1111/j.1525-1594.2012.01543.x] [Medline: 23043484]
12. Khan DZ, Luengo I, Barbarisi S, Addis C, Culshaw L, Dorward NL, et al. Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0). J Neurosurg 2022;137(1):51-58 [FREE Full text] [doi: 10.3171/2021.6.JNS21923] [Medline: 34740198]
13. Ahmed H, Scott A. Pilot training and surgical education: flying towards a brighter future. Postgrad Med J 2021;97(1148):345-348. [doi: 10.1136/postgradmedj-2020-138966] [Medline: 33177209]
14. Abdelsattar JM, Pandian TK, Finnesgard EJ, El Khatib MM, Rowse PG, Buckarma ENH, et al. Do you see what I see? how we use video as an adjunct to general surgery resident education. J Surg Educ 2015;72(6):e145-e150. [doi: 10.1016/j.jsurg.2015.07.012] [Medline: 26454723]

XSL•FO
RenderX

15. Mentis HM, Chellali A, Manser K, Cao CGL, Schwaitzberg SD. A systematic review of the effect of distraction on surgeon performance: directions for operating room policy and surgical training. Surg Endosc 2016;30(5):1713-1724 [FREE Full text] [doi: 10.1007/s00464-015-4443-z] [Medline: 26194261]
16. Lai M, Skyrman S, Shan C, Babic D, Homan R, Edström E, et al. Fusion of augmented reality imaging with the endoscopic view for endonasal skull base surgery; a novel application for surgical navigation based on intraoperative cone beam computed tomography and optical tracking. PLoS One 2020;15(1):e0227312 [FREE Full text] [doi: 10.1371/journal.pone.0227312] [Medline: 31945082]
17. Agha R, Abdall-Razak A, Crossley E, Dowlut N, Iosifidis C, Mathew G, STROCSS Group. STROCSS 2019 guideline: strengthening the reporting of cohort studies in surgery. Int J Surg 2019;72:156-165 [FREE Full text] [doi: 10.1016/j.ijsu.2019.11.002] [Medline: 31704426]
18. Bretonnier M, Michinov E, Morandi X, Riffaud L. Interruptions in surgery: a comprehensive review. J Surg Res 2020;247:190-196 [FREE Full text] [doi: 10.1016/j.jss.2019.10.024] [Medline: 31706542]
19. Keogh S, Laski D. A concern for intraoperative distractions and interference: an observational study identifying, measuring, and quantifying both within the operating theatre. Surg Res Pract 2021;2021:9910290 [FREE Full text] [doi: 10.1155/2021/9910290] [Medline: 34931175]
20. Sevdalis N, Undre S, McDermott J, Giddie J, Diner L, Smith G. Impact of intraoperative distractions on patient safety: a prospective descriptive study using validated instruments. World J Surg 2014;38(4):751-758. [doi: 10.1007/s00268-013-2315-z] [Medline: 24240670]
21. Zheng B, Martinec DV, Cassera MA, Swanström LL. A quantitative study of disruption in the operating room during laparoscopic antireflux surgery. Surg Endosc 2008;22(10):2171-2177. [doi: 10.1007/s00464-008-0017-7] [Medline: 18622567]
22. García-Pérez V, Muñoz-Moreno E, Aja-Fernández S, Alberola-López C. A 3-D collision handling algorithm for surgery simulation based on feedback fuzzy logic. IEEE Trans Inf Technol Biomed 2009;13(4):451-457. [doi: 10.1109/TITB.2009.2016838] [Medline: 19278931]
23. Randhawa N. Preventing instrument clashing in multi-quadrant surgery. In: Coyne P, Khan J, editors. Robotic Colorectal Surgery: Complete Manual of Surgical Techniques. Cham: Springer; 2022:27-30.
24. Kavitha JC, Suruliandi A. Feature extraction using dominant local texture-color patterns (DLTCP) and classification of color images. J Med Syst 2018;42(11):220. [doi: 10.1007/s10916-018-1067-6] [Medline: 30280254]
25. De La Torre AB, Joe S, Lee VS. An evaluation of YouTube videos as a surgical instructional tool for endoscopic endonasal approaches in otolaryngology. Ear Nose Throat J 2021:1455613211062447 [FREE Full text] [doi: 10.1177/01455613211062447] [Medline: 34894801]
26. Wiedmann M, Lashkarivand A, Berg-Johnsen J, Dahlberg D. How I do it: endoscopic endonasal resection of tuberculum sellae meningioma. Acta Neurochir 2021;163(8):2193-2197 [FREE Full text] [doi: 10.1007/s00701-021-04784-5] [Medline: 33665730]
27. Liang SF, Shih YH, Chen PY, Kuo CE. Development of a human-computer collaborative sleep scoring system for polysomnography recordings. PLoS One 2019;14(7):e0218948 [FREE Full text] [doi: 10.1371/journal.pone.0218948] [Medline: 31291270]
28. Terveen LG. Overview of human-computer collaboration. Knowl Based Syst 1995;8(2-3):67-81. [doi: 10.1016/0950-7051(95)98369-h]
29. Lalys F, Riffaud L, Morandi X, Jannin P. Automatic phases recognition in pituitary surgeries by microscope images classification. In: Navab N, Jannin P, editors. Information Processing in Computer-Assisted Interventions. Berlin, Heidelberg: Springer; 2010:34-44.
30. Schlick CJR, Bilimoria KY, Stulberg JJ. Video-based feedback for the improvement of surgical technique: a platform for remote review and improvement of surgical technique. JAMA Surg 2020;155(11):1078-1079. [doi: 10.1001/jamasurg.2020.3286] [Medline: 32902619]
31. Qureshi T, Chaus F, Fogg L, Dasgupta M, Straus D, Byrne RW. Learning curve for the transsphenoidal endoscopic endonasal approach to pituitary tumors. Br J Neurosurg 2016;30(6):637-642 [FREE Full text] [doi: 10.1080/02688697.2016.1199786] [Medline: 27332969]
32. Yan JL, Chang CN, Chen PY. Endoscopic transsphenoidal surgery for resection of pituitary macroadenoma: a retrospective study. PLoS One 2021;16(8):e0255599 [FREE Full text] [doi: 10.1371/journal.pone.0255599] [Medline: 34358251]
33. Hollon TC, Parikh A, Pandian B, Tarpeh J, Orringer DA, Barkan AL, et al. A machine learning approach to predict early outcomes after pituitary adenoma surgery. Neurosurg Focus 2018;45(5):E8. [doi: 10.3171/2018.8.FOCUS18268] [Medline: 30453460]
34. Castle-Kirszbaum M, Wang YY, King J, Goldschlager T. Predictors of visual and endocrine outcomes after endoscopic transsphenoidal surgery for pituitary adenomas. Neurosurg Rev 2022;45(1):843-853. [doi: 10.1007/s10143-021-01617-y] [Medline: 34618249]
35. Niederhauser L, Gunser S, Waser M, Mast FW, Caversaccio M, Anschuetz L. Training and proficiency level in endoscopic sinus surgery change residents' eye movements. Sci Rep 2023;13(1):79 [FREE Full text] [doi: 10.1038/s41598-022-25518-2] [Medline: 36596830]

XSL•FO
RenderX

36. Lamsal R, Panda NB, Wig J. Effect of tranexamic acid on blood loss and the quality of surgical field in transsphenoidal pituitary surgeries: double-blind placebo-controlled randomized control trial. Neurol India 2022;70(3):960-964 [FREE Full text] [doi: 10.4103/0028-3886.349622] [Medline: 35864625]
37. Kim DH, Hong YK, Jeun SS, Park JS, Kim SW, Cho JH, et al. Is septoplasty necessary when using the endoscopic endonasal transsphenoidal approach for a deviated nasal septum? J Neurol Surg B Skull Base 2018;79(6):569-573 [FREE Full text] [doi: 10.1055/s-0038-1641602] [Medline: 30456026]
38. Cheng MZ, Saraswathula A, Qureshi HA, Mukherjee D, Rowan NR. Otolaryngology considerations of pituitary surgery: what an endocrinologist should know. J Endocr Soc 2023;7(7):1-8 [FREE Full text] [doi: 10.1210/jendso/bvad058] [Medline: 37313347]

Abbreviations

AI: artificial intelligence
CV: computer vision
EEA: endonasal endoscopic approach
PA: pituitary adenoma
STROCSS: Strengthening the Reporting of Cohort Studies in Surgery

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 07.01.24; peer-reviewed by C Ordun, F Zhu; comments to author 12.03.24; revised version received 27.03.24; accepted 24.05.24; published 04.07.24.

Please cite as:
Wong CE, Chen PW, Hsu HJ, Cheng SY, Fan CC, Chen YC, Chiu YP, Lee JS, Liang SF
Collaborative Human–Computer Vision Operative Video Analysis Algorithm for Analyzing Surgical Fluency and Surgical Interruptions in Endonasal Endoscopic Pituitary Surgery: Cohort Study
J Med Internet Res 2024;26:e56127
URL: https://www.jmir.org/2024/1/e56127
doi:10.2196/56127
PMID:38963694

©Chia-En Wong, Pei-Wen Chen, Heng-Jui Hsu, Shao-Yang Cheng, Chen-Che Fan, Yen-Chang Chen, Yi-Pei Chiu, Jung-Shun Lee, Sheng-Fu Liang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Electronic Health Record–Oriented Knowledge Graph System for Collaborative Clinical Decision Support Using Multicenter Fragmented Medical Data: Design and Application Study

Yong Shang[1*], PhD; Yu Tian[2,3*], PhD; Kewei Lyu[2,3], PhD; Tianshu Zhou[1], PhD; Ping Zhang[4], MD, PhD; Jianghua Chen[4], MD, PhD; Jingsong Li[1], PhD

[1]Research Center for Data Hub and Security, Zhejiang Laboratory, Hangzhou, China

[2]Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[3]Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[4]Kidney Disease Center, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Jingsong Li, PhD
Research Center for Data Hub and Security
Zhejiang Laboratory
No.1 Kechuang Avenue, Zhongtai Sub-District, Yuhang District
Hangzhou, 310000
China
Phone: 86 0571 58005162
Email: ljs@zju.edu.cn

Abstract

Background: The medical knowledge graph provides explainable decision support, helping clinicians with prompt diagnosis and treatment suggestions. However, in real-world clinical practice, patients visit different hospitals seeking various medical services, resulting in fragmented patient data across hospitals. With data security issues, data fragmentation limits the application of knowledge graphs because single-hospital data cannot provide complete evidence for generating precise decision support and comprehensive explanations. It is important to study new methods for knowledge graph systems to integrate into multicenter, information-sensitive medical environments, using fragmented patient records for decision support while maintaining data privacy and security.

Objective: This study aims to propose an electronic health record (EHR)–oriented knowledge graph system for collaborative reasoning with multicenter fragmented patient medical data, all the while preserving data privacy.

Methods: The study introduced an EHR knowledge graph framework and a novel collaborative reasoning process for utilizing multicenter fragmented information. The system was deployed in each hospital and used a unified semantic structure and Observational Medical Outcomes Partnership (OMOP) vocabulary to standardize the local EHR data set. The system transforms local EHR data into semantic formats and performs semantic reasoning to generate intermediate reasoning findings. The generated intermediate findings used hypernym concepts to isolate original medical data. The intermediate findings and hash-encrypted patient identities were synchronized through a blockchain network. The multicenter intermediate findings were collaborated for final reasoning and clinical decision support without gathering original EHR data.

Results: The system underwent evaluation through an application study involving the utilization of multicenter fragmented EHR data to alert non-nephrology clinicians about overlooked patients with chronic kidney disease (CKD). The study covered 1185 patients in nonnephrology departments from 3 hospitals. The patients visited at least two of the hospitals. Of these, 124 patients were identified as meeting CKD diagnosis criteria through collaborative reasoning using multicenter EHR data, whereas the data from individual hospitals alone could not facilitate the identification of CKD in these patients. The assessment by clinicians indicated that 78/91 (86%) patients were CKD positive.

Conclusions: The proposed system was able to effectively utilize multicenter fragmented EHR data for clinical application. The application study showed the clinical benefits of the system with prompt and comprehensive decision support.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e54263) doi:10.2196/54263

KEYWORDS

knowledge graph; electronic health record; ontology; data fragmentation; data privacy; knowledge graphs; visualization; ontologies; data science; privacy; security; collaborative; collaboration; kidney; CKD; nephrology; EHR; health record; hypernym; encryption; encrypt; encrypted; decision support; semantic; vocabulary; blockchain

Introduction

The fragmentation of patient data across multiple hospitals adversely impacts health care quality. In practice, patients visit different hospitals for various medical services. Previous studies have indicated that up to 26.5% of patients from a hospital have visited other institutions in the past 12 months [1-3]. These visits result in fragmented patient data across different hospitals within various electronic health record (EHR) systems [2,4-6]. Because of the sensitivity of medical data, sharing information between hospitals encounters obstacles related to data privacy and security [7]. As a result, EHR data in each hospital are often incomplete, making collaboration difficult.

The missing information from outside the local institution cannot provide clinicians with complete clinical evidence during routine practice. This potentially affects decision-making and harms health care quality in several aspects [2,8-10]:

- Delayed or missed diagnoses: Missing evidence from other hospitals can lead to unconsidered diseases until apparent symptoms occur.
- Duplicate care or additional tests: Additional tests may be ordered to verify diagnoses, even though records in other hospitals might already contain the needed results.
- Incomprehensive analysis and decisions: Because of incomplete disease history, clinicians may neglect important risk factors during decision-making.

The missing information could adversely affect patients in nearly 50% of cases. Much of the needed data can be found outside local hospitals [9]. Thus, utilizing multicenter fragmented EHR data for comprehensive decision support is essential while maintaining data privacy.

The knowledge graph stands as an explainable artificial intelligence method applicable across numerous domains. Using knowledge graphs to enhance semantic relationships within EHR data and execute deductive reasoning aids in producing understandable results within clinical practice [11]. Recent research on EHR-based knowledge graphs highlights the benefits of integrating medical knowledge into clinical applications (Textbox 1).

Textbox 1. Benefits of integrating medical knowledge into clinical applications.

1. Generating medical knowledge graphs from electronic health record data

- Li et al [12] have introduced systematic methodologies for the semiautomatic construction of medical knowledge graphs using electronic health record (EHR) data. Entity recognition and occurrence-based algorithms play pivotal roles in relation to extraction and ranking.
- Hong et al [13] have introduced a clinical knowledge extraction technique using sparse embedding regression with multicenter EHR data. The embedding vectors derived from multicenter EHR data enhance the robustness of knowledge and facilitate the identification of data set heterogeneity.

2. Knowledge graph–based EHR query

- Thukral et al [14] have pioneered a method to convert tabular format EHR data into a knowledge graph representation, thereby enriching the semantic relationships among EHR data elements. This approach enables the execution of complex data queries using the easily interpretable SPARQL language.
- Xiao et al [15] used Ontology-Based Data Access to establish a virtual fast health care interoperability resources–based knowledge graph derived from Observational Medical Outcomes Partnership (OMOP) EHR data. This approach facilitates data interoperability with exceptional efficiency and generality.
- Although these studies greatly enhance data interoperability, they do not incorporate decision support functions.

3. EHR knowledge graph–based clinical decision support

- Carvalho et al [16] integrated EHR data with knowledge graph embeddings to develop a machine learning model for predicting intensive care unit readmissions. The knowledge embedded within EHR data serves as a feature for model training, resulting in improved predictive performance.
- Liu et al [17] introduced a heterogeneous similarity graph neural network approach for diagnosing predictions based on graph-formatted EHR data. The heterogeneous EHR graph undergoes normalization into multiple homogeneous graphs, which are then fused into a graph neural network to enhance prediction accuracy.

Although the methods described in Textbox 1 have shown enhanced performance, the model lacks both explainability and generality. In our previous studies, we introduced an EHR-oriented knowledge graph system, leveraging medical information often overlooked or underutilized by clinicians. This system aimed to offer decision support for diseases spanning multiple departments [18]. Specifically, it aided nonnephrology clinicians in identifying patients at risk of chronic kidney disease (CKD) who had been overlooked for extended periods. By transcending traditional knowledge barriers, the system tapped into previously underutilized

XSL•FO
RenderX

information to facilitate the early detection of diseases crossing departmental boundaries.

Our previous work focused solely on cross-departmental data within a single hospital. However, in real-world scenarios, fragmented patient records within a single hospital often fail to provide sufficient information for knowledge graphs to conduct comprehensive analyses [10,19,20]. This limitation can result in imprecise decisions or delayed diagnoses, thereby constraining the practical implementation of knowledge graphs. Currently, only a handful of studies on multicenter knowledge graphs address the challenge of data fragmentation during model application phases. The predominant focus of multicenter knowledge graph research lies in 3 key areas: constructing knowledge graphs from diverse sources, completing knowledge graphs using data from multiple centers, and facilitating data interoperability guided by knowledge graphs [21-24]. Challenges in multicenter knowledge graphs encompass data heterogeneity, knowledge inconsistency, and concerns regarding the privacy and security of data sources. To address these challenges, researchers have explored federated knowledge graph embedding methods, allowing model training with multicenter data while upholding data security. For instance, Chen et al [25] introduced FedE, a knowledge graph embedding method leveraging a federated learning framework. In these approaches, each data source learns embedding vectors using its local data and then shares these vectors for model iteration. Peng et al [26] introduced FKGE (Federated Knowledge Graphs Embedding), which enables the learning of embeddings from various knowledge graphs in an asynchronous and peer-to-peer manner while safeguarding privacy. These methods have demonstrated improved performance in link prediction tasks without necessitating the centralization of original data. The studies, however, used multicenter data solely during the training phase of the embedding model and did not address data fragmentation during application. When applied to real-world decision support scenarios, these models are still fed with fragmented patient data from single hospitals only, which can significantly impact the performance of otherwise well-trained models. Therefore, it is crucial to empower knowledge graph systems to leverage multicenter fragmented EHR data for CDS while ensuring the preservation of data privacy, particularly in chronic disease management and long-term decision support applications.

Collaborative research networks such as the Observational Health Data Sciences and Informatics (OHDSI) offer valuable insights into addressing this issue [27,28]. These networks use local analysis results from various institutions, aggregating models, or summarizations to enhance generalizability and mitigate bias. Collaborative research does not necessarily require the centralization of original data, thus ensuring data security. Axfors et al [29] and Baigent et al [30] showcased meta-analyses by incorporating results from multiple randomized control trials and providing additional insights through summarization. Noman et al [31] and Tian et al [32] introduced collaborative methods using federated learning and multivariate aggregation to enhance model accuracy and generalizability. Such collaborative research only necessitates the collection of local analysis results and has demonstrated significant clinical value through the utilization of multicenter data sources.

Taking inspiration from collaborative research, this study introduces an EHR-oriented knowledge graph system designed to effectively harness multicenter fragmented patient EHR data while safeguarding data security. Implemented within each hospital, the system conducts local reasoning based on local EHR data and generates intermediate reasoning results. Through a distribution module, the intermediate results are formulated as an online subgraph with encrypted identities, facilitating multicenter collaboration and alignment. A blockchain network synchronizes the findings across centers, and the collaborated patient clinical evidence is used for final reasoning, enabling comprehensive CDS. Importantly, the original data remain within the local institute to uphold data security. The main contribution of the study is as follows:

- Introducing a novel framework for multicenter collaborative reasoning using fragmented EHR data for comprehensive CDS without the need to share original data. This approach enables knowledge graphs to use intact evidence for CDS purposes.
- Implementing an EHR-oriented knowledge graph system across multiple hospitals to standardize local EHR data and facilitate the local reasoning process. This initiative establishes a standardized semantic environment conducive to multicenter collaborative reasoning.
- Developing a distribution component and online subgraph structure to facilitate the collaboration of intermediate reasoning findings across multiple centers. This initiative addresses data privacy concerns and enhances local systems with the capability for multicenter collaboration.

An application study was conducted to evaluate the system's effectiveness in assisting clinicians in detecting undiagnosed CKD in patients who visited multiple hospitals. The system successfully issued timely CKD warnings, a capability not supported by data from a single hospital alone.

Methods

EHR-Oriented Knowledge Graph System for Multicenter Collaboration

Overall System Architecture

This study presents an EHR-oriented knowledge graph system designed for multicenter collaboration using fragmented patient information. The overall system architecture is depicted in Figure 1. The proposed system uses structured EHR data following the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) for semantic reasoning and clinical applications [33]. The semantic organization of EHR data within the knowledge graph adheres to the structure outlined in the OMOP CDM. The system consists of 3 main components: (1) the local EHR knowledge graph component, which conducts semantic reasoning on local EHR data to generate independent clinical findings; (2) the distribution component, which manages the distribution of intermediate reasoning results and patient alignment for multicenter collaboration; and (3) the blockchain component, which establishes a secure network for multicenter synchronization.

XSL•FO
RenderX

The system is deployed in hospitals, where it conducts local reasoning on local EHR data. The distribution component and blockchain network collaborate on intermediate reasoning results without exposing original data for privacy concerns. Subsequently, fragmented patient information from multiple hospitals is used to generate comprehensive CDS with complete clinical evidence.

Figure 1. The architecture of the EHR-oriented knowledge graph system. (1) The local EHR knowledge graph system performs local reasoning based on local EHR data. (2) The distribution component creates online subgraphs with intermediate reasoning findings to synchronize across hospitals without sharing original EHR data. (3) The blockchain network supports the collaborative process. CDS: clinical decision support; CP-ABE: ciphertext-policy attribute–based encryption. EHR: electronic health record; OMOP: Observational Medical Outcomes Partnership; RDF: resource description framework.

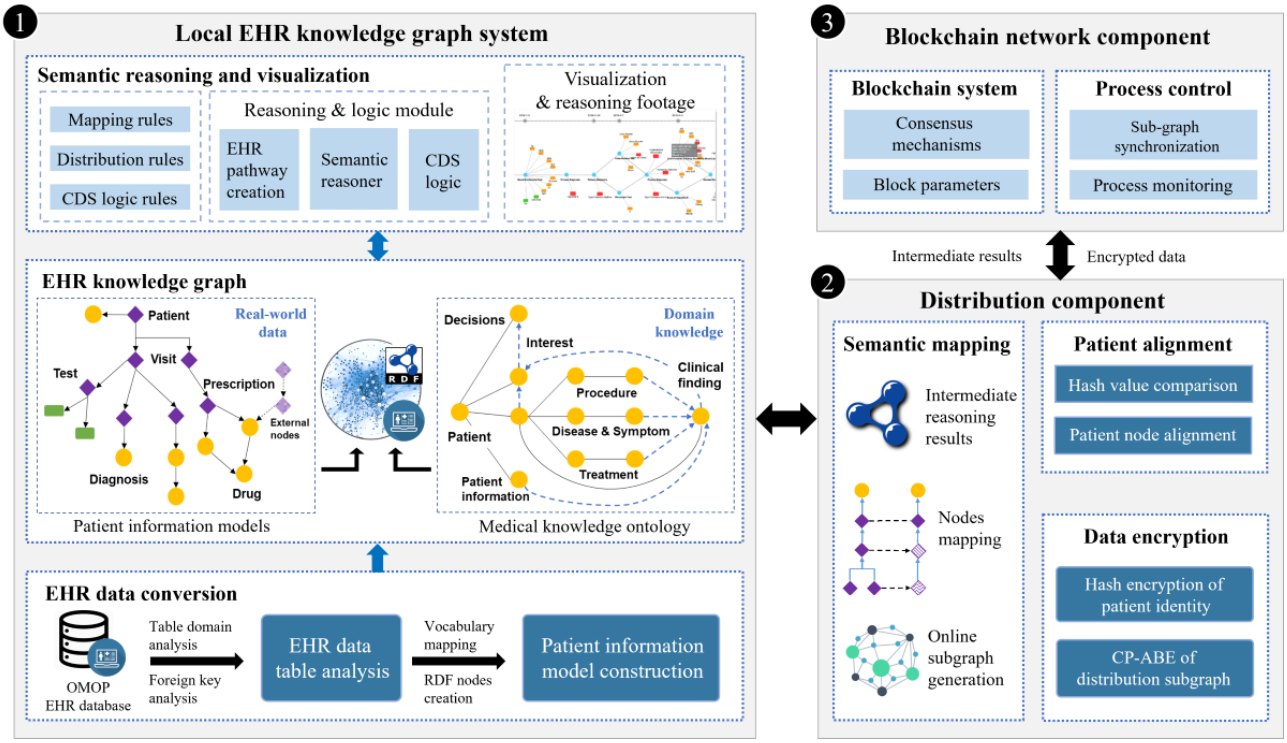

The Local EHR Knowledge Graph Component

The local EHR knowledge graph component offers the capability to leverage local EHR data for semantic reasoning and CDS generation. This component has been adapted from our previous study [18]. The EHR Data Conversion Module is responsible for transforming EHR data into resource description framework (RDF)–type triples to enable semantic querying and reasoning. The module conducts an analysis of the EHR database and aligns table concepts with the entities within the knowledge graph ontology. Within the EHR knowledge graph, EHR data and clinical knowledge entities undergo a semantic transformation, organized under a unified top-level ontology structure. Semantic triples are used within the EHR knowledge graph to represent the clinical information pertaining to each patient. The Semantic Reasoning Module offers rule-based reasoning capabilities on the local knowledge graph to generate CDS-related findings. Additionally, the module establishes an EHR pathway for each patient to facilitate the collaboration of multicenter information. This involves connecting intermediate findings from multiple centers along a virtual timeline using semantic relationships, ultimately contributing to the final reasoning process for CDS. The Visualization and Explanation Module furnishes clinicians with a visualized timeline, aiding in their comprehension of critical medical information and evidence pertinent to the CDS.

To support the collaboration of local reasoning results, the generated findings are transformed into hypernym concept expressions to isolate original EHR data (eg, using “abnormal blood potassium” to represent reasoning findings from hyperkalemia diagnosis, blood potassium test results, or treatment medicines). Other hospitals will learn about the abnormality but not the original examination data or prescriptions. During the construction of the knowledge graph, candidate hypernym concepts are selected based on hierarchical relationships and the cosine similarity of their leaf nodes. Clinical experts review and adjust these candidate hypernym concepts to ensure information accuracy. The identified findings are automatically transformed into hypernym expressions during the knowledge graph reasoning process. Additional technical details are in Multimedia Appendix 1.

The Distribution Component

The distribution component facilitates collaborative reasoning between multicenter EHR knowledge graph systems. It extracts intermediate reasoning results and visit pathway information, encrypts patient identities, and builds an online subgraph to synchronize local patient findings with other institutions. As shown in Figure 2, patient identities are hash encrypted for multicenter patient alignment. The hash codes of patient identities are compared to align the same patient across different hospitals, allowing multicenter findings for the same patient to be matched. Medical data nodes are prohibited from being

XSL•FO
RenderX

distributed and are not used for online subgraph construction. Intermediate reasoning findings, represented by hypernym concepts, are extracted to build the online subgraph. This allows the transfer of a patient's clinical evidence without exposing original EHR data. Other hospitals receive the online subgraph to collaborate on multicenter clinical evidence by loading intermediate findings of the aligned patients. If access control is required, the online subgraph can be encrypted using ciphertext policy attribute–based encryption.

Figure 2. Data access of the Distribution Component. The original EHR data and identities are not acquired for online subgraph creation. The encrypted identity, visit time stamps, and authorized findings are permitted to be constructed as online subgraphs for collaborative reasoning. EHR: electronic health record.

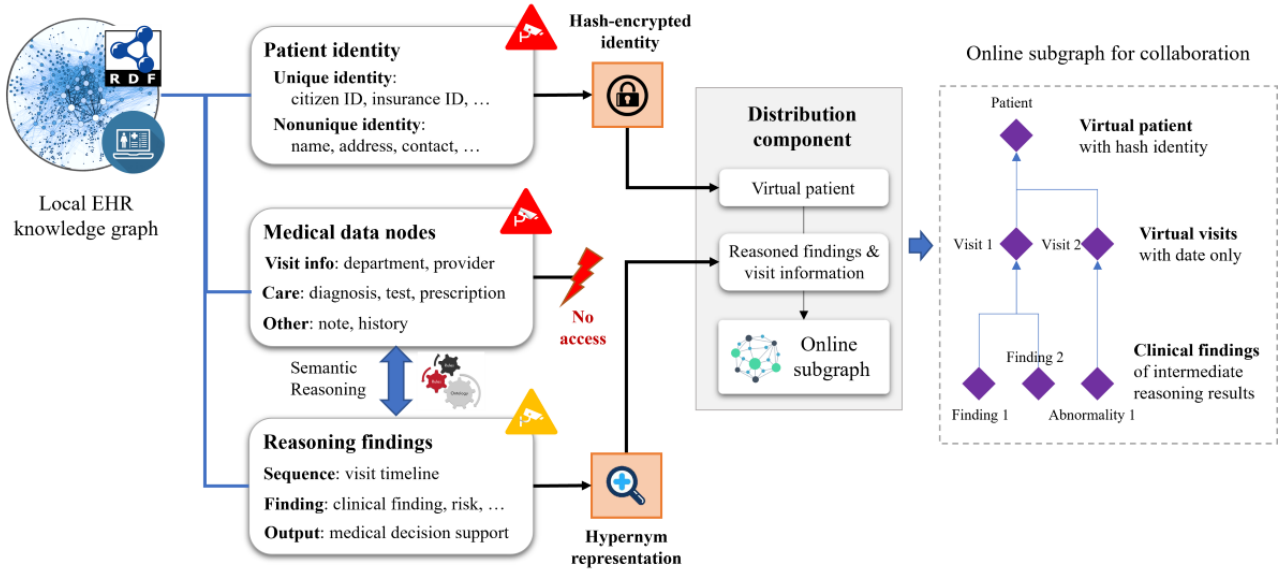

The Blockchain Network Component

The blockchain network component establishes a blockchain node and manages the blockchain network. It securely synchronizes locally generated online subgraphs with other systems and delivers acquired triples from other hospitals to the local system, supporting the collaborative reasoning process. A requirement of the collaborative reasoning process is broadcast through the blockchain network, allowing each node to receive the process's series number. The blockchain method was chosen because it is a proven approach for synchronizing data in distributed systems and is already used in medical domain studies [34,35]. All actions on the blockchain are logged and traceable. In this study, we used Golang [36] and libp2p [37] for blockchain platform construction, with proof of stake as the consensus mechanism.

Multicenter Collaboration Settings of the EHR-Oriented Knowledge Graph System

Deployment Overview

The EHR knowledge graph systems are deployed in each hospital and connected through a blockchain network for collaborative reasoning. Figure 3 illustrates the multicenter collaboration setting of the system. The EHR knowledge graph system is implemented in local hospitals and uses local EHR data sets for reasoning without exposing the original EHR data. Participating hospitals generate intermediate reasoning results, represented by hypernym concepts to isolate them from the original data, and use hash-encrypted identities to build online subgraphs. The sponsoring hospital receives the online subgraphs via the blockchain network and conducts patient alignment by comparing identity hash codes. For every patient, a comprehensive clinical pathway is established by amalgamating local evidence and intermediate reasoning outcomes from participating hospitals. A conclusive summary reasoning process utilizes the entirety of patient data to furnish clinicians with comprehensive CDS. Throughout this process, the original EHR data remain preserved within the local systems.

XSL•FO
RenderX

Figure 3. Setting of the EHR knowledge graph system in the multicenter environment. The participating hospitals perform local reasoning and pass the intermediate reasoning results through generated subgraphs. The sponsoring hospital performs local reasoning based on local EHR data and acquired intermediate findings to generate a comprehensive CDS for application. CDS: clinical decision support; EHR: electronic health record; KG: knowledge graph.

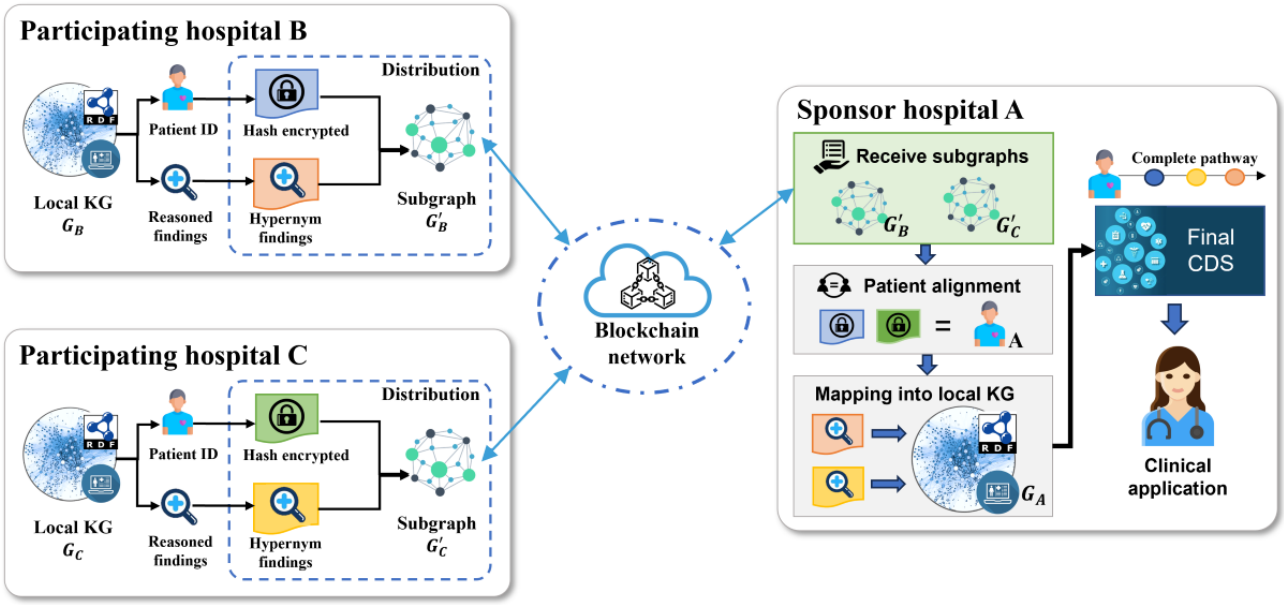

Patient Information Model

The EHR data within an OMOP CDM–based data table undergo transformation into RDF-type triples, thereby adopting a patient-centric information model suitable for querying and semantic reasoning. The structure of this patient information model is depicted on the left side of Figure 4. It is a 3-level, patient-visit-treatment semantic structure. It models each patient's EHR data into a semantic clinical trajectory, facilitating patient-level querying and reasoning. The patient information model transmutes table-based EHR data into semantic graphs, enabling semantic reasoning, with each data element linking to its corresponding knowledge nodes.

Figure 4. The semantic structure of the RDF-type patient EHR data. In the local EHR knowledge graph, a patient-visit-treatment structure defines the semantic structure of EHR information. In the online subgraph, the structure contains only patient nodes with hash identity, virtual visit nodes with visit dates, and virtual finding nodes with finding types and positive labels. EHR: electronic health record; RDF: resource description framework.

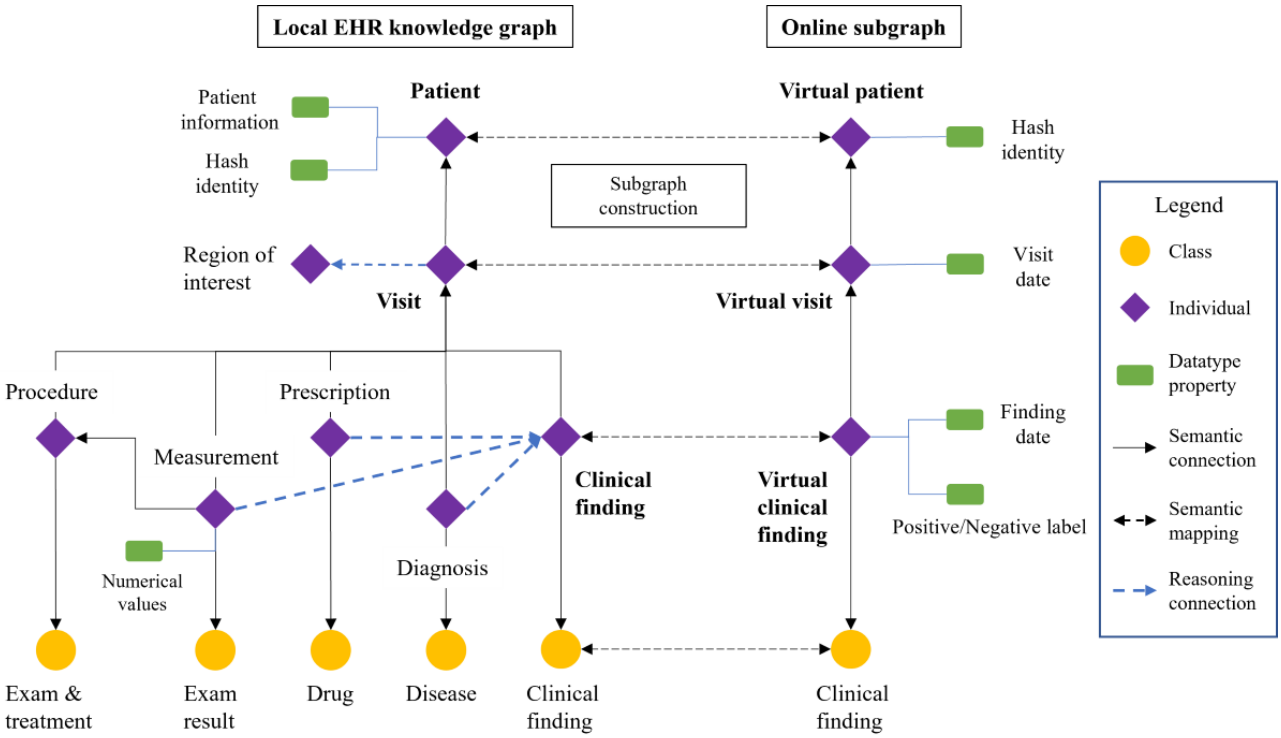

XSL•FO
RenderX

The Online Subgraph

The online subgraph serves as a streamlined patient information model designed for online synchronization among multiple EHR knowledge graph systems. The ontology structure of the online subgraph is illustrated in the right section of Figure 4. The entities within the online subgraph mirror those within the local EHR knowledge graph, focusing solely on information pertinent to collaborative reasoning to conserve network resources. Each patient entity comprises solely hash-encrypted identity values for patient alignment. Virtual visit nodes exclusively feature visit dates to denote visit records from other hospitals. Similarly, virtual clinical finding nodes harbor intermediate reasoning outcomes tailored for collaborative reasoning purposes.

The Multicenter Collaborative Reasoning Process

Purpose

The multicenter collaborative reasoning process delineates a systematic interaction protocol for multicenter systems to engage in collaborative CDS reasoning. Illustrated in Figure 5, the process encompasses multiple steps, including preparing the reasoning cohort, aligning patients, defining the reasoning data period, and conducting semantic reasoning. This procedural framework ensures the efficacy and efficiency of collaborative reasoning endeavors.

Figure 5. The overall process of collaborative reasoning. (1) All the centers identify patients meeting the cohort entry criteria. (2) The sponsoring center aligns patients by hash-encrypted identities. (3) The system creates a complete visit pathway with cohort entry findings for each patient to determine the ROIs for further reasoning. (4) Each center performs semantic reasoning and generates intermediate findings based on local EHR data. (5) The intermediate reasoning findings from multiple hospitals are collaborated for final decisions. CDS: clinical decision support; EHR: electronic health record; ROI: region of interest.

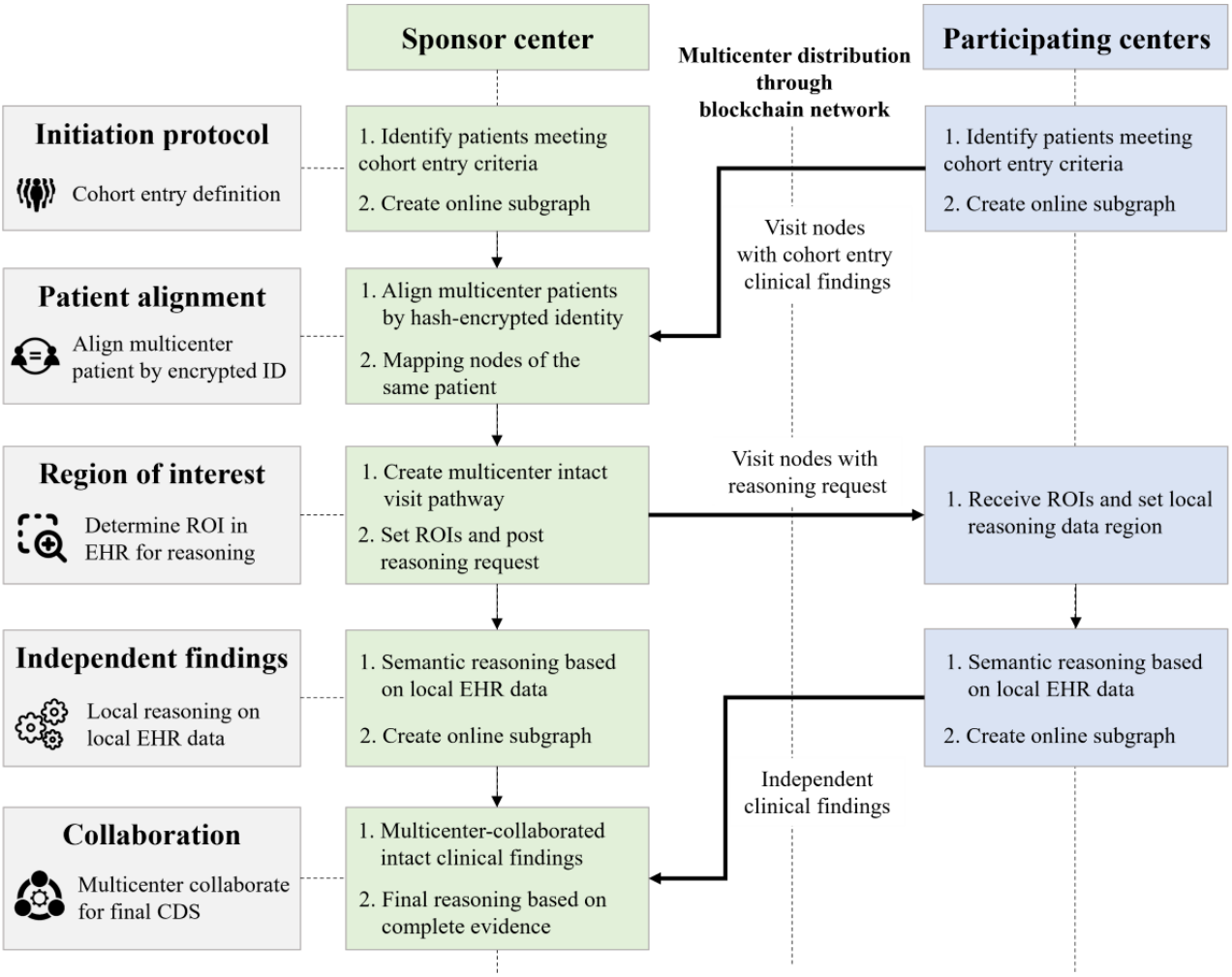

Initiation Protocol

To initiate a collaborative reasoning process, the local system first executes the initiation protocol as a preparatory step. This protocol is tailored for the local system to delineate the reasoning cohort. The process identifies initial clinical evidence to determine whether a patient needs to join the collaborative reasoning. Unrelated patients are ruled out to save resources. The initiation protocol varies from one disease to another.

For example, in an application concerning unconsidered CKD warnings based on multicenter EHR data, the initiation protocol mandates each local system to ascertain whether a patient possesses kidney function test results but has not undergone nephrology visits. If this criterion is met, the patient becomes eligible to participate in the collaborative reasoning process. Subsequently, the local system generates virtual visit nodes containing abnormal kidney function findings for synchronization and collaborative analysis.

XSL•FO
RenderX

Patient Alignment

The patient alignment process is facilitated by the distribution component of the system. Patients are aligned by comparing hash-encrypted identities between local patient nodes and the received online subgraphs. Sensitive patient identities undergo hash encryption, ensuring that matching can be accomplished solely using ciphertext. This approach prevents the leakage of privacy information during data transmission and alignment. The distribution component maintains a table that records the mappings between matched patient node uniform resource identifiers. The records within the table serve as the foundation for mapping intermediate findings in subsequent steps. Subsequently, the reasoning outcomes pertaining to the matched patient are integrated into the local EHR knowledge graph system to facilitate collaborative reasoning.

Region of Interest Designation

The region of interest (ROI) delineates the disease-related observation period within the patient's EHR data, specifying a temporal window for collaborative reasoning. The sponsoring hospital consolidates the results of the initiation protocol reasoning to identify observation periods relevant to disease risks. Collaborative reasoning concentrates on EHR data within this designated period to ensure efficient analysis and excludes irrelevant noise information.

The sponsoring hospital initially obtains visits and clinical findings related to the initiation protocol from other hospitals. Subsequently, the local EHR knowledge graph system constructs a multicenter visit timeline incorporating significant clinical findings. Following this, the system engages in semantic reasoning to ascertain the periods during which the patient's data are pertinent to the disease and require additional collaborative reasoning for CDS. The ROIs are delineated and transmitted to other hospitals. Any nonrelated periods are subsequently excluded to optimize efficiency and conserve computing resources.

Local Reasoning Process

Throughout the collaborative reasoning process, the local reasoning process assumes responsibility for leveraging the local EHR data within the local knowledge graph. Its primary task involves generating atomic, independent clinical findings conducive to synchronization. The generated clinical findings exclusively present conclusions derived from EHR data. Intermediate findings use higher-level concepts to extract reasoning results from the original data, with the selection of these concepts being determined by domain experts and medical professionals. Other hospitals solely receive intermediate reasoned conclusions and are not provided with the corresponding original data or informed about the methodology used to derive the findings. For instance, the identification of abnormal blood potassium might be reasoned based on hyperkalemia, measurement results, or treatment medications, yet the other hospitals remain unaware of the specific source behind the reasoned findings.

The Reasoning Module conducts rule-based semantic reasoning, analyzing diagnoses, medical test results, procedures, and prescriptions separately at each visit to produce independent clinical findings. These findings are then converted into hypernym representation, aligning with semantic relationships within the knowledge graph, to facilitate multicenter collaboration. For instance, a <abnormal kidney function> node might represent a <estimated glomerular filtration rate at g3b stage> node within the knowledge graph.

The Multicenter Distribution and Summarization

The sponsoring hospital conducts summarization reasoning, drawing upon intact clinical evidence to generate final CDS responses. An illustration outlining the distribution and summarization process is provided in Figure 6.

The online subgraphs transmit intermediate reasoning results to the sponsoring hospital for collaborative result synthesis. Upon receiving the subgraphs, the sponsoring hospital maps the incoming visits and clinical findings as virtual visits and virtual clinical findings, respectively. This process culminates in the creation of an integrated medical pathway for each patient, all achieved without the necessity of gathering original medical data. The system engages in semantic reasoning using multicenter collaborated reasoning outcomes. Collaborating on intermediate findings furnishes comprehensive patient clinical evidence, empowering the knowledge graph to produce precise decision support. The resulting CDS responses are presented in a timeline format, accompanied by explanatory reasoning details, allowing clinicians to review and interpret the information effectively.

In instances where decision support necessitates evidence beyond the scope of semantic reasoning, the system offers an interface to interact with other nonreasoning protocols to obtain the requisite evidence. For instance, in the application study, the garbled circuit algorithm is used to compare 2 test results without revealing the actual numerical values [38]. The protocol incorporates its own security mechanism to generate clinical findings in a data-private manner.

XSL•FO
RenderX

Figure 6. (A) Local systems generate intermediate reasoning results and collaborate by a blockchain network. (B) The multicenter collaborated findings support CDS reasoning and provide explainable results to clinicians. CDS: clinical decision support; EHR: electronic health record.

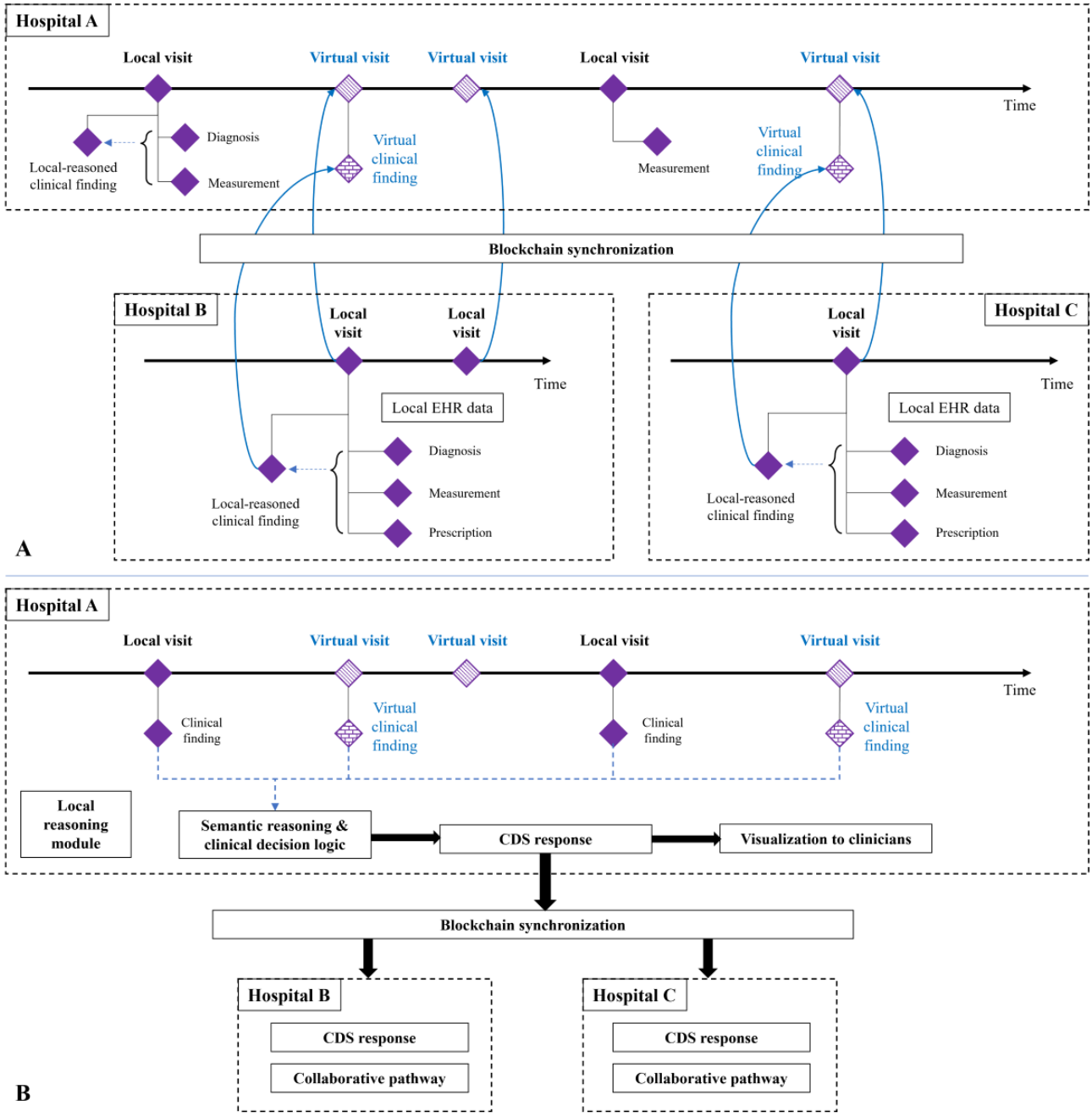

Application to Unconsidered CKD Detection Through Fragmented EHR Data

Background

An application study assesses the performance and clinical value of the proposed system. The EHR knowledge graph system conducts collaborative reasoning on patients' fragmented EHR data to identify their CKD-related risks, which were challenging to recognize using data from individual hospitals alone.

CKD is a prevalent chronic disorder associated with various complications and has seen a significant increase in prevalence in recent decades [39]. Epidemiological research indicates that the prevalence of CKD in China stands at 10.8%, yet only 12.5% of impacted individuals are aware of their condition [40]. The early detection of CKD relies on nonnephrology clinicians; however, it is particularly challenging because early-stage CKD often exhibits fewer symptoms. However, insufficient CKD knowledge among nonnephrology clinicians may result in the oversight of CKD-related risks during routine practice. Moreover, the extended monitoring required for the chronic progression of abnormal kidney function presents challenges for clinicians in timely identifying CKD [41,42]. Patients frequently seek care at multiple hospitals or clinics over a period, resulting in fragmented renal function test results and a disjointed disease history spread across different institutions. For nonnephrology clinicians with access solely to single-center data, identifying overlooked CKD becomes challenging. This situation can lead to delayed diagnosis, the necessity for repeated tests, and potentially inappropriate treatment. Combining

XSL•FO
RenderX

fragmented test results and clinical findings while ensuring data security can facilitate the timely identification of CKD [43].

Application Study Design

In the application study, our system was deployed across 3 tertiary A-level hospitals in Hangzhou: the First Affiliated Hospital, College of Medicine, Zhejiang University (FAHZU); Zhejiang Hospital; and the Affiliated Hospital of Hangzhou Normal University (AHHNU). FAHZU is a comprehensive hospital providing a wide range of general care services. Zhejiang Hospital and AHHNU specialize in providing focused care services. By combining these hospitals, the focus is on addressing the needs of patients who seek care at multiple health care facilities for various types of medical services.

The study conducted collaborative reasoning on fragmented medical information from patients who had visited multiple hospitals, aiming to detect overlooked CKD without necessitating the gathering of original EHR data. The study concentrated on patients in nonnephrology departments and leveraged CKD-related information typically overlooked by nonnephrology clinicians. This approach aimed to facilitate early detection of CKD risks. As illustrated in Figure 7, the collaborative reasoning mainly focused on 2 types of patients meeting the CKD diagnosis criteria (Textbox 2).

A disease-specific local ontology for CKD and semantic reasoning rules were developed based on clinical practice guidelines and CKD management studies [44-47]. Medical experts from the kidney department of FAHZU contributed to the creation of the ontology and semantic rules to ensure accuracy and clinical functionality. The systems analyzed local test results related to kidney function and identified abnormal findings, along with occurrences of visits. Primary CKD–related evidence is detailed in Multimedia Appendix 2.

Figure 7. The multicenter data group focuses on the early detection of unconsidered CKD using collaborated evidence from multiple hospitals. The transferred group focuses on unconsidered CKD warning at the first visit in the transferred hospital by collaborating evidence from previous hospital visits. CKD: chronic kidney disease.

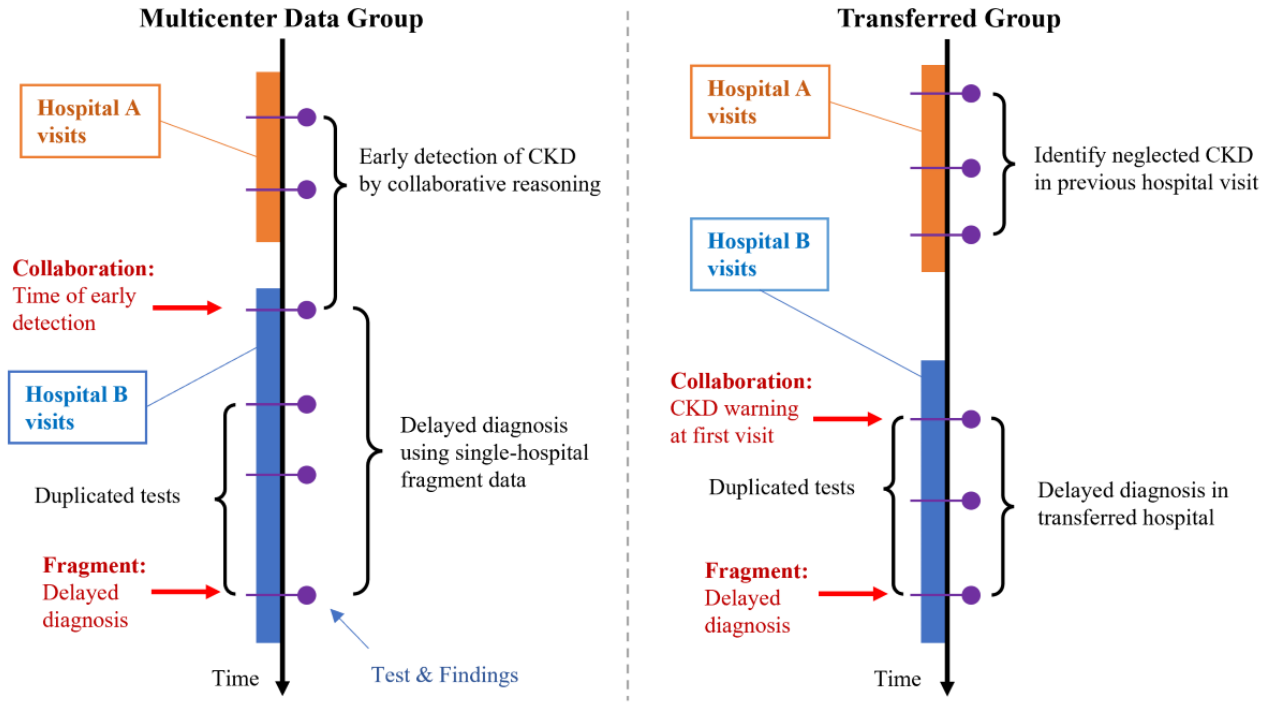

Textbox 2. Focus of the collaborative reasoning.

1. Detection through the combination of information from multiple hospitals

These patients' chronic kidney disease (CKD)–related data are dispersed across multiple hospitals, and only collaborative reasoning that integrates findings from multiple centers can promptly identify CKD. The collaboration of clinical findings could fulfill the 3-month monitoring criteria for chronic development. (This group is labeled the "Multicenter Data Group" in the "Results" section.)

2. Detection by bridging the information gap between hospitals

These patients fulfilled the CKD diagnosis criteria during previous visits to one hospital but were overlooked by clinicians. Subsequently, when transferred to another hospital without access to their prior data, health care providers neglected the patients' CKD risks for an extended period. (This group is labeled the "Transferred Group" in the "Results" section.)

Evaluation and Data Source

The evaluation of the application study encompasses the following aspects of clinical value and system performance: First, the system's capacity to identify overlooked patients with CKD using multicenter EHR data within a secure environment is assessed. Nephrology experts evaluate the reasoning outcomes by examining the comprehensive multicenter EHR data to

XSL•FO
RenderX

determine whether the patients have been confirmed as CKD positive. The assessment involved identifying patients who had subsequent EHR data regarding kidney function after the date recorded in the EHR when the knowledge graph identified the patients as meeting the CKD criteria. In essence, the assessment process utilizes EHR data both from the knowledge graph used for reasoning and subsequent EHR data used as "labels." As the study cohort lacks CKD diagnosis labels, it is crucial to incorporate additional data beyond the utilization of the knowledge graph to ensure the patients truly exhibit CKD positivity "in the future." Relying solely on identical data for assessment merely evaluates whether the system adheres to guideline-based rules and generates appropriate outputs. The selected subset, augmented with additional data, facilitates a more comprehensive assessment and supports the subsequent evaluation aspects.

Second, the advantages offered by collaborative reasoning in terms of discovery lead time, risk coverage, and potential test reduction are evaluated. Discovery lead time highlights the system's capability for early CKD detection, suggesting its potential to mitigate delayed diagnoses. The lead time $t_{\text{lead(M)}}$ of the multicenter data group is computed as the difference between the date t_{CDS} of the visit where the knowledge graph system identified patients meeting the CKD diagnosis criteria and the date $t_{\text{diagnosis}}$ of the assessment when clinicians diagnosed CKD using single-hospital EHR data for these patients, as in the following equation:

$$t_{\text{lead(M)}}=t_{\text{diagnosis}} - t_{\text{CDS}} \quad \textbf{(1)}$$

For patients in the transferred group, the lead time $t_{\text{lead}(T)}$ is calculated as the difference between the date t_{transfer} of the first visit to the transferred hospital and the date $t_{\text{diagnosis}}$ of the assessment when clinicians diagnose CKD using EHR data from the transferred hospital, as in the following equation:

$$t_{\text{lead(T)}}=t_{\text{diagnosis}} - t_{\text{transfer}} \quad \textbf{(2)}$$

The risk coverage demonstrates the system's capability to furnish abundant evidence of CKD for explanation and review by clinicians. Collaborative reasoning yields CKD-related risks r from multiple-hospital EHR data within a 3-month window at the ROI (t_{ROI}). As a baseline, single-hospital reasoning provides risks using EHR data from the latest 3 months (t_{3m}) as a comparative baseline. The risk coverage cov is calculated as a comparison as follows:

The duplicate examination reduction demonstrates the potential of the system to decrease unnecessary renal function tests for CKD diagnosis through multicenter collaborative reasoning. It is calculated based on the additional test records used in the clinician assessment of single-hospital data. The duplicated tests are depicted in Figure 7.

Lastly, the visualization and explanation of decision support were emphasized. A user interface was developed for information review and explanation of CDS results. An example of an overlooked patient with CKD identified through collaborative reasoning was presented to demonstrate the functionality of the visualization and explanation. The evaluation of blockchain performance is detailed in Multimedia Appendix 3 to underscore its suitability for the collaborative reasoning process.

The application study used EHR data from March 2008 to November 2020, provided by FAHZU, Zhejiang Hospital, and AHHNU. The study cohort comprised patients who had visited at least two of these hospitals. Cohort patients were selected and aligned using hash-encrypted identities. Patients with at least one decreased kidney function test result were included in the cohort. Patients who had either a nephrology visit or a kidney disease diagnosis record during the observation period were excluded. The EHR data, initially in the OMOP CDM format, were converted into the local EHR knowledge graph system.

Ethics Approval

The study was approved by the Clinical Research Ethics Committee of FAHZU (approval number 2020-330) and was exempt from informed consent for the following reasons: (1) the identity information of the data was either removed or encrypted before utilization; (2) the study did not involve commercial interests, and the data were not publicly disclosed; and (3) the data were used solely for system evaluation, and the study did not impact the health status of the patients.

Results

Study Cohort Characteristics

The cohort for multicenter reasoning of unconsidered patients with CKD included individuals from FAHZU, Zhejiang Hospital, and AHHNU, spanning from March 2008 to November 2020. All patients had visit records at FAHZU as well as visit records at either Zhejiang Hospital or AHHNU. Patients with a test record of an estimated glomerular filtration rate lower than 60 mL/min or a urine albumin-to-creatinine ratio higher than 30 mg/g were included in the cohort. Exclusion criteria were having a diagnosis record of any kidney disease or having a visit record from the kidney department. The cohort comprised a total of 1185 patients. Table 1 presents the characteristics of the patients at the time of cohort entry.

XSL•FO
RenderX

Table 1. Characteristics of the study cohort.

Characteristics	Values
Target cohort	
Patients, n	1185
Age (years)	
18-59, n (%)	219 (18.48)
>60, n (%)	966 (81.52)
Mean (SD)	68.52 (11.35)
Sex, n (%)	
Female	470 (39.66)
Male	715 (60.34)
Measurement, mean (SD)	
Blood potassium (mmol/L)	4.27 (0.46)
Serum creatinine (μmol/L)	109.24 (59.86)
Estimated glomerular filtration rate (mL/min)	57.20 (16.81)
Blood urine nitrogen (mmol/L)	7.55 (3.64)
Blood glucose (mmol/L)	5.97 (1.92)
High-density lipoprotein cholesterol (mmol/L)	1.21 (0.37)
Low-density lipoprotein cholesterol (mmol/L)	2.53 (0.92)
Total cholesterol (mmol/L)	4.60 (1.13)
Albumin-to-creatinine ratio (mg/mmol)	44.69 (110.36)
Diagnosis, n (%)	
Diabetes mellitus	160 (13.50)
Hypertension	281 (23.71)
Cardiovascular disease	106 (8.95)
Hyperlipidemia	29 (2.45)
Visit department[a] (top 10 most), n (%)	
Cardiovascular medicine	130 (11.43)
Emergency department	119 (10.47)
Gastroenterology	64 (5.63)
Ophthalmology	59 (5.19)
Orthopedics	56 (4.93)
Endocrinology	56 (4.93)
Urology	54 (4.75)
Respiratory medicine	50 (4.40)
Cardiology	46 (4.05)
Infectious disease	39 (3.43)

[a]The percentages were calculated based on the 1137 patients who had a specific visit department at the cohort entry visit. In total, the cohort included visits to 39 different departments.

Multicenter Reasoning of Unconsidered CKD

The evaluation study results are presented in Table 2. The EHR knowledge graph systems performed collaborative reasoning across the 3 hospitals and identified 124 patients who met the CKD diagnosis criteria based on the combination of multicenter medical information. In the multicenter data group, 69 patients met the CKD diagnosis criteria through the collaborative reasoning of fragmented EHR data. The data from either hospital alone would not support the diagnostic criteria for CKD. In the transferred group, 55 patients met the CKD diagnostic criteria during early visits to one hospital, but their CKD positivity was

XSL•FO
RenderX

identified much later at another hospital. Clinicians overlooked their CKD risks during the initial hospital visits. The information gap prevents doctors at subsequent hospitals from making prompt diagnoses. Collaborative reasoning could alert clinicians to previously neglected CKD risks during the initial visit.

A total of 91 patients were assessed by nephrology clinicians. These patients were selected if subsequent EHR data were available beyond what were used by the knowledge graph reasoning. These additional EHR data served as "labels" to further confirm CKD. The results demonstrated that the proposed system effectively identified patients' CKD risks through collaborative reasoning. The false positives in the assessment were primarily due to CKD recovery. These patients had undergone surgery or long-term treatment, resulting in temporarily reduced kidney function. After the treatment ceased and no longer affected the renal system, kidney function recovered, leading to false positives. Nonetheless, these patients required kidney function monitoring at the time of reasoning to prevent chronic risks.

Table 2. Evaluation results of collaborative reasoning of patients with unconsidered CKD[a].

Group	Patients, n	Assessed patients[b], n	Confirmed CKD by assessment, n (%)	Unconfirmed CKD by assessment, n
Patients meeting CKD diagnosis criteria (full cohort)	124	91	78 (86)	13
Multicenter data group[c]	69	40	32 (80)	8
Transferred group[d]	55	51	46 (90)	5

[a]CKD: chronic kidney disease.

[b]The patients' CKD was assessed by clinicians using subsequent electronic health record data that were dated later than the reasoned CKD date in accordance with CKD guidelines.

[c]The CKD status of patients in this group was determined through collaborative reasoning using findings from multiple hospitals.

[d]The CKD of patients in this group was overlooked at one hospital, and the information regarding their CKD risk did not transfer to another hospital during subsequent visits.

Advantages of Multicenter Reasoning

Table 3 presents the benefits of collaborative reasoning compared with single-hospital data analysis, encompassing discovery lead time, risk coverage comparison, and duplicate examination reduction. The discovery lead time and potential examination reduction are calculated based on subsequent EHR data following the data used in the reasoning process. The discovery lead time demonstrated that the system could identify CKD risks early on, long before clinical assessment. By leveraging multicenter fragmented medical information, the system delivered timely CDSs for cross-departmental clinicians. Consequently, it has the potential to reduce delayed or missed diagnoses during routine practice and address information gaps stemming from data fragmentation and security issues.

The CKD-related risk coverage highlights the system's ability to furnish comprehensive information for clinicians to review and assess the significance and progression of CKD risks. Part of the comparison between identified CKD-related risks in multiple-hospital collaborative reasoning and single-center reasoning is depicted in Figure 8. Each column indicates the number of patients found to have the specific risk. For instance, a history of acute kidney injury (an essential risk factor for CKD) from other hospitals may go unnoticed in single-hospital reasoning, leading clinicians to miss a crucial reference point in evaluating the patient's condition.

The potential examination reduction suggests that collaborative reasoning can effectively leverage multicenter fragmented information, using previous tests to identify overlooked CKD and offer decision support. This has the potential to reduce duplicate tests resulting from information gaps and facilitate prompt treatment.

XSL•FO
RenderX

Table 3. Discovery lead time and comparison between multicenter collaborative reasoning and single-center data analysis.

Variables	Multicenter data group	Transferred group
Discovery lead time[a] (days)		
Mean (SD)	434 (363)	208 (219)
Median	364	121
Reduced duplicate examinations[b]		
Tests, mean (SD)	3.34 (2.72)	3.56 (4.12)
Risk coverage comparison[c]		
Ratio, %	133	165

[a]The discovery lead time of the multicenter data group is calculated as the difference between the date of chronic kidney disease reasoning finding and the date of clinician assessment results. Conversely, the discovery lead time of the transferred group was calculated as the difference between the date of the first visit to the transferred hospital and the date of clinician assessment results.

[b]Test reduction refers to the additional tests required for clinicians to assess chronic kidney disease based solely on single-hospital data, compared with using multicenter collaborative reasoning.

[c]Using single-center reasoned chronic kidney disease–related risks as the baseline.

Figure 8. (A) Identified CKD-related risks of the Multicenter Data Group. (B) Identified CKD-related risks of the Transferred Group. The collaborative reasoning identified more CKD-related risk identification, while single-hospital reasoning missed some clinical evidence during decision support. AKI: acute kidney injury; BUN: blood urea nitrogen; CKD: chronic kidney disease; UA: uric acid; UP: urinary protein.

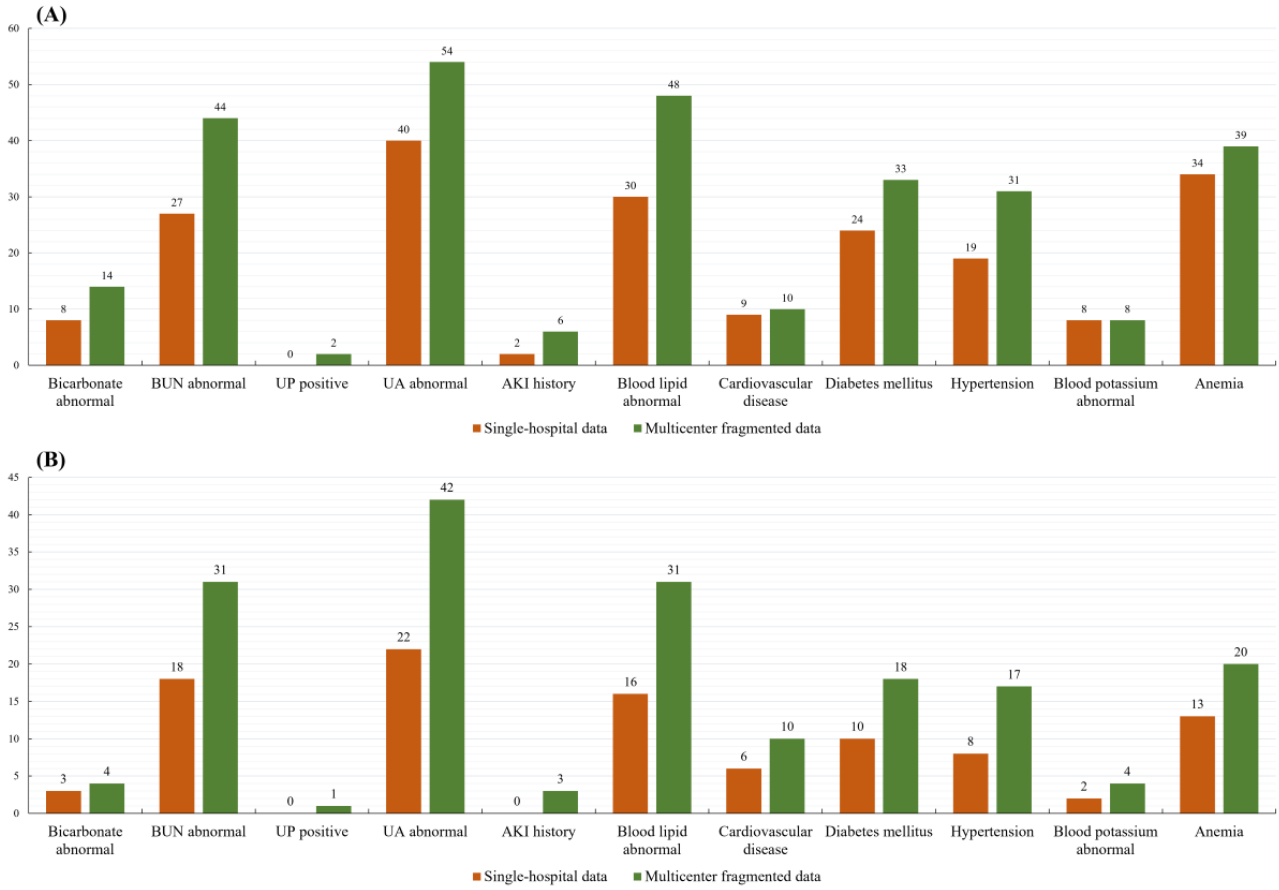

Visualization and Explanation of CDS Results

The user interface depicted in Figure 9 showcases an example of an overlooked patient with CKD identified by the system. At the top of the page, patient information and a table format of local EHR data are presented for review. On the lower part of the page, a graphical timeline illustrates the patient's medical pathway, accompanied by reasoning footage providing an explanation of the CDS for overlooked CKD. Additional interfaces can be found in Multimedia Appendix 4.

The patient with ID 11046406 visited FAHZU in April 2019, May 2019, and May 2020, and visited other hospitals (Zhejiang Hospital in this case) in August 2019 and January 2020. The system established an ROI meeting the CKD diagnosis criteria from April 4, 2019, to August 2, 2020. Local findings and

XSL•FO
RenderX

remote findings are distinguished by different filling styles. The local EHR data revealed abnormal estimated glomerular filtration rate test results and several CKD-related risks during that period. Findings from other hospitals indicated abnormal kidney function and several additional CKD-related risks. The system aggregated these findings and concluded that the patient met the CKD diagnosis criteria, exhibiting several significant risk factors. The reasoning footage and essential risks within the ROI are listed beneath the timeline for review. Despite the limited information provided by the remote system, it aids clinicians in identifying abnormalities and making targeted inquiries.

Figure 9. The system interface with information timeline and the reasoning footage for clinicians to review (translated version).

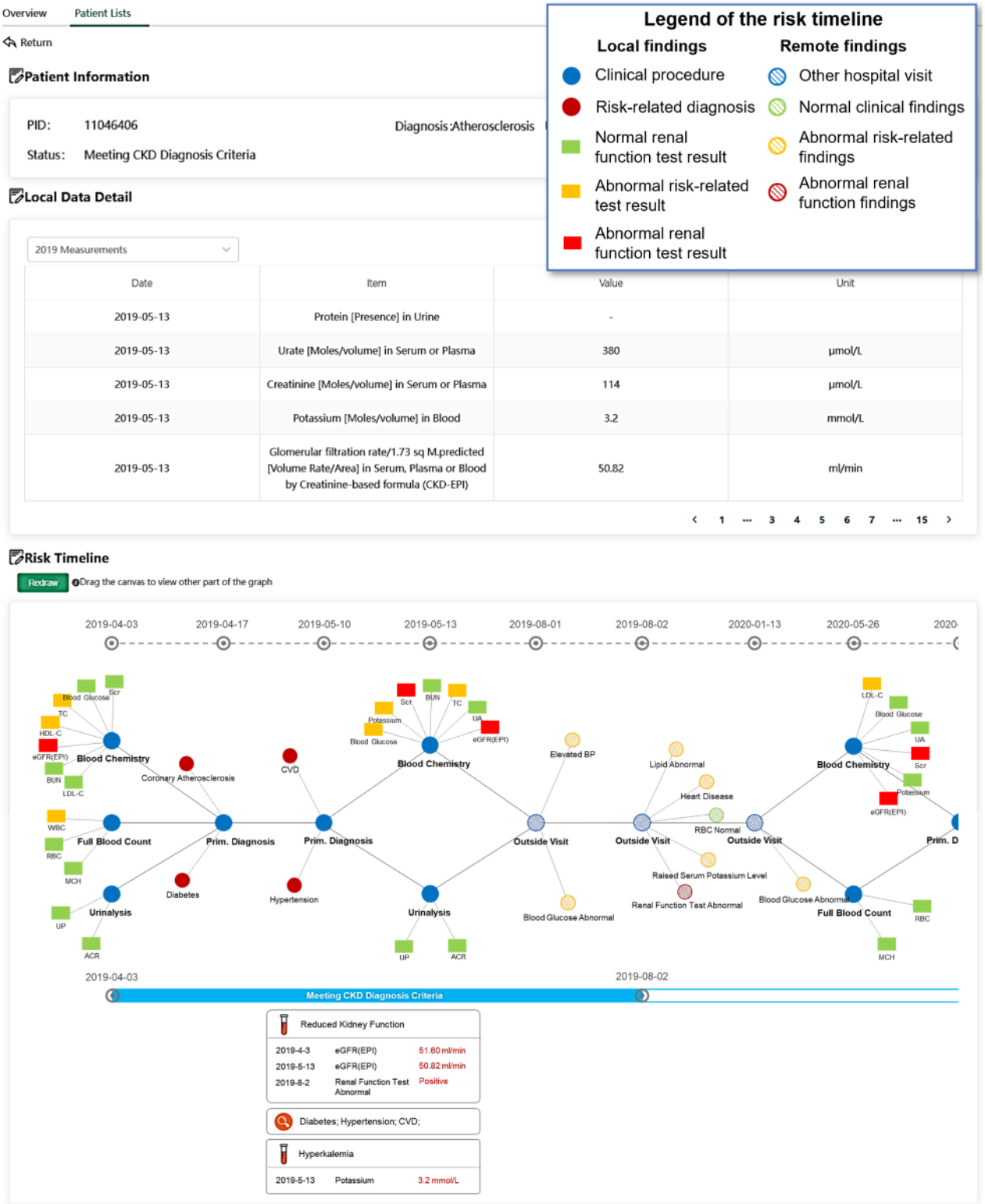

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we introduced an EHR-oriented knowledge graph system designed for use in a medical information-sensitive environment. The proposed framework enables knowledge graph systems to collaborate on clinical evidence across multiple hospitals, facilitating comprehensive CDS without the need for model aggregation. Several studies have highlighted the significance of addressing data fragmentation in clinical decision-making for improving health care outcomes [19,48]. Our proposed pilot framework was designed to address this issue and was implemented in a real-world application scenario. The application demonstrated that the system could identify chronic disease risks using multicenter fragmented information, a capability that single-hospital reasoning alone cannot achieve.

As an enhancement to our previous study, we introduced a collaborative reasoning framework to the system. This framework adopts a decentralized design, eliminating the need for a coordination center and offering flexibility for implementation across hospitals [49]. The knowledge graph within the hospital is a fully functional system equipped with data transformation and semantic reasoning capabilities, enabling local reasoning and CDS. The distribution component and blockchain facilitate the participation of the local system in collaborative reasoning. The local knowledge graph is responsible solely for handling reasoning on local triples, while the distribution component manages the collaboration of the multicenter reasoning process.

The application study on warning unconsidered patients with CKD underscored the clinical value of the proposed system. Fragmented EHR data of a patient present challenges for clinicians in obtaining a comprehensive view of multicenter evidence, potentially resulting in delayed or missed diagnoses [8,9]. The proposed systems were capable of identifying overlooked CKD in advance, particularly benefiting nonnephrology clinicians who might overlook patients' renal risks. The results of the application study indicated that the proposed system could (1) counter delayed or missed diagnoses, (2) potentially reduce redundant tests, and (3) provide complete information for clinicians to review.

During the application study, 69 patients were identified as meeting the CKD diagnostic criteria. Furthermore, clinician assessment revealed that 80% (32/40) of the evaluated patients exhibited positive CKD symptoms and test results. This suggests that a significant number of patients could benefit from the multicenter EHR knowledge graph system for timely CDS, thereby facilitating prompt treatment and enhancing health care quality. Conversely, another group of patients highlights the challenge posed by the information gap between hospitals. The system identified 55 patients with overlooked CKD during previous hospital visits. These patients either remained undiagnosed during subsequent visits to another hospital or received diagnoses much later. Patient transfers between hospitals often result in an information gap. The proposed system has the capability to transmit information between hospitals, alerting clinicians to risks and offering CDS during the initial visit after transfer.

The discovery lead time revealed that patients with overlooked CKD were neglected for an extended period. This underscores the detrimental effects of EHR data fragmentation across institutions. Valuable disease information remains obscured and underutilized due to fragmentation. Without comprehensive evidence, clinicians face limitations in identifying cross-departmental risks, leading to prolonged neglect. Implementing the proposed system across an extensive network encompassing medical centers and primary clinics has the potential to facilitate collaboration and early detection of disease risks by leveraging valuable information. Improved risk coverage can also furnish clinicians with a comprehensive background, enabling them to conduct thorough inquiries and assessments.

The study's main concepts revolve around distributed local reasoning and the collaboration of intermediate reasoning results to facilitate comprehensive CDS. For an evidence-based approach, it is crucial to gather complete findings from multiple centers to elucidate the rationale behind decision support creation. However, information exchanges do occur during the collaborative process. The framework implemented 3 major data security measures: (1) It ensured isolation between reasoning findings and original data. The intermediate findings generated by the local reasoning process do not disclose the source of the data. Remote hospitals only receive analysis results (eg, abnormal blood potassium and abnormal blood glucose findings) without information about how these findings were derived, whether from diagnosis or measurement. (2) The clinical findings used for cross-hospital collaboration are high-level concepts carefully chosen by domain experts during the development of the disease's local ontology. This selection aims to minimize the level of detail in the findings and obscure the relationship between reasoned findings and their original records. (3) The online subgraph is encrypted during the online synchronization phase, ensuring that only authorized hospitals on the network can receive the intermediate findings and protecting against cyberattacks. During clinical practice, clinicians also conduct inquiries to gather medical history from patients. Concerns about information exposure are manageable through patients' authorization of medical record usage and the implementation of proper security measures.

While the application study concentrated on unconsidered CKD warning, the proposed system can be adapted to other application domains for various clinical purposes. For instance, leveraging multicenter information aids in the sensitive and precise identification of type 2 diabetes, while collaborative reasoning offers risk warnings for general practitioners, and so forth. We are committed to further enhancing the proposed system to ensure its reliability and security in real-world applications. Implementing a more implicit collaboration method would foster better system adoption, particularly in data-sensitive environments.

Limitations

This study has its limitations. When deploying the system across an extensive network of hospitals and clinics, communication

XSL•FO
RenderX

efficiency may encounter bottlenecks. Additionally, the network and computational resource costs may escalate due to patient alignment and semantic reasoning of numerous subgraphs. To address these challenges, further systematic design and application of the Hyperledger method could facilitate the widespread deployment of the system. Furthermore, the patient alignment process relies on unique identifiers, which may pose challenges when unique citizen IDs are absent in the records. To enhance system adoption, alternative approaches using nonunique identifiers for similar patient alignment are necessary [4].

Comparison With Prior Work

In this study, we introduced a framework for knowledge graph systems to collaborate on multicenter fragmented clinical evidence to generate comprehensive CDS without sharing original data. First, our method focuses on collaborating local reasoning findings rather than original EHR data. By contrast, existing studies on patient record completion primarily concentrate on securely sharing EHR data through blockchain and selective encryption, encountering challenges related to data privacy and property rights [50,51]. Second, our proposed framework leverages multicenter fragmented information during the CDS application phase. Previous studies on using multicenter EHR data primarily focus on enlarging the model training set through federated learning to enhance model performance. However, these methods often fall short in addressing incomplete patient information from single centers when models are applied in daily practices [31,32,52]. Third, our proposed method uses knowledge graphs for explainable CDS and conducts local reasoning for local clinical findings. Current multicenter knowledge graph studies predominantly emphasize federated embedding learning, which trains embedding models without centralizing diverse knowledge graphs to ensure data security [25,26]. However, these methods also encounter challenges related to data incompleteness during model application.

To our knowledge, only a few studies have addressed the collaboration of fragmented medical information during CDS in practical settings. We introduced a pilot framework and reported clinical application results demonstrating the value of using multicenter fragmented information for CDS. This approach may assist nonnephrology clinicians in identifying patients with CKD risks in advance.

Conclusions

This study introduced an EHR-oriented knowledge graph system for collaborative CDS. The research demonstrated that the system effectively leverages fragmented patient EHR data from multiple hospitals, enabling the generation of CDS with intact clinical evidence without the need to share original data, thus addressing security and privacy concerns. The application study showcased a valuable scenario of detecting overlooked CKD using multicenter clinical information. Patients derived benefits from collaborative CDS for early-stage chronic disease warnings, all while safeguarding data security, an aspect unsupported by single-hospital data.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (grant 2022YFC2504605), the Preferred Funding Program of Postdoctoral Research of Zhejiang Province (grant ZJ2023066), the National Natural Science Foundation of China (grant 82172069), the Key Research Project of Zhejiang Lab Grants (grant 2022ND0AC01), and the Fundamental Research Funds for the Central Universities (grant 226-2023-00050). Generative artificial intelligence was not used in any portion of the manuscript.

Data Availability

The data used in the study are from and stored in 3 different hospitals. Because of the multicenter origin of the data, access to the data requires further approval from these hospitals and is not publicly available. Please contact the corresponding author to make reasonable requests on the data access.

Authors' Contributions

YS and YT conceptualized and designed the study. YS, KL, and TZ contributed to the system design and development. JC and PZ provided expert assessments for the study. YS and YT drafted the manuscript. All authors critically edited the manuscript. JL approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional technical details of the proposed system.
[DOCX File , 310 KB - jmir_v26i1e54263_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

The process of identifying patient with unconsidered CKD warning signs and primary primary evidence related to CKD. CKD: chronic kidney disease.
[DOCX File , 159 KB - jmir_v26i1e54263_app2.docx]

Multimedia Appendix 3
The performance and resource consumption of the blockchain network.
[DOCX File , 16 KB - jmir_v26i1e54263_app3.docx]

Multimedia Appendix 4
Additional user interfaces of the proposed system.
[DOCX File , 2590 KB - jmir_v26i1e54263_app4.docx]

References

1. Kern LM, Grinspan Z, Shapiro JS, Kaushal R. Patients' use of multiple hospitals in a major US city: implications for population management. Popul Health Manag 2017 Apr;20(2):99-102 [FREE Full text] [doi: 10.1089/pop.2016.0021] [Medline: 27268133]
2. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010 Dec 13;170(22):1989-1995. [doi: 10.1001/archinternmed.2010.439] [Medline: 21149756]
3. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annu Symp Proc 2011;2011:409-416 [FREE Full text] [Medline: 22195094]
4. Kho A, Yu J, Bryan M, Gladfelter C, Gordon H, Grannis S. Privacy-preserving record linkage to identify fragmented electronic medical records in the all of us research program. In: Cellier P, Driessens K, editors. Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science (vol 1168). Cham, Switzerland: Springer; Mar 28, 2020:77-87.
5. Cwinn MA, Forster AJ, Cwinn AA, Hebert G, Calder L, Stiell IG. Prevalence of information gaps for seniors transferred from nursing homes to the emergency department. CJEM 2009 Sep;11(5):462-471. [doi: 10.1017/s1481803500011660] [Medline: 19788791]
6. Bansler JP, Havn EC, Schmidt K. A study of the fragmentation of the medical record. In: Infrastructure for Healthcare: Global Heathcare: Proceedings of the 3rd International Workshop 2011. Copenhagen, Denmark: IT-Universitetet i København; 2011:94-97.
7. Kish LJ, Topol EJ. Unpatients-why patients should own their medical data. Nat Biotechnol 2015 Sep;33(9):921-924. [doi: 10.1038/nbt.3340] [Medline: 26348958]
8. Liu C, Talaei-Khoei A, Zowghi D, Daniel J. Data completeness in healthcare: a literature survey. PAJAIS 2017;9(2):75-100 [FREE Full text] [doi: 10.17705/1pais.09204]
9. Smith PC, Araya-Guerra R, Bublitz C, Parnes B, Dickinson LM, Van Vorst R, et al. Missing clinical information during primary care visits. JAMA 2005 Feb 02;293(5):565-571. [doi: 10.1001/jama.293.5.565] [Medline: 15687311]
10. Nicholson Price II W. Riskresilience in health data infrastructure. SSRN. 2017. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2928997 [accessed 2024-05-04]
11. Chen X, Jia S, Xiang Y. A review: knowledge reasoning over knowledge graph. Expert Systems with Applications 2020 Mar;141:112948. [doi: 10.1016/j.eswa.2019.112948]
12. Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med 2020 Mar;103:101817. [doi: 10.1016/j.artmed.2020.101817] [Medline: 32143785]
13. Hong C, Rush E, Liu M, Zhou D, Sun J, Sonabend A, VA Million Veteran Program. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. NPJ Digit Med 2021 Oct 27;4(1):151 [FREE Full text] [doi: 10.1038/s41746-021-00519-z] [Medline: 34707226]
14. Thukral A, Dhiman S, Meher R, Bedi P. Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications. Int J Inf Technol 2023 Jan 03;15(1):53-65. [doi: 10.1007/s41870-022-01145-y]
15. Xiao G, Pfaff E, Prud'hommeaux E, Booth D, Sharma DK, Huo N, et al. FHIR-Ontop-OMOP: building clinical knowledge graphs in FHIR RDF with the OMOP common data model. J Biomed Inform 2022 Oct;134:104201 [FREE Full text] [doi: 10.1016/j.jbi.2022.104201] [Medline: 36089199]
16. Carvalho RMS, Oliveira D, Pesquita C. Knowledge graph embeddings for ICU readmission prediction. BMC Med Inform Decis Mak 2023 Jan 19;23(1):12 [FREE Full text] [doi: 10.1186/s12911-022-02070-7] [Medline: 36658526]
17. Liu Z, Li X, Peng H, He L, Philip S. Heterogeneous similarity graph neural network on electronic health records. New York, NY: IEEE; 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13, 2020; Atlanta, GA p. 1196-1205. [doi: 10.1109/BigData50022.2020.9377795]

XSL•FO
RenderX

18. Shang Y, Tian Y, Zhou M, Zhou T, Lyu K, Wang Z, et al. EHR-oriented knowledge graph system: toward efficient utilization of non-used information buried in routine clinical practice. IEEE J Biomed Health Inform 2021 Jul;25(7):2463-2475. [doi: 10.1109/JBHI.2021.3085003] [Medline: 34057901]
19. Wei W, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19(2):219-224 [FREE Full text] [doi: 10.1136/amiajnl-2011-000597] [Medline: 22249968]
20. Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity (Edinb) 2020 Apr;124(4):525-534 [FREE Full text] [doi: 10.1038/s41437-020-0303-2] [Medline: 32139886]
21. Zhao X, Jia Y, Li A, Jiang R, Song Y. Multi-source knowledge fusion: a survey. World Wide Web 2020 Apr 08;23(4):2567-2592. [doi: 10.1007/s11280-020-00811-0]
22. Shang L, Kou Z, Zhang Y, Chen J, Wang D. A privacy-aware distributed knowledge graph approach to QoIS-driven COVID-19 misinformation detection. New York, NY: IEEE; 2022 Presented at: 2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS); June 10-12, 2022; Oslo, Norway p. 1-10. [doi: 10.1109/iwqos54832.2022.9812879]
23. Zhang F, Yuan N, Lian D, Xie X. Collaborative knowledge base embedding for recommender systems. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery; 2016 Aug 13 Presented at: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Augest 13-17, 2016; San Francisco, CA p. 353-362. [doi: 10.1145/2939672.2939673]
24. González-Beltrán A, Tagger B, Finkelstein A. Federated ontology-based queries over cancer data. BMC Bioinformatics 2012 Jan 25;13 Suppl 1(Suppl 1):S9 [FREE Full text] [doi: 10.1186/1471-2105-13-S1-S9] [Medline: 22373043]
25. Chen M, Zhang W, Yuan Z, Jia Y, Chen H. Fede: embedding knowledge graphs in federated setting. 2022 Jan 24 Presented at: The 10th International Joint Conference on Knowledge Graphs; December 6-8, 2021; Virtual Event, Thailand p. 80-88. [doi: 10.1145/3502223.3502233]
26. Peng H, Li H, Song Y, Zheng V, Li J. Differentially private federated knowledge graphs embedding. In: CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York, NY: Association for Computing Machinery; 2021 Oct 30 Presented at: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management; November 1-5, 2021; Virtual Event, Queensland, Australia p. 1416-1425. [doi: 10.1145/3459637.3482252]
27. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598 [FREE Full text] [doi: 10.1038/s41598-020-69250-1] [Medline: 32724046]
28. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
29. Axfors C, Schmitt AM, Janiaud P, Van't Hooft J, Abd-Elsalam S, Abdo EF, et al. Mortality outcomes with hydroxychloroquine and chloroquine in COVID-19 from an international collaborative meta-analysis of randomized trials. Nat Commun 2021 Apr 15;12(1):2349 [FREE Full text] [doi: 10.1038/s41467-021-22446-z] [Medline: 33859192]
30. Nuffield Department of Population Health Renal Studies Group, SGLT2 inhibitor Meta-Analysis Cardio-Renal Trialists' Consortium. Impact of diabetes on the effects of sodium glucose co-transporter-2 inhibitors on kidney outcomes: collaborative meta-analysis of large placebo-controlled trials. Lancet 2022 Nov 19;400(10365):1788-1801. [doi: 10.1016/S0140-6736(22)02074-8] [Medline: 36351458]
31. Noman AA, Rahaman M, Pranto TH, Rahman RM. Blockchain for medical collaboration: a federated learning-based approach for multi-class respiratory disease classification. Healthcare Analytics 2023 Nov;3:100135. [doi: 10.1016/j.health.2023.100135]
32. Tian Y, Shang Y, Tong D, Chi S, Li J, Kong X, et al. POPCORN: A web service for individual PrognOsis prediction based on multi-center clinical data CollabORatioN without patient-level data sharing. J Biomed Inform 2018 Oct;86:1-14 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.008] [Medline: 30103028]
33. The Book of OHDSI. OHDSI. URL: https://ohdsi.github.io/TheBookOfOhdsi/ [accessed 2022-12-15]
34. Chen Y, Ding S, Xu Z, Zheng H, Yang S. Blockchain-based medical records secure storage and medical service framework. J Med Syst 2018 Nov 22;43(1):5. [doi: 10.1007/s10916-018-1121-4] [Medline: 30467604]
35. Xie Y, Zhang J, Wang H, Liu P, Liu S, Huo T, et al. Applications of blockchain in the medical field: narrative review. J Med Internet Res 2021 Oct 28;23(10):e28613 [FREE Full text] [doi: 10.2196/28613] [Medline: 34533470]
36. Golang. Go. URL: https://go.dev/ [accessed 2024-05-04]
37. libp2p: a modular, P2P, networking library. libp2p. URL: https://libp2p.io/ [accessed 2024-05-04]
38. Yao A. How to generate and exchange secrets. New York, NY: IEEE; 1986 Presented at: 27th Annual Symposium on Foundations of Computer Science (sfcs 1986); October 27-29, 1986; Toronto, ON, Canada p. 162-167. [doi: 10.1109/sfcs.1986.25]

XSL•FO
RenderX

39. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. Lancet 2017 Mar 25;389(10075):1238-1252. [doi: 10.1016/S0140-6736(16)32064-5] [Medline: 27887750]
40. Zhang L, Wang F, Wang L, Wang W, Liu B, Liu J, et al. Prevalence of chronic kidney disease in China: a cross-sectional survey. Lancet 2012 Mar 03;379(9818):815-822. [doi: 10.1016/S0140-6736(12)60033-6] [Medline: 22386035]
41. Israni RK, Shea JA, Joffe MM, Feldman HI. Physician characteristics and knowledge of CKD management. Am J Kidney Dis 2009 Aug;54(2):238-247. [doi: 10.1053/j.ajkd.2009.01.258] [Medline: 19359079]
42. Fox CH, Brooks A, Zayas LE, McClellan W, Murray B. Primary care physicians' knowledge and practice patterns in the treatment of chronic kidney disease: an Upstate New York Practice-based Research Network (UNYNET) study. J Am Board Fam Med 2006;19(1):54-61 [FREE Full text] [doi: 10.3122/jabfm.19.1.54] [Medline: 16492006]
43. Drawz PE, Archdeacon P, McDonald CJ, Powe NR, Smith KA, Norton J, et al. CKD as a model for improving chronic disease care through electronic health records. Clin J Am Soc Nephrol 2015 Aug 07;10(8):1488-1499 [FREE Full text] [doi: 10.2215/CJN.00940115] [Medline: 26111857]
44. NA. Erratum: Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBD Update Work Group. KDIGO 2017 Clinical Practice Guideline Update for the Diagnosis, Evaluation, Prevention, and Treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). . 2017;7:1-59. Kidney Int Suppl (2011) 2017 Dec;7(3):1-150 [FREE Full text] [doi: 10.1016/j.kisu.2017.10.001] [Medline: 30681074]
45. Kellum JA, Lameire N, KDIGO AKI Guideline Work Group. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). Crit Care 2013 Feb 04;17(1):204 [FREE Full text] [doi: 10.1186/cc11454] [Medline: 23394211]
46. Vassalotti JA, Centor R, Turner BJ, Greer RC, Choi M, Sequist TD, National Kidney Foundation Kidney Disease Outcomes Quality Initiative. Practical approach to detection and management of chronic kidney disease for the primary care clinician. Am J Med 2016 Feb;129(2):153-162.e7 [FREE Full text] [doi: 10.1016/j.amjmed.2015.08.025] [Medline: 26391748]
47. Johnson DW, Atai E, Chan M, Phoon RK, Scott C, Toussaint ND, KHA-CARI. KHA-CARI guideline: early chronic kidney disease: detection, prevention and management. Nephrology (Carlton) 2013 May;18(5):340-350. [doi: 10.1111/nep.12052] [Medline: 23506545]
48. Hersh WR, Totten AM, Eden KB, Devine B, Gorman P, Kassakian SZ, et al. Outcomes from health information exchange: systematic review and future research needs. JMIR Med Inform 2015 Dec 15;3(4):e39 [FREE Full text] [doi: 10.2196/medinform.5215] [Medline: 26678413]
49. Westphal E, Seitz H. Digital and decentralized management of patient data in healthcare using blockchain implementations. Front Blockchain 2021 Aug 26;4:732112. [doi: 10.3389/fbloc.2021.732112]
50. Qiu H, Qiu M, Liu M, Memmi G. Secure health data sharing for medical cyber-physical systems for the healthcare 4.0. IEEE J Biomed Health Inform 2020 Sep;24(9):2499-2505. [doi: 10.1109/JBHI.2020.2973467] [Medline: 32071015]
51. Chelladurai U, Pandian S. A novel blockchain based electronic health record automation system for healthcare. J Ambient Intell Human Comput 2021 Mar 23;13(1):693-703. [doi: 10.1007/s12652-021-03163-3]
52. Guo K, Chen T, Ren S, Li N, Hu M, Kang J. Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinform 2022 Jun 23:e1 (forthcoming). [doi: 10.1109/TCBB.2022.3185395] [Medline: 35737631]

Abbreviations

AHHNU: Affiliated Hospital of Hangzhou Normal University
CDM: common data model
CDS: clinical decision support
CKD: chronic kidney disease
EHR: electronic health record
FAHZU: First Affiliated Hospital, College of Medicine, Zhejiang University
FKGE: Federated Knowledge Graphs Embedding
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
RDF: resource description framework
ROI: region of interest

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 03.11.23; peer-reviewed by S Ding, Z Wang; comments to author 30.11.23; revised version received 02.02.24; accepted 16.05.24; published 05.07.24.

Please cite as:
Shang Y, Tian Y, Lyu K, Zhou T, Zhang P, Chen J, Li J
Electronic Health Record–Oriented Knowledge Graph System for Collaborative Clinical Decision Support Using Multicenter Fragmented Medical Data: Design and Application Study
J Med Internet Res 2024;26:e54263
URL: https://www.jmir.org/2024/1/e54263
doi:10.2196/54263
PMID:

©Yong Shang, Yu Tian, Kewei Lyu, Tianshu Zhou, Ping Zhang, Jianghua Chen, Jingsong Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Computed Tomography–Based Fracture Prediction Model With Images of Vertebral Bones and Muscles by Employing Deep Learning: Development and Validation Study

Sung Hye Kong[1*], MD, PhD; Wonwoo Cho[2*], MSc; Sung Bae Park[3], MD, PhD; Jaegul Choo[2*], PhD; Jung Hee Kim[4*], MD, PhD; Sang Wan Kim[5], MD, PhD; Chan Soo Shin[4], MD, PhD

[1]Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[2]Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[3]Department of Neurosurgery, Seoul National University Boramae Hospital, Seoul, Republic of Korea
[4]Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[5]Department of Internal Medicine, Seoul National University Boramae Hospital, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Jung Hee Kim, MD, PhD
Department of Internal Medicine
Seoul National University Hospital
101 Dae-hak ro, Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 220724839
Email: jhee1@snu.ac.kr

Abstract

Background: With the progressive increase in aging populations, the use of opportunistic computed tomography (CT) scanning is increasing, which could be a valuable method for acquiring information on both muscles and bones of aging populations.

Objective: The aim of this study was to develop and externally validate opportunistic CT-based fracture prediction models by using images of vertebral bones and paravertebral muscles.

Methods: The models were developed based on a retrospective longitudinal cohort study of 1214 patients with abdominal CT images between 2010 and 2019. The models were externally validated in 495 patients. The primary outcome of this study was defined as the predictive accuracy for identifying vertebral fracture events within a 5-year follow-up. The image models were developed using an attention convolutional neural network–recurrent neural network model from images of the vertebral bone and paravertebral muscles.

Results: The mean ages of the patients in the development and validation sets were 73 years and 68 years, and 69.1% (839/1214) and 78.8% (390/495) of them were females, respectively. The areas under the receiver operator curve (AUROCs) for predicting vertebral fractures were superior in images of the vertebral bone and paravertebral muscles than those in the bone-only images in the external validation cohort (0.827, 95% CI 0.821-0.833 vs 0.815, 95% CI 0.806-0.824, respectively; $P<.001$). The AUROCs of these image models were higher than those of the fracture risk assessment models (0.810 for major osteoporotic risk, 0.780 for hip fracture risk). For the clinical model using age, sex, BMI, use of steroids, smoking, possible secondary osteoporosis, type 2 diabetes mellitus, HIV, hepatitis C, and renal failure, the AUROC value in the external validation cohort was 0.749 (95% CI 0.736-0.762), which was lower than that of the image model using vertebral bones and muscles ($P<.001$).

Conclusions: The model using the images of the vertebral bone and paravertebral muscle showed better performance than that using the images of the bone-only or clinical variables. Opportunistic CT screening may contribute to identifying patients with a high fracture risk in the future.

(J Med Internet Res 2024;26:e48535) doi:10.2196/48535

XSL•FO
RenderX

KEYWORDS

fracture; bone; bones; muscle; muscles; musculoskeletal; prediction; deep learning; prospective cohort; fracture risk assessment; predict; predictive; machine learning; develop; development; validate; validation; imaging; tomography; scanning

Introduction

The globally aging society has driven an increase in the incidence of fragility fractures and imposed a significant burden on health care systems, societies, and most importantly, on patients and their families [1-3]. Thus, proactively identifying patients with a high risk of fractures is vital. There are well-established methods to evaluate the risk of fractures, such as dual-energy X-ray absorptiometry (DXA) to assess bone mineral density (BMD), which is a reference standard for the diagnosis of osteoporosis [4]. However, a large proportion of patients have never undergone DXA, and 60% of the patients with major osteoporotic fractures do not receive proper treatment to reduce the risk of fractures [5].

Opportunistic computed tomography (CT) scans can be a novel approach for identifying patients with a high risk of fractures. Along with the increase in progressively aging populations, the use of opportunistic CT scanning is increasing, with over 80 million examinations performed each year in the United States [6]. Retrieval of information that can help assess the fracture risks from opportunistic CT scans does not require additional costs, time, or equipment, and data can be retrospectively acquired. Thus, it may help reduce the efforts associated with screening patients with high risks of fractures. Several studies have assessed BMD by using opportunistic CT scans [7], mainly utilizing the attenuation data of the trabecular bone of the spine [8,9].

There have been significant advances in deep learning techniques for medical image analysis, such as the convolutional neural network (CNN) method [10]. The CNN method facilitates the utilization of highly representative, data-driven image features, arranged in a layered hierarchical structure, which are effective in successfully classifying medical images. Various images have been used in CNN to classify patients with a high fracture risk [11-14]. Previous studies have primarily focused on the use of radiographic images for fracture detection, with AUROCs reported in the range of 0.73-0.80 [14,15]. However, there is scarcity of research utilizing CT images, and this is limited to bone texture analysis. Our study may fill this gap by applying CNN techniques to CT scans, which may provide a more accurate assessment of fracture risk due to the detailed and comprehensive nature of CT imaging. Further, as paravertebral muscles are among the critical contributing factors to vertebral fractures [16,17], CT could be a valuable method for acquiring information on both the muscle and vertebral bone. Nevertheless, to our knowledge, no previous study has reported using images of both the vertebral bone and muscle from CT scans by using the CNN method. Therefore, we aimed to develop and externally validate a CT-based fracture prediction model by using images of vertebral bones and muscles by employing a deep learning method. This study may help identify patients having high risk of fractures among those who undergo opportunistic CT scans for screening or other purposes.

Methods

Study Design and Participants

This study was based on a retrospective longitudinal cohort study of 32,435 patients having abdominal CT images at Seoul National University Bundang Hospital between 2010 and 2019. Patients who met all the inclusion criteria were included. Inclusion criteria were as follows: (1) patients who had abdominal CT imaging at Seoul National University Bundang Hospital between 2010 and 2019 and had follow-up images at the 5-year timepoint, (2) those who were aged between 50 and 80 years, and (3) those who were followed up for over a year. Further, patients who met any one of the exclusion criteria were excluded. The exclusion criteria and the number of excluded patients were as follows: (1) patients who were younger than 50 years or older than 80 years (n=2643), (2) those whose follow-up periods were less than a year (n=8029), and (3) those who had compression fractures or spinal surgery at the baseline (n=3258) (Figure 1). Finally, 18,505 patients were included in the analysis. During follow-up, 693 patients experienced vertebral fractures, while the remaining 17,812 patients did not. Among the 693 patients, after excluding 85 patients owing to the poor image quality or inappropriate CT protocols, 608 patients remained as cases.

For the control group, we selected individuals from the same time frame as the fracture cases. Among 17,812 patients who did not experience fracture, after excluding 2141 patients with poor image quality, we selected 606 age-, sex-, and BMI-matched individuals at a ratio of 1 patient to 1 control within a similar follow-up period. The fracture events were determined by reviewing medical records, with efforts to exclude any fractures associated with trauma. If patients had multiple CT scans during the follow-up, the earliest CT scan was used.

As a result, 1214 patients were eligible for analysis and constituted the development set. In addition, we developed an external validation set of 495 patients from Seoul National University Boramae Hospital by using the same protocol but without case-control matching between 2012 and 2013. An external validation set was developed to assess the performance of the intervention. The same protocol was used to ensure consistency in evaluation while allowing for a broader application of the findings in real-world settings.

XSL•FO
RenderX

Figure 1. Flowchart of participant selection for this study on vertebral fracture prediction. CT: computed tomography; SNUBH: Seoul National University Bundang Hospital.

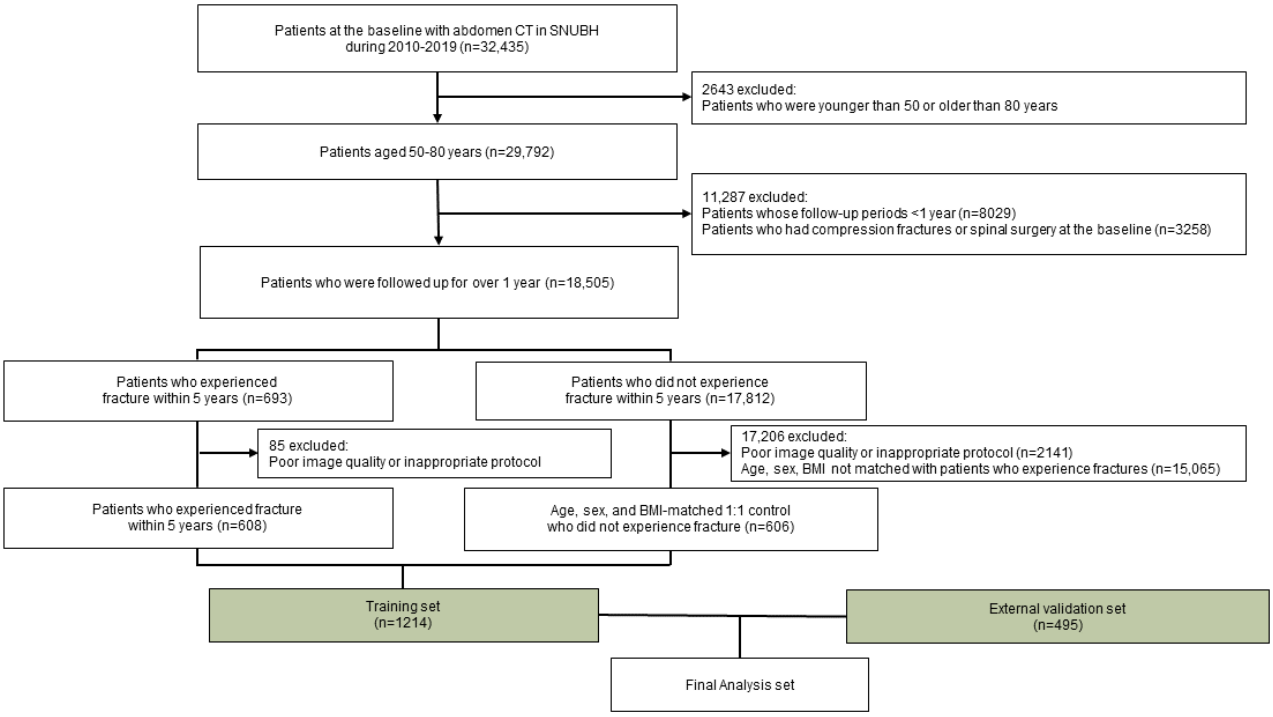

Ethics Approval

This study protocol was approved by the institutional review board of the Seoul National University Bundang Hospital (B-2104-677-402). The requirement for informed consent was waived owing to the retrospective design of this study. This study was conducted in accordance with the ethical standards laid down in the 1964 World Medical Association Declaration of Helsinki and its later amendments. This study also complies with the ethical principles for medical research.

Primary Outcome

The primary outcome of this study was the predictive accuracy of vertebral fracture events between T12 and L4 occurring within 5 years. Vertebral fractures were defined as morphometric fractures and were confirmed using radiographic or CT images. These images were adjudicated by SHK and JHK, who were blinded to any patient information prior to their assessments. Morphometric vertebral fractures were confirmed by radiographic or reconstructed CT images with measurements of anterior (Ha), middle (Hm), and posterior (Hp) height of each vertebral body from T11 to L4 measured. Normal population was classified as having no vertebral fracture by gross visual inspection as being within normal range for vertebral height and shape. The mean (SD) of ratio of normal vertebral height was obtained from patients without incident fractures. We first calculated the anterior to posterior (Ha/Hp), middle to posterior (Hm/Hp), and posterior to posterior above and below (Hpi/Hpi+1 and Hpi/Hpi–1) ratios. Vertebral fracture was defined if any of the following ratios were more than 3 SDs below the normal mean for that vertebral level, which was 0.91 (SD 0.08), as described in previous reports [18,19].

Measurements of Clinical Factors

Sociodemographic factors, including age, sex, and medical history, were obtained from a review of electronic medical records at baseline. Height and body weight were measured using standard methods by trained staff with a scale and wall-mounted extensometer, and the participants wore lightweight clothes. BMI was calculated as weight divided by height in meters squared (kg/m^2). Current smokers were defined as patients who were smoking during the study period, while current alcohol consumers were defined as those who consumed 3 or more units of alcohol daily. The use of glucocorticoids was defined as using oral glucocorticoids or having been exposed to oral glucocorticoids for more than 3 months at a prednisolone dose >5 mg or its equivalent doses. Possible secondary osteoporosis is defined as osteoporosis that occurs due to factors other than primary menopause or age-related causes. It includes patients with osteoporosis and concurrent diagnosis with type 1 diabetes, osteogenesis imperfecta in adulthood, hyperthyroidism, hypogonadism, premature menopause (age<45 years), chronic malnutrition, malabsorption, or chronic liver disease [20].

CT Protocols, Image Preprocessing, and Deep Learning Techniques

Intravenous contrast-enhanced images were obtained using CT scanners with 64 detector rows (Brilliance; Philips Medical Systems). All the patients were placed in a supine position and scanned from the diaphragm to the symphysis pubis. The reference tube current–time product was empirically set, aiming at effective radiation doses of 2 mSv. The effective tube current–time product generally ranges between 25 mA and 40 mA. The actual radiation dose was adjusted according to the body size by automatically modulating the tube current

XSL•FO
RenderX

(Dose-Right; Philips Medical Systems). The values of tube voltage, collimation, rotation speed, and pitch were 120 kVp, 64 mm×0.625 mm, 0.5 seconds, and 0.891, respectively. Patients were administered 2 mL iopromide/kg (Ultravist 370; Schering) intravenously at a rate of 3 mL/s via the antecubital vein, and scanning was initiated 60 seconds after the enhancement of the descending aorta reached 150 HU. From each helical scan, the images were reconstructed using a section thickness of 5 mm.

Consecutive image processing was applied to all the CT images for accurate deep learning–based image analysis. Each axial slice of the abdominal CT scan was resampled to obtain a pixel spacing of 1×1 mm^2. The signal intensity of each CT image was min-max normalized to the –1 to 1 range after windowing the Hounsfield unit values in the range of –200 to 1000. Subsequently, 2 classes of image data, that is, vertebral body only (bone-only) and vertebral body with paravertebral muscles (bone+muscle) were extracted from each CT image. Based on the manual annotations of the vertebral body, excluding the intervertebral disc, the vertebral body regions from T12 to L4 were extracted from the CT images, where each bone-only image had 96×96 pixels in the axial plane. Centered on the vertebral body, the images of the paravertebral muscle were automatically cropped using a rectangular box (96×144 pixels) in the axial plane (Figure S1 in Multimedia Appendix 1).

Deep learning–based image features for the 5-year risk analysis of vertebral fractures were extracted using the attention CNN–recurrent neural network (CNN-RNN) model for image data (Figure 2) [21,22]. In the CNN-RNN model, ImageNet-pretrained ResNeXt-50 and gated recurrent units were employed as the CNN encoder backbone and the RNN recurrent decoder, respectively. As inputs of the model, 14 equidistant axial slices were extracted from the T12 to L4 vertebrae region of each CT image, where the starting slice was randomly selected at each training iteration for data augmentation. In the training phase, the CNN-RNN model was optimized using the Adam optimizer and cross-entropy loss, where the learning rate and batch size were 1e-5 and 128, respectively. In our image-based fracture prediction model, the utilization of CT images was primarily driven by deep learning methods, particularly CNNs. Although specific imaging parameters such as attenuation values and density were not directly used as standalone inputs, the CNN's learning process inherently captured these aspects as part of the comprehensive image analysis. The model processed the entire CT images, extracting deep features that potentially included characteristics related to bone and muscle attenuation and density, among others. This approach allowed for a sophisticated interpretation of the CT scans, identifying nuanced patterns indicative of fracture risk. To incorporate clinical variables into our image-based prediction model, we first standardized the clinical variables to ensure consistency and comparability. Following this, we concatenated the standardized variables to the image features in the final layer of the CNN-RNN model.

To understand how the model identifies and differentiates key areas for predicting vertebral fractures in CT images, we employed the gradient-weighted class activation mapping technique. This approach involves highlighting the most crucial regions within the images, marked by a bright red overlay, thereby revealing the model's decision-making process and focal areas for classification (Figure 3). The image models were developed with a high-performance computing server with 4 NVIDIA GeForce GTX 1080 Ti (NVIDIA) graphic processing units and the Ubuntu 16.04.4 operating system.

Figure 2. Architecture of the convolutional neural network–recurrent neural network model for vertebral fracture prediction. CNN: convolutional neural network; FC, fully connected; RNN: recurrent neural network.

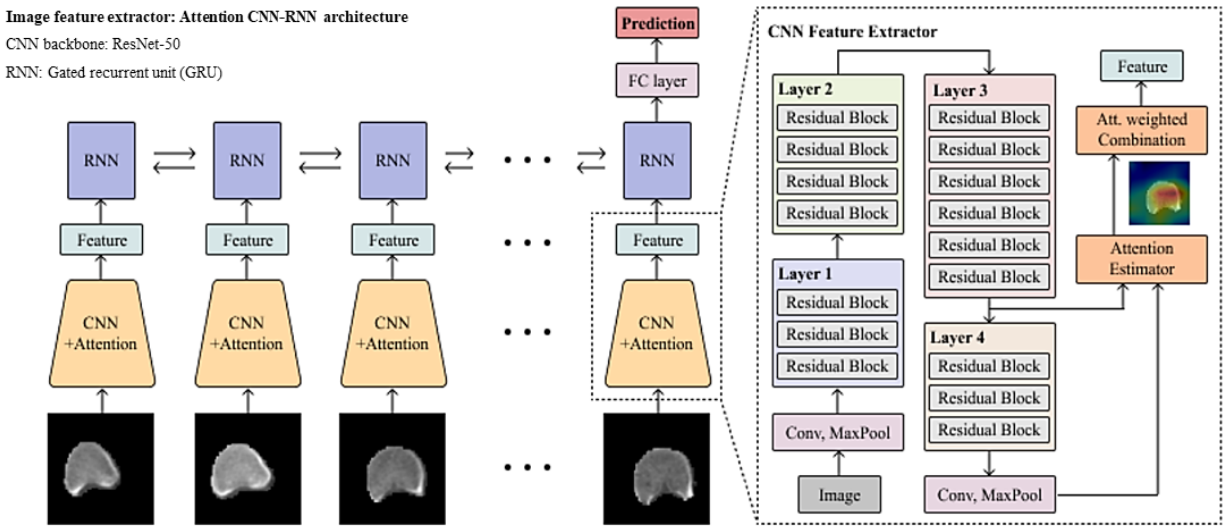

XSL•FO
RenderX

Figure 3. Gradient-weighted class activation mapping heatmaps of image models using (A) bone-only and (B) bone+muscle images. A heatmap was generated by gradient-weighted class activation mapping, a technique used to visualize those parts of an input image that are important for a neural network's decision, especially in convolutional neural networks.

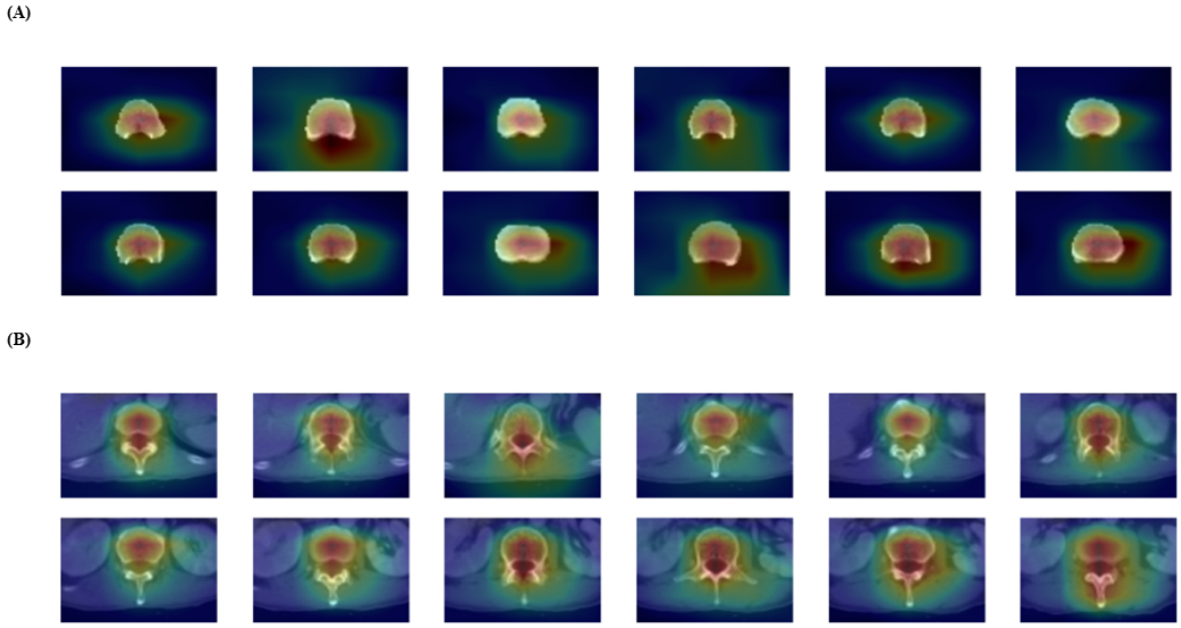

Statistical Analyses

For the baseline characteristics, depending on the distribution, continuous parameters were presented as means with standard deviations and categorical data were presented as proportions. Comparisons between the groups having continuous variables were analyzed using 2-sided Student *t* test, whereas χ^2 test was used for the categorical variables. The area under the receiver operating characteristic curve (AUROC) was calculated to compare the preprocessed images. Cases predicted to have an actual fracture event and experienced it during follow-up were defined as true positive. Those predicted to have but did not experience a fracture was designated as false positive. Cases predicted to be free from fracture events but experienced one during follow-up were defined as false negative. True-negative cases were predicted to be free of fracture events with no fractures during follow-up. Sensitivity and specificity were calculated for each time series as follows: sensitivity = true positive / (true positive + false negative) and specificity = true negative / (true negative + false positive). The risk prediction performance measures were gauged using 10-fold cross-validation.

The image-only 5-year risk analyses of vertebral fractures were conducted by applying a fully connected layer, which generates binary prediction results to the CNN-RNN feature extractor. In addition to the image-only model, clinical models (models A, B, C, and D) were developed by analyzing the corresponding clinical variables via XGBoost. The clinical variables included age, sex, BMI, use of steroids, smoking status, and possible secondary osteoporosis. Model A included age and sex as independent variables; model B additionally incorporated BMI; model C further included clinical variables such as the use of glucocorticoids, history of alcohol consumption, smoking, and possible secondary osteoporosis; and model D included type 2 diabetes mellitus, HIV, hepatitis C infection [23], and renal failure [24].

The matching of cases and controls in our study was conducted based on age and sex. This process was facilitated using the propensity score matching method, implemented through the MatchIt package in R (version 4.1.2; R Foundation). PyTorch and Scikit-learn libraries from Python were used for the analyses. A *P* value <.05 was considered significant. Correction for multiple testing was not performed across models. Statistical analyses were performed using Python (version 3.8.10; Python Software Foundation). The programs used in the experiments were PyCharm (JetBrains s.r.o.) and Visual Studio (Microsoft Corp).

Results

Clinical Characteristics

A total of 1709 individuals were included in the analysis. The participants were divided into a development set from Seoul National University Bundang Hospital (n=1214) and an external validation set (n=495) from Seoul National University Boramae Medical Center. As shown in Table 1, the development set was older (mean 72.5, SD 7.9 years) than the external validation set (mean 67.6, SD 8.6 years), with a statistically significant difference (*P*<.001). The proportion of females in the external validation set (390/495, 78.8%) was higher than that in the development set (839/1214, 69.1%), with this difference also being significant (*P*<.001). However, no significant differences were observed in weight and BMI between the 2 sets. When considering lifestyle factors, there was a higher prevalence of current smokers (284/1214, 23.4% vs 28/495, 5.7%; *P*<.001) and current drinkers (236/1214, 19.4% vs 34/495, 6.9%; *P*<.001) in the development set than those in the external validation set, respectively. The use of steroids was similar across both groups (*P*=.56), while the prevalence of possible secondary osteoporosis

XSL•FO
RenderX

was significantly higher in the development set (122/1214, 10.1% vs 16/495, 3.2%, respectively; *P*<.001). Within 5 years of follow-up, 454 (37.4%) and 61 (12.3%) individuals experienced vertebral fractures in the development and external validation sets, respectively.

In the development set (n=1214), participants were matched based on age, sex, and BMI to compare those with incident fractures (n=608) to those without (n=606). There was a less than 1-year age difference (mean age 72.0, SD 7.6 years in the nonfracture group vs 72.9, SD 8.2 years in the fracture group; *P*<.001) and a BMI difference of less than 0.5 (mean BMI 23.7, SD 3.4 kg/m² in the nonfracture group vs 23.5, SD 3.5 kg/m² in the fracture group; *P*=.39). Gender distribution was balanced between the 2 groups (422/606, 69.6% females in the nonfracture group vs 417/608, 68.6% females in the fracture group; *P*=.96).

Despite these matched parameters, a higher prevalence of current smokers was noted in the fracture group than in the nonfracture group (169/608, 27.8% vs 115/606, 18.9%, respectively; *P*=.001) along with a significantly higher use of steroids (114/608, 18.8% vs 47/606, 7.8%, respectively; *P*<.001) and a greater prevalence of possible secondary osteoporosis (77/608, 12.7% vs 45/606, 7.4%, respectively; *P*=.001). No significant differences were observed in the height, weight, and current drinking status between the 2 groups (Table 2).

Table 1. Clinical characteristics of the development and external validation sets.[a]

	Development set (n=1214)	External validation set (n=495)	*P* value
Age (years), mean (SD)	72.5 (7.9)	67.6 (8.6)	<.001
Female, n (%)	839 (69.1)	390 (78.8)	<.001
Height (cm), mean (SD)	157.0 (8.4)	155.2 (7.9)	<.001
Weight (kg), mean (SD)	58.2 (9.9)	57.6 (9.7)	.22
BMI (kg/m^2), mean (SD)	23.6 (3.5)	23.9 (3.6)	.18
Current smoker, n (%)	284 (23.4)	28 (5.7)	<.001
Current drinker, n (%)	236 (19.4)	34 (6.9)	<.001
Use of steroids,[b] n (%)	161 (13.3)	57 (11.5)	.56
Possible secondary osteoporosis,[c] n (%)	122 (10.1)	16 (3.2)	<.001
Vertebral fracture within 5 years, n (%)	608 (50)	61 (12.3)	<.001

[a]The variables between the groups were compared using the 2-sided Student *t* test for continuous variables and the χ^2 test for categorical variables.

[b]Use of steroids was defined as the use of prednisolone 5 mg daily or equivalent over 3 months.

[c]Possible secondary osteoporosis includes patients with osteoporosis and concurrent diagnosis with type 1 diabetes, osteogenesis imperfecta in adulthood, hyperthyroidism, hypogonadism, premature menopause (<45 years), chronic malnutrition, malabsorption, and chronic liver disease.

Table 2. Baseline clinical characteristics of the development set comparing incident fracture groups.[a]

	Incident fracture (-) (n=606)	Incident fracture (+) (n=608)	*P* value
Age (years), mean (SD)	72.0 (7.6)	72.9 (8.2)	<.001
Females, n (%)	422 (69.6)	417 (68.6)	.96
Height (cm), mean (SD)	156.9 (8.5)	157.1 (8.3)	.32
Weight (kg), mean (SD)	58.29 (9.5)	58.11 (10.4)	.23
BMI (kg/m^2), mean (SD)	23.7 (3.4)	23.5 (3.5)	.39
Current smoker, n (%)	115 (18.9)	169 (27.8)	.001
Current drinker, n (%)	113 (18.7)	123 (20.2)	.92
Use of steroids,[b] n (%)	47 (7.8)	114 (18.8)	<.001
Possible secondary osteoporosis,[c] n (%)	45 (7.4)	77 (12.7)	.001

[a]The variables between the groups were compared using the 2-sided Student *t* test for continuous variables and the χ^2 test for categorical variables. Fracture (-) and (+) groups represent participants who did not and did experience fractures at 5 years of follow-up, respectively.

[b]Use of steroids was defined as the use of prednisolone 5 mg daily or equivalent over 3 months.

[c]Possible secondary osteoporosis includes patients with osteoporosis and concurrent diagnosis with type 1 diabetes, osteogenesis imperfecta in adulthood, hyperthyroidism, hypogonadism, or premature menopause (<45 years), chronic malnutrition, malabsorption, and chronic liver disease.

XSL•FO
RenderX

Comparisons Between the Performances of Image Models in Predicting Vertebral Fractures

As demonstrated in Table 3, for the development set, the models using images that included both vertebral bone and paravertebral muscle showed significantly better AUROC, accuracy, and precision values compared to those using bone-only images. Specifically, the bone-only images had an AUROC of 0.677 (95% CI 0.674-0.680) and accuracy of 0.669 (95% CI 0.665-0.673). In contrast, the images including both bone and muscle exhibited an AUROC of 0.739 (95% CI 0.737-0.741) and accuracy of 0.719 (95% CI 0.715-0.722; all *P*<.001). The fracture risk assessment tool (FRAX) model for major osteoporotic fracture and hip fracture showed lower AUROCs of 0.557 and 0.563, respectively, indicating a significantly better performance of our image model (all *P*<.001).

Similar trends were observed in the external validation set, where bone-only images resulted in an AUROC of 0.815 (95% CI 0.806-0.824) and accuracy of 0.754 (95% CI 0.752-0.756), while the combined bone and muscle images demonstrated an AUROC of 0.827 (95% CI 0.821-0.833) and accuracy of 0.812 (95% CI 0.798-0.826; all *P*<.001), though the specificity value was similar between the 2 groups. The FRAX model for major osteoporotic fracture and hip fracture had AUROCs of 0.810 and 0.780, respectively. Again, these results confirmed the superior predictive capability of our image-based model (all *P*<.001).

Table 3. Performance comparisons of image models in predicting vertebral fractures.

	Development set			External validation set		
	Bone only	Bone + muscle	*P* value	Bone only	Bone + muscle	*P* value
AUROC[a] (95% CI)	0.677 (0.674-0.680)	0.739 (0.737-0.741)	<.001	0.815 (0.806-0.824)	0.827 (0.821-0.833)	.04
Accuracy (95% CI)	0.669 (0.665-0.673)	0.719 (0.715-0.722)	<.001	0.754 (0.752-0.756)	0.812 (0.798-0.826)	<.001
Sensitivity (95% CI)	0.746 (0.739-0.753)	0.761 (0.746-0.776)	.23	0.645 (0.613-0.677)	0.704 (0.675-0.733)	.054
Specificity (95% CI)	0.601 (0.586-0.616)	0.634 (0.625-0.643)	.002	0.844 (0.810-0.877)	0.855 (0.835-0.875)	.43

[a]AUROC: area under the receiver operating characteristic curve.

Comparisons Between the Performances of Image and Clinical Models

Compared to the clinical models, the image model using vertebral bone and muscle showed significantly higher performance than the clinical models in predicting the vertebral fractures during the 5-year follow-up period in the development and external validation sets (Figure 4, Table 4). In the development set, the images that included vertebral bone and muscle had significantly better AUROC and accuracy than the clinical model D, which included age, sex, BMI, history of alcohol consumption, smoking, possible secondary osteoporosis, type 2 diabetes mellitus, HIV, hepatitis C infection status, and renal failure (AUROC 0.667, 95% CI 0.661-0.672 and accuracy 0.640, 95% CI 0.661-0.649; all *P*<.001, Table 4). In addition, the performance did not show a significant change when the clinical variables were added to the image-only model (Table S1 in Multimedia Appendix 1).

Figure 4. Receiver operating characteristic curves comparing models for 5-year vertebral fracture prediction: model A includes age and sex, model B additionally includes BMI, and model C additionally includes a history of drinking, smoking, and possible secondary osteoporosis.

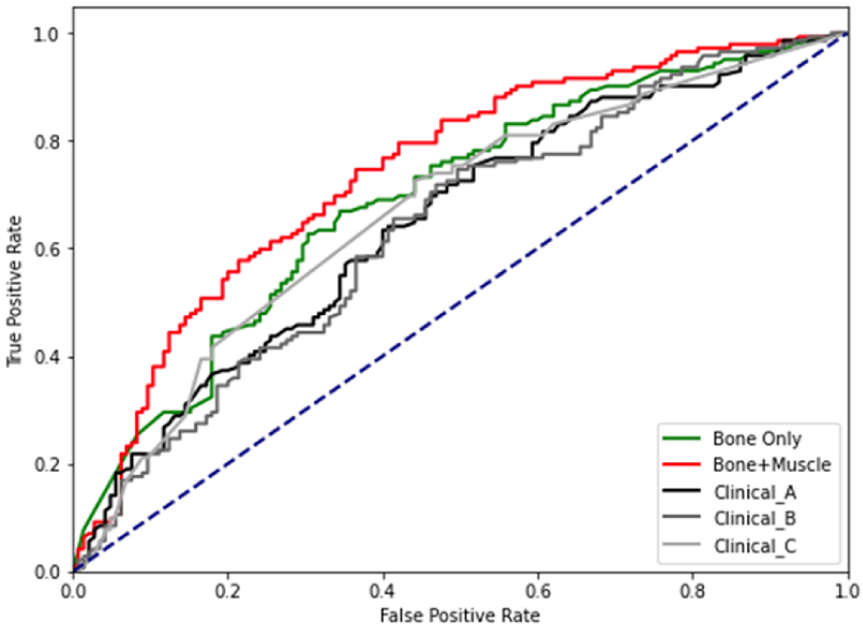

XSL•FO
RenderX

Table 4. Performance comparisons of image and clinical variable models in predicting fractures.

	AUROC[a] (95% CI)	*P* value	Accuracy (95% CI)	*P* value	Sensitivity (95% CI)	*P* value	Specificity (95% CI)	*P* value
Development set								
Image-only[b]	0.739 (0.737-0.741)	Reference	0.719 (0.716-0.722)	Reference	0.761 ± 0.024 (0.746-0.776)	Reference	0.634 (0.625-0.643)	Reference
Clinical model A[c]	0.647 (0.643-0.651)	<.001	0.620 (0.614-0.626)	<.001	0.681 (0.643-0.719)	.03	0.575 0.549-0.601)	<.001
Clinical model B[d]	0.631 (0.626-0.636)	<.001	0.612 (0.610-0.614)	<.001	0.675 (0.639-0.711)	.02	0.558 (0.517-0.598)	.003
Clinical model C[e]	0.663 (0.659-0.667)	<.001	0.637 (0.631-0.643)	<.001	0.723 (0.694-0.752)	.11	0.553 (0.521-0.585)	<.001
Clinical model D[f]	0.667 (0.661-0.672)	<.001	0.640 (0.661-0.649)	<.001	0.729 (0.690-0.768)	.13	0.560 (0.527-0.593)	.005
FRAX[g] (MOF)[h]	0.557[i]	<.001	0.557[i]	<.001	0.442[i]	<.001	0.672[i]	<.001
FRAX (hip)	0.563[i]	<.001	0.556[i]	<.001	0.449[i]	<.001	0.663[i]	<.001
External validation set								
Image-only	0.827 (0.821-0.833)	Reference	0.812 (0.798-0.826)	Reference	0.704 (0.675-0.733)	Reference	0.855 (0.834-0.875)	Reference
Clinical model A	0.731 (0.725-0.737)	<.001	0.651 (0.629-0.673)	<.001	0.715 (0.683-0.747)	<.001	0.656 (0.629-0.683)	<.001
Clinical model B	0.733 (0.725-0.737)	<.001	0.654 (0.625-0.683)	<.001	0.728 (0.673-0.783)	<.001	0.662 (0.621-0.703)	<.001
Clinical model C	0.745 (0.733-0.757)	<.001	0.669 (0.646-0.692)	<.001	0.713 (0.678-0.748)	<.001	0.720 (0.689-0.751)	<.001
Clinical model D	0.749 (0.736-0.762)	<.001	0.675 (0.643-0.707)	<.001	0.729 (0.690-0.768)	<.001	0.686 (0.650-0.722)	<.001
FRAX (MOF)	0.810[i]	<.001	0.810[i]	<.001	0.262[i]	<.001	0.887[i]	<.001
FRAX (hip)	0.780[i]	<.001	0.685[i]	<.001	0.705[i]	<.001	0.682[i]	<.001

[a]AUROC: area under the receiver operating characteristic curve.

[b]Image model represents the model using bone and muscle.

[c]Model A includes age and sex.

[d]Model B additionally includes BMI.

[e]Model C additionally includes history of drinking, smoking, and possible secondary osteoporosis.

[f]Model D includes age, sex, BMI, history of alcohol consumption, smoking, possible secondary osteoporosis, type 2 diabetes mellitus, HIV, hepatitis C infection status, and renal failure.

[g]FRAX: fracture risk assessment tool.

[h]MOF: major osteoporotic fracture.

[i]Since this was calculated for a single data set, there are no 95% CI values.

As depicted in Figure 4, in the external validation set, the images including vertebral bone and muscle showed a significantly better AUROC and accuracy than the clinical model D (AUROC 0.749, 95% CI 0.736-0.762 and accuracy 0.675, 95% CI 0.643-0.707; all *P*<.001). The results were similar for clinical models A, B, C, and D, which showed poorer performance than the image model.

Discussion

In this study, we developed and externally validated a vertebral fracture prediction model by using abdominal CT images. In the development cohort, the performance of predicting vertebral fractures represented by AUROC was 0.688 (SD 0.001) by using images of vertebral bone-only and 0.736 (SD 0.003) by using images of vertebral bone and paravertebral muscle. In the validation cohort, the performances (AUROC) were 0.698 (SD 0.001) and 0.729 (SD 0.002) for images of vertebral bone-only and images of vertebral bone and paravertebral muscle, respectively. In addition, the performance of the model using images of vertebral bone and muscle was significantly better than that of the clinical models using age, sex, BMI, use of steroids, smoking status, and possible secondary osteoporosis, which showed performances of 0.635 (SD 0.002) and 0.698

XSL•FO
RenderX

(SD 0.021), respectively, for the development and validation cohorts.

Our model shows that the image models using vertebral bone and muscle had a better performance than those using images of vertebral bone-only. Osteosarcopenia, defined by combined occurrence of bone loss and sarcopenia, is one of the critical risk factors for osteoporotic fractures [25,26]. The paravertebral muscles are essential components of the vertebral column and are associated with osteoporotic vertebral fractures [27,28]. In previous studies, information retrieved from muscle images, such as cross-sectional area, volume, and degree of fat infiltration in the paravertebral muscle, was correlated with vertebral stability and the risk of fractures [28,29]. Specifically, Kim et al [30] reported lower cross-sectional areas and greater fat infiltration of the paravertebral muscles in patients with vertebral fractures than in those without fractures. This implies that not only the density and quality of the bones are correlated with the risk of fractures but also the quality of the muscles supporting and communicating with the bones [17]. Fat infiltration in the muscles, called myosteatosis, has been reported to be associated with an increased risk of fractures [17,31]. Thus, in line with previous studies, our study results imply that information from the images of the paravertebral muscles in addition to the information from the images of vertebral bones can help predict vertebral fractures more accurately.

Further, the image-based learning model with images of both vertebral bone and muscle showed better performance than the clinical variable–based models. This finding is consistent with a previous report that showed that information from the images of vertebral bones and muscles from CT scans can be used to predict major osteoporotic fractures and is comparable with FRAX [32]. Another group reported different algorithms by using opportunistic CT-based bone assessments for osteoporotic fracture prediction [33]. They showed that CT-based predictors (vertebral compression fractures, simulated DXA T-scores, and lumbar trabecular density) with metadata of age and sex showed better performance in AUROC than FRAX [33]. However, in that model, muscle information was not considered [33], which may further improve the performance. In addition to the attenuation information, we used information from the image itself on the quality of the bone and muscle structure, similar to the trabecular bone score [13]. The trabecular bone score is an algorithm used to calculate the microstructure of the bone based on DXA images [34]. More than 50% of the osteoporotic fractures occur in patients with a normal or osteopenic range of BMD [35], which implies that the microarchitecture of the bone is also a key determinant of bone strength [36]. Similarly, in our study, the model used the information on the qualities of bones and muscles from CT images, demonstrating the potential value of CT images that may include rich and various informative data for the metabolic diseases of bones and muscles.

We also observed that the performance did not significantly change when clinical variables were added to the image-only model. There is a possibility that information such as age and gender could already be reflected to some extent in the image itself [37]. Therefore, there could be an insignificant improvement in the performance because the information poses a redundant input to the model. It is widely accepted that there is a noticeable sex difference in the size of the vertebral body and paravertebral muscles [37], and BMI could be positively correlated with the size of the vertebrae and muscles. In addition, although the model was based on high-resolution peripheral quantitative CT, each bone has different characteristics according to age and sex, such as calcification and size, which could have influenced our analysis [38]. In addition, the vertebral endplate calcification increases with age, implying that age information can be reflected in the image [39]. In addition to that reported in previous studies, smoking and alcohol consumption status can be associated with low muscle mass [40], which may explain why adding simple clinical variables to the image may not significantly improve the model, as the image already contains some clinical information. The results are clinically promising, and they can be utilized in the future, as only opportunistic CT scans without detailed clinical variables may automatically provide the risk of osteoporotic fractures.

To extract pertinent information from each CT scan, we designed an image-only model to prevent overfitting and to focus on the essential regions. Since 3D CNN models, which have a large number of parameters to be optimized, tend to overfit the training data [41], the CNN encoder of our model took consecutive 2D images as its input data while keeping their sequential information with the RNN decoder [42]. The input processing strategy served as a robust data augmentation method because our model could exploit different 2D image sets from a single 3D CT scan at each training iteration. In addition, an attention module was applied to the CNN encoder to further enhance its robustness. The attention module automatically guided the image model to concentrate on essential regions [43] for the prediction of vertebral fractures. Thus, the attention CNN-RNN model avoids making predictions based on background regions, except for the vertebral body and paravertebral muscles. Unlike previous CNN model–based deep learning algorithms, which were limited to 2D X-ray analysis or bone texture analysis, our CNN-RNN model showed robust performance in fracture prediction. Owing to its design to mitigate the overfitting problem of conventional 3D CNN models [42], the CNN-RNN model could extract effective information from 3D CT images, which were intractable in previous approaches. In addition, the attention module forced our model to focus on important regions in the CT images by removing the effects of the background regions [43].

Our study has several limitations. The data set did not contain BMD due to the retrospective study design, which is an essential predictor for osteoporotic fracture. It was difficult to compare the clinical model containing BMD with the image model. The model showed a 5-year fracture prediction model instead of a 10-year model owing to the follow-up duration of the data set, which is relatively short to be utilized in real-world practice. Thus, due to the short time frame, we could not show the results for nonvertebral fractures because the number of cases was too small. In addition, the paravertebral muscles were included without distinction among the psoas, intervertebral, multifidus, longissimus, iliocostalis, and quadratus lumborum muscles. Therefore, it is difficult to interpret the contribution of each

XSL•FO
RenderX

muscle. In addition, the number of images in the development set may not be sufficient for model optimization. Moreover, the utilization could be low in various contrast settings because it was based on contrast CT scans. There was also the disparity in vertebral fracture incidence between the development and the external validation set, which may affect the external validity and generalizability of our fracture prediction model. The retrospective nature inherently carries the potential for selection bias, including confounding by indication. Although we have employed propensity score adjustment to mitigate this bias, it is important to acknowledge that residual bias may still be present. Another limitation is the exclusion of radiographic imaging data with poor quality from our models. This decision might have introduced detection bias such that it may have impacted the diagnostic accuracy of our models in correctly identifying positive versus negative fracture cases. Further, we could not assess the reproducibility of these measurements through interexaminer and intraexaminer κ value assessments, which could be considered a limitation of our study. Future prospective studies could benefit from including such reproducibility assessments.

Our study has several strengths. Our study was longitudinally designed to observe future fracture events in patients who did not have baseline fractures. Furthermore, in the development cohort, we used controls with matched clinical variables, which made it possible to attenuate the effects of the major clinical variables in the model. It was also externally validated, which helped prove the generalizability of the model. In addition, the model used the image itself as an input, which made it possible to utilize the information on vertebral bone and muscle quality and quantity. This inclusion of the muscle image reflected the interplay between muscle health and fracture risk. For instance, factors such as muscle mass and muscle steatosis, which are visible in CT images as darker and more heterogeneous areas compared to normal muscle, could be crucial inputs. These muscle attributes, automatically analyzed by the CNN, contribute significantly to the model's ability to discern patients at higher risk of fractures, offering a more comprehensive view than bone analysis alone. In addition, by sequentially applying bones and muscles to the model, it was possible to check the degree of contribution of muscles and bones to the model performance, thereby increasing the interpretability of the model. In addition, the differences in the clinical characteristics between development and external validation sets were purposefully leveraged to assess the generalizability of our model across populations with varying clinical profiles.

In this study, we showed that a deep learning model of the CNN-RNN structure based on CT images of the muscle and vertebral bone could help predict the risk of vertebral fractures. The model using images of the vertebral bone and muscle showed better performance than the model using images of the vertebral bone-only. This implies that the information from the muscle images provides additional key information for predicting fractures. In addition, the model using images showed better performance than the model using clinical variables, suggesting that images can provide useful information in addition to having known clinical variables. This study has clinical significance in suggesting that opportunistic CT screening with deep learning algorithms utilizing bone and muscle images may contribute to identifying patients with a high fracture risk in the future. Further prospective studies are needed to broaden the applicability of our model.

Acknowledgments

The study was funded by the National Research Foundation of Korea (grants 2020R1A2C2011587 and 2021R1A2C2003410).

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary table and figure.
[DOCX File , 64 KB - jmir_v26i1e48535_app1.docx]

References

1. Tran O, Silverman S, Xu X, Bonafede M, Fox K, McDermott M, et al. Long-term direct and indirect economic burden associated with osteoporotic fracture in US postmenopausal women. Osteoporos Int 2021 Jun;32(6):1195-1205 [FREE Full text] [doi: 10.1007/s00198-020-05769-3] [Medline: 33411007]
2. Williams SA, Daigle SG, Weiss R, Wang Y, Arora T, Curtis JR. Economic burden of osteoporosis-related fractures in the US Medicare population. Ann Pharmacother 2021 Jul;55(7):821-829 [FREE Full text] [doi: 10.1177/1060028020970518] [Medline: 33148010]
3. Ahn SH, Park S, Park SY, Yoo J, Jung H, Nho J, et al. Osteoporosis and osteoporotic fracture fact sheet in Korea. J Bone Metab 2020 Nov;27(4):281-290 [FREE Full text] [doi: 10.11005/jbm.2020.27.4.281] [Medline: 33317231]

XSL•FO
RenderX

4. Stone K, Seeley D, Lui L, Cauley J, Ensrud K, Browner W, Osteoporotic Fractures Research Group. BMD at multiple sites and risk of fracture of multiple types: long-term results from the Study of Osteoporotic Fractures. J Bone Miner Res 2003 Nov;18(11):1947-1954 [FREE Full text] [doi: 10.1359/jbmr.2003.18.11.1947] [Medline: 14606506]
5. Amarnath ALD, Franks P, Robbins JA, Xing G, Fenton JJ. Underuse and overuse of osteoporosis screening in a regional health system: a retrospective cohort study. J Gen Intern Med 2015 Dec;30(12):1733-1740 [FREE Full text] [doi: 10.1007/s11606-015-3349-8] [Medline: 25986135]
6. CT market outlook report. IMV. 2022 Jan 01. URL: https://imvinfo.com/product/2022-ct-market-outlook-report/ [accessed 2024-06-26]
7. Lenchik L, Weaver AA, Ward RJ, Boone JM, Boutin RD. Opportunistic screening for osteoporosis using computed tomography: state of the art and argument for paradigm shift. Curr Rheumatol Rep 2018 Oct 13;20(12):74 [FREE Full text] [doi: 10.1007/s11926-018-0784-7] [Medline: 30317448]
8. Pickhardt PJ, Pooler BD, Lauder T, del Rio AM, Bruce RJ, Binkley N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. Ann Intern Med 2013 Apr 16;158(8):588-595 [FREE Full text] [doi: 10.7326/0003-4819-158-8-201304160-00003] [Medline: 23588747]
9. Alacreu E, Moratal D, Arana E. Opportunistic screening for osteoporosis by routine CT in Southern Europe. Osteoporos Int 2017 Mar;28(3):983-990. [doi: 10.1007/s00198-016-3804-3] [Medline: 28108802]
10. Dutta P, Upadhyay P, De M, Khalkar R. Medical image analysis using deep convolutional neural networks: CNN architectures and transfer learning. 2020 Jun 9 Presented at: 2020 International Conference on Inventive Computation Technologies (ICICT); Feb 26-28; Coimbatore, India p. 175-180. [doi: 10.1109/icict48043.2020.9112469]
11. Kong SH, Lee J, Bae BU, Sung JK, Jung KH, Kim JH, et al. Development of a spine X-ray-based fracture prediction model using a deep learning algorithm. Endocrinol Metab 2022 Aug;37(4):674-683. [doi: 10.3803/enm.2022.1461]
12. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD. Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual x-ray absorptiometry. Radiology 2019 Nov;293(2):405-411. [doi: 10.1148/radiol.2019190201] [Medline: 31526255]
13. Kong SH, Hong N, Kim J, Kim DY, Kim JH. Application of the trabecular bone score in clinical practice. J Bone Metab 2021 May;28(2):101-113 [FREE Full text] [doi: 10.11005/jbm.2021.28.2.101] [Medline: 34130362]
14. Kong SH, Shin CS. Applications of machine learning in bone and mineral research. Endocrinol Metab 2021 Oct;36(5):928-937. [doi: 10.3803/enm.2021.1111]
15. Muehlematter UJ, Mannil M, Becker AS, Vokinger KN, Finkenstaedt T, Osterhoff G, et al. Vertebral body insufficiency fractures: detection of vertebrae at risk on standard CT images using texture analysis and machine learning. Eur Radiol 2019 May;29(5):2207-2217. [doi: 10.1007/s00330-018-5846-8] [Medline: 30519934]
16. Zhang S, Chen H, Xu H, Yi Y, Fang X, Wang S. Computed tomography-based paravertebral muscle density predicts subsequent vertebral fracture risks independently of bone mineral density in postmenopausal women following percutaneous vertebral augmentation. Aging Clin Exp Res 2022 Nov;34(11):2797-2805. [doi: 10.1007/s40520-022-02218-5] [Medline: 36001272]
17. Kim H, Kim C. Quality matters as much as quantity of skeletal muscle: clinical implications of myosteatosis in cardiometabolic health. Endocrinol Metab 2021 Dec;36(6):1161-1174. [doi: 10.3803/enm.2021.1348]
18. Genant H, Wu C, van Kuijk C, Nevitt M. Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res 1993 Sep;8(9):1137-1148 [FREE Full text] [doi: 10.1002/jbmr.5650080915] [Medline: 8237484]
19. Shin CS, Kim MJ, Shim SM, Kim JT, Yu SH, Koo BK, et al. The prevalence and risk factors of vertebral fractures in Korea. J Bone Miner Metab 2012 Mar;30(2):183-192. [doi: 10.1007/s00774-011-0300-x] [Medline: 21773702]
20. Kanis J, Johansson H, McCloskey E, Liu E, Åkesson KE, Anderson F, et al. Previous fracture and subsequent fracture risk: a meta-analysis to update FRAX. Osteoporos Int 2023 Dec;34(12):2027-2045 [FREE Full text] [doi: 10.1007/s00198-023-06870-z] [Medline: 37566158]
21. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W. CNN-RNN: A unified framework for multi-label image classification. 2016 Mar 20 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; March 20; Anaheim, CA. [doi: 10.1109/cvpr.2016.251]
22. Khaki S, Wang L, Archontoulis SV. A CNN-RNN framework for crop yield prediction. Front Plant Sci 2019;10:1750 [FREE Full text] [doi: 10.3389/fpls.2019.01750] [Medline: 32038699]
23. Dong H, Cortés YI, Shiau S, Yin M. Osteoporosis and fractures in HIV/hepatitis C virus coinfection: a systematic review and meta-analysis. AIDS 2014 Sep 10;28(14):2119-2131 [FREE Full text] [doi: 10.1097/QAD.0000000000000363] [Medline: 24977441]
24. Ganesan K, Jandu J, Anastasopoulou C, et al. Secondary osteoporosis. StatPearls [Internet] 2023 Apr 2:1-20. [Medline: 29262237]
25. Teng Z, Zhu Y, Teng Y, Long Q, Hao Q, Yu X, et al. The analysis of osteosarcopenia as a risk factor for fractures, mortality, and falls. Osteoporos Int 2021 Nov;32(11):2173-2183. [doi: 10.1007/s00198-021-05963-x] [Medline: 33877382]
26. Cedeno-Veloz B, López-Dóriga Bonnardeauxa P, Duque G. [Osteosarcopenia: A narrative review]. Rev Esp Geriatr Gerontol 2019;54(2):103-108. [doi: 10.1016/j.regg.2018.09.010] [Medline: 30471719]

XSL•FO
RenderX

27. Yagi M, Hosogane N, Watanabe K, Asazuma T, Matsumoto M, Keio Spine Research Group. The paravertebral muscle and psoas for the maintenance of global spinal alignment in patient with degenerative lumbar scoliosis. Spine J 2016 Apr;16(4):451-458. [doi: 10.1016/j.spinee.2015.07.001] [Medline: 26165478]
28. Habibi H, Takahashi S, Hoshino M, Takayama K, Sasaoka R, Tsujio T, et al. Impact of paravertebral muscle in thoracolumbar and lower lumbar regions on outcomes following osteoporotic vertebral fracture: a multicenter cohort study. Arch Osteoporos 2021 Jan 03;16(1):2. [doi: 10.1007/s11657-020-00866-6] [Medline: 33389230]
29. Choi M, Kim S, Park C, Malla H, Kim S. Cross-sectional area of the lumbar spine trunk muscle and posterior lumbar interbody fusion rate: a retrospective study. Clin Spine Surg 2017 Jul;30(6):E798-E803. [doi: 10.1097/BSD.0000000000000424] [Medline: 27623301]
30. Kim JY, Chae SU, Kim GD, Cha MS. Changes of paraspinal muscles in postmenopausal osteoporotic spinal compression fractures: magnetic resonance imaging study. J Bone Metab 2013 Nov;20(2):75-81 [FREE Full text] [doi: 10.11005/jbm.2013.20.2.75] [Medline: 24524061]
31. Lang T, Cauley J, Tylavsky F, Bauer D, Cummings S, Harris T, Health ABC Study. Computed tomographic measurements of thigh muscle cross-sectional area and attenuation coefficient predict hip fracture: the health, aging, and body composition study. J Bone Miner Res 2010 Mar;25(3):513-519 [FREE Full text] [doi: 10.1359/jbmr.090807] [Medline: 20422623]
32. Pickhardt PJ, Graffy PM, Zea R, Lee SJ, Liu J, Sandfort V, et al. Automated abdominal CT imaging biomarkers for opportunistic prediction of future major osteoporotic fractures in asymptomatic adults. Radiology 2020 Oct;297(1):64-72 [FREE Full text] [doi: 10.1148/radiol.2020200466] [Medline: 32780005]
33. Dagan N, Elnekave E, Barda N, Bregman-Amitai O, Bar A, Orlovsky M, et al. Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. Nat Med 2020 Jan;26(1):77-82. [doi: 10.1038/s41591-019-0720-z] [Medline: 31932801]
34. Pothuaud L, Carceller P, Hans D. Correlations between grey-level variations in 2D projection images (TBS) and 3D microarchitecture: applications in the study of human trabecular bone microarchitecture. Bone 2008 Apr;42(4):775-787. [doi: 10.1016/j.bone.2007.11.018] [Medline: 18234577]
35. Siris ES, Chen Y, Abbott TA, Barrett-Connor E, Miller PD, Wehren LE, et al. Bone mineral density thresholds for pharmacological intervention to prevent fractures. Arch Intern Med 2004 May 24;164(10):1108-1112. [doi: 10.1001/archinte.164.10.1108] [Medline: 15159268]
36. Silva B, Leslie W, Resch H, Lamy O, Lesnyak O, Binkley N, et al. Trabecular bone score: a noninvasive analytical method based upon the DXA image. J Bone Miner Res 2014 Mar;29(3):518-530 [FREE Full text] [doi: 10.1002/jbmr.2176] [Medline: 24443324]
37. Gilsanz V, Boechat MI, Gilsanz R, Loro ML, Roe TF, Goodman WG. Gender differences in vertebral sizes in adults: biomechanical implications. Radiology 1994 Mar;190(3):678-682. [doi: 10.1148/radiology.190.3.8115610] [Medline: 8115610]
38. Kazakia GJ, Nirody JA, Bernstein G, Sode M, Burghardt AJ, Majumdar S. Age- and gender-related differences in cortical geometry and microstructure: Improved sensitivity by regional analysis. Bone 2013 Feb;52(2):623-631 [FREE Full text] [doi: 10.1016/j.bone.2012.10.031] [Medline: 23142360]
39. Bernick S, Cailliet R. Vertebral end-plate changes with aging of human vertebrae. Spine (Phila Pa 1976) 1982;7(2):97-102. [doi: 10.1097/00007632-198203000-00002] [Medline: 7089697]
40. Zhong J, Xie W, Wang X, Dong X, Mo Y, Liu D, et al. The prevalence of sarcopenia among Hunan province community-dwelling adults aged 60 years and older and its relationship with lifestyle: diagnostic criteria from the Asian Working Group for Sarcopenia 2019 update. Medicina (Kaunas) 2022 Oct 30;58(11):1562 [FREE Full text] [doi: 10.3390/medicina58111562] [Medline: 36363519]
41. Kumawat S, Raman S. Lp-3dcnn: Unveiling local phase in 3D convolutional neural networks. 2019 Presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; January 9; Long Beach, CA, USA. [doi: 10.1109/cvpr.2019.00504]
42. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv Preprint posted online on December 11, 2014. [doi: 10.48550/arXiv.1412.3555]
43. Jetley S, Lord N, Lee N, Torr P. Learn to pay attention. ArXiv. Preprint posted online on April 26, 2018 2018. [doi: 10.48550/arXiv.1804.02391]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BMD: bone mineral density
CNN: convolutional neural network
CT: computed tomography
DXA: dual-energy X-ray absorptiometry
FRAX: fracture risk assessment
RNN: recurrent neural network

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.04.23; peer-reviewed by M Liu, SH Lee, R Mpofu; comments to author 07.12.23; revised version received 27.01.24; accepted 30.05.24; published 12.07.24.

Please cite as:
Kong SH, Cho W, Park SB, Choo J, Kim JH, Kim SW, Shin CS
A Computed Tomography–Based Fracture Prediction Model With Images of Vertebral Bones and Muscles by Employing Deep Learning: Development and Validation Study
J Med Internet Res 2024;26:e48535
URL: https://www.jmir.org/2024/1/e48535
doi:10.2196/48535
PMID:38995678

©Sung Hye Kong, Wonwoo Cho, Sung Bae Park, Jaegul Choo, Jung Hee Kim, Sang Wan Kim, Chan Soo Shin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Diagnostic Accuracy of a Mobile AI-Based Symptom Checker and a Web-Based Self-Referral Tool in Rheumatology: Multicenter Randomized Controlled Trial

Johannes Knitza[1,2,3,4], MD; Koray Tascilar[3,4], MD; Franziska Fuchs[3,4], MD; Jacob Mohn[3,4], MD; Sebastian Kuhn[1], MD; Daniela Bohr[3,4], MD; Felix Muehlensiepen[2,5,6], MPH, Dr rer medic; Christina Bergmann[3,4], MD; Hannah Labinsky[3,4], MD; Harriet Morf[3,4], MD; Elizabeth Araujo[3,4], MD; Matthias Englbrecht[7]; Wolfgang Vorbrüggen[7,8], Dr med; Cay-Benedict von der Decken[8,9,10], MD; Stefan Kleinert[8,11], MD; Andreas Ramming[3,4], MD; Jörg H W Distler[3,4,12], MD; Peter Bartz-Bazzanella[8,10], MD; Nicolas Vuillerme[2,13,14], PhD; Georg Schett[3,4], MD; Martin Welcker[8,15*], MD; Axel Hueber[3,4,16*], MD

[1]Institute for Digital Medicine, University Hospital Giessen-Marburg, Philipps University Marburg, Marburg, Germany
[2]AGEIS, Université Grenoble Alpes, Grenoble, France
[3]Department of Internal Medicine 3, Friedrich-Alexander-University Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany
[4]Deutsches Zentrum für Immuntherapie (DZI), Friedrich-Alexander-University Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany
[5]Center for Health Services Research, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany
[6]Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Potsdam, Germany
[7]Verein zur Förderung der Rheumatologie e.V., Würselen, Germany
[8]RheumaDatenRhePort (rhadar), Planegg, Germany
[9]Medizinisches Versorgungszentrum Stolberg, Stolberg, Germany
[10]Klinik für Internistische Rheumatologie, Rhein-Maas Klinikum, Würselen, Germany
[11]Rheumatologische Schwerpunktpraxis, Drs. Kleinert, Rapp, Ronneberger, Schuch u. Wendler, Erlangen, Germany
[12]Department of Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich Heine University, Düsseldorf, Germany
[13]Institut Universitaire de France, Paris, France
[14]LabCom Telecom4Health, Orange Labs & Université Grenoble Alpes, CNRS, Inria, Grenoble INP-UGA, Grenoble, France
[15]MVZ für Rheumatologie Dr. Martin Welcker GmbH, Planegg, Germany
[16]Division of Rheumatology, Klinikum Nürnberg, Paracelsus Medical University, Nürnberg, Germany
[*]these authors contributed equally

Corresponding Author:
Johannes Knitza, MD
Institute for Digital Medicine
University Hospital Giessen-Marburg
Philipps University Marburg
Baldingerstrasse
Marburg, 35043
Germany
Phone: 1 49 6421 ext 58
Email: johannes.knitza@uni-marburg.de

Abstract

Background: The diagnosis of inflammatory rheumatic diseases (IRDs) is often delayed due to unspecific symptoms and a shortage of rheumatologists. Digital diagnostic decision support systems (DDSSs) have the potential to expedite diagnosis and help patients navigate the health care system more efficiently.

Objective: The aim of this study was to assess the diagnostic accuracy of a mobile artificial intelligence (AI)–based symptom checker (Ada) and a web-based self-referral tool (Rheport) regarding IRDs.

XSL•FO
RenderX

Methods: A prospective, multicenter, open-label, crossover randomized controlled trial was conducted with patients newly presenting to 3 rheumatology centers. Participants were randomly assigned to complete a symptom assessment using either Ada or Rheport. The primary outcome was the correct identification of IRDs by the DDSSs, defined as the presence of any IRD in the list of suggested diagnoses by Ada or achieving a prespecified threshold score with Rheport. The gold standard was the diagnosis made by rheumatologists.

Results: A total of 600 patients were included, among whom 214 (35.7%) were diagnosed with an IRD. Most frequent IRD was rheumatoid arthritis with 69 (11.5%) patients. Rheport's disease suggestion and Ada's top 1 (D1) and top 5 (D5) disease suggestions demonstrated overall diagnostic accuracies of 52%, 63%, and 58%, respectively, for IRDs. Rheport showed a sensitivity of 62% and a specificity of 47% for IRDs. Ada's D1 and D5 disease suggestions showed a sensitivity of 52% and 66%, respectively, and a specificity of 68% and 54%, respectively, concerning IRDs. Ada's diagnostic accuracy regarding individual diagnoses was heterogenous, and Ada performed considerably better in identifying rheumatoid arthritis in comparison to other diagnoses (D1: 42%; D5: 64%). The Cohen κ statistic of Rheport for agreement on any rheumatic disease diagnosis with Ada D1 was 0.15 (95% CI 0.08-0.18) and with Ada D5 was 0.08 (95% CI 0.00-0.16), indicating poor agreement for the presence of any rheumatic disease between the 2 DDSSs.

Conclusions: To our knowledge, this is the largest comparative DDSS trial with actual use of DDSSs by patients. The diagnostic accuracies of both DDSSs for IRDs were not promising in this high-prevalence patient population. DDSSs may lead to a misuse of scarce health care resources. Our results underscore the need for stringent regulation and drastic improvements to ensure the safety and efficacy of DDSSs.

Trial Registration: German Register of Clinical Trials DRKS00017642; https://drks.de/search/en/trial/DRKS00017642

(J Med Internet Res 2024;26:e55542) doi:10.2196/55542

KEYWORDS

symptom checker; artificial intelligence; eHealth; diagnostic decision support system; rheumatology; decision support; decision; diagnostic; tool; rheumatologists; symptom assessment; resources; randomized controlled trial; diagnosis; decision support system; support system; support

Introduction

Symptoms caused by inflammatory rheumatic diseases (IRDs) are often unspecific and difficult to correctly interpret for patients [1] and even for experienced rheumatologists [2]. This diagnostic complexity frequently results in significant delay [3], which can diminish treatment efficacy and lead to progressive damage [4].

To address these challenges, a variety of freely available, patient-centered diagnostic decision support systems (DDSSs) have emerged and are increasingly being used by the general public [5] and patients with IRDs [6]. These DDSSs offer disease suggestions and advice for action within a few minutes and without any health care provider contact.

Rheport [7] is a web-based rheumatology referral system used in Germany to automatically triage appointments of new patients according to IRD probability [8,9]. To date, Rheport has been used to schedule more than 3000 appointments [10]. Ada [11], an artificial intelligence (AI)–based chatbot, is one of the most promising DDSS currently available. Multiple case-vignette studies showcased its high diagnostic accuracy [5,12], and more than 30 million symptom assessments have been completed in 130 countries [13].

Despite the expanding usage, little evidence is available regarding the accuracy of DDSSs in rheumatology [14-16]. To our knowledge, Rheport and Ada are among the most widely used DDSSs in rheumatology within Germany [10,13]. However, a direct comparative study between these 2 systems has yet to be conducted. Therefore, the aim of this analysis was to evaluate the diagnostic capability of Ada and Rheport in identifying IRDs.

Methods

Study Design and Participants

The study design for this pragmatic, prospective, multicenter, crossover randomized controlled trial (German Register of Clinical Trials DRKS00017642) has been described elsewhere in detail [9]. Results are presented according to the CONSORT-EHEALTH checklist (Multimedia Appendix 1) [17].

Adult patients with musculoskeletal symptoms who had been referred for the first time to 3 recruiting rheumatology outpatient clinics with a suspected diagnosis of an IRD were consecutively included in this study. Participants were instructed to enter the required data into both Ada and Rheport while waiting for their scheduled appointment with the rheumatologist. Assistance from support staff was available if needed. Patients were randomized 1:1 by computer-generated block randomization into group 1 (first Ada, then Rheport) or group 2 (first Rheport, then Ada), with each block comprising 100 patients. This crossover design was chosen to mitigate potential bias from a priming effect, where completing the first DDSS could influence responses to the second DDSS without the patient's awareness. For instance, a priming effect was previously observed where participants who answered questions about their religiosity before reporting their alcohol consumption indicated fewer drinks on peak drinking occasions [18]. A designated project manager, uninvolved in the recruitment process, was responsible for assigning patients to the intervention arms. The statistician

XSL•FO
RenderX

was kept blinded for group allocation. Assistance from the study personnel was available for DDSS completion when needed.

Ethical Considerations

The study was approved by the ethics committee of the medical faculty of the University of Erlangen-Nürnberg, Germany (106_19 Bc), and was conducted in compliance with the Declaration of Helsinki. All patients provided written informed consent.

The DDSSs

Rheport [7] comprises a static 23-item questionnaire designed to assess symptoms and generate an expert-derived weighted sum score [8,9]. Median completion time was 8 minutes [19]. A higher sum score correlates with an increased probability of an IRD. Rheumatologists utilizing this system can allocate slots for automatic patient scheduling. Based on the calculated IRD probability, patients are offered available appointments categorized into 4 urgency levels. Patients with scores below 1 are considered unlikely to have an IRD and do not receive an appointment. Those with a minimum score of 1 are considered likely to have an IRD and are enabled to book an appointment. The urgency levels are categorized as follows: patients with scores between 1 and 2.4 are considered intermediate, patients with scores between 2.4 and 4 are considered urgent, and patients with scores exceeding 4 are considered very urgent. Patients in the very urgent category should ideally receive appointments within 1 week [19]. Upon the acceptance of a proposed appointment by a patient, the rheumatologist is notified and provided with a structured summary report of the questionnaire. There were no changes to the Rheport algorithm during the study period.

Ada [11] is a native app and certified medical product, designed to cover a broad spectrum of symptoms and diseases. Programmed as a chatbot, Ada mimics traditional history taking by initially requesting basic health information, such as sex and age, followed by current symptoms. Based on these responses, the app generates individualized follow-up questions. Ada's diagnostic suggestions are driven by a Bayesian network that is continuously updated [20]. Upon the completion of symptom querying (median time: 7 minutes [19]), a summary report is generated, including (1) a summary of present, absent, and uncertain symptoms; (2) up to 5 disease suggestions with the corresponding probabilities, triage recommendations (eg, call an ambulance), and symptom relevance; and (3) access to basic information about the suggested diseases. Ada was regularly updated along the course of the study to ensure functionality.

Outcome

The primary end point of the study was concordant detection of any IRD diagnosis (including, eg, rheumatoid arthritis or systemic lupus erythematosus) by the DDSSs and the gold standard, that is, the rheumatologist's final diagnosis, reported on the discharge summary report and adjudicated by the attending head physician of the local rheumatology department. For Rheport, the detection of an IRD by the DDSS was defined as a sum score of 1 or higher. Regarding Ada, we analyzed whether there was an IRD diagnosis and whether the correct diagnosis was listed as the top diagnosis (Ada top 1 [D1]) or was listed at all among all suggested diagnoses (Ada top 5 [D5]).

Statistical Analysis

Descriptive characteristics are presented as median and IQR for interval data and as absolute (n) and relative frequency (percentage) for nominal data. The minimum necessary sample size for this study was 122 in order to detect a specificity and sensitivity of at least 70% for Ada or Rheport, with a type I error of 4.4% and type II error of 19% using a 1-sample test [21] against a benchmark accuracy of 50% based on a previous evaluation of DDSSs [14]. Operating characteristics of Ada and Rheport for a diagnosis of rheumatic disease was evaluated using sensitivity, specificity, negative predictive value, positive predictive value, and overall accuracy with respective 95% CIs. This evaluation was done both separately and for the combined use of the DDSSs. The agreement between the DDSSs was evaluated using the Cohen κ statistic, with values ≤0 indicating no agreement, 0.01-0.20 indicating none to slight agreement, 0.21-0.40 indicating fair agreement, 0.41-0.60 indicating moderate agreement, 0.61-0.80 indicating substantial agreement, and 0.81-1.00 indicating almost perfect agreement [22]. We evaluated the cumulative proportion of correct diagnoses using Ada with exact CIs. We used a binomial regression with a log-link function to calculate the risk ratio for correct identification of any IRD by Rheport in comparison to Ada when the respective DDSS was used first and when it was used after the crossover. We preferred this method over logistic regression since the interpretation of risk ratios is more intuitive than that for odds ratios with high-prevalence binary outcomes. All analyses were conducted using the open-source R software (version 4.1.0; R Foundation for Statistical Computing) running under RStudio (version 1.4.1103; RStudio).

Results

Participants

A total of 755 consecutive patients were approached between September 2019 and April 2021, of whom 654 (87%) agreed to participate and 600 (79.4%) were included in the analysis presented (Figure 1). The participation exceeded the minimal sample size calculation, since patients were very eager to participate and considered the study a welcome distraction during the waiting time for their appointment. Overall, 35.7% (214/600) of the patients were diagnosed with an IRD based on physicians' judgment.

The demographic characteristics of the patients are displayed in Table 1, and the physicians' final diagnoses are presented in Table 2.

XSL•FO
RenderX

Figure 1. Patient flow diagram.

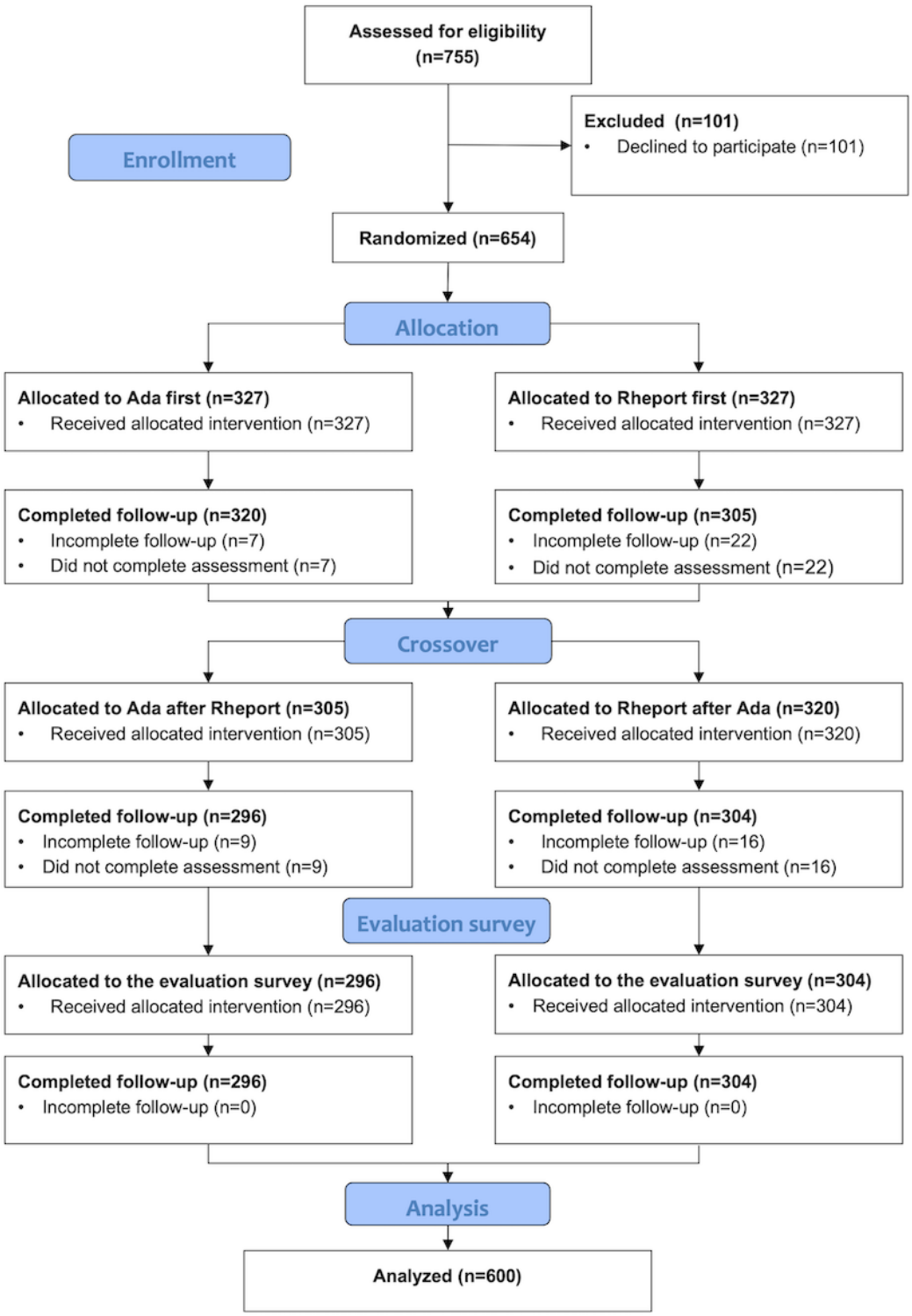

XSL•FO
RenderX

Table 1. Patient demographic characteristics according to the physicians' final diagnosis, reported on the discharge summary report.

Demographics	Final diagnosis		Total sample (n=600)
	IRD[a] (n=214)	Non-IRD (n=386)	
Age (years), median (IQR)	53.0 (38.0-64.0)	51.0 (37.0-59.0)	52.0 (37.0-61.0)
Sex (female), n (%)	131 (61.2)	287 (74.4)	418 (69.7)
Tender joint count 28, median (IQR)	1.0 (0.0-3.0)	0.0 (0.0-1.0)	0.0 (1.0-2.0)
Swollen joint count 28, median (IQR)	0.0 (0.0-2.0)	0.0 (0.0-0.0)	0.0 (0.0-0.0)
VAS[b] patient global (mm), median (IQR)	35.0 (9.5-60.0)	40.0 (10.0-60.0)	40.0 (10.0-60.0)
ESR[c] (mm/h), median (IQR)	13.0 (6.0-25.0)	8.0 (5.0-13.0)	9.0 (5.0-17.0)
CRP[d] (mg/L), median (IQR)	5.0 (5.0-8.8)	5.0 (3.1-5.0)	5.0 (3.2-5.0)
DAS28[e] (CRP), median (IQR)	3.0 (2.2-3.9)	2.2 (1.7-2.9)	2.5 (1.8-3.2)
RF[f] (positive), n/N (%)	36/196 (18.4)	34/350 (9.7)	70/546 (11.7)
ACPA[g] (positive), n/N (%)	20/192 (10.4)	5/336 (1.5)	25/528 (4.2)

[a]IRD: inflammatory rheumatic disease.
[b]VAS: visual analogue scale.
[c]ESR: erythrocyte sedimentation rate.
[d]CRP: C-reactive protein.
[e]DAS28: disease activity score 28.
[f]RF: rheumatoid factor.
[g]ACPA: anticitrullinated protein antibodies.

Table 2. Patients according to diagnostic categories.

Diagnostic categories	Patients (n=600), n (%)
IRD[a]	214 (35.7)
Axial spondyloarthritis	31 (5.2)
Connective tissue disease	22 (3.7)
Crystal arthropathies	8 (1.3)
Peripheral spondyloarthritis	3 (0.5)
Polymyalgia rheumatica	16 (2.7)
Psoriatic arthritis	31 (5.2)
Rheumatoid arthritis	69 (11.5)
Undifferentiated arthritis	19 (3.2)
Vasculitis	8 (1.3)
Other IRD	7 (1.2)
Non-IRD	386 (64.3)
Osteoarthritis	71 (11.8)
Fibromyalgia	37 (6.2)
Other noninflammatory, unclear diagnosis	278 (46.3)

[a]IRD: inflammatory rheumatic disease.

Diagnostic Accuracy of Ada and Rheport

Rheport showed an overall sensitivity of 62% and a specificity of 47% for IRDs (Figure 2). Ada's D1 and D5 disease suggestions showed a sensitivity of 52% and 66%, respectively, and a specificity of 68% and 54%, respectively, concerning IRDs (Table 3). The odds ratio for Rheport correctly suggesting a rheumatic disease diagnosis in comparison to Ada D5 as the first used DDSS was 0.89 (95% CI 0.83-0.97). When the initial DDSS was Ada, the accuracy of Ada D5 was 61% (95% CI 55%-66%) and the accuracy of Rheport was 53% (95% CI 47%-59%), whereas after the crossover, this odds ratio was 0.98

XSL•FO
RenderX

(95% CI 0.91-1.06) with corresponding accuracies of Ada D5 at 56% (95% CI 50%-61%) and of Rheport at 52% (95% CI 46%-58%).

Ada's diagnostic accuracy regarding individual diagnoses was heterogenous. Ada suggested the correct diagnosis of as top suggestion (Ada D1) in 42% (29/69) of patients with rheumatoid arthritis, and the correct diagnosis was suggested overall (Ada D5) in 64% (44/69); moreover, the first suggestion of Ada (Ada D1) was correct in 22% (14/65) of patients with spondyloarthritis (including axial spondyloarthritis, peripheral spondyloarthritis, and psoriatic arthritis), and the correct diagnosis was suggested overall (Ada D5) in 38% (25/65). The findings suggest that Ada performed considerably better in identifying rheumatoid arthritis in comparison to other diagnoses (see Figure 3).

Figure 2. Diagnostic properties of Ada and Rheport regarding IRDs according to the order of usage. Ada D1: Ada's top diagnosis; Ada D5: Ada's top 5 suggestions; IRD: inflammatory rheumatic disease; DDSS: diagnostic decision support system.

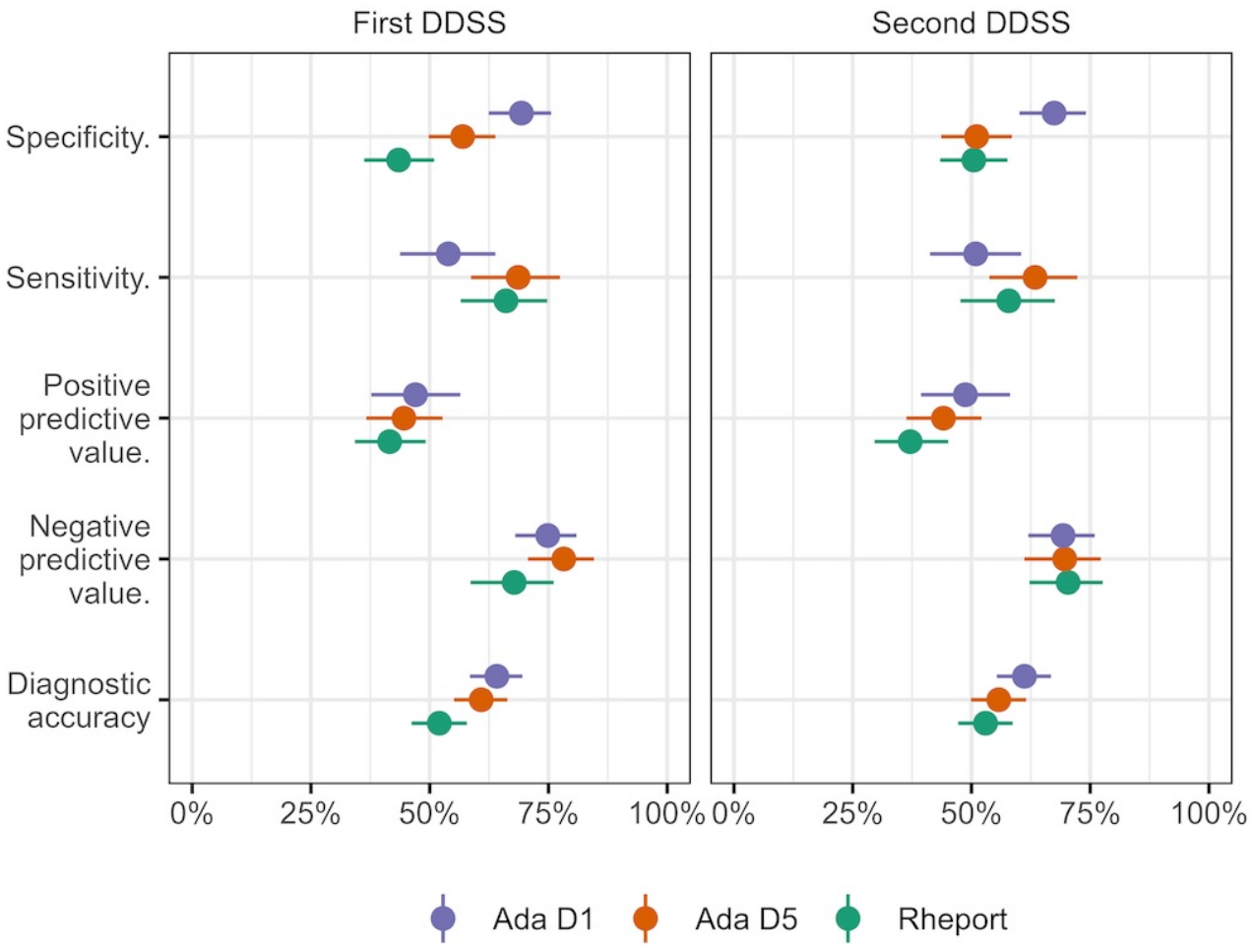

XSL•FO
RenderX

Table 3. Sensitivity, specificity, PPV[a], NPV[b], and overall accuracy of Ada and Rheport for the diagnosis of inflammatory rheumatic diseases including 95% CIs.

Characteristic	Diagnostic decision support system		
	Rheport	Ada D1[c]	Ada D5[d]
Total sample			
Sensitivity (%; 95% CI)	62 (55-69)	52 (45-59)	66 (59-72)
Specificity (%; 95% CI)	47 (42-52)	68 (63-73)	54 (49-59)
PPV (%; 95% CI)	39 (34-45)	48 (41-54)	44 (39-50)
NPV (%; 95% CI)	69 (63-75)	72 (67-77)	74 (69-79)
Accuracy (%; 95% CI)	52 (48-57)	63 (59-67)	58 (54-62)
Ada first			
Sensitivity (%; 95% CI)	58 (48-68)	54 (44-64)	69 (59-77)
Specificity (%; 95% CI)	50 (43-58)	69 (62-76)	57 (50-64)
PPV (%; 95% CI)	37 (30-45)	47 (38-56)	45 (37-53)
NPV (%; 95% CI)	70 (62-78)	75 (68-81)	78 (71-85)
Accuracy (%; 95% CI)	53 (47-59)	64 (58-70)	61 (55-66)
Rheport first			
Sensitivity (%; 95% CI)	66 (57-75)	51 (41-60)	63 (54-72)
Specificity (%; 95% CI)	43 (36-51)	67 (60-74)	51 (44-59)
PPV (%; 95% CI)	42 (34-49)	49 (39-58)	44 (36-52)
NPV (%; 95% CI)	68 (59-76)	69 (62-76)	70 (61-77)
Accuracy (%; 95% CI)	52 (46-58)	61 (55-67)	56 (50-61)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]Ada D1: using Ada's top suggestion only.

[d]Ada D5: using all suggestions provided by Ada.

XSL•FO
RenderX

Figure 3. Cumulative overall diagnostic accuracy of Ada for selected diseases according to diagnostic rank. FM: fibromyalgia; IRD: inflammatory rheumatic diseases; OA: osteoarthritis; overall: all patients with a final medical diagnosis; RA: rheumatoid arthritis; SpA: spondyloarthritis.

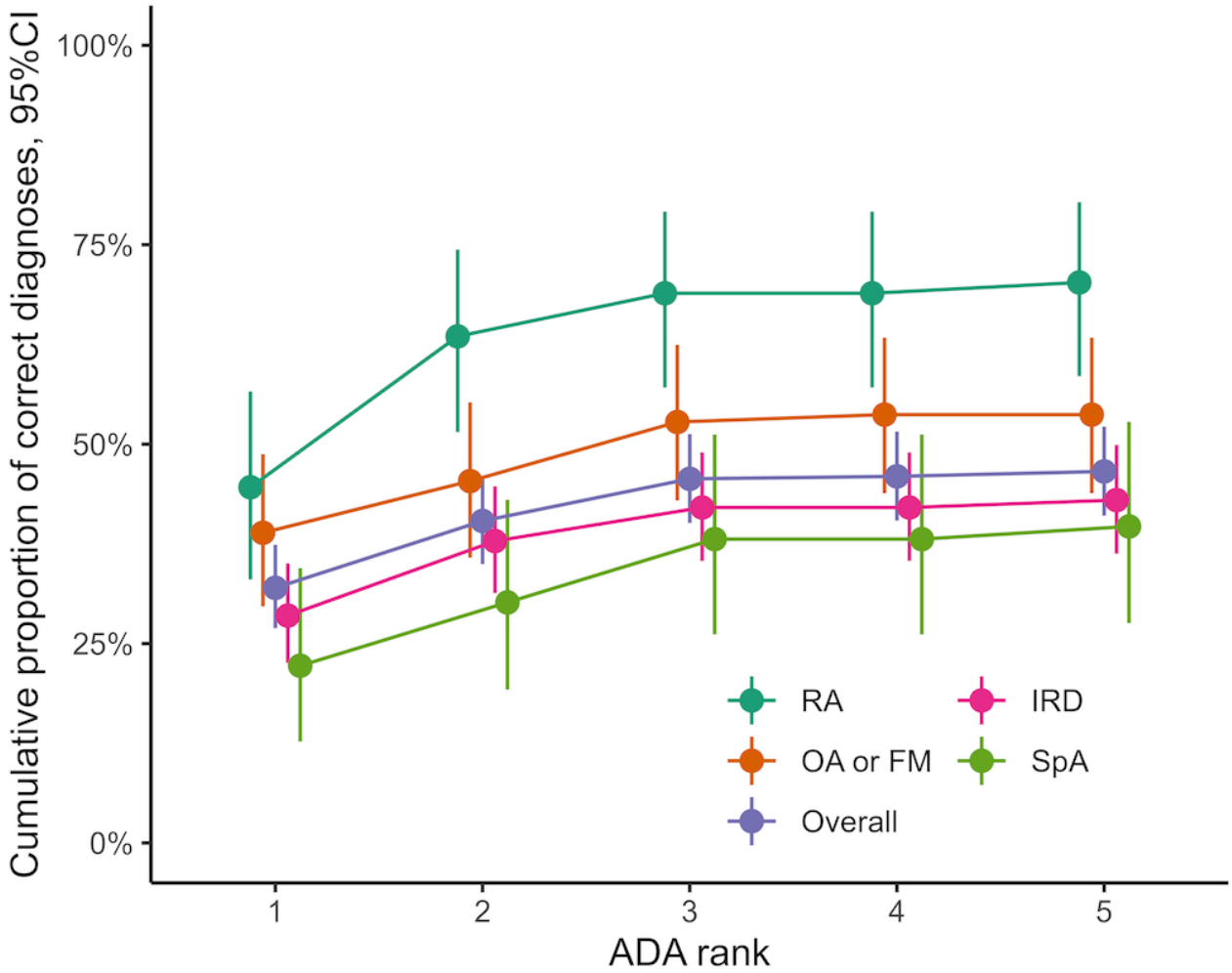

Agreement Between Rheport and Ada

The Cohen κ statistic of Rheport for agreement on any rheumatic disease diagnosis with Ada D1 was 0.15 (95% CI 0.08-0.18) and with Ada D5 was 0.08 (95% CI 0.00-0.16), indicating poor or nonexistent agreement for the presence of any rheumatic disease between the 2 DDSSs.

Discussion

Principal Findings

This prospective, multicenter randomized controlled trial investigated the diagnostic accuracy of 2 DDSSs regarding IRDs. Overall, the diagnostic accuracy of both DDSSs was limited. Rheport was less likely to correctly identify any IRD when used as the first DDSS; this could not be reproduced when it was used as the second DDSS. The diagnostic accuracy was comparably low with both tools, although Ada is an AI-based chatbot whereas Rheport is built on a simple weighted-sum-score questionnaire and expert opinions. The low negative predictive values of both Ada and Rheport suggest frequent errors when used as rheumatic screening tools. The low diagnostic accuracy is all the more alarming, as this study population already had a higher pretest probability. Patients were highly preselected with physicians explicitly only referring patients with a suspected IRD. Overall, these final results confirm the low diagnostic accuracy of both DDSSs observed in the interim analysis, which included 164 patients from a single participating center [9]. This study also confirms the case-dependent variations in the diagnostic accuracy of Ada, which was the highest for rheumatoid arthritis, in line with the results of a previous vignette-based study [23,24].

Two previous studies have demonstrated the strong user dependence of Ada's diagnostic accuracy [23,25]. The low DDSS accuracy for IRDs is in line with previous results from smaller studies. In a pilot study, Powley et al [15] showed that only 19% of patients with IRDs were correctly identified. Similarly, Proft et al [26] recently showed that only 19% of patients using an axial spondyloarthritis self-referral tool were actually correctly diagnosed. A general reason for the low diagnostic accuracy of symptom-based DDSSs in rheumatology could be the lack of available information compared to the physician. Ehrenstein et al [2] previously demonstrated that the diagnostic accuracy of experienced rheumatologists regarding correct identification of IRDs solely based on medical history was only 14%. The low accuracy of DDSSs poses substantial challenges, as inaccurate diagnoses can cause misutilization of scarce health care resources, anxiety among patients, and

XSL•FO
RenderX

frustration among health care professionals. Complementing subjective symptom descriptions by adding objective laboratory values obtained via self-sampling [27,28] could improve the accuracy of DDSS suggestions while preserving remote care advantages. Furthermore, the application of machine learning has proven effective in improving the diagnostic accuracy of the current Rheport algorithm, highlighted by an increase in the area under the receiver operating characteristic curve from 0.534 to 0.737 [29]. The top 5 most significant features identified by the best-performing logistic regression model for IRD classification included finger joint pain, elevated inflammatory marker levels, the presence of psoriasis, symptom duration, and female sex. Additionally, the integration of large language models such as ChatGPT could significantly improve the DDSS performance [30]. In a recent study, we demonstrated that ChatGPT achieved diagnostic accuracy for rheumatic diseases comparable to that of experienced rheumatologists when both were provided with identical summary reports generated by patients using Ada. Additionally, we believe that usability may be improved by incorporating the free-text input and voice-enabled features of ChatGPT. A scoping review has highlighted the poor diagnostic accuracy, lack of evidence, and absence of regulation for patient-facing DDSSs [31]. To address these issues, we echo the calls for the implementation of stricter regulatory frameworks, certification procedures, and ongoing monitoring to close these regulatory gaps [31].

Limitations

A main limitation of the study is the fact that patients were already screened by referring physicians, causing a much higher a priori chance of having an IRD. Furthermore, help from assisting personnel was available if needed for DDSS completion. To our knowledge, this is the largest comparative DDSS trial with real patients; however, the results of this study are not automatically transferable to other disciplines, languages, patient groups, and DDSSs. We therefore call for future studies involving real patients to build more solid evidence.

Conclusions

Overall, the diagnostic accuracy and agreement of both DDSSs regarding IRDs were limited. Improvements are needed to ensure DDSS safety and efficacy. The results suggest that physicians and the complex process of establishing a medical diagnosis cannot be replaced by an algorithm-based or AI-based DDSS. Future studies are needed to evaluate the generalizability of our findings.

Acknowledgments

We thank all the patients who participated as well as all physicians who supported the study. This work is part of the PhD thesis of JK. Open-access funding was provided by the Open Access Publishing Fund of Philipps-Universität Marburg. This study was partially supported by Novartis Pharma GmbH, Nürnberg, Germany (grant 33419272), and the European Health and Digital Executive Agency (HADEA) under grant agreement 101080711 (SPIDeRR) that provided support to JK. Novartis employees were not involved in the design and conduct of the study. They did not contribute to the collection, analysis, and interpretation of data. They did not support the authors in the development of the manuscript.

Conflicts of Interest

JK and MW have received research support from Novartis Pharma GmbH. Qinum and RheumaDatenRhePort developed and hold the rights for Rheport. WV, CBvdD, S Kleinert, PB-B, and MW are members of RheumaDatenRhePort. WV, CBvdD, and PB-B were involved in the development of Rheport. JK is a member of the scientific board of RheumaDatenRhePort. S Kuhn is a founder and shareholder of MED.digital. PB-B has received speaker honoraria from AbbVie, Boehringer Ingelheim, Chugai/Roche, Janssen-Cilag, Novartis, and Pfizer.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1066 KB - jmir_v26i1e55542_app1.pdf]

References

1. Simons G, Belcher J, Morton C, Kumar K, Falahee M, Mallen CD, et al. Symptom recognition and perceived urgency of help-seeking for rheumatoid arthritis and other diseases in the general public: a mixed method approach. Arthritis Care Res (Hoboken) 2017 May;69(5):633-641 [FREE Full text] [doi: 10.1002/acr.22979] [Medline: 27389847]
2. Ehrenstein B, Pongratz G, Fleck M, Hartung W. The ability of rheumatologists blinded to prior workup to diagnose rheumatoid arthritis only by clinical assessment: a cross-sectional study. Rheumatology (Oxford) 2018 Sep 01;57(9):1592-1601. [doi: 10.1093/rheumatology/key127] [Medline: 29850885]
3. Sørensen J, Hetland ML, All Departments of Rheumatology in Denmark. Diagnostic delay in patients with rheumatoid arthritis, psoriatic arthritis and ankylosing spondylitis: results from the Danish nationwide DANBIO registry. Ann Rheum Dis 2015 Mar;74(3):e12 [FREE Full text] [doi: 10.1136/annrheumdis-2013-204867] [Medline: 24534758]
4. Burgers LE, Raza K, van der Helm-van Mil AH. Window of opportunity in rheumatoid arthritis - definitions and supporting evidence: from old to new perspectives. RMD Open 2019;5(1):e000870 [FREE Full text] [doi: 10.1136/rmdopen-2018-000870] [Medline: 31168406]

XSL•FO
RenderX

5. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One 2021 Jul 15;16(7):e0254088 [FREE Full text] [doi: 10.1371/journal.pone.0254088] [Medline: 34265845]
6. Kernder A, Morf H, Klemm P, Vossen D, Haase I, Mucke J, et al. Digital rheumatology in the era of COVID-19: results of a national patient and physician survey. RMD Open 2021 Feb;7(1):e001548 [FREE Full text] [doi: 10.1136/rmdopen-2020-001548] [Medline: 33622673]
7. Rheport. URL: https://www.rheport.de/ [accessed 2024-06-13]
8. Kleinert S, Bartz-Bazzanella P, von der Decken C, Knitza J, Witte T, Fekete SP, et al. A real-world rheumatology registry and research consortium: the German RheumaDatenRhePort (RHADAR) Registry. J Med Internet Res 2021 May 20;23(5):e28164 [FREE Full text] [doi: 10.2196/28164] [Medline: 34014170]
9. Knitza J, Mohn J, Bergmann C, Kampylafka E, Hagen M, Bohr D, et al. Accuracy, patient-perceived usability, and acceptance of two symptom checkers (Ada and Rheport) in rheumatology: interim results from a randomized controlled crossover trial. Arthritis Res Ther 2021 Apr 13;23(1):112 [FREE Full text] [doi: 10.1186/s13075-021-02498-8] [Medline: 33849654]
10. Benesova K, Lorenz H, Lion V, Voigt A, Krause A, Sander O, et al. Early recognition and screening consultation: a necessary way to improve early detection and treatment in rheumatology? : Overview of the early recognition and screening consultation models for rheumatic and musculoskeletal diseases in Germany [Article in German]. Z Rheumatol 2019 Oct;78(8):722-742. [doi: 10.1007/s00393-019-0683-y] [Medline: 31468170]
11. Ada. URL: https://ada.com/ [accessed 2024-06-13]
12. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? a clinical vignettes comparison to GPs. BMJ Open 2020 Dec 16;10(12):e040269 [FREE Full text] [doi: 10.1136/bmjopen-2020-040269] [Medline: 33328258]
13. Butcher M. Ada Health built an AI-driven startup by moving slowly and not breaking things. TechCrunch. 2020 Mar 5. URL: https://techcrunch.com/2020/03/05/move-slow-and-dont-break-things-how-to-build-an-ai-driven-startup/ [accessed 2021-01-29]
14. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 08;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
15. Powley L, McIlroy G, Simons G, Raza K. Are online symptoms checkers useful for patients with inflammatory arthritis? BMC Musculoskelet Disord 2016 Aug 24;17(1):362 [FREE Full text] [doi: 10.1186/s12891-016-1189-2] [Medline: 27553253]
16. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023 Jun 02;25:e43803 [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]
17. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
18. Rodriguez LM, Neighbors C, Foster DW. Priming effects of self-reported drinking and religiosity. Psychol Addict Behav 2014 Mar;28(1):1-9 [FREE Full text] [doi: 10.1037/a0031828] [Medline: 23528191]
19. Knitza J, Muehlensiepen F, Ignatyev Y, Fuchs F, Mohn J, Simon D, et al. Patient's perception of digital symptom assessment technologies in rheumatology: results from a multicentre study. Front Public Health 2022 Feb 22;10:844669 [FREE Full text] [doi: 10.3389/fpubh.2022.844669] [Medline: 35273944]
20. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence-based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. JMIR Hum Factors 2020 Jul 10;7(3):e19713 [FREE Full text] [doi: 10.2196/19713] [Medline: 32540836]
21. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. J Clin Diagn Res 2016 Oct;10(10):YE01-YE06 [FREE Full text] [doi: 10.7860/JCDR/2016/18129.8744] [Medline: 27891446]
22. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
23. Knitza J, Tascilar K, Gruber E, Kaletta H, Hagen M, Liphardt A, et al. Accuracy and usability of a diagnostic decision support system in the diagnosis of three representative rheumatic diseases: a randomized controlled trial among medical students. Arthritis Res Ther 2021 Sep 06;23(1):233 [FREE Full text] [doi: 10.1186/s13075-021-02616-6] [Medline: 34488887]
24. Knitza J, Hueber AJ. Reply to the comment on "Accuracy and usability of a diagnostic decision support system in the diagnosis of three representative rheumatic diseases: a randomized controlled trial among medical students". Arthritis Res Ther 2022 May 24;24(1):123 [FREE Full text] [doi: 10.1186/s13075-022-02806-w] [Medline: 35610629]
25. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res 2019 Oct 29;3(4):e13863 [FREE Full text] [doi: 10.2196/13863] [Medline: 31663858]

XSL•FO
RenderX

26. Proft F, Spiller L, Redeker I, Protopopov M, Rodriguez VR, Muche B, et al. Comparison of an online self-referral tool with a physician-based referral strategy for early recognition of patients with a high probability of axial spa. Semin Arthritis Rheum 2020 Oct;50(5):1015-1021. [doi: 10.1016/j.semarthrit.2020.07.018] [Medline: 32911279]
27. Morf H, Krusche M, Knitza J. Patient self-sampling: a cornerstone of future rheumatology care? Rheumatol Int 2021 Jun 10;41(6):1187-1188 [FREE Full text] [doi: 10.1007/s00296-021-04853-z] [Medline: 33839886]
28. Knitza J, Knevel R, Raza K, Bruce T, Eimer E, Gehring I, JPAST Group. Toward earlier diagnosis using combined ehealth tools in rheumatology: the joint pain assessment scoring tool (JPAST) project. JMIR Mhealth Uhealth 2020 May 15;8(5):e17507 [FREE Full text] [doi: 10.2196/17507] [Medline: 32348258]
29. Knitza J, Janousek L, Kluge F, von der Decken CB, Kleinert S, Vorbrüggen W, et al. Machine learning-based improvement of an online rheumatology referral and triage system. Front Med (Lausanne) 2022 Jul 22;9:954056 [FREE Full text] [doi: 10.3389/fmed.2022.954056] [Medline: 35935756]
30. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024 Feb;44(2):303-306 [FREE Full text] [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]
31. Müller R, Klemmt M, Ehni H, Henking T, Kuhnmünch A, Preiser C, et al. Ethical, legal, and social aspects of symptom checker applications: a scoping review. Med Health Care Philos 2022 Dec;25(4):737-755 [FREE Full text] [doi: 10.1007/s11019-022-10114-y] [Medline: 36181620]

Abbreviations

AI: artificial intelligence
D1: top 1
D5: top 5
DDSS: diagnostic decision support system
IRD: inflammatory rheumatic disease

Edited by E Sezgin; submitted 15.12.23; peer-reviewed by N Martelli, L Klareskog, LI Hanke; comments to author 13.04.24; revised version received 24.05.24; accepted 29.05.24; published 23.07.24.

Please cite as:
Knitza J, Tascilar K, Fuchs F, Mohn J, Kuhn S, Bohr D, Muehlensiepen F, Bergmann C, Labinsky H, Morf H, Araujo E, Englbrecht M, Vorbrüggen W, von der Decken CB, Kleinert S, Ramming A, Distler JHW, Bartz-Bazzanella P, Vuillerme N, Schett G, Welcker M, Hueber A
Diagnostic Accuracy of a Mobile AI-Based Symptom Checker and a Web-Based Self-Referral Tool in Rheumatology: Multicenter Randomized Controlled Trial
J Med Internet Res 2024;26:e55542
URL: https://www.jmir.org/2024/1/e55542
doi:10.2196/55542
PMID:

©Johannes Knitza, Koray Tascilar, Franziska Fuchs, Jacob Mohn, Sebastian Kuhn, Daniela Bohr, Felix Muehlensiepen, Christina Bergmann, Hannah Labinsky, Harriet Morf, Elizabeth Araujo, Matthias Englbrecht, Wolfgang Vorbrüggen, Cay-Benedict von der Decken, Stefan Kleinert, Andreas Ramming, Jörg H W Distler, Peter Bartz-Bazzanella, Nicolas Vuillerme, Georg Schett, Martin Welcker, Axel Hueber. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Building Dual AI Models and Nomograms Using Noninvasive Parameters for Aiding Male Bladder Outlet Obstruction Diagnosis and Minimizing the Need for Invasive Video-Urodynamic Studies: Development and Validation Study

Chung-You Tsai[1,2], PhD; Jing-Hui Tian[3], MS; Chien-Cheng Lee[2], PhD; Hann-Chorng Kuo[4], MD

[1]Divisions of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei, Taiwan

[2]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

[3]Department of Medical Research, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan

[4]Department of Urology, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation and Tzu Chi University, Hualien, Taiwan

Corresponding Author:
Hann-Chorng Kuo, MD
Department of Urology, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation and Tzu Chi University
707, Section 3, Chung-Yang Road
Hualien, 970
Taiwan
Phone: 886 3 856 1825
Fax: 886 3 856 0794
Email: hck@tzuchi.com.tw

Abstract

Background: Diagnosing underlying causes of nonneurogenic male lower urinary tract symptoms associated with bladder outlet obstruction (BOO) is challenging. Video-urodynamic studies (VUDS) and pressure-flow studies (PFS) are both invasive diagnostic methods for BOO. VUDS can more precisely differentiate etiologies of male BOO, such as benign prostatic obstruction, primary bladder neck obstruction, and dysfunctional voiding, potentially outperforming PFS.

Objective: These examinations' invasive nature highlights the need for developing noninvasive predictive models to facilitate BOO diagnosis and reduce the necessity for invasive procedures.

Methods: We conducted a retrospective study with a cohort of men with medication-refractory, nonneurogenic lower urinary tract symptoms suspected of BOO who underwent VUDS from 2001 to 2022. In total, 2 BOO predictive models were developed—1 based on the International Continence Society's definition (International Continence Society–defined bladder outlet obstruction; ICS-BOO) and the other on video-urodynamic studies–diagnosed bladder outlet obstruction (VBOO). The patient cohort was randomly split into training and test sets for analysis. A total of 6 machine learning algorithms, including logistic regression, were used for model development. During model development, we first performed development validation using repeated 5-fold cross-validation on the training set and then test validation to assess the model's performance on an independent test set. Both models were implemented as paper-based nomograms and integrated into a web-based artificial intelligence prediction tool to aid clinical decision-making.

Results: Among 307 patients, 26.7% (n=82) met the ICS-BOO criteria, while 82.1% (n=252) were diagnosed with VBOO. The ICS-BOO prediction model had a mean area under the receiver operating characteristic curve (AUC) of 0.74 (SD 0.09) and mean accuracy of 0.76 (SD 0.04) in development validation and AUC and accuracy of 0.86 and 0.77, respectively, in test validation. The VBOO prediction model yielded a mean AUC of 0.71 (SD 0.06) and mean accuracy of 0.77 (SD 0.06) internally, with AUC and accuracy of 0.72 and 0.76, respectively, externally. When both models' predictions are applied to the same patient, their combined insights can significantly enhance clinical decision-making and simplify the diagnostic pathway. By the dual-model prediction approach, if both models positively predict BOO, suggesting all cases actually resulted from medication-refractory primary bladder neck obstruction or benign prostatic obstruction, surgical intervention may be considered. Thus, VUDS might be unnecessary for 100 (32.6%) patients. Conversely, when ICS-BOO predictions are negative but VBOO predictions are positive, indicating varied etiology, VUDS rather than PFS is advised for precise diagnosis and guiding subsequent therapy, accurately identifying 51.1% (47/92) of patients for VUDS.

XSL•FO
RenderX

Conclusions: The 2 machine learning models predicting ICS-BOO and VBOO, based on 6 noninvasive clinical parameters, demonstrate commendable discrimination performance. Using the dual-model prediction approach, when both models predict positively, VUDS may be avoided, assisting in male BOO diagnosis and reducing the need for such invasive procedures.

(J Med Internet Res 2024;26:e58599) doi:10.2196/58599

KEYWORDS

bladder outlet obstruction; lower urinary tract symptoms; machine learning; nomogram; artificial intelligence; video urodynamic study

Introduction

Diagnosing the underlying causes of nonneurogenic male lower urinary tract symptoms (LUTS) associated with bladder outlet obstruction (BOO) is complex and cannot rely solely on symptomatology [1]. The pathophysiology of male BOO encompasses both anatomical and functional obstruction subtypes, such as benign prostatic obstruction (BPO), urethral stricture (US), primary bladder neck obstruction (PBNO), and dysfunctional voiding (DV) [2]. To accurately determine the pathophysiology behind BOO, it is crucial to conduct a comprehensive urodynamic study (UDS). These studies include 2 invasive diagnostic approaches—pressure-flow studies (PFS) or video-urodynamic studies (VUDS). PFS assesses bladder and outlet dynamics during urination [3,4]. Nevertheless, VUDS is a more precise and advanced examination that synchronously records PFS data plus real-time fluoroscopic imaging (x-ray) during both the storage and emptying phases of the bladder, providing a more detailed understanding of the underlying dysfunction with both functional and anatomical insights [5,6]. However, due to the higher costs associated with VUDS, as well as the requirements for specialized x-ray equipment and radiation protection, it is typically performed in tertiary medical centers and is not routinely available in all hospital settings.

PFS is crucial for evaluating bladder outlet dynamics with measurements of bladder pressure and urinary flow. According to the International Continence Society (ICS), male BOO is confirmed when the BOO index reaches 40 or higher (detailed in the “Methods” section) [7]. However, PFS falls short in distinguishing specific pathophysiologies such as PBNO or DV. VUDS, in contrast, offers a more comprehensive analysis of male LUTS suggestive of BOO, particularly in cases of PBNO or DV [8]. PBNO involves the failure of the bladder neck’s smooth muscle to adequately open during voiding. In contrast, DV is characterized by an intermittent or fluctuating urinary flow, typically due to inadequate or inconsistent relaxation of the external urethral sphincter during the process of voiding in men who do not have neurological disorders [7]. This distinction between PBNO and DV is critical since their treatment strategies differ markedly [9,10]. Beyond medication, PBNO often necessitates surgical intervention, whereas DV is primarily managed with behavioral therapies and physical therapies instead of surgery, highlighting the need for precise diagnostic methods.

From the clinical decision-making perspective, there exists a dilemma for men with medication-refractory nonneurogenic LUTS who are suspected of having BOO. The following question arises: which diagnostic UDS is most appropriate for the patient? (1) Should one take the first pathway and perform PFS only to potentially obtain an accurate diagnosis? (2) Or should one follow the second pathway, starting with PFS and proceeding to VUDS if a definitive diagnosis is not achieved? (3) Alternatively, is the third pathway preferable, where one directly conducts VUDS for a detailed diagnosis?

Given the invasive nature of PFS and VUDS, which can lead to discomfort and the risk of urinary tract infection, their use may not be feasible for all men with LUTS. These procedures are especially impractical for individuals who are physically frail. With these considerations, it is desirable to develop artificial intelligence (AI) models that leverage noninvasive clinical parameters to aid in the decision-making process. The aim is to predict the possible pathophysiology of the patient, reducing the need for invasive examinations, or to precisely determine which patients require VUDS. This approach would reserve VUDS for those who would benefit most from its detailed analysis. Although noninvasive methods and parameters have been proposed to predict the likelihood of BOO in male common or treatment-naïve LUTS [11-17], to our knowledge, no existing studies have developed prediction models using VUDS as the diagnostic standard, especially for those with medication-refractory nonneurogenic LUTS suspected of BOO. Addressing this gap, our goal is to develop and validate 2 clinical nomograms and AI models that can predict both International Continence Society–defined bladder outlet obstruction (ICS-BOO) and video-urodynamic studies–diagnosed bladder outlet obstruction (VBOO) using only noninvasive clinical parameters, potentially streamlining the diagnostic process for this patient group.

Methods

Overview

The study flow diagram and methodological framework for the development of BOO predictive models are illustrated in Figure 1. This outlines the process from the initial patient selection based on inclusion and exclusion criteria, through the development and test validation of various machine learning (ML) algorithms, to the final deployment in the form of both paper-based nomograms and an integrated web-based clinical decision support tool.

XSL•FO
RenderX

Figure 1. Study flow diagram and methodological framework for bladder outlet obstruction predictive model development. This diagram illustrates the methodological approach to constructing predictive models for bladder outlet obstruction. BOO: bladder outlet obstruction; ICS-BOO: International Continence Society–defined bladder outlet obstruction; IPP: intravesical prostatic protrusion; IPSS: International Prostate Symptom Score; LUTS: lower urinary tract symptoms; PSA: prostate-specific antigen; PU angle: prostatic urethral angle; PVR: postvoid residual; Qmax: maximum urinary flow rate; ROC: receiver operating characteristic; TPV: total prostate volume; TZI: transitional zone index; V/S ratio: voiding to storage score ratio; VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction; VUDS: video-urodynamic studies; Vvol: voided volume.

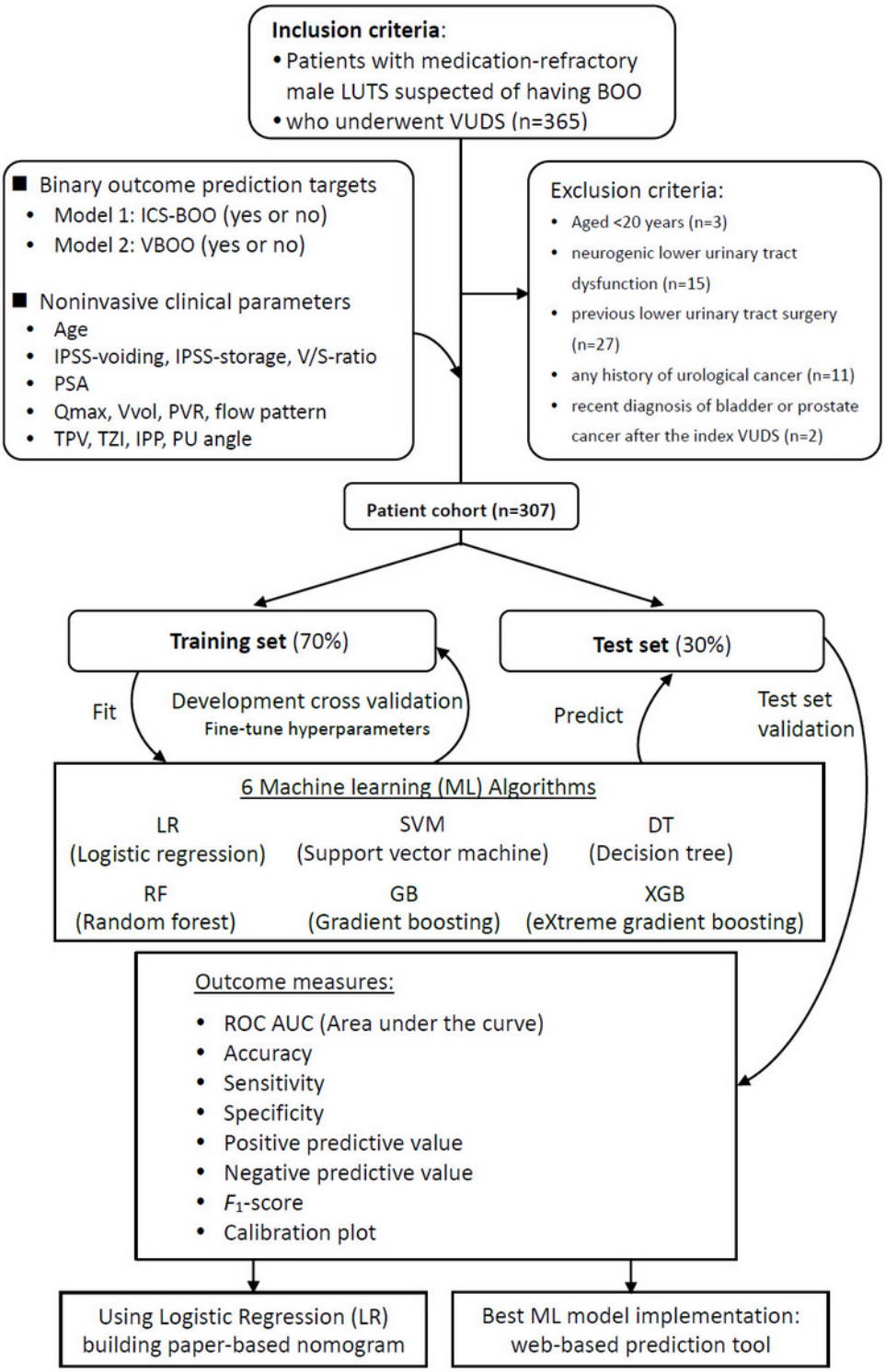

Patient Population

This retrospective study encompassed a cohort of male patients with medication-refractory, nonneurogenic LUTS who were suspected of having BOO and underwent VUDS at a single medical center from January 2001 to May 2022.

The inclusion criteria targeted men who had been on continuous medication for over 3 months without a response to at least one

XSL•FO
RenderX

type of LUTS treatment, including alpha-blockers, 5-alpha-reductase inhibitors, or antimuscarinics, before undergoing VUDS (Multimedia Appendix 1). Exclusion criteria included patients younger than 20 years, those with neurogenic lower urinary tract dysfunction, previous lower urinary tract surgery, any history of urological cancer, or a recent diagnosis of bladder or prostate cancer after the index VUDS.

Noninvasive Variables and VUDS Procedure

In daily clinical practice for men with LUTS, we routinely conduct history-taking, physical examinations, and record the International Prostate Symptom Score (IPSS). We also measure prostate-specific antigen levels and perform a transrectal ultrasound of the prostate along with uroflowmetry. The noninvasive clinical parameters chosen to construct the predictive models, serving as features for ML, are derived from these evaluations. These parameters include uroflowmetry measures such as maximum urinary flow rate (Qmax; measured in mL/s), voided volume (Vvol; measured in mL), postvoid residual (measured in mL), and flow pattern (qualitatively assessed). IPSS components comprise IPSS-voiding, IPSS-storage, and the IPSS-voiding to storage score ratio. Transrectal ultrasound of the prostate measures include total prostate volume (TPV; measured in mL), transitional zone index (defined as the percentage of transitional zone volume to TPV), intravesical prostatic protrusion (IPP; measured in cm), and prostatic urethral angle (PUA; measured in degrees).

During the VUDS procedure, we recorded bladder sensation, capacity, compliance, detrusor activity, and the maximum detrusor pressure at Qmax (Pdet.Qmax). In addition, electromyography and fluoroscopy imaging (x-ray) during both the storage and voiding phases were recorded using the Laborie system. All VUDS reports were collected and reviewed by a single urology professor (HCK) with more than 30 years of experience in VUDS to ensure consistency and reliability in the interpretation of these complex studies.

Definition of 2 Prediction Targets

In this study, 2 distinct prediction targets were defined for the development of the predictive models. The first model, "model 1," is designed to predict ICS-BOO as a binary outcome. This prediction is based on the ICS's criteria for the bladder outlet obstruction index (BOOI) [7], also referred to as the Abrams-Griffiths number. The BOOI is calculated using the formula "Pdet.Qmax-2×Qmax." A BOOI threshold value of 40 or higher is defined as obstructive, thus identifying the presence of ICS-BOO. Furthermore, a BOOI between 20 and 40 falls into an equivocal zone according to the ICS's criteria.

The second model, "model 2," is developed to predict VBOO as a binary outcome. VBOO includes a range of conditions contributing to male BOO, such as BPO, PBNO, DV, and US. Both PBNO and DV are types of functional BOO, as detailed in sections Bladder Outlet Obstruction (BOO), Dysfunctional Voiding, Detrusor Sphincter Dyssynergia, and Primary Bladder Neck Obstruction (nonneurogenic) of the ICS Standardisation Report for adult male LUTS [7]. The ICS Standardisation Reports also indicate that VUDS is necessary for diagnosing PBNO or DV, as these conditions cannot be adequately identified by PFS alone [7]. Some types of PBNO or DV would not reveal a classical "high-pressure, low-flow pattern." This means that patients with a BOOI<40, though not meeting the ICS-BOO criteria, could still be diagnosed with PBNO or DV as BOO through VUDS. Consequently, it is possible that a larger number of patients within the same cohort may be diagnosed with VBOO compared to ICS-BOO.

Model Derivations With Development and Test Validation

In the derivation of our predictive models, the data set from the entire cohort was randomly split once, allocating 70% to the training set and 30% to the test set to support both development and test validation processes. To develop the 2 prediction models, ML approaches using logistic regression (LR) were used. Variable selection for these models was conducted using a backward stepwise approach based on Akaike's information criterion (AIC). The β coefficients from the final regression models were then used to construct nomograms, which provide a paper-based, visual representation of the prediction models and facilitate their application in clinical practice. In addition, web-based prediction tools were also deployed to assist in AI-driven decision-making.

The development validation was rigorously conducted using a repeated 5-fold cross-validation technique within the training set. To ensure the robustness of development validation, the process was repeated 20 times, resulting in 100 iterations of model training and validation. The performance of the models was then evaluated based on the average metrics from these iterations, providing an unbiased estimation of model performance.

For test validation, the best model and hyperparameters derived from the development validation were retrained on the entire training data set and then tested on the independent test set to assess their generalizability and performance in a test cohort. This comprehensive validation approach aimed to bolster the robustness of the predictive models before their potential clinical application.

In addition to LR, 5 other ML algorithms were used to construct models using the same methodology as described previously. These algorithms included support vector machine, decision tree, random forest, gradient boosting, and extreme gradient boosting. Each algorithm was systematically evaluated to assess its performance in predicting BOO, ensuring a comprehensive analysis of different modeling approaches.

Outcome Measures and Statistical Analysis

In assessing the performance of our predictive models, we used several outcome measures to evaluate their effectiveness. Discrimination (the model's ability to distinguish between the presence and absence of an outcome) was measured by the area under the receiver operating characteristic curve (AUC). The AUC provides a single measure summarizing the model's accuracy across all classification thresholds. The confusion matrix provided crucial metrics influenced by the chosen cut-off value, such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the F_1-score. The optimal threshold was determined using the

XSL•FO
RenderX

Youden index. Model agreement was assessed using calibration plots, which visually compare predicted probabilities with observed outcomes, thus ensuring that the model's predictions align closely with actual risk. The Hosmer-Lemeshow test was also used to evaluate the goodness-of-fit of the models [18]. A *P* value of less than .05 was considered statistically significant.

Numerical variables with a normal distribution are reported as means and SDs and compared using an independent *t* test (2-tailed). All categorical demographic variables were compared using the chi-square test. LR models and nomograms were developed using R (version 4.1.1; R Core Team) with the "glmnet" and "rms" packages for model fitting and nomogram visualization [19]. Python's "scikit-learn" library (version 0.24.2) was used for supplementary ML and validation tasks [20]. In addition, SPSS Statistics (version 25; IBM Corp) was used for statistical testing [21].

Ethical Considerations

The study protocol was approved by the Research Ethics Committee of Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, on June 15, 2022 (IRB111-132-B).

Results

Patient Cohort Characteristics and LUTS Diagnoses

This study analyzed a cohort of 307 male patients presenting with medication-refractory nonneurogenic LUTS suspected of BOO, all of whom underwent VUDS. The patient group was further subdivided into a training set of 70% (215) and a test set of 30% (92). The average age of the entire cohort was 67.8 (SD 9.7, range 32.2-91.5) years, which did not differ significantly between the training and test groups (*P*=.93). Clinical parameters such as IPSS-voiding, IPSS-storage scores, IPSS-voiding to storage score ratio, Qmax, Vvol, postvoid residual, TPV, transitional zone index, IPP, and PUA all showed no statistically significant difference when the training set was compared with the test set (Table 1).

XSL•FO
RenderX

Table 1. Characteristics, distribution of video-urodynamic studies diagnoses, and bladder outlet obstruction binary classifications in the patient cohort, with training or test set split.

Cohort	Entire	Training	Test	Train vs test *P* value
Overall, n (%)	307 (100)	215 (70)	92 (30)	__[a]
Characteristics, mean (SD)				
Age (years)	67.8 (9.7), range (32.2-91.5)	67.8 (9.8), range (38.2-91.5)	67.9 (9.5), range (32.2-86.9)	.93[b]
IPSS[c]-voiding	9.5 (5.9)	9.5 (5.7)	9.7 (6.3)	.78[b]
IPSS-storage	8.7 (4.1)	8.6 (4)	9 (4.2)	.45[b]
IPSS-V/S[d] ratio	1.5 (1.6)	1.5 (1.6)	1.5 (1.7)	.99[b]
Qmax[e] (mL/s)	10.2 (6.5)	10.6 (7)	9.4 (5.1)	.11[b]
Voided volume (mL)	196.8 (124.1)	200.1 (124.7)	189.2 (123)	.48[b]
Postvoid residual (mL)	44.1 (83.9)	48.7 (96.3)	33.5 (41.2)	.15[b]
Total prostate volume (mL)	38.1 (19.5)	39.2 (20.9)	35.4 (15.7)	.12[b]
Transitional zone index	0.4 (0.16)	0.4 (0.16)	0.42 (0.14)	.39[b]
IPP[f] (cm)	0.3 (0.6)	0.3 (0.6)	0.3 (0.5)	.39[b]
Prostatic urethral angle (º)	26.3 (18.7)	25.9 (18.8)	27.4 (18.3)	.52[b]
VUDS[g] diagnosis, n (interclass%)				0.50[h]
Obstruction-predominant				
Primary bladder neck obstruction	120 (39.1)	84 (39.1)	36 (39.1)	—
Benign prostatic obstruction	87 (28.3)	61 (28.4)	26 (28.3)	—
Dysfunctional voiding	45 (14.7)	32 (14.9)	13 (14.1)	—
Urethral stricture	0 (0)	0 (0)	0 (0)	—
Bladder dysfunction				
Detrusor underactivity	13 (4.2)	1 (1.1)	12 (5.6)	—
Detrusor overactivity	28 (9.1)	10 (10.9)	18 (8.4)	—
Bladder oversensitivity	6 (2)	2 (2.2)	4 (1.9)	—
Generally normal	8 (2.6)	4 (1.9)	4 (4.3)	—
BOO[i] binary classifications, n (interclass%)				
ICS-BOO[j]				0.872[h]
Yes	82 (26.7)	58 (27)	24 (26.1)	—
No	225 (73.3)	157 (73)	68 (73.9)	—
VBOO[k]				0.866[h]
Yes	252 (82.1)	177 (82.3)	75 (81.5)	—
No	55 (17.9)	38 (17.7)	17 (18.5)	—

[a]Not applicable.

[b]Comparison between training or test groups using independent *t* test.

[c]IPSS: International Prostate Symptom Score.

[d]IPSS-V/S ratio: International Prostate Symptom Score voiding to storage score ratio.

[e]Qmax: maximum urinary flow rate.

[f]IPP: intravesical prostatic protrusion.

[g]VUDS: video-urodynamic studies.

XSL•FO
RenderX

[h]Comparison between training or test groups using chi-square test.

[i]BOO: bladder outlet obstruction.

[j]ICS-BOO: International Continence Society–defined bladder outlet obstruction.

[k]VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction.

In the breakdown of VUDS diagnoses within the cohort, obstruction-predominant conditions were observed, with PBNO accounting for 39.1% (120/307), BPO for 28.3% (87/307), and DV for 14.7% (45/307) of the cases. Notably, there were no instances of US reported. Regarding bladder dysfunction, detrusor underactivity (DU) was present in 4.2% (13/307) of the patients, detrusor overactivity (DO) in 9.1% (28/307), and bladder oversensitivity in 2% (6/307), with 2.6% (8/307) of the cohort deemed generally normal. The distribution of these diagnoses was consistent across both training and test sets, as indicated by the nonsignificant chi-square *P* values, demonstrating a homogeneous division of conditions between the 2 groups (Table 1).

For the ICS-BOO, only 26.7% (82/307) of the entire cohort was classified as having BOO. However, the prevalence of VBOO was much higher, with 82.1% (225/307) of patients being identified as having BOO (McNemar test $P<.001$). This disparity in BOO prevalence between ICS-BOO and VBOO highlights the different diagnostic criteria and approaches between PFS and VUDS. VUDS excels in identifying the causes of nonneurogenic medication-refractory male BOO more precisely than PFS. The train or test division yielded similar proportions of ICS-BOO and VBOO classifications, supporting the reliability of the training set for model development and the test set for subsequent validation (Table 1).

Performance Comparison Across 6 Machine Learning Algorithms

Both models 1 and 2 used 6 different algorithms for their predictive analyses, with comprehensive results presented in Table S1 in Multimedia Appendix 1. LR consistently achieved the highest AUC values for both models across the training and test sets, leading to its selection for developing both the nomogram and the web-based predictive tool.

Model 1: ICS-BOO Prediction Nomogram Development and Performance

For model 1, a multivariate LR was used to construct a nomogram that predicts the probability of ICS-BOO. In developing this model, we included key clinical variables, with the final model retaining Qmax, Vvol, TPV, and PUA as contributory predictors. The odds ratios (ORs) for these variables were calculated, indicating that a decrease in Qmax and Vvol and an increase in TPV and PUA were associated with the presence of ICS-BOO. Specifically, the model detailed a statistically significant OR of 1.022 for each degree increase in PUA ($P=.03$) and 1.017 for each milliliter increase in TPV ($P=.04$), suggesting a strong influence on ICS-BOO risk (Table 2). The predictive capability of the model was further substantiated by a satisfactory Hosmer-Lemeshow test ($P=.36$) and by the calibration plot (Figure S1 in Multimedia Appendix 1), indicating a robust fit with the observed outcomes.

XSL•FO
RenderX

Table 2. ICS-BOO and VBOO prediction models using multivariate logistic regression.

Variables	Model 1[a]: ICS-BOO[b] prediction		Model 2[c]: VBOO[d] prediction	
	OR[e] (95% CI)	*P* value	OR (95% CI)	*P* value
Age	—[f]	—	—	—
IPSS[g]-voiding	—	—	0.934 (0.871-0.999)	.05
IPSS-storage	—	—	—	—
IPSS-V/S[h] ratio	—	—	—	—
Qmax[i] (mL/s)	0.927 (0.858-1.002)	.06	0.908 (0.849-0.966)	.003
Voided volume (mL)	0.997 (0.993-1.001)	.10	1.004 (1-1.008)	.04
PVR[j] (mL)	—	—	—	—
Total prostate volume (mL)	1.017 (1.001-1.034)	.04	—	—
Transitional zone index	—	—	—	—
IPP[k] (cm)	—	—	2.253 (0.874-8.132)	.14
Prostatic urethral angle (°)	1.022 (1.002-1.042)	.03	1.017 (0.994-1.041)	.15

[a]Model-1: variable selection based on Akaike's information criterion (223.45). Hosmer-Lemeshow goodness-of-fit test, *P*=.36.

[b]ICS-BOO: International Continence Society–defined bladder outlet obstruction.

[c]Model-2: variable selection based on Akaike's information criterion (188.75). Hosmer-Lemeshow goodness-of-fit test, *P*=.54.

[d]VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction.

[e]OR: odds ratio.

[f]The variable was not included in the model.

[g]IPSS: International Prostate Symptom Score.

[h]IPSS-V/S ratio: International Prostate Symptom Score voiding to storage score ratio.

[i]Qmax: maximum urinary flow rate.

[j]PVR: postvoid residual.

[k]IPP: intravesical prostatic protrusion.

The nomogram derived from model 1 was visually represented, plotting points for each variable that corresponded to their measured values. All variables, such as Qmax, Vvol, TPV, and PUA, were translated into points, which were then summed to yield a total score. This score was mapped to the corresponding probability of ICS-BOO, with the threshold for predicting a positive ICS-BOO outcome set at a probability of 0.31, as determined by the Youden index (Figure 2A).

XSL•FO
RenderX

Figure 2. Nomogram for International Continence Society–defined bladder outlet obstruction prediction (model 1) with validation. (A) Nomogram for International Continence Society–defined bladder outlet obstruction prediction, (B) mean receiver operating characteristic curve of development validation by 20 iterations of 5-fold cross-validation on the training set, and (C) receiver operating characteristic curve of test validation on the test set. AUC: area under the receiver operating characteristic curve; CV: cross-validation; ICS-BOO: International Continence Society–defined bladder outlet obstruction; PU angle: prostatic urethral angle; Qmax: maximum urinary flow rate; ROC: receiver operating characteristic; TPV: total prostate volume; Vvol: voided volume.

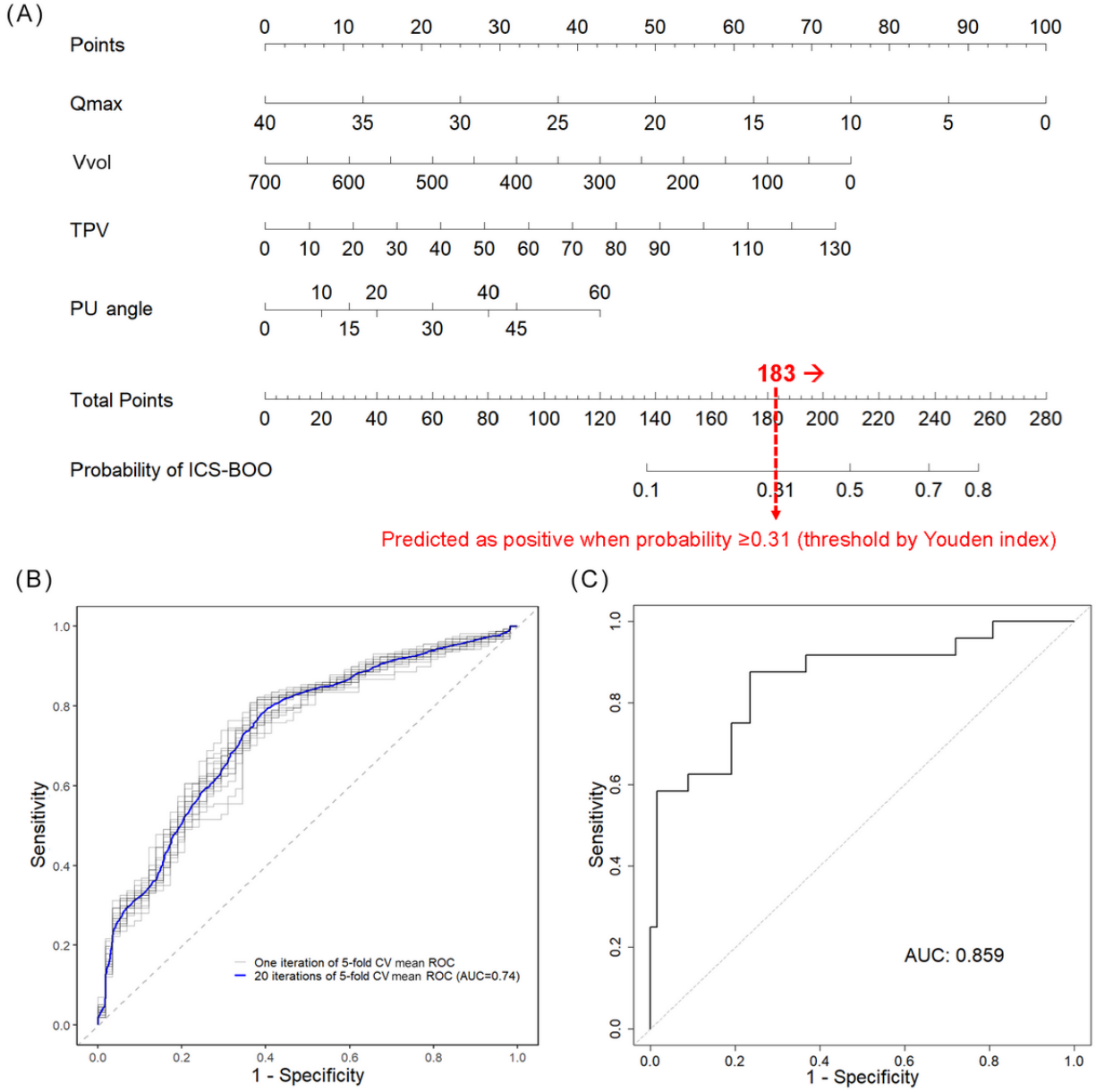

Model 1, designed for ICS-BOO prediction, demonstrated solid performance across various metrics (Table 3 and Figure 2). During the development phase, the model achieved an AUC of 0.76 on the training data set, indicating a good ability to differentiate between the presence and absence of ICS-BOO. The development validation, which involved 20 repetitions of 5-fold cross-validation on the training set, reported an AUC of 0.74 (SD 0.09). The accuracy of the model was consistent as well, with a development validation accuracy of 0.76 (SD 0.04), slightly higher than the developmental phase accuracy of 0.75.

XSL•FO
RenderX

Table 3. Performance metrics of ICS-BOO and VBOO prediction models.

Performance metrics	Mode 1[a]: ICS-BOO[b] prediction			Mode 2[c]: VBOO[d] prediction		
	Development	Development validation	Test validation	Development	Development validation	Test validation
Data set cohort	Training	Training set 20x 5-fold CV[e], mean (SD)	Test	Training	Training set 20x 5-fold CV, mean (SD)	Test
AUC[f]	0.76	0.74 (0.09)	0.86	0.75	0.71 (0.06)	0.72
Accuracy	0.75	0.76 (0.04)	0.77	0.80	0.77 (0.06)	0.76
Sensitivity	0.67	__[g]	0.75	0.85	—	0.87
Specificity	0.78	—	0.78	0.58	—	0.29
PPV[h]	0.53	—	0.55	0.90	—	0.84
NPV[i]	0.87	—	0.90	0.46	—	0.33
F_1-score	0.59	—	0.63	0.87	—	0.85

[a]Binary classification best threshold: Youden index, $P \geq .31$.

[b]ICSS-BOO: International Continence Society–defined bladder outlet obstruction.

[c]Binary classification best threshold: Youden index, $P \geq .73$.

[d]VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction.

[e]CV: cross-validation, presented as mean (SD).

[f]AUC: area under the receiver operating characteristic curve.

[g]Not available.

[h]PPV: positive predictive value.

[i]NPV: negative predictive value.

When applied to the external test set, model 1's performance improved, with the AUC increasing to 0.86. The test validation also revealed a sensitivity of 0.75, a specificity of 0.78, a PPV of 0.55, and a notably high NPV of 0.90, along with an F_1-score of 0.63, balancing PPV and sensitivity. To summarize, the robustness of the ICS-BOO prediction model was confirmed across both development and test validations.

Model 2: Development and Performance of the VBOO Prediction Nomogram

For model 2, a multivariate LR was used to develop a nomogram that predicts the probability of VBOO. In the model selection process based on AIC, 5 variables, including IPSS-voiding (OR 0.934), Qmax (OR 0.908), Vvol (OR 1.004), IPP (OR 2.253), and PUA (OR 1.017) were retained in the final model as contributory predictors. Specifically, the model detailed statistically significant OR in Qmax (P=.003), Vvol (P=.04), and IPSS-voiding (P=.05), suggesting strong associations with VBOO risk (Table 2). The model demonstrated a satisfactory goodness of fit, as indicated by a Hosmer-Lemeshow test P value of .54 and by the calibration plot (Figure S2 in Multimedia Appendix 1).

The nomogram for model 2 translates these variables into a points system, allowing for an intuitive prediction of VBOO risk. For instance, a decrease in IPSS-voiding or Qmax scores increases the risk points on the nomogram, directly correlating with a higher probability of VBOO. The nomogram defines a positive prediction for VBOO when the probability is ≥0.73, as determined by the Youden index. This threshold was chosen to ensure the nomogram effectively balances sensitivity and specificity (Figure 3A).

XSL•FO
RenderX

Figure 3. Nomogram for video-urodynamic studies–diagnosed bladder outlet obstruction prediction (model 2) with validation. (A) Nomogram for video-urodynamic studies–diagnosed bladder outlet obstruction prediction, (B) mean receiver operating characteristic curve of development validation by 20 iterations of 5-fold cross-validation on the training set, and (C) receiver operating characteristic curve of test validation on the test set. AUC: area under the receiver operating characteristic curve; CV: cross-validation; IPP: intravesical prostatic protrusion; IPSS: International Prostate Symptom Score; PU angle: prostatic urethral angle; Qmax: maximum urinary flow rate; ROC: receiver operating characteristic; VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction; Vvol: voided volume.

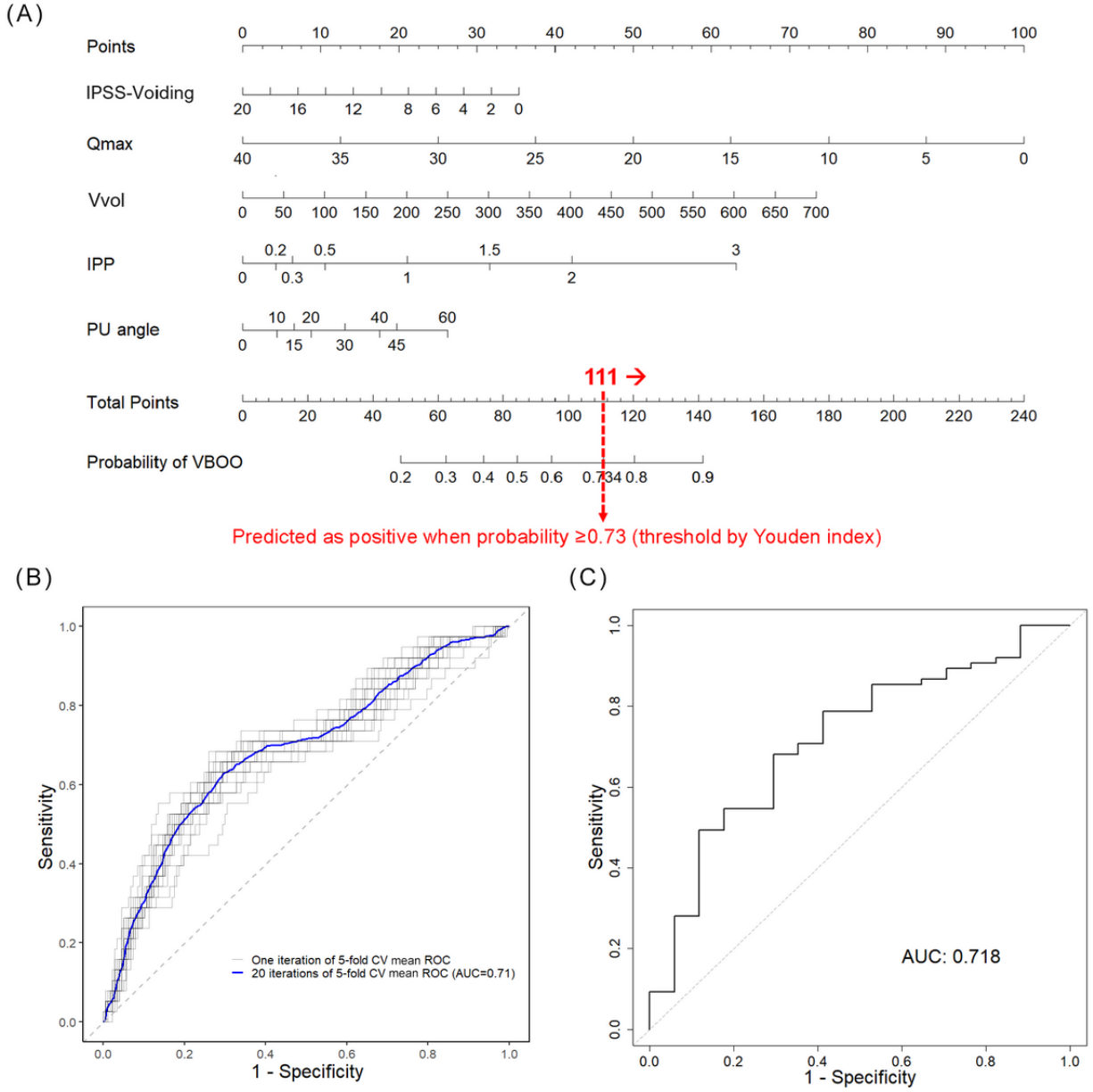

Model 2, aimed at predicting VBOO, exhibited robust performance in both the development and validation phases (Table 3 and Figure 3). During the initial development of the training data set, the model achieved an AUC of 0.75, indicating respectable discriminative ability. The accuracy was notably high at 0.80, with an outstanding sensitivity of 0.85, suggesting the model's effectiveness in identifying patients with VBOO. The specificity, however, was moderate at 0.58, which is reflected in the high PPV of 0.90 and a lower NPV of 0.46. The F_1-score, which harmonizes the PPV and sensitivity, stood at 0.87, highlighting the model's balanced performance.

In the development validation, model 2 maintained an AUC of 0.71 (SD 0.06) and an accuracy of mean 0.77 (SD 0.06), confirming the model's consistency. The test validation on the test set further validated the model, with an AUC of 0.72. The sensitivity was notably high at 0.87 on the test set, surpassing the training performance. Despite this, the external PPV remained high at 0.84, and the NPV improved to 0.33, with an F_1-score comparable to the development phase at 0.85.

Dual Model Predictions Aiding Clinical Decision-Making

The preceding results demonstrate that VUDS can more accurately differentiate between causes of nonneurogenic male

XSL•FO
RenderX

BOO compared with PFS. Although an initial attempt was made to develop a single AI model using noninvasive parameters to predict the exact VUDS diagnosis (a 5-class classification), it resulted in poor accuracy (Multimedia Appendix 1). However, 2 ML models designed to predict ICS-BOO and VBOO in patients with medication-refractory, nonneurogenic male LUTS have shown commendable discrimination performance. When the predictions from both models are applied to the same patient, their combined insights can significantly enhance clinical decision-making and simplify the diagnostic pathway. The algorithm detailing this combined predictive approach is depicted in Figure 4.

Figure 4. Algorithmic flowchart of dual artificial intelligence models predicting bladder outlet obstruction to aid clinical decision-making for nonneurogenic male lower urinary tract symptoms. From the clinical decision-making perspective, there exists a dilemma for men with medication-refractory nonneurogenic lower urinary tract symptoms who are suspected of having bladder outlet obstruction. The following question arises: which diagnostic urodynamic study is most appropriate for the patient? (1) Should one take the first pathway and perform pressure-flow studies only to potentially obtain an accurate diagnosis? (2) Or should one follow the second pathway, starting with pressure-flow studies and proceeding to video-urodynamic studies if a definitive diagnosis is not achieved? (3) Alternatively, is the third pathway preferable, where one directly conducts video-urodynamic studies for a detailed diagnosis? By the dual-model prediction approach, if both models positively predict bladder outlet obstruction, suggesting it is caused by either medication-refractory primary bladder neck obstruction or benign prostatic obstruction, surgical intervention may be considered; thus, video-urodynamic studies or pressure-flow studies might be unnecessary. Conversely, when prediction for International Continence Society–defined bladder outlet obstruction is negative but video-urodynamic studies–diagnosed bladder outlet obstruction is positive, indicating varied etiology, video-urodynamic studies are advised over pressure-flow studies for precise diagnosis and guiding subsequent therapy. AI: artificial intelligence; BOO: bladder outlet obstruction; ICS-BOO: International Continence Society–defined bladder outlet obstruction; IPSS: International Prostate Symptom Score; LUTS: lower urinary tract symptoms; PFS: pressure-flow studies; TRUS-P: transrectal ultrasound of the prostate; UDS: urodynamic study; UFM: uroflowmetry; VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction; VUDS: video-urodynamic studies.

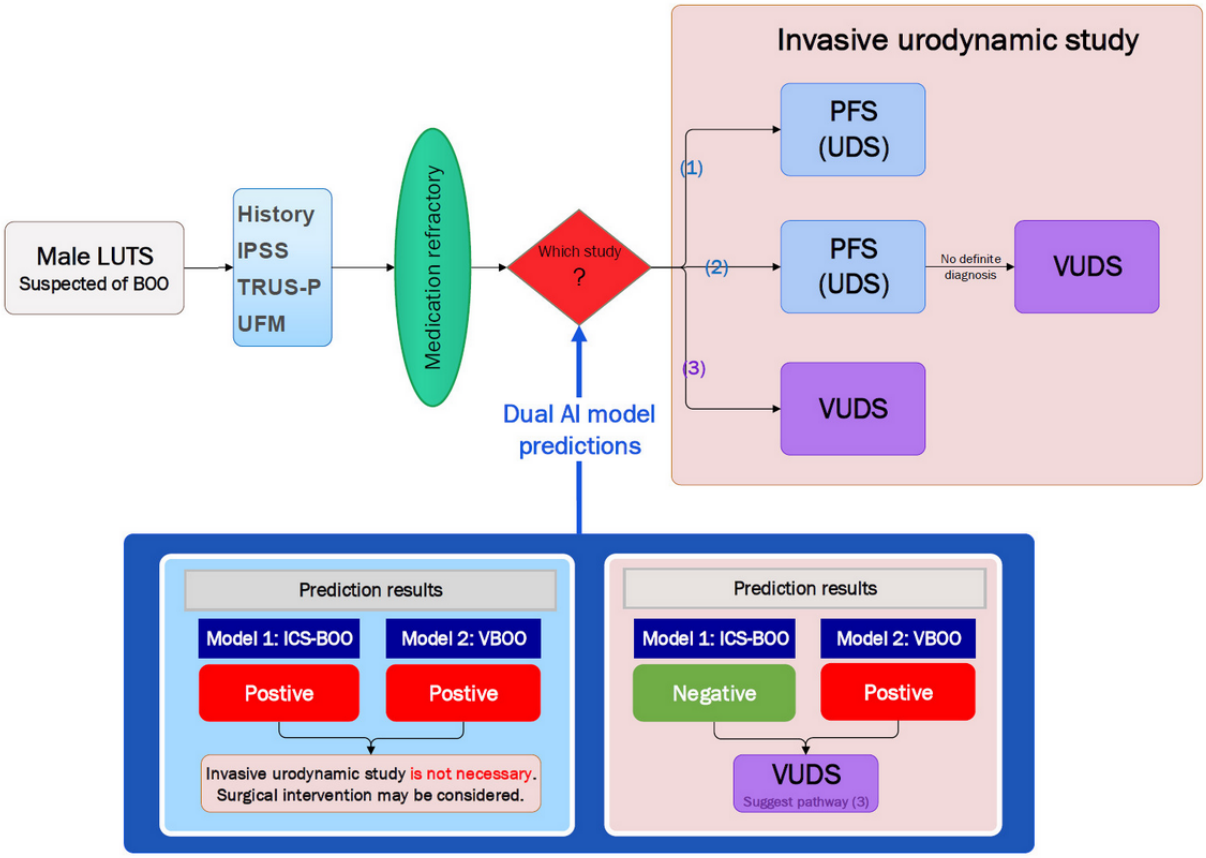

In the validation performed on the test set (n=92), diagnoses were concurrently made for each patient using the ICS-BOO and VBOO criteria, resulting in an actual confusion matrix as illustrated in Figure S3 in Multimedia Appendix 1. In addition, using the dual model prediction approach produced a separate dual predictions confusion matrix. The comparison revealed a similar distribution of BOO across the 2 confusion matrices, suggesting a substantial agreement between the actual BOO diagnoses and the predictions from both the ICS-BOO and VBOO models.

In our analysis, both the ICS-BOO and VBOO models predicted positively in 30 out of 92 patients, which represents 32.6% of the test cohort. Notably, for these 30 patients, the BOO was determined to be exclusively caused by either BPO or PBNO, accounting for 100% of cases within this subgroup, with no cases diagnosed as DV. For this subgroup of patients with medication-refractory BPO or PBNO, similar surgical interventions such as transurethral resection of the prostate (TURP) or transurethral incision of the prostate (TUIP) may be suggested, thus guiding the subsequent surgical treatment. Consequently, VUDS could potentially be avoided for these

XSL•FO
RenderX

patients, aiding in the diagnosis of male BOO and possibly reducing the necessity of such invasive examinations for approximately one-third of the test cohort.

Conversely, in the test cohort of 92 patients, there were 51.1% (n=47) cases where the ICS-BOO model predicted negatively while the VBOO model predicted positively. Within this particular subgroup, the actual diagnoses were divergent—PBNO was identified in 40% (19/92) cases, BPO in 19% (9/92) cases, DV in 15% (7/92) cases, and non-BOO in 26% (12/92) cases. Therefore, for this subgroup, the use of VUDS rather than PFS is strongly recommended for differential diagnosis, which guides varied treatment, thereby precisely targeting 51.1% (47/92) of patients for VUDS (Figure 4).

Web-Based Dual-Model Prediction Tool

We developed a user-friendly web-based online prediction tool that simplifies the dual-model prediction process for BOO. This tool unifies the predictors of both models into 6 parameters, listed as IPSS-voiding score, Qmax, Vvol, TPV, PUA, and IPP, enabling clinicians to input patient-specific data and instantly receive probabilities for both ICS-BOO and VBOO with corresponding decision-making recommendations into one output. This innovation seamlessly incorporates AI-driven predictions into the clinical workflow, helping to discern which patients may require a surgical intervention or a VUDS, thereby facilitating and optimizing the precision of patient care (Figure 5).

Figure 5. Two screenshots of outputs from the web-based dual-model prediction tool. Based on 6 noninvasive clinical parameters from 2 different patients, the prediction results are as follows: (A) demonstrates a scenario where both International Continence Society–defined bladder outlet obstruction and video-urodynamic studies–diagnosed bladder outlet obstruction models predict a positive outcome for bladder outlet obstruction, with probabilities of 54% and 99%, respectively, indicating a high likelihood of bladder outlet obstruction caused by either benign prostatic obstruction or primary bladder neck obstruction and suggesting that invasive studies like video-urodynamic studies or pressure-flow studies may be unnecessary. (B) illustrates a case where the International Continence Society–defined bladder outlet obstruction model predicts a negative outcome (17% probability) while the video-urodynamic studies–diagnosed bladder outlet obstruction model predicts a positive outcome (88% probability), suggesting a complex bladder outlet obstruction etiology that may not present with high-pressure low-flow dynamics and may benefit from further evaluation with video-urodynamic studies. BOO: bladder outlet obstruction; BPO: benign prostatic obstruction; DV: dysfunctional voiding; ICS-BOO: International Continence Society–defined bladder outlet obstruction; IPSS: International Prostate Symptom Score; PBNO: primary bladder neck obstruction; PFS: pressure-flow studies; TUIP: transurethral incision of the prostate; TURP: transurethral resection of the prostate; VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction; VUDS: video-urodynamic studies.

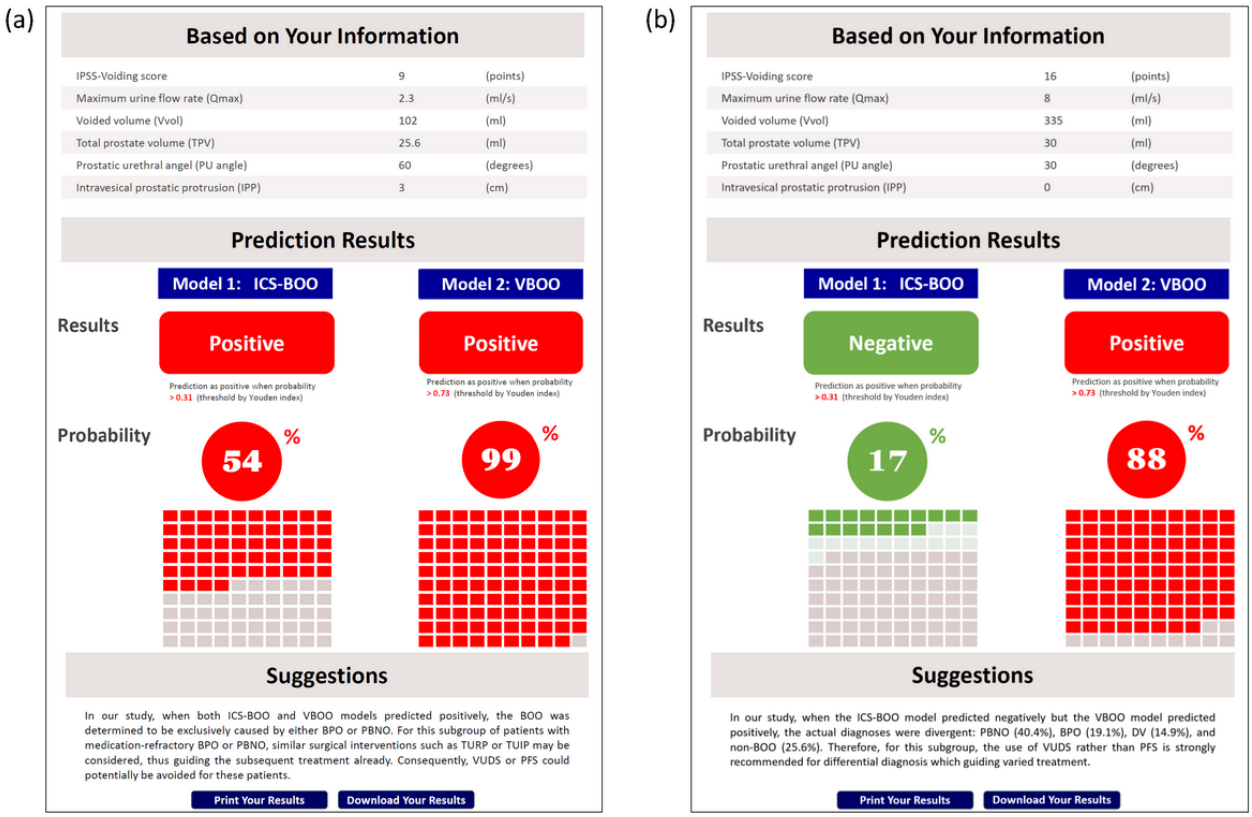

Figure 5 demonstrates 2 screenshots of outputs from the web-based dual-model prediction tool. Using 6 noninvasive clinical parameters from 2 different patients, the tool generates the following predictive results: in Figure 5A, both ICS-BOO and VBOO models make a positive BOO prediction, with probabilities of 54% and 99%, respectively, indicating a high likelihood of BOO caused by either BPO or PBNO and suggesting that surgical intervention may be considered and invasive studies like VUDS or PFS may be unnecessary. Figure 5B illustrates a case where the ICS-BOO model predicts a negative outcome (17% probability) while the VBOO model predicts a positive outcome (88% probability), suggesting a complex BOO etiology that may not present with classical high-pressure low-flow dynamics and may benefit from further evaluation with VUDS (Multimedia Appendix 1).

XSL•FO
RenderX

Discussion

Uniqueness and Contributions

The uniqueness of this study lies in using VUDS to identify the causes of male BOO and LUTS. In the literature review, most studies only use PFS as the standard method for assessing BOO. Our model 1 also adopts this conventional approach by using a BOOI greater than 40 as the diagnostic standard. In contrast, this study uses the more accurate VUDS as the diagnostic standard for BOO in our model 2, enabling the identification of more patients with potential obstructions. Using these data, 2 predictive models are established. To our knowledge, no other studies have been conducted in this area.

Male BOO is a urodynamic condition in which a bladder outlet cannot open during urination. The underlying pathophysiology could be anatomical (due to BPO or US) or functional (PBNO or DV) obstruction. The results revealed that VUDS excels in identifying the causes of nonneurogenic medication-refractory male BOO more precisely than PFS, as evidenced by the higher prevalence of BOO diagnoses made by VUDS (252/307, 82.1%) compared to those identified by PFS (82/307, 26.7%; $P<.001$). For example, patients exhibiting a nonrelaxing bladder neck during emptying—a type of BOO that could only be diagnosed through VUDS, specifically type 2 PBNO [22]—when combined with low Pdet.Qmax might have a BOOI under 40. This falls short of the ICS-BOO standards, according to PFS, leading to misclassification as non-BOO. Such complicated etiologies of male BOO can only be determined by VUDS but not by PFS alone.

Alternatively, our innovative VBOO model, used before any invasive UDS, can identify potential BOO when predictions are positive. Alongside ICS-BOO predictions, whether positive or negative, it aids in determining the need for surgical intervention or further VUDS evaluation. When both VBOO and ICS-BOO predictions are positive, indicating medication-refractory PBNO or BPO, similar surgical interventions such as TURP or TUIP may be considered, thus guiding the subsequent treatment. Therefore, an invasive VUDS or PFS examination might be unnecessary.

Conversely, a positive VBOO prediction with a negative ICS-BOO suggests a complex BOO etiology not typically presenting with high-pressure, low-flow dynamics, possibly indicating a BOOI <40 or within the BOOI equivocal zone, necessitating further VUDS evaluation for these patients. Furthermore, given the high sensitivity (0.86) of the VBOO model, a negative VBOO prediction from such a sensitive test can reliably exclude the BOO disease (as non-BOO), a principle commonly referred to as SnNOUT [23].

The ICS-BOO standard, defined by a BOOI >40, represents the classic high-pressure low-flow pattern but may not capture all obstructive disorders. For example, BOOI <20 indicates nonobstruction, while values between 20 and 40 fall into an "equivocal" zone by the ICS BOOI standard, where PFS alone may not definitively classify a patient's condition. This is depicted in the clinical decision-making "pathway (2)" in Figure 4, where the patients with medication-refractory male LUTS who do not receive a definitive diagnosis from PFS may require further VUDS examination—a situation that occurs in real-world practice and can lead to multiple invasive examinations for the patient. Hence, our VBOO prediction model serves as a clinical decision support tool, particularly beneficial for physicians or hospitals lacking VUDS facilities, aiding treatment decisions based on noninvasive predictions.

Our dual nomograms were developed for and applied to a distinct patient cohort: nonneurogenic patients with medication-refractory male LUTS suspected of BOO, differentiating them from men with treatment-naive or typical LUTS. This specific focus aligns our application more closely with real clinical decision-making scenarios. Typically, PFS or VUDS are not routinely performed for all men with common LUTS but are considered when initial diagnoses and medication trials are ineffective. This approach suggests that the pathophysiology in our patient group might be more complex than in common LUTS cases. A considerable portion might have had other functional BOO pathologies such as PBNO or DV, or non-BOO causes such as DO, DU, or BO. It was evidenced by the fact that our cohort showed a relatively low incidence of ICS-BOO (82/307, 26.7%) compared to the prevalence in common patients with male LUTS, yet a high prevalence of VBOO (252/307, 82.1%). Therefore, this study, which used a relatively challenging-to-diagnose cohort to validate the performance of our ICS-BOO model, should not be directly compared in terms of performance with other nomograms validated using patients with common LUTS [16,24].

Noninvasive Predictors

The ICS-BOO model incorporates Qmax, Vvol, PUA, and TPV as predictors, while the VBOO model includes Qmax, Vvol, PUA, IPSS-voiding, and IPP. Thus, these 2 models share 3 common predictors: Qmax, Vvol, and PUA. Qmax was significantly associated with the presence of ICS-BOO and VBOO, aligning with earlier research on male treatment-naïve or common LUTS [11]. However, the diagnostic reliability of Qmax alone is limited, as conditions like DU may also result in a lower Qmax, thus complicating the prediction of BOO based solely on this parameter [25]. Similarly, the PUA was incorporated into both ICS-BOO and VBOO models showing a positive correlation between the angle degree and the likelihood of BOO, consistent with findings from previous studies [26-28]. Ku et al [28] revealed that patients with higher PUA had a higher BOOI than those with lower PUA, suggesting PUA's utility in PFS-obstruction prediction. Moreover, PUA also contributes to our VBOO prediction, thus indicating that PUA may play an important role associated with the varied etiologies of BOO. Regarding TPV, which is significantly associated with the ICS-BOO model but not included as a predictor in the VBOO model, it suggests that a high TPV is correlated with a high BOOI, aiding in inferring that the etiology could be BPO [29]. Conversely, TPV's lack of contribution to VBOO prediction reasonably implies that it is not useful in detecting functional BOO such as PBNO or DV. Finally, the inclusion of the IPSS-voiding score, rather than the IPSS-storage score, in the VBOO model's prediction aligns well with domain knowledge [30,31].

XSL•FO
RenderX

Clinical Significance

In summary, our dual-model AI application is a pioneering tool in functional urology. The strengths and medical significance of the application are outlined as follows:

1. Aiding clinical decision-making and guiding treatment: our dual-model approach uses noninvasive clinical parameters to predict anatomical or functional BOO, providing actionable recommendations and reducing the need for invasive procedures like VUDS or PFS. Specifically:

- When both ICS-BOO and VBOO predictions are positive, it suggests that surgical interventions such as TURP or TUIP may be considered without further invasive studies.
- A negative ICS-BOO but positive VBOO prediction suggests using VUDS over PFS for differential diagnosis, guiding varied treatment approaches.
- A negative prediction in both models reliably excludes BOO, prompting the evaluation of other non-BOO conditions such as DU, DO, or BO, guiding appropriate treatment.

2. Reduction of invasive procedures: the application potentially reduces the frequency of unnecessary invasive tests by about one-third, enhancing patient comfort and reducing complications and health care costs.

3. Streamlining clinical workflows: the web-based AI tool integrates seamlessly into clinical workflows, enabling quick, data-driven decision-making and personalized patient management.

4. Expanding access to advanced diagnostics: in environments lacking VUDS due to cost or logistical constraints, our tool offers a practical alternative, democratizing advanced diagnostics and potentially improving outcomes.

5. Optimizing treatment for challenging cases: our application aids in managing challenging clinical scenarios, such as those involving patients with medication-refractory male LUTS who have experienced primary treatment failure. A lack of accurate diagnosis in such complex cases can lead to inappropriate treatment and poor outcomes. For example, performing TURP surgery on a patient suspected of BOO could fail to improve symptoms if the actual diagnosis is DV rather than BPO. Our application helps identify which patients might benefit directly from surgery without the need for invasive VUDS or PFS, thus optimizing treatment strategies and improving patient outcomes.

6. Innovative approach over existing tools: unlike traditional models that depend solely on PFS, this study uses VUDS data, addressing the limitations of existing models in handling equivocal BOOI zones. Our dual-model prediction approach represents an innovative advancement in this field. The combination of ICS-BOO and VBOO models sufficiently aids clinical decision-making and guides treatment, surpassing the capabilities of traditional single-prediction models based solely on ICS-BOO criteria.

7. Transparent and interpretable AI models: after fine-tuning the hyperparameters of 6 ML algorithms, LR consistently achieved the highest AUC values for both models. The LR model's selected predictors and their corresponding ORs offer high clinical interpretability, which enhances trust and usability among clinicians.

8. Flexible use from single to dual models, traditional to modern approaches: our tool offers flexibility in clinical application, enabling clinicians to adopt only a single model or a dual-model approach based on their preference. In addition, the paper-based nomogram provides a traditional, user-friendly option that is adaptable for various clinical scenarios.

Limitations

This study has several limitations. First, its retrospective design introduces a potential for selection bias due to the exclusion of patients with incomplete data. Second, using data solely from a single medical institution might limit the generalizability of the findings. To enhance the robustness of our predictive model, future validation with multicenter data is essential. Furthermore, the models developed specifically for patients with medication-refractory, nonneurogenic LUTS might not be suitable for those with initial diagnoses or common LUTS. In addition, although a range of parameters were considered as predictive factors, not including concurrent comorbidities may have limited the control over potential confounding factors.

Conclusions

Nonneurogenic male BOO caused by BPO, PBNO, DV, or US cannot be precisely differentiated solely through PFS or a single AI prediction model. However, 2 ML models predicting ICS-BOO and VBOO among patients with medication-refractory, nonneurogenic male LUTS using noninvasive clinical parameters demonstrate commendable discrimination performance. By our innovative dual-model prediction approach, if both models positively predict BOO, suggesting it is caused by either medication-refractory PBNO or BPO, surgical intervention may be considered, thus VUDS might be unnecessary for about one-third of patients. Conversely, when prediction for ICS-BOO is negative but VBOO is positive, indicating varied etiology, VUDS is advised over PFS for precise diagnosis and guiding subsequent therapy, accurately identifying 51.1% (47/92) of patients for VUDS.

Acknowledgments

We would like to acknowledge the efforts of all colleagues who contributed to data collection.

Data Availability

The data sets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

CYT contributed to conception and design, development of models, analysis and interpretation of data, drafting of the manuscript. JHT contributed to acquisition of data, development of models, statistical analysis, and visualization of data. CCL contributed to implementation of models to web tool. HCK contributed to acquisition of data, interpretation of VUDS, critical revision of the manuscript, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material.
[PDF File (Adobe PDF File), 475 KB - jmir_v26i1e58599_app1.pdf]

Multimedia Appendix 2
Data S1: Dataset for validation of ICS-BOO and VBOO models.
[XLSX File (Microsoft Excel File), 18 KB - jmir_v26i1e58599_app2.xlsx]

References

1. Barry MJ, Fowler FJ, O'leary MP, Bruskewitz RC, Holtgrewe HL, Mebust WK, Measurement Committee of the American Urological Association. The American urological association symptom index for benign prostatic hyperplasia. J Urol 2017;197(2S):S189-S197. [doi: 10.1016/j.juro.2016.10.071] [Medline: 28012747]
2. Sussman RD, Drain A, Brucker BM. Primary bladder neck obstruction. Rev Urol 2019;21(2-3):53-62 [FREE Full text] [Medline: 31768132]
3. Gratzke C, Bachmann A, Descazeaud A, Drake MJ, Madersbacher S, Mamoulakis C, et al. EAU guidelines on the assessment of non-neurogenic male lower urinary tract symptoms including benign prostatic obstruction. Eur Urol 2015;67(6):1099-1109. [doi: 10.1016/j.eururo.2014.12.038] [Medline: 25613154]
4. Cornu JN, Gacci M, Hashim H. EAU guidelines on non-neurogenic male lower urinary tract symptoms (LUTS), including benign prostatic obstruction (BPO). 2023. URL: https://uroweb.org/guidelines/management-of-non-neurogenic-male-luts [accessed 2023-12-02]
5. Jiang YH, Wang CC, Kuo HC. Videourodynamic findings of lower urinary tract dysfunctions in men with persistent storage lower urinary tract symptoms after medical treatment. PLoS One 2018;13(2):e0190704 [FREE Full text] [doi: 10.1371/journal.pone.0190704] [Medline: 29462145]
6. Jiang YH, Chen SF, Kuo HC. Role of videourodynamic study in precision diagnosis and treatment for lower urinary tract dysfunction. Ci Ji Yi Xue Za Zhi 2019;32(2):121-130 [FREE Full text] [doi: 10.4103/tcmj.tcmj_178_19] [Medline: 32269943]
7. D'Ancona C, Haylen B, Oelke M, Abranches-Monteiro L, Arnold E, Goldman H, et al. The International Continence Society (ICS) report on the terminology for adult male lower urinary tract and pelvic floor symptoms and dysfunction. Neurourol Urodyn 2019;38(2):433-477. [doi: 10.1002/nau.23897] [Medline: 30681183]
8. Kuo HC. Videourodynamic analysis of pathophysiology of men with both storage and voiding lower urinary tract symptoms. Urology 2007;70(2):272-276. [doi: 10.1016/j.urology.2007.03.063] [Medline: 17826488]
9. Nitti VW, Lefkowitz G, Ficazzola M, Dixon CM. Lower urinary tract symptoms in young men: videourodynamic findings and correlation with noninvasive measures. J Urol 2002;168(1):135-138. [Medline: 12050507]
10. Padmanabhan P, Nitti VW. Primary bladder neck obstruction in men, women, and children. Curr Urol Rep 2007;8(5):379-384. [doi: 10.1007/s11934-007-0035-3] [Medline: 17880837]
11. Steele GS, Sullivan MP, Sleep DJ, Yalla SV. Combination of symptom score, flow rate and prostate volume for predicting bladder outflow obstruction in men with lower urinary tract symptoms. J Urol 2000;164(2):344-348. [Medline: 10893581]
12. Griffiths CJ, Harding C, Blake C, McIntosh S, Drinnan MJ, Robson WA, et al. A nomogram to classify men with lower urinary tract symptoms using urine flow and noninvasive measurement of bladder pressure. J Urol 2005;174(4 Pt 1):1323-1326. [doi: 10.1097/01.ju.0000173637.07357.9e] [Medline: 16145412]
13. Oelke M, Höfner K, Jonas U, de la Rosette JJ, Ubbink DT, Wijkstra H. Diagnostic accuracy of noninvasive tests to evaluate bladder outlet obstruction in men: detrusor wall thickness, uroflowmetry, postvoid residual urine, and prostate volume. Eur Urol 2007;52(3):827-834. [doi: 10.1016/j.eururo.2006.12.023] [Medline: 17207910]
14. Franco G, De Nunzio C, Leonardo C, Tubaro A, Ciccariello M, De Dominicis C, et al. Ultrasound assessment of intravesical prostatic protrusion and detrusor wall thickness--new standards for noninvasive bladder outlet obstruction diagnosis? J Urol 2010;183(6):2270-2274. [doi: 10.1016/j.juro.2010.02.019] [Medline: 20400138]
15. Yurt M, Süer E, Gülpinar O, Telli O, Arikan N. Diagnosis of bladder outlet obstruction in men with lower urinary tract symptoms: comparison of near infrared spectroscopy algorithm and pressure flow study in a prospective study. Urology 2012;80(1):182-186. [doi: 10.1016/j.urology.2012.03.022] [Medline: 22748876]

XSL•FO
RenderX

16. De Nunzio C, Autorino R, Bachmann A, Briganti A, Carter S, Chun F, EAU Young Academic Urologists BPH Group Arnhem, The Netherlands. The diagnosis of benign prostatic obstruction: development of a clinical nomogram. Neurourol Urodyn 2016;35(2):235-240. [doi: 10.1002/nau.22705] [Medline: 25524269]
17. Lee YJ, Lee JK, Kim JJ, Lee HM, Oh JJ, Lee S, et al. Development and validation of a clinical nomogram predicting bladder outlet obstruction via routine clinical parameters in men with refractory nonneurogenic lower urinary tract symptoms. Asian J Androl 2019;21(5):486-492 [FREE Full text] [doi: 10.4103/aja.aja_127_18] [Medline: 30829291]
18. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. Hoboken, NJ: John Wiley & Sons; 2013.
19. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing URL: https://www.R-project.org/ [accessed 2023-08-20]
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
21. IBM SPSS Statistics for Windows, Version 25. Released 2017. Armonk, NY: IBM Corp; 2017. URL: https://www.ibm.com/products/spss-statistics [accessed 2023-10-05]
22. Nitti VW. Primary bladder neck obstruction in men and women. Rev Urol 2005;7 Suppl 8(Suppl 8):S12-S17 [FREE Full text] [Medline: 16985885]
23. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ 2004;329(7459):209-213 [FREE Full text] [doi: 10.1136/bmj.329.7459.209] [Medline: 15271832]
24. Kang MY, Ku JH, Oh SJ. Non-invasive parameters predicting bladder outlet obstruction in Korean men with lower urinary tract symptoms. J Korean Med Sci 2010;25(2):272-275 [FREE Full text] [doi: 10.3346/jkms.2010.25.2.272] [Medline: 20119582]
25. Abrams P, Chapple C, Khoury S, Roehrborn C, de la Rosette J, International Scientific Committee. Evaluation and treatment of lower urinary tract symptoms in older men. J Urol 2009;181(4):1779-1787. [doi: 10.1016/j.juro.2008.11.127] [Medline: 19233402]
26. Hou CP, Chen CL, Lin YH, Tsai YL, Chang PL, Juang HH, et al. Prostatic urethral angle might be a predictor of treatment efficacy of α-blockers in men with lower urinary tract symptoms. Drug Des Devel Ther 2014;8:937-943 [FREE Full text] [doi: 10.2147/DDDT.S62428] [Medline: 25075177]
27. Kang DH, Lee JY, Hah YS, Chung DY, Lee DH, Cho KS. Correlation of prostatic urethral angle with the severity of urinary symptom and peak flow rate in men with small prostate volume. PLoS One 2014;9(8):e104395 [FREE Full text] [doi: 10.1371/journal.pone.0104395] [Medline: 25127394]
28. Ku JH, Ko DW, Cho JY, Oh SY. Correlation between prostatic urethral angle and bladder outlet obstruction index in patients with lower urinary tract symptoms. Urology 2010;75(6):1467-1471. [doi: 10.1016/j.urology.2009.08.049] [Medline: 19962734]
29. Rosier PF, de la Rosette JJ. Is there a correlation between prostate size and bladder-outlet obstruction? World J Urol 1995;13(1):9-13. [doi: 10.1007/BF00182658] [Medline: 7773319]
30. Reynard JM, Abrams P. Bladder-outlet obstruction--assessment of symptoms. World J Urol 1995;13(1):3-8. [doi: 10.1007/BF00182657] [Medline: 7539678]
31. Jiang YH, Lin VCH, Liao CH, Kuo HC. International Prostatic Symptom Score-voiding/storage subscore ratio in association with total prostatic volume and maximum flow rate is diagnostic of bladder outlet-related lower urinary tract dysfunction in men with lower urinary tract symptoms. PLoS One 2013;8(3):e59176 [FREE Full text] [doi: 10.1371/journal.pone.0059176] [Medline: 23527124]

Abbreviations

AI: artificial intelligence
AIC: Akaike's information criterion
AUC: area under the receiver operating characteristic curve
BOO: bladder outlet obstruction
BOOI: bladder outlet obstruction index
BPO: benign prostatic obstruction
DO: detrusor overactivity
DU: detrusor underactivity
DV: dysfunctional voiding
ICS: International Continence Society
ICS-BOO: International Continence Society–defined bladder outlet obstruction
IPP: intravesical prostatic protrusion
IPSS: International Prostate Symptom Score
LR: logistic regression
LUTS: lower urinary tract symptoms
ML: machine learning

XSL•FO
RenderX

NPV: negative predictive value
OR: odds ratio
PBNO: primary bladder neck obstruction
Pdet.Qmax: maximum detrusor pressure at Qmax
PFS: pressure-flow studies
PPV: positive predictive value
PUA: prostatic urethral angle
Qmax: maximum urinary flow rate
TPV: total prostate volume
TUIP: transurethral incision of the prostate
TURP: transurethral resection of the prostate
UDS: urodynamic study
US: urethral stricture
VBOO: video-urodynamic studies–diagnosed bladder outlet obstruction
VUDS: video-urodynamic studies
Vvol: voided volume

Edited by G Eysenbach, S Ma; submitted 20.03.24; peer-reviewed by HC Yang, LP Hung, WH Liao, SJ Wu, CH Wu; comments to author 16.04.24; revised version received 22.04.24; accepted 14.06.24; published 23.07.24.

Please cite as:
Tsai CY, Tian JH, Lee CC, Kuo HC
Building Dual AI Models and Nomograms Using Noninvasive Parameters for Aiding Male Bladder Outlet Obstruction Diagnosis and Minimizing the Need for Invasive Video-Urodynamic Studies: Development and Validation Study
J Med Internet Res 2024;26:e58599
URL: https://www.jmir.org/2024/1/e58599
doi:10.2196/58599
PMID:

©Chung-You Tsai, Jing-Hui Tian, Chien-Cheng Lee, Hann-Chorng Kuo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automated Behavioral Coding to Enhance the Effectiveness of Motivational Interviewing in a Chat-Based Suicide Prevention Helpline: Secondary Analysis of a Clinical Trial

Mathijs Pellemans[1,2], MSc; Salim Salmi[2,3], MSc; Saskia Mérelle[2,4], PhD; Wilco Janssen[2,4], MSc; Rob van der Mei[1,3], PhD

[1]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[2]113 Suicide Prevention, Amsterdam, Netherlands
[3]Centrum Wiskunde & Informatica, Amsterdam, Netherlands
[4]Department of Psychiatry, Amsterdam UMC, Amsterdam, Netherlands

Corresponding Author:
Mathijs Pellemans, MSc
Department of Mathematics
Vrije Universiteit Amsterdam
De Boelelaan 1111
Amsterdam, 1081 HV
Netherlands
Phone: 31 20 5987700
Email: m.j.pellemans@vu.nl

Abstract

Background: With the rise of computer science and artificial intelligence, analyzing large data sets promises enormous potential in gaining insights for developing and improving evidence-based health interventions. One such intervention is the counseling strategy motivational interviewing (MI), which has been found effective in improving a wide range of health-related behaviors. Despite the simplicity of its principles, MI can be a challenging skill to learn and requires expertise to apply effectively.

Objective: This study aims to investigate the performance of artificial intelligence models in classifying MI behavior and explore the feasibility of using these models in online helplines for mental health as an automated support tool for counselors in clinical practice.

Methods: We used a coded data set of 253 MI counseling chat sessions from the 113 Suicide Prevention helpline. With 23,982 messages coded with the MI Sequential Code for Observing Process Exchanges codebook, we trained and evaluated 4 machine learning models and 1 deep learning model to classify client- and counselor MI behavior based on language use.

Results: The deep learning model BERTje outperformed all machine learning models, accurately predicting counselor behavior (accuracy=0.72, area under the curve [AUC]=0.95, Cohen κ=0.69). It differentiated MI congruent and incongruent counselor behavior (AUC=0.92, κ=0.65) and evocative and nonevocative language (AUC=0.92, κ=0.66). For client behavior, the model achieved an accuracy of 0.70 (AUC=0.89, κ=0.55). The model's interpretable predictions discerned client change talk and sustain talk, counselor affirmations, and reflection types, facilitating valuable counselor feedback.

Conclusions: The results of this study demonstrate that artificial intelligence techniques can accurately classify MI behavior, indicating their potential as a valuable tool for enhancing MI proficiency in online helplines for mental health. Provided that the data set size is sufficiently large with enough training samples for each behavioral code, these methods can be trained and applied to other domains and languages, offering a scalable and cost-effective way to evaluate MI adherence, accelerate behavioral coding, and provide therapists with personalized, quick, and objective feedback.

(J Med Internet Res 2024;26:e53562) doi:10.2196/53562

KEYWORDS

motivational interviewing; behavioral coding; suicide prevention; artificial intelligence; effectiveness; counseling; support tool; online help; mental health

XSL•FO
RenderX

Introduction

Background

Motivational interviewing (MI) is a client-centered counseling style that helps individuals change their behavior by resolving ambivalence using nondirective conversation techniques. It has been found effective in improving a wide range of health-related behaviors [1], such as weight management [2], addictive behaviors [3], and promoting self-management in patients with chronic health conditions [4]. During MI, counselors use a specific set of conversation techniques, most notably open-ended questions, and reflections, to let clients voice their own arguments for a particular behavior change. They are encouraged to elaborate on these reasons. This way, clients reason themselves into changing their behavior, strengthening their intrinsic motivation for the behavior change and avoiding the often-triggered defensive mechanisms when others argue for such a change. MI is a crucial anchor in guiding the counseling process during chat-based conversations at the Dutch national organization for suicide prevention (Dutch: 113 Zelfmoordpreventie).

Despite the simplicity of its principles, MI can be a challenging skill to learn, and MI requires substantial expertise to apply effectively [5]. Earlier research has shown that counselors at 113 applied MI techniques consistently during chat conversations but could not strategically deploy MI techniques to elicit enough change talk from clients to change their behavior intrinsically [6]. Therefore, it becomes imperative to improve the proficiency level of counselors in applying MI techniques to conduct conversations more effectively, especially, since eliciting change talk from clients' accounts for the effectiveness of MI [7].

One way to achieve this is through the automated evaluation of counselor responses to clients' expressed language utterances. By increasing their behavior awareness, counselors can significantly reduce cognitive effort and reflect on MI insights for education. Multiple validated proficiency measures exist for MI [8], and tools are already in development to measure treatment fidelity automatically [5,9]. In the context of chat-based helplines, these tools can provide counselors with immediate feedback during ongoing chats, potentially improving the quality of the service. Chat-based helplines also present a unique opportunity for developing such treatment fidelity tools due to the availability of extensive databases of written conversations. Also, Lundahl et al [1] found that MI is a robust intervention across patient characteristics, which gives these tools broad applicability in numerous health settings.

Enhancing MI Effectiveness Through Artificial Intelligence

Artificial intelligence (AI) has made a significant impact in recent years in many fields, the field of clinical mental health being no exception. AI offers enormous potential to analyze large data sets through machine learning (ML) algorithms. By analyzing data from MI sessions, ML algorithms can identify successful and unsuccessful applications of MI concepts, supporting and training MI practitioners. In addition, counselors can use an AI support tool to evaluate the quality of their sessions. These tools can help improve and assess counselors' MI proficiency cost-effectively and tailor additional training to their needs.

AI can also speed up the coding of MI sessions, making it easier to analyze and provide feedback *during* and *after* a counseling session. Providing counselors with ongoing feedback seemed especially important for learning MI [10]. Besides, immediate feedback has a more powerful impact on skill development than delayed feedback [9].

Although large amounts of data are typically required to train ML models to perform well on complex tasks such as capturing MI behavior, AI has developed techniques that perform well in domains with limited available data, providing insights into developing and improving evidence-based health interventions [11].

Since behavioral coding is often time-consuming, several studies have explored the automated annotation of MI transcripts in counseling sessions using ML techniques. Hasan et al [12] conducted experiments on automating the annotation of weight loss counseling sessions using the MI Sequential Code for Observing Process Exchanges (MI-SCOPE) codebook. They assessed various classification methods, incorporating linguistic, contextual, and semantic features based on linguistic inquiry and word count (LIWC) [13]. Their experiments showed that a support vector machine (SVM) model with these features achieved 75% accuracy in automatically annotating MI transcripts containing 17 behavioral codes. Idalski Carcone et al [14] aimed to develop a classification model to automatically code clinical encounter transcripts about weight loss using the MI-SCOPE behavioral code scheme. Their SVM model achieved a 69.6% F_1-score on 17 classes. Tanana et al [15] introduced 2 ML models for automatically coding MI sessions. The researchers found that the best-performing ML model had a good or higher utterance–level agreement with human coders (Cohen κ>0.60) for open and closed questions, affirmations, and giving information. However, there was a poor agreement for client change talk, client sustain talk, and therapist MI-congruent behaviors. Pérez-Rosas et al [16] presented a model for predicting MI counselor behaviors in multiple medical settings. Their SVM classifier performed well for more frequently encountered behaviors (reflections and questions) using N-grams, syntactic, and semantic LIWC features [13]. However, the performance varied much per predicted class, also obtaining lower performance for emphasizing autonomy and affirmations. Tavabi et al [17] compared the classification performance of client behaviors throughout MI psychotherapy sessions with students having alcohol-related problems using pretrained embeddings and interpretable LIWC features. Their best-performing model (pretrained RoBERTa) achieved an F_1-score of 0.66 in a 3-class classification. Saiyed et al [5] developed a Technology-Assisted Motivational Interviewing Coach incorporating ML models to deliver MI predictions for counseling sessions about tobacco cessation. Using a novel deep learning architecture combining a large fine-tuned language model and graph theory, the automated change talk/sustain

XSL•FO
RenderX

talk/follow-neutral classifier achieved an accuracy of 0.74 and an F_1-score of 0.75.

For a comprehensive overview of research papers using ML to classify MI behavior for assessing treatment fidelity, we refer to Ahmadi et al [18]. Multimedia Appendix 1 provides a schematic overview of the related work, including study context, study size, used fidelity measure, and—if reported—the coding reliability estimate. The application domain significantly varies, which also applies to the reporting of coding reliability estimates. Assessing treatment fidelity and reliability holds enormous relevance for evaluating study quality and the successful integration of MI into practice. A meta-analysis on the effect of MI on medication adherence found that interventions that examined fidelity and provided counselors with feedback on their fidelity were more effective than those that did not [19], indicating that a higher fidelity may lead to improved intervention outcomes. Frost et al [20] highlighted that fidelity is often poorly measured and reported. Moreover, MI adherence and fidelity demonstrated considerable variation across different settings and application domains [20-22].

Despite the promising algorithm performances, predicting MI-congruent counselor behavior and eliciting client change talk was challenging [15,16]. Besides, few studies adhered to best-practice ML guidelines. Although testing methods on unseen data is an essential measure of method performance in ML, only a small proportion of studies tested their methods on holdout data. A holdout subset provides a final estimate of the ML model's performance after it has been trained and validated. Similarly, Ahmadi et al [18] found that almost half of the studies in their review did not describe how they undertook data preprocessing.

For readers to assume that ML methods will generalize on future data, researchers must report these methodological processes clearly and transparently, including robust coding reliability and fidelity measures. Previous studies showed the feasibility of providing feedback to counselors via a support tool [23], consistently measuring fidelity and reporting Krippendorff's alpha estimates for interrater reliability.

This Study

This study aims to investigate the performance of AI models in classifying MI behavior and explore the feasibility of using these models in helplines as an automated support tool for counselors in clinical practice. We use a coded data set of 253 chat-based MI counseling sessions conducted at the chat helpline of 113 Suicide Prevention. We train and compare different AI algorithms to classify client- and counselor MI behavior based on language use to identify the most suitable model for the task.

The key contributions of this paper are as follows: (1) to the best of our knowledge, this is the first research that combines AI and MI with a focus on suicide prevention. (2) We aim to assist counselors in a suicide prevention helpline to overcome the practical challenges of eliciting change talk and enhancing awareness of conversation quality by providing feedback. (3) Our AI approach is described in detail, adhering to the best practices in the field and establishing a benchmark for implementing similar techniques in various settings.

Methods

Data Set

This study used a coded data set of 253 chat conversations (constituting 23,982 chat messages, 12,125 counselor messages, and 11,857 client messages) from chat-based MI counseling sessions conducted at 113 between July 2020 and January 2021. All chats were Dutch language chats and lasted at least 20 minutes. Janssen et al [6] described the exact data collection procedure.

Participants

Participants in the data set contacted the 113 crisis chat service in the Netherlands between 8:30 AM and 10:30 PM. All clients who spoke Dutch, filled out both a pre-and postchat questionnaire, and reported at least some suicidal ideation on the prechat questionnaire (score ≥1 on a 7-point Likert scale) were eligible for participation in the study [6].

Ethical Considerations

The ethics review committee of the VU University Medical Center in Amsterdam reviewed and approved this study (2020.105). The national legislation and institutional requirements did not require written informed consent from the participants. All nonessential identifying details have been omitted.

For analysis, we used only (cleaned) text of the chat messages without any personalized metainformation (including—but not limited to—age, gender, ethnicity, or clinical diagnosis). There was no collection procedure for other additional data *before*, *during*, or *after* a chat conversation. The publication of the results did not have any negative impact on the participants. Participants did not receive any form of compensation.

Procedure

Measures

Practical instruments exist to understand the quality and effectiveness of applying MI in counseling conversations. Researchers coded the data set with the MI-SCOPE coding instrument [24]. Researchers created this tool to explore the relationships between essential theoretical constructs of MI, the therapy process, and client outcomes. The focus is on analyzing the relationship between MI-specific interviewer behaviors and subsequent client behaviors within an MI session. The MI-SCOPE combines 2 successful coding systems: the MISC [25] and the commitment language coding system developed by Amrhein et al [26].

The MI-SCOPE provides 5 indices of treatment integrity, including the percentage of MI-consistent responses, the relative amount of open questions, the proportion of complex reflections, the reflection-to-question ratio, and the proportion of change talk. Hurlocker et al [8] indicated that reliability estimates for the MI-SCOPE are generally fair to excellent.

While most studies have used the MISC only [18] or the well-validated but relatively short MITI, these instruments do not provide information on the amount of change talk and sustain talk expressed by the client, whereas the MI-SCOPE

XSL•FO
RenderX

does. The MI-SCOPE thus covers more aspects of MI, incorporating both client and counselor behavior, and is more time-efficient [27]. Had the MITI or the MISC been used instead of the MI-SCOPE, Janssen et al [6] would not have detected the insufficiency of MI effectiveness in eliciting client change talk. Research by Magill et al [28] and Pas et al [29] also emphasizes the importance of fidelity measures in MI. The studies suggest that therapist adherence to MI techniques can influence client engagement and outcomes, and high-fidelity counseling can improve intervention effectiveness.

Data Set Coding and Reliability

Researchers who coded the data set [6] followed recommendations by O'Connor and Joffe [30] and described the exact coding procedure in their paper. Of the total number of counselor messages, Janssen et al [6] labeled 9177 counselor messages with fine-grained MI behavioral codes and 2948 chat messages with less fine-grained codes, indicating only MI congruency.

The coding process for all MI conversations lasted 4 months, from January 1, 2021, to May 12, 2021. The researchers used the qualitative data analysis tool ATLAS.ti 9 for the coding procedure and assessing reliability estimates. Intercoder reliability was sufficient, as Janssen et al [6] reported a Krippendorff's alpha-binary of 0.82 for the percentage of MI-consistent responses and 0.90 or higher for open questions, closed questions, and reflections. Generally, researchers consider an alpha-binary over 0.90 acceptable in all cases, while an alpha-binary ranging from 0.80 to 0.90 is deemed sufficient.

Code Grouping

To predict counselor behavior congruent with MI, we partnered with a seasoned psychologist at 113 (listed as the fourth author, WJ) to group the annotated MI behavioral codes—as outlined in the MI-SCOPE coding manual [31]—considering the practical challenges within the counseling process.

We combined all closed questions, negative and neutral reflections, and positive reflections (simple and complex), yielding 17 code groups for counselor language. For counselor language, we created 2 groups of the MI-SCOPE codes based on whether counselor language elicited client change talk (7477 nonevocative messages; 1700 evocative messages) and whether it was MI congruent (8765 MI-congruent messages; 3360 MI incongruent messages; see Table 1). We excluded the labels *Raise Concern* and *Direct* from further analysis due to their low occurrence in the data set. We did not assign detailed labels to the 4 client codes (*Ask*, *Follow/Neutral*, *Change Talk*, *and Sustain Talk*). Initial data analysis revealed that only 18.52% of all counselor messages were evocative and 70.33% were MI congruent.

Table 1. MI[a] code groups for counselor language and whether or not a code is assigned evocative or MI congruent.

MI code group	Evocative	MI congruent
Advise with Permission	×	✓
Advise without Permission	×	×
Affirm (Aff)	✓	✓
Closed Question	×	×
Confront (Con)	×	×
Emphasize Control (Econ)	✓	✓
Filler (Fill)	×	×
General Information (GI)	×	×
Open Question (OQ+)	✓	✓
Open Question (OQ−)	×	✓
Open Question (OQ0)	×	✓
Permission Seeking	×	✓
Reflection (+)	✓	✓
Reflection (0−)	×	✓
Self-Disclose (Sdis)	×	×
Structure (Str)	×	✓
Support (Sup)	×	✓

[a]MI: motivational interviewing.

Analytic Strategy

We trained and evaluated 4 ML models and 1 deep learning model to classify client and counselor MI behavior based on language use. ML models benefit from human-extracted features, while deep learning models learn complex patterns without feature selection. Although deep learning models have better performance potential, these models require more data and have less interpretable reasoning. We further describe the

XSL•FO
RenderX

feature selection process, the models, and how we addressed these limitations of the deep learning model.

Feature Selection

Available Features

In total, we extracted 5850 features for each of the MI-coded chat messages. These features included high-level concepts such as topic, grammar, and sentiment, as well as low-level concepts such as counting the occurrence of each word. For a complete list of all feature categories used for the ML models, see Multimedia Appendix 1.

Feature Subsets

To gain insight into the impact of each feature category on the classification performance, we created subsets by adding 1 or more feature categories to the previous subset, resulting in 8 sets of features that we used to train the ML models, starting with the initial subset containing only the basic feature categories and ending with the final set containing all extracted features.

Train-Validation-Test Split

For each classification problem, we split each group of chat messages—stratified by class distribution—80%:10%:10% to create training, validation, and test data sets. To rightly measure model performance, it is essential to hold out data. We used the training set exclusively to train the models, evaluated the training progress on the validation set, and obtained the final performance using the test set. Table S3 in Multimedia Appendix 1 shows the number of classes and instances for each classification problem.

Learning Algorithms

We trained and evaluated 4 different ML models and 1 deep learning model for each classification problem: a random forest, an SVM, k-nearest neighbors, a decision tree, and a *pretrained transformer model*.

We chose a pretrained transformer model to overcome the limitations of regular deep learning model architectures. Transformer models are a type of deep learning network that can be pretrained on a large amount of data and then fine-tuned on a smaller, more specific data set to make predictions. By pretraining on a large data set, the model can learn to understand the structure and patterns of language, making it easier to adapt to new domains, which enables the training of complex models with limited data. Researchers showed that the BERT model [32] suits this approach particularly. We used a variant of BERT (Bidirectional Encoder Representations from Transformers), called BERTje (monolingual Dutch BERT), which already has been pretrained on a large Dutch text corpus [33], and fine-tuned BERTje on our domain-specific data set for each classification problem.

A grid search technique was used to select the best model parameters, as initial testing showed that the parameter values could severely impact model performance. Table S4 in Multimedia Appendix 1 provides an overview of the models and the considered parameters. The final analysis excluded compensating for the imbalance of the class labels in the data, as initial testing also showed that it did not lead to differences in the results. To account for this, we evaluated the models using statistics that can take class imbalances into account.

We used 5-fold cross-validation to validate the models and applied minimum-maximum scaling before training. We implemented all ML models in *Python 3.8* (developed by Python Software Foundation) and implemented the fine-tuning of BERTje using *PyTorch Lightning* (developed by Lightning AI).

Evaluation Metrics

Computing the confusion matrix and conducting an area under the receiver operating characteristic curve analysis allowed us to assess the classifiers and obtain visual and statistical insights into their predictive performance. We also quantified the kappa statistic and accuracy for the best-performing models. We extracted the probability distribution of the predictions from all classifiers to compute the sample average F_1-score. The probability distribution indicates the confidence or likelihood of a specific model prediction. For a detailed explanation of these evaluation metrics, see the study by Zheng [34] and Figures S1 and S2 in Multimedia Appendix 1 [35,36].

Baseline

The baseline score provides a required point of comparison when evaluating all predictive algorithms for a classification task. We consider predicting the majority class as a baseline, meaning that we select the prediction class with the most observations and use it as the outcome for all predictions. We expect the predictive models that learn from the data to perform substantially better.

Validity

We used identical statistics, training, validation, and test samples to evaluate the trained models, making the validation of the results comparable across all models.

Explainability

To interpret the output of the models, we used Shapley Additive Explanations (SHAP) [37] as a method. SHAP provides a way to obtain the contribution of each feature in the model's prediction for a particular input. Values provided by SHAP represent a feature's average marginal contribution toward the difference between the predicted output and the model's expected output. A higher value indicates a higher contribution to the output and interprets it as a more important feature.

Results

Algorithm Performance

In this section, we present a comprehensive evaluation and performance analysis of the ML models and the transformer model BERTje across all 4 classification tasks. We further interpret the model predictions by deploying SHAP and laying out the most occurring word combinations for each prediction class.

XSL•FO
RenderX

Classifying Counselor Behavior

Fine-Grained Predictions

Figure 1 presents a performance comparison of the learning algorithms using the best parameters for classifying counselor behavior. The reported scores represent the average of 5 repeated runs for each model. The SVM model (γ=0.1, C=10) showed the highest F_1-score of 0.63 among all ML models. Random forest and SVM models outperformed the decision tree and k-nearest neighbors. For an overview of the ML model performances on the different feature subsets, see Table S5 in Multimedia Appendix 1.

Figure 1. Learning algorithm performance for predicting client- and counselor behavior. The SD for the results yields 0.015 for counselor behavior and 0.013 for client behavior. CIs for the results are given by performance value ± 2 × SD. DT: decision tree; kNN: k-nearest neighbors; RF; random forest; SVM: support vector machine.

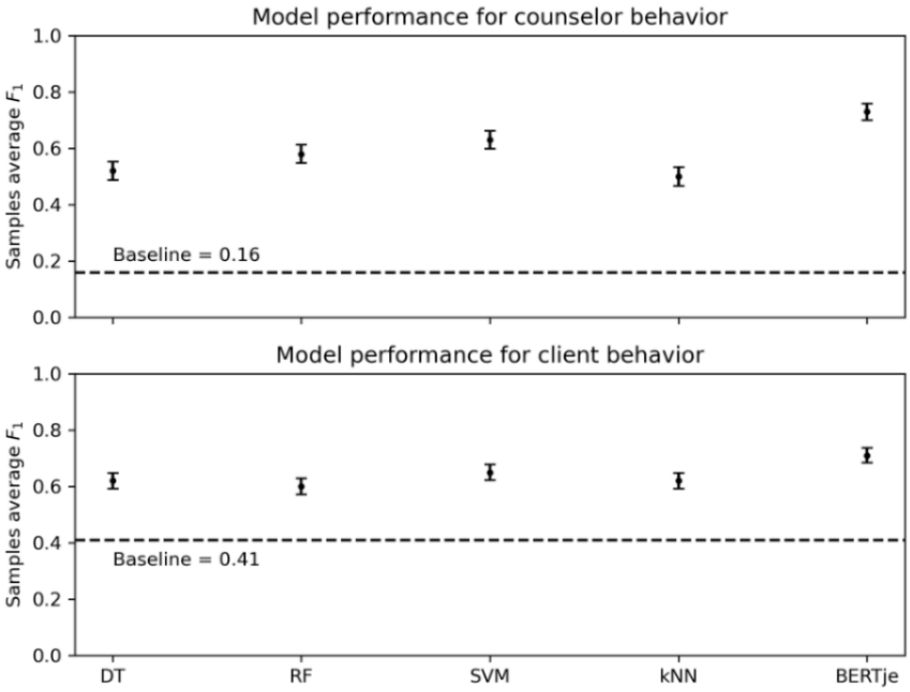

Incorporating textual information and word-embedding features resulted in the highest increase in the performance of the ML algorithms. Among all models tested, the transformer model BERTje achieved the highest performance with an F_1-score of 0.73. Table 2 shows a detailed model performance evaluation of BERTje. With an accuracy of 0.72, kappa statistic of 0.69, and area under the curve (AUC) score of 0.95, its results represent a 350% improvement in accuracy from the baseline. The AUC scores per class ranged from 0.89 to 0.99. The model performed best on fillers and affirmations with an AUC score of 0.99 and lowest on advise without permission and confront with an AUC score of 0.89. Errors mainly occurred when predicting neutral, open questions as positive open questions (14 errors), positive reflections as neutral or negative reflections, and open questions as closed questions (17 errors).

XSL•FO
RenderX

Table 2. BERTje: detailed model performance evaluation on all classification tasks[a].

Classification task	F_1-score	AUC[b]	Sample average F_1-score	Microaverage AUC	Macroaverage AUC	Accuracy	Kappa values
Counselor behavior			0.73	0.96	0.95	0.72	0.69
Advise with Permission (AWP)	0.50	0.94					
Advise without Permission (ADW)	0.39	0.89					
Affirm (Aff)	0.86	0.99					
Closed Question	0.85	0.97					
Confront (Con)	0.31	0.89					
Emphasize Control (Econ)	0.38	0.91					
Filler (Fill)	0.80	0.99					
General Information (GI)	0.67	0.96					
Open Question (OQ+)	0.64	0.95					
Open Question (OQ−)	0.69	0.97					
Open Question (OQ0)	0.72	0.94					
Permission Seeking (Perm)	0.71	0.96					
Reflection (+)	0.51	0.93					
Reflection (0−)	0.76	0.97					
Self-Disclose (Sdis)	0.57	0.96					
Structure (Str)	0.84	0.97					
Support (Sup)	0.69	0.95					
MI[c] congruency			0.88	0.94	0.92	0.87	0.65
MI-Congruent (X MI+)	0.91	0.92					
MI-Incongruent (X MI−)	0.76	0.92					
Evocative language			0.90	0.96	0.92	0.90	0.66
Evocative	0.73	0.92					
Nonevocative	0.94	0.92					
Client behavior			0.71	0.90	0.89	0.70	0.55
Ask	0.81	0.99					
Follow/Neutral (FN)	0.61	0.81					
Change Talk (X Csa+)	0.66	0.87					
Sustain Talk (X Csa−)	0.74	0.89					

[a]Cells with no numerical value indicate "not applicable."

[b]AUC: area under the curve.

[c]MI: motivational interviewing.

MI Congruency

The fine-tuned BERTje model achieved a sample average F_1-score of 0.88 in accurately predicting counselor behavior as either MI-congruent or MI-incongruent (see Table 2). In addition, it demonstrated a high accuracy of 0.87, accompanied by a kappa value of 0.65 and a macroaverage AUC score of 0.92. These results signify an accuracy improvement of 20.8% compared with the baseline performance (accuracy=0.72).

Evocative Language

The fine-tuned BERTje model achieved a sample average F_1-score of 0.90 in accurately predicting whether counselor language is evocative or nonevocative (see Table 2). Moreover, it demonstrated an accuracy of 0.90, a kappa value of 0.65, and a macroaverage AUC score of 0.92. These results signify an accuracy improvement of 9.8% compared with the baseline performance (accuracy=0.82).

XSL•FO
RenderX

Classifying Client Behavior

Figure 1 also shows a performance comparison of the learning algorithms using the best parameters for classifying client behavior. For an overview of the ML model performances on the different feature subsets, see Table S6 in Multimedia Appendix 1. All models improved classification performance compared with the baseline. The best ML model was an SVM model (γ=0.1, C=10), reaching a sample average F_1-score of 0.65. BERTje outperformed all ML algorithms, reaching a sample average F_1-score of 0.71 with an accuracy of 0.70, Cohen κ=0.55, and a macroaverage AUC score of 0.89. These results indicate an accuracy improvement of 70.7% compared with the baseline. The AUC scores per class range from 0.81 to 0.99 (see Table 2). Although the lowest occurrence across the client messages, BERTje predicted the code *Ask* best (AUC=0.99). Follow/Neutral was the hardest to predict (AUC score=0.81). We observed that errors mainly occurred in predicting Follow/Neutral messages as commitment language.

Feature Contributions

According to the SHAP feature importance analysis conducted on the best-performing ML models, word-embedding features held significant dominance. Furthermore, the number of question marks in a message emerged as a consistently influential factor for client- and counselor behaviors. Table S7 in Multimedia Appendix 1 shows the features that contribute most to the predictions of each class individually. Moreover, this table shows the top word combinations reflecting the language character of different client and counselor behaviors. The inferred prediction classes associated with the MI-SCOPE codes are generally interpretable. For example, client ambivalence becomes clear when counselors use reflections with word combinations such as “on one side,” “on the other side,” and “conflicted.” Concerning client commitment language, negative sentiment, and negations contributed to both sustain talk and change talk. When these features were *present*, client language was more likely to be associated with sustain talk rather than change talk. Contrarily, the *absence* of these features indicates more association with client change talk.

Discussion

Interpretation of the Results

The results of this study demonstrate the potential of AI models, particularly the transformer model BERTje, in classifying MI behavior in online mental health helplines. BERTje outperformed all ML models tested, achieving high levels of accuracy across all classification tasks. Although ML models obtained lower performance than the BERTje model, their high explainability adds value for gaining a deeper understanding of language use concerning specific MI behaviors.

The successful application of a fine-tuned transformer model in classifying MI behavior is consistent with other recent studies, such as by Saiyed et al [5] and Tavabi et al [17], who also used a fine-tuned transformer model to classify MI behavior in counseling sessions. Both studies also used some form of model interpretation to understand how the models make predictions and what features or words characterize each class. Our study extends this line of research by using a different data set, coding scheme, and transformer model than the previous studies. Both related studies used the MISC codebook for data annotation and did not provide any estimates of coding reliability. In contrast, our study offers an in-depth account of the procedures and methodology, including reporting on coding reliability and fidelity measures. Studies that also used the MI-SCOPE (eg, Idalski Carcone et al [14]) used a small data set and obtained lower F_1-scores than this study, highlighting the importance of using larger data sets to improve the performance of AI models in predicting MI behavior.

Our study contributes to the growing evidence base for MI as an effective intervention for various health-related behaviors. We showed that the AI models can accurately identify the effective ingredients of MI, such as client change talk and sustain talk, counselor affirmations, and reflection types—facilitating valuable counselor feedback. Furthermore, this study is the first to apply such a model to the domain of suicide prevention, which poses specific challenges and opportunities for MI. For example, counselors in a suicide prevention helpline need to adhere to MI but also balance building rapport; exploring ambivalence; focusing on engagement, collaboration, and empathy; and ensuring client safety.

Strengths and Limitations

This study held several notable strengths. In our methodology, we adhered to the best AI practices. We used a holdout test set to evaluate the performance of the AI models, providing a realistic estimate of their generalization ability. Using diverse statistics to evaluate the model performances, we make the validation process comparable across all models. We clearly described the analytic strategy, ensuring transparency and reproducibility of the research process substantiated by the comprehensive supplemental material.

In the context of generalizability, classifying MI behavior could relatively easily be deployed in other domains and other languages. These days, large language models are being pretrained on many texts in multiple languages. From an AI point of view, implementing these methods in other online mental health helplines is relatively effortless.

While this study holds several strengths, there are also limitations. The data set used in this study is relatively small, which could lead to higher variance in the test set. Some MI code groups were underrepresented in the data set (eg, less evocative statements occurred in the text than nonevocative statements). This limited data set size may restrict the model’s ability to generalize to a broader range of MI conversations. It is important to note that the model’s performance in this specific domain of suicide prevention does not guarantee its effectiveness in other settings. To assess and train the model performances in another domain’s context, it is still necessary to gather domain-specific data.

Another limitation is that human experts coded the data set used for training and evaluation. While the expertise of human coders adds value, it is essential to acknowledge that the labeling process can still be subjective. Although not applicable to this

XSL•FO
RenderX

study (since the intercoder reliability was sufficient, see section "Data Set Coding and Reliability"), individual coders may interpret and classify MI techniques and behaviors differently, potentially leading to inconsistent or inaccurately labeled data. This subjectivity in labeling could impact the performance of the automatic coding system and introduce errors or biases. A possible solution could be *active learning*—in which the AI model can interactively query a domain expert to label new data points with the desired outputs—or *reinforcement learning with human feedback* [38,39]. Reinforcement learning with human feedback has emerged as a powerful technique for refining these models. After initial training, the models receive feedback from human evaluators, enabling them to refine their approach.

A final limitation is the disregard for demographic traits of clients and counselors, such as age and cultural background. These characteristics could influence language use and model predictions. Further research is needed to refine the AI models with these factors, but these are not without ethical concerns [40]. In addition, years of counselor experience and MI proficiency could affect model effectiveness, with less experienced counselors likely benefiting more from the model.

Implications for Clinical Practice

Leveraging AI Models for Clinical Support

There are several potential ways to incorporate AI models into clinical practice to enhance MI proficiency in online helplines for mental health:

By integrating AI models into chat-based counseling platforms, counselors can receive instant feedback on their MI behavior during sessions. This feedback allows counselors to review their generated messages before sending them and make necessary adjustments, such as changing a closed question to an open one. Initial results suggest that counselors find such systems acceptable [41], but more studies are needed to evaluate the reception and impact of these tools in different settings and populations.

By offering postsession feedback and training to counselors, they can reflect on their performance and pinpoint areas where they may require additional training or support. By analyzing data from multiple counseling sessions, AI models can detect patterns or trends in counselor MI behavior and develop tailored training programs that offer recommendations for training or support. This integration can assist counselors in identifying areas where they may need to modify their approach or apply MI techniques more effectively.

Figure 2 shows a schematic overview illustrating the proof of concept of the support tool. Studies can investigate the feasibility of such a tool using a Wizard-of-Oz approach, where an experienced counselor acts as the support tool to simulate a best-case scenario. This setup could serve as a preliminary test for the viability of the support tool without requiring the development of a fully equipped AI tool for examining the potential advantages or disadvantages.

Figure 2. Schematic overview illustrating the proof of concept of the support tool.

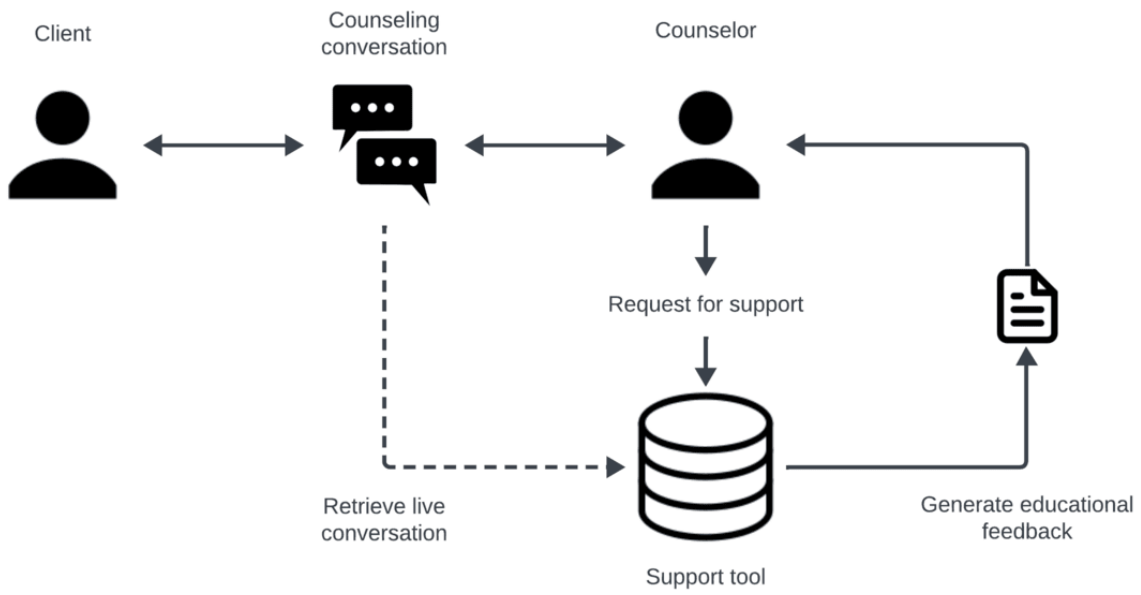

Madeira et al [42] proposed such a tool, and Salmi et al [41] examined it in focus groups with counselors and tested viability in a simulated environment. Studies emphasize the significance of tools that can help ease the workload of counseling [43,44]. Several studies also highlight relevant elements concerning the viability of AI in mental health counseling [43,45,46]. Therefore, a comprehensive evaluation of feasibility is necessary.

Helpline administrators or supervisors could also use AI models to monitor and evaluate the quality of counseling services provided by their organization. Using AI models to classify MI behavior and providing counselors with feedback offer a scalable and cost-effective solution for enhancing MI proficiency in helplines and other counseling settings. Many helplines struggle

XSL•FO
RenderX

to find good staff, and the turnover is high, so reducing the time to proficiency is very valuable.

Evaluating the Effectiveness of AI Models in Clinical Practice

One way to evaluate the effectiveness of AI models in clinical practice would be to conduct pilot studies or randomized controlled trials analyzing changes in counseling outcomes over time. By comparing data from counseling sessions before and after the introduction of AI models, it would be possible to determine whether using these models leads to improved health outcomes for individuals seeking help.

Another way to evaluate the effectiveness of AI models would be to monitor changes in counselor MI behavior over time. By analyzing data from multiple counseling sessions, it would be possible to determine whether counselors who receive feedback from AI models improve their proficiency in applying MI techniques.

Feedback from counselors and clients could also provide valuable insights into the effectiveness of AI models in enhancing MI proficiency in chat-based counseling sessions. Counselors could provide feedback on the usefulness and accuracy of the provided feedback by the AI models, while clients could provide feedback on the quality of their interactions with counselors.

Next Steps to Take

Clinicians or researchers interested in leveraging AI for their specific use case in online helplines for mental health can already take initial steps to get started. A crucial first step in leveraging AI to enhance MI proficiency is to collect data from counseling sessions, such as chat transcripts and relevant metadata. Another important step is connecting with other clinicians and researchers in the field. By joining an active community of professionals working with MI and AI, one can benefit from the knowledge and resources created by others. These resources may encompass pretrained AI models, guidelines for collecting and analyzing data, and opportunities to collaborate and share knowledge with others in the field.

Future Directions

While the results are promising, additional research is needed to evaluate the performance of these models on larger data sets with sufficient representation of each class. Future studies may explore alternative modeling techniques that better capture the conversational structure in classifying MI behavior. For instance, graph-based models can store information about the relationships between messages within and across conversations.

While this work and earlier research successfully quantified and validated the technical aspects of MI, it is also relevant to consider fundamental principles of MI. These principles contain the conversational processes that guide interactions between counselors and clients. Adhering to these processes ensures that counselors do not exceed the level a client is comfortable with and adapt their behavior appropriately based on the specific context. For instance, an appropriate question during the engaging process (building rapport between client and counselor) may become counterproductive during the evoking process, where the main goal is to elicit change talk. In these processes, concepts such as collaboration, engagement, and empathy also play a significant role. Recent work is already exploring integrating these concepts into AI models in the context of online helplines for mental health [44,47,48].

Furthermore, applying the methods used in this study to other languages and institutions could provide valuable additional validation of the study findings. A final point of future work would be to investigate and measure the potential improvement of MI quality in the chat helpline before and after counselors used MI insights during their conversations over time. This could also be combined with a support tool for MI feedback in a randomized controlled trial.

Conclusions

The results of this study demonstrate that AI techniques can accurately classify MI behavior, indicating their potential as a valuable tool for enhancing MI proficiency in online helplines for mental health. Provided that the data set size is sufficiently large with enough training samples for each behavioral code, these methods can be trained and applied to other domains and languages, offering a scalable and cost-effective way to evaluate MI adherence, speed up behavioral coding, and provide therapists with personalized, quick, and objective feedback.

Acknowledgments

The generative AI tool *Grammarly* was used to improve readability, narrative coherence, and grammatical structure. Grammarly was solely used as a writing aid. The scientific content remains the work of the authors.

Data Availability

The data sets, materials, and code used or analyzed during this study are available from the corresponding author upon reasonable request.

Data Transparency Statement

The data reported in this manuscript were collected as part of a larger data collection. Findings from the data collection have been reported in separate manuscripts. MS 1 [6] investigated the impact of training counselors in motivational interviewing on the outcomes of a chat-based suicide prevention helpline using the annotated MI behavioral codes and pre- and postchat questionnaire

XSL•FO
RenderX

scores to measure suicidal ideation. MS 2 (this study) focuses on a subset of the collected data, including the raw text of the chat messages and their annotated MI behavioral codes to explore the use of automated behavioral coding to enhance the effectiveness of motivational interviewing. Unlike MS 1, MS 2 does not use suicidal ideation scores, and the raw text was used only for data annotation in MS 1 and not for analysis.

Authors' Contributions

MP led the conceptualization, formal analysis, investigation, methodology, project administration, software, validation, visualization, and writing—original draft. SS led the conceptualization, data curation, methodology, software, and writing—original draft. SM led the conceptualization, project administration, supervision, and writing—review and editing. WJ led the data curation, methodology, resources, validation, and writing—review and editing. RM led the conceptualization, supervision, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed insights into our methods and findings.
[PDF File (Adobe PDF File), 169 KB - jmir_v26i1e53562_app1.pdf]

References

1. Lundahl B, Moleni T, Burke BL, Butters R, Tollefson D, Butler C, et al. Motivational interviewing in medical care settings: a systematic review and meta-analysis of randomized controlled trials. Patient Educ Couns 2013 Nov;93(2):157-168. [doi: 10.1016/j.pec.2013.07.012] [Medline: 24001658]
2. Michalopoulou M, Ferrey AE, Harmer G, Goddard L, Kebbe M, Theodoulou A, et al. Effectiveness of motivational interviewing in managing overweight and obesity: a systematic review and meta-analysis. Ann Intern Med 2022 Jun;175(6):838-850. [doi: 10.7326/M21-3128] [Medline: 35344379]
3. Lindson N, Thompson TP, Ferrey A, Lambert JD, Aveyard P. Motivational interviewing for smoking cessation. Cochrane Database Syst Rev 2019 Jul 31;7(7):CD006936 [FREE Full text] [doi: 10.1002/14651858.CD006936.pub4] [Medline: 31425622]
4. Phillips R, Hogden A, Greenfield D. Motivational interviewing to promote self-management. In: Promoting Self-Management of Chronic Health Conditions: Theories and Practice. Cary, NC: Oxford University Press; 2017:126-144.
5. Saiyed A, Layton J, Borsari B, Cheng J, Kanzaveli T, Tsvetovat M, et al. Technology-assisted motivational interviewing: developing a scalable framework for promoting engagement with tobacco cessation using NLP and machine learning. Procedia Computer Science 2022;206:121-131. [doi: 10.1016/j.procs.2022.09.091]
6. Janssen W, van Raak J, van der Lucht Y, van Ballegooijen W, Mérelle S. Can outcomes of a chat-based suicide prevention helpline be improved by training counselors in motivational interviewing? A non-randomized controlled trial. Front Digit Health 2022;4:871841 [FREE Full text] [doi: 10.3389/fdgth.2022.871841] [Medline: 35801118]
7. Magill M, Apodaca TR, Borsari B, Gaume J, Hoadley A, Gordon REF, et al. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. J Consult Clin Psychol 2018 Feb;86(2):140-157 [FREE Full text] [doi: 10.1037/ccp0000250] [Medline: 29265832]
8. Hurlocker MC, Madson MB, Schumacher JA. Motivational interviewing quality assurance: A systematic review of assessment tools across research contexts. Clin Psychol Rev 2020 Dec;82:101909 [FREE Full text] [doi: 10.1016/j.cpr.2020.101909] [Medline: 32920371]
9. Imel ZE, Pace BT, Soma CS, Tanana M, Hirsch T, Gibson J, et al. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. Psychotherapy (Chic) 2019 Jun;56(2):318-328. [doi: 10.1037/pst0000221] [Medline: 30958018]
10. Schwalbe CS, Oh HY, Zweben A. Sustaining motivational interviewing: a meta-analysis of training studies. Addiction 2014 Aug;109(8):1287-1294. [doi: 10.1111/add.12558] [Medline: 24661345]
11. Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. Lancet Psychiatry 2022 Mar;9(3):243-252. [doi: 10.1016/S2215-0366(21)00254-6] [Medline: 35183281]
12. Hasan M, Carcone AI, Naar S, Eggly S, Alexander GL, Hartlieb KEB, et al. Identifying effective motivational interviewing communication sequences using automated pattern analysis. J Healthc Inform Res 2019;3(1):86-106 [FREE Full text] [doi: 10.1007/s41666-018-0037-6] [Medline: 31602420]
13. Pennebaker J, Boyd R, Jordan K, Blackburn K. The Development and Psychometric Properties of LIWC2015. The Development and Psychometric Properties of LIWC2015. 2015. URL: http://hdl.handle.net/2152/31333 [accessed 2024-06-20]

XSL•FO
RenderX

14. Idalski Carcone A, Hasan M, Alexander GL, Dong M, Eggly S, Brogan Hartlieb K, et al. Developing machine learning models for behavioral coding. J Pediatr Psychol 2019 Apr 01;44(3):289-299 [FREE Full text] [doi: 10.1093/jpepsy/jsy113] [Medline: 30698755]
15. Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V. A comparison of natural language processing methods for automated coding of motivational interviewing. J Subst Abuse Treat 2016 Jun;65:43-50 [FREE Full text] [doi: 10.1016/j.jsat.2016.01.006] [Medline: 26944234]
16. Pérez-Rosas V, Mihalcea R, Resnicow K, Singh S, An L, Goggin KJ, et al. Predicting counselor behaviors in motivational interviewing encounters. : Association for Computational Linguistics; 2017 Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers; 2024 June 20; Valencia, Spain p. 1128-1137 URL: https://aclanthology.org/E17-1106 [doi: 10.18653/v1/e17-1106]
17. Tavabi L, Tran T, Stefanov K, Borsari B, Woolley JD, Scherer S, et al. Analysis of behavior classification in motivational interviewing. Proc Conf 2021 Jun;2021:110-115 [FREE Full text] [doi: 10.18653/v1/2021.clpsych-1.13] [Medline: 34337617]
18. Ahmadi A, Noetel M, Schellekens M, Parker P, Antczak D, Beauchamp M, et al. A systematic review of machine learning for assessment and feedback of treatment fidelity. Psychosoc Interv 2021;30(3):139-153. [doi: 10.5093/pi2021a4]
19. Palacio A, Garay D, Langer B, Taylor J, Wood BA, Tamariz L. Motivational interviewing improves medication adherence: a systematic review and meta-analysis. J Gen Intern Med 2016 Aug;31(8):929-940 [FREE Full text] [doi: 10.1007/s11606-016-3685-3] [Medline: 27160414]
20. Frost H, Campbell P, Maxwell M, O'Carroll RE, Dombrowski SU, Williams B, et al. Effectiveness of Motivational Interviewing on adult behaviour change in health and social care settings: A systematic review of reviews. PLoS One 2018;13(10):e0204890 [FREE Full text] [doi: 10.1371/journal.pone.0204890] [Medline: 30335780]
21. Hallgren KA, Dembe A, Pace BT, Imel ZE, Lee CM, Atkins DC. Variability in motivational interviewing adherence across sessions, providers, sites, and research contexts. J Subst Abuse Treat 2018 Jan;84:30-41 [FREE Full text] [doi: 10.1016/j.jsat.2017.10.011] [Medline: 29195591]
22. DiClemente CC, Corno CM, Graydon MM, Wiprovnick AE, Knoblach DJ. Motivational interviewing, enhancement, and brief interventions over the last decade: A review of reviews of efficacy and effectiveness. Psychol Addict Behav 2017 Dec;31(8):862-887. [doi: 10.1037/adb0000318] [Medline: 29199843]
23. Flemotomos N, Martinez VR, Chen Z, Singla K, Ardulov V, Peri R, et al. Automated evaluation of psychotherapy skills using speech and language technologies. Behav Res Methods 2022 Apr;54(2):690-711 [FREE Full text] [doi: 10.3758/s13428-021-01623-4] [Medline: 34346043]
24. Moyers TB, Martin T. Therapist influence on client language during motivational interviewing sessions. J Subst Abuse Treat 2006 Apr;30(3):245-251. [doi: 10.1016/j.jsat.2005.12.003] [Medline: 16616169]
25. Moyers T, Martin T, Catley D, Harris K, Ahluwalia J. Assessing the integrity of motivational interviewing interventions: reliability of the motivational interviewing skills code. Behav Cognit Psychother 2003;31(2):177-184. [doi: 10.1017/S1352465803002054]
26. Amrhein PC, Miller WR, Yahne CE, Palmer M, Fulcher L. Client commitment language during motivational interviewing predicts drug use outcomes. J Consult Clin Psychol 2003 Oct;71(5):862-878. [doi: 10.1037/0022-006X.71.5.862] [Medline: 14516235]
27. Lundahl B, Droubay BA, Burke B, Butters RP, Nelford K, Hardy C, et al. Motivational interviewing adherence tools: A scoping review investigating content validity. Patient Educ Couns 2019 Dec;102(12):2145-2155. [doi: 10.1016/j.pec.2019.07.003] [Medline: 31514978]
28. Magill M, Gaume J, Apodaca TR, Walthers J, Mastroleo NR, Borsari B, et al. The technical hypothesis of motivational interviewing: a meta-analysis of MI's key causal model. J Consult Clin Psychol 2014 Dec;82(6):973-983 [FREE Full text] [doi: 10.1037/a0036833] [Medline: 24841862]
29. Pas ET, Borden L, Debnam KJ, De Lucia D, Bradshaw CP. Exploring profiles of coaches' fidelity to Double Check's Motivational Interviewing-embedded coaching: Outcomes associated with fidelity. J Sch Psychol 2022 Jun;92:285-298. [doi: 10.1016/j.jsp.2022.04.003] [Medline: 35618375]
30. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922-160940691989913. [doi: 10.1177/1609406919899220]
31. Martin T, Moyers T, Houck J, Christopher P. Motivational Interviewing Sequential Code for Observing Process Exchanges (MI-SCOPE) coder's manual. 2005. URL: https://casaaweb.unm.edu/assets/docs/scope.pdf [accessed 2024-06-20]
32. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 June; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/n18-2]
33. de Vries W, van Cranenburgh A, Bisazza A, Caselli T, van Noord G, Nissim M. BERTje: a Dutch BERT model. 2019. URL: http://arxiv.org/abs/1912.09582 [accessed 2024-06-20]
34. Zheng A. Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls. Sebastopol, CA: O'Reilly Media; 2015 Sep. URL: https://www.oreilly.com/library/view/evaluating-machine-learning/9781492048756/ [accessed 2024-06-20]

XSL•FO
RenderX

35. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 2016 Jul 02;20(1):37-46. [doi: 10.1177/001316446002000104]
36. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Computational Linguistics 2008 Dec;34(4):555-596. [doi: 10.1162/coli.07-034-R2]
37. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. Vol 30. Red Hook, NY: Curran Associates Inc; 2017 Presented at: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec; Long Beach, CA p. 4768-4777 URL: http://arxiv.org/abs/1705.07874
38. Christiano P, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems. Vol 30. Red Hook, NY: Curran Associates, Inc; 2017 Presented at: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec; Long Beach, CA URL: https://arxiv.org/abs/1706.03741
39. Chung NC, Dyer G, Brocki L. Challenges of large language models for mental health counseling. arXiv preprint. 2023 Nov. URL: https://arxiv.org/abs/2311.13857 [accessed 2024-06-20]
40. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023 May;23(5):4-13 [FREE Full text] [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
41. Salmi S, Mérelle S, Gilissen R, Brinkman WP. Content-based recommender support system for counselors in a suicide prevention chat helpline: design and evaluation study. J Med Internet Res 2021 Jan 07;23(1):e21690 [FREE Full text] [doi: 10.2196/21690] [Medline: 33410755]
42. Madeira T, Bernardino H, Francisco DSJ, Gomide H, Machado NM, da Silva BMP, et al. A framework to assist chat operators of mental healthcare services. Stroudsburg, Pennsylvania, US: Association for Computational Linguistics; 2020 Presented at: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS); 2020 Nov; Online p. 1-7. [doi: 10.18653/v1/2020.nlposs-1.1]
43. Hsu SL, Shah RS, Senthil P, Ashktorab Z, Dugan C. Helping the helper: supporting peer counselors via AI-empowered practice and feedback. arXiv preprint. 2023 May. URL: https://arxiv.org/abs/2305.08982 [accessed 2024-06-20]
44. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023 Jan 23;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
45. De Freitas J, Agarwal S, Schmitt B, Haslam N. Psychological factors underlying attitudes toward AI tools. Nat Hum Behav 2023 Nov;7(11):1845-1854. [doi: 10.1038/s41562-023-01734-2] [Medline: 37985913]
46. Wutz M, Hermes M, Winter V, Köberlein-Neu J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. J Med Internet Res 2023 Sep 26;25:e46548 [FREE Full text] [doi: 10.2196/46548] [Medline: 37751279]
47. Sharma A, Miner A, Atkins D, Althoff T. A computational approach to understanding empathy expressed in text-based mental health support. 2020 Nov Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov; Online p. 5263-5276. [doi: 10.18653/v1/2020.emnlp-main.425]
48. Zech JM, Johnson M, Pullmann MD, Hull TD, Althoff T, Munson SA, et al. An integrative engagement model of digital psychotherapy: exploratory focus group findings. JMIR Form Res 2023 Apr 26;7:e41428 [FREE Full text] [doi: 10.2196/41428] [Medline: 37099363]

Abbreviations

AI: Artificial intelligence
AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
BERTje: monolingual Dutch BERT
LIWC: linguistic inquiry and word count
MI: motivational interviewing
MI-SCOPE: MI Sequential Code for Observing Process Exchanges
ML: machine learning
SHAP: Shapley Additive Explanations
SVM: support vector machine

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.10.23; peer-reviewed by A Carcone, H Szlyk; comments to author 18.01.24; revised version received 01.03.24; accepted 02.06.24; published 01.08.24.

Please cite as:
Pellemans M, Salmi S, Mérelle S, Janssen W, van der Mei R
Automated Behavioral Coding to Enhance the Effectiveness of Motivational Interviewing in a Chat-Based Suicide Prevention Helpline: Secondary Analysis of a Clinical Trial
J Med Internet Res 2024;26:e53562
URL: https://www.jmir.org/2024/1/e53562
doi:10.2196/53562
PMID:39088244

©Mathijs Pellemans, Salim Salmi, Saskia Mérelle, Wilco Janssen, Rob van der Mei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Influencing the Implementation of Digital Advance Care Planning: Qualitative Interview Study

Andy Bradshaw[1], BSc, PhD; Jacqueline Birtwistle[2], BSc, MSc; Catherine J Evans[1,3], BSc, MSc, RGN, SPQDN, PhD; Katherine E Sleeman[1], BSc, MBBS, MRCP, PhD; Suzanne Richards[2], BSc, PhD; Robbie Foy[2], MBChB, MRCGP, MSc, MFPHM, PhD; Pablo Millares Martin[4], LMS, DCS, MSc; Paul Carder[5], BSc; Matthew J Allsop[2], BSc, PhD, CPsychol; Maureen Twiddy[6], BSc, MSc, PhD

[1]Cicely Saunders Institute, Kings College London, London, United Kingdom
[2]Leeds Institute of Health Sciences, University of Leeds, Leeds, United Kingdom
[3]Sussex Community NHS Foundation Trust, Brighton, United Kingdom
[4]Whitehall Surgery, Leeds, United Kingdom
[5]NHS West Yorkshire Integrated Care Board, White Rose House, Wakefield, United Kingdom
[6]Hull York Medical School, Institute of Clinical and Applied Health Research, Allam Medical Building, University of Hull, Hull, United Kingdom

Corresponding Author:
Matthew J Allsop, BSc, PhD, CPsychol
Leeds Institute of Health Sciences
University of Leeds
Worsley Building
Leeds, LS2 9LU
United Kingdom
Phone: 44 1133434185
Email: m.j.allsop@leeds.ac.uk

Abstract

Background: Palliative care aims to improve the quality of life for people with life-limiting illnesses. Advance care planning conversations that establish a patient's wishes and preferences for care are part of a person-centered approach. Internationally, electronic health record systems are digital interventions used to record and share patients' advance care plans across health care services and settings. They aim to provide tools that support electronic information sharing and care coordination. Within the United Kingdom, Electronic Palliative Care Coordination Systems (EPaCCS) are an example of this. Despite over a decade of policy promoting EPaCCS nationally, there has been limited implementation and consistently low levels of use by health professionals.

Objective: The aim of this study is to explore the factors that influence the implementation of EPaCCS into routine clinical practice across different care services and settings in 2 major regions of England.

Methods: A qualitative interview study design was used, guided by Normalization Process Theory (NPT). NPT explores factors affecting the implementation of complex interventions and consists of 4 primary components (coherence, cognitive participation, collective action, and reflexive monitoring). Health care and social care practitioners were purposively sampled based on their professional role and work setting. Individual web-based semistructured interviews were conducted. Data were analyzed using thematic framework analysis to explore issues which affected the implementation of EPaCCS across different settings at individual, team, organizational, and technical levels.

Results: Participants (N=52) representing a range of professional roles were recruited across 6 care settings (hospice, primary care, care home, hospital, ambulatory, and community). In total, 6 themes were developed which mapped onto the 4 primary components of NPT and represented the multilevel influences affecting implementation. At an individual level, these included (1) EPaCCS providing a clear and distinct way of working and (2) collective contributions and buy-in. At a team and organizational level, these included (3) embedding EPaCCS into everyday practice and (4) championing driving implementation. At a technical level, these included (5) electronic functionality, interoperability, and access. Breakdowns in implementation at different levels led to variations in (6) confidence and trust in EPaCCS in terms of record accuracy and availability of access.

Conclusions: EPaCCS implementation is influenced by individual, organizational, and technical factors. Key challenges include problems with access alongside inconsistent use and engagement across care settings. EPaCCS, in their current format as digital

XSL•FO
RenderX

advance care planning systems are not consistently facilitating electronic information sharing and care coordination. A redesign of EPaCCS is likely to be necessary to determine configurations for their optimal implementation across different settings and locations. This includes supporting health care practitioners to document, access, use, and share information across multiple care settings. Lessons learned are relevant to other forms of digital advance care planning approaches being developed internationally.

(J Med Internet Res 2024;26:e50217) doi:10.2196/50217

KEYWORDS

palliative care; electronic palliative care coordination systems; electronic health record systems; advance care planning; end of life care; technology; Normalization Process Theory; NPT; qualitative

Introduction

Palliative care aims to improve the quality of life for people with life-limiting illnesses through a person-centered, multidisciplinary, and holistic approach [1]. The focus on person-centeredness is reflected in health policy both within the United Kingdom and internationally [2-5]. Key to facilitating person-centered care in palliative services is the concept of advance care planning. This involves having planned conversations with a patient around their individual goals, wishes, and preferences for their current and future care [6]. If a person's preferences are documented and shared, there is evidence of beneficial outcomes. Advance care planning has been associated with better quality of care, helping people to be cared for and die in their usual place of residence, and preventing unplanned hospital admissions [7-9]. However, other studies indicate that it has no impact on patient outcomes or quality of life [10-12]. Despite this uncertainty, advance care planning has been adopted by health care systems internationally as a key feature of person-centered care.

While there are benefits associated with advance care planning, delivery of palliative care requires the involvement of, and communication between, multiprofessional services across different settings of care (ie, hospice, general practice, community-based care, out-of-hours services, hospitals, emergency services, care homes, and social care) [13,14]. To overcome the need for information sharing across multiple health care providers and settings, electronic health record systems are increasingly being used to document and share advance care planning information. This approach has been reported across countries that include the United States [15,16], Australia [17], and the United Kingdom. In the United Kingdom, the use of electronic health record systems for documenting and sharing advance care plans are called Electronic Palliative Care Coordination Systems, often referred to using the acronym EPaCCS.

EPaCCS emerged in response to the Department of Health's 2008 End of Life Care Strategy, which advocated for improved coordination of care at the end of life for people with life-limiting conditions (ie, cancer and noncancer conditions, including dementia) [18]. Policy drivers for the widespread use of electronic systems (such as EPaCCS) across health and social care providers have continued to the present day [19,20]. This includes, for example, an expectation that care records for all people living with a long-term condition should include a person's care needs and preferences, and should be shared with all those involved in their care [19]. The development of EPaCCS sought to overcome challenges arising through the fragmentation of health systems that can lead to patients not receiving person-centered care at the right time and in the right place [21,22]. This can result in patient needs not being met, unplanned and avoidable hospital admissions, and patients not being cared for, or dying in a place of their choice [23-25].

EPaCCS have been developed across the United Kingdom since 2008 and multiple variants have arisen. These include standalone web-based electronic registers such as Coordinate my Care which was implemented in London [26], and template forms integrated into already-existing electronic patient records, such as in Leeds [27] and the Key Information Summary in Scotland [28]. Regardless of the mode used for implementing EPaCCS locally across regions, the Palliative and End of Life Care Information Standard specifies the core content that should be recorded and shared (eg, demographic information, diagnosis, medication, advance care planning information, Do Not Attempt Cardiopulmonary Resuscitation decisions, and preferred places of care and death) [29]. The expectation is that once this information is stored, it should be possible to share across all settings involved in the delivery of palliative and end-of-life care, as well as sharing any updates on the care plan.

However, there is widespread variation with regard to how EPaCCS are implemented within local health care systems (eg, who initiates the creation of a record, which settings of care can access and edit information in EPaCCS records) [14] which has resulted in variable levels of interoperability and access [14,30-33]. In part, this may be a factor influencing the low use rates reported for EPaCCS, with 9%-43% of people with palliative care needs having an EPaCCS record created before death [14,27,34,35]. Alongside low use, there is uncertainty about how EPaCCS are being used in routine practice and limited evidence of their impact, inhibiting the development of an evidence base to guide how their implementation might be optimized [13,14,30,36,37]. The aim of this study was to explore health care professionals' perspectives on factors that influence the implementation of EPaCCS in routine clinical practice across different care settings in England.

Methods

Design

We undertook a qualitative interview study. This approach was selected as we sought to develop new insights and knowledge on a relatively understudied topic area [38]. Our study was informed by an interpretative paradigm [39]. That is, we explored the study aim from the standpoint of ontological

XSL•FO
RenderX

relativism (the acceptance of multiple, mind-dependent realities) and epistemological constructionism (an appreciation that knowledge generated during data analysis and write-up reflected interpretations made collectively by the research team) [40]. We reported the research in line with the consolidated criteria for reporting qualitative research (COREQ [consolidated criteria for reporting qualitative research], see Multimedia Appendix 1) [41].

Theoretical Perspective

EPaCCS can be conceptualized as a complex intervention. This is because they comprise multiple interacting components and operate at the interface of different health care professionals, organizations, settings of care, and patients and their families [42]. Normalization Process Theory (NPT) is an implementation theory that is used to explain how complex interventions are normalized (ie, deeply embedded into, and used as part of, routine practice) [43-45].

In explaining the different types of "work" that people do in normalizing complex interventions, NPT consists of 4 interlinked primary constructs (Table 1) [46,47]. We used these primary constructs of NPT as a guiding theoretical framework to guide data collection, analysis, and interpretation in understanding the factors affecting the implementation of EPaCCS into routine clinical practice.

Table 1. An overview of the 4 constructs from the Normalization Process Theory (NPT), definitions of the constructs, and how the constructs were applied and understood in the context of the study.

NPT construct	Definition[a]	Framework applied in this study[b]
Coherence ("sense-making work")	"How do people work together in everyday settings to understand and plan the activities that need to be accomplished to put an intervention and its components into practice?"	The ways in which participants think about distinguishing use of digital systems from other formats for advance care planning, collectively agreeing on the purpose of EPaCCS[c], individually understanding what EPaCCS requires of them, and constructing potential value of EPaCCS for their work.
Cognitive participation ("relational work")	"How do people work together to create networks of participation and communities of practice around interventions and their components?"	The ways in which participants become engaged in understanding what they need to do and support for EPaCCS to be sustained, influencing how EPaCCS use can be sustained, adapting to EPaCCS to support use by themselves and others, and supporting others' engagement with EPaCCS.
Collective action ("operational work")	"How do people work together to enact interventions and their components?"	The ways in which participants perform the tasks required for EPaCCS to support advance care planning, build accountability and maintain confidence in the use of EPaCCS, understand the appropriateness of existing roles and responsibilities relating to EPaCCS use, and view the resources and organizational support for EPaCCS use.
Reflexive monitoring ("appraisal work")	"How do people work together to appraise interventions and their components?"	The ways in which participants appraise the effects of EPaCCS, themselves and with colleagues understand whether EPaCCS are operating well, individually understand and respond to the impact of EPaCCS, and modify their work in response to their appraisal of EPaCCS.

[a]Definitions derived from May et al (2022) [46].

[b]How definitions of NPT constructs were "in the simplest possible terms" [48] and applied to data collection, analysis, and interpretation.

[c]EPaCCS: Electronic Palliative Care Coordination Systems.

Recruitment and Settings

Recruitment took place in 2 UK regions in London (population circa 9 million) and West Yorkshire (population circa 2.3 million). Participants comprised a subsample of respondents to an earlier survey who had agreed to be contacted for follow-up interviews. In West Yorkshire, EPaCCS comprise a template that is embedded within a patient's electronic record, generally in the electronic record system used by primary care providers that can share information across different care settings within a defined geographical area. Within London, at the time of this study, the most used EPaCCS system was "Coordinate My Care" (CMC). This was a standalone system (eg, it operated outside routinely used patient records), and enabled patients to access their own records [14]. Since the conception of this study, it has been superseded by an EPaCCS system called "Universal Care Plan."

Participants were initially selected using purposive maximum variation sampling to gather the widest range of perspectives [49]. This entailed sampling participants based on specific criteria (geographical location, professional role, setting of care, and levels of understanding of or engagement with EPaCCS based on previous survey responses). The logic underpinning this approach was to explore our research aim from diverse perspectives [49]. We worked across primary, secondary, and tertiary care settings in purposefully recruiting doctors, nurses, care home staff, paramedics, and general practitioners. Difficulties in recruitment, however, meant that we supplemented our recruitment approach with convenience sampling using the same criteria.

Participants were approached via email and provided with a participant information sheet. A combination of verbal and written informed consent was obtained, either by AB or JB, prior to interviews being conducted. Recruitment ran concurrently with data collection between November 2021 and June 2022. The concept of "information power" [50] was used to guide decisions on when to halt recruitment and data collection. This entailed several meetings during which research team members (AB, JB, MT, MA, CE, and KS) considered whether and when data collected from our sample held enough relevant and detailed information to comprehensively understand our research aim.

XSL•FO
RenderX

Data Collection

Single, web-based interviews were conducted by 1 of 2 researchers, both of whom had prior experience in qualitative interviewing [AB (male, research fellow) and JB (female, research fellow)]. The topic guide (provided in Multimedia Appendix 2) comprised questions about how participants used EPaCCS, alongside the factors that they perceived affected their implementation. During development, these questions were mapped onto the 4 primary constructs of NPT. Interviews were audio recorded, anonymized, and transcribed verbatim. All participant personal data were stored in a secure cloud storage platform within password-protected files. These data were only accessible to, shared between, and used by members of the research team, using data-sharing agreements.

Analysis

Interview data were managed using NVivo (version 12; Lumivero) [51] and analyzed using thematic Framework analysis [52]. The 4 primary constructs of NPT were used as the theoretical framework to guide the coding and interpretation of data. Data analysis included moving between induction and deduction. We first used NPT to deductively build our initial analytic framework and then supplemented this through inductive coding in which we explored how patterns grounded in the data related to and enriched our analytic framework. Analysis consisted of seven iterative steps: (1) familiarization (through rereading transcripts), (2) coding (by labeling relevant segments of transcripts that aligned with our research aims), (3) creation of an initial analytic framework (by grouping similar codes into categories and categories into themes), (4) indexing (by applying our analytic framework back to raw data and refining it where appropriate), (5) charting (by creating a matrix that explored differences in data across region, role, and setting of care), (6) description (through defining and describing themes), and (7) interpretation (using our theoretical framework to further interpret our findings through the write-up of data). This approach allowed us to conduct within- and between-case pattern matching to explore where participant accounts on the use and implementation of EPaCCS converged or diverged, and how this was influenced by contextual factors (ie, setting, region, and role). Data analysis was led by AB, with fortnightly meetings between authors (AB, JB, MT, and MJA) to review the ongoing coding and analysis.

Multiple approaches were used to ensure rigor during data analysis. Throughout data collection and analysis, the researchers engaged in different forms of reflexivity. This included reflecting introspectively (inward reflections on how they impacted the research process and vice versa) and intersubjectively (reflections on relationships between them and participants) [53]. These were used as a "springboard for interpretations and more general insight" into the ways through which understandings of the research aim were being co-constructed through the research process, including analysis [53]. This included regular discussions with JB, who shared data collection and who was familiar with the data corpus. Moreover, throughout the analytical process, members of the wider interdisciplinary research team (consisting of academics and clinicians with experience in palliative care research and practice from across care settings including hospital, community, and primary care) acted as "critical friends" [54]. This entailed working collaboratively through regular meetings and written feedback in which findings were constructively challenged, reflexivity encouraged, and alternative interpretations of the data proposed. This process took place until the research team agreed that the final analytic framework accurately reflected participant accounts.

Ethical Considerations

The North of Scotland Research Ethics Committee approved the research (reference 21/PR/0428). In this study, we also recognized ethics as a reflexive process through engaging in "ethics in practice" [55]. This approach was used to remain responsive to and navigate ongoing and potentially unexpected ethical issues that may have arisen throughout data collection, analysis, and write-up (eg, by reflecting on how the research might affect professionals' clinical practice and potential impact on patients and carers). All participants provided written consent prior to participating in the study. Before analysis, all interview transcripts were deidentified and stored on a secure cloud storage platform only accessible to the study team. Organizations in which participants were based were offered reimbursement of £75 (US $95.91) for allowing a professional to participate in the study. The level of the incentive payment was based on the cost of 1 hour of a locum doctor in the United Kingdom and was agreed upon by the research ethics committee and study funder.

Results

A total of 52 people (characterized in Table 2) participated out of the 99 people approached for interview participated (characterized in Table 2), from London (n=29) and West Yorkshire (n=23). These participants represented a range of different professionals who work across hospice, primary care, care home, hospital, and community settings.

Six themes were developed and are represented under the corresponding constructs of NPT (coherence, cognitive participation, collective action, and reflexive monitoring). Figure 1 provides an overview of these main themes.

XSL•FO
RenderX

Table 2. Characteristics of study participants (N=52) including location, settings of care, and role.

	London	West Yorkshire	Total
Interviews, n	29	23	52
Setting of care, n			
Hospice	8	4	12
Primary care	7	5	12
Care home	5	4	9
Hospital	5	3	8
Ambulance	3	3	6
Community nurse	1	4	5
Role, n			
Registered nurse	6	7	13
Care home	4	3	6
Community	1	4	5
Hospital	1	1	2
General practitioner	7	5	12
Palliative care consultant	6	4	13
Paramedic	3	3	6
Clinical Nurse Specialist	4	1	5
Care home manager	1	2	3

Figure 1. An overview of the themes and key messages generated from the thematic framework analysis as they align with the 4 constructs of normalization process theory. EPaCCS: Electronic Palliative Care Coordination Systems.

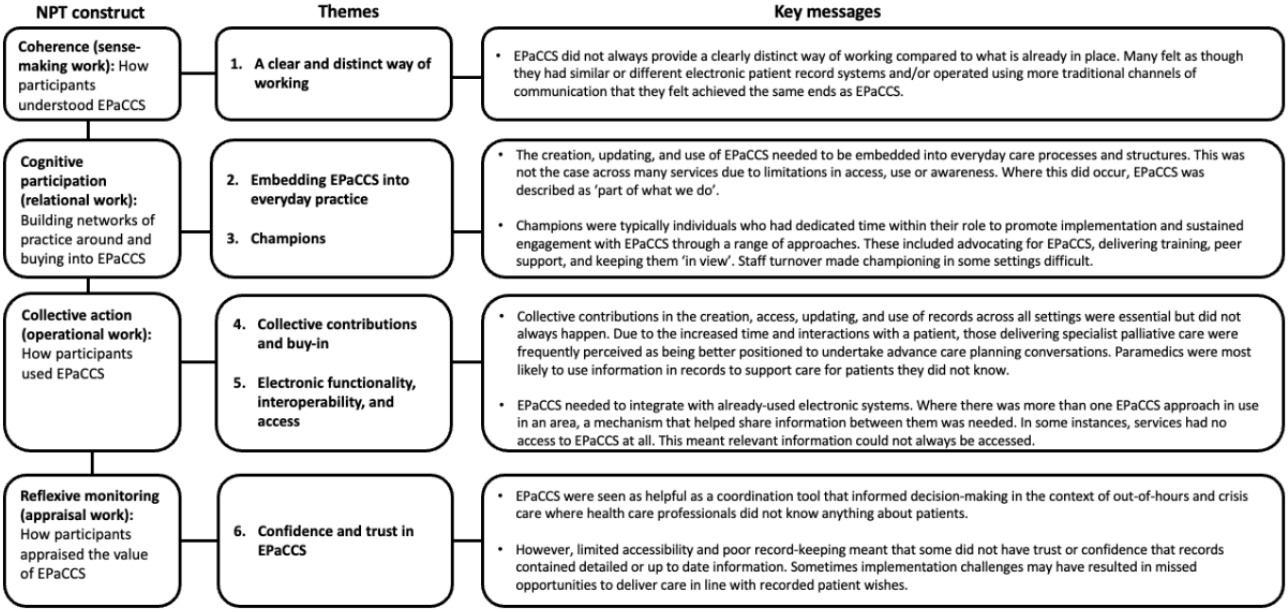

Coherence: How Participants Understood EPaCCS—Theme 1: A Clear and Distinct Way of Working

For a complex intervention to be normalized, it needs to be understood as a clear and distinct way of working that enhances patient care. In the case of EPaCCS, this required health care professionals to appreciate how adding all relevant advance care planning information onto a single digital platform allowed for easier sharing among organizations. Across all settings of care, however, participants did not always see how EPaCCS enhanced patient care and outcomes compared to traditional channels of communication. These included using paper-based discharge summaries, cover letters, face-to-face conversations, emails, and telephone for direct communication of advance care plans across settings:

XSL•FO
RenderX

> *I don't think it [EPaCCS] changes much in terms of our GP engagement with patients … our primary channel of communication is telephone, face to face, perhaps email and EPaCCS doesn't sort of sit with any of those. Our primary function really is to provide the clinical care and record that clinical care and those are traditional methods of doing that. EPaCCS is an add-on, and I think because it's an add-on, it's not the primary form of communication.* [Lon18, GP, London]

Participants working in care homes across London and West Yorkshire reported using a range of different electronic patient record systems for documenting their residents' care plans. These systems facilitated the documentation of patient wishes and preferences that could be easily accessed, reviewed, and updated internally by care home staff. Some electronic record systems used in care homes were also capable of integrating and sharing data with other systems across different settings, but this was often limited to those in general practice and not on the scale envisaged by EPaCCS:

> *The electronic care plan system that we are using, it's something called PCS, so Person Centred Software. You can do your usual stuff like you said, day-to-day care notes and things like that. There also with PCS, the party piece it has is something called GP connect … with PCS if you do kind of speak with them, they are able to kind of merge certain systems together with their [other healthcare settings'] system … as far as I know PCS works with all of them* [WY12&WY13, care home managers, West Yorkshire]

Cognitive Participation: How Participants Built Networks of Practice Around and Bought Into EPaCCS

Theme 2: Embedding EPaCCS Into Everyday Practice

The service-wide embedding of EPaCCS into everyday practice varied and was partly influenced by the extent to which they were an integral part of everyday care processes, structures, and settings. In some contexts, entire services had no access to EPaCCS. Where EPaCCS were present and accessible across multiple settings, implementation was still thwarted because they were not used consistently by some health care professionals:

> *it can be like an easy thing to do and I think it gets ingrained in your normal kind of process of doing a patient's notes … but I think when people just don't know about it or don't know how to access it, it then becomes, it seems more of a challenge to actually set one up whereas once you've got used to it, it just becomes part of your normal routine first assessment, set up CMC and then it's done.* [Lon5, hospice community doctor, London]

Some participants reported that EPaCCS were integrated into everyday clinical routines. In these instances, EPaCCS were referred to as being "part of what we do" [WY2, community/hospice, West Yorkshire]. This included ensuring that EPaCCS management was integrated into key care processes such as initial assessments, caseload reviews, admission and discharge planning, multidisciplinary team meetings, handover sheets, and standard operating procedures:

> *Whenever we take on a new patient onto the caseload, we will do our level best to have a conversation about the current EPaCCS system we use, which is Coordinate My Care, to get consent to put people on that system … it's very much part of our mantra, it's something that we do … We talk about Coordinate My Care at our weekly multidisciplinary team meetings, so we ask people to check that people are on there and if they're not, we try and think of a plan to get somebody on there.* [Lon26, hospice consultant nurse, London]

Theme 3: Championing Driving Implementation

Participants across all professional groups in both regions reflected on the role of colleagues who championed the use of EPaCCS. Champions were typically individuals who had been given dedicated time within their role to promote implementation and who encouraged staff engagement with EPaCCS. Champions used a range of approaches including offering peer support, taking active roles in teaching and education, presenting the potential benefits of EPaCCS, and keeping them present or "in view" in everyday clinical practice:

> *Having local champions who are just, [I] don't want to say checking, but just ensuring that locally they're being completed, that they're demonstrating a difference. You know, there's no point doing it if it's not demonstrating any benefit really is there?* [WY2, community/hospice consultant, West Yorkshire]

There were examples where engaging staff in the use of EPaCCS and learning how different systems work was described as a "constant recurring battle" [Lon06, hospital Clinical Nurse Specialist, London] that required time, dedication, and energy. This view was particularly present in hospital settings and the challenge was mainly attributed to high staff turnover. Different staff meant that the same messages and training had to be repeated continuously for EPaCCS to remain a priority for teams. However, such training did not always translate into increases in health professional use:

> *The reality is, I think that EPaCCS is underused, but generally across the hospital. That's why I have to be out there doing education and encouragement… it's continually education, trial, training, nudging, pushing to get them to use it because, a level of busyness, a churn of staff, you know they're churning staff all the time, they're coming from different Trusts who are not used to SystmOne, never mind EPaCCS, so it's a continual, continual, continual thing and training* [WY7, hospital nurse, West Yorkshire]

Collective Action: How People Enact EPaCCS

Theme 4: Collective Contributions and Buy-in

Collective contributions referred to the extent to which health care professionals across settings of care contributed to the

XSL•FO
RenderX

creation, sharing, updating, and use of EPaCCS records to inform care. Underpinning collective contributions was the degree to which health care professionals "bought into" EPaCCS by seeing them as a legitimate part of their role or as supporting the work of others. There was a general agreement that EPaCCS needed "buy-in *from everyone - not just palliative care teams – for it to work*" [WY2, consultant, community/hospice consultant, West Yorkshire]. Despite this, health care professionals working in specialist palliative care were often the ones creating and updating EPaCCS records.

Participants had different perceptions in terms of the skills and capacity of professionals across different care settings to support advance care planning. Some felt that staff working in specialist palliative care were best placed to initiate sensitive conversations about advance care planning and end-of-life choices. Others reflected that although they believed health care professionals working outside of palliative care settings could broach advance care planning conversations, they did not always have the confidence to do so. Indeed, participant accounts suggested that when palliative care services were not involved in a patient's care, this led to a general lack of clarity over who should do what, when, and how, which often resulted in the ad hoc creation of records.

> *If they're working in specialist palliative care, most of those people will have those skills [for advance care planning]. If we then look at people who don't deal exclusively with palliative care but see a lot of it, so district nurses, elderly medicine doctors, general practitioners, I think there is a lot of skill there. There's not always the skill and there's often a gap in confidence to apply the skills that people have … when people don't have the skills and confidence, that first conversation where we seek, where we explain to the patient where they are in their illness and the fact that they're in a palliative phase of their illness and seeking consent to use an EPaCCS doesn't happen* [WY6, hospice, West Yorkshire]

> *…in [our] Community Trust we've also got a respiratory service who are involved with people that are end stage of respiratory failure and we also have heart failure nurses and diabetic nurse specialists, and you know, those sorts of questions are talked about with patients often. But what we're trying to do as a service in palliative care [non-specialist palliative care in Community Nurse Team] is encourage that to be done because it isn't done as much as it should be really. You know when people are actually reaching sort of end-stage heart failure but yet nobody's actually spoken to them about their wishes at the end of life. They feel it's not their responsibility.* [WY9, community nurse, West Yorkshire]

A concern among participants working in general practice was that it was difficult to contribute to EPaCCS because they did not always fit with their existing ways of working. Given the time constraints and competing priorities in general practice, accessing an EPaCCS record and then conducting and documenting advance care planning conversations was often seen as unmanageable and unrealistic:

> *we've only got a 10- or 15-minute window to see that patient for their current problem, so we don't bother to update the CMC after just because of sheer time. So, unless you're having a special CMC kind of session and you've dedicated a bit of time to go and do a home visit on someone or you've planned it in that you're going to update the CMC and those wishes and concerns etc, that's only when it really gets touched by the GP practice.* [Lon16, GP, London]

Conversely, paramedics with access in London were likely to use EPaCCS records, typically because it helped decision-making around the urgent management of patients that they were hitherto unfamiliar with:

> *…we use it just as part of our decision making … I would say it's a big part of my role… in the main bulk of my role which is in an ambulance setting I use it all the time, it's second nature and it's very valuable.* [Lon29, paramedic, London]

Theme 5: Electronic Functionality, Interoperability, and Access

At a technical level, the integration of EPaCCS within existing electronic systems was important to their implementation. However, according to participants, this process of integration had not always occurred. Across most care settings in London, participants were frustrated that CMC was not seamlessly interoperable with existing electronic patient record systems. This lack of technical interoperability (ie, basic data exchange capabilities between systems) created a restrictive process that resulted in additional work because health care professionals had to remember (and frequently update) log-in details and enter duplicate information across different systems:

> *Coordinate My Care for us is a completely separate system … it doesn't pull data from the current electronic system, you still have to kind of manually enter the patient's name, address, NHS number … it is extra work … that kind of influence[s] how detailed a care record might be. Sometimes we will just put on the basic information that you think's important … if it was integrated into a current kind of electronic system then I guess it would just make it easier.* [Lon5, hospice community doctor, London]

In services across West Yorkshire, problems were caused by the fact that EPaCCS were embedded within several different electronic health record systems. However, because no mechanism in place allowed for the sharing of information between these systems, health care professionals could not always access relevant information from EPaCCS records when they needed it:

> *In this area a lot of folks were using SystmOne and we use EMIS … So, they [hospice services] can't see what we've done on our system and it's a bit messy … I know EPaCCS is supposed to be a document that everyone can access and fill in, but you can't, ours is just on our system and 'cos no one else can fill it*

XSL•FO
RenderX

> *in or see what's been changed that's where it falls down…. it is stupid that you can't share an EPaCCS with anyone, it seems like a bit pointless* [WY17, GP, West Yorkshire]

Reflexive Monitoring ("Appraisal Work"): Appraising the Value of EPaCCS—Theme 6: Confidence and Trust in EPaCCS

Participants across different regions and settings reflected on the potential value of EPaCCS as a tool that could facilitate the coordination of care. Others reflected on first-hand experiences of how EPaCCS were valuable in the context of crisis and out-of-hours care. In particular, paramedics who had access felt that EPaCCS provided vital information needed to effectively support person-centered decision-making with people whom they did not know:

> *If I saw a CMC for example where it was recorded that the patient had a preference for treatment care in the home, that would make it much more likely that I would dispatch one of my colleagues because we already know that that is the patient's preferences and so as far as possible, we're going to work to make that happen … End of life care pushes against the normal direction of paramedic care, [the] normal direction of paramedic care is rescue, save … end of life care obviously isn't about life-saving, it's about dignified death, symptom control management. It's a change of thinking and so a lot of paramedics struggle with that, and I think they will look for anything that will help them guide them in that process and I think CMC is one that people are very familiar with using and generally find quite helpful.* [Lon29, paramedic, London]

> *I've worked with the ambulance service in the time before it [EPaCCS] was standard practice and I would say it's such a necessity now that we've got it, if we lost it I think it would literally be like losing my hand … it cuts out awkward conversations and it also cuts out us doing something that may be an issue would be against their wishes … it takes that anxiety or the uncertainty out of the what are we going to do and see what's best for this patient.* [Lon28, paramedic, London]

However, other professionals reported that EPaCCS records were sometimes of poor quality and that there were frequent problems with accessing the system. Consequently, this reduced their confidence that EPaCCS records contained sufficiently up-to-date information to support decision-making. Concerns were not about the potential of EPaCCS to improve care but regarding implementation issues such as restricted access, shareability, and inconsistent use of these records by staff. There was a fear that such problems could lead to patients receiving interventions and treatments that were against their stated and recorded wishes and preferences. This was particularly the case for paramedics who, without access to records, were more likely to make risk-averse decisions to hospitalize patients in the absence of knowing their wishes:

> *I can tell you categorically that we have not acted in the patient's best interest … I'll have taken somebody into hospital without my knowledge that [they have] an end-of-life care plan somewhere* [WY21, paramedic, West Yorkshire]

Discussion

Principal Results

This study explored factors that influenced the implementation of EPaCCS in routine clinical practice across different care settings in 2 major regions of England. It identifies and elaborates on challenges around the implementation of EPaCCS, including problems with access, and inconsistent use and engagement across settings. A key issue was technological limitations, with separate electronic health records often operating in parallel systems or failing to provide sufficient documentation or access. Such problems have led to the potential value of EPaCCS being unrealized.

Guided by sociotechnical systems theory [56], Figure 2 summarizes these issues by highlighting how interactions between the individual, team, organizational, and technical dimensions of EPaCCS affected implementation. The content of this figure is grounded in the data. It was generated inductively through highlighting relationships between each of our themes alongside how they related to different levels of practice. This allowed us to move beyond description by explaining linkages between themes and bringing them together in a way that tells an overarching story of health care professionals' perspectives on the processes that influenced the implementation of EPaCCS.

XSL•FO
RenderX

Figure 2. An overview of the individual, team, organizational, and technical factors that affect the implementation of digital advance care planning systems (ie, EPaCCS) in routine practice. EPaCCS: Electronic Palliative Care Coordination Systems.

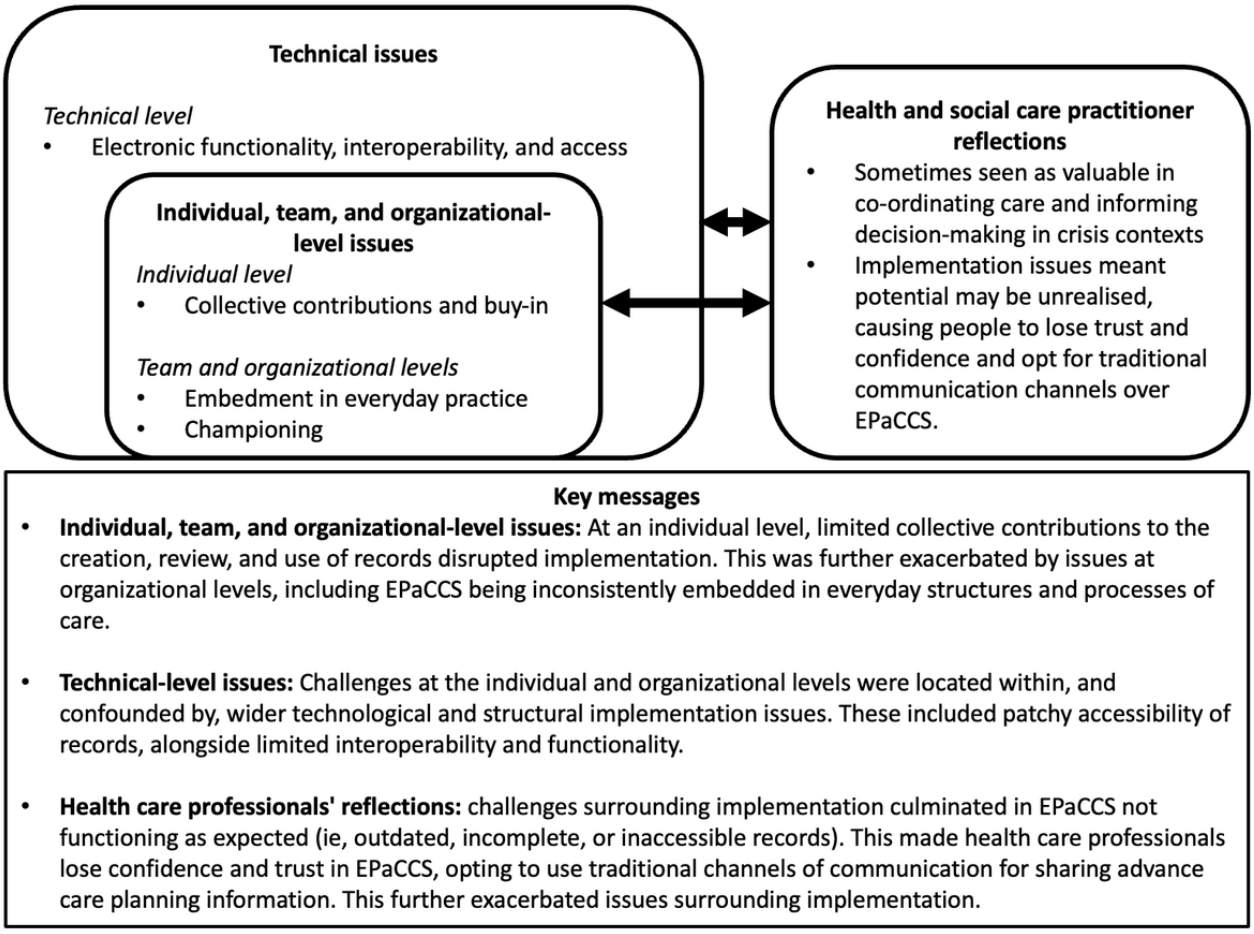

Comparison With Prior Work

A key theme in this study was that limited buy-in and collective contributions to the creation, initiation, and use of EPaCCS across care settings affected implementation. Previous research has primarily used quantitative approaches to explore EPaCCS implementation, such as determining the proportion of EPaCCS records created across a specified population, and the average number of days that EPaCCS records are created before death [27,34]. However, our findings add novel and unique contributions by describing who initiates EPaCCS records, the factors that influence this, and the differences across regions. While some general practitioners and community nurses reported recording advance care planning discussions electronically, our study also highlights contributions made by health care professionals working across specialist palliative care settings. For some participants, providers of specialist palliative care were perceived as being better placed to initiate advance care planning conversations. A referral of a patient to specialist palliative care was perceived as affording greater time to interact with patients and other staff groups, alongside their experience in facilitating advance care planning discussions with patients and their families.

Participant experiences resonate with previously documented challenges related to advance care planning such as perceived lack of time, hesitancy in initiating conversations, and lack of care continuity [57-59]. Lack of clarity over who contributed to records and the timing of these contributions often resulted in poor-quality data. Such issues led to fears that EPaCCS records were incomplete or out of date, with some records being overly detailed or conversely, insufficiently informative to effectively support decision-making in out-of-hours or emergency care. The successful implementation of interventions that work across organizations (such as EPaCCS) relies on “whole systems” thinking [35]. Such implementation entails those involved in the use and implementation of EPaCCS (including commissioners and those responsible for service [re]design) accounting for the needs and influence of people working across different care settings and specialties.

In this study, care home staff reported having detailed discussions regarding residents’ wishes and preferences for care and documenting these within their own electronic systems. This aligns with previous research that the close relationships between care home staff and residents mean that staff are also well placed to have the sensitive and in-depth conversations required for advance care planning [60]. In this study, while electronic systems were reported to be used and engaged across care homes, the information contained within them was mostly inaccessible to external services. Care homes were also largely unable to access or provide helpful and detailed information from EPaCCS that could be used by other services. This highlights a key gap in the “whole systems” approach to EPaCCS. Addressing this gap is especially important given that

XSL•FO
RenderX

care homes are projected to be the most common place of death in England by 2040 [61].

The implementation of EPaCCS at the individual, team, and organizational levels was affected by wider technological challenges. Issues related to the technical dimensions of EPaCCS have been well documented in the literature [13,30,33,36]. However, this study provides novel findings relating to health professional experiences of EPaCCS, further strengthened by its collection of data across 2 large geographical regions. This study builds on previous work and contributes details of how and why issues with access, functionality, and technical interoperability (ie, data integration, presentation, and exchange) [62] affected the implementation of EPaCCS across the United Kingdom. In London, challenges were experienced around the ability to move data between EPaCCS and existing electronic patient record systems (and vice versa) without duplication. Across West Yorkshire, a major interoperability issue was the failure of different electronic patient record systems to exchange advance care planning information. Moreover, while paramedics were most likely to use information within EPaCCS records in London to support decision-making in crisis contexts, they did not have this access across West Yorkshire. While these interoperability limitations affected health care professionals in different ways, they ultimately hindered the harnessing of the electronic information sharing and care coordination that EPaCCS is intended for.

Implications for Policy and Practice

Our findings highlight implications for practice relating to the use and implementation of electronic systems for palliative and end-of-life care. For complex interventions like EPaCCS to become normalized into everyday practice, they must fit within and enhance established systems of care [63]. It is also crucial that users can see their benefits to patient care and clinical practice compared to traditional ways of working [64,65]. However, these findings suggest that, in their current format, EPaCCS are not working as intended for facilitating person-centered care. Implementation challenges which resulted in missed opportunities to deliver care in line with recorded patient wishes, sometimes led to a loss of trust and confidence in EPaCCS, instead staff opted for more traditional means of communication and a return to traditional communication methods. Consequently, a redesign of EPaCCS is likely to be necessary to achieve the optimal configuration for successful implementation across different settings of care and geographical locations. To this end, we have generated a set of questions focusing on factors that were found to be influencing the uptake and use of EPaCCS across different settings of care (see Table 3).

Table 3. Questions for health care practitioners, commissioners of health care services, and policy makers to consider in optimizing the uptake and implementation of electronic information-sharing systems across different settings of care.

Level of action	Relevant to	Questions to consider
Individual or user	Professionals across settings of care involved in the creation, updating, reviewing, and use of electronic information sharing systems to inform decision-making	• What are current levels of staff confidence in broaching and conducting advance care planning conversations with patients who have life-limiting illnesses? • To what extent are professionals across care settings clear on whose responsibility (including their own) it is to engage in advance care planning and documentation of any resultant patient wishes and preferences? • What are current levels of staff awareness of the different stages of electronic information sharing (eg, the creation, update, review, and use to inform decision-making) relating to their role?
Team or organizational	Team leaders and service management	• Is there collective agreement across care teams on how to embed electronic information sharing systems within everyday structures and processes (eg, at admission or discharge, multidisciplinary team meetings, and handovers)? • Is there a dedicated member or group of staff that can help to monitor uptake and promote sustained engagement with electronic information sharing systems both within and between services? • Is there clarity on the way in which electronic systems are intended to be used for the documentation of advance care plans (eg, which patients it is intended for, and when it might be used as part of their care)?
Technical or structural	Commissioners and policy makers	• Do all services that are involved in the care of people with life-limiting illnesses have access to an electronic information sharing system (including paramedics and care homes)? • Have the outcomes of the electronic information-sharing systems been agreed upon, alongside how outcomes can be measured and fed back to users?

While this study was conducted within England and discussed within the policy context of the United Kingdom, the policy implications have international relevance. The global strategy on digital health from the World Health Organization calls for the implementation of functional and interoperable electronic health records that can contribute to informed decision-making and high-quality, person-centered care [66]. Our findings highlight factors to consider when developing electronic systems for use in the delivery of palliative and end-of-life care. The questions presented in Table 3, therefore, are likely to have relevance to policy makers and practitioners seeking to use and implement similar complex digital interventions (including electronic information-sharing systems) across multiple country and health care contexts. With increasing governmental policies on the development and implementation of health information

XSL•FO
RenderX

technologies within the United Kingdom [62], these questions can guide efforts in the context of palliative and end-of-life care.

Strengths and Limitations

A strength and novelty of this work lies in the adoption of NPT to explain the processes across different levels of the health care system which affected whether and how EPaCCS were normalized into everyday practice. Through recruiting a wide range of health care professionals, across 5 settings of care, and 2 major regions within the United Kingdom, naturalistic generalizations [67] may be made from this work. That is, the findings of this study are likely to resonate with the personal experiences of healthcare professionals who use EPaCCS and similar health information technologies across end-of-life settings. We highlight 4 study limitations. First, we struggled to recruit community nurses across London, meaning that the application of study findings to this professional group is likely to be limited. Second, this study only sought the perspectives of health care professionals. Future research should explore patients' preferences on the content, sharing, and accessibility of their electronic records, alongside the impact of such digital interventions on the patient experience and clinical outcomes. Third, in line with other studies [68] throughout analysis, we found that the technical language of NPT and the overlapping of its components made deductively coding and interpreting data using this theoretical framework challenging. Fourth, we also appreciate that constructs comprising NPT focus on specific factors that influenced the implementation of EPaCCS in routine practice. Other implementation theories, models, and frameworks (eg, the Consolidated Framework for Implementation Research [69], Promoting Action on Research Implementation in Health Services [70], Capability, Opportunity, Motivation and Behaviour Theory [71]) may have provided different, yet equally valuable insights into answering the research question.

Conclusions

The successful implementation of digital interventions into routine care depends on the extent to which they enhance established ways of working with minimal disruption. EPaCCS represents just 1 approach to the electronic sharing of advance care plans, and other forms of digital advance care planning exist internationally. Integral to the implementation of digital advance care planning systems for palliative care is ensuring they can allow health care practitioners to document, access, use, and share information across multiple care settings. There also needs to be an effort at individual, team, and organizational levels to make sure that these tools are embedded into everyday care practices. It is paramount that they are championed within and between services, and that people know when, how, and why to use them. Commissioners, health care services, and practitioners should consider these multilevel factors when planning and rolling out digital advance care planning approaches.

Acknowledgments

This study is funded by the NIHR Health and Social Care Delivery Research Programme (NIHR129171). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. MA is a University Academic Fellow, funded through a research fellowship from Yorkshire Cancer Research. We wish to acknowledge Dr Samantha Coster, our colleague at King's College London, for her support and guidance in editing this article. This manuscript is published on behalf of the wider Optimal Care project team. Members of the team who are not included as authors on this manuscript but who were coapplicants and made significant contributions to the conceptualization and design of the project include Prof Michael Bennett (University of Leeds) and Dr Samuel Relton (University of Leeds, United Kingdom). We would also like to extend specific acknowledgment and recognition of our patient and public involvement lead for the Optimal Care project, Mrs Barbara Hibbert, who supported the development of project plans, formed our patient involvement group and facilitated inclusive and insightful patient and public involvement group meetings.

Data Availability

Deidentified transcripts generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MJA is the grant holder and Chief Investigator and was responsible for study conceptualization and development of the study protocol, with critical input from grant coapplicants, SR, RF, MT, KES, CJE, PC, and PMM. AB and JB conducted interviews with study participants. Analysis was led by AB and conducted with critical input and supervision from JB, MT, and MA. All authors had access to all study data, discussed the interpretation of findings, take responsibility for data integrity and analysis, contributed to the analysis plan, and provided critical comments and contributed to the writing and development of the manuscript. All authors reviewed the final manuscript. Artificial Intelligence was not used to generate any portions of the submitted manuscript.

Conflicts of Interest

This study is funded by the NIHR Health and Social Care Delivery Research Programme (NIHR129171). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. KES is the Laing Galazka chair at King’s College London, funded through an endowment from Cicely Saunders International and the Kirby Laing

XSL•FO
RenderX

Foundation. MJA is a University Academic Fellow, funded through a research fellowship from Yorkshire Cancer Research. RF receives funding from the UK National Institute for Health and Care Research and is chair of the National Institute for Health and Care Excellence Implementation Strategy Group. We wish to acknowledge Dr Samantha Coster, our colleague at King's College London, for her support and guidance in editing this article. This manuscript is published on behalf of the wider Optimal Care project team. Members of the team who are not included as authors on this manuscript but who were coapplicants and made significant contributions to the conceptualization and design of the project include Prof Michael Bennett (University of Leeds) and Dr Samuel Relton (University of Leeds, United Kingdom). We would also like to extend specific acknowledgment and recognition of our patient and public involvement lead for the Optimal Care project, Mrs Barbara Hibbert, who supported the development of project plans, formed our patient involvement group and facilitated inclusive and insightful patient and public involvement group meetings.

Multimedia Appendix 1
COREQ (consolidated criteria for reporting qualitative research) checklist.
[DOCX File , 27 KB - jmir_v26i1e50217_app1.docx]

Multimedia Appendix 2
Interview topic guide.
[DOCX File , 82 KB - jmir_v26i1e50217_app2.docx]

References

1. Twycross R. Introducing Palliative Care. United Kingdom: Radcliffe Publishing; 2003.
2. Paparella G. Person-Centred Care in Europe: A Cross-Country Comparison of Health System Performance, Strategies and Structures. England: Picker Institute; 2016.
3. WHO global strategy on integrated people-centred health services 2016-2026: Executive Summary. Geneva, Switzerland: World Health Organisation; 2015.
4. Sleeman KE, Timms A, Gillam J, Anderson JE, Harding R, Sampson EL, et al. Priorities and opportunities for palliative and end of life care in United Kingdom health policies: a national documentary analysis. BMC Palliat Care 2021;20(1):108 [FREE Full text] [doi: 10.1186/s12904-021-00802-6] [Medline: 34261478]
5. Alderwick H, Dixon J. The NHS long term plan. BMJ 2019;364:l84 [FREE Full text] [doi: 10.1136/bmj.l84] [Medline: 30617185]
6. Sudore RL, Lum HD, You JJ, Hanson LC, Meier DE, Pantilat SZ, et al. Defining advance care planning for adults: a consensus definition from a multidisciplinary delphi panel. J Pain Symptom Manage 2017;53(5):821-832 [FREE Full text] [doi: 10.1016/j.jpainsymman.2016.12.331] [Medline: 28062339]
7. Brinkman-Stoppelenburg A, Rietjens JA, van der Heide A. The effects of advance care planning on end-of-life care: a systematic review. Palliat Med 2014;28(8):1000-1025. [doi: 10.1177/0269216314526272] [Medline: 24651708]
8. Klingler C, in der Schmitten J, Marckmann G. Does facilitated advance care planning reduce the costs of care near the end of life? Systematic review and ethical considerations. Palliat Med 2016;30(5):423-433 [FREE Full text] [doi: 10.1177/0269216315601346] [Medline: 26294218]
9. Abel J, Pring A, Rich A, Malik T, Verne J. The impact of advance care planning of place of death, a hospice retrospective cohort study. BMJ Support Palliat Care 2013;3(2):168-173 [FREE Full text] [doi: 10.1136/bmjspcare-2012-000327] [Medline: 23626905]
10. Korfage IJ, Carreras G, Arnfeldt Christensen CM, Billekens P, Bramley L, Briggs L, et al. Advance care planning in patients with advanced cancer: a 6-country, cluster-randomised clinical trial. PLoS Med 2020;17(11):e1003422 [FREE Full text] [doi: 10.1371/journal.pmed.1003422] [Medline: 33186365]
11. Malhotra C, Shafiq M, Batcagan-Abueg APM. What is the evidence for efficacy of advance care planning in improving patient outcomes? A systematic review of randomised controlled trials. BMJ Open 2022;12(7):e060201. [doi: 10.1136/bmjopen-2021-060201]
12. Rietjens J, Korfage I, van der Heide A. Advance care planning: not a panacea. Palliat Med 2016;30(5):421-422. [doi: 10.1177/0269216316642963] [Medline: 27095742]
13. Allsop MJ, Chumbley K, Birtwistle J, Bennett MI, Pocock L. Building on sand: digital technologies for care coordination and advance care planning. BMJ Support Palliat Care 2022;12(2):194-197 [FREE Full text] [doi: 10.1136/bmjspcare-2021-003304] [Medline: 34876456]
14. Birtwistle J, Millares-Martin P, Evans CJ, Foy R, Relton S, Richards S, et al. Mapping and characterising electronic palliative care coordination systems and their intended impact: a national survey of end-of-life care commissioners. PLoS One 2022;17(10):e0275991 [FREE Full text] [doi: 10.1371/journal.pone.0275991] [Medline: 36240254]
15. Huber MT, Highland JD, Krishnamoorthi VR, Tang JW. Utilizing the electronic health record to improve advance care planning: a systematic review. Am J Hosp Palliat Care 2018;35(3):532-541. [doi: 10.1177/1049909117715217] [Medline: 28627287]

XSL•FO
RenderX

16. Lamas D, Panariello N, Henrich N, Hammes B, Hanson LC, Meier DE, et al. Advance care planning documentation in electronic health records: current challenges and recommendations for change. J Palliat Med 2018;21(4):522-528. [doi: 10.1089/jpm.2017.0451] [Medline: 29360417]
17. McCarthy S, Meredith J, Bryant L, Hemsley B. Legal and Ethical Issues Surrounding Advance Care Directives in Australia: Implications for the Advance Care Planning Document in the Australian My Health Record. J Law Med 2017;25(1):136-149. [Medline: 29978629]
18. End of Life Care Strategy promoting high quality care for all adults at the end of life. UK: Department of Health; 2008.
19. National Palliative and End of Life Care Partnership, Ambitions for Palliative and End of Life Care: A National Framework for Local Action 2021-2026. 2021. URL: https://www.england.nhs.uk/wp-content/uploads/2022/02/ambitions-for-palliative-and-end-of-life-care-2nd-edition.pdf [accessed 2024-06-18]
20. Our Commitment to you for end of life care: the Government Response to the Review of Choice in End of Life Care. London: Department of Health; 2016.
21. den Herder-van der Eerden M, van Wijngaarden J, Payne S, Preston N, Linge-Dahl L, Radbruch L, et al. Integrated palliative care is about professional networking rather than standardisation of care: a qualitative study with healthcare professionals in 19 integrated palliative care initiatives in five European countries. Palliat Med 2018;32(6):1091-1102 [FREE Full text] [doi: 10.1177/0269216318758194] [Medline: 29436279]
22. Hsiao CJ, King J, Hing E, Simon AE. The role of health information technology in care coordination in the United States. Med Care 2015;53(2):184-190 [FREE Full text] [doi: 10.1097/MLR.0000000000000276] [Medline: 25464164]
23. Beernaert K, Cohen J, Deliens L, Devroey D, Vanthomme K, Pardon K, et al. Referral to palliative care in COPD and other chronic diseases: a population-based study. Respir Med 2013;107(11):1731-1739 [FREE Full text] [doi: 10.1016/j.rmed.2013.06.003] [Medline: 23810150]
24. Billingham MJ, Billingham S. Congruence between preferred and actual place of death according to the presence of malignant or non-malignant disease: a systematic review and meta-analysis. BMJ Support Palliat Care 2013;3(2):144-154. [doi: 10.1136/bmjspcare-2012-000292] [Medline: 24644562]
25. Ko W, Deliens L, Miccinesi G, Giusti F, Moreels S, Donker GA, EURO IMPACT. Care provided and care setting transitions in the last three months of life of cancer patients: a nationwide monitoring study in four European countries. BMC Cancer 2014;14:960 [FREE Full text] [doi: 10.1186/1471-2407-14-960] [Medline: 25510507]
26. Smith C, Hough L, Cheung C, Millington-Sanders C, Sutton E, Ross JR, et al. Coordinate my care: a clinical service that coordinates care, giving patients choice and improving quality of life. BMJ Support Palliat Care 2012;2(4):301-307. [doi: 10.1136/bmjspcare-2012-000265] [Medline: 24654211]
27. Allsop MJ, Kite S, McDermott S, Penn N, Millares-Martin P, Bennett MI. Electronic palliative care coordination systems: devising and testing a methodology for evaluating documentation. Palliat Med 2017;31(5):475-482 [FREE Full text] [doi: 10.1177/0269216316663881] [Medline: 27507636]
28. Finucane AM, Davydaitis D, Horseman Z, Carduff E, Baughan P, Tapsfield J, et al. Electronic care coordination systems for people with advanced progressive illness: a mixed-methods evaluation in Scottish primary care. Br J Gen Pract 2020;70(690):e20-e28 [FREE Full text] [doi: 10.3399/bjgp19X707117] [Medline: 31848198]
29. Palliative and end of life care information standard. Professional Record Standards Body. 2022. URL: https://prsb2.vercel.app/page/palliative-and-end-of-life-care-information-standard?hsCtaTracking=f03fff77-1ecb-44e3-955b-a56f038c83dd%7C52d83f23-deb4-48a6-99ef-5bd893ec26b6 [accessed 2024-06-18]
30. Standing H, Patterson R, Lee M, Dalkin SM, Lhussier M, Bate A, et al. Information sharing challenges in end-of-life care: a qualitative study of patient, family and professional perspectives on the potential of an electronic palliative care Co-ordination system. BMJ Open 2020;10(10):e037483 [FREE Full text] [doi: 10.1136/bmjopen-2020-037483] [Medline: 33020093]
31. Patterson R, Standing H, Lee M, Dalkin S, Lhussier M, Exley C, et al. Paramedic information needs in end-of-life care: a qualitative interview study exploring access to a shared electronic record as a potential solution. BMC Palliat Care 2019;18(1):108 [FREE Full text] [doi: 10.1186/s12904-019-0498-2] [Medline: 31805917]
32. Electronic palliative care co-ordination systems (EPaCCS). Public Health England. 2013. URL: https://www.gov.uk/government/publications/electronic-palliative-care-co-ordination-systems-epaccs [accessed 2024-06-18]
33. Millares Martin P. Electronic palliative care coordination system (EPaCCS): interoperability is a problem. BMJ Support Palliat Care 2018;8(3):358-359. [doi: 10.1136/bmjspcare-2017-001396] [Medline: 28838930]
34. Pocock L, Morris R, French L, Purdy S. Underutilisation of EPaCCS (electronic palliative care coordination systems) in end-of life-care: a cross-sectional study. BMJ Support Palliat Care 2021:bmjspcare-2020-002798. [doi: 10.1136/bmjspcare-2020-002798] [Medline: 33837112]
35. Wye L, Lasseter G, Simmonds B, Duncan L, Percival J, Purdy S. Electronic palliative care coordinating systems (EPaCCS) may not facilitate home deaths: A mixed methods evaluation of end of life care in two English counties. Journal of Research in Nursing 2016;21(2):96-107. [doi: 10.1177/1744987116628922]
36. Petrova M, Riley J, Abel J, Barclay S. Crash course in EPaCCS (electronic palliative care coordination systems): 8 years of successes and failures in patient data sharing to learn from. BMJ Support Palliat Care 2018;8(4):447-455 [FREE Full text] [doi: 10.1136/bmjspcare-2015-001059] [Medline: 27638631]

XSL•FO
RenderX

37. Leniz J, Weil A, Higginson IJ, Sleeman KE. Electronic palliative care coordination systems (EPaCCS): a systematic review. BMJ Support Palliat Care 2020;10(1):68-78 [FREE Full text] [doi: 10.1136/bmjspcare-2018-001689] [Medline: 31068334]
38. Stebbins RA. Exploratory research in the social sciences. London: Sage Publications; 2001.
39. Bryman A. Social research methods. England: Oxford: Oxford University Press; 2004.
40. Burke S. Rethinking 'validity' and 'trustworthiness' in qualitative inquiry: How might we judge the quality of qualitative research in sport and exercise sciences?, in Routledge Handbook of Qualitative Research in Sport and Exercise. London: Routledge; 2016:330-339.
41. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
42. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of medical research council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
43. May CR, Mair F, Finch T, MacFarlane A, Dowrick C, Treweek S, et al. Development of a theory of implementation and integration: normalization process theory. Implement Sci 2009;4:29 [FREE Full text] [doi: 10.1186/1748-5908-4-29] [Medline: 19460163]
44. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med 2010;8:63 [FREE Full text] [doi: 10.1186/1741-7015-8-63] [Medline: 20961442]
45. May C. Towards a general theory of implementation. Implement Sci 2013;8:18 [FREE Full text] [doi: 10.1186/1748-5908-8-18] [Medline: 23406398]
46. May CR, Albers B, Bracher M, Finch TL, Gilbert A, Girling M, et al. Translational framework for implementation evaluation and research: a normalisation process theory coding manual for qualitative research and instrument development. Implement Sci 2022;17(1):19 [FREE Full text] [doi: 10.1186/s13012-022-01191-x] [Medline: 35193611]
47. McEvoy R, Ballini L, Maltoni S, O'Donnell CA, Mair FS, Macfarlane A. A qualitative systematic review of studies using the normalization process theory to research implementation processes. Implement Sci 2014;9:2 [FREE Full text] [doi: 10.1186/1748-5908-9-2] [Medline: 24383661]
48. May CR, Finch T, Ballini L, MacFarlane A, Mair F, Murray E, et al. Evaluating complex interventions and health technologies using normalization process theory: development of a simplified approach and web-enabled toolkit. BMC Health Serv Res 2011;11:245 [FREE Full text] [doi: 10.1186/1472-6963-11-245] [Medline: 21961827]
49. Etikan I, Musa SA, Alkassim RS. Comparison of convenience sampling and purposive sampling. AJTAS 2016;5(1):1. [doi: 10.11648/j.ajtas.20160501.11]
50. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
51. NVIVO. 2020. URL: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/support-services/nvivo-downloads/older-versions [accessed 2024-06-18]
52. Ritchie L, Lewis J, Nicholls CM, Ormston R. Qualitative research practice: a guide for social science students and researchers. California: Los Angeles, CA: Sage; 2003.
53. Finlay L. Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. Qualitative Research 2016;2(2):209-230. [doi: 10.1177/146879410200200205]
54. Smith B, McGannon KR. Developing rigor in qualitative research: problems and opportunities within sport and exercise psychology. International Review of Sport and Exercise Psychology 2017;11(1):101-121. [doi: 10.1080/1750984x.2017.1317357]
55. Sparkes A, Smith B. Qualitative Research Methods in Sport, Exercise & Health From Process to Product. London: Routledge; 2014.
56. Papagiannidis S. TheoryHub book: A theory resource for students and researchers alike. UK: Theory for Explaining; 2022.
57. Bernacki RE, Block SD, American College of Physicians High Value Care Task Force. Communication about serious illness care goals: a review and synthesis of best practices. JAMA Intern Med 2014;174(12):1994-2003. [doi: 10.1001/jamainternmed.2014.5271] [Medline: 25330167]
58. De Vleminck A, Houttekier D, Pardon K, Deschepper R, Van Audenhove C, Vander Stichele R, et al. Barriers and facilitators for general practitioners to engage in advance care planning: a systematic review. Scand J Prim Health Care 2013;31(4):215-226 [FREE Full text] [doi: 10.3109/02813432.2013.854590] [Medline: 24299046]
59. Jabbarian LJ, Zwakman M, van der Heide A, Kars MC, Janssen DJA, van Delden JJ, et al. Advance care planning for patients with chronic respiratory diseases: a systematic review of preferences and practices. Thorax 2018;73(3):222-230. [doi: 10.1136/thoraxjnl-2016-209806] [Medline: 29109233]
60. Bradshaw A, Ostler S, Goodman C, Batkovskyte I, Ellis-Smith C, Tunnard I, et al. Provision of palliative and end-of-life care in UK care homes during the Covid-19 pandemic: a mixed methods observational study with implications for policy. Front Public Health 2023;11:1058736 [FREE Full text] [doi: 10.3389/fpubh.2023.1058736] [Medline: 36998280]

XSL•FO
RenderX

61. Bone AE, Gomes B, Etkind SN, Verne J, Murtagh FE, Evans CJ, et al. What is the impact of population ageing on the future provision of end-of-life care? Population-based projections of place of death. Palliat Med 2018;32(2):329-336 [FREE Full text] [doi: 10.1177/0269216317734435] [Medline: 29017018]
62. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. Lancet Digit Health 2021;3(6):e383-e396 [FREE Full text] [doi: 10.1016/S2589-7500(21)00005-4] [Medline: 33967002]
63. Dorr D, Bonner LM, Cohen AN, Shoai RS, Perrin R, Chaney E, et al. Informatics systems to promote improved care for chronic illness: a literature review. J Am Med Inform Assoc 2007;14(2):156-163 [FREE Full text] [doi: 10.1197/jamia.M2255] [Medline: 17213491]
64. Hall S, Murchie P, Campbell C, Murray SA. Introducing an electronic palliative care summary (ePCS) in Scotland: patient, carer and professional perspectives. Fam Pract 2012;29(5):576-585. [doi: 10.1093/fampra/cms011] [Medline: 22337868]
65. Carlfjord S, Lindberg M, Bendtsen P, Nilsen P, Andersson A. Key factors influencing adoption of an innovation in primary health care: a qualitative study based on implementation theory. BMC Fam Pract 2010;11:60 [FREE Full text] [doi: 10.1186/1471-2296-11-60] [Medline: 20731817]
66. Global strategy on digital health 2020-2025. California: World Health Organisation; 2020.
67. Smith B. Generalizability in qualitative research: misunderstandings, opportunities and recommendations for the sport and exercise sciences. Qualitative Research in Sport, Exercise and Health 2017;10(1):137-149. [doi: 10.1080/2159676x.2017.1393221]
68. May CR, Cummings A, Girling M, Bracher M, Mair FS, May CM, et al. Using normalization process theory in feasibility studies and process evaluations of complex healthcare interventions: a systematic review. Implement Sci 2018;13(1):80 [FREE Full text] [doi: 10.1186/s13012-018-0758-1] [Medline: 29879986]
69. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
70. Rycroft-Malone J. The PARIHS framework--a framework for guiding the implementation of evidence-based practice. J Nurs Care Qual 2004;19(4):297-304. [doi: 10.1097/00001786-200410000-00002] [Medline: 15535533]
71. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]

Abbreviations

CMC: Coordinate My Care
COREQ: consolidated criteria for reporting qualitative research
EPaCCS: Electronic Palliative Care Coordination Systems
NPT: Normalization process theory

Edited by A Mavragani; submitted 23.06.23; peer-reviewed by C Laranjeira, T Phenwan; comments to author 14.09.23; revised version received 18.10.23; accepted 30.05.24; published 16.08.24.

Please cite as:
Bradshaw A, Birtwistle J, Evans CJ, Sleeman KE, Richards S, Foy R, Millares Martin P, Carder P, Allsop MJ, Twiddy M
Factors Influencing the Implementation of Digital Advance Care Planning: Qualitative Interview Study
J Med Internet Res 2024;26:e50217
URL: https://www.jmir.org/2024/1/e50217
doi:10.2196/50217
PMID:

©Andy Bradshaw, Jacqueline Birtwistle, Catherine J Evans, Katherine E Sleeman, Suzanne Richards, Robbie Foy, Pablo Millares Martin, Paul Carder, Matthew J Allsop, Maureen Twiddy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation

Xuan Zou[1*], MM; Weijie He[2,3,4*], PhD; Yu Huang[5*], ME; Yi Ouyang[5], PhD; Zhen Zhang[1], MM; Yu Wu[1], PhD; Yongsheng Wu[1], BMed; Lili Feng[6], PhD; Sheng Wu[6], BMed; Mengqi Yang[7], BE; Xuyan Chen[6], PhD; Yefeng Zheng[5], PhD; Rui Jiang[4,8], PhD; Ting Chen[2,3,4], PhD

[1]Shenzhen Center for Disease Control and Prevention, Shenzhen, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Institute of Artificial Intelligence, Tsinghua University, Beijing, China
[4]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
[5]Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[6]Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China
[7]Tencent Healthcare, Shenzhen, China
[8]Department of Automation, Tsinghua University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Ting Chen, PhD
Department of Computer Science and Technology
Tsinghua University
Room 3-609, Future Internet Technology Research Center
Tsinghua University
Beijing, 100084
China
Phone: 86 010 62797101
Email: tingchen@tsinghua.edu.cn

Abstract

Background: For medical diagnosis, clinicians typically begin with a patient's chief concerns, followed by questions about symptoms and medical history, physical examinations, and requests for necessary auxiliary examinations to gather comprehensive medical information. This complex medical investigation process has yet to be modeled by existing artificial intelligence (AI) methodologies.

Objective: The aim of this study was to develop an AI-driven medical inquiry assistant for clinical diagnosis that provides inquiry recommendations by simulating clinicians' medical investigating logic via reinforcement learning.

Methods: We compiled multicenter, deidentified outpatient electronic health records from 76 hospitals in Shenzhen, China, spanning the period from July to November 2021. These records consisted of both unstructured textual information and structured laboratory test results. We first performed feature extraction and standardization using natural language processing techniques and then used a reinforcement learning actor-critic framework to explore the rational and effective inquiry logic. To align the inquiry process with actual clinical practice, we segmented the inquiry into 4 stages: inquiring about symptoms and medical history, conducting physical examinations, requesting auxiliary examinations, and terminating the inquiry with a diagnosis. External validation was conducted to validate the inquiry logic of the AI model.

Results: This study focused on 2 retrospective inquiry-and-diagnosis tasks in the emergency and pediatrics departments. The emergency departments provided records of 339,020 consultations including mainly children (median age 5.2, IQR 2.6-26.1 years) with various types of upper respiratory tract infections (250,638/339,020, 73.93%). The pediatrics department provided records of 561,659 consultations, mainly of children (median age 3.8, IQR 2.0-5.7 years) with various types of upper respiratory tract infections (498,408/561,659, 88.73%). When conducting its own inquiries in both scenarios, the AI model demonstrated high diagnostic performance, with areas under the receiver operating characteristic curve of 0.955 (95% CI 0.953-0.956) and 0.943 (95% CI 0.941-0.944), respectively. When the AI model was used in a simulated collaboration with physicians, it notably reduced the average number of physicians' inquiries to 46% (6.037/13.26; 95% CI 6.009-6.064) and 43% (6.245/14.364; 95%

XSL•FO
RenderX

CI 6.225-6.269) while achieving areas under the receiver operating characteristic curve of 0.972 (95% CI 0.970-0.973) and 0.968 (95% CI 0.967-0.969) in the scenarios. External validation revealed a normalized Kendall τ distance of 0.323 (95% CI 0.301-0.346), indicating the inquiry consistency of the AI model with physicians.

Conclusions: This retrospective analysis of predominantly respiratory pediatric presentations in emergency and pediatrics departments demonstrated that an AI-driven diagnostic assistant had high diagnostic performance both in stand-alone use and in simulated collaboration with clinicians. Its investigation process was found to be consistent with the clinicians' medical investigation logic. These findings highlight the diagnostic assistant's promise in assisting the decision-making processes of health care professionals.

(J Med Internet Res 2024;26:e54616) doi:10.2196/54616

KEYWORDS

inquiry and diagnosis; electronic health record; reinforcement learning; natural language processing; artificial intelligence

Introduction

Background

The growing demand for intelligent clinical decision support systems (CDSSs) has become increasingly evident in the health care landscape today [1,2]. The surge in demand can be attributed to advances in medical knowledge, the accumulation of health care data, and the rapid progression of artificial intelligence (AI) and machine learning technologies. Traditionally, health care professionals have relied heavily on their individual experiences and medical expertise to make clinical decisions, which are often susceptible to subjective biases and information gaps [3,4]. Furthermore, the world faces challenges related to insufficient medical resources, particularly in regions where a shortage of health care professionals impedes timely access to care. Consequently, intelligent CDSSs have been developed with the potential to enhance patient care quality, reduce medical costs, and minimize diagnostic errors. These systems were developed by analyzing and training data from electronic health records (EHRs) and medical literature [5,6] using the power of extensive data analysis, machine learning algorithms, and natural language processing techniques.

Our research is focused on advancing CDSSs with a primary emphasis on aiding the diagnostic process before the diagnosis is made, specifically during the information-gathering phase. Existing diagnostic support systems often rely on comprehensive patient information, such as EHRs and medical images [7-13], to provide diagnostic predictions only at the final step of the medical investigation process. Thus, the vital need for decision support during the intermediate steps of the diagnostic process is largely overlooked. The clinical diagnostic process is a complex procedure that involves a series of inquiries and examinations. In this dynamic process, health care professionals continuously gather information, adjust hypotheses, and refine diagnostic reasoning until the optimal diagnosis and treatment are determined. The process is not merely a sequence of isolated steps but a dynamic and evolving interaction between the clinician and the patient. Conventional diagnostic support systems tend to provide little to no guidance during these intermediate steps, falling short of providing meaningful support when it is most needed.

A variety of automatic disease diagnosis techniques have emerged recently as a result of the advancement of AI with the goal of assisting in the middle stages of clinical decision-making [14-21]. However, a common limitation among these methods is their exclusive focus on online disease diagnosis. Online health care platforms have undoubtedly revolutionized medical consultations by facilitating remote access to health care professionals. Patients can seek advice and preliminary assessments for various ailments without visiting a health care facility physically. Unfortunately, there are inherent constraints to these platforms. Online consultations are primarily reliant on textual descriptions, making it challenging to gather essential information that requires palpation, auscultation, or specialized laboratory investigations, which collectively form the foundation of a comprehensive and accurate diagnosis.

Large language models (LLMs) such as ChatGPT [22] have demonstrated remarkable proficiency in following instructions and generating humanlike responses across various domains. LLMs have already attracted a lot of attention regarding their potential applications in health care settings, which include facilitating clinical documentation, summarizing research papers, assisting with medical education [23], or acting as a chatbot to respond to questions from patients about their specific concerns [24,25]. While LLMs have access to extremely large corpora containing medical dialogues, knowledge bases, and the literature, there is a significant gap in access to large-scale real patient records [26]. This lack of actual clinical data significantly impacts their ability to provide tailored inquiry recommendations for patients in varying clinical scenarios. In addition, ethical [24,25] and security [27] concerns further complicate the incorporation of sensitive health records into the training of LLMs.

Objectives

To address the gap in diagnostic decision support, we present MedRIA, a medical reinforcement learning inquiry assistant. MedRIA is designed to facilitate the medical investigation process by intelligently guiding inquiries; symptom assessments; examination recommendations; and, in particular, the transitions between them. To the best of our knowledge, no previous system has successfully managed the entire diagnostic process. MedRIA learns its inquiry capabilities from a large number of processed EHRs. MedRIA uses a reinforcement learning actor-critic framework [28] to explore rational and effective inquiry logic. Following the standard inquiry process of physicians [29], we segmented the inquiry into 4 stages: inquiring about symptoms and medical history, conducting physical examinations (PEs),

XSL•FO
RenderX

requesting auxiliary examinations (AEs), and terminating the inquiry with a diagnosis. MedRIA's ability to tailor inquiry strategies according to patient-specific conditions positions it as a promising solution for regions or populations facing constraints in medical resources.

Methods

Processed Outpatient EHRs

The Shenzhen Center for Disease Control and Prevention in Guangdong Province, China, collected deidentified outpatient EHRs from 76 hospitals in Shenzhen. We selected records of the emergency and pediatrics departments from July 2021 to November 2021. As depicted in Figure 1, these records consist of both unstructured textual information and structured laboratory test results, which underwent feature extraction and standardization using natural language processing techniques. An example of a processed EHR is provided on the right side of Figure 1.

Figure 1. Schematic illustration of the data collection, data processing, and task building process for the development of MedRIA. ADHD: attention-deficit/hyperactivity disorder; EHR: electronic health record.

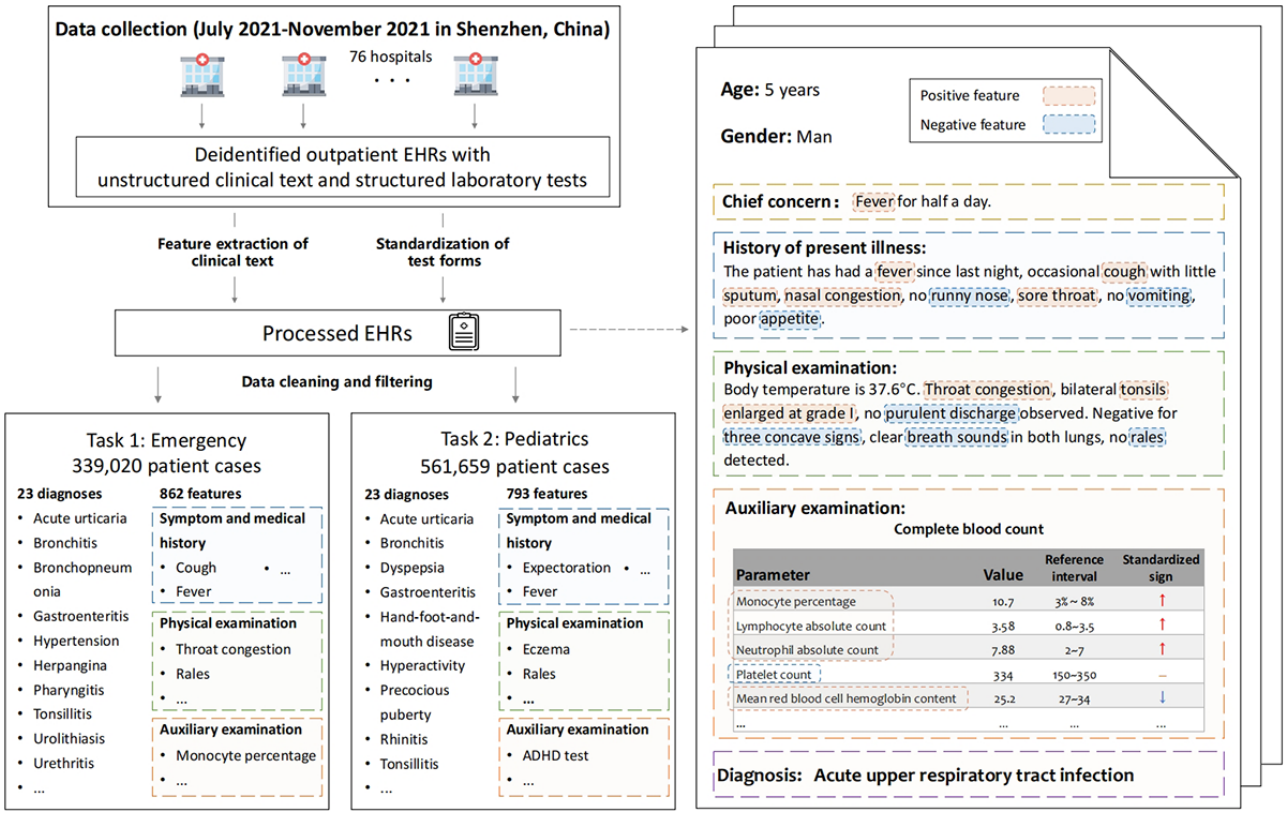

By combining established medical terminologies with domain-specific knowledge from health care professionals, we constructed a comprehensive feature set for the feature extraction process. Initially, we used existing Chinese versions of general medical terminologies, including the *International Classification of Diseases, 10th Revision*, and Systematized Nomenclature of Medicine–Clinical Terms. In addition, we incorporated domain expertise by consulting with health care experts to identify relevant clinical concepts and terminology to the context of our study. The feature sets of the emergency and pediatrics EHRs are presented in Multimedia Appendices 1 and 2, respectively.

As depicted in Figure 1, after extracting features from clinical text and standardizing test forms, each record was converted into a feature set, where each feature was categorized into 1 of 3 types based on its source—symptom and medical history (SMH), PE, and AE—and assigned a value based on its data type. Symptom features (eg, *cough*) and some features observed from examinations (eg, *throat congestion* and *proteinuria*) were assigned binary values indicating whether the patient exhibited that feature. Numerical laboratory test features with specific reference ranges, such as *platelet count*, were categorized as high, low, or within the reference range. The values of descriptive features depend on their meanings. For example, *urine color* was given values such as *yellow*, *colorless*, *white*, *red*, and so on.

In our study, we used the Medical Bidirectional Encoder Representations From Transformers (MedBERT) model for feature extraction on clinical text, which was based on Bidirectional Encoder Representations From Transformers [30]. Using a vast corpus of Chinese clinical text data, including medical textbooks, online consultations, journal article abstracts, and deidentified EHRs, we pretrained the Bidirectional Encoder Representations From Transformers model to obtain MedBERT. We manually constructed 2 labeled data sets to fine-tune MedBERT to perform medical named entity recognition and entity normalization. In our pipeline, we first used fine-tuned MedBERT for medical named entity recognition on clinical

XSL•FO
RenderX

text. Next, we proceeded to normalize the identified entities to our predefined feature set using the fine-tuned MedBERT for entity normalization. Finally, we used a string-matching negation detection mechanism to determine the values of binary features based on the contents of the text. For structured laboratory test results, we used a string-matching algorithm to standardize the test items to our predefined feature set. The feature values could be directly obtained from the structured results.

Following the feature extraction process, we conducted data cleaning and filtering. Details are presented in Multimedia Appendix 3 [31-37]. Finally, we filtered the data sets by department, creating 2 retrospective inquiry-and-diagnosis evaluation tasks for emergency and pediatrics. Emergency and pediatrics are the 2 busiest departments in hospitals and often put tremendous pressure on health care personnel. These departments require effective decision support tools to assist health care professionals in managing the workload efficiently while providing timely and accurate patient care.

Reinforcement Learning Formulation

MedRIA takes as initial input the patient's basic demographic information, including gender and age, along with features extracted from the patient's chief concerns. Subsequently, MedRIA provides recommendations for the physician to inquire about specific features. As the physician conducts the inquiry and gathers patient information, MedRIA updates the patient's status and continues to output the next relevant feature or suggests terminating the inquiry and providing a predicted diagnosis in terms of the probability distribution over potential disease diagnoses. This can be described as a Markov decision process [38], represented by the tuple $M(S, A, T, R, \gamma)$. Here, S (state) is a set of vectors representing the inquiry states that incorporate both observed and unobserved features so far, denoted as x_O and x_U, respectively. A (action) involves the actions available to MedRIA, including recommending specific features or suggesting the termination of the inquiry process. The goal is to determine the optimal inquiry strategy composed of a series of actions, denoted as π_θ. At each step t, the policy $\pi_\theta(a_t|s_t)$ determines the action a_t based on the current state s_t. The transition function T represents the probability distribution of the next state s_{t+1} given the current state s_t and action a_t, denoted as $p(s_{t+1}|s_t, a_t)$. R is a reward function used to assess the benefit of the transition (s_t, a_t, s_{t+1}). γ is the discount factor for accumulating rewards at each step. The state-value function that reinforcement learning maximizes is the expected sum of discounted rewards given the policy π_θ and the state s_t.

Actor-Critic Framework

MedRIA uses the classic reinforcement learning actor-critic framework [28] involving 2 neural networks: the actor and the critic. The actor network determines the next action to take, whereas the critic network is responsible for providing feedback to the actor, evaluating the quality of actions and suggesting adjustments. On the basis of our previous work [39], we used a pretrained variational autoencoder (VAE) [31] as the backbone of the actor network. We used a supervised diagnostic prediction model, denoted as D, to provide the current disease prediction probability distribution $D(s_t)$ based on the state s_t to help decide on each inquiry action.

The actor of MedRIA makes decisions for the next action based on the observed features x_O. It is intuitive to prioritize asking about feature x that has a high conditional probability $p(x|x_O)$. We leveraged the VAE to incorporate the conditional probability distribution between observed features x_O and unobserved features x_U. A VAE defines a generative model of the form $p(x, z)=\prod_i p_\Theta(x_i|z)p(z)$, where features x are generated from latent variables z. The number of dimensions of z is 64. p(z) represents a prior, often a spherical Gaussian distribution. $p_\Theta(x|z)$ is represented by a 4-layer multilayer perceptron (MLP) decoder with parameters Θ. The VAE uses another neural network encoder with parameters Φ to generate the variational approximation of the posterior $q_\Phi(z|x)$. To obtain $p(x_U|x_O)$, we sampled from the VAE encoder as $\sim q_\Phi(z|x)$ and then sampled x_U from the VAE decoder as $p_\Theta(x_U|\ )$. Let e_i represent the embedding vector for the ith observed feature x_i, and let $c_i=[x_i, e_i]$ denote the concatenated input carrying information for x_i. The number of dimensions of e_i is 64. Then, we used a 4-layer MLP to map the input c_i to a Gaussian distribution in latent space, with mean vector μ_i and variance vector V_i. To address arbitrary partial observations of features during the inquiry process, we used a product-of-experts mechanism [40] to calculate the approximate posterior. The VAE was incorporated within the actor network to leverage the conditional probability distribution $p(x_U|x_O)$. To obtain the action a_t, x_O were first fed into the nested VAE, yielding decoded features that contain predictive information regarding x_U. The decoded features were then concatenated with the current diagnostic confidence $D(s_t)$ and the representation of the current timestep n_t. Here, $n_t=t / N_T$ represents the ratio of the number of completed inquiries t over a predefined maximum action count N_T. N_T was set to 21 and 26 for the emergency and pediatrics tasks, respectively. Finally, a 2-layer MLP with softmax activation was used to map the concatenated vector to the action space.

The objective of the critic is to estimate the state-value function to optimize the policy π_θ. Similar to the approach used in the actor, we concatenated the current state s_t with $D(s_t)$ and n_t to form an input vector to aid in estimation. In addition, we obtained the informative latent variable z_t by feeding observed features x_O into the VAE encoder and appended z_t to the input vector. Ultimately, a 5-layer MLP was used to map the input vector to predict the state value.

Reward Shaping

The reward function for evaluating the gain in state transition plays a crucial role in the reinforcement learning process. We start by defining the short-term reward function R_{short} when the action is to inquire about a specific feature x. For MedRIA, the inquiry states can be categorized into 4 ordered stages: SMH (S_{SMH}), PE (S_{PE}), AE (S_{AE}), and termination with diagnosis (S_{TD}). During the inquiry process, there can be multiple states that fall under each stage. For example, the initial state of

XSL•FO
RenderX

MedRIA falls under S_{SMH}, followed by a sequence of states that fall under S_{PE}, and so on. Correspondingly, all features are categorized according to the 3 inquiry stages S_{SMH}, S_{PE}, and S_{AE}. When MedRIA selects a feature that falls under a particular inquiry stage, it signifies a transition from the inquiry stage to that stage. We ensure that MedRIA's inquiry process adheres to the sequential nature of clinical inquiries, where features from previous stages are not considered for selection once the inquiry progresses to a subsequent stage. This is achieved by manually setting the probabilities of actions associated with features from previous stages to 0 in the action probability distribution of the actor.

To measure the consistency of MedRIA's inquiries with those made by physicians, we introduce 3 functions: $M_{SMH}(s_t)$, $M_{PE}(s_t)$, and $M_{AE}(s_t)$. These functions quantify the number of features that physicians have inquired about in the current state s_t but that MedRIA has not investigated. Specifically, $M_{SMH}(s_t)$, $M_{PE}(s_t)$, and $M_{AE}(s_t)$ quantify the number of SMH features, PE features, and AE features, respectively. When calculating $M_{AE}(s_t)$, all features within a laboratory test are considered as a unit. Essentially, $M_{AE}(s_t)$ calculates the number of AEs that physicians have recommended in the current state s_t but that MedRIA has omitted. Correspondingly, when MedRIA selects a feature associated with an AE, it indicates that MedRIA has recommended the entire examination.

To assess the diagnostic quality of inquiries, we define a function $\mathrm{Diff}(s_t, x, s_{t+1})$ to estimate the differential effect of querying the feature x in the transition from state s_t to state s_{t+1}. If x belongs to the extracted features, meaning that the physician also inquired about that feature from the patient, then $\mathrm{Diff}(s_t, x, s_{t+1})$ is defined as $\mathrm{Diff}(s_t, x, s_{t+1})=|D_{KL}[y||D(s_t)]-D_{KL}[y||D(s_{t+1})]|$, where D_{KL} is the Kullback-Leibler divergence between 2 distributions and y is a one-hot vector indicating the final diagnosis made by the physician. The absolute difference between these 2 Kullback-Leibler divergences represents the degree to which the feature x influences the diagnosis results, with the final diagnosis made by the physician as the reference. If x does not belong to the extracted features, then $\mathrm{Diff}(s_t, x, s_{t+1})$ is simply set to 0. Subsequently, we define the short-term reward function $R_{short}(s_t, x, s_{t+1})$ as follows: when s_t belongs to S_{SMH} and s_{t+1} belongs to S_{PE}, $R_{short}(s_t, x, s_{t+1})=\mathrm{Diff}(s_t, x, s_{t+1})+\alpha\cdot I_{phy}(x)-m\cdot M_{SMH}(s_t)$; when s_t belongs to S_{SMH} and s_{t+1} belongs to S_{AE}, $R_{short}(s_t, x, s_{t+1})=\mathrm{Diff}(s_t, x, s_{t+1})+\alpha\cdot I_{phy}(x)-m\cdot(M_{SMH}(s_t)+M_{PE}(s_t))$; when s_t belongs to S_{PE} and s_{t+1} belongs to S_{AE}, $R_{short}(s_t, x, s_{t+1})=\mathrm{Diff}(s_t, x, s_{t+1})+\alpha\cdot I_{phy}(x)-m\cdot M_{AE}(s_t)$; otherwise, $R_{short}(s_t, x, s_{t+1})=\mathrm{Diff}(s_t, x, s_{t+1})+\alpha\cdot I_{phy}(x)$. $I_{phy}(x)$ is an indicator representing whether feature x belongs to the extracted features. If yes, it takes the value of 1; otherwise, it is 0. α and m are small constant factors to balance terms in the expression. We set α to 0.2. m was set to 0.2 in the emergency task and 0.3 in the pediatrics task. In R_{short}, we penalize each hasty stage transition based on the number of features that the actual physician inquired about from the patient in the previous stage but were not inquired about by MedRIA.

Next, we elaborate on the definition of the long-term reward R_{long} for the action of terminating the inquiry. When MedRIA selects the action to stop the inquiry process, if the disease predicted by the diagnostic model *D* matches the physician's diagnosis, R_{long} is set to 2 in the emergency task and 3 in the pediatrics task. If it does not match, similar to the penalty of inquiry consistency in R_{short}, we define $R_{long}(s_t, x, s_{t+1})$ as follows: when s_t belongs to S_{SMH}, $R_{long}(s_t, x, s_{t+1})=-m\cdot(M_{SMH}(s_t)+M_{PE}(s_t)+M_{AE}(s_t))$; when s_t belongs to S_{PE}, $R_{long}(s_t, x, s_{t+1})=-m\cdot(M_{PE}(s_t)+M_{AE}(s_t))$; when s_t belongs to S_{AE}, $R_{long}(s_t, x, s_{t+1})=-m\cdot M_{AE}(s_t)$.

Training Details

During the training process of MedRIA, we first trained a VAE. In particular, we simulated the arbitrary partial observations during the inquiry process by randomly dropping a portion of input features. Next, we trained a diagnostic prediction model *D* based on a 5-layer MLP using the softmax function and cross-entropy loss function. Then, we initialized the parameters of the nested VAE in the actor network of MedRIA with the parameters of the trained VAE. Only the nested VAE decoder was fine-tuned in the follow-up training. We used the proximal policy optimization [32] algorithm to train both the actor and critic networks. To make the predicted diagnostic probability distribution more accurate when dealing with the partially observed features generated by MedRIA, we collected the observed features x_O when the actor chose to terminate the inquiry process during each training epoch. At the end of each epoch, collected data were used to fine-tune *D* to better adapt to MedRIA's inquiry patterns. More details on implementation are presented in Multimedia Appendix 3.

External Validation

We used data from the medical dialogue corpus, IMCS-21 [41], to assess MedRIA's inquiry logic. IMCS-21 contains 4116 online medical consultations between physicians and patients covering 10 pediatric diseases. After data filtering, we selected 950 consultations with clear diagnoses that overlapped with 4 diseases covered in our pediatrics task: bronchitis (286 consultations), acute upper respiratory tract infection (392 consultations), indigestion (233 consultations), and bronchopneumonia (39 consultations). This data set served as an external test set for the trained pediatric MedRIA. We constructed ordered inquiry sequences based on the dialogue contents. These inquiry sequences reflected the typical order of feature queries in the diagnostic process, and they were used to evaluate MedRIA's inquiry sequences. If MedRIA's inquiry sequences are similar to those of physicians, we can conclude that MedRIA's inquiry logic is similar to that of physicians. The similarity is measured by the Kendall τ distance, which treats an inquiry sequence as a permutation of features and quantifies the pairwise disagreements between 2 permutations. In addition, we applied the normalized Kendall τ distance, which scales the Kendall τ distance by the total number of possible feature pairs.

XSL•FO
RenderX

Statistical Analysis

Nonparametric bootstrap sampling was used to calculate a 95% CI. Specifically, we repeatedly drew 1000 bootstrap samples from the test set. Each bootstrap sample was obtained through random sampling with replacement and has the same size as the test set.

Ethical Considerations

All EHR data used in this study were retrospectively collected from EHR systems sourced from routine clinical practice. To ensure patient privacy and confidentiality, all data were deidentified. This study was approved by the ethics committees of the Shenzhen Center for Disease Control and Prevention (SZCDC-IRB2024055). As this study involved no direct patient intervention and was retrospective in nature, individual informed consent was waived.

Results

Data Characteristics

We focused on 2 retrospective inquiry-and-diagnosis evaluation tasks in the emergency and pediatrics departments. The emergency departments provided 339,020 records, and the pediatrics departments provided 561,659 records. We randomly split the data into 3 parts: 75% for training MedRIA, 10% for validation, and 15% for testing.

The emergency EHRs contained 862 features, including 42.8% (369/862) SMH features, 21.2% (183/862) PE features, and 36% (310/862) AE features. There were 18 diagnosed diseases. As some records had 2 diagnosed diseases, we classified each combination as a separate disease category, bringing the total number of distinct diagnoses to 23. Similarly, the pediatrics EHRs covered 793 features, including 44.3% (351/793) SMH features, 22.6% (179/793) PE features, and 33.2% (263/793) AE features. There were 18 diagnosed diseases. When combinations of diagnosed diseases were considered, the total number of different diagnoses increased to 23. The detailed data characteristics are presented in Tables 1 and 2.

XSL•FO
RenderX

Table 1. Characteristics of the data in the emergency task (N=339,020).

Characteristics	Values
Gender, n (%)	
Male	187,992 (55.5)
Female	151,028 (44.5)
Age (years), median (IQR)	5.2 (2.6-26.1)
Diagnosis, n (%)	
AURTI[a]	93,594 (27.6)
Bronchitis	93,564 (27.6)
Bronchitis with rhinitis	2439 (0.7)
AURTI with bronchitis	4923 (1.5)
Bronchopneumonia	3833 (1.1)
Laryngopharyngitis	6095 (1.8)
Pharyngitis	24,146 (7.1)
AURTI with pharyngitis	2639 (0.8)
Pharyngitis with bronchitis	3564 (1.1)
Herpangina	7604 (2.2)
Tonsillitis	10,286 (3)
AURTI with tonsillitis	1784 (0.5)
Gastroenteritis	30,385 (9)
Gastritis	10,142 (3)
Enteritis	2690 (0.8)
Acute appendicitis	1957 (0.6)
Urinary tract stones	13,884 (4.1)
Urethritis	4052 (1.2)
Dermatitis	6239 (1.8)
Acute urticaria	5597 (1.7)
Hand-foot-and-mouth disease	2990 (0.9)
Hypertension	4397 (1.3)
Lumbar disc herniation	2216 (0.7)

[a]AURTI: acute upper respiratory tract infection.

XSL•FO
RenderX

Table 2. Characteristics of the data in the pediatrics task (N=561,659).

Characteristics	Values
Gender, n (%)	
Male	315,074 (56.1)
Female	246,585 (43.9)
Age (years), median (IQR)	3.8 (2.0-5.7)
Diagnosis, n (%)	
Bronchitis	209,918 (37.4)
AURTI[a]	155,181 (27.6)
AURTI with bronchitis	11,734 (2.1)
Bronchopneumonia	8119 (1.4)
Bronchitis with bronchopneumonia	2565 (0.5)
Pharyngitis	41,205 (7.3)
Pharyngitis with bronchitis	5232 (0.9)
AURTI with pharyngitis	2859 (0.5)
Laryngopharyngitis	11,909 (2.1)
Tonsillitis	19,790 (3.5)
Herpangina	15,647 (2.8)
Rhinitis	14,824 (2.6)
Bronchitis with rhinitis	10,109 (1.8)
Gastroenteritis	16,261 (2.9)
Gastritis	5313 (0.9)
Enteritis	3479 (0.6)
Indigestion	8757 (1.6)
Hand-foot-and-mouth disease	5156 (0.9)
Dermatitis	3151 (0.6)
Acute urticaria	2595 (0.5)
Developmental delay	2953 (0.5)
Hyperactivity	2538 (0.5)
Precocious puberty	2364 (0.4)

[a]AURTI: acute upper respiratory tract infection.

Evaluation Criteria

We assessed the inquiry performance of MedRIA along 2 dimensions: the consistency of its inquiries with those of physicians and the contribution of its inquiries to diagnostic accuracy. It should be noted that physicians will inevitably make incorrect diagnoses, conduct inquiries based on individual previous experience, and record the EHRs in different ways. Therefore, we can establish a baseline diagnostic model by training a supervised classification model using the extracted feature sets as input and the physicians' diagnoses as output. This model simulated the physicians' diagnostic ability, and its performance approximates the actual performance of physicians. We aimed to mitigate the influence of individual physician variability and ensure a fair comparison between different inquiry methods based on the extracted feature sets. While the diagnoses in the records could serve as a gold standard, our simulated physician diagnostic model also serves as an important baseline. For training, several machine learning models were used, including logistic regression, random forest, support vector machine, MLP, and light gradient-boosting machine [42]. We chose light gradient-boosting machine for the emergency task and random forest for the pediatrics task based on their performance on the validation set.

To better evaluate the potential of MedRIA as a clinical assistant, we simulated a collaborative inquiry process between MedRIA and a physician. Specifically, MedRIA's actor network would generate probability distributions for potential inquiry features based on the currently collected medical information. These distributions suggest several high-probability features for the physician to inquire about. The physician has the flexibility to accept and incorporate some or all of these suggestions into the inquiries. In addition, the physician can

XSL•FO
RenderX

exercise their professional judgment to either ask additional questions or proceed to examination. In our simulated collaboration, we simplified this process by having MedRIA recommend only the feature with the highest probability at each step. If the inquiry features suggested by MedRIA were in agreement with part of the extracted feature set, it was assumed that the physician accepted the inquiry recommendations. Otherwise, if the suggested features were not found in the extracted feature set, the simulated physician would randomly select an unasked feature from the extracted features as the next inquiry question. This simplification represents a simplistic approximation of a less informed or junior physician's behavior. It serves as a conservative estimate of MedRIA's performance when collaborating with less experienced counterparts, reflecting a lower-bound performance of MedRIA-physician collaborative inquiry. Figure 2 illustrates a vivid example of the simulated collaborative inquiry process.

Figure 2. Schematic illustration of the simulated collaborative inquiry process between a physician and MedRIA. AI: artificial intelligence.

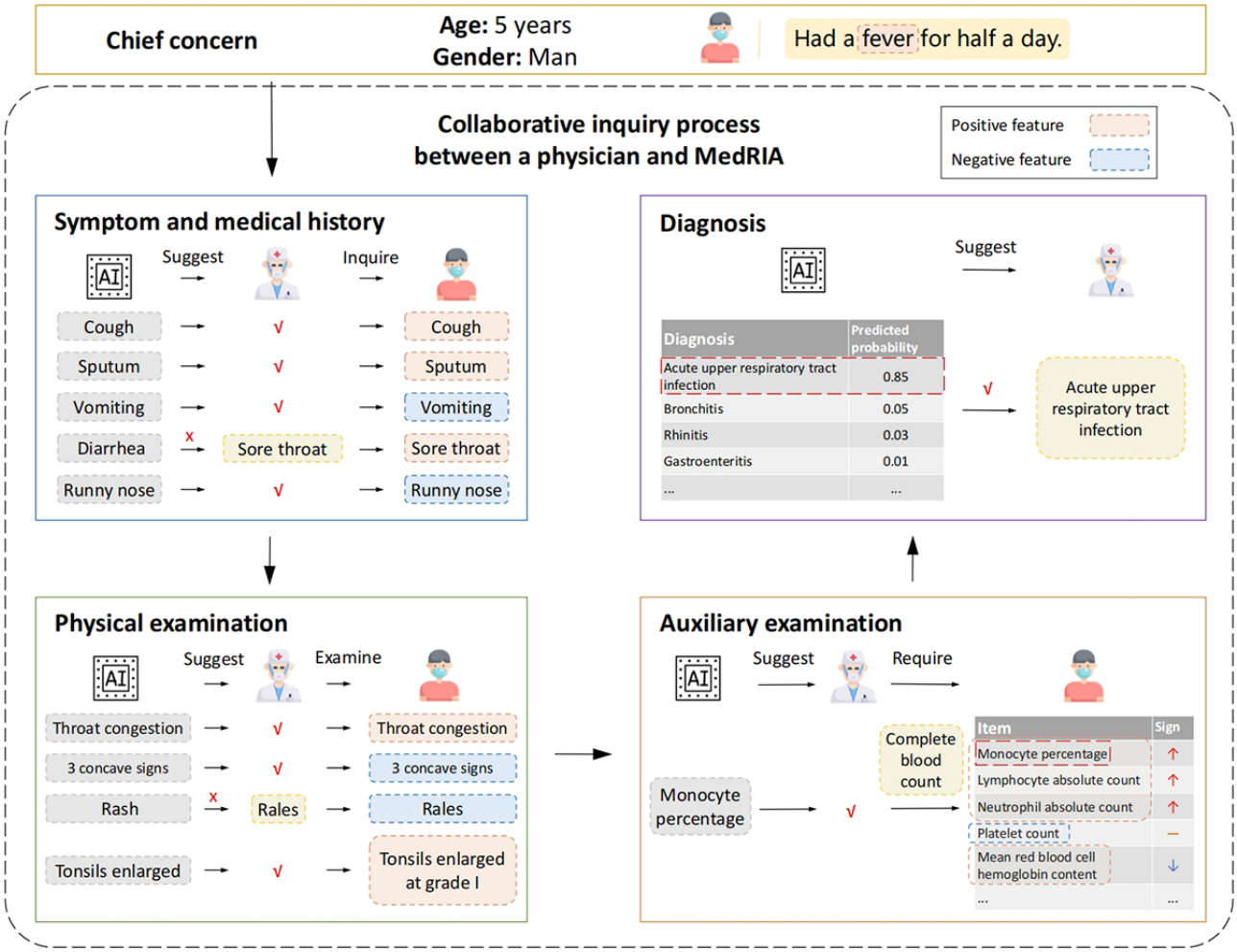

Inquiry Quality Analysis

Table 3 illustrates the performance of MedRIA in the emergency and pediatrics tasks in terms of diagnosis accuracy and inquiry consistency. In the emergency task, physicians used an average of 13.26 (95% CI 13.202-13.317) inquiries to gather 19.488 (95% CI 19.361-19.618) features. This disparity in numbers is because a single inquiry related to a specific laboratory test can yield multiple features. For example, a complete blood count includes several features, such as lymphocyte percentage, platelet count, hemoglobin concentration, and so on. When MedRIA conducted its own inquiries, it averaged 14.24 (95% CI 14.2-14.284) inquiries, with 58% (8.253/14.24; 95% CI 8.212-8.297) of them matching the extracted features. It resulted in the acquisition of an average of 10.628 (95% CI 10.549-10.714) features, accounting for 55% (10.628/19.488) of all extracted features. In the pediatrics task, MedRIA performed, on average, 1.9 more inquiries (16.271, 95% CI 16.231-16.307) than physicians (14.364, 95% CI 14.322-14.404), with 60% (9.81/16.271; 95% CI 9.776-9.842) of them matching 64% (13.168/20.429; 95% CI 13.093-13.244) of all extracted features (20.429, 95% CI 20.338-20.513). When MedRIA collaborated with physicians, it recalled 95% (18.543/19.488; 95% CI 18.425-18.657) and 99% (20.168/20.429; 95% CI 20.080-20.251) of the extracted features in the emergency and pediatrics tasks, respectively. The simulated collaborative inquiry process required only an average of 12.648 (95% CI 12.602-12.696) and 14.146 (95% CI 14.106-14.182) inquiries, respectively, for the emergency and pediatrics tasks, both fewer than the number of inquiries made by physicians. With the assistance of MedRIA, we were able to reduce the number of inquiries made by physicians to 46% (6.037/13.26; 95% CI 6.009-6.064) and 43% (6.245/14.364;

XSL•FO
RenderX

95% CI 6.225-6.269) of those in independent inquiries while recovering nearly all the original inquiry records.

We used multiple metrics to assess final diagnoses, including the area under the receiver operating characteristic curve (AUROC), macro–F_1-score, and accuracy. We also introduced a new metric called general accuracy, which considers both fully and partially correct diagnoses in accuracy calculation because health care professionals from different hospitals may have varying recording habits. A diagnosis was considered fully correct when the core diagnosis was recorded and partially correct when only a constituent element of the core diagnosis was recorded [43,44]. As shown in Table 3, when MedRIA conducted its own inquiries in the emergency task, it achieved an AUROC (0.955, 95% CI 0.953-0.956) comparable to when it acquired complete extracted features (0.960, 95% CI 0.957-0.962) despite obtaining only 55% (10.628/19.488) of the extracted features. In the pediatrics task, MedRIA achieved a higher AUROC (0.943, 95% CI 0.941-0.944) with 64% (13.168/20.429) of the extracted features compared to using complete features (0.930, 95% CI 0.928-0.932). As shown in Figure 3, MedRIA's own inquiry and the collaborative inquiry in the emergency task significantly improved the AUROC of diagnostic predictions for acute appendicitis. Specifically, when compared to physicians' inquiry, both MedRIA's own inquiry and the collaborative inquiry demonstrated substantially lower false negative rates (equal to 1 minus true positive rates) at various false positive rate thresholds. These findings underscore the potential of MedRIA to significantly reduce diagnostic errors, particularly for critical conditions such as acute appendicitis in which even small increases in false negative rates may be unacceptable to patients and health system leaders.

The collaborative inquiry and the physicians' inquiry resulted in similar AUROCs for acute upper respiratory tract infections and bronchopneumonia. As shown in Figure 4, MedRIA's own inquiry resulted in a higher AUROC for rhinitis than the physicians' inquiry in the pediatrics task but a lower AUROC for bronchitis and bronchopneumonia. For these 3 diseases, the collaborative inquiry yielded the highest AUROC. The complete receiver operating characteristic curves for all diseases are shown in Figures S1 and S2 in Multimedia Appendix 4. These results demonstrate that MedRIA understands the importance of inquiring about discriminative features that have a high diagnostic contribution, especially in the pediatrics task.

Table 3. Performance of MedRIA in the emergency and pediatrics tasks.

	AUROC[a] (95% CI)	F_1-score (95% CI)	General accuracy (95% CI)	Accuracy (95% CI)	Average number of inquiries (95% CI)	Average number of inquiries matching extracted features (95% CI)	Average number of inquiries conducted by physicians (95% CI)	Average number of recalled features (95% CI)
Emergency								
Physicians	0.96 (0.957-0.962)	0.63 (0.623-0.638)	0.853 (0.85-0.856)	0.744 (0.741-0.748)	13.26 (13.202-13.317)	13.26 (13.202-13.317)	13.26 (13.202-13.317)	19.488 (19.361-19.618)
MedRIA	0.955 (0.953-0.956)	0.538 (0.531-0.545)	0.806 (0.803-0.81)	0.684 (0.68-0.688)	14.24 (14.2-14.284)	8.253 (8.212-8.297)	0	10.628 (10.549-10.714)
Collaboration	0.972 (0.97-0.973)	0.646 (0.639-0.653)	0.854 (0.851-0.857)	0.746 (0.742-0.75)	12.648 (12.602-12.696)	12.648 (12.602-12.696)	6.037 (6.009-6.064)	18.543 (18.425-18.657)
Pediatrics								
Physicians	0.93 (0.928-0.932)	0.613 (0.608-0.618)	0.827 (0.825-0.83)	0.704 (0.701-0.708)	14.364 (14.322-14.404)	14.364 (14.322-14.404)	14.364 (14.322-14.404)	20.429 (20.338-20.513)
MedRIA	0.943 (0.941-0.944)	0.521 (0.514-0.527)	0.782 (0.779-0.784)	0.656 (0.653-0.66)	16.271 (16.231-16.307)	9.81 (9.776-9.842)	0	13.168 (13.093-13.244)
Collaboration	0.968 (0.967-0.969)	0.633 (0.627-0.638)	0.842 (0.839-0.844)	0.732 (0.729-0.735)	14.146 (14.106-14.182)	14.146 (14.106-14.182)	6.245 (6.225-6.269)	20.168 (20.08-20.251)

[a]AUROC: area under the receiver operating characteristic curve.

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curves for 3 diseases in the emergency task. AUC: area under the curve; AURTI: acute upper respiratory tract infection.

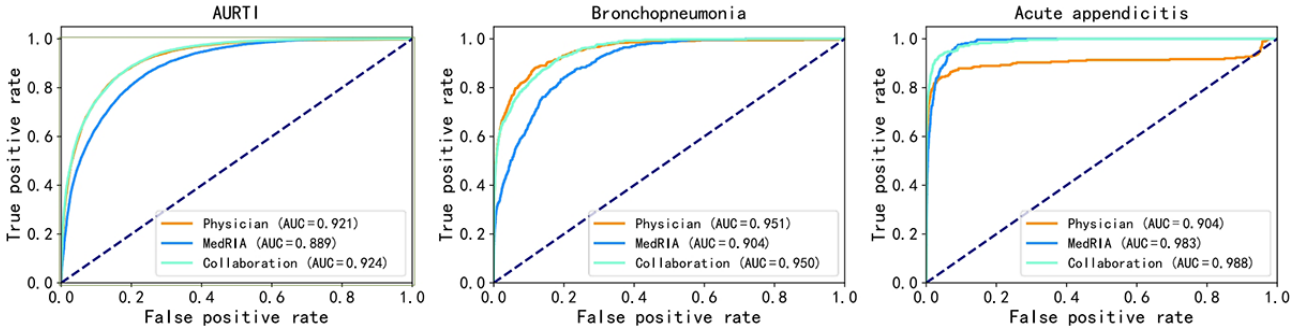

Figure 4. Receiver operating characteristic curves for 3 diseases in the pediatrics task. AUC: area under the curve.

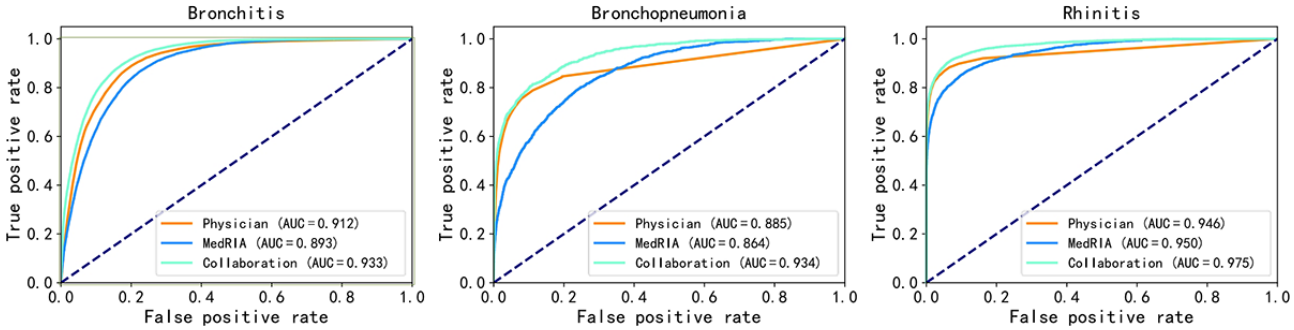

MedRIA may conduct inquiries that were not asked by physicians, so they are excluded from the extracted features of patient EHRs. Hence, when MedRIA conducted its own inquiries, there was a varying degree of decline in all diagnostic metrics when compared to using complete extracted features. On the other hand, MedRIA obtained more information through the guidance of physicians when it collaborated with them. In the emergency task, the F_1-score (0.646, 95% CI 0.639-0.653) of diagnoses significantly improved. Both general accuracy (0.854, 95% CI 0.851-0.857) and accuracy (0.746, 95% CI 0.742-0.750) were comparable to those of physicians' diagnoses. In the pediatrics task, collaborative inquiry led to a significant increase in F_1-score (0.633, 95% CI 0.627-0.638), general accuracy (0.842, 95% CI 0.839-0.844), and accuracy (0.732, 95% CI 0.729-0.735).

Considering the high prevalence of respiratory diseases in our data set, we evaluated MedRIA's performance on a more representative data set to show its generalizability. Specifically, we randomly discarded two-thirds of respiratory cases from the emergency data set and three-quarters from the pediatrics data set. The results are presented in Table S1 in Multimedia Appendix 5. From these results, we can draw similar conclusions to those obtained from the experiments conducted on the complete data set.

Inquiry Logic Analysis

Table 4 provides the number of different types of positive recalled features, which refer to exhibited symptoms or laboratory test results showing out-of-range values. Collecting more positive features under the same number of inquiries often leads to more accurate diagnosis results [16,45]. In the pediatrics task, 36% (7.326/20.429; 95% CI 7.294-7.357) of the features collected by physicians were positive, and MedRIA's own inquiries recalled 66% (4.852/7.326; 95% CI 4.825-4.877) of positive features. Specifically, it recalled 62% (1.665/2.692; 95% CI 1.654-1.675) of SMH features, 79% (2.24/2.837; 95% CI 2.23-2.25) of PE features, and 53% (0.948/1.797; 95% CI 0.931-0.965) of AE features. When MedRIA worked with physicians, the recall rates for all types of features were >95%. Overall, we observed that MedRIA was capable of accurately identifying positive features that were not mentioned in the chief concerns of patients. This ability is crucial during the investigation process because patients may not always have sufficient medical knowledge to fully describe their health conditions.

Multimedia Appendix 6 shows the top 10 physician inquiry features on patients diagnosed with pharyngitis in the emergency task and tonsillitis in the pediatrics task. We observed that MedRIA could accurately inquire about features obtained by physicians. Figures 5 and 6 illustrate when these features were investigated by MedRIA and through collaborative investigation, respectively. First, we observed a clear distinction between when SMH and PE features were acquired, which is in line with our settings for MedRIA's inquiry process. SMH features, including cough, runny nose, convulsions, and chills, were consistently acquired before the seventh step, whereas PE features, including pharyngeal congestion, enlarged tonsils, coarse breath sounds, rash, herpes, and hand and foot rash, were consistently acquired after the seventh step. We also noticed that consecutive steps of inquiries exhibited consistently high frequencies, allowing us to understand MedRIA's inquiry logic clearly. For example, when conducting PEs, MedRIA began with pharyngeal congestion and enlarged tonsils followed by coarse breath sounds and rash. This logical sequence reflects a structured approach to differential diagnosis, ensuring a thorough evaluation of relevant symptoms and signs. Pharyngeal congestion and enlarged tonsils are pivotal features in the

XSL•FO
RenderX

context of pharyngitis. Coarse breath sounds may suggest respiratory issues or lung conditions, whereas rash can be associated with a wide range of dermatological, allergic, or infectious disorders. As shown in Figures 5 and 6, when collaborating with physicians, the refinement of MedRIA's inquiry logic is evident in the increased stability of determining which features to inquire about at each step. Table 5 provides concrete examples of MedRIA's inquiry sequences, where the aforementioned logic can be traced back. Figure S1 in Multimedia Appendix 7 presents the tree diagram depicting the order of MedRIA's inquiry features for 4 patients diagnosed with pharyngitis in the emergency task. These evidences demonstrate MedRIA's coherent, stable, and well-reasoned inquiry logic. The characteristics of inquiry features for all diseases in both tasks are shown in Figures S1-S6 in Multimedia Appendix 8.

Table 4. Number of recalled features from the extracted features by MedRIA in the emergency and pediatrics tasks.

	Average SMH[a] features (95% CI)	Average PE[b] features (95% CI)	Average AE[c] features (95% CI)	Average positive features[d] (95% CI)	Average positive SMH features (95% CI)	Average positive PE features (95% CI)	Average positive AE features (95% CI)
Emergency							
Physicians	7.295 (7.26-7.332)	5.291 (5.265-5.32)	6.902 (6.793-7.011)	6.83 (6.788-6.87)	2.52 (2.503-2.538)	2.527 (2.514-2.541)	1.782 (1.749-1.813)
MedRIA	4.31 (4.29-4.331)	3.835 (3.808-3.865)	2.483 (2.421-2.546)	3.923 (3.897-3.954)	1.363 (1.353-1.375)	1.907 (1.894-1.92)	0.653 (0.635-0.672)
Collaboration	7.01 (6.978-7.043)	5.222 (5.196-5.25)	6.311 (6.206-6.415)	6.543 (6.505-6.581)	2.435 (2.418-2.452)	2.501 (2.489-2.515)	1.607 (1.576-1.636)
Pediatrics							
Physicians	7.522 (7.494-7.546)	6.29 (6.269-6.312)	6.617 (6.546-6.687)	7.326 (7.294-7.357)	2.692 (2.677-2.707)	2.837 (2.826-2.848)	1.797 (1.776-1.82)
MedRIA	5.021 (5.001-5.04)	4.639 (4.619-4.66)	3.508 (3.452-3.565)	4.852 (4.825-4.877)	1.665 (1.654-1.675)	2.24 (2.23-2.25)	0.948 (0.931-0.965)
Collaboration	7.461 (7.435-7.484)	6.28 (6.258-6.302)	6.427 (6.356-6.496)	7.241 (7.209-7.27)	2.672 (2.658-2.687)	2.833 (2.822-2.844)	1.736 (1.714-1.758)

[a]SMH: symptom and medical history.

[b]PE: physical examination.

[c]AE: auxiliary examination.

[d]Refers to exhibited symptoms or laboratory test results showing out-of-range values.

Figure 5. MedRIA and collaborative inquiry heat maps of the top 10 physician inquiry features in patients diagnosed with pharyngitis in the emergency task.

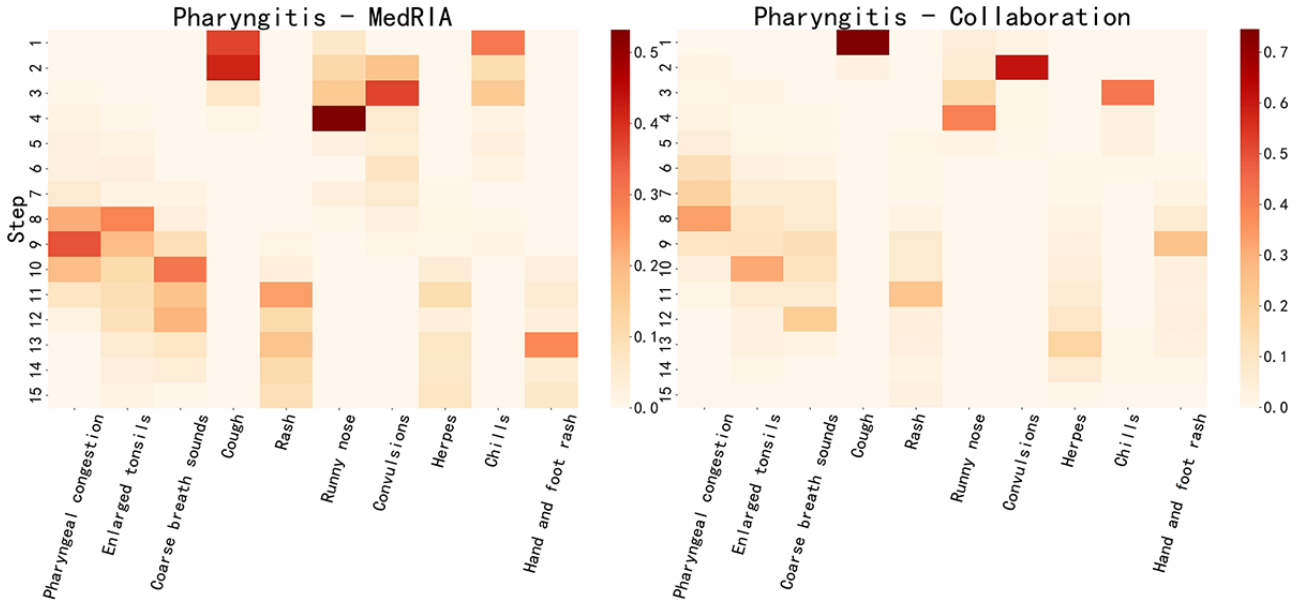

XSL•FO
RenderX

Figure 6. MedRIA and collaborative inquiry heat maps of the top 10 physician inquiry features in patients diagnosed with tonsillitis in the pediatrics task.

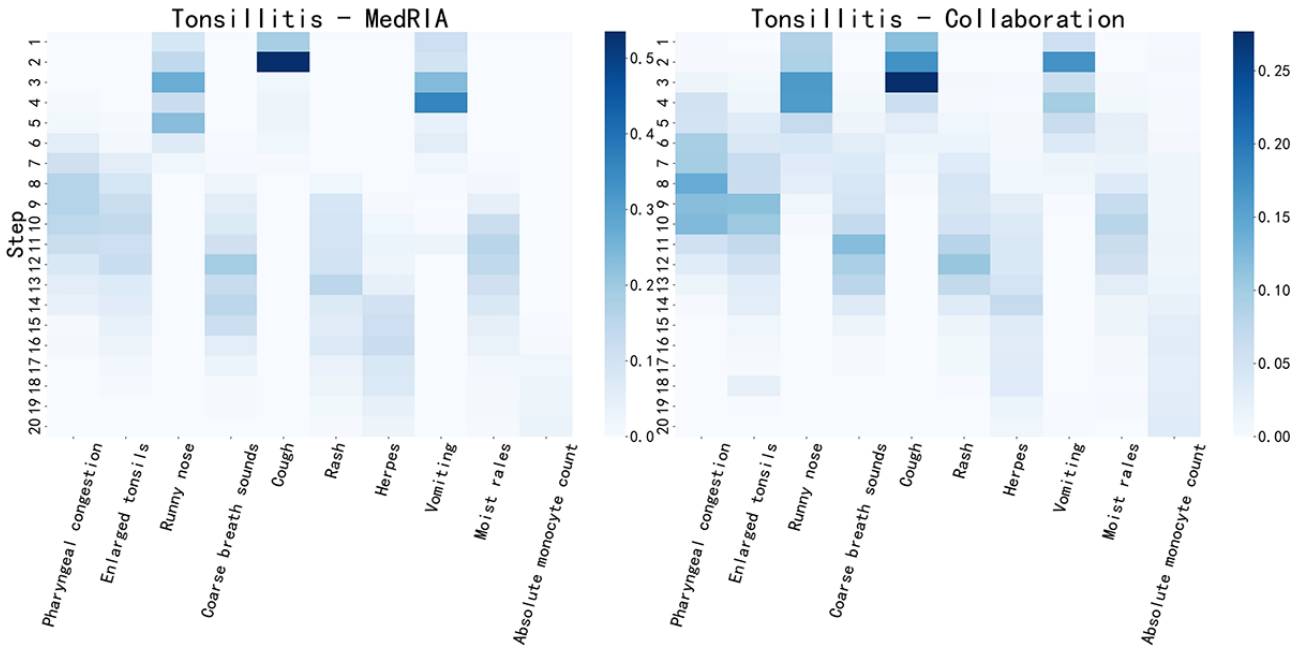

Table 5. Examples of the inquiry sequences of MedRIA on 3 patients diagnosed with pharyngitis.

Gender	Age (y)	Chief concern	Inquiry sequences of SMH[a]	Inquiry sequences of PE[b]	Inquiry sequences of AE[c]
Male	26	*Sore throat* [d]	*Fever(0)*[e], *cough(0)*, *runny nose(0)*, *nasal congestion(0)*, *chest pain(0)*, *coughing up phlegm(0)*, abdominal pain, diarrhea	*Pharyngeal congestion(1)*, *enlarged tonsils(0)*, *moist rales(0)*, *dry rales(0)*, lower limb edema, *coarse breath sounds(0)*	*Absolute monocyte count [complete blood count]* [f]
Male	50	*Pharyngeal foreign body sensation*	Fever, *cough(1)*, *sore throat(0)*, runny nose, *coughing up phlegm(0)*, abdominal pain, diarrhea, vomiting, dizziness	*Pharyngeal congestion(1)*, *enlarged tonsils(0)*, *moist rales(0)*, *dry rales(0)*, *coarse breath sounds(0)*, rash	Absolute monocyte count
Male	4	*Fever*	*Cough(1)*, runny nose, *vomiting(0)*, coughing up phlegm, *diarrhea(0)*, *convulsions(0)*, *shortness of breath(0)*	*Pharyngeal congestion(1)*, *coarse breath sounds(1)*, *enlarged tonsils(0)*, *hand and foot Rash(0)*, *dry rales(0)*, normal muscle tone, *moist rales(0)*	*Absolute monocyte count [complete blood count]*

[a]SMH: symptom and medical history.

[b]PE: physical examination.

[c]AE: auxiliary examination.

[d]Features belonging to the extracted features are italicized.

[e]Feature values are indicated in parentheses.

[f]Contents within square brackets represent the associated laboratory tests documented in the patient's health record.

External Validation

We used external medical dialogues to evaluate MedRIA's inquiry logic. These dialogues consist of inquiry sequences generated by physicians during medical consultations. Limited by the online scenario, the average length of the inquiry sequences was 3.905 (95% CI 3.744-4.076), consisting of 3.043 (95% CI 2.895-3.195) inquiries about SMH and PE features and 0.862 (95% CI 0.814-0.907) recommended AEs. Diagnoses based on such few features were unreliable, so we primarily focused on evaluating MedRIA's own inquiry order. As a result, MedRIA matched 64% (1.946/3.043; 95% CI 1.835-2.058) of the inquiries in the inquiry sequences and 28% (0.244/0.862; 95% CI 0.218-0.269) of recommended AEs. We compared the order of MedRIA's inquiries to the inquiry sequences of physicians. The normalized Kendall τ distance was 0.323 (95% CI 0.301-0.346). It indicated that only 32.3% of pairs of inquiry features differed in ordering, demonstrating MedRIA's consistency with physicians and reasonable inquiry logic.

Discussion

Principal Findings

In this study, we proposed a reinforcement learning medical inquiry assistant, MedRIA, aiming to provide inquiry recommendations for medical investigation processes. The variety of questions that can be asked during the medical investigation process, especially when it incorporates PEs and

XSL•FO
RenderX

AEs, poses a significant challenge. In addition, patients vary in their understanding of their health status and their ability to articulate it. To address these issues, MedRIA uses an actor-critic framework [28]. The actor incorporates a pretrained VAE [31] to recommend actions based on conditional probability distributions between observed and unobserved patient information. MedRIA uses a supervised diagnostic prediction model to aid in action selection, ensuring that the actor suggests inquiries that contribute to accurate diagnoses. In addition, MedRIA places significant emphasis on crafting a reward function to guide the reinforcement learning process.

The superiority of these designs in MedRIA was verified in our retrospective experiments. When conducting its own inquiries in both emergency and pediatrics scenarios, MedRIA demonstrated comparable diagnostic prediction performance on AUROCs with that of physicians and even achieved a higher AUROC in the pediatrics setting. This was accomplished with MedRIA obtaining only 55% (10.628/19.488) and 64% (13.168/20.429) of the information that physicians received in making the final diagnosis in the emergency and pediatrics tasks, respectively. It is important to consider that AI, similarly to MedRIA, has the capacity to identify subtle patterns within vast data sets, which physicians may overlook. This inherent capability often leads to divergence from the inquiry logic of individual clinicians, as evidenced in our experiments. Given that MedRIA has demonstrated its ability to identify important diagnostic features, such divergence should be considered acceptable as it may provide new insights. In addition, it is worth noting that our experiments were based on EHRs. Therefore, when MedRIA conducted its own inquiries, its performance might very well have been underestimated due to the unavailability of patient information that was not documented in the records.

When MedRIA worked in simulated collaboration with physicians, it notably reduced the number of inquiries made by physicians to 46% (6.037/13.26; 95% CI 6.009-6.064) and 43% (6.245/14.364; 95% CI 6.225-6.269) while maintaining feature recall rates of 95% (18.543/19.488; 95% CI 18.425-18.657) and 99% (20.168/20.429; 95% CI 20.080-20.251) in the emergency and pediatrics scenarios, respectively. This reduction suggests that MedRIA's recommendations are sufficiently informative to assist physicians in considering aspects they may have overlooked. The additional insights provided by AI recommendations have a significant impact on medical decision-making [46]. By leveraging the insights provided by MedRIA, physicians can potentially make more informed decisions and arrive at more accurate diagnoses.

The demonstration and analysis of MedRIA's inquiry feature frequencies, as well as concrete inquiry sequences, reveal its clear inquiry logic for different diseases, which resembles the approach of physicians. The ability of MedRIA to emulate the diagnostic process of human physicians suggests that it has the potential to be a useful tool in medical education, particularly in regions or populations with limited medical resources. Junior health care practitioners can leverage MedRIA to enhance their diagnostic skills, benefiting from its structured and logical inquiry process.

Limitations

Despite these promising results, it is essential to acknowledge certain limitations. First, the feature extraction process of EHRs requires careful feature engineering, which could be further refined. Second, while MedRIA excelled in adapting its inquiry logic, it is not immune to instances of missed inquiries. For example, the reliance on AEs of MedRIA is significantly lower than that of physicians. One potential solution is to enhance MedRIA by modeling the significance of inquiries for disease treatment. This is because the purpose of examinations may not be for diagnosis but to determine the severity of the disease to formulate appropriate treatment.

Third, the composition of patient presentations can vary significantly between countries. For example, in some Western countries, only 5% to 20% of emergency [47] or pediatrics [48] outpatients are due to upper respiratory tract infections. The proportion in our data set was many times greater due to seasonal variations. We speculate that this may have arisen due to limited access to primary care, seasonal respiratory peaks during the autumn and winter months, greater COVID-19 awareness and fear, and our method of prioritizing the most frequent diagnoses. Changes in these factors, such as using summer data in China or winter data elsewhere, may influence accuracy. Furthermore, while MedRIA's workflow is language independent, the evaluation of our system has been limited to a Chinese context. Further validation in other languages and health care settings is necessary to establish its broader applicability and effectiveness.

Fourth, our AI system would be biased toward excess certainty if deployed in the real world as it lacks the humility to recognize its diagnostic limitations. Future work should include the AI system's degree of confidence and manage corresponding clinical risks. Fifth, decision system developers must consider adopting safe integration systems to mitigate potential errors, especially for life-threatening conditions such as appendicitis, where risk aversion may be greater. Sixth, future work could incorporate MedRIA with the latest developments in LLMs [49], which will assist in its future deployment and testing in clinical workflows.

Conclusions

In conclusion, MedRIA is a significant step toward enhancing health care decision-making processes. Its proficiency in medical inquiries, ability to identify important diagnostic features, and adaptability underscore its value as an AI-driven diagnostic assistant. While there are challenges to overcome, the results of this study show a prototype for sophisticated and effective AI applied in health care, ultimately benefiting both health care professionals and patients.

XSL•FO
RenderX

Acknowledgments

This work is supported by the National Key R&D Program of China (2021YFF1201303 and 2022YFC2703105), Guoqiang Institute of Tsinghua University, and Beijing National Research Center for Information Science and Technology. The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The feature set of the emergency task.
[XLSX File (Microsoft Excel File), 26 KB - jmir_v26i1e54616_app1.xlsx]

Multimedia Appendix 2
The feature set of the pediatrics task.
[XLSX File (Microsoft Excel File), 28 KB - jmir_v26i1e54616_app2.xlsx]

Multimedia Appendix 3
Implementation details of MedRIA.
[DOCX File , 22 KB - jmir_v26i1e54616_app3.docx]

Multimedia Appendix 4
Receiver operating characteristic curves of all diseases in the emergency and pediatrics tasks.
[DOCX File , 2682 KB - jmir_v26i1e54616_app4.docx]

Multimedia Appendix 5
Performance of MedRIA in more generalizable emergency and pediatrics tasks.
[DOCX File , 21 KB - jmir_v26i1e54616_app5.docx]

Multimedia Appendix 6
Inquiry frequencies of the top 10 physician inquiry features in patients diagnosed with pharyngitis in the emergency task and tonsillitis in the pediatrics task.
[PNG File , 92 KB - jmir_v26i1e54616_app6.png]

Multimedia Appendix 7
Tree diagram of the inquiry features of MedRIA for 4 patients diagnosed with pharyngitis in the emergency task. These 4 patients shared similar characteristics, all being male with ages ranging from 31 to 32 years and complaining of sore throat. The numbers in brackets indicate the timesteps of inquiries.
[DOCX File , 215 KB - jmir_v26i1e54616_app7.docx]

Multimedia Appendix 8
Inquiry frequencies of the physician and MedRIA and inquiry heat maps of MedRIA and the collaborative inquiry in the emergency and pediatrics tasks.
[DOCX File , 8348 KB - jmir_v26i1e54616_app8.docx]

References

1. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Cham, Switzerland: Springer; 2021:795-840.
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
3. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med 2008 May;121(5 Suppl):S2-23. [doi: 10.1016/j.amjmed.2008.01.001] [Medline: 18440350]
4. Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. J Eval Clin Pract 2001 May;7(2):97-107. [doi: 10.1046/j.1365-2753.2001.00284.x] [Medline: 11489035]

XSL•FO
RenderX

5. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan 20;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
7. Liang H, Tsui BY, Ni H, Valentim CC, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019 Mar;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
8. Zhou HY, Yu Y, Wang C, Zhang S, Gao Y, Pan J, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng 2023 Jun 12;7(6):743-755. [doi: 10.1038/s41551-023-01045-x] [Medline: 37308585]
9. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 2020 Oct 16;3(1):136 [FREE Full text] [doi: 10.1038/s41746-020-00341-z] [Medline: 33083571]
10. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1(1):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
11. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017 Jun 21;19:221-248 [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
12. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Dec;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
13. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis 2017 Dec;42:60-88. [doi: 10.1016/J.MEDIA.2017.07.005]
14. He W, Mao X, Ma C, Huang Y, Hernàndez-Lobato J, Chen T. BSODA: a bipartite scalable framework for online disease diagnosis. In: Proceedings of the 2022 ACM Web Conference. 2022 Presented at: WWW '22; April 25-29, 2022; Virtual Event p. 2511-2521 URL: https://dl.acm.org/doi/10.1145/3485447.3512123 [doi: 10.1145/3485447.3512123]
15. Chen J, Li D, Chen Q, Zhou W, Liu X. Diaformer: automatic diagnosis via symptoms sequence generation. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022 Presented at: AAAI '22; February 22–March 1, 2022; Virtual Event p. 4432-4440 URL: https://cdn.aaai.org/ojs/20365/20365-13-24378-1-2-20220628.pdf [doi: 10.1609/aaai.v36i4.20365]
16. Liu Z, Li Y, Sun X, Wang F, Hu G, Xie G. Dialogue based disease screening through domain customized reinforcement learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021 Presented at: KDD '21; August 14-18, 2021; Virtual Event p. 1120-1128 URL: https://dl.acm.org/doi/10.1145/3447548.3467255 [doi: 10.1145/3447548.3467255]
17. Xia Y, Zhou J, Shi Z, Lu C, Huang H. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. Proc AAAI Conf Artif Intell 2020 Apr 03;34(01):1062-1069. [doi: 10.1609/aaai.v34i01.5456]
18. Zhong C, Liao K, Chen W, Liu Q, Peng B, Huang X, et al. Hierarchical reinforcement learning for automatic disease diagnosis. Bioinformatics 2022 Aug 10;38(16):3995-4001. [doi: 10.1093/bioinformatics/btac408] [Medline: 35775965]
19. Liu W, Cheng Y, Wang H, Tang J, Liu Y, Zhao R, et al. “My nose is running.” “Are you also coughing?”: building a medical diagnosis agent with interpretable inquiry logics. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. 2022 Presented at: IJCAI '22; July 23-29, 2022; Vienna, Austria p. 4266-4272 URL: https://www.ijcai.org/proceedings/2022/0592.pdf [doi: 10.24963/ijcai.2022/592]
20. Wang Z, Wen R, Chen X, Cao SL, Huang S, Qian B. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In: Proceedings of the 2021 Web Conference. 2021 Presented at: WWW '21; April 19-23, 2021; Ljubljana, Slovenia p. 3349-3358 URL: https://dl.acm.org/doi/10.1145/3442381.3449795 [doi: 10.1145/3442381.3449795]
21. Lin J, Wang K, Chen Z, Liang X, Lin L, Lin L. Towards causality-aware inferring: a sequential discriminative approach for medical diagnosis. IEEE Trans Pattern Anal Mach Intell 2021:1-12. [doi: 10.1109/tpami.2023.3292363]
22. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv. Preprint posted online March 4, 2022 [FREE Full text]
23. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
24. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
25. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
26. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135 [FREE Full text] [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
27. Taylor J. ChatGPT’s alter ego, Dan: users jailbreak AI program to get around ethical safeguards. The Guardian. URL: https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards [accessed 2024-04-29]

XSL•FO
RenderX

28. Mnih V, Badia A, Mirza M, Graves A, Lillicrap T, Harley TP, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning. 2016 Presented at: ICML'16; June 19-24, 2016; New York, NY p. 1928-1937 URL: https://dl.acm.org/doi/10.5555/3045390.3045594
29. Balogh EP, Miller BT, Ball JR. Improving Diagnosis in Health Care. New York, NY: The National Academies Press; 2015.
30. Kenton JD, Toutanova LK. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL-HLT '19; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf [doi: 10.18653/v1/N19-1423]
31. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv. Preprint posted online December 20, 2013 [FREE Full text] [doi: 10.61603/ceas.v2i1.33]
32. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv. Preprint posted online July 20, 2017 [FREE Full text]
33. Konda VR, Tsitsiklis JN. OnActor-critic algorithms. SIAM J Control Optim 2003 Jan;42(4):1143-1166. [doi: 10.1137/S0363012901385691]
34. Schulman J, Moritz P, Levine S, Jordan M, Abbeel P. High-dimensional continuous control using generalized advantage estimation. arXiv. Preprint posted online June 8, 2015 [FREE Full text] [doi: 10.48550/arXiv.1506.02438]
35. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online December 22, 2014 [FREE Full text]
36. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. arXiv. Preprint posted online December 3, 2019 [FREE Full text] [doi: 10.7551/mitpress/11474.003.0014]
37. Weng J, Chen H, Yan D, You K, Duburcq A, Zhang M. Tianshou: a highly modularized deep reinforcement learning library. J Mach Learn Res 2022;23(267):1-6 [FREE Full text]
38. Bellman R. A markovian decision process. Indiana Univ Math J 1957;6(4):679-684. [doi: 10.1512/iumj.1957.6.56038]
39. He W, Chen T. Scalable online disease diagnosis via multi-model-fused actor-critic reinforcement learning. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022 Presented at: KDD '22; August 14-18, 2022; Washington, DC p. 4695-4703 URL: https://dl.acm.org/doi/10.1145/3534678.3542672 [doi: 10.1145/3534678.3542672]
40. Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. arXiv. Preprint posted online February 14, 2018 [FREE Full text]
41. Chen W, Li Z, Fang H, Yao Q, Zhong C, Hao J, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. Bioinformatics 2023 Jan 01;39(1):btac817 [FREE Full text] [doi: 10.1093/bioinformatics/btac817] [Medline: 36539203]
42. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 3149-3157 URL: https://dl.acm.org/doi/10.5555/3294996.3295074
43. Schmidt HG, Van Gog T, Schuit SC, Van den Berge K, Van Daele PL, Bueving H, et al. Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. BMJ Qual Saf 2017 Jan 07;26(1):19-23. [doi: 10.1136/bmjqs-2015-004109] [Medline: 26951795]
44. Mamede S, van Gog T, van den Berge K, Rikers RM, van Saase JL, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. JAMA 2010 Sep 15;304(11):1198-1203. [doi: 10.1001/jama.2010.1276] [Medline: 20841533]
45. Peng Y, Tang K, Lin H, Chang E. REFUEL: exploring sparse features in deep reinforcement learning for fast disease diagnosis. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018 Presented at: NIPS'18; December 3-8, 2018; Montréal, QC p. 7333-7342 URL: https://dl.acm.org/doi/10.5555/3327757.3327834 [doi: 10.7551/mitpress/11474.003.0014]
46. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digit Med 2023 Nov 07;6(1):206 [FREE Full text] [doi: 10.1038/s41746-023-00955-z] [Medline: 37935953]
47. Nørgaard B, Mogensen CB, Teglbjærg LS, Brabrand M, Lassen AT. Diagnostic packages can be assigned accurately in emergency departments. A multi-centre cohort study. Dan Med J 2016 Jun;63(6):A5240 [FREE Full text] [Medline: 27264941]
48. Hiscock H, Roberts G, Efron D, Sewell JR, Bryson HE, Price AM, et al. Children attending paediatricians study: a national prospective audit of outpatient practice from the Australian paediatric research network. Med J Aust 2011 Apr 18;194(8):392-397. [doi: 10.5694/j.1326-5377.2011.tb03028.x] [Medline: 21495938]
49. Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023 Apr 12;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]

XSL•FO
RenderX

Abbreviations

AE: auxiliary examination
AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
CDSS: clinical decision support system
EHR: electronic health record
LLM: large language model
MedBERT: Medical Bidirectional Encoder Representations From Transformers
MLP: multilayer perceptron
PE: physical examination
SMH: symptom and medical history
VAE: variational autoencoder

Edited by S Ma; submitted 17.11.23; peer-reviewed by H Lee, T Tillmann; comments to author 14.03.24; revised version received 13.04.24; accepted 04.07.24; published 23.08.24.

Please cite as:
Zou X, He W, Huang Y, Ouyang Y, Zhang Z, Wu Y, Wu Y, Feng L, Wu S, Yang M, Chen X, Zheng Y, Jiang R, Chen T
AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation
J Med Internet Res 2024;26:e54616
URL: https://www.jmir.org/2024/1/e54616
doi:10.2196/54616
PMID:

©Xuan Zou, Weijie He, Yu Huang, Yi Ouyang, Zhen Zhang, Yu Wu, Yongsheng Wu, Lili Feng, Sheng Wu, Mengqi Yang, Xuyan Chen, Yefeng Zheng, Rui Jiang, Ting Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Creating a Modified Version of the Cambridge Multimorbidity Score to Predict Mortality in People Older Than 16 Years: Model Development and Validation

Debasish Kar[1,2*], MD, FRCP; Kathryn S Taylor[1*], PhD; Mark Joy[1], PhD; Sudhir Venkatesan[3], PhD; Wilhelmine Meeraus[4], PhD; Sylvia Taylor[4], PhD; Sneha N Anand[1], PhD; Filipa Ferreira[1], PhD; Gavin Jamie[1], MSc, MBBS; Xuejuan Fan[1], PhD; Simon de Lusignan[1,5], MBBS

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

[2]Peninsula Medical School, University of Plymouth, Plymouth, United Kingdom

[3]Medical & Payer Evidence Statistics, BioPharmaceuticals Medical, AstraZeneca PLC, Cambridge, United Kingdom

[4]Medical Evidence, Vaccines and Immune Therapies, AstraZeneca PLC, Cambridge, United Kingdom

[5]Royal College of General Practitioners of the United Kingdom, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Kathryn S Taylor, PhD
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Primary Care Building, Radcliffe Observatory Quarter
Woodstock Road
Oxford, OX2 6GG
United Kingdom
Phone: 44 1865617855
Email: kathryn.taylor@phc.ox.ac.uk

Abstract

Background: No single multimorbidity measure is validated for use in NHS (National Health Service) England's General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR), the nationwide primary care data set created for COVID-19 pandemic research. The Cambridge Multimorbidity Score (CMMS) is a validated tool for predicting mortality risk, with 37 conditions defined by Read Codes. The GDPPR uses the more internationally used Systematized Nomenclature of Medicine clinical terms (SNOMED CT). We previously developed a modified version of the CMMS using SNOMED CT, but the number of terms for the GDPPR data set is limited making it impossible to use this version.

Objective: We aimed to develop and validate a modified version of CMMS using the clinical terms available for the GDPPR.

Methods: We used pseudonymized data from the Oxford-Royal College of General Practitioners Research and Surveillance Centre (RSC), which has an extensive SNOMED CT list. From the 37 conditions in the original CMMS model, we selected conditions either with (1) high prevalence ratio (≥85%), calculated as the prevalence in the RSC data set but using the GDPPR set of SNOMED CT codes, divided by the prevalence included in the RSC SNOMED CT codes or (2) conditions with lower prevalence ratios but with high predictive value. The resulting set of conditions was included in Cox proportional hazard models to determine the 1-year mortality risk in a development data set (n=500,000) and construct a new CMMS model, following the methods for the original CMMS study, with variable reduction and parsimony, achieved by backward elimination and the Akaike information stopping criterion. Model validation involved obtaining 1-year mortality estimates for a synchronous data set (n=250,000) and 1-year and 5-year mortality estimates for an asynchronous data set (n=250,000). We compared the performance with that of the original CMMS and the modified CMMS that we previously developed using RSC data.

Results: The initial model contained 22 conditions and our final model included 17 conditions. The conditions overlapped with those of the modified CMMS using the more extensive SNOMED CT list. For 1-year mortality, discrimination was high in both the derivation and validation data sets (Harrell C=0.92) and 5-year mortality was slightly lower (Harrell C=0.90). Calibration was reasonable following an adjustment for overfitting. The performance was similar to that of both the original and previous modified CMMS models.

XSL•FO
RenderX

Conclusions: The new modified version of the CMMS can be used on the GDPPR, a nationwide primary care data set of 54 million people, to enable adjustment for multimorbidity in predicting mortality in people in real-world vaccine effectiveness, pandemic planning, and other research studies. It requires 17 variables to produce a comparable performance with our previous modification of CMMS to enable it to be used in routine data using SNOMED CT.

(J Med Internet Res 2024;26:e56042) doi:10.2196/56042

KEYWORDS

pandemics; COVID-19; multimorbidity; prevalence; predictive model; discrimination; calibration; systematized nomenclature of medicine; computerized medical records; systems

Introduction

People with multimorbidity, defined by those with 2 or more long-term conditions (LTCs) [1-6], have complex needs and impose increasing demands on primary care services given the aging population. Multimorbidity is associated with reduced life expectancy [7], lower quality of life [8], and an increased risk of hospitalization and death due to COVID-19 [9]. In clinical trials, vaccination against COVID-19 showed reduced risk of hospitalization and death in all groups [10,11]. However, in real-world studies, people with multimorbidity benefited less from vaccination [12] and were at increased risk of mortality, morbidity, and hospitalization, compared to those without multimorbidity [13]. People with 5 or more LTCs had a more than 4-fold higher risk of severe COVID-19 outcomes than those with less than 5 LTCs [12].

Developing a single comorbidity measure is challenging [14]. The Charlson Comorbidity Index (CCI) is a commonly used tool to predict mortality over time [15]. However, CCI is based on hospital data, therefore, its applicability to primary care data is limited and not readily implementable [16]. The Cambridge Multimorbidity Score (CMMS) addressed this limitation and is an established measure of multimorbidity in primary care data. The original CMMS used 37 LTCs from routine primary care data in computerized medical records to predict the risk of primary care consultations, unplanned hospital admissions, and mortality [17]. It was developed and validated using the Clinical Practice Research Datalink [18] from the codes of Read version 2. However, Read version 2 is no longer used in England and is not updated since 2018 [19]. Additionally, the original CMMS model excluded people younger than 21 years of age, which somewhat restricted its applicability to the general population.

To overcome these limitations, we have already developed and validated a modified CMMS, replacing Read version 2 with the Systematized Nomenclature of Medicine clinical terms (SNOMED CT) [20] and using pseudonymized data from the Oxford Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) sentinel network of individuals aged 16 years or older [21]. Established in 1967, the RSC is an internationally renowned source of primary care data [22]. It has been used for influenza and respiratory disease monitoring for the last 50 years [23]. During the COVID-19 pandemic, with linkage to existing NHS (National Health service) England data sets, RSC data were also used to understand its epidemiology and assess vaccine effectiveness and safety [24-27]. This modified version of the CMMS was used to assess the real-world effectiveness of the Oxford-AstraZeneca COVID-19 vaccine in England (RAVEN) study, which was run on the RSC using linked data from NHS England [28].

The RAVEN study also used primary care data from the larger and nationwide General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR) data source maintained by NHS England, providing pseudonymized data for over 54 million people in England [29]. GDPPR is linked at the individual patient level to hospital, death, vaccine exposure, and test results. However, while substantial, its primary care data collection is incomplete. The primary care data were created from the existing list of conditions comprising 56,319 different SNOMED CT codes. This is a large number, but it is less than 20% of all SNOMED CT codes. These covered some clinical conditions well (eg, diabetes) and some less so (eg, psychoactive substance disorder). This study aimed to develop and validate a modified version of the CMMS, which could be used for the population aged 16 years and older in this new English NHS nationwide data set (GDPPR).

Methods

UK Primary Care Data

In the United Kingdom, each patient registers with a single general practitioner practice. Information about their primary care consultations, prescriptions, investigation results, and certified sickness and mortality data are recorded in computerized medical records systems. Each patient has a unique identifier, the NHS number, which allows data linkage with other data sets, including the hospital data, Hospital Episode Statistics, death certificate data provided by the Office for National Statistics and the NHS prescribing data set [30].

Data Sources

We used pseudonymized data from the RSC to construct and validate a revised version of CMMS based on the limited set of SNOMED CT codes (we refer to this as the GDPPR-modified CMMS). RSC data are stored in the Oxford RCGP Digital Informatics Hub (ORCHID) trusted research environment. The RCGP RSC extracts data from just under 2000 general practices in England [31] and provides a data set that is representative of individuals in England.

We applied the same analytical approach and the same inclusion criteria as in our previous study, where we developed and validated a new CMMS for RSC using the more extensive list of SNOMED CT codes (we refer to this as the RSC-modified CMMS) [21]. We included people aged 16 years or older on the index date registered with a practice for 12 months or longer.

XSL•FO
RenderX

Three separate data sets were sampled from the RSC as described previously (Figures S1 and S2 in Multimedia Appendix 1 [21]) and they are (1) derivation data set (n=500,000); (2) validation data set 1 (n=250,000) with the same study start and study end date as the derivation set (synchronous outcome); and (3) validation data set 2 (n=250,000) with 12-month outcome at a different time point to the derivation data set (asynchronous outcome), and 60-month outcome occurring at the same time point as the 12-month outcome of the derivation data set (synchronous outcome), as illustrated in Figure 1 [21]. These 3 data sets were generally comparable in terms of age, sex, number of conditions, and follow-up time.

Figure 1. Study design for the development and validation of the GDPPR-version of the CMMS, involving 3 cohorts, with index dates set at 12 months after their respective study start dates, and synchronous and asynchronous outcomes at 1 year and 5 years (adapted from Tsang et al [21], which is published under Creative Commons Attribution 4.0 International License). CMMS: Cambridge Multimorbidity Score; GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.

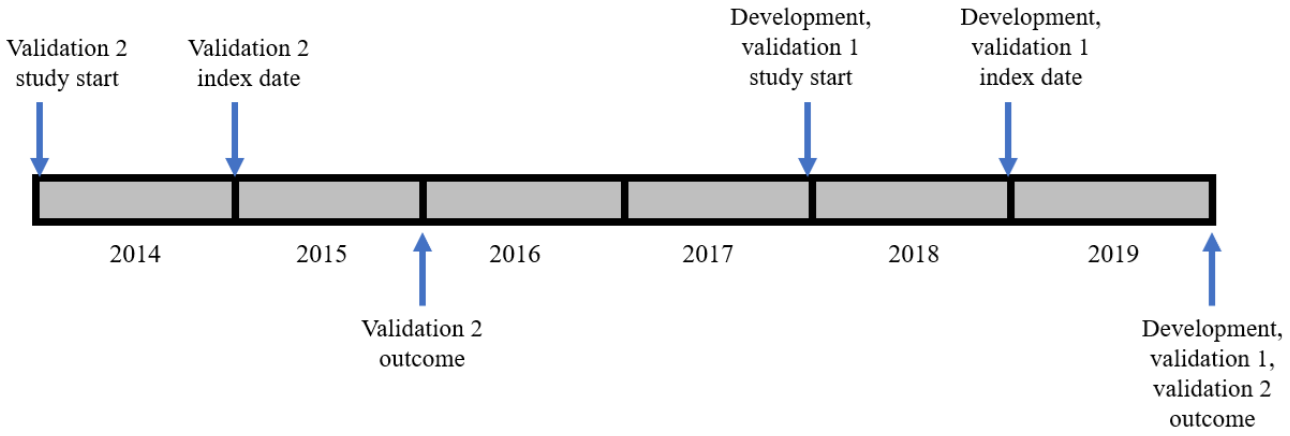

Curating and Selecting Individual CMMS Component Variables

In selecting conditions for a new modified CMMS model, we first considered all 37 conditions that were included in the original CMMS development and validation and used the same definitions and prescriptions [17]. Following the approach we previously developed for the RSC-modified CMMS with SNOMED CT [21], we carefully curated the conditions within the limited set of SNOMED CT codes in the GDPPR. We performed a confirmatory study concerning SNOMED CT coverage in the GDPPR by carrying out a statistical and clinical matching of the conditions between the GDPPR and RSC (Figure 2). A matching percentage for each condition was defined as the prevalence ratio, that is, the prevalence in the RSC data set but included in the GDPPR set of SNOMED CT codes divided by the prevalence included in the RSC set of SNOMED CT codes. Therefore, a ratio less than 100% indicated a higher prevalence of that condition within the RSC data set using its set of RSC SNOMED CT codes, compared to the prevalence using the GDPPR SNOMED CT list on the same data set. We set an 85% threshold for inclusion in the development of the GDPPR-modified CMMS model unless there was a clinical reason to accept a lower threshold.

As the RSC provides a data set that is representative of England, we assumed that the actual prevalence of each condition in the RSC data set is similar to that in the GDPPR data set, and therefore, the RCS data set provided a suitable environment to develop the GDPPR-modified CMMS. We developed this GDPPR-modified CMMS in the RSC data set because it offered a complete set of SNOMED CT codes for each clinical concept, and we could replicate the reduced data set within GDPPR and then compare the case finding with each approach.

XSL•FO
RenderX

Figure 2. Selection of candidate CMMS conditions for the GDPPR-modified CMMS from the 37 conditions included in the original CMMS and inclusion of conditions based on prevalence defined by both the SNOMED CT lists available in the GDPPR and RSC, or knowledge that the condition is known to have high predictive value. CMMS: Cambridge Multimorbidity Score; GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research; RSC: Research and Surveillance Centre; SNOMED CT: Systematized Nomenclature of Medicine clinical terms.

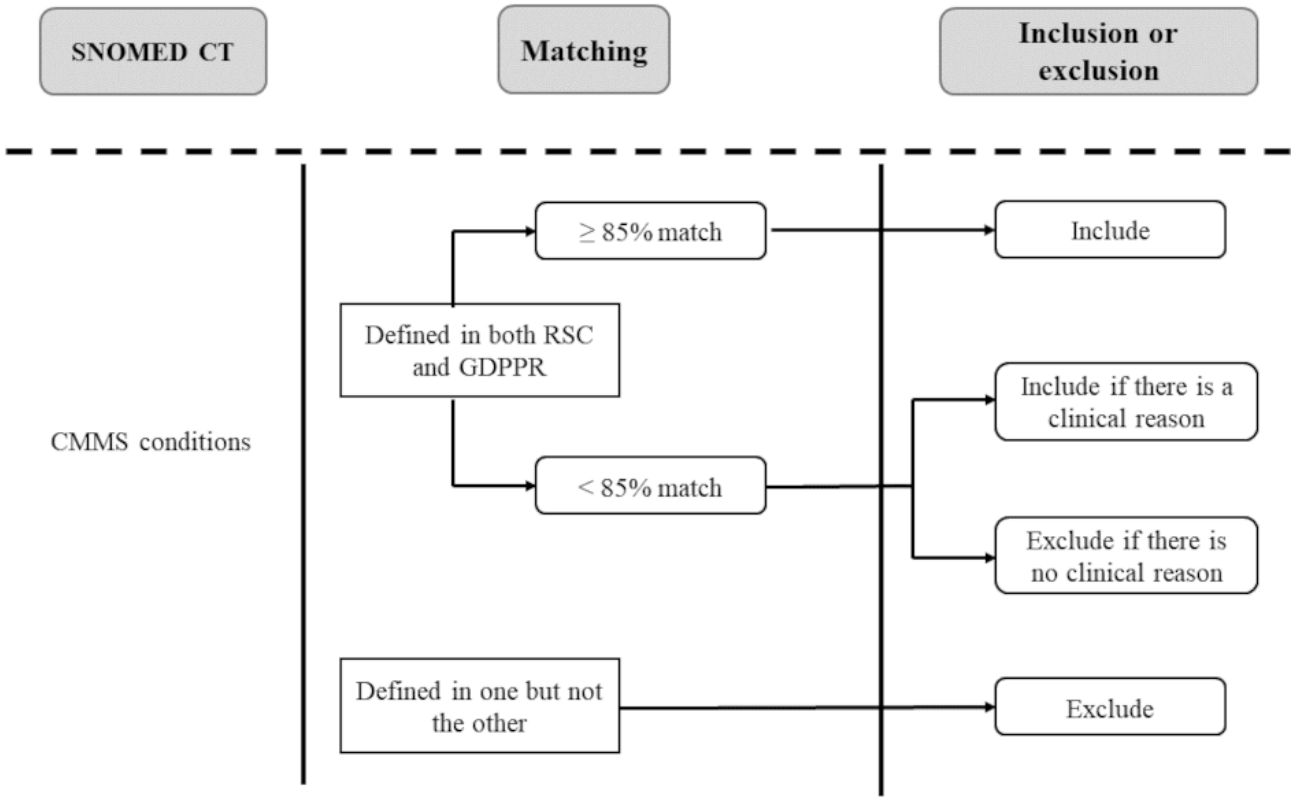

Statistical Analyses

Using the previously described method [21], we used time-to-mortality Cox proportional hazards models based on the development data set. We first included the conditions as binary indicators with sex and age (in decades) and a quadratic age term as covariates. We then carried out a variable reduction process via backward elimination and using the Akaike information criteria as the stopping criterion [32]. The goal was to be parsimonious with the number of variables needed for implementation. This was carried out using the "fastbw" function from the rms R package. Model performance evaluation was based on discrimination and calibration. Discrimination was assessed by the pseudo R^2, Somers D, and Harrell C [33]. Model calibration was evaluated by plotting a calibration curve and recalibration was carried out by resampling using cross-validation to correct for optimism or overfitting. This was implemented using the "calibrate" function in the rms R package. The model was developed, performance was evaluated on the derivation data set, and we then evaluated the performance of the models on the 2 validation data sets. All data preparation and analyses were conducted in R (version 4.1.0; R Core Team) [34], using the following R packages lme4 (version 1.1-27) [35], lubridate (version 1.7.10) [36,37], randomizr (version 0.20.0) [37], rms (version 6.2-0) [38], survival (version 3.2-11) [39,40], tableone (version 0.12.0) [41], and tidyverse (version 1.3.1) [42].

Ethical Considerations

This development of the CMMS for use in GDPPR was performed as part of the RAVEN study which received ethical approval (Integrate Research Application Service number 300259) and was approved by the Health Research Authority's Bromley Research Ethics Committee reference 21/HRA/1971, on October 8, 2021. NHS England hosts the national safe haven for patient data. The legal basis for this is Regulation 3 of the Health Service (Control of Patient Information Regulations) 2002 [43]. Pseudonymized data extracted from the practices are kept in a secured server at ORCHID, which is an NHS England policy-compliant trusted research environment (organization code EE133863-MSD-NDPCHS).

Results

Of the 37 conditions in the original CMMS model, 22 were included in developing the GDPPR-modified CMMS (Table 1). This involved 5 conditions based on clinical judgement—alcohol problems, chronic liver disease and viral hepatitis, stroke and transient ischemic attack, thyroid disorders, and dementia—which were included for clinical reasons alone (Figure 2), as they were likely to have a good predictive value, notwithstanding their lower prevalence ratios. All but thyroid disorders remained in the model after variable reduction in the final 17-condition model (Table 1).

XSL•FO
RenderX

Table 1. Inclusion of the original CMMS[a] set of 37 medical conditions in the 2 modified versions of the CMMS, for use in the RSC[b] and for use in the GDPPR[c], using the SNOMED CT[d] list available for the 2 respective data sets, ordered by prevalence of these conditions in the RSC data set.

Condition	Patients in RSC data set as defined by RSC SNOMED CT codes (n=7,555,767), n	Prevalence ratio (%)	Final RSC-modified model[e] (n=21)	Initial GDPPR-modified model (n=22)	Final GDPPR-modified model[e] (n=17)
Diabetes	412,960	100.0	✓	✓	✓
Chronic kidney disease	260,270	100.0	✓	✓	✓
Chronis sinusitis	103,638	100.0		✓	
Bronchiectasis	34,050	99.9		✓	
Atrial fibrillation	198,949	99.7	✓	✓	✓
Asthma currently treated	542,001	99.0		✓	
Chronic obstructive pulmonary disease	147,191	99.0	✓	✓	✓
Schizophrenia or bipolar disease	48,975	97.9	✓	✓	✓
Epilepsy	54,964	95.9	✓	✓	✓
Parkinsonism	17,156	95.8	✓	✓	✓
Constipation	123,999	95.7	✓	✓	✓
Hypertension	1,280,958	94.7		✓	
Learning disability	44,100	93.4	✓	✓	✓
Heart failure	95,828	93.0	✓	✓	✓
Cancer in the last 5 years	168,577	91.1	✓	✓	✓
Peripheral vascular disease	38,946	87.3	✓	✓	✓
Coronary heart disease	322,338	84.7		✓	✓
Dementia	80,156	84.2	✓	✓	✓
Thyroid disorders	420,681	83.3		✓	
Stroke and transient ischemic attack	182,979	81.0		✓	✓
Migraine	33,719	68.4			
Connective tissue disorder or rheumatoid arthritis	153,350	51.8			
Chronic liver disease and viral hepatitis	52,265	49.3	✓	✓	✓
Painful conditions	810,458	43.6	✓		
Alcohol problems	190,439	33.0	✓	✓	✓
Psoriasis or eczema	63,556	26.5			
Inflammatory bowel disease	50,370	25.6			
Anxiety or depression	919,962	25.1	✓		
Disorders of prostate	201,559	25.0	✓		
Blindness and low vision	88,398	17.5			
Diverticular disease of intestine	229,536	17.2			
Anorexia	52,742	16.7			
Hearing loss	581,804	12.9			
Peptic ulcer disease	98,487	9.5			
Irritable bowel syndrome	407,880	0.0	✓		
Psychoactive substance misuse	92,944	0.0	✓		

XSL•FO
RenderX

Condition	Patients in RSC data set as defined by RSC SNOMED CT codes (n=7,555,767), n	Prevalence ratio (%)	Final RSC-modified model[e] (n=21)	Initial GDPPR-modified model (n=22)	Final GDPPR-modified model[e] (n=17)
Multiple sclerosis	15,717	0.0	✓		

[a]CMMS: Cambridge Multimorbidity Score.

[b]RSC: Research and Surveillance Centre.

[c]GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.

[d]SNOMED CT: Systematized Nomenclature of Medicine clinical terms.

[e]After variable reduction.

There were few differences in the development and validation data sets in terms of age, sex, number of conditions and follow-up time (Table 2).

The prevalence of the 22 conditions in the model derivation data set is presented in Table 3. These prevalences and their rankings were generally similar to those reported in the original CMMS study [17], and our previous study [21]. There was only 1 exception, coronary heart disease, which ranked lower (7163/500,000, 1.4%; original CMMS study—15,887/300,000, 4.8%; and previous RSC study—15,887/300,000, 5.3%).

The model performance of the 22-condition and 17-condition models was almost identical and similar to those in the original CMMS study and the RSC-modified CMMS (Tables 4 and 5).

Table 2. Descriptive statistics of 3 data sets sampled from the RSC[a] for deriving and validating a modified version of the CMMS[b] for use in the GDPPR[c] data set, within the constraints of its limited set of SNOMED CT[d] codes.

	Derivation (2019)	Validation 1 (2019)	Validation 2 (2015)
Male, n (%)	247,807 (49.6)	124,514 (49.8)	123,541 (49.4)
Age at index date in years			
Mean (SD)	49.03 (19.29)	48.11 (19.09)	46.0 (19.34)
Range	16-95	16-95	16-95
65-84, n (%)	103,587 (22)	48,016 (0.21)	46,834 (0.20)
85 or older, n (%)	16,387 (4)	7,513 (0.03)	7,608 (0.03)
Number of conditions			
Mean (SD)	0.72 (1.19)	0.69 (1.17)	0.70 (1.0)
Range	0-11	0-11	0-11
0, n (%)	309,089 (62)	157,981 (63)	161,941 (65)
1, n (%)	101,547 (20)	49,463 (20)	47,530 (19)
2 or more, n (%)	89,364 (18)	42,556 (17)	40,529 (16)
Number of deaths in follow-up, n	5104	2392	2408/11,948
Mean follow-up time[e] (days), n	352.8	351.9	350.4/1538.0
Total person years[e,f], n	482,885.5	240,859.6	239,859.2/1,052,567
Mortality rate (per 1000 person years)[e], n	10.57	9.93	10.04/11.35

[a]RSC: Research and Surveillance Centre.

[b]CMMS: Cambridge Multimorbidity Score.

[c]GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.

[d]SNOMED CT: Systematized Nomenclature of Medicine clinical terms.

[e]1-year follow-up for validation 1 and 1- and 5-year follow-up for validation 2.

[f]Calculated as number of person-days divided by 365.25.

XSL•FO
RenderX

Table 3. Prevalence in individuals in the model derivation data set of the 22 candidate conditions in the GDPPR[a]-modified CMMS[b] model before variable reduction and weights for the final set of 17 conditions after variable reduction, with conditions ordered by prevalence.

Condition	Value (n=500,000), n (%)	Weight
Hypertension	98,849 (19.8)	N/A[c]
Asthma currently treated	36,951 (7.4)	N/A
Diabetes	33,312 (6.7)	0.2623
Thyroid disorders	28,891 (5.8)	N/A
Chronic kidney disease	23,145 (4.6)	0.1286
Atrial fibrillation	15,041 (3.0)	0.2779
Cancer in the last 5 years	13,059 (2.6)	1.1876
Chronic obstructive pulmonary disease	12,734 (2.5)	0.6638
Alcohol problems	12,132 (2.4)	0.5670
Stroke and transient ischemic attack	12,118 (2.4)	0.2299
Constipation	8698 (1.7)	0.5889
Chronis sinusitis	8195 (1.6)	N/A
Coronary heart disease	21,897 (1.4)	0.1201
Heart failure	7163 (1.4)	0.5022
Dementia	5884 (1.2)	0.9815
Epilepsy	4114 (0.8)	0.6714
Schizophrenia or bipolar disorder	3819 (0.8)	0.5621
Learning disability	2857 (0.6)	1.0992
Peripheral vascular disease	2963 (0.6)	0.3519
Bronchiectasis	2563 (0.5)	N/A
Chronic liver disease and viral hepatitis	1890 (0.4)	1.0844
Parkinsonism	1409 (0.3)	0.5339

[a]GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.
[b]CMMS: Cambridge Multimorbidity Score.
[c]N/A: not applicable.

Table 4. Model discrimination for the final RSC[a]-modified and GDPPR[b]-modified versions of the CMMS[c], after variable reduction, and compared with a full 37-condition model using RSC data and SNOMED CT[d].

	37-condition model [21]	Final RSC-modified 21-condition model [21]	Final GDPPR-modified 17-condition model
Pseudo R^2	0.153	0.153	0.140
Somers D	0.851	0.851	0.833
Harrell C (SE)			
Derivation	0.925 (0.002)	0.926 (0.002)	0.916 (0.002)
Validation 1	0.920 (0.004)	0.921 (0.004)	0.922 (0.003)
Validation 2, 1-year follow-up	0.920 (0.003)	0.920 (0.003)	0.915 (0.003)
Validation 2, 5-year follow-up	0.907 (0.002)	0.907 (0.002)	0.902 (0.001)

[a]RSC: Research and Surveillance Centre.
[b]GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.
[c]CMMS: Cambridge Multimorbidity Score.
[d]SNOMED CT: Systematized Nomenclature of Medicine clinical terms.

XSL•FO
RenderX

Table 5. Hazard ratios (95 CIs) of the predictors for the final RSC[a]-modified and GDPPR[b]-modified versions of the CMMS[c], after variable reduction, and compared with a full 37-condition model using RSC data and SNOMED CT[d].

	37-condition model HR[e] (95% CI) [21]	RSC-modified 21-condition model HR (95% CI) [21]	GDPPR-modified 17-condition model HR (95% CI)
Age (10 years)	1.22 (1.02-1.47)	N/A[f]	1.02 (1.01-1.04)
[Age_(10 years)]^2	1.05 (1.03-1.06)	1.06 (1.06-1.06)	1.00 (1.00-1.00)
Sex (Male)	1.33 (1.23-1.45)	1.34 (1.24-1.46)	1.14 (1.08-1.21)
Cancer in the last 5 years	3.31 (2.99-3.67)	3.33 (3.00-3.69)	3.28 (3.06-3.52)
Dementia	2.57 (2.33-2.84)	2.55 (2.32-2.82)	2.67 (2.47-2.88)
Alcohol problems	2.17 (1.84-2.55)	2.21 (1.88-2.60)	1.76 (1.53-2.03)
Multiple sclerosis	2.13 (1.32-3.44)	2.14 (1.33-3.46)	N/A
Chronic liver disease and viral hepatitis	1.98 (1.57-2.49)	1.99 (1.58-2.50)	2.96 (2.38-3.68)
Chronic obstructive pulmonary disease	1.96 (1.76-2.18)	2.02 (1.83-2.23)	1.94 (1.80-2.10)
Learning disability	1.88 (1.14-3.10)	1.89 (1.15-3.11)	3.00 (2.18-4.13)
Parkinsonism	1.71 (1.39-2.11)	1.73 (1.40-2.13)	1.71 (1.43-2.04)
Heart failure	1.66 (1.49-1.85)	1.66 (1.49-1.84)	1.65 (1.51-1.80)
Epilepsy	1.59 (1.25-2.02)	1.61 (1.27-2.04)	1.96 (1.62-2.37)
Schizophrenia or bipolar disorder	1.59 (1.22-2.06)	1.62 (1.25-2.10)	1.75 (1.42-2.17)
Psychoactive substance abuse	1.57 (1.20-2.04)	1.57 (1.20-2.04)	N/A
Painful condition	1.55 (1.42-1.68)	1.56 (1.44-1.69)	N/A
Constipation	1.47 (1.33-1.62)	1.47 (1.33-1.62)	1.80 (1.67-1.95)
Atrial fibrillation	1.39 (1.27-1.53)	1.40 (1.27-1.53)	1.32 (1.23-1.42)
Peripheral vascular disease	1.39 (1.07-1.81)	1.40 (1.08-1.82)	1.42 (1.24-1.63)
Anxiety or depression	1.38 (1.27-1.50)	1.38 (1.27-1.50)	N/A
Diabetes	1.31 (1.20-1.43)	1.34 (1.23-1.46)	1.30 (1.21-1.39)
Psoriasis or eczema	1.27 (1.03-1.57)	N/A	N/A
Chronic kidney disease	1.24 (1.14-1.35)	1.24 (1.14-1.35)	1.14 (1.06-1.21)
Anorexia or bulimia	1.22 (0.66-2.28)	N/A	N/A
Peptic ulcer	1.13 (0.98-1.30)	N/A	N/A
Bronchiectasis	1.11 (0.87-1.41)	N/A	N/A
Stroke and transient ischemic attack	1.11 (1.00-1.24)	N/A	1.26 (1.16-1.36)
Asthma currently treated	1.05 (0.93-1.18)	N/A	N/A
Hypertension	1.04 (0.96-1.13)	N/A	N/A
Thyroid disorder	1.03 (0.92-1.14)	N/A	N/A
Coronary heart disease	1.00 (0.91-1.09)	N/A	1.13 (1.05-1.21)
Chronic sinusitis	0.98 (1.57-2.49)	N/A	N/A
Rheumatoid arthritis	0.98 (0.85-1.12)	N/A	N/A
Blindness and low vision	0.96 (0.84-1.11)	N/A	N/A
Diverticular disease of intestine	0.92 (0.82-1.02)	N/A	N/A
Hearing loss	0.92 (0.95-1.00)	N/A	N/A
Disorder of the prostate	0.83 (0.74-0.93)	0.83 (0.74-0.93)	N/A
Irritable bowel syndrome	0.83 (0.71-0.95)	0.82 (0.71-0.94)	N/A
Inflammatory bowel disease	0.65 (0.43-0.97)	N/A	N/A
Migraine	0.59 (0.25-1.42)	N/A	N/A

XSL•FO
RenderX

[a]RSC: Research and Surveillance Centre.

[b]GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.

[c]CMMS: Cambridge Multimorbidity Score.

[d]SNOMED CT: Systematized Nomenclature of Medicine clinical terms.

[e]HR: hazard ratio.

[f]N/A: not applicable.

For 1-year mortality, discrimination was high in both the derivation and validation data sets (Harrell C=0.92) and for 5-year mortality, it was slightly lower (Harrell C=0.90). The model calibration displayed underprediction at lower risks (<60%), and the calibration improved with the adjustment for optimism or overfitting (Figure 3).

Figure 3. Calibration curve for the final 17 condition GDPPR-modified CMMS model. Black: observed; blue: optimism corrected; gray: ideal. Mean (SE) is 0.07 (0.9). Quantile is 0.009. CMMS: Cambridge Multimorbidity Score; GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research.

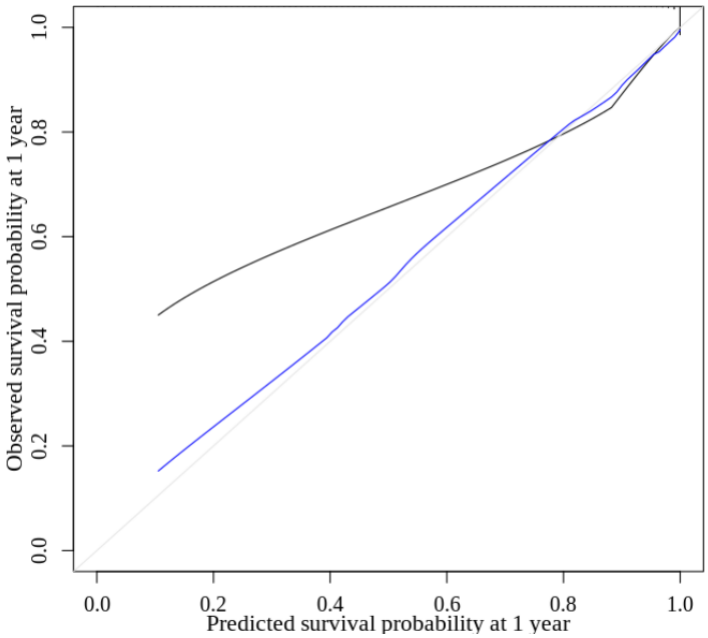

Discussion

Principal Findings

In this study, we developed and validated a modified version of a single measure of multimorbidity, CMMS, for use within a national data set created during the pandemic from all existing primary care data collections, GDPPR, and using its limited set of SNOMED CT codes. The initial model included 22 conditions from the set of 37 in the original CMMS model and the reduced 17-condition model showed an identical performance in predicting mortality to the 22-condition model and a similar performance compared to the original 37-condition CMMS and the previous modification which was based on an extensive SNOMED CT data set.

Interpretations and Implications

The GDPPR database remains available and is listed in the NHS Data Model and Dictionary [44]. It is now one of the data collections available through the NHS England Secure Data Environment [45], which was created as part of NHS England's Data Saves Lives policy, following the Goldacre report [46].

This new single measure of multimorbidity will help us measure vaccine effectiveness using GDPPR, NHS England's nationwide database. We will use this GDPPR-modified CMMS score in the RAVEN study to build on the existing evidence base [47,48]. Several observational studies have shown that the effectiveness of vaccination could be suboptimal in people with multimorbidity [49], and thus it is important to be able to explore and adjust for multimorbidity.

This tool may also be useful in a wider range of studies of people with multimorbidity, including vaccine and post-authorization safety studies. The risk of hospitalization, admission to intensive care unit beds, and mortality in people with multimorbidity are significantly higher than in the general population [50,51]. In an observational study of hospitalized patients in the United Kingdom with COVID-19, the crude mortality in people with multimorbidity, compared to single comorbidity, after adjusting for the demographic factors, was more than double (1492/3961, 37.7% vs 341/1971, 17.3%) [52]. It was estimated to reduce 63.5% (1905/3000) of deaths by prioritizing people with multimorbidity [53]. Therefore, people with multimorbidity were prioritized for vaccine rollout [54].

Our previous study showed that reducing the original CMMS variables from 37 to 21 did not compromise mortality predictability in people with multimorbidity [21]. In this study, we have demonstrated that the number of conditions can be reduced to 17 to match the data available in GDPPR and still can be a very good predictor of mortality.

XSL•FO
RenderX

Comparison With Prior Work

We focused on mortality and this is the outcome most often reported in development studies of comorbidity indices [55]. There are many published comorbidity indices and they vary according to their time of development (and thus the number of modifications), derivation population, conditions (predictors), prediction horizon, outcome predicted, and data source [52]. Historically, the mortality indices have been designed for people in hospitals, using secondary care coding systems and they have provided predictions of in-hospital mortality and mortality between 6 months and 5 years [55]. The GDPPR-modified CMMS is the latest adaptation to the original CMMS [17] for predicting mortality in primary care. The predictions for the CMMS and its modified versions are for the same prediction horizons of 1 year and 5 years. The previous modification adapted the CMMS to conditions defined by the internationally recognized SNOMED CT coding system [21], as the original CMMS was based on a population in the United States and conditions defined by the Read clinical terminology, which is no longer used in England. Both modifications of the CMMS have been developed on English populations with a lower minimum age compared to that for the original version (16 years as opposed to 21 years).

This GDPPR-modified version produces a multimorbidity index for mortality based on conditions defined by the limited SNOMED CT list of the GDPPR. The RSC provided the development data set for both modifications of the CMMS. The RCS has a complete set of SNOMED CT codes. Its data set includes people registered in a fraction of English general practices, and the assumption of this study is that the underlying prevalence of each CMMS condition in people in the RSC data set is the same as those in the larger GDPPR data set. While the RSC is recruited to be nationally representative, there may inevitably be differences [24].

The GDPPR includes the English primary care data used by the British Heart Foundation's Data Science Centre for their COVID-19 and cardiovascular diseases Consortium's work. Our version of CMMS could be deployed by this and other groups using GDPPR or underlying primary care (General Practice Extraction Service) data [56].

Strengths

The main strength of this study is that it built on our expertise in developing a version of CMMS that could be applied to routine clinical data recorded using SNOMED CT. This terminology is used internationally [21]. We overcame the limitations of the relatively limited number of SNOMED clinical terms in GDPPR and demonstrated that a 17-condition CMMS could run in the GDDPR data set of 54 million individuals.

Limitations

There are several limitations of this study. We only predicted the mortality risk and did not predict the hospitalization or intensive care unit admission risk. The conditions were a subset of those included in the original CMMS study, which arose from a review of multimorbidity literature at the time of its development. The nature of disease and treatments, as to population characteristics, will change over time. Hence, the new CMMS versions will need to be updated regularly, and this may involve adding conditions that were not included in the original CMMS. Although we split our initial data set randomly into development and validation data sets, we have performed simple temporal external validation [57]. A more robust form of external validation would involve investigating the generalizability to other countries (geographical validation) or other settings (domain validation), but neither is relevant in this case as we are considering a national data set, and we have developed the new CMMS using what we assume to be a representative sample of the adult English population. The generalizability could be tested further on other primary care data such as Clinical Practice Research Datalink [18], which provides a database of anonymized health records for another sample of English general practitioner practices.

Conclusions

This latest modification of the CMMS provides a new validated single multimorbidity measure, which was generated through a combination of unique access to data and expertise in validation. The RSC provided nationally representative and comprehensive primary care data. The study team had experience in developing a validated CMMS version to use within SNOMED CT. This combination meant that it was possible to develop and validate a new version of CMMS for use in the national English data set, the GDPPR. Our previous study showed that reducing the original CMMS variables from 37 to 21 did not compromise mortality predictability in people with multimorbidity [21]. In this study, we have demonstrated that the number of conditions can be reduced to 17 to match the data available in GDPPR and still can be a very good predictor of mortality. Therefore, researchers using this national database, or looking for a further reduced CMMS measure, can use this 17-component single measure of comorbidity.

The approach used in this study could also be applied in other contexts. Our approach has been to replicate a validated multimorbidity measure in a smaller, but complete and high data quality sentinel network database, the RSC. Within the RSC we could ensure the model performs as well as the one run on complete data [21]. Additionally, developing and validating this reduced CMMS model in the RSC required less processing time. This may make this reduced version more attractive to other users, should processing time be at a premium.

Acknowledgments

We would like to acknowledge that Ruby Tsang prepared the R script that we modified, Rashmi Wimalaratna helped with the data curation, and Julian Sherlock was also involved in preparing the data. This study was conducted to support the real-world effectiveness of the Oxford-AstraZeneca COVID-19 vaccine in England (RAVEN) study which was funded by AstraZeneca.

XSL•FO
RenderX

The sponsors were involved in the Systematized Nomenclature of Medicine (SNOMED) matching process (SV), and in reviewing and approving the paper (SV, WM, and ST).

Data Availability

The data sets generated and analyzed during this study are not publicly available as they have to be analyzed within a secure network, after appropriate training and subject to ethical approval, and this can be arranged by the corresponding author on reasonable request.

Authors' Contributions

SdL conceived this research project and played a supervisory role. XF was involved in extracting data from the Research and Surveillance Centre. DK curated the variables. DK compared the SNOMED CT lists and DK and SdL conducted the clinical sign off of conditions to be included in the analysis. KST, SV, and MJ were involved in the statistical analyses. SNA and FF were responsible for managing the project. DK and KST drafted the paper. All authors have reviewed and approved the final paper.

Conflicts of Interest

SdL is the principal investigator for real-world effectiveness of the Oxford-AstraZeneca COVID-19 vaccine in England (RAVEN; EUPAS43571) funded by AstraZeneca. SdL is also the Director of the Oxford-Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC), which is included in his academic role at the University of Oxford. He has received research funding through his University from AstraZeneca, GlaxoSmithKline (GSK), Lily, Moderna, Medical Science Division (MSD), Sanofi, Seqirus, and Takeda. He has also served as an advisory board member for AstraZeneca, GSK, Sanofi, Seqirus, and Pfizer. SV, WM and ST are employees of AstraZeneca and may own stock/shares. ST reports ownership of GSK stocks/shares.

Multimedia Appendix 1
Flowcharts showing the selection of practices and individuals for inclusion in the analysis (adapted from Tsang et al [21], which is published under Creative Commons Attribution 4.0 International License).
[DOCX File , 118 KB - jmir_v26i1e56042_app1.docx]

References

1. Aubert CE, Schnipper JL, Roumet M, Marques-Vidal P, Stirnemann J, Auerbach AD, et al. Best definitions of multimorbidity to identify patients with high health care resource utilization. Mayo Clin Proc Innov Qual Outcomes 2020;4(1):40-49 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2019.09.002] [Medline: 32055770]
2. Dambha-Miller H, Simpson G, Hobson L, Roderick P, Little P, Everitt H, et al. Integrated primary care and social services for older adults with multimorbidity in England: a scoping review. BMC Geriatr 2021;21(1):674 [FREE Full text] [doi: 10.1186/s12877-021-02618-8] [Medline: 34861831]
3. Hanlon P, Nicholl BI, Jani BD, Lee D, McQueenie R, Mair FS. Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality: a prospective analysis of 493 737 UK biobank participants. Lancet Public Health 2018;3(7):e323-e332 [FREE Full text] [doi: 10.1016/S2468-2667(18)30091-4] [Medline: 29908859]
4. Smith SM, Wallace E, O'Dowd T, Fortin M. Interventions for improving outcomes in patients with multimorbidity in primary care and community settings. Cochrane Database Syst Rev 2016;3(3):CD006560 [FREE Full text] [doi: 10.1002/14651858.CD006560.pub3] [Medline: 26976529]
5. Pearson-Stuttard J, Ezzati M, Gregg EW. Multimorbidity-a defining challenge for health systems. Lancet Public Health 2019;4(12):e599-e600 [FREE Full text] [doi: 10.1016/S2468-2667(19)30222-1] [Medline: 31812234]
6. Cassell A, Edwards D, Harshfield A, Rhodes K, Brimicombe J, Payne R, et al. The epidemiology of multimorbidity in primary care: a retrospective cohort study. Br J Gen Pract 2018;68(669):e245-e251 [FREE Full text] [doi: 10.3399/bjgp18X695465] [Medline: 29530918]
7. Chudasama YV, Khunti K, Gillies CL, Dhalwani NN, Davies MJ, Yates T, et al. Healthy lifestyle and life expectancy in people with multimorbidity in the UK biobank: a longitudinal cohort study. PLoS Med 2020;17(9):e1003332 [FREE Full text] [doi: 10.1371/journal.pmed.1003332] [Medline: 32960883]
8. Carretero-Bravo J, Ramos-Fiol B, Ortega-Martín E, Suárez-Lledó V, Salazar A, O'Ferrall-González C, et al. Multimorbidity patterns and their association with social determinants, mental and physical health during the COVID-19 pandemic. Int J Environ Res Public Health 2022;19(24):16839 [FREE Full text] [doi: 10.3390/ijerph192416839] [Medline: 36554719]
9. Chudasama YV, Zaccardi F, Gillies CL, Razieh C, Yates T, Kloecker DE, et al. Patterns of multimorbidity and risk of severe SARS-CoV-2 infection: an observational study in the U.K. BMC Infect Dis 2021;21(1):908 [FREE Full text] [doi: 10.1186/s12879-021-06600-y] [Medline: 34481456]
10. Graña C, Ghosn L, Evrenoglou T, Jarde A, Minozzi S, Bergman H, et al. Efficacy and safety of COVID-19 vaccines. Cochrane Database Syst Rev 2022;12(12):CD015477 [FREE Full text] [doi: 10.1002/14651858.CD015477] [Medline: 36473651]

XSL•FO
RenderX

11. Knoll MD, Wonodi C. Oxford-AstraZeneca COVID-19 vaccine efficacy. Lancet 2021;397(10269):72-74 [FREE Full text] [doi: 10.1016/S0140-6736(20)32623-4] [Medline: 33306990]
12. Agrawal U, Bedston S, McCowan C, Oke J, Patterson L, Robertson C, et al. Severe COVID-19 outcomes after full vaccination of primary schedule and initial boosters: pooled analysis of national prospective cohort studies of 30 million individuals in England, Northern Ireland, Scotland, and Wales. Lancet 2022;400(10360):1305-1320 [FREE Full text] [doi: 10.1016/S0140-6736(22)01656-7] [Medline: 36244382]
13. Lai FTT, Huang L, Chui CSL, Wan EYF, Li X, Wong CKH, et al. Multimorbidity and adverse events of special interest associated with Covid-19 vaccines in Hong Kong. Nat Commun 2022;13(1):411 [FREE Full text] [doi: 10.1038/s41467-022-28068-3] [Medline: 35058463]
14. Ho IS, Azcoaga-Lorenzo A, Akbari A, Black C, Davies J, Hodgins P, et al. Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies. Lancet Public Health 2021;6(8):e587-e597 [FREE Full text] [doi: 10.1016/S2468-2667(21)00107-9] [Medline: 34166630]
15. Fraccaro P, Kontopantelis E, Sperrin M, Peek N, Mallen C, Urban P, et al. Predicting mortality from change-over-time in the Charlson Comorbidity Index: a retrospective cohort study in a data-intensive UK health system. Medicine (Baltimore) 2016;95(43):e4973 [FREE Full text] [doi: 10.1097/MD.0000000000004973] [Medline: 27787358]
16. Drosdowsky A, Gough K. The Charlson Comorbidity Index: problems with use in epidemiological research. J Clin Epidemiol 2022;148:174-177. [doi: 10.1016/j.jclinepi.2022.03.022] [Medline: 35395393]
17. Payne RA, Mendonca SC, Elliott MN, Saunders CL, Edwards DA, Marshall M, et al. Development and validation of the Cambridge multimorbidity score. CMAJ 2020;192(5):E107-E114 [FREE Full text] [doi: 10.1503/cmaj.190757] [Medline: 32015079]
18. CPRD: UK data driving real-world evidence. Medicines & Healthcare products Regulatory Agency. URL: https://www.cprd.com/ [accessed 2023-07-04]
19. Retirement of Read version 2 and clinical terms version 3. NHS England. URL: https://digital.nhs.uk/services/terminology-and-classifications/read-codes [accessed 2023-07-04]
20. SNOMED CT. NHS England. URL: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct [accessed 2023-07-04]
21. Tsang RS, Joy M, Whitaker H, Sheppard JP, Williams J, Sherlock J, et al. Development of a modified Cambridge multimorbidity score for use with SNOMED CT: an observational English primary care sentinel network study. Br J Gen Pract 2023;73(731):e435-e442 [FREE Full text] [doi: 10.3399/BJGP.2022.0235] [Medline: 37130611]
22. Correa A, Hinton W, McGovern A, van Vlymen J, Yonova I, Jones S, et al. Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) sentinel network: a cohort profile. BMJ Open 2016;6(4):e011092 [FREE Full text] [doi: 10.1136/bmjopen-2016-011092] [Medline: 27098827]
23. de Lusignan S, Correa A, Smith GE, Yonova I, Pebody R, Ferreira F, et al. RCGP research and surveillance centre: 50 years' surveillance of influenza, infections, and respiratory conditions. Br J Gen Pract 2017;67(663):440-441 [FREE Full text] [doi: 10.3399/bjgp17X692645] [Medline: 28963401]
24. Leston M, Elson WH, Watson C, Lakhani A, Aspden C, Bankhead CR, et al. Representativeness, vaccination uptake, and COVID-19 clinical outcomes 2020-2021 in the UK Oxford-Royal College of General Practitioners Research and Surveillance Network: cohort profile summary. JMIR Public Health Surveill 2022;8(12):e39141 [FREE Full text] [doi: 10.2196/39141] [Medline: 36534462]
25. de Lusignan S, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, et al. Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance centre primary care network: a cross-sectional study. Lancet Infect Dis 2020;20(9):1034-1042 [FREE Full text] [doi: 10.1016/S1473-3099(20)30371-6] [Medline: 32422204]
26. Whitaker HJ, Tsang RSM, Byford R, Andrews NJ, Sherlock J, Pillai PS, et al. Pfizer-BioNTech and Oxford astrazeneca COVID-19 vaccine effectiveness and immune response amongst individuals in clinical risk groups. J Infect 2022;84(5):675-683 [FREE Full text] [doi: 10.1016/j.jinf.2021.12.044] [Medline: 34990709]
27. Tsang RS, Joy M, Byford R, Robertson C, Anand SN, Hinton W, et al. Adverse events following first and second dose COVID-19 vaccination in England, October 2020 to September 2021: a national vaccine surveillance platform self-controlled case series study. Euro Surveill 2023;28(3):2200195 [FREE Full text] [doi: 10.2807/1560-7917.ES.2023.28.3.2200195] [Medline: 36695484]
28. Meeraus W, Joy M, Ouwens M, Taylor KS, Venkatesan S, Dennis J, et al. AZD1222 effectiveness against severe COVID-19 in individuals with comorbidity or frailty: the RAVEN cohort study. J Infect 2024;88(4):106129 [FREE Full text] [doi: 10.1016/j.jinf.2024.106129] [Medline: 38431156]
29. COVID-19 General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR). NHS England. URL: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/gpes-data-for-pandemic-planning-and-research-gdppr [accessed 2023-07-04]
30. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract 2006;23(2):253-263. [doi: 10.1093/fampra/cmi106] [Medline: 16368704]

XSL•FO
RenderX

31. de Lusignan S, Jones N, Dorward J, Byford R, Liyanage H, Briggs J, et al. The Oxford Royal College of General Practitioners clinical informatics digital hub: protocol to develop extended COVID-19 surveillance and trial platforms. JMIR Public Health Surveill 2020;6(3):e19773 [FREE Full text] [doi: 10.2196/19773] [Medline: 32484782]
32. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974;19(6):716-723. [doi: 10.1109/TAC.1974.1100705]
33. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247(18):2543-2546. [Medline: 7069920]
34. R: a language and environment for statistical computing, Vienna, R Foundation for Statistical Computing. R Core Team. URL: https://www.R-project.org/ [accessed 2024-07-13]
35. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1):1-48.
36. Grolemund G, Wickham H. Dates and times made easy with lubridate. J Stat Soft 2011;40(3):1-25.
37. Coppock A, Cooper J. Randomizr: easy-to-use tools for common forms of random assignment and sampling. R Package. URL: https://CRAN.R-project.org/package=randomizr [accessed 2024-07-13]
38. Harrell FE. rms: Regression Modeling Strategies. URL: https://cran.r-project.org/web/packages/rms/index.html [accessed 2023-04-24]
39. Therneau TM. survival: Survival Analysis. URL: https://cran.r-project.org/web/packages/survival/index.html [accessed 2023-04-24]
40. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
41. Yoshida K, Bartel A. Create 'Table 1' to describe baseline characteristics with or without propensity score weights. tableone. URL: https://CRAN.R-project.org/package=tableone [accessed 2024-07-13]
42. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686.
43. Taylor MJ. Legal bases for disclosing confidential patient information for public health: distinguishing between health protection and health improvement. Med Law Rev 2015;23(3):348-374 [FREE Full text] [doi: 10.1093/medlaw/fwv018] [Medline: 25995294]
44. General Practice Data for Planning and Research (GPDfPR). NHS Data Model and Dictionary. URL: https://www.datadictionary.nhs.uk/GPDataSet/index.html [accessed 2023-07-04]
45. The NHS England Secure Data Environment. NHS Digital. URL: https://digital.nhs.uk/services/secure-data-environment-service [accessed 2023-07-04]
46. de Lusignan S, Leston M, Ikpoh M, Howsam G. Data saves lives: bottom-up, professionally-led endorsement would increase the chance of success. Br J Gen Pract 2022;72(724):512-513 [FREE Full text] [doi: 10.3399/bjgp22X720965] [Medline: 36302671]
47. Rotshild V, Hirsh-Raccah B, Miskin I, Muszkat M, Matok I. Comparing the clinical efficacy of COVID-19 vaccines: a systematic review and network meta-analysis. Sci Rep 2021;11(1):22777 [FREE Full text] [doi: 10.1038/s41598-021-02321-z] [Medline: 34815503]
48. Mills EJ, Reis G. Evaluating COVID-19 vaccines in the real world. Lancet 2022;399(10331):1205-1206 [FREE Full text] [doi: 10.1016/S0140-6736(22)00194-5] [Medline: 35303472]
49. Choi WS, Cheong HJ. COVID-19 vaccination for people with comorbidities. Infect Chemother 2021;53(1):155-158 [FREE Full text] [doi: 10.3947/ic.2021.0302] [Medline: 34409789]
50. Russell CD, Lone NI, Baillie JK. Comorbidities, multimorbidity and COVID-19. Nat Med 2023;29(2):334-343. [doi: 10.1038/s41591-022-02156-9] [Medline: 36797482]
51. McQueenie R, Foster HME, Jani BD, Katikireddi SV, Sattar N, Pell JP, et al. Multimorbidity, polypharmacy, and COVID-19 infection within the UK biobank cohort. PLoS One 2020;15(8):e0238091 [FREE Full text] [doi: 10.1371/journal.pone.0238091] [Medline: 32817712]
52. Agrawal U, Azcoaga-Lorenzo A, Fagbamigbe AF, Vasileiou E, Henery P, Simpson CR, et al. Association between multimorbidity and mortality in a cohort of patients admitted to hospital with COVID-19 in Scotland. J R Soc Med 2022;115(1):22-30 [FREE Full text] [doi: 10.1177/01410768211051715] [Medline: 34672832]
53. Ioannou GN, Green P, Fan VS, Dominitz JA, O'Hare AM, Backus LI, et al. Development of COVIDVax model to estimate the risk of SARS-CoV-2-related death among 7.6 million US veterans for use in vaccination prioritization. JAMA Netw Open 2021;4(4):e214347 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.4347] [Medline: 33822066]
54. Russo AG, Decarli A, Valsecchi MG. Strategy to identify priority groups for COVID-19 vaccination: a population based cohort study. Vaccine 2021;39(18):2517-2525 [FREE Full text] [doi: 10.1016/j.vaccine.2021.03.076] [Medline: 33824037]
55. Yurkovich M, Avina-Zubieta J, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. J Clin Epidemiol 2015;68(1):3-14. [doi: 10.1016/j.jclinepi.2014.09.010] [Medline: 25441702]
56. Abbasizanjani H, Torabi F, Bedston S, Bolton T, Davies G, Denaxas S, CVD-COVID-UK/COVID-IMPACT Consortium. Harmonising electronic health records for reproducible research: challenges, solutions and recommendations from a UK-wide COVID-19 research collaboration. BMC Med Inform Decis Mak 2023;23(1):8 [FREE Full text] [doi: 10.1186/s12911-022-02093-0] [Medline: 36647111]

XSL•FO
RenderX

57. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. external validation, model updating, and impact assessment. Heart 2012;98(9):691-698. [doi: 10.1136/heartjnl-2011-301247] [Medline: 22397946]

Abbreviations

CCI: Charlson Comorbidity Index
CMMS: Cambridge Multimorbidity Score
GDPPR: General Practice Extraction Service Data for Pandemic Planning and Research
LTC: long-term condition
NHS: National Health Service
ORCHID: Oxford Royal College of General Practitioners Digital Informatics Hub
RAVEN: real-world effectiveness of the Oxford-AstraZeneca COVID-19 vaccine in England
RCGP: Royal College of General Practitioners
RSC: Research and Surveillance Centre
SNOMED CT: Systematized Nomenclature of Medicine clinical terms

Edited by A Mavragani; submitted 03.01.24; peer-reviewed by PH Liao, A Azcoaga-Lorenzo, C Okusi; comments to author 29.02.24; revised version received 17.05.24; accepted 20.05.24; published 26.08.24.

Please cite as:
Kar D, Taylor KS, Joy M, Venkatesan S, Meeraus W, Taylor S, Anand SN, Ferreira F, Jamie G, Fan X, de Lusignan S
Creating a Modified Version of the Cambridge Multimorbidity Score to Predict Mortality in People Older Than 16 Years: Model Development and Validation
J Med Internet Res 2024;26:e56042
URL: https://www.jmir.org/2024/1/e56042
doi:10.2196/56042
PMID:

©Debasish Kar, Kathryn S Taylor, Mark Joy, Sudhir Venkatesan, Wilhelmine Meeraus, Sylvia Taylor, Sneha N Anand, Filipa Ferreira, Gavin Jamie, Xuejuan Fan, Simon de Lusignan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Combining Clinical-Radiomics Features With Machine Learning Methods for Building Models to Predict Postoperative Recurrence in Patients With Chronic Subdural Hematoma: Retrospective Cohort Study

Cheng Fang[1*], MD; Xiao Ji[2*], MM; Yifeng Pan[3*], MSE; Guanchao Xie[4*], MM; Hongsheng Zhang[4*], MM; Sai Li[4], MM; Jinghai Wan[1,4], MD

[1]Department of Neurosurgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Department of Neurosurgery, The First Affiliated Hospital of Anhui Medical University, Anhui Public Health Clinical Center, Hefei, China

[3]The School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, China

[4]Department of Neurosurgery, The Second Affiliated Hospital of Anhui Medical University, Anhui Medical University, Hefei, China

[*]these authors contributed equally

Corresponding Author:
Jinghai Wan, MD
Department of Neurosurgery
National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital
Chinese Academy of Medical Sciences and Peking Union Medical College
No. 17 Nanli, Panjiayuan
Chaoyang District
Beijing, 100021
China
Phone: 86 13426261848
Email: wanjinghai@sina.com

Abstract

Background: Chronic subdural hematoma (CSDH) represents a prevalent medical condition, posing substantial challenges in postoperative management due to risks of recurrence. Such recurrences not only cause physical suffering to the patient but also add to the financial burden on the family and the health care system. Currently, prognosis determination largely depends on clinician expertise, revealing a dearth of precise prediction models in clinical settings.

Objective: This study aims to use machine learning (ML) techniques for the construction of predictive models to assess the likelihood of CSDH recurrence after surgery, which leads to greater benefits for patients and the health care system.

Methods: Data from 133 patients were amassed and partitioned into a training set (n=93) and a test set (n=40). Radiomics features were extracted from preoperative cranial computed tomography scans using 3D Slicer software. These features, in conjunction with clinical data and composite clinical-radiomics features, served as input variables for model development. Four distinct ML algorithms were used to build predictive models, and their performance was rigorously evaluated via accuracy, area under the curve (AUC), and recall metrics. The optimal model was identified, followed by recursive feature elimination for feature selection, leading to enhanced predictive efficacy. External validation was conducted using data sets from additional health care facilities.

Results: Following rigorous experimental analysis, the support vector machine model, predicated on clinical-radiomics features, emerged as the most efficacious for predicting postoperative recurrence in patients with CSDH. Subsequent to feature selection, key variables exerting significant impact on the model were incorporated as the input set, thereby augmenting its predictive accuracy. The model demonstrated robust performance, with metrics including accuracy of 92.72%, AUC of 91.34%, and recall of 93.16%. External validation further substantiated its effectiveness, yielding an accuracy of 90.32%, AUC of 91.32%, and recall of 88.37%, affirming its clinical applicability.

Conclusions: This study substantiates the feasibility and clinical relevance of an ML-based predictive model, using clinical-radiomics features, for relatively accurate prognostication of postoperative recurrence in patients with CSDH. If the

XSL•FO
RenderX

model is integrated into clinical practice, it will be of great significance in enhancing the quality and efficiency of clinical decision-making processes, which can improve the accuracy of diagnosis and treatment, reduce unnecessary tests and surgeries, and reduce the waste of medical resources.

(J Med Internet Res 2024;26:e54944) doi:10.2196/54944

KEYWORDS

chronic subdural hematoma; convolutional neural network; machine learning; neurosurgery; radiomics; support vector machine

Introduction

Background

Chronic subdural hematoma (CSDH) is a prevalent neurosurgical pathology, disproportionately affecting middle-aged and older adults. Epidemiological data indicate incidence rates of 13.5/100,000, escalating to 58.1/100,000 in individuals aged ≥65 years [1,2]. Manifestations commonly include headache, nausea, vomiting, and diplopia, indicative of elevated intracranial pressure. Diagnosis is generally confirmed through cranial computed tomography (CT) or magnetic resonance imaging. Established as a medical condition since 1857, surgical intervention remains a proven, efficacious treatment modality for CSDH. However, postoperative recurrence serves as a critical metric for evaluating surgical success [3]. Such recurrence imposes not only physical suffering on patients but also accentuates financial burden on families and health care systems. For patients who are older, have a history of multiple surgeries, or have other complications, an accurate predictive tool can help physicians identify these high-risk patients in advance, allowing for a more precise treatment and follow-up plan. Each patient with CSDH has a different condition and clinical background, and thus requires a personalized treatment plan that can assess the risk of postoperative recurrence based on the patient's specific clinical information (age, sex, medical history, symptoms, signs, and imaging manifestations). Given the increasing strain on medical resources, optimizing the allocation and use of medical resources and improving the operational efficiency of hospitals has become an important issue. Hence, in the above clinical context, the development of a predictive tool for postoperative recurrence risk is integral for informed clinical decision-making and optimized treatment outcomes, which can bring greater benefits to both patients and the health care system.

Recent advancements in computer technology have facilitated the construction of predictive models anchored on clinically pertinent data. Machine learning (ML) has emerged as a particularly robust paradigm, capable of delineating complex, nonlinear relationships between variables and outcomes. A plethora of studies substantiate the impressive levels of accuracy and reliability achieved through ML applications [4-7]. This study uses 4 ML methodologies—convolutional neural networks (CNNs), support vector machines (SVMs), random forest (RF), and linear regression (LR)—each enjoying widespread academic acceptance and demonstrated applicability in predictive research across various domains, including health care and food sciences [6,8,9].

By surveying existing literature, it is evident that ML models integrated with radiomics are garnering increased scholarly attention [10,11]. Radiomics constitutes a novel approach in medical image analysis, principally centered on quantitative feature extraction. This technique transforms medical imagery into high-dimensional structures, facilitating the comprehensive analysis of regions of interest (ROIs) in conjunction with relevant clinical, diagnostic, and prognostic data. A typical radiomics workflow encompasses stages of image acquisition, reconstruction, preprocessing and processing, feature extraction, selection, and eventually, classification or regression modeling [12]. Although previous research has melded radiomics and ML for diagnostic and prognostic applications in other medical specialties such as dermatology, oncology, and cardiology [13-15], studies targeting CSDH remain comparatively scant.

Objective

The objective of this investigation is to amalgamate ML algorithms with radiomics and clinical variables for the construction of a predictive model aimed at gauging the risk of CSDH recurrence after surgery. The study will rigorously compare various methodologies and models to identify the most efficacious predictive framework for CSDH recurrence.

Methods

Ethical Considerations

The retrospective study was approved (approval no PJ-YX2024-021) by the Ethics Committee of the First Affiliated Hospital of Anhui Medical University (Anhui Public Health Clinical Center).

Participants

We compiled clinical and radiological data from patients diagnosed with CSDH who were treated at the neurosurgery department of the Second Affiliated Hospital of Anhui Medical University between May 2012 and May 2022. The inclusion criteria were as follows: (1) confirmed clinical diagnosis of CSDH; (2) participants who underwent surgical intervention, either Burr hole craniostomy or craniotomy; and (3) availability of comprehensive clinical records, encompassing treatment histories, preoperative and postoperative imaging examinations, laboratory analyses, among other pertinent data. Exclusion criteria included the following: (1) patients who exhibited symptomatic improvement via pharmacological intervention, obviating the need for surgical treatment; (2) any history of prior neurosurgical procedures that could potentially induce CSDH; and (3) cases with incomplete follow-up data or where recurrence after surgery was undetermined. Following these criteria, 133 patients were incorporated into the study, and there are no missing values for all case data in this study. The process of patient selection and enrollment is delineated in Figure 1.

XSL•FO
RenderX

Furthermore, an external validation set was generated by screening data from 20 patients with CSDH who underwent treatment at the First Affiliated Hospital of Anhui Medical University.

Figure 1. Patient selection and enrollment process. CSDH: chronic subdural hematoma.

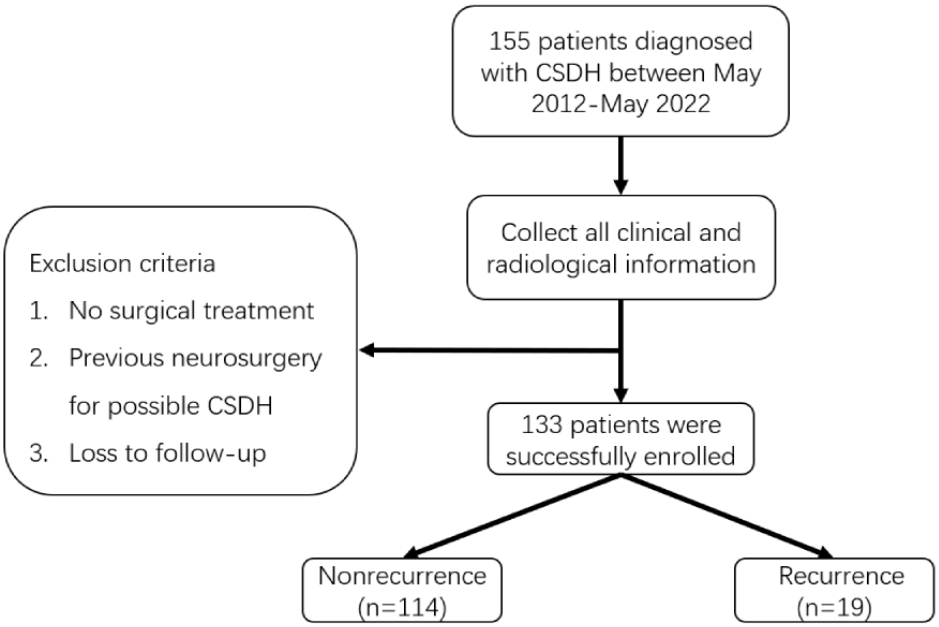

Clinical Data

Upon a rigorous review of existing literature and consultations with experienced neurosurgeons, specific data parameters were established for model construction. These parameters were extracted from the electronic case files of the participants and included: patient demographics (age and sex), pertinent clinical history (smoking or drinking habits, prior medical history, history of head trauma, and history of antiplatelet or anticoagulant therapy), preoperative clinical grading based on the Markwalder Grading Scale (MGS), and duration of hospitalization. Imaging data encompassed variables such as the location of the CSDH (unilateral or bilateral), hematoma classification, preoperative hematoma volume, preoperative midline shift, postoperative midline improvement, and cranial CT scans. Subsequent to the follow-up period, 19 patients exhibited postoperative recurrence. These patients were bifurcated into 2 cohorts: those with recurrence and those without recurrence. Clinical indices used in model construction are elaborated in Table 1. Continuous variables were subjected to 2-tailed *t* test analyses and are represented as mean (SD). Categorical variables were analyzed using chi-square test and are conveyed as percentages. Statistical analysis was performed using SPSS Statistics (version 27.0; IBM Corp).

In Textbox 1, we delineate the grading criteria and definitions associated with the MGS [16]. Initially proposed in the 1980s, extensive research has validated the MGS as a robust metric for evaluating postoperative neurological recovery and prognosis in patients with CSDH. Specifically, a grade of 0 indicates normal neurological function, grades 1 to 2 signify good neurological function, and grades 3 to 4 represent poor neurological function.

XSL•FO
RenderX

Table 1. Clinical variables used to construct the model.

Variables	Nonrecurrence (n=114)	Recurrence (n=19)	*P* value
Age (y), mean (SD)	70.1 (9.4)	73.7 (8.3)	.13
Sex, n (%)			.50
Male	91 (79.8)	17 (89)	
Female	23 (20.2)	2 (11)	
Smoking or drinking , n (%)			.12
Yes	56 (49.1)	13 (68)	
No	58 (50.9)	6 (32)	
Hypertension , n (%)			.10
Yes	43 (37.7)	11 (58)	
No	71 (62.3)	8 (42)	
Cerebral infarction, n (%)			.02
Yes	11 (9.6)	6 (32)	
No	103 (90.4)	13 (68)	
History of head trauma, n (%)			.34
Yes	73 (64.0)	10 (53)	
No	41 (36.0)	9 (47)	
History of antiplatelet or anticoagulation , n (%)			.003
Yes	10 (8.8)	7 (37)	
No	104 (91.2)	12 (63)	
Markwalder Grading Scale, n (%)			.004
0	1 (0.9)	0 (0)	
1	38 (33.3)	0 (0)	
2	56 (49.1)	12 (63)	
3	18 (15.8)	7 (37)	
4	1 (0.9)	0 (0)	
Length of stay in hospital (d), mean (SD)	14.0 (4.3)	14.5 (3.2)	.64
CSDH[a] location, n (%)			.01
Unilateral	104 (91.2)	13 (68)	
Bilateral	10 (8.8)	6 (32)	
Classification of hematoma （%）			.02
Homogeneous	50 (43.9)	2 (11)	
Laminar	12 (10.5)	1 (5)	
Separated	35 (30.7)	8 (42)	
Trabecular type	17 (14.9)	8 (42)	
Preoperative hematoma volume (mL), mean (SD)	103.4 (25.0)	127.1 (49.3)	.05
Preoperative midline shift (cm), mean (SD)	1.1 (0.3)	0.9 (0.3)	.01
Postoperative midline improvement (cm), mean (SD)	0.6 (0.2)	0.4 (0.2)	.01

[a]CSDH: chronic subdural hematoma.

XSL•FO
RenderX

Textbox 1. Descriptions of the categories of the Markwalder Grading Scale [16].

Grading score and description

- Grade 0: neurologically normal
- Grade 1: alert and orientated: absence of mild symptoms such as headache, or mild neurological deficit such as reflex asymmetry
- Grade 2: drowsy or disorientated, or variable neurological deficit such as hemiparesis
- Grade 3: stuporous, but responding appropriately to noxious stimuli, several focal signs such as hemiplegia
- Grade 4: comatose with absent motor responses to painful stimuli, decerebrate or decorticate posturing

Image Data

Complementing the comprehensive clinical data set, preoperative cranial CT scans were acquired for all enrolled participants. Hematoma images were systematically categorized into 4 distinct types: homogeneous, laminar, separated, and trabecular. Representative CT scans for each category are furnished in Figure 2. This taxonomic approach to hematoma classification was initially delineated by Nakaguchi et al [17]. Their work posited that hematomas with more irregular structures correlated with elevated recurrence rates. Subsequently, this classification schema has been integrated into a scoring system aimed at assessing recurrence risk [18]. All procured CT images were stored in DICOM format.

Figure 2. Classification of hematoma.

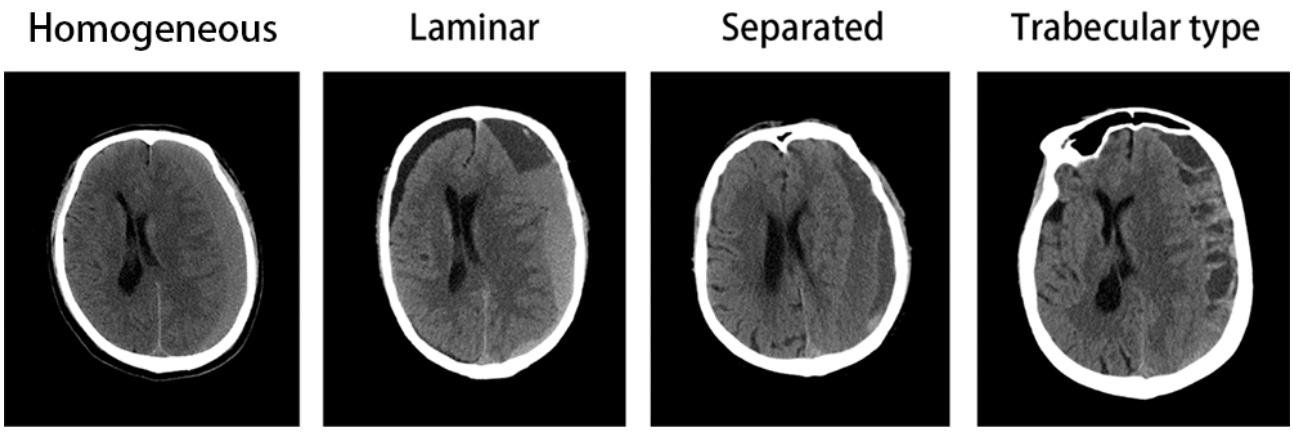

Radiomics

Hematoma segmentation was executed through semiautomatic techniques using the 3D Slicer open-source software platform (version 4.10). The ROI, specifically the hematoma, was further segmented using the PyRadiomics package, an open-source plug-in available on the 3D Slicer platform (Figure 3). The use of open-source software enabled the direct computation of 3D features without the necessity for slice-wise combination or averaging. A total of 107 radiomic features were manually extracted from each patient's CT images. These features were allocated to 7 distinct feature categories: 18 were first-order statistics, 14 were shape based, 24 were derived from gray-level co-occurrence matrices, 16 were derived from gray-level run-length matrices, 16 were derived from gray-level size-zone matrices, 5 were derived from neighboring gray tone difference matrices, and 14 were derived from gray-level dependence matrices. This conversion from image-based to data-driven features optimizes the data set for subsequent computational analyses and research endeavors.

XSL•FO
RenderX

Figure 3. Hand-delineated regions of interest.

Modeling

Overview

Neuroscientists participated comprehensively in this research, inclusive of data collection and categorical predictive model development. To safeguard patient confidentiality and adhere to the ethical guidelines stipulated by the ethics committee, patient data were anonymized through numerical coding. This approach also ensured the clinical applicability of the study's findings. We complied with Items of reporting predictive models in biomedical research in this study (Multimedia Appendix 1).

Four ML algorithms were used for model development: CNN, SVM, RF, and LR. Hyperparameter tuning was conducted via a grid search algorithm to optimize model performance. The main components of the CNN structure include an input layer, 3 convolution-pooling layers, 1 flat layer, 2 fully connected layers, and 1 output layer. Textboxes 2 and 3 detail the parameter configurations of the CNN and SVM. The data set was stratified into a training set, comprising 69.9% (93/133) of the samples, and a test set accounting for the remaining 30.1% (40/133). Given the data set's limited sample size, 5-fold cross-validation was executed on the training set to ensure robustness and validity. This cross-validation technique is a standard practice in ML for its ability to produce reliable performance metrics, mitigate the risks of overfitting and underfitting, and assess the model's generalization capability. The models were developed using the scikit-learn framework and implemented in the Python 3.9 programming environment.

Textbox 2. Setting hyperparameters for convolutional neural network.

Hyperparameter and setting

- Activation function: rectified linear unit
- Optimizer: Adam
- Batch size: 64
- Dropout: 0.5
- Loss: binary cross-entropy

XSL•FO
RenderX

Textbox 3. Setting hyperparameters for support vector machine.

Hyperparameter and setting

- Kernel: radial basis function
- Gamma: 0.01
- C: 50

Input Data Set Selection

Multiple input data set configurations were assessed to optimize the predictive model. Initial models were constructed using 3 distinct types of data: clinical data, CT images, and radiomics from the enrolled patient cohort. These data types were then aggregated in 2 specific combinations to generate models based on mixed input data, specifically, clinical-CT and clinical-radiomics features. Furthermore, a composite model using clinical data, CT images, and histological imaging was also developed. Comparative analysis of these data set configurations was performed to identify the algorithms and input data sets most conducive for predictive modeling.

Feature Selection

Overview

Our evaluation indicated that the SVM model, when configured with the clinical-radiomics data set, demonstrated superior predictive efficacy. However, the complexity of the input variables compromised both the model's performance and computational efficiency, thereby not meeting the study's predefined objectives. To address this, a feature selection strategy was used with the principal aim of isolating the most impactful variables. This was anticipated to enhance model performance, increase computational efficiency, and minimize algorithmic complexity. For this task, the recursive feature elimination (RFE) method was selected, supported by references [19-21]. Specific steps for the RFE implementation will be outlined in the subsequent sections.

Initial Feature Subset Evaluation

The composite data set of radiomic and clinical variables serves as the initial feature subset for the SVM model. Each feature's importance is quantified, and the classification accuracy of this initial feature set is assessed through cross-validation techniques.

Iterative Feature Removal and Recalculation

he least impactful feature is excised from the current feature subset, creating a modified feature set. This new set is subsequently input into the SVM model. Feature importance is recalculated and the modified feature subset's classification accuracy is evaluated using cross-validation methods.

Optimizing Feature Selection

The procedure delineated in step 2 is recursively applied until no features remain in the subset. Through this iterative process, a total of 15 distinct feature subsets are generated, each comprising a varying number of features. The feature subset yielding the highest classification accuracy is identified as the optimal feature combination for the predictive model.

Clinical Settings and Modeling Background

The clinical settings for our predictive model include the following: (1) Facility type—it is mainly applied to neurosurgery wards in general hospitals or neurosurgery specialty hospitals, which are capable of handling complex neurosurgical procedures. (2) Size—it is more suitable to be implemented in large- or medium-sized hospitals because these hospitals usually have more case data and experience, which is conducive to the training and validation of the model, and it can be generalized to smaller hospitals after it passes the clinical practice. The modeling background of the prediction model includes the following: (1) data duration—long-term clinical data, covering relevant case information over the past 10 years, are needed to ensure the stability and accuracy of the model. (2) Data characteristics—the data come from multiple sources, including medical records, imaging studies, and laboratory tests, reflecting the multiple and complex factors affecting recurrence. (3) Modeling purpose—to improve the accuracy of recurrence prediction through ML, help physicians develop more personalized treatment plans, and optimize the allocation of hospital resources. In summary, the clinical environment of the target prediction model is mainly set in the neurosurgical wards of large- or medium-sized hospitals, and the modeling background involves long-term clinical data collection and analysis, aiming to improve prediction accuracy and optimize the allocation of medical resources.

Results

Overview

The primary objective of this study was to streamline the clinical application of the predictive model, specifically by enabling direct input of cranial CT images for generating predictive outcomes. Contrary to expectations, models using CT images as the sole input data type demonstrated suboptimal performance across all evaluation metrics, irrespective of the ML algorithm used. Moreover, models incorporating both clinical data and CT images as input data yielded prediction outcomes significantly inferior to those relying solely on clinical data. Upon using 3D Slicer for radiological feature extraction from the CT images, the resultant model performance exhibited notable improvement. This underscores the inadequacy of current ML algorithms in directly using CT images for clinical research; while predictive results can be generated, they remain unsatisfactory (Figure 4 and Table 2). Consequently, the study abandoned the notion of using image-based input data. In contrast, radiomics data are already a distillation and abstraction of the CT image information, and using the original image as input again may lead to information redundancy and even introduce noise, thus affecting the performance of the model. From the perspective of computational efficiency, direct

XSL•FO
RenderX

processing of raw CT images usually requires more computational resources and time, whereas radiomics data, as a more compact and higher-level feature representation, can significantly improve the training and prediction speed of the model. Therefore, in this study, we did not choose to try radiomics data+CT images as input data for model validation. Instead, an exploration into the viability of using clinical data, radiomic features, or a combination thereof as input data sets was conducted to optimize predictive model performance.

Figure 4. Classification results of different data sets in 4 machine learning models. ACC: accuracy; AUC: area under the curve; CNN: convolutional neural network; CT: computed tomography; LR: linear regression; RF: random forest; SVM: support vector machine.

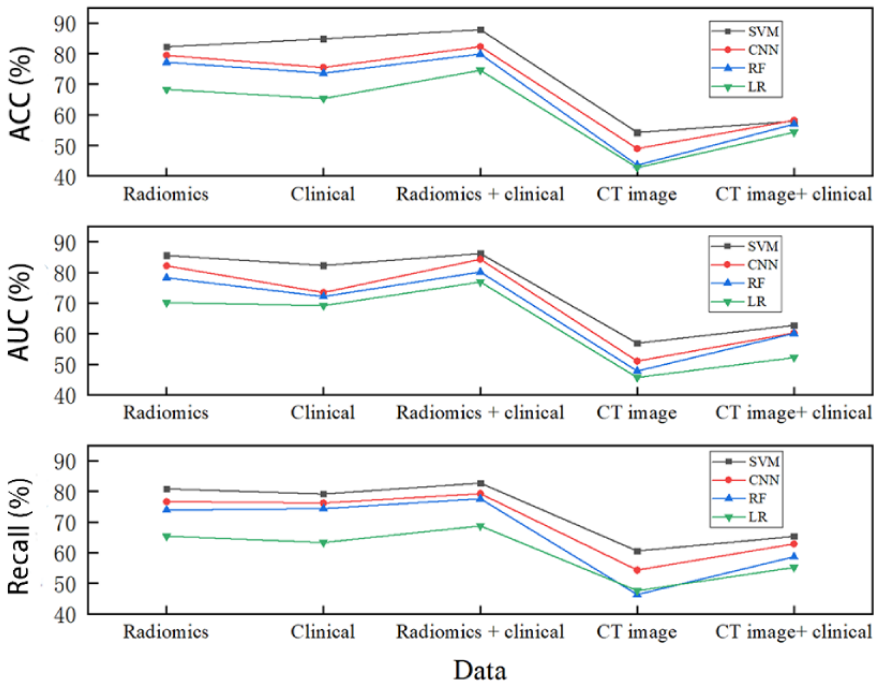

XSL•FO
RenderX

Table 2. Classification results of different data sets in 4 machine learning models.

Model	Radiomics	Clinical	Clinical-radiomics	CT[a] image	CT image-clinical
SVM[b]					
Accuracy (%)	82.19	84.82	87.79	54.27	57.86
AUC[c] (%)	85.57	82.29	86.13	56.91	62.75
Recall (%)	80.82	79.16	82.70	60.54	65.32
CNN[d]					
Accuracy (%)	79.42	75.44	82.27	48.95	58.19
AUC (%)	82.16	73.45	84.32	51.07	60.24
Recall (%)	76.64	76.18	79.23	54.29	62.85
RF[e]					
Accuracy (%)	77.13	73.58	79.85	43.55	56.97
AUC (%)	78.27	72.16	80.13	47.82	60.11
Recall (%)	73.89	74.37	77.62	46.31	58.63
LR[f]					
Accuracy (%)	68.32	65.32	74.56	42.75	54.32
AUC (%)	70.13	69.17	76.91	45.68	52.19
Recall (%)	65.34	63.28	68.76	47.54	55.13

[a]CT: computed tomography.
[b]SVM: support vector machine.
[c]AUC: area under the curve.
[d]CNN: convolutional neural network.
[e]RF: random forest.
[f]LR: linear regression.

Predictive Model Evaluation

This study uses accuracy, AUC, and recall as evaluative metrics for the predictive models, with corresponding results delineated in Figure 4. Accuracy serves as a direct indicator of the model's consistency in aligning predictive and actual outcomes. Specifically, accuracy represents the ratio of correctly classified samples to the overall sample pool, offering both intuitive understanding and straightforward implementation. It is principally used to assess the model's ability to accurately categorize target variables in the predictive outcomes. However, accuracy possesses limitations as it exclusively considers the classification of positive samples, thereby omitting negative samples. This lack of comprehensiveness limits accuracy's capacity to measure the overlap between predicted and true outcomes. To address this limitation, AUC is incorporated as it holistically considers both positive and negative samples, thereby providing a more nuanced evaluation of model performance. Recall, another metric used, is particularly pertinent given the study's objective to predict the recurrence of CSDH. Recall quantifies the model's proficiency in accurately identifying positive samples, focusing on true positive cases. It is especially vital for this study, as it emphasizes the model's ability to correctly predict patient recurrence. Distinct from accuracy and AUC, recall remains unaffected by the selection of a decision threshold, rendering it more appropriate for comparing various models, particularly when the decision threshold is ambiguous or challenging to ascertain. In summary, accuracy, AUC, and recall are deployed as multifaceted evaluative metrics to assess the predictive models constructed in this study.

In a comprehensive evaluation across all designated metrics—accuracy, AUC, and recall—the SVM model consistently outperformed the CNN, RF, and LR models. This was observed irrespective of the input data set used, be it clinical data, radiomics features, or a hybrid of both. For models using radiomics features, SVM demonstrated improvements of 2.77%, 5.06%, and 13.87% in accuracy; 3.41%, 7.3%, and 15.37% in AUC; and 4.18%, 6.93%, and 15.48% in recall compared to CNN, RF, and LR models, respectively. Similarly, when clinical data served as the input, SVM enhanced accuracy by 9.38%, 11.24%, and 19.5%; AUC by 8.84%, 10.13%, and 13.12%; and recall by 2.98%, 4.79%, and 15.88% in comparison to CNN, RF, and LR models, respectively. Interestingly, a combination of clinical and radiomics features as input to the SVM model resulted in further performance gains: accuracy improved by 5.52%, 7.94%, and 13.23%; AUC by 1.81%, 6%, and 9.22%; and recall by 3.47%, 5.08%, and 13.94% in comparison to CNN, RF, and LR models, respectively. These outcomes substantiate the efficacy of the SVM model in predicting postoperative recurrence in patients with CSDH. Moreover, it was observed that the hybrid input set comprising both clinical and radiomics

XSL•FO
RenderX

data enhanced the performance of the SVM model itself. Specifically, the accuracy, AUC, and recall were higher by 5.6% and 2.97%; 0.56% and 3.84%; and 1.88% and 3.54%, respectively, when compared to SVM models that used either radiomics features or clinical data as stand-alone inputs. In conclusion, the SVM model, when constructed based on a fusion of clinical and radiomics features, exhibited superior predictive capabilities, making it the optimal choice for this study.

While the SVM model using clinical-radiomics features demonstrated superior performance, it did not meet our predefined target of exceeding 90% across key evaluation metrics. To address this, a feature selection process was implemented to refine the input variables for the predictive model. Our analysis identified the top 5 influential variables impacting postoperative recurrence in patients with CSDH as history of head trauma, MGS, postoperative midline improvement, preoperative midline shift, and radiomics features (Multimedia Appendix 2). History of head trauma is one of the main factors leading to the formation of CSDH; MGS is used to assess the severity of CSDH, including hematoma volume, midline shift, state of consciousness, and other factors; postoperative midline improvement is an important index for assessing surgical results and patient's recovery; preoperative midline shift reflects the degree of compression of hematoma on brain tissue, if postoperative midline improvement is poor or larger midline shift usually indicates a poor prognosis and a higher risk of recurrence; and radiomics features can provide detailed information about CSDH, which can help physicians more accurately assess the disease and predict the risk of recurrence. These features are valuable in predicting the risk of recurrence after CSDH and should be emphasized by clinicians during diagnosis.

Subsequent feature selection experiments, conducted using the RFE method, indicated an optimal combination of 12 variables (Figure 5). When variables such as length of stay in hospital, sex, and age were excluded, the model yielded the highest performance across accuracy, AUC, and recall, registering 92.72%, 91.34%, and 93.16%, respectively. Consequently, these 12 features were incorporated into the refined SVM model to yield an optimal predictive tool for assessing the likelihood of postoperative recurrence in patients with CSDH.

Figure 5. Results of accuracy (ACC), area under the curve (AUC), and recall after selecting the number of features by cross-validation.

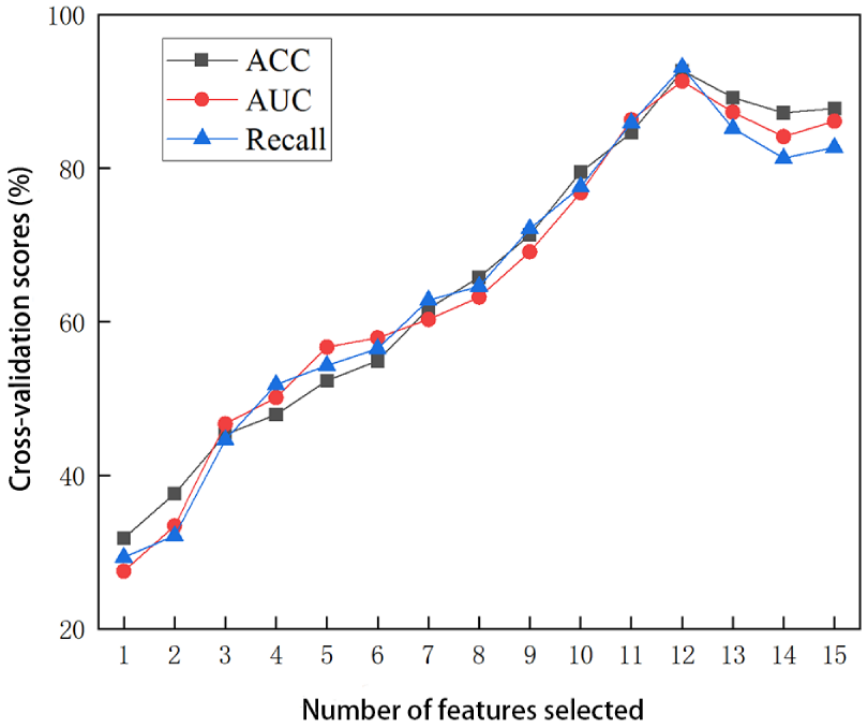

External Data Set Validation

To corroborate the reliability and generalizability of the developed SVM model, an external validation was performed using clinical-radiomics feature data from 20 patients with CSDH, sourced from the First Affiliated Hospital of Anhui Medical University. The inclusion criteria, exclusion criteria, outcome metrics, and predictors for the data set used for external validation (including clinical data and preoperative head CT images) were identical to those for the modeling data set (Table 3). There was no significant difference between the distribution of significant variables between the model validation data set and the model development data set when comparing Tables 1 and 3. The obtained data were fed into the 4 preestablished models, and the outcomes are depicted in Figure 6. The accuracy metrics for the 4 models (SVM, CNN, RF, and LR) registered at 90.32%, 84.67%, 81.3%, and 75.78%, respectively. The AUC outcomes were 91.32%, 86.73%, 82.15%, and 72.16%, respectively. Recall rates were recorded at 88.37%, 87.12%, 80.01%, and 74.68%, respectively. Across all evaluation parameters, the SVM model consistently exhibited superior performance. With context, we found that the SVM model constructed based on the fusion of clinical and radiomics features has consistent results in both internal validation and

XSL•FO
RenderX

external testing, and performs best among the 4 models. Consequently, these results reconfirm that the SVM model is the most effective predictive tool for assessing postoperative recurrence in patients with CSDH, as further substantiated by this external data set validation.

The data we used for modeling came from the Second Affiliated Hospital of Anhui Medical University, which mainly serves local patients, so its medical record data may reflect more of the disease characteristics and treatment experiences of local patients. The data used for the external validation of the model came from the First Affiliated Hospital of Anhui Medical University, which attracts patients from all over the province and even the neighboring regions due to the hospital's reputation and geographical location. These patients may have different cultural backgrounds, living habits, and medical needs. Therefore, the 2 aforementioned hospitals provide a context for the differences between the internal and external data sets, ensuring the general applicability of the prediction model.

XSL•FO
RenderX

Table 3. Clinical variables used to validate the model.

Variables	Nonrecurrence (n=17)	Recurrence (n=3)	*P* value
Age (y), mean (SD)	72.3 (10.6)	68.0 (9.6)	.52
Sex , n (%)			.53
Male	14 (82)	3 (100)	
Female	3 (18)	0 (0)	
Smoking or drinking, n (%)			.89
Yes	12 (71)	2 (67)	
No	5 (29)	1 (33)	
Hypertension, n (%)			.72
Yes	4 (23)	1 (33)	
No	13 (77)	2 (67)	
Cerebral infarction, n (%)			__[a]
Yes	0 (0)	0 (0)	
No	17 (100)	3 (100)	
History of head trauma, n (%)			.95
Yes	10 (59)	2 (67)	
No	7 (41)	1 (33)	
History of antiplatelet or anticoagulation , n (%)			.33
Yes	2 (11.8)	1 (33.3)	
No	15 (88.2)	2 (66.7)	
Markwalder Grading Scale, n (%)			.31
0	0 (0)	0 (0)	
1	5 (29)	0 (0)	
2	9 (53)	3 (100)	
3	3 (18)	0 (0)	
4	0 (0.0)	0 (0)	
Length of stay in hospital (d), mean (SD)	14.8 (5.7)	11.7 (3.8)	.38
CSDH[b] location, n (%)			.58
Unilateral	13 (77)	3 (100)	
Bilateral	4 (23)	0 (0)	
Classification of hematoma, n (%)			.11
Homogeneous	4 (23)	0 (0.0)	
Laminar	3 (18)	0 (0.0)	
Separated	6 (35)	1 (33)	
Trabecular type	4 (23)	2 (67)	
Preoperative hematoma volume (mL), mean (SD)	105.7 (27.3)	126.2 (14.6)	.23
Preoperative midline shift (cm), mean (SD)	1.1 (0.4)	1.3 (0.2)	.39
Postoperative midline improvement (cm), mean (SD)	0.5 (0.3)	0.6 (0.3)	.59

[a]Not available.

[b]CSDH: chronic subdural hematoma.

XSL•FO
RenderX

Figure 6. Accuracy (ACC), area under the curve (AUC), and recall for the 4 models in the external validation set. CNN: convolutional neural network; LR: linear regression; RF: random forest; SVM: support vector machine.

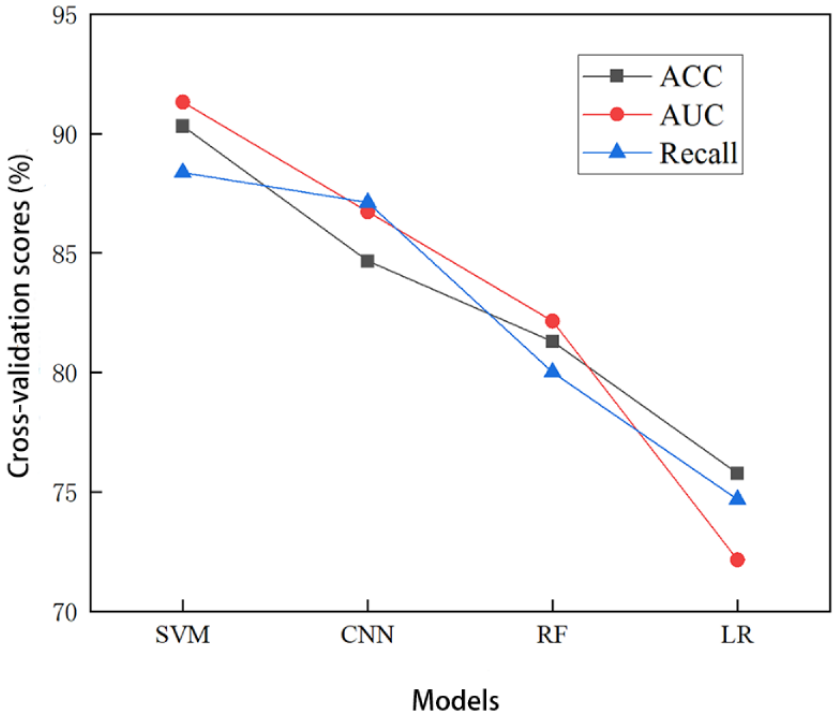

Discussion

Principal Findings

In this study, the observed postoperative recurrence rate for patients with CSDH was 14.3%, a figure that aligns closely with previously published data, ranging from 5% to 26% [22,23]. This congruence lends credibility to the representativeness of our data set and suggests that the study's findings have broad applicability.

Within the neurosurgical landscape, CSDH has received limited investigative focus. This is likely due to its relatively high prevalence and standardized treatment approach, coupled with its lower mortality risk compared to other neurosurgical conditions. However, given the global demographic shift toward an older population, the incidence of CSDH—and consequently, its recurrence rate—is witnessing a steady uptick. This escalating trend underscores the need for ongoing research attention, a need that remains largely unmet.

Using extensive data collection, processing, and iterative model optimization, we arrived at an SVM model predicated on clinical-radiomics features that exhibited optimal predictive performance. The final model not only met our predetermined efficacy criteria but also carries clinical utility. To our understanding, this constitutes the inaugural effort to leverage clinical-radiomics features in tandem with ML methodologies for the relatively accurate forecasting of postoperative recurrence in patients with CSDH. Thus, this study paves the way for future research, offering a novel paradigm for evaluating surgical outcomes in this patient cohort.

Our review of existing literature reveals that although there are studies incorporating ML with medical imaging (CT or magnetic resonance imaging) for clinical applications, the prevailing approach does not typically leverage raw images for model construction. Instead, researchers use various software tools or techniques to extract radiomics features from these images, upon which subsequent models are built. Consistent with these findings, our empirical results confirm that predictive models constructed directly from unprocessed images perform suboptimally. The limitations are primarily attributable to the inherent difficulty that traditional ML algorithms face in extracting key lesion characteristics—such as location, size, and morphology—from unprocessed CT images. To address these limitations, we used the open-source software, 3D Slicer, which offers an array of algorithms for feature extraction, including edge and corner detection, as well as texture analysis. By segmenting the ROI and extracting features using 3D Slicer, we acquired meaningful mathematical attributes, such as gradient and curvature. This enhanced feature set enabled more effective computerized analysis of lesions and improved predictive outcomes. Consequently, we shifted our strategy from using raw CT images to combining extracted radiomics features with clinical data for patients with CSDH as input variables in our model construction. The comparative evaluation of models developed through 4 ML algorithms (SVM, CNN, RF, and LR) revealed the superiority of the hybrid data approach over models built solely on clinical or radiomics data. Further, a performance matrix based on metrics such as accuracy, AUC, and recall indicated marked variations among the 4 methods, with the SVM model outperforming CNN, RF, and LR on all 3 metrics (accuracy: 2.77%, 5.06%, and 13.87%; AUC: 3.41%, 7.3%, and 15.44%; and recall: 4.18%, 6.93%, and 15.48%, respectively). Therefore, our analysis corroborates the high reliability of the SVM-based predictive model constructed using the amalgamated data set.

Our analysis of the reasons for the poor performance of other models is as follows. (1) Mismatch between data characteristics

XSL•FO
RenderX

and model complexity: some highly complex models, such as CNN for deep learning, may encounter overfitting problems when confronted with data sets that are relatively small or not rich enough in feature dimensions; this means that the model performs well on training data, but performs poorly on new, unseen data. (2) Feature extraction and exploitation capabilities: while RF and LR have some advantages in dealing with nonlinear problems and feature selection, they may not be as good as especially designed algorithms such as SVM with appropriate kernel functions. (3) Sensitivity to unbalanced data: models such as RF and LR may be biased toward the majority class if the positive and negative samples are unevenly distributed, which is a common problem in medical image analysis. This may lead to a decrease in the recall of the model and weak identification of the minority class (recurrent cases in this study). And the main reason why SVM performs well in predicting postoperative recurrence in patients with CSDH is that its principle of maximum interval (SVM improves the generalization ability of the model by maximizing the interval between the decision boundary and the support vectors) and the kernel trick (SVM can map the input space to a high-dimensional feature space through the kernel function, in which the nonlinear problem may become linearly differentiable) provide powerful support for dealing with small samples, high dimensionality, and unbalanced medical data provide powerful support.

Our model achieved 90.32% accuracy, 91.32% AUC, and 88.37% recall on an independent external validation set. These metrics indicate that the model has high accuracy in predicting postoperative recurrence in patients with CSDH. For socioeconomic purposes, accurate prediction can reduce unnecessary examinations and follow-up visits, which can save a large amount of medical resources every year; accurate prediction can also help physicians to take interventions in advance to reduce the incidence of postoperative complications and readmission rate of patients. This not only reduces patient suffering but also lowers hospital readmission costs. For patients, it reduces their burden by decreasing unnecessary examinations and follow-up visits, and timely interventions based on the predicted results can help reduce the occurrence of complications and improve their quality of life; in addition, through the model prediction, physicians can provide patients with more personalized treatment plans and care recommendations, which can improve patients' satisfaction and trust. In summary, our model has excellent performance and significant impact on clinical practice and economic benefits.

As early as 2009, Abouzari et al [24] explored the use of ML algorithms, specifically artificial neural networks and LR, for prognosis prediction in patients with CSDH. Given the technological limitations of that era, these models exhibited low accuracy and questionable evaluation metrics. However, their pioneering work served as a catalyst for us to implement contemporary ML techniques in this research domain.

With respect to postoperative recurrence in patients with CSDH, extensive studies have been conducted on hematoma staging using CT imaging. Historically, hematomas were simplistically categorized into 4 density-based types: low density, isodense, high density, and mixed density [25]. However, Tsutsumi et al [26] found no statistically significant difference in postoperative recurrence rates when using these classifications. Subsequently, Nakaguchi et al [17] introduced an alternative, more nuanced, classification—comprising homogeneous, laminar, separated, and trabecular types—which garnered wide acceptance in the research community.

In the current investigation, we adhered to this latter classification scheme when analyzing CT images. Notably, our data analysis revealed that the "separated" type constituted a greater fraction of the recurrence group, aligning with prior research findings. However, the proportion of cases classified as "trabecular" diverged from existing literature. We hypothesize that this discrepancy may be attributable to selection bias arising from our limited data set.

In the realm of CSDH postoperative recurrence, the scholarly focus has predominantly been on surgical methodologies—recently emphasizing middle meningeal artery embolization—patient age, and the administration of antiplatelet or anticoagulant medications [27-29]. This narrow concentration likely stems from the ubiquity of CSDH and the established efficacy of existing surgical treatments, which generally yield favorable outcomes without posing immediate life-threatening risks to patients. Consequently, research has largely remained at the clinical echelon. However, as technological advancements continue to pervade medical practice, the incorporation of these innovations not only streamlines clinical operations but also enhances patient outcomes, thereby advancing the objective of precision medicine.

In the current investigation, we diverged from the conventional practice of using either clinical data or imaging histology data exclusively. Rather, we integrated both data types and, through comparative analysis, substantiated the superior predictive performance of combined clinical-radiomics features. Furthermore, SVM was identified as an efficacious classification algorithm particularly suited for the unique characteristics of medical imaging data, which are high-dimensional and often limited in sample size. SVM achieves classification by constructing a hyperplane that appropriately segregates distinct feature sets in medical imaging data, thereby facilitating more accurate identification and prediction of postoperative recurrence in patients with CSDH. In addition, SVM exhibits robustness in mitigating the influence of noise and outliers commonly present in radiomics features, thus bolstering the reliability of model predictions.

In the realm of medical research, feature selection predominantly uses filtering methods, including but not limited to correlation coefficients, chi-square tests, and mutual information, to identify variables that highly correlate with the target outcome. Filtering methods excel in computational efficiency, capable of swiftly processing large data sets and thereby reducing dimensionality. These methods also offer adaptability, accommodating user-defined criteria for application-specific scenarios.

Nevertheless, this study uses RFE in lieu of filtering methods, and for several substantive reasons: Capability to manage highly correlated features—filtering methods struggle with the presence of a multitude of highly correlated variables, a challenge more effectively navigated by RFE. Distributional assumption

XSL•FO
RenderX

sensitivity—filtering methods often rest on certain statistical distribution assumptions (eg, normal, t-distribution), which, if incorrect, compromise feature selection accuracy. Conversely, RFE operates independently of such assumptions. Computational efficiency—contrary to common perception, filtering methods, while efficient with smaller data sets, may demand substantial computational resources and time when applied to larger data sets. RFE, on the other hand, demonstrates superior computational efficiency in such contexts. Applicability to nonlinear problems—filtering methods generally rely on linear models, limiting their efficacy for nonlinear challenges. RFE exhibits no such constraint. Automated and robust feature selection—unique to RFE is its ability to automatically discern the most pertinent feature subset, obviating the need for manual selection. This automation further minimizes overfitting risks and enhances model interpretability by focusing on the most salient features [30,31].

Given these advantages, RFE was selected as the feature selection methodology for this study.

Upon implementing RFE for feature selection, the predictive model demonstrated robust performance metrics, including an accuracy of 92.72%, AUC of 91.34%, and recall rate of 93.16%. Experimental outcomes identified the top 5 variables influencing postoperative recurrence in patients with CSDH as follows: history of head trauma, MGS, postoperative midline improvement, preoperative midline shift, and radiomics. Notably, the substantial impact of postoperative midline improvement and preoperative midline shift on the prognosis of CSDH has not been highlighted in extant literature. Consequently, we advocate for the inclusion of these novel factors in future CSDH studies given their potential clinical relevance.

In the clinical setting, the specific steps we take to implement the model prediction function are as follows: (1) integrate the predictive model into an existing health care information system, such as an electronic medical record system or medical image processing software; (2) input data into the system, which consists of the patient's medical images and relevant clinical information; (3) predictive results are presented to the physician in an easy-to-understand manner, such as a percentage of probability representation; and (4) regularly update and maintain the prediction model. This can be done by collecting new clinical data, optimizing algorithm parameters, or adjusting feature selection strategies. Through the abovementioned steps, patients can be provided with personalized treatment plans, reducing unnecessary tests and surgeries and improving the efficient use of medical resources. For patients with a higher risk of recurrence, physicians can take treatment measures earlier, thus improving the patient's prognosis. In the event of a discrepancy between the model prediction and the physician's judgment, the physician should first adopt a conservative treatment strategy to ensure patient safety. For example, patients whose model predictions are at low risk of recurrence but whom physicians believe are at higher risk should continue to be closely monitored and followed up. The physician can then compare the model predictions with the patient's actual treatment results and feed this information back to the model developer. Through continuous data feedback and model optimization, the predictive accuracy and generalization ability of the model can be improved. For now, ML models are only supplementary tools, and physicians should always make the final decision in conjunction with their own expertise and experience.

To realize the effective application of predictive models in clinical practice, we need to establish a stable and reliable data pipeline. First, we need to collect clinical data from patients with CSDH, including radiomic features, medical record information, and surgical records. Then, we need to preprocess and feature extract the data to feed it into a predictive model. Next, we need to train and validate the model using ML algorithms to ensure its predictive accuracy and reliability. Finally, we need to integrate the predictive model into an existing health care information system so that it can automatically receive and process patients' clinical data and generate predictions. During the establishment of the data pipeline, we need to consider the quality, integrity, and security of the data. We need to ensure the accuracy and consistency of the data to avoid adverse effects on the predicted outcomes. At the same time, we need to ensure data security and privacy to protect patients' privacy rights.

While our findings hold considerable clinical utility and prospective applicability, it is imperative to acknowledge the following limitations of the study: Assumed input and output data formats—the ML model used in this study is based on a specific input data format (eg, radiological features and clinical data extracted via 3D Slicer software) and assumes that these data are fully representative of the patient's health status and subdural hematoma characteristics. However, this assumption may omit other important biomarkers or unquantified clinical parameters [19,20], which may have an impact on the predictive power of the model. The output data format is assumed to be measured in terms of specific predictive accuracy metrics (eg, accuracy, AUC, and recall), which may not adequately reflect the utility and sensitivity of the model in different clinical settings. Potential pitfalls in interpreting the model—although the SVM model showed good performance in this study, SVMs and other ML models are often considered "black-box" models, in which the model's decision-making process may not be transparent, and this lack of interpretability may produce a lack of trust in settings where the model is used to guide clinical decision-making. Potential bias of the data used in modeling—the study was conducted based on a retrospective data set from a specific health care organization, which may be subject to selection bias (eg, only patients who received surgical treatments were included) and informational bias (data records may not be completely accurate). In addition, due to the relatively small sample size (133 patients), the complex relationship between CSDH recurrence and multiple underlying factors may not have been adequately captured, which may have affected the model's ability to generalize and predict accuracy. Generalizability of the data—although the study was externally validated, the validation set consisted of only 20 patients from another health care facility, which may not be sufficient to comprehensively assess the ability of the model to generalize across populations and geographical regions. Patient populations in different regions may have different clinical characteristics, such as different treatment modalities and different health care

XSL•FO
RenderX

resources, all of which may affect the generalizability and accuracy of the model.

Our subsequent research will work to address the above issues.

Conclusions

In this study, we constructed 4 models to predict postoperative recurrence in patients with CSDH, using ML algorithms and an amalgamated data set comprising both radiomics attributes and clinical variables. Comparative evaluation revealed that the SVM model, using this integrated data set, demonstrated superior predictive accuracy. The model not only outperforms previously established methods but also provides a more specific and comprehensive framework for predicting outcomes. These predictive findings enable health care teams to refine clinical decision-making and offer individualized treatment plans. Moreover, patients can engage in proactive follow-up and informed participation in their treatment protocols based on these results. The developed method offers the advantage of real-time updates and holds considerable clinical implications.

Acknowledgments

This research was funded by the Natural Science Foundation of Anhui Xinhua University (2023zr013), the Anhui Provincial University Students Innovation and Entrepreneurship Program (S202312216119 and S202212216022), and the CAMS Innovation Fund for Medical Sciences (2022-I2M-C&T-B-063).

Authors' Contributions

JW and XJ designed the study and revised the manuscript and were responsible for the study. SL, GX, and HZ collected the data. CF screened and checked the data. YP built the models. CF and YP drafted the manuscript. All authors made substantial contributions to the study and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Items of reporting predictive models in biomedical research.
[DOCX File , 21 KB - jmir_v26i1e54944_app1.docx]

Multimedia Appendix 2
Ranking the importance of these 12 features screened by the recursive feature elimination method in descending order of importance. CSDH: chronic subdural hematoma.
[PNG File , 60 KB - jmir_v26i1e54944_app2.png]

References

1. Kudo H, Kuwamura K, Izawa I, Sawa H, Tamaki N. Chronic subdural hematoma in elderly people: present status on Awaji Island and epidemiological prospect. Neurol Med Chir (Tokyo) 1992 Apr;32(4):207-209 [FREE Full text] [doi: 10.2176/nmc.32.207] [Medline: 1378564]
2. Rust T, Kiemer N, Erasmus A. Chronic subdural haematomas and anticoagulation or anti-thrombotic therapy. J Clin Neurosci 2006 Oct;13(8):823-827. [doi: 10.1016/j.jocn.2004.12.013] [Medline: 16997707]
3. Feghali J, Yang W, Huang J. Updates in chronic subdural hematoma: epidemiology, etiology, pathogenesis, treatment, and outcome. World Neurosurg 2020 Sep;141:339-345. [doi: 10.1016/j.wneu.2020.06.140] [Medline: 32593768]
4. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930. [doi: 10.1161/circulationaha.115.001593]
5. Li C, Liu M, Zhang Y, Wang Y, Li J, Sun S, et al. Novel models by machine learning to predict prognosis of breast cancer brain metastases. J Transl Med 2023 Jun 21;21(1):404 [FREE Full text] [doi: 10.1186/s12967-023-04277-2] [Medline: 37344847]
6. Wang X, Zhong J, Lei T, Chen D, Wang H, Zhu L, et al. An artificial neural network prediction model for posttraumatic epilepsy: retrospective cohort study. J Med Internet Res 2021 Aug 19;23(8):e25090 [FREE Full text] [doi: 10.2196/25090] [Medline: 34420931]
7. Yao Q, Jia W, Chen S, Wang Q, Liu Z, Liu D, et al. Machine learning was used to predict risk factors for distant metastasis of pancreatic cancer and prognosis analysis. J Cancer Res Clin Oncol 2023 Sep 06;149(12):10279-10291. [doi: 10.1007/s00432-023-04903-y] [Medline: 37278826]
8. Monteiro M, Newcombe VF, Mathieu F, Adatia K, Kamnitsas K, Ferrante E, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. Lancet Digit Health 2020 Jun;2(6):e314-e322 [FREE Full text] [doi: 10.1016/S2589-7500(20)30085-6] [Medline: 33328125]

XSL•FO
RenderX

9. Fang C, Pan Y, Zhao L, Niu Z, Guo Q, Zhao B. A machine learning-based approach to predict prognosis and length of hospital stay in adults and children with traumatic brain injury: retrospective cohort study. J Med Internet Res 2022 Dec 09;24(12):e41819 [FREE Full text] [doi: 10.2196/41819] [Medline: 36485032]
10. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016 Feb;278(2):563-577 [FREE Full text] [doi: 10.1148/radiol.2015151169] [Medline: 26579733]
11. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017 Dec 4;14(12):749-762. [doi: 10.1038/nrclinonc.2017.141] [Medline: 28975929]
12. Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. J Nucl Med 2018 Feb 24;59(2):189-193 [FREE Full text] [doi: 10.2967/jnumed.117.200501] [Medline: 29175982]
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
14. Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 2019 Jul 01;9(1):9441 [FREE Full text] [doi: 10.1038/s41598-019-45766-z] [Medline: 31263116]
15. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep 2016 Mar 24;6(1):23428 [FREE Full text] [doi: 10.1038/srep23428] [Medline: 27009765]
16. Markwalder TM. Chronic subdural hematomas: a review. J Neurosurg 1981 May;54(5):637-645. [doi: 10.3171/jns.1981.54.5.0637] [Medline: 7014792]
17. Nakaguchi H, Tanishima T, Yoshimasu N. Factors in the natural history of chronic subdural hematomas that influence their postoperative recurrence. J Neurosurg 2001 Aug;95(2):256-262. [doi: 10.3171/jns.2001.95.2.0256] [Medline: 11780895]
18. Stanišic M, Pripp AH. A reliable grading system for prediction of chronic subdural hematoma recurrence requiring reoperation after initial burr-hole surgery. Neurosurgery 2017 Nov 01;81(5):752-760 [FREE Full text] [doi: 10.1093/neuros/nyx090] [Medline: 28379528]
19. Hirai S, Yagi K, Hara K, Kanda E, Matsubara S, Uno M. Postoperative recurrence of chronic subdural hematoma is more frequent in patients with blood type A. J Neurosurg 2021 Jan 01;135(4):1203-1207. [doi: 10.3171/2020.7.JNS202330] [Medline: 33385994]
20. Matsubara M, Yagi K, Minami Y, Kanda E, Sunada Y, Tao Y, et al. Preoperative elevated eosinophils in peripheral blood for prediction of postoperative recurrence of chronic subdural hematoma. J Neurosurg 2023 Sep 01;139(3):708-713. [doi: 10.3171/2022.12.JNS222432] [Medline: 36640094]
21. Pan Y, Fang C, Zhu X, Wan J. Construction of a predictive model based on MIV-SVR for prognosis and length of stay in patients with traumatic brain injury: retrospective cohort study. Digit Health 2023 Dec 03;9:20552076231217814 [FREE Full text] [doi: 10.1177/20552076231217814] [Medline: 38053736]
22. Martinez-Perez R, Tsimpas A, Rayo N, Cepeda S, Lagares A. Role of the patient comorbidity in the recurrence of chronic subdural hematomas. Neurosurg Rev 2021 Apr 07;44(2):971-976. [doi: 10.1007/s10143-020-01274-7] [Medline: 32146611]
23. Weigel R, Schmiedek P, Krauss JK. Outcome of contemporary surgery for chronic subdural haematoma: evidence based review. J Neurol Neurosurg Psychiatry 2003 Jul 01;74(7):937-943 [FREE Full text] [doi: 10.1136/jnnp.74.7.937] [Medline: 12810784]
24. Abouzari M, Rashidi A, Zandi-Toghani M, Behzadi M, Asadollahi M. Chronic subdural hematoma outcome prediction using logistic regression and an artificial neural network. Neurosurg Rev 2009 Oct 4;32(4):479-484. [doi: 10.1007/s10143-009-0215-3] [Medline: 19653019]
25. Nomura S, Kashiwagi S, Fujisawa H, Ito H, Nakamura K. Characterization of local hyperfibrinolysis in chronic subdural hematomas by SDS-PAGE and immunoblot. J Neurosurg 1994 Dec;81(6):910-913. [doi: 10.3171/jns.1994.81.6.0910] [Medline: 7965121]
26. Tsutsumi K, Maeda K, Iijima A, Usui M, Okada Y, Kirino T. The relationship of preoperative magnetic resonance imaging findings and closed system drainage in the recurrence of chronic subdural hematoma. J Neurosurg 1997 Dec;87(6):870-875. [doi: 10.3171/jns.1997.87.6.0870] [Medline: 9384397]
27. Cofano F, Pesce A, Vercelli G, Mammi M, Massara A, Minardi M, et al. Risk of recurrence of chronic subdural hematomas after surgery: a multicenter observational cohort study. Front Neurol 2020 Nov 24;11:560269 [FREE Full text] [doi: 10.3389/fneur.2020.560269] [Medline: 33329304]
28. Shotar E, Meyblum L, Premat K, Lenck S, Degos V, Grand T, et al. Middle meningeal artery embolization reduces the post-operative recurrence rate of at-risk chronic subdural hematoma. J Neurointerv Surg 2020 Dec 21;12(12):1209-1213. [doi: 10.1136/neurintsurg-2020-016048] [Medline: 32439812]
29. Andersen-Ranberg NC, Poulsen FR, Bergholt B, Hundsholt T, Fugleholm K. Bilateral chronic subdural hematoma: unilateral or bilateral drainage? J Neurosurg 2017 Jun;126(6):1905-1911. [doi: 10.3171/2016.4.JNS152642] [Medline: 27392267]
30. Yan K, Zhang D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sens Actuators B Chem 2015 Jun;212:353-363. [doi: 10.1016/j.snb.2015.02.025]

XSL•FO
RenderX

31. Chen XW, Jeong JC. Enhanced recursive feature elimination. In: Proceedings of the 6th International Conference on Machine Learning and Applications. 2007 Presented at: ICMLA '07; December 13-15, 2007; Cincinnati, OH p. 429-435 URL: https://ieeexplore.ieee.org/document/4457268 [doi: 10.1109/icmla.2007.35]

Abbreviations

AUC: area under the curve
CNN: convolutional neural network
CSDH: chronic subdural hematoma
CT: computed tomography
LR: linear regression
MGS: Markwalder Grading Scale
ML: machine learning
RF: random forest
RFE: recursive feature elimination
ROI: region of interest
SVM: support vector machine

Edited by S Ma, T Leung; submitted 28.11.23; peer-reviewed by Y Jing, F Zhu; comments to author 14.03.24; revised version received 11.04.24; accepted 26.06.24; published 28.08.24.

Please cite as:
Fang C, Ji X, Pan Y, Xie G, Zhang H, Li S, Wan J
Combining Clinical-Radiomics Features With Machine Learning Methods for Building Models to Predict Postoperative Recurrence in Patients With Chronic Subdural Hematoma: Retrospective Cohort Study
J Med Internet Res 2024;26:e54944
URL: https://www.jmir.org/2024/1/e54944
doi:10.2196/54944
PMID:

©Cheng Fang, Xiao Ji, Yifeng Pan, Guanchao Xie, Hongsheng Zhang, Sai Li, Jinghai Wan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study

Yun Kwan Kim[1,2], MSc; Won-Doo Seo[1], MSc; Sun Jung Lee[1], PhD; Ja Hyung Koo[1], PhD; Gyung Chul Kim[1], MSE; Hee Seok Song[1], MSE; Minji Lee[3], PhD

[1]Technology Development, Seers Technology Co. Ltd., Pyeongtaek-si, Gyeonggi-do, Republic of Korea

[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

[3]Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Minji Lee, PhD
Department of Biomedical Software Engineering
The Catholic University of Korea
43, Jibong-ro
Bucheon-si, Gyeonggi-do, 14662
Republic of Korea
Phone: 82 2 2164 4364
Email: minjilee@catholic.ac.kr

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e67135

Abstract

Background: Cardiac arrest (CA) is one of the leading causes of death among patients in the intensive care unit (ICU). Although many CA prediction models with high sensitivity have been developed to anticipate CA, their practical application has been challenging due to a lack of generalization and validation. Additionally, the heterogeneity among patients in different ICU subtypes has not been adequately addressed.

Objective: This study aims to propose a clinically interpretable ensemble approach for the timely and accurate prediction of CA within 24 hours, regardless of patient heterogeneity, including variations across different populations and ICU subtypes. Additionally, we conducted patient-independent evaluations to emphasize the model's generalization performance and analyzed interpretable results that can be readily adopted by clinicians in real-time.

Methods: Patients were retrospectively analyzed using data from the Medical Information Mart for Intensive Care-IV (MIMIC-IV) and the eICU-Collaborative Research Database (eICU-CRD). To address the problem of underperformance, we constructed our framework using feature sets based on vital signs, multiresolution statistical analysis, and the Gini index, with a 12-hour window to capture the unique characteristics of CA. We extracted 3 types of features from each database to compare the performance of CA prediction between high-risk patient groups from MIMIC-IV and patients without CA from eICU-CRD. After feature extraction, we developed a tabular network (TabNet) model using feature screening with cost-sensitive learning. To assess real-time CA prediction performance, we used 10-fold leave-one-patient-out cross-validation and a cross–data set method. We evaluated MIMIC-IV and eICU-CRD across different cohort populations and subtypes of ICU within each database. Finally, external validation using the eICU-CRD and MIMIC-IV databases was conducted to assess the model's generalization ability. The decision mask of the proposed method was used to capture the interpretability of the model.

Results: The proposed method outperformed conventional approaches across different cohort populations in both MIMIC-IV and eICU-CRD. Additionally, it achieved higher accuracy than baseline models for various ICU subtypes within both databases. The interpretable prediction results can enhance clinicians' understanding of CA prediction by serving as a statistical comparison between non-CA and CA groups. Next, we tested the eICU-CRD and MIMIC-IV data sets using models trained on MIMIC-IV and eICU-CRD, respectively, to evaluate generalization ability. The results demonstrated superior performance compared with baseline models.

XSL•FO
RenderX

Conclusions: Our novel framework for learning unique features provides stable predictive power across different ICU environments. Most of the interpretable global information reveals statistical differences between CA and non-CA groups, demonstrating its utility as an indicator for clinical decisions. Consequently, the proposed CA prediction system is a clinically validated algorithm that enables clinicians to intervene early based on CA prediction information and can be applied to clinical trials in digital health.

(J Med Internet Res 2024;26:e62890) doi:10.2196/62890

KEYWORDS

early cardiac arrest warning system; electric medical record; explainable clinical decision support system; pseudo-real-time evaluation; ensemble learning; cost-sensitive learning

Introduction

Critical illness is defined as the presence or potential development of organ dysfunction. Cardiac arrest (CA), a critical condition that impacts patient safety, refers to the sudden cessation of cardiac function due to specific abnormal events, such as ventricular arrhythmia, asystole, and pulseless electrical activity [1]. At least one abnormal sign, such as respiratory distress or hemodynamic instability, occurs in 59.4% of patients within 1-4 hours before the onset of CA [2]. Early identification of the causes of CA improves patient survival by approximately 29% within the first hour of the episode and 19% at discharge [3]. Therefore, early prediction of CA is crucial to allow for more time for clinical intervention, thereby reducing mortality.

Clinical decision support systems (CDSSs) are clinical computer systems that apply algorithms to patient information, use machine learning to evaluate clinical data, and provide clinical decision support [4,5]. These systems, developed using electronic medical records, utilize various paradigms—such as predicting early cardiac events, heart failure (HF), and critical illness—to enable rapid response through real-time patient monitoring [6-9]. To enhance the quality and speed of medical services, CA prediction and warning systems have been developed for use in intensive care units (ICUs) within the field of CDSSs [6]. These computer-based CA prediction algorithms offer new opportunities for clinicians to improve the accuracy of predicting CA events [10].

Traditional score–based methods, including the Simplified Acute Physiology Score (SAPS)-II, Sequential Organ Failure Assessment (SOFA), and Modified Early Warning Score (MEWS), are tools used by in-hospital care teams to identify early indicators of CA and initiate early intervention and therapy [11-15]. However, these score-based systems suffer from low sensitivity or a high false alarm rate [16]. To address these issues, machine learning methods have been used in CA prediction [17,18], leading to significant improvements in performance.

Churpek et al [19] proposed the use of a random forest (RF) classifier, based on clinical information extracted from a multicenter data set, achieving an area under the receiver operating characteristic curve (AUROC) of 0.83. Similarly, Hong et al [20] implemented an RF model using a clinical data set from a retrospective study, attaining an AUROC of 0.97 and an area under the precision-recall curve (AUPRC) of 0.86. Although the authors achieved accurate CA prediction results, their methodology relied heavily on features not commonly used during hospitalization and did not offer real-time predictions. To address this, Layeghian Javan et al [21] proposed a stacking method that combines RF, balanced bagging, and logistic regression to predict CA 1 hour in advance. Layeghian Javan et al [21] achieved an AUROC of 0.82 using the Medical Information Mart for Intensive Care (MIMIC)-III data set. Kwon et al [17] proposed a deep learning–based early warning system that utilizes a recurrent neural network (RNN) to assess risk scores from input vectors measured over an 8-hour period. Their system, based on vital signs extracted from a retrospective multicenter cohort data set, resulted in AUROC and AUPRC values of 0.85 and 0.04, respectively. Additionally, Kim et al [18] developed an ensemble-based CA prediction system using the light gradient boosting method (LGBM) to predict CA 1 hour in advance, obtaining AUROC and AUPRC values of 0.86 and 0.58, respectively, using the MIMIC-IV data set.

As mentioned earlier, artificial intelligence (AI) has been applied to CA prediction solutions and has demonstrated high predictive power in several studies [22]. However, hospitals often group patients with similar conditions and illness severity within the same unit for more efficient treatment. Specifically, ICUs are divided into subtypes, such as general ICUs and cardiac ICUs, to optimize care. In this context, previous studies on CA prediction focused on the entire ICU without accounting for the heterogeneity within subtypes. As a result, the performance of CA prediction models may vary depending on the distinct characteristics of each group [23].

Although numerous studies have applied AI to CA prediction [17,18], challenges persist in their practical application. First, CA prediction studies must confirm clinical validity through multicenter studies [2]. However, clinical maturity for CA prediction has not been established when monitored in real-time, as validation was typically performed using representative events extracted from the validation site. Second, patients grouped into different ICU subtypes exhibit varying characteristics and likelihood of developing CA. However, the performance of CA prediction models across these subtypes has not been validated. Third, while interpreting the results of prediction models is crucial for clinicians to make informed decisions [13], an interpretable model capable of providing this information in real-time monitoring—especially among deep learning–based models—has yet to be developed.

This study proposes a framework for early and accurate prediction of CA across diverse clinical settings, accounting for heterogeneity. We aim to validate the clinical maturity, safety,

XSL•FO
RenderX

and effectiveness of the proposed framework by comparing it with existing trigger systems and machine learning methods using a pseudo real-time CA evaluation system. We propose a novel framework that learns patient-independent and subtype-specific characteristics in the ICU to improve CA prediction and reduce the false alarm rate. As a deep learning–based model optimized for tabular data, such as tabular network (TabNet), it can address the overfitting and performance limitations of existing tree-based models [24]. In addition, a cost-sensitive learning approach was applied to address the class imbalance in CA events. We then used the MIMIC-IV and eICU-Collaborative Research Database (eICU-CRD) to evaluate clinical maturity across various patient populations and ICU subtypes [25,26]. To illustrate changes in feature importance over time for clinical decisions, we utilized the MIMIC-IV data set [25]. Therefore, the proposed CA prediction framework can offer clinicians a reliable warning of CA occurrence within 24 hours. It also provides interpretable information about CA alarms and insights for rapid response.

Methods

Data Source

We used 2 databases: MIMIC-IV and eICU-CRD. The MIMIC-IV database, which includes information on vital signs, laboratory tests, and procedural events for ICU patients, was utilized to develop and validate a CA prediction model using multivariate vital sign time-series data from patients with HF. Specifically, MIMIC-IV is a well-known single-center database containing information on 46,520 patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019. The database includes demographic data, International Classification of Diseases (ICD) codes, clinical modification codes, hourly vital signs, inputs and outputs, laboratory test and microbiological culture results, imaging data, treatment methods, medication administration, and survival statistics. In addition, MIMIC-IV includes data from the clinical information system iMDsoft MetaVision. Compared with MIMIC-III, which extracts data from heterogeneous sources, MIMIC-IV provides more comprehensive patient data and detailed information on procedural events, serving as a primary source of clinical information in ICUs [25]. Consequently, MIMIC-IV data are more homogeneous compared with MIMIC-III data [25].

The eICU-CRD contains data from over 200,000 ICU admissions monitored across the United States through the eICU-CRD program developed by Philips Healthcare. This collaborative database includes information on patients admitted to the ICU in 2014 and 2015 [26].

Ethical Considerations

The MIMIC-IV database and eICU-CRD are deidentified, transformed, and made available to researchers who have completed human research training and signed a data use agreement. The Institutional Review Board at the BIDMC granted a waiver of informed consent and approved the sharing of the MIMIC-IV database. Similarly, the eICU-CRD data were exempt from Institutional Review Board approval and were also granted a waiver of informed consent [25,26].

To enhance the system for doctors and patients, AI has addressed concerns related to bias and fairness in health care. Our strategies to mitigate these issues are as follows:

Bias in AI generally arises from 2 primary sources: the data used for algorithmic training (data bias) and the intrinsic design or learning mechanisms of the algorithm itself (algorithmic bias). In health care settings, the involvement of human interaction and decision-making can introduce additional bias due to the inherently complex nature of the process [27]. To mitigate the impact of data bias, we used a patient-centered data set rather than relying solely on representative event data. Additionally, we conducted a subgroup analysis to identify potential biases in various environments. Key factors contributing to algorithmic bias are label bias and cohort bias [28,29]. Label bias has been addressed through updates to the MIMIC-IV and eICU-CRD databases [25,26]. To counter cohort bias, which may arise when different group levels are not adequately considered, we evaluated both patients with heart disease and the broader population using different databases [30].

Fairness in health care is multidimensional, involving the equitable distribution of resources, opportunities, and outcomes among diverse patient populations. Health care systems must ensure access to quality care for all individuals without discrimination [31]. To uphold this fairness, we selected MIMIC-IV and eICU-CRD—2 representative databases in critical care—and proceeded with AI development only after minimizing bias in each database. Additionally, we assessed explainability and the Brier score to address potential errors, harmful outcomes, and biases in AI-generated predictions.

The use of AI-driven predictions in critical care decisions can introduce various biases. Bias related to clinician interaction may be common in CDSS systems, with risks including overconfidence in the AI system or desensitization to real-world events due to excessive alerts [32,33]. To mitigate these biases, we evaluated the false alarm rate, event recall, and sensitivity. Additionally, future deployments will require clinician training on inherent biases and regular monitoring of the algorithm.

Problem Definition

The task of the study is to predict CA events within 24 hours. The input data include the patient's vital signs and the MEWS over a 12-hour time window. We then generate continuous labels every hour regarding the risk of CA within the next 24 hours and calculate the alarm rate based on the correctness of the alarms. The primary outcome was the AUROC, used to quantitatively assess the prediction results for CA events within 24 hours. The alarm rate, including false alarm rate and event recall, was calculated as a secondary outcome to evaluate alarm fatigue. Sensitivity was also assessed as a secondary outcome to identify any reductions in false alarms or missed CA events. Additionally, we provided clinically interpretable decision support information.

XSL•FO
RenderX

Prediction Model Framework

Overview

We propose a framework for predicting CA events within 24 hours in advance. As illustrated in Figure 1, the framework consists of 6 components: data preparation, data preprocessing and extraction, feature generation, feature aggregation and CA event labeling, model development, and evaluation. Details about the open-source tools and development code used are provided in Multimedia Appendix 1 [34].

After applying the inclusion and exclusion criteria, we extracted vital signs and calculated the MEWS based on these vital signs. In step 2, we processed and normalized the features after resampling the vital signs and MEWS to a 1-hour resolution. In step 3, we generated multiresolution statistical features and Gini index–based features. The multiresolution statistical features were created using a sliding window approach to segment each vital sign into 4-, 6-, and 12-hour intervals. Next, we generated continuous labels every hour indicating the risk of CA within the next 24 hours. In step 4, we aggregated multiresolution statistical features, Gini index–based features, and labels. In step 5, we developed a TabNet classifier, known for its effectiveness in various classification tasks with tabular data [24], and incorporated different cost weights for each class. Finally, in step 6, we evaluated the performance of the proposed model using sensitivity, false alarm rate, event recall, and AUROC.

Figure 1. Patient inclusion and exclusion flow diagram for the MIMIC-IV and eICU-CRD. (A) MIMIC-IV, (B) eICU-CRD. CA: cardiac arrest; eICU-CRD: eICU-Collaborative Research Database; HF: heart failure; ICU: intensive care unit; MIMIC: Medical Information Mart for Intensive Care; SAPS: Simplified Acute Physiology Score; SOFA: Sequential Organ Failure Assessment.

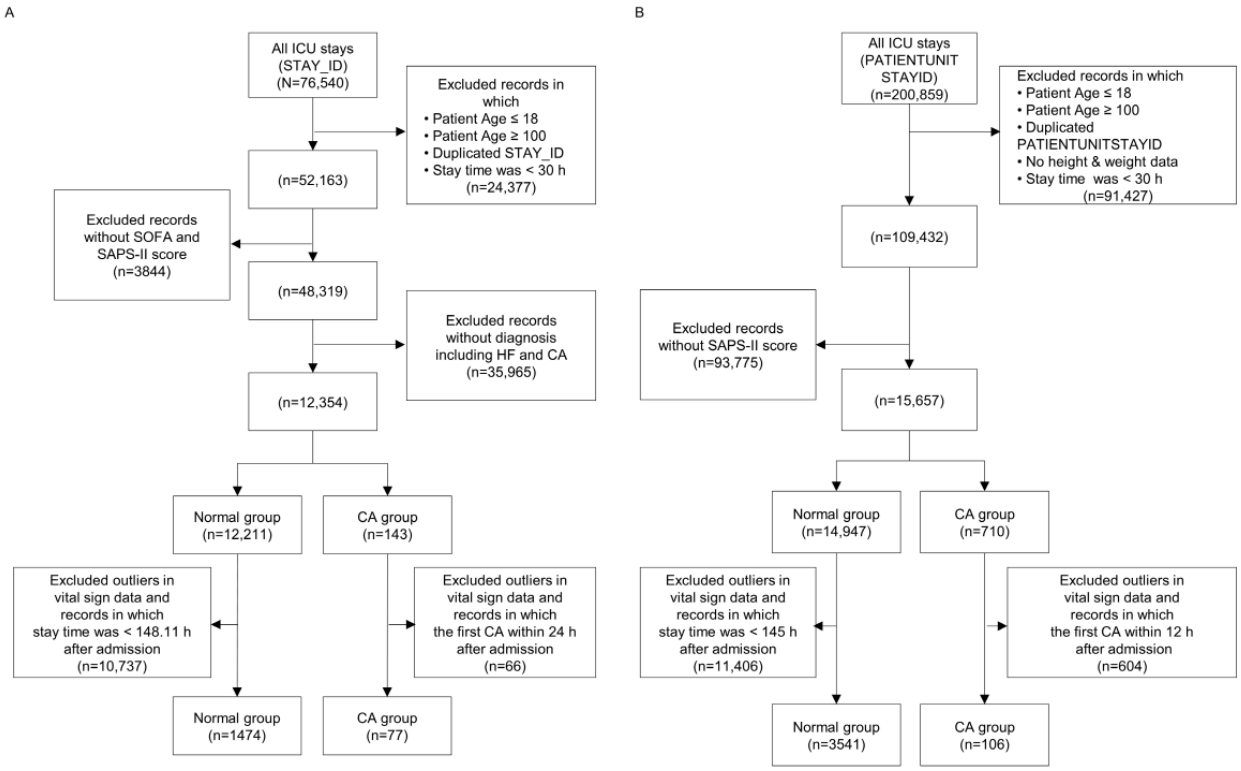

Step 1: Data Preparation

Data were obtained from the MIMIC-IV and eICU-CRD databases to construct cohorts meeting the inclusion and exclusion criteria [25]. The target populations of the 2 databases differed: MIMIC-IV includes only patients with cardiac issues in the ICU, while eICU-CRD encompasses all ICU patients. The number of CA events per patient also varied between the databases. MIMIC-IV typically has 1 CA event per patient, whereas eICU-CRD often records multiple CA events per patient. As CA events can occur multiple times per patient in a clinical setting, we validated multiple events in the eICU-CRD. Finally, we performed an analysis that accounted for differences across databases to compare the performance of CA prediction between high-risk patient groups and those without CA, as well as across different clinical settings, using the proposed framework, as shown in Figure 2.

For the inclusion and exclusion processing of MIMIC-IV, we applied the criteria to select the study cohort. Patients aged over 18 and under 100 years were included. Records of patients without SOFA and SAPS-II scores were excluded, as these scores were used to compare the prediction performance of the prognostic scales with the proposed framework. HF is a major risk factor for sudden CA and a significant contributor to CA-related mortality. CA is more prevalent in patients with a history of HF or previous CA. Therefore, we included ICU stays of patients with these cardiovascular conditions in the cohort study. For the CA group, we included ICU data if the vital sign data were not outliers and if any events occurred within 1 hour before the CA within 24 hours of patient admission.

XSL•FO
RenderX

For the inclusion and exclusion processing of the eICU-CRD, we acknowledged the differences in target population characteristics as mentioned above. Therefore, the inclusion and exclusion criteria were the same as those for MIMIC-IV, except for the criteria related to patients with high-risk CA.

MIMIC-IV includes only patients with cardiac issues in the ICU, while eICU-CRD encompasses all ICU patients. Consequently, the MIMIC-IV data set included 77 patients in the CA group and 1474 patients in the normal group, whereas the eICU-CRD data set included 106 patients in the CA group and 3641 patients in the normal group.

Figure 2. Overview of the proposed framework. This is composed of 6 steps including data preparation; data preprocessing and extraction; feature generation; feature aggregation and CA event labeling; model development; and evaluation. Three components make up TabNet, including feature transformer, attentive transformer, and feature masking. A split block separates the processed representation for the overall output and is used by the attentive transformer of the next phase. The feature selection mask provides comprehensible details about the functioning of the model for each step, and the masks can be combined to produce global feature important attribution. BN: batch normalization; CA: cardiac arrest; DBP: diastolic blood pressure; EWS: early warning score; HR: heart rate; MBP: mean blood pressure; MEWS: Modified Early Warning Score; ReLU: rectified linear unit; RFE: recursive feature elimination; RR: respiratory rate; SBP: systolic blood pressure; SpO_2: oxyhemoglobin saturation; TabNet: tabular network; TEMP: temperature.

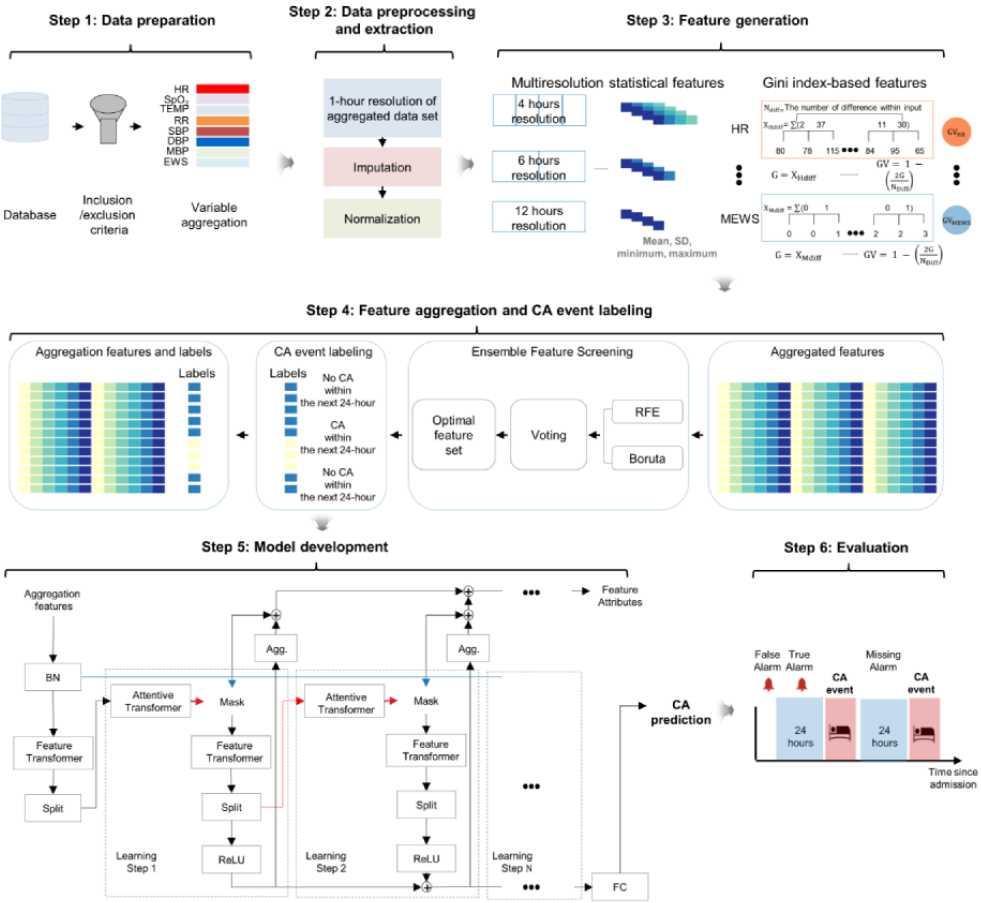

Step 2: Data Preprocessing and Extraction

We collected data on vital sign parameters, including heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), temperature, respiratory rate (RR), and oxyhemoglobin saturation (SpO_2) from the experimental database. These vital sign parameters may be recorded with irregular time-series data due to equipment malfunctions and varying patient responses [35]. Prediction models are not designed to classify data with irregular time-series samples between groups. To address this issue, the models require data collected at regular time intervals. We used a bucketing technique to manage the irregularities in the time series [21]. We divided the 12-hour time windows into 12 sequential 1-hour buckets, and the measured values within each bucket were averaged. Consequently, each time series consisted of 12 values at regular 1-hour intervals. If there were no values in a bucket, it was marked as null. To address missing values, we used the last observation carried forward (LOCF) and last observation carried backward (LOCB) imputation techniques [36]. In the LOCF method, missing values are filled by carrying forward the most recent nonmissing values. Similarly, in the LOCB method, missing values are filled by carrying backward the subsequent nonmissing values. Although we primarily used the LOCB method to impute missing values, the LOCF method was applied when subsequent values were missing, filling in missing values with the most recent nonmissing values.

Additionally, we extracted the early warning scores (EWS) for vital signs. We used the MEWS [15], a composite score commonly used by medical staff to assess illness severity. EWS observations were assigned scores ranging from 0 to 3. The

XSL•FO
RenderX

EWS was calculated every hour. To remove outliers, we determined the acceptable range for each variable based on the input from medical experts. Values falling outside this range were eliminated. We normalized each feature using the minimum and maximum values within the abnormal range for each vital sign, as each feature column had a different scale. We converted the database into an hourly time series with 12-hour intervals. Subsequently, we combined the CA and non-CA groups to perform the imputation task.

Step 3: Feature Generation and Aggregation

Feature Extraction and Processing for Cardiac Arrest Prediction

After applying the inclusion and exclusion criteria, we extracted vital signs—such as HR, SBP, DBP, temperature, RR, and SpO_2—and calculated the MEWS based on these vital signs. The features were then processed and normalized after resampling both the vital signs and MEWS to a 1-hour resolution [37]. The database was organized into an hourly time series with 12-hour intervals. Finally, we combined the CA and non-CA groups to perform the imputation task.

We generated 3 types of features within a 12-hour time window: vital sign–based features, multiresolution statistical features, and Gini index–based features. These features were designed to capture meaningful changes for predicting the occurrence of CA by identifying temporal patterns in vital signs, statistical variations across different resolutions, and the degree of information imbalance. The method for generating these features is outlined in the following sections.

Vital Sign–Based Features

To extract the pattern of vital signs, we used normalized vital signs and MEWS within a 12-hour time window.

Multiresolution Statistical Features

To capture statistical changes across different sections, we created time windows of increasing sizes and extracted summary statistics from these multiresolution sliding windows. For the multiresolution sliding window–based statistical features, the input data were segmented into resolutions of 0-4 hours, 0-6 hours, and 0-12 hours. Each time-series segment of the vital sign data was then aggregated to calculate the mean, median, minimum, maximum, and SD for each feature.

Specifically, the temporal patterns of each biological signal over a 12-hour period, with representative values for each 1-hour segment, show distinctive characteristics between the group that experienced CA and the group that did not. However, compressing the 5 statistical features mentioned above for the entire input window into a single statistical summary may not fully capture the differences in patterns between the groups. Consequently, 5 statistical values were derived by shifting a 4-hour window across the 12-hour input window. This approach allowed us to obtain statistical values for each section using both 6- and 12-hour windows. This feature extraction method, previously used in our research, was shown to enhance CA prediction accuracy by providing a condensed statistical summary of various sections [18].

Gini Index–Based Features

Inspired by the Gini index, which measures statistical variance to indicate income inequality in economics, we propose a method to calculate the imbalance of patterns within each vital sign over the input time steps. This method calculates inequality for each vital sign and uses it as a feature to distinguish between situations where CA occurred and did not occur. Previous research suggests that significant changes in temporal patterns often precede CA, making this imbalance a valuable characteristic variable. The index-based feature formulation is expressed as follows:

$$GV = 1 - (2G/N_{Diff})$$

where G is the index; GV indicates index-based features of each vital sign; x_i and x_j are the values of vital signs within the input range in each vital sign; and N_{Diff} is the number of intervals in the input vital sign. We calculated the Gini index to assess the impurity of each vital sign variation and then performed the normalization step, denoted as GV.

For instance, if the Gini index–based feature value is relatively low, it indicates that the values within the input window are stable. Conversely, if there is a rapid change in HR before CA, the Gini index–based feature value will increase. This pattern change is captured by the Gini index, which measures statistical dispersion.

Step 4: Feature Aggregation and CA Event Labeling

We aggregated multiresolution statistical features, Gini index–based features, and labels to enhance temporal features and achieve better inter-ICU generalizations from the model, utilizing vital signs and specific clinical latent scores.

To select and screen the most relevant and nonredundant features, we used 2 feature selection methods: recursive feature elimination and the Boruta method. Recursive feature elimination identifies the most relevant features for predicting the target by recursively eliminating a small number of features in each iteration [38]. This process helps eliminate collinearity within the proposed framework. The Boruta method assesses the relevance of each feature using statistical testing and shadow features [39].

To overcome the limitations of conventional feature selection methods, we used an ensemble feature screening approach using a majority voting mechanism. This method was applied to a total of 653 features, resulting in the selection of 86 features from the MIMIC-IV database and 94 features from the eICU-CRD database. In this approach, each feature screening method casts 1 vote for a selected feature, and a feature will receive 2 votes if both feature screening methods choose it.

We then aggregated the variables into binary indicators to denote the presence or absence of CA in each class.

Step 5: Model Development

We used a TabNet classifier with features from a 12-hour time window to predict CA events within a 24-hour period. We generated continuous labels every hour indicating the risk of

XSL•FO
RenderX

CA in the next 24 hours and calculated the alarming rate based on whether the alarm was correct or incorrect. Additionally, we applied cost-sensitive learning as an algorithm-level approach to address the extreme imbalances in the MIMIC-IV and eICU-CRD data sets [40]. Cost-sensitive learning was applied to penalize errors in the minority class (the CA group). The TabNet classifier, using cost-sensitive learning, helps reduce bias or variance and improve the stability of machine learning algorithms [41,42]. The minority classes from the MIMIC-IV and eICU-CRD data sets were penalized with a weight of 100. Finally, we set the weight for the CA class to 100 and the learning rate to 0.01.

Step 6: Evaluation

Leave-One-Patient-Out K-Fold Internal Validation

We used leave-one-patient-out (LOPO) K-fold validation to evaluate the individual patient performance of CA prediction and to provide a realistic estimate of model performance on new patients [43]. This method is a variant of K-fold cross-validation, where the folds consist of individual patients. Specifically, we used a 10-fold LOPO validation [44].

Additionally, we established a pseudo-real-time CA evaluation system to assess and compare the real-time CA prediction performance for hospitalized patients, simulating an ICU environment. This system used the LOPO K-fold validation approach to predict CA events within a 24-hour period.

Cross-Data Set External Validation

In this experimental setting, both the proposed method and comparison methods were trained on 1 database and tested on another to evaluate their generalization ability [45]. We alternately used MIMIC-IV and eICU-CRD as the source and target databases to assess generalization performance. To ensure consistency in feature properties across databases, we excluded SAPS-II, which includes laboratory tests. The experiment consisted of 2 steps. First, we trained the models, including the proposed method and comparison models, using MIMIC-IV as the source database. Next, we tested the CA prediction performance on eICU-CRD using the trained models and compared their performance.

To assess the generalization ability and minimize the impact of specific group characteristics in the learning data, we conducted another cross-data set external validation. We trained the models using eICU-CRD as the source database and evaluated CA prediction performance on MIMIC-IV as the target database.

Subgroup Analysis

We evaluated the CA prediction performance for each ICU subtype to determine if there were differences in performance between the proposed method and the comparative models for each ICU cluster. The common ICU subtypes across MIMIC-IV and eICU-CRD were general, cardiac, neuro, and trauma. Neuro and trauma ICUs were excluded from this analysis due to the low number or absence of admissions of patients with CA. We used 10-fold LOPO cross-validation for subgroup evaluation. In this process, the entire data set assigned to each fold was used for training, and during testing, performance was evaluated by categorizing patients according to their ICU subtype.

Baseline Models

To evaluate the performance of the proposed method, we used the following baseline models: National Early Warning Score (NEWS), SOFA, SAPS-II, logistic regression, k-nearest neighbors, multilayer perceptron, LGBM, RNN, and reverse time attention. Details of these baseline models can be found in Multimedia Appendix 2.

Evaluation Metrics

We assessed the performance of the proposed and baseline methods using AUROC, event recall, false alarm rate, and sensitivity. Our goal was to evaluate the proposed method in a clinically relevant context by focusing on the percentage of CA events detected and the rate of false alarms. We specifically measured event recall [46], which quantifies whether the CA prediction system correctly triggered an alarm in the period preceding a CA event.

$$\text{ER} = N_{\text{Captured}}/N_{\text{Total}}$$

where ER is event recall; N_{Captured} implies the number of captured events; and N_{Total} indicates the number of total CA events.

Next, the false alarm rate was defined as the fraction of alarms that failed to detect an actual event, to investigate whether operational costs were being wasted. This concept is analogous to exon prediction in gene discovery. The false alarm rate is evaluated as follows:

$$\text{FAR} = 1 - (N_{\text{True}}/N_{\text{Alarm}})$$

where FAR is the false alarm rate; N_{True} is the number of true alarms; and N_{Alarm} is the number of total alarms in the CA prediction system.

Explainable Predictions

We extracted both local and global interpretability information by examining the decision masks of TabNet. After determining the impact of each feature using the proposed model, we summarized and visualized the top 25 features with the highest mean values. Additionally, we visualized the impact of features over time using a heatmap and tracked changes in the features with the highest values. To compare the differences in impact between non-CA and CA groups, we conducted a statistical test. Specifically, an independent *t* test with false discovery rate (FDR) correction was used to assess the differences in interpretability information between the 2 groups.

Statistical Analysis

Differences in patient characteristics, such as age, ICU length of stay, and vital signs, between the non-CA and CA groups were evaluated using independent *t* tests. To compare performance metrics between the baseline and proposed models, we used the Kruskal-Wallis test, followed by the honestly significant difference (HSD) test for post hoc analysis. The differences in interpretable information between the non-CA and CA groups were evaluated using an independent *t* test with FDR correction. A significance level of 5% ($P<.05$) was used for all analyses.

XSL•FO
RenderX

Results

Patient Characteristics

The patient characteristics are presented as means and SDs in Table 1.

In the 12-hour time window for MIMIC-IV, age did not differ significantly between the CA and non-CA groups. However, the ICU length of stay was statistically different between the 2 groups (*P*=.02). Significant differences were observed in HR (*P*<.001), RR (*P*<.001), SBP (*P*<.001), DBP (*P*=.01), SpO_2 (*P*<.001), and temperature (*P*<.001). In the eICU-CRD data set, age (*P*=.11) and ICU length of stay (*P*=.21) were not considered significant because their *P* values were greater than .05. Except for HR, which was not significant in either group (<.001 in the MIMIC-IV data set and .41 in the eICU-CRD data set), the other variables had significance levels (ie, *P*<.05; see full data in Table 1).

Next, we provided the patient characteristics for each ICU subtype, specifically general and cardiac ICUs, as shown in Multimedia Appendices 3 and 4. In the 12-hour time window for MIMIC-IV, the characteristics of general and cardiac ICUs were similar to those of the overall ICU population between the CA and non-CA groups, except for ICU length of stay (hours). In the general ICU, the length of stay was significantly longer in the CA group (*P*=.006). In the 12-hour time window for eICU-CRD, notable discrepancies were observed in the characteristics of general and cardiac ICUs between the CA and non-CA groups with respect to HR. In the general ICU, the CA group had a significantly higher HR compared with the non-CA group (*P*<.001). By contrast, the non-CA group in the cardiac ICU had a lower HR, although this difference was not statistically significant (*P*=.41).

Table 1. Demographic information of patients from MIMIC-IV[a] and eICU-CRD[b].

Characteristic	MIMIC-IV			eICU-CRD		
	Cardiac arrest (n[c]=77)	Noncardiac arrest (n=1474)	*P* value	Cardiac arrest (n=106)	Noncardiac arrest (n=3541)	*P* value
Age (year), mean (SD)	68.53 (13.63)	67.64 (13.57)	.74	60.03 (16.52)	62.65 (15.83)	.11
Intensive care unit length of stay (hours), mean (SD)	318.90 (346.82)	273.26 (142.96)	.02	199.82 (198.01)	175.28 (144.14)	.21
Vital signs, mean (SD)						
Heart rate (beats/minute)	88.79 (17.60)	87.10 (17.30)	<.001	87.45 (19.54)	87.35 (17.30)	.41
Respiratory rate (breaths/minute)	21.26 (5.68)	20.99 (5.70)	<.001	20.10 (5.90)	19.76 (4.85)	<.001
Systolic blood pressure (mmHg)	111.21 (22.60)	118.03 (21.27)	<.001	118.88 (21.21)	125.68 (21.73)	<.001
Diastolic blood pressure (mmHg)	59.41 (14.17)	59.64 (13.70)	.01	63.47 (14.06)	68.49 (13.86)	<.001
Oxyhemoglobin saturation (SpO_2)	97.22 (3.57)	96.92 (2.87)	<.001	96.94 (4.01)	96.55 (2.82)	<.001
Temperature (°C)	36.88 (0.91)	37.12 (0.64)	<.001	36.94 (0.94)	36.90 (0.58)	<.001

[a]MIMIC: Medical Information Mart for Intensive Care.

[b]eICU-CRD: eICU-Collaborative Research Database.

[c]n: number of ICU stays.

Feature Screening Strategy

Table 2 illustrates the efficacy of the proposed methodology, both in its original form and when combined with the ensemble feature screening process. Initially, the proposed framework was trained and validated using all features for CA prediction, with validation performed through a 10-fold LOPO cross-validation approach. The framework achieved AUROC values of 0.75, 0.99, 0.80, and 0.80 for event recall, false alarm rate, sensitivity, and specificity, respectively, on the MIMIC-IV data set. For the eICU-CRD data set, the AUROC values were 0.78, 0.99, 0.45, and 0.99, respectively, for event recall, false alarm rate, sensitivity, and specificity.

To minimize the risk of overfitting in the proposed method, feature screening was essential. An ensemble feature screening method was used to identify the optimal feature set for the best results. The selected feature sets from the ensemble screening on the MIMIC-IV data set were then incorporated into the proposed framework. This adjustment led to AUROC values of 0.79, event recall of 0.99, false alarm rate of 0.77, and sensitivity of 0.89 for the proposed framework. The selected feature sets from the ensemble feature screening on the eICU-CRD data set were incorporated into the proposed framework. This adjustment resulted in AUROC values of 0.80, event recall of 0.99, false alarm rate of 0.36, and sensitivity of 0.99. The proposed ensemble feature screening approach demonstrated superior performance compared with using all feature sets.

XSL•FO
RenderX

Table 2. Performance comparison without and with ensemble feature screening methods along with the proposed framework using MIMIC-IV[a] and eICU-CRD[b].

Method	MIMIC-IV			eICU-CRD		
	Event recall (↑[c])	False alarm rate (↓[d])	Sensitivity (↑)	Event recall (↑)	False alarm rate (↓)	Sensitivity (↓)
Proposed method, mean (SD)	0.99 (0.00)	0.80 (0.04)	0.80 (0.09)	0.99 (0.01)	0.45 (0.14)	0.99 (0.01)
Proposed method with feature screening, mean (SD)	0.99 (0.00)	0.77 (0.05)	0.89 (0.06)	0.99 (0.00)	0.36 (0.16)	0.99 (0.00)

[a]MIMIC: Medical Information Mart for Intensive Care.

[b]eICU-CRD: eICU-Collaborative Research Database.

[c]The ↑ symbol indicates that a higher value for the evaluation metric corresponds to a more meaningful or effective model.

[d]The ↓ symbol indicates that a lower value for the evaluation metric corresponds to more meaningful or effective model performance.

Predictive Performance

This section presents the results of CA predictive performance. We evaluated the performance using metrics including AUROC, event recall, false alarm rate, and sensitivity.

In the 12-hour time window from the MIMIC-IV database, we compared the AUROC of the proposed framework with that of baseline methods to investigate CA predictive performance. The proposed method achieved a higher overall AUROC value compared with the baseline methods, as shown in Figure 3. The AUROC results using the proposed method were statistically higher than those of the comparison methods (χ^2_{15}=68.67), as determined by the Kruskal-Wallis test with HSD post hoc analysis (see Multimedia Appendix 5). Additionally, we compared other performance metrics, including event recall, false alarm rate, and sensitivity, to assess effectiveness in a clinically relevant context. This evaluation focused on detecting CA events within a 24-hour period and minimizing false alarm rates, as shown in Table 3. The proposed method achieved statistically higher performance in event recall and sensitivity (χ^2_{15}=90.34 for event recall and χ^2_{15}=38.70 for sensitivity), as shown in Multimedia Appendices 6 and 7. Additionally, the proposed method demonstrated a statistically lower false alarm rate (χ^2_{15}=110.00), as detailed in Multimedia Appendix 8.

We compared the AUROC of the comparison methods and the proposed framework using the 12-hour time window from the eICU-CRD data set. The proposed method achieved statistically higher performance, with an overall AUROC value as shown in Figure 3 and Multimedia Appendix 9 (χ^2_{14}=81.38). Additionally, we evaluated other performance metrics, including event recall, false alarm rate, and sensitivity, as detailed in Table 4. The proposed method achieved statistically higher values for event recall and sensitivity compared with other methods, as demonstrated by the Kruskal-Wallis test with HSD (χ^2_{14}=90.75 for event recall, χ^2_{14}=100.86 for sensitivity), as shown in Multimedia Appendices 10 and 11. The proposed method also achieved a lower false alarm rate than the comparison methods, except for SAPS-II and LGBM.

Figure 3. Comparison of AUROC performance among baseline models and the proposed method from MIMIC-IV and eICU-CRD. (A) AUROC from MIMIC-IV and (B) AUROC from eICU-CRD. AUROC: area under the receiver operating characteristic curve; DEWS: Deep Early Warning Score; eICU-CRD: eICU-Collaborative Research Database; FS: feature screening; KNN: K-nearest neighbors; LGBM: light gradient boosting method; LR: logistic regression; MIMIC: Medical Information Mart for Intensive Care; MLP: multilayer perceptron; NEWS: National Early Warning Score; RETAIN: reverse time attention; SAPS: Simplified Acute Physiology Score; SOFA: Sequential Organ Failure Assessment.

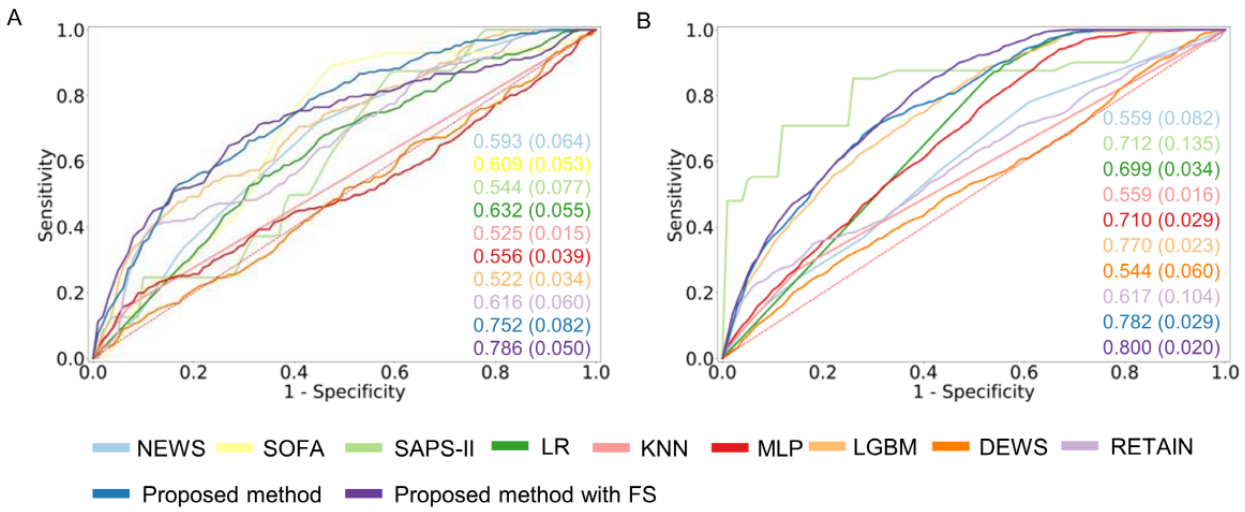

XSL•FO
RenderX

Table 3. Comparison of LOPO[a] cross-validation performance using MIMIC-IV[b].

Model	Event recall (↑[c])	False alarm rate (↓[d])	Sensitivity (↑)
National Early Warning Score ≥5, mean (SD)	0.87 (0.12)	0.89 (0.09)	0.39 (0.10)
Sequential Organ Failure Assessment ≥6, mean (SD)	0.71 (0.13)	0.90 (0.08)	0.59 (0.15)
Simplified Acute Physiology Score-II ≥32, mean (SD)	0.91 (0.10)	0.90 (0.10)	0.91 (0.10)
Logistic regression, mean (SD)	0.99 (0.04)	0.90 (0.10)	0.69 (0.21)
k-Nearest neighbors, mean (SD)	0.05 (0.06)	0.96 (0.10)	0.01 (0.01)
Multilayer perceptron, mean (SD)	0.36 (0.13)	0.88 (0.04)	0.03 (0.02)
Light gradient boosting method, mean (SD)	0.51 (0.17)	0.87 (0.01)	0.17 (0.08)
DEWS[e] ≥2.9, mean (SD)	0.91 (0.11)	0.92 (0.08)	0.44 (0.10)
DEWS ≥3, mean (SD)	0.91 (0.11)	0.92 (0.02)	0.44 (0.10)
DEWS ≥7.1, mean (SD)	0.85 (0.10)	0.92 (0.03)	0.37 (0.09)
DEWS ≥8, mean (SD)	0.84 (0.10)	0.92 (0.05)	0.35 (0.09)
DEWS ≥18.2, mean (SD)	0.83 (0.12)	0.92 (0.09)	0.29 (0.09)
DEWS ≥52.8, mean (SD)	0.69 (0.13)	0.92 (0.07)	0.18 (0.06)
Reverse time attention, mean (SD)	0.98 (0.05)	0.92 (0.03)	0.94 (0.10)
Proposed method, mean (SD)	0.99 (0.00)	0.80 (0.04)	0.80 (0.09)
Proposed method with feature screening, mean (SD)	0.99 (0.00)	0.77 (0.05)	0.89 (0.06)

[a]LOPO: leave-one-patient-out.

[b]MIMIC: Medical Information Mart for Intensive Care.

[c]The ↑ symbol indicates that a higher value for the evaluation metric corresponds to a more meaningful or effective model.

[d]The ↓ symbol indicates that a lower value for the evaluation metric corresponds to more meaningful or effective model performance.

[e]DEWS: Deep Learning–Based Early Warning Score.

XSL•FO
RenderX

Table 4. Comparison of LOPO[a] cross-validation performance using eICU-CRD[b].

Model	Event recall (↑[c])	False alarm rate (↓[d])	Sensitivity (↑)
National Early Warning Score ≥5, mean (SD)	0.99 (0.02)	0.51 (0.16)	0.68 (0.12)
Simplified Acute Physiology Score-II ≥32, mean (SD)	0.70 (0.21)	0.44 (0.19)	0.70 (0.24)
Logistic regression, mean (SD)	0.99 (0.02)	0.48 (0.14)	0.86 (0.20)
k-Nearest neighbors, mean (SD)	0.19 (0.08)	0.43 (0.27)	0.02 (0.01)
Multilayer perceptron, mean (SD)	0.54 (0.11)	0.49 (0.14)	0.08 (0.02)
Light gradient boosting method, mean (SD)	0.94 (0.05)	0.42 (0.14)	0.67 (0.05)
DEWS[e] ≥2.9, mean (SD)	0.98 (0.03)	0.56 (0.18)	0.63 (0.09)
DEWS ≥3, mean (SD)	0.98 (0.03)	0.56 (0.18)	0.63 (0.09)
DEWS ≥7.1, mean (SD)	0.96 (0.04)	0.55 (0.18)	0.55 (0.09)
DEWS ≥8, mean (SD)	0.96 (0.04)	0.55 (0.18)	0.54 (0.09)
DEWS ≥18.2, mean (SD)	0.92 (0.06)	0.55 (0.18)	0.45 (0.08)
DEWS ≥52.8, mean (SD)	0.86 (0.09)	0.55 (0.18)	0.29 (0.06)
Reverse time attention, mean (SD)	0.99 (0.00)	0.50 (0.14)	0.99 (0.01)
Proposed method, mean (SD)	0.99 (0.01)	0.45 (0.14)	0.99 (0.01)
Proposed method with feature screening, mean (SD)	0.99 (0.00)	0.36 (0.16)	0.99 (0.01)

[a]LOPO: leave-one-patient-out.

[b]eICU-CRD: eICU-Collaborative Research Database.

[c]The ↑ symbol indicates that a higher value for the evaluation metric corresponds to a more meaningful or effective model.

[d]The ↓ symbol indicates that a lower value for the evaluation metric corresponds to more meaningful or effective model performance.

[e]DEWS: Deep Learning–Based Early Warning Score.

Subgroup Analysis

We evaluated the performance of the comparison models and the proposed framework across different ICU types, including general and cardiac ICUs. Most ICU types showed similar performance, except for patients in the cardiac ICU within the eICU-CRD data set. As shown in Figure 4, the proposed method demonstrated statistically higher performance compared with the comparative models across all ICU types in both MIMIC-IV and eICU-CRD data sets. The comparisons by ICU type are as follows: general ICU in MIMIC-IV (χ^2_8=29.67), cardiac ICU in MIMIC-IV (χ^2_8=44.22), and cardiac ICU in eICU-CRD (χ^2_8=45.07). Detailed statistical comparison results are presented in Multimedia Appendices 12-14.

XSL•FO
RenderX

Figure 4. Model performance in difference patient cohorts from MIMIC-IV and eICU-CRD. (A) AUROC on ICU types of MIMIC-IV. (B) AUROC on ICU types of eICU-CRD. Boxes in the box plot show IQR and the cross marks are outliers with values that lie outside the minimum and maximum ranges of the whiskers, where minimum = Q1 - 1.5 × IQR and maximum = Q3 + 1.5 × IQR. * Statistically significant (*P*<.05). AUROC: area under the receiver operating characteristic curve; DEWS: Deep Learning–Based Early Warning Score; eICU-CRD: eICU-Collaborative Research Database; ICU: intensive care unit; KNN: k-nearest neighbors; LGBM: light gradient boosting method; LR: logistic regression; MIMIC: Medical Information Mart for Intensive Care; MLP: Multilayer perceptron; NEWS: National Early Warning Score; Q1: first quartile; Q3: third quartile; RETAIN: reverse time attention; SAPS: Simplified Acute Physiology Score.

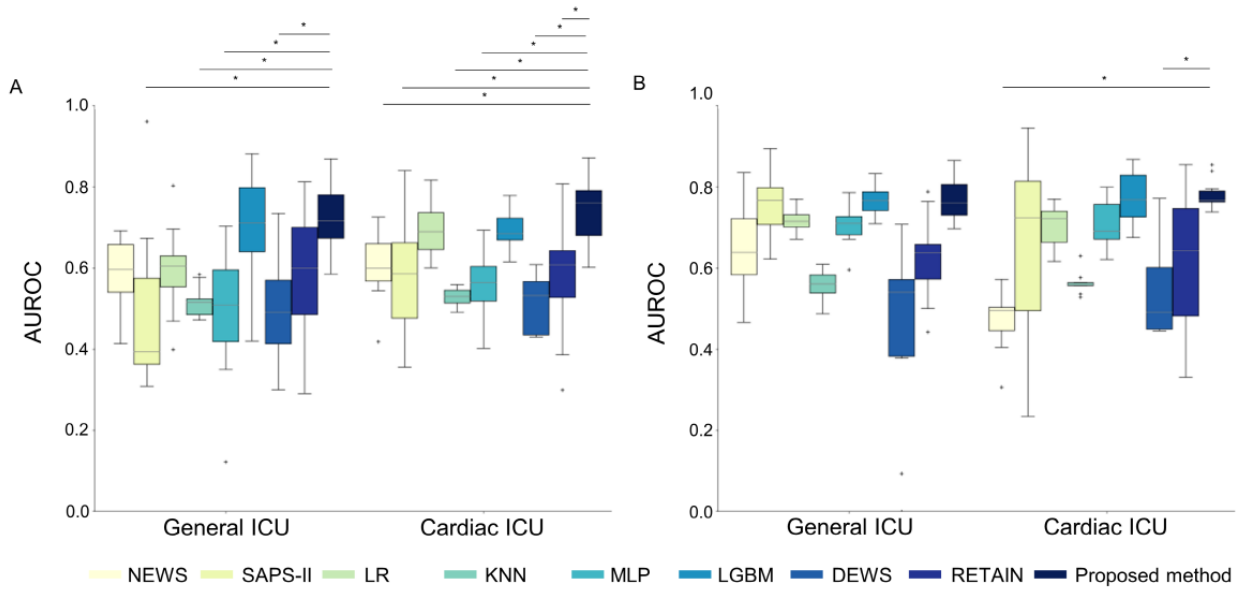

External Validation

We conducted cross-data set external validation to assess the generalization ability of the proposed method and comparison models. After training on the MIMIC-IV data set, we evaluated the clinical validity of predicting CA within 24 hours using the eICU-CRD data set as the test set. Figure 5 and Table 5 present the external validation results for conventional systems (including NEWS, SOFA, and SAPS-II), machine learning–based comparison methods, and deep learning–based scoring systems. The proposed method achieved higher AUROC, event recall, and a lower false alarm rate compared with the comparison methods.

Conversely, we tested the proposed framework by evaluating a cohort from a general hospital setting (eICU-CRD) and a cohort with heart disease (MIMIC-IV). The results showed that the proposed framework achieved superior performance in AUROC, false alarm rate, and sensitivity, as detailed in Table 6.

Figure 5. Cross–data set external validation AUROC performance. (A) eICU after training MIMIC-IV. (B) MIMIC-IV after training eICU-CRD. AUROC: area under the receiver operating characteristic curve; DEWS: Deep Learning–Based Early Warning Score; eICU-CRD: eICU-Collaborative Research Database; KNN: k-nearest neighbors; LGBM: light gradient boosting method; LR: logistic regression; MIMIC: Medical Information Mart for Intensive Care; MLP: multilayer perceptron; NEWS: National Early Warning Score; RETAIN: reverse time attention.

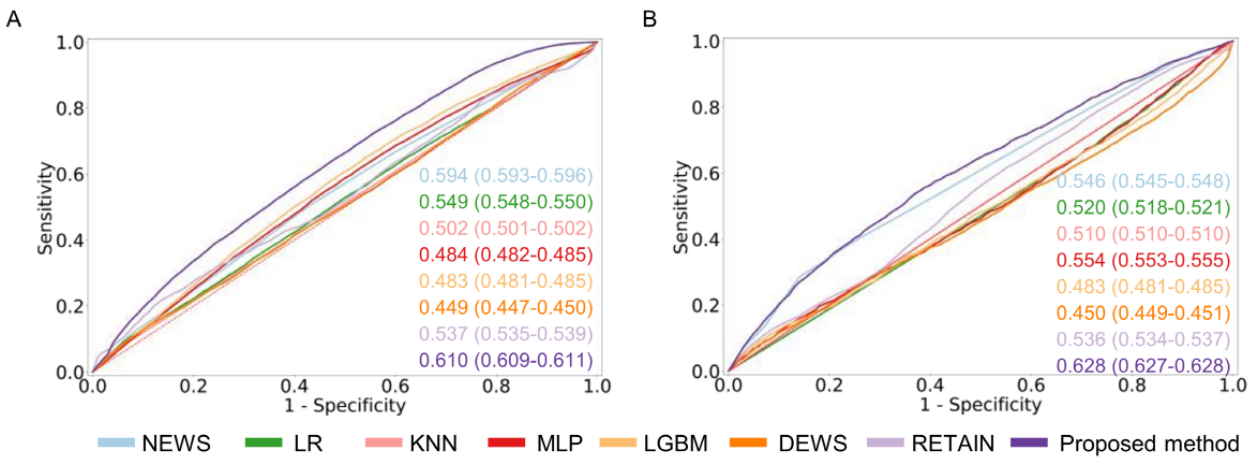

XSL•FO
RenderX

Table 5. Cross-data set external validation performance using eICU-CRD[a] after training MIMIC-IV[b].

Model	Event recall (↑[c]) (95% CI)	False alarm rate (↓[d]) (95% CI)	Sensitivity (↑) (95% CI)	Brier score (95% CI)
National Early Warning Score ≥5	0.64 (0.64-0.64)	0.93 (0.93-0.93)	0.20 (0.20-0.20)	0.18 (0.18-0.18)
Logistic regression	0.74 (0.74-0.75)	0.94 (0.94-0.94)	0.25 (0.25-0.26)	0.23 (0.22-0.23)
k-Nearest neighbors	0.02 (0.02-0.02)	0.89 (0.89-0.90)	0.00 (0.00-0.00)	0.05 (0.05-0.05)
Multilayer perceptron	0.08 (0.08-0.09)	0.93 (0.93-0.93)	0.01 (0.01-0.01)	0.06 (0.06-0.06)
Light gradient boosting method	0.14 (0.14-0.14)	0.92 (0.91-0.92)	0.03 (0.03-0.03)	0.06 (0.06-0.06)
DEWS[e] ≥2.9	0.84 (0.84-0.84)	0.97 (0.97-0.97)	0.32 (0.32-0.32)	0.09 (0.08-0.09)
DEWS ≥3	0.84 (0.84-0.84)	0.97 (0.97-0.97)	0.32 (0.31-0.32)	0.09 (0.08-0.09)
DEWS ≥7.1	0.78 (0.78-0.79)	0.97 (0.97-0.97)	0.25 (0.25-0.25)	0.09 (0.08-0.09)
DEWS ≥8	0.77 (0.77-0.77)	0.97 (0.97-0.97)	0.24 (0.24-0.24)	0.09 (0.08-0.09)
DEWS ≥18.2	0.69 (0.68-0.69)	0.97 (0.97-0.97)	0.17 (0.17-0.17)	0.08 (0.08-0.09)
DEWS ≥52.8	0.54 (0.54-0.54)	0.97 (0.97-0.97)	0.09 (0.09-0.10)	0.08 (0.08-0.09)
Reverse time attention	0.94 (0.97-0.98)	0.95 (0.95-0.95)	0.98 (0.98-0.98)	0.40 (0.40-0.40)
Proposed method	0.98 (0.98-0.99)	0.91 (0.90-0.91)	0.99 (0.99-0100)	0.30 (0.30-0.30)

[a]eICU-CRD: eICU-Collaborative Research Database.

[b]MIMIC: Medical Information Mart for Intensive Care.

[c]The ↑ symbol indicates that a higher value for the evaluation metric corresponds to a more meaningful or effective model.

[d]The ↓ symbol indicates that a lower value for the evaluation metric corresponds to more meaningful or effective model performance.

[e]DEWS: Deep Learning–Based Early Warning Score.

Table 6. Cross-data set external validation performance using MIMIC-IV[a] after training eICU-CRD[b].

Model	Event recall (↑[c]) (95% CI)	False alarm rate (↓[d]) (95% CI)	Sensitivity (↑) (95% CI)	Brier score (95% CI)
National Early Warning Score ≥5	0.99 (0.98-0.99)	0.98 (0.98-0.98)	0.84 (0.84-0.84)	0.75 (0.75-0.75)
Logistic regression	0.99 (0.98-0.99)	0.98 (0.98-0.98)	0.95 (0.95-0.95)	0.91 (0.91-0.91)
k-Nearest neighbors	0.20 (0.19-0.20)	0.98 (0.98-0.98)	0.01 (0.01-0.01)	0.04 (0.04-0.04)
Multilayer perceptron	0.69 (0.68-0.69)	0.98 (0.98-0.98)	0.10 (0.09-0.10)	0.08 (0.08-0.08)
Light gradient boosting method	0.86 (0.86-0.86)	0.98 (0.98-0.98)	0.58 (0.57-0.58)	0.40 (0.40-0.40)
DEWS[e] ≥2.9	0.95 (0.94-0.95)	0.98 (0.98-0.98)	0.60 (0.60-0.60)	0.25 (0.25-0.25)
DEWS ≥3	0.95 (0.95-0.95)	0.98 (0.98-0.98)	0.60 (0.60-0.60)	0.25 (0.25-0.25)
DEWS ≥7.1	0.92 (0.92-0.93)	0.98 (0.98-0.98)	0.54 (0.54-0.54)	0.25 (0.25-0.25)
DEWS ≥8	0.92 (0.92-0.92)	0.98 (0.98-0.98)	0.52 (0.52-0.53)	0.25 (0.25-0.25)
DEWS ≥18.2	0.91 (0.91-0.91)	0.98 (0.98-0.98)	0.45 (0.45-0.45)	0.25 (0.25-0.25)
DEWS ≥52.8	0.83 (0.83-0.83)	0.98 (0.98-0.98)	0.29 (0.29-0.29)	0.25 (0.25-0.25)
Reverse time attention	0.99 (0.98-0.99)	0.98 (0.98-0.98)	0.99 (0.98-0.99)	0.76 (0.76-0.76)
Proposed method	0.99 (0.98-0.99)	0.50 (0.50-0.50)	0.99 (0.98-0.99)	0.31 (0.31-0.31)

[a]MIMIC: Medical Information Mart for Intensive Care.

[b]eICU-CRD: eICU-Collaborative Research Database.

[c]The ↑ symbol indicates that a higher value for the evaluation metric corresponds to a more meaningful or effective model.

[d]The ↓ symbol indicates that a lower value for the evaluation metric corresponds to more meaningful or effective model performance.

[e]DEWS: Deep Learning–Based Early Warning Score.

XSL•FO
RenderX

Clinical Interpretability

We utilized local and global interpretability information from TabNet to assess the impact of each feature on the proposed model's output. Positive importance values indicated features that increased prediction scores, while negative values indicated features that decreased prediction scores. Figure 6 displays the top 25 features of the proposed model.

Specifically, the features EWS_SpO_2_1_6 h_Min (P<.001), Temp_2 h (P<.001), and HR_5_8 h_Skewness (P<.001) had relatively significant impacts on the performance of the proposed method, as shown in Figure 6A. Additionally, most of the top 25 influential features were based on vital signs and multiresolution statistical features. Of these, 20 (all features except DBP_9_12h_Skewness, EWS_HR_8h, and EWS_SBP_9_12h_Max: P<.001; DBP_9_12h_Skewness: P=.02; EWS_HR_8h: P=.03; and EWS_SBP_9_12h_Max: P=.01) out of the 25 global interpretability features showed statistically significant differences between the non-CA and CA groups, as determined by an independent t test with FDR correction, detailed in Multimedia Appendix 15. In the multiresolution statistical features based on the sliding window, temperature (all P<.001), EWS-SpO_2 (all P<.001), EWS-RR (all P<.001), EWS-total (EWS_9_12 h_max: P<.001), EWS-SBP (all P<.001), HR (all P<.001), TEMP (all P<.001), SpO_2 (SpO_2_5_8 h_Min: P<.001), and DBP (P=.02) had statistically significant effects on the proposed model, as shown in Figure 6C [47]. Additionally, the Gini index–based function, calculated over a 12-hour window for SpO_2, emerged as the most important indicator for both the non-CA and CA groups, as shown in Figure 6D [48].

XSL•FO
RenderX

Figure 6. Feature inspection on MIMIC-IV. (A) Global feature impact values produced by the proposed method. (B) Vital sign–based feature set between the non-CA and CA groups. (C) Multiresolution feature set between the non-CA and CA groups. (D) Gini index–based feature set between the non-CA and CA groups. CA: cardiac arrest; DBP: diastolic blood pressure; EWS: Early warning score; HR: heart rate; Max: maximum; Min: minimum; MIMIC: Medical Information Mart for Intensive Care; MR: medical record; RR: respiratory rate; SBP: systolic blood pressure; SpO_2: oxyhemoglobin saturation; TEMP: temperature.

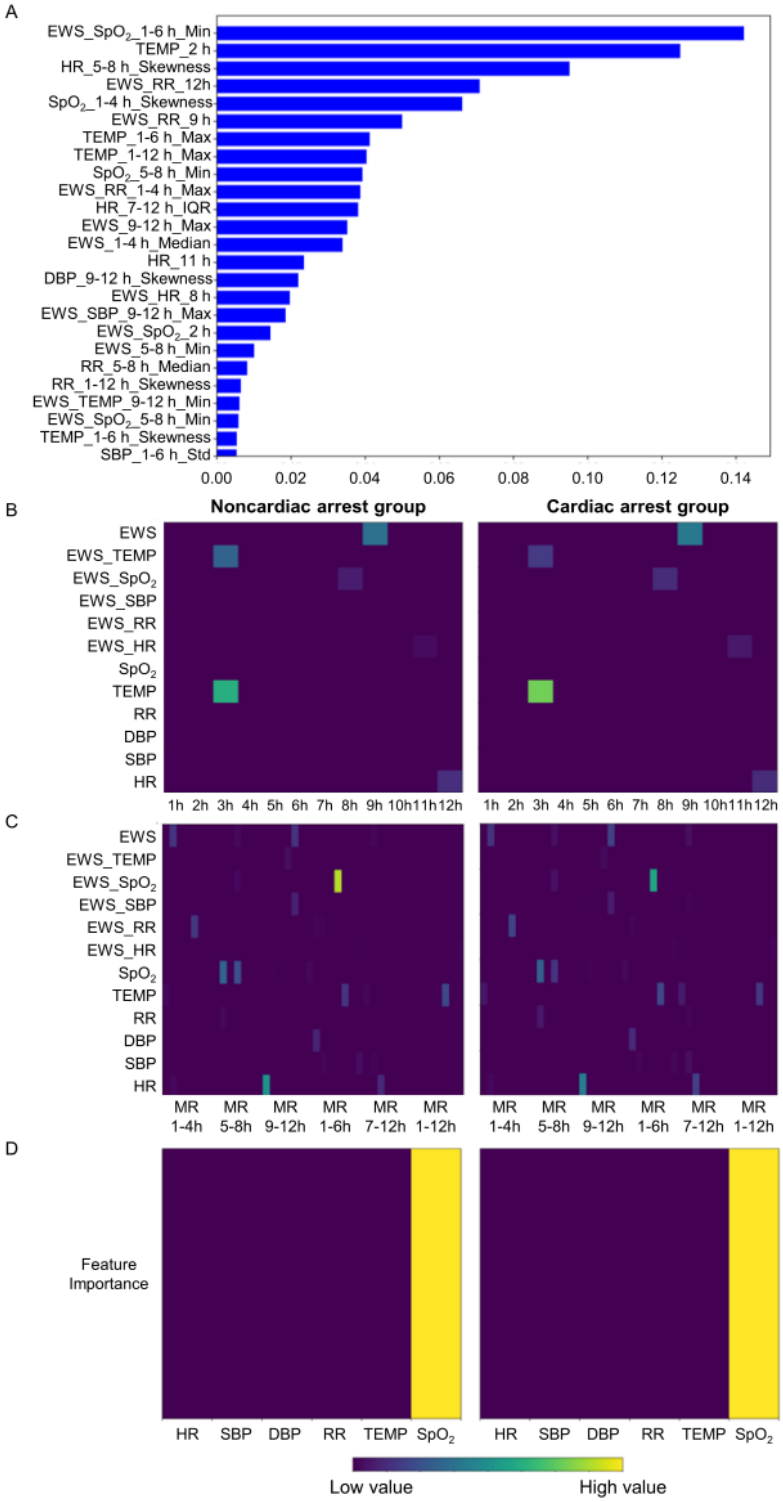

Computational Cost Analysis

We analyzed the computational complexity and time required for model training and prediction of both the comparison models and the proposed framework to assess their applicability in real-time critical care settings. Computational complexity measures the amount of computing resources consumed by an algorithm [49] and can be expressed in terms of both time and space. In essence, computational complexity measures how quickly or slowly an algorithm performs relative to a given input size. It is specified in terms of time complexity and space complexity. Time complexity refers to the amount of

XSL•FO
RenderX

computational time an algorithm requires, while space complexity describes the amount of additional memory needed to execute the algorithm. Both types of complexity are expressed relative to a specific input size, which includes factors such as the number of training examples, the number of features, the depth and number of trees, and the length and dimension of the input.

Table 7 illustrates the computational complexity of both the comparative models and the proposed framework, categorized by time and space complexity. Additionally, the table depicts the actual training time and the time required to generate results with the trained model, expressed as training and inference time, respectively. As a result, the proposed method required 31.82 and 60.51 seconds of training time per epoch on the MIMIC-IV and eICU-CRD data sets, respectively. Additionally, the prediction time for the proposed method was 0.25 and 0.54 seconds on the MIMIC-IV and eICU-CRD data sets, respectively. In comparison, LGBM demonstrated the shortest training time and fastest prediction time, making it the most lightweight algorithm. The proposed method required the least training and prediction time among the deep learning–based methods. Although the training time was longer than that of LGBM, the prediction time was not significantly different from that of LGBM (MIMIC-IV: P=.11, eICU-CRD: P=.08).

Table 7. Comparisons of computational complexity and actual training and inference time.

Method	Computational complexity		MIMIC-IV[a]		eICU-CRD[b]	
	Time	Space	Time per training (seconds)	Inference time (seconds)	Time per training (seconds)	Inference time (seconds)
National Early Warning Score	—[c]	—	—	—	—	—
Simplified Acute Physiology Score-II	—	—	—	—	—	—
Logistic regression	$O(n^{d}f^{e})$	$O(f)$	27.69	0.02	52.62	0.03
k-Nearest neighbors	$O[fn\log(n)]$	$O(nf)$	0.04	33.18	0.08	136.24
Multilayer perceptron	$O(nfl^{2}{}_{oi})$	—	624.98	2.51	5061.21	6.48
Light gradient boosting method	$O[n\log(n)]k^{f}$	$O[n\log(n)]$	4.94	0.06	7.17	0.09
Deep Learning–Based Early Warning Score	$O(N^{g}D^{2h})$	$O(ND^{2})$	1384.37	0.87	2907.94	1.33
Reverse time attention	$O(ND^{2})$	$O(ND^{2})$	1497.30	0.06	2906.91	0.10
Proposed method	$O(ND^{2})$	$O(ND^{2})$	31.82	0.25	60.51	0.54

[a]MIMIC: Medical Information Mart for Intensive Care.
[b]eICU-CRD: eICU-Collaborative Research Database.
[c]Not available.
[d]n: number of training examples.
[e]f: number of features.
[f]k: number of trees.
[g]N: the length of the input.
[h]D: the dimension of the input.

Discussion

Principal Findings

Clinicians can use the proposed model to make consistent clinical decisions for patients in various ICUs with similar organ failure types, facilitating efficient management [50]. This study developed and validated an ensemble-based model capable of predicting CA events 24 hours in advance, regardless of different hospitalization settings, patient populations, and ICU subtypes within each database, as well as across external validations. Therefore, the proposed method can help reduce the number of CA events and ICU deaths without compromising performance due to heterogeneity. Additionally, it lowers the false alarm rate compared with existing methods, providing clinicians with more reliable CA event alarms and better preparedness in hospitals.

For CA prediction across different patient populations, the proposed method, utilizing 3 types of features, demonstrated statistically superior performance in AUROC, sensitivity, false alarm rate, and event recall compared with the comparison models. The proposed method achieved significant performance improvements in predicting CA events for the entire ICU patient group (AUROC: P<.001; sensitivity: P<.001; and event recall: P<.001) and showed even greater statistical differences in high-risk patients with CA (AUROC: P<.001; sensitivity: P<.001; false alarm rate: P<.001; and event recall: P<.001). Therefore, the proposed method is crucial for predicting significant CA events in both the high-risk CA group and the entire ICU population.

We evaluated the performance of both the comparison models and the proposed framework across different ICU types, including general and cardiac ICUs. We found similar

XSL•FO
RenderX

performance levels across most ICU types, except for patients in the cardiac ICU of the eICU-CRD.

For the MIMIC-IV data set, which includes patients with cardiac-related diseases in the ICU, the class distribution rate between the CA and non-CA groups is 3.2% (24/738; CA group, n=24 and non-CA group, n=738) in the general ICU. In the cardiac ICU, the ratio is 5.8% (43/741; CA group, n=43 and non-CA group, n=741), which is slightly higher than in the general ICU. The AUROC performance improved in all cases except for LGBM in the cardiac ICU, where the distribution trend increased, although LGBM's performance was similar between the general and cardiac ICUs. Overall, the AUROC performance for each ICU type, including both general and cardiac ICUs, was consistent with the performance observed across all patients. The AUROC performance of NEWS and SAPS-II in the cardiac ICU improved by 0.02 and 0.10, respectively, compared with their performance in the general ICU. This finding aligns with the analysis results in MIMIC-IV, which indicated that vital sign–based features had a strong influence on performance, as mentioned in the results of the analysis of the effect of the feature set on performance in MIMIC-IV.

When analyzing the AUROC for each ICU type in the eICU-CRD, which includes all hospitalized ICU patients, the performance patterns differed from those observed in the MIMIC-IV data set. The class distribution ratio between the CA and non-CA groups in both the general and cardiac ICUs in the eICU-CRD is similar to that in MIMIC-IV. Specifically, in the general ICU of eICU-CRD, the class distribution rate between the CA and non-CA groups is 2.35% (60/2554; CA group, n=60 and non-CA group, n=2554). In the cardiac ICU, the ratio is 5.80% (43/741; CA group, n=43 and non-CA group, n=741). However, the AUROC performance of NEWS and SAPS-II decreased by 0.17 and 0.10, respectively. Additionally, the AUROC performance of the other comparison models remained similar, despite the increased class ratio of the CA group in the cardiac ICU of the eICU-CRD.

Strengths

This study has several strengths. First, we assessed the performance of predicting CA across different patient populations and hospital settings. Moreover, we evaluated the predictive performance of CA over the entire hospitalization period for each patient, rather than relying on event-based assessments that represent the patient's CA or normal condition. Second, we optimized the hyperparameter values of each machine learning model through iterative grid search. Hyperparameter tuning has been shown to improve the performance of these models. We proposed an interpretable and calibrated ensemble approach using TabNet with different cost weights for each class to predict CA events within 24 hours. The proposed feature sets, defined according to the AI development cycle, and TabNet demonstrated significantly higher CA prediction accuracy compared with the baseline models (all $P<.001$). Therefore, clinicians would have sufficient time to respond to CA events when using the proposed model. Third, we examined the extent to which bias in the data influenced our framework. Ultimately, our proposed framework outperformed all comparative models when trained and tested on these diverse populations. This demonstrates that the framework has strong generalization capabilities and is not overly affected by the confounding effects of specific group characteristics in the learning data.

We analyzed the computational complexity and time required for model training and prediction to determine whether the comparative models and the proposed framework could be applied in real-time within the critical care field. Although the training time for the proposed method was longer than that for LGBM, the prediction time was comparable. Therefore, the proposed method can deliver prediction results without delay in response time, making it suitable for clinical use in critical care settings and integration into existing systems.

Limitations

This study has several limitations. Although the proposed model demonstrated superior performance in external validation, further validation using independent data sets is necessary. Additionally, the ensemble approach based on TabNet was developed without feature screening. However, as the high-performance results indicate, this did not greatly impact the model's effectiveness. Although the proposed method demonstrated higher precision, sensitivity, specificity, AUROC, and a lower false alarm rate, the false alarm rate remains high, which is a limitation of our study. Further research is needed to develop feature-generation methods and models that can further reduce the false alarm rate. In the future, the proposed model could be optimized to include feature screening. Nevertheless, as discussed, the prediction model in this study shows strong potential for clinical application in CDSSs and early interventions. Accessibility and user experience can be enhanced by implementing a user-centered CDSS or web-based application according to the proposed model. Lastly, although we extensively validated the framework on various data sets and subgroups through retrospective analysis, achieving clinical maturity will require real-time prospective studies. Therefore, further research is necessary to ensure clinical maturity through prospective clinical studies in critical care environments.

Conclusions

In this study, we evaluated the performance of an interpretable AI alert model across different ICU populations in a simulated real-time clinical setting. As these results were tested in real-time clinical scenarios based on various databases for each patient group, we anticipate that similar outcomes will be observed in actual clinical trials. Consequently, the test results are expected to be valuable and applicable in real clinical settings. We then used the interpretable information from the proposed method to describe both global and time-specific relevance. This clinical descriptive information aids clinicians in decision-making by highlighting the associations between predicted outcomes and patient characteristics. Consequently, our CA prediction system is considered to have achieved clinical validity and is now being used and validated for routine clinical applications.

XSL•FO
RenderX

Acknowledgments

This research was partly supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant HI21C1074) and an Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT; grant RS-2021-II212068, Artificial Intelligence Innovation Hub).

Data Availability

MIMIC-IV is available online upon request at the repository PhysioNet [51]. The eICU-CRD is also freely available through PhysioNet to access [52]. The source code generated and used for this study is publicly available for download in Multimedia Appendix 1. The source code for Scikit-learn of Python is freely available at [53]. The source code for TabNet is freely available at [54].

Authors' Contributions

YKK and ML conceived and planned the study and wrote the manuscript. YKK, WS, SJL, and JHK analyzed the data. All the authors contributed to the interpretation of the results, provided critical feedback, and supported the research, analysis, and writing process involved in preparing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Source code information.
[DOCX File , 15 KB - jmir_v26i1e62890_app1.docx]

Multimedia Appendix 2
Details about the hyperparameters of the baseline models.
[DOCX File , 20 KB - jmir_v26i1e62890_app2.docx]

Multimedia Appendix 3
Demographic information of patients from subgroups of MIMIC-IV. MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 21 KB - jmir_v26i1e62890_app3.docx]

Multimedia Appendix 4
Demographic information of patients from subgroups of eICU-CRD. eICU-CRD: eICU-Collaborative Research Database.
[DOCX File , 21 KB - jmir_v26i1e62890_app4.docx]

Multimedia Appendix 5
Statistical comparison of the overall area under the receiver operating characteristic curve between the proposed method and baseline methods on the MIMIC-IV data set. MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 21 KB - jmir_v26i1e62890_app5.docx]

Multimedia Appendix 6
Statistical comparison of the overall event recall between the proposed method and baseline methods on the MIMIC-IV. MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 22 KB - jmir_v26i1e62890_app6.docx]

Multimedia Appendix 7
Statistical comparison of the overall sensitivity between the proposed method and baseline methods on the MIMIC-IV. MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 25 KB - jmir_v26i1e62890_app7.docx]

Multimedia Appendix 8

XSL•FO
RenderX

Statistical comparison of the overall false alarm rate between the proposed method and baseline methods on the MIMIC-IV. MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 25 KB - jmir_v26i1e62890_app8.docx]

Multimedia Appendix 9
Statistical comparison of the overall area under the receiver operating characteristic curve between the proposed method and baseline methods on the eICU-CRD. eICU-CRD: eICU-Collaborative Research Database.
[DOCX File , 25 KB - jmir_v26i1e62890_app9.docx]

Multimedia Appendix 10
Statistical comparison of the overall event recall between the proposed method and baseline methods on the eICU-CRD. eICU-CRD: eICU-Collaborative Research Database.
[DOCX File , 25 KB - jmir_v26i1e62890_app10.docx]

Multimedia Appendix 11
Statistical comparison of the overall sensitivity between the proposed method and baseline methods on the eICU-CRD. eICU-CRD: eICU-Collaborative Research Database.
[DOCX File , 25 KB - jmir_v26i1e62890_app11.docx]

Multimedia Appendix 12
Statistical comparison of the overall area under the receiver operating characteristic curve between the proposed method and baseline methods on the MIMIC-IV in cardiac ICU. ICU: intensive care unit; MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 20 KB - jmir_v26i1e62890_app12.docx]

Multimedia Appendix 13
Statistical comparison of the overall area under the receiver operating characteristic curve between the proposed method and baseline methods on the MIMIC-IV in general ICU. ICU: intensive care unit; MIMIC: Medical Information Mart for Intensive Care.
[DOCX File , 20 KB - jmir_v26i1e62890_app13.docx]

Multimedia Appendix 14
Statistical comparison of the overall area under the receiver operating characteristic curve between the proposed method and baseline methods on the eICU-CRD in cardiac ICU. eICU-CRD: eICU-Collaborative Research Database; ICU: intensive care unit.
[DOCX File , 20 KB - jmir_v26i1e62890_app14.docx]

Multimedia Appendix 15
Statistical comparison between Gini index–based feature of SpO_2 and the top 25 global importance features in the proposed method. SpO_2: oxyhemoglobin saturation.
[DOCX File , 28 KB - jmir_v26i1e62890_app15.docx]

References

1. Nolan JP, Berg RA, Andersen LW, Bhanji F, Chan PS, Donnino MW, et al. Cardiac arrest and cardiopulmonary resuscitation outcome reports: update of the Utstein Resuscitation Registry Template for in-hospital cardiac arrest: a consensus report from a Task Force of the International Liaison Committee on Resuscitation (American Heart Association, European Resuscitation Council, Australian and New Zealand Council on Resuscitation, Heart and Stroke Foundation of Canada, InterAmerican Heart Foundation, Resuscitation Council of Southern Africa, Resuscitation Council of Asia). Circulation 2019 Oct 29;140(18):1286-1300. [doi: 10.1161/cir.0000000000000710]
2. Andersen LW, Kim WY, Chase M, Berg KM, Mortensen SJ, Moskowitz A, American Heart Association's Get With the Guidelines(®) – Resuscitation Investigators. The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest. Resuscitation 2016 Jan;98:112-117 [FREE Full text] [doi: 10.1016/j.resuscitation.2015.08.016] [Medline: 26362486]
3. Bergum D, Haugen BO, Nordseth T, Mjølstad OC, Skogvoll E. Recognizing the causes of in-hospital cardiac arrest--a survival benefit. Resuscitation 2015 Dec;97:91-96 [FREE Full text] [doi: 10.1016/j.resuscitation.2015.09.395] [Medline: 26449872]

XSL•FO
RenderX

4. Guidi G, Pettenati MC, Melillo P, Iadanza E. A machine learning system to improve heart failure patient assistance. IEEE J Biomed Health Inform 2014 Nov;18(6):1750-1756. [doi: 10.1109/jbhi.2014.2337752]
5. Jeong J, Cho J, Lee B, Lee S, Jeong J. Real-time deep neurolinguistic learning enhances noninvasive neural language decoding for brain–machine interaction. IEEE Trans Cybern 2023 Dec;53(12):7469-7482. [doi: 10.1109/tcyb.2022.3211694]
6. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3(1):17-17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
7. Zhou C, Li A, Hou A, Zhang Z, Zhang Z, Dai P, et al. Modeling methodology for early warning of chronic heart failure based on real medical big data. Expert Syst Appl 2020 Aug;151:113361. [doi: 10.1016/j.eswa.2020.113361]
8. Jain A, Chandra Sekhara Rao A, Kumar Jain P, Hu Y. Optimized levy flight model for heart disease prediction using CNN framework in big data application. Expert Syst Appl 2023 Aug;223:119859. [doi: 10.1016/j.eswa.2023.119859]
9. Alsalem M, Alamoodi A, Albahri O, Albahri A, Martínez L, Yera R, et al. Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. Expert Syst Appl 2024 Jul;246:123066. [doi: 10.1016/j.eswa.2023.123066]
10. Sidek KA, Khalil I, Jelinek HF. ECG biometric with abnormal cardiac conditions in remote monitoring system. IEEE Trans Syst Man Cybern Syst 2014 Nov;44(11):1498-1509. [doi: 10.1109/tsmc.2014.2336842]
11. Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B. SAPS-II revisited. Intensive Care Med 2005 Jan 28;31(3):416-423. [doi: 10.1007/s00134-005-2557-9]
12. Spångfors M, Molt M, Samuelson K. In-hospital cardiac arrest and preceding National Early Warning Score (NEWS): a retrospective case-control study. Clin Med (Lond) 2020 Jan;20(1):55-60 [FREE Full text] [doi: 10.7861/clinmed.2019-0137] [Medline: 31941734]
13. Yijing L, Wenyu Y, Kang Y, Shengyu Z, Xianliang H, Xingliang J, et al. Prediction of cardiac arrest in critically ill patients based on bedside vital signs monitoring. Comput Methods Programs Biomed 2022 Feb;214:106568. [doi: 10.1016/j.cmpb.2021.106568] [Medline: 34883382]
14. Vincent J, Moreno R, Takala J, Willatts S, Mendonça AD, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996 Jul 1;22(7):707-710. [doi: 10.1007/s001340050156]
15. Subbe C, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM 2001 Oct;94(10):521-526. [doi: 10.1093/qjmed/94.10.521] [Medline: 11588210]
16. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI, Higgins B. A review, and performance evaluation, of single-parameter "track and trigger" systems. Resuscitation 2008 Oct;79(1):11-21. [doi: 10.1016/j.resuscitation.2008.05.004] [Medline: 18620794]
17. Kwon J, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in - hospital cardiac arrest. J Am Heart Assoc 2018 Jul 03;7(13):e008678. [doi: 10.1161/jaha.118.008678]
18. Kim YK, Koo JH, Lee SJ, Song HS, Lee M. Explainable artificial intelligence warning model using an ensemble approach for in-hospital cardiac arrest prediction: retrospective cohort study. J Med Internet Res 2023 Dec 22;25:e48244 [FREE Full text] [doi: 10.2196/48244] [Medline: 38133922]
19. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical Care Med 2016;44(2):368-374. [doi: 10.1097/ccm.0000000000001571]
20. Hong S, Lee S, Lee J, Cha WC, Kim K. Prediction of cardiac arrest in the emergency department based on machine learning and sequential characteristics: model development and retrospective clinical validation study. JMIR Med Inform 2020 Aug 04;8(8):e15932 [FREE Full text] [doi: 10.2196/15932] [Medline: 32749227]
21. Layeghian Javan S, Sepehri MM, Layeghian Javan M, Khatibi T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. Comput Methods Programs Biomed 2019 Sep;178:47-58. [doi: 10.1016/j.cmpb.2019.06.010] [Medline: 31416562]
22. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial Intelligence Applications in Health Care Practice: Scoping Review. J Med Internet Res 2022 Oct 5;24(10):e40238. [doi: 10.2196/40238]
23. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
24. Arik S, Pfister T. TabNet: attentive interpretable tabular learning. Washington, D.C: AAAI; 2021 May 18 Presented at: The 35th AAAI Conference on Artificial Intelligence; 2021-05-15; Virtual Conference p. 6679-6687. [doi: 10.1609/aaai.v35i8.16826]
25. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L, Mark R. MIMIC-IV (version 1). PhysioNet. URL: https://physionet.org/content/mimiciv/1.0/ [accessed 2023-08-23]
26. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5(1):180178-180113 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]

XSL•FO
RenderX

27. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 04;169(12):866. [doi: 10.7326/m18-1990]
28. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun ACM 2021 Nov 19;64(12):86-92. [doi: 10.1145/3458723]
29. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron L, et al. Transparency and reproducibility in artificial intelligence. Nature 2020 Oct 14;586(7829):E14-E16 [FREE Full text] [doi: 10.1038/s41586-020-2766-y] [Medline: 33057217]
30. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol 2024 Jan 04;42(1):3-15 [FREE Full text] [doi: 10.1007/s11604-023-01474-3] [Medline: 37540463]
31. Marmot M, Bell R. Fair society, healthy lives. Public Health 2012 Sep;126 Suppl 1:S4-S10. [doi: 10.1016/j.puhe.2012.05.014] [Medline: 22784581]
32. Anderson M, Anderson S. How should AI be developed, validated, and implemented in patient care? AMA J Ethics 2019 Feb 01;21(2):E125-E130 [FREE Full text] [doi: 10.1001/amajethics.2019.125] [Medline: 30794121]
33. Walsh C, Chaudhry B, Dua P, Goodman K, Kaplan B, Kavuluru R, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open 2020 Apr;3(1):9-15 [FREE Full text] [doi: 10.1093/jamiaopen/ooz054] [Medline: 32607482]
34. GitHub. URL: https://github.com/yunkwankim/SCAP [accessed 2024-08-29]
35. Barnard J, Meng X. Applications of multiple imputation in medical studies: from AIDS to NHANES. Stat Methods Med Res 1999 Mar 02;8(1):17-36. [doi: 10.1177/096228029900800103] [Medline: 10347858]
36. Kenward MG, Molenberghs G. Last observation carried forward: a crystal ball? J Biopharm Stat 2009 Sep 07;19(5):872-888. [doi: 10.1080/10543400903105406] [Medline: 20183449]
37. Goodwin T, Demner-Fushman D. A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. J Am Med Inform Assoc 2020 Apr 01;27(4):567-576 [FREE Full text] [doi: 10.1093/jamia/ocaa004] [Medline: 32065628]
38. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 2018 Sep 17;19(Suppl 1):65 [FREE Full text] [doi: 10.1186/s12863-018-0633-8] [Medline: 30255764]
39. Manikandan G, Pragadeesh B, Manojkumar V, Karthikeyan A, Manikandan R, Gandomi AH. Classification models combined with Boruta feature selection for heart disease prediction. Inform Med Unlocked 2024;44:101442. [doi: 10.1016/j.imu.2023.101442]
40. Kotsiantis S, Pintelas P. Mixture of expert agents for handling imbalanced data sets. Ann Math Comput Teleinform 2003;1(1):46-55 [FREE Full text]
41. Li F, Zhang X, Zhang X, Du C, Xu Y, Tian Y. Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. Inform Sci 2018 Jan;422:242-256. [doi: 10.1016/j.ins.2017.09.013]
42. Naseem U, Rashid J, Ali L, Kim J, Haq QEU, Awan MJ, et al. An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers. IEEE Access 2022;10:78242-78252. [doi: 10.1109/access.2022.3174599]
43. Gholamiangonabadi D, Kiselov N, Grolinger K. Deep neural networks for human activity recognition with wearable sensors: leave-one-subject-out cross-validation for model selection. IEEE Access 2020;8:133982-133994. [doi: 10.1109/access.2020.3010715]
44. Pauli M, Pohl C, Golz M. Balanced leave-one-subject-out cross-validation for microsleep classification. Curr Dir Biomed Eng 2021 Oct;7(2):147-150. [doi: 10.1515/cdbme-2021-2038]
45. Newman-Griffis D, Divita G, Desmet B, Zirikly A, Rosé CP, Fosler-Lussier E. Ambiguity in medical concept normalization: an analysis of types and coverage in electronic health record datasets. J Am Med Inform Assoc 2021 Mar 01;28(3):516-532 [FREE Full text] [doi: 10.1093/jamia/ocaa269] [Medline: 33319905]
46. Graham R, Margaret AM, Andrea MS. Strategies to Improve Cardiac Arrest Survival: A Time to Act. Washington, DC: The National Academies Press; 2015.
47. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020 Mar;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
48. Lowry A, Futterman C, Gazit A. Acute vital signs changes are underrepresented by a conventional electronic health record when compared with automatically acquired data in a single-center tertiary pediatric cardiac intensive care unit. J Am Med Inform Assoc 2022 Jun 14;29(7):1183-1190 [FREE Full text] [doi: 10.1093/jamia/ocac033] [Medline: 35301538]
49. Papadimitriou C. Computational complexity. Ency Comput Sci 2003 Jan:260-265 [FREE Full text]
50. Rossetti S, Knaplund C, Albers D, Dykes PC, Kang MJ, Korach TZ, et al. Healthcare process modeling to phenotype clinician behaviors for exploiting the signal gain of clinical expertise (HPM-ExpertSignals): development and evaluation of a conceptual framework. J Am Med Inform Assoc 2021 Jun 12;28(6):1242-1251 [FREE Full text] [doi: 10.1093/jamia/ocab006] [Medline: 33624765]
51. PhysioNet. URL: https://physionet.org/content/mimiciv/2.2/ [accessed 2024-08-29]

XSL•FO
RenderX

52. eICU Collaborative Research Database - MIT. URL: https://eicu-crd.mit.edu/ [accessed 2024-08-29]
53. Scikit-learn. URL: https://scikit-learn.org/stable/ [accessed 2024-08-29]
54. GitHub. URL: https://github.com/topics/pytorch-tabnet [accessed 2024-08-29]

Abbreviations

AI: artificial intelligence
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
BIDMC: Beth Israel Deaconess Medical Center
CA: cardiac arrest
CDSS: clinical decision support system
DBP: diastolic blood pressure
eICU-CRD: eICU-Collaborative Research Database
EWS: early warning score
FDR: false discovery rate
HF: heart failure
HR: heart rate
HSD: honestly significant difference
ICU: intensive care unit
LGBM: light gradient boosting method
LOCB: last observation carried backward
LOCF: last observation carried forward
LOPO: leave-one-patient-out
MEWS: Modified Early Warning Score
MIMIC: Medical Information Mart for Intensive Care
NEWS: National Early Warning Score
RF: random forest
RNN: recurrent neural network
RR: respiratory rate
SAPS: Simplified Acute Physiology Score
SBP: systolic blood pressure
SOFA: Sequential Organ Failure Assessment
SpO_2: oxyhemoglobin saturation
TabNet: tabular network

Edited by G Eysenbach, A Coristine, A Mavragani; submitted 04.06.24; peer-reviewed by U Sinha, M Wu, IW Pulantara; comments to author 09.07.24; revised version received 30.07.24; accepted 21.08.24; published 17.09.24.

Please cite as:
Kim YK, Seo WD, Lee SJ, Koo JH, Kim GC, Song HS, Lee M
Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study
J Med Internet Res 2024;26:e62890
URL: https://www.jmir.org/2024/1/e62890
doi:10.2196/62890
PMID:

©Yun Kwan Kim, Won-Doo Seo, Sun Jung Lee, Ja Hyung Koo, Gyung Chul Kim, Hee Seok Song, Minji Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating a Natural Language Processing–Driven, AI-Assisted International Classification of Diseases, 10th Revision, Clinical Modification, Coding System for Diagnosis Related Groups in a Real Hospital Environment: Algorithm Development and Validation Study

Hong-Jie Dai[1,2,3], PhD; Chen-Kai Wang[1,4,5], MSc; Chien-Chang Chen[6], PhD; Chong-Sin Liou[1], MSc; An-Tai Lu[7], MS; Chia-Hsin Lai[1], BS; Bo-Tsz Shain[1], BS; Cheng-Rong Ke[1], BS; William Yu Chung Wang[8], PhD; Tatheer Hussain Mir[1], MS; Mutiara Simanjuntak[1], MS; Hao-Yun Kao[9*], PhD; Ming-Ju Tsai[10*], MD, PhD; Vincent S Tseng[4*], PhD

[1]Intelligent System Lab, College of Electrical Engineering and Computer Science, Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[2]National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan

[3]Center for Big Data Research, Kaohsiung Medical University, Kaohsiung, Taiwan

[4]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[5]Advanced Technology Laboratory, Chunghwa Telecom Laboratories, Taoyuan, Taiwan

[6]Electromagnetic Sensing Control and AI Computing System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[7]School of Post-Baccalaureate Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[8]Waikato Management School, University of Waikato, Hamilton, New Zealand

[9]Department of Healthcare Administration and Medical Informatics, College of Health Sciences, Kaohsiung Medical University, Kaohsiung, Taiwan

[10]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[*]these authors contributed equally

Corresponding Author:
Ming-Ju Tsai, MD, PhD
Division of Pulmonary and Critical Care Medicine
Department of Internal Medicine
Kaohsiung Medical University Hospital, Kaohsiung Medical University
No 100, Tzyou 1st Road
Sanmin District
Kaohsiung, 80756
Taiwan
Phone: 886 73121101 ext 4660035
Email: mjt@kmu.edu.tw

Abstract

Background: International Classification of Diseases codes are widely used to describe diagnosis information, but manual coding relies heavily on human interpretation, which can be expensive, time consuming, and prone to errors. With the transition from the *International Classification of Diseases, Ninth Revision*, to the *International Classification of Diseases, Tenth Revision* (*ICD-10*), the coding process has become more complex, highlighting the need for automated approaches to enhance coding efficiency and accuracy. Inaccurate coding can result in substantial financial losses for hospitals, and a precise assessment of outcomes generated by a natural language processing (NLP)–driven autocoding system thus assumes a critical role in safeguarding the accuracy of the Taiwan diagnosis related groups (Tw-DRGs).

Objective: This study aims to evaluate the feasibility of applying an *International Classification of Diseases, Tenth Revision, Clinical Modification* (*ICD-10-CM*), autocoding system that can automatically determine diagnoses and codes based on free-text

XSL•FO
RenderX

discharge summaries to facilitate the assessment of Tw-DRGs, specifically principal diagnosis and major diagnostic categories (MDCs).

Methods: By using the patient discharge summaries from Kaohsiung Medical University Chung-Ho Memorial Hospital (KMUCHH) from April 2019 to December 2020 as a reference data set we developed artificial intelligence (AI)–assisted *ICD-10-CM* coding systems based on deep learning models. We constructed a web-based user interface for the AI-assisted coding system and deployed the system to the workflow of the certified coding specialists (CCSs) of KMUCHH. The data used for the assessment of Tw-DRGs were manually curated by a CCS with the principal diagnosis and MDC was determined from discharge summaries collected at KMUCHH from February 2023 to April 2023.

Results: Both the reference data set and real hospital data were used to assess performance in determining *ICD-10-CM* coding, principal diagnosis, and MDC for Tw-DRGs. Among all methods, the GPT-2 (OpenAI)-based model achieved the highest F_1-score, 0.667 (F_1-score 0.851 for the top 50 codes), on the KMUCHH test set and a slightly lower F_1-score, 0.621, in real hospital data. Cohen κ evaluation for the agreement of MDC between the models and the CCS revealed that the overall average κ value for GPT-2 (κ=0.714) was approximately 12.2 percentage points higher than that of the hierarchy attention network (κ=0.592). GPT-2 demonstrated superior agreement with the CCS across 6 categories of MDC, with an average κ value of approximately 0.869 (SD 0.033), underscoring the effectiveness of the developed AI-assisted coding system in supporting the work of CCSs.

Conclusions: An NLP-driven AI-assisted coding system can assist CCSs in *ICD-10-CM* coding by offering coding references via a user interface, demonstrating the potential to reduce the manual workload and expedite Tw-DRG assessment. Consistency in performance affirmed the effectiveness of the system in supporting CCSs in *ICD-10-CM* coding and the judgment of Tw-DRGs.

(J Med Internet Res 2024;26:e58278) doi:10.2196/58278

KEYWORDS

natural language processing; International Classification of Diseases; deep learning; electronic medical record; Taiwan diagnosis related groups

Introduction

Background

The International Classification of Diseases (ICD) [1], established by the World Health Organization, is a crucial medical classification system that defines the universe of diseases, disorders, injuries, and other related health conditions. Since its first publication in 1893, the ICD has become a widely adopted standard across various health care facilities and settings globally, providing consistency and accuracy in disease diagnosis and classification. In 1992, the World Health Organization published the *International Classification of Diseases, Tenth Revision* (*ICD-10*), which has since been widely adopted worldwide [2,3]. Many countries extended and customized the *ICD-10* classification system for their country-specific reporting purposes, such as *International Classification of Diseases, Tenth Revision, Clinical Modification* (*ICD-10-CM*) in the United States, *International Classification of Diseases, Tenth Revision, Canadian Modification* in Canada, *International Classification of Diseases, Tenth Revision, German Modification* in Germany, and *International Classification of Diseases, Tenth Revision, Australian Modification* that is followed by Australia as well as 15 other countries including Ireland, Singapore, and Saudi Arabia.

The *ICD-10-CM* is an ICD system that classifies patients according to the type of illness, severity, and the location of the disease that was developed to describe more clinical details with the increasing number of diagnoses and procedural codes applied in payment methodologies [4,5]. As a result, the *ICD-10-CM* coding task has become a crucial element in various fields, such as disease surveillance [6], health services management [7], and clinical research [8]. For the National Health Insurance Administration (NHIA) in Taiwan, *ICD-10-CM* holds immense significance by serving as a standardized coding system for the statistical analysis of disease diagnosis, surgical treatment for patients admitted to hospital, and payment of health insurance. In 2016, the NHIA in Taiwan followed the global trend and transitioned from the International *Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*), to *ICD-10-CM*, which expanded the number of codes available. Previously, the *ICD-9-CM* contained approximately 14,000 diagnosis codes, while the *ICD-10-CM* had approximately 69,000 diagnosis codes [5]. The NHIA is currently using the 2014 version of *ICD-10-CM*, which consists of approximately 71,900 diagnosis codes. Nowadays, many medical institutions are relying on licensed certified coding specialists (CCSs) to manually assign *ICD-10* codes to inpatients. These coders spend a significant amount of time reviewing various medical materials to accurately diagnose and code each patient's condition. Due to the complexity of the ICD-10 structure and coding rules, the task of coding is significantly more labor-intensive and time-consuming than *ICD-9*, even when performed by a skilled CCS who typically dedicates approximately 30 minutes per case on average. In response to these challenges, investigators [9-11] have applied both rule-based algorithms and machine learning methods, such as recurrent neural network, long short-term memory, and bidirectional encoder representations from transformers (BERT), to classify patients with specific conditions. However, the performance of these approaches remains limited.

By contrast, to control the rising health care costs, health authorities in many countries have implemented the diagnosis related group (DRG) payment system. The NHIA in Taiwan has implemented the Taiwan DRG (Tw-DRG) payment systems since 2010 [12] to consolidate related DRGs for the purpose of

XSL•FO
RenderX

determining payment standards [13]. Tw-DRGs are classified by major diagnostic categories (MDCs) that depend on the assigned *ICD-10-CM* codes. Consequently, accurate *ICD-10-CM* coding is critical for the accurate generation of Tw-DRGs, as coding errors can lead to inappropriate treatment options, delayed reimbursement processes, and significant financial losses for hospitals [14,15]. Factors including incomplete information, incorrect data entry, and insufficient coder expertise can lead to inaccurate coding [16]. In addition, errors may also arise from incorrect human perception [17], the complex technical nature of the work [18], and human fatigue from heavy workloads [19]. While natural language processing (NLP)–driven autocoding systems have the potential to enhance the quality of the manual coding results and expedite code assignment, it is imperative to assess their accuracy to ensure cost savings for the hospitals [20]. In summary, precise coding according to the *ICD-10-CM* system is essential for accurately categorizing Tw-DRGs, as any fault in this process can lead to misclassification and subsequently impact health care reimbursements.

Study Overview

In this study, we developed 2 deep learning–based models, the hierarchical attention network (HAN) and the GPT-2 (OpenAI), in the manner of multi-label supervised learning for *ICD-10-CM* coding. The former is a conventional classification model, and the latter is categorized as a generative model.

To facilitate the coding process in the real hospital environment, we established an NLP-driven *ICD-10-CM* autocoding system. This system includes a user-friendly visual interface to display the predicted coding results, CIs, and relevant medical record keywords. We integrated the system into the coding procedure protocolled at Kaohsiung Medical University Chung-Ho Memorial Hospital (KMUCHH) to assist the workflow of clinical coders to expedite the efficiency of disease coding. We then compared the consistency of the principal diagnosis and MDC coding between the autocoding system and the data curated by the CCS.

Methods

Data Set and Task Definition

Data Collection and Preparation

This study used a total of 136,841 unstructured discharge summaries of patients who were hospitalized, recorded in the KMUCHH from April 1, 2019, to December 31, 2020, as the primary data source for the development of the *ICD-10-CM* coding system. After deploying the developed system in the workflow of the KMUCHH CSS, we collected an additional data set containing 2632 discharge cases processed by the system from February 2023, to April 2023, to assess the performance of the *ICD-10-CM* coding system in the real hospital environment and the feasibility of assisting the process of Tw-DRGs. The original data set contains 15,756 discharge cases. After excluding 7541 (47.86%) non–Tw-DRGs cases and 5583 (35.43%) with incomplete electronic medical records (EMRs), a data set of 2632 (16.7%) discharge cases was used for our evaluation.

To maximize the amount of data available for the training phase of our system, we composed a test set by selecting the latest 1000 discharge summaries from the raw data sorted by time stamp. The remaining summaries were then allocated for use as the training set. During the training phase, we randomly sampled 5% (6842/136,841) of the discharge summaries to form the validation set consisting of 6842 summaries. As depicted in Table 1, a total of 129,000 (94.27%) of 136,841 discharge summaries were assigned to the training set (comprising a total of 567,957 *ICD-10-CM* codes, with 11,494 (2.02%) being unique), 6842 (5%) discharge summaries were assigned to the validation set (with a total of 36,205 codes, including 4038, 11.15% unique ones), and 1000 (0.73%) discharge summaries were included in the test set with a total of 10,412 codes and 1482 (14.23%) of them being unique.

XSL•FO
RenderX

Table 1. Prevalence of the 2014 International Classification of Diseases, Tenth Revision, Clinical Modification codes used in the compiled data set consisting of 136,841 discharge summaries.

Chapters	Blocks	Definitions	Codes in training set (n=567,957), n (%)	Codes in validation set (n=36,205), n (%)	Codes in test set (n=10,412), n (%)
I	A00-B99	Certain infectious and parasitic diseases	35,708 (6.29)	2233 (6.17)	341 (3.28)
II	C00-D48	Neoplasms	82,248 (14.48)	4838 (13.36)	827 (7.94)
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	17,347 (3.05)	1281 (3.54)	168 (1.61)
IV	E00-E90	Endocrine, nutritional, and metabolic diseases	63,545 (11.19)	4077 (11.26)	301 (2.89)
V	F00-F99	Mental and behavioral disorders	8835 (1.56)	565 (1.56)	289 (2.78)
VI	G00-G99	Diseases of the nervous system	13,236 (2.33)	835 (2.31)	395 (3.79)
VII	H00-H59	Diseases of the eye and adnexa	8520 (1.5)	421 (1.16)	661 (6.35)
VIII	H60-H95	Diseases of the ear and mastoid process	1304 (0.23)	120 (0.33)	193 (1.85)
IX	I00-I99	Diseases of the circulatory system	82,005 (14.44)	5066 (13.99)	636 (6.11)
X	J00-J99	Diseases of the respiratory system	31,233 (5.5)	2091 (5.78)	285 (2.74)
XI	K00-K93	Diseases of the digestive system	48,827 (8.6)	3098 (8.56)	514 (4.94)
XII	L00-L99	Diseases of the skin and subcutaneous tissue	7469 (1.32)	488 (1.35)	372 (3.57)
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue	20,312 (3.58)	1296 (3.58)	1191 (11.44)
XIV	N00-N99	Diseases of the genitourinary system	39,504 (6.96)	2498 (6.9)	328 (3.15)
XV	O00-O99	Pregnancy, childbirth, and the puerperium	3887 (0.68)	255 (0.70)	394 (3.78)
XVI	P00-P96	Certain conditions originating in the perinatal period	4297 (0.76)	271 (0.75)	202 (1.94)
XVII	Q00-Q99	Congenital malformations, deformations, and chromosomal abnormalities	2918 (0.51)	171 (0.47)	315 (3.03)
XVIII	R00-R99	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	24,544 (4.32)	1636 (4.52)	392 (3.76)
XIX	S00-T98	Injury, poisoning, and certain other consequences of external causes	20,850 (3.67)	1783 (4.92)	1661 (15.95)
XX	V01-Y98	External causes of morbidity and mortality	11,791 (2.08)	711 (1.96)	530 (5.09)
XXI	Z00-Z99	Factors influencing health status and contact with health services	39,577 (6.97)	2471 (6.83)	417 (4)

ICD-10-CM Coding Task Formulation

The *ICD-10-CM* coding task presented in this study aims to develop an NLP autocoding system for generating *ICD-10-CM* codes from a patient's discharge summary. An *ICD-10-CM* code consists of 3 to 7 characters, and each code begins with an alphabetic character that signifies the relevant classification chapter. The first 3 characters in the code designate the category of the diagnosis, while the subsequent 3 characters correspond to the related etiology. The seventh character provides the related extensions. As shown in Table 1, there are 21 chapters in *ICD-10-CM*. The ground-truth *ICD-10-CM* codes for the compiled 13,6841 discharge summaries were annotated by specialized CCSs at KMUCHH, culminating in a grand total of 11,653 unique codes. Textbox 1 displays an example of the discharge summary from the KMUCHH. The output of the coding system for each summary encompasses a main code along with multiple other codes, thereby formulating the task as a multi-labeling classification problem. Within our data set, the unique count for main codes and other codes is 5835 and 10,393, respectively. Moreover, the generated main codes will serve as the principal diagnosis, while the other codes will be used as secondary diagnoses in the Tw-DRGs estimation process.

XSL•FO
RenderX

Textbox 1. An example of the discharge summary from Kaohsiung Medical University Chung-Ho Memorial Hospital.

Chief complaint:

Acute urine retention after discharge today

Impression on Admission:

Acute urine retention.

Bladder stone s/p op.

Hypertension.

Discharge Diagnosis:

--underlying--

#Acute urine retention.

#Hypertension.

#Hyperlipidemia

History on Admission:

This 71 y/o male has history of hypertension and under regular medical control. He was just discharged from our ward bladder stone and accepted surgical intervention (Endoscopic cystolithotripsy) in last admission. This time, he ...

Data Preprocessing

During the preprocessing stage, delimiters (eg, "--underlying--" shown in Textbox 1) and prefix symbols (such as "#") in the unstructured discharge summaries were filtered out. Subsequently, the clinical NLP toolkit was leveraged to preprocess the summaries by applying sentence splitting and tokenization [21].

Deep Learning Models

As an effective approach to the coding task, we formulated it as a multi-label supervised learning problem and applied 2 deep learning–based methods to the compiled data set. The first model is grounded in HAN, a neural network architecture specifically crafted to tackle the complexities of modeling hierarchical structures within text data [22]. HAN leverages attention mechanisms to capture fine-grained information at both the document and sentence levels. The second model is a generative network built upon GPT-2; a causal language model released by OpenAI which was pretrained on large-scale text data [23]. The architecture is also featured with an attention mechanism that enables it to comprehend and generate natural language. HAN was selected for its capacity to provide high readability and interpretability of text through visualization [24]. It leverages hierarchical text representation and attention mechanisms to effectively highlight important words and sentences at multiple levels of the text. This enhances the interpretability of the model to allow a better understanding of its internal workings and gain insights into the model's decision-making process with a strengthened trust in its outputs [25]. Furthermore, considering the growing significance of generative artificial intelligence (AI) in research and landing applications, large language models such as GPT-2 are poised to bring about significant transformation in clinical medicine and health care and will be ubiquitous in these fields. However, concerns regarding data sensitivity, inference speed, hardware requirements, as well as the ease of deployment and difficulty of maintenance of the system are paramount [26]. In light of these considerations, we opted for the GPT-2 model due to its balance of performance and practicality in our real clinical environment setting.

Our HAN implementation adheres to the original network architecture and is tailored specifically for the multi-labeling classification task of *ICD-10-CM* codes. The customization involves the application of a fully connected layer to transform the attention results of HAN into the desired number of target ICD codes as illustrated in Figure 1. The process begins in the embedding layer, where we use pretrained global vector embeddings containing 300D vectors trained on a corpus of 6 billion tokens [27] to extract essential information from the textual data. Following the embedding layer, we execute word-level and sentence-level encoding procedures, which use an attention mechanism to capture crucial words and sentences within the text. Subsequently, we implement a fully connected layer to generate a set of 11,653 unique *ICD-10-CM* codes. Finally, the output values are transformed into a range between 0 and 1 using a sigmoid layer, which represents the probability of the ICD code being related to the summary. The loss function used for the HAN model is the cross-entropy loss defined in Equation (1), which measures the sum of the negative log-likelihood of the probabilities of the actual labels. A large deviation between and the actual label cause a greater value in the loss and thus is penalized more during training.

XSL•FO
RenderX

Figure 1. Architecture of the hierarchical attention network implementation in this study. ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification.

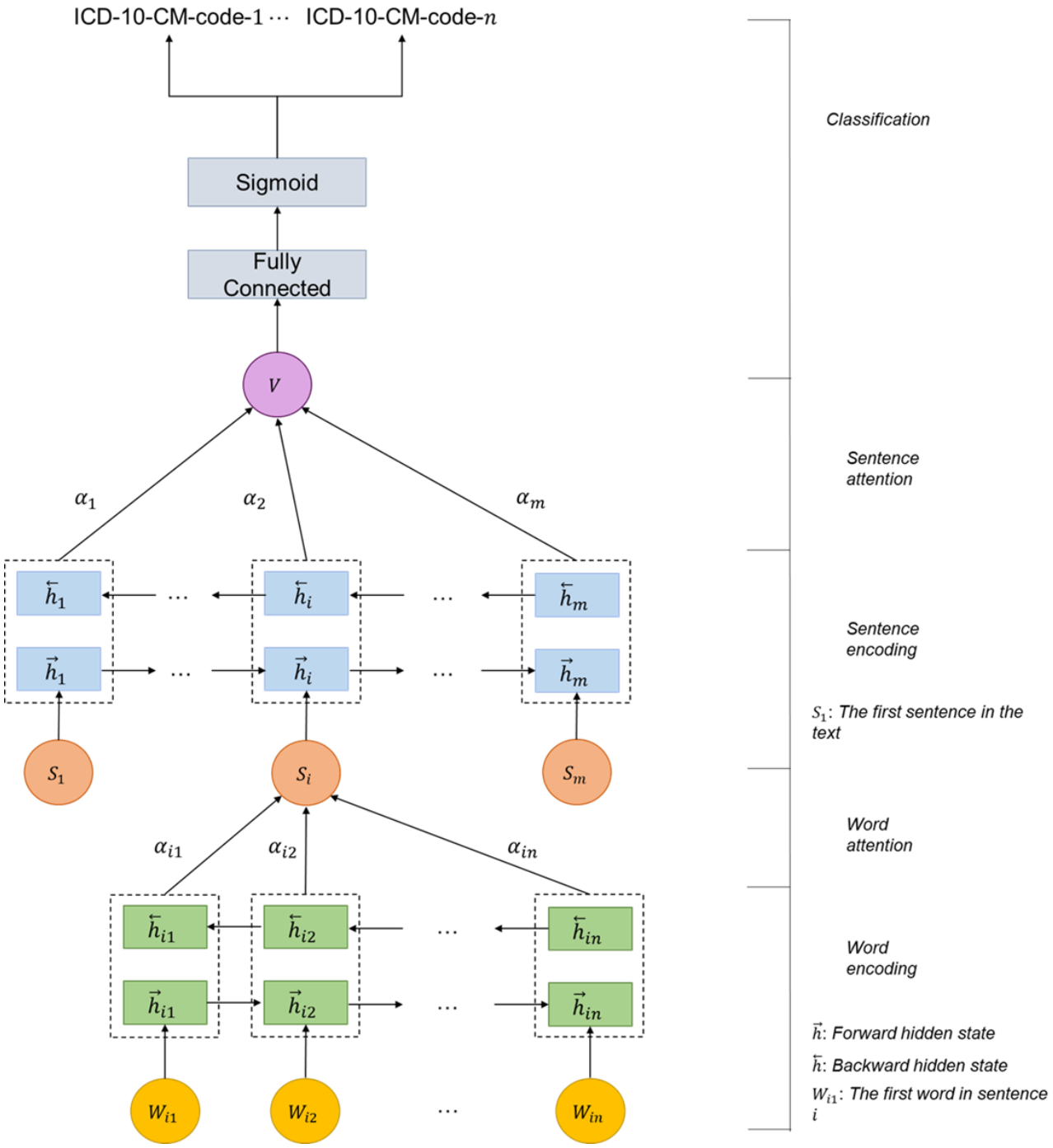

For inference, a threshold is set to assign the main code and other codes to the document based on the estimated label probability set following equations 2 and 3.

argmax(p) **(2)**

As shown in the aforementioned equations, the threshold *t* is applied to determine the predicted probability for *ICD-10-CM* codes, with the code having the highest probability selected as the *ICD-10-CM* main code. The main code is subsequently used as the principal diagnosis in the Tw-DRG process. After conducting multiple experiments, we set the threshold *t* in equation (3) as 0.5 as it has been determined that this setting achieves a better balance between the precision and recall.

In our GPT-2 implementation, we used a pretrained model developed by Papanikolaou and Pierleoni [28], which was a model based on the GPT-2 architecture fine-tuned using 0.5 million PubMed abstracts with 355 million parameters. This model was adopted instead of the original GPT-2 released by OpenAI to address the issue of the variety of medical

XSL•FO
RenderX

vocabularies conveyed in the summaries, which include diverse clinical data such as symptoms, diseases, and medications written in varying styles. To further fine-tune the pretrained model on our data set, we added special tokens to the model's tokenizer to help the model understand the structure of the sequence to complete the coding tasks. A total of 3 special tokens including "CLS," "SEP," and "#@#" were defined, which represent the starting position symbol of the input, the ending position symbol of the sentence, and the separator between *ICD-10-CM* main code (MAINCODE) and other codes (OTHERCODE), respectively. An example of a fine-tuned instance after processing is shown in Figure 2. The coding task is then formulated in a generative manner that is given an input text (*eg*, the prompt part shown in Figure 2) with a sequence of n tokens = $[x_1,x_2,...,x_n]$, and the target output sequence y = $[y_1,y_2,...,y_m]$. The objective is to maximize the conditional probability in the auto-regressive formulation represented in equation (4).

Figure 2. An example of fine-tuned data for GPT-2.

Prompt:
DischgDiag:#Acuteurine retention.<NL>#Hypertension.<NL>#Hyperlipidemia"

Completion:
MAINCODE: R33.9#@#OTHERCODE: I10, E78.5

During inference, the probability of each generated code is estimated by averaging the accumulated conditional probabilities for the tokens of the generated code. In instances where discharge summaries exceed the 1024 token limit of the GPT-2 model, we truncate the text during preprocessing to ensure it stays within the specified limit.

Finally, in addition to HAN and GPT-2, we implemented 2 baselines for performance comparison, one is a BERT-based model proposed by Devlin et al [29] and the other is based on the bidirectional gated recurrent unit along with the BERT-based word representation architecture proposed by Chen et al [30].

Integration of the Developed NLP-Driven AI-Assisted Coding System at KMUCHH

To integrate the developed models within the workflow of the CSS at KMUCHH, a user interface as illustrated in Figure 3 was implemented [31]. Various components can be identified within this interface. Area (1) serves as the patient EMRs display area. It conveniently showcases the relevant unstructured text content from the EMRs with 14 selectable sections. Area (2) features 3 functional buttons from left to right. The first button is used for attention visualization as presented in Figure 4. To display this figure, we analyze the results of the attention layer of the deployed model to highlight key medical terms from discharge summaries.

The second button initiates the *ICD-10-CM* automatic coding, which exploits the developed model to generate *ICD-10-CM* codes based on the content shown in area (1). The third button is the save button, which is used to store the results of *ICD-10-CM* autocoding, user selections, and system operation time stamps in the database. The search bar is located in area (3) and can be used by the CCS to manually look up *ICD-10-CM* codes or keywords in English to assist in manually adding missing codes. Any newly added codes are combined with the selected codes and displayed in area (4). Finally, area (5) provides a list of *ICD-10-CM* codes recommended by the deployed model through the coding assistant process. These results are presented in a checklist format as illustrated in Figure 3, including *ICD-10-CM* codes, their corresponding Chinese and English descriptions, and the confidence assigned by the deep learning model.

XSL•FO
RenderX

Figure 3. User interface of a natural language processing–driven artificial intelligence–assisted coding system at Kaohsiung Medical University Chung-Ho Memorial Hospital.

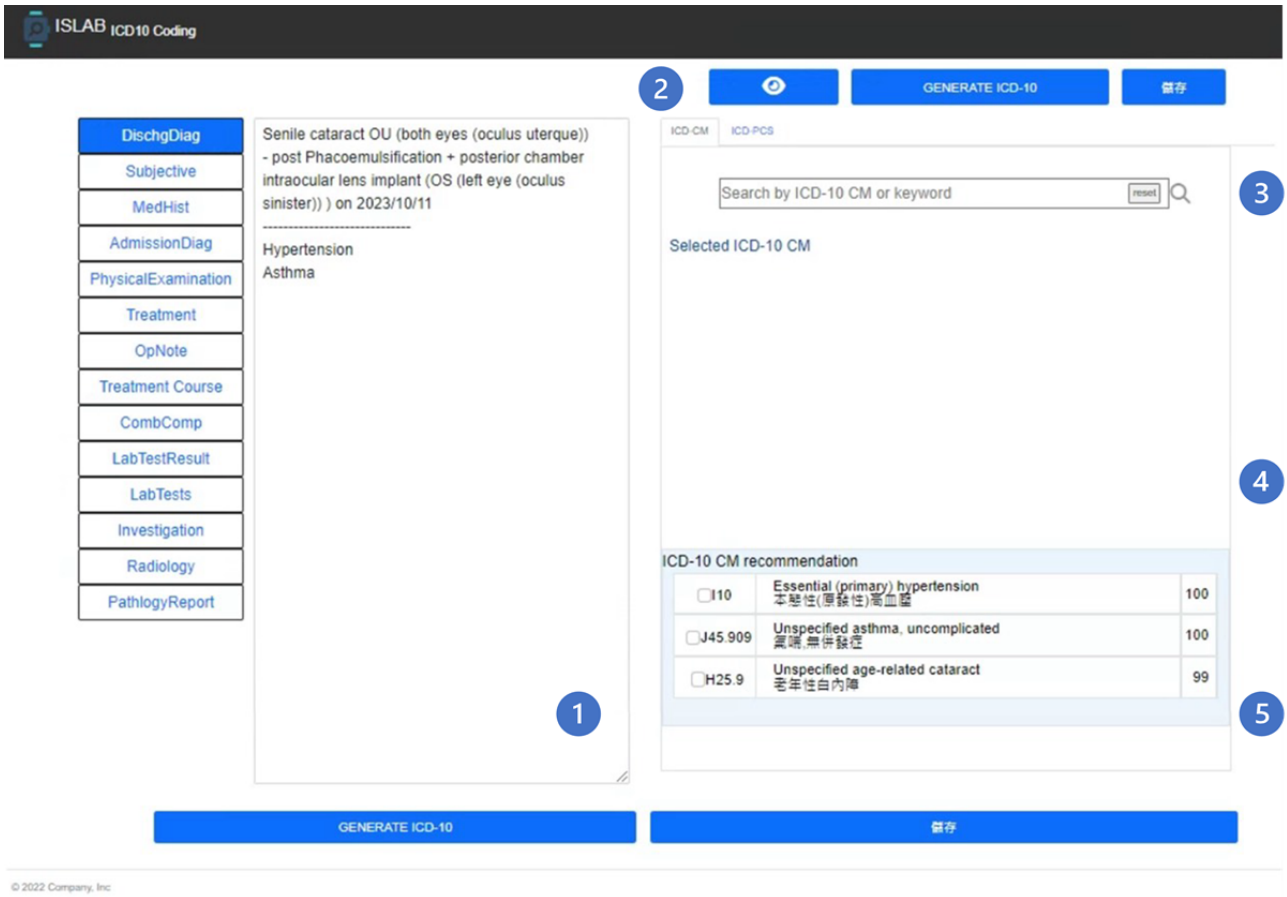

Figure 4. Visualization of the attentive key terms highlighted in a given discharge summary by the attention mechanism.

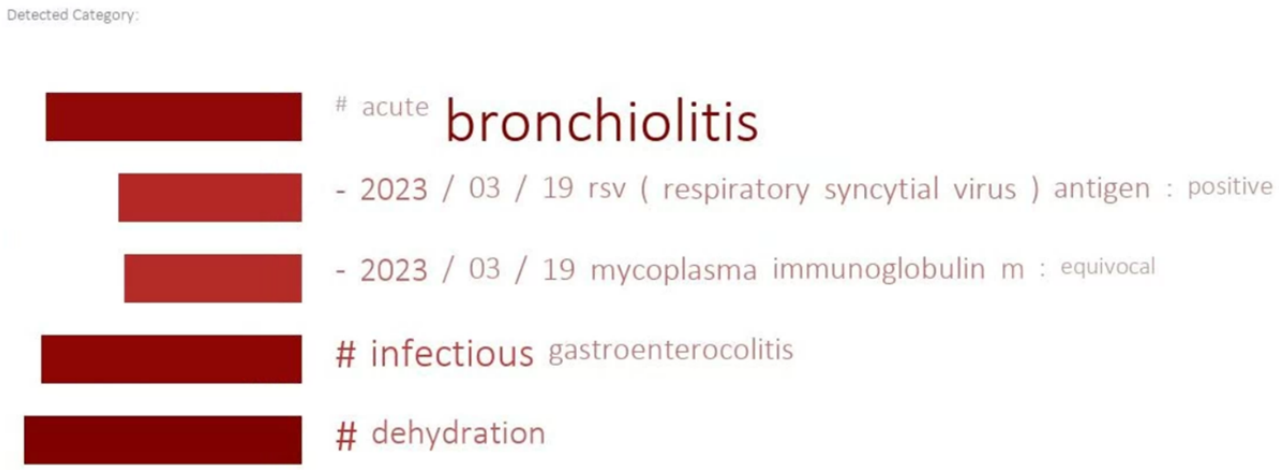

The user interface and coding system were deployed and tested in January 2023 at KMUCHH. An overview of the integrated workflow is displayed in Figure 5. First, the CCS uses the user interface developed for the AI-assisted coding system to send requests to the hospital information system. Once the hospital information system accepts the request, the requested EMRs are transferred to the structured query language server database specifically developed for the study. Subsequently, a notification is dispatched to the user interface. Upon receiving the notification, the user interface retrieves the corresponding medical record from the database and uses the developed model to generate suggested *ICD-10-CM* codes. The recommendations are then presented in area (5) of the user interface as shown in Figure 3. Finally, the CCS reviews the recommendations and selects the final *ICD-10-CM* codes in area (4) of Figure 3, which are subsequently saved in the database.

XSL•FO
RenderX

Figure 5. Workflow of the natural language processing–driven artificial intelligence–assisted coding system deployed in Kaohsiung Medical University Chung-Ho Memorial Hospital. ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SQL: structured query language.

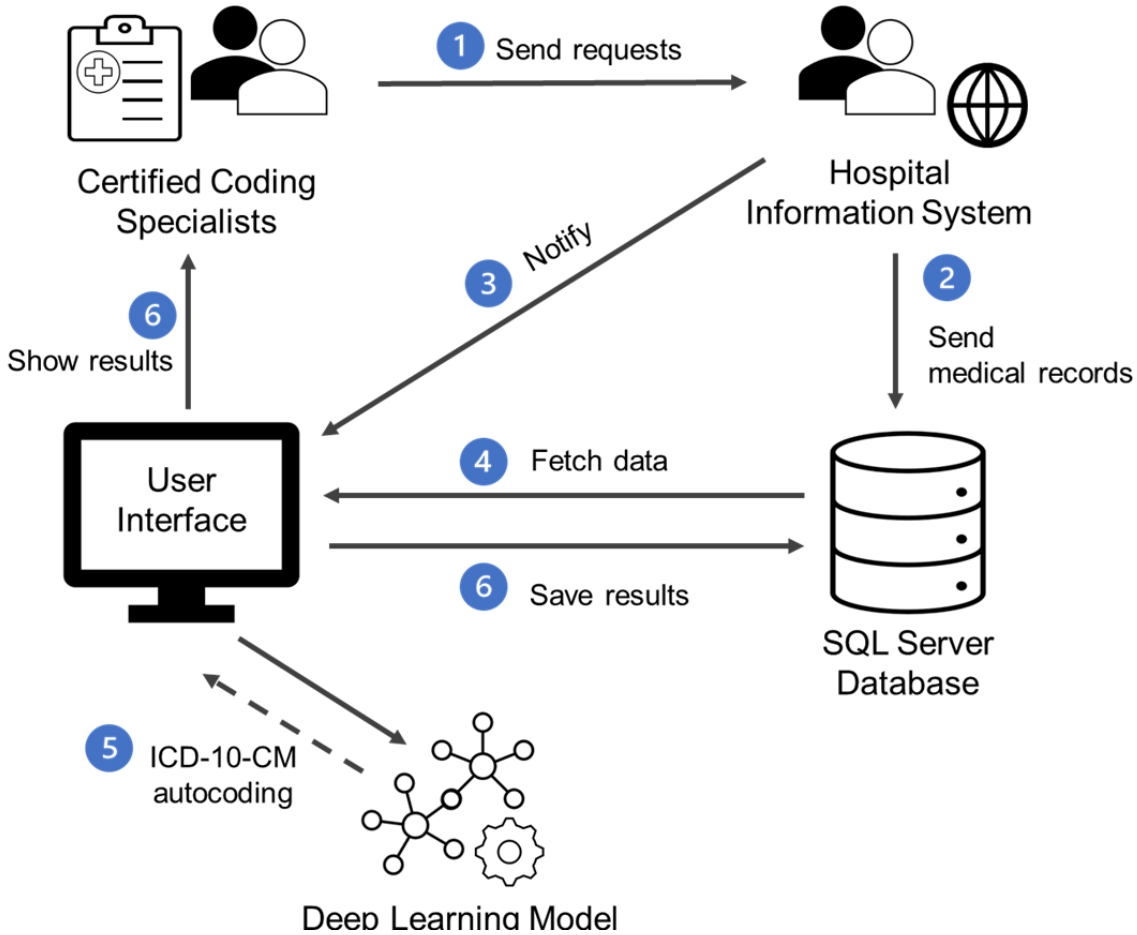

Tw-DRG Payment System

Figure S1 in Multimedia Appendix 1 depicts the current implementation practice of the Tw-DRG payment system. The initial step of this process involves the conversion of diagnoses from physician-provided discharge summaries into *ICD-10-CM* codes and follows the principal diagnosis selection principle defined by NHIA to use the main cause of hospitalization as the principal diagnosis. In cases where multiple principal diagnoses are identified simultaneously, the one with the highest medical cost will be selected. Our NLP-driven AI-assisted coding system plays a pivotal role in the first and second steps of the process, automating the *ICD-10-CM* coding process based on discharge summaries and determining the principal diagnosis.

As listed in Table 2, the cause of a patient's hospitalization identified by our system as the principal diagnosis is subsequently categorized into the corresponding MDC under the Tw-DRG calculation software program. The MDC schematic classification for inpatient cases consists of 26 distinct categories, spanning from pre-MDC to MDC 1 through MDC 25. Once the principal diagnosis is determined for a patient, similar therapeutic diseases or procedures are further divided into multiple DRGs. This division considers various factors including the patient's age and gender, the presence of comorbidities or complications (secondary diagnoses), discharge status, etc. The prospective inpatient costs that the NHIA should reimburse hospitals are calculated by leveraging historical data from the health care industry as a foundational reference. In sum, the DRG provides essential information for this specific hospitalization, including details related to health insurance reimbursement, relative weight, and the presence of comorbidities and complications.

XSL•FO
RenderX

Table 2. The 26 major diagnostic categories (MDC) in Taiwan diagnosis related groups.

MDC	Title
Pre	• Heart Transplant • Liver Transplant • Bone Marrow Transplant • Tracheostomy • Lung Transplant • Pancreatic Transplant
1	• Diseases and Disorders of the Nervous System
2	• Diseases and Disorders of the Eye
3	• Diseases and Disorders of the Ear, Nose, Mouth and Throat
4	• Diseases and Disorders of the Respiratory System
5	• Diseases and Disorders of the Circulatory System
6	• Diseases and Disorders of the Digestive System
7	• Diseases and Disorders of the Hepatobiliary System and Pancreas
8	• Diseases and Disorders of the Musculoskeletal System and Connective Tissue
9	• Diseases and Disorders of the Skin, Subcutaneous Tissue and Breast
10	• Endocrine, Nutritional and Metabolic Diseases and Disorders
11	• Diseases and Disorders of the Kidney and Urinary Tract
12	• Diseases and Disorders of the Male Reproductive System
13	• Diseases and Disorders of the Female Reproductive System
14	• Pregnancy, Childbirth and the Puerperium
15	• Newborns and Other Neonates with Conditions Originating in the Perinatal Period
16	• Diseases and Disorders of the Blood and Blood Forming Organs and Immunological Disorders
17	• Myeloproliferative Diseases and Disorders, and Poorly Differentiated Neoplasm
18	• Infectious and Parasitic Diseases (Systemic or Unspecified Sites)
19	• Mental Diseases and Disorders
20	• Alcohol/Drug Use and Alcohol/Drug Induced Organic Mental Disorders
21	• Injuries, Poisonings and Toxic Effects of Drugs
22	• Burns
23	• Factors Influencing Health Status and Other Contacts with Health Services
24	• Multiple Significant Trauma
25	• Human Immunodeficiency Virus Infection

XSL•FO
RenderX

Experiment Configurations

Both deep learning models were implemented using CUDA 12.0 (Nvidia Corp) and the PyTorch (Meta Platforms) libraries trained on machines equipped with an Intel i7-13700 processor (Intel Corporation), 64 GB of RAM, and an NVIDIA GeForce RTX 4090 24-GB graphics card (Nvidia Corp). During the training of these 2 models, different hyper-parameter configurations were used. For HAN, we set the batch size to 32, the learning rate to 1e-3, and the number of epochs to 500, and used the Adam optimizer [32] for optimization. During the training process, we implemented an early stopping strategy with a patience value of 50, which was triggered if there was no improvement in the F_1-score or loss, or if the loss on the validation set reached 0. For GPT-2, we configured the learning rate to be 1.5e-4, set the number of epochs to 10, and set the batch size to 4 to prevent out-of-memory issues. We used the AdamW optimizer [33] for parameter optimization, setting the epsilon to 1.0e-09 to ensure that the model stops updating when the learning rate drops below this threshold. In addition, to align with the maximum input length in the GPT-2 architecture, we set the maximum input length for each training instance to 1024.

The evaluation metric used to assess the performance of the developed models is the *F*-measure, calculated from the harmonic mean of the precision and recall. Precision and recall are computed based on the counts of correctly predicted *ICD-10-CM* codes (true positives), incorrectly predicted *ICD-10-CM* codes (false positives), and undetected *ICD-10-CM* codes (false negatives). These values are determined by comparing the model's predictions with the ICD codes assigned by the CCS. These metrics are calculated using the following equations 5, 6 and 7 for the precision, recall and F_1-scores.

It is worth noting that in the evaluation step, in addition to assessing the performance of the autocoding systems on the original compiled corpus, we also compiled a new Tw-DRGs data set consisting of EMRs collected from the real KMUCHH environment. The new data set was processed by the aforementioned autocoding systems in series with NHIA's DRG calculation software. In addition, to assess the reliability of our NLP-driven AI-assisted coding system in the Tw-DRG inpatient payment process, we used the κ measure as an indicator of agreement for MDC between our system and CCS curated results. Specifically, we used the Cohen κ [34] to estimate the κ values. For this comparison, we used the Tw-DRG calculation program provided by the NHIA in Taiwan. This program uses inputs, such as the predicted main code, other codes, gender, and date of birth to perform MDC classification and DRG grouping. We developed a program to analyze the output from the Tw-DRG calculation program for the comparison study.

Ethical Considerations

This study was approved by the KMUCHH Institutional Review Board (protocol title: "Developing Artificial Intelligence model to support *ICD-10-CM* and PCS coding and comparing the performance between machine coding and manual coding"; institutional review board number: KMUHIRB-E(II)-20230214). As the study utilized secondary data collected under the original institutional review board approved protocol, no additional informed consent was required. All data collected from KMUCHH was aggregated for research purposes adhering to fair use principles. Participant privacy was rigorously protected through anonymization, and stringent confidentiality measures were implemented. Given that the study involved only the analysis of existing, deidentified data, there was no direct interaction with participants, and thus, no compensation was provided. Additionally, the study strictly adhered to all applicable local, national, regional, and international laws and regulations concerning the protection of personal information, privacy, and human rights.

Results

Overview

In the following subsections, we demonstrated the effectiveness of deploying NLP-driven AI-assisted coding systems, specifically using HAN and GPT-2 models, in the context of *ICD-10-CM* coding and Tw-DRGs process. We provide a comparative analysis of the AI-assisted coding results with CCS curated results for Tw-DRGs using the data set gathered from the deployment of the developed AI-assistant system in the hospital environment. This analysis validates the potential of NLP-driven AI-assisted coding in facilitating the Tw-DRGs process, highlighting its effectiveness in improving coding.

Performance Comparison of the Developed Deep Learning Models on the ICD-10-CM Coding Task

Tables 3 and 4 offer a comprehensive analysis of the performance of the baseline models as well as the HAN and GPT-2 models in *ICD-10-CM* autocoding on the test set. In addition to the full code results, Table 3 shows the performance of each model on the top-50 codes (Top-50 F) which is often reported in previous research papers. The list of the top-50 *ICD-10-CM* codes can be found in Multimedia Appendix 2. The results presented in Table 3 highlight the superior performance of the GPT-2 model compared to the other models. GPT-2 exhibited higher precision, recall, and *F*-measure in its overall performance compared to HAN, with improvements of 0.134, 0.077, and 0.107, respectively. Additionally, the boost in the F_1-score for the main code further emphasizes the advantage of applying the GPT-2 model over the other models. Both HAN and GPT-2 models achieved satisfactory top-50 F_1-scores but obtained significantly lower F_1-scores for the full codes owing to the long-tailed distribution of the *ICD-10-CM* codes assigned by the CCS.

XSL•FO
RenderX

Table 3. Overall performance evaluation of hierarchical attention network (HAN) and GPT-2 models in the International Classification of Diseases, 10th Revision, Clinical Modification, autocoding on the test set.

Model	Precision	Recall	*F*-measure	Top-50 (F)	Main code (F)
HAN	0.529	0.594	0.560	0.818	0.429
GPT-2	0.663	*0.671* [a]	*0.667* [a]	*0.851* [a]	*0.575* [a]
BiGRU[b]	*0.765* [a]	0.451	0.567	0.717	0.250
BERT[c]	0.759	0.366	0.494	0.698	0.215

[a]The best value for each performance metric is italicized.

[b]BiGRU: bidirectional gated recurrent unit.

[c]BERT: bidirectional encoder representations from transformers.

Table 4 provides a more detailed breakdown of the *ICD-10-CM* subcategory performance for the 2 models. The results underscore the consistent superiority of GPT-2 over HAN across all categories, particularly in situations with limited data (<10%). Notably, GPT-2 acquired better F_1-scores in the T, H, and O categories compared to HAN with an increase of 0.387, 0.279, and 0.269, respectively. This accentuates the advantage of GPT-2 due to its pretrained nature, which enables it to extract crucial features in a more efficient manner compared to HAN when only limited training data are provided.

By contrast, both models revealed low performance in the X and Y categories. These 2 categories suffered from a lack of data, as they have the lowest number of training instances as outlined in the last column of Table 4. The restricted amount of training data poses a challenge to the models in achieving better performances in these specific categories.

XSL•FO
RenderX

Table 4. The International Classification of Diseases, Tenth Revision, Clinical Modification, category-specific performance comparison of the hierarchical attention network (HAN) and GPT-2 on the test set.

Categories (block)	Chapters	Models (*F*-measure)		Discharge summaries in the training set (n=105,101), n (%)
		HAN	GPT-2	
A	I	0.851	*0.955* [a]	15,267 (14.53)
B	I	0.703	*0.889* [a]	20,441 (19.45)
C	II	0.812	*0.890* [a]	71,202 (67.75)
D	II, III	0.702	*0.814* [a]	28,393 (27.01)
E	IV	0.797	*0.888* [a]	63,545 (60.46)
F	V	0.611	*0.833* [a]	8835 (8.41)
G	VI	0.566	*0.794* [a]	13,236 (12.59)
H	VII, VIII	0.576	*0.855* [a]	9824 (9.35)
I	IX	0.781	*0.833* [a]	82,005 (78.02)
J	X	0.756	*0.880* [a]	31,233 (29.72)
K	XI	0.714	*0.843* [a]	48,827 (46.46)
L	XII	0.496	*0.730* [a]	7369 (7.01)
M	XIII	0.549	*0.764* [a]	20,312 (19.33)
N	XIV	0.783	*0.894* [a]	39,504 (37.59)
O	XV	0.583	*0.852* [a]	3887 (3.7)
P	XVI	0.738	*0.857* [a]	4297 (4.09)
Q	XVII	0.571	*0.652* [a]	2918 (2.78)
R	XVIII	0.545	*0.617* [a]	24,544 (23.35)
S	XIX	0.661	*0.742* [a]	20,850 (19.84)
T	XIX	0.297	*0.684* [a]	8505 (8.09)
V	XX	0.446	*0.584* [a]	3613 (3.44)
W	XX	0.459	*0.619* [a]	3558 (3.39)
X	XX	0.270	*0.444* [a]	1968 (1.87)
Y	XX	0.425	*0.458* [a]	2652 (2.52)
Z	XXI	0.657	*0.747* [a]	39,577 (37.66)

[a]The best value for each performance metric is italicized.

Comparative Analysis of NLP-Driven AI-Assisted Coding Results With CCS for Tw-DRGs in the Real Hospital Environment

In this subsection, we provide a comparative analysis of the performance of the deployed systems from the perspective of Tw-DRGs. The performance of Tw-DRGs was estimated on the additional data set of 2632 discharge cases stored in the SQL server database indicated in Figure 5, which was collected from February 2023 to April 2023.

As described in the previous section on the Tw-DRGs payment system, the principal diagnosis is the crucial factor in determining the MDC, while secondary diagnoses are only used to determine different Tw-DRGs distributions within the same MDC. The principal diagnosis is determined based on the reason for the patient's hospitalization and only a single disease can be selected. If there are multiple diseases for which the patient is receiving treatment upon admission, selecting any one of them as the principal diagnosis is not considered an error but may affect the results of Tw-DRGs. Hence, the principal diagnosis selection cannot be assessed by solely comparing the system output against the *ICD-10-CM* main codes initially assigned by the CCS. Instead, it necessitates the expertise of the CCS to reevaluate both the EMRs and the codes proposed by the developed systems. Therefore, to provide a comparative

XSL•FO
RenderX

analysis of the 2 developed models in terms of their performance in defining the principal diagnosis as shown in Table 5, the senior CCS, author ATL, who is the third author of the paper, examined the discharged cases to determine the accuracy of principal diagnoses suggested by the AI-assisted coding systems based on the HAN and GPT-2 models.

Table 5. Comparison of performance between the deep learning models and certified coding specialist–curated results for principal diagnosis in discharge cases.

Models and conditions	February 2023 (n=748), n (%)	March 2023 (n=991), n (%)	April 2023 (n=893), n (%)
HAN[a] (*F*-measure=0.524)			
All correct[b]	79 (10.6)	181 (18.3)	177 (19.8)
Principal diagnosis with incorrect secondary diagnoses[c]	462 (61.8)	477 (48.1)	369 (41.3)
No principal diagnosis[d]	181 (24.2)	277 (27)	285 (31.9)
All incorrect[e]	26 (3.5)	56 (5.5)	62 (6.9)
GPT-2 (*F*-measure=0.621)			
All correct[b]	130 (17.4)	161 (16.2)	144 (16.1)
Principal diagnosis with incorrect secondary diagnoses[c]	443 (59.2)	599 (60.4)	524 (58.7)
No principal diagnosis[d]	165 (22.1)	186 (18.8)	183 (20.5)
All incorrect[e]	24 (3.2)	29 (2.9)	43 (4.8)

[a]HAN: hierarchical attention network.

[b]Autocoding results were identical to certified coding specialists.

[c]Autocoding system correctly identified the principal diagnosis, but discrepancies exist in ≥1 secondary diagnosis codes assigned by certified coding specialists.

[d]Autocoding system's principal diagnosis was different from certified coding specialists.

[e]Autocoding results were entirely different from certified coding specialists.

The first column of Table 5 shows the *F*-measure estimates for HAN (0.524) and GPT-2 (0.621) in their suggestion for main codes in comparison with the CCS coding results. The reported *F*-measures align with our observation on the test set and validate GPT-2 as a more reliable model. The manual reassessment of the results performed by the senior CCS according to the definition of principal diagnosis is displayed in the second to fourth columns of the table. Notably, most (1308/2632, 49.7%) of the cases of the 2 deep learning models when reviewed against the manual coding of the CCS fall under the category "principal diagnosis with incorrect secondary diagnoses," followed by the category "no principal diagnosis." The "all incorrect" cases had the lowest proportion among all categories. Predictions of GPT-2 achieved a better "all correct" category coverage and once again demonstrate superiority over HAN in facilitating the ICD coding task. The respective overall correct rate for principal diagnosis for HAN and GPT-2, which takes into account both categories "all correct" and "principal diagnosis with incorrect secondary diagnoses," was 0.663 and 0.7602.

In Table 6, we extended the comparison to assess the agreement of MDC between the 2 developed models and the CCS. The first, second, and third columns of Table 6 show the MDC estimated results for the codes assigned by CCS, HAN, and GPT-2. The last 2 columns show the estimated κ values corresponding to the 2 models against the CCS.

XSL•FO
RenderX

Table 6. Comparison of the κ values between 2 deep learning models and a senior certified coding specialist (CCS) for major diagnostic category (MDC) in discharge cases spanning from February 2023 to April 2023 at Kaohsiung Medical University Chung-Ho Memorial Hospital.

MDC	CCS	Model		κ value[a]	
	Manual coding (n=2632), n (%)	HAN[b] (n=2632), n (%)	GPT-2 (n=2632), n (%)	HAN (average 0.592)	GPT-2 (average 0.714)
Pre	0 (0)	0 (0)	0 (0)	__[c]	—
1	509 (19.34)	280 (10.64)	497 (18.88)	0.670[d]	0.803[d]
2	38 (1.44)	9 (0.34)	28 (1.06)	0.300	0.724[d]
3	132 (5.02)	113 (4.29)	124 (4.71)	0.689[d]	*0.852* [e]
4	302 (11.47)	309 (11.74)	287 (10.90)	*0.845* [e]	*0.826* [e]
5	184 (6.99)	310 (11.78)	213 (8.09)	0.607	0.786[d]
6	229 (8.7)	217 (8.24)	222 (8.43)	0.775[d]	*0.818* [e]
7	84 (3.19)	90 (3.42)	77 (2.93)	0.710[d]	*0.827* [e]
8	87 (3.31)	66 (2.51)	87 (3.31)	0.576	0.786[d]
9	116 (4.41)	83 (3.15)	95 (3.61)	0.692[d]	0.729[d]
10	132 (5.02)	237 (9)	167 (6.34)	0.505	0.642[d]
11	120 (4.56)	205 (7.79)	99 (3.76)	0.648[d]	0.738[d]
12	3 (0.11)	7 (0.27)	3 (0.11)	0.362	0.666[d]
13	8 (0.30)	8 (0.30)	10 (0.38)	0.749[d]	*0.889* [e]
14	41 (1.56)	11 (0.42)	38 (1.44)	0.419	*1.000* [e]
15	15 (0.57)	9 (0.34)	16 (0.61)	0.635[d]	0.708[d]
16	57 (2.17)	64 (2.43)	44 (1.67)	0.624[d]	0.728[d]
17	2 (0.08)	3 (0.11)	1 (0.04)	0.399	0.666[d]
18	505 (19.19)	465 (17.67)	483 (18.35)	*0.870* [e]	0.777[d]
19	0 (0)	23 (0.87)	3 (0.11)	—	—
20	0 (0)	2 (0.08)	2 (0.08)	—	—
21	21 (0.8)	24 (0.91)	16 (0.61)	0.404	0.646[d]
22	1 (0.04)	0 (0)	0 (0)	—	—
23	42 (1.6)	24 (0.91)	43 (1.63)	0.366	0.398
24	4 (0.15)	6 (0.23)	3 (0.11)	0.599	–0.010
25	0 (0)	2 (0.08)	0 (0)	—	—
None	0 (0)	65 (2.47)	74 (2.81)	—	—

[a]The average κ values are calculated using arithmetic averaging of the MDC categories, excluding the blank cells.

[b]HAN: hierarchical attention network.

[c]Not applicable.

[d]The κ value falls within the range of 0.61 to 0.80, indicating substantial reliability.

[e]The κ value falls within the range of 0.81 to 1.00, indicating almost perfect reliability.

Our analysis revealed that the CCS did not assign any cases to pre-MDC, MDC 19, MDC 20, and MDC 25 for the curated patients. By contrast, HAN erroneously assigned 23 (0.87%) of the 2632 cases to MDC 19, 2 (0.08%) cases to MDC 20, and 2 (0.08%) cases to MDC 25. GPT-2 made fewer mistakes in classifying 3 (0.11%) of the 2632 cases to MDC 19 and 2 (0.08%) cases to MDC 20. Following an analysis conducted by using SPSS (version 19; IBM Corp), outputs of both models along with the coding results manually curated by the CCS underwent Cohen κ evaluation for agreement. The results indicated that the average κ value for GPT-2 (0.714/Substantial) was approximately 12.2 percentage points higher than that of

XSL•FO
RenderX

HAN (0.592/Moderate). Notably, the average κ value of GPT-2 exceeded 0.81 (highlighted in italics in Table 6) across 6 categories of the MDC, indicating almost perfect agreement between the CCS and GPT-2. These observations exhibit the effectiveness of NLP-driven systems in supporting the work of CCSs. It is worth mentioning that the κ value of GPT-2 in MDC 24 (pertaining to injuries involving 2 or more body systems) was recorded as –0.01 (poor). A close examination of the 7 cases of the compiled data set in MDC 24 reveals that while 6 (86%) cases shared consistent principal diagnoses, discrepancies arose in the secondary diagnoses. These discrepancies involved instances of overcoding for lung contusion and traumatic hemothorax, as well as undercoding for hepatic contusion and cervical spinal cord injury. Consequently, these inconsistencies resulted in a complete mismatch between the MDC results of GPT-2 and CCS-curated results.

The Effectiveness of Applying Models Acquired With Biomedical Knowledge

We conducted an ablation study comparing the performance of 2 fine-tuned GPT-2 models: one fine-tuned on PubMed documents and the other being the original GPT-2 model released by OpenAI. This study aimed to assess the necessity of using models fine-tuned on biomedical documents. The results indicate that the PubMed fine-tuned model achieves slightly better performance than the original GPT-2 model, with F_1-score improvements of 0.05 for full coding and 0.03 for main coding.

In addition, we used the following prompt template to assess the knowledge of the 2 GPT-2 models for the tail-50 ICD codes.

> "<*ICD-10-CM* Code> < *ICD-10-CM* Description> The condition involves..."

From the generated texts, we observed that both GPT-2 models generally possess a considerable understanding of *ICD-10-CM* terms. For instance, when prompted with the *ICD-10-CM* code "C9502: Acute leukemia of unspecified cell type, in relapse," the GPT-2 model generated the following response: "This condition involves erythroid and myeloid cells and is associated with a poor prognosis." This response accurately describes a challenging scenario in leukemia management, where the disease has relapsed with the involvement of both erythroid and myeloid cell lines, often indicating a poor prognosis.

Discussion

Error Analysis

Ji et al [35] discussed several challenges encountered during the implementation of automated clinical coding procedures. One significant challenge is that electronic health records often contain a variety of professional medical vocabularies alongside noisy information, including nonstandard synonyms and misspellings. To address this issue, our study used a variation of GPT-2 fine-tuned using 0.5 million PubMed abstracts to enhance its ability to recognize medical terminologies. The observations from the prompting results, as described in the Methods section, highlight the biomedical knowledge acquired by the pretrained GPT-2 model. This enhancement enabled the GPT-2 model perform better than the HAN model, which was not pretrained on biomedical data, particularly in categories, such as XIX-T, VII, VIII, and XV, where training instances were limited. Our analysis also reveals that HAN tends to attend to noisy information and sometimes generates completely irrelevant codes, which can be frustrating for CCSs. The result underscores the GPT-2 model's adeptness in comprehending medical context and its effectiveness in mitigating the challenges associated with clinical coding procedures.

Another problem highlighted by Ji et al [35] is the high-dimensionality of medical codes and the long-tailed distribution, which results in limited corresponding training instances that are necessary for effective model training. EMRs associated with multiple diagnoses are particularly complex issues known as a multi-label extreme classification problem characterized by a vast label set, which was also encountered by our HAN model. In our implementation of HAN, the high-dimensional label space comprises 11,653 codes, magnifying the complexity of the task. On the contrary, while the high-dimensionality issue does not significantly impact the GPT-2 model, the scarcity of training instances still presents a considerable challenge. For instance, codes associated with the rarity of congenital conditions, such as those in category XVII (congenital malformations, deformations, and chromosomal abnormalities), are less frequent in the data set. This rarity results in fewer cases for the models to learn from, thereby hindering accurate predictions. In addition, we noted that certain codes manually labeled by CCSs, such as D489, K5100, Z8616, and others, were not present in the training set. Consequently, both models struggled to predict these unlearned codes, illustrating a critical limitation in the presented training data.

In addition, there are categories with abundant training instances where both models' performance remains unsatisfactory. Notably, codes from category XVIII appear in 23.35% of the training set summaries, yet the F_1-scores for both models are below 0.65. Upon analyzing the predicted results, we observed that while codes from the chapter, symptoms, signs, and abnormal clinical and laboratory findings, frequently appear in discharge summaries, disease classification rules often do not require separate coding for symptoms and signs that are related to a diagnosed disease. This discrepancy can lead to confusion for the developed models, causing them to misinterpret these entries.

To tackle this issue, we intend to augment the training set by incorporating additional coded content curated by the CCS in future iterations. This will enable our system to learn novel coding content that the system did not previously come across. Moreover, we plan to explore the use of ICD code representation methods, as proposed by Vu et al [36] and Wu et al [37], into our models. By integrating these methods, we aim to further enhance our system's performance and robustness in addressing the challenges associated with automated clinical coding.

Moreover, although models with neural attentions learned to infer implicit relationships in discharge summaries by interpreting contextual expressions with weighted attentions, there are instances where pertinent information required for the coding judgment criteria is absent from the discharge summaries, for example, for chapters XIX and XX (injury, poisoning, and

XSL•FO
RenderX

certain other consequences of external causes and external causes of morbidity and mortality). According to the classification rules, codes from these chapters are often used in conjunction with other codes. The data sources for these codes are not limited to discharge summaries but also include nursing records, imaging reports, and emergency department records. The developed models' poor performance in these chapters may be due to the complexity and the need to integrate information from various sources, which is currently unavailable in the current implementation.

Finally, we noticed that certain codes only existed in the test set. These scenarios can lead to incorrect *ICD-10-CM* codes generated by the model. Some examples as such are listed subsequently.

In the discharge diagnosis, the developed system predicted the *ICD-10-CM* code for "# Heart failure" as I509:

> *# Dyspnea with desaturation, focus on HAP (hospital-acquired pneumonia) and coronavirus disease of 2019 # Sepsis, focus on HAP (hospital-acquired pneumonia) and catheter related UTI (urinary tract infection) # Hyponatremia # Heart failure.*

However, the correct coding is I5020. This discrepancy is primarily due to supplementary information recorded in the medical history on admission, specifically mentioning "# chronic systolic heart failure." Following discussions with the CCS, it was agreed that future enhancements will attempt to merge the content from the history on admission to enable the model to learn from a broader range of medical history information, thus ensuring more accurate coding outcomes.

Comparison With Prior Work

Li et al [38] introduced the DeepLabeler, a deep learning architecture based on a combination of the convolutional neural network (CNN) with the document-to-vector technique [39] to extract and encode local and global features for *ICD-9* coding. Their approach achieved micro *F*-measures of 0.335 and 0.408 in the public multiparameter intelligent monitoring in intensive care (MIMIC)-II and MIMIC-III data sets, respectively. Zeng et al [40] transferred the knowledge learned from the Medical Subject Headings indexing domain using the large-scale biomedical semantic indexing competition challenge data set [41] to enhance the performance of the developed multi-scale CNN for automatic *ICD-9* coding. Their approach achieved a micro *F*-measure of 0.420 on the public MIMIC-III data set. Chen et al [30] used diagnostic records from the National Taiwan University Hospital to build a data set with a total of 1,043,124 labels (using 14,602 unique codes as prediction candidates) and developed a deep neural network classification model based on the bidirectional gated recurrent unit along with the BERT-based word representation method. This model obtained an *F*-measure of 0.715 for *ICD-10-CM* coding on their test set. Wu et al [37] proposed a pseudo label-wise attention mechanism aimed at automatically combining attention modes of similar ICD codes to tackle the issue of unbalanced *multi-label* classification in ICD coding. Their methodology involved using a bidirectional long short-term memory in tandem with the pseudo label-wise attention mechanism to represent EMRs as vectors. They then used a BERT-based pretrained model to determine the vector representations of the *ICD-10* codes. Finally, they calculated the similarity between EMR vectors and ICD vectors to determine the assigned codes. This approach yielded a micro *F*-measure of 0.806 on their private Chinese Xiangya data sets. In a separate study, Bhutto et al [42] proposed a deep recurrent-CNN architecture with a lambda-scaled attention module. Their approach yielded micro *F*-measures of 0.862 and 0.705 on a private Pakistan clinical notes. In comparison with these studies, the GPT-2 model in our research demonstrates competitive performance, particularly excelling in cases where the training data for specific categories is limited, showcasing its robustness in handling diverse and challenging scenarios.

In addition, certain previous research also explored the impact of erroneous *ICD-10-CM* coding on hospital finances. Zafirah et al [43] studied the potential loss attributable to clinical coding errors in a Malaysian teaching hospital. Their findings indicate a high prevalence of error coding in medical records, particularly concerning secondary diagnosis codes, which reached 81.3% (377/464). The estimated financial impact on the medical discipline in this hospital amounted to a potential profit loss of RM 85,804.92 (US $19,617.05). Toner et al [15] conducted a retrospective comparative analysis of case records for patients with the M966 diagnosis code (periprosthetic fracture) in a district general hospital. Their work revealed that *ICD-10-CM* coding errors resulted in a loss of £25,000 (US $33,029.97) when compared to the actual hospital revenue. In contrast with relevant studies, our examination of the consistency assessment of the autocoding system and CCS coding indicates promising prospects in reducing manual workload and providing coding references to minimize human errors. The error rate of CCS detected with the assistance of the AI-assisted coding system is 1.9% (50/2632; number of coding errors/total cases).

Pivotally, our work serves as the first endeavor to examine the feasibility of the combination of *ICD-10* coding with DRGs in the real hospital environment, which indicates a significant advancement in the area. In conclusion, the implementation of NLP-driven AI-assisted coding systems contributes to a reduction in CCS coding errors and manual workload, thereby enhancing the overall efficiency of the coding process, lowering error rates, and mitigating financial losses.

Limitations

While our study demonstrates the potential of NLP-driven AI-assisted coding systems in improving *ICD-10-CM* coding accuracy and efficiency, several limitations should be acknowledged. First, the data set used for this study was sourced from a single hospital, which may limit the generalizability of the presented results. Future studies should include data from multiple hospitals, covering diverse geographical regions and varying patient demographics, to ensure broader applicability of the findings and to validate the robustness of the models across different settings.

Second, this study involved only 1 senior CCS to conduct the κ analysis for the agreements among the developed models. This limitation means that the results may not necessarily extend

XSL•FO
RenderX

to other CCSs, particularly those with different levels of experience. Future works should consider recruiting a larger and more diverse group of CCSs, including both senior and junior coders, to evaluate the helpfulness of the proposed models comprehensively. In addition, studying the reduction in coding time and the impact on workflow efficiency in practical settings would provide valuable insights into the real-world benefits of AI-assisted coding systems.

Third, this study concentrated solely on Tw-DRGs, which are specific to Taiwan's health care system. As a result, the findings may not be directly generalizable to other DRG systems used in different countries or regions, or to DRGs estimated in different periods. Further research is needed to ascertain the applicability of our conclusions to other DRG systems worldwide. Investigating the performance of the models in different international contexts and updating the models to reflect changes in DRG systems over time would enhance the relevance and utility of the findings.

Conclusions

In summary, our study has demonstrated the effectiveness of deploying NLP-driven AI-assisted coding systems, specifically using HAN and GPT-2 models, in the context of *ICD-10-CM* coding and Tw-DRGs process. The comparative analysis revealed that GPT-2 consistently performed better than HAN and exhibited higher precision, recall, and F_1-scores. This superior performance was particularly evident in scenarios with limited data, highlighting the robustness of GPT-2 in extracting vital features from discharge summaries. Moreover, the evaluation of principal diagnosis and MDC in Tw-DRGs showcased the utility of the developed models. GPT-2, in particular, acquired higher agreement values and made fewer mistakes in MDC classification. In the practical deployment of the system within a hospital environment, the comparative analysis with CCS validated the potential of NLP-driven autocoding in the Tw-DRGs process. Despite encountering certain discrepancies, our study demonstrates the significant value of the implemented models as tools that offer insights and support to CCSs during the coding process. The deployment of AI-assisted coding systems has the potential to enhance coding accuracy while simultaneously reducing manual workload, leading to improved process efficiency, lower error rates, and ultimately, a decrease in financial losses.

While our proposed system effectively alleviates the manual workload of CCSs, our error analysis has also unveiled notable challenges. These include the absence of coding judgment information in discharge summaries, presence of coding answers not included in the training set, and the need for suggesting 1 or more main codes for the development of *ICD-10* coding system to assist in the DRG process. These findings underscore areas for further improvement and refinement in future iterations of our system. Addressing these challenges will be pivotal in enhancing the efficacy and reliability of automated coding systems, thereby maximizing their potential to support and streamline both clinical coding and DRG processes.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data for the process of Taiwan diagnosis related groups.
[DOCX File , 323 KB - jmir_v26i1e58278_app1.docx]

Multimedia Appendix 2
Supplementary data for the top 50 International Classification of Diseases, Tenth Revision, Clinical Modification, codes.
[DOCX File , 15 KB - jmir_v26i1e58278_app2.docx]

References

1. International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index. World Health Organization. 1978. URL: https://iris.who.int/handle/10665/39473 [accessed 2024-04-29]
2. Sivashankaran S, Borsi JP, Yoho A. Have ICD-10 coding practices changed since 2015? AMIA Annu Symp Proc 2019;2019:804-811 [FREE Full text] [Medline: 32308876]
3. de Lissovoy G. Codes, coding, and COVID-19. Med Care 2020 Dec;58(12):1035-1036. [doi: 10.1097/MLR.0000000000001430] [Medline: 33003050]
4. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. J Am Med Inform Assoc 2010 May 01;17(3):274-282 [FREE Full text] [doi: 10.1136/jamia.2009.001230] [Medline: 20442144]
5. Mills RE, Butler RR, McCullough EC, Bao MZ, Averill RF. Impact of the transition to ICD-10 on Medicare inpatient hospital payments. Medicare Medicaid Res Rev 2011 Jun 06;1(2):E1-13 [FREE Full text] [doi: 10.5600/mmrr.001.02.a02] [Medline: 22340773]
6. Lee LM, Thacker SB. Public health surveillance and knowing about health in the context of growing sources of health data. Am J Prev Med 2011 Dec;41(6):636-640. [doi: 10.1016/j.amepre.2011.08.015] [Medline: 22099242]

XSL•FO
RenderX

7. Uzkuraitis C, Hastings K, Torney B. Casemix funding optimisation: working together to make the most of every episode. Health Inf Manag 2010 Oct 01;39(3):47-49. [doi: 10.1177/183335831003900309] [Medline: 28683680]
8. Ho C, Guilcher SJ, McKenzie N, Mouneimne M, Williams A, Voth J, et al. Validation of algorithm to identify persons with non-traumatic spinal cord dysfunction in Canada using administrative health data. Top Spinal Cord Inj Rehabil 2017 Oct;23(4):333-342 [FREE Full text] [doi: 10.1310/sci2304-333] [Medline: 29339909]
9. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014 Mar 01;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
10. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015 Apr 24;350(apr24 11):h1885 [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
11. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018 Jul 20;1(1):53-68 [FREE Full text] [doi: 10.1146/annurev-biodatasci-080917-013315] [Medline: 31218278]
12. Wu JJ. Implementation and outcome of Taiwan diagnosis-related group (DRG) payment system. Georgia State University. 2015. URL: https://scholarworks.gsu.edu/iph_theses/357/ [accessed 2024-04-29]
13. Wu CY, Chien LC, Lin CC, Ma HM, Hu RH, Chen CL, et al. Original article: the impacts of DRG payment system on financial balance of multiple trauma: experiences of three trauma centers in Taiwan. Injury 2023 Sep;54(9):110703. [doi: 10.1016/j.injury.2023.03.031] [Medline: 37045657]
14. Ayub S, Scali ST, Richter J, Huber TS, Beck AW, Fatima J, et al. Financial implications of coding inaccuracies in patients undergoing elective endovascular abdominal aortic aneurysm repair. J Vasc Surg 2019 Jan;69(1):210-218 [FREE Full text] [doi: 10.1016/j.jvs.2018.04.027] [Medline: 29937283]
15. Toner E, Khaled A, Ramesh A, Qureshi MK, Al Suyyagh K, Dunkow P. Financial impact of inaccurate coding plus cost-effectiveness analysis for surgically managed patients with periprosthetic fractures. Cureus 2021 Feb 01;13(2):e13060 [FREE Full text] [doi: 10.7759/cureus.13060] [Medline: 33680602]
16. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005 Oct 11;40(5 Pt 2):1620-1639 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00444.x] [Medline: 16178999]
17. Kaur R. Distributed knowledge based clinical auto-coding system. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2019 Presented at: ACL '19; July 28-August 2, 2019; Florence, Italy p. 1-9 URL: https://aclanthology.org/P19-2001.pdf [doi: 10.18653/v1/p19-2001]
18. Donaldson MS, Corrigan JM, Kohn LT. To Err is Human: Building a Safer Health System. Washington, DC: National Academies Press; 2000.
19. Yeow JA, Ng PK, Tan KS, Chin TS, Lim WY. Effects of stress, repetition, fatigue and work environment on human error in manufacturing industries. J Appl Sci 2014 Dec 1;14(24):3464-3471. [doi: 10.3923/JAS.2014.3464.3471]
20. International statistical classification of diseases and related health problems: alphabetical index. World Health Organization. 2004. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2024-04-29]
21. Chang NW, Dai HJ, Jonnagaddala J, Chen CW, Tsai RT, Hsu WL. A context-aware approach for progression tracking of medical concepts in electronic medical records. J Biomed Inform 2015 Dec;58(Suppl):S150-S157 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.013] [Medline: 26432355]
22. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classificatio. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: NAACL '16; June 12-17, 2016; San Diego, CA p. 1480-1489 URL: https://aclanthology.org/N16-1174.pdf [doi: 10.18653/v1/n16-1174]
23. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog 2019;1(8):9 [FREE Full text]
24. Dong H, Suárez-Paniagua V, Whiteley W, Wu H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. J Biomed Inform 2021 Apr;116:103728 [FREE Full text] [doi: 10.1016/j.jbi.2021.103728] [Medline: 33711543]
25. Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimed Syst 2022 Jun 25;28(6):2335-2355 [FREE Full text] [doi: 10.1007/s00530-022-00960-4] [Medline: 35789785]
26. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DS, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023 Aug 24;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61] [Medline: 38939520]
27. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: EMNLP '14; October 25-29, 2014; Doha, Qatar p. 1532-1543 URL: https://aclanthology.org/D14-1162.pdf [doi: 10.3115/v1/d14-1162]
28. Papanikolaou Y, Pierleoni A. Dare: data augmented relation extraction with gpt-2. arXiv Preprint posted online April 6, 2020 [FREE Full text]

XSL•FO
RenderX

29. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL-HLT '19; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf
30. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. JMIR Med Inform 2021 Aug 31;9(8):e23230 [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
31. Lai CH, Shian BT, Ke CR. Demonstration website for the developed NLP-driven AI-assisted ICD-10-CM coding system. ISLAB. 2024. URL: http://140.127.114.22/Patient/Detail [accessed 2024-08-22]
32. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv Preprint posted online December 22, 2014 [FREE Full text]
33. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv Preprint posted online November 14, 2017 [FREE Full text]
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
35. Ji S, Li X, Sun W, Dong H, Taalas A, Zhang Y, et al. A unified review of deep learning for automated medical coding. ACM Comput Surv 2024 May 17:1-38. [doi: 10.1145/3664615]
36. Vu T, Nguyen DQ, Nguyen A. A label attention model for ICD coding from clinical text. In: Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. 2020 Presented at: IJCAI 20; January 7-15, 2021; Yokohama, Japan p. 3335-3341 URL: https://dl.acm.org/doi/abs/10.5555/3491440.3491901 [doi: 10.24963/ijcai.2020/461]
37. Wu Y, Zeng M, Yu Y, Li Y, Li M. A pseudo label-wise attention network for automatic ICD coding. IEEE J Biomed Health Inform 2022 Oct;26(10):5201-5212. [doi: 10.1109/jbhi.2022.3193291]
38. Li M, Fei Z, Zeng M, Wu FX, Li Y, Pan Y, et al. Automated ICD-9 coding via a deep learning approach. IEEE/ACM Trans Comput Biol Bioinform 2019;16(4):1193-1202. [doi: 10.1109/TCBB.2018.2817488] [Medline: 29994157]
39. Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning. 2014 Presented at: ICML '14; June 21-26, 2014; Beijing, China p. 1188-1196 URL: https://dl.acm.org/doi/10.5555/3044805.3045025
40. Zeng M, Li M, Fei Z, Yu Y, Pan Y, Wang J. Automatic ICD-9 coding via deep transfer learning. Neurocomputing 2019 Jan;324:43-50. [doi: 10.1016/J.NEUCOM.2018.04.081]
41. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 2015 Apr 30;16(1):138 [FREE Full text] [doi: 10.1186/s12859-015-0564-6] [Medline: 25925131]
42. Bhutto SR, Zeng M, Niu K, Khoso S, Umar M, Lalley G, et al. Automatic ICD-10-CM coding via Lambda-Scaled attention based deep learning model. Methods 2024 Feb;222:19-27. [doi: 10.1016/j.ymeth.2023.11.017] [Medline: 38141869]
43. Zafirah SA, Nur AM, Puteh SE, Aljunid SM. Potential loss of revenue due to errors in clinical coding during the implementation of the Malaysia diagnosis related group (MY-DRG) Casemix system in a teaching hospital in Malaysia. BMC Health Serv Res 2018 Jan 25;18(1):38 [FREE Full text] [doi: 10.1186/s12913-018-2843-1] [Medline: 29370785]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
CCS: certified coding specialist
CNN: convolutional neural network
DRG: diagnosis related group
EMR: electronic medical record
HAN: hierarchical attention network
ICD: International Classification of Diseases
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
KMUCHH: Kaohsiung Medical University Chung-Ho Memorial Hospital
MDC: major diagnostic category
MIMIC: multiparameter intelligent monitoring in intensive care
NHIA: National Health Insurance Administration
NLP: natural language processing
Tw-DRG: Taiwan Diagnosis Related group

XSL•FO
RenderX

Edited by T Leung, F Xie; submitted 13.03.24; peer-reviewed by PF Chen, J Mistry, DA Southern; comments to author 24.05.24; revised version received 28.06.24; accepted 17.07.24; published 20.09.24.

Please cite as:
Dai HJ, Wang CK, Chen CC, Liou CS, Lu AT, Lai CH, Shain BT, Ke CR, Wang WYC, Mir TH, Simanjuntak M, Kao HY, Tsai MJ, Tseng VS
Evaluating a Natural Language Processing–Driven, AI-Assisted International Classification of Diseases, 10th Revision, Clinical Modification, Coding System for Diagnosis Related Groups in a Real Hospital Environment: Algorithm Development and Validation Study
J Med Internet Res 2024;26:e58278
URL: https://www.jmir.org/2024/1/e58278
doi:10.2196/58278
PMID:

©Hong-Jie Dai, Chen-Kai Wang, Chien-Chang Chen, Chong-Sin Liou, An-Tai Lu, Chia-Hsin Lai, Bo-Tsz Shain, Cheng-Rong Ke, William Yu Chung Wang, Tatheer Hussain Mir, Mutiara Simanjuntak, Hao-Yun Kao, Ming-Ju Tsai, Vincent S Tseng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of a Clinical Decision Support System for Antimicrobial Prescribing in Sub-Saharan Africa: Multisectoral Qualitative Study

Nathan Peiffer-Smadja[1,2,3], MD, PhD; Sophie Descousse[1], MSc; Elsa Courrèges[1], MSc; Audrey Nganbou[1], MD; Pauline Jeanmougin[2], MD; Gabriel Birgand[4], MD, PhD; Séverin Lénaud[5], MSc; Anne-Lise Beaumont[3], MD; Claire Durand[1], MD; Tristan Delory[2,6], MD; Josselin Le Bel[1,2,7], MD; Elisabeth Bouvet[2,3], MD, PhD; Sylvie Lariven[2,3], MD; Eric D'Ortenzio[3,8], MD; Issa Konaté[9], MD; Marielle Karine Bouyou-Akotet[10], MD, PhD; Abdoul-Salam Ouedraogo[11], PhD, PharmD; Gisèle Affoue Kouakou[12], MD, PhD; Armel Poda[11], MsPH, MD; Corinne Akpovo[12], MD; François-Xavier Lescure[1,2,3], MD, PhD; Aristophane Tanon[12], MD, PhD

[1]Université Paris Cité et Université Sorbonne Paris Nord, Inserm, IAME, Paris, France
[2]Antibioclic Steering Committee, Paris, France
[3]Infectious Diseases Department, Bichat-Claude Bernard Hospital, Assistance-Publique Hôpitaux de Paris, Paris, France
[4]CPias, Centre Hospitalo-Universitaire de Nantes, Nantes, France
[5]CHU de Treichville, Centre Hospitalo-universitaire de Treichville, Abidjan, Cote D'Ivoire
[6]Innovation and Clinical Research Unit, Annecy-Genevois Hospital, Epagny-Metz-Tessy, France
[7]Department of General Practice, Université Paris Diderot, Université de Paris, Paris, France
[8]ANRS - Maladies infectieuses émergentes, INSERM, Paris, France
[9]CHU Point G, Bamako, Mali
[10]Université des Sciences de la Santé, Libreville, Gabon
[11]CHU Bobo Dioulasso, Bobo Dioulasso, Burkina Faso
[12]CHU de Treichville, Abidjan, Cote D'Ivoire

Corresponding Author:
Nathan Peiffer-Smadja, MD, PhD
Université Paris Cité et Université Sorbonne Paris Nord
Inserm
IAME
16 rue Henri Huchard
Paris, 75018
France
Phone: 33 140257311
Fax: 33 140258825
Email: nathan.psmadja@gmail.com

Abstract

Background: Suboptimal use of antimicrobials is a driver of antimicrobial resistance in West Africa. Clinical decision support systems (CDSSs) can facilitate access to updated and reliable recommendations.

Objective: This study aimed to assess contextual factors that could facilitate the implementation of a CDSS for antimicrobial prescribing in West Africa and Central Africa and to identify tailored implementation strategies.

Methods: This qualitative study was conducted through 21 semistructured individual interviews via videoconference with health care professionals between September and December 2020. Participants were recruited using purposive sampling in a transnational capacity-building network for hospital preparedness in West Africa. The interview guide included multiple constructs derived from the Consolidated Framework for Implementation Research. Interviews were transcribed, and data were analyzed using thematic analysis.

Results: The panel of participants included health practitioners (12/21, 57%), health actors trained in engineering (2/21, 10%), project managers (3/21, 14%), antimicrobial resistance research experts (2/21, 10%), a clinical microbiologist (1/21, 5%), and

XSL•FO
RenderX

an anthropologist (1/21, 5%). Contextual factors influencing the implementation of eHealth tools existed at the individual, health care system, and national levels. At the individual level, the main challenge was to design a user-centered CDSS adapted to the prescriber's clinical routine and structural constraints. Most of the participants stated that the CDSS should not only target physicians in academic hospitals who can use their network to disseminate the tool but also general practitioners, primary care nurses, midwives, and other health care workers who are the main prescribers of antimicrobials in rural areas of West Africa. The heterogeneity in antimicrobial prescribing training among prescribers was a significant challenge to the use of a common CDSS. At the country level, weak pharmaceutical regulations, the lack of official guidelines for antimicrobial prescribing, limited access to clinical microbiology laboratories, self-medication, and disparity in health care coverage lead to inappropriate antimicrobial use and could limit the implementation and diffusion of CDSS for antimicrobial prescribing. Participants emphasized the importance of building a solid eHealth ecosystem in their countries by establishing academic partnerships, developing physician networks, and involving diverse stakeholders to address challenges. Additional implementation strategies included conducting a local needs assessment, identifying early adopters, promoting network weaving, using implementation advisers, and creating a learning collaborative. Participants noted that a CDSS for antimicrobial prescribing could be a powerful tool for the development and dissemination of official guidelines for infectious diseases in West Africa.

Conclusions: These results suggest that a CDSS for antimicrobial prescribing adapted for nonspecialized prescribers could have a role in improving clinical decisions. They also confirm the relevance of adopting a cross-disciplinary approach with participants from different backgrounds to assess contextual factors, including social, political, and economic determinants.

(J Med Internet Res 2024;26:e45122) doi:10.2196/45122

KEYWORDS

antimicrobial resistance; implementation research; Consolidated Framework for Implementation Research; CDSS; mobile health; mHealth; eHealth; mobile phone

Introduction

Background

Antimicrobial resistance poses a global public health threat that affects both high-income countries and low- and middle-income countries (LMICs). In West African countries, the misuse and overuse of antimicrobials are further exacerbated by contextual factors, such as inappropriate antimicrobial prescribing practices, over-the-counter availability of antimicrobials, and low levels of public awareness. In addition, a specific socioeconomic context marked by disparities in health coverage, poor regulation in the pharmaceutical sector, and health care systems with limited resources also contribute to the misuse of antimicrobials [1]. Deviations from existing guidelines on antimicrobial use, including delays and overuse, are linked to both patient and health care professional behaviors [2,3].

Clinical decision support systems (CDSSs) could address the suboptimal use of antimicrobials by providing physicians with updated clinical practice guidelines and assisting them in prescribing appropriate antimicrobial treatments. For instance, ALMANACH, a CDSS designed for low-resource settings, has been shown to reduce by 80% the volume of antibiotics prescribed to children with acute illnesses in Tanzania [4]. Antibioclic [5], a not-for-profit CDSS developed by French university clinicians for antimicrobial prescribing in primary care, allows clinicians to obtain personalized recommendations for diagnostic and therapeutic management in just a few clicks, tailored to each clinical situation and patient characteristic [6,7].

A preimplementation study was carried out in Burkina Faso to assess the feasibility of adapting a CDSS for antimicrobial prescribing to the West African context [6]. In April 2020, Antibioclic Afrique was adapted from Antibioclic and co-designed to implement COVID-19 guidelines in various West African countries. It was codeveloped by clinicians from Ivory Coast, Gabon, France, Burkina Faso, Senegal, and Mali in partnership with the Antibioclic committee, the African Society for Infectious Diseases (SAPI), and the French Society for Infectious Diseases (SPILF). Antibioclic Afrique [8,9] is a Francophone nonprofit-oriented CDSS that is freely available on the web and as a smartphone app on iOS (Apple Inc) and Android (Google LLC). The recommendations used in the CDSS were derived from national clinical practice guidelines from the 5 partner countries or the World Health Organization (WHO) guidelines if national guidelines were not available and were gradually transformed into decision trees in Antibioclic Afrique. In a more illustrative manner, after the clinicians have made the diagnosis, the tool enables them to access specific pathologies during their consultation, such as "purulent meningitis," and to input clinical elements related to the diagnosis. This input leads to a final page recommending the prescription of a particular type, duration, and dosage of antimicrobial tailored to the patient [7].

The implementation of a CDSS is influenced by the existing digital health environment at the national and global levels. In 2005, the WHO adopted a resolution establishing an eHealth strategy and recognized its role in achieving Millennium Development Goals, particularly in health system strengthening and extending universal health care coverage [10]. The WHO has developed several tools and guidelines to support and evaluate eHealth initiatives and provide technical assistance [11-13].

Concerning eHealth policies among the 5 partner countries, Côte d'Ivoire, Burkina Faso, Mali, and Senegal have developed strategies to develop eHealth and included it in national health development plans [14-16]. Some countries have a robust ecosystem of eHealth stakeholders and have quickly developed eHealth tools, including tools to fight COVID-19. For example,

XSL•FO
RenderX

the Burkina Faso Ministry of Health developed 3 mobile phone apps to detect suspected COVID-19 cases, follow them up, and search for unreported contact cases [17]. Senegal developed the DAANCOVID19 initiative to create and share relevant content related to COVID-19 through social media. Moreover, a special committee is responsible for designing innovative eHealth solutions to combat COVID-19 [18]. Regarding human resources, training in telehealth and medical informatics is gradually being introduced in some partner countries, and a diploma in eHealth has been established in partnership with universities in Mali, Senegal, and Ivory Coast [19]. However, the effective implementation of eHealth initiatives remains at a pioneer stage for many countries, with many development paths underway. Some countries have identified the insufficient financing of eHealth, an insufficient number of qualified human health professionals in ICTs (information and communication technologies), and a lack of eHealth priority setting from the ministry of health as barriers to the effective implementation of eHealth initiatives. Despite a global agenda for action in eHealth, the scope of deployment of eHealth initiatives at the country level, aligned with national health priorities, is still relatively low. According to the WHO, only 6.6% of West African countries have effectively implemented a digital health strategy [14]. While many eHealth projects have been developed in sub-Saharan Africa, few have been scaled up due to poor adherence and low uptake rates.

The existing literature highlights various conceptual frameworks that researchers can use to understand potential contextual factors involved in CDSS implementation and classify them into broad axes or domains. These frameworks include the technology acceptance model, the resource-based theory, and the Consolidated Framework for Implementation Research (CFIR) [20]. Although their approaches differ, they all propose a classification of factors that can be useful and adaptable to each implementation project. Overall, the axes integrate the organization or environment in which the CDSS is implemented, the individual behaviors related to the prescriber or patient, and the CDSS itself and its functionalities. Within these broad axes, there are numerous factors, such as practitioner acceptance of new technology, integration of the CDSS into workflows, and access to technical support [21-23]. These factors guide the analysis of potential barriers to CDSS implementation. Many challenges are cited in the literature, such as the poor acceptance of behavior change, the lack of CDSS integration into clinical workflows, the loss of autonomy experienced by the clinician, or the fragmentation of initiatives [23].

However, there is limited literature that focuses specifically on the contexts of the West African countries, as most studies have centered on projects developed in Europe or North America [14,24]. In addition, there is a lack of preimplementation studies that assess microlevel and macrolevel influences, which could be crucial in strengthening the sustainability of eHealth projects by bridging the gap between the digital solution and the reality on the ground. Given the weight of socioeconomic factors in the misuse of antimicrobials, multidimensional studies in this area are underexplored and urgently needed [25-27]. Cross-cultural research that brings together international research centers is also encouraged to overcome these barriers [24].

Objectives

The main objective of the study was to identify contextual barriers or facilitators directly or indirectly affecting the co-design and implementation of a CDSS for antimicrobial prescribing in 5 sub-Saharan countries (ie, Ivory Coast, Burkina Faso, Mali, Gabon, and Senegal). To anticipate its concrete implementation among health care professionals, the secondary objectives were to analyze prospective implementation strategies, to specify which prescribers and health care structures should be targeted by the CDSS, and to find ways to ensure the sustainability of the CDSS.

Methods

Study Design and Participants

This study adopted a qualitative approach through the use of semistructured interviews. The study is reported in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist (Multimedia Appendix 1) [28]. The objective of the recruitment process was to provide a comprehensive understanding of the challenges faced during the development and implementation phases of the CDSS by ensuring a variability of approaches and perspectives. Purposive sampling targeted stakeholders, health actors, scientists, physicians, biologists, anthropologists, project managers, and funders with expertise in antimicrobial stewardship. To this aim, potential participants were identified through a large transnational capacity-building hospital preparedness network in the context of the operational response to COVID-19 in 5 sub-Saharan African countries: Senegal, Gabon, Ivory Coast, Burkina Faso, and Mali [29]. All professionals participating in the network and having a practical or professional experience in the field of antimicrobial stewardship were contacted by email and invited to participate in the study. A reminder email was sent 2 weeks after the first email. Among the 43 potential participants invited to participate voluntarily in a videoconference interview, 21 (49%) answered positively. The individuals who received an email but did not participate in the study either never responded to the contact email or responded once to the email but did not confirm their participation when an interview date was suggested. None of the participants who were interviewed dropped out of the study. The interviews were conducted in French using the Zoom (Zoom Video Communications, Inc) videoconferencing application between September and December 2020. The choice of semistructured interviews appeared more suitable than a stakeholder roundtable for various reasons. First, there existed potential professional interactions among some interviewees as well as hierarchical relationships. These hierarchical relationships hold significant importance and could have led to self-censorship among participants. Second, besides hierarchical relationships, participants worked in different contexts and organizations that do not always collaborate together, and we think there might have been less freedom of expression regarding health system functioning or field experiences. Third, organizing roundtable

XSL•FO
RenderX

discussions during the COVID-19 pandemic was judged to be more complicated.

Interview Schedule

A multidisciplinary team, including infectious diseases clinicians (NP-S and CA), an implementation sciences methodology expert (GB), and eHealth experts in West Africa (EC, SD, and AN), co-designed the qualitative interview guide (Multimedia Appendix 2) based on constructs derived from the CFIR, such as assessing context and needs, local conditions, partnership and connections, available resources, and implementation facilitators [20]. The interview guide was pretested with 3 participants to verify the formulation and the sufficient degree of openness of the questions, as well as the participants' understanding. All one-to-one interviews were carried out by the same qualitative researcher (EC, a female researcher with a Master of Science degree who was in charge of cross-country coordination for operational and research activities at the time) using a similar approach and began with an introductory presentation of the Antibioclic Afrique CDSS prototype, which was not available for clinicians at the time but already operational and presentable to participants. A simulation demonstrating the use of the tool was presented to them as though they were future users of the tool. This entailed selecting specific decision trees for diseases to ensure they understood how the CDSS could be used. After this presentation, the interviews were aimed at discussing the challenges around antimicrobial prescription, including the barriers to and facilitators of the use of the CDSS, and at identifying tailored implementation strategies.

After identifying the barriers and facilitators according to the CFIR conceptual framework, participants were asked to identify implementation strategies that could help implement Antibioclic Afrique. These strategies were then classified according to the Expert Recommendation for Implementing Change (ERIC) framework, which is a compilation of 73 implementation strategies [30]. These implementation strategies are tailored to address barriers identified with the CFIR framework [31]. The process of linking strategies proposed by the participants to ERIC implementation strategies was continuous and involved discussion and validation during steering committee and staff meetings, as well as identification and debate during individual interviews. On average, the interviews lasted from 45 minutes to 1 hour. Data saturation was collectively discussed after 18 interviews. Participants, while providing specific insights at times depending on their position or location, generally expressed themselves homogeneously on the topics discussed. A total of 3 additional interviews were conducted before closing the interview phase.

Data Analysis

The interviews were transcribed verbatim by a third-party transcription company, and the data were analyzed through a thematic analysis by a researcher (EC). Key categories were identified and developed into themes through coding. Two other authors (SD and AN) independently coded the transcripts, and the 3 researchers collectively reviewed and reached a consensus on the themes through independent coding and group discussion. The research team then reviewed and agreed upon the major and minor themes. No coding software was used. Preliminary findings and the interview guide were sent to participants and discussed to ensure that conclusions were solidified and to avoid bias in data analysis. Participants found that there was coherence between the collected data and the proposed results.

Ethical Considerations

This project was approved by the Ethical Committee of Côte d'Ivoire (Comité National d'Éthique des Sciences de la Vie et de la Santé, reference 068-20/MSHP/CNESVS-kp) in June 2020 and the Ethical Committee of Burkina Faso (Comité Éthique pour la Recherche en Santé au Burkina Faso, ref 2021-07-163) in July 2021. No compensation was offered to participants. The research protocol was approved by the Ethics Committee in Infectiology in France (CER-MIT), and a research declaration was obtained from the CNIL (Commission Nationale de l'Informatique et des Libertés). All participants signed informed consent forms before the interviews and received a preinterview information sheet. No personal identifiable information was recorded, and all data were analyzed anonymously.

Results

Sample Characteristics

Of the 43 professionals who were contacted, a total of 21 (49%) individuals participated in one-to-one interviews. These 21 participants comprised 12 (57%) men and 9 (43%) women who were working in Ivory Coast (n=3, 14%), Burkina Faso (n=3, 14%), Gabon (n=4, 19%), Mali (n=1, 5%), or Senegal (n=1, 5%). Some participants were working in multiple countries in West Africa (9/21, 43%) and were employed by West African organizations or international organizations. One-third (7/21, 33%) of the participants worked in university hospitals, and nearly one-third (6/21, 29%) worked in nongovernmental organizations (NGOs) specializing in health care. The remaining participants worked in public health agencies (5/21, 24%), foundations (2/21, 10%), and dispensaries (1/21, 5%). Most participants (12/21, 57%) were physicians, while others were health actors, such as technical coordinators (2/21, 10%), project managers (3/21, 14%), experts on antimicrobial resistance (AMR) research (2/21, 10%), a clinical microbiologist (1/21, 5%) and an anthropologist (1/21, 5%). Table 1 provides details on the participants' characteristics.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of the participants (N=21)[a].

Characteristics	Values, n (%)
Gender	
Men	12 (57)
Women	9 (43)
Area of practice	
Multiple countries in West Africa	9 (43)
Gabon	4 (19)
Ivory Coast	3 (14)
Burkina Faso	3 (14)
Mali	1 (5)
Senegal	1 (5)
Working structure	
University hospitals	7 (33)
Nongovernmental organizations	6 (29)
Public health agencies	5 (24)
Foundations	2 (10)
Dispensaries	1 (5)
Job title	
Physician	12 (57)
Project manager	3 (14)
AMR[b] expert	2 (10)
Technical coordinator	2 (10)
Clinical microbiologist	1 (5)
Anthropologist	1 (5)

[a]Participants in this qualitative study were antimicrobial stewardship experts selected through a transnational capacity-building hospital preparedness network in Senegal, Gabon, Ivory Coast, Burkina Faso, and Mali. Participants had expertise in antimicrobial stewardship and participated in an interview to identify contextual barriers or facilitators directly or indirectly affecting the co-design and implementation of a clinical decision support system for antimicrobial prescribing in their countries.

[b]AMR: antimicrobial resistance.

Outcomes

CDSS Design and Implementation Challenges

The challenges that were identified have been categorized into three distinct levels in Figure 1: (1) the individual level, where the sustainable adoption of the CDSS relies on behavioral patterns and interactions among the public, health care professionals, and patients; (2) the health system level, where factors such as human resource training and the organization of health facilities must be considered to adapt the CDSS; and (3) the regional and country level, where broader political and economic influences that could either facilitate or hinder the implementation of the CDSS must be evaluated.

XSL•FO
RenderX

Figure 1. Global challenges in the design and implementation of Antibioclic Afrique. CDSS: clinical decision support system.

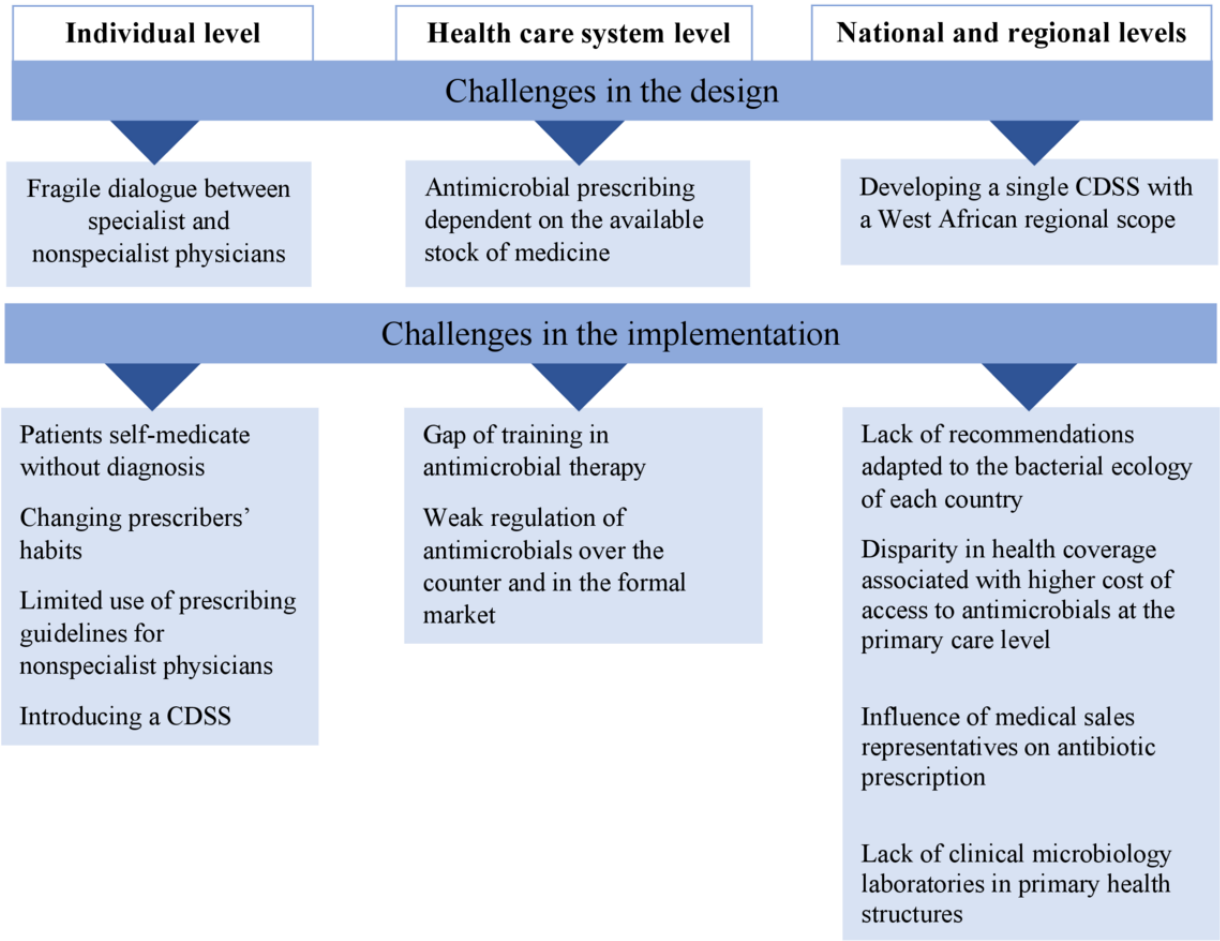

At the individual level, the participants reported deeply ingrained behaviors and habits both among patients and clinicians regarding the use and misuse of antibiotics, which can pose a significant challenge in moving toward changing practices regarding antimicrobials. The main challenges identified by participants were behavioral factors, such as a lack of common stance in antimicrobial prescribing between specialists and nonspecialists, the belief among many patients in the greater efficacy of antibiotics compared to other types of medications, and the difficulty faced by certain generations of doctors in changing or questioning their professional habits.

Talking about antimicrobial prescribing in clinical practice, a participant noted the following:

> *As physician-researchers, we don't have that approach compared to clinicians. Because we are used to seeking guidelines to be sure of what we are doing [with antimicrobials], whereas clinicians, especially the older ones, are not used to looking for references. And I think the problem would not only come from the patients, the sick, but even from the caregivers who could actually not use it [the CDSS] ... Just because it's part of our daily lives doesn't mean everyone knows how to prescribe an antibiotic today. People need to accept that they don't know how to prescribe an antibiotic, that it follows rules, and that these rules are subject to change.* [Infectious disease specialist]

Talking about the importance of patients' beliefs associated with the supposed effectiveness of certain treatments such as antibiotics, a participant said the following:

> *The only thing that will be complicated is the "no antibiotics" approach. Not prescribing antibiotics for the patient should not simply mean "no treatment." There must be a prescription for a medication. For example, in Casamance where we intervene, traditional medicine is still very much present. We need to consider whether the alternative could be prescribing plants or other remedies, but I believe that patients should leave the consultation with something prescribed. And it's so strong to the point that, for example, there is a huge placebo effect of the needle, of the infusion. Meaning, you give the same medication, you know perfectly well that the effectiveness is the same, whether it's intravenous or oral, but in this region, for patients, the symbolism of the needle in the infusion, of the injection, is extremely powerful. It's the same with antibiotics.* [NGO project manager]

XSL•FO
RenderX

The participants also identified challenges at the level of the health care system itself, such as the lack of consensus between infectious disease specialists and nonspecialist physicians and the fluctuating availability and cost of certain molecules. The weak regulation of antimicrobials over the counter was also reported to be a variable and may lead to prescribers switching from one class of antimicrobials to another or even not using the CDSS at all.

Talking about the clinicians, a participant expressed the following:

> *The limiting factor that I see is that if what is recommended is not a molecule available in the central hospital pharmacy or in a pharmacy, or a molecule that is easily accessible, they will not actually prescribe it. And they'll stick to their current practice of saying, "I've got another antibiotic on the side."* [NGO project manager]

Talking about the specificity of molecules and its impact on CDSS use by clinicians, a participant said the following:

> *If your recommendations suggest molecules that are not readily available, they [the prescribers] will actually turn away the second, third, fourth time they are recommended. If it's not a molecule they can order, they will give up.*

In addition to the variability in the availability of antimicrobials, these drugs are also widely available over the counter in pharmacies, allowing patients to self-prescribe without consultation or diagnosis. Some participants noted that the COVID-19 pandemic has exacerbated this issue, with an increase in informal and unregulated suppliers of antimicrobials. This could potentially undermine the impact of CDSSs for antimicrobial prescribing:

> *The second major prescriber in these regions is the pharmacist because what happens for the patient is that instead of paying for a consultation and then going to the pharmacy, they go directly to the pharmacy. And the pharmacist prescribes, I won't say for every other patient, but it's significant. And from what I've seen anyway, because it could be the pharmacist or the pharmacist's assistant. People come to the pharmacy to get their prescription.* [NGO project manager]

> *There is also a problem of availability and even cost. There is no Social Security, so it's the patients who pay. A very illustrative example: when you take an antibiotic like Ceftriaxone which should be the second- or third-line antibiotic, it costs less than one euro compared to the antibiotic like amoxicillin which should be the first line. These are quite a few reasons that can impact on the quality of antibiotic prescribing.* [Infectious disease specialist]

At the national level, participants identified a key challenge: the absence of national guidelines for prescribing antimicrobials, coupled with a lack of understanding of the local bacterial ecology. They also highlighted the influential role of sales representatives in antimicrobial prescribing:

> *One of the barriers we see in our context, due to the absence of human resources, is that sometimes there are no recommendations for the prescription of certain antibiotics.* [Infectious disease specialist]

> *If you see someone using recommendations for antibiotic treatment, they are French recommendations. For someone who reads a lot, they will seek out American or British recommendations or those from other countries. They will try to compare all recommendations and then draw something from them, in a personal and individual way. So, we don't have data that are built on the basis of our local ecology. So, probabilistically, in Europe, yes, we could treat a patient arriving at the emergency room with a third-generation cephalosporin as a first-line treatment, because in Europe, the resistance level of E. coli to third-generation cephalosporins is less than 20%. But here, I'm not sure that's the case. That's the challenge.* [General practitioner]

> *The profile of fevers, both in outpatient and inpatient settings, has changed dramatically in West Africa. I assure you that bacterial infections do not even represent 30% of fevers in hospitals. And even less in outpatient clinics. But antibiotic prescribing, I think it's over 50% to 60% of prescriptions that include at least one antibiotic.* [Infectious disease specialist]

Regarding the influence of the private sector in prescribing, a participant said the following:

> *It is much more the medical representatives who greatly influence the prescriptions of doctors, especially non-infectious disease specialists. They are really influenced by medical representatives, and that is a big problem.* [General practitioner]

The participants also emphasized the challenge of motivating countries to take further action at the national level for better regulation of prescriptions.

Participants suggested that an eHealth tool could serve as a catalyst for change by incentivizing priority setting in health strengthening policies and promoting regional governance to combat AMR:

> *Antibioclic Afrique could be a springboard for SAPI [African Society for Infectious Diseases] to initiate a reform on algorithms, protocols and recommendations concerning antibiotic therapy through the appropriation of the tool by different actors. SAPI has a major role to play because the majority of French-speaking African countries are represented in the organization. We must not stop only at our scientific level but go precisely to the level of decision-makers, politicians, regulatory authorities to see how to try to change attitudes and to formalize the algorithms, the recommendations for the management of infectious diseases, the different protocols and ensure that they are applied at the national level by each of the countries.* [Infectious disease specialist]

XSL•FO
RenderX

Identifying Potential Future Users of the CDSS

After discussing the general challenges in implementing a CDSS, the participants addressed the issue of its target users. They recognized the need to analyze which health care professionals would benefit the most from the CDSS to support its dissemination and adoption. Figure 2 depicts the barriers and facilitators that need to be considered in CDSS implementation at different levels of care. The organization of health care has been represented as a pyramid of care, with each level corresponding to the number of patients cared for at each level: primary, secondary, and tertiary care. Each level is associated with specific types of physicians (general practitioners or specialized physicians), paramedics (midwives or nurses), or health workers and specific types of health structures [32].

Figure 2. Barriers and facilitators at the different levels of the health pyramid. CDSS: clinical decision support system.

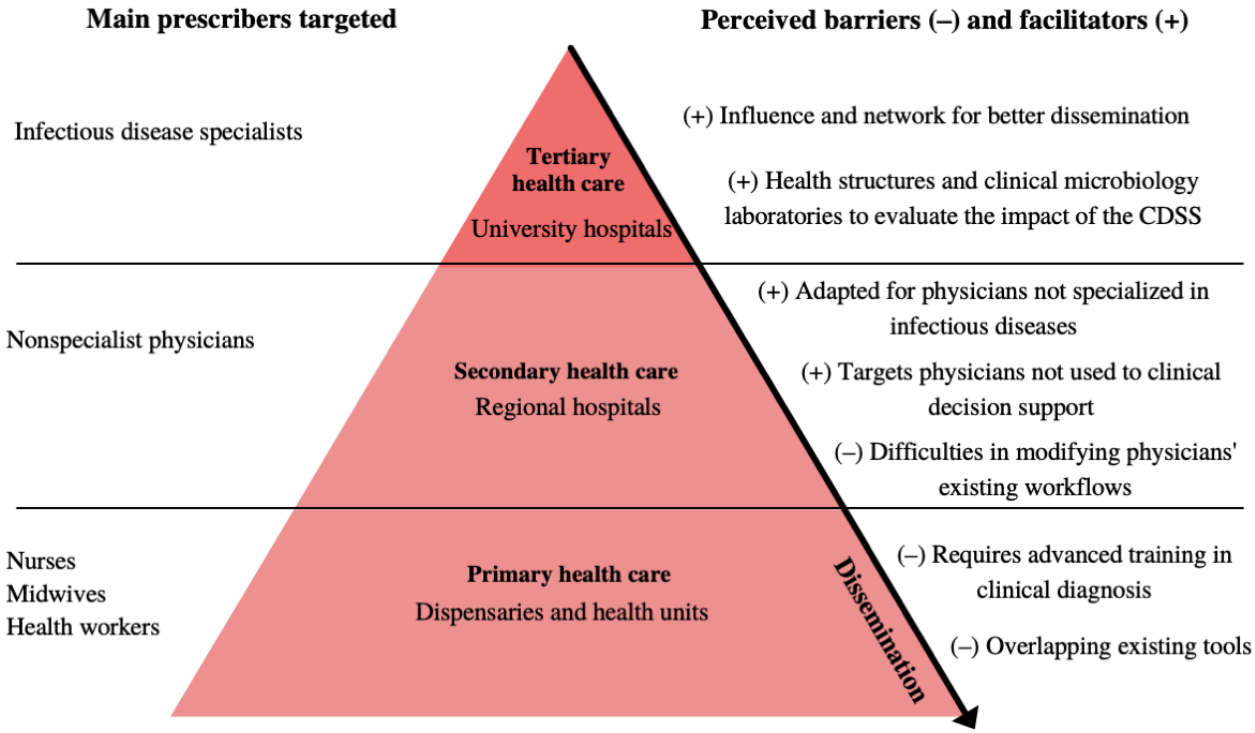

Participants discussed the issue of CDSS target users and shared the assessment that midwives, nurses, and community health workers are key prescribers at the primary health care level:

> *In Mali, whether officially or unofficially, nurses are sometimes allowed to prescribe medication. ... There are health centers run solely by nurses or often by nursing assistants, which are in distant areas, not close to the city centre.* [General practitioner 1]

However, most of them stated that the current CDSS architecture was not adapted to their basic training in antimicrobial therapy and clinical diagnosis. The participants identified a major barrier in that a single tool could not be adapted to different types of prescribers with varying levels of training in antimicrobial therapy:

> *The nurse is trained fundamentally, it's true, to provide primary care, but they are trained to carry out treatments. The midwife as well. The doctor is trained to analyse a clinical situation and eventually make a decision; they have a diagnostic approach, so the practice is not the same. The tool is great, but it's essential in its design to consider the user category. It's easy to say when we're among ourselves as doctors, but with nurses, with midwives, it needs to be explained well. So, if we're going to use the tool and make it available to them, it needs to be done from basic training so that they understand the logic of the exercise.* [Infectious disease specialist 3]

> *The way the tool is presented, I think it's more addressed to doctors than to nurses. Nurses in their everyday practice, working with doctors, learn certain things; they start prescribing because they've seen doctors do it that way, so they do it that way too.* [General practitioner 2]

According to participants, the CDSS would be more appropriate for general practitioners, especially in health centers where there are no antimicrobial therapy specialists:

> *At the prescriber's profile level, the goal is also to enable individuals who are not specialists in infectious diseases to have sufficient answers, and to have a guide that allows them to recommend or not recommend antibiotics. ... There are antibiotics that normally should not be prescribed by general practitioners, especially imipenem, and it is prescribed haphazardly for any infection without considering antibiotic susceptibility testing. Without any thought, they are prescribed. ... At general hospitals, the likelihood of finding specialists is low; there are mostly general practitioners. There are very*

XSL•FO
RenderX

> *few general hospitals where infectious diseases are not encountered. Even the gynaecologist who prescribes an antibiotic behaves like a general practitioner in terms of antibiotic therapy.* [Infectious disease specialist 3]

Regarding facilitators of CDSS implementation, participants indicated that the initial targets should be nonspecialist physicians in the secondary and tertiary levels of care, where the CDSS could be monitored and better used. Participants also suggested disseminating the tool, starting at the tertiary and secondary levels, so that specialist and nonspecialist physicians can introduce it to their peers in the primary sector and possibly to paramedics. In addition, infectious diseases specialists have an extensive network of connections with public health agencies and health establishments, which can help disseminate the tool:

> *We know that at the intermediate level, in particular at the level of general hospitals, there is a minimum of technical equipment that already exists, which can be a basis for guiding the decision, unlike at the primary level in rural dispensaries or in health facilities. Because behind a therapeutic decision lies a diagnostic presumption, and even after the treatment has been administered, especially regarding antibiotic therapy, there are monitoring criteria to evaluate effectiveness.* [Infectious disease specialist 3]

> *The physician, clinician, researcher, they are always seeking information. ... They will have arguments to encourage their colleagues to adopt the tool. It is necessary to identify healthcare personnel, guide them, train them, whether through webinars or tutorials, and ensure that these trained personnel are willing to transmit the information or knowledge to others, to train other individuals, so that it can be done in cascade.* [General practitioner 1]

> *I really think it should be the district chief physician who trains all the district nurses in a single session.* [NGO project manager]

They also confirmed the need for the CDSS to mitigate antimicrobial misuse in several departments of teaching and regional hospitals:

> *I was on a test panel last Friday, it was a pilot survey on the use of antibiotics in hospitals, and what came out was that in more than 90% of the departments that were surveyed, an antibiotic had been prescribed, and the main prescribers were in the surgical departments. This to say that this CDSS would come at the right time, because we handle antibiotics, but are we using them correctly?* [Infectious disease specialist 3]

However, despite these significant challenges, participants were convinced that targeting paramedics as potential regular users of the CDSS is necessary. This will require efforts to adapt the tool and additional implementation work with nurses and midwives, in particular. At this stage, nonspecialist physicians were identified as the primary targets for the CDSS.

Implementation Strategies for the CDSS

Given the difficulty in changing primary care prescribers' routines and the different levels of knowledge in antimicrobial therapy, academic partnerships were cited as one of the most sustainable implementation strategies. A participant suggested that the CDSS could be disseminated through the interuniversity diploma dedicated to antimicrobial therapy [33]:

> *We already have an Internet site at the level of the University Diploma and we are very willing to disseminate Antibioclic Afrique by popularizing it, already with 150 University Diploma trainees who will each be able to present it to their community.* [Infectious disease specialist 2]

> *Even experts contribute to over-prescribing. Antibioclic Afrique will allow us to give them a way of learning, updating existing knowledge and even of applying what they have learned, that is more accessible, according to their context and the context of Android phones, the Internet, etc. Young people like that, so they take it up more easily.* [Infectious disease specialist 1]

Some participants suggested that creating internal and external networks that connect nonspecialist physicians with antimicrobial therapy experts could be helpful in addressing complex clinical cases and promoting collaborative learning. The dissemination through scientific channels, such as gatherings of the medical community, is also an interesting avenue:

> *I think that Antibioclic Afrique can really be an opportunity to develop a network of exchanges. We have specialists. We are lucky in Burkina Faso, we have training in antibiotic therapy, practically all the structures have doctors who have been trained here.* [Infectious disease specialist 4]

Talking about an annual health congress organized in Mali, a participant noted the following:

> *Utilizing each slot to present the tool is a plus; there are many more specialist doctors, general practitioners, and professors who can also support the transmission of tool usage. They come from all over Mali to attend the conference, so it really allows for a wide dissemination of the message.* [General practitioner 1]

The participants suggested another measure to support the sustainability of the CDSS, which was to organize continuing education sessions on antimicrobial therapy for health care personnel. Such sessions would not only update their knowledge but also increase the awareness of the risks of AMR for patients and the health care system. These actions would complement the dissemination of the CDSS to potential users and help ensure its long-term viability. Table 2 summarizes the implementation strategies that could be used to promote the adoption of the CDSS according to ERIC categories.

XSL•FO
RenderX

Table 2. Identification and classification of implementation strategies suggested during the interviews[a].

ERIC[b] strategies	Corresponding specific implementation strategies for the CDSS[c]
Promote network weaving	Develop intrahospital and external networks connecting nonspecialist physicians with trained physicians in antimicrobial prescription
Create a learning collaborative	Identify and train key physicians to develop a network of ambassadors for the CDSS dissemination
Identify early adopters	At the initial phase of implementation, prioritize the intermediary and central levels (university hospitals, general hospitals, and private structures)
Work with educational institutions	Presentation of the CDSS during academic courses for sixth year medical students
Develop academic partnerships	Possible partnership with training degree specialized in antimicrobial prescription
Use an implementation adviser	Experts in implementation sciences involved in the methodology for the CDSS implementation
Use advisory boards and workgroups	Steering committee with representatives of the 5 countries involved
Use advisory boards and workgroups	A total of 21 interviews conducted with health actors to identify implementation strategies of the CDSS
Start a dissemination organization and conduct a local needs assessment	Begin with a pilot site in Ivory Coast, which is responsible for conducting a local needs assessment and disseminating the clinical innovation

[a]Specific implementation strategies for clinical decision support systems (CDSSs) for antimicrobial prescribing in West Africa were identified through interviews with antimicrobial stewardship experts (N=21) and categorized according to the ERIC framework.

[b]ERIC: Expert Recommendations for Implementing Change.

[c]CDSS: clinical decision support system.

Discussion

Principal Findings

Participants highlighted the importance of the CDSS for both nonspecialist physicians and primary health care prescribers. They stressed the need for a CDSS that guides nonexpert physicians, given the inappropriate and unregulated antimicrobial prescribing practices observed in various health care settings. Participants also noted that hospitals' technical platforms could facilitate better monitoring and impact assessment of the CDSS. However, they acknowledged the difficulty in implementing a CDSS that is usable by all primary care prescribers, given their varying levels of training in antimicrobial prescribing.

The main barriers to CDSS implementation cited by the participants were behavioral factors, such as the resistance to changing the physician's workflow, patients' self-medication practices, the lack of regulation in antimicrobial dispensing in pharmacies, and the absence of guidelines. To overcome these barriers, participants proposed implementation strategies, such as promoting network weaving, creating a learning collaborative, and identifying early adopters of the CDSS. In addition, they recommended organizing continuing education sessions on antimicrobial therapy for health care personnel to increase the awareness of the risk of AMR and to support the sustainability of the CDSS's dissemination.

Comparison With Prior Work

A study conducted in Burkina Faso in 2019 [6] identified the main characteristics required for a CDSS for antimicrobial prescribing, as well as the barriers to and facilitators of its implementation. The study confirmed the importance of a CDSS at the primary care level that is tailored to the local epidemiology of infectious diseases, is available through a smartphone, and includes an offline mode. Co-designing the CDSS with diverse stakeholders was identified as a key strategy to ensure sustainable adoption. Another study found that mobile health (mHealth), a component of eHealth referring to the use of mobile communication technologies to promote health by supporting health care practices, was more accepted and adapted to existing infrastructure and local context due to the rising number of mobile subscriptions in Africa [34].

Beyond a context favorable to the emergence of mHealth tools in LMICs and in Africa more specifically, contextual factors often remain underestimated in the implementation process and need to be identified to optimize successful uptake. Preimplementation studies and holistic evaluations are needed to assess the readiness for eHealth in LMICs, which are still lacking and contribute to the limited uptake of effective clinical innovations [35,36]. Questioning the design of a CDSS necessarily involves understanding the specific context in which it is implemented. Although this study focuses on factors related to the social, economic, and political contexts, such as the level of government involvement in health coverage and health programs, regardless of the country or continent concerned, it is important not to forget other factors. The CFIR conceptual framework, among other conceptual frameworks, offers a useful classification to identify the factors that should be considered to foster the adoption of CDSSs. Factors related to the behavior of patients and health professionals; the functionalities offered by the CDSS; and the specific environment in which it is developed, including the type of infrastructure and level of

XSL•FO
RenderX

technology provided, are key elements to evaluate or include in CDSS design [22,23].

In 2014, the Integrated e-Diagnostic Approach (IeDA) was developed and scaled up in Burkina Faso, which is a CDSS for the management of childhood illness at the primary care level. In the preimplementation study of IeDA, participants pointed out that the CDSS must include medicines that are available in the given country and the health care setting. This is necessary to reduce the risk of dropout [37]. In this same project, a multipronged evaluation identified contextual factors that may have been underestimated and limited the implementation of the CDSS. These factors include self-medication, the lack of official guidelines, and difficult access to prescribers [38]. Similarly, ePOCT+ is a clinical decision support system recently developed in Tanzania, Senegal, Rwanda, Kenya, and India for the care of pediatric outpatients in LMIC settings. Its aim is to reduce antimicrobial prescribing. However, 2 limitations of the system are the lack of national guidelines to adapt the algorithms to the local microbial ecology and the difficulty of adherence to the guidelines by practitioners [39]. In Mali, a team is currently developing another CDSS called the Diarrheal Etiology Prediction app. It is available on mobile phones and helps clinicians identify viral childhood diarrhea to reduce antimicrobial prescribing. Preimplementation studies have shown that it is crucial to assess clinicians' adherence to new technology. This is important to anticipate their hesitation, understand their expectations, and thus promote the sustainability of a CDSS [40]. As existing literature shows, many CDSSs developed in sub-Saharan Africa are addressed to pediatric patients. Conversely, adults, pregnant women, and older adults are underrepresented in algorithms for infectious diseases.

Identifying the health care professionals who will benefit the most from the tool was a crucial aspect of the future CDSS. The results confirmed that, given the gap in training on antimicrobial therapy among health care professionals, the CDSS is more relevant for nonspecialized physicians who mostly work in secondary and tertiary health care facilities. Developing a CDSS in LMICs dedicated to primary health care would require complementary advanced training modules in clinical diagnosis and antimicrobial therapy. Another study conducted on different continents confirmed that financial constraints influence the uneven presence of formally trained practitioners and informally trained practitioners [41]. With regard to behavioral interventions, the same study showed that multifaceted interventions are more effective than educational interventions alone in improving antimicrobial prescribing [41]. Multifaceted interventions that include bundle interventions, regulation enforcement, face-to-face education, and peer influence have the potential to address the behavioral factors cited by the participants.

Some participants who specialize in project management have already developed eHealth tools and shared their best practices and pitfalls to avoid for designing and implementing a CDSS. Other participants shared their global vision of the governance landscape in eHealth and AMR. Regarding the identification of implementation strategies, most of them emphasized the importance of building a solid ecosystem that involves diverse stakeholders from different backgrounds, such as hospital directors, policy makers, associations of health professionals, and the national clinical referent for epidemic management. In the context of the IeDA, a dedicated study recommended designing a stakeholder analysis to tailor the approach and involve each actor in the project [42]. Partnerships with the Health Ministry, health managers, NGOs, and social entrepreneurs were contributing factors to scaling the project. Multisectoral coordination and follow-up, along with incentives for stakeholders, could also be a means to enhance regulatory enforcement for antimicrobial prescription [43,44].

Limitations

First, most participants in the study were clinicians working at university hospitals. As highlighted by the participants themselves, it would have been interesting to also include nurses, midwives, health workers, and pharmacists who prescribe antimicrobials in primary care in West Africa to provide a broader overview of antimicrobial prescribing. On the one hand, this is explained by the fact that the participants were identified through a network focusing on hospital preparedness. However, these physicians had extensive experience as practitioners, were involved in research projects, and had a global vision on antimicrobial prescribing. On the other hand, although other health care professionals may represent antimicrobial prescribers in primary care in West Africa, the CDSS and its architecture were initially designed for physicians, as the CDSS requires the selection of a specific diagnosis before providing guidance. This explains why recruitment mainly focused on physicians rather than other health care professionals during the preimplementation phase. The CDSS is intended to evolve, and the inclusion of other health care professionals has been considered in the design of future studies. Second, some factors were examined only in 1 country; therefore, their generalizability to the entire western and central African region may be limited. Third, the role of a CDSS for antimicrobial prescribing will highly rely on the existence of local and national antimicrobial prescribing guidelines in West Africa. Indeed, CDSSs such as Antibioclic Afrique may provide targeted recommendations only when guidelines are available. However, the development of guidelines in West Africa is a significant challenge. Fourth, this study may not be generalizable to other CDSSs concerning other diseases or medical specialties. Even though there are some challenges (eg, fragile dialogue between specialist and nonspecialist physicians) and perceived barriers and facilitators (eg, difficulties in modifying physicians' existing workflows) that might be applicable to the implementation of CDSSs in other medical specialties, this hypothesis would need confirmation studies. Fifth, the thematic analysis of the interviews was conducted by a pluridisciplinary group of qualitative researchers, scientists, and physicians, some of whom were involved in the development of the academic CDSS for antimicrobial prescribing. This might lead to a specific interpretation of the interview data, and their involvement in the design of the CDSS may influence their perception of the obstacles and implementation strategies examined. However, measures have been taken to minimize any potential bias, including ensuring rigorous data analysis and encouraging open and critical discussion among research team members and

XSL•FO
RenderX

interview participants. Sixth, this study only examined the use of an expert system to support clinical decision-making based on the guidelines established by the country in which it is used. Evaluating the design and implementation of a machine learning CDSS would require the consideration of other contextual factors, including database management, automatic data entry, and interoperability, which may present significant challenges in LMICs.

Conclusions

The inappropriate prescribing of antimicrobials and the need for updated knowledge in antimicrobial therapy have been identified as issues that need to be addressed. A CDSS for antimicrobial prescribing might interest West African prescribers. However, significant challenges exist in this context, such as the gap in training for antimicrobial prescribing, antimicrobial prescribers with diverse profiles and levels of expertise, the limited availability of national guidelines, and antimicrobial shortages. A potential implementation strategy for a CDSS for antimicrobial prescribing could be to target physicians in academic hospitals who could then use their network to further disseminate the tool in secondary and primary care. Multilevel partnerships with the government, universities, and medical corporations were identified as crucial means to ensure the sustainable implementation of the CDSS. Further work is needed to explore different contextual factors and settings while addressing the issue of designing one tool for different prescribers, including general practitioners, primary care nurses, midwives, and other health care workers. An upcoming study focusing on user feedback in pilot sites will deepen these results with a focused approach on primary level actors.

Acknowledgments

The authors would like to thank the Antibioclic committee, the APHRO-CoV (support for hospital preparedness in the operational response to COVID-19) network, African Society for Infectious Diseases (SAPI), and the members of the French Society for Infectious Diseases (SPILF) for their participation in the study. Antibioclic Afrique is publicly financed by Agence Nationale Recherche Sida (ANRS COV003), an autonomous agency from Inserm in charge of evaluating, coordinating, and financing research on infectious and emerging diseases.

The template of Antibioclic was freely provided to co-design its West African adaptation, Antibioclic Afrique. No generative artificial intelligence tools were used to write this paper.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. Full transcripts of the interview are available upon request (in French).

Authors' Contributions

NP-S, SD, EC, and FXL contributed to conceptualization. EC, SD, and AN contributed to data curation. FXL, AT, and NP-S contributed to funding acquisition. NP-S, SD, and EC contributed to redaction. EC, SD, NP-S, and AN contributed to formal analysis. EC contributed to investigation. GB, EC, and NP-S contributed to methodology. FXL, AT, and NP-S contributed to project administration. NP-S and CA contributed to supervision. SD contributed to visualization. SD and EC contributed to writing the original draft. SD, NP-S, and ALB contributed to reviewing and editing. CD, JLB, TD, IK, ASO, MB-A, AP, and PJ contributed to validation. EO, EB, and S Lariven contributed to resource acquisition. S Lénaud contributed to software use.

Conflicts of Interest

Some co-authors of this study, namely NP-S, AT, FXL, and GK, participated in the development of the free academic Antibioclic Afrique app, which is the subject of the qualitative evaluation conducted in this work. These authors are employees of the public health sector in their countries, and none of them have a financial conflict of interest with Antibioclic Afrique, which is a non–profit-oriented decision support system. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist.
[PDF File (Adobe PDF File), 752 KB - jmir_v26i1e45122_app1.pdf]

Multimedia Appendix 2
Interview guide.
[DOCX File , 13 KB - jmir_v26i1e45122_app2.docx]

References

XSL•FO
RenderX

1. Ouedraogo AS, Jean Pierre H, Bañuls AL, Ouédraogo R, Godreuil S. Emergence and spread of antibiotic resistance in West Africa: contributing factors and threat assessment. Med Sante Trop 2017 Jun 01;27(2):147-154 [FREE Full text] [doi: 10.1684/mst.2017.0678] [Medline: 28655675]
2. Carrara E, Pfeffer I, Zusman O, Leibovici L, Paul M. Determinants of inappropriate empirical antibiotic treatment: systematic review and meta-analysis. Int J Antimicrob Agents 2018 Apr;51(4):548-553. [doi: 10.1016/j.ijantimicag.2017.12.013] [Medline: 29277528]
3. Limmathurotsakul D, Sandoe JA, Barrett DC, Corley M, Hsu LY, Mendelson M, et al. 'Antibiotic footprint' as a communication tool to aid reduction of antibiotic consumption. J Antimicrob Chemother 2019 Aug 01;74(8):2122-2127 [FREE Full text] [doi: 10.1093/jac/dkz185] [Medline: 31074489]
4. Shao AF, Rambaud-Althaus C, Samaka J, Faustine AF, Perri-Moore S, Swai N, et al. New algorithm for managing childhood illness using mobile technology (ALMANACH): a controlled non-inferiority study on clinical outcome and antibiotic use in Tanzania. PLoS One 2015 Jul 10;10(7):e0132316 [FREE Full text] [doi: 10.1371/journal.pone.0132316] [Medline: 26161535]
5. Antibioclic. Collège de la Médecine Générale, Société de Pathologie Infectieuse de Langue Francaise, Collège National des Généralistes Enseignants. URL: https://antibioclic.com [accessed 2024-04-29]
6. Peiffer-Smadja N, Poda A, Ouedraogo AS, Guiard-Schmid J, Delory T, Le Bel J, et al. Paving the way for the implementation of a decision support system for antibiotic prescribing in primary care in West Africa: preimplementation and co-design workshop with physicians. J Med Internet Res 2020 Jul 20;22(7):e17940 [FREE Full text] [doi: 10.2196/17940] [Medline: 32442155]
7. Delory T, Jeanmougin P, Lariven S, Aubert J, Peiffer-Smadja N, Boëlle PY, et al. A computerized decision support system (CDSS) for antibiotic prescription in primary care-Antibioclic: implementation, adoption and sustainable use in the era of extended antimicrobial resistance. J Antimicrob Chemother 2020 Aug 01;75(8):2353-2362. [doi: 10.1093/jac/dkaa167] [Medline: 32357226]
8. Antibioclic Afrique. Université Paris Diderot Department of General Medicine. URL: https://antibioclic-afrique.com [accessed 2024-04-29]
9. Home page. Antibioclic Afrique. URL: https://antibioclic-afrique.com/ [accessed 2024-08-06]
10. Al-Shorbaji N. The World Health Assembly resolutions on eHealth: eHealth in support of universal health coverage. Methods Inf Med 2013;52(6):463-466. [Medline: 24310395]
11. The MAPS toolkit: mHealth assessment and planning for scale. World Health Organization. 2015. URL: https://apps.who.int/iris/handle/10665/185238 [accessed 2024-04-29]
12. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/252183 [accessed 2024-04-29]
13. Classification of digital health interventions?: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf%3Bjsessionid%3DE865DD370BD42DB4B6BBCB2E246C4C9E?sequence=1 [accessed 2024-04-29]
14. White paper for the 1st digital health meeting in West Africa. AMREF Health Africa. URL: https://amref.fr/wp-content/uploads/2019/03/Amref-Health-Africa-White-Paper-DHM-2018.pdf [accessed 2024-04-29]
15. National health information system and 2010 - 2020. Government of, Burkina Fasos, Ministry of Health. URL: https://static1.squarespace.com/static/548487dce4b08bf981fe60d5/t/60354eef1ec0a67d735b4a1f/1614106365430/BurkinaFaso_Report_Final.pdf [accessed 2024-04-29]
16. e-Gabon. The World Bank. URL: https://projects.worldbank.org/en/projects-operations/project-detail/P132824 [accessed 2024-04-29]
17. COVID-19 au Burkina Faso: trois applications mobiles lancées. Terre des hommes. URL: https://www.tdh.ch/fr/revue-presse/covid-19-burkina-faso-trois-applications-mobiles-lancees [accessed 2024-04-29]
18. #DAAN COVID-19?: la riposte digitale. Conseil National du Patronat du Sénégal. 2020. URL: https://www.ioe-emp.org/index.php?eID=dumpFile&t=f&f=147293&token=3897456da266cc6255458475e67d1fd00adc6bf3 [accessed 2024-04-29]
19. Création d'un Diplôme en E-santé avec 3 universités africaines. Fondation Pierre Fabre. 2019. URL: https://www.fondationpierrefabre.org/fr/carte-fabre/creation-dun-diplome-en-e-sante-avec-3-universites-africaines/ [accessed 2024-04-29]
20. Consolidated framework for implementation research constructs. CFIR Research Team-Center for Clinical Management Research. 2022. URL: https://cfirguide.org/wp-content/uploads/2019/08/cfirconstructs.pdf [accessed 2024-04-29]
21. Shi Y, Amill-Rosario A, Rudin RS, Fischer SH, Shekelle P, Scanlon DP, et al. Barriers to using clinical decision support in ambulatory care: do clinics in health systems fare better? J Am Med Inform Assoc 2021 Jul 30;28(8):1667-1675 [FREE Full text] [doi: 10.1093/jamia/ocab064] [Medline: 33895828]
22. Moxey A, Robertson J, Newby D, Hains I, Williamson M, Pearson SA. Computerized clinical decision support for prescribing: provision does not guarantee uptake. J Am Med Inform Assoc 2010;17(1):25-33 [FREE Full text] [doi: 10.1197/jamia.M3170] [Medline: 20064798]
23. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]

XSL•FO
RenderX

24. Adeloye D, Adigun T, Misra S, Omoregbe N. Assessing the coverage of e-health services in Sub-Saharan Africa. Methods Inf Med 2018 Jan 24;56(03):189-199. [doi: 10.3414/me16-05-0012]
25. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
26. Kilsdonk E, Peute LW, Jaspers MW. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform 2017 Feb;98:56-64. [doi: 10.1016/j.ijmedinf.2016.12.001] [Medline: 28034413]
27. Tilahun B, Smillie K, Bardosh KL, Murray M, Fitzgerald M, Cook V, et al. Identifying barriers and facilitators of 13 mHealth projects in North America and Africa: protocol for a 5-year implementation science study. JMIR Res Protoc 2018 Jul 03;7(7):e162 [FREE Full text] [doi: 10.2196/resprot.9633] [Medline: 29970360]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Appui à la préparation des hôpitaux dans la réponse opérationnelle face à la COVID-19 - Burkina Faso. APRHO-CoV. URL: https://aphro-cov.com/ [accessed 2024-04-29]
30. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
31. Waltz TJ, Powell BJ, Fernández ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. Implement Sci 2019 Apr 29;14(1):42 [FREE Full text] [doi: 10.1186/s13012-019-0892-4] [Medline: 31036028]
32. Frieden TR. A framework for public health action: the health impact pyramid. Am J Public Health 2010 Apr;100(4):590-595. [doi: 10.2105/ajph.2009.185652]
33. Diplôme inter universitaire: antibiologie et antibiothérapie en Afrique Sub-saharienne. Université Nazi Boni. URL: https://www.diu-antibio.org/ [accessed 2024-04-29]
34. Aranda-Jan CB, Mohutsiwa-Dibe N, Loukanova S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC Public Health 2014 Feb 21;14:188 [FREE Full text] [doi: 10.1186/1471-2458-14-188] [Medline: 24555733]
35. Kiberu VM, Mars M, Scott RE. Barriers and opportunities to implementation of sustainable e-Health programmes in Uganda: a literature review. Afr J Prim Health Care Fam Med 2017 May 29;9(1):e1-10 [FREE Full text] [doi: 10.4102/phcfm.v9i1.1277] [Medline: 28582996]
36. Bakibinga P, Kamande E, Kisia L, Omuya M, Matanda DJ, Kyobutungi C. Challenges and prospects for implementation of community health volunteers' digital health solutions in Kenya: a qualitative study. BMC Health Serv Res 2020 Sep 21;20(1):888 [FREE Full text] [doi: 10.1186/s12913-020-05711-7] [Medline: 32957958]
37. Blanchet K, Lewis JJ, Pozo-Martin F, Satouro A, Somda S, Ilboudo P, et al. A mixed methods protocol to evaluate the effect and cost-effectiveness of an integrated electronic diagnosis approach (IeDA) for the management of childhood illnesses at primary health facilities in Burkina Faso. Implement Sci 2016 Aug 04;11(1):111 [FREE Full text] [doi: 10.1186/s13012-016-0476-5] [Medline: 27488566]
38. Sarrassat S, Lewis JJ, Some AS, Somda S, Cousens S, Blanchet K. An Integrated eDiagnosis Approach (IeDA) versus standard IMCI for assessing and managing childhood illness in Burkina Faso: a stepped-wedge cluster randomised trial. BMC Health Serv Res 2021 Apr 16;21(1):354 [FREE Full text] [doi: 10.1186/s12913-021-06317-3] [Medline: 33863326]
39. Tan R, Cobuccio L, Beynon F, Levine GA, Vaezipour N, Luwanda LB, et al. ePOCT+ and the medAL-suite: development of an electronic clinical decision support algorithm and digital platform for pediatric outpatients in low- and middle-income countries. PLOS Digit Health 2023 Jan 19;2(1):e0000170 [FREE Full text] [doi: 10.1371/journal.pdig.0000170] [Medline: 36812607]
40. Keita AM, Brintz BJ, Khan AI, Taufiqul Islam M, Khan ZH, Keita Y, et al. Introduction of an electronic clinical decision support tool to inform prescribing for pediatric diarrhea in Bangladesh and Mali: do provider expectations predict experiences? Am J Trop Med Hyg 2022 Jul 13;107(1):32-34 [FREE Full text] [doi: 10.4269/ajtmh.21-1248] [Medline: 35895365]
41. Nair MM, Mahajan R, Burza S, Zeegers MP. Behavioural interventions to address rational use of antibiotics in outpatient settings of low-income and lower-middle-income countries. Trop Med Int Health 2021 May 28;26(5):504-517 [FREE Full text] [doi: 10.1111/tmi.13550] [Medline: 33452857]
42. Lampariello R, Ancellin-Panzani S. Mastering stakeholders' engagement to reach national scale, sustainability and wide adoption of digital health initiatives: lessons learnt from Burkina Faso. Fam Med Community Health 2021 Jun 18;9(3):e000959 [FREE Full text] [doi: 10.1136/fmch-2021-000959] [Medline: 34144970]
43. Goodman C, Kachur SP, Abdulla S, Bloland P, Mills A. Drug shop regulation and malaria treatment in Tanzania--why do shops break the rules, and does it matter? Health Policy Plan 2007 Nov 20;22(6):393-403 [FREE Full text] [doi: 10.1093/heapol/czm033] [Medline: 17921151]

XSL•FO
RenderX

44. Porter G, Joshi J, Bhullar L, Kotwani A, Smart Regulation of Antibiotic Use in India: Understanding, InnovatingImproving Compliance Project Group. Using 'smart regulation' to tackle antimicrobial resistance in low-income and middle-income countries. BMJ Glob Health 2020 Jan 30;5(1):e001864 [FREE Full text] [doi: 10.1136/bmjgh-2019-001864] [Medline: 32133168]

Abbreviations

AMR: antimicrobial resistance
CDSS: clinical decision support system
CFIR: Consolidated Framework for Implementation Research
CNIL: Commission Nationale de l'Informatique et des Libertés
COREQ: Consolidated Criteria for Reporting Qualitative Research
ERIC: Expert Recommendations for Implementing Change
ICT: information and communication technology
IeDA: Integrated e-Diagnostic Approach
LMIC: low- and middle-income country
mHealth: mobile health
SAPI: African Society for Infectious Diseases
SPILF: French Society for Infectious Diseases
WHO: World Health Organization

Edited by T Leung, A Mavragani; submitted 21.12.22; peer-reviewed by Z Huang; comments to author 02.03.23; revised version received 10.04.23; accepted 20.06.24; published 07.10.24.

Please cite as:
Peiffer-Smadja N, Descousse S, Courrèges E, Nganbou A, Jeanmougin P, Birgand G, Lénaud S, Beaumont AL, Durand C, Delory T, Le Bel J, Bouvet E, Lariven S, D'Ortenzio E, Konaté I, Bouyou-Akotet MK, Ouedraogo AS, Kouakou GA, Poda A, Akpovo C, Lescure FX, Tanon A
Implementation of a Clinical Decision Support System for Antimicrobial Prescribing in Sub-Saharan Africa: Multisectoral Qualitative Study
J Med Internet Res 2024;26:e45122
URL: https://www.jmir.org/2024/1/e45122
doi:10.2196/45122
PMID:39374065

©Nathan Peiffer-Smadja, Sophie Descousse, Elsa Courrèges, Audrey Nganbou, Pauline Jeanmougin, Gabriel Birgand, Séverin Lénaud, Anne-Lise Beaumont, Claire Durand, Tristan Delory, Josselin Le Bel, Elisabeth Bouvet, Sylvie Lariven, Eric D'Ortenzio, Issa Konaté, Marielle Karine Bouyou-Akotet, Abdoul-Salam Ouedraogo, Gisèle Affoue Kouakou, Armel Poda, Corinne Akpovo, François-Xavier Lescure, Aristophane Tanon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reinforcement Learning to Optimize Ventilator Settings for Patients on Invasive Mechanical Ventilation: Retrospective Study

Siqi Liu[1*], PhD; Qianyi Xu[2*], BEng; Zhuoyang Xu[3], MSc; Zhuo Liu[3], MSc; Xingzhi Sun[3], PhD; Guotong Xie[3], PhD; Mengling Feng[2,4], PhD; Kay Choong See[5], MBBS

[1]National University of Singapore Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore, Singapore
[2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore
[3]Ping An Healthcare Technology, Beijing, China
[4]Institute of Data Science, National University of Singapore, Singapore, Singapore
[5]Division of Respiratory and Critical Care Medicine, Department of Medicine, National University Hospital, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Mengling Feng, PhD
Saw Swee Hock School of Public Health
National University of Singapore
12 Science Drive 2
Singapore, 117549
Singapore
Phone: 65 65164984
Email: ephfm@nus.edu.sg

Abstract

Background: One of the significant changes in intensive care medicine over the past 2 decades is the acknowledgment that improper mechanical ventilation settings substantially contribute to pulmonary injury in critically ill patients. Artificial intelligence (AI) solutions can optimize mechanical ventilation settings in intensive care units (ICUs) and improve patient outcomes. Specifically, machine learning algorithms can be trained on large datasets of patient information and mechanical ventilation settings. These algorithms can then predict patient responses to different ventilation strategies and suggest personalized ventilation settings for individual patients.

Objective: In this study, we aimed to design and evaluate an AI solution that could tailor an optimal ventilator strategy for each critically ill patient who requires mechanical ventilation.

Methods: We proposed a reinforcement learning–based AI solution using observational data from multiple ICUs in the United States. The primary outcome was hospital mortality. Secondary outcomes were the proportion of optimal oxygen saturation and the proportion of optimal mean arterial blood pressure. We trained our AI agent to recommend low, medium, and high levels of 3 ventilator settings—positive end-expiratory pressure, fraction of inspired oxygen, and ideal body weight–adjusted tidal volume—according to patients' health conditions. We defined a policy as rules guiding ventilator setting changes given specific clinical scenarios. Off-policy evaluation metrics were applied to evaluate the AI policy.

Results: We studied 21,595 and 5105 patients' ICU stays from the e-Intensive Care Unit Collaborative Research (eICU) and Medical Information Mart for Intensive Care IV (MIMIC-IV) databases, respectively. Using the learned AI policy, we estimated the hospital mortality rate (eICU 12.1%, SD 3.1%; MIMIC-IV 29.1%, SD 0.9%), the proportion of optimal oxygen saturation (eICU 58.7%, SD 4.7%; MIMIC-IV 49%, SD 1%), and the proportion of optimal mean arterial blood pressure (eICU 31.1%, SD 4.5%; MIMIC-IV 41.2%, SD 1%). Based on multiple quantitative and qualitative evaluation metrics, our proposed AI solution outperformed observed clinical practice.

Conclusions: Our study found that customizing ventilation settings for individual patients led to lower estimated hospital mortality rates compared to actual rates. This highlights the potential effectiveness of using reinforcement learning methodology to develop AI models that analyze complex clinical data for optimizing treatment parameters. Additionally, our findings suggest the integration of this model into a clinical decision support system for refining ventilation settings, supporting the need for prospective validation trials.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e44494) doi:10.2196/44494

KEYWORDS

mechanical ventilation; reinforcement learning; artificial intelligence; validation study; critical care; treatment; intensive care unit; critically ill; patient; monitoring; database; mortality rate; decision support; support tool; survival; prognosis; respiratory support

Introduction

Mechanical ventilation is the foundation of critical care medicine and is one of the most common interventions for patients admitted to intensive care units (ICUs). Studies showed that approximately one-third of ICU patients require mechanical ventilation in the United States [1]. In recent years, due to the COVID-19 pandemic and aging populations in many countries, mechanical ventilation in ICU use has been constantly increasing.

Despite decades of research, choosing the optimal ventilator strategy for a patient remains imprecise. Appropriate ventilator settings are important but complicated by significant interpatient variability. Current clinical guidelines provide one-size-fits-all recommendations but do not personalize the treatment for different ICU patients. In particular, existing clinical guidelines do not address personalized optimal settings for mechanical ventilation, including positive end-expiratory pressure (PEEP) level, fraction of inspired oxygen (FiO_2), and ideal body weight–adjusted tidal volume [2]. With the understanding that mechanical ventilation itself can cause and potentiate lung injury, it is important to choose appropriate ventilatory strategies to mitigate ventilator-induced lung injury [3]. Nonetheless, even guideline recommendations may not be adhered to, as a wide discrepancy in practice exists and evidence-based interventions are underused for the task [4].

The drive to discover an effective solution capable of managing the intricate ICU environment and providing personalized treatment to each patient is a compelling motivator. One particularly promising approach is the use of reinforcement learning (RL) for formulating treatment recommendations, supported by the following reasons. First, RL is a decision-making tool that can learn complex sequential decisions, making it a natural fit for critical care applications. Second, RL can take individual patients' health conditions and disease histories into account, hence providing more personalized treatment decisions that have the potential to surpass existing clinical practices. However, the RL method for mechanical ventilation guidance needs further evaluation before committing resources for prospective clinical studies. We therefore aimed to test the concept that RL can optimize ventilator settings for patients on invasive mechanical ventilation, by applying RL on existing large ICU databases.

Methods

Overview of the Methods

For our study, we named the RL-based artificial intelligence (AI) solution "EZ-Vent." The framework of the proposed solution is shown in Figure 1. We first collected mechanically ventilated patients' health data and intensivists' treatment actions from 2 large electronic health record (EHR) datasets in the United States. We then trained a type of RL-based model, named the Batch Constrained Deep Q-learning (BCQ), to learn from physicians' treatment actions and to develop an optimal strategy for setting mechanical ventilation. This type of learning is commonly referred to as batch learning in RL. However, many traditional RL algorithms have been unsuccessful in the batch setting, while the models they produced often suffered from overestimation and exhibited poor performance when presented with data not included in the provided batch. In contrast to traditional RL algorithms, the BCQ algorithm imposes constraints to ensure that the learned policy remains reasonably close to physicians' policy. For this reason, we chose to implement BCQ in our solution due to its capacity to develop a safe policy from observational data. Given the crucial significance of safe policy learning in health care applications, the proposed AI solution may then be integrated as a component of a clinical decision support system, assisting intensivists in making optimal decisions for critically ill patients who require mechanical ventilation.

Our proposed AI solution recommends optimal ventilator settings for PEEP, FiO_2, and tidal volume levels by considering the individual patients' conditions including their demographic features, physiological status, and multiple comorbidities. Compared to the existing guidelines, the proposed solution can adjust treatment recommendations based on changes in a patient's condition. Moreover, we developed a set of flags designed to detect sudden changes in patients' health and leveraged the timing of these flags to partition patients' trajectories into discrete time-varying intervals. We anticipated that these timings correspond to critical decision points for physicians to intervene in practice. If our model were to be implemented at the bedside in real time, it has the potential to assist intensivists in making more informed and optimized decisions. Studies have reported ICU mortality rates as high as 86% to 97% for invasive mechanical ventilation [5-7], and improved ventilator settings would greatly benefit critically ill patients. As such, our proposed AI solution holds promise for improving patient outcomes in ICUs for those requiring mechanical ventilation.

Although general improvements in ICU outcomes and changes in ventilation practices over time would positively affect the model in the training process, the proposed BCQ model would automatically learn from the good practices and avoid bad practices to achieve the best long-term return, which is the survival of the patients.

XSL•FO
RenderX

Figure 1. Proposed EZ-Vent framework. We collected data on ventilated patients from EHR and trained a policy to recommend optimal ventilator settings. EHR: electronic health record.

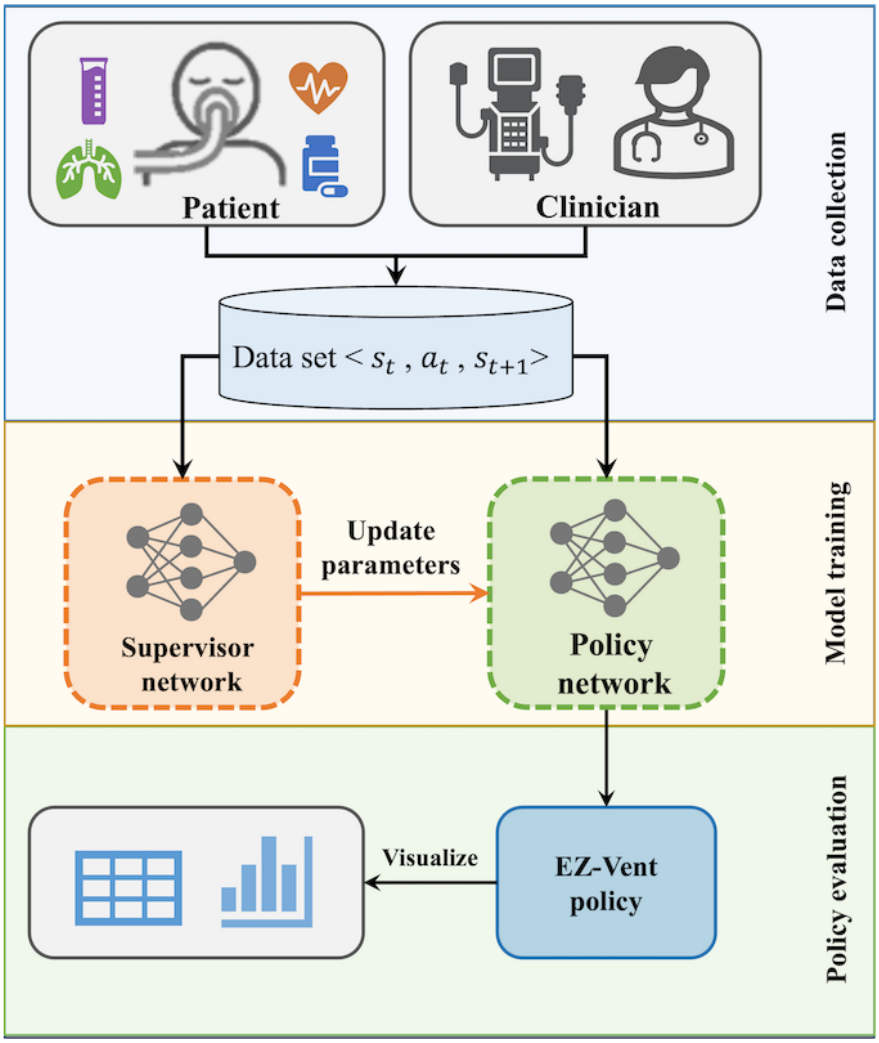

Study Population and Datasets

The observational data for mechanically ventilated patients were extracted from 2 large EHR databases in the United States: the Medical Information Mart for Intensive Care IV (MIMIC-IV) database [8] and the e-Intensive Care Unit Collaborative Research (eICU) database [9]. We included patients who were aged 16 years and older, and whose ventilation duration was at least 24 hours. Only the first ICU admission for each patient was considered, and we studied the first 48 hours of ventilation data.

We excluded patients who did not have data for mortality, height, or sex. In addition, we excluded patients whose mechanical ventilation duration exceeded 2 weeks, because patients who required long-term mechanical ventilation may not be representative of the general population of patients who require mechanical ventilation. Lastly, we excluded patients who have missing ventilation settings of PEEP, FiO_2, and tidal volume recorded for the entire ventilation duration.

After those exclusions, 5105 patients in the MIMIC-IV dataset and 21,595 patients in the eICU dataset remained. The flowchart for cohort selection is shown in Figure 2. We conducted a 5-fold cross-validation on the eICU dataset. The full MIMIC-IV dataset was held out as the testing set.

XSL•FO
RenderX

Figure 2. Overview of exclusion criteria and the number of patients left after each round of selection. eICU: e-Intensive Care Unit Collaborative Research; FiO2: fraction of inspired oxygen; ICU: intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care; MV: mechanical ventilation; PEEP: positive end-expiratory pressure.

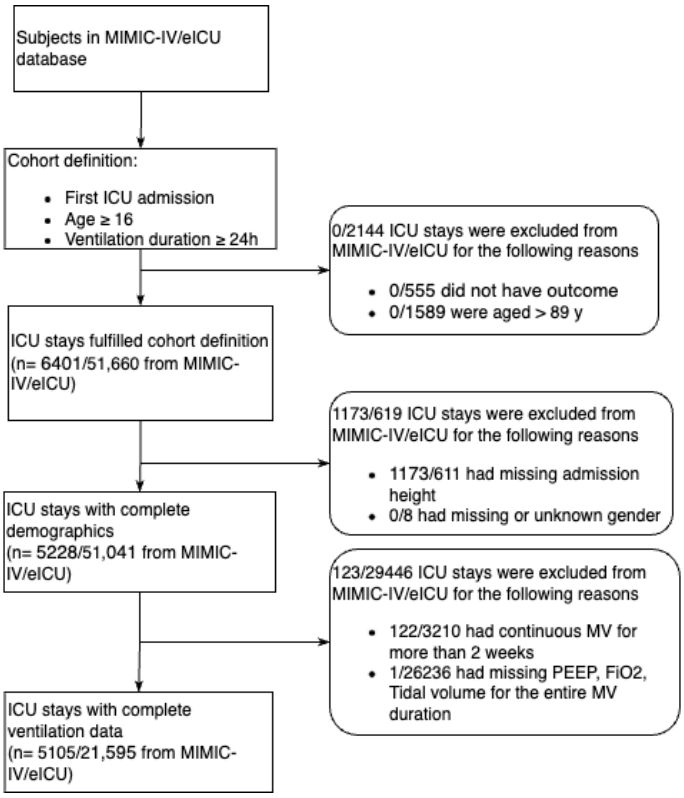

Outcome Variables

The primary outcome was hospital mortality. Mortality outcomes are the most important outcomes for patients in the ICUs, given the high mortality for patients in ICUs in general. Mortality is a definitive measure of success in interventions, given the ultimate goal of ICU care is to save lives.

Secondary outcomes were the proportion of optimal oxygen saturation (SpO_2) and the proportion of optimal mean arterial blood pressure (MBP). The optimal ranges of SpO_2 and MBP were defined as follows: 94%<SpO_2<98% [10] and 70 mm Hg≤MBP≤80mm Hg [11].

RL: A Primer

RL is a goal-oriented AI method where a computer agent, acting as a decision maker, analyzes available data within its defined environment, derives a policy for taking actions, and optimizes long-term rewards. The agent is the computational model we want to develop. In general, an agent obtains evaluative feedback (reward) about the performance of its action at each consecutive time step, allowing it to improve the performance of subsequent actions by trial and error. Mathematically, the sequential decision-making process is called the Markov Decision Process (MDP). We define the 5-tuple MDP as (S, A, P, R, γ):

- State: A state $s_t \in S$ is the state at time t in state space S. In this study, it represents the health status of a patient at each timestamp. We constructed the patient's state by using 40 relevant physiological features containing demographics, laboratory values, and vital signs (see Table S1 in Multimedia Appendix 1 for the full list).
- Action: An action $a_t \in A$ is the treatment option that the agent takes at each time step t, which influences the next state s_{t+1}. In our study, the action space was constructed as 18 possible discrete actions from combinations of low, medium, and high levels of the 3 ventilator settings: PEEP, FiO_2, and ideal body weight–adjusted tidal volume (Figure S1 in Multimedia Appendix 1.
- Transition probability: $P(s_t|a_t) \rightarrow s_{t+1}$ is the probability of transiting from the state s_t to the next state s_{t+1} given an action a_t.
- Reward: R is the observed feedback given a state-action (s_t, a_t) pair. The reward of our model reflected the objective of an RL agent, which was to improve survival and achieve oxygen saturation and mean arterial pressure within their respective optimal ranges. Hospital mortality was used as the terminal reward, whereas SpO_2 and MBP were applied as intermittent rewards.
- $\gamma \in \{0,1\}$ is the discount factor.

We assume the process of ventilator adjustment has Markov property. That is, the state space is completely observable, the state transition probability P is only related to the last state and the last action, and the immediate reward is only related to the state and the action taken in the corresponding step.

The solution of the MDP is an optimized set of rules, that is, the RL policy. The ability of RL to learn complex sequential decisions makes it suitable for critical care applications, and we hope to use its capability of learning to provide individualized treatment policies that could improve the survival of patients who are mechanically ventilated.

XSL•FO
RenderX

Specifically, we aim to train an RL policy π: $S \times A$ {0,1}, which specifies the probability of taking each action in each state through Q-learning, where $Q^{\pi}(s, a)$ is the value of taking action a in state s using policy π and is defined as the expected sum of future rewards, discounted by γ at each time step as follows:

$$Q^{\pi}(s, a) = E\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... |\pi(s, a)\}$$

RL: Time-Varying Intervals and Flags

We applied time-varying intervals to discretized state and action pairs. For each patient's treatment trajectory, we analyzed the data from the first hour of using mechanical ventilation until the 48th hour or until ventilator weaning, whichever came earlier. Then we cut the trajectories into 4-hour time steps, except for cases when flags were raised to further cut the trajectories. We designed the set of flags as follows: (1) the SpO_2 dropped under 90%, (2) the partial pressure of oxygen (PaO_2) dropped under 60 mm Hg, and (3) pH <7.25 or pH >7.45. When any flag was raised, the trajectories would be further cut into shorter time steps. We designed this set of flags to reflect real-world conditions, where such flags would prompt changes in ventilator settings. Next, we selectively merged time steps if they were too short, with a minimum time interval of no less than 1 hour. For multiple values in 1 time step, we computed a time-weighted average value. Patients that had at least 1 of the 3 actions empty for all the time steps were removed.

Data imputation is common in EHR data analysis as some data are manually entered by the doctors, which may cause them to be recorded less frequently. Similar ways such as k-nearest neighbor imputation and time-windowed sample-and-hold method have been applied to handle data imputation in previous works [12,13]. In the literature that reported the percentage of imputed data, the percentage of imputation was as high as 95% [13]. In our study, 66.28% (n=14,314/21,595) of patients in the eICU database had data requiring imputation. Among these patients, a median 87% (IQR 64%-95%) of data were imputed. Data in the MIMIC-IV database were more complete, with 33.8% (n=1724/5105) of patients requiring imputation. Among these patients, a median 72% (IQR 47%-89%) of data were imputed. The distribution of data that require imputation was reported in Figure S6 for eICU and S7 for MIMIC-IV in Multimedia Appendix 1. Missing values were imputed with the nearest value before the time step, and if this was not available, missing values were imputed with the value from the next time step. Binary state variables were represented using 0 or 1. Continuous state variables were normalized or log-normalized to (0, 1) as appropriate.

RL: Model Development

One major drawback of using historical EHR data to train an RL model is extrapolation error. The term extrapolation is a statistical technique for estimating values that extend beyond a particular collection of data or observations. Extrapolation error is caused by the mismatch between the data distribution in the offline dataset and future observations. To mitigate this problem, we applied the BCQ model in this study, which is an RL model that has the advantage over other RL algorithms in the batch setting by addressing the extrapolation error [14].

The BCQ model consists of 2 modules, a supervisor network and a policy network. The supervisor network is used to mimic the physicians' policy from the observational data. The policy network fits the optimal policy under the constraint of the supervisor network, where only the actions likely to be taken in the observational data are considered and evaluated. The final optimized policy is then expected to lead to good future outcomes as well as to select a safe action.

In this study, the loss of the BCQ model is defined as the combination of 2 loss functions: $L = L_Q + \beta L_P$, where L_Q is the value loss and is defined as:

L_P is the probability loss and is defined as $L_P = -\log(P(s_t, a_t))$. The final RL policy is defined as:

In the training stage, when BCQ receives the training sample, the supervisor network will first learn the mapping from state to action, that is, which action would be taken based on historical data. Then, the policy network will optimize its policy with the reward information and the output of the supervisor network. This training process is iterated several times until we derive the final AI policy.

For the supervisor network, we adopted a fully connected stream with 2 hidden layers of 256 units to infer the action value $Q^{\pi}(s, a)$ function and a fully connected stream with 2 hidden layers of 256 units to infer the state-action probability $P(s_t, a_t)$. Each hidden layer contained the rectified linear unit activation. The policy network had the same structure as the supervisor network. The learning rate was 0.0003, the discount factor γ was 0.99, the batch size was 32, the tracking rate α was 0.01, the extrapolation threshold τ was 0.05, and the trade-off factor of 2 kinds of loss functions β was 1. We trained the RL model using the Adam optimizer.

We designed a clinically guided reward function that produced a reward (penalty) when the patients' state improved (deteriorated) based on short-term and long-term health outcomes. The short-term health indicators were the patients' MBP as well as SpO_2. At each intermittent (ie, nonterminal) time step of a patient's trajectory, the patient would receive a positive short-term reward b if MBP fell within the range of 70-80 mm Hg] or c if SpO_2 fell within 94% to 98% based on the literature on maintaining optimal levels of vital signs and blood gases, and the patient would receive a penalty (negative reward) of $-b/2$ or $-c/2$ when MBP and SpO_2 fell out of the range. We applied the range (94%, 98%) of SpO_2 for the intermittent reward design due to the following reasons: conservative O_2 therapy has been variably defined in various randomized controlled trials (RCTs). RCTs investigating the lower SpO_2 threshold have found evidence of increased mortality from SpO_2 <93% [15]. Other RCTs did not show any increased mortality in their conservative groups if these groups attained an SpO_2 (or equivalent PaO_2) within 94% to 98% (eg, a conservative group of ICU randomized trial comparing 2

XSL•FO
RenderX

approaches to oxygen therapy [16] had time-weighted PaO_2 ~80 mm Hg and the lowest SpO_2 group of Pragmatic Investigation of Optimal Oxygen Targets [17] had SpO_2 around 94%). At terminal time steps, each patient would receive a final reward a (or penalty $-a/2$) if a patient survived (or became deceased) at discharge. The overall reward function was defined as follows:

where S_T, S_{sur}, and S_{dec} sets represented terminal, survived, and deceased patient states, and a, b, and c were parameters that were tuned during training. The reward information at each time step would help BCQ learn those action patterns from physicians that lead to good short-term and long-term outcomes.

To account for potential cointerventions that would affect the MBP and survival, we included the maximum dose of vasopressor (Table S1 in Multimedia Appendix 1) over every 4-hour time window of mechanical ventilation within patient states for our RL model. In addition, as differences in illness severity could also modify mortality risk, we included the patients' Elixhauser score, sequential organ failure assessment (SOFA) score, number of systemic inflammatory response syndrome criteria, vital signs, and laboratory test results (Table S1 in Multimedia Appendix 1) in the patient state to reflect differences in patients' illness severity over time. The treatment action from the model was conditioned on all the state variables so that state differences were handled in the model.

RL: Benchmark Policies

We evaluated our RL-based policy by comparison with 3 benchmark policies:

- Random policy: All 18 discrete actions have equal probabilities to be chosen.
- One-size-fits-all policy: The action with the highest probability in the cohort is always chosen.
- Physicians' policy: The actual observed policy in the validation and testing sets.

RL: Evaluation Metrics

We used extensive quantitative and qualitative analyses to evaluate the performance of the learned AI policy and benchmarks. First, to understand the relationship between the expected return of the learned policies and the clinical outcomes, we mapped the expected return to the estimated outcome occurrence. We sorted the expected returns of the physicians' policy into discrete bins and obtained the average empirical mortality rate from the patients in each bin. The empirical mortality estimate was used to derive a relationship between the range of computed returns of the AI policy against the observed mortality. This estimation process was performed for secondary outcomes too.

Treatment recommendation is an off-policy learning problem, which aims to learn an optimal policy using trajectories from an observed behavior policy (physicians' policy). Evaluating the learned policy with off-policy estimation (OPE) methods is crucial for health care applications to avoid the high risk of failure or negative impact. OPE methods use examples from the behavior policy to evaluate the performance of the learned policy. Precise evaluation with OPE remains a challenging problem. Previous studies have used the V-curve method [18,19], observed mortality [19,20], importance sampling evaluation [21], and eligibility traces [22]. In this work, we adopted multiple evaluation metrics. We followed the V-curve method to qualitatively evaluate changes in mortality with action differences. We also quantitatively estimated the mortality rate and performed importance sampling with one type of importance sampling estimator, namely Consistent Weighted Per-Decision Importance Sampling (CWPDIS) [23], which is defined as:

where D_{beh} is a retrospective trajectory set generated by physician policy π_{cl}, and $n = (S_{n0}, a_{n0}, r_{n1}, s_{n1}, a_{n1}, r_{n2}, ..., s_{nT-1}, a_{nT-1}, r_{nT})$ is a specific trajectory with state, action, and reward in each time step. Note that CWPDIS is based on important sampling, which is a general technique for accomplishing OPE. Compared with other OPE methods, CWPDIS could make unbiased evaluations with higher sampling efficiency. In addition, we used a random forest classification model to rank the importance of various predictors for the actions under the physicians' policy (Figures S3-S5 in Multimedia Appendix 1). This allowed us to understand physicians' behavior regarding the choice of ventilator settings.

Ethical Considerations

The collection of patient information and creation of the research resource in the MIMIC-IV database was reviewed by the Institutional Review Board at the Beth Israel Deaconess Medical Center (number 2001-P-001699/14), which granted a waiver of informed consent and approved the data-sharing initiative. For the eICU database, it has been approved by the Institutional Review Board of the Massachusetts Institute of Technology. After completing the National Institutes of Health's online training course and the Protection of Human Research Participants Examination, we had the access to extract data from both the MIMIC-IV and the eICU databases. The study data are anonymous and deidentified.

Results

Patient Characteristics

Patient characteristics of the selected cohorts from the MIMIC-IV and eICU datasets are provided in Table 1. There were no statistically significant differences in age, sex, or body weight. However, we observed that patients in the MIMIC-IV dataset had greater illness severity compared with those in the eICU dataset. Patients in the MIMIC-IV dataset had higher Elixhauser scores (5.0 [0.0, 12.0] vs 3.0 [0.0, 7.0]), higher reintubation rate (30.7% vs 16.7%), longer hospital stay (291.0 hours [171.0, 477.0] vs 191.4 hours [120.1, 307.5]), and higher hospital mortality rate (31.1% vs 18.2%) compared to patients in the eICU dataset.

XSL•FO
RenderX

Table 1. Patient characteristics.

Variables	MIMIC-IV (N=5105)	eICU (N=21,595)
Female (%)	2154 (42.2)	9244 (42.8)
Age (years), median (IQR)[a]	65.0 (53.0, 76.0)	64.0 (53.0, 74.0)
Body weight (kg), median (IQR)[a]	80.9 (67.6, 97.2)	81.9 (68.0, 99.7)
Reintubation (%)	1565 (30.7)	3661 (16.7)
Elixhauser score, median (IQR)[a]	5.0 (0.0, 12.0)	3.0 (0.0, 7.0)
First SOFA[b] score, median (IQR)[a]	2.0 (0.0, 4.0)	2.0 (0.0, 4.0)
Hospital LOS[c] (h), median (IQR)[a]	291.0 (171.0, 477.0)	191.4 (120.1, 307.5)
Hospital mortality (%)	1590 (31.1)	3934 (18.2)
PEEP[d], (cmH2O), median (IQR)[a]	5.3 (5.0, 10.0)	5.0 (5.0, 5.0)
FiO_2 (%), median (IQR)[a]	50.0 (40.0, 60.0)	49.2 (40.0, 61.1)
Tidal volume[e] (ml/kg IBW[f]), median (IQR)[a]	7.2 (6.4, 8.3)	7.8 (6.9, 8.8)

[a]25th percentile, 75th percentile.
[b]SOFA: sequential organ failure assessment.
[c]LOS: length of stay.
[d]PEEP: positive end-expiratory pressure.
[e]Tidal volume: ideal weight-adjusted tidal volume.
[f]IBW: ideal body weight.

Performance of the RL Method

We plotted the action frequency distributions of the physicians' policy and the learned AI policy. We compared the learned policy against the physicians' policy for low (<5), medium (5-15), and high (>15) SOFA score levels for patients in the eICU (Figure 3) and the MIMIC-IV databases (Figure S3 in Multimedia Appendix 1). For each SOFA group, we counted the number of actions for the 3 action categories: PEEP, FiO_2, and tidal volume. Actions taken by physicians were different from those suggested by AI. The learned policy recommended more low-level actions for PEEP and FiO_2 and high-level actions for tidal volume. Relationships between the range of computed returns of the learned policy against various outcomes for both the eICU validation set and the MIMIC-IV test set are shown in Figure 4. The figures show that policies with higher returns were associated with lower mortality and higher proportions of optimal SpO_2 and MBP.

The OPE performance of the learned policy is shown in Table 2. We estimated the hospital mortality rate (eICU 12.1%, SD 3.1%; MIMIC-IV 29.1%, SD 0.9%), the proportion of optimal SpO_2 (eICU 58.7%, SD 4.1%; MIMIC-IV 49%, SD 1%), and the proportion of optimal MBP (eICU 31.1%, SD 4.5%; MIMIC-IV 41.2%, SD 1%) for the learned policy. We also report outcomes for the physicians' policy, including the observed mortality rate (eICU 14.3%; MIMIC-IV 30.6%), the proportion of optimal SpO_2 (eICU 47.8%; MIMIC-IV 40.5%), and the proportion of optimal MBP (eICU 28.2%; MIMIC-IV 37.1%) in the 2 datasets, respectively. We also performed *t* tests (2-tailed) for proportions of optimal SpO_2 and MBP, and Fisher exact tests for hospital mortality rate, and calculated the *P* values compared with the physicians' policy. The results from all 3 policies achieved *P* values of <.001, which indicates that the differences were very unlikely to arise from randomness. Overall, the AI policy achieved a longer duration within optimal SpO_2 and MBP ranges with lower mortality compared to the physicians' policy. To examine the effectiveness of using time-varying intervals in the action setting, we visualize a representative patient case in Figure 5. The relationship between mortality and discrepancy between AI and physicians' ventilator settings is illustrated in Figure 6.

We report the feature importance with regards to choosing ventilator settings under the physicians' policy in Figures S3-S5 in Multimedia Appendix 1. The top 10 important features included the following: PaO_2/FiO_2 ratio, $PaCO_2$, PaO_2, creatinine level, lactate, prothrombin time, base excess, age, admission weight, and Richmond Agitation Sedation Scale score.

XSL•FO
RenderX

Figure 3. Comparative action distributions for physicians (blue) and learned policy (red) in the MIMIC-IV test set. Each panel represents actions taken for different SOFA score levels: low (SOFA<5), medium (5≤SOFA<15), and high (SOFA>15). Actions taken by physicians were different from those suggested by AI. The learned policy recommended more low-level actions for PEEP and FiO2 and high-level actions for tidal volume. AI: artificial intelligence; FiO2: fraction of inspired oxygen; Med: medium; Mid: middle; MIMIC-IV: Medical Information Mart for Intensive Care; PEEP: positive end-expiratory pressure; SOFA: sequential organ failure assessment.

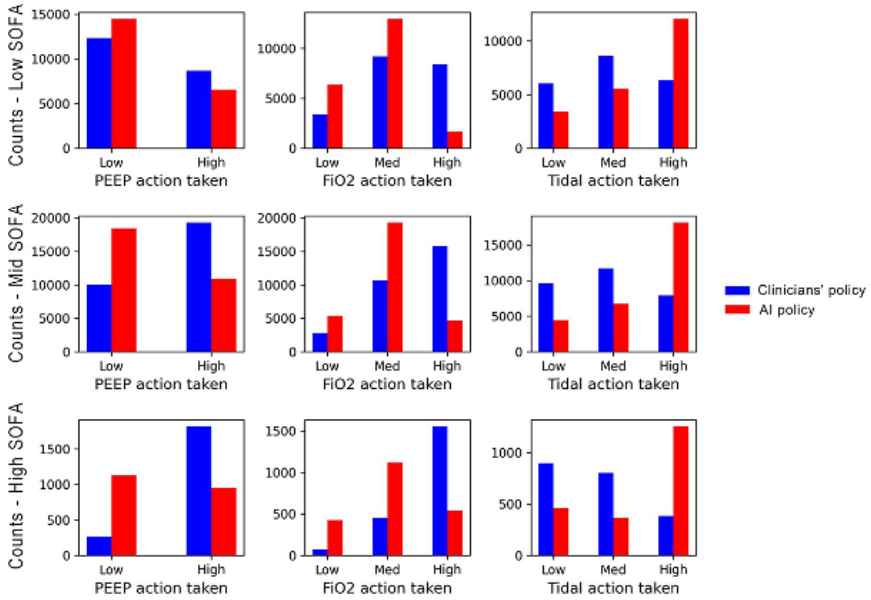

Figure 4. Changes in observed mortality (red), proportion of optimal SpO2 (green), and proportion of optimal MBP (blue) versus the expected return curves for learned policies in the eICU validation set and MIMIC-IV test set. The proportion of mortality (red) against the returns showed inverse relationships, whereas the proportion of optimal SpO2 (green) and the proportion of optimal MBP (blue) against returns showed overall positive relationships in both data sets. Overall, the figures show that policies with higher returns were associated with lower mortality and higher proportions of optimal SpO2 and MBP. eICU: e-Intensive Care Unit Collaborative Research; MBP: mean arterial blood pressure; MIMIC-IV: Medical Information Mart for Intensive Care; SpO2: optimal oxygen saturation.

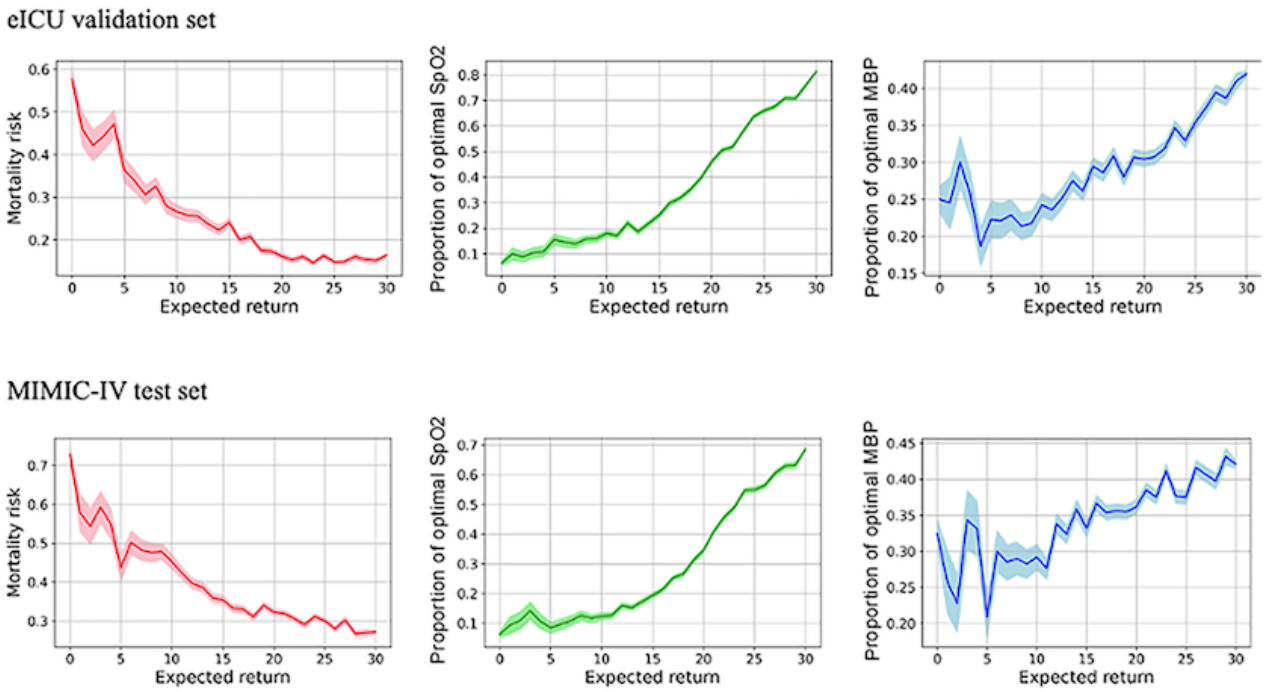

XSL•FO
RenderX

Table 2. Outcomes using the validation set (eICU[a]) and test set (MIMIC-IV[b]).[c]

Dataset	Policy	Proportion of time within SpO_2[d] range (%), mean (SD)	Proportion of time within MBP target[e] range (%), mean (SD)	Observed mortality (%), mean (SD)	*P* value	95% CI
eICU						
	Physician[f]	47.8 (5.0)	28.2 (4.5)	14.3 (3.6)	__[g]	—
	Random[h]	49.9 (4.8*)	30.7 (4.5*)	15.2 (3.5*)	<.001	15.17-15.23
	One-size-fit-all[i]	53.2 (4.9*)	29.4 (4.5*)	17.5 (3.8*)	<.001	17.47-17.53
	AI[j]	58.7 (4.7*)	31.1 (4.5*)	12.1 (3.1*)	<.001	12.07-12.13
MIMIC-IV						
	Physician	40.5 (1.0)	37.1 (1.0)	30.6 (0.9)	—	—
	Random	34.6 (1.0*)	36.2 (1.0*)	32.3 (0.9*)	<.001	32.29-32.31
	One-size-fit-all	40.9 (1.0*)	38.5 (1.0*)	32.0 (0.9*)	<.001	31.99-32.01
	AI	49.0 (1.0*)	41.2 (1.0*)	29.1 (0.9*)	<.001	29.09-29.11

[a]eICU: e-Intensive Care Unit Collaborative Research.

[b]MIMIC-IV: Medical Information Mart for Intensive Care IV.

[c]**P* value <.001 when compared to physicians' policy.

[d]SpO_2 target range: 94%<SpO_2<98%.

[e]MBP target range: 70 mm Hg ≤ MBP ≤ 80 mm Hg.

[f]Physician: the actual observed policy in the validation and testing set.

[g]Not applicable.

[h]Random: all the 18 discrete actions have equal probabilities to be chosen.

[i]One-size-fit-all: the action with the highest probability in the cohort is always chosen.

[j]AI: artificial intelligence policy from Batch Constrained Deep Q-learning model.

Figure 5. Visualization of representative patient cases in raw, fixed, time-varying intervals for mechanical ventilator action setting. Visualization of representative case study for mechanical ventilator settings of PEEP (left), FiO2 (middle), and ideal body weighted-adjusted tidal volume (right) using time-varying interval (red), raw data (blue), and fixed 4-hour time interval (green). Flags to cut intervals in the time-varying setting are shown as vertical dotted lines with yellow shadows. The flags could catch the changes in the ventilator settings. FiO2: fraction of inspired oxygen; PaO2: partial pressure of oxygen; PEEP: positive end-expiratory pressure.

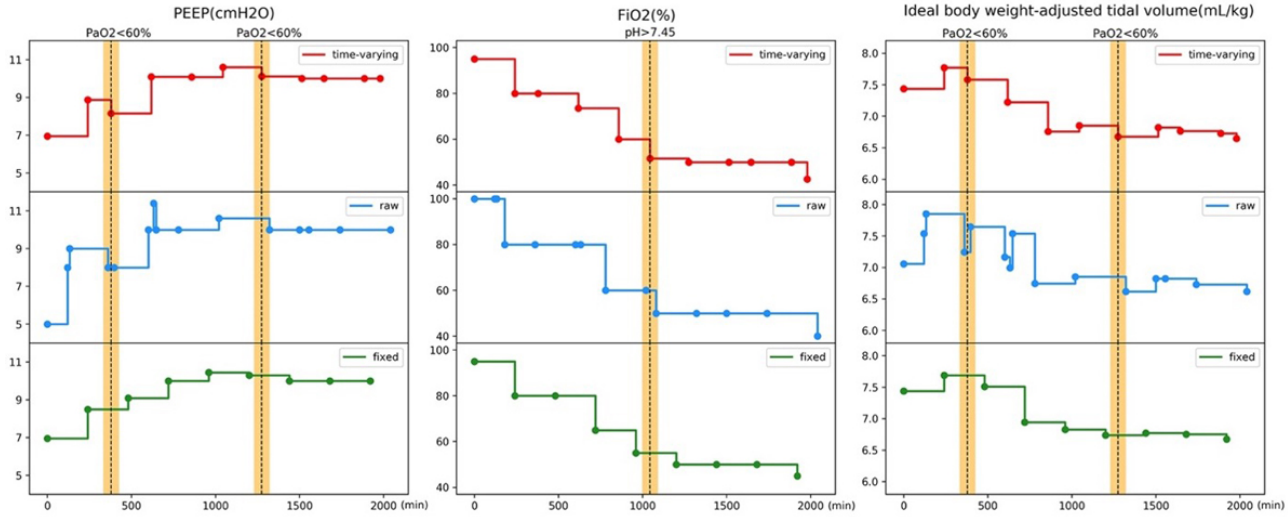

XSL•FO
RenderX

Figure 6. Changes in observed mortality (y-axis) versus the difference between the mechanical ventilation settings recommended by the optimal policy and the settings administered by physicians (x-axis) on the eICU validation set and MIMIC-IV test set. The x-axis indicates the differences in the quantile number. The plots show a v-shape which indicates that mortality is the minimum when we follow the actions suggested by the policy. eICU: e-Intensive Care Unit Collaborative Research; FiO2: fraction of inspired oxygen; MIMIC-IV: Medical Information Mart for Intensive Care; PEEP: positive end-expiratory pressure.

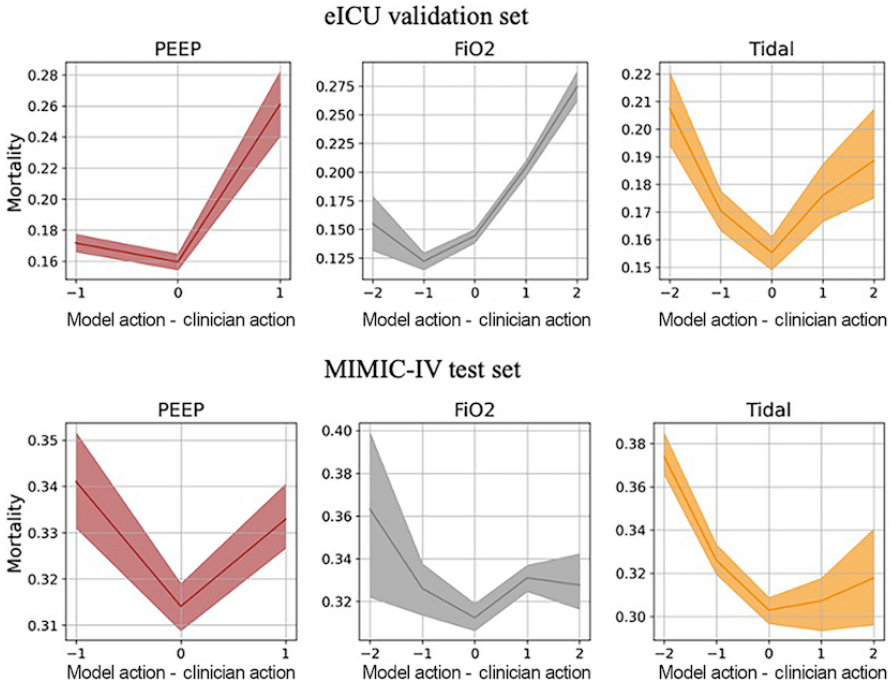

Discussion

Overview

In this study, we used an RL-based AI model (BCQ) to learn the optimal ventilation policy customized for patients who are critically ill and require mechanical ventilation. We validated the policy using 2 large public datasets from the United States: the eICU and MIMIC-IV datasets. In both datasets, the learned policy had superior performance compared to the observed physicians' policy, based on several quantitative and qualitative evaluation metrics.

Principle Findings

In the MIMIC-IV dataset patients exhibited a higher severity of illness relative to those in the eICU dataset. However, this presented an opportunity to evaluate the extrapolation capacity of the BCQ model. The BCQ model-derived RL policy consistently demonstrated superior performance to physicians' policy in both datasets. Consequently, we surmised that the BCQ model's extrapolation ability was acceptable.

We formulated the clinical problem of choosing optimal ventilator settings in the ICU as an RL problem. We then used relevant physiological variables to represent patients' health status as states and cut the ventilator treatment trajectories into time-varying steps to reflect the changes in patients' conditions. We designed a set of flags to capture the sudden changes in patients' health and used the flag timings to further cut the trajectory because such timings were the likely decision points for physicians to make necessary interventions. From the visualization of time-varying intervals in Figure 5, we observed that when the flags were raised (vertical dotted line), time-varying interval setting (red lines) can better reflect the changes in raw data (blue lines) of ventilator settings promptly compared to fixed 4-hour time intervals (green lines).

The AI policy used a "penalty" and "reward" function to regulate SpO_2 and MBP within their optimal ranges. Notably, the policy tried to avoid hyperoxemia due to evidence of its harmful effects, as demonstrated by randomized trials in adults and younger patients [24,25]. While evidence of harm in hypertension was not as strong as in hyperoxemia, physicians could avoid overdosing on vasopressors to minimize the risk of arrhythmia [26]. Nonetheless, caution should be exercised in the use of the model, and it should not be relied upon as a standalone tool. On the contrary, it was designed as a decision support tool that provides more personalized guidance and might lead to better treatment plans. Physicians should evaluate the recommendations from the model carefully and balance between AI predictions and established treatment protocols.

From the action frequency distribution plot (Figure 3) for patients in MIMIC-IV, we found that the actions from physicians (red) and the actions recommended by AI policy (blue) have some discrepancies in all ventilator settings. This result is desirable because the supervisor network in the BCQ model does not aim to duplicate physicians' choices. On the contrary, the supervisor network was used to learn good action patterns from physicians and limit the choice of actions with constraints. In addition, we found the learned policy recommended low-level PEEP and high-level ideal body weight-adjusted tidal volume more frequently compared to physicians' current practices for all the SOFA groups. This finding suggests that the high PEEP-low tidal volume strategy for acute respiratory distress syndrome [2,27] may not be optimal for all mechanically

XSL•FO
RenderX

ventilated patients (eg, patients with focal acute respiratory distress syndrome [28]) and should not be applied as a one-size-fits-all approach. For the management of FiO_2, the learned policy suggested more frequent use of low and medium levels and avoided high levels of FiO_2 for all SOFA groups. This policy suggestion is in line with the known harm from excessive oxygenation, which has been found across different types of critical illness [10,29,30].

We computed the learned policy's expected return, and we plotted it against mortality risk in Figure 4. We observed inverse relationships between expected return and mortality (red) in both validation and testing datasets. This indicates that the optimal policy (high return) results in lower mortality for patients. For the secondary outcomes related to maintaining SpO_2 and MBP within their respective optimal ranges, the expected return showed positive relationships (green for SpO_2 and blue for MBP). This indicates that the optimal policy (high return) leads to higher proportions of SpO_2 and MBP within their respective optimal ranges.

Figure 6 highlights the mortality differences associated with discrepancies between AI-driven and clinician-determined ventilator settings. An effective policy has the lowest mortality when the recommended and administered ventilator settings coincide (the x-axis value is zero), indicating that when the practice strictly followed the AI policy, it had the lowest mortality. At the same time, for an effective policy, the observed mortality should increase as the administered ventilator settings deviate from the recommended settings of the AI policy. Accordingly, an effective policy should have a V-shaped curve with a minimum of 0, which we observed for the AI policy under all the 3 action groups (PEEP, FiO_2, and tidal volume).

From the quantitative evaluation using CWPDIS, we found the learned policy had the lowest observed mortality compared with all 3 benchmark policies. At the same time, the learned policy achieved the highest proportion of optimal SpO_2 and MBP in both datasets. Intuitively and as expected, the random policy had the worst outcome among all the policies.

Comparison to Prior Work

Many recent works have provided RL methods to address treatment recommendation problems [18-21,31-33]. Deep Q Network is one of the most popular RL models used in the literature, which is a powerful model that can handle high-dimensional state spaces, and noisy and incomplete input data. However, the Deep Q Network would not perform well when certain states are rarely observed in the historical data, and it tends to make random treatment assignments in such scenarios. This is referred to as extrapolation error, when a model is used to make predictions outside the range of the input data during training. On the contrary, BCQ is an effective tool to avoid extrapolation error, because it is designed to be more robust to the distribution of the data. This is achieved using a regularization term that encourages the policy to remain close to the initial policy during training when rare states are observed. Other works [20,33] focus on the novel algorithms increasing the model complexity and do not pay enough attention to the extrapolation error.

In addition, previous RL methods used for ICU care used fixed 4-hour intervals in the action setting [18,19]. In this work, we propose a time-varying intervals setting to capture fine-grained treatment assignments, which is more in line with real-world clinical practice.

Limitations

Although our study harnessed 2 large databases for derivation and external validation of an RL model, several limitations remain. First, the ICU environment is highly complex, with many interacting variables that may not be fully captured in the EHR data, thus making it challenging for the RL model to deliver accurate and effective policy. We tried to exploit the data using advanced modeling techniques to capture high-dimensional state spaces' characteristics. Second, our study is retrospective, and the results require prospective validation to ensure safety before deploying it in the ICUs. Third, our study trained on a patient cohort in the United States, which is a high-income country with advanced medical care. Whether the RL model would perform similarly in a lower-resourced country is unknown and the model performance may not be generalizable to such resource-limited settings. Future validation should therefore be done in countries belonging to various World Bank income groups.

Future Directions

Despite the above limitations, our study highlights the potential of AI (specifically RL) to personalize medical care by accounting for the myriad variations in patients' clinical features and tailoring treatment recommendations according to those variations. By using the RL model, different actions could be used to quantify and compare the quality of the current treatment options for a given patient, concerning mortality rate. The proposed solution allows physicians to collaborate with the AI agent while retaining physician control of the decision-making process. Our method may also be applied to complex clinical decision-making beyond mechanical ventilation, such as sepsis management [21] and drug dosing [33].

To ensure the reliability and generalizability of our findings, we conducted thorough validation, both internally and externally, using 2 separate datasets. Despite these efforts, we acknowledge the limitation of using retrospective data to model clinical benefits. To confirm our preliminary results, randomized trials comparing AI-guided management with usual care and protocolized care can be done. The method of AI deployment can also be tested under different conditions: as an advisory versus strict implementation [34].

Conclusions

The clinical implications of using RL models to suggest mechanical ventilation settings in ICU settings are of considerable importance. In this study, the RL model was trained to learn from patient data and to adjust mechanical ventilation settings, thereby optimizing patient outcomes. As the model is capable of continuously adapting to the patient's evolving needs, the AI policy has the potential to outperform current clinical interventions and optimize personalized care for patients who are critically ill. One possible development is the integration of the AI agent into a clinical decision support system to optimize

XSL•FO
RenderX

ventilation settings. However, before this can be done, prospective validation of this method will be needed in various ICU settings.

Acknowledgments

This research is supported by A*STAR, CISCO Systems Pte Ltd, and the National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (award I21001E0002) and MOE Tier 1 Grant through NUS Saw Swee Hock School of Public Health.

Data Availability

The data that support the findings of this study are openly available in the MIMIC-IV Database [8] and the eICU database [9]. Our code is available online [35].

Authors' Contributions

All authors contributed to this study's design. KCS contributed to the literature search. SL collected the data. SL, QX, ZX, and ZL contributed to data analysis, figures, and tables. All authors contributed to the writing and approval of this paper.

Conflicts of Interest

KCS declares receipt of honoraria from GE Healthcare and Medtronic.

Multimedia Appendix 1
Tables and figures for features and action details.
[DOCX File , 985 KB - jmir_v26i1e44494_app1.docx]

References

1. Wunsch H, Wagner J, Herlim M, Chong DH, Kramer AA, Halpern SD. ICU occupancy and mechanical ventilator use in the United States. Critical Care Medicine 2013;41(12):2712-2719. [doi: 10.1097/ccm.0b013e318298a139] [Medline: 23963122]
2. NA. Erratum: an official American Thoracic Society/European Society of Intensive Care Medicine/Society of Critical Care Medicine clinical practice guideline: mechanical ventilation in adult patients with acute respiratory distress syndrome. Am J Respir Crit Care Med 2017;195(11):1540. [doi: 10.1164/rccm.19511erratum] [Medline: 28569586]
3. Slutsky AS, Ranieri VM. Ventilator-induced lung injury. N Engl J Med 2013;369(22):2126-2136. [doi: 10.1056/NEJMra1208707] [Medline: 24283226]
4. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. JAMA 2016;315(8):788-800. [doi: 10.1001/jama.2016.0291] [Medline: 26903337]
5. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA 2020;323(20):2052-2059 [FREE Full text] [doi: 10.1001/jama.2020.6775] [Medline: 32320003]
6. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395(10229):1054-1062 [FREE Full text] [doi: 10.1016/S0140-6736(20)30566-3] [Medline: 32171076]
7. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020;8(5):475-481 [FREE Full text] [doi: 10.1016/S2213-2600(20)30079-5] [Medline: 32105632]
8. Johnson A, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 03;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
9. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
10. van den Boom W, Hoy M, Sankaran J, Liu M, Chahed H, Feng M, et al. The search for optimal oxygen saturation targets in critically ill patients: observational data from large ICU databases. Chest 2020;157(3):566-573. [doi: 10.1016/j.chest.2019.09.015] [Medline: 31589844]
11. Khanna AK, Kinoshita T, Natarajan A, Schwager E, Linn DD, Dong J, et al. Association of systolic, diastolic, mean, and pulse pressure with morbidity and mortality in septic ICU patients: a nationwide observational study. Ann Intensive Care 2023;13(1):9 [FREE Full text] [doi: 10.1186/s13613-023-01101-4] [Medline: 36807233]

XSL•FO
RenderX

12. Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med 2021;4(1):32 [FREE Full text] [doi: 10.1038/s41746-021-00388-6] [Medline: 33608661]
13. Kondrup F, Jiralerspong T, Lau E, de Lara N, Shkrob J, Tran MD, et al. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. AAAI 2023;37(13):15696-15702. [doi: 10.1609/aaai.v37i13.26862]
14. Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. 2019 Presented at: PMLR 97: 36th International Conference on Machine Learning; June 9-15, 2019; Long Beach, CA p. 2052-2062 URL: https://proceedings.mlr.press/v97/fujimoto19a.html
15. Barrot L, Asfar P, Mauny F, Winiszewski H, Montini F, Badie J, et al. Liberal or conservative oxygen therapy for acute respiratory distress syndrome. N Engl J Med 2020;382(11):999-1008. [doi: 10.1056/NEJMoa1916431] [Medline: 32160661]
16. ICU-ROX Investigators and the Australian and New Zealand Intensive Care Society Clinical Trials Group, Mackle D, Bellomo R, Bailey M, Beasley R, Deane A, et al. Conservative oxygen therapy during mechanical ventilation in the ICU. N Engl J Med 2020;382(11):989-998. [doi: 10.1056/NEJMoa1903297] [Medline: 31613432]
17. Semler MW, Casey JD, Lloyd BD, Hastings PG, Hays MA, Stollings JL, et al. Oxygen-saturation targets for critically Ill adults receiving mechanical ventilation. N Engl J Med 2022;387(19):1759-1769. [doi: 10.1056/NEJMoa2208415] [Medline: 36278971]
18. Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M. Deep reinforcement learning for sepsis treatment. arXiv Preprint posted online on November 27, 2017. [doi: 10.48550/arXiv.1711.09602]
19. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. 2017 Presented at: PMLR 68: 2nd Machine Learning for Healthcare Conference; August 18-19, 2017; Boston, MA p. 147-163 URL: https://proceedings.mlr.press/v68/raghu17a.html
20. Wang L, Zhang W, He X, Zha H. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. 2018 Presented at: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; August 19-23, 2018; London, United Kingdom. [doi: 10.1145/3219819.3219961]
21. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24(11):1716-1720 [FREE Full text] [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
22. Precup D, Sutton RS, Singh S. Eligibility traces for off-policy policy evaluation. In: Computer Science Department Faculty Publication Series. 2000 Presented at: ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning; June 29 to July 2, 2000; Stanford, CA p. 759-766 URL: https://dl.acm.org/doi/10.5555/645529.658134
23. Thomas PS. Safe reinforcement learning. Papers With Code. 2015. URL: https://paperswithcode.com/task/safe-reinforcement-learning [accessed 2024-09-05]
24. Cumpstey AF, Oldman AH, Martin DS, Smith A, Grocott MPW. Oxygen targets during mechanical ventilation in the ICU: a systematic review and meta-analysis. Crit Care Explor 2022;4(4):e0652 [FREE Full text] [doi: 10.1097/CCE.0000000000000652] [Medline: 35506014]
25. Lilien TA, Groeneveld NS, van Etten-Jamaludin F, Peters MJ, Buysse CMP, Ralston SL, et al. Association of arterial hyperoxia with outcomes in critically ill children: A systematic review and meta-analysis. JAMA Netw Open 2022;5(1):e2142105 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.42105] [Medline: 34985516]
26. Hylands M, Moller MH, Asfar P, Toma A, Frenette AJ, Beaudoin N, et al. A systematic review of vasopressor blood pressure targets in critically ill adults with hypotension. Can J Anaesth 2017;64(7):703-715. [doi: 10.1007/s12630-017-0877-1] [Medline: 28497426]
27. Briel M, Meade M, Mercat A, Brower RG, Talmor D, Walter SD, et al. Higher vs lower positive end-expiratory pressure in patients with acute lung injury and acute respiratory distress syndrome: Systematic review and meta-analysis. JAMA 2010;303(9):865-873. [doi: 10.1001/jama.2010.218] [Medline: 20197533]
28. Constantin JM, Jabaudon M, Lefrant JY, Jaber S, Quenot JP, Langeron O, et al. Personalised mechanical ventilation tailored to lung morphology versus low positive end-expiratory pressure for patients with acute respiratory distress syndrome in France (the LIVE study): a multicentre, single-blind, randomised controlled trial. Lancet Respir Med 2019;7(10):870-880. [doi: 10.1016/S2213-2600(19)30138-9] [Medline: 31399381]
29. Asfar P, Schortgen F, Boisramé-Helms J, Charpentier J, Guérot E, Megarbane B, et al. Hyperoxia and hypertonic saline in patients with septic shock (HYPERS2S): a two-by-two factorial, multicentre, randomised, clinical trial. Lancet Respir Med 2017;5(3):180-190. [doi: 10.1016/S2213-2600(17)30046-2] [Medline: 28219612]
30. Girardis M, Busani S, Damiani E, Donati A, Rinaldi L, Marudi A, et al. Effect of conservative vs conventional oxygen therapy on mortality among patients in an intensive care unit: the oxygen-ICU randomized clinical trial. JAMA 2016;316(15):1583-1589 [FREE Full text] [doi: 10.1001/jama.2016.11993] [Medline: 27706466]
31. Peng X, Ding Y, Wihl D, Gottesman O, Komorowski M, Lehman LWH, et al. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. AMIA Annu Symp Proc 2018;2018:887-896 [FREE Full text] [Medline: 30815131]

XSL•FO
RenderX

32. Lopez-Martinez D, Eschenfeldt P, Ostvar S, Ingram M, Hur C, Picard R. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep Q networks. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:3960-3963. [doi: 10.1109/EMBC.2019.8857295] [Medline: 31946739]
33. Zhao Y, Zeng D, Socinski MA, Kosorok MR. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics 2011;67(4):1422-1433 [FREE Full text] [doi: 10.1111/j.1541-0420.2011.01572.x] [Medline: 21385164]
34. Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: Comprehensive review. J Med Internet Res 2020;22(7):e18477 [FREE Full text] [doi: 10.2196/18477] [Medline: 32706670]
35. Reinforcement learning to optimize ventilator settings for patients on invasive mechanical ventilation: a retrospective study. GitHub. URL: https://github.com/nus-mornin-lab/Mechanical_ventilation_with_RL [accessed 2024-09-07]

Abbreviations

AI: artificial intelligence
BCQ: Batch Constrained Deep Q-learning
CWPDIS: Consistent Weighted Per-Decision Important Sampling
EHR: electronic health record
eICU: e-Intensive Care Unit Collaborative Research
FiO2: fraction of inspired oxygen
ICU: intensive care unit
MBP: mean arterial blood pressure
MDP: Markov Decision Process
MIMIC-IV: Medical Information Mart for Intensive Care
OPE: off-policy estimation
PaO2: partial pressure of oxygen
PEEP: positive end-expiratory pressure
RCT: randomized controlled trial
RL: reinforcement learning
SOFA: sequential organ failure assessment
SpO2: optimal oxygen saturation

Edited by T Leung, Y Li; submitted 21.11.22; peer-reviewed by SC Tan, F Balzer; comments to author 16.01.23; revised version received 24.02.23; accepted 18.08.24; published 16.10.24.

Please cite as:
Liu S, Xu Q, Xu Z, Liu Z, Sun X, Xie G, Feng M, See KC
Reinforcement Learning to Optimize Ventilator Settings for Patients on Invasive Mechanical Ventilation: Retrospective Study
J Med Internet Res 2024;26:e44494
URL: https://www.jmir.org/2024/1/e44494
doi:10.2196/44494
PMID:39219230

©Siqi Liu, Qianyi Xu, Zhuoyang Xu, Zhuo Liu, Xingzhi Sun, Guotong Xie, Mengling Feng, Kay Choong See. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Implementing a Barcode Information Management System on Operating Room Staff: Comparative Study

Chia-Yen Li[1,2], MSN; Mei-Hui Huang[3], BSN; Yu-Shiue Lin[1,4*], MSN; Chi-Ming Chu[1,5,6,7,8,9*], PhD; Hsueh-Hsing Pan[1,3,10*], PhD

[1]Graduate Institute of Medical Sciences, National Defense Medical Center, Taipei, Taiwan
[2]Department of Planning and Management, Tri-Service General Hospital, Taipei, Taiwan
[3]Department of Nursing, Tri-Service General Hospital, Taipei, Taiwan
[4]Department of Anesthesia, Tri-Service General Hospital, Taipei, Taiwan
[5]Graduate Institute of Life Sciences, School of Public Health, National Defense Medical Center, Taipei, Taiwan
[6]Big Data Research Center, College of Medicine, Fu-Jen Catholic University, New Taipei City, Taiwan
[7]Department of Public Health, China Medical University, Taichung, Taiwan
[8]Department of Healthcare, Department of Healthcare Administration and Medical Informatics, College of Health Sciences, Kaohsiung Medical University, Kaohsiung, Taiwan
[9]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan
[10]School of Nursing, National Defense Medical Center, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Hsueh-Hsing Pan, PhD
School of Nursing
National Defense Medical Center
No 161, Section 6, Minquan East Road
Neihu District
Taipei, 114201
Taiwan
Phone: 886 2 87923100 ext 18166
Email: pshing2001@gmail.com

Abstract

Background: Barcode information management systems (BIMS) have been implemented in operating rooms to improve the quality of medical care and administrative efficiency. Previous research has demonstrated that the Agile development model is extensively used in the development and management of information systems. However, the effect of information systems on staff acceptance has not been examined within the context of clinical medical information management systems.

Objective: This study aimed to explore the effects and acceptance of implementing a BIMS in comparison to the original information system (OIS) among operating and supply room staff.

Methods: This study was a comparative cohort design. A total of 80 staff members from the operating and supply rooms of a Northern Taiwan medical center were recruited. Data collection, conducted from January 2020 to August 2020 using a mobile-based structured questionnaire, included participant characteristics and the Information Management System Scale. SPSS (version 20.0, IBM Corp) for Windows (Microsoft Corporation) was used for data analysis. Descriptive statistics included mean, SD, frequency, and percentage. Differences between groups were analyzed using the Mann-Whitney *U* test and Kruskal-Wallis test, with a *P* value <.05 considered statistically significant.

Results: The results indicated that the BIMS generally achieved higher scores in key elements of system success, system quality, information quality, perceived system use, perceived ease of use, perceived usefulness, and overall quality score; none of these differences were statistically significant ($P>.05$), with the system quality subscale being closest to significance ($P=.06$). Nurses showed significantly better perceived system use than technicians (1.58, SD 4.78 vs –1.19, SD 6.24; $P=.02$). Significant differences in perceived usefulness were found based on educational level ($P=.04$) and experience with OIS ($P=.03$), with junior college-educated nurses and those with over 6 years of OIS experience reporting the highest perceived usefulness.

XSL•FO
RenderX

Conclusions: The study demonstrates that using the Agile development model for BIMS is advantageous for clinical environments. The high acceptance among operating room staff underscores its practicality and broader adoption potential. It advocates for continued exploration of technology-driven solutions to enhance health care delivery and optimize clinical workflows.

(J Med Internet Res 2024;26:e56192) doi:10.2196/56192

KEYWORDS

barcode information management system; barcode; information system; operation management information system; Agile development model

Introduction

The implementation of barcode information management systems (BIMS) in clinical practice is widely recognized as a method to enhance health care quality and safety [1]. BIMS replaces manual processes, thereby improving efficiency, quality management, and reducing costs [2]. They are used for patient identification, medication administration, and specimen transportation, contributing to reduced nursing turnover and increased patient satisfaction [3-5]. As health care quality becomes more important, organizations are upgrading management practices.

Despite the initial costs and potential staff resistance, technologies such as barcodes have demonstrated significant improvements in hospital services [6]. Health information technology (HIT) is significant for effectively managing and improving medical care quality, reducing costs, and addressing various challenges. HIT encompasses a range of technologies used to manage health information, including electronic health records, clinical decision support systems, and BIMS [7,8]. Effective HIT systems ensure that different technologies can communicate and share data seamlessly, which is critical for coordinated and efficient care [9,10]. However, challenges associated with HIT include cost, interoperability, privacy and security, data quality, and workflow integration [7,8].

The advantages of HIT are well-documented, including the ease of reading and interpreting physiological values; reduced time for completing electronic records; and decreased errors in drug administration, blood drawing, and specimen transportation. Consequently, HIT can improve patient safety, enhance nursing efficiency, and increase both nurse and patient satisfaction [3,4,11]. A previous study has shown that barcode systems can reduce surgical instrument packaging time and costs [12], and further research has demonstrated that HIT advances the safety and quality of surgical care, reduces staff stress, and enhances service quality [3,13].

Operating rooms are high-cost, high-skill units where complex and critical issues can significantly impact patient safety. It is essential for operating room nurses to maintain a positive view of patient care, ensure the physical safety of patients, and consider patient vulnerabilities [14]. The quality and efficiency of services in operating rooms are of paramount concern, as surgeries are invasive procedures that require stringent sterilization of medical equipment to prevent nosocomial infections. Such infections can seriously compromise patient safety, prolong hospitalization, and increase medical costs, thereby affecting a hospital's reputation [11]. The National Healthcare Safety Network [15] reported that surgical site infections are the costliest health care–associated infections, accounting for 20% of such infections and increasing the risk of death significantly. Improving surgical safety is one of Taiwan's patient safety goals, aimed at enhancing high-quality medical care and avoiding unnecessary patient harm [16].

In the software industry, Agile methods are widely adopted for efficient product and service development [17]. Agile software development (ASD) has become the primary approach for managing information systems implementation, which is crucial for modern digital health software. ASD facilitates customer feedback and revisions, emphasizing teamwork and collective decision-making in cross-functional teams [18,19]. The Agile transition process can be modified to adapt to the original standard process but may face challenges primarily due to human factors [19,20]. Issues such as a lack of understanding of the development model, direct implementation without comprehension, adherence to familiar practices, and disagreement with development values can lead to resistance and a failure to consider the entire process [19].

The relationship between BIMS, HIT, and ASD is interconnected, with each component playing a crucial role in modernizing and optimizing health care delivery. BIMS, as a part of HIT, relies on the robust infrastructure provided by HIT systems to ensure seamless integration and effective data use [21,22]. ASD supports the development and refinement of BIMS by emphasizing flexibility, user feedback, and iterative improvements, ensuring that BIMS solutions are continuously optimized based on real-world usage and feedback from health care staff [23]. Effective assessment is crucial when using ASD in operational engineering, as it helps identify team issues and facilitates improvements to enhance user experience. In the context of surgery, information systems manage logistics and materials, involving various professionals such as nurses and technicians. Therefore, this study explores the effects of implementing ASD with BIMS on operating room staff, providing insights into operational management and the comparative impact on staff experiences and efficiencies.

Methods

Study Design, Setting, and Participants

This study uses a comparative cohort design to evaluate the effects of implementing a BIMS on operating room staff, including nurses and technicians. The study was conducted between January 2020 and October 2020.

Participants were recruited from the operating and supply rooms of a medical center in Northern Taiwan. A total of 126 staff

XSL•FO
RenderX

members, including nurses and technicians responsible for handling sterile device materials or packaging and checking expiration dates, were eligible to participate. Therefore, a census approach was used for this group. Ultimately, 80 staff members, comprising 48 nurses and 32 technicians, agreed to participate and were included in the analysis.

Operation Management Information System

In surgical procedures, various clinical personnel collaborate to ensure seamless operations. The supply room, a vital medical logistics unit, manages the cleaning, packaging, sterilization, storage, and supply of medical devices, including expiration date management. Staffed by nurses and technicians, this unit ensures adherence to infection control principles for patient safety and efficient surgical procedures. The introductions of the original information system (OIS) and BIMS was as follows.

Original Information System

Under the OIS, the outpatient and emergency information systems are not connected to the surgical information system, requiring doctors to manually enter the surgical schedule repeatedly. The day before surgery, an operating room nurse selects the surgery case cart, specialized equipment, and packaging based on the surgery requirements. When additional packages, specialized tools, equipment, or materials are needed during surgery, nurses must select items and quantities in the original system or generate a handwritten list. They then confirm the material dynamics of temporary selections with the supply room by phone. The lack of immediate access to packaging sterilization information and inventory levels necessitates the estimation of item locations and sterilization processes through paper records, resulting in significant time spent tracking for the supply room.

Barcode Information Management System

The BIMS significantly enhances these operations. A doctor imports the surgery schedule and selects the specialist package and unique device identification system final rule. The day before surgery, an operating room nurse confirms the patient's case cart, special package, and required materials. During surgery, if additional packages or specialized tools are needed, nurses use BIMS to select the necessary items and quantities. Technicians then receive urgent messages from BIMS, acquire the supplies, and deliver them to the operating room.

BIMS streamlines communication and tracking, allowing nurses to monitor material processing status conveniently without needing to write application forms or confirm over the phone. It provides detailed information about sterilization equipment and the usage history of sterilized packages, enabling instant tracking of item locations and sterilization processes, along with managing package validity periods. This eliminates the need for phone confirmation with the supply room. If a system warning indicates insufficient supply allocation and the technician cannot resolve it directly, it is reported back to nurses. They use the system to understand device status dynamics and coordinate the process with the clinician.

Nurses oversee sterilization quality control, communicate with clinical units, and guide technical staff. They coordinate staff, manage instruments, handle supply alerts, assist in surgeries, provide patient care, and ensure the sterilization of equipment. Technicians are responsible for cleaning, packaging, sterilizing, and supplying instruments. This collaboration, facilitated by BIMS, is crucial for maintaining efficient and safe surgical procedures. Figure 1 describes the comparison between OIS and BIMS in the surgical procedure.

XSL•FO
RenderX

Figure 1. Comparison between the original information system and barcode information management system.

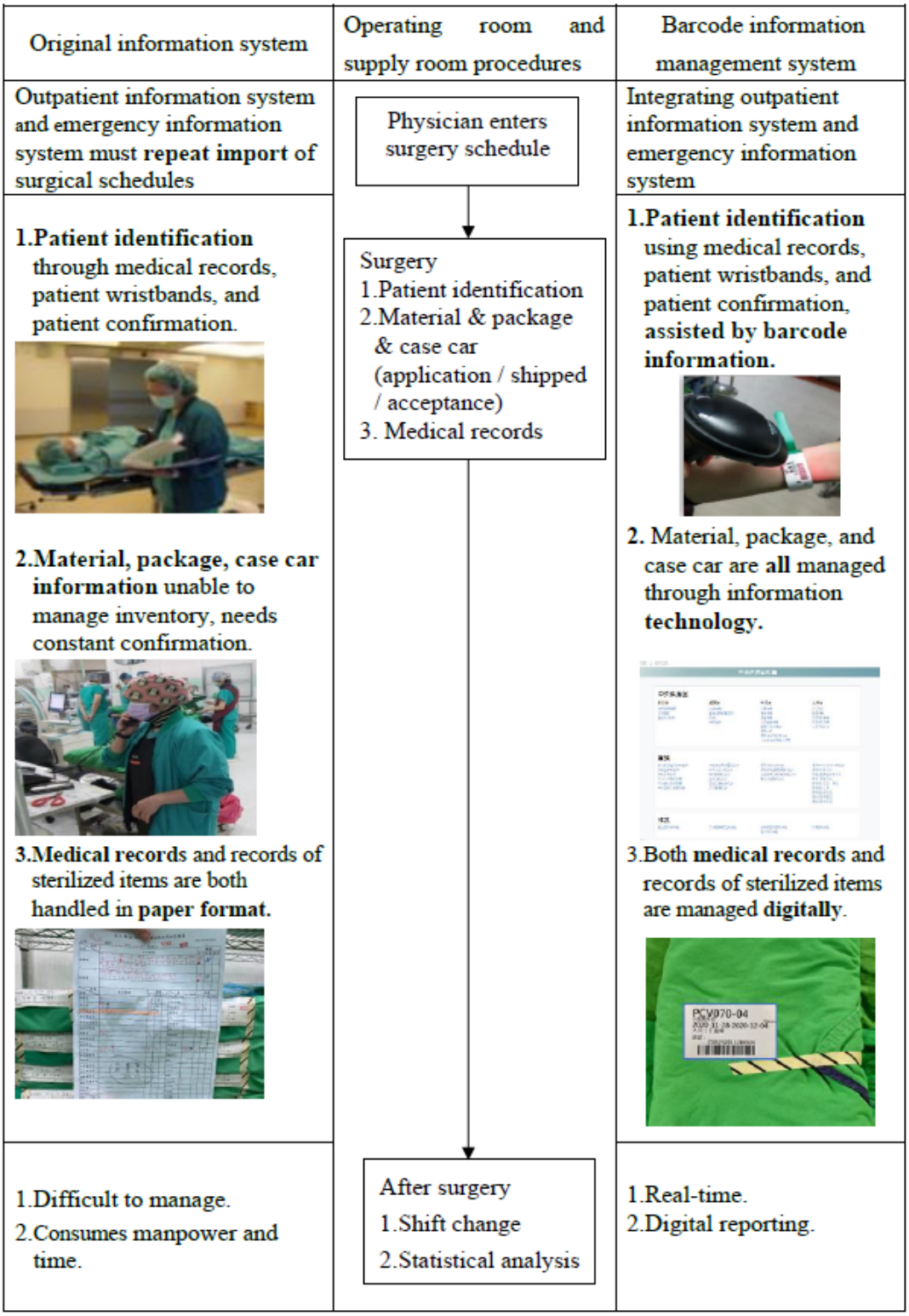

Measurements

Characteristics of Participants

The characteristics of the participants included the following variables: identity, sex, and age (categorized into 3 groups: young [19-34 years], early middle-aged [35-49 years], and late middle-aged adults [50-65 years]) [24], educational level (below senior high school, specialist, and junior college), working years, and experience using the original system.

Information Management System Scale

The Information Management System Scale (IMSS) developed by Chiang and Lee [25] was used to evaluate the effectiveness of the nursing staff information system. The IMSS comprises 28 questions divided into 3 main parts.

The "key elements of system success" section includes 2 subscales: system quality, with 7 questions focusing on screen simplicity, data classification, reading speed, response time,

XSL•FO
RenderX

screen transition, function availability, and patient care information integration; and information quality, with 7 questions on team communication efficiency, time savings, information accuracy, and display quality.

The "perceived system use" section also has 2 subscales: perceived ease of use, with 5 questions on system operation, use, proficiency attainment, and the system's significance in improving clinical work efficiency and team connectivity; and perceived usefulness, with 5 questions on the system's convenience in clinical work, control of patient conditions, work quality improvement, procedure simplification, and usefulness of patient information.

The "attitude toward system use" section includes 4 questions assessing the information system as a valuable tool for the medical team and its role in improving care quality.

A Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) was used, with higher scores indicating greater satisfaction, efficiency improvement, and positive attitudes. The scale was tested with 185 nurses, achieving an overall Cronbach α of 0.973, with subscale Cronbach α values between 0.897 and 0.920 [25]. This study invited an information supervisor, an information engineer, and a senior nurse to validate the assessments on a scale of 1 to 5 for correctness, suitability, and completeness of the questionnaire contents, resulting in a content validity index of 0.94. We recruited 80 participants for this study, achieving an overall Cronbach α of 0.984, with each subscale having a Cronbach α ranging from 0.925 to 0.969, indicating high reliability and validity.

Statistical Analysis

The SPSS (version 20; IBM Corp) software for Windows (Microsoft Corporation) was used to analyze the data. Continuous variables were presented as mean and SD, while categorical variables were presented as frequencies and percentages. The Mann-Whitney *U* test and Kruskal-Wallis test were used for univariate analysis to compare differences in participants' data. A *P* value of <.05 was considered statistically significant.

Ethical Considerations

This study was approved by the Institutional Review Board of Tri-Services General Hospital (No. C202005035). The purpose of the study was explained by the same researcher to the operating room and supply room staff of a medical center in northern Taiwan during morning meetings. Out of 126 eligible participants, 80 agreed and signed the informed consent to participate in the study. Data was collected both before (using OIS) and after the implementation of the BIMS. Baseline data was collected in January 2020, and post-implementation data were collected in August 2020. This approach allowed for a comprehensive comparison of the effects of BIMS on the operating and supply room staff, including nurses and technicians. The details are shown in Figure 2.

Figure 2. Flowchart of the study population.

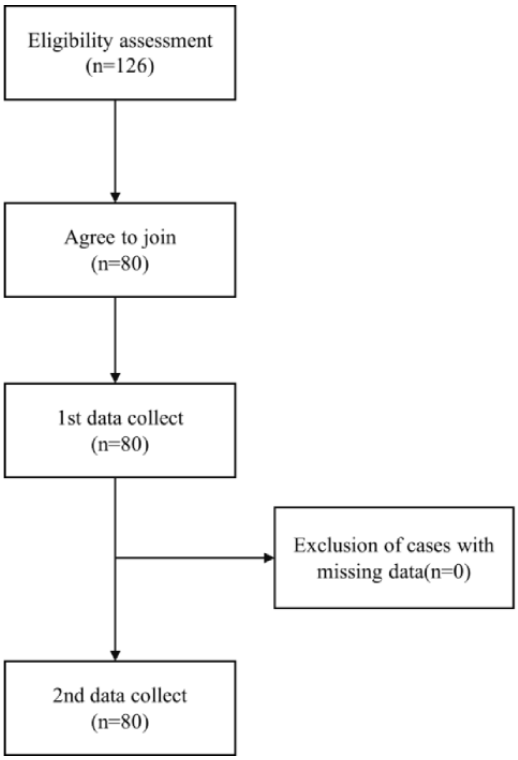

Results

Characteristics of Participants

A total of 80 staff members from the operating and supply rooms participated in the study, comprising 42 nurses (60%) and 38 technicians (40%). The mean age of the participants was 47.1 (SD 10.4) years. Most participants were female (n=64, 80%) and had a bachelor's degree or above (n=34, 42%). In addition, 46% had over 20 years of service experience, and 71% (n=57) had more than 6 years of experience using the OIS. Table 1 presents these details.

XSL•FO
RenderX

Table 1. Characteristics of participants (N=80).

Variables	Participants
Identity, n (%)	
Nurse	48 (60)
Technician	32 (40)
Sex n (%)	
Male	16 (20)
Female	64 (80)
Age (years), mean (SD)	47.1 (10.4)
Young adults (19-34), n (%)	8 (10)
Early middle-aged adults (35-49), n (%)	39 (49)
Late middle-aged adults (50-65), n (%)	33 (41)
Educational level, n (%)	
Below senior high school	24 (30)
Junior college	22 (28)
Bachelor and above	34 (42)
Length of service year (years), mean (SD)	18.3 (12)
<10, n (%)	25 (31)
10-20, n (%)	18 (23)
>20, n (%)	37 (46)
Experience of using original information systems (years), n (%)	
<3	19 (24)
3-6	4 (5)
>6	57 (71)

Difference Between Original Information System and Barcode Information Management Systems in Information Management System Scale Scores

The results, as presented in Table 2, show that the BIMS generally achieved higher scores compared with the OIS across various metrics. While BIMS showed improvements in key elements of system success, system quality, information quality, perceived system use, perceived ease of use, perceived usefulness, and overall quality score, none of these differences were statistically significant ($P>.05$). The closest to significance was the system quality subscale ($P=.06$). Attitudes toward system use were similarly positive for both systems. Overall, BIMS demonstrated a trend toward better performance, but without significant statistical differences from OIS.

XSL•FO
RenderX

Table 2. Comparison of original information system and barcode information management systems Information Management System Scale scores (N=80).

Variable (item)	Original information system		Barcode information management systems		*z*	*P* value
	Total, mean (SD)	Average, mean (SD)	Total, mean (SD)	Average, mean (SD)		
Informational Management System Scale						
Key elements of system success	48.51 (9.96)	3.46 (0.71)	49.88 (8.55)	3.56 (0.61)	–1.38	.17
System quality (7)	24.13 (4.78)	3.45 (0.68)	25.01 (4.11)	3.57 (0.58)	–1.92	.06
Information quality (7)	24.37 (5.38)	3.48 (0.77)	24.87 (4.57)	3.55 (0.65)	–0.81	.42
Perceived system use	35.46 (7.11)	3.54 (0.71)	35.94 (5.76)	3.59 (0.57)	–0.79	.43
Perceived ease of use (5)	17.71 (3.69)	3.54 (0.73)	17.93 (3.04)	3.58 (0.60)	–0.58	.56
Perceived usefulness (5)	17.75 (3.63)	3.46 (0.71)	18.02 (2.91)	3.60 (0.58)	–0.81	.42
Attitudes of system use (4)	18.25 (3.74)	4.56 (0.94)	17.96 (3.42)	4.49 (0.86)	–0.58	.56
Overall quality score	102.23 (19.83)	3.65 (0.71)	103.78 (17.15)	3.71 (0.61)	–0.85	.39

Difference Between Original Information System and Barcode Information Management Systems According to Participant Characteristics, Key Elements of System Success, Perceived System Use, and Attitude Toward System Use

The difference between the OIS and BIMS was calculated by deducting the OIS score from the BIMS score across various participant characteristics and key elements of system success, perceived system use, and attitudes toward system use among 80 participants. A positive difference indicates a better information management system using BIMS. Significant differences were observed in participant identity and perceived system use, with nurses showing better perceived system use than technicians (1.58, SD 4.78 vs –1.19, SD 6.24; *P*=.02). No statistically significant differences were observed in gender, age, educational level, length of service, or experience in using an OIS (*P*>.05). Table 3 lists these details.

Table 4 shows the differences between the BIMS and OIS scores according to participant characteristics and the 2 subscales, perceived ease of use and perceived usefulness, of the perceived system use measure, were also investigated. The results revealed significant differences between the participant's identity and perceived usefulness. Nurses had better perceived the system's usefulness than technicians (0.85, SD 2.46 vs –0.63, SD 3.39; *P*=.03). No statistically significant differences were observed in the demographics, perceived ease of use, and perceived usefulness (*P*>.05).

Table 5 lists the differences in participant characteristics and perceived usefulness between nurses and technicians using BIMS and OIS. The results showed significant differences in perceived usefulness based on educational level (*P*=.04) and experience in using the OIS (*P*=.03). Junior college–educated nurses reported the highest perceived usefulness. In addition, nurses with more than 6 years of experience using the OIS reported higher perceived usefulness compared with technicians.

XSL•FO
RenderX

Table 3. Comparison of original information system and barcode information management systems scores for participant characteristics and key elements of system success, perceived system use, and attitude toward system use (N=80).

Variable	Key elements of system success[a]		Perceived system use[a]		Attitudes of system use[a]	
	Mean (SD)	*P* values	Mean (SD)	*P* values	Mean (SD)	*P* values
Identity		.14[b]		.02[b]		.19[b]
Nurse	2.27 (7.04)		1.58 (4.78)		0.12 (2.92)	
Technician	0.03 (9.98)		-1.19 (6.24)		-0.91 (4.11)	
Sex		.47[b]		.53[b]		.26[b]
Male	–0.19 (8.26)		–1.00 (6.25)		–1.31 (4.09)	
Female	1.77 (8.39)		0.84 (5.34)		–0.03 (3.27)	
Age		.71[c]		.30[c]		.90[c]
Young adults	2.75 (68.6)		1.75 (4.13)		0.13 (2.17)	
Early middle-aged adults	1.69 (7.30)		0.89 (5.56)		–0.62 (3.64)	
Late middle-aged adults	0.67 (9.87)		–0.33 (5.85)		0.00 (3.54)	
Educational level		.83[c]		.24[c]		.27[c]
Below senior high school	1.29 (10.88)		–0.75 (6.64)		–0.71 (4.53)	
Junior college	2.23 (6.99)		1.82 (4.56)		0.50 (2.81)	
Bachelor and above	0.88 (7.26)		0.47 (5.24)		–0.50 (2.97)	
Length of service years		.55[c]		.24[c]		.18[c]
<10 years	–0.16 (6.99)		–1.12 (5.53)		–1.36 (3.43)	
10-20 years	1.44 (8.35)		1.59 (5.62)		–0.26 (3.32)	
>20 years	2.68 (9.46)		0.82 (5.36)		0.64 (3.46)	
Experience of using original information system		.43[c]		.11[c]		.16[c]
<3 years	–0.58 (7.51)		–1.63 (5.91)		–1.68 (3.59)	
3-6 years	–1.00 (2.94)		–1.50 (2.38)		–1.25 (2.62)	
>6 years	2.19 (8.80)		1.32 (5.42)		0.24 (3.37)	

[a]Difference between original information system and barcode information management system.

[b]Mann-Whitney *U* test.

[c]Kruskal-Wallis test.

XSL•FO
RenderX

Table 4. Comparison of original information system and barcode information management system scores for participant characteristics and perceived system use, including perceived ease of use and perceived usefulness (N=80).

Variables	Perceived system use[a]			
	Perceived ease of use		Perceived usefulness	
	Mean (SD)	*P* values	Mean (SD)	*P* values
Identity		.09[b]		.03[b]
Nurse	0.73 (2.84)		0.85 (2.46)	
Technician	–0.56 (3.32)		–0.63 (3.39)	
Sex		.99[b]		.45[b]
Male	–0.18 (3.51)		–0.81 (3.27)	
Female	0.31 (2.99)		0.53 (2.82)	
Age (years)		.34[c]		.41[c]
Young adults	0.87 (2.99)		0.87 (1.45)	
Early middle-aged adults	0.54 (3.01)		0.36 (3.05)	
Late middle-aged adults	–0.33 (3.21)		0.00 (3.11)	
Educational level		.67[c]		.08[c]
Below senior high school	–0.33 (3.53)		–0.42 (3.59)	
Junior college	0.50 (2.81)		1.31 (2.15)	
Bachelor and above	0.41 (2.97)		0.06 (2.76)	
Length of service year		.39[c]		.36[c]
<10 years	–0.48 (3.20)		–0.64 (2.87)	
10-20 years	1.07 (2.97)		0.52 (3.03)	
>20 years	0.00 (3.01)		0.82 (2.83)	
Experience of using original information system		.29[c]		.12[c]
<3 years	–0.57 (3.33)		–1.05 (3.08)	
3-6 years	–1.50 (2.38)		0.00 (0.00)	
>6 years	0.59 (2.99)		0.72 (2.89)	

[a]Difference between original information system and barcode information management system.

[b]Mann-Whitney *U* test.

[c]Kruskal-Wallis test.

XSL•FO
RenderX

Table 5. Comparison of participant characteristics and perceived usefulness between nurses and technicians (N=80).

Variables	Perceived usefulness[a]		
	Nurse (n=48) mean (SD)	Technician (n=32) mean (SD)	*P* values
Sex			.08[b]
Male	0.00 (0.00)	–0.87 (3.37)	
Female	0.87 (2.48)	–0.41 (3.50)	
Age (years)			.08[c]
Young adults	1.00 (1.67)	0.50 (0.71)	
Early middle-aged adults	1.04 (2.69)	–1.17 (3.35)	
Late middle-aged adults	0.47 (2.36)	–0.39 (3.64)	
Educational level			.04[c]
Below senior high school	–0.50 (2.12)	–0.41 (3.73)	
Junior college	1.71 (2.23)	0.00 (1.22)	
Bachelor and above	0.45 (2.53)	–2.20 (3.27)	
Length of service year			.05[c]
<10 years	0.85 (1.57)	–1.22 (3.08)	
10-20 years	0.58 (2.76)	0.37 (3.81)	
>20 years	1.09 (2.48)	–0.16 (3.97)	
Experience of using original information system			.03[c]
<3 years	0.40 (0.89)	–1.57 (3.43)	
3-6 years	0.00 (0.00)	0.00 (0.00)	
>6 years	0.90 (2.58)	0.14 (3.73)	

[a]Difference between original information system and barcode information management system.

[b]Mann-Whitney U test.

[c]Kruskal-Wallis test.

Discussion

Principal Findings

This study used an ASD information framework to manage the implementation of a BIMS for operating and supply room staff, focusing on the management of surgical packages. While there was no statistically significant difference in the IMSS scores between the OIS and BIMS, the overall IMSS quality score of the BIMS was higher than that of the OIS. The results demonstrated that 71.2% of the participants had a substantial length of service and over 6 years of experience using the OIS. Previous research has shown that information systems with user-friendly interfaces and ease of operation are advantageous. These benefits should facilitate acceptance by staff in the working environment [26].

This study provided valuable insights into whether nurses perceived system use better than technicians. The findings revealed that nurses indeed had a better perception of system use compared with technicians. Previous research has indicated that Agile transformation is user-based, with attitudes, norms, and self-efficacy being key determinants of intention and healthy behavior [19,27]. These factors are primary concerns in people's perceptions of Agile transformation. Interventions based on self-determination theory in health settings have been shown to effectively promote the adoption and maintenance of health-related behaviors [28]. Nurses' better perception of system use compared with technicians can be attributed to their extensive experience, with most nurses in this study having over 20 years of service. In addition, the BIMS is user-friendly, with simplified operations and improved work efficiency. The system's traceability management of equipment packages also contributed to a stronger sense of identity in clinical care. This enhanced the willingness to use BIMS and encouraged proactive suggestions for system improvements to better meet actual demands.

Regarding perceived usefulness, the subgroup analysis between nurses and technicians showed that educational level and experience using an OIS were statistically significant factors. Effective software development must consider perceived system usefulness to motivate usage [29]. Therefore, the system should align with users' past experiences, needs, and educational levels to improve outcomes. Operating and supply room nurses are responsible for directing and supervising technicians in performing various sterilization techniques. Nurses involved in developing a surgical BIMS considered standard operating procedures and clinical practice experience. Given that most

XSL•FO
RenderX

staff in this study had over 20 years of service and were accustomed to the daily workflow, it was challenging for information engineers to modify the information program immediately during work changes, causing staff pressure. The HIT should improve efficiency and operator satisfaction by aligning workflow and equipment at a reasonable cost [30].

The use of information technology in medical material management significantly reduces staff workload. Acceptance of this technology is influenced by software and hardware availability, user age, and education level. Younger users, with better information skills due to recent education and widespread technology exposure, tend to be more accepting [31,32]. However, insufficient training, concerns about system security, and interdepartmental communication pressures can hinder usage. In addition, changes in workflow, competition for computer access, and extra time required to handle information can reduce interpersonal interactions, further obstructing implementation [32,33]. In the supply room, nurses and technicians collaboratively manage logistics and medical materials. Nurses are responsible for education, guidance, and quality control to ensure the safety of medical materials for patients. BIMS operational functions vary by authority level, with nurses using their system privileges to monitor instrument flow and supply levels, coordinating with clinical processes to enhance recognition and support of clinical care. Literature highlights the importance of defining human-machine collaboration, interactive empowerment, and a digital ecosystem [34]. A strategic management approach to digital quality management, guided by principles of humanistic care, is essential. Continuous development of digital technologies aims to achieve optimal care quality [34,35]. Nursing managers, who also serve as educators, communicators, and supervisors, facilitated the smooth implementation of Agile transition processes through continuous team discussions, workflow adjustments, and ongoing staff education and training.

Limitations and Recommendations

This study had several limitations. First, a self-reported structured questionnaire was used to measure the effectiveness of the information system. Second, participants were recruited from the operating and supply rooms of a single medical center. Finally, the completion rate was 63%. Based on these limitations, the following recommendations are presented. First, objective instruments should be used to measure the effectiveness of the information system, and qualitative interviews can be conducted to better understand user experience. Second, participants from multiple institutions should be recruited to enhance the external validity of the findings. Third, various strategies should be implemented to increase the completion rates of surveys and questionnaires. Finally, while this study collected quantitative data from nurses and technicians, it is recommended to gather qualitative data in a subsequent phase to validate and complement these findings.

Conclusions

This study developed a BIMS to replace handwritten information with barcode data. A traceability information management system was established for surgical instruments and packages, facilitating cooperation among the operating room, supply room, software information, and infection control teams to formulate information-based operating specifications. The study re-examined and refined the Agile transformation process through interteam communication. In addition, staff education was emphasized to enhance the acceptability and efficiency of BIMS.

Acknowledgments

We would like to thank the surgical team staff and technicians who participated in this study, with special appreciation to Dr Chia-Cheng Lee for his valuable contributions. This study was supported by research grants from Tri-Service General Hospital in 2020 (TSGH-D-109156, TSGH-E-112246).

Authors' Contributions

CYL contributed to conceptualization, methodology, data curation, formal analysis, investigation, and writing of the original draft. MHH contributed to conceptualization, methodology, and data curation. YSL handled methodology and data curation. CMC managed conceptualization, data curation, and original draft preparation. HHP contributed to conceptualization, methodology, data curation, and original draft preparation.

Conflicts of Interest

None declared.

References

1. Alotaibi YK, Federico F. The impact of health information technology on patient safety. Saudi Med J 2017;38(12):1173-1180 [FREE Full text] [doi: 10.15537/smj.2017.12.20631] [Medline: 29209664]
2. Goh MM, Tan ABH, Leong MHW. Bar code-based management to enhance efficiency of a sterile supply unit in Singapore. AORN J 2016;103(4):407-413. [doi: 10.1016/j.aorn.2016.01.018] [Medline: 27004503]
3. Lee T, Sun G, Kou L, Yeh M. The use of information technology to enhance patient safety and nursing efficiency. Technol Health Care 2017;25(5):917-928. [doi: 10.3233/THC-170848] [Medline: 28826193]
4. Hanna MG, Pantanowitz L. Bar coding and tracking in pathology. Surg Pathol Clin 2015;8(2):123-135. [doi: 10.1016/j.path.2015.02.017] [Medline: 26065787]

XSL•FO
RenderX

5. Berdot S, Boussadi A, Vilfaillot A, Depoisson M, Guihaire C, Durieux P, et al. Integration of a commercial barcode-assisted medication dispensing system in a teaching hospital. Appl Clin Inform 2019;10(4):615-624 [FREE Full text] [doi: 10.1055/s-0039-1694749] [Medline: 31434161]
6. Ehteshami A. Barcode technology acceptance and utilization in health information management department at academic hospitals according to technology acceptance model. Acta Inform Med 2017;25(1):4-8 [FREE Full text] [doi: 10.5455/aim.2017.25.4-8] [Medline: 28484289]
7. Park Y, Kim YS, Yi B, Kim SM. Clinical decision support functions and digitalization of clinical documents of electronic medical record systems. Healthc Inform Res 2019;25(2):115-123 [FREE Full text] [doi: 10.4258/hir.2019.25.2.115] [Medline: 31131146]
8. Abdelaziz T, Rosa EM. Electronic health records with decision support systems for sharper diagnoses: bibliometric analysis. Int J Adv Appl Sci 2024;13(2):411 [FREE Full text]
9. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023;23(1):18 [FREE Full text] [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
10. Steichen O, Gregg W. Health information technology coordination to support patient-centered care coordination. Yearb Med Inform 2015;10(1):34-37 [FREE Full text] [doi: 10.15265/IY-2015-027] [Medline: 26293848]
11. Chu J, Hsieh CH, Shih YN, Wu CC, Singaravelan A, Hung LP, et al. Operating room usage time estimation with machine learning models. Healthcare (Basel) 2022;10(8):1518 [FREE Full text] [doi: 10.3390/healthcare10081518] [Medline: 36011177]
12. Wu SL, Liu YC, Shih WH, Wu SC, Lee HF, Lin CT. [Computerized barcode operational system for package of surgical instruments in operating room]. Hu Li Za Zhi 2008;55(5):56-63. [Medline: 18836975]
13. Ellner SJ, Joyner PW. Information technologies and patient safety. Surg Clin North Am 2012;92(1):79-87. [doi: 10.1016/j.suc.2011.11.002] [Medline: 22269262]
14. Bastami M, Imani B, Koosha M. Operating room nurses' experience about patient cares for laparotomy surgeries: a phenomenological study. J Family Med Prim Care 2022;11(4):1282-1287 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_1085_21] [Medline: 35516685]
15. Surgical site infection event (SSI). Centers for Disease Control and Prevention. 2024 Feb 02. URL: https://www.cdc.gov/nhsn/pdfs/pscmanual/9pscssicurrent.pdf [accessed 2024-08-09]
16. Patient safety goals. Joint Commission of Taiwan. 2022. URL: https://www.jct.org.tw/cp-1351-8370-3fe36-2.html [accessed 2024-08-09]
17. Ahmed A, Ahmad S, Ehsan N, Mirza E, Sarwar SZ. Agile software development: impact on productivity and quality. 2010 Presented at: 2010 IEEE International Conference on Management of Innovation & Technology; June 2-5, 2010; Singapore p. 287-291 URL: https://www.semanticscholar.org/paper/Agile-software-development%3A-Impact-on-productivity-Ahmed-Ahmad/c1af452e58200cab7102aa2ba54d11f2e4a7c8a0
18. Kokol P, Blažun Vošner H, Kokol M, Završnik J. Role of agile in digital public health transformation. Front Public Health 2022;10:899874 [FREE Full text] [doi: 10.3389/fpubh.2022.899874] [Medline: 35646754]
19. Javdani Gandomani T, Ziaei Nafchi M. Agile transition and adoption human-related challenges and issues: a grounded theory approach. Comput Human Behav 2016;62:257-266. [doi: 10.1016/j.chb.2016.04.009]
20. Pikkarainen M, Salo O, Kuusela R, Abrahamsson P. Strengths and barriers behind the successful agile deployment—insights from the three software intensive companies in Finland. Empir Software Eng 2011 Oct 27;17(6):675-702. [doi: 10.1007/s10664-011-9185-5]
21. Hachesu PR, Zyaei L, Hassankhani H. Recommendations for using barcode in hospital process. Acta Inform Med 2016;24(3):206-210 [FREE Full text] [doi: 10.5455/aim.2016.24.206-210] [Medline: 27482137]
22. He Q, Bao M, Hass K, Lin W, Qin P, Du K. Perspective of molecular diagnosis in healthcare: from barcode to pattern recognition. Diagnostics (Basel) 2019;9(3):75 [FREE Full text] [doi: 10.3390/diagnostics9030075] [Medline: 31337082]
23. Barros L, Tam C, Varajão J. Agile software development projects–Unveiling the human-related critical success factors. Inf Softw Technol 2024;170:107432. [doi: 10.1016/j.infsof.2024.107432]
24. Franssen T, Stijnen M, Hamers F, Schneider F. Age differences in demographic, social and health-related factors associated with loneliness across the adult life span (19-65 years): a cross-sectional study in the Netherlands. BMC Public Health 2020;20(1):1118 [FREE Full text] [doi: 10.1186/s12889-020-09208-0] [Medline: 32758200]
25. Chiang YC, Lee TT. The effectiveness of the use by nurses of the interdisciplinary care information system. Hu Li Za Zhi 2018;65(4):49-59. [doi: 10.6224/JN.201808_65(4).08] [Medline: 30066323]
26. Chao YL, Lee TT. A study of the satisfaction of nurses with the clinical information system in intensive care units. J Nurs Healthc Res 2015;11(2):109-118.
27. Sheeran P, Maki A, Montanaro E, Avishai-Yitshak A, Bryan A, Klein WMP, et al. The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behavior: a meta-analysis. Health Psychol 2016;35(11):1178-1188. [doi: 10.1037/hea0000387] [Medline: 27280365]

XSL•FO
RenderX

28. Teixeira PJ, Marques MM, Silva MN, Brunet J, Duda JL, Haerens L, et al. A classification of motivation and behavior change techniques used in self-determination theory-based interventions in health contexts. Motivation Science 2020;6(4):438-455. [doi: 10.1037/mot0000172]
29. Al-Bashayreh M, Almajali D, Altamimi A, Masa'deh R, Al-Okaily M. An empirical investigation of reasons influencing student acceptance and rejection of mobile learning apps usage. Sustainability 2022;14(7):4325. [doi: 10.3390/su14074325]
30. Pahontu R, Will A, Bergh B. Towards communication requirements in the operating room and clinic IT. Stud Health Technol Inform 2015;212:1-8. [Medline: 26063250]
31. Chen YT, Chiu YC, Teng ML, Liao PH. The effect of medical material management system app on nursing workload and stress. BMC Nurs 2022;21(1):19 [FREE Full text] [doi: 10.1186/s12912-022-00806-4] [Medline: 35039036]
32. Lee TT. Nursing information: users' experiences of a system in Taiwan one year after its implementation. J Clin Nurs 2008;17(6):763-771. [doi: 10.1111/j.1365-2702.2007.02041.x]
33. Arboit É, Freitas EO, Balsanelli AP, da Silva RM, Camponogara S. Work intensification from nursing workers' perspective. Texto contexto - enferm 2023;32(6). [doi: 10.1590/1980-265x-tce-2023-0146en]
34. Han L, Bai X, Zhang H, An N. Digital nursing quality management: a concept analysis. medRxiv. Preprint posted online on July 17, 2023 2023.
35. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023;20(4):3407 [FREE Full text] [doi: 10.3390/ijerph20043407] [Medline: 36834105]

Abbreviations

ASD: Agile software development
BIMS: barcode information management systems
HIT: health information technology
IMSS: Information Management System Scale
IRB: Institutional Review Board
OIS: original information system

Edited by G Tsafnat; submitted 26.01.24; peer-reviewed by O Ntsweng, Z Maleki; comments to author 18.05.24; revised version received 30.06.24; accepted 07.07.24; published 17.10.24.

Please cite as:
Li CY, Huang MH, Lin YS, Chu CM, Pan HH
Effects of Implementing a Barcode Information Management System on Operating Room Staff: Comparative Study
J Med Internet Res 2024;26:e56192
URL: https://www.jmir.org/2024/1/e56192
doi:10.2196/56192
PMID:

©Chia-Yen Li, Mei-Hui Huang, Yu-Shiue Lin, Chi-Ming Chu, Hsueh-Hsing Pan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Establishing Medical Intelligence—Leveraging Fast Healthcare Interoperability Resources to Improve Clinical Management: Retrospective Cohort and Clinical Implementation Study

Alexander Brehmer[1*], MSc; Christopher Martin Sauer[1,2*], MD, MPH, PhD; Jayson Salazar Rodríguez[1*], MSc; Kelsey Herrmann[1], MD; Moon Kim[1], MD; Julius Keyl[1,3], MD; Fin Hendrik Bahnsen[1], MSc; Benedikt Frank[4], MD; Martin Köhrmann[4], Prof Dr Med; Tienush Rassaf[5], Prof Dr Med; Amir-Abbas Mahabadi[5], MD; Boris Hadaschik[6], MD; Christopher Darr[6], MD; Ken Herrmann[7], Prof Dr Med; Susanne Tan[8], Prof Dr Med; Jan Buer[9], Prof Dr Med; Thorsten Brenner[10], Prof Dr Med; Hans Christian Reinhardt[2], Prof Dr Med; Felix Nensa[1], PhD, Prof Dr Med; Michael Gertz[11], Prof Dr; Jan Egger[1], PhD; Jens Kleesiek[1], PhD, Prof Dr Med

[1]Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany

[2]Deptartment of Hematology and Stem Cell Transplantation, West German Cancer Center, German Cancer Consortium Partner Site Essen, Center for Molecular Biotechnology, University Hospital Essen, Essen, Germany

[3]Institute of Pathology, University Hospital Essen, Essen, Germany

[4]Department of Neurology, University Hospital Essen, Essen, Germany

[5]Department of Cardiology and Vascular Medicine, West German Heart- and Vascular Center, University Hospital Essen, Essen, Germany

[6]Department of Urology and German Cancer Consortium (DKTK) Partner Site, University Hospital Essen, Essen, Germany

[7]Department of Radiotherapy, University Hospital Essen, Essen, Germany

[8]Department of Endocrinology, Diabetes and Metabolism, University Hospital Essen, Essen, Germany

[9]Institute of Medical Microbiology, University Hospital Essen, Essen, Germany

[10]Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen, Essen, Germany

[11]Institute of Computer Science, Heidelberg University, Heidelberg, Germany

[*]these authors contributed equally

Corresponding Author:
Jens Kleesiek, PhD, Prof Dr Med
Institute for Artificial Intelligence in Medicine
University Hospital Essen
Hufelandstr. 55
Essen, 45147
Germany
Phone: 49 0201723 ext 77801
Email: jens.kleesiek@uk-essen.de

Abstract

Background: FHIR (Fast Healthcare Interoperability Resources) has been proposed to enable health data interoperability. So far, its applicability has been demonstrated for selected research projects with limited data.

Objective: This study aimed to design and implement a conceptual medical intelligence framework to leverage real-world care data for clinical decision-making.

Methods: A Python package for the use of multimodal FHIR data (FHIRPACK [FHIR Python Analysis Conversion Kit]) was developed and pioneered in 5 real-world clinical use cases, that is, myocardial infarction, stroke, diabetes, sepsis, and prostate cancer. Patients were identified based on the *ICD-10* (*International Classification of Diseases, Tenth Revision*) codes, and outcomes were derived from laboratory tests, prescriptions, procedures, and diagnostic reports. Results were provided as browser-based dashboards.

Results: For 2022, a total of 1,302,988 patient encounters were analyzed. (1) Myocardial infarction: in 72.7% (261/359) of cases, medication regimens fulfilled guideline recommendations. (2) Stroke: out of 1277 patients, 165 received thrombolysis and 108 thrombectomy. (3) Diabetes: in 443,866 serum glucose and 16,180 glycated hemoglobin A_{1c} measurements from 35,494

XSL•FO
RenderX

unique patients, the prevalence of dysglycemic findings was 39% (13,887/35,494). Among those with dysglycemia, diagnosis was coded in 44.2% (6138/13,887) of the patients. (4) Sepsis: In 1803 patients, *Staphylococcus epidermidis* was the primarily isolated pathogen (773/2672, 28.9%) and piperacillin and tazobactam was the primarily prescribed antibiotic (593/1593, 37.2%). (5) PC: out of 54, three patients who received radical prostatectomy were identified as cases with prostate-specific antigen persistence or biochemical recurrence.

Conclusions: Leveraging FHIR data through large-scale analytics can enhance health care quality and improve patient outcomes across 5 clinical specialties. We identified (1) patients with sepsis requiring less broad antibiotic therapy, (2) patients with myocardial infarction who could benefit from statin and antiplatelet therapy, (3) patients who had a stroke with longer than recommended times to intervention, (4) patients with hyperglycemia who could benefit from specialist referral, and (5) patients with PC with early increases in cancer markers.

(J Med Internet Res 2024;26:e55148) doi:10.2196/55148

KEYWORDS

clinical informatics; FHIR; real-world evidence; medical intelligence; interoperability; data exchange; clinical management; clinical decision-making; electronic health records; quality of care; quality improvement

Introduction

Electronic health records (EHRs) were developed in the 1970s and have evolved and grown in usage over the years [1,2]. An important milestone in the use of routinely collected clinical data was the publication of Medical Information Mart for the Intensive Care (MIMIC-III). Now in its fourth version, it was the first, large, granular publicly available patient-level dataset and includes more than 50,000 unique intensive care unit patients [3,4]. While numerous other publicly available intensive care unit databases have recently been released, the use of hospital-wide EHR data is strikingly less common [5]. More recently, implementation of EHRs has been mandated in the United States and they are the accepted standard in many other countries [6,7].

Although there have been many improvements in the EHRs, various challenges using EHR data remain. The data are large-scale (a large volume), heterogeneous (drawn from different resources), multimodal (multiple data modalities), temporal (collected over time), high-dimensional (thousands of distinct medical events), and often uncurated (not carefully chosen and thoughtfully organized or presented), poor quality (rarely subject to data quality audits), sparse (many zero values), and incomplete (missing values) [8]. Furthermore, not all patient data are located within a unified EHR. Even within a single hospital, the information is distributed across many primary systems, for example, laboratory, pathology, or medication systems—to name just a few. Within an average hospital, the number of these primary systems can easily add up to hundreds.

FHIR (Fast Healthcare Interoperability Resources) has been proposed to unify information exchange. It is a standardized set of application programming interfaces that enables the secure and efficient exchange of health care data across platforms and operating systems [9]. It is rapidly gaining traction in the health care industry and is being adopted by a growing number of health care organizations around the world. FHIR has emerged as the de facto standard for interoperability in health care, such that it is anchored within the Office of the National Coordinator for Health Information Technology 21st Century Cures Act [10].

Using FHIR to exchange medical data may provide potential benefits in many different areas, such as clinical decision support, precision medicine, mobile health apps, wearable devices, big data analytics, and using EHRs for clinical research [11]. Current reviews demonstrate that its application thus far is limited to (clinical) research with selected narrow use cases [12-14].

While FHIR holds the potential to standardize data, various challenges persist. Most frequently named is the implementation of FHIR as an application, the complexity of the FHIR standard (including its nested structure), and the representational state transfer (RESTful) approach [14]. In particular, the complexity of the data structure makes it not readily available for processing and easy access to the end user. Various solutions have been proposed, including SMART (Substitutable Medical Applications and Reusable Technologies) on FHIR, which is gaining momentum [15,16].

In this study, we evaluate the application of FHIR for a comprehensive analysis across medical disciplines to support multimodal decision management as part of routine clinical care. We propose to call this paradigm medical intelligence, akin to the expression business intelligence, which is defined as “strategies and technologies used by enterprises for the data analysis and management of business information [17,18].

We developed a standardized data pipeline using Germany’s largest FHIR database located at the University Hospital Essen. Together with department heads and (senior) physicians, we built and iteratively evaluated intuitive dashboards that provide health care workers with relevant information in a one-stop, intuitive format. As a start, we chose 5 common hospital diseases with a high disease burden, that is, myocardial infarction (MI), prostate cancer (PC), sepsis, stroke, and diabetes. To demonstrate clinical use, we iteratively tested and evaluated the dashboards with real-world data from our hospital.

Methods

FHIR

The resources infrastructure of FHIR forms the basic building blocks of the standard, representing health care information

XSL•FO
RenderX

such as patient demographics, medication requests and administrations, or diagnostic reports. Each resource is defined in a standard way, with a consistent structure and a set of standardized elements. FHIR resources are designed to be modular, meaning that they can be combined in various ways to represent more complex information or workflows. For example, a medication order (MedicationRequest in FHIR terminology) can be linked to a patient resource to indicate who the medication is prescribed for, and a practitioner resource can be linked to indicate who prescribed the medication [19]. This modular design makes it easy to create custom workflows and exchange information granularly. While FHIR offers many benefits, implementing and using the standard can still be challenging, especially when extracting and filtering the stored data.

FHIR Python Analysis Conversion Kit

The clinical use cases presented technical challenges regarding the identification, retrieval, clean-up, linkage, and subsequent distillation of FHIR data. To extract meaningful insights in an abstract, user-friendly, flexible, and reproducible manner we developed an open-source, interoperable Python client, and analysis toolkit for FHIR servers (FHIRPACK [FHIR Python Analysis Conversion Kit]) [20]. FHIRPACK enables clinicians and data scientists in finding, querying, and transforming FHIR resources and generating alternative representations with better processing properties such as powerful Python data analysis toolkit (pandas) data frames [21]. These can be used to carry out data analyses, generate static data visualizations, or implement dynamic dashboards that support medical processes such as continuous monitoring or complex, data-hungry research tasks such as hypothesis generation and statistical verification.

Cohort and Covariates

FHIRPACK was used to access and extract relevant information from the University Hospital Essen FHIR database with over 2 billion resources. Among these are over 1.4 million unique patients—constantly growing as the EHR system is in routine use. Our extraction workflow followed a similar structure for each use case, starting with the extraction of condition resources labeled with *ICD* (*International Classification of Diseases*) codes provided by clinicians. Next, patient cohorts were built using subject references in the observational data. We then obtained specific patient information through targeted queries from their respective FHIR resource, such as laboratory values (Observation), procedures (Procedure), diagnostic reports (DiagnosticReports), drug administration (MedicationAdministration), and vital signs (Observation; Figure 1).

Figure 1. Summary of the cohort building process and outcomes studied. Gray circles indicate patient counts for each use case. CRP: C-reactive protein; FHIR: Fast Healthcare Interoperability Resources; HBA1c: glycated hemoglobin A1c; ICD: International Classification of Diseases; PSA: prostate-specific antigen.

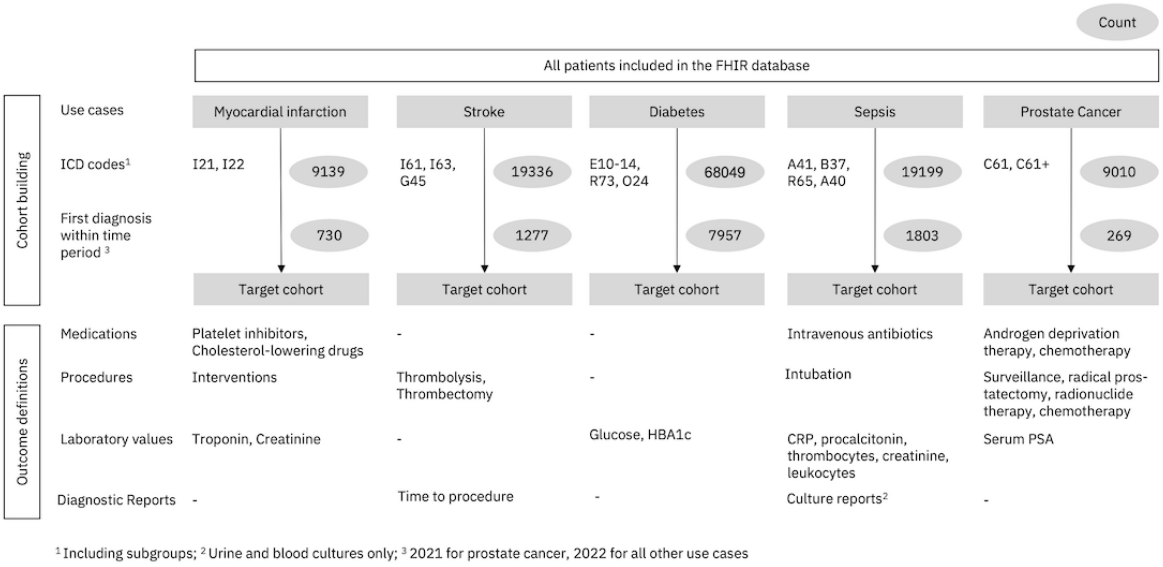

To enhance the speed and accuracy of our data evaluation, we restricted our cohort to a single calendar year (January 1-December 31, 2022). However, due to longer observation periods required for PC, we included all cases with diagnoses in 2021 and treatments in 2021 or 2022 for this use case. We used the FHIRPACK [20] and the Streamlit packages [22] to implement the dashboards at the point of care (PoC), accessible to clinicians in a web browser. Furthermore, we leveraged the same data and analysis processes to prepare the figures found in this manuscript using Matplotlib, another traditional Python visualization library. Details on the cohort building and covariate definitions are available in Multimedia Appendix 1.

Ethical Considerations

This study was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (22-10881-BO). No informed consent was required as routinely collected health care data were used.

Results

Cohort Description

We identified a total of 1,302,988 patient encounters in Essen University Hospital's FHIR database for 2022 (Figure 1). Based

XSL•FO
RenderX

on the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes, we identified between 269 (PC) and 7957 (diabetes) newly diagnosed patients across the 5 clinical use cases. All outcomes were derived from medication prescriptions, procedures, laboratory examinations, or diagnostic reports. While numerous data points could be reported, we focused on key indicators derived from clinicians' inputs through iterative focus groups. The clinical rationale, research question, and implications will be discussed for each use case separately.

MI

Background

An acute MI is an event of heart muscle necrosis caused by insufficient blood and oxygen supply. The incidence of acute MI or death is approximately 805,000 people in the United States alone, and accounts for approximately 20% of all causes of death [23]. Acute MI is classified based on the presence or absence of ST-segment elevation (ST elevation MI [STEMI] vs non–ST elevation MI [NSTEMI]) on electrocardiogram, while serial troponin levels differentiate a non-ST segment MI from unstable angina. Treatment includes coronary reperfusion with percutaneous coronary intervention, intravenous fibrinolytic therapy, or surgery. Depending on the bleeding risk and other patient factors, different antithrombotic drug combinations exist and are recommended in the guidelines. Further management focuses on risk reduction, including lipid-lowering drugs such as statins [24].

Clinical Question

How many patients were admitted with a new diagnosis of an NSTEMI or STEMI, received an interventional coronary revascularization therapy, and which antithrombotic and lipid-lowering drugs were prescribed?

Results

We identified 730 patients in 2022 with a diagnosis of acute MI. Of these, 359 patients were coded as STEMI or received a coronary intervention (Figure 2). Antiplatelet therapy was started in 71.6% (257/359), with most being on dual-antiplatelet therapy, while partly single antiplatelet therapy was administered in combination with existing anticoagulation therapy (not shown). Independent of antiplatelet therapy, most patients received a statin (227/359, 63.2%), with only a minority also receiving ezetimibe (97/359, 27%). Taken together, 72.7% (261/359) of patients received guideline-recommended therapy. Meanwhile, it should be appreciated that additional patients were considered for therapy, yet not administered due to contraindications, or received bridging therapy.

Figure 2. Sankey Diagram summarizing the antiplatelet (left bars) and lipid-lowering medications (right bars) prescribed to patients with STEMI or those receiving an intervention. Most patients not receiving antiplatelet therapy also did not receive lipid-lowering drugs (n=98). STEMI: ST elevation myocardial infarction.

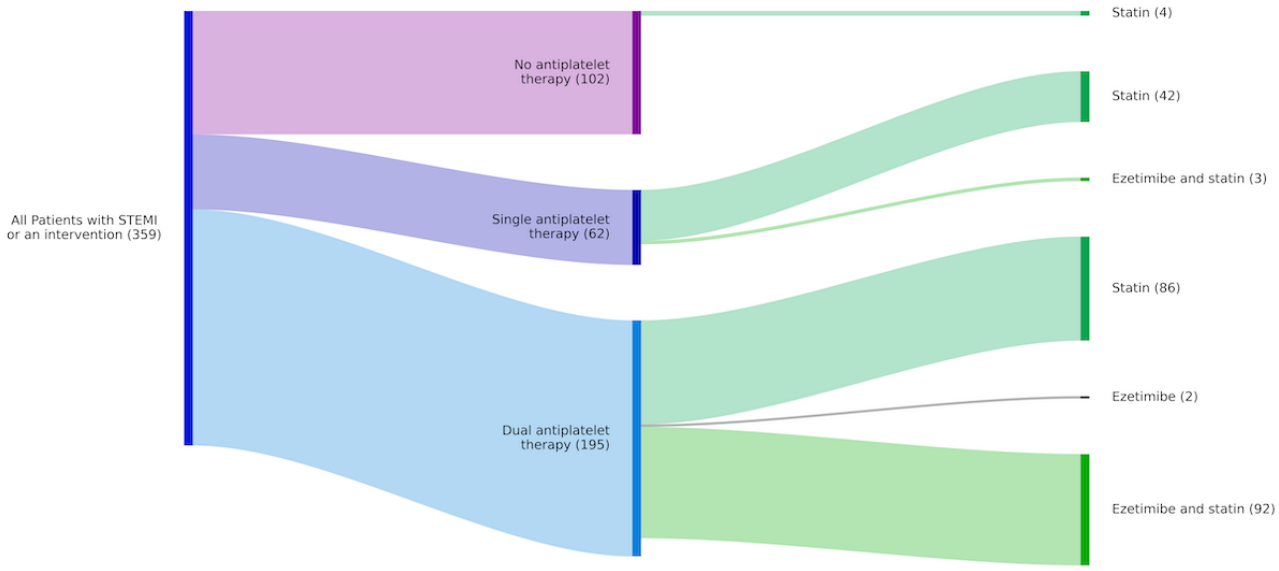

Clinical Implication

Platelet inhibitors and statins play an important role in the prevention of MI and mortality. Implementation of a clinical dashboard highlighting patients not receiving all recommended therapies can help to initiate these cost-efficient drugs and substantially improve patient care.

Stroke

Background

Ischemic stroke is a cardiovascular disease estimated to affect approximately 5.5 million patients across the globe in 2016. With approximately 116 million disease-adjusted life years lost globally, it carries a high mortality and morbidity [25]. A stroke occurs when a circulatory block results in a lack of oxygen delivery thus resulting in the subsequent death of brain tissue. In the case of an ischemic stroke, the main treatment is the resolution of the blood clot, either by drugs or mechanically by catheterization. The time from symptom onset to the resolution of the clot is one of the most important predictors of outcomes and an important indicator of improved patient care [26]. Huge efforts are made to reduce the time of the rescue chain to a minimum ("time is brain"). Along this way, the reduction of the time from presentation in the emergency room to initiation

XSL•FO
RenderX

of treatment, that is, thrombolysis ("Door to Needle") or thrombectomy ("Door to Groin"), is an important contribution of the intrahospital treatment [26]. Even though the outcome after acute ischemic stroke is generally assessed after 90 days, early neurological improvement is a surrogate indicator for a 90-day outcome. It is often defined as the percentage change in the National Institutes of Health Stroke Scale (NIHSS) score from baseline to 24 hours thereafter [27].

Clinical Questions

Are there fluctuations in the process quality, as measured by the waiting time between emergency room arrival and treatment application from the time of arrival in the emergency room until treatment? Does the time between the onset of symptoms until the start of a procedure correlate with early neurologic improvement?

Results

In the study period, we identified 1277 patients with a stroke diagnosis treated in the department of neurology, of which 1088 were ischemic and 189 were hemorrhagic. Out of 165 patients treated with intravenous thrombolysis, 143 (86.7%) received this therapy within the guideline-recommended time window of less than 60 minutes (Figure 3). Similarly, we identified 108 patients who received thrombectomy, of which 82 (75.9%) were within the guideline-recommended time window of 90 minutes. For both therapies, our cohort of a referral stroke center has a significant number of patients presenting in an extended time window and with the need of previous coagulation testing.

Figure 3. (A) Visualization of individual patients' time from admission to the administration of thrombolysis; green area indicates the recommended time of 60 minutes and (B) correlation of early neurological improvement with time from symptom onset to needle. (C) Individual patients' time from admission to thrombectomy; green area indicates the recommended time of 90 minutes and (D) correlation of early neurological improvement with time from symptom onset to groin puncture time. NIHSS: National Institutes of Health Stroke Scale.

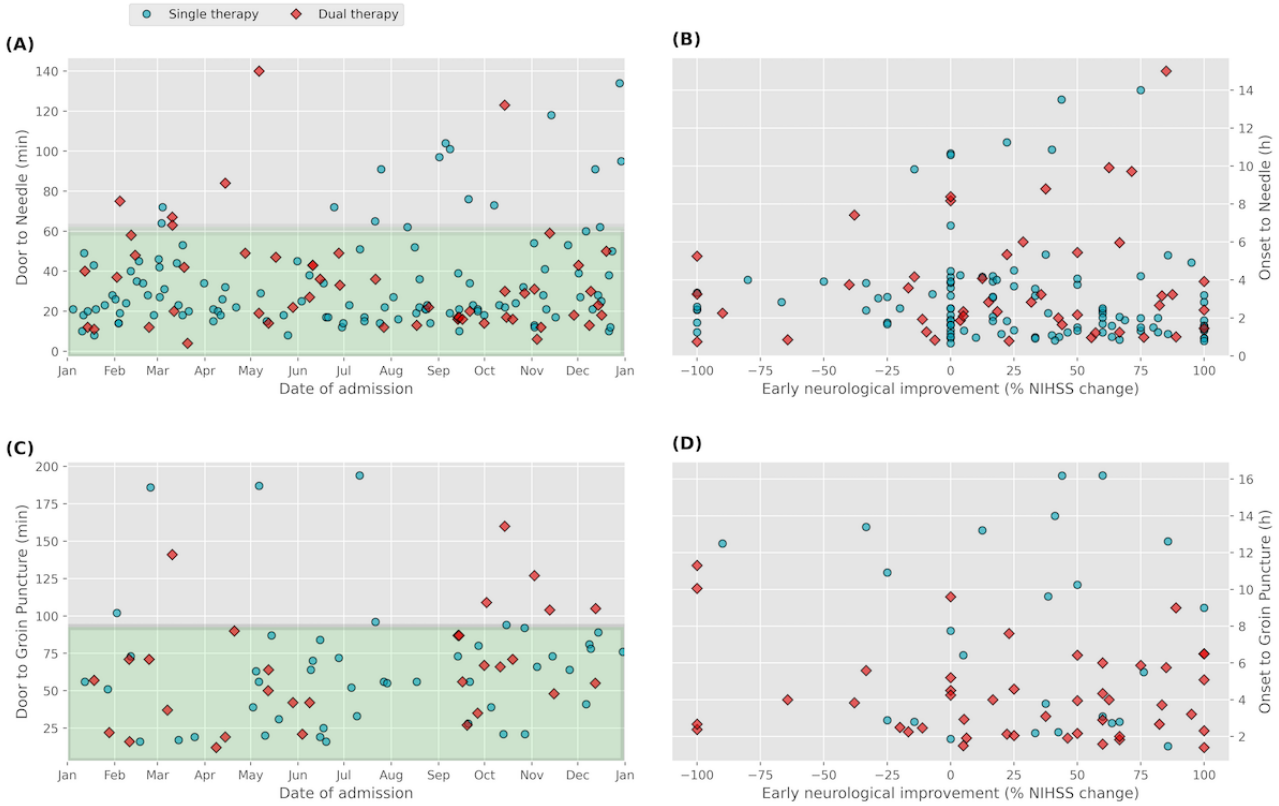

Furthermore, we correlated the time from symptom onset (witnessed onset or last known-well time) to intervention with early neurological improvement (NIHSS score) from 157 patients receiving thrombolysis and 101 receiving thrombectomy (Figures 3B and 3D). Among the patients receiving thrombolysis, 63.3% (101/157) displayed a decrease in NIHSS score (ie, clinical improvement), 15.9% (25/157) showed no change, and 21.7% (31/157) exhibited an increase in NIHSS score. Among the patients who received thrombectomy, 63% (48/76) demonstrated a decrease in NIHSS score, while 7% (5/76) showed no change, and 30% (23/76) exhibited an increase in NIHSS score. In total, 25 patients were excluded from the analysis due to ongoing weaning from mechanical ventilation after thrombectomy.

Clinical Implications

Monitoring of clinical workflows can identify cases with longer-than-recommended turnaround times. While many reasons for delays exist, for example, need of previous coagulation assessment in patients on oral anticoagulation or need of perfusion imaging in extended-time windows, and not all may be modifiable, some can potentially be addressed, and workflows adjusted. Structured analysis of this data is the first step for improving the care of future patients.

In all figures, patients who underwent a single therapy (either thrombolysis in Figures 3A and 3B or thrombectomy in Figures 3C and 3D) are color-coded in blue, and patients who received a dual therapy (both thrombolysis and thrombectomy) in red.

XSL•FO
RenderX

Diabetes

Background

Diabetes mellitus is a metabolic disorder with a high disease burden affecting 540 million people worldwide [28]. In hospitals, diabetes is one of the most common but undetected comorbidities increasing the length of stay and risk for nosocomial complications [29]. Identifying dysglycemia is a critically necessary step to introduce affected patients to dedicated in-hospital diabetes management and offer primary prevention strategies to those with prediabetes [29].

Clinical Question

How many inpatients are affected by dysglycemia and are cases identified appropriately?

Results

In total, 443,866 serum glucose and 16,180 glycated hemoglobin A_{1c} (HbA_{1c}) readings from 35,494 unique patients in 2022 were found. Further, 13,887 (39%) individuals, out of 35,494, were identified as having at least one pathologic glycemic finding. Overall, 30.5% (10,835/35,494) of the total cohort were at least hyperglycemic, while the prevalence of prediabetes and diabetes was 11.8% (4176/35,494) and 15.8% (5601/35,494; of which 3270 were identified by serum glucose, 941 by HbA_{1c} levels, and 1390 by both criteria), respectively (Figure 4A). Per definition, mean serum glucose and HbA_{1c} levels discriminated patients with normoglycemia from those with dysglycemia with a rise of mean glucose from 103.1 mg/dL to 112.3 mg/dL and 130.9 mg/dL to 167.1 mg/dL, respectively, and HbA_{1c} from 5.6% to 5.7% and 6% to 7.2%, respectively, in patients with normoglycemia, hyperglycemia, prediabetes, and diabetes (Figures 4B and 4D). *ICD-10* code documenting case identification for dysglycemia was found in 44.2% (6138/13,887) cases among those with biochemically proven hyperglycemia, prediabetes, or diabetes. Consequently, 55.8% (7749/13,887) of patients with dysglycemia would not have been identified by *ICD-10* coding alone (Figure 4C).

Figure 4. (A) Proportion of patients fulfilling the biochemical criteria for hyperglycemia (10,835/35,494, 30.5%), prediabetes (4176/35,494, 11.8%), and diabetes (5601/35,494, 15.8%), based on glucose and HbA1c readings. (C) Of all patients identified with dysglycemia (hyperglycemia, prediabetes, and diabetes), only 44.2% (6138/13,887) had a documented ICD-10 code for dysglycemia, while 55.8% (7749/13,887) of patients were not coded accordingly. Comparison of the distribution of (B) serum glucose and (D) HbA1c measurements of patients who do (blue) and do not fulfill the biochemical criteria of dysglycemia (red). Percentage of patients with biochemically proven dysglycemia with (blue; 6138/13,887, 44.2%) and without (red; 7749/13,887, 55.8%) a documented ICD-10 code. ICD-10: International Statistical Classification of Diseases, Tenth Revision.

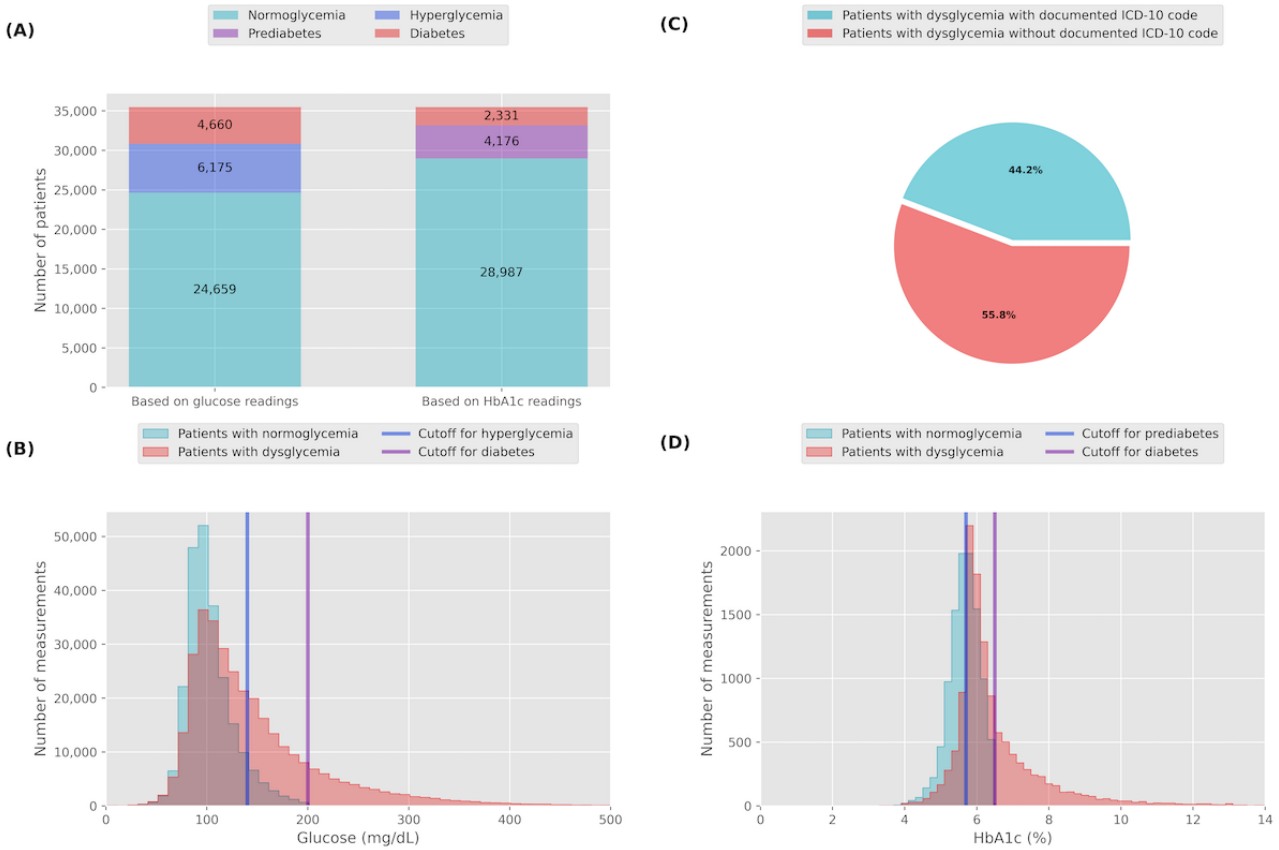

Clinical Implications

A dashboard visualizing in-hospital glucose metabolism data improves the dysglycemia detection rate. Future automated communication of identified cases to health care professionals may improve inpatient outcome by simplifying access by an algorithm-based referral to specialized diabetes teams as the first step of a digitalized in-hospital diabetes management program triggering dedicated diabetes care.

XSL•FO
RenderX

Sepsis

Background

Sepsis is a clinically heterogeneous syndrome defined by a dysregulated host response to infection resulting in life-threatening (multi-)organ dysfunction [30]. Its primary treatment is the timely administration of broad-spectrum antibiotics. Meanwhile, with an ever-increasing prevalence of multi–drug-resistant pathogens, it is paramount to treat as targeted and shortly as possible, to avoid the selection of further resistant strains. To this end, hospital-level insights into current antibiotic use, isolated pathogens, and the prevalence of multi–drug-resistant pathogens are critical to tailoring antibiotic practice to local conditions.

Clinical Question

Which antibiotics are currently used in patients with sepsis? Which pathogens are isolated from urine and blood cultures and how do the inflammatory biomarkers, that is, C-reactive protein and procalcitonin, change during therapy?

Results

We identified a total of 1803 patients with a sepsis diagnosis in 2022. Out of 213 distinct isolated pathogen strains, the top 10 most frequently isolated pathogen included *Staphylococcus epidermidis* (773/2672, 28.9%), a frequent contaminant with low pathogenic potential, followed by *Escherichia coli* (479/2672, 17.9%) and *Enterococcus faecium* (348/2672, 13%) strains (Figure 5A). Of the 10 most prescribed antibiotic piperacillin and tazobactam (593/1593, 37.2%), a broad-spectrum penicillin, followed by carbapenems (219/1593, 13.7%) and glycopeptides (eg, vancomycin, 186/1593, 11.7%) were most common (Figure 5B). For the clinical dashboard, we also provide a patient-level view showing the change in inflammatory markers, isolated pathogens, and antibiotics prescribed (Figure 5C).

Figure 5. (A) Pie chart of the 10 most common antibiotics prescriptions for patients with a sepsis diagnosis. (B) Distribution of the ten most frequently isolated microorganisms. (C) Example of a patient-level graph showing changes in the inflammatory markers C-reactive protein (purple), procalcitonin (blue), isolated pathogens (red icons), and prescribed antibiotics (turquoise bars). CRP: C-reactive protein; PCT: procalcitonin.

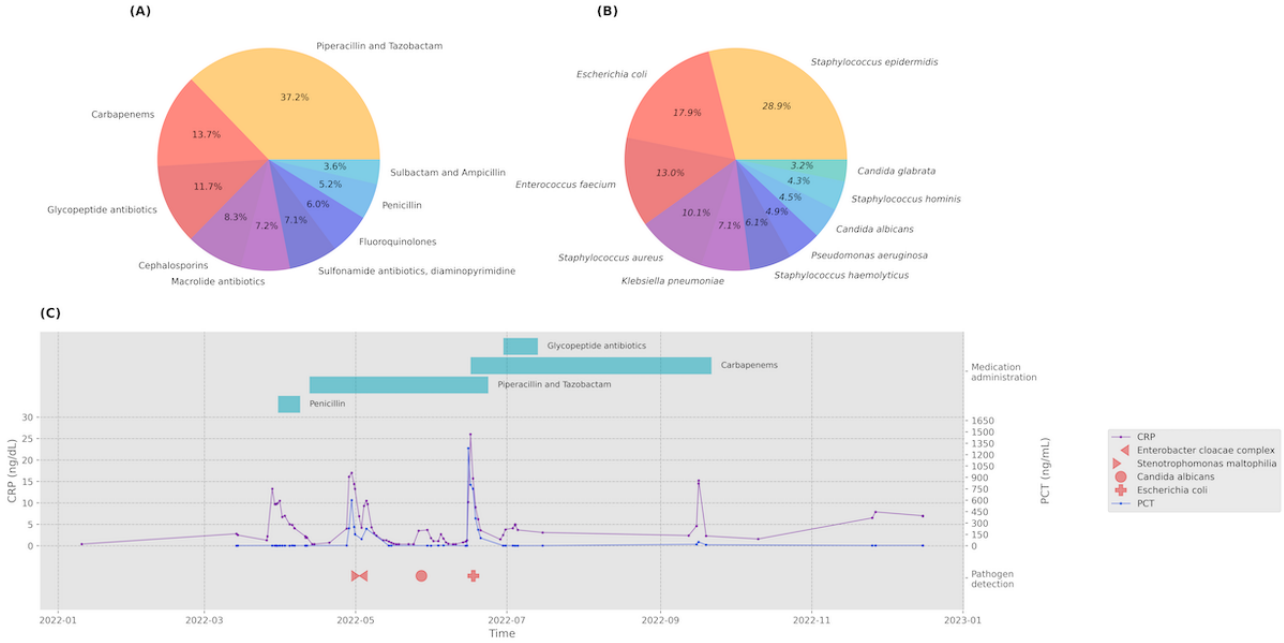

Clinical Implications

The dashboard provides a summary of the evolution of inflammatory markers and antibiotic prescriptions. It can help to identify changes in the causal pathogens and monitor antibiotic prescriptions to avoid harmful over- or undertreatment.

PC

Background

PC is the most common solid cancer in men in most of the Western World [31]. Various treatment options exist to achieve prolonged disease-free periods and cures. Important predictors for survival include the Tumor, Node, Metastasis stage and pathological features (International Society for Urological Pathology [ISUP] Gleason Grade Groups), which are summarized as a clinical stage (I-IV) as well as synchronous versus metachronous metastases. To assist in the diagnosis and monitor disease progression, prostate-specific antigen (PSA) is commonly used.

Clinical Questions

What is the current treatment landscape for PC by ISUP grades? What is the evolution of PSA levels before and after treatments depending on surgical resection status?

Results

We identified a total of 269 patients with newly diagnosed PC in 2021 who received treatment between 2021 and 2022. Among these, localized tumors (T1N0M0) with low ISUP scores were most common at biopsy. The most common treatment approach was active surveillance (n=29, 72.5%), followed by surgery (n=8, 20%; Figure 6A). For 54 patients treated with radical prostatectomy, PSA values were monitored longitudinally,

XSL•FO
RenderX

identifying 3 patients with PSA persistence or biochemical recurrence (Figure 6B).

Figure 6. (A) Schematic summary of the Tumor, Node, Metastasis stage, International Society for Urological Pathology grade and therapies received. The color of the lines represents the International Society for Urological Pathology grades (red=1, blue=2, green=3, purple=4, and yellow=5). (B) Longitudinal evolution of prostate-specific antigen values for all (n=54) patients treated with radical prostatectomy by pathological resection margins. Normalization of the temporal axis is achieved by ranking measurements according to their corresponding timestamp, where the average time between prostate-specific antigen–measurements across patients was found to be 110 days. Zoom-in highlighting cases with rising prostate-specific antigen values after treatment, indicating biochemical persistence. PSA: prostate-specific antigen.

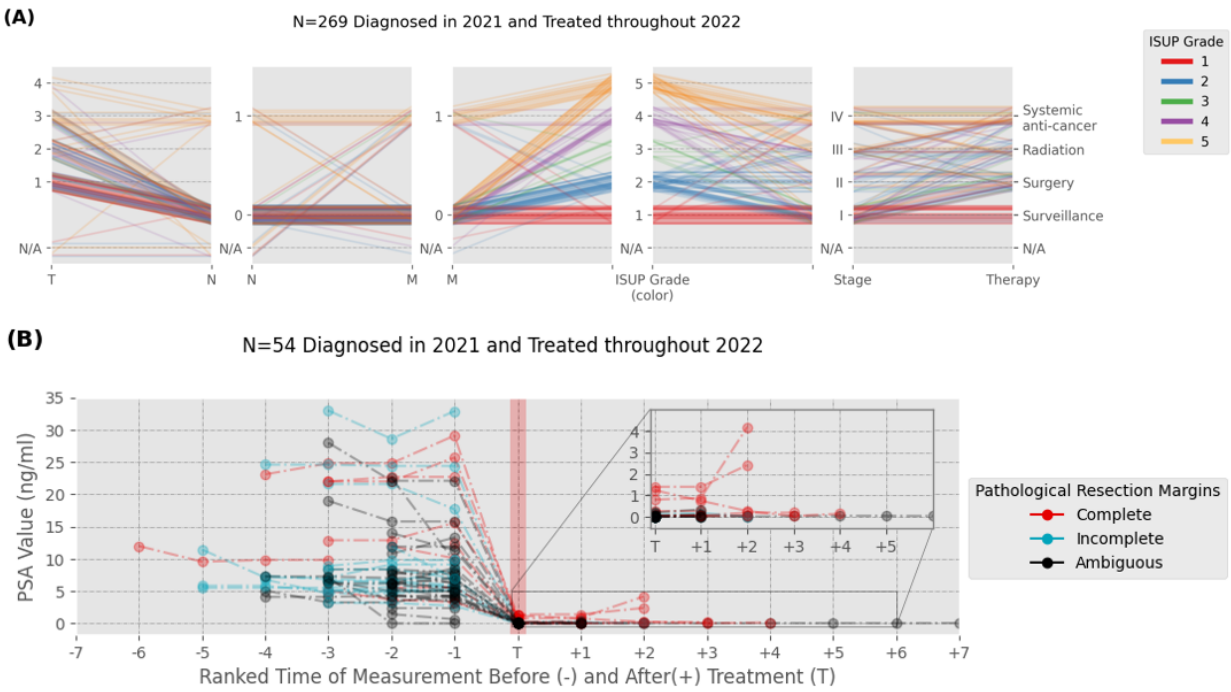

Clinical Implications

Increases in PSA values after treatment indicate disease relapse or progression. Especially in the context of easy access to sensitive PSA tests, earlier detection of biochemical recurrence is warranted. Identification and monitoring can streamline treatment approaches, potentially improving patient outcomes by initiating early salvage therapy. Furthermore, with recent improvements in staging through molecular imaging with prostate-specific membrane antigen positron emission tomography or computed tomography (PSMA-PET/CT) and novel combination treatment regimens, centralized monitoring of shifts in routine practice is an essential first step to draw conclusions about the real-world effectiveness of these approaches.

Discussion

With the introduction of EHRs, significantly more information has become available to physicians. Yet, this flood of information does not automatically result in better care, as pointedly summarized in the saying, “We are drowning in information but *starved* for knowledge” [32]. To address this, we establish a conceptual “medical intelligence” framework for leveraging real-world data for clinical monitoring and decision-making across medical specialties and at the PoC. We demonstrate the applicability of this flexible conceptual framework using 5 conditions with a high disease burden and major causes of hospitalization.

The use cases presented in this paper provide new insights for clinicians and health care managers. In the MI use case, we identified patients who did not receive lipid-lowering or antiplatelet therapy, allowing for an internal review. As for the PC use case, we detected patients with early increases in PSA values indicating tumor persistence after therapy with curative intent. We found higher than recommended use of reserve, broad-spectrum antibiotics in the sepsis use case. For patients who had a stroke, we identified some with turnaround times above guideline-recommended thresholds. Meanwhile, the diabetes use case highlighted a high prevalence of dysglycemia and diabetes in the hospital population, while coding for diabetes—and therefore potential referral rates to endocrinologists—was significantly lower.

These use cases demonstrate the applicability and benefits of a unified approach based on the interoperability of FHIR and a Python library that enables easy access to data. The approach can be quickly scaled and applied to different types of FHIR data, providing a foundation for hospitals to truly harness the power of patient data across departments, specialties, and locations. It enables easier creation of dashboards that can be used to monitor key performance indicators, summarize information for clinicians, support clinical decision-making, and ultimately facilitate better patient care. By analogy with business intelligence, where data support business decisions, our approach demonstrates that data can also support clinical decisions. We show that our approach can lead to new insights into patient care, both at the patient and hospital levels.

XSL•FO
RenderX

While the results presented are descriptive, it is possible to integrate predictive and artificial intelligence algorithms. Regardless of the use case and methodological technique, we believe that a multidisciplinary approach is key, and that successful deployment and adoption will depend on updated training for the modern physician. These "new" physicians have been called "information specialists" [33]. Information management and processing will become increasingly important with the increasing complexity and availability of data from novel diagnostic tools. Our proposed medical intelligence approach is the first step in closing the implementation gap that currently hinders clinical progress, since it provides an interoperable data structure (FHIR) with a publicly available analysis framework (this paper) [34].

The conceptual framework and implementation show that validated and highly interdisciplinary knowledge acquisition from FHIR data is possible. However, several key challenges remain. First, while FHIR provides a standardized format to exchange data, it does not dictate the content or structure of the data itself. To achieve this, FHIR profiles can be leveraged, which are structured, user-defined, and derived FHIR resources. However, these are used only infrequently or selectively and typically not hospital-wide, resulting in residual differences in how health care organizations store and structure their data. This can create challenges for interoperability and external validation of results. Second, the FHIR standard and the data it allows health care institutions to gather, and store are complex, thus requiring joint technical and medical expertise to implement. This can be a barrier to adoption for health care organizations, particularly smaller ones, that may not have the in-house expertise to develop their own organizational, conceptual, and programmatic approaches. Third, the accuracy and completeness of the data exchanged through FHIR depends on the quality of the data entered into the primary systems. To this end, incomplete or incorrect data, as is frequently the case with ICD codes, may negatively impact analytical results. Most data exist in an unstructured format that can be accessed through FHIR but requires significant postacquisition structuring.

This study provides the technical basis for pioneering medical intelligence for all institutions working with FHIR data and the results presented are directly applicable to clinical practice. However, a prospective impact on clinical decision-making remains to be demonstrated.

In summary, we developed and implemented a conceptual framework for FHIR data analysis that enables researchers and clinicians to derive insights from hospital EHR data. This approach has the potential to enable medical intelligence and improve clinical care by supporting decision-making at the PoC.

Acknowledgments

CMS is supported by the German Research Foundation–funded UMEA Clinician Scientist Program, (grant FU356/12-2).

JSR is supported by the Helmholtz Association under the joint research school "HIDSS4Health Helmholtz Information and Data Science School for Health."

TB received research funding from the German Research Foundation, Dietmar Hopp Stiftung, Stiftung Universitätsmedizin Essen, Heidelberger Stiftung Chirurgie, and Innovationsfonds des Gemeinsamen Bundesausschusses.

This project received funding from the Cancer Research Center Cologne Essen, the Helmholtz Information & Data Science School for Health, and KI Translation Essen.

This study received funding from the Cancer Research Center Cologne Essen, The Helmholtz Information & Data Science School for Health, KI Translation Essen, the German Research Foundation, and the UMEA Clinician Scientist Program.

Data Availability

The datasets analyzed during the current study are not publicly available as per the German legal code. However, and upon reasonable request, we will provide interested researchers with a pseudonymized version of the different datasets on an individual basis.

The FHIR data processing Python software package, FHIRPACK, is freely available on the internet, packaged, versioned, as well as citable [20]. This is the first central component to all use cases presented in this work, as it was used to retrieve, filter, process, and persist the different cohorts and their corresponding datasets. Separately, we provide the data-processing notebooks associated with each case, which produced the data and images presented herein. For reviewing purposes, the data-processing notebooks, visualizations as well as a copy of FHIRPACK's documentation are provided [35].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methods.
[DOCX File , 85 KB - jmir_v26i1e55148_app1.docx]

XSL•FO
RenderX

References

1. Haux R. Health information systems - past, present, future. Int J Med Inform 2006;75(3-4):268-281. [doi: 10.1016/j.ijmedinf.2005.08.002] [Medline: 16169771]
2. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
3. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
4. Moody G, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. Comput Cardiol 2000;27:179-182. [Medline: 14632011]
5. Sauer C, Dam TA, Celi LA, Faltys M, de la Hoz MAA, Adhikari L, et al. Systematic review and comparison of publicly available ICU data sets-a decision guide for clinicians and data scientists. Crit Care Med 2022;50(6):e581-e588 [FREE Full text] [doi: 10.1097/CCM.0000000000005517] [Medline: 35234175]
6. Aguirre RR, Suarez O, Fuentes M, Sanchez-Gonzalez MA. Electronic health record implementation: a review of resources and tools. Cureus 2019;11(9):e5649 [FREE Full text] [doi: 10.7759/cureus.5649] [Medline: 31700751]
7. Adoption of Electronic Health Records by Hospital Service Type 2019-2021. URL: https://www.healthit.gov/data/quickstats/adoption-electronic-health-records-hospital-service-type-2019-2021 [accessed 2024-09-21]
8. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, et al. Deep representation learning of patient data from electronic health records (EHR): a systematic review. J Biomed Inform 2021;115:103671 [FREE Full text] [doi: 10.1016/j.jbi.2020.103671] [Medline: 33387683]
9. Index - FHIR v4.3.0. URL: https://hl7.org/fhir/ [accessed 2024-09-21]
10. Federal register. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program. 2020. URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2024-09-21]
11. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. Stud Health Technol Inform 2019;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
12. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
13. Duda SN, Kennedy N, Conway D, Cheng AC, Nguyen V, Zayas-Cabán T, et al. HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. J Am Med Inform Assoc 2022;29(9):1642-1653 [FREE Full text] [doi: 10.1093/jamia/ocac105] [Medline: 35818340]
14. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
15. Mandl KD, Kohane IS. No small change for the health information economy. N Engl J Med 2009;360(13):1278-1281. [doi: 10.1056/NEJMp0900411] [Medline: 19321867]
16. SMART Health IT. URL: https://smarthealthit.org/ [accessed 2024-09-21]
17. Olszak CM, Ziemba E. Business intelligence systems in the holistic infrastructure development supporting decision making in organisations. Interdiscip J Inf Knowl Manag 2006;1:47-58. [doi: 10.28945/113]
18. Business intelligence. 2023. URL: https://en.wikipedia.org/wiki/Business_intelligence [accessed 2024-09-21]
19. Mandel J, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
20. Salazar J. FHIRPACK (FHIR Python Analysis Client and Kit). France: Zenodo; 2022.
21. Reback J, McKinney W, Jbrockmendel, Van Den Bossche J, Augspurger T, Cloud P, Gfyoung, Sinhrks, MomIsBestFriend, et al. pandas-dev/pandas: Pandas 1.1.3. France: Zenodo; 2020.
22. Streamlit: A faster way to build and share data apps. 2021. URL: https://streamlit.io/undefined/ [accessed 2024-09-21]
23. Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation. 2022. URL: https://www.ahajournals.org/doi/10.1161/CIR.0000000000001052 [accessed 2024-09-21]
24. Collet JP, Thiele H, Barbato E, Barthélémy O, Bauersachs J, Bhatt DL, ESC Scientific Document Group. 2020 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Eur Heart J 2021;42(14):1289-1367. [doi: 10.1093/eurheartj/ehaa575] [Medline: 32860058]
25. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol 2019;18(5):439-458 [FREE Full text] [doi: 10.1016/S1474-4422(19)30034-1] [Medline: 30871944]
26. Campbell BCV, Khatri P. Stroke. Lancet 2020;396(10244):129-142. [doi: 10.1016/S0140-6736(20)31179-X] [Medline: 32653056]

XSL•FO
RenderX

27. Kobeissi H, Ghozy S, Bilgin C, Kadirvel R, Kallmes DF. Early neurological improvement as a predictor of outcomes after endovascular thrombectomy for stroke: a systematic review and meta-analysis. J Neurointerv Surg 2023;15(6):547-551. [doi: 10.1136/neurintsurg-2022-019008] [Medline: 35636948]
28. GBD 2019 Diabetes in the Americas Collaborators. Burden of diabetes and hyperglycaemia in adults in the Americas, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Diabetes Endocrinol 2022;10(9):655-667 [FREE Full text] [doi: 10.1016/S2213-8587(22)00186-3] [Medline: 35850129]
29. Valent F, Tonutti L, Grimaldi F. Does diabetes mellitus comorbidity affect in-hospital mortality and length of stay? Analysis of administrative data in an Italian academic hospital. Acta Diabetol 2017;54(12):1081-1090. [doi: 10.1007/s00592-017-1050-6] [Medline: 28916936]
30. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
31. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7-33 [FREE Full text] [doi: 10.3322/caac.21708] [Medline: 35020204]
32. Naisbitt J. Megatrends: Ten New Directions Transforming Our Lives. New York: Warner Books; 1982.
33. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA 2016;316(22):2353-2354. [doi: 10.1001/jama.2016.17438] [Medline: 27898975]
34. Weiss CH, Krishnan JA, Au DH, Bender BG, Carson SS, Cattamanchi A, ATS Ad Hoc Committee on Implementation Science. An official American Thoracic Society research statement: implementation science in pulmonary, critical care, and sleep medicine. Am J Respir Crit Care Med 2016;194(8):1015-1025 [FREE Full text] [doi: 10.1164/rccm.201608-1690ST] [Medline: 27739895]
35. Leveraging FHIR to Improve Clinical Management. Establishing Medical Intelligence. URL: https://fhirpack.gitlab.io/establishing-medical-intelligence/examples/medical-intelligence.html [accessed 2024-09-30]

Abbreviations

EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
FHIRPACK: FHIR Python Analysis Conversion Kit
HBA1c: glycated hemoglobin A1c
ICD: International Classification of Diseases
ICD-10: International Statistical Classification of Diseases, 10th Revision
ISUP: International Society for Urological Pathology
MI: myocardial infarction
MIMIC-III: Medical Information Mart for the Intensive Care
NIHSS: National Institutes of Health Stroke Scale
NSTEMI: non–ST elevation myocardial infarction
pandas: powerful Python data analysis toolkit
PC: prostate cancer
PoC: point of care
PSA: prostate-specific antigen
PSMA-PET/CT: prostate-specific membrane antigen positron emission tomography or computed tomography
SMART: Substitutable Medical Applications and Reusable Technologies
STEMI: ST elevation myocardial infarction

Edited by F Xie; submitted 04.12.23; peer-reviewed by C Asuzu, Y Chu, K Nolen, A Hidki; comments to author 26.06.24; revised version received 20.07.24; accepted 20.08.24; published 31.10.24.

Please cite as:
Brehmer A, Sauer CM, Salazar Rodríguez J, Herrmann K, Kim M, Keyl J, Bahnsen FH, Frank B, Köhrmann M, Rassaf T, Mahabadi AA, Hadaschik B, Darr C, Herrmann K, Tan S, Buer J, Brenner T, Reinhardt HC, Nensa F, Gertz M, Egger J, Kleesiek J
Establishing Medical Intelligence—Leveraging Fast Healthcare Interoperability Resources to Improve Clinical Management: Retrospective Cohort and Clinical Implementation Study
J Med Internet Res 2024;26:e55148
URL: https://www.jmir.org/2024/1/e55148
doi:10.2196/55148
PMID:39240144

XSL•FO
RenderX

©Alexander Brehmer, Christopher Martin Sauer, Jayson Salazar Rodríguez, Kelsey Herrmann, Moon Kim, Julius Keyl, Fin Hendrik Bahnsen, Benedikt Frank, Martin Köhrmann, Tienush Rassaf, Amir-Abbas Mahabadi, Boris Hadaschik, Christopher Darr, Ken Herrmann, Susanne Tan, Jan Buer, Thorsten Brenner, Hans Christian Reinhardt, Felix Nensa, Michael Gertz, Jan Egger, Jens Kleesiek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study

Minseok Hong[1,2], MD; Ri-Ra Kang[3], BS; Jeong Hun Yang[2,4], MD, PhD; Sang Jin Rhee[1], MD, PhD; Hyunju Lee[1], MD, PhD; Yong-gyom Kim[3], BS; KangYoon Lee[3,5], PhD; HongGi Kim[6], BS; Yu Sang Lee[7], MD, PhD; Tak Youn[8,9], MD, PhD; Se Hyun Kim[1,2], MD, PhD; Yong Min Ahn[1,2,10], MD, PhD

[1]Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea
[2]Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea
[3]Department of IT Convergence Engineering, Gachon University, Seongnam-si, Republic of Korea
[4]Department of Psychiatry, Chungnam National University Sejong Hospital, Sejong, Republic of Korea
[5]Department of Computer Engineering, Gachon University, Seongnam-si, Republic of Korea
[6]Healthconnect Co. Ltd., Seoul, Republic of Korea
[7]Department of Psychiatry, Yong-In Mental Hospital, Yongin-si, Republic of Korea
[8]Department of Psychiatry and Electroconvulsive Therapy Center, Dongguk University International Hospital, Goyang-si, Republic of Korea
[9]Institute of Buddhism and Medicine, Dongguk University, Seoul, Republic of Korea
[10]Institute of Human Behavioral Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea

Corresponding Author:
Yong Min Ahn, MD, PhD
Department of Neuropsychiatry
Seoul National University Hospital
101 Daehak-ro
Jongno-Gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 2450
Fax: 82 2 766 2450
Email: aym@snu.ac.kr

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e69042

Abstract

Background: Assessing the complex and multifaceted symptoms of patients with acute psychiatric disorders proves to be significantly challenging for clinicians. Moreover, the staff in acute psychiatric wards face high work intensity and risk of burnout, yet research on the introduction of digital technologies in this field remains limited. The combination of continuous and objective wearable sensor data acquired from patients with deep learning techniques holds the potential to overcome the limitations of traditional psychiatric assessments and support clinical decision-making.

Objective: This study aimed to develop and validate wearable-based deep learning models to comprehensively predict patient symptoms across various acute psychiatric wards in South Korea.

Methods: Participants diagnosed with schizophrenia and mood disorders were recruited from 4 wards across 3 hospitals and prospectively observed using wrist-worn wearable devices during their admission period. Trained raters conducted periodic clinical assessments using the Brief Psychiatric Rating Scale, Hamilton Anxiety Rating Scale, Montgomery-Asberg Depression Rating Scale, and Young Mania Rating Scale. Wearable devices collected patients' heart rate, accelerometer, and location data. Deep learning models were developed to predict psychiatric symptoms using 2 distinct approaches: single symptoms individually (Single) and multiple symptoms simultaneously via multitask learning (Multi). These models further addressed 2 problems:

XSL•FO
RenderX

within-subject relative changes (Deterioration) and between-subject absolute severity (Score). Four configurations were consequently developed for each scale: Single-Deterioration, Single-Score, Multi-Deterioration, and Multi-Score. Data of participants recruited before May 1, 2024, underwent cross-validation, and the resulting fine-tuned models were then externally validated using data from the remaining participants.

Results: Of the 244 enrolled participants, 191 (78.3%; 3954 person-days) were included in the final analysis after applying the exclusion criteria. The demographic and clinical characteristics of participants, as well as the distribution of sensor data, showed considerable variations across wards and hospitals. Data of 139 participants were used for cross-validation, while data of 52 participants were used for external validation. The Single-Deterioration and Multi-Deterioration models achieved similar overall accuracy values of 0.75 in cross-validation and 0.73 in external validation. The Single-Score and Multi-Score models attained overall R^2 values of 0.78 and 0.83 in cross-validation and 0.66 and 0.74 in external validation, respectively, with the Multi-Score model demonstrating superior performance.

Conclusions: Deep learning models based on wearable sensor data effectively classified symptom deterioration and predicted symptom severity in participants in acute psychiatric wards. Despite lower computational costs, Multi models demonstrated equivalent or superior performance to Single models, suggesting that multitask learning is a promising approach for comprehensive symptom prediction. However, significant variations were observed across wards, which present a key challenge for developing clinical decision support systems in acute psychiatric wards. Future studies may benefit from recurring local validation or federated learning to address generalizability issues.

(J Med Internet Res 2024;26:e65994) doi:10.2196/65994

KEYWORDS

digital phenotype; mental health monitoring; smart hospital; clinical decision support system; multitask learning; wearable sensor; local validation; mental health facility; deep learning

Introduction

Assessing psychiatric symptoms of patients with acute psychiatric disorders remains challenging [1-3]. Patients with psychosis, mania, and severe depression who require hospitalization often struggle to accurately report their symptoms. Poor insight, a short attention span, impaired cognition, paranoia, and severe avolition can result in underreporting, minimization, and impairment of the verbalization process [3,4]. Additionally, the inherent subjectivity in assessing psychopathology presents reliability concerns and requires experienced professionals [5,6]. Even well-staffed and equipped wards cannot provide continuous patient observation by trained medical personnel.

Recent advancements in wearable sensor technology and artificial intelligence (AI) have enabled the collection and analysis of vast amounts of data [7-9]. These data encompass characteristics related to human behavior, cognition, and mood associated with mental disorders [10] and are often referred to as digital phenotyping [11-13]. The continuous and objective measurement capabilities of sensor data can overcome the limitations of traditional psychiatric assessments [14-16]. Numerous studies have reported that sensor data from wearables and mobile phones can significantly predict health outcomes, such as disease risk and mortality, by unobtrusively monitoring sleep [17-20], circadian rhythms [21-23], and physical activity [24,25].

Using sensor data in acute psychiatric wards is expected to produce significant advantages. Acute psychiatric wards must always be prepared for the risks of aggression and self-harm, resulting in high staffing demands [26,27]. Health care workers in these settings are repeatedly exposed to such patient behaviors, introducing a significant risk of burnout [28,29]. Additionally, human rights standards and the legislation required for the care of psychiatric inpatients are progressing globally [30]. Consequently, digital technology adoption remains a pressing issue in addressing these challenges in psychiatric wards [31]. Furthermore, psychiatric wards provide a highly controlled environment, enabling accurate and continuous collection of wearable sensor data and detailed clinical information. This creates suitable conditions for digital phenotyping research [32].

Despite these expected advantages, only a few studies have presented AI models to support medical personnel in the context of acute psychiatric wards. To the best of our knowledge, the task of symptom prediction has not yet been addressed. Psychiatric wards exhibit a wide range of architectural, staffing, and organizational variations [33-35], likely contributing to limited research in these contexts. Furthermore, psychiatric inpatients have diverse diagnoses and a wide symptom range [36,37]. Predicting comprehensive symptoms across multi-institutional pandiagnostic groups is necessary to provide a robust foundation for AI-assisted clinical decision support systems (AI-CDSSs) [38].

Therefore, we aimed to develop and validate a comprehensive symptom prediction model for patients in acute psychiatric wards using sensor data obtained using wearable devices and deep learning–based prediction models. This prospective, longitudinal, observational study was conducted across multiple institutions, representing various regions and hospital types. We constructed and evaluated models for predicting single symptoms individually and multiple symptoms simultaneously, as well as for predicting relative changes within subjects and absolute severity between subjects. Moreover, we explored the future directions and challenges of AI-CDSSs in psychiatric ward settings.

XSL•FO
RenderX

Methods

Overview

The study was presented according to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [39], as well as the TRIPOD+AI guideline (version February 7, 2024) [40].

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki [41]. The study protocol was reviewed and approved by the Institutional Review Board of the Seoul National University Hospital (2210-073-1368). All participants provided informed consent. Participants received compensation of KRW 100,000 (approximately US $70) for each week of participation. Research data were stored on servers within each hospital, separate from personally identifiable information except for study identification numbers. Access to and analysis of the data were restricted to preapproved researchers only. No identifiable images or personal information of participants are included in this manuscript or any supplementary materials.

Participants

The Sensor Application for Early Response in Closed Wards (SAFER) project aims to build a generalizable wearable-based AI-CDSS for acute psychiatric wards. This project is ongoing until 2024, and interim data were used in this study.

Four wards from 3 hospitals representing diverse regions and hospital types participated in the SAFER study: Seoul National University Hospital, a tertiary general hospital in the capital city with a single participating ward; Yongin Mental Hospital, a suburban psychiatric hospital with 2 wards, one for each sex; and Dongguk University Ilsan Hospital, a general hospital in a newly developed city with one participating ward.

Inpatients aged 13 years or older, diagnosed with a mood disorder (major depressive or bipolar disorder) or schizophrenia spectrum disorder, were recruited from each ward. For vulnerable participants, including minors and individuals whose symptoms might impair their ability to understand and consent, additional consent was obtained from their legal guardians. Exclusion criteria included dementia, intellectual disability, organic brain disorders, and any physical condition causing difficulty in data collection (eg, wrists too thin owing to low body weight of participants with anorexia nervosa). However, other coexisting diagnoses were included to ensure a pandiagnostic study population. As one of the analysis targets was a comparison with previous scores of the same individual, the analysis included only participants who had undergone at least 2 assessments, including the baseline assessment.

Measurements

Each participant was provided with a wrist-worn device called URBAN HR (Partron Co, Ltd). Participants were instructed to wear the device continuously, except when showering or leaving the ward, although they could remove it freely if desired. The wearable devices collected data on heart rate, 3-axis acceleration (which measures acceleration along 3 perpendicular axes in space), and location. Using in-device algorithms, these metrics were used to calculate calories burned, steps walked, distance moved, and sleep index. Among these, the sleep index is used to analyze participants' sleep and is calculated based on heart rate and acceleration data using algorithms provided by the manufacturer. This sleep index quantifies the cumulative daily sleep duration, converting it into a score ranging from 30 (2.25 h or less) to 100 (7.5 h or more). The generated data were transmitted in real time to the server at each hospital via Bluetooth gateways installed in each patient's room and common areas.

The participants were assessed weekly for symptoms by raters using clinician-rated scales. Additional assessments were conducted following injection, seclusion, and restraint interventions: the Brief Psychiatric Rating Scale (BPRS) was used to evaluate psychotic and general psychiatric symptoms, the Hamilton Anxiety Scale (HAM-A) was used to assess anxiety, the Montgomery-Asberg Depression Rating Scale (MADRS) was used to evaluate depressive symptoms, and the Young Mania Rating Scale (YMRS) was used to measure manic symptoms. The raters were registered nurses with a minimum of 51 months and an average of 11 years of clinical experience. All raters completed 40 hours of specialized training for the study, and their ratings were periodically reviewed at research conferences with psychiatrists.

Feature Engineering

A 1-hour sliding window was used to synthesize time-series values from various sensors. The 3-axis accelerations were converted to a total acceleration, which is the magnitude of the summed acceleration. The first value, last value, mean, median, maximum, minimum, SD, and number of unique values within the 1-hour sliding window were calculated for total acceleration and heart rate. Calories burned, sleep index, number of steps walked, and distance traveled were calculated as cumulative values over this window. Regarding location data, the variance of the coordinate values and most frequent semantic location (room, hallway, other, and no signal) were computed over the sliding window.

Location entropy quantifies the diversity and unpredictability of a participant's location patterns, which have demonstrated predictive value for mental health correlates [42-44]. We computed entropy values over 24-hour and 8-hour periods, delineated by midnight, 8 AM, and 4 PM. This approach was chosen as meaningful location entropy assessment requires longer periods than the 1-hour sliding window.

Missing values in the sensor data were imputed using the *softImpute* method [45]. The participant was excluded if missing values exceeded 50% of their total timespan. Sensor data from the 4 weeks preceding each symptom assessment were used in the model. In cases where available data preceding the assessment were less than 4 weeks, zero-padding was applied to extend the data to the required length [46]. The size of the input layer was 672 hours with 31 features (672 × 31) including 2 nonsensor features (age and sex). t-distributed Stochastic Neighbor Embedding (t-SNE) was used to visualize the distribution of these 31 features by ward.

XSL•FO
RenderX

Prediction Model and Performance Evaluation

A model designed to predict symptoms encompassing the BPRS, HAM-A, MADRS, and YMRS was constructed using a deep learning architecture that incorporates 1D convolutional layers and gated recurrent units (Figure 1). All models shared an identical structure, except for the terminal fully connected layer and output layer.

Figure 1. The architecture of deep learning models for predicting comprehensive symptoms in inpatients with acute psychiatric disorders. Original sequence data were adjusted to match the required length by removing data older than 4 weeks; shorter data were zero-padded to form the input layer. 1D-Conv, GRU, and fully connected layers were used sequentially. The model's performance was compared based on various output layer configurations. 1D-Conv: 1D convolutional; GRU: gated recurrent unit.

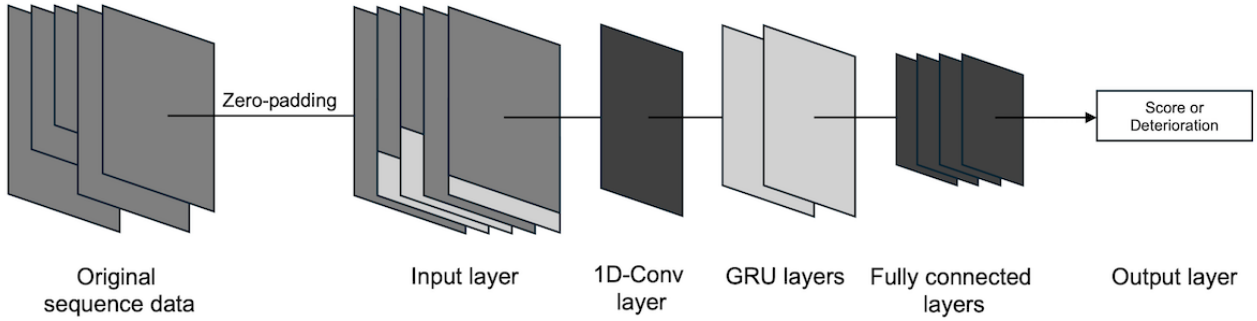

These models were differentiated based on the configuration of the output layer according to 2 criteria. First, the models were classified based on whether they predicted a single scale at a time (Single) or predicted all scales simultaneously (Multi). Multi models used a multitask learning (MTL) approach [47]. Second, the models were distinguished based on whether they classified instances where the symptom scale score increased from the previous assessment within an individual (Deterioration) or predicted the absolute severity score of the symptoms (Score). Deterioration models corresponded to a within-subject design, while Score models aligned with a between-subject design [48]. Consequently, 4 types of models were developed for each scale by intersecting these 2 criteria: Single-Deterioration, Single-Score, Multi-Deterioration, and Multi-Score.

Participants recruited prior to May 1, 2024, were allocated to the internal cross-validation set, while those recruited on or after May 1, 2024, were assigned to the external validation set. The internal cross-validation set underwent a 5-fold cross-validation. Hyperparameters were optimized using the random search method [49]. The final model was evaluated on the external validation set. The performance of Deterioration models was evaluated using accuracy, area under the curve (AUC), and receiver operating characteristic (ROC) curve. In the cross-validation, the ROC curve was generated by vertically pooling the curves from each fold. The performance of Score models was assessed using R^2 and normalized root mean squared error (NRMSE), with NRMSE chosen to account for varying score ranges across scales. Permutation feature importance was measured for features in the external validation set. The deep learning model was built using *PyTorch* (version 2.2) [50], and model performance evaluation and t-SNE were performed using *scikit-learn* (version 1.3) [51].

Results

Overall, 244 participants were enrolled in the SAFER study between May 26, 2023, and August 5, 2024. Sixteen participants were excluded based on the exclusion criteria or consent withdrawal after initially agreeing to participate. Additionally, 37 participants were excluded from the analysis because they lacked at least 1 assessment or had > 50% missing data from wearable sensors. Ultimately, data from 191 participants, encompassing 3954 person-days of observation, were analyzed. Participants were observed for a mean of 20.7 (SD 17.5) days, with 4.3 (SD 2.9) assessments per participant. In the final sample, 57.6% (110/191) were female, and the most frequent diagnosis was a mood disorder with a current depressive episode (93/191, 48.7%). For validation purposes, 139 participants were assigned to the cross-validation set and 52 participants were assigned to the external validation set.

Significant differences were observed in sex, age, and household income across hospitals (all $P<.001$). The first psychiatric admission ($P=.17$) did not show any significant differences nor did BPRS ($P=.45$) or HAM-A ($P=.27$) scores, suggesting general clinical severity was similar. However, significant differences were observed in the number of assessments ($P=.02$), days of observation ($P=.007$), diagnostic groups ($P=.005$), and MADRS ($P=.002$) and YMRS ($P<.001$) scores across the hospitals (Table 1).

XSL•FO
RenderX

Table 1. Basic characteristics and statistical comparison of participants by hospital. The Kruskal-Wallis test was performed for continuous variables, and the chi-square test was performed for categorical variables.

	Total (N=191)	Hospital 1[a] (n=73)	Hospital 2[b] (n=88)	Hospital 3[c] (n=30)	Statistics		*P* value
					Chi-square (*df*)	Kruskal-Wallis test (*H*)	
Sex, n (%)					19.6 (2)	__[d]	<.001
Male	81 (42.4)	17 (23.3)	51 (58.0)	13 (43.3)			
Female	110 (57.6)	56 (76.7)	37 (42.0)	17 (56.7)			
Age (years)					—	21.5	<.001
Mean (SD)	32.6 (13.8)	27.2 (10.7)	35.0 (13.5)	38.7 (17.1)			
Median (IQR)	27 (21-40)	24 (19-29)	30 (22-45)	32 (24-52)			
Diagnostic group, n (%)					14.8 (4)	—	.005
Schizophrenia spectrum	72 (37.7)	19 (26.0)	34 (38.6)	19 (63.3)			
Mood-depressive[e]	93 (48.7)	44 (60.3)	39 (44.3)	10 (33.3)			
Mood-manic[f]	26 (13.6)	10 (13.7)	15 (17.0)	1 (3.3)			
Household income (KRW 5 million=US $3600), n (%)					24.1 (2)	—	<.001
>KRW 5 million	78 (40.8)	32 (43.8)	23 (26.1)	23 (76.7)			
≤KRW 5 million	113 (59.2)	41 (56.2)	65 (73.9)	7 (23.3)			
Education level, n (%)					8.5 (4)	—	.07
≥College	48 (25.1)	16 (21.9)	25 (28.4)	7 (23.3)			
High school	118 (61.8)	41 (56.2)	57 (64.8)	20 (66.7)			
≤Middle school	25 (13.1)	16 (21.9)	6 (6.8)	3 (10.0)			
Number of assessments					—	7.7	.02
Mean (SD)	4.3 (2.9)	4.1 (2.8)	4.6 (3.1)	3.5 (2.8)			
Observational days					—	9.9	.007
Mean (SD)	20.7 (17.5)	19.4 (14.5)	23.4 (19.3)	15.8 (17.5)			
Baseline symptom score, mean (SD)							
BPRS[g]	21.5 (9.1)	22.8 (10.9)	20.6 (8.1)	21.2 (6.7)	—	1.6	.45
HAM-A[h]	11.2 (6.4)	10.2 (4.6)	12.0 (8.2)	11.5 (3.3)	—	2.6	.27
MADRS[i]	18.8 (10.9)	22.0 (9.7)	16.8 (12.4)	16.7 (6.2)	—	12.2	.002
YMRS[j]	11.1 (8.3)	9.8 (8.5)	10.7 (8.4)	15.5 (6.1)	—	17.5	<.001
First psychiatric admission, n (%)					3.5 (2)	—	.17
Yes	66 (34.6)	26 (35.6)	34 (38.6)	6 (20.0)			
No	125 (65.4)	47 (64.4)	54 (61.4)	24 (80.0)			

[a]Hospital 1: Seoul National University Hospital.
[b]Hospital 2: Yongin Mental Hospital.
[c]Hospital 3: Dongguk University Ilsan Hospital.
[d]Not applicable.
[e]Mood-depressive: Mood disorder with current depressive episode.
[f]Mood-manic: Mood disorder with current manic episode.
[g]BPRS: Brief Psychiatric Rating Scale.
[h]HAM-A: Hamilton Anxiety Rating Scale.
[i]MADRS: Montgomery-Asberg Depression Rating Scale.
[j]YMRS: Young Mania Rating Scale.

XSL•FO
RenderX

The group-level mean for the scale scores were 15.9 (SD 9.5) for the BPRS, 8.8 (SD 6.0) for the HAM-A, 14.1 (SD 10.5) for the MADRS, and 8.1 (SD 7.1) for the YMRS, with no significant differences between the cross-validation and external validation sets (all *P*>.05; Table S1 in Multimedia Appendix 1). The deterioration of symptom cases, where the assessment score was increased to that of the previous assessment for an individual, were 22.9% (143/624) for the BPRS, 26.1% (163/624) for the HAM-A, 28.8% (180/624) for the MADRS, and 28.2% (176/624) for the YMRS; no significant differences were observed between the cross-validation and external validation sets (all *P*>.05; Table S2 in Multimedia Appendix 1).

The distribution of the multidimensional sensor data visualized using t-SNE revealed several clusters. Many clusters comprised data originating from 1 or 2 wards. Clusters with overlapping data from all wards were rare (Figure 2).

Figure 2. Visualization of the distribution of sensor data using t-SNE. Each point is color-coded to differentiate between hospital wards—hospital 1: Seoul National University Hospital; hospital 2: male (M) and female (F) wards in Yongin Mental Hospital; and hospital 3: Dongguk University Ilsan Hospital. t-SNE: t-distributed stochastic neighbor embedding.

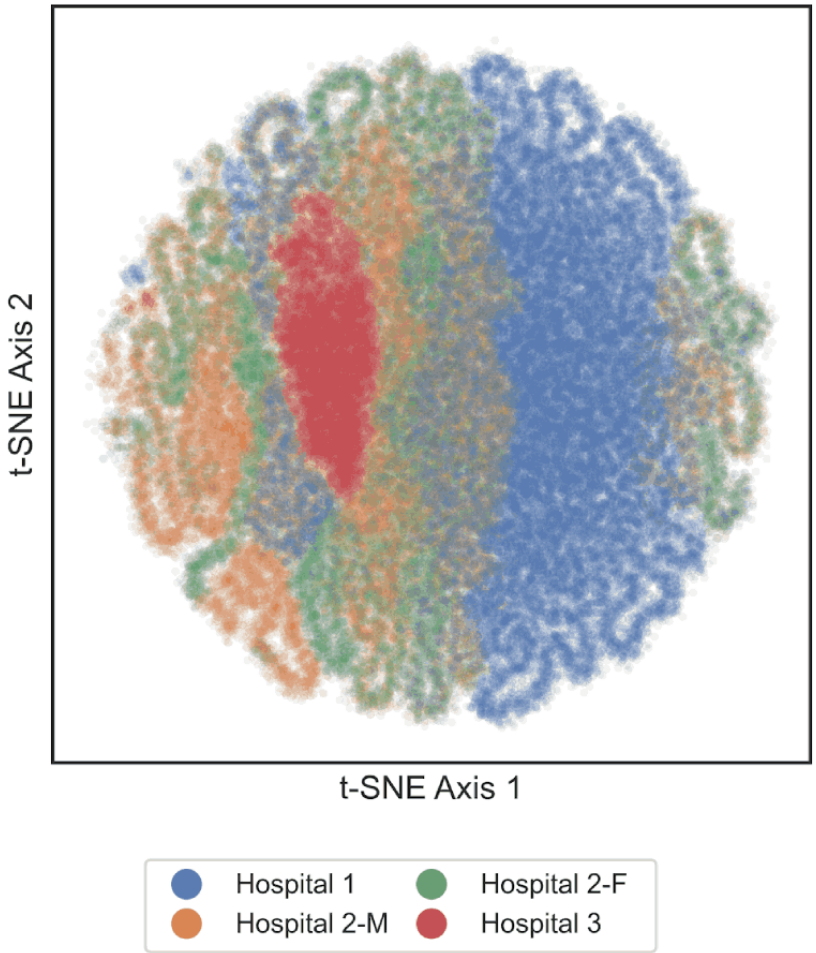

In the Deterioration models, the Single-Deterioration and Multi-Deterioration models showed similar overall performance. Cross-validation revealed that both models achieved an accuracy of 0.75, with an AUC of 0.74 and 0.76, respectively. External validation revealed an accuracy of 0.73 for both models, with an AUC of 0.71 and 0.74, respectively. Among the scales, the BPRS demonstrated the highest accuracy across all models and sets, with accuracy ranging from 0.75 to 0.79 and AUC from 0.76 to 0.82. The MADRS performed the least accurately in cross-validation and the the YMRS performed the least accurately in external validation, regardless of model type (Table 2). The ROC curve for the BPRS in cross-validation showed no clear superiority between Single-Deterioration and Multi-Deterioration models within 1 SD. External validation indicated that the 2 models' curves intersected at both ends of the false positive rate range, with the Multi-Deterioration model generally outperforming in the central portion (0.1 to 0.6; Figure 3). Similar patterns were observed for other scales (Figures S1-3 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 2. Performance of the Deterioration models measured by accuracy and area under the curve (AUC). The Deterioration models predict whether scale scores increased compared with the previous assessment.

	Cross-validation, mean (SD)		External validation	
	Single	Multi	Single	Multi
Accuracy				
Overall	0.75 (0.02)	0.75 (0.03)	0.73	0.73
BPRS[a]	0.79 (0.02)	0.79 (0.03)	0.79	0.75
HAM-A[b]	0.76 (0.05)	0.75 (0.05)	0.76	0.75
MADRS[c]	0.72 (0.03)	0.73 (0.05)	0.74	0.75
YMRS[d]	0.73 (0.04)	0.74 (0.02)	0.65	0.69
AUC				
Overall	0.74 (0.04)	0.76 (0.04)	0.71	0.74
BPRS	0.79 (0.07)	0.80 (0.06)	0.76	0.82
HAM-A	0.75 (0.06)	0.77 (0.05)	0.73	0.75
MADRS	0.69 (0.03)	0.72 (0.05)	0.69	0.73
YMRS	0.73 (0.05)	0.76 (0.04)	0.64	0.67

[a]BPRS: Brief Psychiatric Rating Scale.

[b]HAM-A: Hamilton Anxiety Rating Scale.

[c]MADRS: Montgomery-Asberg Depression Rating Scale.

[d]YMRS: Young Mania Rating Scale.

Figure 3. Receiver operating characteristic curve of the Deterioration models with respect to BPRS scores: (A) cross-validation and (B) external validation. The Deterioration models predict whether BPRS scores increased compared with the previous assessment. Colored areas in cross-validation represent the range of 1 SD. BPRS: Brief Psychiatric Rating Scale.

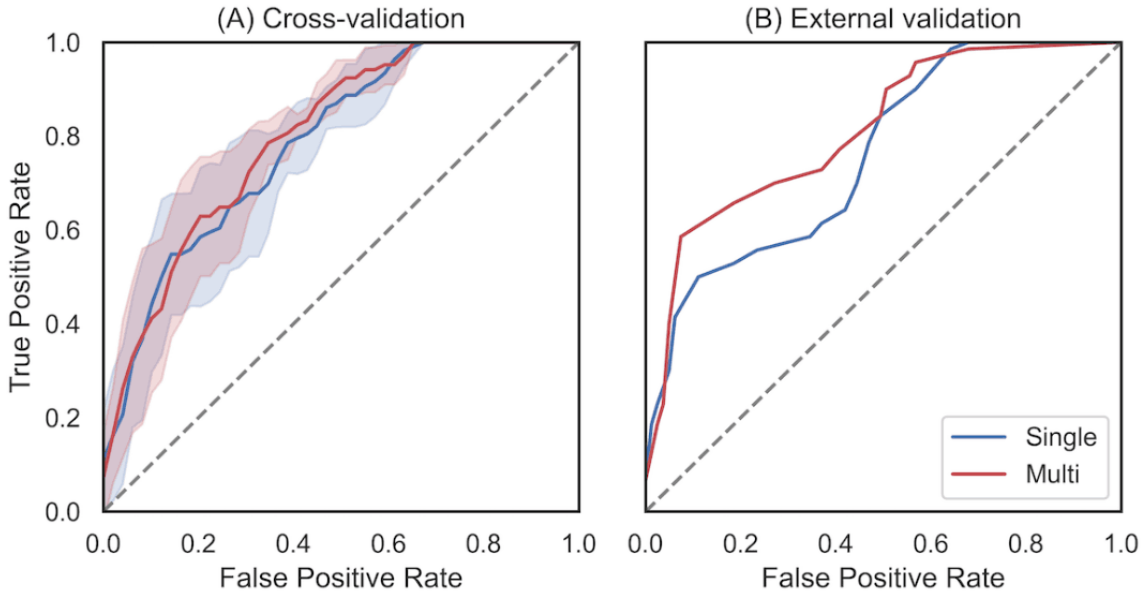

The Multi-Score model outperformed the Single-Score model in both R^2 and NRMSE. In the cross-validation set, the Single-Score and Multi-Score models achieved R^2 values of 0.78 and 0.83 and NRMSE values of 0.072 and 0.064, respectively. External validation showed that the Single-Score and Multi-Score models achieved R^2 values of 0.66 and 0.74 and NRMSE values of 0.067 and 0.056, respectively. Among the scales, the BPRS, which was predicted well in the Deterioration models, demonstrated the lowest R^2 values during external validation: 0.47 for Single-Score and 0.57 for Multi-Score. In contrast, the MADRS and YMRS exhibited R^2 values exceeding 0.85 in the external validation of the Multi-Score model (Figure 4).

XSL•FO
RenderX

Figure 4. Performance of the Score model measured by R2 and NRMSE: (A) cross-validation and (B) external validation. The Score models predict scale scores; the error bar represents an interval of 1 SD. The dotted line indicates an R2 value of 0.7. BPRS: Brief Psychiatric Rating Scale; HAM-A: Hamilton Anxiety Rating Scale; MADRS: Montgomery-Asberg Depression Rating Scale; Multi: model predicting multiple symptoms simultaneously; NRMSE: normalized root mean squared error; Single: model predicting single symptoms individually; YMRS: Young Mania Rating Scale.

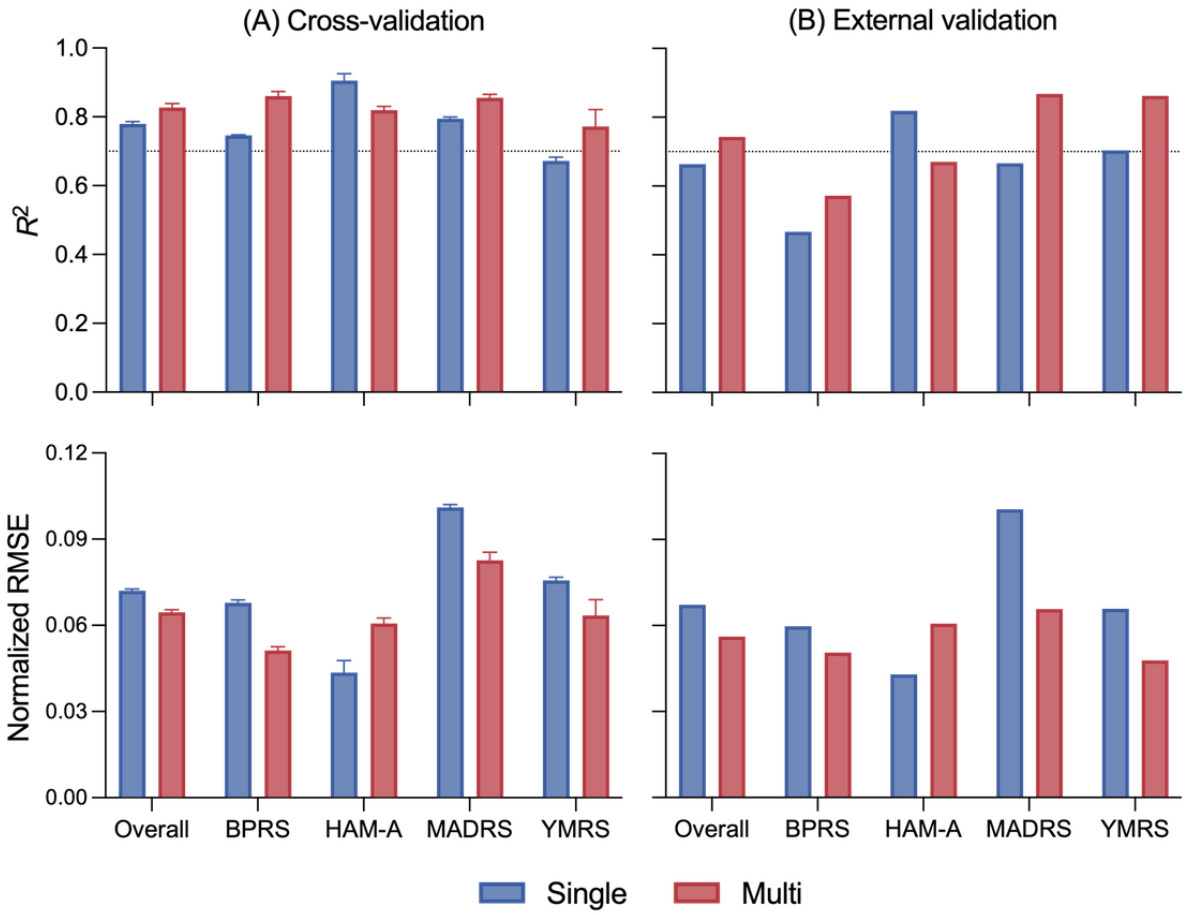

The permutation feature importance analysis indicated that time spent in places other than rooms or hallways (*place_other*, 5 models), the number of unique heart rate values (*nunique_HEARTBEAT*, 4 models), and 2 location entropy indicators (*Location_Entropy_Daily* and *Location_Normalized_Entropy_8hr*, 4 models each) were among the top 5 important features in 4 or more models. Notably, no single feature consistently ranked in the top 5 across more than half of the models. Nonsensor features such as age and sex did not rank in the top 5 important features for the Deterioration models but were frequently included in the Score models (Figure 5).

XSL•FO
RenderX

Figure 5. Permutation feature importance in external validation. The horizontal axis represents 31 individual features, and the vertical axis represents the prediction models. To ensure comparability, the importance values are converted to ranks, with the top 5 annotated. BPRS: Brief Psychiatric Rating Scale; HAM-A: Hamilton Anxiety Rating Scale; MADRS: Montgomery-Asberg Depression Rating Scale; Multi: model predicting multiple symptoms simultaneously; Single: model predicting single symptoms individually; YMRS: Young Mania Rating Scale.

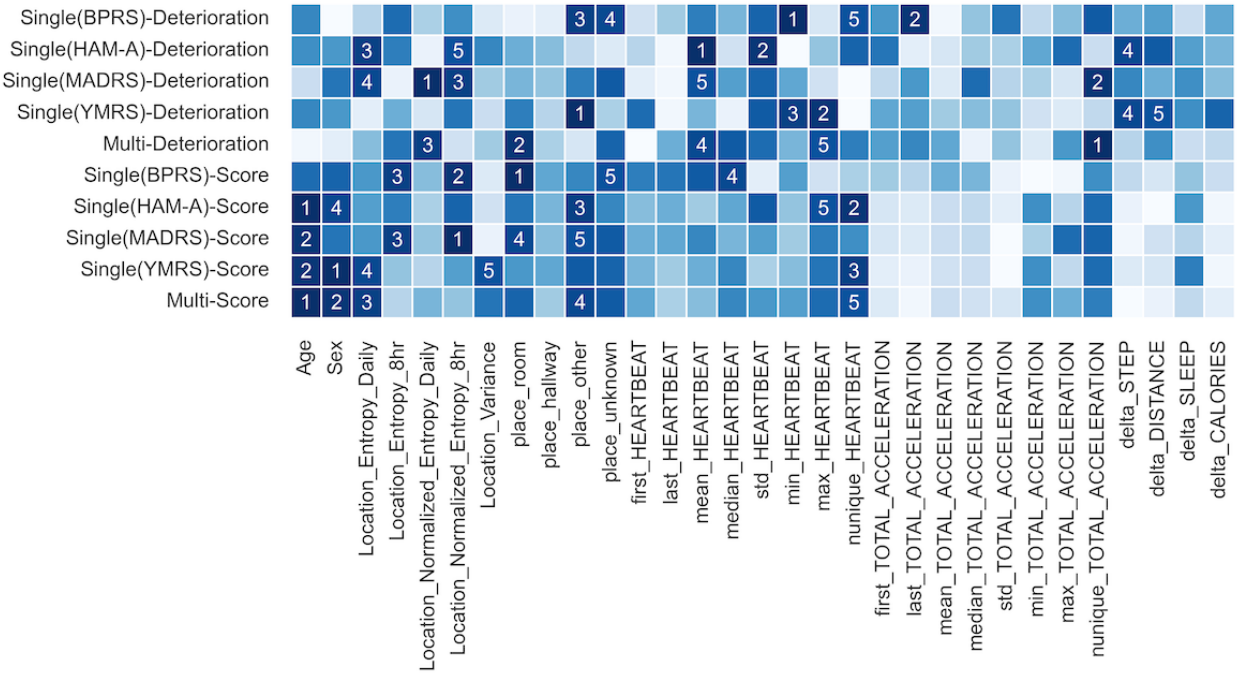

Discussion

Principal Findings

We conducted a longitudinal, prospective, multicenter study of inpatients using wearable devices to develop and validate deep learning models for predicting comprehensive symptoms. To our knowledge, this is the first study to predict clinical rating scale scores of pandiagnostic cases in acute psychiatric wards using wearable devices and deep learning models. While direct benchmarks are unavailable owing to our inpatient setting, we can compare our results to those of similar studies. A meta-analysis of depression prediction studies using wearable devices reported a pooled mean accuracy ranging from 0.70 to 0.89 [52]. Similarly, a meta-analysis of anxiety prediction studies using wearable devices reported a 95% CI for pooled mean accuracy of 0.71 to 0.89 [53]. Our Single-Deterioration and Multi-Deterioration models for MADRS, HAM-A, and overall performance fall within these ranges, demonstrating comparable performance. The Single-Score and Multi-Score models showed overall R^2 values of approximately 0.8 in cross-validation and 0.7 in external validation, indicating substantial explanatory power. In summary, our wearable-based deep learning models effectively predict comprehensive psychiatric symptoms.

Single Versus Multi Models

Our Multi models used an MTL framework to simultaneously predict 4 different scale scores. While Single models might be expected to outperform Multi models, given that they require approximately 4 times the computational resources and parameters to learn and solve the same tasks [47], our findings suggest otherwise. The Single-Deterioration and Multi-Deterioration models, which predict changes from previous assessments for an individual, showed no material differences in both accuracy and AUC. Moreover, Multi-Deterioration models demonstrated a superior balance between sensitivity and specificity, as the AUC value and ROC curve indicated. Considering the class imbalance in symptom deterioration cases, this suggests that MTL-based models provide a more balanced predictive performance. Furthermore, the Single-Score and Multi-Score models, which predict absolute symptom severity, showed that the Multi-Score model demonstrated better explanatory power and lower errors in both cross-validation and external validation.

This study is not the first to attempt to predict mental health indicators using MTL-based models. Several studies have reported significant success using this approach with data from social media [54], electrocardiography [55], functional magnetic resonance imaging [56], and so on. Interestingly, while Harvey et al [57] argued that thousands of samples might be necessary for MTL to show benefits over single-task learning in predicting psychiatric diagnoses using functional magnetic resonance imaging, our study demonstrated these benefits with fewer samples. Although the comparability is limited due to the difference between cross-sectional brain imaging and longitudinal wearable sensor data, a notable distinction is that they predicted diagnoses, whereas we predicted symptoms.

Many contemporary psychopathologists agree that individual psychiatric symptoms form a complex network [58,59]. From this perspective, symptoms are not merely passive manifestations of specific diagnoses but active entities influencing other symptoms [60]. Thus, information learned in predicting one symptom can inform the prediction of others. In

XSL•FO
RenderX

contrast, psychiatric diagnoses are defined through operational criteria of symptom constellations [61]. Consequently, learning shared latent information for multiple symptom prediction may be more feasible and may require fewer samples than those required for predicting multiple diagnoses. A similar study predicting symptoms in individuals with schizophrenia using smartphone passive data found benefits from MTL with only 61 samples [19]. This was likely possible because the prediction targets were individual symptoms rather than diagnoses. Although there are not yet many studies directly comparable with the present research, our study findings suggest that MTL is an effective framework for psychiatric symptom prediction tasks.

Deterioration Versus Score Models

The Deterioration and Score models exhibit several notable differences. The purpose of the Deterioration models was to detect symptom changes within individuals compared with their previous state, whereas the Score models aimed to predict the absolute symptom severity between individuals. For feature importance, no shared patterns of important features were observed between the Deterioration and Score models predicting the same symptom scale. For instance, age and sex emerged as important in most Score models but not in Deterioration models. This is likely because predicting absolute symptom severity relies more on basic demographic information as crucial calibration data compared with predicting relative changes. Performance patterns during external validation also varied across scales. The BPRS demonstrated the best performance among all scales in both cross-validation and external validation for the Deterioration models, but the Score models showed a considerable R^2 reduction in external validation. Conversely, the YMRS exhibited the most significant performance decline in external validation for the Deterioration models but not for the Score models. This indicates that the Deterioration and Score models, even when predicting the same symptom scale, differ in their required information and generalization challenges. This serves as another example demonstrating the distinction between within-subject and between-subject designs [62]. Future studies on predicting symptoms in acute psychiatric inpatients should consider treating relative change and absolute severity as distinct analyses.

Challenges to Generalizability in AI-CDSSs for Acute Psychiatric Wards

The generalizability of models is a critical issue in developing AI models for predicting mental health indicators [63,64]. SAFER's ultimate goal was to develop a predictive model that can serve as the foundation for an AI-CDSS in acute psychiatric wards with broad generalizability. In this study, external validation based on time point maintained considerable performance compared with internal cross-validation, demonstrating the model's temporal robustness at least for the study period. However, recent studies have pointed out that external validation alone cannot fully guarantee reliability and generalizability in clinical prediction models [65]. In this study, challenges to generalizability remain in 2 aspects.

One challenge is interward variations. In this study, participant characteristics and sensor data distributions showed considerable differences across institutions. This was partly inevitable owing to the study's aim to include diverse regions and hospital types. Nevertheless, the observed substantial interward variation suggests that applying an AI-CDSS based on this study to any other facility may encounter significant generalization difficulties. Studies on other clinical prediction models have noted that even models developed from large international multisite clinical trials often see their performance drop to chance levels when subjected to completely independent tests [66]. A potential solution to this issue is a method called recurring local validation or targeted validation [67,68]. This method involves validation and adjustment of prediction models to fit the specific settings, where they will be deployed. Future research should explore whether such ward-specific models could replace a single universal model.

The other challenge is the issue of selection bias [69]. In this study, we recruited pandiagnostic cases by minimizing the exclusion criteria based on clinical conditions and accepting various comorbidities. One reason for needing such an inclusive sample is that the end users of an AI-CDSS in acute psychiatric wards are primarily the ward's medical staff [31]; thus, its usability can only be ensured if it can be applied with minimal restrictions to patients in the ward. However, participants in this study were recruited based on informed consent. Even with minimized exclusion criteria, patients with acute psychiatric symptoms may be subject to selection bias owing to differences in their ability to consent caused by their symptoms, making it difficult to recruit participants in a completely inclusive manner, like those with stroke or Alzheimer disease [70,71]. Therefore, in the future deployment of this model, validation based on data collected close to usual care, similar to the concept of pragmatic trials, will be necessary [72]. Waivers of the requirement for informed consent could be helpful in this case, but they require various ethical and regulatory considerations [73]. One of these is that the collected sensor data contains sensitive personal information, which makes it difficult to analyze through data centralization in one place. A promising solution to this could be engineering techniques such as federated learning, which allows for learning and validation without exporting raw data outside the institution [74].

Limitations

First, although external validation was performed, it was conducted by dividing participants based on recruitment periods rather than by institutions. As confirmed in this study, acute psychiatric facilities varied significantly from one another, resulting in inherent generalizability issues. Thus, while the significance of the external validation is somewhat limited, we discussed various challenges and potential solutions to improve the model's generalizability. Second, this study was conducted in acute psychiatric wards within the context of South Korea's specific ethnic, sociocultural, legislative, and reimbursement systems. These contextual factors surrounding mental health facilities differ considerably between countries [30,75]. Therefore, developing a comprehensive symptom prediction model based on this study in other countries would require substantial consideration of these factors. Third, the average duration of the study was <3 weeks, which is relatively short. Therefore, patients with shorter treatment durations may have

XSL•FO
RenderX

been overrepresented, while relatively few participants provided data over a sufficiently long observation period. Fourth, the sample size was insufficient to fully cover the objectives of this study, which aimed at a diverse population across various settings. In particular, the number of participants older than 40 years or with low educational attainment was small, and the usefulness of our clinical prediction model may be reduced for participants with these demographic characteristics. Fifth, different raters were responsible for clinical assessments at each hospital. Efforts were made to improve measurement reliability through specialized training courses and periodic research conferences; however, interrater errors were not statistically measured or adjusted.

Conclusions

We developed and validated a model to predict multidimensional symptoms in various acute psychiatric wards using wearable sensor data and deep learning models. The constructed model proved effective in predicting a single symptom individually and multiple symptoms simultaneously. Notably, a model that predicted multiple symptoms simultaneously demonstrated more balanced within-subject classification and better between-subject symptom severity prediction. Substantial interward variations were also found in this study, suggesting that generalizability is a key issue. By discussing challenges and solutions for generalizability, we have contemplated the future direction of AI-CDSSs in acute psychiatric wards.

Acknowledgments

The authors appreciate the technical assistance provided by Dr Yunsun Kim at Hallym University Kangnam Sacred Heart Hospital in preparing the figures for this manuscript. This work was supported by the Mental Health Research and Development Project of the Korea Health Industry Development Institute (grant RS-2022-KH129879). The funding body had no role in the study design, data collection, analysis, interpretation, writing of the report, or decision to submit the article for publication. During the preparation of this work, the authors used ChatGPT 4o (OpenAI) [76] to enhance their English style. After using this tool/service, the authors reviewed and edited the content as needed. The authors assume full responsibility for the content of the published article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality assurances given during the informed consent process, as the study involves sensitive personal data from vulnerable participants. The analysis code is publicly available [77].

Authors' Contributions

MH contributed to the conceptualization, methodology, investigation, drafting of the original manuscript, and visualization. RRK was responsible for validation, formal analysis, and manuscript review and editing. JHY assisted with methodology and reviewed and edited the manuscript. SJR contributed to the validation, investigation, and review and editing of the manuscript. HL participated in the investigation and manuscript review and editing. YGK handled formal analysis and reviewed and edited the manuscript. KL provided resources and contributed to manuscript review and editing. HK was responsible for software, resources, data curation, and manuscript review and editing. YSL, TY, and SHK participated in the investigation and manuscript review and editing, with SHK also providing supervision. YMA was involved in conceptualization, supervision, project administration, funding acquisition, and manuscript review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 594 KB - jmir_v26i1e65994_app1.docx]

References

1. Liu F, Gao M, Wu Q, Yan M, Wu R, Shao P, et al. Diagnostic stability in psychiatric patients from hospital admission to discharge: a 10-year retrospective study. Psychiatry Investig 2023 May;20(5):461-470. [doi: 10.30773/pi.2022.0349] [Medline: 37253472]
2. Thrasher TW, Rolli M, Redwood RS, Peterson MJ, Schneider J, Maurer L, et al. 'Medical clearance' of patients with acute mental health needs in the emergency department: a literature review and practice recommendations. WMJ 2019 Dec;118(4):156-163. [Medline: 31978283]
3. Ho BC, Flaum M, Hubbard W, Arndt S, Andreasen NC. Validity of symptom assessment in psychotic disorders: information variance across different sources of history. Schizophr Res 2004 Jun 01;68(2-3):299-307. [doi: 10.1016/j.schres.2003.07.006] [Medline: 15099611]

XSL•FO
RenderX

4. Carpels A, de Smet L, Desplenter S, de Hert M. Falls among psychiatric inpatients: a systematic review of literature. Alpha Psychiatry 2022 Sep;23(5):217-222. [doi: 10.5152/alphapsychiatry.2022.21735] [Medline: 36426266]
5. Kang UG. Psychopathology: A Conceptual Approach. Seoul, South Korea: SNU Press; Dec 10, 2021.
6. Strauss J. Subjectivity and severe psychiatric disorders. Schizophr Bull 2010 Oct 20;37(1):8-13. [doi: 10.1093/schbul/sbq116] [Medline: 20961994]
7. de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago J. Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. Sensors (Basel) 2016 Sep 21;16(9):1538. [doi: 10.3390/s16091538] [Medline: 27657081]
8. Chen L, Xia C, Zhao Z, Fu H, Chen Y. AI-driven sensing technology: review. Sensors (Basel) 2024 May 07;24(10):2958. [doi: 10.3390/s24102958] [Medline: 38793814]
9. Kim D, Min J, Ko SH. Recent developments and future directions of wearable skin biosignal sensors. Adv Sens Res 2023 Oct 24;3(2):2300118. [doi: 10.1002/adsr.202300118]
10. Chia AZR, Zhang MWB. Digital phenotyping in psychiatry: a scoping review. Technol Health Care 2022;30(6):1331-1342. [doi: 10.3233/thc-213648] [Medline: 35661034]
11. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol 2015;33(5):462-463. [doi: 10.1038/nbt.3223] [Medline: 25965751]
12. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318(13):1215-1216 This renowned article by Insel suggests the need and potential for digital phenotyping in acute psychiatric wards. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
13. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023 Dec 13;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]
14. Gurrin C, Smeaton AF, Doherty AR. LifeLogging: Personal Big Data. Norwell, MA: Now Publishers; 2014.
15. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023;25:e44502 [FREE Full text] [doi: 10.2196/44502] [Medline: 37792430]
16. Hsin H, Fromer M, Peterson B, Walter C, Fleck M, Campbell A, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. NPJ Digit Med 2018;1(1):37 [FREE Full text] [doi: 10.1038/s41746-018-0046-0] [Medline: 31304319]
17. Niemeijer K, Mestdagh M, Kuppens P. Tracking subjective sleep quality and mood with mobile sensing: multiverse study. J Med Internet Res 2022;24(3):e25643 [FREE Full text] [doi: 10.2196/25643] [Medline: 35302502]
18. Sakal C, Li T, Li J, Yang C, Li X. Association between sleep efficiency variability and cognition among older adults: cross-sectional accelerometer study. JMIR Aging 2024;7:e54353 [FREE Full text] [doi: 10.2196/54353] [Medline: 38596863]
19. Tseng VWS, Sano A, Ben-Zeev D, Brian R, Campbell AT, Hauser M, et al. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. Sci Rep 2020;10(1):15100 [FREE Full text] [doi: 10.1038/s41598-020-71689-1] [Medline: 32934246]
20. Abd-Alrazaq A, Aslam H, AlSaad R, Alsahli M, Ahmed A, Damseh R, et al. Detection of sleep apnea using wearable AI: systematic review and meta-analysis. J Med Internet Res 2024;26:e58187 [FREE Full text] [doi: 10.2196/58187] [Medline: 39255014]
21. Li X, Kane M, Zhang Y, Sun W, Song Y, Dong S, et al. Circadian rhythm analysis using wearable device data: novel penalized machine learning approach. J Med Internet Res 2021;23(10):e18403 [FREE Full text] [doi: 10.2196/18403] [Medline: 34647895]
22. Gao C, Haghayegh S, Wagner M, Cai R, Hu K, Gao L, et al. Approaches for assessing circadian rest-activity patterns using actigraphy in cohort and population-based studies. Curr Sleep Medicine Rep 2023;9(4):247-256. [doi: 10.1007/s40675-023-00267-4]
23. Haghayegh S, Gao C, Sugg E, Zheng X, Yang HW, Saxena R, et al. Association of rest-activity rhythm and risk of developing dementia or mild cognitive impairment in the middle-aged and older population: prospective cohort study. JMIR Public Health Surveill 2024;10:e55211 [FREE Full text] [doi: 10.2196/55211] [Medline: 38713911]
24. Feng H, Yang L, Liang YY, Ai S, Liu Y, Liu Y, et al. Associations of timing of physical activity with all-cause and cause-specific mortality in a prospective cohort study. Nat Commun 2023;14(1):930 [FREE Full text] [doi: 10.1038/s41467-023-36546-5] [Medline: 36805455]
25. Cao Z, Min J, Chen H, Hou Y, Yang H, Si K, et al. Accelerometer-derived physical activity and mortality in individuals with type 2 diabetes. Nat Commun 2024;15(1):5164 [FREE Full text] [doi: 10.1038/s41467-024-49542-0] [Medline: 38886353]
26. Selenius H, Leppänen Östman S, Strand S. Self-harm as a risk factor for inpatient aggression among women admitted to forensic psychiatric care. Nord J Psychiatry 2016;70(7):554-560. [doi: 10.1080/08039488.2016.1183707] [Medline: 27224513]
27. Brennan G, Flood C, Bowers L. Constraints and blocks to change and improvement on acute psychiatric wards - lessons from the city nurses project. J Psychiatr Ment Health Nurs 2006;13(5):475-482. [doi: 10.1111/j.1365-2850.2006.00956.x] [Medline: 16965464]

XSL•FO
RenderX

28. Sófi G, Törő K, Csikós G, Fliegauf G. Perceiving violence against healthcare workers in a child and adolescent emergency psychiatric ward in Hungary: a qualitative pilot study. Front Psychol 2023;14:1220183 [FREE Full text] [doi: 10.3389/fpsyg.2023.1220183] [Medline: 38023022]
29. Laker C, Cella M, Callard F, Wykes T. Why is change a challenge in acute mental health wards? a cross-sectional investigation of the relationships between burnout, occupational status and nurses' perceptions of barriers to change. Int J Ment Health Nurs 2019;28(1):190-198 [FREE Full text] [doi: 10.1111/inm.12517] [Medline: 29993168]
30. Mental Health, Human Rights and Legislation: Guidance and Practice. Geneva, Switzerland: World Health Organization and the United Nations (represented by the Office of the United Nations High Commissioner for Human Rights); 2023.
31. Westheimer JL, Moukaddam N, Lindsay JA, Sabharwal A, Najafi B, Iacobelli PA, et al. Technology implementation for mental health end users: a model to guide digital transformation for inpatient mental health professionals. JMIR Ment Health 2023;10:e40429 [FREE Full text] [doi: 10.2196/40429] [Medline: 37023415]
32. Langholm C, Kowatsch T, Bucci S, Cipriani A, Torous J. Exploring the potential of apple sensorKit and digital phenotyping data as new digital biomarkers for mental health research. Digit Biomark 2023;7(1):104-114 [FREE Full text] [doi: 10.1159/000530698] [Medline: 37901364]
33. Chrysikou E. Architecture for Psychiatric Environments and Therapeutic Spaces. Amsterdam: IOS Press; 2014.
34. Staggs VS. National trends and variation in nurse staffing on inpatient psychiatric units. Res Nurs Health 2019;42(5):410-415. [doi: 10.1002/nur.21979] [Medline: 31429481]
35. Park S, Park S, Lee YJ, Park CS, Jung YC, Kim S. Nurse staffing and health outcomes of psychiatric inpatients: a secondary analysis of national health insurance claims data. J Korean Acad Nurs 2020;50(3):333-348. [doi: 10.4040/jkan.19203] [Medline: 32632069]
36. Degli Esposti M, Ziauddeen H, Bowes L, Reeves A, Chekroud AM, Humphreys DK, et al. Trends in inpatient care for psychiatric disorders in NHS hospitals across England, 1998/99-2019/20: an observational time series analysis. Soc Psychiatry Psychiatr Epidemiol 2022;57(5):993-1006 [FREE Full text] [doi: 10.1007/s00127-021-02215-5] [Medline: 34951652]
37. Sohn JH, Lee BC, Park S, Oh GH, Kim HR, Kim NY, et al. Health insurance issues and future prospects related to Korean mental health care: adult mental health and hospitalization services. J Korean Neuropsychiatr Assoc 2024;63(2):83-95. [doi: 10.4306/jknpa.2024.63.2.83]
38. Golden G, Popescu C, Israel S, Perlman K, Armstrong C, Fratila R, et al. Applying artificial intelligence to clinical decision support in mental health: what have we learned? Health Policy Technol 2024;13(2):100844. [doi: 10.1016/j.hlpt.2024.100844]
39. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
40. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378 [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
41. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
42. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
43. Raugh IM, James SH, Gonzalez CM, Chapman HC, Cohen AS, Kirkpatrick B, et al. Geolocation as a digital phenotyping measure of negative symptoms and functional outcome. Schizophr Bull 2020;46(6):1596-1607 [FREE Full text] [doi: 10.1093/schbul/sbaa121] [Medline: 32851401]
44. Balliu B, Douglas C, Seok D, Shenhav L, Wu Y, Chatzopoulou D, et al. Personalized mood prediction from patterns of behavior collected with smartphones. NPJ Digit Med 2024;7(1):49 [FREE Full text] [doi: 10.1038/s41746-024-01035-6] [Medline: 38418551]
45. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 2010;11:2287-2322 [FREE Full text] [Medline: 21552465]
46. Jlassi O, Dixon PC. The effect of time normalization and biomechanical signal processing techniques of ground reaction force curves on deep-learning model performance. J Biomech 2024;168:112116 [FREE Full text] [doi: 10.1016/j.jbiomech.2024.112116] [Medline: 38677026]
47. Zhang Y, Yang Q. An overview of multi-task learning. Natl Sci Rev 2018;5(1):30-43. [doi: 10.1093/nsr/nwx105]
48. Edmonds WA, Kennedy TD. An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods. Thousand Oaks, CA: SAGE Publications, Inc; 2024:392.
49. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281-305.
50. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G. PyTorch: an imperative style, high-performance deep learning library. 2019 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2009); Dec 8-14, 2019; Vancouver, BC p. 08-14 URL: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

XSL•FO
RenderX

51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12(85):2825-2830.
52. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023;6(1):84 [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
53. Abd-Alrazaq A, AlSaad R, Harfouche M, Aziz S, Ahmed A, Damseh R, et al. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. J Med Internet Res 2023;25:e48754 [FREE Full text] [doi: 10.2196/48754] [Medline: 37938883]
54. Benton A, Mitchell M, Hovy D. Multi-task learning for mental health using social media text. 2017 Presented at: Proceedings of the 15th Conference of the EACL; December 10, 2017; Valencia, Spain p. 152-162.
55. Zanna K, Sridhar K, Yu H, Sano A. Bias reducing multitask learning on mental health prediction. 2022 Presented at: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII; October 21, 2022; Nara, Japan p. 18-21. [doi: 10.1109/acii55700.2022.9953850]
56. Huang ZA, Liu R, Zhu Z, Tan KC. Multitask learning for joint diagnosis of multiple mental disorders in resting-state fMRI. IEEE Trans Neural Netw Learn Syst 2024;35(6):8161-8175. [doi: 10.1109/TNNLS.2022.3225179] [Medline: 36459608]
57. Harvey A, Moreau CA, Kumar K, Huguet G, Urchs SGW, Sharmarke H. Challenges in multi-task learning for fMRI-based diagnosis: benefits for psychiatric conditions and CNVs would likely require thousands of patients. Imaging Neurosci 2024:1-20.
58. Hofmann SG, Curtiss J, McNally RJ. A complex network perspective on clinical science. Perspect Psychol Sci 2016;11(5):597-605 [FREE Full text] [doi: 10.1177/1745691616639283] [Medline: 27694457]
59. McNally RJ. Network analysis of psychopathology: controversies and challenges. Annu Rev Clin Psychol 2021;17(1):31-53 [FREE Full text] [doi: 10.1146/annurev-clinpsy-081219-092850] [Medline: 33228401]
60. Robinaugh DJ, Hoekstra RHA, Toner ER, Borsboom D. The network approach to psychopathology: a review of the literature 2008-2018 and an agenda for future research. Psychol Med 2020;50(3):353-366 [FREE Full text] [doi: 10.1017/S0033291719003404] [Medline: 31875792]
61. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). Washington, DC: American Psychiatric Association; 2022.
62. Hoffman L. Longitudinal Analysis: Modeling Within-Person Fluctuation and Change. New York, NY: Routledge; 2015.
63. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. Proc ACM Interact Mob Wearable Ubiquitous Technol 2023;6(4):1-34. [doi: 10.1145/3569485]
64. Goetz L, Seedat N, Vandersluis R, van der Schaar M. Generalization-a key challenge for responsible AI in patient-facing clinical applications. NPJ Digit Med 2024;7(1):126 [FREE Full text] [doi: 10.1038/s41746-024-01127-3] [Medline: 38773304]
65. van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023;21(1):70 [FREE Full text] [doi: 10.1186/s12916-023-02779-w] [Medline: 36829188]
66. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science 2024;383(6679):164-167 This paper provides empirical evidence suggesting the need for alternatives to external validation, particularly in the field of mental health AI. [doi: 10.1126/science.adg8538] [Medline: 38207039]
67. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nat Med 2023;29(11):2686-2687. [doi: 10.1038/s41591-023-02540-z] [Medline: 37853136]
68. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res 2022;6(1):24 [FREE Full text] [doi: 10.1186/s41512-022-00136-8] [Medline: 36550534]
69. Wan B, Caffo B, Vedula SS. A unified framework on generalizability of clinical prediction models. Front Artif Intell 2022;5:872720 [FREE Full text] [doi: 10.3389/frai.2022.872720] [Medline: 35573904]
70. Hotter B, Ulm L, Hoffmann S, Katan M, Montaner J, Bustamante A, et al. Selection bias in clinical stroke trials depending on ability to consent. BMC Neurol 2017;17(1):206 [FREE Full text] [doi: 10.1186/s12883-017-0989-9] [Medline: 29202730]
71. Wilkins JM, Forester BP. Informed consent, therapeutic misconception, and clinical trials for alzheimer's disease. Int J Geriatr Psychiatry 2020;35(5):430-435 [FREE Full text] [doi: 10.1002/gps.5262] [Medline: 31895474]
72. Ford I, Norrie J. Pragmatic trials. N Engl J Med 2016;375(5):454-463 [FREE Full text] [doi: 10.1056/NEJMra1510059] [Medline: 27518663]
73. Largent EA, Joffe S, Dickert NW, Morain SR. The ethical value of consulting community members in non-emergency trials conducted with waivers of informed consent for research. Clin Trials 2024:17407745241259360 [FREE Full text] [doi: 10.1177/17407745241259360] [Medline: 38916109]
74. Kim YG, Yang S, Lee K. Multicenter analysis of emergency patient severity through local model evaluation client selection: optimizing client selection based on local model evaluation. Appl Sci 2024;14(16):6876. [doi: 10.3390/app14166876]
75. Joe KH, Jang M. Comparison of mental health act about involuntary admission among 4 East Asian countries. J Korean Neuropsychiatr Assoc 2019;58(4):297-313. [doi: 10.4306/jknpa.2019.58.4.297]

XSL•FO
RenderX

76. ChatGPT. OpenAI. URL: https://chatgpt.com/?model=gpt-4o [accessed 2024-11-04]
77. Source code repository for "Comprehensive Symptom Prediction in Acute Psychiatric Inpatients Using Wearable-Based Deep Learning Models: Development and Validation Study". GitHub. URL: https://github.com/gachon-CCLab/SAFER-Symptom-Prediction.git [accessed 2024-10-30]

Abbreviations

AI: artificial intelligence
AI-CDSS: artificial intelligence–assisted clinical decision support system
AUC: area under the curve
BPRS: Brief Psychiatric Rating Scale
HAM-A: Hamilton Anxiety Scale
MADRS: Montgomery-Asberg Depression Rating Scale
MTL: multitask learning
NRMSE: normalized root mean squared error
ROC: receiver operating characteristic
SAFER: Sensor Application for Early Response in Closed Wards
t-SNE: t-distributed stochastic neighbor embedding
YMRS: Young Mania Rating Scale

Edited by G Eysenbach, A Coristine; submitted 31.08.24; peer-reviewed by C Sakal, L Yang; comments to author 30.09.24; revised version received 20.10.24; accepted 20.10.24; published 13.11.24.

Please cite as:
Hong M, Kang RR, Yang JH, Rhee SJ, Lee H, Kim YG, Lee K, Kim H, Lee YS, Youn T, Kim SH, Ahn YM
Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study
J Med Internet Res 2024;26:e65994
URL: https://www.jmir.org/2024/1/e65994
doi:10.2196/65994
PMID:

©Minseok Hong, Ri-Ra Kang, Jeong Hun Yang, Sang Jin Rhee, Hyunju Lee, Yong-gyom Kim, KangYoon Lee, HongGi Kim, Yu Sang Lee, Tak Youn, Se Hyun Kim, Yong Min Ahn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Machine Learning–Based Early Warning Model for Lichenoid Vulvar Disease: Prediction Model Development Study

Jian Meng[1], MM; Xiaoyu Niu[1], MD, PhD; Can Luo[1], MM; Yueyue Chen[1], MD; Qiao Li[1], MM; Dongmei Wei[1], MD

Department of Obstetrics and Gynecology, West China Second Hospital, Sichuan University, Chengdu, China

Corresponding Author:
Dongmei Wei, MD
Department of Obstetrics and Gynecology
West China Second Hospital
Sichuan University
No.20, section 3, South Renmin Road
Wuhou District, Sichuan Province
Chengdu, 610041
China
Phone: 86 18982150655
Email: weidongmei@scu.edu.cn

Abstract

Background: Given the complexity and diversity of lichenoid vulvar disease (LVD) risk factors, it is crucial to actively explore these factors and construct personalized warning models using relevant clinical variables to assess disease risk in patients. Yet, to date, there has been insufficient research, both nationwide and internationally, on risk factors and warning models for LVD. In light of these gaps, this study represents the first systematic exploration of the risk factors associated with LVD.

Objective: The risk factors of LVD in women were explored and a medically evidence-based warning model was constructed to provide an early alert tool for the high-risk target population. The model can be applied in the clinic to identify high-risk patients and evaluate its accuracy and practicality in predicting LVD in women. Simultaneously, it can also enhance the diagnostic and treatment proficiency of medical personnel in primary community health service centers, which is of great significance in reducing overall health care spending and disease burden.

Methods: A total of 2990 patients who attended West China Second Hospital of Sichuan University from January 2013 to December 2017 were selected as the study candidates and were divided into 1218 cases in the normal vulvovagina group (group 0) and 1772 cases in the lichenoid vulvar disease group (group 1) according to the results of the case examination. We investigated and collected routine examination data from patients for intergroup comparisons, included factors with significant differences in multifactorial analysis, and constructed logistic regression, random forests, gradient boosting machine (GBM), adaboost, eXtreme Gradient Boosting, and Categorical Boosting analysis models. The predictive efficacy of these six models was evaluated using receiver operating characteristic curve and area under the curve.

Results: Univariate analysis revealed that vaginitis, urinary incontinence, humidity of the long-term residential environment, spicy dietary habits, regular intake of coffee or caffeinated beverages, daily sleep duration, diabetes mellitus, smoking history, presence of autoimmune diseases, menopausal status, and hypertension were all significant risk factors affecting female LVD. Furthermore, the area under the receiver operating characteristic curve, accuracy, sensitivity, and F_1-score of the GBM warning model were notably higher than the other 5 predictive analysis models. The GBM analysis model indicated that menopausal status had the strongest impact on female LVD, showing a positive correlation, followed by the presence of autoimmune diseases, which also displayed a positive dependency.

Conclusions: In accordance with evidence-based medicine, the construction of a predictive warning model for female LVD can be used to identify high-risk populations at an early stage, aiding in the formulation of effective preventive measures, which is of paramount importance for reducing the incidence of LVD in women.

(J Med Internet Res 2024;26:e55734) doi:10.2196/55734

XSL•FO
RenderX

KEYWORDS

female; lichenoid vulvar disease; risk factors; evidence-based medicine; early warning model

Introduction

In different time periods, lichenoid vulvar disease (LVD) has been variously named such as white spot disease, or vulvar white lesions. The term LVD was coined by the International Society for the Study of Vulvovaginal Disease in 1975. LVD is classified into 3 pathological types, namely vulvar squamous cell hyperplasia type, lichen sclerosus type, and mixed type [1,2]. It primarily manifests as mucocutaneous hypopigmentation of vulva to gray-brown or whitening, roughening of skin with lichenification or reduction of elasticity with atrophic thinning and chapping, and introital stenosis. It is characterized by pruritus or pain, regardless of seasons or day-night variations, with exacerbation often occurring at night [3]. The lesion locations include the labia majora and minora, the clitoral region, perianal area, and the pubic hairline. LVD is often recurrent, and in the later stage, it can lead to vaginal stenosis, causing dyspareunia and even carrying the risk of carcinoma, which severely impacts the physical and psychological well-being of the patients [4-6]. It is evident that Chinese women are often reluctant to seek medical care from formal health care facilities due to economic, cultural and traditional concept, and other factors. Clinical practice, both domestically and internationally, has shown that early warning, timely diagnosis, prompt intervention, and immediate treatment are effective strategies in preventing the development of LVD [6,7]. The early identification of high-risk individuals and proactive prevention of LVD have garnered increasing attention. However, current research on the early identification of risk factors for LVD remains inadequate, lacking systematic and comprehensive approaches. For instance, Sideri et al [8] conducted a case-control study with 75 cases and 225 age-matched controls to assess risk factors for LVD. The study collected data on personal characteristics and habits, gynecologic and obstetric history, general sexual habits, and selected dietary habits. The findings indicated a higher risk of vulvar lichen sclerosis (VLS) in parous women compared with nulliparous women, with no significant increase in risk correlating with the number of births. Smith and Haefner [9] highlighted the association of VLS with autoimmune conditions, genetic factors, hormone levels, and infections. These studies were limited to identifying high-risk factors and did not develop early warning models.

Given the complexity and diversity of LVD risk factors, it is crucial to actively explore these factors and construct personalized warning models using relevant clinical variables to assess disease risk in patients. Yet, to date, there has been insufficient research by scholars, both domestically and internationally, on the risk factors and warning models for LVD. In light of these gaps, this study represents the first systematic exploration of the risk factors associated with LVD. Using an evidence-based medicine approach, a warning model has been constructed, offering a tool for the early identification of high-risk individuals. Furthermore, this model has the potential to enhance the diagnostic and treatment capabilities of medical personnel in primary community health service centers. This holds great significance in mitigating overall health care expenditures and the disease burden. More importantly, our research is the first to apply machine learning techniques to predict LVD, filling a novel research gap in the field.

Methods

Study Candidates

The study incorporated 2634 female patients who attended the West China Second Hospital of Sichuan University from January 2013 to December 2017. Based on the results of their medical examinations, these patients were divided into 2 groups, that is, the normal vulvovagina group (group 0) consisting of 1080 cases and the LVD group (group 1) comprising 1554 cases. Inclusion criteria for the study encompassed female patients who were definitively diagnosed with LVD, were fully informed, voluntarily participated in the study, and could undergo long-term follow-ups. Exclusion criteria were as follows: (1) patients with mental illness history, cerebral infarction, or other conditions that may lead to communication difficulties; (2) patients in the acute phase of the disease; and (3) patients diagnosed with malignant tumors.

Diagnosis of LVD

Per the 2021 European Guideline for the Management of Vulval Conditions, patients can be diagnosed with LVD according to the characteristic clinical manifestations and signs, including symptoms such as vulvar pain and itching (with severe cases leading to scratching), a burning sensation in the vulvovaginal area, abnormal vaginal discharge, dyspareunia, or even dysuria, along with signs such as vulvar skin keratinization hypertrophy, coarsening and thickening of the cortex or with part atrophy thinning, reduced or absent skin elasticity, and in severe cases, chapping and festering, hypopigmentation, or whitening of the skin [10]. In this study, however, the diagnostic criteria for LVD involve not only the characteristic clinical appearances but also require confirmation through biopsy. Conducting a biopsy before treatment is advantageous for establishing a definitive histopathological record for the diagnosis.

Collection of Clinical Information

Based on previous domestic and international research findings combined with clinical experience, we selected and collected clinical information related to female LVD [11-14]. This included factors such as age, educational attainment, vaginitis, urinary incontinence, humidity of the long-term residential environment, type of long-term residence (urban or rural), spicy dietary habits, regular intake of coffee or caffeinated beverages, good sleep habits, daily sleep duration (<8 hours or ≥8 hours), diabetes mellitus, smoking history (≥100 cigarettes smoked in lifetime) [15], alcohol consumption, presence of autoimmune diseases, human papilloma virus status, presence of allergic diseases, menopausal status, hypertension, frequent seafood consumption, and income level.

XSL•FO
RenderX

Statistical Analysis

In this study, the K-Nearest Neighbor imputer method was applied to fill in missing variables. To ensure the accuracy of the results after filling, variables with a missing rate greater than 50% (such as occupation, income level, educational attainment, age, and frequent consumption of seafood or high-fat foods) were excluded from modeling. Risk factors associated with female LVD were obtained after univariate analysis of all variables on the postfilled data. Variables that exhibited significance in the univariate statistical test were included in the recursive feature elimination for multivariate selection, and the area under the curve (AUC) value was highest when 10 variables were included [16]. Subsequently, these selected variables were used for machine learning modeling, resulting in the construction of logistic regression (LR), random forests (RF), gradient boosting machine (GBM), adaptive boosting (ADA), eXtreme gradient boosting (XGB), and categorical boosting (CatBoost) analysis models. The rationale for selecting these models lies in their exemplification of fundamental and archetypal algorithmic types within machine learning, LR for classification model, RF for bagging model, ADA for boosting model, and GBM for gradient boosting tree model. Widely applied across diverse domains and tasks, these algorithms consistently demonstrate robust performance. Furthermore, XGB and CatBoost represent advanced models that refine and enhance these classic algorithms, showcasing notable prowess, especially within the gradient boosting tree category (Multimedia Appendix 1) [17]. The predictive efficacy of these 6 models was assessed using receiver operating characteristic (ROC) curves, accuracy, sensitivity, and specificity. Based on the above best model, the correlation ranking between each variable and the target variables was investigated using SHapley Additive exPlanations (SHAP) plots. The details are shown in the study design workflow (Figure 1).

Figure 1. Study design. The early warning system model was built on the retrospective cohort (n=2990) and validated on the prospective cohort (n=356). LVD: Lichenoid Vulvar Disease; LR: Logistic Regression; RF: Random Forest; GBM: Gradient Boosting Machine; ADA: AdaBoost; XGB: Extreme Gradient Boosting; CatBoost: Categorical Boosting; AUC: Area Under the Curve.

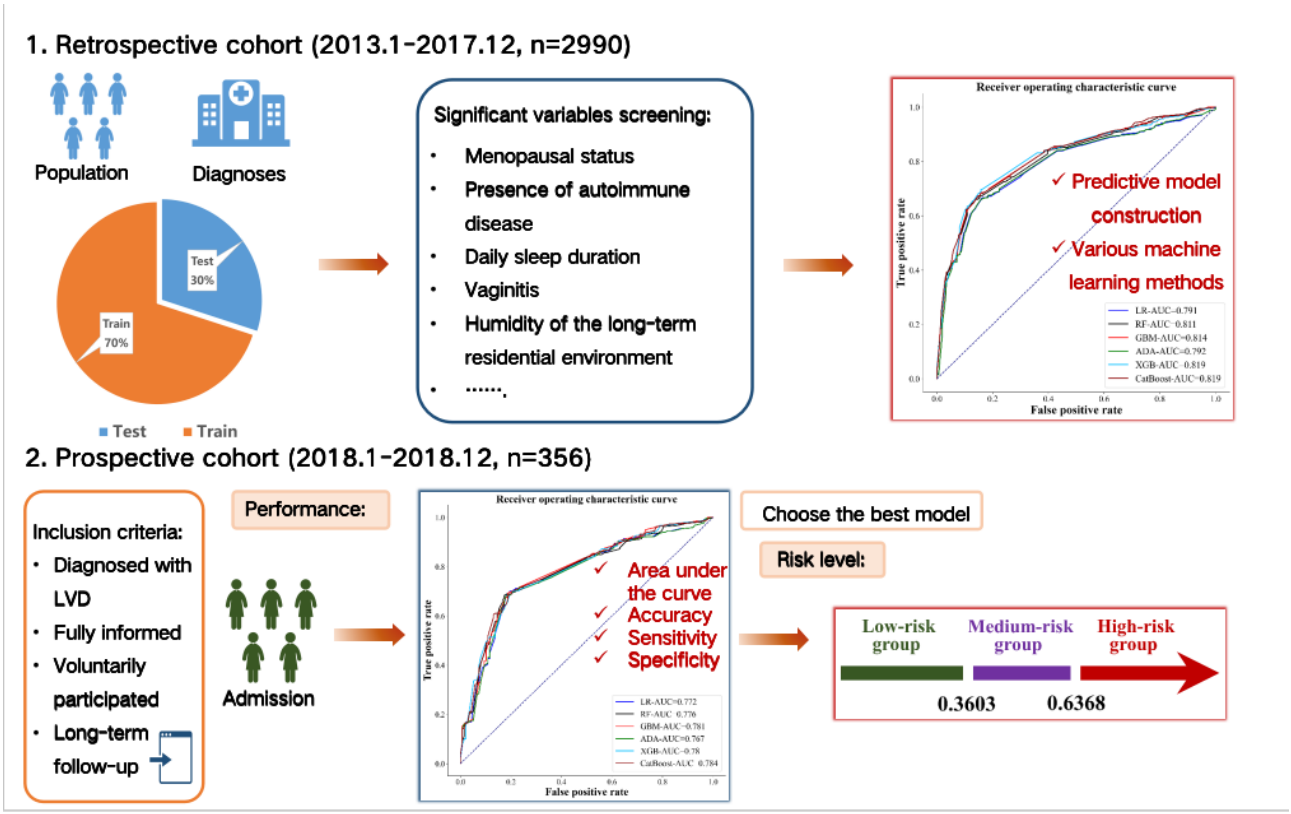

Ethical Considerations

Ethical approval was granted by the Ethics Committee of West China Second Hospital of Sichuan University (2023179) on September 21, 2023. Informed consent was obtained from all participants. All procedures were performed in accordance with relevant guidelines.

Results

Univariate Analysis of Risk Factors for LVD

According to the results of case examinations, the study cohort was stratified into 2 groups, the normal vulvovagina group (group 0) with 1080 cases and the LVD group (group 1) with 1554 cases. The univariate analysis was performed, revealing statistically significant differences between the 2 groups in various factors, including vaginitis (P<.001), humidity of the long-term residential environment (P=.01), spicy dietary habits (P=.00), regular intake of coffee or caffeinated beverages (P=.00), daily sleep duration (P<.001), diabetes mellitus

XSL•FO
RenderX

(*P*<.001), smoking history (*P*<.001), presence of autoimmune diseases (*P*<.001), menopausal status (*P*<.001), and hypertension (*P*<.001). Conversely, no statistically significant discrepancies were observed between the 2 groups in other baseline data, such as urinary incontinence (*P*=.07), type of long-term residence (*P*=.11), good sleep habits (*P*=.24), alcohol consumption (*P*=.42), and presence of allergic diseases (*P*=.52). Detailed findings are presented in Table 1.

Table 1. Univariate analysis of risk factors for lichenoid vulvar disease.

Parameter	Total (N=2634), n (%)	Group 0 (n=1080), n (%)	Group 1 (n=1554), n (%)	Chi-square	*P* value
Vaginitis	860 (32.65)	396 (36.667)	464 (29.858)	13.1	<.001
Urinary incontinence	275 (10.44)	98 (9.074)	177 (11.39)	3.4	.07
Humidity of the long-term residential environment	282 (10.706)	136 (12.593)	146 (9.395)	6.4	<.01
Type of long-term residence	2356 (89.446)	953 (88.241)	1403 (90.283)	2.6	.11
Spicy dietary habits	1365 (51.822)	601 (55.648)	764 (49.163)	10.4	<.001
Regular intake of coffee or caffeinated beverages	235 (8.922)	119 (11.019)	116 (7.465)	9.4	<.001
Good sleep habits	1607 (61.01)	644 (59.63)	963 (61.969)	1.3	.24
Daily sleep duration	921 (34.966)	263 (24.352)	658 (42.342)	89.9	<.001
Diabetes mellitus	112 (4.252)	15 (1.389)	97 (6.242)	35.7	<.001
Smoking history	86 (3.265)	55 (5.093)	31 (1.995)	18.4	<.001
Alcohol consumption	300 (11.39)	130 (12.037)	170 (10.94)	0.6	.42
Presence of autoimmune diseases	2070 (78.588)	689 (63.796)	1381 (88.867)	236.5	<.001
Presence of allergic diseases	229 (8.694)	99 (9.167)	130 (8.366)	0.4	.52
Menopausal status	896 (34.017)	141 (13.056)	755 (48.584)	356.7	<.001
Hypertension	50 (1.898)	31 (2.87)	19 (1.223)	8.4	<.001

A Risk Model to Predict LVD

The establishment of the warning model begins with the foundation of massive data provided by large-sample retrospective clinical profiles, as the basis for evaluating numerous variables and screening out truly meaningful risk factors. Subsequently, the construction of machine learning models serves as the cornerstone for developing the early warning model. Further validation of the model's accuracy is performed in clinical follow-up cohorts and prospective studies to establish the genuinely significant warning model. Since 2013, the pelvic floor group in West China Second Hospital of Sichuan University, has set up a big data cloud platform encompassing data from tens of thousands of patients with LVD, which provides a substantially big data foundation for the development of early warning and diagnostic model for LVD.

The variables that showed significance in the univariate statistical tests were included in the recursive feature elimination multifactorial screening. As depicted in Figure 2, AUC was used as the evaluation metric, and the highest AUC score with 0.794 was achieved when 10 variables were included. In Table 2, the predictive models were constructed using various machine learning methods, including LR, RF, GBM, ADA, XGB, and CatBoost analysis models, based on the above selected 10 variables. Optimal parameters for each model were determined using grid search method combined with 5-fold cross validation. A comprehensive elucidation of the model principles and parameter selections is provided in the appendix. Results indicated superior performance of the GBM model, yielding an accuracy value of 0.747, sensitivity of 0.703, and specificity of 0.818. The ROC curves demonstrated that the AUC score of the GBM model was 0.814, as illustrated in Figure 3. In summary, the GBM warning model exhibited superior predictive efficacy, with higher values for the area under the ROC curve, accuracy, sensitivity, and specificity compared with the other 5 warning models.

XSL•FO
RenderX

Figure 2. Recursive feature elimination variables screening curve for lichenoid vulvar disease (defined as group 1). AUC: Area under the curve.

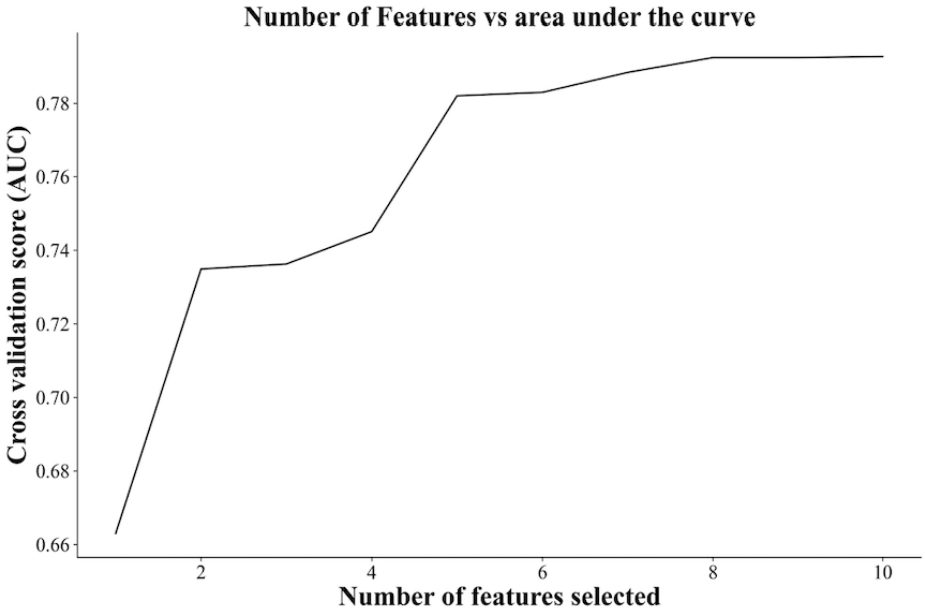

Table 2. Results of the prediction efficacy assessment of 6 lichenoid vulvar disease prediction models.

Model[a]	Area under the curve, mean (95% CI)	Accuracy, mean (95% CI)	Sensitivity, mean (95% CI)	Specificity, mean (95% CI)	F_1-score, mean (95% CI)
LR[b]	0.79 (0.765-0.818)	0.733 (0.703-0.752)	0.667 (0.629-0.703)	0.814 (0.777-0.86)	0.748 (0.717-0.768)
RF[c]	0.811 (0.784-0.839)	0.738 (0.712-0.767)	0.741 (0.703-0.773)	0.738 (0.691-0.78)	0.771 (0.735-0.796)
GBM[d]	0.814 (0.786-0.842)	0.747 (0.711-0.772)	0.703 (0.642-0.732)	0.818 (0.779-0.857)	0.766 (0.726-0.79)
ADA[e]	0.793 (0.763-0.825)	0.728 (0.705-0.764)	0.695 (0.661-0.733)	0.776 (0.736-0.815)	0.753 (0.721-0.785)
XGB[f]	0.818 (0.788-0.845)	0.751 (0.712-0.776)	0.672 (0.628-0.71)	0.862 (0.829-0.895)	0.761 (0.729-0.793)
CatBoost[g]	0.82 (0.791-0.847)	0.75 (0.722-0.778)	0.678 (0.648-0.713)	0.846 (0.818-0.882)	0.760 (0.73-0.787)

[a]The warning model was constructed using prediction models such as LR, RF, GBM, ADA, XGB, and CatBoost. The dataset was divided into a training set and a test set in a 7:3 ratio. The models were trained using the training set and evaluated using the test set. The numbers in parentheses represent the 95% CIs derived from 100 bootstrap iterations.

[b]LR: logistic regression.

[c]RF: random forests.

[d]GBM: gradient boosting machine.

[e]ADA: adaptive boosting.

[f]XGB: eXtreme gradient boosting.

[g]CatBoost: categorical boosting.

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curves for 6 lichenoid vulvar disease prediction models. LR: Logistic Regression; RF: Random Forest; GBM: Gradient Boosting Machine; ADA: AdaBoost; XGB: Extreme Gradient Boosting; CatBoost: Categorical Boosting; AUC: Area Under the Curve.

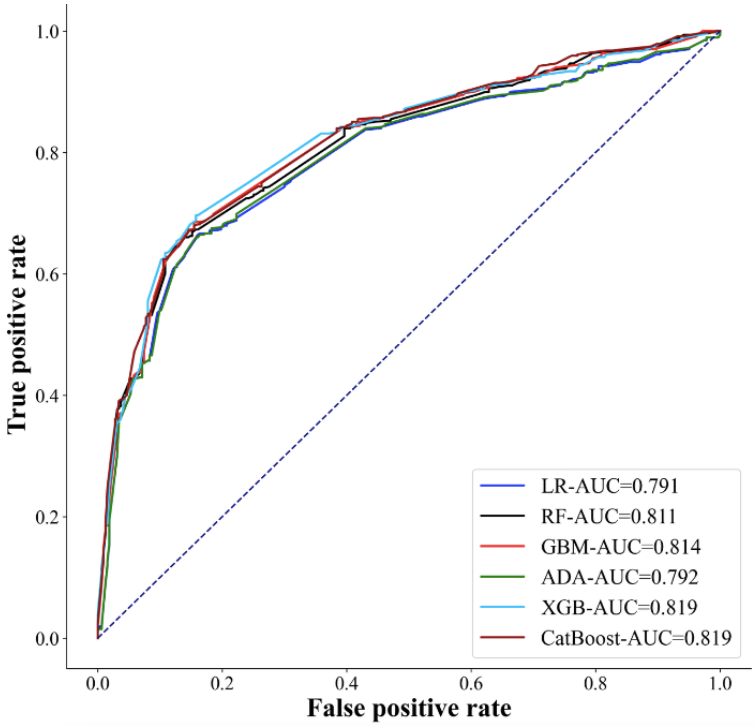

Shapley Additive Explanations Correlation Analysis

To further analyze the impact of multiple factors on the outcome variable within the GBM model, SHAP plots were generated. These plots were constructed to interpret the correlations between various factors and the target variables. Notably, it was observed that menopausal status exhibited the strongest positive dependence with the target variables. Following this, the presence of autoimmune diseases also displayed a positive relevance. Daily sleep duration was positively correlated with LVD, while vaginitis showed a negative association. Further details of the correlation results are presented in Figure 4.

Figure 4. (A) Shapley additive explanations (SHAP) summary plot of relevant factors affecting lichenoid vulvar disease based on the gradient boosting machine model. (B) Importance ranking of relevant factors affecting lichenoid vulvar disease.

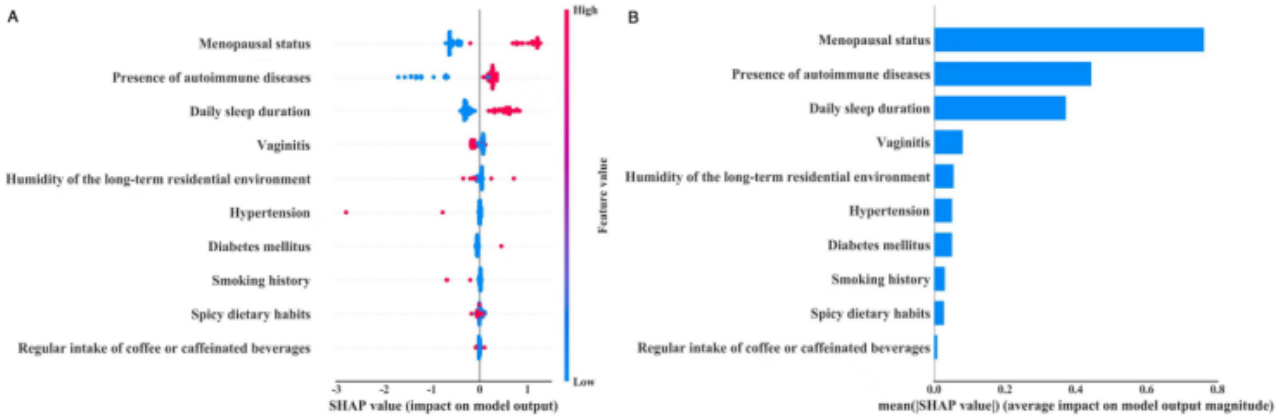

External Validation of the LVD Warning Model

To verify the effectiveness of the model, it was first internally validated with the ROC curve, and the results indicated the AUC value of 0.814, as illustrated in Table 2. Furthermore, external validation was performed using baseline data from 356 patients in 2018, and the results showed that the corresponding GBM model achieved an AUC score of 0.778, an accuracy of 0.732, sensitivity of 0.713, and specificity of 0.767. In comparison with the internal validation, the AUC, accuracy, and specificity values only decreased by 0.036, 0.015, and 0.051, respectively, while the sensitivity value improved by 0.1. Therefore, it is evident that the constructed risk warning model exhibits favorable predictive efficacy, as depicted in Figure 5 and Table 3. Afterwards, the GBM warning model was applied to predict the training dataset and obtain predictive probabilities. Subsequently, k-means algorithm was used to cluster those probabilities, yielding 3 cluster centers, 0.45529467, 0.81832383, and 0.26538643. The thresholds for 3 risk zones were calculated as 0.3603 and 0.6368. In short, Figure 6 displays the GBM warning model's classification of individuals into 3 groups based on predictions, those below 0.3603 as the low-risk

XSL•FO
RenderX

group, those above 0.6368 as the high-risk group, and those within the range of 0.3603 to 0.6368 as the medium-risk group. The use of population-related variables as input for the alert model led to the generation of predicted values, depicting varying degrees of risk, as demonstrated in Figure 7.

Figure 5. Receiver operating characteristic curves for 6 lichenoid vulvar disease prediction models with external validation. LR: Logistic Regression; RF: Random Forest; GBM: Gradient Boosting Machine; ADA: AdaBoost; XGB: Extreme Gradient Boosting; CatBoost: Categorical Boosting; AUC: Area Under the Curve.

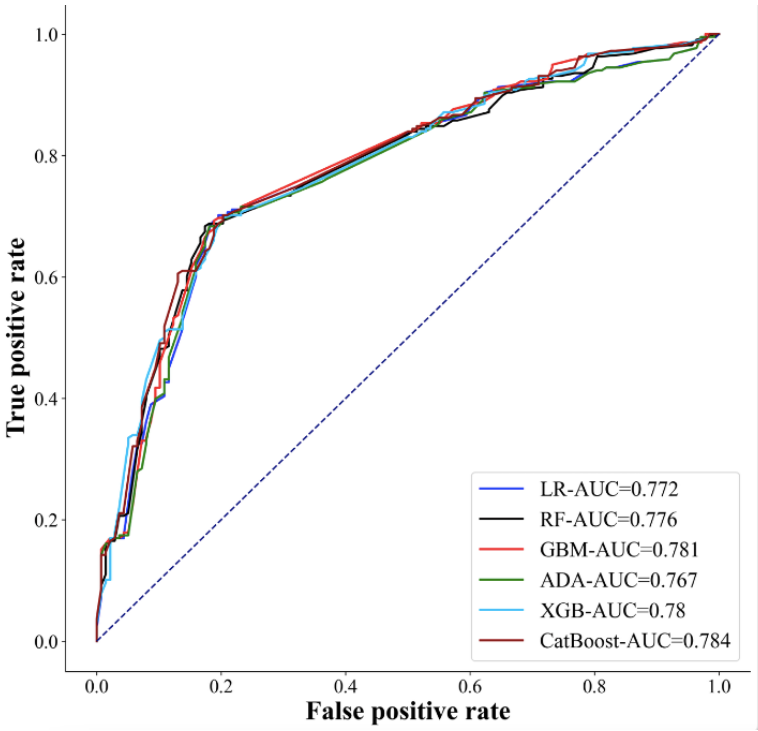

Table 3. Results of the prediction efficacy assessment of 6 lichenoid vulvar disease prediction models.

Model	Area under the curve, mean (95% CI)	Accuracy, mean (95% CI)	Sensitivity, mean (95% CI)	Specificity, mean (95% CI)	F_1-score, mean (95% CI)
LR[a]	0.778 (0.72-0.818)	0.742 (0.701-0.784)	0.703 (0.645-0.767)	0.808 (0.734-0.864)	0.771 (0.729-0.809)
RF[b]	0.776 (0.735-0.822)	0.722 (0.685-0.764)	0.743 (0.674-0.808)	0.682 (0.628-0.771)	0.765 (0.72-0.808)
GBM[c]	0.778 (0.727-0.822)	0.732 (0.684-0.77)	0.713 (0.667-0.765)	0.767 (0.678-0.826)	0.764 (0.719-0.803)
ADA[d]	0.766 (0.715-0.816)	0.73 (0.682-0.772)	0.713 (0.651-0.772)	0.755 (0.681-0.831)	0.767 (0.714-0.804)
XGB[e]	0.772 (0.722-0.819)	0.712 (0.664-0.763)	0.65 (0.59-0.722)	0.808 (0.744-0.879)	0.734 (0.688-0.784)
Cat-Boost[f]	0.787 (0.741-0.829)	0.737 (0.681-0.783)	0.696 (0.62-0.754)	0.802 (0.731-0.869)	0.764 (0.708-0.809)

[a]LR: logistic regression.
[b]RF: random forest.
[c]GBM: gradient boosting machine.
[d]ADA: adaptive boosting.
[e]XGB: eXtreme gradient boosting.
[f]CatBoost: categorical boosting.

XSL•FO
RenderX

Figure 6. The predictive outcome range of the gradient boosting machine warning model.

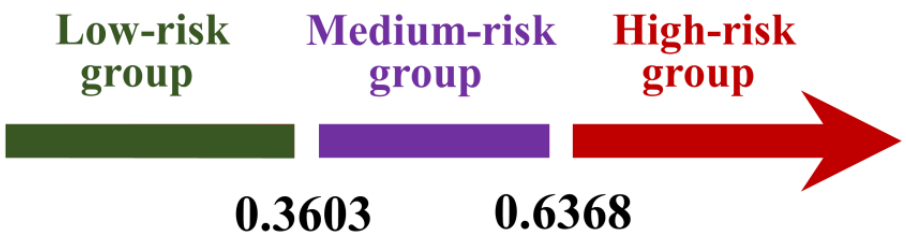

Figure 7. (A) Prediction of low-risk group. (B) Prediction of medium-risk group. (C) Prediction of high-risk group.

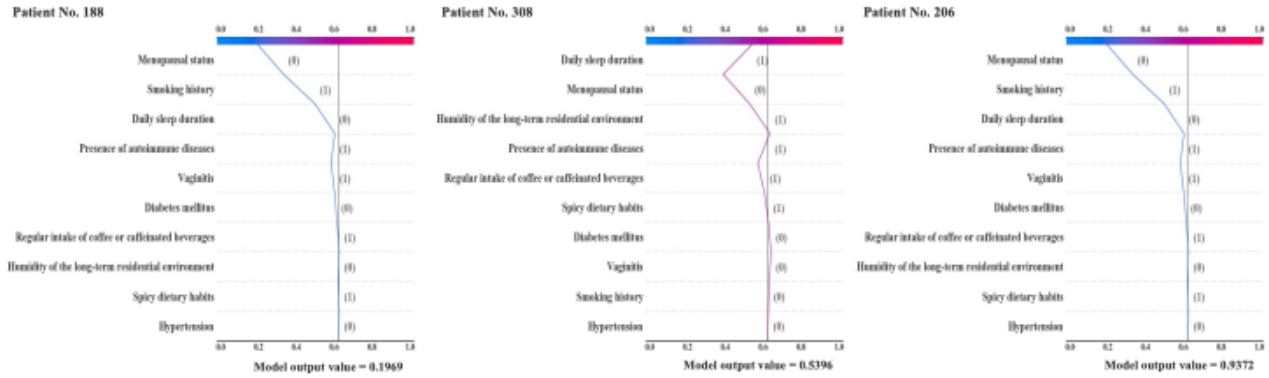

Discussion

Principal Findings

LVD is a common chronic gynecological condition, clinically characterized by vulvar hypopigmentation and pruritus. In middle and later stage, it may lead to dysuria, dyspareunia, introital stenosis, and vulvovaginal atrophy, causing considerable physical and psychological anguish to affected women [18,19]. Many patients experience delayed diagnosis and treatment due to the sensitive nature of the condition, which can cause embarrassment and feelings of shame, or due to misdiagnosis resulting from symptomatic variations. Early misdiagnosis of LVD, particularly VLS, can result in delayed treatment, leading to irreversible vulvar structural abnormalities, scarring, and, in severe cases, genital cancer [20]. Currently, there is no gold-standard treatment for LVD. Conventional first-line therapies are associated with adverse reactions and frequent disease recurrence. Emerging approaches, such as energy-based therapies and platelet-rich plasma injections, have yet to be fully integrated into clinical practice. Therefore, it is of paramount importance to identify factors potentially contributing to LVD in women and establish a predictive model to warn about its risk. To our knowledge, this study is the first systematic investigation of risk factors associated with LVD in women and the development of a predictive model rooted in the principles of evidence-based medicine. Univariate analysis revealed that vaginitis, urinary incontinence, humidity of the long-term residential environment, spicy dietary habits, regular intake of coffee or caffeinated beverages, daily sleep duration, diabetes mellitus, smoking history, presence of autoimmune diseases, menopausal status, and hypertension were all significant risk factors affecting female LVD. Furthermore, the GBM warning model outperforms the other 5 predictive models, demonstrating superior performance with higher ROC curve area, accuracy, sensitivity, and F_1-score. The GBM model reveals that menopausal status has the strongest positive correlation with female LVD, followed by the presence of autoimmune diseases, which also displayed a positive dependency.

Comparison With Previous Work

The GBM algorithm, a type of boosting algorithm, is capable of exploring the relationships between observed outcomes and various target variables, making it useful for clinical analysis of high-risk factors contributing to inducing the disease [21,22]. In this study, the GBM model was used to construct SHAP plots for analyzing the correlations between multiple factors and the target variables. The results revealed that menopausal status exerted the strongest positive correlation with the development of LVD in women, exhibiting a positive dependency. This finding aligns with the clinical observation that perimenopausal women have a higher incidence of LVD [23]. Günthert et al [24] have reported that the disease may be attributed to endocrine factors, the dearth of sex hormones, as the sex hormone levels in serum of affected individuals are lower than those of normal women of the same age, suggesting that sex hormone may be one of the contributing factors to the pathogenesis of LVD. Furthermore, the presence of autoimmune systemic diseases, daily sleep duration, diabetes, and humidity of the long-term residential environment are all positively correlated with LVD. The skin is a peripheral lymphoid organ with immunological functions that is closely connected to the systemic immune system. A substantial body of evidence has established a close association between LVD and autoimmune disorders. This relationship is particularly evident in cases of vulvar lichen sclerosus, a lymphocyte-mediated chronic localized skin disease, where both cellular and humoral immunity are involved in the pathogenesis. Therefore, patients

XSL•FO
RenderX

with LVD often present with various autoimmune diseases, such as vitiligo and hypothyroidism, among others [11,25]. Furthermore, some researches have indicated that diabetes is a systemic metabolic disorder often associated with various skin lesions [26,27]. In particular, individuals with type 2 diabetes are more likely to develop a range of skin disorders. In addition, diabetes tends to exacerbate the pruritus severity in patients with LVD. The intensity of itching is positively correlated with blood glucose levels, indicating that higher blood glucose values result in more pronounced itching symptoms. Similarly, excessive daily sleep duration has been linked to an increased risk of developing type 2 diabetes, thereby further raising the risk of LVD [28,29]. On the other hand, vaginitis showed a negative association with LVD. This could be attributed to the fact that women with vaginitis typically receive prompt and effective diagnosis and treatment, which reduces the risk of overusing medications, for some cases of LVD are due to the overuse of drugs mainly containing aromatics.

Strengths and Limitations

To sum up, it is crucial for women to cultivate healthy lifestyles and dietary habits, as well as to be informed about the factors related to LVD in order to reduce the incidence of this condition. Furthermore, the study demonstrates that the warning model exhibits strong predictive capabilities and discriminatory effect. The characteristics and strengths of the risk warning model primarily manifest in the following aspects: (1) it is convenient and easy to implement in clinical practice and (2) it can promptly identify the high-risk individuals and provide preventive interventions. To enhance the model's clinical use, we devised a web-based calculator and mobile app featuring easily accessible functionalities for patients. As a self-diagnostic tool, it offers early detection and timely intervention for LVD.

For female patients identified as low risk for predicted outcomes, it is advisable to undergo appropriate treatment for autoimmune conditions such as thyroid disorders, systemic lupus erythematosus, psoriasis, rheumatism, or arthritis. In addition, managing blood sugar levels, incorporating moderate physical activity, and ensuring adequate sleep duration are recommended. Furthermore, avoiding tight fitting garments, using mild soaps during bathing (avoiding its application on the vulva and instead cleansing the area with water only), gently patting the vulva dry post bathing, and using 100% cotton, unscented or fragrance-free menstrual pads are suggested. Emphasizing guidance on vulvovaginal medications, particularly avoiding products containing fragrances and preservatives, alongside health education, is paramount. Elevating awareness and self-management skills regarding LVD, especially recognizing the risk of neoplastic change associated with the absence of standardized treatment and follow-up, is crucial [7]. Low-risk individuals should maintain annual follow-up appointments.

For individuals categorized as medium-risk, extending lifestyle interventions akin to those recommended for low-risk individuals is advisable. Furthermore, incorporating general management strategies for perimenopause and hormone replacement therapy. In addition, applying preservative-free, unscented or fragrance-free emollient to hold moisture in the skin and improve barrier function is recommended [10]. Medium-risk individuals should maintain biannual follow-up appointments.

For high-risk patients, a higher frequency of follow-ups is recommended, ideally every 3 months. Depending on the situation, timely referral to a higher-level hospital should be considered to implement more precise and personalized treatment and monitoring. In cases where vulvar symptoms are significant, such as persistent itching and lack of self-relief, prompt medical attention is advised. Routine secretion tests should be conducted to rule out herpes simplex and candidal infections, with additional tests for autoimmune diseases, thyroid disease, and diabetes performed as necessary. Dermatoscopy can be used when feasible. After excluding conditions such as vaginitis, vulvar pathological examination may be warranted. In addition, it is advised to avoid any genital skin irritants and allergens, such as cleansing products, frequent water exposure, incontinence, and tight-fitting clothing. Some activities, such as cycling or horse riding, may exacerbate symptoms. Regular application of a barrier emollient to affected areas can protect against local irritants such as urine and menstrual blood [20]. During intercourse, it is recommended to use adequate lubrication, with unscented or fragrance-free silicone-based lubricants. Patients should be advised to closely monitor their vulvar condition and contact their doctor if they notice any changes in appearance or texture, such as lumps, ulceration, or skin hardening, or if there are significant changes in symptoms [10].

Limitations

This study has a few limitations. It is a single-center cohort study based on data from a single ethnic group, which may limit the generalizability of the findings. In addition, the relatively small sample size could affect the accuracy of the model's predictions. Future research should aim to include larger sample sizes and to optimize and validate the LVD warning model through multi-center studies involving diverse ethnic and racial groups.

Conclusions

In conclusion, developing an early warning model for female LVD on the basis of evidence-based medicine, promptly identifying high-risk factors for the condition, and implementing risk-stratified population management are crucial measures to enhance the effectiveness of preventing and treating female LVD. By using the warning model to classify individuals into high-risk and low-risk groups, it is conductive to explore the tailored health intervention strategies, and afford personalized health management recommendations, encompassing adjustments to individual lifestyle habits and the implementation of improved conservative treatment protocols. This contributes to an effect evaluation of grading management mode, facilitating the realization of personalized and long-term pelvic health control patterns. Early detection and prompt treatment of LVD can thus be achieved, thereby ameliorating the quality of life and overall well-being for women.

XSL•FO
RenderX

Acknowledgments

This study was supported by the foundation of Sichuan University (21H0103), Science and Technology Department of Sichuan Province (2019YFS0008), the National Key Research and Development Program of China (2021YFC2009100), Sichuan Provincial Health Commission Project (21ZD002), Research Projects of Sichuan Science and Technology Department (2023YFQ0070), Key Research Projects of Sichuan Science and Technology Department (2023YFG0128, 2023YFS0024, 2023NSFSC1606, and 23ZDYF2049). The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript. A patent application for the warning model developed in this study is planned, with ownership of the patent rights assigned to the authors.

Data Availability

Data are available upon request by the corresponding author.

Authors' Contributions

JM wrote this manuscript and collected the clinical data. CL was responsible for the data collection, project management, and manuscript editing. YC contributed to the project development and manuscript editing. QL contributed to the project development and manuscript editing. XN participated in the design, project development, and manuscript revision. DW was accountable for the conceptualization and design of this study, project development, and manuscript revision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the included machine learning algorithms and the rationale for their selection.
[DOCX File , 12 KB - jmir_v26i1e55734_app1.docx]

References

1. Fruchter R, Melnick L, Pomeranz MK. Lichenoid vulvar disease: a review. Int J Womens Dermatol 2017;3(1):58-64 [FREE Full text] [doi: 10.1016/j.ijwd.2017.02.017] [Medline: 28492056]
2. Lee A, Fischer G. Diagnosis and treatment of vulvar lichen sclerosus: an update for dermatologists. Am J Clin Dermatol 2018;19(5):695-706. [doi: 10.1007/s40257-018-0364-7] [Medline: 29987650]
3. Corazza M, Schettini N, Zedde P, Borghi A. Vulvar lichen sclerosus from pathophysiology to therapeutic approaches: evidence and prospects. Biomedicines 2021;9(8):950 [FREE Full text] [doi: 10.3390/biomedicines9080950] [Medline: 34440154]
4. Ling M, Dongmei W, Yueyue C, Yueting Z, Tao C, Jian M, et al. Vaginal microbiota changes in the vulvar lichen simplex chronicus. Clin Exp Obstet Gynecol 2023;50(1):17 [FREE Full text] [doi: 10.31083/j.ceog5001017]
5. Leis M, Singh A, Li C, Ahluwalia R, Fleming P, Lynde CW. Risk of vulvar squamous cell carcinoma in lichen sclerosus and lichen planus: a systematic review. J Obstet Gynaecol Can 2022;44(2):182-192. [doi: 10.1016/j.jogc.2021.09.023] [Medline: 34678521]
6. Steben M. Lichen sclerosus: why do most women struggle with their diagnosis? J Obstet Gynaecol Can 2022;44(2):119-120.e1. [doi: 10.1016/j.jogc.2021.12.001] [Medline: 35181006]
7. American College of Obstetricians Gynecologists' Committee on Practice Bulletins—Gynecology. Diagnosis and management of vulvar skin disorders: ACOG practice bulletin, number 224. Obstet Gynecol 2020;136(1):e1-e14. [doi: 10.1097/AOG.0000000000003944] [Medline: 32590724]
8. Sideri M, Parazzini F, Rognoni MT, La Vecchia C, Negri E, Garsia S, et al. Risk factors for vulvar lichen sclerosus. Am J Obstet Gynecol 1989;161(1):38-42. [doi: 10.1016/0002-9378(89)90228-7] [Medline: 2750821]
9. Smith YR, Haefner HK. Vulvar lichen sclerosus : pathophysiology and treatment. Am J Clin Dermatol 2004;5(2):105-125. [doi: 10.2165/00128071-200405020-00005] [Medline: 15109275]
10. van der Meijden WI, Boffa MJ, Ter Harmsel B, Kirtschig G, Lewis F, Moyal-Barracco M, et al. 2021 European guideline for the management of vulval conditions. J Eur Acad Dermatol Venereol 2022;36(7):952-972. [doi: 10.1111/jdv.18102] [Medline: 35411963]
11. Tran DA, Tan X, Macri CJ, Goldstein AT, Fu SW. Lichen sclerosus: an autoimmunopathogenic and genomic enigma with emerging genetic and immune targets. Int J Biol Sci 2019;15(7):1429-1439 [FREE Full text] [doi: 10.7150/ijbs.34613] [Medline: 31337973]
12. Virgili A, Borghi A, Cazzaniga S, Di Landro A, Naldi L, Minghetti S, GLS Italian Study Group. New insights into potential risk factors and associations in genital lichen sclerosus: Data from a multicentre Italian study on 729 consecutive cases. J Eur Acad Dermatol Venereol 2017;31(4):699-704. [doi: 10.1111/jdv.13867] [Medline: 27515901]

XSL•FO
RenderX

13. Fergus KB, Lee AW, Baradaran N, Cohen AJ, Stohr BA, Erickson BA, et al. Pathophysiology, clinical manifestations, and treatment of lichen sclerosus: a systematic review. Urology 2020;135:11-19. [doi: 10.1016/j.urology.2019.09.034] [Medline: 31605681]
14. Gulin SJ, Lundin F, Seifert O. Comorbidity in patients with Lichen sclerosus: a retrospective cohort study. Eur J Med Res 2023;28(1):338 [FREE Full text] [doi: 10.1186/s40001-023-01335-9] [Medline: 37697418]
15. National Center for Health Statistics. National Health Interview Survey.: Centers for Disease Control and Prevention URL: https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm [accessed 2024-09-23]
16. Damm MMB, Jensen TSR, Mahmood B, Lundh M, Poulsen SS, Bindslev N, et al. Acetylcholine-related proteins in non-neoplastic appearing colonic mucosa from patients with colorectal neoplasia. Mol Carcinog 2017;56(10):2223-2233. [doi: 10.1002/mc.22675] [Medline: 28544328]
17. Lee C, Jo B, Woo H, Im Y, Park RW, Park C. Chronic disease prediction using the common data model: development study. JMIR AI 2022;1(1):e41030 [FREE Full text] [doi: 10.2196/41030] [Medline: 38875545]
18. McCluggage WG. Recent developments in vulvovaginal pathology. Histopathology 2009;54(2):156-173. [doi: 10.1111/j.1365-2559.2008.03098.x] [Medline: 18637148]
19. Baggish MS. Fractional CO2 laser treatment for vaginal atrophy and vulvar lichen sclerosus. Journal of Gynecologic Surgery 2016;32(6):309-317 [FREE Full text] [doi: 10.1089/gyn.2016.0099]
20. Kirtschig G, Kinberger M, Kreuter A, Simpson R, Günthert A, van Hees C, et al. EuroGuiderm guideline on lichen sclerosus-treatment of lichen sclerosus. J Eur Acad Dermatol Venereol 2024. [doi: 10.1111/jdv.20083] [Medline: 38822598]
21. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, et al. JOURNAL CLUB: use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. AJR Am J Roentgenol 2019;212(1):44-51. [doi: 10.2214/AJR.18.20260] [Medline: 30354266]
22. Zhang X, Xue Y, Su X, Chen S, Liu K, Chen W, et al. A transfer learning approach to correct the temporal performance drift of clinical prediction models: retrospective cohort study. JMIR Med Inform 2022;10(11):e38053 [FREE Full text] [doi: 10.2196/38053] [Medline: 36350705]
23. Nair PA. Vulvar lichen sclerosus et atrophicus. J Midlife Health 2017;8(2):55-62 [FREE Full text] [doi: 10.4103/jmh.JMH_13_17] [Medline: 28706405]
24. Günthert AR, Faber M, Knappe G, Hellriegel S, Emons G. Early onset vulvar lichen sclerosus in premenopausal women and oral contraceptives. Eur J Obstet Gynecol Reprod Biol 2008;137(1):56-60. [doi: 10.1016/j.ejogrb.2007.10.005] [Medline: 18055095]
25. Sherman V, McPherson T, Baldo M, Salim A, Gao XH, Wojnarowska F. The high rate of familial lichen sclerosus suggests a genetic contribution: an observational cohort study. J Eur Acad Dermatol Venereol 2010;24(9):1031-1034. [doi: 10.1111/j.1468-3083.2010.03572.x] [Medline: 20202060]
26. Makrantonaki E, Jiang D, Hossini AM, Nikolakis G, Wlaschek M, Scharffetter-Kochanek K, et al. Diabetes mellitus and the skin. Rev Endocr Metab Disord 2016;17(3):269-282. [doi: 10.1007/s11154-016-9373-0] [Medline: 27432328]
27. de Macedo GMC, Nunes S, Barreto T. Skin disorders in diabetes mellitus: an epidemiology and physiopathology review. Diabetol Metab Syndr 2016;8(1):63 [FREE Full text] [doi: 10.1186/s13098-016-0176-y] [Medline: 27583022]
28. Zizi F, Jean-Louis G, Brown CD, Ogedegbe G, Boutin-Foster C, McFarlane SI. Sleep duration and the risk of diabetes mellitus: epidemiologic evidence and pathophysiologic insights. Curr Diab Rep 2010;10(1):43-47 [FREE Full text] [doi: 10.1007/s11892-009-0082-x] [Medline: 20425066]
29. Briançon-Marjollet A, Weiszenstein M, Henri M, Thomas A, Godin-Ribuot D, Polak J. The impact of sleep disorders on glucose metabolism: endocrine and molecular mechanisms. Diabetol Metab Syndr 2015;7:25 [FREE Full text] [doi: 10.1186/s13098-015-0018-3] [Medline: 25834642]

Abbreviations

ADA: adaptive boosting
AUC: area under the curve
CatBoost: categorical boosting
GBM: gradient boosting machine
LR: logistic regression
LVD: lichenoid vulvar disease
RF: random forest
ROC: receiver operating characteristic
SHAP: SHapley Additive exPlanations
VLS: vulvar lichen sclerosus
XGB: eXtreme gradient boosting

XSL•FO
RenderX

Edited by Y Li; submitted 22.12.23; peer-reviewed by B Chang, H Lu, H Zhang, B Chen; comments to author 26.05.24; revised version received 18.06.24; accepted 01.09.24; published 22.11.24.

Please cite as:
Meng J, Niu X, Luo C, Chen Y, Li Q, Wei D
Development and Validation of a Machine Learning–Based Early Warning Model for Lichenoid Vulvar Disease: Prediction Model Development Study
J Med Internet Res 2024;26:e55734
URL: https://www.jmir.org/2024/1/e55734
doi:10.2196/55734
PMID:

©Jian Meng, Xiaoyu Niu, Can Luo, Yueyue Chen, Qiao Li, Dongmei Wei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identification of a Susceptible and High-Risk Population for Postoperative Systemic Inflammatory Response Syndrome in Older Adults: Machine Learning–Based Predictive Model

Haiyan Mai[1*], BS; Yaxin Lu[2*], MS; Yu Fu[1*], MS; Tongsen Luo[2], MS; Xiaoyue Li[2], MS; Yihan Zhang[2], PhD; Zifeng Liu[2], PhD; Yuenong Zhang[3], MS; Shaoli Zhou[2], PhD; Chaojin Chen[2,4], PhD

[1]Department of Pharmacy, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[2]Department of Anesthesiology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[3]Department of Surgery and Anesthesia, The Third Affiliated Hospital of Sun Yat-sen University Yuedong Hospital, Meizhou, China

[4]Big Data and Artificial Intelligence Center, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Chaojin Chen, PhD
Department of Anesthesiology
The Third Affiliated Hospital of Sun Yat-sen University
600 Tianhe Road
Tianhe District
Guangzhou, 510631
China
Phone: 86 020 85253333
Email: chenchj28@mail.sysu.edu.cn

Abstract

Background: Systemic inflammatory response syndrome (SIRS) is a serious postoperative complication among older adult surgical patients that frequently develops into sepsis or even death. Notably, the incidences of SIRS and sepsis steadily increase with age. It is important to identify the risk of postoperative SIRS for older adult patients at a sufficiently early stage, which would allow preemptive individualized enhanced therapy to be conducted to improve the prognosis of older adult patients. In recent years, machine learning (ML) models have been deployed by researchers for many tasks, including disease prediction and risk stratification, exhibiting good application potential.

Objective: We aimed to develop and validate an individualized predictive model to identify susceptible and high-risk populations for SIRS in older adult patients to instruct appropriate early interventions.

Methods: Data for surgical patients aged ≥65 years from September 2015 to September 2020 in 3 independent medical centers were retrieved and analyzed. The eligible patient cohort in the Third Affiliated Hospital of Sun Yat-sen University was randomly separated into an 80% training set (2882 patients) and a 20% internal validation set (720 patients). We developed 4 ML models to predict postoperative SIRS. The area under the receiver operating curve (AUC), F_1 score, Brier score, and calibration curve were used to evaluate the model performance. The model with the best performance was further validated in the other 2 independent data sets involving 844 and 307 cases, respectively.

Results: The incidences of SIRS in the 3 medical centers were 24.3% (876/3602), 29.6% (250/844), and 6.5% (20/307), respectively. We identified 15 variables that were significantly associated with postoperative SIRS and used in 4 ML models to predict postoperative SIRS. A balanced cutoff between sensitivity and specificity was chosen to ensure as high a true positive as possible. The random forest classifier (RF) model showed the best overall performance to predict postoperative SIRS, with an AUC of 0.751 (95% CI 0.709-0.793), sensitivity of 0.682, specificity of 0.681, and F_1 score of 0.508 in the internal validation set and higher AUCs in the external validation-1 set (0.759, 95% CI 0.723-0.795) and external validation-2 set (0.804, 95% CI 0.746-0.863).

Conclusions: We developed and validated a generalizable RF model to predict postoperative SIRS in older adult patients, enabling clinicians to screen susceptible and high-risk patients and implement early individualized interventions. An online risk calculator to make the RF model accessible to anesthesiologists and peers around the world was developed.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57486) doi:10.2196/57486

KEYWORDS

older adult patients; postoperative SIRS; sepsis; machine learning; prediction model

Introduction

Systemic inflammatory response syndrome (SIRS) is a nonhomeostatic, self-destructive, and uncontrollable inflammatory response of the whole body triggered by infection, trauma, or major operations [1]. Recognizing SIRS has been a prerequisite of suspecting potential sepsis and implementing decisions such as sample culturing for the source of infection, escalating antibiotic regimens, and the level of patient monitor and care [2,3]. Although there is a tendency to apply criteria including the Sequential Organ Failure Assessment (SOFA) score or quick SOFA score to identify the possibility of sepsis [4], SIRS criteria has demonstrated higher sensitivity than the quick SOFA score [3], and it has served as both useful inclusion criteria and therapeutic target of trials aiming to treat sepsis [5]. Early identification of patients who will develop postoperative SIRS may enable clinicians to provide timely interventions to prevent sepsis and improve outcomes. It has been reported that the incidence of postoperative SIRS could be as high as 89% [6,7] in patients undergoing abdominal surgery, and the condition frequently developed into sepsis and even multiple organ dysfunction syndrome [8]. A 13-fold increase in mortality was reported in patients with postoperative SIRS compared with those without SIRS [9]. Notably, the incidences of SIRS and sepsis steadily increase with age, and octogenarians are almost twice as likely to develop sepsis than those aged less than 50 years [10,11]. Although standardized preoperative antibiotic prophylaxis has been recommended and clinically applied for older adult populations with a high risk of postoperative infection and SIRS, nearly one-quarter of older adult patients still develop SIRS within 3 days after surgery [12]. Thus, it is important to identify the risk of postoperative SIRS for older adult patients at a sufficiently early stage, which would allow preemptive individualized enhanced therapy to be conducted to improve the prognosis of older adult patients.

Compared with traditional biostatistical methods, machine learning (ML) methods hold the advantages of flexibility, scalability, and the ability to analyze diverse data types, which can be deployed for many tasks, such as risk stratification, diagnosis and classification, and survival predictions [13-15]. In recent years, there have been many ML models used to predict sepsis [16-22], the majority of which were developed in similar populations such as patients in the intensive care unit [23]. However, only a few have focused on older adult surgical patients, and an ML model to predict postoperative SIRS has been rarely reported.

The goal of our study was to use ML methods to develop an individualized predictive model for older adult surgical patients to screen for a susceptible population at a high risk for SIRS in order to instruct appropriate early interventions. Meanwhile, the generalizability of the model was validated with the data sets from the other 2 medical centers.

Methods

Setting, Dates, and Population

The study was performed based on the electronic health record (EHR) systems of 3 independent medical centers including the Third Affiliated Hospital of Sun Yat-sen University (Guangzhou, China), Lingnan Hospital of Sun Yat-sen University, and Yuedong Hospital of Sun Yat-sen University using data from surgical patients aged ≥65 years from 2015 to 2020.

During the retrospective enrollment, the inclusion criteria included (1) age ≥65 years, (2) patients who underwent general anesthetic with endotracheal intubation or laryngeal mask, and (3) patients who had preoperative antimicrobial prophylaxis. The exclusion criteria included (1) patients whose total intraoperative infusion volumes, fluid loss, or American Society of Anesthesiologists (ASA) classifications were not recorded and (2) patients diagnosed with SIRS before surgery or undergoing infectious surgeries.

Data Sources

The data comes from the EHR system of the 3 hospitals, which were established by extracting medical records from the hospital information system, laboratory information system, picture archiving and communication system (PACS), and Docare Anesthesia System (2005-2020 Medical System Co, Ltd), which enabled access to a comprehensive data set collected during hospital admissions, inpatient stays, and post-hospital follow-up visits, including demographic characteristics, daily documentation, laboratory tests, imaging results, and anesthesia records.

Outcome

SIRS was diagnosed according to the American College of Chest Physicians [24]. It was defined when 2 or more of the following criteria were present: (1) temperature <36 °C or ≥38 °C, (2) heart rate ≥90 bpm, (3) respiratory rate ≥20 bpm or arterial carbon dioxide tension <32 mm Hg, and (4) white blood cell count $<4\times10^9$/L, $\geq12\times10^9$/L, or >10% immature forms. The incidence of SIRS within 3 postoperative days was recorded in the study.

Data Sets

Data sets from the 3 medical centers were created separately and included: (1) patient demographics such as age, gender, and smoking history; (2) preoperative complications, including diabetes, hypertension, and fever; (3) anesthesia records, including ASA, administration of ulinastatin, dexamethasone, methylprednisolone, and dexmedetomidine; (4) liquid management including volume of fluid loss, blood loss, and colloid input; (5) duration of surgery; (6) laboratory parameters, including alanine aminotransferase (ALT), white blood cell count, hemoglobin, creatinine, albumin, high-sensitivity

XSL•FO
RenderX

C-reactive protein (hs-CRP), activated partial thromboplastin time, glucose, low-density lipoprotein, high-density lipoprotein, blood urea nitrogen (BUN), fibrinogen, thrombin time, lymphocyte, red blood cell count, indirect bilirubin. These variables were selected based on EHR availability and their relevance to SIRS risk according to the literature and clinical experience [25].

For the purpose of developing and validating the ML models for risk prediction, the cohort of eligible patients from the Third Affiliated Hospital of Sun Yat-sen University was randomly divided into an 80% training set (2882 patients) and a 20% internal validation set (720 patients). In addition, 2 validation data sets were created with the eligible patients from the Lingnan Hospital and Yuedong Hospital for external validation. Based on the events per variable (EPV)-10 principle [26], this study has a sufficient sample size.

Statistical Analysis and ML Model Training

This study was reported as per the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [27]. Categorical variables are expressed as number (%), and continuous variables that follow a normal distribution are expressed as mean (SD) or otherwise as median (IQR). The missing rates of the variables are shown in Table S1 in Multimedia Appendix 1. Missing values were imputed on each of the 3 cohorts (hospitals) independently: Categorical variables were filled in using mode counts, and continuous variables were filled in using means. The outliers were eliminated based on clinical experience, and 0 variance variables were removed.

We trained 4 different classical ML algorithms trained on the training set, including random forest (RF), XGBoost, logistic regression (LR), and multilayer perceptron (MLP) models. RF represents a bagging parallel integration algorithm, XGBoost represents a boosting algorithm; LR represents the most classical linear model, and MLP represents a neural network algorithm. Grid search was applied based on 5-fold cross-validation to find the optimal hyperparameters for each algorithm based on the training set. The formulas for the evaluation indicators in the models are provided in Multimedia Appendix 1.

The performance of the 4 algorithms was compared, and the optimal model was selected for this study. A comparison between the optimal model and nomogram we established before [28] and was also conducted with the internal validation set. Note that the prepared data sets incurred a data imbalance problem. To address this problem, we used the synthetic minority oversampling technique (SMOTE) provided by the performanceEstimation package [29,30].

The statistical analyses were all done using Python 3.7 [31-34] and R-3.6.2 [35]. All results were considered statistically significant at $P<.05$.

Ethical Considerations

The study protocol followed the principles of the Declaration of Helsinki and was approved by the Institutional Ethics Committee of the Third Affiliated Hospital of Sun Yat-sen University on July 27, 2022 (number [2019]02-609-04). The requirements for informed consent and clinical trial registration were waived by the committee.

Results

Study Cohorts and Characteristics

Among the 16,141 patients aged ≥65 years accessed from the EHR system, only 3602 patients who met the inclusion criteria were included in the development cohort, with 876 (24.3%) postoperative SIRS events. The development cohort was then randomly separated into the training set and the internal validation set, which consisted of 2882 and 720 patients, respectively. Meanwhile, 844 and 307 patients were finally included in the external validation-1 set (Lingnan Hospital) and the external validation-2 set (Yuedong Hospital), respectively (Figure 1).

XSL•FO
RenderX

Figure 1. Study design and flowchart. ASA: American Society of Anesthesiologists; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; MLP: multilayer perceptron; RF: random forest; SIRS: systemic inflammatory response syndrome.

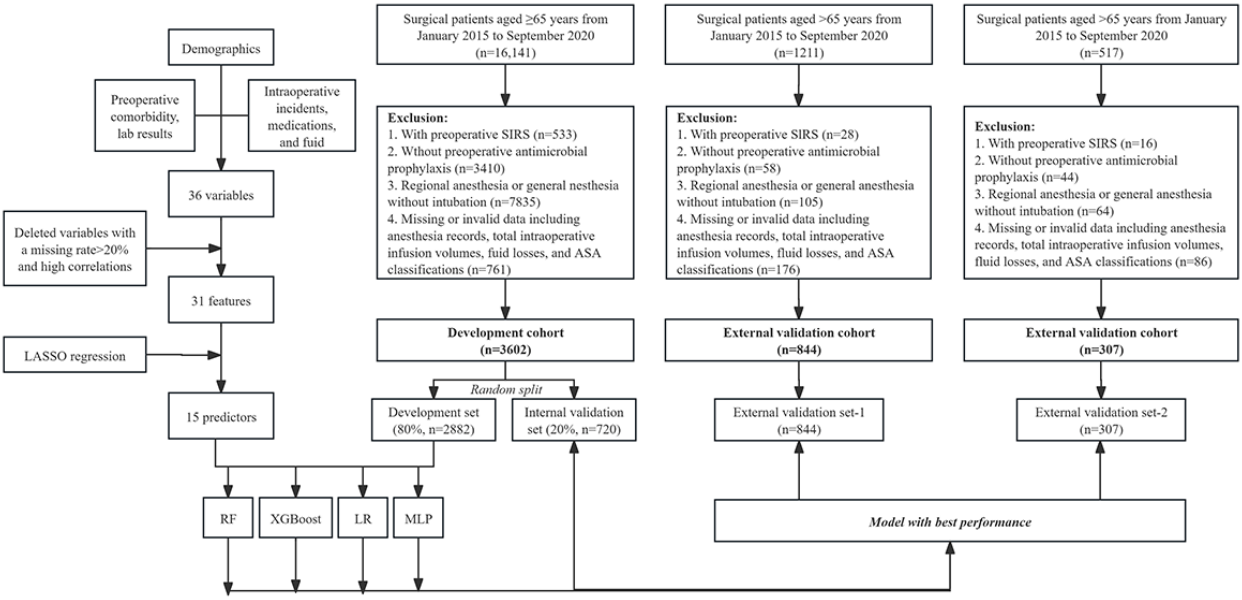

The characteristics of the 3 cohorts from different medical centers are shown in Table 1. The incidence rates of postoperative SIRS in each of the 3 medical centers were 24.3% (876/3602), 29.6% (250/844), and 6.5% (20/307), respectively. The different incidences of SIRS across the 3 medical centers may result from factors such as different regional characteristics, economic development levels, patient characteristics, surgical procedures, and the medical and nursing team.

Among the 3602 patients in the development cohort with a mean age of 70.0 years, 2144 (59.5%) were women, 2152 (59.7%) had comorbid hypertension, 1140 (31.6%) had comorbid diabetes, and 437 (12.1%) were smokers. In addition, 420 (11.7%) patients had preoperative fever, and 1330 (36.9%) were assessed as ASA III/IV/V preoperatively.

The characteristics of patients with or without postoperative SIRS in the development cohort are shown in Table 2. Patients who ended up with SIRS (SIRS group vs the non-SIRS group) were mainly women (573/876, 65.4% vs 1571/2726, 57.6%, $P<.001$) and more likely to have been diagnosed with diabetes (334/876, 38.1% vs 806/2726, 29.6%, $P<.001$) and assessed as ASA III/IV/V preoperatively (472/876, 53.9% vs 858/2726, 31.5%, $P<.001$). With evident incidence of preoperative fever (182/876, 20.8% vs 238/2726, 8.7%, $P<.001$), the total volume of fluid loss, volume of blood loss, and duration of surgery were also significantly higher and longer in the SIRS group than in the non-SIRS group (700 [330-1200] mL vs 430 [200-800] mL, 100 [50.0-200] mL vs 50 [20.0-100] mL, 207 [133-310] min vs 150 [90-232] min, respectively; all $P<.001$). Additionally, laboratory indicators including hs-CRP, BUN, direct bilirubin (DBILI), and prothrombin time (PT) were higher in SIRS patients than in non-SIRS patients, whereas albumin was lower in SIRS patients (all $P<.001$; Table 2).

XSL•FO
RenderX

Table 1. Baseline characteristics of patients.

Characteristics	Development cohort (n=3602)	External validation-1 (n=844)	External validation-2 (n=307)
Postoperative SIRS [a], n (%)			
No	2726 (75.7)	594 (70.4)	287 (93.5)
Yes	876 (24.3)	250 (29.6)	20 (6.5)
Demographics			
Age, (years), median (IQR)	70.0 (67.0-75.0)	70.0 (67.0-76.0)	72.0 (67.2-78.8)
Gender, n (%)			
Female	2144 (59.5)	468 (55.5)	169 (55)
Male	1458 (40.5)	376 (44.5)	138 (45)
Hypertension, n (%)			
No	1450 (40.3)	313 (37.1)	173 (56.4)
Yes	2152 (59.7)	531 (62.9)	134 (43.6)
Diabetes, n (%)			
No	2462 (68.4)	526 (62.3)	51 (54.8)
Yes	1140 (31.6)	318 (37.7)	42 (45.2)
History of smoking, n (%)			
No	3165 (87.9)	755 (89.5)	257 (86.2)
Yes	437 (12.1)	89 (10.5)	41 (13.8)
Preoperative fever, n (%)			
No	3182 (88.3)	642 (76.1)	290 (94.5)
Yes	420 (11.7)	202 (23.9)	17 (5.54)
ASA [b] classification, n (%)			
I/II	2272 (63.1)	509 (60.3)	236 (76.9)
III/IV/V	1330 (36.9)	335 (39.7)	71 (23.1)
Preoperative variables			
WBC[c] (10^9/L), median (IQR)	6.41 (5.15-8.11)	6.57 (5.38-8.66)	6.69 (5.57-8.53)
Lymphocytes (10^9/L), median (IQR)	1.60 (1.20-2.05)	1.50 (1.09-1.96)	1.73 (1.32-2.19)
RBC[d] (10^{12}/L), median (IQR)	4.27 (3.85-4.66)	4.15 (3.74-4.62)	4.29 (3.91-4.72)
HGB[e] (g/L), median (IQR)	127 (113-138)	124 (111-137)	125 (111-136)
RDW-CV[f] (%), median (IQR)	0.13 (0.13-0.14)	0.13 (0.12-0.14)	13.5 (12.9-14.4)
hs-CRP[g] (mg/L), median (IQR)	6.41 (5.10-8.20)	6.56 (5.30-8.66)	10.0 (4.17-34.9)
Albumin (g/L), median (IQR)	39.6 (36.2-42.7)	39.0 (36.0-41.7)	37.8 (34.3-40.8)
ALT[h] (U/L), median (IQR)	17.0 (13.0-26.0)	18.0 (13.0-28.0)	19.0 (13.0-28.0)
TBILI[i] (umol/L), median (IQR)	9.80 (7.00-13.9)	10.3 (7.30-14.7)	14.6 (11.2-19.5)
DBILI[j] (umol/L), median (IQR)	3.10 (2.10-4.90)	3.51 (2.30-5.55)	3.40 (2.60-4.85)
IBILI[k] (umol/L), median (IQR)	6.50 (4.50-9.10)	6.50 (4.60-9.20)	10.8 (8.15-14.4)
Glucose (mmol/L), median (IQR)	5.42 (4.87-6.48)	5.52 (4.83-6.69)	5.43 (4.92-6.48)
Creatinine (umol/L), median (IQR)	75.0 (61.0-90.0)	79.0 (64.0-96.0)	74.0 (62.4-89.4)
BUN[l] (mmol/L), median (IQR)	5.67 (4.60-7.02)	5.83 (4.70-7.58)	5.66 (4.50-6.82)
LDL[m] (mmol/L), median (IQR)	2.90 (2.24-3.59)	2.87 (2.24-3.53)	2.55 (2.06-2.99)

XSL•FO
RenderX

Characteristics	Development cohort (n=3602)	External validation-1 (n=844)	External validation-2 (n=307)
HDL[n] (mmol/L), median (IQR)	1.05 (0.86-1.26)	1.11 (0.92-1.33)	1.21 (0.98-1.56)
PT[o] (seconds), median (IQR)	13.2 (12.8-13.9)	13.1 (12.5-13.8)	11.1 (10.6-11.6)
APTT[p] (seconds), median (IQR)	37.4 (34.9-40.0)	36.3 (33.5-39.1)	31.1 (29.2-33.7)
Fibrinogen (g/L), median (IQR)	3.65 (3.08-4.52)	3.67 (3.08-4.58)	3.23 (2.83-3.92)
TT[q] (seconds), median (IQR)	17.5 (17.5-17.5)	17.0 (17.0-17.0)	15.7 (14.9-16.4)
PTINR[r], median (IQR)	1.00 (0.96-1.06)	0.99 (0.94-1.07)	1.01 (0.96-1.05)
Intraoperative variables			
Ulinastatin, n (%)			
No	2817 (78.2)	627 (74.3)	305 (99.3)
Yes	785 (21.8)	217 (25.7)	2 (0.65)
Dexamethasone, n (%)			
No	3222 (89.5)	762 (90.3)	303 (98.7)
Yes	380 (10.5)	82 (9.72)	4 (1.3)
Dexmedetomidine, n (%)			
No	2270 (63)	566 (67.1)	110 (35.8)
Yes	1332 (37)	278 (32.9)	197 (64.2)
Methylprednisolone (mg), median (IQR)	0.00 (0.00-0.00)	0.00 (0.00-40.0)	0.00 (0.00-40.0)
Total volume of fluid loss (mL), median (IQR)	500 (220-900)	505 (250-900)	420 (220-808)
Volume of blood loss (mL), median (IQR)	50.0 (20.0-150)	100 (30.0-200)	20.0 (10.0-80.0)
Intraoperative colloid (mL), median (IQR)	500 (500-500)	500 (500-500)	500 (500-500)
Duration of surgery (min), median (IQR)	163 (95.0-250)	170 (105-253)	140 (87.8-210)

[a]SIRS: systemic inflammatory response syndrome.
[b]ASA: American Society of Anesthesiologists.
[c]WBC: white blood cell count.
[d]RBC: red blood cell count.
[e]HGB: hemoglobin.
[f]RDW-CV: red blood cell distribution width-coefficient of variation.
[g]hs-CRP: high sensitivity C-reactive protein.
[h]ALT: alanine aminotransferase.
[i]TBILI: total bilirubin.
[j]DBILI: direct bilirubin.
[k]IBILI: indirect bilirubin.
[l]BUN: blood urea nitrogen.
[m]LDL: low-density lipoprotein.
[n]HDL: high-density lipoprotein.
[o]PT: prothrombin time.
[p]APTT: activated partial thromboplastin time.
[q]TT: thrombin time.
[r]PTINR: international normalized ratio of prothrombin time.

XSL•FO
RenderX

Table 2. Characteristics of non-systemic inflammatory response syndrome (SIRS) and SIRS groups in the development cohort.

Characteristics	Development cohort			
	Total (n=3602)	Non-SIRS (n=2726)	SIRS (n=876)	*P* value
Demographics				
Age (years), median (IQR)	70.0 (67.0-75.0)	70.0 (67.0-75.0)	71.0 (67.0-76.2)	<.001
Gender, n (%)				<.001
Female	2144 (59.5)	1571 (57.6)	573 (65.4)	
Male	1458 (40.5)	1155 (42.4)	303 (34.6)	
Hypertension, n (%)				.76
No	1450 (40.3)	1093 (40.1)	357 (40.8)	
Yes	2152 (59.7)	1633 (59.9)	519 (59.2)	
Diabetes, n (%)				<.001
No	2462 (68.4)	1920 (70.4)	542 (61.9)	
Yes	1140 (31.6)	806 (29.6)	334 (38.1)	
History of smoking, n (%)				.001
No	3165 (87.9)	2423 (88.9)	742 (84.7)	
Yes	437 (12.1)	303 (11.1)	134 (15.3)	
Preoperative fever, n (%)				<.001
No	3182 (88.3)	2488 (91.3)	694 (79.2)	
Yes	420 (11.7)	238 (8.7)	182 (20.8)	
ASA [a] classification, n (%)				<.001
I/II	2272 (63.1)	1868 (68.5)	404 (46.1)	
III/IV/V	1330 (36.9)	858 (31.5)	472 (53.9)	
Preoperative variables				
WBC[b] (10^9/L), median (IQR)	6.41 (5.15-8.11)	6.32 (5.07-7.88)	6.80 (5.44-8.94)	<.001
Lymphocytes (10^9/L), median (IQR)	1.60 (1.20-2.05)	1.63 (1.23-2.05)	1.54 (1.12-2.05)	.007
RBC[c] (10^{12}/L), median (IQR)	4.27 (3.85-4.66)	4.28 (3.88-4.66)	4.22 (3.73-4.65)	.006
HGB[d] (g/L), median (IQR)	127 (113-138)	127 (115-138)	126 (110-139)	.04
RDW-CV[e] (%), median (IQR)	0.13 (0.13-0.14)	0.13 (0.12-0.14)	0.13 (0.13-0.14)	<.001
hs-CRP[f] (mg/L), median (IQR)	6.41 (5.10-8.20)	6.31 (5.03-7.93)	6.83 (5.46-9.20)	<.001
Albumin (g/L), median (IQR)	39.6 (36.2-42.7)	40.0 (36.6-42.9)	38.2 (35.0-41.7)	<.001
ALT[g] (U/L), median (IQR)	17.0 (13.0-26.0)	17.0 (12.0-25.0)	19.0 (14.0-32.0)	<.001
TBILI[h] (umol/L), median (IQR)	9.80 (7.00-13.9)	9.70 (6.90-13.5)	10.2 (7.30-15.2)	<.001
DBILI[i] (umol/L), median (IQR)	3.10 (2.10-4.90)	3.00 (2.10-4.70)	3.50 (2.30-5.50)	<.001
IBILI[j] (umol/L), median (IQR)	6.50 (4.50-9.10)	6.40 (4.60-8.90)	6.60 (4.50-9.50)	.16
Glucose (mmol/L), median (IQR)	5.42 (4.87-6.48)	5.41 (4.87-6.41)	5.43 (4.88-6.75)	.31
Creatinine (umol/L), median (IQR)	75.0 (61.0-90.0)	73.0 (60.0-88.0)	78.0 (63.0-96.0)	<.001
BUN[k] (mmol/L), median (IQR)	5.67 (4.60-7.02)	5.59 (4.56-6.90)	5.96 (4.69-7.42)	<.001
LDL[l] (mmol/L), median (IQR)	2.90 (2.24-3.59)	2.92 (2.26-3.59)	2.81 (2.19-3.58)	.055
HDL[m] (mmol/L), median (IQR)	1.05 (0.86-1.26)	1.07 (0.88-1.28)	1.02 (0.81-1.24)	<.001
PT[n] (seconds), median (IQR)	13.2 (12.8-13.9)	13.2 (12.7-13.8)	13.4 (12.9-14.2)	<.001

XSL•FO
RenderX

Characteristics	Development cohort			
	Total (n=3602)	Non-SIRS (n=2726)	SIRS (n=876)	*P* value
APTT[o] (seconds), median (IQR)	37.4 (34.9-40.0)	37.2 (34.9-39.9)	38.0 (35.0-41.3)	.001
Fibrinogen (g/L), median (IQR)	3.65 (3.08-4.52)	3.63 (3.09-4.44)	3.77 (3.05-4.75)	.03
TT[p] (seconds); median (IQR)]	17.5 (17.5-17.5)	17.5 (17.5-17.5)	17.5 (17.5-17.5)	.055
PTINR[q], median (IQR)	1.00 (0.96-1.06)	1.00 (0.96-1.06)	1.02 (0.97-1.09)	<.001
Intraoperative variables				
Ulinastatin, n (%)				<.001
No	2817 (78.2)	2214 (81.2)	603 (68.8)	
Yes	785 (21.8)	512 (18.8)	273 (31.2)	
Dexamethasone, n (%)				.81
No	3222 (89.5)	2436 (89.4)	786 (89.7)	
Yes	380 (10.5)	290 (10.6)	90 (10.3)	
Dexmedetomidine, n (%)				.01
No	2270 (63)	1749 (64.2)	521 (59.5)	
Yes	1332 (37)	977 (35.8)	355 (40.5)	
Methylprednisolone (mg), median (IQR)	0	0	0	<.001
Total volume of fluid loss (mL), median (IQR)	500 (220-900)	430 (200-800)	700 (330-1200)	<.001
Volume of blood loss (mL), median (IQR)	50.0 (20.0-150)	50.0 (20.0-100)	100 (50.0-200)	<.001
Intraoperative colloid (mL), median (IQR)	500 (500-500)	500 (500-500)	500 (500-1000)	<.001
Duration of surgery (min), median (IQR)	163 (95.0-250)	150 (90.0-232)	207 (133-310)	<.001

[a]ASA: American Society of Anesthesiologists.
[b]WBC: white blood cell count.
[c]RBC: red blood cell count.
[d]HGB: hemoglobin.
[e]RDW-CV: red blood cell distribution width-coefficient of variation.
[f]hs-CRP: high sensitivity C-reactive protein.
[g]ALT: alanine aminotransferase.
[h]TBILI: total bilirubin.
[i]DBILI: direct bilirubin.
[j]IBILI: indirect bilirubin.
[k]BUN: blood urea nitrogen.
[l]LDL: low-density lipoprotein.
[m]HDL: high-density lipoprotein.
[n]PT: prothrombin time.
[o]APTT: activated partial thromboplastin time.
[p]TT: thrombin time.
[q]PTINR: international normalized ratio of prothrombin time.

The prognoses of the non-SIRS and SIRS groups in the development cohort are shown in Table 3. Compared with the non-SIRS group, patients in the SIRS group were significantly more likely to develop postoperative complications that included hemorrhage (446/876, 50.9% vs 939/2726, 34.4%, *P*<.001), acute respiratory distress syndrome (14/876, 1.6% vs 2/2726, 0.1%, *P*<.001), cardiac arrest (10/876, 1.1% vs 5/2726, 0.2%, *P*<.001), agitation and delirium (65/876, 7.4% vs 32/2726, 1.2%, *P*<.001), coma (60/876, 6.8% vs 7/2726, 0.3%, *P*<.001), and acute kidney injury (81/876, 9.2% vs 55/2726, 2%, *P*<.001). Furthermore, SIRS patients had a longer median postoperative hospitalization (12.0 [8.0-19.0] days vs 7.0 [5.0-10.0] days, *P*<.001), higher median cost (¥91,846 [¥61,765-¥135.5] vs ¥56,542 [¥34,273-¥76,956], *P*<.001; a currency exchange rate of ¥1=US $0.14 is applicable), higher risk of postoperative intensive care unit admissions (273/876, 37.1% vs 79/2726, 3.3%, *P*<.001), and higher in-hospital mortality rate (21/876, 2.4% vs 10/2726, 0.4%, *P*<.001) than non-SIRS patients.

XSL•FO
RenderX

Table 3. Prognoses of the non-systemic inflammatory response syndrome (SIRS) and SIRS groups in the development cohort.

Prognoses	Total cohort (n=3602)	Non-SIRS (n=2726)	SIRS (n=876)	*P* value
Postoperative complications, n (%)				
Hemorrhage	1385 (38.5)	939 (34.4)	446 (50.9)	<.001
ARDS[a]	16 (0.4)	2 (0.1)	14 (1.6)	<.001
Cardiac arrest	15 (0.4)	5 (0.2)	10 (1.1)	<.001
Agitation and delirium	97 (2.7)	32 (1.2)	65 (7.4)	<.001
Coma	67 (1.9)	7 (0.3)	60 (6.8)	<.001
Acute kidney injury	136 (3.8)	55 (2)	81 (9.2)	<.001
Postoperative ICU[b] admission, n (%)	352 (11.4)	79 (3.3)	273 (37.1)	<.001
In-hospital death, n (%)	31 (0.9)	10 (0.4)	21 (2.4)	<.001
Postoperative hospital stay (days), median (Q1-Q3)	8.0 (5.0-12.0)	7.0 (5.0-10.0)	12.0 (8.0-19.0)	<.001
Total hospital stay (days), median (Q1-Q3)	16.0 (11.0-22.0)	15.0 (10.0-20.0)	21.5 (15.0-32.0)	<.001
Total cost (¥)[c], median (Q1-Q3)	62,434 (39,886-87,937)	56,542 (34,273-76,956)	91,846 (61,675-135,548)	<.001

[a]ARDS: acute respiratory distress syndrome.

[b]ICU: intensive care unit.

[c]A currency exchange rate of ¥1=US $0.14 is applicable.

Variable Selection

Details of the missing rates in each center for the 36 variables selected based on the literature and our clinical experience are shown in Table S1 in Multimedia Appendix 1, and the highly correlated variables with a correlation coefficient higher than 0.7 were deleted to avoid collinearity (based on the principle of keeping the variable with the highest area under the receiver-operating curve [AUC] in the independent prediction of postoperative SIRS; Figure S1 in Multimedia Appendix 1).

Since partially relevant or less important features may negatively affect the performance of ML models, feature selection was performed on the development cohort using least absolute shrinkage and selection operator (LASSO) regression methods. As shown in Figure 2, the dimensionality was reduced to 15 features after LASSO, including preoperative fever, ASA, PT, hsCRP, BUN, diabetes, duration of surgery, ulinastatin, methylprednisolone, alanine aminotransferase, total volume of fluid loss, volume of blood loss, DBILI, albumin, and gender. Except albumin and gender, which were negatively associated with postoperative SIRS, all other features were positively associated with postoperative SIRS in older adult patients (Table 4).

XSL•FO
RenderX

Figure 2. Feature selection using least absolute shrinkage and selection operator (LASSO) regression.

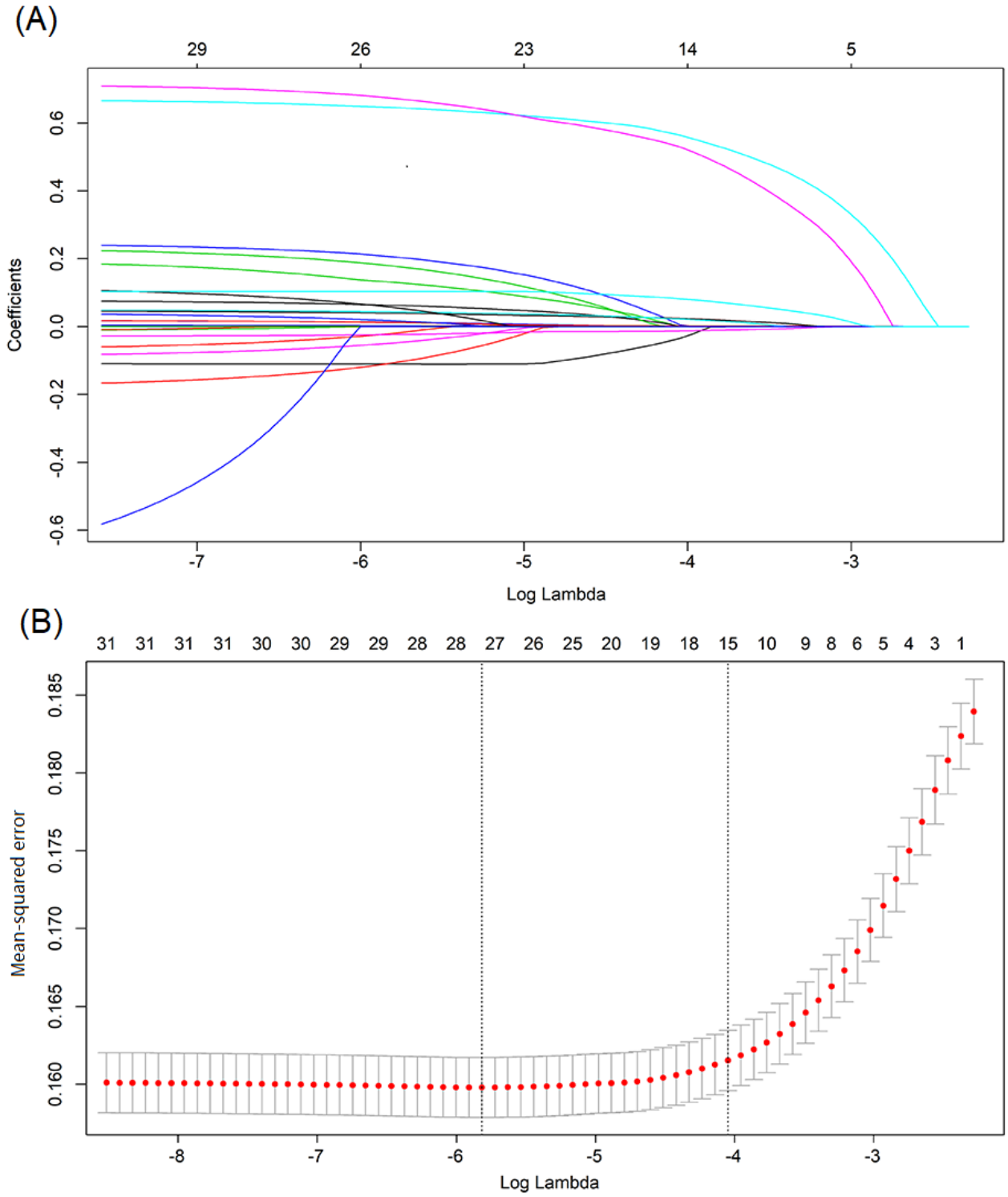

XSL•FO
RenderX

Table 4. Correlation coefficients between postoperative systemic inflammatory response syndrome (SIRS) and the 15 selected features using least absolute shrinkage and selection operator (LASSO) regression methods.

Characteristics	Correlation coefficient
Preoperative fever	0.104
ASA[a] classification	0.101
PT[b]	0.016
hs-CRP[c]	0.005
BUN[d]	0.004
Diabetes mellitus	0.001
Duration of surgery	<0.001
Ulinastatin	<0.001
Methylprednisone	<0.001
ALT[e]	<0.001
Total volume of fluid loss	<0.001
Volume of blood loss	<0.001
DBILI[f]	<0.001
Albumin	–0.002
Gender	–0.005

[a]ASA: American Society of Anesthesiologists.
[b]PT: prothrombin time.
[c]hs-CRP: high-sensitivity C-reactive protein.
[d]BUN: blood urea nitrogen.
[e]ALT: alanine aminotransferase.
[f]DBILI: direct bilirubin.

Model Construction, Internal Validation, and Horizontal Comparison

Finally, the 15 selected predictors were used in the 4 ML models, including RF, XGBoost, LR, and MLP, to predict postoperative SIRS. The performance of the different ML algorithms in the internal validation set are shown in Figure 3, and the calibration curves are presented in Figure 4. The AUC of the RF model was 0.751 (95% CI 0.709-0.793) with the highest sensitivity of 0.682 and specificity of 0.681. The F_1 score of the RF model (0.508) was the highest among the 4 ML models, and the Brier score (0.153) was also relatively higher. As a result, upon general consideration of the AUC, F_1 score, Brier score, calibration curve, and the ability of subsequent cross-center promotion and application, we thought that the RF model demonstrated the best performance. Additionally, the performance of RF model in the internal validation set was compared with the nomogram [23] we once established to predict postoperative SIRS in older patients (Figure 3). The results showed that the RF model had a significantly higher AUC (AUC=0.751) than the nomogram (AUC=0.671) in the internal validation set, which further proved the generalizability of the RF model (Table 5). The internal validation results after applying SMOTE are shown in Table S2 in Multimedia Appendix 1, where the AUC values increased for almost all ML algorithms except for LR and RF still exhibited the best performance with an AUC of 0.923.

XSL•FO
RenderX

Figure 3. Performance of different machine learning algorithms in the internal validation set. LR: logistic regression; MLP: multilayer perceptron; RF: random forest.

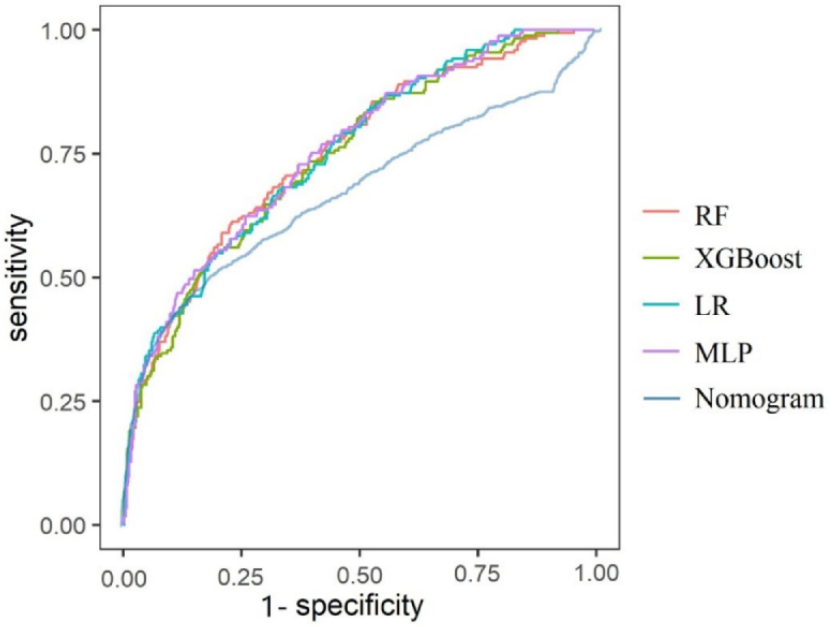

Figure 4. The calibration curves in the internal validation set for the (A) random forest model, (B) XGBoost model, (C) logistic regression model, and (D) multilayer perceptron model.

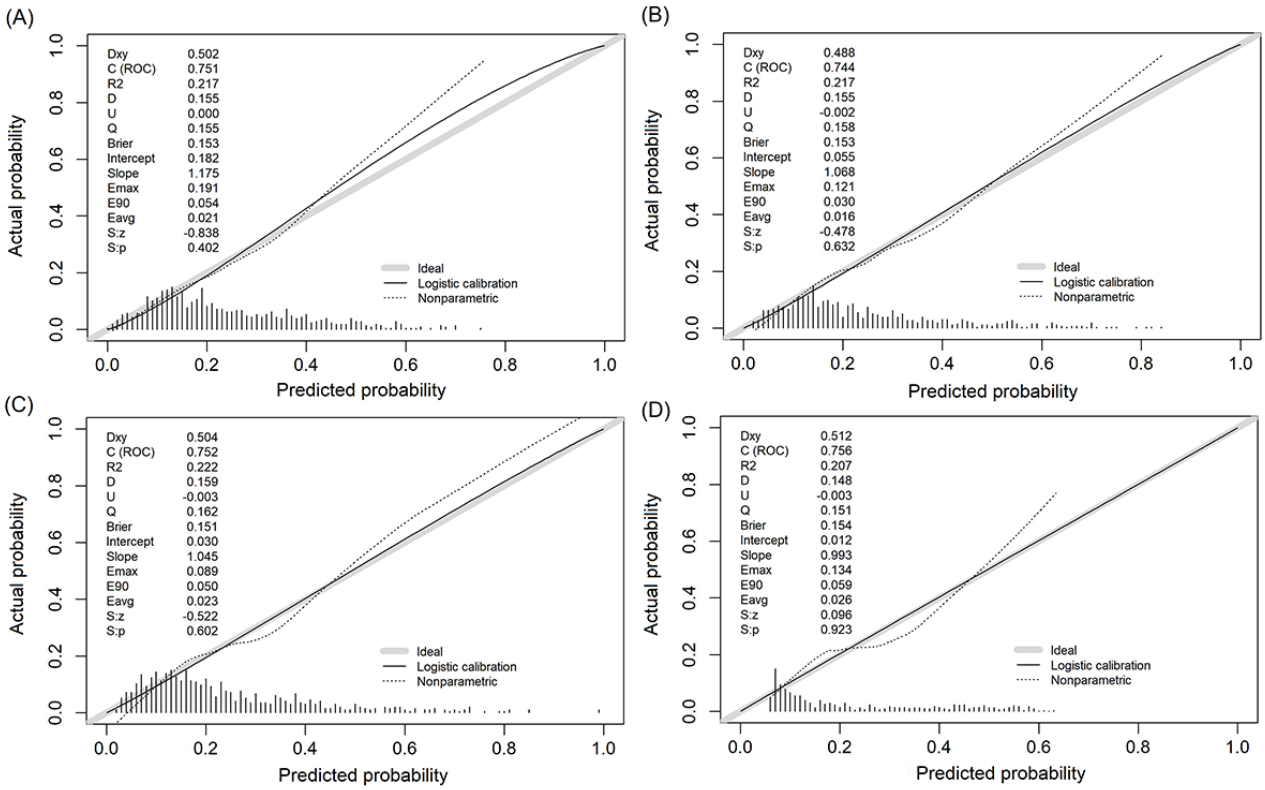

XSL•FO
RenderX

Table 5. Internal validation performance of different machine learning models.

Metrics	Random forest	XGBoost	Logistic regression	Multilayer perceptron	Nomogram
AUC[a] (95% CI)	0.751 (0.709-0.793)	0.744 (0.702-0.786)	0.752 (0.711-0.793)	0.756 (0.715-0.797)	0.671 (0.648-0.694)
Cutoff	0.237	0.226	0.230	0.216	0.276
Sensitivity	0.682	0.665	0.671	0.665	0.626
Specificity	0.681	0.667	0.672	0.665	0.625
Accuracy	0.682	0.666	0.672	0.665	0.626
F_1 score	0.508	0.489	0.496	0.488	0.448
PPV[b]	0.404	0.387	0.393	0.386	0.349
NPV[c]	0.871	0.863	0.866	0.862	0.839
Brier score	0.153	0.153	0.151	0.154	0.174

[a]AUC: area under the receiver operating curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

External Validation Performance

External validation of the developed RF algorithm was conducted with the eligible data from patients from Lingnan Hospital (external validation-1) and Yuedong Hospital (external validation-2). As shown in Table 6, the RF model achieved relatively higher AUCs for the external validation-1 (0.759, 95% CI 0.723-0.795) and external validation-2 (0.804, 95% CI 0.746-0.863) sets. Moreover, the external validation-2 set demonstrated much higher sensitivity (lower false-negative rate) than the internal validation set (0.800 vs 0.682), and the specificity of the model was improved in both external validation sets. On the other hand, the negative predictive value was high, and the positive predictive value (PPV) was relatively low for all 3 validation sets since we chose a balanced cutoff between sensitivity and specificity to ensure as high a true positive (TP) as possible, which also results in a high false positive (FP) and low false negative. Notably, the PPV and F_1 score of external validation-2 was very low due to the low incidence rate of postoperative SIRS (20/307, 6.5%) at Yuedong Hospital.

After applying SMOTE, the performance of the RF model in the external validation set was basically the same as the model without SMOTE, with AUCs of 0.783 vs 0.759 in external validation set 1 and 0.784 vs 0.804 in external validation set 2, which further proves that the RF model has good generalizability to different validation populations and different data processing methods (Table S3 in Multimedia Appendix 1).

Table 6. External validation performance of the random forest model.

Metrics	Internal validation	External validation-1	External validation-2
AUC[a] (95% CI)	0.751 (0.709-0.793)	0.759 (0.723-0.795)	0.804 (0.746-0.863)
Cutoff[b]	0.237	0.237	0.237
Sensitivity	0.682	0.680	0.800
Specificity	0.681	0.689	0.690
Accuracy	0.682	0.686	0.697
F_1 score	0.508	0.562	0.256
PPV[c]	0.404	0.479	0.152
NPV[d]	0.871	0.836	0.980

[a]AUC: area under the receiver operating curve.

[b]We chose a cutoff balanced between sensitivity and specificity.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

Feature Importance Weight

The feature importance permutation was used to rank the levels of feature importance, which is defined as the decrease in a model score when a single feature value is randomly shuffled (Figure 5). The results showed that the total volume of fluid loss, duration of surgery, ASA, and volume of blood loss had a significant impact on the outcome, with higher importance weight values of 0.121, 0.119, 0.118, and 0.101, respectively (Table 7).

XSL•FO
RenderX

Figure 5. Feature importance weight of the random forest algorithm. ALB: albumin; ALT: alanine aminotransferase; ASA: American Society of Anesthesiologists; BUN: blood urea nitrogen; DBILI: direct bilirubin; hsCRP: high sensitivity C-reactive protein; PT: prothrombin time.

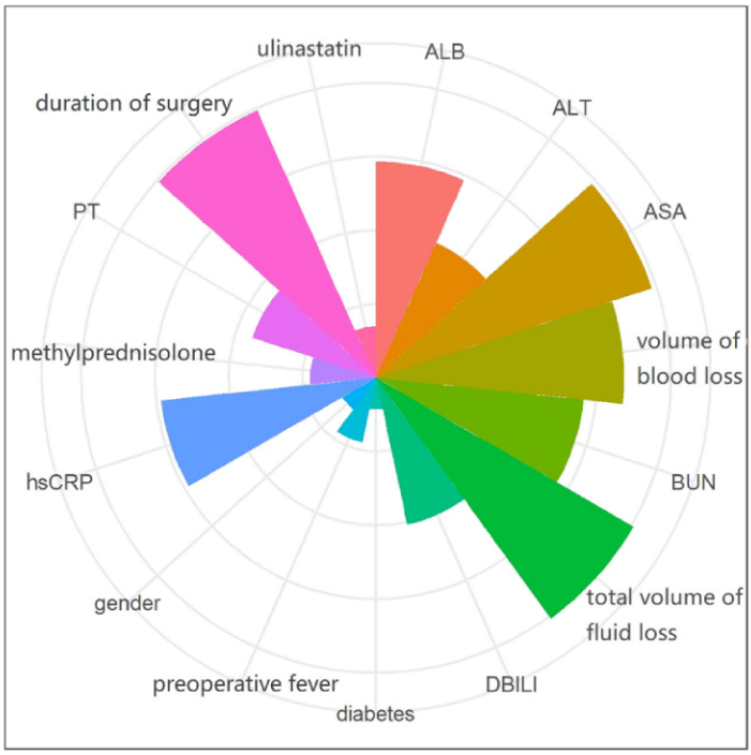

Table 7. Feature importance weight of the random forest algorithm.

Characteristics	Permutation importance scores, mean (SD)	Feature importance weight
Total volume of fluid loss	0.047 (0.002)	0.121
Duration of surgery	0.047 (0.001)	0.119
ASA[a]	0.046 (0.003)	0.118
Volume of blood loss	0.039 (0.001)	0.101
hs-CRP[b]	0.035 (0.002)	0.088
Albumin	0.034 (0.002)	0.088
BUN[c]	0.033 (0.000)	0.085
DBILI[d]	0.024 (0.001)	0.061
ALT[e]	0.023 (0.001)	0.060
PT[f]	0.021 (0.001)	0.053
Methylprednisolone	0.011 (0.001)	0.027
Preoperative fever	0.011 (0.001)	0.027
Ulinastatin	0.008 (0.001)	0.021
Gender	0.006 (0.000)	0.016
Diabetes mellitus	0.005 (0.000)	0.013

[a]ASA: American Society of Anesthesiologists.
[b]hs-CRP: high sensitivity C-reactive protein.
[c]BUN: blood urea nitrogen.
[d]DBILI: direct bilirubin.
[e]ALT: alanine aminotransferase.
[f]PT: prothrombin time.

XSL•FO
RenderX

Online Application

Figure S2 in Multimedia Appendix 1 shows the online calculator for our RF model to predict the risk of postoperative SIRS in older adult patients, which can be found at [36]. Users can quickly obtain the predicted probability of the patient's postoperative SIRS risk by entering the values of the 15 predictors.

Discussion

Principal Findings

In this study, we evaluated the ability of 4 ML algorithms including RF, XGBoost, LR, and MLP to predict postoperative SIRS in older adult patients based on eligible data from EHR systems and concluded that the RF model has moderately better performance than the other ML algorithms, with an AUC value of 0.751, highest sensitivity of 0.682, and specificity of 0.681. The RF model also exhibited better performance in the internal validation set than the nomogram we once established to predict postoperative SIRS in older patients. Furthermore, the applicability of the RF model was proved by external validation in the other 2 independent medical centers, with relatively higher AUC values of 0.759 and 0.804, indicating good reproducibility and generalizability of the model in older adult patients. Notably, the model exhibited a high negative predictive value and relatively low PPV because we chose a balanced cutoff between sensitivity and specificity to ensure as high a TP as possible, which inevitably leads to a high FP and low false negative. Especially for external validation-2 (Yuedong Hospital), which had a very low incidence rate of postoperative SIRS (20/307, 6.5%), 16 TP were successfully predicted with a high sensitivity of 0.800, which comes with 89 FP and results in a low PPV (0.152) and F_1 score (0.256).

The model indicators used in this study include preoperative fever, ASA, PT, hsCRP, BUN, diabetes, duration of surgery, ulinastatin, methylprednisolone, ALT, total volume of fluid loss, volume of blood loss, DBILI, albumin, and gender. These indicators are mainly related to the individual disease characteristics and are not affected by the level of different medical institutions or regional characteristics, which makes the model more generalizable. The results of our study further prove that the model has a good generalization ability for populations with different SIRS incidences. The high recall of our model is important for reminders to check patients early and prevent SIRS in advance; this demonstrates a more responsible attitude toward patients.

Currently, given the more severe mortality and adverse prognosis, most studies have adopted sepsis as a clinical endpoint [2]. In our study, we used SIRS as the primary outcome because it has been an acknowledged criterion that is easier to identify and can help physicians notice the possibility of sepsis and prescribe tests to examine whether infection truly exists.

As the population ages, the surgical population is also aging faster than the general population, with higher morbidity and mortality rates [37]. Notably, older adults are predisposed to postoperative infections, SIRS, and even sepsis due to preexisting comorbidities, repeated and prolonged hospitalizations, immune dysregulation, and functional limitations [38]. Although preoperative antimicrobial prophylaxis has been routinely administered for older adult patients in various specialized operations, the persistent high incidence of postoperative SIRS suggests that it cannot be effectively prevented in this way [12]. In fact, for older patients undergoing different types of surgery, there is still a lack of effective tools to identify high-risk patients with postoperative SIRS and assist with decision-making on the need for individualized interventions.

In this study, we used ML methods to analyze diverse data types [13] due to the advantages of flexibility and scalability. A predictive model based on an RF algorithm was developed to identify older adult patients with a high risk of postoperative SIRS. Internal validation and two external validations confirmed that the established model could predict postoperative SIRS with high accuracy and specificity. The results might be important derivatives and supplements to the current perioperative prevention and management programs, which enable surgeons and anesthesiologists to identify older adult patients who are suspected of having postoperative SIRS. In addition to screening high-risk groups, this model can also help prevent and treat postoperative SIRS in older adult patients more accurately and in a timelier manner by using various drug or nondrug means under the guideline of enhanced recovery after surgery, so as to promote the short-term and long-term prognoses of patients [39]. Finally, an online risk calculator was developed to improve clinical usability and make our model accessible to anesthesiologists and peers around the world.

Comparison With Prior Work

Distinguished from previous investigations that mainly focused on a single indicator or a single surgery, we fully used the perioperative data of patients and constructed an optimal combination of risk factors to predict postoperative SIRS in older adult patients. In this study, 15 variables were identified to be significantly associated with postoperative SIRS, including preoperative fever, ASA [40], PT, hsCRP, BUN, diabetes [41], duration of surgery [42], ulinastatin, methylprednisolone, ALT, total volume of fluid loss, volume of blood loss, DBILI [43], albumin, and gender [44]. These variables have also been reported to be associated with postoperative SIRS and sepsis in earlier studies, adding clinical credibility to our model. Meanwhile, all the variables were routinely recorded and widely used in clinical practice, which makes the model more feasible and can be widely used in different hospitals.

Methylprednisolone and ulinastatin have also been associated with postoperative SIRS, which may be attributed to the fact that both medicines are widely used among high-risk surgical patients with postoperative infection according to anesthesiologists' clinical experience [45-47]. Additionally, intraoperative fluid loss ranks first in the feature importance weight of the RF model, which indicates that fluid management such as intraoperative intravenous infusion should be given higher priority during the perioperative period to prevent postoperative SIRS, and this was consistent with the consensus on early management of sepsis [48].

XSL•FO
RenderX

Limitations

Notably, several limitations must be listed here, including that (1) this is a retrospective cohort study with collection and entry bias, as well as possible residual confounding, which requires future prospective studies to validate the model. In addition, some indicators for SOFA scoring were missing, so we could not perform a comparative study between our model and SOFA scoring. (2) We selected 36 variables based on the availability in the EHR, correlation with SIRS risk, and predictive potential, but it should be noted that we may have missed other indicators with more predictive performance in postoperative SIRS due to the limitation of our data. (3) The model demonstrated a low PPV because we wanted the model to have a relatively high recall rate to predict more patients who were truly at a high risk of postoperative SIRS, which would inevitably lead to more FPs, but from a health economics perspective, the economic burden associated with enhanced perioperative management of high-risk patients is much less than that resulting from the progression of the patient to SIRS or even after sepsis, resulting in medical-economic stress. (4) The generalizability of this model in different ethnic groups and different regions also needs further validation. (5) It should be emphasized that estimates in our model are predictive and should not be interpreted as causal [49], such as the association between intraoperative use of methylprednisolone and ulinastatin and the higher SIRS incidence rate. Intraoperative use of such drugs is likely a marker of risk stratification with a lower risk of infection patients using these drugs than those at high risk.

Conclusions

We enrolled 3 independent cohorts to develop and validate a generalizable RF model for the prediction of postoperative SIRS in older adult patients that enables surgeons and anesthesiologists to screen susceptible and high-risk populations for SIRS in older adult surgical patients and to implement early individualized interventions based on existing prevention and management programs.

Acknowledgments

The authors thank all the patients who kindly participated in this study.

This study was supported partly by the Key-Area Research and Development Program of Guangdong Province (number 2020B0101130015); Special Support Project of Guangdong Province (0720240209); the Natural Science Foundation of Guangdong Province (grant number 2022A1515012603); Science and Technology Program of Guangzhou, China (grant number 2024A04J4246); and the “Five and five” Project of the Third Affiliated Hospital of Sun Yat-Sen University (grant number 2023WW501) and Science and Technology Planning Project of Guangdong Province-Regional Innovation Capacity and Support System Construction (grant number 2023B110006).

Data Availability

The raw data supporting the conclusions of this article will be made available by the corresponding authors via email.

Authors' Contributions

CC, ZL, and YZ conceived the study. CC, ZL, YZ, and HM designed the study. HM, YF, TL, XL, YZ, and SZ collected and analyzed the data. CC, HM, YF, XL, and YZ wrote the manuscript. All authors edited and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Missing rates, formulas for evaluation indicators, deleted variables, internal validation results after applying SMOTE, external validation, and the online calculator.
[DOCX File , 1358 KB - jmir_v26i1e57486_app1.docx]

References

1. Jeong H, Cha BG, Kang D, Kim DY, Yang W, Ki S, et al. Ceria nanoparticles fabricated with 6-aminohexanoic acid that overcome systemic inflammatory response syndrome. Adv Healthc Mater 2019 May;8(9):e1801548. [doi: 10.1002/adhm.201801548] [Medline: 30843374]
2. Simpson SQ. SIRS in the time of sepsis-3. Chest 2018 Jan;153(1):34-38. [doi: 10.1016/j.chest.2017.10.006] [Medline: 29037526]
3. Serafim R, Gomes JA, Salluh J, Póvoa P. A comparison of the quick-SOFA and systemic inflammatory response syndrome criteria for the diagnosis of sepsis and prediction of mortality: a systematic review and meta-analysis. Chest 2018 Mar;153(3):646-655. [doi: 10.1016/j.chest.2017.12.015] [Medline: 29289687]

XSL•FO
RenderX

4. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
5. Balk RA. Systemic inflammatory response syndrome (SIRS): where did it come from and is it still relevant today? Virulence 2014 Jan 01;5(1):20-26 [FREE Full text] [doi: 10.4161/viru.27135] [Medline: 24280933]
6. Bown MJ, Nicholson ML, Bell PRF, Sayers RD. The systemic inflammatory response syndrome, organ failure, and mortality after abdominal aortic aneurysm repair. J Vasc Surg 2003 Mar;37(3):600-606 [FREE Full text] [doi: 10.1067/mva.2003.39] [Medline: 12618699]
7. Wu W, Zhang D, Jin T, Lu T, Zhou F. Progress in the study of biomarkers for early prediction of systemic inflammatory response syndrome after percutaneous nephrolithotomy. Front Immunol 2023;14:1142346 [FREE Full text] [doi: 10.3389/fimmu.2023.1142346] [Medline: 37063849]
8. Liu D, Huang S, Sun J, Zhang H, Cai Q, Gao C, et al. Sepsis-induced immunosuppression: mechanisms, diagnosis and current treatment options. Mil Med Res 2022 Oct 09;9(1):56 [FREE Full text] [doi: 10.1186/s40779-022-00422-y] [Medline: 36209190]
9. Ferraris VA, Ballert EQ, Mahan A. The relationship between intraoperative blood transfusion and postoperative systemic inflammatory response syndrome. Am J Surg 2013 Apr;205(4):457-465. [doi: 10.1016/j.amjsurg.2012.07.042] [Medline: 23414633]
10. Vogel TR, Dombrovskiy VY, Carson JL, Graham AM, Lowry SF. Postoperative sepsis in the United States. Ann Surg 2010 Dec;252(6):1065-1071 [FREE Full text] [doi: 10.1097/SLA.0b013e3181dcf36e] [Medline: 20571363]
11. Boonmee P, Ruangsomboon O, Limsuwat C, Chakorn T. Predictors of mortality in elderly and very elderly emergency patients with sepsis: a retrospective study. West J Emerg Med 2020 Oct 06;21(6):210-218 [FREE Full text] [doi: 10.5811/westjem.2020.7.47405] [Medline: 33207168]
12. Chen C, Chen X, Chen J, Xing J, Hei Z, Zhang Q, et al. Association between preoperative hs-CRP/albumin ratio and postoperative SIRS in elderly patients: a retrospective observational cohort study. J Nutr Health Aging 2022;26(4):352-359 [FREE Full text] [doi: 10.1007/s12603-022-1761-4] [Medline: 35450991]
13. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]
14. Zhu Y, Liu X, Li Y, Yi B. The applications and prospects of big data in perioperative anesthetic management. APS 2024 Aug 26;2(3):1. [doi: 10.1007/s44254-024-00068-0]
15. Li C, Zhou R, Chen G, Hao X, Zhu T. Knowledge mapping and research hotspots of artificial intelligence on ICU and Anesthesia: from a global bibliometric perspective. APS 2023 Oct 23;1(4):1. [doi: 10.1007/s44254-023-00031-5]
16. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. J Am Med Inform Assoc 2022 Jan 29;29(3):559-575 [FREE Full text] [doi: 10.1093/jamia/ocab236] [Medline: 34897469]
17. Zheng Y, Wang J, Ling Z, Zhang J, Zeng Y, Wang K, et al. A diagnostic model for sepsis-induced acute lung injury using a consensus machine learning approach and its therapeutic implications. J Transl Med 2023 Sep 12;21(1):620 [FREE Full text] [doi: 10.1186/s12967-023-04499-4] [Medline: 37700323]
18. Fan Z, Jiang J, Xiao C, Chen Y, Xia Q, Wang J, et al. Construction and validation of prognostic models in critically Ill patients with sepsis-associated acute kidney injury: interpretable machine learning approach. J Transl Med 2023 Jun 22;21(1):406 [FREE Full text] [doi: 10.1186/s12967-023-04205-4] [Medline: 37349774]
19. Faisal M, Scally A, Richardson D, Beatson K, Howes R, Speed K, et al. Development and external validation of an automated computer-aided risk score for predicting sepsis in emergency medical admissions using the patient's first electronically recorded vital signs and blood test results. Crit Care Med 2018 Apr;46(4):612-618. [doi: 10.1097/CCM.0000000000002967] [Medline: 29369828]
20. Hu C, Li L, Huang W, Wu T, Xu Q, Liu J, et al. Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. Infect Dis Ther 2022 Jun 10;11(3):1117-1132 [FREE Full text] [doi: 10.1007/s40121-022-00628-6] [Medline: 35399146]
21. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018 Jan 26;8(1):e017833 [FREE Full text] [doi: 10.1136/bmjopen-2017-017833] [Medline: 29374661]
22. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
23. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020 Mar;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
24. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, SCCM/ESICM/ACCP/ATS/SIS. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. Crit Care Med 2003 Apr;31(4):1250-1256. [doi: 10.1097/01.CCM.0000050454.01978.3B] [Medline: 12682500]

XSL•FO
RenderX

25. Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. NPJ Digit Med 2020 Sep 23;3(1):125 [FREE Full text] [doi: 10.1038/s41746-020-00331-1] [Medline: 33043149]
26. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020 Mar 18;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
27. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
28. Li X, Lu Y, Chen C, Luo T, Chen J, Zhang Q, et al. Development and validation of a patient-specific model to predict postoperative SIRS in older patients: A two-center study. Front Public Health 2023;11:1145013 [FREE Full text] [doi: 10.3389/fpubh.2023.1145013] [Medline: 37139371]
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002 Jun 01;16:321-357 [FREE Full text] [doi: 10.1613/jair.953]
30. Torgo L. An infra-structure for performance estimation and experimental comparison of predictive models in R. arXiv 2015:1 [FREE Full text] [doi: 10.48550/arXiv.1412.0436]
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011;12:2825-2830 [FREE Full text]
32. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 Sep;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
33. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016:785-794. [doi: 10.1145/2939672.2939785]
34. McKinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011. URL: https://docslib.org/doc/4231522/a-foundational-python-library-for-data-analysis-and-statistics [accessed 2024-11-12]
35. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. URL: https://www.R-project.org/ [accessed 2024-11-12]
36. Caculation tool of Random Forest model to estimate SIRS. The Third Affiliated Hospital, Sun-Yat-Sen University. URL: https://analysis.aidcloud.cn/cn/sirs/ [accessed 2024-11-11]
37. Fowler AJ, Abbott TEF, Prowle J, Pearse RM. Age of patients undergoing surgery. Br J Surg 2019 Jul;106(8):1012-1018. [doi: 10.1002/bjs.11148] [Medline: 31115918]
38. Nasa P, Juneja D, Singh O. Severe sepsis and septic shock in the elderly: An overview. World J Crit Care Med 2012 Feb 04;1(1):23-30 [FREE Full text] [doi: 10.5492/wjccm.v1.i1.23] [Medline: 24701398]
39. Chen C, Li X, Luo G, Zhou S, Hei Z. Realising the full potential of anaesthesiology to promote enhanced recovery after surgery programmes in China. Br J Anaesth 2021 May;126(5):e157-e159 [FREE Full text] [doi: 10.1016/j.bja.2021.01.016] [Medline: 33618827]
40. Sood A, Abdollah F, Sammon JD, Arora N, Weeks M, Peabody JO, et al. Postoperative sepsis prediction in patients undergoing major cancer surgery. J Surg Res 2017 Mar;209:60-69. [doi: 10.1016/j.jss.2016.09.059] [Medline: 28032572]
41. Blair LJ, Huntington CR, Cox TC, Prasad T, Lincourt AE, Gersin KS, et al. Risk factors for postoperative sepsis in laparoscopic gastric bypass. Surg Endosc 2016 Apr;30(4):1287-1293. [doi: 10.1007/s00464-015-4349-9] [Medline: 26130133]
42. Kreydin EI, Eisner BH. Risk factors for sepsis after percutaneous renal stone surgery. Nat Rev Urol 2013 Oct;10(10):598-605. [doi: 10.1038/nrurol.2013.183] [Medline: 23999583]
43. Zhang H, Meng F, Lu S. Nomograms predicting the occurrence of sepsis in patients following major hepatobiliary and pancreatic surgery. Gastroenterol Res Pract 2020;2020:9761878 [FREE Full text] [doi: 10.1155/2020/9761878] [Medline: 32802049]
44. Zuckerman SL, Lakomkin N, Stannard BP, Hadjipanayis CG, Shaffrey CI, Smith JS, et al. Incidence and predictive factors of sepsis following adult spinal deformity surgery. Neurosurgery 2018 Nov 01;83(5):965-972. [doi: 10.1093/neuros/nyx578] [Medline: 29281103]
45. Takaki M, Ichikado K, Kawamura K, Gushima Y, Suga M. The negative effect of initial high-dose methylprednisolone and tapering regimen for acute respiratory distress syndrome: a retrospective propensity matched cohort study. Crit Care 2017 Jun 08;21(1):135 [FREE Full text] [doi: 10.1186/s13054-017-1723-0] [Medline: 28592332]
46. Shao Y, Zhang L, Deng L, Yao H. [Clinical study on effects of ulinastatin on patients with systemic inflammatory response syndrome]. Zhongguo Wei Zhong Bing Ji Jiu Yi Xue 2005 Apr;17(4):228-230. [Medline: 15836828]
47. Shu H, Liu K, He Q, Zhong F, Yang L, Li Q, et al. Ulinastatin, a protease inhibitor, may inhibit allogeneic blood transfusion-associated pro-inflammatory cytokines and systemic inflammatory response syndrome and improve postoperative recovery. Blood Transfus 2014 Jan;12 Suppl 1(Suppl 1):s109-s118 [FREE Full text] [doi: 10.2450/2013.0224-12] [Medline: 23736923]
48. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Med 2021 Nov;47(11):1181-1247 [FREE Full text] [doi: 10.1007/s00134-021-06506-y] [Medline: 34599691]

XSL•FO
RenderX

49. Adibi A, Sin DD, Safari A, Johnson KM, Aaron SD, FitzGerald JM, et al. The Acute COPD Exacerbation Prediction Tool (ACCEPT): a modelling study. Lancet Respir Med 2020 Oct;8(10):1013-1021. [doi: 10.1016/S2213-2600(19)30397-2] [Medline: 32178776]

Abbreviations

ALT: alanine aminotransferase
ASA: American Society of Anesthesiologists
AUC: area under the receiver operating curve
BUN: blood urea nitrogen
DBILI: direct bilirubin
EHR: electronic health record
EPV: events per variable
FP: false positive
hs-CRP: high-sensitivity C-reactive protein
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
ML: machine learning
MLP: multilayer perceptron
PACS: picture archiving and communication system
PPV: positive predictive value
PT: prothrombin time
RF: random forest
SIRS: systemic inflammatory response syndrome
SMOTE: synthetic minority oversampling technique
SOFA: Sequential Organ Failure Assessment
TP: true positive

Edited by A Coristine; submitted 19.02.24; peer-reviewed by B Yi, X Xu; comments to author 07.05.24; revised version received 20.05.24; accepted 04.11.24; published 22.11.24.

Please cite as:
Mai H, Lu Y, Fu Y, Luo T, Li X, Zhang Y, Liu Z, Zhang Y, Zhou S, Chen C
Identification of a Susceptible and High-Risk Population for Postoperative Systemic Inflammatory Response Syndrome in Older Adults: Machine Learning–Based Predictive Model
J Med Internet Res 2024;26:e57486
URL: https://www.jmir.org/2024/1/e57486
doi:10.2196/57486
PMID:

©Haiyan Mai, Yaxin Lu, Yu Fu, Tongsen Luo, Xiaoyue Li, Yihan Zhang, Zifeng Liu, Yuenong Zhang, Shaoli Zhou, Chaojin Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Prediction Model to Identify Clinically Relevant Medication Discrepancies at the Emergency Department (MED-REC Predictor): Development and Validation Study

Greet Van De Sijpe[1,2], PharmD; Matthias Gijsen[1,2], PharmD, PhD; Lorenz Van der Linden[1,2], PharmD, PhD; Stephanie Strouven[3], MSc; Eline Simons[1], PharmD; Emily Martens[4], PharmD; Nele Persan[4], PharmD; Veerle Grootaert[4], PharmD; Veerle Foulon[2], PharmD, PhD; Minne Casteels[2], MD, PhD; Sandra Verelst[5], MD, PhD; Peter Vanbrabant[5], MD; Sabrina De Winter[1*], PharmD, PhD; Isabel Spriet[1,2*], PharmD, PhD

[1]Pharmacy Department, University Hospitals Leuven, Leuven, Belgium

[2]Clinical Pharmacology and Pharmacotherapy, Department of Pharmaceutical and Pharmacological Sciences, KU Leuven, Leuven, Belgium

[3]Department of Gynaecology and Obstetrics, University Hospitals Leuven, Leuven, Belgium

[4]Pharmacy Department, General Hospital Sint-Jan Brugge-Oostende AV, Brugge, Belgium

[5]Department of Emergency Medicine, University Hospitals Leuven, Leuven, Belgium

[*]these authors contributed equally

Corresponding Author:
Greet Van De Sijpe, PharmD
Pharmacy Department
University Hospitals Leuven
Herestraat 49
Leuven, 3000
Belgium
Phone: 32 16345969
Email: greet.vandesijpe@uzleuven.be

Abstract

Background: Many patients do not receive a comprehensive medication reconciliation, mostly owing to limited resources. We hence need an approach to identify those patients at the emergency department (ED) who are at increased risk for clinically relevant discrepancies.

Objective: The aim of our study was to develop and externally validate a prediction model to identify patients at risk for at least 1 clinically relevant medication discrepancy upon ED presentation.

Methods: A prospective, multicenter, observational study was conducted at the University Hospitals Leuven and General Hospital Sint-Jan Brugge-Oostende AV, Belgium. Medication histories were obtained from patients admitted to the ED between November 2017 and May 2022, and clinically relevant medication discrepancies were identified. Three distinct datasets were created for model development, temporal external validation, and geographic external validation. Multivariable logistic regression with backward stepwise selection was used to select the final model. The presence of at least 1 clinically relevant discrepancy was the dependent variable. The model was evaluated by measuring calibration, discrimination, classification, and net benefit.

Results: We included 824, 350, and 119 patients in the development, temporal validation, and geographic validation dataset, respectively. The final model contained 8 predictors, for example, age, residence before admission, number of drugs, and number of drugs of certain drug classes based on Anatomical Therapeutic Chemical coding. Temporal validation showed excellent calibration with a slope of 1.09 and an intercept of 0.18. Discrimination was moderate with a *c*-index (concordance index) of 0.67 (95% CI 0.61-0.73). In the geographic validation dataset, the calibration slope and intercept were 1.35 and 0.83, respectively, and the *c*-index was 0.68 (95% CI 0.58-0.78). The model showed net benefit over a range of clinically reasonable threshold probabilities and outperformed other selection criteria.

Conclusions: Our software-implemented prediction model shows moderate performance, outperforming random or typical selection criteria for medication reconciliation. Depending on available resources, the probability threshold can be customized to increase either the specificity or the sensitivity of the model.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55185) doi:10.2196/55185

KEYWORDS

medication reconciliation; medication discrepancy; emergency department; prediction model; risk stratification; MED-REC predictor; MED-REC; predictor; patient; medication; hospital; software-implemented prediction model; software; geographic validation; geographic

Introduction

Medication discrepancies occur frequently upon admission to the emergency department (ED) and are a major source of avoidable harm [1-5]. Taking an accurate medication history early on during the patient trajectory is the typical first step needed to tackle medication safety in a multifaceted approach. Medication reconciliation is necessary to identify drug-related problems, rendering it possible to assess therapy appropriateness. Accordingly, an accurate medication history is crucial in ensuring correct medication prescriptions during hospital stay and upon discharge [6]. Conversely, medication history errors may result in iatrogenic patient harm, leading to prolonged length of stay or even death [7-9]. Obtaining an accurate medication history early on during hospitalization, for example, while the patient is still in the ED, is clearly the preferred approach. Fixing medication errors downstream has largely been considered to be both unsafe and inefficient [10].

However, the medication reconciliation process is prone to many errors and is labor-intensive [10]. The hectic environment along with competing medical priorities among physicians and nurses, unreliable information due to the acute illness of the patient, and the nonfamiliarity of drug names for the patient or family members lead to incomplete and inaccurate medication histories [11,12]. The MARQUIS study found that pharmacist involvement was a key component in improving the reliability of medication reconciliation [13]. Unfortunately, this requires a level of staffing resources beyond usual care, which is often not available [10,14]. As a result, many patients simply do not receive a thorough and complete medication reconciliation. Therefore, an approach to prioritize patients who are most likely to benefit from medication reconciliation would be of great benefit.

To date, literature on medication reconciliation risk stratification tools is scarce [1,15-18]. Our research group published a prediction model almost a decade ago that identified patients at risk for medication discrepancies upon ED presentation. The model was based on 13 parameters that can be easily extracted from the electronic health record (EHR). The downside of this prediction model was its inability to identify clinically relevant discrepancies, thereby limiting the usefulness of this model [19].

Hence, we need an approach enabling us to identify those patients at the ED who are at increased risk for clinically relevant discrepancies. As such, priority for a pharmacist-led medication reconciliation can be given to these patients. The aim of this study was to develop and externally validate a prediction model, the MED-REC predictor, that can be programmed in the EHR to identify patients at risk for at least 1 clinically relevant medication discrepancy upon ED presentation.

Methods

Appropriate reporting of our prediction model was performed based on the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guideline [20].

Study Design and Setting

A prospective, multicenter study was carried out at the EDs of University Hospitals Leuven (UZ Leuven) and General Hospital Sint-Jan Brugge-Oostende AV (AZ Sint-Jan Brugge-Oostende), Belgium, from November 2017 to May 2022.

UZ Leuven is a tertiary 1995-bed teaching hospital. The ED consists of an admission and a treatment area (12 boxes), an ambulatory zone, a pediatric zone, and 3 observation care units (30 beds of which 6 are equipped as intensive care beds). On average, 197 patients per day visit the ED, of which approximately 30% (60/197) are hospitalized.

AZ Sint-Jan Brugge-Oostende is a 1182-bed general hospital. The ED consists of an admission and treatment area (11 boxes), an ambulatory zone, a pediatric zone, and 3 beds reserved for critically ill patients. Approximately 92 patients per day visit the ED, of which 38% (35/92) are hospitalized.

Depending on the presumed medical diagnosis, patients are treated by a physician specialized in internal medicine, surgery, emergency medicine, pediatrics, or psychiatry. As part of the obligatory patient assessment, a medication history is obtained by the physician in all patients within 24 hours of the ED visit and is entered in a dedicated medication history module in the EHR. Subsequently, the entered medication history should be electronically validated once considered complete.

Patients

Pharmacy, biomedical sciences, or hospital pharmacy students trained in medication reconciliation enrolled planned admission patients on weekdays between 8:30 AM and 5 PM, starting with the patient present the longest and proceeding in descending order. Medication reconciliation was performed based on student availability during their internship. Medication reconciliation was performed independently from the information obtained by physicians. While medication reconciliation was performed between 8:30 AM and 5 PM on weekdays, the sample included patients arriving to the ED during the evening, night, and weekend as a result of the need for observation in the ED or waiting for inpatient beds. Exclusion criteria were (1) patients aged younger than 18 years; (2) patients receiving end-of-life care; (3) patients transferred from another hospital or ward; (4) patients not speaking Dutch, French, or English; (5) patients

XSL•FO
RenderX

discharged home or deceased before the medication history was obtained; (6) patients with intentional intoxication; or (7) patients in isolation due to a possible COVID-19 infection.

Training

Every student performing medication reconciliation received formal training by a clinical pharmacist (SDW or GVDS) prior to patient recruitment. Interrater reliability for medication reconciliation was evaluated using a similar methodology as previously described by Pippins et al [1]. Every student had to obtain 4 fictional medication histories through role-play. Results were compared with the best possible medication history as composed by the research team. A cutoff of 90% agreement was set for each fictive case.

Data Collection

Overview

Three distinct datasets were collected. The first dataset (UZ Leuven; from November 2017 to September 2019) was used to develop and internally validate the MED-REC predictor. The second dataset (UZ Leuven; from October 2021 to April 2022) and the third dataset (AZ Sint-Jan Brugge-Oostende; from April to May 2022) were used for temporal and geographic external validation, respectively. Temporal validation assesses the performance of the model on patients from a later period, and geographic validation uses data from a different hospital [20].

Medication Reconciliation

A structured form—including a checklist, a table, and a standardized list of questions—and a protocol were used to guide the students in obtaining the best possible medication history (Multimedia Appendix 1) [19]. The protocol included 3 steps. The first step consisted of a conversation with the patient. The patient was explicitly asked about prescription and nonprescription drugs easily forgotten, such as transdermal patches, inhalers, and dietary supplements. If the patient was unable to communicate or was unaware of his or her current medication intake, a close relative was interviewed. Other sources were used as well to retrieve information including verification of medication boxes if brought to the ED, previous medical records, and referral notes. The second step consisted of contacting the community pharmacist. The third step involved contacting the general practitioner if there was a discrepancy between the list obtained from the patient or his or her relative and the list obtained from the community pharmacist, or if the community pharmacist had no overview of the patient's current medication intake.

When resources were insufficient to verify the medication schedule, for example, when neither the community pharmacist nor the general practitioner could be reached and the gold standard was thus unavailable, the patient was excluded.

Discrepancies and Their Clinical Relevance

The medication history obtained by the trained student was defined as the gold standard. A discrepancy was defined as any difference between the physician-acquired medication history and the gold standard. Discrepancies consisted of a drug omission or commission, an incorrect dose or dosage regimen.

A classification was developed to categorize discrepancies according to perceived clinical relevance. The classification was defined by an expert panel, consisting of 4 clinical pharmacists (SDW, LVdL, ES, and IS) and 1 physician (PV) with substantial expertise in medication management. SDW and LVdL are experienced ED clinical pharmacists, ES is a junior clinical pharmacist, and IS is a senior clinical pharmacist. PV is a senior physician specializing in both internal medicine as well as emergency medicine, with 20 years of experience. The expert panel independently assigned clinical relevance categories to the major drug classes listed in the national (Belgian) drug formulary. In case of the absence of consensus, SDW made the final decision. Three categories were used, which were based on the methodology proposed by Cornish et al [2]. Only discrepancies that may lead to severe discomfort or clinical deterioration were included for the development of the MED-REC predictor and are further specified in this work as "clinically relevant discrepancies." An overview is presented in Table 1. Discrepancies involving one of these drugs are assumed to affect treatment or urgent procedures in the ED and should be corrected before the patient is discharged from the ED. "As needed" medication was never considered clinically relevant.

XSL•FO
RenderX

Table 1. Classification of clinically relevant medication discrepancies in medication histories. A drug omission or commission and an incorrect dose or dosage regimen concerning these drugs were considered clinically relevant discrepancies. Groups were based on the World Health Organization's Anatomical Therapeutic Chemical (ATC) classification system.

Anatomical or pharmacological group	Chemical, pharmacological, or therapeutic subgroup
A. Alimentary tract and metabolism	• A10 drugs used in diabetes
B. Blood and blood forming organs	• B01A antithrombotic agents • B02 antihemorrhagics
C. Cardiovascular system	• C01 cardiac therapy • C02 antihypertensives • C03 diuretics • C04 peripheral vasodilators • C07 beta blocking agents • C08 calcium channel blockers • C09 agents acting on the renin-angiotensin system
H. Systemic hormonal preparations. (excluding sex hormones and insulins)	• H01 pituitary and hypothalamic hormones and analogs • H02 corticosteroids for systemic use
J. Anti-infectives for systemic use	• J01 antibacterials for systemic use • J02 antimycotics for systemic use • J04 antimycobacterials • J05 antivirals for systemic use • J06 immune sera and immunoglobulins
L. Antineoplastic and immunomodulating agents	• L01 antineoplastic agents • L03 immunostimulants • L04 immunosuppressants
N. Nervous system	• N02A opioids • N03 antiepileptics • N05 psycholeptics • N07A parasympathomimetics • N07B drugs used in addictive disorders • N07X other nervous system drugs
P. Antiparasitic products, insecticides, and repellents	• P01 antiprotozoals
R. Respiratory system	• R03 drugs for obstructive airway disease

Patient-, Drug-, ED Visit–, and Medication Reconciliation–Related Factors

A structured form was used to collect patient-, drug-, ED visit–, and medication reconciliation–related factors. Details are provided in Multimedia Appendix 2.

Model Development

Variables were considered for the MED-REC predictor if (1) a significant association between the variable and the occurrence of medication discrepancies had been shown in the literature [21] or if (2) the variable was considered of interest as discussed by our multidisciplinary research team. Furthermore, EHR availability of the variable upon ED visit was required.

Multivariable logistic regression with backward stepwise selection according to the Akaike information criterion was used to select the final model. The presence of at least 1 clinically relevant discrepancy was the dependent variable. Sex, age, language (Dutch, French, or English), residence before ED admission (home, nursing home, or others), transport to the ED (patient's own transport, ambulance, or emergency physician transport vehicle), ED triage acuity scale using the Emergency Severity Index (ESI), number of drugs documented by the ED physician in the medication history module, number of drugs documented by the ED physician in the medication history module of a certain first WHO (World Health Organization) Anatomical Therapeutic Chemical (ATC) level (ATC A, ATC B, etc), number of high-risk drugs documented by the ED physician in the medication history module, time of ED visit (morning, afternoon, evening, or night), specialty of the physician obtaining the medication history (emergency medicine, internal medicine, psychiatry, or surgery), availability of a regionally shared electronic medication record, and electronic validation of the medication history were investigated as predictor variables.

Continuous predictors were evaluated for nonlinearity using scatter plots, and various transformations were applied to address any nonlinear relationships. However, these transformations did not improve model performance compared with simpler models. Therefore, we decided to forego transforming continuous predictors to maintain interpretability and simplicity.

XSL•FO
RenderX

Missing data were present for the ED triage acuity scale (160/824, 19.4%) and were handled using dummy variable adjustment.

Internal and External Validation

The predictive performance of the model was assessed by measuring calibration, discrimination, and classification in all datasets. Calibration reflects the agreement between model-based predictions and observed outcomes, and was assessed using calibration plots with a target slope and intercept of 1 and 0, respectively. Discrimination refers to the ability of the prediction model to differentiate between those experiencing at least 1 clinically relevant discrepancy and those who do not. Discrimination was evaluated using the concordance index (*c*-index), for example, the area under the receiver operating characteristic curve (AUROC) for models with binary end points. Bootstrap 95% CIs were calculated. We assessed the following classification measures in various scenarios relevant to daily clinical practice: specificity, sensitivity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio, and alert rate. Scenarios were defined by specific probability thresholds, which were used to classify patients predicted to show at least 1 clinically relevant medication discrepancy or not. We investigated the Youden index (maximized sensitivity and specificity), with a sensitivity of at least 80% for hospitals with sufficient staffing resources and a specificity of at least 80% for hospitals with limited staffing resources.

Model Update

Commonly, the development and validation datasets differ in proportion to outcome events, yielding poor calibration of the original model when applied to new data. Accordingly, adjusting the intercept of the original model to the validation sample can improve calibration [20]. In such a model update, the correction factor for the intercept is estimated in the validation dataset and should be added to the intercept of the original model when applying the model to new patients [22].

Decision Curve Analysis

The utility of the model for decision-making was evaluated using decision curve analysis [23-25]. In brief, decision curve analysis calculates a "net benefit" for a prediction model by comparing it to (at least) 2 default strategies: exposing either all patients or no patients to an intervention—in this study, medication reconciliation. The net benefit weighs the true positives against the false positives and is calculated using the following equation [23]:

Where:

- n is the total number of patients
- P_t is the threshold probability, which reflects the probability at which a health care provider would choose to perform a medication reconciliation to detect clinically relevant medication discrepancies.
- true positive count/n represents the proportion of patients correctly identified as needing a medication reconciliation.
- false positive count/n represents the proportion of patients incorrectly identified as needing a medication reconciliation
- $P_t/(1–P_t)$ is given a weighting factor based on the threshold probability P_t. It adjusts the impact of false positives by reflecting how tolerable the false positives are relative to the true positives.
    - For instance, a higher P_t—and hence higher $P_t/(1–P_t)$—increases the importance of false positives in estimating the net benefit of a certain strategy.
    - Conversely, when assuming a lower P_t, the importance of false positives decreases, putting more emphasis on identifying true positives.

The net benefit is plotted across a range of threshold probabilities, generating the "decision curve."

Comparison of the MED-REC Predictor With Existing Selection Algorithms

The MED-REC predictor was compared with two existing patient selection algorithms, as used in practice by ED pharmacists and physicians: (1) random selection and (2) patients aged ≥75 years and taking ≥5 drugs.

Statistical Analysis

Data analysis was performed using R (version 4.1.1; R Core Team). The R packages *givitiR*, *pROC* [26], and *dcurves* were used to assess calibration, discrimination, and net benefit, respectively. Descriptive statistics were presented as frequency with percentage for categorical data and as median with IQR for continuous data. Baseline characteristics of the 3 cohorts were compared by chi-square test for categorical data and by the Kruskal-Wallis test for continuous data.

No sample size calculation was performed for model development as we anticipated a large number of inclusions and a high number of discrepancies [19]. A sample size calculation was performed for the temporal validation to estimate the absolute agreement rate with a prespecified precision, defined as the maximum width of the associated 95% CI. In this calculation, we assumed an absolute agreement rate of 70% between the MED-RED predictor and the gold standard, to be estimated with a 95% CI that should have a maximal half-width of 5%. The required sample size was estimated at 333 patients, which also exceeded the minimum of 100 events and nonevents as suggested by the TRIPOD guidelines, assuming a prevalence of 35% as observed in the development dataset [20]. For the geographic validation, we intended to include a pragmatic sample of 100 patients, based on resources available at the external hospital.

Ethical Considerations

The study was approved by the Ethics Committee Research UZ/KU Leuven and AZ Sint-Jan Brugge Oostende (S60638). Written informed consent was obtained prior to inclusion from each patient or relative when the patient was deemed incapacitated. All study data were deidentified. Participation was voluntary and not financially compensated.

XSL•FO
RenderX

Results

Overview

In total, 824, 350, and 119 patients were included in the development, temporal validation, and geographic validation dataset, respectively. A flowchart of the study participants is presented in Figure 1. Patient characteristics, drug-related factors, and factors related to ED visits and the medication reconciliation process that were considered candidate variables for the MED-REC predictor are presented in Table 2. Data on the number of drugs before and after medication reconciliation and data on observed medication discrepancies are shown in Table 3. All students achieved the cutoff level of interrater reliability with a mean percent agreement of 96%.

Figure 1. Study flow diagram.

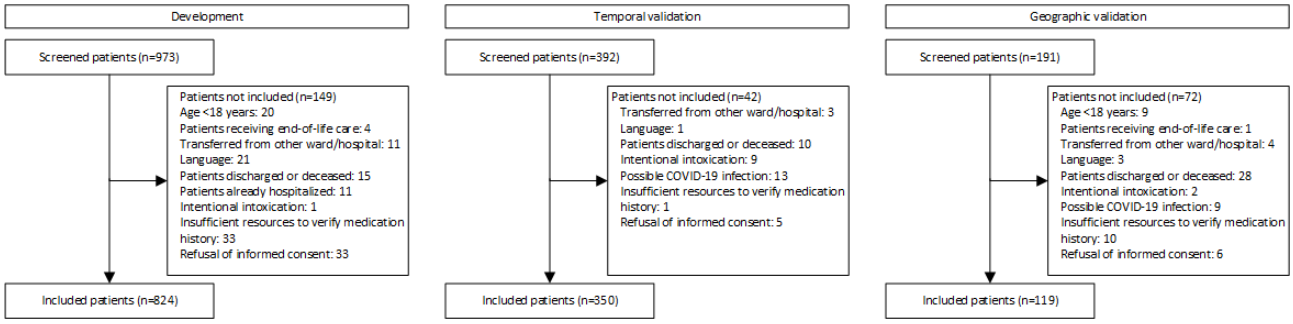

XSL•FO
RenderX

Table 2. Description of the development and external validation cohorts. Baseline characteristics were compared by the chi-square test for categorical data and by the Kruskal-Wallis test for continuous data.

	Development (n=824)	Temporal validation (n=350)	Geographic validation (n=119)	*P* value
Study period	2017-2019	2021-2022	2022	__[a]
Patient-related factors				
Sex, n (%)				
Female	385 (46.7)	199 (56.9)	66 (55.5)	.003
Age (years), median (IQR)	69 (55-80)	67 (55-79)	69 (55-78)	.91
Language, n (%)				.68
Dutch	810 (98.3)	347 (99.1)	118 (99.2)	
French	11 (1.3)	3 (0.9)	1 (0.8)	
English	3 (0.4)	0 (0)	0 (0)	
Residence before admission, n (%)				.37
Home	798 (95.8)	341 (97.4)	116 (97.5)	
Nursing home	29 (3.5)	9 (2.6)	2 (1.7)	
Other	6 (0.7)	0 (0)	1 (0.8)	
ESI[b]triage score, n (%)				<.001
1	1 (0.1)	5 (1.4)	0 (0)	
2	117 (14.2)	102 (29.1)	14 (11.8)	
3	175 (21.2)	116 (33.1)	91 (76.5)	
4	14 (21.7)	19 (5.4)	14 (11.8)	
5	3 (0.4)	1 (0.3)	0 (0)	
Unknown	75 (9.1)	107 (30.6)	0 (0)	
Transport to the ED[c], n (%)				<.001
Patient's own transport	706 (85.7)	243 (69.4)	76 (63.9)	
Ambulance	97 (11.8)	80 (22.9)	25 (21.0)	
Emergency physician vehicle transport	21 (2.5)	27 (7.7)	18 (15.1)	
Drug-related factors				
Number of medications reported by the ED physician per patient at time of ED visit, median (IQR)	6 (3-10)	5 (3-9)	6 (3-9)	.23
Number of medications reported by the ED physician per patient at time of ED visit by first-level ATC[d] group, median (IQR)				
ATC A drugs (alimentary tract and metabolism)	1 (0-2)	1 (0-3)	1 (0-2)	.74
ATC B (blood and blood forming agents)	1 (0-1)	1 (0-1)	0 (0-1)	.18
ATC C (cardiovascular system)	1 (0-3)	1 (0-2)	1 (0-2)	.88
ATC D (dermatologicals)	0 (0-0)	0 (0-0)	0 (0-0)	.08
ATC G (genito urinary system and sex hormones)	0 (0-0)	0 (0-0)	0 (0-0)	.34
ATC H (systemic hormonal preparations, excl. sex hormones and insulins)	0 (0-0)	0 (0-0)	0 (0-0)	.57
ATC J (anti-infectives for systemic use)	0 (0-0)	0 (0-0)	0 (0-0)	.004
ATC L (antineoplastic and immunomodulating agents)	0 (0-0)	0 (0-0)	0 (0-0)	.60
ATC M (musculo-skeletal system)	0 (0-0)	0 (0-0)	0 (0-0)	.04
ATC N (nervous system)	1 (0-2)	1 (0-2)	1 (0-2)	.28
ATC P (antiparasitic products, insecticides, and repellents)	0 (0-0)	0 (0-0)	0 (0-0)	.20
ATC R (respiratory system)	0 (0-0)	0 (0-0)	0 (0-0)	.72

XSL•FO
RenderX

	Development (n=824)	Temporal validation (n=350)	Geographic validation (n=119)	*P* value
ATC S (sensory organs)	0 (0-0)	0 (0-0)	0 (0-0)	.007
ATC V (various)	0 (0-0)	0 (0-0)	0 (0-0)	.39
ED visit related factors				
Specialty of the ED physician, n (%)				<.001
Emergency medicine	107 (13.0)	61 (17.4)	19 (16.0)	
Internal medicine	569 (69.0)	254 (72.6)	77 (64.7)	
Surgery	148 (18.0)	35 (10)	21 (17.6)	
Psychiatry	0 (0)	0 (0)	2 (1.7)	
Time of ED visit, n (%)				<.001
Morning	340 (41.3)	157 (44.9)	81 (68.0)	
Afternoon	137 (16.6)	33 (9.4)	16 (13.4)	
Evening	260 (31.5)	60 (17.1)	1 (0.8)	
Night	87 (10.5)	100 (28.6)	21 (17.6)	
Medication reconciliation-related factors, n (%)				
Availability of regional shared electronic medication record (yes)	111 (13.5)	80 (22.9)	46 (38.6)	<.001
Electronic validation of the medication history by the ED physician (yes)	155 (18.8)	63 (18.0)	31 (26.0)	.14
Outcome, n (%)				
At least 1 clinically relevant medication discrepancy, (yes)	287 (35)	131 (37)	58 (49)	.01

[a]Not applicable.

[b]ESI: Emergency Severity Index.

[c]ED: emergency department.

[d]ATC: Anatomical Therapeutic Chemical.

Table 3. The number of drugs before and after medication reconciliation and observed medication discrepancies in the development and external validation cohorts.

	Development (n=824)	Temporal validation (n=350)	Geographic validation (n=119)
Total number of drugs reported by the ED[a] physician, n	5486	2266	734
Number of drugs reported by the ED physician per patient, median (IQR)	6 (3-10)	5 (3-9)	6 (3-9)
Total number of drugs after medication reconciliation	6456	2614	926
Number of drugs per patient after medication reconciliation, median (IQR)	7 (4-11)	7 (4-10)	8 (4-10)
Total number of discrepancies, n	1604	801	451
Number of discrepancies per patient, median (IQR)	1 (0-3)	2 (1-3)	3 (2-5)
Total number of clinically relevant discrepancies, n	475	205	116
Number of clinically relevant discrepancies per patient, median (IQR)	0 (0-1)	0 (0-1)	1 (0-2)
At least 1 clinically relevant discrepancy (yes), n (%)	287 (35)	131 (37)	58 (49)

[a]ED: emergency department.

Model Development

Out of the 824 included patients, 287 (35%) had at least 1 clinically relevant discrepancy, with a total of 475 clinically relevant discrepancies.

After feature selection based on 26 initial variables, the final prediction model contained 8 variables (Table 4). The retained predictor variables were found in similar proportions across all datasets (Table 2). Figure 2A shows that our MED-REC predictor was well calibrated over a broad range of probabilities. The calibration slope was 0.72 and the intercept was 0.03. Discrimination was moderate with an AUROC of 0.66 (95% CI 0.62-0.70; Figure 2D). Classification measures are shown in Table 5. The Youden index, a sensitivity of at least 80%, and

XSL•FO
RenderX

a specificity of at least 80% were found at probability thresholds of 0.31, 0.25, and 0.45, respectively.

Table 4. Equation of the MED-REC predictor estimating the probability of having at least 1 clinically relevant discrepancy. The shape of the formula is [illegible], where P is the probability of having at least 1 clinically relevant discrepancy, c is the intercept, xi is the predictor variable and βi is the corresponding β coefficient. For each predictor variable, the β coefficient, odds ratio, 95% CI, and the *P* value are presented.

Variable	Coefficient	OR[a] (95% CI)	*P* value
Intercept	–1.82	0.16 (0.08 to 0.30)	<.001
Age (years)	0.012	1.01 (1.00 to 1.02)	.02
Residence of the patient before admission (reference category: home)			
Residence of the patient (nursing home)	–0.96	0.38 (0.15 to 0.90)	.03
Residence of the patient (other)	–0.79	0.45 (0.06 to 2.25)	.36
Number of drugs reported by the ED[b] physician	–0.094	0.91 (0.83 to 0.99)	.04
Number of “ATC A”[c] drugs reported by the ED physician (alimentary tract and metabolism)	0.276	1.32 (1.13 to 1.55)	<.001
Number of “ATC C” drugs reported by the ED physician (cardiovascular system)	0.167	1.18 (1.02 to 1.37)	.03
Number of “ATC N” drugs reported by the ED physician (nervous system)	0.206	1.23 (1.06 to 1.42)	.005
Number of “ATC P” drugs reported by the ED physician (antiparasitic products insecticides and repellents)	1.712	5.54 (1.07 to 41.4)	.05
Number of “ATC R” drugs reported by the ED physician (respiratory system)	0.269	1.31 (1.09 to 1.58)	.004

[a]OR: odds ratio.

[b]ED: emergency department.

[c]ATC: Anatomical Therapeutic Chemical.

XSL•FO
RenderX

Figure 2. Calibration plots for the MED-REC predictor in the (A) development, (B) temporal validation, and (C) and geographic validation dataset, as well as AUROC curves for the MED-REC predictor in the (D) development, (E) temporal validation, (F) and geographic validation dataset. The red line represents perfect calibration. The shaded area represents the 95% CI. AUROC: area under the receiver operating characteristic curve.

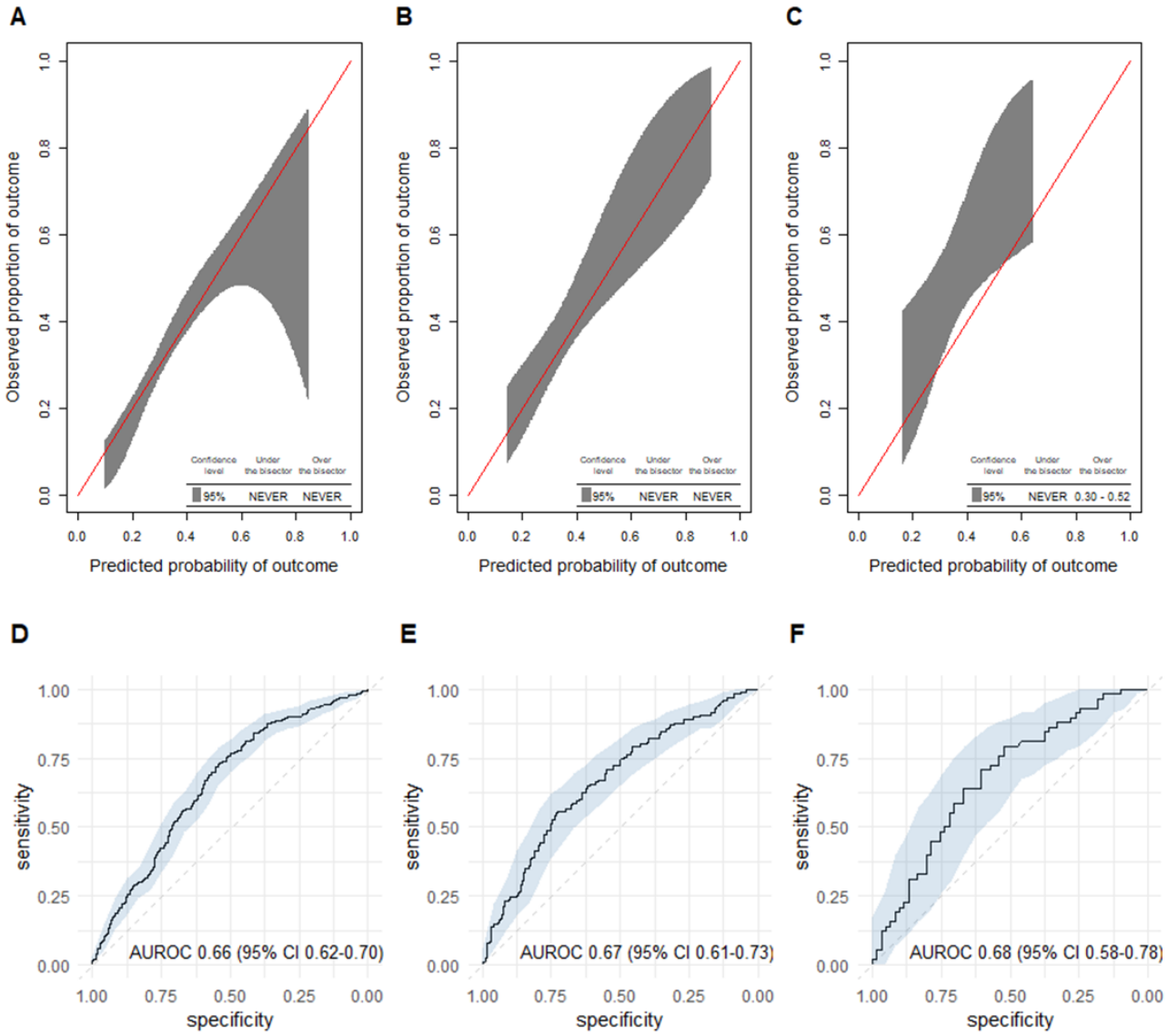

XSL•FO
RenderX

Table 5. Classification measures in each cohort using 3 different probability thresholds (0.25, 0.31, and 0.45). The output of the logistic regression model is a score between 0 and 1. To make a binary classification decision, a probability threshold is applied. If the calculated probability exceeds this threshold, it is categorized as having at least 1 clinically relevant discrepancy. The probability threshold is a crucial parameter that affects the trade-off between sensitivity and specificity.

Datasets and probability threshold	Specificity (%)	Sensitivity (%)	PPV[a] (%)	NPV[b] (%)	LR+[c]	LR–[d]	Alert rate (%)
Development dataset							
0.25	26	90	40	84	1.23	0.37	79
0.31	54	74	46	78	1.61	0.48	56
0.45	84	29	49	69	1.83	0.85	20
Temporal validation dataset							
0.25	26	89	42	80	1.20	0.42	80
0.31	51	73	47	76	1.48	0.54	58
0.45	88	25	55	66	2.04	0.85	17
Geographic validation dataset							
0.25	26	88	53	70	1.19	0.46	81
0.31	49	79	60	71	1.56	0.42	65
0.45	87	24	64	55	1.84	0.87	18

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]LR+: positive likelihood ratio.

[d]LR–: negative likelihood ratio.

Temporal Validation

Out of 350 included patients, 131 (37%) had at least 1 clinically relevant discrepancy, with a total of 205 clinically relevant discrepancies. Excellent calibration was found over the whole range of probabilities with a slope of 1.09 and an intercept of 0.18 (Figure 2B). Discrimination was moderate with an AUROC of 0.67 (95% CI 0.61-0.73), which was not different compared with the development dataset (Figure 2E). Classification measures are shown in Table 5.

Geographic Validation

Out of 119 included patients, 58 (49%) had at least 1 clinically relevant discrepancy, with a total of 116 clinically relevant discrepancies. The calibration curve is shown in Figure 2C. A slope of 1.35 and an intercept of 0.83 were found. The predictions underestimated the observed outcome probability of having at least 1 clinically relevant medication discrepancy for estimated probabilities between 0.3 and 0.52. Discrimination was moderate with an AUROC of 0.68 (95% CI 0.58-0.78), which was not different compared with the development dataset (Figure 2F). Classification measures are shown in Table 5.

Model Update

The incidence in the geographic validation dataset was 49% and the mean predicted risk was 35%. Adjustment of the intercept was performed and calibration and discrimination were reassessed. The correction factor was 0.58. The calibration plot of the updated model in the geographic validation dataset showed excellent calibration over the whole range of probabilities (Figure 3). Discrimination of the updated model did not differ from that of the original model (AUROC 0.68).

XSL•FO
RenderX

Figure 3. Calibration plot for the updated MED-REC predictor in the geographic validation dataset. The red line represents perfect calibration. The shaded area represents the 95% CI.

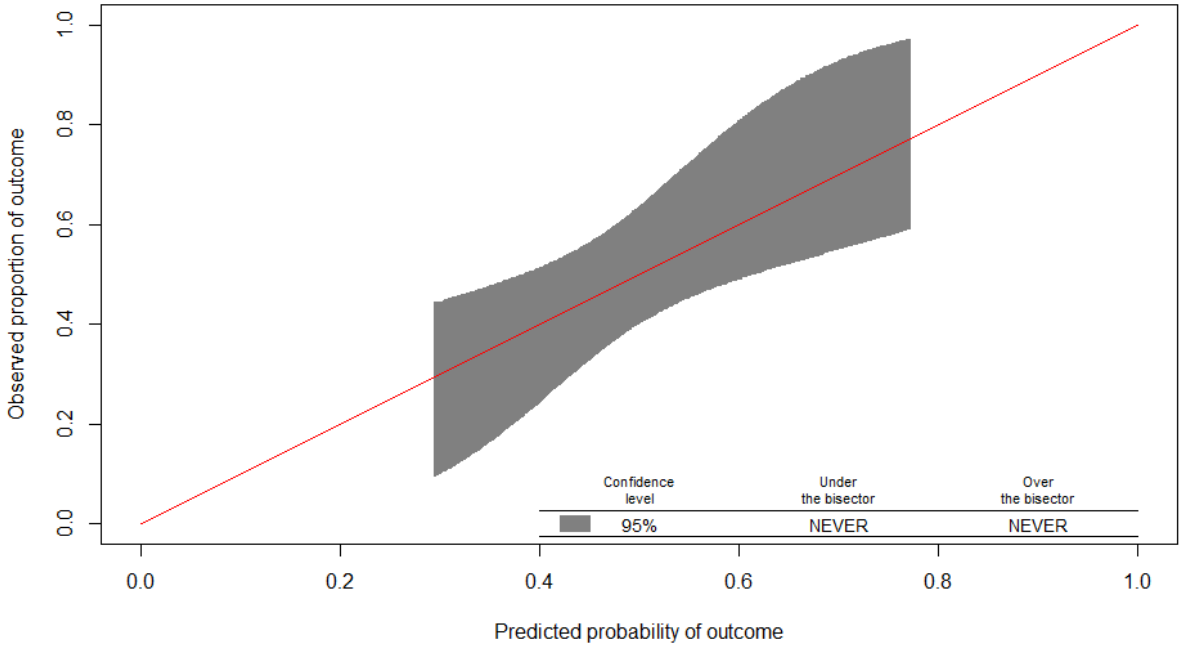

Decision Curve Analysis

The MED-REC predictor showed net benefit over the other strategies (treat all, treat none, and patients that are ≥75 years and taking ≥5 drugs) over a range of clinically reasonable threshold probabilities (20%-50%) in the development cohort (Figure 4). Decision curve analysis confirmed the net benefit of the MED-REC predictor over a similar range of threshold probabilities in the temporal validation cohort (Figure S1 in Multimedia Appendix 3). In the geographic validation cohort, the net benefit for the updated MED-REC predictor was demonstrated at higher threshold probabilities, starting from 30% (Figure S2 in Multimedia Appendix 3). The miscalibrated MED-REC predictor did not show a net benefit (Figure S3 in Multimedia Appendix 3).

Figure 4. Decision curve analysis for the MED-REC predictor in the development dataset. Red line (treat all): all patients receive medication reconciliation. Green line (treat none): none of the patients receives medication reconciliation. Blue line: selection of patients who are ≥75 years and take ≥5 drugs. Purple line: MED-REC predictor. The MED-REC predictor shows net benefit over the range of clinically reasonable threshold probabilities. Classifying a patient as having at least 1 clinically relevant medication discrepancy will trigger performing a comprehensive medication reconciliation. Accordingly, lower threshold probabilities are preferred if one is worried about missing clinically relevant discrepancies. Higher threshold probabilities are preferred if one is worried about additional costs or resources associated with medication reconciliation. For instance, a probability threshold of 20% implies that you expect to find at least 1 clinically relevant discrepancy in 1 out of 5 reconciled patients.

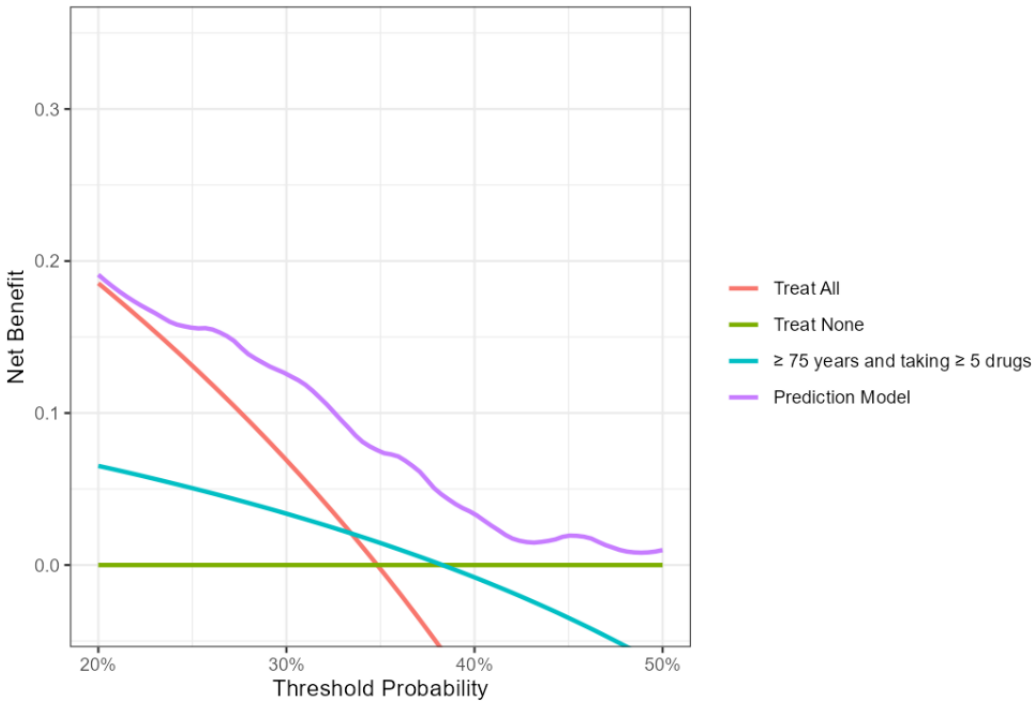

XSL•FO
RenderX

Comparison of the MED-REC Predictor With Existing Selection Algorithms

Random Selection

A comparison of the MED-REC predictor to random selection is presented in Figure 5.

Figure 5. Comparison of the (A) MED-REC predictor to (B) random selection. Calculations are based on the classification measures of the temporal validation. A probability threshold of 0.45 was used, considering limited staffing resources for medication reconciliation in our hospitals. Patients with at least 1 clinically relevant discrepancy are presented in yellow. Patients without clinically relevant discrepancies are presented in white. Medication reconciliation is performed for patients in green. (A) If the MED-REC predictor is applied to 100 patients, 17 will be identified as high risk, alerting the pharmacist to perform a medication reconciliation. For 9 of these patients (17×0.55), at least 1 clinically relevant medication discrepancy will be found. (B) When a medication reconciliation is performed for 17 randomly selected patients, at least 1 clinically relevant medication discrepancy will be found in 6 patients (17×0.37). Eleven patients, as opposed to 8 with the MED-REC predictor, would not have needed the pharmacy staff's intervention.

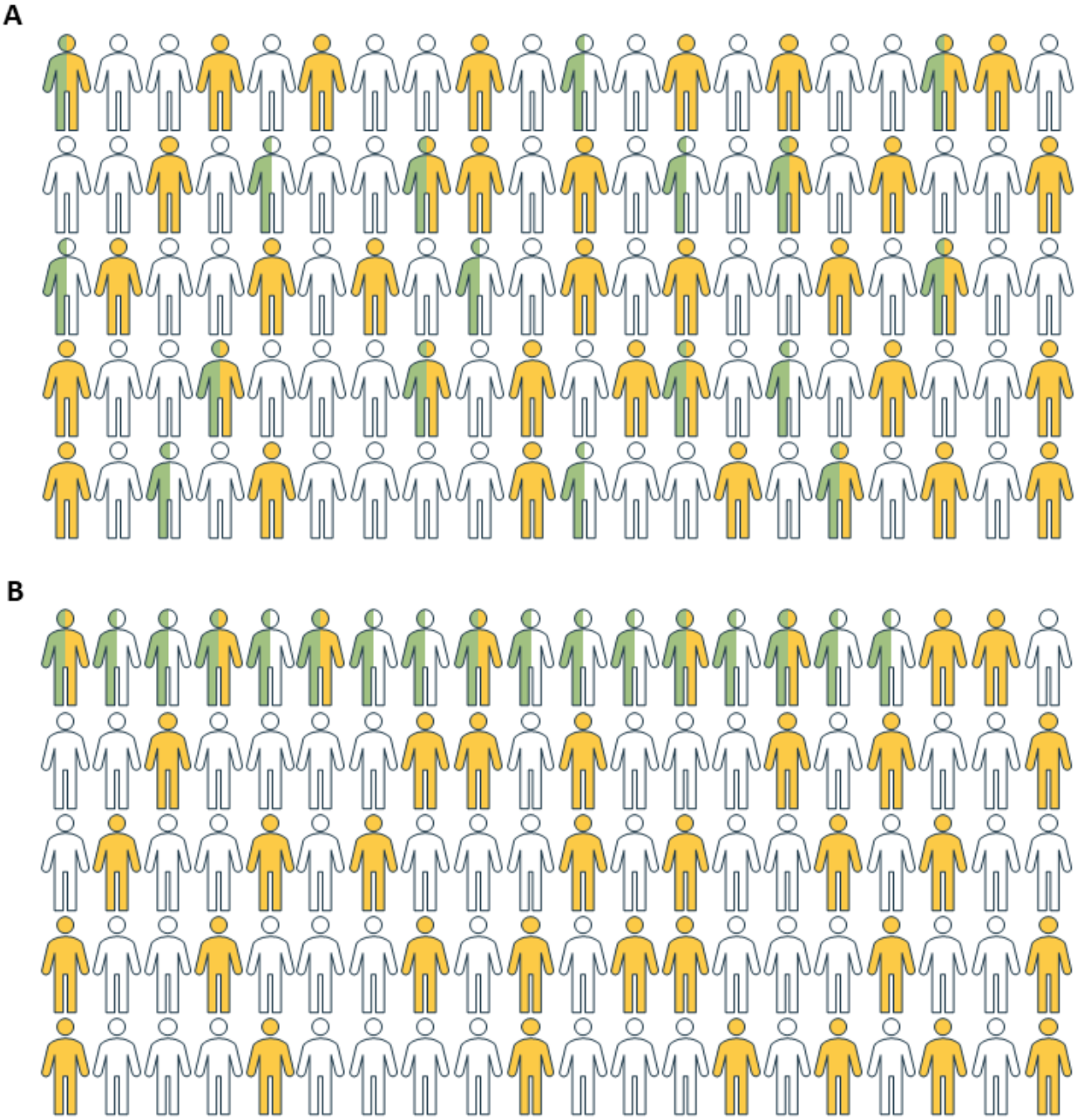

Patients Aged ≥75 Years and Taking ≥5 Drugs

The AUROC curves for this selection algorithm showed no discrimination across the 3 datasets. The AUROCs were 0.52 (95% CI 0.48-0.55), 0.58 (95% CI 0.53-0.63), and 0.54 (95% CI 0.46-0.63) for the development, temporal validation, and geographic validation dataset, respectively. The MED-REC predictor discriminated significantly better across the 3 datasets (P<.001, P=.002, and P=.003, respectively). Additionally, the net benefit analysis indicated that the MED-REC predictor provided the highest net benefit (Figure 4).

XSL•FO
RenderX

Discussion

Principal Findings

We developed and externally validated the MED-REC predictor, according to the current standards of predictive modeling [20]. The final model contains 8 variables that can be easily extracted from the EHR upon ED presentation. The MED-REC predictor identifies patients at risk for at least 1 clinically relevant discrepancy with moderate performance and clearly outperforms existing selection strategies. The clinical usefulness of the model was demonstrated through decision curve analysis, which showed net benefit for the MED-REC predictor compared with default strategies and typical selection criteria. While there is no question that conducting a comprehensive medication reconciliation for each hospital admission is of the utmost importance, the stark reality is that this objective is frequently untenable in clinical practice. Hence, our MED-REC predictor can be used to guide the rational use of limited resources at the ED and is especially useful in countries where resources are insufficient.

As an outcome variable, we opted for clinically relevant discrepancies, allowing us to focus on patients with the highest need for medication reconciliation. The current prediction model contrasts with our previously developed model, which screened for any medication discrepancy [19]. After applying the former model in clinical practice, we experienced its lack of clinical usefulness as the model identified many patients with nonclinically relevant discrepancies, for example, discrepancies concerning vitamin supplements.

We validated the MED-REC predictor in both a temporal and geographic validation dataset. Discrimination was moderate and was retained in both external validation datasets, demonstrating the robustness of our MED-REC predictor. The lack of excellent discrimination might be due to other unknown or unmeasured variables not included in our model. However, upon ED presentation, readily available patient information is limited. Calibration was excellent in the temporal validation dataset, whereas underestimation of the observed outcome by the model predictions was observed in the geographic validation dataset, due to a higher prevalence of the outcome in the geographic validation dataset. A simple model update improved calibration in this setting [22,27].

Net benefit was demonstrated over a range of clinically reasonable threshold probabilities (20%-50%) in both the development and temporal validation datasets. Notably, in the geographic validation dataset, the net benefit was observed starting from a higher probability threshold of 30%, which can be attributed to the higher prevalence of the outcome in this cohort (Figure S2 in Multimedia Appendix 3).

Implementation of the MED-REC predictor, a relatively simple prediction model based on readily available variables, is feasible in clinical practice. A precondition is the structural availability of the variables in the EHR in order for the MED-REC predictor to work optimally. The probability threshold can be easily customized to accommodate the available staffing resources. A lower probability threshold will lead to a higher alert rate and fewer false negative results, which might be interesting for hospitals with a large staff available to perform medication reconciliation. A higher probability threshold will be better suited for hospitals with limited staff, hence decreasing the false positive results. The decision curve analysis illustrates how the probability threshold impacts clinical decision-making (Figure 4). The threshold reflects the trade-off between the benefits of true positives and the potential harms of false positives. For instance, a probability threshold of 20% suggests a willingness to perform medication reconciliation on 5 patients to identify the outcome, at least 1 clinically relevant discrepancy, in one patient. This threshold implies that the benefit of detecting one true positive is deemed 4 times greater than the potential (financial) harm of performing unnecessary medication reconciliations [25].

Literature on medication reconciliation risk stratification tools is scarce. Moreover, most published models are not validated at all or are limited to discrimination measurement. Pippins et al [1] published the potential adverse drug event tool based on an observational study on general medical wards. An exploratory model was developed to predict the number of potential adverse drug events, but this model was not validated. Besides, automatization of their tool is limited because not all included predictors are structurally available in the EHR, for example, patient understanding of preadmission medications, and family members or caregivers as the source of preadmission medication information. Ebbens et al [16] developed and validated a risk prediction model to identify patients at risk for medication discrepancies at planned hospital admission. Retained variables were the number of drugs and the presence of cardiovascular and respiratory comorbidities. Good discrimination was found in the development cohort (AUROC 0.75) but was not retained in the external validation cohort (AUROC 0.54). Unfortunately, calibration was not assessed. Recently, Chu et al [15] reported a case series on the implementation of medication reconciliation prediction models in clinical practice. These models were initially developed to identify medication errors upon discharge as part of the MARQUIS2 [28] study and were adapted by removing variables that are not available upon ED presentation. However, the models were not developed based on predictive modeling standards nor was their performance assessed to account for this new setting. Thus, we cannot draw any conclusions regarding the robustness or generalizability of these models. Similar to our study, Damlien et al [17] developed and tested a prediction model to identify patients at risk for clinically relevant medication discrepancies upon ED admission. The authors included a limited sample of 276 patients over the course of 7 weeks, which was used for both development and validation of their model. Four variables were included, for example, sex, age ≥60 years, hospital admission in the last 12 months, and admission reason (malfunction, surgical, or cancer). Using a fixed cutoff value, the model classified patients as having a high versus low risk for discrepancies. An AUROC of 0.7 was found for the development cohort but was not reported for the validation cohort. Again, calibration was not assessed.

Our study has several strengths. First, we included large datasets and applied the current state-of-the-art of predictive modeling. Next to the model development, we performed extensive

XSL•FO
RenderX

external validation, including both temporal and geographic validation. As such, we have developed a robust prediction model to identify patients who will benefit from comprehensive medication reconciliation upon ED presentation. Second, we showed how a simple model update can be performed to increase the performance of the model in a distinct setting. Third, we only included candidate predictors that are widely and consistently available within the EHR and accessible upon ED visit. As a result, the MED-REC predictor can easily be implemented in clinical practice. Finally, we chose the presence of at least 1 clinically relevant discrepancy as the outcome variable, thereby increasing the clinical relevance of our model, and allowing us to focus on patients with the highest need for medication reconciliation.

Our study also has some limitations. First, clinical relevance of discrepancies was defined by an in-house expert panel and is therefore open to discussion. Currently, there is a notable lack of consensus regarding the classification of clinically relevant discrepancies. Leading international organizations such as the WHO and accreditation bodies like the Joint Commission International have yet to establish standardized guidelines. Therefore, we emphasize the need for international consensus or standardized indicators to guide the development of future risk prediction models. While our work represents a first step in defining clinically relevant medication discrepancies, external validation by additional experts or institutions (eg, the World Academic Council of Emergency Medicine) is warranted. Second, the requirement of informed consent introduced the possibility of volunteer bias, where individuals who choose to participate may differ systematically from those who do not, possibly affecting the generalizability of our results. However, only a minority of screened patients refused informed consent (ie, 3.4%, 1.3%, and 5% in the development, temporal, and geographic validation cohort, respectively). Third, our enrollment protocol, for example, starting with the patients who have been present the longest, may have introduced selection bias, as they might present with more complex pathology, which might reduce the generalizability of our results. We chose this approach to ensure the inclusion of patients presenting during nighttime or on weekends, aiming to reduce potential bias related to the timing of ED presentation. Although we recognize that this approach may introduce other forms of bias, we think it remains limited. For instance, very complex patients can be both transferred very quickly (eg, a patient having a major cardiovascular event that is transferred immediately to the catheterization laboratory) as well as stay in the ED for an extended duration (eg, a patient with prostatitis who clinically deteriorates and is challenging to transfer). Finally, additional variables, which might further increase the performance of our model, may be missing. Variables that can be assessed in future research include the presence of specific comorbidities and multimorbidity [29], ED crowding, the day of ED presentation and whether it is a holiday or weekend, the number of prior hospital visits and revisits, and the date of the last medication history validation. We did not include these variables in this study because they were either not structurally available in our EHR, not documented, or not prevalent. The presence of multiple comorbidities would likely increase the risk of medication discrepancies due to the complexity of the patient's medical regimen and the likelihood of polypharmacy. Nonetheless, the number of preadmission drugs and preadmission drugs per ATC code might have partially captured comorbidities in our study. High patient volumes and staff workload during periods of crowding might lead to rushed or incomplete medication reconciliation, increasing the risk of discrepancies.

Conclusions

The probability of having at least 1 clinically relevant medication discrepancy can be calculated by the MED-REC predictor with moderate performance. The MED-REC predictor outperforms medication reconciliation at random or based on other typical selection criteria, offering a guide for the rational use of limited resources at the ED. Depending on available resources, different probability thresholds can be applied to increase either the specificity or the sensitivity of the MED-REC predictor.

Acknowledgments

The authors would like to thank Ann-Sophie Jacob, Lies Braeken, Julie Veys, Fien Loosveldt, Ali Mehmet, Nadjia Amini, Sanne Nijs, Charlotte Lauweryns, Heleen Reumers, Tess Horemans, Elise Vanthomme, Hannah De Baere, Laute Leenders, Margaux Vansteenkiste, Lisa Berlanger, Anne-Sophie Depuydt, and Yurith Naeyaert, who are master's-level students from KU Leuven, for their contribution to this research.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured form for medication reconciliation.
[DOCX File , 22 KB - jmir_v26i1e55185_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Supplementary methods: collection of patient-, drug-, emergency department visit–, and medication reconciliation–related factors. [DOCX File , 18 KB - jmir_v26i1e55185_app2.docx]

Multimedia Appendix 3
Supplementary figures: decision curve analyses for the MED-REC predictor in the temporal and geographic validation dataset. [DOCX File , 162 KB - jmir_v26i1e55185_app3.docx]

References

1. Pippins JR, Gandhi TK, Hamann C, Ndumele CD, Labonville SA, Diedrichsen EK, et al. Classifying and predicting errors of inpatient medication reconciliation. J Gen Intern Med 2008;23(9):1414-1422 [FREE Full text] [doi: 10.1007/s11606-008-0687-9] [Medline: 18563493]
2. Cornish PL, Knowles SR, Marchesano R, Tam V, Shadowitz S, Juurlink DN, et al. Unintended medication discrepancies at the time of hospital admission. Arch Intern Med 2005;165(4):424-429. [doi: 10.1001/archinte.165.4.424] [Medline: 15738372]
3. Tam VC, Knowles SR, Cornish PL, Fine N, Marchesano R, Etchells EE. Frequency, type and clinical importance of medication history errors at admission to hospital: a systematic review. CMAJ 2005;173(5):510-515 [FREE Full text] [doi: 10.1503/cmaj.045311] [Medline: 16129874]
4. Wai A, Salib M, Aran S, Edwards J, Patanwala AE. Accuracy of patient self-administered medication history forms in the emergency department. Am J Emerg Med 2020;38(1):50-54. [doi: 10.1016/j.ajem.2019.04.016] [Medline: 31005394]
5. van Cauwenberghe L, van Kemseke S, Oudaert E, Pauwels S, Steurbaut S, van Laere S, et al. Interception of chronic medication discrepancies by the clinical pharmacist in the emergency department. Eur J Emerg Med 2023;30(1):7-14. [doi: 10.1097/MEJ.0000000000000961] [Medline: 35861664]
6. Etchells E. Admitting medication errors: five critical concepts. Qual Saf Health Care 2010;19(5):369-370. [doi: 10.1136/qshc.2010.041020] [Medline: 20977992]
7. Mueller SK, Sponsler KC, Kripalani S, Schnipper JL. Hospital-based medication reconciliation practices: a systematic review. Arch Intern Med 2012;172(14):1057-1069. [doi: 10.1001/archinternmed.2012.2246] [Medline: 22733210]
8. Kwan JL, Lo L, Sampson M, Shojania KG. Medication reconciliation during transitions of care as a patient safety strategy: a systematic review. Ann Intern Med 2013;158(5 Pt 2):397-403. [doi: 10.7326/0003-4819-158-5-201303051-00006] [Medline: 23460096]
9. Leguelinel-Blache G, Arnaud F, Bouvet S, Dubois F, Castelli C, Roux-Marson C, et al. Impact of admission medication reconciliation performed by clinical pharmacists on medication safety. Eur J Intern Med 2014;25(9):808-814. [doi: 10.1016/j.ejim.2014.09.012] [Medline: 25277510]
10. Pevnick JM, Shane R, Schnipper JL. The problem with medication reconciliation. BMJ Qual Saf 2016;25(9):726-730 [FREE Full text] [doi: 10.1136/bmjqs-2015-004734] [Medline: 26795914]
11. Sponsler KC, Neal EB, Kripalani S. Improving medication safety during hospital-based transitions of care. Cleve Clin J Med 2015;82(6):351-360 [FREE Full text] [doi: 10.3949/ccjm.82a.14025] [Medline: 26086494]
12. Stockton KR, Wickham ME, Lai S, Badke K, Dahri K, Villanyi D, et al. Incidence of clinically relevant medication errors in the era of electronically prepopulated medication reconciliation forms: a retrospective chart review. CMAJ Open 2017;5(2):E345-E353 [FREE Full text] [doi: 10.9778/cmajo.20170023] [Medline: 28476877]
13. Schnipper JL, Mixon A, Stein J, Wetterneck TB, Kaboli PJ, Mueller S, et al. Effects of a multifaceted medication reconciliation quality improvement intervention on patient safety: final results of the MARQUIS study. BMJ Qual Saf 2018;27(12):954-964. [doi: 10.1136/bmjqs-2018-008233] [Medline: 30126891]
14. Mekonnen AB, McLachlan AJ, Brien JE. Effectiveness of pharmacist-led medication reconciliation programmes on clinical outcomes at hospital transitions: a systematic review and meta-analysis. BMJ Open 2016;6(2):e010003 [FREE Full text] [doi: 10.1136/bmjopen-2015-010003] [Medline: 26908524]
15. Chu ES, El-Kareh R, Biondo A, Chang J, Hartman S, Huynh T, et al. Implementation of a medication reconciliation risk stratification tool integrated within an electronic health record: a case series of three academic medical centers. Healthc (Amst) 2022;10(4):100654. [doi: 10.1016/j.hjdsi.2022.100654] [Medline: 36126418]
16. Ebbens MM, Laar SAV, Wesselink EJ, Gombert-Handoko KB, van den Bemt PMLA. Prospective validation of a risk prediction model to identify high-risk patients for medication errors at hospital admission. Ann Pharmacother 2018;52(12):1211-1217. [doi: 10.1177/1060028018784905] [Medline: 29923419]
17. Damlien L, Davidsen N, Nilsen M, Godø A, Moger TA, Viktil KK. Drug safety at admission to emergency department: an innovative model for PRIOritizing patients for MEdication Reconciliation (PRIOMER). Eur J Emerg Med 2017;24(5):333-339. [doi: 10.1097/MEJ.0000000000000355] [Medline: 26657211]
18. Falconer N, Nand S, Liow D, Jackson A, Seddon M. Development of an electronic patient prioritization tool for clinical pharmacist interventions. Am J Health Syst Pharm 2014;71(4):311-320. [doi: 10.2146/ajhp130247] [Medline: 24481156]

XSL•FO
RenderX

19. de Winter S, Spriet I, Indevuyst C, Vanbrabant P, Desruelles D, Sabbe M, et al. Pharmacist- versus physician-acquired medication history: a prospective study at the emergency department. Qual Saf Health Care 2010;19(5):371-375. [doi: 10.1136/qshc.2009.035014] [Medline: 20595717]
20. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015;13(1):1 [FREE Full text] [doi: 10.1186/s12916-014-0241-z] [Medline: 25563062]
21. Hias J, van der Linden L, Spriet I, Vanbrabant P, Willems L, Tournoy J, et al. Predictors for unintentional medication reconciliation discrepancies in preadmission medication: a systematic review. Eur J Clin Pharmacol 2017;73(11):1355-1377. [doi: 10.1007/s00228-017-2308-1] [Medline: 28744584]
22. Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth 2009;56(3):194-201. [doi: 10.1007/s12630-009-9041-x] [Medline: 19247740]
23. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-574 [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]
24. Vickers AJ, van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6 [FREE Full text] [doi: 10.1136/bmj.i6] [Medline: 26810254]
25. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3:18 [FREE Full text] [doi: 10.1186/s41512-019-0064-7] [Medline: 31592444]
26. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
27. Steyerberg E. Clinical Prediction Models. New York, NY: Springer; 2009.
28. Schnipper JL, Reyes Nieva H, Mallouk M, Mixon A, Rennke S, Chu E, et al. Effects of a refined evidence-based toolkit and mentored implementation on medication reconciliation at 18 hospitals: results of the MARQUIS2 study. BMJ Qual Saf 2022;31(4):278-286 [FREE Full text] [doi: 10.1136/bmjqs-2020-012709] [Medline: 33927025]
29. Schnipper J. MARQUIS implementation manual: a guide for medication reconciliation quality improvement. Society of Hospital Medicine. 2011. URL: https://www.hospitalmedicine.org/globalassets/clinical-topics/clinical-pdf/shm_medication_reconciliation_guide.pdf [accessed 2024-08-09]

Abbreviations

ATC: Anatomical Therapeutic Chemical
AUROC: area under the receiver operating characteristic curve
AZ Sint-Jan Brugge-Oostende: General Hospital Sint-Jan Brugge-Oostende AV
c-index: concordance index
ED: emergency department
EHR: electronic health record
ESI: Emergency Severity Index
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UZ Leuven: University Hospitals Leuven
WHO: World Health Organization

Edited by A Coristine, A Mavragani; submitted 05.12.23; peer-reviewed by K Omae, E Bai; comments to author 23.07.24; revised version received 09.08.24; accepted 20.09.24; published 27.11.24.

Please cite as:
Van De Sijpe G, Gijsen M, Van der Linden L, Strouven S, Simons E, Martens E, Persan N, Grootaert V, Foulon V, Casteels M, Verelst S, Vanbrabant P, De Winter S, Spriet I
A Prediction Model to Identify Clinically Relevant Medication Discrepancies at the Emergency Department (MED-REC Predictor): Development and Validation Study
J Med Internet Res 2024;26:e55185
URL: https://www.jmir.org/2024/1/e55185
doi:10.2196/55185
PMID:

©Greet Van De Sijpe, Matthias Gijsen, Lorenz Van der Linden, Stephanie Strouven, Eline Simons, Emily Martens, Nele Persan, Veerle Grootaert, Veerle Foulon, Minne Casteels, Sandra Verelst, Peter Vanbrabant, Sabrina De Winter, Isabel Spriet. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study

Scott Silvey[1], MS; Jinze Liu[1], PhD

Department of Biostatistics, School of Public Health, Virginia Commonwealth University, Richmond, VA, United States

Corresponding Author:
Scott Silvey, MS
Department of Biostatistics
School of Public Health
Virginia Commonwealth University
830 East Main Street
Richmond, VA, 23219
United States
Phone: 1 4348256974
Email: silveys@vcu.edu

Abstract

Background: The performance of a classification algorithm eventually reaches a point of diminishing returns, where the additional sample added does not improve the results. Thus, there is a need to determine an optimal sample size that maximizes performance while accounting for computational burden or budgetary concerns.

Objective: This study aimed to determine optimal sample sizes and the relationships between sample size and dataset-level characteristics over a variety of binary classification algorithms.

Methods: A total of 16 large open-source datasets were collected, each containing a binary clinical outcome. Furthermore, 4 machine learning algorithms were assessed: XGBoost (XGB), random forest (RF), logistic regression (LR), and neural networks (NNs). For each dataset, the cross-validated area under the curve (AUC) was calculated at increasing sample sizes, and learning curves were fit. Sample sizes needed to reach the observed full–dataset AUC minus 2 points (0.02) were calculated from the fitted learning curves and compared across the datasets and algorithms. Dataset–level characteristics, minority class proportion, full–dataset AUC, number of features, type of features, and degree of nonlinearity were examined. Negative binomial regression models were used to quantify relationships between these characteristics and expected sample sizes within each algorithm. A total of 4 multivariable models were constructed, which selected the best-fitting combination of dataset–level characteristics.

Results: Among the 16 datasets (full-dataset sample sizes ranging from 70,000-1,000,000), median sample sizes were 9960 (XGB), 3404 (RF), 696 (LR), and 12,298 (NN) to reach AUC stability. For all 4 algorithms, more balanced classes (multiplier: 0.93-0.96 for a 1% increase in minority class proportion) were associated with decreased sample size. Other characteristics varied in importance across algorithms—in general, more features, weaker features, and more complex relationships between the predictors and the response increased expected sample sizes. In multivariable analysis, the top selected predictors were minority class proportion among all 4 algorithms assessed, full–dataset AUC (XGB, RF, and NN), and dataset nonlinearity (XGB, RF, and NN). For LR, the top predictors were minority class proportion, percentage of strong linear features, and number of features. Final multivariable sample size models had high goodness-of-fit, with dataset–level predictors explaining a majority (66.5%-84.5%) of the total deviance in the data among all 4 models.

Conclusions: The sample sizes needed to reach AUC stability among 4 popular classification algorithms vary by dataset and method and are associated with dataset–level characteristics that can be influenced or estimated before the start of a research study.

(J Med Internet Res 2024;26:e60231) doi:10.2196/60231

KEYWORDS

medical informatics; machine learning; sample size; research design; decision trees; classification algorithm; clinical research; learning-curve analysis; analysis; analyses; guidelines; ML; decision making; algorithm; curve analysis; dataset

XSL•FO
RenderX

Introduction

Background

Machine learning (ML) is becoming increasingly popular within the domain of health care data analysis and clinical decision-making [1]. The lack of a fixed model specification and distributional assumptions allows for these methods to learn complex relationships that are not necessarily linear in nature, such as high-order interactions and polynomial effects. Due to this, most popular machine-learning algorithms require much larger sample sizes than traditional statistical methods [2]. However, exact amounts are not clear, and there are many different ML algorithms, each containing its own limitations and properties [3]. Furthermore, in traditional statistical analysis, we can often analytically derive equations that measure how the sample size needed to detect a certain prespecified effect will behave under certain assumptions [4]. Due to the data-driven and algorithmic nature of ML methods, which rely on computational approaches rather than statistical theory to capture relationships, an empirical approach is necessary in order to understand the behavior of these methods under varying conditions.

It is known that, for any given dataset, there is a point where adding additional samples will not increase the performance metrics of the model considerably [5]. Thus, it becomes important to collect enough data to optimize these metrics while also accounting for this performance ceiling and the budgetary or computational concerns that may arise when collecting substantial amounts of unnecessary data.

Another reason for the difficulty in selecting a proper sample size when applying ML is the lack of a true end point or common metric of interest. As discussed previously, the traditional target for sample size determination methods is the statistical power to detect a certain effect size [4]. In ML, since predictive performance rather than parameter estimation is usually of interest, this end point becomes unclear. A commonly used metric of predictive performance is prediction accuracy, defined as the proportion of correct classifications made [6]. However, the prediction accuracy is related to the distribution of the outcome; for a rare event, accuracy can be high even with a completely noninformative model [7]. As a result, a fairer performance metric is the area under the receiver operating characteristic curve (area under the curve [AUC]), which evaluates model predictions over a range of probability thresholds from 0 to 1 [8]. AUC is widely used to evaluate the performance of an ML algorithm and has several desirable properties. First is interpretability—a higher AUC indicates a higher degree of separability, and an AUC of 0.5 implies a completely random prediction, while an AUC of 1.0 indicates perfect classification. A second desirable property of AUC is insensitivity to the proportion of cases versus controls in the dataset [9]; because the entire range of probability thresholds is considered, AUC can also be considered the "average" sensitivity (true-positive rate) over all possible values of specificity (true-negative rate). While the AUC is commonly used to evaluate the initial performance of an ML algorithm, other metrics, such as calibration [10], may also be preferred once modeling reaches later stages. However, it is necessary to ensure that a trained ML model can first make stable predictions before assessing further metrics such as calibration or threshold-selection.

Related Works

The concept of empirically estimating the performance of a classification algorithm as the training set size increases has been widely explored in a variety of different settings. This is typically done by creating a "learning curve," measuring a metric (such as classification accuracy) as a function of sample size [11]. Perlich et al [12] compared logistic regression (LR) approaches versus decision-tree–based approaches, demonstrating that LR often outperforms tree induction in small samples, but decision trees excel as the sample size becomes large. Mukherjee et al [13] developed a method to assess the error rate of a classifier as a function of sample size using an inverse power-law model. Their method was introduced in the context of DNA microarray data, which often contains a large amount of features and limited access to samples due to cost restraints. Figueroa et al [14] modified the original learning curve fitting process by using nonlinear weighted least squares to favor future predictions, using 3 moderately sized datasets to demonstrate their algorithm. Provost et al [15] used learning curves and efficient progressive sampling to show that classification algorithms eventually converge to a stable accuracy with increasing sample size, mainly focusing on the methodology of the sampling scheme. More recently, van der Ploeg et al [16] used several clinical datasets and a simulation-based approach to show that modern classification algorithms such as neural networks (NNs) and random forest (RF) require at least 200 events per variable to reach a stable AUC. Richter and Khoshgoftaar [17] experimented with learning curves on biomedical big data with limited labels and heavy class imbalance, using 1% of the full dataset AUC as their stopping rule. Because the cost of labeling certain types of data is expensive, it is important to maximize the quality of the data while minimizing costs. They found that a semisupervised approach and pseudolabeled data generated from a small amount of actual data could accurately predict future performance.

Study Aim

Previous contributions have focused on the methodology of learning curve fitting or estimating future performance from an already-collected sample. In those that have examined similar end points (ie, AUC plateau or stability over a variety of algorithms), the number of real-world datasets included was small, modern gradient boosting techniques (XGBoost [XGB], etc) were not examined because they had not yet been developed, and the impact of dataset–level characteristics on sample sizes was not extensively studied [15,16]. Previous literature has also mostly used small datasets in the context of -omics type data. In general, a focused clinical study often contains fewer features, a wide variety of variable types (ie, numeric, categorical, ordinal, etc), and fewer correlated features than -omics data [18]. This study aims to develop algorithm-specific sample size guidelines using dataset–level variables that can be estimated or manipulated by researchers before any data has been collected, analogous to a sample size

XSL•FO
RenderX

calculation performed in a traditional power analysis. The focus of these guidelines was on the stable internal validity of each method, which is typically the first benchmark used to assess performance when attempting to develop a predictive model. We examined 4 popular binary classification algorithms in the context of clinical research, where the aim is to predict a health-related outcome such as a disease state or event. The contributions of this study include a learning curve analysis of 16 real-world datasets, an examination of modern gradient-boosting methods (XGB) within this analysis, and concrete sample size guidelines based on dataset-level characteristics.

Methods

Dataset Description

We have collected 16 public-access clinical datasets ranging from sample sizes of 70,000-1,000,000. A detailed description of dataset sources, variables included, and outcomes can be found in Multimedia Appendix 1. It should be noted that 8 of these 16 datasets were artificially created from smaller real-life datasets using Bayesian Network Generation, and their details have been previously discussed [19]. All datasets contained a single binary outcome, such as a disease state, with a combination of continuous numeric, discrete numeric, or binary predictors. Continuous numeric features were considered to have at least 10 unique values.

A detailed description of specific data preprocessing steps can be found in Multimedia Appendix 1. In summary, nominal variables were converted to binary variables based on arbitrary binning rules, and variables containing text values (ie, “gender”: male vs female) were also converted to binary variables. Missing data was present in 3 datasets only (CDC Heart Disease [2022], Diabetes130, and COVID-19), although the amount of missingness was quite low among these sets (1.3%, <1%, and <1%, respectively). Without knowing additional information regarding the nature of these missing values, we considered them missing completely at random (MCAR) and performed mean imputation [20].

Classification Algorithms

We examined the following binary classifiers on each dataset: Logistic Regression (LR) [21], Random Forest (RF) [22], XGBoost (XGB) [23], and Neural Networks (NNs) [24]. These algorithms were selected due to their widespread and popular use in clinical data analysis [25]. We performed LR by fitting a multivariable model using all predictors without any variable selection or regularization methods. For the RF and XGB algorithms, hyperparameters were left at their default values, which can be found in the R documentation [22,23]. For NNs, we used the R (R Foundation for Statistical Computing) package *h2o* [26] to perform our analyses; we considered one hidden layer with 20 units and 10 epochs of data training. The type of NN used by “h2o” is a multilayer feedforward artificial NN, also known as multilayer perceptron (MLP). The activation function used in the hidden layers was the default linear rectifier, with softmax activation in the final output nodes for probability estimation and classification. Other NN hyperparameters were again left at their default values, which can be found in the documentation [26].

Learning Curve Approach

From the 16 datasets studied, we evaluated the cross-validated area under the curve (CV-AUC) as a function of increasing sample size. Cross-validation is a method of assessing the internal validity of a classifier; it works by splitting the entire training dataset into *k* folds and fitting *k* models, with a single fold left out for evaluation in each model [27]. The final cross-validated metric (in our case, AUC) is calculated by taking the average performance over all *k* of the held-out folds. As a result, the entire training dataset is used to generate an estimate of out-of-sample performance. The learning curve approach is detailed below:

1. Create a list of proposed training set sizes.
2. At each point in the sample size interval, randomly sample 10 subdatasets of size n from the full dataset.
3. In each of the 10 subdatasets, estimate the (5-fold, outcome-stratified) CV-AUC on the proposed algorithm of choice. Average the ten CV-AUC values to generate an estimate of out-of-sample performance at a given n.
4. Repeat at the next n in the list.

For the first step, the training set size list usually consisted of 10 evenly spaced points ranging from n=500 to n=50,000, but if stability was not reached by n=50,000, the end point was extended. For LR, the final n was lower, as the AUC from these models typically became stable much earlier than more complex ML algorithms. A full description of the sample size intervals used for each dataset and each algorithm can be found in Multimedia Appendix 1 (Table S1). Stability was defined as the smallest n where the CV-AUC was within 2 points (0.02) of the observed full–dataset AUC. For example, if the full–dataset AUC was 0.85, we would obtain the smallest n where a CV-AUC of 0.83 was first surpassed. The full–dataset AUC for each classification algorithm was calculated using 5-fold stratified cross-validation (CV) on the entire dataset. We chose this stopping point of 0.02 because, although arbitrary, we believed that it provided the most reasonable trade-off between high performance and computational burden. Specifically, as can be seen visually from figures in the Learning Curve Results section, this stopping point typically marks the beginning of the “point of diminishing returns,” where the power law curves begin to plateau, and the amount of additional sample needed to make further improvements increases exponentially. It is important to note that once the learning curve equations are estimated, this choice of stopping point can be freely altered. Therefore, although we report 0.02 in this study as our stopping rule of interest, the final equations derived below can be re-estimated with any user-specified stopping point (for example, 0.01 or 0.05). We present sample size results using alternative thresholds of 0.01 and 0.05 AUC points from the full–dataset AUC in Multimedia Appendix 1.

Once the raw data was generated, estimated learning curves were fit using nonlinear least squares optimization [28], following the power law equation: $AUC_{(n)} = an^b+c$, where a and b were estimated, and c was either fixed to be the

XSL•FO
RenderX

full–dataset AUC or was also estimated, depending on the quality of the fit. For some datasets and algorithms, the power law function did not fit the data well. These were typically scenarios where the dataset required a relatively larger sample size to become stable. In these scenarios, we instead fit the learning curves using a logarithmic function, $AUC_{(n)}=\beta_0+\beta_1*\log(n)$, where β_0 and β_1 were estimated using ordinary least-squares [29].

Sample Size Determination and Guidelines

Following the learning curve analysis of the 4 selected algorithms on our datasets, we examined the effects of 6 dataset–level factors on the sample sizes needed for AUC stability. These included minority class proportion (maximum value of 50%, indicating no class imbalance), separability (defined as the full–dataset AUC itself), the total number of features, the percentage of features that were continuous (versus binary or discrete numeric), the percentage of "core linear" features, and "dataset nonlinearity." Core linear features were determined by adding an L1 (LASSO [least absolute shrinkage and selection operator]) penalty to the LR model for each full dataset [30]. The percentage of variables that did not shrink to zero when this penalty was added were defined as core linear features. Dataset nonlinearity was a rough measure of the degree of nonlinear or interactive relationships between the predictors and the outcomes that were present in the data. This was defined as the point difference in the full–dataset AUC when using a complex algorithm (XGB) compared with LR. For example, if LR yielded a full–dataset AUC of 0.90 and XGB yielded a full–dataset AUC of 0.95, the dataset nonlinearity would be calculated as 5.0. For the purpose of calculating these values, XGB hyperparameters were left at their default values [26].

Within the context of each algorithm, the relationship between these dataset–level variables and the n required for AUC stability was examined. Since the estimated sample sizes were discrete and right-skewed numeric values, we used negative binomial regression models [31] to quantify the strength and significance of each dataset characteristic on predicted sample sizes, which produce coefficients in terms of log-expected counts. Then, in multivariable negative binomial regression models for each algorithm, we selected up to 3 dataset–level predictors that together minimized the Akaike Information Criterion, which evaluates how well the model fits the data while penalizing for the number of parameters estimated [32]. A maximum of 3 predictors per model were considered in order to avoid potential overfitting.

We also calculated adjusted deviance-based pseudo-R^2 statistics [33,34], which further quantified each model's goodness-of-fit and proportion of deviance explained by the predictors. The final model equations were reported and discussed for each algorithm, and visualizations of the model predictions at varying levels of each dataset-level characteristic were generated. Statistical significance was set to α=.05 for all hypothesis tests considered, and RStudio (version 4.2.3; R Foundation for Statistical Computing) was used for all analyses.

Results

A workflow and overview of the study's aims and end points can be seen in Figure 1.

We gathered 16 datasets with sample sizes ranging from 70,000 to 1,000,000 (Table 1). Out of the 4 classification algorithms examined, XGB performed the best or tied for the best performance on 14/16 (87.5%) datasets, while RF performed the best on two. Full dataset AUCs (separability) ranged from 0.608-0.979 (XGB), 0.609-0.976 (RF), 0.596-0.949 (LR), and 0.603-0.974 (NN; Table 2). As expected, LR models generally performed the worst, with full–dataset AUCs that were 0.028 points lower on average compared with XGB.

XSL•FO
RenderX

Figure 1. Flowchart of study aims and deliverables. AUC: area under the curve.

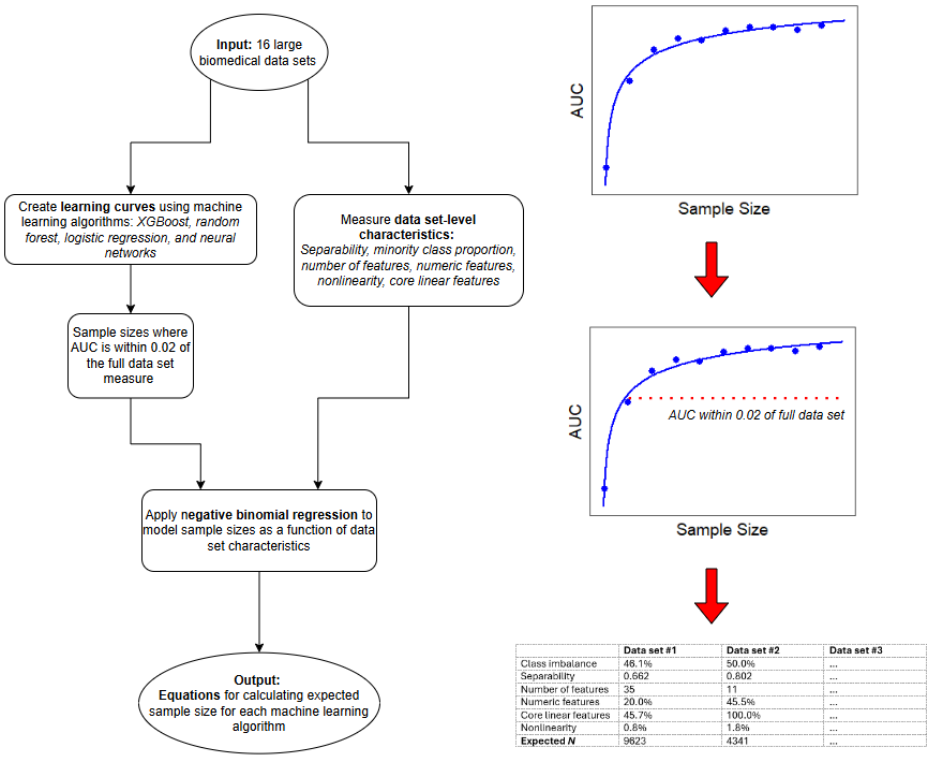

Table 1. Dataset–level characteristics.

Dataset name	Source	Full dataset size, n	Minority class proportion (%)	Features, n	Core linear features (%)	Continuous features (%)	Nonlinearity
Cardio	OpenML	70,000	50	11	100	45.5	1.8
Diabetes130	OpenML	101,766	46.1	35	45.7	20	0.8
NoShow	OpenML	110,527	20.2	8	25	12.5	1.2
BreastTumor	OpenML	116,640	34.6	9	44.4	33.3	9.5
Diabetes	UCI	253,680	13.9	21	85.7	19	0.7
COVID-19	OpenML	263,007	39	16	37.5	6.25	2.2
LOS	OpenML	318,438	2.1	11	36.4	27.3	1.9
CDC Heart Disease (2020)	Kaggle	319,795	4.4	17	76.5	23.5	0.5
CDC Heart Disease (2022)	Kaggle	394,509	8.6	39	59	15.4	0.6
Heart	OpenML	1,000,000	44.4	13	92.3	46.2	1.6
Hepatitis	OpenML	1,000,000	20.8	19	94.7	31.6	4.0
Lymph	OpenML	1,000,000	45.7	18	83.3	5.6	2.3
Pharynx	OpenML	1,000,000	25.6	11	72.2	18.2	1.5
Cholesterol	OpenML	1,000,000	16.5	13	61.5	30.8	6.7
Dermatology	OpenML	1,000,000	13.2	33	84.8	3	3.6
PBC	OpenML	1,000,000	17.8	18	83.3	55.6	6.2

XSL•FO
RenderX

Table 2. Full–dataset AUC (separability) for each algorithm.

Dataset name	XGB[a]	RF[b]	LR[c]	NN[d]
Cardio	0.802	0.796	0.784	0.795
Diabetes130	0.662	0.661	0.654	0.661
NoShow	0.608	0.609	0.596	0.603
BreastTumor	0.777	0.780	0.682	0.730
Diabetes	0.829	0.822	0.822	0.826
COVID-19	0.664	0.661	0.642	0.661
LOS	0.917	0.915	0.898	0.901
CDC Heart Disease (2020)	0.815	0.810	0.810	0.809
CDC Heart Disease (2022)	0.815	0.801	0.809	0.801
Heart	0.965	0.963	0.949	0.962
Hepatitis	0.979	0.976	0.939	0.974
Lymph	0.957	0.957	0.934	0.956
Pharynx	0.858	0.858	0.843	0.856
Cholesterol	0.736	0.728	0.669	0.714
Dermatology	0.859	0.857	0.823	0.852
PBC	0.850	0.850	0.788	0.823

[a]XGB: XGBoost.

[b]RF: random forest.

[c]LR: logistic regression.

[d]NN: neural network.

Learning Curve Results

Learning curves were fit to the 16 collected datasets. Table 3 and Figure 2 contain a full summary and visualization of estimated sample sizes across each classification algorithm and dataset. NNs required the largest sample sizes to reach stability and also had the most variability among the datasets (median 12,298, range 1824-180,835). LR required the smallest sample size to reach stability and also was the least variable (median 696, range 204-6798). XGB required approximately 3 times the sample size compared with RF, but the range of estimated sample sizes generated from RF models was nearly twice as wide (Table 2). Figure 3 shows the fitted learning curves for each algorithm generated within each dataset, with a marker indicating the earliest sample size where the CV-AUC was within 2 points of the full–dataset AUC.

XSL•FO
RenderX

Table 3. Sample sizes are needed to reach AUC stability from the learning curve analysis.

Dataset name	XGB[a]	RF[b]	LR[c]	NN[d]
Cardio, n	4341	1476	363	4349
Diabetes130, n	9623	3544	1822	16,823
NoShow, n	12,114	2241	742	8084
BreastTumor, n	19,668	17,383	558	28,424
Diabetes, n	9306	2261	1140	8556
COVID-19, n	7026	4750	543	4241
LOS, n	18,239	15,381	2555	14,085
CDC Heart Disease (2020), n	15,177	4995	2243	10,510
CDC Heart Disease (2022), n	30,534	16,355	6768	25,120
Heart, n	960	250	204	1824
Hepatitis, n	3513	3265	425	15,302
Lymph, n	1409	1992	276	4470
Pharynx, n	10,296	2488	317	5260
Cholesterol, n	65,556	140,499	1368	180,835
Dermatology, n	7979	3103	1696	47,489
PBC, n	31,897	71,194	650	53,453
Median (Range)	9960 (960-65,556)	3404 (250-140,499)	696 (204-6798)	12,298 (1824-180,835)
Mean (SD), Log-Transform	9.16 (1.11)	8.57 (1.55)	6.75 (0.96)	9.47 (1.17)

[a]XGB: XGBoost.

[b]RF: random forest.

[c]LR: logistic regression.

[d]NN: neural network.

Figure 2. Visualization of expected sample sizes calculated from the learning-curve analysis of 16 data sets. LR: Logistic Regression. NN: Neural Networks. RF: Random Forest. XGB: XGBoost.

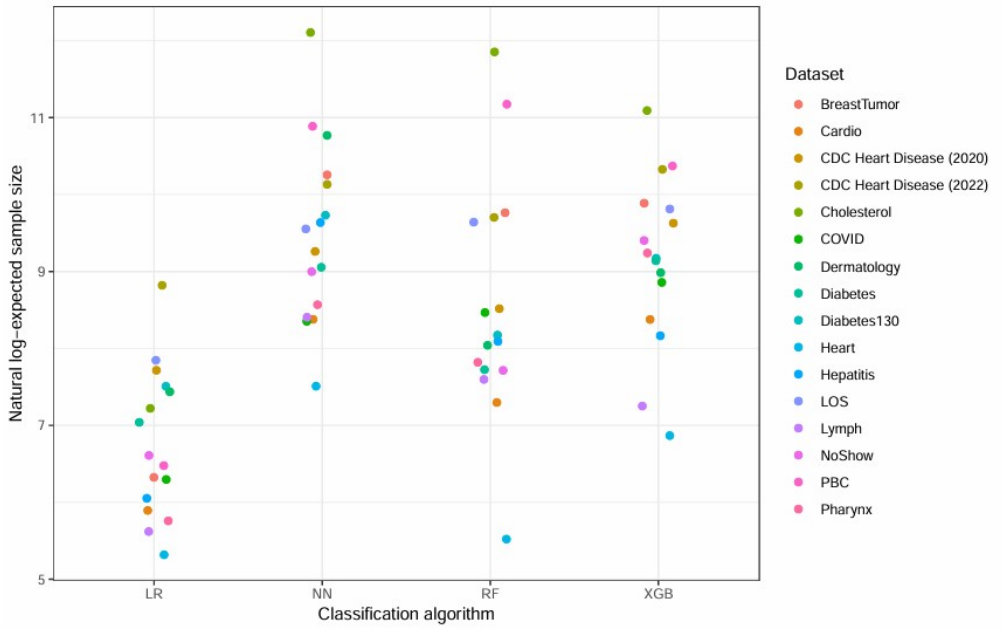

XSL•FO
RenderX

Figure 3. Fitted learning curves for 16 data sets across 4 classification algorithms. Different colors were chosen for different algorithms Each black "X" represents the point where the AUC at size n first comes within 2-points (or, 0.02) of the asymptotic (ie full-data set) AUC. LR: Logistic Regression. NN: Neural Networks. RF: Random Forest. XGB: XGBoost. AUC: Area-under the receiver operating characteristic curve.

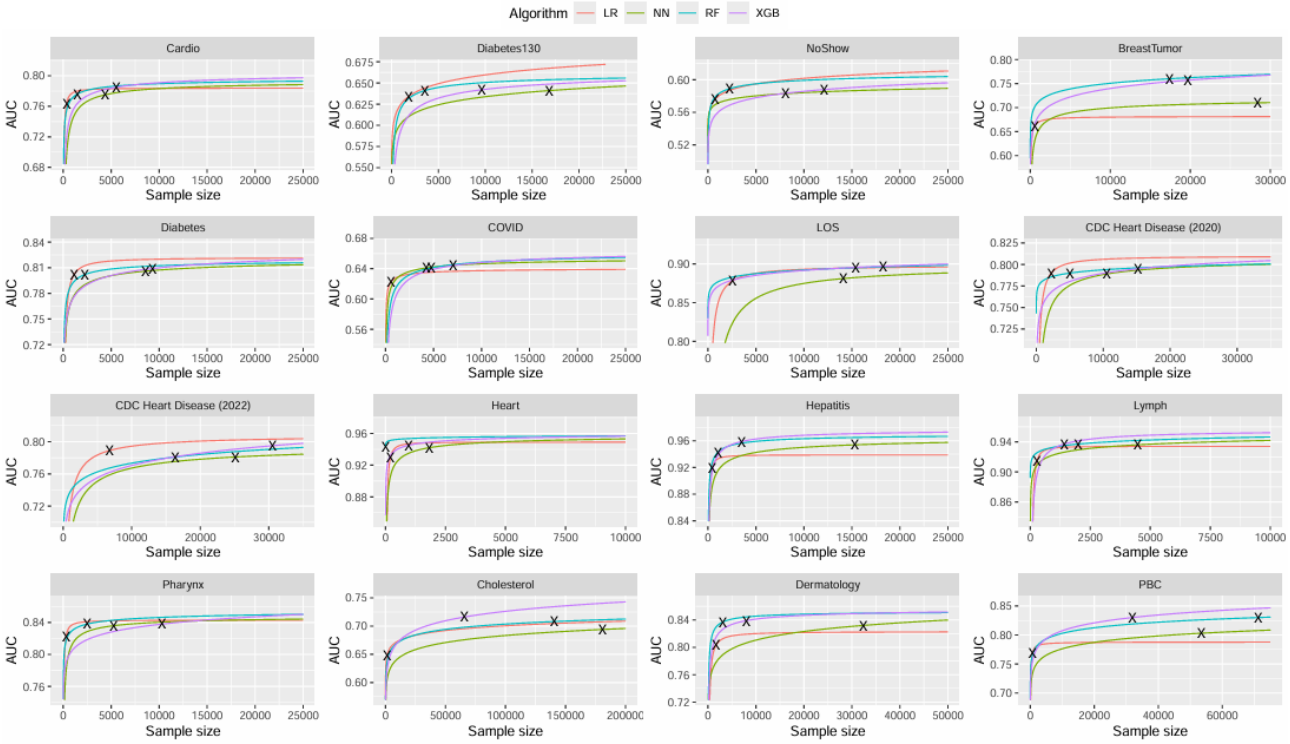

Dataset–Level Characteristics

Dataset–level characteristics were examined (Table 1). The average minority class percentage was 25.18% (SD 15.93%), and the average number of features was 18 (SD 9). The average percentage of continuous numeric features was 24.6% (SD 15.31%), and the average percentage of core linear features was 67.69% (SD 23.73%). The median dataset nonlinearity was 1.85 (range 0.50-9.50). Most datasets (13/16, 81.3%) had nonlinearity values under 5.0. Thus, for the purpose of model fitting, this was converted into a binary variable indicating either "high" (≥) or "low" (<5) nonlinearity. Scatterplots examining the visual relationships between log-expected sample sizes and each dataset–level characteristic can be found in Figure 4.

Negative binomial regression models were fitted, examining the individual associations between each of the dataset–level characteristics and predicted sample sizes (Table 4). In these models, separability (full–dataset AUC) was multiplied by 100 for easier interpretation. For example, an AUC of 0.80 was entered as 80.0 in the models. For XGB, minority class proportion and separability were both inversely related to sample size; for every 1-unit increase in separability (where 50.0 was the baseline value), estimated sample sizes were affected by a multiplier of 0.955 (P=.02). For every 1% increase in minority class proportion, estimated sample sizes were affected by a multiplier of 0.959 (P<.001). In datasets with high (≥5.0) values of nonlinearity, estimated sample sizes were affected by a multiplier of 3.888 (P=.005). In the RF analyses, results were similar for minority class proportion (0.931× multiplier for 1% increase, P<.001), separability (0.939× multiplier for each 1-unit increase over 50.0, P=.047), and nonlinearity (15.984× multiplier for those with high values, P<.001). However, the percentage of continuous numeric features (1.065× multiplier for every 1% increase, P=.003) was also individually statistically significant. For LR, minority class proportion (0.963× multiplier for every 1% increase, P=.001), the number of features (1.056× multiplier for each additional feature, P=.006), the percentage of core features (0.982× multiplier for 1% increase, P=.046), and the percentage of continuous numeric features (0.971× multiplier for 1% increase, P=.04) were significantly associated with sample size. Again, a more balanced ratio of classes reduced the needed sample size, while more features increased the sample size. However, a higher percentage of core linear features and a higher percentage of continuous numeric features lowered the sample size. Finally, for NNs, results were similar to XGB; minority class proportion (0.953× multiplier for 1% increase, P=.003), full–dataset AUC (0.950× multiplier for each 1% increase over 50.0, P=.03), and nonlinearity (6.85× multiplier for high values, P<.001) were all individually statistically significant.

XSL•FO
RenderX

Figure 4. Relationships between each data set–level characteristic and expected sample sizes. The y-axis represents the natural-log transformed sample size values, while each x-axis represents varying levels of each data set–level characteristic. Separability multiplied by 100. All values representing proportions were multiplied by 100 so that 0 indicates 0% and 100 indicates 100%. Nonlinearity "low": <5, "high": ≥5. LR: Logistic Regression. NN: Neural Networks. RF: Random Forest. XGB: XGBoost.

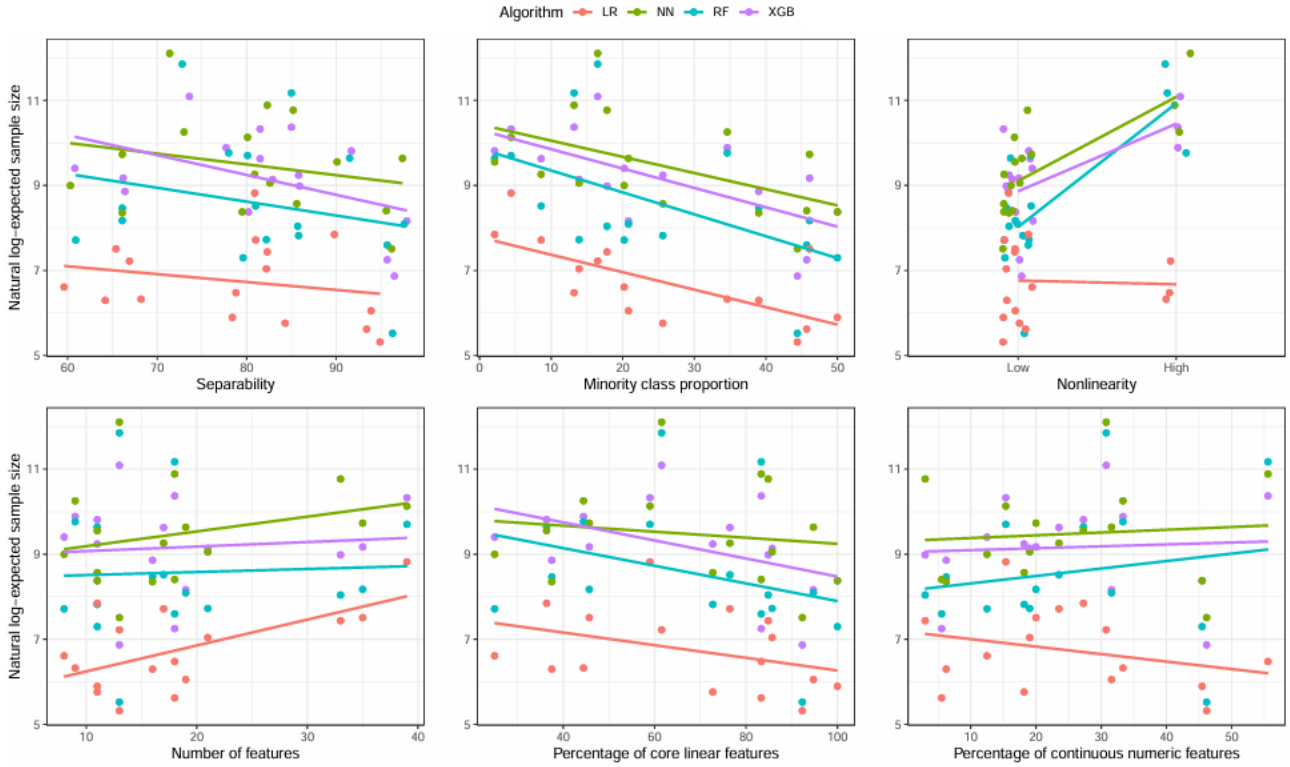

Table 4. Univariable association of each dataset–level characteristics with predicted sample size.

Variable	XGB[a]		RF[b]		LR[c]		NN[d]	
	Estimate (95% Confidence Interval)	*P* value	Estimate (95% Confidence Interval)	*P* value	Estimate (95% Confidence Interval)	*P* value	Estimate (95% Confidence Interval)	*P* value
Minority class proportion	0.959 (0.934-0.985)	<.001	0.931 (0.891-0.977)	<.001	0.963 (0.945-0.983)	.001	0.953 (0.921-0.990)	.003
Separability	0.955 (0.903-1.000)	.02	0.939 (0.861-1.028)	.047	0.999 (0.946-1.051)	.98	0.950 (0.898-1.007)	.03
Number of features	1.000 (0.958-1.053)	.99	0.966 (0.904-1.059)	.35	1.056 (1.021-1.097)	.006	1.001 (0.947-1.071)	.98
Continuous features (%)	1.019 (0.986-1.057)	.21	1.065 (1.013-1.121)	.003	0.971 (0.942-1.006)	.04	1.019 (0.982-1.059)	.04
Core linear Features (%)	0.983 (0.958-1.001)	.07	0.988 (0.941-1.034)	.41	0.982 (0.959-1.004)	.046	0.994 (0.960-1.028)	.61
Dataset nonlinearity	3.889 (1.631-11.209)	.005	15.985 (5.914-55.754)	<.001	0.585 (0.211-2.112)	.35	6.853 (2.763-20.947)	<.001

[a]XGB: XGBoost.

[b]RF: random forest.

[c]LR: logistic regression.

[d]NN: neural network.

In multivariable models for each algorithm, we selected the set of 3 predictors that minimized the AIC. The equation below, as well as Table 5, presents a summary of each algorithm-specific model, which shows the adjusted contribution of each predictor to the expected sample size.

Equation 1: Empirically derived sample size equations for XGB, RF, LR, and NN algorithms.

1.
2.
3.
4.

XSL•FO
RenderX

Table 5. Multivariable negative binomial regression—data-level characteristics effect on predicted sample size.

Variable	XGB[a]		RF[b]		LR[c]		NN[d]	
	Estimate (95% Confidence Interval)	*P* value	Estimate (95% Confidence Interval)	*P* value	Estimate (95% Confidence Interval)	*P* value	Estimate (95% Confidence Interval)	*P* value
Intercept	121,967 (57,883-262,108)	<.001	26,872 (8118-92,378)	<.001	1801 (904-3,809)	<.001	36,819 (18,691-76,970)	<.001
Minority class proportion	0.956 (0.946-0.967)	<.001	0.957 (0.940-0.975)	<.001	0.968 (0.957-0.979)	<.001	0.976 (0.957-0.996)	.02
Separability	0.952 (0.934-0.970)	<.001	0.975 (0.947-1.004)	.07	__[e]	—	—	—
Number of features	—	—	—	—	1.054 (1.034-1.076)	<.001	—	—
Continuous features (%)	—	—	—	—	—	—	0.973 (0.950-0.997)	.02
Core linear features (%)	—	—	—	—	0.988 (0.980-0.996)	.005	—	—
Dataset nonlinearity	3.091 (2.011-4.922)	<.001	12.298 (5.826-28.791),	<.001	—	—	10.209 (4.274-26.569)	<.001

[a]XGB: XGBoost; adjusted pseudo-R^2=0.845.

[b]RF: random forest; adjusted pseudo-R^2=0.808.

[c]LR: logistic regression; adjusted pseudo-R^2=0.798.

[d]NN: neural network; adjusted pseudo-R^2=0.665.

[e]Not available.

For XGB and RF, minority class proportion, separability, and nonlinearity were the top 3 variables selected. For LR, minority class proportion, number of features, and percentage of core features were the top 3 variables. For NN, the top 3 variables were minority class proportion, number of features, and nonlinearity. The direction and magnitude of coefficient estimates from multivariable models were similar to those obtained from univariable models (Table 5). Deviance-based R^2 statistics, adjusted for the number of predictors added, were 0.845 (XGB), 0.808 (RF), 0.798 (LR), and 0.665 (NN; Table 5). This indicated that the dataset–level predictors explained a majority (66.5%-84.5%) of the total deviance in the data among all 4 models, although the NN model was weaker than the other 3. Figure 5 shows the predicted sample sizes estimated from each algorithm-specific model at a variety of levels for each predictor. As can be seen, for all 4 classification algorithms, a balanced class ratio (50% cases versus 50% controls) resulted in the lowest predicted sample sizes.

XSL•FO
RenderX

Figure 5. Fitted values derived from the 4 final negative binomial regression models for each classification algorithm. Shaded lines represent 95% CI. Imbalance = minority class proportion × 100. Separability = Full–data set AUC × 100. Nonlinearity "low": <5, "high": ≥5. LR: Logistic Regression. NN: Neural Networks. RF: Random Forest. XGB: XGBoost.

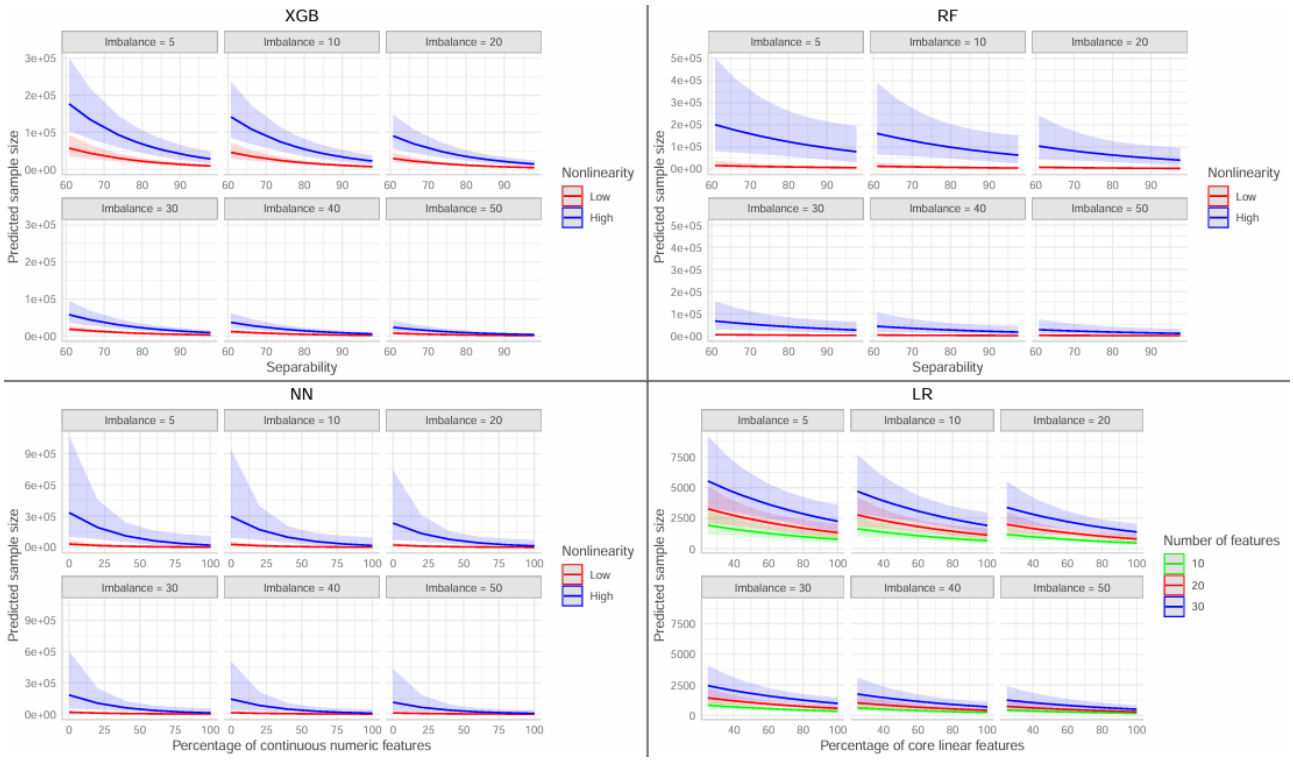

Discussion

Conclusions

In this study, we performed a learning curve analysis of 16 datasets over 4 different classification algorithms. From this, we identified the expected samples needed to reach AUCs within 2 points of those measured in the full dataset. We then examined the effects of dataset–level characteristics on expected sample sizes and provided formulas that can be used to predict the necessary sample size in a new dataset. We found that LR required the smallest sample size (median 696, range 204-6798) but performed slightly worse, on average, compared with more complicated algorithms. RF (median 3404, range 250-140,499) and XGB algorithms (median 9960, range 960-65,556) required larger sample sizes, as expected. NNs required the largest sample size (median 12,998, IQR 1824-180,835) and also had the most variability over the 16 datasets. This was an interesting finding, as our implementation of a deep learning approach was a basic architecture including only one hidden layer and 20 nodes. The fact that NNs required the largest median sample size of >12,000 and were the most variable in terms of expected sample size demonstrates that implementation of deep learning methods—especially more complex deep-learning schemes than what we have provided—should probably be reserved for extremely large datasets for optimal performance and adequate discriminative stability. In addition, NNs had a weaker performance than XGB in every dataset. However, it is possible that a more expressive deep learning approach may yield higher AUCs compared with the best-performing ML solutions considered in this study, although to the best of our knowledge, no other studies have provided deep learning results that outperform ML approaches in these specific datasets. These results support current literature that suggests deep learning may not be optimal for tabular data analysis compared with tree-based methods [35] when weighing accuracy trade-offs versus computational burden.

Our results are consistent with Perlich et al [12], which showed that LR might be optimal in small samples, but tree-based methods eventually provide the best performance in large datasets. Van der Ploeg et al [16] determined that LR required a much lower number of events-per-variable for AUC stability, defined as CV-AUC within 0.01 of the full–dataset performance compared to RF and NN, which required >200 events-per-variable. We can convert our expected sample sizes to events-per-variable by taking the predicted n, multiplying it by the minority class proportion of the dataset, and then dividing it by the number of features. In our study, LR required an average of 11 events-per-variable, XGB: 205, RF: 231, and NN: 342, which supports this notion that modern modeling techniques are "data-hungry" [16].

In summary, this study provides a simple framework for determining sample size in the context of 4 popular ML algorithms. Dataset–level variables that altered expected sample sizes varied by algorithm, but the class imbalance of the outcome, the strength and number of features, and the nonlinearity of the predictors were among the most influential characteristics. Most of these dataset–level characteristics can be reasonably guessed or influenced before the study begins. For example, researchers can examine previous studies in their field of interest to determine a reasonable range for separability and minority class proportion. For minority class proportion, which was a key selected feature for all 4 models, we observed

XSL•FO
RenderX

that an optimal class balance (50% cases, 50% controls) led to the lowest predicted sample sizes, with each additional percentage point of balance decreasing the needed n by a multiplier of 0.96-0.98.

In addition, researchers can use feature engineering to control the quality and overall number of predictors included in their models. As we have determined in this study, a smaller number of strong predictors will generally require less sample size than a large and noisy predictor set, supporting the idea that more features are not always ideal [36]. Dataset nonlinearity is less intuitive to guess before data collection. In general, we found that datasets with nonlinearity values of at least 5.0 required approximately 3-12 times the amount of sample to reach stability, depending on the algorithm. However, in this study, 13/16 (81.3%) of the datasets had values under 5.0, which means that high values of dataset nonlinearity may be uncommon. Again, previous studies where both simple (LR) and complex (NN, RF, and XGB) methods are compared can help researchers determine if this value will be high. As a last resort, researchers can simply calculate expected sample sizes for both scenarios (<5.0 and ≥5.0) using the model equations presented in this study and discuss the implications. It is also important to note that the effect of nonlinearity (and other dataset–level characteristics) on estimated sample sizes is diminished when the class imbalance is optimized. This is due to the multiplicative nature of the negative binomial regression models, which is illustrated in Figure 5. Thus, first and foremost, it is critical that researchers aim to collect a sample with the most balance between cases and controls.

To our knowledge, no previous study has presented specific formulas for calculating sample size within the context of ML using multiple dataset–level characteristics. Although other studies have provided estimates of the needed sample size (or number of events per variable) to reach performance stability over a variety of classification algorithms, these works used simulation approaches or a limited number of real-life datasets and did not consider multiple specific dataset–level characteristics in calculation of these estimates [15,16].

Limitations and Future Work

One limitation of our study was the relatively small sample size of only 16 datasets to develop our final models. Although this number is a relatively large amount in this area of research—similar learning curve analyses have typically examined less than 10 [14,16,17]—assessment of more would strengthen these models and provide clearer insight into dataset–level effects on expected sample size. However, the fact that we still observed many statistically significant relationships even with this small effective sample size is a strength of the study. In addition, the datasets we examined were all tabular in nature and had relatively low (<50) numbers of features—generalization of the formulas presented in this study may not extrapolate to datasets with larger amounts of features or data arising from medical imaging or nontabular sources. Finally, it is important to note that ML, specifically algorithms like RF and XGB, can still outperform traditional parametric methods even if the sample size is limited (ie, under n=5000) and when hyperparameter tuning is implemented [37,38]. Therefore, these guidelines should serve as a supplement to be used in the first stage of predictive modeling, giving a general idea of how much sample is expected to reach a point of "diminishing returns," where large amounts of additional data will only increase the AUC marginally.

Future research in this area could examine different outcome types, such as regression, multiclass, or survival end points—or different performance metrics, such as area under the precision-recall curve or Brier score for probability calibration [10]. In addition, a more in-depth examination of XGB, RF, and NN hyperparameters would be impactful, as all of the equations developed in this study considered only the default hyperparameter values, which could limit the generalizability of the results. However, in practice, it would be extremely difficult to guess plausible values of hyperparameters before data collection, so examination of different configurations would mostly be educational in nature and impractical to incorporate in sample size equations before any data collection. Finally, stacked ML methods [39], or different gradient-boosted tree algorithms such as CatBoost [40] or LightGBM [41] could be investigated.

Data Availability

The datasets generated during and/or analyzed during this study are publicly available and can be accessed using links found in Multimedia Appendix 1. In addition, the datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. An RShiny app is publicly available where researchers can calculate the expected sample size based on our formulas with user-input dataset-level characteristics and can be found at [42].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information: data processing steps and additional tables and figures.
[DOCX File , 35 KB - jmir_v26i1e60231_app1.docx]

References

XSL•FO
RenderX

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
2. Cho H, She J, De Marchi D, El-Zaatari H, Barnes EL, Kahkoska AR, et al. Machine learning and health science research: tutorial. J Med Internet Res 2024;26:e50890 [FREE Full text] [doi: 10.2196/50890] [Medline: 38289657]
3. Bonaccorso G. Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning. Birmingham: Packt Publishing; 2018.
4. Dorey FJ. Statistics in brief: statistical power: what is it and when should it be used? Clin Orthop Relat Res 2011;469(2):619-620 [FREE Full text] [doi: 10.1007/s11999-010-1435-0] [Medline: 20585913]
5. Devijver P, Kittler J. Pattern Recognition: A Statistical Approach. New Jersey, United States: Prentice-Hall; 1982.
6. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2009.
7. He H, Garcia EA. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng 2009;21(9):1263-1284. [doi: 10.1109/tkde.2008.239]
8. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
9. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997;30(7):1145-1159. [doi: 10.1016/s0031-3203(96)00142-2]
10. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020;27(4):621-633 [FREE Full text] [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
11. Webb GI, Sammut C, Perlich C, Horváth T, Wrobel S, Korb KB. Learning curves in machine learning. Encyclopedia of Machine Learning 2011:577-580. [doi: 10.1007/978-0-387-30164-8_452]
12. Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: a learning-curve analysis. Journal of Machine Learning Research 2003:211-255 [FREE Full text]
13. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, et al. Estimating dataset size requirements for classifying DNA microarray data. J Comput Biol 2003;10(2):119-142. [doi: 10.1089/106652703321825928] [Medline: 12804087]
14. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak 2012;12(1):8 [FREE Full text] [doi: 10.1186/1472-6947-12-8] [Medline: 22336388]
15. Provost F, Jensen D, Oates T. Efficient progressive sampling. URL: https://dl.acm.org/doi/pdf/10.1145/312129.312188 [accessed 2024-11-19]
16. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137 [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
17. Richter AN, Khoshgoftaar TM. Sample size determination for biomedical big data with limited labels. Netw Model Anal Health Inform Bioinforma 2020;9(1). [doi: 10.1007/s13721-020-0218-0]
18. Micheel CM, Nass SJ, Omenn GS. Evolution of Translational Omics Lessons Learned and the Path Forward. Washington, DC: National Academies Press; 2012.
19. Rijn JN, Holmes G, Bernhard P, Vanschoren J. Algorithm selection on data streams. Lecture Notes in Computer Science 2014:325-336. [doi: 10.1007/978-3-319-11812-3_28]
20. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken, Nj: John Wiley & Sons, Inc; 2020.
21. Cramer J. The origins of logistic regression. SSRN Electronic Journal 2003:16. [doi: 10.2139/ssrn.360300]
22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD-16; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: 10.1145/2939672.2939785]
23. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Soft 2017;77(1):1-17. [doi: 10.18637/jss.v077.i01]
24. Mall PK, Singh PK, Srivastav S, Narayan V, Paprzycki M, Jaworska T, et al. A comprehensive review of deep neural networks for medical image processing: recent developments and future opportunities. Healthcare Analytics 2023;4:100216. [doi: 10.1016/j.health.2023.100216]
25. Nwanosike EM, Conway BR, Merchant HA, Hasan SS. Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review. Int J Med Inform 2022;159:104679. [doi: 10.1016/j.ijmedinf.2021.104679] [Medline: 34990939]
26. Foundation for Open Access Statistics. Fast scalable R with H20. 2015. URL: https://h2o.ai/ [accessed 2024-11-25]
27. Berrar D. Cross-validation. Encyclopedia of Bioinformatics and Computational Biology 2019;1:542-545. [doi: 10.1016/b978-0-12-809633-8.20349-x]
28. Bates DM, Watts DG. Nonlinear Regression Analysis and Its Applications. Hoboken, New Jersey: Wiley-Interscience; 2007.
29. Kutner M, Nachtsheim C, Neter J. Applied Linear Regression Models. London: Mcgraw-Hill Education - Europe; 2004.

XSL•FO
RenderX

30. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1-22 [FREE Full text] [Medline: 20808728]
31. Hilbe J. Negative Binomial Regression. Cambridge, England: Cambridge University Press; 2007.
32. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika 1987;52(3):345-370. [doi: 10.1007/bf02294361]
33. Vanegas L, Rondón L, Paula G. _glmtoolbox: Set of tools to data analysis using generalized linear models_. 2024. URL: https://CRAN.R-project.org/package=glmtoolbox [accessed 2024-11-19]
34. Veall MR, Zimmermann KF. Pseudo-R2 measures for some common limited dependent variable models. Journal of Economic Surveys 2006;10(3):241-259. [doi: 10.1111/j.1467-6419.1996.tb00013.x]
35. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. arXiv 2021 [FREE Full text] [doi: 10.1016/j.inffus.2021.11.011]
36. Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021;4(1):153 [FREE Full text] [doi: 10.1038/s41746-021-00521-5] [Medline: 34711924]
37. Papini S, Norman SB, Campbell-Sills L, Sun X, He F, Kessler RC, et al. Development and validation of a machine learning prediction model of posttraumatic stress disorder after military deployment. JAMA Netw Open 2023;6(6):e2321273 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.21273] [Medline: 37389870]
38. Abe D, Inaji M, Hase T, Takahashi S, Sakai R, Ayabe F, et al. A prehospital triage system to detect traumatic intracranial hemorrhage using machine learning algorithms. JAMA Netw Open 2022;5(6):e2216393 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.16393] [Medline: 35687335]
39. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007;6:Article25. [doi: 10.2202/1544-6115.1309] [Medline: 17910531]
40. Liudmila OP, Gleb G, Aleksandr V, Anna VD, Andrey G. CatBoost: unbiased boosting with categorical features. arXiv (Cornell University) 2021 [FREE Full text]
41. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 2017;30:3146-3154 [FREE Full text]
42. RShiny app. URL: https://silveys.shinyapps.io/shiny_app_aim_1/ [accessed 2024-11-25]

Abbreviations

AUC: area under the curve
CV: cross-validation
CV-AUC: cross-validated area under the curve
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
MCAR: missing completely at random
ML: machine learning
MLP: multilayer perceptron
NN: neural network
RF: random forest
XGB: XGBoost

Edited by G Tsafnat; submitted 06.05.24; peer-reviewed by R Figueroa, AG Gad, MR Hasan; comments to author 27.07.24; revised version received 20.09.24; accepted 20.10.24; published 17.12.24.

Please cite as:
Silvey S, Liu J
Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study
J Med Internet Res 2024;26:e60231
URL: https://www.jmir.org/2024/1/e60231
doi:10.2196/60231
PMID:

©Scott Silvey, Jinze Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Feedback Visit and a Clinical Decision Support System Based on Antibiotic Prescription Audit in Primary Care: Multiarm Cluster-Randomized Controlled Trial

Pauline Jeanmougin[1,2,3], MD; Stéphanie Larramendy[1], MD, PhD; Jean-Pascal Fournier[1,3], MD, PhD; Aurélie Gaultier[1,4], MSc; Cédric Rat[1,3], MD, PhD

[1]Department of General Practice, Faculty of Medicine, Nantes University, Nantes, France
[2]Antibioclic Steering Committee, Paris, France
[3]POPS - SFR ICAT, University of Angers, Angers, France
[4]Methodology and Biostatistics Platform, Nantes University Hospital, Nantes, France

Corresponding Author:
Pauline Jeanmougin, MD
Department of General Practice
Faculty of Medicine
Nantes University
1, rue Gaston-Veil
Nantes, 44000
France
Phone: 33 02 40 41 11 29
Email: pauline.jeanmougin@univ-nantes.fr

Abstract

Background: While numerous antimicrobial stewardship programs aim to decrease inappropriate antibiotic prescriptions, evidence of their positive impact is needed to optimize future interventions.

Objective: This study aimed to evaluate 2 multifaceted antibiotic stewardship interventions for inappropriate systemic antibiotic prescription in primary care.

Methods: An open-label, cluster-randomized controlled trial of 2501 general practitioners (GPs) working in western France was conducted from July 2019 to January 2021. Two interventions were studied: the standard intervention, consisting of a visit by a health insurance representative who gave prescription feedback and provided a leaflet for treating cystitis and tonsillitis; and a clinical decision support system (CDSS)–based intervention, consisting of a visit with prescription feedback and a CDSS demonstration on antibiotic prescribing. The control group received no intervention. Data on systemic antibiotic dispensing was obtained from the National Health Insurance System (*Système National d'Information Inter-Régimes de l'Assurance Maladie*) database. The overall antibiotic volume dispensed per GP at 12 months was compared between arms using a 2-level hierarchical analysis of covariance adjusted for annual antibiotic prescription volume at baseline.

Results: Overall, 2501 GPs were randomized (n=1099, 43.9% women). At 12 months, the mean volume of systemic antibiotics per GP decreased by 219.2 (SD 61.4; 95% CI −339.5 to −98.8; $P<.001$) defined daily doses in the CDSS-based visit group compared with the control group. The decrease in the mean volume of systemic antibiotics dispensed per GP was not significantly different between the standard visit group and the control group (−109.7, SD 62.4; 95% CI −232.0 to 12.5 defined daily doses; $P=.08$).

Conclusions: A visit by a health insurance representative combining feedback and a CDSS demonstration resulted in a 4.4% (-219.2/4930) reduction in the total volume of systemic antibiotic prescriptions in 12 months.

Trial Registration: ClinicalTrials.gov NCT04028830; https://clinicaltrials.gov/study/NCT04028830

(J Med Internet Res 2024;26:e60535) doi:10.2196/60535

KEYWORDS

antibacterial agents; feedback; clinical decision support system; prescriptions; primary health care; clinical decision; antibiotic prescription; antimicrobial; antibiotic stewardship; interventions; health insurance; systematic antibiotic prescriptions

XSL•FO
RenderX

Introduction

Antibiotic resistance is a worldwide threat to public health and has major consequences for human health and health care systems [1,2]. Recently, the literature has advocated antimicrobial stewardship programs to reduce inappropriate antibiotic prescriptions and improve adherence to medical guidelines in primary care [3]. Several levers have been proposed to change behavior: active or passive training [4], audits and clinical practice feedback [5], patient-focused actions, nudge interventions [6], and financial incentives [7]. Another effective intervention combines an audit of clinical practice and promoting appropriate and necessary antibiotic prescription [8-10]. Although the literature advocates for multifaceted interventions, results vary depending on how these interventions are implemented [7,8,11-13].

Obstacles to implementing relevant interventions include reaching health care professionals and identifying the specific barriers and facilitators that drive behavioral change [14,15]. In France, this task is the responsibility of health insurance representatives (HIRs) who work for the National Health Care Insurance Organization (NHIS) known as Assurance Maladie. This organization provides health care insurance to approximately 90% of French citizens of all ages. HIRs visit general practitioners (GPs) 3 or 4 times a year to promote actions related to public health priorities. However, evidence is lacking on whether these visits do, in practice, have a positive effect on prescribing habits among primary care providers.

In this context, recent studies suggest that clinical decision support systems (CDSS) can help physicians make appropriate decisions when prescribing antibiotics [16,17]. However, their effectiveness in reducing the volume of antibiotics prescribed, and whether they improve the overall quality of prescriptions, has rarely been evaluated at the population scale. Among the several CDSSs available [18-20] to optimize antibiotic prescribing, French GPs most often use Antibioclic [21-23]. The system was developed by French academics and released in 2011, and it provides easy access to best practice recommendations.

This study aimed to evaluate the effect of a multifaceted intervention involving an HIR visit, clinical practice feedback, and the presentation of an online decision aid (the CDSS-based visit group) on antibiotic prescribing among GPs in western France. The intervention was compared with (1) a standard intervention involving a HIR visit, clinical practice feedback, and a prescribing information leaflet (the standard visit group) and (2) no intervention (the control group).

Methods

Study Design

The study was an open-label, cluster-randomized controlled trial. It was conducted between July 2019 and January 2021 among GPs in western France (the Pays de la Loire geographic area, with 3,832,120 inhabitants). The academic research team worked in collaboration with HIRs working for the regional division of the NHIS.

This study followed the CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines (Multimedia Appendix 1) and was registered with ClinicalTrials.gov under NCT04028830 (version 5, April 19, 2024).

Participants and Setting

All 2758 GPs working in private practice in the Pays de la Loire geographical area were considered eligible if they had been practicing in April 2019 and had seen at least 100 different patients in 2018. GPs were excluded if they practiced alternative medicine, including acupuncture, allergology, or angiology; they were participating in another national antibiotic resistance project; or they were involved in developing the intervention.

Interventions

In total, 2 multifaceted interventions were evaluated in this study. The group of HIRs were collectively trained on June 17, 2019, before the interventions.

In the CDSS-based visit group, the intervention was carried out by the regional HIR at the GPs' practice and consisted of (1) providing information about antibiotic resistance, good antibiotic use, and prescription practices; (2) giving feedback based on individual, regional, and national antibiotic prescription rates; and (3) providing a presentation on how to use the CDSS in the treatment of cystitis and tonsillitis. The CDSS is presented in Multimedia Appendices 2 and 3 The user selects the pathology (not limited to tonsilitis or urinary tract infections [21]), and the tool suggests a therapeutic strategy adapted to French national recommendations. Access to Antibioclic is free of charge.

The standard intervention (standard visit group) was also carried out by the HIR at the GPs' practice. It consisted of (1) providing information about antibiotic resistance, good antibiotic use, and prescription practices; (2) giving feedback based on individual, regional, and national antibiotic prescription rates; and (3) providing an information leaflet about the appropriate antibiotic treatment for cystitis and tonsillitis (Multimedia Appendix 4). The control group received a routine visit by the regional HIR, but the discussion focused on a health priority other than antibiotic prescription.

Visits were planned between July 2019 and January 2020.

Randomization

Once eligible GPs had been included, they were randomly assigned at a 1:1:1 ratio to either the CDSS-based visit group, the standard visit group, or the control group (in July 2019). GPs were clustered within practices to avoid contamination bias stemming from shared tracking mechanisms and communication among GPs within the same practice.

The regional division of the NHIS was responsible for selecting GPs according to the inclusion criteria, grouping practices according to their location, and assigning GPs to the randomization arm to which their practice had been allocated.

The assessment of outcomes may be considered formally blinded, as all trial-relevant data were collected with automated processes used to record health insurance claims information.

XSL•FO
RenderX

Outcomes

The primary outcome was the total volume of systemic antibiotics dispensed as defined daily doses (DDD; according to the World Health Organization) per participating GP at the end of 12 months of follow-up. All systemic antibiotics listed in the 2017 Anatomical Therapeutic Chemical classification system were included (Anatomical Therapeutic Chemical class J01) [24]. Patients treated with topical antibiotics, antituberculosis agents, antiparasitic agents, or antimycotic agents were excluded.

Secondary outcomes were the total volume of systemic antibiotics dispensed after 3 months of follow-up; the volume of antibiotics considered critical at the European level according to the Aware classification [25,26] (third-generation cephalosporins, fluoroquinolones, and amoxicillin-clavulanic acid); and the volume of antibiotics dispensed in 2 specific patient subgroups (people aged >65 years and children aged <6 years). In France, these 2 populations are prescribed a high volume of antibiotics [27].

Data Collection

All data were extracted from the NHIS database (*Système National d'Information Inter-Régimes de l'Assurance Maladie*, SNIIRAM) using routine procedures.

To describe the population included in the study, the following variables characterizing GPs were extracted at baseline: (1) age, sex, practice location, type of practice (group or sole practitioner), and patient base (number of patients on their list at the beginning of the intervention), and (2) the number of consultations and antibiotic dispensing data over the year preceding the start of follow-up.

Antibiotic dispensing data (required to calculate primary and secondary outcome measures) were collected over a 12-month period starting from the date of the HIR visit. The SNIIRAM database holds no clinical data (ie, there is no information about a patient's diagnosis or clinical indications).

For GPs who did not receive an HIR visit (refusal, unavailability, etc), a follow-up start date was selected at random from among the actual dates of visits made by HIRs to other GPs.

Statistical Analysis

The minimum number of participants was calculated on the basis of preliminary data. The number of eligible GPs was estimated at 2400. The average volume of antibiotics dispensed per GP was estimated at 7671 (SD 5360) DDD. A previous study [28] that investigated a subdivision of the study region found that the average number of GPs per practice was 1.9. We assumed an intraclass correlation coefficient of 0.05. With an α risk of 5% and a power of 80%, a significant result could be demonstrated if an average difference of 450 DDD per GP was identified between the 2 arms.

The analysis was performed on an intention-to-treat basis among all GPs practicing at the time of randomization (refer to the research protocol, Multimedia Appendix 5). The aim was to conduct a "real-life" study that considered the fluctuating impact of the same intervention received by GPs.

The data collection process minimized missing data. The overall volume of antibiotics dispensed after 12 months of follow-up for GPs who ceased their activity (retirement, moved house, death, etc) during the follow-up period was imputed using a multiple imputation method.

The statistical unit was the GP. GPs and their patients were described using numbers and percentages of each modality for qualitative variables, and by means and SD for quantitative variables, both overall and according to the 3 randomization arms.

For the primary end point, a hierarchical procedure was used to compare the 3 arms, while maintaining a 5% α risk. First, we tested for differences between the CDSS-based visit group and the control group. If this test was significant, a second test was performed to compare the standard visit group with the control group.

Bivariate analyses were used to select variables to be included in the multiple imputation. These concerned the characteristics of GPs at inclusion and when they ceased their activity, and between the characteristics of GPs at inclusion and the main criterion. All significant variables and the randomization arm were included in the model. The number of imputations performed corresponded to the percentage of data to be imputed. Multiple imputation was performed using chained equations and predictive mean matching (*mice* library).

A sensitivity analysis of the per-protocol population was also performed, excluding GPs who had ceased practice and those who had not received a visit from the HIR.

The overall volume of antibiotics dispensed at the end of the 12-month follow-up, per participating GP, was compared between arms using a 2-level hierarchical analysis of covariance (ANCOVA) adjusted for annual antibiotic prescription volume at baseline, with practice as a random effect.

The volumes of antibiotics dispensed after three months to patients considered critical (aged >65 years or <6 years) were also compared between groups using a 2-level hierarchical ANCOVA.

All tests were 2-tailed, with significance defined as $P<.05$. Model assumptions were verified. Analyses were performed using R software (version 3.6.0; Foundation for Statistical Computing).

Ethical Considerations

Nantes University Hospital has adopted the MR004 Reference Methodology, which this protocol complies with. HIR visits are standard practice in the French health insurance system and seek to promote best practices regarding public health. The need to obtain consent from participating GPs and their patients was waived, according to rules governing the exceptional use of health data without the formal consent of participants [29]. Thus, GPs in the intervention groups were unaware that they had been involved in an intervention study, and GPs in the control group

XSL•FO
RenderX

were not informed that their antibiotic prescription was monitored for the duration of the trial.

GP confidentiality was ensured with pseudonymized identifiers created by independent data managers of the regional division of the NHIS.

On July 18, 2019, the protocol (Multimedia Appendix 5) was approved under 110719107 by the Ethics Committee of the National College of Teaching General Practitioners (IRB00010804; Multimedia Appendix 6).

Results

Overview

Among the 2501 GPs included in the study, 835 were randomly assigned to the CDSS-based visit group, 847 to the standard visit group, and 819 to the control group (Figure 1).

Demographic, professional, and antibiotic prescribing characteristics of GPs are provided in Table 1.

The monthly volume of antibiotics dispensed (DDD) per participating GP is shown in Figure 2.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart. CDSS: clinical decision support system; HIR: health insurance representative.

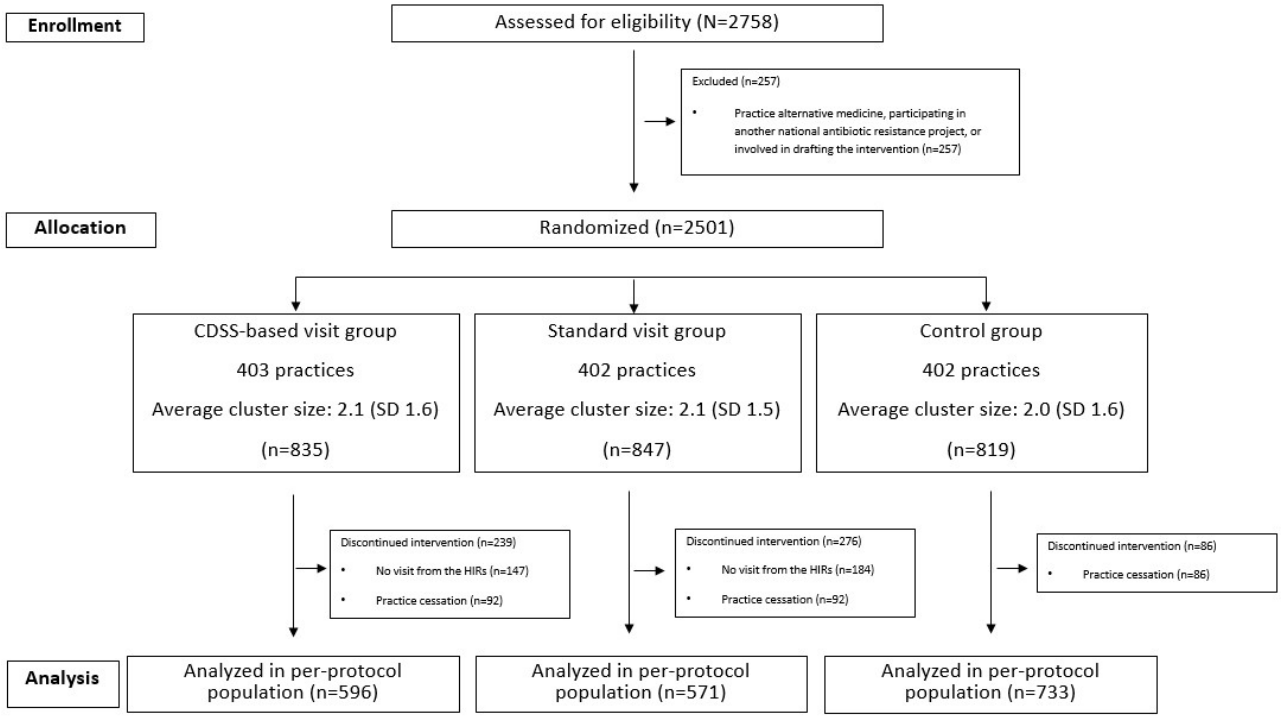

XSL•FO
RenderX

Table 1. Baseline characteristics of participating general practitioners.

Interventions	CDSS[a]-based visit group (n=835)	Standard visit group (n=847)	Control group (n=819)
Sociodemographic characteristics, n (%)			
Age group (years)			
25-35	59 (7.0)	70 (8.3)	77 (9.4)
36-45	99 (11.9)	97 (11.4)	107 (13.1)
46-55	185 (22.2)	181 (21.4)	168 (20.5)
56-65	176 (21.1)	203 (24)	179 (21.9)
≥65	316 (37.8)	296 (35)	287 (35.1)
Not available	0 (0)	0 (0)	1 (0.1)
Woman	362 (43.3)	378 (44.6)	359 (43.8)
Workplace characteristics			
Group practitioner, n (%)	639 (76.5)	646 (76.3)	611 (74.6)
Sole practitioner, n (%)	196 (23.5)	201 (23.7)	208 (25.4)
Consultations per year, mean (SD)	4834.7 (1921)	4758 (2016)	4828.7 (1971)
Patients, mean (SD)	934 (424)	901.2 (438)	903.9 (436)
Activity ceased in the follow-up year, n (%)	92 (11)	92 (10.9)	86 (10.5)
Annual volume of systemic antibiotic prescriptions (DDD[b]) over the year preceding the start of the follow-up			
All, mean (SD)	5714.2 (3870.7)	5506.4 (4271.1)	5661.7 (3675.2)
Critical antibiotics, mean (SD)	1534.3 (1348.5)	1424.0 (1334.6)	1539.2 (1261.1)
Health Insurance Representatives visits conducted, n (%)	648 (77.6)	621 (73.3)	__[c]

[a]CDSS: clinical decision support system.

[b]DDD: defined daily dose.

[c]Not applicable.

Figure 2. The mean volume of systemic antibiotics dispensed per month. Error bars indicate 95% CI. N is defined as general practitioners in continuous practice during the follow-up period. CDSS: clinical decision support system; D0: day 0 intervention; DDD: defined daily doses; M: month.

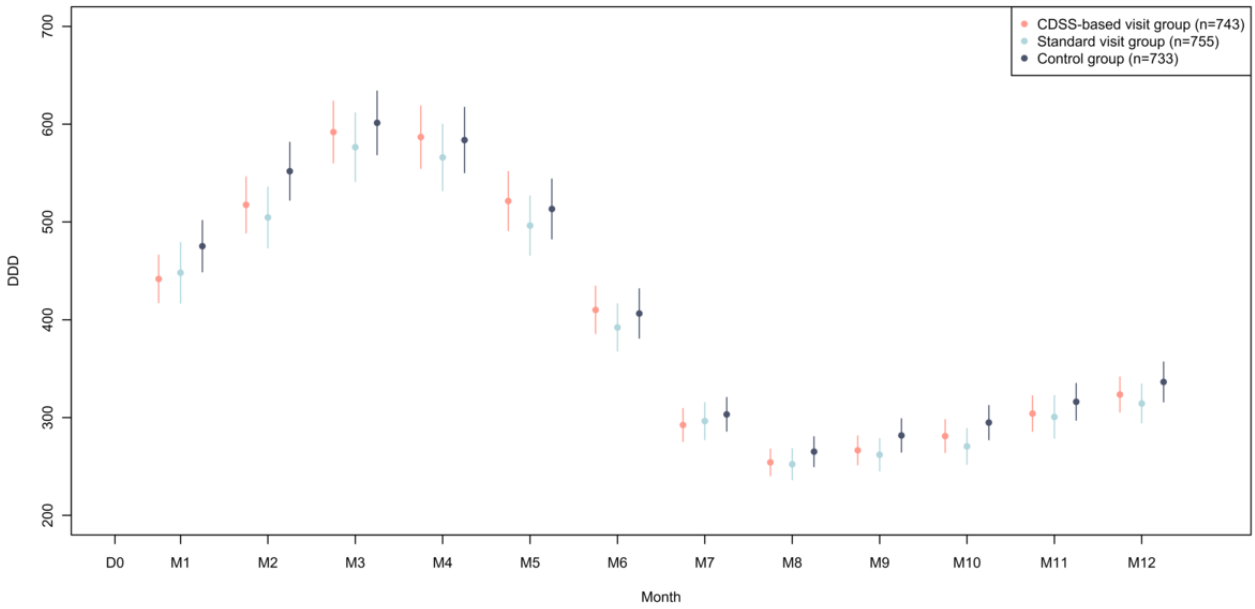

XSL•FO
RenderX

Primary Outcome

At 12 months follow-up, the mean volume of systemic antibiotics per GP decreased by 219.2 (SD 61.4; 95% CI −339.5 to −98.8; *P*<.001) DDD in the CDSS-based visit group compared with the control group (Table 2).

The decrease in the mean volume of antibiotics dispensed per GP was not significantly different between the standard visit group and the control group (−109.7, SD 62.4; 95% CI −232.0 to 12.5 DDD; *P*=.08).

The per-protocol analysis is presented in Multimedia Appendix 7.

Table 2. Total volume of systemic antibiotics dispensed (DDD) per participating general practitioner (GP) at 12-month follow-up.

Arm	GPs in continuous practice during the follow-up period (available data), n	Total volume of systemic antibiotics dispensed, mean (SD)[a]	All GPs, n	Absolute difference (95% CI)[b]	Absolute standardized mean difference[b]	*P* values[b]
CDSS[c]-based visit group	743	4791 (3353.7)	835	−219.2 (−339.5 to −98.8)	−3.57	<.001
Standard visit group	755	4680 (3754.5)	847	−109.7 (−232.0 to12.5)	−1.76	.08
Control	733	4930 (3467.1)	819	Reference	Reference	Reference

[a]Based on available data.

[b]Based on all randomized GPs after multiple imputations. The total volume of systemic antibiotics dispensed was compared between arms using a 2-level hierarchical analysis of covariance adjusted for annual antibiotic prescription volume at baseline, with practice as a random effect.

[c]CDSS: clinical decision support system.

Secondary Outcomes

A reduction in the volume of systemic antibiotic prescriptions was not observed at a 3-month follow-up for either group (Figure 3).

At 12 months, there was a reduction in both intervention groups (CDSS based and standard) in the volume of prescriptions for critical antibiotics, particularly cephalosporins, quinolones, and amoxicillin-clavulanic acid. In the CDSS-based group, there was also a reduction in the volume of prescriptions for patients aged >65 years and <6 years.

Figure 3. Subgroup analysis of the volumes delivered. Error bars indicate 95% CI. CDSS: clinical decision support system; DDD: defined daily doses; M: month.

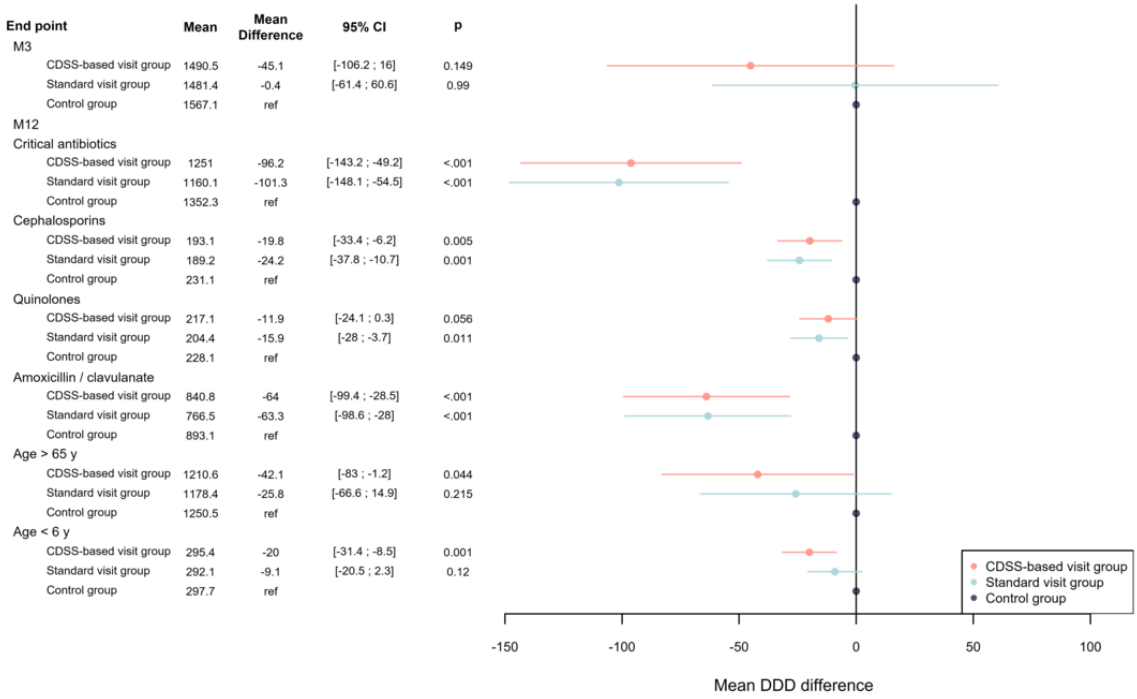

XSL•FO
RenderX

Discussion

Principal Findings

This study is the first randomized controlled trial to demonstrate the positive effect of an antibiotic stewardship intervention involving visits by HIRs, feedback, and the presentation of a CDSS dedicated to antibiotic prescribing at the population scale in France. Considering that France is the fourth-largest country in Europe in terms of antibiotic consumption [30], our findings, which show a 4.4% (-219.2/4930) reduction in the overall volume of consumption over 12 months, are very encouraging.

Our study shows that combining feedback, a face-to-face visit to the GP's practice, and the promotion of an easy-to-use CDSS leads to a reduction in antibiotic prescribing in primary care. These results support and enrich those previously reported in the literature. For example, Ivers et al [31] and Daneman et al [10] showed the effectiveness of auditing and feedback in changing professional practices. Conversely, Aghlmandi et al [32] reported that audits and feedback had no effect. One reason for these conflicting results might be that their intervention involved email (mailing), while our intervention included an in-person HIR visit to GPs. The results of interventions involving the use of a CDSS have also been reported. In hospitals, Nachtigal et al [18] and Carracedo-Martinez et al [33] demonstrated better adherence to recommendations and reduced exposure to antibiotics through the use of a CDSS. In a systematic review targeting primary care, Holstiege et al [19] found a moderate impact of a CDSS on antibiotic prescribing. The effectiveness of our intervention is undoubtedly due to its multifaceted nature, which combines a CDSS demonstration with feedback.

It is likely that a HIR visit to the GP's practice facilitates behavioral change, especially when combined with an easy-to-use CDSS, thus improving adherence to recommendations [20]. The success of the CDSS visit may also be related to the level of CDSS use in our region. Many GPs in the region were already CDSS users [34], and most knew how it worked, which probably made it easier to adopt. Our study suggests that these findings might be generalizable to the primary care setting. The positive effect of both interventions (an HIR visit, feedback, and either a CDSS demonstration or an information leaflet) on antibiotic prescribing (notably regarding broad-spectrum antibiotics and cephalosporins) is also relevant, and this result is consistent with the findings of Høgli et al [8]. The appropriate prescription of antibiotics is a vital issue in the context of increasing antibiotic resistance. The addition of the CDSS demonstration led to fewer antibiotics being prescribed in subgroups of the population that are highly exposed, notably patients aged >65 years and <6 years.

Finally, it should be noted that while, overall, the impact of the intervention was significant, this was not the case at 3 months, which suggests that there was a familiarization and learning effect regarding the CDSS. One hypothesis is that GPs gradually get into the habit of using the CDSS when they prescribe antibiotics for different infections and, thus, improve the overall quality of their prescriptions, in line with national recommendations. A learning effect has already been described when using Antibioclic in clinical studies [23]. It reported no effect on prescriptions from a recently used computer program. The success of the CDSS visit may also be related to the level of CDSS use in our region. The GPs in the region were already heavy CDSS users [34], and most knew how it worked, which probably made it easier to adopt.

A future study in France could examine how the CDSS is used by the biggest prescribers of antibiotics. In Canada, Schwartz et al [35] have shown that an intervention with this type of prescriber can be effective.

Strengths and Weaknesses

Our study has various strengths. Its pragmatic design allowed the intervention to be evaluated under real-life conditions. As a regional project, it involved a large number of GPs. The data sources used to measure outcomes were robust, reliable health care administration databases. Considerable thought and care went into designing this study, and it is in line with recent recommendations on antimicrobial stewardship intervention evaluation [36]. The impact of the intervention, in terms of better practices, is likely to go beyond the 2 pathologies studied here, as the Antibioclic CDSS can be applied to other pathologies. We know, for example, that over 50% of CDSS queries concern 6 pathologies, including cystitis and tonsilitis [21]. The effect on practices may therefore have been underestimated. Our findings are interesting, as there is no consensus in the literature regarding the positive effects of multimodal computerized interventions [37].

Our study also has various limitations. We have no data on the implementation, or actual use of the CDSS by GPs, as this was not measurable. Furthermore, clinical indications for antibiotic prescriptions are not available in the SNIIRAM database. It is therefore impossible to conclude that there may be an improvement in appropriate prescriptions. In addition, it would have been interesting to assess the sustainability of the effect of the intervention beyond 12 months. However, cost and feasibility limitations meant that we were unable to carry out the necessary follow-up. Finally, even if the intervention can be easily reproduced on a national scale since HIR visits are routine practice in all regions in France and are managed by the NHIS, the generalization of our findings to the international context would require the creation of HIRs in other health insurance systems.

Although the study design is based on a randomized controlled trial, it is possible that the periods of lockdown related to the COVID-19 pandemic, beginning on March 17, 2020, could have affected our results. During this period, there was a historic decrease (17%) in antibiotic consumption in France [27]. While antibiotic prescription volumes decreased significantly in each of our 3 groups, this decrease was greater in the CDSS and standard visit groups than in the control group. The gap between the arms may have been even greater if the study had not been conducted during the pandemic. Consequently, it would be interesting to replicate the study in a nonpandemic context.

Conclusion

Our study found that the combination of a HIR visit and feedback, together with a presentation of the CDSS Antibioclic,

XSL•FO
RenderX

led to a 4.4% (-219.2/4930) reduction in the volume of systemic antibiotics prescribed after 12 months. In addition, visits that provided feedback, both with and without the presentation of the CDSS, led to a reduction in the volume of critical antibiotics (broad-spectrum and cephalosporins) prescriptions.

Acknowledgments

The authors would like to thank the Antibioclic Steering Committee (Josselin Le Bel, Tristan Delory, Elisabeth Bouvet, and Sylvie Lariven) for their help during protocol development; ACTION Research group (Willy Boutfol, Solène Schirr-Bonnans, Emmanuelle Bataille, France Nanin, David Boutoille, Sonia Thibaut) for their help during protocol development, protocol reflection, and HIR training; and the Caisse régionale d'Assurance maladie des pays de la Loire and Solène Samson for their help with data collection. Medical writing was provided by Amy Whereat of Speak the Speech Consulting. The authors thank Elaine Seery for the translation. The French network of University Hospitals (Hôpitaux Universitaires du Grand Ouest) funded this individual's contribution to this work. This study was funded by a grant from the French Ministry of Health (PREPS-19-0175).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

PJ, SL, AG, and CR contributed to the concept and design. PJ, SL, AG, JPF, and CR handled the acquisition, analysis, or interpretation of data. PJ managed the drafting of the manuscript. SL, AG, JPF, and CR performed critical revision of the manuscript for important intellectual content. AG managed statistical analysis. PJ, SL, and CR obtained funding. CR performed supervision.

Conflicts of Interest

PJ is the the coordinator of CDSS Antibioclic.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 148 KB - jmir_v26i1e60535_app1.pdf]

Multimedia Appendix 2
Tonsilitis, choice of criteria.
[PNG File , 47 KB - jmir_v26i1e60535_app2.png]

Multimedia Appendix 3
Treatment of tonsillitis.
[PNG File , 57 KB - jmir_v26i1e60535_app3.png]

Multimedia Appendix 4
Leaflet used in the "standard visit group" intervention.
[PNG File , 600 KB - jmir_v26i1e60535_app4.png]

Multimedia Appendix 5
Research protocol.
[DOCX File , 100 KB - jmir_v26i1e60535_app5.docx]

Multimedia Appendix 6
Ethical approval.
[PDF File (Adobe PDF File), 162 KB - jmir_v26i1e60535_app6.pdf]

Multimedia Appendix 7
Total volume of systemic antibiotics dispensed in defined daily doses (DDD) per participating general practitioner after 12-month follow-up (per-protocol analysis).
[DOCX File , 15 KB - jmir_v26i1e60535_app7.docx]

References

XSL•FO
RenderX

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 2022;399(10325):629-655 [FREE Full text] [doi: 10.1016/S0140-6736(21)02724-0] [Medline: 35065702]
2. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European economic Area in 2015: a population-level modelling analysis. Lancet Infect Dis 2019;19(1):56-66 [FREE Full text] [doi: 10.1016/S1473-3099(18)30605-4] [Medline: 30409683]
3. Sadeq AA, Hasan SS, AbouKhater N, Conway BR, Abdelsalam AE, Shamseddine JM, et al. Exploring antimicrobial stewardship influential interventions on improving antibiotic utilization in outpatient and inpatient settings: a systematic review and meta-analysis. Antibiotics (Basel) 2022;11(10):1306 [FREE Full text] [doi: 10.3390/antibiotics11101306] [Medline: 36289964]
4. McNulty C, Hawking M, Lecky D, Jones L, Owens R, Charlett A, et al. Effects of primary care antimicrobial stewardship outreach on antibiotic use by general practice staff: pragmatic randomized controlled trial of the TARGET antibiotics workshop. J Antimicrob Chemother 2018;73(5):1423-1432 [FREE Full text] [doi: 10.1093/jac/dky004] [Medline: 29514268]
5. Carney G, Maclure M, Patrick DM, Fisher A, Stanley D, Bassett K, et al. A cluster randomized trial assessing the impact of personalized prescribing feedback on antibiotic prescribing for uncomplicated acute cystitis to family physicians. PLoS One 2023;18(7):e0280096 [FREE Full text] [doi: 10.1371/journal.pone.0280096] [Medline: 37523381]
6. Raban MZ, Gonzalez G, Nguyen AD, Newell BR, Li L, Seaman KL, et al. Nudge interventions to reduce unnecessary antibiotic prescribing in primary care: a systematic review. BMJ Open 2023;13(1):e062688 [FREE Full text] [doi: 10.1136/bmjopen-2022-062688] [Medline: 36657758]
7. Arnold SR, Straus SE. Interventions to improve antibiotic prescribing practices in ambulatory care. Cochrane Database Syst Rev 2005;2005(4):CD003539 [FREE Full text] [doi: 10.1002/14651858.CD003539.pub2] [Medline: 16235325]
8. Høgli JU, Garcia BH, Skjold F, Skogen V, Småbrekke L. An audit and feedback intervention study increased adherence to antibiotic prescribing guidelines at a Norwegian hospital. BMC Infect Dis 2016;16:96 [FREE Full text] [doi: 10.1186/s12879-016-1426-1] [Medline: 26920549]
9. Grigoryan L, Zoorob R, Germanos G, Sidani M, Horsfield M, Khan F, et al. Case-based audit and feedback around a decision aid improved antibiotic choice and duration for uncomplicated cystitis in primary care clinics. Fam Med Community Health 2021;9(3):e000834 [FREE Full text] [doi: 10.1136/fmch-2020-000834] [Medline: 34244304]
10. Daneman N, Lee SM, Bai H, Bell CM, Bronskill SE, Campitelli MA, et al. Population-wide peer comparison audit and feedback to reduce antibiotic initiation and duration in long-term care facilities with embedded randomized controlled trial. Clin Infect Dis 2021;73(6):e1296-e1304 [FREE Full text] [doi: 10.1093/cid/ciab256] [Medline: 33754632]
11. Liu E, Linder KE, Kuti JL. Antimicrobial stewardship at transitions of care to outpatient settings: synopsis and strategies. Antibiotics (Basel) 2022;11(8):1027 [FREE Full text] [doi: 10.3390/antibiotics11081027] [Medline: 36009896]
12. Steinman MA, Ranji SR, Shojania KG, Gonzales R. Improving antibiotic selection: a systematic review and quantitative analysis of quality improvement strategies. Med Care 2006;44(7):617-628. [doi: 10.1097/01.mlr.0000215846.25591.22] [Medline: 16799356]
13. Tonkin-Crine SK, Tan PS, van Hecke O, Wang K, Roberts NW, McCullough A, et al. Clinician-targeted interventions to influence antibiotic prescribing behaviour for acute respiratory infections in primary care: an overview of systematic reviews. Cochrane Database Syst Rev 2017;9(9):CD012252 [FREE Full text] [doi: 10.1002/14651858.CD012252.pub2] [Medline: 28881002]
14. Chammat M, Litvine L, McMinigal T. The potential benefits of applying behavioural insights to reduce antibiotic overconsumption. Bulletin Epidémiologique Hebdomadaire 2021 Nov 16;18-19:370-373 [FREE Full text]
15. Suttels V, van Singer M, Clack LC, Plüss-Suard C, Niquille A, Mueller Y, et al. Factors influencing the implementation of antimicrobial stewardship in primary care: a narrative review. Antibiotics (Basel) 2022;12(1):30 [FREE Full text] [doi: 10.3390/antibiotics12010030] [Medline: 36671230]
16. Zhuo C, Wei X, Zhang Z, Hicks JP, Zheng J, Chen Z, et al. An antibiotic stewardship programme to reduce inappropriate antibiotic prescribing for acute respiratory infections in rural Chinese primary care facilities: study protocol for a clustered randomised controlled trial. Trials 2020;21(1):394 [FREE Full text] [doi: 10.1186/s13063-020-04303-4] [Medline: 32398065]
17. Müller L, Srinivasan A, Abeles SR, Rajagopal A, Torriani FJ, Aronoff-Spencer E. A risk-based clinical decision support system for patient-specific antimicrobial therapy (iBiogram): design and retrospective analysis. J Med Internet Res 2021;23(12):e23571 [FREE Full text] [doi: 10.2196/23571] [Medline: 34870601]
18. Nachtigall I, Tafelski S, Deja M, Halle E, Grebe MC, Tamarkin A, et al. Long-term effect of computer-assisted decision support for antibiotic treatment in critically ill patients: a prospective 'before/after' cohort study. BMJ Open 2014;4(12):e005370 [FREE Full text] [doi: 10.1136/bmjopen-2014-005370] [Medline: 25534209]
19. Holstiege J, Mathes T, Pieper D. Effects of computer-aided clinical decision support systems in improving antibiotic prescribing by primary care providers: a systematic review. J Am Med Inform Assoc 2015;22(1):236-242 [FREE Full text] [doi: 10.1136/amiajnl-2014-002886] [Medline: 25125688]
20. Demonchy E, Dufour JC, Gaudart J, Cervetti E, Michelet P, Poussard N, et al. Impact of a computerized decision support system on compliance with guidelines on antibiotics prescribed for urinary tract infections in emergency departments: a

XSL•FO
RenderX

multicentre prospective before-and-after controlled interventional study. J Antimicrob Chemother 2014;69(10):2857-2863. [doi: 10.1093/jac/dku191] [Medline: 24898019]
21. Delory T, Jeanmougin P, Lariven S, Aubert JP, Peiffer-Smadja N, Boëlle PY, et al. A computerized decision support system (CDSS) for antibiotic prescription in primary care-Antibioclic: implementation, adoption and sustainable use in the era of extended antimicrobial resistance. J Antimicrob Chemother 2020;75(8):2353-2362. [doi: 10.1093/jac/dkaa167] [Medline: 32357226]
22. Durand C, Alfandari S, Béraud G, Tsopra R, Lescure FX, Peiffer-Smadja N. Clinical decision support systems for antibiotic prescribing: an inventory of current French language tools. Antibiotics (Basel) 2022;11(3):384 [FREE Full text] [doi: 10.3390/antibiotics11030384] [Medline: 35326847]
23. Delory T, Maillard A, Tubach F, Böelle PY, Bouvet E, Lariven S, et al. Appropriateness of intended antibiotic prescribing using clinical case vignettes in primary care, and related factors. Eur J Gen Pract 2024;30(1):2351811 [FREE Full text] [doi: 10.1080/13814788.2024.2351811] [Medline: 38766775]
24. ATC index with DDDs. WHO Collaborating Centre for Drug Statistics Methodology. URL: https://www.whocc.no/atc_ddd_index_and_guidelines/atc_ddd_index/ [accessed 2024-02-22]
25. Budd E, Cramp E, Sharland M, Hand K, Howard P, Wilson P, et al. Adaptation of the WHO essential medicines list for national antibiotic stewardship policy in England: being AWaRe. J Antimicrob Chemother 2019;74(11):3384-3389 [FREE Full text] [doi: 10.1093/jac/dkz321] [Medline: 31361000]
26. Sharland M, Pulcini C, Harbarth S, Zeng M, Gandra S, Mathur S, et al. Classifying antibiotics in the WHO essential medicines list for optimal use-be AWaRe. Lancet Infect Dis 2018;18(1):18-20. [doi: 10.1016/S1473-3099(17)30724-7] [Medline: 29303731]
27. NA. Urban Antibiotic Consumption in France, 2011-2021. Saint-Maurice, France: Santé Publique France; 2011:19.
28. Rat C, Pogu C, Le Donné D, Latour C, Bianco G, Nanin F, et al. Effect of physician notification regarding nonadherence to colorectal cancer screening on patient participation in fecal immunochemical test cancer screening: a randomized clinical trial. JAMA 2017;318(9):816-824 [FREE Full text] [doi: 10.1001/jama.2017.11387] [Medline: 28873160]
29. Office for Human Research Protections. What are the regulatory bases for waiving or altering some or all of the required elements of informed consent or parental permission? US Department of Health and Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/informed-consent/index.html [accessed 2023-07-14]
30. Antimicrobial consumption in the EU/EEA (ESAC-Net) - annual epidemiological report 2022. European Centre for Disease Prevention and Control. 2023. URL: https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-consumption-europe-2022 [accessed 2024-11-29]
31. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012;2012(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
32. Aghlmandi S, Halbeisen FS, Saccilotto R, Godet P, Signorell A, Sigrist S, et al. Effect of antibiotic prescription audit and feedback on antibiotic prescribing in primary care: a randomized clinical trial. JAMA Intern Med 2023;183(3):213-220 [FREE Full text] [doi: 10.1001/jamainternmed.2022.6529] [Medline: 36745412]
33. Carracedo-Martinez E, Gonzalez-Gonzalez C, Teixeira-Rodrigues A, Prego-Dominguez J, Takkouche B, Herdeiro MT, Galician Pharmacoepidemiology Research Group. Computerized clinical decision support systems and antibiotic prescribing: a systematic review and meta-analysis. Clin Ther 2019;41(3):552-581. [doi: 10.1016/j.clinthera.2019.01.018] [Medline: 30826093]
34. Verger P, Fressard L, Jacquemot AF, Bergeat M, Vergier N. Every second GP is confronted with antimicrobial resistance problems. DREES Etudes Résultats 2022;1217:6.
35. Schwartz KL, Ivers N, Langford BJ, Taljaard M, Neish D, Brown KA, et al. Effect of antibiotic-prescribing feedback to high-volume primary care physicians on number of antibiotic prescriptions: a randomized clinical trial. JAMA Intern Med 2021;181(9):1165-1173 [FREE Full text] [doi: 10.1001/jamainternmed.2021.2790] [Medline: 34228086]
36. Schweitzer VA, van Werkhoven CH, Rodríguez Baño J, Bielicki J, Harbarth S, Hulscher M, et al. Optimizing design of research to evaluate antibiotic stewardship interventions: consensus recommendations of a multinational working group. Clin Microbiol Infect 2020;26(1):41-50 [FREE Full text] [doi: 10.1016/j.cmi.2019.08.017] [Medline: 31493472]
37. Catho G, Sauser J, Coray V, Da Silva S, Elzi L, Harbarth S, et al. Impact of interactive computerised decision support for hospital antibiotic use (COMPASS): an open-label, cluster-randomised trial in three Swiss hospitals. Lancet Infect Dis 2022;22(10):1493-1502 [FREE Full text] [doi: 10.1016/S1473-3099(22)00308-5] [Medline: 35870478]

Abbreviations

ANCOVA: analysis of covariance
CDSS: clinical decision support system
CONSORT: Consolidated Standards of Reporting Trials
DDD: defined daily doses
GP: general practitioner

XSL•FO
RenderX

HIR: health insurance representative
NHIS: National Health Care Insurance Organization
SNIIRAM: Système National d'Information Inter-Régimes de l'Assurance Maladie

Edited by Y Zhuang; submitted 14.05.24; peer-reviewed by M François, G Carney, S MacBride-Stewart; comments to author 05.08.24; revised version received 30.09.24; accepted 05.10.24; published 18.12.24.

Please cite as:
Jeanmougin P, Larramendy S, Fournier JP, Gaultier A, Rat C
Effect of a Feedback Visit and a Clinical Decision Support System Based on Antibiotic Prescription Audit in Primary Care: Multiarm Cluster-Randomized Controlled Trial
J Med Internet Res 2024;26:e60535
URL: https://www.jmir.org/2024/1/e60535
doi:10.2196/60535
PMID:39693139

©Pauline Jeanmougin, Stéphanie Larramendy, Jean-Pascal Fournier, Aurélie Gaultier, Cédric Rat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Machine Learning–Based Prediction Model for Acute Kidney Injury in Patients With Community-Acquired Pneumonia: Multicenter Validation Study

Mengqing Ma[1*], MD; Caimei Chen[2*], MD; Dawei Chen[1], MD; Hao Zhang[1], MD; Xia Du[3], MD; Qing Sun[3], MD; Li Fan[1], MD; Huiping Kong[3], MD; Xueting Chen[4], MD; Changchun Cao[3*], MD, PhD; Xin Wan[1], MD

[1]Department of Nephrology, Nanjing First Hospital, Nanjing Medical University, Nanjing, China
[2]Department of Nephrology, Wuxi People's Hospital Affiliated with Nanjing Medical University, Wuxi, China
[3]Department of Nephrology, Sir Run Run Hospital, Nanjing Medical University, Nanjing, China
[4]Department of Nephrology, Xinyi people's Hospital, Xuzhou, China
[*]these authors contributed equally

Corresponding Author:
Xin Wan, MD
Department of Nephrology
Nanjing First Hospital, Nanjing Medical University
Changle Road 68, Nanjing
Nanjing, 210006
China
Phone: 86 18951670991
Email: wanxin@njmu.edu.cn

Abstract

Background: Acute kidney injury (AKI) is common in patients with community-acquired pneumonia (CAP) and is associated with increased morbidity and mortality.

Objective: This study aimed to establish and validate predictive models for AKI in hospitalized patients with CAP based on machine learning algorithms.

Methods: We trained and externally validated 5 machine learning algorithms, including logistic regression, support vector machine, random forest, extreme gradient boosting, and deep forest (DF). Feature selection was conducted using the sliding window forward feature selection technique. Shapley additive explanations and local interpretable model-agnostic explanation techniques were applied to the optimal model for visual interpretation.

Results: A total of 6371 patients with CAP met the inclusion criteria. The development of CAP-associated AKI (CAP-AKI) was recognized in 1006 (15.8%) patients. The 11 selected indicators were sex, temperature, breathing rate, diastolic blood pressure, C-reactive protein, albumin, white blood cell, hemoglobin, platelet, blood urea nitrogen, and neutrophil count. The DF model achieved the best area under the receiver operating characteristic curve (AUC) and accuracy in the internal (AUC=0.89, accuracy=0.90) and external validation sets (AUC=0.87, accuracy=0.83). Furthermore, the DF model had the best calibration among all models. In addition, a web-based prediction platform was developed to predict CAP-AKI.

Conclusions: The model described in this study is the first multicenter-validated AKI prediction model that accurately predicts CAP-AKI during hospitalization. The web-based prediction platform embedded with the DF model serves as a user-friendly tool for early identification of high-risk patients.

(J Med Internet Res 2024;26:e51255) doi:10.2196/51255

KEYWORDS

acute kidney injury; community-acquired; pneumonia; machine learning; prediction model

XSL•FO
RenderX

Introduction

Acute kidney injury (AKI) is common in patients with community-acquired pneumonia (CAP), which is associated with higher morbidity and mortality, longer hospital stays, and increased financial costs. In recent years, outbreaks of pneumonia caused by new pathogens, including SARS in 2002, Middle East respiratory syndrome in 2012, and the COVID-19 pandemic in 2019, have severely strained resources and had significant societal impacts. In addition, a considerable percentage of patients with CAP, ranging from 4.3% (429/10,069) to 34.4% (631/1836), develop AKI [1-3]. Identifying and intervening in AKI early can facilitate the timely restoration of renal function, ultimately enhancing survival rates and aiding patients in their recovery post discharge [4,5]. However, AKI is primarily detected through elevated levels of serum creatinine (Scr), which is a delayed indicator of AKI. Although various novel urine and plasma biomarkers have shown promising results in diagnosing AKI, only a few are presently used in clinical practice [6]. Furthermore, the cost associated with biomarker assessment and the challenges in interpreting the results have limited their widespread adoption.

In recent years, machine learning (ML) techniques have offered the potential to develop accurate prediction models for AKI, enabling early risk stratification and personalized interventions. A retrospective study published in *Nature* in 2019 included a total of 703,782 adult patients from 172 hospitals and 1062 outpatient facilities. By using the recurrent neural network algorithm, this study demonstrated the ability to predict 55.8% (18,837/33,759) of patients with AKI 48 hours in advance, with an impressive area under the receiver operating characteristic curve (AUC) of 92% [7]. These findings suggest that ML can accurately predict the occurrence of AKI at an early stage.

Numerous studies have focused on ML models for predicting AKI in patients. However, these models have primarily focused on vigorous patients, patients with sepsis, patients with cardiac surgery, patients with cancer, and other specific populations. Currently, there is a lack of models for predicting AKI in patients with CAP. Furthermore, most of the studies have relied on databases like MIMIC-III (The Medical Information Mart for Intensive Care III), with a scarcity of local patient databases and a lack of external validation. Therefore, there is an urgent need to develop AKI prediction models specifically tailored to the population of patients with CAP to provide targeted guidance for diagnosis and treatment. This study aimed to use ML methods to develop models that accurately predict CAP-associated AKI (CAP-AKI).

Methods

Study Population

This retrospective cohort study involved 3 independent tertiary hospitals—Wuxi People's Hospital, Nanjing First Hospital, and Sir Run Run Hospital affiliated with Nanjing Medical University. Data were collected from January 2016 to December 2017. The inclusion criteria included patients aged 18 years and older who were diagnosed with CAP during hospitalization. The study exclusion criteria were (1) patients diagnosed with stage 4-5 chronic kidney disease (CKD) or receiving regular renal replacement therapy, (2) patients with a history of kidney transplant, (3) patients with a hospital stay of less than 24 hours or exceeding 90 days, (4) patients who had been treated by other medical institutions within 14 days before hospitalization, (5) patients who relied on mechanical ventilation and palliative care for advanced tumors on a chronic basis, (6) occurrence of AKI within 24 hours of admission or 7 days thereafter, and (7) patients with peak Scr levels less than 53 μmol/L or Scr measurements fewer than 2 times between 1 year before hospitalization and discharge or fewer than 1 time during hospitalization.

Primary Outcome

The primary outcome for developing the ML models was AKI in patients with CAP. AKI is identified based on the Scr level, as stated by the Kidney Disease Improving Global Outcomes (KDIGO) [8]. It is defined as an increase in Scr to ≥1.5 times baseline within 7 days of admission, or an increase in Scr ≥0.3 mg/dL within 48 hours. The baseline Scr level was determined as the average within 7 to 365 days before hospitalization. If baseline Scr is not recorded, the first Scr value upon admission is used as the baseline Scr. Urine output was not considered for AKI diagnosis due to missing data [9]. Estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation [10]. Proteinuria is defined as "The presence of at least 1+ protein in urinalysis."

Training and Validation Data

The dataset was divided into 2 parts: one from Wuxi People's Hospital for training and internal validation (2568/3211, 80% for model development and 643/3211, 20% for validation) and the other from the multicenter patients for external validation.

Data Processing

Baseline characteristics and clinical data of patients within 24 hours of admission, including 62 indicators, were obtained from the electronic record system. Variables with over 15% (9/62) missing values were excluded, and the remaining variables were imputed using the missForest method. Outliers were identified and treated as missing values. The data of each feature were standardized using the *z* score method. Variables with strong collinearity (variance inflation factor [VIF] value≥10) were discarded using the VIF. The synthetic minority over-sampling technique (SMOTE) was applied to address data imbalance and improve the prediction efficiency of the models.

Feature Selection and Machine Learning Models Establishment

Feature importance was evaluated using the *sklearn* library method. Features were sorted in descending order based on their scores. A sliding windows sequential forward feature selection technique method based on random forest (RF) and out-of-bag (OOB) filtering was used to select features related to AKI. The optimal number of features is determined by finding the subset that results in the lowest OOB error (or the highest OOB score). After feature selection, 5 ML algorithms, including logistic regression (LR), RF, support vector machine, extreme gradient boosting (XGBoost), and deep forest (DF) models, were used

XSL•FO
RenderX

to construct the prediction models. A detailed description of ML methods was given in Multimedia Appendix 1.

Model Validation and Performance

External datasets were used to validate the models, and their performances were evaluated using various metrics such as the AUC, specificity, sensitivity, positive predictive value (PPV), negative predictive value, F_1-score, expected calibration error (ECE), calibration curve, and decision curve. A 1000-bootstrap method was used to determine CIs by sampling with replacement on the prediction indices.

Explanation of the Model

Shapley additive explanations (SHAP) is a method that constructs an additive interpretation model. It considers all features as "contributors" and calculates the marginal contribution of each feature when added to the model output. SHAP aims to provide explanations for "black box models" at both global and local levels, making the model more understandable and applicable.

On the other hand, local interpretable model-agnostic explanations (LIME) modify a single data sample by adjusting the eigenvalues and observing its effect on the output. LIME acts as an "interpreter" by explaining the prediction of each data sample. In LIME, a set of interpretations represents the contribution of each feature to the prediction of a single sample.

Statistical Analysis

To compare baseline characteristics across different datasets, we used 1-way ANOVA and chi-square tests. Statistical significance was determined using a 2-sided *P* value of less than .05. The ML models were created using Python (version 3.9.7; Python Software Foundation) with packages such as *scikit-learn*, *NumPy*, *SHAP*, and *pandas*, as well as R (version 4.1.0; R Core Team) with the *glmnet* package.

Ethical Considerations

The study was approved by the institutional review boards of Wuxi People's Hospital (KY22073), Sir Run Run Hospital (2021-SR-042), and Nanjing First Hospital (KY20230410-01-KS-01). The study was conducted in accordance with the declaration of Helsinki. Due to its retrospective design, the ethics committees from Wuxi People's Hospital, Sir Run Run Hospital, and Nanjing First Hospital waived the need for informed consent. Data can be analyzed secondary without additional consent. The personal information of patients was kept confidential throughout the entire study. Since the experimental design did not involve any intervention for the patients, no compensation was paid to the patients in the study.

Results

Baseline Characteristics

The study comprised a total of 6371 patients, with 3211 (50.4%) patients assigned to the training and internal validation groups, and 3160 (49.6%) patients assigned to the external validation group (Figure 1). Out of the total patients, 2566 (40.2%) were female, with a mean age of 67.0 (SD 17.6) years and a baseline Scr level of 72.5 (SD 31.1) µmol/L. Table 1 illustrates the differences between the training and validation sets.

Figure 1. Flowchart for patient selection. CAP: community-acquired pneumonia; CAP-AKI: community-acquired pneumonia–associated acute kidney injury.

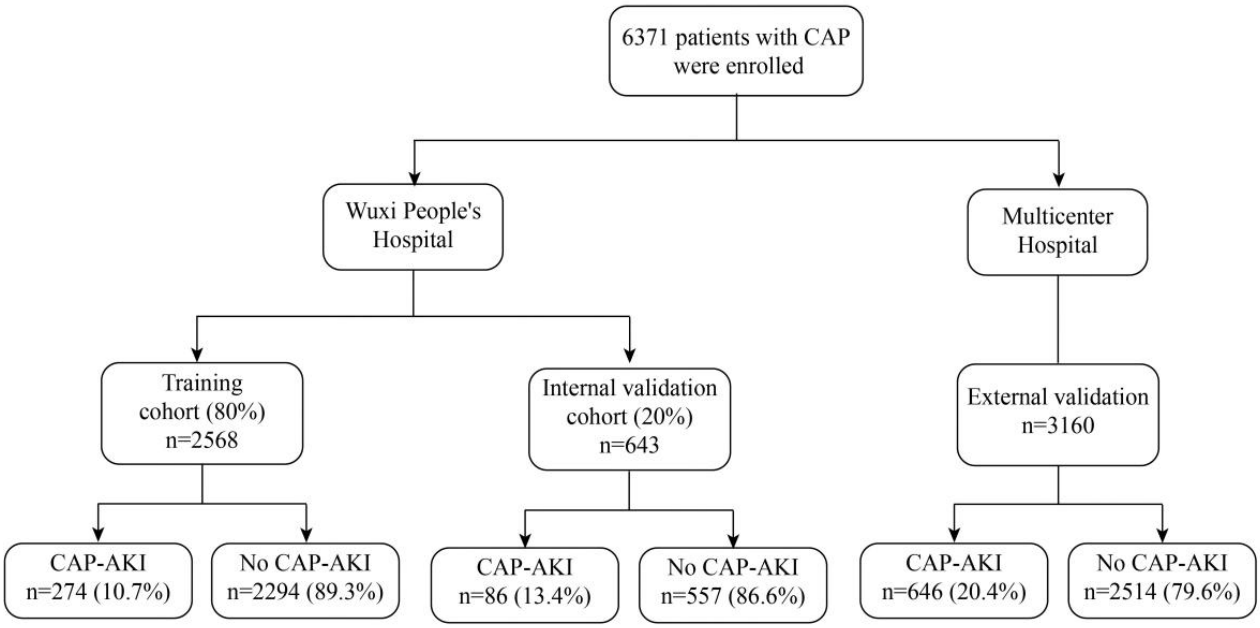

XSL•FO
RenderX

Table 1. Baseline characteristics of the patients with community-acquired pneumonia.

Variables	All (n=6371)	Training set (n=2568)	Internal validation set (n=643)	External validation set (n=3160)	*P* value
Female, n (%)	2566 (40.2)	996 (38.8)	227 (35.3)	1343 (42.5)	<.001
Age (years), mean (SD)	67.0 (17.6)	65.0 (17.9)	65.7 (17.2)	68.8 (17.3)	<.001
Hypertension, n (%)	3007 (47.1)	1227 (47.8)	296 (46)	1484 (46.9)	.68
Diabetes, n (%)	1416 (22.2)	720 (28)	201 (31.3)	495 (15.6)	<.001
CKD[a], n (%)	420 (6.6)	209 (8.1)	61 (9.5)	150 (4.7)	<.001
CVD[b], n (%)	1322 (20.7)	484 (18.8)	127 (19.8)	711 (22.5)	.003
Cerebrovascular disease, n (%)	977 (15.3)	379 (14.7)	94 (14.6)	504 (15.9)	.40
Chronic lung disease, n (%)	1965 (30.8)	774 (30.1)	195 (30.3)	996 (31.5)	.51
Cancer, n (%)	469 (7.4)	205 (8)	49 (7.6)	215 (6.8)	.23
Mechanical ventilation, n (%)	345 (5.4)	127 (4.9)	30 (4.7)	188 (5.9)	.17
Temperature[c] (°C), mean (SD)	37.3 (1.2)	37.8 (1.2)	37.8 (1.2)	36.9 (1.8)	<.001
Heart rate[d] (beats/minute), mean (SD)	84.3 (14.9)	83.7 (14.2)	83.6 (14.2)	85.0 (15.6)	.001
Breathing rate[e] (births/minute), mean (SD)	19.6 (4.0)	19.5 (3.7)	19.3 (3.1)	19.7 (4.3)	.02
Systolic BP[f,g] (mm Hg), mean (SD)	126.6 (18.3)	127.6 (19.7)	126.9 (19.5)	125.7 (16.4)	<.001
Diastolic BP[h] (mm Hg), mean (SD)	76.1 (10.9)	77.4 (19.3)	76.9 (10.7)	75.1 (11)	<.001
Acute respiratory failure, n (%)	598 (9.4)	320 (12.5)	79 (12.3)	199 (6.2)	<.001
Acute cardiac dysfunction, n (%)	354 (5.6)	200 (7.8)	55 (8.6)	99 (3.1)	<.001
ACEI[i] or ARB[j], n (%)	910 (14.3)	369 (14.4)	93 (14.5)	447 (14.1)	.97
Diuretic, n (%)	1452 (22.7)	414 (16.1)	110 (17.1)	928 (29.3)	<.001
CCB[k], n (%)	1251 (19.6)	546 (21.2)	129 (20)	576 (18.2)	.02
β-block, n (%)	364 (5.7)	149 (5.8)	40 (6.2)	175 (5.5)	.77
Baseline Scr[l] (μmol/L), mean (SD)	72.5 (31.1)	72.5 (22.5)	73.7 (24.3)	72.2 (37.3)	.52
eGFR[m] (mL/min×1.73m^2), mean (SD)	84.9 (27.6)	84.3 (25.7)	82.8 (25.4)	85.8 (29.3)	.02
WBC[n] (×10^9/L), median (IQR)	6.9 (5.2-9.5)	6.9 (5.3-9.5)	7.0 (5.2-9.5)	6.4 (4.9-8.8)	.03
Neutrophil count (×10^9/L), median (IQR)	4.9 (3.4-7.4)	4.8 (3.3-7.4)	4.8 (3.4-7.1)	5.0 (3.5-7.6)	.004
Platelet (×10^9/L), mean (SD)	210.0 (87.9)	211.7 (86.0)	214.9 (107.4)	208.1 (88.4)	.22
CRP[o] (mg/L), median (IQR)	21.9 (4.0-74.2)	14.0 (2.0-48.9)	13.9 (2.0-48.3)	19.1 (6.4-73.9)	<.001
Hemoglobin (g/L), mean (SD)	121.1 (20.0)	123.0 (18.7)	122.6 (19.2)	119.3 (20.9)	<.001
Albumin (g/L), mean (SD)	32.9 (5.7)	32.0 (5.9)	31.7 (5.9)	33.9 (5.3)	<.001
BUN[p] (mmol/L), median (IQR)	5.1 (3.7-4.7)	5.0 (3.5-6.3)	4.9 (3.2-6.2)	5.4 (4.3-8.0)	<.001
Potassium (mmol/L), mean (SD)	3.8 (0.5)	3.8 (0.4)	3.8 (0.5)	3.7 (0.5)	<.001
Natrium (mmol/L), mean (SD)	138.8 (5.1)	138.8 (3.7)	138.5 (3.9)	138.9 (6.0)	.08
Chloride (mmol/L), mean (SD)	101.9 (5.6)	101.5 (4.6)	101.2 (5.3)	102.4 (6.0)	<.001
Proteinuria[q], n (%)	1927 (30.2)	942 (36.7)	239 (37.2)	746 (23.6)	<.001

[a]CKD: chronic kidney disease.

[b]CVD: cardiovascular disease.

[c]The peak temperature on the day of admission.

[d]The immediate admission heart rate.

[e]The immediate admission respiratory rate, or if mechanical ventilation has been initiated, select the respiratory rate upon emergency department

XSL•FO
RenderX

admission.

[f]BP: blood pressure.

[g]The systolic blood pressure at admission, or if vasopressor medication has been administered, select the systolic blood pressure upon emergency department admission.

[h]the diastolic blood pressure at admission, or if vasopressor medication has been administered, select the diastolic blood pressure upon emergency department admission.

[i]ACEI: angiotensin-converting enzyme inhibitor.

[j]ARB: angiotensin receptor antagonists.

[k]CCB: calcium-channel blockers.

[l]Scr: serum creatinine.

[m]eGFR: estimated glomerular filtration rate.

[n]WBC: white blood cell.

[o]CRP: C-reactive protein.

[p]BUN: blood urea nitrogen.

[q]Proteinuria: positive urine protein.

Among the 6371 patients, 1006 (15.8%) developed AKI, with the median time for AKI diagnosis being day 3 (IQR 2-4) post admission. A comparison between patients with and without AKI is presented in Table 2. Patients with AKI exhibited higher age, temperature, heart rate, breathing rate, and lower diastolic blood pressure levels. They also had a higher incidence of hypertension, CKD, cardiovascular disease, cerebrovascular diseases, acute respiratory failure, and acute cardiac dysfunction and a higher rate of mechanical ventilation compared with those without AKI. Furthermore, patients with CAP having AKI demonstrated significantly elevated baseline Scr, white blood cell count, neutrophil count, C-reactive protein (CRP), and blood urea nitrogen (BUN) levels compared with those without AKI (all $P<.05$). Conversely, baseline eGFR, platelet count, hemoglobin, and blood albumin levels were substantially lower in the non-AKI group (all $P<.05$).

XSL•FO
RenderX

Table 2. Clinical features of patients in the AKI[a] and non-AKI groups.

Variables	All (n=6371)	AKI group (n=1006)	Non-AKI group (n=5365)	*P* value
Female, n (%)	2566 (40.2)	287 (28.5)	2279 (42.4)	<.001
Age (years), mean (SD)	67.0 (17.6)	77.3 (13.6)	65.1 (17.7)	<.001
Hypertension, n (%)	3007 (47.1)	617 (61.3)	2390 (44.5)	<.001
Diabetes, n (%)	1416 (22.2)	200 (19.8)	1216 (22.6)	.05
CKD[b], n (%)	420 (6.6)	213 (21.2)	207 (3.9)	<.001
CVD[c], n (%)	1322 (20.7)	347 (34.5)	975 (18.2)	<.001
Cerebrovascular disease, n (%)	977 (15.3)	192 (19.1)	785 (14.6)	<.001
Chronic lung disease, n (%)	1965 (30.8)	305 (30.3)	1660 (30.9)	.71
Cancer, n (%)	469 (7.4)	76 (7.5)	393 (7.3)	.79
Mechanical ventilation, n (%)	345 (5.4)	187 (18.6)	158 (2.9)	<.001
Temperature (℃), mean (SD)	37.8 (1.2)	38.0 (1.2)	37.7 (1.2)	.002
Heart Rate (beats/minute), mean (SD)	84.3 (14.9)	89.9 (18.8)	83.3 (13.9)	<.001
Breathing rate (births/minute), mean (SD)	19.6 (4.0)	21.6 (5.3)	19.2 (3.6)	<.001
Systolic BP[d] (mm Hg), mean (SD)	126.6 (18.3)	126.4 (19.7)	126.6 (18.0)	.71
Diastolic BP (mm Hg), mean (SD)	76.1 (10.9)	72.4 (13.7)	76.8 (10.2)	<.001
Acute respiratory failure, n (%)	598 (9.4)	200 (19.9)	398 (7.4)	<.001
Acute cardiac dysfunction, n (%)	354 (5.6)	125 (12.4)	229 (4.2)	<.001
ACEI[e] or ARB[f], n (%)	910 (14.3)	188 (18.7)	722 (13.5)	<.001
Diuretic, n (%)	1452 (22.7)	598 (59.4)	854 (15.9)	<.001
CCB[g], n (%)	1251 (19.6)	210 (20.8)	1041 (19.4)	.28
β-block, n (%)	364 (5.7)	62 (6.1)	302 (5.6)	.51
Baseline Scr[h] (μmol/L), mean (SD)	72.5 (31.1)	85.3 (52.1)	70.1 (24.6)	<.001
eGFR[i] (ml/min×1.73m^2), mean (SD)	84.9 (27.6)	71.9 (30.7)	87.4 (26.3)	<.001
WBC[j] (×10^9/L), median (IQR)	6.9 (5.2-9.5)	8.8 (6.3-12.7)	6.7 (5.1-9.0)	.03
Neutrophil count (×10^9/L), median (IQR)	4.9 (3.4-7.4)	7.2 (4.9-11.1)	4.6 (3.2-6.8)	.004
Platelet (×10^9/L), mean (SD)	210.0 (87.9)	180.4 (92.8)	215.5 (84.4)	<.001
CRP[k] (mg/L), median (IQR)	21.9 (4.0-74.2)	67.6 (18.8-131.0)	17.2 (3.2-57.8)	<.001
Hemoglobin (g/L), mean (SD)	121.1 (20.0)	112.1 (25.0)	122.8 (18.4)	<.001
Albumin (g/L), mean (SD)	32.9 (5.7)	29.5 (5.6)	33.6 (5.5)	<.001
BUN[l] (mmol/L), median (IQR)	5.1 (3.7-4.7)	9.5 (6.6-14.2)	4.8 (3.6-6.1)	<.001
Potassium (mmol/L), mean (SD)	3.8 (0.5)	3.9 (0.7)	3.7 (0.4)	<.001
Natrium (mmol/L), mean (SD)	138.8 (5.1)	138.9 (7.3)	138.8 (4.6)	.85
Chloride (mmol/L), mean (SD)	101.9 (5.6)	102.0 (5.6)	102.0 (5.1)	.84
Proteinuria, n (%)	1927 (30.2)	536 (53.3)	1391 (25.9)	<.001

[a]AKI: acute kidney injury.
[b]CKD, chronic kidney disease.
[c]CVD: cardiovascular disease.
[d]BP: blood pressure.
[e]ACEI: angiotensin-converting enzyme inhibitor.
[f]ARB: angiotensin receptor antagonists.
[g]CCB: calcium-channel blockers.

XSL•FO
RenderX

[h]Scr: serum creatinine.

[i]eGFR: estimated glomerular filtration rate.

[j]WBC: white blood cell.

[k]CRP: C-reactive protein.

[l]BUN: blood urea nitrogen.

Model Validation and Performance

In the training model, 9 variables with over 15% (9/62) missing values were excluded from the final cohort. These variables included blood lactic acid, ferritin, glycosylated hemoglobin C, blood oxygen partial pressure, blood carbon dioxide partial pressure, sputum pathogens, complement C3, complement C4, and procalcitonin. Among the remaining 53 variables, 3 had VIF values≥10. These variables were inhaled oxygen concentration, oxygen inhalation status, and vasoactive drug application. After eliminating variables with collinearity, the remaining 50 features (including age, sex, smoking status, and severe pneumonia), comorbidities (including history of chronic lung disease, CKD, malignancy, hypertension, diabetes, cardiovascular disease, chronic liver disease, and cerebrovascular disease), complications (such as septic shock, acute respiratory failure, and acute cardiac dysfunction), vital signs (including temperature, heart rate, respiratory rate, systolic blood pressure, and diastolic blood pressure), treatment on the day of admission (including noninvasive ventilation, mechanical ventilation, antiplatelet therapy, anticoagulation, renin-angiotensin-aldosterone system [RAAS] inhibitors, proton pump inhibitors, statins, diuretics, nonsteroidal anti-inflammatory drugs, insulin, calcium channel blockers, beta-blockers, and corticosteroids), and laboratory tests (including baseline Scr, baseline eGFR, CRP, urine protein, serum albumin, white blood cell count, platelet count, hemoglobin, neutrophil count, monocyte count, lymphocyte count, BUN, alanine aminotransferase [ALT], aspartate aminotransferase [AST], serum potassium, serum chloride, and serum sodium) were evaluated using the sliding windows sequential forward feature selection technique method.

The model generated an OOB value for each included feature. Figure 2 indicates that as the number of features increased to 11, the OOB value (0.579773) decreased substantially, and subsequently, the OOB value remained stable with increasing features. When the number of features reached 40, the OOB value (0.0340017) reached its lowest point. Therefore, to facilitate clinical application, we selected the first 11 features as the final model variables. These variables included sex, temperature, breathing rate, diastolic blood pressure, CRP, albumin, white blood cell count, platelet count, hemoglobin, neutrophil count, and BUN. Figure 3 illustrates that the BUN and neutrophil count consistently ranked among the top 5 indicators in all 5 models, suggesting their significant contribution to the occurrence of AKI.

Figure 2. Feature selection based on the sliding windows sequential forward feature selection technique process. The optimal number of features was determined according to the OOB value. A lower OOB value and fewer variables included results in the optimal combination of features (11 at the red circle). OOB: out-of-bag.

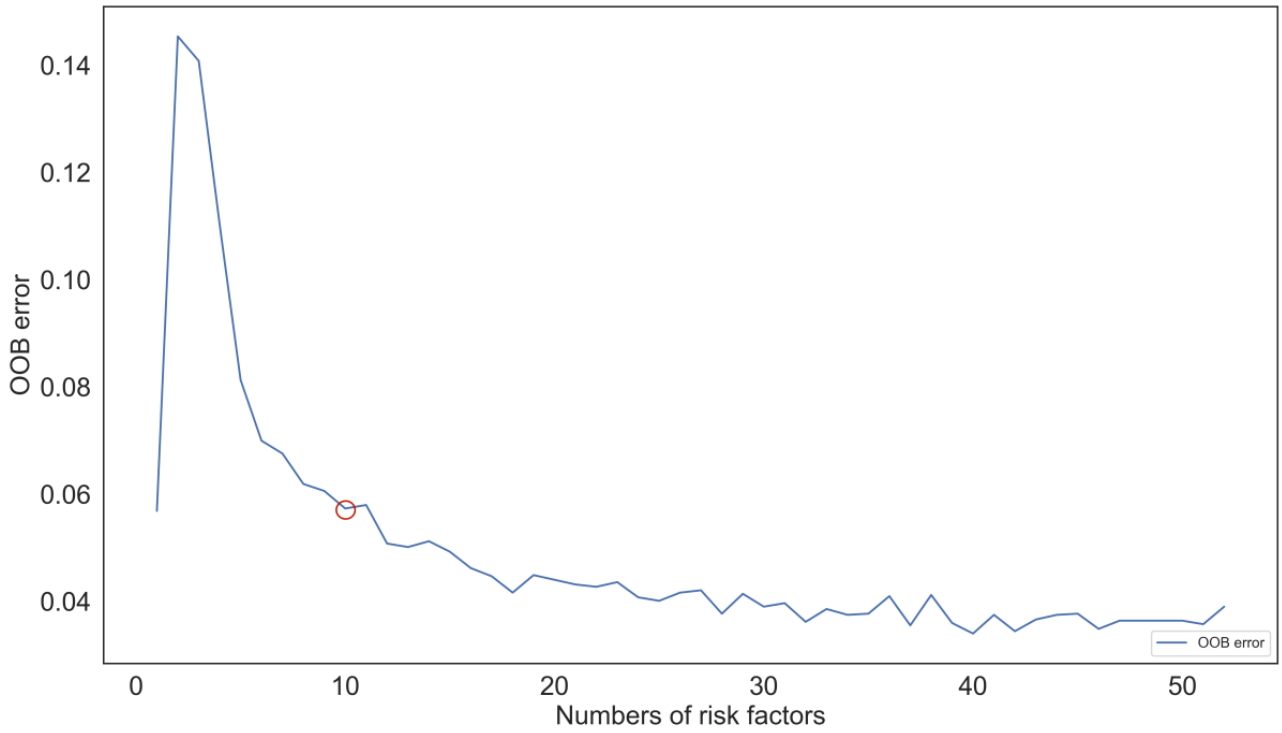

XSL•FO
RenderX

Figure 3. Feature importance derived from the machine learning models. (A) logistic regression model, (B) random forest model, (C) support vector machine model, (D) extreme gradient boosting model, and (E) deep forest model. BUN: blood urea nitrogen; CRP: C-reactive protein; DiasBP: diastolic blood pressure; WBC: white blood cells.

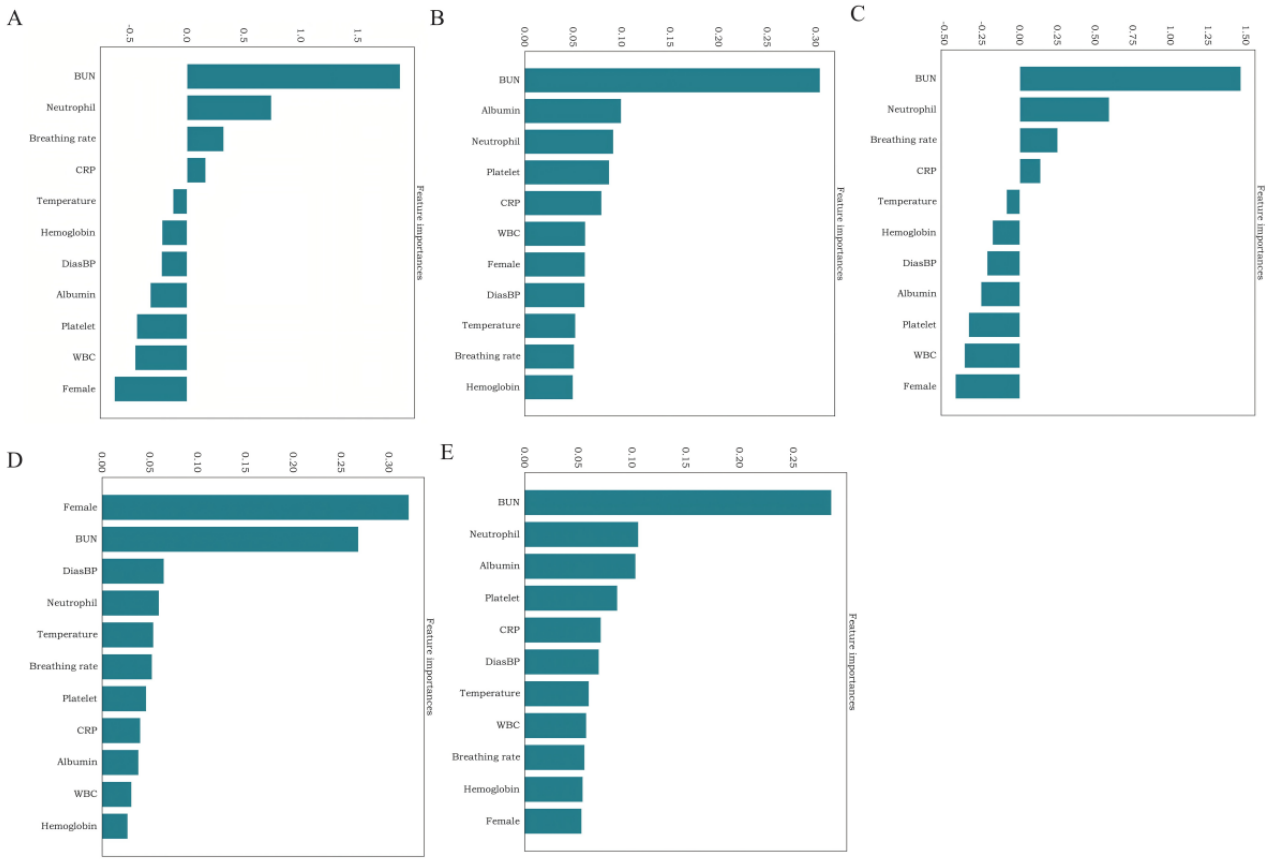

Internal Validation of the Model

A total of 643 patients were included in the internal validation, with an AKI incidence of 13.5% (n=87). Figure 4A illustrates the AUC of the 5 models in the internal validation set. Among them, the DF model exhibited the highest discrimination compared with the other 4 models, achieving an AUC of 0.89 (95% CI 0.84-0.92). The XGBoost model followed closely with an AUC of 0.88 (95% CI 0.84-0.92). Detailed prediction performance of the 5 ML models is presented in Table 3. The DF model demonstrated the best specificity prediction for AKI (0.99, 95% CI 0.98-0.99). Compared with the other models except DF, the XGBoost model showed higher accuracy (0.92, 95% CI 0.90-0.94), PPV (0.78, 95% CI 0.68-0.86), and F_1-score (0.63, 95% CI 0.54-0.70). Calibration curves of the 5 models are depicted in Figure 4B. The results indicated that the DF model exhibited the lowest ECE value of 0.03, signifying superior calibration. The XGBoost model followed with an ECE of 0.05. Decision curve analysis of the prediction models is presented in Figure 4C. The findings revealed that the DF, XGBoost, and RF models yielded net benefits within the threshold range of 0-0.8. In the internal validation set, the DF model demonstrated the highest predictive ability for AKI in patients with CAP.

XSL•FO
RenderX

Figure 4. The performance of machine learning models in the internal validation set. (A) area under receiver operating characteristic curve, (B) calibration curve, and (C) decision curve analysis curve. AUC: area under the receiver operating characteristic curve; DF: deep forest; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBT: extreme gradient boosting.

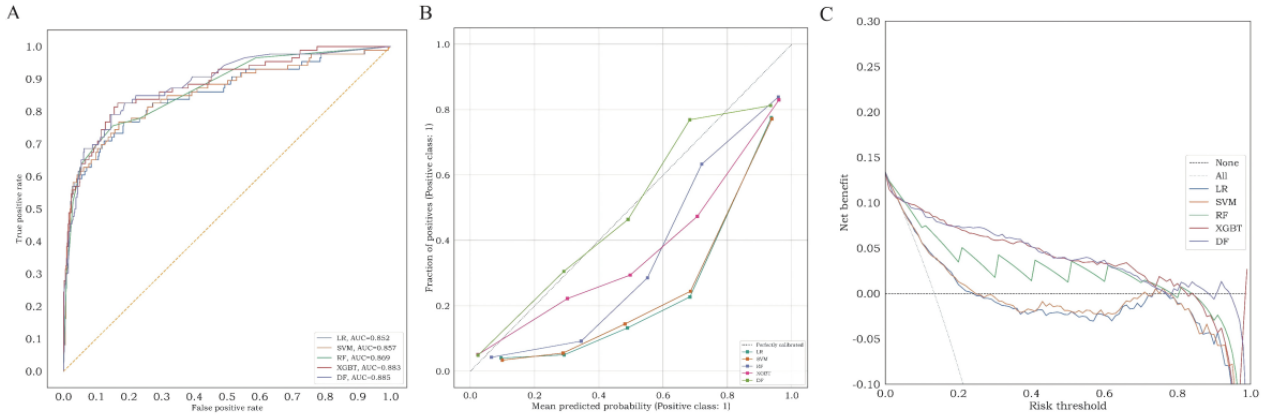

Table 3. Detailed prediction performance of 5 machine learning models in the internal validation set.

Models	AUC[a], (95% CI)	Accuracy, (95% CI)	Sensitivity, (95% CI)	Specificity, (95% CI)	PPV[b], (95% CI)	NPV[c], (95% CI)	F_1-score, (95% CI)	ECE[d]
LR[e]	0.85 (0.81-0.90)	0.75 (0.72-0.77)	0.81 (0.74-0.88)	0.73 (0.70-0.77)	0.32 (0.27-0.37)	0.96 (0.95-0.98)	0.46 (0.40-0.52)	0.12
SVM[f]	0.86(0.81-0.90)	0.76 (0.72-0.78)	0.79 (0.71-0.87)	0.75 (0.72-0.78)	0.33 (0.28-0.38)	0.96 (0.94-0.97)	0.46 (0.40-0.52)	0.12
RF[g]	0.87 (0.83-0.90)	0.91 (0.89-0.93)	0.52 (0.43-0.61)	0.97 (0.96-0.98)	0.74 (0.65-0.83)	0.93 (0.91-0.95)	0.61 (0.53-0.69)	0.08
XGBoost[h]	0.88 (0.84-0.92)	0.92 (0.90-0.94)	0.52 (0.43-0.61)	0.98 (0.97-0.99)	0.78 (0.68-0.86)	0.93 (0.91-0.95)	0.63 (0.54-0.70)	0.05
DF[i]	0.89 (0.84-0.92)	0.90 (0.88-0.91)	0.31 (0.23-0.39)	0.99 (0.98-0.99)	0.77 (0.65-0.89)	0.90 (0.88-0.92)	0.45 (0.34-0.53)	0.03

[a]AUC: area under the receiver operating characteristic curve.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]ECE: expected calibration error.
[e]LR: logistic regression.
[f]SVM: support vector machine.
[g]RF: random forest.
[h]XGBoost: extreme gradient boosting.
[i]DF: deep forest.

External Validation of the Model

The data for the external validation set were obtained from Nanjing First Hospital and Sir Run Run Hospital affiliated with Nanjing Medical University, encompassing 3160 patients with CAP. The incidence of AKI was 20.4% (n=645) and the baseline clinical features are outlined in Table 1. Figure 5A presents the AUC of the 5 models in the external validation set. Among them, the DF model exhibited the most favorable predictive performance, achieving an AUC of 0.87 (95% CI 0.85-0.88), specificity of 0.93 (95% CI 0.93-0.94), and PPV of 0.63 (95% CI 0.59-0.67). The XGBoost model closely followed with an AUC of 0.86 (95% CI 0.85-0.88), accuracy of 0.84 (95% CI 0.83-0.85), and F_1-score of 0.65 (95% CI 0.63-0.68). Detailed performance metrics are presented in Table 4. Calibration curves of the prediction models are depicted in Figure 5B. The results demonstrated that the DF model exhibited the best calibration ability in the external validation set. Figure 5C presents the decision curve analysis curves of the 5 models, indicating that the DF, XGBoost, and RF models yielded net benefits within the threshold range of 0-0.6. In the external validation set, the DF model showcased the highest predictive ability for AKI in patients with CAP.

XSL•FO
RenderX

Figure 5. The performance of machine learning models in the external validation set. (A) area under receiver operating characteristic curve, (B) calibration curve, (C) decision curve analysis curve. AUC: area under the receiver operating characteristic curve; DF: deep forest; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBT: extreme gradient boosting.

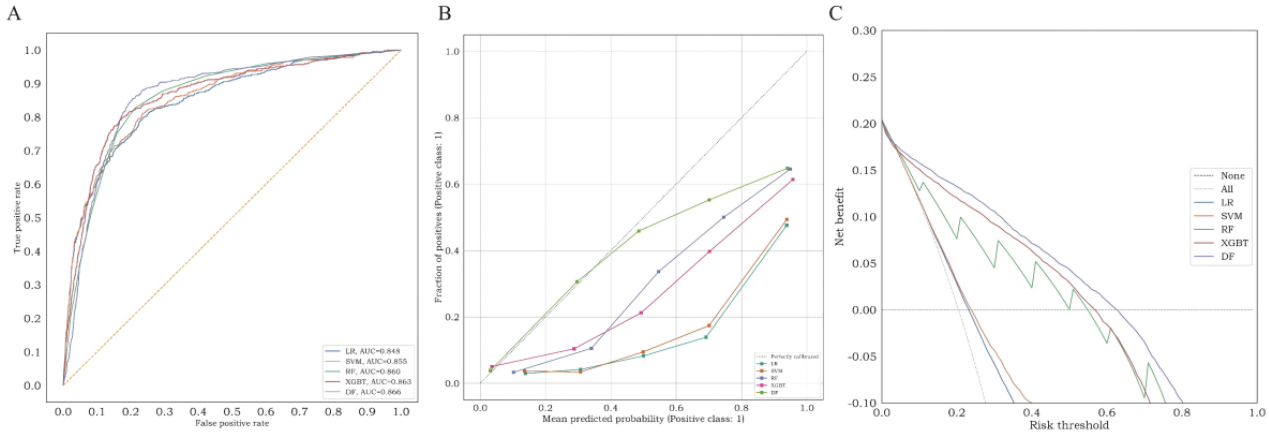

Table 4. Detailed prediction performance of 5 machine learning models in external validation sets.

Models	AUC[a], (95% CI)	Accuracy, (95% CI)	Sensitivity, (95% CI)	Specificity, (95% CI)	PPV[b], (95% CI)	NPV[c], (95% CI)	F_1-score, (95% CI)	ECE[d]
LR[e]	0.85 (0.83-0.86)	0.45 (0.43-0.46)	0.97 (0.95-0.98)	0.32 (0.30-0.33)	0.27 (0.25-0.28)	0.97 (0.96-0.98)	0.42 (0.40-0.44)	0.38
SVM[f]	0.86 (0.84-0.87)	0.51 (0.49-0.52)	0.95 (0.94-0.96)	0.39 (0.38-0.41)	0.29 (0.27-0.30)	0.97 (0.96-0.98)	0.44 (0.42-0.46)	0.36
RF[g]	0.86 (0.85-0.87)	0.83 (0.82-0.84)	0.65 (0.62-0.68)	0.88 (0.87-0.90)	0.58 (0.55-0.61)	0.91 (0.90-0.92)	0.61 (0.59-0.64)	0.16
XGBoost[h]	0.86 (0.85-0.88)	0.84 (0.83-0.85)	0.73 (0.70-0.76)	0.87 (0.86-0.88)	0.59 (0.56-0.62)	0.93 (0.92-0.94)	0.65 (0.63-0.68)	0.13
DF[i]	0.87 (0.85-0.88)	0.83 (0.82-0.84)	0.44 (0.41-0.47)	0.93 (0.93-0.94)	0.63 (0.59-0.67)	0.87 (0.86-0.88)	0.52 (0.49-0.55)	0.05

[a]AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]ECE: expected calibration error.

[e]LR: logistic regression.

[f]SVM: support vector machine.

[g]RF: random forest.

[h]XGBoost: extreme gradient boosting.

[i]DF: deep forest.

SHAP Values Evaluate Feature Importance

We used SHAP to determine the importance ranking of the 11 features, as shown in Figure 6A. The significance of SHAP features is depicted in Figure 6B, where the x-axis represents the positive or negative contribution of a single feature to the model. The y-axis displays the 11 features in descending order of importance based on the absolute value of SHAP. A higher absolute value of SHAP indicates a greater contribution of the feature to AKI risk. The results revealed that BUN, neutrophil count, CRP, white blood cell count, breathing rate, and temperature made positive contributions, implying that higher values increased the risk of AKI. Conversely, female sex, albumin, platelet count, diastolic blood pressure, and hemoglobin made negative contributions, indicating that higher values decreased the risk of AKI.

XSL•FO
RenderX

Figure 6. Shapley additive explanations (SHAP) interpretation and visualization of the deep forest model. (A) SHAP analysis shows a vital ranking of 11 features, and (B) SHAP analysis shows point estimation of features to model output. Each point represents a single patient in the data set. BUN: blood urea nitrogen; CRP: C-reactive protein; DiasBP: diastolic blood pressure; WBC: white blood cells.

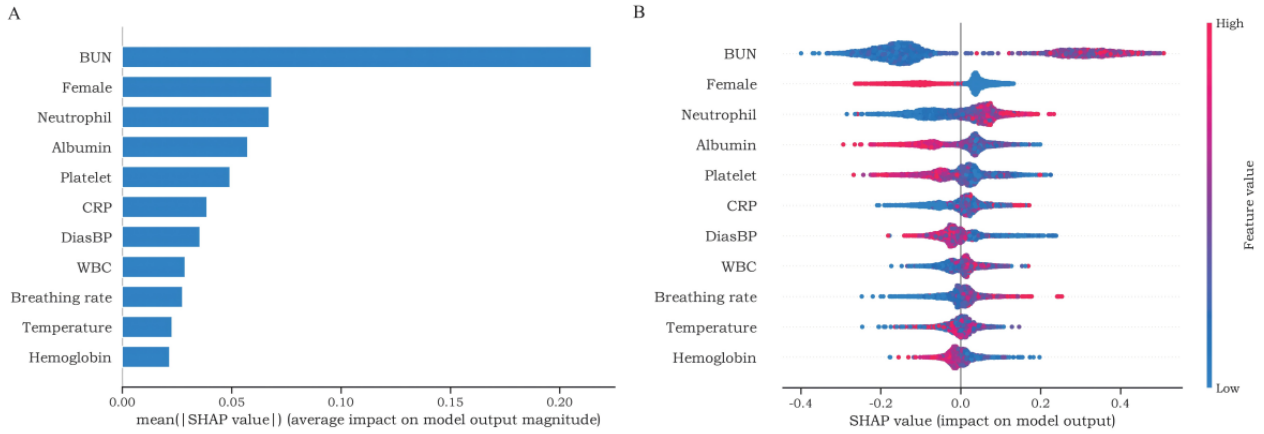

In addition to explaining the overall contribution of the model using SHAP, we used LIME to individualize the contribution of the last 11 features in the prediction process. Figure 7A illustrates the predictive process for a patient without AKI. First, the 11 patient characteristics were ranked by importance. Blue indicates the contribution to the prediction as non-AKI, orange represents the contribution as AKI, and the corresponding prediction probability is displayed on the left. The specific values of the patient's 11 features are presented on the right. In this case, the patient had a BUN level of 5.0 mmol/L, neutrophil count of 9.52×10^9/L, breathing rate of 20 breaths/min, temperature of 36.8°C, white blood cell count of 10.6×10^9/L, albumin level of 24.6 g/L, CRP level of 0.5 mg/L, diastolic blood pressure level of 85 mm Hg, hemoglobin level of 110 g/L, and platelet count of 161×10^9/L. The probability of predicting non-AKI for this patient was 1.0.

Figure 7. Local interpretable model-agnostic explanations (LIME) explaining individual's prediction results. The left part of the figure shows predicted results using LIME. The middle part presents the impact of the 11 variables on acute kidney injury (AKI) or no AKI. The length of the bar for each feature indicates the importance (weight) of that feature in making the prediction. A longer bar indicates a feature that contributes more to (A) No AKI or (B) AKI. The right panel shows the critical values of these 11 variables when they had the greatest impact on (A) No AKI or (B) AKI. AKI: acute kidney injury; BUN: blood urea nitrogen; CRP: C-reactive protein; DiasBP: diastolic blood pressure; SHAP: SHapley Additive exPlanations; WBC: white blood cells.

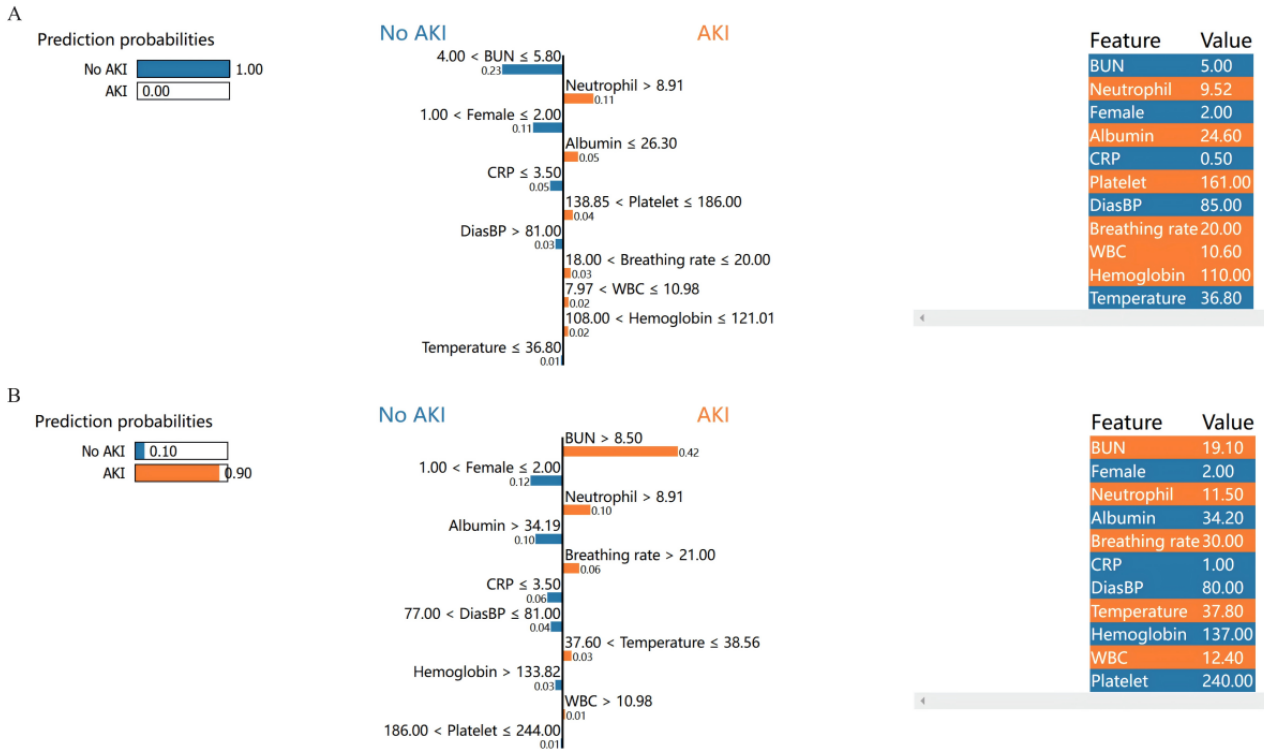

Furthermore, Figure 7B demonstrates the predictive process for a patient with AKI. With a BUN level of 19.1 mmol/L, neutrophil count of 11.5×10^9/L, breathing rate of 30 breaths/min, temperature of 37.8°C, white blood cell count of 12.4×10^9/L,

XSL•FO
RenderX

albumin level of 34.2 g/L, CRP level of 1 mg/L, diastolic blood pressure level of 80 mm Hg, hemoglobin level of 137 g/L, and platelet count of 240×10^9/L, the probability of predicting AKI for this patient was 0.9.

Establishment of a Web-Based Prediction Platform

We developed a web-based prediction platform specifically designed for predicting AKI in patients with CAP (Figure 8). This platform uses the 11 selected variables as inputs to generate a risk value for predicting CAP-AKI. If the predicted risk value exceeds the predefined threshold of 74.5% set by the system, the result is considered positive, indicating that the patient falls into the high-risk category for developing CAP-AKI.

Figure 8. Web-based prediction platform of community-acquired pneumonia–associated acute kidney injury (CAP-AKI). (A) When 11 variables of a patient with community-acquired pneumonia were entered, the risk of CAP-AKI was calculated to be 56.25 %, and the result was 0, that is, the patient was not a high-risk group of CAP-AKI. (B) When 11 variables of a patient with community-acquired pneumonia were entered, the risk of CAP-AKI was calculated to be 82.75 %, and the result was 1, that is, the patient was a high-risk group of CAP-AKI. We set the risk threshold of our web calculator was 74.50% based on the best F1-score. When the predictive probability exceeds the risk threshold, the output is 1, otherwise the output is 0.

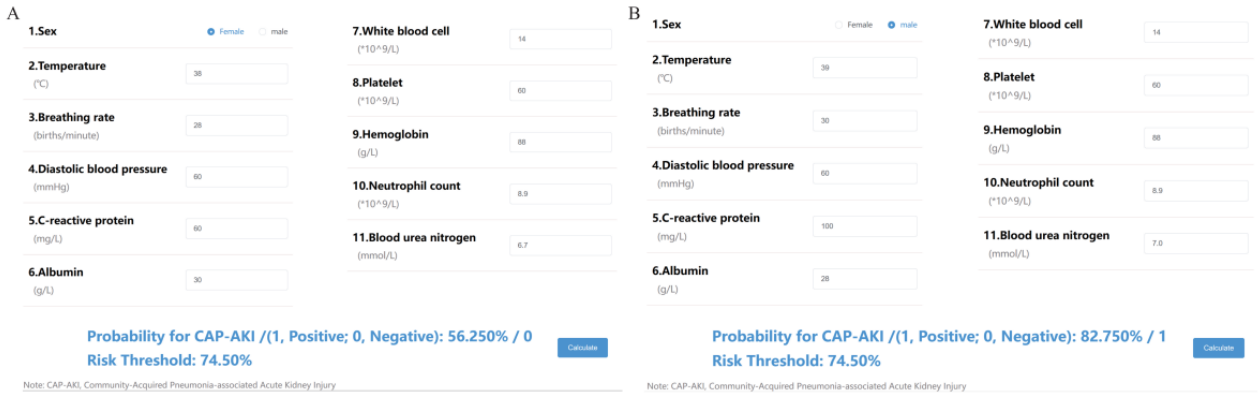

Discussion

Principal Findings

In this study, we used ML algorithms to develop an innovative prediction tool for CAP-AKI. CAP stands as a prevalent infectious disease globally, entailing considerable morbidity and mortality [11-13]. Among the array of complications linked with CAP, AKI has surfaced as a notable concern [14,15]. Our study unveils a noteworthy incidence of 15.8% for AKI among patients with CAP, indicating a substantial incidence of AKI within this population. Pinpointing the risk factors associated with AKI and crafting a predictive model are pivotal endeavors, enabling effective risk stratification and timely intervention for AKI. In this study, we identified the top 11 important predictors, including BUN, neutrophil count, respiratory rate, CRP, temperature, albumin, diastolic blood pressure, hemoglobin, platelet count, white blood cell count, and sex, as these variables can be easily collected in medical practice. In addition, our DF model showed good performance in both internal (AUC 0.89, 95% CI 0.84-0.92) and external (AUC 0.87, 95% CI 0.85-0.88) validation sets, indicating strong generalization ability.

ML techniques offer a promising path by harnessing the extensive pool of clinical data to offer insightful early-stage predictions of AKI. This approach potentially eliminates the necessity for additional AKI biomarker testing [16,17]. Hsu et al [18] constructed an AKI prediction model using 10 preprocedural indicators through an XGBoost algorithm, revealing an AUC of 0.767 on derivation and 0.761 on validation for any stage of AKI. Other researchers also found that XGBoost [19] and recurrent neural network [20] could perform well in predicting cardiac surgery–associated AKI and in-hospital AKI.

Due to the distinct etiology of AKI in patients with CAP compared with other causes such as cardiac surgery–related AKI or sepsis-induced AKI, there are differences in the characteristics and factors required for model establishment. Some unique features need to be incorporated, including respiratory status, body temperature, blood pressure, and inflammatory markers, which have a significant impact on the occurrence of AKI. These parameters may differ from those in AKI models related to other etiologies. This aspect has been further elaborated in the discussion section.

In this study, we used ML algorithms to develop an innovative prediction tool for CAP-AKI. Furthermore, we determined the top 11 important predictors in the DF model as prediction model variables as these variables can be collected easily in medical activities. These features comprehensively consider the patient's basic characteristics (such as age and sex), disease status (such as blood pressure and Scr), and potential causes (such as infections and chronic diseases), providing a well-rounded informational foundation for the model from multiple perspectives. Similarly, AUC for AKI in patients with CAP of the internal validation and external validation cohorts was shown to be 0.86 (95% CI 0.84-0.92) and 0.87 (95% CI 0.85-0.88), respectively. In addition, we constructed SHAP to provide personalized interpretation for each patient. Subsequently, an online web risk calculator model for AKI in patients with CAP was established to predict the occurrence of AKI within 7 days upon patients' hospital admission. The median time for diagnosing AKI in our study was the third day. Although there was a certain time difference between the diagnosis time of AKI and the time of the indicators we included, our results showed that the model still exhibited good predictive performance in the validation set, and we believed that this time difference

XSL•FO
RenderX

could be used as the best time for early intervention in high-risk patients who may develop AKI.

Among the 5 ML models used in this study, LR is commonly used in clinical settings. However, its performance is suboptimal when dealing with large-sample and high-dimensional data. RF exhibits good generalization ability and can prevent overfitting. XGBoost derived from RF is unaffected by multicollinearity and possesses characteristics such as flexibility and efficiency. DF is a novel ensemble-based method proposed in recent years, serving as an alternative to deep neural networks. It combines various ensemble-based methods with indivisible modules. Compared with deep neural network, DF has the advantage of achieving the optimal model with a small amount of training data, without the need for backpropagation, and has low computational costs. Zhang et al [21] found that the DF model achieved an AUC of 0.881 for predicting AKI after cardiac surgery, outperforming other ML models, including LR, XGBoost, and RF. Since its proposal, DF has been widely applied and has demonstrated good predictive performance in various fields such as radiological diagnosis of COVID-19 and emotion recognition [22,23].

In this study, we conducted a ranking analysis of the included feature variables and observed that several factors, namely BUN, neutrophil count, respiratory rate, CRP, temperature, albumin, diastolic blood pressure, hemoglobin, platelet count, white blood cell count, and sex, exerted a significant impact on the final model. Among these features, BUN emerged as the most crucial indicator for AKI. Similar to Scr, BUN serves as a reflection of renal function and carries substantial predictive weight for AKI occurrence. These findings align with the diagnostic criteria recommended by the KDIGO and are consistent with previous clinical investigations. Elevated levels of white blood cells, neutrophils, and CRP are also risk factors for AKI in patients. AKI is closely associated with renal and systemic inflammation. In the presence of infection, inflammatory factors play a critical role in combating microbial pathogens and facilitating tissue repair. White blood cells, including neutrophils, can infiltrate the injured kidneys through the circulatory system, triggering the production of cytokines, chemokines, and other inflammatory mediators, ultimately leading to renal damage [24]. CRP, as a plasma protein synthesized in the liver, serves as a nonspecific marker for systemic inflammatory response during the acute phase and is closely related to the severity of patients with CAP. Consistent with the findings of Wang et al [25] in patients with AKI after myocardial infarction, we also found that elevated CRP levels were a risk factor for AKI in the cohort of patients with CAP. Hypoalbuminemia is another risk factor for AKI in patients with CAP. Animal experiments have shown that albumin can counteract the decline in arterial vascular reactivity caused by endotoxemia, alleviate ischemia-reperfusion injury, and possess anti-inflammatory effects. Therefore, reduced albumin levels significantly reduce a patient's resistance to infection and stress damage [26]. Similar conclusions have been reached by Bang et al [27]. The relationship between hemoglobin levels and AKI has been widely studied, with anemia being an independent risk factor for AKI [28-30]. The findings of our study provide compelling evidence that platelet levels act as a protective factor for AKI in patients with CAP. These results align with the previous investigations conducted by Zhang et al [31], thereby reinforcing the notion that a decreased platelet count is significantly associated with an elevated risk of AKI occurrence.

The prediction process of ML algorithms can be likened to a "black box" [32], where the inner workings may not be readily understandable to clinical physicians. Therefore, it is crucial for physicians to not solely rely on predictive models but also use their expertise in making appropriate diagnostic and therapeutic decisions. In this study, we used SHAP and LIME techniques to visualize the overall and individual contributions of each indicator to the model's predictive performance, providing additional insights to clinical physicians. By quantifying the individual risks of AKI occurrence through a web-based prediction platform, we enhanced the clinical utility of the predictive model and demonstrated its significant potential in AKI risk prediction.

In this study, a comprehensive model was developed to predict the occurrence of AKI in patients with CAP by integrating multiple feature variables, resulting in robust predictive performance in both internal and external datasets. In addition, to enhance the interpretability of the best-performing DF model, visual explanations were provided using SHAP and LIME techniques.

Furthermore, a web-based prediction platform was established to augment the clinical utility of the predictive model, offering practical applications in real-world health care settings. The threshold of 74.5% was chosen based on the method of selecting the maximum F_1-score on the training set. In a real-world clinical setting, a model with low sensitivity means that a significant number of patients with AKI might not be identified. This could lead to missed opportunities for early intervention and treatment, potentially worsening patient outcomes. However, the higher PPV suggests that when the model does predict AKI, it is more likely to be correct. This reduces the burden of false positives on clinicians, who might otherwise need to follow up on many non-AKI cases, saving time and resources.

The choice of threshold reflects a preference for higher precision, which might be suitable in situations where the cost of false positives is high, such as in scenarios with limited resources or when unnecessary treatments pose risks to patients. However, the trade-off with lower sensitivity means clinicians must be aware that not all cases of AKI will be caught by the model, and they should continue to use their clinical judgment and other diagnostic tools to ensure comprehensive patient care. Remarkably, the DF model outperformed others in both the internal and external validation cohorts.

However, it is important to acknowledge the limitations of this study. First, it should be noted that this was a retrospective study, which may have introduced certain limitations in terms of the completeness, accuracy, and homogeneity of data recording compared with prospective studies. In addition, the presence of selection bias cannot be ruled out, highlighting the need for further validation through prospective research. Second, the collection of urine output data posed challenges, resulting in cases of incomplete data collection. Furthermore, the

XSL•FO
RenderX

consideration of urine output standards was not incorporated in this study, which could have provided additional valuable insights. Finally, while the constructed risk prediction model shows promise, it remains uncertain whether its implementation in clinical practice would translate into tangible benefits for patients. Therefore, prospective and multicenter studies are necessary to comprehensively evaluate the effectiveness and real-world impact of the model. Acknowledging and addressing these limitations will be essential in further refining the understanding and application of the findings in future research.

In summary, this study presents a valuable early prediction tool utilizing ML techniques to identify patients with high risk of AKI among patients with CAP. The developed predictive model has the potential to aid physicians in the early identification of patients at a higher risk of AKI during their hospitalization, enabling prioritized treatment interventions and ultimately improving patient outcomes. Moving forward, our future research endeavors will focus on prospective evaluation of the effectiveness of our AKI prediction model and determining its impact on the prognosis of patients with AKI in real-world clinical practice. In addition, the model could be applied to other etiologies of AKI, such as sepsis-induced AKI, to assess its generalizability.

Conclusion

This study developed a ML-based prediction model for CAP-AKI, demonstrating excellent predictive performance in both internal and external validation cohorts. By integrating easily accessible clinical variables, the DF model achieved high accuracy and interpretability, offering robust support for early risk assessment and targeted intervention in patients with CAP-AKI. In the future, we hope to reduce the risk of CAP-AKI through our online prediction platform.

Acknowledgments

The authors acknowledge the technical support of the information query department of Wuxi People's Hospital, Sir Run Run Hospital, and Nanjing First Hospital.

This study was supported by the National Science Foundation of China (82170698), the major project of Jiangsu Provincial Health Commission Medical Research (ZDB2020023), the general project of Jiangsu Provincial Health Commission Medical Research (H2023057), and Nanjing Municipal Health Science and Technology Development Special Fund Key Project (ZKX22035).

Authors' Contributions

Research idea and study design were contributed by MM, C Cao, and XW. Data acquisition was contributed by C Chen, DC, QS, HZ, XD, LF, XC, and HK. Data analysis and interpretation was performed by MM, C Chen, and DX. Supervision or mentorship was performed by C Cao and XW. Each author contributed important intellectual content during manuscript drafting or revision and agrees to be personally accountable for the individual's own contributions and to ensure that questions pertaining to the accuracy or integrity of any portion of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of the machine learning methods.
[DOCX File , 13 KB - jmir_v26i1e51255_app1.docx]

References

1. Chawla L, Amdur R, Faselis C, Li P, Kimmel P, Palant C. Impact of acute kidney injury in patients hospitalized with pneumonia. Crit Care Med 2017;45(4):600-606. [doi: 10.1097/CCM.0000000000002245] [Medline: 28291091]
2. Lin TY, Chen YG, Lin CL, Kao CH. Increased Risk of Acute Kidney Injury following Pneumococcal Pneumonia: A Nationwide Cohort Study. PLoS One 2016;11(6):e0158501 [FREE Full text] [doi: 10.1371/journal.pone.0158501] [Medline: 27362355]
3. Srisawat N, Murugan R, Lee M, Kong L, Carter M, Angus DC, GeneticInflammatory Markers of Sepsis (GenIMS) Study Investigators. Plasma neutrophil gelatinase-associated lipocalin predicts recovery from acute kidney injury following community-acquired pneumonia. Kidney Int 2011 Sep;80(5):545-552 [FREE Full text] [doi: 10.1038/ki.2011.160] [Medline: 21677634]
4. Fiorentino M, Tohme FA, Wang S, Murugan R, Angus DC, Kellum JA. Long-term survival in patients with septic acute kidney injury is strongly influenced by renal recovery. PLoS One 2018;13(6):e0198269 [FREE Full text] [doi: 10.1371/journal.pone.0198269] [Medline: 29870535]
5. Vijayan A, Abdel-Rahman EM, Liu KD, Goldstein SL, Agarwal A, Okusa MD, et al. Recovery after critical illness and acute kidney injury. Clin J Am Soc Nephrol 2021;16(10):1601-1609 [FREE Full text] [doi: 10.2215/CJN.19601220] [Medline: 34462285]

XSL•FO
RenderX

6. Alge JL, Arthur JM. Biomarkers of AKI: a review of mechanistic relevance and potential therapeutic implications. Clin J Am Soc Nephrol 2015 Jan 07;10(1):147-155 [FREE Full text] [doi: 10.2215/CJN.12191213] [Medline: 25092601]
7. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
8. Kellum JA, Lameire N, KDIGO AKI Guideline Work Group. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). Crit Care 2013;17(1):204 [FREE Full text] [doi: 10.1186/cc11454] [Medline: 23394211]
9. Yang L, Xing G, Wang L, Wu Y, Li S, Xu G, et al. Acute kidney injury in China: a cross-sectional survey. Lancet 2015 Oct 10;386(10002):1465-1471. [doi: 10.1016/S0140-6736(15)00344-X] [Medline: 26466051]
10. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 05;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
11. Rothberg MB. Community-Acquired Pneumonia. Ann Intern Med 2022 Apr;175(4):ITC49-ITC64. [doi: 10.7326/AITC202204190] [Medline: 35404672]
12. Niederman M, Torres A. Severe community-acquired pneumonia. Eur Respir Rev 2022;31(166):220123 [FREE Full text] [doi: 10.1183/16000617.0123-2022] [Medline: 36517046]
13. Peghin M, Bouza E. Community-acquired pneumonia: is less more? Lancet Infect Dis 2022;22(2):159-161. [doi: 10.1016/S1473-3099(21)00401-1] [Medline: 34627497]
14. Hoşgün D, Aydemir S. Factors affecting 90-day mortality in community and hospital acquired pneumonia patients with or without acute kidney injury. Afr Health Sci 2022;22(3):567-577 [FREE Full text] [doi: 10.4314/ahs.v22i3.61] [Medline: 36910350]
15. Latief M, Ahmed Para R, Shafi O, Hassan Z, Farooq S, Abbas F. Incidence and risk factors of acute kidney injury in patients hospitalized with pneumonia: a prospective observational study. Med J Islam Repub Iran 2021;35:150 [FREE Full text] [doi: 10.47176/mjiri.35.150] [Medline: 35321362]
16. Dong J, Feng T, Thapa-Chhetry B, Cho B, Shum T, Inwald D, et al. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. Crit Care 2021;25(1):288 [FREE Full text] [doi: 10.1186/s13054-021-03724-0] [Medline: 34376222]
17. Yue S, Li S, Huang X, Liu J, Hou X, Zhao Y, et al. Machine learning for the prediction of acute kidney injury in patients with sepsis. J Transl Med 2022;20(1):215 [FREE Full text] [doi: 10.1186/s12967-022-03364-0] [Medline: 35562803]
18. Hsu C, Liu C, Tain Y, Kuo C, Lin Y. Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: development and validation study. J Med Internet Res 2020;22(8):e16903 [FREE Full text] [doi: 10.2196/16903] [Medline: 32749223]
19. Luo X, Kang Y, Duan S, Yan P, Song G, Zhang N, et al. Machine learning-based prediction of acute kidney injury following pediatric cardiac surgery: model development and validation study. J Med Internet Res 2023;25:e41142 [FREE Full text] [doi: 10.2196/41142] [Medline: 36603200]
20. Kim K, Yang H, Yi J, Son H, Ryu J, Kim Y, et al. Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: external validation and model interpretation. J Med Internet Res 2021;23(4):e24120 [FREE Full text] [doi: 10.2196/24120] [Medline: 33861200]
21. Zhang H, Wang Z, Tang Y, Chen X, You D, Wu Y, et al. Prediction of acute kidney injury after cardiac surgery: model development using a Chinese electronic health record dataset. J Transl Med 2022;20(1):166 [FREE Full text] [doi: 10.1186/s12967-022-03351-5] [Medline: 35397573]
22. Sun L, Mo Z, Yan F, Xia L, Shan F, Ding Z, et al. Adaptive feature selection guided deep forest for COVID-19 classification with chest CT. IEEE J Biomed Health Inform 2020;24(10):2798-2805 [FREE Full text] [doi: 10.1109/JBHI.2020.3019505] [Medline: 32845849]
23. Cheng J, Chen M, Li C, Liu Y, Song R, Liu A, et al. Emotion recognition from multi-channel EEG via deep forest. IEEE J Biomed Health Inform 2021;25(2):453-464. [doi: 10.1109/JBHI.2020.2995767] [Medline: 32750905]
24. Akcay A, Nguyen Q, Edelstein C. Mediators of inflammation in acute kidney injury. Mediators Inflamm 2009;2009:137072 [FREE Full text] [doi: 10.1155/2009/137072] [Medline: 20182538]
25. Wang W, Xie Y, Huang X, Zhou Y, Luo L. The value of N-terminal pro-brain natriuretic peptide and hs-CRP in predicting acute kidney injury after acute myocardial infarction. Am J Transl Res 2022;14(8):5501-5510 [FREE Full text] [Medline: 36105028]
26. Artigas A, Wernerman J, Arroyo V, Vincent J, Levy M. Role of albumin in diseases associated with severe systemic inflammation: pathophysiologic and clinical evidence in sepsis and in decompensated cirrhosis. J Crit Care 2016;33:62-70. [doi: 10.1016/j.jcrc.2015.12.019] [Medline: 26831575]
27. Bang J, Kim S, Kim S, Song J, Kang J, Kim J, et al. Impact of the serum albumin level on acute kidney injury after cerebral artery aneurysm clipping. PLoS One 2018;13(11):e0206731 [FREE Full text] [doi: 10.1371/journal.pone.0206731] [Medline: 30395651]

XSL•FO
RenderX

28. Tseng P, Chen Y, Wang C, Chiu K, Peng Y, Hsu S, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020;24(1):478 [FREE Full text] [doi: 10.1186/s13054-020-03179-9] [Medline: 32736589]
29. Kuno T, Numasawa Y, Mikami T, Niimi N, Sawano M, Kodaira M, et al. Association of decreasing hemoglobin levels with the incidence of acute kidney injury after percutaneous coronary intervention: a prospective multi-center study. Heart Vessels 2021;36(3):330-336. [doi: 10.1007/s00380-020-01706-w] [Medline: 33034713]
30. Kim-Campbell N, Gretchen C, Callaway C, Felmet K, Kochanek P, Maul T, et al. Cell-free plasma hemoglobin and male gender are risk factors for acute kidney injury in low risk children undergoing cardiopulmonary bypass. Crit Care Med 2017;45(11):e1123-e1130 [FREE Full text] [doi: 10.1097/CCM.0000000000002703] [Medline: 28863013]
31. Zhang C, Wang G, Zhou H, Lei G, Yang L, Fang Z, et al. Preoperative platelet count, preoperative hemoglobin concentration and deep hypothermic circulatory arrest duration are risk factors for acute kidney injury after pulmonary endarterectomy: a retrospective cohort study. J Cardiothorac Surg 2019;14(1):220 [FREE Full text] [doi: 10.1186/s13019-019-1026-4] [Medline: 31888760]
32. Seetharam K, Balla S, Bianco C, Cheung J, Pachulski R, Asti D, et al. Applications of machine learning in cardiology. Cardiol Ther 2022;11(3):355-368 [FREE Full text] [doi: 10.1007/s40119-022-00273-7] [Medline: 35829916]

Abbreviations

AKI: acute kidney injury
ALT: alanine aminotransferase
AST: aspartate aminotransferase
AUC: area under the receiver operating characteristic curve
BUN: blood urea nitrogen
CAP: community-acquired pneumonia
CAP-AKI: community-acquired pneumonia–associated acute kidney injury
CKD: chronic kidney disease
CKD-EPI: Chronic Kidney Disease Epidemiology Collaboration
CRP: C-reactive protein
DF: deep forest
ECE: expected calibration error
eGFR: estimated glomerular filtration rate
KDIGO: Kidney Disease Improving Global Outcomes
LIME: local interpretable model-agnostic explanation
LR: logistic regression
MIMIC-III: The Medical Information Mart for Intensive Care III
ML: machine learning
OOB: out-of-bag
PPV: positive predictive value
RAAS: renin-angiotensin-aldosterone system
RF: random forest
Scr: serum creatinine
SHAP: Shapley additive explanation
SMOTE: synthetic minority over-sampling technique
VIF: variance inflation factor
XGBoost: extreme gradient boosting

Edited by T de Azevedo Cardoso; submitted 28.07.23; peer-reviewed by HR Chua, R He, Y Xin, B Liu, W Xiang; comments to author 20.02.24; revised version received 31.05.24; accepted 30.09.24; published 19.12.24.

Please cite as:
Ma M, Chen C, Chen D, Zhang H, Du X, Sun Q, Fan L, Kong H, Chen X, Cao C, Wan X
A Machine Learning–Based Prediction Model for Acute Kidney Injury in Patients With Community-Acquired Pneumonia: Multicenter Validation Study
J Med Internet Res 2024;26:e51255
URL: https://www.jmir.org/2024/1/e51255
doi:10.2196/51255
PMID:39699941

XSL•FO
RenderX

©Mengqing Ma, Caimei Chen, Dawei Chen, Hao Zhang, Xia Du, Qing Sun, Li Fan, Huiping Kong, Xueting Chen, Changchun Cao, Xin Wan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficient Screening in Obstructive Sleep Apnea Using Sequential Machine Learning Models, Questionnaires, and Pulse Oximetry Signals: Mixed Methods Study

Nai-Yu Kuo[1,2,3], MS; Hsin-Jung Tsai[2], PhD; Shih-Jen Tsai[2,3], MD; Albert C Yang[1,2,4,5], MD, PhD

[1]Sleep Medicine Center, Taipei Veterans General Hospital, Taipei, Taiwan
[2]Institute of Brain Science, National YangMing Chiao Tung University, Taipei, Taiwan
[3]Department of Psychiatry, Taipei Veterans General Hospital, Taipei, Taiwan
[4]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan
[5]Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

Corresponding Author:
Albert C Yang, MD, PhD
Digital Medicine and Smart Healthcare Research Center
National Yang Ming Chiao Tung University
No. 155 Sec. 2 Linong Street
Beitou District
Taipei, 11221
Taiwan
Phone: 886 02 28267000 ext 66555
Email: accyang@gmail.com

Abstract

Background: Obstructive sleep apnea (OSA) is a prevalent sleep disorder characterized by frequent pauses or shallow breathing during sleep. Polysomnography, the gold standard for OSA assessment, is time consuming and labor intensive, thus limiting diagnostic efficiency.

Objective: This study aims to develop 2 sequential machine learning models to efficiently screen and differentiate OSA.

Methods: We used 2 datasets comprising 8444 cases from the Sleep Heart Health Study (SHHS) and 1229 cases from Taipei Veterans General Hospital (TVGH). The Questionnaire Model (Model-Questionnaire) was designed to distinguish OSA from primary insomnia using demographic information and Pittsburgh Sleep Quality Index questionnaires, while the Saturation Model (Model-Saturation) categorized OSA severity based on multiple blood oxygen saturation parameters. The performance of the sequential machine learning models in screening and assessing the severity of OSA was evaluated using an independent test set derived from TVGH.

Results: The Model-Questionnaire achieved an F_1-score of 0.86, incorporating demographic data and the Pittsburgh Sleep Quality Index. Model-Saturation training by the SHHS dataset displayed an F_1-score of 0.82 when using the power spectrum of blood oxygen saturation signals and reached the highest F_1-score of 0.85 when considering all saturation-related parameters. Model-saturation training by the TVGH dataset displayed an F_1-score of 0.82. The independent test set showed stable results for Model-Questionnaire and Model-Saturation training by the TVGH dataset, but with a slightly decreased F_1-score (0.78) in Model-Saturation training by the SHHS dataset. Despite reduced model accuracy across different datasets, precision remained at 0.89 for screening moderate to severe OSA.

Conclusions: Although a composite model using multiple saturation parameters exhibits higher accuracy, optimizing this model by identifying key factors is essential. Both models demonstrated adequate at-home screening capabilities for sleep disorders, particularly for patients unsuitable for in-laboratory sleep studies.

(J Med Internet Res 2024;26:e51615) doi:10.2196/51615

KEYWORDS

sleep apnea; machine learning; questionnaire; oxygen saturation; polysomnography; screening; sleep disorder; insomnia; utilization; dataset; training; diagnostic

XSL•FO
RenderX

Introduction

Obstructive sleep apnea (OSA) is a prevalent sleep-related breathing disorder resulting from reduced muscle tension and collapsed soft tissue, causing airway blockage or limited airflow. This leads to fragmented sleep [1], nocturnal hypoxia [2], intrathoracic pressure fluctuations [3], daytime sleepiness [4], cognitive decline [5,6], and increased cardiovascular and metabolic disease risk [7-9]. OSA negatively impacts patients' long-term health, and quality of life, and increases medical and societal burdens [10]. An 11-year analysis by the US Veterans Health Administration identified sleep apnea and insomnia as the most common sleep disorders, at 47% and 26%, respectively [11]. Nearly 1 billion people worldwide experience from sleep apnea, with 430 million experiencing moderate to severe cases [12]. Prevalence increases with age, male gender, and higher BMI [13].

Polysomnography is the gold standard for diagnosing sleep apnea, requiring overnight monitoring in a sleep center with sensors detecting physiological changes. Respiratory signals, such as airflow, respiratory effort, snoring, and blood oxygen saturation (SpO_2), are analyzed by sleep technicians to calculate the Apnea-Hypopnea Index (AHI), which classifies apnea severity. Moderate to severe cases (ie, AHI values ≥ 15) require further treatment and management.

The prevalence of sleep apnea is quite high, yet low diagnosis rates persist. In a telephone survey of 4011 Taiwanese people, 51.9% reported snoring symptoms, and 2.6% were observed to have sleep apnea [14], indicating many potential patients may still be undiagnosed and untreated compared with other Asian regions [15]. Low diagnosis rates may be due to polysomnography limitations. Unfamiliar environments and equipment can cause discomfort and affect sleep patterns, potentially leading to misdiagnosis. The labor-intensive process requires trained technicians to analyze only three patients per night, causing longer waiting times. Manual analysis introduces variability in interpretation, with past research showing significant discrepancies in AHI values [16]. COVID-19 has further complicated hospital visits for testing, reducing confidence in diagnosis and increasing health care system burdens.

Recent studies aim to provide comfortable and convenient sleep apnea screening options, using subjective methods like questionnaires (eg, STOP-Bang, an acronym for Snoring, Tiredness, Observed apnea, high Blood Pressure, BMI, age, neck circumference, and gender) or objective methods involving wearable devices. The STOP-Bang questionnaire has high sensitivity but low specificity [17], potentially misclassifying patients. Objective methods analyze signals from wearables, such as electrocardiogram signals combined with cardiopulmonary coupling and cyclic variation of heart rate, achieving 89% sensitivity and 79% specificity [18]. Pulse oximeters are popular in research due to their simplicity and affordability [19].

The primary objective of this study is therefore to create a patient-friendly, clinically applicable tool for detecting OSA using machine learning techniques. This tool will combine subjective clinical questionnaires with objective blood oxygen concentration signals. In total, 2 distinct models have been developed: the Questionnaire Model (Model-Questionnaire) and the Saturation Model (Model-Saturation).

The Model-Questionnaire collects relevant questionnaires from a database, which includes demographic information and the Pittsburgh Sleep Quality Index (PSQI). Machine learning is then performed on these selected items. The Model-Saturation, on the other hand, uses SpO_2 signals. This model involves preprocessing to eliminate noise interference and uses various methods for feature extraction, such as time-domain analysis, frequency-domain analysis, nonlinear analysis, and estimation of oxygen desaturation index.

In this study, we used 2 extensive polysomnography datasets, the Sleep Heart Health Study (SHHS) [20,21] and polysomnography records from the Sleep Medicine Center of Taipei Veterans General Hospital (TVGH), for the training, testing, and validation of both models. We justified using datasets from different geographic areas to enhance the robustness and generalizability of their machine learning models across diverse populations, addressing potential biases and improving the study's external validity. Furthermore, we will assess these models using independent test data derived from TVGH to confirm the combined process's effectiveness and applicability.

This approach presents several advantages. For example, self-administered questionnaires can be implemented across various medical settings, increasing awareness among high-risk patients who might be unaware of their sleep apnea. In addition, pulse oximeters offer a noninvasive, cost-effective, and convenient solution for long-term monitoring, serving as an alternative to standard polysomnography. The unique aspect of this study lies in the integration of these methods for preliminary screening and triaging patients with OSA, ultimately promoting more efficient and accessible diagnosis and treatment options.

Methods

Sleep Data

We used 2 primary databases for this purpose: the clinical sleep data from the Sleep Medicine Center at TVGH, and the SHHS from the National Sleep Research Resource in the United States.

The Sleep Medicine Center's clinical sleep database, operated by the Department of Psychiatry at TVGH, contains clinical questionnaires and full-night standard polysomnography data for 4132 patients collected between December 1, 2012, and December 31, 2019. These questionnaires included clinical information, the PSQI, the Insomnia Severity Index, the Epworth Sleepiness Scale, and the Beck Depression Inventory-II. Sleep studies were conducted at the Sleep Medicine Center of Taipei Veterans General Hospital using Embla N7000 equipment (Natus Medical Inc), and data analysis was performed using RemLogic Version 3.4.1 software (Natus Medical Inc).

The SHHS is a multicenter cohort study designed to test whether sleep-related breathing is associated with an increased risk of cardiovascular disease. The study was conducted using the

XSL•FO
RenderX

home-based type II polysomnography device between November 1, 1995, and January 31, 1998, and involved 6441 men and women aged 40 years and above for the initial evaluation and sleep study. Between January 2001 and June 2003, 3295 participants underwent a second polysomnography. The SHHS database provides 8444 records of home-based polysomnography data with rigorous scoring. For the purposes of this study, we used the provided 8444 records and extracted the SpO_2 signals for Model-Saturation training.

Data Preparation and Preprocessing

To develop 2 models for predicting the presence and severity of OSA, the data will be divided into three groups: (1) the TVGH training set and the SHHS dataset, (2) both used for constructing the models, and (3) a TVGH independent test set for evaluating the models' performance, as illustrated in Figure 1.

Figure 1. Structure of Polysomnographic Data from Sleep Heart Health Study and Taipei Veterans Hospital. ModelQues: Model-Questionnaire; ModelSat: Model-Saturation. SHHS: Sleep Heart Health Study; TVGH: Taipei Veterans General Hospital.

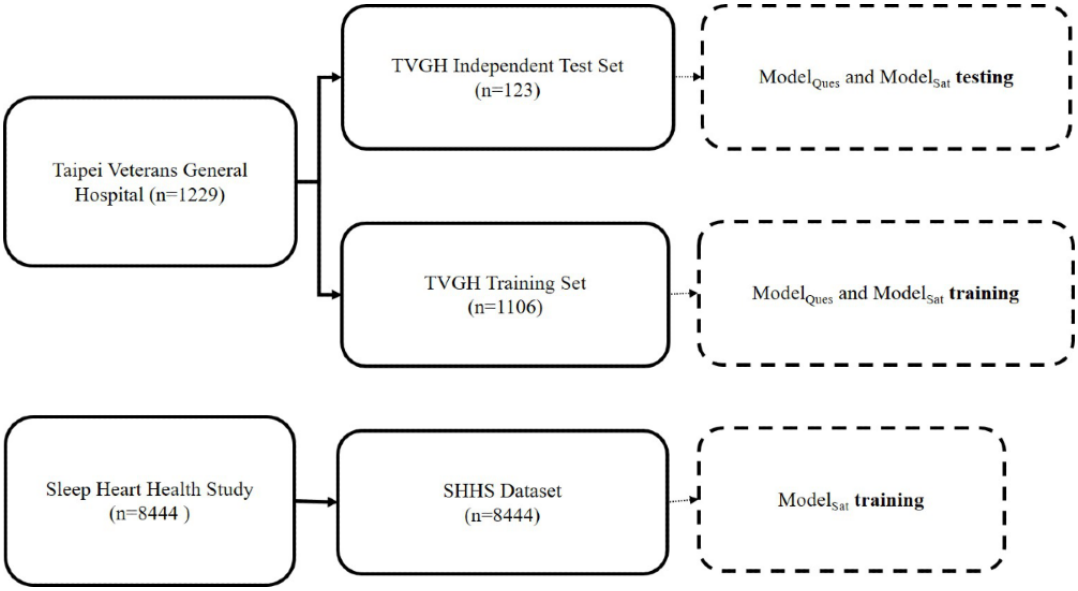

Initially, 1229 patients were selected from the TVGH database, comprising 822 patients diagnosed with OSA and 407 patients with primary insomnia through polysomnography. These patients were split into the TVGH Model-Questionnaire training set and the TVGH independent test set at a 9:1 ratio. The cases were categorized into 3 groups based on the severity of their condition: no sleep apnea, mild sleep apnea, and moderate-to-severe sleep apnea. To ensure a more uniform distribution of AHI across both datasets, participants in each group were sorted by age and AHI. Every 10th participant was chosen in sequence, resulting in 123 participants for the TVGH independent test set and 1106 participants for the TVGH Model-Questionnaire training set.

All 8444 sleep study records in the SHHS database served as the SHHS Model-Saturation training set. In addition, the polysomnography data obtained from these 1106 participants at TVGH also served as a training dataset for the Model-Saturation, which was trained separately from the SHHS database. SpO_2 signals were obtained from both the SHHS and TVGH databases. However, due to the different equipment used, the sampling frequency of SpO_2 signals varied between the 2 databases, 1 Hz in the SHHS database and 10 Hz in the TVGH database. To achieve signal consistency, the blood oxygen concentration signal in the TVGH database was down-sampled to 1 Hz. After standardizing the sampling frequency, SpO_2 signals underwent a three-step preprocessing procedure, which involved (1) excluding data points with blood oxygen concentrations below 50; (2) eliminating data exhibiting abrupt drops in blood oxygen concentration signals exceeding 4% within 1 second; and (3) excluding noncontinuous blood oxygen concentration signals for over 1 minute, to eliminate noise arising from instrument disconnection or participant movement during monitoring.

Feature Extraction

Model-Questionnaire

The Model-Questionnaire used questionnaires as the screening process for determining the presence of sleep apnea. Clinical data collected by the questionnaires and the PSQI served as feature values. Clinical data included gender, age, BMI, and the patient's history of hypertension, diabetes, and cardiovascular disease, resulting in a total of 6 items.

The PSQI is a retrospective questionnaire that prompts participants to recall their sleep patterns over the past month. It consists of 18 self-reported questions. Once completed, a specific calculation method is used to derive 7 subscores representing subjective sleep quality, sleep latency, sleep duration, sleep efficiency, sleep disturbance, medication use, and daytime dysfunction. The subscores are then summed to produce the final PSQI score. This study used 18 self-reported questions, 7 subscores, and the final PSQI score, a total of 26 items, as feature values.

Model-Saturation

The Model-Saturation used SpO_2 signals to classify patients' sleep apnea severity. After preprocessing the blood oxygen

XSL•FO
RenderX

concentration signal, four methods were used to extract signal features:

Saturation Distribution

Sleep-disordered breathing indirectly influences blood oxygen concentration changes. The average blood oxygen concentration distribution graph, plotted after grouping participants with varying apnea severities, indicates changes in blood oxygen concentration distribution based on severity. This study calculated the mean, median, maximum, minimum, kurtosis, variance, and skewness of SpO_2 signal from the collected overnight signals to describe the blood oxygen concentration distribution and use these 7 parameters as feature values.

Power Spectral Density

Sleep-disordered breathing induces fluctuations in blood oxygen concentration. Therefore, frequency analysis can be used to assess the oxygen desaturation index. By analyzing the power spectral density of the signal from the time domain to the frequency domain, the proportion of different frequencies can be obtained. This study will use 100 feature values acquired at intervals of 0.001 Hz from 0 to 0.1 Hz.

Multiscale Entropy

This study also attempted to use multiscale entropy (MSE) to detect changes in blood oxygen concentration. As a nonlinear analysis method, MSE observes the complexity of the signal at different time scales by using data points of varying lengths and averaging their values [22]. It then calculates the sample entropy [23] at different scales. This study used 20 scales and adopted these 20 items as feature values.

Approximation of Oxygen Desaturation Index

The oxygen desaturation index is defined as the average number of times that blood oxygen concentration drops by more than 4% per hour during an entire night's sleep. However, this study only used a single signal and cannot determine the patient's sleep and wake times. Therefore, the total number of times blood oxygen concentration decreases by 4% is divided by the total recording time to obtain an approximate value of the oxygen desaturation index.

In summary, the Model-Saturation adopts a total of 128 feature values, including 7 features from saturation distribution, 100 features from Power spectral density (PSD), 20 features from MSE, and 1 feature from the approximation of oxygen desaturation index (ODI).

Ground Truth Labels for Machine Learning Models

Model-Questionnaire

This study used the database from TVGH for training and testing the Model-Questionnaire. Patients with an AHI ≥ 5 after undergoing overnight polysomnography were labeled as having OSA. In contrast, those diagnosed with primary insomnia by a physician, without any other sleep disorders detected during polysomnography, were labeled as having primary insomnia. The Model-Questionnaire used 6 features from clinical data and 26 features from the PSQI, targeting OSA and primary insomnia as the ground truth values obtained from polysomnography results.

Model-Saturation

The Model-Saturation used preprocessed and feature-extracted SpO_2 signals to further categorize the severity of OSA in patients identified by the Model-Questionnaire. In this study, the AHI defined OSA severity. Patients with an AHI of <5 were labeled as without sleep apnea, those with an AHI of ≥5 but <15 were labeled as having mild sleep apnea, and those with an AHI of ≥15 were labeled as having moderate to severe sleep apnea. To ensure consistency among different databases, this study adopted the criteria published by the American Academy of Sleep Medicine (AASM) in 2012 to calculate the AHI. Apnea was defined as a reduction in airflow by ≥90% lasting for at least 10 seconds, while hypopnea was defined as a reduction in airflow by ≥30% lasting for at least 10 seconds accompanied by ≥4% oxygen desaturation. The total number of apnea and hypopnea events was divided by the total sleep time to obtain the AHI.

Machine Learning Models

In building the machine learning models, this study used a machine learning training platform (Invent AI by Chunghwa Telecom in Taiwan) to train and establish the Model-Questionnaire and Model-Saturation. TVGH independent test set was used to test the process after creating the 2 models. Three algorithms, including CatBoost, Random Forest, and Extra Trees were used in the AI platform for this study.

During the modeling process, the training sets for both the Model-Questionnaire and the Model-Saturation were divided into 3 sets by a fixed ratio. Specifically, 70% (Model-Questionnaire: 774/1106, Model-Saturation: 5910/8444) was allocated for training, 15% (Model-Questionnaire: 166/1106, Model-Saturation: 1267/8444) for validation, and the remaining 15% (Model-Questionnaire: 166/1106, Model-Saturation: 1267/8444) of the data was designated as the testing set for each model. Synthesized Minority Oversampling Technique method, which is an oversampling technique that artificially increases the data by interpolating among the minority samples, was used to deal with the issue of data imbalance. After the model building was completed, the reserved TVGH independent test set of 123 samples was used to evaluate whether the 2 models could classify OSA well in a sequential manner. First, the questionnaire data in the TVGH independent test set was inputted into the established Model-Questionnaire, resulting in the classification of primary insomnia and OSA. Then, the SpO_2 signal from cases predicted by Model-Questionnaire to have OSA was used in Model-Saturation, resulting in the final classification into three categories: (1) without sleep apnea, (2) mild sleep apnea, and (3) moderate to severe sleep apnea.

Statistical Analysis and Model Evaluation

As the data for this study were obtained from 2 different databases and was subsequently divided into 3 datasets for model training and testing, descriptive statistics were used to calculate the mean and SD of age, BMI, and AHI for each dataset. The number and proportion of cases by gender and severity of OSA were also calculated.

XSL•FO
RenderX

To evaluate the machine learning models, this study generated a confusion matrix by comparing the predicted values generated by the model with the actual classifications. F_1-score, precision, recall, accuracy, and area under the curve (AUC) were calculated to assess the predictive ability of the classifiers. The independent test dataset was used to validate the established model-questionnaire and model-saturation, and the results were used to generate a confusion matrix. F_1-score, precision, recall, and accuracy were calculated to evaluate the model's performance.

In this study, we also presented the ranking of feature importance based on the weights in the algorithms. The weights, which indicate the number of times a feature appears across all trees, were retrieved from the model to evaluate the feature importance, thereby yielding the average gain associated with each feature and highlighting how each contributes to the model's decision-making process.

Ethical Considerations

This study was approved by the institutional review board of Taipei Veterans General Hospital (approval number 2020-09-008BC) for the TVGH dataset, and approval for the use of data from the National Sleep Research Resource (NSRR R24 HL114473: NHLBI National Sleep Research Resource) was also obtained. Both databases do not contain any personally identifiable information or sensitive individual data, and the researchers were unable to determine the identities of the participants or contact them.

Results

Data Characteristics

This study used 2 large databases and divided them into 3 datasets for training and validation of the models. Table 1 displays the demographic data for the 3 datasets. The TVGH training set and independent testing datasets were both from the TVGH database. The SHHS dataset had slight differences in gender, age, and AHI distribution compared with the other 2 datasets. This was because the SHHS database included participants over the age of 40, while the TVGH database included participants over the age of 18 years who underwent sleep studies for sleep disorders. These differences in data collection purpose can affect the results but also provide insight into whether using a Model-Saturation trained from the SHHS dataset can predict well for Taiwanese data.

As shown in Table 1, both databases had data imbalance issues. The proportions of patients with primary insomnia and those with OSA in the TVGH training set were 33.1% (366/1106) and 66.9% (740/1106), respectively, while the SHHS dataset included 46.2% (3904/8444) of no sleep apnea, 31.2% (2630/8444) of mild sleep apnea, and 22.6% (1910/8444) of moderate-to-severe sleep apnea. Since data imbalance can affect the model's prediction, we employed the Synthesized Minority Oversampling Technique method to balance the data.

Table 1. Demographic data of 3 datasets.

	SHHS[a] dataset (n=8444)	TVGH[b] training set (n=1106)	TVGH independent test det (n=123)
Sex, n (%)			
Male	3986 (47.2)	651 (58.9)	69 (56.1)
Female	4458 (52.8)	455 (41.1)	54 (43.9)
Age (years), mean (SD)	64.55 (11.17)	52.35 (15.61)	51.80 (15.16)
BMI (kg/m^2), mean (SD)	28.21 (5.79)	25.04 (4.82)	25.20 (4.56)
AHI[c] (events/hour), mean (SD)	10.67 (13.78)	17.43 (21.08)	16.84 (19.88)
Severity of Sleep Apnea, n (%)			
Normal (AHI<5)	3904 (46.2)	366 (33.1)	41 (33.3)
Mild (5≦AHI<15)	2630 (31.2)	325 (29.4)	36 (29.3)
Moderate (15≦AHI<30)	1218 (14.4)	194 (17.5)	22 (17.9)
Severe (30≦AHI)	692 (8.2)	221 (20.0)	24 (19.5)

[a]SHHS: Sleep Heart Health Study.

[b]TVGH: Taipei Veterans General Hospital.

[c]AHI: Apnea-Hypopnea Index.

Model-Questionnaire

The training results of Model-Questionnaire for different algorithms and feature selection are presented in Table 2. Using all features for model training resulted in the same accuracy for all three algorithms. However, CatBoost had a better F_1-score of 0.85 compared with the other 2 algorithms. The feature importance ranking showed that BMI, age, self-reported snoring,

XSL•FO
RenderX

gender, and hypertension history were the most frequently occurring 5 features. Reducing the features to only these 5 and training the model again resulted in a slightly improved F_1-score and accuracy compared with using all features. The best algorithm for this reduced feature model was ExtraTrees, which may be related to the preselection of features.

Table 2. Results of Model-Questionnaire.

Model name	F_1-score	Precision	Recall	Accuracy (%)	AUC[a]	Feature importance				
						High				Low
All features										
CatBoost	0.85[b]	0.86	0.86	86.49%	0.84	BMI	Age	PSQI[c]	Gender	PSQI[d]
ExtraTrees	0.85	0.87	0.86	86.49%	0.85	Age	PSQI[a]	Gender	BMI	HBP[e]
Random Forest	0.84	0.88	0.86	86.49%	0.86	BMI	Age	PSQI[c]	Gender	HBP
5 features										
CatBoost	0.86	0.86	0.86	86.49%	0.84	Age	BMI	PSQI[c]	Gender	HBP
ExtraTrees	0.86[b]	0.87	0.87	87.39%	0.86	Age	PSQI[a]	BMI	Gender	HBP
Random Forest	0.86	0.87	0.87	87.39%	0.86	BMI	Age	PSQI[c]	Gender	HBP

[a]AUC: Area Under Curve.

[b]The machine learning model with the highest F_1-score.

[c]PSQI: PSQI question 5.5 (During the past month, how often have you had trouble sleeping because you cough or snore loudly?).

[d]PSQI: PSQI question 5.4 (During the past month, how often have you had trouble sleeping because you cannot breathe comfortably?).

[e]HBP: high blood pressure.

Model-Saturation

Model Training Using the Sleep Heart Health Study Dataset

The training results of Model-Saturation for different algorithms and feature selection are presented in Table 3. Using all features for model training, CatBoost had the best F_1-score of 0.85 and an accuracy of 84.5% among the 3 algorithms. The feature importance ranking showed that the estimated oxygen desaturation index, MSE, and spectral density analysis were all important features. However, after removing the SpO_2 distribution–related parameters, the model performance was not as good as using all features. Using single feature analysis for model training (Table 4), the model using spectral density analysis with Random Forest had the best prediction performance, with an F_1-score of 0.82 and an accuracy of 82%.

XSL•FO
RenderX

Table 3. Results of Model-Saturation by Sleep Heart Health Study dataset.

Model name	F_1-score	Precision	Recall	Accuracy (%)	AUC[a]	Feature importance				
						High				Low
All features										
CatBoost	0.85[b]	0.85	0.84	84.47%	0.95	ODI[c]	MSE1[d]	PSD[e] (0.013 Hz)	PSD (0.020 Hz)	PSD (0.017 Hz)
ExtraTrees	0.81	0.81	0.81	80.55%	0.94	ODI	MSE1	PSD (0.013 Hz)	PSD (0.016 Hz)	PSD (0.021 Hz)
Random Forest	0.83	0.83	0.83	83.19%	0.95	PSD (0.020 Hz)	PSD (0.019 Hz)	PSD (0.022 Hz)	PSD (0.021 Hz)	PSD (0.017 Hz)
ODI+MSE+PSD										
CatBoost	0.84[b]	0.84	0.84	83.87%	0.95	ODI4	MSE1	PSD (0.022 Hz)	PSD (0.017 Hz)	PSD (0.019 Hz)
ExtraTrees	0.82	0.82	0.82	81.57%	0.94	MSE1	ODI4	PSD (0.019 Hz)	MSE3	PSD (0.020 Hz)
Random Forest	0.79	0.80	0.79	78.58%	0.93	PSD (0.020 Hz)	PSD (0.019 Hz)	PSD (0.021 Hz)	PSD (0.024 Hz)	PSD (0.017 Hz)

[a]AUC: area under curve.

[b]The machine learning model with the highest F_1-score.

[c]ODI: oxygen desaturation index.

[d]MSE1: multiscale entropy.

[e]PSD: power spectral density.

Table 4. Model training results of SpO_2[a] signal by single feature analysis

Model name	F_1-score	Precision	Recall	Accuracy, %	Area under curve
Power spectrum density					
CatBoost	0.81	0.81	0.81	80.89	0.94
ExtraTrees	0.79	0.80	0.79	79.10	0.93
Random Forest	0.82[b]	0.82	0.82	82.00	0.94
Multiscale entropy					
CatBoost	0.70	0.70	0.70	69.54	0.86
ExtraTrees	0.69	0.69	0.69	68.77	0.86
Random Forest	0.71[b]	0.72	0.71	71.25	0.87
Saturation distribution					
CatBoost	0.67[b]	0.67	0.67	67.32	0.84
ExtraTrees	0.65	0.66	0.65	65.27	0.83
Random Forest	0.66	0.67	0.67	67.06	0.84
Approximation of oxygen desaturation index					
CatBoost	0.67[b]	0.67	0.66	66.30	0.83
ExtraTrees	0.66	0.67	0.66	66.30	0.82
Random Forest	0.66	0.67	0.66	66.21	0.83

[a]SpO_2: blood oxygen saturation.

[b]The machine learning model with the highest F_1-score.

XSL•FO
RenderX

Model Training Using the Taipei Veterans General Hospital Dataset

Table 5 presents the Model-Saturation built for 1106 cases in the TVGH training dataset. When using all features, the CatBoost algorithm achieves the best F_1-score of 0.83 and an accuracy of 0.82. The feature importance, in order, includes ODI, power spectral density, and variance as well as kurtosis of SpO_2 signal.

Table 5. Results of Model-Saturation using the Taipei Veterans General Hospital training set.

Model name	F_1-score	Precision	Recall	Accuracy (%)	AUC[a]	Feature importance				
						High				Low
CatBoost	0.82[b]	0.83	0.82	82.40%	0.94	ODI[c]	PSD[d] (0.022 Hz)	SpO_2 Variance	PSD (0.014 Hz)	SpO_2 Kurtosis
ExtraTrees	0.81	0.83	0.80	80.80%	0.92	ODI	PSD (0.022 Hz)	MSE9	PSD (0.039 Hz)	MSE2
Random Forest	0.82	0.83	0.81	81.60%	0.93	ODI	PSD (0.020 Hz)	SpO_2 Variance	PSD (0.014 Hz)	PSD (0.025 Hz)

[a]AUC: area under the curve.

[b]The machine learning model with the highest F_1-score.

[c]ODI: Oxygen desaturation index.

[d]PSD: Power spectral density.

Validation of Models Using TVGH Independent Test Data

The ultimate purpose of this study was to use 2 machine learning models to screen patients who may have OSA for further diagnosis and triage. Therefore, the best Model-Questionnaire and Model-Saturation were selected from the training models for use in the final validation. Based on the results in Table 2, the Extra Trees algorithm using the 5 features had good F_1-score, accuracy, and recall, and was chosen as the final Model-Questionnaire for the screening process. The Model-Saturation using all features with CatBoost was chosen for the second step of OSA triage, as it had the best performance in terms of recall, precision, accuracy, and F_1-score compared with other combinations.

The TVGH independent test dataset contains 123 cases, all of which included questionnaire data and SpO_2 signal. For the Model-Questionnaire, the results using the independent testing dataset are shown in Table 6. Although the accuracy and precision are lower than those in Table 2, the recall rate is 0.94, indicating that the model has a high sensitivity to the independent testing dataset. The F_1-score is also slightly higher. The confusion matrix in Figure 2a shows that although almost half of the 41 patients with primary insomnia were classified as having sleep apnea, 77 out of 82 patients with true sleep apnea were correctly classified, which suggests the model has high sensitivity despite the lower precision.

For the Model-Saturation, 97 cases classified as having sleep apnea by the Model-Questionnaire were further classified according to the severity of sleep apnea. Table 6 shows that although the accuracy of the Model-Saturation for moderate-to-severe sleep apnea is only 65.98% (64/97), which is significantly lower than the result obtained from the SHHS dataset (as shown in Table 3), this may be due to differences between the databases. However, the confusion matrix (Figure 2b) shows that, among those classified as having moderate-to-severe sleep apnea, all but four patients who actually had mild sleep apnea were correctly classified. The precision also slightly increased to 0.89, indicating that if a patient is classified as having moderate-to-severe sleep apnea by the Model-Saturation, there is a high possibility of having moderate-to-severe sleep apnea. Therefore, this Model-Saturation is considered effective in achieving the initial screening goal of moderate-to-severe sleep apnea even when using databases from different sources.

Finally, when tested on a TVGH-independent test set, the Model-Saturation trained using the TVGH training set showed an improved F_1-score from 0.775 to 0.828 (Table 6) compared with the Model-Saturation trained with the SHHS database. The results were closer to the accuracy obtained from the TVGH training set, and as shown in Figure 2c, the Model-Saturation trained with TVGH data demonstrates better performance in classifying patients with mild and moderate-severe sleep apnea compared with the Model-Saturation trained on the SHHS database.

XSL•FO
RenderX

Table 6. Testing results of two models by independent testing dataset.

Model	F_1-score	Precision	Recall	Accuracy, %
Model-Questionnaire				
ExtraTrees	0.8603	0.7938	0.939	79.67
Model-Saturation using the Sleep Heart Health Study dataset				
CatBoost	0.775	0.8857	0.6889	65.98
Model-Saturation using the Taipei Veterans General Hospital dataset				
CatBoost	0.8276	0.8571	0.8000	77.32

Figure 2. Confusion Matrix of Two Models: (a) Confusion matrix of Model-Questionnaire (b) Confusion matrix of Model-Saturation by the Sleep Heart Health Study data set (c) Confusion matrix of Model-Saturation by the Taipei Veterans General Hospital training set. PI: Primary insomnia; SA: Sleep apnea; 0: Without sleep apnea; 1: Mild sleep apnea; 2: Moderate and severe sleep apnea. SHHS: Sleep Heart Health Study; TVGH: Taipei Veterans General Hospital.

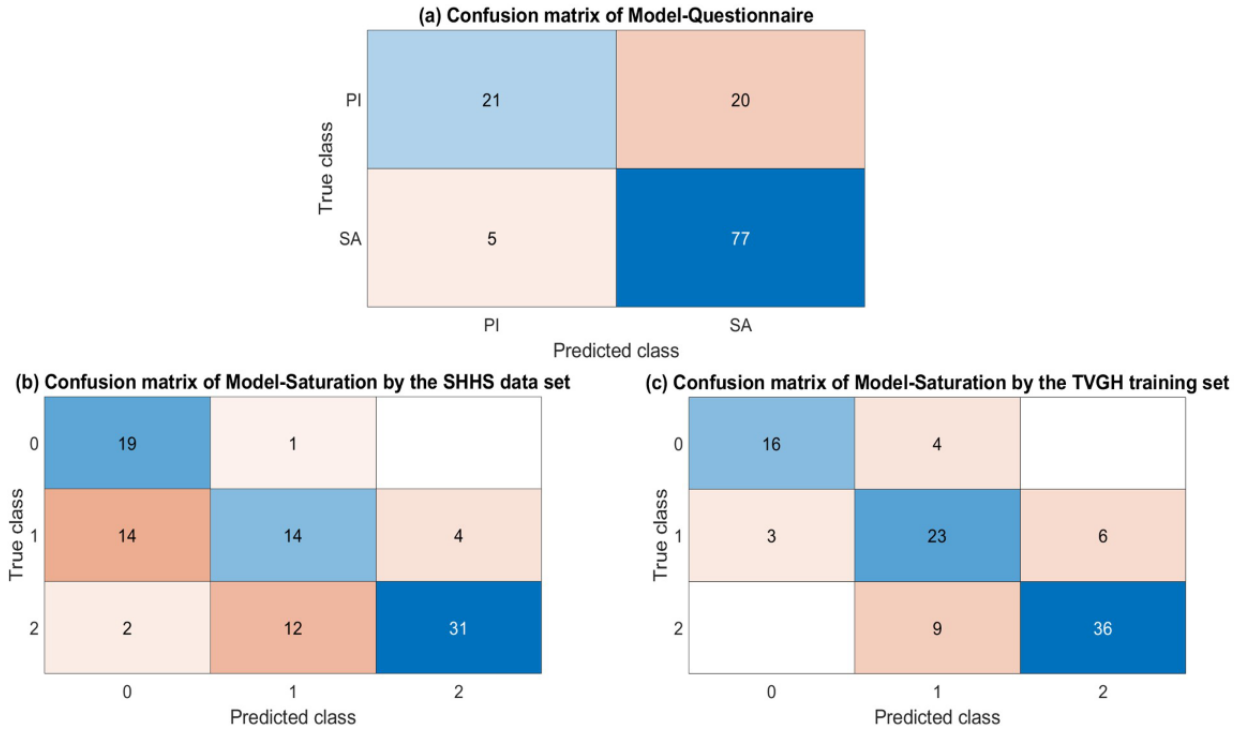

Discussion

Principal Findings

To the best of our knowledge, this study is the first to use a large amount of polysomnography data to establish sequential questionnaire and saturation models as a preliminary screening process for OSA. Previous studies mostly used a combination of questionnaires and blood oxygen signal features for prediction rather than building 2 models separately. For example, Mashaqi et al [24] used the STOP-Bang questionnaire combined with ODI to predict sleep apnea but achieved better classification only for severe cases.

However, if we can use a questionnaire to quickly and easily classify potential patients and then use SpO_2 signals to classify patients with moderate to severe sleep apnea who need to be diagnosed through polysomnography tests, it can improve the efficiency of sleep examinations and reduce unnecessary medical resource consumption. Therefore, compared with previous studies that used a model combining questionnaire and blood oxygen concentration for prediction, this study constructed a process consisting of 2 machine learning models that are more suitable for clinical applications.

Model-Questionnaire

The Model-Questionnaire was established using PSQI questionnaire data collected from the TVGH and was mainly used to distinguish between patients with sleep apnea and those with primary insomnia. As these are the 2 most common sleep disorders, the established model can cover a wide range of clinical applications. Based on the results in Table 2, using only the 5 features with the highest importance for the ExtraTrees algorithm, namely age, gender, BMI, snoring status, and history of hypertension, resulted in the highest accuracy for the questionnaire model while also improving model efficiency. These 5 features have been proven in previous studies to be risk factors for sleep apnea and are similar to STOP-Bang questionnaire items [17].

XSL•FO
RenderX

Although the STOP-Bang questionnaire has high sensitivity but low accuracy in detecting sleep apnea, with an AUC of only 0.56 in East Asian populations [25], our Model-Questionnaire achieved an AUC of 0.86, indicating better performance. Furthermore, since the database used for the questionnaire model only included data for sleep apnea and primary insomnia diagnoses, the trained model classified cases into these 2 categories. However, based on the important features for screening sleep apnea exhibited in the Model-Questionnaire, it may be possible to more broadly classify patients as having or not having sleep apnea.

Model-Saturation

Since the Model-Questionnaire has high sensitivity but low accuracy, it may result in misclassifying patients without sleep apnea. Therefore, it is important to further classify patients using physiological signals, which led to the development of the Model-Saturation in this study. Four different analyses were used to obtain features from the SpO_2 signal, and sleep apnea was classified into 3 different severity levels.

According to the feature importance in Table 3, the estimated ODI is the most important parameter in the model. Hang et al [26] used a support vector machine algorithm with ODI to predict moderate to severe sleep apnea and achieved an accuracy rate of 87.3%-87.8%. However, the accuracy rate of this study's model using only ODI as a feature was only 66.3%, which may be due to differences in signal quality and feature analysis methods. The data used in the study by Hang et al [26] were obtained from a single sleep center and SpO_2 signals were collected through traditional sleep tests. In contrast, this study used the SHHS database, which is a multicenter home sleep test that cannot control for artifacts and other interferences, making it more challenging to estimate ODI from blood oxygen signals. However, this approach is more representative of real-world clinical data collected from patients at home. In addition, the calculation of ODI in this study only considered oxygen desaturation of 4%, while Hang et al. included both 2% and 4% desaturation. This difference in methodology may be worth considering in future studies to improve the results.

In terms of feature importance, sample entropy (ie, entropy obtained in the original time series calculation under a time scale of 1) was found to be the second most important feature after ODI. Previous studies have also used MSE to study pediatric sleep apnea and calculated sample entropy over a range of time scales from 1 to 6. However, the final model only included sample entropy calculated under a time scale of 1 [27]. Based on the results of this study and previous literature, using sample entropy directly calculated from the original signal may be more efficient than using a time-consuming preprocessing method such as MSE.

Table 4 shows that the highest accuracy rate was achieved using spectral density analysis, with an accuracy rate of 82% and an AUC of 0.94. This feature also ranked high in the feature importance analysis in Table 3. Alvarez et al also found that spectral density analysis had better accuracy in classifying sleep apnea compared with other time-domain and nonlinear analysis methods. Among the features obtained through spectral density analysis, the peak amplitude between 0.014 Hz and 0.033 Hz had the highest accuracy rate of 83.1% [28]. In this study, using an interval of 0.001 Hz, the most important frequencies in the top 5 features were found to be between 0.013-0.014 Hz, indicating that the duration of blood oxygen desaturation of around 71.4-76.9 seconds is a discriminating factor for classifying the severity of sleep apnea.

In addition, from Table 5, which presents the Model-Saturation built for the 1106 cases at TVGH, it is evident that the importance of the variance of SpO_2 signal has significantly increased. In spectral analysis, the feature importance has slightly increased, and the frequency adopted is somewhat higher compared with that of the SHHS database. Comparing the validation results of the TVGH independent test set for Model-Saturation trained on 2 different databases in Table 6, it can be seen that using the TVGH database yields better predictions than the SHHS database. This suggests that different predictive models may be needed for different populations.

Finally, the change in feature importance reveals that variance has more significant importance in distinguishing sleep apnea in the TVGH database, implying that the blood oxygen fluctuations in patients with sleep apnea are more notable than those without sleep apnea. Furthermore, the slight increase in frequency from the spectral analysis indicates that the duration of blood oxygen fluctuations is shorter.

Limitations

There are several limitations that need to be addressed in this study. First, using SpO_2 signals as a physiological signal for OSA detection is intuitive and widely used because repetitive airway obstruction can cause fluctuations in blood oxygen levels. However, not all respiratory events result in changes in blood oxygen levels. According to the definition of apnea by the AASM, it is based on the degree of airflow reduction, not changes in blood oxygen levels. Therefore, using a single blood oxygen concentration signal may miss such events.

Second, this study used the acceptable definition published by the AASM in 2012 to estimate the AHI. Compared with the recommended definition, the main difference is in the evaluation of hypopnea. According to the acceptable definition, hypopnea is defined as a decrease in airflow of 30% or more for at least 10 seconds, accompanied by a decrease in blood oxygen levels of 4% or more. The recommended definition is defined as a decrease in airflow of 30% or more for at least 10 seconds, accompanied by a decrease in blood oxygen levels of 3% or more or awakening from sleep due to brain wave changes. Due to the differences in definitions, the Model-Saturation may provide relatively lenient results compared to the current standards, resulting in an underestimation of the severity of sleep apnea.

Third, the Model-Saturation was established using the SHHS database, which was collected in the United States and includes middle-aged and elderly participants older than 40 years. Although the advantage of this database is a large amount of home-based polysomnography data, there is still concern about whether the model established from this database is suitable for the Taiwanese population, in terms of social determinants of

XSL•FO
RenderX

health and access to health care resources, as well as a generalization to young and middle-aged adults. However, there is currently no similar large-scale home-based polysomnography database in Taiwan for training purposes. If there are future integration efforts for relevant sleep examination data, it may be possible to train models that are more tailored to the needs of Taiwanese patients and more accurate.

Fourth, the Model-Questionnaire trained using the TVGH dataset lacked data from a healthy control group. This dataset consisted exclusively of patients diagnosed with OSA and primary insomnia, which may affect the model's ability to accurately distinguish healthy participants. Despite this limitation, the PSQI questionnaire, incorporated within our model, served a critical role. In practical clinical settings, it enables the preliminary differentiation between patients and potentially healthy individuals through the identification of nonapnea cases, specifically by using a PSQI score below 5 as a screening criterion.

Conclusions

This study aimed to develop an efficient and cost-effective sleep apnea screening process by combining 2 machine learning models based on subjective questionnaires and objective SpO_2 signals. The Model-Questionnaire achieved an F_1-score of 0.86 using key features and the ExtraTrees algorithm, while the Model-Saturation scored 0.85 using CatBoost. Both models demonstrated good predictive ability.

The screening process of combined models started with a questionnaire to identify patients with potential sleep apnea, followed by a SpO_2 analysis for those classified with sleep apnea. The Model-Saturation showed high precision in classifying moderate to severe sleep apnea cases. This combined approach enhanced screening accuracy and reduced resource waste compared with using solely subjective or objective methods, potentially benefiting practical clinical applications and optimizing resources.

Acknowledgments

This study was supported by the National Science and Technology Council of Taiwan with grants to HJ Tsai (grant ID: NSTC-112-2314-B- A49-012-MY3) and AC Yang (grant ID:MOST-111-2622-8-075-002-IE; MOST-111-2321-B-A49-007; MOST-111-2628-B-A49-012; MOST-111- 2321-B-A49-011; and NSTC-111-2634-F-A49-014). AC Yang was also supported by the Mt. Jade Young Scholarship Award from the Ministry of Education, Taiwan, as well as the Brain Research Center, National Yang-Ming Chiao Tung University, and the Ministry of Education (Aim for the Top University Plan), Taipei, Taiwan. All the funding resources provided financial support only and had no other role in study design, data collection, analysis, or interpretation.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473 and 75N92019R002).

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to a confidentiality agreement required by the institutional review board but are available from the corresponding author on reasonable request.

Authors' Contributions

NYK analyzed the data, developed the machine learning models, performed the computational analysis, and wrote the manuscript. ACY designed the study and obtained the necessary funding resources. HJT and SJT gave critical comments and contributed to manuscript writing. All authors discussed the results and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Kimoff RJ. Sleep fragmentation in obstructive sleep apnea. Sleep 1996;19(9 Suppl):S61-S66. [doi: 10.1093/sleep/19.suppl_9.s61] [Medline: 9122574]
2. Bédard MA, Montplaisir J, Richer F, Malo J. Nocturnal hypoxemia as a determinant of vigilance impairment in sleep apnea syndrome. Chest 1991;100(2):367-370. [doi: 10.1378/chest.100.2.367] [Medline: 1864107]
3. Kohler M, Stradling JR. Mechanisms of vascular damage in obstructive sleep apnea. Nat Rev Cardiol 2010;7(12):677-685. [doi: 10.1038/nrcardio.2010.145] [Medline: 21079639]
4. Johns MW. Daytime sleepiness, snoring, and obstructive sleep apnea. The Epworth Sleepiness Scale. Chest 1993;103(1):30-36. [doi: 10.1378/chest.103.1.30] [Medline: 8417909]

XSL•FO
RenderX

5. Beebe DW, Groesz L, Wells C, Nichols A, McGee K. The neuropsychological effects of obstructive sleep apnea: a meta-analysis of norm-referenced and case-controlled data. Sleep 2003;26(3):298-307. [doi: 10.1093/sleep/26.3.298] [Medline: 12749549]
6. Engleman HM, Kingshott RN, Martin SE, Douglas NJ. Cognitive function in the sleep apnea/hypopnea syndrome (SAHS). Sleep 2000;23 Suppl 4:S102-S108. [Medline: 10893080]
7. Vgontzas AN, Papanicolaou DA, Bixler EO, Hopper K, Lotsikas A, Lin HM, et al. Sleep apnea and daytime sleepiness and fatigue: relation to visceral obesity, insulin resistance, and hypercytokinemia. J Clin Endocrinol Metab 2000;85(3):1151-1158. [doi: 10.1210/jcem.85.3.6484] [Medline: 10720054]
8. Tkacova R, McNicholas WT, Javorsky M, Fietze I, Sliwinski P, Parati G, European Sleep Apnoea Database study collaborators. Nocturnal intermittent hypoxia predicts prevalent hypertension in the European sleep apnoea database cohort study. Eur Respir J 2014;44(4):931-941 [FREE Full text] [doi: 10.1183/09031936.00225113] [Medline: 25102963]
9. McNicholas WT, Bonsigore MR, Management Committee of EU COST ACTION B26. Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities. Eur Respir J 2007;29(1):156-178 [FREE Full text] [doi: 10.1183/09031936.00027406] [Medline: 17197482]
10. Lyons MM, Bhatt NY, Pack AI, Magalang UJ. Global burden of sleep-disordered breathing and its implications. Respirology 2020;25(7):690-702 [FREE Full text] [doi: 10.1111/resp.13838] [Medline: 32436658]
11. Alexander M, Ray MA, Hébert JR, Youngstedt SD, Zhang H, Steck SE, et al. The national veteran sleep disorder study: descriptive epidemiology and secular trends, 2000-2010. Sleep 2016;39(7):1399-1410 [FREE Full text] [doi: 10.5665/sleep.5972] [Medline: 27091538]
12. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019;7(8):687-698 [FREE Full text] [doi: 10.1016/S2213-2600(19)30198-5] [Medline: 31300334]
13. Senaratna CV, Perret JL, Lodge CJ, Lowe AJ, Campbell BE, Matheson MC, et al. Prevalence of obstructive sleep apnea in the general population: a systematic review. Sleep Med Rev 2017;34:70-81. [doi: 10.1016/j.smrv.2016.07.002] [Medline: 27568340]
14. Chuang LP, Hsu SC, Lin SW, Ko WS, Chen NH, Tsai YH. Prevalence of snoring and witnessed apnea in Taiwanese adults. Chang Gung Med J 2008;31(2):175-181. [Medline: 18567418]
15. Mirrakhimov AE, Sooronbaev T, Mirrakhimov EM. Prevalence of obstructive sleep apnea in Asian adults: a systematic review of the literature. BMC Pulm Med 2013;13:10 [FREE Full text] [doi: 10.1186/1471-2466-13-10] [Medline: 23433391]
16. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med 2002;3(1):43-47. [doi: 10.1016/s1389-9457(01)00115-0] [Medline: 14592253]
17. Chung F, Abdullah HR, Liao P. STOP-bang questionnaire: a practical approach to screen for obstructive sleep apnea. Chest 2016;149(3):631-638 [FREE Full text] [doi: 10.1378/chest.15-0903] [Medline: 26378880]
18. Magnusdottir S, Hilmisson H. Ambulatory screening tool for sleep apnea: analyzing a single-lead electrocardiogram signal (ECG). Sleep Breath 2018;22(2):421-429. [doi: 10.1007/s11325-017-1566-6] [Medline: 28884264]
19. Netzer N, Eliasson AH, Netzer C, Kristo DA. Overnight pulse oximetry for sleep-disordered breathing in adults: a review. Chest 2001;120(2):625-633. [doi: 10.1378/chest.120.2.625] [Medline: 11502669]
20. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. The national sleep research resource: towards a sleep data commons. J Am Med Inform Assoc 2018;25(10):1351-1358 [FREE Full text] [doi: 10.1093/jamia/ocy064] [Medline: 29860441]
21. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, et al. The sleep heart health study: design, rationale, and methods. Sleep 1997;20(12):1077-1085. [doi: 10.1093/sleep/20.12.1077]
22. Costa M, Goldberger AL, Peng CK. Multiscale entropy analysis of complex physiologic time series. Phys Rev Lett 2002;89(6):068102. [doi: 10.1103/PhysRevLett.89.068102] [Medline: 12190613]
23. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278(6):H2039-H2049 [FREE Full text] [doi: 10.1152/ajpheart.2000.278.6.H2039] [Medline: 10843903]
24. Mashaqi S, Staebler D, Mehra R. Combined nocturnal pulse oximetry and questionnaire-based obstructive sleep apnea screening - A cohort study. Sleep Med 2020;72:157-163. [doi: 10.1016/j.sleep.2020.03.027] [Medline: 32653766]
25. Pivetta B, Chen L, Nagappa M, Saripella A, Waseem R, Englesakis M, et al. Use and performance of the STOP-bang questionnaire for obstructive sleep apnea screening across geographic regions: a systematic review and meta-analysis. JAMA Netw Open 2021;4(3):e211009 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1009] [Medline: 33683333]
26. Hang LW, Wang HL, Chen JH, Hsu JC, Lin HH, Chung WS, et al. Validation of overnight oximetry to diagnose patients with moderate to severe obstructive sleep apnea. BMC Pulm Med 2015;15:24 [FREE Full text] [doi: 10.1186/s12890-015-0017-z] [Medline: 25880649]
27. Crespo A, Álvarez D, Gutiérrez-Tobal GC, Vaquerizo-Villar F, Barroso-García V, Alonso-Álvarez ML, et al. Multiscale entropy analysis of unattended oximetric recordings to assist in the screening of paediatric sleep apnoea at home. Entropy 2017;19(6):284. [doi: 10.3390/e19060284]

XSL•FO
RenderX

28. Alvarez D, Hornero R, Marcos JV, del Campo F. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. IEEE Trans Biomed Eng 2010;57(12):2816-2824. [doi: 10.1109/TBME.2010.2056924] [Medline: 20624698]

Abbreviations

AASM: American Academy of Sleep Medicine
AHI: apnea-hypopnea index
AUC: area under curve
MSE: multiscale entropy
ODI: oxygen desaturation index
OSA: obstructive sleep apnea
PSD: power spectral density
PSQI: Pittsburgh Sleep Quality Index
SHHS: Sleep Heart Health Study
SMOTE: Synthesized Minority Oversampling Technique
SpO_2: blood oxygen saturation
STOP: Snoring, Tiredness, Observed apnea, high Blood Pressure
TVGH: Taipei Veterans General Hospital

Edited by A Schwartz; submitted 05.08.23; peer-reviewed by JM Aerts, L Guo; comments to author 02.02.24; revised version received 27.03.24; accepted 08.10.24; published 19.12.24.

Please cite as:
Kuo NY, Tsai HJ, Tsai SJ, Yang AC
Efficient Screening in Obstructive Sleep Apnea Using Sequential Machine Learning Models, Questionnaires, and Pulse Oximetry Signals: Mixed Methods Study
J Med Internet Res 2024;26:e51615
URL: https://www.jmir.org/2024/1/e51615
doi:10.2196/51615
PMID:39699950

©Nai-Yu Kuo, Hsin-Jung Tsai, Shih-Jen Tsai, Albert C Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey

Xiaoli Tang[1,2,3,4*], MSN; Xiaochen Yang[2,4*], BSN; Jiajun Yuan[1,3*], MPH; Jie Yang[5,6], MD; Qian Jin[2], BSN; Hanting Zhang[2], BSN; Liebin Zhao[1], DM, MM; Weiwei Guo[1,2,6,7], MM

[1]Shanghai Engineering Research Center of Intelligence Pediatrics, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[2]Neonatology Department, Shanghai Children's Medical Center, National Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[3]Child Health Advocacy Institute, China Hospital Development Institute, Shanghai, China

[4]Shanghai Jiao Tong University School of Nursing, Shanghai, China

[5]Neonatology Department, Nanfang Hospital, Southern Medical University, Guangzhou, China

[6]Neonatal Respiratory Group of Chinese Physicians Association, Guangzhou, China

[7]Administrative Office, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Weiwei Guo, MM
Shanghai Engineering Research Center of Intelligence Pediatrics
Shanghai Children's Medical Center, School of Medicine
Shanghai Jiao Tong University
1678 Dongfang Road, Pudong New Area
Shanghai, 200127
China
Phone: 86 18930830578
Email: guoweiwei@scmc.com.cn

Abstract

Background: Previous studies have shown that electrocardiographic (ECG) alarms have high sensitivity and low specificity, have underreported adverse events, and may cause neonatal intensive care unit (NICU) staff fatigue or alarm ignoring. Moreover, prolonged noise stimuli in hospitalized neonates can disrupt neonatal development.

Objective: The aim of the study is to conduct a nationwide, multicenter, large-sample cross-sectional survey to identify current practices and investigate the decision-making requirements of health care providers regarding ECG alarms.

Methods: We conducted a nationwide, cross-sectional survey of NICU staff working in grade III level A hospitals in 27 Chinese provinces to investigate current clinical practices, perceptions, decision-making processes, and decision-support requirements for clinical ECG alarms. A comparative analysis was conducted on the results using the chi-square, Kruskal-Wallis, or Mann-Whitney *U* tests.

Results: In total, 1019 respondents participated in this study. NICU staff reported experiencing a significant number of nuisance alarms and negative perceptions as well as practices regarding ECG alarms. Compared to nurses, physicians had more negative perceptions. Individuals with higher education levels and job titles had more negative perceptions of alarm systems than those with lower education levels and job titles. The mean difficulty score for decision-making about ECG alarms was 2.96 (SD 0.27) of 5. A total of 62.32% (n=635) respondents reported difficulty in resetting or modifying alarm parameters. Intelligent module–assisted decision support systems were perceived as the most popular form of decision support.

Conclusions: This study highlights the negative perceptions and strong decision-making requirements of NICU staff related to ECG alarm handling. Health care policy makers must draw attention to the decision-making requirements and provide adequate decision support in different forms.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e60944) doi:10.2196/60944

KEYWORDS

ECG alarm; electrocardiographic; perception; practice; decision-making; neonatal intensive care unit; health care providers; cross-sectional survey; nationwide

Introduction

Neonatal intensive care units (NICUs) are clinical settings equipped with over 40 types of medical devices with alarm functions [1,2]. Several studies have reported that alarm problems associated with electrocardiographic (ECG) monitors were the most prominent among all devices with high sensitivity and low specificity in clinical settings [3]. An observational study by Cho et al [4] revealed that ECG alarms accounted for 81.9% of the total 2184 alarms, far more than the number of alarms triggered by other devices. Li et al [5] found that each ECG monitors in the NICU generated over 177 alarms daily, of which over 95% were deemed false alarms. It is worth noting that several other studies have shown that existing clinical management strategies are insufficient in enhancing the precision of ECG alarms and providing clinical staff with effective decision support related to ECG alarms [6-9].

The prolonged exposure of health care providers to "high sensitivity and low specificity" alarms will disrupt their work and lead to emotional and work-related stress [3]. The alarms can also cause alarm fatigue and ignoring of signals in health care providers [10], which may increase health risks for patients or even death [11]. According to a national survey, up to 57.1% of adverse events were related to vital signs [12]. The Joint Commission published 98 cases of alarm-related accidents over a 3-year period, including 80 deaths and 13 cases of permanent loss of function. Meanwhile, it is worth noting that this result only represented less than 10% of the hospitals in the actual occurrence of alarm-related hazards [13].

In addition, sound exposure from frequent alarms can also have negative effects on neonatal outcomes, including growth and development. Due to the neonates' immature systems, prolonged noise stimuli can not only disrupt the development of the auditory cortex, leading to hearing loss [14], but can also imbalance the balance of excitation and inhibition in the cerebral cortex, causing neurological disorders and affecting brain development [15]. The American Academy of Pediatrics recommends that sound exposure in the NICU should not exceed 45 dB [16]. However, the noise levels in NICU wards can be up to 70 dB in general, with a maximum noise of up to 100 dB [17-19], which is significantly beyond the international threshold [20]. Zhang et al [20] further measured the sources of noise in the NICU and found that the ECG monitor was the instrument with the highest alarm volume (55.7-86.3 dB) in the NICU setting. Thus, the problem of ECG alarms in NICUs is closely related to the safety and health of newborns.

The Joint Commission's newly released Patient Safety Goals for 2022 included the safe use of device alarms as one of its goals [21]. The Chinese Hospital Association's Patient Safety Goals released in 2019 also specified the need to "enhance medical device safety and alarm management" [22]. The relevant system evaluation pointed out that false alarms of ECG monitors in NICU are mainly caused by irregular alarm management behaviors such as irrational setting of alarm values and parameters and answering or delayed answering, which depends on whether health care providers can make scientific decisions and self-management during the alarm setting and after they have occurred [9]. Decision-making behavior refers to the complex process of making a decision based on cognition, personal preferences, values, and other trade-offs when confronted with a problem; however, to the best of our knowledge, there has been no study investigating decision-making with regard to the ECG alarm. On the other hand, although a few studies have reported that alarm improvement management programs using alarms, algorithmic optimization software, and clinical decision support systems (CDSSs) [23] can promote decision-making about alarms, decrease false alarms [9,24-26], reduce alarm fatigue [27,28], and enhance alarm response speeds [29-31], these studies had limited sample numbers, and it was difficult to extend their results. Additionally, no studies have assessed the clinical implementation of ECG alarm practices [32].

Therefore, the objective of this study was to conduct a comprehensive nationwide, multicenter, large-sample cross-sectional survey to identify the current practices and investigate the decision-making requirements of health care providers regarding ECG alarms. The findings of this study serve as a solid foundation for enhancing decision support and developing strategies that address the challenges associated with ECG alarms.

Methods

Study Design and Participants

This was a strategic research and consulting project launched by the Respiratory Group of the Neonatologists' Branch of the Chinese Medical Doctors Association, China. The inclusion criteria were as follows: (1) the number of ECG monitors in the NICU was ≥50, (2) on-the-job registered physicians and nurses, and (3) years of work experience ≥3 years. The exclusion criterion was medical staff in training. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines, ensuring transparent and comprehensive reporting of the research findings.

Ethical Considerations

Prior to the commencement of the study, ethics approval was obtained from the ethics committee of Shanghai Children's Medical Center (approval SCMCIRB-K2024014-1). This study strictly adhered to ethical guidelines, with all participants signing informed consent forms to ensure their awareness of the research objectives and procedures. All data were anonymized to protect participant privacy.

XSL•FO
RenderX

Sampling Method

We first screened all hospitals within the Chinese Physicians Association, which encompasses the majority of hospitals with NICUs nationwide, to identify those that met the inclusion criteria. In each NICU, an equal number of health care providers were surveyed. An anonymous web-based questionnaire was administered to physicians and nurses in a 1:2 ratio, that is, the numbers of physicians and nurses selected from each NICU were 4 and 8, respectively, with a total of 12 respondents. Both physician and nurse subgroups were selected by convenience sampling according to a 1:1:2 ratio of professional titles (senior, intermediate, and junior titles) [33] (Figure 1).

Figure 1. Sampling method.

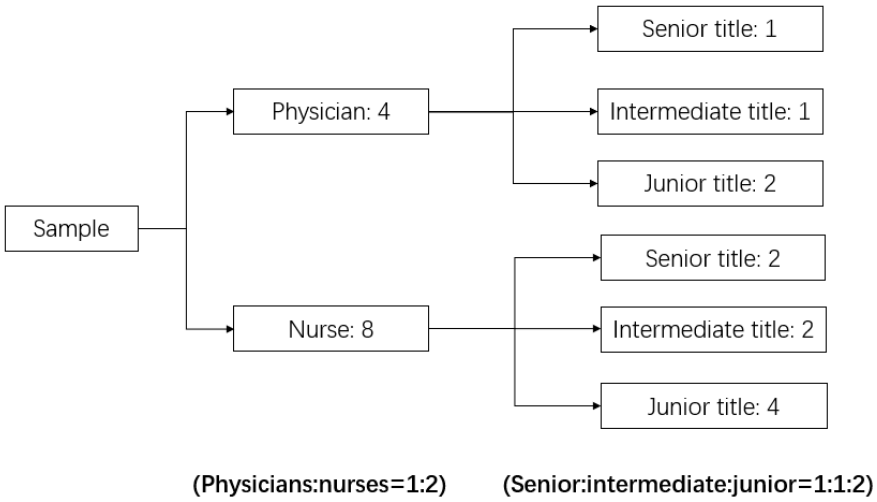

Data Collection

We used the Wenjuanxing website, a secure web-based platform designed to capture data from web-based investigations. This platform was used to create a survey project located on a static URL that could be remotely accessed via a smartphone browser regardless of the geographical position of the proxy. There was a web-based informed consent page for obtaining e-consent prior to the web-based study questionnaire page, and consent questions were added to the page in yes or no format. On the first page of the questionnaire, the aims, objectives, instructions, and questions were designed to be mandatory, so participants who did not complete all the questions received a warning before submission. Therefore, all the collected questionnaires were completed without missing data. The questionnaire response time was approximately 10 minutes. According to the sampling method mentioned earlier, the link to the e-questionnaire was distributed to the responsible personnel at each selected hospital via email. Those who expressed willingness to participate were provided with a web-based anonymous questionnaire by the designated person. The platform's settings ensured that multiple entries from the same individual were prevented by tracking unique IP addresses.

Measurements

Sociodemographic Information

Sociodemographic characteristics included sex, age, educational level, marital status, hospital, job title, profession, position, and years of work experience. The job-title classification system is widely acknowledged and implemented throughout the Chinese medicine and health care system, ranking medical professionals as junior, intermediate, and senior. Physicians and nurses can advance to their respective titles if they meet the necessary conditions, including educational requirements, clinical competency criteria, and passing the National Uniform Title Examination.

Survey Questionnaire Regarding Clinical Alarms on Medical Devices (Chinese Version)

The original questionnaire was initially administered by the Healthcare Technology Foundation in 2005 and was revised in 2011 [34] and 2016 [35]. In 2023, the 2016 version was translated from English to Chinese by Wang [36]. The questionnaire was designed to assess perceptions and practices of medical staff regarding clinical alarms across 7 dimensions (nuisance alarms, experience of alarm systems, alarm notifications, smart alarms, hospital policies and procedures, alarm management improvements, and alarm-related adverse events). It contained 31 items, of which 16 items consisted of agreement levels on alarms and alarm management. These were rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), with higher scores indicating stronger agreement with the item statements. Items 6, 7, and 9 in dimension II were reverse-scored. In addition, 8 items concerned practices related to clinical alarms (response options including "no," "yes," and "not sure"). Each item was analyzed based on the proportion of respondents selecting each option, with a higher proportion indicating that the practice was more commonly used. The remaining 7 items were open-ended questions for each dimension that were used to collect the respondent's subjective opinion (see Part II in Multimedia Appendix 1). The survey showed sound internal consistency with a Cronbach α of 0.766, an item content validity index ranging from 0.86 to 1.00, and a scale content validity index of 0.98 [36]. The questionnaire exhibited good construct validity.

XSL•FO
RenderX

Questionnaire on Decision-Making and Decision-Support Regarding ECG Alarms

A questionnaire was developed to assess the decision-making processes and decision-support requirements of medical staff regarding clinical alarms. First, we created an initial version of the questionnaire by developing items based on the self-evaluation section conducted by He et al [37]. Second, an expert panel comprising 2 medical scientists, 2 senior nurses, and a data scientist was consulted to refine these items. Third, 3 doctors and 7 nurses took a pretest of the questionnaire, and modifications were made according to their feedback to obtain the final version of the questionnaire.

The final questionnaire contained the following three sections:

1. Confidence and difficulty of decision-making: This section included 8 items, each rated on a 5-point Likert scale (1=low and 5=high). Higher scores indicated that the health care providers had greater confidence or more difficulty in their decision-making process.
2. Decision-making basis: This section included 3 items. The first item investigated the basis for decision-making and included 9 multiple-choice items. A higher proportion indicated that the basis for making decisions about ECG alarms was more commonly used. The second and third items were designed to evaluate the effectiveness and accessibility of the existing decision-making bases, respectively, which were rated on a 5-point Likert scale. Higher scores indicated that the basis was perceived as more effective or accessible.
3. Decision support needs: This section included 6 items, aimed at exploring the preferred forms and specific functions of decision support. Items 12 and 13 were multiple-choice questions regarding the preferred forms by health care providers. The remaining items were rated on a 5-point Likert scale, with higher scores indicating stronger agreement with the specific function requirements.

Statistical Analysis

According to our pilot study, the mean confidence and difficulty score for the decision-making about ECG alarms were 4.04 (SD 0.51) and 2.88 (SD 0.40) of 5, respectively. A sample size of 570 and 920 will produce a 2-sided 95% CI with a distance from the mean to the limits that is less than or equal to 0.04 if the population SD is estimated to be 0.4 and 0.51 by a previous sample of size 30, respectively. Considering around a 5% dropout rate, the expected sample size in this study was 969 participants. Descriptive statistics were used to describe the respondents' sociodemographic information as well as their responses to the questions. A comparison analysis was conducted to test whether the perception differed by educational level, job title, and profession using a chi-square test, Kruskal-Wallis test, or Mann-Whitney *U* test. All statistical analyses in this study were 2-tailed, a *P* value of less than .05 was considered statistically significant, and all analyses were carried out using SPSS statistical software (version 20.0; IBM Corp).

Results

General Information

In this cross-sectional study, data were collected through a web-based survey conducted between October 1, 2023, and December 30, 2023. The survey aimed to cover nearly all geographical regions of China, and the percentages of sample size distributed in each region are detailed in Table 1. The respondents were from 99 hospitals representing 27 province-level regions in China (Figure 2). Finally, 1019 respondents were included in the data analysis with a response rate of 85.77%. All respondents worked in grade III level A hospitals. The detailed demographic characteristics of the respondents are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of the respondents (N=1019).

Characteristics	Values
Sex, n (%)	
Male	105 (10.30)
Female	914 (89.70)
Education background, n (%)	
Associate degree or below	94 (9.23)
Bachelor degree	674 (66.14)
Master degree or above	251 (24.63)
Marital status, n (%)	
Single	238 (23.36)
Married	763 (74.87)
Other	18 (1.77)
Area of the hospital, n (%)	
East China	415 (40.73)
North China	70 (6.87)
Northeastern China	70 (6.87)
Southern China	108 (10.60)
Northwestern China	85 (8.34)
Southwest China	153 (15.01)
Central China	108 (10.60)
Other	10 (0.98)
Type of hospital, n (%)	
General hospital	586 (57.51)
Specialized hospital	423 (41.51)
Other	10 (0.98)
Age (years), mean (SD)	34.65 (6.93)
Children, mean (SD)	0.94 (0.76)
Length of work experience (years), mean (SD)	11.42 (7.44)
Length of work experience in intensive care unit (years), mean (SD)	8.53 (6.20)
Daily working hours, mean (SD)	8.96 (2.22)
Profession, n (%)	
Nurse	666 (65.36)
Physician	287 (28.16)
Other	66 (6.48)
Position, n (%)	
Archiater	96 (9.42)
Physician-in-charge	141 (13.84)
In-training resident	50 (4.91)
Head nurse	70 (6.87)
Specialist nurse	138 (13.54)
Bedside nurse	458 (44.94)
Other	66 (6.48)
Job title, n (%)	

XSL•FO
RenderX

Characteristics	Values
Senior	186 (18.25)
Intermediate	415 (40.73)
Junior or below	418 (41.02)

Figure 2. Geographical region distribution of the hospitals.

Perceptions and Practices Regarding Clinical ECG Alarms

Dimensions I and II indicate health care providers' clinical negative perceptions related to ECG alarms. Additionally, dimensions III to VII indicate health care providers' perceptions and practices of ECG alarm management initiatives. The detailed scores and the choice proportion from dimensions I to VII are found in Tables 2 and 3.

XSL•FO
RenderX

Table 2. Clinical perceptions related to electrocardiographic alarms (N=1019).

Dimension and item	Scores, mean (SD)	"Agree" or "strongly agree," n (%)
Dimension I: nuisance alarms		
2.1 Nuisance alarms occur frequently	3.00 (1.12)	391 (38.37)
2.2 Nuisance alarms disrupt patient care	3.11 (1.15)	513 (50.34)
2.3 Nuisance alarms reduce trust in alarms and cause caregivers to inappropriately turn alarms off at times other than setup or procedural events	2.32 (1.19)	230 (22.57)
Mean score for each item in dimension I	2.81 (0.90)	—[a]
Dimension II: experience of alarm systems		
2.5 Properly setting alarm parameters and alerts is overly complex in existing devices	2.53 (1.03)	221 (21.69)
2.6 Newer monitoring systems have solved most of the previous problems we experienced with clinical alarms	2.49 (0.93)	570 (55.94)
2.7 The alarms used on my floor/area of the hospital are adequate to alert staff of potential or actual changes in a patient's condition	2.05 (0.80)	827 (81.16)
2.8 There have been frequent instances where alarms could not be heard and were missed	2.21 (0.97)	134 (13.15)
2.9 Clinical staff is sensitive to alarms and responds quickly	1.80 (0.66)	933 (91.56)
2.10 When a number of devices are used with a patient, it can be confusing to determine which device is in an alarm condition	2.53 (1.15)	274 (26.89)
2.11 Environmental background noise has interfered with alarm recognition	2.85 (1.15)	400 (39.25)
Mean score for each item in dimension II	2.35 (0.58)	—
Dimension III: alarm notification		
2.14 Alarm integration and communication systems are useful for improving alarm management and response	3.89 (0.77)	758 (74.39)
2.16 Central alarm management staff responsible for receiving alarm messages and alerting appropriate staff is helpful	3.85 (0.83)	730 (71.64)
Mean score for each item in dimension III	3.87 (0.69)	—
Dimension IV: smart alarms		
2.19 Smart alarms would be effective to use for reducing false alarms	3.75 (0.79)	648 (63.59)
2.20 Smart alarms would be effective to use for improving clinical response to important patient alarms	3.83 (0.76)	711 (69.77)
Mean score for each item in dimension IV	3.89 (0.73)	—
Dimension V: hospital policies and procedures		
2.24 Clinical policies and procedures regarding alarm management are effectively used in my facility	3.79 (0.81)	699 (68.60)
Dimension VII: alarm-related adverse events		
2.30 Since the implementation of the 2019 Patient Safety Goals issued by the Chinese Hospital Association regarding the alarm management program, your hospital has experienced a decrease in the occurrence of adverse events related to alarms	3.77 (0.75)	640 (62.81)

[a]Not applicable.

XSL•FO
RenderX

Table 3. Clinical practices of electrocardiographic alarm management initiatives (N=1019).

Dimensions and items	Yes, n (%)	No, n (%)	Not sure, n (%)
Dimension III: alarm notification			
2.13 Does your institution use alarm integration and communication systems (eg, pagers, cell phones, and other wireless devices) to notify alarms?	458 (44.95)	415 (40.73)	146 (14.33)
2.15 Does your institution use "monitor watchers" in a central viewing area to observe and communicate alarm conditions to caregivers?	336 (32.97)	499 (48.97)	184 (18.06)
Dimension IV: smart alarms			
2.18 Does your institution use Smart Alarm Systems?	368 (36.11)	407 (39.94)	244 (23.95)
Dimension V: hospital policies and procedures			
2.22 Have you received education on the setup and operation of device alarms?	687 (67.42)	243 (23.85)	89 (8.73)
2.23 Does your institution have a requirement to document that the alarms are set and are appropriate for each patient?	425 (41.71)	412 (40.43)	182 (17.86)
Dimension VI: alarm management improvements			
2.26 Has your institution developed clinical alarm improvement initiatives over the past 2 years?	554 (54.37)	184 (18.06)	281 (27.58)
2.27 Does your institution have a program in place to improve the security of alarm management?	647 (63.49)	142 (13.94)	230 (22.57)
Dimension VII: alarm-related adverse events			
2.29 Has your institution experienced adverse patient events in the last 2 years related to clinical alarm problems?	171 (16.78)	576 (56.53)	272 (26.69)

Different Personal Characteristics of Health Care Providers Toward Clinical Perceptions of ECG Alarms

Sex

Comparing female health care providers toward clinical perceptions of ECG alarms, male health care providers had more negative perceptions of experiences with alarm systems (P=.008). Female health care providers had more positive attitudes toward improving the current status of ECG alarms through adhering to hospital regulations (P<.001; Tables 4 and 5).

XSL•FO
RenderX

Table 4. Descriptive analysis of different personal characteristics of health care providers toward clinical perceptions of electrocardiographic alarms (N=1019).

Variables	Values, n (%)	Dimension I: nuisance alarms			Dimension II: experience of alarm systems		
		Mean (SD)	*t* test (*df*)	*P* value	Mean (SD)	*t* test (*df*)	*P* value
Sex			0.16 (1017)	.87		2.67 (1017)	.008[a]
Male	105 (10.30)	2.82 (0.91)			2.49 (0.56)		
Female	914 (89.70)	2.81 (0.90)			2.34 (0.57)		
Education background			4.50 (2, 1016)[a]	.01[b]		11.84 (2, 1016)[a]	<.001
Master degree or above	251 (24.63)	2.90 (0.87)			2.49 (0.54)		
Bachelor degree	674 (66.14)	2.81 (0.91)			2.32 (0.57)		
Associate degree or below	94 (9.22)	2.58 (0.93)			2.20 (0.62)		
Job title			4.61 (2, 1016)[a]	.01[b]		8.11 (2, 1016)[a]	<.001
Senior	186 (18.25)	2.95 (0.87)			2.48 (0.53)		
Intermediate	415 (40.73)	2.84 (0.90)			2.37 (0.58)		
Junior or below	418 (41.02)	2.72 (0.92)			2.28 (0.59)		
Profession			–1.99 (951)	.047[b]		–5.59 (951)	<.001
Nurse	666 (65.35)	2.80 (0.93)			2.29 (0.58)		
Physician	287 (28.16)	2.92 (0.84)			2.52 (0.53)		

[a]*F* test.

[b]$P<.05$.

XSL•FO
RenderX

Table 5. Descriptive analysis of the electrocardiographic alarm management initiatives among health care providers with different personal characteristics (N=1019).

Variables	Values, n (%)	Dimension III: alarm notification			Dimension IV: smart alarms			Dimension V: hospital policies and procedures		
		Mean (SD)	*t* test (*df*)	*P* value	Mean (SD)	*t* test (*df*)	*P* value	Mean (SD)	*t* test (*df*)	*P* value
Sex			–0.81 (1017)	.42		–0.29 (1017)	.78		–3.74 (1017)	<.001
Male	105 (10.30)	3.81 (0.63)			3.77 (0.70)			3.51 (0.82)		
Female	914 (89.70)	3.88 (0.70)			3.79 (0.74)			3.82 (0.80)		
Education background			6.48 (2, 1016)[a]	.002[b]		1.91 (2, 1016)[a]	.15		12.81 (2, 1016)[a]	<.001
Master degree or above	251 (24.63)	3.80 (0.68)			3.75 (0.70)			3.60 (0.77)		
Bachelor degree	674 (66.14)	3.86 (0.69			3.78 (0.74			3.83 (0.81		
Associate degree or below	94 (9.22)	4.10 (0.71)			3.92 (0.74)			4.04 (0.80)		
Job title			1.61 (2, 1016)[a]	.20		4.44 (2, 1016)[a]	.012[b]		7.43 (2, 1016)[a]	.001[b]
Senior	186 (18.25)	3.84 (0.68)			3.67 (0.74)			3.62 (0.82)		
Intermediate	415 (40.73)	3.84 (0.71)			3.76 (0.74)			3.77 (0.82)		
Junior	418 (41.02)	3.92 (0.67)			3.86 (0.72)			3.89 (0.77)		
Profession			1.14 (951)	.26		1.46 (951)	.14		6.04 (567.42)	<.001
Nurse	666 (65.36)	3.87 (0.71)			3.80 (0.76)			3.89 (0.81)		
Physician	287 (28.16)	3.82 (0.65)			3.72 (0.69)			3.56 (0.77)		

[a]*F* test.

[b]*P*<.05.

Job Title and Education Level

Individuals with higher education levels (*P*=.01; *P*<.001) and job titles (*P*=.01; *P*<.001) had more negative perceptions of nuisance alarms and experiences with alarm systems than those with lower education levels and job titles (Table 4). In addition, individuals with lower job titles had more positive perceptions toward improving the current status of ECG alarms by adhering to hospital regulations and the application of smart alarms (*P*=.001; *P*=.012; Table 5).

Profession

Compared to the nurses, physicians had more negative perceptions with ECG monitors with regard to nuisance alarms and alarm systems (*P*=.047; *P*<.001). Additionally, nurses were more likely to agree that alarm management could be improved through policy implementation than physicians (*P*<.001).

Difficulty and Confidence in Decision-Making About ECG Alarms

In total, the mean confidence and difficulty scores for decision-making about ECG alarms were 3.99 (SD 0.63) and 2.96 (SD 0.27) of 5, respectively. Furthermore, 70.07% (n=714) of the respondents reported difficulty or great difficulty in the response and disposition alarms after they occurred, while 62.32% (n=635) reported difficulty in the timing of resetting or modifying alarm parameters (Figure 3). The rankings of confidence scores are in contrast to the difficulty scores, except for “response and disposition after an alarm occurs” (Table 6).

XSL•FO
RenderX

Figure 3. Difficulty of decision-making regarding electrocardiographic alarms.

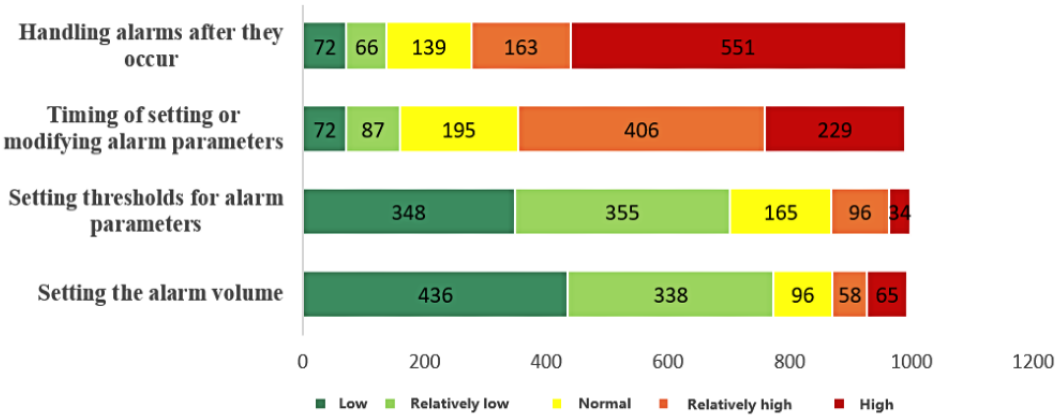

Table 6. Difficulty and confidence scores of decision-making about electrocardiographic alarms.

Items	Difficulty score, mean (SD)	"High" or "relatively high," n (%)	Confidence score, mean (SD)	"High" or "relatively high," n (%)
Response and disposition after an alarm occurs	4.06 (1.27)	714 (70.07)	4.06 (0.69)	870 (85.38)
Timing of resetting or modifying alarm parameters	3.64 (1.14)	635 (62.32)	3.88 (0.76)	761 (74.68)
Setting thresholds for alarm parameters	2.11 (1.09)	130 (12.76)	4.00 (0.71)	850 (83.41)
Setting the alarm volume	1.97 (1.17)	123 (12.07)	4.03 (0.74)	852 (83.61)
Mean score	2.96 (0.27)	__[a]	3.99 (0.63)	—

[a]Not applicable.

Basis of Decision-Making Regarding ECG Alarms

Colleagues and leaders (n=891, 87.44%) were considered the most accessible decision-making bases, and clinical guidelines were considered the most effective (n=954, 93.62%). Table 7 shows the details of the selection of each decision-making basis.

Table 7. Accessibility and effectiveness scores of decision-making basis.

	Effectiveness score, mean (SD)	"Effective" or "very effective," n (%)	Accessibility score, mean (SD)	"Accessible" or "very accessible," n (%)
Primary study	3.99 (0.84)	776 (76.15)	3.73 (0.96)	641 (62.90)
Academic conference	4.06 (0.82)	817 (80.18)	3.75 (0.96)	660 (64.77)
Thematic training	4.08 (0.79)	831 (81.55)	3.76 (0.95)	663 (65.06)
Textbook	4.15 (0.81)	870 (85.38)	4.00 (0.92)	810 (79.49)
Colleagues and leaders	4.14 (0.76)	873 (85.67)	4.28 (0.75)	891 (87.44)
Systematic review	4.22 (0.74)	893 (87.63)	3.97 (0.87)	761 (74.68)
Standard operating procedure	4.30 (0.74)	924 (90.68)	4.15 (0.79)	858 (84.20)
Instruction book	4.37 (0.73)	925 (90.78)	4.26 (0.79)	885 (86.85)
Clinical guideline	4.42 (0.69)	954 (93.62)	4.21 (0.75)	884 (86.75)

Needs for Decision Support in ECG Alarm Management

The preferred forms of decision support by health care providers are presented in Table 8. "Intelligent module–assisted decision support systems" was the most popular form of decision support. Further, Table 9 revealed specific function development regarding intelligent module–assisted decision support systems for ECG monitors.

XSL•FO
RenderX

Table 8. Forms of decision support for electrocardiographic monitors.

Items	Decision support forms on alarm settings, n (%)	Decision support forms on responding to and handling alarms after they occur, n (%)
Intelligent module–assisted decision support systems	722 (70.85)	710 (69.68)
Consultation with professionals	684 (67.12)	689 (67.62)
Knowledge repository	595 (58.39)	578 (56.72)
eBook	534 (52.40)	532 (52.21)
Paper-based book	462 (45.34)	442 (43.38)

Table 9. Specific function development for electrocardiographic (ECG) monitors.

Item	"Important" or "very important," n (%)	Mean (SD)
Develop a standardized device interface to integrate ECG monitors with other devices	833 (81.75)	4.17 (0.78)
Intelligent decision support for alarm threshold setting for newborns of different birth gestational ages module	820 (80.47)	4.10 (0.79)
Optimizing algorithm models to achieve accurate alarms	814 (79.88)	4.09 (0.75)
Functional modules for grading alarms based on the severity of the patient's condition	806 (79.10)	4.04 (0.76)

Discussion

Strengths

This was the first national-scale survey that aimed to examine clinical response and management regarding ECG alarms at both the practical and perceptual levels. The study covered more than 1019 medical staff members from 27 provinces in China, which is the largest sample size to date. The overall sample size was representative and large, which can help extend our results to reflect clinical practice as well as decision-making processes and needs related to ECG alarms. Overall, the study found that most departments have implemented initiatives for ECG alarm management. However, health care providers reported experiencing a significant number of nuisance alarms and held negative perceptions regarding alarms in clinical settings. Moreover, health care providers continue to face challenges in decision-making and accessing decision support for alarm settings and handling. These findings provide valuable insights for developing future strategies to enhance ECG alarm management.

NICU Health Care Providers Have Alarm-Related Negative Perceptions and Practices

In this study, we found that more than half of the respondents reported that "nuisance alarms disrupt patient care," and 38.37% (n=391) of them reported that "nuisance alarms occur frequently." Altogether, 16.78% (n=171) of respondents reported that alarm-related adverse events related to ECG alarms still occurred in the past 2 years. These results show that there is currently a relative lack of initiative in the management and clinical response to ECG alarms. Most previous studies focused on nurses, with less attention paid to physicians' perceptions and practices regarding ECG alarms [32,36,37]. Our study found that physicians had higher negative perceptions of ECG alarms than nurses.

In addition, we found that health care providers with higher education levels and job titles had more negative perceptions of nuisance alarms and experiences with alarm systems. This may be due to the fact that this population usually has a greater sense of responsibility and vigilance toward ECG alarms [24]. Interestingly, we found that health care providers with lower job titles had more positive attitudes toward improving the current status of ECG alarms through adhering to hospital regulations and the application of smart alarms but had fewer negative perceptions of nuisance alarms for ECG monitors than those with higher job titles. These 2 slightly contradictory results are thought-provoking. They suggested that health care providers with low job titles may be less vigilant about clinical alarms and may not pay as much attention to them, which may pose a hidden risk to clinical safety.

Moreover, in our study, the proportion of female health care providers was much higher than that of male health care providers, which was in line with the general characteristics of the sex ratio of medical staff in maternal and child hospitals in China [38,39]. We found that male health care providers had more negative perceptions regarding experiences with alarm systems, whereas female health care providers had more positive attitudes toward improving the current status of ECG alarms by adhering to hospital regulations. This might be because male health care providers tended to display more impatient and careless personality traits compared to female health care providers. These personality traits make male health care providers less likely to recognize alarm reminders and more prone to negative emotional reactions in their clinical work. In contrast, female health care providers generally demonstrate greater patience, which makes them more receptive to alarm management policies. However, the sample size of male participants included in this study was limited. Future research should investigate the perceptions of clinical practices among male health care providers in larger populations.

XSL•FO
RenderX

NICU Health Care Providers Have Clinical Decision-Making Difficulties With ECG Monitor Parameter Settings

The effectiveness and accessibility of evidence have important impacts on the decision-making support for health care providers [40]. We found that although clinical guidelines were considered to be the most effective decision-making basis for ECG alarm parameter settings, 13.25% (n=135) of the respondents still had difficulties of availability in the clinical settings. It is worth noting that consulting colleagues and leaders remained the most convenient and accessible method for most health care providers, despite the reported limited effectiveness. Hence, the basis of decision-making should be provided in more friendly and convenient ways.

A meta-analysis by Shen and Xiong [41] showed that the parameter setting is an important factor influencing the occurrence of ECG alarm artifacts and poor response to alarms in neonates. This study revealed that decision-making difficulties related to ECG alarms have been significantly underestimated and inadequately addressed, particularly when it comes to setting or modifying the alarm parameters. In addition, we found that health care providers considered "responding to and handling alarms after they occur" to be the most difficult, but interestingly, they also had the highest level of confidence in it. This indicates that the current clinical training is effective in improving the competency of health care providers to respond to and manage alarms [42-44]. Accordingly, we hypothesized that the current clinical training has a limited effect on improving health care providers' competency regarding "timing of setting or changing alarm parameters" and "setting thresholds for alarm parameters."

On the other hand, it is worth noting that the ability to change alarm settings in a timely manner relies on the objective judgment of changes in the patient's condition. This is highly dependent on the professional competence of health care providers, and it could be a fatal weakness especially in health providers with lower job titles. Hence, it is essential to create specialized decision-making support systems or improve decision-making procedures to facilitate this intricate process.

High Demand for Intelligent Clinical Decision-Making Support Systems Among NICU Health Care Providers

Several guidelines and evidence summaries have included intelligent alarm management in their recommendations for ECG monitoring [26,45]. The CDSS uses artificial intelligence techniques to collect and analyze patient data from electronic medical records, based on established guidelines and high-quality clinical evidence [46]. This study showed that intelligent module–assisted decision support systems have become the most popular decision-making support method, and the selection rate is much higher than that of traditional decision applications.

Specific features should include categorizing alarms by priority, differentiating alarm notifications, and filtering unnecessary alarms [47]. In this study, we found that the highest demand was to "develop a standardized device interface to integrate ECG monitors with other devices." This indicates that ECG monitors should be integrated with ventilators and other equipment to provide clinically meaningful alarms and to reduce the incidence of repeated alarms or alarm artifacts. Additionally, we found a high demand for the development of "intelligent decision support for alarm threshold settings for newborns of different birth gestational ages module." This demand was even higher than that for "optimizing algorithm models to reduce alarm artifacts." These results provide important evidence to provide sustainable technical solutions for clinical responses to ECG alarms.

Limitations

This study had several limitations. First, all respondents volunteered to complete the questionnaires and may thus be subject to participant bias. Second, the data relied on self-reported measurements. Therefore, further research should be conducted to design observational studies or qualitative interviews that provide more robust evidence.

Comparison With Prior Work

Evidence Before This Study

Alarm problems associated with ECG monitors were found to be the most prominent among all devices in the NICU. We conducted a comprehensive search of academic databases, including PubMed, Web of Science, Wanfang Data, CNKI, and China Science and Technology Journal Database, using specific keywords "electrocardio/ECG/EKG" and "NICU/ICU/intensive care" in combination with terms such as "clinical alarm/false alarm/alarm management/alarm improvement/alarm fatigue/alarm rejection/alarm apathy." Our search was limited to original publications between January 1, 2020, and August 20, 2023. Previous studies have consistently demonstrated that ECG alarms in the NICU exhibit a characteristic pattern of "high sensitivity and low specificity." Specifically, more than 95% of ECG alarms in this setting were identified as false alarms [5]. Prolonged exposure to such alarms can lead to alarm fatigue among health care providers, potentially resulting in disregarding important signals and an increased risk of adverse events for patients. Moreover, repeated exposure to alarm sounds can have detrimental effects on neonatal infants, including disruptions in the development of the auditory cortex, neurological disorders, and impairments in brain development. While previous studies have investigated the knowledge of clinical staff regarding ECG alarms, they have often neglected to examine their actual clinical practices. Consequently, the challenges and decision-making processes involved in ECG alarm management remain unclear.

Added Value of This Study

This was a cross-sectional survey covering most provinces and all geographical regions of China. To our knowledge, the study was the first national study to investigate both physicians and nurses, which provided reliable data for assessing practices and perceptions of clinical ECG alarms and summarized the related characteristics. It was also noted that decision-making difficulties and decision-support requirements of ECG alarms are greatly underrecognized and undertreated, especially for resetting or modifying alarm parameters. Additionally,

XSL•FO
RenderX

intelligent module–assisted decision support systems have become the most popular form of decision support.

Implications of All the Available Evidence

The findings demonstrate that all NICU staff members reported negative experiences with ECG alarms, encompassing both clinical handling and management initiatives. Nevertheless, physicians had more negative perceptions toward ECG alarms than nurses, emphasizing the need for greater attention toward the physician population. Additionally, the study highlighted that over half of the NICU staff reported difficulties in making decisions related to resetting or modifying alarm parameters. There was a significant demand for the implementation of intelligent module–assisted decision support systems to address this challenge.

Conclusions

This cross-sectional survey highlights the negative perceptions and practices as well as the strong decision-making requirements of NICU staff related to ECG alarm handling. Despite existing routine management measures, decision-making support related to ECG alarms remains a significant concern in the NICU. Intelligent CDSSs may become the most favored tool for overcoming the timing and threshold of ECG parameter settings.

Acknowledgments

The authors would like to thank the Neonatal Respiratory Group of the Chinese Physicians Association as well as Professors Li Xie and Bin Sun for their experimental support and statistical and professional consultation. This study was supported by the research project from Shanghai Jiao Tong University “Medical-Engineering Cross Research Fund—Clinical Transformation Project” (YG2023LC12), Shanghai Pudong New Area Science and Technology Development Fund Special Project for People’s Livelihood Scientific Research (PKJ2021-Y42), the Fundamental Research Funds for the Central Universities (YG2022QN095), Hainan Provincial Natural Science Foundation of China (824QN411), Nursing Development Program from Shanghai Jiao Tong University School of Medicine (SJTUHLXK22023), Fujian Provincial Office of Science and Technology Foundation of China (2024Y0040), Shanghai Public Health Excellent Talent Project (GWVI-11.2-YQ58), and high-level local university construction project founded by Shanghai Municipal Education Commission. The funding source played no role in the study design, data collection, data analysis, data interpretation, or manuscript writing. The corresponding author had full access to all data in the study and had the final responsibility for the decision to submit the data for publication.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy and confidentiality concerns but are available from the corresponding author on reasonable request.

Authors' Contributions

WG, XT, and LZ contributed to the conceptualization and design of the study. WG, XT, J Yang, HZ, and QJ were responsible for data collection and data cleaning. XT, XY, and HZ performed the statistical analyses and were involved in data organization and presentation. XT, XY, and J Yuan drafted the manuscript. WG and LZ critically reviewed the manuscript, providing suggestions for improvements and revisions. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A survey on the current clinical status of ECG alarm management among NICU healthcare providers.
[DOCX File , 34 KB - jmir_v26i1e60944_app1.docx]

References

1. Zhao HX, Wang W. Advances in the study of cardiac monitor alarm fatigue in healthcare workers. J Nurs 2015;17:26-29. [doi: 10.1542/gr.34-3-27]
2. Hayhurst C. Right alert, right time: developing clinically meaningful, actionable alarm notifications. Biomed Instrum Technol 2020;54(1):12-20. [doi: 10.2345/0899-8205-54.1.12] [Medline: 31961741]
3. Yang XK, Wang Y. Status quo and research progress of alarm in clinical monitoring instruments. Chin Nurs Res 2018;32:506-509. [doi: 10.3969/j.issn.1009-6493.2018.04.002]
4. Cho OM, Kim H, Lee YW, Cho I. Clinical alarms in intensive care units: perceived obstacles of alarm management and alarm fatigue in nurses. Healthc Inform Res 2016;22(1):46-53 [FREE Full text] [doi: 10.4258/hir.2016.22.1.46] [Medline: 26893950]
5. Li T, Matsushima M, Timpson W, Young S, Miedema D, Gupta M, et al. Epidemiology of patient monitoring alarms in the neonatal intensive care unit. J Perinatol 2018;38(8):1030-1038 [FREE Full text] [doi: 10.1038/s41372-018-0095-x] [Medline: 29740183]

XSL•FO
RenderX

6. He X, Liu M, Liu S, Sun Z, Wang C, Li Y. Research on Patient Monitoring Alarms in ICU and NICU. Zhongguo Yi Liao Qi Xie Za Zhi 2021;45(5):585-590. [doi: 10.3969/j.issn.1671-7104.2021.05.024] [Medline: 34628778]
7. Chen T, Wang XL, Wu H. Alarm data analysis of monitor physiological parameters. Chin Med Equip 2022 Apr:43-73. [doi: 10.19745/j.1003-8868.2022080]
8. Li MZ, Rao XH, Zhang ZH, Tu JQ, Peng MH. The influential factors on bedside management of medical equipment alarm among ICU nurses: a qualitative study. J Nurs 2020;35(23):49-51. [doi: 10.3870/j.issn.1001-4152.2020.23.049]
9. Yao MX, Wu GL, Peng ZH, Chen JJ, Lu XX. Behavioral intention of alarm management in ECG monitor of nurses in neonatology department from the perspective of planned behavior theory: a qualitative study. Chin Nurs Manag 2023;23(2):215-219. [doi: 10.3969/j.issn.1672-1756.2023.02.012]
10. Lewandowska K, Weisbrot M, Cieloszyk A, Mędrzycka-Dąbrowska W, Krupa S, Ozga D. Impact of alarm fatigue on the work of nurses in an intensive care environment—a systematic review. Int J Environ Res Public Health 2020;17(22):8409 [FREE Full text] [doi: 10.3390/ijerph17228409] [Medline: 33202907]
11. Joint Commission. Medical device alarm safety in hospitals. Sentinel Event Alert 2013(50):1-3. [Medline: 23767076]
12. Jia L, Wang H, Gao Y, Liu H, Yu K. High incidence of adverse events during intra-hospital transport of critically ill patients and new related risk factors: a prospective, multicenter study in China. Crit Care 2016;20:12 [FREE Full text] [doi: 10.1186/s13054-016-1183-y] [Medline: 26781179]
13. Mitka M. Joint commission warns of alarm fatigue: multitude of alarms from monitoring devices problematic. JAMA 2013;309(22):2315-23156. [doi: 10.1001/jama.2013.6032] [Medline: 23757063]
14. Al-Ani RM. Various aspects of hearing loss in newborns: a narrative review. World J Clin Pediatr 2023;12(3):86-96 [FREE Full text] [doi: 10.5409/wjcp.v12.i3.86] [Medline: 37342452]
15. Venkataraman R, Kamaluddeen M, Amin H, Lodha A. Is less noise, light and parental/caregiver stress in the neonatal intensive care unit better for neonates? Indian Pediatr 2018;55(1):17-21. [doi: 10.1007/s13312-018-1220-9]
16. Committee on Environmental Health. Noise: a hazard for the fetus and newborn. Pediatrics 1997;100:724-727. [doi: 10.1542/peds.100.4.724]
17. Bonan KCSC, Pimentel Filho C, Tristão RM, Jesus JAL, Campos Junior D. Sleep deprivation, pain and prematurity: a review study. Arq Neuropsiquiatr 2015;73(2):147-154 [FREE Full text] [doi: 10.1590/0004-282X20140214] [Medline: 25742585]
18. Beken S, Önal E, Gündüz B, Çakir U, Karagöz İ, Kemaloğlu YK. Negative effects of noise on NICU graduates' cochlear functions. Fetal Pediatr Pathol 2021;40(4):295-304. [doi: 10.1080/15513815.2019.1710788] [Medline: 31984823]
19. Capriolo C, Viscardi RM, Broderick KA, Nassebeh S, Kochan M, Solanki NS, et al. Assessment of neonatal intensive care unit sound exposure using a smartphone application. Am J Perinatol 2022;39(2):189-194. [doi: 10.1055/s-0040-1714679] [Medline: 32702769]
20. Zhang SW, Li LL, Dou YL, Lv TC, Cao Y, Yu L, et al. Investigation on noise status of neonatal intensive care unit. Chin Nurs Res 2022:1036-1098. [doi: 10.12102/j.issn.1009-6493.2022.06.028]
21. The Joint Commission. 2022 National Patient Safety Goals. 2022. URL: https://tinyurl.com/5acaruy6 [accessed 2023-12-20]
22. Patient Safety Goals of the Chinese Hospital Association (2017 Version). The Chinese Hospital Association. URL: https://www.cha.org.cn/site/content/9773fc6a8ebec5fdb57dc97ddf282e14.html [accessed 2023-12-20]
23. Xu JQ, Gao YY, Zhang WT, Li SF, Qin MR, Zou BM, et al. Status quo of the application of clinical decision support system in intensive care. Chin Nurs Res 2022;36(20):3650-3654. [doi: 10.12102/j.issn.1009-6493.2022.20.015]
24. Wang L, He W, Chen Y, Wu Q, Du X, Li Q, et al. Intensive care unit nurses' perceptions and practices regarding clinical alarms: a descriptive study. Nurs Open 2023;10(8):5531-5540 [FREE Full text] [doi: 10.1002/nop2.1792] [Medline: 37101342]
25. Leigher D, Kemppainen P, Neyens DM. Skin preparation and electrode replacement to reduce alarm fatigue in a community hospital intensive care unit. Am J Crit Care 2020;29(5):390-395. [doi: 10.4037/ajcc2020120] [Medline: 32869068]
26. Zhang Y, Chen XP, Shao S, Zhuang YY, Lao YW, Ding ZY, et al. Evidence summary for management of bedside electrocardiogram monitor alarms. Chin J Nurs 2021;56(3):445-451. [doi: 10.3761/j.issn.0254-1769.2021.03.023]
27. Tang WQ, Ma J, Feng X, Zhang ZY, Zhou WJ. Mediating Effect of Mental Resilience on Alarm Fatigue and Occupational Stress in ICU Nurses. China medical herald 2022 Dec;19(36):19-81. [doi: 10.20047/j.issn1673-7210.2022.36.16]
28. Wang Y, Sha LY, Liu Y. The influence of working environment alarm fatigue and emotional labor on intensive care unit nurses' job burnout. Chin Nurs Manag 2022;22(9):1405-1410. [doi: 10.3969/j.issn.1672-1756.2022.09.026]
29. Wong PK, Leung KS, Wong ML. Probabilistic grammar-based neuroevolution for physiological signal classification of ventricular tachycardia. Expert Syst Appl 2019;135:237-248. [doi: 10.1016/j.eswa.2019.06.012]
30. Zhou Y, Zhao G, Li J, Sun G, Qian X, Moody B, et al. A contrastive learning approach for ICU false arrhythmia alarm reduction. Sci Rep 2022;12(1):4689 [FREE Full text] [doi: 10.1038/s41598-022-07761-9] [Medline: 35304473]
31. Manna T, Swetapadma A, Abdar M. Decision tree predictive learner-based approach for false alarm detection in ICU. J Med Syst 2019;43(7):191. [doi: 10.1007/s10916-019-1337-y] [Medline: 31115734]
32. Yang XK. Study on knowledge, attitude and practice of nurses about electrocardiogram monitor alarm. Huazhong University of Science and Technology. 2019. URL: https://tinyurl.com/yc7zcs9z [accessed 2023-11-05]

XSL•FO
RenderX

33. Annual data from National Bureau of Statistics. National Bureau of Statistics. 2023. URL: https://www.stats.gov.cn/sj/ndsj/2023/indexch.htm [accessed 2023-11-05]
34. Funk M, Clark JT, Bauld TJ, Ott JC, Coss P. Attitudes and practices related to clinical alarms. Am J Crit Care 2014;23(3):e9-e18. [doi: 10.4037/ajcc2014315] [Medline: 24786820]
35. Ruppel H, Funk M, Clark JT, Gieras I, David Y, Bauld TJ, et al. Attitudes and practices related to clinical alarms: a follow-up survey. Am J Crit Care 2018;27(2):114-123. [doi: 10.4037/ajcc2018185] [Medline: 29496767]
36. Wang L. Study on the current situations of medical devices clinical alarm management and alarm fatigue among ICU nurses. Army Medical University. 2023. URL: https://tinyurl.com/bdh4apjt [accessed 2023-11-05]
37. He M, Huang Q, Lu H, Gu Y, Hu Y, Zhang X. Call for decision support for high-alert medication administration among pediatric nurses: findings from a large, multicenter, cross-sectional survey in China. Front Pharmacol 2022;13:860438 [FREE Full text] [doi: 10.3389/fphar.2022.860438] [Medline: 35928259]
38. Fu YL, Zhang XH, Li LQ, Zhou M, Wei L. An assessment of NICU healthcare staff members' understanding and application of family participatory nursing. J Chin Nurs 2021;56(12):1812. [doi: 10.3761/j.issn.0254-1769.2021.12.008]
39. National Health Commission of China. 2022 China Health Statistics Yearbook. URL: https://tinyurl.com/nt7764r8 [accessed 2024-10-02]
40. Lu W, Chen XY. Influencing factors of usage intention to nursing decision supporting system. J Nurs 2023;30:18-22 [FREE Full text]
41. Shen Q, Xiong J. Meta-analysis of the influencing factors of false alarm of ECG monitoring in newborn babies. J Nurs Sci 2020;35(23):45-48. [doi: 10.3870/j.issn.1001-4152.2020.23.045]
42. Yang XK, Wang Y, Guo XB, Lu ZH. Knowledge of electrocardiogram monitor alarms among nurses and the influencing factors. J Nurs Sci 2020;35(4):44-47. [doi: 10.3870/j.issn.1001-4152.2020.04.044]
43. Shen F. Construction of evaluation index system of ICU nurses? Core competence and preliminary application. Shandong University. 2020. URL: https://tinyurl.com/3ddam7f3 [accessed 2023-11-05]
44. Nyarko BA, Nie H, Yin Z, Chai X, Yue L. The effect of educational interventions in managing nurses' alarm fatigue: an integrative review. J Clin Nurs 2023;32(14):2985-2997. [doi: 10.1111/jocn.16479] [Medline: 35968774]
45. Li T, Hu HL, Wu X. Summary of evidence for threshold setting of multi-parameter electrocardiograph monitor in intensive care unit. Zhonghua Wei Zhong Bing Ji Jiu Yi Xue 2023;35(6):643-650. [doi: 10.3760/cma.j.cn121430-20230228-00121] [Medline: 37366133]
46. Gong TT, Jin L, Zhen YN, Ding M, Ding Y. Construction and application of early warning model of neonatal diagnosis and treatment based on clinical decision support system. J Nurs Train 2021:636-609. [doi: 10.16821/j.cnki.hsjx.2021.07.007]
47. Burdick KJ, Gupta M, Sangari A, Schlesinger JJ. Improved patient monitoring with a novel multisensory smartwatch application. J Med Syst 2022;46(12):83 [FREE Full text] [doi: 10.1007/s10916-022-01869-1] [Medline: 36261739]

Abbreviations

CDSS: clinical decision support system
ECG: electrocardiographic
NICU: neonatal intensive care unit
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by G Tsafnat; submitted 27.05.24; peer-reviewed by X Lihua, W Ahmed; comments to author 13.08.24; revised version received 06.10.24; accepted 02.11.24; published 20.12.24.

Please cite as:
Tang X, Yang X, Yuan J, Yang J, Jin Q, Zhang H, Zhao L, Guo W
Call for Decision Support for Electrocardiographic Alarm Administration Among Neonatal Intensive Care Unit Staff: Multicenter, Cross-Sectional Survey
J Med Internet Res 2024;26:e60944
URL: https://www.jmir.org/2024/1/e60944
doi:10.2196/60944
PMID:

©Xiaoli Tang, Xiaochen Yang, Jiajun Yuan, Jie Yang, Qian Jin, Hanting Zhang, Liebin Zhao, Weiwei Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Nomogram-Based Model to Predict Primary Hypertension Within the Next Year in Children and Adolescents: Retrospective Cohort Study

Chenlong Qin[1,2], MEng; Li Peng[3], MD; Yun Liu[1,2,4], MD; Xiaoliang Zhang[1,4,5], MEng; Shumei Miao[1,4], MEng; Zhiyuan Wei[1], MD; Wei Feng[1], MD; Hongjian Zhang[1], MEng; Cheng Wan[1,2], MEng; Yun Yu[1,2], MEng; Shan Lu[6], MD; Ruochen Huang[1,2], PhD; Xin Zhang[1,2,4], MEng

[1]Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

[2]Institute of Medical Informatics and Management, Nanjing Medical University, Nanjing, China

[3]Department of Endocrinology and Metabolism, the Fourth Affiliated Hospital, Nanjing Medical University, Nanjing, China

[4]Department of Information, the First Affiliated Hospital, Nanjing Medical University, Nanjing, China

[5]Suqian Hospital, Jiangsu Province Hospital, Suqian, China

[6]Women and Children Department of the First Affiliated Hospital of Nanjing Medical University, Nanjing, China

Corresponding Author:
Xin Zhang, MEng
Department of Information, the First Affiliated Hospital
Nanjing Medical University
No.300 Guang Zhou Road, Nanjing, Jiangsu, 210029, China
Nanjing
China
Phone: 86 13584027490
Email: zx6800@njmu.edu.cn

Abstract

Background: Primary hypertension (PH) poses significant risks to children and adolescents. Few prediction models for the risk of PH in children and adolescents currently exist, posing a challenge for doctors in making informed clinical decisions.

Objective: This study aimed to investigate the incidence and risk factors of PH in Chinese children and adolescents. It also aimed to establish and validate a nomogram-based model for predicting the next year's PH risk.

Methods: A training cohort (n=3938, between January 1, 2008, and December 31, 2020) and a validation cohort (n=1269, between January 1, 2021, and July 1, 2023) were established for model training and validation. An independent cohort of 576 individuals was established for external validation of the model. The result of the least absolute shrinkage and selection operator regression technique was used to select the optimal predictive features, and multivariate logistic regression to construct the nomogram. The performance of the nomogram underwent assessment and validation through the area under the receiver operating characteristic curve, concordance index, calibration curves, decision curve analysis, clinical impact curves, and sensitivity analysis.

Results: The PH risk factors that we have ultimately identified include gender (odds ratio [OR] 3.34, 95% CI 2.88 to 3.86; $P<.001$), age (OR 1.11, 95% CI 1.08 to 1.14; $P<.001$), family history of hypertension (OR 42.74, 95% CI 23.07 to 79.19; $P<.001$), fasting blood glucose (OR 6.07, 95% CI 4.74 to 7.78; $P<.001$), low-density lipoprotein cholesterol (OR 2.03, 95% CI 1.60 to 2.57; $P<.001$), and uric acid (OR 1.01, 95% CI 1.01 to 1.01; $P<.001$), while factor breastfeeding (OR 0.04, 95% CI 0.03 to 0.05; $P<.001$) has been identified as a protective factor. Subsequently, a nomogram has been constructed incorporating these factors. Areas under the receiver operating characteristic curves of the nomogram were 0.892 in the training cohort, 0.808 in the validation cohort, and 0.790 in the external validation cohort. Concordance indexes of the nomogram were 0.892 in the training cohort, 0.808 in the validation cohort, and 0.790 in the external validation cohort. The nomogram has been proven to have good clinical benefits and stability in calibration curves, decision curve analysis, clinical impact curves, and sensitivity analysis. Finally, we observed noteworthy differences in uric acid levels and family history of hypertension among various subgroups, demonstrating a high correlation with PH. Moreover, the web-based calculator of the nomogram was built online.

XSL•FO
RenderX

Conclusions: We have developed and validated a stable and reliable nomogram that can accurately predict PH risk within the next year among children and adolescents in primary care and offer effective and cost-efficient support for clinical decisions for the risk prediction of PH.

(J Med Internet Res 2024;26:e58686) doi:10.2196/58686

KEYWORDS

independent risk factors; prediction model; primary hypertension; clinical applicability; development; validation; pediatrics; electronic health records

Introduction

In the past 20 years, primary hypertension (PH) has become the dominant cause of arterial hypertension (AH) in children aged older than 6 years, especially in adolescents [1]. The American Heart Association suggested that pediatric PH is a condition that is severely underrecognized in the latest scientific statement [2]. In an analysis of a large pediatric health care claims database, PH was found to be approximately 10 times more prevalent than secondary hypertension (SH; 0.2% versus 0.02%) [3]. PH in children and adolescents is the early phase of a condition that exists on a continuum across the life course, with higher BP exposure over time contributing to subclinical outcomes in childhood and CVD events later in life [2,3]. However, the diagnosis of PH is a diagnosis of exclusion, the etiology of PH has not been elucidated yet [2].

For newly diagnosed children and adolescents with AH, screening for PH is challenging and time-consuming. According to the latest guideline proposed by a consensus panel composed of multiple institutions such as the European Society of Cardiology Associations and Councils, the differential diagnosis between primary and SH should include detailed family history, physical examination, and laboratory tests [4]. In case of abnormal examination results, the consensus panel agrees that further diagnostic investigations are needed [4]. This contributes to a time-consuming and costly diagnostic process for PH. The current situation is that inexperienced clinicians cannot make suspected diagnoses to recommend proper examinations and tend to recommend too many examinations, which leads to a waste of medical resources and delayed intervention measures [5]. Therefore, the American Heart Association recommends using clinical decision-support tools for detecting hypertension in children and adolescents to achieve the goal of simple, accurate, and cost-effective risk prediction [2,6,7].

However, all of the following studies have the same deficiencies [8-10]. (1) Neglecting PH: the above studies focused on the disease of AH in children and adolescents, neglecting that PH has become the predominant subtype of AH in children and adolescents over the past two decades, with a more complex diagnostic process and severe long-term effects on organs in adulthood. (2) Inadequate use of electronic health record (EHR) data: the risk prediction models discussed in the above studies were developed using survey data and ignored that EHR data encompasses a broader array of both confirmed and potential PH-related risk factors, which allows for a more accurate reflection of the physical state of children and adolescents. This omission could be one of the reasons behind the suboptimal prediction results observed in the prediction models. (3) None of the aforementioned studies have undergone external validation or prospective validation, which significantly impairs the usability of the prediction model.

To address the above limitations, we analyzed the medical condition of PH in children and adolescents and aimed to develop an accurate, rapid, and cost-effective nomogram to predict their PH risk within the next year based on their EHR data. The nomogram was trained and validated on the data of the First Affiliated Hospital of Nanjing Medical University and performed external validation on the data of the Fourth Affiliated Hospital of Nanjing Medical University and the Suqian Hospital. Additionally, we have developed a free online tool implementing nomograms to assist doctors in accurate and low-cost identification of high-risk populations for PH in a timely manner.

Methods

Study Population and Data Collection

Data from the EHR database of the First Affiliated Hospital of Nanjing Medical University were used for model development. As shown in Figure 1, we identified a cohort of patients aged 6-18 years who had at least two primary clinical visits between January 1, 2008, and July 1, 2023 (5245 participants). Patients were excluded if they had any diagnosis of hypertension, pregnancy, or were using antihypertensive drugs at cohort entry. Participants who had >30% missingness (38 participants) were also excluded. Overall, a total of 5207 participants were finally enrolled after rigorous screening. For the prediction modeling cohort, cases in the training population referred to patients from January 1, 2008, to December 31, 2020 (3938 participants), whereas cases in the validation cohort were patients from January 1, 2021, to July 1, 2023 (1269 participants), and were used for model validation.

An independent cohort from the Fourth Affiliated Hospital of Nanjing Medical University and the Suqian Hospital was used for external validation. In brief, we screened a total of 53 children and adolescents with PH and 523 children and adolescents without PH from the EHR databases of these two hospitals using the same screening process as when modeling.

The demographic and clinical parameters from all the EHR databases were collected, including diagnosis of PH, gender, age, family history of hypertension, family history of diabetes, breastfeeding situation, white blood cell count, lymphocyte count (LYC), lymphocyte percentage (LYP), platelet distribution (PDW), mean platelet volume, plateletcrit, platelet count, red blood cell distribution width-coefficient of variation, mean cell hemoglobin concentration, mean corpuscular hemoglobin, mean

XSL•FO
RenderX

corpuscular volume (MCV), hematocrit, hemoglobin, red blood cell count, basophil count, basophil percentage, eosinophil count, eosinophil percentage, neutrophil count, neutrophil percentage (NEUTP), monocyte count (MONOC), monocyte percentage, retinol-binding protein, kalium, natrium, chlorine, calcium, phosphorus, magnesium, fasting blood glucose (FBG), creatinine, urea, uric acid (UA), total protein, albumin, triglycerides, aspartate aminotransferase, alanine aminotransferase, lactate dehydrogenase, γ-glutamyl transpeptidase, alkaline phosphatase, total bilirubin, direct bilirubin, indirect bilirubin (IBIL), total cholesterol (TC), high-density lipoprotein cholesterol, low-density lipoprotein cholesterol (LDL-C), lipoprotein(a), creatine kinase, adenosine deaminase, white/bulb ratio, globulin, α-hydroxybutyrate dehydrogenase. The family history of diabetes and hypertension was defined by the reported history of diabetes and hypertension in immediate family members.

For participants with an outcome event, the index date is defined as the date of the first PH diagnosis. For those without an outcome event, the index date is set as the date of the last non-PH diagnosis. The EHR data from all participants within the year preceding the respective index date is averaged for model training, validation, and external validation.

Figure 1. Flowchart of the procedure. A total of 5245 participants aged 6-18 years in the EHR database of the First Affiliated Hospital of Nanjing Medical University from January 1, 2008, to July 1, 2023, were enrolled in this study. A total of 3938 participants were finally enrolled in the training cohort to construct the nomogram and 1269 participants were finally enrolled in the validation cohort for nomogram validation, respectively. A total of 576 participants aged 6-18 years in the EHR database of the Fourth Affiliated Hospital of Nanjing Medical University and the Suqian Hospital from January 1, 2008, to July 1, 2023, were finally enrolled for independent external validation. EHR: electronic health record.

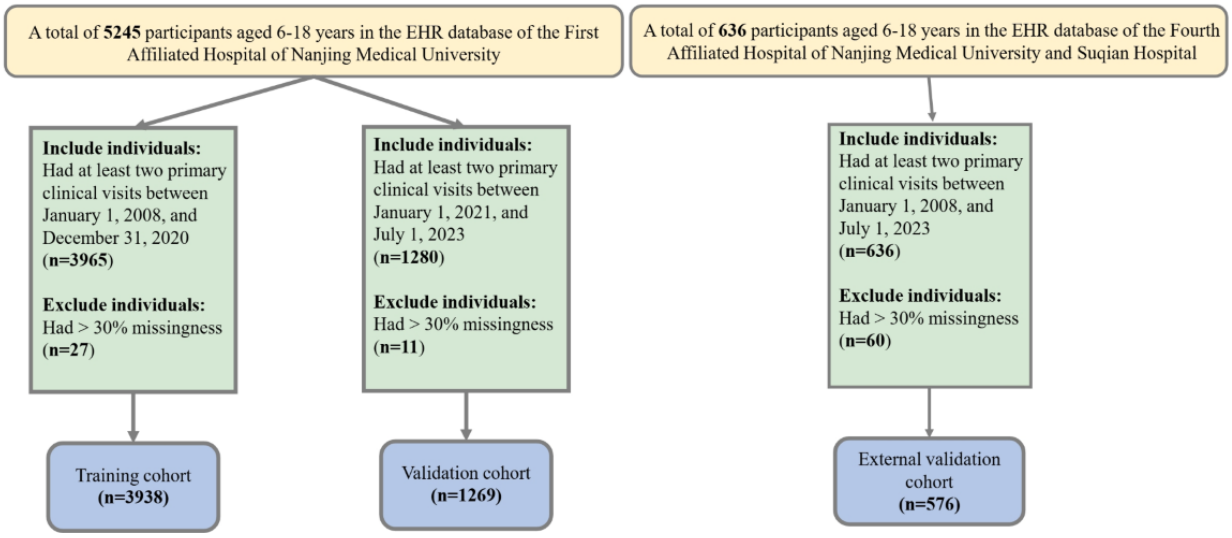

Ethical Considerations

Our study was performed per the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement [11]. All procedures were reviewed and approved by the Ethics Committee of Nanjing Medical University (2023-SR-500). Written informed consent for participation was not required for this study per the national legislation and the institutional requirements. All personal privacy information was well protected and removed during the process of analysis and publication.

Data Preprocessing

A total of 58 variables were reviewed (Table S1 in Multimedia Appendix 1), and 14 variables with missing data exceeding 20% were excluded from the analysis. To address the missing data for the remaining 44 variables, we used chained random forests with predictive mean matching [12]. After imputation, continuous variables were transformed into standardized scores (*z* scores) by subtracting their respective means and dividing by their SDs. Finally, we apply the RandomOverSampler algorithm to the training dataset to solve the problem of sample imbalance [13].

Definition of PH

Our outcome was PH, with the adoption of the *ICD-10-CM* (*International Classification of Diseases, 10th Revision, Clinical Modification*) since October 1, 2015, the diagnosis of PH was defined using the *ICD-10-CM* category I10.

Statistical Analysis

Continuous variables following a normal distribution were reported as means (SD) and analyzed with the Student *t* test (2-tailed) to assess differences between the training and validation cohorts. For continuous variables with skewed distributions, they were described as median (25th percentile, 75th percentile) and analyzed with the Mann-Whitney *U* test. Categorical data were represented as numbers (percent) and analyzed through the chi-square test or Fisher exact test to make comparisons.

We first used the least absolute shrinkage and selection operator (Lasso) regression technique to identify significant features in the training cohort, which can be obtained from the R (R Foundation) package *glmnet* [14]. Then, the multivariable logistic regression model was constructed through sequentially selected candidate variables in the Lasso regression. Predictors were eliminated using the forward stepwise regression strategy based on their respective *P* values to construct the final model

XSL•FO
RenderX

[15]. The nomogram was developed using factors that exhibited a two-sided *P* value <.05 in the multivariate analysis, implemented by the R package *rms*. We used the area under the receiver operating characteristic curve (AUC) and concordance index (C-index) to assess the predictive accuracy and discriminatory ability of the model using the R package *pROC* and *Hmisc*, respectively. The calibration curve was used to evaluate the consistency between nomogram-predicted and observed risks. The decision curve analysis (DCA) was conducted to gauge the net benefit of identifying true high-risk patients that ought to have intervention and the net reduction of unnecessary interventions using the R package *rmda*. The clinical utility of the nomogram was also measured by the clinical impact curves for a population size of 1000. Stratified analyses were displayed by forest plots using the R package *forestploter*.

Statistical analyses were performed with R software (version 4.2.3; RRID:SCR_001905). Two-sided *P* value <.05 was considered to be statistically significant.

Results

Baseline Characteristics

A total of 5207 participants from the EHR database of the First Affiliated Hospital of Nanjing Medical University were finally enrolled after rigorous screening. Table S1 in Multimedia Appendix 2 shows the baseline characteristics of participants in the training and validation cohort. For model training, 3938 participants between January 1, 2008, and December 31, 2020, were used to build the training cohort, and 1269 participants between January 1, 2021, and July 1, 2023, were used to build the validation cohort for model validation. The total prevalence of PH in the whole cohort was 10% (521 participants). The training cohort consisted of 3938 participants, while the validation cohort had 1269 participants assigned to it. The prevalence of PH in the training cohort was 9.8% (385 participants) and 10.7% (136 participants) in the validation cohort, respectively. Of the training cohort, the median age of children and adolescents was 14 years with a range from 6 to 17, and 15 years with a range from 6 to 17 in the validation cohort. More than half of the participants were girls both in the training cohort (55.51%, 2186 participants) and validation cohort (54.5%, 692 participants). There were no significant differences in the characteristics of diagnosis of PH, family history of hypertension, family history of diabetes, gender, age, albumin, calcium, FBG, hematocrit, hemoglobin, IBIL, LDL-C, LYC, LYP, MCV, MONOC, NEUTP, PDW, red blood cell count, TC, triglycerides, UA, and urea between 2 cohorts.

A total of 576 participants from the EHR database of the Fourth Affiliated Hospital of Nanjing Medical University and the Suqian Hospital were used for external validation. There were 53 participants with PH and 523 participants without PH in the external validation cohort. Table S2 in Multimedia Appendix 1 shows the baseline characteristics of participants in the training cohort and external validation cohort.

Development and Construction of Nomogram

According to the Lasso regression model, 38 candidate predictors had nonzero coefficients including gender, age, family history of hypertension, family history of diabetes, breastfeeding, PDW, mean platelet volume, plateletcrit, platelet count, red blood cell distribution width-coefficient of variation, mean cell hemoglobin concentration, MCV, hematocrit, hemoglobin, neutrophil count, NEUTP, MONOC, monocyte percentage, LYC, LYP, kalium, natrium, calcium, FBG, creatinine, UA, albumin, lactate dehydrogenase, total bilirubin, direct bilirubin, TC, high-density lipoprotein cholesterol, LDL-C, lipoprotein(a), white/bulb ratio, globulin, IBIL, and creatine kinase (Figures 2A and 2B). Out of these candidate predictors, we used the forward stepwise regression strategy to construct the final multivariate logistic regression model, which revealed that gender, age, family history of hypertension, breastfeeding, FBG, LDL-C, and UA were independent risk factors for PH (Table 1). Using the coefficient estimates from this final logistic model, nomogram was defined as:

$$\text{Nomogram}_{\text{Score}} = -257.5045 + 45.4545 \times \text{FBG} + 17.8043 \times \text{LDL–C} + 0.2351 \times \text{UA} + 30.3525 \times \text{Gender} + 2.5831 \times \text{AGE} + 94.6279 \times \text{Family history of hypertension} - 83.3383 \times \text{Breastfeeding}$$

These 7 independent factors were used to construct the nomogram (Figure 3).

XSL•FO
RenderX

Figure 2. Predictor selection using the Lasso binary logistic regression model. (A) The optimal penalization coefficient λ in the Lasso model was identified through 10-fold cross-validation in the training cohort. (B) Lasso coefficient profiles of all predictors. The trajectory of each coefficient related to PH features was observed within the Lasso coefficient profiles as the λ parameter changed in the Lasso algorithm. Lasso: least absolute shrinkage and selection operator.

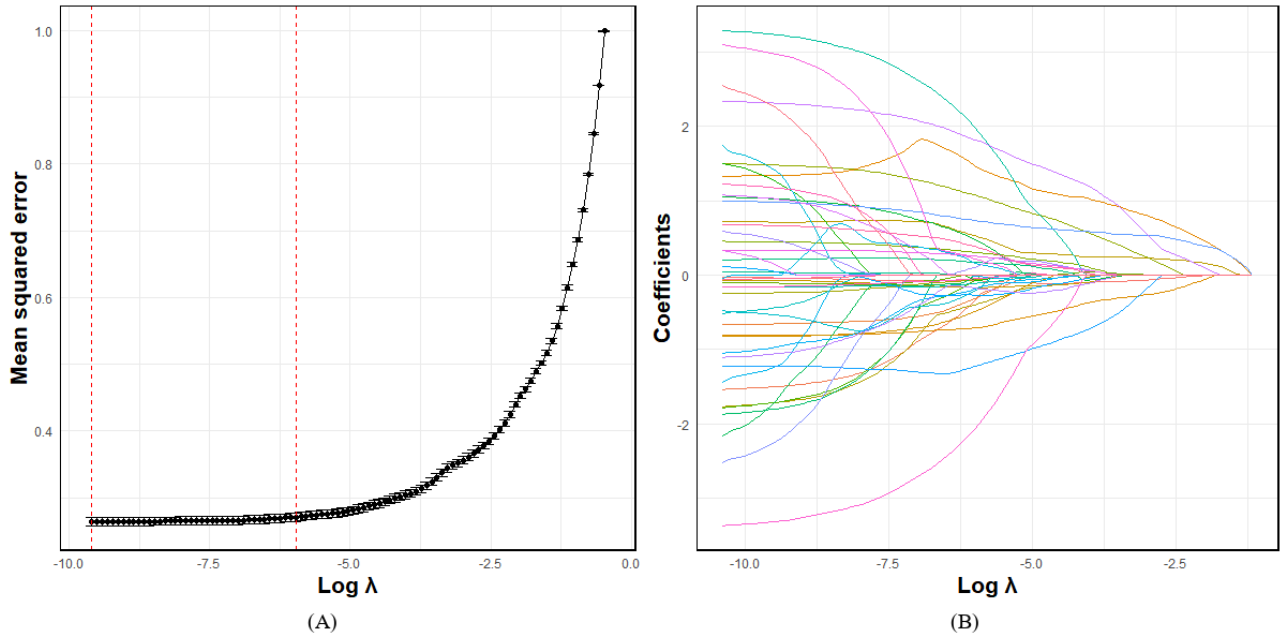

Table 1. Risk factors for primary hypertension in the training cohort.

Predictors	β coefficient	Odds ratio (95% CI)	*P* values
Family history of hypertension			
No	Reference	—[a]	—
Yes	3.76	42.74 (23.07 to 79.19)	<.001
Breastfeeding			
No	Reference	—	—
Yes	–3.31	0.04 (0.03 to 0.05)	<.001
Gender			
Female	Reference	—	—
Male	1.2	3.34 (2.88 to 3.86)	<.001
Age (years)	0.1	1.11 (1.08 to 1.14)	<.001
Fasting blood glucose	1.8	6.07 (4.74 to 7.78)	<.001
Low-density lipoprotein cholesterol	0.71	2.03 (1.6 to 2.57)	<.001
Uric acid	0.01	1.01 (1.01 to 1.01)	<.001

[a]Not applicable.

XSL•FO
RenderX

Figure 3. Nomogram for the prediction of primary hypertension. To use the nomogram, add up the scores for each feature of the participants to calculate the total score. Then, draw a vertical line at the corresponding position on the total score axis to determine the associated risk of PH. For example, a girl aged 10 years with a family history of hypertension and a history of breastfeeding, FBG 6 mmol/L, LDL-C 2 mmol/L, and UA 400 umol/L has a total score of 10+0+95+0+91+10+59=265. This indicates that the risk of developing PH is 83%. FBG: fasting blood glucose; LDL-C: low-density lipoprotein cholesterol; PH: primary hypertension; UA: uric acid.

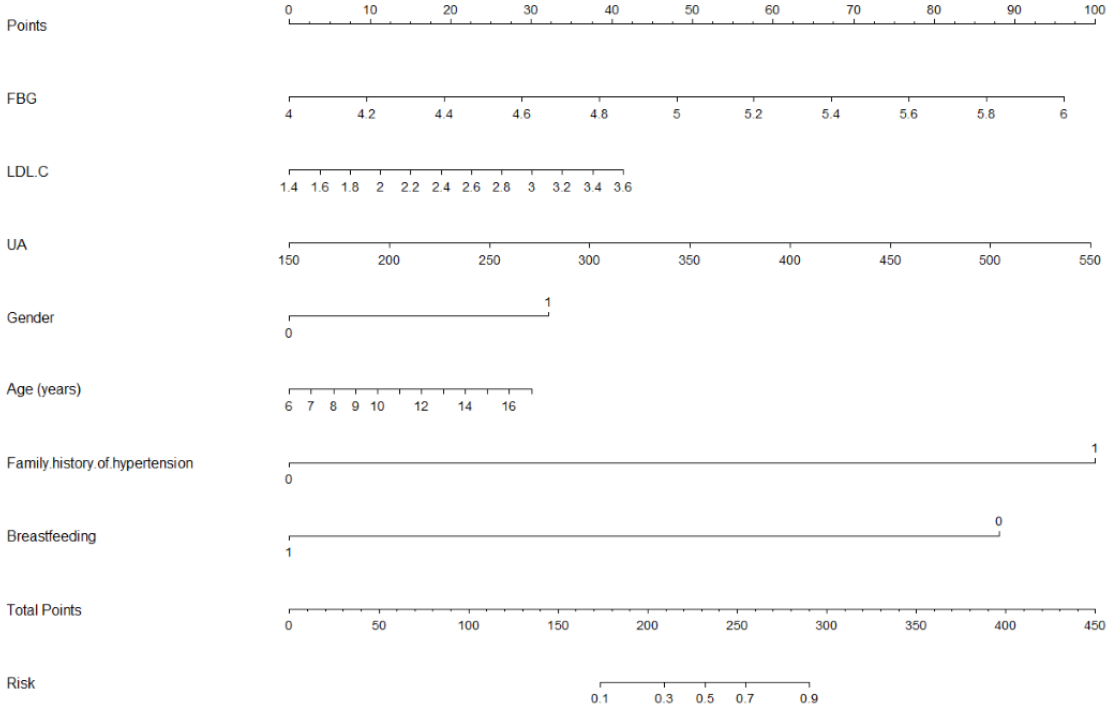

Assessment of the Nomogram

The nomogram was constructed to predict the risk of PH in children and adolescents by using gender, age, family history of hypertension, breastfeeding, FBG, LDL-C, and UA. First, we calculated the total nomogram points for each participant and divided the points into two subgroups according to the median. Figure 4 shows that the risk of PH increased with the total nomogram points, and participants in the high point subgroup (total points: 97.61-303.29) had a significantly higher PH risk than those in the low point subgroup. We next constructed the AUC and C-index to assess the performance of the model. The AUC value of the nomogram in the training cohort was 0.892 (95% CI 0.884-0.899), and the AUCs of gender, age, family history of hypertension, breastfeeding, FBG, LDL-C, and UA were 0.692 (95% CI 0.681-0.702), 0.744 (95% CI 0.733-0.756), 0.555 (95% CI 0.549-0.560), 0.735 (95% CI 0.727-0.744), 0.567 (95% CI 0.553-0.581), 0.670 (95% CI 0.657-0.684), and 0.782 (95% CI 0.771-0.794), respectively (Figure 5A). The C-index of the nomogram was 0.892 (95% CI 0.884-0.899) in the training cohort, signifying its robust predictive discriminatory ability. Moreover, the calibration curve of the training cohort revealed a high consistency between prediction and actual observation (Figure 6A).

Figure 4. Association between the total points of the nomogram and primary hypertension.

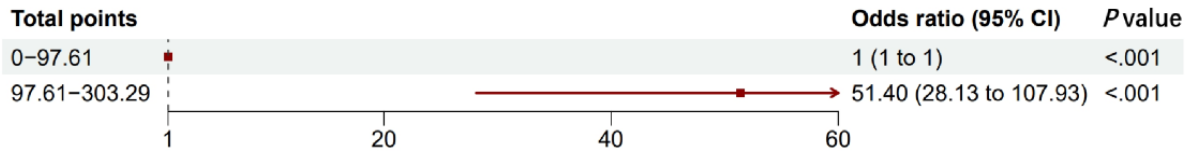

XSL•FO
RenderX

Figure 5. ROC curves for the prediction of primary hypertension in the training, validation, and external validation cohort. (A) ROC curves of the factors and nomogram in the training cohort. (B) ROC curves of the factors and nomogram in the validation cohort. (C) ROC curves of the factors and nomogram in the external validation cohort. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.

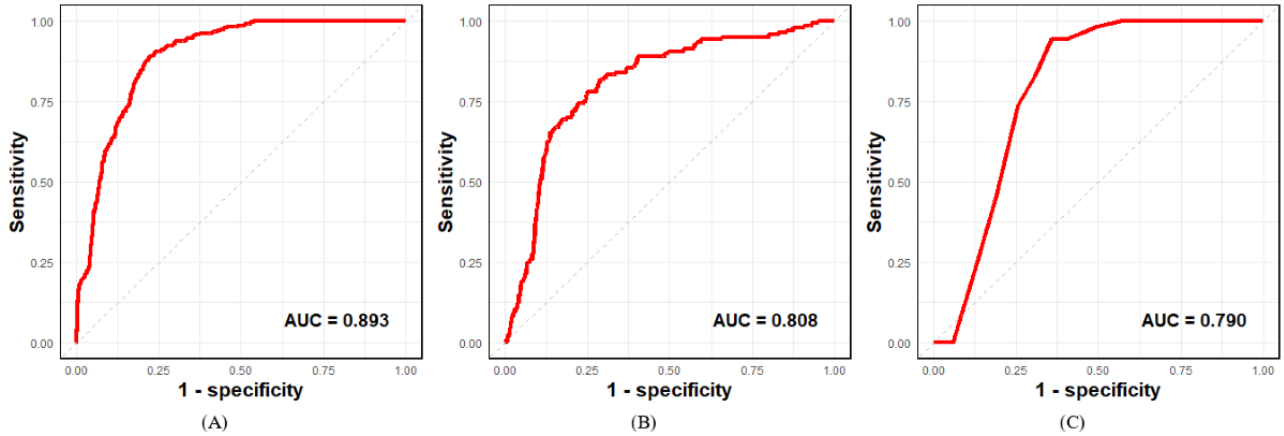

Figure 6. Calibration curves of the nomogram prediction in the training, validation, and external validation cohort. (A) Calibration curves of nomogram prediction in the training cohort. (B) Calibration curves of nomogram prediction in the validation cohort. (C) Calibration curves of nomogram prediction in the external validation cohort. PH: primary hypertension.

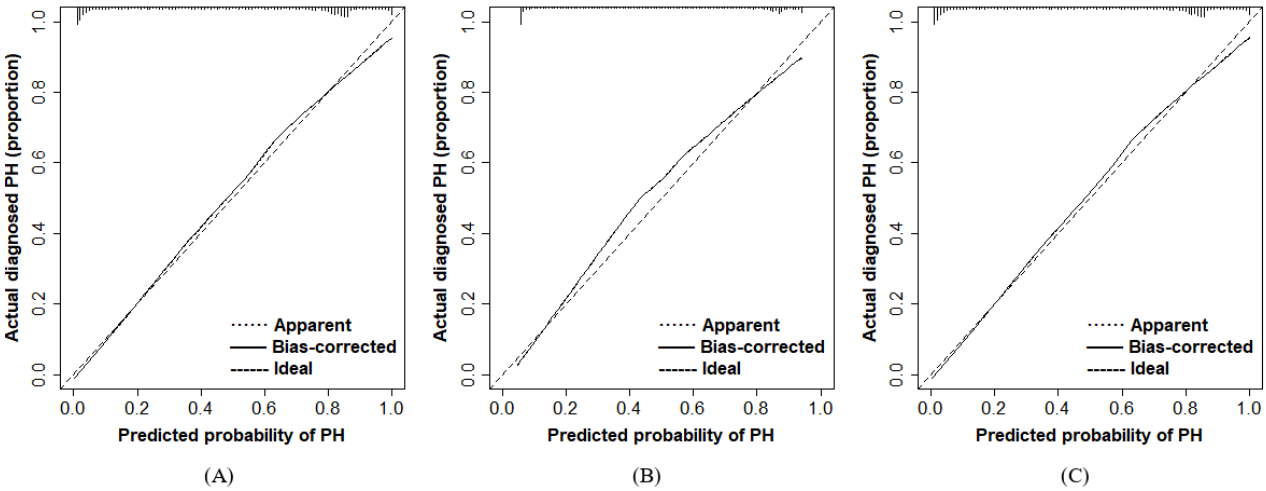

Validation of the Nomogram

The data of the validation cohort were used to independently validate the nomogram. In the validation cohort, the AUC value of the nomogram was 0.808 (95% CI 0.769-0.846; Figure 5B), and the AUC of UA was 0.742 (95% CI 0.692-0.791). The nomogram exhibited a C-index of 0.808 (95% CI 0.770-0.846) in the validation cohort, indicating commendable predictive discrimination within the model. However, the calibration curve revealed a relatively modest consistency between predictions and actual observations (Figure 6B).

We have performed an external validation of the nomogram using an independent cohort. The AUC value of the nomogram was 0.790 (95% CI 0.751-0.830), and the C-index of the nomogram was 0.790 (95% CI 0.764-0.849) in the external validation (Figure 5C). Moreover, the calibration curve revealed a high consistency between prediction and actual observation in the external validation (Figure 6C).

Clinical Net Benefits With Nomogram

The DCA illustrated that the nomogram conferred more clinical net benefits than several competing intervention strategies, namely, intervention for all and intervention for none, as depicted in Figure 7A. Specifically, with a threshold probability of 0.8, the nomogram presented the best net benefit. Similar findings were observed in the validation cohort represented in Figure 7B and the external validation cohort represented in Figure 7C. The clinical impact curves for the nomogram revealed a close alignment between predicted and actual probabilities in the training cohort, as illustrated in Figure 8A. Comparable results were observed in the validation cohort (Figure 8B) and the external validation cohort (Figure 8C).

XSL•FO
RenderX

Figure 7. DCA of the nomogram prediction in the training, validation, and external validation cohort. (A) DCA of nomogram prediction in the training cohort. (B) DCA of nomogram prediction in the validation cohort. (C) DCA of nomogram prediction in the external validation cohort. DCA: decision curve analysis.

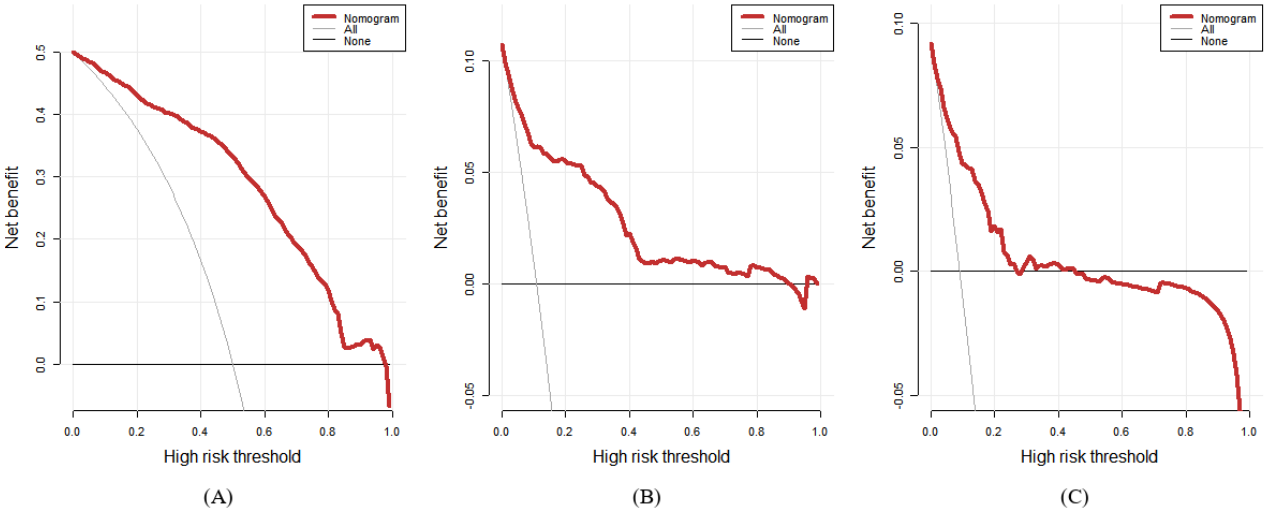

Figure 8. Clinical impact curves of the nomogram prediction in the training, validation, and external validation cohort. (A) Clinical impact curves of nomogram prediction in the training cohort. (B) Clinical impact curves of nomogram prediction in the validation cohort. (C) Clinical impact curves of nomogram prediction in the external validation cohort.

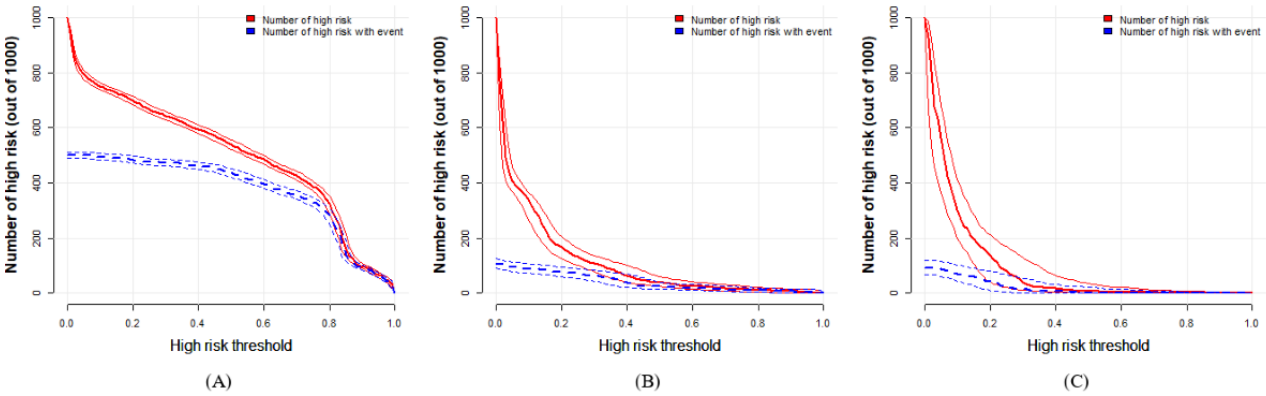

Sensitivity Analysis of Nomogram Prediction

To assess the robustness of the nomogram, we performed a series of subgroup analyses with subgroups defined by age, gender, and PH status. The nomogram had reasonable AUCs in all of these subpopulations, ranging from 0.796 (95% CI 0.773-0.819) to 0.960 (95% CI 0.931-0.990) in the training cohort, 0.730 (95% CI 0.695-0.764) to 0.933 (95% CI 0.851-0.999) in the validation cohort, and 0.716 (95% CI 0.694-0.832) to 0.900 (95% CI 0.808-0.926) in the external validation cohort (Figures 9A-9C). It is worth noting that the nomogram performed better in the population aged 11 years or younger than in the population aged older than 11 years in the sensitivity analysis across all three cohorts.

XSL•FO
RenderX

Figure 9. Subgroup analyses of nomogram. (A) Prediction accuracy for PH in the training cohort. (B) Prediction accuracy for PH in the validation cohort. (C) Prediction accuracy for PH in the external validation cohort. PH: primary hypertension.

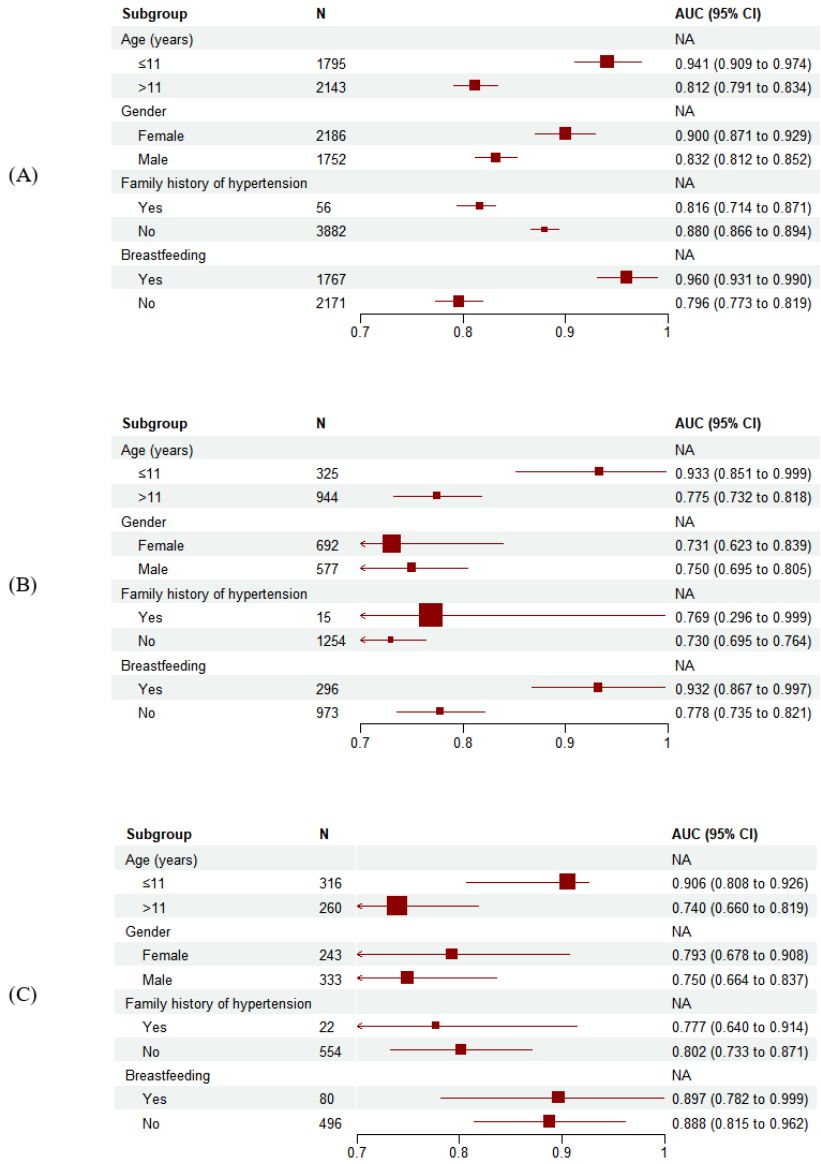

Website of Nomogram

A web-based nomogram calculator for PH has been developed and is freely accessible online [16]. This tool is designed to provide diagnostic probability, facilitating guardians and physicians in the user-friendly identification of PH among children and adolescents.

Discussion

Principal Results

In this study, we have constructed and validated a risk prediction model of future 1-year incident PH in children and adolescents. The EHR-based nomogram achieved 0.892, 0.808, and 0.790 AUCs in training, validation, and external validation cohorts, respectively. Moreover, the DCA and clinical impact curve indicated that a substantial proportion of the threshold probabilities in this model yielded favorable net benefits. In sensitivity analysis, the nomogram showed relatively stable and consistent results in subgroups of age, gender, and PH status, making it the most sensitive predictive tool for PH risk while ensuring accuracy.

Accurately, timely, and low-cost identification of children and adolescents at high risk of PH is of great significance. Previous studies have preliminarily constructed risk prediction models targeting the hazards of AH [8-10]. However, as the leading type of AH in children and adolescents, the differential diagnosis of PH is more challenging and time-consuming. PH children are highly likely to become PH adults and to have measurable target organ injury if timely intervention is not applied. It is therefore suggested to build a model to identify children and adolescents at high risk of PH and intervene promptly to minimize harm as much as possible.

Our study differed from previous research in that we constructed a nomogram using EHR data to achieve initial screening of children and adolescents at high PH risk. Compared to survey data, EHR data encompasses a broader array of both confirmed and potential PH-related risk factors, which allows for a more accurate reflection of the physical state of children and

XSL•FO
RenderX

adolescents. Furthermore, EHR data is readily accessible in clinical settings, providing a reliable and objective source of information unaffected by subjective influences from patients. Meanwhile, EHR data can be shared among different medical institutions, which is beneficial for the external validation of predictive models. The application of the EHR-based nomogram in primary care institutions can effectively improve the diagnosis of pediatric PH, reducing its difficulty and time consumption. This suggests that our model can serve as a clinical decision-support tool to assist clinicians in screening children and adolescents with high PH risk.

We used Lasso as a variable selection technique to reduce the number of parameters, thereby reducing the overfitting and complexity of the model, making it easier to interpret, and reducing operational complexity. Previous studies have demonstrated that Lasso regression is a widely used method for high dimensional predictor selection and nomogram construction [10,17], it shrinks the size of the coefficients of the independent variables depending on their predictive power. Some coefficients may shrink down to zero, allowing us to restrict the model to variables with nonzero coefficients. In our study, we calculated the mean squared error for each variable individually. The λ that provided the minimum mean squared error (minMSE) on the data was 0.00007 and the λ with the minMSE + 1 SE of minMSE (minMSE + 1 SE) was 0.00308. Considering that after the value of λ reached a certain threshold, further increasing the number of independent variables in the model did not significantly improve the model's performance. Therefore, we decided to use the minMSE + 1 SE criterion to select the best predictive feature. This approach allowed us to maintain model performance while simultaneously reducing model complexity and mitigating overfitting.

The nomogram incorporated 7 parameters including gender, age, family history of hypertension, breastfeeding, FBG, LDL-C, and UA. Consistent with prior research findings, our study also identified gender and age as potential high-risk factors for PH [1,18,19]. Notably, there was a more pronounced sex-related disparity, indicating a higher ratio of boys to girls (3-4:1) among adolescents with PH (Figure S1 in Multimedia Appendix 1). To delve deeper into the variations in PH influenced by gender and age, we conducted a subgroup analysis. We first categorized the participants into two subgroups based on their PH status both in the training cohort and validation cohort (Table S3 in Multimedia Appendix 1). In the training cohort, males constituted a substantial majority among PH patients, comprising 77.7% (299 participants), in stark contrast to females who accounted for only 22% (86 participants). Additionally, individuals without PH were generally younger than their counterparts with PH. These trends were consistently observed in the validation cohort. Then, we divided the participants into two subgroups using the age of 11 years as the cutoff point in the training and validation cohorts (Table S4 in Multimedia Appendix 1). In the training cohort, diagnoses with PH were recorded for 1% (n=25) of participants aged 11 years or younger and 16.8% (360 participants) of participants aged older than 11 years, the results in the validation cohort are approximately similar. Finally, we divided both the training and validation cohorts into male and female subgroups based on gender (Table S5 in Multimedia Appendix 1). In the training cohort, participants with PH were recorded for 4% (86 participants) of females and 17.1% (299 participants) of males, and 3% (21 participants) of females and 19.9% (115 participants) of males in the validation cohort, respectively.

We observed significant differences in UA levels among different subgroups in all 3 subgroup analyses. Accumulating evidence supports the notion that elevated UA levels in children and adolescents are an independent risk factor for hypertension. In 2000, Fujiwara et al [20] confirmed that elevated UA levels are a typical characteristic of PH in children and adolescents. A recent report from the Study of High Blood Pressure in Pediatrics, Adult Hypertension Onset in Youth highlighted that mean UA concentrations increased from 5.3 to 5.9 mg/dL as blood pressure values escalated from below the 80th to above the 90th percentile [21]. Moreover, a cohort study with an average follow-up time of 5.7 years also confirmed that higher baseline UA levels increased the risk of incident hypertension (hazard ratio 1.19, 95% CI 1.03-1.38) [22]. This aligns with the results observed in our research, further reinforcing the association between elevated UA levels and an increased risk of PH in children and adolescents.

Our research findings indicate that children and adolescents who were breastfed exhibit a lower risk of PH compared to those who were not breastfed. Several studies have demonstrated that breastfeeding, considered an environmental factor, exerts a protective effect on blood pressure [23-25]. Moreover, our study revealed that children and adolescents with a positive family history of hypertension face an elevated risk of developing PH. While there is a widespread consensus on the genetic factors contributing to PH in children and adolescents, the specific underlying mechanism still requires further research [26-28]. In our study, we observed a noteworthy escalation in the risk of PH among children and adolescents with rising levels of FBG. The reason may be that insulin resistance and hyperinsulinemia can trigger the activation of the sympathetic nervous system, representing key mechanisms associated with the pathophysiology of obesity-related hypertension in children and adolescents [29-32].

Evidence has been raised that childhood obesity is linked not only to the onset of PH during childhood but also to an elevated risk of cardiovascular disease in adulthood [33-36]. Giussani et al [37] also confirmed elevated BMI is the most potent risk factor for PH in children and adolescents in one of their studies. In 2012, an analysis included data from 63,025 overweight and obese children and adolescents found that high LDL-C levels are related to an increased BMI. In 2012, an analysis incorporating data from 63,025 overweight and obese children and adolescents revealed an association between elevated levels of LDL-C with increased BMI [38]. In our research, we observed a correlation between obesity and PH in children and adolescents. Our findings indicated that with the elevation of LDL-C values, the risk of PH in children and adolescents also increased. This further substantiates the robust connection between obesity and PH in this demographic.

Importantly, all of these predictors are routinely accessible through standard clinical examinations. This implies that the

XSL•FO
RenderX

model we have established in our study could serve as a potentially valuable tool for the rapid assessment of PH in children and adolescents, surpassing the capabilities of previous models. In the macro-view, identifying the PH predictors for youths would relieve the burden of monitoring and administering youth health and economizing facilities and resources. Furthermore, these identifications can facilitate the implementation of early interventions for those at risk. These interventions encompass a range of strategies, including but not limited to instilling disease awareness in children and adolescents from an early age and preemptively averting the transition to other potential hazards, such as adverse cardiovascular events.

Limitations

There are also some limitations in our study. First, due to various factors such as cuff size, poor compliance, and rapid changes in the body during growth, medical institutions often struggle to record accurate BP in children and adolescents [39-42]. Bearing this in mind, we decided not to include BP in our study. Second, as previously mentioned, distinguishing between PH and SH can be challenging; thus, it is possible that some individuals diagnosed with SH may have been inadvertently included in this study. Moreover, in practical applications, the model faces limitations regarding input and output data formats. Third, the predictive accuracy of the nomogram can still be improved, which may imply the need for the inclusion of additional factors such as physical examination results and survey data. The prediction accuracy could perhaps be improved in further studies with a wider population coverage, larger sample size, and more variables.

Conclusions

To promptly and accurately identify children and adolescents at high risk of PH with maximum efficiency and at a low cost, aiming to minimize the potential harm to this population, we develop and validate a straightforward and dependable nomogram. The nomogram uses conventional EHR data to predict risk with strong accuracy, discrimination, and clinical utility in both training and validation cohorts, signifying its robust performance in practical applications. This visual model and accompanying website will assist patients and physicians in forecasting the PH risk and enhance clinical management in a timely manner.

Acknowledgments

This study was supported by the Ministry of Science and Technology of the People's Republic of China National Key R&D Program (2023YFC3605800), the Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX23_1983), the External Collaboration Project (2022), the Nanjing Science and Technology Bureau Project (202205053), and the Social Development Plan of the Provincial Department of Science and Technology in Jiangsu Province (BE2023781).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables and figures for the study.
[DOCX File , 143 KB - jmir_v26i1e58686_app1.docx]

Multimedia Appendix 2
Table of baseline characteristics of individuals in the training cohort and validation cohort.
[DOCX File , 19 KB - jmir_v26i1e58686_app2.docx]

References

1. Gupta-Malhotra M, Banker A, Shete S, Hashmi SS, Tyson JE, Barratt MS, et al. Essential hypertension vs. secondary hypertension among children. Am J Hypertens 2015;28(1):73-80 [FREE Full text] [doi: 10.1093/ajh/hpu083] [Medline: 24842390]
2. Falkner B, Gidding SS, Baker-Smith CM, Brady TM, Flynn JT, Malle LM, On behalf of the American heart association council on hypertension; council on lifelong congenital heart disease and heart health in the young; council on kidney in cardiovascular disease; council on lifestyle and cardiometabolic health; and council on cardiovascular and stroke nursing. Pediatric primary hypertension: an underrecognized condition: a scientific statement from the American Heart Association. Hypertension 2023;80(6):e101-e111 [FREE Full text] [doi: 10.1161/HYP.0000000000000228] [Medline: 36994715]
3. Welch WP, Yang W, Taylor-Zapata P, Flynn JT. Antihypertensive drug use by children: are the drugs labeled and indicated? J Clin Hypertens (Greenwich) 2012;14(6):388-395 [FREE Full text] [doi: 10.1111/j.1751-7176.2012.00656.x] [Medline: 22672093]
4. De Simone G, Mancusi C, Hanssen H, Genovesi S, Lurbe E, Parati G, et al. Hypertension in children and adolescents. Eur Heart J 2022;43(35):3290-3301. [doi: 10.1093/eurheartj/ehac328] [Medline: 35896123]
5. Rimoldi SF, Scherrer U, Messerli FH. Secondary arterial hypertension: when, who, and how to screen? Eur Heart J 2014;35(19):1245-1254 [FREE Full text] [doi: 10.1093/eurheartj/eht534] [Medline: 24366917]

XSL•FO
RenderX

6. Lee H, Cho JK, Park J, Lee H, Fond G, Boyer L, et al. Machine learning-based prediction of suicidality in adolescents with allergic rhinitis: derivation and validation in 2 independent nationwide cohorts. J Med Internet Res 2024;26:e51473 [FREE Full text] [doi: 10.2196/51473] [Medline: 38354043]
7. He F, Ng Yin Ling C, Nusinovici S, Cheng CY, Wong TY, Li J, et al. Development and external validation of machine learning models for diabetic microvascular complications: cross-sectional study with metabolites. J Med Internet Res 2024;26:e41065 [FREE Full text] [doi: 10.2196/41065] [Medline: 38546730]
8. Vos LE, Oren A, Bots ML, Gorissen WH, Grobbee DE, Uiterwaal CS. Does a routinely measured blood pressure in young adolescence accurately predict hypertension and total cardiovascular risk in young adulthood? J Hypertens 2003;21(11):2027-2034. [doi: 10.1097/00004872-200311000-00011] [Medline: 14597845]
9. Juhola J, Oikonen M, Magnussen CG, Mikkilä V, Siitonen N, Jokinen E, et al. Childhood physical, environmental, and genetic predictors of adult hypertension: the cardiovascular risk in Young Finns study. Circulation 2012;126(4):402-409. [doi: 10.1161/circulationaha.111.085977]
10. Liang JH, Zhao Y, Chen YC, Huang S, Zhang SX, Jiang N, et al. Development and validation of a nomogram-based prognostic model to predict high blood pressure in children and adolescents-findings from 342,736 individuals in China. Front Cardiovasc Med 2022;9:884508 [FREE Full text] [doi: 10.3389/fcvm.2022.884508] [Medline: 35811689]
11. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015;13:1 [FREE Full text] [doi: 10.1186/s12916-014-0241-z] [Medline: 25563062]
12. Wright MN, Ziegler A. A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Software 2017;77(1):1-17. [doi: 10.18637/jss.v077.i01]
13. Han H, Wang WY, Mao BH, Borderline SMOTE. In: Huang DS, Zhang XP, Huang GB, editors. A New Over-Sampling Method in Imbalanced Data Sets Learning. Berlin, Heidelberg: Springer; 2005.
14. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
15. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54(8):774-781. [doi: 10.1016/s0895-4356(01)00341-9] [Medline: 11470385]
16. Qin C. Online calculation tool for primary hypertension risk in children and adolescents. URL: https://qinchenlong.github.io/ [accessed 2024-11-06]
17. Deng X, Hou H, Wang X, Li Q, Li X, Yang Z, et al. Development and validation of a nomogram to better predict hypertension based on a 10-year retrospective cohort study in china. URL: https://elifesciences.org/articles/66419 [accessed 2021-05-28]
18. Litwin M, Obrycki Ł, Niemirska A, Sarnecki J, Kułaga Z. Central systolic blood pressure and central pulse pressure predict left ventricular hypertrophy in hypertensive children. Pediatr Nephrol 2019;34(4):703-712 [FREE Full text] [doi: 10.1007/s00467-018-4136-7] [Medline: 30426220]
19. Kit BK, Kuklina E, Carroll MD, Ostchega Y, Freedman DS, Ogden CL. Prevalence of and trends in dyslipidemia and blood pressure among US children and adolescents, 1999-2012. JAMA Pediatr 2015;169(3):272-279 [FREE Full text] [doi: 10.1001/jamapediatrics.2014.3216] [Medline: 25599372]
20. Fujiwara N, Osanai T, Kamada T, Katoh T, Takahashi K, Okumura K. Study on the relationship between plasma nitrite and nitrate level and salt sensitivity in human hypertension : modulation of nitric oxide synthesis by salt intake. Circulation 2000;101(8):856-861. [doi: 10.1161/01.cir.101.8.856] [Medline: 10694524]
21. Urbina EM, Mendizábal B, Becker RC, Daniels SR, Falkner BE, Hamdani G, et al. Association of blood pressure level with left ventricular mass in adolescents. Hypertension 2019;74(3):590-596. [doi: 10.1161/hypertensionaha.119.13027]
22. Bjornstad P, Laffel L, Lynch J, El Ghormli L, Weinstock RS, Tollefsen SE, et al. Elevated serum uric acid is associated with greater risk for hypertension and diabetic kidney diseases in obese adolescents with type 2 diabetes: an observational analysis from the treatment options for type 2 diabetes in adolescents and youth (TODAY) study. Diabetes Care 2019;42(6):1120-1128 [FREE Full text] [doi: 10.2337/dc18-2147] [Medline: 30967435]
23. Martin RM, Gunnell D, Smith GD. Breastfeeding in infancy and blood pressure in later life: systematic review and meta-analysis. Am J Epidemiol 2005;161(1):15-26. [doi: 10.1093/aje/kwh338] [Medline: 15615909]
24. Kelishadi R, Ardalan G, Gheiratmand R, Majdzadeh R, Delavari A, Heshmat R, et al. Blood pressure and its influencing factors in a national representative sample of Iranian children and adolescents: the CASPIAN Study. Eur J Cardiovasc Prev Rehabil 2006;13(6):956-963. [doi: 10.1097/01.hjr.0000219109.17791.b6] [Medline: 17143128]
25. Zarrati M, Shidfar F, Razmpoosh E, Nezhad FN, Keivani H, Hemami MR, et al. Does low birth weight predict hypertension and obesity in schoolchildren? Ann Nutr Metab 2013;63(1-2):69-76. [doi: 10.1159/000351869] [Medline: 23942014]
26. Camci L, Kilic Z, Dinleyici EC, Muslumanoglu H, Tepeli E, Ucar B. Angiotensin-converting enzyme gene insertion/deletion polymorphism frequency in normotensive children with a positive family history of essential hypertension. J Paediatr Child Health 2009;45(12):742-746. [doi: 10.1111/j.1440-1754.2009.01605.x] [Medline: 19863706]
27. Bigazzi R, Zagato L, Lanzani C, Fontana S, Messaggio E, Delli Carpini S, et al. Hypertension in high school students: genetic and environmental factors: the HYGEF study. Hypertension 2020;75(1):71-78 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.119.13818] [Medline: 31760884]

XSL•FO
RenderX

28. Olczak KJ, Taylor-Bateman V, Nicholls HL, Traylor M, Cabrera CP, Munroe PB. Hypertension genetics past, present and future applications. J Intern Med 2021;290(6):1130-1152 [FREE Full text] [doi: 10.1111/joim.13352] [Medline: 34166551]
29. Flynn J. The changing face of pediatric hypertension in the era of the childhood obesity epidemic. Pediatr Nephrol 2013;28(7):1059-1066. [doi: 10.1007/s00467-012-2344-0] [Medline: 23138756]
30. Brady TM. Obesity-related hypertension in children. Front Pediatr 2017;5:197 [FREE Full text] [doi: 10.3389/fped.2017.00197] [Medline: 28993801]
31. Wühl E. Hypertension in childhood obesity. Acta Paediatr 2019;108(1):37-43. [doi: 10.1111/apa.14551] [Medline: 30144170]
32. Ouchi N, Parker JL, Lugus JJ, Walsh K. Adipokines in inflammation and metabolic disease. Nat Rev Immunol 2011;11(2):85-97 [FREE Full text] [doi: 10.1038/nri2921] [Medline: 21252989]
33. Ribeiro J, Guerra S, Pinto A, Oliveira J, Duarte J, Mota J. Overweight and obesity in children and adolescents: relationship with blood pressure, and physical activity. Ann Hum Biol 2003;30(2):203-113. [doi: 10.1080/0301446021000054587] [Medline: 12637195]
34. Angelopoulos PD, Milionis HJ, Moschonis G, Manios Y. Relations between obesity and hypertension: preliminary data from a cross-sectional study in primary schoolchildren: the children study. Eur J Clin Nutr 2006;60(10):1226-1234. [doi: 10.1038/sj.ejcn.1602442] [Medline: 16708067]
35. Tu W, Eckert GJ, DiMeglio LA, Yu Z, Jung J, Pratt JH. Intensified effect of adiposity on blood pressure in overweight and obese children. Hypertension 2011;58(5):818-824 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.111.175695] [Medline: 21968752]
36. Cheung EL, Bell CS, Samuel JP, Poffenbarger T, Redwine KM, Samuels JA. Race and obesity in adolescent hypertension. Pediatrics 2017;139(5):e20161433 [FREE Full text] [doi: 10.1542/peds.2016-1433] [Medline: 28557717]
37. Giussani M, Antolini L, Brambilla P, Pagani M, Zuccotti G, Valsecchi MG, et al. Cardiovascular risk assessment in children: role of physical activity, family history and parental smoking on BMI and blood pressure. J Hypertens 2013;31(5):983-992. [doi: 10.1097/HJH.0b013e32835f17c7] [Medline: 23425707]
38. Simonetti GD, Mohaupt MG, Bianchetti MG. Monogenic forms of hypertension. Eur J Pediatr 2012;171(10):1433-1439 [FREE Full text] [doi: 10.1007/s00431-011-1440-7] [Medline: 21404100]
39. National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. Pediatrics 2004;114(2):555-576. [Medline: 15286277]
40. Hart JT. Blood pressure in children: is screening for essential hypertension in children essential? BMJ 2008;336(7659):1452 [FREE Full text] [doi: 10.1136/bmj.39619.490475.3A] [Medline: 18583652]
41. Bijlsma MW, Blufpand HN, Kaspers GJL, Bökenkamp A. Why pediatricians fail to diagnose hypertension: a multicenter survey. J Pediatr 2014;164(1):173-177.e7. [doi: 10.1016/j.jpeds.2013.08.066] [Medline: 24120124]
42. Genovesi S, Parati G, Giussani M, Bona G, Fava C, Maffeis C, et al. How to apply European and American guidelines on high blood pressure in children and adolescents. a position paper endorsed by the italian society of hypertension and the italian society of pediatrics. High Blood Press Cardiovasc Prev 2020;27(3):183-193. [doi: 10.1007/s40292-020-00369-y] [Medline: 32170711]

Abbreviations

AH: arterial hypertension
AUC: area under the receiver operating characteristic curve
C-index: concordance index
DCA: decision curve analysis
EHR: electronic health record
FBG: fasting blood glucose
IBIL: indirect bilirubin
ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification Lasso: least absolute shrinkage and selection operator
LDL-C: low-density lipoprotein cholesterol
LYC: lymphocyte count
LYP: lymphocyte percentage
MCV: mean corpuscular volume
minMSE: minimum mean squared error
MONOC: monocyte count
NEUTP: neutrophil percentage
PDW: platelet distribution
PH: primary hypertension
OR: odds ratio
SH: secondary hypertension

XSL•FO
RenderX

TC: total cholesterol
UA: uric acid

Edited by T de Azevedo Cardoso, S Ma; submitted 22.03.24; peer-reviewed by J Mistry, K Wang; comments to author 06.05.24; revised version received 25.06.24; accepted 25.10.24; published 30.12.24.

Please cite as:
Qin C, Peng L, Liu Y, Zhang X, Miao S, Wei Z, Feng W, Zhang H, Wan C, Yu Y, Lu S, Huang R, Zhang X
Development and Validation of a Nomogram-Based Model to Predict Primary Hypertension Within the Next Year in Children and Adolescents: Retrospective Cohort Study
J Med Internet Res 2024;26:e58686
URL: https://www.jmir.org/2024/1/e58686
doi:10.2196/58686
PMID:

©Chenlong Qin, Li Peng, Yun Liu, Xiaoliang Zhang, Shumei Miao, Zhiyuan Wei, Wei Feng, Hongjian Zhang, Cheng Wan, Yun Yu, Shan Lu, Ruochen Huang, Xin Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leveraging Machine Learning to Identify Subgroups of Misclassified Patients in the Emergency Department: Multicenter Proof-of-Concept Study

Sage Wyatt[1], MSc; Dagfinn Lunde Markussen[2,3], MD; Mounir Haizoune[4], MSc; Anders Strand Vestbø[5], PhD; Yeneabeba Tilahun Sima[1], MD, MPH, PhD; Maria Ilene Sandboe[6]; Marcus Landschulze[7], PhD; Hauke Bartsch[8], PhD; Christopher Martin Sauer[9,10,11], MD, MPH, PhD

[1]Department of Global Public Health, Faculty of Medicine, University of Bergen, Bergen, Norway
[2]Department of Emergency Medicine, Haukeland University Hospital, Bergen, Norway
[3]Department of Clinical Science, Faculty of Medicine, University of Bergen, Bergen, Norway
[4]Helse Vest IKT, Bergen, Norway
[5]Department of Research and Development, Haukeland University Hospital, Bergen, Norway
[6]Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[7]Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences (HVL), Bergen, Norway
[8]Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen, Norway
[9]Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany
[10]MIT Critical Data, Massachusetts Institute of Technology, Boston, MA, United States
[11]Department of Hematology and Stem Cell Transplantation, University Hospital Essen, Essen, Germany

Corresponding Author:
Christopher Martin Sauer, MD, MPH, PhD
Institute for Artificial Intelligence in Medicine
University Hospital Essen
Girardetstraße 2
Essen, 45131
Germany
Phone: 49 201 723 0
Email: sauerc@mit.edu

Abstract

Background: Hospitals use triage systems to prioritize the needs of patients within available resources. Misclassification of a patient can lead to either adverse outcomes in a patient who did not receive appropriate care in the case of undertriage or a waste of hospital resources in the case of overtriage. Recent advances in machine learning algorithms allow for the quantification of variables important to under- and overtriage.

Objective: This study aimed to identify clinical features most strongly associated with triage misclassification using a machine learning classification model to capture nonlinear relationships.

Methods: Multicenter retrospective cohort data from 2 big regional hospitals in Norway were extracted. The South African Triage System is used at Bergen University Hospital, and the Rapid Emergency Triage and Treatment System is used at Trondheim University Hospital. Variables included triage score, age, sex, arrival time, subject area affiliation, reason for emergency department contact, discharge location, level of care, and time of death were retrieved. Random forest classification models were used to identify features with the strongest association with overtriage and undertriage in clinical practice in Bergen and Trondheim. We reported variable importance as SHAP (SHapley Additive exPlanations)-values.

Results: We collected data on 205,488 patient records from Bergen University Hospital and 304,997 patient records from Trondheim University Hospital. Overall, overtriage was very uncommon at both hospitals (all <0.1%), with undertriage differing between both locations, with 0.8% at Bergen and 0.2% at Trondheim University Hospital. Demographics were similar for both hospitals. However, the percentage given a high-priority triage score (red or orange) was higher in Bergen (24%) compared with

XSL•FO
RenderX

9% in Trondheim. The clinical referral department was found to be the variable with the strongest association with undertriage (mean SHAP +0.62 and +0.37 for Bergen and Trondheim, respectively).

Conclusions: We identified subgroups of patients consistently undertriaged using 2 common triage systems. While the importance of clinical patient characteristics to triage misclassification varies by triage system and location, we found consistent evidence between the two locations that the clinical referral department is the most important variable associated with triage misclassification. Replication of this approach at other centers could help to further improve triage scoring systems and improve patient care worldwide.

(J Med Internet Res 2024;26:e56382) doi:10.2196/56382

KEYWORDS

emergency department; triage; machine learning; real world evidence; random forest; classification; subgroup; misclassification; patient; multi-center; proof-of-concept; hospital; clinical feature; Norway; retrospective; cohort study; electronic health system; electronic health record

Introduction

A triage system is a standardized system for rapid patient decision-making used in emergency departments (EDs) worldwide. Most systems classify patients into different emergency levels based on symptoms and clinical signs [1]. Triage systems are frequently established on the basis of expert opinion and may not consistently undergo validation [2]. Preventing triage misclassification is the main objective of triage. Overprioritizing patients with mild conditions is not the best distribution of limited hospital resources, whereas underprioritizing severe cases can detrimentally impact patient outcomes.

Validation of a triage system can be challenging without a gold-standard assessment of patient urgency. Previous studies have used patient mortality and admission to intensive care units (ICUs) as measures of high urgency and discharge from ED as low urgency [1,3]. All triage systems aim to classify patients by urgency of the condition and availability of resources, but they may vary by the importance of clinical criteria like pain and definitions of time to care for patients classified as urgent. The sensitivity and specificity of different triage systems vary greatly. Most perform moderately well in identifying high-urgency patients (sensitivity between 58% and 100%) but perform significantly worse in identifying low-urgency patients (sensitivity between 8% and 70%). The literature on the validity of triage systems has been limited so far, and no triage system has been identified as clearly outperforming others [1,4]. Triage systems also have variable performance across contexts, reasons for presenting to the ED, patient age, and patient race [5-7].

In Norway, no triage system is used consistently nationwide, with different regions either using the Manchester Triage System (MTS) [8], the Rapid Emergency Triage and Treatment System (RETTS) [9], or a modified version of the South African Triage Scale (SATS) [10]. MTS is among the best-studied systems worldwide, and there is much more limited evaluation of RETTS and SATS, especially considering what factors influence their performance [1,3,4].

So far, most studies evaluating triage systems have relied on domain knowledge to evaluate the importance of factors contributing to triage misclassification [1,4]. While research driven by qualitative insights and posterior probabilities is important, it may also be valuable to evaluate the importance of variables without prior assumptions. In this study, we validated the performance of the modified SATS and RETTS used at Bergen and Trondheim University Hospital and aimed to identify patient clusters that are misclassified using these systems. To capture nonlinear relationships, we used various machine learning (ML) methods and used SHAP-values to establish feature importance.

Methods

Data Source

We conducted a multicenter retrospective cohort study at the main ED at Haukeland University Hospital, Bergen, and St. Olav's University Hospital in Trondheim, Norway. Haukeland University Hospital functions as a referral center for about a million inhabitants in the Bergen metropolitan area [10] and had an annual ED admission ranging from 33,000 to 38,000 during the study period. The ED manages patients with medical, surgical, and neurologic conditions, excluding children with medical issues and pregnant women with obstetric conditions who are treated elsewhere. For the period between 2013 and 2017, comprehensive data on all patients treated in the ED was gathered. This data encompasses administrative details like time of admission, department, and source of admission, as well as clinical information such as age, sex, and triage level. The data collection was conducted using the electronic health system used in the ED (Akuttdatabasen, Helse-Vest IKT, version 1.5.5., Stavanger).

St. Olav's, a tertiary medical center, serves as the primary health care facility for a local population of 300,000 residents and functions as the regional hospital for the Trondheim metropolitan area in central Norway, catering to over 700,000 residents. The hospital manages around 22,000 emergency department admissions annually [11]. Data from all patient contacts in the emergency department between 2012 and 2022 were extracted. The extraction was based on the ED's patient database (Akuttdatabasen, Helse Vest IKT, version 1.5.5, Stavanger, Norway). Following this, the extracted data was linked to patient administrative hospital data, with no exclusions made for patient contacts. Variables with incomplete registration or linkage between databases were omitted from individual variable results but retained in the overall dataset. Information

XSL•FO
RenderX

encompassing age, sex, arrival time, subject area affiliation, reason for contact, discharge location, level of care, and time of death was accessed in the database. This study received approval from the data protection officer (ESA-no 16/9114).

Missing Data

Patient information was collected prospectively at both hospitals, and we received access to the data in 2022. The Bergen dataset included all patients who presented at the ED from January 2012 through September 2017, while the Trondheim dataset encompassed patients presenting to the ED from 2012 to 2022. Data in Bergen could not be accessed after 2017 due to changes in the data storage protocol. For both datasets, we excluded patients with missing national identity numbers and no triage scores. For missing categorical data, we included a category "missing" in the model.

Triage Definitions

The SATS protocol (Bergen) categorizes patients into triage levels: 1 is red (emergency), 2 is orange (very urgent), 3 is yellow (urgent), 4 is green (not urgent), or 5 is blue (can wait). The physician should assess the patient immediately if triaged to level red, within 10 minutes to level orange, within 60 minutes to level yellow, and within 120 minutes to level green [1]. The RETTS protocol (Trondheim) also categorizes patients into triage levels 1 to 5 but has different recommended times until care for each level: immediately for level red, 20 minutes for level orange, 60 minutes for level yellow, and 240 minutes for level green [12]. Blue triage encompasses patients arriving for administrative reasons or scheduled visits in both systems.

Outcome Definition

The composite endpoint used to define high acuity was defined as (1) death within 24 hours after presentation to the emergency room, (2) transfer to the ICU from the ED, or (3) transfer to the surgical operating theater or for coronary angiography directly from the ED. The composite of these outcomes is hereafter referred to as "severe illness." Discharge from ED (ie, patients not admitted to the hospital) was the reference standard for low acuity. Undertriage was defined as a patient who died or was admitted to the ICU within 24 hours of presenting to the ED and was given a triage score of level 3, 4, or 5. Overtriage was defined as a patient who was discharged from the ED and was given a triage score of 1 or 2.

Classification Model

Separate classification models were built to identify variables associated with either under- or overtriage. The 2 subgroups consisting of patients with high or low acuity conditions were used as the training and test dataset, where the target variable was defined as undertriage (yes or no) or overtriage (yes or no), respectively. The models with the higher area under the curve score were used, a Random Forest for the Trondheim dataset and XGBM classifier for the Bergen dataset. The best-performing model was evaluated using a 5-fold cross-validation approach and GridSearch for parameter optimization. To handle the unbalanced data problem, SMOTE (synthetic minority oversampling technique) oversampling was applied to the data, and adjusted class weight parameters of the algorithm (ie, class weight for LR/RF/DT, scale_pos_weight for XGB) were applied to the models. The resulting models were then scored on the test data (30% of the data), and a receiver operating characteristic (ROC) curve score was computed. The feature importance and contributions were analyzed by calculating the SHAP (SHapley Additive exPlanations)-values [13].

Ethical Considerations

Approval for this study and a waiver of written informed consent was obtained from the Regional Committee for Medical and Health Research Ethics in Western Norway for Bergen University Hospital (case number 2018/2128). At Trondheim University Hospital, this study was classified as a quality assurance study, and a waiver from the Regional Committee for Medical and Health Research Ethics was granted (2016/1813/REK).

Results

Overview

Records for 205,488 patients from Bergen University Hospital and 304,997 records from Trondheim University Hospital were included in the final analysis (Table 1). There were few demographic differences between Bergen and Trondheim University Hospital, though a greater percentage of patients in Bergen (24.1%) were given low-priority triage scores than patients in Trondheim (9%, Table 1). Undertriage occurred in 1579 patients in Bergen and 736 patients in Trondheim, while overtriage was observed in 7 patients in Bergen and 22 patients in Trondheim. Missing information was overall low, with a maximum of 2.7% for age in the Bergen dataset.

XSL•FO
RenderX

Table 1. Demographic and clinical characteristics of the patients included in the final cohort.

	Bergen University Hospital (N=205,488), n (%)	Trondheim University Hospital (N=304,997), n (%)
Age groups		
<18 years old	14,307 (7)	7362 (2.5)
18-65 years old	104,634 (50.9)	155,233 (52.3)
>65 years old	86,321 (42)	134,268 (45.2)
Missing	226 (0.1)	8134 (2.7)
Sex		
Female	97,535 (47.5)	149,111 (48.9)
Male	107,945 (52.5)	148,353 (48.6)
Missing	8 (<0.1)	7533 (2.5)
Day of week		
Weekday	145,952 (7)	217,909 (71.4)
Weekend	59,535 (29)	87,088 (28.6)
Missing	1 (<0.1)	0 (0)
Time of day		
Early morning (4:00 to 7:59)	10,781 (5.2)	14,272 (4.7)
Morning (8:00 to 11:59)	40,242 (19.6)	57,988 (19)
Noon (12:00 to 15:59)	61,815 (30.1)	97,621 (32)
Evening (16:00 to 19:59)	44,766 (21.8)	61,456 (20.1)
Night (20:00 to 23:59)	32,190 (15.7)	46,972 (15.4)
Late night (00:00 to 3:59)	15,693 (7.6)	26,688 (8.8)
Missing	1 (<0.1)	0 (0)
Clinical referral department		
Internal Medicine	97,058 (47.2)	169,619 (55.6)
Surgical	55,491 (27)	78,825 (25.8)
Others	52,939 (25.8)	56,553 (18.5)
Missing	0 (0)	0 (0)
Triage score		
Low (1-2, Blue or Green)	49,583 (24.1)	27,300 (9)
High (3-5, Yellow, Orange, or Red)	155,904 (75.9)	276,933 (90.8)
Missing	1 (<0.1)	764 (0.3)
Status at 24 hours		
Alive	204,416 (99.5)	303,483 (99.5)
Deceased	1072 (0.5)	1514 (0.5)
Missing	0 (0)	0 (0)

Model Performance

All 4 considered statistic models performed similarly for a specific dataset, and little differences between SMOTE and class weight parameter adjustment were found (Table 2). The random forest model was chosen for Bergen University Hospital, and the XGBoost classifier for Trondheim University Hospital. Model performance for undertriage was overall lower for Trondheim University Hospital due to a high number of patients receiving high triage scores but having nonsevere outcomes (Figure 1). ROC curves for the chosen models are provided in Figure 2.

XSL•FO
RenderX

Table 2. Undertriage classifier ROC[a] scores. XGBoost was chosen as the final model for Bergen and Random Forest for Trondheim University Hospital.

	Bergen University Hospital		Trondheim University Hospital	
	SMOTE[b]	Class weight parameter	SMOTE	Class weight parameter
Logistic regression (LR)	0.78	0.78	0.60	0.61
Random Forest (RF)	0.78	0.78	0.59	0.61
XGB Classifier (XGB)	0.75	0.79	0.53	0.60
DecisionTree Classifier	0.77	0.77	0.53	0.60

[a]ROC: receiver operating characteristic.

[b]SMOTE: synthetic minority oversampling technique.

Figure 1. Undertriage classifier confusion matrix for the Bergen University Hospital XGB classifier (left) and Trondheim University Hospital RF classifier (right). XGB: XGBoost; RF: random forest.

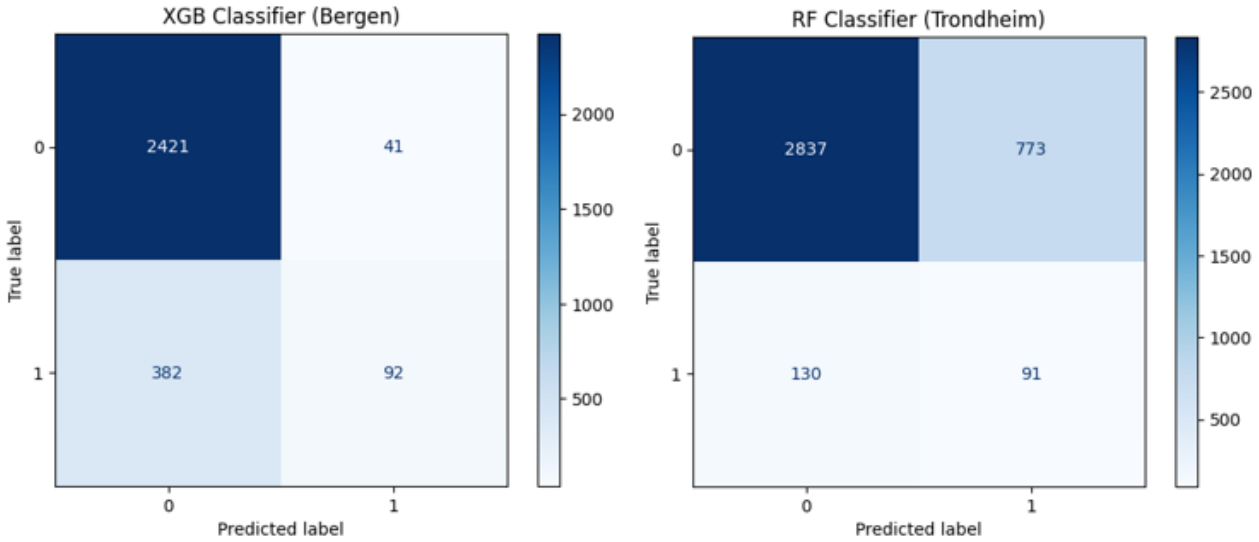

Figure 2. Undertriage classifier area under the receiver operator curve for XGB with Bergen University Data (left) and RF for Trondheim University data (right). XGB: XGBoost; RF: random forest.

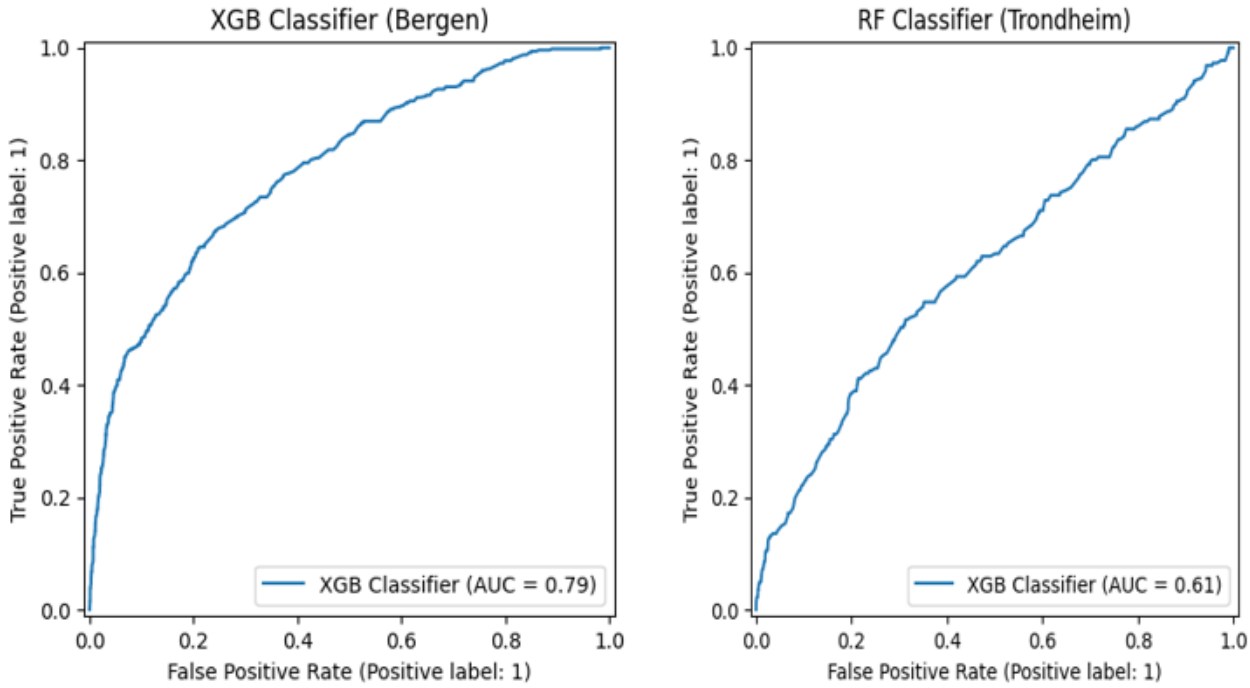

XSL•FO
RenderX

Undertriage

In both the Bergen and Trondheim datasets, the most influential features associated with undertriage were the clinical referral department, time of day of admission, and patient age (Figure 3). The clinical referral department was the only statistically significant variable in the model for the Bergen dataset, while in Trondheim, patient age and time of day were also associated with undertriage (Figures 5 and 6). A bee plot showing all variables included in the models is provided in the supplementary figure (Multimedia Appendices 1 and 2).

Figure 3. SHapley Additive exPlanations (SHAP)-values for undertriage in the Bergen (A) and Trondheim (B) classification model.

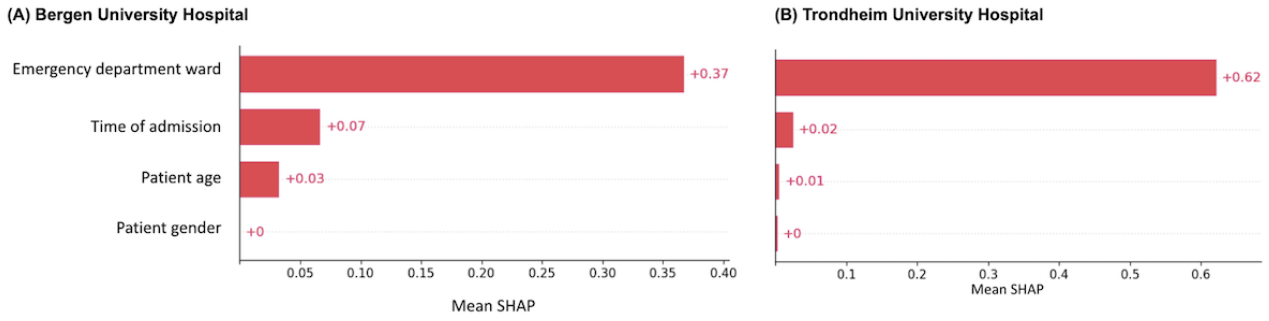

In the Bergen dataset, orthopedics and plastic surgery clinical assignment categories (Figure 4A) show a shift towards higher SHAP-values, indicating a higher probability of undertriage, while the trauma category shows a shift toward lower SHAP-values, indicating a higher probability of correct triage. There was no specific category of patient age or time of day that was important to the model.

In the Trondheim dataset clinical referral departments such as orthopedics (Figure 4B) and late-night admission time (Figure 6B) have higher SHAP-values indicating a higher probability of undertriage, while patients aged 0 to 17 (Figure 7B) and admission times in the morning showed a trend towards lower SHAP-values indicating a higher probability of correct triage.

Figure 4. SHapley Additive exPlanations (SHAP)-values for undertriage by clinical assignment category for the Bergen (A) and Trondheim (B) data set. NA: not available.

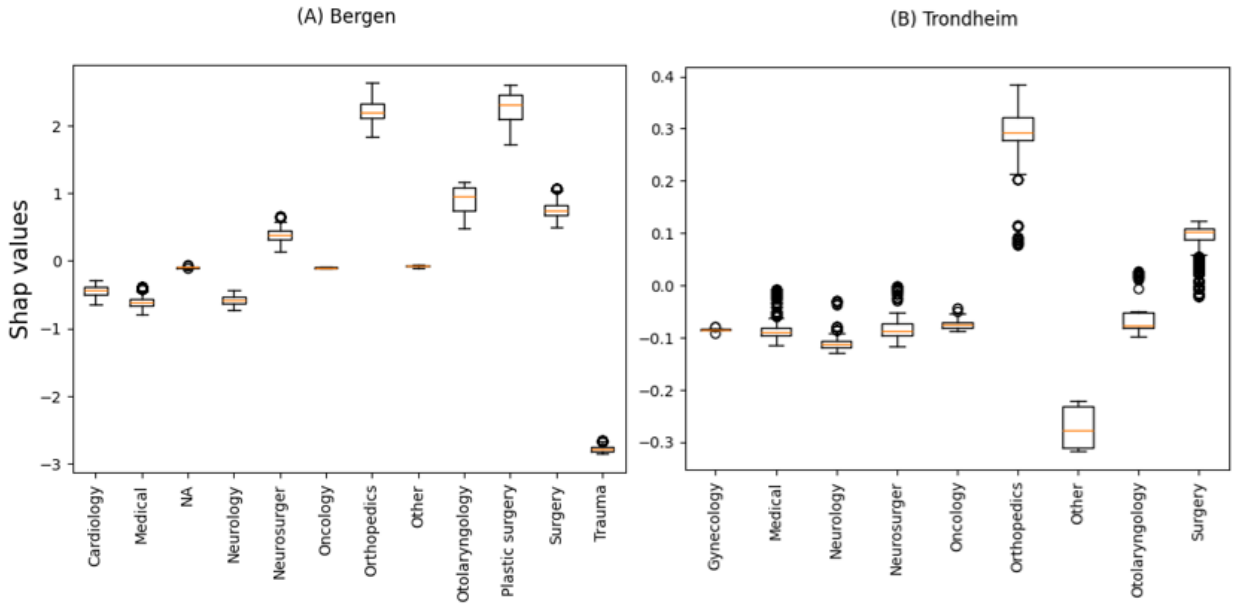

XSL•FO
RenderX

Figure 5. SHapley Additive exPlanations (SHAP)-values for undertriage by age group in the Bergen (A) and Trondheim (B) data set.

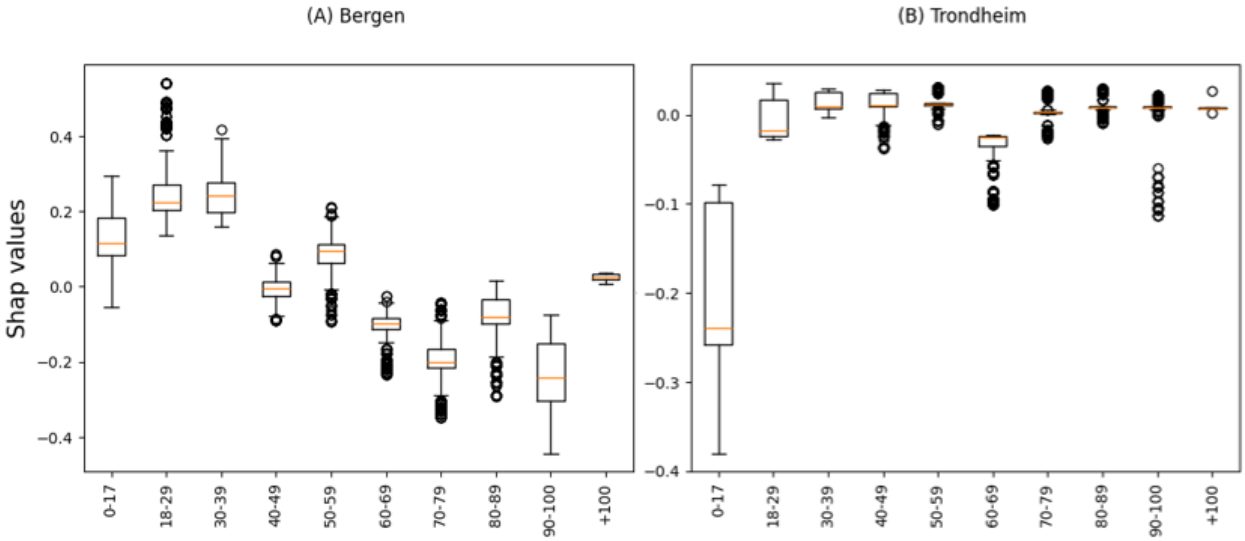

Figure 6. SHapley Additive exPlanations (SHAP)-values for undertriage by time of admission in the Bergen (A) and Trondheim (B) data set.

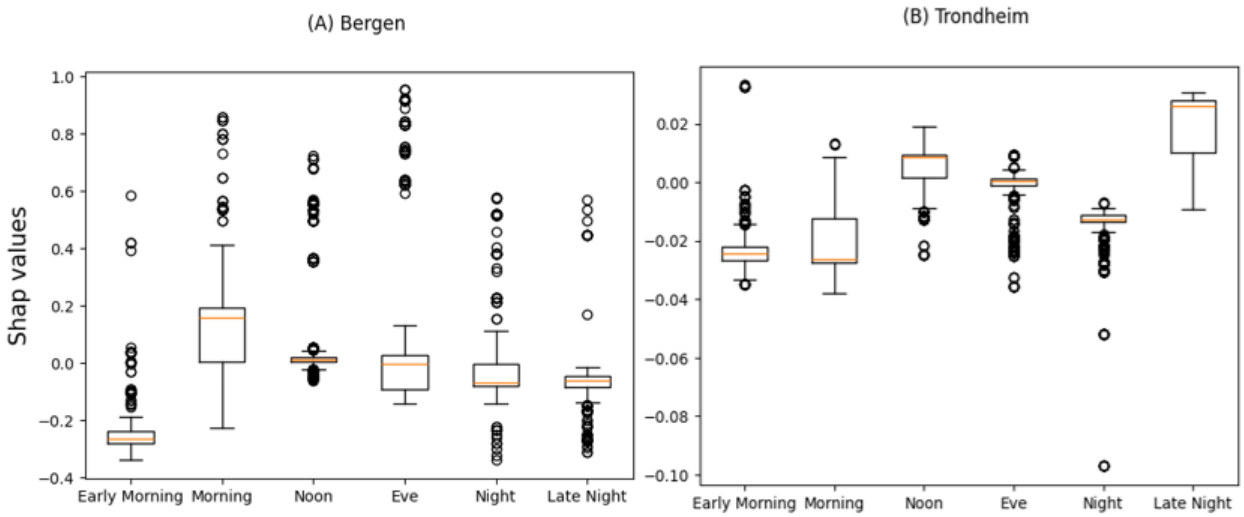

Figure 7. SHapley Additive exPlanations (SHAP)-values for undertriage by sex in the Bergen (A) and Trondheim (B) data set.

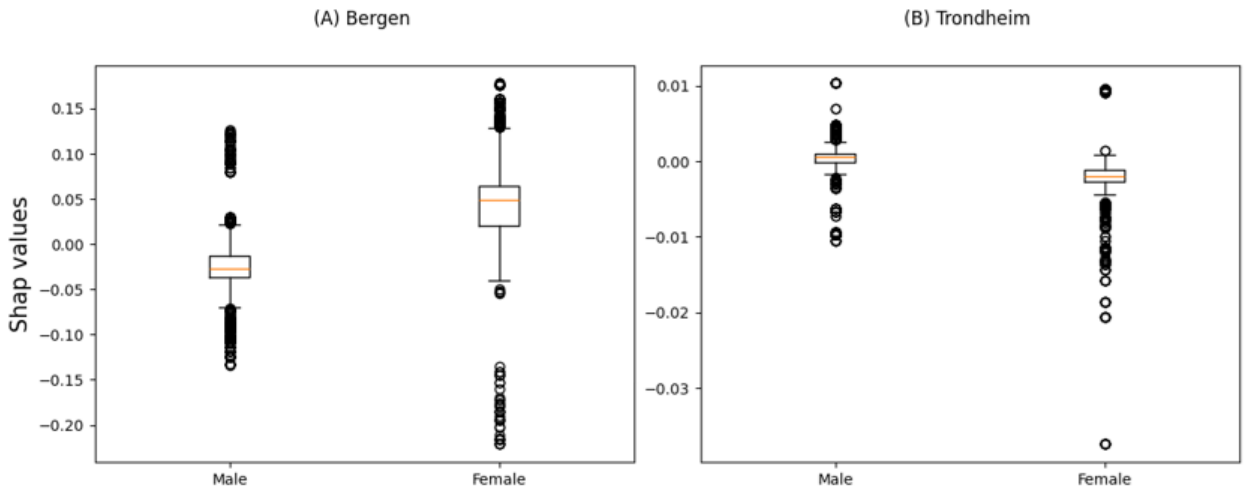

XSL•FO
RenderX

Overtriage

Overtriage was a rare event, only occurring in a total of 29 patients. Results for these exploratory analyses are available from Multimedia Appendix 3.

Discussion

Principal Findings

Our study provides a proof of concept of how ML can be leveraged to identify over- and undertriaged patients in the emergency room. As both hospitals function as regional medical centers for surrounding rural areas, they are at high risk of receiving overtriaged transfer patients from smaller regional hospitals without the confidence for definitive treatment [14]. However, the proportion of overtriage was very low (<0.1%), while we found a significant proportion of undertriage (0.2%-0.8%). Improvement of undertriage may, therefore, benefit patients and prevent adverse consequences. However, the increased frequency of undertriage within the orthopedic and surgical emergency groups could also be attributed to the specific definition of high urgency we used. This definition, encompassing factors like direct transfer to the surgical operating theater, aligns with cases requiring prompt surgical intervention but not necessarily immediate life-saving measures, for example, in case of ankle fractures triggering prompt transfers to surgical theaters. In Trondheim, the association between lower age and undertriage is consistent with previous literature on undertriage in children. Most triage systems are refined to adult standards, and triage systems for children can be inaccurate when predicting morbidity and mortality [15].

For overtriaged patients, there was some evidence of the increased frequency of overtriage by the time of day and sex. Both factors can be indicators of health care worker mood and attitudes, and be reflective of biases within the hospital systems [16,17]. However, due to the low number of overtriaged patients (<0.1%), results carry high statistical uncertainty, and the results presented in Multimedia Appendix 3 should be considered explorative only.

New Applications of ML Methodology

ML is a topic of growing interest in the field of health care, and the number of articles published about ML, specifically in diagnostics, has steadily increased annually since 2000 [18]. ML offers advantages over many conventional statistical methods because they can model nonlinear associations. Previous research considering performance measures only suggests that random forest performs better than the more conventional method of logistic regression in most datasets [19]. Statistical interactions and model multiplicity can also be better captured by ML methods than conventional methods. For optimal usage, appropriate methods must be tailored to the specific research context, and common pitfalls need to be avoided [20].

In this setting, ML methods allowed us to consider the importance of many complex factors that impact triage classification simultaneously, providing more nuanced results than conventional methods. A previous study using conventional methods in Bergen found that overtriage was most prevalent in patients younger than 18 years [21]. Yet the automated variable selection based on SHAP-values in this study revealed that age might not be the most important characteristic contributing to overtriage in Bergen, as might have been an assumption in a study directed by domain knowledge. Instead, with ML methods, we identified that the clinical referral department and *ICD* (*International Classification of Diseases*) diagnostic codes are more important factors associated with overtriage in the Bergen dataset. Our results are in agreement with conventional methods in undertriage, which both showed the triage system as having lower sensitivity to surgical patients than other patients [21].

Differences Between Triage Systems

The differences seen in overtriage between the Trondheim and Bergen dataset might be due to the usage of different triage systems, as demonstrated in a recent Swedish study [22]. The difference in health system and context between the two Norwegian hospitals is likely not substantial. In the context of health systems, there is a need for more external validation of ML methods against conventional methods in diverse contexts and data [23,24].

Even between 2 hospitals with similar populations and health systems, we could not generalize characteristics important to triage misclassification between the two triage systems. All triage systems are subject to misclassification, but information about the strengths of each system is necessary to create new systems with greater validity. More research is needed in the future about triage systems and new applications of ML methods, such as automated triage classification systems [25]. Identifying which patient groups are at risk of misclassification is a crucial step to reduce health care resource waste and enhance patient safety. Furthermore, this may help to address and reduce health biases.

Strengths and Limitations

The strengths of this study include large, high-quality data from two different emergency department datasets. Our findings are novel in that we focused our analysis on a comparison in misclassification between two triage systems rather than in the isolated context of one hospital. We are the first to use ML methods to investigate characteristics contributing to misclassification in Norway, yet our approach is largely generalizable beyond Norway and may trigger additional studies investigating the undertriage of patient subgroups.

One limitation of this study pertains to the temporal context of our data acquisition. In 2017, Haukeland Hospital underwent a significant transformation in its computer system used for triage data recording. This shift led to the availability of only aggregated data, devoid of the granular characteristics of individual patient visits that could be effectively linked to other data sources. This transition restricted our ability to work with individual-level data and prompted us to work with somewhat older data for this specific hospital. Despite this challenge, we maintain that the data used remains valid and representative of the patient population that frequents the emergency department. While the more recent transition to aggregated data may have limited our analysis scope, the fundamental characteristics of

XSL•FO
RenderX

the patient population, their triage experiences, and the overarching triage dynamics are captured faithfully in our study.

A critical factor in the successful application of ML to ED triage systems is the availability of granular, detailed data. The precision and effectiveness of ML algorithms are directly influenced by the depth and breadth of the data they process. In the context of ED triage, this means having access to comprehensive patient data, including specific reasons for ED visits, presenting symptoms, and vital clinical parameters like blood pressure, heart rate, and respiratory rate. Such detailed data enables the development of more accurate and nuanced ML models capable of making informed triage decisions that closely align with the complexities of real-world clinical scenarios. In addition, granular data is essential in identifying and addressing potential biases inherent in ML models, ensuring that these systems are equitable and effective for diverse patient populations [26].

It is worth acknowledging that our dataset had certain limitations in terms of clinical parameters. While we had access to fundamental data like sex, age, and clinical referral department, our dataset lacked detailed information on the specific reason for ED presentation, presenting symptoms, and clinical parameters like blood pressure, heart rate, or respiratory rate. These clinical details play a pivotal role in the accurate assessment and classification of patient urgency. The absence of these granular clinical parameters may have influenced the precision of our analysis and the generalizability of our findings to a more comprehensive clinical context. Future research would benefit from access to a more comprehensive dataset encompassing these key clinical parameters. As for the risk of inclusion bias, we believe that due to Norway's highly inclusive and public health care system, this is overall low. However, the underrepresentation of underserved minorities cannot be finally excluded.

The study's insights underscore the importance of updated and accessible data for robust research in health care. The shift to aggregated data in one of the hospitals due to computer system changes in 2017 serves as a reminder of the dynamic nature of data availability. The ability to work with up-to-date and comprehensive datasets, coupled with the capacity to integrate data from diverse sources, is pivotal in enabling a holistic understanding of complex health care scenarios like emergency department triage.

Conclusions

In conclusion, we provide a machine-learning framework to identify subgroups of patients that are undertriaged using two common triage scoring systems. Combining clinical knowledge, routinely collected clinical data, and ML can improve patient care and the efficacy of health care delivery. As we move forward, the emphasis on data access, integration, and real-time updating becomes ever more paramount in advancing clinical care.

Acknowledgments

We gratefully acknowledge the invaluable help of Sarah King for the provision of the Trondheim dataset. We thank the organizers of the Eitri Datathon 2022, in particular, Torleif Markussen Lunde and Dr Leo Celi, for bringing us together and for their valuable feedback in preparing this manuscript.

SW and YTS are supported by the European Research Council (grant 833076). CM Sauer is supported by the German Research Foundation (DFG) funded UMEA Clinician Scientist Program (grant FU356/12-2). We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Data Availability

The code underlying these analyses is available from [27]. The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request after signing a data transfer agreement with the receiving institution to ensure the protection of personal information.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Beeplot showing the SHAP (SHapley Additive exPlanations)-values for all included variables at Bergen (A) und Trondheim University Hospital (B).
[PNG File , 286 KB - jmir_v26i1e56382_app1.png]

Multimedia Appendix 2
Beeplot showing the SHAP (SHapley Additive exPlanations)-values for all included variables at Bergen (A) und Trondheim University Hospital (B).
[PNG File , 289 KB - jmir_v26i1e56382_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3
Results for the exploratory findings on overtriage for the Bergen and Trondheim data sets.
[DOCX File , 214 KB - jmir_v26i1e56382_app3.docx]

References

1. Hinson JS, Martinez DA, Cabral S, George K, Whalen M, Hansoti B, et al. Triage performance in emergency medicine: a systematic review. Ann Emerg Med 2019 Jul;74(1):140-152. [doi: 10.1016/j.annemergmed.2018.09.022] [Medline: 30470513]
2. Moll HA. Challenges in the validation of triage systems at emergency departments. J Clin Epidemiol 2010 Apr;63(4):384-388. [doi: 10.1016/j.jclinepi.2009.07.009] [Medline: 19875271]
3. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019 May 28;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
4. Kuriyama A, Urushidani S, Nakayama T. Five-level emergency triage systems: variation in assessment of validity. Emerg Med J 2017 Nov;34(11):703-710. [doi: 10.1136/emermed-2016-206295] [Medline: 28751363]
5. Ginsburg A, Oliveira JSL, Mullan A, Mhayamaguru K, Bellolio F. 92 emergency severity index and older adults: should age be incorporated into the triage algorithm? Ann Emerg Med 2020;76(4):S36. [doi: 10.1016/j.annemergmed.2020.09.102]
6. Zook HG, Kharbanda AB, Flood A, Harmon B, Puumala SE, Payne NR. Racial differences in pediatric emergency department triage scores. J Emerg Med 2016 May;50(5):720-727 [FREE Full text] [doi: 10.1016/j.jemermed.2015.02.056] [Medline: 26899520]
7. Gräff I, Goldschmidt B, Glien P, Bogdanow M, Fimmers R, Hoeft A, et al. The German Version of the Manchester Triage System and its quality criteria--first assessment of validity and reliability. PLoS One 2014;9(2):e88995 [FREE Full text] [doi: 10.1371/journal.pone.0088995] [Medline: 24586477]
8. Mackway‐Jones K, Marsden J, Windle J. The triage method. In: Emergency Triage: Manchester Triage Group. Hoboken, New Jersey: John Wiley & Sons; 2013:11-24.
9. Widgren BR, Jourak M. Medical Emergency Triage and Treatment System (METTS): a new protocol in primary triage and secondary priority decision in emergency medicine. J Emerg Med 2011 Jun;40(6):623-628. [doi: 10.1016/j.jemermed.2008.04.003] [Medline: 18930373]
10. Brevik H, Eide M, Engan M, Aalvik R. Standardised emergency medicine assessment and prioritisation (triage) tool. Helse Vest.: SATS-N; 2017. URL: https://helse-bergen.no/seksjon/mottaksklinikken/Documents/SATS-N%20users%20manual%20version%203.02.pdf [accessed 2023-03-03]
11. Bjørnsen LP, Uleberg O, Dale J. Patient visits to the emergency department at a Norwegian university hospital: variations in patient gender and age, timing of visits, and patient acuity. Emerg Med J 2013 Jun;30(6):462-466. [doi: 10.1136/emermed-2012-201191] [Medline: 22736718]
12. Tran T. Triage I pasientmottak for barnn studie av validiteten avtriagesystemet ”RETTS pediatrisk” ved.: St Olavs Hospital URL: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2419684/Tran.PDF?sequence=1 [accessed 2024-12-04]
13. A unified approach to interpreting model predictions. URL: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [accessed 2023-12-21]
14. Parikh PP, Parikh P, Mamer L, McCarthy MC, Sakran JV. Association of system-level factors with secondary overtriage in trauma patients. JAMA Surg 2019 Jan 01;154(1):19. [doi: 10.1001/jamasurg.2018.3209]
15. Mora MC, Veras L, Burke RV, Cassidy LD, Christopherson N, Cunningham A, et al. Pediatric trauma triage: A Pediatric Trauma Society Research Committee systematic review. J Trauma Acute Care Surg 2020 Apr 15;89(4):623-630. [doi: 10.1097/ta.0000000000002713]
16. Larsson E, Lindström A, Eriksson M, Oldner A. Impact of gender on post- traumatic intensive care and outcomes. Scand J Trauma Resusc Emerg Med 2019 Dec 23;27(1):115 [FREE Full text] [doi: 10.1186/s13049-019-0693-4] [Medline: 31870391]
17. Zajic P, Bauer P, Rhodes A, Moreno R, Fellinger T, Metnitz B, et al. Time of day and its association with risk of death and chance of discharge in critically ill patients: a retrospective study. Sci Rep 2019 Aug 29;9(1):12533 [FREE Full text] [doi: 10.1038/s41598-019-48947-y] [Medline: 31467390]
18. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina (Kaunas) 2020 Sep 08;56(9):455 [FREE Full text] [doi: 10.3390/medicina56090455] [Medline: 32911665]
19. Couronné R, Probst P, Boulesteix A. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 2018 Jul 17;19(1):270. [doi: 10.1186/s12859-018-2264-5]
20. Sauer CM, Chen L, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. Lancet Digit Health 2022 Dec;4(12):e893-e898 [FREE Full text] [doi: 10.1016/S2589-7500(22)00154-6] [Medline: 36154811]

XSL•FO
RenderX

21. Markussen DL, Brevik HS, Bjørneklett RO, Engan M. Validation of a modified South African triage scale in a high-resource setting: a retrospective cohort study. Scand J Trauma Resusc Emerg Med 2023 Mar 20;31(1):13 [FREE Full text] [doi: 10.1186/s13049-023-01076-y] [Medline: 36941710]
22. Habbouche S, Carlson T, Johansson D, Kjaerbeck S, Malm M, Svensson P, et al. Comparison of the novel WEst coast System for Triage (WEST) with Rapid Emergency Triage and Treatment System (RETTS©): an observational pilot study. Int J Emerg Med 2022 Sep 12;15(1):47 [FREE Full text] [doi: 10.1186/s12245-022-00452-2] [Medline: 36096726]
23. Upshaw TL, Craig-Neil A, Macklin J, Gray CS, Chan TCY, Gibson J, et al. Priorities for artificial intelligence applications in primary care: a canadian deliberative dialogue with patients, providers, and health system leaders. J Am Board Fam Med 2023 Apr 03;36(2):210-220 [FREE Full text] [doi: 10.3122/jabfm.2022.220171R1] [Medline: 36948537]
24. Rose S. Intersections of machine learning and epidemiological methods for health services research. Int J Epidemiol 2021 Jan 23;49(6):1763-1770 [FREE Full text] [doi: 10.1093/ije/dyaa035] [Medline: 32236476]
25. Sánchez-Salmerón R, Gómez-Urquiza JL, Albendín-García L, Correa-Rodríguez M, Martos-Cabrera MB, Velando-Soriano A, et al. Machine learning methods applied to triage in emergency services: A systematic review. Int Emerg Nurs 2022 Jan;60:101109. [doi: 10.1016/j.ienj.2021.101109] [Medline: 34952482]
26. Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW, et al. A translational perspective towards clinical AI fairness. npj Digit. Med 2023 Sep 14;6(1):172. [doi: 10.1038/s41746-023-00918-4]
27. GitHub. mounirHai/triage. URL: https://github.com/mounirHai/triage [accessed 2024-12-09]

Abbreviations

ED: emergency department
ICD: International Classification of Diseases
ICU: intensive care units
ML: machine learning
MTS: Manchester Triage System
RETTS: Rapid Emergency Triage and Treatment System
ROC: receiver operating characteristic
SATS: South African Triage Scale
SHAP: SHapley Additive exPlanations
SMOTE: synthetic minority oversampling technique

Edited by A Mavragani; submitted 15.01.24; peer-reviewed by L Agha-Mir-Salim, M Taghavi; comments to author 15.03.24; revised version received 04.05.24; accepted 16.09.24; published 31.12.24.

Please cite as:
Wyatt S, Lunde Markussen D, Haizoune M, Vestbø AS, Sima YT, Sandboe MI, Landschulze M, Bartsch H, Sauer CM
Leveraging Machine Learning to Identify Subgroups of Misclassified Patients in the Emergency Department: Multicenter Proof-of-Concept Study
J Med Internet Res 2024;26:e56382
URL: https://www.jmir.org/2024/1/e56382
doi:10.2196/56382
PMID:39451101

©Sage Wyatt, Dagfinn Lunde Markussen, Mounir Haizoune, Anders Strand Vestbø, Yeneabeba Tilahun Sima, Maria Ilene Sandboe, Marcus Landschulze, Hauke Bartsch, Christopher Martin Sauer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multiple Perspectives on the Adoption of SMART Technologies for Improving Care of Older People: Mixed Methods Study

Steriani Elavsky[1], PhD; Lenka Knapova[1], MA; Kamil Janiš[1], PhD; Richard Cimler[2], ING, PhD; Jitka Kuhnova[3], PhD; Tomas Cernicky[4], PhD

[1]Department of Human Movement Studies, University of Ostrava, Ostrava, Czech Republic
[2]Anume Ltd, Hradec Kralové, Czech Republic
[3]Faculty of Science, University of Hradec Karlove, Hradec Kralove, Czech Republic
[4]SeneCura SeniorCentrum HŠH Inc, Prague, Czech Republic

Corresponding Author:
Steriani Elavsky, PhD
Department of Human Movement Studies
University of Ostrava
Varenska 40a
Ostrava, 70200
Czech Republic
Phone: 420 553 46 2588
Email: steriani.elavsky@osu.cz

Abstract

Background: Despite the ever-increasing offering of SMART technologies (ie, computer-controlled devices acting intelligently and capable of monitoring, analyzing or reporting), a wide gap exists between the development of new technological innovations and their adoption in everyday care for older adults.

Objective: This study aims to explore the barriers and concerns related to the adoption of SMART technologies among different groups of stakeholders.

Methods: Data from 4 sources were used: semistructured in-person or internet-based interviews with professional caregivers (n=12), structured email interviews with experts in the area of aging (n=9), a web-based survey of older adults (>55 years) attending the Virtual University of the Third Age (n=369), and a case study on the adoption of new technology by an older adult care facility.

Results: Although all stakeholders noted the potential of SMART technologies to improve older adult care, multiple barriers to their adoption were identified. Caregivers perceived older adults as disinterested or incompetent in using technology, reported preferring known strategies over new technologies, and noted own fears of using technology. Experts viewed technologies as essential but expressed concerns about cost, low digital competency of older adults, and lack of support or willingness to implement technologies in older adult care. Older adults reported few concerns overall, but among the mentioned concerns were lack of ability or interest, misuse of data, and limited usefulness (in specific subgroups or situations). In addition, older adults' ratings of the usefulness of different technologies correlated with their self-rating of digital competency (r=0.258; P<.001).

Conclusions: Older adults appeared to have more positive views of various technologies than professional caregivers; however, their concerns varied by the type of technology. Lack of competence and lack of support were among the common themes, suggesting that educationally oriented programs for both older adults and their caregivers should be pursued.

(J Med Internet Res 2024;26:e45492) doi:10.2196/45492

KEYWORDS

adults; older; technologies; technological; caregivers; SMART; mobile phone

XSL•FO
RenderX

Introduction

Background

In the Czech Republic, similar to the rest of the world, adults aged ≥65 years belong to the fastest-growing segment of the population, accounting for 19.2% of the Czech population. People aged >80 years belong to the fastest-growing subgroup, and their share among those aged >60 years is expected to increase from 11% to 19% by 2050 [1]. However, with increasing age, the likelihood of comorbidities, functional limitations, and disabilities increases, with negative consequences for self-sufficiency and mobility. The probability of at least partial dependence on the help and care of others or of transition to care facilities increases with age [2], which has implications for the growing demands on the care system in the Czech Republic.

The Czech Republic (together with Poland and Greece) belongs to the European Union countries with the greatest preference for family care in old age [2]. However, although most older adults wish to spend the last years of their lives at home, only a minority fulfill this wish. Although 80% of Czech older adults remain in homecare, where they are primarily cared for by their relatives [3], and 60% wish to live their lives at home, only 21% of deaths in the Czech Republic occur at home [4]. However, preferences for family care decrease with increasing health problems and functional limitations or with the need for specialized care (eg, for chronic diseases or dementia) [2]. Unfortunately, the Czech Republic spends proportionately less on health care and provides only limited support in long-term care compared with other European Union countries [3]. Likewise, the field of private care is not yet fully developed in the Czech Republic, and people are not used to paying extra for care. Apart from family care at home, the state care system for the older adults dominates in the form of either field care services (ie, assistive services or homecare) or state-supported homes for the older adults [5,6].

Modern (SMART) technologies are one of the tools that can be helpful in supporting dignified aging both at home and in the institutional environment. SMART technologies refer to computer-controlled devices that appear to not only act intelligently [7] but also describe technologies capable of monitoring, analyzing, and reporting the status of an object. In the health sciences literature, this topic falls under the eHealth domain. The term “eHealth” is specifically used to describe approaches and devices that promote health or healthy behaviors that use the internet. The narrower field of mobile health refers to the use of mobile handheld devices to support medical and public health practices [8]. eHealth or mobile health technologies also include integrated or connected sensors that provide ways to monitor and improve health (or factors in the environment that affect health), healthy lifestyles, or overall well-being. Technologies specifically developed for older adults or aimed for use by aging populations are referred to as gerontechnology [9].

The application of gerontechnology can help compensate in selected areas of care and support selected needs of older adults [10]. At the same time, informal caregivers can benefit from the use of technology when providing care at home, and technology can help facilitate the care provided in residential or other specialized facilities for older adults [11]. The COVID-19 pandemic further exposed areas in which greater involvement of technology could facilitate care or support the quality of life of older adults [12]. Despite this vast potential, however, a wide gap remains between the innovations and development of new technologies and their adoption by target populations and settings [13-15]. Among the facilitating factors for technology use are motivation, support and feedback received from others, and usability or accessibility, whereas the most commonly cited barriers to the adoption of new technologies are sociodemographic factors (age, education, and skills); personal factors (such as lack of time or other priorities); lack of support; lack of resources or unclear guidance on how to use the technology [16,17]. Different technology acceptance models have also been proposed to help explain the use of technology at the individual level such as the Technology Acceptance Model (see the review by Marangunić and Granić [18]) or the Unified Theory of Acceptance and Use of Technology (see the review by Khechine et al [19]). Although these theoretical models are useful in describing individual differences in technology use or explaining the adoption of a concrete technology, they are not well suited for studying technology adoption in a broader context and describing attitudes toward novel technological advancements. This is particularly true when users lack sufficient familiarity with the technology under investigation, making it challenging for them to express their attitudes or assess predictor variables due to their limited experience or knowledge.

Objectives

In this study, we focused on evaluating the potential of SMART technologies in the care of older adults. Specifically, we focused on multiple perspectives on the adoption of SMART technologies in care for older adults. We concentrated on inputs from different groups of stakeholders. First, we interviewed professional caregivers from a variety of settings providing care to older adults. Second, we conducted a cross-sectional web-based survey with adults attending the Virtual University of the Third Age (V-U3A). Third, we interviewed experts on aging on their views. Finally, we included a case study focusing on the collaboration between a university start-up company and a private care provider with respect to the adoption of a specific technology, the ANUME smart bed system. The overarching research questions were as follows:

1. What are the barriers to the adoption of SMART technologies in the care of older adults?
2. How can these barriers be overcome?

Methods

Overall Conceptual Framework

To effectively represent the spectrum of applications in which the use of technology intersects older adult populations, in this study, we applied a triangulation method, combining quantitative (cross-sectional) and qualitative data from several sources [20,21]. The selection of target groups and type of data used reflected several overlapping (and perhaps underrepresented)

XSL•FO
RenderX

realities. First, there is an interplay among technology use, perceived competence, and attitudes or motivation in using technology, especially among older adults or those who perceive themselves as generally less technologically savvy [22]. Despite the general contemporary increase in the use of information and communications technologies across all age groups, in the context of the Czech Republic, the use of technologies by older adults remains low [6,7]. This makes it challenging to effectively evaluate the attitudes toward technologies among this target population, as they lack the necessary frame of self-reference (ie, experience with said technologies) to do so. Relatedly, the probability of at least a partial dependence on the help of caregivers or of a transition to care facilities increases with age [2], making it likely that most older adults will be cared for through external resources (ie, family members and professionals). As such, caregivers thus represent the target market for care-related technology or operate as advocates recommending the technology to other users, including older adults themselves or those caring for them. Simultaneously, they bear the heavy burden associated with caring for an aging person (often being individuals with limited function and autonomy) and are therefore positioned to benefit the most from using technology as part of the care they provide. In light of these complex intersections between technology use and aging, and in an effort to more effectively examine those with an advanced aptitude toward technology in this framework of relations, we chose to focus on two primary target groups: (1) people who care for older adults in a variety of settings and (2) older adults who are also attendees of the V-U3A (ie, representing more technologically "savvy" older adults who are the most likely users of SMART technologies). To supplement the views of these care providers and older adults themselves, we also used insights from an invited group of experts in the field of aging, with the belief that their perspectives would help contextualize the experiences of care workers and older adults in a broader context. Finally, to illuminate the dialogic processes that comprise the research-application continuum, we have included the details surrounding a concrete example of the adoption of technology in older adult care. Concretely, we have provided a case study of implementation that further highlights the "real-world" application process by recounting the process of cooperation between researchers, a commercial entity (a university-based start-up company), and an application partner (the SeneCura company) that provided continuous feedback on the partial and evolving functionalities of the specific technology (ANUME smart bed), communicating both the specific expectations of future users (eg, in terms of function, operation, and safety) and the subsequent experience of technology adoption, underscoring the ways in which intention and application are always in process.

Study 1: Perspectives of Professional Caregivers

Sample and Recruitment

We conducted semistructured interviews with 12 professional care providers: 5 (42%) workers from older adult homes, 3 (25%) from nursing services, 3 (25%) from social services, and 1 (8%) from a medical facility. Recruitment was performed through selected cooperating facilities and through chain referral using purposive sampling strategy (Multimedia Appendix 1).

Procedure

Conducted by the study coauthor KJ, a Doctor of Philosophy researcher (male) with extensive experience in qualitative methods and the area of aging, the interviews were performed from September 2020 to March 2021, that is, during the period of the COVID-19 pandemic, which complicated recruitment efforts to some extent and affected the form of the interviews (some interviews took place face to face, some were internet-based via Zoom [Zoom Video Communications, Inc] or Skype [Skype Technologies]). All respondents provided informed consent and gave permission for the interviews to be audio or video recorded for the purpose of obtaining verbatim transcripts of the interviews to be analyzed.

Measures

The interviews followed a semistructured "interview guide" (Multimedia Appendix 2).

Study 2: Perspectives of Older Adults

Sample and Recruitment

To examine the perspective of older adults, we conducted a web-based cross-sectional survey with adults aged ≥50 years. The recruitment strategy focused on students at the V-U3A, which is organized by the Faculty of Business Economics of the Czech University of Life Sciences in Prague. The V-U3A allows interested parties from all over the Czech Republic to pursue higher education at V-U3A as part of an adult continuing education (attendees must be Czech citizens of retirement age). Education within the V-U3A takes place through a combination of internet-based and collective "in-person" methods of education. The listeners meet once every 14 days in a so-called consultation center (mostly a city or municipal office, library, school, or older adults' club in different cities or towns), where they have access to a PC with an internet connection, among other things. As a target group, V-U3A listeners therefore included older adults with at least basic digital literacy and an assumed ability to handle participation in a web-based questionnaire survey. The focus on V-U3A listeners was also partly related to the complicated epidemiological situation when there was no face-to-face teaching, and most of the events aimed at the target group of older adults did not take place in person.

Procedure

Respondents were contacted via a mass email, which was distributed upon request by the relevant coordinators in individual regions of the Czech Republic (4 coordination centers were randomly selected from each region and were invited to participate). The recruitment email described the study and contained a link to the web-based questionnaire. Informed consent was presented as the first page of the survey, and subsequent completion of the survey indicated consent with participation. Data were collected between January 2021 and April 2022.

Measures

The survey was conducted on the web using the Qualtrics platform (**Qualtrics** International Inc). It included basic sociodemographic questions and assessed the level of digital literacy with a single item "How would you rate your skills in

XSL•FO
RenderX

using IT technologies (eg, computer, smartphone, tablet)?" rated on a scale from 1 ("beginner") to 7 ("expert"). The respondents then rated 21 different technologies or SMART devices marketed commercially to older adults and their caregivers. The evaluated smart technologies included technologies that can be used to satisfy various needs of older adults or caregivers (related to everyday practical tasks, social and emotional support, health monitoring or managing, and compensatory assistance rehabilitation) [23] and spanned the entire spectrum of technologies from simple ones (such as smart thermometer) to complex smart technologies (internet-based assistants and smart home systems). The individual technologies and the purpose of their use were first briefly described (together with a photo), and the respondents then evaluated the extent to which they found them useful (on a scale of 1 ["minimally useful"] to 10 ["maximum useful"] with the option to choose "don't know"). For each technology, the respondents had the opportunity to describe their concerns in an open answer using the given technology. The questionnaire also included questions for caregivers who care for older adults aged >65 years. The survey was conducted anonymously.

Study 3: Perspectives of Experts in the Area of Aging

Sample and Recruitment

Email interviews were conducted with experts in the area of aging including researchers, clinicians, and aging business or nonprofit representatives. The goal was to incorporate perspectives of various stakeholders, including but not limited to academic researchers. Participants were selected based on convenience and prominence in the aging sector in the Czech Republic across the fields.

Procedure

A total of 28 aging experts were contacted via email by the research project principal investigator in 2 waves during the summer and fall of 2021: 11 during the first wave in July 2021, followed by 18 in the second wave in September 2021. A reminder email was sent 5 to 8 weeks after the first invitation. The email invitation included a file containing the survey questions to be returned to the researchers via email. Of the 28 contacted experts, 15 (54%) were primarily academic researchers, 7 (25%) were representatives of older adult or aging organizations, and 6 (21%) were clinicians. Participants provided textual answers to 5 open-ended questions. The answers were of varying length, detail, and comprehensiveness, as no guidelines on the detail and extent of the answers were provided by the researchers.

Measures

With respect to time constraints of the respondents, the email interview was constructed to be brief with 5 open-ended questions on SMART or modern technology for the older adults. The questions covered the following areas: (1) opportunities of SMART or modern technologies for older adults, (2) barriers to the adoption of SMART or modern technologies for older adults, (3) good practice examples, (4) technologies that should be commonplace in older adult care, and (5) recommendations to stakeholders researching and implementing technologies for older adults (Multimedia Appendix 2).

Case Study: Perspective of Care Providers

Sample and Recruitment

For the case study, SeneCura SeniorCentrum Hradec Králové, located in Hradec Králové, Czech Republic, entered the project as an application partner on the grant project application along with the University of Hradec Králové start-up company ANUME Ltd. The SeneCura SeniorCentrum is a modern private residential facility for social services offering approximately 150 beds. The facility provides the following 2 services in terms of the Social Services Act: home for older adults and home with a special regime. Most of the clients of the home with a special regime have dementia, most commonly Alzheimer disease. Furthermore, most clients are polymorbid and require 24-hour care. SeneCura currently operates 15 similar facilities in the Czech Republic, with a total capacity of approximately 2100 beds. The ANUME smart bed product (developed by DeepLab Ltd., and later renamed to ANUME, Ltd., a second application partner in the project) consists of pads measuring the microvibrations of the human body, which are placed under the mattresses in the clients' bed. These pads can be used to determine the presence of the person on the bed and the elapsed time since the last positioning or repositioning, and thanks to special algorithms, the pulse and breathing frequency of the person lying on the bed can also be obtained from the microvibrations. Nursing staff have tablets on which this information is displayed, and problematic conditions are notified through visual and audio alerts.

Procedure

Cooperation with the application partner included regular meetings every 14 days, with group discussions based on the staff feedback and fine-tuning to meet the specific needs of care providers. Two of these meetings were in the format of a half-day "workshop" (July 2 to September 23, 2020), where there was an in-depth debate on the specific functionalities of the ANUME smart bed system. Initially, 10 ANUME smart bed systems were placed in the SeneCura care facility, where regular consultations and tuning of the product took place across 12 months. Later, 5 additional ANUME smart bed systems were placed in 5 homes where family members cared for older adults.

Measures

The process of implementation included interviews with nursing staff, clients, and management to ensure that the product effectively targeted everyone's needs. The main outputs from the interactions were summarized and agreed upon by all stakeholders before proceeding with system adjustments.

Data Analysis

Data from study 1 (in-person interviews) and open-ended responses from the web-based survey in study 2 were analyzed using content analysis and group interpretation. Each interview was read by 2 independent people (SE, the researcher, read all interviews; LK and KJ read 6 interviews each). During the reading, each researcher took notes in their Excel (Microsoft Corp) file, which was structured to align with the "interview guide." This was followed by a group discussion and interpretation of the most important themes. For analysis of the open-ended responses in study 2, the data were coded by 3

XSL•FO
RenderX

independent coders based on a predetermined coding scheme. The coding scheme was developed based on a preliminary content analysis of 50% of the data by the lead study author (SE) and refined in the process of subsequent coding. Interrater reliability was evaluated on randomly selected data that were coded by all 3 coders (about 10%) and ranged from 0.42 to 0.99 for the checked categories. Inconsistencies in coding were then discussed among the 3 coders and the first study author and resolved them after clarification and upon mutual agreement. This process was repeated 3 times to increase the precision and consistency of coding throughout the coding process. Frequency analysis on the coded data was performed. In addition, quantitative analysis of web-based survey data in study 2 included basic descriptive statistics and Pearson correlation coefficient calculated using SPSS (IBM Corp). Missing data were handled through listwise deletion. For study 3, due to the varying length of the answers in the email interviews and the inability of the researchers to elicit further information on the topic, no advanced qualitative data analysis (such as thematic analysis) was used. Individual answers to each of the 5 questions were merged and subsequently synthesized by the researchers. For the case study, we summarized the near 2-year experience of a 3-way collaboration between a start-up technology company, private care provider, and academic researchers.

Ethics Approval

The project and study procedures were approved by the Ethics Committee at the University of Ostrava (application number: OU-78256/90-2020).

Results

Perspective of Professional Caregivers

Sample Description

The average age of the interviewed professional caregivers was 44 (SD 9.2; range 27-56) years. Most of the interviewees (8/12, 67%) had a university education, 25% (3/12) had a high school diploma, and 8% (1/12) had vocational training without a high school diploma. The respondents had been in their current job for an average of 12 (SD 9; range 0.5-30) years, and 7 respondents had previous relevant work experience with older adults before working in their current positions. The average length of the interview was 41 (SD 10.5; range 27-63) minutes.

Findings

On the basis of the caregivers' reports, they used a limited number of technologies in their work. The primary method of documentation was paper and pencil. Nursing home managers and social workers mentioned the use of a computer with a specialized software application for documenting care, and some care workers mentioned the use of readers to record care activities (even in these cases, however, they duplicated the records by using the "paper-pencil" method if the reader failed). Other "tools" mentioned included the use of a weighing chair, monitoring bracelets with alarms at the entrance (all used in one specific older adult home), and an interactive table. Respondents also mentioned that they used mobile phones for work and that their activization workers used tablets when working with clients. In these cases, the clients were passive users of tablets or phones, as the activization worker would search for music or photos and then share them with the clients. Tablets were also used to communicate with family (mentioned by 1 worker in an older adult home).

When discussing the specific technologies, although several are already commonly available, their penetration into common practice in care for older adults was seen as problematic. This was partly related to the technologies being perceived as costly and out of reach for a Czech older adult or in the context of "typical" care provided. However, the main problem that was identified was the social service care workers' perceived rigidity to change:

> *It's just that everyone here is like that, I don't know, like that, they're just slow to open to these technologies and just see the problem behind everything, yeah [...]; [...]that if somewhere I don't know something new electronic, the workers in social services are terribly irritated by it, that it will be extra work again* [...]

Although the caregivers saw the advantages of a particular technology that would facilitate some of the tasks that they were performing or help satisfy the needs of older adults, they were not willing to consider including it in their work routine. Existing procedures and approaches were assessed as sufficient and easier as they were used to them. This attitude was also partially supported by the caregivers' assessment of their clients' digital literacy or their own negative relationship with technology. Caregivers presumed the digital literacy of older adults to be low, or they expected that physical and other limitations would prevent older adults from using them:

> *You know what, it's just that I can't, like, when I drive around these families and the elderly, they don't have any modern technology at home, yes, you can see an old television, for example, and I can't imagine that he would understand this, control it like and that is not to demean them, please, yes, but I don't know how they would at all with such modern technology, well, I don't know.*

The interviews also showed that the caregivers would be reluctant to implement some smart technologies in their work even if they could, especially when it came to technologies that were perceived as "displacing" older adult needs. A typical example was the perception of a social robot as a replacement for personal contact and not as a supplementary technology aimed at increasing socialization and entertainment at moments when personal contact might be scarce. The following was one respondent's reaction to the usability of the interactive robot:

> *[...] Like, good, I would be afraid of that too, I wouldn't want it either, like. Although it looks, like, peaceful. But I, I would be afraid of that [...].*

Additional themes that emerged in the conversations included the lack of trust of old people in anything new, including technology and privacy concerns in general (eg, when it comes to constant surveillance in the case of motion detection systems); specifics of use (control, charging, etc) and the expected problems associated with them; the need for simple control or

XSL•FO
RenderX

operation not only when used by the older adults themselves but also by the caring staff; and differences in the perception of technology by the type of care facility (small and family-oriented facilities perceived to have a sufficient number of workers to provide quality care even without technology solutions) or depending on the condition or diagnosis of the client (eg, for clients with cognitive deficits or dementia, any "wearable" technology was perceived as useless and not helpful).

Despite the identification of many barriers or nuances in the use of technology, there were positive comments that underlined the potential of technology to improve the quality of life of older adults and reduce the workload of the care staff:

> *Exactly, and it will also make life easier for the seniors, like the feeling that I am able to do something for myself again, I go back to that spoon, to eat soup, that is so important for them, but so terribly important, they just here they will fully appreciate such a step with a smile, with enthusiasm, that it will be better again, that they will simply take a step forward again, and that is just great.*

Perspective of Older Adults

Sample Description

In total, 521 respondents clicked on the survey link. Of them, 145 (27.8%) respondents were excluded: 60 (41.4%) respondents did not answer a single question, 60 (41.4%) answered one question but did not complete the ratings of technologies, and 25 (17.2%) did not provide key sociodemographic information (age). Of the remaining 376 respondents, 7 (1.9%) stated that they were aged <50 years and were excluded. Of the remaining 369 in the final sample, 58 (15.7%) had some missing data: 2 (3%) <1% missing; 5 (9%) <20%; 7 (12%) <50%; 6 (10%) <75%; 38 (66%) ≥75%. The final sample of adults consisted of 369 respondents aged >55 (mean 71.1, SD 5.4; range 57-95) years, with a completion rate of 65%. Most respondents (n=313, 84.8%) were female, and most (n=355, 96.2%) were retired. Most respondents reported having secondary education with (n=186, 50.4%) or without a diploma (n=21, 5.9%), and 33.3% (n=123) had a university degree. The respondents (n=259, 70.2%) rated their economic situation as average. Their perceived digital competence varied and rated on average 3.67 (SD 1.01) on a scale from 1 (beginner) to 7 (expert). Moreover, 16% (n=59) of the respondents said that they provided care to someone aged >65 years. Therefore, these respondents were able to provide the perspective of a caregiver in addition to the perspective of older adults. Descriptive characteristics of the sample are depicted in Table 1.

XSL•FO
RenderX

Table 1. Description of survey respondents (n=369).

Variable	Values, n (%)
Age (years)	
55-59	2 (0.5)
60-64	24 (6.5)
65-69	138 (37.4)
70-74	111 (30.1)
75-79	65 (17.6)
80-84	24 (6.5)
≥85	5 (1.4)
Gender	
Female	313 (84.8)
Male	55 (14.9)
Missing or not reported	1 (0.3)
Marital status	
Single	6 (1.6)
Married	206 (55.8)
Partner relationship	12 (3.3)
Divorced	50 (13.6)
Widowed	95 (25.7)
Education	
Elementary (including unfinished)	7 (1.9)
Vocational training without diploma	21 (5.7)
High school education with diploma	186 (50.4)
Higher education	28 (7.6)
University education	123 (33.3)
Missing or not reported	4 (1.1)
Employment status	
Retired	355 (96.2)
Employed	15 (4.1)
Private business	6 (1.6)
In household	2 (0.5)
Full-time caregiver	3 (0.8)
Other	8 (2.2)
Self-rated economic situation	
Below average	6 (1.6)
Rather below average	32 (8.7)
Average	259 (70.2)
Rather above average	56 (15.2)
Above average	5 (1.4)
Missing or not reported	11 (3)
Self-rated digital skills	
1 (beginner)	10 (2.7)
2	33 (8.9)

XSL•FO
RenderX

Variable	Values, n (%)
3	113 (30.6)
4	124 (33.6)
5	63 (17.1)
6	13 (3.5)
7 (expert)	1 (0.3)
Missing or not reported	12 (3.3)

Findings

Among the 3 highest-rated technologies with respect to their usefulness for older adult care were wearable fall detection sensors (SOS buttons for calling for help; mean 8.9, SD 1.8), tablets (mean 8.8, SD 1.8), and smartphones (mean 8.7, SD 2.0). Among the lowest-rated technologies were interactive robots (mean 4.5, SD 2.9), followed by virtual reality (mean 6.2, SD 3.0) and smart alarm clock (mean 6.3, SD 2.8). The average rating across all technologies (mean 7.4, SD 1.5) was positively related to self-rated information and communications technology abilities (mean 3.7, SD 1.1; r=0.238; P<.001).

Overall, the older adults had no or few concerns about the technologies, with 69.8% of the respondents stating that they had no concerns when averaged across the technologies. The most frequently mentioned concerns across technologies were as follows: certain technologies seemed as unnecessary, redundant, or suitable only for certain subgroups of older adults (eg, those with health problems or dependent on the care of others), the fear of loss of privacy or misuse of the acquired data, the inability or unwillingness of the older population to learn how to handle new technologies or fully use all their functionalities, or worries about making mistakes when using the technology. However, many cited advantages of various technologies, commenting that they had personal experience with them and considered some to be essential (eg, smartphones). Detailed results are presented in Table 2 and Multimedia Appendix 2.

Table 2. Ratings of technologies by survey respondents (n=369).

	Participants, n (%)	Values, mean (SD)
Fall detector	310 (84)	8.88 (1.75)
Tablet	314 (85.1)	8.78 (1.77)
Smartphone	313 (84.8)	8.73 (2.05)
Smart bed	291 (78.9)	8.65 (1.98)
Smart contactless thermometer	301 (81.6)	8.59 (1.96)
Smart spoon	295 (79.9)	8.42 (2.21)
Interactive desk	297 (80.5)	8.07 (2.18)
Smart pill dispenser	290 (78.6)	7.95 (2.46)
Smart lightbulb	300 (81.3)	7.87 (2.45)
Witrack	276 (74.8)	7.83 (2.41)
Environmental sensors	283 (76.7)	7.77 (2.5)
Mean usefulness across technologies	334 (90.5)	7.43 (1.54)
Mobile apps	226 (61.2)	7.11 (2.79)
Fitness bracelet	305 (82.7)	7.10 (2.68)
Virtual assistants	244 (66.1)	7.04 (2.66)
Smart cup	284 (77)	6.69 (2.91)
Smart scale	292 (79.1)	6.65 (2.77)
RFID[a] chip	301 (81.6)	6.5 (2.62)
Barcode reader	285 (77.2)	6.36 (2.86)
Smart alarm clock	282 (76.4)	6.29 (2.78)
Virtual reality	260 (70.5)	6.2 (2.98)
Interactive robots	238 (64.5)	4.55 (2.91)

[a]RFID: radio frequency identification.

XSL•FO
RenderX

Perspective of Aging Experts

Sample Description

A total number of 12 experts responded to the invitation by October 2021. However, 3 of the responses did not include any answers to the survey: 1 participant reported not working in a relevant field, 1 participant was worried about sharing their "know how," and 1 participant was only interested in "serious collaboration." A total of 9 experts provided answers to the survey questions. Of them, 8 (89%) were female participants, and 1 (11%) was a male participant; 7 (78%) were academic researchers, 1 (11%) was a clinician, 1 (11%) was a representative of an older adult or aging organization. Overall, the response rate was approximately 50% for academic researchers and as low as 15% for clinicians and representatives of older adult organizations, which have proven to be very difficult-to-reach populations.

Findings

There was an agreement among the experts that SMART technologies can bring about large benefits to older adults, especially in terms of increasing or prolonging self-sufficiency and autonomy (eg, through the use of assistive technology and web-based cognitive rehabilitation) and increasing quality of life and decreasing loneliness (eg, through web-based communication tools). Respondents also mentioned that the use of SMART technologies could lead to improved and more efficient care for the older adults and decreased caregiver burden through remote monitoring, shared or distributed responsibilities among caregivers, or web-based rehabilitation options. In addition, several respondents highlighted the facilitation of communication and interaction via technology.

However, a series of barriers to the use of SMART technologies for older adult care was perceived. These could be grouped into (1) obstacles on the side of older adults, (2) objective barriers, and (3) lack of a facilitator. On the side of older adults, respondents mentioned low digital literacy, fear, and negative attitudes, as well as declined cognitive abilities (eg, learning and memory). Objective barriers that were voiced were high costs, nonavailability of SMART technologies to all older adults in all areas, and technologies that were too complex and not customized for their needs. Importantly, respondents mentioned the lack of a facilitator who would introduce the technology to older adults, provide support in case of technical issues, and serve as an intermediary to caregivers or clinicians. The following quotes illustrate this issue:

> *Furthermore. The absence of experts[...] who could convey information about possible technologies. Their benefits and limits. Social workers play a key role in the introduction of technology for the elderly and especially in their support.*
>
> *There is a lack of specifically "senior" help desks and people who would physically come and look at the computer problem. so that the senior is not dependent only on the children.*

When asked to provide examples of good practice of the use of SMART technologies, some experts mentioned specific technologies, whereas others mentioned systemic practices aimed at implementing technologies for older adult care. In particular, a (social) system that enables older adults to live autonomously at home for as long as possible as well as educates caregivers and social workers about technology was envisioned. Specific technologies perceived as "good practice" ranged from security or monitoring systems and SOS buttons through web-based communication tools to robots for various purposes such as socialization and help with personal hygiene.

Similarly, the level of complexity and technological advancement varied in the experts' views on technologies that should be commonplace in older adult care. The recommended technologies varied from personal hygiene robots, voice-operated smart beds, or technology-enhanced cognitive rehabilitation, and smart tools of daily living to SOS buttons, movement detectors, and monitors of health indicators. In addition, web-based communication tools including telemedicine were often mentioned.

Finally, the experts provided recommendations to stakeholders researching and implementing technologies for older adults. There was an agreement on the need to consider the whole spectrum of older adults with varied cognitive and physical abilities and contexts and to consider the specifics of this group with relation to technology (eg, insecurity and need to preserve dignity). Importantly, they stressed the need to develop the technologies directly for older adults according to their needs. Technologies should then be tested with all involved parties: older adults, experts, and caregivers. One expert summarized this as follows:

> *Show them what you can do and get advice on what you should be able to do...*

Perspective of Care Providers

Case Description

The main motivation of the application partner, SeneCura, was the use of SMART technology to reduce the care burden of their workers and improve the quality of care in their facility. The cooperation was established based on a specific product, the ANUME smart bed developed by ANUME Ltd.; however, the application partner also showed interest in other relevant technologies that could provide solutions for additional needs (eg, temperature-monitoring solutions and GPS location or detection for leaving the "safe" zone of the device).

The main target groups were social workers and nurses who would use the information provided by the ANUME smart bed. Targeting caregivers in direct care was not a priority and was also perceived as potentially problematic due to the frequent turnover of staff members and the perceived limited ability of the staff (who are often foreign workers) to competently handle technology. During the implementation process, SeneCura identified an additional target group, namely, physicians who take over the care of the older adults in case of transfer to hospital, emergency, or specialized care.

Findings

The collaboration between the researchers, start-up ANUME Ltd., and the SeneCura care facility led to continuous improvement of the ANUME smart bed product. Feedback

XSL•FO
RenderX

received from the staff and facility management led to both hardware and software adjustments as well as continuous improvement of the application that displayed the monitored data. For example, the original intention was to install only the individual sensors under the mattress and place the computing and communication unit together with the display on the bed frame, which was sufficient in terms of functionality, but the problem was the need to be able to "disinfect" the sensors (despite not being in direct contact with the client but under the mattress) and possible repositioning of the sensors. This need was also intensified during the COVID-19 pandemic. Feedback from the nursing staff also revealed the uselessness of a display placed directly on the bed (staff expressed that they were not using it and the light emitted, even in subdued mode, was bothering some clients). The display was later removed based on this feedback and for economic reasons. The most suitable way to place ANUME in beds proved to be embedding the sensors and all the necessary elements of the system, that is, the sensors themselves and the computing and communication module, in a thin pad. This pad is made of washable material that is intended for use in a hospital environment and is fully disinfectable. It is created from a material that allows the mattress to breathe despite being waterproof. In addition, the user interface underwent changes from a stand-alone bedside monitor to an internet-based web interface. Finally, the tablet application was continuously modified based on the staff and management feedback. As a result of these modifications, the management began to make extensive use of positioning reports (generated based on ANUME system data), which allowed for monitoring of client positioning or the lack thereof. At present, management of the SeneCura facility regularly uses positioning incident reports to evaluate the quality of care of individual shifts and at higher levels of management to compare individual workplaces. At the staff level, night shifts especially appreciated the constant overview of the presence of clients in bed and the staff's ability to respond in time when the system alerted them that the client had left. The system settings allow adjustment of the alarm schedule to, for example, not alert during a shorter stay out of bed such as a visit to the toilet, but alert in the case of a longer stay, which may be indicative of a risk situation such as a fall.

Discussion

Principal Findings

This study used mixed methods and multiple sources of data to examine the potential of SMART technologies to improve care for older adults, identify key barriers to their adoption, and demonstrate how barriers can be overcome to facilitate the use of technology in caregiving contexts. Perspectives from different groups of stakeholders (ie, professional caregivers, older adults, and aging experts) and a concrete example of the implementation of a specific type of technology have demonstrated that despite the potential SMART technologies offer for improving older adult care, multiple barriers persist to their broader adoption. The key barriers were perceived to be the inability or unwillingness of at least some older adults to use SMART technology, partly due to low perceived competence and fear of misuse of the data it provides (mentioned by all stakeholders). On the side of the caregivers, a lack of knowledge about technologies, low digital literacy, rigidity to change, and inaccessibility were identified as important barriers by both caregivers and aging experts. Older adults and aging experts agreed that a potentially useful way to overcoming these barriers could be the inclusion of a facilitator or tech mediation worker who would provide support not only during the initial adoption phase but also throughout the use of the technology (the main themes are represented in Figure 1). In addition, the presented case study indicated that meaningful improvements in technology development can be achieved and new opportunities for the involvement of new technologies can be identified through a collaborative process involving all stakeholders, thus enhancing the potential for technology adoption.

Figure 1. Conceptual representation of barriers to the adoption of SMART technologies from different perspectives. Note: caregivers, experts, and older adults in italics indicate which group of respondents mentioned this barrier (at least to some extent).

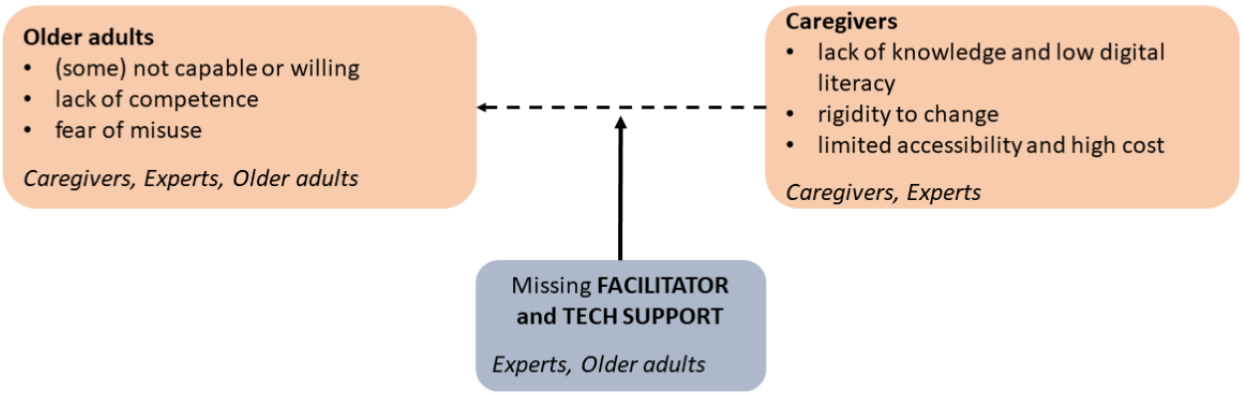

Note. Caregivers, Experts, Older adults in italics indicate which group of respondents mentioned this barrier (at least to some extent).

XSL•FO
RenderX

Comparison With Prior Work

Previous research has shown that age plays an important role in the use and adoption of new technologies [24-27]. Although older adults use new technologies less and perceive more barriers to their use, the use of new technologies is also related to the level of education [25-28]. Research also shows that older adults are able and willing to use new technologies if they find them useful and simple enough to use [18]. Thus, we should not approach older adults as a monolithic group and anticipate reluctance or fear of interacting with technology without offering opportunities for older adults to try new technologies and proper support in the process of their adoption (eg, through caregivers or relatives who can introduce technology to older adults and teach them to work with it) [29].

It was evident from the responses by older adults that they found SMART technologies highly useful (7.4 on a 0-10 scale), and they mentioned few concerns in relation to the different SMART technologies surveyed. Among the specifically mentioned concerns were fear of loss of privacy or data abuse, which can be considered as relatively universal concerns, and although they are prevalent, they do not typically lead to abandonment of technologies [30]. Views of the older adults also indicated that various SMART technologies might be perceived very differently not only based on their type and function but also possibly the amount of exposure that older adults have had to the given technology and their related understanding of its functionality. For some technologies, older adults did not see the point of using them personally and rather found them useful for specific populations "in need;" for example, SMART voice-operated lightbulbs were seen as useful predominantly for immobile individuals. Arguably, however, some of these technologies could help fulfill the needs of any person irrespective of age (eg, various indoor sensors and internet-based assistants to aid in activities of daily living). It should be noted that this assessment may reflect the inability of the generally healthy and high-functioning older adults in this study to accept the eventual probability of needing different forms of specialized care in later years (an analogous process has been documented with respect to perceptions of frailty) [31]. This further underscores the need to introduce and explain potential SMART technologies to older adults and provide specific relatable scenarios where they could be useful for them personally.

Older adults and experts were rather consistent in what should already be commonplace or what are the good examples of SMART technologies for older adults: fall detectors, SOS buttons, smart beds, some smart tools of daily living, etc. One technology that experts saw a great benefit in, but older adults did not find very useful and voiced concerns over, was social or interactive robots. This sentiment was also echoed by professional caregivers who saw these technologies as potentially displacing important needs of older adults (ie, need for personal contact). This was a surprising finding, given the literature that generally emphasizes the advantages of social robots and supports their social benefits [32].

Taken altogether, the views expressed by professional caregivers, older adults, and aging experts indicated that once objective barriers such as cost and accessibility are successfully confronted, a key prerequisite to higher adoption of SMART technologies in older adult care is the overall increase in digital literacy both at the caregiver and older adult level. Increasing knowledge of the benefits of digital technologies that can be linked to the specific needs of older adults and their caregivers should be of utmost importance. These changes will partly occur naturally, without significant efforts from the outside, because of the aging of the population for whom it is natural to use SMART technologies in everyday life. Nevertheless, this process could be potentially accelerated in several ways. For instance, "safe spaces" could be created for older adults and their caregivers to try out technologies, gain personal experience, and subsequently make an informed decision on purchase and use. Similarly, potential users of SMART technologies for older adult care would benefit from an overview of the technologies available in the market, ideally linked to the needs of older adults and caregivers that could be fulfilled by using these technologies. In addition, educational programs directed both at caregivers and older adults would be useful tools for improving digital literacy, attitudes toward technology, and eventual use of technology [33].

Our case study also shows that it is important to find and communicate "selling points" of SMART technologies to the facilities and caregivers and that the original use case and functionality may not prove to be the most useful for the customer. For example, when the ANUME smart bed system was first considered by SeneCura, the application partner, its most desired feature was considered to be the monitoring of vital signs while in bed. However, the most recent iteration of the technology does not even include real-time data on vital signs; instead, information on bed exits has been prioritized by the caring staff, and the positioning incident report has been adopted as the primary tool by the management staff. These gradually transforming and emerging needs would not have been captured without an ongoing shared dialog between the researchers, the technology developers, and the care facility staff and management. In addition, SeneCura communicated that this iterative reciprocal process helped overcome the initial skepticism of some of the staff toward the technology and made them use the technology more (ie, their needs were "heard").

Strengths and Limitations

The main strengths of this study lie in incorporating multiple perspectives on the use and adoption of SMART technologies for improving care of older adults and presenting a concrete example of technology implementation in a caregiving context, which is rarely done in one comprehensive study. Nonetheless, this study has several limitations. First, it was cross-sectional and thus limited in terms of causality. Second, the study included perspectives of only a small sample of professional caregivers and a selective group of older adults (attendees of the V-U3A). These more technologically savvy older adults could have a more positive view of technology than the general older adult population. Similarly, the comparatively more "negative" evaluation of older adults' abilities by professional caregivers could reflect the specific clientele they serve (ie, older adults in older adult homes or in need of care). Third, despite the successful example of collaboration with an industry partner, SeneCura, it must be noted that the management of SeneCura

XSL•FO
RenderX

was very receptive to innovations to begin with, which may have influenced the findings of this study. According to the Diffusion of Innovations theory [34,35], innovators and early adopters are more likely to embrace new technologies compared with the majority of the population. In the context of a research study, if a care provider falls into the category of innovators or early adopters and exhibits a high level of enthusiasm and openness toward technology, their adoption behavior may not accurately represent the attitudes and behaviors of the broader population. To avoid this potential pitfall in future research studies, several strategies such as diverse participant recruitment (ie, a broad range of care providers with varied levels of technology receptiveness); separate data analyses for different segments of care providers (ie, based on their technology adoption profiles); or contextual exploration (organizational culture, resource availability, training opportunities, etc) can be used.

Future Directions

Regarding the driving force of adoption of SMART technologies, with significant barriers to adoption at multiple levels, it is unrealistic to expect that caregivers or older adults would be the primary or sole actors in this process. Even if SMART technology product owners, developers, and designers strive to develop products that are usable for older adults with adjusted interfaces and operation where necessary (a need that was underlined by the experts in this study), this by itself will likely not be sufficient to overcome the abovementioned barriers and convince the potentially reluctant end users. Synergies must be created among efforts on the side of the researchers, caregivers, adopting settings (facilities), and potential business partners to ensure successful adoption of SMART technologies and their long-term use in care for older adults. Experts explicitly pointed at a “missing link” represented by a facilitator that would assist with the implementation of the technology as well as technical support in the process of use. If SMART technologies for older adults show the promise of profit, the adoption might be naturally facilitated by interested business parties. However, with institutionalized care for older adults being a difficult segment to navigate (ie, many regulations related to health services, slow-moving) and home care being an underfunded and underrated segment, the risk of limited interest of business partners to enter these market segments, develop specifically for them, and additionally take on the difficult facilitating and support role is evident. This is where nonprofit organizations and government-funded agencies might prove highly beneficial in facilitating the connection between researchers, business partners and end users, be it older adults, care facilities, or individual caregivers; providing educational programs on technology; and creating the “safe spaces” to gain a firsthand experience with the technology. Future research studies on this topic should incorporate longitudinal designs to track technology adoption over time and ideally in combination with a mixed methods approach to capture not only the statistical trends but also the richness of the underlying motivations, barriers, and experiences of different stakeholders to provide a more comprehensive understanding of technology adoption. Theoretical frameworks related to the diffusion or spread of technologies (such as the Diffusions of Innovations theory) can help provide a theoretical foundation for such work.

Acknowledgments

This paper is a partial result of the Technology Agency of the Czech Republic (TAČR) ÉTA project number: TL03000520, which was implemented between 2020 and 2023 with the financial support of TAČR. The preparation of the manuscript was supported by the project “Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583” which is cofinanced by the European Union.

Data Availability

Data used in this study are available upon request from the first author.

Authors' Contributions

SE conceptualized the study; SE and LK contributed to methodology (KJ collaborated on design of the interview protocol; RC and JK supervised modifications to ANUME); SE, LK, and KJ performed the data collection; SE, LK, and KJ prepared the original draft; RC, JK, and TC contributed to writing—review and editing; and SE supervised the study and acquired the funding. All authors have read and agreed to the published version of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

RC is a partner in the ANUME Ltd. that provided the smart bed system tested in this study. The manuscript is the outcome of a joint project funded by the Technological Agency of the Czech Republic where ANUME Ltd. was an official project partner.

Multimedia Appendix 1
Data collection instruments.
[DOCX File , 43 KB - jmir_v26i1e45492_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Evaluation of technologies by older adults.
[DOCX File , 40 KB - jmir_v26i1e45492_app2.docx]

References

1. Proměny věkového složení obyvatelstva ČR - 2001-2050. Český Statistický Úřad. URL: https://www.czso.cz/csu/czso/promeny-vekoveho-slozeni-obyvatelstva-cr-2001-2050 [accessed 2022-01-14]
2. Mair CA, Quiñones AR, Pasha MA. Care preferences among middle-aged and older adults with chronic disease in Europe: individual health care needs and national health care infrastructure. Gerontologist 2016 Aug;56(4):687-701 [FREE Full text] [doi: 10.1093/geront/gnu119] [Medline: 26035873]
3. Andel R. Aging in the Czech Republic. Gerontologist 2014 Dec;54(6):893-900. [doi: 10.1093/geront/gnu047] [Medline: 24855314]
4. Vaculíková J, Vávrová S. Exploring the meaning of old age from the Czech adult perspective: a quantitative research study. Kontakt 2019 Sep;21(3):326-333. [doi: 10.32725/kont.2019.038]
5. Vrabková I, Ertingerová I, Kukuliač P. Determination of gaps in the spatial accessibility of nursing services for persons over the age of 65 with lowered self-sufficiency: evidence from the Czech Republic. PLoS One 2021 Jan 11;16(1):e0244991 [FREE Full text] [doi: 10.1371/journal.pone.0244991] [Medline: 33428655]
6. Kubalčíková K, Havlíková J. Current developments in social care services for older adults in the Czech Republic: trends towards deinstitutionalization and marketization. J Soc Serv Res 2016 Mar 07;42(2):180-198 [FREE Full text] [doi: 10.1080/01488376.2015.1129014]
7. Oxford Learner's Dictionaries homepage. Oxford Learner's Dictionaries. URL: https://www.oxfordlearnersdictionaries.com/ [accessed 2022-04-08]
8. International classification of health interventions (ICHI). World Health Organization. URL: https://www.who.int/standards/classifications/international-classification-of-health-interventions [accessed 2021-01-18]
9. Chen LK. Gerontechnology and artificial intelligence: better care for older people. Arch Gerontol Geriatr 2020 Sep 10;91:104252 [FREE Full text] [doi: 10.1016/j.archger.2020.104252] [Medline: 32979638]
10. Pedrozo Campos Antunes T, Souza Bulle de Oliveira A, Hudec R, Brusque Crocetta T, Ferreira de Lima Antão JY, de Almeida Barbosa RT, et al. Assistive technology for communication of older adults: a systematic review. Aging Ment Health 2019 Apr;23(4):417-427. [doi: 10.1080/13607863.2018.1426718] [Medline: 29451399]
11. Choi YK, Lazar A, Demiris G, Thompson HJ. Emerging smart home technologies to facilitate engaging with aging. J Gerontol Nurs 2019 Dec 01;45(12):41-48. [doi: 10.3928/00989134-20191105-06] [Medline: 31755542]
12. Clipper B. The influence of the COVID-19 pandemic on technology: adoption in health care. Nurse Lead 2020 Oct;18(5):500-503 [FREE Full text] [doi: 10.1016/j.mnl.2020.06.008] [Medline: 32837346]
13. Bert F, Giacometti M, Gualano MR, Siliquini R. Smartphones and health promotion: a review of the evidence. J Med Syst 2014 Jan;38(1):9995 [FREE Full text] [doi: 10.1007/s10916-013-9995-7] [Medline: 24346929]
14. Hale TM, Goldner M, Stern M, Drentea P, Cotton SR. Patterns of online health searching 2002-2010: implications for social capital, health disparities and the de-professionalization of medical knowledge. Res Sociol Health Care 2014;32:35-60. [doi: 10.1108/S0275-495920140000032016]
15. Liang TH. Matrix analysis of the digital divide in eHealth services using awareness, want, and adoption gap. J Med Internet Res 2012 Feb 13;14(1):e11 [FREE Full text] [doi: 10.2196/jmir.1670] [Medline: 22329958]
16. Kruse CS, Mileski M, Moreno J. Mobile health solutions for the aging population: a systematic narrative analysis. J Telemed Telecare 2017 May;23(4):439-451 [FREE Full text] [doi: 10.1177/1357633X16649790] [Medline: 27255207]
17. Kampmeijer R, Pavlova M, Tambor M, Golinowska S, Groot W. The use of e-health and m-health tools in health promotion and primary prevention among older adults: a systematic literature review. BMC Health Serv Res 2016 Sep 05;16 Suppl 5(Suppl 5):290 [FREE Full text] [doi: 10.1186/s12913-016-1522-3] [Medline: 27608677]
18. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2014 Feb 16;14:81-95. [doi: 10.1007/s10209-014-0348-1]
19. Khechine H, Lakhal S, Ndjambou P. A meta‐analysis of the UTAUT model: eleven years later. Can J Adm Sci 2016 Jun 06;33(2):138-152. [doi: 10.1002/cjas.1381]
20. Olsen WK, Haralambos M, Holborn M. Triangulation in social research: : qualitative and quantitative methods can really be mixed. In: Developments in Sociology. Harlow, UK: Causeway Press Ltd; 2004.
21. Blaikie NW. A critique of the use of triangulation in social research. Qual Quant 1991 May;25:115-136. [doi: 10.1007/bf00145701]
22. Terp R, Kayser L, Lindhardt T. Older patients' competence, preferences, and attitudes toward digital technology use: explorative study. JMIR Hum Factors 2021 May 14;8(2):e27005 [FREE Full text] [doi: 10.2196/27005] [Medline: 33988512]
23. Lim HT, Han JW. A user-centered classification framework for digital service innovation: case for elderly care service. Int J Content 2018 Mar;14(1):7-11. [doi: 10.5392/IJoC.2018.14.1.007]

XSL•FO
RenderX

24. Magsamen-Conrad K, Upadhyaya S, Joa CY, Dowd J. Bridging the divide: using UTAUT to predict multigenerational tablet adoption practices. Comput Human Behav 2015 Sep 01;50:186-196 [FREE Full text] [doi: 10.1016/j.chb.2015.03.032] [Medline: 25937699]
25. Vroman KG, Arthanat S, Lysack C. “Who over 65 is online?” Older adults’ dispositions toward information communication technology. Comput Hum Behav 2015 Feb;43:156-166. [doi: 10.1016/j.chb.2014.10.018]
26. Wagner N, Hassanein K, Head M. Computer use by older adults: a multi-disciplinary review. Comput Hum Behav 2010 Sep;26(5):870-882. [doi: 10.1016/j.chb.2010.03.029]
27. Olson KE, O'Brien MA, Rogers WA, Charness N. Diffusion of technology: frequency of use for younger and older adults. Ageing Int 2011 Mar;36(1):123-145 [FREE Full text] [doi: 10.1007/s12126-010-9077-9] [Medline: 22685360]
28. Selwyn N, Gorard S, Furlong J, Madden L. Older adults' use of information and communications technology in everyday life. Ageing Soc 2003 Sep 02;23(5):561-582. [doi: 10.1017/S0144686X03001302]
29. Knapova L, Klocek A, Elavsky S. The role of psychological factors in older adults' readiness to use eHealth technology: cross-sectional questionnaire study. J Med Internet Res 2020 May 28;22(5):e14670 [FREE Full text] [doi: 10.2196/14670] [Medline: 32348251]
30. Schroeder T, Haug M, Gewald H. Data privacy concerns using mHealth apps and smart speakers: comparative interview study among mature adults. JMIR Form Res 2022 Jun 13;6(6):e28025 [FREE Full text] [doi: 10.2196/28025] [Medline: 35699993]
31. Schoenborn NL, Van Pilsum Rasmussen SE, Xue QL, Walston JD, McAdams-Demarco MA, Segev DL, et al. Older adults' perceptions and informational needs regarding frailty. BMC Geriatr 2018 Feb 13;18(1):46 [FREE Full text] [doi: 10.1186/s12877-018-0741-3] [Medline: 29433426]
32. Lutz C, Tamó-Larrieux A. The robot privacy paradox: understanding how privacy concerns shape intentions to use social robots. Hum Mach Commun 2020;1:87-111. [doi: 10.30658/hmc.1.6]
33. Juárez MR, González VM, Favela J. Effect of technology on aging perception. Health Informatics J 2018 Jun;24(2):171-181 [FREE Full text] [doi: 10.1177/1460458216661863] [Medline: 27496861]
34. Cain M, Mittman R. Diffusion of innovation in health care. Institute for the Future. 2002 May. URL: https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101155468-pdf [accessed 2023-12-20]
35. Dearing JW, Cox JG. Diffusion of innovations theory, principles, and practice. Health Aff (Millwood) 2018 Feb;37(2):183-190. [doi: 10.1377/hlthaff.2017.1104] [Medline: 29401011]

Abbreviations

V-U3A: Virtual University of the Third Age

Edited by T Leung, G Eysenbach; submitted 04.01.23; peer-reviewed by N Kaufman, J Ford Ii; comments to author 17.06.23; revised version received 04.07.23; accepted 25.11.23; published 07.02.24.

Please cite as:
Elavsky S, Knapova L, Janiš K, Cimler R, Kuhnova J, Cernicky T
Multiple Perspectives on the Adoption of SMART Technologies for Improving Care of Older People: Mixed Methods Study
J Med Internet Res 2024;26:e45492
URL: https://www.jmir.org/2024/1/e45492
doi:10.2196/45492
PMID:38324345

©Steriani Elavsky, Lenka Knapova, Kamil Janiš, Richard Cimler, Jitka Kuhnova, Tomas Cernicky. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study

Amy Thomas[1], MD; Andrea Asnes[1], MSW, MD; Kyle Libby[2], BA, MS; Allen Hsiao[1], MD, FAAP, FAMIA; Gunjan Tiyyagura[1], MHS, MD

[1]Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States

[2]3M | M*Modal, 3M Health Information Systems, 3M Company, Maplewood, MN, United States

Corresponding Author:
Gunjan Tiyyagura, MHS, MD
Department of Pediatrics
Yale University School of Medicine
Pediatric Emergency Medicine, PO Box 208064
New Haven, CT, 06520
United States
Phone: 1 203 464 6343
Fax: 1 203 737 7447
Email: gunjan.tiyyagura@yale.edu

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e60444

Abstract

Background: Despite the impact of physical abuse on children, it is often underdiagnosed, especially among children evaluated in emergency departments (EDs). Electronic clinical decision support (CDS) can improve the recognition of child physical abuse.

Objective: We aimed to develop and test the usability of a natural language processing–based child abuse CDS system, known as the Child Abuse Clinical Decision Support (CA-CDS), to alert ED clinicians about high-risk injuries suggestive of abuse in infants' charts.

Methods: Informed by available evidence, a multidisciplinary team, including an expert in user design, developed the CA-CDS prototype that provided evidence-based recommendations for the evaluation and management of suspected child abuse when triggered by documentation of a high-risk injury. Content was customized for medical versus nursing providers and initial versus subsequent exposure to the alert. To assess the usability of and refine the CA-CDS, we interviewed 24 clinicians from 4 EDs about their interactions with the prototype. Interview transcripts were coded and analyzed using conventional content analysis.

Results: Overall, 5 main categories of themes emerged from the study. CA-CDS benefits included providing an extra layer of protection, providing evidence-based recommendations, and alerting the entire clinical ED team. The user-centered, workflow-compatible design included soft-stop alert configuration, editable and automatic documentation, and attention-grabbing formatting. Recommendations for improvement included consolidating content, clearer design elements, and adding a hyperlink with additional resources. Barriers to future implementation included alert fatigue, hesitancy to change, and concerns regarding documentation. Facilitators of future implementation included stakeholder buy-in, provider education, and sharing the test characteristics. On the basis of user feedback, iterative modifications were made to the prototype.

Conclusions: With its user-centered design and evidence-based content, the CA-CDS can aid providers in the real-time recognition and evaluation of infant physical abuse and has the potential to reduce the number of missed cases.

(J Med Internet Res 2024;26:e51058) doi:10.2196/51058

KEYWORDS

child abuse; clinical decision support; CDS; pediatrics; child; children; natural language processing; usability; clinical decision support system; physical abuse

XSL•FO
RenderX

Introduction

Background

Child physical abuse is commonly missed by emergency department (ED) providers, leading to escalating injuries and death [1]. More than 30% of children with serious injuries resulting from physical abuse have been previously evaluated for injuries that were not recognized as abusive [2-5]. This is amplified in general EDs, where most children receive emergency care and abuse is more frequently missed than in pediatric EDs [6-8]. Evaluation and reporting of child abuse are also impacted by provider biases [2,9-12]. Children belonging to racial or ethnic minority groups are more often evaluated for abusive head trauma than White or non-Hispanic children, and children with public insurance undergo increased testing and are reported more often to Child Protective Services (CPS) than privately insured children [9,12]. These findings highlight the need for systems that standardize care, improve clinical outcomes, and reduce bias.

Clinical decision support (CDS) integrated into the electronic health record (EHR) can present intelligently filtered, individualized, and timely information to enhance clinical decision-making [13]. Child abuse–specific CDS systems may improve outcomes and reduce bias in the evaluation and reporting of suspected abuse. Experts have shared consensus recommendations regarding developing, disseminating, and sustaining EHR-embedded child abuse CDS systems in the ED [14]. Key recommendations included universal, routine implementation of a child abuse CDS system in general and pediatric EDs for children aged <4 years; use of active alerts that share their reason for triggering; integration of a standardized system for reports to CPS; use of data warehouse reports to evaluate the CDS system's efficacy; integration of a system that is feasible, sustainable, and easily disseminated; and personalized usability testing to ensure seamless integration of the system [14].

While reviewing the existing child abuse CDS systems [15-32], we found that a common limitation is their inability to be triggered by free text in an EHR encounter. To address this gap, our team previously developed and validated a natural language processing (NLP) algorithm that automatically and methodically examined the free text in the notes of nursing providers, medical providers, and social workers (SWs) to identify high-risk injuries associated with possible abuse in children [33]. The NLP algorithm would provide a positive alert when it identified preselected combinations of written terms associated with fractures, intracranial injuries, abdominal injuries, burns, bruising, or oral injuries. It was targeted to identify high-risk injuries in children aged <1 year (ie, infants) specifically given that infants are more than twice as likely to experience maltreatment and thrice as likely to experience fatality from maltreatment compared to older children of any age group [1]. Developing a novel child abuse CDS system triggered by this validated NLP algorithm may further increase the tool's potential to reduce the number of missed cases and mitigate bias.

To change providers' practice using CDS, it is crucial to understand the providers' needs and priorities before development and implementation. Evaluating a system's usability involves the assessment of its accommodation of users' needs, ease of mastery, effects on workflow, and achievement of goals. Conducting evaluations during the design process is also important to identify shortcomings and incorporate user-centered modifications [34,35]. Usability testing can include direct observation, recording of user-system interactions, *think-aloud* sessions where users verbalize their thoughts while interacting with the system, *near-live* sessions where users test the system with simulated patient interactions, live testing, and quantitative measures [34,36]. However, to date, only 1 study has described the usability testing of child abuse CDS in local settings [31].

Objective

Therefore, in this study, we aimed to develop a novel child abuse CDS system—hereafter referred to as the Child Abuse Clinical Decision Support (CA-CDS)—which is triggered by a validated NLP algorithm and that both alerts ED providers to high-risk injuries in infants and provides evidence-based recommendations for evaluation and management. We also sought to test the usability of the CA-CDS and refine the system based on user feedback.

Methods

Study Design

The study consisted of 3 phases informed by the Guideline Implementation with Decision Support (GUIDES) checklist by Van de Velde et al [37], which describes factors relevant to the development of successful guideline-based CDS. The phases included the (1) development of a prototype, (2) mixed methods usability testing, and (3) iterative refinement of the CA-CDS based on stakeholder feedback. Participants were stakeholders from 4 EDs, including 1 (25%) academic pediatric ED (Yale New Haven Children's Hospital) and 3 (75%) community pediatric and general EDs (Bridgeport Hospital, Lawrence + Memorial Hospital, and Saint Raphael Campus). All campuses use Epic (Epic Systems Corporation) as their EHR and can use the M*Modal Fluency Direct speech recognition technology and Natural Language Understanding platform (3M Company), which hosts the NLP algorithm and presents the CA-CDS via built-in computer-assisted physician documentation functionality.

Development of the Prototype

The initial CA-CDS was developed after literature review and discussions with local experts in child abuse, pediatric emergency medicine, and health informatics. The issues discussed included target users (medical and nursing providers in EDs), appropriate language, recommendations considering the local context (eg, using order sets vs consulting the local child protection team [CPT]), and degree of interruption (ie, hard-stop vs soft-stop alert in which the former requires alert completion to proceed with one's workflow). The prototype, as depicted in Figures 1-3, consisted of a *card* and *protocol* that appeared in the EHR once the NLP algorithm identified a

XSL•FO
RenderX

high-risk injury within a note's free text. A smaller card would first appear, stating that a high-risk injury was found, with the triggering language presented in a tooltip. The card would then allow providers to open a larger protocol that presented further information about the triggering language, suggested questions for evaluation, and suggested actions for management. Users could then select between 2 acknowledgment options regarding the likelihood of child abuse or neglect and select the actions taken, which would automatically be entered into an editable documentation field. Finally, they could click *submit response* to add the documentation to the bottom of their note or *later* to minimize the CA-CDS such that it no longer blocked the provider's view of the EHR but remained accessible via the Fluency Direct pop-up bar. The CA-CDS was designed as a soft-stop alert such that completion was not required and workflow was not permanently interrupted.

To tailor the CA-CDS to the needs of medical versus nursing providers, the content was customized for each provider type. For instance, for medical providers (Figure 1), the suggested questions were based on the MORE (Mechanism, Others present, Review of development, and Examination details) mnemonic. The components of the mnemonic ("Mechanism: additional details about history and injury mechanism; Others present: witnesses to injury and history corroboration; Review of development: developmental ability; and Examination details: disrobed exam, specifically to examine for sentinel injuries, and additional details related to the physical examination") aids providers in differentiating between accidental and abusive injuries [38]. For nurses (Figure 2), the content was simplified and asked whether the history was consistent with the injury. The suggested actions included recommendations to contact the local CPT (known as the Detection, Assessment, Referral, and Treatment or DART team) and file a report with Connecticut's CPS agency known as the Department of Children and Families (DCF) as appropriate. The nursing version also recommended discussing concerns with medical providers. A subsequent-provider version was also designed to be received after another provider had already submitted their response (Figure 3). This version was equivalent to the first-provider version, except for text indicating that another provider had responded and its modified acknowledgment options, allowing the subsequent provider to disagree with the previous provider's selection regarding the likelihood of abuse, agree without further action, or agree and take additional action.

A web-based prototype of the CA-CDS in a model EHR was designed using the InVision platform (InVisionApp Inc) for usability testing with the abovementioned features (Figures 1-3). The model EHR showed a clinical vignette of an infant presenting for emergency care for whom a provider documented a high-risk injury that triggered the CA-CDS (Table S1 in Multimedia Appendix 1). For this study, the CA-CDS was tested solely in a model EHR before future live implementation.

Figure 1. Initial first-provider version of the medical provider–specific Child Abuse Clinical Decision Support (CA-CDS) prototype within a model electronic health record. DART: Detection, Assessment, Referral, and Treatment; DCF: Department of Children and Families; ED: emergency department; M: male; MORE: Mechanism, Others present, Review of development, and Examination details.

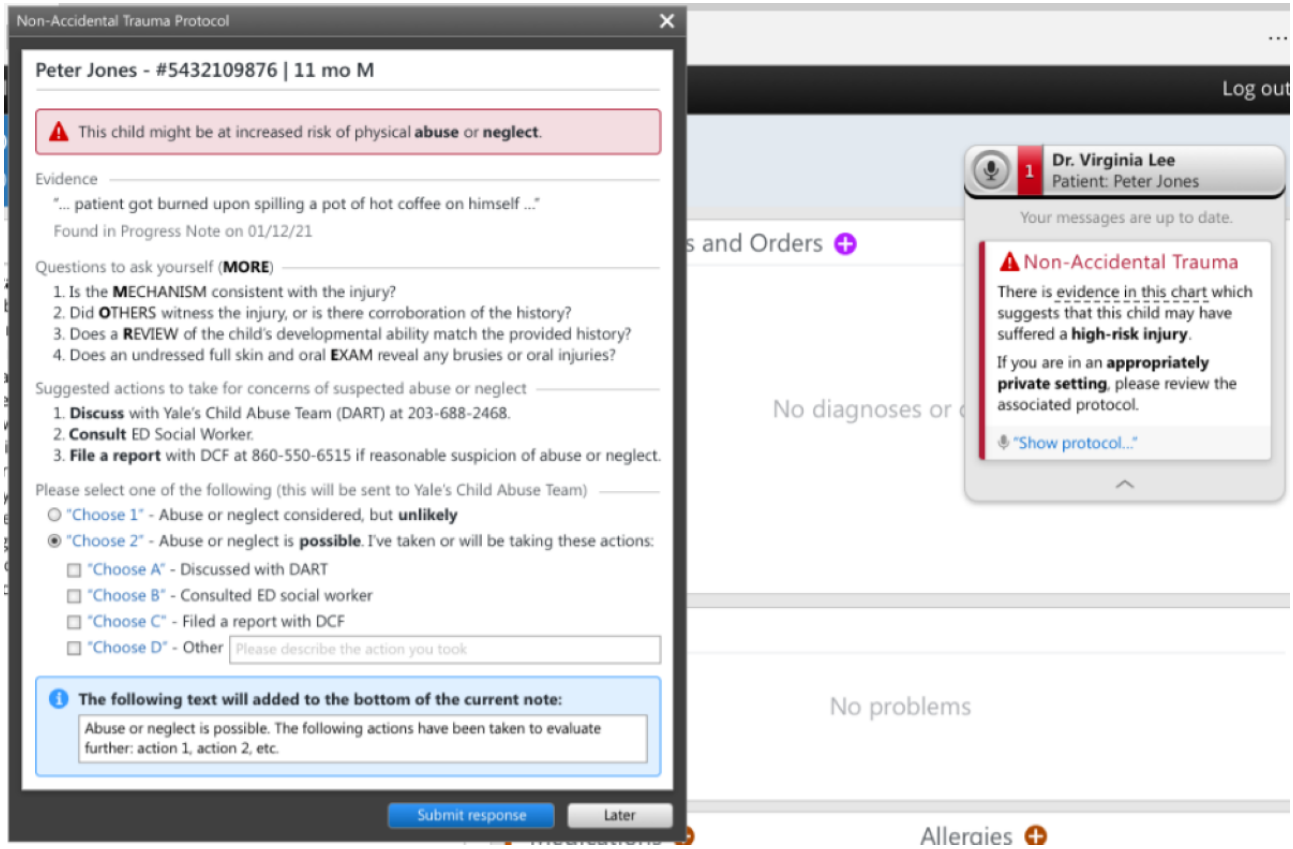

XSL•FO
RenderX

Figure 2. Initial first-provider version of the nursing provider–specific Child Abuse Clinical Decision Support (CA-CDS) prototype. DART: Detection, Assessment, Referral, and Treatment; DCF: Department of Children and Families; M: male.

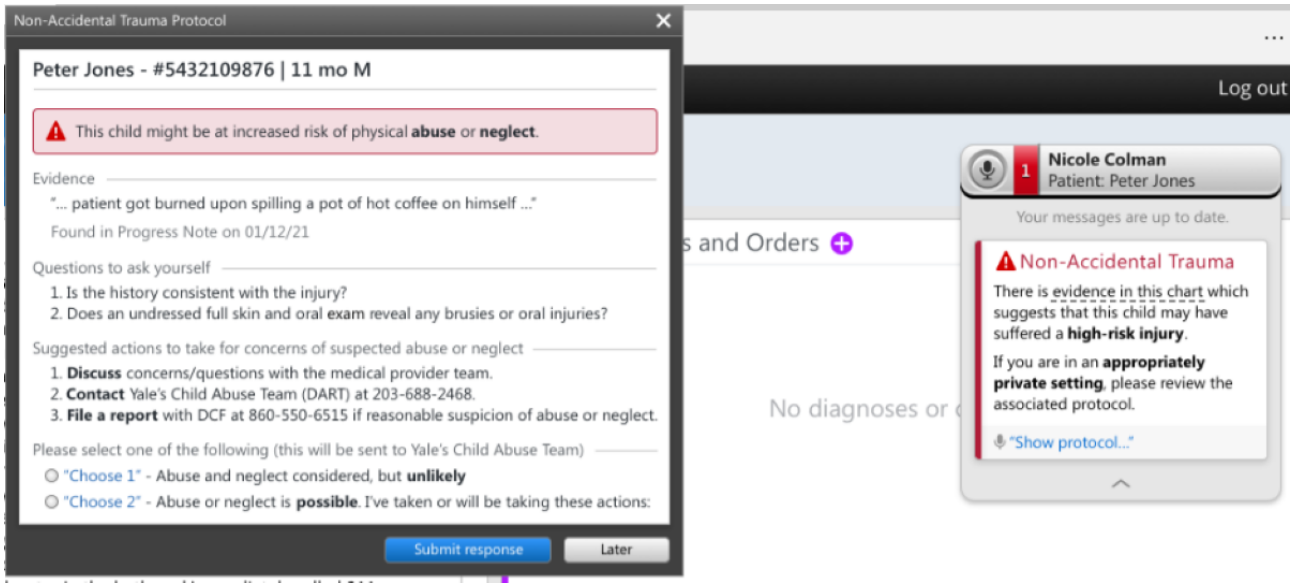

Figure 3. Initial subsequent-provider version of the medical provider–specific Child Abuse Clinical Decision Support (CA-CDS) prototype. DART: Detection, Assessment, Referral, and Treatment; DCF: Department of Children and Families; ED: emergency department; F: female; MORE: Mechanism, Others present, Review of development, and Examination details.

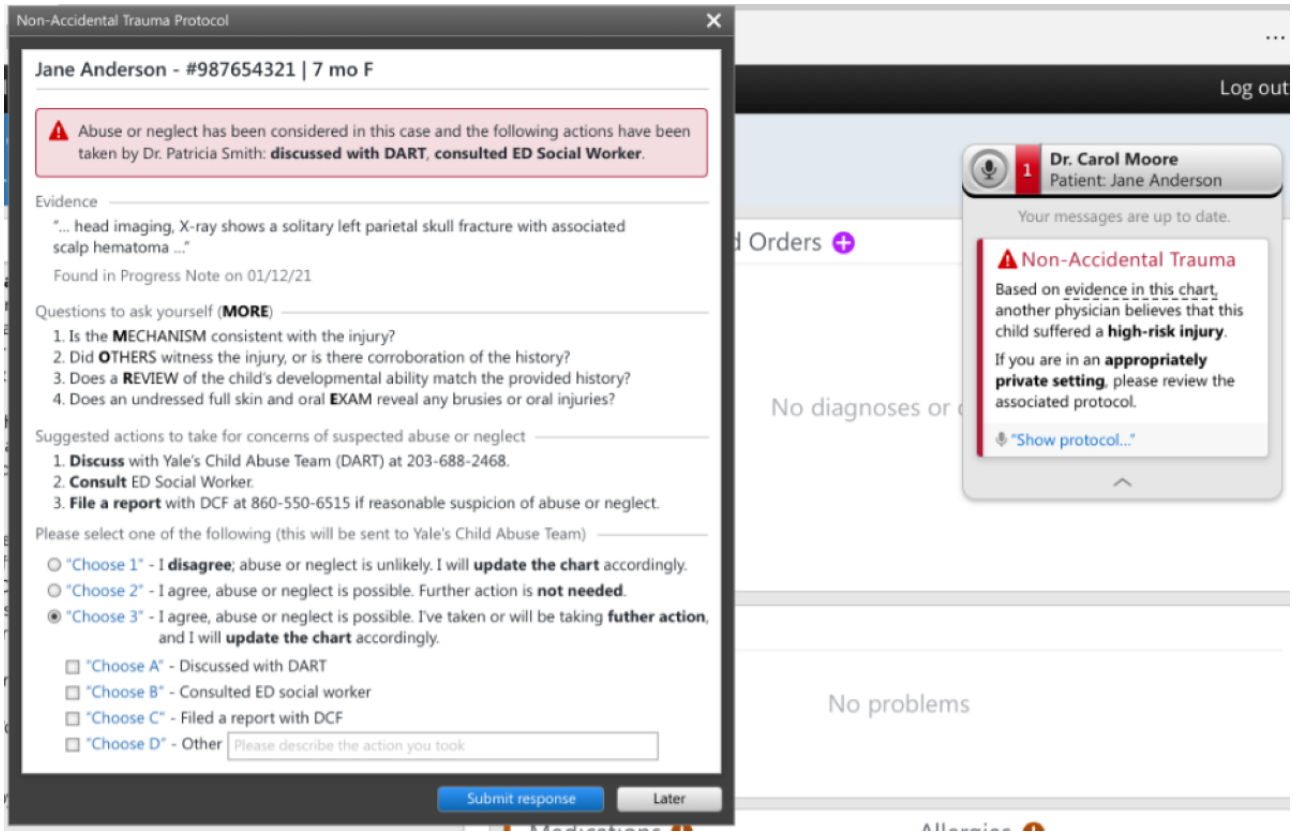

Usability Testing

We tested the CA-CDS's usability through a mixed methods approach. The research team, including a user design expert (KL) and researchers with qualitative research expertise (GT and AA), developed and iteratively refined an interview guide (Table S2 in Multimedia Appendix 1) with open-ended questions about topics including the CA-CDS's design, strengths, deficits, and recommendations for improvement and future implementation. Purposive recruitment of stakeholders for interviews who represented the CA-CDS's end users and local champions in child abuse care was conducted via email, in person, and through ED section meetings. Overall, 3 rounds of interviews were conducted by GT and AT, with audiovisual recording for documenting user-system interactions and transcript generation (Figure 4). Interviews were conducted until thematic sufficiency was achieved [39-41].

XSL•FO
RenderX

Figure 4. Flow diagram of usability testing. CA-CDS: Child Abuse Clinical Decision Support.

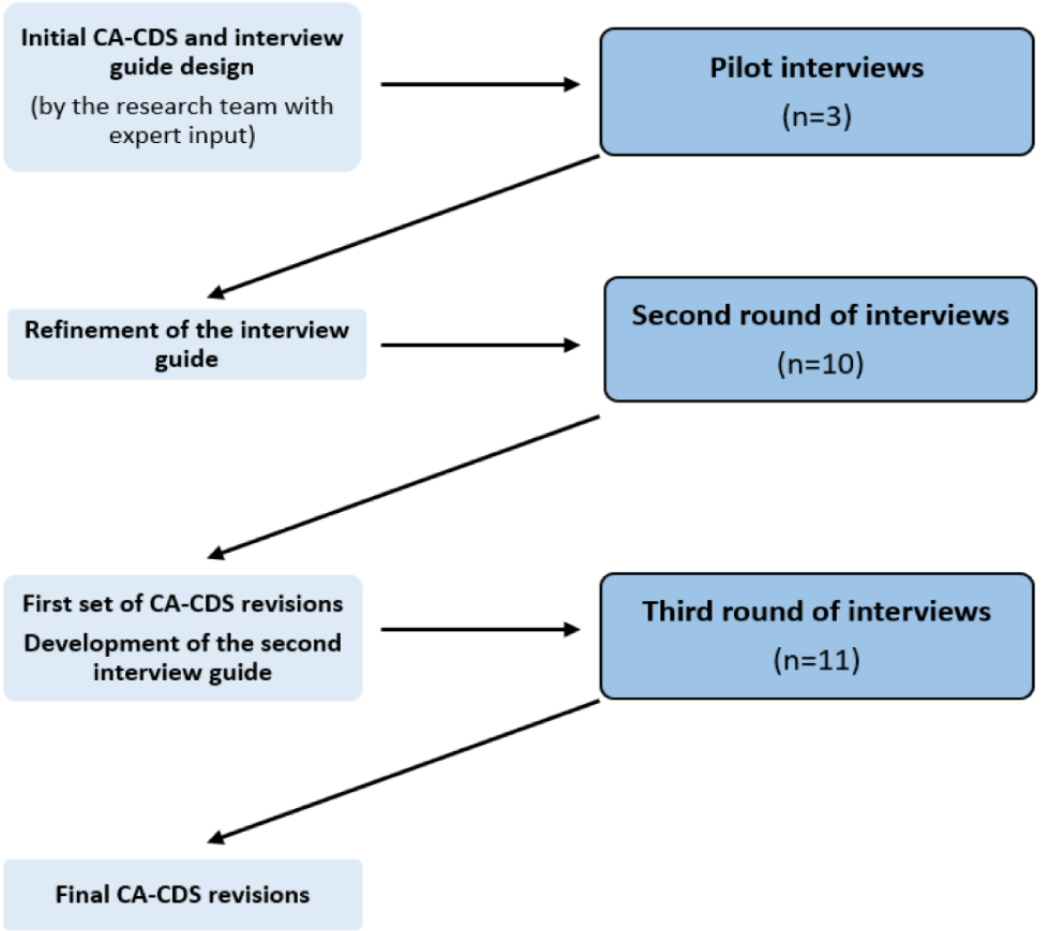

The web-based prototype (accessed via hyperlink) and interview guide were pilot-tested through in-person, think-aloud interviews with 3 ED providers (Figure 4). After further refinement of the interview guide, interviews were conducted via teleconferencing with 10 additional providers. Participants were instructed to think aloud while interacting with the prototype and then asked targeted questions using the interview guide. On the basis of the findings from the initial interviews, the CA-CDS was refined, and usability of the updated prototype was assessed with another round of interviews. Here, we sought to address topics that we felt needed more exploration such as preferred resources, documentation-related concerns, and target users for the subsequent-provider alert, and thus, we designed a more targeted interview guide (Table S2 in Multimedia Appendix 1). 11 additional ED providers were recruited in person in the ED to participate in a final round of interviews. The updated prototype was provided as a multipage PDF document on the interviewer's tablet. The prototype was further refined based on these interviews.

Following each interview, participants were asked to complete a survey to capture demographic and quantitative usability data with adequate time and privacy for completion (Figure S1 in Multimedia Appendix 1). The survey requested the participants' ID numbers to anonymously link their survey and interview transcript, profession and years in their role, employment site, and experience with suspected child abuse cases. To assess usability quantitatively, we used the System Usability Scale (SUS), which is a 5-point Likert scale with 10 questions exploring the different aspects of a tool's usability and learnability. The SUS is a validated, frequently used scale that provides a quick, standardized, and easily interpretable measure for reporting and comparing a product's usability [42,43].

Analysis

The interview transcripts were anonymized and independently reviewed by a coding team consisting of 3 researchers, including 1 experienced in the analysis of qualitative data (GT) and 1 experienced in usability testing (KL), using conventional content analysis [44]. Team members applied codes to categorize data. Researchers then met to discuss the codes until consensus was reached, and a code list was subsequently generated and iteratively revised as new interviews were discussed [41]. The codes were then clustered into recurrent categories.

Regarding surveys, each participant's SUS responses were scored, with the score ranging between 0 and 100 [45]. A cumulative SUS score for the CA-CDS was then obtained by calculating the median and IQR of the data set of participants' scores. This cumulative score was compared against the curved grading scale developed by Lewis and Sauro [46] in which an SUS score of 68 corresponded to the 50th percentile of the range

XSL•FO
RenderX

of scores included in their study and thus a "C" letter grade [46,47]. In their study and industry, an SUS score ≥80 indicated an above-average user experience.

Ethical Considerations

All participants provided verbal informed consent to be interviewed and recorded before starting the interviews. Participants received no compensation. This study was approved by the Yale Human Investigations Committee (2000029566).

Results

Participant Characteristics

In total, 24 participants were interviewed in the study, and 23 participants completed the demographic survey. Most were physicians (13/23, 57%), from Yale New Haven Children's Hospital (19/23, 83%), and held their current roles for >6 years (13/23, 57%; Table 1).

Table 1. Demographics of interviewees (n=23a).

Demographics	Values, n (%)
Profession	
Medical provider	
Physician assistant	2 (9)
Nurse practitioner	1 (4)
Physician (attending or fellow)	10 (43)
Physician (resident)	3 (13)
Nursing provider	7 (30)
Primary hospital affiliation	
Yale New Haven Children's Hospital	19 (83)
Saint Raphael Campus	1 (4)
Lawrence + Memorial Hospital	2 (9)
Bridgeport Hospital	1 (4)
Duration in the role (y)	
<1	1 (4)
1-5	9 (39)
6-10	4 (17)
≥11	9 (39)
Average monthly exposure to cases with suspicion for child abuse (number of patients)	
0-1	9 (39)
2-5	11 (48)
6-10	2 (9)
>10	1 (4)

[a]In total, 24 participants were interviewed, but 1 (4%) participant was unable to complete the Qualtrics survey that requested demographic data due to conflicting clinical obligations.

Emerging Themes

Overview

Analysis of the interviews revealed 5 main categories of themes. These themes, along with sample subcategories and representative quotations, are presented in Table 2.

XSL•FO
RenderX

Table 2. Categories of themes emerging from the coding of interview transcripts.

Categories and sample subcategories	Representative quotes
CA-CDS[a] benefits	
Extra layer of protection	"The advantage is that you're going to try to reduce the needle in the haystack phenomenon of missing these very occult cases of child abuse."
Inclusion of evidence-based recommendations	"I think the trigger questions are...helping us better interview patients and families, so that when we called DART[b], we have a better history to relate to them...It's helpful that there's phone numbers and...that it kind of prompts you through the next steps."
Alerting the entire clinical ED[c] team	"It gives each individual provider a chance to document what they need to document about their concerns for each individual patient."
User-centered, workflow-compatible design	
Customization based on provider type	"From a nursing end, obviously discussing it with the care team, with all of us being involved and voicing our concerns on the child."
Soft-stop alert configuration	"Hard stops...could be in the wrong place...I think soft is good. I think being able to close [the alert] because [the flow] is just so unpredictable."
Editable and automatic documentation	"I think as long as those options can be documented in the chart just to save somebody the step of writing...That would be helpful."
Triggers from multiple providers' notes	"You getting to the point of actually doing documentation is midway or further down. So, if it's also looking at the nursing notes..., then on [opening the] chart, I would get this message."
Clear presentation of alert trigger	"[The evidence] is factual...They can get everybody on the same page with what the concern is, and what everybody needs to be aware of that's caring for that child."
Attention-grabbing design elements	"I like that it has...the exclamation point with the red triangle that, kind of, alerts you to pay attention to it."
Accessible recording of the previous provider's actions	"I think [the previous provider's actions box] is good because that's the sign-out. So, I think that's reiterated as what should be signed out from the prior provider."
Recommendations for improvement	
Consolidating the content	"As opposed to having so much in this 'Please select one of the following, this will be sent to the Yale child abuse team' stuff popping up, [I wonder] if the suggested actions for concern of suspected abuse and neglect...could be clickable themselves."
Clearer design elements	"I don't know if [the evidence tooltip] was a different color or, like, you know like a hyperlink is in an email...how it shows up as a different color and underlined...I just wasn't aware that's what it was representing."
Adding a hyperlink to additional resources	"The diagnosis is a fracture. Then here, for all this stuff, there can be direct links whether it's either mandated, like happens automatically, or they can link to it to see what the evidence actually is."
Adding further information about the trigger source	"But if it's not your note, is there a way to more readily send to the specific note?...A progress note, that's pretty generic...but it might be nice if there's a way to get a little more specific, that it was the triage note or."
Modifications to better reflect the provider workflow	"I submit that, then this just completely disappearing...it would seem to me, maybe having it be completely disappeared, and if you want, at the point of discharge, to maybe one more time give me the opportunity to say, 'Are you really? Just think about this one more time'."
Barriers to future implementation	
Alert fatigue	"The drawback is that you'll...probably have to trigger 10 alerts for every kid that is a true positive...So, there might be fatigue with the alert."
Infringement of provider autonomy	"There are some sentiments that this is being forced on somebody without any evidence...I don't want to call them naysayers, but they don't really believe the bibliographic evidence."
Hesitancy to change	"It seems like every week we're getting told to do something new. So honestly, I really feel like as much as people don't like change, we're just getting told this is the way it is, and you need to do it."

XSL•FO
RenderX

Categories and sample subcategories	Representative quotes
Concerns regarding documentation	"If I am going to contact DCF[d], and my husband's the one that's beating my kid, and now it's available per the Cures Act, they know that DCF was contacted. You might be putting me and my kid at risk...Just putting it in the notes...[which] are being released to everybody now who has access, that could be a safety concern."
Facilitators of future implementation	
Stakeholder buy-in	"To communicate with physicians, practitioners, you need to do it five different ways. So, I think probably asking to come to staff meetings...that system ED leadership committee...to do high-level show-and-tell."
Provider education	"I think as long as there is some kind of super user that can educate everybody on it...It's a matter of just getting shown once how to do something, and then it kind of sticks."
Sharing the system's test characteristics	"I would love to have some evidence to show that that's benefiting somebody, hopefully the patient...Just some retrospective data to show the veracity or the utility of your system, however you define it, either accuracy or times that you picked up something that the physicians should have or didn't."

[a]CA-CDS: Child Abuse Clinical Decision Support.

[b]DART: Detection, Assessment, Referral, and Treatment.

[c]ED: emergency department.

[d]DCF: Department of Children and Families.

CA-CDS Benefits

Participants discussed the challenges of recognizing abusive injuries, especially those that were subtle or "minor." They expressed that the CA-CDS could provide an extra layer of protection against missing abuse by reminding providers in real time to consider abuse in their differential diagnosis. Users also appreciated the CA-CDS's evidence-based recommendations for evaluation and management that included guidance about important historical information to be collected and about using the expertise of specialists. Specifically, they found the MORE mnemonic to be clear, memorable, and helpful to improve information gathering, decision-making, and documentation. Participants also valued the emphasis on consulting specialists to determine the appropriate workup as it enabled the CA-CDS to remain simple but adaptable despite case-specific variations. In addition, the users appreciated that the CA-CDS alerted both medical and nursing providers, allowing for open communication of concerns among the entire clinical team.

User-Centered, Workflow-Compatible Design

Participants discussed the elements of the CA-CDS that would optimize their workflow. First, they valued the CA-CDS's customization based on provider type, which reflected workflow differences, and preferred that nurses and medical providers submit independent CA-CDS responses. For instance, nurses favored the recommendation to discuss concerns with the medical team rather than consulting the CPT directly as it reflected the typical nursing workflow. Second, users felt that the CA-CDS's soft-stop alert configuration, which could be minimized and reaccessed on demand, would be more flexible around providers' unpredictable workflow. Third, users expressed that the documentation component, which automatically populated the selected actions into the note while also remaining editable for providers to share their own decision-making, would avoid redundancy. Fourth, participants appreciated that the CA-CDS could be triggered by injuries in various providers' notes. In particular, they expressed that by including nursing notes, which are often created before other clinicians' notes, as a trigger source, the CA-CDS could allow for more timely evaluation. Fifth, users shared that having the alert explicitly identify the triggering documentation enabled all team members to quickly be on the same page regarding the specific injury causing concern for abuse and allowed providers to assess if their documentation was being construed as intended. Sixth, participants appreciated the CA-CDS's attention-grabbing formatting. Features such as bold text, colorful symbols, and high-risk injury phrasing helped emphasize the alert's significance. Finally, regarding the subsequent-provider CA-CDS version, participants valued that the protocol clearly displayed the actions selected by the provider who initially submitted a response. They found this helpful for handoffs, highlighting the salient concerns and workup for all team members.

Recommendations for Improvement

Users made several recommendations to improve the CA-CDS's usability. Participants recommended consolidating the content to reduce information overload. For instance, they suggested removing the acknowledgment section to reduce redundant text, allow more flexibility for documentation, and circumvent the potential legal implications of documenting disagreement with previous providers. To improve clarity, users also suggested using obvious underlining and bold colors to highlight design elements such as hyperlinks and editable text fields. In addition, to improve providers' case-specific and general knowledge about abuse, participants suggested adding a link to additional resources that providers could access for further support.

Next, users requested further information about the source of the triggering documentation, including its author and location, to better find and assess the triggering content. Finally, providers suggested modifications to improve their workflow and use of the CA-CDS. For example, nurses appreciated the card's reminder to be in an appropriately private setting as they often charted near patients, whereas medical providers supported

XSL•FO
RenderX

removing this component which was less relevant to their workflow. They also shared that the reappearance of the CA-CDS at discharge could serve as a reminder to those who had missed, ignored, or initially not felt ready to complete the alert and as an opportunity to add more information and reconsider abuse in their differential.

Barriers to Future Implementation

First, participants warned about the potential for alert fatigue, especially if there were several false positives or if excessive effort was required for completion. They discussed how alert fatigue may lead providers to ignore the alert or seek work-arounds such as documenting in a manner to avoid triggering the alert. Second, they shared that the CA-CDS, with its interruptive alert and recommendations to consult specialists, may be perceived by some providers as infringing on their autonomy. Third, users counseled that providers may be accustomed to a particular manner of providing care and hesitant to change.

Finally, users warned that providers may be wary of using the CA-CDS, especially its documentation component, given the potential consequences of documenting concerns about abuse in notes that are accessible to caregivers. These included liability, inadequate patient-sensitive language, caregivers learning about concerns before discussions with the medical team, and caregivers purposefully obstructing care or inflicting further harm. However, users also responded that they tried to document objectively, being mindful about how their documentation could be interpreted by caregivers. On the basis of these discussions, the following text was added to the protocol: “Consider ‘unsharing’ the note ‘to prevent substantive harm to patient or another person.’”

Facilitators of Future Implementation

Participants recommended several strategies to optimize the CA-CDS’s implementation. Users felt that stakeholder and leadership buy-in and support for the system would promote future use and sustainability. In addition, participants stated the importance of educating providers regarding how to use the system to appropriately manage the cases of potential abuse and how to approach caregivers based on evidence-based recommendations. Users also recommended communicating the accuracy of the CA-CDS and its triggering NLP algorithm by sharing the system’s validation data and instances where the system could have made a difference.

Prototype Revisions

On the basis of the interviews and feedback from our team of experts, multiple rounds of modifications were performed to create our final prototype (Figures 5 and 6). All modifications are listed in Table S3 in Multimedia Appendix 1.

Figure 5. Final prototype of the first-provider version of the medical provider–specific Child Abuse Clinical Decision Support (CA-CDS). DART: Detection, Assessment, Referral, and Treatment; DCF: Department of Children and Families; ED: emergency department; M: male; MORE: Mechanism, Others present, Review of development, and Examination details.

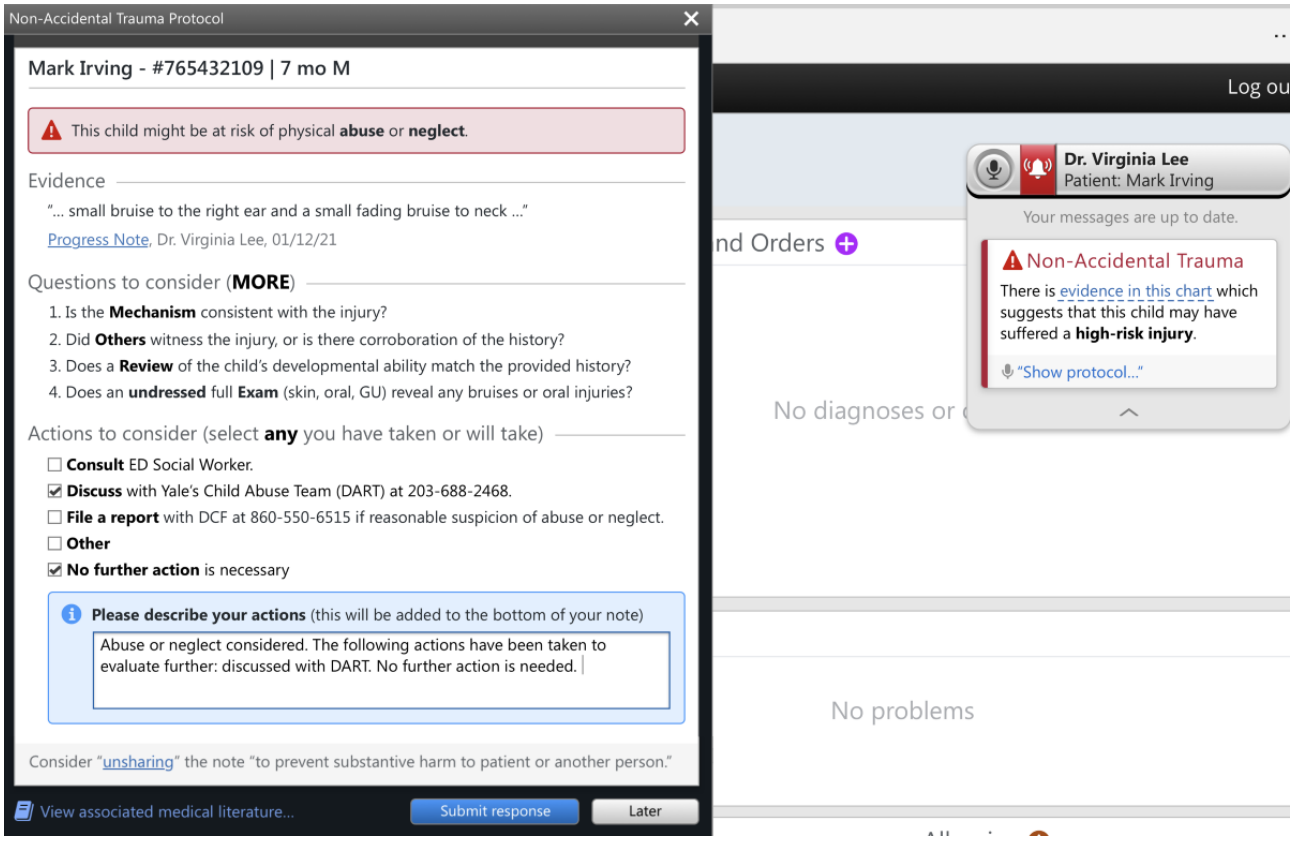

XSL•FO
RenderX

Figure 6. Final prototype of the subsequent-provider version of the medical provider–specific Child Abuse Clinical Decision Support (CA-CDS). DART: Detection, Assessment, Referral, and Treatment; DCF: Department of Children and Families; ED: emergency department; F: female; MORE: Mechanism, Others present, Review of development, and Examination details.

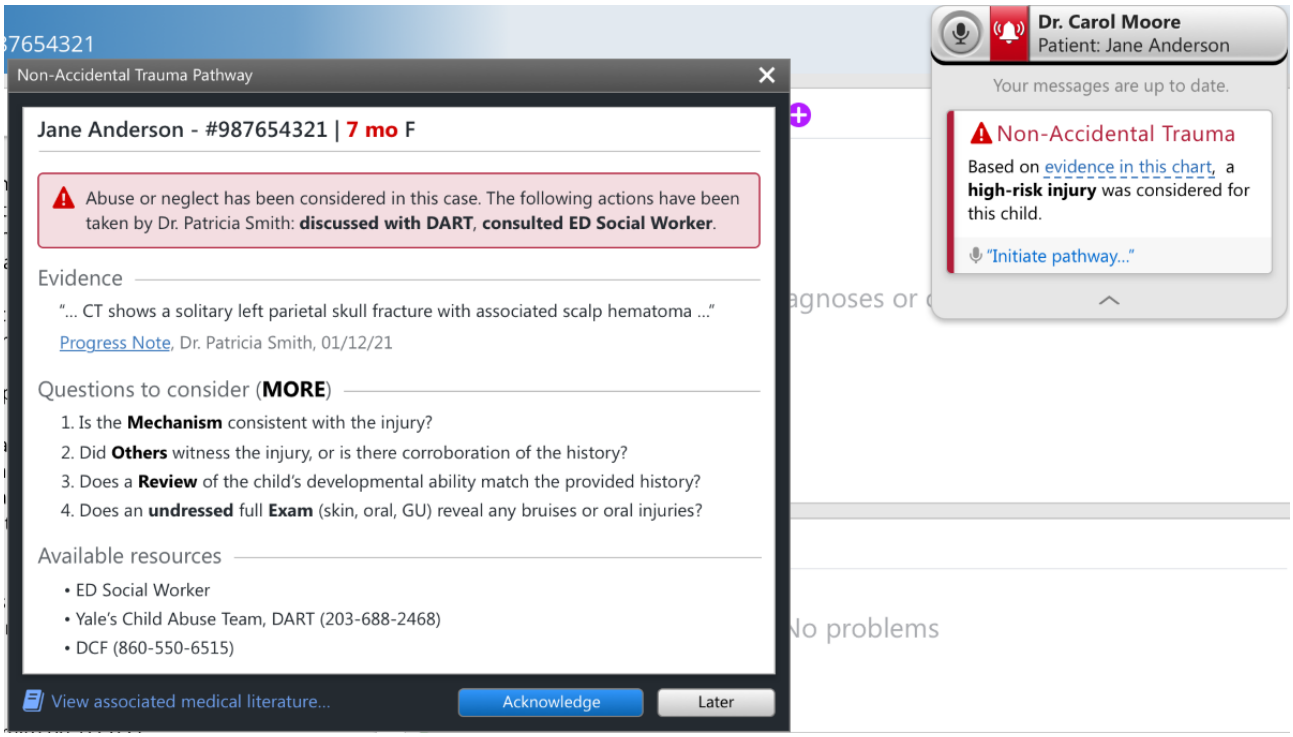

SUS Scores

Of the 24 interviewees, 23 (96%) completed the SUS. Scores ranged from 62.5 to 100, with a median of 80 (IQR 75-92.5). Compared with Lewis and Sauro's [46,47] curved grading scale, our median corresponded to the 85th to 89th percentile and an A− letter grade.

Discussion

Principal Findings

Usability testing of the CA-CDS revealed several key findings. Users valued the additional protection against missing abuse that is offered by the alert to the entire clinical team and the presence of evidence-based recommendations for the evaluation and management of suspected abuse. Users also appreciated the CA-CDS's user-centered, workflow-compatible design elements that captured the user's attention to provide timely, provider-specific information while minimizing interruptions and redundancy. However, they recommended improving the system's clarity and brevity, highlighting critical features such as the triggering documentation's source, and further supporting the users by offering additional resources and alert reappearance at discharge. User recommendations informed the iterative refinements of the CA-CDS prototype. Future studies will be directed toward the implementation and live testing of the revised, user-centric CA-CDS within our hospital system's EHR.

Comparing the CA-CDS With Existing Systems

Researchers have described the development, implementation, and evaluation of a child abuse CDS system for pediatric and general EDs that identified high-risk injuries through a variety of alert triggers including specific screening results, orders, and discharge documentation [26-31]. Their CDS system notified providers about the concern for abuse and recommended direct connection to an age-appropriate, injury-specific order set or a CPS referral. While the CA-CDS similarly aimed to identify high-risk injuries and provide CDS regarding the evaluation of suspected abuse, there were numerous differences. The current CA-CDS was triggered via an NLP algorithm that examined all the free text in the notes of medical providers, nursing providers, and SWs, whereas previous CDS systems were triggered primarily by discrete fields, active screening such as those completed by nurses upon evaluation, or limited NLP function that could only examine the free text within the chief complaint and focused assessment fields [33]. An entirely NLP-triggered CDS system may allow for minimal interruptions to the workflow; be more acceptable to frontline providers; and allow the CA-CDS to be triggered as soon as there is any documentation, even as early as triage, without requiring actions outside the normal workflow.

While many existing CDS systems connect users to standardized order sets [26-31] and recent consensus guidelines also recommended the use of a physical abuse order set with consistent and evidence-based actions [14], most of our users (13/24, 54%) preferred simpler suggested actions with reminders to consult a SW or the CPT to aid in nuanced decision-making. Consultation with these specialists may facilitate appropriate decision-making around performing additional testing or reporting to CPS and reduce bias in evaluation and reporting of suspected child abuse [48,49]. However, to acknowledge the importance of autonomy for users, the CA-CDS was designed as a nonmandatory, soft-stop alert that included a free-text response option and a hyperlink to additional resources including

XSL•FO
RenderX

local clinical pathway guidelines to provide either support or an avenue for independent decision-making depending on the provider's needs. Next steps include comparing the outcomes of systems that recommend standardized order sets to those that recommend consulting clinicians.

Finally, the CA-CDS was hosted on external software from 3M Company and designed to be subsequently integrated into the EHR, rather than being directly built into the EHR. This design aligns with the Fast Healthcare Interoperability Resources (FHIR) data standard that standardizes how information is stored, used, and exchanged between computer systems and thereby streamlines software development to support health care needs [50,51]. EHRs with FHIR-enabled technology allow for the packaging of information from the EHR into discrete, standardized units that can be interpreted and acted upon by external applications including CDS systems. FHIR-based applications, such as those used in this study, allow the results of a CDS system to trigger the opening of order sets or to directly provide CDS within the EHR and may realistically solve the problem of 1 child abuse CDS system communicating with multiple EHRs [31,52,53]. With 84% of hospitals in the United States having adopted FHIR-enabled technology and 3M's connection with hundreds of EHR systems [51,54,55], the CA-CDS's FHIR-based application design may facilitate the system's dissemination across numerous institutions and EHRs.

Examining the Rigor of the CA-CDS

Consistent with the recommendations for successful, guideline-based, computerized CDS as described by the GUIDES checklist and child abuse expert consensus recommendations [14,22,37], the CA-CDS was developed by a team of local experts to provide evidence-based guidelines for the evaluation and management of high-risk injuries, reflective of recent studies in the field. In addition, the CA-CDS was integrated with an objective and internally validated NLP algorithm that captured data widely in the notes of ED SWs, nursing providers, and medical providers [33]. Given that the system is triggered independent of the providers' gestalt and background, the CA-CDS may improve the standardization of patient care and reduce the impact of providers' implicit biases [49,56].

The CA-CDS met the recommended design standards in several ways. Users' feedback demonstrated the system's usability, with users finding the CA-CDS to be user friendly, concise, and clear. The CA-CDS was intentionally refined based on feedback to reflect the users' preferences. Considerable effort was made to integrate the system into providers' clinical and EHR workflow to minimize interruptions and redundancy. The CA-CDS also provided ample flexibility around decision-making through features such as editable and automatic documentation and soft-stop alert design. Interestingly, in contrast to experts' recommendation to incorporate automated referrals, standardized CPS reporting, and a multidisciplinary audience [14,22], most users preferred to keep the CA-CDS simple without these features (19/24, 79%) and limit the alert's recipients to the primary clinical team (7/13, 54%). Next steps include examining the real-time use of the CA-CDS by ED clinicians.

Similar to the consensus recommendations, participants discussed the importance of planning for future implementation [14,37]. They identified the facilitators of future implementation, such as stakeholder buy-in, education about the CA-CDS and the accuracy of the underlying trigger (ie, the NLP algorithm), and iterative refinement of the system based on user feedback. While participants discussed alert fatigue, or provider desensitization owing to excessive alerts [57], as a potential barrier to future implementation, the NLP algorithm's relatively high specificity and the limited patient population may minimize this concern. However, continual improvement of any rule-based algorithm is critical to maintain its quality. While participants did not discuss the implications of receiving a CA-CDS alert after a patient's discharge if a provider completes the documentation after a patient's ED visit, encounters with injuries identified by the NLP algorithm undergo weekly routine case surveillance by the CPT [58]. This is especially important for cases in which documentation is completed after a patient's discharge to assure that the identified injuries are not concerning for missed abuse. Such a monitoring system may facilitate the identification of cases that might have been missed in real time during the ED encounter [14].

Patient Access to Electronic Health Information

An important but underexplored aspect of child abuse CDS systems is the impact of the 21st Century Cures Act on providers' EHR interactions. The Cures Act is a federal law that came into effect in April 2021, mandating the free, timely release of electronic health information to patients and their guardians unless the practice meets the condition of select exceptions, one of which is preventing harm in contexts such as child abuse [59-61]. Concurrently, adoption and use of the patient portal dramatically increased as a result of the COVID-19 pandemic, with much of patient care and communication shifting to electronic mediums. Given the increased ease of patient access to EHR content, it is especially important to understand how provider perspectives about documentation of suspected child abuse have been affected by the new law. This study was uniquely timed to explore these concerns following the Cures Act. Although users worried about the potential repercussions of using the CA-CDS to document suspicions about abuse in caregiver-accessible notes, their unease was alleviated by the clarity provided by the protocol regarding the destination of the automatic documentation into the note and the addition of a reminder to unshare the note if appropriate. Next steps include exploring caregivers' perspectives about the documentation of suspected child abuse.

Limitations

This study had at least 3 limitations. First, although we tried to recruit representative participants, few community ED providers (4/23, 17%) versus pediatric ED providers (19/23, 83%) participated in the usability testing. As providers who work at sites that often see most of a community's pediatric population and more often underdiagnose child abuse [6-8], feedback from community ED clinicians is uniquely valuable. Future system testing would benefit from having a more balanced or

XSL•FO
RenderX

community-focused participant pool. Second, we modified the CA-CDS based on the majority's preferences. As such, there may have been modifications desired by a notable percentage of our users that were not implemented, which may limit their interaction with the CA-CDS in the future. However, the high-risk injuries identified by the CA-CDS will be routinely reviewed by the CPT to assure that cases are not misdiagnosed. Third, while this system was developed by a local team of experts and through iterative usability testing with providers at different sites, the CA-CDS's recommended management may be institution specific. Further usability testing may be required if the system is disseminated to other hospitals, especially those with limited resources.

Conclusions

In summary, with its user-centered design and evidence-based content, the CA-CDS offers a novel method to aid ED medical and nursing providers in the real-time recognition, evaluation, and management of infant physical abuse. Our system has the potential to reduce the number of missed cases and increase the provision of less biased and evidence-based care to all infants.

Acknowledgments

This study was supported in part by funds from the National Institute of Child Health and Human Development (grant K23HD107178 [GT]). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data underlying this paper will be shared upon reasonable request to the corresponding author. The technical components used in this study were provided by 3M Company under license and will be shared upon request to the corresponding author with permission from 3M Company.

Authors' Contributions

AT, GT, and KL were involved in all steps of the study including conducting literature review, communicating with experts, developing the study material such as the interview guides and surveys, conducting interviews, analyzing the interview transcripts, and designing and refining the Child Abuse Clinical Decision Support (CA-CDS) prototypes. GT was also responsible for initial project conception, institutional review board submission, suggestion of candidates to participate in the study, and presentation of material at Yale New Haven Health stakeholder meetings. KL was also responsible for the technical components, designing and editing the prototype, and serving as our liaison with other 3M Company staff. AA and AH provided guidance and shared their clinical expertise in pediatric emergency medicine, child abuse, and health informatics as we developed and refined the CA-CDS. All authors collaborated to draft and review this paper.

Conflicts of Interest

KL serves as a consultant for 3M Health Information Systems Inc, 3M Company, and AH serves as a member of the Strategic Advisory Board for Johnson & Johnson. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary tables and figures including clinical vignettes, interview guide, demographic and System Usability Scale survey, and prototype modifications.
[DOCX File , 2506 KB - jmir_v26i1e51058_app1.docx]

References

1. Child maltreatment 2021. U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. 2023. URL: https://www.acf.hhs.gov/sites/default/files/documents/cb/cm2021.pdf [accessed 2024-03-17]
2. Jenny C, Hymel KP, Ritzen A, Reinert SE, Hay TC. Analysis of missed cases of abusive head trauma. JAMA 1999 Feb 17;281(7):621-626. [doi: 10.1001/jama.281.7.621] [Medline: 10029123]
3. Letson MM, Cooper JN, Deans KJ, Scribano PV, Makoroff KL, Feldman KW, et al. Prior opportunities to identify abuse in children with abusive head trauma. Child Abuse Negl 2016 Oct;60:36-45. [doi: 10.1016/j.chiabu.2016.09.001] [Medline: 27680755]
4. Thorpe EL, Zuckerbraun NS, Wolford JE, Berger RP. Missed opportunities to diagnose child physical abuse. Pediatr Emerg Care 2014 Nov;30(11):771-776. [doi: 10.1097/PEC.0000000000000257] [Medline: 25343739]
5. King WK, Kiesel EL, Simon HK. Child abuse fatalities: are we missing opportunities for intervention? Pediatr Emerg Care 2006 Apr;22(4):211-214. [doi: 10.1097/01.pec.0000208180.94166.dd] [Medline: 16651907]
6. Ravichandiran N, Schuh S, Bejuk M, Al-Harthy N, Shouldice M, Au H, et al. Delayed identification of pediatric abuse-related fractures. Pediatrics 2010 Jan;125(1):60-66. [doi: 10.1542/peds.2008-3794] [Medline: 19948569]

XSL•FO
RenderX

7. Trokel M, Waddimba A, Griffith J, Sege R. Variation in the diagnosis of child abuse in severely injured infants. Pediatrics 2006 Mar;117(3):722-728. [doi: 10.1542/peds.2004-2731] [Medline: 16510652]
8. Ziegler DS, Sammut J, Piper AC. Assessment and follow-up of suspected child abuse in preschool children with fractures seen in a general hospital emergency department. J Paediatr Child Health 2005;41(5-6):251-255. [doi: 10.1111/j.1440-1754.2005.00605.x] [Medline: 15953323]
9. Hymel KP, Laskey AL, Crowell KR, Wang M, Armijo-Garcia V, Frazier TN, et al. Racial and ethnic disparities and bias in the evaluation and reporting of abusive head trauma. J Pediatr 2018 Jul;198:137-43.e1 [FREE Full text] [doi: 10.1016/j.jpeds.2018.01.048] [Medline: 29606408]
10. Lane WG, Dubowitz H. What factors affect the identification and reporting of child abuse-related fractures? Clin Orthop Relat Res 2007 Aug;461:219-225. [doi: 10.1097/BLO.0b013e31805c0849] [Medline: 17415005]
11. Wood JN, Christian CW, Adams CM, Rubin DM. Skeletal surveys in infants with isolated skull fractures. Pediatrics 2009 Feb;123(2):e247-e252. [doi: 10.1542/peds.2008-2467] [Medline: 19171576]
12. Wood JN, Hall M, Schilling S, Keren R, Mitra N, Rubin DM. Disparities in the evaluation and diagnosis of abuse among infants with traumatic brain injury. Pediatrics 2010 Sep;126(3):408-414. [doi: 10.1542/peds.2010-0031] [Medline: 20713477]
13. Clinical decision support. HealthIT.gov.: Office of the National Coordinator for Health Information Technology URL: https://www.healthit.gov/topic/safety/clinical-decision-support [accessed 2021-12-23]
14. Suresh S, Barata I, Feldstein D, Heineman E, Lindberg DM, Bimber T, et al. Clinical decision support for child abuse: recommendations from a consensus conference. J Pediatr 2023 Jan;252:213-8.e5. [doi: 10.1016/j.jpeds.2022.06.039] [Medline: 35817134]
15. Low D. The Child Protection Information Sharing Project (CP-IS): electronic information sharing to improve the assessment of known vulnerable or at-risk children. Adopt Foster 2016;40(3):293-296. [doi: 10.1177/0308575916664626]
16. Child Protection - Information Sharing (CP-IS) service. National Health Service England.: National Health Service England URL: https://digital.nhs.uk/services/child-protection-information-sharing-project [accessed 2021-12-24]
17. Deutsch SA, Henry MK, Lin W, Valentine KJ, Valente C, Callahan JM, et al. Quality improvement initiative to improve abuse screening among infants with extremity fractures. Pediatr Emerg Care 2019 Sep;35(9):643-650. [doi: 10.1097/PEC.0000000000001671] [Medline: 30586037]
18. Erlanger AC, Heyman RE, Slep AM. Creating and testing the reliability of a family maltreatment severity classification system. J Interpers Violence 2022 Apr;37(7-8):NP5649-NP5668. [doi: 10.1177/0886260520961866] [Medline: 32990144]
19. Kelly P, Chan C, Reed P, Ritchie M. The national child protection alert system in New Zealand: a prospective multi-centre study of inter-rater agreement. Child Youth Serv Rev 2020 Sep;116:105174. [doi: 10.1016/j.childyouth.2020.105174]
20. Thraen IM, Frasier L, Cochella C, Yaffe J, Goede P. The use of TeleCAM as a remote web-based application for child maltreatment assessment, peer review, and case documentation. Child Maltreat 2008 Nov;13(4):368-376. [doi: 10.1177/1077559508318068] [Medline: 18483209]
21. Child-at-risk electronic medical record alert. Agency for Clinical Innovation, New South Wales Government.: New South Wales Government; 2017 Aug 25. URL: https://aci.health.nsw.gov.au/ie/projects/child-at-risk [accessed 2021-12-24]
22. Gonzalez DO, Deans KJ. Hospital-based screening tools in the identification of non-accidental trauma. Semin Pediatr Surg 2017 Feb;26(1):43-46. [doi: 10.1053/j.sempedsurg.2017.01.002] [Medline: 28302284]
23. Escobar MAJ, Pflugeisen BM, Duralde Y, Morris CJ, Haferbecker D, Amoroso PJ, et al. Development of a systematic protocol to identify victims of non-accidental trauma. Pediatr Surg Int 2016 Apr;32(4):377-386. [doi: 10.1007/s00383-016-3863-8] [Medline: 26809669]
24. Riney LC, Frey TM, Fain ET, Duma EM, Bennett BL, Murtagh Kurowski E. Standardizing the evaluation of nonaccidental trauma in a large pediatric emergency department. Pediatrics 2018 Jan;141(1):e20171994. [doi: 10.1542/peds.2017-1994] [Medline: 29212880]
25. Luo S, Botash AS. Designing and developing a mobile app for clinical decision support: an interprofessional collaboration. Comput Inform Nurs 2018 Oct;36(10):467-472. [doi: 10.1097/CIN.0000000000000487] [Medline: 30300186]
26. Berger RP, Saladino RA, Fromkin J, Heineman E, Suresh S, McGinn T. Development of an electronic medical record-based child physical abuse alert system. J Am Med Inform Assoc 2018 Feb 01;25(2):142-149 [FREE Full text] [doi: 10.1093/jamia/ocx063] [Medline: 28641385]
27. Suresh S, Heineman E, Meyer L, Richichi R, Conger S, Young S, et al. Improved detection of child maltreatment with routine screening in a tertiary care pediatric hospital. J Pediatr 2022 Apr;243:181-7.e2. [doi: 10.1016/j.jpeds.2021.11.073] [Medline: 34929244]
28. Suresh S, Saladino RA, Fromkin J, Heineman E, McGinn T, Richichi R, et al. Integration of physical abuse clinical decision support into the electronic health record at a Tertiary Care Children's Hospital. J Am Med Inform Assoc 2018 Jul 01;25(7):833-840 [FREE Full text] [doi: 10.1093/jamia/ocy025] [Medline: 29659856]
29. Rumball-Smith J, Fromkin J, Rosenthal B, Shane D, Skrbin J, Bimber T, et al. Implementation of routine electronic health record-based child abuse screening in General Emergency Departments. Child Abuse Negl 2018 Nov;85:58-67. [doi: 10.1016/j.chiabu.2018.08.008] [Medline: 30170921]

XSL•FO
RenderX

30. Rosenthal B, Skrbin J, Fromkin J, Heineman E, McGinn T, Richichi R, et al. Integration of physical abuse clinical decision support at 2 general emergency departments. J Am Med Inform Assoc 2019 Oct 01;26(10):1020-1029 [FREE Full text] [doi: 10.1093/jamia/ocz069] [Medline: 31197358]
31. McGinn T, Feldstein DA, Barata I, Heineman E, Ross J, Kaplan D, et al. Dissemination of child abuse clinical decision support: moving beyond a single electronic health record. Int J Med Inform 2021 Mar;147:104349 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104349] [Medline: 33360791]
32. Luo S, Botash AS. Testing a mobile app for child abuse treatment: a mixed methods study. Int J Nurs Sci 2020 Jun 23;7(3):320-329 [FREE Full text] [doi: 10.1016/j.ijnss.2020.06.008] [Medline: 32817855]
33. Tiyyagura G, Asnes AG, Leventhal JM, Shapiro ED, Auerbach M, Teng W, et al. Development and validation of a natural language processing tool to identify injuries in infants associated with abuse. Acad Pediatr 2022 Aug;22(6):981-988 [FREE Full text] [doi: 10.1016/j.acap.2021.11.004] [Medline: 34780997]
34. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004 Feb;37(1):56-76 [FREE Full text] [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
35. Usability. Interaction Design Foundation. URL: https://www.interaction-design.org/literature/topics/usability [accessed 2021-12-26]
36. Mann DM, Chokshi SK, Kushniruk A. Bridging the gap between academic research and pragmatic needs in usability: a hybrid approach to usability evaluation of health care information systems. JMIR Hum Factors 2018 Nov 28;5(4):e10721 [FREE Full text] [doi: 10.2196/10721] [Medline: 30487119]
37. Van de Velde S, Kunnamo I, Roshanov P, Kortteisto T, Aertgeerts B, Vandvik PO, et al. The GUIDES checklist: development of a tool to improve the successful use of guideline-based computerised clinical decision support. Implement Sci 2018 Jun 25;13(1):86 [FREE Full text] [doi: 10.1186/s13012-018-0772-3] [Medline: 29941007]
38. Shum M, Asnes A, Leventhal JM, Bechtel K, Gaither JR, Tiyyagura G. The use of experts to evaluate a child abuse guideline in community emergency departments. Acad Pediatr 2021 Apr;21(3):521-528. [doi: 10.1016/j.acap.2020.11.001] [Medline: 33160081]
39. Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. Med Educ 2017 Jan;51(1):40-50. [doi: 10.1111/medu.13124] [Medline: 27981658]
40. LaDonna KA, Artino ARJ, Balmer DF. Beyond the guise of saturation: rigor and qualitative interview data. J Grad Med Educ 2021 Oct;13(5):607-611 [FREE Full text] [doi: 10.4300/JGME-D-21-00752.1] [Medline: 34721785]
41. Hanson JL, Balmer DF, Giardino AP. Qualitative research methods for medical educators. Acad Pediatr 2011;11(5):375-386. [doi: 10.1016/j.acap.2011.05.001] [Medline: 21783450]
42. Johnson CM, Johnston D, Crowley PK, Culbertson H, Rippen HE, Damico DJ, et al. EHR usability toolkit: a background report on usability and electronic health records. Agency for Healthcare Research and Quality.: U.S. Department of Health and Human Services; 2011 Aug. URL: https://digital.ahrq.gov/sites/default/files/docs/citation/EHR_Usability_Toolkit_Background_Report.pdf [accessed 2024-03-11]
43. Brooke J. SUS: a retrospective. J User Exp 2013;8(2):29-40 [FREE Full text]
44. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
45. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
46. Lewis JR, Sauro J. Item benchmarks for the system usability scale. J Usab Stud 2018 May 1;13(3):158-167 [FREE Full text]
47. Sauro J. A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Denver, CO: Measuring Usability LLC; 2011.
48. Powers E, Tiyyagura G, Asnes AG, Leventhal JM, Moles R, Christison-Lagay E, et al. Early involvement of the child protection team in the care of injured infants in a pediatric emergency department. J Emerg Med 2019 Jun;56(6):592-600. [doi: 10.1016/j.jemermed.2019.01.030] [Medline: 30879856]
49. Tiyyagura G, Emerson B, Gaither JR, Bechtel K, Leventhal JM, Becker H, et al. Child protection team consultation for injuries potentially due to child abuse in community emergency departments. Acad Emerg Med 2021 Jan;28(1):70-81 [FREE Full text] [doi: 10.1111/acem.14132] [Medline: 32931628]
50. What is FHIR®? HealthIT.gov.: The Office of the National Coordinator for Health Information Technology; 2019. URL: https://www.healthit.gov/sites/default/files/2019-08/ONCFHIRFSWhatIsFHIR.pdf [accessed 2024-03-11]
51. Yan E. Catching FHIR: healthcare interoperability in 2023. Keysight Technologies. 2023 Jan 16. URL: https://www.keysight.com/blogs/en/tech/software-testing/2023/1/16/fhir-healthcare-interoperability-testing [accessed 2024-03-11]
52. The FHIR® API. HealthIT.gov.: The Office of the National Coordinator for Health Information Technology URL: https://www.healthit.gov/sites/default/files/page/2021-04/FHIR%20API%20Fact%20Sheet.pdf [accessed 2024-03-11]
53. Vorisek CN, Lehne M, Klopfenstein SA, Mayer PJ, Bartschke A, Haese T, et al. Fast healthcare interoperability resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]

XSL•FO
RenderX

54. Barker W, Posnack S. The heat is on: US caught FHIR in 2019. Health IT Buzz.: Office of the National Coordinator for Health Information Technology; 2021 Jul 29. URL: https://www.healthit.gov/buzz-blog/health-it/the-heat-is-on-us-caught-fhir-in-2019 [accessed 2024-03-11]
55. 3M™ M*Modal fluency direct. 3M Company. URL: https://www.3m.com/3M/en_US/health-information-systems-us/create-time-to-care/clinician-solutions/speech-recognition/fluency-direct/ [accessed 2022-01-17]
56. Rangel EL, Cook BS, Bennett BL, Shebesta K, Ying J, Falcone RA. Eliminating disparity in evaluation for abuse in infants with head injury: use of a screening guideline. J Pediatr Surg 2009 Jun;44(6):1229-1235. [doi: 10.1016/j.jpedsurg.2009.02.044] [Medline: 19524746]
57. Alert fatigue. Patient Safety Network.: Agency for Healthcare Research and Quality; 2019 Sep 07. URL: https://tinyurl.com/467dssaw [accessed 2024-03-10]
58. Shum M, Hsiao A, Teng W, Asnes A, Amrhein J, Tiyyagura G. Natural language processing - a surveillance stepping stone to identify child abuse. Acad Pediatr 2024;24(1):92-96. [doi: 10.1016/j.acap.2023.08.015] [Medline: 37652162]
59. 21st century cures act. U.S. Food & Drug Administration. 2020 Jan 31. URL: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act [accessed 2022-01-17]
60. ONC's Cures Act Final Rule. HealthIT.gov.: The Office of the National Coordinator for Health Information Technology URL: https://www.healthit.gov/topic/oncs-cures-act-final-rule [accessed 2022-01-17]
61. Anthony ES. The cures act final rule: interoperability-focused policies that empower patients and support providers. Health IT Buzz.: The Office of the National Coordinator for Health Information Technology; 2020 Mar 09. URL: https://www.healthit.gov/buzz-blog/21st-century-cures-act/the-cures-final-rule [accessed 2022-01-16]

Abbreviations

CA-CDS: Child Abuse Clinical Decision Support
CDS: clinical decision support
CPS: Child Protective Services
CPT: child protection team
DART: Detection, Assessment, Referral, and Treatment
DCF: Department of Children and Families
ED: emergency department
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
GUIDES: Guideline Implementation with Decision Support
MORE: Mechanism, Others present, Review of development, and Examination details
NLP: natural language processing
SUS: System Usability Scale
SW: social worker

Edited by G Tsafnat; submitted 20.07.23; peer-reviewed by E Heiman, D Listman; comments to author 08.02.24; revised version received 23.02.24; accepted 23.02.24; published 29.03.24.

Please cite as:
Thomas A, Asnes A, Libby K, Hsiao A, Tiyyagura G
Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study
J Med Internet Res 2024;26:e51058
URL: https://www.jmir.org/2024/1/e51058
doi:10.2196/51058
PMID:38551639

©Amy Thomas, Andrea Asnes, Kyle Libby, Allen Hsiao, Gunjan Tiyyagura. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Differences in Physicians' Ratings of Work Stressors and Resources Associated With Digital Transformation: Cross-Sectional Study

Magdalena Katharina Wekenborg[1,2], PhD; Katharina Förster[3], PhD; Florian Schweden[4], PhD; Robin Weidemann[5], PhD; Felix von Bechtolsheim[6,7], PhD; Clemens Kirschbaum[1], PhD; Jürgen Weitz[6,7], PhD; Beate Ditzen[8], PhD

[1]Chair of Biopsychology, Faculty of Psychology, TU Dresden, Dresden, Germany

[2]Else Kröner Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany

[3]Clinical Psychology and Behavioural Neuroscience, Faculty of Psychology, TU Dresden, Dresden, Germany

[4]Institute for Work Design and Organizational Development INAGO, Hamburg, Germany

[5]Clinic of Internal Medicine, Krankenhaus St. Joseph-Stift, Dresden, Germany

[6]Department of Gastrointestinal, Thoracic, and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany

[7]Center for Tactile Internet with Human-in-the-Loop (CeTI), TU Dresden, Dresden, Germany

[8]Institute of Medical Psychology, Heidelberg University, Heidelberg, Germany

Corresponding Author:
Magdalena Katharina Wekenborg, PhD
Else Kröner Fresenius Center for Digital Health
Faculty of Medicine and University Hospital Carl Gustav Carus
TU Dresden
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 0351 45819581
Email: magdalena.wekenborg@tu-dresden.de

Abstract

Background: The emergence of the COVID-19 pandemic rapidly accelerated the need and implementation of digital innovations, especially in medicine.

Objective: To gain a better understanding of the stress associated with digital transformation in physicians, this study aims to identify working conditions that are stress relevant for physicians and differ in dependence on digital transformation. In addition, we examined the potential role of individual characteristics (ie, age, gender, and actual implementation of a digital innovation within the last 3 years) in digitalization-associated differences in these working conditions.

Methods: Cross-sectional web-based questionnaire data of 268 physicians (mean age 40.9, SD 12.3 y; n=150, 56% women) in Germany were analyzed. Physicians rated their chronic stress level and 11 relevant working conditions (ie, work stressors such as time pressure and work resources such as influence on sequence) both before and after either a fictional or real implementation of a relevant digital transformation at their workplace. In addition, a subsample of individuals (60; n=33, 55% women) submitted self-collected hair samples for cortisol analysis.

Results: The stress relevance of the selected working conditions was confirmed by significant correlations with self-rated chronic stress and hair cortisol levels (hair F) within the sample, all of them in the expected direction (*P* values between .01 and <.001). Multilevel modeling revealed significant differences associated with digital transformation in the rating of 8 (73%) out of 11 working conditions. More precisely, digital transformation was associated with potentially stress-enhancing effects in 6 working conditions (ie, *influence on procedures* and *complexity of tasks*) and stress-reducing effects in 2 other working conditions (ie, *perceived workload* and *time pressure*). Younger individuals, women, and individuals whose workplaces have implemented digital innovations tended to perceive digitalization-related differences in working conditions as rather stress-reducing.

XSL•FO
RenderX

Conclusions: Our study lays the foundation for future hypothesis-based longitudinal research by identifying those working conditions that are stress relevant for physicians and prone to differ as a function of digital transformation and individual characteristics.

(J Med Internet Res 2024;26:e49581) doi:10.2196/49581

KEYWORDS

physicians; digital transformation; chronic stress; hair cortisol concentration; work stressors; work resources

Introduction

Background

The COVID-19 pandemic convincingly demonstrated the urgent need for employable physicians for the functionality of many aspects of society. Given this crucial role of physicians in society, the widespread prevalence of chronic stress among them [1] is alarming. Chronic stress is experienced when perceived resources are enduringly outweighed by demands [2]. Keeping in mind that chronic stress and its sequelae (ie, burnout symptoms) have been associated with risks for patients (eg, heightened risk for medical errors [3]), as well as a significant increase in rates of sickness absence and incapacity for work [4], chronic stress poses a serious threat to the health and employability of physicians [5].

In theory, the increasing digital transformation of the health system could help to reduce the chronic stress of physicians because the potential of improvement of working conditions of health care professionals by IT has been emphasized repeatedly [6]. Thereby, digital transformation defined as an automatization of tasks [7] can be distinguished from digitization (ie, technical process of converting analog signals into a digital form [7,8]) and digitalization (ie, the process of adopting and using this technology in broader contexts [9]). As described in a review by Topol [10], the importance of top digital health care technologies, namely, digital medicine (eg, telemedicine), genomics (eg, reading the genome), artificial intelligence (AI), and robotics (eg, automated image interpretation using AI), for the work of physicians will increase significantly within the next decades.

Indeed, theoretically, digital transformation offers a variety of ways to reduce workplace stress for physicians, including simplifying time-consuming bureaucracy and relieving certain tasks and responsibilities through digital decision-making tools [6]. Empirically, there is evidence for the stress-reducing effects of digital innovations, for example, with respect to robot-assisted surgery on surgeons' stress load [11] and mental effort and workload [12].

However, the few existing empirical studies do not consistently support the stress-buffering effects of digital transformation on perceived stress in physicians. Particularly, electronic health records (EHRs) and comparable information systems have been implicated as a factor that might enhance physicians' chronic stress, both in cross-sectional [13-15] and longitudinal studies [16].

Although EHRs receive significant attention, research indicates that increased use of other digital technologies, such as new surgical technologies in operating rooms [17] and telemedicine [18], may also contribute to chronic stress among physicians. These findings implicate that very different types of digital transformation can have stress-enhancing effects for the user.

One of the most established models to explain the development of stress as a consequence of digital transformation is the technostress model, which was originally introduced by Brod [19]. On the basis of the transactional model of stress and health [2], Brod [19] defined technostress as the result of an inadequate ability to cope with the requirements of the use of computer technology. More precisely, he defined the following 5 components crucially related to the development of technostress, namely, techno-invasion (employee can be contacted at any time), techno-overload (technology forces one to work harder), techno-complexity (complexity requires learning efforts), techno-insecurity (one is afraid of losing one's job because of technology), and techno-uncertainty (continuous changes requires constant relearning).

The technostress model by Brod [19] has undoubtedly made a decisive contribution to the description and study of the phenomenon. However, its further development is highly relevant for a better understanding of the process by which digital transformation affects stress. Although there have been important expansions to the original model (eg, by Gimpel et al [20] and Ragu-Nathan et al [21]), empirical evidence on which particular stress-relevant working conditions mediate the association between digital transformation and enhanced stress is still lacking. Identifying the working conditions that contribute to physician stress and are susceptible to change would be highly relevant for monitoring and designing health care digitalization.

When it comes to the selection of these stress-relevant working conditions, different theoretical conceptualizations and empirical evidence exist [22-26]. Thereby, the working conditions identified by Rau and Buyken [27] within a meta-analysis appear to be especially suited with regard to the aim of this study, as they have been shown to be relevant in *physicians* [27,28]. The working conditions identified by this meta-analysis can be roughly subdivided into *work resources* potentially reducing stress load at work (eg, learning new skills and job control) and *work stressors* potentially enhancing stress load at work (eg, time pressure).

Objectives

To complement existing models on technostress in physicians and enabling future hypothesis-based research, this study set out to examine (1) which of those working conditions identified by Rau and Buyken [27] are stress relevant for physicians and (2) which differ as a function of digital transformation. Thereby, in contrast to the prevailing approach in prior research, our study acknowledges stress as a biopsychological phenomenon. We

XSL•FO
RenderX

do this by operationalizing stress using both self-report measures as well as a biological marker of stress, namely, hair cortisol concentrations (hair F). Shortly summarized, hair F has been shown to be a valid and solid index of the functional status of one of the central stress pathways of the body, the hypothalamus-pituitary-adrenal (HPA) axis, with high intraindividual stability and test-retest reliability, which, in contrast to other forms of cortisol quantification, represents an aggregated measure of chronic stress over a period of months [29].

To answer these research questions, the following hypotheses are tested:

- *H1*: The selected working conditions are associated with psychological and physiological markers of chronic stress in physicians. More precisely, *work resources* (ie, *influence on sequence of activities; influence on workload and procedures; learning new skills; use of knowledge, skills, and abilities; visibility of task accomplishment*; and *consideration of employee input*) are associated with reduced levels of chronic stress in physicians, and *work stressors* (ie, *workload, time pressure, excessive complexity of tasks, excessive demands on concentration*, and *interruptions of workflow*) are associated with enhanced levels of chronic stress in physicians..
- *H2*: The ratings of the selected working conditions carried out by physicians differ as a function of digital transformation
  - *H2.1*: Age and gender influence differences in these ratings.
  - *H2.2*: Finally, we were interested in how actual experiences with digital transformation at the workplace might moderate differences in the ratings of these working conditions before and after digital transformation.

Methods

Recruitment

Physicians were recruited throughout Germany via social networks, web-based platforms, and medical associations with a special focus on the University Hospital Dresden. Practicing physicians from all disciplines were included in the study.

A total of 437 participants started the web-based questionnaire. Of the 437 participants, 271 (62%) completed the questionnaire. Of these 271 participants, 3 (1.1%) defined their gender as "diverse." They were not included in the analyses of the core sample because gender was included as a control variable in all subsequent analyses, and this number is too small for statistical processing, leaving a final sample of 268 participants (core sample: mean age 40.79, SD 12.3 y; n=150, 56% women). Hair samples were available from 24% (60/268) of these participants (subsample). Sample characteristics of the core sample as well as the subsample that provided hair samples are presented in Table 1.

Table 1. Characteristics of the core sample and the subsample.

Characteristic	Core sample (N=268)	Subsample (n=60)
Age (y), mean (SD)	40.8 (12.3)	40.5 (12.1)
Work experience (y), mean (SD)	13.98 (12.4)	13.4 (12.3)
Gender (women), n (%)	150 (56)	33 (55)
Specialty[a], n (%)		
Internal medicine	79 (29.5)	16 (27)
General medicine	23 (8.6)	3 (5)
Anesthesiology	21 (7.8)	6 (10)
Surgery	20 (7.5)	3 (5)
Psychiatry	17 (6.3)	7 (12)
Pediatrics	15 (5.6)	6 (10)
PSS-4[b], mean (SD)	6.3 (2.9)	6.1 (3)
Hair F[c], mean (SD)	6.3 (5)	6.3 (5)

[a]Most frequently mentioned specialties.

[b]PSS-4: Perceived Stress Scale 4.

[c]Hair F: hair cortisol concentration.

The most commonly reported medical specialty was by far internal medicine, followed by general medicine, anesthesiology, surgery, psychiatry, and pediatrics.

Overall, 57.1% (153/268) of the participants reported that a digital innovation was implemented at their workplace within the last 3 years, whereas 21.3% (57/268) of the individuals referred to a digital transformation that would be presumably implemented within the upcoming 3 years, and 21.6% (58/268) of the participants used our provided fictional example. Facing the large variety of different digital health care technologies mentioned by our study participants, we used three categories introduced by Topol [30] to categorize them: (1) *digital medicine* (ie, digital products and services that are intended for use in

XSL•FO
RenderX

diagnosis, prevention, monitoring, and treatment such as EHRs and wearables), (2) *robotics* (ie, construction, operation, and application of intelligent machines, eg, robot-assisted surgery), and (3) *AI* (ie, methods that can be used to analyze, interpret, and make predictions using these data source, for example, AI-based diagnostics). Figure 1 presents the participant flow.

Figure 1. Participant flow through the study.

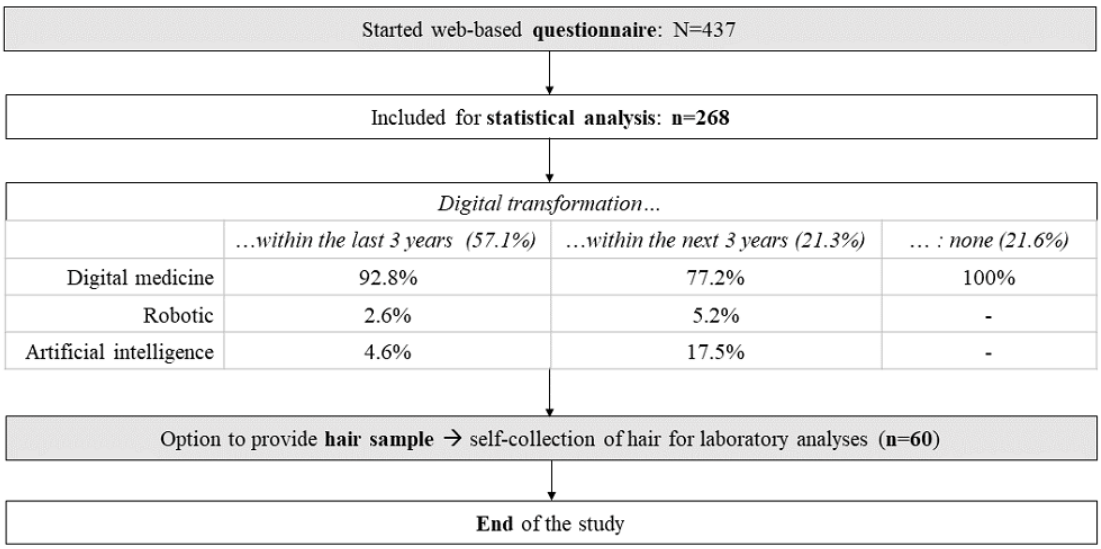

Measures

This study had a cross-sectional, quantitative research design. The main part of the study was web-based questionnaire that was administered via LimeSurvey, encompassing questions regarding stress-relevant working conditions, sociodemographic factors (age, gender [men, women, nonbinary]), occupational characteristics (eg, specialization and years of work experiences), the actual implementation of a digital innovation at the workplace within the last 3 years (yes or no), and chronic stress. Sociodemographic factors and occupational characteristics were assessed using self-generated items. In addition, participants could send in self-collected hair samples as a biological marker of chronic stress.

Stress-relevant working conditions identified by Rau and Buyken [27] were assessed with the Short Questionnaire for Workplace Analysis (KFZA [31]). The KFZA seems especially suited as it fulfills the recommendations of the Joint German Occupational Safety and Health Strategy [32] for psychosocial workplace risk assessment and has been empirically proven to be generally applicable in hospital settings [33-36]. The KFZA originally consists of 26 items.

To take into account its applicability in a sample that is characterized by constant presence of time pressure, we shortened the KFZA to 11 items to enhance its applicability in physicians. Selection of items was based on the findings of Appel et al [33] as well as consensus of several experts. Within the original KFZA, the 26 items can be subsumed under 4 aspects of work. For each of these work aspects ≥1 items of the KFZA were included within this study: (1) “Job Content” (items in this study: *learning new skills*; *use of knowledge, skills, and abilities*; and *visibility of task accomplishment*), (2) “Resources” (items in this study: *influence on sequence of work activities and influence on workload and procedures*), (3) “Stressors” (items in this study: *workload*, *time pressure*,*excessive complexity of tasks*,*excessive demands on concentration*, and *interruptions to workflow*), and (4) “Organizational Culture” (item in this study: *consideration of employee input* [adjusted by referring to “clinic management” instead of originally referring to company management]). Items that are summarized under the work aspects 1, 2, and 4 have been theoretically and empirically associated with reduced work stress experiences (referred to as *work resources* in the following sections), whereas items that are summarized under work aspect 3 have been associated with enhanced stress experiences (referred to as *work stressors* in the following sections) [31,33]. Participants rated the extent to which they agree with each of these items on a 5-pont Likert scale, ranging from 1 (fully disagree) to 5 (fully agree). The internal consistencies (ie, Cronbach α) of the work aspects were calculated based on the current rating of the respective KFZA items (ie, preimplementation ratings of individuals without the actual implementation of a digital innovation and postimplementation ratings of individuals with the actual implementation of a digital innovation at one’s workplace) and ranged from acceptable to very good (work aspect 1: α=0.71, work aspect 2: α=0.89, and work aspect 3: α=0.77). No Cronbach α could be calculated for “Organizational Culture,” as this work factor was operationalized using only 1 item. As recommended by the authors of the KFZA, no sum scores over the different work aspects were calculated [31]. Instead, all items were analyzed separately.

Chronic self-reported stress was assessed using the German version [37] of the short form of the Perceived Stress Scale 4 (PSS-4 [38]). Respondents answered the 4 items on a 5-point Likert scale, ranging from 0 (never) to 4 (very often). Two items were reversed coded; thus, they were recoded before a total score was calculated, with higher values reflecting more stress (range 0 to 16). The internal consistency of the PSS-4 was, in accordance with the original publication [39], acceptable (Cronbach α=0.66).

XSL•FO
RenderX

Psychobiological indicators of stress were assessed via levels of the central hormone of the HPA axis, namely, cortisol, in the scalp hair. Cortisol mediates a number of biological, cognitive, and behavioral stress responses such as enhancing metabolic actions that increase energy level. These responses are necessary for adequately dealing with stressful situations and inhibiting stress-irrelevant body functions (eg, digestion [40]), making accumulated cortisol level a valid marker of HPA axis activity and therefore the chronic stress level of an individual [41]. In contrast to traditional measures of cortisol (blood, saliva, or urinary samples), the assessment of relevant HPA axis hormones in scalp hair retrospectively reflects an integrated secretion over several months [41], making hair F a valid and reliable biological marker of chronic stress [42]. Hair F concentration was determined from the 3 cm segment most proximal to the scalp. Given an average hair growth of 1 cm per month [43], this segment represents the cumulated cortisol secretion over a 3-month prior sampling episode. In the laboratory, handling and extraction of hair F were conducted in accordance with the laboratory protocol by Gao et al [44]. All samples were analyzed by liquid chromatography coupled with tandem mass spectrometry. The lower detection limit of the liquid chromatography coupled with tandem mass spectrometry protocol was 0.3 pg/mg for cortisol. All samples were processed in a single batch. The intraassay coefficients of variance was 8.2%.

Ethical Considerations

The study was designed according to the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975 (revised in 2008) and approved by the ethics committee of Technische Universität Dresden (identifier EK222052019). All participants provided informed consent. Data were pseudonymized during data collection using individually generated codes and anonymized after the data collection period.

Procedure

After opening the link to the web-based questionnaire, participants were presented information about the study, and the participants provided informed consent. They were then asked about their medical specialty.

Next, participants read a short text, which explained the term digital transformation in medicine and included some examples. They were then asked to name and briefly describe a digital innovation at their workplace, which had either been implemented within the last 3 years or would be presumably implemented within the coming 3 years and which affected them personally and significantly altered their work. If no digital transformation matching these criteria existed, they were provided with a short description of the EHR, as EHRs are particularly widespread in all medical specialties. As depicted in Figure 1, most participants described digital health care technologies from the category of digital medicine (within the last 3 years: n=142, 53% and within the next 3 years: n=44, 16.4%); AI was the second most common category (within the last 3 years: n=7, 2.6% and within the next 3 years: n=10, 3.7%), ahead of robotics (within the last 3 years: n=4, 1.5% and within the next 3 years: n=3, 1.1%).

Thereafter, participants were asked to rate each of the KFZA items, regarding the situation both before and after the introduction of the respective digital transformation (exact wording: “Please rate how strongly the following statements apply to your work. As you do so, take turns imagining that the digital innovation you have just described does or does not affect you.”). In a final part of the web-based questionnaire, participants responded to items assessing perceived chronic psychological stress, sociodemographic variables, and further constructs, which are not part of this study.

Afterward, participants could receive information on how to self-collect hair samples, store them, and send them to the laboratory via mail. Previous research indicated that the self-collection of hair in domestic settings is a viable and economical method for measuring long-term steroid concentration in hair [45]. All individuals with hair >3 cm were invited to send in hair samples and were instructed to cut hair strains as close as possible to the scalp from the posterior vertex position.

As compensation, all participants were eligible to enter a lottery for 5 activity vouchers, each valued at €200 (US $215), and received personalized feedback on their chronic stress levels.

Data Processing and Statistical Analyses

Hair F turned out to be positively skewed and was transformed on the natural logarithm scale to approach a normal distribution. For descriptive purposes, hair data in text and tables are reported in original units (pg/mg); however, for statistical analyses, log-transformed hair data were used.

First, we calculated descriptive statistics of study characteristics. Second, we examined potential differences in the chronic stress level between physicians from the most frequently mentioned medical specialties using an analysis of covariance (ANCOVA), with the PSS-4 as dependent and the medical specialty as independent variable, while adjusting for age and gender. We chose an ANCOVA as it allowed adjustment for the covariates gender and age. Due to the small number of hair samples per medical specialty, we abstained from calculating a respective ANCOVA for hair F. Third, we tested our first hypothesis that our selected KFZA items were stress relevant for physicians by investigating their associations with biopsychological markers of chronic stress. More precisely, we used nonparametrical partial correlation analyses controlling for age and gender to examine the association between the PSS-4, hair F, and the current rating of the respective KFZA items independent of any digital transformation. We opted for nonparametric partial correlation analyses to account for the different scale levels of the included variables. The ANCOVA and the nonparametrical partial correlation analyses were conducted using SPSS Statistics (version 28; IBM Corp).

Fourth, we tested our research hypotheses 2, 2.1, and 2.2 with multilevel linear mixed effects models nested within the person using the *nlme* function implemented in R (R Foundation for Statistical Computing) [46]. In all these models, time was included as within-subject variable, and age, gender, and the implementation of a digital innovation at one’s workplace within the last 3 years were included as level 2 predictors (model 1).

XSL•FO
RenderX

To test our hypothesis 2 of differences in the rating of stress-relevant working conditions associated with digital transformation, separate multilevel linear mixed effects models were calculated for each KFZA item (dependent variable) to account for high collinearity (model 1). To test our hypothesis 2.1 on the influence of demographical variables on differences in the rating of stress-relevant working conditions associated with digital transformation, interaction terms for age and time as well as gender and time were added to model 1 (model 2). Here too, separate multilevel linear mixed effects models were calculated with each of the KFZA items serving as dependent variable. Hypothesis 2.2 on the influence of actual experiences with digital innovations at one's workplace on differences in the rating of stress-relevant working conditions associated with digital transformation was tested by including an interaction term for this variable and time in model 2. We decided to use multilevel linear mixed effects models for testing our research hypotheses 2, 2.1, and 2.2 mainly because they provide a powerful method for examining complex data structures, which in our case enabled to differentiate between trait-like (intercepts) and digitalization-associated state-like (slopes) individual ratings of the KFZA items. In addition, multilevel linear mixed effects models allow for the consideration of covariates, and they provide better handling of missing values than most other common statistical methods.

For all analyses, we used the standard $P<.05$ criteria for determining if the results are significantly different from those expected if the null hypothesis were correct.

Results

Descriptive Results

In terms of chronic stress levels, these medical specialties significantly differed ($F_{5,167}=2.43$; $P=.04$; $\eta^2=0.068$), with highest perceived stress levels (PSS-4) in internal medicine (mean 7.03, SD 2.87) and the lowest in anesthesiology (mean 5.24, SD 3.14).

An overview of the before and after ratings of all stress-relevant working conditions is given in Figure 2 [31]. Following the specifications made by Prümper et al [31], mean ratings >3 indicate a high perceived presence and mean ratings <3 indicate a low perceived presence of the respective working condition. Applying these criteria, the perceived working environment of the physicians in this sample is generally characterized by a relatively high amount of *interruption of their workflow*; a relatively high *perceived workload*; a relatively high *perceived visibility of task accomplishment*; and a relatively high perceived *use of knowledge, skills, and abilities*. In contrast, the physicians in the present sample perceived the *possibilities for their participation* in modification processes as rather low.

Figure 2. Mean ratings of stress-relevant working conditions operationalized using the Short Questionnaire for Workplace Analysis in the core sample. Error bars indicate SEs. *Indicates significant differences in the ratings before the digital transformation compared to after with $P<.05$ (for exact P values, refer to Tables 3 and 4).

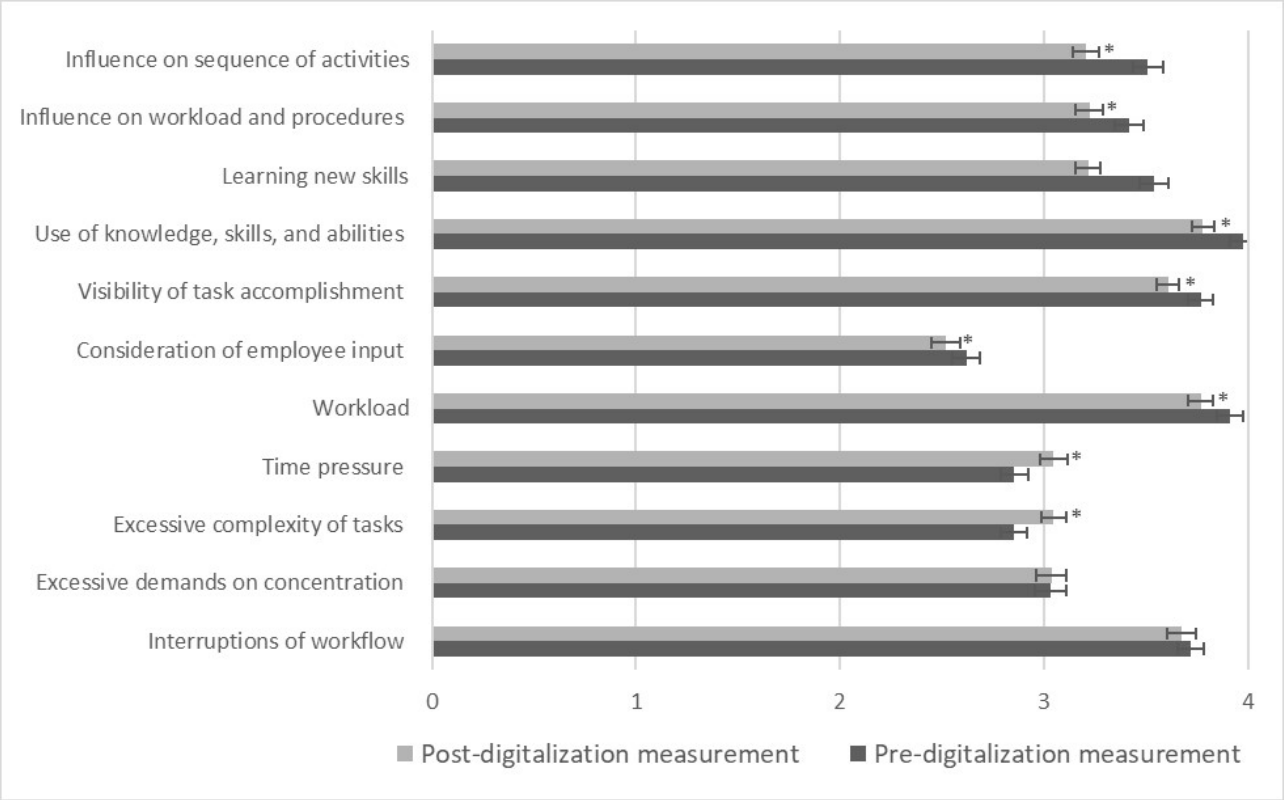

XSL•FO
RenderX

Associations Between Chronic Stress and the Selected Working Conditions (H1)

To test our first hypothesis regarding the stress relevance of the selected working conditions, associations between the current rating of the KFZA items and chronic self-reported stress as well as hair F were evaluated. We operationalized the current ratings of KFZA items by comparing preimplementation ratings of individuals without the digital innovation and postimplementation ratings of individuals with the innovation in their workplace. These ratings reflect their current work situation. As expected, *work stressors* were positively associated with biopsychological markers of chronic stress, and *work resources* were negatively associated with biopsychological markers of chronic stress (Table 2). Thereby, chronic self-reported stress, operationalized using the PSS-4, was significantly associated with all working conditions, except for *visibility of task accomplishment*, which, however, only narrowly missed statistical significance (P=.05). Hair F depicted significant positive associations only with *excessive demands on concentration* and *consideration of employee input*. H1 can therefore be retained.

XSL•FO
RenderX

Table 2. Partial nonparametric correlations between markers of chronic stress and working conditions adjusting for age and gender.

Variable	PSS-4[a] (N=268)	Hair F[b] (n=60)
Learning new skills		
r	−0.284	−0.113
P value	<.001	.40
Use of knowledge, skills, and abilities		
r	−0.193	−0.104
P value	.002	.44
Visibility of task accomplishment		
r	−0.117	−0.078
P value	.06	.56
Influence on sequence of work activities		
r	−0.206	0.076
P value	.001	.57
Influence on workload and procedures		
r	−0.222	0.095
P value	<.001	.48
Workload		
r	0.338	0.066
P value	<.001	.62
Time pressure		
r	0.261	0.029
P value	<.001	.83
Excessive complexity of tasks		
r	0.275	0.08
P value	<.001	.55
Excessive demands on concentration		
r	0.268	0.333
P value	<.001	.01
Interruptions to workflow		
r	0.241	0.134
P value	<.001	.32
Consideration of employee input		
r	−0.163	0.223
P value	.008	.09

[a]PSS-4: Perceived Stress Scale 4.

[b]Hair F: hair cortisol concentration.

Digitalization-Associated Differences in the Perception of Working Conditions (H2)

Our hypothesis 2 was that digital transformation would be associated with differences in the perception of stress-relevant working conditions, operationalized using the KFZA. Separate mixed linear effects model for each of the selected KFZA items were conducted. Results are depicted in Table 3 (*work resources*) and Table 4 (*work stressors*).

XSL•FO
RenderX

Table 3. Results from multilevel analysis on work resources (dependent variable) before and after digital transformations (within-subject variable) considering age, gender, and the implementation of a digital innovation at one's workplace within the last 3 years.

	Stress-reducing working conditions (dependent variables), β coefficients (SE)					
	Influence on sequence of activities	Influence on workload and procedures	Learning new skills	Use of knowledge, skills, and abilities	Visibility of task accomplishment	Consideration of employee input
Model 1						
Intercept	2.878 (0.244)**	2.756 (0.252)**	3.483 (0.244)**	3.487 (0.214)**	3.383 (0.205)**	3.051 (0.268)**
Age	0.012 (0.005)*	0.012 (0.005)*	–0.004 (0.005)	0.010 (0.004)*	0.010 (0.004)*	–0.009 (0.006)
Gender	0.086 (0.123)	0.146 (0.127)	0.152 (0.123)	–0.030 (0.108)	0.033 (0.103)	–0.254 (0.135)
Impl dig[a]	0.184 (0.120)	0.132 (0.124)	0.217 (0.120)	0.202 (0.105)	–0.060 (0.101)	0.170 (0.132)
Within-subject variable						
Time	–0.302 (0.065)**	–0.194 (0.060)**	–0.093 (0.056)	–0.198 (0.057)**	–0.160 (0.055)**	-0.101 (0.042)*
Model 2: Model 1+interactions						
Time×age	–0.021 (0.005)**	–0.022 (0.005)**	–0.012 (0.005)*	–0.016 (0.005)**	–0.012 (0.005)*	–0.008 (0.003)*
Time×gender	0.181 (0.129)	0.145 (0.120)	0.035 (0.116)	–0.079 (0.116)	0.126 (0.113)	0.181 (0.085)*
Time×impl dig	–0.112 (0.127)	–0.012 (0.118)	–0.182 (0.113)	–0.038 (0.113)	0.071 (0.110)	0.059 (0.083)
χ^{2b} (*df*)	23.8 (11)**	26.2 (11)**	11.2 (11)*	11.7 (11)**	9.9 (11)*	12.9 (11)**

[a]Impl dig: implementation of a digital innovation at one's workplace within the last 3 years.

[b]χ^2 compares the lower level model with the respective next level model (ie, model 1 vs model 2).

*P<.05.

**P<.01.

XSL•FO
RenderX

Table 4. Results from multilevel analysis on work stressors (dependent variable) before and after digital transformations (within-subject variable) considering age, gender, and the implementation of a digital innovation at one's workplace within the last 3 years.

	Stress-evoking working conditions (dependent variables), β coefficients (SE)				
	Workload	Time pressure	Excessive complexity of tasks	Excessive demands on concentration	Interruptions of work-flow
Model 1					
Intercept	3.975 (0.247)**	3.992 (0.238)**	3.402 (0.245)**	2.899 (0.293)**	4.889 (0.242)**
Control variables					
Age	–0.002 (0.005)	–0.002 (0.005)	–0.018 (0.005)**	0.001 (0.006)	–0.029 (0.005)**
Gender	0.007 (0.125)	0.085 (0.120)	0.112 (0.123)	0.190 (0.148)	–0.056 (0.122)
Impl dig[a]	0.058 (0.122)	0.148 (0.117)	0.203 (0.120)	-0.046 (0.145)	0.069 (0.119)
Within-subject variable					
Time	–0.146 (0.048)**	–0.175 (0.048)**	0.194 (0.062)**	0.004 (0.041)	–0.045 (0.054)
Model 2: model 1+interactions					
Time×age	0.007 (0.004)	0.008 (0.004)*	0.016 (0.005)**	0.011 (0.003)**	0.014 (0.004)**
Time×gender	–0.078 (0.098)	–0.100 (0.096)	0.042 (0.125)	–0.132 (0.083)	<0.001 (0.108)
Time×impl dig	0.249 (0.096)**	0.436 (0.093)**	0.272 (0.123)*	0.038 (0.081)	0.330 (0.106)**
χ^{2b}	13.61 (11)**	31.43 (11)**	17.11 (11)**	17.77 (11)**	23.62 (11)**

[a]Impl dig: implementation of a digital innovation at one's own workplace within the last 3 years.

[b]χ^2 compares the lower level model with the respective next level model (ie, model 1 vs model 2).

*P<.05.

**P<.01.

Significant main effects of time were revealed for 8 (72%) of the 11 KFZA items (refer to Table 3 for *work resources* and Table 4 for *work stressors*; model 2), which implies that H2 can, at least partly, be retained. For 6 (55%) of these 8 KFZA items for which significant main effects of time were revealed, digital transformation was associated with potentially stress-enhancing effects. More precisely, 5 *work resources* were rated lower (ie, *influence on sequence of work activities*: t_{267}=−4.67; P<.001; *influence on workload and procedures*:t_{267}=−3.21; P=.002; *use of knowledge, skills, and abilities*:t_{267}=−3.49; P<.001; *visibility of task accomplishment*:t_{267}=−2.92; P=.004; and *consideration of employee input*:t_{267}=−2.41; P=.02), and one *work stressor* was rated higher (ie, *excessive complexity of tasks*:t_{267}=3.14; P=.002) after the digital transformation compared to before. With regard to the KFZA item *consideration of employee input*, it should be noted that the effect does not stand up to a conservative Bonferroni correction (P<.005).

With respect to the 2 remaining KFZA items for which significant main effects were revealed, digital transformation was associated with potentially stress-reducing effects, as these *work stressors* (ie, *workload*: t_{267}=−3.04; P=.003; *time pressure*: t_{267}=−3.63; P<.001) were rated lower after the digital transformation compared to before. No significant main effects of time were revealed for the work resource *learning new skills* and the 2 work stressors *excessive demands on concentration* and *interruptions of workflow*.

Influence of Age and Gender on Digitalization-Associated Differences in the Perception of Working Conditions (H2.1 and H2.2)

Results regarding our hypotheses H2.1 and H2.2 are depicted in Table 3 for *work resources* and in Table 4 for *work stressors* (model 3).

With respect to age, significant main effects on the rating of stress-relevant working conditions were revealed (model 1; Tables 3 and 4). More precisely, a higher age was associated with a less stressful perception of work, as indicated by higher ratings of *work resources* (ie, *influence on sequence of activities*; *influence on workload and procedures*; *use of knowledge, skills, and abilities*; and *visibility of task accomplishment*; model 1; Table 3) and lower ratings of *work stressors* (ie, *excessive complexity of tasks* and *interruptions of workflow*; model 1; Table 4).

The examination of our hypothesis H2.1 revealed significant interaction effects of time and age on all *work resources* and all *work stressors* (model 2; Tables 3 and 4). Thereby, a higher age was associated with lower ratings of *work resources* and higher ratings of *work stressors* after the digital transformation compared to before (Figure 3 [31]), indicating that older individuals rated differences associated with digital transformation in a stress-enhancing way compared to younger individuals.

XSL•FO
RenderX

Figure 3. Significant interaction effects of time (before the digital transformation compared to after) and age on the perception of stress-relevant working conditions operationalized using the Short Questionnaire for Workplace Analysis.

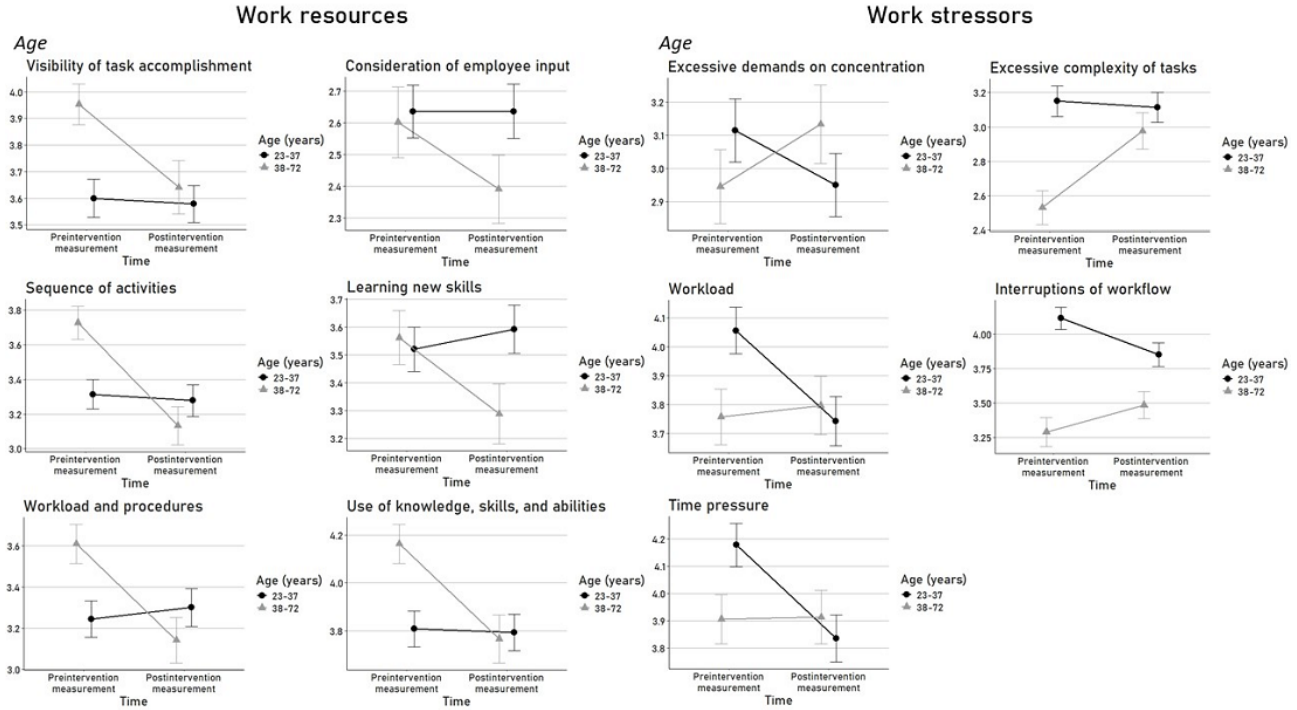

No significant main effects of gender on ratings of stress-relevant working conditions were revealed (model 1; Tables 3 and 4).

The testing of our hypothesis H2.1 revealed one significant interaction effect of time and gender on the rating of the KFZA item *consideration of employee input* (model 2; Tables 3 and 4). More precisely, compared to women, men rated this *work resource* lower after the digital transformation compared to before, indicating that men perceive differences associated with digital transformation in a stress-enhancing way (Figure 4 [31]).

In summary, it can be concluded that hypothesis 2.1, stating that age and gender influence the perceived differences in ratings of stress-relevant working conditions before and after the digital transformation, can be retained.

Figure 4. Significant interaction effects of time (before and after the digital transformation), gender (A), and previous experiences with digital innovations (B) on the perception of stress-relevant working conditions.

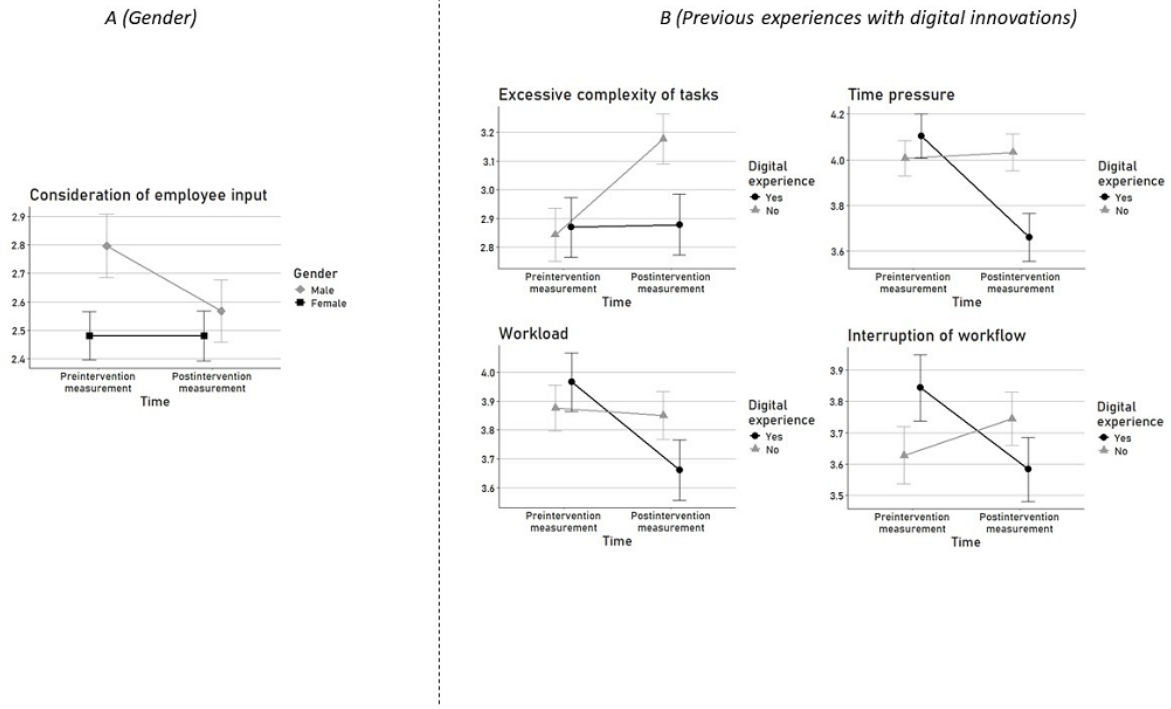

XSL•FO
RenderX

Influence of the Actual Implementation of a Digital Innovation at One's Workplace on Digitalization-Associated Differences in the Perception of Working Conditions

No main effect of the actual implementation of a digital innovation at one's workplace within the last 3 years on the rating of stress-relevant working conditions was revealed (model 1; Tables 3 and 4).

The testing of our hypothesis H2.2 revealed significant interaction effects between time and implementation of a digital innovation at one's workplace on the ratings of 4 *work stressors* (ie, *workload*, *time pressure*, *excessive complexity of tasks*, and *interruptions of workflow*; model 2; Table 4). On the basis of these results, our hypothesis H2.2 that the actual implementation of a digital innovation at one's workplace is associated with differences in the perception of working conditions can be, at least with respect to 4 of the examined working conditions, retained. Thereby, individuals who actually experienced the implementation of a digital innovation at their workplace rated these *work stressors* higher post compared to pre digital transformation (Figure 4 [31]), indicating that actually experiencing the implementation of a digital innovation is associated with a rather stress-enhancing perception of digital transformation, compared to individuals who only imagined how these work stressors would differ in dependence of digital transformation.

Discussion

Principal Findings

The main goal of this study was to identify stress-related working conditions in physicians (H1), which are potentially prone to differ as a function of digital transformation in physicians (H2). In addition, we tested if these potential differences would be influenced by demographic variables (ie, age and gender; H2.1) and the actual experience of the implementation of a digital innovation at one's workplace within the last 3 years (H2.2).

With respect to our first hypothesis, our results support, at least partly, the theoretically derived assumption that the selected working conditions (ie, KFZA items) were stress relevant for physicians, as 10 (90%) of the 11 KFZA items were significantly associated with either a psychological or a biological stress marker. Moreover, the direction of the revealed correlations confirms our categorization of the selected working conditions into *work resources* and *work stressors*. Our finding of associations with working conditions being found mainly with the psychological (PSS-4) and not with the biological (hair F) markers of chronic stress is consistent with a constantly revealed divergence between questionnaire-based measures and cortisol with respect to both hair F (for review, refer to the study by Stalder et al [29]) as well as other, more traditional cortisol measures, such as the cortisol awakening response [47], and cortisol (stress) reactivity [48,49]. Study-specific reasons for the lack of significant associations between the selected working conditions and hair F could be the reduced power in these analyses, as only a small number of participants could be included in these analyses, as well as shared method variance between the rating of the selected working conditions and the PSS-4.

In addition, our hypothesis H2 can be, at least partly, maintained, as our data suggest that digital transformation is associated with significant differences in the perception of 8 (73%) out of 11 stress-relevant working conditions. However, when interpreting these findings, it should be noted that one of these effects did not hold up to a Bonferroni corrector for multiple testing. Interestingly, the vast majority of these differences indicated a stress-enhancing effect of digital transformation (ie, enhanced *work stressors* and reduced *work resources* post compared to pre digital transformation). It should be noted, however, that we also found stress-reducing effects associated with digital transformation (ie, reduced *perceived workload* and reduced *time pressure*). In line with previous research [50,51], this finding contradicts single-sided views that describe processes of digital transformation at work either as an unambiguous savior from chronic stress or a fundamental negative occurrence with mainly negative effects. The stress-enhancing effects associated with digital transformation observed in this study were mainly conveyed by reductions in the perception of *work resources* post compared to pre digital transformation (ie, reduced *perceived own influence possibility on sequence of work activities*; lower *influence on workload and procedures*; less *use of knowledge, skills, and abilities*; less *visibility of task accomplishment*; and less *consideration of employee input*) and not by enhanced perception of *work stressors* (ie, only *excessive complexity of tasks* was rated higher post compared to pre digital transformation). This overall pattern is in line with previous research that suggested that digitalization-associated stress at work might rather be the result of a loss of resources than an increase in workplace stressors [13,16]. Moreover, our results reaffirm previous findings regarding the significance of resources in shaping how individuals perceive and cope with workplace stress [52,53].

Our findings provide important implications for the further development of theoretical models explaining the emergence of technostress [19,20]. They suggest that digital transformation is associated with differences in specific stress-relevant working conditions. Those working conditions should, if confirmed by larger longitudinal studies, be considered as potentially mediating variables. Furthermore, our data provide important insights on which working conditions to focus on for the health-promoting design of implementation processes of digital innovations. In this study, the types of digital transformation differed between participants. This suggests that very different types of digital transformations (ie, digital medicine, robotic, and AI) might result in similar differences in the perception of stress-relevant working conditions.

The Role of Age and Gender

Our hypothesis 2.1 focused on the influence of central demographic variables on the perception of differences in stress-relevant working conditions associated with digital transformation.

Independent of digitalization, a higher age was associated with a general tendency to perceive working conditions in a less

XSL•FO
RenderX

stressful way (ie, higher ratings of work resources and lower ratings of work stressors). Prior research on age-related work stress attributes this phenomenon to several factors. These include the progression of career development, with older workers typically occupying more favorable positions associated with lower chronic stress levels, as well as the development of enhanced stress coping skills over time. In addition, there may be a selection bias favoring stress-resistant employees, as individuals who perceive their work as stressful are more likely to withdraw from their jobs over the years compared to those who do not [54].

In terms of age-dependent differences in the perception of digital transformation at work, the results of this study indicate that older age was linked to a perception of working conditions as more stressful after the digital transformation compared to before. Finding age differences in technostress is in line with previous studies outside the medical context [55]. However, findings regarding the direction of this age effect are mixed, with some studies supporting our findings by revealing a positive association between an increasing age and an increasing technostress [56,57], whereas others showed the opposite pattern [21,58]. Several explanations for an age-dependent increase in technostress have been suggested, such as a diminishing acceptance and use of technology [59], a reduced ability to adapt to new technologies [60,61], less computer experience [55], less computer self-efficacy [55], or enhanced appraisal of difficulties to use digital applications [56,62] in older age.

With respect to gender, our findings indicate that men perceived differences in working conditions in a rather stress-enhancing way compared to women (ie, reduced rating of the work resource *consideration of employee input*). At first glance, this finding conflicts with the cultural tendency to understand technology as a masculine area [63], which would logically result in a more favorable perception of digital transformations for men compared to women. However, the findings of this study are in line with previous notions of women being more positive about the potential of digital transformation [64-66]. As the perception of digital transformation has been shown to influence the willingness to use these technologies [60], the results of this study might as well make an important contribution to accelerate the digitalization of the health system in a gender- and age-adapted manner.

The Role of One's Experience With Digital Innovations

With respect to our hypothesis 2.2, our results suggest that personal experiences with the implementation of digital innovations at one's workplace resulted in a rather stress-enhancing evaluation of digital transformation (ie, higher ratings of 4 *work stressors* post compared to pre digital transformation) compared to only imagining these consequences. To the best of our knowledge, no previous study investigated differences between expected and experienced differences in the rating of stress-related working conditions. Therefore, further research is needed to evaluate the general validity of this finding. If confirmed, our finding suggests that, at least partly, the actual experiences with digital innovations was more negative than the expectations of physicians with no such experience, which would underline the need to evaluate the usefulness of digital innovations, preferably before they are introduced.

Strengths and Limitations

The following limitations challenge the generalizability of the revealed results.

First, the cross-sectional design of this study (pre- and postdigital transformation ratings of the stress-related working conditions were collected at 1 time point only) does not allow to draw conclusions about the causality and the long-term stability of the revealed associations. The conclusion of the literature review by Berg-Beckhoff et al [50] that positive associations between digital technologies at work and stress being mostly found in cross-sectional studies instead of intervention studies underlines the importance of validating the revealed effects within methodically sound longitudinal field studies.

Second, to ensure the highest possible ecological validity, we chose not to focus on one single medical discipline and one specific type of digital transformation, which resulted in a relatively heterogeneous sample composition. The fact that significant effects were found even in such a heterogeneous sample basically speaks for the robustness of the revealed effects. One explanation for these cross-disciplinary effects could be that in everyday medical work the medical disciplines show a similar work distribution, workflows, and workload. For example, surgeons and urologists both operate using similar approaches (open, laparoscopic, and robotic-assisted surgery), both disciplines treat emergency patients additionally to elective cases, and both disciplines manage wards with inpatients. In addition, digital innovations such as the EHR are usually the same for all medical disciplines hospital wide. However, due to the small sample size in the respective conditions or subgroups, we were unable to examine stress-relevant effects of more specific types of digital transformation and specific medical disciplines. Future research with larger sample sizes is needed to examine possible differences in the effects with regard to these variables.

Third, data collection via a web-based survey might have influenced sample composition. More precisely, it seems plausible that the general attitude toward and familiarity with information technologies influences the willingness and capability to participate in a web-based survey, thereby limiting the generalizability of our results.

Conclusions, Implications, and Outlook

Digital transformation processes are omnipresent and substantially alter work in medicine and physician-patient interactions. Therefore, it seems important to systematically evaluate the consequences that these processes may have for stress in physicians. In this regard, our study makes an important contribution by identifying those work stressors and resources that are potentially prone to differ between pre- and postdigital transformation in physicians, depending on age, gender, and previous experiences. If confirmed by comprehensive longitudinal studies, our results not only serve as a valuable addition to theoretical models elucidating the development of technostress in physicians but also carry significant practical

XSL•FO
RenderX

implications for effectively navigating digital transformations. First, our findings hold concrete recommendations on which working conditions to focus on during design and implementation of digital innovation to reduce chronic stress in physicians. Moreover, our finding of a rather stress-enhancing effect of digital transformations emphasizes the importance of carefully evaluating in advance which digital innovations are truly beneficial for the user and which are more likely to deteriorate the working environment. Furthermore, our observation that this potentially stress-enhancing impact is largely driven by reductions in resources could imply the necessity for workplace interventions aimed at conserving resources during digital transformation processes and beyond. Finally, the age- and gender-dependent variations in the perception of digital transformation at work, revealed by this study, suggest that digitalization-associated training and support opportunities at work tailored to specific subgroups might be beneficial.

Acknowledgments

MKW was funded by the Postdoc Starter Kit of the Graduate Academy, Technische Universität Dresden. FvB was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy—EXC 2050/1—Project ID 390696704—Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop," Technische Universität Dresden. The authors would like to thank Josi Harzbecker, Annika Walz, Hanna Singer, Anna Katharina Metzler, Pia Grabbe, and Elisabeth Eisele for their contribution to manuscript preparation.

Conflicts of Interest

None declared.

References

1. Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S, et al. Prevalence of burnout among physicians: a systematic review. JAMA 2018 Sep 18;320(11):1131-1150 [FREE Full text] [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
2. Lazarus RS, Folkman S. Stress, Appraisal, and Coping. Cham, Switzerland: Springer; 1984.
3. Shanafelt TD, Balch CM, Bechamps G, Russell T, Dyrbye L, Satele D, et al. Burnout and medical errors among American surgeons. Ann Surg 2010 Jun;251(6):995-1000. [doi: 10.1097/SLA.0b013e3181bfdab3] [Medline: 19934755]
4. Han S, Shanafelt TD, Sinsky CA, Awad KM, Dyrbye LN, Fiscus LC, et al. Estimating the attributable cost of physician burnout in the United States. Ann Intern Med 2019 Jun 04;170(11):784-790. [doi: 10.7326/M18-1422] [Medline: 31132791]
5. Ritsma A, Forrest L. Causes of chronic stress and impact on physician health. In: Hategan A, Saperson K, Harms S, Waters H, editors. Humanism and Resilience in Residency Training. Cham, Switzerland: Springer International Publishing; 2020.
6. McKee M, van Schalkwyk MC, Stuckler D. The second information revolution: digitalization brings opportunities and concerns for public health. Eur J Public Health 2019 Oct 01;29(Supplement_3):3-6 [FREE Full text] [doi: 10.1093/eurpub/ckz160] [Medline: 31738440]
7. Wiesböck F, Hess T. Digital innovations. Electron Mark 2019 Oct 25;30(1):75-86. [doi: 10.1007/s12525-019-00364-9]
8. Tilson D, Lyytinen K, Sørensen C. Research commentary—digital infrastructures: the missing is research agenda. Inf Syst Res 2010 Dec;21(4):748-759. [doi: 10.1287/isre.1100.0318]
9. Legner C, Eymann T, Hess T, Matt C, Böhmann T, Drews P, et al. Digitalization: opportunity and challenge for the business and information systems engineering community. Bus Inf Syst Eng 2017 Jul 4;59(4):301-308. [doi: 10.1007/s12599-017-0484-2]
10. Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future. National Health Service Health Education England. 2019. URL: https://topol.hee.nhs.uk/the-topol-review/ [accessed 2024-05-09]
11. Grochola LF, Soll C, Zehnder A, Wyss R, Herzog P, Breitenstein S. Robot-assisted single-site compared with laparoscopic single-incision cholecystectomy for benign gallbladder disease: protocol for a randomized controlled trial. BMC Surg 2017 Feb 09;17(1):13 [FREE Full text] [doi: 10.1186/s12893-017-0206-1] [Medline: 28183345]
12. Moore LJ, Wilson MR, McGrath JS, Waine E, Masters RS, Vine SJ. Surgeons' display reduced mental effort and workload while performing robotically assisted surgical tasks, when compared to conventional laparoscopy. Surg Endosc 2015 Sep 27;29(9):2553-2560. [doi: 10.1007/s00464-014-3967-y] [Medline: 25427414]
13. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO study. J Am Med Inform Assoc 2014 Feb 01;21(e1):e100-e106 [FREE Full text] [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]
14. Golz C, Peter KA, Zwakhalen SM, Hahn S. Technostress among health professionals - a multilevel model and group comparisons between settings and professions. Inform Health Soc Care 2021 Jun 02;46(2):136-147. [doi: 10.1080/17538157.2021.1872579] [Medline: 33646891]

XSL•FO
RenderX

15. Heponiemi T, Hyppönen H, Vehko T, Kujala S, Aalto AM, Vänskä J, et al. Finnish physicians' stress related to information systems keeps increasing: a longitudinal three-wave survey study. BMC Med Inform Decis Mak 2017 Oct 17;17(1):147 [FREE Full text] [doi: 10.1186/s12911-017-0545-y] [Medline: 29041971]
16. Heponiemi T, Hyppönen H, Kujala S, Aalto AM, Vehko T, Vänskä J, et al. Predictors of physicians' stress related to information systems: a nine-year follow-up survey study. BMC Health Serv Res 2018 Apr 13;18(1):284 [FREE Full text] [doi: 10.1186/s12913-018-3094-x] [Medline: 29653530]
17. Stahl JE, Egan MT, Goldman JM, Tenney D, Wiklund RA, Sandberg WS, et al. Introducing new technology into the operating room: measuring the impact on job performance and satisfaction. Surgery 2005 May;137(5):518-526. [doi: 10.1016/j.surg.2004.12.015] [Medline: 15855924]
18. Haimi M, Brammli-Greenberg S, Waisman Y, Baron-Epel O. Physicians' experiences, attitudes and challenges in a Pediatric Telemedicine Service. Pediatr Res 2018 Nov;84(5):650-656. [doi: 10.1038/s41390-018-0117-6] [Medline: 30120402]
19. Brod C. Technostress: The Human Cost Of The Computer Revolution. New York, NY: Basic Books; 1984.
20. Gimpel H, Lanzl J, Manner-Romberg T, Nüske N. Digitaler stress in Deutschland: eine befragung von erwerbstätigen zu belastung und beanspruchung durch arbeit mit digitalen technologien. In: Digitalisierung und Gesundheit. Baden-Baden, Germany: Nomos Publishing House; 2022:265-302.
21. Ragu-Nathan TS, Tarafdar M, Ragu-Nathan BS, Tu Q. The consequences of technostress for end users in organizations: conceptual development and empirical validation. Inf Syst Res 2008 Dec;19(4):417-433. [doi: 10.1287/isre.1070.0165]
22. Kubicek B, Paškvan M, Prem R, Schöllbauer J, Till M, Cabrita J, et al. Working conditions and workers' health. European Union. 2019. URL: https://data.europa.eu/doi/10.2806/909840 [accessed 2024-02-28]
23. Cabrita J, Cerf C, Foden D. Overtime in Europe: regulation and practice. Publications Office of the European Union. 2022. URL: https://data.europa.eu/doi/10.2806/095550 [accessed 2024-02-28]
24. Netterstrøm B, Conrad N, Bech P, Fink P, Olsen O, Rugulies R, et al. The relation between work-related psychosocial factors and the development of depression. Epidemiol Rev 2008 May 14;30(1):118-132. [doi: 10.1093/epirev/mxn004] [Medline: 18587142]
25. Niedhammer I, Bertrais S, Witt K. Psychosocial work exposures and health outcomes: a meta-review of 72 literature reviews with meta-analysis. Scand J Work Environ Health 2021 Oct 01;47(7):489-508 [FREE Full text] [doi: 10.5271/sjweh.3968] [Medline: 34042163]
26. Nixon AE, Mazzola JJ, Bauer J, Krueger JR, Spector PE. Can work make you sick? A meta-analysis of the relationships between job stressors and physical symptoms. Work Stress 2011 Jan;25(1):1-22. [doi: 10.1080/02678373.2011.569175]
27. Rau R, Buyken D. Der aktuelle kenntnisstand über erkrankungsrisiken durch psychische arbeitsbelastungen. Zeitschrift Arbeits- Organisationspsychologie A&O 2015 Jun 16;59(3):113-129 [FREE Full text] [doi: 10.1026/0932-4089/a000186]
28. Rau R. Lern- und gesundheitsförderliche arbeitsgestaltung: eine empirische studie. Zeitschrift Arbeits Organisationspsychologie A&O 2004 Oct;48(4):181-192 [FREE Full text] [doi: 10.1026/0932-4089.48.4.181]
29. Stalder T, Steudte-Schmiedgen S, Alexander N, Klucken T, Vater A, Wichmann S, et al. Stress-related and basic determinants of hair cortisol in humans: a meta-analysis. Psychoneuroendocrinology 2017 Mar;77:261-274. [doi: 10.1016/j.psyneuen.2016.12.017] [Medline: 28135674]
30. Preparing the healthcare workforce to deliver the digital future: an independent report on behalf of the Secretary of State for Health and Social Care, February 2019. National Health Service. 2019. URL: https://topol.hee.nhs.uk/the-topol-review/#:~:text=About%20the%20Topol%20Review&text=The%20Topol%20Review%2C%20led%20by,and%20three%20Expert%20Advisory%20Panels [accessed 2024-05-09]
31. Prümper J, Hartmannsgruber K, Frese M. KFZA. Kurz-fragebogen zur arbeitsanalyse. Zeitschrift Arbeits Organisationspsychologie A&O 1995 Jan;39(3):125-131.
32. Beck D, Berger S, Breutmann N, Fergen A, Gregersen S, Morschhäuser M, et al. Recommendations of the institutions of the Joint German Occupational Safety and Health Strategy (GDA) for implementing psychosocial risk assessment. Federal Ministry of Labour and Social Affairs, Berlin. 2014. URL: http://www.gda-psyche.de/SharedDocs/Publikationen/EN/Recommendations%20for%20implementing%20psychoso cial%20risk%20assessment.html [accessed 2024-05-09]
33. Appel P, Schuler M, Vogel H, Oezelsel A, Faller H. Short Questionnaire for Workplace Analysis (KFZA): factorial validation in physicians and nurses working in hospital settings. J Occup Med Toxicol 2017 May 12;12(1):11 [FREE Full text] [doi: 10.1186/s12995-017-0157-6] [Medline: 28507593]
34. Bauer J, Bendels MH, Groneberg DA. [Subjective job strain and job satisfaction among neurologists in German hospitals]. Nervenarzt 2016 Jun 4;87(6):629-633. [doi: 10.1007/s00115-016-0076-2] [Medline: 26847570]
35. Buddeberg-Fischer B, Klaghofer R, Buddeberg C. Die motivation zur weiterbildung in psychiatrie und psychotherapie: eine untersuchung bei jungen ärztinnen und ärzten in der deutschschweiz. Schweiz Arch Neurol Psychiatr 2008;159(01):34-41. [doi: 10.4414/sanp.2008.01921]
36. Christoph Eichert H. Supervision und Ressourcenentwicklung: Eine Untersuchung zur Supervision in der Stationären Psychiatrie. Riga, Latvia: VDM Verlag; 2013.
37. Klein EM, Brähler E, Dreier M, Reinecke L, Müller KW, Schmutzer G, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16:159 [FREE Full text] [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]

XSL•FO
RenderX

38. Warttig SL, Forshaw MJ, South J, White AK. New, normative, English-sample data for the short form Perceived Stress Scale (PSS-4). J Health Psychol 2013 Dec;18(12):1617-1628. [doi: 10.1177/1359105313508346] [Medline: 24155195]
39. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404]
40. Kudielka BM, Wüst S. Human models in acute and chronic stress: assessing determinants of individual hypothalamus-pituitary-adrenal axis activity and reactivity. Stress 2010 Jan;13(1):1-14. [doi: 10.3109/10253890902874913] [Medline: 20105052]
41. Stalder T, Kirschbaum C. Analysis of cortisol in hair--state of the art and future directions. Brain Behav Immun 2012 Oct;26(7):1019-1029. [doi: 10.1016/j.bbi.2012.02.002] [Medline: 22366690]
42. Gao W, Kirschbaum C, Grass J, Stalder T. LC-MS based analysis of endogenous steroid hormones in human hair. J Steroid Biochem Mol Biol 2016 Sep;162:92-99. [doi: 10.1016/j.jsbmb.2015.12.022] [Medline: 26718873]
43. Wennig R. Potential problems with the interpretation of hair analysis results. Forensic Sci Int 2000 Jan 10;107(1-3):5-12. [doi: 10.1016/s0379-0738(99)00146-2] [Medline: 10689559]
44. Gao W, Stalder T, Foley P, Rauh M, Deng H, Kirschbaum C. Quantitative analysis of steroid hormones in human hair using a column-switching LC-APCI-MS/MS assay. J Chromatogr B Analyt Technol Biomed Life Sci 2013 Jun 01;928:1-8. [doi: 10.1016/j.jchromb.2013.03.008] [Medline: 23584040]
45. Enge S, Fleischhauer M, Hadj-Abo A, Butt F, Kirschbaum C, Schmidt K, et al. Comparison of hair cortisol concentrations between self- and professionally-collected hair samples and the role of five-factor personality traits as potential moderators. Psychoneuroendocrinology 2020 Dec;122:104859 [FREE Full text] [doi: 10.1016/j.psyneuen.2020.104859] [Medline: 32992135]
46. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: linear and nonlinear mixed effects models. R package version 3. The Comprehensive R Archive Network. 2018. URL: https://www.scirp.org/reference/referencespapers?referenceid=3025117 [accessed 2024-05-10]
47. Chida Y, Steptoe A. Cortisol awakening response and psychosocial factors: a systematic review and meta-analysis. Biol Psychol 2009 Mar;80(3):265-278. [doi: 10.1016/j.biopsycho.2008.10.004] [Medline: 19022335]
48. Campbell J, Ehlert U. Acute psychosocial stress: does the emotional stress response correspond with physiological responses? Psychoneuroendocrinology 2012 Aug;37(8):1111-1134. [doi: 10.1016/j.psyneuen.2011.12.010] [Medline: 22260938]
49. Feldman PJ, Cohen S, Lepore SJ, Matthews KA, Kamarck TW, Marsland AL. Negative emotions and acute physiological responses to stress. Ann Behav Med 1999;21(3):216-22; discussion 223-6. [doi: 10.1007/BF02884836] [Medline: 10626027]
50. Berg-Beckhoff G, Nielsen G, Ladekjær Larsen E. Use of information communication technology and stress, burnout, and mental health in older, middle-aged, and younger workers - results from a systematic review. Int J Occup Environ Health 2017 Apr;23(2):160-171 [FREE Full text] [doi: 10.1080/10773525.2018.1436015] [Medline: 29460697]
51. Christensen JO, Finne LB, Garde AH, Nielsen MB, Sørensen K, Vleeshouwes J. The influence of digitalization and new technologies on psychosocial work environment and employee health: a literature review. STAMI. 2019 Nov. URL: https://hdl.handle.net/11250/2723779 [accessed 2024-05-10]
52. Bakker AB, Demerouti E, Euwema MC. Job resources buffer the impact of job demands on burnout. J Occup Health Psychol 2005 Apr;10(2):170-180. [doi: 10.1037/1076-8998.10.2.170] [Medline: 15826226]
53. Xanthopoulou D, Bakker AB, Demerouti E, Schaufeli WB. The role of personal resources in the job demands-resources model. Int J Stress Manag 2007;14(2):121-141. [doi: 10.1037/1072-5245.14.2.121]
54. Hsu HC. Age differences in work stress, exhaustion, well-being, and related factors from an ecological perspective. Int J Environ Res Public Health 2018 Dec 25;16(1):50 [FREE Full text] [doi: 10.3390/ijerph16010050] [Medline: 30585250]
55. Tams S, Thatcher JB, Grover V. Concentration, competence, confidence, and capture: an experimental study of age, interruption-based technostress, and task performance. J Assoc Inf Syst 2018 Jan;19(9):857-908 [FREE Full text] [doi: 10.17705/1jais.00511]
56. Burton-Jones A, Hubona GS. Individual differences and usage behavior: revisiting a technology acceptance model assumption. ACM SIGMIS Database Adv Inf Syst 2005 Jun 07;36(2):58-77. [doi: 10.1145/1066149.1066155]
57. Tu Q, Wang K, Shu Q. Computer-related technostress in China. Commun ACM 2005 Apr 01;48(4):77-81. [doi: 10.1145/1053291.1053323]
58. Hsiao KL, Lee CH, Chiang HS, Wang JY. Exploring the antecedents of technostress and compulsive mobile application usage: personality perspectives. In: Proceedings of the International Symposium on Interactive Technology and Ageing Populations. 2016 Presented at: ITAP 2016; October 20-22, 2016; Kochi, Japan. [doi: 10.1007/978-3-319-39943-0_31]
59. Magsamen-Conrad K, Upadhyaya S, Joa CY, Dowd J. Bridging the divide: using UTAUT to predict multigenerational tablet adoption practices. Comput Human Behav 2015 Sep 01;50:186-196 [FREE Full text] [doi: 10.1016/j.chb.2015.03.032] [Medline: 25937699]
60. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]

XSL•FO
RenderX

61. O'Driscoll MP, Brough P, Timms C, Sawang S. Engagement with information and communication technology and psychological well-being. In: Perrewé PL, Ganster DC, editors. New Developments in Theoretical and Conceptual Approaches to Job Stress (Research in Occupational Stress and Well Being. Leeds, England: Emerald Group Publishing Limited; 2010.
62. Day A, Scott N, Kevin Kelloway E. Information and communication technology: implications for job stress and employee well-being. In: Perrewé PL, Ganster DC, editors. New Developments in Theoretical and Conceptual Approaches to Job Stress (Research in Occupational Stress and Well Being. Leeds, England: Emerald Group Publishing Limited; 2010.
63. Salminen‐Karlsson M. Women who learn computing like men: different gender positions on basic computer courses in adult education. J Vocat Educ Train 2009 May 15;61(2):151-168. [doi: 10.1080/13636820902933254]
64. Liff S. Clerical workers and information technology: gender relations and occupational change. New Technol Work Employ 1990 Mar;5(1):44-55. [doi: 10.1111/j.1468-005X.1990.tb00118.x]
65. Norris DF. Gender, job, and effects of microcomputers in public organizations. State Local Gov Rev 1992;24(2):65-70.
66. Pérez MP, de-Luis P, Sánchez AM. Differential effects of gender on perceptions of teleworking by human resources managers. Women Manag Rev 2002 Sep;17(6):262-275. [doi: 10.1108/09649420210441914]

Abbreviations

AI: artificial intelligence
ANCOVA: analysis of covariance
EHR: electronic health record
HPA: hypothalamus-pituitary-adrenal
KFZA: Short Questionnaire for Workplace Analysis
PSS-4: Perceived Stress Scale 4

Edited by G Tsafnat; submitted 02.06.23; peer-reviewed by J Cecil, K Muroi; comments to author 20.02.24; revised version received 20.03.24; accepted 06.04.24; published 17.06.24.

Please cite as:
Wekenborg MK, Förster K, Schweden F, Weidemann R, Bechtolsheim FV, Kirschbaum C, Weitz J, Ditzen B
Differences in Physicians' Ratings of Work Stressors and Resources Associated With Digital Transformation: Cross-Sectional Study
J Med Internet Res 2024;26:e49581
URL: https://www.jmir.org/2024/1/e49581
doi:10.2196/49581
PMID:38885014

©Magdalena Katharina Wekenborg, Katharina Förster, Florian Schweden, Robin Weidemann, Felix von Bechtolsheim, Clemens Kirschbaum, Jürgen Weitz, Beate Ditzen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patients' and Clinicians' Perceptions of the Clinical Utility of Predictive Risk Models for Chemotherapy-Related Symptom Management: Qualitative Exploration Using Focus Groups and Interviews

Morven Miller[1], BA (Hons), RGN, MSc, PhD; Lisa McCann[1], MSc, PhD; Liane Lewis[2], BSc, PhD; Christine Miaskowski[3], RN, PhD; Emma Ream[4], MSc, PhD; Andrew Darley[5], BA(Mod), MSc, PhD; Jenny Harris[4], MSc, GradCert, PhD; Grigorios Kotronoulas[6], MSc, BSN, RN, PhD; Geir V Berg[7], RN, MSc, PhD; Simone Lubowitzki[8], MA, PhD; Jo Armes[4], RGN, MSc, PhD; Elizabeth Patiraki[9], RN, BSc, PhD; Eileen Furlong[10], MMed Sc, PhD; Patricia Fox[10], BNS, MA, PhD; Alexander Gaiger[8], MD; Antonella Cardone[11], MSc, MBA; Dawn Orr[12], MSc, PhD; Adrian Flowerday[13], BSc, CEng; Stylianos Katsaragakis[9], BA, MSc, PhD; Simon Skene[14], MSc, PhD; Margaret Moore[1], BSc (Hons); Paul McCrone[15], BA (Hons), MSc, PhD; Nicosha De Souza[16], MSc, PhD; Peter T Donnan[16], BA, MSc, PhD; Roma Maguire[1], RGN, BN, MSci, PhD

[1]Computer & Information Sciences, University of Strathclyde, Glasgow, United Kingdom
[2]Johnson and Johnson Medical, Norderstedt, Germany
[3]University of California, San Francisco, CA, United States
[4]School of Health Sciences, University of Surrey, Guildford, United Kingdom
[5]School of Medicine, University College Dublin, Dublin, Ireland
[6]School of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, United Kingdom
[7]Department of Health Sciences, Norwegian University of Science and Technology, Gjøvik, Norway
[8]Department of Internal Medicine 1, Division of Hematology and Hemostaseology, Medical University of Vienna, Vienna, Austria
[9]School of Health Sciences, National and Kapodistrian University of Athens, Athens, Greece
[10]School of Nursing, Midwifery and Health Systems, University College Dublin, Dublin, Ireland
[11]European Cancer Patient Coalition, Brussels, Belgium
[12]NHS 24, Glasgow, United Kingdom
[13]Docobo Limited, Leatherhead, United Kingdom
[14]Surrey Clinical Trials Unit, University of Surrey, Guildford, United Kingdom
[15]Department of Health Services and Population Research, Institute of Psychiatry Psychology and Neuroscience, King's College London, London, United Kingdom
[16]Population Health and Genomics, Medical School, University of Dundee, Dundee, United Kingdom

Corresponding Author:
Roma Maguire, RGN, BN, MSci, PhD
Computer & Information Sciences
University of Strathclyde
13.15 Livingston Tower
26 Richmond Street
Glasgow, G1 1XQ
United Kingdom
Phone: 44 141 548 4101
Email: roma.maguire@strath.ac.uk

Abstract

Background: Interest in the application of predictive risk models (PRMs) in health care to identify people most likely to experience disease and treatment-related complications is increasing. In cancer care, these techniques are focused primarily on the prediction of survival or life-threatening toxicities (eg, febrile neutropenia). Fewer studies focus on the use of PRMs for

XSL•FO
RenderX

symptoms or supportive care needs. The application of PRMs to chemotherapy-related symptoms (CRS) would enable earlier identification and initiation of prompt, personalized, and tailored interventions. While some PRMs exist for CRS, few were translated into clinical practice, and human factors associated with their use were not reported.

Objective: We aim to explore patients' and clinicians' perspectives of the utility and real-world application of PRMs to improve the management of CRS.

Methods: Focus groups (N=10) and interviews (N=5) were conducted with patients (N=28) and clinicians (N=26) across 5 European countries. Interactions were audio-recorded, transcribed verbatim, and analyzed thematically.

Results: Both clinicians and patients recognized the value of having individualized risk predictions for CRS and appreciated how this type of information would facilitate the provision of tailored preventative treatments or supportive care interactions. However, cautious and skeptical attitudes toward the use of PRMs in clinical care were noted by both groups, particularly in relationship to the uncertainty regarding how the information would be generated. Visualization and presentation of PRM information in a usable and useful format for both patients and clinicians was identified as a challenge to their successful implementation in clinical care.

Conclusions: Findings from this study provide information on clinicians' and patients' perspectives on the clinical use of PRMs for the management of CRS. These international perspectives are important because they provide insight into the risks and benefits of using PRMs to evaluate CRS. In addition, they highlight the need to find ways to more effectively present and use this information in clinical practice. Further research that explores the best ways to incorporate this type of information while maintaining the human side of care is warranted.

Trial Registration: ClinicalTrials.gov NCT02356081; https://clinicaltrials.gov/study/NCT02356081

(J Med Internet Res 2024;26:e49309) doi:10.2196/49309

KEYWORDS

chemotherapy; digital health; education; predictive risk models; qualitative research methods; symptoms; symptom cluster; tailored information

Introduction

Background

Over 19 million people worldwide were diagnosed with cancer in 2020 [1]. By 2040, the global burden is expected to grow to 29.4 million new cancer cases annually, which reflects the growth and aging of the global population [2]. While chemotherapy is one of the most effective cancer treatments [3-5], some patients experience severe and potentially life-threatening symptoms [6-8]. Poor symptom management can result in symptom escalation [9,10], poor adherence to treatment [11,12], reduced health-related quality of life (QoL) [13,14], and treatment-related mortality [9,15]. The value of early detection and prompt symptom management cannot be underestimated given that decreases in symptom severity and rapid resolution are associated with improved patient experiences and treatment outcomes. Moreover, symptoms can persist beyond treatment into survivorship [16], increase the long-term burden of cancer on individuals and their families [6,17,18], and generate significant costs to the health care system and society [19].

Symptom Management and Information Provision

Accurate prediction of patients who are most likely to experience high levels of symptom burden is clinically challenging [20]. These challenges relate to variability in patients' demographic (eg, age and gender), clinical (eg, comorbidities), genetic, and socioeconomic characteristics that may be key moderators of patients' symptom experience, and psychosocial adjustment to cancer treatment [18,21,22].

Traditional approaches to managing symptoms in patients receiving chemotherapy on an outpatient basis rely on patients' recollection when they attend for subsequent treatments. However, approaches that depend on patients' recollection are subject to both recall bias and underreporting [23]. Alternatively, people present to acute oncology services with severe symptoms. The ideal scenario would be to predict, pre-empt and, when possible, treat chemotherapy-related symptoms (CRS) prophylactically in addition to effective self-care.

Prediction of Future Symptom Experiences

During the last decade, enhanced computing power and the availability of large amounts of data have prompted the practical use of artificial intelligence (AI) in health care [24]. AI has the potential to transform health care practices with its increasing ability to translate the uncertainty and complexity of data into actionable clinical decisions [25]. Predictive models have been used to identify patients who are most likely to experience specific disease and treatment-related events [26,27].

One of the first studies to explore the use of predictive risk modeling (PRM) in supportive care developed and tested a symptom risk modeling tool for patients with breast cancer undergoing adjuvant chemotherapy and demonstrated a high level of accuracy in predicting 4 out of 6 symptoms [28]. Recent studies advanced the state of the art by employing person-centered analytical approaches to gain insights into the complex nature of co-occurring symptoms and symptom clusters and the identification of subgroups of patients with distinct symptom profiles [20,29,30]. These findings highlight some of the potential that risk prediction can play in the delivery of person-centered care through the identification of high-risk patients with multiple co-occurring symptoms.

XSL•FO
RenderX

While an evidence base is emerging around the development of PRMs for supportive cancer care, information about clinicians' and patients' perspectives on their implementation, utility, and real-world application is limited. This significant gap in the literature warrants evaluation [25,31,32] so that PRMs can be implemented in ways that will improve the quality of care and patient outcomes [33].

Aim of This Study

This paper reports on an exploratory secondary objective from a larger program of work Electronic Symptom Management Using the Advanced Symptom Management System (eSMART; ClinicalTrials.gov NCT02356081). The primary study was a randomized controlled trial (RCT) that evaluated the effects of real-time remote monitoring on symptom burden, QoL, supportive care needs, anxiety, self-efficacy, and work limitations experienced by patients with cancer receiving chemotherapy [34-37]. eSMART was funded through the European Union's Seventh Framework Programme for research, technological development, and demonstration (602289EU) and was conducted across 12 clinical sites within 5 European countries (ie, Austria, Greece, Republic of Ireland, Norway, and the United Kingdom). The purpose of the current analysis was to explore patients' and clinicians' perceptions of the utility and real-world application of PRMs for CRS.

Methods

Participants and Setting

Purposive groups of patients and clinicians were recruited from 12 clinical sites in Austria, Greece, Ireland, Norway, and the United Kingdom. Participants (ie, patients and clinicians) were not required to have participated in the RCT component of eSMART. Given that both patients' and clinicians' perspectives were equally important, adequate representation from both groups was sought.

Potential participants were purposively identified and recruited by members of the local health care team to have a sample that was as varied as possible per age, gender, diagnoses (for patients), and clinical roles (for clinicians) to capture a wide range of viewpoints. Patients were eligible if they were diagnosed with breast or colorectal cancer or Hodgkin or non-Hodgkin lymphoma and had received chemotherapy within the past year. Clinicians were eligible if they were members of clinical teams involved in the provision of supportive care for patients with breast, colorectal, or any hematological cancer receiving chemotherapy.

Ethical Considerations

This study was approved by NHS Lothian South East Scotland Research Ethics Committee 02 (14/SS/1062) and sponsored by the University of Strathclyde. This study received ethical approval from each of the study's sites. All participants provided written informed consent.

Data Collection

Patient and clinician interviews and focus groups were conducted separately. Interview guides were developed for each sample (patient and clinician) that aimed to explore and understand patients' and clinicians' perspectives on the utility and real-world application of the PRMs on CRS. These guides were devised by the study team to ensure that they focused on key questions. While patient and clinician topic guides addressed the same issues and followed the same format, minor changes were made for each target group (Figures 1 and 2). Given the novel nature of the concepts under discussion, visual aids were used to facilitate discussions (Figure 3). These visual aids provided examples of ways in which risk predictions of CRS could be presented to support understanding of the concept being discussed and to explore ways in which risks could be visualized and understood by patients and clinicians.

Data were collected between January and April 2019. The main study team provided education and training to each of the clinical sites to conduct the focus groups and interviews in the participants' native language. The education included concepts to be discussed and the interview guides and visual aids to be used. In addition, the clinical sites received all of the material that would be needed to conduct the interviews or focus groups. Data collection involved 8 interviewers: 5 female and 3 male. Further, 2 of these interviewers were members of the main study team while the remainder were principal or local investigators at the clinical sites. While some interviewers may have been familiar to the clinician participants, they had no direct clinical or line management relationships with them. The interviewers were entirely unknown to the patient participants. For each site, the initial interview or focus group was evaluated by a member of the main study team and feedback was provided to optimize data collection. Except for 1 interview that was conducted by telephone, the other 4 interviews and 10 focus groups were conducted in person.

XSL•FO
RenderX

Figure 1. Topic Guide for Clinicians and Patients.

Patient Topic Guide	Clinician Topic Guide
Show Slide 1 – and talk through the concept of predictive risk models	*Show Slide 1 – and talk through the concept of predictive risk models*
As a patient, if you had information on what chemotherapy related symptoms you as an **individual** patient would experience in advance of starting chemotherapy…. • Would you want this information – why/why not? • Would you find this information helpful? • How would you use this information? • Would this information change the way that you managed your symptoms?	As a clinician, if you had information on what chemotherapy related symptoms an **individual** patient would experience in advance of starting chemotherapy…. • Would you want this information – why/why not? • Would you find this information helpful? • How would you use this information? • Would this information change the way that you currently manage patients?
Show slide 2 – this explains the 3 dimensions of symptoms – occurrence, severity and distress	*Show slide 2 – this explains the 3 dimensions of symptoms – occurrence, severity and distress*
Would you want predictions of the different dimensions of the symptom experience? For example, would you want a diagram predicting: • The likelihood of you having the symptom – why? • The predicted severity you would experience of the symptom – why? • The predicted level of distress that you would experience from the symptom – why?	Would you want predictions of the different dimensions of the symptom experience? For example, would you want a diagram predicting: • The likelihood of the person having the symptom – why? • The predicted severity of the symptom – why? • The predicted level of distress that the person will experience from the symptom – why? • Of all 3 dimensions of the

Figure 2. Topic Guides for Clinicians and Patients.

Would you want predictions of symptoms to show: • Predicted risk of symptoms on a daily basis – why? • Predicted risk of symptoms per cycle of chemotherapy – why? • Predicted risk of symptoms for the overall duration of chemotherapy treatment – why?	Would you want predictions of symptoms to show: • Predicted risk of symptoms on a daily basis – why? • Predicted risk of symptoms per cycle of chemotherapy – why? • Predicted risk of symptoms for the overall duration of chemotherapy treatment – why?
Option to show slides 3 & 4 – these give examples of how predictive risk models can be presented – feel free to hold off on this if you feel that the group/individual could have their own ideas	*Option to show slides 3 & 4 – these give examples of how predictive risk models can be presented – feel free to hold off on this if you feel that the group/individual could have their own ideas*
How do you think PRMs could be presented to you as a patient that would make them most useful for you? (graphs, charts, tables, anything else…?)	How do you think PRMs could be presented to you as clinicians that would make them most useful in your daily/clinical practice? (graphs, charts, tables, anything else…?)
Show slides 5-7 – these show diagrams of how symptoms (occurrence, severity and distress) could be presented to you?	*Show slides 5-7 – these show diagrams of how symptoms (occurrence, severity and distress) could be presented to you?*
Do these make sense to you? • Are the associations between symptoms as you would expect? Yes/no – please explain? • Are there any associations that are surprising to you? Yes/no – why?	Do these make sense clinically? • Are the associations between symptoms as expected? Yes/no – please explain? • Are there any associations that are surprising? Yes/no – why?
Have you any other thoughts or comments about Predictive Risk Models that you would like to add to the discussion today?	Have you any other thoughts or comments about Predictive Risk Models that you would like to add to the discussion today?

XSL•FO
RenderX

Figure 3. Visual aids used alongside topic guides.

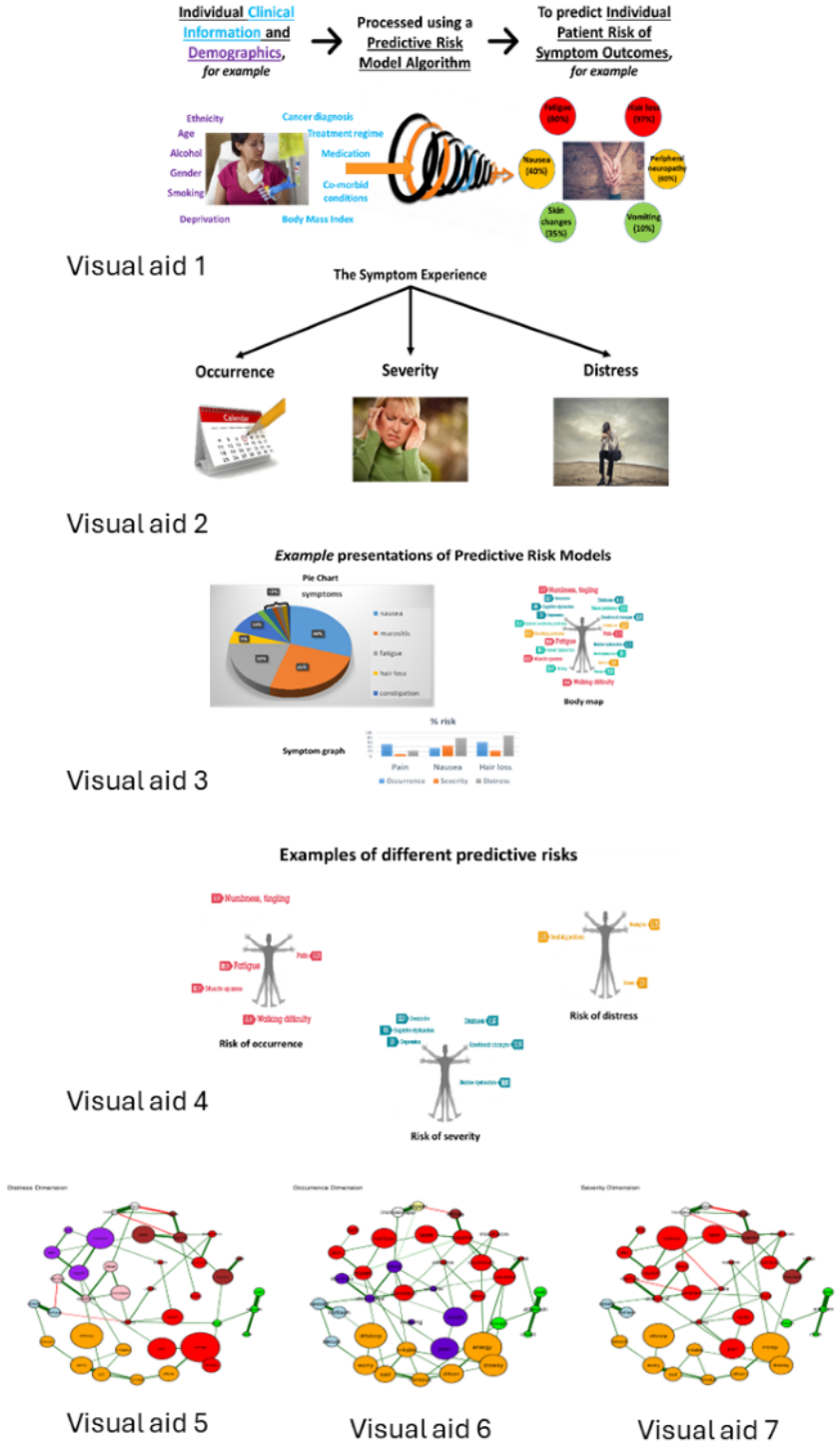

Data Synthesis

Interviews and focus groups were recorded on encrypted digital recording devices and uploaded to a secure cloud-based server by all clinical sites. The recordings were transcribed verbatim by a transcription company. For the non-English recordings, the transcriber listened to them in the original language and then transcribed them into English. The correctness of the transcriptions was verified by a member of this study's team whose native language was German, Greek, or Norwegian who listened to the recordings and read the English translation simultaneously. Any errors were corrected in the transcript before analysis.

Data Analysis

Qualitative data were analyzed using NVivo software (Lumivero). Thematic analysis, a method of identifying, analyzing, organizing, describing, and reporting themes found

XSL•FO
RenderX

within a data set [38], was used to answer this study's aims. The initial coding framework was based on the patient and clinician interview guides and data from the 2 groups was analyzed together. Common overarching themes were identified in both the patient and clinician data. However, within the subthemes, both convergent and divergent themes emerged between the patients and the clinicians.

Data analysis, coding, and findings were reviewed at weekly team meetings of the lead team of the eSMART RCT to encourage collaborative discussion about the data and facilitate the interpretation of findings. Further, 2 researchers (RM and M Miller) were responsible for completing the thematic analysis individually and then meeting to discuss. A review of the full data set was conducted by a further member of this study's team (LL) to verify the completed analysis. The main study team concluded that saturation was reached with this sample.

Results

Overview

In total, 23 patients participated in 4 focus groups in the United Kingdom, Austria, Norway, and Greece. These focus groups ranged in size from 4 to 6 participants. All 5 patients from Ireland opted for individual interviews due to logistic concerns related to travel, challenges in finding a mutually convenient time, and ongoing symptoms (such as fatigue) following chemotherapy. In total, 26 clinicians participated in 6 focus groups across all 5 countries. These focus groups ranged in size from 2 to 6 participants.

Participants

Table 1 describes the patients' and clinicians' characteristics concerning country, number, age, and diagnosis (patients only). Details on the clinicians' roles are presented in Table 2.

Table 1. Participant demographic characteristics.

	Patient sample						Clinician sample	
	Breast cancer		Colorectal cancer		Hematological cancer		n	Mean (SD)
	n	Mean (SD)	n	Mean (SD)	n	Mean (SD)		
United Kingdom (3 sites)	5	51 (10.7)	3	57 (5.7)	—[a]	—	8	41 (10)
Ireland (1 site)	4	51 (15.8)	1	70	—	—	6	—
Greece (1 site)	1	42	1	70	2	69 (8.5)	5	41 (10.1)
Austria (1 site)	4	55 (11.3)	—	—	2	56 (3.5)	4	43 (7.9)
Norway (1 site)	4	53 (4.2)	1	31	—	—	3	—

[a]—: not available.

Table 2. Clinician roles.

Roles	Participants, n
Doctor	2
Nurse (staff nurse, oncology nurse, and chemotherapy nurse)	9
Nurse consultant	3
Clinical nurse manager	3
Clinical nurse specialist	4
Head nurse	2
Patient information lead	1
Liaison nurse	2

Themes Identified

From the data analyses, while overarching themes (ie, perceived benefits of PRMs, negative perceptions of risk, and using risk prediction data) were common across both groups, distinct differences emerged in some of the subthemes by group. Therefore, patients' and clinicians' findings are reported separately.

Patient Themes

Analysis of the patient data identified 3 main themes and several subthemes (Table 3). Specific details for each theme are presented in Table 3.

XSL•FO
RenderX

Table 3. Patient themes and subthemes.

Themes	Subthemes
Positive perceptions of predicting risk	• Managing expectations • Enabling self-care • Planning, preparing, and being in control • Educating friends and family members
Negative perceptions of predicting risk	• Skeptical of concept • Negative mindset • Self-fulfilling prophecy
Use of risk prediction data	• Visualizing the data • Frequency of data • Preferences for technology

Positive Perceptions of Predicting Risk

Overview

Patients were able to see positive uses for information provided through the use of PRMs for CRS. These uses included helping them to manage symptom expectations; enabling self-care; planning, preparing, being in control; and educating family members and friends.

Managing Expectations

Most patients spoke positively about being able to access personalized predictions of symptoms that they were most likely to experience at the start of chemotherapy. They perceived that this information would help them to better manage their expectations of treatment, reduce associated distress, and improve adjustments to treatment and its impact. Further, 1 patient spoke about how having this individualized information could make them feel "a bit more special...":

> *Because I would be prepared for [me], not for what chemotherapy does to people as a whole….Treatment cannot have the same symptoms to everyone. I would want the targeted information to be for [me] and not based on a protocol, and the symptoms would be accurate. Just like I know the treatment, I would want to know the symptoms for me. I think that this would be the best.* [site 5, patient 4]

> *I was dreading nausea so much because that's the worst thing I know…[it] would have been really good for me to know that the prognosis shows that I'm likely to not get it – that would have saved me from a lot of dreading.* [site 7, patient 3]

Patients highlighted how personalized predictions of CRS would help them distinguish between symptoms that were or were not related to their chemotherapy. They perceived that they would be able to distinguish the cause of symptoms and decipher if they were "normal" for them or not. In addition, they suggested that these predictions would have a role in ensuring that patients report "relevant" symptoms to clinicians in a timely manner:

> *I think it would be easier to handle. Knowing that something is a symptom of something makes it also easier to talk about. Otherwise, it's like this, 'I don't know… I feel that… is it normal or not.'* [site 6, patient 1]

> *Merely to be able to self-assess my situation and say how significant is this, is this normal? This is what used to bother me the most. I had some side effects, I felt something and the uncertainty of whether I should report it, whether it's something that needed some specific course of action or whether it was something completely normal.* [site 6, patient 6]

Enabling Self-Care

Patients spoke about how predictive information would allow them to better plan and prepare and enable them to proactively care for themselves:

> *Okay, I can get some creams in for that and have them ready, rather than going shopping for the stuff...when you—when it happens. So, you have everything under the sun ready to go just in case.* [site 1, patient 2]

> *If I'm going to be sore on this day, maybe I could take a pain killer or something that would, that would stop it, like.* [site 4, patient 1]

Planning, Preparing, and Being in Control

A few patients highlighted how PRM information would support them in their daily lives. For example, this information would enable them to work during treatment and manage childcare responsibilities.

> *To know if I would be able to go to work, if I could get two-three days sick leave, to schedule my life, my daily routine, to take painkillers beforehand.* [site 5, patient 4]

> *I think, for me, it would have allowed me to plan a lot better. Because I'm at home with just my son, I'm a single parent. And it would have allowed me to get, or to have more support around me.* [site 1, patient 4]

Educating Friends and Family Members

An interesting perspective was the way that a few patients spoke about how they would share this information to help educate friends and family members. This risk prediction would convey to others what symptoms they, the patient, would likely

XSL•FO
RenderX

experience and allow those individuals who were close to them to prepare themselves in advance:

> *Then the person you're living with can also be prepared.* [site 7, patient 1]
>
> *It's never just about you… you're bringing people with you, partners, family, parents and whatever. And having to explain these are the things I might suffer from, they're able to see it visually... that might be useful.* [site 4, patient 5]

Negative Perceptions of Predicting Risk

Overview

In addition to positive perceptions of PRMs, several patients had some negative perceptions of them.

Skepticism About the Concept

Some patients were skeptical of the overall concept. They did not believe that it was possible to accurately predict the likelihood of CRS:

> *It sounds a bit like science fiction that you could say this in advance.* [site 6, patient 1]
>
> *I mean, they're [PRMs] not going to say you are definitely going to get that [symptom], can they?* [site 1, patient 1]

Negative Mindset

Other patients spoke about how PRM information could be detrimental to their well-being and result in a negative mindset. These patients preferred taking an uninformed approach to CRS:

> *See that's the deal, some [people] could be completely negatively affected with this information, and others are not affected at all. And if the probability is that you will be affected by something with an average risk of 90%, of course it's going to impact your mind negatively.* [site 6, patient 2]
>
> *It could be a worry more than desired, as far as I'm concerned.* [site 7, patient 2]

Self-Fulfilling Prophecy

Linked to these perceptions of PRMs creating a negative mindset, a few patients spoke about predicting risk as a self-fulfilling prophecy. They perceived that being given personalized predictions of likely CRS would affect them psychologically and result in them experiencing these symptoms simply because they were told that they were at increased risk:

> *Beforehand? I would not want that, because I am easily influenced and I would experience them, even if I was not meant to. From a psychological point of view alone, I would not want that, no.* [site 5, patient 2]
>
> *If you're told you're going to be ill on a certain day…you think…oh, I don't feel good.* [site 1, patient 3]

Using the Data From PRMs

Visualizing the Data

Visual aids provided examples of how PRM information about CRS could be presented (Figure 1). When asked, patients expressed their preferences for different approaches that could be used to visualize and facilitate their understanding of PRMs. Several patients spoke positively about risk predictions being presented within a cluster of symptoms, seeing the benefit in visualizing how their predicted symptoms "joined up." They felt that this type of visual aid would help them understand why they experienced certain groups of symptoms and their interdependencies:

> *I quite like it because it shows that if you get one, you're more, you know, like the energy, the drowsiness, the, uh, difficulty in sleeping and the fact that they all go together.* [site 1, patient 1]
>
> *It was partially related and depending on one another, because once you have nausea and constipation, you lose your appetite.* [site 6, patient 3]

Conversely, some patients commented that the cluster visualization was too complicated and better suited for clinicians:

> *For me, presented in this manner, it just doesn't help me, it doesn't make sense.* [site 1, patient 4]
>
> *I mean if this is for the doctors then fine, but I wouldn't like to have it as a patient.* [site 3, patient 4]

Some patients suggested that the risk predictions could be presented in a simpler format and proposed a "traffic light" priority system that would enable them to decipher what symptoms were problematic and what ones were not:

> *I think…the traffic light system is a good system. Because red, we all know that red is danger and not good, so if you have got something in there what you might get it's something that you will look out for.* [site 3, patient 1]

Frequency of Data

In terms of the timing of the receipt of PRM data, most patients favored receiving information at the start of each cycle of chemotherapy—with a sense that being given daily information would be overwhelming. Having new information with each cycle of chemotherapy was perceived as useful because patients acknowledged that symptom experiences could change over time:

> *I think, for me, it would be too much for me to handle if I'm getting a……for example, if it's a text or alert coming through ... today you might vomit [laughter]. Great! Have a nice day [laughter]!* [site 1, patient 4]
>
> *After every treatment there are more symptoms you see. It takes a while for the symptoms to manifest themselves because it is a build-up I suppose… I do think it is a build-up and the tiredness, that's a build-up as well, at first you're not too bad and then…*

XSL•FO
RenderX

> *but yeah, I would like to know [symptom information] ongoing.* [site 4, patient 4]

> *I think having such a prognosis once per month, at the beginning of the month, that would be enough. It should just contain things like, 'This can happen in this timeframe'. And the second month there is a probability that other symptoms occur. That you have this in advance but not get bombarded with this information all the time.* [site 6, patient 3]

Preferences for Technology

Patients identified the usefulness of having access to the visual representation of risk data on mobile technology (eg, smartphones). They suggested that being able to interact with the data (eg, tap on a symptom to view self-care information) would add value for them:

> *And I can imagine this to be on a smartphone because everyone has a smartphone anyway. And you see a notification with a new message. And you get this maybe every 2 weeks, 4 weeks, or so. And if this happens, the probability is this and that, take this and that.* [site 6, patient 4]

> *I like the idea of being able to either hover over it or tap on a certain thing…and get more information… it's that positive reinforcement of you're doing the right thing. So, it's good to have that little bit of background, that little bit of information that is kind of either telling you that or telling you something a little bit new.* [site 1, patient 2]

However, despite all of their positive suggestions about the usefulness of PRM information, patients still wanted the reassurance of knowing they could contact a clinician when required:

> *And I am assuming that the app wouldn't sort of take away from the fact that there is a person that you can call if you wanted to? Because if it ruined that possibility then it would be a bit worrying I think.* [site 3, patient 2]

Clinicians' Themes

Analysis of the data identified 3 main and several subthemes from the clinicians' focus groups (Table 4). Each theme is presented in greater depth in Table 4.

Table 4. Clinicians' themes and subthemes.

Themes	Subthemes
Positive perceptions of predicting risk	• Anticipatory, preventative, and targeted care • "Unique" patient education
Negative perceptions of predicting risk	• Trust (in the concept) • No additional benefits of artificial intelligence
Use of predictive risk data	• Visualizing the data • Frequency of data • Preferences for technology

Positive Perceptions of Predicting Risk

Overview

Clinicians perceived several ways in which PRMs would positively influence their practice.

Anticipatory, Preventative, and Targeted Care

According to the clinicians, one of the main benefits of having PRM information was the way it could promote a more anticipatory and preventative approach to CRS. This approach could decrease patients' distress and enable them to feel more in control during treatment:

> *I can recognise problems earlier and intervene earlier, in the sense that I can prevent problems instead of solving them. So I don't want to wait until the patient has to go the ER on the weekend, but I can take preventive measures in terms of higher dosage and medication management. It's useful and saves me a lot of work and saves the patient pain and eventually saves the entire medical system costs.* [site 6, clinician 1]

> *If the information is there and it's predicted then you can actually reduce the possibility of unpleasant experiences for patients. Because once they experience one unpleasant experience then the subsequent treatment will become quite intolerable for them, because they anticipate the undesirable reaction.* [site 2, clinician 2]

In addition, clinicians suggested that PRM information would enable them to better target and manage the side effects of treatment, optimize patient outcomes, and create efficiencies in services:

> *Knowing the severity levels and the burdening levels with individual patients, and what the stress is, so these three components, it is quite targeted. Because if I take any action, I can have the most beneficial effect, so on one side it is a very efficient measure, and on the other side, it has the advantage that I can prove better that it really is useful. And with this it is easier to divide the resources, and can save resources to be able to target care to this certain patient group.* [site 6, clinician 3]

XSL•FO
RenderX

> *For example, if we knew ahead of time that they [the patient] was at a higher risk for permanent hair loss then we might say encourage scalp cooling before they start, in case the scalp cooling helps save what hair they have….because, we have had some ladies in the past who have had permanent hair loss and they've said that had they known ahead of time they would never have chosen the adjuvant chemotherapy, because, it [hair loss] restricts their life so much now because of their distress over the symptom.* [site 2, clinician 1]

"Unique" Patient Education

The availability of specific predictions of CRS was perceived by clinicians as an effective way to provide patients with individualized and targeted education about self-care. Clinicians viewed this personalized approach as having the potential to be more effective than general standardized approaches:

> *This is very important as far as the patient counselling goes. That you don't have to tell them about all the possible symptoms that can happen, but instead, just thoroughly discuss the ones, three let's say, that are most likely in his/her case. And just focus on those, and certainly, if some of the others come up, they can get it touch to discuss those. It's better for the patient because they would go home not with less information but more targeted information.* [site 6, clinician 3]

> *I would want that [PRM] information because it would influence the information that I then needed to give them [the patient] to help prepare. Because right now, say in a pre-chemotherapy setting we have to tell them about all the different side effects they might get…But, that means that some patients like in the past have come back and have said that was much worse than I expected or that was not as I expected… You know, because we are gauging it to, not specifically for them.* [site 2, clinician 1]

Negative Perceptions of Predicting Risk

Overview

Despite their positive perceptions of PRMs, clinicians had some negative comments about the use of PRMs in clinical practice.

Trust (in the Concept)

Similar to the patients' perspectives, some clinicians were skeptical that PRMs would be able to accurately predict individual patient's symptom experiences. They questioned whether what was proposed was actually possible. Others questioned the evidence base for these types of models and their reliability.

> *I don't know until how far a predictive model could go. Until which point can we predict? Until which phase of therapy and patient monitoring we can predict? Can we do that?* [site 5, clinician 2]

> *I don't know if I could see it working* [site 4, clinician 4]

While recognizing several benefits, a few clinicians raised concerns that these predictions would result in patients having increased anxiety which resulted in more frequent contacts with clinical services.

> *I think there can sometimes be a certain aspect of…if you're saying to a patient in their pre-chemo consult this is a 90% likelihood that this is going to make you sick, they're going to be hypervigilant for that. And, any sign kind of sign of sickness they are going to be straight on the phone and feeling more anxious about it.* [site 1, clinician 3]

> *It has happened to me, that a patient, a young one, even though young people usually ask more questions, and she told me that she did not want to know, because she would be expecting the symptoms.* [site 5, clinician 6]

Unable to Deal With all of the Complexity

The ability of PRMs to deal with the complexity of cancer care and the multiple factors that contribute to patients' symptom experiences was questioned by some clinicians. Some clinicians suggested that the models would not account for human complexity:

> *Whether it is because of the complexity of the cancer patients, not all of them are well before they start treatment, by the time they start treatment they could be a second line, or third line treatment, they are so unwell. So, my concern would be that those predictors occurrence and so on, would they disregard the general health before the treatment?* [site 2, clinician 2]

> *I don't think you can predict [symptom] distress because it depends upon the person's coping and ability to cope with what is happening.* [site 2, clinician 1]

Clinicians doubted that knowledge derived from PRMs could equal the skills of an expert oncology clinician to adequately identify and predict the complex issues associated with caring for patients with cancer:

> *The only problem with algorithms and AI is that it needs to learn, at the moment it is not complex enough to actually deal with a good chemotherapy nurse. And also, it doesn't take into account at the moment the patient holistic baselines.* [site 1, clinician 3]

Several clinicians stated that they already followed anticipatory and preventative models of care provision for symptoms. Therefore, they questioned whether PRMs would add any value to current or future clinical practice:

> *I think it is used already…in the pre-treatment consultations to target what patients should be aware of, that are most likely to happen in a lot of cases. We use it as well when someone rings up if we know what treatment they're having, then we can see what is normal.* [site 1, clinician 2]

> *We do work preventatively; we do that with many.* [site 7, clinician 3]

XSL•FO
RenderX

> *I tend to work with that anyway, with the kind of evidence base on the side effects that we know occur with chemotherapy. And, I suppose from clinical experience then you know that certain patients will have more severe symptoms if that makes sense.* [site 1, clinician 1]

Using the Data From PRMs

Visualization of the Data

Similar to patients' perspectives, clinicians were able to see the benefits of presenting the risk predictions within a cluster of related symptoms as an acknowledgment of the relationships between or among specific symptoms.

> *I think it's quite useful, it might help with your assessment something like this. Or if somebody rings up with say taste change, kind of helps you to structure the questions so you might ask around it. So for example, is it affecting your appetite, have you lost any weight, is your mouth dry, you know, so it does help with that assessment actually.* [site 1, clinician 3]

A few of the clinicians liked the way that risk predictions of CRS as a cluster could be presented based on the different dimensions of the symptom experience (eg, severity and distress):

> *What we see a lot, is the data about the severity [of symptoms] in a lot of scientific work, but about the distress, we actually don't know very much.* [site 4, clinician 2]

However, concern was expressed about the challenges of viewing predictions of the risk of chemotherapy symptoms based on their distress alone:

> *So, it can actually be quite deceiving to see what distresses the patient, because to a patient maybe hair loss is the most important factor, but yet for a healthcare professional, the diarrhoea …they may have actually under-estimated the diarrhoea that could be actually life threatening… they have been having diarrhoea for so long but they didn't think it was as important as the hair loss.* [site 2, clinician 2]

Being able to visualize a higher level of risk for a specific symptom within a cluster by the size of its "bubble" (Figure 1) and using the size of the bubbles and links to decipher levels of interdependencies between associated symptoms was viewed as a positive aspect of this approach:

> *…you would pick the three biggest bubbles if that is all you could address and then they will show you the relationship between the larger bubbles and other things that are influencing it. So, in that sense it is helpful. Right, because you can see what's maybe influencing the larger bubbles.* [site 2, clinician 1]

However, concerns were voiced that focusing on symptoms within a cluster that were depicted as having the highest level of risk, may draw attention away from more important or life-threatening symptoms:

> *So if my attention is being drawn to the things that have the biggest bubble then I am missing the subtle things that are happening that I actually need to pay attention to, because of the type of treatment.* [site 2, clinician 1]

Clinicians recognized that being able to visualize predictions would be helpful for less experienced clinicians during their interactions with patients:

> *I think it is useful when you are starting out in a role like this, when you don't have the experience. I would have found something like this useful, yes.* [site 1, clinician 3]

> *I have noticed a lot of new nurses coming on, they quite like models of care. They actually look at temperature and follow the flowchart and not actually looking at the patient. So some [new] staff might actually like this predictive model because it does help them.* [site 2, clinician 2]

Clinicians recommended that predictive risk information be presented in a simple, visual, way, that patients could easily use and that was intuitive to understand:

> *So visualisation to quickly see what it is. Visual graphic information goes very quickly into the brain, so if we could have a 'flash' graphic display, that would be cool.* [site 6, clinician 1]

> *I think it would be good if you could see the whole cycle and then be able to like narrow down on each day. Then you could see change in symptoms, those sort of things or measure against it because they might experience more and more severe at this stage.* [site 1, clinician 3]

> *Traffic light system because it is more perceptual, isn't it? So green is safe, amber is in between, and red is not so good, so it might be something like that.* [site 2, clinician 2]

Frequency of Data

When asked about the frequency of presentations of the predictive risk information, clinicians stated that this information would be most helpful if it were tailored and provided at each cycle of chemotherapy:

> *I think it is more important for the patients to know at which days during which cycle the symptoms can occur. For example, the first 5-7 days fatigue, such things, so the patient can prepare himself, and orientate themselves towards it. That would be most beneficial to the patients.* [site 6, clinician 3]

Preferences for Technology

Interactivity and incorporation of PRMs into technology was identified by some clinicians as a potential way to integrate this information in practice:

> *We are in the digital age now... Then you can have the information where you just click, like a tailored bit, this is day two there is a symptom you're highly likely have that.* [site 1, clinician 4]

XSL•FO
RenderX

> *The more graphic the better, a graphic overview, a running chart and having distinct indicators for on-edge situations. Having an alarm-like indicator and a pre-alarm indicator, so to speak.* [site 6, clinician 2]

Discussion

Principal Results

The focus groups and interviews with patients with cancer and clinicians across 5 European countries generated valuable and novel insights into their perceptions about the utility and usefulness of CRS predictions in daily life and clinical practice. Patients and clinicians commented on the practical and clinical advantages of using PRMs to prevent and manage CRS. However, with some reservations about the validity of PRMs, they highlighted that how this information was presented would impact its usefulness and usability in clinical practice. Both groups suggested alternative ideas for how to visualize PRM information. They emphasized the importance of simple and user-friendly formats to maximize benefits. These types of explorations of data visualization and associated usefulness and usability of PRM information for CRS management warrant additional investigation. Given that no differences in perceptions, positive or negative, were evident among patients, clinicians, clinical sites, and countries suggest that perceptions about the use of PRMs for CRS are similar.

Comparison With Prior Work

Patients' Perspectives

Patients perceived that predictive risk information would help them manage expectations, enable them to proactively plan, prepare, and engage in self-care activities, feel in control, and educate family members and friends about the symptoms they were most likely to experience. These findings are consistent with studies that reported positive perceptions of the use of approaches such as machine learning in health care [39]. Providing patients with accurate information that allows them to anticipate symptoms, and engage in coping strategies to minimize their impact, has the potential to reduce patients' distress [40] and experiences of anxiety and depression [41] as well as increase their satisfaction with information [42]. In addition, increasing patients' confidence in performing self-care improves their QoL during chemotherapy [43]. Therefore, the provision of individualized predictive risk information about likely symptoms to patients is likely to improve a range of patient outcomes and experiences. However, this approach may not be useful for all patients.

Some patients were skeptical that the PRMs would be able to accurately predict their symptoms. In addition, some commented that these predictions would negatively impact on their lives. Considering the high levels of anxiety and depression that some patients experience [44], clinicians need to assess whether patients want this type of information during their treatment [45]. Patient-centered communication is linked to greater patient satisfaction and needs to be tailored to individual patient's needs and preferences [46,47]. Clinicians need to work with patients to explain the potential benefits of PRMs and determine their preference for using this type of information. Given that a trusting relationship between patients and clinicians facilitates communication and health care decision-making, reduces patients' fears, and results in increased adherence to treatment [48], a shift in trust toward technology may be required. Patients may be more likely to view PRMs positively if they feel that their clinicians also do so.

Clinicians' Perspectives

Clinicians were positive about the information derived from PRMs of CRS, particularly in the delivery of anticipatory, preventative, and targeted cancer care. They appreciated the way that they would be able to use this information to provide truly personalized patient education. Given that the burden of cancer care will increase, the ability to predict, identify, and target patients who are at increased risk of a higher symptom burden will assist with the allocation of health care resources. The oncology specialty is facing several concurrent challenges including an affordability crisis [49] and challenges from the COVID-19 pandemic (eg, delayed and altered treatment pathways and reductions in specialist reviews) that are associated with poorer patient outcomes [50]. Given these growing pressures on various health care systems, the cancer community needs to work collaboratively to find ways to mitigate the potential negative impacts and make use of the best available resources to deliver the highest quality cancer care. The use of PRMs of CRS to plan patients' care and support during chemotherapy provides 1 way to facilitate the most appropriate allocation of resources.

However, as with some patients, some clinicians expressed concerns about the ability of PRMs to accurately predict symptom experiences. Trust is a crucial factor that influences human beings' interactions with technology [25]. For digital health care technologies to be accepted as trustworthy and evidence-based, they should be based on robust, resilient, reliable, and effective systems [51]. However, given that a lack of trust in technology by both patients and clinicians continues, additional research on the human factors of technology is required [52]. While providing promising opportunities to support clinicians to achieve the best outcomes for patients, methods to ensure that clinicians view predictive risks for CRS as trusted reliable sources of information warrant additional research and education. Patients and clinicians need to be able to trust that the technologies approved for use in health care systems meet robust standards [51,53]. Consequently, a need exists to apply a framework of trust in health-related machine learning based on sound research and conceptual rigor to ensure that developments, such as risk prediction, will be accepted and implemented in clinical practice [54].

Some clinicians questioned whether PRMs provided any additional information. They expressed that, as expert clinicians, they were able to predict symptoms and initiate preventative and timely interventions. An attitudinal shift is necessary, acknowledging the opportunities offered through health-related technology, that can be used to augment clinical expertise for the benefit of all those who provide or deliver health care. For example, endorsement by a trusted authority like the National

XSL•FO
RenderX

Institute for Health and Care Excellence, may facilitate acceptance of PRMs by clinicians [53].

Future Implications—the Importance of Data Visualization

In this study, patients valued the potential of visualizing how symptoms linked with each other. This visualization helped them understand why certain symptoms occurred together. In addition, for clinicians, the visual clustering of symptoms was perceived to be helpful for their assessment processes. However, participants noted that the visual examples of PRM information that they viewed during focus groups and interviews needed to be revised and made less complex. Given that information may influence patients' decisions and outcomes, clinicians need to use representations that increase patients' insights and turn data into something meaningful, relevant, and useful. This approach can be a platform for informed discussions about the significance and burden of risks and the implications for an individual and family members [55,56].

The optimization of visualizations of PRMs for CRS is key because it will improve patients' and clinicians' understanding of this information and allow them to see symptom patterns and relationships between or among symptoms more clearly. In addition, this increased understanding will enhance their decision-making abilities about symptoms and what proactive measures to take to manage symptoms in a timely fashion. Optimization of visualizations and understanding of predictive risks have the potential to improve patients' engagement in their care and facilitate communication and collaboration between patients and clinicians.

Sharing complex health-related information in a way that ensures that users can understand and act on it is a challenge in the field of digital health data [42]. Data visualization, and in particular data visualization that patients can easily use, is a more effective approach to communicating key messages and increasing patient and clinician understanding and engagement [57,58]. Indeed, concerning the PRMs for CRS evaluated in this study, patients' and clinicians' preferences for the way forward included visual, usable, *and* informative formats. The development of these types of usable and useful visualizations of CRS has the potential to result in profound and positive changes in patients' experiences, as well as the provision of individualized supportive care delivery. Future work needs to co-design the presentation of PRM information with clinicians, patients, and family caregivers to ensure that the way that information is presented is engaging, meaningful, usable, and useful for all of the intended recipients [59].

Strengths and Limitations

This study was conducted across 5 European countries. The consistent positive and negative perspectives of patients and clinicians from different countries who have different cultural backgrounds and lived experiences strengthen the generalizability of these findings. That said, it is important to note that the findings may not apply to countries outside of Europe or to different cancer diagnoses and treatments.

Another strength of this study is the fact that patients' perceptions of the usability and usefulness of PRMs were grounded in real-life experiences as this study recruited patients who had received chemotherapy. While their comments may have been influenced by the severity of their previous symptom experiences, they could draw on their personal knowledge to visualize ways in which this type of information would have an impact on them (both positively and negatively) during their chemotherapy treatment, validating this study's results.

The decision on whether to use visual prompts during the focus groups was carefully considered in case so doing would influence participants' suggestions about ways in which this information could be presented. Some of the visual aids were necessary because they depicted real data, while other images facilitated discussion about novel concepts. The individuals who conducted the focus groups were advised to "hold off" on showing the visual aids if conversations and ideas were flowing well and to use them "when necessary" to explain novel concepts. Given that all researchers opted to use all of the visual aids suggests that they were useful to facilitate meaningful discussions.

A limitation of this study is the lack of information about clinicians' level and duration of experience with chemotherapy services. However, given that the roles of these participants (eg, chemotherapy nurse specialist, nurse consultant, and head nurse) reflect a level of seniority, it is possible that the interviewees were highly skilled and experienced in chemotherapy treatment and care. These clinicians may have different perspectives about PRMs compared to less experienced colleagues. Comparing perspectives about PRMs among clinicians with different levels of clinical experience may provide new and valuable insights.

Another limitation is that the data were not gathered in a way that allowed for subgroup analysis. Perceptions may be influenced by other factors (eg, health or technical literacy) that were not assessed in this study. However, participants provided positive and negative comments about PRMs. Future research needs to explore potential differences in perceptions held by various subgroups.

The inclusion of family caregivers in future research may provide additional insights about the use of PRMs. In this study, patients identified that a potential benefit of PRMs would be to help them explain their expected symptom experiences to family members and friends. Enhanced understanding of symptom patterns using technology has empowered family members to plan and maintain their own lives during their relative's chemotherapy [60].

Conclusions

This study provides novel insights into the perceptions of patients and clinicians about the clinical usefulness and usability of PRMs in the management of CRS. While the potential benefits of PRMs for CRS were recognized, additional strategies to increase users' confidence and trust in the ability of PRMs to support tailored and predictive interventions and care need to be developed and implemented. The presentation of PRM information in usable, useful, and engaging ways would be a way to facilitate implementation, optimize its uptake, and make it meaningful, relevant, and useful to patients and clinicians as well as family caregivers.

XSL•FO
RenderX

Additional research that is focused on building and validating the processes to develop PRMs is essential. However, advances in the analytical processes must be matched with how this knowledge can be integrated into clinical practice. While incorporating the use of PRMs to provide person-centered, effective management of CRS is important, clinicians must maintain the human side of care that includes the fundamental values of empathy, compassion, and trust [61].

Acknowledgments

This study was funded by the European Commission (EC; via the FP7-HEALTH-2013-INNOVATION-1 grant award scheme (602289). The trial was sponsored by the University of Strathclyde. This study's team would like to thank all those patients and clinicians across all clinical sites (Medical University Vienna Comprehensive Cancer Centre, Austria; Agioi Anargiri Cancer Hospital, Greece; Metropolitan Hospital, Greece; Air Force General Hospital, Greece; St James' Hospital, Ireland; St Vincent's Private Hospital, Ireland; St Vincent's University Hospital, Ireland; Innlandet Hospital Trust Division Gjøvik-Lillehammer, Norway; Guy's and St Thomas' NHS Foundation Trust, United Kingdom; Mount Vernon Cancer Centre, United Kingdom; St George's Healthcare Trust, United Kingdom; and University College London Hospitals, United Kingdom) who gave their time to participate in data collection.

Authors' Contributions

All authors contributed to the conceptualization, writing, and editing of the paper.

Conflicts of Interest

None declared.

References

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M. Global cancer observatory: cancer today. International Agency for Research on Cancer. 2020. URL: https://gco.iarc.fr/today/en [accessed 2024-03-15]
2. The burden of cancer. The Cancer Atlas. URL: https://canceratlas.cancer.org/the-burden/the-burden-of-cancer/ [accessed 2023-03-23]
3. Phillips AA, Smith DA. Health disparities and the global landscape of lymphoma care today. Am Soc Clin Oncol Educ Book 2017;37:526-534 [FREE Full text] [doi: 10.1200/EDBK_175444] [Medline: 28561692]
4. Tong CWS, Wu M, Cho WCS, To KKW. Recent advances in the treatment of breast cancer. Front Oncol 2018;8:227 [FREE Full text] [doi: 10.3389/fonc.2018.00227] [Medline: 29963498]
5. Xie YH, Chen YX, Fang JY. Comprehensive review of targeted therapy for colorectal cancer. Signal Transduct Target Ther 2020;5(1):22 [FREE Full text] [doi: 10.1038/s41392-020-0116-z] [Medline: 32296018]
6. Pathak R, Giri S, Aryal MR, Karmacharya P, Bhatt VR, Martin MG. Mortality, length of stay, and health care costs of febrile neutropenia-related hospitalizations among patients with breast cancer in the United States. Support Care Cancer 2015;23(3):615-617. [doi: 10.1007/s00520-014-2553-0] [Medline: 25556610]
7. Pearce A, Haas M, Viney R, Pearson SA, Haywood P, Brown C, et al. Incidence and severity of self-reported chemotherapy side effects in routine care: a prospective cohort study. PLoS One 2017;12(10):e0184360 [FREE Full text] [doi: 10.1371/journal.pone.0184360] [Medline: 29016607]
8. Nurgali K, Jagoe RT, Abalo R. Editorial: adverse effects of cancer chemotherapy: anything new to improve tolerance and reduce sequelae? Front Pharmacol 2018;9:245 [FREE Full text] [doi: 10.3389/fphar.2018.00245] [Medline: 29623040]
9. McKenzie H, Hayes L, White K, Cox K, Fethney J, Boughton M, et al. Chemotherapy outpatients' unplanned presentations to hospital: a retrospective study. Support Care Cancer 2011;19(7):963-969. [doi: 10.1007/s00520-010-0913-y] [Medline: 20499108]
10. Patel MI, Williams DC, Wohlforth C, Fisher G, Wakelee HA, Blayney DW. Are patients with thoracic malignancies at risk for uncontrolled symptoms? J Oncol Pract 2015;11(1):e98-e102 [FREE Full text] [doi: 10.1200/JOP.2014.001502] [Medline: 25271246]
11. Trajkovic-Vidakovic M, de Graeff A, Voest EE, Teunissen SCCM. Symptoms tell it all: a systematic review of the value of symptom assessment to predict survival in advanced cancer patients. Crit Rev Oncol Hematol 2012;84(1):130-148. [doi: 10.1016/j.critrevonc.2012.02.011] [Medline: 22465016]
12. Puts MTE, Tu HA, Tourangeau A, Howell D, Fitch M, Springall E, et al. Factors influencing adherence to cancer treatment in older adults with cancer: a systematic review. Ann Oncol 2014;25(3):564-577 [FREE Full text] [doi: 10.1093/annonc/mdt433] [Medline: 24285020]
13. Sommariva S, Pongiglione B, Tarricone R. Impact of chemotherapy-induced nausea and vomiting on health-related quality of life and resource utilization: a systematic review. Crit Rev Oncol Hematol 2016;99:13-36. [doi: 10.1016/j.critrevonc.2015.12.001] [Medline: 26697988]

XSL•FO
RenderX

14. Binotto M, Reinert T, Werutsky G, Zaffaroni F, Schwartsmann G. Health-related quality of life before and during chemotherapy in patients with early-stage breast cancer. Ecancermedicalscience 2020;14:1007 [FREE Full text] [doi: 10.3332/ecancer.2020.1007] [Medline: 32104209]
15. Jawed I, Wilkerson J, Burotto Pichun ME, Duffy AG, Fojo AT. Effect of chemotherapy on the progress in colorectal cancer survival in the past two decades. J Clin Oncol 2014;32(15_suppl):3642 [FREE Full text] [doi: 10.1200/jco.2014.32.15_suppl.3642]
16. Nurgali K, Jagoe RT, Abalo R. Editorial: adverse effects of cancer chemotherapy: anything new to improve tolerance and reduce sequelae? Front Pharmacol 2018;9:245 [FREE Full text] [doi: 10.3389/fphar.2018.00245] [Medline: 29623040]
17. Duijts SFA, van Egmond MP, Spelten E, van Muijen P, Anema JR, van der Beek AJ. Physical and psychosocial problems in cancer survivors beyond return to work: a systematic review. Psycho-Oncology 2014;23(5):481-492. [doi: 10.1002/pon.3467] [Medline: 24375630]
18. Götze H, Taubenheim S, Dietz A, Lordick F, Mehnert A. Comorbid conditions and health-related quality of life in long-term cancer survivors: associations with demographic and medical characteristics. J Cancer Surviv 2018;12(5):712-720. [doi: 10.1007/s11764-018-0708-6] [Medline: 30097854]
19. Kline RM, Arora NK, Bradley CJ, Brauer ER, Graves DL, Lunsford NB, et al. Long-term survivorship care after cancer treatment: summary of a 2017 National Cancer Policy Forum workshop. J Natl Cancer Inst 2018;110(12):1300-1310 [FREE Full text] [doi: 10.1093/jnci/djy176] [Medline: 30496448]
20. Papachristou N, Puschmann D, Barnaghi P, Cooper B, Hu X, Maguire R, et al. Learning from data to predict future symptoms of oncology patients. PLoS One 2018;13(12):e0208808 [FREE Full text] [doi: 10.1371/journal.pone.0208808] [Medline: 30596658]
21. Kolankiewicz ACB, de Souza Magnago TSB, Dos Santos Dullius AI, De Domenico EBL. Association of demographic, economic and clinical variables in daily activities and symptoms presented by patients in cancer treatment. Can Oncol Nurs J 2017;27(4):365-374 [FREE Full text] [doi: 10.5737/23688076274365374] [Medline: 31148740]
22. George M, Smith A, Sabesan S, Ranmuthugala G. Physical comorbidities and their relationship with cancer treatment and its outcomes in older adult populations: systematic review. JMIR Cancer 2021;7(4):e26425 [FREE Full text] [doi: 10.2196/26425] [Medline: 34643534]
23. Coolbrandt A, Van den Heede K, Vanhove E, De Bom A, Milisen K, Wildiers H. Immediate versus delayed self-reporting of symptoms and side effects during chemotherapy: does timing matter? Eur J Oncol Nurs 2011;15(2):130-136. [doi: 10.1016/j.ejon.2010.06.010] [Medline: 20685164]
24. Goodman K, Zandi D, Reis A, Vayena E. Balancing risks and benefits of artificial intelligence in the health sector. Bull World Health Organ 2020;98(4):230-230A [FREE Full text] [doi: 10.2471/BLT.20.253823] [Medline: 32284640]
25. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
26. Falconer N, Barras M, Cottrell N. Systematic review of predictive risk models for adverse drug events in hospitalized patients. Br J Clin Pharmacol 2018;84(5):846-864 [FREE Full text] [doi: 10.1111/bcp.13514] [Medline: 29337387]
27. Wu R, Ai S, Cai J, Zhang S, Qian ZM, Zhang Y, et al. Predictive model and risk factors for case fatality of COVID-19: a cohort of 21,392 cases in Hubei, China. Innovation (Camb) 2020;1(2):100022 [FREE Full text] [doi: 10.1016/j.xinn.2020.100022] [Medline: 33521759]
28. Maguire R, Cowie J, Leadbetter C, McCall K, Swingler K, McCann L, et al. The development of a side effect risk assessment tool (ASyMS©-SERAT) for use in patients with breast cancer undergoing adjuvant chemotherapy. J Res Nurs 2009;14(1):27-40. [doi: 10.1177/1744987108099235]
29. Papachristou N, Barnaghi P, Cooper BA, Hu X, Maguire R, Apostolidis K, et al. Congruence between latent class and K-modes analyses in the identification of oncology patients with distinct symptom experiences. J Pain Symptom Manage 2018;55(2):318-333.e4 [FREE Full text] [doi: 10.1016/j.jpainsymman.2017.08.020] [Medline: 28859882]
30. Papachristou N, Barnaghi P, Cooper B, Kober KM, Maguire R, Paul SM, et al. Network analysis of the multidimensional symptom experience of oncology. Sci Rep 2019;9(1):2258 [FREE Full text] [doi: 10.1038/s41598-018-36973-1] [Medline: 30783135]
31. Sujan M, Furniss D, Grundy K, Grundy H, Nelson D, Elliott M, et al. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform 2019;26(1):e100081 [FREE Full text] [doi: 10.1136/bmjhci-2019-100081] [Medline: 31780459]
32. Kurniawan J, Duffy VG. Systematic review of the importance of human factors in incorporating healthcare automation. In: Duffy VG, editor. Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: AI, Product and Service. Cham, Switzerland: Springer; 2021:96-110.
33. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
34. Fox P, Darley A, Furlong E, Miaskowski C, Patiraki E, Armes J, et al. The assessment and management of chemotherapy-related toxicities in patients with breast cancer, colorectal cancer, and Hodgkin's and non-Hodgkin's

XSL•FO
RenderX

lymphomas: a scoping review. Eur J Oncol Nurs 2017;26:63-82 [FREE Full text] [doi: 10.1016/j.ejon.2016.12.008] [Medline: 28069154]
35. Maguire R, Fox PA, McCann L, Miaskowski C, Kotronoulas G, Miller M, et al. The eSMART study protocol: a randomised controlled trial to evaluate electronic symptom management using the advanced symptom management system (ASyMS) remote technology for patients with cancer. BMJ Open 2017;7(5):e015016 [FREE Full text] [doi: 10.1136/bmjopen-2016-015016] [Medline: 28592577]
36. Furlong E, Darley A, Fox P, Buick A, Kotronoulas G, Miller M, et al. Adaptation and implementation of a mobile phone-based remote symptom monitoring system for people with cancer in Europe. JMIR Cancer 2019;5(1):e10813 [FREE Full text] [doi: 10.2196/10813] [Medline: 30869641]
37. Maguire R, McCann L, Kotronoulas G, Kearney N, Ream E, Armes J, et al. Real time remote symptom monitoring during chemotherapy for cancer: European multicentre randomised controlled trial (eSMART). BMJ 2021;374:n1647 [FREE Full text] [doi: 10.1136/bmj.n1647] [Medline: 34289996]
38. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
39. Tran VT, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Digit Med 2019;2(1):53 [FREE Full text] [doi: 10.1038/s41746-019-0132-y] [Medline: 31304399]
40. Gibbons A, Groarke A. Coping with chemotherapy for breast cancer: asking women what works. Eur J Oncol Nurs 2018;35:85-91. [doi: 10.1016/j.ejon.2018.06.003] [Medline: 30057089]
41. D'Souza V, Blouin E, Zeitouni A, Muller K, Allison PJ. An investigation of the effect of tailored information on symptoms of anxiety and depression in head and neck cancer patients. Oral Oncol 2013;49(5):431-437. [doi: 10.1016/j.oraloncology.2012.12.001] [Medline: 23295073]
42. O'Connor G, Coates V, O'Neill S. Randomised controlled trial of a tailored information pack for patients undergoing surgery and treatment for rectal cancer. Eur J Oncol Nurs 2014;18(2):183-191. [doi: 10.1016/j.ejon.2013.10.011] [Medline: 24275208]
43. Akin S, Kas Guner C. Investigation of the relationship among fatigue, self-efficacy and quality of life during chemotherapy in patients with breast, lung or gastrointestinal cancer. Eur J Cancer Care (Engl) 2019;28(1):e12898. [doi: 10.1111/ecc.12898] [Medline: 30039883]
44. Pitman A, Suleman S, Hyde N, Hodgkiss A. Depression and anxiety in patients with cancer. BMJ 2018;361:k1415 [FREE Full text] [doi: 10.1136/bmj.k1415] [Medline: 29695476]
45. Patch C, Middleton A. Genetic counselling in the era of genomic medicine. Br Med Bull 2018;126(1):27-36 [FREE Full text] [doi: 10.1093/bmb/ldy008] [Medline: 29617718]
46. Gilligan T, Coyle N, Frankel RM, Berry DL, Bohlke K, Epstein RM, et al. Patient-clinician communication: American society of clinical oncology consensus guideline. J Clin Oncol 2017;35(31):3618-3632 [FREE Full text] [doi: 10.1200/JCO.2017.75.2311] [Medline: 28892432]
47. Lehmann V, Labrie NHM, van Weert JCM, van Dulmen S, de Haes HJCJM, Kersten MJ, et al. Tailoring the amount of treatment information to cancer patients' and survivors' preferences: effects on patient-reported outcomes. Patient Educ Couns 2020;103(3):514-520. [doi: 10.1016/j.pec.2019.09.024] [Medline: 31585818]
48. Hillen MA, de Haes HCJM, Smets EMA. Cancer patients' trust in their physician: a review. Psycho-Oncology 2011;20(3):227-241. [doi: 10.1002/pon.1745] [Medline: 20878840]
49. Wilking N, Lopes G, Meier K, Simoens S, van Harten W, Vulto A. Can we continue to afford access to cancer treatment? Eur Oncol Haematol 2017;13(2):114-119. [doi: 10.17925/eoh.2017.13.02.114]
50. Chazan G, Franchini F, Alexander M, Banerjee S, Mileshkin L, Blinman P, et al. Impact of COVID-19 on cancer service delivery: a follow-up international survey of oncology clinicians. ESMO Open 2021;6(5):100224 [FREE Full text] [doi: 10.1016/j.esmoop.2021.100224] [Medline: 34461484]
51. Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future: an independent report on behalf of the secretary of state for health and social care. The NHS Constitution. 2019. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2024-03-15]
52. Harwich E, Laycock K. Thinking on its own: AI in the NHS. REFORM. 2018. URL: https://reform.uk/wp-content/uploads/2018/11/AI-in-Healthcare-report_WEB.pdf [accessed 2024-03-15]
53. Archer S, Babb de Villiers C, Scheibl F, Carver T, Hartley S, Lee A, et al. Evaluating clinician acceptability of the prototype CanRisk tool for predicting risk of breast and ovarian cancer: a multi-methods study. PLoS One 2020;15(3):e0229999 [FREE Full text] [doi: 10.1371/journal.pone.0229999] [Medline: 32142536]
54. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. Intell-Based Med 2020;1-2:100001 [FREE Full text] [doi: 10.1016/j.ibmed.2020.100001]
55. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ 2003;327(7417):741-744 [FREE Full text] [doi: 10.1136/bmj.327.7417.741] [Medline: 14512488]
56. Naik G, Ahmed H, Edwards AGK. Communicating risk to patients and the public. Br J Gen Pract 2012;62(597):213-216 [FREE Full text] [doi: 10.3399/bjgp12X636236] [Medline: 22520906]

XSL•FO
RenderX

57. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
58. Murray CJL, Alamro NMS, Hwang H, Lee U. Digital public health and COVID-19. Lancet Public Health 2020;5(9):e469-e470 [FREE Full text] [doi: 10.1016/s2468-2667(20)30187-0]
59. Shoop-Worrall SJW, Cresswell K, Bolger I, Dillon B, Hyrich KL, Geifman N, Members of the CLUSTER consortium. Nothing about us without us: involving patient collaborators for machine learning applications in rheumatology. Ann Rheum Dis 2021;80(12):1505-1510 [FREE Full text] [doi: 10.1136/annrheumdis-2021-220454] [Medline: 34226185]
60. Darley A, Coughlan B, Maguire R, McCann L, Furlong E. A bridge from uncertainty to understanding: the meaning of symptom management digital health technology during cancer treatment. Digit Health 2023;9:20552076231152163 [FREE Full text] [doi: 10.1177/20552076231152163] [Medline: 36714543]
61. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bull World Health Organ 2020;98(4):245-250 [FREE Full text] [doi: 10.2471/BLT.19.237198] [Medline: 32284647]

Abbreviations

AI: artificial intelligence
CRS: chemotherapy-related symptoms
eSMART: Electronic Symptom Management Using the Advanced Symptom Management System
PRM: predictive risk modeling
QoL: quality of life
RCT: randomized controlled trial

Edited by T de Azevedo Cardoso; submitted 01.06.23; peer-reviewed by I Wilson, C Hathaway; comments to author 13.10.23; revised version received 22.11.23; accepted 06.03.24; published 20.06.24.

Please cite as:
Miller M, McCann L, Lewis L, Miaskowski C, Ream E, Darley A, Harris J, Kotronoulas G, V Berg G, Lubowitzki S, Armes J, Patiraki E, Furlong E, Fox P, Gaiger A, Cardone A, Orr D, Flowerday A, Katsaragakis S, Skene S, Moore M, McCrone P, De Souza N, Donnan PT, Maguire R
Patients' and Clinicians' Perceptions of the Clinical Utility of Predictive Risk Models for Chemotherapy-Related Symptom Management: Qualitative Exploration Using Focus Groups and Interviews
J Med Internet Res 2024;26:e49309
URL: https://www.jmir.org/2024/1/e49309
doi:10.2196/49309
PMID:

©Morven Miller, Lisa McCann, Liane Lewis, Christine Miaskowski, Emma Ream, Andrew Darley, Jenny Harris, Grigorios Kotronoulas, Geir V Berg, Simone Lubowitzki, Jo Armes, Elizabeth Patiraki, Eileen Furlong, Patricia Fox, Alexander Gaiger, Antonella Cardone, Dawn Orr, Adrian Flowerday, Stylianos Katsaragakis, Simon Skene, Margaret Moore, Paul McCrone, Nicosha De Souza, Peter T Donnan, Roma Maguire. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Facilitators of and Barriers to Teachers' Engagement With Consumer Technologies for Stress Management: Qualitative Study

Julia B Manning[1], BSc, PhD; Ann Blandford[2], CEng, PhD; Julian Edbrooke-Childs[3], PhD

[1]UCL Interaction Centre, Department of Computer Science, University College London, London, United Kingdom

[2]UCL Institute of Healthcare Engineering, University College London, London, United Kingdom

[3]Evidence-based Practice Unit, University College London and Anna Freud Centre, London, United Kingdom

Corresponding Author:
Julia B Manning, BSc, PhD
UCL Interaction Centre
Department of Computer Science
University College London
66-72 Gower Street
London, WC1E 6EA
United Kingdom
Phone: 44 02076792000
Email: julia.manning@ucl.ac.uk

Abstract

Background: Consumer technology is increasingly being adopted to support personal stress management, including by teachers. Multidisciplinary research has contributed some knowledge of design and features that can help detect and manage workplace stress. However, there is less understanding of what facilitates engagement with ubiquitous "off the shelf" technologies, particularly in a specific occupational setting. An understanding of features that facilitate or inhibit technology use, and the influences of contexts on the manner of interaction, could improve teachers' stress-management opportunities.

Objective: The aim of the study was to investigate the interaction features that facilitated or inhibited engagement with 4 consumer technologies chosen by teachers for stress management, as well as the influence of the educational contexts on their engagement. We also examined how use of well-being technology could be better supported in the school.

Methods: The choice of consumer technologies was categorized in a taxonomy for English secondary school teachers according to stress-management strategies and digital features. Due to the COVID-19 pandemic, we adapted the study so that working from home in the summer could be contrasted with being back in school. Thus, a longitudinal study intended for 6 weeks in the summer term (in 2020) was extended into the autumn term, lasting up to 27 weeks. Teachers chose to use either a Withings smartwatch or Wysa, Daylio, or Teacher Tapp apps. Two semistructured interviews and web-based surveys were conducted with 8 teachers in England in the summer term, and 6 (75%) of them took part in a third interview in the autumn term. Interviews were analyzed using reflexive thematic analysis informed by interpretive phenomenological analysis.

Results: Technology elements and characteristics such as passive data collation, brevity of interaction, discreet appearance, reminders, and data visualization were described by teachers as facilitators. Lack of instructions and information on features, connectivity, extended interaction requirements, and nondifferentiation of activity and exercise data were described as barriers. Mesocontextual barriers to engagement were also reported, particularly when teachers were back on school premises, including temporal constraints, social stigma, and lack of private space to de-stress. Teachers had ideas for feature improvements and how educational leadership normalizing teachers' stress management with consumer technologies could benefit the school culture.

Conclusions: Having preselected their stress-management strategies, teachers were able to harness design features to support themselves over an extended period. There could be an important role for digital interventions as part of teachers' stress management, which the school leadership would need to leverage to maximize their potential. The findings add to the holistic understanding of situated self-care and should inform developers' considerations for occupational digital stress support.

(J Med Internet Res 2024;26:e50457) doi:10.2196/50457

XSL•FO
RenderX

KEYWORDS

teachers; stress; workplace; self-management; digital health; technology; qualitative; context; high schools; wearables; apps; human-computer interaction; HCI; personal informatics

Introduction

Background

Workplace stress was reported nearly 200 years ago and has been described by Jackson [1] as "psychological pressures generated by the unfettered growth of industrial and technological capitalism." When these pressures overwhelm, they cause distress (hereafter just referred to as stress) and stress continues as a pervasive phenomenon today. More than twice as many people report stress, depression, or anxiety causing them to miss work compared with any other reason, with teachers demonstrating nearly double the incidence rates of the general population [2]. Stress is a determinant of well-being and can be defined as "the adverse reaction people have to excessive pressures or other types of demand placed on them" [3]. Persistent, excessive stress is a risk factor for illness, including affective disorders such as anxiety and depression [4] and cardiovascular disease [5]; the latter is associated with burnout in even young female teachers [6]. Therefore, prevention or early intervention to reduce the effects of stress remains of huge importance.

Stress management interventions have proliferated, motivated by challenges including therapy costs, accessibility, and unmet needs of employees [7-11] with over 1000 stress apps found in web-based stores [12]. Increasingly, research shows digital tools can also be harnessed to detect, reduce, and manage stress. These technologies have used a variety of methods, including approaches broadly classified as personal informatics [13] that involve gathering self-report or automatic sensor data from everyday life to support reflection [14-16]. Other approaches include condition-based education [17], self-management through guided therapy [18,19], peer or social support [20,21], and gaming or entertainment [22,23].

Digital platforms mean that personal stress management interventions are now accessible throughout the working day via web- and computer-based programs [24], apps [25-28], and wearables [29,30]. In addition, popular websites have been repurposed to provide stress management interventions [10].

People's ability to access stress management interventions is also highly reliant on their contexts. Methods such as contextual inquiry explore how technology fits into everyday life and tend to consider the context of the organization or sector [31,32]. Also important are the contextual influences on the ability to engage with digital interventions such as social norms, culture, time, and location [33]. Human-computer interaction analysis only occasionally extends to a distal understanding of the value chain in how interventions work in context and why [34]. Yet, this is important for people using stress management interventions in the workplace, both with novel designs [35-40] and consumer devices [41-44], as such an understanding of contextual influences on interaction with features would inform evaluation, iteration, and use. Such real-world studies are few, especially outside office environments.

There have been some multidisciplinary reviews of digital workplace interventions, yielding little insight into influences and features, but stress is usually grouped with other conditions in the continuum of mental health and well-being, and few studies are situated in the public sector [45-48]. Many studies in these reviews focus on novel system design or the evaluation of a single digital solution. Less is known about facilitators for and barriers to employees using consumer technologies to manage occupational stress, despite the proliferation of commercial tracking tools and the consequent amenability of studying their real-world use [16,49].

Aims

We know that many teachers already have their own stress management strategies [50,51] and some are already using consumer technologies to support them [52]. This study took a phenomenological approach to understand how digital technology could facilitate teachers' stress management in their real-world, complex layered contexts of the school setting [53]. Our research questions were as follows:

1. What were the facilitators and barriers to teachers engaging with their technology?
2. How could the school support the use of consumer technology for stress self-management? Coincidentally, this study commenced as the global COVID-19 pandemic took hold. Most teachers were suddenly working remotely from home at first, so we had the opportunity to compare facilitators and barriers to technology use between the home and school environments.

In this paper, we first describe digital detection and management of workplace stress and then review what is known about facilitators and barriers to interaction with such digital tools. We then present our study, describing which technologies were used by the teachers and presenting the research findings. We discuss what features teachers found to help them with stress management and what were the impediments to self-care. Finally, we consider what this means for school leadership in the context of education.

Related Work

Stress Detection

Digital tools can enable both self-report and automatic detection of stress. Manually, this can be achieved through ecological momentary assessment [54], and popular consumer apps such as Daylio and Moodkit (Thriveport LLC) allow reflection on captured feelings or symptoms associated with stress and well-being [49,55]. Automatic detection can be through wearable sensors using galvanic skin response (GSR), also known as electrodermal activity [43], through photoplethysmography (PPG) to detect heart rate variation [30,56], and from smartphone text language [57]. Equivalence of stress detection in the real world has been demonstrated between manual self-reports and both automated GSR or voice-based analysis [58]. Heart rate and heart rate variation are a burgeoning area

XSL•FO
RenderX

of research [59,60]. Automatic sensor data enable reflection on both stress-associated heart rate directly and on associated stress variables such as mood, activity, and sleep [16,60-63]. Caution is required for stress detection, both because detection itself can be a stressor [36] and due to discrepancy in perception between automated reports and self-evaluation, both in terms of not only severity [64] but also lack of differentiation between distress and eustress (the latter being a positive motivator for growth, development, and change) [4,65].

By definition, employees' stress is strongly linked to the contexts in which they work. Some workplace studies have illustrated occupational associations—for instance, a study of desk-based office workers' changes in the use of a mouse or keyboard showed it to be a valid proxy for stress [66]; in quiet environments, smartphone sensors have also detected stress through the human voice [58,67]; and phone use data are also correlated with stress [68]. However, these examples also show their contextual sensitivity and irrelevance for teachers in their noisy, dynamic, and didactic environments. Few studies have examined stress detection among teachers. Fitbits (Google Fitbit) using PPG to capture heart rate data were used to help teachers pinpoint stressors [69] and Philips DTI-2 devices (Philips) using GSR were worn by vocational school teachers to explore associations between data and stressors [70]. Research on the use of stress-tracking (personal informatics) technologies in other real-world contexts is sparse [16,41], although it is of increasing interest as machine learning is applied to sensor data, enabling prediction as well as detection [35,71].

Given that "context" is used to describe everything from high-level surroundings to individual-level characteristics, we drew on multidisciplinary domains of education, human factor, and informatics literature [72-75] to define the education workplace as the macrocontext; define the temporal, physical, social, and cultural factors as the mesocontexts; and define the individual's personally chosen strategy as the microcontext.

Stress Management in the Workplace

Detecting stress and understanding that contexts influence experiences of stress is one side of the equation. Assisting people to then manage this stress is the other side, and critical to this is not only the digital medium or concept but what features are appropriate for the context. For the purposes of this paper, features are regarded as mechanisms of interaction or engagement, including reflection, therapy, and social or professional support [33,76,77] with facilitating elements such as information, notifications, or passive data collection and characteristics such as the medium and esthetics [78]. We have little information on what appropriate features are for school teachers, although there may be some relevant insight from other workplaces. Some studies have reported that participants would recommend the intervention to a friend but did not give any reasons as to why or when [29]. They also omitted key information, such as whether interventions were accessed at work, particularly in busy contexts, such as health care [79-81].

Facilitating Features

Personal data collection has facilitated correlations between stress and real-world stressors [49], including for teachers from their wearable data [69,70], and enabled anticipation of stress for mitigation planning among students [82]. One population-specific stress intervention, a role-playing game, showed design could be relatable to a specific culture [83], although role-playing games for stress management generally appear to be untested [84].

Timing and choice of stress intervention or prompts have given indicators of user receptivity. Desk-based workers preferred prompts at the start of a task or day, during lunch breaks, or during the postlunch circadian rhythm dip of 2 PM to 3 PM [38]. Autonomy to choose interventions has been linked to enhanced motivation and adherence to mental health interventions [85]. Where cognitive behavioral therapy (CBT), mindfulness, and tailored stress management therapies have been delivered digitally to employees from a variety of occupations, they have fared better in the workplace when supported or guided than when not supported or guided [46,85,86].

In education, despite some extensive trials with teachers [50,69,70,87-91], qualitative insights on facilitating features of digital stress interventions are few. However, temporal flexibility of intervention [20,91], informational support [92], and anonymity and peer social support [20] have been credited as facilitating features. In other workplaces, interesting content and interactivity, short completion time, progress tracking for reflection and reminders [93], and personalization in apps were appreciated [8,94]; and flexibility, goals, and calming strategies were desired [95]. Synthesizing this workplace literature, facilitating themes for digitally supported stress management emerge as a personal choice over timing, supported reflection, and relevance.

Barriers to Use and Interaction

Barriers to digital workplace stress interventions are from heterogeneous sources but include concepts relevant to teaching. They can include a lack of information: not understanding how to respond to data or lack of instructions and the need for "communities of practice" [41,70,96,97]. Mesocontextual barriers have included time pressures [95,98,99], lack of privacy [97] and cultural stigma from having to admit a need or the intervention being for mental health,' or lack of active management support for intervention deployment [93,97,100,101]. Personal factors include being unmotivated [93] or lacking autonomy or choice over design interventions [100,102].

Methods

Overview

The goal of this study was to investigate how consumer technologies could facilitate stress management in the school environment.

A longitudinal study was planned that used semistructured interviews, as well as open and closed-question surveys. Before the first interview, workshops were conducted to help teachers make an informed choice of consumer technology based on understanding their personal stress management strategy and technology preferences. Consumer technologies were framed

XSL•FO
RenderX

as digital companions when presented to teachers in workshops. This framing drew on the framing of apps as "life companions" by Klasnja and Pratt [103], acknowledging users' autonomy [104], emphasizing the mediating and collaborative role of the technology [105], and avoiding any commercial connotations of simply being a consumer Referred to by name or simply as technologies in this paper, those chosen by teachers for this study were the Withings Steel HR smartwatch (that uses PPG to detect heart rate changes); Wysa (Wysa Ltd), an artificial intelligence chatbot based on CBT techniques; Daylio, a mood diary and activity tracking app; and Teacher Tapp, a daily teachers survey with an educational blog. Support was given to the teachers in acquiring their choice of technology from those populating a rigorously compiled taxonomy based on suitability, availability, evaluation, security, validity, and cost; the design of which is described elsewhere [106]. Reasons for teachers' choices are described in detail in a companion paper on their experiences but include apparent ease of use, to inform stress correlations; enable passive collation of data; and give insight into sleeping patterns [107].

The interview data were supplemented with feedback questionnaires, memos and note taking, and email responses (including from nonparticipants, with their permission). The details, rationale, and elements of these methods are now described.

The orientation of this research was experiential qualitative research [108]. The approach was not guided by an existing theory but combined inductive (bottom-up from the data), cognitive, and social constructivist (the participants' viewpoints) reflexive thematic analysis [109,110]. The constructivist paradigm is familiar in both human-computer interaction and education research, thus enabling epistemic consistency.

Teacher Recruitment

The initial selection of participants was purposive. The inclusion criteria were to be a head of year or assistant head with pastoral responsibilities. Willingness to both talk about stress and to trial their chosen technology to support their stress management strategy; and not currently receiving treatment for a clinical stress disorder such as posttraumatic stress disorder or acute stress disorders were specified in the information sheet. The latter was stipulated to avoid interference between this study and potential prescribed interventions. No prior training in stress interventions was required.

Recruitment was from schools in one multi-academy trust (MAT) based in South London. The Trust had been approached for several reasons. First, MATs have a single governance structure, providing some reduction of internal organizational variation, but each has its own leadership team. Second, the MAT had 3 secondary schools in the South London region whose principals were willing to participate, providing a geographic locus and city context for the research. While there are always some differences in the local housing and labor markets, the good transport links, a predominance of business and service sector jobs along with low unemployment, gave some cohesion to area characteristics [111]. Third, the MAT had a track record of interest in staff well-being. By working in the same city with one MAT and teachers with the same role (head of year), the aim was to ensure some contextual consistency.

Of the 16 teachers who originally volunteered to take part before the pandemic (having been invited by the MAT directors and principals), 8 (50%) were still able to participate after the COVID-19 pandemic precipitated a national lockdown, with participation now facilitated on the web. None were previously known to the researcher. Four (N=16, 25%) of the teachers who dropped out gave reasons and consented to have them noted. These included extra childcare responsibilities, their partner working for the National Health Service and being overwhelmed, and an increase in teaching load. The other 4 did not communicate their reasons.

Ethical Considerations

This research was approved by the University College London Research Ethics Committee (UCLIC_1920_004_Staff_BlandfordManning). Information and consent forms were sent electronically to teachers before the workshops, followed up with reminder emails if they had not been received by the day before the workshop. The consent and information forms specified the purpose of the workshop and subsequent interviews and survey should they choose to take part, as well as the interviewer's (JBM) department and university, and that the data would contribute to a final study being undertaken as part of her PhD. The data collected were anonymized, with personal names, identifying remarks, or identifiable topics mentioned in the interview audio recordings removed from the transcripts. Participation was voluntary, but a flat participation fee of £150 (US $183.10) was paid to cover the costs of the participant obtaining their chosen technology. Participants had the right to withdraw from the study at any time.

Due to the pandemic, only 2 participants spent more time in school than working from home during the summer leg of the study. Six (75%) of the 8 teachers extended their involvement from the "intended" study into the autumn term back in school. The 2 (25%) who did not continue gave no reason, but there was still a huge pressure on schools at that time. This autumn interview is described as the "extended" period in the results.

Workshops

The choice of consumer technology was facilitated by the presentation of a taxonomy of stress management interventions [106]. The workshop was piloted and refined with 3 teachers not involved in the study. Workshops used Zoom (Zoom Video Communications), PowerPoint (Microsoft), and Mentimeter and were arranged at a convenient time for the teachers. Three teachers attended the first workshop, 2 the second, and 3 the third. Two people ran the workshops, a teacher and the researcher (JBM), who is an experienced clinician and policy researcher interested in professionals' stress.

The workshops were designed to first introduce the study concept. Second, they were designed to allow teachers the time and information to understand the different ways in which they consciously or subconsciously seek to manage stress. This was intended to facilitate the identification of personal stress management strategies. Third, the workshops were designed to introduce teachers to the different ways in which technology

XSL•FO
RenderX

could conceptually support their stress management strategies. Finally, they were designed to describe each of the technology options open to them for the forthcoming study. Mentimeter, an interactive presentation software, was adopted to enable anonymous crowdsourcing of ideas from teachers. The software aggregates responses to questions and displays them in real time as a word cloud.

Data were gathered both through feedback forms and from the workshop recordings. Transcripts were made of the latter, but the data were mainly reflections on understanding stress and contextual information relating to COVID-19 pandemic changes. The feedback forms allowed teachers to record their choice of technology, along with why they chose it, and when and where they anticipated using it. Participants were reminded only to submit their feedback form if they wished to proceed. All participants returned their forms. The feedback form also captured some demographic data and contextual feedback, largely related to changes brought about by the pandemic. These data along with the comments from the workshop allowed for some tailoring of the semistructured interviews according to technology choice and the altered teaching arrangements due to the COVID-19 pandemic.

After receiving the feedback form, the researcher contacted the participant to arrange (1) payment of the flat participation fee and (2) schedule the time of their first interview. The staggered timings of the workshops meant that some teachers had their technology ahead of others. One teacher had considered trying an online CBT program but on examination reported it as too medical. Another had hoped to use the Fit2Teach Pro app but at the time, technology purchases needed to be made and the app had not been updated within the previous 6 months, so it was excluded as an option. Another teacher wanted to try both the Withings watch and Wysa app, but as the cost exceeded the flat participation fee, she opted for the Withings watch only (Table 1).

Table 1. Teachers' technology choices and summary findings.

Teacher number	Technology choice	Time as teacher (years)	Chosen stress management strategy and technology-supported concept (first or second)	Duration of technology use (from download or receipt and nighttime use)	Extra features desired (or already present but undiscovered)
T1	Daylio and Wysa apps	>10	Cognitive and reflection; cognitive and self-management	20 weeks continuous then a few missed days before the autumn half term	• Better data visualization • Congratulatory messaging
T2	Withings watch—wearable	>10	Physiological and reflection or peer-to-peer social	22 weeks (including nights)	• Ability to link raised heart rate with correlations in a diary • Raised heart rate warning
T3	Withings watch—wearable	>20	Physiological and reflection or peer-to-peer social	22 weeks continuous (nightwear quickly discontinued)	• Ability to track gardening or not just sports activity
T4	Withings watch—wearable	<10	Physiological and reflection or peer-to-peer social	15 weeks continuous then 4× per week for the next 7 weeks (nightwear quickly discontinued)	• Link between calorie expenditure and nutritional advice • Slimmer-too bulky to wear at night • Naming alarms
T5	Teacher Tapp	>10	Educational and reflection	7 weeks 2 days continuous, then occasional use for 7 weeks in the autumn term	• Mindfulness section and link to a yoga workout as a reminder to take breaks
T6	Withings watch—wearable	>20	Physiological and reflection or peer-to-peer social	27 weeks continuous (including nights) included planned health leave for 6 weeks	• Longer-range Bluetooth connectivity • More than 3 alarms • Naming alarms
T7	Withings watch—wearable	>10	Physiological and reflection or peer-to-peer social	7 weeks of intermittent wear; also picked up Teacher Tapp intermittently	• None
T8	Withings watch—wearable	<10	Physiological and reflection or peer-to-peer social	8 weeks continuous	• Indicators of normal parameters with focus on solutions or signposting in the Withings app (eg, raised heart rate warning)

XSL•FO
RenderX

Semistructured Interview Details

Eight teachers took part in 2 one-to-one interviews planned for the intended study; 6 (75%) teachers went on to participate in a further interview for the extended study in the autumn. At each interview, an overview of the study was first recapped and the teacher's current circumstances were discussed. Open-ended questions were then framed around the teacher's personal choice of technology followed by probing of contexts (eg, physical, social, and cultural). Questions tailored to the technology allowed in-depth explanations of the facilitation of and barriers to the use of features and their ideas for improvement. After the interview, each participant's responses were copied into a file and added to a pseudonymized case study folder for that participant in NVivo (Lumivero). The length of each interview ranged from 32 to 76 minutes.

Open and Closed Survey Questions

In the intended study, spanning 6 to 7 weeks of the summer term, surveys were sent after about 4 weeks to prompt thinking about activity, data gathered, frequency of use, and stress symptoms. Three surveys were set up using Microsoft Forms, and each participant was sent a link by email and staggered through the week. The questions were arranged around themes and included branching from responses to ensure subsequent questions were logical and related to their technology choice. Closed yes or no, multiple-choice, Likert scale, and open responses were included. The response rate varied from 100% (8/8, first and last day of the week) to 75% (6/8, midweek). The answers informed the second and third interview questions.

In the extended study, 14 closed questions were asked at the end of the interview as a way of capturing final participant thoughts on technology features and the influence of their occupational contexts on use. These were simple agree-disagree responses that allowed for greater confidence in the analysis. All survey responses also provided more data for triangulation.

Data Analysis

All interview recordings were transcribed verbatim and identifiers were removed. Transcripts were not returned to participants, as (1) the researcher did not want to add a time burden to the teachers taking part, (2) the evidence suggests that doing so adds little to the accuracy of the transcript [112], and (3) this retains the responses precisely as they had been given at the time of the interview when the teacher had the time and space to respond naturally. Transcribed interviews were checked against the audio recordings for accuracy.

Reflexive thematic analysis was undertaken of the whole data set (transcripts, surveys, researchers notes, and emails) [113] and informed by interpretive phenomenological analysis. NVivo (version 12) and SimpleMind (versions 1.22 and 1.31) software were used for the organization and development of codes, categories, relationships, and themes. Codes were shared with the coauthors (AB and JE-C), and a codebook was generated. In line with the advice of Braun and Clarke [110], a self-reflexive approach was taken throughout the research, including creating a rolling memo of reflections from the time of the interviews until the end of the study. These notes were shared with the coauthors during the stages of thematic analysis.

Table 1 summarizes the teachers' technology choice and supporting concepts, duration of use, strategy and mechanisms, and extra features desired.

Results

Overview

A total of 8 teachers completed the summer interviews, and 6 (75%) were involved in the extended study in the autumn of 2020. Most teachers were working from home during the summer term due to COVID-19 pandemic restrictions. They were all back in school in the autumn but in "bubbles" segregated by year group, moving between classes instead of students moving around the school and with extended school days due to staggered student arrival times. The sample was ethnically diverse, with just 1 male participant in both summer and autumn studies.

Teachers described how they began to use their technologies and which features facilitated their interaction with the technology and awareness of their stress. Quick and discreet interactions were complemented by passive data collection, making use simple, convenient, and private. Automated prompts and data collation facilitated easy engagement and awareness of stress patterns and associations. However, the teachers described having to be persistent with adoption, as inadequate instructions required them to experiment and use social media to discover more features. There was a particular issue reported with the Withings watch measuring sleep when they were awake and a desire that this wearable could automatically detect activity related to their job or their stress-relieving exercise. For fuller use and benefits to be realized, teachers described creating a culture in which it is socially acceptable to manage and talk about stress, where privacy was facilitated and self-care with technology was normal and personalized from the start of a teacher's career. This would legitimize experiences of pressure and permit self-care.

The participant code is included at the end of quotations in the form of a teacher number (T1, T2, etc); the technology; and the term (summer or autumn) from which the quotation was taken.

Theme 1: Technology Design Facilitated Interaction for Teachers

This theme captured how the design and features supported rapid technology use. Teachers conveyed a sense of confidence that interaction could therefore take place without anyone knowing.

Subtheme 1.1: Passive Collection of Data Is Convenient

Although the Withings watch can be set to record activities, it also automatically registers when the wearer goes for a walk. Teachers appreciated this convenience and the option to either reflect on data in the moment or at a time of their choosing. Teachers using the Withings watch liked the fact that it was "just there" on their wrist, and this in itself could motivate activity:

> *I go from here to work and walk or I do a walk, it will still register when you download it as a walk. Or you can then programme it and it will still register it as*

XSL•FO
RenderX

> *a walk. So that's good in the fact that it knows it's not just general walking around, it does it as an activity even though I haven't put it and logged it as an activity.* [T3, Withings, summer]

In-the-moment reflection was facilitated by setting haptic notifications; one teacher described being able to set an alert triggered by her heart rate:

> *I know if I'm getting totally stressed, I'll have a look at [the watch] now. And I'll go, hold on, it's not worth this, just slow down... you can set an alarm when your heart rate goes over a certain amount... Just a buzz, and it tells you where your reading is.* [T6, Withings, autumn]

Subtheme 1.2: Brevity of Interaction Simplifies Usability

Several teachers spoke not only of the convenience of passive data collection but also of this being complemented by the brevity with which data could be viewed. This cognitive ease of viewing the Withings watch was true for 2 apps as well, even though the data collection design required active engagement.

The contrast between working from home and in school can be seen in a teacher's report of using her app in the summer and then toward the end of the autumn half term, when even the brief interaction could be overwhelmed by ongoing workplace demands:

> *As soon as [I] turn my phone [on] you see all the faces that come up straight away... And [later] I hear the brrrrng...I look down...I just do it quickly.* [T1, Daylio, summer]

> *But I know that I've missed days, not from not wanting to do it, but I forgot to do it. I'm using it then I say okay, I'll do it in a minute.... I think the last two weeks, I've been really tired, really, really tired.* [T1, Daylio, autumn]

Another contrast could be seen between apps that offer tracking or treatment. While both were accessible, only the former was usable in school, although the understanding facilitated by the latter was appreciated:

> *I like Wysa. But.... You can't just do a quick fix.... I was using Wysa, and it must have been about 45 minutes later [I thought] I've got to go to bed!...as a result of that...I don't use Wysa as much.* [T1, Wysa, summer]

Subtheme 1.3: Discreet Appearance Encourages Use

The concept of cognitive ease also included the ability for discreet use, including using the messaging feature of the Withings watch to triage notifications, with the sender and start of message view enabling teachers to decide whether the whole message or email needed immediate attention or not. The subtle design of both the wearable and tracking app was commented on by most participants:

> *I'm not typing in anything. It's very simple to use.... You can actually press the buttons and then press okay and then it just goes away straight away. Nobody needs to know why you are using [it] or what that's for.* [T1, Daylio, summer]

The brevity of interaction or unobtrusive haptic notification built into the Withings watch worked together with discreet design to facilitate use:

> *I know now if my arm's vibrating, I can just take a quick look. And if I think it is something that I need to address, then I can, and if it's not, just ignore it. I got into trouble for having my phone out before.* [T2, Withings, autumn]

Theme 2: Data Presentation Enabled Teachers' Stress Awareness

Overview

The essence of the first theme was cognitive ease, and this was also true for data presentation. Use was further described in terms of how prompts and patterns made interacting with data simple. Teachers reported their data were a useful measure of quantification of their stress and they could see information relevant to educators.

Subtheme 2.1: Automated Reminder for Cognition or Reflection

As teachers used their technologies, both automation of questions linked to data collation and seeing raw data with the chance to capture other activities were appreciated as prompts to take further action. For the teachers using Daylio and Teacher Tapp, the automated reminders to record their emotional state or answer education-oriented survey questions facilitated the opportunity to stop and think, although the frequency of doing this was significantly curtailed when they were back in school full time. Some attributed this entirely to the ongoing pandemic:

> *Looking at it, and then it'll prompt you, if you like, so how many more steps do you want to do, what have you done? It congratulates you, it just does nice stuff on it.* [T6, Withings, summer]

> *Teacher Tapp was asking you your thoughts, or...there were a couple of times it asked you specific stress questions as well...that made me think, okay, what am I doing? Should I try some mindfulness? Should I take some time off? Am I reflecting on my work?... During COVID and being [out of] school, [I used Teacher Tapp] almost 99–100%. Now that school has started, completely the opposite, I'd say I've probably used it ten times...if I was back to school and it wasn't a pandemic, I'm still adamant I would use it.* [T5, Teacher Tapp, autumn]

Subtheme 2.2: Correlating Patterns in the Macrocontext of Education

Throughout the study, many teachers mentioned the patterns they could see in their data as being informative and relevant. This could simply be the visual display of their activity readings or tracking display, as shown in the Withing's Health Mate companion app or on the Daylio app. It could be data showing a raised heart rate and making post hoc correlations between it

XSL•FO
RenderX

and a stressful situation during the day or episodes of a particular in-school activity:

> *I'm looking at the data like steps, heart rate, because you can track it during the day when you go onto the app, and you can see your heart rate throughout the day. You can see where things, you can say, yes, that was where I was...dealing with a stressful situation.* [T6, Withings, autumn]

As the Daylio app allowed the tracking through emoticons, the teacher using it was able to see the connections between a less satisfactory day and her physical well-being:

> *I actually started to note the days that I'm not feeling physically well are the days, and I note it as a "meh" day, so you can see where the pattern is when I do that.* [T1, Daylio, summer]

Awareness of these patterns in the data was interlinked with the theme of relating data to fellow teachers.

Subtheme 2.3: Relevance of Information to Situation

Teachers described the data not only quantifying their stress but also facilitating education about their colleagues' stress as well. Two described Teacher Tapp survey results as providing relevant data on the broader educational system. The teacher using the chatbot described how the interactions equipped her with the language to talk to fellow staff and this was an unanticipated benefit:

> *What Wysa has helped me with is managing other people because of all the tools that it gives you. So, even today, somebody was feeling a bit down.... And I said I'm sorry that you feel like that. It's something that Wysa would say!... Wysa is very educational, it's been more of an educational tool of helping me to understand.* [T1, Wysa, autumn]

Theme 3: Barriers With Instructions, Accuracy, and Other Limitations

Overview

Contrasting with the facilitating features of design, barriers were first evident from descriptions of the designers' apparent assumptions that consumer technologies use could be entirely intuitive. This created some friction in their use. The usefulness of the technology was hindered by a lack of information, with some participants believing that features were not available when they were and vice versa. Barriers also included a lack of automatic data differentiation and connectivity limitations. Once again, these subthemes are interconnected.

Subtheme 3.1: Inadequate Information Supplied on Technology Adoption

Five out of 6 teachers who used the Withings wearable commented on the lack of information provided with the watch. They all looked for more information on YouTube. Some teachers had found a feature that others thought was missing, but there was a strong sense that clearer instructions for all the technologies would have been appreciated:

> *I think for me, doing this has been more of a trial and error, and working out, oh this is what you do. But if there was a better instruction manual, it'd be easier to work through things.... The YouTube things are great because I'll watch [and] they can [show you how to] do this and that.* [T1, Daylio, summer]

The inadequacy of instruction was related not only to the functional use of the Withings watch but also to what data constituted "normal" for the user and how data could vary. Some teachers were worried by their raised heart rate or commented that seeing their personal data and variation in isolation was worrying:

> *Also, the only thing with the watch and its limitations is, you don't really have a national comparison.... As in, I find there's ten beats per minute difference between me at the weekend, and me at school. So, on the weekend I average 73, 74 resting heart rate. Whereas in school, it's 83, 84. And I'm like, is that normal?...Obviously, I'm more stressed, you'd expect that, but is that the same average (for other) people?* [T8, Withings, summer]

Subtheme 3.2: Insufficient Instructions on Technology Features and Limited Strategy Options

Alternative stress management features, narrowness of existing features, and lack of signposting were described by teachers as insufficiencies in their technology. The desire for a greater plurality of features was often related to alternative stress management strategies that teachers described when on the school premises. Instructional insufficiency was seen when one teacher, asked about improving the Withings watch, said that she would like there to be a food diary. However, this was already a feature on the companion app that another teacher had mentioned. Another example was where 2 teachers described the haptic notification function and how they would appreciate being able to set it to alert if their heart rate was raised, yet another teacher described that she had done this. However, the teacher who had managed this also expressed disappointment that the alarms she had set up as reminders to drink could not be named, yet other teachers had described naming their alarms:

> *What would I like the watch to be able to do?... Set multiple alarms more than I'm doing now, I think. You can set multiple alarms [but I would want to name] an alarm to say, have a water break.... I don't set it for anything else [now] other than [as a] cue to have a drink, make sure I'm drinking.* [T6, Withings, autumn]

A point was also made that having exercise limited to sport was a missed opportunity to label data on activities, such as gardening or cleaning. Sometimes secondary technology concepts supported alternative stress management strategies, such as social connection, unanticipated before the social isolation of lockdown. However, some of the teachers wanted more strategies supported by their technology:

> *Maybe Teacher Tapp could be changed.... I just found as someone that likes to read and get information it...brought me into that mindset that I was still*

XSL•FO
RenderX

> *working.... I'd definitely make a part where there's yoga or mindfulness sections allowing people to get away, just a different strategy.* [T5, Teacher Tapp, autumn]

Subtheme 3.3: Variation in Usefulness, Accuracy, and Accessibility

Comments on the accuracy of the data collected were mostly positive, with 1 participant who had previously used another activity tracker stating that she "couldn't cheat" with the Withings watch, as she had been able to increase her activity reading simply through shaking her old tracker. However, inaccuracies were mentioned by several teachers, including the Withings watch recording them as sleeping when they were not:

> *One thing that I was confused about.... It measured me sleeping when I actually wasn't.* [T7, Withings, summer]

Accessibility of the technology was a theme that meant different things to teachers. Teachers with the wearable reported appreciating that the watch was "just there" on their wrists and easy to view during the working day, but one needed spectacles to see it. Using a smartphone app during the school day required forward planning for proximity, either to simply access the app or for the short-range Bluetooth required for message transmission between the phone and watch:

> *Normally I used to just put [my phone] in my bag. But what I've done, I've bought a pouch... so if I'm wearing a dress or it's something that hasn't got pockets, then I can put it in my pouch and keep it on me.* [T1, Daylio, summer]

> *I haven't got the connection [in school] to [receive messages]... You have to be within a certain distance, and I'm never within that distance...maybe... if I put my phone in my laptop bag...so that it's with me, but the kids don't know I've got it. Because I don't like to have my phone around.* [T6, Withings, autumn]

Most teachers liked the presentation of data, but it was observed that a more visual display of data could be more compelling and relatable to the associated activity:

> *It [Daylio] gives you all the data, it breaks it down, say, how many times you've done this, how many times you've done this, etc...but I guess it would be good to have a visual...where you have a bar chart or something that's blatantly obvious... a sunshine, or something, I don't know. Feel good. Happy. Oh, wow, there's that sunshine, that rainbow, wow...So, you've had these good days because you've done X, Y, Z.* [T1, Daylio, autumn]

Subtheme 3.4: Need for Differentiation for Data Relating to Stress Management Strategy

Descriptions of how the data were relevant to participants' working lives as teachers included capturing the demands of the day. However, for some, differentiation between stress self-management activity from activity intrinsic to the job required remembering to activate specific tracking as it was not automatic. One teacher described it as an additional task, although possibly unrealistic, as 5 (83%) out of 6 teachers described a decrease in interaction with technologies overall in the autumn term:

> *I think for me...[the technology] was a really useful reminder for me to keep fit, to keep active...I think that really made a big difference to my mental wellbeing in the lockdown...[but] there's no breakdown on where that activity is coming from in the day...I need to get better at...turn[ing] the activity on...to differentiate how much of my activity is actually coming from something good, and then how much of it is from walking around.* [T4, Withings, autumn]

The decrease in interaction in the autumn relates to the final theme of managing technology use in the mesocontexts of school life.

Theme 4: Managing Mesocontextual Barriers Could Offer a Framework for a Caring School Culture

Overview

Barriers to engagement were also external to the technology. All the teachers using their technology in the physical school environment expressed a clear sense of competition between technology interaction and occupational demands. In addition, the ongoing pandemic meant that routines were not the same and the extra or different demands on teachers' time sometimes eclipsed the use of the technology. Some wondered how their interactions would have been different had they begun using their technology while on school premises. Despite the mesocontextual barriers of time, space, and social and cultural norms, teachers' raised awareness of stress management through technology notifications or data meant some did still try to activate stress management strategies during the school day.

Teachers considered that by leveraging personal technologies, schools could embed self-care from the start of a teacher's career. The aim would be a school culture and environment in which self-management of stress became normative.

Subtheme 4.1: Competing Workplace Demands Can Overwhelm Use

Teachers described a frenetic working environment in school in the autumn, which had not been the case during lockdown and working from home. Four (66%) teachers talked about responding to technology alerts or notifications on the school premises; 3 of whom described this being done "on the run" between activities—if they responded at all. This was despite participant unanimity on the data-facilitated opportunity to reflect on personal well-being. This temporal barrier to interaction was possibly more noticeable due to teachers experiencing the contrast of job demands between working from home and working back in school:

> *I feel motivated when it beeps to tell me that I've done so many thousands of steps in the day, but I'm running around like a headless chicken so much that it [the technology] could be ringing or [alerting] emails or*

XSL•FO
RenderX

> *texts, and I don't even notice it.* [T2, Withings, autumn]

Teachers still used their technology but most interacted with it less when back on school premises, some putting this down to the ongoing disrupting effects of COVID-19. Wysa was described as needing a "quiet space," and both Wysa and Teacher Tapp needed time and "work" for interaction:

> *But because there's a pandemic and things arise, you don't have the routine...The [Teacher Tapp] alarm goes off at 3.30 but I'm always caught in the middle of something...with people getting sent home or staggered departures or parents wanting to meet you, it just fell behind.* [T5, Teacher Tapp, autumn]

Subtheme 4.2: A Physical Environment That Facilitates Taking a Break and Privacy

Staff talked about the need to have a brief respite from being in public view or always "performing" but having no staff room to facilitate this. To gain just a few minutes' peace, 5 (63%) out of 8 teachers in the summer described that before the pandemic they had hid in the toilets, in the comer of a classroom, or in a colleague's office or gone outdoors where they could not be found. One teacher said there was no point in trying to have a break during the school day as she would always be disturbed. Now that they wanted to use their technology or act on the data to de-stress, they described being more aware of the physical barriers to doing so. Others talked about making an extra effort in the autumn to act on the knowledge gained from their technology in the summer when there were few physical constraints:

> *I think things that I learnt over the lockdown, and with the technology, I definitely have brought into this term.... [I am] trying to take a lunch break.... I'm just trying to take at least 15 minutes of the day to just sit down because I wasn't even sitting down, really, before. Ever.* [T4, Withings, autumn]

For one teacher, teaching in a new environment having moved school in the autumn, the added structures in place meant that having a break was more realistic:

> *I think just better structures, so different things happening at the different times that you know about and everybody just sticks to what they're supposed to do...But break, lunch, I am on duty sometimes but I make sure I'm supporting the staff. But yes, to get a break, again, that's down to structure and routine so I'm more able to get a break [now].* [T5, Teacher Tapp, autumn]

Subtheme 4.3: Making It Socially Acceptable to Talk About Stress

From the start of the study, teachers had reported a reluctance to talk about stress in school or to share sentiments of stress with some of their colleagues:

> *I think teachers' stress is still quite taboo, to be honest, within the workplace.... I think it can be seen as a sign of weakness.* [T4, Withings, summer]

Having used their technology, some teachers considered that data could be used to demonstrate objectively to the leadership where situations in school were more stressful. If leadership encouraged discussion of stress based on data this could facilitate both technology use and targeted problem-solving:

> *I think if you could show them the value of using it, [say] with the tool, these [data show] where I'm most stressed, this is where I'm not... if you put me in here this is where I stress. ...if we could show that there's ways of tracking stress, I don't know, they might value it.* [T6, Withings, autumn]

Subtheme 4.4: Normalizing and Personalizing Self-Care With Technology From the Start

Some participants described technologies as an opportunity to introduce self-care right from the start of the profession. If the school leadership facilitated this it would send the cultural message from the outset that teachers' self-care is important:

> *Yes...I think that we should start to use them [technologies] earlier on in [our] teacher career, because then it helps you to focus on your wellbeing as an individual, and help you with your stress, your work/life balance.... As a teacher, you don't actually look after you, you look after everybody else's people. But if you can't look after yourself, it's going to be very difficult for you to look after others.* [T1, Dalyio and Wysa, autumn]

Some considered that using technology for self-care would also remind teachers that they had a responsibility to look after themselves, not that they bore the entire burden of stress management but that they had an important part to play:

> *I think there's a lot to be done, and I think technology is certainly a way that we could manage that, because I think it's very easy, and I recognise this in myself big time...to blame the job, blame the role, blame SLT, blame the marking, blame the kids, which obviously is a huge factor. But I think actually there's a lot more we could do as individuals, just to take that time to dedicate to ourselves, reduce our stress levels in whatever way that might be.* [T4, Withings, summer]

Discussion

Principal Findings

To the best of our knowledge, this is the first study to describe facilitators and barriers to personally chosen technology support for stress self-management in the educational context, along with insight into how schools could facilitate such interventions. The findings complement those of teachers' experiences, which demonstrate how a more holistic understanding of stress, and permission to manage it is generated as teachers found meaning in their data and interactions [107].

As shown in our findings, the themes generated indicate specific features, data collation, and technology-supported concepts that facilitated teachers' use during their working day. In contrast, the barriers to use centered around a lack of information and signposting and external, school-related mesocontexts that

XSL•FO
RenderX

inhibited use or putting their stress-relieving action into practice as planned. These themes are now unpacked further.

Insight Into Design Facilitators of Technology Use by Teachers

Teachers expressed their appreciation of the wearable appearing to be a watch, not an obvious tracking device. They also liked the rapidity with which Withings watch data could be viewed and the Daylio app could be used, thus preventing potentially stigmatizing exposure. These themes reflect the persistent taboo of experiencing stress, and teachers' fear of not being perceived to be able to cope. Discreet "design" has been noted previously as a facilitator for recording feelings [49,114] and for the use of technologies by employees for stress management [93,115]. It has also been observed that the esthetic of design of wearables is important [116]. Added to the knowledge that short interventions are important for stress management for providers and users [95], including in constrained workplaces [98], we suggest designers should ensure that where the perceived risk of stigma in the workplace remains, design camouflage would encourage engagement for teachers.

Immediacy is increasingly being facilitated by wearables and sensors [117]. The Withings watch enabled stress awareness with cognitive ease both through passively collated data and through haptic notifications. Haptic stimulation has been shown to have a positive influence on attitude to accompanying messages and such a physical cue is motivational [118]. Our findings suggest that for teachers in their polychronically demanding environments, such a physical cue facilitates engagement up to a point, dependent on the individual and competing demands. The exception to immediacy was Wysa, which required deliberate, sustained engagement. Temporal constraints precluded use in school, but Wysa (the chatbot) was described as a "best friend" for giving contextually applicable advice that was both educational and transferable. Such high-effort, problem-solving interventions have been found to be more effective for IT workers than low-effort, positive distractions [8]. Therefore, we suggest that cognitive ease can also enable digitally supported stress management by virtue of ease of application of learning, not just the immediacy of information.

This applicability of the chatbot advice also links to the finding of the importance of contextually mediated reflection on patterns for engagement. This aligns with research hinting that more contextualization facilitates engagement [119] and the noted lack of contextualization in mood-tracking apps used by the general population [49]. A combination of data patterns that could be correlated by the teacher to their circumstances, and the legitimacy of the intervention as facilitated by the school leadership, enabled teachers to reflect on their stress associations and plan mitigating actions.

Teachers identified automated prompts for taking more steps or meditating, or for messages of affirmation when an activity had been undertaken, as helpful. This complements teachers' reports of appreciating congratulatory text messages of support via WhatsApp [120]. Elsewhere, police officers using wearables to track activity [121] and working-age women with technologies meeting self-set goals [122] have also reported the motivational aspect of digital messages of encouragement. Prompts or reminders for stress management can be preplanned or automated through algorithmic analysis of data [35,82]. Conversely, when prompts are not received, stress awareness among Chinese adults was reduced as was their ability to manage their stress [41]. Even for people in less-restricted roles than teaching, receptivity to automated supportive messaging is highly dependent on contextual variables [123]. Future research could consider how messaging could be further enhanced to enhance stress management among employees in temporally demanding roles.

Barriers to Technology Use Described by the Teachers

This study validated the established need for good data visualization with improvements suggested for the Daylio app and readability of the Withings watch's digital display screen. Stress data visualization has been tested at the organizational level [39], but this would be difficult to translate to schools. The paucity of setup information supplied with the Withings watch was commented on by 5 (83%) out of the 6 users. Teachers resorted to YouTube (not the Withings website) for guidance, although not knowing what constituted "normal" data remained a problem for 2 of them. Social sensemaking with personal informatics has demonstrated that value can be gained from contextualizing data through comparison with peers [124], but the COVID-19 pandemic would have been a barrier to this. The app users also cited apparent design assumptions of intuitive use, particularly with Daylio, which delayed full adoption. These barriers to adoption extend the finding of inadequate technical instruction in a digital mental health review [125], adding a specific example of a wearable consumer technology adopted for stress management (out of 208 papers, only n=2, 0.01% were for wearables in the review by Borghouts et al [125], and these were for tracking serious mental illness). Wearable connectivity was found to be limited by Bluetooth range, a challenge noted generally with digital stress support [126], which for teachers specifically related to the proximity of their smartphone, often not kept with them [52]. Designers therefore need to remember that users still need comprehensive navigational advice and that they cannot assume proximity between wearables and smartphones for all occupations.

Some teachers were disappointed that the data indicated sleep when they were awake, despite Withings' reputation for accuracy [127]. This has been noted in studies of Fitbit [61,128]. Liang et al [128] suggested the companion app should allow the user to edit their inaccurate data, an important consideration because inaccurate data is generally a reason for abandonment [129]. This remains a consideration as users perceive stress algorithms to be at odds with their subjective experiences [64].

No other inaccuracies were noted by teachers in this study. Several teachers wished that their technology offered alternative strategies, for example, relaxation as well as education, partly because stress symptoms can change over time. This adds to similar reports for a greater plurality of features or strategies from other studies [93,130-132]. However, this must be balanced against the extent of choice [133], as too much choice in digital tools could be overwhelming [134].

XSL•FO
RenderX

Temporal constraints led to reduced frequency of interaction with technology for most of the staff when back in school, a barrier to interaction that can only be discerned in real-world evaluations [135]. Notably, this was only detected because of the extraordinary circumstances of teachers able to compare working from home with working back in school and because sustained engagement was still demonstrated months after initiation. The text-heavy Wysa (never accessed in school) and Teacher Tapp (significantly reduced interaction in school) apps were the most affected. Counson et al [98] found that vocational intensity was a barrier to engagement with a stress management app cocreated with doctors and professional demands constrained time and enthusiasm for digital health interventions generally [136]. Thus, organization-level solutions are likely to be required as well, connecting to the final finding that teachers' desire is for a workplace culture that facilitates self-care and caring. Such a culture requires the facilitation of mesocontexts, such as time and physical environment. Although other studies have used machine learning to judge the best timings for stress interventions for students and in offices [10,38,82,131], the polychronic nature of teachers' work would make finding patterns in behavior and just-in-time interventions more challenging. In addition, during the school day, stress management can require privacy, for example, in the form of a staff room, with lack of privacy being a known deterrent to interaction generally for digital mental health [125].

Implications for the School Workplace

The findings on both facilitators and barriers to teachers' engagement suggest that school leadership could leverage technology to support stress management. Some alerts can prompt teachers' just-in-time stress relief strategies or simple self-care such as having a drink. Yet, teachers also need supportive mesocontexts such as a private space in which to relax, protected time when they are undisturbed, and a culture in which stress is seen as a warning and not a weakness. This need for a positive organizational culture for stress support has been noted before by office workers considering the use of personal digital tracking and e-coaching [137]. As several teachers commented, what is the point of drinking water to keep hydrated, if there is no time or provision (in their building) or even acceptance that they will then need to use the bathroom.

Strengths and Limitations

Rich, contextualized data on the adoption of and influences on technology use were collected over an extended period of up to 27 weeks, enabling deep and detailed insights into teachers' engagement. All 8 heads of year completed the summer study and just 2 (25%) were lost to follow-up in the extended period. This longitudinal study provides evidence of the acceptability of consumer technology to teachers for stress management support and which features were of value for self-care in an occupation renowned for its stressfulness.

Taking part in the research inevitably meant there was some extrinsic motivation and would have raised awareness of stress and self-care; therefore, there could have been a digital placebo effect [138-140]. The pandemic led to the exceptional circumstances of both more attention being given to teachers' well-being by the school leadership and the teachers beginning to use their technologies while working from home. This meant that all participants had time to adopt their technology in a quieter environment and could compare facilitators and barriers between the 2 settings. Further research could test digital stress support against an active control.

The choice of technology was framed by a specific taxonomy which enabled choice based on strategy and conceptualization of interaction. Given the pandemic, workshops were held on the web which prevented hands-on exploration of technologies before selection. However, the main author was available for queries throughout the study. Therefore, another limitation was that autonomy was limited to the technologies that populated the taxonomy, selected based on suitability, availability, usability, security, validity, and cost. In addition, there was no classification of technologies according to detection, management or prevention, just-in-time recovery, or resilience building. These concepts and characteristics could be tested to see if they assist in technology selection and self-care. Apart from Wysa, no technology was chosen that facilitated treatment or guided self-management nor was entertainment based. The reasons for this as well as encouraging a wider variety of interventions could be investigated further.

Another strength of this longitudinal study was that it was undertaken with teachers who taught many socioeconomically deprived students, typically a more stressful setting. Teachers in more affluent areas or working with more economically advantaged children are known to experience less stress [141-143]. Thus, the findings should benefit teachers in more challenging situations. It was poignant to hear teachers in the autumn term talking about children returning to school after the pandemic without even a pen, let alone lunch money.

The participant sample cannot be claimed to be representative of all secondary school heads of year, although this study makes an important contribution, given the dearth of research with middle managers. The sample was self-selecting and only one head of year was male, but the racial mix was diverse.

Conclusions

Consumer technologies are quick, easy, and discreet to use; offer encouragement; and enable personalization. They could have an explicit role in facilitating teachers' understanding and managing their stress. Developers should not make design assumptions of users intuitively discovering all features, and the ability to manually amend some data should be facilitated. School leadership has an important part to play in facilitating stress management, and technology use could contribute to a culture in which self-care is normative and discussions of stress are socially acceptable. The convenience and ease of technology engagement could be leveraged at the organizational level to integrate stress support. This invites the proposition that a teacher's ability to act on technology data is a proxy measure for a caring school culture. This would require both protected time and space on the school premises. We hope our findings are of interest to policy makers and encourage developers of digital interventions to work with educators to ensure contextual insights inform the design of health and well-being technology.

XSL•FO
RenderX

Acknowledgments

This research received funding from the Engineering and Physical Sciences Research Council Doctoral Training Programme (UCL EP/R513143/1). No generative artificial intelligence was used in any portion of this manuscript. The authors are extremely grateful to the teachers who took part in this study despite the unprecedented disruption to their lives from having to adapt to teaching remotely during the pandemic, both domestically and occupationally.

Data Availability

The data generated and analyzed during this study are restricted because the transcripts in their entirety could potentially identify participants and are confidential.

Authors' Contributions

JBM was responsible for the formulation of research goals and aims, undertook the investigation and analysis, and wrote the first draft. All the authors reviewed and provided feedback on the final manuscript.

Conflicts of Interest

None declared.

References

1. Jackson M. The stress of life: a modern complaint? Lancet 2014 Jan 25;383(9914):300-301 [FREE Full text] [doi: 10.1016/s0140-6736(14)60093-3] [Medline: 24475483]
2. Health and safety statistics. Health and Safety Executive. URL: https://www.hse.gov.uk/statistics/ [accessed 2023-06-30]
3. Mental wellbeing at work. National Institute for Health and Care Excellence. 2022 Mar 2. URL: https://www.nice.org.uk/guidance/ng212 [accessed 2024-09-22]
4. McEwen BS, Akil H. Revisiting the stress concept: implications for affective disorders. J Neurosci 2020 Jan 02;40(1):12-21 [FREE Full text] [doi: 10.1523/JNEUROSCI.0733-19.2019] [Medline: 31896560]
5. Steptoe A, Kivimäki M. Stress and cardiovascular disease. Nat Rev Cardiol 2012 Apr 03;9(6):360-370. [doi: 10.1038/nrcardio.2012.45] [Medline: 22473079]
6. May RW, Seibert GS, Sanchez-Gonzalez MA, Fincham FD. School burnout and heart rate variability: risk of cardiovascular disease and hypertension in young adult females. Stress 2018 May 30;21(3):211-216. [doi: 10.1080/10253890.2018.1433161] [Medline: 29382258]
7. Heber E, Lehr D, Ebert DD, Berking M, Riper H. Web-based and mobile stress management intervention for employees: a randomized controlled trial. J Med Internet Res 2016 Jan 27;18(1):e21 [FREE Full text] [doi: 10.2196/jmir.5112] [Medline: 26818683]
8. Howe E, Suh J, Morshed MB, McDuff D, Rowan K, Hernandez J, et al. Design of digital workplace stress-reduction intervention systems: effects of intervention type and timing. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3502027]
9. Ebert DD, Lehr D, Heber E, Riper H, Cuijpers P, Berking M. Internet- and mobile-based stress management for employees with adherence-focused guidance: efficacy and mechanism of change. Scand J Work Environ Health 2016 Sep 01;42(5):382-394 [FREE Full text] [doi: 10.5271/sjweh.3573] [Medline: 27249161]
10. Paredes P, Gilad-Bachrach R, Czerwinski M, Roseway A, Rowan K, Hernandez J. PopTherapy: coping with stress through pop-culture. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. 2014 Presented at: PervasiveHealth '14; May 20-23, 2014; Oldenburg, Germany. [doi: 10.4108/icst.pervasivehealth.2014.255070]
11. Krusche A, Cyhlarova E, King S, Williams JM. Mindfulness online: a preliminary evaluation of the feasibility of a web-based mindfulness course and the impact on stress. BMJ Open 2012 Nov 29;2(3):e003498 [FREE Full text] [doi: 10.1136/bmjopen-2011-000803] [Medline: 22614170]
12. Lau N, O'Daffer A, Colt S, Yi-Frazier JP, Palermo TM, McCauley E, et al. Android and iPhone mobile apps for psychosocial wellness and stress management: systematic search in app stores and literature review. JMIR Mhealth Uhealth 2020 May 22;8(5):e17798 [FREE Full text] [doi: 10.2196/17798] [Medline: 32357125]
13. Li I, Dey A, Forlizzi J. A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010 Presented at: CHI '10; April 10-15, 2010; Atlanta, GA. [doi: 10.1145/1753326.1753409]

XSL•FO
RenderX

14. Epstein DA, Ping A, Fogarty J, Munson SA. A lived informatics model of personal informatics. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan. [doi: 10.1145/2750858.2804250]
15. Rooksby J, Rost M, Morrison A, Chalmers M. Personal tracking as lived informatics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON. [doi: 10.1145/2556288.2557039]
16. Epstein DA, Caldeira C, Figueiredo MC, Lu X, Silva LM, Williams L, et al. Mapping and taking stock of the personal informatics literature. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Dec 18;4(4):1-38. [doi: 10.1145/3432231]
17. Weber S, Lorenz C, Hemmings N. Improving stress and positive mental health at work via an app-based intervention: a large-scale multi-center randomized control trial. Front Psychol 2019 Dec 6;10:2745 [FREE Full text] [doi: 10.3389/fpsyg.2019.02745] [Medline: 31866915]
18. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
19. Morris ME, Kathawala Q, Leen TK, Gorenstein EE, Guilak F, Labhard M, et al. Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. J Med Internet Res 2010 Apr 30;12(2):e10 [FREE Full text] [doi: 10.2196/jmir.1371] [Medline: 20439251]
20. Leung SS, Chiang VC, Chui YY, Lee AC, Mak YW. Feasibility and potentials of online support for stress management among secondary school teachers. Stress Health 2010 Aug 18;27(3):e282-e286. [doi: 10.1002/smi.1347]
21. Paredes P, Chan M. CalmMeNow: exploratory research and design of stress mitigating mobile interventions. In: CHI '11 Extended Abstracts on Human Factors in Computing Systems. 2011 Presented at: CHI EA '11; May 7-12, 2011; Vancouver, BC. [doi: 10.1145/1979742.1979831]
22. Litvin S, Saunders R, Maier MA, Lüttke S. Gamification as an approach to improve resilience and reduce attrition in mobile mental health interventions: a randomized controlled trial. PLoS One 2020 Sep 2;15(9):e0237220 [FREE Full text] [doi: 10.1371/journal.pone.0237220] [Medline: 32877425]
23. Hoffmann A, Faust-Christmann CA, Zolynski G, Bleser G. Gamification of a stress management app: results of a user study. In: Proceedings of the 8th International Conference on Design, User Experience, and Usability. Application Domains. 2019 Presented at: DUXU 2019; July 26-31, 2019; Orlando, FL. [doi: 10.1007/978-3-030-23538-3_23]
24. Heber E, Ebert DD, Lehr D, Cuijpers P, Berking M, Nobis S, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017 Feb 17;19(2):e32 [FREE Full text] [doi: 10.2196/jmir.5774] [Medline: 28213341]
25. Ly KH, Asplund K, Andersson G. Stress management for middle managers via an acceptance and commitment-based smartphone application: a randomized controlled trial. Internet Interv 2014 Jul;1(3):95-101. [doi: 10.1016/j.invent.2014.06.003]
26. Hoffmann A, Christmann CA, Bleser G. Gamification in stress management apps: a critical app review. JMIR Serious Games 2017 Jun 07;5(2):e13 [FREE Full text] [doi: 10.2196/games.7216] [Medline: 28592397]
27. Hwang H, Kim SM, Netterstrøm B, Han DH. The efficacy of a smartphone-based app on stress reduction: randomized controlled trial. J Med Internet Res 2022 Feb 15;24(2):e28703 [FREE Full text] [doi: 10.2196/28703] [Medline: 35166687]
28. Ptakauskaite N, Cox AL, Berthouze N. Knowing what you're doing or knowing what to do: how stress management apps support reflection and behaviour change. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI 2018; April 21-26, 2018; Montreal, Canada.
29. Smith EN, Santoro E, Moraveji N, Susi M, Crum AJ. Integrating wearables in stress management interventions: promising evidence from a randomized trial. Int J Stress Manage 2020 May;27(2):172-182. [doi: 10.1037/str0000137]
30. Hao T, Walter KN, Ball MJ, Chang HY, Sun S, Zhu X. StressHacker: towards practical stress monitoring in the wild with smartwatches. AMIA Annu Symp Proc 2017;2017:830-838 [FREE Full text] [Medline: 29854149]
31. Beyer H, Holtzblatt K. Contextual Design: Defining Customer-Centered Systems. San Francisco, CA: Morgan Kaufmann Publishers Inc; 1997.
32. Rogers Y, Sharp H, Preece J. Interaction Design: Beyond Human-Computer Interaction. Hoboken, NJ: John Wiley & Sons; 2019.
33. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
34. Smith W, Wadley G, Webber S, Ploderer B, Lederman R. Unbounding the interaction design problem: the contribution of HCI in three interventions for well-being. In: Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design. 2014 Presented at: OzCHI '14; December 2-5, 2014; Sydney, Australia. [doi: 10.1145/2686612.2686672]
35. Kim T, Kim H, Lee HY, Goh H, Abdigapporov S, Jeong M, et al. Prediction for retrospection: integrating algorithmic stress prediction into personal informatics systems for college students' mental health. In: Proceedings of the 2022 CHI Conference

XSL•FO
RenderX

on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3517701]
36. MacLean D, Roseway A, Czerwinski M. MoodWings: a wearable biofeedback device for real-time stress intervention. In: Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments. 2013 Presented at: PETRA '13; May 29-31, 2013; Rhodes, Greece. [doi: 10.1145/2504335.2504406]
37. Sanches P, Hook K, Vaara E, Weymann C, Bylund M, Ferreira P, et al. Mind the body!: designing a mobile stress management application encouraging personal reflection. In: Proceedings of the 8th ACM Conference on Designing Interactive Systems. 2010 Presented at: DIS '10; August 16-20, 2010; Aarhus, Denmark. [doi: 10.1145/1858171.1858182]
38. Sano A, Johns P, Czerwinski M. Designing opportune stress intervention delivery timing using multi-modal data. In: Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017 Presented at: ACII 2017; October 23-26, 2017; San Antonio, TX. [doi: 10.1109/acii.2017.8273623]
39. Stepanovic S, Mozgovoy V, Mettler T. Designing visualizations for workplace stress management: results of a pilot study at a swiss municipality. In: Proceedings of the 18th IFIP WG 8.5 International Conference. 2019 Presented at: EGOV 2019; September 2-4, 2019; San Benedetto Del Tronto, Italy. [doi: 10.1007/978-3-030-27325-5_8]
40. Xue M, Liang RH, Yu B, Funk M, Hu J, Feijs L. AffectiveWall: designing collective stress-related physiological data visualization for reflection. IEEE Access 2019;7:131289-131303. [doi: 10.1109/access.2019.2940866]
41. Ding X, Wei S, Gui X, Gu N, Zhang P. Data engagement reconsidered: a study of automatic stress tracking technology in use. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445763]
42. Pakhomov SV, Thuras PD, Finzel R, Eppel J, Kotlyar M. Using consumer-wearable technology for remote assessment of physiological response to stress in the naturalistic environment. PLoS One 2020 Mar 25;15(3):e0229942 [FREE Full text] [doi: 10.1371/journal.pone.0229942] [Medline: 32210441]
43. Sano A, Picard RW. Stress recognition using wearable sensors and mobile phones. In: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013 Presented at: ACII 2013; September 2-5, 2013; Geneva, Switzerland. [doi: 10.1109/acii.2013.117]
44. Gjoreski M, Luštrek M, Gams M, Gjoreski H. Monitoring stress with a wrist device using context. J Biomed Inform 2017 Sep;73:159-170 [FREE Full text] [doi: 10.1016/j.jbi.2017.08.006] [Medline: 28803947]
45. Carolan S, Harris PR, Cavanagh K. Improving employee well-being and effectiveness: systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. J Med Internet Res 2017 Jul 26;19(7):e271 [FREE Full text] [doi: 10.2196/jmir.7583] [Medline: 28747293]
46. Stratton E, Lampit A, Choi I, Calvo RA, Harvey SB, Glozier N. Effectiveness of eHealth interventions for reducing mental health conditions in employees: a systematic review and meta-analysis. PLoS One 2017 Dec 21;12(12):e0189904 [FREE Full text] [doi: 10.1371/journal.pone.0189904] [Medline: 29267334]
47. Phillips EA, Gordeev VS, Schreyögg J. Effectiveness of occupational e-mental health interventions: a systematic review and meta-analysis of randomized controlled trials. Scand J Work Environ Health 2019 Nov 01;45(6):560-576 [FREE Full text] [doi: 10.5271/sjweh.3839] [Medline: 31184758]
48. Ryan JC, Williams G, Wiggins BW, Flitton AJ, McIntosh JT, Carmen MJ, et al. Exploring the active ingredients of workplace physical and psychological wellbeing programs: a systematic review. Transl Behav Med 2021 May 25;11(5):1127-1141. [doi: 10.1093/tbm/ibab003] [Medline: 33677571]
49. Schueller SM, Neary M, Lai J, Epstein DA. Understanding people's use of and perspectives on mood-tracking apps: interview study. JMIR Ment Health 2021 Aug 11;8(8):e29368 [FREE Full text] [doi: 10.2196/29368] [Medline: 34383678]
50. Leung SS, Wah Mak Y, Yu Chui Y, Chiang VC, Lee AC. Occupational stress, mental health status and stress management behaviors among secondary school teachers in Hong Kong. Health Educ J 2010 Mar 15;68(4):328-343. [doi: 10.1177/0017896909349255]
51. Austin V, Shah S, Muncer S. Teacher stress and coping strategies used to reduce stress. Occup Ther Int 2005;12(2):63-80. [doi: 10.1002/oti.16] [Medline: 16136865]
52. Manning JB, Blandford A, Edbrooke-Childs J, Marshall P. How contextual constraints shape midcareer high school teachers' stress management and use of digital support tools: qualitative study. JMIR Ment Health 2020 Apr 27;7(4):e15416 [FREE Full text] [doi: 10.2196/15416] [Medline: 32338623]
53. Ayobi A, Marshall P, Cox AL. Reflections on 5 years of personal informatics: rising concerns and emerging directions. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2016 Presented at: CHI EA '16; May 7-12, 2016; San Jose, CA. [doi: 10.1145/2851581.2892406]
54. Chan L, Swain VD, Kelley C, de Barbaro K, Abowd GD, Wilcox L. Students' experiences with ecological momentary assessment tools to report on emotional well-being. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar 26;2(1):1-20. [doi: 10.1145/3191735]
55. Potapov K, Vasalou A, Lee V, Marshall P. What do teens make of personal informatics?: young people's responses to self-tracking practices for self-determined motives. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445239]

XSL•FO
RenderX

56. Pinheiro N, Couceiro R, Henriques J, Muehlsteff J, Quintal I, Goncalves L, et al. Can PPG be used for HRV analysis? Annu Int Conf IEEE Eng Med Biol Soc 2016 Aug;2016:2945-2949. [doi: 10.1109/EMBC.2016.7591347] [Medline: 28268930]
57. Byrne ML, Lind MN, Horn SR, Mills KL, Nelson BW, Barnes ML, et al. Using mobile sensing data to assess stress: associations with perceived and lifetime stress, mental health, sleep, and inflammation. Digit Health 2021 Aug 27;7:20552076211037227 [FREE Full text] [doi: 10.1177/20552076211037227] [Medline: 34777852]
58. Adams P, Rabbi M, Rahman T, Matthews M, Voida A, Gay G, et al. Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. 2014 Presented at: PervasiveHealth '14; May 20-23, 2014; Oldenburg, Germany. [doi: 10.4108/icst.pervasivehealth.2014.254959]
59. Nelson BW, Low CA, Jacobson N, Areán P, Torous J, Allen NB. Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. NPJ Digit Med 2020 Jun 26;3:90 [FREE Full text] [doi: 10.1038/s41746-020-0297-4] [Medline: 32613085]
60. van Dijk ET, Westerink JH, Beute F, IJsselsteijn WA. In sync: the effect of physiology feedback on the match between heart rate and self-reported stress. Biomed Res Int 2015;2015:134606 [FREE Full text] [doi: 10.1155/2015/134606] [Medline: 26146611]
61. Liang Z, Ploderer B. Sleep tracking in the real world: a qualitative study into barriers for improving sleep. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction. 2016 Presented at: OzCHI '16; November 29-December 2, 2016; Tasmania, Australia. [doi: 10.1145/3010915.3010988]
62. Kersten-van Dijk ET, Westerink JH, Beute F, IJsselsteijn WA. Personal informatics, self-insight, and behavior change: a critical review of current literature. Hum Comput Interact 2017 Feb 17;32(5-6):268-296. [doi: 10.1080/07370024.2016.1276456]
63. Harrison DB. The Self-Tracker's Journey: situated engagement and non-engagement with personal informatics systems over time. University College London. 2020. URL: https://discovery.ucl.ac.uk/id/eprint/10094630/ [accessed 2024-09-23]
64. Smitt MS, Montakhabi M, Morton J, van Leeuwen C, Bombeke K, Jacobs A. Users' perceptions of a digital stress self-monitoring application: research insights to design a practical innovation. In: Proceedings of the 16th International Conference on Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies. 2022 Presented at: UAHCI 2022; June 26-July 1, 2022; Virtual Event. [doi: 10.1007/978-3-031-05028-2_22]
65. Selye H. Stress without distress. In: Serban G, editor. Psychopathology of Human Adaptation. Boston, MA: Springer; 1976.
66. Hernandez J, Paredes P, Roseway A, Czerwinski M. Under pressure: sensing stress of computer users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON. [doi: 10.1145/2556288.2557165]
67. Lu H, Frauendorfer D, Rabbi M, Mast MS, Chittaranjan GT, Campbell AT, et al. StressSense: detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. 2012 Presented at: UbiComp '12; September 5-8, 2012; Pittsburgh, PA. [doi: 10.1145/2370216.2370270]
68. Vildjiounaite E, Kallio J, Kyllönen V, Nieminen M, Määttänen I, Lindholm M, et al. Unobtrusive stress detection on the basis of smartphone usage data. Pers Ubiquit Comput 2018 Jan 17;22:671-688. [doi: 10.1007/s00779-017-1108-z]
69. Runge N, Haarman S, Fisher MH. Using Fitbit fitness trackers to measure teacher stress and coping. Int J Soc Policy Educ 2020;2(3):56-70.
70. Kocielnik R, Sidorova N. Personalized stress management: enabling stress monitoring with LifelogExplorer. Künstl Intell 2015 Jan 23;29:115-122. [doi: 10.1007/s13218-015-0348-1]
71. Magtibay K, Umapathy K. A review of tools and methods for detection, analysis, and prediction of allostatic load due to workplace stress. IEEE Trans Affective Comput 2024;15(1):357-375. [doi: 10.1109/taffc.2023.3273201]
72. Holden RJ, Valdez RS, Schubert CC, Thompson MJ, Hundt AS. Macroergonomic factors in the patient work system: examining the context of patients with chronic illness. Ergonomics 2017 Jan;60(1):26-43 [FREE Full text] [doi: 10.1080/00140139.2016.1168529] [Medline: 27164171]
73. Carayon P. Human factors of complex sociotechnical systems. Appl Ergon 2006 Jul;37(4):525-535. [doi: 10.1016/j.apergo.2006.04.011] [Medline: 16756937]
74. Jumisko-Pyykkö S, Vainio T. Framing the context of use for mobile HCI. Int J Mobile Hum Comput Interact 2010;2(4):1-28. [doi: 10.4018/jmhci.2010100101]
75. Ball SJ, Maguire M, Braun A. How Schools Do Policy: Policy Enactments in Secondary Schools. Milton Park, UK: Taylor & Francis; 2011.
76. Nunes F, Verdezoto N, Fitzpatrick G, Kyng M, Grönvall E, Storni C. Self-care technologies in HCI: trends, tensions, and opportunities. ACM Trans Comput Hum Interact 2015 Dec 14;22(6):1-45. [doi: 10.1145/2803173]
77. Oinas-Kukkonen HA. A foundation for the study of behavior change support systems. Pers Ubiquit Comput 2012 Jul 19;17:1223-1235. [doi: 10.1007/s00779-012-0591-5]
78. Mohr DC, Schueller SM, Montague E, Burns MN, Rashidi P. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J Med Internet Res 2014 Jun 05;16(6):e146 [FREE Full text] [doi: 10.2196/jmir.3077] [Medline: 24905070]

XSL•FO
RenderX

79. Ketelaar SM, Nieuwenhuijsen K, Bolier L, Smeets O, Sluiter JK. Improving work functioning and mental health of health care employees using an e-mental health approach to workers' health surveillance: pretest-posttest study. Saf Health Work 2014 Dec;5(4):216-221 [FREE Full text] [doi: 10.1016/j.shaw.2014.08.002] [Medline: 25516815]
80. Deitz D, Cook RF, Hersch RK, Leaf S. Heart healthy online: an innovative approach to risk reduction in the workplace. J Occup Environ Med 2014 May;56(5):547-553 [FREE Full text] [doi: 10.1097/JOM.0000000000000148] [Medline: 24806568]
81. Bolier L, Ketelaar SM, Nieuwenhuijsen K, Smeets O, Gärtner FR, Sluiter JK. Workplace mental health promotion online to enhance well-being of nurses and allied health professionals: a cluster-randomized controlled trial. Internet Interv 2014 Oct;1(4):196-204. [doi: 10.1016/j.invent.2014.10.002]
82. Lee K, Cho H, Toshnazarov K, Narziev N, Rhim SY, Han K, et al. Toward future-centric personal informatics: expecting stressful events and preparing personalized interventions in stress management. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376475]
83. Lee MD, Kang X, Hanrahan N. Addressing cultural contexts in the management of stress via narrative and mobile technology. Stud Health Technol Inform 2014;199:173-177. [Medline: 24875715]
84. Arenas DL, Viduani A, Araujo RB. Therapeutic use of role-playing game (RPG) in mental health: a scoping review. Simul Gaming 2022 Mar 02;53(3):285-311. [doi: 10.1177/10468781211073720]
85. Renfrew ME, Morton DP, Morton JK, Przybylko G. The influence of human support on the effectiveness of digital mental health promotion interventions for the general population. Front Psychol 2021 Aug 19;12:716106 [FREE Full text] [doi: 10.3389/fpsyg.2021.716106] [Medline: 34489818]
86. Zarski AC, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146 [FREE Full text] [doi: 10.2196/jmir.4493] [Medline: 27357528]
87. Thiart H, Lehr D, Ebert DD, Sieland B, Berking M, Riper H. Log in and breathe out: efficacy and cost-effectiveness of an online sleep training for teachers affected by work-related strain--study protocol for a randomized controlled trial. Trials 2013 Jun 11;14:169 [FREE Full text] [doi: 10.1186/1745-6215-14-169] [Medline: 23759035]
88. Thiart H, Lehr D, Ebert DD, Berking M, Riper H. Log in and breathe out: internet-based recovery training for sleepless employees with work-related strain - results of a randomized controlled trial. Scand J Work Environ Health 2015 Mar 15;41(2):164-174 [FREE Full text] [doi: 10.5271/sjweh.3478] [Medline: 25590336]
89. Ebert DD, Lehr D, Boß L, Riper H, Cuijpers P, Andersson G, et al. Efficacy of an internet-based problem-solving training for teachers: results of a randomized controlled trial. Scand J Work Environ Health 2014 Nov 13;40(6):582-596 [FREE Full text] [doi: 10.5271/sjweh.3449] [Medline: 25121986]
90. Persson Asplund R, Dagöö J, Fjellström I, Niemi L, Hansson K, Zeraati F, et al. Internet-based stress management for distressed managers: results from a randomised controlled trial. Occup Environ Med 2018 Feb 30;75(2):105-113 [FREE Full text] [doi: 10.1136/oemed-2017-104458] [Medline: 28855344]
91. Ansley BM, Houchins DE, Varjas K, Roach A, Patterson D, Hendrick R. The impact of an online stress intervention on burnout and teacher efficacy. Teach Teach Educ 2021 Feb;98:103251. [doi: 10.1016/j.tate.2020.103251]
92. Klap J. Supporting and transforming school principals' wellbeing and leadership. Murdoch University. 2020. URL: https://researchportal.murdoch.edu.au/esploro/outputs/doctoral/Supporting-and-transforming-school-principals-wellbeing/991005544202907891#metrics [accessed 2023-06-30]
93. Carolan S, de Visser RO. Employees' perspectives on the facilitators and barriers to engaging with digital mental health interventions in the workplace: qualitative study. JMIR Ment Health 2018 Jan 19;5(1):e8 [FREE Full text] [doi: 10.2196/mental.9146] [Medline: 29351900]
94. Alhasani M, Mulchandani D, Oyebode O, Baghaei N, Orji R. A systematic and comparative review of behavior change strategies in stress management apps: opportunities for improvement. Front Public Health 2022 Feb 24;10:777567 [FREE Full text] [doi: 10.3389/fpubh.2022.777567] [Medline: 35284368]
95. Blankenhagel KJ, Theilig MM, Koch H, Witte AK, Zarnekow R. Challenges for preventive digital stress management systems - identifying requirements by conducting qualitative interviews. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019 Presented at: HICSS 2019; January 8-11, 2019; Maui, HI. [doi: 10.24251/hicss.2019.461]
96. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. Cambridge, UK: Cambridge University Press; 1991.
97. Tucker SJ, Lanningham-Foster LM, Murphy JN, Thompson WG, Weymiller AJ, Lohse C, et al. Effects of a worksite physical activity intervention for hospital nurses who are working mothers. AAOHN J 2011 Sep;59(9):377-386. [doi: 10.3928/08910162-20110825-01] [Medline: 21877670]
98. Counson I, Bartholomew A, Crawford J, Petrie K, Basarkod G, Moynihan V, et al. Development of the shift smartphone app to support the emotional well-being of junior physicians: design of a prototype and results of usability and acceptability testing. JMIR Form Res 2021 Dec 02;5(12):e26370 [FREE Full text] [doi: 10.2196/26370] [Medline: 34860662]

XSL•FO
RenderX

99. Sanatkar S, Counson I, Mackinnon A, Bartholomew A, Glozier N, Harvey S. Preliminary investigation of shift, a novel smartphone app to support junior doctors' mental health and well-being: examination of symptom progression, usability, and acceptability after 1 month of use. J Med Internet Res 2022 Sep 21;24(9):e38497 [FREE Full text] [doi: 10.2196/38497] [Medline: 36129745]
100. Havermans BM, Boot CR, Brouwers EP, Houtman IL, Anema JR, van der Beek AJ. Process evaluation of a digital platform-based implementation strategy aimed at work stress prevention in a health care organization. J Occup Environ Med 2018 Sep;60(9):e484-e491. [doi: 10.1097/JOM.0000000000001402] [Medline: 30199413]
101. Peters D, Deady M, Glozier N, Harvey S, Calvo RA. Worker preferences for a mental health app within male-dominated industries: participatory study. JMIR Ment Health 2018 Apr 25;5(2):e30 [FREE Full text] [doi: 10.2196/mental.8999] [Medline: 29695371]
102. Stratton E, Choi I, Peters D, Calvo RA, Harvey SB, Glozier N. Co-designing a web-based decision aid tool for employees disclosure of mental health conditions: a participatory study design using employee and organizational preferences. JMIR Form Res 2020 Nov 06;4(11):e23337 [FREE Full text] [doi: 10.2196/23337] [Medline: 33155982]
103. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. J Biomed Inform 2012 Feb;45(1):184-198 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.017] [Medline: 21925288]
104. Calvo RA, Peters D. Positive Computing: Technology for Wellbeing and Human Potential. Cambridge, MA: MIT Press; 2014.
105. Morley J, Floridi L. The limits of empowerment: how to reframe the role of mHealth tools in the healthcare ecosystem. Sci Eng Ethics 2020 Jun;26(3):1159-1183 [FREE Full text] [doi: 10.1007/s11948-019-00115-1] [Medline: 31172424]
106. Manning JB, Blandford A, Edbrooke-Childs J. Digital companion choice to support teachers' stress self-management: systematic approach through taxonomy creation. JMIR Form Res 2022 Feb 16;6(2):e32312 [FREE Full text] [doi: 10.2196/32312] [Medline: 35171106]
107. Manning JB, Blandford A, Edbrooke-Childs J. High school teachers' experiences of consumer technologies for stress management during the COVID-19 pandemic: qualitative study. JMIR Form Res 2023 Nov 15;7:e50460 [FREE Full text] [doi: 10.2196/50460] [Medline: 37966873]
108. Clarke V, Braun V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: SAGE Publications; 2013.
109. Duffy TM, Jonassen DH. Constructivism and the Technology of Instruction: A Conversation. Milton Park, UK: Routledge; 1992.
110. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
111. Thrupp M, Lupton R. Taking school contexts more seriously: the social justice challenge. Br J Educ Stud 2006;54(3):308-328. [doi: 10.1111/j.1467-8527.2006.00348.x]
112. Baker SE, Edwards R. How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research. National Centre for Research Methods. 2012. URL: https://eprints.ncrm.ac.uk/id/eprint/2273/4/how_many_interviews.pdf [accessed 2019-01-25]
113. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
114. Barker M. Designing personal informatics technologies for meaningful engagement with experiences during everyday life. The Open University. 2021. URL: https://oro.open.ac.uk/79470/ [accessed 2024-09-23]
115. Carolan S, Harris PR, Greenwood K, Cavanagh K. Increasing engagement with an occupational digital stress management program through the use of an online facilitated discussion group: results of a pilot randomised controlled trial. Internet Interv 2017 Aug 24;10:1-11 [FREE Full text] [doi: 10.1016/j.invent.2017.08.001] [Medline: 30135747]
116. Yang H, Yu J, Zo H, Choi M. User acceptance of wearable devices: an extended perspective of perceived value. Telemat Inform 2016 May;33(2):256-269. [doi: 10.1016/j.tele.2015.08.007]
117. Woodward K, Kanjo E, Brown DJ, McGinnity TM, Inkster B, Macintyre DJ, et al. Beyond mobile apps: a survey of technologies for mental well-being. IEEE Trans Affective Comput 2022;13(3):1216-1235. [doi: 10.1109/taffc.2020.3015018]
118. Hadi R, Valenzuela A. Good vibrations: consumer responses to technology-mediated haptic feedback. J Consum Res 2020 Aug;47(2):256-271. [doi: 10.1093/jcr/ucz039]
119. Park S, Choi J, Lee S, Oh C, Kim C, La S, et al. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. J Med Internet Res 2019 Apr 16;21(4):e12231 [FREE Full text] [doi: 10.2196/12231] [Medline: 30990463]
120. Cansoy R. Teachers' professional development: the case of WhatsApp. J Educ Learn 2017 Aug 08;6(4):285-293. [doi: 10.5539/jel.v6n4p285]
121. Buckingham SA, Morrissey K, Williams AJ, Price L, Harrison J. The physical activity wearables in the police force (PAW-force) study: acceptability and impact. BMC Public Health 2020 Nov 03;20(1):1645 [FREE Full text] [doi: 10.1186/s12889-020-09776-1] [Medline: 33143665]
122. Lupton D, Maslen S. How women use digital technologies for health: qualitative interview and focus group study. J Med Internet Res 2019 Jan 25;21(1):e11481 [FREE Full text] [doi: 10.2196/11481] [Medline: 30681963]

XSL•FO
RenderX

123. Bhattacharjee A, Williams JJ, Meyerhoff J, Kumar H, Mariakakis A, Kornfield R. Investigating the role of context in the delivery of text messages for supporting psychological wellbeing. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3580774]
124. Puussaar A, Clear AK, Wright P. Enhancing personal informatics through social sensemaking. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO. [doi: 10.1145/3025453.3025804]
125. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
126. Goodday SM, Friend S. Unlocking stress and forecasting its consequences with digital technology. NPJ Digit Med 2019 Jul 31;2:75 [FREE Full text] [doi: 10.1038/s41746-019-0151-8] [Medline: 31372508]
127. Kaewkannate K, Kim S. A comparison of wearable fitness devices. BMC Public Health 2016 May 24;16:433 [FREE Full text] [doi: 10.1186/s12889-016-3059-0] [Medline: 27220855]
128. Liang Z, Ploderer B, Chapa-Martell MA. Is fitbit fit for sleep-tracking?: sources of measurement errors and proposed countermeasures. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2017 Presented at: PervasiveHealth '17; May 23-26, 2017; Barcelona, Spain. [doi: 10.1145/3154862.3154897]
129. Attig C, Franke T. Abandonment of personal quantification: a review and empirical study investigating reasons for wearable activity tracking attrition. Comput Hum Behav 2020 Jan;102:223-237. [doi: 10.1016/j.chb.2019.08.025]
130. Christmann CA, Hoffmann A, Bleser G. Stress management apps with regard to emotion-focused coping and behavior change techniques: a content analysis. JMIR Mhealth Uhealth 2017 Feb 23;5(2):e22 [FREE Full text] [doi: 10.2196/mhealth.6471] [Medline: 28232299]
131. Lee K, Hong H. MindNavigator: exploring the stress and self-interventions for mental wellness. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC. [doi: 10.1145/3173574.3174146]
132. Konrad A, Bellotti V, Crenshaw N, Tucker S, Nelson L, Du H, et al. Finding the adaptive sweet spot: balancing compliance and achievement in automated stress reduction. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15; April 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702512]
133. Schwartz B. Self-determination: the tyranny of freedom. Am Psychol 2000;55(1):79-88. [doi: 10.1037/0003-066x.55.1.79]
134. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
135. Sarker H, Sharmin M, Ali AA, Rahman MM, Bari R, Hossain SM, et al. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014 Presented at: UbiComp '14; September 13-17, 2014; Seattle, WA. [doi: 10.1145/2632048.2636082]
136. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
137. Lentferink A, Polstra L, de Groot M, Oldenhuis H, Velthuijsen H, van Gemert-Pijnen L. The values of self-tracking and persuasive eCoaching according to employees and human resource advisors for a workplace stress management application: a qualitative study. In: Proceedings of the 13th International Conference on Persuasive Technology. 2018 Presented at: PERSUASIVE 2018; April 18-19, 2018; Waterloo, ON. [doi: 10.1007/978-3-319-78978-1_13]
138. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
139. Jones NM, Johnson M, Sathappan AV, Torous J. Benefits and limitations of implementing mental health apps among the working population. Psychiatr Ann 2021 Feb;51(2):76-83. [doi: 10.3928/00485713-20210112-01]
140. Stratton E, Jones N, Peters SE, Torous J, Glozier N. Digital mHealth interventions for employees: systematic review and meta-analysis of their effects on workplace outcomes. J Occup Environ Med 2021 Aug 01;63(8):e512-e525. [doi: 10.1097/JOM.0000000000002267] [Medline: 34050095]
141. Shernoff ES, Mehta TG, Atkins MS, Torf R, Spencer J. A qualitative study of the sources and impact of stress among urban teachers. School Ment Health 2011 Mar 19;3(2):59-69. [doi: 10.1007/s12310-011-9051-z]
142. Abel MH, Sewell J. Stress and burnout in rural and urban secondary school teachers. J Educ Res 1999 May;92(5):287-293. [doi: 10.1080/00220679909597608]
143. Camacho DA, Parham B. Urban teacher challenges: what they are and what we can learn from them. Teach Teach Educ 2019 Oct;85:160-174. [doi: 10.1016/j.tate.2019.06.014]

XSL•FO
RenderX

Abbreviations

CBT: cognitive behavioral therapy
GSR: galvanic skin response
MAT: multi-academy trust
PPG: photoplethysmography

Edited by T de Azevedo Cardoso; submitted 01.07.23; peer-reviewed by F Moir, C Pankhurst; comments to author 19.04.24; revised version received 25.07.24; accepted 30.08.24; published 22.10.24.

Please cite as:
Manning JB, Blandford A, Edbrooke-Childs J
Facilitators of and Barriers to Teachers' Engagement With Consumer Technologies for Stress Management: Qualitative Study
J Med Internet Res 2024;26:e50457
URL: https://www.jmir.org/2024/1/e50457
doi:10.2196/50457
PMID:

©Julia B Manning, Ann Blandford, Julian Edbrooke-Childs. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of the Response to an Electronic Versus a Traditional Informed Consent Procedure in Terms of Clinical Patient Characteristics: Observational Study

Anna G M Zondag[1], MSc; Marieke J Hollestelle[2], PhD; Rieke van der Graaf[2], PhD; Hendrik M Nathoe[3], MD, PhD; Wouter W van Solinge[1], PhD; Michiel L Bots[2], MD, PhD; Robin W M Vernooij[2,4], PhD; Saskia Haitjema[1], MD, PhD; UCC-CVRM study group[5]

[1]Central Diagnostic Laboratory, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[3]Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[4]Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht, Netherlands
[5]See Acknowledgements

Corresponding Author:
Anna G M Zondag, MSc
Central Diagnostic Laboratory
University Medical Center Utrecht
Utrecht University
Heidelberglaan 100
Utrecht, 3584 CX
Netherlands
Phone: 31 631117922
Email: a.g.m.zondag@umcutrecht.nl

Abstract

Background: Electronic informed consent (eIC) is increasingly used in clinical research due to several benefits including increased enrollment and improved efficiency. Within a learning health care system, a pilot was conducted with an eIC for linking data from electronic health records with national registries, general practitioners, and other hospitals.

Objective: We evaluated the eIC pilot by comparing the response to the eIC with the former traditional paper-based informed consent (IC). We assessed whether the use of eIC resulted in a different study population by comparing the clinical patient characteristics between the response categories of the eIC and former face-to-face IC procedure.

Methods: All patients with increased cardiovascular risk visiting the University Medical Center Utrecht, the Netherlands, were eligible for the learning health care system. From November 2021 to August 2022, an eIC was piloted at the cardiology outpatient clinic. Prior to the pilot, a traditional face-to-face paper-based IC approach was used. Responses (ie, consent, no consent, or nonresponse) were assessed and compared between the eIC and face-to-face IC cohorts. Clinical characteristics of consenting and nonresponding patients were compared between and within the eIC and the face-to-face cohorts using multivariable regression analyses.

Results: A total of 2254 patients were included in the face-to-face IC cohort and 885 patients in the eIC cohort. Full consent was more often obtained in the eIC than in the face-to-face cohort (415/885, 46.9% vs 876/2254, 38.9%, respectively). Apart from lower mean hemoglobin in the full consent group of the eIC cohort (8.5 vs 8.8; P=.0021), the characteristics of the full consenting patients did not differ between the eIC and face-to-face IC cohorts. In the eIC cohort, only age differed between the full consent and the nonresponse group (median 60 vs 56; P=.0002, respectively), whereas in the face-to-face IC cohort, the full consent group seemed healthier (ie, higher hemoglobin, lower glycated hemoglobin [HbA_{1c}], lower C-reactive protein levels) than the nonresponse group.

Conclusions: More patients provided full consent using an eIC. In addition, the study population remained broadly similar. The face-to-face IC approach seemed to result in a healthier study population (ie, full consenting patients) than the patients without IC, while in the eIC cohort, the characteristics between consent groups were comparable. Thus, an eIC may lead to a better representation of the target population, increasing the generalizability of results.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e54867) doi:10.2196/54867

KEYWORDS

informed consent; learning health care system; e-consent; cardiovascular risk management; digital health; research ethics

Introduction

The use of electronic informed consent (eIC) procedures in clinical research is increasing due to several benefits, including increased enrollment and improved efficiency, by reducing the need for on-site research staff and the associated paperwork [1-3]. eICs have the potential to improve the patient experience (eg, patient understanding and confidence) of the informed consent (IC) process, in part because of the ability to include digital multimedia [4,5]. Alternatively, concerns were raised about whether eIC forms are easily accessible to an elderly population and those with limited digital literacy [6], making it more difficult to assess whether the patient has fully understood the IC form, one of the requirements for a valid IC [7]. Standardized best practices for eIC procedures are still lacking [5,8].

In 2014, the University Medical Center (UMC) Utrecht in the Netherlands initiated the Utrecht Cardiovascular Cohort-CardioVascular Risk Management (UCC-CVRM) as a learning health care system (LHS). The UCC-CVRM LHS aims to improve uniform assessment and registration of cardiovascular risk indicators, based on Dutch national guidelines, in electronic health records (EHRs) for all patients referred to the UMC Utrecht for cardiovascular evaluation [9]. In an LHS, care and research are integrated in such a way that health care activities are continuously analyzed and the knowledge gained from these analyses is used to improve care by changing health care practices [10]. In the case of UCC-CVRM LHS, a traditional face-to-face IC procedure was used for blood sample storage in a biobank and the reuse of routine care data for scientific research purposes including linkage of data from EHRs to national registries [9]. In 2020, during the COVID-19 pandemic, the UCC-CVRM steering committee evaluated the study including the IC procedure [11,12]. In short, less than half of the patients who were invited to participate, 41.5% (2378/5730), provided written IC [12]. Next, patients who did consent differed in clinical characteristics from those nonconsenting or nonresponding, clearly leading to a selection of patients not representable for all eligible patients. For example, consenting patients had a lower cardiovascular disease burden than nonconsenting patients [11,12]. In addition, structured registration of cardiovascular risk management (CVRM) indicators in the EHR was worse compared with consenting patients. This selection is detrimental to an LHS, as the population included in the LHS may be less representative of the target population as a whole [11,12]. Finally, eligible patients were not invited, mainly because of the time-consuming and unsustainable IC procedure due to changes in personnel and changes in priorities during peak periods (eg, the COVID-19 pandemic) [13].

Therefore, the UCC-CVRM steering committee decided to alter the approach. Identification of eligible patients for the LHS was to be automated and CVRM data, to be assessed regularly in patients at higher cardiovascular risk, were extracted from structured fields in the EHR. To still enable the linkage of this patient information to data from national registries, general practitioners (GPs), and other hospitals, an eIC procedure was piloted. This study aims to evaluate the eIC procedure by studying the response to the IC form. In addition, we aim to assess whether the change in the IC procedure leads to a different study population by investigating potential differences in clinical characteristics between the response categories of the eIC compared with the former face-to-face IC procedure.

Methods

We used the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement as a reporting guideline for this study.

Study Setting

Patients visiting the UMC Utrecht for the first time for the evaluation of cardiovascular disease or risk factors were eligible for inclusion in the UCC-CVRM LHS. The full rationale of UCC-CVRM has been described elsewhere [9]. The eIC pilot was conducted between November 2021 and August 2022. During this period, all patients (18 years and older) visiting the cardiology outpatient clinic for first-time evaluation automatically received an email. This email notified these patients about UCC-CVRM and the associated eIC form that was available for completion in the UMC Utrecht patient portal. The IC procedures of the face-to-face IC and eIC are illustrated in Figure 1. The full details eIC form presented to the patients in the patient portal are included in Multimedia Appendix 1.

In both the face-to-face IC and the eIC forms, IC was asked for linkage with national registries, GPs, and other hospitals through the following two statements: (1) I consent to future requests to link with various international or national registries, such as the Central Bureau of Statistics (also known as Statistics Netherlands) where all causes of death are registered, the Dutch Cancer Registration (NKR), where all people with cancer are registered, the National Basic Registration of Hospital Care (LBZ), where all hospital admissions are registered, the Foundation for Pharmaceutical Statistics (SFK), where all people who use medicines are registered, and other regional and national registries. (2) I consent to the retrieval of my medical information from my GP, my pharmacy, and any other hospitals where I have been treated in the past.

XSL•FO
RenderX

Figure 1. Informed consent procedure of the face-to-face UCC-CVRM (left) and the eIC pilot (right). CVD: cardiovascular disease; EHR: electronic health record; eIC: electronic informed consent; UCC-CVRM: Utrecht Cardiovascular Cohort-CardioVascular Risk Management; UMC: University Medical Center.

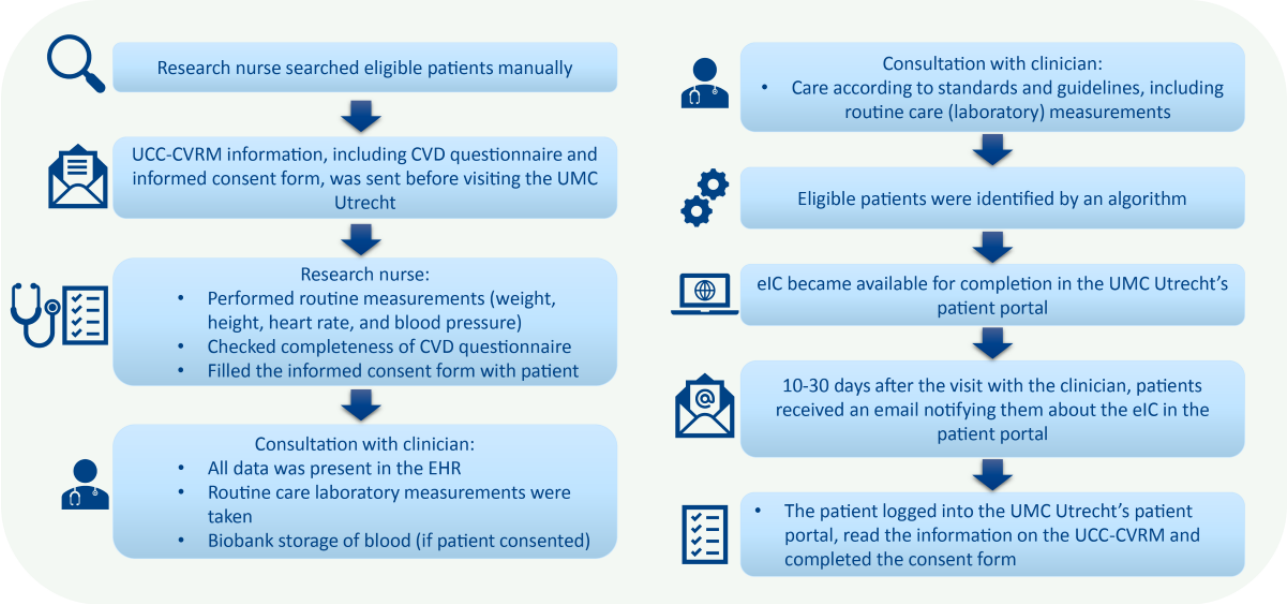

Data Collection

We collected data from all patients aged 18 years or older referred to a cardiology outpatient clinic. We used the Utrecht Patient Oriented Database to collect data from the patients who participated in the eIC pilot, referred to as the "eIC cohort." The Utrecht Patient Oriented Database comprises data on, among others, patient characteristics and laboratory tests for all patients treated at the UMC Utrecht since 2004 [14]. We collected routine care data related to the patient's demographics and cardiovascular risk, namely blood pressure, BMI, and laboratory measurements (serum lipids, glycated hemoglobin [HbA_{1c}], hemoglobin, and renal function). These data were also collected for the cardiology patients who were invited during the period in which a face-to-face IC procedure was in place, referred to as the "face-to-face IC cohort." From the face-to-face IC cohort, only patients invited up until December 31, 2019, were included because the COVID-19 pandemic significantly hampered the face-to-face IC procedure.

Variables

All measurements were extracted from structured fields in the EHR. Blood pressure values were extracted from the EHR ±7 days from the date of the visit at the cardiology outpatient clinic. For other measurements, the closest value, within ±21 days of the visit date, was extracted. If no measurements were found within these cutoffs, the measurement was considered missing. An overview of missingness per variable is added as Multimedia Appendix 2. Age was calculated by subtracting the date of the visit from the patient's date of birth. The estimated glomerular filtration rate was calculated using the Chronic Kidney Disease Epidemiology Collaboration equation and used as a measure of renal function [15].

Data Analyses

We presented the yield for both the eIC and face-to-face IC cohort as counts and the percentages of patients who, (1) consented to the linkage of their data with their GP, pharmacy, and other hospitals, and linkage with national registries (ie, "full consent"); (2) did not consent for the linkage of their data with their GP, pharmacy, and other hospitals, nor to the linkage with national registries (ie, "nonconsent"); (3) consented to only 1 of the 2; and (4) the percentage of patients who did not respond at all or provided an answer for only 1 of the 2 statements (ie, "nonresponse"). Due to the limited number of observations in some IC response groups (n<25), especially in the eIC cohort, other than the full consent (n=1291) and nonresponse (n=1477) group, further analyses were restricted to the full consent and nonresponse group only.

To assess differences in patient groups, characteristics of the patients with full consent were stratified by cohort (ie, eIC cohort vs face-to-face IC cohort). As a supplement, we also explored the characteristics of the nonresponders by cohort. Finally, we assessed the differences in patient characteristics between the response categories within each cohort (ie, full consent versus nonresponse).

Clinical characteristics were presented as means with corresponding SDs, medians with corresponding interquartile ranges, or counts and percentages, as appropriate. To quantify differences in characteristics between cohorts, or response categories within cohorts, we performed multivariable linear regression analyses, adjusted for age, categorized into 4 categories with an approximately equal number of observations (18 to 47, 48 to 60, 61 to 70, and 71 to 95 years old), and sex. The assumptions of linear regression (eg, approximate normal distribution of the error terms, homoscedasticity of errors) were assessed. Where needed, we used the Box-Cox method to estimate the most appropriate transformation of the dependent variable to stabilize the variance and improve the accuracy of our estimations [16]. Similarly, multivariable linear regression was used to assess the difference in age (as a continuous variable) between groups, adjusted for sex. Multivariable logistic regression was used to assess the difference in sex between the

XSL•FO
RenderX

groups (ie, between the 2 cohorts and between the response categories within each cohort), adjusted for age (categorized into 4 categories with an approximately equal number of observations). As a sensitivity analysis, we repeated the adjusted regression analyses with age as a continuous variable, to assess whether the categorization led to different results.

We used the Bonferroni correction to reduce the risk of a type I error resulting from the multiple tests [17]. Thus, the α that we considered as cutoff, .05, was divided by the number of analyses (N=13) performed per comparison. Therefore, a *P* value ≤.0038 was considered statistically significant.

All statistical analyses were performed using R software (version 4.0.5; The R Foundation) [18].

Ethical Considerations

We obtained an additional ethical waiver (number 19/641) from the Research Ethics Committee Utrecht to examine the characteristics of patients in all IC response categories. Patients who objected to the use of their clinical data for research purposes via the UMC Utrecht opt-out procedure were excluded from this study. Data were pseudonymized and the patients did not receive any compensation for their participation in this study.

Results

Yield of the IC Procedure

In total, 3139 patients participated in this study, of whom 885 (28.2%) participated in the eIC pilot cohort and 2254 (71.8%) in the face-to-face IC cohort (Multimedia Appendix 3). Of all patients from the eIC cohort, 49.9% (442/885) completed the eIC form, 50.1% (443/885) did not respond. Of all patients who completed the eIC form, we obtained full consent for linkage with GPs, hospitals, and national registries from 93.9% (415/442) of the patients. In the face-to-face IC cohort, 54.1% (1220/2254) of all patients completed the IC form, and 45.9% (1034/2254) patients did not respond. The percentage of responding patients with full consent was higher in the eIC cohort as compared with the face-to-face IC cohort (415/442, 93.9% vs 876/1220, 71.8%, respectively).

Differences in Characteristics Between the eIC and Face-to-Face Cohorts

Overall, fully consenting patients had similar patient characteristics (Table 1). Adjusted for sex and age, the eIC cohort had lower hemoglobin levels and higher HbA_{1c} levels than the face-to-face IC cohort, reaching the multiple testing threshold for statistical significance for hemoglobin (*P*=.0021).

Similarly, we compared the (clinical) characteristics of the nonresponding patients between cohorts, shown in Multimedia Appendix 4. Adjusted for sex, the nonresponders of the eIC cohort were significantly younger and had, adjusted for age and sex, lower c-reactive protein values than the nonresponders of the face-to-face cohort. No other differences were observed.

XSL•FO
RenderX

Table 1. Differences between patients with full consent by cohort, adjusted for age and sex.

Variable	Full consent		*P* value
	eIC[a] (n=415)	F2F IC[b] (n=876)	
Age (years), median (IQR)	60.0 (48.0-70.0)	61.0 (50.0-69.0)	.2529
Sex, n (%)			
Male	237 (57.1)	476 (54.3)	__[c]
Female	178 (42.9)	400 (45.7)	.3239
BMI (kg/m^2), mean (SD)	26.6 (5.2)	26.7 (5.7)	.8981
SAP[d] (mm Hg), mean (SD)	132.1 (19.4)	137.6 (19.6)	.0586
Hemoglobin (mmol/L), mean (SD)	8.5 (1.4)	8.8 (0.9)	.0021
HbA_{1c}[e] (mmol/mol), median (IQR)	37.5 (34.0-44.0)	37.0 (34.0-40.0)	.0454
Cholesterol (mmol/L), mean (SD)	4.8 (1.2)	5.1 (1.3)	.1266
HDL[f]-cholesterol (mmol/L), mean (SD)	1.3 (0.4)	1.4 (0.4)	.0676
LDL[g]-cholesterol (mmol/L), mean (SD)	2.7 (1.1)	2.9 (1.1)	.1086
Triglycerides (mmol/L), median (IQR)	1.7 (1.1-2.6)	1.6 (1.0-2.1)	.3023
CRP[h] (mg/L), median (IQR)	2.0 (0.5-10.0)	2.6 (1.1-8.5)	.6666
Creatinine (µmol/L), median (IQR)	76.0 (64.2-94.0)	74.0 (64.0-88.0)	.7760
eGFR CKD-EPI[i] (mL/min/1.73 m^2), mean (SD)	83.3 (23.1)	84.5 (22.3)	.7068

[a]eIC: electronic informed consent.
[b]F2F IC: face-to-face informed consent.
[c]Reference group.
[d]SAP: systolic arterial blood pressure.
[e]HbA_{1c}: glycated hemoglobin.
[f]HDL: high-density lipoprotein.
[g]LDL: low-density lipoprotein.
[h]CRP: c-reactive protein.
[i]eGFR CKD-EPI: estimated glomerular filtration rate calculated using the Chronic Kidney Disease Epidemiology Collaboration equation.

Differences in Characteristics Between Response Categories

Within each cohort, we assessed whether there were differences in characteristics between the response categories (ie, full consent vs nonresponse). In the eIC cohort, the nonresponse group was significantly younger than the full consent group (Table 2). Other than that, the clinical characteristics of the full consent group were similar to those of the nonresponse group.

More differences were found between the response categories of the face-to-face IC cohort. Adjusted for age and sex, patients in the full consent group had higher hemoglobin, but lower HbA_{1c} and c-reactive protein values than the nonresponse group (Table 3).

XSL•FO
RenderX

Table 2. Differences between the response categories in the electronic informed consent cohort, adjusted for age and sex.

Variable	Full consent (n=415)	Nonresponse (n=443)	*P* value
Age (years), median (IQR)	60.0 (48.0-70.0)	56.0 (28.0-72.0)	.0002
Sex, n (%)			
Male	237 (57.1)	222 (50.1)	__[a]
Female	178 (42.9)	221 (49.9)	.0420
BMI (kg/m^2), mean (SD)	26.6 (5.2)	26.0 (4.9)	.3673
SAP[b] (mm Hg), mean (SD)	132.1 (19.4)	130.4 (19.6)	.4168
Hemoglobin (mmol/L), mean (SD)	8.5 (1.4)	8.4 (1.3)	.2397
HbA_{1c}[c] (mmol/mol), median (IQR)	37.5 (34.0-44.0)	37.5 (34.0-40.2)	.1940
Cholesterol (mmol/L), mean (SD)	4.8 (1.2)	4.6 (1.5)	.2852
HDL[d]-cholesterol (mmol/L), mean (SD)	1.3 (0.4)	1.2 (0.5)	.3371
LDL[e]-cholesterol (mmol/L), mean (SD)	2.7 (1.1)	2.6 (0.9)	.9304
Triglycerides (mmol/L), median (IQR)	1.7 (1.1-2.6)	1.4 (1.1-2.0)	.4167
CRP[f] (mg/L), median (IQR)	2.0 (0.5-10.0)	3.0 (0.5-12.0)	.5922
Creatinine (µmol/L), median (IQR)	76.0 (64.2-94.0)	79.0 (64.0-100.5)	.0897
eGFR CKD-EPI[g] (mL/min/1.73 m^2), mean (SD)	83.3 (23.1)	82.0 (30.7)	.1103

[a]Reference group.
[b]SAP: systolic arterial blood pressure.
[c]HbA_{1c}: glycated hemoglobin.
[d]HDL: high-density lipoprotein.
[e]LDL: low-density lipoprotein.
[f]CRP: c-reactive protein.
[g]eGFR CKD-EPI: estimated glomerular filtration rate calculated using the Chronic Kidney Disease Epidemiology Collaboration equation.

XSL•FO
RenderX

Table 3. Differences between the response categories of the face-to-face informed consent cohort, adjusted for age and sex.

Variable	Full consent (n=876)	Nonresponse (n=1034)	*P* value
Age, median (IQR)	61.0 (50.0-69.0)	61.0 (48.0-71.0)	.9461
Sex, n (%)			
Male	476 (54.3)	552 (53.4)	__[a]
Female	400 (45.7)	482 (46.6)	.7859
BMI (kg/m^2), mean (SD)	26.7 (5.7)	26.2 (5.5)	.1063
SAP[b] (mm Hg), mean (SD)	137.6 (19.6)	136.3 (22.0)	.1093
Hemoglobin (mmol/L), mean (SD)	8.8 (0.9)	8.3 (1.2)	<.0001
HbA_{1c}[c] (mmol/mol), median (IQR)	37.0 (34.0-40.0)	38.0 (34.0-42.0)	.0001
Cholesterol (mmol/L), mean (SD)	5.1 (1.3)	5.0 (1.4)	.4493
HDL[d]-cholesterol (mmol/L), mean (SD)	1.4 (0.4)	1.3 (0.4)	.0898
LDL[e]-cholesterol (mmol/L), mean (SD)	2.9 (1.1)	2.9 (1.1)	.2754
Triglycerides (mmol/L), median (IQR)	1.6 (1.0-2.1)	1.6 (1.0-2.4)	.0435
CRP[f] (mg/L), median (IQR)	2.6 (1.1-8.5)	8.1 (2.0-38.2)	<.0001
Creatinine (µmol/L), median (IQR)	74.0 (64.0-88.0)	75.0 (63.0-92.0)	.4361
eGFR CKD-EPI[g] (mL/min/1.73 m^2), mean (SD)	84.5 (22.3)	81.3 (29.0)	.0946

[a]Reference group.

[b]SAP: systolic arterial blood pressure.

[c]HbA_{1c}: glycated hemoglobin.

[d]HDL: high-density lipoprotein.

[e]LDL: low-density lipoprotein.

[f]CRP: c-reactive protein.

[g]eGFR CKD-EPI: estimated glomerular filtration rate calculated using the Chronic Kidney Disease Epidemiology Collaboration equation.

Sensitivity Analysis

We repeated the regression analyses adjusted for sex and age. In these regression analyses, age was maintained continuous instead of categorized, to assess whether the categorization of age led to different results. The results were similar (Multimedia Appendix 5).

Discussion

Principal Results

We showed that by using an eIC in an LHS, patients more often provided full consent to link their data to national registries, GPs, and other hospitals compared with a face-to-face IC procedure. The clinical characteristics of patients with full consent remained largely similar after changing the IC procedure to an eIC. Except for age, we did not find any differences between the response categories of the eIC cohort, whereas in the face-to-face cohort, several differences were found. These differences potentially suggest a higher (cardiovascular) disease burden in the nonresponse group compared with the full consent group, indicative of a potentially more pronounced selection in the face-to-face approach.

A possible explanation for the differences in characteristics between the response categories in the face-to-face cohort is that patients may have been too ill or frail to attend the physical appointment with the research nurse to discuss and sign the IC form, resulting in nonresponse. The inability to attend the appointment was probably less of an issue in the eIC cohort, as patients were able to access the eIC form remotely. The finding suggests that the use of eIC results in a study population (ie, those who give full consent) that is more representative of the full target population. Our findings agree with a previous study showing that providing computer-based clinical study information leads to more willingness to participate [19], as the increased willingness to participate is consistent with the higher full consent rates found in the eIC group compared with the face-to-face IC group in our study.

Frequently Mentioned Challenges of eICs in the Literature

Concerns have been raised about whether consent given via an eIC is truly an “informed” consent [8]. According to the principles of the Declaration of Helsinki [7], potential participants must be adequately informed about various aspects of the study, such as its purpose, sources of funding, the anticipated benefits and potential risks, and the right to refuse or withdraw consent to participate without giving a reason [7]. According to previous research, comprehension assessment is more challenging when an eIC procedure is used as there is no direct interaction between the potential participant and

XSL•FO
RenderX

researcher [6]. As a result, patients might provide consent without fully understanding what they are consenting to, or, conversely, patients may be less likely to consent because of the lack of personal interaction with the researcher or clinician, especially those who were already doubtful about participating in the first place. However, our findings indicate that the latter might not have been the case in our pilot study, as we observed a higher percentage of patients with full consent in the eIC cohort compared with the face-to-face IC cohort.

Another frequently mentioned concern is that studies using an eIC procedure could become inaccessible to patients who lack the digital literacy needed to access and understand the eIC form [6]. In 2021, the Netherlands had the highest percentage (ie, 79%) of 17- to 74-year-olds with at least basic digital skills in Europe [20]. Therefore, incomprehension of the eIC due to limited digital literacy may appear less of an issue in our study. However, the percentage of persons with basic digital skills varied considerably by age, with older people being less literate [20]. A sensitivity analysis showed that the age distribution of responding patients was similar between the eIC and the face-to-face approach (Multimedia Appendix 6), indicating that the eIC was not less accessible than the face-to-face IC for certain age groups. However, accessibility may be an issue for geriatric patients, who are generally older than cardiology patients and often have geriatric syndromes that sometimes affect comprehension and literacy [21]. These syndromes generally make it difficult to obtain IC from the elderly [21]. eIC could, therefore, also be seen as an opportunity. Unlike paper-based ICs, multiple formats can be used to inform the patient about the purpose of the eIC and to provide technical support, for example, by using instructional videos or audio. The use of multiple formats in IC forms for the elderly has been recommended by, among others, Barron et al [22]. Furthermore, UCC-CVRM's eIC form is available in UMC Utrecht's long-existing patient portal. In the portal, patients have the opportunity to, among others, ask questions to their clinician via an e-consult, which can be used if parts of the eIC are unclear [23]. Another possibility would be a hybrid format, allowing patients who prefer correspondence by regular mail to respond using a paper-based IC form. However, it is questionable whether this would be helpful and it would negate the positive aspects of the eIC highlighted in this study (eg, less pronounced selection).

Legislation and Regulation Regarding eIC

Since July 2022, eICs have been permitted in the Netherlands when certain conditions are met [24]. A total of 6 conditions are described in the guideline written by the Central Committee on Research Involving Human Subjects (Centrale Commissie Mensgebonden Onderzoek) and the Dutch Association of Medical Research Ethics Committees (Nederlandse Vereniging voor Medisch-Ethische toetsingscommissies) [25]. The most important conditions are (1) eIC must be appropriate for the study, meaning that the study is associated with low potential risk and burden for the patient, (2) the eIC process must be sufficiently reliable and confidential, guaranteed by an electronic system that is compliant to the Dutch General Data Protection Regulation (UAVG in Dutch) and ensures the validity of the electronic signatures, and (3) the eIC procedure must be described in the study protocol [24,25]. The implementation of an eIC seems appropriate in the case of the UCC-CVRM, as no potential risk or burden for the patient is involved. Furthermore, in the eIC of the UCC-CVRM, data security, identity verification, and the validity of the electronic signature are ensured by the Dutch digital ID, an identification method for accessing web-based services [26]. Regarding the third condition, an amendment to the UCC-CVRM approach, including the eIC, was submitted and approved by the Research Ethics Committee.

Clinical Implications

Based on the results of our study, the use of eIC to obtain IC might be a sustainable and adequate way to enable researchers to link with national registries, GPs, and other hospitals. The use of the eIC seemed to have resulted in a population with consent that is more similar to the target population compared with the face-to-face IC, which is of great importance in an LHS. Results from the LHS would be more generalizable to the target population, namely to all patients at higher cardiovascular risk. Yet, one may argue whether ≈50% response to both the electronic and face-to-face IC for an LHS approach is sufficient. In addition, it should be noted that the extractability of CVRM indicators from structured fields in the EHR was much lower in the eIC cohort compared with the face-to-face IC cohort. Groenhof et al [13] showed that the former, protocolized, face-to-face UCC-CVRM approach led to more systematic registration of the cardiovascular risk profile in the EHR, which had a positive effect on CVRM guideline adherence in consenting patients, compared with the situation before UCC-CVRM was introduced [13]. The substantial missingness in the eIC cohort of our study may suggest that these improvements are at risk when the approach is automated, as deviations from the initial protocol are made, potentially leading to suboptimal CVRM in clinical care.

Exploring the views and experiences of patients could help to further improve the eIC form. Therefore, we recommend further qualitative research into the accessibility and understandability of eICs used for similar purposes and in similar settings as the UCC-CVRM LHS from a patient's perspective.

Strengths and Limitations

To the best of our knowledge, we are among the first to investigate the differences in clinical patient characteristics between response categories of an eIC compared with those of a traditional face-to-face IC, specifically in the context of a cardiovascular LHS in a large sample of patients. Our uniqueness, however, limits the ability to compare our findings to the literature, as most research on eIC has focused on user perspectives, experiences, and the ethical considerations of eICs. For example, Chen et al [5] showed that in most included studies, participants had a better understanding of the information when using an eIC compared with a traditional paper-based face-to-face IC, while others found no difference [5]. Nevertheless, they [5] and others [2,6,27] indicated that face-to-face interaction should remain part of the IC process, especially for more complex and higher-risk studies. However, as the UCC-CVRM LHS is not a complex or high-risk study, the face-to-face interaction may be less necessary. Furthermore,

XSL•FO
RenderX

the nonresponders in the eIC cohort may not be fully comparable to the nonresponders in the face-to-face IC cohort because, in the eIC cohort, patients received the eIC after their appointment at the cardiology outpatient clinic, whereas in the face-to-face IC cohort, cardiology patients were identified as eligible and received information about the UCC-CVRM LHS prior to their appointment. This means that patients who, for example, canceled their appointment at the last minute would still be included in the face-to-face cohort as nonresponders. It may be that patients who did not attend their appointment at all had different characteristics to those who attended but did not respond to the eIC, potentially affecting the validity of the comparisons made. Finally, the eIC form was piloted in the patient population of the cardiology outpatient clinic only. Although our results indicated that there were only minor differences (ie, hemoglobin) between patients providing full consent using the eIC compared with the face-to-face IC, it remains to be seen whether this would still be the case after implementation of the eIC in other clinical departments.

Conclusions

To conclude, our findings suggest that using an eIC may lead to a better representation of the target population by consenting patients. This increases the generalizability of results from studies using the data collected within the LHS from consenting patients.

Acknowledgments

The Utrecht Cardiovascular Cohort-Cardiovascular Risk Management (UCC-CVRM) is primarily financed by the University Medical Center (UMC) Utrecht (contact information of UCC-CVRM is ucc@umcutrecht.nl). AGMZ was supported by a grant from the European Union's Horizon 2020 research and innovation program (grant agreement number 101017331; ODIN). MJH and RvdG were supported by the ZonMw, ETHMIRE project (grant agreement number 91217027). The funding sources were not involved in the design of the study, the analysis and interpretation of the data, the writing of the manuscript, and the decision to submit the manuscript for publication. Members of the UCC-CVRM study group were the following: GJ de Borst, Department of Vascular Surgery; ML Bots (chair), Julius Center for Health Sciences and Primary Care; M Hollander, Julius Center for Health Sciences and Primary Care; MH Emmelot, Department of Geriatrics; PA de Jong, Department of Radiology; AT Lely, Department of Obstetrics/Gynecology; HM Nathoe, Department of Cardiology; IE Hoefer, Central Diagnostic Laboratory; NP van der Kaaij, Department of Cardiothoracic Surgery; YM Ruigrok, Department of Neurology; and MC Verhaar, Department of Nephrology and Hypertension, FLJ Visseren, Department of Vascular Medicine, University Medical Center Utrecht and Utrecht University.

Authors' Contributions

AGMZ, HMN, WWvS, MLB, SH, and RWMV contributed to the conceptualization of the project. AGMZ, RWMV, SH, WWvS, and MLB contributed to the methodology of the project. AGMZ analyzed the data and drafted the manuscript. AGMZ, MJH, RvdG, HMN, WWvS, MLB, SH, and RWMV contributed substantially to the interpretation of the data. The final manuscript was critically reviewed and edited by all authors. Approval of the final manuscript was obtained by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The electronic informed consent form as presented in the patient portal of the UMC (University Medical Center) Utrecht (translated from Dutch to English).
[DOC File , 25 KB - jmir_v26i1e54867_app1.doc]

Multimedia Appendix 2
Missingness per variable in count and percentage, by cohort and informed consent response strata.
[DOC File , 42 KB - jmir_v26i1e54867_app2.doc]

Multimedia Appendix 3
Yield (ie, response to the informed consent invitation), by type of informed consent. eIC: electronic informed consent; GP: general practitioner.
[DOC File , 26 KB - jmir_v26i1e54867_app3.doc]

Multimedia Appendix 4
Differences between patients who did not respond, by cohort, adjusted for age and sex.
[DOC File , 46 KB - jmir_v26i1e54867_app4.doc]

Multimedia Appendix 5

XSL•FO
RenderX

Results of the sensitivity analysis in which age is treated as a continuous variable instead of categorical variable.
[DOC File , 118 KB - jmir_v26i1e54867_app5.doc]

Multimedia Appendix 6
Age distribution of patients who completed the informed consent form, stratified by cohort.
[DOC File , 109 KB - jmir_v26i1e54867_app6.doc]

References

1. Simon CM, Klein DW, Schartz HA. Traditional and electronic informed consent for biobanking: a survey of U.S. biobanks. Biopreserv Biobank 2014;12(6):423-429. [doi: 10.1089/bio.2014.0045] [Medline: 25496155]
2. Skelton E, Drey N, Rutherford M, Ayers S, Malamateniou C. Electronic consenting for conducting research remotely: a review of current practice and key recommendations for using e-consenting. Int J Med Inform 2020;143:104271 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104271] [Medline: 32979650]
3. Phillippi JC, Doersam JK, Neal JL, Roumie CL. Electronic informed consent to facilitate recruitment of pregnant women into research. J Obstet Gynecol Neonatal Nurs 2018;47(4):529-534. [doi: 10.1016/j.jogn.2018.04.134] [Medline: 29758172]
4. Boutin NT, Mathieu K, Hoffnagle AG, Allen NL, Castro VM, Morash M, et al. Implementation of electronic consent at a biobank: an opportunity for precision medicine research. J Pers Med 2016;6(2):17 [FREE Full text] [doi: 10.3390/jpm6020017] [Medline: 27294961]
5. Chen C, Lee P, Pain KJ, Delgado D, Cole CL, Campion TR. Replacing paper informed consent with electronic informed consent for research in academic medical centers: a scoping review. AMIA Jt Summits Transl Sci Proc 2020;2020:80-88 [FREE Full text] [Medline: 32477626]
6. Yusof MYPM, Teo CH, Ng CJ. Electronic informed consent criteria for research ethics review: a scoping review. BMC Med Ethics 2022;23(1):117 [FREE Full text] [doi: 10.1186/s12910-022-00849-x] [Medline: 36414962]
7. Declaration of helsinki - ethical principles for medical research involving human subjects. World Medical Association. 2022. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [accessed 2023-05-05]
8. Lunt H, Connor S, Skinner H, Brogden G. Electronic informed consent: the need to redesign the consent process for the digital age. Intern Med J 2019;49(7):923-929. [doi: 10.1111/imj.14339] [Medline: 31295775]
9. Asselbergs FW, Visseren FL, Bots ML, de Borst GJ, Buijsrogge MP, Dieleman JM, et al. Uniform data collection in routine clinical practice in cardiovascular patients for optimal care, quality control and research: the Utrecht cardiovascular cohort. Eur J Prev Cardiol 2017;24(8):840-847. [doi: 10.1177/2047487317690284] [Medline: 28128643]
10. Wouters RHP, van der Graaf R, Voest EE, Bredenoord AL. Learning health care systems: highly needed but challenging. Learn Health Syst 2020;4(3):e10211 [FREE Full text] [doi: 10.1002/lrh2.10211] [Medline: 32685681]
11. Groenhof TKJ, Mostert M, Lea NC, Haitjema S, de Vries MC, van Dijk WB, et al. How traditional informed consent impairs inclusivity in a learning healthcare system: lessons learned from the Utrecht cardiovascular cohort. J Clin Epidemiol 2022;149:190-194. [doi: 10.1016/j.jclinepi.2022.04.025] [Medline: 35504525]
12. Zondag AGM, Groenhof TKJ, van der Graaf R, van Solinge WW, Bots ML, Haitjema S, UCC-CVRM study group. Asking informed consent may lead to significant participation bias and suboptimal cardiovascular risk management in learning healthcare systems. BMC Med Res Methodol 2023;23(1):98 [FREE Full text] [doi: 10.1186/s12874-023-01924-6] [Medline: 37087415]
13. Groenhof TKJ, Haitjema S, Lely AT, Grobbee DE, Asselbergs FW, Bots ML, UCC-CVRMUPOD Study groups. Optimizing cardiovascular risk assessment and registration in a developing cardiovascular learning health care system: women benefit most. PLOS Digit Health 2023;2(2):e0000190 [FREE Full text] [doi: 10.1371/journal.pdig.0000190] [Medline: 36812613]
14. ten Berg MJ, Huisman A, van den Bemt PMLA, Schobben AFAM, Egberts ACG, van Solinge WW. Linking laboratory and medication data: new opportunities for pharmacoepidemiological research. Clin Chem Lab Med 2007;45(1):13-19. [doi: 10.1515/CCLM.2007.009] [Medline: 17243908]
15. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
16. Glen S. Box cox transformation: definition, examples. Statistics How To. URL: https://www.statisticshowto.com/probability-and-statistics/normal-distributions/box-cox-transformation/ [accessed 2023-08-09]
17. VanderWeele TJ, Mathur MB. Some desirable properties of the bonferroni correction: is the bonferroni correction really so bad? Am J Epidemiol 2019;188(3):617-618 [FREE Full text] [doi: 10.1093/aje/kwy250] [Medline: 30452538]
18. R: a language and environment for statistical computing. R foundation for statistical computing. 2021. URL: https://www.R-project.org/ [accessed 2022-08-31]
19. Karunaratne AS, Korenman SG, Thomas SL, Myles PS, Komesaroff PA. Improving communication when seeking informed consent: a randomised controlled study of a computer-based method for providing information to prospective clinical trial participants. Med J Aust 2010 Apr;192(7):388-392. [doi: 10.5694/j.1326-5377.2010.tb03561.x] [Medline: 20367586]

XSL•FO
RenderX

20. Dutch digital skills at the top in Europe. CBS. 2022. URL: https://www.cbs.nl/en-gb/news/2022/19/dutch-digital-skills-at-the-top-in-europe [accessed 2023-04-20]
21. Cherubini A, Gasperini B. How to increase the participation of older subjects in research: good practices and more evidence are needed!. Age Ageing 2017;46(6):878-881. [doi: 10.1093/ageing/afx123] [Medline: 29088365]
22. Barron JS, Duffey PL, Byrd LJ, Campbell R, Ferrucci L. Informed consent for research participation in frail older persons. Aging Clin Exp Res 2004;16(1):79-85. [doi: 10.1007/BF03324536] [Medline: 15132296]
23. My UMC Utrecht patient portal. UMC Utrecht. URL: https://www.umcutrecht.nl/nl/inloggen-patientenportaal?lang=en [accessed 2023-08-10]
24. Consent: on paper or electronic (in Dutch only). CCMO. URL: https://tinyurl.com/bdzach6u [accessed 2023-04-11]
25. Guide to granting electronic consent for participation in medical-scientific research (in Dutch only). CCMO, NVMETC. 2022. URL: https://www.ccmo.nl/metcs/publicaties/publicaties/2022/08/31/handreiking-elektronische-toestemmingsverlening [accessed 2023-04-11]
26. What is DigiD? Netherlands Worldwide. 2024. URL: https://www.netherlandsworldwide.nl/digid-abroad/what-is-digid [accessed 2023-07-26]
27. De Sutter E, Zaçe D, Boccia S, Di Pietro ML, Geerts D, Borry P, et al. Implementation of electronic informed consent in biomedical research and stakeholders' perspectives: systematic review. J Med Internet Res 2020;22(10):e19129 [FREE Full text] [doi: 10.2196/19129] [Medline: 33030440]

Abbreviations

CVRM: cardiovascular risk management
EHR: electronic health record
eIC: electronic informed consent
GP: general practitioner
HbA1c: glycated hemoglobin
IC: informed consent
LHS: learning health care system
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UCC-CVRM: Utrecht Cardiovascular Cohort-Cardiovascular Risk Management
UMC: University Medical Center

Edited by A Mavragani; submitted 29.11.23; peer-reviewed by CMJ Wong, H Kondylakis; comments to author 28.02.24; revised version received 15.04.24; accepted 10.05.24; published 11.07.24.

Please cite as:
Zondag AGM, Hollestelle MJ, van der Graaf R, Nathoe HM, van Solinge WW, Bots ML, Vernooij RWM, Haitjema S, UCC-CVRM study group
Comparison of the Response to an Electronic Versus a Traditional Informed Consent Procedure in Terms of Clinical Patient Characteristics: Observational Study
J Med Internet Res 2024;26:e54867
URL: https://www.jmir.org/2024/1/e54867
doi:10.2196/54867
PMID:38990640

©Anna G M Zondag, Marieke J Hollestelle, Rieke van der Graaf, Hendrik M Nathoe, Wouter W van Solinge, Michiel L Bots, Robin W M Vernooij, Saskia Haitjema, UCC-CVRM study group. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Short-Form Videos to Get Clinical Trial Newcomers to Sign Up: Message-Testing Experiment

Sisi Hu[1*], PhD; Ciera E Kirkpatrick[2*], PhD; Namyeon Lee[3*], PhD; Yoorim Hong[4*], MA; Sungkyoung Lee[4*], PhD; Amanda Hinnant[4*], PhD

[1]School of Journalism and Strategic Media, University of Arkansas, Fayetteville, AR, United States

[2]College of Journalism & Mass Communications, University of Nebraska-Lincoln, Lincoln, NE, United States

[3]Department of Mass Communication, University of North Carolina at Pembroke, Pembroke, NC, United States

[4]School of Journalism, University of Missouri, Columbia, MO, United States

[*]all authors contributed equally

Corresponding Author:
Sisi Hu, PhD
School of Journalism and Strategic Media
University of Arkansas
129 Kimpel Hall, 280 N. McIlroy Avenue
Fayetteville, AR, 72701
United States
Phone: 1 479 575 6717
Email: sisih@uark.edu

Abstract

Background: Recruiting participants for clinical trials poses challenges. Major barriers to participation include psychological factors (eg, fear and mistrust) and logistical constraints (eg, transportation, cost, and scheduling). The strategic design of clinical trial messaging can help overcome these barriers. While strategic communication can be done through various channels (eg, recruitment advertisements), health care providers on the internet have been found to be key sources for communicating clinical trial information to US adults in the social media era.

Objective: This study aims to examine how communication source (ie, medical doctors and peers) and message framing of TikTok videos (ie, psychological and logistical framing) influence clinical trial–related attitudes, perceptions, and sign-up behaviors under the guidance of the integrated behavioral model.

Methods: This study used a 2 (source: doctor vs peer) × 2 (framing: psychological vs logistical) between-participant factorial design web-based experiment targeting adults in the United States who had never participated in clinical trials (ie, newcomers). A Qualtrics panel was used to recruit and compensate the study respondents (n=561). Participants viewed short-form videos with doctors or peers, using psychological or logistical framing. The main outcome measures included perceived source credibility, self-efficacy, attitude toward clinical trial participation, behavioral intention, and sign-up behavior. Structural equation modeling was used to analyze the direct and indirect effects of message factors on the outcome variables. Source (doctor=1; peer=0) and framing (psychological=1; logistical=0) were dummy-coded.

Results: Doctor-featured messages led to greater perceived source credibility (β=.31, *P*<.001), leading to greater self-efficacy (95% CI 0.13-0.30), which in turn enhanced behavioral intention (95% CI 0.12-0.29) and clinical trial sign-up behavior (95% CI 0.02-0.04). Logistical barrier–framed messages led to greater self-efficacy (β=–.09, *P*=.02), resulting in higher intention to participate in clinical trials (95% CI –0.38 to –0.03) and improved sign-up behavior (95% CI –0.06 to –0.004). Logistical barrier–framed messages were also directly associated with an increased likelihood of signing up for a clinical trial (β=–.08, *P*=.03). The model accounted for 21% of the variance in clinical trial sign-up behavior. Attitude did not significantly affect behavioral intention in this study (β=.08, *P*=.14), and psychological and logistical barrier–framed messages did not significantly differ in attitudes toward clinical trial participation (β=–.04, *P*=.09).

Conclusions: These findings advance our understanding of how people process popular message characteristics in short-form videos and lend practical guidance for communicators. We encourage medical professionals to consider short-form video sites (eg, TikTok and Instagram Reels) as effective tools for discussing clinical trials and participation opportunities. Specifically,

XSL•FO
RenderX

featuring doctors discussing efforts to reduce logistical barriers is recommended. Our measuring of actual behavior as an outcome is a rare and noteworthy contribution to this research.

(J Med Internet Res 2024;26:e49600) doi:10.2196/49600

KEYWORDS

clinical trial recruitment; TikTok; source credibility; framing; psychological barriers; logistical barriers; integrated behavioral model; short-form videos; social media use; clinical trial; recruitment

Introduction

Background

The development and advancement of medical treatments rely heavily on the willingness of individuals to participate in clinical trial research. However, clinical trials have historically had low enrollment rates, especially among underrepresented groups [1,2]. Research aimed at increasing willingness to participate in clinical trials has examined several factors impeding participation [3,4], including a lack of access [4,5], psychological barriers (such as issues related to fear or a lack of trust in research, doctors, and the process), and a lack of understanding of the process and value of clinical trials [6]. Others are deterred by logistical barriers, including financial constraints, time commitment, travel difficulties, and interference with other obligations (eg, work and family) [3,4]. Each barrier requires different solutions and communication strategies to effectively increase overall clinical trial participation [3]. For instance, Clark et al [3] explain that providing compensation, transportation, and flexible hours can help to lessen time and resource-related constraints (ie, logistical barriers); whereas barriers related to fear and a lack of trust (ie, psychological barriers) can be alleviated by explaining the voluntary nature of participation, emphasizing participant safety, and clarifying the participation process. Such actionable recommendations require adjustments by trial administrators (eg, ensuring compensation or help with scheduling and transportation), but also have important communication implications as the information related to overcoming the barriers must be shared with prospective participants.

While communication to prospective participants can be done through various channels (eg, recruitment ads), health care providers on the internet have been found to be critical sources for communicating clinical trial information to US adults [7]. Social media use has saturated society, with 70% of the US population using at least one form of social media [8]. More specifically, short-form video has widely been used for social media apps such as TikTok, a video sharing social media app that allows users to create and host short video content (the worldwide number of TikTok users doubled from 291.4 million in 2019 to 655.9 million in 2021) [9]. Short-form videos, typically lasting from a few seconds to a few minutes, encompass a wide variety of entertaining content, including comedy, dance, music, cooking, fitness, and other daily activities [10]. With social media skyrocketing in popularity, short-form videos are believed to have public health implications as people use TikTok, specifically, to disseminate health-related content [11]. This includes many conversations related to clinical trial research. A May 2023 search indicates that since the platform's worldwide launch in 2018, the hashtags "#clinicaltrial" and "#clinicalresearch" have received 43.4 million and 22.4 million views, respectively. TikTok has made it increasingly cost-effective for communicators of all types to disseminate information, as anyone with a smartphone can easily record and edit videos directly in the application.

Communication Source

The most popular sources communicating health information on TikTok are health care professionals (eg, doctors and clinical researchers) and general TikTok users [12,13]. Doctors and other health care professionals have used the platform to cover a variety of topics, including COVID-19, vaping, chronic pulmonary obstructive disease, and diabetes [11-13]. On the topic of clinical trials, doctors on TikTok have shared information on topics such as finding trials [14], upcoming trials [15], the importance of trials [16], and what the participation process is like [17]. Much of the content aims to help potential participants feel more comfortable. For instance, Dr Donald Garcia, medical director at Austin Clinical Trial Partners, explains in his video that institutional review boards are in place to help keep participants safe [17].

Likewise, some general TikTok users have taken to the platform to share their firsthand experience of being in a clinical trial themselves. Some record themselves within the actual clinical trial setting to provide an inside look at what the environment and procedures are like [18], while others detail their experience from home and sometimes answer questions that other users have about the experience [19]. In many cases, these videos advocate for others to participate, too. This sometimes includes the participants sharing details of the monetary compensation they have made from clinical trials and encouraging others to participate for this reason [20]. Testimonials from TikTok users serve as peer-to-peer exchanges of experience-based health information.

The effectiveness of these 2 dominant source types is likely to depend on how the audience perceives each source's credibility. McCroskey and Teven [21] suggested that 3 key components comprise credibility: competence, trustworthiness, and goodwill. As prior clinical trial participants have firsthand experience with participation, they may be perceived to have high trustworthiness and believability, which are both important for message acceptance [22]. Furthermore, as peer-type sources have similarities to the audience, this can also help to improve message perceptions and effectiveness [23].

Doctors, on the other hand, will likely generate high levels of perceived credibility because of their extensive education and experience (ie, competence). Merely having a respected, authoritative title of "Dr," alone, can lead others to respect them

XSL•FO
RenderX

and listen to what they have to say [24]. Besides their competence, health care providers have been found to be the most trusted source of information about clinical trials in comparison to government health agencies, health organizations, and others such as support groups, drug companies, families, and friends [1]. In addition, research examining the effects of various sources communicating general health information via health podcasts showed that doctors were found to elicit greater perceived source competence and trustworthiness in comparison to peers (individuals with firsthand experience with the health topic) [25]. In addition to competence and trust, doctors and peers discussing clinical trials on TikTok may also generate perceptions of goodwill. Sharing information on TikTok is voluntary, which may help users feel that the doctors and peers have positive intent, truly wanting to help and care for others.

Altogether, given doctors' high levels of competence (relative to peer sources) and their likelihood of generating perceptions of trust and goodwill, this study predicts that clinical trial messaging on TikTok coming from doctors will result in greater perceptions of source credibility than similar messages that come from general TikTok peers (hypothesis 1 [H1]): doctor-featured messages will lead to greater perceived source credibility toward clinical trial participation than peer-featured messages.

Attitude change can occur through the peripheral route or central route based on individuals' levels of ability and motivation to think (elaboration likelihood model [ELM] [26]). Source credibility has been found to influence attitude through the peripheral route because it is easy to process [27]. In the case of doctors on TikTok, their credibility (conveyed through their title, introduction, appearance, etc) can serve as peripheral cues (eg, "experts are generally correct") that improve message acceptance [26]. Since perceived credibility affects message processing and message acceptance [22,28], doctor-featured messages are also expected to cultivate improved audience attitudes and perceptions of self-efficacy. Generally, people rely on sources they trust and reject information from sources they do not trust [29], and acceptance of content influences one's behavior [30]. Thus, if the source is more credible, individuals will be more likely to perceive the recommended health behavior from the source as effective and feasible. Within the context of TikTok and health content, Song et al [13] found that the information quality of peer TikTok videos was low and engagement (eg, likes, shares) with health professional videos was high. Based on the findings, they suspected that the health professionals conveyed a greater level of expertise that then generated greater credibility perceptions, which could, in turn, improve the adoption of the message recommendations. Research has also shown that increased source credibility leads to favorable attitudes [31] and increased self-efficacy [32] about suggested health behavior. Based on these findings, we hypothesize that as the doctor-featured clinical trial messages increase perceptions of credibility, this will lead to more favorable attitudes and increased self-efficacy (hypothesis 2 [H2]): Doctor-featured messages will lead to (1) more favorable attitudes and (2) greater self-efficacy toward clinical trial participation than peer-featured messages through increased source credibility.

Then, as attitude and self-efficacy are larger drivers of behavioral intention (integrated behavioral model [IBM] [33,34]), the increased attitude and self-efficacy (generated by the doctor-featured messages) will lead to greater intentions to participate in a future trial (hypothesis 3 [H3]): Doctor-featured messages will lead to greater behavioral intention toward clinical trial participation than peer-featured messages through increased (1) attitudes and (2) self-efficacy.

Message Framing

Within the clinical trial messaging produced by both doctors and prior clinical trial participants, there is an opportunity to emphasize particular information (eg, information about the benefits of participating and the safety of participating). Such emphasis is a form of message framing and is a means of making particular information salient to the audience [35-37]. Framing has been recommended as a strategy for helping to overcome barriers associated with clinical trial participation [3,38,39]. In specific, emphasis framing—which involves designing messages to focus on particular subsets of information related to an overall topic [40]—has been recommended to help prospective participants see how relevant barriers can be overcome (eg, emphasizing safety precautions may help reduce psychological barriers to participation) [3]. Framing can help improve message processing by providing a simplified structure focused on the most important information [41,42]. This leads the audience to focus on and better remember the emphasized information [43,44]. Thus, in the case of clinical trial messaging, when the emphasis is on information that reduces concerns related to the common barriers, the audience will be more likely to process this information, which can then lead to the message favorably affecting their attitudes and behaviors [45]. In addition to having a more favorable attitude toward participation, their self-efficacy for participating should be improved as well, given that the emphasized information provides solutions to the barriers, which should help the barriers seem like less of a deterrent.

In this study, the TikTok messaging is designed to make salient information related to overcoming psychological and logistical barriers. Half of the messages talk about clinical trials with message frames focusing on overcoming psychological barriers, highlighting participants' well-being and safety during a clinical trial to help prospective participants feel more comfortable about the process (per recommendations from Clark et al [3]). Meanwhile, the logistical barrier–framed messages use the suggestions of Clark et al [3] of focusing on overcoming common logistical issues that impede participation (eg, transportation, scheduling, and finances) to help prospective participants envision how joining a trial could be possible for them [3]. That said, research investigating framing in this context is still new. Logistical barrier–framed messages have not been tested in this manner, and TikTok is still a new, understudied media platform. Specifically, with a diverse sample of individuals who have never participated in a trial before, it is not clear which barrier will be more relevant to them, and thus which barrier-framing will be more effective. Psychological and logistical barrier–framed messaging may be equally effective at improving attitudes and self-efficacy related to clinical trials, or 1 of the 2 framing types might have greater persuasive influence. To explore this, we ask the following

XSL•FO
RenderX

question (research question 1 [RQ1]): How does framing (focused on psychological barriers vs logistical barriers) influence (1) attitudes and (2) self-efficacy toward clinical trial participation?

If framing influences attitudes and self-efficacy, the framing may also affect intentions related to participation (through attitudes and self-efficacy) given the relationship between attitudes, self-efficacy, and behavior (IBM [33,34]). Thus, we also ask the following question (research question 2 [RQ2]): How does framing influence intention toward clinical trial participation through (1) attitudes and (2) self-efficacy toward clinical trial participation?

Pathways to Sign-Up Behavior

In addition to behavioral intention, in this study, we aim to also measure the actual behavior of our participants—whether or not they choose to be redirected (after the study) to a page where they can sign up to participate in future clinical trials. Measuring an actual behavior is valuable because it helps to better understand the persuasive effect of the messages.

While IBM focuses heavily on the formation of behavioral intention, it also explains the importance of environmental constraints and an individual's knowledge or skills (related to behavior performance) in impacting actual behavior change [46,47]. Environmental constraints are contextual factors that either help or hinder behavior performance [47]. Fishbein and Yzer [48] suggested that health educators should evaluate the potential effectiveness of addressing underlying beliefs or tackling environmental constraints, and then utilize the most appropriate one for the situation. For example, a prospective participant may not intend to participate in a trial because of their beliefs that trials are unsafe (eg, psychological barriers). In this case, a message strategy should be devised to focus on improving the factors influencing intention (eg, attitudes toward the behavior). On the other hand, if a prospective participant has the intrinsic motivation to participate in a trial but might encounter an environmental constraint like transportation issues (eg, logistical barriers), then an effective message strategy is to focus on addressing how to overcome the environmental constraint. In this case, by removing environmental constraints, the performance of the behavior (ie, clinical trial participation) is more likely [49].

Therefore, in this context, there may be indirect pathways in which the source and framing of the messages influence willingness to sign up for a future clinical trial. For instance, a doctor-featured message may increase perceived source credibility which may then improve attitude and self-efficacy which may then improve behavioral intentions and ultimately their behavior. Additionally, there may also be a direct effect of environmental constraints (ie, the contextual factors hindering the behavior) on behavior. For instance, if an individual does not have the literal means (eg, transportation) to get to a clinical trial, they will not participate. Based on this, we predict that our logistical barrier–framed messages (which provide information about overcoming common logistical issues) may have a direct effect on participants' behavior.

To explore the indirect pathways, we ask the following question (research question 3 [RQ3]): What are the indirect pathways that source and framing influence signing up for a clinical trial?

To examine the direct pathway to sign-up behavior, we pose the following hypothesis (hypothesis 4 [H4]): Logistical barrier–framed messages will lead to a greater likelihood of signing up for a clinical trial than psychological barrier–framed messages.

Combined Effects of Communication Source and Message Framing

Previous research has examined the interaction between communication sources and message framing [31,50,51]. However, these studies primarily focused on gain and loss frames (or relatedly, positive and negative frames), and the results were inconsistent. For example, Jones et al [31] found the expert source or positive-framed messages generated more behavioral intention than other conditions (eg, nonexpert/positive), whereas Borah and Xiao [50] and Huang and Liu [51] did not find significant interaction effects between framing and source. Considering the novelty of our study, which examines psychological and logistical barrier–framed messages for the first time, it is unclear which barrier frame would be more effective and how it interacts with the message source. Therefore, we ask the following question (research question 4 [RQ4]): How do framing (focused on psychological barriers vs logistical barriers) and message source (doctor vs peer) jointly influence attitudes and self-efficacy, subsequently affecting behavioral intention and sign-up behavior regarding clinical trial participation?

Study Overview

This study aims to examine the effects of short-form video's message features on clinical trial–related attitudes and behaviors. We test the effects of the 2 communication sources, doctors and peers, commonly seen on TikTok and other social media, as well as the effects of discussing 2 different barriers to clinical trials (psychological and logical framing) in the videos. Informed by the integrated behavioral model [34], the ELM [26], and framing [36,37], this study uses structural equational modeling to test our predictions: The source and message framing of videos indirectly influence behaviors related to clinical trial participation through attitude and self-efficacy, and message framing will directly affect signing up for a clinical trial. We also capture the actual behavior of our participants' willingness to join a sign-up list for a clinical trial and be contacted by clinical researchers about upcoming clinical participation opportunities. The inclusion of such an outcome is novel and allows us to avoid relying solely on behavioral intention as an indicator of actual behavior outcomes within this context. Figure 1 illustrates the conceptual framework for the above-proposed hypotheses and research questions.

XSL•FO
RenderX

Figure 1. Conceptual framework. CT: clinical trial; H1: hypothesis 1; H2: hypothesis 2; H3: hypothesis 3; H4: hypothesis 4; RQ1: research question 1; RQ2: research question 2; RQ3: research question 3; RQ4: research question 4.

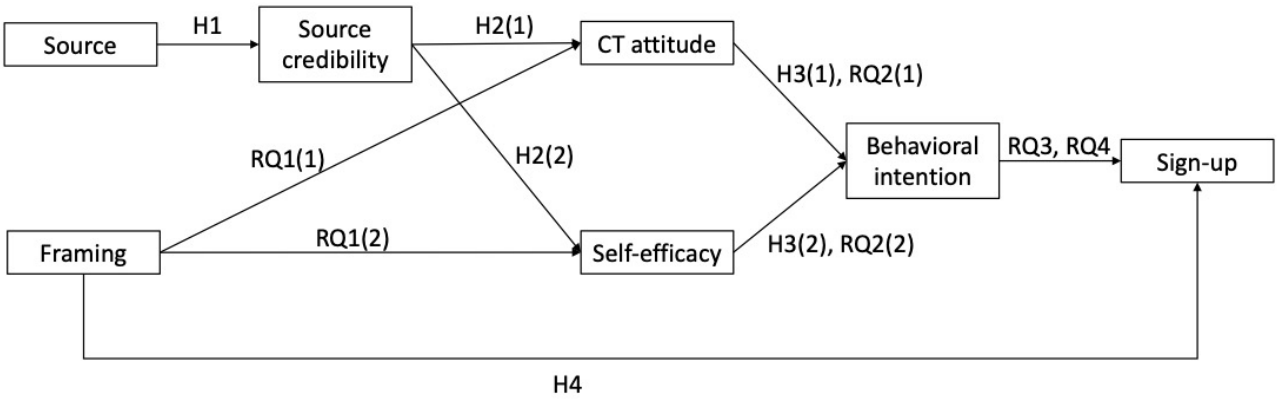

Methods

Experimental Design

This study used a 2 (source: doctor vs peer) × 2 (framing: psychological vs logistical) between-subject factorial design web-based experiment targeting adults in the United States who have never participated in clinical trials. A total of 4 conditions were formed (ie, doctor source/psychological barrier–framed, doctor source/logistical barrier–framed, peer source/psychological barrier–framed, and peer source/logistical barrier–framed). For each condition, 3 messages—featuring 3 different clinical trial topics (ie, sleep, stress, and caffeine)—were used to account for message variance [52].

Respondents and Procedure

Participants were eligible for this study if they were aged 18 years or older and had no prior clinical trial experience. A Qualtrics panel was used to recruit and compensate the study respondents in June 2022. The questionnaire began with informed consent, demographics (ie, age, gender, education, income, political ideology, and race), a question regarding their prior clinical trial participation, and a video test to ensure respondents could see and hear videos. After passing the video test, respondents answered questions measuring their preexisting attitude toward clinical trial participation and were then randomly assigned to 1 of the 4 experimental conditions where they viewed 3 TikTok videos in that condition. The presentation order of the 3 videos was fully randomized to eliminate response bias related to the order of the video presentation. After each of the 3 videos, respondents answered questions checking the manipulation and measuring source credibility and attitudes toward clinical trial participation. After viewing all 3 videos, respondents answered questions measuring self-efficacy toward clinical trial participation and intention to participate in a clinical trial. Last, an IRB-approved fabricated scenario was presented, and respondents were asked to choose whether they were willing to sign up to join a clinical trial participant list and be contacted by clinical researchers about upcoming clinical trial participation opportunities right after this survey. After answering this question, respondents saw the debriefing statement. In the questionnaire, items associated with the same variable were randomized to prevent order-related bias. Each questionnaire page contained between 1 and 5 questions, with a maximum of 67 pages (screens) a respondent could view. Respondents could not review or change their answers during the survey, but they could withdraw their responses at any time.

Qualtrics panel experts handled visitor tracking, participation rates, completion rates, cookies, and IP address duplication checks. The authors double-checked the final data received from the Qualtrics panel. After removing the low-quality responses (n=7), which included speeders (respondents who spent less than half-median amount of time on the survey based on Qualtrics panel expert criteria in the industry), straightliners (respondents who consistently gave identical responses, such as all “7” or “4,” to a series of questions using the same response scale), and respondents with missing data, a total of 561 respondents were used in the data analysis (doctor source/psychological barrier–framed=144, doctor source/logistical barrier–framed=139, peer source/psychological barrier–framed=144, and peer source/logistical barrier–framed=134).

Stimuli and Manipulation

The stimuli were TikTok videos recorded by 3 female actors on 3 different clinical trial topics (ie, sleep, stress, and caffeine). Each actor focused on one topic and recorded a video for each of the 4 experimental conditions. Each video includes 5 shots: introduction, clinical trial topic overview, mention of barriers, overcoming the barriers, and recommendations. Each shot varied to reflect the 4 experimental conditions. The key messages were also added as captions to each video. All other video features were kept similar across the 4 experimental conditions, including the video length (around 90 seconds), TikTok logo, and caption color and font. All 12 transcripts were edited by a professional news editor, and all 12 videos were reviewed by the research team. Examples of message transcripts are provided in Multimedia Appendix 1.

The source referred to the type of speakers in the video, including doctor source (ie, clinical researcher) and peer source (ie, prior participant). In the doctor-featured videos, actors wore professional white lab coats in an office setting with a blank wall background, while in the peer-featured videos, actors wore casual clothes in a home setting. Specifically, in each doctor-featured message, the actor introduced herself with the “Dr” honorific and as a clinical trial researcher and then brought up the clinical trial topic she worked on. She then talked about the barriers people may have in terms of clinical trial

XSL•FO
RenderX

participation (ie, psychological or logistical) and how clinical researchers attempt to address the barriers. Last, the actor recommended that the audience sign up for clinical trials and reiterated that they attempted to address the barriers. In a peer-featured message, the actor introduced herself as a clinical trial participant and then shared her experience of clinical trial participation. She started by mentioning her barriers before clinical trial participation, and then she talked about how her barriers had been resolved during the clinical trial participation. Finally, the actor recommended the audience sign up for clinical trials and reiterated the clinical researchers' attempts to address the barriers.

Framing referred to the type of barrier information made salient in the post, including psychological barriers and logistical barriers. Psychological barrier–framed videos focused on addressing psychological barriers to participating in clinical trials (eg, fear about clinical trial participation and medical mistrust) by highlighting participants' well-being and safety during a clinical trial. Logistical barrier–framed videos centered on overcoming logistical barriers to clinical trial participation (eg, cost and flexibility) by emphasizing the monetary reward and flexibility of participation. These frame wordings were created based on the suggestion from Clark et al [3]. The framing was manipulated in the mention of barriers, overcoming the barriers, and recommendations sections in the videos. A manipulation check was performed to ensure the 2 framing categories were distinguished in the videos. After each video, respondents were asked to indicate what the video emphasized (1=participants' well-being and safety, 5=Monetary reward and flexibility of participation). An independent samples *t* test showed a significant difference between the psychological barrier–framed videos (mean 1.54, SD 0.84; n=288) and the logistical barrier–framed videos (mean 3.88, SD 0.94; n=273), $t_{545.55}=-31.04$ (2-tailed), $P<.001$. Thus, the psychological barrier–framed videos were perceived as emphasizing participants' well-being and safety more whereas the logistical barrier–framed videos were perceived as emphasizing monetary reward and flexibility of participation more.

Measurements

Preexisting attitude toward clinical trial participation assessed respondents' positive or negative feelings toward clinical trial participation before exposure to the stimuli. Respondents were asked to rate the degree of their perception of clinical trial participation on a 5-item, 5-point semantic differential scale adapted from Kang and Lee [53], such as bad/good and negative/positive. (across the conditions, Cronbach α=0.87-0.92.)

Perceived source credibility assessed respondents' perceptions of the video speaker's competence, goodwill, and trustworthiness with an 18-item scale adapted from McCroskey and Teven [21]. Respondents were asked to rate their perception of the speaker on a 5-point bipolar scale, including items such as untrained/trained, cares about me/does not care about me, and untrustworthy/trustworthy (across the conditions, Cronbach α=0.96-0.98).

Self-efficacy measured respondents' belief in their capabilities to participate in clinical trials with a 5-point Likert scale of 2 items adapted from Lee et al [54]. Items included, "for me, to participate in a clinical trial would be difficult (1) to easy (5)" and "How certain are you that you could participate in a clinical trial? (1=Not at all certain; 5=Very certain)" (across the conditions, Cronbach α=0.82-0.86).

Attitude toward clinical trial participation assessed respondents' positive or negative feelings toward clinical trial participation after clinical trial to the stimuli with the same scale used to measure preexisting attitudes toward clinical trial participation. (across the conditions, Cronbach α=0.96-0.97).

Behavioral intention assessed the likelihood that respondents would participate in clinical trials with a 5-point Likert scale of 2 items adapted from Chen et al [55]. Respondents were asked to rate their opinion on 2 items from strongly disagree (1) to strongly agree (5), including "I plan on joining a clinical trial" and "I am willing to join a clinical trial" (across the conditions, Cronbach α=0.83-0.91).

Sign-up behavior was measured by whether the respondents clicked "Yes" to move forward to a sign-up page to participate in future clinical trials. Although this question did not measure participants' behavior beyond clicking "Yes" to share their email address, the behavior of choosing to go to the sign-up page is a reasonably accurate indicator for actual sign-up behavior (Yes=45.6%, 256/561). After participants answered this question, they were not directed to a sign-up page. Instead, they saw a debriefing statement that explained the true purpose of this question, and they were given information on how to find clinical trial participation options.

Ethical Considerations

This study received approval from the institutional review board of the University of Missouri (IRB #2054423). Informed consent was obtained from participants at the beginning of the survey. The consent form explained that the survey would take approximately 15-20 minutes, emphasized voluntary and anonymous participation, assured participants that their data would be stored securely in a password-protected electronic format, and clarified that researchers would only report aggregate data. Participants who did not agree with the consent form had the option to opt out. A Qualtrics panel was used to recruit and compensate the study participants at the agreed-upon rate established between the participants and Qualtrics.

Statistical Analysis

The hypothesized model was tested using structural equation modeling with the lavaan package for R [56]. All models were estimated using robust maximum likelihood unless bootstrapping was used, in which case ML estimation was adopted. Robust maximum likelihood was used because this model includes a binary outcome variable (ie, sign-up) [57]. Following Kline's (2015) 2-step process, a measurement model was first fit to verify the factor structure of clinical trial attitude, efficacy, and behavioral intention. Subsequently, to test the hypotheses, a structural model was fit in which source credibility, clinical trial attitudes, and self-efficacy were regressed on dummy-coded message source (doctor=1, peer=0), framing (psychological barrier frame=1, logistical barrier frame=0), and the interaction term of source and framing. The behavioral intention was

XSL•FO
RenderX

regressed on clinical trial attitudes and self-efficacy. Sign-up behavior was regressed on behavioral intention, message source, framing, and the interaction term of source and framing. All variables in the model were regressed on the control variable (ie, preexisting attitudes toward clinical trial participation). To maintain the model simplicity, however, the interaction term of source and framing was omitted from the final model as it did not exert any significant effect.

Results

Sample Overview

The average age of the sample was 49 (SD 17.52) years, and there were more females (301/561, 53.7%) than males (256/561, 45.6%) and 4 participants did not disclose. Respondents were primarily White (433/561, 77.2%), followed by Black or African American (52/461, 9.3%), Asian (39/561, 7%), and others (30/561, 5.3%). Approximately half of the respondents completed high school or some college (286/561, 51%), 13.7% (n=77) respondents had an associate's degree, and 32.8% (n=184) respondents had a bachelor degree or above. From liberal (1) to conservative (7), the sample leaned conservative (mean 3.93, SD 1.59).

Modeling Results

Model fit for the measurement model was acceptable based on the criteria from MacCallum et al [58] and Little [59] (χ^2_{24}=92.6, *P*<.001); robust root-mean-square error of approximation=0.063 (90% CI 0.079-0.097); robust comparative fit index=0.984; robust nonnormed fit index/Tucker Lewis index=0.975; and standardized root mean residual=0.026. The final structural model (Figure 2) achieved good model fit based on the same criteria as above (χ^2_{65}=206.01, *P*<.001); robust root-mean-square error of approximation=0.066 (90% CI 0.056-0.076); robust comparative fit index=0.974; robust nonnormed fit index/Tucker Lewis index=0.965; and standardized root mean residual=0.04.

Figure 2. Final structural model (N=561). The model controlled for preexisting attitudes toward clinical trial participation. Path coefficients are standardized (β). Dashed lines represent nonsignificant paths. **P*<.05, ***P*<.01, and ****P*<.001 (specific *P* values are reported in the results texts to minimize repetitive reporting).

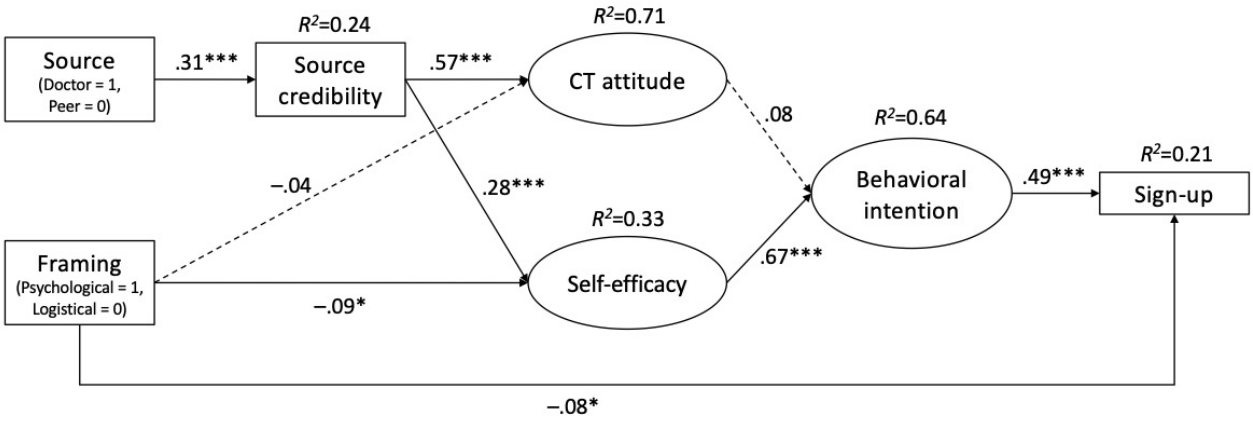

Hypotheses Testing

H1 predicted that doctor-featured messages would lead to greater perceived source credibility toward clinical trial participation than peer-featured messages. Consistent with this prediction, compared with peer-featured messages, doctor-featured messages were significantly associated with increased perceived source credibility (β=.31, *P*<.001).

H2 predicted that doctor-featured messages would lead to (1) more favorable attitudes and (2) greater self-efficacy toward clinical trial participation than peer-featured messages through increased source credibility. In support of this hypothesis, there were significant positive indirect effects of doctor-featured messages on clinical trial attitudes (95% CI 0.48-0.84) and self-efficacy toward clinical trial participation (95% CI 0.13-0.30) through the increased source credibility, as evidenced by a 5000 bootstrapped 95% CI that did not contain zero.

H3 hypothesized that doctor-featured messages would lead to greater behavioral intention toward clinical trial participation than peer-featured messages through increased (1) attitudes and (2) self-efficacy. The model showed clinical trial attitudes were not significantly associated with the increased intention to participate in clinical trials (β=.08, *P*=.14); thus, inconsistent with H3(1), doctor-featured messages did not lead to greater clinical trial participation intention than peer-featured messages through increased attitudes. However, self-efficacy was significantly positively associated with clinical trial participation intention (β=.67, *P*<.001); therefore, in support of H3(2), doctor-featured messages led to greater clinical trial participation intention than peer-featured messages through increased perceived source credibility and self-efficacy (95% CI 0.12-0.29).

RQ1 examined the influence of framing on (1) attitudes and (2) self-efficacy toward clinical trial participation. The results showed the effects of psychological barrier–framed messages and logistical barrier–framed messages did not significantly differ in attitudes toward clinical trial participation (β=–.04, *P*=.09). However, logistical barrier–framed messages led to significantly greater self-efficacy toward clinical trial participation than psychological barrier–framed messages (β=–.09, *P*=.02).

RQ2 investigated the indirect effects of framing on the intention to participate in clinical trials through (1) attitudes and (2) self-efficacy. Because different framing foci did not lead to

XSL•FO
RenderX

significantly different direct effects on attitudes (shown in RQ1(1)), the indirect effect of framing on intention to participate in clinical trials through attitudes could not be significant (RQ2(1)). For RQ2(2), however, logistical barrier–framed messages led to significantly greater intention to participate in clinical trials than psychological barrier–framed messages through increased self-efficacy (95% CI –0.38 to –0.03).

RQ3 explored the indirect effects of source and framing on sign-up behavior. There are 2 significant pathways. First, doctor-featured messages led to greater perceived source credibility, which, in turn, led to greater self-efficacy and then increased behavioral intention, which, last, boosted the clinical trial sign-up behavior (95% CI 0.02-0.04). Second, logistical barrier–framed messages led to greater self-efficacy, which, in turn, increased intention to participate in clinical trials, which, last, improved the clinical trial sign-up behavior (95% CI –0.06 to –0.004).

H4 posited a direct effect of framing on sign-up behavior, such that logistical barrier–framed messages would lead to a greater likelihood of signing up for a clinical trial than psychological barrier–framed messages. Consistent with this prediction, compared with psychological barrier–framed messages, logistical barrier–framed messages were associated with an increased likelihood of signing up for a clinical trial (β=–.08, P=.03).

RQ4 explored the joint effects of framing and sources. There was no significant interaction between these 2 variables, as tested in the modeling section. However, their combined effects were notable. The results revealed that framing and sources independently influenced sign-up behavior. Together, the model accounted for 21% of the variance in the likelihood of signing up for a clinical trial.

Discussion

Summary of Findings

This research offers several important findings that illuminate strategic ways to use short-form social media videos, such as those on TikTok, to improve perceptions of clinical trials and increase enrollment. The findings suggested that doctor-featured messages led to greater perceived source credibility, leading to greater self-efficacy, subsequently enhancing behavioral intention and clinical trial sign-up behavior. Logistical barrier–framed messages led to greater self-efficacy, resulting in higher intention to participate in clinical trials and improved sign-up behavior. Logistical barrier–framed messages were also directly associated with an increased likelihood of signing up for a clinical trial. The theoretical and practical implications, limitations, and future directions of this study are discussed below.

Finding Implications

First, using doctors (ie, clinical researchers) as the source sharing the information made a difference in how participants assessed credibility (H1). The doctors were perceived as more credible than the peer clinical-trial participants, and the credibility afforded to the doctors led to better attitudes about clinical trials and an increased feeling of self-efficacy to engage in clinical trials among the participants (H2). This finding accords with previous ELM research that found higher perceived credibility resulted in enhanced favorable attitudes [22,28]. Extending the path even further, the perceived self-efficacy, cultivated in part by the doctors' credibility, also led to greater intention to sign up for clinical trials (H3(2)) and ultimately increased clinical trial sign-up behavior (RQ3). This finding confirmed the significant mediating role of self-efficacy in the relationship between source credibility and behavioral intention. Unexpectedly, attitudes were found not to mediate this relationship (H3(1)), supporting the idea that the ELM is more a model of attitude change than persuasion as strong attitudes do not necessarily induce behavioral change particularly if self-efficacy is lacking [60]. Petty et al [61] stressed the importance of attitude change in eliciting behavioral change by explaining that self-efficacy can be "positive attitudes toward the self" (distinct from attitudes toward a behavior) and suggested scholars should identify the most important type of attitudes for predicting a particular health behavior. Despite varying explanations, our findings indicate that future scholars should incorporate self-efficacy into the ELM, either as an additional construct or a critical type of attitude in health communication. It is noteworthy that, as far as the authors are aware, this is the first ELM study that captures persuasive effects on a type of behavioral enactment in the form of clicking to sign up for future clinical trials.

Second, this research contributes to IBM research in that it demonstrated the effects of the messages on attitude, self-efficacy (RQ1), behavioral intention (RQ2), and sign-up behavior (RQ3/H4) on TikTok within the context of health communication. Fishbein and Yzer [48] suggested that changing the psychological determinants to perform a behavior would be more efficient than changing skills and environmental constraints because the latter is often difficult to change. However, this study found that certain environmental constraints, such as transportation issues (ie, logistical barriers), can be communicated and can directly lead to the desired behavioral performance. As IBM affords the flexibility to communicate about a wide range of determinants, a proper IBM-based research agenda could allow communicators to target a health behavior that is stymied by both psychological and logistical barriers.

Third, our findings about framing revealed that the logistical frame outperformed the psychological frame (RQ2(2), H4), which accords with expectations that different frames would lead to different reactions [36,37]. More specifically, framing according to perceived barriers to clinical trials and their solutions, was predicted to have persuasive power [3,38,39]. Regardless of source type, the messages that were framed to address how to overcome logistical barriers had a direct effect on increasing participants' clinical trial sign-up behavior (H4). This represents an original contribution, given that logistical barrier–framing has not yet been tested. The power of focusing on logistical problems and how to overcome them is promising for both theory and practice.

Finally, despite the lack of interaction between source and framing, the finding demonstrated doctor sourcing and logistical barrier–framing independently but cumulatively contributed to

XSL•FO
RenderX

increasing sign-up behavior (RQ4). This finding, along with the abovementioned findings regarding the independent effects of source and framing, suggests that incorporating doctor sources and logistical barrier–framing in health communication, particularly through short-form videos, can yield more successful outcomes in terms of behavioral change.

Limitations and Future Research

One limitation of this research is that it is difficult to measure full commitment to participate in a clinical trial. We designed the sign-up option to replicate the first behavioral step that a person could take when seeking out clinical trials. It is one of several steps a person would need to take to fully participate in a clinical trial, thereby representing behavior change. Another limitation of this study is the omission of perceived norms toward clinical trials in the integrated behavioral model. This decision assumed that clinical trials are less prevalent compared with other health behaviors, and therefore perceived norms may not exert significant variance. However, it is important to note that exposure to videos featuring prior clinical trial participants may enhance perceived norms by increasing the perception that other people with similar barriers are engaging in the behavior (ie, descriptive norms). Thus, future studies should explore the influence of source and framing on perceived norms and their subsequent influence on behavioral intentions and actual behaviors. Last, it is important to note that the sample in this study mainly consisted of White individuals with a conservative leaning. We suggest that scholars exercise caution when generalizing the current findings to significantly different racial and ideological groups.

Future research should also create similar stimuli to test the same framing and sourcing patterns with different health topics that might involve psychological and logistical barriers, such as blood donation. Furthermore, investigating source and framing preferences among diverse racial and age groups is also essential, given disparities in clinical trial participation within specific demographics [2]. Additionally, future research could examine whether one's level of familiarity and experience with TikTok (and other related platforms) moderate the effects of these message features.

Conclusion and Practical Recommendations

Given the effectiveness of the doctor-sourced videos in this study, we recommend that medical professionals take to short-form video sites such as TikTok to discuss clinical trials and participation opportunities. Recent recommendations have advised health professionals to leverage social media platforms like TikTok to disseminate health information [11,13], and this study provides evidence for this recommendation within this specific context of clinical trials. As none of the participants in this research had ever been part of a clinical trial before, this research is especially useful for outreach to people who have likely given clinical trials little thought or who have considered participating but did not follow through. This study mirrors one step in the recruiting process, which is getting potential participants to agree to receive information about future clinical trials. Recruiters could see a similar willingness to sign up for clinical trials if their promotional material involved doctors speaking about how they have worked to reduce logistical barriers. With TikTok users already turning to this platform to share clinical-trial experiences and seek related content (as evidenced by the popularity of the #clinicaltrial and #clinicalresearch hashtags), TikTok may be an especially useful mode of communication. This study serves to inform those interested in taking advantage of this new modality for public health messaging.

Acknowledgments

This study was supported by the University of Missouri Research Council Grant (URC-21-030).

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author upon request.

Authors' Contributions

All authors contributed to the conceptualization and methodology of the study. Data curation was conducted by NL and SH. SH performed the formal analysis and visualization. AH and SL were responsible for funding acquisition, project administration, and supervision. The investigation involved all authors, and they also provided the necessary resources and validation. SH and CEK wrote the original draft, and all authors participated in the review and editing of the draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Message transcripts for the sleep topic.
[PDF File (Adobe PDF File), 110 KB - jmir_v26i1e49600_app1.pdf]

References

XSL•FO
RenderX

1. National Cancer Institute. Clinical trial participation among US adults. Health Information National Trends Survey HINTS Briefs. 2022. URL: https://hints.cancer.gov/docs/Briefs/HINTS_Brief_48.pdf [accessed 2023-05-27]
2. Clinical trial diversity. US Food and Drug Administration. URL: https://www.fda.gov/consumers/minority-health-and-health-equity/clinical-trial-diversity [accessed 2023-05-27]
3. Clark LT, Watkins L, Piña IL, Elmer M, Akinboboye O, Gorham M, et al. Increasing diversity in clinical trials: overcoming critical barriers. Curr Probl Cardiol 2019;44(5):148-172 [FREE Full text] [doi: 10.1016/j.cpcardiol.2018.11.002] [Medline: 30545650]
4. Meng J, McLaughlin M, Pariera K, Murphy S. A comparison between Caucasians and African Americans in willingness to participate in cancer clinical trials: the roles of knowledge, distrust, information sources, and religiosity. J Health Commun 2016;21(6):669-677. [doi: 10.1080/10810730.2016.1153760] [Medline: 27175604]
5. Kim SH, Tanner A, Friedman DB, Foster C, Bergeron C. Barriers to clinical trial participation: comparing perceptions and knowledge of African American and White South Carolinians. J Health Commun 2015;20(7):816-826. [doi: 10.1080/10810730.2015.1018599] [Medline: 26042496]
6. Tanner A, Kim SH, Friedman DB, Foster C, Bergeron CD. Barriers to medical research participation as perceived by clinical trial investigators: communicating with rural and African American communities. J Health Commun 2015;20(1):88-96. [doi: 10.1080/10810730.2014.908985] [Medline: 25204763]
7. Yadav S, Todd A, Patel K, Tabriz AA, Nguyen O, Turner K, et al. Public knowledge and information sources for clinical trials among adults in the USA: evidence from a health information national trends survey in 2020. Clin Med (Lond) 2022;22(5):416-422 [FREE Full text] [doi: 10.7861/clinmed.2022-0107] [Medline: 36507814]
8. Social media fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2023-05-27]
9. Yuen M. TikTok users worldwide (2020-2025). Insider Intelligence. URL: https://www.insiderintelligence.com/charts/global-tiktok-user-stats/ [accessed 2023-05-27]
10. Zhang X, Wu Y, Liu S. Exploring short-form video application addiction: socio-technical and attachment perspectives. Telemat Inform 2019;42:101243. [doi: 10.1016/j.tele.2019.101243]
11. Comp G, Dyer S, Gottlieb M. Is TikTok the next social media frontier for medicine? AEM Educ Train 2021;5(3):1-4 [FREE Full text] [doi: 10.1002/aet2.10532] [Medline: 34095694]
12. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021;23(9):e30409 [FREE Full text] [doi: 10.2196/30409] [Medline: 34468327]
13. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021;23(12):e28318 [FREE Full text] [doi: 10.2196/28318] [Medline: 34931996]
14. 1. Google search Pharma clinical trial list 2. search www.clinicaltrials.gov #obesity #clinicaltrials #obesitymedicine #greenscreen #greenscreenvideo. TikTok. URL: https://www.tiktok.com/@michaelalbertmd/video/7176312644490562859?is_copy_url=1&is_from_webapp=v1&item_id=7176312644490562859&q=%23clinicaltrial&t=1672869700860 [accessed 2023-05-27]
15. Have you heard of the new MS clinical trial, ATA-188?! for my FULL review – check my YouTube channel!. TikTok. URL: https://www.tiktok.com/@drgretchenpt/video/7111861106326637870 [accessed 2023-05-27]
16. There are many successful clinical trials! let's talk about one! #clinicaltrial #cancermedicine #tiktokoncdoc #oncologist #curecancer #cll #cance. TikTok. URL: https://www.tiktok.com/@draljazayrly/video/7025666462362750213 [accessed 2023-05-27]
17. As a clinical trial participant, you are not a “guinea pig” or “lab rat” but an integral part of the research team #clinical trial #mental health #fyp. TikTok. URL: https://www.tiktok.com/@atxctp/video/7170355183497825582 [accessed 2023-05-27]
18. This is my first attempt at a voiceover lol but yeah this is my first day doing a clinical trial, follow for more updates #flucamp #clinicaltrial #nhs #london. TikTok. URL: https://www.tiktok.com/@faithlarks/video/7184495430183554310 [accessed 2023-05-27]
19. Part 1 #covid19 #clinicaltrial #vaccine participant timeline of Pfizer’s clinical trial. TikTok. URL: https://www.tiktok.com/@essdeedub/video/6970333799423036678 [accessed 2023-05-27]
20. I did my first clinical trial in 5 years! just google clinical trials to find one near you! it was 6 days in there then I returned for 4 out patient visits.it was an appetite suppressant. use your money wisely! #clinicaltrial #studies #gatekeep #nogatekeeping. TikTok. URL: https://www.tiktok.com/@chinyereogazi/video/7146248914302225706 [accessed 2023-05-27]
21. McCroskey JC, Teven JJ. Goodwill: a reexamination of the construct and its measurement. Commun Monogr 1999;66(1):90-103. [doi: 10.1080/03637759909376464]
22. Dong Z. How to persuade adolescents to use nutrition labels: effects of health consciousness, argument quality, and source credibility. Asian J Commun 2015;25(1):84-101. [doi: 10.1080/01292986.2014.989241]
23. Paek HJ, Hove T, Ju Jeong H, Kim M. Peer or expert? Int J Advert 2011;30(1):161-188. [doi: 10.2501/ija-30-1-161-188]
24. Rhoads KVL, Cialdini RB. The business of influence: principles that lead to success in commercial settings. In: Dillard JP, Pfau M, editors. The Persuasion Handbook: Developments in Theory and Practice. Thousand Oaks, California: Sage; 2002:513-542.

XSL•FO
RenderX

25. Kirkpatrick CE, Lee S. The impact of source and message relevance on audience responses to health podcasts. Commun Rep 2021;34(2):78-91. [doi: 10.1080/08934215.2021.1912129]
26. Petty R, Cacioppo J. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Berlin/Heidelberg, Germany: Springer-Verlag; 1986.
27. Petty RE, Cacioppo JT, Goldman R. Personal involvement as a determinant of argument-based persuasion. J Pers Soc Psychol 1981;41(5):847-855. [doi: 10.1037//0022-3514.41.5.847]
28. Wilson EJ, Sherrell DL. Source effects in communication and persuasion research: a meta-analysis of effect size. JAMS 1993;21:101-112. [doi: 10.1007/BF02894421]
29. Malka A, Krosnick JA, Langer G. The association of knowledge with concern about global warming: trusted information sources shape public thinking. Risk Anal 2009;29(5):633-647. [doi: 10.1111/j.1539-6924.2009.01220.x] [Medline: 19302280]
30. Dutta-Bergman MJ. The linear interaction model of personality effects in health communication. Health Commun 2003;15(1):101-115. [doi: 10.1207/S15327027HC1501_5] [Medline: 12553779]
31. Jones LW, Sinclair RC, Courneya KS. The effects of source credibility and message framing on exercise intentions, behaviors, and attitudes: an integration of the elaboration likelihood model and prospect theory. J Appl Soc Pyschol 2006;33(1):179-196. [doi: 10.1111/j.1559-1816.2003.tb02078.x]
32. De Meulenaer S, De Pelsmacker P, Dens N. Power distance, uncertainty avoidance, and the effects of source credibility on health risk message compliance. Health Commun 2018;33(3):291-298. [doi: 10.1080/10410236.2016.1266573] [Medline: 28094542]
33. Fishbein M. The role of theory in HIV prevention. AIDS Care 2000;12(3):273-278. [doi: 10.1080/09540120050042918] [Medline: 10928203]
34. Fishbein M. An integrative model for behavioral predictionits application to health promotion. In: DiClemente RJ, Crosby RA, Kegler MC, editors. Emerging Theories in Health Promotion Practice and Research 2nd ed. New York City, United States: Jossey-Bass/Wiley; 2009:215-234.
35. Entman RM. Framing: toward clarification of a fractured paradigm. J Commun 1993;43(4):51-58. [doi: 10.1111/j.1460-2466.1993.tb01304.x]
36. Kelly BJ, Hornik RC. Effects of framing health messages in terms of benefits to loved ones or others: an experimental study. Health Commun 2016;31(10):1284-1290. [doi: 10.1080/10410236.2015.1062976] [Medline: 26940483]
37. Uskul AK, Oyserman D. When message-frame fits salient cultural-frame, messages feel more persuasive. Psychol Health 2010;25(3):321-337 [FREE Full text] [doi: 10.1080/08870440902759156] [Medline: 20204971]
38. Hu S, Kirkpatrick CE, Hong Y, Lee N, Lee S, Hinnant A. Improving rural white men's attitudes toward clinical trial messaging and participation: effects of framing, exemplars and trust. Health Educ Res 2022;37(6):476-494 [FREE Full text] [doi: 10.1093/her/cyac026] [Medline: 36169357]
39. Kirkpatrick CE, Hu S, Lee N, Hong Y, Lee S, Hinnant A. Overcoming black Americans' psychological and cognitive barriers to clinical trial participation: effects of news framing and exemplars. Health Commun 2023;38(12):2663-2675. [doi: 10.1080/10410236.2022.2105619] [Medline: 35924326]
40. Druckman JN. The implications of framing effects for citizen competence. Polit Behav 2001;23(3):225-256. [doi: 10.1023/a:1015006907312]
41. Davies J, Mabin VJ. Knowledge management and the framing of information: a contribution to OR/MS practice and pedagogy. J Oper Res Soc 2017;52(8):856-872. [doi: 10.1057/palgrave.jors.2601171]
42. Zaller J, Feldman S. A simple theory of the survey response: answering questions versus revealing preferences. Am J Political Sci 1992;36(3):579-616. [doi: 10.2307/2111583]
43. Campbell HA, Hawk D. Al jazeera's framing of social media during the arab spring. CyberOrient 2020;6(1):34-51. [doi: 10.1002/j.cyo2.20120601.0003]
44. Wasike BS. Framing news in 140 characters: how social media editors frame the news and interact with audiences via twitter. Glob Media J Can Ed 2013;6(1):5-23.
45. Tewksbury D, Scheufele D. News framing theory and research. In: Oliver MB, Raney AA, Bryant J, editors. Media effects. 4 ed. London, United Kingdom: Routledge; 2019:51-68.
46. Fishbein M, Cappella JN. The role of theory in developing effective health communications. J Commun 2006 Aug;56(s1):S1-S17. [doi: 10.1111/j.1460-2466.2006.00280.x]
47. Yzer M. The integrative model of behavior prediction as a tool for designing health messages. In: Designing Messages for Health Communication Campaigns: Theory and Practice. United Kingdom: Sage; 2012:21-40.
48. Fishbein M, Yzer MC. Using theory to design effective health behavior interventions. Commun Theory 2003;13(2):164-183. [doi: 10.1111/j.1468-2885.2003.tb00287.x]
49. Montaño D, Kasprzyk D. Theory of reasoned action, theory of planned behavior,the integrated behavioral model. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior: Theory, Research, and Practice. 5 ed. Hoboken , NJ: Jossey-Bass; 2015:168-222.
50. Borah P, Xiao X. The importance of 'likes': the interplay of message framing, source, and social endorsement on credibility perceptions of health information on facebook. J Health Commun 2018;23(4):399-411. [doi: 10.1080/10810730.2018.1455770] [Medline: 29601271]

XSL•FO
RenderX

51. Huang Y, Liu W. Promoting COVID-19 vaccination: the interplay of message framing, psychological uncertainty, and public agency as a message source. Sci Commun 2021;44(1):3-29. [doi: 10.1177/10755470211048192]
52. Thorson E, Wicks R, Leshner G. Experimental methodology in journalism and mass communication research. J Mass Commun Q 2012;89(1):112-124. [doi: 10.1177/1077699011430066]
53. Kang H, Lee MJ. Designing anti-binge drinking prevention messages: message framing vs. evidence type. Health Commun 2018;33(12):1494-1502. [doi: 10.1080/10410236.2017.1372046] [Medline: 28952800]
54. Lee S, Cappella JN, Lerman C, Strasser AA. Effects of smoking cues and argument strength of antismoking advertisements on former smokers' self-efficacy, attitude, and intention to refrain from smoking. Nicotine Tob Res 2013;15(2):527-533 [FREE Full text] [doi: 10.1093/ntr/nts171] [Medline: 22949578]
55. Chen T, Dai M, Xia S, Zhou Y. Do messages matter? investigating the combined effects of framing, outcome uncertainty, and number format on COVID-19 vaccination attitudes and intention. Health Commun 2022;37(8):944-951. [doi: 10.1080/10410236.2021.1876814] [Medline: 33499677]
56. Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Softw 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
57. Bandalos DL. Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. Struct Equ Model: Multidiscip J 2014;21(1):102-116. [doi: 10.1080/10705511.2014.859510]
58. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1(2):130-149. [doi: 10.1037/1082-989X.1.2.130]
59. Little TD. Longitudinal Structural Equation Modeling. New York, NY: The Guilford Press; 2013.
60. O'Keefe D. The elaboration likelihood model. In: Dillard JP, Shen L, editors. The SAGE Handbook of Persuasion: Developments in Theory and Practice. London, UK: Sage Publications; 2013:137-149.
61. Petty R, Barden J, Wheeler S. The elaboration likelihood model of persuasion: developing health promotions for sustained behavioral change. In: DiClemente RJ, Crosby RA, Kegler M, editors. Emerging Theories in Health Promotion Practice and Research. San Francisco, CA: Jossey-Bass/Wiley; 2009:185-214.

Abbreviations

ELM: elaboration likelihood model
H1: hypothesis 1
H2: hypothesis 2
H3: hypothesis 3
H4: hypothesis 4
IBM: integrated behavioral model
RQ1: research question 1
RQ2: research question 2
RQ3: research question 3
RQ4: research question 4

Edited by T de Azevedo Cardoso; submitted 03.06.23; peer-reviewed by K Turner, M El Tantawi; comments to author 13.10.23; revised version received 20.10.23; accepted 12.04.24; published 15.08.24.

Please cite as:
Hu S, Kirkpatrick CE, Lee N, Hong Y, Lee S, Hinnant A
Using Short-Form Videos to Get Clinical Trial Newcomers to Sign Up: Message-Testing Experiment
J Med Internet Res 2024;26:e49600
URL: https://www.jmir.org/2024/1/e49600
doi:10.2196/49600
PMID:

©Sisi Hu, Ciera E Kirkpatrick, Namyeon Lee, Yoorim Hong, Sungkyoung Lee, Amanda Hinnant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptance of Social Media Recruitment for Clinical Studies Among Patients With Hepatitis B: Mixed Methods Study

Theresa Willem[1,2,3*], MA; Bettina M Zimmermann[1,2,4*], PhD; Nina Matthes[2]; Michael Rost[5], PhD; Alena Buyx[2], Prof Dr Med

[1]Institute of Molecular Immunology, Klinikum Rechts der Isar, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany
[2]Institute of History and Ethics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany
[3]Department of Science, Technology and Society (STS), School of Social Sciences and Technology, Technical University of Munich, Munich, Germany
[4]Institute of Philosophy, Multidisciplinary Center for Infectious Diseases, University of Bern, Bern, Switzerland
[5]Institute for Biomedical Ethics, University of Basel, Basel, Switzerland
[*]these authors contributed equally

Corresponding Author:
Bettina M Zimmermann, PhD
Institute of History and Ethics in Medicine
TUM School of Medicine and Health
Technical University of Munich
Ismaninger Str. 22
Munich, 81675
Germany
Phone: 49 89 4140 4041
Email: bettina.zimmermann@tum.de

Abstract

Background: Social media platforms are increasingly used to recruit patients for clinical studies. Yet, patients' attitudes regarding social media recruitment are underexplored.

Objective: This mixed methods study aims to assess predictors of the acceptance of social media recruitment among patients with hepatitis B, a patient population that is considered particularly vulnerable in this context.

Methods: Using a mixed methods approach, the hypotheses for our survey were developed based on a qualitative interview study with 6 patients with hepatitis B and 30 multidisciplinary experts. Thematic analysis was applied to qualitative interview analysis. For the cross-sectional survey, we additionally recruited 195 patients with hepatitis B from 3 clinical centers in Germany. Adult patients capable of judgment with a hepatitis B diagnosis who understood German and visited 1 of the 3 study centers during the data collection period were eligible to participate. Data analysis was conducted using SPSS (version 28; IBM Corp), including descriptive statistics and regression analysis.

Results: On the basis of the qualitative interview analysis, we hypothesized that 6 factors were associated with acceptance of social media recruitment: using social media in the context of hepatitis B (hypothesis 1), digital literacy (hypothesis 2), interest in clinical studies (hypothesis 3), trust in nonmedical (hypothesis 4a) and medical (hypothesis 4b) information sources, perceiving the hepatitis B diagnosis as a secret (hypothesis 5a), attitudes toward data privacy in the social media context (hypothesis 5b), and perceived stigma (hypothesis 6). Regression analysis revealed that the higher the social media use for hepatitis B (hypothesis 1), the higher the interest in clinical studies (hypothesis 3), the more trust in nonmedical information sources (hypothesis 4a), and the less secrecy around a hepatitis B diagnosis (hypothesis 5a), the higher the acceptance of social media as a recruitment tool for clinical hepatitis B studies.

Conclusions: This mixed methods study provides the first quantitative insights into social media acceptance for clinical study recruitment among patients with hepatitis B. The study was limited to patients with hepatitis B in Germany but sets out to be a reference point for future studies assessing the attitudes toward and acceptance of social media recruitment for clinical studies. Such empirical inquiries can facilitate the work of researchers designing clinical studies as well as ethics review boards in balancing the risks and benefits of social media recruitment in a context-specific manner.

(J Med Internet Res 2024;26:e54034) doi:10.2196/54034

XSL•FO
RenderX

KEYWORDS

Facebook; Twitter; social media; clinical trial; enrollment; health technology acceptance; ethics; infectious diseases; privacy; data protection; stigma; discrimination

Introduction

Benefits and Risks of Using Social Media Recruitment for Clinical Studies

Recruiting clinical study participants through social media has the potential to increase the recruitment accrual in a cost-effective way [1]. Consequently, social media recruitment has been increasingly applied for clinical studies, often in parallel with other recruitment strategies. However, social media recruitment still bears a host of challenges. First, maintaining a social media presence and community management can be resource intensive. Second, when used as a stand-alone recruiting method, it might yield a cohort of limited demographic representativeness. Finally, social media recruitment comes with ethical issues, particularly when used to recruit for clinical studies [2]. Because social media recruitment includes reaching potential research participants outside a clinical setting and in a public online space without direct personal contact, risks related to social stigma, privacy infringement, loss of trust, and psychological harm have been discussed [3]. To mitigate some of these risks, prioritizing investigator transparency and obtaining explicit consent when recruiting from others' social network was suggested [4]. Yet, because the activities of social media platforms are primarily unregulated and partly belong to large global technology companies, activities conducted on social media, including study recruitment, can never be fully controlled by researchers or institutions. Remaining privacy-infringing risks include hidden data collection and profiling, particularly problematic for patients carrying vulnerable characteristics [5].

Early studies assessing social media recruitment for clinical studies focused on the effectiveness of the method. For example, Frandsen et al [3] used social media recruitment for a smoking cessation trial and compared their cohort recruited from a Facebook-based approach to cohorts resulting from other recruitment methods. They found no differences between the cohorts regarding socioeconomic or smoking characteristics, except that participants recruited via Facebook were significantly younger. Wisk et al [4] recruited college students with type 1 diabetes, a hard-to-reach population, using a variety of outreach channels, including social media. They found that Facebook was the most successful recruitment method. Guthrie et al [5] found that Facebook advertising was significantly cheaper than recruiting via mail. While these studies allow insights into the utility of social media recruitment from the perspective of researchers, studies assessing patients' perspectives and attitudes toward social media for clinical study recruitment are lacking. This study aims to deliver first evidence on patient attitudes toward social media recruitment, focusing on patients with hepatitis B.

Patients With Hepatitis B and Social Media

Patients with hepatitis B are a particularly interesting cohort to study acceptance for social media recruitment as the particularities of the disease exhibit potentially confounding factors for their attitudes toward social media recruitment. First, there is robust empirical evidence that patients with hepatitis B can be subject to social stigma [6-10]. Therefore, the risk of public exposure to hepatitis B diagnosis on social media renders them—and patients with other stigmatized traits and conditions—particularly vulnerable in the context of social media recruitment [11]. Second, hepatitis B in Europe is particularly prevalent in certain immigrant populations, which are at risk of being neglected for clinical studies due to language barriers and lack of health care access. Social media recruitment can help include patient populations who otherwise would be disregarded for clinical studies or are hard to reach [12-14].

Study Rationale and Objectives

However, the effectiveness of social media recruitment crucially hinges on technology acceptance. To date, the attitudes of patients regarding social media recruitment are underexplored. Addressing this gap, this mixed methods study assesses factors predicting the acceptance of social media recruitment among patients with hepatitis B. On the basis of qualitative individual interviews with 6 patients with hepatitis B and 30 multidisciplinary experts and a literature review, we hypothesized that general social media use (hypothesis 1), social media literacy (hypothesis 2), interest in clinical studies (hypothesis 3), trust (hypothesis 4), privacy needs (hypothesis 5), and perceived stigma (hypothesis 6) are associated with acceptance of social media recruitment.

Methods

Study Design

This study is part of the European Union–funded international research consortium "TherVacB—A Therapeutic Vaccine to Cure Hepatitis B," work package 6 (ethical, legal, and social aspects of social media recruitment). Using a mixed methods design, we first conducted an explorative qualitative multistakeholder interview study assessing the ethical, legal, social, and practical implications of social media recruitment for clinical studies [2]. The hypotheses investigated in this paper are based on these interviews and a conceptual literature review mapping the ethical implications of social media recruitment [11]. The reporting of this study followed the Strengthening the Reporting of Observational Studies in Epidemiology guidelines [15].

Survey Recruitment

On the basis of preliminary statistical power analysis and pragmatic considerations of available study participants, we aimed for 200 responses in a recruitment period of 7 months. Due to administrative constraints, including the COVID-19 pandemic, the overall recruitment period was prolonged by 5 months (total recruitment period 12 months, June 4, 2022, to May 31, 2023), and the recruitment period varied among the recruiting clinics (Multimedia Appendix 1).

XSL•FO
RenderX

Adult, German-speaking patients diagnosed with acute or chronic viral hepatitis B were recruited from 3 large university hospitals in Germany. We chose such a venue-based recruitment methodology because it is considered one of the best options to recruit representative samples from hard-to-reach populations [16]. The clinical staff was instructed to distribute the study information leaflet to every eligible patient in the study period, explaining the implications of the study and inviting them to fill out the questionnaire. To limit recruitment bias and enhance sample representativeness, study nurses were briefed to avoid self-selected restrictions in recruitment and, if possible, to give a questionnaire to every incoming patient with hepatitis B who understood German sufficiently well. However, because of the administrative burden of the clinical staff, only 30.4% (285/939) of the estimated eligible incoming patients received the questionnaire (Multimedia Appendix 1). Because this low distribution number results from administrative burden in the clinic, we do not expect this to have a relevant impact on representativeness (refer to the Limitations subsection under the Discussion section). Completed questionnaires (207/285, 72.6% of the distributed questionnaires; Multimedia Appendix 1) were collected in the recruiting hospital and sent to the authors via mail.

Survey Construction

The dependent variable (acceptance of social media recruitment) was constructed based on the Technology Acceptance Model [17,18], involving the dimensions of perceived usefulness; perceived ease of use, intentions, and problem awareness; and proved good internal consistency (Cronbach α=0.863). Possible predictors for social media recruitment acceptance were identified based on the abovementioned hypotheses and operationalized by, if possible, existing validated questionnaires. For 3 (33%) of the 9 independent variables, we used existing validated questionnaires that were found to be of excellent reliability: the social media literacy scale (14 items, Cronbach α=0.947) [19], the Berger HIV Stigma Scale for use among patients with hepatitis C virus (6 items, Cronbach α=0.931) [20], and the Privacy Attitude Questionnaire [21]. For the latter, we included a shortened version that covered the dimensions developed in the Privacy Attitude Questionnaire but targeted it toward the hepatitis B context. From these dimensions, 2 subscales were created: secrecy of hepatitis B diagnosis (2 items, Cronbach α=0.623) and data privacy needs regarding hepatitis B diagnosis (2 items, Cronbach α=0.587).

For the remaining variables, no validated tools existed. Hence, we developed new scales for each variable of interest. As indicated by internal consistency, these were of moderate, good, or excellent reliability: general social media use (8 items, Cronbach α=0.676), hepatitis B–related social media use (6 items, Cronbach α=0.906), interest in clinical studies (2 items, Cronbach α=0.895), and trust in information sources regarding hepatitis B (11 items, Cronbach α=0.905; 2 subscales were created: trust in medical information sources—4 items, Cronbach α=0.784 and trust in nonmedical information sources, ie, traditional media, social media, other patients, poster advertisements, etc—7 items, Cronbach α=0.881). In addition to these adapted and self-developed scales, we included 4 demographic variables in the regression model (age, gender, education, and mother tongue as an indicator of migration background). A preliminary version of the questionnaire was discussed with 3 experts from the fields of infectiology and bioethics and then adapted and shortened based on their comments. We then performed cognitive pretesting [22] with 6 patients with hepatitis B, leading to minor changes. The full questionnaire is provided in Multimedia Appendix 2.

Statistical Analysis

Using SPSS (version 28.0; IBM Corp), we (1) performed descriptive analyses, (2) determined independent factors associated with participants' acceptance of social media as a recruitment tool for clinical hepatitis B studies through multiple linear regression analysis, and (3) performed additional exploratory bivariate analyses of hepatitis B–related stigma (ie, correlation, independent 2-tailed *t* test). The statistical significance level was set at $P<.05$. For multiple linear regression analysis, assumption checks were performed before the interpretation of the model (Multimedia Appendix 3). For the scale measuring the frequency of social media use, missing values were replaced by "0" (ie, "never"), assuming that participants did not tick a box, as they did not know the respective social media platform. Overall, 71.3% (139/195) of the participants completed all items, resulting in 3.66% (478/13,065) missing values and 81% (54/67) incomplete variables.

For the linear regression analysis, theoretical considerations and hypotheses derived from our previous qualitative study determined predictor selection. In addition, the sample size or predictor ratio a priori determines variable selection for regression modeling. According to Harrell [23], a fitted regression model is likely to be reliable when $p<m/10$ or $p<m/20$ (average requirement: $p<m/15$), where p is the number of predictors and m is the sample size. Applying this requirement to our sample size (N=195) and having missing data, we preliminarily limited the number of included predictors to 11. The following 11 predictors were included in the regression model: general social media use, social media literacy, hepatitis B–related social media use, interest in clinical studies, trust in medical information sources regarding hepatitis B (dichotomized to meet assumption of linearity), trust in nonmedical information sources regarding hepatitis B, secrecy of hepatitis B (dichotomized to meet assumption of linearity), data privacy needs regarding hepatitis B (dichotomized to meet assumption of linearity), perceived stigma, age, and education. Assumptions checks for regression analyses are presented in Multimedia Appendix 3.

Ethical Considerations

For study consent, participants were asked to confirm having read and understood the study information and to consent to the study participation by checking a consent box at the beginning of the questionnaire. Only questionnaires with this box checked were included in the analysis (12/207, 5.8% of the questionnaires were excluded for that reason; Multimedia Appendix 1). The ethics committees from the Technical University of Munich (12/22-S-NP), Hannover Medical School (10368_BO_K_2022), and University Clinic Leipzig (189/22-lk) approved the study.

XSL•FO
RenderX

Results

Deriving Hypotheses

Overview

After conducting an in-depth literature review on the ethical and social challenges surrounding social media recruitment for clinical studies [11], we developed 2 semistructured interview guides, one targeted at patients with hepatitis B and the other targeted at multidisciplinary experts. On the basis of interviews with 6 patients that were triangulated with findings from 30 interviews with experts, we qualitatively assessed what factors could be associated with the acceptance of social media recruitment for clinical hepatitis B studies. On the basis of these findings, we derived hypotheses to be tested quantitatively in a survey among patients with hepatitis B in Germany (Textbox 1).

Textbox 1. Hypotheses derived from qualitative interviews regarding factors potentially associated with the acceptance of social media recruitment for clinical studies among patients with hepatitis B.

Hypotheses

- Hypothesis 1: The more patients use social media for hepatitis B, the higher their acceptance of using social media as a recruitment tool for clinical hepatitis B studies.
- Hypothesis 2: Digital literacy is associated with social media acceptance.
- Hypothesis 3: The higher the general interest in clinical study participation, the higher the acceptance of social media recruitment for clinical studies.
- Hypothesis 4: The more patients trust information sources for hepatitis B, the higher their acceptance of social media recruitment.
- Hypothesis 5: The more patients value privacy, the lower their acceptance of using social media as a recruitment tool for clinical hepatitis B studies.
- Hypothesis 6: The higher the perceived stigma of patients with hepatitis B, the lower their acceptance of social media as a recruitment tool for clinical studies.

Intensity of Using Social Media in the Context of Hepatitis B

Most of the patients we talked with were rejecting the idea of being recruited for a clinical hepatitis B study via social media. However, patients who were more actively involved in their own recruitment tended to have more accepting attitudes. For example, patients who described using social media as a tool for informing themselves about potential clinical studies related to their disease were less opposed to being recruited via the same channel. One patient included search engines in their definition of social media and mentioned the following:

> *You can also advertise on Google. That is quasi/I think it's better if I [as a patient] search for a study. For example, I search for a study related to psoriasis and enter that term in Google—when the advertisement for a psoriasis study is then made so that it shows up as the first suggestion...I think that's better because in these instances I'm already searching, so I take the first step, I search for the study. And then the study, or the advertisement must be done in such a way that I can find it. So, I take the first step and then I land on the study.* [Patient 3]

Similarly, patients who joined shared interest groups, such as patient groups on Facebook, which gather people who deliberately want to share their own experiences with the disease and learn from others' experiences, were more open toward the idea of being approached and recruited within such groups.

These insights indicate that patients who were already active on social media and found it useful for their personal disease management were more open to being recruited via social media. This led us to the following hypothesis: (H1) The more patients use social media (for hepatitis B), the higher their acceptance of using social media as a recruitment tool for clinical (hepatitis B) studies.

Digital Literacy

The patients we interviewed represented a variety of levels regarding social media literacy. While some patients have had very limited contact with social media, others were very active on social media. One patient even described social media content management as part of their daily job. Another had conducted a research web-based questionnaire for which they were recruiting on the web. Analyzing the interviewees' accounts about their experience with social media, and partially their use habits, we found a scattered connection to social media recruitment acceptance: those who were considered to have higher digital literacy skills were, in some instances, likely to accept social media as a recruitment tool for clinical hepatitis B studies because they perceived other forms of recruitment as outdated:

> *I think we are living in a time that you have to use social media because if you don't use it...sending a letter or put[ting it] in the newspaper, will not help you.* [Patient 6]

On the other end of the spectrum, however, patients with very low digital literacy skills and relatedly very little reported use of social media, or digital media in general, in some instances had difficulties delimiting the concept of social media as such. Presumably, their less nuanced understanding of social media as a concept makes them less strictly opposed to being recruited for a clinical study via social media. One patient, for example, favored personal contact for study recruitment at first but then

XSL•FO
RenderX

revised their statement and reported that being helped was even more important than personal contact:

> *Yes definitely. If it was something important it would be best if we met at a clinic, or I don't know where this study is being done.... But even via Facebook or Messenger.... Yeah, actually never mind, I don't care actually.* [Patient 2]

While the interviews suggested a connection between the acceptance of social media recruitment for clinical hepatitis B studies and digital literacy, it remained unclear whether acceptance was higher with high or low digital literacy. Consequently, we formulated the nondirectional hypothesis that (H2) digital literacy is associated with social media acceptance (SMA).

Interest in Clinical Studies

Some participating patients expressed particularly high interest in participating in clinical studies about hepatitis B. One patient explained to us that they were "very, very happy to support studies" (patient 5), and another patient told us the following: "I actually want to help. So, that's why I get in" (patient 6). Patients like this, who reported an increased willingness to participate in clinical studies in general, seemed more susceptible to social media as a recruitment tool, too.

Another patient perceived it as beneficial that online recruitment made them less dependent on their physician to refer them to the study:

> *I don't know if my physician is even internet-savvy, he's a bit older. And well, then I thought, I have to see for myself because I'm not sure how competent he is with such things. What I mean is, it would be nicer if I...could google for [a clinical trial], land on a platform, search for [relevant studies], see all the information and can get in touch right away and say: "Hey, I am interested in your study. I would like to participate." Because in my case, the...specialists didn't even know that this [study] existed.... That's stupid and got me pretty upset."* [Patient 3]

None of the patients interviewed reported that they were generally against participation in clinical studies. This is likely a recruitment bias of this qualitative interview study, which made it difficult to interrogate if patients who are less accepting of clinical studies are also less accepting of social media recruitment. Yet, based on the apparent influence of this aspect in 2 (33%) of 6 patient interviews, we formulated the following hypothesis: (H3) The more patients are interested in clinical studies, the more they accept social media as a recruitment tool for clinical hepatitis B studies.

Trust

The role of trust in health care professionals, social media platforms, and other recruitment channels was a very salient aspect of all interviews. Illustrating this, one participating patient with hepatitis B stated the following as a reason for being against social media recruitment:

> *I just feel such a distrust of social media. Any information I share there, I'm not completely comfortable with/It's just not a safe way for me to share information.* [Patient 4]

Other patients were more open to social media recruitment if they knew the source of the advertisement and assigned relevant expertise to them:

> *It would be okay for me [if someone would contact me on social media to ask whether I would like to meet for a clinical study, as long as] the person is qualified in that direction and is well versed in this expertise.* [Patient 2]

> *[R]ecruiting is normally working if the person that suggests it is a person that you trust or you know. So because she was a person I knew from [redacted], then I clicked the link and I got in. Normally we know, of course, that social media is also a trap for many, I don't know, viruses and this kind of thing. So you don't open everything if you don't trust the link.... If I would see it on, I don't know, social media and as we know, because you have these cookies that you accept, then immediately, they know that you have something or you are looking for some article. Then this kind of things will pop up. Again, it's all about trusting links. I'm not sure how much I will get in something that is suggesting from just because I click on a link.* [Patient 6]

More implicitly, another patient emphasized that the clinical setting was the place for them to discuss things in the context of hepatitis B, not social media:

> *This channel through the [clinic in Germany]... I have a very good opinion of the hospital and I have always been well taken care of there. That is the only channel through which I would talk about my condition and about my/yes.* [Patient 1]

We analyze the aspect of trust in a separate publication (Willem, T, et al, unpublished data, January 2024) in detail and hypothesize the following: (H4) The more patients trust information sources, the higher their acceptance of social media recruitment. The hypothesis was operationalized for trust in medical information sources (H4a) and trust in nonmedical information sources (H4b).

Privacy

A particular concern of most patients we spoke with was their privacy. Privacy is a multifaceted and complex concept, and we found that participants referred to different dimensions of privacy: (1) *data privacy*, defined as the general attitude toward protective measures that empower patients or users to make their own decisions about who can process their data for which purpose; and (2) privacy related to the *perceived secrecy* of the hepatitis B diagnosis.

First, regarding data privacy, several patients perceived recruitment via social media as dubious and suspected some form of data leakage or malicious data collection goals behind the reach outs. This view applied irrespectively to how they would be approached on social media (eg, advertisement banners in their social media timelines or personal contact requests via social media messengers by health care professionals). For

XSL•FO
RenderX

example, a patient who reported on being in the process of decreasing their social media use to protect their privacy also said that if someone contacted them on social media regarding clinical study participation, they would "find that very strange, because [I] would ask [my]self, where did they get this information?" and reported that they would feel that this "would rob quite a lot of privacy" (patient 5). Another patient, who reported using WhatsApp as their only social media, explained that by saying that they "consider social media to be useful in some instances;" however, they continued, "It's too risky for me with my private data and so much advertising. This, for me, trumps all advantages of social media recruitment" (patient 4).

Regarding the second privacy dimension, secrecy, several patients commented on their hepatitis B diagnosis being a very private, intimate matter:

> *This condition is in my most private, intimate sphere…. And you might be right, I never thought about it in this way, but [my avoiding engaging on social media regarding hepatitis B] may be related to the fact that content I pass on via WhatsApp can be passed on thousands of times with one click.* [Patient 1]

One patient replied to a question regarding their attitude toward being contacted by a study center via social media that they "would find that difficult". As a reason, this patient explained the following:

> *[T]hat's just the problem: it ends up on social media. See, if someone writes: "Hey, I would like to ask you about your hepatitis B, whether you would participate in a study?" Then this information is out there on social media.... That's why I had a very, very good feeling when my doctor approached me about [this interview study] and that it just went through the clinic. If she had said, "Look, someone is approaching you via social media," or something, then I would have said no, right? Because I wouldn't have wanted to, because these data/social media make money because they have data. They run the ads based on your data and what you type in there or what you say or whatever. And I don't want that associated with my disease.* [Patient 5]

These findings led us to the following hypothesis: (H5) The more patients value privacy, the lower their acceptance of using social media as a recruitment tool for clinical hepatitis B studies. The hypothesis was operationalized for secrecy (H5a) and data privacy (H5b).

Perceived Stigma

Several interviewed patients with hepatitis B reported fear of being stigmatized if their social environment found out about their diagnosis as an important reason against social media recruitment. One patient, who mentioned that only their closest family members knew about their diagnosis, expressed fear that other people learning the diagnosis would lead to social exclusion:

> *A broken leg or surgery on the knee or hip. This is apparent to everyone. And everyone assumes that it will heal at some point and that there is no potential infectious danger from these people. Whereas in the case of infectious diseases, no one can assess that, and people get socially excluded very quickly.... And this is why I am so cautious with my data.* [Patient 1]

A similar view was shared by patient 5. Another patient added that perception of stigma differed depending on the context:

> *I come from [Eastern European country], I have moved to Germany. So here the mentality is a little bit different. If you say to someone, I have Hepatitis, he is okay with it. He says: "Oh, is not a problem. Normally here we are vaccinated against it." If you are going to [Eastern European country] and say: "I have Hepatitis B," it's like you have a huge disease that can just be taken by a handshake [laughs]. And so I think that's why I'm going on the conservative site.* [Patient 6]

The connection between the stigma connected to hepatitis B and the social media–connected perceived privacy risks established by several interview participants led us to the following hypothesis: (H6) The higher the perceived stigma of patients, the lower their acceptance of social media as a recruitment tool for clinical hepatitis B studies.

Survey Results

Participant Characteristics

A total number of 195 eligible questionnaires were included in the statistical analysis of the survey study. Table 1 displays the characteristics of the patients with hepatitis B who participated in the study: more than half of the participants (108/195, 55.4%) were aged between 30 and 49 years. Just above half (110/195, 56.4%) reported having lower educational degrees than Abitur (German equivalent to a high school degree). More than half of the participants (111/195, 56.9%) had another mother tongue than German (only). All participants had a chronic hepatitis B infection, as per the inclusion criterion of this study.

XSL•FO
RenderX

Table 1. Participant characteristics (N=195).

Characteristics	Participants, n (%)
Gender	
Male	101 (51.8)
Female	88 (45.1)
No answer	6 (3.1)
Age (y)	
18-29	16 (8.2)
30-39	50 (25.6)
40-49	58 (29.7)
50-59	38 (19.5)
>60	24 (12.3)
No answer	9 (4.6)
Education: high school diploma	
Yes	71 (36.4)
No	110 (56.4)
No answer	14 (7.2)
Mother tongue (multiple answers possible)	
German	101 (51.8)
Other	111 (56.9)
No answer	12 (6.2)

Description of Scales

The questionnaire included 7 scales that were measured through several items (Table 2 and Multimedia Appendices 1 and 4).

The level of acceptance for social media recruitment was measured through the SMA scale, which was calculated based on 4 questionnaire items (P6.01 to P6.04; Multimedia Appendix 4). Each item was measured by a 5-point Likert scale, ranging from 0 (completely disagree) to 4 (completely agree). Items P6.01 (“Social media are well suited to make patients aware of studies on new hepatitis B treatments”) and P6.02 (“Social media increase the likelihood of success in hepatitis B clinical trials”) formed the subscale of the perceived usefulness of social media recruitment and received moderate agreement (P6.01: mean 1.99, SD 1.23; P6.02: mean 1.81, SD 1.12). Items P6.03 and P6.04 formed the SMA subscale on the perceived usefulness of social media recruitment. Item P6.03 (“I would be recruited via social media for a hepatitis B clinical trial”) received particularly low acceptance (mean 1.13, SD 1.13; Multimedia Appendix 4). P6.04 (I would use social media to learn about hepatitis B clinical trials) received a higher mean acceptance score than P6.03 (mean 1.58, SD 1.23; Multimedia Appendix 4).

The overall SMA score was calculated by summarizing the scores from items 6.01 to 6.04 and ranged from 0 (no acceptance) to 16 (full acceptance; mean 6.48, SD 3.03; Table 2). While 28.7% (56/195) of the respondents rejected social media recruitment with an SMA score of <5, only 10.2% (20/195) of the respondents accepted social media recruitment with an SMA score of >11 (Table 3).

XSL•FO
RenderX

Table 2. Description of scales.

	Valid, n (%)	Items, n (%)	Scale, median (range[a])	Values, mean (SD)
General social media use	195 (100)	8 (15)	11 (0-32)	11.22 (6.51)
Social media literacy (hypothesis 2)	174 (89.2)	14 (25)	41 (0-56)	37.58 (14.60)
Hepatitis B–related social media use (hypothesis 1)	181 (92.8)	6 (11)	3 (0-24)	5.22 (5.61)
Interest in clinical studies (hypothesis 3)	187 (95.9)	2 (4)	6 (0-8)	5.53 (2.45)
Trust in medical information sources	180 (92.3)	4 (7)	11 (0-16)	10.27 (3.64)
Trust in nonmedical information sources (hypothesis 4)	175 (89.7)	7 (13)	8.5 (0-28)	8.36 (5,76)
Acceptance of social media recruitment (dependent variable)	178 (91.3)	4 (7)	6 (0-16)	6.48 (3.93)
Secrecy (hypothesis 5a)	185 (94.9)	2 (4)	2 (0-8)	2.25 (2.09)
Data privacy (hypothesis 5b)	186 (95.4)	2 (4)	7 (0-8)	6.25 (2.10)
Perceived stigma (hypothesis 6)	180 (92.3)	6 (11)	3.5 (0-24)	5.52 (6.02)

[a]Items were measured through a 5-point Likert scale, ranging from 0 (completely disagree) to 4 (completely agree).

Table 3. Social media acceptance among patients with hepatitis B in Germany. The higher the score, the higher the acceptance (n=195).

Social media acceptance score	Responses, n (%)
0	20 (10.3)
1	4 (2.1)
2	6 (3.1)
3	8 (4.1)
4	18 (9.2)
5	14 (7.2)
6	20 (10.3)
7	20 (10.3)
8	17 (8.7)
9	12 (6.2)
10	8 (4.1)
11	11 (5.6)
12	7 (3.6)
13	7 (3.6)
14	2 (1)
15	1 (0.5)
16	3 (1.5)
Missing	17 (8.7)

Regression Analysis

Using multiple linear regression analyses, we evaluated the predictors of participants' acceptance of social media as a recruitment tool for clinical hepatitis B studies. Testing the statistical significance of the overall model fit, the *F* test indicated that the predictors included in the model substantially contributed to the explanation of the dependent variable (Table 4). Regression analysis revealed that social media use for hepatitis B, interest in clinical studies, trust in nonmedical information sources, and hepatitis B secrecy independently predicted acceptance of social media as a recruitment tool for clinical hepatitis B studies. More precisely, the higher the social media use for hepatitis B, the higher the interest in clinical studies, the more trust in nonmedical information sources, and the less secret hepatitis B, the higher the acceptance of social media as a recruitment tool for clinical hepatitis B studies (Table 4).

XSL•FO
RenderX

Table 4. Multiple linear regression analysis[a].

	Unstandardized coefficients B (SE)	β	*t* test (*df*)	*P* value	Tolerance	VIF[b]
Constant	4.007 (1.935)	__[c]	2.071 (127)	.04	—	—
General social media use	0.060 (0.051)	.098	1.175 (127)	.24	.628	1.593
Social media literacy	–0.002 (0.025)	–.008	–0.096 (127)	.92	.600	1.668
Hepatitis B–related social media use	0.279 (0.053)	.391	5.299 (127)	<.001	.804	1.234
Interest clinical studies	0.283 (0.127)	.171	2.217 (127)	.03	.732	1.366
Trust medical information sources	–0.601 (0.683)	–.079	–0.879 (127)	.38	.546	1.830
Trust in nonmedical information sources	0.252 (0.058)	.359	4.307 (127)	<.001	.632	1.583
Secrecy	–1.299 (0.542)	–.171	–2.399 (127)	.02	.861	1.161
Data privacy	–0.765 (0.577)	–.099	–1.326 (127)	.19	.792	1.262
Perceived stigma	–0.003 (0.048)	–.004	–0.057 (127)	.95	.770	1.299
Age	–0.052 (0.028)	–.151	–1.842 (127)	.07	.648	1.543
Education	0.770 (0.567)	.102	1.357 (127)	.18	.782	1.278

[a]Overall model fit: $F_{11,127}$=9.221, P<.001; R^2=0.444; N=139.

[b]VIF: variance inflation factor.

[c]Not applicable.

Discussion

Principal Findings

We present the first empirical study investigating how adult patients with hepatitis B accept social media recruitment for clinical studies. Social media have been suggested to increase recruitment accrual, particularly for hard-to-reach populations [13,14,24]. Our study provides a more fine-grained contextualization of this potential. We find that acceptance of social media recruitment among patients with hepatitis B is associated with higher ongoing activity on social media with regard to hepatitis B (confirming H1), a generally high interest in participating in clinical studies for hepatitis B (confirming H3), and high trust recruitment channels outside the clinical setting (confirming H4a). Patients with these characteristics are, consequently, recruitable via social media under the assumptions that (1) patients are most effectively recruited via social media if they accept this channel as a recruitment method and (2) people who do not accept this recruitment channel should also not be recruited in this way.

Yet, 54 (27.7%) out of 195 participants reported an acceptance score of <5 and, thus, rejected being recruited via social media. Moreover, only 20 (10.3%) out of 195 participants reported an acceptance score >11, indicating high acceptance. These findings indicate that recruitment success via social media might be limited among patients with hepatitis B in Germany and underline the importance of using multiple recruitment channels to facilitate diversity and equitable health care access, particularly for patient groups considered vulnerable [11].

Contrary to what we had hypothesized, SMA was not associated with digital literacy (rejecting H2), data privacy needs (rejecting H5b), and perceived hepatitis B–related stigma (rejecting H6), although reported secrecy around hepatitis B diagnosis was a predictor (confirming H5a). Moreover, trust in medical information sources and demographic variables (age and education) as well as the overall frequency of using social media were not associated with SMA. The results for H2 and H4b are not surprising, as the preceding qualitative interviews did not explicitly indicate a linear connection between digital literacy and social media recruitment acceptance. Our study cannot exclude the possibility that there might be a potential nonlinear association, but another survey study also found that digital literacy did not directly affect the intention to use digital technology [25]. Furthermore, trust is a multifaceted concept [26,27], which is why the subjects of trust were split into medical information sources and other advertisement channels. Hence, it is not unexpected that trust in medical information sources is not associated with SMA.

The rejection of H5b (data privacy) was more surprising, particularly because the qualitative interviews indicated strong connections between data privacy and SMA. In addition, the scholarly debate around data privacy issues has been very salient: data ethicists have repeatedly emphasized the issues related to data privacy and transparency in the context of social media use in the research context [12,28,29]. In addition, the European General Data Protection Regulation emphasizes the transparent use of data and the rights of data subjects [30]. Moreover, various scandals (eg, related to the US presidential election in 2016 and the UK Brexit referendum) diminished users' trust in social media platforms and increased awareness of data privacy in that context [31,32]. A recent population survey conducted in Germany, the United Kingdom, and the United States confirmed high levels of concern regarding data privacy in all included countries [33]. Given these public discussions about social media activities being problematic for data privacy, it is particularly astonishing that data privacy concerns (as operationalized in our study) were not predicting

XSL•FO
RenderX

SMA. The findings align with discussions around the privacy paradox. It was confirmed in numerous studies that social media users display limited data protection behavior despite being concerned about their privacy [34-36]. In line with this, the aforementioned scandals have not resulted in a decline in Facebook users [37,38]. Other studies suggest a poor user awareness of online privacy [39] and fatigue in engaging with privacy-related risks [40]. It seems that the surveyed population with hepatitis B in Germany are also affected by this privacy paradox.

The rejection of H6 (association of stigma) was surprising, too, particularly because of the strong association between hepatitis B and stigma in other studies. An Indian survey study found that most surveyed patients with hepatitis B were subject to severe stigma and moderate to severe discrimination, with gender identification as men, unemployment, and illiteracy being predictors of discrimination [6]. Other survey studies from Australia, Turkey, and Serbia confirmed the presence of self-reported perception of stigma in 35% to 47% of patients with hepatitis B and 60% to 65% of patients with hepatitis C [10,41,42]. An Iranian qualitative study found that patients with hepatitis B conceptualized stigma as both extrinsic (eg, discrimination, public embarrassment, or blame) and intrinsic (eg, perceived rejection, social isolation, and frustration) [8]. Although this empirical evidence illustrates the relative importance of stigma in the context of hepatitis B, this did not predict patients' acceptance of social media recruitment in our study. Instead, our findings suggest that the perceived secrecy of a hepatitis B diagnosis, which seems to be unrelated to the perception of stigma, is informative on social media recruitment acceptance. This indicates that perceptions of stigma in other stigmatized diseases (eg, sexually transmitted diseases, and psychiatric disorders) might not influence patient acceptance to be recruited via social media for clinical studies. However, empirical studies within these populations need to confirm this.

Limitations and Further Research

Our survey showed a relatively balanced representation of genders. This aligns with a German serological study from 2011, which indicated no statistically significant difference in the prevalence of acute or chronic hepatitis B infection in men and women [43]. In terms of age distribution, the survey study covered a diverse range of age groups, mirroring the distribution found in the German serological study [43]. On the basis of these observations, the survey sample overall is representative of the population with hepatitis B in Germany regarding gender and age.

However, it is essential to consider potential limitations and sources of bias. The recruitment strategy used, primarily relying on venue-based recruitment within a clinical setting, might introduce selection bias, as it may not fully capture the diverse population that may exist outside such settings. In addition, only 30.4% (285/939) of estimated incoming patients received the questionnaire, which might introduce an additional selection bias. We attempted to mitigate this by explicitly briefing the study nurses to avoid self-selection when distributing the survey. The low distribution rate has been mainly caused by administrative burden, resulting in weeks during which no questionnaires were distributed. Thus, we do not expect this to have a large impact on selection bias.

In addition, the study's restriction to the German language may have impaired the accessibility of the questionnaire for participants who do not have German as their mother tongue. In addition, the exclusive focus on a German setting may limit the generalizability of the findings to a broader international context, potentially impacting the study's external validity. Finally, it is important to note that we have shortened the questionnaire in comparison to its original length after discussion with clinical colleagues, who provided the feedback that the questionnaire was too long. As part of this shortening, some validated scales were replaced by self-developed scales, which may have implications for the comprehensiveness and depth of the data collected.

Consequently, the attitudes of patients in other medical conditions toward social media recruitment, and a comparison to the attitudes of patients with hepatitis B assessed in this study, should be subject to further research. Similarly, it will be important to study how the different social media platforms, their underlying logic, use patterns, and other factors might influence patients' acceptance of social media recruitment over time.

Conclusions

This study provides the first quantitative data on the acceptance of social media as a recruitment channel for clinical studies. In the context of hepatitis B in Germany, acceptance of being recruited via social media was very limited. More than 1 (28.7%) in 4 participants rejected this recruitment channel. The study sets out to be a reference point for future studies assessing the attitudes and acceptance of social media recruitment for clinical studies. Such empirical inquiries can facilitate the work of researchers designing clinical studies as well as ethics review boards in balancing the risks and benefits of social media recruitment in a context-specific manner. Moreover, this study provides guidance for researchers considering using social media recruitment and ethics review boards judging such undertakings, by cautioning against the potentially low acceptance rates social media–based recruitment might yield for some patient populations. These should be weighed against the risks of social media recruitment for the target populations.

Similarly relevant for practice, the findings indicate that social media recruitment is particularly accepted in patient populations with high interest in participating in clinical studies. This is particularly the case for diseases with insufficient treatment options and historically neglected diseases with high unmet needs [44]. Using social media as a recruitment channel for studies targeting these patient groups might thus encounter higher acceptance levels than in this study. There was no statistically significant role associated with perceived stigma and data privacy needs among patients, suggesting that these concerns are unrelated to social media recruitment acceptance.

XSL•FO
RenderX

Acknowledgments

This study received funding from the European Union's Horizon 2020 research and innovation program (848223; TherVacB). This publication reflects only the authors' views, and the European Commission is not liable for any use that may be made of the information contained therein. The authors would like to thank all TherVacB clinical project partners who helped recruit for this study and provided feedback on the questionnaire for their kind collaboration. The authors would also like to thank all patients with hepatitis B who took the time to participate in the survey.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Response rate information.
[DOCX File , 14 KB - jmir_v26i1e54034_app1.docx]

Multimedia Appendix 2
Questionnaire.
[PDF File (Adobe PDF File), 184 KB - jmir_v26i1e54034_app2.pdf]

Multimedia Appendix 3
Assumptions checks for regression analyses.
[DOCX File , 13 KB - jmir_v26i1e54034_app3.docx]

Multimedia Appendix 4
Description of each item of the questionnaire.
[DOCX File , 29 KB - jmir_v26i1e54034_app4.docx]

References

1. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020 Oct 26;22(10):e22810 [FREE Full text] [doi: 10.2196/22810] [Medline: 33104015]
2. Goldman N, Willem T, Buyx A, Zimmermann BM. Practical benefits, challenges, and recommendations on social media recruitment: multi-stakeholder interview study. J Med Internet Res 2023 May 22;25:e44587 [FREE Full text] [doi: 10.2196/44587] [Medline: 37213177]
3. Frandsen M, Thow M, Ferguson SG. The effectiveness of social media (Facebook) compared with more traditional advertising methods for recruiting eligible participants to health research studies: a randomized, controlled clinical trial. JMIR Res Protoc 2016 Aug 10;5(3):e161 [FREE Full text] [doi: 10.2196/resprot.5747] [Medline: 27511829]
4. Wisk LE, Nelson EB, Magane KM, Weitzman ER. Clinical trial recruitment and retention of college students with type 1 diabetes via social media: an implementation case study. J Diabetes Sci Technol 2019 May;13(3):445-456 [FREE Full text] [doi: 10.1177/1932296819839503] [Medline: 31010315]
5. Guthrie KA, Caan B, Diem S, Ensrud KE, Greaves SR, Larson JC, et al. Facebook advertising for recruitment of midlife women with bothersome vaginal symptoms: a pilot study. Clin Trials 2019 Oct;16(5):476-480 [FREE Full text] [doi: 10.1177/1740774519846862] [Medline: 31055949]
6. Behera MK, Nath P, Behera SK, Padhi PK, Singh A, Singh SP. Unemployment and illiteracy are predictors of hepatitis B virus-related stigma and discrimination. J Clin Exp Hepatol 2022;12(3):767-773 [FREE Full text] [doi: 10.1016/j.jceh.2021.12.006] [Medline: 35677504]
7. Freeland C, Farrell S, Kumar P, Kamischke M, Jackson M, Bodor S, et al. Common concerns, barriers to care, and the lived experience of individuals with hepatitis B: a qualitative study. BMC Public Health 2021 May 28;21(1):1004 [FREE Full text] [doi: 10.1186/s12889-021-11093-0] [Medline: 34044808]
8. Valizadeh L, Zamanzadeh V, Bayani M, Zabihi A. The social stigma experience in patients with hepatitis B infection: a qualitative study. Gastroenterol Nurs 2017;40(2):143-150. [doi: 10.1097/SGA.0000000000000223] [Medline: 28362662]
9. Le TV, Vu TT, Mai HT, Nguyen LH, Truong NT, Hoang CL, et al. Social determinants of stigma and discrimination in Vietnamese patients with chronic hepatitis B. Int J Environ Res Public Health 2019 Jan 31;16(3):398 [FREE Full text] [doi: 10.3390/ijerph16030398] [Medline: 30708943]
10. Yozgat A, Can G, Can H, Ekmen N, Akyol T, Kasapoglu B, et al. Social stigmatization in Turkish patients with chronic hepatitis B and C. Gastroenterol Hepatol 2021 May;44(5):330-336. [doi: 10.1016/j.gastrohep.2020.08.008] [Medline: 33213938]

XSL•FO
RenderX

11. Zimmermann BM, Willem T, Bredthauer CJ, Buyx A. Ethical issues in social media recruitment for clinical studies: ethical analysis and framework. J Med Internet Res 2022 May 03;24(5):e31231 [FREE Full text] [doi: 10.2196/31231] [Medline: 35503247]
12. Caplan A, Friesen P. Health disparities and clinical trial recruitment: is there a duty to tweet? PLoS Biol 2017 Mar 01;15(3):e2002040 [FREE Full text] [doi: 10.1371/journal.pbio.2002040] [Medline: 28249024]
13. Lohse B. Facebook is an effective strategy to recruit low-income women to online nutrition education. J Nutr Educ Behav 2013;45(1):69-76. [doi: 10.1016/j.jneb.2012.06.006] [Medline: 23305805]
14. Burton-Chase AM, Parker WM, Hennig K, Sisson F, Bruzzone LL. The use of social media to recruit participants with rare conditions: lynch syndrome as an example. JMIR Res Protoc 2017 Jan 23;6(1):e12 [FREE Full text] [doi: 10.2196/resprot.6066] [Medline: 28115298]
15. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
16. Wagner J, Lee S. Sampling rare populations. In: Johnson TP, editor. Health Survey Methods. Hoboken, NJ: John Wiley & Sons; Oct 17, 2014.
17. Arndt S. Evaluierung der Akzeptanz von Fahrerassistenzsystemen: Modell zum Kaufverhalten von Endkunden. Wiesbaden, Germany: Springer VS; 2011.
18. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
19. Tandoc ECJ, Yee AZ, Ong J, Lee JC. Developing a perceived social media literacy scale: evidence from Singapore. Int J Commun 2021 Jun;15:2484-2505.
20. Saine ME, Moore TM, Szymczak JE, Bamford LP, Barg FK, Mitra N, et al. Validation of a modified Berger HIV stigma scale for use among patients with hepatitis C virus (HCV) infection. PLoS One 2020 Feb 5;15(2):e0228471 [FREE Full text] [doi: 10.1371/journal.pone.0228471] [Medline: 32023310]
21. Chignell MH, Quan-Haase A, Gwizdka J. The privacy attitudes questionnaire (PAQ): initial development and validation. Proc Hum Factors Ergon Soc Annu Meet 2003;47(11):1326-1330. [doi: 10.1177/154193120304701102]
22. Willis G. Pretesting of Health Survey Questionnaires: Cognitive Interviewing, Usability Testing, and Behavior Coding. In: Leenes TP, editor. Health Survey Methods. Hoboken, New Jersey: Wiley (Wiley handbooks in survey methodology); 2014:217-242.
23. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Cham, Switzerland: Springer; 2015.
24. King DB, O'Rourke N, DeLongis A. Social media recruitment and online data collection: a beginner's guide and best practices for accessing low-prevalence and hard-to-reach populations. Can Psychol 2014;55(4):240-249. [doi: 10.1037/a0038087]
25. Kabakus AK, Bahcekapili E, Ayaz A. The effect of digital literacy on technology acceptance: an evaluation on administrative staff in higher education. J Inf Sci 2023 Mar 15:1-12. [doi: 10.1177/01655515231160028]
26. Taddeo M. Defining Trust and E-Trust: From Old Theories to New Problems. Hershey, PA: IGI Global; 2009.
27. Taddeo M. Trusting digital technologies correctly. Minds Mach 2017 Nov 15;27(4):565-568. [doi: 10.1007/s11023-017-9450-5]
28. Gelinas L, Pierce R, Winkler S, Cohen IG, Lynch HF, Bierer BE. Using social media as a research recruitment tool: ethical issues and recommendations. Am J Bioeth 2017 Mar;17(3):3-14 [FREE Full text] [doi: 10.1080/15265161.2016.1276644] [Medline: 28207365]
29. Mühlhoff R, Willem T. Social media advertising for clinical studies: ethical and data protection implications of online targeting. Big Data Soc 2023 Feb 21;10(1):1-15. [doi: 10.1177/20539517231156127]
30. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. 2016 May 04. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 [accessed 2024-07-30]
31. Reisach U. The responsibility of social media in times of societal and political manipulation. Eur J Oper Res 2021 Jun 16;291(3):906-917 [FREE Full text] [doi: 10.1016/j.ejor.2020.09.020] [Medline: 32982027]
32. Bay M. Social media ethics: a Rawlsian approach to hypertargeting and psychometrics in political and commercial campaigns. ACM Trans Soc Comput 2018 Dec 21;1(4):1-14. [doi: 10.1145/3281450]
33. Kozyreva A, Lorenz-Spreen P, Hertwig R, Lewandowsky S, Herzog SM. Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, Great Britain, and the United States. Humanit Soc Sci Commun 2021 May 14;8(1):117. [doi: 10.1057/s41599-021-00787-w]
34. Barth S, de Jong MD. The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. Telemat Inform 2017 Nov;34(7):1038-1058. [doi: 10.1016/j.tele.2017.04.013]
35. Kokolakis S. Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. Comput Security 2017 Jan;64:122-134. [doi: 10.1016/j.cose.2015.07.002]

XSL•FO
RenderX

36. Gerber N, Gerber P, Volkamer M. Explaining the privacy paradox: a systematic review of literature investigating privacy attitude and behavior. Comput Security 2018 Aug;77:226-261. [doi: 10.1016/j.cose.2018.04.002]
37. Afriat H, Dvir-Gvirsman S, Tsuriel K, Ivan L. “This is capitalism. It is not illegal”: users’ attitudes toward institutional privacy following the Cambridge analytica scandal. Inf Soc 2021 Mar 11;37(2):115-127. [doi: 10.1080/01972243.2020.1870596]
38. Hinds J, Williams EJ, Joinson AN. “It wouldn't happen to me”: privacy concerns and perspectives following the Cambridge Analytica scandal. Int J Hum Comput Stud 2020 Nov;143:102498. [doi: 10.1016/j.ijhcs.2020.102498]
39. Trepte S, Teutsch D, Masur PK, Eicher C, Fischer M, Hennhöfer A, et al. Do people know about privacy and data protection strategies? Towards the “online privacy literacy scale” (OPLIS). In: Gutwirth S, Leenes R, de Hert P, editors. Reforming European Data Protection Law. Dordrecht, The Netherlands: Springer; Oct 04, 2014.
40. Choi H, Park J, Jung Y. The role of privacy fatigue in online privacy behavior. Comput Hum Behav 2018 Apr;81:42-51. [doi: 10.1016/j.chb.2017.12.001]
41. Drazic YN, Caltabiano ML. Chronic hepatitis B and C: exploring perceived stigma, disease information, and health-related quality of life. Nurs Health Sci 2013 Jun;15(2):172-178. [doi: 10.1111/nhs.12009] [Medline: 23171324]
42. Kostic M, Kocic B, Tiodorovic B. Stigmatization and discrimination of patients with chronic hepatitis C. Vojnosanit Pregl 2016;73(12):1116-1124. [doi: 10.2298/vsp150511135k]
43. Poethko-Müller C, Zimmermann R, Hamouda O, Faber M, Stark K, Ross RS, et al. [Epidemiology of hepatitis A, B, and C among adults in Germany: results of the German Health Interview and Examination Survey for Adults (DEGS1)]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2013 May;56(5-6):707-715. [doi: 10.1007/s00103-013-1673-x] [Medline: 23703489]
44. Durant RW, Wenzel JA, Scarinci IC, Paterniti DA, Fouad MN, Hurd TC, et al. Perspectives on barriers and facilitators to minority recruitment for clinical trials among cancer center leaders, investigators, research staff, and referring clinicians: enhancing minority participation in clinical trials (EMPaCT). Cancer 2014 Apr 01;120 Suppl 7(0 7):1097-1105 [FREE Full text] [doi: 10.1002/cncr.28574] [Medline: 24643647]

Abbreviations

SMA: social media acceptance

Edited by A Mavragani; submitted 27.10.23; peer-reviewed by D Kukadiya, WB Lee; comments to author 26.02.24; revised version received 08.03.24; accepted 03.06.24; published 26.08.24.

Please cite as:
Willem T, Zimmermann BM, Matthes N, Rost M, Buyx A
Acceptance of Social Media Recruitment for Clinical Studies Among Patients With Hepatitis B: Mixed Methods Study
J Med Internet Res 2024;26:e54034
URL: https://www.jmir.org/2024/1/e54034
doi:10.2196/54034
PMID:

©Theresa Willem, Bettina M Zimmermann, Nina Matthes, Michael Rost, Alena Buyx. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Strategies for Identifying and Recruiting Women at High Risk for Breast Cancer for Research Outside of Clinical Settings: Observational Study

Claire C Conley[1], PhD; Jennifer D Rodriguez[1], BA; McKenzie McIntyre[2], MPH; Bethany L Niell[3,4], MD, PhD; Suzanne C O'Neill[1], PhD; Susan T Vadaparampil[2], MPH, PhD

[1]Department of Oncology, Georgetown University, Washington, DC, United States
[2]Health Outcomes and Behavior Program, Moffitt Cancer Center, Tampa, FL, United States
[3]Department of Diagnostic Imaging and Interventional Radiology, Moffitt Cancer Center, Tampa, FL, United States
[4]Department of Oncologic Sciences, University of South Florida, Tampa, FL, United States

Corresponding Author:
Claire C Conley, PhD
Department of Oncology
Georgetown University
2115 Wisconsin Ave NW
Suite 300
Washington, DC, 20007
United States
Phone: 1 2026875086
Email: claire.conley@georgetown.edu

Abstract

Background: Research is needed to understand and address barriers to risk management for women at high (≥20% lifetime) risk for breast cancer, but recruiting this population for research studies is challenging.

Objective: This paper compares a variety of recruitment strategies used for a cross-sectional, observational study of high-risk women.

Methods: Eligible participants were assigned female at birth, aged 25-85 years, English-speaking, living in the United States, and at high risk for breast cancer as defined by the American College of Radiology. Individuals were excluded if they had a personal history of breast cancer, prior bilateral mastectomy, medical contraindications for magnetic resonance imaging, or were not up-to-date on screening mammography per American College of Radiology guidelines. Participants were recruited from August 2020 to January 2021 using the following mechanisms: targeted Facebook advertisements, Twitter posts, ResearchMatch (a web-based research recruitment database), community partner promotions, paper flyers, and community outreach events. Interested individuals were directed to a secure website with eligibility screening questions. Participants self-reported method of recruitment during the eligibility screening. For each recruitment strategy, we calculated the rate of eligible respondents and completed surveys, costs per eligible participant, and participant demographics.

Results: We received 1566 unique responses to the eligibility screener. Participants most often reported recruitment via Facebook advertisements (724/1566, 46%) and ResearchMatch (646/1566, 41%). Community partner promotions resulted in the highest proportion of eligible respondents (24/46, 52%), while ResearchMatch had the lowest proportion of eligible respondents (73/646, 11%). Word of mouth was the most cost-effective recruitment strategy (US $4.66 per completed survey response) and paper flyers were the least cost-effective (US $1448.13 per completed survey response). The demographic characteristics of eligible respondents varied by recruitment strategy: Twitter posts and community outreach events resulted in the highest proportion of Hispanic or Latina women (1/4, 25% and 2/6, 33%, respectively), and community partner promotions resulted in the highest proportion of non-Hispanic Black women (4/24, 17%).

Conclusions: Although recruitment strategies varied in their yield of study participants, results overall support the feasibility of identifying and recruiting women at high risk for breast cancer outside of clinical settings. Researchers must balance the associated costs and participant yield of various recruitment strategies in planning future studies focused on high-risk women.

(J Med Internet Res 2024;26:e54450) doi:10.2196/54450

XSL•FO
RenderX

KEYWORDS

breast cancer; high-risk populations; risk management; recruitment; woman; women; high risk; observational study; cross-sectional; Facebook; Twitter; flyer; flyers; community events; community event; genetic mutation

Introduction

Breast cancer is the most common non–skin cancer diagnosed among women in the United States [1]. One in 8 US women will develop breast cancer during her lifetime, equivalent to a 13% lifetime risk [1]. However, a subset of women is at high risk for developing breast cancer during their lifetime. Specifically, women who carry a pathogenic variant in genes associated with hereditary breast cancer (ie, *ATM*, *BARD1*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *NF1*, *PALB2*, *PTEN*, *RAD51C*, and *RAD51D*), who received thoracic radiation between 10 and 30 years of age, or who have an estimated lifetime breast cancer risk of ≥20% are considered to be at high risk for breast cancer [2-4].

Once identified, high-risk women may consider their options for breast cancer risk management, including risk-reducing medication (tamoxifen or raloxifene) and supplemental screening with breast magnetic resonance imaging (MRI) [2-4]. However, uptake of breast cancer risk management options among high-risk women is low. Less than 5% (161/4055, 4.0% for raloxifene; 16/4055, 0.4% for tamoxifen) of eligible women use raloxifene or tamoxifen for breast cancer risk reduction [5], and an estimated 7%-21% (158/2403, 7% observed by Miles et al [6]; 2147/10715, 20% observed by Wernli et al [7]; and 49/228, 21% observed by Ter-Minassian et al [8]) of high-risk women have received a screening breast MRI. Research is urgently needed to understand and address barriers to breast cancer risk management among high-risk women [9].

However, identifying high-risk women for research studies can be challenging. Women at high risk for breast cancer represent a small proportion of the general population; an estimated 1%-15% (5468/42,2406, 1% observed by Miles et al [6]; 25,620/12,30,363, 2% observed by Wernli et al [7]; 183/4266, 4% observed by Weisstock et al [10]; 342/5894, 6% observed by Ozanne et al [11]; 309/2881, 11% observed by Morman et al [12]; and 440/3503, 14% observed by Niell et al [13]) of women presenting for screening mammography are identified as high-risk [6,10-14]. Recruiting high-risk women in clinical settings may be facilitated by routine assessment of breast cancer risk factors [15]. Identifying high-risk women outside of clinical settings presents additional challenges. For example, women may not self-identify as “high-risk” despite the presence of factors that increase their breast cancer risk [16,17] and thus may not respond to study advertisements targeting “high-risk” women. Despite these challenges, there are several significant benefits of recruiting outside of traditional clinical settings. Clinic-based recruitment inherently requires that individuals are connected to the health care system. In contrast, recruiting outside of clinical settings offers the potential for broader reach to groups that have less contact with the health care system [18,19]. Thus, data demonstrating the most effective ways to identify and recruit high-risk women outside of clinical settings would support the recruitment of more diverse, underserved women at high risk of breast cancer in future studies targeting this population.

We successfully recruited women at high risk for breast cancer for an observational study of barriers to screening breast MRI [20,21]. This paper describes recruitment strategies employed for this study and provides metrics (eg, response rate, eligibility rate, cost per eligible participant, and eligible participant demographics) for each strategy.

Methods

Procedures and Participants

Eligible participants were assigned female at birth, aged 25-85 years, English speaking, living in the United States, and at high risk for breast cancer. For the purpose of this study, high-risk groups were defined according to criteria from the American College of Radiology (ACR) [4]. Participants were considered at high risk if they had (1) a pathogenic genetic mutation associated with increased risk for breast cancer in self or a first-degree relative; (2) a history of lobular carcinoma in situ (LCIS); (3) received thoracic radiation between 10 and 30 years of age; or (4) an estimated lifetime breast cancer risk of ≥20% per the National Cancer Institute (NCI) Breast Cancer Risk Assessment Tool (BCRAT) [22]. Although other, lengthier risk assessment models (eg, Breast Cancer Surveillance Consortium, Tyrer-Cuzick) have better predictive ability [13,23,24], we chose to use the relatively short BCRAT to estimate lifetime breast cancer risk due to concerns about participant burden.

Exclusion criteria included personal history of breast cancer, prior bilateral mastectomy, medical contraindications for MRI, and not up-to-date on screening mammography per ACR guidelines [25]. Given that our study provided an incentive to participants, we were vulnerable to fraudulent responses by ineligible individuals motivated to harvest the incentive. For this reason, consistent with best practices for detecting fraudulent responses to web-based research studies [26], we excluded individuals who made multiple attempts at completing the web-based eligibility screener. These individuals were identified by IP address, which was captured by our survey platform (Qualtrics) upon submission of the eligibility screener.

Participant recruitment took place between August 2020 and January 2021. Participants were recruited using a variety of mechanisms including targeted Facebook advertisements, Twitter posts, ResearchMatch, community partners, paper flyers, and community outreach events. Details for each recruitment mechanism are provided in the section titled “Recruitment Strategies.” All advertisements stated that the study was focused on experiences with breast cancer screening, that women aged 25-85 years may be eligible to participate, and that participants would be compensated for survey completion. An example study advertisement is shown in Figure 1. Interested individuals were directed to a secure website (Qualtrics) with eligibility screening questions. Those deemed eligible based on initial

XSL•FO
RenderX

screener responses (ie, pathogenic genetic mutation carrier or relative, history of LCIS, or history of thoracic radiation) were able to continue to the web-based survey immediately. For all other respondents, the research team used data elements from the screener to calculate estimated lifetime risk via the NCI BCRAT. Specifically, a member of the research team downloaded Qualtrics data weekly and calculated BCRAT scores for these respondents using the available SAS macro (National Institutes of Health, National Cancer Institute) [27]. Those deemed eligible based on BCRAT scores were emailed a link to the web-based survey.

Figure 1. Example study advertisement.

Recruitment Strategies

Participants self-reported method of recruitment during the eligibility screener. Response options included targeted Facebook advertisements, Twitter posts, ResearchMatch, community partners, paper flyers, and community outreach events, "word of mouth," "other," and "prefer not to answer."

Targeted Facebook Advertisements

Targeted Facebook advertisements were distributed via the Georgetown-Howard Universities Center for Clinical and Translational Science [28]. Facebook allows advertisers to "target" individuals meeting certain demographic characteristics and who express interest in specific topics. For this study, Facebook advertisements were targeted to users: located in the United States, aged 18 years and older, whose language settings were set to English, and whose Facebook "interests" included a breast cancer research or advocacy group (eg, Breast Cancer Research Foundation, Susan G. Komen) or event (eg, National Breast Cancer Awareness Month, Making Strides Against Breast Cancer).

Twitter Posts

The study advertisement was distributed via the Georgetown Lombardi Comprehensive Cancer Center Twitter account [29]. Tweets are public by default, meaning that any Twitter user can view and interact with them. Unlike advertisements distributed via Facebook, advertisements posted on Twitter were not targeted to any particular demographic. In addition, these tweets were not "promoted," meaning that we did not pay for their placement on Twitter. Rather, tweets were displayed to users who "follow" the Georgetown Lombardi Comprehensive Cancer Center Twitter account and other users who the Twitter algorithm deemed may be interested in the topic [30].

ResearchMatch

ResearchMatch [31] is a national electronic, web-based research recruitment database funded in part by the National Institutes of Health Clinical and Translational Science Award program and hosted at Vanderbilt University as an institutional review board (IRB)–approved data repository (VUMC IRB #090207). Study staff searched the ResearchMatch database for participants who were female, aged 25-85 years, and who did not report a personal history of breast cancer. Participants meeting these criteria were sent study invitation emails through the ResearchMatch web portal (Multimedia Appendix 1). Interested participants released their contact information to the study team via the ResearchMatch platform. Study staff then followed up via email with a link to the web-based eligibility screener.

Community Partners

To further expand our reach, we worked with 2 community organizations serving women at high risk for breast cancer to promote the study. Organization A is a Florida-based nonprofit organization with the goal of supporting people at risk for or living with hereditary cancer in general. Organization B is a nonprofit organization with multiple locations across the United States with the goal of supporting people living with or at increased genetic risk for breast or ovarian cancer specifically. Both organizations distributed information about the study via email and posted about it on social media (ie, Facebook).

Paper Flyers

We developed flyers describing the study goals and activities, which included contact information for study staff and a QR code directing interested individuals to the web-based eligibility screener (Multimedia Appendix 2). We mailed copies of the flyer to 279 ACR-accredited mammography clinics within our institutions' catchment areas. These mailings included a cover letter explaining the purpose of the study and requesting that

XSL•FO
RenderX

the flyer be posted in an area where potential participants might view it (ie, waiting rooms).

Community Outreach Events

NCI-designated Comprehensive Cancer Centers engage in community outreach and engagement (COE) activities, such as health fairs and other educational events. In the hopes of diversifying our study sample, we worked with COE staff to promote this study at community outreach events, which often target underserved populations (eg, under- or uninsured, racial and ethnic minority groups) within our institutional catchment areas. Specifically, COE staff at both sites brought paper copies of study flyers (described in the section "Paper Flyers") to distribute at community outreach events.

Statistical Analysis

All analyses were conducted using SPSS (version 28; IBM Corp). First, descriptive statistics were used to calculate rates of eligible respondents and completed surveys by recruitment strategy. Second, *z* tests were used to compare the proportion of eligible respondents by recruitment strategy. For independent-samples proportions tests, it has been suggested that both events and nonevents should occur at least 10 times in both samples [32]. Thus, we compared only recruitment strategies that identified at least 10 eligible participants and 10 ineligible participants. *z* tests were 2-tailed, and a Bonferroni correction was applied to adjust for multiple comparisons. As 10 *z* tests were conducted, significance was specified as $\alpha<.005$.

Finally, descriptive statistics were used to characterize the cost per eligible participant (ie, the amount spent on a recruitment strategy divided by the number of eligible participants resulting from that recruitment strategy) and participant demographic information. Cost was defined as direct costs (ie, the amount paid for access to a given recruitment mechanism), personnel costs (ie, the person-hours spent on a given recruitment mechanism multiplied by the hourly staff wage), and overhead costs (ie, fixed costs such as rent, utilities, and phone). Person-hours included time spent on setup (eg, developing flyers and advertisements, meeting with community partners, and posting advertisements on social media), contacting potential participants, screening for eligibility, and following up with eligible participants. Person-hours were estimated retrospectively at the conclusion of the study based on study records of contact with potential participants, conservatively assuming an average of 5 minutes per email contact. The hourly staff wage used to calculate personnel costs was US $23.90 per hour, based on the median wage for social science research assistants in 2021 [33]. Overhead costs were estimated based on our institutional indirect rate in 2021 (56%).

Ethical Considerations

Study procedures were reviewed by the Advarra IRB (protocol #00000971) and determined exempt from IRB oversight due to minimal risk. The Advarra IRB also provided a waiver of informed consent for this study (protocol #00000971) due to minimal risk. All data were deidentified and participants received a US $10 gift card upon survey completion.

Results

Between August 2020 and January 2021, we received 1566 unique responses to the eligibility screener (Table 1). The most commonly reported recruitment methods were Facebook advertisements (724/1566, 46%) and emails from ResearchMatch (646/1566, 41%). Recruitment via COE events (17/1566, 1%) and paper flyers (14/1566, 1%) was least commonly reported. Notably, 36 participants reported hearing about the study via "other" methods. "Other" recruitment methods reported included "email" (without any indication of the email's source), "social media" (without any indication of the specific social media platform), and "private Facebook group." Interestingly, the "private Facebook group" referred to in open-ended responses was a group for parents of children with ataxia telangiectasia (AT), a rare genetic condition. Parents of children with AT are carriers of a mutation in the *ATM* gene, which is associated with increased risk for breast cancer [34,35]. A member of this group saw our targeted Facebook advertisement and shared it with other group members. This "snowball" recruitment [36] via social media was unplanned and unexpected. In interpreting the results, it is important to note that the "other" category includes this unique source of participants.

While Facebook advertisements yielded the greatest number of eligible respondents (n=258), this represented only 36% (258/724) of respondents recruited via Facebook (Table 1). Although they were associated with fewer eligibility screener responses, the recruitment methods with the greatest proportion of eligible responses were "other" (19/36, 53% eligible), community partners (24/46, 52% eligible), and "prefer not to answer" (7/18, 39% eligible). ResearchMatch had the lowest proportion of eligible respondents, with only 11% (73/646) of respondents being eligible for the study.

Overall, participants were most commonly ineligible because they were not at high risk for breast cancer (702/1162, 60% of ineligible respondents). However, this did vary by recruitment strategy (Table 2). For example, individuals who were recruited via Facebook were most frequently ineligible due to a personal history of breast cancer. In addition, 23% (5/22) of ineligible respondents recruited via community partners and 18% (3/17) of ineligible respondents recruited via "other" methods were ineligible due to a prior bilateral mastectomy. This was much higher than respondents recruited via other methods (0%-5% of ineligible respondents; 0/30 for Twitter, 0/11 for "prefer not to answer," 0/11 for COE events, 0/11 for paper flyers, 2/573 for ResearchMatch, 7/466 for Facebook, and 1/21 for "word of mouth"). As noted in the section titled "Procedures and Participants," we excluded individuals who made multiple attempts at completing the web-based eligibility screener, as this is an indicator of fraudulent survey responses [26]. Participants who were recruited via Facebook and ResearchMatch had the highest rates of ineligibility due to multiple attempts (49/466, 11% and 25/573, 4%, respectively).

The eligibility rate for participants identified via ResearchMatch (73/646, 11%) was significantly lower than for participants identified via Facebook (258/724, 36%; $z=10.50$; $P<.001$),

XSL•FO
RenderX

community partners (24/46, 52%; z=–7.72; P<.001), "other" recruitment strategies (19/36, 53%; z=–7.09; P<.001), and word of mouth (10/31, 32%; z=–3.48; P<.001) (Table 3). No other recruitment strategies significantly differed in the proportion of respondents who were eligible for the study.

The average age for eligible respondents was 43 years, and the majority were non-Hispanic White (337/404, 83%). However, the demographic characteristics of eligible respondents varied by recruitment strategy (Table 4). For example, participants recruited via ResearchMatch tended to be older than the sample average (mean 45, SD 12.7), while those recruited via Twitter, COE events, and paper flyers tended to be younger than the sample average (mean 36, SD 7.1; mean 34, SD 10.8; and mean 39, SD 7.6, respectively). Although few eligible respondents were identified via Twitter and COE events, these strategies did identify a higher proportion of Hispanic or Latina women than other recruitment strategies (1/4, 25% and 2/6, 33% of eligible respondents were Hispanic or Latina, respectively). The strategy that identified the greatest proportion of non-Hispanic Black respondents was recruiting via community partners (4/24, 17% of eligible respondents were non-Hispanic Black).

Once identified as eligible, participants had the option of continuing on to complete the study survey. The recruitment strategies with the greatest survey completion rates were ResearchMatch (61/73, 84% of eligible respondents completed the survey), "other" (16/19, 84% of eligible respondents completed the survey), and word of mouth (8/10, 80% of eligible respondents completed the survey; Table 1). In contrast, COE events had the lowest rate of survey completion; none of the 6 eligible participants completed the survey.

Three of the recruitment strategies had direct costs: Facebook advertisements, community partners, and paper flyers (Table 5). The other recruitment strategies (ResearchMatch, Twitter, word of mouth, and COE events) incurred personnel costs only. The cost per eligible participant averaged US $105.80 (range: US $3.73 to US $482.71), and the cost per completed survey response averaged US $302.54 (range: US $4.66 to US $1448.13). Word of mouth was the most cost-effective recruitment strategy and paper flyers were the least cost-effective recruitment strategy.

Table 1. Respondent eligibility and survey completion rates by recruitment strategy (N=1566).

Recruitment strategy	All responses	Ineligible, n (% of total)	Eligible, n (% of total)	Completed, n (% of eligible)
Facebook	724	466 (64)	258 (36)	124 (48)
ResearchMatch	646	573 (90)	73 (11)	61 (84)
Community partners	46	22 (48)	24 (52)	15 (63)
Other	36	17 (47)	19 (53)	16 (84)
Twitter	34	30 (88)	4 (12)	2 (50)
Word of mouth	31	21 (68)	10 (32)	8 (80)
Prefer not to answer	18	11 (61)	7 (39)	5 (71)
COE[a] events	17	11 (65)	6 (35)	0 (0)
Paper flyers	14	11 (79)	3 (21)	1 (33)
Total	1566	1162 (74)	404 (26)	232 (57)

[a]COE: community outreach and engagement.

XSL•FO
RenderX

Table 2. Reasons for respondent ineligibility by recruitment strategy (N=1162)[a].

	Total	Facebook	Research-Match	Community partners	Other	Twitter	Word of mouth	Prefer not to answer	COE[b] events	Paper flyers
Not high risk, n (%)	702 (60)	164 (35)	474 (83)	6 (27)	7 (41)	26 (87)	8 (38)	6 (55)	4 (36)	7 (64)
History of breast cancer, n (%)	212 (18)	195 (42)	2 (<1)	6 (27)	1 (6)	0 (0)	4 (19)	3 (27)	1 (9)	0 (0)
Mammogram out of date, n (%)	90 (8)	15 (3)	65 (11)	1 (5)	2 (12)	2 (7)	2 (10)	2 (18)	1 (9)	0 (0)
Multiple attempts at eligibility screener, n (%)	82 (7)	49 (11)	25 (4)	1 (5)	0 (0)	1 (3)	3 (14)	0 (0)	1 (9)	2 (18)
Outside United States, n (%)	44 (4)	26 (6)	5 (1)	3 (14)	3 (18)	1 (3)	1 (5)	0 (0)	4 (36)	1 (9)
Prior bilateral mastectomy, n (%)	18 (2)	7 (2)	2 (<1)	5 (23)	3 (18)	0 (0)	1 (5)	0 (0)	0 (0)	0 (0)
Age <25 years or >85 years, n (%)	8 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	2 (10)	0 (0)	0 (0)	1 (9)
Male, n (%)	4 (<1)	3 (<1)	0 (0)	0 (0)	1 (6)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Medical contraindication for MRI[c], n (%)	2 (<1)	2 (<1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	1162	466	573	22	17	30	21	11	11	11

[a]Percentages indicate proportion of the column total.

[b]COE: community outreach and engagement.

[c]MRI: magnetic resonance imaging.

Table 3. Results of z tests for differences in proportion of eligible respondents by recruitment strategy.

	ResearchMatch	Community partners	Other	Word of mouth
Facebook	10.50 (*P*<.001)	–2.26 (*P*=.024)	–2.09 (*P*=.037)	0.39 (*P*=.700)
ResearchMatch	N/A[a]	–7.72 (*P*<.001)	–7.09 (*P*<.001)	–3.48 (*P*<.001)
Community partners	–7.72 (*P*<.001)	N/A[a]	–0.05 (*P*=.957)	1.73 (*P*=.084)
Other	–7.09 (*P*<.001)	–0.05 (*P*=.957)	N/A[a]	1.69 (*P*=.091)

[a]Not applicable.

XSL•FO
RenderX

Table 4. Demographic characteristics for eligible participants by recruitment strategy (N=404).

	Age (years), mean (SD)	Race and ethnicity, n (%)							
		NHW[a]	NHB[b]	Latina	Asian	AI/AN[c]	NH/PI[d]	Multiple	Other
Facebook	43 (13.6)	219 (85)	18 (7)	11 (4)	2 (1)	4 (2)	1 (<1)	2 (1)	1 (<1)
ResearchMatch	45 (12.7)	56 (77)	8 (11)	5 (7)	2 (3)	0 (0)	0 (0)	2 (3)	0 (0)
Community partners	43 (11.4)	18 (75)	4 (17)	2 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	43 (12.2)	18 (95)	0 (0)	1 (5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Twitter	36 (7.1)	3 (75)	0 (0)	1 (25)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Word of mouth	44 (9.8)	10 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Prefer not to answer	40 (13.9)	6 (86)	0 (0)	0 (0)	0 (0)	0 (0)	1 (14)	0 (0)	0 (0)
COE[e] events	34 (10.8)	4 (67)	0 (0)	2 (33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Paper flyers	39 (7.6)	3 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	43 (13.1)	337 (83)	30 (7)	22 (5)	4 (1)	4 (1)	2 (<1)	4 (1)	1 (<1)

[a]NHW: non-Hispanic White.

[b]NHB: non-Hispanic Black.

[c]AI/AN: American Indian/Alaska Native.

[d]NH/PI: Native Hawaiian/Pacific Islander.

[e]COE: community outreach and engagement.

Table 5. Cost (in US $) per eligible participant and completed survey response.

Recruitment Strategy	Cost				Cost per eligible participant	Cost per completed survey
	Direct	Personnel	Overhead	Total		
Facebook	1050.00	577.58	911.45	2539.03	9.84	20.48
ResearchMatch	0.00	2459.71	1377.44	3837.15	52.56	62.90
Community partners	2000.00	53.78	1134.50	3160.39	131.68	210.69
Twitter	0.00	87.63	49.07	136.71	34.18	68.35
Word of mouth	0.00	23.90	13.38	37.28	3.73	4.66
COE[a] events	0.00	99.58	55.77	155.35	25.89	__[b]
Paper flyers	356.68	571.61	519.84	1448.13	482.71	1448.13
Total	3406.68	3873.79	4061.45	11,314.04	Mean 105.80	Mean 302.54

[a]COE: community outreach and engagement.

[b]Unable to be calculated, denominator is 0.

Discussion

Principal Findings

This paper outlines steps taken in recruitment for a web-based, survey-based observational study of women at high risk for breast cancer. We present data on a variety of different recruitment strategies, including cost-effectiveness. Although recruitment strategies varied in their yield of study participants, our study results support the feasibility of identifying and recruiting high-risk populations outside of clinical settings.

In this study, the overall rate of survey completion among eligible volunteers was 57% (232/404). Although this is comparable with other observational studies in cancer prevention and early detection that have used broad advertising strategies (eg, 191/282, 68% completion among individuals eligible for lung cancer screening recruited via targeted Facebook advertisements [37]), it is significantly higher than previously reported recruitment rates for women at high risk for breast cancer. Padamsee et al [38] described the creation of a community-based cohort of diverse women at high risk for breast cancer. Several different recruitment strategies (including ResearchMatch and targeted Facebook advertising) yielded 3275 eligibility screener responses; of those, 22% (717/3275) were deemed to be eligible, valid, and at high risk for breast cancer. McGuinness et al [39] described strategies for recruiting

XSL•FO
RenderX

women at high risk for breast cancer for a randomized controlled trial of a web-based tool to support informed decision-making about chemoprevention. They used a combination of in-clinic and web-based recruitment and reported that only 9% (300/3459) of contacted individuals consented to study participation. The participant yield was much higher (18/54, 33%) among participants recruited online and via posted study flyers but still lower than the yield in our study. One potential explanation for the higher recruitment rate observed in this study is our differing inclusion criteria. Padamsee et al [38] enrolled women with ≥20% lifetime risk according to 1 of 3 risk prediction models (ie, Gail, Claus, and Tyrer-Cuzick). Given their focus on chemoprevention, McGuinness et al [39] sought to recruit women with a history of LCIS or elevated 5-year breast cancer risk (ie, ≥1.67%). Our inclusion of women with pathogenic genetic mutations associated with increased breast cancer risk, who are often highly motivated to participate in research [40], may have bolstered our recruitment rate.

In our study, targeted Facebook advertisements resulted in the most completed survey responses. However, Facebook advertisements were not the most efficient recruitment strategy. Advertising via community partners and "other" recruitment strategies resulted in the greatest proportions of eligible participants as well as high conversion rates to completed surveys. If we consider that participants reporting "other" recruitment strategies included members of Facebook groups for parents of children with AT, both of these recruitment strategies focused on audiences that were enriched with eligible individuals. Identifying forums where eligible individuals may view study advertisements may be one method that researchers can use to increase the efficiency of recruitment in future studies. However, it must be noted that recruiting via community partners was also relatively expensive per eligible participant and completed survey; Facebook advertisements were a cheaper option. Researchers should consider available resources—including financial resources, personnel resources, and time—when selecting strategies for recruiting high-risk women for future studies.

Furthermore, recruiting exclusively via community partners and dedicated support groups may result in samples that systematically differ from the larger population of interest. In order to view study advertisements posted by our community partners, individuals must subscribe to the organization's mailing list or "follow" them on social media. It is very likely that these individuals not only identify with the organization's mission but are also highly engaged and informed. Such individuals may also differ from the larger target population in terms of sociodemographic characteristics. Web-based recruitment for research studies can result in samples that are less racially and ethnically diverse than samples recruited using traditional methods [41]. As has been observed in prior studies, our sample was primarily non-Hispanic White. Yet web-based recruitment does not necessarily mean that resulting samples will be demographically homogenous. The subsample of high-risk women recruited on the web by McGuinness et al [39] was 29% (15/51) non-Hispanic Black and 27% (14/51) Hispanic or Latina. Padamsee et al [38] recruited high-risk women online, 35% (251/717) of whom were Black. Therefore, it is possible to identify and recruit racially and ethnically diverse high-risk women via web-based strategies. However, strategies for advertising must be carefully developed, preferably with input from members of the target group(s), and specific resources must be allocated for advertising and recruitment [42,43].

Unexpectedly, the recruitment strategy with the lowest rate of eligible participants was the National Institutes of Health–sponsored registry ResearchMatch. The proportion of eligible participants identified via ResearchMatch was significantly lower than those identified via Facebook advertisements, community partners, "other" methods, and word of mouth. The primary reason for ineligibility of participants identified via ResearchMatch was that they were not at high risk for breast cancer as defined by this study (474/573, 83% of ineligible respondents). Unlike other recruitment strategies (eg, targeted Facebook advertisements and community partners), ResearchMatch did not allow us to focus on users with specific interests; we were able to specify only the age range and medical conditions of the participants. Therefore, it is likely that we reached a more general audience with ResearchMatch than other recruitment strategies. However, ResearchMatch was associated with one of the highest rates of survey completion among eligible participants, perhaps reflecting the high level of motivation to participate in research among individuals who have chosen to register for a research participant database. In short, recruiting via ResearchMatch required our team to screen nearly 600 ineligible participants, incurring personnel costs of about US $53 per eligible participant and about US $63 per completed survey response. Again, this highlights how researchers must balance cost, effort, and potential yield of various recruitment strategies.

Taken together, these results suggest that respondents identified through different recruitment mechanisms effectively represent different subpopulations. For example, respondents who heard about the study via Facebook were most likely to have a personal history of breast cancer. In contrast, respondents who heard about the study via ResearchMatch and Twitter were most often at average risk for breast cancer. Regarding demographic characteristics, individuals recruited via COE events and community partners were more racially and ethnically diverse than respondents recruited via other methods. These patterns shed light on which recruitment methods would be most effective for reaching which groups, akin to audience segmentation approaches that are ubiquitous in marketing research and have been increasingly applied in health settings [44,45]. By identifying the unique characteristics of population segments, future research can focus recruitment resources on specific subpopulations of interest.

The results presented here must be interpreted cautiously in light of limitations. First, we did not track the number of personnel hours associated with each recruitment strategy during the recruitment period but estimated them retrospectively at the end of the study. Thus, our estimate may over- or underrepresent the number of person-hours associated with using various recruitment platforms. In future studies, personnel time should be tracked during the recruitment period in order to accurately estimate the costs associated with recruitment. Relatedly, we estimated overhead costs based on our institutional indirect rate.

XSL•FO
RenderX

This is an approximate cost of conducting this study in our setting, which is fairly generalizable to other academic and nonprofit settings, but may not be generalizable to all business contexts. Second, although the study team reviewed each survey response for validity, it is possible that some of the survey responses were fraudulent [26]. This is particularly common in studies that provide compensation for survey responses, such as this one. Future studies might incorporate additional methods to ensure data integrity, such as attention checks or CAPTCHA algorithms [26,46,47]. Third, given that recruitment occurred outside of clinical settings, breast cancer risk factors were assessed via self-report. Our assessment of eligibility thus may be subject to reporting biases. Relatedly, lifetime risk—a key eligibility criterion—was estimated using the relatively short NCI BCRAT due to concerns about participant burden [22]. Other, more in-depth risk assessment models have better predictive ability [13,23,24] and might be used in future research. Fourth, results cannot be generalized to breast cancer survivors, who may be at high risk for breast cancer, but were excluded from this study. Fifth, data were collected during the COVID-19 pandemic, which significantly affected recruitment of participants for research studies. For example, our community partners adjusted their activities and priorities in response to the emergent pandemic-related needs. Similarly, our institution was restricted in the types and frequency of community outreach events that could be hosted. We also attempted to recruit participants via paper flyers in mammography clinic settings, but receipt of cancer screening significantly declined during the early months of the COVID-19 pandemic [48,49]. While rates of screening mammography rebounded by mid-2021 [50], it is still unclear whether the results presented here would generalize outside of the acute pandemic timeframe, and replication of this study may be needed. Finally, conclusions about the effectiveness of recruitment strategies were drawn based on proportion of eligible respondents and the number of participants who ultimately completed the study survey; we did not capture individuals' perspectives on advertising materials or strategies, nor the proportion of responders versus nonresponders among individuals who viewed the advertisements. Future studies are needed to more thoroughly evaluate barriers and facilitators to recruitment of high-risk women for research studies.

Conclusions

Using a variety of web-based and in-person methods, we successfully identified and recruited women at high risk for breast cancer outside of clinical settings, supporting the feasibility of recruiting and retaining this unique population for participation in behavioral research. However, additional research is needed to identify best practices for recruiting a more demographically diverse group of high-risk women. Although our study focused on women at high risk for breast cancer, results may also provide insight into identification and recruitment of other high-risk populations eligible for risk-based cancer screening (eg, lung cancer screening). Researchers seeking to recruit individuals at high risk for cancer may choose from a variety of recruitment strategies but must balance the associated costs and participant yield.

Acknowledgments

This work was supported by grants from the Breast Cancer Research Foundation (ASPO-19-002, principal investigator [PI]: CCC) and the National Cancer Institute (K08CA270402, PI: CCC). This work was also supported by the Participant Research, Interventions, and Measurement (PRISM) Core at the H. Lee Moffitt Cancer Center & Research Institute, a comprehensive cancer center designated by the National Cancer Institute and funded in part by Moffitt's Cancer Center Support grant (P30CA076292, PI: Cleveland).

Data Availability

Data are publicly available in the Open Science Framework [51].

Authors' Contributions

CCC participated in the conceptualization, funding acquisition, formal analysis, visualization, interpretation, and writing—original draft. JDR and MM participated in the investigation, data curation, project administration, and writing—review and editing. BLN participated in the conceptualization, resources, supervision, and writing—review and editing. SCO participated in the conceptualization, supervision, and writing—review and editing. STV participated in the conceptualization, funding acquisition, resources, supervision, and writing—review and editing.

Conflicts of Interest

CCC has received research funding from Pfizer. BLN has a research equipment loan agreement with Hologic Inc. and has received reimbursement for business-related travel from the National Comprehensive Cancer Network. SCO has received research funding from Pfizer and Gilead Sciences, Inc. No other authors have conflicts of interest to disclose.

Multimedia Appendix 1
ResearchMatch recruitment email.
[PDF File (Adobe PDF File), 92 KB - jmir_v26i1e54450_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Example study recruitment flyer.
[PDF File (Adobe PDF File), 332 KB - jmir_v26i1e54450_app2.pdf]

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73(1):17-48 [FREE Full text] [doi: 10.3322/caac.21763] [Medline: 36633525]
2. Saslow D, Boetes C, Burke W, Harms S, Leach MO, Lehman CD, American Cancer Society Breast Cancer Advisory Group. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. CA Cancer J Clin 2007;57(2):75-89 [FREE Full text] [doi: 10.3322/canjclin.57.2.75] [Medline: 17392385]
3. National Comprehensive Cancer Network (NCCN). Genetic/familial high-risk assessment: breast, ovarian, and pancreatic. 2023. URL: https://www.nccn.org/professionals/physician_gls/pdf/genetics_bop.pdf [accessed 2024-02-12]
4. Monticciolo DL, Newell MS, Moy L, Lee CS, Destounis SV. Breast cancer screening for women at higher-than-average risk: updated recommendations from the ACR. J Am Coll Radiol 2023;20(9):902-914. [doi: 10.1016/j.jacr.2023.04.002] [Medline: 37150275]
5. Pinsky PF, Miller E, Heckman-Stoddard B, Minasian L. Use of raloxifene and tamoxifen by breast cancer risk level in a medicare-eligible cohort. Am J Obstet Gynecol 2018;218(6):606.e1-606.e9 [FREE Full text] [doi: 10.1016/j.ajog.2018.03.031] [Medline: 29630889]
6. Miles R, Wan F, Onega TL, Lenderink-Carpenter A, O'Meara ES, Zhu W, et al. Underutilization of supplemental magnetic resonance imaging screening among patients at high breast cancer risk. J Womens Health (Larchmt) 2018;27(6):748-754 [FREE Full text] [doi: 10.1089/jwh.2017.6623] [Medline: 29341851]
7. Wernli KJ, Callaway KA, Henderson LM, Kerlikowske K, Lee JM, Ross-Degnan D, et al. Trends in screening breast magnetic resonance imaging use among US women, 2006 to 2016. Cancer 2020;126(24):5293-5302 [FREE Full text] [doi: 10.1002/cncr.33140] [Medline: 32985694]
8. Ter-Minassian M, Schaeffer ML, Jefferson CR, Shapiro SC, Suwannarat P, Visvanathan K. Screening and preventative strategies for patients at high risk for breast cancer. JCO Oncol Pract 2021;17(4):e575-e581. [doi: 10.1200/OP.20.00262] [Medline: 33428469]
9. Padamsee TJ, Wills CE, Yee LD, Paskett ED. Decision making for breast cancer prevention among women at elevated risk. Breast Cancer Res 2017;19(1):34 [FREE Full text] [doi: 10.1186/s13058-017-0826-5] [Medline: 28340626]
10. Weisstock CR, Rajapakshe R, Bitgood C, McAvoy S, Gordon PB, Coldman AJ, et al. Assessing the breast cancer risk distribution for women undergoing screening in British Columbia. Cancer Prev Res (Phila) 2013;6(10):1084-1092. [doi: 10.1158/1940-6207.CAPR-13-0027] [Medline: 23963801]
11. Ozanne EM, Drohan B, Bosinoff P, Semine A, Jellinek M, Cronin C, et al. Which risk model to use? Clinical implications of the ACS MRI screening guidelines. Cancer Epidemiol Biomarkers Prev 2013;22(1):146-149. [doi: 10.1158/1055-9965.EPI-12-0570] [Medline: 23093547]
12. Morman NA, Byrne L, Collins C, Reynolds K, Bell JG. Breast cancer risk assessment at the time of screening mammography: perceptions and clinical management outcomes for women at high risk. J Genet Couns 2017;26(4):776-784. [doi: 10.1007/s10897-016-0050-y] [Medline: 28124179]
13. Niell BL, Augusto B, McIntyre M, Conley CC, Gerke T, Roetzheim R, et al. Distribution of estimated lifetime breast cancer risk among women undergoing screening mammography. AJR Am J Roentgenol 2021;217(1):48-55 [FREE Full text] [doi: 10.2214/AJR.20.23333] [Medline: 33978450]
14. Wernli KJ, DeMartini WB, Ichikawa L, Lehman CD, Onega T, Kerlikowske K, Breast Cancer Surveillance Consortium. Patterns of breast magnetic resonance imaging use in community practice. JAMA Intern Med 2014;174(1):125-132 [FREE Full text] [doi: 10.1001/jamainternmed.2013.11963] [Medline: 24247555]
15. Conley CC, Niell BL, Augusto BM, McIntyre M, Roetzheim R, Funaro K, et al. Uptake of risk appropriate behaviors after breast cancer risk stratification in the mammography screening population. J Am Coll Radiol 2020;17(10):1285-1288 [FREE Full text] [doi: 10.1016/j.jacr.2020.02.019] [Medline: 32229274]
16. Haas JS, Kaplan CP, Des Jarlais G, Gildengoin V, Pérez-Stable EJ, Kerlikowske K. Perceived risk of breast cancer among women at average and increased risk. J Womens Health (Larchmt) 2005;14(9):845-851. [doi: 10.1089/jwh.2005.14.845] [Medline: 16313212]
17. Fehniger J, Livaudais-Toman J, Karliner L, Kerlikowske K, Tice JA, Quinn J, et al. Perceived versus objective breast cancer risk in diverse women. J Womens Health (Larchmt) 2014;23(5):420-427 [FREE Full text] [doi: 10.1089/jwh.2013.4516] [Medline: 24372085]
18. Benedict C, Hahn AL, Diefenbach MA, Ford JS. Recruitment via social media: advantages and potential biases. Digit Health 2019;5:2055207619867223 [FREE Full text] [doi: 10.1177/2055207619867223] [Medline: 31431827]
19. Wozney L, Turner K, Rose-Davis B, McGrath PJ. Facebook ads to the rescue? Recruiting a hard to reach population into an Internet-based behavioral health intervention trial. Internet Interv 2019;17:100246 [FREE Full text] [doi: 10.1016/j.invent.2019.100246] [Medline: 31080751]

XSL•FO
RenderX

20. Conley CC, Rodriguez JD, Brownstein NC, O'Neill SC, Vadaparampil ST. Characteristics associated with healthcare disruptions during the COVID-19 pandemic for women in the United States at high risk for breast cancer. Prev Med Rep 2022;30:101975 [FREE Full text] [doi: 10.1016/j.pmedr.2022.101975] [Medline: 36090472]
21. Conley CC, Rodriguez JD, McIntyre M, Brownstein NC, Niell BL, O'Neill SC, et al. Self-reported barriers to screening breast MRI among women at high risk for breast cancer. Breast Cancer Res Treat 2023;202(2):345-355. [doi: 10.1007/s10549-023-07085-w] [Medline: 37640965]
22. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91(18):1541-1548. [doi: 10.1093/jnci/91.18.1541] [Medline: 10491430]
23. Terry MB, Liao Y, Whittemore AS, Leoce N, Buchsbaum R, Zeinomar N, et al. 10-year performance of four models of breast cancer risk: a validation study. Lancet Oncol 2019;20(4):504-517. [doi: 10.1016/S1470-2045(18)30902-1] [Medline: 30799262]
24. Meadows RJ, Figueroa W, Shane-Carson KP, Padamsee TJ. Predicting breast cancer risk in a racially diverse, community-based sample of potentially high-risk women. Cancer Med 2022;11(21):4043-4052 [FREE Full text] [doi: 10.1002/cam4.4721] [Medline: 35388639]
25. Monticciolo DL, Malak SF, Friedewald SM, Eby PR, Newell MS, Moy L, et al. Breast cancer screening recommendations inclusive of all women at average risk: update from the ACR and society of breast imaging. J Am Coll Radiol 2021;18(9):1280-1288. [doi: 10.1016/j.jacr.2021.04.021] [Medline: 34154984]
26. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133 [FREE Full text] [doi: 10.1111/jlme.12200] [Medline: 25846043]
27. National Cancer Institute. Breast Cancer Risk Assessment SAS Macro (Version 4, Gail Model). Bethesda, MD: Division of Cancer Epidemiology & Genetics, National Cancer Institute; 2018.
28. Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) Facebook page. URL: https://www.facebook.com/GHUCCTSCommunity [accessed 2023-09-16]
29. Georgetown Lombardi Comprehensive Cancer Center (GLCCC) Twitter profile. URL: https://x.com/LombardiCancer [accessed 2023-09-16]
30. Twitter. Twitter's recommendation algorithm. 2023. URL: https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm [accessed 2023-07-06]
31. About us. ResearchMatch. URL: https://www.researchmatch.org/about/ [accessed 2023-09-16]
32. Agresti A, Franklin C, Klingenberg B. Statistics: The Art and Science of Learning From Data. Harlow, England: Pearson Education; 2021.
33. 19-4061 Social science research assistants. US Bureau of Labor Statistics. 2022. URL: https://www.bls.gov/oes/2021/may/oes194061.htm [accessed 2024-08-08]
34. Moslemi M, Moradi Y, Dehghanbanadaki H, Afkhami H, Khaledi M, Sedighimehr N, et al. The association between ATM variants and risk of breast cancer: a systematic review and meta-analysis. BMC Cancer 2021;21(1):27 [FREE Full text] [doi: 10.1186/s12885-020-07749-6] [Medline: 33402103]
35. Hall MJ, Bernhisel R, Hughes E, Larson K, Rosenthal ET, Singh NA, et al. Germline pathogenic variants in the ataxia telangiectasia mutated (ATM) gene are associated with high and moderate risks for multiple cancers. Cancer Prev Res (Phila) 2021;14(4):433-440 [FREE Full text] [doi: 10.1158/1940-6207.CAPR-20-0448] [Medline: 33509806]
36. Naderifar M, Goli H, Ghaljaie F. Snowball sampling: a purposeful method of sampling in qualitative research. Strides Dev Med Educ 2017;14(3). [doi: 10.5812/sdme.67670]
37. Walsh LE, Carter-Bawa L. Using social media to recruit individuals for health-related research: feasibility and lessons learned. J Health Psychol 2023;28(7):599-606 [FREE Full text] [doi: 10.1177/13591053221136928] [Medline: 36416197]
38. Padamsee TJ, Bijou C, Swinehart-Hord P, Hils M, Muraveva A, Meadows RJ, et al. Risk-management decision-making data from a community-based sample of racially diverse women at high risk of breast cancer: rationale, methods, and sample characteristics of the Daughter Sister Mother Project survey. Breast Cancer Res 2024;26(1):8 [FREE Full text] [doi: 10.1186/s13058-023-01753-x] [Medline: 38212792]
39. McGuinness JE, Bhatkhande G, Amenta J, Silverman T, Mata J, Guzman A, et al. Strategies to identify and recruit women at high risk for breast cancer to a randomized controlled trial of web-based decision support tools. Cancer Prev Res (Phila) 2022;15(6):399-406 [FREE Full text] [doi: 10.1158/1940-6207.CAPR-21-0593] [Medline: 35412592]
40. Hurley RM, Suman V, Daly M, Mandrekar S. Assessment of interest for breast cancer prevention trial participation among BRCA mutation carriers. Hereditary Genet 2014;3(127):2161. [doi: 10.4172/2161-1041.1000127]
41. Topolovec-Vranic J, Natarajan K. The use of social media in recruitment for medical research studies: a scoping review. J Med Internet Res 2016;18(11):e286 [FREE Full text] [doi: 10.2196/jmir.5698] [Medline: 27821383]
42. Pii KH, Schou LH, Piil K, Jarden M. Current trends in patient and public involvement in cancer research: a systematic review. Health Expect 2019;22(1):3-20 [FREE Full text] [doi: 10.1111/hex.12841] [Medline: 30378234]

XSL•FO
RenderX

43. Harrison JD, Auerbach AD, Anderson W, Fagan M, Carnie M, Hanson C, et al. Patient stakeholder engagement in research: a narrative review to describe foundational principles and best practice activities. Health Expect 2019;22(3):307-316 [FREE Full text] [doi: 10.1111/hex.12873] [Medline: 30761699]
44. Slater MD. Theory and method in health audience segmentation. J Health Commun 1996;1(3):267-283. [doi: 10.1080/108107396128059] [Medline: 10947364]
45. Forthofer MS, Bryant CA. Using audience-segmentation techniques to tailor health behavior change strategies. Am J Health Behav 2000;24(1):36-43. [doi: 10.5993/ajhb.24.1.6]
46. Fiedler K, McCaughey L, Prager J. Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. Perspect Psychol Sci 2021;16(4):816-826 [FREE Full text] [doi: 10.1177/1745691620970602] [Medline: 33440127]
47. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. Quant Methods Psychol 2020;16(5):472-481. [doi: 10.20982/tqmp.16.5.p472]
48. Miller MM, Meneveau MO, Rochman CM, Schroen AT, Lattimore CM, Gaspard PA, et al. Impact of the COVID-19 pandemic on breast cancer screening volumes and patient screening behaviors. Breast Cancer Res Treat 2021;189(1):237-246 [FREE Full text] [doi: 10.1007/s10549-021-06252-1] [Medline: 34032985]
49. Smetherman DH. Breast cancer screening and the COVID-19 pandemic. J Breast Imaging 2021;3(1):3-11 [FREE Full text] [doi: 10.1093/jbi/wbaa106] [Medline: 38424843]
50. McBain RK, Cantor JH, Jena AB, Pera MF, Bravata DM, Whaley CM. Decline and rebound in routine cancer screening rates during the COVID-19 pandemic. J Gen Intern Med 2021;36(6):1829-1831 [FREE Full text] [doi: 10.1007/s11606-021-06660-5] [Medline: 33742300]
51. Conley C, Open Science Framework. Barriers to screening breast MRI utilization among patients at elevated risk (B-SUPER). 2022. URL: https://doi.org/10.17605/OSF.IO/XPZ35 [accessed 2024-08-12]

Abbreviations

ACR: American College of Radiology
AT: ataxia telangiectasia
BCRAT: Breast Cancer Risk Assessment Tool
COE: community outreach and engagement
IRB: institutional review board
LCIS: lobular carcinoma in situ
MRI: magnetic resonance imaging
NCI: National Cancer Institute

Edited by T de Azevedo Cardoso; submitted 09.11.23; peer-reviewed by J McGuinness, T Padamsee; comments to author 28.02.24; revised version received 22.04.24; accepted 17.07.24; published 02.09.24.

Please cite as:
Conley CC, Rodriguez JD, McIntyre M, Niell BL, O'Neill SC, Vadaparampil ST
Strategies for Identifying and Recruiting Women at High Risk for Breast Cancer for Research Outside of Clinical Settings: Observational Study
J Med Internet Res 2024;26:e54450
URL: https://www.jmir.org/2024/1/e54450
doi:10.2196/54450
PMID:39222344

©Claire C Conley, Jennifer D Rodriguez, McKenzie McIntyre, Bethany L Niell, Suzanne C O'Neill, Susan T Vadaparampil. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Challenges and Approaches to Recruitment for and Retention in a Dyad-Focused eHealth Intervention During COVID-19: Randomized Controlled Trial

Chunxuan Ma[1], MSc; Rachel H Adler[1], PhD; Daria B Neidre[1], PhD; Ronald C Chen[2], MD, MPH; Laurel L Northouse[3], PhD; Christine Rini[4], PhD; Xianming Tan[5], PhD; Lixin Song[6], PhD

[1]School of Nursing, University of Texas Health Science Center, San Antonio, TX, United States

[2]University of Kansas Medical Center, Kansas City, KS, United States

[3]School of Nursing, University of Michigan, Ann Arbor, MI, United States

[4]Department of Medical Social Sciences, Robert H Lurie Comprehensive Cancer Center of Northwestern University, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[5]Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[6]School of Nursing, Mays Cancer Center, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

Corresponding Author:
Lixin Song, PhD
School of Nursing
Mays Cancer Center
University of Texas Health Science Center at San Antonio
7703 Floyd Curl Dr
San Antonio, TX, 78229
United States
Phone: 1 210 450 8561
Email: songl2@uthscsa.edu

Abstract

Background: Family-based randomized controlled trials (RCTs) encounter recruitment and retention challenges. Cancer-focused RCTs typically recruit convenience samples from local cancer centers and hospitals.

Objective: This study aimed to examine the recruitment and retention of a population-based, patient-partner dyad cohort in an RCT testing a dyadic eHealth intervention to improve the quality of life in patients with prostate cancer and their partners.

Methods: In this 2-arm, parallel-group RCT, men who recently completed treatment for localized prostate cancer statewide were recruited through North Carolina Central Cancer Registry rapid case ascertainment between April 2018 and April 2021, coinciding with the COVID-19 pandemic. Patient-partner dyads underwent baseline assessments and were randomly assigned to either the intervention or control groups. Follow-up surveys were conducted at 4, 8, and 12 months after baseline. Descriptive and logistic regression analyses were used to achieve the study's aims.

Results: Of the 3078 patients referred from rapid case ascertainment, 2899 were screened. A total of 357 partners were approached after obtaining the eligible patients' permission, 280 dyads completed baseline assessments and were randomized (dyad enrollment rate: 85.11%, 95% CI 81.3%-88.9%), and 221 dyads completed the 12-month follow-up (retention rate: 78.93%, 95% CI 74.2%-83.7%). Regarding the factors associated with retention, compared with White participants, people self-reporting as "other races" (including American Indian, Asian, and multiracial) were more likely to drop out of the study (odds ratio 2.78, 95% CI 1.10-7.04), and older participants were less likely to withdraw (odds ratio 0.96, 95% CI 0.92-0.99).

Conclusions: Despite the negative impact of the pandemic, we successfully recruited enough patient-partner dyads to test our RCT hypotheses. Our recruitment and retention rates were equivalent to or higher than those in most dyadic intervention studies. A well-functioning research team and specific strategies (eg, eHealth intervention, internet phone, and online surveys) facilitated the recruitment and retention of patients with prostate cancer and their partners during the unprecedented pandemic.

Trial Registration: ClinicalTrials.gov NCT03489057; https://clinicaltrials.gov/study/NCT03489057

International Registered Report Identifier (IRRID): RR2-https://doi.org/10.1186/s13063-021-05948-5

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51877) doi:10.2196/51877

KEYWORDS

randomized controlled trials; RCT; prostate cancer; accrual; retention; COVID-19 pandemic; family-based research

Introduction

Participant recruitment and retention for cancer survivorship research have always been innately challenging for family-focused research, especially for randomized controlled trials (RCTs), due to the geographical distance between participants and the trial location [1]; patients' poor functional status, high symptom burden, and health deterioration [2,3]; and caregiver unavailability [4]. Most studies have used convenience samples of patients with cancer and caregivers recruited from one or a few cancer centers, hospitals, and community settings, leading to small sample sizes and low generalizability of study results [4].

Social isolation during the COVID-19 pandemic compounded the difficulties with family-based intervention RCTs, especially the challenges of participant recruitment and retention [5-7]. Long-term "shelter in place" orders negated in-person strategies and forced researchers and participants to use online or telehealth options to continue research studies [8,9]. More importantly, for older adults and patients who are immunocompromised, like many patients with cancer and chronic illnesses, isolation was imperative to protect their health. Supportive care for patients with cancer also changed with the pandemic, from in-person meetings and interactions to becoming partially or solely reliant on online resources [10,11]. These issues have brought critical challenges to the recruitment and retention of adult patients with cancer and their partners in family-based RCTs.

We conducted a dyadic RCT (ClinicalTrials.gov NCT03489057, principal investigator LS) that innovatively used North Carolina Central Cancer Registry (NCCCR) rapid case ascertainment (RCA) to recruit a population-based group of patients and their partners from across the state. In this RCT, we test the efficacy of a couple-focused eHealth intervention, titled the Prostate Cancer Education and Resources for Couples (PERC) program, which provides online education and skills training, an online forum for professional and peer support, and local and national resources for symptom management after cancer treatment for patients with prostate cancer and their intimate partners. However, the RCT timeline coincided with the COVID-19 pandemic, which introduced significant barriers to cancer diagnosis, treatment, and care [12]. Patients with prostate cancer are usually older adults (median age 66 years) with multiple comorbid conditions. They are at an increased risk for COVID-19 infection and prone to the adverse effects of the COVID-19 pandemic–related health problems. These characteristics complicated the already challenging recruitment and retention of patients with prostate cancer and their partners. This study examines the challenges, strategies, and outcomes of the RCT recruitment and retention of patients with prostate cancer and their intimate partners during the pandemic.

Methods

Study Design

We used a 2-arm, randomized, repeated measures, longitudinal design to test the efficacy of the eHealth PERC intervention among survivors with localized prostate cancer and their intimate partners [13]. We randomly assigned the patient-partner dyads who completed baseline (T1) surveys to either the intervention group (PERC) or the control group (National Cancer Institute [NCI] prostate cancer website). Dyads then completed follow-up surveys at 4 (T2), 8 (T3), and 12 (T4) months after baseline. All research activities were conducted virtually, through telephone and email, and the PERC intervention itself was a web-based program (see our protocol for details) to reduce the burden for patients, their partners, and the research team in this family-based RCT [13]. After the onset of the pandemic, we started to use a HIPAA (Health Insurance Portability and Accountability Act)–compliant, cloud-based phone system to contact study participants remotely. We also changed in-person weekly team meetings to Zoom (Zoom Video Communications) virtual meetings to continue our discussion of program-related issues and develop appropriate strategies to enhance recruitment, retention, and intervention fidelity.

Participants

Patients were eligible if (1) they were English-speaking men aged 40-75 years who were within 15 weeks of completing treatment with curative intent for their localized prostate cancer (ie, surgery, radiation therapy with or without hormonal therapy); (2) they did not have another cancer diagnosis within the previous 2 years; and (3) they were not receiving current treatment for another form of cancer. We only included the eligible patients if they had an intimate partner willing to participate in the study. Intimate partners were recruited if they were English speakers over 18 years old, had no previous diagnosis of cancer themselves, or had not received cancer treatment within the last year. Patients and their partners were excluded from the study if either exhibited severe cognitive impairment.

Recruitment Process

We recruited patient-partner dyads through the NCCCR RCA [14]. RCA accelerates the regular reporting timeline for the NCCCR; it allows researchers to identify incident patients with cancer within 1 week of diagnosis. This RCT targeted the 36 counties with the highest proportions of ethnic populations and people living in poverty to maximize the study's ability to recruit a diverse cohort. RCA staff screened the electronic medical records at cancer centers and hospitals in these counties and identified patients who were 40-75 years old and had been diagnosed with nonmetastatic prostate cancer.

Institutional review board (IRB) approval was obtained from the University of North Carolina-Chapel Hill before any research

XSL•FO
RenderX

activities. Following the principal investigator's (LS) relocation, we obtained IRB approval from the University of Texas Health Center at San Antonio for the analysis of deidentified data from the RCT. The NCCCR sent consent information to the potentially eligible patients and informed them about the PERC RCT. RCA staff then extracted the cancer-related data (diagnosis, time of diagnosis, stage, and biopsy results) and contact information of patients and their health care providers. RCA staff uploaded these data to an encrypted, password-protected data transfer hub for the research team to download biweekly; the trained research assistants downloaded and transferred the data to REDCap (Research Electronic Data Capture; Vanderbilt University), a password-protected, secure, encrypted, HIPAA-compliant database for data entry and management [15]. Within a total of 6 weeks after RCA staff sent the patients' NCCCR information, the research assistants mailed study information packages (including a descriptive introductory letter, brochure, and informed consent information) to each patient and his physician, giving each group up to 3 weeks to opt out of the RCT before they screened the prospective patients and partners. The project coordinator monitored recruitment, and the research assistants tracked recruitment and other research activities using REDCap.

Enrollment Process

Using the university-designated phones, team email, and their own emails, the research assistants assessed the eligibility of patients and partners using screening surveys and answered any questions; patients and partners who met the inclusion criteria provided consent separately. The informed consent was completed through telephone; audio recorded; and saved in an encrypted, password-protected database separate from the REDCap surveys. The research assistants then scheduled the patients and partners to complete their baseline (T1) survey separately. Automated reminder emails and voice mails were sent to consented dyads on days 3 and 5 if they still needed to answer T1 survey phone calls. On day 14, a research assistant sent a personal follow-up email. After patients and partners completed the informed consent and T1 survey, the dyads were randomly assigned to the PERC or NCI control group (1:1). These administrative and survey data were entered into the REDCap separately. The survey data were deidentified using an assigned unique ID number linking to the identifiable tracking administrative database, which only authorized study personnel could access.

Administering the surveys by telephone was initially the only planned remote data collection option. We added the online survey options after the onset of the pandemic in response to requests of the participants for more flexibility in completing the surveys due to increased demands at work and home. The biostatistician assistant of the study helped research assistants generate personalized URLs using the REDCap online survey panel; research assistants emailed the URL to the enrolled participants. Enrolled patients and partners could choose to complete either a phone or online survey throughout the remainder of the study.

Retention Process

Patients and their partners separately completed the follow-up surveys after dyads completed the PERC or use of the NCI prostate cancer website at 4, 8, and 12 months after baseline (T2, T3, and T4, respectively). Two weeks before a follow-up survey was due, the study staff sent a reminder letter to dyads' homes with a survey answer key example; the study staff then called the participants up to 12 times during a 4-week period, rotating morning, afternoon, and evening during weekday and weekends, to schedule survey completion through the phone or online. If patients or partners preferred to complete the online survey, the study staff sent the link in 3 automated emails on days 0, 3, and 6 (at T2, T3, and T4). We considered the dyads who did not respond to any previous contact attempts by day 60 as "lost to follow up" and discontinued contact. To ensure privacy and confidentiality, all phone calls were behind closed doors in private rooms. Research assistants and coordinators used group emails and their personal secured university emails to contact participants.

We offered each patient and partner gift cards of US $20, US $30, US $30, and US $50 for completing the surveys at T1, T2, T3, and T4, respectively, with up to US $130 in financial incentives for completing the entire study. In addition, we sent each participant retention gifts, that is, a fanny and a duffel bag at 6 and 10 months, respectively. We also sent postcards to each dyad over the holidays (eg, Thanksgiving, Christmas, and New Year's Day).

Retaining patient-partner dyads during the 4-month PERC intervention (or use of the NCI website) was just as crucial as follow-up survey completion. The interventionist was a master's degree–prepared nurse clinician with decades of experience working with patients with genitourinary diseases, including prostate cancer. In preparation for her work on the study, she received 20 hours of PERC intervention training. REDCap automatically notified the interventionist through email after a dyad was randomized to the PERC or NCI control group. Patients and partners received their usernames and passwords for a central study website through email and mail, which blinded the study participants and the research staff. After logging in, the dyads were directed to either the NCI prostate cancer web page or the PERC website. The research staff provided ongoing technical assistance.

In addition, we informed participants in both groups that they could contact the study team at any time (including the interventionist) through telephone (including a toll-free number), email, and the "Contact Us" link on the study website. The research team responded to participants' inquiries and requests within 24-48 hours. The team also sent all participants study updates, including midterm outcome findings or publications, through email (the details about the PERC intervention and the NCI control can be found in our protocol paper [13] and outcome paper that is forthcoming).

Measurements

Outcomes

To calculate the enrollment and retention rates, we obtained the numbers of participants who refused or withdrew from the study,

XSL•FO
RenderX

as well as those who enrolled and remained in the RCT, after exporting and deidentifying the relevant administrative data in REDCap.

Enrollment Rate

We defined the enrollment rate as the number of dyads enrolled (both the patient and partner consented and completed) at T1 divided by the number of dyads contacted and accepted (neither the patient nor partner declined). In this dyadic study, dyads were not considered enrolled if one partner consented and completed T1 but the other partner did not consent, or if 1 partner consented without completing T1 by the time the other partner consented and completed T1.

Retention Rate

We defined the retention rate at each follow-up as the number of dyads who completed corresponding surveys divided by those who enrolled.

Other Variables

We collected Gleason scores from RCA referral documents and categorized the patients into low-to-intermediate (Gleason score<8) and high grades (Gleason score8) [16]. Race and comorbidity were collected separately from patients' and partners' self-reports at baseline. We categorized the race as White, Black, and others (including American Indian, Asian, and multiracial). Comorbidity was measured using the Charlson Comorbidity Index (CCI). CCI calculated one's comorbidity score by summing up the comorbid conditions the individuals reported as "yes" [17].

Data Analyses

We conducted descriptive analyses to examine the recruitment and retention rates and presented the 95% CI values for these rates. Chi-square tests assessed differences in survey completion (phone vs online) among patients and their partners, as well as overall response rates at follow-ups T2 to T4. Logistic regression was used to explore the impact of demographic and clinical factors—specifically, age, race, CCI of patients and partners, and patients' Gleason scores—on retention, while adjusting for group assignment. The Mann-Whitney *U* test was used to examine age differences between participants opting for online versus telephone surveys at each assessment point (T1, T2, T3, and T4). All tests were conducted at a 2-sided significance level of α=.05, using SAS software (version 9.4, SAS Institute Inc.), and all confidence intervals are 95% CIs.

Ethical Considerations

The study was approved by the Institutional Review Boards at the University of North Carolina at Chapel Hill (IRB# 17-0482) and the Institutional Review Boards at the University of Texas Health Science Center at San Antonio (IRB# 23-0191). The team obtained written or verbal consent from all patients with prostate cancer and their partners enrolled in this study. The verbal consent was recorded. Written and verbal consent was saved in a password-protected secured folder on an encrypted network separate from the survey data. Only authorized research staff had access to the recordings and data.

Results

Overview

The mean age of the patients was approximately 64 years, and their partners' age averaged 61 years. Their average relationship length was nearly 33 years. While all patients were male, most partners (279/280, 99.6%) were female. Approximately 76.79% (215/280) of patients identified as White patients, and 23.21% (65/280) identified to Black or other racial groups. About 70% (196/280) of patients had less than a college or bachelor's degree, with nearly 49% (135/280) reporting a family income exceeding US $90,000. More than half of the patients and partners (127/280, 45.36%) were not presently working. Approximately three-quarters of patients (208/280, 74.29%) underwent surgery.

Recruitment

Between April 2018 and April 2021, the NCCCR RCA referred 3078 patients with prostate cancer who were 40-75 years old and potentially had a partner, as indicated in their electronic medical record. The RCA staff uploaded an average of 49 referrals every other week between April 2018 and December 2019 in the prepandemic period and 31 referrals from January 2020 to April 2021 during the COVID-19, representing a 36.73% decrease due to the reduced number of patients receiving active treatment because of the closure and lockdown of health care facilities.

The research team sent 3077 physician opt-out letters (1 patient was removed due to incorrect mailing information) and screened 2899 prospective patients for eligibility (Figure 1 [18]). We successfully contacted the partners of 357 prospective patients who met the inclusion criteria and provided us with their partner's contact information. In total, we excluded 2721 patients from the RCA patient referral pool for reasons such as refusal (ie, opt-out letters returned: n=83, 3% physicians and n=405, 14.9% patients; invalid contact information: n=835, 30.7%; and ineligibility: n=507, 18.6%).

XSL•FO
RenderX

Figure 1. CONSORT patient recruit flow diagram. Our study focuses on both patients and partners. *Discrepancies in the dyads' completion number at each time point were due to dyads contacting our team to skip 1 or 2 follow-up surveys. CONSORT: Consolidated Standards for Reporting Trials.

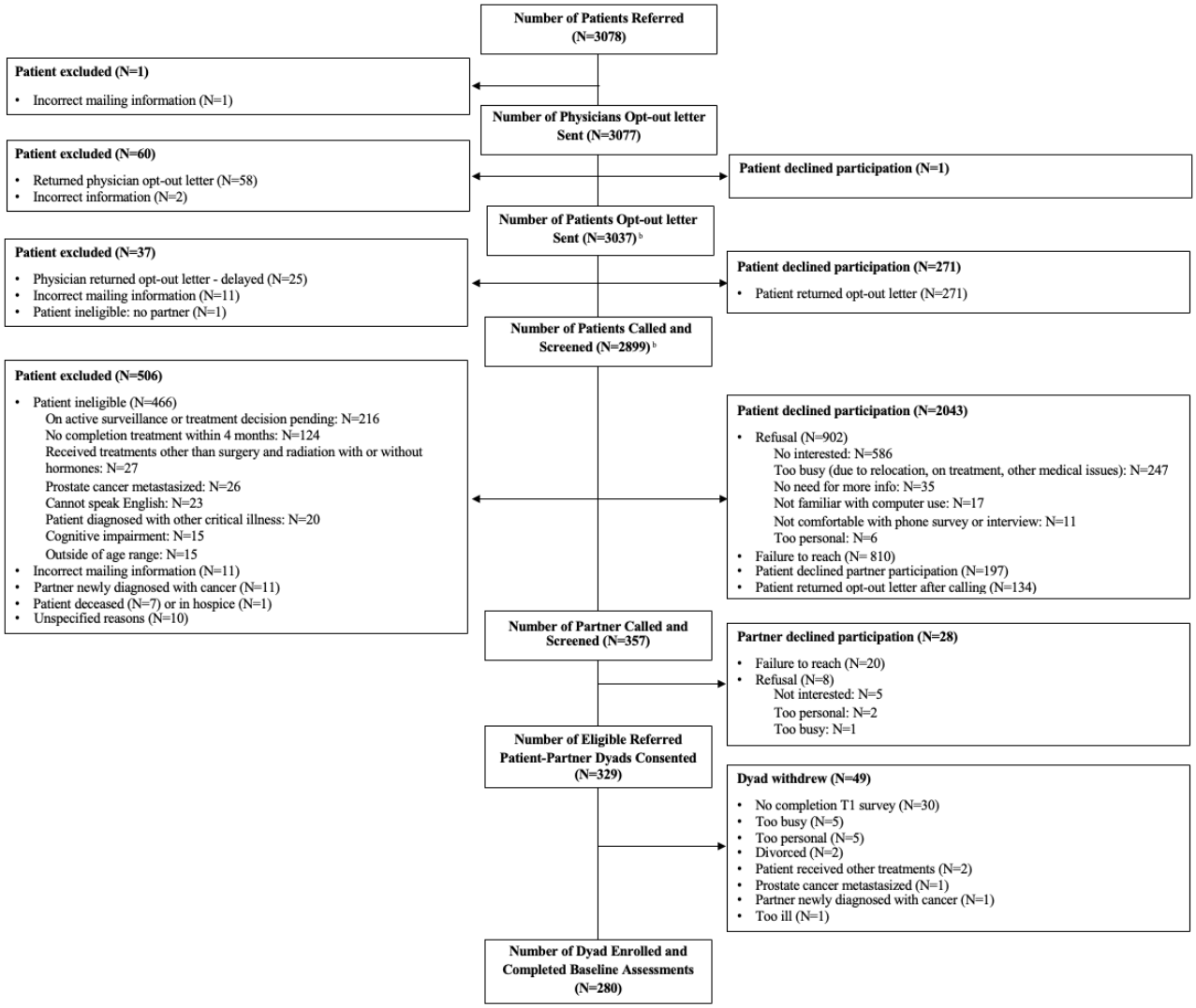

Among the 357 patient-partner dyads contacted, we obtained consent from 329 dyads after 28 partners declined participation because of failure to reach (n=20), lack of interest (n=5), too personal (n=2), and having no time (n=1). A total of 49 dyads withdrew after consenting because they did not complete the baseline survey (n=30), were too busy (n=5), felt the study was too personal (n=5), or were divorced (n=2). The other drop-out reasons also included the patient having received new treatment (n=2) or cancer metastasized (n=1), the partner having a new cancer diagnosis (n=1), or the patient being too ill (n=1; Figure 1).

Enrollment

As displayed in Table 1 and Figure 2, we enrolled 280 patient-partner dyads, including 1 same-sex couple, 20.7% (58/280) Black dyads, and 2.5% (7/280) Asian and Native American dyads. The mean age of enrolled dyads was 63.08 (SD 7.25) years. We randomly assigned the dyads to the PERC intervention (n=141) or the NCI control group (n=139). The demographics were reported in our previous paper [19]. Among the 3078 referred patients, 11.6% (n=357) were eligible and consented to the research team to contact their partners. Our dyad enrollment rate was 85.1%, that is, the number of dyads who consented and completed T1 (n=280) / (number of dyads contacted [n=357] – number of dyads refused [n=28]). The trend in the number of consented and enrolled dyads is presented in Figure 3.

XSL•FO
RenderX

Table 1. Recruitment and retention by race and group.

Group	Total, N	White[a], n (%)	Retention rate, %	Black[a], n (%)	Retention rate, %	*P* value[b]	Others, n (%)	Retention rate, %
RCA [c] referred	3078	1992 (64.72)	__[d]	696 (22.61)	—	—	390 (12.67)	—
Enrolled	280	215 (76.79)	—	58 (20.71)	—	—	7 (2.5)	—
T1								
PERC[e]	141	107 (75.89)	—	28 (19.86)	—	—	6 (4.26)	—
NCI[f] attention control	139	108 (77.7)	—	30 (21.58)	—	—	1 (0.72)	—
T2								
PERC	113	88 (77.88)	82.24	24 (21.24)	85.71	.78	1 (0.88)	16.67
NCI attention control	112	85 (75.89)	78.7	26 (23.21)	86.76	.44	1 (0.89)	100
T3								
PERC	104	83 (79.81)	77.57	19 (18.27)	67.86	.29	2 (1.92)	33.33
NCI attention control	113	86 (76.11)	79.63	26 (23.01)	86.76	.44	1 (0.88)	100
T4								
PERC	106	82 (77.36)	76.64	22 (20.75)	78.57	.83	2 (1.98)	33.33
NCI attention control	115	88 (76.52)	81.48	26 (22.61)	86.76	.60	1 (0.87)	100

[a]We used the patient's race to represent the dyad because there was high congruence in patient and partner self-reported races: 92.3% (205/215) of partners of White patients identified as White, and 91.4% (53/58) of partners of Black patients identified as Black.

[b]*P* value obtained from Fisher tests and chi-square tests. The difference between White and other races was not calculated because of the small sample size. A *P* value of .05 is considered significant.

[c]RCA: rapid case ascertainment.

[d]Not available.

[e]PERC: Prostate Cancer Education and Resources for Couples.

[f]NCI: National Cancer Institute.

Figure 2. CONSORT dyads recruit and follow-up flow diagram. CONSORT: Consolidated Standards for Reporting Trials; NCI: National Cancer Institute; PERC: Prostate Cancer Education and Resources for Couples.

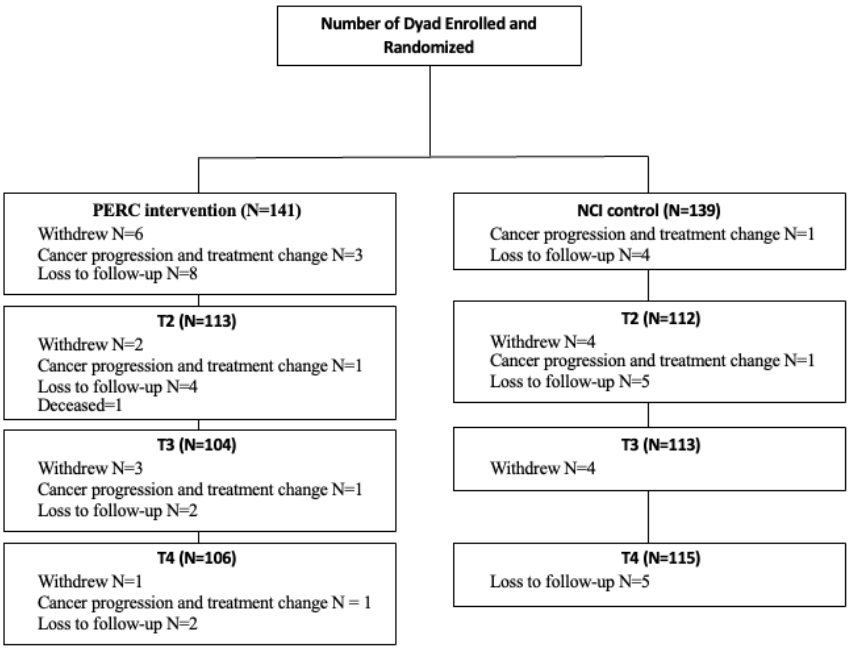

XSL•FO
RenderX

Figure 3. Dyad enrollment (consented and completed T1) quarterly trend between 2018 and 2021.

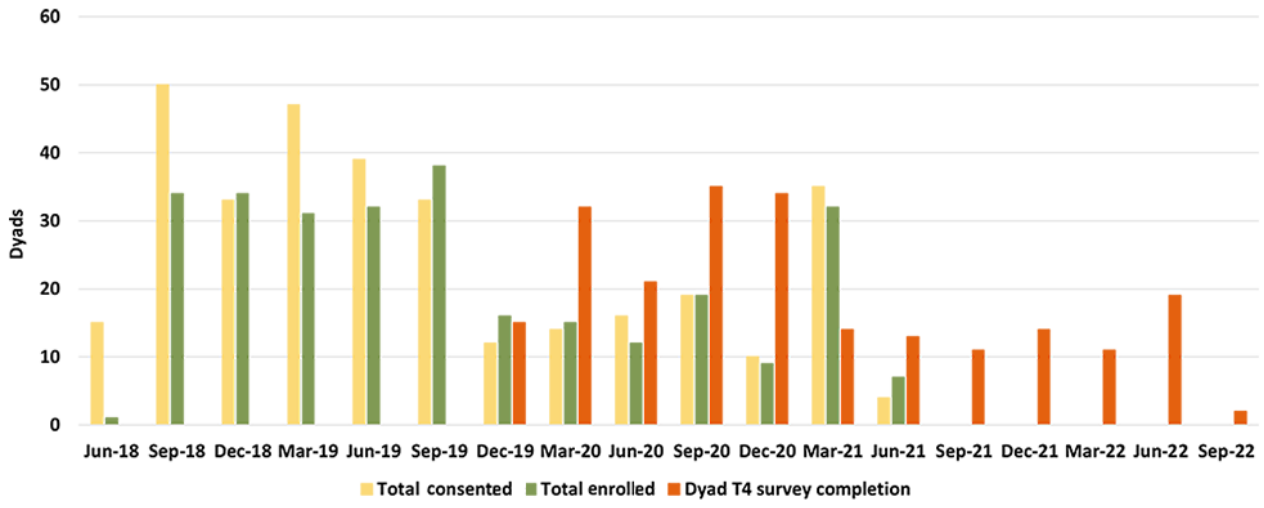

Figure 3. Dyad enrollment (consented and completed T1) quarterly trend between 2018 to 2021.

Retention

Figure 2 presents a CONSORT (Consolidated Standards for Reporting Trials) diagram of randomization and follow-up surveys [18]. Our retention rates were 80.36%, 77.5%, and 78.93% at T2, T3, and T4, respectively. The T2 and T3 retention rates were lower than expected (~90% and 85% at T2 and T3, respectively), whereas the T4 retention rate was close to expected (80% at 12 months after baseline). There was no racial difference in participant retention rates at different time points (Table 1).

Furthermore, 94/280 dyads (33.6%), 128/225 dyads (56.9%), 168/217 dyads (77.4%), and 185/200 dyads (92.5%) completed the baseline T1, T2, T3, and T4 surveys, respectively, during the COVID-19 pandemic (January 2020 to March 2022). The total number of completions peaked in the first 3 months of 2020 and subsequently fluctuated, but with an overall downward trend (Figure 4). The percentages of participants who chose the online versus telephone surveys increased significantly throughout the study (P<.001): 23.93% (134/559), 34.89% (157/450), 57.14% (248/434), and 73.3% (324/441) for T1, T2, T3, and T4 (Table 2), respectively. There is a significant age difference between participants who used phone versus online surveys at T1, T2, and T3; participants of older age were more likely to complete phone surveys at T1 to T3 (P values <.01), but no age difference was found at T4. Compared with White participants, people self-reporting as “other races” (including American Indian, Asian, and multiracial) were more likely to drop out of the study (odds ratio 2.78, 95% CI 1.10-7.04), and older participants were less likely to withdraw (odds ratio 0.96, 95% CI 0.92-0.99).

XSL•FO
RenderX

Figure 4. Dyads T1 to T4 completion (every 3-calendar months during 2018-2022).

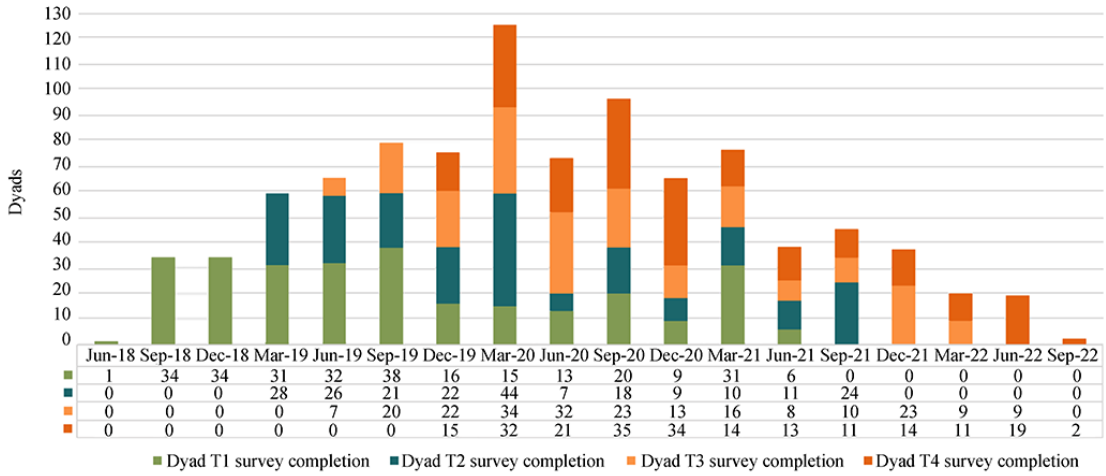

Table 2. Frequency of survey type by group.

Group and role	T1[a] (n=559), n (%)		T2 (n=450), n (%)		T3 (n=434), n (%)		T4[a] (n=441), n (%)	
	Phone survey	Online survey	Phone survey	Online survey	Phone survey	Online survey	Phone survey	Online survey
PERC [b]								
Patient	105 (18.75)	36 (6.43)	71 (15.78)	42 (9.33)	38 (8.76)	66 (15.21)	27 (6.11)	78 (17.65)
Partner	104 (18.57)	37 (6.61)	70 (15.56)	43 (9.56)	38 (8.76)	66 (15.21)	25 (5.66)	81 (18.33)
NCI [c]								
Patient	108 (19.29)	30 (5.36)	78 (17.33)	34 (7.56)	54 (12.44)	59 (13.59)	32 (7.24)	83 (18.78)
Partner	108 (19.29)	31 (5.54)	74 (16.44)	38 (8.44)	56 (12.9)	57 (13.13)	33 (7.47)	82 (18.55)
Total	425 (75.89)	134 (23.93)	293 (65.11)	157 (34.89)	186 (42.86)	248 (57.14)	117 (26.47)	324 (73.3)

[a]Participants completed surveys using mixed methods (half phone survey and half online survey) at T1 (n=1) and at T4 (n=1).

[b]PERC: Prostate Cancer Education and Resources for Couples.

[c]NCI: National Cancer Institute.

Discussion

Principal Findings

To our knowledge, the PERC study is the first to use a cancer registry RCA to facilitate the recruitment of a population-based sample of patient-partner dyads who have been recently treated for prostate cancer for a symptom management eHealth intervention RCT. We completed participant recruitment and follow-up data collection despite the COVID-19 pandemic. We received 3087 referrals from RCA. Following a thorough screening process, we obtained permission from 357 eligible patients to contact their partners for consent. Subsequently, 280 patients and partners consented, completed baseline assessments, and were randomized, achieving an enrollment rate of 85.11%. Among these, 221 dyads completed the 12-month follow-up, reflecting a robust retention rate of 78.93%. Previous reviews have indicated that 86% of RCTs fail to finish on time, mainly due to insufficient enrollment [20]. Patient participation in RCTs decreases with advanced age, and ethnic and minority groups are underrepresented [21]. Coupled with the difficulties inherent to family research, as discussed previously, these challenges indicate that patients with prostate cancer, with a median age of 66 years, are among the hardest-to-reach populations. The outbreak of the COVID-19 pandemic in 2020 and the subsequent shelter-in-place for extended periods further complicated the dyad-focused PERC RCT. Despite these challenges, we proactively and adaptively used a series of strategies, including contacting participants using telephone, email, and regular mail; using telephone and online surveys; and implementing an eHealth intervention program, to ensure completion of the proposed study as planned. Although significantly impacted by the COVID-19 pandemic lockdowns, we successfully accrued enough study participants (280 dyads) to test our RCT hypotheses; our enrollment and retention rates in this NCCCR RCA–facilitated, dyad-focused, eHealth RCT are superior to the family-based intervention studies that recruited from cancer centers, hospitals, and community settings [4]. We also found that people who were younger and who self-reporting as other races were more likely to drop off from the RCT as compared with their White counterparts, and older participants were more likely to complete surveys through phone during T1, T2, and T3.

XSL•FO
RenderX

Participant Recruitment

Our study is unique in that we used a cancer registry to recruit patient-partner dyads for a symptom management eHealth intervention RCT. Cancer registries have been used extensively in cancer surveillance and epidemiologic research [22,23] and in research that aims to improve cancer detection and treatment [24-26], mostly among patients with cancer [22-26] and occasionally among patients and clinicians [27]. We used a novel NCCCR RCA–facilitated approach to expand our geographic catchment area and enroll patients with newly treated prostate cancer and their intimate partners. To facilitate the recruitment of participants across the state of North Carolina and to reduce costs, we targeted the 36 counties with the highest proportions of ethnic populations and the highest rates of poverty and designed our protocol to contact patients through phone, text messaging, email, and ground mail, without in-person meetings. Upon the onset of the pandemic, we quickly adjusted the work schedule for our research team members to adapt to the lockdown orders and shelter-in-place mandates without affecting the overall flow of the study or requiring significant protocol amendments. As a result, we continued to recruit throughout the pandemic and successfully enrolled 280 prostate cancer patient-partner dyads. Our enrollment rate of prostate cancer patient-partner dyads (85.11%) is significantly higher than the mean and median enrollment rates (33% and 23%, respectively) of studies recruited from cancer hospitals and centers using in-person and clinician referral approaches [4]. The racial composition of our study participants (215/280, 76.79% White) is similar to that (75.9%) of family-based intervention studies conducted in the United States and the United Kingdom [4].

Although recruitment using the NCCCR RCA allowed us to overcome many of the challenges when in-person contact was impossible during the pandemic, we excluded approximately 43% (1331/3096) of the referred patients because of incorrect contact information and patient ineligibility, indicating a need to improve the accuracy of data in electronic medical records and during data extraction (eg, updated contact information, diagnosis, and treatment). Researchers could benefit from working closely with the cancer registrars to refine initial screening criteria and reduce the referral of ineligible patients.

We have identified the reasons for nonparticipation in this dyad-focused eHealth RCT. Like previous patient-focused RCTs [21,28], the main reasons for nonparticipation included a lack of interest, privacy concerns about discussing personal issues, and challenges related to patients' cancer and treatment options. It is worth noting that 197 patients declined partner participation, making them ineligible for this dyad-focused RCT. Prostate cancer is known as a "couple's illness" because cancer-related symptoms and stress negatively affect the quality of life of patients and their partners, and partners often have worse quality of life than the patients [29]. Dyad-focused interventions and RCTs have become increasingly common in managing health issues and quality of life challenges for patients with cancer or their caregivers (eg, partners); this is because the improvement in symptoms and quality of life involves psychosocial processes and communication between patients and caregivers, and resolving complex issues require couples to work together as a team [30-32]. When patients act as gatekeepers and decline partner participation, patients and partners both miss the opportunity to engage in and benefit from a dyad-focused symptom management program. Future research is needed to understand why patients decline their partners' participation and develop strategies to engage them, as members of dyads, in psychosocial behavioral interventions to improve both people's quality of life. Future RCTs can also stratify randomization based on partner availability.

It is worth noting that due to the intimate and private nature of prostate cancer, the privacy of cancer-related information and hesitancy to discuss topics of a personal nature were central challenges for many dyads. Regardless of assurances from the study team and explanations of measures taken to protect privacy and confidentiality, many patients and partners in the recruitment phase refused to participate over concerns about access to their personal, medical, and psychosocial information. During the RCT, we noted participant hesitancy or refusal to discuss intimate issues such as sexual function or sexual symptoms. Researchers must consider participants' desire to maintain privacy and confidentiality, earn their trust, and engage in best practices for addressing their legitimate privacy concerns during recruitment and throughout study implementation, especially when in-person contact is impossible.

Improve Communication and Build Trust During Recruitment and Retention

Because trust is the foundation of recruitment and retention [33], especially in digital health research [34], we used different strategies to build trust among dyads in the RCT. Although providing flexibility and convenience, contacting prospective participants through telephone, email, and mail without face-to-face meetings could jeopardize participant accrual and retention, because cold-calling and remote contacts reduced trust. The sensitive nature of prostate cancer could exacerbate these concerns for patients and partners. Furthermore, prospects and participants may not have felt the continued desire to communicate or ask questions as time passed since treatment completion and as the pandemic continued. In this RCT, we adopted the following strategies built upon previous research [35-37] to acknowledge the vital role of trust and communication in recruitment and retention (Textbox 1).

XSL•FO
RenderX

Textbox 1. Strategies to acknowledge in recruitment and retention.

- We provided opt-out letters for both the patients and their physicians, which enhanced patient autonomy (eg, offering choice and control), beneficence (eg, reducing selection bias), and justice (eg, reducing lack of generalizability) [38].
- We formed a diverse research team that comprised members of different demographic backgrounds (eg, Black, Asian, and White American; Hispanics and non-Hispanics; and young and older males and females) so that patients and partners felt comfortable communicating with the research team. We included pictures of and information about the research team members on the study website hosted at a research University and in the physician and patient opt-out letters to help develop familiarity with the team members.
- We provided extensive training for the research staff and used role play to practice interviews for recruitment, consent, and surveys.
- When contacting participants, our staff used the university-designated team email and their email accounts through their dedicated university office phone numbers and, later a Health Insurance Portability and Accountability Act–compliant, cloud-based phone system, which displayed the University of North Carolina name; hence, participants knew who was calling them.
- We ensured that each participant was contacted for all surveys by the same research staff throughout the study period unless participants requested a different staff member. Doing so helped the research staff and study participants become familiar with each other, have consistent interpersonal interaction, and build rapport; this made the participants willing to share their concerns and feel comfortable requesting scheduling changes rather than canceling the research activities altogether.
- To ensure participant convenience, we asked their preferred time for being contacted and accommodated mornings, afternoons, and evenings during both weekdays and weekends.
- We responded to participants' phone calls and emails within 24-48 hours. Research staff contacted clinicians and coinvestigators for assistance when they could not answer the questions or address participants' concerns.
- We planned regular reminder emails, follow-up emails, and phone calls consistent with the recruitment schedules, follow-up surveys, and intervention activities.
- We provided gift cards after each survey and retention gifts with our team logo between follow-up surveys.

These approaches were vital to maintaining positive communication and establishing trust with our participants, from our early recruitment stage, when we helped prospective participants develop an interest in the study, to follow-ups when participants were not responsive. Consistently using these strategies ultimately led us to achieve our recruitment and retention goals for this dyad-focused eHealth intervention RCT despite the extended period of the stressful pandemic. Despite these strategies having been used to improve the representation of minority participants in previous research [35-37], our study population included one same-sex dyad. The overall percentages of Black American (58/280, 20.71%,) and other racial group patients (7/280, 2.5%, mainly Asian and Native American) in our RCT are also slightly lower than those in the North Carolina population (22.1% and 3.7%, respectively) [39]. The high incident and mortality rates of COVID-19 among Black and Asian participants might have negatively affected the participation of racially diverse minority participants because the percentage of Black patients dropped from >25% before COVID-19. Future research must explore more effective, rigorous strategies to promote racial diversity in family-based intervention research.

Participant Retention

A positive influence on our retention rates was that in our protocol, the method of participant contact was flexible, as we contacted participants through phone, text messages, email, and ground mail. Although before the pandemic, many studies used in-person or phone surveys and interviews [33], we built more flexibility into our method of study implementation (eg, surveys conducted through telephone and later online). What is significant to participant retention is that the PERC intervention is an online platform. We intentionally developed PERC to address the needs of patients who live in rural areas and to accommodate the private nature of prostate cancer symptoms and complications (eg, sexual dysfunction). The online medium provides patients and family caregivers easy access to posttreatment supportive care whenever and wherever they need it. In addition, we provided training for participants unfamiliar with online platforms. We loaned participants a tablet computer and hotspot internet connectivity if they could not access the internet.

Meanwhile, the pandemic and associated lockdowns changed the research environment as the public became familiar with online technology for social interactions, meetings, and schooling during the pandemic. Thus, as a result, online surveys have become more widely used in research [40]. Many of our participants also requested more flexibility in completing survey assessments due to increased demands at home and work. After team discussion and consultation with our psychometrician, we added the online survey option to meet participants' needs. The percentage of participants who completed online surveys increased significantly over time, ranging from 23.93% (134/559) for T1 to 73.3% (324/441) for T4. Unlike other RCTs for which follow-up retention rates decrease over time, our retention rates at different time points were flat (ie, 80.36%, 77.5%, and 78.93% at T2, T3, and T4, respectively). Adding the online survey option might have eased the negative impact of pandemic-related chaos and stress on our participants' completion of study surveys during 2020 and early 2021, which might have reduced the effects of the pandemic lockdown on the RCT. It is worth noting that our retention rates were similar in both the PERC and control groups across different time points for different racial groups; these retention rates are equivalent to or better than most family-based intervention studies (69%-70% on average) [4].

XSL•FO
RenderX

Limitations

This study has several limitations. First, we excluded a large percentage of the patients referred by the cancer registry RCA because of incorrect contact information and ineligibility, which may have caused sample selection bias. Second, the percentages of Black American patients and patients from other non-White groups were lower than projected despite the various strategies we used. Anecdotally, some dyads from minority groups declined RCT participation because the partners worked multiple jobs while the patients were out of work due to prostate cancer and related treatment. Future research must examine effective, innovative strategies to reduce system barriers and engage nonintimate caregivers from minority groups. Finally, verifying participants' responses to online surveys was difficult when they entered incorrect data. We ensured data quality and accuracy for the telephone survey by randomly selecting up to 50% of all data to check against the phone survey recording and conducting double data entry. For the online surveys, our database manager immediately examined the data after the surveys were submitted; we contacted the respondents if the data were incomplete or included unusual values. Despite controversies in the strengths and weaknesses of online versus telephone surveys [41,42], online surveys have been shown, as compared with telephone surveys, to produce more reliable and complete data and to be cheaper and less time-consuming to conduct [42]. We provided free internet access (ie, iPad [Apple] and JackPot [Microgaming]) and training to increase the participation of people with limited ability to engage in digital health research. Following the recommendation for recruiting a diverse study population [43], we combined telephone and online surveys to ensure complete and correct data reporting, especially during the unprecedented disruption of the COVID-19 pandemic. Finally, the retention efforts of the study team, while effective for the RCT, may not translate well to intervention implementation when the team may not be available for participant engagement. However, many strategies can help enhance communication between the research team and participants, build trust, and ultimately, benefit participant enrollment and retention in family-based research projects.

Conclusions

This study details the recruitment and retention challenges of our dyad-focused eHealth intervention RCT, related strategies that our team used, and outcomes of participant recruitment and retention during the pandemic. We innovatively used the NCCCR RCA to identify potential participants, which enabled us to continue to recruit during the COVID-19 pandemic, as we could complete the RCT without in-person human contact. Similarly, our use of online surveys, eHealth intervention, and other alternatives to in-person interactions, as well as our dedication to persistent and regular communication with participants, positively participant recruitment and retention during the unprecedented pandemic.

Data Availability

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

The project is sponsored by the National Institute of Health-National Institute of Nursing Research (NR016990, principal investigator LS). We thank the patients with prostate cancer, their partners, health care professionals, and our research staff for making the study possible.

Authors' Contributions

LS, RCC, LN, and CR contributed to the conceptualization and design of this study. LS oversaw the project. CM performed data analysis. LS and CM interpreted the results. LS, CM, RHA, and DBN contributed to the first draft of this manuscript. All authors made critical and intellectual comments on the various versions of the manuscript. All authors approved the final version and agreed to be accountable for all aspects of this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 395 KB - jmir_v26i1e51877_app1.pdf]

References

1. Fredman SJ, Baucom DH, Gremore TM, Castellani AM, Kallman TA, Porter LS, et al. Quantifying the recruitment challenges with couple-based interventions for cancer: applications to early-stage breast cancer. Psychooncology 2009;18(6):667-673 [FREE Full text] [doi: 10.1002/pon.1477] [Medline: 19061201]
2. Kleine A, Hallensleben N, Mehnert A, Hönig K, Ernst J. Psychological interventions targeting partners of cancer patients: a systematic review. Crit Rev Oncol Hematol 2019;140:52-66. [doi: 10.1016/j.critrevonc.2019.05.008] [Medline: 31176272]
3. Teixeira RJ, Applebaum AJ, Bhatia S, Brandão T. The impact of coping strategies of cancer caregivers on psychophysiological outcomes: an integrative review. Psychol Res Behav Manag 2018;11:207-215 [FREE Full text] [doi: 10.2147/PRBM.S164946] [Medline: 29872357]

XSL•FO
RenderX

4. Song L, Qan'ir Y, Guan T, Guo P, Xu S, Jung A, et al. The challenges of enrollment and retention: a systematic review of psychosocial behavioral interventions for patients with cancer and their family caregivers. J Pain Symptom Manage 2021;62(3):e279-e304 [FREE Full text] [doi: 10.1016/j.jpainsymman.2021.04.019] [Medline: 33933618]
5. Tom M, Cohen EJ, LopezJimenez C, Hochman JS, Troxel AB, Jeng BH. Meeting the challenges of retention and enrollment of study participants in clinical trials during the COVID-19 pandemic from the study leadership perspective: experience from the zoster eye disease study (ZEDS). Contemp Clin Trials Commun 2021;24:100875 [FREE Full text] [doi: 10.1016/j.conctc.2021.100875] [Medline: 34805615]
6. Pembroke S, Rogerson S, Coyne I. Conducting a randomised controlled trial of a psychosocial intervention for adolescents with type 1 diabetes during COVID-19: recommendations to overcome the challenges complicated by inconsistent public health guidelines on research. Trials 2022;23(1):362 [FREE Full text] [doi: 10.1186/s13063-022-06314-9] [Medline: 35477434]
7. Audisio K, Lia H, Robinson N, Rahouma M, Soletti G, Cancelli G, et al. Impact of the COVID-19 pandemic on non-COVID-19 clinical trials. J Cardiovasc Dev Dis 2022;9(1):19 [FREE Full text] [doi: 10.3390/jcdd9010019] [Medline: 35050229]
8. Farmer N, Tuason RT, Kazmi N, Flynn S, Mitchell V, Middleton K, et al. Going virtual during the COVID-19 pandemic: adaptation of a mixed-methods dietary behavior study within a community-based participatory research study of African-American adults at risk for cardiovascular disease. BMC Med Res Methodol 2022;22(1):1-10 [FREE Full text] [doi: 10.1186/s12874-022-01806-3] [Medline: 36550396]
9. Walker CM, Williams LA, Bowdre T. Lessons learned in abruptly switching from in-person to remote data collection in light of the COVID-19 pandemic. TQR 2021;26(8):2382-2389. [doi: 10.46743/2160-3715/2021.4745]
10. Wurz A, Janzen A, Ellis K, Lesser I, Arshad N. Describing supportive care programming access and comfort gathering through the COVID-19 pandemic: an observational mixed methods study with adults affected by cancer. Curr Oncol 2023;30(3):2598-2612 [FREE Full text] [doi: 10.3390/curroncol30030198] [Medline: 36975411]
11. Evered J, Andersen L, Foxwell A, Iroegbu C, Whitney C. The impact and implications of virtual supportive cancer care during the COVID-19 pandemic: integrating patient and clinician perspectives. Support Care Cancer 2022;30(12):9945-9952 [FREE Full text] [doi: 10.1007/s00520-022-07393-9] [Medline: 36208318]
12. Rucinska M, Nawrocki S. COVID-19 Pandemic: impact on cancer patients. Int J Environ Res Public Health 2022;19(19):12470 [FREE Full text] [doi: 10.3390/ijerph191912470] [Medline: 36231769]
13. Song L, Nielsen ME, Chen RC, Rini C, Keyserling TC, Idiagbonya E, et al. Testing the efficacy of a couple-focused, tailored eHealth intervention for symptom self-management among men with prostate cancer and their partners: the study protocol. Trials 2022;23(1):12 [FREE Full text] [doi: 10.1186/s13063-021-05948-5] [Medline: 34983621]
14. Rapid case ascertainment. UNC Lineberger Comprehensive Cancer Center. 2023. URL: https://unclineberger.org/rapid-case-ascertainment/ [accessed 2024-09-19]
15. Patridge EF, Bardyn TP. Research Electronic Data Capture (REDCap). JMLA 2018;106(1):1-3. [doi: 10.5195/jmla.2018.319]
16. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A contemporary prostate cancer grading system: a validated alternative to the gleason score. Eur Urol 2016;69(3):428-435 [FREE Full text] [doi: 10.1016/j.eururo.2015.06.046] [Medline: 26166626]
17. Charlson ME, Pompei P, Ales KL, MacKenzie C. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
18. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. J Pharmacol Pharmacother 2010 Jul;1(2):100-107 [FREE Full text] [doi: 10.4103/0976-500X.72352] [Medline: 21350618]
19. Song L, Keyserling TC, Chen RC, Ma C, Xu S, Shieh K, et al. Role, race, and place: prostate cancer disparities in patients' and partners' health outcomes and psychosocial factors. Cancer Med 2023;12(8):9857-9867 [FREE Full text] [doi: 10.1002/cam4.5646] [Medline: 36748581]
20. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018;11:156-164 [FREE Full text] [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
21. Unger J, Hershman DL, Till C, Minasian LM, Osarogiagbon RU, Fleury ME, et al. "When Offered to Participate": a systematic review and meta-analysis of patient agreement to participate in cancer clinical trials. J Natl Cancer Inst 2021;113(3):244-257 [FREE Full text] [doi: 10.1093/jnci/djaa155] [Medline: 33022716]
22. White MC, Babcock F, Hayes NS, Mariotto AB, Wong FL, Kohler BA, et al. The history and use of cancer registry data by public health cancer control programs in the United States. Cancer 2017;123 Suppl 24(Suppl 24):4969-4976 [FREE Full text] [doi: 10.1002/cncr.30905] [Medline: 29205307]
23. Parkin DM. The evolution of the population-based cancer registry. Nat Rev Cancer 2006;6(8):603-612. [doi: 10.1038/nrc1948] [Medline: 16862191]
24. Islami F, Siegel RL, Jemal A. The changing landscape of cancer in the USA - opportunities for advancing prevention and treatment. Nat Rev Clin Oncol 2020;17(10):631-649. [doi: 10.1038/s41571-020-0378-y] [Medline: 32467620]

XSL•FO
RenderX

25. Bacchus MW, McKee B, Gwede CK, Cogle CR. Implementation of cancer plans in the United States: a review. Healthcare (Basel) 2021;9(3):291 [FREE Full text] [doi: 10.3390/healthcare9030291] [Medline: 33800077]
26. Millar MM, Elena JW, Gallicchio L, Edwards SL, Carter ME, Herget KA, et al. The feasibility of web surveys for obtaining patient-reported outcomes from cancer survivors: a randomized experiment comparing survey modes and brochure enclosures. BMC Med Res Methodol 2019;19(1):208 [FREE Full text] [doi: 10.1186/s12874-019-0859-9] [Medline: 31730474]
27. Prang K, Karanatsios B, Zhang A, Verbunt E, Wong H, Wong V, et al. "Nothing to lose and the possibility of gaining": a qualitative study on the feasibility and acceptability of registry-based randomised controlled trials among cancer patients and clinicians. Trials 2023;24(1):92 [FREE Full text] [doi: 10.1186/s13063-023-07109-2] [Medline: 36747274]
28. Cartmell K, Bonilha HS, Simpson KN, Ford ME, Bryant DC, Alberg AJ. Patient barriers to cancer clinical trial participation and navigator activities to assist. Adv Cancer Res 2020;146:139-166 [FREE Full text] [doi: 10.1016/bs.acr.2020.01.008] [Medline: 32241387]
29. Song L, Northouse LL, Braun TM, Zhang L, Cimprich B, Ronis DL, et al. Assessing longitudinal quality of life in prostate cancer patients and their spouses: a multilevel modeling approach. Qual Life Res 2011;20(3):371-381 [FREE Full text] [doi: 10.1007/s11136-010-9753-y] [Medline: 20927648]
30. Tiete J, Delvaux N, Liénard A, Razavi D. Efficacy of a dyadic intervention to improve communication between patients with cancer and their caregivers: a randomized pilot trial. Patient Educ Couns 2021;104(3):563-570. [doi: 10.1016/j.pec.2020.08.024] [Medline: 33129628]
31. Hu Y, Liu T, Li F. Association between dyadic interventions and outcomes in cancer patients: a meta-analysis. Support Care Cancer 2019;27(3):745-761. [doi: 10.1007/s00520-018-4556-8] [Medline: 30604008]
32. Badr H, Bakhshaie J, Chhabria K. Dyadic interventions for cancer survivors and caregivers: state of the science and new directions. Semin Oncol Nurs 2019;35(4):337-341 [FREE Full text] [doi: 10.1016/j.soncn.2019.06.004] [Medline: 31248677]
33. Andrews L, Davies TH. Participant recruitment and retention from vulnerable populations in clinical trials is a matter of trust. Contemp Clin Trials 2022;123:106969. [doi: 10.1016/j.cct.2022.106969] [Medline: 36273802]
34. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
35. Chalela P, Suarez L, Muñoz E, Gallion KJ, Pollock BH, Weitman SD, et al. Promoting factors and barriers to participation in early phase clinical trials: patients perspectives. J Community Med Health Educ 2014;4(281):1000281 [FREE Full text] [doi: 10.4172/2161-0711.1000281] [Medline: 25077043]
36. Otado J, Kwagyan J, Edwards D, Ukaegbu A, Rockcliffe F, Osafo N. Culturally competent strategies for recruitment and retention of African American populations into clinical trials. Clin Transl Sci 2015;8(5):460-466 [FREE Full text] [doi: 10.1111/cts.12285] [Medline: 25974328]
37. Tseng CF, Miller DL, Huerta P, PettyJohn ME, Yzaguirre MM, Agundez JC, et al. Recruitment strategies in couple intervention studies: a systematic review of recruitment methods and sample characteristics in the United States from 2015 to 2020. Fam Process 2023;62(3):880-898. [doi: 10.1111/famp.12885] [Medline: 37086013]
38. Pittman T, Bell L, Jones S, Brown K, Kirchoff K, Flume P. Enhancing study recruitment through implementation of an opt-out, cold contact process with consideration for autonomy, beneficence and justice. J Clin Transl Sci 2023;7(1):e63 [FREE Full text] [doi: 10.1017/cts.2023.21] [Medline: 37008607]
39. Quick facts: North Carolina. United States Census Bureau. 2022. URL: https://www.census.gov/quickfacts/NC [accessed 2024-09-19]
40. Hlatshwako TG, Shah SJ, Kosana P, Adebayo E, Hendriks J, Larsson EC, et al. Online health survey research during COVID-19. Lancet Digit Health 2021;3(2):e76-e77 [FREE Full text] [doi: 10.1016/S2589-7500(21)00002-9] [Medline: 33509387]
41. Hollier LP, Pettigrew S, Slevin T, Strickland M, Minto C. Comparing online and telephone survey results in the context of a skin cancer prevention campaign evaluation. J Public Health (Oxf) 2017;39(1):193-201. [doi: 10.1093/pubmed/fdw018] [Medline: 26968482]
42. Braunsberger K, Wybenga H, Gates R. A comparison of reliability between telephone and web-based surveys. Journal of Business Research 2007;60(7):758-764. [doi: 10.1016/j.jbusres.2007.02.015]
43. Szolnoki G, Hoffmann D. Online, face-to-face and telephone surveys—comparing different sampling methods in wine consumer research. Wine Economics and Policy 2013;2(2):57-66. [doi: 10.1016/j.wep.2013.10.001]

Abbreviations

CCI: Charlson Comorbidity Index
CONSORT: Consolidated Standards for Reporting Trials
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
NCCCR: North Carolina Central Cancer Registry
NCI: National Cancer Institute
PERC: Prostate Cancer Education and Resources for Couples

XSL•FO
RenderX

RCA: rapid case ascertainment
REDCap: Research Electronic Data Capture
RCT: randomized controlled trial

Edited by T de Azevedo Cardoso; submitted 15.08.23; peer-reviewed by Y Su, K Tan; comments to author 28.11.23; revised version received 16.01.24; accepted 30.08.24; published 03.12.24.

Please cite as:
Ma C, Adler RH, Neidre DB, Chen RC, Northouse LL, Rini C, Tan X, Song L
Challenges and Approaches to Recruitment for and Retention in a Dyad-Focused eHealth Intervention During COVID-19: Randomized Controlled Trial
J Med Internet Res 2024;26:e51877
URL: https://www.jmir.org/2024/1/e51877
doi:10.2196/51877
PMID:39625741

©Chunxuan Ma, Rachel H Adler, Daria B Neidre, Ronald C Chen, Laurel L Northouse, Christine Rini, Xianming Tan, Lixin Song. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Improving Participant Recruitment in Clinical Trials: Comparative Analysis of Innovative Digital Platforms

Alexia Georgia Bikou[1*], MSc; Elena Deligianni[1*], MSc; Foteini Dermiki-Gkana[1*], MSc; Nikolaos Liappas[2], PhD; José Gabriel Teriús-Padrón[2], PhD; Maria Eugenia Beltrán Jaunsarás[2], PhD; Maria Fernanda Cabrera-Umpiérrez[2], PhD; Christos Kontogiorgis[1], PhD

[1]Department of Medicine, Democritus University of Thrace, Alexandroupolis, Greece

[2]Life Supporting Technologies (LifeSTech), Superior Technical School of Telecommunication Engineers, Universidad Politécnica de Madrid (UPM), Madrid, Spain

[*]these authors contributed equally

Corresponding Author:
Christos Kontogiorgis, PhD
Department of Medicine
Democritus University of Thrace
6th kilometer Dragana
Alexandroupolis, 68100
Greece
Phone: 30 2551030601
Fax: 30 302551030601
Email: ckontogi@med.duth.gr

Abstract

Background: Pharmaceutical product development relies on thorough and costly clinical trials. Participant recruitment and monitoring can be challenging. The incorporation of cutting-edge technologies such as blockchain and artificial intelligence has revolutionized clinical research (particularly in the recruitment stage), enhanced secure data storage and analysis, and facilitated participant monitoring while protecting their personal information.

Objective: This study aims to investigate the use of novel digital platforms and their features, such as e-recruitment, e-consent, and matching, aiming to optimize and expedite clinical research.

Methods: A review with a systematic approach was conducted encompassing literature from January 2000 to October 2024. The MEDLINE, ScienceDirect, Scopus, and Google Scholar databases were examined thoroughly using a customized search string. Inclusion criteria focused on digital platforms involving clinical trial recruitment phases that were in English and had international presence, scientific validation, regulatory approval, and no geographic limitations. Literature reviews and unvalidated digital platforms were excluded. The selected studies underwent meticulous screening by the research team, ensuring a thorough analysis of novel digital platforms and their use and features for clinical trials.

Results: A total of 24 digital platforms were identified that supported clinical trial recruitment phases. In general, most of them (n=22, 80%) are headquartered and operating in the United States, providing a range of functionalities including electronic consent (n=14, 60% of the platforms), participant matching, and monitoring of patients' health status. These supplementary features enhance the overall effectiveness of the platforms in facilitating the recruitment process for clinical trials. The analysis and digital platform findings refer to a specific time frame when the investigation took place, and a notable surge was observed in the adoption of these novel digital tools, particularly following the COVID-19 outbreak.

Conclusions: This study underscores the vital role of the identified digital platforms in clinical trials, aiding in recruitment, enhancing patient engagement, accelerating procedures, and personalizing vital sign monitoring. Despite their impact, challenges in accessibility, compatibility, and transparency require careful consideration. Addressing these challenges is crucial for optimizing digital tool integration into clinical research, allowing researchers to harness the benefits while managing the associated risks effectively.

(J Med Internet Res 2024;26:e60504) doi:10.2196/60504

XSL•FO
RenderX

KEYWORDS

clinical research; e-recruitment; patient matching; clinical trials; digital platforms; enrollment; electronic consent

Introduction

Background

Developing a new pharmaceutical product or revising and indicating potential new uses for existing products is a meticulous, time-consuming, and costly process [1]. Clinical trials ultimately represent a key tool in clinical research for advancing medical scientific knowledge and patient care [2]. Recruiting the appropriate target and number of participants remains one of the significant challenges in clinical trials [3]. The recruitment stage is a critical factor in ensuring the success and quality of a clinical trial [4,5]. Recent statistics now indicate that approximately 80% of clinical trials face delays due to patient recruitment issues [6]. Many of these delays last between 1 and 6 months, with some trials facing even longer setbacks. In addition, a significant number of research sites enroll very few or no patients, worsening the delays [6]. The success of clinical trials heavily depends on the prompt identification and recruitment of individuals meeting the study's inclusion criteria [7]. Over the years, a face-to-face approach has been used in various stages of clinical trial research, including recruitment, consent for enrollment, retention, delivery of interventions, and data collection [8,9]. However, it has been widely observed that a significant number of individuals and patients interested in participating in a clinical trial are not adequately informed about how to identify and approach clinical trials that align with their needs and interests [10]. In addition, numerous factors such as the number of patients to be screened, the participating sites, clinical institutions' infrastructures, and data protection laws may significantly affect recruitment [11]. The lack of accessibility to funds and the high costs associated with conducting clinical trials are also notable barriers to the recruitment stage and the overall trial processes [12]. Despite the performance of clinical trials worldwide, the time-consuming manual process of selecting the participants and matching them to the right study remains a major challenge in the trial recruitment stage [13]. Given the increasing number of accessible clinical trials and the complexity of their designs, the aforementioned processes demand in-depth understanding of patient characteristics and eligibility requirements based on the inclusion criteria [14].

In recent years, the integration of cutting-edge technologies into clinical trials has significantly influenced the successful and expedited conduct of various clinical trial phases, especially the recruitment process [15,16]. The expansion of web access; computational tools; and portable devices such as smartphones, tablets, and wearables has had a positive influence on health, research, and development innovation [17]. More specifically, technology's impact on clinical research has led to a shift from traditional recruitment strategies to the adoption of new methods [18]. This transition is closely related to the use of advanced technology tools, including social media and web platforms, for patient recruitment [19]. Consequently, a new category of clinical trials, commonly referred to as digital clinical trials, has arisen [20]. These trials use digital recruitment, electronic consent, health data collection formatted in electronic medical records, and advanced analysis methodologies driven by artificial intelligence (AI) [21]. This integration enhances the automation and acceleration of various clinical trial stages [22]. Nowadays, numerous patient-centric digital platforms and applications have been created and introduced with the aim of recruiting and matching participants or patients with the most fitting clinical trials based on their individual profiles, conditions, and needs [23]. Furthermore, certain platforms have a diverse purpose by providing digital recruitment solutions to health care professionals, academic researchers, and clinical trial sponsors involved in clinical research [24]. Several platforms leveraging AI technology are specifically designed to offer an advanced clinical trial matching system [25]. This system enables the seamless integration of patient recruitment into the most suitable trial based on their characteristics and needs [21]. The primary aim of these patient-driven platforms and mobile apps is to facilitate the recruitment process and increase the willingness of patients and healthy volunteers to participate in ongoing clinical studies [26].

One of the latest evolutions in the field of clinical research involves leveraging blockchain technology and AI in health care [27,28]. For instance, blockchain defines and forms a decentralized and distributed digital ledger that records data in the form of transactions across the entire network in a secure and transparent way [29]. Each transaction is recorded in a block, and when the blocks are completed, they are cryptographically linked, creating a chain [30]. Moreover, distributed ledger technology (DLT) is a broader term that includes blockchain technology [31]. However, not all DLTs are blockchain based. Other DLTs include distributed databases maintained by the participants of a distributed network [32]. On the other hand, AI refers to computer systems capable of replicating human behavior and learning to perform tasks through experience [33]. This is accomplished by simulating human cognitive functions, enabling machines to make decisions and perform tasks that typically require human intelligence [34]. Both AI and blockchain technologies share the advantages of immutability, transparency, tamper-resistant records, and no need for a central authority [35]. These technologies offer the capability to improve secure storage and analysis of vast amounts of data, particularly in organizing electronic medical records [36]. As a result, they ensure the highest level of security for handling sensitive datasets [37]. In addition, these technologies ease the process of clinical trial recruitment by offering the prospect to connect, enroll, and allocate appropriate patients into clinical trials matching their requirements on an anonymous basis, securing their personal sensitive information [27]. Therefore, clinical research teams are increasingly focused on choosing the appropriate clinical trial participants according to the settled inclusion and exclusion criteria, leading to the successful conduct of a study as well as protecting participants' personal sensitive information [38]. In this direction, pharmaceutical companies have turned their investment interest on AI technologies and big data analytics or blockchain as they

XSL•FO
RenderX

provide novel solutions by reducing costs and leading the research and development process at the same time toward innovative new paths [39]. Moreover, the COVID-19 outbreak seemed to play a significant role in this transition from manual to virtual clinical trial functions as the remote conduct of clinical trial stages was essential and increased during this period [40,41]. In fact, the COVID-19 pandemic acted as a driving force to simplify the processes of clinical trials through the introduction and development of many important features in electronic format that play a crucial role in clinical trial implementation, such as web-based recruitment, e-consent, and electronic patient matching. These features play a crucial role in clinical trial implementation by improving operational efficiency ensuring benefits for participants [42].

One of the most important principles in medical ethics and the conduct of research following good clinical practice guidelines is the concept of informed consent [43]. Informed consent is required to clearly provide all the details concerning a clinical study to be read carefully and be well understood by all the participants willing to enroll in a clinical trial [44,45]. It serves 2 crucial purposes in clinical research: it promotes and respects participants' willingness and decision-making regarding enrollment in a clinical trial and also protects individuals from exposure to potential harm [46]. To proceed with the enrollment of participants or patients in clinical trials, written informed consent is an internationally acknowledged requirement [47]. The key elements of informed consent include voluntariness, capacity, disclosure, understanding, and decision [48]. Informed consent in clinical research requires that participants make voluntary decisions by having the capacity to understand well the information provided and be fully informed of all the stages and the aim of the research [49]. Recent studies have further investigated a novel format of this feature called electronic consenting (e-consent). However, it is important to note that electronic consenting is not a new phenomenon. e-Consent has been used in clinical trials for approximately 15 years. Initially adopted slowly, its use has grown significantly due to its benefits in improving patient understanding, enrollment speed, and data accuracy [50]. The term e-consent includes the legally guaranteed participation of patients and volunteers in clinical trials by following accelerated signature processes and avoiding time-consuming submission of files and documents as it is currently used in traditional in-clinic recruitment strategies, known as offline recruitment [51]. The future of clinical research is strongly associated with the use of innovative technologies and novel features that can play an important role in leading to the digitalization of research and development [52].

Another innovative characteristic in clinical trials is the continuous monitoring of patients [3]. It is a critical process that refers to the actual monitoring of patients' health state and providing health care professionals and researchers with access to important insights and information about patients, such as vital signs, therapeutic response to treatments or medical interventions, and compliance with medical therapeutic schemes, aiming at advanced therapeutic approaches and overall improved patient outcomes [4]. Recent studies have shown that there is a notable dropout rate in clinical trials (30%), raising concerns [22]. Thus, patient monitoring and health care professional counseling can reduce this burden by improving early detection of potential problems and, consequently, minimizing patient dropout rates in ongoing clinical studies [53]. The use of advanced information and communications technologies in combination with novel medical device technologies offers new approaches leading to the development of power-efficient, real-time, and personalized patient-monitoring systems in mobile devices [54]. Technological innovations for advanced measurements like electrocardiography, pulse oximetry, and noninvasive hemoglobin monitoring, as well as the use of smartphones, provide the ability to remotely monitor patients effectively and with an approximate accuracy [55,56]. In summary, the integration of patient health–monitoring features into clinical studies has the potential to significantly elevate the quality and efficiency of the digital tools and innovative approaches designed to conduct clinical trials in a novel digital health care era [17].

Objectives

The objectives of this study were to (1) provide a comprehensive overview of existing innovative digital tools such as web platforms and applications that focus on patient recruitment; (2) list the tools that also incorporate the continuous monitoring of patients' health status; (3) illustrate how technology-based interventions can ensure and facilitate the matching of patients to the appropriate trial based on individual characteristics and specific inclusion criteria; (4) present the advantages, challenges, and limitations of these novel technology tools and features in clinical trial recruitment strategies; and (5) highlight the future directions regarding the use of e-technologies and novel features in clinical research.

Methods

Study Design

A review with a systematic research approach was conducted. The research team searched the MEDLINE, ScienceDirect, Scopus, and Google Scholar databases to identify relevant studies from January 2000 to October 2024. The search strategy devised for MEDLINE was refined to encompass various combinations of terms related to clinical trials, clinical research, and digital methodologies. Specifically, the search string used was as follows: *{Clinical trials OR Clinical research OR clinical studies} AND e-consent, {Clinical trials OR clinical research OR clinical studies} AND e-matching, {Clinical trials OR clinical research OR clinical studies} AND e-recruitment, Clinical trials AND web platforms, Clinical research AND web platforms, Clinical Research AND mobile applications, Clinical Trials AND artificial intelligence, Clinical Trials AND machine learning*. The aforementioned string was adjusted accordingly for other databases using a combination of Medical Subject Heading algorithms and keywords such as *clinical trials*, *clinical research*, *online recruitment*, *patient matching*, *electronic consent*, and *digital platforms*. Relevant studies were identified and selected through citation searches, as well as by reviewing abstracts, full texts, and peer-reviewed articles.

XSL•FO
RenderX

Inclusion and Exclusion Criteria

The inclusion of manuscripts in this study was based on the criteria listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion Criteria

- Timeline of companies' operation concerning the digital platforms and tools of interest from 2000 to 2022
- Digital platforms and tools operating internationally without geographical limitations
- Digital tools and platforms that have been scientifically validated through relevant research study publications or peer-reviewed articles in scientific journals
- Digital platforms and tools that retain approval to function, demonstrated through testing and use in clinical trial processes
- Digital platforms and tools with approval from relevant regulatory authorities concerning data safety and transparent function
- Digital tools and platforms in English or with translated websites (and with manuscripts also available in English)
- Digital platforms and tools created by entities from the pharmaceutical industry or academia
- Digital platforms and tools focused on the clinical trial recruitment phase, including features related to patient matching or electronic consent

Exclusion Criteria

- Literature reviews
- Digital platforms or tools designed and launched for future testing and operation but not yet validated through scientific publication in peer-reviewed journal

Study Selection and Data Extraction

In total, 2 independent investigators (AGB and FD-G) initially screened the articles, with any discrepancies resolved through consensus at each step with a third independent investigator (ED). The search results and outcomes were imported into the reference management software Mendeley (Elsevier) for recording, managing, and generating the reference list, including the removal of duplicates. Finally, the findings of the studies were evaluated, summarized, and recorded. All references from the selected studies were retrieved and manually reviewed using the snowball effect [57].

Results

Comparison of Platforms for Recruitment

A flowchart detailing the study selection procedure is presented in Figure 1. Table 1 shows the innovative digital platforms used for clinical trials focusing on the recruitment stage, with the features that they provide, their main characteristics, whom the digital tool is addressed to, and the source from the literature review. These digital tools according to their characteristics were classified into 2 main categories considering the way in which they were launched as solely digital web platforms or mobile apps or web applications. They were also categorized according to whom they were addressed to for use, such as patients or sponsors (eg, health care practitioners, industry professionals, or researchers involved in clinical trial conduct). The key attributes outlined here play a pivotal role in the participant recruitment phase for clinical trials, specifically focusing on the features of electronic consent, web-based recruitment, and patient-matching technologies. In addition, these digital platforms and tools are characterized by several core aspects: the start year of the companies behind these innovations, the total count of patients or users engaged via these technologies, and the geographic presence of each digital tool.

XSL•FO
RenderX

Figure 1. Flow chart showing the study selection procedure.

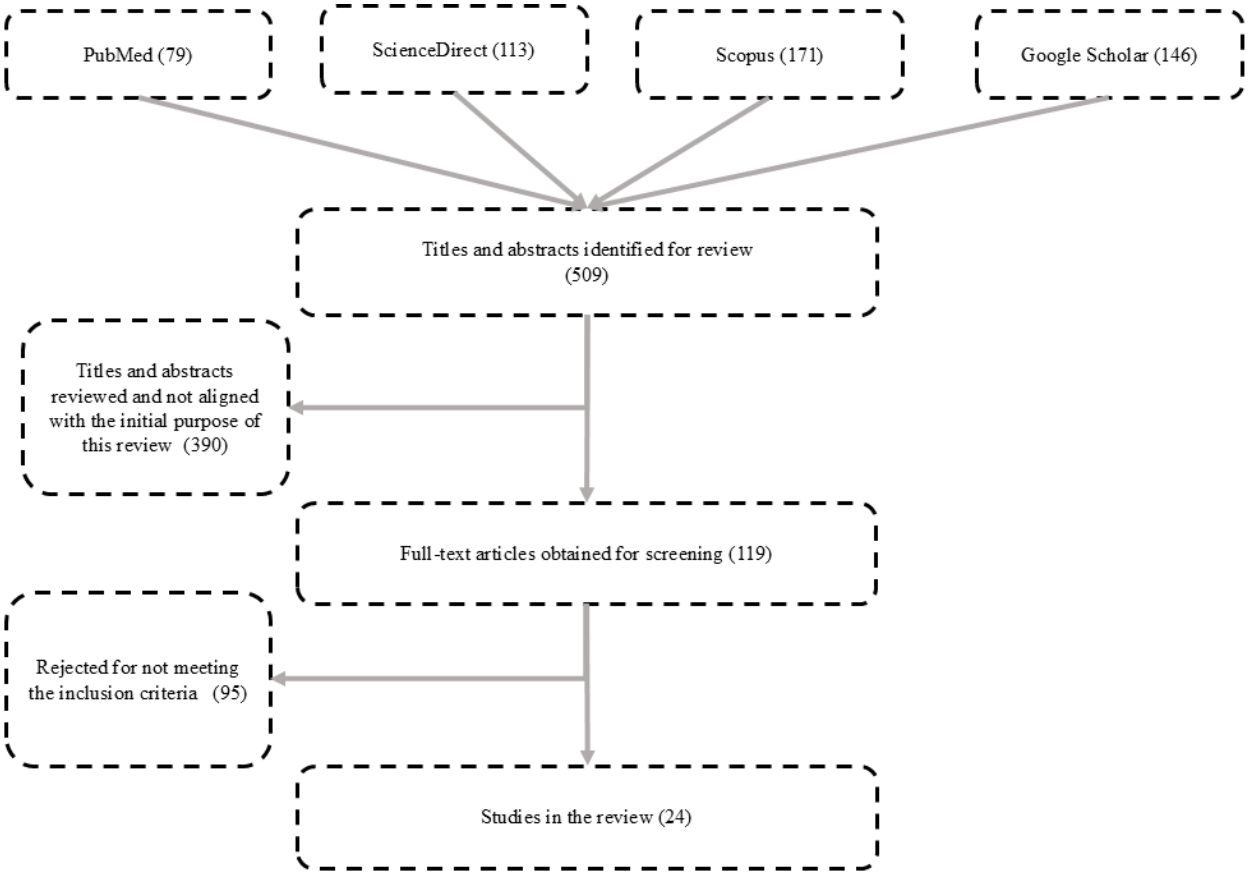

After the classification of the platforms, further investigation and review of the literature was carried out to determine the capabilities of each platform, organize their characteristics into groups, and indicate their pros and cons. The focus was on specific criteria, such as some potential limitations in geographic operation areas, therapeutic fields, approach toward patients or sponsors, educational material availability, patient engagement, scientific evidence, the tools' ease of use, and financial-related barriers.

Table 2 shows the respective competencies and constraints of each platform identified in Table 1.

XSL•FO
RenderX

Table 1. Digital tools supporting the clinical trial recruitment stage and their features and main characteristics.

Name	Type of digital tool	Addressed to patients or sponsors	Main features (e-consent, matching, and e-recruitment)	Platform characteristics (company start year, number of users or patients enrolled, and geographic location)
Accelerated Enrollment Solutions [58]	Digital platform	Sponsors	Matching and e-recruitment	2018; 100-million-patient database (20 million prescreened); United States
Antidote [59]	Digital platform	Patients and sponsors	e-Consent, matching, and e-recruitment	2016; approximately 1 million users; United States
BreastCancerTrials [60]	Digital platform	Patients and sponsors	Matching and e-recruitment	2008; number of users not available; United States
Clara Health [61]	Digital platform	Patients and sponsors	Matching and e-recruitment	2016; 10,000 patients in matching and 50,000 in clinical trials; United States
Clinical Connection [62]	Digital platform	Patients	e-Consent, matching, and e-recruitment	2000; approximately 850,000 patients; United States
Elligo [63]	Digital platform	Sponsors	e-Consent, matching, and e-recruitment	2016; approximately 150 million patients and approximately 1000 studies; United States
Inspire [64]	Application and digital platform	Patients and sponsors	e-Recruitment	2005; approximately 2 million members; United States and India
KU[a] Cancer Center Clinical Trial Finder [65]	App	Patients	Matching and e-recruitment	2020; approximately 300 app downloads in 2022; United States
Labcorp—Xcellerate Trial Management [66]	Digital platform	Sponsors	e-Recruitment	2011; ≥65,000 patients internationally; United States
MatchMiner [67]	Digital platform	Sponsors	e-Consent, matching, and e-recruitment	2016; approximately 200 patients enrolled; United States
Mendel Health [68]	Digital platform	Patients and sponsors	Matching	2016; approximately 1 million patient health records; United States
ResearchMatch [69]	Digital platform	Patients and sponsors	Matching and e-recruitment	2009; 144,340 volunteers; United States
Trialbee (Hive and Honey) [70]	Digital platform	Patients and sponsors	e-Consent, matching, and e-recruitment	2010; approximately 500,000 patients; Sweden
Trialfacts [71]	Digital platform	Patients and sponsors	Matching and e-recruitment	2006; approximately 20,000 patients and 600 studies; United States

[a]KU: University of Kansas.

XSL•FO
RenderX

Table 2. Platforms' strengths and limitations.

Name	Pros	Cons
Accelerated Enrollment Solutions [58]	• Patient-centric approach; strong demographic, psychographic, and behavioral information available • Various clinical research therapeutic areas • Patient engagement via regular prescreening processes and site follow-ups • Telehealth options • Education and information available about clinical trials and patients' treatment options • Advanced data modeling system predicting patients most likely to enroll in each clinical trial	• A strong presence mainly in the United States but also operating worldwide, with sites in many countries • Primarily targeted at sponsors to enhance patient recruitment
Antidote [59]	• Patient-centric approach and user-friendly platform design • Provision of educational podcasts and webinars • Community engagement through encouraging patients to share their health journeys • International reach, with operating offices in the United Kingdom and the United States	• Despite the educational character on clinical trial awareness, no focus on sharing publicly detailed information on clinical trial outcomes for a more holistic perspective • There are clinical trials that are not listed on the platform and may need to use additional resources
BreastCancerTrials [60]	• Patient-friendly platform with plain language • Clinical trials related to all stages of cancer • Emphasis on the rights of patients regarding clinical trial participation; strong medical ethics • Provision of educational resources for patients (scientific articles and educational videos)	• Geographic limitations as it is a platform primarily focusing on trials available in the United States • Focusing mainly on clinical trials of a specific cancer type (breast cancer) • A search engine greatly supporting recruitment but not guaranteeing clinical study enrollment, which is particularly crucial for cancer patients, as eligibility often depends on their cancer stage.
Clara Health [61]	• Easy to use • Patient engagement • Focus on public health challenges; supported clinical trials for diagnostics, treatment, and vaccine discovery during the COVID-19 pandemic • Personalized support • Podcast series about clinical trial experiences	• Limited to certain types of medical treatments depending on available partnerships and network • Limited integration with other clinical tools or software, such as electronic medical records or laboratory systems • May pose some cost barriers according to insurance types
Clinical Connection [62]	• Patient-centric approach • Telemedicine options • Wide range of clinical trials covering a variety of therapeutic areas • User-friendly digital environment • Options for remote clinical trial participation	• Potential geographic limitations; primarily operating in the United States, but there may also be trials extended internationally • Limitations in matching of patients with clinical trials based on their personalized characteristics
Elligo [63]	• Strong connections with clinical sites and physicians • Highly technologically oriented, keeping electronic health records and simplifying clinical trial processes • Enhanced data management tools and high-quality services for sponsors	• Potential compatibility issues due to advanced technology features in case of concurrent use of other tools by sponsors or researchers • Potential financial barriers for some advanced, high-technology services • Electronic health record system may pose some privacy concerns for patients or users • Limited direct engagement with patients

XSL•FO
RenderX

Name	Pros	Cons
Inspire [64]	• Acceleration and enhancement of enrollment stage • Support for the recruitment process by focusing on location, condition, and demographics of participants • Educational character, providing information on clinical trial awareness • Wide range of trials covering different therapeutic areas • Support for patients at decision-making points • Community engagement; sharing patients' and caregivers' thoughts and experiences • Offer of services to sponsors, such as real-world evidence data, patient insights, and digital advertising options • Available in the United States and India, offering potential for collaboration with underrepresented and vulnerable populations	• Limited description of clinical trial recruitment features; mainly functioning as a patient and caregiver community platform
KU[a] Cancer Center Clinical Trial Finder [65]	• Academic research oriented • Prompt allocation of patients to clinical trials • "Make a trial referral" feature for HCPs'b and physicians' ease in their practice • A cancer-focused institution • Patient support features	• Potential geographic limitations due to the operation of the center within the University of Kansas area in the United States • Primarily targeted only to patients with cancer, although it is a cancer research institution • Mainly available via app format; some restrictions in use
Labcorp—Xcellerate Trial Management [66]	• Strong global perspective (≥30 countries involved) • Diversity in therapeutic areas • Targeted patient recruitment and engagement with the clinical trial process • Personalized and user-friendly experience • Predictability; identifying potential issues before they occur in a study • Medical review evaluation	• Potential challenges in integration compatibility with some other electronic systems and software
MatchMiner [67]	• Provision of educational material to patients about clinical trials • Option for specific genomic profiles related to cancer precision medicine trials • Great focus on scientific publications • Easy access to the platform without requiring registration or log-in • High rates of successful matching • Use of open-source AIc technologies	• No direct engagement with patients but still facilitates interaction between clinical trial sites and patients • Might pose some technology barriers with software complexity for some users with less technical experience
Mendel Health [68]	• Outcome validation through AI "intelligent trials" • Use of deep learning technology • Continuous update of results • Focus on real-world evidence analysis and data	• Currently focused solely on oncology and populations of patients with cancer • Focus primarily on automated eligibility and screening rather than patient recruitment strategies
ResearchMatch [69]	• High quality of demographic data and support for diversity of participants • Strong scientific impact • Option for children registration • Flexibility in patient enrollment; no health insurance required	• Geographic limitations; only in the United States • Mainly focusing on processing the matching stage of a clinical trial; further effort required for the filtering process regarding eligibility criteria for clinical trial participation
Trialbee (Hive and Honey) [70]	• Global-oriented platform • Focus on patient engagement • Patient oriented; enhancing the success of the matching process	• Mainly focused on sponsors • Some potential limitations in clinical trial options; mainly listing sites or trials that have established partnerships with the platform
Trialfacts [71]		• Potential barriers with advanced high-cost services • Focusing mainly on sponsors' needs; not a patient-centric platform but making efforts for more accessible trials for patients and engagement of stakeholders

XSL•FO
RenderX

Name	Pros	Cons
	• Offers educational resources • Promotion of communication or collaboration among researchers • Extensive experience in clinical trial management • Combination of on-site clinical trials and virtual-based studies • Access to a global network of sites and patients • Consulting services to researchers on how to avoid recruitment errors	

[a]KU: University of Kansas.

[b]HCP: health care professional.

[c]AI: artificial intelligence.

Description of the Digital Tools' Characteristics

Accelerated Enrollment Solutions [58] is a cutting-edge digital platform designed to enhance patient recruitment and increase the successful rate of patient matching in clinical trials. Accelerated Enrollment Solutions offers a diverse range of therapeutic areas and features, including patient engagement, screening processes, location tracking, and telehealth options. The platform also provides advanced data modeling that predicts the likelihood of patient enrollment. However, it should be noted that this digital tool mainly operates in the United States, posing some geographical limitations, and focuses on sponsors (ie, health care providers and researchers) rather than patients.

Antidote [59] is a patient-centered platform that offers e-recruitment, matching, and e-consent features. The platform is easily accessed through web-based searches. One of its characteristics is the emphasis on patient education and information, which is achieved through informative interactive resources such as podcasts, webinars, and videos featuring patient stories. Patients widely accept Antidote for its convenience and user-friendly interface. However, despite the tool's educational character, there is limited information available regarding clinical trial outcomes for a more holistic comprehension of the clinical study process.

BreastCancerTrials [60] is a user-friendly platform providing features such as e-recruitment and matching in clinical trials focusing on all the stages of breast cancer, prioritizing patient rights, and providing educational content that contributes to patients' informed decisions regarding their participation in clinical trials. There was no clear evidence of the existence of an electronic consent feature on the interface of BreastCancerTrials. Its main limitations lie in both geographic location (clinical trials conducted mainly in the United States) and type of disease (only breast cancer). In addition, the digital platform system is primarily a search engine and may pose some limitations on up-to-date information, especially for eligibility criteria for clinical trials. Hence, patient enrollment is not guaranteed, which poses a significant barrier for patients with cancer due to challenges concerning their cancer stage and time.

Clara Health [61] is a user-friendly digital tool that has a patient-centric approach providing personalized support. It has released a limited podcast series delving into the clinical trial experience. It is an innovative platform that aims to recruit patients using advanced technology tools and match them to the appropriate clinical trial, although there is no clear information available on whether the e-consent feature is provided by Clara Health. One of its unique features is its focus on public health issues. In the case of the COVID-19 pandemic, Clara Health was very active regarding clinical trials for diagnostics, treatment, and vaccine discovery. Nevertheless, it is worth noting that this platform covers certain types of diseases based on available network partnerships and has limitations in integration with other advanced clinical software, such as electronic medical records or laboratory systems.

Clinical Connection [62] is a digital patient-addressed tool that offers electronic patient recruitment, matching, and electronic informed consent features. It promotes patient engagement and health care provider interaction and enables remote participation in a wide range of clinical trials across various therapeutic areas. Despite its vast potential, Clinical Connection has limited use as it operates primarily in the United States, posing some geographical limitations for patients, but may also offer extended trials internationally. Unlike some other platforms, it lacks a personalized matching function for participants to be promptly allocated to trials based on their personal medical history and preferences.

Elligo [63] as a cutting-edge platform addressed to sponsors for clinical trial processes leverages electronic health records and offers seamless connectivity with clinical sites and physicians, advanced data management capabilities, and high-quality services to sponsors. The platform's e-recruitment and matching features, as well as its e-consent functions, demonstrate its commitment to streamlining the clinical trial process and enhancing patient engagement. The platform's vast reach of approximately 150 million patients demonstrates its potential impact on clinical research. However, there are also potential limitations and challenges to this platform, such as compatibility issues with other advanced software and cost barriers due to its high technological orientation. Some patients may also have privacy and security concerns due to the use of electronic health records.

Inspire [64] is a user-friendly digital tool that enrolls patients for clinical trials based on their location, condition, and demographics, covering various therapeutic areas. It offers educational resources and supports patients at crucial decision-making points from diagnosis through treatment. The

XSL•FO
RenderX

operations of Inspire in India and the United States enhance collaboration and research with underrepresented and vulnerable populations. Furthermore, it may also present some limitations due to the direct marketing involvement of brand partners on pharmaceutical options and products. Moreover, while it can be used for participant recruitment in clinical trials, it is primarily designed as a patient community platform, which may have some drawbacks compared to other digital tools in the recruitment stage of clinical studies.

University of Kansas Cancer Center Clinical Trial Finder [65] is one of the first observed research-oriented applications that successfully aims to recruit, guide, and match patients with cancer with the most suitable clinical trial for them according to their needs. This platform explicitly states that its target user group includes only patients. The application includes features such as "Make a trial referral" to support health care professionals and physicians in their practice. However, it has certain use limitations as it can be used only as an application and it is addressed also to a specific patient group. Moreover, it may present geographical limitations as it only addresses cancer patients based in the USA. This may also contribute to the limited total number of enrolled patients in the app (300 downloads).

Labcorp's Xcellerate Trial Management [66] tool is a widely used digital platform that includes >30 countries and targets >50 therapeutic areas. It offers personalized and user-friendly patient recruitment features along with an interactive patient community to direct sponsors (health care professionals and researchers) toward clinical trial success. The platform provides medical review evaluation to ensure patient safety monitoring and identify potential issues before they occur. However, compatibility issues with other advanced technological software may arise. It is important to note that Xcellerate Trial Management primarily focuses on clinical trial management and may not be as effective in other stages of clinical trials, such as patient recruitment or data analysis.

MatchMiner [67] use open AI sources to facilitate patient recruitment for clinical trials, with a special focus on cancer precision trials, including the option to investigate specific genomic profiles, advancing personalized medicine approaches. It also serves as an educational resource for patients, emphasizing as well the publication of scientific articles. The platform is easily accessible, with no registration or log-in requirements, but it does not involve direct patient engagement. MatchMiner offers all the features concerning the recruitment process, such as e-recruitment, matching, and e-consent. There may be platform use barriers and software implementation issues for some users with less technological experience.

Mendel Health [68] uses AI and natural language processing technology to validate its recruitment outcomes through the "intelligent trials" feature, with a focus on analyzing real-world evidence and data. It is the only included platform that offers the matching feature as its main service. Using deep learning technology, the web-based platform offers features that include searching medical literature and patient health records and suggesting evidence-based treatments. Despite its reliance on advanced technology, Mendel Health has a focus solely on oncology and populations of patients with cancer. Its approach is centered on automated eligibility criteria and patient screening strategies rather than on traditional patient recruitment methods.

ResearchMatch [69] is a digital tool that mainly focuses on the recruitment and matching stages of clinical studies. It integrates high-quality demographic data into the recruitment process, ensuring participant and patient diversity, including the incorporation of children as well. Moreover, the platform is financially affordable for researchers and does not require health insurance documentation during patient enrollment. However, it is important to note that, currently, only individuals living in the United States and Puerto Rico have access to this platform, posing geographical limitations. In addition, ResearchMatch excels mainly at the matching stage of clinical trials; hence, researchers need to make organized efforts to filter eligible patients based on study-specific inclusion criteria. It is not clearly indicated whether any type of electronic consent is provided through this tool.

Trialbee (Hive and Honey) [70] is a patient-centric platform that possesses all 3 features concerning the recruitment process (e-recruitment, matching, and e-consent). With a global reach that leverages patient information, Trialbee endeavors to improve the matching process. While it targets patient needs, Trialbee primarily serves health care providers and researchers, mainly referring to sites or trials with whom it has established partnerships, potentially excluding other relevant clinical trials available.

Trialfacts [71] is a comprehensive platform that focuses on the preparatory stages of conducting a clinical study, emphasizing patient recruitment and matching processes. One of its primary objectives is to combine on-site clinical trials with virtual-based studies. The platform provides valuable educational resources and promotes strong interaction among researchers. With extensive experience in clinical trial management, Trialfacts offers access to a global network of sites and patients. It provides consulting services to researchers to avoid recruitment errors. However, it does not guarantee clinical trial enrollment, and the platform lacks direct patient outreach as a function.

Comparison of Platforms for Recruitment That Emphasize Monitoring Features

Table 3 shows digital tools that additionally provide the function of participant or patient monitoring. Their main features, characteristics, and the enrolled patients and users are shown. Monitoring entails tracking the patients' health status (compliance with treatments or interventions and vital sign measurements) during the trial and following its completion. The vital signs of patients are recorded through wearable sensors, also linked with other fitness- or health-tracking apps, and online diaries that focus on medication adherence. These tools enable health care professionals and researchers to monitor patients' progress accurately, identify potential adverse effects, and assess the effectiveness of the intervention in a clinical study. The monitoring features of these digital platforms are crucial in ensuring the safety and well-being of patients during and following a clinical study, contributing to the overall success of a clinical trial.

XSL•FO
RenderX

Table 3. Digital tools facilitating the recruitment stage of clinical trials that incorporate the feature of monitoring patients' health status.

Name	Type of digital tool	Addressed to patients or sponsors	Main features (e-consent, matching, and e-recruitment)	Platform characteristics (company start date, number of users and patients enrolled, and geographic location)
Advarra [72]	Digital platform	Sponsors	e-Consent and e-recruitment	2020; approximately 56,000 participants; United States
Castor [73]	Application and digital platform	Patients and sponsors	e-Consent, matching, and e-recruitment	2012; approximately 7 million patients and 147,000 users; United States and EU[a] (Amsterdam)
Clinical Ink [74]	Application and digital platform	Sponsors	e-Consent and e-recruitment	2020; approximately 2 million patients; United States
Clinical Research IO [75]	Digital platform	Sponsors	e-Consent and e-recruitment	2015; approximately 7 million patients; United States
Clinpal [76]	Application and digital platform	Sponsors	e-Consent, matching, and e-recruitment	2012; approximately 10,000 patients; United Kingdom
Spark [77]	Digital platform	Sponsors	e-Consent and e-recruitment	2009; number of users and patients not available; United States
Eureka [78]	Application and digital platform	Patients and sponsors	e-Consent and e-recruitment	2018; 83,000 participants (during the COVID-19 pandemic); United States
Massive Bio [79]	Application and digital platform	Patients and sponsors	Matching and e-recruitment	2015; ≥100,000 patients; United States
Medable [80]	Digital platform	Sponsors	e-Consent, matching, and e-recruitment	2012; approximately 1 million patients; United States
Veeva Systems [81]	Application and digital platform	Patients and sponsors	e-Consent, matching, and e-recruitment	2007; number of users and patients not available; United States

[a]EU: European Union.

Table 4 shows the respective competencies and constraints of each platform identified in Table 3.

XSL•FO
RenderX

Table 4. Platforms' capabilities and constraints.

Name	Pros	Cons
Advarra [72]	• Telehealth, remote data collection, and virtual patient visits • Enhanced engagement through mobile-compatible tools for remote participation	• No information about patient-matching feature, potentially limiting recruitment capabilities
Castor [73]	• A robust e-consent system and participant recruitment process • User-friendly electronic data capture • Support for decentralized clinical trials with tools for remote patient monitoring and ePROsa • Strong patient engagement (>7 million patients and 147,000 users) • Global presence	• May require additional integration of tools for trial management • New users may face barriers with more advanced features
Clinical Ink [74]	• Advanced e-consent and e-recruitment • Support for decentralized and hybrid clinical trial real-time data capture; advanced monitoring via sensors, wearables, and televisits • Strong focus on patient engagement	• Limited information on matching features • Technology complexity • Potential connectivity issues due to cloud-based functionalities • Primarily patient focused; further tools may be needed for sponsors • Limited global reach; mainly in the United States
Clinical Research IO [75]	• A robust participant recruitment process, simplifying enrollment • Support for decentralized clinical trials with tools for remote patient monitoring and ePROs • Advanced patient-matching tools and comprehensive monitoring • Strong patient engagement	• More focused on sponsors • Limitations in financial tracking–related issues for clinical trial conduct
Clinpal [76]	• Provision of learning and advanced training mechanisms for researchers • Suitable for remote (virtual), hybrid, and direct-to-patient clinical studies • Cloud-based software as a service platform • Easy-to-use platform • Real-time compliance or performance metrics through the integration of external technology devices • Support for patients in the overall clinical trial management process with engagement and reminder mechanisms	• More focused on sponsors • The app is not widely used • Potential technical limitations
Spark [77]	• Limitations in financial tracking–related issues for clinical trial conduct	• More focused on sponsors • Limited detailed information on specific patient-matching features • Limited updated information on clinical trial outcomes and number of users enrolled
Eureka [78]	• Academic research oriented • Remote monitoring • Creation of electronic medical records • Focus on rare diseases and minority populations • Active research during public health crises such as the COVID-19 pandemic • Global perspective; involving international patients • Provision of personalized health services • Strong publication record on high-impact scientific journals • Educational material and podcasts available	• Matching feature not clearly described • Potential technical limitations

XSL•FO
RenderX

Name	Pros	Cons
Massive Bio [79]	• Strong engaging community between patients and health care professionals • A digital environment for personalized and precision medicine solutions • Compliance with and security certifications from regulatory authorities (eg, EMAb) • AIc advanced technology for the matching stage (patient prescreening) and data analysis	• Geographic limitations; mainly in the United States • Offer of clinical trials for patients with cancer; no variety regarding other diseases • Potential barriers due to costs
Medable [80]	• Comprehensive e-consent solution • Efficient recruitment and patient matching • Integration with wearables and sensors • Patient engagement features	• More focused on oncology and vaccines
Veeva Systems [81]	• Strong global reach; operates in key regions worldwide • Robust patient engagement tools • Integration flexibility with other clinical trial management systems • Decentralized clinical trial support via advanced e-recruitment	• Services and advanced tools that may be challenging; potential complexity in navigation for new users • Potential financial barriers

[a]ePRO: electronic patient-reported outcome.

[b]EMA: European Medicines Agency.

[c]AI: artificial intelligence.

Description of the Characteristics of the Digital Tools That Emphasize Monitoring Features

The Advarra [72] platform is a comprehensive digital tool designed to support decentralized clinical trials. It provides a wide range of features, including e-consent, telehealth, remote data collection, and virtual patient visits, enabling clinical trials to be conducted seamlessly across multiple locations. These capabilities help improve engagement with participants and streamline trial management for stakeholders. However, a limitation of the platform is that it does not include a dedicated patient-matching feature, which could hinder recruitment efficiency in certain studies.

Castor [73] is a comprehensive platform dedicated to clinical trial management, offering advanced features for e-consent and patient recruitment and monitoring. It is a useful tool for sponsors and patients. According to its records, it may support >7 million patients and 147,000 users. Its tools offer advanced support for decentralized trials, patient matching, electronic patient-reported outcomes, and remote monitoring. However, while Castor excels in patient engagement and data capture, more advanced tracking activities may require integration with other tools. It has a global presence, with offices in New York and the European Union (Amsterdam), providing reliable support for trials across multiple regions.

Clinical Ink [74] is a robust clinical trial platform with advanced features in patient recruitment, engagement, and monitoring with cutting-edge technology such as e-consent, sensors, wearables, and televisits. Despite being a relatively new platform, it serves approximately 2 million patients. The platform's strengths are optimizing patient experience, supporting decentralized clinical trials, and enabling real-time data collection through wearables and sensors. It has a patient-centered approach, but it may require some additional tools for overall clinical trial management and especially for sponsors. This platform shows a great potential for supporting clinical studies, although it is primarily available in the United States, posing some geographic limitations.

Clinical Research IO (CRIO) [75] is a comprehensive digital platform designed to enhance the management and execution of clinical trials. The platform offers a wide array of features, including centralized patient recruitment, advanced e-consent forms, and tools that foster strong patient monitoring. These capabilities help streamline operations and improve data quality for single sites, sponsors, and academic partners. While CRIO excels at simplifying workflows for these stakeholders, there are limited references to patient-matching features. Despite this, CRIO remains a valuable solution for clinical trial management.

Clinpal [76] is a digital tool that is accessible on web and mobile platforms, providing e-consent and health-monitoring features in addition to its other necessary existing functions for clinical studies. It tracks real-time compliance with treatments or interventions but also performance metrics and integrates with external devices for clinical trials. It also offers remote, hybrid, and direct-to-patient clinical study participation through a cloud-based software platform and assists patients with engagement and reminders for the overall clinical trial process. Clinpal focuses mainly on researchers and health care providers. Relative bias may occur due to the nature of self-monitoring, the sensitivity of wearable sensor measurements, and the reliance on self-reports from patients.

DrugDev's Spark [77] platform offers a unified suite of solutions aimed at optimizing clinical trial operations. Its key features include e-consent, site selection tools, and data-tracking tools such as eTrackers that automate patient retention, enrollment diversity monitoring, and trial management processes. There is

XSL•FO
RenderX

limited transparency on the exact number of patients enrolled through the platform, and it is unclear whether the platform provides specific patient-matching features.

Eureka [78] is an academic-oriented platform that aims to enhance clinical studies of rare diseases in minority populations. This platform has been actively facing public health challenges such as the case of the COVID-19 pandemic, with their research involving approximately 83,000 participants (enrolled number). It aims to enhance patient recruitment stages and provides e-consent features to accelerate enrollment processes. It also supports the remote monitoring of clinical trial participants or patients, data collection from wearables and apps, and the creation of electronic health records, advancing a personalized medicine approach to clinical trials. Eureka has an impressive record of publications in high-impact scientific journals and available educational material and podcasts concerning clinical trials. However, the recruitment process's efficacy may have limitations in some aspects due to the unclear statement of the matching feature.

Massive Bio [79] is a commonly used e-recruitment and matching digital tool with a considerably large number of participants. It uses AI technology to assess patient data, identify individuals who meet the eligibility criteria for specific clinical trials, and conduct comprehensive data analysis to streamline the recruitment process. It provides personalized medicine and a precise environment for patients, including the exchange of patient experiences and their journeys through its Patient Ambassador Program. Massive Bio is compliant with and certified by regulatory authorities such as the European Medicines Agency, but it is solely available to the US population and targets patients with cancer.

The Medable [80] platform is a unified, cloud-based solution designed to support decentralized, hybrid, and traditional clinical trials. It offers a comprehensive set of features such as e-consent, electronic clinical outcome assessments, virtual patient visits, and e-recruitment. However, while the platform supports recruitment indirectly, it does not offer a dedicated patient-matching feature, which may limit its recruitment capabilities for some studies.

Veeva Systems [81] is a comprehensive platform for clinical trials with a strong global presence in different continents, such as North and South America, Europe, and Asia, offering a variety of tools for e-recruitment, patient engagement, and decentralized clinical trial support. However, although Veeva Systems is rich in features, it can also be quite complex in navigation, requiring more advanced use, particularly in the case of e-consent. Its integration capabilities and global reach make it ideal for large-scale multicenter trials, though its cost can present a potential challenge

Discussion

In this section, our study findings, strengths, and limitations are discussed. In addition, concept technological solutions are presented based on blockchain and AI, along with the perspectives that support them.

Comparison of the Digital Platforms

The platforms presented in this study and their features refer to a specific time frame when the investigation took place. According to our findings, there is an adequate number of digital platforms (N=24) focused on orienting the early stages of clinical trials, specifically focusing on matching, e-consent, and patient recruitment. Most of these digital platforms (n=22, 80%) are headquartered in the United States, but most of them (n=18, 75%) also provide global access to their features, whereas 25% (n=6) primarily provide services within the United States. These platforms attempt to provide a well-informed digital environment for patients and simplify the clinical trial processes. This review revealed that 8% (n=2) of the platforms have enrolled >100 million participants, 29% (n=7) have >1 million participants, 4% (n=1) have approximately 200 participants, and participant numbers were unclear for 17% (n=4) of the platforms. Outcomes indicated that these digital platforms and applications have an adequate number of participants, representing an acceptable sample size, also assisting in the successful conduct of clinical trials in rare diseases. It was also shown that many of the platforms (n=14, 60%) have clearly incorporated an electronic consent form as a feature that promotes autonomous decision-making and simultaneously protects human rights. In addition, most of the digital platforms seem to serve different purposes and scopes for both researchers or sponsors and patients. More specifically, 8% (n=2) of them leverage patient demographics to achieve an accurate match of patients with clinical studies, whereas 38% (n=9) of the platforms offer sponsors or researchers a clinical trial management plan for the conduct of their research. Furthermore, 50% (n=12) of the platforms offer dual functionality, addressing the needs of both patients and sponsors, providing a more integrated approach to managing clinical trials. Most of the platforms (n=10, 42%) incorporate educational features, providing valuable information to patients and explaining the purpose of clinical trials, encouraging the interaction and exchange of ideas in the patient community. The platforms differed in scope, with 21% (n=5) focusing primarily on oncology, whereas others addressed broader populations. The flourishing and increasing use of such digital tools appeared to arise since the COVID-19 outbreak, when the barriers that were created due to the lockdown restrictions posed a great burden to the efficient conduct of clinical studies. In contrast, while these digital tools may offer multiple benefits for a digital new era in clinical trials, such as increased patient participation, faster progression through clinical trial stages, and personalized monitoring features there are also many challenges. Accessibility, compatibility, and accuracy are just a few examples of issues that can arise. Older individuals and those unfamiliar with newer technologies may have limited access to and use of these tools. Furthermore, due to the prevalence of the development, design, and availability of these digital tools in specific geographic regions such as the United States, cultural and language barriers may arise and limit the access and use of these technologies in other regions of the world. Another challenge is the potential for transparency issues if e-consent features are unclear and not accompanied by detailed information on all patient rights. Finally, it is important to note that integrating telehealth options into some of these clinical

XSL•FO
RenderX

research tools entails risks, including the potential for invalidity as self-monitoring systems still need improvement. By considering carefully the integration of innovative digital tools into clinical research, researchers can maximize the benefits of these technologies while minimizing the risks and challenges associated with their use.

Novel Technologies (AI Based) and Software Underlying the Development and Functioning of Innovative Digital Platforms Contributing to Clinical Research

Advancements in digital health tools and emerging technologies play a crucial role in revolutionizing the design, development, and operation of systems supporting the clinical trial recruitment process, thereby significantly enhancing the value of these systems within the clinical research domain. Blockchain or DLT and AI combined can provide the best technical elements of both worlds. Sponsors, clinicians, and researchers can make decisions with greater ease owing to machine learning models, which can analyze massive amounts of data and offer sufficient decision-making tools. For instance, finding the best candidates and recruiting patients are crucial steps in the clinical research process. AI and machine learning reduce the time and cost of identifying these patients by analyzing both structured and unstructured electronic health record data. For example, in a clinical trial, the sponsors can send a matching request to the blockchain system, which, supported by AI, can automatically match the potential patients, saving time for principal investigators, researchers, and patients. The blockchain or DLTs can ensure the validity of a clinical trial and the sponsor's identity. In this context, patient identification speeds up and leads to improved and more accessible and efficient clinical trials [82]. On the other hand, finding the best researchers and candidates that can lead and perform a clinical trial is also a time-consuming issue that may be accelerated using deep learning techniques [83]. Identifying highly skilled and experienced investigators is crucial for successful drug development, implementing cost-effective strategies, and selecting the most appropriate patient candidates for clinical trials. Combining these values from the AI domain with the security, privacy, and transparency features of blockchain leads to advanced clinical research systems able to produce robust data management in clinical trials. Modern applications using DLTs offer maximum data management and security capabilities. More specifically, clinical trial data that are recorded on the blockchain are protected against unauthorized access and alterations or tampering. In addition, DLT features such as decentralization, immutability, data provenance, and auditability increase trust and awareness. The highest level of trust between software and users (eg, patients, physicians, sponsors, and researchers) can be achieved by relying on the data recorded in blockchain. Data integrity through transparency reinforces the trust among all the stakeholders involved in a clinical trial process [84]. Other values that blockchain contributes to clinical trial tasks are data management, analysis and reporting, patient privacy, patient retention, and regulatory compliance. Solving the problem of trust for the entire clinical trial ecosystem fosters seamless collaboration among stakeholders and people willing to collaborate easily, trusting the system in a decentralized way rather than trusting a centralized system owned by one entity. Most individuals are willing to share their medical data anonymously if privacy and security are guaranteed [85]. Community-driven clinical trials represent a new era, characterized by trusted environments, privacy by design, voluntary participation, user-controlled data sharing, anonymized raw data, secondary studies, and metadata analysis. These pillars are supported by DLT, enabling secure and transparent processes. Nonetheless, ensuring the ethical exploitation of AI and blockchain tools is yet another priority to address when using such technologies. In summary, the collaboration among AI, blockchain, and other digital health platforms able to support clinical research can reduce time-costly procedures and bureaucracy burdens, provide privacy and data control, and empower participants to be more active.

The PharmaLedger Innovative Web Platform and PharmaLedger Association

PharmaLedger is an example of how blockchain technology can be applied in various use cases of the health care domain. It has revolutionized the way data are handled and stored in ledgers and how privacy and data ownership are controlled in a decentralized way. PharmaLedger [86], as a European Union and Innovative Medicines Initiative project, exploited this blockchain technology to bring together experts in technical, legal, and regulatory fields, as well as pharmaceutical companies and patient organizations to accelerate the innovations in health care and benefit specifically the patients and the entire health care ecosystems. The overall target of the project was to leverage blockchain technology to address critical points of friction, fraud, and waste management in pharmaceutical development and distribution. The project developed several use cases in the pharmaceutical and health domains:

- e-Leaflet or electronic product information. This serves as the digital counterpart to traditional paper leaflets in medicine, offering enhanced traceability of the product, data integrity, interoperability, and patient empowerment.
- Supply chain management. Applying blockchain technology provides many benefits, such as traceability of medical products in a tamper-resistant way, authentication and anticounterfeiting, data integrity, efficient recalls, and smart payments.
- Clinical trial recruitment. This use case addressed many challenges, such as time-costly patient recruitment procedures, enrollment times, financial problems, and accessibility and diversity issues.
- Internet of Things and personalized medicine. This approach collects data from medical devices of patients who agree to participate in a clinical trial in an automated way, stores them in the ledger in encrypted ways, and gives the user control to let their data be used by other researchers. These cases exploit a novel dynamic permission algorithm that provides real-time permission to researchers to perform secondary studies. Patients can see how their data are used, and they can revoke permission at any time and may withdraw their participation completely.
- e-Consent. Electronic consent is for certain applications in clinical research, to create value; work in a digital way; and

XSL•FO
RenderX

provide immutability, digital signatures, and all the essential elements required for modern platforms of clinical research.

The core architecture behind the PharmaLedger project is an open-source platform called OpenDSU [87]. The PharmaLedger project is considered a blockchain of blockchains network or a multiblockchain architecture (hierarchical) and is mainly characterized as a blockchain-agnostic network. In other words, using OpenDSU helped create a big ecosystem of trust that can work with any other permission or public blockchain. Most of the use cases required off-chain data management due to the limitations of the transaction size in the ledger, the complexity of the smart contracts, the execution time, the finality time, and the cost of writing data to the ledger. To avoid these problems, OpenDSU provides an extra layer of transactions that separates the off-chain data from the on-chain data and maintains only the essential transactions with their hashes on the ledger. Security, privacy, and confidentiality are embedded into the architecture with user wallets and client-side encryption along the self-sovereign applications. Providing a flexible architecture and keeping it blockchain agnostic creates valuable support for all the different use cases. PharmaLedger uses key design principles, such as (1) hierarchical blockchain structure, (2) blockchain agnosticism, (3) code integrity, (4) use case flexibility, (5) efficient smart contract use, and (6) decentralized identities and flexible deployment procedures, among others. As previously mentioned, data-sharing units (DSUs) can be thought of as standard entities in computer science, but they are encrypted. However, they can converse, and they carry arbitrary code and data as well. Each DSU is then divided into many smaller parts called “bricks,” and they form a brick map that has the links to all bricks. A person reconstructs the DSU object from the brick map when interacting with a DSU. Because no one can control it, anyone can host DSUs anywhere and at any time without worrying about privacy. The code and data are only visible to the owners. Finally, the PharmaLedger Association (PLA) is the continuation of the successful PharmaLedger project [88]. It evolved as a nonprofit association to deliver the Digital Trust Ecosystem in Healthcare. The focus is being shifted to the patients following the principles of PharmaLedger, such as neutrality, inclusivity, simplicity in coding, transparency, and open source. The main lines of development and research are decentralized trials, product trust, and supply chain. Many companies from the public to the private sector have already joined the association and exploited this new ecosystem. Previous use cases from the PharmaLedger project were transferred to the PLA XLAB, which stands for the innovation arm of the PLA and allows for the incubation and development of prototypes and demonstrators.

The Future of Novel Technology Digital Tools in Clinical Research and What to Expect

The future of emerging technologies in clinical research will witness transformative advancements driven by AI, blockchain, and innovative tools that are able to provide personalized solutions to all the stakeholders involved in the health care domain. The algorithms based on AI are revolutionizing the identification of suitable candidates and reducing the costs and time required for such procedures. The integration of blockchain into these technologies is a good fit for clinical trials and especially for remote clinical trials. These technologies not only maximize security but also enhance the privacy elements that were missing up to this point and ensure the transparency of clinical trials. Data security is ensured through the trial processes and transactions, as well as the accuracy and integrity of data exchanged among stakeholders during the clinical trial. The blockchain can provide improved data management and real-time analysis with the actual consent of patients, enabling patient empowerment and assured decisions. With new user-friendly interfaces, mobile apps, and other novel technology mechanisms involving blockchain, patient engagement will be advanced even further. Transparency, privacy, and ethical considerations will be the primary forces to be considered when creating such solutions. Better clinical research instruments and prompt trial schedules are anticipated in the future, which will promote patients’ comfort and provide them with better clinical trial experiences. The future of clinical research will be shaped by the integration of AI systems, such as chatbots, virtual assistants, predictive analytics, and data-driven decision-making tools, which will redefine the current tools in use. A new era is dawning on the conduct of clinical trials characterized by heightened standards and cutting-edge technologies. This shift contributes to the ongoing expansion, progression, and advancement of this field.

Strengths and Limitations of This Study

Ours is one of the first studies conducted at a European level reviewing and providing valuable information on innovative digital platforms available internationally focusing on features that enhance and accelerate complex clinical research processes such as the clinical trial recruitment stage. The review and selection of the studies was conducted with a careful systematic approach and validated by a team of researchers. One researcher conducted study collection; the other 2 reviewed the studies; and they all conducted the final screening, selection, and decision-making regarding the platforms included in the final tables. Concerning limitations, there were many digital platforms available on the internet that did not match the inclusion criteria for this review. One of the main issues was that many novel platforms had not been reviewed or tested enough in real time by researchers or had not been validated via scientific reports or published studies.

Conclusions

This study highlights the pivotal role of the identified innovative digital platforms in clinical trials, contributing significantly to recruitment efficacy, patient engagement enhancement, trial stage acceleration, and personalized patient vital sign monitoring. The positive impact of these platforms underscores their potential to revolutionize the landscape of clinical research. However, many challenges remain, such as issues of accessibility, digital compatibility constraints, and transparency issues that necessitate careful consideration of these tools’ use. Addressing these challenges is not only vital for the seamless integration of digital tools into clinical trials but is also crucial for ensuring the conduct of trials following all the appropriate ethical considerations. Achieving a balance between innovation and ethics is essential to optimize the benefits derived from digital tools while effectively managing the associated risks.

XSL•FO
RenderX

Future advancements in digital technology and ongoing efforts to address challenges are poised to further enhance the efficiency and ethical standards of clinical trials, ultimately advancing medical research and patient care.

Acknowledgments

This research work was conducted as part of the European Union–funded Horizon 2020 PharmaLedger project, which aimed at implementing blockchain technology to improve clinical research. The author team would like to express gratitude toward all the partners from the PharmaLedger project consortium and from the PharmaLedger Association [88], a nonprofit organization that continues and evolves the work started during the PharmaLedger project with the aim of supporting collaborative innovation in a health care digital trust ecosystem. They would also like to thank all the users that contributed significantly during the pilot-testing period of the PharmaLedger platform or application and actively participated in the project development with their valuable feedback. The authors worked as researchers for their academic institutions, which were partner members of the Horizon 2020 PharmaLedger project, which focused on the integration of blockchain technology into health care and the pharmaceutical industry. The research team of authors conceptualized the idea to conduct a review investigating the available innovative digital platforms used as novel technology tools for various clinical trial stages, especially for the recruitment process. PharmaLedger is a project characterized by innovative use cases, such as clinical trial e-recruitment, Product Leaflet, and personalized medicine features.

Authors' Contributions

AGB, ED, FD-G, CK, NL, and MFC-U contributed to the conceptualization of the study ED and CK contributed to the methodology. AGB, ED, and FD-G performed the formal analysis. AGB, FD-G, NL, JGT-P, MEBJ, and MFC-U conducted the investigation. AGB and FD-G were responsible for the collection of resources, FD-G, and ED curated data. CK, MEBJ, and MFC-U acquired funding for the project. AGB, ED, FD-G, and CK prepared the original draft. AGB, ED, FD-G, NL, JGT-P, MEBJ, MFC-U, and CK reviewed and edited the manuscript. MFC-U and CK supervised the visualization. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Lipsky MS, Sharp LK. From idea to market: the drug approval process. J Am Board Fam Pract 2001;14(5):362-367 [FREE Full text] [Medline: 11572541]
2. Sacristán JA. Clinical research and medical care: towards effective and complete integration. BMC Med Res Methodol 2015 Jan 09;15(1):4 [FREE Full text] [doi: 10.1186/1471-2288-15-4] [Medline: 25575454]
3. Poncette AS, Spies C, Mosch L, Schieler M, Weber-Carstens S, Krampe H, et al. Clinical requirements of future patient monitoring in the intensive care unit: qualitative study. JMIR Med Inform 2019 Apr 30;7(2):e13064 [FREE Full text] [doi: 10.2196/13064] [Medline: 31038467]
4. Kipnis E, Ramsingh D, Bhargava M, Dincer E, Cannesson M, Broccard A, et al. Monitoring in the intensive care. Crit Care Res Pract 2012;2012:473507 [FREE Full text] [doi: 10.1155/2012/473507] [Medline: 22970356]
5. Fitzer K, Haeuslschmid R, Blasini R, Altun FB, Hampf C, Freiesleben S, et al. Patient recruitment system for clinical trials: mixed methods study about requirements at ten university hospitals. JMIR Med Inform 2022 Apr 20;10(4):e28696 [FREE Full text] [doi: 10.2196/28696] [Medline: 35442203]
6. Johnson O. An evidence-based approach to conducting clinical trial feasibility assessments. Clin Investig 2015 May;5(5):491-499 [FREE Full text] [doi: 10.4155/CLI.14.139]
7. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018 Sep;11:156-164 [FREE Full text] [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
8. Baker TB, Gustafson DH, Shah D. How can research keep up with eHealth? Ten strategies for increasing the timeliness and usefulness of eHealth research. J Med Internet Res 2014 Feb 19;16(2):e36 [FREE Full text] [doi: 10.2196/jmir.2925] [Medline: 24554442]
9. Riley WT, Glasgow RE, Etheredge L, Abernethy AP. Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise. Clin Transl Med 2013 May 10;2(1):10 [FREE Full text] [doi: 10.1186/2001-1326-2-10] [Medline: 23663660]
10. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ 1999 Jul 17;319(7203):185. [doi: 10.1136/bmj.319.7203.185] [Medline: 10406763]

XSL•FO
RenderX

11. Brintnall-Karabelas J, Sung S, Cadman ME, Squires C, Whorton K, Pao M. Improving recruitment in clinical trials: why eligible participants decline. J Empir Res Hum Res Ethics 2011 Mar;6(1):69-74. [doi: 10.1525/jer.2011.6.1.69] [Medline: 21460590]
12. Watson JM, Torgerson DJ. Increasing recruitment to randomised trials: a review of randomised controlled trials. BMC Med Res Methodol 2006 Jul 19;6(1):34 [FREE Full text] [doi: 10.1186/1471-2288-6-34] [Medline: 16854229]
13. Desai M. Recruitment and retention of participants in clinical studies: critical issues and challenges. Perspect Clin Res 2020;11(2):51-53. [doi: 10.4103/picr.picr_6_20]
14. Califf R, Rutherford JD. Reflections on the clinical research enterprise: past, present and future. Circulation 2018 Oct 23;138(17):1765-1770. [doi: 10.1161/circulationaha.118.037900]
15. Steinhubl SR, Wolff-Hughes DL, Nilsen W, Iturriaga E, Califf RM. Digital clinical trials: creating a vision for the future. NPJ Digit Med 2019 Dec 12;2(1):126 [FREE Full text] [doi: 10.1038/s41746-019-0203-0] [Medline: 31872066]
16. Cascade E, Marr P, Winslow M, Burgess A, Nixon M. Conducting research on the internet: medical record data integration with patient-reported outcomes. J Med Internet Res 2012 Oct 11;14(5):e137 [FREE Full text] [doi: 10.2196/jmir.2202] [Medline: 23060427]
17. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020 Dec 3;2020:8830200 [FREE Full text] [doi: 10.1155/2020/8830200] [Medline: 33343657]
18. Alzubaidi H, Austin Z, Saidawi W, Rees VW. Exploring the quality of smoking cessation in community pharmacies: a simulated patient study. Res Social Adm Pharm 2022 Jun;18(6):2997-3003. [doi: 10.1016/j.sapharm.2021.07.013] [Medline: 34284972]
19. Yuan P, Bare MG, Johnson MO, Saberi P. Using online social media for recruitment of human immunodeficiency virus-positive participants: a cross-sectional survey. J Med Internet Res 2014 May 01;16(5):e117 [FREE Full text] [doi: 10.2196/jmir.3229] [Medline: 24784982]
20. Inan OT, Tenaerts P, Prindiville SA, Reynolds HR, Dizon DS, Cooper-Arnold K, et al. Digitizing clinical trials. NPJ Digit Med 2020 Jul 31;3(1):101 [FREE Full text] [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]
21. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial intelligence applied to clinical trials: opportunities and challenges. Health Technol (Berl) 2023;13(2):203-213 [FREE Full text] [doi: 10.1007/s12553-023-00738-2] [Medline: 36923325]
22. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019 Aug;40(8):577-591 [FREE Full text] [doi: 10.1016/j.tips.2019.05.005] [Medline: 31326235]
23. Algorri M, Cauchon NS, Christian T, O'Connell C, Vaidya P. Patient-centric product development: a summary of select regulatory CMC and device considerations. J Pharm Sci 2023 Apr;112(4):922-936 [FREE Full text] [doi: 10.1016/j.xphs.2023.01.029] [Medline: 36739904]
24. Aiyegbusi OL, Davies EH, Myles P, Williams T, Frost C, Haroon S, et al. Digitally enabled decentralised research: opportunities to improve the efficiency of clinical trials and observational studies. BMJ Evid Based Med 2023 Oct 21;28(5):328-331 [FREE Full text] [doi: 10.1136/bmjebm-2023-112253] [Medline: 36810190]
25. Helgeson J, Rammage M, Urman A, Roebuck MC, Coverdill S, Pomerleau K, et al. Clinical performance pilot using cognitive computing for clinical trial matching at Mayo Clinic. J Clin Oncol 2018 May 20;36(15_suppl):e18598. [doi: 10.1200/jco.2018.36.15_suppl.e18598]
26. Betcheva L, Kim JY, Erhun F, Oraiopoulos N, Getz K. Applying systems thinking to inform decentralized clinical trial planning and deployment. Ther Innov Regul Sci 2023 Sep 30;57(5):1081-1098. [doi: 10.1007/s43441-023-00540-2] [Medline: 37389795]
27. Tagde P, Tagde S, Bhattacharya T, Tagde P, Chopra H, Akter R, et al. Blockchain and artificial intelligence technology in e-Health. Environ Sci Pollut Res Int 2021 Oct 02;28(38):52810-52831 [FREE Full text] [doi: 10.1007/s11356-021-16223-0] [Medline: 34476701]
28. Lee K, Liu Z, Mai Y, Jun T, Ma M, Wang T, et al. Optimizing clinical trial eligibility design using natural language processing models and real-world data: algorithm development and validation. JMIR AI 2024 Jul 29;3:e50800 [FREE Full text] [doi: 10.2196/50800] [Medline: 39073872]
29. Panwar A, Bhatnagar V. Distributed ledger technology (DLT): the beginning of a technological revolution for blockchain. In: Proceedings of the 2nd International Conference on Data, Engineering and Applications. 2020 Presented at: IDEA 2020; February 28-29, 2020; Bhopal, India. [doi: 10.1109/idea49133.2020.9170699]
30. Risius M, Spohrer K. A blockchain research framework. Bus Inf Syst Eng 2017 Dec 5;59(6):385-409. [doi: 10.1007/s12599-017-0506-0]
31. Yuan Y, Wang FY. Towards blockchain-based intelligent transportation systems. In: Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems. 2016 Presented at: ITSC 2016; November 1-4, 2016; Rio de Janeiro, Brazil. [doi: 10.1109/itsc.2016.7795984]
32. Ballamudi KR. Blockchain as a type of distributed ledger technology. Asian J Humanity Art Lit 2016;3(2):127-136. [doi: 10.18034/ajhal.v3i2.528]
33. Amin R, Mandapuram M. CMS - intelligent machine translation with adaptation and AI. ABC J Adv Res 2021;10(2):199-206. [doi: 10.18034/abcjar.v10i2.693]

XSL•FO
RenderX

34. Gutlapalli SS. Commercial applications of blockchain and distributed ledger technology. Eng Int 2016 Dec 31;4(2):89-94. [doi: 10.18034/ei.v4i2.653]
35. Ramos S, Ellul J. Blockchain for Artificial Intelligence (AI): enhancing compliance with the EU AI Act through distributed ledger technology. A cybersecurity perspective. Int Cybersecur Law Rev 2024 Jan 25;5(1):1-20. [doi: 10.1365/s43439-023-00107-9]
36. Haleem A, Javaid M, Singh RP, Suman R, Rab S. Blockchain technology applications in healthcare: an overview. Int J Intell Netw 2021;2:130-139. [doi: 10.1016/j.ijin.2021.09.005]
37. Zarour M, Ansari MT, Alenezi M, Sarkar AK, Faizan M, Agrawal A, et al. Evaluating the impact of blockchain models for secure and trustworthy electronic healthcare records. IEEE Access 2020;8:157959-157953. [doi: 10.1109/access.2020.3019829]
38. Kim HJ, Kim HH, Ku H, Yoo KD, Lee S, Park JI, et al. Smart decentralization of personal health records with physician apps and helper agents on blockchain: platform design and implementation study. JMIR Med Inform 2021 Jun 07;9(6):e26230 [FREE Full text] [doi: 10.2196/26230] [Medline: 34096877]
39. Vora LK, Gholap AD, Jetha K, Thakur RR, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics 2023 Jul 10;15(7):1916 [FREE Full text] [doi: 10.3390/pharmaceutics15071916] [Medline: 37514102]
40. Loucks TL, Tyson C, Dorr D, Garovic VD, Hill J, McSwain SD, et al. Clinical research during the COVID-19 pandemic: the role of virtual visits and digital approaches. J Clin Transl Sci 2021 Mar 08;5(1):e102 [FREE Full text] [doi: 10.1017/cts.2021.19] [Medline: 34192057]
41. Dron L, Dillman A, Zoratti MJ, Haggstrom J, Mills EJ, Park JJ. Clinical trial data sharing for COVID-19-related research. J Med Internet Res 2021 Mar 12;23(3):e26718 [FREE Full text] [doi: 10.2196/26718] [Medline: 33684053]
42. Robison L, Cho Y, Viecelli AK, Johnson DW, Hawley CM, Valks A, et al. Conducting clinical trials during the COVID-19 pandemic-a collaborative trial network response. Trials 2021 Apr 14;22(1):278 [FREE Full text] [doi: 10.1186/s13063-021-05200-0] [Medline: 33853661]
43. Will JF. A brief historical and theoretical perspective on patient autonomy and medical decision making: part II: the autonomy model. Chest 2011 Jun;139(6):1491-1497. [doi: 10.1378/chest.11-0516] [Medline: 21652559]
44. Gupta UC. Informed consent in clinical research: revisiting few concepts and areas. Perspect Clin Res 2013 Jan;4(1):26-32 [FREE Full text] [doi: 10.4103/2229-3485.106373] [Medline: 23533976]
45. De Sutter E, Zaçe D, Boccia S, Di Pietro ML, Geerts D, Borry P, et al. Implementation of electronic informed consent in biomedical research and stakeholders' perspectives: systematic review. J Med Internet Res 2020 Oct 08;22(10):e19129 [FREE Full text] [doi: 10.2196/19129] [Medline: 33030440]
46. Pietrzykowski T, Smilowska K. The reality of informed consent: empirical studies on patient comprehension-systematic review. Trials 2021 Jan 14;22(1):57 [FREE Full text] [doi: 10.1186/s13063-020-04969-w] [Medline: 33446265]
47. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
48. del Carmen MG, Joffe S. Informed consent for medical treatment and research: a review. Oncologist 2005 Sep;10(8):636-641 [FREE Full text] [doi: 10.1634/theoncologist.10-8-636] [Medline: 16177288]
49. Sreenivasan G. Does informed consent to research require comprehension? Lancet 2003 Dec 13;362(9400):2016-2018. [doi: 10.1016/S0140-6736(03)15025-8] [Medline: 14683665]
50. Chimonas S, Lipitz-Snyderman A, Gaffney K, Kuperman GJ. Electronic consent at US cancer centers: a survey of practices, challenges, and opportunities. JCO Clin Cancer Inform 2023 Jan;7. [doi: 10.1200/cci.22.00122]
51. Almeida-Magana R, Maroof H, Grierson J, Clow R, Dinneen E, Al-Hammouri T, et al. E-Consent-a guide to maintain recruitment in clinical trials during the COVID-19 pandemic. Trials 2022 May 12;23(1):388 [FREE Full text] [doi: 10.1186/s13063-022-06333-6] [Medline: 35550639]
52. Rosa C, Marsch LA, Winstanley EL, Brunner M, Campbell AN. Using digital technologies in clinical trials: current and future applications. Contemp Clin Trials 2021 Jan;100:106219 [FREE Full text] [doi: 10.1016/j.cct.2020.106219] [Medline: 33212293]
53. Chun-Yun Kang G. Technology-based interventions to improve adherence to antihypertensive medications - an evidence-based review. Digit Health 2022 May 03;8:20552076221089725 [FREE Full text] [doi: 10.1177/20552076221089725] [Medline: 35531090]
54. Iqbal SM, Mahgoub I, Du E, Leavitt MA, Asghar W. Advances in healthcare wearable devices. npj Flex Electron 2021 Apr 12;5:9. [doi: 10.1038/s41528-021-00107-x]
55. Breteler MJ, Huizinga E, van Loon K, Leenen LP, Dohmen DA, Kalkman CJ, et al. Reliability of wireless monitoring using a wearable patch sensor in high-risk surgical patients at a step-down unit in the Netherlands: a clinical validation study. BMJ Open 2018 Feb 27;8(2):e020162 [FREE Full text] [doi: 10.1136/bmjopen-2017-020162] [Medline: 29487076]
56. Weller RS, Foard KL, Harwood TN. Evaluation of a wireless, portable, wearable multi-parameter vital signs monitor in hospitalized neurological and neurosurgical patients. J Clin Monit Comput 2018 Oct;32(5):945-951. [doi: 10.1007/s10877-017-0085-0] [Medline: 29214598]

XSL•FO
RenderX

57. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005 Nov 05;331(7524):1064-1065. [doi: 10.1136/bmj.38636.593461.68] [Medline: 16230312]
58. Wessels DH, Rosenberg Z. Awareness of non-alcoholic steatohepatitis and treatment guidelines: what are physicians telling us? World J Hepatol 2021 Feb 27;13(2):233-241 [FREE Full text] [doi: 10.4254/wjh.v13.i2.233] [Medline: 33708352]
59. Geist R, Militello M, Albrecht JM, Presley CL, Anderson JB, Laughter M, et al. Social media and clinical research in dermatology. Curr Dermatol Rep 2021 Nov 19;10(4):105-111 [FREE Full text] [doi: 10.1007/s13671-021-00350-5] [Medline: 34840860]
60. Cohen E, Belkora J, Tyler J, Schreiner J, Deering MJ, Grama L, et al. Adoption, acceptability, and accuracy of an online clinical trial matching website for breast cancer. J Med Internet Res 2012 Jul 11;14(4):e97 [FREE Full text] [doi: 10.2196/jmir.1855] [Medline: 22784878]
61. Gaba P, Bhatt DL. The COVID-19 pandemic: a catalyst to improve clinical trials. Nat Rev Cardiol 2020 Nov 01;17(11):673-675 [FREE Full text] [doi: 10.1038/s41569-020-00439-7] [Medline: 32873977]
62. Park M, Bannuru RR, Price LL, Harvey WF, Driban JB, Wang C. Effective recruitment strategies in an exercise trial for patients with fibromyalgia. Trials 2021 Aug 21;22(1):557 [FREE Full text] [doi: 10.1186/s13063-021-05502-3] [Medline: 34419131]
63. Enos G. Provider organization in North Carolina seeks to establish major research center. Mental Health Weekly 2019 Jul 28;29(29):1-5. [doi: 10.1002/mhw.31991]
64. Takeda C, Guyonnet S, Sumi Y, Vellas B, Araujo de Carvalho I. Integrated care for older people and the implementation in the INSPIRE care cohort. J Prev Alzheimers Dis 2020;7(2):70-74. [doi: 10.14283/jpad.2020.8] [Medline: 32236394]
65. Mudaranthakam DP, Alsup AM, Murakonda V, Lin T, Thompson J, Gajewski B, et al. Accelerating cancer patient recruitment through a mobile application (Clinical Trial Finder). Cancer Inform 2022 Jan 25;21:11769351211073114 [FREE Full text] [doi: 10.1177/11769351211073114] [Medline: 35095270]
66. Norton L, Ciervo J, Lobanov VS, Agrafiotis DK. Xcellerate investigator portal: a new web-based tool for online delivery of central laboratory data, reports, and communications to clinical sites. SLAS Technol 2020 Oct;25(5):427-435 [FREE Full text] [doi: 10.1177/2472630320942200] [Medline: 32726559]
67. Maslove DM, Klein J, Brohman K, Martin P. Using blockchain technology to manage clinical trials data: a proof-of-concept study. JMIR Med Inform 2018 Dec 21;6(4):e11949 [FREE Full text] [doi: 10.2196/11949] [Medline: 30578196]
68. -. FDA approves stroke-detecting AI software. Nat Biotechnol 2018 Apr 05;36(4):290. [doi: 10.1038/nbt0418-290] [Medline: 29621226]
69. Jain NM, Culley A, Knoop T, Micheel C, Osterman T, Levy M. Conceptual framework to support clinical trial optimization and end-to-end enrollment workflow. JCO Clin Cancer Inform 2019 Jun 21;3:1-10. [doi: 10.1200/cci.19.00033]
70. Storrar W, Brown T, Balls L, Chauhan A, Fogg C. Recruitment to clinical trials - the use of social media. Trials 2015 Nov 16;16(Suppl 2):O77. [doi: 10.1186/1745-6215-16-s2-o77]
71. Ilesanmi-Oyelere BL, Schollum L, Kuhn-Sherlock B, McConnell M, Mros S, Coad J, et al. Inflammatory markers and bone health in postmenopausal women: a cross-sectional overview. Immun Ageing 2019 Jul 10;16(1):15 [FREE Full text] [doi: 10.1186/s12979-019-0155-x] [Medline: 31333751]
72. Thurman A, Simon J, Kingsberg SA, Cornell KA, Hatheway J, Friend DR, et al. (123) preliminary efficacy of sildenafil cream, 3.6% in an exploratory, randomized, placebo-controlled, phase 2b trial for the treatment of female sexual arousal disorder. J Sex Med 2024 Feb;21(Supplement_1):qdae001.117. [doi: 10.1093/jsxmed/qdae001.117]
73. Copland RR, Hanke S, Rogers A, Mpaltadoros L, Lazarou I, Zeltsi A, et al. The digital platform and its emerging role in decentralized clinical trials. J Med Internet Res 2024 Sep 03;26:e47882 [FREE Full text] [doi: 10.2196/47882] [Medline: 39226549]
74. Saha B, Mazumder C, Laha A, Mitra Thakur GS, Mukherjee D. ImmunoCharge: phase 4 clinical trial app for pandemic vaccination. engrxiv Preprint posted online on December 13, 2021 [FREE Full text] [doi: 10.31224/osf.io/gm46z]
75. Christofides EA, Tian A, Mastacouris N, Zhu A, Dennis O. Comparison of electronic health records and electronic source data in clinical research trials: a retrospective review. J Soc Clin Data Manag 2024;4(1). [doi: 10.47912/jscdm.224]
76. Raevuori A, Vahlberg T, Korhonen T, Hilgert O, Aittakumpu-Hyden R, Forman-Hoffman V. A therapist-guided smartphone app for major depression in young adults: a randomized clinical trial. J Affect Disord 2021 May 01;286:228-238 [FREE Full text] [doi: 10.1016/j.jad.2021.02.007] [Medline: 33743385]
77. Hulstaert L, Twick I, Sarsour K, Verstraete H. Enhancing site selection strategies in clinical trial recruitment using real-world data modeling. PLoS One 2024 Mar 11;19(3):e0300109 [FREE Full text] [doi: 10.1371/journal.pone.0300109] [Medline: 38466688]
78. Peyser ND, Marcus GM, Beatty AL, Olgin JE, Pletcher MJ. Digital platforms for clinical trials: the Eureka experience. Contemp Clin Trials 2022 Apr;115:106710 [FREE Full text] [doi: 10.1016/j.cct.2022.106710] [Medline: 35183763]
79. Beck JT, Rammage M, Jackson GP, Preininger AM, Dankwa-Mullan I, Roebuck MC, et al. Artificial intelligence tool for optimizing eligibility screening for clinical trials in a large community cancer center. JCO Clin Cancer Inform 2020 Jan;4:50-59 [FREE Full text] [doi: 10.1200/CCI.19.00079] [Medline: 31977254]

XSL•FO
RenderX

80. DiMasi JA, Smith Z, Oakley-Girvan I, Mackinnon A, Costello M, Tenaerts P, et al. Assessing the financial value of decentralized clinical trials. Ther Innov Regul Sci 2023 Mar;57(2):209-219 [FREE Full text] [doi: 10.1007/s43441-022-00454-5] [Medline: 36104654]
81. Kumar BR. Case 25 EPAM Systems Inc. In: Rising Stars. Management for Professionals. Cham, Switzerland: Springer; 2024.
82. Cai T, Cai F, Dahal KP, Cremone G, Lam E, Golnik C, et al. Improving the efficiency of clinical trial recruitment using an ensemble machine learning to assist with eligibility screening. ACR Open Rheumatol 2021 Sep 23;3(9):593-600 [FREE Full text] [doi: 10.1002/acr2.11289] [Medline: 34296815]
83. Gligorijevic J, Gligorijevic D, Pavlovski M, Milkovits E, Glass L, Grier K, et al. Optimizing clinical trials recruitment via deep learning. J Am Med Inform Assoc 2019 Nov 01;26(11):1195-1202 [FREE Full text] [doi: 10.1093/jamia/ocz064] [Medline: 31188432]
84. Omar IA, Jayaraman R, Salah K, Yaqoob I, Ellahham S. Applications of blockchain technology in clinical trials: review and open challenges. Arab J Sci Eng 2020 Oct 14;46(4):3001-3015. [doi: 10.1007/s13369-020-04989-3]
85. Pletscher F, Mändli Lerch K, Glinz D. Willingness to share anonymised routinely collected clinical health data in Switzerland: a cross-sectional survey. Swiss Med Wkly 2022 Jun 20;152(2324):w30182 [FREE Full text] [doi: 10.4414/smw.2022.w30182] [Medline: 35752970]
86. PharmaLedger. European Commission. URL: https://cordis.europa.eu/project/id/853992 [accessed 2024-12-06]
87. OpenDSU homepage. OpenDSU Documentation. URL: https://www.opendsu.org/ [accessed 2024-12-06]
88. PharmaLedger homepage. PharmaLedger. URL: https://pharmaledger.org/ [accessed 2024-12-06]

Abbreviations

AI: artificial intelligence
CRIO: Clinical Research IO
DLT: distributed ledger technology
DSU: data-sharing unit
PLA: PharmaLedger Association

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by PFW Chien, M Popovic, N Howe; comments to author 02.08.24; revised version received 21.10.24; accepted 06.11.24; published 18.12.24.

Please cite as:
Bikou AG, Deligianni E, Dermiki-Gkana F, Liappas N, Teriús-Padrón JG, Beltrán Jaunsarás ME, Cabrera-Umpiérrez MF, Kontogiorgis C
Improving Participant Recruitment in Clinical Trials: Comparative Analysis of Innovative Digital Platforms
J Med Internet Res 2024;26:e60504
URL: https://www.jmir.org/2024/1/e60504
doi:10.2196/60504
PMID:

©Alexia Georgia Bikou, Elena Deligianni, Foteini Dermiki-Gkana, Nikolaos Liappas, José Gabriel Teriús-Padrón, Maria Eugenia Beltrán Jaunsarás, Maria Fernanda Cabrera-Umpiérrez, Christos Kontogiorgis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining the Type, Quality, and Content of Web-Based Information for People With Chronic Pain Interested in Spinal Cord Stimulation: Social Listening Study

Maarten Moens[1,2,3,4,5], MD, PhD; Leen Van Doorslaer[2], MSc; Maxime Billot[6], MSc, PhD; Edgard Eeckman[7], MSc, PhD; Manuel Roulaud[6], MSc; Philippe Rigoard[6,8,9], MD, PhD; Maaike Fobelets[10,11], MSc, PhD; Lisa Goudman[1,2,3,5,12], MSc, PhD

[1]Department of Neurosurgery, Universitair Ziekenhuis Brussel, Brussels, Belgium

[2]STIMULUS (reSearch and TeachIng neuroModULation Uz bruSsel) Research Group, Vrije Universiteit Brussel, Brussels, Belgium

[3]Center for Neurosciences, Vrije Universiteit Brussel, Brussels, Belgium

[4]Department of Radiology, Universitair Ziekenhuis Brussel, Brussels, Belgium

[5]Pain in Motion Research Group, Department of Physiotherapy, Human Physiology and Anatomy, Vrije Universiteit Brussel, Brussels, Belgium

[6]PRISMATICS (Predictive Research in Spine/Neuromodulation Management and Thoracic Innovation/Cardiac Surgery) Lab, Poitiers University Hospital, Poitiers, France

[7]ECHO Research Group, Vrije Universiteit Brussel, Brussels, Belgium

[8]Department of Spine Surgery & Neuromodulation, Poitiers University Hospital, Poitiers, France

[9]Pprime Institute UPR 3346, CNRS, ISAE-ENSMA, University of Poitiers, Chasseneuil-du-Poitou, France

[10]Biostatistics and Medical Informatics Research Group, Vrije Universiteit Brussel, Brussels, Belgium

[11]Brussels Institute for Teacher Education, Vrije Universiteit Brussel, Brussels, Belgium

[12]Research Foundation—Flanders (FWO), Brussels, Belgium

Corresponding Author:
Lisa Goudman, MSc, PhD
STIMULUS (reSearch and TeachIng neuroModULation Uz bruSsel) Research Group
Vrije Universiteit Brussel
Laarbeeklaan 103
Brussels, 1090
Belgium
Phone: 32 472412507
Email: lisa.goudman@gmail.com

Abstract

Background: The increased availability of web-based medical information has encouraged patients with chronic pain to seek health care information from multiple sources, such as consultation with health care providers combined with web-based information. The type and quality of information that is available on the web is very heterogeneous, in terms of content, reliability, and trustworthiness. To date, no studies have evaluated what information is available about neuromodulation on the web for patients with chronic pain.

Objective: This study aims to explore the type, quality, and content of web-based information regarding spinal cord stimulation (SCS) for chronic pain that is freely available and targeted at health care consumers.

Methods: The social listening tool Awario was used to search Facebook (Meta Platforms, Inc), Twitter (Twitter, Inc), YouTube (Google LLC), Instagram (Meta Platforms, Inc), blogs, and the web for suitable hits with “pain” and “neuromodulation” as keywords. Quality appraisal of the extracted information was performed using the DISCERN instrument. A thematic analysis through inductive coding was conducted.

Results: The initial search identified 2174 entries, of which 630 (28.98%) entries were eventually withheld, which could be categorized as web pages, including news and blogs (114/630, 18.1%); Reddit (Reddit, Inc) posts (32/630, 5.1%); Vimeo (Vimeo, Inc) hits (38/630, 6%); or YouTube (Google LLC) hits (446/630, 70.8%). Most posts originated in the United States (519/630, 82.4%). Regarding the content of information, 66.2% (383/579) of the entries discussed (fully discussed or partially discussed) how SCS works. In total, 55.6% (322/579) of the entries did not elaborate on the fact that there may be >1 potential treatment

XSL•FO
RenderX

choice and 47.7% (276/579) did not discuss the influence of SCS on the overall quality of life. The inductive coding revealed 4 main themes. The first theme of pain and the burden of pain (1274/8886, 14.34% coding references) explained about pain, pain management, individual impact of pain, and patient experiences. The second theme included neuromodulation as a treatment approach (3258/8886, 36.66% coding references), incorporating the background on neuromodulation, patient-centered care, SCS therapy, and risks. Third, several device-related aspects (1722/8886, 19.38% coding references) were presented. As a final theme, patient benefits and testimonials of treatment with SCS (2632/8886, 29.62% coding references) were revealed with subthemes regarding patient benefits, eligibility, and testimonials and expectations.

Conclusions: Health care consumers have access to web-based information about SCS, where details about the surgical procedures, the type of material, working mechanisms, risks, patient expectations, testimonials, and the potential benefits of this therapy are discussed. The reliability, trustworthiness, and correctness of web-based sources should be carefully considered before automatically relying on the content.

(J Med Internet Res 2024;26:e48599) doi:10.2196/48599

KEYWORDS

online information; social listening; neuromodulation; patient care; chronic pain; web-based data

Introduction

Spinal Cord Stimulation for Chronic Pain

One of the treatment options for a variety of chronic pain conditions is spinal cord stimulation (SCS), a neuromodulation technique that delivers electrical current via an implanted lead to the spinal cord [1,2]. SCS is a well-known pain management option for patients with chronic lower back pain or leg pain who are refractory to conventional management, those with complex regional pain syndrome, those with neuropathic pain syndromes, or those experiencing ischemic pain syndromes to decrease pain intensity scores and to improve the functional status of these patients [3-7]. In addition, SCS can provide greater pain relief than pharmacological treatment, improve the quality of life, and enable the return to normalized and work-related activities in some patients [8,9]. Specific to the context of persistent spinal pain syndrome and generalizable to the context of chronic pain, an interdisciplinary approach has been proposed for pain management [10,11]; therefore, several health care providers are involved in the management of these patients [12]. Most interventions in the management of chronic pain include goal setting (ie, the process of selecting treatment goals), action planning, ways of monitoring treatment progress, feedback procedures, and compliance evaluations [13,14]. Goal setting refers to the process in which goals are identified and agreed between the medical team and the patient [15]. Interviews with patients before SCS implantation revealed that all patients had goals on the level of pain reduction and walking abilities [16] and treatment expectations about improvement in the quality of life with SCS [17]. When expectations regarding SCS were explored, six themes were revealed: (1) physical well-being, (2) social well-being, (3) material well-being, (4) emotional well-being, (5) development and activity, and (6) constraints of the SCS procedure [17]. These goals and expectations are different for every patient and will highly influence patient decision-making when seeking medical help [18].

Web-Based Health Information

In addition, as has been demonstrated in patients with other chronic diseases such as Parkinson disease, patients and their caregivers increasingly rely on the internet to gain information and access health information [19]. Previously, a gap has been identified between the availability of medical information and the translation of this knowledge to consumers of the health system [20]. Research has shown that consumers with lower back pain want clear, concise, and easy-to-understand information regarding medical intervention, physiotherapy, and other therapeutic interventions [21] and most often seek and receive information from multiple sources [20]. An increase has been documented in the number of consumers who access web-based information through web pages to search for information regarding the proposed treatments to inform personal choice. Nevertheless, there is considerable variability in the type and quality of information that is available on the web [22]. As health care professionals, it is important to be aware of the type and quality of information that is currently available to consumers on the web [23,24]. A recent study showed that approximately half of the consumers looking for health information for themselves reported that decisions regarding their treatment choices were influenced by the results of their web-based searches [22]. Moreover, web health information made resources exceedingly available and increased patients' capabilities of influence during a consultation with a health care provider [25]. As such, the availability of web-based inquiries can complement and be used in synergy during consultations with health care providers and provide options for exploring answers to additional or forgotten questions, sensitive situations, and similar patient testimonials [26]. In contrast, quality control remains challenging, and patients' health information literacy is highly variable [27]. Furthermore, misleading information, emotionally charged stories, and unscientific health practices are available, potentially leading to detrimental decisions in susceptible persons [26] as they experience a painful condition.

Web-Based Information About SCS

To date, no studies have identified and evaluated what information is available over the web about SCS for patients with chronic pain. Therefore, this study applied a social listening approach, defined as an active process of attending to, observing, interpreting, and responding to a variety of stimuli through mediated, electronic, and social channels to explore web-based information about SCS [28]. The primary aim of this study was

XSL•FO
RenderX

to explore the type and quality of web-based information regarding SCS for chronic pain that is freely available over the web and targeted at consumers of health care. The secondary aim was to explore the content of web-based information.

Methods

Search Strategy

For this study, only publicly available information from the internet was used. To collect data from the internet, a social listening tool, Awario, was consulted [29], and data were collected until July 31, 2022. Awario is a social listening software that is able to continually search various web-based platforms for keywords or phrases in new or previously made posts, with >13 billion web pages searched daily [23]. Awario used the same search term strategy as literature databases; therefore, the research question was composed according to the population, intervention, comparison, and outcome framework [30]. As keywords, we searched for "pain" and "neuromodulation" with the Boolean operator AND to refine results in 3 different languages: English, French, and Dutch. This search strategy was used to search Facebook (Meta Platforms, Inc), Twitter (Twitter, Inc, subsequently rebranded X), YouTube (Google LLC), Instagram (Meta Platforms, Inc), blogs, and the web using the following search terms for each keyword: *("pain" OR pijn OR douleur OR #pain) AND (neurostimulation OR #neurostimulation OR neurostimulatie OR neuromodulation OR #neuromodulation OR "spinal cord stimulation" OR #spinalcordstimulation OR neuromodulatie OR ruggenmergstimulatie OR SCS OR #SCS)* to collect web-based information regarding neuromodulation for chronic pain, targeted at consumers.

Selection Criteria

All extracted mentions were read in full by 2 independent reviewers to ensure their suitability for inclusion. This study assessed the type and quality of information available for people with chronic pain who aimed to learn more about SCS as a treatment modality. Therefore, web pages including Twitter (Twitter, Inc) posts, Facebook (Meta Platforms, Inc) posts, YouTube (Google LLC) videos, Vimeo (Vimeo, Inc) videos, Instagram (Meta Platforms, Inc) posts, blogs, and HTML web pages were included if they met the following criteria: (1) the term SCS was discussed; (2) the setting was chronic pain; and (3) the pages were written in English, French, or Dutch. Language restrictions were set for the researchers to fully comprehend the information and correctly perform quality appraisal and thematic analysis. Posts were excluded if they were not aimed at consumers (ie, patients) or were solely advertisements for specific physicians or hospitals. If discrepancies occurred between the 2 reviewers, consensus was sought through consultation and discussion with a third independent reviewer.

Ethical Considerations

The study protocol was approved by the Ethics Committee of Universitair Ziekenhuis Brussel (B.U.N. 1432022000130) on June 22, 2022. All data used in this study are of public nature, relying on internet-based data. We did not rely on artificial intelligence to analyze the data in this study.

Quality Appraisal and Data Extraction

After the search period, all results were extracted from Awario and imported into an Excel (Microsoft Corporation) spreadsheet to allow for data extraction. For each extracted mention, the following information was systematically extracted: location by country, type of post (eg, a web page, blog, or tweet), industry sponsorship, and information on the web page host (eg, physician, health service provider, or patient support group).

Quality appraisal of the extracted information was performed using the DISCERN instrument [31], which was specifically designed to assess written information regarding treatment choices. More specifically, the instrument provides internet users with a method to rate the quality of the obtained information [32]. The instrument consists of 15 key questions and an overall quality rating question. The first 8 questions evaluate reliability to determine whether the source of information can be trusted. The following 7 questions focus on the details of the information about the treatment options (ie, SCS in this case). The final question evaluated the overall quality rating. All questions are scored on a 5-point scale, where 1 indicates that the quality criterion has not been fulfilled and 5 indicates that the quality criterion has been completely fulfilled.

Evidence Synthesis

All analyses were performed using the R Studio software (version 4.0.5; R Foundation of Statistical Computing). At first, a world map was created with the aid of the *rnaturalearth* package. Entries for which the country of origin could not be derived were omitted from this analysis. Entries that were created for a specific region (eg, Europe, Middle East, and Africa) were counted in the country of the headquarters of that region. NVivo Transcription (QSR International) was used to obtain the transcripts for each entry. All transcripts were screened, and manual corrections were performed where necessary. A thematic analysis was performed in NVivo 12 (QSR International), in which data were categorized under different themes and subthemes through inductive coding according to standard thematic analysis guidelines. The applied thematic analysis consisted of familiarization, coding sentences, categorizing codes, and developing themes and subthemes. During the first cycle of coding, mostly descriptive and holistic coding methods were used, and to a minor degree, structural coding methods were assessed. During the second cycle of coding, a focused coding method was used to develop themes and subthemes. A word cloud of the most frequently used words was calculated after removing common stop words (eg, slang, abbreviations, or ironic unconventional written expressions).

Results

Search Results

Using Awario, we identified 2174 web-based entries, of which 45 (2.07%) duplicate entries could be removed immediately. The remaining 2129 (97.93%) entries were evaluated by 2 reviewers independently, where the percentage of agreement

XSL•FO
RenderX

between both reviewers for full-entry screening was 93.85%. After this full-entry screening of 2129 entries, 1 (0.05%) entry was excluded because there was no free access, 587 (27.57%) were excluded because of wrong intervention (no SCS), 37 (1.74%) were excluded because these only consisted of a referral to a specific center or physician or referral to SCS without explanation about the treatment, and 101 (4.74%) were excluded because of wrong population. Of 2129 entries, the format was not applied to patients for 202 (9.45%) entries, and 328 (15.41%) entries were excluded owing to the wrong topic that was discussed. In addition, 152 (7.14%) entries were inaccessible, and 90 (4.23%) were excluded based on language. One additional duplicate was found during the full screening. Figure 1 presents the flowchart of this study. Thus, after screening, 1499 (70.41%) entries were excluded and 630 (29.59%) were included in this study. For quality appraisal and thematic analysis, entries with a different URL address but with the same content as another entry were removed from this analysis. Therefore, quality appraisal was performed for 579 (27.19%) entries. Within the thematic analysis, 6 entries were excluded because of the lack of spoken or written material.

The included 630 web-based entries could be categorized as web pages, including news and blogs (114/630, 18.1%); Reddit (Reddit, Inc) posts (32/630, 5.1%); Vimeo (Vimeo, Inc) hits (38/630, 6%); or YouTube (Google LLC) hits (446/630, 70.8%). The upload date of the entries ranged from October 30, 2007, to July 27, 2022, with the median date on November 1, 2019 (quartile 1 to quartile 3: June 11, 2016, to July 8, 2021). The entries were created by institutions (ie, medical centers, expert institutes, expert societies, and universities; 311/630, 49.4%), industry (133/630, 21.1%), patients (73/630, 11.6%), health care providers (49/630, 7.8%), television shows (43/630, 6.8%), and others (21/630, 3.3%). Figure 2 presents the country of origin for all entries on the world map. Most posts originated in the United States (519/630, 82.4%). Furthermore, posts originated in the following countries: United Kingdom (26/630, 4.1%), the Netherlands (16/630, 2.5%), France (11/630, 1.8%), Australia (11/630, 1.8%), Canada (9/630, 1.4%), and India (7/630, 1.1%). To a lesser degree, posts originated in Switzerland, Belgium, Germany, Denmark, Spain, Ireland, Japan, Luxembourg, Pakistan, Sweden, Singapore, and Taiwan. For 10 (1.6%) out of 630 entries, the location could not be derived. Of 630 entries, 274 (43.5%) were created by an industrial partner of neuromodulation devices or denoted a specific industry partner, whereas 356 (56.5%) were not associated with an industrial partner. World maps and entry types based on industry-related and non–industry-related entries are provided in Multimedia Appendix 1. World maps presenting the locations of the 4 prominent companies using neuromodulation devices are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. Flowchart representing the screening process for this study in which web-based information about spinal cord stimulation (SCS) for chronic pain was retrieved with a social listening design. Blue boxes indicate the flow of included entries, whereas orange boxes give indications on excluded entries.

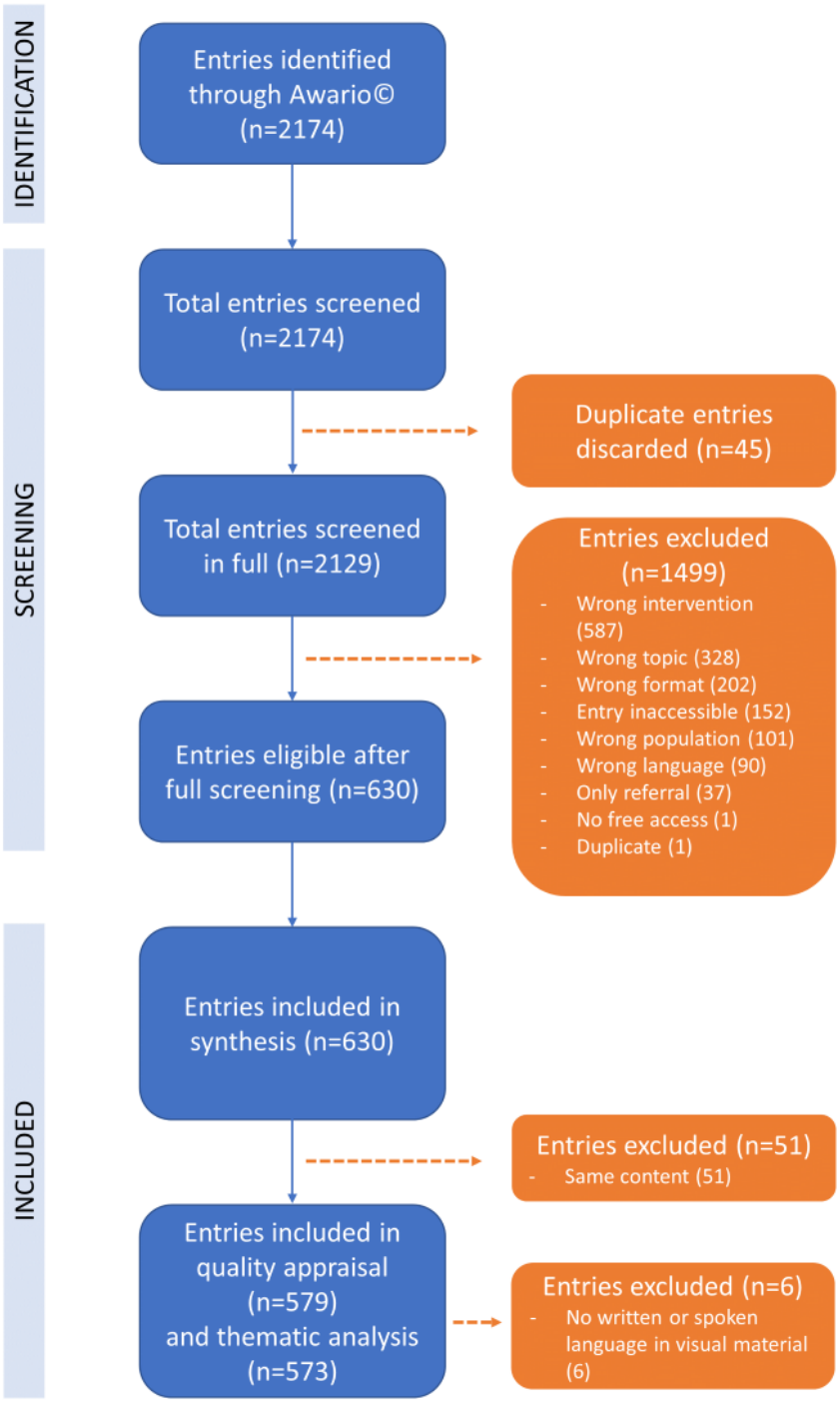

XSL•FO
RenderX

Figure 2. World map indicating the country of origin of the included web-based entries that were available about spinal cord stimulation for chronic pain. Darker red color indicates a higher count of entries. White color indicates that no entries originated from that respective country. For 10 entries, no country could be derived; therefore, 620 entries are represented on this world map.

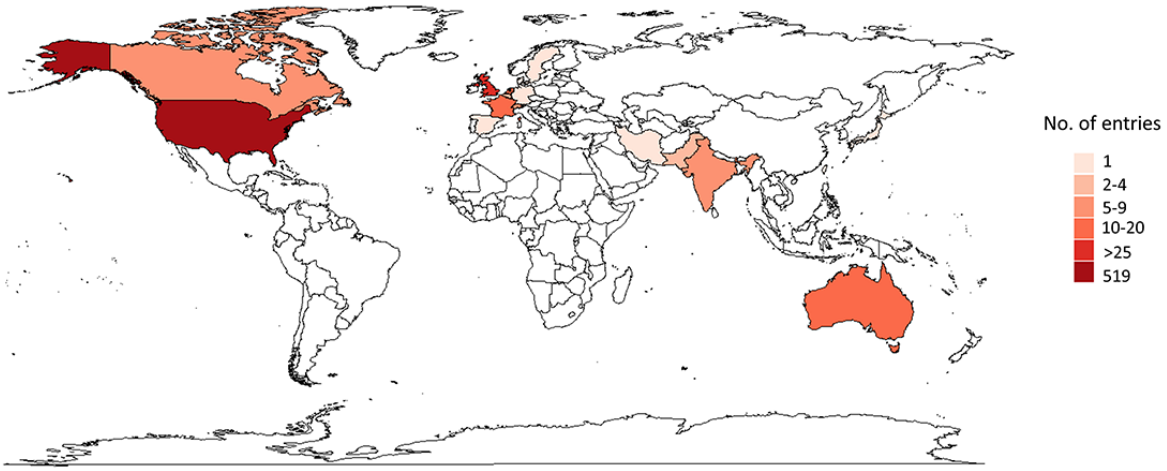

In Multimedia Appendix 3, the word cloud is presented (80 most frequent words) with a clear dominance of the word "pain." As most of the entries were written in English, with few articles in French or Dutch, the most frequently stated terms were in English.

Quality Appraisal

Regarding the content of information, 66.2% (383/579) of the entries discussed (at least partially) how SCS works. The risks of treatment were only discussed in a limited number of entries (142/579, 24.5%). In total, 55.6% (322/579) of the entries did not elaborate on the fact that there may be >1 potential treatment choice and 47.7% (276/579) did not discuss the influence of SCS on the overall quality of life. Regarding the reliability of information, 72.2% (418/579) of the entries did not clearly describe the aims that they aimed to achieve. In 94.8% (549/579) of the entries, it was clear or partially clear when the information in the entry was produced. Areas of uncertainty were not discussed in most entries (313/579, 54.1%). The overall quality of entries was 2.5 on a scale of 5, in which Reddit (Reddit, Inc) posts scored 2.6, Vimeo (Vimeo, Inc) videos 2.5, web pages 2.8, and YouTube (Google LLC) hits 2.4 out of 5, respectively. The complete quality appraisal, including the number of entries that did not fulfill the quality criteria (score 1), partially fulfilled the quality criteria (scores 2-4), or completely fulfilled the quality criteria (score 5), for each question of the DISCERN instrument is presented in Table 1. In addition, the average scores are shown for all entries and separated by the type of entry.

XSL•FO
RenderX

Table 1. Quality appraisal scoring of the available web-based information concerning spinal cord stimulation for chronic pain using the DISCERN instrument (N=579).

Category and questions	No[a] (score 1), n (%)	Partially[b] (scores 2-4), n (%)	Yes[c] (score 5), n (%)	Total (N=579), mean (SD)	Reddit (Reddit, Inc) posts (n=32), mean (SD)	Vimeo (Vimeo, Inc) videos (n=35), mean (SD)	Web pages (n=105), mean (SD)	YouTube (Google LLC) hits (n=407), mean (SD)
Reliability								
Are the aims clear?	*418 (72.2)* [d]	134 (23.1)	27 (4.7)	1.6 (1.1)	1.9 (1.0)	1.5 (1.0)	2.2 (1.4)	1.5 (1.0)
Does the publication achieve its aims?[e]	11 (6.8)	*134 (83.2)*	16 (9.9)	3.1 (0.9)	3.1 (0.3)	3.0 (0.0)	3.3 (1.2)	3.0 (0.8)
Is the publication relevant?	70 (12.1)	*416 (71.9)*	93 (16.1)	3.0 (1.2)	3.6 (1.2)	3.6 (0.9)	3.7 (1.2)	2.7 (1.0)
Is it clear what sources of information were used to compile the publication?	105 (18.1)	*414 (71.5)*	60 (10.4)	2.9 (1.1)	2.8 (0.8)	3.0 (1.0)	2.7 (1.6)	2.9 (0.9)
Is it clear when the information used or reported in the publication was produced?	30 (5.2)	*502 (86.7)*	47 (8.1)	3.1 (0.7)	3.1 (0.3)	3.2 (0.6)	3.1 (1.1)	3.1 (0.7)
Is the publication balanced and unbiased?	15 (2.6)	243 (41.9)	*321 (55.4)*	4.1 (1.1)	4.1 (1.3)	4.5 (0.9)	4.3 (1.0)	4.1 (1.1)
Does the publication provide details of additional sources of support and information?	215 (37.1)	*325 (56.1)*	39 (6.7)	2.4 (1.2)	1.3 (0.9)	2.3 (1.1)	2.9 (1.5)	2.4 (1.0)
Does the publication refer to areas of uncertainty?	*313 (54.1)*	103 (17.8)	163 (28.2)	2.5 (1.7)	3.8 (1.6)	1.9 (1.5)	2.6 (1.7)	2.4 (1.8)
Content of information								
Does the publication describe how each treatment works?	196 (33.9)	146 (25.2)	*237 (40.9)*	3.1 (1.7)	1.2 (0.8)	2.5 (1.7)	3.9 (1.5)	3.2 (1.7)
Does the publication describe the benefits of each treatment?	*267 (46.1)*	253 (43.7)	59 (10.2)	2.2 (1.3)	1.3 (0.7)	2.2 (1.1)	2.5 (1.5)	2.3 (1.3)
Does the publication describe the risks of each treatment?	*437 (75.5)*	89 (15.4)	53 (9.2)	1.7 (1.3)	1.8 (1.4)	1.7 (1.2)	2.2 (1.6)	1.5 (1.1)
Does the publication describe what would happen if no treatment is used?	*540 (93.3)*	30 (5.2)	9 (1.6)	1.2 (0.7)	1.9 (1.5)	1.2 (0.6)	1.1 (0.6)	1.1 (0.5)
Does the publication describe how the treatment choices affect overall quality of life?	*276 (47.7)*	99 (17.1)	204 (35.2)	2.7 (1.8)	2.8 (1.7)	2.9 (1.9)	2.1 (1.6)	2.9 (1.8)
Is it clear that there may be >1 possible treatment choice?	*322 (55.6)*	119 (20.6)	138 (23.8)	2.4 (1.7)	3.3 (1.7)	2.2 (1.5)	3.1 (1.8)	2.1 (1.6)
Does the publication provide support for shared decision-making?	*316 (54.6)*	139 (24)	124 (21.4)	2.3 (1.6)	3.6 (1.5)	2.7 (1.7)	2.3 (1.6)	2.2 (1.6)
Overall quality								
On the basis of the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices	9 (1.6)	*570 (98.5)*	0 (0)	2.5 (0.6)	2.6 (0.6)	2.5 (0.5)	2.8 (0.7)	2.4 (0.6)

[a]Entries that did not fulfill the quality criteria.

[b]Entries that partially fulfilled the quality criteria.

XSL•FO
RenderX

[c]Entries that completely fulfilled the quality criteria.

[d]Italicization indicate the most frequent answer.

[e]This item was not scored if item 1 was scored as no.

Thematic Analysis

Overview

During inductive coding, four main themes were revealed: (1) pain and the burden of pain with 14.34% (1274/8886) of coding references, (2) neuromodulation as a treatment approach with 36.66% (3258/8886) of coding references, (3) device-related aspects with 19.38% (1722/8886) of coding references, and (4) patient benefits and testimonials of treatment with SCS with 29.62% (2632/8886) of coding references.

Pain and the Burden of Pain

In the first theme, pain and the burden of pain were described with 14.34% (1274/8886) of coding references from 56.7% (325/573) of coded entries. The first subtheme discussed pain and chronic pain with information on pain awareness, how to objectify pain, the definition of chronic pain, the duration of pain, potential causes of pain, the incidence of pain, a closer look at neuropathic pain, other types of pain, and the heterogeneity of pain. The second subtheme elaborated on other pain management techniques compared with SCS in which pain medication, epidural infiltrations or nerve blocks, spinal surgery, physical therapy, psychological therapy, or even a combination of techniques were described. The third subtheme evaluated how pain can affect an individual, with a focus on symptoms of pain, the misinterpretation of input signals as a painful experience during chronic pain, and potential implications on social life, daily living, bodily functions, and emotional distress. The final subtheme reported on patient experiences and stories of living with chronic pain. Within this subtheme, individual stories about pain physicians and quotes on how chronic pain feels are discussed. In addition, the burden of the time delay in finding the right help and the misunderstanding of other people who do not understand pain were categorized under this subtheme.

Neuromodulation as Treatment Approach

In the second theme, information about neuromodulation, in general, was pooled with 36.66% (3258/8886) of coding references from 86.9% (498/573) of coded entries. The first subtheme that was identified focused on access to neuromodulation, different techniques, targets of neuromodulation, limitations of these therapy options, history, underlying working mechanisms (eg, gate control theory), and the fact that neuromodulation masks pain and does not heal it. Patient-centered care was revealed as the second subtheme, in which a multidisciplinary approach including patient empowerment, independence, and education, with a holistic evaluation and shared decision-making with physicians and family members was discussed. Another subtheme consisted of SCS therapy, in which the definition of SCS was stated and the uncertainties with response to success rate (eg, trial pain relief vs pain relief after implantable pulse generator [IPG], responder rates, SCS outcome variables, and patient selection) and the combination of multiple neuromodulation devices was discussed. Furthermore, the SCS procedure itself was classified as a distinct subtheme and is presented in Figure 3. The final subtheme discussed the risks of the treatment (Table 2).

Figure 3. Identified topics about the spinal cord stimulation (SCS) procedure for chronic pain for which web-based information is available, as identified during the thematic analysis. IPG: implantable pulse generator.

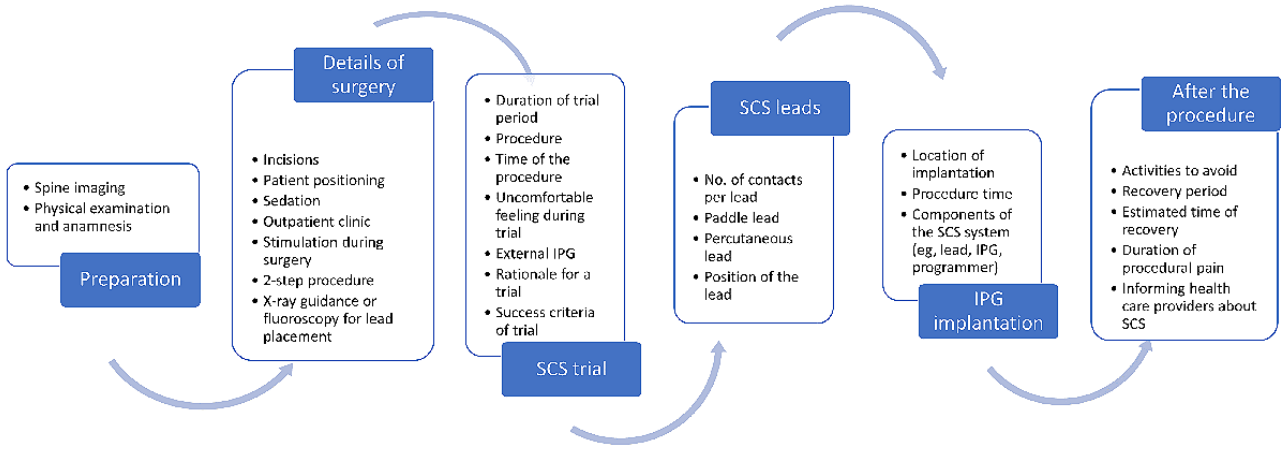

XSL•FO
RenderX

Table 2. Reported risks of spinal cord stimulation (SCS) for chronic pain that were revealed during the thematic analysis of the available web-based information. Counts present the number of entries that discussed each risk of SCS compared with the number of web-based entries in the thematic analysis (N=573).

Topic and risks of SCS	Frequency, n (%)
Pain	
Additional pain after procedure	2 (0.4)
Inadequate pain relief or loss of therapy effect	48 (8.4)
Adverse events and serious adverse events	
Allergic reactions	13 (2.3)
Bleeding	20 (3.5)
Dural puncture	4 (0.7)
Fluid accumulation	6 (1.1)
Scar tissue	7 (1.2)
Headache	7 (1.2)
Infection	77 (13.4)
Injuries to spinal cord	7 (1.2)
Nerve injury	13 (2.3)
Neurological deficit	5 (0.9)
Paralysis	21 (3.7)
Poor wound healing	1 (0.2)
Reaction to anesthetic agent	1 (0.2)
Skin erosion	1 (0.2)
Soreness or bruising	3 (0.5)
Mental distress	1 (0.2)
Death or injury	2 (0.4)
Device-related risks	
Positional sensitivity (ie, distance between electrode and spinal cord)	1 (0.2)
Hardware issues	22 (3.8)
Fracture of electrode	4 (0.7)
Electromagnetic interference	26 (4.5)
Disconnection electrode–IPG[a]	5 (0.9)
Lead migration	68 (11.9)
Pain at implantation site	34 (5.9)
Uncomfortable stimulation	13 (2.3)
General risks	
Complications are extremely rare	7 (1.2)

[a]IPG: implantable pulse generator.

Device-Related Aspects

In the third theme, device-related aspects were discussed (1722/8886, 19.38% of coding references from 379/573, 66.1% of coded entries), in which the charger and remote programming, the different device manufacturers of SCS, disadvantages of SCS, innovations, programming possibilities, types of stimulation, and scientific studies were included. Specifically focusing on the potential disadvantages of SCS, the following aspects were noted: difficulties with charging the IPG, magnetic resonance imaging (MRI) restrictions, restrictions with driving and operating machinery, insufficient coverage of the targeted region, implications on other medical procedures owing to the presence of SCS device, the need for remote control (ie, the lack of control through mobile phone), replacement of nonrechargeable IPG, detection of the SCS system at security gates, the potential of nausea, and the therapy burden in general (eg, need for a registration card and recharging process). The innovative aspects highlighted were as follows: accelerometry,

XSL•FO
RenderX

closed-loop stimulation, constantly evolving technology, improved battery life, increased competition in the market, increased market value of SCS manufacturing companies, MRI safety, the possibility of multiple waveform programming, new waveforms and software, more precise stimulation targets (ie, better coverage), rechargeable and nonrechargeable devices, running multiple SCS stimulation programs simultaneously, x-ray and computed tomography safety, and various sizes of IPGs.

Patient Benefits and Testimonials of Treatment With SCS

In the last theme (2632/8886, 29.62% of coding references from 514/573, 89.7% of coded entries), 3 subthemes were identified: benefits of SCS through patient experiences, candidates for SCS, and patient testimonials. The benefits of SCS that were reported are presented in Figure 4. The most frequently stated aspects were alternatives to opioids, pain relief, improved daily functioning, reduction in pain medication use, reversible therapy, and a minimally invasive approach. In the second subtheme about patient selection and candidate selection, contraindications were identified, including candidate checklists, insurance coverage, pain location, conditions that are suitable for SCS, pain description, pain relief during the trial period, psychological screening, selection process, tobacco use, and SCS as the last resort therapy. In the final subtheme, patient testimonials were combined, in which doubts about stimulation, expectations before SCS, the exchange of experiences and information, SCS coverage, living with an implanted device, pain scores, postoperative pain, and patient goals were discussed. Finally, the positive and negative experiences reported in the testimonials are presented in Textbox 1.

Table 3 presents a selection of quotes for each theme. Table 4 presents the frequency of themes and subthemes.

Figure 4. Benefits of a treatment trajectory with spinal cord stimulation (SCS) for chronic pain that were found during the thematic analysis of the web-based information. IPG: implantable pulse generator.

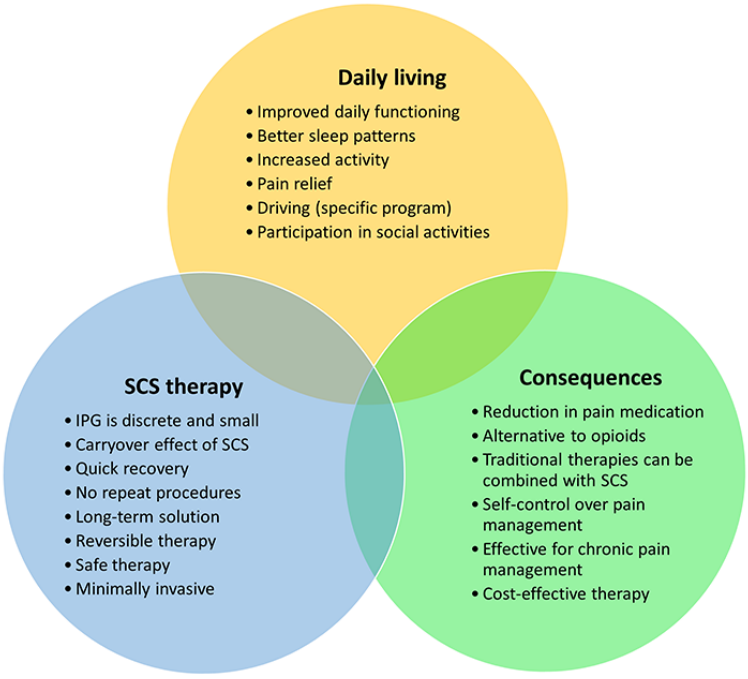

XSL•FO
RenderX

Textbox 1. Negative and positive experiences with spinal cord stimulation for chronic pain that were revealed during thematic analysis in patient testimonials. Counts present the number of entries that expressed these experiences.

Negative experiences

- Adjustments in stimulation are difficult (n=1)
- Lack of follow-up after surgery by physicians (n=2)
- Negative patient experiences (n=5)
- Negative experiences with insurance or compensation (n=6)
- Negative thoughts on spinal cord stimulation by patients who were never implanted (n=5)
- Reduced trust in health care (n=10)
- Wrong stimulation location (n=2)

Positive experiences

- Increase in activities of daily living (n=50)
- Increase in functionality (n=25)
- Return to work (n=7)
- Life-changing therapy (n=73)

XSL•FO
RenderX

Table 3. A selection of quotes that was revealed during the thematic analysis of available web-based information about spinal cord stimulation (SCS) for chronic pain.

Theme, subtheme, and context	Quote
Pain and the burden of pain	
Pain and chronic pain	
Definition of chronic pain presented by a health care provider.	"Chronic pain as pain that persists beyond the expected recovery period or pain that accompanies a chronic health condition. Because this pain is not productive and is not the result of an ongoing injury, it is referred to as pathological and is therefore treated as a condition, not just as a symptom."
How pain can affect an individual	
Penny, a woman who had pain for >10 years explained how the pain impacted her functioning.	"I rarely left the house. I could only just cope with the very basic way of living. I just really felt like I wanted to give up. It was just all too much."
Patient experiences and stories on living with chronic pain	
While reviewing her SCS system, Sarah explained how chronic pain makes her feel.	"But I mean, there's been lots of times that I have locked myself in my room and cried all night just because I couldn't imagine dealing with this pain one more hour. Obviously, it's not like that every day, but sometimes it is."
Neuromodulation as a treatment approach	
Patient-centered care	
Soula discussed about her involvement in her treatment and working together with the company representatives.	"And of course, I can have as many programs put on it. I basically work it out together with the programmers. Recently, for example, I went back and said: 'ah my feet are really sore. I really need more stimulation on my feet.' And so I had two new programs popped on there, and they've been fantastic."
Furthermore, Soula elaborated on her independence as a patient.	"That, to me, makes me feel I've got options of long term options that I don't have these big health appointment calendar that I have to meet and that all the things that I need for my management are tucked into my body look like who they are. You then have to go anywhere or reach for anything to some difficult. It really is just in me that makes me feel quite independent, makes me feel really strong about the way that I manage day to day."
Device-related aspects	
Potential disadvantages	
Reddit (Reddit, Inc) video	"Keep that thing charged. I've noticed the closer I get to empty, the less effective it is."
Pam discussed these elements in her video.	"Remember, it is a life changing experience in more than one way. It may help you a great deal. But the other issue that you have to deal with is having something implanted inside you that you that you have to carry a card to make sure that if you ever have a medical procedure like an MRI, that it's understood what kind of device you have and can you have an MRI if you travel, if you go off on a weekend, you have to be able to charge this device so you're tethered to a device."
Innovative aspects	
Example from a physician and researcher	"There is no limit for these indications, this is why I love neuromodulation, it's an evolving field and every six months at the latest we have something new not only in the spinal cord stimulation but also in all other options."
Patient benefits and testimonials of treatment with SCS	
Patient testimonials	
Eric's quote expressed a positive feeling.	"The way I'm living right now, if you would have known me five or six years ago, you would have never guessed. Before I couldn't walk from my front room to the kitchen and now running marathons and enjoying holiday company from everybody else."
Tom's quote represented some doubts.	"At first, I was a little reluctant. I didn't know how I would feel about something I was going to be, you know, inserted into my body. Is too much, is it too fast? That I realized that after the last 47 years, I really didn't have anything to risk. I wanted to feel better."
A Reddit (Reddit, Inc) post expressed a negative experience.	"So... Now it's been 3 months since surgery. It's not even coming close to relieving any pain."

XSL•FO
RenderX

Table 4. Frequency of themes and subthemes identified in web-based information about spinal cord stimulation (SCS) for chronic pain during the thematic analysis expressed by aggregated coding references (N=8886) and aggregated coded entries (n=573).

Themes and subthemes	Coding references, n (%)	Coded entries, n (%)
Pain and the burden of pain	1274 (14.3)	325 (56.7)
Pain and chronic pain	230 (2.6)	125 (21.8)
Pain management techniques	601 (6.8)	218 (38)
How pain can affect an individual	398 (4.5)	183 (31.9)
Patient experiences and stories on living with chronic pain	45 (0.5)	36 (6.3)
Neuromodulation as a treatment approach	3258 (36.7)	498 (86.9)
Neuromodulation	513 (5.8)	201 (35.1)
Patient-centered care	345 (3.9)	195 (34)
SCS therapy	480 (5.4)	303 (52.9)
SCS procedure	1437 (16.2)	320 (55.8)
Risks of the treatment	483 (5.4)	127 (22.2)
Device-related aspects	1722 (19.4)	379 (66.1)
Charger and remote programming	109 (1.2)	76 (13.3)
Different device manufacturers for SCS	190 (2.1)	113 (19.7)
Disadvantages of SCS	157 (1.8)	95 (16.6)
Innovations	554 (6.2)	208 (36.3)
Programming possibilities	220 (2.5)	117 (20.4)
Type of stimulation	407 (4.6)	200 (34.9)
Scientific studies	85 (1)	40 (7)
Patient benefits and testimonials of a treatment with SCS	2632 (29.6)	514 (89.7)
Benefits of SCS through patient experiences	1148 (12.9)	376 (65.6)
Candidates for SCS	1109 (12.5)	365 (63.7)
Patient testimonials	375 (4.2)	187 (32.6)

Discussion

Principal Findings

Patient-centered care, described as shared decision-making through patient autonomy, has drastically increased during the last decade [33]. Medical and health information contribute to equalizing the balance of power between patients and health care providers and thus to patient empowerment [25]. Through this approach, patients are supported to make informed and active choices concerning their health, compared with fulfilling the role of a passive recipient of care [34]. Empowered decision-making in treatment selection is inevitably related to appropriate communication and information tools and greater patient involvement in the decision-making process [35]. Therefore, this study evaluated the availability and content of information on the internet regarding SCS, which is a crucial prerequisite to enhance empowered decision-making in the field of neuromodulation.

The results of this study clearly indicated that most of the information originated in the United States (519/630, 82.4%). Besides the dominance of the United States, posts also originated in the United Kingdom (26/630, 4.1%), the Netherlands (16/630, 2.5%), France (11/630, 1.8%), Australia (11/630, 1.8%), Canada (9/630, 1.4%), and India (7/630, 1.1%). None of the included entries could be linked back to Africa or South America, potentially because of the language restrictions. The field of neuromodulation is driven by technical innovations by device manufacturers, with company headquarters located in the United States. This may be one of the potential explanations for the dominance of information in this region.

During thematic analysis, 4 major themes were identified, namely pain, neuromodulation as a treatment approach, device-related aspects, and patient benefits and testimonials. In these themes, important aspects in the field of neuromodulation were discussed, from patient selection, over the SCS trial period and IPG implantation, implanted materials, risks, benefits, and expectations to the place of neuromodulation within the treatment arsenal of patients with chronic pain. There were no gaps identified regarding crucial information that was lacking. The content of information that is available for patients is closely aligned with the latest research developments, for example, continued technological advancements [36], holistic evaluations [37], and ethical considerations, such as MRI conditional devices [38]. This indicates that recent research developments have been timely translated into layperson terms to make them available to the public, potentially because of the strong valorization of

XSL•FO
RenderX

project plans and effective marketing strategies, in contrast to the lack of or limited knowledge translation that is often observed [39,40]. Nevertheless, the availability of web-based information does not mean that this information is easily accessible to health care consumers. Previous research on patients with persistent spinal pain syndrome type 2 indicated that 85.3% had a low position in the social hierarchy (ie, a low social gradient of health) [41]. On the basis of qualitative research, patients are particularly interested in having information in layperson terms with concise language, contextualization of values, and definitions of medical terms to increase their understanding [42]. Nevertheless, in health care, it is rare to find a layperson knowledgeable in medical terminology, potentially leading to poor understanding of the medical conditions and potential treatments [43]. This emphasizes the need for layperson terms to ensure that populations with a low position in the social hierarchy have equal opportunities to enhance patient empowerment. Especially in the current digital era, patients develop toward so-called e-patients and are more prone to obtaining complementary web-based health information. Nevertheless, this evolution may give rise to larger interindividual differences based on socioeconomic status in cases where patients do not have access to health resources, eventually creating less-informed patients and, as such, enlarging disparities. In addition, the availability of information does not imply that information is centralized; therefore, real-time capturing of the web-based information-seeking process should be conducted [44]. This advocates further evaluation of the web-based information-seeking process to determine whether patients have access to information and whether the information is effectively understandable by the target population.

Besides an evaluation of the availability of information, the provided information should also be reliable to trust the obtained information. The DISCERN instrument was used to conduct a quality appraisal [45], where poor scores were mainly because of a lower scoring on the content of information than on the reliability. Overall, information was deemed reliable, where creators of new content should primarily focus on clearly defining the aims of their posts, videos, or web pages and the areas of uncertainty to further enhance the reliability of web-based information. During the evaluation of the content, the broader perspective of the therapeutic intervention was explored (ie, alternative treatment options, risks of every treatment option, what if treatment is not applied, and shared decision-making). Neuromodulation techniques, including SCS, are often considered a last resort therapy after all other pain management procedures fail to alleviate the pain. It may be hypothesized that alternative treatments and risks, quality of life, and benefits of other treatments were less frequently discussed as SCS is a final pain management (ie, pain control) option and not a therapeutic (ie, healing) option, where several techniques could be applied to treat a specific condition. Nevertheless, despite the poor scoring of many entries on the content of information, there are sources where the information is available because the results of the content analysis could not identify an obvious knowledge gap. Combining these results, the field may benefit from more centralized pages or complete overviews of neuromodulation techniques in perspective to the complete therapeutic arsenal of pain management techniques.

The relationship among health care providers, patients, and industrial device representatives is a *conditio sine qua non* in the field of neuromodulation. Despite the disputable role of industry-sponsored randomized controlled trials of drugs and medical devices toward more favorable results compared with non–industry-sponsored trials [46] and the influence of medical promotional tools provided by medical representatives on physicians' prescribing practices [47], in total, 43.5% of the retrieved information referred to a specific industrial partner (directly or indirectly). Notwithstanding the considerable involvement of device manufacturers, web-based information is not limited to commercial information about the different types of materials and IPGs, as revealed by the heterogeneity of detected themes and subthemes in the content analysis. This clearly indicates that there is a difference in bias introduced by industry funding versus information that is available to the public.

This study indicated that a lot of information about SCS is available on the web, where it should be clearly stated that this information only provides an additional resource for patients besides the traditional health care system and consultation with multidisciplinary pain management teams. While interpreting these results, a couple of limitations should be considered. Notably, only web-based information that was provided in English, French, or Dutch was retrieved from our search. In addition, web-based information that is spread through LinkedIn (Microsoft Corporation) was underrepresented because information retrieval from LinkedIn is not fully supported owing to the limitations of the official application programming interface (retrieved through personal communication with an Awario representative). In addition, social listening only reflects the narratives and discourse occurring in public, internet-based spaces and may not mirror information from closed platforms and non–internet-based channels [48]. Finally, within the search strategy, the most common synonyms of SCS were listed; however, materials that extend beyond the terms used will not have been retracted with the current search strategy.

Conclusions

This social listening study regarding the availability of web-based information about SCS demonstrated that health care consumers have the possibility to access web-based information about SCS, including details about the surgical procedures, type of material, working mechanisms, risks, patient expectations, testimonials, and the potential benefits of this therapy option. The availability of information, however, does not evaluate whether information is effectively retrieved by consumers of health care and whether it is easily accessible, which needs to be further evaluated. In addition, quality control of web-based information is extremely difficult, and therefore, reliability, trustworthiness, and correctness of sources should be carefully considered before automatically relying on the content of the sources.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Jonas Callens for screening entries as a second reviewer.

Data Availability

The data set generated and analyzed during this study is available from the corresponding author upon reasonable request.

Conflicts of Interest

MM received speaker fees from Medtronic, Nevro, and Saluda Medical. PR reports grants and consultant fees from Medtronic, Abbott, and Boston Scientific outside the submitted work. LG is a postdoctoral research fellow funded by the Research Foundation of Flanders, Belgium (project number 12ZF622N). STIMULUS research group received independent research grants from Medtronic. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
World maps and entry types for industry-related and non–industry-related entries.
[DOCX File , 244 KB - jmir_v26i1e48599_app1.docx]

Multimedia Appendix 2
World maps present the locations of the 4 prominent companies using neuromodulation devices.
[DOCX File , 427 KB - jmir_v26i1e48599_app2.docx]

Multimedia Appendix 3
Word cloud based on the transcripts of the included entries from the social listening study about spinal cord stimulation for chronic pain.
[PNG File , 1073 KB - jmir_v26i1e48599_app3.png]

References

1. Grider JS, Manchikanti L, Carayannopoulos A, Sharma ML, Balog CC, Harned ME, et al. Effectiveness of spinal cord stimulation in chronic spinal pain: a systematic review. Pain Physician 2016 Jan;19(1):E33-E54 [FREE Full text] [Medline: 26752493]
2. Hofmeister M, Memedovich A, Brown S, Saini M, Dowsett LE, Lorenzetti DL, et al. Effectiveness of neurostimulation technologies for the management of chronic pain: a systematic review. Neuromodulation 2020 Feb;23(2):150-157. [doi: 10.1111/ner.13020] [Medline: 31310417]
3. Kumar K, Taylor RS, Jacques L, Eldabe S, Meglio M, Molet J, et al. The effects of spinal cord stimulation in neuropathic pain are sustained: a 24-month follow-up of the prospective randomized controlled multicenter trial of the effectiveness of spinal cord stimulation. Neurosurgery 2008 Oct;63(4):762-70; discussion 770. [doi: 10.1227/01.NEU.0000325731.46702.D9] [Medline: 18981888]
4. North RB, Kidd DH, Farrokhi F, Piantadosi SA. Spinal cord stimulation versus repeated lumbosacral spine surgery for chronic pain: a randomized, controlled trial. Neurosurgery 2005;56(1):98-106; discussion 106. [doi: 10.1227/01.neu.0000144839.65524.e0] [Medline: 15617591]
5. Hunter CW, Carlson J, Yang A, Deer T. Spinal cord stimulation for the treatment of failed neck surgery syndrome: outcome of a prospective case series. Neuromodulation 2018 Jul;21(5):495-503. [doi: 10.1111/ner.12769] [Medline: 29566313]
6. Lamer TJ, Moeschler SM, Gazelka HM, Hooten WM, Bendel MA, Murad MH. Spinal stimulation for the treatment of intractable spine and limb pain: a systematic review of RCTs and meta-analysis. Mayo Clin Proc 2019 Aug;94(8):1475-1487. [doi: 10.1016/j.mayocp.2018.12.037] [Medline: 31279543]
7. Thomson S, Huygen F, Prangnell S, De Andrés J, Baranidharan G, Belaïd H, et al. Appropriate referral and selection of patients with chronic pain for spinal cord stimulation: European consensus recommendations and e-health tool. Eur J Pain 2020 Jul;24(6):1169-1181 [FREE Full text] [doi: 10.1002/ejp.1562] [Medline: 32187774]
8. Kumar K, Taylor RS, Jacques L, Eldabe S, Meglio M, Molet J, et al. Spinal cord stimulation versus conventional medical management for neuropathic pain: a multicentre randomised controlled trial in patients with failed back surgery syndrome. Pain 2007 Nov;132(1-2):179-188. [doi: 10.1016/j.pain.2007.07.028] [Medline: 17845835]
9. Zhang TC, Janik JJ, Grill WM. Mechanisms and models of spinal cord stimulation for the treatment of neuropathic pain. Brain Res 2014 Jun 20;1569:19-31. [doi: 10.1016/j.brainres.2014.04.039] [Medline: 24802658]
10. Rigoard P, Gatzinsky K, Deneuville JP, Duyvendak W, Naiditch N, Van Buyten JP, et al. Optimizing the management and outcomes of failed back surgery syndrome: a consensus statement on definition and outlines for patient assessment. Pain Res Manag 2019;2019:3126464. [doi: 10.1155/2019/3126464] [Medline: 30911339]

XSL•FO
RenderX

11. Connell NB, Prathivadi P, Lorenz KA, Zupanc SN, Singer SJ, Krebs EE, et al. Teaming in interdisciplinary chronic pain management interventions in primary care: a systematic review of randomized controlled trials. J Gen Intern Med 2022 May;37(6):1501-1512 [FREE Full text] [doi: 10.1007/s11606-021-07255-w] [Medline: 35239110]
12. Gatzinsky K, Eldabe S, Deneuville JP, Duyvendak W, Naiditch N, Van Buyten JP, et al. Optimizing the management and outcomes of failed back surgery syndrome: a proposal of a standardized multidisciplinary team care pathway. Pain Res Manag 2019;2019:8184592 [FREE Full text] [doi: 10.1155/2019/8184592] [Medline: 31360272]
13. Levack WM, Weatherall M, Hay-Smith EJ, Dean SG, McPherson K, Siegert RJ. Goal setting and strategies to enhance goal pursuit for adults with acquired disability participating in rehabilitation. Cochrane Database Syst Rev 2015 Jul 20;2015(7):CD009727 [FREE Full text] [doi: 10.1002/14651858.CD009727.pub2] [Medline: 26189709]
14. Werbrouck A, Swinnen E, Kerckhofs E, Buyl R, Beckwée D, De Wit L. How to empower patients? A systematic review and meta-analysis. Transl Behav Med 2018 Sep 08;8(5):660-674. [doi: 10.1093/tbm/iby064] [Medline: 29982675]
15. Austin JT, Vancouver JB. Goal constructs in psychology: structure, process, and content. Psychol Bull 1996 Nov 21;120(3):338-375. [doi: 10.1037/0033-2909.120.3.338]
16. Goudman L, Bruzzo A, van de Sande J, Moens M. Goal identification before spinal cord stimulation: a qualitative exploration in potential candidates. Pain Pract 2020 Mar;20(3):247-254. [doi: 10.1111/papr.12845] [Medline: 31610085]
17. Henssen DJ, Scheepers N, Kurt E, Arnts I, Steegers M, Vissers K, et al. Patients' expectations on spinal cord stimulation for failed back surgery syndrome: a qualitative exploration. Pain Pract 2018 Apr;18(4):452-462. [doi: 10.1111/papr.12617] [Medline: 28742245]
18. El-Haddad C, Hegazi I, Hu W. Understanding patient expectations of health care: a qualitative study. J Patient Exp 2020 Dec 28;7(6):1724-1731 [FREE Full text] [doi: 10.1177/2374373520921692] [Medline: 33457636]
19. Fitzsimmons PR, Michael BD, Hulley JL, Scott GO. A readability assessment of online Parkinson's disease information. J R Coll Physicians Edinb 2010 Dec;40(4):292-296. [doi: 10.4997/JRCPE.2010.401] [Medline: 21132132]
20. Tanenbaum SJ. Perspectives on evidence-based practice from consumers in the US public mental health system. J Eval Clin Pract 2008 Oct;14(5):699-706. [doi: 10.1111/j.1365-2753.2008.01020.x] [Medline: 19018898]
21. Lim YZ, Chou L, Au RT, Seneviwickrama KM, Cicuttini FM, Briggs AM, et al. People with low back pain want clear, consistent and personalised information on prognosis, treatment options and self-management strategies: a systematic review. J Physiother 2019 Jul;65(3):124-135 [FREE Full text] [doi: 10.1016/j.jphys.2019.05.010] [Medline: 31227280]
22. Cline RJ, Haynes KM. Consumer health information seeking on the internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
23. Terrens AF, Soh SE, Morgan P. What web-based information is available for people with Parkinson's disease interested in aquatic physiotherapy? A social listening study. BMC Neurol 2022 May 05;22(1):170 [FREE Full text] [doi: 10.1186/s12883-022-02669-3] [Medline: 35513789]
24. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 2014 Dec 02;16(12):e262 [FREE Full text] [doi: 10.2196/jmir.3706] [Medline: 25470306]
25. Eeckman E. Power to the patient? Studying the balance of power between patient and GP in relation top web health information. In: Murru MF, Colombo F, Peja L, Tosoni S, Kilborn R, Kunelius R, et al, editors. Communication as the Intersection of the Old and the New: The Intellectual Work of the 2018 European Media and Communication Doctoral Summer School. Bremen, Germany: Edition Lumière; 2019.
26. Tonsaker T, Bartlett G, Trpkov C. Health information on the internet: gold mine or minefield? Can Fam Physician 2014 May;60(5):407-408 [FREE Full text] [Medline: 24828994]
27. Battineni G, Baldoni S, Chintalapudi N, Sagaro GG, Pallotta G, Nittari G, et al. Factors affecting the quality and reliability of online health information. Digit Health 2020;6:2055207620948996 [FREE Full text] [doi: 10.1177/2055207620948996] [Medline: 32944269]
28. Stewart MC, Arnold CL. Defining social listening: recognizing an emerging dimension of listening. Int J Listening 2017 Jun 14;32(2):85-100. [doi: 10.1080/10904018.2017.1330656]
29. Awario home page. Awario. 2022 Jul 31. URL: https://awario.com [accessed 2024-01-15]
30. O'Sullivan D, Wilk S, Michalowski W, Farion K. Using PICO to align medical evidence with MDs decision making models. Stud Health Technol Inform 2013;192:1057. [Medline: 23920831]
31. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
32. Hong SW, Kang JH, Park JH, Park HJ, Kim E. Quality and readability of online information on hand osteoarthritis. Health Informatics J 2023;29(1):14604582231169297 [FREE Full text] [doi: 10.1177/14604582231169297] [Medline: 36995242]
33. Grover S, Fitzpatrick A, Azim FT, Ariza-Vega P, Bellwood P, Burns J, et al. Defining and implementing patient-centered care: an umbrella review. Patient Educ Couns 2022 Jul;105(7):1679-1688. [doi: 10.1016/j.pec.2021.11.004] [Medline: 34848112]

XSL•FO
RenderX

34. Kane PM, Murtagh FE, Ryan K, Mahon NG, McAdam B, McQuillan R, et al. The gap between policy and practice: a systematic review of patient-centred care interventions in chronic heart failure. Heart Fail Rev 2015 Nov;20(6):673-687 [FREE Full text] [doi: 10.1007/s10741-015-9508-5] [Medline: 26435042]
35. Burtchell J, Clemmons D, Clemmons J, Sabutis T, Rosenberg A, Graves J, et al. A targeted literature search and phenomenological review of perspectives of people with multiple sclerosis and healthcare professionals of the immunology of disease-modifying therapies. Neurol Ther 2022 Sep;11(3):955-979 [FREE Full text] [doi: 10.1007/s40120-022-00349-5] [Medline: 35608740]
36. Lempka SF, Patil PG. Innovations in spinal cord stimulation for pain. Curr Opin Biomed Eng 2018 Dec;8:51-60 [FREE Full text] [doi: 10.1016/j.cobme.2018.10.005] [Medline: 30911705]
37. Levy RM, Mekhail N, Abd-Elsayed A, Abejón D, Anitescu M, Deer TR, et al. Holistic treatment response: an international expert panel definition and criteria for a new paradigm in the assessment of clinical outcomes of spinal cord stimulation. Neuromodulation 2023 Jul;26(5):1015-1022 [FREE Full text] [doi: 10.1016/j.neurom.2022.11.011] [Medline: 36604242]
38. Goudman L, Rigoard P, Billot M, Duarte RV, Eldabe S, Moens M. Patient selection for spinal cord stimulation in treatment of pain: sequential decision-making model - a narrative review. J Pain Res 2022;15:1163-1171 [FREE Full text] [doi: 10.2147/JPR.S250455] [Medline: 35478997]
39. Straus SE, Tetroe J, Graham I. Defining knowledge translation. CMAJ 2009 Aug 04;181(3-4):165-168 [FREE Full text] [doi: 10.1503/cmaj.081229] [Medline: 19620273]
40. Connelly S, Vanderhoven D, Rutherfoord R, Richardson L, Matthews P. Translating research for policy: the importance of equivalence, function, and loyalty. Humanit Soc Sci Commun 2021 Aug 03;8(1):191. [doi: 10.1057/s41599-021-00873-z]
41. Naiditch N, Billot M, Moens M, Goudman L, Cornet P, Le Breton D, et al. Persistent spinal pain syndrome type 2 (PSPS-T2), a social pain? Advocacy for a social gradient of health approach to chronic pain. J Clin Med 2021 Jun 25;10(13):2817 [FREE Full text] [doi: 10.3390/jcm10132817] [Medline: 34202362]
42. Perlis N, Finelli A, Lovas M, Berlin A, Papadakos J, Ghai S, et al. Creating patient-centered radiology reports to empower patients undergoing prostate magnetic resonance imaging. Can Urol Assoc J 2021 Apr;15(4):108-113 [FREE Full text] [doi: 10.5489/cuaj.6585] [Medline: 33007175]
43. Ibrahim M, Gauch S, Salman O, Alqahtani M. An automated method to enrich consumer health vocabularies using GloVe word embeddings and an auxiliary lexical resource. PeerJ Comput Sci 2021 Aug 9;7:e668 [FREE Full text] [doi: 10.7717/peerj-cs.668] [Medline: 34458573]
44. Marcu A, Muller C, Ream E, Whitaker KL. Online information-seeking about potential breast cancer symptoms: capturing online behavior with an internet browsing tracking tool. J Med Internet Res 2019 Feb 06;21(2):e12400 [FREE Full text] [doi: 10.2196/12400] [Medline: 30724741]
45. Charnock D, Shepperd S. Learning to DISCERN online: applying an appraisal tool to health websites in a workshop setting. Health Educ Res 2004 Aug;19(4):440-446. [doi: 10.1093/her/cyg046] [Medline: 15155597]
46. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev 2017 Feb 16;2(2):MR000033 [FREE Full text] [doi: 10.1002/14651858.MR000033.pub3] [Medline: 28207928]
47. Ali KE, Naser AY, Al-Rousan R, Alwafi H, AbuAlhommos AK, Alsairafi ZK, et al. The attitude and acceptability towards medical promotional tools and their influence on physicians' prescribing practices in Jordan and Iraq: a cross-sectional study. BMC Health Serv Res 2022 Jan 25;22(1):105 [FREE Full text] [doi: 10.1186/s12913-022-07525-1] [Medline: 35078461]
48. Silva M, Anaba U, Jani Tulsani N, Sripad P, Walker J, Aisiri A. Gender-based violence narratives in internet-based conversations in Nigeria: social listening study. J Med Internet Res 2023 Sep 15;25:e46814 [FREE Full text] [doi: 10.2196/46814] [Medline: 37713260]

Abbreviations

IPG: implantable pulse generator
MRI: magnetic resonance imaging
SCS: spinal cord stimulation

XSL•FO
RenderX

Edited by A Mavragani; submitted 29.04.23; peer-reviewed by M Barbosa, D Carvalho; comments to author 02.11.23; revised version received 03.11.23; accepted 28.11.23; published 30.01.24.

Please cite as:
Moens M, Van Doorslaer L, Billot M, Eeckman E, Roulaud M, Rigoard P, Fobelets M, Goudman L
Examining the Type, Quality, and Content of Web-Based Information for People With Chronic Pain Interested in Spinal Cord Stimulation: Social Listening Study
J Med Internet Res 2024;26:e48599
URL: https://www.jmir.org/2024/1/e48599
doi:10.2196/48599
PMID:38289645

©Maarten Moens, Leen Van Doorslaer, Maxime Billot, Edgard Eeckman, Manuel Roulaud, Philippe Rigoard, Maaike Fobelets, Lisa Goudman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Patient-Centered eHealth Information on Erosive Tooth Wear: Systematic Search and Evaluation of Websites and YouTube Videos

Lena Holland[1]; Amelie Friederike Kanzow[2], MEd; Annette Wiegand[1], Prof Dr med dent; Philipp Kanzow[1], MSc, Dr rer medic, PD Dr med dent

[1]Department of Preventive Dentistry, Periodontology and Cariology, University Medical Center Göttingen, Göttingen, Germany

[2]Study Deanery, University Medical Center Göttingen, Göttingen, Germany

Corresponding Author:
Philipp Kanzow, MSc, Dr rer medic, PD Dr med dent
Department of Preventive Dentistry, Periodontology and Cariology
University Medical Center Göttingen
Robert-Koch-Str 40
Göttingen, 37075
Germany
Phone: 49 551 3960870
Fax: 49 551 3960869
Email: philipp.kanzow@med.uni-goettingen.de

Abstract

Background: Due to the declining prevalence of dental caries, noncarious tooth defects such as erosive tooth wear have gained increased attention over the past decades. While patients more frequently search the internet for health-related information, the quality of patient-centered, web-based health information on erosive tooth wear is currently unknown.

Objective: This study aimed to assess the quality of patient-centered, web-based health information (websites and YouTube videos) on erosive tooth wear.

Methods: German-language websites were systematically identified through 3 electronic search engines (google.de, bing.de or yahoo.de, and duckduckgo.com) in September 2021. Eligible websites were independently assessed for (1) technical and functional aspects via the LIDA instrument, (2) readability via the Flesch reading-ease score, (3) comprehensiveness of information via a structured checklist, and (4) generic quality and risk of bias via the DISCERN instrument by 2 different reviewers. An overall quality score (ie, higher scores being favored) generated from all 4 domains was used as the primary outcome. Quality scores from each domain were separately analyzed as secondary outcomes and compared by the Friedman test. The effect of practice-specific variables on quality scores of websites from private dental offices was assessed using generalized linear modeling. Eligible YouTube videos were judged based on (1) the comprehensiveness of information, (2) viewers' interaction, and (3) viewing rate. The comprehensiveness of information was compared between websites and YouTube videos using the Wilcoxon rank-sum test.

Results: Overall, 231 eligible websites and 7 YouTube videos were identified and assessed. The median overall quality of the websites was 33.6% (IQR 29.8%-39.2%). Secondary outcome scores amounted to 64.3% (IQR 59.8%-69.0%) for technical and functional aspects, 40.0% (IQR 34.0%-49.0%) for readability, 11.5% (IQR 3.9%-26.9%) for comprehensiveness of information, and 16.7% (IQR 8.3%-23.3%) for generic quality. While the comprehensiveness of information and generic quality received low scores, technical and functional aspects as well as readability resulted in higher scores (both $P_{adjusted}$<.001). Regarding practice-specific variables, websites from private dental offices outside Germany (P=.04; B=–6.64, 95% CI –12.85 to –0.42) or from dentists who are a dental society member (P=.049; B=–3.55, 95% CI –7.09 to –0.01) resulted in lower readability scores (ie, were more difficult to read), while a shorter time since dentists' examination resulted in higher readability scores (P=.01; B=0.24 per year, 95% CI 0.05-0.43). The comprehensiveness of information from YouTube videos was 34.6% (IQR 13.5%-38.5%). However, the comprehensiveness of information did not vary between websites and YouTube videos (P=.09). Additionally, viewers' interaction (1.7%, IQR 0.7%-3.4%) and viewing rates (101%, IQR 54.6%-112.6%) were low.

XSL•FO
RenderX

Conclusions: The quality of German-language, patient-centered, web-based information on erosive tooth wear was limited. Especially, the comprehensiveness and trustworthiness of the available information were insufficient. Web-based information on erosive tooth wear requires improvement to inform patients comprehensively and reliably.

(J Med Internet Res 2024;26:e49514) doi:10.2196/49514

KEYWORDS

consumer health information; dental erosion; dental sciences; digital media; erosive tooth wear; evidence-based dentistry; health education; information quality; internet; shared decision making

Introduction

Patients search the internet for health-related information to an increased extent [1], and web-based health information impact the physician-patient relationship [2]. High-quality content empowers patients to participate in an informed shared decision-making process [3]. To be beneficial for patients, web-based health information should be easily accessible and readable, comprehensive, and trustworthy (ie, being credible and unbiased).

In addition, dental patients often search for dental or oral conditions and treatment procedures using the internet [4,5]. Within the context of preventive and restorative dentistry, multiple studies focused on patient-centered, web-based health information. Each study assessed information on specific oral conditions and treatment procedures, including endodontics [6-11], dental caries [12-17], pit and fissure sealant application [18], restoration repair [19], and periodontitis [20-24].

Due to the declining prevalence of dental caries over the past decades [25], noncarious tooth defects such as erosive tooth wear have gained increased attention. Erosive tooth wear is defined as tooth wear with dental erosion as the primary etiological factor. While tooth decay is caused by acids derived from oral microorganisms, dental erosion is the loss of mineralized tooth substance due to exposure to acids not derived from oral bacteria [26]. The erosive demineralization might be caused either by extrinsic (eg, acids in food or beverages) or intrinsic (ie, endogenous acid) acids [27]. A recent review estimated a global prevalence between 20% and 45% for permanent teeth [28]. In Germany, a prevalence of 44.8% in adults aged 35 to 44 years was reported [29]. Advanced erosive tooth wear often requires restorative management by direct or indirect dental restorations to reduce pain and dentine hypersensitivity and to restore esthetic and function [30].

To the best of our knowledge, no previous study assessed patient-centered, web-based information on noncarious tooth defects such as erosive tooth wear. Consequently, it is currently unknown whether patients might find comprehensive and trustworthy information on erosive tooth wear using the internet.

Therefore, this study aimed to analyze patient-centered, web-based health information (ie, websites and YouTube videos) on erosive tooth wear published by different content providers and to assess its quality and credibility. Websites were scored across four domains: (1) technical and functional aspects, (2) readability, (3) comprehensiveness of information, and (4) generic quality and risk of bias. YouTube videos were judged based on (1) comprehensiveness of information, (2) viewers' interaction, and (3) viewing rate. Another objective was to identify practice-specific variables, which might impact the quality of websites provided by private dental offices. The null hypotheses were that information quality and credibility do not vary across different content providers, and practice-specific variables of private dental offices do not impact information quality and credibility.

Methods

Ethical Considerations

The reporting of this study is in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and the ENTREQ (Enhancing Transparency in Reporting the Synthesis of Qualitative Research) statements [31,32]. The study was approved by the local ethics committee of the University Medical Center Göttingen (approval 3/11/23). Only publicly available, nonsensitive data were used.

Eligible Sources

Freely accessible websites and YouTube videos in German language published from (1) private dental offices; (2) corporate dental offices or private hospital groups resembling juridical entities (ie, Medizinisches Versorgungszentrum, Gesellschaft mit beschränkter Haftung, and Aktiengesellschaft); (3) public dental clinics or dental schools; (4) dental societies, dental regulatory bodies, public bodies, or insurance companies; or (5) information services (ie, other commercial or nonprofit content providers) containing patient-centered health information on erosive tooth wear were regarded as eligible sources.

Websites and YouTube videos from other content providers (eg, dental laboratories, dental supply and material companies, and manufacturers of dental hygiene products), advertisements, articles, books, research agencies without focus on patient care, and forums or blogs operated by nondentists were excluded. In addition, YouTube videos without sound or captions and videos with a duration of >15 minutes were not considered as eligible sources.

Search Strategy

Websites were identified through 3 electronic search engines (google.de, bing.de or yahoo.de, and duckduckgo.com). The searches were performed on September 21 and 22, 2021, using 6 different search terms representing different German synonyms for male and female dentists, dentists, dental offices, dental erosion, erosive tooth wear, and acid-related tooth destruction, in both professional and lay language (Table 1). The same search terms were used to identify YouTube videos from Youtube.de on September 29, 2021.

XSL•FO
RenderX

Table 1. Search terms used and number of results for each search engine.

Search term[a]	Google (google.de), n	Bing or Yahoo (bing.de or yahoo.de), n	DuckDuckGo (duckduckgo.com), n
(Zahnärztin OR Zahnarzt OR Zahnärzte OR Zahnarztpraxis) Erosion	200	602	206
(Zahnärztin OR Zahnarzt OR Zahnärzte OR Zahnarztpraxis) Erosionen	200	470	224
(Zahnärztin OR Zahnarzt OR Zahnärzte OR Zahnarztpraxis) “erosive Zahnhartsubstanzdefekte”	34	27	30
(Zahnärztin OR Zahnarzt OR Zahnärzte OR Zahnarztpraxis) “erosiver Zahnhartsubstanzverlust”	25	18	15
(Zahnärztin OR Zahnarzt OR Zahnärzte OR Zahnarztpraxis) “Säureschädigung”	41	62	40
(Zahnärztin OR Zahnarzt OR Zahnärzte OR Zahnarztpraxis) “Säureschäden”	157	283	177

[a]Search terms represent different German synonyms for male and female dentists, dentists, dental offices, dental erosion, erosive tooth wear, and acid-related tooth destruction, in both professional and lay language.

For all searches, a computer running Windows 10 Home (Microsoft Inc) and Firefox version 92.0 (Mozilla Foundation) connected to the internet in Germany was used. Prior to each search, the browser cache, cookies, and browser history were cleared. All searches were performed using search engines’ standard settings. Only web pages being displayed as “most relevant” were considered.

Based on the search results and shown previews, web pages not meeting the inclusion criteria were excluded, and potentially eligible web pages were screened in full. Subsequently, noneligible web pages and duplicates were removed. Multiple web pages published under the same domain were jointly assessed as 1 website. The search and inclusion of websites was performed by 1 author (LH) and verified by another author (PK). Any uncertainties were resolved through discussion.

Data Extraction

The following data were extracted from each website or YouTube video using a pilot-tested spreadsheet (Multimedia Appendix 1), if available: (1) content provider’s name, (2) URL, (3) content provider’s country (ie, Germany or foreign country), (4) content provider’s location (ie, rural, town [<100,000 inhabitants], or city [≥100,000 inhabitants]), (5) content provider (ie, private dental office, corporate dental office or private hospital group, public dental clinic or dental school, dental society or dental regulatory body or public body or insurance company, or information service).

If a website or YouTube video was published from a private dental office, the following data were also extracted: (1) practice setting (ie, single practitioner or multiple dentists), (2) sex (ie, female, male, or mixed [in case of multiple dentists]), (3) dental society memberships, and (4) years of examination (averaged in case of multiple dentists). In addition, in the case of YouTube videos, (1) upload date, (2) duration (minutes), (3) number of likes, (4) number of dislikes, and (5) number of comments were extracted. Regarding dental society memberships, no differentiation between different fields (eg, restorative dentistry) was made.

If information on memberships in dental associations was missing or unclear, it was compared with the publicly available membership data of the German Society of Dentistry and Oral Medicine (*Deutsche Gesellschaft für Zahn-, Mund- und Kieferheilkunde*; DGZMK) [33] and the Swiss Dental Association (SSO) [34]. In addition, the years of examination were cross-referenced from older versions of the respective websites accessed via the Internet Archive’s Wayback Machine [35], the information provided from dental regulatory bodies, or publicly available curriculum vitae (ie, published in dentists’ dissertations, social networking sites XING or LinkedIn, profiles at doktor.ch [36], or directories of professional representations). Data extraction was independently performed by 2 authors (LH and PK). Subsequently, any discrepancies were resolved by repeated data extraction.

Outcomes

Quality and credibility of health information on websites were assessed across four domains: (1) technical and functional aspects, (2) readability, (3) comprehensiveness of information, and (4) generic quality and risk of bias. Domains 1, 2, and 4 were scored using established and validated tools. Domain 1 was scored using the LIDA validation instrument (version 1.2; Minervation) for judging the accessibility (1.1), usability (1.2), and reliability (1.3) of health care websites [37]. Domain 2 was scored according to the following equation:

which represents the Flesch reading-ease score (FRES) [38], adapted by Amstad [39] for the German language. Domain 4 was scored using the DISCERN instrument for judging the reliability (4.1) and quality (4.2) of written consumer health information [40].

For domain 3, a structured checklist with 4 subdomains focusing on the etiology and pathogenesis (3.1), clinical signs (3.2), preventive measures (3.3), and therapy options (3.4) of erosive tooth wear was developed by 2 expert clinicians (AW and PK).

The manually scored domains (ie, domains 1, 3, and 4) were independently evaluated by multiple authors (domains 1, 3, and 4: LH; domains 1 and 4: AFK; and domain 3: PK) by applying the scoring guidelines and definitions of each tool. For each tool, the originally published scales and scoring criteria were used. Finally, any discrepancies were resolved through discussion. Domain 2 was evaluated using an automated free

XSL•FO
RenderX

web-based tool [41] calculating the FRES on a scale ranging from 0 to 100.

For YouTube videos, the comprehensiveness of information was also explored. Again, 2 reviewers (LH and PK) independently assessed each video based on the same structured checklist. Furthermore, viewers' interaction and the viewing rate were calculated according to the following equations:

which have been described previously [42].

Statistical Analysis

Statistical analysis was performed using SPSS Statistics (Macintosh version 29.0.0.0; IBM Corp). The level of statistical significance was set at $P<.05$.

For all manually scored domains (ie, domains 1, 3, and 4), interrater reliability was analyzed using 2-way agreement, average intraclass correlation (ICC[A,2]) [43]. As the scales and numbers of items differed between each domain, relative percentage scores of the maximum possible score sums were used to allow for comparison between domains. While an overall quality score for each website was calculated by averaging the results from all 4 domains (primary outcome), relative percentage scores from each domain were used as secondary outcomes.

Descriptive analysis of quality scores consisted of median, IQR, and range between minimum and maximum scores. Potential differences between each quality domain were assessed using the Friedman test followed by the Dunn-Bonferroni post hoc tests. In addition, statistical differences across content providers (ie, private dental offices, corporate dental offices or private hospital groups, public dental clinics or dental schools, dental societies or dental regulatory bodies or public bodies or insurance companies, or information services) were analyzed by Kruskal-Wallis rank-sum tests followed by Dunn-Bonferroni post hoc tests. For websites posted by private dental offices, the impact of practice-specific variables on both the primary and secondary outcomes was assessed using generalized linear modeling (GLM). Within the multivariable analysis, covariates were entered simultaneously, and only main effects without interaction terms were tested. In case of missing information regarding dental society memberships and years of examination for all dentists, the dental offices' respective variables were either scored as "no" or websites were excluded from the multivariable analysis (n=9). Finally, the comprehensiveness of information was compared between websites and YouTube videos using the Wilcoxon rank-sum test.

Results

Overview

A total of 231 eligible websites and 7 YouTube videos were identified. A flowchart representing the search is shown in Figure 1.

Figure 1. Flowchart representing the systematic workflow of the evaluation.

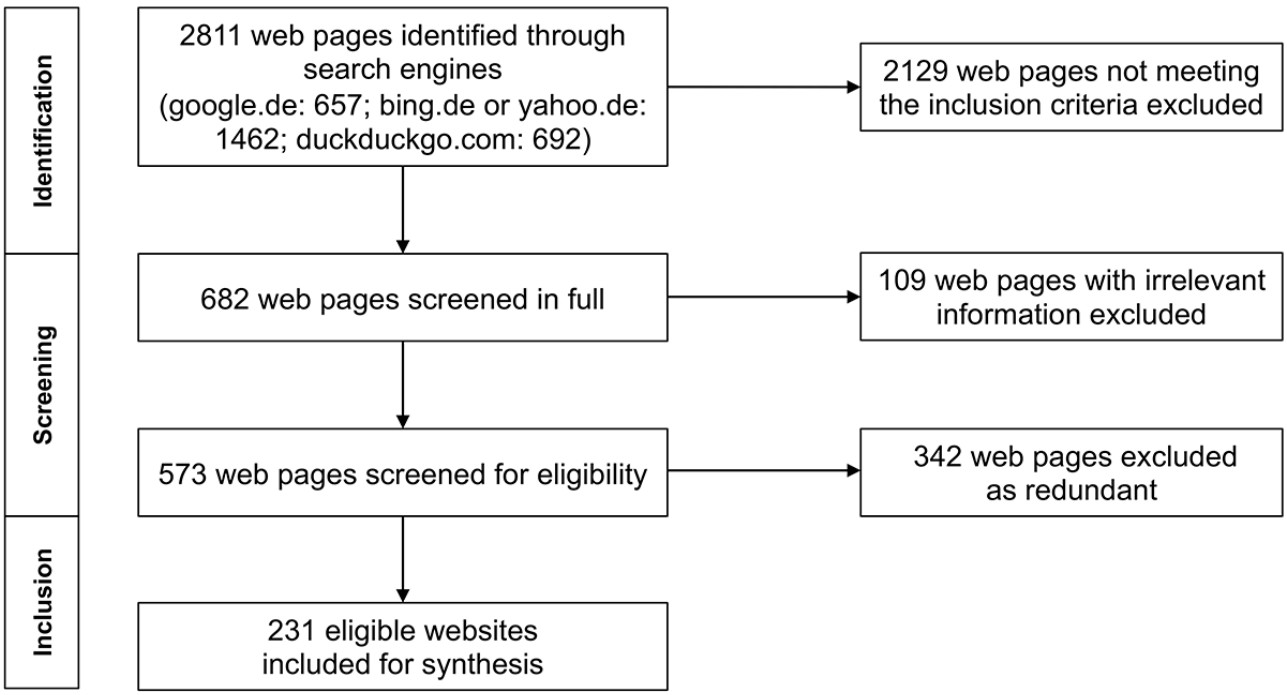

Demographic Data

Most websites and YouTube videos were operated by private dental offices (181 websites and 3 YouTube videos). Characteristics of included websites and YouTube videos are shown in Table 2.

XSL•FO
RenderX

Table 2. Characteristics of included websites and YouTube videos.

Variables and attributes	Websites (n=231), n (%)	YouTube videos (n=7), n (%)
Country		
Germany	202 (87.4)	5 (71.4)
Foreign	29 (12.6)	2 (28.6)
Content provider		
Private dental office	181 (78.4)	3 (42.9)
Corporate dental office or private hospital group	20 (8.7)	2 (28.6)
Public dental clinic or dental school	6 (2.6)	0 (0)
Dental society, dental regulatory body, public body, or insurance company	16 (6.9)	0 (0)
Information service	8 (3.5)	2 (28.6)

Among the websites (n=181) published by private dental offices, most dental offices were located in cities with ≥100,000 inhabitants (n=76, 42%), and multiple dentists worked together (n=109, 60.2%) in a mixed sex setting (n=80, 44.2%). In 61.9% (n=112) of the private dental offices, at least 1 dentist was a member of a dental society. Further details on dentists' demographics are shown in Table 3.

Table 3. Dentists' demographics of included websites and YouTube videos operated by private dental offices.

Variable and attributes	Websites (n=181)	YouTube videos (n=3)
Country, n (%)		
Germany	166 (91.7)	3 (100)
Foreign	15 (8.3)	0 (0)
Practice location, n (%)		
Rural	35 (19.3)	1 (33)
Town (<100,000 inhabitants)	70 (38.7)	2 (67)
City (≥100,000 inhabitants)	76 (42)	0 (0)
Practice setting, n (%)		
Single practitioner	72 (39.8)	1 (33)
Multiple dentists	109 (60.2)	2 (67)
Sex, n (%)		
Female	29 (16)	1 (33)
Male	72 (39.8)	1 (33)
Mixed	80 (44.2)	1 (33)
Dental society membership, n (%)		
No or unknown	69 (38.1)	1 (33)
Yes	112 (61.9)	2 (67)
Dentists' year of examination[a,b], mean (SD)	1997.7 (9.9)	2003.0 (14.6)

[a]In the case of multiple dentists, the years of examination were averaged, if available.

[b]There were 9 missing values (ie, information regarding dentists' years of examination was missing for all dentists of the respective private dental office).

Websites

The median score regarding the technical and functional aspects (domain 1) was 64.3% (IQR 59.8%-69.0%; range 25.3%-82.1%). For this domain, the interrater reliability was 0.953, which indicates an excellent agreement [32]. Further details on the checklist's items and results for each subdomain are shown in Multimedia Appendix 2. The median FRES was 40.0% (IQR 34.0%-49.0%; range 0%-64%). Therefore, the readability of most websites can be classified as "difficult" [38].

The median score regarding the comprehensiveness of information on dental erosion (domain 3) was 11.5% (IQR 3.9%-26.9%; range 0.0%-76.9%). For this domain, the interrater reliability was 0.922, which indicates an excellent agreement

XSL•FO
RenderX

[44]. Further details regarding the checklist's items and results for each subdomain are shown in Multimedia Appendix 3.

The median score regarding the generic quality (domain 4) was 16.7% (IQR 8.3%-23.3%; range 0.0%-53.3%). The interrater reliability of this domain was 0.923, which indicates an excellent agreement [44]. Further details regarding the checklist's items and results for each subdomain are shown in Multimedia Appendix 4.

As a combination of all 4 domains, the overall quality score was 33.6% (IQR 29.8%-39.2%; range 15.3%-58.8%). Individual scores from different domains varied between domains (χ^2_3=517.7; P<.001 for the Friedman test) as shown in Figure 2. Domain 3 and domain 4 resulted in lower scores than domains 1 and 2 (both $P_{adjusted}$<.001).

Figure 2. For all websites, the overall quality score (averaging the results from all 4 domains) and quality scores of each domain (relative percentage of maximum possible score sum) are shown. Domain 1: technical and functional aspects (LIDA instrument, version 1.2; Minervation); domain 2: readability (Flesch reading-ease score, adapted by Amstad for the German language); domain 3: comprehensiveness of information; domain 4: generic quality and risk of bias (DISCERN instrument). Significant differences between domains are marked by different letters (P<.001 for the Friedman test followed by Dunn-Bonferroni post hoc tests). Outliers are marked with an asterisk (*).

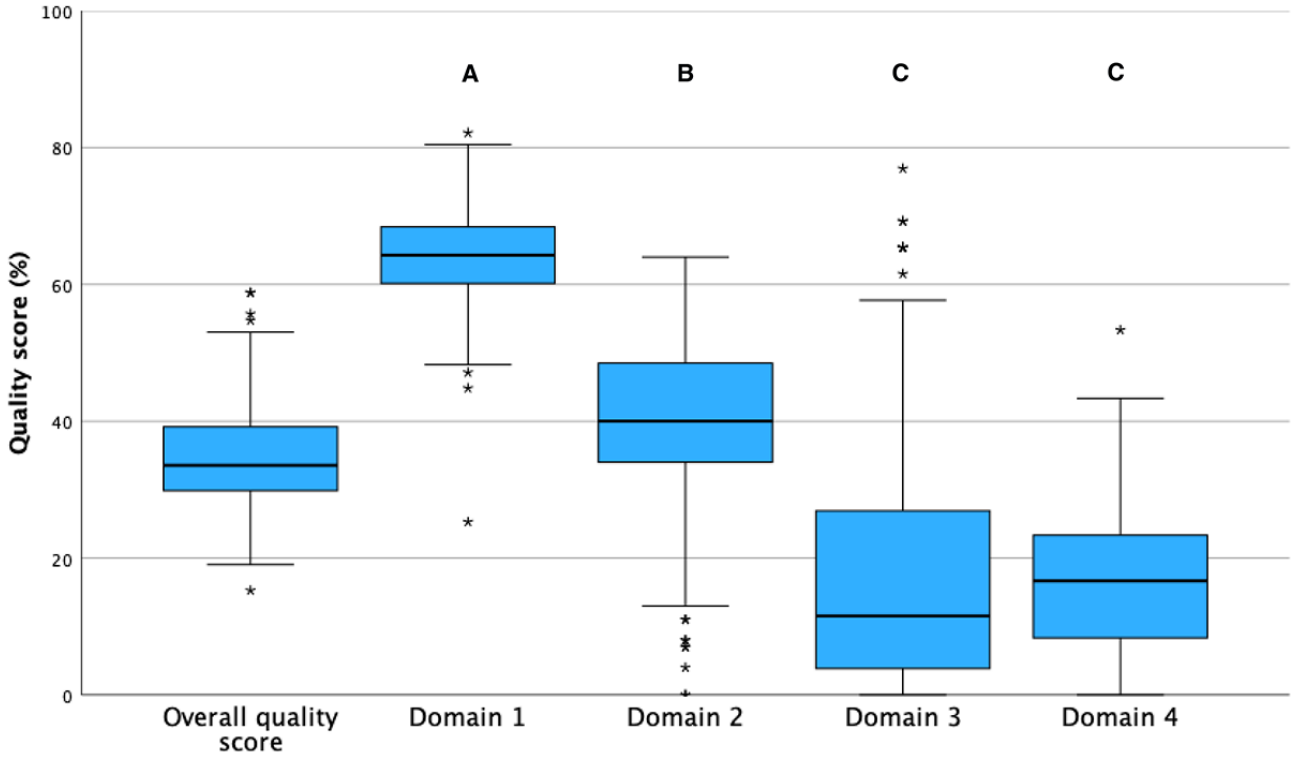

The overall quality score as well as individual scores from domain 1 and domain 3 varied between website providers (P≤.005 for Kruskal-Wallis rank-sum tests) as shown in Table 4. Websites from private dental offices resulted in lower overall quality scores, lower scores in domain 1, and lower scores in domain 3 than websites published by dental societies, dental regulatory bodies, public bodies, or insurance companies ($P_{adjusted}$=.003; $P_{adjusted}$<.001; $P_{adjusted}$=.01, respectively). In addition, for domain 1, websites from information services resulted in lower scores than websites published by dental societies, dental regulatory bodies, public bodies, or insurance companies ($P_{adjusted}$=.003).

XSL•FO
RenderX

Table 4. Quality scores of all domains and overall quality scores across different website providers.

Website provider	Overall quality score		Domain 1[a]		Domain 2[b]		Domain 3[c]		Domain 4[d]	
	Median (IQR)	Range	Median (IQR)	Range	Median (IQR)	Range	Median (IQR)	Range	Median (IQR)	Range
Private dental offices	33.3 (29.4-37.6)	15.3-58.7	63.2 (59.5-67.8)	25.3-80.5	40.0 (34.0-50.0)	0.0-64.0	7.7 (3.9-23.1)	0.0-69.2	16.1 (8.3-23.3)	0.0-53.3
Corporate dental offices or private hospital groups	33.4 (30.7-40.0)	19.6-48.1	66.1 (61.9-69.0)	57.5-73.6	40.0 (33.5-48.0)	0.0-53.0	13.5 (7.7-25.0)	3.9-53.9	12.1 (7.1-22.4)	5.0-41.1
Public dental clinics or dental schools	31.8 (26.0-40.2)	24.2-42.2	63.8 (60.9-69.0)	52.4-73.6	33.0 (24.0-43.0)	21.0-44.0	17.3 (3.9-34.6)	3.9-42.3	14.3 (7.1-21.7)	3.6-23.2
Dental societies, dental regulatory bodies, public bodies, or insurance companies	40.8 (34.2-50.4)	28.9-58.8	69.7 (67.2-75.3)	59.8-82.1	43.0 (40.3-50.5)	31.5-53.0	26.9 (13.5-51.9)	0.0-76.9	19.6 (10.3-36.2)	5.4-41.7
Information services	38.5 (34.3-40.0)	26.7-42.8	60.5 (57.9-64.2)	53.1-69.0	36.5 (31.0-38.0)	27.0-42.0	32.7 (17.3-40.4)	0.0-42.3	26.6 (19.6-30.4)	5.4-31.7
P value[e]	.005	__[f]	<.001	—	.09	—	.002	—	.11	—

[a]Technical and functional aspects (LIDA instrument, version 1.2; Minervation [37]).

[b]Readability (Flesch reading-ease score [38], adapted by Amstad [39] for the German language).

[c]Comprehensiveness of information.

[d]Generic quality and risk of bias (DISCERN instrument [40]).

[e]*P* values from Kruskal-Wallis rank-sum tests.

[f]Not applicable.

Regarding the relation between practice-specific variables of private dental offices and the overall quality score as well as individual scores from each domain, only domain 2 was associated with practice-specific variables as shown in Table 5. German-language websites from non-German private dental offices (*P*=.04 for GLM; B=–6.64, 95% CI –12.85 to –0.42) or from dentists who are a dental society member (*P*=.049 for GLM; B=–3.55, 95% CI –7.09 to –0.01) resulted in lower scores (ie, were more difficult to read), while a shorter time since examination resulted in higher scores (*P*=.01 for GLM; B=0.24 per year, 95% CI 0.05-0.43).

Table 5. Association between practice-specific variables and readability (Flesch reading-ease score) of websites provided by private dental offices. Model fit: χ^2_8=19.6; *P*=.01 (Likelihood ratio test).

Variable	Estimate B (95% CI)	*P* value
Foreign country (reference Germany)	–6.64 (–12.85 to –0.42)	.04
Practice location		
Town (reference rural)	2.92 (–1.73 to 7.58)	.22
City (reference rural)	2.18 (–2.39 to 6.74)	.35
Multiple dentists (reference single practitioner)	–0.67 (–5.81 to 4.47)	.80
Sex		
Male only (reference female only)	–2.51 (–7.63 to 2.61)	.34
Mixed (reference female only)	–1.14 (–7.29 to 5.00)	.72
Dental society membership (reference no or unknown)	–3.55 (–7.09 to –0.01)	.049
Dentists' year of examination[a]	0.24 (0.05 to 0.43)	.01

[a]In the case of multiple dentists, the years of examination were averaged, if available.

YouTube Videos

The median length of YouTube videos was 5.4 (IQR 4.0-7.1) minutes, and the median age was 2.8 (IQR 1.3-3.5) years. However, viewers' interaction and viewing rates of YouTube videos were low. Further characteristics of included YouTube videos are shown in Table 6.

XSL•FO
RenderX

Table 6. Characteristics of included YouTube videos.

Variable	Value	
	Median (IQR)	Range
Age of video (years)	2.8 (1.3-3.5)	0.6-6.1
Duration (minutes)	5.4 (4.0-7.1)	0.5-10.8
Number of views	591 (523.5-1372.5)	105-2711
Number of likes	12 (6.5-18.5)	0-33
Number of dislikes	0 (0-0)	0-1
Number of comments	0 (0-1)	0-30
Comprehensiveness of information (%)	34.6 (13.5-38.5)	3.8-50.0
Viewers' interaction (%)	1.71 (0.70-3.36)	0.00-4.04
Viewing rate (%)	101.0 (54.6-112.6)	23.5-525.0

Among YouTube videos, the median score regarding the comprehensiveness of information on dental erosion was 34.6% (IQR 13.5%-38.5%; range 3.8%-50.0%). Interrater reliability was 0.852, which indicates an excellent agreement [44]. However, the comprehensiveness of information did not vary between websites and YouTube videos (P=.09 for the Wilcoxon rank-sum test).

Discussion

Principal Findings

A total of 231 German-language websites and 7 YouTube videos containing patient-centered, web-based health information on erosive tooth wear were identified and included in the analysis. However, based on the total number of dentists in the DACH (Deutschland [Germany], Austria, and Confoederatio Helvetica [Switzerland]) countries, only a minority of dentists provide web-based health information on erosive tooth wear to (their) patients.

While most websites achieved sufficient scores regarding technical and functional aspects (median score 64.3%, IQR 59.8%-69.0%), the provided text presented a high reading difficulty (median score 40.0%, IQR 34.0%-49.0%). In addition, the comprehensiveness of information on dental erosion was low (median score 11.5%, IQR 3.9%-26.9%), and the generic quality was low resulting in a high risk of bias (median score 16.7%, IQR 8.3%-23.3%). For the calculation of an overall quality score (median score 33.6%, IQR 29.8%-39.2%), all 4 domains were assumed to be of equal importance, and individual domains' quality scores were averaged. The overall quality score as well as individual scores from domain 1 and domain 3 varied between websites' content providers. Thus, the first null hypothesis must be rejected.

Practice-specific variables of private dental offices (eg, years of examination and dental society membership) impacted information quality. As erosive tooth wear gained increased attention over the past decades, only dentists who graduated in recent years (ie, 10 to 15 years) can be expected to be educated about erosive tooth wear during their undergraduate dental curriculum. In addition, dentists who graduated recently are younger and possibly show a greater affinity for the web than older dentists. Dentists who are a dental society member can be expected to be more involved in continuing education and better informed about new developments. In addition, dental societies might spend more money on creating professional patient-centered information than individual dentists. The resulting information is likely of higher quality than self-made texts from individual dentists, and dental societies might provide the created information to their members to be shared with their patients (eg, on dental offices' websites). As a result, dentists who graduated in recent years or dentists who are dental society members are more likely to provide information on erosive tooth wear to their patients and that information might be of higher quality. Regarding the effect of practice-specific variables of private dental offices on the overall quality score as well as individual scores, only scores from domain 2 were associated with practice-specific variables. Thus, the second null hypothesis must be rejected.

The number of available YouTube videos was limited. Among the assessed videos, comprehensiveness of information on dental erosion was also low (median score 34.6%; IQR 13.5%-38.5%). However, content providers of YouTube videos were limited to private dental offices, corporate dental offices, or private hospital groups. No YouTube videos published by public dental clinics, dental schools, dental societies, dental regulatory bodies, public bodies, or insurance companies were found.

Comparison to Prior Work

Previous studies concerning the quality of web-based health information within the context of preventive and restorative dentistry either assessed websites [6,12,13,17,19-24] or YouTube videos [7-11,14-16,18] only. This study evaluated both formats of web-based health information on the same topic (ie, erosive tooth wear) and included a higher number of websites than the majority of the previously performed studies [6,12,13,17,19,20,22-24]. Furthermore, eligible sources were extended beyond websites published by dental offices. While patients in need of dental care are likely to primarily focus on dentists' websites, patients might also find information on websites operated by public dental clinics or dental schools, dental societies, dental regulatory bodies, public bodies, insurance companies, or information services helpful to gain

XSL•FO
RenderX

information and to participate in informed shared decision-making.

While some of the previous studies used different tools (ie, JAMA benchmarks, the presence of HONcode certification, and Global Quality Score), health information on endodontics, dental caries, restoration repair, and periodontitis resulted in overall LIDA scores between 55% and 72% [6,13,19,23], overall DISCERN scores between 4% and 61% [6,9-13,15,17,19,20,22,23], and overall FRES between 60% and 97% [12,13,17,20]. However, this study used the validated LIDA [37] and DISCERN [40] instruments as originally published instead of adjusting the assessment and scoring as it was done in previous studies within the context of preventive and restorative dentistry [9-11,15,19,23].

Some of the previous studies only assessed formal quality or readability [13,17,20-22]. In this study, the comprehensiveness of content-specific information regarding erosive tooth wear was also evaluated. Technical and functional aspects resulted in significantly higher scores than the comprehensiveness of information and generic quality. These findings are in line with previous studies, which also found lower scores regarding the comprehensiveness of the respective disease-specific information than scores regarding formal quality criteria [6,15,19,23]. However, a direct comparison between the scores on the comprehensiveness of information is not possible due to different checklists or qualitative analyses being applied in some studies.

Practical Implications

Websites' information quality on tooth wear was found to be limited. In addition, some information was scientifically inaccurate or misleading, and the available information was neither comprehensive nor trustworthy. As a result, dentists' web-based information on erosive tooth wear requires improvement to inform patients comprehensively and reliably. Until then, patients should be aware of the limitations or gain information elsewhere.

This study found patient-centered, web-based health information on erosive tooth wear provided by dental societies, dental regulatory bodies, public bodies, or insurance companies as well as information services to be more comprehensive than information provided by dental offices. The quality of dentists' web-based health information on erosive tooth wear might be increased by different interventions. First, higher information standards might be enforced by regulatory and legislative bodies. Second, dental societies, dental regulatory bodies, or public bodies might provide validated and high-quality content to dentists, which might be adopted by the individual dentists on their websites. Third, dentists might add links to external websites containing validated and high-quality content published elsewhere (eg, by dental societies, dental regulatory bodies, or public bodies).

Strengths and Limitations

This study has several strengths increasing the overall reliability of our findings. First, websites were identified through a systematic search using 3 different consumer search engines. The selection of the used search engines (google.de, bing.de or yahoo.de, and duckduckgo.com) was based on the current usage data among German internet users at the time of this study [45]. Second, a high number of websites in total and websites operated by different service providers (eg, dental offices, dental societies, dental regulatory bodies, public bodies, insurance companies, and information services) were included. Third, 3 out of 4 domains were scored using established and validated tools [37,38,40]. Fourth, all domains that required manual scoring were independently assessed by 2 different raters. Interrater reliability was found to be excellent (≥0.852) throughout all domains [44]. Fifth, technical and functional aspects were tested using a variety of browsers (Google Chrome, Firefox, Microsoft Edge, and Apple Safari). Again, the selection of the browsers was based on their current usage frequency among German internet users at the time of this study [46].

However, several limitations are present. First, only websites and YouTube videos published in German were assessed. Websites in other languages might show a higher quality or divergent results. However, previous studies regarding the quality of websites within the context of periodontitis found a low quality of information for both German- and English-language websites [21,23]. Second, most websites were published by service providers located in the DACH countries. Therefore, the transferability and generalizability of the findings to other countries or health care systems are unknown. Third, the comprehensiveness of information on dental erosion was assessed using a checklist developed by the authors. As this checklist was not formally tested, a different checklist might have yielded different results. Fourth, only the comprehensiveness of information regarding dental erosion but not its scientific accuracy was evaluated. For example, numerous websites recommended a waiting time between the consumption of erosive food or drinks and toothbrushing, but current evidence suggests that delayed toothbrushing is not capable of preventing erosive tooth wear [47]. Fifth, while the accessibility and usability of websites were assessed using the LIDA instrument, no dedicated tools focusing on accessibility to people with disabilities (eg, Web Content Accessibility Guidelines 2.0) were used.

Future Directions

More dentists should offer web-based information on erosive tooth wear to (their) patients. In addition, the content should be more comprehensive, trustworthy, and of higher quality than the currently available web-based health information on erosive tooth wear. To increase the number of reliable websites, dental societies, dental regulatory bodies, or public bodies might assist dentists by providing high-quality content for further adaptation by the individual dentist.

Conclusions

The quality of German-language, patient-centered, web-based information on erosive tooth wear was found to be limited. The available information was neither comprehensive nor trustworthy. However, quality differed between content providers, and practice-specific variables impacted the readability of websites from private dental offices. Web-based information on erosive tooth wear requires improvement to

XSL•FO
RenderX

inform patients comprehensively and reliably and to allow for patients' participation in informed shared decision-making.

Acknowledgments

The authors acknowledge support from the Open Access Publication Funds of Göttingen University. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Generative artificial intelligence was not used in any portion of this manuscript.

Data Availability

All data generated during or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

AW and PK contributed to the study's conception and design. LH and PK performed the systematic search and data extraction and drafted the manuscript. LH, AFK, and PK evaluated the different domains. PK performed statistical analyses. All authors interpreted the data, critically revised the manuscript, and approved the final version of the manuscript.

Conflicts of Interest

PK is Associate Editor of JMIR Medical Education at the time of this publication. Other authors have no competing interests to declare.

Multimedia Appendix 1
List of included websites and YouTube videos with information on erosive tooth wear and dental erosion.
[DOC File , 787 KB - jmir_v26i1e49514_app1.doc]

Multimedia Appendix 2
Results for each subdomain regarding technical and functional aspects (domain 1: LIDA instrument, version 1.2; Minervation).
[DOC File , 61 KB - jmir_v26i1e49514_app2.doc]

Multimedia Appendix 3
Results for each subdomain regarding the comprehensiveness of information on dental erosion (domain 3).
[DOC File , 41 KB - jmir_v26i1e49514_app3.doc]

Multimedia Appendix 4
Results for each subdomain regarding generic quality and risk of bias (domain 4: DISCERN instrument).
[DOC File , 46 KB - jmir_v26i1e49514_app4.doc]

References

1. Rainie L, Fox S. The online health care revolution. Pew Research Center. 2000. URL: https://www.pewresearch.org/internet/2000/11/26/the-online-health-care-revolution/ [accessed 2022-02-18]
2. Luo A, Qin L, Yuan Y, Yang Z, Liu F, Huang P, et al. The effect of online health information seeking on physician-patient relationships: systematic review. J Med Internet Res 2022;24(2):e23354 [FREE Full text] [doi: 10.2196/23354] [Medline: 35142620]
3. Wald HS, Dube CE, Anthony DC. Untangling the web—the impact of internet use on health care and the physician-patient relationship. Patient Educ Couns 2007;68(3):218-224. [doi: 10.1016/j.pec.2007.05.016] [Medline: 17920226]
4. Ní Ríordáin R, McCreary C. Dental patients' use of the internet. Br Dent J 2009;207(12):583-586; 575 [FREE Full text] [doi: 10.1038/sj.bdj.2009.1137] [Medline: 20019727]
5. Verploegen VJN, Schuller AA. Erosive tooth wear: knowledge among young adults and their preferred information sources. Int J Dent Hyg 2019;17(1):85-92. [doi: 10.1111/idh.12367] [Medline: 30216663]
6. Rossi-Fedele G, Musu D, Cotti E, Doğramacı EJ. Root canal treatment versus single-tooth implant: a systematic review of internet content. J Endod 2016;42(6):846-853. [doi: 10.1016/j.joen.2016.02.011] [Medline: 27080114]
7. Nason K, Donnelly A, Duncan HF. YouTube as a patient-information source for root canal treatment. Int Endod J 2016;49(12):1194-1200. [doi: 10.1111/iej.12575] [Medline: 26551481]
8. Özbay Y, Çırakoğlu NY. YouTube as an information source for instrument separation in root canal treatment. Restor Dent Endod 2021;46(1):e8 [FREE Full text] [doi: 10.5395/rde.2021.46.e8] [Medline: 33680897]

XSL•FO
RenderX

9. McLean S, Cook N, Rovira-Wilde A, Patel S, Kanagasingam S. Evaluating YouTube as a patient information source for the risks of root canal treatment. J Endod 2023;49(2):155-161 [FREE Full text] [doi: 10.1016/j.joen.2022.09.003] [Medline: 36526106]
10. Kodonas K, Fardi A. YouTube as a source of information about pulpotomy and pulp capping: a cross sectional reliability analysis. Restor Dent Endod 2021;46(3):e40 [FREE Full text] [doi: 10.5395/rde.2021.46.e40] [Medline: 34513646]
11. Jung MJ, Seo MS. Assessment of reliability and information quality of YouTube videos about root canal treatment after 2016. BMC Oral Health 2022;22(1):494 [FREE Full text] [doi: 10.1186/s12903-022-02540-4] [Medline: 36384745]
12. Aguirre PEA, Coelho MM, Rios D, Machado MAAM, Cruvinel AFP, Cruvinel T. Evaluating the dental caries-related information on Brazilian websites: qualitative study. J Med Internet Res 2017;19(12):e415 [FREE Full text] [doi: 10.2196/jmir.7681] [Medline: 29237585]
13. Blizniuk A, Furukawa S, Ueno M, Kawaguchi Y. Evaluation of English websites on dental caries by using consumer evaluation tools. Oral Health Prev Dent 2016;14(4):363-369. [doi: 10.3290/j.ohpd.a35746] [Medline: 26973989]
14. ElKarmi R, Hassona Y, Taimeh D, Scully C. YouTube as a source for parents' education on early childhood caries. Int J Paediatr Dent 2017;27(6):437-443. [doi: 10.1111/ipd.12277] [Medline: 27882621]
15. Rachmawati YL, Putri DW, Hariyani N, Bahar A, Maharani DA. Analysis of quality, usefulness, reliability, visibility, and popularity of videos about dental caries on YouTube: a cross-sectional analysis. J Int Soc Prev Community Dent 2022;12(2):245-251 [FREE Full text] [doi: 10.4103/jispcd.JISPCD_246_21] [Medline: 35462746]
16. Long M, Forbes LE, Papagerakis P, Lieffers JRL. YouTube videos on nutrition and dental caries: content analysis. JMIR Infodemiology 2023;3:e40003 [FREE Full text] [doi: 10.2196/40003] [Medline: 37561564]
17. Aboalshamat K. Quality and readability of web-based information on dental caries in Arabic: an infodemiological study. BMC Oral Health 2023;23(1):797 [FREE Full text] [doi: 10.1186/s12903-023-03547-1] [Medline: 37880640]
18. Erturk-Avunduk AT, Delikan E. Evaluation of the quality of YouTube™ videos about pit and fissure sealant applications. Int J Dent Hyg 2023;21(3):590-598. [doi: 10.1111/idh.12646] [Medline: 36480358]
19. Kanzow P, Büttcher AF, Wiegand A, Schwendicke F. Quality of information regarding repair restorations on dentist websites: systematic search and analysis. J Med Internet Res 2020;22(4):e17250 [FREE Full text] [doi: 10.2196/17250] [Medline: 32062595]
20. Al-Ak'hali MS, Fageeh HN, Halboub E, Alhajj MN, Ariffin Z. Quality and readability of web-based Arabic health information on periodontal disease. BMC Med Inform Decis Mak 2021;21(1):41 [FREE Full text] [doi: 10.1186/s12911-021-01413-0] [Medline: 33541345]
21. Bizzi I, Ghezzi P, Paudyal P. Health information quality of websites on periodontology. J Clin Periodontol 2017;44(3):308-314. [doi: 10.1111/jcpe.12668] [Medline: 28005268]
22. Kanmaz B, Buduneli N. Evaluation of information quality on the internet for periodontal disease patients. Oral Dis 2021;27(2):348-356. [doi: 10.1111/odi.13546] [Medline: 32671923]
23. Schwendicke F, Stange J, Stange C, Graetz C. German dentists' websites on periodontitis have low quality of information. BMC Med Inform Decis Mak 2017;17(1):114 [FREE Full text] [doi: 10.1186/s12911-017-0511-8] [Medline: 28768511]
24. Chestnutt IG. The nature and quality of periodontal related patient information on the world-wide web. Br Dent J 2002;193(11):657-659 [FREE Full text] [doi: 10.1038/sj.bdj.4801653] [Medline: 12607624]
25. Lagerweij MD, van Loveren C. Declining caries trends: are we satisfied? Curr Oral Health Rep 2015;2(4):212-217 [FREE Full text] [doi: 10.1007/s40496-015-0064-9] [Medline: 26523247]
26. Schlueter N, Amaechi BT, Bartlett D, Buzalaf MAR, Carvalho TS, Ganss C, et al. Terminology of erosive tooth wear: consensus report of a workshop organized by the ORCA and the Cariology Research Group of the IADR. Caries Res 2020;54(1):2-6 [FREE Full text] [doi: 10.1159/000503308] [Medline: 31610535]
27. Kanzow P, Wegehaupt FJ, Attin T, Wiegand A. Etiology and pathogenesis of dental erosion. Quintessence Int 2016;47(4):275-278. [doi: 10.3290/j.qi.a35625] [Medline: 27022647]
28. Schlueter N, Luka B. Erosive tooth wear—a review on global prevalence and on its prevalence in risk groups. Br Dent J 2018;224(5):364-370 [FREE Full text] [doi: 10.1038/sj.bdj.2018.167] [Medline: 29495027]
29. Jordan AR, Micheelis W. Fünfte Deutsche Mundgesundheitsstudie (DMS V) [The Fifth German Oral Health Study (DMS V)]. Cologne: Deutscher Ärzte-Verlag; 2016.
30. Carvalho TS, Colon P, Ganss C, Huysmans MC, Lussi A, Schlueter N, et al. Consensus report of the European Federation of Conservative Dentistry: erosive tooth wear—diagnosis and management. Clin Oral Investig 2015;19(7):1557-1561 [FREE Full text] [doi: 10.1007/s00784-015-1511-7] [Medline: 26121968]
31. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012;12:181 [FREE Full text] [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Zahnarztsuche in Ihrer Stadt—Zahnmedizinische Patienteninformation [Dentist search in your city—dental patient information]. German Society of Dentistry and Oral Medicine. URL: https://www.zahnmedizinische-patienteninformationen.de/zahnarztsuche [accessed 2021-04-09]

XSL•FO
RenderX

34. Finden Sie Ihren passenden Zahnarzt mit unserer Zahnarztsuche [Find the right dentist for you with our dentist search]. Swiss Dental Association. URL: https://www.sso.ch/zahnarztsuche [accessed 2021-10-12]
35. Wayback Machine. Internet Archive. URL: https://archive.org/web/ [accessed 2021-04-09]
36. health.ch AG. Das Schweizer Ärzteportal, seit 1998 [The Swiss doctors portal, since 1998]. doktor.ch. URL: https://www.doktor.ch [accessed 2021-04-09]
37. The LIDA instrument: minervation validation instrument for health care web sites, version 1.2. Minervation. 2011. URL: http://www.minervation.com/wp-content/uploads/2011/04/Minervation-LIDA-instrument-v1-2.pdf [accessed 2022-02-18]
38. Flesch R. A new readability yardstick. J Appl Psychol 1948;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
39. Amstad T. Wie verständlich sind unsere Zeitungen? [How understandable are our newspapers?]. Zurich, Switzerland: Studenten-Schreib-Service; 1978.
40. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
41. Schöll P. Berechnung—Fleschindex [Calculation—Flesch reading-ease score]. URL: https://fleschindex.de/berechnen/ [accessed 2021-04-09]
42. Hassona Y, Taimeh D, Marahleh A, Scully C. YouTube as a source of information on mouth (oral) cancer. Oral Dis 2016;22(3):202-208. [doi: 10.1111/odi.12434] [Medline: 26718020]
43. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1(1):30-46. [doi: 10.1037/1082-989X.1.1.30]
44. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in Psychology. Psychol Assess 1994;6(4):284-290. [doi: 10.1037/1040-3590.6.4.284]
45. Welche Suchmaschine haben Sie in den letzten 4 Wochen genutzt? [Which search engine did you used in the past 4 weeks?]. Statista Global Consumer Survey. 2021. URL: https://de.statista.com/statistik/daten/studie/436243/umfrage/meistgenutzte-browser-im-internet-in-deutschland/ [accessed 2021-03-24]
46. Markanteile der meistgenutzten Browserversionen in Deutschland im Mai 2021 [Market share of the most frequently used browser versions in Germany in May 2021]. StatCounter. 2021. URL: https://de.statista.com/statistik/daten/studie/436243/umfrage/meistgenutzte-browser-im-internet-in-deutschland/ [accessed 2021-06-16]
47. Hong DW, Lin XJ, Wiegand A, Yu H. Does delayed toothbrushing after the consumption of erosive foodstuffs or beverages decrease erosive tooth wear? A systematic review and meta-analysis. Clin Oral Investig 2020;24(12):4169-4183. [doi: 10.1007/s00784-020-03614-9] [Medline: 33052542]

Abbreviations

DACH: Deutschland (Germany), Austria, and Confoederatio Helvetica (Switzerland)
DGZMK: Deutsche Gesellschaft für Zahn-, Mund- und Kieferheilkunde (German Society of Dentistry and Oral Medicine)
ENTREQ: Enhancing Transparency in Reporting the Synthesis of Qualitative Research
FRES: Flesch reading-ease score
GLM: generalized linear modeling
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SSO: Swiss Dental Association

Edited by A Mavragani; submitted 31.05.23; peer-reviewed by J Jeong, B Luka, J Schmidt, K Scholz; comments to author 26.10.23; revised version received 01.11.23; accepted 03.01.24; published 31.01.24.

Please cite as:
Holland L, Kanzow AF, Wiegand A, Kanzow P
Quality of Patient-Centered eHealth Information on Erosive Tooth Wear: Systematic Search and Evaluation of Websites and YouTube Videos
J Med Internet Res 2024;26:e49514
URL: https://www.jmir.org/2024/1/e49514
doi:10.2196/49514
PMID:38167299

©Lena Holland, Amelie Friederike Kanzow, Annette Wiegand, Philipp Kanzow. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and

XSL•FO
RenderX

reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Retrospecting Digital Media Use, Negative Emotions, and Trust Gaps During the COVID-19 Pandemic in China: Cross-Sectional Web-Based Survey

Lu Wei[1,2], PhD; Qing Huang[2], PhD

[1]School of Journalism and Communication, Communication University of Zhejiang, Hangzhou, China

[2]International Communication Institute, College of Media and International Culture, Zhejiang University, Hangzhou, China

Corresponding Author:
Qing Huang, PhD
International Communication Institute
College of Media and International Culture
Zhejiang University
866 Yuhangtang Road
Zijingang Campus
Hangzhou, 310058
China
Phone: 86 0571 87075102
Email: qing_huang@zju.edu.cn

Abstract

Background: Retrospecting the trust gaps and their dynamics during the pandemic is crucial for understanding the root causes of postpandemic challenges and offers valuable insights into preparing for future public health emergencies. The COVID-19 pandemic eroded people's trust in strangers and acquaintances, while their trust in family members remained relatively stable. This resulted in 2 trust gaps, namely, the family members–strangers trust gap and the family members–acquaintances trust gap. Widening trust gaps impede social integration and undermine the effective management of public health crises. However, little is known about how digital media use shaped trust gaps during a pandemic.

Objective: This study aims to investigate the relationships between digital media use, negative emotions, the family members–strangers trust gap, and the family members–acquaintances trust gap during the COVID-19 pandemic in China. We test the mediating role of negative emotions between digital media use and 2 trust gaps and compare the indirect effect of digital media use on 2 trust gaps through negative emotions.

Methods: A cross-sectional web-based survey was conducted in China between January 31, 2020, and February 9, 2020. A total of 1568 adults participated in the survey. Questions related to digital media use, negative emotions, trust in family members, trust in acquaintances, and trust in strangers during the pandemic were asked. Regression analyses were performed to test the associations between the examined variables. We used a 95% bootstrap CI approach to estimate the mediation effects.

Results: Digital media use was positively associated with negative emotions (B=0.17, SE 0.03; P<.001), which in turn were positively associated with the family members–strangers trust gap (B=0.15, SE 0.03; P<.001). Likewise, digital media use was positively associated with negative emotions (B=0.17, SE 0.03; P<.001), while negative emotions were positively associated with the family members–acquaintances trust gap (B=0.08, SE 0.03; P=.01). Moreover, the indirect effect of digital media use on the family members–strangers trust gap (B=0.03, SE 0.01; 95% CI 0.01-0.04) was stronger than that on the family members–acquaintances trust gap (B=0.01, SE 0.01; 95% CI 0.003-0.027).

Conclusions: The results demonstrate that negative emotions resulting from the frequent use of digital media are a key factor that accounts for the widening trust gaps. Considering the increasing reliance on digital media, the findings indicate that the appropriate use of digital media can prevent the overamplification of negative emotions and curb the enlargement of trust gaps. This may help restore social trust and prepare for future public health crises in the postpandemic era.

(J Med Internet Res 2024;26:e49422) doi:10.2196/49422

XSL•FO
RenderX

KEYWORDS

digital media use; negative emotions; family members–strangers trust gap; family members–acquaintances trust gap; mediation effect; COVID-19

Introduction

Background

During the period from early 2020 to May 5, 2023, marked by the World Health Organization's declaration of the end of the COVID-19 global health emergency [1], the pandemic caused increased societal division and declined social trust across many countries [2]. The erosion of social trust during a crisis tends to engender enduring repercussions, such as individuals' psychological distress, public dissent, and a decrease in community resilience [3,4]. This, in turn, impedes the processes of social cohesion and the maintenance of a robust society. Moreover, against the backdrop of a high-risk environment [5], large-scale infectious diseases might break out in the future. Therefore, retrospecting social trust dynamics and examining the influence of various factors on trust-related outcomes during the pandemic have dual implications: (1) to understand the root causes of postpandemic trust-related issues; and (2) to prepare for future public health emergencies, given that trust serves as the bedrock of effective crisis management [6,7].

We used the individual-level approach to understand social trust [8-10] and define it as an individual's trust in others, including family members, acquaintances (eg, coworkers, classmates, and casual friends), and strangers. Noticeably, the differential mode of association (*chaxugeju*) in Chinese culture indicates that as social distance between individuals increases, trust decreases [11,12]. Accordingly, trust is highest and most stable in family members, followed by acquaintances, while trust in strangers is relatively low and more susceptible to change. This phenomenon suggests 2 trust gaps: the family members–strangers trust gap and the family members–acquaintances trust gap. Moreover, evidence suggests that these gaps existed during the COVID-19 pandemic [13,14]. Given that increasing trust gaps are linked to social estrangement and other challenges [15], it is essential to examine how the trust gaps widened during the pandemic.

Information that circulates on digital media platforms is a major source that influences trust gaps. At the early stage of the pandemic, information on digital media involved a considerable amount of uncivil comments about general others [16-18]. Thus, an individual's exposure to such information might erode his or her trust in strangers and in acquaintances, whereas their trust in family members was less likely to be affected. This enlarged the family members–strangers trust gap and the family members–acquaintances trust gap. Besides, an individual's exposure to information on digital media that highlighted the susceptibility and severity of the coronavirus and the social consequences of the pandemic provoked his or her negative emotions, such as fear, anxiety, sadness, anger, and hostility [19,20], which in turn weakened their trust in others [21,22], including trust in strangers and acquaintances. Therefore, digital media use is a major informational source that influences the 2 trust gaps, while negative emotions are a psychological response that may increase these gaps.

This study aims to unravel the relationship between digital media use, negative emotions, the family members–strangers trust gap, and the family members–acquaintances trust gap during the pandemic. Despite the extensive body of studies that examined media use and its associated outcomes during the COVID-19 pandemic [23-25], this study is, to our knowledge, the first to propose the concept of *trust gap* and explore the mechanisms that contribute to the increasing gaps. Moreover, our study is an empirically grounded theoretical exploration of the trust gaps and its associated dynamics, rather than policy research. Therefore, although the pandemic has ended, our findings and conclusions may advance the understanding of the effect of media use on trust-related challenges in other health and risk situations beyond the COVID-19 pandemic.

Negative Emotions Mediate Between Digital Media Use and 2 Trust Gaps

Digital media refer to a wide range of computer-mediated and mobile technology–based applications. These encompass websites, search engines, web-based forums, wikis, blogs, social networking sites, news feeds, video-sharing sites, and so forth [26,27]. People use digital media to satisfy certain needs, such as cognition, entertainment, interactivity, and agency [28-30]. Among various purposes of digital media use, information seeking serves as the basic motive to meet people's cognitive need [31-33]. In particular, acquiring mediated information and staying informed of the latest updates is of great significance during a public health emergency [34,35]. Drawing upon previous conceptions of media use for information seeking [36,37], we define digital media use as the frequency that people use different digital media tools to acquire information related to the coronavirus. This information mainly involves the health risks of the virus and the socioeconomic consequences of the pandemic.

Despite the existence of positive emotions, people tend to feel negative emotions more commonly when a contagious disease breaks out. For instance, people experienced worry, anxiety, fear, distress, anger, and sadness at the early stage of the Middle East respiratory syndrome (MERS), and H1N1 influenza and H7N9 influenza [35,38,39]. Given that the coronavirus is characterized by high infection rate and severe consequences, people tended to feel fear, anxiety, and sadness during early phase of the pandemic. Moreover, the unsettled disputes over the origin of the pandemic and the widespread use of controversial, xenophobia, and stigmatized terms associated with the coronavirus incited people's anger and hostility [16-18]. Thus, we define negative emotions as people's undesirable feelings toward the coronavirus outbreak and its consequences, such as fear, anxiety, sadness, anger, and hostility.

A considerable amount of literature has discussed the association between people's digital media use and their negative emotions

XSL•FO
RenderX

amid public health emergencies. For example, studies demonstrated that the more frequently individuals exposed themselves to social media information, the stronger their negative emotions were, either in the MERS outbreak in South Korea [38] or during the H7N9 influenza in China [39]. Moreover, research demonstrated that Chinese netizens' frequent digital media use significantly amplified their negative emotions at the early stage of the COVID-19 outbreak, especially their anxiety, depression, and stress [19,20]. When the coronavirus first broke out in China, information about increasing number of confirmed cases, threat to public health, disruption to daily routines, and economic downturn circulated widely across a variety of digital media [40]. Through exposure to information about the coronavirus's unpredictable health threat, people felt anxiety, fear, and sadness. Similarly, consuming information that emphasized the socioeconomic consequences of the pandemic was likely to trigger people's anger and hostility toward those who were attributed with the responsibility of accelerating the spread of the virus. Accordingly, the more frequently people used digital media to acquire information, the stronger their negative emotions were.

When individuals are exposed to a crisis, their negative emotions motivate them to attribute the causality to certain actors [22]. In circumstances in which the cause of an event remains unclear, an individual's attribution-dependent negative emotions are more likely to encourage him or her to attribute the responsibility to socially distant ones [41]. Given that the disputes over the origin and cause of the coronavirus have remained unsettled so far [42], the pandemic can be seen as a health crisis without an explicit or agreed-upon cause. Thus, people's negative emotions about the pandemic tend to motivate them to attribute part of the responsibility to socially distant others. According to the attribution theory, people who are attributed to responsibility are considered distrustful [21]. Moreover, evidence showed that negative emotions provoked by contagious diseases led people to be distrustful of out-group members [43-45]. Because strangers and acquaintances were regarded as socially distant others or out-group members during the pandemic, the negative emotions tended to decrease one's trust in them, thereby enlarging the family members–strangers trust gap and the family members–acquaintances trust gap.

Taken together, the literature reviewed in this section indicates the mediating role of negative emotions between digital media use and the 2 trust gaps. Thus, we posit the following hypotheses:

Hypothesis 1: negative emotions mediate the association between digital media use and the family members–strangers trust gap.

Hypothesis 2: negative emotions mediate the association between digital media use and the family members–acquaintances trust gap.

Moreover, because strangers were viewed as more socially distant others than acquaintances [12], it was likely that individuals attributed more responsibility to strangers than to acquaintances when experiencing negative emotions elicited by digital media use, thus leading to varying degrees of increase in the 2 trust gaps. Hence, we put forward the following hypothesis:

Hypothesis 3: the mediation effect of negative emotions between digital media use and the family members–strangers trust gap is stronger than the mediation effect of negative emotions between digital media use and the family members–acquaintances trust gap. Figure 1 depicts the conceptual model.

Figure 1. The conceptual model.

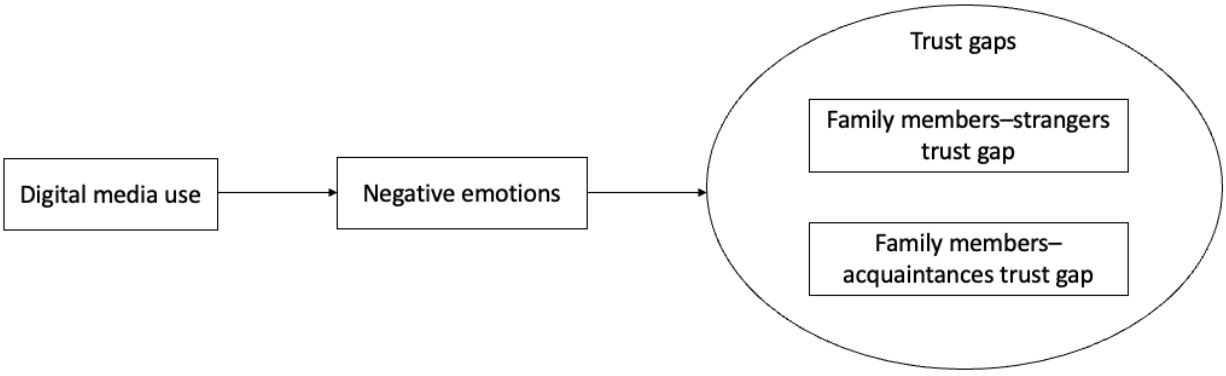

Methods

Participants and Procedures

A cross-sectional web-based survey was conducted from January 31, 2020, to February 9, 2020, an early phase of the COVID-19 outbreak in China. We chose this time span for several reasons. First, a considerable amount of information related to COVID-19 was circulating on a variety of digital media platforms during this period of time, which guaranteed that respondents had adequate informational resources to consume. Second, given the important role of negative emotions in the conceptual model, we attempted to capture people's negative emotions at an early stage to avoid the emotional burnout that might occur at a later stage of the pandemic. Third, public understanding of the coronavirus was very limited at the very beginning of the outbreak. Insufficient knowledge led the public to experience a high degree of uncertainty and a strong sense of insecurity, which potentially impacted their social trust [46,47]. This warranted the investigation into the relationship between digital media use and trust gaps at the very early phase compared with other phases of the health crisis.

XSL•FO
RenderX

We commissioned Sojump, a professional web-based survey company in China, to collect survey data. The company provides a sampling service of 6.2 million registered respondents distributed throughout China. We used a random sampling strategy within the survey pool. Given the emergency nature of the COVID-19 outbreak, this sampling strategy was an efficient and timely manner to explore initial public reactions and attitudes toward the pandemic. Moreover, previous studies have used this sampling strategy to examine major health risk issues in China [48-50].

In terms of the sampling procedure, Sojump first randomly selected 2840 users from the 6.2 million-user survey pool and asked them to participate in the web-based survey through an email invitation. A total of 1656 respondents finished the questionnaire, with a response rate of 58.3% (1656/2840). After deleting the cases with missing values or those that did not pass the attention checks, we obtained a sample of 1568 valid cases for data analysis.

The sample covered 31 provinces, municipalities, and autonomous regions across the Chinese mainland. Because we focused on people's digital media use and the associated effects, Chinese netizens were regarded as the population [51]. Our sample was representative of the population on gender (779/1568, 49.7% were females) and fairly representative on age (739/1568, 47.1% were younger than 30 years) and income (579/1568, 37.0% with an income of 3000-8000 Chinese Yuan (CNY; US $412.60-$1100.28). However, because college students and graduates constituted a major part of the survey pool, we oversampled those with greater education attainment (1237/1568, 78.9% were college students or graduates or above). Table 1 compares the demographic features of our sample with those of the population.

Table 1. Sample demographics compared with the demographic features of Chinese netizens[a].

Variables	Sample, n (%)	Proportion of the population, %
Gender (female)	779 (49.7)	48.1
Age (years)		
<30	739 (47.1)	44.7
30-49	754 (48.1)	38.4
>50	75 (4.8)	16.9
Income (CNY[b])		
≤1000	128 (8.2)	27.9
1001-3000	644 (41.0)	23.1
3,001-5000	332 (21.2)	21.5
5001-8000	247 (15.8)	14.3
>8000	217 (13.9)	13.3
Education		
Middle school or below	54 (3.4)	58.3
High school/technical school	277 (17.7)	22.2
College or above	1237 (78.9)	19.5

[a]The demographic features of Chinese netizens were retrieved from the 45th Statistical Report on China's Internet Development issued in April 2020.
[b]CNY: Chinese Yuan; 1 CNY=US $0.14.

Ethical Considerations

The institutional review board of Zhejiang University approved the data collection protocol (2020-056). Voluntary informed consent was obtained from the participants before the web-based survey. Each participant was compensated with CNY 12 (US $1.6) for participating in the study. The final data set was anonymized. We ensured that no identifiable private information was linked to the participants.

Measures

Trust Gaps

Trust gaps included the family members–strangers trust gap and the family members–acquaintances trust gap. First, we separately measured respondents' trust in their family members, acquaintances, and strangers. Drawing upon the extant measurement [52], trust in family members (mean 4.53, SD 0.71), trust in acquaintances (mean 3.85, SD 0.75), and trust in strangers (mean 2.35, SD 0.81) were measured with a single item asking respondents the degree to which they trusted their family members, acquaintances, and strangers during the COVID-19 outbreak, respectively. A 5-point Likert scale that ranged from 1 (totally distrust) to 5 (totally trust) was used. Then, the family members–strangers trust gap was measured by subtracting trust in strangers from trust in family members (mean 2.18, SD 1.02), and the family members–acquaintances trust gap was measured by subtracting trust in acquaintances from trust in family members (mean 0.68, SD 0.93).

XSL•FO
RenderX

Digital Media Use

Drawing upon the measurement of digital media use for informational purpose [37], we measured digital media use as the frequency of people's exposure to COVID-19 information through a wide range of digital media tools. Based on people's habits of digital media use during the pandemic [40], we asked respondents how frequently they exposed themselves to information about COVID-19 via the nine digital media tools on a 5-point Likert scale ranging from 1 (never) to 5 (very frequently): (1) portal websites, (2) social networking sites, (3) news apps, (4) video apps, (5) web-based question-and-answer communities, (6) search engines, (7) government news apps, (8) medical apps, and (9) netizens-run digital media. Nine items were averaged to form a composite index of digital media use (mean 3.35, SD 0.63; Cronbach α=0.70).

Negative Emotions

According to the commonly experienced negative emotions during public health emergencies [38,53], this study measured negative emotions by asking respondents to indicate the degree to which they felt, on a 5-point Likert scale ranging from 1 (not at all) to 5 (very strongly), the following five emotions toward the COVID-19 outbreak: (1) fear, (2) anxiety, (3) sadness, (4) anger, and (5) hostility. The 5 items were averaged to create an additive index (mean 2.43, SD 0.81; Cronbach α=0.79).

Control Variables

Respondents' demographic characteristics, such as age, gender, income, and education, were included as control variables. In addition, because risk perception, knowledge about COVID-19, and institutional trust tended to influence people's negative emotions [54,55] and thereby affected trust gaps, they were also entered as covariates. Risk perception involves the judgment of the likelihood of being infected by the coronavirus and the assessment of the severity of an infection [56]. Knowledge is defined in two ways: (1) objective knowledge, which reflects the actual amount of information an individual has about a topic, and (2) subjective knowledge, which shows an individual's self-evaluation of the acquired knowledge [57]. Given that respondents might overestimate their subjective knowledge of the coronavirus [58], this study focused on one's objective knowledge about COVID-19. Institutional trust refers to people's positive judgment of the trustworthiness of institutionalized organizations [59]. In public health emergencies, key institutional actors include government agencies, medical systems, and scientists. Thus, institutional trust in this study consists of people's trust in governments, medical systems, and scientists.

Consistent with the established measurement of risk perception [53], risk perception of COVID-19 was calculated by multiplying perceived susceptibility by perceived severity (mean 7.22, SD 4.11). These two items were measured on a 5-point Likert scale: (1) perceived susceptibility: the likelihood of getting infected by COVID-19 between now and the near future (1=very unlikely, 5=very likely), and (2) perceived severity: the threat of COVID-19 infection to one's health and life (1=not at all serious, 5=very serious). Referring to the measurement of MERS knowledge [38], knowledge about COVID-19 was measured by 5 quiz-type questions, such as "COVID-19 patients all have symptoms such as fever and cough" (wrong), and "patients who have underlying health conditions are more likely to suffer from severe COVID-19" (right). Correct answers were coded as 1, and incorrect answers were coded as 0 for each question. The answers to 5 questions were added to create the COVID-19 knowledge index (mean 4.57, SD 0.63).

Five-point Likert scales ranging from 1 (strongly disagree) to 5 (strongly agree) were used to measure 3 types of institutional trust. Based on the previous measurement [59], trust in government was measured by 6 items regarding the central and local governments, such as "What the central government has done so far in containing the coronavirus is trustworthy," "I trust the local government's capability to cope with the coronavirus." Six items were averaged to create a composite index (mean 3.98, SD 0.66; Cronbach α=0.85). Five items were averaged to develop the index of trust in medical systems (mean 3.51, SD 0.75; Cronbach α=0.76). Example items included medical professionals "have high rates of correct diagnosis," and "provide effective health care for patients" [60]. Trust in scientists was measured by respondents' evaluation of scientists in terms of their trustworthiness, competence, benevolence, and general credibility [61] (mean 4.35, SD 0.56; Cronbach α=0.80).

Statistical Analyses

We used PROCESS, an SPSS macro developed by Hayes [62], to test the hypotheses. One of the advantages of PROCESS is that it implements the recommended asymptotic and bootstrapping method to test hypotheses involving mediation effects that have few requirements for model assumption [63]. The research hypotheses constitute 2 mediation effects, in which negative emotions mediate the association between digital media use and the 2 trust gaps, respectively. Consequently, we chose model 4 in the PROCESS templates.

We separately ran 2 mediation tests. In the first mediation test, family members–strangers trust gap was entered as the dependent variable, digital media use as the independent variable, and negative emotions as the mediator variable. Age, gender, income, education, risk perception, COVID-19 knowledge, trust in government, trust in medical systems, and trust in scientists were entered as covariates. In the second mediation test, the family members–acquaintances trust gap was entered as the dependent variable, while other variables remained the same as the first test.

Results

Preliminary Analysis

Digital media use, negative emotions, and the control variables explained 2.5% of the variance in the family members–strangers trust gap ($F_{11,1556}$=3.67; P<.001; R^2=0.025) and 4.6% of the variance in the family members–acquaintances trust gap ($F_{11,1556}$=6.86; P<.001; R^2=0.046). Among the control variables, trust in scientists was positively associated with the family members–strangers trust gap (B=0.15, SE 0.05; P=.004), indicating that a higher level of trust in scientists during the pandemic tended to increase the gap between one's trust in

XSL•FO
RenderX

family members and in strangers. At the early stage of the outbreak, scientists advised people to practice social distancing to prevent the rapid spread of the coronavirus. By believing in scientists' advice and frequently practicing social distancing, people's trust in strangers decreased and thereby enlarged the family members–strangers trust gap. Besides, age (*B*=0.02, SE 0.003; *P*<.001), education (*B*=–0.07, SE 0.03; *P*=.01), and COVID-19 knowledge (*B*=–0.08, SE 0.04; *P*=.02) were significantly associated with the family members–acquaintances trust gap. In other words, older people tended to be more skeptical about acquaintances than their younger counterparts did. Besides, people who had greater education attainment and more COVID-19 knowledge might use a more rational approach to view the coronavirus than those with less education attainment or COVID-19 knowledge did, which lowered the former group's distrust or skepticism in acquaintances.

Testing Mediation Effects

After controlling for the effects of the control variables, digital media use was positively associated with negative emotions (*B*=0.17, SE 0.03; *P*<.001), and negative emotions were positively associated with the family members–strangers trust gap (*B*=0.15, SE 0.03; *P*<.001). According to the results of the bootstrap method at 95% CI, the indirect effect of digital media use on the family members–strangers trust gap via negative emotions was significant (*B*=0.03, SE 0.01; 95% CI 0.01-0.04), which supported hypothesis 1. However, the direct effect of digital media use on the family members–strangers trust gap was not significant (*B*=–0.03, SE 0.04, 95% CI –0.11 to 0.05).

In terms of the second hypothesis, digital media use was positively associated with negative emotions (*B*=0.17, SE 0.03; *P*<.001), while negative emotions were positively associated with the family members–acquaintances trust gap (*B*=0.08, SE 0.03; *P*=.01). The indirect effect of digital media use on the family members–acquaintances trust gap through negative emotions was significant (*B*=0.01, SE 0.01; 95% CI 0.003-0.027). Thus, hypothesis 2 was supported. Nevertheless, we did not observe a significant direct effect of digital media use on the family members–acquaintances trust gap (*B*=–0.04, SE 0.04; 95% CI –0.12 to 0.03).

These results revealed that negative emotions fully mediated the association between digital media use and the 2 trust gaps. Moreover, the mediation tests also demonstrated that the mediation effect of negative emotions on the family members–strangers trust gap was stronger than that on the family members–acquaintances trust gap, showing support for hypothesis 3. Tables 2 and 3 and Figure 2 demonstrate the statistical results and the final model, respectively.

Table 2. Regressing family members–strangers trust gap on antecedents[a].

	Negative emotions		Family members–strangers trust gap	
	B (SE[b])	*P* values	*B* (SE)	*P* values
Control variables				
Age (years)	–0.003 (0.002)	–.07	0.004 (0.003)	.22
Gender[c]	–0.14 (0.04)	<.001	0.02 (0.05)	.68
Income	–0.003 (0.01)	.76	–0.004 (0.01)	.80
Education	0.06 (0.02)	.009	–0.05 (0.03)	.09
Risk perception	0.04 (0.005)	<.001	–0.01 (0.01)	.32
COVID-19 knowledge	0.06 (0.03)	.05	–0.03 (0.04)	.40
Trust in government	–0.20 (0.03)	<.001	0.07 (0.05)	.11
Trust in medical systems	–0.18 (0.03)	<.001	0.02 (0.04)	.67
Trust in scientists	0.04 (0.04)	.29	0.15 (0.05)	.003
Antecedents				
Digital media use	0.17 (0.03)	<.001	–0.03 (0.04)	.51
Negative emotions	__[d,e]	—	0.15 (0.03)[f,g]	<.001

[a]Unstandardized regression coefficients were reported.

[b]Standard errors are within parentheses; bootstrap sample size = 10,000.

[c]Gender: 0=female, 1=male.

[d]Not applicable.

[e]R^2=16.6%; $F_{10,1557}$=30.98; *P*<.001.

[f]R^2=2.5%; $F_{11,1556}$=3.67.

[g]Boot effects: *B*=0.03 (SE 0.01, 95% CI 0.01-0.04) for indirect effects; *B*=–0.03 (0.04, 95% CI –0.11 to 0.05) for direct effects.

XSL•FO
RenderX

Table 3. Regressing family members–acquaintances trust gap on antecedents[a].

	Negative emotions		Family members–acquaintances trust gap	
	B (SE[b])	*P* values	*B* (SE)	*P* values
Control variables				
Age (years)	–0.003 (0.002)	.26	0.02 (0.003)	<.001
Gender[c]	–0.14 (0.04)	<.001	0.08 (0.05)	.08
Income	–0.003 (0.01)	.76	0.005 (0.01)	.72
Education	0.06 (0.02)	.009	–0.07 (0.03)	.01
Risk perception	0.04 (0.005)	<.001	–0.006 (0.01)	.29
COVID-19 knowledge	0.06 (0.03)	.05	–0.08 (0.04)	.02
Trust in government	–0.20 (0.03)	<.001	0.06 (0.04)	.15
Trust in medical systems	–0.18 (0.03)	<.001	0.05 (0.04)	.18
Trust in scientists	0.04 (0.04)	.29	–0.03 (0.05)	.56
Antecedents				
Digital media use	0.17 (0.03)	<.001	–0.04 (0.04)	.29
Negative emotions	__[d,e]	—	0.08 (0.03)[f,g]	.01

[a]Unstandardized regression coefficients were reported.

[b]Standard errors are within parentheses; bootstrap sample size = 10,000.

[c]Gender: 0=female, 1=male.

[d]Not applicable.

[e]R^2=16.6%; $F_{10,1557}$=30.98; P<.001.

[f]R^2=4.6%, $F_{11,1556}$=6.86.

[g]Boot effects: B=0.01 (SE 0.01, 95% CI 0.003-0.027) for indirect effects; B=–0.04 (0.04, 95% CI –0.12 to 0.03) for direct effects.

Figure 2. Diagrams of mediation effects. Unstandardized regression coefficients are reported.

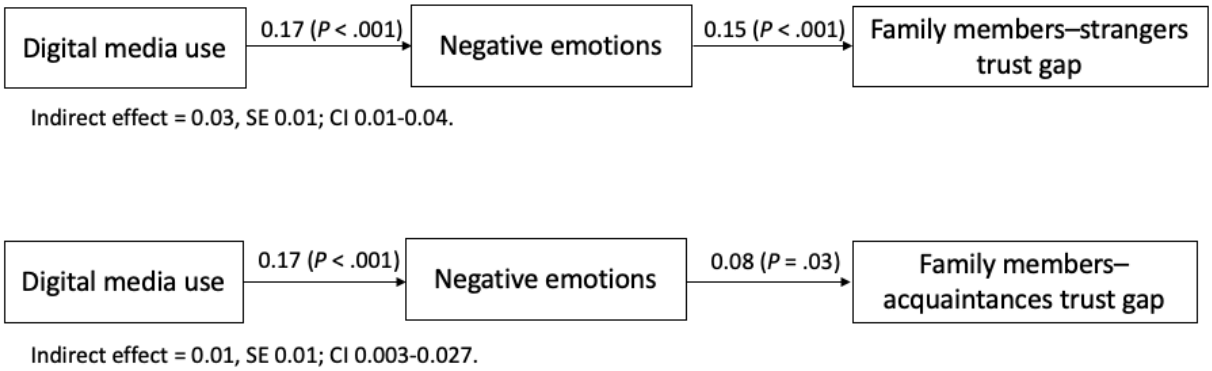

Discussion

Principal Findings

This study assessed the relationship between digital media use, negative emotions, the family members–strangers trust gap, and the family members–acquaintances trust gap during the COVID-19 pandemic in China. A majority of prior research used macrolevel perspectives to explain trust gaps. For instance, the differential mode of association (*chaxugeju*) emphasized that the kinship-based Chinese culture was the major cause of trust gaps, in which people trusted their family members more than others [11]. Besides, the rapid marketization and increasing mobility in China over the past few decades have enlarged the social distance between strangers and unfamiliar ones [64-66], thereby increasing trust gaps. Compared with this line of research, this study used a microlevel perspective and showed that individuals' digital media use was an important predictor of trust gaps. Notably, the Chinese character *trust* (信) is composed of *people* (亻) and *words* (言), which indicates that what people say makes a difference to trust [67]. Moreover, due to the highly contagious nature of the coronavirus, people relied on digital media to acquire information and post comments during the pandemic. These findings indicate that what people said on digital media contributed to the change in their trust in others. Hence, in the digital society, digital media use serves as a valid approach to explicate trust gaps during public health emergencies.

XSL•FO
RenderX

The results demonstrated the mediating role of negative emotions in the associations between digital media use and 2 trust gaps. On the one hand, the positive association between digital media use and negative emotions was consistent with previous studies that focused on public health emergencies [19,20,38,39]. These studies showed that the wide use of digital media facilitated the spread of the virus-related information, which not only amplified people's negative emotions but also resulted in emotion contagion among the population [68,69]. On the other hand, the positive associations between negative emotions and 2 trust gaps highlighted the role of negative emotions in motivating an individual to attribute the responsibility of a crisis to socially distant ones [41]. Negative emotions such as anxiety, anger, and hostility tend to create a sentiment of distrust in others, including strangers and acquaintances, thereby increasing the trust gaps. Prior research mainly focused on the impact of emotion contagion in social media on individual behaviors during public health crises [69,70]. In comparison, by examining the impact of negative emotions on 2 trust gaps, our study adds some insights into understanding how emotion contagion influences interpersonal ties in the digital environment.

By comparing the indirect effect of digital media use on 2 trust gaps via negative emotions, we found that negative emotions tended to enlarge the family members–strangers trust gap to a larger extent than the family members–acquaintances trust gap. These findings not only echo that strong aversive emotions result in individuals' distrust in strangers or loosely connected ones [43-45] but also indicate that people tend to attribute more responsibility to socially distant others, thereby leading to more distrust in strangers than in acquaintances. In recent years, strangers have been increasingly distanced as untrustworthy others, which suggests a decrease in social trust [66,71]. These findings remind us to stay alert to the widening family members–strangers trust gap caused by frequent digital media use and associated negative emotions.

In the past few years, the COVID-19 pandemic has exerted a profound influence on the ways people think, behave, and interact with others. Although the World Health Organization ended global health emergency declaration for the COVID-19 pandemic on May 5, 2023 [1], large-scale infectious diseases might break out in the future [5]. Moreover, the postpandemic era is increasingly shaped by the use of various digital tools [72]. As long as biased and uncivil comments about out-group members circulate on digital media, people's exposure to this information tends to amplify their negative emotions, which in turn enlarges 2 trust gaps. Thus, the findings of this study have implications for curbing the enlargement of trust gaps in the postpandemic era.

We suggest policy makers to take measures to pursue an emotionally balanced digital media environment with less uncivil, hostile, or sensational messages about out-group members when an infectious disease breaks out. For instance, the emotionally laden web-based misinformation and disinformation should be identified and labeled to remind users to keep alert of these contents. Besides, algorithms that push diverse viewpoints to users can be used to avoid emotional polarization. These efforts may ease people's negative emotions during their digital media use, thereby curbing the enlargement of trust gaps. Meanwhile, we advise digital media users to improve their verification skills. When encountering a piece of news on digital media, users should first verify the credentials of the authors, such as whether the authors are specialized in the field covered by the news. Then, users should check the source by asking these questions: Are there any cited sources in the article? How many sources are cited? How credible are the sources? Finally, users should examine whether the news article includes biased or sensational contents. The 3-step verification behavior helps users avoid the overamplification of negative emotions and prevent the enlargement of trust gaps, even if they encounter fake news on digital media.

Limitations and Future Directions

This study has several limitations. First, the measurement of digital media use might be too general to capture its precise impact on negative emotions and trust gaps. Although frequency of exposure to information may provide much information about the influence of digital media use on users' perceptions [73], researchers could analyze the content related to negative emotions and distrust in others on digital media to replicate this study. In addition, although prior studies have demonstrated that people primarily experience negative emotions at the early stage of a public health emergency [35,38,39], they may also feel positive emotions, such as hope, to navigate them through uncertainties [74,75]. Accordingly, future research could incorporate positive emotions into the model to get a more nuanced and comprehensive understanding of how digital media use influences social trust–related outcomes through negative and positive emotions. Besides, although the single-item measurement of trust in certain groups has been frequently used [76,77], multiple items should be used to measure one's trust in family members, acquaintances, and strangers in the future to improve the construct reliability.

Second, in terms of data collection, the cross-sectional web-based survey provided empirical evidence for testing the associations between the examined variables. However, we cannot claim causality in the proposed model. It is possible that participants who experienced negative emotions and distrusted strangers and acquaintances selected to expose themselves to biased and uncivil contents to confirm their existing perceptions. Moreover, given that negative emotions were highly volatile over the course of the pandemic [78], researchers are advised to use multiple waves of survey that focus on different stages of a public health crisis in the future. This longitudinal design not only enables a researcher to trace the evolvement of negative emotions but also warrants the examination of the causal relationships between digital media use, negative emotions, and trust gaps.

Finally, our survey was conducted at the onset of the pandemic. The early stage of a pandemic was characterized by high public demand for information [79], strong negative emotions [80], and a fragile state of social trust [81]. Conducting the survey at the middle or late stages of the pandemic might have yielded different results. Therefore, the findings should be interpreted as specific to the initial stage of a public health emergency, instead of being generalized to other phases. However, as an

XSL•FO
RenderX

exploratory study, the current model may serve as a link between the early-stage trust dynamics and their long-term impacts in the postpandemic era. Researchers can use our findings to explicate trust-related outcomes in the present days.

Conclusions

The widening trust gaps amid public health emergencies not only undermine social integration but also hinder postpandemic recovery. This study examines trust gaps in the early phase of the COVID-19 pandemic in China and demonstrates that frequent digital media use was positively associated with negative emotions. These emotions, in turn, were positively associated with 2 trust gaps: the family members–strangers trust gap and the family members–acquaintances trust gap. Moreover, the mediation effect of negative emotions between digital media use and the family members–strangers trust gap was stronger than that between digital media use and the family members–acquaintances trust gap. Compared with previous macro-level studies that focused on factors such as culture and modernity to explain the decline in social trust [11,82], our study contributes in two ways: (1) delineating the structure of decreasing social trust into 2 distinct trust gaps, and (2) using a media effect approach to unveiling the individual-level mechanism that contributes to the widening trust gaps. Against the backdrop of the mediatized era and global risk society, the findings underscore the importance of the appropriate use of digital media and the strategic management of public emotions to curb the widening trust gaps. This, in turn, can facilitate effective public health crisis intervention in the postpandemic era.

Acknowledgments

This study is supported by the National Social Science Fund of China under grant (project number 19ZDA325).

Authors' Contributions

LW acquired the funding, designed the study, wrote the main body of the manuscript, and edited the manuscript. QH participated in the study design, analyzed the data, and wrote the draft of the manuscript.

Conflicts of Interest

None declared.

References

1. World Health Organization. WHO director-general's opening remarks at the media briefing—5 May 2023. 2023 May 05. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023 [accessed 2023-05-26]
2. Silver L, Connaughton A, Pew Research Center. Partisanship colors views of Covid-19 handling across advanced economies. 2022 Aug 11. URL: https://tinyurl.com/yuaajsts [accessed 2024-05-15]
3. Shahbazi M, Bunker D. Social media trust: fighting misinformation in the time of crisis. Int J Inf Manage 2024;77:102780. [doi: 10.1016/j.ijinfomgt.2024.102780]
4. Li Y, Li H, Ruan J. Do long-term natural disasters influence social trust? empirical evidence from China. Int J Environ Res Public Health 2021;18(14):7280 [FREE Full text] [doi: 10.3390/ijerph18147280] [Medline: 34299730]
5. Sozinova AA, Popkova EG. Dataset analysis of pandemic risks and risk management prospects based on management and marketing in conditions of Covid-19 recession. Risks 2023;11(2):37. [doi: 10.3390/risks11020037]
6. Pitas N, Ehmer C. Social capital in the response to Covid-19. Am J Health Promot 2020;34(8):942-944 [FREE Full text] [doi: 10.1177/0890117120924531] [Medline: 32394721]
7. Yuan H, Long Q, Huang G, Huang L, Luo S. Different roles of interpersonal trust and institutional trust in Covid-19 pandemic control. Soc Sci Med 2022;293:114677 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114677] [Medline: 35101260]
8. Bäck M, Christensen HS. When trust matters—a multilevel analysis of the effect of generalized trust on political participation in 25 European democracies. J Civil Soc 2016;12(2):178-197. [doi: 10.1080/17448689.2016.1176730]
9. Putnam RD. Bowling alone: America's declining social capital. In: Crothers L, Lockhart C, editors. Culture and Politics. New York, NY: Palgrave Macmillan; 2000:223-234.
10. Beaudoin CE, Thorson E. Social capital in rural and urban communities: testing differences in media effects and models. J Mass Commun Q 2016;81(2):378-399 [FREE Full text] [doi: 10.1177/107769900408100210]
11. Fei X. From the Soil: The Foundations of Chinese Society. Berkeley, CA: University of California Press; 1992.
12. Song F, Cadsby C, Bi Y. Trust, reciprocity, and guanxi in China: an experimental investigation. Manag Organ Rev 2012;8(2):397-421. [doi: 10.1111/j.1740-8784.2011.00272.x]
13. Fell L. Trust and Covid-19: implications for interpersonal, workplace, institutional, and information-based trust. Digit Gov Res Pract 2020;2(1):1-5. [doi: 10.1145/3428472]
14. Wei Q. Stay vigilant of the panic and other negative impacts of the Covid-19 outbreak. 2020. URL: http://finance.sina.com.cn/zl/china/2020-02-27/zl-iimxxstf4772092.shtml [accessed 2021-08-19]

XSL•FO
RenderX

15. Bierman A, Schieman S. Social estrangement and psychological distress before and during the Covid-19 pandemic: patterns of change in Canadian workers. J Health Soc Behav 2020;61(4):398-417. [doi: 10.1177/0022146520970190] [Medline: 33211540]
16. Karmakar S, Das S. Evaluating the impact of Covid-19 on cyberbullying through Bayesian trend analysis. In: EICC '20: Proceedings of the 2020 European Interdisciplinary Cybersecurity Conference. 2020 Presented at: Association for Computing Machinery; 2021; New York, NY p. 1-6. [doi: 10.1145/3424954.3424960]
17. Kim B. Effects of social grooming on incivility in Covid-19. Cyberpsychol Behav Soc Netw 2020;23(8):519-525. [doi: 10.1089/cyber.2020.0201] [Medline: 32267163]
18. Xu J, Sun G, Cao W, Fan W, Pan Z, Yao Z, et al. Stigma, discrimination, and hate crimes in Chinese-speaking world amid Covid-19 pandemic. Asian J Criminol 2021;16(1):51-74 [FREE Full text] [doi: 10.1007/s11417-020-09339-8] [Medline: 33425062]
19. Chao M, Xue D, Liu T, Yang H, Hall BJ. Media use and acute psychological outcomes during Covid-19 outbreak in China. J Anxiety Disord 2020;74:102248 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102248] [Medline: 32505918]
20. Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S, et al. Mental health problems and social media exposure during Covid-19 outbreak. PLoS One 2020;15(4):e0231924 [FREE Full text] [doi: 10.1371/journal.pone.0231924] [Medline: 32298385]
21. Fischer AH, Manstead ASR. Social functions of emotion and emotion regulation. In: Handbook of Emotions, 4th Edition. New York, NY: Guilford Press; 2016:424-439.
22. Jin Y, Liu BF, Anagondahalli D, Austin L. Scale development for measuring publics' emotions in organizational crises. Public Relat Rev 2014;40(3):509-518. [doi: 10.1016/j.pubrev.2014.04.007]
23. Liu PL. Covid-19 information seeking on digital media and preventive behaviors: the mediation role of worry. Cyberpsychol Behav Soc Netw 2020;23(10):677-682. [doi: 10.1089/cyber.2020.0250] [Medline: 32498549]
24. Melki J, Tamim H, Hadid D, Farhat S, Makki M, Ghandour L, et al. Media exposure and health behavior during pandemics: the mediating effect of perceived knowledge and fear on compliance with Covid-19 prevention measures. Health Commun 2022;37(5):586-596. [doi: 10.1080/10410236.2020.1858564] [Medline: 33327785]
25. Ren W, Zhu X, Hu Y. Differential effects of traditional and social media use on Covid-19 preventive behaviors: the mediating role of risk and efficacy perceptions. J Health Psychol 2022;27(8):1861-1874. [doi: 10.1177/13591053211003125] [Medline: 33909510]
26. Huang YC, Wu F, Huang Q. Does research on digital public relations indicate a paradigm shift? An analysis and critique of recent trends. Telematics Inform 2017;34(7):1364-1376. [doi: 10.1016/j.tele.2016.08.012]
27. Mou Y, Atkin D, Fu H, Lin CA, Lau T. The influence of online forum and SNS use on online political discussion in China: assessing "spirals of trust". Telematics Inform 2013;30(4):359-369. [doi: 10.1016/j.tele.2013.04.002]
28. Leung L. Generational differences in content generation in social media: the roles of the gratifications sought and of narcissism. Comput Human Behav 2013;29(3):997-1006. [doi: 10.1016/j.chb.2012.12.028]
29. Leung L, Wei R. More than just talk on the move: uses and gratifications of the cellular phone. J Mass Commun Q 2016;77(2):308-320. [doi: 10.1177/107769900007700206]
30. Wang R, Yang F, Zheng S, Sundar SS. Why do we pin? New gratifications explain unique activities in pinterest. Social Media Soc 2016;2(3):205630511666217. [doi: 10.1177/2056305116662173]
31. Islm T, Meng H, Pitafi AH, Ullah Zafar A, Sheikh Z, Shujaat Mubarik M, et al. Why do citizens engage in government social media accounts during Covid-19 pandemic? A comparative study. Telemat Inform 2021;62:101619 [FREE Full text] [doi: 10.1016/j.tele.2021.101619] [Medline: 36568845]
32. Shim H, You KH, Lee JK, Go E. Why do people access news with mobile devices? Exploring the role of suitability perception and motives on mobile news use. Telematics Inform 2015;32(1):108-117. [doi: 10.1016/j.tele.2014.05.002]
33. Yamamoto M, Morey AC. Incidental news exposure on social media: a campaign communication mediation approach. Social Media Soc 2019;5(2):205630511984361. [doi: 10.1177/2056305119843619]
34. Hu B, Zhang D. Channel selection and knowledge acquisition during the 2009 Beijing H1N1 flu crisis: a media system dependency theory perspective. Chinese J Commun 2014;7(3):299-318. [doi: 10.1080/17544750.2014.926951]
35. Yang ZJ. Too scared or too capable? Why do college students stay away from the H1N1 vaccine? Risk Anal 2012;32(10):1703-1716. [doi: 10.1111/j.1539-6924.2012.01799.x] [Medline: 22404650]
36. Fletcher R, Nielsen RK. Are people incidentally exposed to news on social media? A comparative analysis. New Media Soc 2017;20(7):2450-2468. [doi: 10.1177/1461444817724170]
37. Kim Y, Chen HT, Gil de Zúñiga H. Stumbling upon news on the internet: effects of incidental news exposure and relative entertainment use on political engagement. Comput in Hum Behav 2013;29(6):2607-2614. [doi: 10.1016/j.chb.2013.06.005]
38. Seo M. Amplifying panic and facilitating prevention: multifaceted effects of traditional and social media use during the 2015 MERS crisis in South Korea. J Mass Commun Q 2019;98(1):221-240. [doi: 10.1177/1077699019857693]
39. Qiu J, Xu L, Wang J, Gu W. Mutual influences between message volume and emotion intensity on emerging infectious diseases: an investigation with microblog data. Inf Manage 2020;57(4):103217. [doi: 10.1016/j.im.2019.103217]
40. A report on public attitudes toward Covid-19 and circulation of associated information. National Information Center & Nanjing University. 2020. URL: http://www.sic.gov.cn/News/79/10414.htm [accessed 2021-08-20]

XSL•FO
RenderX

41. Dunn JR, Schweitzer ME. Feeling and believing: the influence of emotion on trust. J Pers Soc Psychol 2005;88(5):736-748. [doi: 10.1037/0022-3514.88.5.736] [Medline: 15898872]
42. Bloom JD, Chan YA, Baric RS, Bjorkman PJ, Cobey S, Deverman BE, et al. Investigate the origins of Covid-19. Science 2021;372(6543):694 [FREE Full text] [doi: 10.1126/science.abj0016] [Medline: 33986172]
43. Huang JY, Sedlovskaya A, Ackerman JM, Bargh JA. Immunizing against prejudice: effects of disease protection on attitudes toward out-groups. Psychol Sci 2011;22(12):1550-1556. [doi: 10.1177/0956797611417261] [Medline: 22058107]
44. Prati G, Pietrantoni L. Knowledge, risk perceptions, and xenophobic attitudes: evidence from Italy during the ebola outbreak. Risk Anal 2016;36(10):2000-2010. [doi: 10.1111/risa.12537] [Medline: 26916803]
45. Tsai J, Phua J, Pan S, Yang C. Intergroup contact, Covid-19 news consumption, and the moderating role of digital media trust on prejudice toward Asians in the United States: cross-sectional study. J Med Internet Res 2020;22(9):e22767 [FREE Full text] [doi: 10.2196/22767] [Medline: 32924948]
46. Bavel JJV, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, et al. Using social and behavioural science to support Covid-19 pandemic response. Nat Hum Behav 2020;4(5):460-471. [doi: 10.1038/s41562-020-0884-z] [Medline: 32355299]
47. Au AKY, Ng JCK, Wu WCH, Chen SX. Who do we trust and how do we cope with Covid-19? A mixed-methods sequential exploratory approach to understanding supportive messages across 35 cultures. Humanit Soc Sci Commun 2023;10(1):272 [FREE Full text] [doi: 10.1057/s41599-023-01747-2] [Medline: 37273414]
48. Chen X, Chen H. Differences in preventive behaviors of Covid-19 between urban and rural residents: lessons learned from a cross-sectional study in China. Int J Environ Res Public Health 2020;17(12):4437 [FREE Full text] [doi: 10.3390/ijerph17124437] [Medline: 32575700]
49. Huang Q. How does news media exposure amplify publics' perceived health risks about air pollution in China? A conditional media effect approach. Int J Commun 2020;14:1705-1724.
50. Wang K, Li S. Who is willing to engage in social gatherings during the coronavirus disease 2019 lockdown? A curvilinear relationship between age and heuristic processing. Front Psychol 2020;11:586408 [FREE Full text] [doi: 10.3389/fpsyg.2020.586408] [Medline: 33343459]
51. The 45th China statistical report on internet development. China Internet Network Information Center. 2020 Apr 28. URL: https://www.gov.cn/xinwen/2020-04/28/content_5506903.htm [accessed 2020-12-04]
52. Näsi M, Räsänen P, Hawdon J, Holkeri E, Oksanen A. Exposure to online hate material and social trust among finnish youth. Inf Technol People Emerald Group Publishing Limited 2015;28(3):622. [doi: 10.1108/itp-09-2014-0198]
53. Kim HK, Niederdeppe J. The role of emotional response during an H1N1 influenza pandemic on a college campus. J Public Relat Res 2013;25(1):30-50. [doi: 10.1080/1062726x.2013.739100]
54. Cheung C, Tse JW. Institutional trust as a determinant of anxiety during the SARS crisis in Hong Kong. Soc Work Public Health 2008;23(5):41-54. [doi: 10.1080/19371910802053224] [Medline: 19301537]
55. Mou Y, Lin CA. Communicating food safety via the social media. Sci Commun 2014;36(5):593-616. [doi: 10.1177/1075547014549480]
56. Yue RPH, Lau BHP, Chan CLW, Ng SM. Risk perception as a double-edged sword in policy compliance in Covid-19 pandemic? A two-phase evaluation from Hong Kong. J Risk Res 2021;25(9):1131-1145. [doi: 10.1080/13669877.2021.1936612]
57. Hou SI. Objective and subjective knowledge and HIV testing among college students. Am J Health Educ 2013;35(6):328-337. [doi: 10.1080/19325037.2004.10604772]
58. Berhe NM, Van de Velde S, Rabiee-Khan F, van der Heijde C, Vonk P, Buffel V, et al. Knowledge deficit and fear of Covid-19 among higher education students during the first wave of the pandemic and implications for public health: a multi-country cross-sectional survey. BMC Public Health 2022;22(1):1144 [FREE Full text] [doi: 10.1186/s12889-022-13511-3] [Medline: 35672723]
59. Huang Q. Public trust in local governments and environmental risks in China: the effects of media use, perceived dread, and perceived inequality. Chin J Commun 2017;11(1):88-104. [doi: 10.1080/17544750.2017.1328452]
60. Shea JA, Micco E, Dean LT, McMurphy S, Schwartz JS, Armstrong K. Development of a revised health care system distrust scale. J Gen Intern Med 2008;23(6):727-732 [FREE Full text] [doi: 10.1007/s11606-008-0575-3] [Medline: 18369678]
61. Hmielowski JD, Feldman L, Myers TA, Leiserowitz A, Maibach E. An attack on science? Media use, trust in scientists, and perceptions of global warming. Public Underst Sci 2014;23(7):866-883. [doi: 10.1177/0963662513480091] [Medline: 23825287]
62. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach: Second Edition. New York, NY: Guilford Publications; 2013.
63. Preacher KJ, Rucker DD, Hayes AF. Addressing moderated mediation hypotheses: theory, methods, and prescriptions. Multivariate Behav Res 2007;42(1):185-227. [doi: 10.1080/00273170701341316] [Medline: 26821081]
64. Jiang F. Space-time changes of hierarchical trust and its typology (in Chinese). Zhejiang Soc Sci 2017;10:66-73.
65. Xin Z, Xin S. Marketization process predicts trust decline in China. J Econ Psychol 2017;62:120-129. [doi: 10.1016/j.joep.2017.07.001]
66. Yang Z, Xin Z. Income inequality and interpersonal trust in China. Asian J Soc Psychol 2019;23(3):253-263. [doi: 10.1111/ajsp.12399]

XSL•FO
RenderX

67. Zhai X. Social Trust Among Chinese People: A Relational Approach (In Chinese). Beijing: The Commercial Press; 2022.
68. Goldenberg A, Gross JJ. Digital emotion contagion. Trends Cogn Sci 2020;24(4):316-328. [doi: 10.1016/j.tics.2020.01.009] [Medline: 32160568]
69. Wheaton M, Prikhidko A, Messner G. Is fear of Covid-19 contagious? The effects of emotion contagion and social media use on anxiety in response to the coronavirus pandemic. Front Psychol 2021;11:567379 [FREE Full text] [doi: 10.3389/fpsyg.2020.567379] [Medline: 33469434]
70. Shao R, Shi Z, Zhang D. Social media and emotional burnout regulation during the Covid-19 pandemic: multilevel approach. J Med Internet Res 2021;23(3):e27015 [FREE Full text] [doi: 10.2196/27015] [Medline: 33661753]
71. Wang Y, Zhou X. A hierarchy of social trust in urban China (in Chinese). New Horiz Tianfu 2009;146(2):108-111.
72. Jiang H, Zhang Y, Guo W, Cheng W, Peng J. Online collaborative documents as media logic: the mediatization of risk response in the post-pandemic era. Front Psychol 2022;13:892569 [FREE Full text] [doi: 10.3389/fpsyg.2022.892569] [Medline: 35814134]
73. Wong FHC, Liu T, Leung DKY, Zhang AY, Au WSH, Kwok WW, et al. Consuming information related to Covid-19 on social media among older adults and its association with anxiety, social trust in information, and Covid-safe behaviors: cross-sectional telephone survey. J Med Internet Res 2021;23(2):e26570 [FREE Full text] [doi: 10.2196/26570] [Medline: 33523825]
74. Russell PS, Frackowiak M, Cohen-Chen S, Rusconi P, Fasoli F. Induced gratitude and hope, and experienced fear, but not experienced disgust, facilitate Covid-19 prevention. Cogn Emot 2023;37(2):196-219. [doi: 10.1080/02699931.2022.2157377] [Medline: 36628460]
75. Leslie-Miller CJ, Waugh CE, Cole VT. Coping with Covid-19: the benefits of anticipating future positive events and maintaining optimism. Front Psychol 2021;12:646047 [FREE Full text] [doi: 10.3389/fpsyg.2021.646047] [Medline: 33897550]
76. Su Z, Ye Y, He J, Huang W. Constructed hierarchical government trust in China: formation mechanism and political effects. Pac Aff 2016;89(4):771-794. [doi: 10.5509/2016894771]
77. Zhang W, Chia SC. The effects of mass media use and social capital on civic and political participation. Commun Stud 2006;57(3):277-297. [doi: 10.1080/10510970600666974]
78. Hagen D, Lai AY, Goldmann E. Trends in negative emotions throughout the Covid-19 pandemic in the United States. Public Health 2022;212:4-6 [FREE Full text] [doi: 10.1016/j.puhe.2022.08.009] [Medline: 36162396]
79. Tang S, Wu X, Chen J, Lu F, Zhang Z, Xu Y, et al. Release and demand of public health information in social media during the outbreak of Covid-19 in China. Front Public Health 2021;9:829589 [FREE Full text] [doi: 10.3389/fpubh.2021.829589] [Medline: 35223765]
80. Meng G, Li Q, Yuan X, Zheng Y, Hu K, Dai B, et al. The roles of risk perception, negative emotions and perceived efficacy in the association between Covid-19 infection cues and preventive behaviors: a moderated mediation model. BMC Public Health 2023;23(1):109 [FREE Full text] [doi: 10.1186/s12889-022-14870-7] [Medline: 36647034]
81. Abdalla SM, Koya SF, Jamieson M, Verma M, Haldane V, Jung AS, et al. Investing in trust and community resilience: lessons from the early months of the first digital pandemic. BMJ 2021;375:e067487 [FREE Full text] [doi: 10.1136/bmj-2021-067487] [Medline: 34840130]
82. Anthony G. The Consequences of Modernity. Redwood City, CA: Stanford University Press; 1991.

Abbreviations

CNY: Chinese Yuan
MERS: Middle East respiratory syndrome

Edited by A Mavragani; submitted 28.05.23; peer-reviewed by X Wang, L Heidi; comments to author 19.01.24; revised version received 01.02.24; accepted 24.05.24; published 10.07.24.

Please cite as:
Wei L, Huang Q
Retrospecting Digital Media Use, Negative Emotions, and Trust Gaps During the COVID-19 Pandemic in China: Cross-Sectional Web-Based Survey
J Med Internet Res 2024;26:e49422
URL: https://www.jmir.org/2024/1/e49422
doi:10.2196/49422
PMID:

XSL•FO
RenderX

©Lu Wei, Qing Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality Assessment of TikTok as a Source of Information About Mitral Valve Regurgitation in China: Cross-Sectional Study

Nannan Cui[1*], MD; Yuting Lu[2*], MD; Yelin Cao[1], MD; Xiaofan Chen[3], PhD; Shuiqiao Fu[1], MD; Qun Su[1], MD

[1]Department of Surgical Intensive Care Unit, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Ultrasonography, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[3]Department of Cardiovascular Surgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Qun Su, MD
Department of Surgical Intensive Care Unit, The First Affiliated Hospital
Zhejiang University School of Medicine
#79 Qingchun Road
Hangzhou, 310003
China
Phone: 86 13588045704
Email: sqicu@zju.edu.cn

Abstract

Background: In China, mitral valve regurgitation (MR) is the most common cardiovascular valve disease. However, patients in China typically experience a high incidence of this condition, coupled with a low level of health knowledge and a relatively low rate of surgical treatment. TikTok hosts a vast amount of content related to diseases and health knowledge, providing viewers with access to relevant information. However, there has been no investigation or evaluation of the quality of videos specifically addressing MR.

Objective: This study aims to assess the quality of videos about MR on TikTok in China.

Methods: A cross-sectional study was conducted on the Chinese version of TikTok on September 9, 2023. The top 100 videos on MR were included and evaluated using quantitative scoring tools such as the modified DISCERN (mDISCERN), the Journal of the American Medical Association (JAMA) benchmark criteria, the Global Quality Score (GQS), and the Patient Education Materials Assessment Tool for Audio-Visual Content (PEMAT-A/V). Correlation and stepwise regression analyses were performed to examine the relationships between video quality and various characteristics.

Results: We obtained 88 valid video files, of which most (n=81, 92%) were uploaded by certified physicians, primarily cardiac surgeons, and cardiologists. News agencies/organizations and physicians had higher GQS scores compared with individuals (news agencies/organizations vs individuals, P=.001; physicians vs individuals, P=.03). Additionally, news agencies/organizations had higher PEMAT understandability scores than individuals (P=.01). Videos focused on disease knowledge scored higher in GQS (P<.001), PEMAT understandability (P<.001), and PEMAT actionability (P<.001) compared with videos covering surgical cases. PEMAT actionability scores were higher for outpatient cases compared with surgical cases (P<.001). Additionally, videos focused on surgical techniques had lower PEMAT actionability scores than those about disease knowledge (P=.04). The strongest correlations observed were between thumbs up and comments (r=0.92, P<.001), thumbs up and favorites (r=0.89, P<.001), thumbs up and shares (r=0.87, P<.001), comments and favorites (r=0.81, P<.001), comments and shares (r=0.87, P<.001), and favorites and shares (r=0.83, P<.001). Stepwise regression analysis identified “length (P<.001),” “content (P<.001),” and “physicians (P=.004)” as significant predictors of GQS. The final model (model 3) explained 50.1% of the variance in GQSs. The predictive equation for GQS is as follows: GQS = 3.230 − 0.294 × content − 0.274 × physicians + 0.005 × length. This model was statistically significant (P=.004) and showed no issues with multicollinearity or autocorrelation.

Conclusions: Our study reveals that while most MR-related videos on TikTok were uploaded by certified physicians, ensuring professional and scientific content, the overall quality scores were suboptimal. Despite the educational value of these videos, the guidance provided was often insufficient. The predictive equation for GQS developed from our analysis offers valuable insights but should be applied with caution beyond the study context. It suggests that creators should focus on improving both the content and presentation of their videos to enhance the quality of health information shared on social media.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55403) doi:10.2196/55403

KEYWORDS

mitral valve regurgitation; video quality; TikTok; Journal of American Medical Association; JAMA; Global Quality Score; GQS; PEMAT- A/V; Spearman correlation analysis; Poisson regression analysis

Introduction

Mitral valve regurgitation (MR) is a condition characterized by the incomplete closure of the mitral valve, allowing blood to flow backward from the left ventricle into the left atrium during systole. MR is one of the most common heart valve diseases, affecting 2%-3% of the population, and its prevalence and severity increase with age. Among individuals older than 75 years, the prevalence of moderate or severe MR is over 10% [1,2]. The prognosis of MR depends on factors such as left ventricular function, etiology, severity, and duration of the disease.

MR is the second-most prevalent heart valve disease requiring surgical intervention in Europe [1]. In the United States, surveys indicate that nearly 2 million patients have moderate to severe MR, and this number is expected to increase to 4 million by 2030 [2]. In China, the detection rate of moderate and severe MR is 2.2% [3]. A large cross-sectional survey of a Chinese hospital population suggests that MR is the most common valvular heart disease in the country [4,5]. The proportion of patients with secondary MR gradually increases with age. A sampling survey showed MR rates of 18.4% among 35-year-old respondents and 25.2% among 65-year-old respondents. Additionally, secondary etiologies accounted for 51.7% of MR cases in 75-year-old respondents [5]. It is estimated that approximately 10 million patients with MR currently need treatment in China [6]. Despite the relative prevalence of MR, Chinese patients are often characterized by a high incidence of the disease, low levels of health awareness, and a low rate of surgical treatment. Fewer than 20% of patients have confirmed valvular disease before hospital admission, and only about 33% undergo invasive treatment [4]. Studies have shown that early diagnosis and intervention significantly improve survival rates [2,7-10]. Therefore, medical health education plays a crucial role in guiding people to recognize early symptoms, seek timely medical treatment, and gain access to early interventions.

Video-sharing media platforms, such as YouTube (Google LLC) and TikTok (ByteDance), have become integral to both personal and professional life. Videos featuring cartoons or documentary content are particularly popular, especially those related to occupational health and disease knowledge [11]. Many people access medical information online, and it is common for patients to look up information both before and after seeking medical treatment [12]. TikTok, a short-video–sharing social media platform that has gained significant popularity in recent years, offers a wide range of content across nearly all fields, including numerous health care–related videos. With up to 600 million active users in China, TikTok serves as a potentially crucial channel for distributing health information to consumers [13]. While TikTok holds great potential for disseminating public health information, the quality of disease-related videos on the platform varies widely. Researchers have evaluated the quality of TikTok videos related to cholelithiasis, COVID-19, diabetes, and chronic obstructive pulmonary disease [14-17]. However, the quality of videos about magnetic resonance (MR) imaging available on TikTok has not been assessed. Therefore, we investigated MR-related videos on TikTok to identify their upload sources, content, and features. We utilized quantitative scoring tools such as DISCERN, the Journal of the American Medical Association (JAMA) benchmark criteria, and the Global Quality Score (GQS) to conduct this evaluation. We also evaluated the educational guidance provided to the audience using the Patient Education Materials Assessment Tool for Audio-Visual Content (PEMAT-A/V) method.

Methods

Ethical Considerations

No clinical data, human specimens, or laboratory animals were involved in this study. Although our study is primarily based on a secondary analysis of mitral valve videos on TikTok, we placed significant emphasis on ethical considerations. We ensured that all research data were anonymized or deidentified to fully protect participant privacy. For any information that could potentially lead to personal identification, we implemented strict confidentiality measures, and access to the data was only provided to the research team.

Retrieval Strategy

On September 9, 2023, from 09:00 AM to 11:00 AM, we conducted a search on the Chinese version of TikTok using the term “二尖瓣反流” (mitral regurgitation). To minimize research bias, we logged out of all personal accounts and refrained from using any filters. The first 100 retrieved videos were included in the study. We limited our analysis to the top 100 videos, as several studies [18,19] have confirmed that videos beyond this range do not significantly affect analysis outcomes. Videos featuring animals (n=2) and those consisting solely of images (n=12) were excluded from the study. For all included videos, the following features were recorded and analyzed: title, URL, uploader, uploader’s identity, website authentication, video duration (seconds), upload date, number of likes, number of comments, number of favorites, number of shares, and upload days (days since upload).

Visual Classification

We classified video uploaders into 4 categories: doctors, individuals (eg, nonmedical professionals), news agencies (eg, online media, newspapers, television stations, and radio stations), and organizations (eg, hospitals, health authorities, research groups, universities, or colleges). The doctors were further classified as cardiac surgeons, cardiologists, or other health care professionals. The videos were categorized based on their content into the following types: disease knowledge,

XSL•FO
RenderX

outpatient medical records, surgical cases, surgical techniques, and news/advertising.

Video Content and Quality Assessment

The reliability of the videos was evaluated using the JAMA benchmark criteria and the modified DISCERN (mDISCERN) tool, while the quality of information was assessed using the GQS. Additionally, the educational impact of the video materials on the general public was measured using the PEMAT-A/V score. The quality assessment was conducted for all videos that met the eligibility criteria.

First, the JAMA benchmark criteria [20] were utilized to assess video reliability, encompassing 4 distinct criteria: (1) authorship; (2) attribution, including copyright information, references, and sources of content; (3) currency, including the initial date and subsequent updates; and (4) disclosure of conflicts of interest, funding, sponsorship, advertising support, and video ownership. Each criterion scores 1 point, with higher scores indicating greater reliability.

Second, the mDISCERN score, derived from the DISCERN tool, was used to analyze the reliability and quality of videos. It has been proven effective in assessing health-related video materials on other platforms, such as YouTube [21-24]. The mDISCERN score consists of 5 questions: (1) Are the aims of the video clear and achieved? (2) Are reliable sources of information used? (3) Is the information presented in a balanced and unbiased manner? (4) Are additional sources of information listed for patient reference? (5) Are areas of uncertainty mentioned? Each question is scored as 1 for “yes” and 0 for “no.” Higher scores indicate that the video is more reliable.

Third, we used the GQS to assess the quality of information in the videos. The GQS is widely recognized for evaluating the quality of health information on online video platforms [15,17,25]. It consists of 5 criteria: (1) poor quality (poor flow, most information missing, and not useful for patients); (2) generally poor quality (poor flow, some information provided but many important topics missing, of very limited use to patients); (3) moderate quality (suboptimal flow, some important information adequately discussed while others are poorly covered, somewhat useful for patients); (4) good quality (generally good flow, most relevant information listed, though some topics are not covered, useful for patients); and (5) excellent quality (excellent flow and very useful information for patients). Higher scores indicate higher-quality videos.

Finally, we used the PEMAT-A/V [26] to evaluate the educational impact of the video materials on the public. The PEMAT-A/V is specifically designed for assessing audiovisual materials and consists of 17 questions: 13 questions evaluate the understandability of the health information provided, while 4 questions assess the actionability of the recommendations. Responses are scored as “agree” (1), “disagree” (0), or “N/A.” The overall scores, as well as scores for the understandability and actionability sections, are calculated using the formula: “total points/total possible points × 100,” with higher scores indicating better understandability or actionability or both.

All authors are senior physicians with extensive clinical expertise in cardiology-related specialties. One person (NC) collected and downloaded all the videos. Two authors (NC and YC) evaluated the videos using the mDISCERN tool, JAMA score, GQS, and PEMAT-A/V. Scores were determined through discussion between them. In cases of disagreement, an arbitrator (XC) made the final decision. All authors subsequently approved each rating.

Data Analysis

Data are expressed as means (SD) or medians (ranges). The Kruskal-Wallis test was used to compare data among multiple groups, while the Mann-Whitney *U* test was used for comparisons between 2 groups. The Dunnett multiple comparison test was applied for pairwise comparisons following a 2-way analysis of variance.

Because of the categorical nature of our data set and the nonnormal distribution of the data, we used the Spearman correlation coefficient to assess interparameter correlations. The correlation strength was categorized as follows: less than 0.25 indicates a poor relationship; 0.25–0.50 signifies a moderate relationship; 0.50–0.75 denotes a good relationship; and 0.75–1.00 represents an excellent relationship.

The Spearman correlation analysis revealed a significant (Figure 6) association between video parameters and the GQS. To assess the predictive capability of these video parameters for GQS, we used stepwise regression analysis, considering data type, characteristics, and other relevant factors. Initially, collinearity detection was performed to identify and remove variables with a variance inflation factor (VIF) greater than 5. Subsequently, GQS was used as the dependent variable, with uploaders, physicians, titles, content, duration, and length as independent variables in a stepwise approach. A *P* value of less than .05 was considered statistically significant.

We used the intraclass correlation coefficient (ICC) with a 2-way fixed-effects model to evaluate the scores between the reviewers (NNC and YLC). ICC values range from 0 to 1, where values less than 0.5 indicate poor consistency, 0.5-0.75 indicate moderate consistency, 0.75-0.90 indicate good consistency, and values greater than 0.90 indicate excellent consistency.

Statistical analysis was conducted using IBM SPSS version 19.0 software (IBM Corp.). Figures were created with GraphPad Prism version 9.5.1 (GraphPad Software) and OriginPro 2021 software (Origin Laboratories).

Results

Video Characteristics

In the final analysis, our study identified 88 eligible video files. The descriptive statistics for these videos are summarized in Tables 1 and 2, showing a total of 476,511 likes, 33,123 comments, 39,781 favorites, and 49,314 shares. The median video length was 255 seconds, with durations ranging from 4 to 1197 seconds, reflecting a wide spectrum of video lengths. Additionally, the median time since video upload was 61 (range 4-491) days, indicating the varied timing of content dissemination.

XSL•FO
RenderX

Table 1. General characteristics of the videos.

Characteristics	Mean (SD)	Median (range)
Video length (seconds)	336.71 (273.99)	255 (4-1197)
Duration on TikTok (days)	76.22 (68.87)	61 (4-491)
Thumbs up	5728.02 (23,785.95)	433 (23-206,000)
Thumbs up/day	20.07 (107.09)	2.22 (0.05-853.88)
Comments	396.20 (1506.29)	46 (1-13,000)
Comments/day	1.91 (6.69)	0.23 (0-53.89)
Favorites	479.29 (1382.07)	98 (0-9522)
Favorites/day	4.11 (15.03)	0.31 (0-107.51)
Sharing (counts)	581.64 (1607.70)	61 (0-12,000)
Sharing/day	2.76 (8.46)	0.55 (0-49.74)
JAMA[a] score (0-4)	1.99 (0.11)	2 (1-2)
GQS[b] score (0-5)	2.75 (0.79)	3 (1-5)
mDISCERN[c] (0-5)	3.06 (0.23)	3 (3-4)
PEMAT[d] understandability score, %	70.14 (17.74)	73.87 (16.67-100)
PEMAT actionability score, %	54.55 (45.26)	66.67 (0-100)

[a]JAMA: Journal of the American Medical Association.
[b]GQS: Global Quality Score.
[c]mDISCERN: modified DISCERN.
[d]PEMAT: Patient Education Materials Assessment Tool.

XSL•FO
RenderX

Table 2. Detailed characteristics of videos based on uploaders and content[a].

Characteristic	n	Length (seconds)	Duration (days)	Thumbs up	Comments	Favorites	Sharing (counts)
Uploaders							
Doctors	81	63 (4-491)	256.28 (4-1958)	421 (23-206,000)	46 (1-13,000)	92 (0-9522)	66 (0-12,000)
Individuals	4	25.5 (874)	375.37 (74-949)	323 (140-1480)	38 (24-57)	48 (2-311)	7 (0-221)
News agency/organizations	3	201 (55-277)	650.5 (490.2-724.6)	399 (255-1451)	87 (13-145)	77.5 (0-155)	343 (0-401)
P value		.049	.17	.93	.94	.46	.22
Doctors							
Cardiac surgeons	39	63 (5-491)	214 (8-1197)	766 (29-39,000)	94 (3-2733)	152 (0-2149)	165 (0-3206)
Cardiologists	37	61 (4-210)	258 (4-1958)	270 (23-20,600)	32 (1-13,000)	50 (0-9522)	48 (1-12,000)
Other health care professionals	5	79.4 (33.1)	337.8 (308.7)	349.6 (303.5)	43.4 (38.8)	67.4 (64.3)	57 (1-419)
P value		.70	.81	.10	.17	.22	.36
Professional titles							
Chief physician	58	63.5 (4-491)	255.5 (4-966)	545 (36-39,000)	65.5 (3-2733)	118.5 (0-6410)	86 (1-7263)
Associate chief physician	14	56.86 (49.6)	314.9 (245.5)	103 (23-1416)	15 (1237)	18 (0-399)	8.5 (0-892)
Attending/registered physician	9	113.67 (55.0)	705.33 (671.4)	300 (46-206,000)	54 (5-13,000)	20 (0-9522)	282 (1-12,000)
P value		.03	.41	.04	.01	.008	.008
Content							
Disease knowledge	50	66.5 (15-210)	73.1 (13.98)	699.5 (23-20,600)	79.5 (1-13,000)	164 (0-9522)	204.5 (0-12,000)
Outpatient cases	19	66 (4-491)	64.74 (13.00)	226 (43-25,000)	16 (3-2695)	35 (1-2149)	19 (0-2007)
Surgical cases	11	26.73 (19.43)	46.36 (15.60)	169 (29-9642)	22 (3-769)	20 (0-137)	12 (1-275)
Surgical techniques	4	86.75 (38.95)	63.75 (35.80)	7740 (54-39,000)	482 (15-2733)	523 (24-1730)	693 (34-1783)
News/advertising	4	126 (134.16)	64.75 (23.08)	303.5 (67-399)	50 (12-202)	5 (0-31)	8.5 (0-343)
P value		.003	<.001	.03	.04	<.001	<.001

[a]Values are expressed as mean (SD) for continuous variables and median (range) for ordinal and discrete variables.

Initially, we obtained 100 videos. Of these, 84 were uploaded by physicians certified by TikTok, 13 by individual users, 2 by a news agency (online media), and 1 by an organization (hospital public account). After excluding videos that did not meet the inclusion criteria, 88 videos remained: 81 were uploaded by physicians, 4 by individual users, 2 by news agencies (online media), and 1 by an organization (hospital public account; Figure 1A). Among the physicians who uploaded videos, 39 were cardiac surgeons, 37 were cardiologists, 3 were sonographers, 1 was a radiologist, and 1 was an anorectal surgeon Figure 1B). Of the 81 videos uploaded by physicians, 58 were uploaded by chief physicians, 14 by associate chief physicians, 7 by attending physicians, and 2 by residents (Figure 1C). The videos were categorized by content (Figure 1D), with the most common topic being disease knowledge (n=50), followed by outpatient cases (n=19), surgical cases (n=11), surgical techniques (n=4), and news/advertising (n=4).

XSL•FO
RenderX

Figure 1. Percentage of videos on mitral valve regurgitation from different sources and with different contents on TikTok. (A) Video distribution based on uploaders. (B) Physician distribution. (C) Physicians' professional and technical titles. (D) Video distribution based on content.

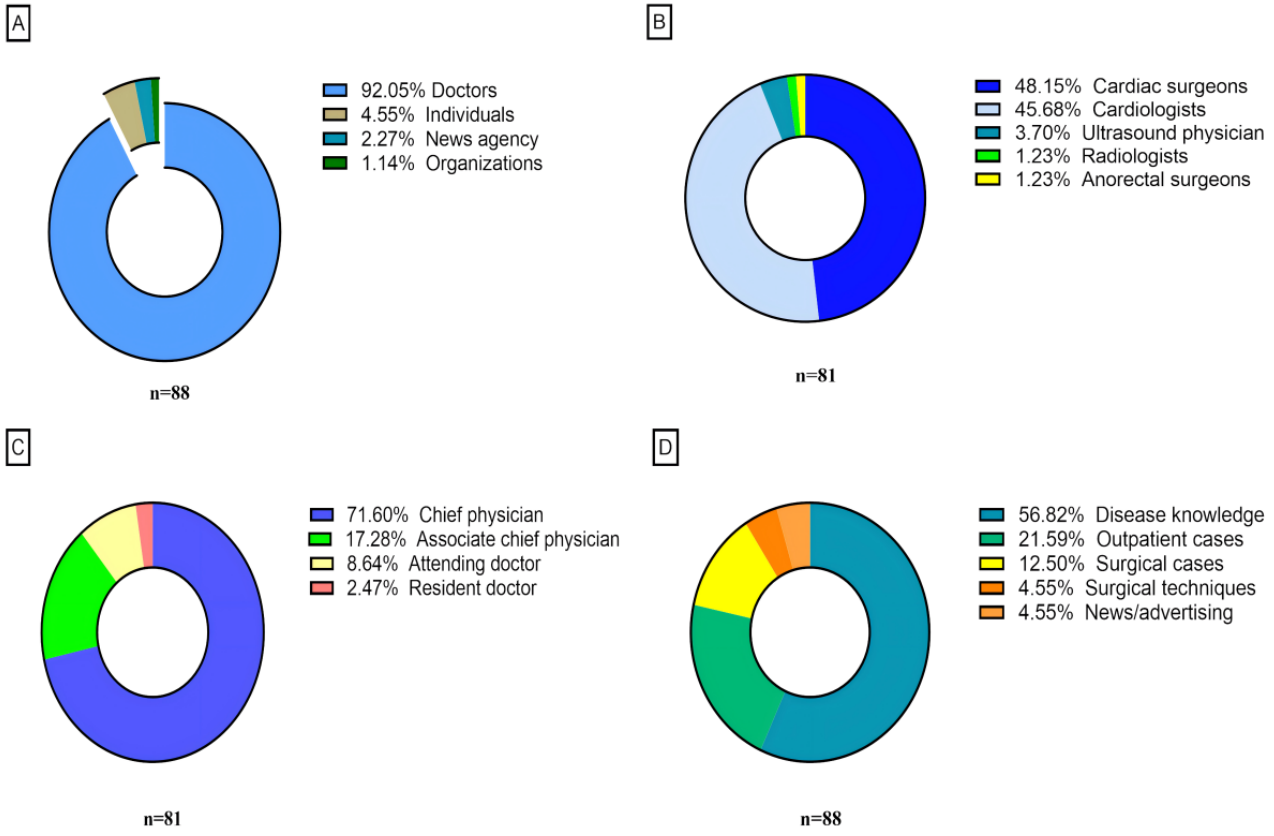

Comparison of General Data

Table 2 provides a detailed account of the intergroup comparisons based on uploader category, physician type, academic title, and video content. In the uploader category, only video length showed a statistically significant difference, with videos from news agencies/organizations being significantly longer than those from other groups (*P*=.049). In terms of physician technical level, pairwise comparison analysis revealed significant differences in video length (chief physician vs associate chief physician, *P*=.02), number of likes (chief physician vs associate chief physician, *P*=.003), number of comments (chief physician vs associate chief physician, *P*=.01), number of favorites (chief physician vs associate chief physician, *P*=.006), and number of shares (chief physician vs associate chief physician, *P*=.009; attending/registered physician vs associate chief physician, *P*=.047). The number of upload days did not show any significant differences (*P*=.41). When classified by video content, statistical analysis revealed differences in all parameters among the groups (Table 2). Pairwise comparisons indicated significant differences in video duration (disease knowledge vs outpatient cases, *P*=.01), video length (disease knowledge vs outpatient cases, *P*=.001; outpatient cases vs surgical cases, *P*=.01), number of favorites (disease knowledge vs surgical cases, *P*=.007), and number of shares (disease knowledge vs outpatient cases, *P*=.006; disease knowledge vs surgical cases, *P*=.007). Because of the small sample size, pairwise comparisons for the number of likes and comments were not feasible. No significant statistical differences were observed in other classifications (Table 2).

Video Quality and Reliability Assessments

We evaluated the videos using quantitative scoring tools such as mDISCERN, JAMA benchmark criteria, and GQS. Most videos (87/88) received a JAMA rating of 2, while 1 video received a rating of 1. Given that the JAMA criteria (range 0-4) were unable to effectively classify and assess video quality, they were excluded from the analyses of video quality and correlations.

The videos had a mean mDISCERN score of 3.06 (range 3-4) and a mean GQS of 2.75 (range 1-5), indicating that the TikTok videos exhibited fair quality and reliability. The PEMAT understandability score was 70.14 (range 16.67-100), and the PEMAT actionability score was 54.54 (range 0-100; Tables 1 and 3). This suggests that the TikTok videos were generally easy to understand but lacked effective implementation of recommendations, often providing either no suggestions or only general advice. The reviewers demonstrated good agreement on the mDISCERN scores (ICC 0.830, 95% CI 0.752–0.885), GQS scores (ICC 0.808, 95% CI 0.716–0.872), PEMAT understandability scores (ICC 0.966, 95% CI 0.946–0.978), and PEMAT actionability scores (ICC 0.829, 95% CI 0.748–0.885).

XSL•FO
RenderX

Table 3. The Global Quality Scores, mDISCERN[a] scores, and PEMAT[b] scores of the videos.

Scale, score	Values (N=88), n (%)
mDISCERN	
1	0 (0)
2	0 (0)
3	83 (94)
4	5 (6)
5	0 (0)
Global Quality Score	
1	2 (2)
2	31 (35)
3	46 (52)
4	5 (6)
5	4 (5)
PEMAT—Understandability score	
0-60	26 (30)
61-80	32 (36)
81-100	30 (34)
PEMAT—Actionability score	
0	30 (34)
33.33-66.67	15 (17)
100	43 (49)

[a]mDISCERN: modified DISCERN.

[b]PEMAT: Patient Education Materials Assessment Tool.

Subgroup Analysis

We classified the videos based on uploaders, physician type, title, and content, and performed statistical analyses on the characteristic parameters of each video group. Subgroup comparisons are detailed in Table 4 and Figures 2-5. Differences were observed in GQS scores and PEMAT understandability scores among the subgroups. News agencies/organizations and doctors had higher GQS scores compared with individuals (news agencies/organizations vs individuals, P=.001; doctors vs individuals, P=.03). News agencies/organizations had higher PEMAT understandability scores compared with individuals (P=.01). No significant differences were found between subgroups based on physician type or title (Table 4). Analysis by video content classification revealed differences in GQS scores, PEMAT understandability scores, and PEMAT actionability scores. Videos focused on disease knowledge had higher GQS scores, PEMAT understandability scores, and PEMAT actionability scores compared with videos on surgical cases (P<.001 for all comparisons). Additionally, PEMAT actionability scores were higher for outpatient cases than for surgical cases (P<.001), and videos on surgical techniques had lower PEMAT actionability scores compared with videos on disease knowledge (P=.04).

XSL•FO
RenderX

Table 4. Quality assessment of videos based on uploaders and content[a].

Characteristic	n	mDISCERN[b]	GQS[c]	PEMAT[d] understandability score (%)	PEMAT actionability score (%)
Uploaders					
Doctors	81	3 (3-4)	3 (1-5)	75 (25-100)	66.67 (0-100)
Individuals	4	3 (3-3)	2 (1-2)	45.83 (16.67-66.67)	0 (0-33.33)
News agency/organizations	3	3 (3-4)	5 (3-5)	83.33 (83.33-91.67)	100 (0-100)
P value		.10	.002	.01	.12
Doctors					
Cardiac surgeons	39	3 (3-4)	3 (1-5)	75 (25-100)	100 (0-100)
Cardiologists	37	3 (3-4)	3 (2-3)	66.67 (41.67-100)	66.67 (0-100)
Other health care professionals	5	3 (3-4)	3 (2-3)	75 (58.33-91.67)	100 (0-100)
P value		.25	.11	.43	.77
Professional titles					
Chief physician	58	3 (3-4)	3 (1-5)	75 (41.67)	83.34 (0-100)
Associate chief physician	14	3 (3-3)	2.5 (2-4)	70.84 (25-91.67)	16.67 (0-100)
Attending/resident doctor	9	3 (3-3)	3 (2-3)	66.67 (58.33-91.67)	100 (0-100)
P value		.44	.52	.69	.01
Content					
Disease knowledge	50	3 (3-4)	3 (2-5)	76.09 (13.98)	100 (0-100)
Outpatient cases	19	3 (2-4)	3 (1-5)	67.82 (13.04)	70.18 (39.90)
Surgical cases	11	3 (3-3)	2 (1-2)	49.24 (15.57)	0 (0-33.33)
Surgical techniques	4	3 (3-3)	2 (2-2)	66.67 (36)	0 (0-0)
News/advertising	4	3 (2-4)	3.5 (2-5)	67.63 (23.72)	0 (0-100)
P value		.98	<.001	<.001	<.001

[a]Values are expressed as mean (SD) for continuous variables and median (range) for ordinal and discrete variables.

[b]mDISCERN: modified DISCERN.

[c]GQS: Global Quality Score.

[d]PEMAT: Patient Education Materials Assessment Tool.

Figure 2. Quality assessment of videos based on uploaders. (A) Comparison of mDICERN scores among different uploaders. (B) Comparison of GQS between different uploaders. (C) Comparison of PEMAT-V scores among different uploaders. (D) Comparison of PEMAT-A scores among different uploaders. GQS: Global Quality Score; mDISCERN: modified DISCERN; PEMAT-A: Patient Education Materials Assessment Tool for Audio Content; PEMAT-V: Patient Education Materials Assessment Tool for Visual Content.

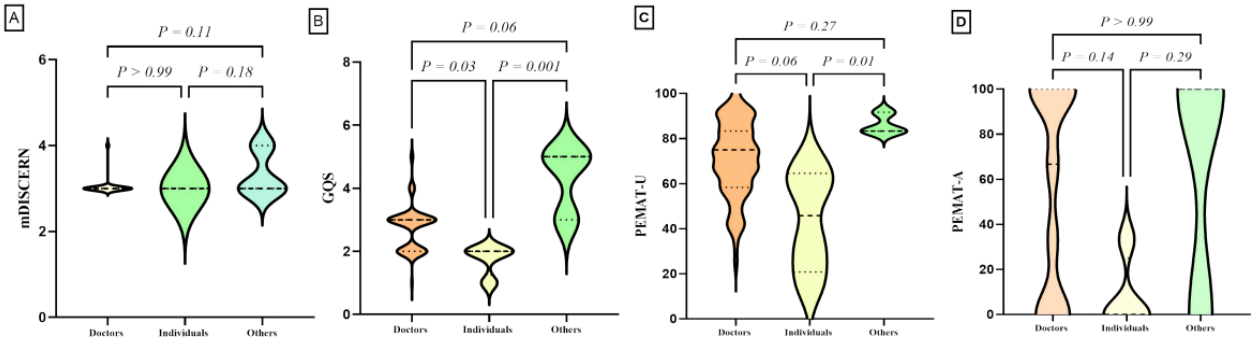

XSL•FO
RenderX

Figure 3. Quality assessment of videos based on doctors. (A) Comparison of mDICERN scores among different doctors. (B) Comparison of GQS between different doctors. (C) Comparison of PEMAT-V scores among different doctors. (D) Comparison of PEMAT-A scores among different doctors. GQS: Global Quality Score; mDISCERN: modified DISCERN; PEMAT-A: Patient Education Materials Assessment Tool for Audio Content; PEMAT-V: Patient Education Materials Assessment Tool for Visual Content.

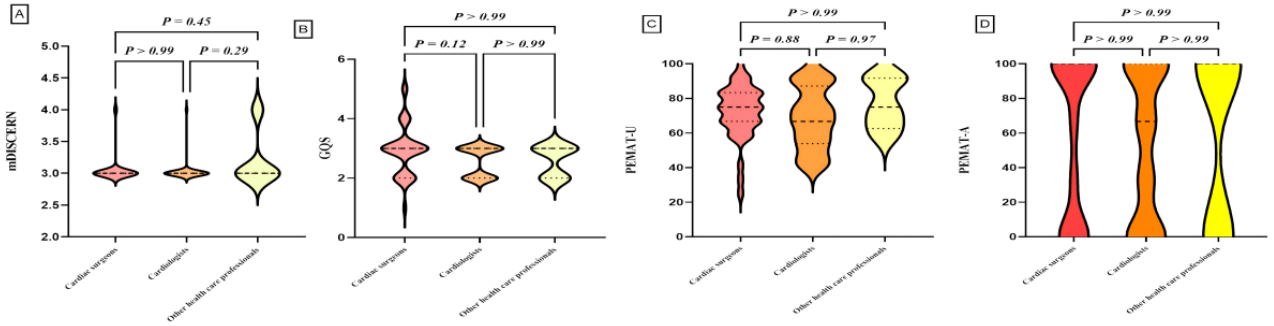

Figure 4. Quality assessment of videos based on professional titles. (A) Comparison of mDICERN scores among different professional titles. (B) Comparison of GQS between different professional titles. (C) Comparison of PEMAT-V scores among different professional titles. (D) Comparison of PEMAT-A scores among different professional titles. GQS: Global Quality Score; mDISCERN: modified DISCERN; PEMAT-A: Patient Education Materials Assessment Tool for Audio Content; PEMAT-V: Patient Education Materials Assessment Tool for Visual Content.

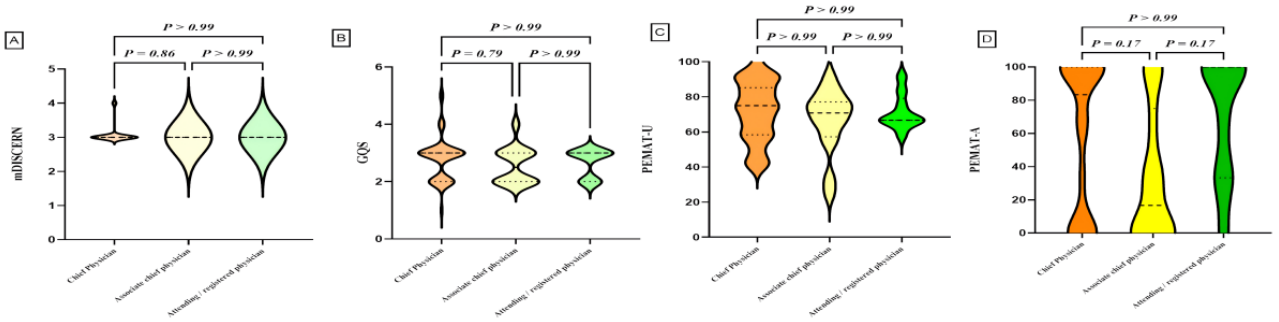

Figure 5. Spearman correlation analysis between video variables and the scores. GQS: Global Quality Score; mDISCERN: modified DISCERN; PEMAT-A: Patient Education Materials Assessment Tool for Audio Content; PEMAT-V: Patient Education Materials Assessment Tool for Visual Content.

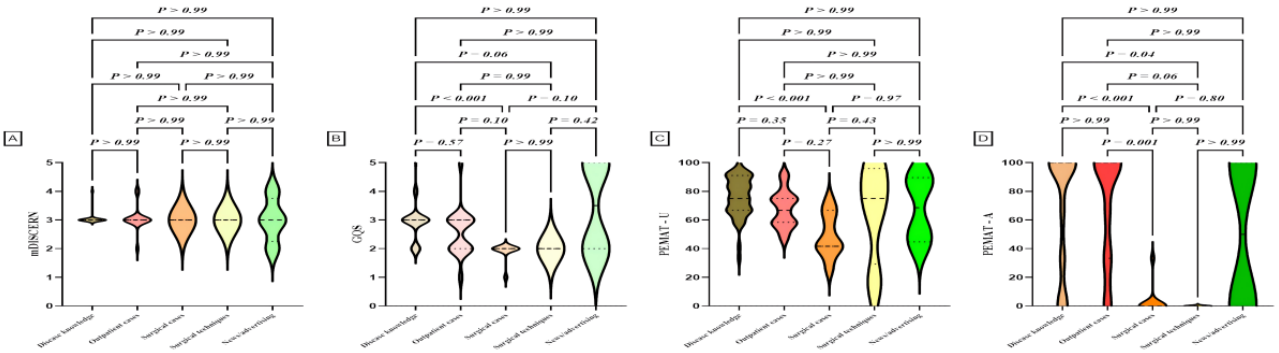

Correlation and Stepwise Regression Analysis

To explore correlations among various video parameters, we conducted a Spearman correlation analysis due to the presence of categorical variables and nonnormally distributed data (Figure 6). The analysis revealed strong correlations between video length, duration on TikTok, thumbs up, comments, favorites, and shares. Specifically, significant correlations were found between thumbs up and comments (r=0.92, P<.001), thumbs up and favorites (r=0.89, P<.001), thumbs up and shares (r=0.87, P<.001), comments and favorites (r=0.81, P<.001), comments and shares (r=0.87, P<.001), and favorites and shares (r=0.83, P<.001). However, the correlation between these video parameters and the evaluation tools is relatively weak (0.5>r>0.3). Specifically, the correlations are as follows: video length and GQS (r=0.48, P<.001), thumbs up and GQS (r=0.30, P<.001), and shares and GQS (r=0.37, P<.001). Additionally, we observed negative correlations between several video parameters: uploaders and physicians (r=–0.51, P<.01), uploaders and titles (r=–0.55, P<.01), physicians and content (r=–0.24, P<.05), content and comments (r=–0.23, P<.05), content and favorites (r=–0.28, P<.05), content and shares (r=–0.37, P<.01), content and GQS (r=–0.41, P<.01), content and PEMAT-Understandability (r=–0.26, P<.05), and content and PEMAT-Actionability (r=–0.34, P<.01).

The stepwise regression analysis produced 3 models, each incorporating additional predictors progressively (Tables 5 and 6). The final model (model 3), which included "length,"

XSL•FO
RenderX

"content," and "physicians" as predictors, demonstrated the strongest predictive power with an R^2 of 0.501, indicating that these variables explain 50.1% of the variance in GQS. The adjusted R^2 value was 0.482, and the model was statistically significant with an $F_{1,77}$ ratio of 8.663 (P=.004). The Durbin-Watson statistic for model 3 was 1.960, indicating no significant autocorrelation in the residuals. The collinearity diagnostics revealed no multicollinearity issues, with tolerance values exceeding 0.9 and VIFs below 2.0 for all predictors in model 3. The predictive equation for GQS, based on model 3, is as follows: GQS = 3.230 − 0.294 × content – 0.274 × physicians + 0.005 × length.

Figure 6. Quality assessment of videos based on contents. (A) Comparison of mDICERN scores among different contents. (B) Comparison of GQS between different contents. (C) Comparison of PEMAT-U scores among different contents. (D) Comparison of PEMAT-A scores among different contents. GQS: Global Quality Score; mDISCERN: modified DISCERN; PEMAT-A: Patient Education Materials Assessment Tool for Audio Content; PEMAT-V: Patient Education Materials Assessment Tool for Visual Content.

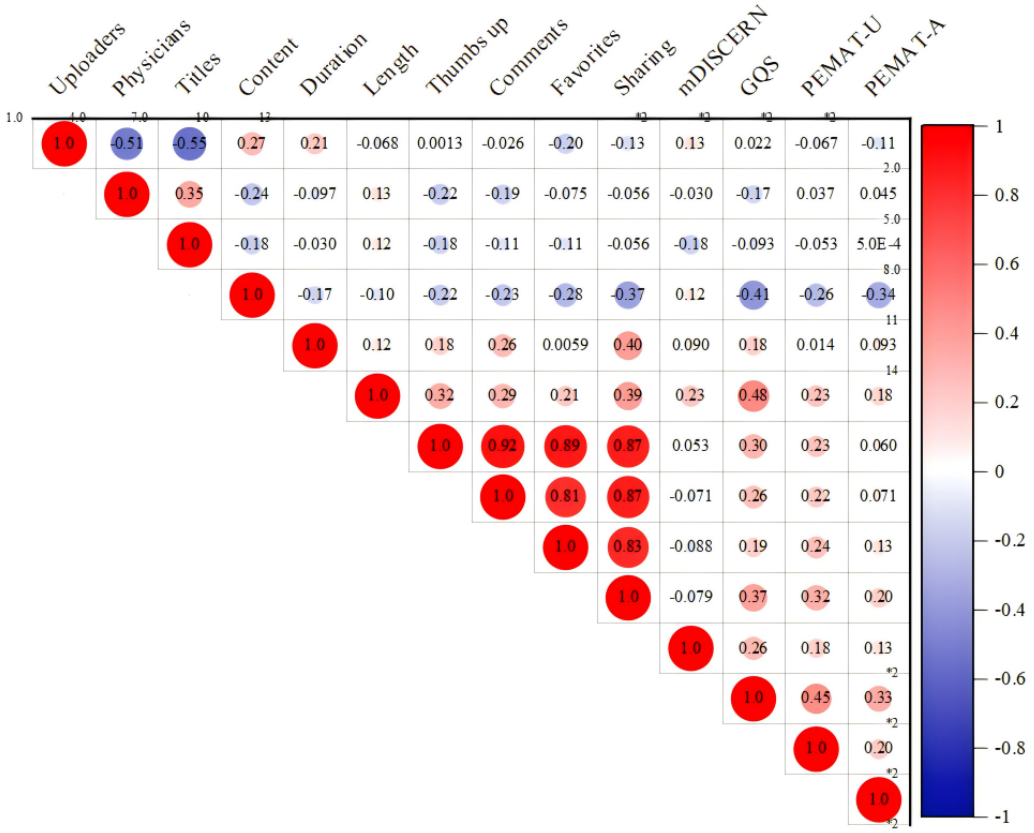

Table 5. Stepwise regression analysis summary.

Model	Predictors included	R^2	Adjusted R^2	SE of the estimate	F change (df)	P value	Durbin-Watson	ANOVA	
								F test (df)	P value
1	Constant and length	0.285	0.276	0.59888	31.563 (1, 79)	<.001	2.194	31.563 (1, 79)	<.001
2	Constant, length, and content	0.445	0.431	0.53112	22.444 (1, 78)	<.001	2.004	31.287 (1, 78)	<.001
3	Constant, length, content, and physicians	0.501	0.482	0.50681	8.663 (1, 77)	.004	1.96	25.795 (1, 77)	<.001

XSL•FO
RenderX

Table 6. Stepwise regression coefficients, statistical significance, and collinearity assessment.

Model and predictors	Unstandardized coefficients		Standardized coefficients (β)	*t* value (*df*)	*P* value	95% CI for B	Collinearity statistics	
	B	SE					Tolerance	VIF[a]
1								
Constant	2.241	0.103	N/A[b]	21.849 (79)	<.001	2.036 to 2.445	N/A	N/A
Length	0.006	0.001	0.534	5.618 (79)	<.001	0.004 to 0.008	1.000	1.000
2								
Constant	2.779	0.146	N/A	19.087 (78)	<.001	2.489 to 3.069	N/A	N/A
Length	0.005	0.001	0.464	5.412 (78)	<.001	0.003 to 0.007	0.970	1.031
Content	–0.286	0.060	–0.406	–4.738 (78)	<.001	–0.407 to –0.166	0.970	1.031
3								
Constant	3.230	0.207	N/A	15.618 (77)	<.001	2.818 to 3.642	N/A	N/A
Length	0.005	0.001	0.457	5.586 (77)	<.001	0.003 to 0.006	0.969	1.032
Content	–0.294	0.058	–0.416	–5.089 (77)	<.001	–0.409 to –0.179	0.968	1.033
Physicians	–0.274	0.093	–0.237	–2.943 (77)	.004	–0.459 to –0.089	0.998	1.002

[a]VIF: variance inflation factor.

[b]N/A: not applicable.

Discussion

Principal Findings

This cross-sectional study revealed that most MR-related videos on TikTok were uploaded by certified physicians, predominantly cardiac surgeons or cardiologists. Notably, all physician identities were verified by TikTok to ensure professionalism and scientific accuracy. The video content quality, assessed using the mDISCERN tool, yielded satisfactory scores; however, the GQS scores did not meet the expected standards. The evaluation of the educational impact of the videos showed that a significant majority achieved PEMAT understandability scores above the threshold, indicating high levels of comprehension and acceptance among viewers. However, there was considerable variation in the PEMAT actionability scores, with approximately half of the videos falling below satisfactory levels. This suggests that many videos have limited practical applicability. Our analyses yielded several noteworthy observations.

Mitral Regurgitation Video Characteristics and Audience Interaction Analysis

Our study found that the 88 videos received a total of 476,511 thumbs up, 33,123 comments, 39,781 favorites, and 49,314 shares. This high level of viewer engagement highlights significant interest in MR content and a notable degree of interactivity, which contrasts with findings from previous studies [27,28]. Most videos were uploaded by cardiologists certified by TikTok, primarily cardiac surgeons and cardiologists, with some contributions from the ultrasound department. This indicates a strong willingness among medical professionals to share their expertise on the TikTok platform. In addition to videos uploaded by doctors, there were contributions from news media, nonprofit medical official accounts, and individuals. The content spanned various topics, including disease knowledge, outpatient cases, surgical cases, surgical techniques, and personal experiences. This diversity addresses the needs of different audiences for medical information and has a positive impact on public health awareness.

Furthermore, our results suggest that video length is a significant distinguishing factor, consistent with previous findings [12]. Analysis by different classification criteria revealed statistical differences among groups (Table 2). Specifically, videos uploaded by chief physicians and those focused on disease knowledge were notably longer and received more thumbs up, comments, favorites, and shares. A detailed examination of the videos revealed that the discrepancy may be attributed to the public's greater trust in physicians with higher technical ranks and their interest in understanding disease knowledge, often seeking assistance through consultation. Notably, 1 video by an attending physician achieved exceptionally high engagement, with 206,000 thumbs up, 13,000 comments, and 12,000 shares. This video stood out because the physician used colloquial language to explain disease knowledge in an accessible and sincere manner, offering valuable insights for content creators on how to produce higher-quality videos.

Video Ratings and Quality Evaluation

Our study used various assessment tools to evaluate video quality. First, we used the modified DISCERN tool to gauge the reliability of the videos. The mDISCERN results indicated that the videos had moderate credibility—acceptable but not entirely satisfactory. Subgroup analysis did not reveal significant differences among the groups. Upon closer examination, we observed that video creators often neglected to provide sources of information, and TikTok's lack of a mandatory review mechanism meant that the question regarding additional sources for patient reference (Are additional sources of information

XSL•FO
RenderX

listed for patient reference?”) was not scored. This highlights the importance for content creators to include clear and reliable sources to enhance the credibility of their content.

Second, we assessed the overall quality of the video content using the GQS. The average GQS score was 2.75 (range 0-5), which is near the threshold for moderate quality. This finding aligns with Collà Ruvolo et al [28] but is notably lower than the results reported by Morra S et al [27]. Subgroup analysis revealed that videos from doctors and news agencies/organizations had higher GQS scores compared with those from individuals (Figure 2B). Additionally, videos focusing on disease knowledge received higher GQS scores than those covering surgical cases (Figure 5B).

Lastly, we used the PEMAT-A/V tool to assess the educational significance of the videos. The PEMAT-A/V results showed that the understandability score was 70.14%, while the actionability score was 54.54%. According to Shoemaker et al [26], a PEMAT score exceeding 70% indicates that the content is sufficiently understandable and actionable. Our findings suggest that while the videos are generally understandable, they fall short in terms of actionability. This observation is consistent with the findings of Kanner et al [29], who reported similar issues with TikTok and YouTube videos on overactive bladder. Subgroup analysis revealed that videos from media organizations scored higher on the PEMAT-A/V understandability and actionability compared with those from individuals (Figure 2C). Additionally, videos focused on disease knowledge had higher PEMAT-A/V understandability scores compared with those on surgical cases. Both disease knowledge and outpatient cases achieved higher scores than surgical cases in terms of understandability and actionability.

Integrating these assessment outcomes, we found that videos produced by doctors or news agencies/organizations with a medical background, particularly those focusing on disease knowledge, tend to be of higher quality compared with others. Patients interested in MR or similar topics are advised to prioritize watching these videos for their informative and reliable content.

Correlation and Stepwise Regression Analysis Between Video Quality and Video Characteristics

We conducted an in-depth analysis of MR video content on TikTok. Using Spearman correlation analysis, we assessed the relationship between video characteristics and quality metrics. Additionally, we used stepwise regression analysis to evaluate the predictive capacity of video parameters on quality scores—an innovative approach not previously explored in similar studies. Our correlation and stepwise regression analyses have provided a comprehensive view of the factors influencing video quality in the context of MR content on TikTok. However, interpreting our results, especially the observed negative correlations, should be done with an understanding of the statistical methods used.

The strong positive correlations observed among video length, duration, and engagement metrics—such as thumbs up, comments, favorites, and shares—highlight a significant interrelationship between these variables. This finding supports existing literature, which suggests that content with higher audience engagement is more likely to be widely disseminated, a concept explored by Weng et al [30] in the context of social media network effects. The weaker correlation between video characteristics and GQS scores suggests that viewer engagement, while significant, does not necessarily reflect the intrinsic quality of the content. This observation aligns with Berger and Milkman [31], who found that emotional appeal and content novelty are key drivers of engagement, regardless of the content’s educational or informational value. Additionally, the negative correlations between certain video parameters, such as the uploader’s professional role and content type, may stem from the statistical methodology used rather than indicating a true inverse relationship.

The stepwise regression analysis highlighted “length,” “content,” and “physicians” as significant predictors of GQS scores, with the final model demonstrating substantial predictive strength. The model suggests that longer videos and those featuring physicians are associated with higher GQS scores, while certain content types may be linked to lower scores. However, the negative coefficients for “content” and “physicians” should be interpreted with caution, as they may not directly indicate causal relationships.

Practical Significance

Mild MR often goes unnoticed and is frequently detected through ultrasound scans. In the digital age, individuals increasingly rely on video platforms to seek health information and share experiences. The Chinese government acknowledges the value of these platforms in health education and encourages medical professionals to actively disseminate health science knowledge [32]. This initiative is especially pertinent on TikTok, a leading short-video platform in China that is brimming with disease-related content. Our study reveals that cardiologists’ contributions to TikTok offer professional and scientifically sound content on MR, providing valuable guidance to patients and potentially reducing unnecessary medical consultations. However, challenges persist, including limitations in TikTok’s recommendation algorithms and constraints on video length, which may restrict the depth of information conveyed. To establish a reliable source of high-quality medical content, it is crucial for TikTok to enhance its certification processes for content creators and implement robust content review mechanisms. These measures will ensure that viewers receive accurate and comprehensive health information that meets their needs and supports informed decision-making.

Strengths and Limitations

Our study provides robust insights into health information dissemination on TikTok, highlighting the importance of content validation by certified medical professionals to ensure the reliability of health messages. By comprehensively measuring audience engagement through metrics such as likes, comments, favorites, and shares, we gain a multidimensional understanding of video impact. Additionally, the innovative use of stepwise regression analysis to predict video quality based on specific characteristics introduces a novel approach to digital health communication. The methodological rigor, demonstrated through the use of Spearman correlation and regression models,

XSL•FO
RenderX

strengthens the study's findings. Additionally, the focus on practical implications provides valuable guidance for both content creators and viewers.

Despite these strengths, the study has several limitations. One limitation is the potential bias in the statistical methodology, which may affect the interpretation of negative correlations. The content analyzed might not fully represent the entire range of MR topics, potentially limiting the applicability of the findings. Additionally, the study does not account for the influence of TikTok's recommendation algorithms, which could significantly impact video visibility and engagement levels. The diverse demographics of TikTok users and their varying interactions with health content are not fully accounted for, which could impact the study's conclusions. As a correlational study, it does not infer causality, and the findings may not be generalizable to other platforms or contexts. Furthermore, the study's scope is limited to a specific period, which may not capture the dynamic nature of content evolution and user engagement on social media platforms over time. This temporal limitation may affect the long-term applicability and stability of the findings.

Conclusions

Currently, most MR-related videos on TikTok are uploaded by certified physicians, which ensures their professional and scientific reliability. However, the overall video-quality scores were suboptimal. While many videos provided educational value, the guidance offered was insufficient, and video production quality needs improvement. Notably, video prevalence was only weakly correlated with quality. The predictive equation for GQS developed from our analysis provides valuable insights but should be applied cautiously beyond the context of this study. This suggests that creators should focus on enhancing both the content and presentation of their videos to improve the quality of health information shared on social media. Additionally, TikTok should bolster its review mechanisms to elevate the standard of medical content, thereby benefiting users and strengthening the integrity of online health communication.

Disclosure of Artificial Intelligence Tool Use

During the period of our manuscript creation, we did not use ChatGPT as we did not have access to it in our country. However, we utilized YoDayo AI (NetEase), which is a language translation software, to enhance the overall readability and clarity of the manuscript. Following its use, we conducted a comprehensive review and editing of all content, taking full responsibility for the published material.

Data Availability

The data for this study, derived from the TikTok platform, have been anonymized to protect privacy, with URLs and titles retained for research integrity. Because of privacy concerns, these data are not publicly accessible. However, we are committed to sharing the data upon request for legitimate research purposes, in accordance with privacy protection principles and data sharing policies. Interested researchers should contact the corresponding author for access.

Authors' Contributions

NC conceived and designed the study. NC and YC were responsible for reviewing and scoring the videos. YL and XC collected and analyzed the data. NC wrote the original draft. SF and QS contributed equally to the research and manuscript preparation, and are considered co-corresponding authors. All the authors contributed to manuscript writing and editing and approved the final draft for submission.

Conflicts of Interest

None declared.

References

1. Iung B, Baron G, Butchart EG, Delahaye F, Gohlke-Bärwolf C, Levang OW, et al. A prospective survey of patients with valvular heart disease in Europe: The Euro Heart Survey on Valvular Heart Disease. Eur Heart J 2003 Jul;24(13):1231-1243. [doi: 10.1016/s0195-668x(03)00201-x] [Medline: 12831818]
2. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. Lancet 2006 Sep 16;368(9540):1005-1011. [doi: 10.1016/S0140-6736(06)69208-8] [Medline: 16980116]
3. Nie JY, Jing-yu LF, Chen Z, Wang X, Guo M, Tian Y, et al. Prevalence and risk factors of mitral regurgitation in the population aged≥35 years. Chinese Circulation Journal 2017 May;32:5-9 [FREE Full text] [doi: 10.3969/j.issn.1000-3614.2017.05.013]
4. Qi X, Xu H, Liu Q, Zhu QF. Analysis of diagnosis and treatment status of hospitalized patients with degenerative heart valve disease in China. Chinese Circulation Journal 2019 Aug 24;34(8):771-776 [FREE Full text] [doi: 10.3969/j.issn.1000-3614.2019.08.007]

XSL•FO
RenderX

5. Hu P, Liu X, Liang J, Zhu Q, Pu C, Tang M, et al. A hospital-based survey of patients with severe valvular heart disease in China. Int J Cardiol 2017 Mar 15;231:244-247. [doi: 10.1016/j.ijcard.2016.11.301] [Medline: 27908608]
6. Lin XZ, Zhu QF, Liu XB, Wang JA. Current status and progress of transcatheter mitral regurgitation interventional therapy. Zhonghua Xin Xue Guan Bing Za Zhi 2021 May 24;49(5):425-431. [doi: 10.3760/cma.j.cn112148-20210211-00148] [Medline: 34034376]
7. Iung B, Vahanian A. Epidemiology of acquired valvular heart disease. Can J Cardiol 2014 Sep;30(9):962-970. [doi: 10.1016/j.cjca.2014.03.022] [Medline: 24986049]
8. Coffey S, Cairns BJ, Iung B. The modern epidemiology of heart valve disease. Heart 2016 Jan;102(1):75-85. [doi: 10.1136/heartjnl-2014-307020] [Medline: 26541169]
9. Iung B, Vahanian A. Epidemiology of valvular heart disease in the adult. Nat Rev Cardiol 2011 Mar;8(3):162-172. [doi: 10.1038/nrcardio.2010.202] [Medline: 21263455]
10. Wang K, Geng B, Shen Q, Wang Y, Shi J, Dong N. Global, regional, and national incidence, mortality, and disability-adjusted life years of non-rheumatic valvular heart disease and trend analysis from 1990 to 2019: Results from the Global Burden of Disease study 2019. Asian Cardiovasc Thorac Ann 2023 Oct;31(8):706-722. [doi: 10.1177/02184923231200695] [Medline: 37674443]
11. Zhu C, Xu X, Zhang W, Chen J, Evans R. How health communication via Tik Tok makes a difference: a content analysis of Tik Tok accounts run by Chinese provincial health committees. Int J Environ Res Public Health 2019 Dec 27;17(1):192. [doi: 10.3390/ijerph17010192] [Medline: 31892122]
12. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
13. Song S, Zhao YC, Yao X, Ba Z, Zhu Q. Serious information in hedonic social applications: affordances, self-determination and health information adoption in TikTok. JD 2021 Nov 19;78(4):890-911. [doi: 10.1108/jd-08-2021-0158]
14. Ostrovsky AM, Chen JR. TikTok and its role in COVID-19 information propagation. J Adolesc Health 2020 Nov;67(5):730 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.07.039] [Medline: 32873499]
15. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021 Dec 20;23(12):e28318 [FREE Full text] [doi: 10.2196/28318] [Medline: 34931996]
16. Sun F, Zheng S, Wu J. Quality of information in gallstone disease videos on TikTok: cross-sectional study. J Med Internet Res 2023 Feb 08;25:e39162 [FREE Full text] [doi: 10.2196/39162] [Medline: 36753307]
17. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021 Sep 01;23(9):e30409 [FREE Full text] [doi: 10.2196/30409] [Medline: 34468327]
18. Mueller SM, Hongler VNS, Jungo P, Cajacob L, Schwegler S, Steveling EH, et al. Fiction, falsehoods, and few facts: cross-sectional study on the content-related quality of atopic eczema-related videos on YouTube. J Med Internet Res 2020 Apr 24;22(4):e15599 [FREE Full text] [doi: 10.2196/15599] [Medline: 32329744]
19. Ferhatoglu MF, Kartal A, Ekici U, Gurkan A. Evaluation of the reliability, utility, and quality of the information in sleeve gastrectomy videos shared on open access video sharing platform YouTube. Obes Surg 2019 May;29(5):1477-1484. [doi: 10.1007/s11695-019-03738-2] [Medline: 30706318]
20. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the internet: caveant lector et viewor--let the reader and viewer beware. JAMA 1997 Apr 16;277(15):1244-1245. [Medline: 9103351]
21. Demirtas MS, Alici N. The reliability and quality of YouTube videos as a source of breath holding spell. Ital J Pediatr 2024 Jan 18;50(1):8 [FREE Full text] [doi: 10.1186/s13052-023-01570-0] [Medline: 38238792]
22. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis--a wakeup call? J Rheumatol 2012 May;39(5):899-903. [doi: 10.3899/jrheum.111114] [Medline: 22467934]
23. Delli K, Livas C, Vissink A, Spijkervet F. Is YouTube useful as a source of information for Sjögren's syndrome? Oral Dis 2016 Apr;22(3):196-201. [doi: 10.1111/odi.12404] [Medline: 26602325]
24. Liu H, Peng J, Li L, Deng A, Huang X, Yin G, et al. Assessment of the reliability and quality of breast cancer related videos on TikTok and Bilibili: cross-sectional study in China. Front Public Health 2023;11:1296386 [FREE Full text] [doi: 10.3389/fpubh.2023.1296386] [Medline: 38317686]
25. Castillo J, Wassef C, Wassef A, Stormes K, Berry AE. YouTube as a source of patient information for prenatal repair of myelomeningocele. Am J Perinatol 2021 Jan;38(2):140-144. [doi: 10.1055/s-0039-1694786] [Medline: 31430813]
26. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
27. Morra S, Collà Ruvolo C, Napolitano L, La Rocca R, Celentano G, Califano G, et al. YouTube as a source of information on bladder pain syndrome: a contemporary analysis. Neurourol Urodyn 2022 Jan;41(1):237-245. [doi: 10.1002/nau.24802] [Medline: 34559920]

XSL•FO
RenderX

28. Collà Ruvolo C, Califano G, Tuccillo A, Tolentino S, Cancelliere E, Di Bello F, et al. "YouTube™ as a source of information on placenta accreta: a quality analysis". Eur J Obstet Gynecol Reprod Biol 2022 May;272:82-87. [doi: 10.1016/j.ejogrb.2022.03.015] [Medline: 35290877]
29. Kanner J, Waghmarae S, Nemirovsky A, Wang S, Loeb S, Malik R. TikTok and YouTube videos on overactive bladder exhibit poor quality and diversity. Urol Pract 2023 Sep;10(5):493-500. [doi: 10.1097/UPJ.0000000000000423] [Medline: 37347790]
30. Weng L, Menczer F, Ahn Y. Virality prediction and community structure in social networks. Sci Rep 2013;3:2522. [doi: 10.1038/srep02522] [Medline: 23982106]
31. Berger J, Milkman KL. What makes online content viral? Journal of Marketing Research 2012 Apr 01;49(2):192-205. [doi: 10.1509/jmr.10.0353]
32. Chinese Government. China’s 14th five-year plan. Central People's Government of the People's Republic of China. 2022. URL: https://www.gov.cn/zhengce/content/2022-05/20/content_5691424.htm [accessed 2022-04-27]

Abbreviations

GQS: Global Quality Score
ICC: intraclass correlation coefficient
JAMA: The Journal of the American Medical Association
mDISCERN: modified DISCERN
MR: mitral valve regurgitation
PEMAT-A/V: Patient Education Materials Assessment Tool for Audio-Visual Content
VIF: variance inflation factor

Edited by A Mavragani; submitted 12.12.23; peer-reviewed by S Morra, MF Zuljevic; comments to author 30.04.24; revised version received 19.06.24; accepted 04.07.24; published 20.08.24.

Please cite as:
Cui N, Lu Y, Cao Y, Chen X, Fu S, Su Q
Quality Assessment of TikTok as a Source of Information About Mitral Valve Regurgitation in China: Cross-Sectional Study
J Med Internet Res 2024;26:e55403
URL: https://www.jmir.org/2024/1/e55403
doi:10.2196/55403
PMID:39163110

©Nannan Cui, Yuting Lu, Yelin Cao, Xiaofan Chen, Shuiqiao Fu, Qun Su. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Videos in Short-Video Sharing Platforms as Sources of Information on Colorectal Polyps: Cross-Sectional Content Analysis Study

Jia-Lun Guan[1], MD; Su-Hong Xia[1], PhD; Kai Zhao[1], PhD; Li-Na Feng[1], PhD; Ying-Ying Han[1], MD; Ji-Yan Li[1], MD; Jia-Zhi Liao[1*], MD, PhD; Pei-Yuan Li[1,2*], MD, PhD

[1]Department of Gastroenterology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]Department of Gastroenterology, Wenchang People's Hospital, Wenchang, China

[*]these authors contributed equally

Corresponding Author:
Pei-Yuan Li, MD, PhD
Department of Gastroenterology
Tongji Hospital, Tongji Medical College
Huazhong University of Science and Technology
1095 Jiefang Avenue
Wuhan, 430030
China
Phone: 86 27 83662831
Fax: 86 27 83662831
Email: pyli@tjh.tjmu.edu.cn

Abstract

Background: Short videos have demonstrated huge potential in disseminating health information in recent years. However, to our knowledge, no study has examined information about colorectal polyps on short-video sharing platforms.

Objective: This study aimed to analyze the content and quality of colorectal polyps-related videos on short-video sharing platforms.

Methods: The terms “肠息肉” (intestinal polyps) or “结肠息肉” (colonic polyps) or “直肠息肉” (rectal polyps) or “结直肠息肉” (colorectal polyps) or “大肠息肉” (polyps of large intestine) were used to search in TikTok (ByteDance), WeChat (Tencent Holdings Limited), and Xiaohongshu (Xingyin Information Technology Limited) between May 26 and June 8, 2024, and then the top 100 videos for each search term on different platforms were included and recorded. The *Journal of American Medical Association* (JAMA) score, the Global Quality Scale (GQS), the modified DISCERN, and the Patient Education Materials Assessment Tool (PEMAT) were used to evaluate the content and quality of selected videos by 2 independent researchers. SPSS (version 22.0; IBM Corp) and GraphPad Prism (version 9.0; Dotmatics) were used for analyzing the data. Descriptive statistics were generated, and the differences between groups were compared. Spearman correlation analysis was used to evaluate the relationship between quantitative variables.

Results: A total of 816 eligible videos were included for further analysis, which mainly conveyed disease-related knowledge (n=635, 77.8%). Most videos were uploaded by physicians (n=709, 86.9%). These videos had an average JAMA score of 2.0 (SD 0.6), GQS score of 2.5 (SD 0.8), modified DISCERN score of 2.5 (SD 0.8), understandability of 80.4% (SD 15.6%), and actionability of 42.2% (SD 36.1%). Videos uploaded by news agencies were of higher quality and received more likes and comments (all $P<.05$). The number of collections and shares of videos about posttreatment caveats were more than those for other content ($P=.03$ and $P=.006$). There was a positive correlation between the number of likes, comments, collections, and shares (all $P<.001$). The duration and the number of fans were positively correlated with the quality of videos (all $P<.05$).

Conclusions: There are numerous videos about colorectal polyps on short-video sharing platforms, but the reliability and quality of these videos are not good enough and need to be improved.

(J Med Internet Res 2024;26:e51655) doi:10.2196/51655

KEYWORDS

colorectal polyps; short videos; health information; quality assessment; reliability

XSL•FO
RenderX

Introduction

Colorectal cancer (CRC) is the third most common cancer and the second cause of cancer-related death worldwide [1]. Colorectal polyps are abnormal growths and protrusions on the colorectal surface, and adenomatous colorectal polyps are the precursors of the majority of CRC [2]. With the improvement of living conditions and the changes in diet habits, the incidences of colorectal polyps and CRC are gradually increasing [3]. Most colorectal polyps are asymptomatic, and some may quietly develop into malignancies, which are more likely to be ignored. The detection and treatment of colorectal polyps is a critical way to reduce the incidence of CRC and its subsequent morbidity and mortality [4]. At the same time, lifestyle adjustments and regular examination could prevent the occurrence or recurrence of colorectal polyps [5-7]. Hence, it is necessary for the population to receive appropriate health information and self-management education.

With the development of the internet and mobile phone communication, information propagates quickly along social networks. People are increasingly seeking and accessing health information and knowledge on mobile internet. Short videos increase the efficiency of people's scattered free time as well as the convenience of enjoyment, thus receiving wide popularity. In 2021, TikTok (ByteDance) had more than 1.6 billion monthly active users, and the number of monthly active users in WeChat (Tencent Holdings Limited) is as high as 1 billion [8,9]. At the same time, Xiaohongshu, a free social networking website, has also attracted millions of users in China in recent years [8]. On these platforms, patients can obtain a large number of health videos by searching related keywords without any cost. According to a report from Tencent, at the end of 2020, 73% of users had seen health information–related short videos and 42% of users would watch related short videos 1-3 times per week for acquiring health information or seeking help for family members [10]. In TikTok, there were more than 200 million users who acquired health information every day. From December 2022 to January 2023, the reading amount of health information–related short videos was more than 45.4 billion [11]. Hence, these short-video sharing platforms played important roles in improving the population's basic knowledge of diseases. However, there is still no strict limitation on the video content and the uploaders on these platforms. Some inaccurate information may be conveyed to the users, leading to misleading and even influencing personal health decisions. Therefore, evaluating the content, accuracy, and quality of the information is essential.

However, to our knowledge, no study has examined information about colorectal polyps on short-video sharing platforms. Therefore, this study aimed to evaluate the quality and reliability of colorectal polyps–related videos on several widely used short-video sharing platforms (TikTok, WeChat, and Xiaohongshu), and then provide some fact-based recommendations for improving the quality and popularity of health-related videos.

Methods

Ethical Considerations

No ethics approval was applied for because the study only explored publicly available data on short-video sharing platforms and did not conduct any experiments on human subjects. In addition, there is no identifiable information about individual users or IDs in this study.

Search Strategy and Data Collection

The search was performed on TikTok, WeChat, and Xiaohongshu between May 26 and June 8, 2024. To minimize the bias introduced by personal recommendation algorithms, we registered a new account, then used "肠息肉" (intestinal polyps) or "结肠息肉" (colonic polyps) or "直肠息肉" (rectal polyps) or "结直肠息肉" (colorectal polyps) or "大肠息肉" (polyps of large intestine) as the keywords to search in above 3 short-video sharing platforms. Previous studies showed that most viewers tend to focus only on the top few pages of search results found online [12,13]. Our study mainly aimed to analyze the colorectal polyps–related health information that most viewers could access from short-video sharing platforms. Therefore, we only collected and recorded the top 100 videos for each search term in different platforms. The inclusion criteria were (1) videos with colorectal polyp content and (2) videos in Chinese. The exclusion criteria were (1) duplicated content, (2) advertisements, and (3) videos with irrelevant contents (polyps in other locations, hemorrhoids, rectal prolapse, etc). For included videos, the characteristics recorded and analyzed were titles, the number of likes, the number of comments, the number of collections, the number of shares, days since upload, video duration, video sources, the number of uploaders' fans, video presentation forms, and video content.

Assessment of Content and Quality of Videos

The *Journal of American Medical Association* (JAMA) benchmark criteria is a well-known tool for evaluating the reliability of information obtained from health-related websites, which consists of 4 evaluation dimensions, that is, author, attribution, disclosure, and currency [14,15]. The Global Quality Scale (GQS) is a 5-point Likert scale that can subjectively rate the overall quality of videos and also considers the flow and ease of website usage. It consisted of 5 criteria ranging from 1 point (poor quality) to 5 points (excellent flow and quality) [16,17]. The modified DISCERN is a valid tool and is recommended in the literature as a measure of the reliability and quality in web-based sources [18,19]. It consisted of 5 questions, and each question was given a weight of 1 point. The JAMA score, GQS score, and modified DISCERN score are used extensively in research [20-22]. To further assess the understandability and actionability of these videos, the Patient Education Materials Assessment Tool (PEMAT), developed by the Agency for Healthcare Research and Quality, was also used in this study [23]. The specific scoring criteria of the above score systems are shown in Table S1-S4 in Multimedia Appendix 1.

Scoring was performed by 2 researchers (J-LG and S-HX) independently, and discussions were carried out with a third

XSL•FO
RenderX

author (P-YL) to solve any uncertainties. Each video was scored 3 times—twice by J-LG, once by S-HX. This design permitted us to assess interrater and intrarater reliability with the intraclass correlation coefficient (ICC). Reliability was defined as poor (ICC<0.50), moderate (ICC=0.50-0.75), good (ICC=0.75-0.90), or excellent (ICC>0.90) [24]. The 2 researchers had good agreement for the JAMA score (ICC=0.92, 95% CI 0.91-0.93), GQS score (ICC=0.81, 95% CI 0.78-0.83), modified DISCERN score (ICC=0.75, 95% CI 0.72-0.78), understandability (ICC=0.81, 95% CI 0.78-0.83), and actionability (ICC=0.85, 95% CI 0.83-0.87). The intrarater reliability was also good overall (JAMA score=0.98, GQS score=0.85, modified DISCERN score=0.78, understandability=0.78, and actionability=0.86).

Statistical Analysis

SPSS (version 22.0; IBM Corp) and GraphPad Prism (version 9.0; Dotmatics) were used for analyzing the data. Categorical variables were described as frequencies and percentages. Normally distributed data were presented as a mean (SD), and nonnormally distributed data were expressed as median and range. Comparisons between 2 groups were performed using the Mann-Whitney *U* test or using the Student *t* test, and comparisons among 3 or more groups were performed by the Kruskal-Wallis test or one-way ANOVA. The Spearman correlation analysis was used to evaluate the relationship between quantitative variables. *P* value less than .05 (*P*<.05) was considered statistically significant.

Results

The General Characteristics of Videos

As shown in Figure 1, after excluding advertisements, duplicated, and irrelevant videos, a total of 816 colorectal polyps–related videos were included for further analysis. Among 816 videos, 301 videos came from TikTok, 230 videos were retrieved from WeChat, and 285 videos came from Xiaohongshu. The number of likes, comments, shares, and collections of videos from TikTok was higher than videos from WeChat and Xiaohongshu (all *P*<.001). The JAMA score, GQS score, and modified DISCERN score of videos from Tiktok and WeChat were higher than videos from Xiaohongshu (mean 2.3, SD 0.5 vs mean 2.2, SD 0.6 vs mean 1.6. SD 0.5; *P*<.001; mean 2.6, SD 0.8 vs mean 2.6, SD 0.8 vs mean 2.3, SD 0.8; *P*<.001; mean 2.9, SD 0.6 vs mean 2.9, SD 0.6 vs mean 1.8, SD 0.7; *P*<.001). There were statistical differences among the understandability and actionability of videos from different platforms (*P*<.001; Table 1). The detailed characteristics of colorectal polyps–related videos in different platforms are shown in Table 1.

Figure 1. Flowchart of the study.

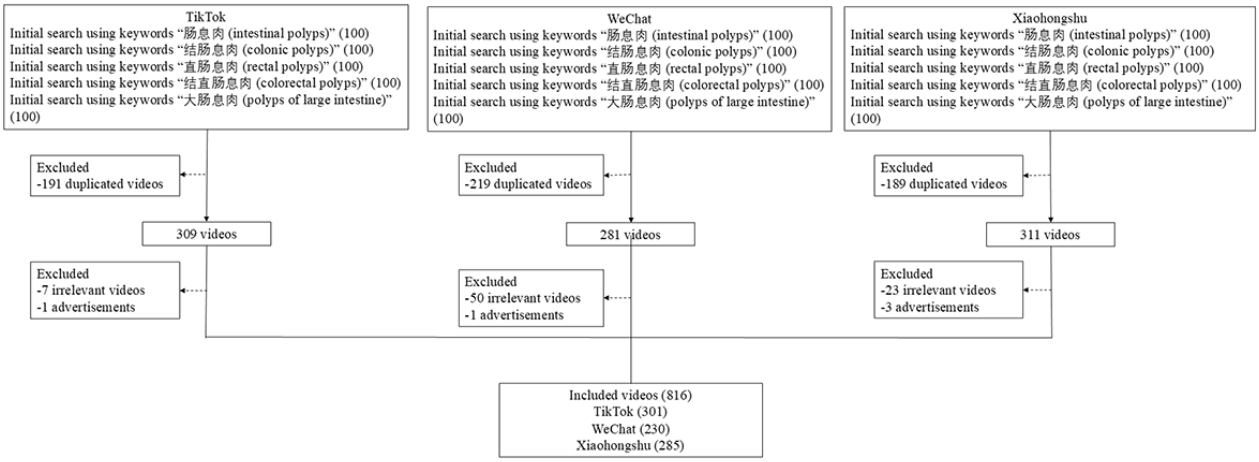

XSL•FO
RenderX

Table 1. The general characteristics and scores of the colorectal polyps–related videos.

Parameters	Total (N=816)	TikTok (n=301)	WeChat (n=230)	Xiaohongshu (n=285)	*P* value
Duration (seconds), median (range)	59 (5-1417)	72 (5-1417)	67 (6-727)	46 (5-1005)	<.001
Likes, median (range)	55 (0-201,000)	644 (16-201,000)	19 (0-1050)	13 (0-3638)	<.001
Comments, median (range)	5 (0-7205)	52 (0-7205)	1 (0-206)	1 (0-211)	<.001
Collections, median (range)	31 (0-25,000)	149 (1-25,000)	21 (0-3046)	9 (0-3986)	<.001
Shares, median (range)	78 (0-23,000)	128 (0-23,000)	42.5 (0-3168)	—[a]	<.001
Days since upload, median (range)	176 (0-1522)	280 (0-1487)	137 (0-1358)	137 (7-1522)	<.001
Fans of video uploaders, median (range)	13500 (1-62.2×10^5)	13.2×10^4 (239-62.2×10^5)	—	5612 (1-17.7×10^5)	<.001
JAMA[b] score, mean (SD)	2.0 (0.6)	2.3 (0.5)	2.2 (0.6)	1.6 (0.5)	<.001
GQS[c] score, mean (SD)	2.5 (0.8)	2.6 (0.8)	2.6 (0.8)	2.3 (0.8)	<.001
Modified DISCERN score, mean (SD)	2.5 (0.8)	2.9 (0.6)	2.9 (0.6)	1.8 (0.7)	<.001
Understandability, mean (SD)	80.4% (15.6%)	77.8% (17.6%)	84.9% (13%)	79.5% (14.3%)	<.001
Actionability, mean (SD)	42.2% (36.1%)	53.7% (34.4%)	42.6% (37.9%)	30.3% (32.2%)	<.001

[a]Not available.

[b]JAMA: Journal of American Medical Association.

[c]GQS: Global Quality Scale.

Video Source and Content

Table 2 shows the source and content of videos regarding colorectal polyps. Physicians were the main video uploaders (709/816, 86.9%), and video content mainly included disease knowledge (635/816, 77.8%) and outpatient scenarios (178/816, 21.8%). Disease knowledge in videos mainly focused on treatments (252/816, 30.9%), preventions (157/816, 19.2%), and carcinogenesis (151/816, 18.5%). Expert monologue (546/816, 66.9%) and dialogue (182/816, 22.3%) were dominated in the video presentation forms.

XSL•FO
RenderX

Table 2. The sources and content of the colorectal polyps–related videos.

Variable	Total (N=816), n (%)	TikTok (n=301), n (%)	WeChat (n=230), n (%)	Xiaohongshu (n=285), n (%)	*P* value
Video source					<.001
Physicians	709 (86.9)	294 (97.7)	148 (64.3)	267 (93.7)	
Hospital	38 (4.7)	1 (0.3)	36 (15.7)	1 (0.4)	
News agencies	34 (4.2)	3 (1)	31 (13.5)	0 (0)	
Independent users	26 (3.2)	3 (1)	10 (4.3)	13 (4.6)	
Others	9 (1.1)	0 (0)	5 (2.1)	4 (1.5)	
Different medical specialties					<.001
TCM[a] practitioner	352 (49.6)	105 (35.7)	81 (54.7)	166 (62.2)	
Western medicine practitioner	357 (50.4)	189 (64.3)	67 (45.3)	101 (37.8)	
Video content					.006
Disease knowledge	635 (77.8)	210 (69.8)	200 (87)	225 (78.9)	
Outpatient scenarios	178 (21.8)	89 (29.6)	30 (13)	59 (20.7)	
Personal experience	3 (0.4)	2 (0.6)	0 (0)	1 (0.4)	
Different disease knowledge					.61
Treatment	252 (30.9)	81 (26.9)	84 (36.5)	87 (30.5)	
Prevention	157 (19.2)	49 (16.3)	62 (27)	46 (16.1)	
Carcinogenesis	151(18.5)	59 (19.6)	47 (20.4)	45 (15.8)	
Symptom	98 (12)	39 (13)	34 (14.8)	25 (8.8)	
Definition	45 (5.5)	18 (6)	13 (5.7)	14 (4.9)	
Posttreatment caveats	41 (5)	18 (6)	12 (5.2)	11 (3.9)	
Reexamination	40 (4.9)	14 (4.7)	10 (4.3)	16 (5.6)	
Others	69 (8.5)	20 (6.6)	27 (11.7)	22 (7.7)	
Video presentation form					<.001
Expert monologue	546 (66.9)	183 (60.8)	174 (75.7)	189 (66.3)	
Dialogue	182 (22.3)	79 (26.2)	36 (15.7)	67 (23.5)	
Visual pictures and literature	32 (3.9)	13 (4.3)	4 (1.7)	15 (5.3)	
Endoscope operation demonstration	31 (3.8)	21 (7)	5 (2.2)	5 (1.8)	
Animation	14 (1.7)	2 (0.7)	7 (3)	5 (1.8)	
Vlogs of patients	6 (0.7)	2 (0.7)	0 (0)	4 (1.4)	
Others	5 (0.6)	1 (0.3)	4 (1.7)	0 (0)	

[a]TCM: traditional Chinese medicine.

The Quality and Popularity of Videos From Different Sources With Different Contents and Different Presentation Forms

The JAMA scores and modified DISCERN scores of videos uploaded by independent users were significantly lower than videos uploaded by physicians' hospitals and news agencies (all *P*<.001; Figures 2A and 2C). The GQS score and understandability of videos uploaded by hospitals and news agencies were higher than those uploaded by physicians (*P*<.001, *P*<.001, *P*=.01, and *P*=.009, respectively) and independent users (*P*<.001, *P*=.004, *P*<.001, and *P*=.03, respectively; Figures 2B and 2D). In addition, the actionability of videos uploaded by news agencies was better than physicians and independent users (*P*=.02 and *P*=.01; Figure 2E). Among videos uploaded by physicians, the GQS score, modified DISCERN score, and actionability of videos uploaded by traditional Chinese medicine (TCM) practitioners were lower than videos uploaded by Western medicine practitioners (*P*<.001 and *P*<.001, and *P*=.006, respectively), while there were no significant differences in the JAMA score and understandability between the 2 groups (Figures 3A-3E). The JAMA score, GQS score, modified DISCERN score, understandability, and actionability of videos about disease knowledge were

XSL•FO
RenderX

significantly higher than videos about outpatient scenarios (*P*<.001, *P*=.02, *P*<.001, *P*<.001, and *P*<.001, respectively; Figures 4A-4E).

Among videos about disease knowledge, videos about symptoms had higher JAMA scores than videos about treatment, prevention, and carcinogenesis (*P*=.001, *P*=.02, and *P*=.048, respectively; Figure 5A). Compared with other disease knowledge, videos about treatment had lower GQS scores than prevention, carcinogenesis, symptom, and posttreatment caveats (*P*<.001, *P*=.003, *P*=.001, and *P*<.001, respectively; Figure 5B). There was no significant difference in the modified DISCERN score among videos about different disease knowledge (Figure 5C). Meanwhile, videos conveying treatment had poorer understandability than those conveying prevention, symptoms, posttreatment caveats, and reexamination (*P*=.005, *P*=.02, *P*=.04, and *P*<.001, respectively). As for actionability, videos about prevention, posttreatment caveats, and reexamination were better than those about treatment (*P*<.001, *P*<.001, and *P*=.001) and carcinogenesis (*P*<.001, *P*<.001, and *P*=.04; Figures 5D and 5E). Among different video presentation forms, videos presented in vlogs of patients had lower JAMA scores and modified DISCERN scores. Videos presented in expert monologues or animation had higher GQS scores than those presented in dialogue (*P*<.001 and *P*<.001), endoscope operation demonstration (*P*<.001 and *P*=.006), and vlogs (*P*<.001 and *P*=.001). In addition, the understandability of the forms of expert monologues and animation were better than the forms of endoscope operation demonstration (*P*<.001 and *P*=.02), visual pictures and literature (*P*<.001 and *P*=.01), and vlogs (*P*=.004 and *P*=.03). The form of expert monologues had better actionability than dialogue (*P*<.001), endoscope operation demonstration (*P*<.001), visual pictures and literature (*P*<.001; Figures 6A-6E).

We also compared the popularity (likes, comments, collections, and shares) of videos uploaded by different sources with different contents and different presentation forms (Table 3). Among different video sources, physicians, especially Western medicine practitioners, received more likes and comments (*P*≤.001). Videos conveying posttreatment caveats and reexamination were more likely to be collected and shared (*P*=.03 and *P*=.006).

Figure 2. The Journal of American Medical Association (JAMA) score, Global Quality Scale (GQS) score, modified DISCERN score, Patient Education Materials Assessment Tool (PEMAT)–understandability, and PEMAT-actionability of videos on colorectal polyps from different sources. (A) the JAMA score, (B) the GQS score, (C) the modified DISCERN score, (D) PEMAT-understandability, and (E) PEMAT-actionability. **P*<.05, ***P*<.01, ****P*<.001.

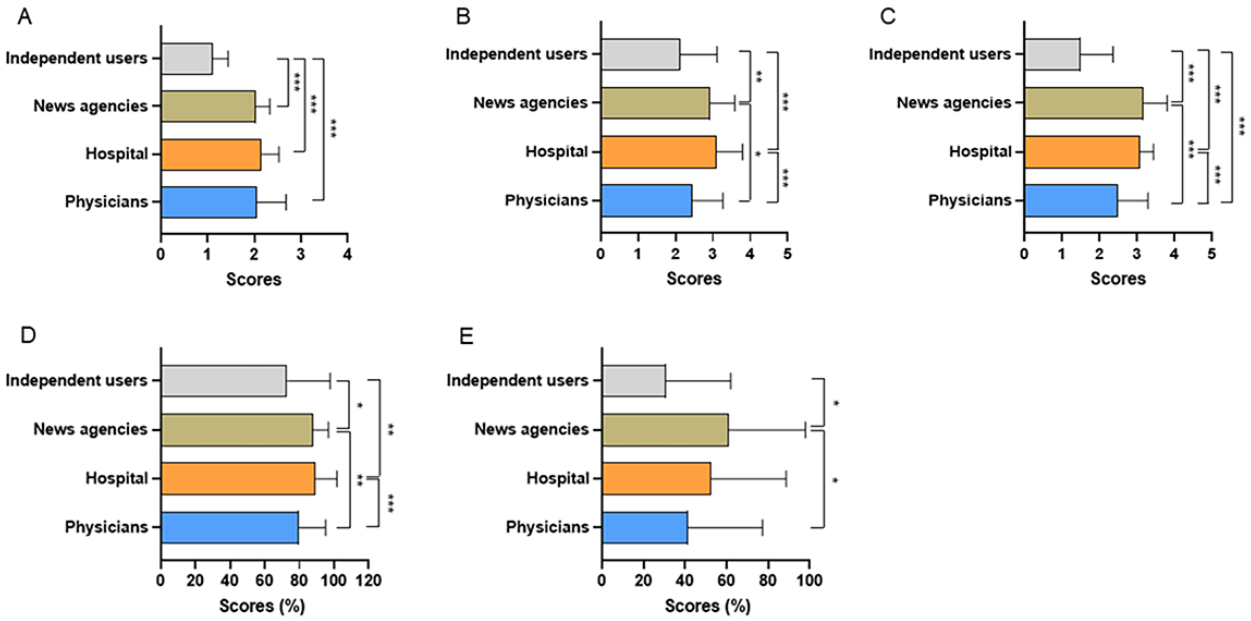

XSL•FO
RenderX

Figure 3. The Journal of American Medical Association (JAMA) score, Global Quality Scale (GQS) score, modified DISCERN score, Patient Education Materials Assessment Tool (PEMAT)–understandability, and PEMAT-actionability of videos on colorectal polyps of Western medicine practitioners and traditional Chinese medicine practitioners. (A) the JAMA score, (B) the GQS score, (C) the modified DISCERN score, (D) PEMAT-understandability, (E) PEMAT-actionability. ns: not significant; TCM: traditional Chinese medicine. **P*<.01, ***P*<.001.

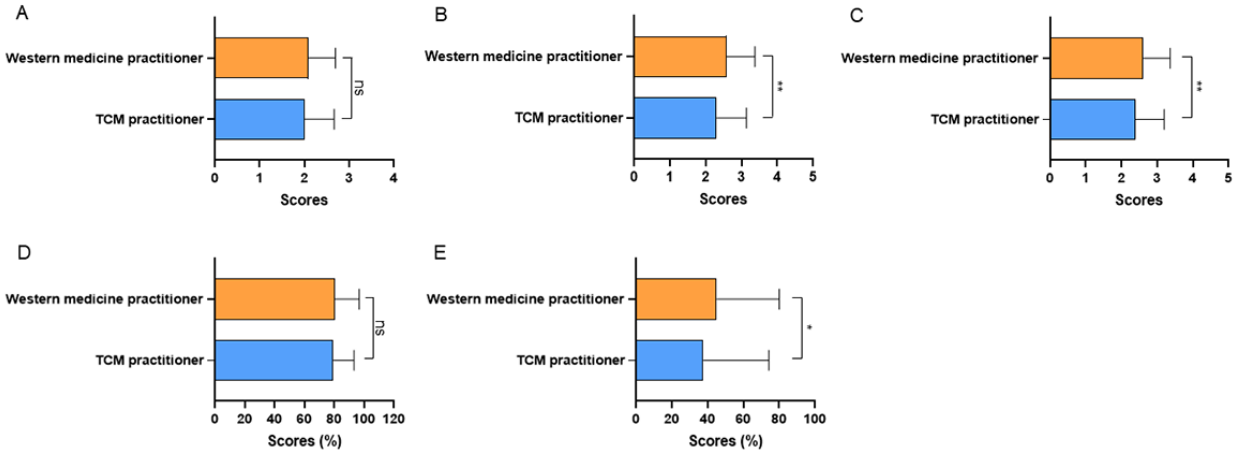

Figure 4. The Journal of American Medical Association (JAMA) score, Global Quality Scale (GQS) score, modified DISCERN score, Patient Education Materials Assessment Tool (PEMAT)–understandability, and PEMAT-actionability of videos on colorectal polyps with different contents. (A) the JAMA score, (B) the GQS score, (C) the modified DISCERN score, (D) PEMAT-understandability, and (E) PEMAT-actionability. **P*<.05, ***P*<.01, ****P*<.001.

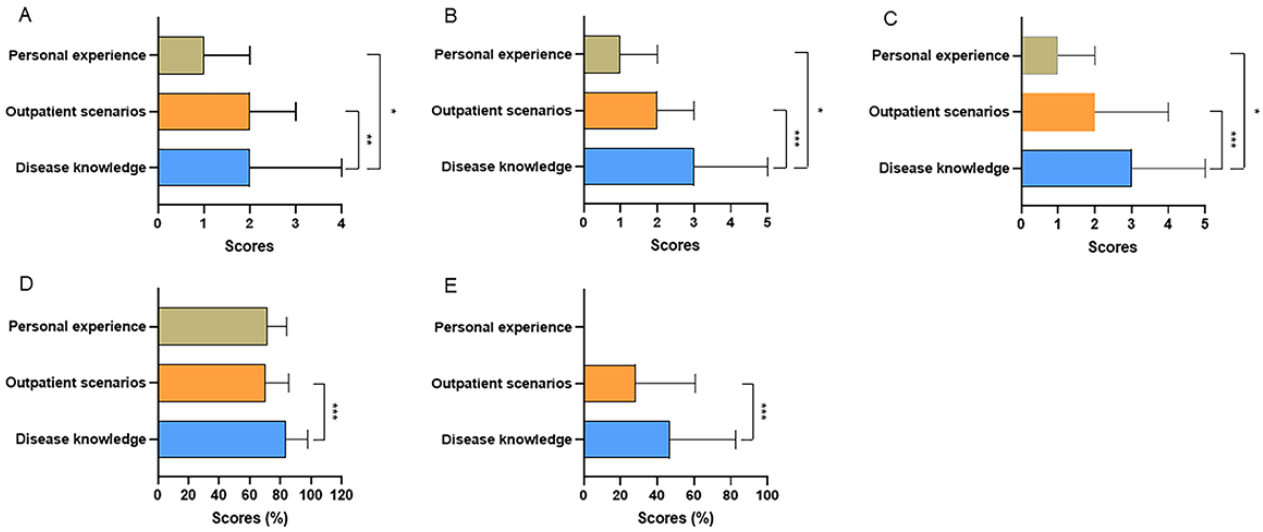

XSL•FO
RenderX

Figure 5. The Journal of American Medical Association (JAMA) score, Global Quality Scale (GQS) score, modified DISCERN score, Patient Education Materials Assessment Tool (PEMAT)–understandability, and PEMAT-actionability of videos on colorectal polyps with different disease knowledge. (A) the JAMA score, (B) the GQS score, (C) the modified DISCERN score, (D) PEMAT-understandability, and (E) PEMAT-actionability. *P<.05, **P<.01, ***P<.001.

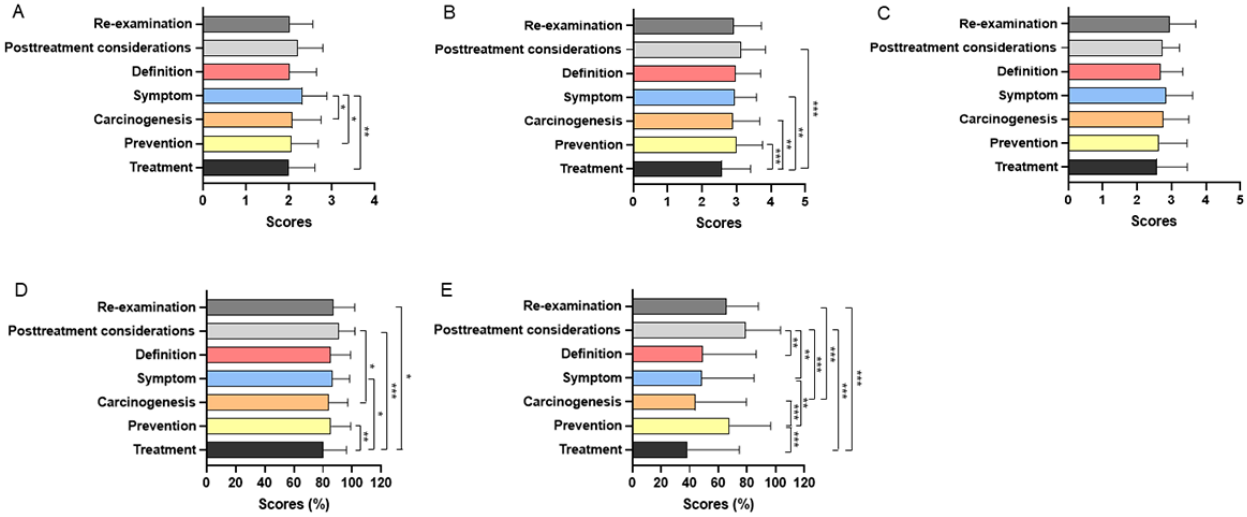

Figure 6. The Journal of American Medical Association (JAMA) score, Global Quality Scale (GQS) score, modified DISCERN score, Patient Education Materials Assessment Tool (PEMAT)–understandability, and PEMAT-actionability of videos on colorectal polyps with different presentation forms. (A) the JAMA score, (B) the GQS score, (C) the modified DISCERN score, (D) PEMAT-understandability, and (E) PEMAT-actionability. *P<.05, **P<.01, ***P<.001.

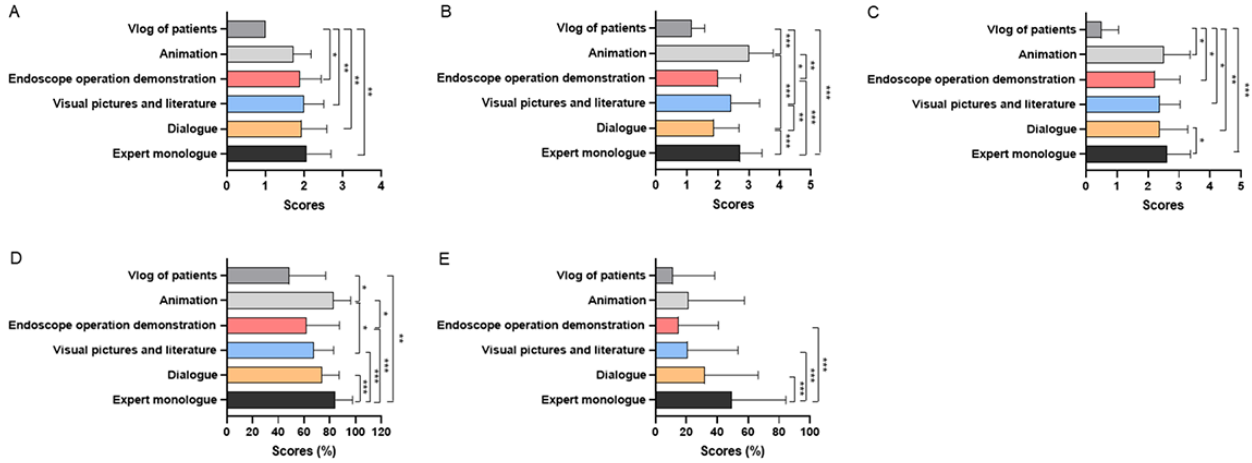

XSL•FO
RenderX

Table 3. The popularity of videos from different sources with different contents and different presentation forms.

Variables	Likes	Comments	Collections	Shares
Video source				
Physician, mean (range)	64 (0-201,000)	6 (0-7205)	33 (0-25,000)	83 (0-23,000)
Hospital, mean (range)	23 (0-4019)	0 (0-177)	24 (0-1227)	38 (1-934)
News agency, mean (range)	62 (0-3744)	3 (0-109)	32 (1-1246)	114 (1-2842)
Independent user, mean (range)	13 (0-536)	3 (0-383)	15 (0-288)	29 (4-110)
Others, mean (range)	11 (1-71)	0 (0-12)	9 (2-320)	52 (15-649)
P value	.001	.001	.06	.81
Different medical specialties				
TCM[a] practitioner, mean (range)	27 (0-35,000)	3 (0-1692)	21 (0-7863)	73 (1-5828)
Western medicine practitioner, mean (range)	152 (0-201,000)	13 (0-7205)	54 (0-25,000)	90 (0-23,000)
P value	<.001	<.001	<.001	.06
Video content				
Disease knowledge, mean (range)	52 (0-201,000)	4 (0-7205)	33 (0-25,000)	89 (0-23,000)
Outpatient scenarios, mean (range)	70 (0-27,000)	11 (0-1692)	27 (0-2582)	50 (1-5100)
Personal experience, mean (range)	95 (3-186)	37 (0-73)	8 (1-15)	5 (5-5)
P value	.38	.02	.16	.02
Different disease knowledge				
Treatment, mean (range)	50 (0-201,000)	4 (0-7205)	31 (0-25,000)	74 (1-23,000)
Prevention, mean (range)	68 (0-139,000)	4 (0-4752)	39 (0-20,000)	125 (0-23,000)
Carcinogenesis, mean (range)	60 (0-201,000)	4 (0-7205)	35 (0-25,000)	90 (0-23,000)
Symptom, mean (range)	58 (0-43,000)	4 (0-1197)	29 (0-7637)	61 (2-4947)
Definition, mean (range)	63 (0-5857)	3 (0-1146)	28 (0-1962)	54 (0-4947)
Posttreatment caveats, mean (range)	106 (2-32,000)	10 (0-1834)	94 (1-11,000)	200 (10-20,000)
Reexamination, mean (range)	80 (1-139,000)	6 (0-4752)	70 (0-20,000)	117 (0-23,000)
Others, mean (range)	47 (0-7897)	3 (0-1146)	26 (0-1686)	51 (1-4947)
P value	.25	.31	.03	.006
Video presentation form				
Expert monologue, mean (range)	50 (0-108,000)	4 (0-4362)	33 (0-14,000)	82 (0-20,000)
Dialogue, mean (range)	55 (0-201,000)	8 (0-7205)	27 (0-25,000)	56 (1-23,000)
Endoscope operation demonstration, mean (range)	340 (2-3314)	42 (0-1416)	93 (1-584)	120 (2-1140)
Visual pictures and literature, mean (range)	78 (1-13,000)	4 (0-1305)	43 (0-3986)	91 (3-2409)
Animation, mean (range)	12 (0-949)	0 (0-211)	15 (0-364)	54 (1-249)
Vlog of patients, mean (range)	14 (0-536)	9 (0-383)	4 (0-48)	35 (5-65)
Others, mean (range)	66 (13-177)	4 (2-7)	28 (12-73)	49 (18-121)
P value	.02	<.001	.048	.39

[a]TCM: traditional Chinese medicine.

XSL•FO
RenderX

Correlation Analysis

Spearman correlation (ρ) analysis revealed the relationship among different video variables. It was shown that there were positive correlations among the number of likes, comments, collections, shares, and fans (all *P*<.001), while the duration of videos was not associated with the number of likes, comments, collections, and shares (all *P*>.05; Table 4). The number of likes, comments, collections, and fans was positively correlated with the JAMA score, GQS score, and modified DISCERN score. Videos with higher GQS scores and modified DISCERN scores were associated with more shares (*P*<.001 and *P*<.001, respectively) and longer duration (*P*<.001 and *P*=.04, respectively). Actionability was positively correlated with the number of collections, shares, fans, and longer duration (*P*=.003, *P*=.03, *P*<.001, and *P*<.001, respectively), while understandability had a negative correlation with the duration of videos (*P*=.001; Table 5).

Table 4. The correlation analysis between the video variables.

Variables	Likes	Comments	Collections	Shares
Likes				
ρ	1	0.898	0.939	0.810
P value	__[a]	<.001	<.001	<.001
Comments				
ρ	0.898	1	0.861	0.840
P value	<.001	—	<.001	<.001
Collections				
ρ	0.939	0.861	1	0.885
P value	<.001	<.001	—	<.001
Shares				
ρ	0.810	0.840	0.885	1
P value	<.001	<.001	<.001	—
Duration				
ρ	0.032	0.036	0.054	0.020
P value	.40	.31	.13	.64
Days since uploaded				
ρ	0.048	0.113	0.046	0.101
P value	.17	.001	.19	.02
Fans				
ρ	0.442	0.439	0.436	0.410
P value	<.001	<.001	<.001	<.001

[a]Not applicable.

XSL•FO
RenderX

Table 5. The correlation analysis between video variables and the video quality.

Variables	JAMA[a]	GQS[b]	Modified DISCERN	Understandability	Actionability
Likes					
ρ	0.118	0.135	0.167	–0.006	0.057
P value	.001	<.001	<.001	.85	.10
Comments					
ρ	0.102	0.107	0.189	–0.036	0.053
P value	.004	.002	<.001	.30	.13
Collections					
ρ	0.125	0.158	0.174	–0.016	0.106
P value	<.001	<.001	<.001	.65	.003
Shares					
ρ	0.055	0.185	0.177	0.012	0.093
P value	.20	<.001	<.001	.78	.03
Duration					
ρ	0.042	0.247	0.073	–0.116	0.168
P value	.23	<.001	.04	.001	<.001
Fans					
ρ	0.182	0.127	0.269	0.001	0.149
P value	<.001	.002	<.001	.98	<.001

[a]JAMA: Journal of American Medical Association.

[b]GQS: Global Quality Scale.

Discussion

Principal Findings

Health problems are crucial concerns in our daily lives, which require significant attention, accurate assessment, and prompt intervention. With the increasing popularity of mobile internet, it has become one of the most popular ways for obtaining health and medical information. A survey showed that 70% of network users first acquired health information from the internet [25]. There have been studies that evaluated the quality of health information on several diseases, like gallstone disease, chronic obstructive pulmonary disease, and diabetes on the internet. It was found that the quality of these popular science videos on different diseases varied significantly [17,26,27]. So, how about the reliability and quality of colorectal polyps–related videos online? In this study, we analyzed the content, reliability, and quality of videos regarding colorectal polyps from 3 popular short-video sharing platforms—TikTok, WeChat, and Xiaohongshu. Among these, videos on TikTok were more popular and received more likes, comments, collections, and shares. Notably, the overall quality of short videos related to colorectal polyps was unsatisfactory, especially videos from Xiaohongshu. Physicians were the main uploaders, and most videos aimed to convey disease knowledge directly in the form of expert monologue. Disease knowledge about treatments, preventions, and carcinogenesis were the most frequently discussed in colorectal polyps–related videos. In addition, videos from physicians and news agencies were heavily preferred, and videos from news agencies and hospitals had higher quality. As for video content, videos primarily concerned with disease knowledge are more popular and of higher quality.

Factors Correlated With the Popularity of Videos

The number of likes of videos on colorectal polyps attained 201,000 at most, and the number of collections and shares of some videos was more than 10,000, which showed that colorectal polyps have drawn much attention. It was reported that the number of likes, comments, collections, and shares could partly reflect the popularity of videos [28,29]. In this study, we found that videos uploaded by Western medicine practitioners and TCM practitioners were almost in equal numbers, but videos from Western medicine practitioners were more popular. It was possibly because the current prominence of the Western medicine may result in people having fewer opportunities to be exposed to TCM, especially for young people, the main users of the internet. The current study also showed that most videos from short-video sharing platforms aimed to convey disease knowledge, and videos about posttreatment caveats and reexamination were more likely to be collected and shared by viewers. It was reported that the risk of bleeding and perforation increased approximately 7 times among those who had a polypectomy compared with those who did not [30]. Hence, knowing the caveats after polypectomy is conducive to postoperative recovery and to avoid adverse outcomes. A meta-analysis confirmed the risk of recurrence of colorectal polyps after standard endoscopic mucosal resection [31], and some of them eventually develop into CRC [4,32].

XSL•FO
RenderX

However, many patients ignore the importance of reexamination in clinical practice. In addition, in the case of different types of polyps, the time interval for reexamination may change. Videos about reexamination could inform patients about the importance of reexamination and instruct them on how to reexamine. Furthermore, there was a positive correlation between likes, comments, collections, and shares, which indicated that a video receiving more likes and comments would be more likely to be collected and shared. The number of fans played an important role in the popularity of videos, thus the influence of these uploaders with large numbers of fans cannot be ignored in conveying health-related information.

Factors Correlated With Video Quality

Although the majority of videos regarding colorectal polyps were uploaded by physicians and mainly conveyed disease knowledge, the overall reliability and quality of these videos regarding colorectal polyps were not good enough. This was consistent with several previous studies. For example, Lee et al [33] analyzed the videos regarding gallstone disease on YouTube (Google) and were surprised to find that more than half of the videos were misleading and presented a risk of harmful consequences. Kong et al [27] assessed the quality of the information in diabetes-related videos on TikTok and found that the overall quality of the information from these videos did not fully meet the health information needs of patients with diabetes. In this study, we also found that a part of uploaders claimed to be medical staff, but did not authenticate their identities. It may be possible that some uploaders pretend to be physicians to increase trust and video views, which affected the overall reliability of colorectal polyps–related videos. In addition, some uploaders may convey inaccurate information or exaggerate the effect of treatment for building a more professional image. However, incorrect health information in short videos may increase patients' risks who may make health decisions based on inaccurate information. Hence, people should exercise caution when acquiring health information online.

The quality of videos uploaded by news agencies and hospitals were higher than those uploaded by physicians. These results are compatible with some studies evaluating health information videos [15,34]. This could be because of videos from the official accounts of institutions are in general created by professional media personnel with the participation of medical professionals. It is noteworthy that video duration demonstrated a positive correlation with video quality, which was consistent with previous studies [21,35,36]. Indeed, half of the videos in our study lasted for less than 1 minute, even a few or a dozen seconds, making it difficult to provide sufficient information, which may greatly affect the video quality. Also, video presentation forms were associated with the quality of videos. Compared with dialogue and vlogs, expert monologue could convey more health information in a short time, and animation could make disease knowledge easy to understand. In addition, the number of likes, shares, and comments was positively correlated with the quality of videos. That is to say, the reliability and substantiality of video contents are important factors that increase the popularity of the videos.

Recommendations Based on Our Results

There is no doubt that health-related content is more complicated, professional, and less attractive than other entertainment videos, therefore it is a challenge to attract more viewers. To increase the influence of related videos, choosing physicians and news agencies with large numbers of fans to further spread correct medical information may be a good idea. Furthermore, building up a medical classification module could further assist users in accessing health information easily. Also, providing standardized criteria for medical videos on short-video sharing platforms, and hiring professional medical workers to conduct reviews before uploading the videos could improve the quality and reliability of health information. In addition, platform administrators need to enhance the management of identity authentication to increase the users' recognition and dissemination of professional information. Videos that are very short cannot provide abundant information, while very long videos may increase the difficulty of understanding. Hence, video uploaders need to think about how to offer adequate information of high quality within a reasonable duration. On the other hand, videos on how to treat colorectal polyps occupied a large proportion, but the quality was quite unsatisfactory and urged to be improved.

The Significance of This Study

Improving the public's general understanding of colorectal polyps can aid in earlier recognition, prevention, and management of colorectal polyps and CRC. Short videos on the internet are important sources for the public to access health-related information, particularly in a medical resource–limited setting. However, the quality varies across short videos containing knowledge about different diseases. This is the first study to evaluate the quality and reliability of short videos regarding colorectal polyps using multiple tools (JAMA score, GQS score, modified DISCERN score, and PEMAT). Our study analyzed the popularity, and the quality of different sources, contents, and presentation forms of videos regarding colorectal polyps. We found that the quality of related videos needed to be improved. Based on the results, we provided some suggestions on how to improve the popularity and quality of short videos regarding colorectal polyps. It would be helpful to produce the right, reliable, and high-quality information source and improve the public's understanding of colorectal polyps.

Limitations

There are some limitations in this study. First, the video search was only performed on Chinese video-sharing platforms, so the findings may not apply to platforms in other languages. Second, only the top 100 videos of each search term rather than all searched videos were included, which may not reflect the overall situation. But most users usually only view the first few pages of search results, our results may be more in line with the information acquired by most users. Third, the present scoring criteria of video quality are mainly based on subjective judgment. Although scoring was performed independently by 2 researchers and they had good agreement, subjective differences still could not be ignored. Finally, although we registered and logged in new accounts on each platform, the

XSL•FO
RenderX

bias in the selection and evaluation of the videos may still have existed.

Conclusion

Colorectal polyps–related videos on short-video sharing platforms were mainly uploaded by physicians and mainly conveyed information regarding disease knowledge in the form of an expert monologue. However, the overall reliability and quality of these videos were not good enough and need to be improved. Users should remain cautious and selective when watching videos regarding colorectal polyps on short-video sharing platforms.

Acknowledgments

The study is supported by Tongji Hospital Clinical Research Fund (2024TJCR019), Hainan Provincial Natural Science Foundation of China (823QN372 and 821RC736), and Key R&D Program of Hubei Province of China (2023BCB003).

Authors' Contributions

All authors contributed to the study conception and design. J-LG, S-HX, and P-YL were involved in the acquisition and interpretation of data. J-LG and S-HX were involved in the analysis of data. J-LG, KZ, and L-NF drafted the manuscript. Y-YH and J-YL critically revised the manuscript. J-LG, P-YL, and J-ZL were involved in the study design and supervision. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File , 18 KB - jmir_v26i1e51655_app1.docx]

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E, Maake C, Rehrauer H, et al. Transcriptome profile of human colorectal adenomas. Mol Cancer Res 2007 Dec;5(12):1263-1275. [doi: 10.1158/1541-7786.MCR-07-0267] [Medline: 18171984]
3. Onyoh EF, Hsu WF, Chang LC, Lee YC, Wu MS, Chiu HM. The rise of colorectal cancer in Asia: epidemiology, screening, and management. Curr Gastroenterol Rep 2019 Jul 10;21(8):36. [doi: 10.1007/s11894-019-0703-8] [Medline: 31289917]
4. Zauber AG, Winawer SJ, O'Brien MJ, Lansdorp-Vogelaar I, van Ballegooijen M, Hankey BF, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 2012 Feb 23;366(8):687-696 [FREE Full text] [doi: 10.1056/NEJMoa1100370] [Medline: 22356322]
5. Yang Y, Han Z, Li X, Huang A, Shi J, Gu J. Epidemiology and risk factors of colorectal cancer in China. Chin J Cancer Res 2020 Dec 31;32(6):729-741 [FREE Full text] [doi: 10.21147/j.issn.1000-9604.2020.06.06] [Medline: 33446996]
6. Cottet V, Jooste V, Fournel I, Bouvier A, Faivre J, Bonithon-Kopp C. Long-term risk of colorectal cancer after adenoma removal: a population-based cohort study. Gut 2012 Aug;61(8):1180-1186. [doi: 10.1136/gutjnl-2011-300295] [Medline: 22110052]
7. Bailie L, Loughrey MB, Coleman HG. Lifestyle risk factors for serrated colorectal polyps: a systematic review and meta-analysis. Gastroenterology 2017 Jan;152(1):92-104 [FREE Full text] [doi: 10.1053/j.gastro.2016.09.003] [Medline: 27639804]
8. The new media ecological insight in 2023.: QuestMobile URL: https://www.questmobile.com.cn/research/report/1726888249161519105 [accessed 2024-07-02]
9. The new media content of ecological data report in 2022.: Newrank URL: https://data.newrank.cn/article/article-detail/e0dadc20dbc445cd [accessed 2024-07-02]
10. Medical popularscience short video and live insight report in 2021.: Tencent URL: https://www.renrendoc.com/paper/228868968.html [accessed 2024-07-02]
11. Health popular science-realted data report.: Tiktok URL: https://www.douyin.com/search/%E5%81%A5%E5%BA%B7%E7%A7%91%E6%99%AE%E6%95%B0%E6%8D%AE%E6%8A%A5%E5%91%8A [accessed 2024-07-02]
12. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsychol Behav 2004 Oct;7(5):497-510. [doi: 10.1089/cpb.2004.7.497] [Medline: 15667044]
13. Chitika C. The value of google results positioning. URL: http://chitika.com/google-positioning-value [accessed 2024-07-02]

XSL•FO
RenderX

14. Kwan JY, Stocco F, Scott DJA, Bailey MA, Coughlin PA. Assessment of internet-based information on statin therapy. Eur J Cardiovasc Nurs 2024 Mar 12;23(2):115-121. [doi: 10.1093/eurjcn/zvad061] [Medline: 37367216]
15. Ergenç M, Uprak TK. YouTube as a source of information on Helicobacter pylori: Content and quality analysis. Helicobacter 2023 Aug;28(4):e12971. [doi: 10.1111/hel.12971] [Medline: 36942858]
16. Chen Z, Pan S, Zuo S. TikTok and YouTube as sources of information on anal fissure: A comparative analysis. Front Public Health 2022;10:1000338 [FREE Full text] [doi: 10.3389/fpubh.2022.1000338] [Medline: 36407987]
17. Sun F, Zheng S, Wu J. Quality of information in gallstone disease videos on TikTok: cross-sectional study. J Med Internet Res 2023 Feb 08;25:e39162 [FREE Full text] [doi: 10.2196/39162] [Medline: 36753307]
18. He Z, Wang Z, Song Y, Liu Y, Kang L, Fang X, et al. The reliability and quality of short videos as a source of dietary guidance for inflammatory bowel disease: cross-sectional study. J Med Internet Res 2023 Feb 09;25:e41518 [FREE Full text] [doi: 10.2196/41518] [Medline: 36757757]
19. Moon H, Lee GH. Evaluation of Korean-language COVID-19-related medical information on YouTube: cross-sectional infodemiology study. J Med Internet Res 2020 Aug 12;22(8):e20775 [FREE Full text] [doi: 10.2196/20775] [Medline: 32730221]
20. Du RC, Zhang Y, Wang MH, Lu NH, Hu Y. TikTok and Bilibili as sources of information on Helicobacter pylori in China: A content and quality analysis. Helicobacter 2023 Oct;28(5):e13007. [doi: 10.1111/hel.13007] [Medline: 37452727]
21. Barlas T, Ecem Avci D, Cinici B, Ozkilicaslan H, Muhittin Yalcin M, Eroglu Altinova A. The quality and reliability analysis of YouTube videos about insulin resistance. Int J Med Inform 2023 Feb;170:104960. [doi: 10.1016/j.ijmedinf.2022.104960] [Medline: 36525801]
22. Zhang X, Yang Y, Shen Y, Zhang K, Ma L, Ding C, et al. Quality of online video resources concerning patient education for neck pain: A YouTube-based quality-control study. Front Public Health 2022;10:972348 [FREE Full text] [doi: 10.3389/fpubh.2022.972348] [Medline: 36211682]
23. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
24. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
25. Prestin A, Vieux SN, Chou WS. Is online health activity alive and well or flatlining? Findings from 10 years of the Health Information National Trends Survey. J Health Commun 2015;20(7):790-798. [doi: 10.1080/10810730.2015.1018590] [Medline: 26042588]
26. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021 Dec 20;23(12):e28318 [FREE Full text] [doi: 10.2196/28318] [Medline: 34931996]
27. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021 Sep 01;23(9):e30409 [FREE Full text] [doi: 10.2196/30409] [Medline: 34468327]
28. Yu JS, Carr JB, Thomas J, Kostas J, Wang Z, Khilnani T, et al. Trends in patient, physician, and public perception of ulnar collateral ligament reconstruction using social media analytics. Orthop J Sports Med 2021 Mar;9(3):2325967121990052 [FREE Full text] [doi: 10.1177/2325967121990052] [Medline: 34250162]
29. Zhu F, Jiao J, Huang Y, Xiao F, Zuo W, Chen M, et al. A preliminary study of the surgical approach for posterior tibial plateau fractures: Based on posterior fragment segment classification. Injury 2022 Nov;53(11):3820-3827. [doi: 10.1016/j.injury.2022.09.009] [Medline: 36116959]
30. Rabeneck L, Paszat LF, Hilsden RJ, Saskin R, Leddin D, Grunfeld E, et al. Bleeding and perforation after outpatient colonoscopy and their risk factors in usual clinical practice. Gastroenterology 2008 Dec;135(6):1899-1906.e1. [doi: 10.1053/j.gastro.2008.08.058] [Medline: 18938166]
31. Rotermund C, Djinbachian R, Taghiakbari M, Enderle MD, Eickhoff A, von Renteln D. Recurrence rates after endoscopic resection of large colorectal polyps: A systematic review and meta-analysis. World J Gastroenterol 2022 Aug 07;28(29):4007-4018 [FREE Full text] [doi: 10.3748/wjg.v28.i29.4007] [Medline: 36157546]
32. Bobe G, Murphy G, Albert PS, Sansbury LB, Lanza E, Schatzkin A, et al. Dietary lignan and proanthocyanidin consumption and colorectal adenoma recurrence in the Polyp Prevention Trial. Int J Cancer 2012 Apr 01;130(7):1649-1659 [FREE Full text] [doi: 10.1002/ijc.26184] [Medline: 21618513]
33. Lee JS, Seo HS, Hong TH. YouTube as a source of patient information on gallstone disease. World J Gastroenterol 2014 Apr 14;20(14):4066-4070 [FREE Full text] [doi: 10.3748/wjg.v20.i14.4066] [Medline: 24744597]
34. Uprak TK, Ergenç M. Assessment of esophagectomy videos on YouTube: is peer review necessary for quality? J Surg Res 2022 Nov;279:368-373. [doi: 10.1016/j.jss.2022.06.037] [Medline: 35820318]
35. Ozsoy-Unubol T, Alanbay-Yagci E. YouTube as a source of information on fibromyalgia. Int J Rheum Dis 2021 Feb;24(2):197-202. [doi: 10.1111/1756-185X.14043] [Medline: 33355406]
36. Gaş S, Zincir Ö, Bozkurt AP. Are YouTube videos useful for patients interested in botulinum toxin for bruxism? J Oral Maxillofac Surg 2019 Sep;77(9):1776-1783. [doi: 10.1016/j.joms.2019.04.004] [Medline: 31077674]

XSL•FO
RenderX

Abbreviations

CRC: colorectal cancer
GQS: Global Quality Scale
ICC: intraclass correlation coefficient
JAMA: Journal of American Medical Association
PEMAT: Patient Education Materials Assessment Tool
TCM: traditional Chinese medicine

Edited by A Schwartz; submitted 07.08.23; peer-reviewed by P Ebrahimi, D Boatman, A Khosla; comments to author 02.05.24; revised version received 22.07.24; accepted 03.09.24; published 29.10.24.

Please cite as:
Guan JL, Xia SH, Zhao K, Feng LN, Han YY, Li JY, Liao JZ, Li PY
Videos in Short-Video Sharing Platforms as Sources of Information on Colorectal Polyps: Cross-Sectional Content Analysis Study
J Med Internet Res 2024;26:e51655
URL: https://www.jmir.org/2024/1/e51655
doi:10.2196/51655
PMID:39470708

©Jia-Lun Guan, Su-Hong Xia, Kai Zhao, Li-Na Feng, Ying-Ying Han, Ji-Yan Li, Jia-Zhi Liao, Pei-Yuan Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceived Experiences and Needs of Digital Resources Among Postpartum Women in the United Arab Emirates: Qualitative Focus Group Study

Nivine Hanach[1], PhD; Roba Saqan[2], MSc; Hadia Radwan[3], PhD; Wegdan Baniissa[4], PhD; Nanne de Vries[1], PhD

[1]Faculty of Health, Medicine, and Life Sciences, Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

[2]Health Promotion Group, Research Institute of Medical and Health Sciences, University of Sharjah, Sharjah, United Arab Emirates

[3]Department of Clinical Nutrition and Dietetics, College of Health Sciences, Research Institute of Medical and Health Sciences, University of Sharjah, Sharjah, United Arab Emirates

[4]Department of Nursing, College of Health Sciences, Research Institute of Medical and Health Sciences, University of Sharjah, Sharjah, United Arab Emirates

Corresponding Author:
Nivine Hanach, PhD
Faculty of Health, Medicine, and Life Sciences, Care and Public Health Research Institute
Maastricht University
Universiteitssingel 40, 6229 ER
Maastricht
Netherlands
Phone: 31 43 388 5655
Email: n.hanach@maastrichtuniversity.nl

Abstract

Background: The postpartum period is a critical phase in a woman's life, marked by various physical, psychological, and social challenges. In light of the rapid proliferation and uptake of digital technologies, particularly in the United Arab Emirates (UAE), mothers increasingly seek informational and emotional support from digital resources. No previous study has thoroughly explored the usage of various digital resources beyond telehealth services in the UAE. This literature gap is particularly relevant for the postpartum period, which remains largely understudied in the UAE.

Objective: This study aims to delve into the digital experiences of postpartum women in the UAE by exploring the types of resources they navigate and the purposes those resources serve. In addition, it seeks to identify their perspectives and needs regarding digital resources that support their postpartum journey.

Methods: Four focus groups were conducted synchronously on the web, involving a total of 27 multicultural mothers (mean age 32.47, SD 4.56 years), between 2 and 12 months post partum and living in the UAE. Descriptive interpretive thematic analysis was used to analyze the data.

Results: Sixteen out of 27 women exhibited severe depressive symptoms at the time of the discussions (Edinburgh Postnatal Depression Scale score of >12). Two main themes were generated from the analysis: (1) Mothers' Experiences with Digital Resources: Participants valued digital resources for providing immediate information, convenience, and support. They primarily used these resources to seek information on infant health, parenting advice, and emotional support through web-based communities. However, the abundance of conflicting information and the pressure to conform to health recommendations often created stress and anxiety. (2) The Perceived Need for Digital Resources: Despite their extensive use of digital resources, mothers articulated the need for a reliable UAE government digital platform tailored specifically to postpartum care, offering trusted information on infant health and postpartum mental well-being. They also emphasized the need for tailored postpartum telemedicine services and moderated web-based discussion forums to foster peer support among mothers.

Conclusions: This study reveals the multifaceted role of digital resources in supporting mothers during the postpartum period, highlighting unmet needs that present opportunities for advancing postpartum care in the UAE. It demonstrates the importance of developing reliable digital solutions for postpartum women, especially regarding mental health and to enhance access to care through tailored telemedicine services. Collaborative efforts are required to ensure the implementation of user-centered digital platforms. Future research should focus on the diverse needs of postpartum women, including cultural sensitivity, the feasibility of telemedicine services, and the integration of partner support in digital interventions to improve maternal health outcomes.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53720) doi:10.2196/53720

KEYWORDS

digital health; social support; telemedicine; postpartum women; focus group; maternal health; postpartum mental health; postpartum depression; emotional support; health information

Introduction

The postpartum period is a critical phase in a woman's life and is characterized by a multitude of physical, psychological, and social challenges [1]. As they adapt and adjust to the demands of this transitional period, mothers may experience ambivalence, anxiety, and depression [2,3]. Previous research has found that receiving sufficient postpartum social support improves maternal and child outcomes [4,5]. Traditionally, mothers would turn to their families, friends, and health care providers for informational and emotional support [6]. When mothers perceive that social support is more readily available, they have greater self-efficacy in their competence to meet the needs of their children [7]. In light of the rapid proliferation and uptake of digital technologies, studies have shown that digital resources are an important growing source of immediate and concise information as well as reassurance and support for postpartum mothers [8,9]. These resources include social media platforms, websites, mobile health (mHealth) apps, short message services, and emails, which may also encompass telemedicine services—the remote administration of health care to patients via telecommunication technologies [10-12].

The integration of digital technologies into postpartum care offers several practical advantages. These include facilitating the understanding of health information presented in different formats, improving connectivity, increasing health care accessibility, and ensuring convenient and timely support for mothers after childbirth [13]. Importantly, digital health care technologies could therefore mitigate geographic obstacles and disparities in health care access, particularly for low-income postpartum women [12,13]. Nevertheless, existing evidence reveals that while digital health may offer a novel opportunity for postpartum women to address their concerns, it also brings about certain drawbacks and considerations. Postpartum women often encounter an overwhelming volume of conflicting information from diverse digital platforms, causing confusion and increased anxiety [14]. Another critical consideration is the digital literacy skills of postpartum women. Effective usage of digital resources relies on the ability to navigate and assess web-based health information accurately to make informed decisions [6]. Moreover, prior studies examining the digital technology needs of postpartum women consistently uncovered new mothers' desire to have improved streamlined communication with their health care providers, a local community with other mothers of similar experiences, and personalized digital features to ensure tailored and relevant resources [15,16].

The United Arab Emirates (UAE) is at the forefront when it comes to the use of digital technologies in the Middle East. In 2024, there were 10.14 million web users in the UAE, representing 99% of the total population [17]. Specifically, the number of mobile phone subscribers with web access reached 9.55 million out of more than 22 million mobile subscribers [17,18].

Telehealth is defined as the delivery and facilitation of health and health-related services via telecommunications and digital communication technologies [10]. A recent study demonstrated that the pandemic has substantially accelerated the adoption of telehealth services in the UAE, with a notable preference for usage among women [19]. Conversely, Ahmed et al [20] revealed remarkable disparities in the usability of telehealth technologies, using information and communication technologies, particularly among Arab and non-Arab women. Arab women exhibited lower levels of usability across all the usability subscales. This implies that the distinctive diversity of the UAE population [21] further adds complexity to the usage of digital technologies and emphasizes the requirement for a comprehensive examination of users' experiences and perceptions to cater digital technologies to their specific needs and cultural norms.

Furthermore, to the best of our knowledge, no previous study has thoroughly explored the usage of various digital resources beyond telehealth services in the UAE. This gap in the literature is particularly relevant for the postpartum period, which remains largely understudied in the UAE and extends to the relatively unexplored realm of usage patterns and needs for digital resources among postpartum women. Therefore, this study aims to delve into the digital experiences of postpartum women in the UAE, including the types of resources navigated and the purposes they serve, in addition to identifying their perspectives and needs for digital resources during the postpartum period. This could ensure alignment between the development of targeted digital interventions and the needs of postpartum women, which in turn could contribute to the enhancement of postpartum care in the digital age.

Methods

Study Design

This study is part of an exploratory, multipurpose exploratory, descriptive focus group study conducted between September 2022 and January 2023. Its primary aim is to comprehensively investigate various aspects of postpartum experiences among mothers in the UAE. Specifically, in this study, we delve into the usage of digital resources and the perceived needs of postpartum women. The choice of an exploratory focus group design is supported by previous research, indicating its suitability for addressing sensitive topics targeting mothers [22] and establishing an understanding of the phenomena of interest [23,24]. Furthermore, the study conforms to Krueger and Casey's [25] guidelines for designing and developing focus group interviews and adheres to the established COREQ (Consolidated Criteria for Reporting Qualitative Research) [26]. In order to bolster the credibility and reliability of the findings

XSL•FO
RenderX

of the studies, researchers have adopted a reflexive approach to the research process, as described by Dodgson [27].

Ethical Considerations

The ethical committee for research ethics at the University of Sharjah approved this study (REC-22-08-29). Written informed consent was obtained from all participants prior to the commencement of the study. Measures were implemented to ensure participant confidentiality, including anonymizing data and securely storing all audio recordings and transcripts. Participants did not receive compensation or direct benefits for their involvement in the study.

Study Setting and Sampling

In this study, participants were purposively recruited from diverse sources including the maternity wards of 2 hospitals in Dubai and Sharjah, where potential participants were initially approached by a nurse after childbirth. Eligible women who showed interest in the study and agreed to be referred to the research team were then informed about the study's objectives, timeline, and process of their involvement by the study researchers. Upon their consent, recruitment was confirmed. Other recruitment sources included social media platforms, such as Facebook and Instagram, where targeted posts and advertisements were used to reach potential participants. In addition, web-based communities and parenting forums that cater to mothers served as platforms for sharing information about the study, allowing for direct engagement with interested participants. Word-of-mouth referrals from participating mothers further facilitated recruitment, and those interested were invited to attend the focus group discussions, thereby ensuring a diverse range of perspectives and experiences were represented.

Participants

Women aged 18-45 years, who had given birth within the last year, and were either UAE citizens or residents of other nationalities were eligible for participation. Women who gave birth prematurely (before 37 weeks of gestation) or whose newborns required intensive care throughout the research period were excluded. There was no restriction on whether the mother had a singleton or multiple pregnancies to establish a thorough understanding of the various postpartum experiences. The study's researchers outlined the study's aims, objectives, and intended involvement process to eligible women before distributing the participant information sheet and written consent form. Three out of 30 women withdrew from the study due to availability and time conflicts. The participants' demographics and characteristics are shown in Table 1.

XSL•FO
RenderX

Table 1. The sociodemographic characteristics of postpartum women in the United Arab Emirates from the 4 focus group discussions.

Parameters	Values
Focus group 1, n	6
Focus group 2, n	8
Focus group 3, n	6
Focus group 4, n	7
EPDS[a] score, n (%)	
Normal (<9)	8 (29.6)
Moderate depressive symptoms (10-12)	3 (11.1)
Severe depressive symptoms (>12)	16 (59.3)
Age (years), mean (SD)	32.47 (4.56)
Parity, n (%)	
Primiparous	14 (51.9)
Multiparous	13 (48.1)
Number of siblings, median (range)	1 (0-3)
Nationality, n (%)	
Arab	14 (51.9)
Western	9 (33.3)
South Asian	4 (14.8)
Expatriate, n (%)	
Yes	24 (88.9)
No	3 (11.1)
Highest education level, n (%)	
High school	3 (11.1)
Technical diploma	2 (7.4)
University degree	22 (81.5)
Employment status, n (%)	
Not employed	9 (33.3)
Part-time	3 (11.1)
Full-time	12 (44.4)
Self-employed	3 (11.1)
Maternity leave , n (%)	
Not employed	9 (33.3)
<3 months	10 (37)
≥3 months	8 (29.6)
Husband's employment status, n (%)	
Not employed	1 (3.7)
Full-time	23 (85.1)
Part-time	__[b]
Self-employed	3 (11.1)
Husband's education level , n (%)	
High school	1 (3.7)
Technical diploma	2 (7.4)

XSL•FO
RenderX

Parameters	Values
University degree	24 (88.9)
Family income, n (%)	
5000 to <10,000 AED (US $1361 to <$2722)	2 (7.4)
10,000 to 15,000 AED (US $2722 to $4087)	__[b]
>15,000 AED (>US $4087)	25 (92.5)
Perception of family income, n (%)	
Not enough	3 (11.1)
Enough	16 (59.3)
More than enough	8 (29.6)
Infant sex, n (%)	
Boy	15 (55.5)
Girl	10 (37)
Twin girls	2 (7.4)
Childbirth type, n (%)	
Vaginal	8 (29.6)
Cesarean	19 (70.4)
Current feeding practices, n (%)	
Breastfeeding	6 (22.2)
Formula feeding	14 (51.8)
Both	7 (26)
Having a nanny, n (%)	
Yes	16 (59.3)
No	11 (40.7)
Having a house helper, n (%)	
Yes	19 (70.3)
No	8 (29.6)
History of psychological disorder, n (%)	
No	23 (85.2)
Yes	4 (14.8)
History of depression, n (%)	
No	18 (66.7)
Yes	9 (33.3)
Complications during childbirth, n (%)	
No	17 (63)
Yes	10 (37)
Complications during pregnancy, n (%)	
Breech presentation	2 (7.4)
Prelabor rupture of membranes	2 (7.4)
Intrauterine growth restriction	1 (3.7)
Gestational diabetes mellitus	3 (11.1)
None	19 (70.4)
Relationship with husband, n (%)	

XSL•FO
RenderX

Parameters	Values
Poor	2 (7.4)
Average	8 (29.6)
Excellent	17 (63)
Primary childcare support, n (%)	
Father	8 (29.6)
Family	3 (11.1)
Nanny	16 (59.3)
Husband support, n (%)	
Rarely	5 (18.5)
Sometimes	6 (22.2)
Usually	5 (18.5)
Always	11 (40.7)

[a]EPDS: Edinburgh Postnatal Depression Scale.
[b]Not available.

Processes

Data Collection

A web-based questionnaire via Google Forms was completed by the participants to obtain an overview of their characteristics including sociodemographic information (eg, age, nationality, education, employment status, financial status, and marital status), obstetric data (type of delivery, complications during pregnancy and childbirth, infant sex, and feeding practices), history of medical and psychological conditions, perceived relationship with the husband, and husband's support during the postpartum period. To identify women exhibiting postpartum depression symptomatology, the questionnaire also included the Edinburgh Postnatal Depression Scale (EPDS) [28]. Each item is assessed on a 4-point scale with a range of 0-3, giving a total score between 0 and 30. Higher scores indicate increased depression symptomatology. A cutoff point of >12 identifies women with depression symptomatology; however, a cutoff point of >9 has been suggested to increase sensitivity for the purposes of community screening [29]. A validated Arabic version of the EPDS with adequate psychometric properties [30] was used with Arabic-speaking participants. To mitigate any potential labeling harm, mothers with high EPDS scores were assured of the nondiagnostic nature of the measure [31], and they were encouraged to seek further assessment in clinical settings, if needed. In addition, it was emphasized that confidentiality is rigorously upheld at all times, and depression scores are neither disclosed nor discussed in the focus groups.

Focus Group Interviews

In accordance with the preference of participating mothers owing to their schedule constraints, the focus group interviews were hosted through a computer video interface using Zoom Software (Zoom Video Communications). Four focus groups were conducted, 2 of which comprised 6 women each, while the remaining groups consisted of 8 and 7 women, respectively. Depending on the number of participants and the extent of discussion required to comprehensively address the emerging topics, the length of the focus group interviews ranged between 90 and 120 minutes. All participants were informed that the interviews would be video recorded before participating in the discussions. In addition, participants were informed that they had the right to withdraw from the study at any time or to omit any of the interviewer's questions without any negative consequences or penalty. Each session was initiated with the facilitator, who has extensive experience in qualitative research involving postpartum women (NH, PhD, female) reviewing the informed consent form and outlining the discussion's ground rules. The facilitator emphasized the importance of each participant's comfort by affirming that there is no "right" or "wrong" response and that all viewpoints are valid. In each focus group, the facilitator (NH) introduced the cofacilitator, an experienced researcher (RS, MSc, female), and acknowledged their roles in the study. The facilitator and cofacilitator were from different cultural backgrounds who maintained neutrality throughout the discussions. This was intended to encourage participants to share their experiences and perspectives openly. Participants were informed that the discussions and data collection were kept strictly confidential and used solely for research purposes. The language spoken in the group discussions was English, as it is extensively used and understood in the UAE, although Arab participants were advised to opt for Arabic if they felt more comfortable expressing themselves in it (however, none did). Based on the existing literature, the research team (NH, NdV, HR, RS, and WBI) collectively developed a semistructured interview guide that covered a series of questions that transitioned from open questions about the mothers' experiences and use of digital resources to more specific questions about their perceived need for digital resources during the postpartum period (Multimedia Appendix 1). Furthermore, the facilitator used probing questions and reflecting statements to deepen the discussion and clarify the participants' responses such as "Could you further explain? Can you tell me more about this experience? Can you give me an example? Let me repeat what you have said." The facilitator then gave a short summary of key points that had been discussed

XSL•FO
RenderX

after each focus group meeting. Before concluding the session, participants were invited to share any additional thoughts and encouraged to ask questions. A concurrent analysis was undertaken, and data collection continued until saturation was achieved and no new themes were identified [32].

Data Analysis

Descriptive quantitative analysis (frequency count and percentages) was performed using Statistical Package for the Social Sciences software (version 26.0; IBM Corp). The discussions were recorded and professionally transcribed verbatim, with additional context provided by field notes taken by the cofacilitator. Thematic analysis was chosen as the method of analysis for the focus groups because of its versatility in accommodating both inductive (data-driven) and deductive (theory-driven) approaches [33]. This allowed the identification of both anticipated and unanticipated themes. Six phases of Braun and Clarke's [34] reflective thematic analysis were used. Two researchers (NH and RS) independently analyzed and developed codes and themes. They read the transcripts multiple times and familiarized themselves with the data. A color-coding method was then applied within ATLAS.ti (version 23; ATLAS.ti Scientific Software Development GmbH) to identify distinct concepts and patterns and form initial codes. Related codes were collated to generate preliminary themes and subthemes and then reviewed by a third researcher (NdV). Any discrepancies were resolved through discussions and consensus (systematic debriefings) among all the study researchers (NH, NdV, HR, WBI, and RS), who were from different cultural backgrounds, ages, sexes, and parity status. By integrating diverse perspectives and interpretations, this has facilitated rigorous and thorough analysis of the data, which in turn has increased the reliability and validity of the study findings. The themes were then refined, defined, and tabulated. The final phase involved writing the report, which provided further clarity for each theme.

Member Checking

To ensure the dependability and transferability of the results, member checking was carried out. After the completion of thematic analysis, each participant was given an opportunity to review the derived themes and subthemes and provide additional insights. All the participants corroborated the findings, which were found to be robust and representative of their experiences; no objections or refutations were raised.

Results

Twenty-seven women, who were between 2 and 12 months post partum took part in 1 of the 4 focus group discussions. Overall, the majority (16/27, 59.2%) exhibited severe depressive symptoms. The demographic and descriptive data for the participating women are outlined in Table 1. Two main themes were identified: (1) mothers' experience with the use of digital resources and (2) the perceived need for digital resources. A thematic map presenting the main themes, subthemes, and participants' insights was derived from the thematic analysis (Multimedia Appendix 2).

Mothers' Experience With the Use of Digital Resources

All participants in the study reported using at least one of the following digital resources during the postpartum period: Google search engine, mobile apps, government websites, and social media platforms. The mothers valued the inherent convenience offered by these digital resources in terms of accessibility and timeliness. A key finding of the focus group discussions was the use of these digital resources to seek information related to parenting and infant concerns or emotional support and validation in web-based communities.

Digital Information–Seeking for Parenting and Infant Concerns

The majority of mothers reported using the Google search engine to obtain immediate health information related to parenting or infant health, especially when they were worried and had limited access to health care providers, such as in the middle of the night. However, some mothers deliberately avoided using the search engine because it provided general advice rather than tailored recommendations and reported having had previous negative experiences that resulted in unnecessary stress and anxiety.

> *[…] What should I do? Where should I start? I googled. And then it showed that my son could be autistic, but it turned out to be a simple developmental delay and now he is completely fine. It put me under a lot of stress and that's why I stopped googling.* [Focus group 3]

Some multiparous mothers expressed that they did not search for as much information as when they had their first child as they perceived themselves to be more experienced.

> *With my second son I don't even read honestly, I don't even need. It's a totally different journey, I don't know, like we get more confident with being a mom, for example I know what's happening, I know that it is a phase and it's gonna pass.* [Focus group 1]

Some expatriate mothers reported the frequent use of government-funded educational websites for parenting and infant health information. Examples of these were the government-sponsored Australian parenting website "Raising Children" and the United Kingdom's National Health Service. They described these websites as noncommercial owing to their provision of credible information by experts about various stages of infant development and infant health. They praised the reliability and trustworthiness of the information on these websites.

> *I haven't lived in Australia for almost 15 years, but there's like a website called Raising children, a government sponsored health site that I feel familiar with and trust its information.* [Focus group 2]

Most mothers also reported the use of mobile apps for tracking developmental milestones, newborn sleep and routine, and receiving advice about the overall infant's health. Some of the highlighted mobile apps were "Little ones," "wonder week," what to expect," and "BabyCenter." These apps were valued for their instantaneous access to information, user-friendly

XSL•FO
RenderX

interface, and personalized features such as the ability to log key details about the infant's sleep, feedings, and daily activities.

> *You get immediate information, and you can log your baby's routine so it indirectly helped my mental health in the sense that it helped me stay organized.* [Focus group 1]

Mothers also turned to social media platforms to seek parenting and infant-related information, with Instagram and TikTok being particularly popular choices. Notably, some mothers indicated that they exclusively followed the accounts of experts and health care providers to ensure access to accurate information, especially from those who shared similar advice as their personal health care providers (eg, pediatricians). Their rationale for this approach was to establish trust in the content they consumed. Conversely, other mothers disclosed that they selectively followed accounts of mothers or experts who aligned with a "similar philosophy of motherhood." They expressed concerns that conflicting viewpoints could significantly affect their mental health by either creating or exacerbating anxiety.

> *Good pediatricians get on TikTok and since every video is about 30 seconds long, they provide quick pieces of information about your baby. You're like, "Oh my God, that makes sense." So, what happens is that when I realize that a specific person, I'm following shares a lot of views with, my daughter's pediatrician, then I start to build trust with them, and I begin to listen closely…not following blindly, but listening closely.* [Focus group 4]

Across all group discussions, participants consistently recognized the potential usefulness of digital resources in accessing regular information. However, they also voiced feelings of overwhelming uncertainty, particularly when discrepancies emerged in shared information on different digital resources. Some mothers further expressed feelings of stress, guilt, disappointment, and a sense of inadequacy when they could not meet the health recommendations, such as exclusive breastfeeding, infant sleep patterns, and routine. Mothers expressed amplified concerns about their infants not meeting expected developmental milestones. They felt pressured to adhere strictly to institutional guidelines, especially first-time mothers. On the other hand, some mothers took a cautious stance by critically evaluating web-based information and relying primarily on their health care providers for guidance.

Emotional Support and Validation in Web-Based Communities

Parenting mobile apps were deemed particularly helpful when they included chat rooms or discussion forums with other mothers who shared similar interests. In these forums, mothers could exchange their experiences and seek advice, if needed. The ability to maintain anonymity within these discussions was highly valued by mothers, as it increased their comfort level and facilitated disclosure of their own experiences. Even mothers who passively consumed information without engaging in discussions found web-based forums to be highly beneficial.

> *And even sometimes I literally go on the app when I need to confirm something, I read the discussions because it's just a group of moms who gave birth around the same time as I did. So, I find that pretty helpful.* [Focus group 4]

One mother reported having joined a paid web-based mother's academy platform led by a clinical dietitian and lactation consultant, which she described as beneficial as it not only provided credible information but also included a mother's forum where they could share their infant feeding experiences, and the health care provider would address any of their concerns. Other mothers reported using Facebook to establish connections and develop social friendships by joining groups of mothers who also shared similar characteristics, such as mothers of twins in Dubai, Australian mothers, and British mothers. This was represented as a form of dealing with isolation and a lack of social support, especially among expatriate mothers who have no family or friends living nearby to reach out to.

> *I joined Facebook groups because I don't have friends here and it has been good because it's people that just come from the same kind of mindset and culture as me and we understand each other.* [Focus group 1]

When mothers were asked about the common topics or areas of interest discussed on these mothers' forums, they mentioned breastfeeding, infant routines, sleeping patterns, postpartum mental health, and occasionally, sensitive personal topics.

The Perceived Need for Digital Resources

Mothers were asked about their perceived need for digital resources in terms of the type, platforms, content, and potential interactive features. The group discussions shed light on the preferences and priorities of various digital tools for supporting their postpartum journey.

Government-Led Digital Resources

Across the 4 focus group discussions, mothers expressed their need to have a well-established UAE governmental digital platform, in the form of either a website or a mobile app, that could serve as a reliable source of information to support their decision to care for their infant.

> *Educational information related to child development and health, so I can use that information to justify my decision to people around me.* [Focus group 2]

Mothers who shared their experiences with the use of government-backed websites in their respective countries were asked about the specific utilities offered by these websites. A key feature of these resources is their accessibility and trustworthiness. They valued the credibility of information, knowing that it had been reviewed by medical experts. In addition, these mothers highlighted the use of the web-based symptom checker. They clarified that this service offers guidance, rather than formal medical diagnoses, for individuals experiencing any symptoms and directs them to a health adviser as needed. Although not explicitly tailored to postpartum mothers, this service was perceived as highly beneficial as it could address and alleviate concerns related to infant health commonly experienced by mothers during the postpartum period. Therefore, mothers emphasized the importance of implementing a comparable governmental digital platform for

XSL•FO
RenderX

postpartum care within the UAE, specifically tailored to meet the unique needs of postpartum women.

When mothers were probed about the specific information they are intrigued to have access to on such hypothetical governmental websites during the postpartum period, they emphasized the need for educational material in the form of concise articles addressing topics related to their own mental well-being, offering guidance on adapting to changes in the postpartum period, and promoting self-care, especially when experiencing "self-identity crises." Mothers notably expressed the desire to gain knowledge about when and where to seek mental health support in the postpartum period.

> *More awareness about mental health and what a mother goes through during the postpartum period, when things might be more alarming when you need to seek support? Professional Support?* [Focus group 3]

Postpartum Telemedicine Services

Mothers mentioned additional digital features and interactive components as imperative to support their postpartum experience. The ability to have remote access to a health care provider, particularly via web-based consultation, came up often across all group discussions. Some mothers expressed that their ideal preference would be to have a "*one-stop-shop*" mobile app with access to information and an embedded virtual consultation service with a UAE health care provider. They further declared that having this option can save their transportation time to the clinic, especially for minor inquiries when no physical examination is needed. One mother pointed out that her child's pediatrician is frequently fully booked and difficult to reach, making virtual consultations extremely valuable whenever she has health-related questions. Some mothers' desire for a virtual consultation with a health care provider stems from the fact that this service has been highly promoted during the COVID-19 pandemic. However, these mothers conveyed their aspiration for the extension of such virtual health care accessibility for postpartum women.

> *Because of living in Dubai, we have learned something from the online world in COVID which was truly helpful. Being at home and having the ability to access doctors by phone or even working from home which made life more convenient because sometimes we simply don't have the time.* [Focus group 4]

> *Time is incredibly valuable here; we are constantly juggling many tasks. Saving even just 30 minutes can make a difference. For example, I calculate that it takes me 30 minutes to get to the doctor and another 30 minutes to return, so that's an hour spent on travel alone. So, it is two hours and in two hours I can accomplish a lot.* [Focus group 3]

Furthermore, some mothers suggested that having a live chat with a health care provider offering immediate responses to mothers' concerns and urgent inquiries would support their postpartum journey and ease their worry, especially during nighttime hours when access to health care services can be challenging. Others proposed that for general inquiries that do not require an instant response, an "ask the expert" forum where they post their questions and receive reliable answers from a health care provider would be extremely beneficial.

> *The ability to ask a question, not in a one-on-one consultation, potentially in like a Q&A, a question that you're struggling to find an answer for, to actually have it answered by a professional.* [Focus group 1]

Mothers expressed a preference for virtual access to various health care providers, including pediatricians, lactation consultants, clinical dietitians, child psychologists, speech therapists, and occupational therapists. The demand for lactation consultants arose from the challenges many mothers faced in receiving adequate breastfeeding support and finding local lactation consultation services. Mothers highlighted the importance of having access to clinical dietitians to assist with infant feeding practices such as breastfeeding and complementary feeding, as well as postpartum weight management. In addition, mothers frequently mentioned their desire to have access to a therapist who specializes in postpartum mental health. Given that some mothers perceived therapy to be relatively expensive in the UAE, they underlined the cost-effectiveness of such access.

> *Going and seeking out therapy is like, so difficult. It's expensive, first of all, but also like, it is just like the last thing on your mind when you have a new baby. So, I think that would be really important.* [Focus group 4]

Some mothers also noted that having access to a therapist through a mobile app could surmount cultural barriers and stigma associated with seeking mental health support. Others suggested the creation of a digital directory of UAE mental health professionals who specialize in maternal mental health, in case virtual consultation is not a feasible service, as they often find it difficult to locate specialized therapists. To enhance the usability of such a directory, mothers recommended including the cultural backgrounds of the therapists, which would facilitate finding a therapist whose cultural background aligns with their own.

> *I won't be judged if I am speaking with a therapist because they might not even know.* [Focus group 4]

Web-Based Discussion Forums With Other Mothers

The need for web-based discussion forums with other mothers was frequently mentioned in all 4 focus group discussions. These forums were valued for maintaining a sense of connection and fostering a community of peer support among mothers. They explained that having access to a group of mothers stratified by interest or characteristics such as infant age, feeding practices, cultural background, and so forth, is comforting and offers rewarding emotional support. One mother articulated a preference for joining a group of experienced mothers instead of reaching out to a health care provider via a web-based consultation, as she believed that a health care provider is not always "*emotionally supportive*" particularly when something

XSL•FO
RenderX

contradicts the guidelines. She further elaborates by giving the example of cosleeping.

> *I'll give you an example. Okay, so, I co-sleep with my baby, he will not sleep in any other way. And I have to sleep with him literally, a pediatrician will tell you that's wrong, it's against safe sleep, like all these things, right? pediatricians don't believe in that kind of stuff. Only another mom who is co-sleeping her child would truly understand.* [Focus group 4]

Conversely, 2 mothers expressed reservations about mothers' groups, such as chat rooms or forums, highlighting some potential drawbacks. They mentioned that these web-based communities often disseminate misleading information and unsolicited advice, which can sometimes lead to a competition among mothers.

> *Talking with other mothers in places like chat rooms or forums ends up with a lot of spiel and competition talk. It's kind of like what you see on Instagram, you know? They are not experts.* [Focus group 1]

Discussion

Principal Findings

This study aimed to explore the digital experiences of postpartum women in the UAE and identify their needs for digital resources during the postpartum period. The findings revealed that mothers valued digital resources for providing immediate information, convenience, and support. They used them to seek infant health information and parenting advice and to obtain emotional support through web-based communities. In addition, mothers expressed the need for a reliable UAE government digital platform tailored to postpartum care, offering trusted information on infant health and postpartum mental well-being. They also emphasized the importance of accessing web-based discussion forums with other mothers to foster peer support and the need for postpartum telemedicine services.

Consistent with previous qualitative studies [35], the need for information is prominent during the first year following childbirth. Our findings demonstrate that mothers' primary use of digital resources is to access information related to infant care and infant health. This is particularly pertinent among primiparous women, who may perceive a greater lack of information than multiparous women, whose need for information appears to decrease and become more specific for subsequent children [8]. In addition, Slomian et al [36] revealed that mothers tend to turn to the web for information significantly more frequently during the postpartum period than during pregnancy. It is hypothesized that pregnant women initially seek and discuss information with their midwives and gynecologists. This discrepancy in information-seeking behavior implies a potential gap in the provision of information by health care providers throughout the antenatal period [36].

Socioeconomic status is asserted to play a moderating role in the use of digital resources. Higher socioeconomic status is associated with greater access to web-based health information [37]. Conversely, Guerra-Reyes et al [38] found that low-income postpartum women rely on web-based sources to obtain self-care and infant health information. However, a particular concern arises in the form of the digital divide, which may also occur because of lower digital literacy in lower-educated populations [39]. Postpartum women with lower educational levels tend to seek reassuring health information, whereas those with higher educational levels tend to exhibit more discernible web-based information-seeking behaviors by searching for trustworthy information on accredited and authoritative websites [38]. The latter could also explain the extensive need of our participating mothers, who mainly had high levels of education, for trustworthy and credible information led by government authorities or health professionals. However, it is important to acknowledge that when the need for information in response to a health concern is perceived as pressing, mothers become less concerned about misinformation by merely searching for validation or guidance, irrespective of their education level [6]. Therefore, promoting science and health literacy should be prioritized to empower postpartum women with the necessary skills to effectively navigate evidence-based web-based health information and make informed health decisions.

While the mothers in this study emphasized the need for informational support during the postpartum period, they equally described their need for emotional, affirmational, and social network support. To address these needs, mothers' support-seeking behaviors expanded to web-based communities and support groups facilitated by various digital platforms, including mobile apps' postpartum forums and Facebook groups. Peer support is defined as the exchange of resources and experiences between like-minded peers with similar experiences [40]. Accessibility to web-based peer support groups is deemed particularly helpful for mothers when navigating the increased demands of infant care after childbirth [41]. Besides offering flexibility, web-based support groups can empower mothers by fostering a sense of normalcy and competence, improving their feelings of self-determination, and promoting social engagement [42,43]. Although the mothers in this study highly valued these virtual support groups, they also drew attention to potential negative aspects, including the dissemination of inaccurate information, misleading advice, and intense arguments. Accordingly, Xie et al [43] suggested that midwives, trained on communication and social skills, moderate web-based parenting communities with the aim of providing both informational and emotional support. Nonetheless, despite the "drama," mothers perceive that the benefits of these web-based forums outweigh the disadvantages [41]. However, it is worth noting that further research is needed to determine effective strategies to moderate and enhance the quality of virtual mother groups.

During group discussions, as mothers elaborated on their use of digital resources during the postpartum period, they highlighted their extensive access to information on infant care and health. Interestingly, when the conversation shifted toward their perceived need for digital resources, mothers expressed a paucity of information on postpartum mental health. They conveyed their desire to acquire sufficient knowledge on this topic, including guidance on when and how to seek local professional support. These insights were similarly shared by mothers with postpartum depression, who identified a

XSL•FO
RenderX

knowledge gap regarding postpartum depression and available care options [15]. Given that the majority of women in our study also exhibited severe depressive symptoms (EPDS >12) at the time of discussion, this could potentially demonstrate the unmet needs of postpartum women in terms of their mental health. The lack of knowledge and understanding of postpartum mental health, as well as attitudes and beliefs surrounding treatment, has been consistently linked to stigma, which poses a significant barrier to mothers' help-seeking behavior [44]. A counterargument to the previously uncovered observations is that numerous eHealth and mHealth resources are available on the web for postpartum mental health. While this assertion holds partial validity, the underlying issue pertains primarily to the quality of web-based information on postpartum mental health. An assessment of 14 smartphone apps using the Silberg Scale revealed an average score of 3.0 (SD 1.52) out of 9 points, indicating poor information quality about postpartum depression and anxiety [45]. Similarly, when more than 20 postpartum mental health websites were evaluated, the vast majority scored low to moderate on the DISCERN tool [46,47]. This was attributed to the absence of information sources and insufficiently detailed descriptions of postpartum mental health treatment options and their benefits [47]. Furthermore, low actionability scores indicate a lack of focus on providing tools and clear instructions for supporting action-taking and informed decision-making among postpartum women [46]. Therefore, it is imperative to consider the psychoeducational needs of postpartum women when designing and developing digital interventions that focus on usability and information quality.

In this study, virtual access to health care professionals was underlined as a prominent digital need among postpartum women. This finding is not surprising in the context of the UAE, where telehealth services have gained substantial recognition and rapid adoption, particularly since the onset of the COVID-19 pandemic [48]. Recent UAE studies have evaluated the usability of telemedicine consultations and patient perceptions in the general population, revealing high satisfaction levels [20,49,50]. However, significant disparities in satisfaction levels emerged when different demographic and contextual factors were considered. For instance, non-Arab participants reported higher levels of satisfaction than Arab participants in terms of the usefulness, ease, and effectiveness of teleconsultations [20]. In addition, specific factors, such as sex, education, frequency of use, delivery format, and teleconsultation duration, contribute to variations in patient satisfaction [20,49,50]. When comparative survey research was conducted in the United States among postpartum women regarding the use and quality of virtual care following childbirth, more than 80% (184/231) of postpartum women reported a high quality of care defined as convenient, easy, safe, and with good information [51]. However, specific concerns have been highlighted that impact their satisfaction, including feeling less connected to their providers, not getting the same level of care as they would in person [51,52], and the unavailability of their providers to discuss postpartum depression, lactation challenges, and contraception [53]. Interestingly, this could be a plausible explanation for our findings, where mothers frequently emphasized the need for virtual access to specialized therapists and lactation consultants, but none mentioned the need for access to obstetrician-gynecologists after childbirth. Unfortunately, this situation parallels that in actual care, as prior research has shown that mothers often experience disappointment and dissatisfaction with their health care providers in addressing their postpartum mental health and emotional needs, causing them to refrain from discussing their concerns [54]. In the postpartum period, virtual consultations can improve continuity of care, lactation support, and counseling on contraception, in addition to screening and treatment for mental health problems [55,56]. To date, no telemedicine interventions have been specifically designed for postpartum care in the UAE. Therefore, it is essential to implement and evaluate telemedicine interventions that address the specific needs of mothers and account for potential cultural and demographic disparities to ensure equitable access and satisfaction and, in turn, optimize the delivery of virtual care in the country.

Strengths and Limitations

To the best of our knowledge, this study represents the first investigation in the UAE to explore the digital needs of postpartum women. Furthermore, this study serves as a continuation of a qualitative inquiry into the perceived mental health experiences and needs of mothers after childbirth. By using synchronous web-based focus group discussions via videoconferencing, we were able to ensure geographic accessibility and recruit participants from various emirates, including Sharjah, Dubai, and Abu Dhabi. This inclusiveness allows for a comprehensive understanding of diverse experiences, considering the variations in health care systems across these regions. The web-based format was convenient and flexible for the participating mothers, allowing them to engage from their preferred locations and even breastfeed their infants during discussions. Such consideration of participant comfort and uninterrupted participation enhances the credibility and dependability of the study's findings.

Several limitations should also be addressed when interpreting the findings of the study. First, the analysis incorporated the perceptions of mothers who were healthy with no medical complications (eg, cardiovascular disease and diabetes), who were married, who were mostly of a high sociodemographic status in terms of education and income, and who gave birth to full-term infants. Therefore, representativeness for all mothers in the UAE cannot be claimed. Second, a significant limitation of this study is the restricted diversity among the enrolled participants. For instance, women of South Asian ethnicity were underrepresented (4/27, 14.8%), despite South Asians comprising 59.4% of the UAE population [57]. In addition, the mothers' unique needs were not disentangled according to the defined postnatal stages (ie, early, intermediate, and late post partum) and severity of postpartum depression symptomatology, which could have influenced their interactions with various digital resources. Another limitation is that the majority of the participants exhibited severe depressive symptoms, potentially leading to different digital resource needs compared with those with no depressive symptoms, thus limiting the generalizability of the findings.

XSL•FO
RenderX

Conclusions, Recommendations, and Implications

This study provided in-depth insights into the experiences and needs of mothers using digital resources during the postpartum period. It also highlights the multifaceted role of digital resources in supporting mothers' postpartum journeys. Postpartum women described a multitude of unmet needs, which subsequently present opportunities to bridge gaps and address these requirements. Of primary interest is the development and implementation of a dedicated digital platform for postpartum care by the UAE government, which aims to improve access to accurate and reliable information, especially concerning postpartum mental health. Collective efforts from health care providers, researchers, science communicators, and policy makers are essential to ensure the design and development of evidence-based content and user-centered digital platforms that align with the preferences and usability requirements of mothers. The study also suggests that future research should investigate the needs of postpartum women from diverse ethnic and cultural backgrounds to develop digital platforms that are not only culturally sensitive but also linguistically adapted, thus ensuring their relevance and effectiveness in serving the UAE's heterogeneous population. In addition, future research should evaluate the digital needs of a more diverse sample of postpartum women to ensure comprehensive and inclusive postpartum digital resources, including those who have experienced preterm birth, have medical conditions, and are socioeconomically disadvantaged. Furthermore, given the evident demand for postpartum telemedicine services, it is imperative to evaluate the feasibility, cost-utility, and cost-effectiveness of offering virtual consultations to postpartum women. This could help estimate the long-term use of telemedicine services and their benefits for maternal health and infant outcomes. Finally, with the significant protective effect of partner support against postpartum mental health problems [58], future studies should investigate the effectiveness of integrating both postpartum women and their partners within digital interventions as a means to enhance partner education and support.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. Data will be available in the form of deidentified transcripts.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group interview guide.
[DOCX File , 20 KB - jmir_v26i1e53720_app1.docx]

Multimedia Appendix 2
A thematic map presenting the main themes, subthemes, and participants' insights derived from the thematic analysis.
[PNG File , 416 KB - jmir_v26i1e53720_app2.png]

References

1. Paladine HL, Blenning CE, Strangas Y. Postpartum care: an approach to the fourth trimester. Am Fam Physician 2019;100(8):485-491 [FREE Full text] [Medline: 31613576]
2. Lupton D. The use and value of digital media for information about pregnancy and early motherhood: a focus group study. BMC Pregnancy Childbirth 2016;16(1):171 [FREE Full text] [doi: 10.1186/s12884-016-0971-3] [Medline: 27435182]
3. Jaks R, Baumann I, Juvalta S, Dratva J. Parental digital health information seeking behavior in Switzerland: a cross-sectional study. BMC Public Health 2019;19(1):225 [FREE Full text] [doi: 10.1186/s12889-019-6524-8] [Medline: 30791927]
4. Sufredini F, Catling C, Zugai J, Chang S. The effects of social support on depression and anxiety in the perinatal period: a mixed-methods systematic review. J Affect Disord 2022;319:119-141. [doi: 10.1016/j.jad.2022.09.005] [Medline: 36108877]
5. White LK, Kornfield SL, Himes MM, Forkpa M, Waller R, Njoroge WFM, et al. The impact of postpartum social support on postpartum mental health outcomes during the COVID-19 pandemic. Arch Womens Ment Health 2023;26(4):531-541 [FREE Full text] [doi: 10.1007/s00737-023-01330-3] [Medline: 37268777]
6. Donelle L, Hall J, Hiebert B, Jackson K, Stoyanovich E, LaChance J, et al. Investigation of digital technology use in the transition to parenting: qualitative study. JMIR Pediatr Parent 2021;4(1):e25388 [FREE Full text] [doi: 10.2196/25388] [Medline: 33595440]
7. Chang YE. Pathways from mothers' early social support to children's language development at age 3. Infant Child Dev 2017;26(6):e2025. [doi: 10.1002/icd.2025]
8. Slomian J, Bruyère O, Reginster JY, Emonts P. The internet as a source of information used by women after childbirth to meet their need for information: a web-based survey. Midwifery 2017;48:46-52. [doi: 10.1016/j.midw.2017.03.005] [Medline: 28324809]

XSL•FO
RenderX

9. Moon RY, Mathews A, Oden R, Carlin R. Mothers' perceptions of the internet and social media as sources of parenting and health information: qualitative study. J Med Internet Res 2019;21(7):e14289 [FREE Full text] [doi: 10.2196/14289] [Medline: 31290403]
10. What is telehealth? NEJM Catal. 2018. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0268 [accessed 2018-02-01]
11. Eijkelboom MCLC, Kalee MM, de Kleijn RAMR, van Wijngaarden JJJ, de Jonge RRR, van der Schaaf MFM, et al. Making knowledge clips with patients: what learning mechanisms are triggered in medical students? Patient Educ Couns 2022;105(10):3096-3102 [FREE Full text] [doi: 10.1016/j.pec.2022.06.008] [Medline: 35725527]
12. Schnitman G, Wang T, Kundu S, Turkdogan S, Gotlieb R, How J, et al. The role of digital patient education in maternal health: a systematic review. Patient Educ Couns 2022;105(3):586-593. [doi: 10.1016/j.pec.2021.06.019] [Medline: 34183217]
13. Lau Y, Wong SH, Cheng LJ, Lau ST. Exploring experiences and needs of perinatal women in digital healthcare: a meta-ethnography of qualitative evidence. Int J Med Inform 2023;169:104929. [doi: 10.1016/j.ijmedinf.2022.104929] [Medline: 36435014]
14. Price SL, Aston M, Monaghan J, Sim M, Tomblin Murphy G, Etowa J, et al. Maternal knowing and social networks: understanding first-time mothers' search for information and support through online and offline social networks. Qual Health Res 2018;28(10):1552-1563. [doi: 10.1177/1049732317748314] [Medline: 29281945]
15. Lackie ME, Parrilla JS, Lavery BM, Kennedy AL, Ryan D, Shulman B, et al. Digital health needs of women with postpartum depression: focus group study. J Med Internet Res 2021;23(1):e18934 [FREE Full text] [doi: 10.2196/18934] [Medline: 33404506]
16. Zingg A, Carter L, Rogith D, Franklin A, Selvaraj S, Refuerzo J, et al. Digital technology needs in maternal mental health: a qualitative inquiry. Stud Health Technol Inform 2021;281:979-983. [doi: 10.3233/SHTI210324] [Medline: 34042819]
17. Official Global Media Insights. UAE internet Stat. 2024. URL: https://www.globalmediainsight.com/blog/uae-internet-statistics/ [accessed 2024-03-13]
18. Open data sets. Telecommunications and Digital Government Regulatory Authority. 2024. URL: https://tdra.gov.ae/en/open-data [accessed 2024-11-16]
19. Swidan A, Alnoon NA, Makki I, Zidan M, Alhammadi H, Rahmani N, et al. Effect of COVID-19 pandemic on patient utilization of the telemedicine services in Dubai. Dubai Med J 2022;5(2):110-116. [doi: 10.1159/000522346]
20. Ahmed F, Zabin RG, Abukoush RS, Alnahoum AA, Alkhawaldeh MY, Al-Yateem N, et al. Usability differences of telehealth technologies by multi-linguistic and multi-cultural users in UAE. 2023 Presented at: IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC); 2023 June 23-26; Torino, Italy p. 569-579. [doi: 10.1109/compsac57700.2023.00082]
21. Fact sheet. The Official Portal of the UAE Government. 2024. URL: https://u.ae/en/about-the-uae/fact-sheet [accessed 2024-11-27]
22. Skelton K, Evans R, LaChenaye J, Amsbary J, Wingate M, Talbott L. Utilization of online focus groups to include mothers: a use-case design, reflection, and recommendations. Digit Health 2018;4:2055207618777675 [FREE Full text] [doi: 10.1177/2055207618777675] [Medline: 29942638]
23. Stebbins RA. Exploratory Research in the Social Sciences. Washington, DC: Sage; 2001.
24. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
25. Krueger RA, Casey MA. Designing and Conducting Focus Group Interviews. Buford, MN: University of Minnesota; 2002.
26. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Dodgson JE. Reflexivity in qualitative research. J Hum Lact 2019;35(2):220-222. [doi: 10.1177/0890334419830990] [Medline: 30849272]
28. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh postnatal depression scale. Br J Psychiatry 1987;150:782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
29. Murray L, Carothers AD. The validation of the edinburgh post-natal depression scale on a community sample. Br J Psychiatry 1990;157:288-290. [doi: 10.1192/bjp.157.2.288] [Medline: 2224383]
30. Abou-Saleh MT, Ghubash R. The prevalence of early postpartum psychiatric morbidity in Dubai: a transcultural perspective. Acta Psychiatr Scand 1997;95(5):428-432. [doi: 10.1111/j.1600-0447.1997.tb09657.x] [Medline: 9197909]
31. Sheehan AM, McGee H. Screening for depression in medical research: ethical challenges and recommendations. BMC Med Ethics 2013;14:4 [FREE Full text] [doi: 10.1186/1472-6939-14-4] [Medline: 23298315]
32. Glaser BG, Strauss A. The Discovery of Grounded Theory. In: Strategies for Qualitative Research. Chicago: Aldine; 1967.
33. Boyatzis RE. Transforming qualitative information. In: Thematic Analysis and Code Development. Washington, DC: Sage; 1998.
34. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]
35. Finlayson K, Crossland N, Bonet M, Downe S. What matters to women in the postnatal period: a meta-synthesis of qualitative studies. PLoS One 2020;15(4):e0231415. [doi: 10.1371/journal.pone.0231415] [Medline: 32320424]

XSL•FO
RenderX

36. Slomian J, Reginster J, Emonts P, Bruyère O. Identifying maternal needs following childbirth: comparison between pregnant women and recent mothers. BMC Pregnancy Childbirth 2021;21(1):405 [FREE Full text] [doi: 10.1186/s12884-021-03858-7] [Medline: 34049520]
37. Nölke L, Mensing M, Krämer A, Hornberg C. Sociodemographic and health-(care-)related characteristics of online health information seekers: a cross-sectional German study. BMC Public Health 2015;15:31 [FREE Full text] [doi: 10.1186/s12889-015-1423-0] [Medline: 25631456]
38. Guerra-Reyes L, Christie VM, Prabhakar A, Harris AL, Siek KA. Postpartum health information seeking using mobile phones: experiences of low-income mothers. Matern Child Health J 2016;20(Suppl 1):13-21 [FREE Full text] [doi: 10.1007/s10995-016-2185-8] [Medline: 27639571]
39. van Deursen A, van Dijk J. Internet skills and the digital divide. New Media Soc 2010;13(6):893-911. [doi: 10.1177/1461444810386774]
40. Shalaby RAH, Agyapong VIO. Peer support in mental health: literature review. JMIR Ment Health 2020;7(6):e15572 [FREE Full text] [doi: 10.2196/15572] [Medline: 32357127]
41. Teaford D, McNiesh S, Goyal D. New mothers' experiences with online postpartum forums. MCN Am J Matern Child Nurs 2019;44(1):40-45. [doi: 10.1097/NMC.0000000000000489] [Medline: 30444739]
42. Bîră M, Buzoianu C, Tudorie G. Social support mediated by technology. A netnographic study of an online community for mothers. Rom J Commun Public Relat 2020;22(2):57-71. [doi: 10.21018/rjcpr.2020.2.300]
43. Xie J, He Z, Burnett G, Cheng Y. How do mothers exchange parenting-related information in online communities? A meta-synthesis. Comput Hum Behav 2021;115:106631. [doi: 10.1016/j.chb.2020.106631]
44. Jones A. Help seeking in the perinatal period: a review of barriers and facilitators. Soc Work Public Health 2019;34(7):596-605. [doi: 10.1080/19371918.2019.1635947] [Medline: 31242074]
45. Zhang MW, Ho RC, Loh A, Wing T, Wynne O, Chan SWC, et al. Current status of postnatal depression smartphone applications available on application stores: an information quality analysis. BMJ Open 2017;7(11):e015655 [FREE Full text] [doi: 10.1136/bmjopen-2016-015655] [Medline: 29138195]
46. Kirby PL, Reynolds KA, Walker JR, Furer P, Pryor TAM. Evaluating the quality of perinatal anxiety information available online. Arch Womens Ment Health 2018;21(6):813-820. [doi: 10.1007/s00737-018-0875-5] [Medline: 29931445]
47. Hardman MP, Reynolds KA, Petty SK, Pryor TAM, Pierce SK, Bernstein MT, et al. An evaluation of the quality of online perinatal depression information. BMC Pregnancy Childbirth 2022;22(1):209 [FREE Full text] [doi: 10.1186/s12884-021-04320-4] [Medline: 35291974]
48. Telemedicine. The Official Portal of the UAE Government. e; 2023. URL: https://u.ae/en/information-and-services/health-and-fitness/telemedicine [accessed 2024-11-16]
49. Al-Sharif GA, Almulla AA, AlMerashi E, Alqutami R, Almoosa M, Hegazi MZ, et al. Telehealth to the rescue during COVID-19: a convergent mixed methods study investigating patients' perception. Front Public Health 2021;9:730647 [FREE Full text] [doi: 10.3389/fpubh.2021.730647] [Medline: 34917570]
50. Al Meslamani AZ, Aldulaymi R, El Sharu H, Alwarawrah Z, Ibrahim OM, Al Mazrouei N. The patterns and determinants of telemedicine use during the COVID-19 crisis: a nationwide study. J Am Pharm Assoc (2003) 2022;62(6):1778-1785 [FREE Full text] [doi: 10.1016/j.japh.2022.05.020] [Medline: 35710898]
51. Marshall C, Gutierrez S, Hecht H, Logan R, Kerns J, Diamond-Smith N. Quality of prenatal and postpartum telehealth visits during COVID-19 and preferences for future care. AJOG Glob Rep 2023;3(1):100139 [FREE Full text] [doi: 10.1016/j.xagr.2022.100139] [Medline: 36465317]
52. Sullivan MW, Kanbergs AN, Burdette ER, Silberman J, Dolisca S, Scarry J, et al. Acceptability of virtual prenatal care: thinking beyond the pandemic. J Matern Fetal Neonatal Med 2022;35(25):8472-8475 [FREE Full text] [doi: 10.1080/14767058.2021.1980534] [Medline: 34554895]
53. Davis A, Bradley D. Telemedicine utilization and perceived quality of virtual care among pregnant and postpartum women during the COVID-19 pandemic. J Telemed Telecare 2024;30(8):1261-1269 [FREE Full text] [doi: 10.1177/1357633X221133862] [Medline: 36398415]
54. Nan Y, Zhang J, Nisar A, Huo L, Yang L, Yin J, et al. Professional support during the postpartum period: primiparous mothers' views on professional services and their expectations, and barriers to utilizing professional help. BMC Pregnancy Childbirth 2020;20(1):402 [FREE Full text] [doi: 10.1186/s12884-020-03087-4] [Medline: 32652965]
55. Hawkins SS. Telehealth in the prenatal and postpartum periods. J Obstet Gynecol Neonatal Nurs 2023;52(4):264-275 [FREE Full text] [doi: 10.1016/j.jogn.2023.05.113] [Medline: 37302795]
56. Hanach N, de Vries N, Radwan H, Bissani N. The effectiveness of telemedicine interventions, delivered exclusively during the postnatal period, on postpartum depression in mothers without history or existing mental disorders: a systematic review and meta-analysis. Midwifery 2021;94:102906 [FREE Full text] [doi: 10.1016/j.midw.2020.102906] [Medline: 33360589]
57. United Arab Emirates-Country Summary. Central Intelligence Agency. 2023. URL: https://www.cia.gov/the-world-factbook/countries/united-arab-emirates/summaries [accessed 2023-06-19]
58. Zorer PB, Akbulut ST, Dirik G. Role of attachment patterns and partner support in postpartum depression. Psikiyatr Guncel Yaklasimlar 2019;11(2):154-166. [doi: 10.18863/pgy.387288]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
EPDS: Edinburgh Postnatal Depression Scale
mHealth: mobile health
UAE: United Arab Emirates

Edited by A Mavragani; submitted 17.10.23; peer-reviewed by E Motrico, G Schnitman, F Medina; comments to author 10.03.24; revised version received 01.04.24; accepted 09.10.24; published 16.12.24.

Please cite as:
Hanach N, Saqan R, Radwan H, Baniissa W, de Vries N
Perceived Experiences and Needs of Digital Resources Among Postpartum Women in the United Arab Emirates: Qualitative Focus Group Study
J Med Internet Res 2024;26:e53720
URL: https://www.jmir.org/2024/1/e53720
doi:10.2196/53720
PMID:39680428

©Nivine Hanach, Roba Saqan, Hadia Radwan, Wegdan Baniissa, Nanne de Vries. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Costs of Anonymization: Case Study Using Clinical Data

Lisa Pilgram[1,2], MD; Thierry Meurers[3], MSc; Bradley Malin[4], PhD; Elke Schaeffner[5], MSc, MD; Kai-Uwe Eckardt[2,6], MD; Fabian Prasser[3], PhD; GCKD Investigators[7]

[1]Junior Digital Clinician Scientist Program, Biomedical Innovation Academy, Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Berlin, Germany

[2]Department of Nephrology and Medical Intensive Care, Charité—Universitätsmedizin Berlin, Berlin, Germany

[3]Medical Informatics Group, Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Berlin, Germany

[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[5]Institute of Public Health, Charité—Universitätsmedizin Berlin, Berlin, Germany

[6]Department of Nephrology and Hypertension, Universitätsklinikum Erlangen, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

[7]See Acknowledgments

Corresponding Author:
Lisa Pilgram, MD
Junior Digital Clinician Scientist Program
Biomedical Innovation Academy
Berlin Institute of Health at Charité—Universitätsmedizin Berlin
Charitéplatz 1
Berlin, 10117
Germany
Phone: 49 30 450543049
Fax: 49 30 450 7 553 9
Email: lisa.pilgram@charite.de

Abstract

Background: Sharing data from clinical studies can accelerate scientific progress, improve transparency, and increase the potential for innovation and collaboration. However, privacy concerns remain a barrier to data sharing. Certain concerns, such as reidentification risk, can be addressed through the application of anonymization algorithms, whereby data are altered so that it is no longer reasonably related to a person. Yet, such alterations have the potential to influence the data set's statistical properties, such that the privacy-utility trade-off must be considered. This has been studied in theory, but evidence based on real-world individual-level clinical data is rare, and anonymization has not broadly been adopted in clinical practice.

Objective: The goal of this study is to contribute to a better understanding of anonymization in the real world by comprehensively evaluating the privacy-utility trade-off of differently anonymized data using data and scientific results from the German Chronic Kidney Disease (GCKD) study.

Methods: The GCKD data set extracted for this study consists of 5217 records and 70 variables. A 2-step procedure was followed to determine which variables constituted reidentification risks. To capture a large portion of the risk-utility space, we decided on risk thresholds ranging from 0.02 to 1. The data were then transformed via generalization and suppression, and the anonymization process was varied using a generic and a use case–specific configuration. To assess the utility of the anonymized GCKD data, general-purpose metrics (ie, data granularity and entropy), as well as use case–specific metrics (ie, reproducibility), were applied. Reproducibility was assessed by measuring the overlap of the 95% CI lengths between anonymized and original results.

Results: Reproducibility measured by 95% CI overlap was higher than utility obtained from general-purpose metrics. For example, granularity varied between 68.2% and 87.6%, and entropy varied between 25.5% and 46.2%, whereas the average 95% CI overlap was above 90% for all risk thresholds applied. A nonoverlapping 95% CI was detected in 6 estimates across all analyses, but the overwhelming majority of estimates exhibited an overlap over 50%. The use case–specific configuration outperformed the generic one in terms of actual utility (ie, reproducibility) at the same level of privacy.

Conclusions: Our results illustrate the challenges that anonymization faces when aiming to support multiple likely and possibly competing uses, while use case–specific anonymization can provide greater utility. This aspect should be taken into account when evaluating the associated costs of anonymized data and attempting to maintain sufficiently high levels of privacy for anonymized data.

XSL•FO
RenderX

Trial Registration: German Clinical Trials Register DRKS00003971; https://drks.de/search/en/trial/DRKS00003971
International Registered Report Identifier (IRRID): RR2-10.1093/ndt/gfr456

(J Med Internet Res 2024;26:e49445) doi:10.2196/49445

KEYWORDS
data sharing; anonymization; deidentification; privacy-utility trade-off; privacy-enhancing technologies; medical informatics; privacy; anonymized; security; identification; confidentiality; data science

Introduction

Sharing data from clinical studies can accelerate scientific progress, improve transparency, and increase the potential for innovation and collaboration [1]. Scientific data sharing has been encouraged by a range of regulatory agencies [2] and is required by many scientific journals [3]. However, there are various challenges to realizing data sharing in practice. For example, the data should satisfy FAIR (findable, accessible, interoperable, and reusable) principles [4], while sharing policies need to comply with relevant privacy laws, such as the European General Data Protection Regulation [5]. Uncertainty in handling personal data is one of the major challenges to collaborative research [6-9].

Privacy-enhancing technologies, including anonymization algorithms, can maintain the privacy of study participants when sharing data [10,11]. Anonymization reduces privacy risks by altering data in a manner such that it is highly unlikely that it can be related to a person. Anonymization can be performed using various transformation mechanisms, such as suppression, randomization, or generalization. Software-enabled solutions have been developed with implementations of published algorithms to support this process [12]. Yet, there is an inherent trade-off between the reduction of privacy risks and the utility of the data that can be shared [13]. In this respect, a key concern is that the amendments needed to maintain privacy at a certain level may adversely influence the inherent statistical properties of the data.

This challenge has been studied extensively in theory [14], and the evidence for utility-preserving anonymization is growing [12,15-18]. However, anonymization has not been broadly adopted in clinical practice. Multiple studies report substantial gaps in data availability and stress the lack of practical guidance [8,9,19-22]. The need for a better understanding is also supported by a review that found most reported successful disclosure attacks on anonymized data were enabled by incorrectly applying anonymization algorithms [23]. In addition, while many approaches have been developed for capturing and reducing privacy risks, these are typically evaluated using general-purpose utility measures and only rarely real-world individual-level clinical data, providing little insights into their performance in real-world applications [21,22,24]. Metrics based on such applications are comparatively less reported [15,25-28], yet greatly needed to gain a better understanding of privacy-utility trade-offs as well as to provide targeted recommendations for data providers.

In this study, we aim to provide a better understanding of the opportunities for sharing individual-level data from clinical studies. Specifically, we investigate how different anonymization algorithms affect the utility in a real-world application using data and scientific results from the German Chronic Kidney Disease (GCKD) study [29].

Methods

Data and Real-World Application

The GCKD study is a nationwide prospective observational cohort to study the natural course of chronic kidney disease (CKD) and to better understand associations between patient characteristics and disease progression [29]. More than 150 outpatient and 11 university-hospital study sites contributed to the recruitment of 5217 patients between March 2010 and March 2012 and subsequent follow-up. Data collection resulted in a high-dimensional data set of more than 4000 variables.

To assess the utility of anonymized data in a real-world application, we studied its performance in downstream analyses. These aimed at describing the disease burden and risk profile of patients with CKD at baseline as previously published by Titze et al [30]. The variables relevant for this application were selected, aggregated, and calculated through multiple preprocessing steps (Figure S1 in Multimedia Appendix 1). The final curated data set was composed of 70 variables and referred to as original for the remainder of this paper.

Threat Model

Based on the assumption that the data will be disclosed through some web-based platform for sharing data from clinical studies [31-33] with additional measures of control in place (ie, data use agreements and compatible legal environment), we assumed a controlled access scenario [34]. In this context, the aim of anonymization was to provide safeguards in case of an accidental disclosure, for example, a breach of the recipient's local security measures [35].

In line with guidelines and recommendations on clinical trial data sharing, we focused on protecting the data from linkage and recognition of subject identities (ie, reidentification) [10]. To detect variables that could be used for reidentifying study participants, a 2-step procedure was adopted. First, a qualitative risk assessment was performed based on international guidelines that document lists of potentially linkable variables [36-39]. Next, a semiquantitative risk assessment was performed by studying the variables' availability (variables likely to be known to adversaries), replicability (variables that occur repeatedly in relationship to the individual), and distinguishability (variables that make, alone or in combination, individuals unique) [40]. This method has been successfully applied for several real-world data sets [41-43]. The scoring system was adapted to our

XSL•FO
RenderX

real-world application according to literature and expert knowledge. In brief, availability, replicability, and distinguishability were quantified from low (1) to high (3), and the sum was calculated as the score per variable. A score of greater than 5 was applied as a threshold for the recognition of a "risky" variable. Overall, we determined that 6 of the 70 variables needed to be protected against reidentification: age, gender, height, weight, BMI, and history of renal biopsy. The underlying reasons and results of the 2-step procedure are provided in Table S1 in Multimedia Appendix 1 [40,44].

We further screened for interdependent relationships between variables in the data set (eg, height, weight, and BMI). Transforming them independently can result in 2 issues. First, the transformed values may no longer be logically consistent. Second, back-calculation may narrow down intended generalization intervals and leak information that undermines established risk thresholds. Among our variables, this was true for the anthropometric data height, weight, and BMI. To account for this, we removed either height and weight or BMI from the data depending on the configuration scenario (see Data Transformation section).

Reidentification Risk Assessment and Thresholds

Following guidelines for clinical trial data anonymization, we quantified and reduced reidentification risks according to probabilistic prosecutor risk (PR) and marketer risk (MR) models [38]. The prosecutor model provides risk estimates under the assumption that the data recipient attempts reidentification against a specific record for which he or she already knows its membership in the data set. Protecting against such attacks also protects the data from reidentification by less knowledgeable data recipients who do not have prior knowledge about membership. By contrast, the marketer model provides an estimate of the average success probability that can be expected of such a less knowledgeable data recipient. Both risk estimates can be calculated from the distinguishability of records in the data set regarding the risky variables. Let $u(r)$ be the number of records indistinguishable from a record r regarding the risky variables (including r itself). Then, the risk of each record is $1/u(r)$. Data set–level risk estimates can be derived from the distribution of the risks of all records, with the PR referring to the maximum and MR to the average of this distribution [45]. A data set can then be protected from reidentification by transforming it in a way that those risk estimates fall below given thresholds. The privacy model aiming at the PR is typically called k-anonymity, whereas the privacy model that addresses a combined view of the PR and MR is called strict-average risk [38]. We studied anonymized data sets with PR and MR thresholds ranging from 1 to 0.02, respectively.

In our analysis, we put a special focus on three risk thresholds: (1) 0.5 (ie, a group size of 2), as this is the greatest risk smaller than 1 that can be measured in approaches built upon distinguishability; (2) 0.09 (ie, a group size of 11), as this is a threshold that has been recommended for sharing data from clinical studies [42,43]; and (3) 0.03 (ie, a group size of 33), as this is the smallest threshold that could be enforced without resulting in additional variables being fully censored in the anonymization process. From these thresholds, 4 privacy levels from moderate to very strict that we highlight in our analyses were derived. We denoted the privacy levels as percentages: (1) 50% PR combined with 9.09% MR, denoted 50% PR+9.09% MR, (2) 50% PR+3.03% MR, (3) 9.09% PR, and (4) 3.03% PR.

Data Transformation

The anonymized data sets were realized by generalizing and suppressing variables using the open-source tool ARX (Institute of Medical Informatics, Statistics and Epidemiology at Technical University of Munich and Medical Informatics Group at the Berlin Institute of Health, Charité—Universitätsmedizin Berlin) [46]. Generalization categorized continuous data into intervals of different sizes (hierarchies) to prevent distinguishability. Its configuration included the definition of hierarchies, grouping factors, and maximum and minimum values (Table 1). Values of 1 variable were transformed consistently (ie, to the same hierarchy level). We chose this process because it simplifies downstream statistical analyses. According to generally accepted rates of missing data for statistical analyses, an overall limit of 10% on the number of records that could be suppressed was specified [47,48]. In the dichotomous variables (ie, gender and renal biopsy), only suppression was applied.

Two different configurations were investigated: (1) a generic scenario that aims to support multiple general medical uses without restriction in the generalization hierarchies applied and (2) a use case–specific scenario in which generalization was restricted in variables that were important for our real-world application [30]. Different strategies were also followed to account for the interdependent relationship between the anthropometric data. In the generic scenario, we transformed height and weight and removed BMI from the data to simulate a situation where it was unknown if BMI would be of relevance to the study. In the use case–specific scenario, we took the relevance of BMI into account and removed height and weight from the data but preserved BMI. Table 1 illustrates the characteristics of the 2 scenarios. In total, we created 200 anonymized data sets based on 100 different risk thresholds in 2 configuration scenarios (Figure S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Differences in generalization between the generic and the use case–specific scenario.[a]

	Generic scenario		Use case–specific scenario	
	Minimum-maximum value	Maximum generalization	Minimum-maximum value	Maximum generalization
Age (years)	15-80	Not defined	18-80	10-year intervals
Height (cm)	20-280	Not defined	Removed	Removed
Weight (kg)	0-160	Not defined	Removed	Removed
BMI (kg/m^2)	Removed	Removed	15-70	<25.0[b]; 25.0-29.9[c]; ≥30.0[d]

[a]To account for collinearity, we generalized height and weight and removed BMI in the generic scenario. In the use case–specific scenario, BMI was generalized, and height and weight were removed. In this scenario, explicit minimum and maximum values were extracted from the original data set, and generalization was restricted in variables that were relevant to our real-world application. The hierarchies have been archived on the web [49].

[b]Underweight or normal weight.

[c]Overweight.

[d]Obesity.

Privacy Assessment of Anonymized Data

Due to the consistent transformation of variables, predefined thresholds did not necessarily translate into the actual risk. We therefore calculated the empirical PR and MR and screened for differences from our predefined thresholds (ie, overprotection). In our figures, we present privacy as a spectrum from 1–maximum PR to 1–average PR (ie, MR) and 1–minimum PR. Apart from maximum PR, we chose to include average and minimum PR in our assessment. Considering the potential to overestimate reidentification risks [23], the average and minimum PR represent important additional guiding factors when implementing real-world anonymization algorithms [43].

Utility Assessment of Anonymized Data

We analyzed general-purpose (ie, generic) utility of the anonymized data sets as well as the degree to which they could be relied upon to reproduce results from the original data describing the disease burden and risk profile of patients with CKD at baseline (use case–specific utility) [30].

To determine general-purpose utility (1) at the cell level, we measured the granularity of the variables in the data set and (2) on the variable level, we applied the nonuniform entropy model that measures deviations in variable distributions. Both were compiled into data set–level measures by averaging their results across all records or all variables, respectively [50-52]. All results were normalized into the range of 0% (all information removed) to 100% (no information modified at all).

To evaluate use case–specific utility, all analyses were performed on the original and the 200 anonymized data sets. To measure reproducibility, we made use of the estimate agreement as described in the context of real-world evidence versus randomized controlled trials [53]. As an estimate agreement, we introduced the relative overlap in 95% CI lengths of the numbers and percentages between the original and anonymized data sets. For this purpose, the proportion or mean 95% CI was determined by the Wilson score interval and 2-tailed *t* test, respectively, and the 95% CI lengths in the anonymized data sets were compared to those in the original data set as proposed by Karr et al [24]. We compiled the relative overlap in 95% CI lengths into table-level measures by averaging all table cells and into data set–level measures (overall average 95% CI overlap) by averaging all analyses including the ones covering only variables that were not affected by the anonymization procedure.

The use case–specific metrics based on 95% CI overlap neglected variables with scale transformation through the anonymization process (ie, age, height, weight, and BMI). In these variables, we compared resulting hierarchy levels and presented their effect on the results visually.

Technical Implementation

ARX (version 3.9.1; published November 2022) was used for anonymization of the data. The data management, analyses, and visualizations were performed using R (version 4.1.0; R Foundation for Statistical Computing), Python (version 3.11; Python Software Foundation), and built-in functions of ARX for general-purpose metrics and the risk models [46].

Ethical Considerations

All methods were carried out in accordance with the Declaration of Helsinki. The GCKD study was approved by local ethics committees (Friedrich-Alexander University Erlangen-Nürnberg, Germany, 3831) and registered in the national registry for clinical studies (DRKS 00003971). Informed consent was obtained from all participants prior to enrollment. The participants did not receive any form of compensation. Approval for this study was covered by the approval of the ethics committees of the GCKD study. Participants' data are stored in pseudonymized form in the study database. The database is on a server at the Regional Computer Centre of the University Hospital in Erlangen. All aspects of data backup and security are based on relevant guidelines and in accordance with the German Federal Data Protection Act.

Results

Empirical Residual Risks

Prior to anonymization, 5112 (98%) records in the original data set were unique regarding the variables that could be used for reidentification. Table 2 presents the empirical PR and MR after having transformed the data set considering different risk thresholds. As can be seen, the process of consistent

XSL•FO
RenderX

transformation often resulted in overprotection, with minimal and average PR (ie, MR) being below the specified thresholds.

Table 2. Empirical minimum, average, and maximum prosecutor risk (PR) and marketer risk (MR).[a]

	Generic scenario			Use case–specific scenario		
	Maximum PR, %	Average PR, % (ie, MR)	Minimum PR, %	Maximum PR, %	Average PR, % (ie, MR)	Minimum PR, %
50% PR+9.09% MR	50	8.7	0.7	50	6.9	0.5
50% PR+3.03% MR	33.3	3	0.2	50	2.5	0.3
9.09% PR	9.1	1.4	0.1	9.1	1.6	0.3
3.03% PR	2.9	0.6	0.1	3	0.9	0.2

[a]We report results for the following four risk thresholds: (1) 50% PR+9.09% MR, (2) 50% PR+3.03% MR, (3) 9.09% PR, and (4) 3.03% PR. It can be seen that empirical risks can be lower than the specified risk thresholds due to consistent data transformation (ie, overgeneralization).

Privacy-Utility Trade-Off

Next, we studied how well the anonymization approach enabled trading off data privacy against utility. Figure 1 presents privacy-utility trade-off curves when scaling risk thresholds for PR against granularity and entropy as general-purpose utility metrics. Privacy was calculated as 1–empirical PR for minimum, average (ie, MR), and maximum PR, respectively. Figure 2 presents analogous curves for use case–specific utility metrics (ie, 95% CI overlap). When scaling risk thresholds for MR in 50% PR+MR, similar results were observed, such that they are deferred to Figures S2 and S3 in Multimedia Appendix 1.

As can be seen from the results shown in Figure 1, the curve is flatter between the 50% PR and 9.09% PR threshold than between 9.09% PR and 3.03% PR. Thus, a gain in privacy was accompanied by a comparatively lower loss in utility across this risk-utility space. At lower privacy levels than the 50% PR threshold, a high initial loss was observed in entropy but not in granularity. It can be seen from these results that the process had a nontrivial impact on variable distributions, pushing it toward a greater amount of privacy than (general-purpose) utility. For example, in the generic scenario, granularity varied between 87.6% (50% PR+9.09% MR) and 68.2% (3.03% PR), while entropy was generally lower with estimates between 46.2% and 25.5%, respectively.

The use case–specific utility is presented as an overall 95% CI overlap and as an 95% CI overlap on the analysis level in Figure 2 when scaling thresholds for PR and in Figure S3 in Multimedia Appendix 1 for 50% PR+MR. Compared to Figure 1 (general-purpose utility), privacy gain could be achieved with a minor impact on utility in this case. The overall 95% CI overlap in the generic scenario varied from 98.4% at 50% PR+9.09% MR to 96.7% at 3.03% PR.

Figure 3 illustrates the differences between the applied utility metrics to point out the multidimensionality of utility. In our real-world application, results from use case–specific metrics were consistently above the ones from general-purpose metrics.

Figure 1. Privacy-utility curves based on general-purpose utility metrics. Granularity and nonuniform entropy served as general-purpose utility metrics. Privacy is demonstrated as 1–empirical maximum PR, average PR (ie, MR), and minimum PR. We used the anonymization processes implementing thresholds on PR for generating the points on the curve: 50% PR, 9.09% PR, and 3.03% PR. Results of granularity in the (A) generic and (C) use case–specific anonymized data sets and results of entropy in the (B) generic and (D) use case–specific anonymized data sets are shown. Results for 50% PR+MR were analogous and are illustrated in Figure S2 in Multimedia Appendix 1. The extreme points at (0,100) and (100,0) have been added to the graph but were not directly measured. MR: marketer risk; PR: prosecutor risk.

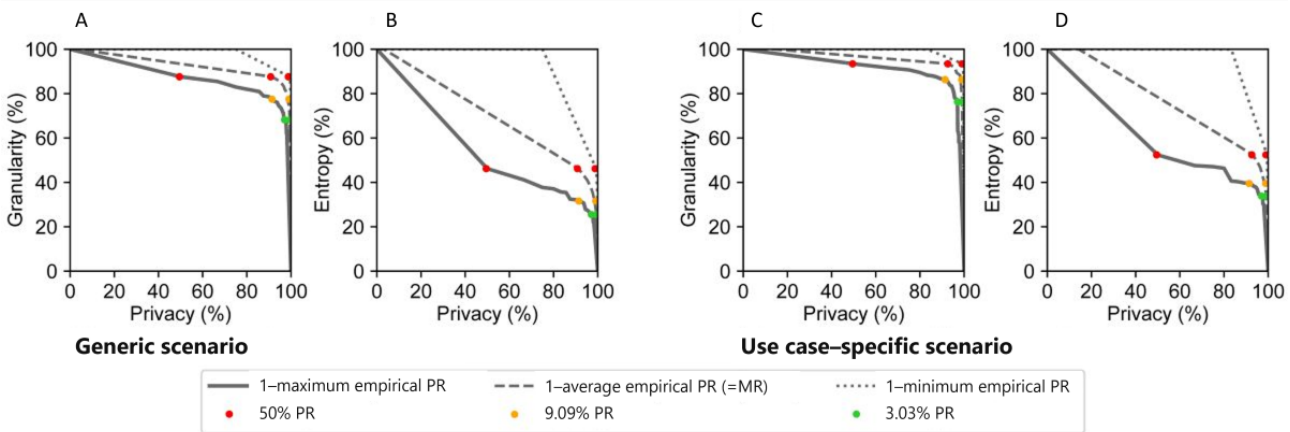

XSL•FO
RenderX

Figure 2. Privacy-utility curves using use case–specific utility metrics based on 95% CIO. 95% CIO was calculated on the data set level (overall 95% CIO) and analysis level. Two analyses (glomerular filtration rate and albuminuria categories and comparison of estimated glomerular filtration rate equations) were not affected by anonymization at all (100% overlap) and are therefore not displayed separately. Privacy is demonstrated as 1–maximum PR, average PR (ie, MR), and minimum PR. We used the anonymization processes implementing thresholds on PR for generating the points on the curve: 50% PR, 9.09% PR, and 3.03% PR. Results of the overall 95% CIO in the (A) generic and (G) use case–specific anonymized data sets and results of the 95% CIOs on analysis level in the (B-F) generic and (H-L) use case–specific anonymized data sets are shown. Results at the estimate level are shown in Tables S2-S10 in Multimedia Appendix 1. Results for 50% PR+MR were analogous and are illustrated in Figure S3 in Multimedia Appendix 1. The extreme points at (0,100) and (100,0) have been added to the graph and were not directly measured. CIO: CI overlap; MR: marketer risk; PR: prosecutor risk.

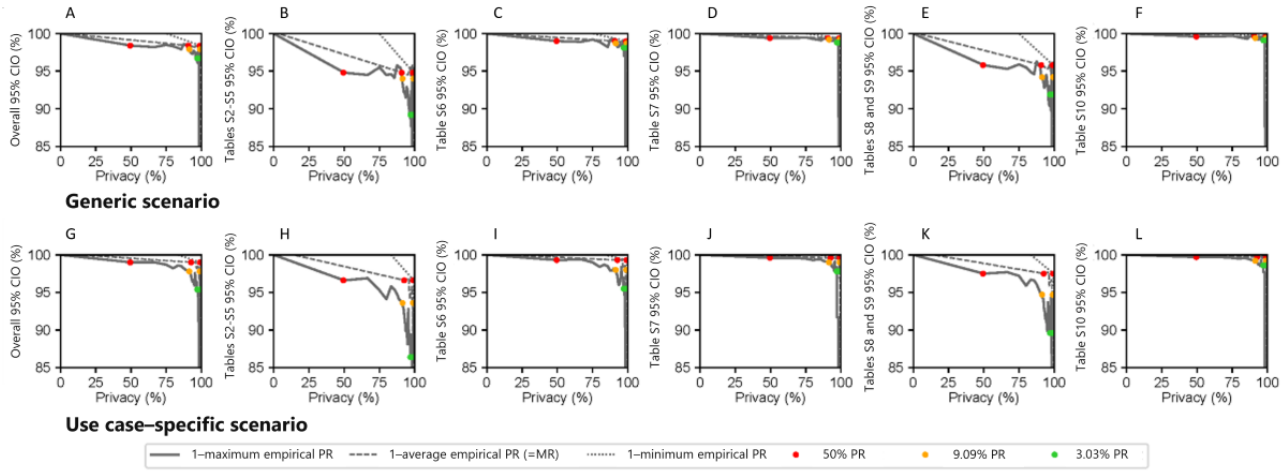

Figure 3. Generic and use case–specific utility metrics. Calculated utility metrics are illustrated in comparison. Granularity and nonuniform entropy served as general-purpose utility metrics. 95% CIO was calculated on the data set level (overall 95% CIO) and analysis level. The latter is exemplary illustrated for the main analysis (Tables S2-S5 in Multimedia Appendix 1, 95% CIO). 95% CIO excluded variables with scale transformation. Privacy is illustrated as 1–maximum PR. We calculated metrics in (A) generic and (B) use case–specific anonymized data sets. Results for 50% PR+MR were analogous and can be drawn from the privacy-utility curves in Figures S2 and S3 in Multimedia Appendix 1. The extreme points at (0,100) and (100,0) have been added to the graph and were not directly measured. CIO: CI overlap; PR: prosecutor risk.

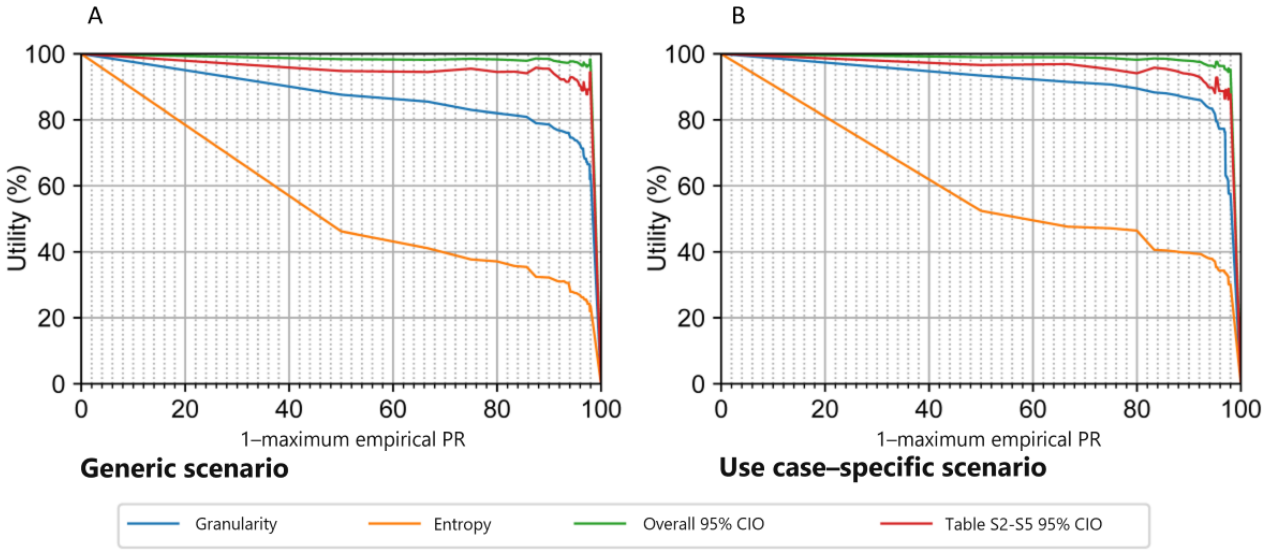

Reproducibility of Prior Study Results

A more detailed analysis of reproducibility was performed by comparing analyses on estimate level for the selected privacy levels: (1) 50% PR+9.09% MR, (2) 50% PR+3.03% MR, (3) 9.09% PR, and (4) 3.03% PR. We conducted 7 analyses to describe the disease burden and risk profile of patients with CKD at baseline: the disease burden and risk profile stratified by gender and presence of diabetes mellitus (main results, Tables S2-S5 in Multimedia Appendix 1 [24]), characteristics stratified by inclusion criteria (Table S6 in Multimedia Appendix 1), biopsy rates per leading cause (Table S7 in Multimedia Appendix 1), cardiovascular disease burden stratified by gender and the presence of diabetes mellitus (Tables S8 and S9 in Multimedia Appendix 1), the characteristics stratified by diabetes mellitus and diabetic nephropathy (Table S10 in Multimedia Appendix 1), the distribution of glomerular filtration rate and albuminuria categories, and the comparison of estimated glomerular filtration rate (eGFR) equations. The last 2 analyses are not shown as they only covered variables that were not affected by the anonymization procedure resulting in a 100% 95% CI overlap. Additional information on the presumed cause

XSL•FO
RenderX

of CKD, patient awareness, and the age distribution of patients stratified by gender and the presence of diabetes mellitus was calculated and illustrated as figures. While no variable was affected by the anonymization process in the first 2 analyses (results not shown), the effects of anonymization for the last one are depicted in Figure 4.

We focused on the reproducibility of the main results (Tables S2-S5 in Multimedia Appendix 1). This included the 95% CI overlap and whether the result of anonymized data was within the original 95% CI (Figure 5 and Tables S2-S5 in Multimedia Appendix 1). The main results were stratified by gender and the presence of diabetes mellitus. For the subset of female participants who did not have diabetes, the results of the 95% CI overlaps at estimate level are shown in Tables S2 and S3 in Multimedia Appendix 1 as well as Figure 5. The original data set included 1462 female participants who did not have diabetes. Due to suppression in the anonymization process, the number decreased at (1) 50% PR+9.09% MR, (2) 50% PR+3.03% MR, (3) 9.09% PR, and (4) 3.03% PR to 1385, 1407, 1360, and 1309, respectively, in the generic scenario and to 1414, 1451, 1342, and 1218, respectively, in the use case–specific scenario. We detected modestly overlapping 95% CIs (<50%) in estimates of the variables renal biopsy (12.4%), urine albumin-to-creatinine ratio (UACR)>300 mg/g (46.2%), and eGFR (47.4%) at 3.03% PR in the generic scenario. In the use case–specific scenario, this was true for estimates of the variables UACR>300 mg/g (19.6%), UACR<30 mg/g (23.9%), eGFR≥60 mL/min (41.6%), eGFR (21.6%), and systolic blood pressure (41.7%) at 3.03% PR. At this privacy level, there was also a nonoverlapping 95% CI (0%) measured for renal biopsy (95% CI 22.2-27.1 vs 29.7-34.5). The modest and nonoverlapping 95% CIs were accompanied by results that were not within the original 95% CI. At lower privacy levels, there were no such constraints. When looking at the other subsets (Tables S4 and S5 in Multimedia Appendix 1), the 95% CI overlap at the estimate level revealed similar results with sporadic modestly (n=7) and nonoverlapping (n=3) 95% CIs.

Considering all analyses, the main results and the results on cardiovascular disease burden were most influenced by a lower 95% CI overlap (Figure 2B, E, H, and K). This is most likely due to the stratification by the variable gender and the large amount of further modified variables in these analyses. Completely overlapping 95% CIs (100%) were reached for 2 of the 7 analyses (glomerular filtration rate and albuminuria categories and comparison of eGFR equations) and for 2 of the 3 figures (presumed cause of CKD and patient awareness). In these analyses, there was no variable affected by the anonymization process, and the results are therefore not shown separately. The trends of affected tables (Tables S6-S10 in Multimedia Appendix 1) at the estimate level were similar to the main results. Two analyses (Tables S7-S9 in Multimedia Appendix 1) could be replicated without modestly overlapping 95% CIs. Within the other tables, estimates exhibited sporadic modestly (n=11) and nonoverlapping (n=2) 95% CIs but by far the majority of estimates exhibited a 95% CI overlap of over 50% across all privacy levels.

The age, height, weight, and BMI variables were converted from a numerical to a categorical scale during the anonymization process. In the use case–specific scenario, the degree of generalization in scale transformation was preconfigured to preserve relevant information. In the generic scenario, there were no restrictions in the generalization hierarchies. The resulting loss of information is visualized in Figure 6 for age and BMI for the subset of female participants who did not have diabetes. In the generic scenario, generalization of age leads to 20-year intervals at 50% PR+3.03% MR, 9.09% PR, and 3.03% PR. In contrast, as predefined, the intervals did not exceed 10 years in the use case–specific scenario. For BMI, generalization was more complex. In the generic scenario, BMI was calculated using the generalized data of height and weight (Figure S5 in Multimedia Appendix 1), which resulted in diverse, partly overlapping, intervals. As demonstrated in Figure S4 in Multimedia Appendix 1, the number of intervals decreased with increasing protection, while their length increased. This resulted in relevant information loss with intervals covering a range from normal weight (24.7 kg/cm^2) to severe obesity (≥40 kg/cm^2). As for age, the use case–specific scenario had restrictions in generalization hierarchies for BMI, which resulted in commonly accepted categories (normal weight, overweight, and obesity) and a good approximation to the original distribution. Reasonable semantics were maintained even at 3.03% PR. Thus, use case–specific configurations were important to obtain reasonable semantics of the variables. Height and weight were considered less relevant for the research focus and therefore removed in favor of preserving the BMI in the use case–specific scenario.

We additionally plotted information on the age distribution stratified by gender and the presence of diabetes mellitus for the original and the anonymized data at the selected privacy levels (Figure 4). In the generic scenario, the figure could only be replicated at 50% PR+9.09% MR due to the large intervals when stricter risk thresholds were enforced. By contrast, the use case–specific scenario maintained the original interval length.

XSL•FO
RenderX

Figure 4. Age distribution stratified by gender and the presence of diabetes mellitus in the original and anonymized data sets. Anonymization was applied as defined in the (A and B) generic and (C and D) use case–specific scenario. Bar plots illustrate counts for anonymized data at selected privacy level: 9.09% MR+50% PR, 3.03% MR+50% PR, 9.09% PR, and 3.03% PR. The figure derived from the original data is illustrated in gray. MR: marketer risk; PR: prosecutor risk.

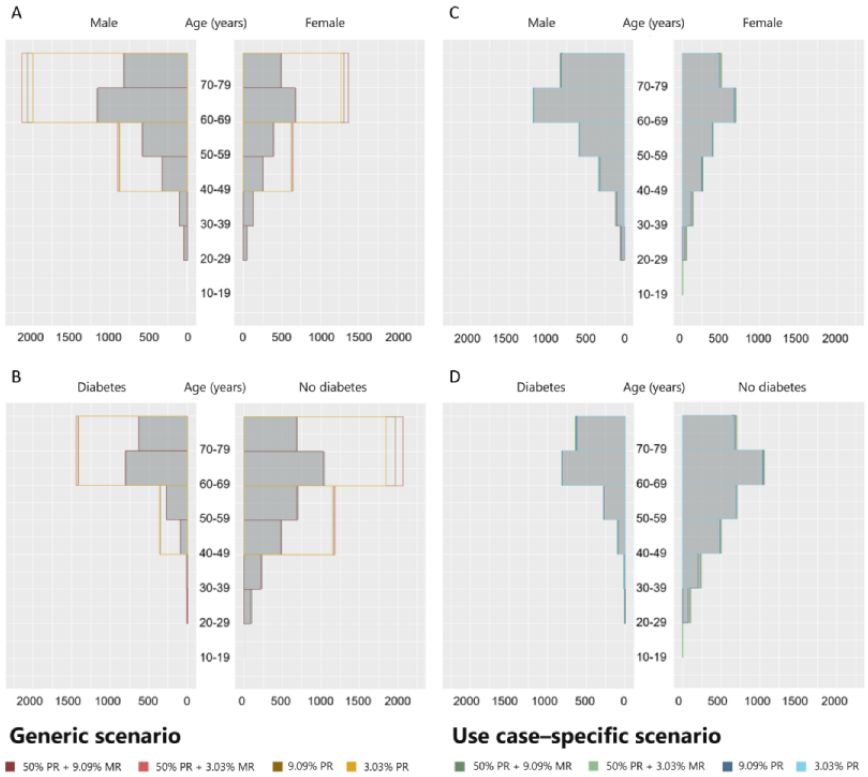

Figure 5. Proportion, CIs, and overlap in the interval lengths for descriptive analyses of the subset of female participants who did not have diabetes. Anonymization was applied as defined in the (A) generic and (B) use case–specific scenario. Results are shown for selected privacy levels: 9.09% MR+50% PR, 3.03% MR+50% PR, 9.09% PR, and 3.03% PR. Only categorical parameters are presented as percentages referred to the numbers excluding missing with proportion 95% CI. 95% CI for both original and anonymized data were calculated based on the Wilson score interval and are displayed in the figure. For the original data, 95% CI is illustrated in gray, and for anonymized data, colors can be depicted from the legend. ACE: angiotensin-converting enzyme; ARBs: Angiotensin II receptor blockers; BP: blood pressure; eGFR: estimated glomerular filtration rate; MR: marketer risk; PR: prosecutor risk; UACR: urine albumin-to-creatinine ratio.

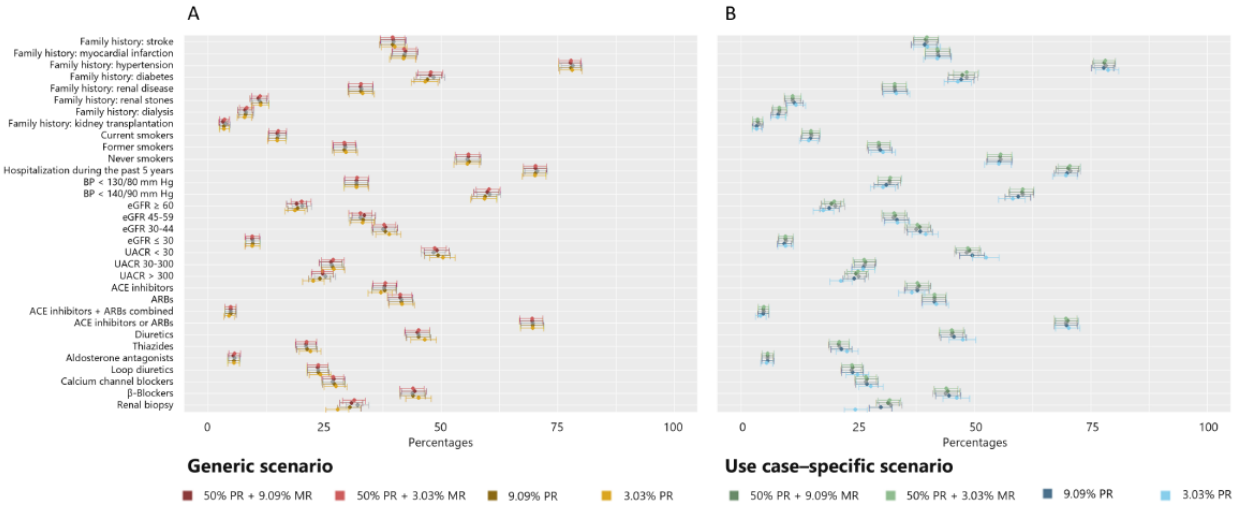

XSL•FO
RenderX

Figure 6. Illustration of age and BMI of female participants who did not have diabetes in the original and anonymized data sets. Anonymization was applied as defined in the (A and B) generic and (C and D) use case–specific scenario. Bar plots illustrate counts for anonymized data at selected privacy level: 9.09% MR+50% PR, 3.03% MR+50% PR, 9.09% PR, and 3.03% PR. The original data are illustrated as a density plot in gray. In the generic scenario, BMI was calculated using the generalized data of height and weight. MR: marketer risk; PR: prosecutor risk.

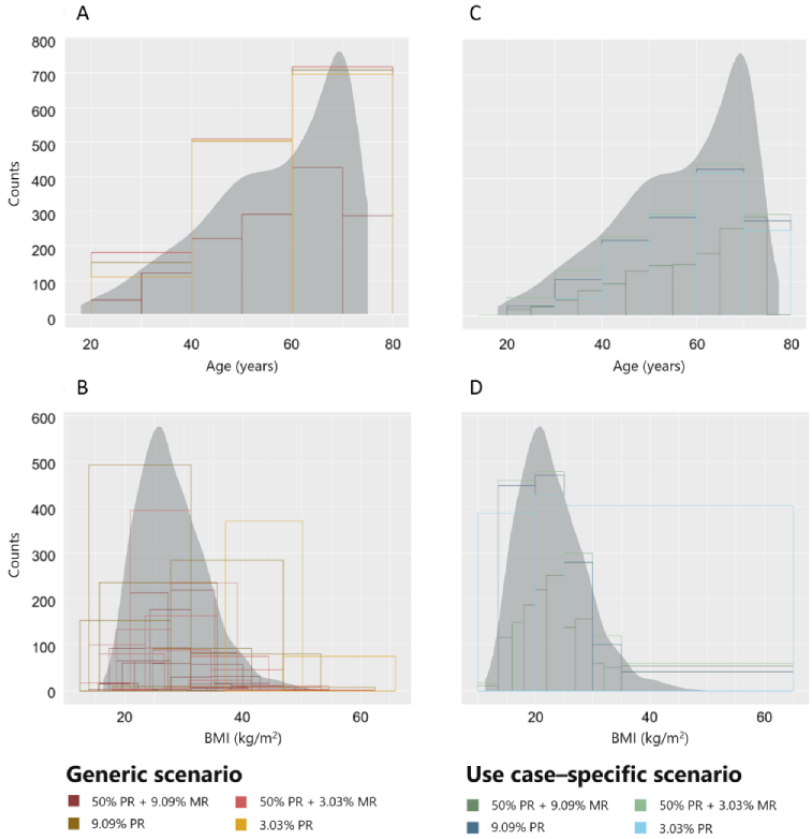

Discussion

Principal Results

This study provides an in-depth view into the use of anonymization processes for sharing data from clinical studies. Based on a state-of-the-art threat modeling methodology and established risk models, a wide range of anonymization configurations was compared to study the privacy-utility trade-off. We further considered a use case–specific anonymization approach tailored toward our real-world application to optimize anonymization.

Our results exhibited quite high average privacy achieved for all records (<10% empirical average PR, ie, MR) even at the highest predefined thresholds for PR. At the same time, use case–specific utility at the data set–level was high (>90%) across all thresholds. Individual results disagreed in the sense of nonoverlapping 95% CIs, but this was rare and mainly occurred at 3.03% PR. We would not consider them as relevant in our descriptive analyses where no direct implications were drawn from individual estimates. General-purpose metrics, in contrast, underestimated the actual utility in our real-world application. The 95% CI overlap at the data set–level therefore seems to be a useful proxy for actual utility in descriptive analyses.

Based on our investigation, it is evident that use case–specific tailoring had a positive effect on the reproducibility. In assessing anthropometric data, for example, use case–specific tailoring unfolded its potential, BMI represents a screening tool for chronic disease and mortality, while weight and in particular height are not pathological factors by themselves [54]. Consequently, reasonable and use case–specific preprocessing (removal of weight and height) resulted in preserved utility, while generic configuration (removal of BMI) lost almost all the information.

Comparison With Prior Work

While an increasing number of examples of real-world applications of anonymization algorithms are published [12,15,16], we did not come across any investigations that measured the reproducibility (eg, by 95% CI overlap) of descriptive real-world analyses except for prior work on the GCKD study. However, several studies focusing on preserving the utility of anonymized data for descriptive real-world analyses without explicitly introducing use case–specific measures have been published. For instance, in the Lean European Open Survey on patients infected with SARS-CoV-2, an anonymization pipeline using 9.09% PR as a threshold was established [25]. The anonymized data set was evaluated in selected clinical parameters with reported maximum frequency differences of only 0.11%. In addition, a real-world analysis on patients with stroke presented with low error rates [15]. Interestingly, the authors of this study evaluated a new method that limits the degree to which generalization is applied. Analogously to what we observed in the use case–specific scenario, these predefined settings resulted in anonymized data that are closer to the original data. In addition, prior work on anonymized data of the GCKD study demonstrated preserved descriptive characteristics

XSL•FO
RenderX

at 2 selected privacy levels and highlighted the limitations of general-purpose utility metrics [28]. Similar to this study, the 95% CI overlap was used to confirm reproducibility, but it was not evaluated throughout an entire research project or across different anonymization processes, and only a small proportion of the risk-utility space was covered.

In inferential statistics in general, there is evidence for a lack of reproducibility. The evaluation of a low dimensionality data set—that contains only 2 variables that needed to be protected—concluded there were biased results across differently anonymized data [27]. The authors reported a decreasing accuracy of relative risk estimates by clustering analyses independent of the applied privacy model. Similarly, a use case–specific evaluation focusing on machine learning models for early acute kidney injury risk prediction identified a statistically relevant discrepancy in individual performance measures while at the same time preserving overall prediction accuracy [26]. This discrepancy points toward a need for multidimensional utility assessment. As stated earlier and shown in these published examples, individual estimate disagreements might or might not result in false implications depending on the affection of outcomes or potential confounders.

Data sharing has been mandated by several regulatory agencies [2] and is desirable for many reasons (eg, transparency, reproducibility, collaboration, and innovation). It is often subject to institutional policies and laws, such as the General Data Protection Regulation [5]. To promote data sharing, technical conditions to satisfy FAIR need to be realized. However, at the same time, privacy-enhancing technologies should be thoroughly assessed. In this context, utility concerns can pose a real threat and should be as much a part of discussion as privacy concerns. We want to encourage data sharing in a way that does not compromise patients' privacy or research quality. The research communicated through our project can contribute to a better understanding of anonymization and potential pitfalls.

We also encourage regulators, policy makers, and society to openly discuss the costs—in terms of domain expert knowledge, time, technical requirements, and utility—that all stakeholders are willing to bear to maintain high levels of privacy. Our results highlight the weakness of generic anonymization when aiming to support disparate uses of the data when high levels of privacy need to be maintained. Conclusions might be either taking the extra costs of use case–specific tailoring and additional measures of control, agreeing on a lower privacy level in favor of high-quality research, or accepting the limitations of studies conducted on a generic anonymized data set.

Limitations

While our evaluation included a comprehensive assessment, there are several limitations to this investigation. First, this study focused on measuring and reducing specific privacy concerns, namely, prosecutor and marketer reidentification risk, as well as a certain type of anonymization framework that uses generalization and suppression. There are certainly other privacy risks and anonymization algorithms, and it is possible that some may provide a better privacy-utility trade-off in certain scenarios. Second, we did not protect the data against sensitive attribute inference where confidential information is accessed indirectly through inference. We made this decision because we assumed a controlled access setting and the relevance of this risk as well as its possible countermeasures are controversial [55]. Third, the threat modeling approach we relied upon requires assumptions about the goals and possibilities of potential adversaries. Other researchers aiming to apply our technique could consider running a structured assessment among a panel of experts (eg, using the Delphi technique) to strengthen the reliability of the threat modeling. Finally, while descriptive analyses are a basic feature of almost any study, anonymization must also stand up to more complex statistics. Individual nonoverlapping 95% CIs as detected in this study might relevantly affect inferential statistics. In this context, the estimate agreement (ie, by 95% CI overlap) and direction of effect and statistical significance need to be considered [53].

Conclusions

Against the background of increasing data sharing initiatives, it should be highlighted that utility concerns should be as much a part of discussion as privacy concerns. Our results highlight the weakness of generic anonymization when high levels of privacy are maintained. An anonymized data set aiming to support multiple disparate and possibly competing likely uses might allow exploratory analyses but may not be appropriate for drawing conclusions from individual analyses. This underscores the merit of applying case-specific tailoring and may justify its extra costs, for example, in terms of time and additional measures of control. However, the discussion about the acceptable costs, both financial and in terms of utility, that are required to uphold high levels of privacy should involve a broad spectrum of stakeholders. It should include domain experts, regulators, policy makers, patient representatives, and society at large.

Acknowledgments

LP is a participant in the Junior Digital Clinician Scientist Program funded by the Charité—Universitätsmedizin Berlin and the Berlin Institute of Health at Charité. The authors would like to thank the program for the support in conducting the research project. The authors acknowledge financial support from the Open Access Publication Fund of Charité—Universitätsmedizin Berlin. The authors are very grateful for the willingness and time of all study participants of the German Chronic Kidney Disease (GCKD) study. The enormous effort of the study personnel at the regional centers is highly appreciated. The authors also thank a large number of nephrologists for their support of the GCKD study (the list of nephrologists currently collaborating with the GCKD study is available on the web) [56].

XSL•FO
RenderX

Current GCKD investigators and collaborators with the GCKD study are University of Erlangen-Nürnberg: KUE, Heike Meiselbach, Markus P Schneider, Mario Schiffer, Hans-Ulrich Prokosch, Barbara Bärthlein, Andreas Beck, Detlef Kraska, André Reis, Arif B Ekici, Susanne Becker, Ulrike Alberth-Schmidt, Sabine Marschall, Eugenia Schefler, and Anke Weigel; University of Freiburg: Gerd Walz, Anna Köttgen, Ulla T Schultheiß, Fruzsina Kotsis, Simone Meder, Erna Mitsch, and Ursula Reinhard; RWTH Aachen University: Jürgen Floege, Turgay Saritas, and Alice Gross; Charité—Universitätsmedizin Berlin: ES, Seema Baid-Agrawal, and Kerstin Theisen; Hannover Medical School: Kai Schmidt-Ott; University Hospital, Renal Center, Heidelberg: Martin Zeier, Claudia Sommerer, and Mehtap Aykac; University Hospital Jena: Gunter Wolf, Martin Busch, and Rainer Paul; Ludwig-Maximilians University of München: Thomas Sitter; University of Würzburg: Christoph Wanner, Vera Krane, Antje Börner-Klein, and Britta Bauer; Division of Genetic Epidemiology, Medical University of Innsbruck: Florian Kronenberg, Julia Raschenberger, Barbara Kollerits, Lukas Forer, Sebastian Schönherr, and Hansi Weissensteiner; Institute of Functional Genomics, University of Regensburg: Peter Oefner and Wolfram Gronwald; and Department of Medical Biometry, Informatics and Epidemiology (IMBIE), University Hospital of Bonn: Matthias Schmid and Jennifer Nadal.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to privacy risks but are available from the corresponding author on reasonable request.

Authors' Contributions

LP, TM, ES, KUE, and FP contributed to the conception and design of the study. TM developed a software pipeline to automate the anonymization. LP performed the analyses. LP, TM, and FP drafted the paper. BM, ES, KUE, and FP further supervised the project. The German Chronic Kidney Disease investigators contributed to patient recruitment and data collection. No generative artificial intelligence was used in writing this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on methods and additional results.
[DOCX File , 789 KB - jmir_v26i1e49445_app1.docx]

References

1. Mansmann U, Locher C, Prasser F, Weissgerber T, Sax U, Posch M, et al. Implementing clinical trial data sharing requires training a new generation of biomedical researchers. Nat Med 2023;29(2):298-301 [FREE Full text] [doi: 10.1038/s41591-022-02080-y] [Medline: 36732626]
2. Egilman AC, Kapczynski A, McCarthy ME, Luxkaranayagam AT, Morten CJ, Herder M, et al. Transparency of regulatory data across the European Medicines Agency, Health Canada, and US Food and Drug Administration. J Law Med Ethics 2021;49(3):456-485. [doi: 10.1017/jme.2021.67] [Medline: 34665102]
3. Naudet F, Siebert M, Pellen C, Gaba J, Axfors C, Cristea I, et al. Medical journal requirements for clinical trial data sharing: ripe for improvement. PLoS Med 2021;18(10):e1003844 [FREE Full text] [doi: 10.1371/journal.pmed.1003844] [Medline: 34695113]
4. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
5. EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679, Article 5(1)(c). Dsgvo-Portal. 2016. URL: https://www.dsgvo-portal.de/gdpr_article_5.php [accessed 2023-04-22]
6. Tacconelli E, Gorska A, Carrara E, Davis RJ, Bonten M, Friedrich AW, et al. Challenges of data sharing in European Covid-19 projects: a learning opportunity for advancing pandemic preparedness and response. Lancet Reg Health Eur 2022;21:100467 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100467] [Medline: 35942201]
7. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014;14:1144 [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
8. Kalkman S, Mostert M, Gerlinger C, van Delden JJM, van Thiel GJMW. Responsible data sharing in international health research: a systematic review of principles and norms. BMC Med Ethics 2019;20(1):21 [FREE Full text] [doi: 10.1186/s12910-019-0359-9] [Medline: 30922290]
9. Read KB, Ganshorn H, Rutley S, Scott DR. Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis. CMAJ Open 2021;9(4):E980-E987 [FREE Full text] [doi: 10.9778/cmajo.20200303] [Medline: 34753787]

XSL•FO
RenderX

10. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015;350:h1139 [FREE Full text] [doi: 10.1136/bmj.h1139] [Medline: 25794882]
11. Heurix J, Zimmermann P, Neubauer T, Fenz S. A taxonomy for privacy enhancing technologies. Comput Secur 2015;53:1-17. [doi: 10.1016/j.cose.2015.05.002]
12. Haber AC, Sax U, Prasser F, NFDI4Health Consortium. Open tools for quantitative anonymization of tabular phenotype data: literature review. Brief Bioinform 2022;23(6):bbac440 [FREE Full text] [doi: 10.1093/bib/bbac440] [Medline: 36215114]
13. Willenborg LCRJ, de Waal T. Elements of Statistical Disclosure Control. New York: Springer-Verlag; 2001.
14. Xia W, Heatherly R, Ding X, Li J, Malin B. R-U policy frontiers for health data de-identification. J Am Med Inform Assoc 2015;22(5):1029-1041 [FREE Full text] [doi: 10.1093/jamia/ocv004] [Medline: 25911674]
15. Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. BMC Med Inform Decis Mak 2017;17(1):104 [FREE Full text] [doi: 10.1186/s12911-017-0499-0] [Medline: 28693480]
16. Liu Y, Conway D, Wan Z, Kantarcioglu M, Vorobeychik Y, Malin BA. De-identifying socioeconomic data at the census tract level for medical research through constraint-based clustering. AMIA Annu Symp Proc 2021;2021:793-802 [FREE Full text] [Medline: 35309009]
17. Prasser F, Bild R, Eicher J, Spengler H, Kohlmayer F, Kuhn KA. Lightning: utility-driven anonymization of high-dimensional data. Trans Data Priv 2016;9(2):161-185 [FREE Full text]
18. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci U S A 2010;107(17):7898-7903 [FREE Full text] [doi: 10.1073/pnas.0911686107] [Medline: 20385806]
19. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. BMJ 2018;360:k400 [FREE Full text] [doi: 10.1136/bmj.k400] [Medline: 29440066]
20. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
21. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. J Med Internet Res 2019;21(5):e13484 [FREE Full text] [doi: 10.2196/13484] [Medline: 31152528]
22. Zuo Z, Watson M, Budgen D, Hall R, Kennelly C, Al Moubayed N. Data anonymization for pervasive health care: systematic literature mapping study. JMIR Med Inform 2021;9(10):e29871 [FREE Full text] [doi: 10.2196/29871] [Medline: 34652278]
23. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071 [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
24. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. Am Stat 2006;60(3):224-232. [doi: 10.1198/000313006x124640]
25. Jakob CEM, Kohlmayer F, Meurers T, Vehreschild JJ, Prasser F. Design and evaluation of a data anonymization pipeline to promote open science on COVID-19. Sci Data 2020;7(1):435 [FREE Full text] [doi: 10.1038/s41597-020-00773-y] [Medline: 33303746]
26. Song X, Waitman LR, Hu Y, Luo B, Li F, Liu M. The impact of medical big data anonymization on early acute kidney injury risk prediction. AMIA Jt Summits Transl Sci Proc 2020;2020:617-625 [FREE Full text] [Medline: 32477684]
27. Pilgram L, Schäffner E, Eckardt KU, Prasser F, GCKD Investigators. Utility-preserving anonymization in a real-world scenario: evidence from the German Chronic Kidney Disease (GCKD) study. Stud Health Technol Inform 2023;302:28-32. [doi: 10.3233/SHTI230058] [Medline: 37203603]
28. Ferrão ME, Prata P, Fazendeiro P. Utility-driven assessment of anonymized data via clustering. Sci Data 2022;9(1):456 [FREE Full text] [doi: 10.1038/s41597-022-01561-6] [Medline: 35907927]
29. Eckardt KU, Bärthlein B, Baid-Agrawal S, Beck A, Busch M, Eitner F, et al. The German Chronic Kidney Disease (GCKD) study: design and methods. Nephrol Dial Transplant 2012;27(4):1454-1460 [FREE Full text] [doi: 10.1093/ndt/gfr456] [Medline: 21862458]
30. Titze S, Schmid M, Köttgen A, Busch M, Floege J, Wanner C, et al. Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the German Chronic Kidney Disease (GCKD) cohort. Nephrol Dial Transplant 2015;30(3):441-451 [FREE Full text] [doi: 10.1093/ndt/gfu294] [Medline: 25271006]
31. Hede K. Project data sphere to make cancer clinical trial data publicly available. J Natl Cancer Inst 2013;105(16):1159-1160 [FREE Full text] [doi: 10.1093/jnci/djt232] [Medline: 23904505]
32. Yale School of Medicine. The YODA Project. Center for Outcomes Research & Evaluation (CORE). URL: https://yoda.yale.edu/ [accessed 2023-04-22]
33. Clinical data. European Medicines Agency (EMA). URL: https://clinicaldata.Ema.europa.eu/web/cdp/home [accessed 2023-04-22]
34. Keerie C, Tuck C, Milne G, Eldridge S, Wright N, Lewis SC. Data sharing in clinical trials—practical guidance on anonymising trial datasets. Trials 2018;19(1):25 [FREE Full text] [doi: 10.1186/s13063-017-2382-9] [Medline: 29321053]

XSL•FO
RenderX

35. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform 2018;22(2):611-622. [doi: 10.1109/JBHI.2017.2676880] [Medline: 28358693]
36. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. European Medicines Agency (EMA). 2018. URL: https://www.ema.europa.eu/en/clinical-data-publication/support-industry-clinical-data-publication/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use [accessed 2018-10-15]
37. Garfinkel SL. De-identification of personally identifiable information. National Institute of Standards and Technology (NIST). 2015. URL: https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf [accessed 2024-03-22]
38. Institute of Medicine, Board on Health Sciences Policy, Committee on Strategies for Responsible Sharing of Clinical Trial Data. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk: Appendix B; Concepts and Methods for De-Identifying Clinical Trial Data. Washington, DC: National Academies Press; 2015.
39. PHUSE Data Transparency Working Group. Data anonymisation and risk assessment automation. PHUSE. 2020. URL: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf [accessed 2022-09-12]
40. Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. Hum Genet 2011;130(3):383-392 [FREE Full text] [doi: 10.1007/s00439-011-1042-5] [Medline: 21739176]
41. Pilgram L, Schons M, Jakob CEM, Claßen AY, Franke B, Tscharntke L, et al. The COVID-19 pandemic as an opportunity and challenge for registries in health services research: lessons learned from the Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS). Gesundheitswesen 2021;83(S 01):S45-S53. [doi: 10.1055/a-1655-8705] [Medline: 34731893]
42. Schons M, Pilgram L, Reese JP, Stecher M, Anton G, Appel KS, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. Eur J Epidemiol 2022;37(8):849-870 [FREE Full text] [doi: 10.1007/s10654-022-00896-z] [Medline: 35904671]
43. Xia W, Basford M, Carroll R, Clayton EW, Harris P, Kantacioglu M, et al. Managing re-identification risks while providing access to the All of Us research program. J Am Med Inform Assoc 2023;30(5):907-914 [FREE Full text] [doi: 10.1093/jamia/ocad021] [Medline: 36809550]
44. Krzyzanowski B, Manson SM. Twenty years of the Health Insurance Portability and Accountability Act safe harbor provision: unsolved challenges and ways forward. JMIR Med Inform 2022 Aug 03;10(8):e37756 [FREE Full text] [doi: 10.2196/37756] [Medline: 35921140]
45. Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst 2002;10(05):557-570. [doi: 10.1142/S0218488502001648]
46. Prasser F, Eicher J, Spengler H, Bild R, Kuhn K. Flexible data anonymization using ARX—current status and challenges ahead. Softw Pract Exp 2020;50(7):1277-1304 [FREE Full text] [doi: 10.1002/spe.2812]
47. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8(1):3-15. [doi: 10.1177/096228029900800102] [Medline: 10347857]
48. Bennett DA. How can I deal with missing data in my study? Aust N Z J Public Health 2001;25(5):464-469. [Medline: 11688629]
49. Pilgram L. pilgraml/UseCaseSpecificAnonymization: published generalization hierarchies. Zenodo. 2023. URL: https://zenodo.org/records/7978630 [accessed 2024-03-22]
50. El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, et al. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc 2009;16(5):670-682 [FREE Full text] [doi: 10.1197/jamia.M3144] [Medline: 19567795]
51. Iyengar VS. Transforming data to satisfy privacy constraints. 2002 Presented at: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 23-26, 2002; Edmonton, AB p. 279-288. [doi: 10.1145/775047.775089]
52. Prasser F, Bild R, Kuhn KA. A generic method for assessing the quality of de-identified health data. Stud Health Technol Inform 2016;228:312-316. [Medline: 27577394]
53. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, et al. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. Clin Pharmacol Ther 2020;107(4):817-826. [doi: 10.1002/cpt.1633] [Medline: 31541454]
54. Flegal KM, Graubard BI. Estimates of excess deaths associated with body mass index and other anthropometric variables. Am J Clin Nutr 2009;89(4):1213-1219 [FREE Full text] [doi: 10.3945/ajcn.2008.26698] [Medline: 19190072]
55. Wachter S, Mittelstadt B. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. Colum Bus L Rev 2019;2:14-18 [FREE Full text] [doi: 10.31228/osf.io/mu2kf]
56. German Chronic Kidney Disease (GCKD). Universitaetsklinikum Erlangen, GCKD Study center. URL: http://www.gckd.org [accessed 2023-05-29]

Abbreviations

CKD: chronic kidney disease

XSL•FO
RenderX

eGFR: estimated glomerular filtration rate
FAIR: findable, accessible, interoperable, and reusable
GCKD: German Chronic Kidney Disease
MR: marketer risk
PR: prosecutor risk
UACR: urine albumin-to-creatinine ratio

Edited by A Mavragani; submitted 30.05.23; peer-reviewed by F Ritchie, R Hendricks-Sturrup; comments to author 29.12.23; revised version received 14.01.24; accepted 13.02.24; published 24.04.24.

Please cite as:
Pilgram L, Meurers T, Malin B, Schaeffner E, Eckardt KU, Prasser F, GCKD Investigators
The Costs of Anonymization: Case Study Using Clinical Data
J Med Internet Res 2024;26:e49445
URL: https://www.jmir.org/2024/1/e49445
doi:10.2196/49445
PMID:38657232

©Lisa Pilgram, Thierry Meurers, Bradley Malin, Elke Schaeffner, Kai-Uwe Eckardt, Fabian Prasser, GCKD Investigators. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Finding Medical Photographs of Patients Online: Randomized, Cross-Sectional Study

Zack Marshall[1,2], MSW, PhD; Maushumi Bhattacharjee[3], LLB, LLM; Meng Wang[1], PhD; Abdul Cadri[4], MPH; Hannah James[5]; Shabnam Asghari[6], MD, MPH, PhD; Rene Peltekian[7], MSW; Veronica Benz[1], MSW; Vanessa Finley-Roy[8], SW, MSc; Brynna Childs[2,9], BA, MISt; Lauren Asaad[1], BHSc; Michelle Swab[10], MA, MLIS; Vivian Welch[11], PhD; Fern Brunger[12], PhD; Chris Kaposy[12], PhD

[1]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[2]School of Social Work, McGill University, Montreal, QC, Canada
[3]Faculty of Law, McGill University, Montreal, QC, Canada
[4]Department of Family Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[5]Department of Anatomy & Cell Biology, McGill University, Montreal, QC, Canada
[6]Faculty of Medicine, Memorial University of Newfoundland, St John's, NL, Canada
[7]Renison University College, University of Waterloo, Waterloo, ON, Canada
[8]Faculty of Medicine, Universite de Montreal, Montreal, QC, Canada
[9]CKUT 90.3FM Radio McGill, Montreal, QC, Canada
[10]Health Sciences Library, Memorial University, St John's, NL, Canada
[11]Bruyere Research Institute, Ottawa, ON, Canada
[12]Division of Population Health and Applied Health Sciences, Faculty of Medicine, Memorial University, St John's, NL, Canada

Corresponding Author:
Zack Marshall, MSW, PhD
Department of Community Health Sciences
Cumming School of Medicine
University of Calgary
3280 Hospital Drive NW
Calgary, AB, T2N 4Z6
Canada
Phone: 1 4032206940
Email: zack.marshall@ucalgary.ca

Abstract

Background: Photographs from medical case reports published in academic journals have previously been found in online image search results. This means that patient photographs circulate beyond the original journal website and can be freely accessed online. While this raises ethical and legal concerns, no systematic study has documented how often this occurs.

Objective: The aim of this cross-sectional study was to provide systematic evidence that patient photographs from case reports published in medical journals appear in Google Images search results. Research questions included the following: (1) what percentage of patient medical photographs published in case reports were found in Google Images search results? (2) what was the relationship between open access publication status and image availability? and (3) did the odds of finding patient photographs on third-party websites differ between searches conducted in 2020 and 2022?

Methods: The main outcome measure assessed whether at least 1 photograph from each case report was found on Google Images when using a structured search. Secondary outcome variables included the image source and the availability of images on third-party websites over time. The characteristics of medical images were described using summary statistics. The association between the source of full-text availability and image availability on Google Images was tested using logistic regressions. Finally, we examined the trend of finding patient photographs using generalized estimating equations.

Results: From a random sample of 585 case reports indexed in PubMed, 186 contained patient photographs, for a total of 598 distinct images. For 142 (76.3%) out of 186 case reports, at least 1 photograph was found in Google Images search results. A total of 18.3% (110/598) of photographs included eye, face, or full body, including 10.9% (65/598) that could potentially identify

XSL•FO
RenderX

the patient. The odds of finding an image from the case report online were higher if the full-text paper was available on ResearchGate (odds ratio [OR] 9.16, 95% CI 2.71-31.02), PubMed Central (OR 7.90, 95% CI 2.33-26.77), or Google Scholar (OR 6.07, 95% CI 2.77-13.29) than if the full-text was available solely through an open access journal (OR 5.33, 95% CI 2.31-12.28). However, all factors contributed to an increased risk of locating patient images online. Compared with the search in 2020, patient photographs were less likely to be found on third-party websites based on the 2022 search results (OR 0.61, 95% Cl 0.43-0.87).

Conclusions: A high proportion of medical photographs from case reports was found on Google Images, raising ethical concerns with policy and practice implications. Journal publishers and corporations such as Google are best positioned to develop an effective remedy. Until then, it is crucial that patients are adequately informed about the potential risks and benefits of providing consent for clinicians to publish their images in medical journals.

(J Med Internet Res 2024;26:e55352) doi:10.2196/55352

KEYWORDS

patient photographs; privacy; informed consent; publication ethics; case reports

Introduction

Case reports are an important tool for medical, scientific, and educational purposes [1]. Written by practicing clinicians, peer-reviewed case reports provide relevant and timely medical information that contributes to evidence-based practice [1,2]. A large number of case reports are published each year; for example, 74,270 case reports were published in 2022 and indexed in PubMed.

Case reports often include images, including patient photographs [3,4]. Guidelines related to the publication of medical photographs in case reports often refer to overarching statements such as the Declaration of Helsinki, or slightly more specific policies such as the guidelines outlined by the Committee on Publication Ethics (COPE), or the Case Report (CARE) guidelines for case reports [5-7]. The Declaration of Helsinki states that research participants must be fully informed of any potential risks and benefits associated with the relevant study [8]. COPE guidelines provide clear recommendations for publishers, editors, and various research institutes on the topic of publication ethics [9]. Meanwhile, CARE guidelines are specific to case reports and seek to promote and improve their transparency, accuracy, and usefulness [10]. The CARE guidelines include a checklist with items such as deidentified patient information, informed consent, and patient perspective on the treatment they received [10]. While multiple guidelines exist, adherence is not mandatory; 1 study found that out of 50 journals, 76% did not adhere to any guidelines for publication of personal information [11]. Another study investigating CARE guideline adherence in 36 Indian medical journals found that only a third exhibited average adherence and that overall there was poor reporting of subject-informed consent [7].

With the growth of online publishing and advancements in technology, case reports from academic journals are widely available as web-based publications, and their reach has expanded to a larger audience [4]. While increased access to medical case reports may be beneficial, photographs from case reports published in academic journals are now also available in online image search results such as Google Images [3,4]. In such cases, patient photographs circulate beyond the original journal website and can be accessed by anyone using the internet. This raises ethical and legal concerns regarding patients' informed consent and privacy of health information.

In the original study on this topic, drawing on a sample of case reports with patients who are transgender published between 2008 and 2015, at least 1 patient photograph was available on Google Images for 37% of the medical case reports in the sample [3]. Curious about whether the results would be the same for a random sample of medical case reports published more recently, the aim of this cross-sectional study was to provide systematic evidence that patient photographs from case reports published in medical journals appear in Google Images search results. Research questions for this study were (1) what percentage of patient medical photographs published in case reports are found in Google Images search results? (2) what is the relationship between open access publication status and image availability? and (3) do the odds of finding patient photographs on third-party websites differ between searches conducted in 2020 and 2022?

Methods

Study Design

The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) cross-sectional checklist was used when writing up results [12] (see Multimedia Appendix 1).

Ethical Considerations

Research ethics approval was not required because the data were collected from case reports published in medical journals.

Data Source and Study Population

PubMed includes a diverse range of medical journals and is "the most widely used database with biomedicine-related article abstracts" with over 36 million entries [13]. The efficient identification of a random sample was facilitated by the ways medical case reports are identified within PubMed. A structured search of PubMed was conducted to identify all indexed medical case reports published within a 1-year period between July 1, 2017, and June 30, 2018 (Search: "2017/07/01" [Date—Publication]: "2018/06/30" (Date—Publication) Filters: Case Reports). The search produced 23,589 results (the search was conducted on August 15, 2018).

Sample Size

To determine sample size, a pilot study was conducted to inform an estimate of effect size and power. All medical case reports indexed in PubMed for the month of February 2018 were

XSL•FO
RenderX

identified. This search produced 955 references. Full-text PDFs were retrieved for each reference and the documents were visually checked to see whether each case study included photographic images of patients. Of the 955 case reports published in English in February 2018 and indexed in PubMed, 370 (38.7%) included patient photographs. Based on the original study, it was anticipated that approximately 37% of the case reports with photographs would include at least 1 image found online. Using a 95% Cl and a 4% margin of error, a sample size of 585 was required.

Data Collection and Measures

To identify the random sample for this study, the list of 23,589 case report references was exported from PubMed to Microsoft Excel (Microsoft Corp). A random number generator was used to assign a number to each of the references, and then the list of references was rank-ordered from the smallest to the largest random number. The first 585 references were selected in order, imported into EPPI-Reviewer (EPPI Centre) [14], and then full-text papers were uploaded for each reference (Figure 1).

For the 585 references, the full text of each case report was examined to determine whether the publication included clinical photographs of patients or not. The photographs from each publication were consecutively numbered on a hard copy, and then the information was entered into a Microsoft Excel spreadsheet, with a unique number for each case report and photograph. A total of 186 case reports included patient photographs, with a total of 598 patient photographs in the sample.

Two categories of data were collected—data at the case report level and data at the image level. At the image level, details were documented related to the specific part of the patient's body that was photographed (eg, eye, face, or torso); the timing of the photograph (eg, pre- or posttreatment or during surgery); the gender and age of the patient as described in the body of the case report; whether the photograph was in color or not; and whether the authors had attempted to anonymize the photograph using image blurring or bars covering parts of the image. For each case report, 1 member of the team entered data about the images into the Excel spreadsheet. All data were then independently verified by a second team member.

At the case report level, data collection included author information, year of publication, open access status, and availability on Google Scholar, ResearchGate, and PubMed Central. To document the open access format for each of the case reports, 1 member of the team searched for each paper in the open database Unpaywall. The open access status of case reports is classified using colors [15]. The color classifications are (1) gold—published in an open access journal that is indexed by the Directory of Open Access Journals; (2) green—toll-access on the publisher page, but there is a free copy in an open access repository; (3) hybrid—free under an open license in a toll-access journal; (4) bronze—free to read on the publisher page, but without a clearly identifiable license; and (5) closed—all other papers, including those shared only on academic social networks or Sci-Hub. For the purposes of this analysis, open access included papers categorized as gold, green, hybrid, and bronze.

Figure 1. Identification of a random sample of case reports with patient photographs.

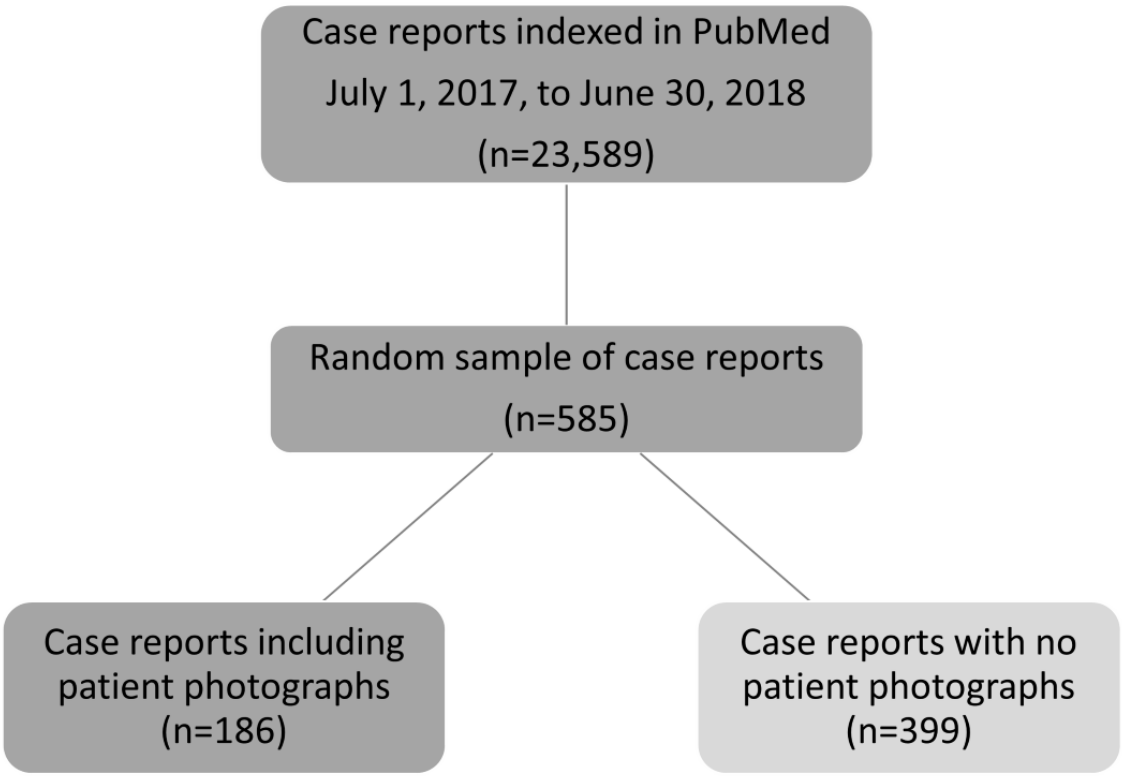

XSL•FO
RenderX

Google Images Search

To determine whether it was possible to find photographs from the case reports on Google Images, searches were carried out for each of the 186 case reports that included patient photographs. Searches were conducted on a yearly basis from 2019 to 2022, using an approach referred to as algorithmic probing [16]. This analysis focuses on the results of the most recent searches conducted in 2022.

Three members of the research team (AC, HJ, and ZM) conducted manual searches on Google Images for each paper in the study sample using the same strategy first developed by Marshall et al [3]. For each reference, the researcher used the title of the case report in quotation marks as the text key. These searches were conducted using a Tor browser, "a proxy that masks the location information and browsing history of the user, allowing for anonymous use of the Internet" [17]. This browser was used to minimize the influence of Google's personalization strategies to help prevent results from being skewed by historical searches conducted by team members [18]. Images of the search result pages were saved in PDF by date. One member of the research team then manually compared the search results in the PDF to the photographs in the published medical case report, circling the matching image using PDF editing software. Google Images search results also include a link directly under the image to the original source of the photograph. In Figure 2, for example, the first 3 images are from a case report [19] and include links to BMJ Case Reports, and Europe PMC. The link was extracted for each image and then each source was coded as a journal website, publisher website, research database (eg, Semantic Scholar), research repository (eg, ResearchGate), social media, professional association, or other. A second member of the team verified the results.

Figure 2. Sample of source links included in Google Images search results.

Data Management

Data were entered into Microsoft Excel and screened for misentries (eg, spelling errors, empty cells, or shifted cells). The primary outcome variable was the availability of medical photographs on Google Images and was coded as "0" not found and "1" found. Secondary outcome variables included the image source and the availability of images on third-party websites over time. Missing data analyses were performed to screen the data for entry errors.

Statistical Analysis

Sample characteristics were described using means, SDs, quantiles, and frequency distributions. The level of analysis is individual case reports rather than individual photographs. This is necessary for 2 reasons. First, some published figures contain more than 1 patient photograph. For example, a figure may include 4 images of a patient taken from different perspectives. Second, case reports included a range of 1 to 33 patient photographs. The relationship between multiple images found in 1 case report is different from multiple images found in separate case reports, and as a result, each photograph cannot be treated independently.

To better understand whether the characteristics of case reports (such as full-text availability on ResearchGate, PubMed Central, or Google Scholar, and open access status) were related to the availability of medical images on Google Images, chi-square

XSL•FO
RenderX

tests and simple logistic regressions were conducted. To test if there is any trend for finding images on third-party websites over different searches over time, generalized estimating equations with a logit link and robust sandwich estimators were used. Odds ratios (OR) with 95% Cl were reported. $P \leq .05$ was considered significant. Analyses were conducted using SAS (version 9.4; SAS Institute).

Results

Sample Demographics

From the sample of 585 case reports, 186 (31.7%) case reports had at least 1 patient clinical photograph. A total of 598 images were identified in these 186 medical case reports. Individual photographs were coded into five broad categories (1) the specific body part that was photographed; (2) patient sex as identified in the case report; (3) patient age (adult vs child); (4) the timing of the photograph (pretreatment, during surgery, autopsy, etc); and (5) whether the photograph was anonymized or not.

From the 186 case reports with 598 photographs, 309 (51.7%) were photographs of women, 278 (46.2%) were photographs of men, and 3 (0.5%) were photographs of trans women. Information about patient sex or gender was not provided for 8 (1.3%) of the photographs. Patients who were photographed ranged in age from 2 days to 93 years. A total of 412 (68.9%) photographs were taken of adult patients (older than 18 years), and 176 (29.4%) were photographs of infants, children, or teenagers younger than 18 years of age. Information about age was not provided for patients in 10 (1.7%) photographs.

Patient photographs most often included internal organs (eg, endoscopy, laparoscopy, or bronchoscopy; n=151 images, 25.3%). Other common types of photographs included limbs such as legs, arms, hands, or feet (n=109 images, 18.3%), or images of the abdomen or torso (n=53 images, 8.9%). A total of 110 (18.3%) out of 598 photographs included eye, face, and full-body photographs, including 65 (10.9%) that could potentially identify the patient. In terms of the context of when the photograph was taken, 403 (67.4%) were photographs of the patient's condition pre- or posttreatment whereas 144 (24%) were photographs taken during surgery (Table 1).

XSL•FO
RenderX

Table 1. Patient demographic characteristics of medical images.

Patient demographic characteristics	Values, n (%)
Gender	
Women	309 (51.7)
Men	278 (46.2)
Trans women	3 (0.5)
Unknown	8 (1.3)
Age (years)	
Adult (18 years or older)	412 (68.2)
Infant, child, or teenager (younger than 18 years)	176 (29.4)
Unknown	10 (1.7)
Type of photograph	
Internal organs or endoscopy	151 (25.3)
Limbs (legs, arms, feet, or hands)	109 (18.2)
Mouth	62 (10.4)
Torso or abdomen	53 (8.9)
Face	44 (7.4)
Eyes	41 (6.9)
Breasts or chest	36 (6.0)
Full body	25 (4.2)
Genitals	22 (3.7)
Ears	13 (2.2)
Head	13 (2.2)
Nose	2 (0.3)
Context of photograph	
Photograph of condition	403 (67.4)
Presurgery	7 (1.2)
During surgery	144 (24.1)
Specimen	39 (6.5)
Autopsy	3 (0.5)
Other	4 (0.7)

Open Access Status of Case Reports With Medical Images

Of the 186 case reports, 102 (54.8%) were closed access; among the closed-access reports, 66 (65%) case reports had at least 1 image found on Google Images. Of the 83 case reports that were open access, 76 (92%) had at least 1 image found on Google Images. From crude comparisons ($P<.001$), it appears that case reports with open access were more likely to have medical images visible as Google Images.

Image Availability

For 76.3% (142/186) of the case reports, at least 1 image was found on Google Images. The odds were higher of finding an image from the case report online if the full-text paper was available on ResearchGate (OR 9.16, 95% Cl 2.71-31.02), PubMed Central (OR 7.90, 95% Cl 2.33-26.77), or Google Scholar (OR 6.07, 95% Cl 2.77-13.29) than if full-text was available solely through an open access journal (OR 5.33, 95% Cl 2.31-12.28), but all factors contribute to increased odds of locating patient images online (Figure 3).

XSL•FO
RenderX

Figure 3. The relationship between the characteristics of case reports and availability of medical photographs by unadjusted ORs (simple logistic regression). OR: odds ratio.

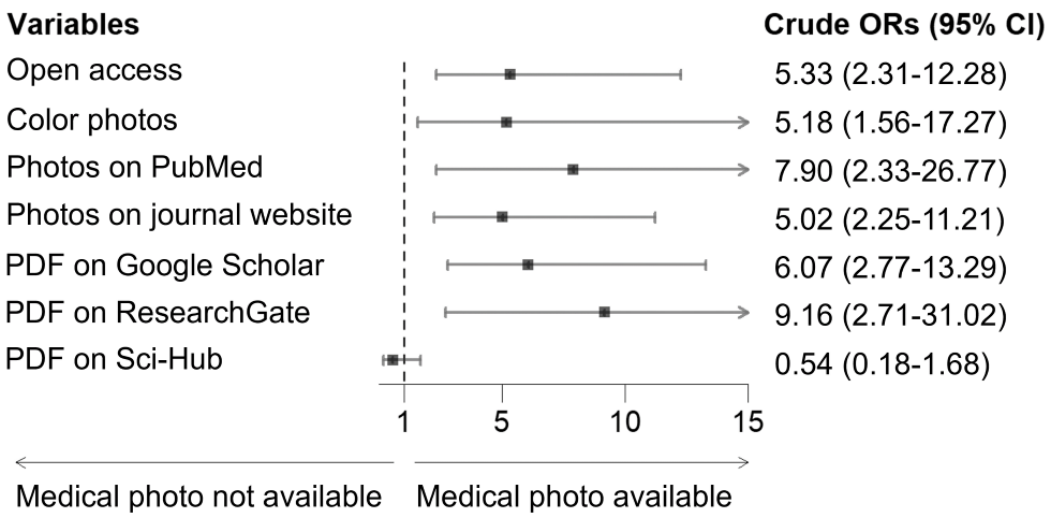

Image Source

To better understand where Google Images is obtaining patient photographs, information about data sources was extracted from the hyperlink under each of the images that were found online. Raw image sources included the journal website, publisher website, research database (eg, Semantic Scholar), research repository (eg, ResearchGate), social media, and professional associations. These were grouped into 2 main categories—journal websites or other websites (any third-party sources). A total of 51.0% of photographs came from the journal website, and 49.0% were from a third-party site. In 2021, 51.1% were from journal websites, and 48.9% from third-party sites. In 2022, the number of images from journal websites increased to 63.4%, while the number from third-party sites was 36.6%.

Trend Over Time

Based on generalized estimating equations, after adjusting for individual study differences, compared with the search in 2020, patient photographs were less likely to be found on third-party websites based on the 2022 search results. Specifically, the odds of finding a patient photograph on a third-party site in 2022 were about 40% less likely, compared with the search done in 2020. This finding was statistically significant with OR 0.61, 95% Cl 0.43-0.87. The likelihood of finding a patient photograph on a third-party website was not significantly different between the search in 2021 and the search in 2020 (Table 2).

Table 2. Google Images findings over time based on generalized estimating equations.

Search	Odd ratio (95% CI)	*P* value
2021 vs 2020	1.04 (0.78-1.40)	.77
2022 vs 2020	0.61 (0.43-0.87)	.006

Discussion

Principal Results

The aims of this study were to identify what percentage of patient photographs published in medical case reports were found in Google Images search results, to better understand the relationship between open access publication status and image availability, and to verify whether there is a trend over time for finding patient photographs on third-party websites. Out of the 186 case reports that included clinical photographs, at least 1 photograph from the case report was available on Google Images for 142 (76.3%) references. The odds of finding an image from the case report online were higher if the full-text paper was available on ResearchGate (OR 9.16, 95% CI 2.71-31.02), PubMed Central (OR 7.90, 95% CI 2.33-26.77), or Google Scholar (OR 6.07, 95% CI 2.77-13.29) than if full-text was available solely through an open access journal (OR 5.33, 95% CI 2.31-12.28), but all factors contributed to an increased risk of locating patient images online. This study is the first of its kind to search Google Images for medical photographs from a random sample of case reports; as such there are no studies with which to compare results.

Findings from this study are notably higher than the results from earlier research, where 34 (37%) out of 94 case reports had at least 1 photograph accessible on Google Images [3]. While the difference in sample population may partially account for the disparity in outcomes, this study identified several additional variables that influenced the availability or unavailability of

XSL•FO
RenderX

patient photographs on Google Images. For instance, finding images from the case reports online was more likely if the full-text paper was also available on ResearchGate, PubMed Central, or Google Scholar, compared to case reports solely accessible through open access publications.

To better understand how Google retrieves the images, the image source was recorded for all photographs found on Google Images and these results were compared over a 3-year time period. From 2020 to 2022, there was a notable change in where images were sourced, with a significant decrease in photographs housed on third-party websites such as ResearchGate and Semantic Scholar. This change may be linked to a recent legal judgment where Google was held liable for copyright infringement for displaying content with links to a third-party infringer's website which was not the original publisher and owner of the copyrighted content [20].

Limitations

The systematic, documented approach to searching for patient medical photographs on Google Images is a strength of this study. The primary challenge is that Google Images search results are not stable. Although the team attempted to manage as many factors as possible, including using the Tor browser to control for the influence of team member search histories, search results changed. Investigating the same data set yearly for over 3 years, sometimes the photographs were never found, while others were consistently located. The primary findings in this paper are based on the most recent searches in 2022, as the purpose of this study was not to demonstrate the ways search results change over time, but whether the images were found or not. Search results from 2020, 2021, and 2022 are available on request.

A further limitation is that the team did not investigate other image search engines or social media platforms where patient photographs might also appear. While the team was able to provide clear evidence using Google Images it would be an interesting avenue for future research to explore some alternate image search engines and platforms. In addition, the use of the Tor browser to minimize personalization in search results may not completely replicate the typical user experience and may have introduced a form of selection bias.

Conclusions

From a clinical standpoint, the availability of patient photographs on Google Images presents both advantages and risks. Results demonstrated a high proportion of medical photographs from case reports on Google Images. While this concentration allows for wider accessibility and educational benefits, the public availability of these sensitive images online also raises ethical concerns with respect to the privacy of personal health information. Patients should be adequately informed about the possible impacts of providing consent for clinicians to publish their images in medical journals. Even if clinicians seek consent for their publication in case reports, it is not clear whether patients are informed about the possibility of photographs becoming available on Google Images and reaching unintended audiences, including the media and the general public. Similarly, it is not known whether clinicians themselves are aware of these risks. As such, they may not be in a position to ensure informed consent from their patients regarding the potential availability of their clinical images online. A recent content analysis of journal consent forms for the publication of patient photographs found that 55.5% (10/18) of consent forms related to 132 journals mentioned photographs being available to an audience outside of the journal website, but only 16.7% (3/18) addressed the possibility of the patient's images being linked to journal or publisher social media platforms [21].

A lack of standardized guidelines poses a challenge to obtaining patient consent for publishing case reports with photographs. In addition to the policy and practice recommendations highlighted in earlier research, current findings underline the need for increased dialogue among academics, patients, governments, and industry. Discussions should focus on improving the consent process and establishing consistent practices and policies for publishing case reports with patient photographs. Study findings indicate that patient photographs are accessible on Google Images, even when published in closed-access case reports. Engagement with Google and other major online image repositories is critical to raise awareness of this issue and to seek input regarding the underlying causes and potential solutions. New policies should be implemented to ensure that patients are protected and that all stakeholders are aware of the risks involved in submitting clinical photographs to online medical journals. Accordingly, the next phase of this study focuses on qualitative interviews with case report authors, journal editors, publishers, and patients. The goal is to identify potential solutions to this complex ethical challenge, including responsive policies that will influence practices across academic publishing to maintain patient privacy.

Acknowledgments

The Natural Sciences and Engineering Research Council of Canada (554764-2021) provided student scholarship funding. The funders had no role in study design, in collection or interpretation of data, in writing the report, or in the decision to submit the paper for publication.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

ZM was responsible for conceptualization (lead), investigation (lead), methodology (equal), supervision (lead), writing—original draft preparation (lead), and writing—review and editing (lead). MB carried out the investigation (equal), project administration (equal), writing—original draft preparation (equal), and writing—review and editing (equal). MW contributed to the formal analysis (lead), methodology (equal), visualization (lead), writing—original draft preparation (equal), and writing—review and editing (supporting). AC did the investigation (equal), writing—original draft preparation (equal), and writing—review and editing (supporting). HJ did the investigation and writing—review and editing (supporting). SA aided in conceptualization (supporting), methodology (equal), writing—original draft preparation (equal), and writing—review and editing (supporting). RP contributed to the investigation (equal), project administration (equal), and writing—review and editing (supporting). VB performed the investigation (equal), project administration (equal), writing—original draft preparation (equal), and writing—review and editing (supporting). VFR did the investigation (equal), project administration (equal), and writing—review and editing (supporting). BC performed the investigation (equal) and writing—review and editing (supporting). LA performed the investigation (equal) and writing—review and editing (supporting). MS contributed to the methodology (equal) and writing—review and editing (supporting). VW aided in conceptualization (supporting), methodology (equal), and writing—review and editing (supporting). FB was involved in conceptualization (supporting) and writing—review and editing (supporting). CK contributed to the conceptualization (supporting) and writing—review and editing (supporting).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) cross-sectional checklist.
[PDF File (Adobe PDF File), 123 KB - jmir_v26i1e55352_app1.pdf]

References

1. Marshall Z, Brunger F, Welch V, Asghari S, Kaposy C. Open availability of patient medical photographs in Google Images search results: cross-sectional study of transgender research. J Med Internet Res 2018;20(2):e70 [FREE Full text] [doi: 10.2196/jmir.8787] [Medline: 29483069]
2. Neely JG, Karni RJ, Nussenbaum B, Paniello RC, Fraley PL, Wang EW, et al. Practical guide to understanding the value of case reports. Otolaryngol Head Neck Surg 2008;138(3):261-264. [doi: 10.1016/j.otohns.2007.12.014] [Medline: 18312868]
3. Roguljić M, Šimunović D, Peričić TP, Viđak M, Utrobičić A, Marušić M, et al. Publishing identifiable patient photographs in scientific journals: scoping review of policies and practices. J Med Internet Res 2022;24(8):e37594 [FREE Full text] [doi: 10.2196/37594] [Medline: 36044262]
4. Roguljić M, Peričić TP, Gelemanović A, Jukić A, Šimunović D, Buljan I, et al. What patients, students and doctors think about permission to publish patient photographs in academic journals: a cross-sectional survey in Croatia. Sci Eng Ethics 2020;26(3):1229-1247. [doi: 10.1007/s11948-019-00134-y] [Medline: 31541413]
5. Rees M, Corson SL. Publication of medical case reports and consent. Case Rep Womens Health 2017;15:A1 [FREE Full text] [doi: 10.1016/j.crwh.2017.07.002] [Medline: 29593998]
6. Neavyn M, Murphy C. Coming to a consensus on informed consent for case reports. J Med Toxicol 2014;10(4):337-339 [FREE Full text] [doi: 10.1007/s13181-014-0421-4] [Medline: 25135309]
7. Ravi R, Mulkalwar A, Thatte UM, Gogtay NJ. Medical case reports published in PubMed-indexed Indian journals in 2015: adherence to 2013 CARE guidelines. Indian J Med Ethics 2018;3(3):192-195 [FREE Full text] [doi: 10.20529/IJME.2018.036] [Medline: 29699955]
8. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
9. Promoting integrity in research and its publication. Committee on Publication Ethics. URL: https://publicationethics.org/ [accessed 2024-03-22]
10. Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D, CARE Group. The CARE guidelines: consensus-based clinical case report guideline development. J Clin Epidemiol 2014;67(1):46-51. [doi: 10.1016/j.jclinepi.2013.08.003] [Medline: 24035173]
11. Taheri A, Adibi P, Jafari MS, Saeedizadeh M, Rahimi A, Abbasi A. The reporting requirements of case reports and adherence of case report reporting guidelines in medical journals: an analysis of the authors' guide sections. J Med Case Rep 2023;17(1):2 [FREE Full text] [doi: 10.1186/s13256-022-03710-2] [Medline: 36604759]
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
13. Novoa J, Chagoyen M, Benito C, Moreno FJ, Pazos F. PMIDigest: interactive review of large collections of PubMed entries to distill relevant information. Genes 2023;14(4):942 [FREE Full text] [doi: 10.3390/genes14040942] [Medline: 37107700]

XSL•FO
RenderX

14. Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M, et al. EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. EPPI Centre. URL: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2967 [accessed 2024-04-07]
15. Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, et al. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. PeerJ 2018;6:e4375 [FREE Full text] [doi: 10.7717/peerj.4375] [Medline: 29456894]
16. Rogers R. Algorithmic probing: Prompting offensive Google results and their moderation. Big Data Soc 2023;10(1). [doi: 10.1177/20539517231176228]
17. Macrina A. Accidental technologist: the Tor browser and intellectual freedom in the digital age. Ref User Serv Q 2015;54(4):17. [doi: 10.5860/rusq.54n4.17]
18. Bozdag E. Bias in algorithmic filtering and personalization. Ethics Inf Technol 2013;15(3):209-227. [doi: 10.1007/s10676-013-9321-6]
19. van Vonderen JJ, Stol K, Buddingh EP, van der Kaay DCM. Herpes simplex transmission to chest and face through autoinoculation in an infant. BMJ Case Rep 2017;2017:00 [FREE Full text] [doi: 10.1136/bcr-2017-220447] [Medline: 28827297]
20. Bhattacharjee M, Kaposy C, Grossman MR, Marshall Z. Patient photographs on Google Images: a commentary on informed consent, copyright, and privacy laws. Law Innov Technol 2023;15(2):536-557. [doi: 10.1080/17579961.2023.2245684]
21. Asaad L, Bhattacharjee M, Marshall Z. A content analysis of patient medical image consent forms. 2024 May 29 Presented at: International Conference on Clinical Ethics and Consultation & Canadian Bioethics Society; May 29-31, 2024; Montreal, Canada URL: https://event.fourwaves.com/iccec2024/abstracts/d53bdd80-d964-4811-985d-910793a762b1

Abbreviations

CARE: Case Report
COPE: Committee on Publication Ethics
OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by T de Azevedo Cardoso; submitted 13.12.23; peer-reviewed by M Roguljić, D Singh; comments to author 08.03.24; revised version received 08.04.24; accepted 24.04.24; published 24.06.24.

Please cite as:
Marshall Z, Bhattacharjee M, Wang M, Cadri A, James H, Asghari S, Peltekian R, Benz V, Finley-Roy V, Childs B, Asaad L, Swab M, Welch V, Brunger F, Kaposy C
Finding Medical Photographs of Patients Online: Randomized, Cross-Sectional Study
J Med Internet Res 2024;26:e55352
URL: https://www.jmir.org/2024/1/e55352
doi:10.2196/55352
PMID:38913416

©Zack Marshall, Maushumi Bhattacharjee, Meng Wang, Abdul Cadri, Hannah James, Shabnam Asghari, Rene Peltekian, Veronica Benz, Vanessa Finley-Roy, Brynna Childs, Lauren Asaad, Michelle Swab, Vivian Welch, Fern Brunger, Chris Kaposy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ethical, Legal, and Practical Concerns Surrounding the Implemention of New Forms of Consent for Health Data Research: Qualitative Interview Study

Svenja Wiertz[1*], DPhil; Joachim Boldt[1*], PhD, PD

Department of Medical Ethics and the History of Medicine, University of Freiburg, Freiburg, Germany

[*]all authors contributed equally

Corresponding Author:
Svenja Wiertz, DPhil
Department of Medical Ethics and the History of Medicine
University of Freiburg
Stefan-Meier-Str. 26
Freiburg, 79104
Germany
Phone: 49 761 203 5044
Fax: 49 761 203 5039
Email: wiertz@egm.uni-freiburg.de

Abstract

Background: In Europe, within the scope of the General Data Protection Regulation, more and more digital infrastructures are created to allow for large-scale access to patients' health data and their use for research. When the research is performed on the basis of patient consent, traditional study-specific consent appears too cumbersome for many researchers. Alternative models of consent are currently being discussed and introduced in different contexts.

Objective: This study explores stakeholder perspectives on ethical, legal, and practical concerns regarding models of consent for health data research at German university medical centers.

Methods: Semistructured focus group interviews were conducted with medical researchers at German university medical centers, health IT specialists, data protection officers, and patient representatives. The interviews were analyzed using a software-supported structuring qualitative content analysis.

Results: Stakeholders regarded broad consent to be only marginally less laborious to implement and manage than tiered consent. Patient representatives favored specific consent, with tiered consent as a possible alternative. All stakeholders lamented that information material was difficult to understand. Oral information and videos were mentioned as a means of improvement. Patient representatives doubted that researchers had a sufficient degree of data security expertise to act as sole information providers. They were afraid of undue pressure if obtaining health data research consent were part of medical appointments. IT specialists and other stakeholders regarded the withdrawal of consent to be a major challenge and called for digital consent management solutions. On the one hand, the transfer of health data to non-European countries and for-profit organizations is seen as a necessity for research. On the other hand, there are data security concerns with regard to these actors. Research without consent is legally possible under certain conditions but deemed problematic by all stakeholder groups, albeit for differing reasons and to different degrees.

Conclusions: More efforts should be made to determine which options of choice should be included in health data research consent. Digital tools could improve patient information and facilitate consent management. A unified and strict regulation for research without consent is required at the national and European Union level. Obtaining consent for health data research should be independent of medical appointments, and additional personnel should be trained in data security to provide information on health data research.

(J Med Internet Res 2024;26:e52180) doi:10.2196/52180

KEYWORDS

health data; health research; informed consent; broad consent; tiered consent; consent management; digital infrastructure; data safety; GDPR

XSL•FO
RenderX

Introduction

Background

In the European Union (EU), the outlines of a European health data space are currently being developed and discussed. It aims to facilitate accessing health data for research purposes, among other features [1]. In some countries, such as Estonia and Finland, fully digitalized health data infrastructures that also allow for secondary use of health data for research have already been implemented. In other countries, among them Germany, developments are ongoing [2-7]. In Germany, a central example of such developments is the government's medical informatics initiative (MII; *Medizininformatik-Initiative*) [8]. Its purpose is to create data integration centers at German university medical centers and to establish centralized access to the data.

In regard to these infrastructures, it is an important question from ethical, legal, and social perspectives whether and how patients are informed about the use of their data and how they consent to this use [9-11]. Informed consent consists of 2 elements, as the term indicates: information ensures that patients know what their health data are used for. From an ethical point of view, information can be regarded as a demand for the principle of transparency. By contrast, consent safeguards the ability to control what one's data are used for and serves to retain privacy. Together, transparency and control allow autonomy rights to be exercised meaningfully.

Various models of informed consent could be used for health data research. *Specific consent*, the standard in research involving humans, demands that trial participants are informed about and consent to the specific aims and methods of the trial that they take part in. Applied to health data research, this would mean that patients are informed about and consent to each health data research project that accesses and uses the patient's data. Since this would place major obstacles in the way of research, among other drawbacks (compare in the Previous Research section), a number of alternatives are discussed, such as broad consent, tiered consent, meta-consent, and dynamic consent.

In *broad consent* patients consent to the use of their health data for research purposes in unspecified future research. The model of *tiered consent* enables patients to tailor individual consent profiles through the selection of broad and narrow options for data use (*tiers*) and thus allows them to adjust their preferences regarding the kinds of future research to which they are willing to consent. The *dynamic consent* framework facilitates study-specific consent by implementing consent requests and information on a web-based platform. Similarly, *meta-consent* implements tiered consent on a web-based platform, including options to ask for study-specific consent or to opt for broad consent [12-15].

The General Data Protection Regulation (GDPR), which was introduced in the EU in 2018, constitutes an important part of existing governance frameworks. Although it contributes to harmonizing legal approaches to data sharing in member states, it does not prescribe a specific model of consent when consent is used as a basis to legally justify the use of health data for research. In fact, under certain conditions, the GDPR permits accessing health data for research without consent [16,17]. Article 9, paragraph 2, of the GDPR allows EU member states to issue such exemptions for scientific research. The prerequisite is that the processing of personal health data is necessary for the purpose of the research and that safeguards are in place to protect the rights of the "data subjects" in accordance with Article 89, paragraph 1. In addition, the processing must be proportionate to the objective pursued. In Germany, section 27 of the German Federal Data Protection Act constitutes such a national regulation.

The World Medical Association's Declaration of Taipei introduces consent as a necessary condition for using health data in research involving humans with few exemptions and lists information that must be included in informed consent for data sharing and biobanking [18]. These stipulations allow for but do not require specific consent for data stored in repositories that are designed for "multiple and indefinite uses." They are compatible with other models of consent as well.

The German MII continues to consider consent as the appropriate legal basis for authorizing the secondary use of health data for research purposes, as is standard in many other EU countries [3,8]. More specifically, the initiative opts for a standardized model of broad consent. All participating medical centers are asked to use the same information and consent form, to which optional modules (eg, concerning biobanking) can be added according to local needs. These developments raise the question of how informed consent should be implemented in health data research infrastructure and which model of consent is to be favored. Moreover, the first experiences of stakeholders in this implementation process can provide valuable insight and information on the challenges connected to operationalizing informed consent in the context of health data research.

Previous Research

Informed consent safeguards transparency and enables patients to exert control over the use of their personal health data and retain privacy. In the ethical debate on health data research, informed consent thus plays an important role. In addition, in the GDPR, obtaining consent is the standard way to legally justify the use of personal health data for research, although consent can be waived if member states fulfill certain conditions (compare in the Background section).

In the current scholarly ethical debate, there is wide agreement that specific informed consent, an ethical cornerstone of clinical research, has significant drawbacks when applied to biobank and health data research [12]. For a number of reasons, staying with specific consent in the case of health data research appears problematic. Since valuable data research will often involve data from a large number of patients, obtaining consent will be cumbersome and delay or prevent studies [12]. This may impede studies and diminish the positive impact that these studies could have had on health and society. Implementing specific consent on a digital platform may mitigate these problems, but it will still constitute an obstacle to large-scale data research. In addition, patients themselves might be irritated by being asked to consent to data research for single studies if approached repeatedly for a large number of studies [19,20]. As some surveys show, many patients and citizens generally deem the

XSL•FO
RenderX

use of their health data for research acceptable. These surveys also show that while patients and citizens might prefer traditional specific consent, they deem other forms of consent acceptable [21-24]. Finally, although there are serious risks connected to the use of health data, such as loss of data and misuse, these risks are different and arguably less severe than the risks of clinical research [12].

Accordingly, in the conceptual ethical debate about alternative consent models for data- or biospecimen-based research, *broad consent* to unspecified future research is suggested as an alternative to specific consent [12,15]. The main issues highlighted as problematic in this debate so far regard the problem of consenting to research that is as of yet unknown. The information conveyed at the time of obtaining consent is necessarily restricted to general information on risks and opportunities of accessing and storing health data, pseudonymization, and anonymization procedures, data security principles, and features of typical health data research projects. By contrast, no information can be supplied concerning specific research projects, their goals and methods, and their risks. Since the validity of consent is based on the knowledge of what one consents to and since broad consent information is necessarily scarce, the validity of this consent is called into question [14,25,26]. Indeed, the provocative question of whether broad consent can be counted as informed consent at all has been raised [27].

Various other proposals have been made as to how the goal of relieving researchers and research participants of the need for frequent and time-consuming information and consent processes can be achieved while at the same time maintaining a high standard of informational self-determination [13]. Making use of digital infrastructure for collecting and managing consent has been a prominent suggestion in this debate. The *dynamic consent* framework has been discussed as a possible way forward that could allow for study-specific consent through a web-based platform, which would facilitate the process of obtaining consent. The suggested framework of meta-consent combines this idea with the principle of tiered consent, in which patients are enabled to tailor their individual consent profiles through the selection of broad and narrow options for data use (tiers), and allowed to adjust their preferences flexibly from home through a web-based platform [28-32] (for our interviews, we have separated the issues of consent models from their digital or analog implementation. We are thus not discussing meta-consent and dynamic consent by name but are referring to the same concepts when we discuss options of a digital implementation of specific or tiered consent).

With a focus on Europe, it has been discussed how GDPR-based rights can be safeguarded and implemented as part of biobank consent management procedures. These rights include the right to be informed about the collection and use of personal data, the right to rectification of wrong personal data, and the right to erasure of personal data [33]. Since research on specimens always also involves the use of health data, the GDPR is a relevant legal framework for this kind of research as well as for research solely on health data. The findings of this debate thus are generally relevant for health data research as well. Nevertheless, the GDPR biobank debate does not touch upon possible alternative forms of consent. In biobank research today, broad consent is the standard and widely accepted model of consent [15].

Attitudes toward data sharing for health research have been studied in the United States as well as in Europe, and some meta-studies have attempted to summarize the results [22,34-37]. They not only report widespread support for data sharing for research purposes under certain conditions (eg, data security, trust, control, and privacy) but also highlight ethnocultural differences in attitude [38]. Concerns identified relate to sharing of data (or specimens) with for-profit organizations [22] as well as transferring data across borders, especially from EU countries to non-EU countries [37].

For Germany in particular, acceptance of broad consent has been studied among outpatients of a university hospital in northern Germany [23] as well as among participants of a cancer registry in the state of Baden-Württemberg [24]. Both studies observe differing levels of acceptance of broad consent (58.9% vs 86.9%, respectively). In a qualitative study from Germany on the perception of risks and opportunities of health data research in general, stakeholder groups from health research, health care, medical informatics, patient advocacy groups, and politics were interviewed. Regarding informed consent, the study concludes that some stakeholders are afraid that patients may be unduly influenced to provide consent if they are, as patients, under distress or wrongly assume that health data research will directly benefit them [39]. The study does not differentiate between models of consent. Overall, although, on a conceptual level, the pros and cons of models of consent regarding health data research have been debated, there are no studies empirically examining the acceptance and perceptions of advantages, disadvantages, and implementation challenges of different models of consent from stakeholder perspectives.

Objective

In our study, we build on and complement the aforementioned studies by focusing on different models of consent and consent procedures. Different models of consent for health data research have been developed on the basis of conceptual and normative arguments. So far, empirical studies and surveys have not taken up these models and compared their acceptance and perception of drawbacks and opportunities from the perspective of stakeholder groups. We aimed to explore the perception of drawbacks and opportunities of different models of consent among stakeholders from patient associations, data protection, medical informatics, and medical research.

The study is explorative. We assume that the perspectives of stakeholders who are dealing with health data research and informed consent in different roles and with different tasks can help to uncover blind spots in the debate and challenges in practical application that have not received adequate attention in the conceptual, ethical debate on consent models for health data research so far.

XSL•FO
RenderX

Methods

Overview

We interviewed stakeholders who are and will be affected by the implementation of informed consent regarding health data research and can be assumed to have a perspective on the drawbacks and opportunities of consent models. Medical researchers will use the relevant infrastructures and data and will have to abide by consent provisions. Patients are the ones whose individual rights and interests are meant to be protected by consent. Data protection officers safeguard that data handling procedures, including consent procedures, conform to the law. Health IT specialists implement and run the IT infrastructure for storing and managing health data and research requests, as well as consent specifications. We included 3 to 4 participants from each stakeholder group, 13 participants in total.

For the profession-based groups, we focused on stakeholders with ties to German university medical centers. These centers are currently implementing broad consent as devised by the MII. Their employees can thus be assumed to have a level of experiential and expert knowledge that enables them to take a stance on models of consent and their implementation.

Representatives of these groups were identified via a web-based search and contacted by publicly available email addresses. The selection criteria were to include participants who are familiar with the topic of informed consent (documented statements, publications, participation in relevant conferences or workshops, etc) and to ensure diversity of individuals and perspectives within the groups (geographical location, specific area of expertise, and gender). Our group of medical researchers comprised 2 physician researchers and 1 epidemiologist.

To recruit patient representatives familiar with the topics of consent and research, as well as patients' attitudes beyond their own, we checked and contacted several established German patient associations and invited individuals, on the basis of either publicly available records of experience or recommendations by board members of the associations.

Ethical Considerations

The study was approved by the ethics commission of the University of Freiburg (approval number 21-1701). It was conducted in accordance with applicable German and EU data protection regulations. Participation was voluntary and not compensated. All interviewees were informed about the aims, methods, benefits, and risks of participating in the study and gave documented informed consent. Any information allowing personal identification was removed from the interview transcripts before analysis. All data in the publication, as well as any data shared on request, were anonymized.

Interviews

We conducted qualitative, semistructured interviews with the respective stakeholders. We opted for group interviews in order to profit from the effect of homogeneous focus groups to illuminate the topic in breadth and depth [40].

Of all the persons interviewed (N=13), the patient representatives (n=3, 23%), the medical researchers (n=3, 23%), and the health IT specialists (n=4, 31%) were interviewed in groups as planned. For the data protection officers (n=3, 23%), 3 individual interviews had to be conducted due to organizational reasons.

The interviews took place between May and December 2022 in a videoconference format. All interviews were jointly conducted by both authors. The interview guide consisted of a series of general questions and some group-specific questions. One interview session lasted between 60 and 150 minutes. The interviews were recorded and subsequently transcribed by an external service provider.

The interview guide was prefaced by a short explanation of broad consent, tiered consent, specific consent, and digital consent platforms (Multimedia Appendix 1). The guide included 8 questions that were addressed to all groups equally, plus 4 to 8 group-specific questions. The explanatory text and the questions were sent to the participants by email several days before the interview.

Data Analysis

The transcriptions were analyzed in an iterative process using qualitative content analysis. We applied the structuring qualitative content analysis as developed by Mayring [41] and Kuckartz [42]. The software MAXQDA (version 22.61; VERBI Software GmbH, 2022) was used for coding. Given the small number of participants in each group, no quantitative analysis was carried out. No interrater reliability or percent level of agreement were calculated.

The coding categories of the material were initially systematized according to action steps in the process of devising and carrying out a data research project: from planning the research to obtaining consent, adapting and withdrawing consent, to handling of consent data and treatment data (in research contexts), to handling of research results, and to addressing data loss and misuse. The category scheme according to which the material was sorted was developed before the analysis and applied deductively. Where necessary, it was later adapted inductively based on the content of the material. The identification of topics, problem areas, and proposed solutions was carried out inductively using the sorted material.

All items were annotated with a 1-sentence summary. On the basis of these summaries, a table was created to grant an overview of the topics of the interview (Multimedia Appendices 2 and 3). For this purpose, suggestions for solutions as well as objections in regard to either the problem itself or a suggested solution were grouped with the problems they addressed. In some cases, solutions were offered to implied problems without the problems themselves being named. The implied problems were added to the table where needed. This sorting was conducted by SW, and JB reviewed all resulting categories. Any disagreements regarding either the placement of a particular statement in a category or the usefulness or exact borders of the categories were discussed between the 2 researchers until a consensus was reached.

XSL•FO
RenderX

Results

Overview

Our interviewees brought up a large number of problems, concerns, and possible solutions regarding models of consent and their implementation. A complete overview of the results can be found in Table S1 in Multimedia Appendix 2 and Table S1 in Multimedia Appendix 3.

The following representation of results highlights topics that either are addressed repeatedly by different stakeholder groups or have not received much attention in the existing ethical debate on models of consent and can thus be considered new and potentially worthy of further consideration.

Attitudes on Specific and Broad Consent

When discussing models of consent for health data research, our interviewees first focused on broad consent and specific consent. The expressed points of view coincide in many respects with the current state of the scholarly debate on this topic. For example, patient representatives doubted that broad consent conveyed all relevant information and prima facie favored specific consent. Data protection officers agreed that specific consent was better suited than broad consent to inform patients adequately, while they deemed broad consent more advantageous for researchers. Researchers pointed to the amount of work needed to design and obtain specific consent and to obtain project-specific ethics approval. They claimed that in order to be workable, biobanks and large health data registries presupposed broad consent solutions for research:

> *Also natürlich, vielmehr auch spezifisch eigentlich spricht mich prinzipiell am meisten an, weil ich dann am ehesten die Klarheit habe, wofür genau, für welches Ziel und wer und in welchem Rahmen und so.* [Patient representative 1]
>
> *[So of course, it's specific that appeals to me the most in principle, because that's when I have the most clarity about what exactly, for what goal and who and in what context and so on.]* [Patient representative 1, translation by SW]
>
> *Umgekehrt haben wir natürlich bei den projektspezifischen Dingen allein so viel Reibungs- oder Zeitverluste, dass sie durch die Ethik durchmüssen, das dreimal überarbeiten müssen und darüber verliert man auch viel. Und deswegen ist die spezifische Aufklärung manchmal schon durchaus auch ein Problem...* [Medical researcher 3]
>
> *[Conversely, of course, with project-specific things alone we have so much friction and loss of time; they have to pass through ethics, they have to be revised 3 times and thereby there is a lot of loss as well. And therefore specific consent is indeed sometimes also a problem...]* [Medical researcher 3, translation by SW]

As a less obvious point, both data protection officers and IT specialists remarked that designing and implementing broad consent could be very complex and time-consuming as well. Data protection officers mentioned the standard broad consent form of the MII as an example, as it contains a number of modules that can but need not be included in local applications. They pointed out that MII broad consent came in regularly updated versions that needed to be tracked and offered some yes or no options for patients (MII yes or no options include transferring data to countries without EU adequacy decision, access to data from outpatient care, transferring data to and from other medical research facilities, and being contacted in case of incidental findings). They also referred to requests to data use and access committees, pseudonymization, and related data management issues that remain laborious when using broad consent. IT specialists stressed the costs of implementing broad consent and the labor needed to establish consent procedures that were not project specific. They also stressed that once established, the efficiency of digital consent implementation of either broad consent or other consent models could be high:

> *Gerade, wenn wir jetzt Medizininformatik noch mal berücksichtigen, die verschiedenen Einwilligungsversionen, also das macht es extrem komplex. Ich habe einen Patienten, der hat meinetwegen die 1.6D-Variante unterschrieben, hat aber nicht mehr die 1.72 unterschrieben. Das muss ich alles irgendwo ab- oder möchte man abbilden, muss man aus meiner Sicht auch abbilden.* [Data protection officer 1]
>
> *[In particular, if we look at the medical informatics initiative once again, the different versions, so that makes it extremely complex. I have a patient, who has, let's assume, signed version 1.6D, but hasn't signed 1.72 anymore. One has to—or one wants to, has to in my opinion, map that.]* [Data protection officer 1, translation by SW]
>
> *Aber ich meine, die meisten Lösungen, die ich kenne, die versuchen jetzt alles, also sowohl die klinischen als auch die studienspezifischen als auch den Broad Consent über so eine selbe Plattformlösung abzubilden, und dann hat man natürlich vereinheitlichte Prozesse. Dann kann man da sehr effektiv mit umgehen.* [Health IT specialist 3]
>
> *[But I think, most of the solutions I know, they attempt to do it all, so the clinical, as well as the study specific, as well as the broad consent—to map it all over the same platform solution. And then one has standardized processes. Then one can work very effectively.]* [Health IT specialist 3, translation by SW]

Tiered Consent

Tiered consent is the least well-known model among our participants. Nevertheless, it was regarded by some as the best option to allow for something broader than specific consent while still providing patients with a good degree of control over their data:

> *ist aus meiner Sicht das gestufte Modell natürlich, also wenn man schon, sagen wir mal so, in die Forschung grundsätzlich einwilligt als Patient, dann ist das gestufte Modell natürlich das mit der besten*

XSL•FO
RenderX

> *Flexibilität, gegebenenfalls mit der höchsten Transparenz.* [Patient representative 2]
>
> *[From my point of view, of course, if one, let's say, generally consents to research as a patient, then the tiered model is of course the one with the best flexibility, possibly with the highest transparency.]* [Patient representative 2, translation by SW]

It was identified as somewhat less favorable for data research assuming the resulting average permission for the use of data would be narrower than for broad consent. Concerns regarding the efforts needed to manage individual consent profiles were raised. Broad consent, as suggested by the MII in Germany, is seen as not far removed from a model of tiered consent, as it offers options for patients to express preferences (compare in the Attitudes on Specific and Broad Consent section).

Comprehensibility

Lack of comprehensibility of information material was named by all groups as an important obstacle to obtaining adequate consent, regardless of which consent model is chosen. IT representatives and researchers generally ascribed a lack of comprehensibility to legal requirements, including GDPR requirements:

> *Also die Vorgaben...die...für Einwilligungserklärungen vorgegeben werden, die sind nicht mehr leicht verständlich. Die sind sehr komplex. Da sind sehr viele Rechtstexte drin. Also das braucht mir niemand erzählen, dass es Teilnehmer gibt, die sich das bis zum Ende wirklich exakt durchlesen und dann auch noch verstanden haben, das glaube ich nicht.* [Health IT specialist 2]
>
> *[As in, those guidelines...for consent forms, they are no longer easy to understand. They are highly complex. There is a lot of legal text in them. Like, no one has to tell me that there are participants who actually read this in detail, through to the end, and after that have really understood it. I don't believe it.]* [Health IT Specialist 2, translation by SW]

One interviewed data protection officer pointed out that while data protection information may often, by habit, make use of complicated legal terminology, the GDPR explicitly asks for comprehensible information. Solutions offered include oral information by specifically trained personnel alone or together with researchers (for more details concerning this point, see the Expertise and Professional Background of Persons Obtaining Consent section).

Videos on specific data protection issues (such as pseudonymization and anonymization) were regarded by patient representatives and researchers as valuable supplements to in-person information, especially because these materials provide the opportunity for patients to become acquainted with relevant issues at home or in other places distinct from the clinical and therapeutic context.

Finally, IT specialists and data protection officers welcomed the option of web-based patient information sites on health data research as a tool to ameliorate information transfer. Data protection officers argued that such a platform could be used to recruit patients for health data research as well, which could lead to more active involvement of patients:

> *Und dann kann man das noch größer denken. Dann kann man eine solche Plattform, Transparenz, kann man so gestalten, dass die Bürger sich erkundigen können: Was gibt es für Forschungsprojekte? Dann können wiederum Forscher in dieser Plattform ihre Forschungsprojekte präsentieren. Und die Bürger können sich aktiv bewerben darum und sagen: Kann ich da nicht mitmachen bei dem Forschungsprojekt?* [Data protection officer 2]
>
> *[And then one can think that on a bigger scale. One could create such a platform, transparency, fashion it in a way that the citizens can inform themselves: what research projects are there? And then the researchers can use the platform to present their projects. And citizens can actively apply for them and say: Can I join in this research project?]* [Data protection officer 2, translation by SW]

Expertise and Professional Background of Persons Obtaining Consent

Patient representatives were reluctant to accept researcher physicians as the sole contact persons for obtaining in-person health data research information. This is due to suspected partiality and lack of relevant data protection expertise. As a consequence, patient representatives in our interviews preferred researcher-independent information providers with specific data expertise in addition to researchers providing information:

> *Also ich stelle mir schon vor, dass die Aufklärung sehr gut ist und differenziert ist und verständlich ist, also dass mir ein Datenschutzmensch und ein Computermensch erklären, worin die Chancen und Risiken bestehen, dass mir ein medizinisch geschulter Mensch erklärt, was der Sinn dieser ganzen Sache ist und auch mir den Ablauf erklären kann…Aber für mich ist immer wichtig, dass es auch aus kritischer Seite beleuchtet wird. Wenn da nur Befürworter sozusagen Propaganda mir liefern [lacht], dann fühle ich mich nicht wirklich aufgeklärt.* [Patient representative 1]
>
> *[So, I imagine it like this, that the disclosure of information is very good and differentiated and intelligible, as in that a data protection person and an IT person explain where the risks and opportunities lie, and that a medically informed person tells me what the purpose of the whole thing is and also explains the process to me.... But for me, it is always important that the critical aspects are also highlighted. If there are advocates delivering propaganda [laughs] than I do not really feel well informed.]* [Patient representative 1, translation by SW]

Researchers also pointed out that compared to physicians, specifically trained personnel or reception desk personnel are better options for providing information in the informed consent procedure. The main reason invoked is the amount of time

XSL•FO
RenderX

needed to communicate in person: scarce time for researcher physicians that ought to be used for what they deem to be more important tasks. Researchers assume that the amount of time needed to provide information in person will be huge, not only with regard to specific consent but also regarding broad consent. In addition to calling on other professions, researchers also mentioned web-based information and consent management platforms as possible sources of information that may help to reduce the amount of time that they must spend on data information and consent.

Undue Pressure to Consent

Patient representatives referred to undue time pressure when data consent is tied to a clinical appointment. One representative referred to the mental overload of being asked to consent to health data research in a situation when patients seek treatment. Patient representatives generally agreed on the advantages of a data information and consent procedure scheduled independently of clinical appointments: at the clinic after treatment, at home before an appointment, or at home at a point of time completely at one's disposal:

> *Und das ist letztlich auch eine Überforderung in dieser Situation, dass man wirklich noch gut abwägen könnte. Das heißt, man muss erst in einem Zustand sein der Ruhe oder der guten Überlegung, wo man wieder sein Hirn normal zur Verfügung hat, um dann zu entscheiden.* [Patient representative 1]
>
> *[And that is in the end also an overburdening, in that situation, that you could really weigh your options. I mean, one needs to be in a state of calm, or careful consideration, where you can access your brain normally, to decide.]* [Patient representative 1, translation by SW]

Withdrawal of Consent

A topic that was discussed broadly in our interviews that has so far seen little attention in the ethical literature on consent concerns the implementation of withdrawal of consent. Researchers and data protection officers not only highlighted how important they consider this option to be but also pointed out that hospitals have not yet established smooth procedures to deal with withdrawal requests. Up until now, it has been difficult and time-consuming for them to handle such withdrawals. IT specialists were very much aware of this topic as well and highlighted that all digital solutions that are currently in development also aim to make changes to consent and withdrawal of consent transparent and easy to handle. From the perspective of patients, this topic was touched upon when they discussed whether and how their legal rights can be assured. Patient representatives were concerned that patients have no means to verify that their data have been erased (or samples have actually been destroyed) in response to their request:

> *Es muss ein Management geben, das die Einwilligungserklärungen widerrufbar macht. Wo geht der Widerruf ein von dem Patienten? Was wird mit dem Widerruf gemacht? Wie wird der umgesetzt? Wer ist dafür zuständig? In welcher Frist wird das gemacht? Was ist, wenn der Patient Auskunft haben möchte? Genau dasselbe Spiel. Eine Einwilligungslösung, eine Einwilligung bedeutet nicht ein Stück Papier, was im Schrank steht, abgeheftet wird und niemand hat damit mehr was zu tun, sondern all das müsste dann auch sichergestellt sein.* [Data protection officer 2]
>
> *[There needs to be a management, that allows for withdrawal of consent. Where is the withdrawal of a patient registered? What happens with the withdrawal? How is it implemented? Who is responsible for that? In what timeframe? What if the patient is asking for disclosure? The same thing again. A solution for consent, a consent, isn't just a piece of paper that is put into the cabinet, filed away, and no one will ever interact with it anymore, but all of this needs to be ensured.]* [Data protection officer 2, translation by SW]

Consent Management

IT specialists and the data protection officers agreed that not only for a model of tiered consent but also for the wide introduction of broad consent in German university medical centers, a digital system for consent management is required. Having to deal with a vast number of paper-based consent forms does not appear to be viable in any way. They pointed out that even in the case of broad consent, a given consent does not equal another given consent since, first, the MII broad consent offers options for patients to specify consent, and, second, the standardized MII consent sample text is regularly updated and thus comes in a number of versions that need to be tracked (as referred in the Attitudes on Specific and Broad Consent section). If any given consent is to be respected in regard to what it does or does not cover, a system that allows researchers or data trustees to access and manage consent forms is a necessity. Many of our interview partners opined that without this management, broad consent, or any other model of consent, cannot be handled efficiently. It was pointed out that the development of a software application for consent management would likely prove to be too expensive for smaller research projects since there are no such systems readily available on the market yet:

> *Nur ich sage mal, also das Medium an sich ist jetzt wahrscheinlich dann gar nicht so relevant, ob das auf Papier oder digital ist. Wichtig ist, glaube ich, nur, dass man größere Mengen dann einfach nicht mehr auf Papier händeln kann, was dann aber eher eine Rolle spielt in Richtung Datenherausgabe/Probenherausgabe und solche Sachen, dass man da einfach schon gezwungen ist, digital diese Consente abzubilden in irgendeiner Form.* [Health IT specialist 2]
>
> *[Well, I'd say, like, the medium itself probably isn't all that relevant, whether it is paper or digital. What is important, I think, is that with larger numbers you cannot handle it on paper anymore, which in that case is more important in regard to handing out data or specimens and such things, that you are forced in*

XSL•FO
RenderX

that case, to map the consent digitally in some form.] [Health IT specialist 2, translation by SW]

Transfer of Health Data

Statements regarding the transfer of health data to for-profit organizations or non-EU countries revealed a dilemma. Concerns were voiced not only predominantly by patient representatives but also by the other groups that actors in the private sector have to be considered as less trustworthy. Industry partners cannot always be expected to act in the best interest of the public; data security standards might be lower in countries beyond the EU, and, therefore, patients should have the possibility to exclude the transfer of their data to industry partners or third countries:

> *Und ganz schwierig finde ich es, sage ich mal, wenn solche Daten halt, wie gesagt, an die Industrie weitergeleitet werden. Dann sind sie vollkommen außerhalb des Einflussbereiches der Uniklinika und so weiter und so fort. Und da kann man mir noch so viel andere Sachen halt erzählen, es geht in der Privatindustrie immer darum, Geld zu verdienen, immer möglichst viel Geld zu verdienen. Altruistische Motive sehe ich da in der Regel nicht am Werke.* [Data protection officer 3]

> *[And I think it is really problematic, let's say, if such data, as we said, is handed over to industry. Then they are completely outside the area of influence of university medical centers and so on and so forth. And you can tell me different as much as you like, in the private industry, it is always about money. Making as much money as possible. I generally do not see any altruistic motivations at work there.]* [Data protection officer 3, translation by SW]

Regarding the contemporary research landscape, the necessity of exactly these kinds of cooperations was highlighted. The lack of permission to share data with non-European countries (or countries without an EU adequacy decision) was named as a showstopper for biobanks. As an example, the problematic legal status of the sensible wish of US-based companies to extract data from their robotic surgery systems used in German clinics in order to further develop their technologies was mentioned. It was pointed out that even publishing research results in US-based journals might be legally challenging, given the current regulation, if this requires uploading nonanonymized health data to US servers. With regard to broad consent in particular, a consent type that excludes these types of cooperation is regarded to significantly hinder valuable research:

> *Es gab, ich glaube, die Version 1.6D der Einwilligung...die aber die Einschränkung oder die wesentliche Einschränkung hatte, dass der Drittlandstransfer ausgeschlossen war, und dass eine Kooperation oder Zusammenarbeit mit Industrie ausgeschlossen war. ..., da kam für uns auch das Feedback: Ja Entschuldigung, damit, mit diesem Informed Consent können wir eigentlich im Grunde nichts anfangen oder sind extrem beschnitten.* [Data protection officer 1]

> *[There was, I believe, that version 1.6D of the consent...but that had the essential limitation that transfer of data to third countries was excluded, and that also a cooperation with industry was excluded....so we got the feedback: Well, excuse me, but ultimately, we can't make use of this consent or we are extremely limited.]* [Data protection officer 1, translation by SW]

Research Without Consent

One topic that was discussed controversially in regard to the legal framework regulating consent for health data research concerns the need for consent in general. Researchers voiced the opinion that collecting consent, no matter in which form, is a time-consuming process. They claimed that more research could be accomplished if research was possible without consent. While some researchers considered this as a desirable scenario, others highlighted the need to obtain consent and involve patients to ensure their continuous support. Beyond personal attitudes, all the researchers were pessimistic about research without consent becoming a viable option in Germany in the near future, notwithstanding the fact that the GDPR does allow for research without consent under certain conditions (compare with the information in the Background section):

> *Fände ich super. Das funktioniert nicht mit der Organspende und das funktioniert nicht mit dem Broad Consent für Forschung, weil da gibt es viel zu viele, die da... Nein, never ever. Aber finde ich gut.* [Medical researcher 1]

> *[That would be great. But that doesn't work for organ donation and that doesn't work for broad consent to research. Because, there are way too many.... No, never ever. But I think it would be good.]* [Medical researcher 1, translation by SW]

Referring to the GDPR, data protection officers did mention research options without consent. On the basis of specific state laws, local regulations exist which enable research without consent, most typically in cases of retrospective studies that only make use of health data stored at the hospital at which the research takes place (so-called *Eigenforschung*). However, the differences between federal state regulations were mentioned as a hindrance to multicenter cross-state research. A national regulation for health data sharing was suggested as a helpful means to achieve legal and governance clarity on this issue:

> *wenn Sie das mit einer Einwilligung versuchen, wird es immer Personen geben, die aus den verschiedensten Gründen sagen: Möchte ich nicht. Und wenn wir dann aber über ein gesundheitspolitisches oder gesellschaftspolitisches Gesundheitsmanagement sprechen, dann wäre es natürlich schon hilfreich, wenn unter klar definierten Rahmenbedingungen Patientendaten, also die digital vorhandenen Daten, für definierte Forschungsvorhaben verwendet werden dürften.* [Data protection officer 1]

> *[If you try that with consent, there will always be someone saying no, for different reasons. I don't want*

XSL•FO
RenderX

> *to. And if we are talking about health policy, or sociopolitical management of public health, then it would of course be helpful if under clearly defined conditions patient data, data available in digital form, could be used for specific research purposes.]* [Data protection officer 1, translation by SW]

IT specialists were less concerned with this topic, but some voiced the opinion that any regulation that does not leave at least the possibility to opt out would stand in stark opposition to informational self-determination. This assessment was shared by some of the data protection officers.

This topic did not receive much explicit attention from patient representatives. However, when it was mentioned, it was not seen favorable at all. Options of control for patients in regard to the use of their data were rated as highly important, as expressed through a preference for models of specific or tiered consent, while even broad consent was rejected as not giving sufficient degrees of control:

> *wir reden ja zum Glück nicht um einwilligungsfreie Forschung, das ist ja nicht das Thema* [Patient representative 2]
>
> *[Luckily we are not talking about research without consent. It is not our topic here.]* [Patient representative 2, translation by SW]
>
> *oder dass man das [die Einwilligung] dann teilweise gar nicht braucht, gerade wenn man noch nach Europa schaut. Da werden ganz andere Sachen vorgeschlagen, inklusive Genom-Daten, auch der selbstverständlichen Freigabe ohne Rückholmöglichkeit. Also wenn wir jetzt schon in einem Ethikdiskurs sind, finde ich das irgendwie empörend.* [Patient representative 2]
>
> *[So that partially, it [consent] is not needed, in particular if we look at Europe. There are completely different things being suggested, including genomic data, as well as the releasing of data without an option to retract, as a matter of course. If we are already in an ethical discourse here, I find that somewhat outrageous.]* [Patient representative 2, translation by SW]

Discussion

Importance of Findings

In light of the current EU efforts to establish a European health data space and initiatives in other countries, including Germany, to develop digital infrastructures that allow large-scale access to health data for research, determining ethically, legally, and socially acceptable consent procedures is a task of utmost importance [1,3-8].

The attitudes expressed by our sample of German stakeholders provide some insight into how workable and acceptable consent may look like in Germany. As developments with regard to consent within the European health data space and in other countries are still in flux, these findings can prove valuable for other national contexts and the EU as well.

Limitations

Our study has an explorative character, and the significance of the results is necessarily limited due to the small size of the study.

The stakeholders interviewed were limited to the 4 groups of medical researchers at university medical centers, health IT specialists, data protection officers, and patient representatives, all from Germany. With the European health data space in the making, interviews with stakeholders from other European countries on consent and models of consent would be of immense value. Perspectives from other stakeholder groups, such as for-profit health research companies, clinic administrative employees, and members of hospital executive boards, could provide further valuable insights.

Regarding patient attitudes, qualitative and quantitative studies on a larger scale would be needed to provide a clearer picture of interests and concerns.

Options of Choice Within Consent

Our interviews do not reveal unequivocal endorsement of 1 specific model of consent across stakeholder groups. While researchers favored broad consent, patient representatives stressed that only specific consent opens up the possibility of receiving detailed information on planned research and the use of data within this research, with tiered consent being regarded as a possibly acceptable alternative. IT specialists and data protection officers were more neutral regarding this controversy. However, IT specialists pointed out that broad consent (at least as devised by the German MII) has limited advantages over tiered consent in regard to management effort as it comes in versions, contains optional modules for research locations, and provides some options for patients to tailor their consent.

In practice, with regard to implementation and management efforts, broad consent thus might often come close to tiered consent. If this holds true and if one wants to harness the advantage of tiered consent to be able to mirror patient preferences, more efforts should be made to determine which options of choice should actually be included. As it appears now, options are dictated mainly by legal requirements.

This issue appears important as patient representatives in our sample also raised concerns about for-profit organizations accessing their data. Currently, although the MII consent includes some options of choice for patients such as data being transferred to non-EU countries, data transfer to for-profit organizations is not among them. It is obligatorily included in the scope of the MII consent.

Digital Tools and Platforms

There is no doubt that efficient data sharing across borders and between different institutions within the EU presupposes a well-functioning digital infrastructure for health data as well as consent data.

In addition, a variety of issues point toward including digital formats and tools as part of the information and consenting process as well. Researchers are reluctant to spend much time on informing patients and obtaining consent. Conflating clinical

XSL•FO
RenderX

appointments with providing consent to health data research might result in mental overload and undue pressure. Patient information is often incomprehensible and does not enable patients to become well-informed. Digital tools could diminish researcher workload and, at the same time, provide patients with better opportunities to absorb relevant information. All these issues highlight the importance of adequate place, time, and format for providing information. Whether or not researchers should still be involved in this process in person remains a disputed issue.

Ensuring that withdrawal of consent can be handled smoothly and effectively is not as trivial as it may sound. Appropriate workflows and processes should still be implemented in many German hospitals. Digital consent management solutions will have to be able to deal with this challenge. Patient representatives in our sample doubted that it would be possible for patients to obtain information on which consent they have granted to an institution and which personal information is stored and used. They also expressed doubt whether a request for data erasure, a key GDPR right of persons whose data are used, would be fulfilled. These concerns and challenges could be mitigated by digital interfaces and platforms that are especially geared toward handling consent and patient requests concerning erasure, rectification, and withdrawal and are directly linked to the systems storing health data and processing research requests, as already in place in, for example, Estonia [43].

Research Without Consent

The GDPR leaves room for the use of health data for research without consent (compare with the Background section). In specific cases and under specific conditions, German hospitals make use of this option. Due to the local character of these agreements and the diverse landscape of the federal state regulations in Germany, it is not transparent where and under which conditions such research takes place. In addition, the lack of unified regulation renders multicenter collaborations with centers in different states difficult. As a consequence, national regulations on health data research without consent appear desirable.

The stakeholder groups were divided on whether research on health data without consent could become more of a standard approach, compared with the few cases today. While data protection officers pointed out that the GDPR leaves room for this kind of research if national regulation is in place, medical researchers were pessimistic with regard to societal acceptance. Patient representatives in our sample were highly critical of this option.

Opt-out solutions, potentially tiered, in combination with easily accessible information about health data research and the possibility of checking for patients which studies make use of their data may be met with more acceptance since such a framework would be transparent and allow for a certain degree of control.

Implementation of Consent and Consent Management

Implementation of the current procedures to obtain consent is criticized for various reasons, most importantly for lack of comprehensibility of information and implicit pressure to consent due to the integration of the data information and consent process in regular medical appointments. In addition, the impartiality of researchers and their competence regarding data protection are called into question. For researchers, providing information on data research is time-consuming and goes at the expense of other duties.

Detaching consent procedures from the clinical setting; providing additional schooled personnel for informing and obtaining consent; and providing alternative forms of information material, such as videos and well-structured web pages, seem promising avenues to mitigate the concerns and drawbacks of current consent implementation.

Conclusions

Politically, in the EU and countries such as Germany, broad consent and, of lately, opt-out solutions are discussed and pushed forward as part of initiatives to build large health data research infrastructures. Our results indicate that these solutions as such may not do justice to the concerns and demands of patients. From the perspectives of participants in our study, both broad consent and opt-out solutions should include tiered options of consenting and opting out, respectively. These tiered options should mirror major patient concerns.

Regarding the implementation of consent procedures, including broad consent, digital management appears to be indispensable in enabling the handling of requests for information disclosure, erasure, rectification, and withdrawal of information. Participants in our study assumed that digital implementation holds great potential for improving workflows and diminishing researcher workload.

Participants in our study pointed out that having access to regularly updated information on ongoing studies is important. This issue should not be underestimated. It is relevant for both broad consent and opt-out solutions. Again, digital data management may allow the implementation of patient information modules that provide easily accessible, timely, and adequate information about ongoing research.

Finally, participants in our study stressed that it would be desirable to remove health data consent procedures from admission to the hospital and the clinical context in order to allow for a well-considered decision. Although digital consent management could allow for such a separation, implementation of consent in the German MII, for example, still connects hospital admission and consent.

Acknowledgments

Our research was supported by the research project DaDuHealth—Data Access and Data Use in Medical Institutional and Consumer Health Settings. An Ethical, Legal, and Social Analysis, funded by the German Ministry of Education and Research (grant 01GP1902).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide (German).
[PDF File (Adobe PDF File), 123 KB - jmir_v26i1e52180_app1.pdf]

Multimedia Appendix 2
Overview of interview results (English version).
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e52180_app2.xlsx]

Multimedia Appendix 3
Overview of interview results (German version).
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e52180_app3.xlsx]

References

1. European health data space. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2024-04-29]
2. Truyers C, Goderis G, Dewitte H, Akker MV, Buntinx F. The Intego database: background, methods and basic results of a Flemish general practice-based continuous morbidity registration project. BMC Med Inform Decis Mak 2014 Jun 06;14:48 [FREE Full text] [doi: 10.1186/1472-6947-14-48] [Medline: 24906941]
3. Abboud L, Cosgrove S, Kesisoglou I. Country factsheets: mapping health data management systems through country visits: development, needs and expectations of the EHDS. The European Union. 2023. URL: https://tehdas.eu/app/uploads/2023/04/tehdas-mapping-health-data-management-systems-through-country-visits.pdf [accessed 2024-04-29]
4. The guidance function closes report. Én Indgang til sundhedsdata. URL: https://www.vejledningsfunktionen.dk/en/homepage/ [accessed 2024-04-29]
5. InnoHealth datalake project. InnoHealth. URL: https://innohealth.eu/en/datalake/project/ [accessed 2024-04-29]
6. The PHIRI project. The Population Health Information Research Infrastructure. URL: https://www.phiri.eu/project [accessed 2024-04-29]
7. ECRIN overview. European Clinical Research Infrastructure Network. URL: https://ecrin.org/ecrin-overview [accessed 2024-04-29]
8. Über die initiative. Medizin Informatik Initiative. URL: https://www.medizininformatik-initiative.de/de/ueber-die-initiative [accessed 2024-04-29]
9. Big Data und Gesundheit: Datensouveränität als Informationelle Freiheitsgestaltung. Berlin, Germany: Deutscher Ethikrat; 2018.
10. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics 2016 Apr;22(2):303-341. [doi: 10.1007/s11948-015-9652-2] [Medline: 26002496]
11. Ploug T, Holm S. The 'expiry Problem' of broad consent for biobank research - and why a meta consent model solves it. J Med Ethics 2020 Sep;46(9):629-631. [doi: 10.1136/medethics-2020-106117] [Medline: 32098907]
12. Mikkelsen RB, Gjerris M, Waldemar G, Sandøe P. Broad consent for biobanks is best - provided it is also deep. BMC Med Ethics 2019 Oct 15;20(1):71 [FREE Full text] [doi: 10.1186/s12910-019-0414-6] [Medline: 31615491]
13. Wiertz S, Boldt J. Evaluating models of consent in changing health research environments. Med Health Care Philos 2022 Jun;25(2):269-280 [FREE Full text] [doi: 10.1007/s11019-022-10074-3] [Medline: 35286521]
14. Caulfield T, Murdoch B. Genes, cells, and biobanks: yes, there's still a consent problem. PLoS Biol 2017 Jul;15(7):e2002654 [FREE Full text] [doi: 10.1371/journal.pbio.2002654] [Medline: 28742850]
15. Rothstein MA, Harrell HL, Saulnier KM, Dove ES, Fan CT, Hung TH, et al. Broad consent for future research: international perspectives. IRB 2018 Nov 01;40(6):7-12. [doi: 10.1002/eahr.406002]
16. Donnelly M, McDonagh M. Health research, consent and the GDPR exemption. Eur J Health Law 2019 Apr 24;26(2):97-119 [FREE Full text] [doi: 10.1163/15718093-12262427] [Medline: 31059476]
17. Dove ES, Chen J. Should consent for data processing be privileged in health research? A comparative legal analysis. Int Data Priv Law 2020;10(2):117-131. [doi: 10.1093/idpl/ipz023]
18. WMA declaration of Taipei on ethical considerations regarding health databases and biobanks: adopted by the 53rd WMA general assembly, Washington, DC, USA, October 2002 and revised by the 67th WMA general assembly, Taipei, Taiwan, October 2016. World Medical Association. 2016. URL: https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/ [accessed 2024-04-29]
19. Ploug T, Holm S. Informed consent and routinisation. J Med Ethics 2013 Apr;39(4):214-218. [doi: 10.1136/medethics-2012-101056] [Medline: 23222145]

XSL•FO
RenderX

20. Cambon-Thomsen A. The social and ethical issues of post-genomic human biobanks. Nat Rev Genet 2004 Nov;5(11):866-873. [doi: 10.1038/nrg1473] [Medline: 15520796]
21. Morain SR, Largent EA. Public attitudes toward consent when research is integrated into care-any "ought" from all the "is"? Hastings Cent Rep 2021 Mar;51(2):22-32 [FREE Full text] [doi: 10.1002/hast.1242] [Medline: 33840104]
22. Warner TD, Weil CJ, Andry C, Degenholtz HB, Parker L, Carithers LJ, et al. Broad consent for research on biospecimens: the views of actual donors at four U.S. medical centers. J Empir Res Hum Res Ethics 2018 Apr;13(2):115-124 [FREE Full text] [doi: 10.1177/1556264617751204] [Medline: 29390947]
23. Richter G, Krawczak M, Lieb W, Wolff L, Schreiber S, Buyx A. Broad consent for health care-embedded biobanking: understanding and reasons to donate in a large patient sample. Genet Med 2018 Jan;20(1):76-82 [FREE Full text] [doi: 10.1038/gim.2017.82] [Medline: 28640237]
24. Köngeter A, Schickhardt C, Jungkunz M, Bergbold S, Mehlis K, Winkler EC. Patients' willingness to provide their clinical data for research purposes and acceptance of different consent models: findings from a representative survey of patients with cancer. J Med Internet Res 2022 Aug 25;24(8):e37665 [FREE Full text] [doi: 10.2196/37665] [Medline: 36006690]
25. Wiertz S. Die zeitliche dimension des broad consent. Ethik Med 2022 Aug 06;34(4):645-667. [doi: 10.1007/s00481-022-00715-9]
26. Master Z, Nelson E, Murdoch B, Caulfield T. Biobanks, consent and claims of consensus. Nat Methods 2012 Sep;9(9):885-888. [doi: 10.1038/nmeth.2142] [Medline: 22936169]
27. Hofmann B. Broadening consent--and diluting ethics? J Med Ethics 2009 Mar;35(2):125-129. [doi: 10.1136/jme.2008.024851] [Medline: 19181887]
28. Budin-Ljøsne I, Teare HJ, Kaye J, Beck S, Bentzen HB, Caenazzo L, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics 2017 Jan 25;18(1):4 [FREE Full text] [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]
29. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-146 [FREE Full text] [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
30. Ploug T, Holm S. Meta consent: a flexible and autonomous way of obtaining informed consent for secondary research. BMJ 2015 May 07;350:h2146. [doi: 10.1136/bmj.h2146] [Medline: 25952952]
31. Prictor M, Teare HJ, Kaye J. Equitable participation in biobanks: the risks and benefits of a "dynamic consent" approach. Front Public Health 2018;6:253 [FREE Full text] [doi: 10.3389/fpubh.2018.00253] [Medline: 30234093]
32. Nembaware V, Johnston K, Diallo AA, Kotze MJ, Matimba A, Moodley K, et al. A framework for tiered informed consent for health genomic research in Africa. Nat Genet 2019 Nov;51(11):1566-1571 [FREE Full text] [doi: 10.1038/s41588-019-0520-x] [Medline: 31659323]
33. Slokenberga S, Tzortzatou O, Reichel J. GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe. Cham, Switzerland: Springer Nature; 2021.
34. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics 2019 Nov 12;48(1):3-13 [FREE Full text] [doi: 10.1136/medethics-2019-105651] [Medline: 31719155]
35. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73 [FREE Full text] [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]
36. Shah N, Coathup V, Teare H, Forgie I, Giordano GN, Hansen TH, et al. Motivations for data sharing-views of research participants from four European countries: a DIRECT study. Eur J Hum Genet 2019 May;27(5):721-729. [doi: 10.1038/s41431-019-0344-2] [Medline: 30700834]
37. Richter G, Borzikowsky C, Hoyer BF, Laudes M, Krawczak M. Secondary research use of personal medical data: patient attitudes towards data donation. BMC Med Ethics 2021 Dec 15;22(1):164 [FREE Full text] [doi: 10.1186/s12910-021-00728-x] [Medline: 34911502]
38. Brown KM, Drake BF, Gehlert S, Wolf LE, DuBois J, Seo J, et al. Differences in preferences for models of consent for biobanks between black and white women. J Community Genet 2016 Jan;7(1):41-49 [FREE Full text] [doi: 10.1007/s12687-015-0248-y] [Medline: 26304495]
39. Köngeter A, Jungkunz M, Winkler EC, Schickhardt C, Mehlis K. Sekundärnutzung klinischer Daten aus der Patientenversorgung für Forschungszwecke – eine qualitative Interviewstudie zu Nutzen- und Risikopotenzialen aus Sicht von Expertinnen und Experten für den deutschen Forschungskontext. In: Richter G, Loh W, Buyx A, von Kielmansegg SG, editors. Datenreiche Medizin und das Problem der Einwilligung: Ethische, rechtliche und sozialwissenschaftliche Persp. Cham, Switzerland: Springer; 2022:185-210.
40. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995 Jul 29;311(7000):299-302 [FREE Full text] [Medline: 7633241]
41. Mayring P. Qualitative inhaltsanalyse. In: Boehm A, Mengel A, Muhr T, editors. Texte Verstehen: Konzepte, Methoden, Werkzeuge. Konstank, Germany: Universitätsverlag Konstanz; 1994:159-175.
42. Kuckartz U. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. Weinheim, Basel: Beltz Juventa; 2018.

XSL•FO
RenderX

43. Metsallik J, Ross P, Draheim D, Piho G. Ten years of the e-health system in Estonia. CEUR Workshop Proceedings. 2018. URL: https://ceur-ws.org/Vol-2336/MMHS2018_invited.pdf [accessed 2024-07-17]

Abbreviations

EU: European Union
GDPR: General Data Protection Regulation
MII: medical informatics initiative

Edited by S Ma; submitted 29.08.23; peer-reviewed by G Richter, R Hendricks-Sturrup; comments to author 28.11.23; revised version received 15.01.24; accepted 31.05.24; published 07.08.24.

Please cite as:
Wiertz S, Boldt J
Ethical, Legal, and Practical Concerns Surrounding the Implemention of New Forms of Consent for Health Data Research: Qualitative Interview Study
J Med Internet Res 2024;26:e52180
URL: https://www.jmir.org/2024/1/e52180
doi:10.2196/52180
PMID:

©Svenja Wiertz, Joachim Boldt. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Pediatric and Young Adult Household Transmission of the Initial Waves of SARS-CoV-2 in the United States: Administrative Claims Study

Ming Kei Chung[1,2,3*], PhD; Brian Hart[4], PhD; Mauricio Santillana[5], PhD; Chirag J Patel[1*], PhD

[1]Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston, MA, United States

[2]Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Institute of Environment, Energy, and Sustainability, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[4]Optum Labs, Eden Prairie, MN, United States

[5]Machine Intelligence Group for the Betterment of Health and the Environment, Network Science Institute, Northeastern University, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Chirag J Patel, PhD
Department of Biomedical Informatics
Harvard Medical School
Harvard University
10 Shattuck Street, Room 302
Boston, MA, 02115
United States
Phone: 1 617 432 1195
Email: chirag_patel@hms.harvard.edu

Abstract

Background: The correlates responsible for the temporal changes of intrahousehold SARS-CoV-2 transmission in the United States have been understudied mainly due to a lack of available surveillance data. Specifically, early analyses of SARS-CoV-2 household secondary attack rates (SARs) were small in sample size and conducted cross-sectionally at single time points. From these limited data, it has been difficult to assess the role that different risk factors have had on intrahousehold disease transmission in different stages of the ongoing COVID-19 pandemic, particularly in children and youth.

Objective: This study aimed to estimate the transmission dynamic and infectivity of SARS-CoV-2 among pediatric and young adult index cases (age 0 to 25 years) in the United States through the initial waves of the pandemic.

Methods: Using administrative claims, we analyzed 19 million SARS-CoV-2 test records between January 2020 and February 2021. We identified 36,241 households with pediatric index cases and calculated household SARs utilizing complete case information. Using a retrospective cohort design, we estimated the household SARS-CoV-2 transmission between 4 index age groups (0 to 4 years, 5 to 11 years, 12 to 17 years, and 18 to 25 years) while adjusting for sex, family size, quarter of first SARS-CoV-2 positive record, and residential regions of the index cases.

Results: After filtering all household records for greater than one member in a household and missing information, only 36,241 (0.85%) of 4,270,130 households with a pediatric case remained in the analysis. Index cases aged between 0 and 17 years were a minority of the total index cases (n=11,484, 11%). The overall SAR of SARS-CoV-2 was 23.04% (95% CI 21.88-24.19). As a comparison, the SAR for all ages (0 to 65+ years) was 32.4% (95% CI 32.1-32.8), higher than the SAR for the population between 0 and 25 years of age. The highest SAR of 38.3% was observed in April 2020 (95% CI 31.6-45), while the lowest SAR of 15.6% was observed in September 2020 (95% CI 13.9-17.3). It consistently decreased from 32% to 21.1% as the age of index groups increased. In a multiple logistic regression analysis, we found that the youngest pediatric age group (0 to 4 years) had 1.69 times (95% CI 1.42-2.00) the odds of SARS-CoV-2 transmission to any family members when compared with the oldest group (18 to 25 years). Family size was significantly associated with household viral transmission (odds ratio 2.66, 95% CI 2.58-2.74).

Conclusions: Using retrospective claims data, the pediatric index transmission of SARS-CoV-2 during the initial waves of the COVID-19 pandemic in the United States was associated with location and family characteristics. Pediatric SAR (0 to 25 years)

XSL•FO
RenderX

was less than the SAR for all age other groups. Less than 1% (n=36,241) of all household data were retained in the retrospective study for complete case analysis, perhaps biasing our findings. We have provided measures of baseline household pediatric transmission for tracking and comparing the infectivity of later SARS-CoV-2 variants.

(J Med Internet Res 2024;26:e44249) doi:10.2196/44249

KEYWORDS

household transmission; infectivity; pediatric; COVID-19; children; claims data; administrative data; transmission; risk factor; logistic regression; regression; multivariable logistics regression; multiple logistic regression statistics; cohort; retrospective cohort; laboratory; LOINC; infant; toddler; newborn

Introduction

Since the outbreak of COVID-19 in early 2020, over 500 million cases and 6 million deaths have been recorded by the World Health Organization [1]. In the United States alone, over 81 million people have been infected, and almost 1 million people have died due to SARS-CoV-2 [1]. Characterizing the transmission dynamics of the evolving virus provides important evidence to assist in formulating effective public health interventions to fight the pandemic. For example, recent research utilizing data from over 80, 000 household contacts estimated the infectivity of the Omicron and Delta variants of SARS-CoV-2 [2].

The Centers for Disease Control and Prevention (CDC) is responsible for keeping track of the COVID-19 pandemic and providing vital national statistics about COVID-19 in the United States, including new cases, deaths, hospitalizations, and vaccination rates [3]. However, location and family-related information is typically not collected; thus, it is difficult to estimate infectivity, such as the household secondary attack rate (SAR) and odds of SARS-CoV-2 transmission to family members. In European countries, such as England and Denmark, it is possible to link individual SARS-CoV-2 test results to the national administrative records in centralized health care systems and registers and thereby obtain address information [4-6]. Currently, the National COVID Cohort Collaborative is one of the largest United States counterparts with data from over 13 million patients for COVID research [7]. Nevertheless, because of privacy regulations, location information is not readily accessible for estimating the infectivity of SARS-CoV-2 to family members.

There are limited reports on household transmission of SARS-CoV-2 in the United States, especially during the first year of the COVID-19 pandemic. Most of the studies were small in sample size (between 100 and 400), sampled in a particular state, and spanned 1 month to a few months [8-11]. Therefore, the estimated transmission rate of SARS-CoV-2 had a large variation with reported values of SARs between 25% and 38%, and the source of the variation is a challenge to pinpoint. In addition, the relationship between transmission risk and family size has been inconsistent [4,12-14]. A few studies compared the SARs between young adult (age 18 years or older) and child (age below 18 years) contacts and reported that adults were generally more susceptible to infection (between 28% and 48%) than children (between 16% and 28%) [8-10]. Currently, little is known about the relative infectivity of children and young adult index cases in a household setting in the United States [12,13]. Therefore, a full understanding of the baseline transmissibility in the initial waves of SARS-CoV-2 (ie, pre–SARS-CoV-2 Delta variant) in the United States can fill an important knowledge gap. This understanding serves as a crucial complement, enabling informed assessments of the infectivity and effectiveness of vaccination against other variants of concern, such as Omicron and Delta [15].

In this study, we hypothesized that the pediatric household transmission of SARS-CoV-2 in the United States was higher in the first outbreak than in the subsequent outbreaks. We aimed to repurpose administrative data to estimate pediatric household transmission of SARS-CoV-2 in the United States between January 2020 and February 2021, a critical time in the trajectory of the pandemic. We leveraged a large nationwide administrative claims data source and estimated the temporal trend of SARs. We also compared the relative infectivity (as the odds ratios [ORs] of SARS-CoV-2 transmission) between 4 pediatric index case age groups (0 to 4 years, 5 to 11 years, 12 to 17 years, and 18 to 25 years) while adjusting for other factors such as family size, quarter of infection, and residential regions to understand the early SARS-CoV-2 transmission dynamics.

Methods

Data Source

We conducted a retrospective cohort analysis on the household transmission of SARS-CoV-2 using deidentified administrative claims. The data source contained comprehensive inpatient and outpatient claims records, laboratory test results, and individual and various aggregated demographic information. In this study, we leveraged the outpatient laboratory data set for SARS-CoV-2 to assemble the cohort.

Ethical Considerations

The Harvard Medical School Institutional Review Board waived the approval requirement for this study as our database analysis is not regarded as human participants research (IRB20-0935).

Study Population

Identification of Households

We identified familial relationships through a family identification variable available in the demographic data table. Because of the nature of the commercial administrative data, only residential private households were included. Since household identification was based on the relationships between employee subscribers and the beneficiary members, most of the households were single- or 2-generation families.

XSL•FO
RenderX

SARS-CoV-2 Infection

We used the Logical Observation Identifiers Names and Codes (LOINC) information submitted by the laboratory vendors to populate the SARS-CoV-2 records, including test dates, types, and results in the outpatient laboratory data set. An individual could be tested at multiple time points with different test results. We defined a SARS-CoV-2 case as an individual tested positive with polymerase chain reaction (PCR) assay. Among the 19,028,401 rows of records, over 78.30% (n=14,899,238) were PCR tested and 10.22% (n=1,522,923) were positive.

Outcomes

We used SAR to measure household transmission of SARS-CoV-2, which can be interpreted as the probability of new infection among susceptible household members caused by a single case. We defined pediatric index cases (children and young adults) as those aged between 0 and 25 years who had the earliest record of SARS-CoV-2 positives in the households. We defined household contacts as all the individuals within the same household excluding the index cases. For secondary cases, we classified those individuals with SARS-CoV-2 positive 2 to 14 days after the index cases [4].

Assembling the Cohort

We created the analysis cohort as described in Figure 1. We first started with all records tested between January 1, 2020, and February 28, 2021, in the laboratory data set because we were only able to obtain complete data within this period at the time of the analysis. We did not restrict subscribers to have continuous enrollment in their health plans because the window of our analysis was short (2 weeks). There were 19,028,401 rows of records, of which we excluded 11,633,630 (61.14%) mostly because of missing family information to identify the corresponding household members. We tentatively identified 4,270,130 households and further excluded (1) 3,781,308 (88.55%) households because only a single member could be found, (2) 367,961 (8.62%) households because of missing SARS-CoV-2 cases, and (3) 15,208 (0.36%) households with more than 1 index case to avoid complicating the estimation of the relationship between index case and household transmission. After including only households with pediatric index cases, we were able to retain 36,241 effective households for the analysis (0.85% of the 4,270,130 households). We found that 31,177 (86.02%) of the households had no secondary case and 5,064 (13.97%) had at least 1 secondary case.

Figure 1. Workflow to filter administrative claims records and assemble the study cohort. PCR: polymerase chain reaction.

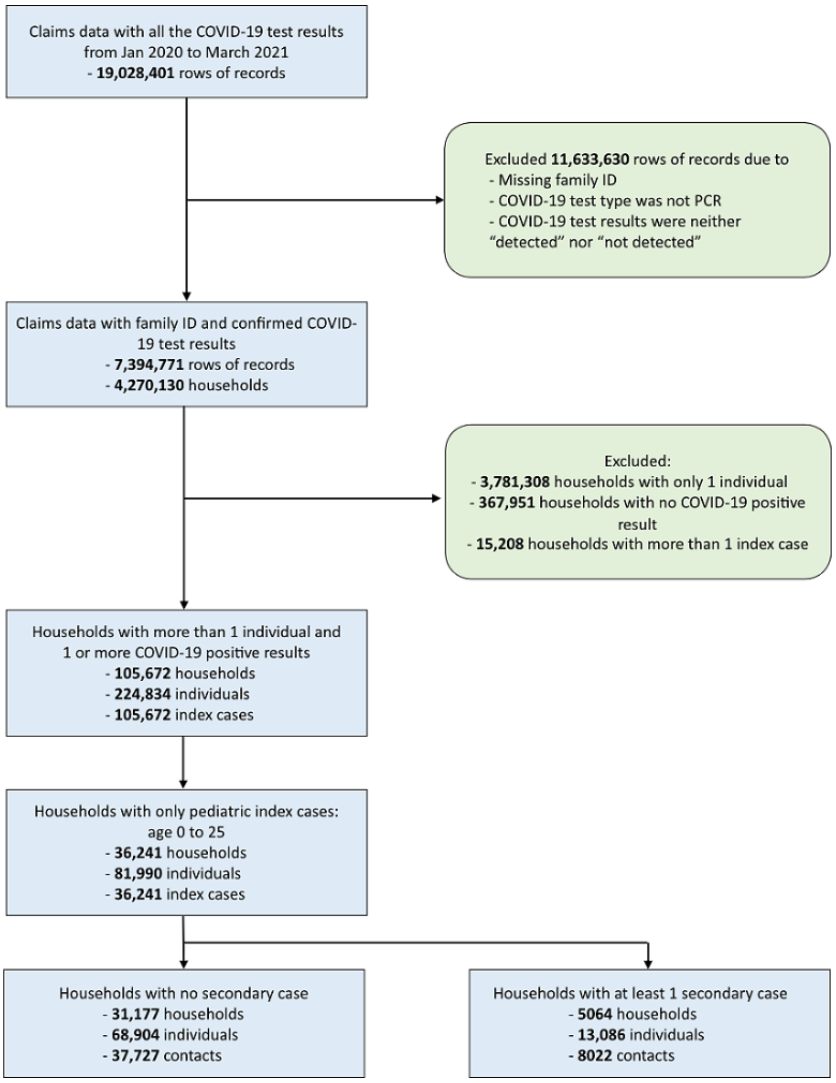

XSL•FO
RenderX

Study Variables

We derived the age variable based on the date of birth information in the administrative claims. We derived the month and quarter of the year of SARS-CoV-2 positive variables based on the date of the laboratory test records. Next, we estimated the family size in each household by counting individuals with the same family identification number. We categorized the date of the first SARS-CoV-2 positive into 5 quarters (between 2020 to 2021). Finally, we used the Census Bureau regions and divisions classification and the residential state information to create the regional division variable with 10 categories.

Statistical Analysis

We calculated SAR as the total number of secondary cases divided by the total number of household contacts in the population. To estimate monthly SARs, we grouped households based on the monthly SARS-CoV-2 test records of the pediatric index case (ie, the first to the last day of the corresponding months). We estimated the standard errors by bootstrapping and calculated the corresponding 95% CIs of the mean SAR.

To understand the dynamic of household transmissions in the pediatric population, we conducted a stratified analysis to estimate the SARs in the age groups of pediatric index cases. Specifically, we created 4 pediatric age groups (0 to 4 years, 5 to 11 years, 12 to 17 years, and 18 to 25 years) and calculated the mean SARs and 95% CIs as described previously.

We conducted a descriptive analysis of the pediatric index cases by the 4 age groups. To quantify the associations between pediatric index case age groups and the household transmission of SARS-CoV-2 to contacts, we used logistic multiple regression to obtain the ORs using the oldest pediatric group (18 to 25 years) as the reference. We modeled the outcome as a binary variable of whether the index case is linked with any secondary cases in the corresponding household. We ran three different regression models to gauge the robustness of our analysis [16]: (1) a crude model with only the pediatric index case age groups as the predictors; (2) an adjusted model that accounted for the influences of sex, season of SARS-CoV-2 test of the pediatric index cases, and family size; and (3) a model further adjusted with the residential region of the pediatric index cases. We set a 2-sided statistical significance level of P=.05 for all the analyses. We executed our analyses using R software (version 4.0.2; R Foundation for Statistical Computing) and used the *boot* package for the bootstrapping step.

Results

We started with over 19 million SARS-CoV-2 test records, which were collected in the first 16 months of the COVID-19 pandemic in the United States We filtered the records and identified 36, 241 eligible households with pediatric index cases for the study. Most of these households (n=31,177, 86.02%) did not report any secondary cases (Figure 1).

We outline the major demographic and associated characteristics of the index case age groups in Table 1. The number of individuals in each age group is correlated with the order of the age groups. The oldest age group (18 to 25 years) had 21.9 times the individuals found in the youngest one (0 to 4 years). In each age category, we found roughly equal numbers of female and male individuals. However, the difference was larger in the oldest age group, which comprised 54.1% (n=13,394) female individuals versus 45.9% (n=11,363) male individuals. The majority of the SARS-CoV-2 cases were detected in the fourth quarter of 2020, followed by the first quarter of 2021. Over half the index cases were found in these 2 quarters. The median household size was 2 across all the age groups. The number of index cases varied greatly across 10 regional divisions, from around 2% in the East South-Central states (Alabama, Kentucky, Mississippi, and Tennessee) to 25% in the South Atlantic states (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington, DC, and West Virginia).

We plot the temporal changes of monthly SARs in Figure 2. Detailed numeric values for the graph can be found in Multimedia Appendix 1. Although our data source contained SARS-CoV-2 test records from January 1, 2020, to March 2021, we could only identify valid households between March 2020 and February 2021. The overall SAR in this period was 23.04% (95% CI 21.88-24.19). Monthly SARs varied from the lowest in September 2020 at 15.6% (95% CI 13.9-17.3) to the highest in April 2020 at 38.3% (95% CI 31.6-45). We observed 3 distinctive trends in the studied period, in that the SAR first generally decreased from April to September 2020, followed by a rebound to a peak SAR of 29% (95% CI 27.8-30.2) in December 2020 before starting to drop again until February 2021.

We show the household transmission rate of SARS-CoV-2 in the pediatric index case age groups in Multimedia Appendix 2. We found a decreasing trend of household transmission, from a SAR of 32.0% to 21.1%, as the age of the index groups increased. Specifically, the SARs were 24.5% for the 12 to 17 years age group (95% CI 23.4-25.6) and 21.1% for the 18 to 25 years age group (95% CI 20.5-21.7). These percentages were significantly lower compared to the SAR of 32.0% for the age group of 0 to 4 years (95% CI 28.8-35.2)

We presented the logistic regression results of the associations between household transmission and pediatric age groups in Table 2. First, we compared the absolute and percentage counts of index cases with and without linkage to any secondary cases by each of the predictors. All the predictors had less than 5% differences between 2 groups other than “2020 third quarter,” “2020 fourth quarter,” and “0-4 years” index age. Our analyses showed that the relationships between the age of the pediatric index cases and household SARS-CoV-2 transmission were consistent across all the investigated models—the younger the index cases were, the higher the odds of transmission. In model 2, we adjusted for all the available relevant predictors and found that the youngest index case group (0 to 4 years) had 1.69 times the odds of transmitting the SARS-CoV-2 to other household contacts (95% CI 1.42-2.00) compared to the oldest reference group (0 to 18 years).

XSL•FO
RenderX

Table 1. Characteristics of the index cases by age groups (N=36,241).

Characteristics of index cases	Age groups			
	0-4 years (n=1132)	5-11 years (n=3378)	12-17 years (n=6974)	18-25 years (n=24,757)
Age (years), median (IQR)	3 (2)	8 (4)	15 (3)	21 (3)
Sex, n (%)				
Female	550 (48.6)	1642 (48.6)	3480 (49.9)	13,394 (54.1)
Male	582 (51.4)	1736 (51.4)	3494 (50.1)	11,363 (45.9)
Quarter of first SARS-CoV-2 positive, n (%)				
2020 first quarter	4 (0.4)	5 (0.1)	11 (0.2)	237 (1)
2020 second quarter	70 (6.2)	182 (5.4)	326 (4.7)	3130 (12.6)
2020 third quarter	117 (10.3)	382 (11.3)	865 (12.4)	6627 (26.8)
2020 fourth quarter	522 (46.1)	1670 (49.4)	3651 (52.4)	10,121 (40.9)
2021 first quarter	419 (37)	1139 (33.7)	2121 (30.4)	4642 (18.8)
Family size, median (IQR)	2 (1)	2 (1)	2 (1)	2 (1)
Regional division, n (%)[a]				
New England	50 (4.4)	129 (3.8)	252 (3.6)	1145 (4.6)
Mid-Atlantic	224 (19.8)	710 (21)	1140 (16.3)	4012 (16.2)
East North Central	56 (4.9)	220 (6.5)	690 (9.9)	2491 (10.1)
West North Central	91 (8)	246 (7.3)	625 (9)	1853 (7.5)
South Atlantic	296 (26.1)	856 (25.3)	1497 (21.5)	6740 (27.2)
East South Central	19 (1.7)	59 (1.7)	193 (2.8)	556 (2.2)
West South Central	136 (12)	454 (13.4)	1123 (16.1)	3209 (13)
Mountain	157 (13.9)	497 (14.7)	1076 (15.4)	3008 (12.2)
Pacific	62 (5.5)	100 (3)	209 (3)	1015 (4.1)
Others	41 (3.6)	107 (3.2)	169 (2.4)	728 (2.9)

[a]States included for the regional divisions are New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont); Mid-Atlantic (New Jersey, New York, and Pennsylvania); East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin); West North Central (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota); South Atlantic (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington, DC, and West Virginia); East South Central (Alabama, Kentucky, Mississippi, and Tennessee); West South Central (Arkansas, Louisiana, Oklahoma, and Texas); Mountain (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming); Pacific (Alaska, California, Hawaii, Oregon, and Washington).

XSL•FO
RenderX

Figure 2. Monthly household secondary attack rates of SARS-CoV-2 for the pediatric index cases in the United States between March 2020 and February 2021. The shaded area represents the 95% CI for each month.

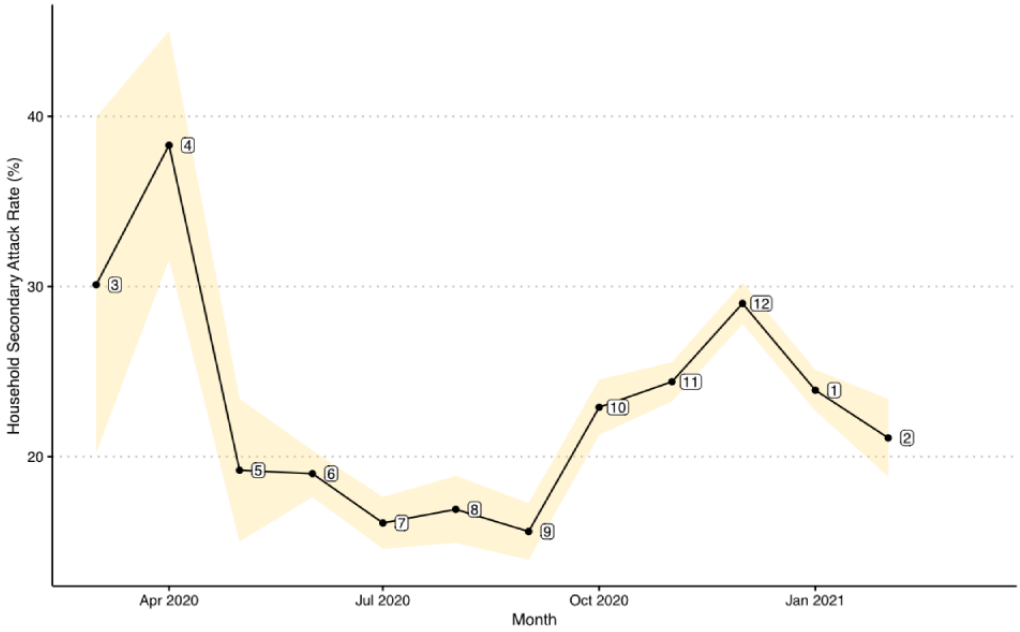

XSL•FO
RenderX

Table 2. Associations between index age groups and SARS-CoV-2 transmission to household contacts.

Variables[a]	Index cases		Odds ratios (95% CI)		
	Associated with no secondary cases	Associated with any secondary cases	Crude model	Model 1	Model 2
Age (years), n (%)					
0-4	934 (3)	198 (3.9)	1.45 (1.23-1.69)	1.69 (1.42-2.01)	1.69 (1.42-2.00)
5-11	2797 (9)	581 (11.5)	1.42 (1.29-1.56)	1.30 (1.17-1.45)	1.29 (1.16-1.44)
12-17	5849 (18.8)	1125 (22.2)	1.31 (1.22-1.41)	1.08 (0.99-1.17)	1.04 (0.96-1.13)
18-25	21,597 (69.3)	3160 (62.4)	1 [Reference]	1 [Reference]	1 [Reference]
Sex, n (%)					
Female	16,461 (52.8)	2603 (51.4)	__[b]	1 [Reference]	1 [Reference]
Male	14,716 (47.2)	2461 (48.6)	—	0.99 (0.93-1.06)	0.99 (0.93-1.06)
Quarter of first SARS-CoV-2 positive, n (%)					
2020 first quarter	234 (0.8)	23 (0.5)	—	0.94 (0.58-1.45)	0.99 (0.61-1.53)
2020 second quarter	3177 (10.2)	531 (10.5)	—	0.81 (0.72-0.90)	0.80 (0.71-0.90)
2020 third quarter	7250 (23.3)	741 (14.6)	—	0.57 (0.52-0.63)	0.57 (0.52-0.63)
2020 fourth quarter	13,322 (42.7)	2642 (52.2)	—	1 [Reference]	1 [Reference]
2021 first quarter	7194 (23.1)	1127 (22.3)	—	0.84 (0.78-0.92)	0.89 (0.82-0.97)
Family size, median (IQR)	2 (1)	2 (1)	—	2.63 (2.55-2.71)	2.66 (2.58-2.74)
Regional divisions[c]					
New England	1370 (4.4)	206 (4.1)	—	—	1 [Reference]
Mid-Atlantic	5259 (16.9)	827 (16.3)	—	—	1.32 (1.10-1.59)
East North Central	2928 (9.4)	529 (10.4)	—	—	1.71 (1.41-2.08)
West North Central	2418 (7.8)	397 (7.8)	—	—	1.83 (1.50-2.25)
South Atlantic	8231 (26.4)	1158 (22.9)	—	—	1.34 (1.12-1.60)
East South Central	720 (2.3)	107 (2.1)	—	—	1.55 (1.17-2.04)
West South Central	4280 (13.7)	642 (12.7)	—	—	1.67 (1.39-2.02)
Mountain	3880 (12.4)	858 (16.9)	—	—	1.80 (1.50-2.17)
Pacific	1192 (3.8)	194 (3.8)	—	—	1.48 (1.17-1.87)
Others	899 (2.9)	146 (2.9)	—	—	1.62 (1.25-2.08)

[a]Variables adjusted in the model are characteristics of the index cases.

[b]Indicates that only the age variable is included in the crude model and regional division is not included in model 1.

[c]States included for the regional divisions are New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont); Mid-Atlantic (New Jersey, New York, and Pennsylvania); East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin); West North Central (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota); South Atlantic (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington, DC, and West Virginia); East South Central (Alabama, Kentucky, Mississippi, and Tennessee); West South Central (Arkansas, Louisiana, Oklahoma, and Texas); Mountain (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming); Pacific (Alaska, California, Hawaii, Oregon, and Washington).

Discussion

Principal Findings

In this study, we leveraged a large deidentified claims database to investigate the COVID-19 pandemic in the United States from January 2020 to March 2021. We identified valid households in the data and estimated the household transmission of SARS-CoV-2 by pediatric to young adult index cases aged 0 to 25 years. We calculated the overall SAR as 23.04% (95% CI 21.88-24.19). In a separate analysis of the same period, we found the overall SAR for all index ages (0 to 65+ years) to be 32.4% (95% CI 32.1-32.8). We observed infection peaks in April and December 2020. In the stratified analysis with pediatric index case age groups, we observed an inverse linear trend between the age groups and SARs—the oldest index case age group (18 to 25 years) had the lowest household transmission, which was statistically significant compared with the youngest group (0 to 4 years). When we analyzed the odds of household transmission of SARS-CoV-2 with a logistic

XSL•FO
RenderX

multiple regression model, we obtained results similar to those found in the SAR analysis. Using the pediatric age group 18 to 25 years as the reference, we found that household transmission of SARS-CoV-2 in the 0 to 4 years age group was 1.69 times the odds of that in the reference group (95% CI 1.42-2.00). The differences between these 2 comparison groups were statistically significant across the 3 investigated models.

Comparison to Prior Work

We found that pediatric-aged individuals did not make up the majority of the household index cases in the early waves of the COVID-19 pandemic. Only 10.9% (11,484/105,672) had pediatric index cases (age 0 to 17 years), which increased to 34% (36,241/105,672) if we included individuals aged 0 to 25 years. Similar proportions were also reported by others worldwide. A United States study reported that 14% (14/101) of the households investigated had pediatric index cases [12], while studies in Asia generally reported a lower percentage of pediatric index cases (age 0 to 20 years) in infected households: between 1% and 5% in China [17,18] and 3% in South Korea [19]. In Europe, the proportions were more comparable to those found in the United States: Denmark reported 5% (index age 0 to 20 years [20]); Greece reported 9% (index age 0 to 18 years [21]), and Switzerland reported 8%, (index age 0 to 16 years [22]). A comprehensive meta-analysis study suggested that the proportion was between 3% and 19% [23]. We emphasize that the SAR of 32.4% estimated for all age groups (0 to 65+ years) was larger than the overall SAR for the 0-to-25-year-old population. Additionally, it was more than 10% higher than the SAR for the young adult (18 to 25 years) population.

The reported household transmission of SARS-CoV-2 by pediatric and young adult index cases in the first year of the pandemic varied. In the United States, SARs in Tennessee and Wisconsin were 53% (95% CI 31-47) and 38% (95% CI 23-56) for index cases aged 0 to 12 years and 12 to 17 years, respectively [12]. The SAR was even higher in Los Angeles, California, at 81.9% (95% CI 72.1-91.9) for index children aged 0 to 18 years [24], which we speculate could be due to the inclusion of households with multiple index cases in the study. In a Denmark study, the SARs for young index case groups were lower, at 19% for the age group 0 to 10 years and 20% for the age group 10 to 20 years). Moreover, the ORs of viral transmission increased significantly as the age of the children index cases decreased [5]: age 0 to 5 years had an OR of 2.17 (95% CI 1.87-2.51), age 5 to 10 years had an OR of 1.66 (95% CI 1.50-1.83), and age 10 to 15 years had an OR of 1.37 (95% CI 1.26-1.49), with the reference age group being 15 to 20 years. Similar inverse relationships between the age of the pediatric index and SARS-CoV-2 transmission were reported in Norway [25], where the SARs were as follows: age 0 to 6 years, 24% (95% CI 20-28); age 7 to 12 years, 14% (95% CI 12-15); age 13 to 16 years, 14% (95% CI 13-16); and age 17 to 20 years, 11% (95% CI 10-13). Meanwhile, in Canada, the ORs were as follows: age 0 to 3 years, 1.43 (95% CI 1.17-1.75); age 4 to 8 years, 1.4 (95% CI 1.18-1.67); and age 9 to 13 years, 1.13 (95% CI 0.97-1.32), with the reference age group being 14 to 17 years.

However, this relationship was not evident in a South Korean investigation [19], where SARs were reported as 5.3% for ages 0 to 9 years (95% CI 1.3-13.7) and 18.6% for ages 10 to 19 years (95% CI 14-24). Upon further investigation, we posit that the SAR in South Korea for the age group spanning 0 to 9 years may not be reliable as it was calculated based on only 3 secondary cases (3/57). Recently, a meta-analysis of 45 studies showed that there was no significant difference in SARs between the age 0 to 19 years and age 20+ years index cases [23].

We found greater household infectivity of SARS-CoV-2 by the youngest index cases relative to their older counterparts (18 to 25 years) in this study. Although viral load is correlated with the spread of COVID-19 [26], several studies have reported that the amount of viral RNA in children is similar, or even higher, than that measured in adults [5,27,28]. Children are generally more likely to be asymptomatic after infection than adults [21,29], and asymptomatic cases are less capable of infecting susceptible individuals than symptomatic cases [18,23,30]. These observations contradict the results from our study and others [5,13,25]. Behavioral factors could play a bigger role in household SARS-CoV-2 transmission. Younger children require frequent and close contact care from their parents, and it may be difficult to quarantine children alone. Together, these factors could increase the infectivity of young children to parents and caregivers significantly when compared to the older, more independent, and self-caring index age group, such as those aged 18 to 25 years.

SARS-CoV-2 transmission is influenced by many factors, such as individuals' demographic profiles and geographic locations [31]. Our regression analysis showed strong temporal and regional variations of the viral transmission in the United States SARs were high and over 30% in March and April 2020. It remained below 20% from May and September and rose to a peak value of approximately 30% in December 2020. We believe that these patterns could be explained partially by a few key events. Initially, the public was not prepared for the SARS-CoV-2, which caused high SARs. Later, national efforts were engaged, including state-level lockdowns and mask-wearing guidelines, to mitigate the risk of SARS-CoV-2 infection and transmission in the community [32]. From September onward to February 2021, the SAR pattern resembled that of the 7-day average of the national new cases [3]. During this period, changes in temperature, an increase in social activities related to the presidential election, and the availability of COVID-19 vaccines could be important factors explaining the observed trends. Among all the United States regions, New England had the lowest household transmission. This could be attributed to its high level of educational attainment, which also generally correlates with socioeconomic status [33].

Household size is a proxy for the number of susceptible contacts in estimating viral transmission. The association between household size and household transmission of SARS-CoV-2 is not consistent in the literature. Like our results, a study conducted in Canada reported a positive relationship (OR 1.63, 95% CI 1.43-1.86) [13]. However, investigations in the United States [12], England [4], and a meta-analysis of 6 international studies reported an inverse trend between the two [14]. Apart from the geographical and temporal factors, variations in the statistical modeling of the transmission outcome could also play a role in the discrepancy. Positive associations between

XSL•FO
RenderX

household size and SARS-CoV-2 transmission were reported for studies defining the outcome as "any secondary cases in the household," whereas a negative relationship was found in those studies modeling the outcome at the individual level.

In addition to the predictors investigated in our analysis, a wide range of factors are known to affect COVID-19 incidence. These include socioeconomic status, ethnicity, household composition, and numerous environmental factors like air pollution and meteorological conditions. For instance, the incidence rate ratios of the Social Vulnerability Index (SVI) to COVID-19 incidence and mortality were 1.14 (95% CI 1.13-1.16) and 1.14 (95% CI 1.12-1.16), respectively [34]. Similar findings were also observed for the subindexes of SVI. A meta-analysis also showed a weak correlation (0.36, 95% CI 0.02-0.62) between the average temperature and daily COVID-19 cases across cities from different countries [35].

Strengths and Limitations

We acknowledge that there are 4 major limitations in our analysis of the administrative claims data. First, because we used data from a private health insurer, we were only able to include working individuals in the United States with insurance coverage. Additionally, adding to the potential bias, only a small proportion of valid households for estimating household SAR were filtered from the raw data (n=35,241, 0.85%; Figure 1). It may not be appropriate to compare the SARs obtained only from the excluded populations, such as families with self-employed and unemployed members, because there could be unique factors affecting SARs in those populations. Nevertheless, prior studies and our preliminary comparisons between the trends of national statistics estimated by claims data and those reported by the CDC are comparable [36,37]. Second, ethnicity is a risk factor for SARS-CoV-2 infection [38,39], and we were unable to account for its influence in estimating the SAR and modeling the household transmission in the logistic regression. Our estimations can only be interpreted as national averages and substantial differences between different ethnic groups are possible. Third, although we used a family identification variable to identify household members, we lacked the capability to fully identify multigenerational households (eg, with at least 1 adult aged 65 years or above, 1 adult in the working group aged 26 to 64 years, and 1 pediatric member aged 0 to 25 years). Commercial insurance subscribers are mostly working adults with spouses and children as dependents in their health plans. However, it is known that older adults (age 65 years and above) are more susceptible to SARS-CoV-2 infection and complications, and the SARs and regression estimates of transmission will likely only be larger and stronger than our study if such information is available. Finally, it is difficult to assess why individuals in a household might seek testing. One reason might be symptoms of COVID-19 among individuals; therefore, the rates will be biased toward estimating the SARs for symptomatic index cases.

Our study also has 2 strengths. First, we used data from the first year of the COVID-19 pandemic, and it is reasonable to assume that most of the contacts were not previously infected with SARS-CoV-2. Therefore, we can obtain an estimation of SARS-CoV-2 household transmission without information about prior history of infection or antibody test results. In addition, compared with the transmission estimations from meta-analysis, we used a single large cohort and were able to minimize the influence of analytic variation such as using a case definition, data filtering scheme, and model with the same set of predictors to estimate the range and heterogeneity of SAR across the time of the pandemic and various age groups of index cases.

Conclusions

Our retrospective administrative claims study of viral transmission with 45,749 household contacts showed that large variations existed across months and regions during the initial waves of SARS-CoV-2 in the United States. The pediatric household SAR decreased from 32.0% to 21.1% as the age of the index cases increased and the SAR for all age groups was higher than the SAR for ages 0 to 25 years. There was also a positive association between family size and household transmission rate.

Acknowledgments

We thank Sarah Daugherty, Caleb Kennedy, Amy Gilmer, Teodi Enrik Racho, Callahan Clark, Michael Lahm, and Elizabeth Erickson for assisting with accessing the administrative claims data. This study was supported by grants from the United States National Institutes of Health through the National Institute for Environmental Health Sciences (ES032470, P30ES000002) and from the United States National Science Foundation through the Northeast Big Data Innovation Hub.

Data Availability

In keeping with the user agreement with the data provider concerning the access to and use of the health research database, individual participant data used in this study and the data dictionary will not be available to the public. Researchers who are interested in the study and able to provide a scientifically sound proposal should contact the corresponding author (CJP) for collaboration. For reproducibility purposes, all analytic code will be provided upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Graph and table of monthly household secondary attack rates of SARS-CoV-2 for pediatric index cases in the United States between March 2020 and February 2021.
[DOCX File , 223 KB - jmir_v26i1e44249_app1.docx]

Multimedia Appendix 2
Household secondary attack rates of SARS-CoV-2 stratified by the pediatric index age groups.
[DOCX File , 96 KB - jmir_v26i1e44249_app2.docx]

References

1. Kim SY, Yeniova A. Global, regional, and national incidence and mortality of COVID-19 in 237 countries and territories, January 2022: a systematic analysis for World Health Organization COVID-19 Dashboard. Life Cycle 2022 May 12;2. [doi: 10.54724/lc.2022.e10]
2. Jørgensen SB, Nygård K, Kacelnik O, Telle K. Secondary attack rates for Omicron and Delta variants of SARS-CoV-2 in Norwegian households. JAMA 2022 Apr 26;327(16):1610-1611 [FREE Full text] [doi: 10.1001/jama.2022.3780] [Medline: 35254379]
3. COVID data tracker. Centers for Disease Control and Prevention. URL: https://covid.cdc.gov/covid-data-tracker [accessed 2022-05-09]
4. Harris RJ, Hall JA, Zaidi A, Andrews NJ, Dunbar JK, Dabrera G. Effect of vaccination on household transmission of SARS-CoV-2 in England. N Engl J Med 2021 Aug 19;385(8):759-760. [doi: 10.1056/nejmc2107717]
5. Lyngse F, Mølbak K, Franck K, Nielsen C, Skov R, Voldstedlund M, et al. Association between SARS-CoV-2 transmissibility, viral load, and age in households. MedRxiv Preprint posted June 4, 2021. [FREE Full text] [doi: 10.1101/2021.02.28.21252608]
6. Lyngse F, Kirkeby C, Halasa T, Andreasen V, Skov R, Møller F, et al. Nationwide study on SARS-CoV-2 transmission within households from lockdown to reopening, Denmark, 27 February 2020 to 1 August 2020. Eurosurveillance 2022 Feb 10;27(6):2001800. [doi: 10.2807/1560-7917.es.2022.27.6.2001800]
7. Madhusoodanan J. Health data for all. Nature 2022 May 03;605(7908):182-183. [doi: 10.1038/d41586-022-01205-0] [Medline: 35505184]
8. Rosenberg ES, Dufort EM, Udo T, Wilberschied LA, Kumar J, Tesoriero J, et al. Association of treatment with hydroxychloroquine or azithromycin with in-hospital mortality in patients with COVID-19 in New York State. JAMA 2020 Jun 23;323(24):2493-2502 [FREE Full text] [doi: 10.1001/jama.2020.8630] [Medline: 32392282]
9. Teherani M, Kao C, Camacho-Gonzalez A, Banskota S, Shane A, Linam W, et al. Burden of illness in households with severe acute respiratory syndrome coronavirus 2-infected children. J Pediatric Infect Dis Soc 2020 Nov 10;9(5):613-616 [FREE Full text] [doi: 10.1093/jpids/piaa097] [Medline: 32780809]
10. Lewis N, Chu V, Ye D, Conners E, Gharpure R, Laws R, et al. Household transmission of severe acute respiratory syndrome coronavirus-2 in the United States. Clin Infect Dis 2021 Oct 05;73(7):1805-1813 [FREE Full text] [doi: 10.1093/cid/ciaa1166] [Medline: 33185244]
11. Dawson P, Rabold E, Laws R, Conners E, Gharpure R, Yin S, et al. Loss of taste and smell as distinguishing symptoms of coronavirus disease 2019. Clin Infect Dis 2021 Feb 16;72(4):682-685 [FREE Full text] [doi: 10.1093/cid/ciaa799] [Medline: 32562541]
12. Grijalva CG, Rolfes MA, Zhu Y, McLean HQ, Hanson KE, Belongia EA, et al. Transmission of SARS-COV-2 infections in households - Tennessee and Wisconsin, April-September 2020. MMWR Morb Mortal Wkly Rep 2020 Nov 06;69(44):1631-1634 [FREE Full text] [doi: 10.15585/mmwr.mm6944e1] [Medline: 33151916]
13. Paul LA, Daneman N, Schwartz KL, Science M, Brown KA, Whelan M, et al. Association of age and pediatric household transmission of SARS-CoV-2 infection. JAMA Pediatr 2021 Nov 01;175(11):1151-1158 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.2770] [Medline: 34398179]
14. Madewell ZJ, Yang Y, Longini IM, Halloran ME, Dean NE. Household transmission of SARS-CoV-2: a systematic review and meta-analysis. JAMA Netw Open 2020 Dec 01;3(12):e2031756 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.31756] [Medline: 33315116]
15. Madewell ZJ, Yang Y, Longini IM, Halloran ME, Dean NE. Household secondary attack rates of sars-CoV-2 by variant and vaccination status: an updated systematic review and meta-analysis. JAMA Netw Open 2022 Apr 01;5(4):e229317 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.9317] [Medline: 35482308]
16. Lee SW. Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee. Life Cycle 2022 Feb 19;2. [doi: 10.54724/lc.2022.e3]
17. Jing Q, Liu M, Zhang Z, Fang L, Yuan J, Zhang A, et al. Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study. Lancet Infect Dis 2020 Oct;20(10):1141-1150 [FREE Full text] [doi: 10.1016/S1473-3099(20)30471-0] [Medline: 32562601]

XSL•FO
RenderX

18. Li F, Li Y, Liu M, Fang L, Dean NE, Wong GWK, et al. Household transmission of SARS-CoV-2 and risk factors for susceptibility and infectivity in Wuhan: a retrospective observational study. Lancet Infect Dis 2021 May;21(5):617-628 [FREE Full text] [doi: 10.1016/S1473-3099(20)30981-6] [Medline: 33476567]
19. Park YJ, Choe YJ, Park O, Park SY, Kim Y, Kim J, COVID-19 National Emergency Response Center, EpidemiologyCase Management Team. Contact tracing during coronavirus disease outbreak, South Korea, 2020. Emerg Infect Dis 2020 Oct;26(10):2465-2468 [FREE Full text] [doi: 10.3201/eid2610.201315] [Medline: 32673193]
20. Lyngse F, Kirkeby C, Halasa T, Andreasen V, Skov R, Møller F, et al. COVID-19 Transmission Within Danish Households: A Nationwide Study from Lockdown to Reopening. MedRxiv Preprint posted online September 9, 2020. [FREE Full text] [doi: 10.1101/2020.09.09.20191239]
21. Maltezou HC, Vorou R, Papadima K, Kossyvakis A, Spanakis N, Gioula G, et al. Transmission dynamics of SARS-CoV-2 within families with children in Greece: A study of 23 clusters. J Med Virol 2021 Mar 26;93(3):1414-1420 [FREE Full text] [doi: 10.1002/jmv.26394] [Medline: 32767703]
22. Posfay-Barbe K, Wagner N, Gauthey M, Moussaoui D, Loevy N, Diana A, et al. COVID-19 in children and the dynamics of infection in families. Pediatrics 2020 Aug;146(2):e20201576. [doi: 10.1542/peds.2020-1576] [Medline: 32457213]
23. Thompson H, Mousa A, Dighe A, Fu H, Arnedo-Pena A, Barrett P, et al. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) setting-specific transmission rates: a systematic review and meta-analysis. Clin Infect Dis 2021 Aug 02;73(3):e754-e764 [FREE Full text] [doi: 10.1093/cid/ciab100] [Medline: 33560412]
24. Tanaka M, Marentes Ruiz CJ, Malhotra S, Turner L, Peralta A, Lee Y, et al. SARS-CoV-2 transmission dynamics in households with children, Los Angeles, California. Front Pediatr 2021;9:752993 [FREE Full text] [doi: 10.3389/fped.2021.752993] [Medline: 35071125]
25. Telle K, Jørgensen SB, Hart R, Greve-Isdahl M, Kacelnik O. Secondary attack rates of COVID-19 in Norwegian families: a nation-wide register-based study. Eur J Epidemiol 2021 Jul 25;36(7):741-748 [FREE Full text] [doi: 10.1007/s10654-021-00760-6] [Medline: 34036466]
26. Marks M, Millat-Martinez P, Ouchi D, Roberts CH, Alemany A, Corbacho-Monné M, et al. Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. Lancet Infect Dis 2021 May;21(5):629-636. [doi: 10.1016/s1473-3099(20)30985-3]
27. Heald-Sargent T, Muller WJ, Zheng X, Rippe J, Patel AB, Kociolek LK. Age-Related differences in nasopharyngeal severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) levels in patients with mild to moderate coronavirus disease 2019 (COVID-19). JAMA Pediatr 2020 Jul 30. [doi: 10.1001/jamapediatrics.2020.3651] [Medline: 32745201]
28. Yonker LM, Neilan AM, Bartsch Y, Patel AB, Regan J, Arya P, et al. Pediatric severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): clinical presentation, infectivity, and immune responses. J Pediatr 2020 Dec;227:45-52.e5 [FREE Full text] [doi: 10.1016/j.jpeds.2020.08.037] [Medline: 32827525]
29. Soriano-Arandes A, Gatell A, Serrano P, Biosca M, Campillo F, Capdevila R, et al. COVID-19 Pediatric Disease in Catalonia Research Group. Household Severe Acute Respiratory Syndrome Coronavirus 2 Transmission and Children: A Network Prospective Study. Clinical Infectious Diseases 2021 Sep 15 2021 Sep 15;73(6):e1269. [doi: 10.20944/preprints202010.0308.v1]
30. Zhu Y, Bloxham C, Hulme K, Sinclair J, Tong Z, Steele L, et al. A meta-analysis on the role of children in severe acute respiratory syndrome coronavirus 2 in household transmission clusters. Clin Infect Dis 2021 Jun 15;72(12):e1146-e1153 [FREE Full text] [doi: 10.1093/cid/ciaa1825] [Medline: 33283240]
31. Wang D, Wu X, Li C, Han J, Yin J. The impact of geo-environmental factors on global COVID-19 transmission: A review of evidence and methodology. Sci Total Environ 2022 Jun 20;826:154182 [FREE Full text] [doi: 10.1016/j.scitotenv.2022.154182] [Medline: 35231530]
32. CDC museum COVID-19 timeline. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/museum/timeline/covid19.html [accessed 2022-05-09]
33. Education levels for each state and county in the United States. US Department of Agriculture. URL: https://data.ers.usda.gov/reports.aspx?ID=17829 [accessed 2022-05-09]
34. Karmakar M, Lantz PM, Tipirneni R. Association of social and demographic factors with COVID-19 incidence and death rates in the US. JAMA Netw Open 2021 Jan 04;4(1):e2036462 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.36462] [Medline: 33512520]
35. Han J, Yin J, Wu X, Wang D, Li C. Environment and COVID-19 incidence: A critical review. J Environ Sci (China) 2023 Feb;124:933-951 [FREE Full text] [doi: 10.1016/j.jes.2022.02.016] [Medline: 36182196]
36. Chung MK, Caboni M, Strandwitz P, D'Onofrio A, Lewis K, Patel CJ. Systematic comparisons between Lyme disease and post-treatment Lyme disease syndrome in the US with administrative claims data. EBioMedicine 2023 Apr;90:104524 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104524] [Medline: 36958992]
37. Schwartz AM, Kugeler KJ, Nelson CA, Marx GE, Hinckley AF. Use of commercial claims data for evaluating trends in Lyme disease diagnoses, United States, 2010-2018. Emerg Infect Dis 2021;27(2):499-507 [FREE Full text] [doi: 10.3201/eid2702.202728] [Medline: 33496238]

XSL•FO
RenderX

38. Mackey K, Ayers CK, Kondo KK, Saha S, Advani SM, Young S, et al. Racial and ethnic disparities in COVID-19-related infections, hospitalizations, and deaths: a systematic review. Ann Intern Med 2021 Mar;174(3):362-373 [FREE Full text] [doi: 10.7326/M20-6306] [Medline: 33253040]
39. Yang T, Emily Choi S, Sun F. COVID-19 cases in US counties: roles of racial/ethnic density and residential segregation. Ethn Health 2021 Jan;26(1):11-21. [doi: 10.1080/13557858.2020.1830036] [Medline: 33059471]

Abbreviations

CDC: Centers for Disease Control and Prevention
LOINC: Logical Observation Identifiers Names and Codes
OR: odds ratio
PCR: polymerase chain reaction
SAR: secondary attack rate
SVI: Social Vulnerability Index

Edited by G Eysenbach, T Leung; submitted 11.11.22; peer-reviewed by DK Yon, JR Medina; comments to author 09.05.23; revised version received 18.07.23; accepted 29.10.23; published 04.01.24.

Please cite as:
Chung MK, Hart B, Santillana M, Patel CJ
Pediatric and Young Adult Household Transmission of the Initial Waves of SARS-CoV-2 in the United States: Administrative Claims Study
J Med Internet Res 2024;26:e44249
URL: https://www.jmir.org/2024/1/e44249
doi:10.2196/44249
PMID:37967280

©Ming Kei Chung, Brian Hart, Mauricio Santillana, Chirag J Patel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Consumption of Health-Related Videos and Human Papillomavirus Awareness: Cross-Sectional Analyses of a US National Survey and YouTube From the Urban-Rural Context

Ashvita Garg[1,2], PhD; Alan G Nyitray[3,4], PhD; James R Roberts[1], MD, MPH; Nicholas Shungu[1,2], MD, MPH; Kenneth J Ruggiero[1], PhD; Jessica Chandler[1,5], PhD; Haluk Damgacioglu[1,2], PhD; Yenan Zhu[1,2], PhD; Naomi C Brownstein[1], PhD; Katherine R Sterba[1,2], PhD; Ashish A Deshmukh[1,2], PhD; Kalyani Sonawane[1,2], PhD

[1]Medical University of South Carolina, Charleston, SC, United States
[2]Hollings Cancer Center, Charleston, SC, United States
[3]Medical College of Wisconsin, Milwaukee, IL, United States
[4]Medical College of Wisconsin Cancer Center, Milwaukee, IL, United States
[5]Applications Center for Healthful Lifestyles, Charleston, SC, United States

Corresponding Author:
Kalyani Sonawane, PhD
Medical University of South Carolina
132 Cannon St
Charleston, SC, 29425
United States
Phone: 1 8438761100
Email: sonawane@musc.edu

Abstract

Background: Nearly 70% of Americans use the internet as their first source of information for health-related questions. Contemporary data on the consumption of web-based videos containing health information among American adults by urbanity or rurality is currently unavailable, and its link with health topic awareness, particularly for human papillomavirus (HPV), is not known.

Objective: We aim to describe trends and patterns in the consumption of health-related videos on social media from an urban-rural context, examine the association between exposure to health-related videos on social media and awareness of health topics (ie, HPV and HPV vaccine), and understand public interest in HPV-related video content through search terms and engagement analytics.

Methods: We conducted a cross-sectional analysis of the US Health Information National Trends Survey 6, a nationally representative survey that collects data from civilian, noninstitutionalized adults aged 18 years or older residing in the United States. Bivariable analyses were used to estimate the prevalence of consumption of health-related videos on social media among US adults overall and by urbanity or rurality. Multivariable logistic regression models were used to examine the association between the consumption of health-related videos and HPV awareness among urban and rural adults. To provide additional context on the public's interest in HPV-specific video content, we examined search volumes (quantitative) and related query searches (qualitative) for the terms "HPV" and "HPV vaccine" on YouTube.

Results: In 2022, 59.6% of US adults (152.3 million) consumed health-related videos on social media, an increase of nearly 100% from 2017 to 2022. Prevalence increased among adults living in both urban (from 31.4% in 2017 to 59.8% in 2022; $P<.001$) and rural (from 22.4% in 2017 to 58% in 2022; $P<.001$) regions. Within the urban and rural groups, consumption of health-related videos on social media was most prevalent among adults aged between 18 and 40 years and college graduates or higher-educated adults. Among both urban and rural groups, adults who consumed health-related videos had a significantly higher probability of being aware of HPV and the HPV vaccine compared with those who did not watch health videos on the internet. The term "HPV" was more frequently searched on YouTube compared with "HPV vaccine." Individuals were most commonly searching for videos that covered content about the HPV vaccine, HPV in males, and side effects of the HPV vaccine.

Conclusions: The consumption of health-related videos on social media in the United States increased dramatically between 2017 and 2022. The rise was prominent among both urban and rural adults. Watching a health-related video on social media was

XSL•FO
RenderX

associated with a greater probability of being aware of HPV and the HPV vaccine. Additional research on designing and developing social media strategies is needed to increase public awareness of health topics.

(J Med Internet Res 2024;26:e49749) doi:10.2196/49749

KEYWORDS

awareness; health awareness; health information; health videos; HINTS; HPV vaccine; HPV; information behavior; information behaviors; information seeking; online information; reproductive health; rural; sexual health; sexually transmitted; social media; STD; STI; urban; video; videos; YouTube

Introduction

Access to the internet is deemed to be an important social determinant of health [1,2]. Nearly 80% of Americans seek health information on the web, and almost 70% are reportedly using the internet as their first source of information for health-related questions [3-5]. Although data suggest that urban-rural disparities in access to the internet continue to persist, with the advent and wide availability of broadband cellular network technology, the "digital divide" is likely to narrow in the future, making web-based health content more accessible to rural and underserved communities [6,7]. There are a plethora of avenues (blogs, informational webpages, social media, and health support groups) and various formats (links, status updates, polls, pictures, and videos) for sharing health information. Literature suggests that user engagement with web-based health content is generally greater if it includes a visual component such as a photo or video [8-11]. Trials evaluating digital interventions that included educational or health-promotional videos have reported positive outcomes (increased knowledge, intentions, and actions in individuals) for disease prevention (human papillomavirus [HPV] vaccination, and cancer, HIV, and fall prevention) [12-15].

The popularity of web-based video-sharing platforms, such as YouTube, has grown in recent years. While internet accessibility has been extensively compared between urban and rural geographies, contemporary US data quantifying geographical differences in exposure to health-related videos on social media are not available. It is important to quantify these data in the context of "place," as the expansion of the cellular network has improved accessibility to health information available on the web in urban as well as rural areas. In this study, we estimate the prevalence of watching health-related videos among rural and urban American adults using the nationally representative Health Information National Trends Survey (HINTS). Furthermore, according to the structural model of health communication, urbanity or rurality is a key social determinant that impacts health outcomes (ie, knowledge, comprehension, health beliefs, health behaviors, and prevention), and its effect is moderated by health media exposure [16,17]. Although empirical research has established that individuals from rural communities have poor health awareness and experience worse disease outcomes [18-21], the intersection of urbanity or rurality with health media consumption and its association with health awareness has not been established using a representative sample of urban and rural adults. We leveraged the HPV (a very common infection that can cause genital warts and cancers) awareness questionnaire from the survey to examine whether watching health-related videos decreases (or increases) the probability of HPV awareness among urban- and rural-dwelling adults. In addition, we evaluated YouTube (the largest video-sharing website) data alongside the survey data to provide context on the public's interest in HPV-related videos.

Methods

Data Sources

This study used the 2022 HINTS 6 data (administered between March and November 2022) to derive the latest and most up-to-date estimates [22]. The 2017 HINTS 5 was used to compute the change in prevalence only. The survey represents a nationally representative sample, and it is conducted by the US National Cancer Institutes (NCI) to monitor changes in the rapidly evolving fields of health communication and health information technology. Data are collected from civilian, noninstitutionalized adults aged 18 years or older residing in the united States. Details regarding the survey design and methodology are available elsewhere [22]. Data on the popularity and viewership of web-based videos were extracted from 2 publicly available sources: the Google Trends and the YouTube Data Application Programming Interface (API; version 3), respectively [23,24]. Google Trends provides access to a largely unfiltered sample of search requests made to the Google search engine. Data are anonymized, categorized by topic for a search query, and can be restricted to provide searches specific to social media platforms. Public interest in a particular topic from around the globe and from more granular areas (country, state, and metropolitan urban areas) is available. Google Trends provides volume-standardized scores of keyword searches; that is, the highest volume of searches originating for a given spatial unit and time is assigned a value of 100, and all subsequent volumes are assigned relative scores. For instance, a score of 50 means that the term was half as popular, and a score of 0 means there was not enough data. The YouTube Data API is the official public-facing API of the web-based video-sharing and social media platform YouTube. The API is used in this study to capture the viewership and engagement of the public with web-based video content on HPV topics. The API provides researchers or users access to the YouTube data server [24]. Researchers can submit queries for topics of interest through the interface using function calls to access information (views, likes, comments, captions, subscribers, and sponsors) for YouTube videos, channels, or playlists.

Variables of Interest

Outcomes of interest were watching a health-related video on social media (primary outcome) and HPV topic awareness

XSL•FO
RenderX

(secondary outcome). The primary outcome was a self-reported indicator of watching health-related web-based videos ("yes" or "no"). Participants were asked, "In the last 12 months, how often did you watch a health-related video on a social media site (for example, YouTube)?" Participants who watched health-related videos on social media in the past year were coded as "yes," and those who reported never watching health videos were coded as "no." Awareness of HPV as a health topic was captured through HPV awareness (Have you ever heard of HPV?) and HPV vaccine awareness (Before today, have you ever heard of the cervical cancer vaccine or HPV shot?) questions with dichotomous response options ("yes" or "no"). The rurality variable was based on the Rural-Urban Commuting Area (RUCA) codes; areas with RUCA codes ≥4 were identified as rural.

We identified the following demographic variables: age (18-40 years, 41-60 years, 61-70 years, and ≥71 years), sex assigned at birth (female or male), and race or ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, and others). We also identified other sociodemographic variables: education (less than high school, high school graduate, some college, and college graduate or higher), annual household income (less than US $50,000, US $50,000-US $74,999, US $75,000-US $99,999, US $100,000-US $199,999, and ≥US $200,000), and cancer history.

To capture the public's interest in HPV-related content, we identified patterns in the search for videos containing HPV information shared on the YouTube platform in the 3 years (2019-2021) preceding the HINTS 6. The standardized search volume overall, by each month, and by metropolitan statistical area were identified for the term "HPV." To help put the search volume about the HPV topic into perspective and get a sense of relative size, we reported data on the HPV vaccine as a comparison term. We restricted the data originating from the United States to searches that were determined to be queried in the context of health. The top related queries, that is, terms that users typed when searching for HPV or HPV vaccine–related videos on YouTube, were identified. The YouTube data API was searched for the keyword terms "HPV" and "HPV vaccine," restricted to videos published before December 2022. The API identified the top 500 videos tagged "HPV" or "HPV vaccine" and provided metrics on views (ie, the total number of views) and engagement (ie, the total number of likes and comments) on those videos.

Statistical Analyses

Descriptive statistics were used to present the sociodemographic characteristics of the final sample by urbanity or rurality and by web-based health video consumption. Bivariable analyses were performed to estimate and compare weighted proportions for HPV and HPV vaccine awareness by consumption of health-related videos on social media; differences between groups (urban vs rural) were tested using the Wald *F* statistic, accounting for multiple comparisons. Similarly, we compared HPV and HPV vaccine awareness between those who watched versus those who did not watch health-related videos among the urban and rural groups using the Wald *F* statistic. Multivariable logistic regression models were constructed to estimate the predicted probability (PP) and difference in predicted probability (dPP) for HPV and HPV vaccine awareness among those who watched health-related videos, compared with those who did not watch health-related videos as the reference group. Standardized search volumes and frequency of engagement (views, likes, and comments) were used to describe patterns in query search and audience engagement with HPV- and HPV vaccine–related videos on YouTube. We used PROC SURVEY procedures to adjust for the complex survey design and weights. All HINTS analyses were conducted using SAS (version 9.4; SAS Institute) statistical software. We used the R 4.2.1 (R Project for Statistical Computing) programming language to derive estimates from the public YouTube data API.

Ethical Considerations

This study was deemed exempt from review by the institutional review board at the Medical University of South Carolina (Pro00128988). All respondents provided electronic consent and received a US $10 Amazon e-gift card from HINTS after completing the survey.

Results

A total of 6252 participants (5442 urban and 811 rural) were identified from 2022 HINTS 6 (Figure S1 and Table S1 in Multimedia Appendix 1). Information on watching web-related videos was available for 6158 participants (5358 urban and 800 rural). In 2022, 59.6% (3473/6158) of the participants, representative of an estimated 152.3 million Americans, watched health-related videos on social media (Multimedia Appendix 2).

The prevalence of watching health-related videos increased by nearly 100% (from 30.1% to 59.6%, $P<.001$) between 2017 and 2022. The increase in prevalence in both urban (from 31.4% in 2017 to 59.8% in 2022; $P<.001$) and rural (from 22.4% in 2017 to 58.0% in 2022; $P<.001$) US adults was statistically significant. The 9%-point difference between urban versus rural adults who watched videos observed during 2017 ($P=.02$) reduced to 1.5%-point in 2022 ($P=.52$).

Table 1 describes the prevalence of watching health-related social media videos among urban and rural adults across sociodemographic strata**.** In the urban group, watching health-related videos on social media was most prevalent among participants in the 18-40 years (1018/1329, 72.5%) age group; for "others," it was 74.9% (303/427); for college graduates or higher-educated adults, it was 69.3% (1563/2427); and for individuals living in households with annual income ≥US $200,000, it was 65.8% (294/453). The prevalence in the rural group was higher among individuals in the 18-40 years age group (100/129, 76.7%) and was highest for college graduates or higher-educated adults (171/282, 68.3%). A table describing the characteristics of participants who watched versus those who did not watch health-related videos on the web is presented in Table S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Prevalence of watching health-related videos on social media among urban and rural adults by sociodemographic strata, 2022 Health Information National Trends Survey (HINTS) 6.

Characteristics[a]	Urban[b] (n=5,358)				*P* value	Rural[b] (n=800)				*P* value
	Watched (n=3069)		Did not watch (n=2289)			Watched (n=404)		Did not watch (n=396)		
	Unweighted, n	Weighted, N (%)	Unweighted, n	Weighted, N (%)		Unweighted, n	Weighted, N (%)	Unweighted, n	Weighted, N (%)	
Age (years)					<.001					<.001
18-40	1018	59,860,846 (72.5)	311	22,650,452 (27.5)		100	6,214,335 (76.7)	29	1,884,450 (23.3)	
41-60	1064	48,666,504 (61.7)	606	30,158,575 (38.3)		152	7,980,602 (66.4)	96	4,039,696 (33.6)	
61-70	589	16,077,457 (48.7)	596	16,949,155 (51.3)		88	2,543,072 (42)	121	3,510,612 (58)	
≥71	376	8,666,125 (31.5)	731	18,823,670 (68.5)		59	1,417,340 (28.6)	138	3,536,450 (71.4)	
Sex					.24					.62
Female	1723	64,714,544 (61.1)	1296	41,153,419 (38.9)		254	8,987,216 (58.4)	220	6,413,763 (41.6)	
Male	1124	60,378,555 (58.2)	876	43,323,733 (41.8)		129	8,193,155 (56.1)	162	6,415,736 (43.9)	
Race or ethnicity					<.001					.07
Non-Hispanic White	1407	66,027,091 (54.7)	1232	54,622,063 (45.3)		279	12,302,142 (56.4)	269	9,527,994 (43.6)	
Non-Hispanic Black	487	16,301,275 (68.3)	312	7,572,056 (31.7)		36	672,640 (42.6)	39	906,835 (57.4)	
Hispanic	589	24,522,106 (65.4)	353	12,987,442 (34.6)		24	1,346,027 (68.5)	20	618,243 (31.5)	
Others	303	16,539,325 (74.9)	124	5,539,062 (25.1)		25	2,227,711 (75.2)	16	733,214 (24.8)	
Education					<.001					.008
Less than high school	121	5,430,530 (40)	197	8,150,012 (60)		15	1,098,344 (43.1)	40	1,449,996 (56.9)	
High school graduate	369	20,727,370 (49.2)	475	21,365,065 (50.8)		75	4,550,256 (49.4)	129	4,668,751 (50.6)	
Some college	798	48,663,876 (59.6)	634	33,008,983 (40.4)		122	6,536,964 (59.1)	103	4,517,706 (40.9)	
College graduate or higher	1563	49,263,948 (69.3)	864	21,842,753 (30.7)		171	4,791,613 (68.3)	111	2,228,229 (31.7)	
Annual household income (US $)					.002					.40
0-49,999	1192	45,185,976 (53.8)	1149	38,727,795 (46.2)		200	8,668,914 (53.5)	248	7,549,338 (46.5)	
50,000-74,999	524	23,538,835 (60.3)	347	15,529,334 (39.7)		62	2,557,549 (54.1)	61	2,167,305 (45.9)	
75,000-99,999	409	17,362,615 (60.4)	272	11,371,208 (39.6)		54	3,144,168 (68.8)	35	1,426,496 (31.2)	
100,000-199,999	646	31,712,428 (65.7)	357	16,548,180 (34.3)		67	2,869,756 (65.2)	39	1,528,962 (34.8)	

XSL•FO
RenderX

Characteristics[a]	Urban[b] (n=5,358)				*P* value	Rural[b] (n=800)				*P* value
	Watched (n=3069)		Did not watch (n=2289)			Watched (n=404)		Did not watch (n=396)		
	Unweighted, n	Weighted, N (%)	Unweighted, n	Weighted, N (%)		Unweighted, n	Weighted, N (%)	Unweighted, n	Weighted, N (%)	
≥200,000	294	14,974,547 (65.8)	159	7,775,563 (34.2)		19	754,263 (64.5)	11	414,392 (35.5)	
Cancer history (yes)	329	9,802,776 (48.8)	411	10,298,077 (51.2)	<.001	65	1,989,773 (49.6)	75	2,020,963 (50.4)	.21

[a]Sociodemographic characteristics were self-reported by participants. Information on age (n=98), sex (n=410), race or ethnicity (n=687), education (n=404), income (n=17), and cancer history (n=370) were missing for some participants.

[b]Urban-rural designations were identified from Rural-Urban Commuting Area Code (RUCA) codes. RUCA ≥4 were classified as rural based on the US Census Bureau and the Rural Health Information Hub.

Awareness of HPV and the HPV vaccine was higher in those who watched versus those who did not watch health-related YouTube videos among both urban and rural groups (Multimedia Appendix 3). Within the urban group, awareness of HPV was 26.1% higher (*P*<.001) and that of the HPV vaccine was 15.2% (*P*<.001) higher in those who reportedly watched health-related videos. Similarly, in the rural group, awareness of HPV (33.5% higher; *P*<.001) and HPV vaccine (23.7% higher; *P*<.001) was greater in those who watched a health-related video in the past 12 months than in those who did not watch videos. The findings were consistent with adjusted estimates from multivariable models (Table S3 in Multimedia Appendix 1). In the rural group, probability of being aware of the HPV (difference in predicted probability [dPP] 21.2%, 95% CI 18.5%-24.0%) and the HPV vaccine (dPP 14.6%, 95% CI 11.7%-17.6%) was greater for individuals who watched a health-related video compared with those who do not watch health videos on the internet. Similarly, the probability of being aware of the HPV (dPP 18.3%, 95% CI 17.1%-19.5%) and the HPV vaccine (dPP 11.8%, 95% CI 10.4%-13.2%) was greater among urban adults who watched versus those who did not watch health-related videos on social media.

Figure 1 illustrates space-time patterns in web-based searches for HPV and HPV vaccine–related videos on YouTube preceding the 2022 HINTS 6. Users predominantly searched for HPV content using the keyword “HPV” and less frequently using the term “HPV vaccine,” as suggested by the average standardized search score of 31 and 5, respectively, over the 3 years. Across all US urban areas, the highest relative volume of searches for HPV and HPV vaccine content on YouTube originated from Ottumwa, IA; Kirksville, MO; and Topeka, KS, respectively. The most popular queries users searched in combination with “HPV” were “HPV vaccine” and “HPV men,” and for “HPV vaccine,” the related query “HPV vaccine side effects” was most popular. The top 500 videos related to HPV garnered a total of 54.8 million views, 769,276 likes, and 68,933 comments since publication, and the top 500 HPV vaccine videos garnered a total of 13.6 million views, 77,813 likes, and 25,277 comments (Figure S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Patterns in seeking human papillomavirus (HPV) and HPV vaccine content on YouTube. Panels (A-C) illustrate the patterns in HPV (blue) and HPV vaccine (red) query searches on YouTube in the years 2019-2021 preceding the 2022 Health Information National Trends Survey (HINTS) 6. (A) This illustrates time trends in query search standardized relative to peak popularity during 2019-2021. (B) This shows popularity across large urban metropolitan areas in the United States. (C) This lists the top 3 most popular terms users searched with "HPV" and "HPV vaccine." Trends in query search on YouTube were examined using Google Trends restricted to 3 years before the 2022 HINTS 6 (ie, 2019-2021).

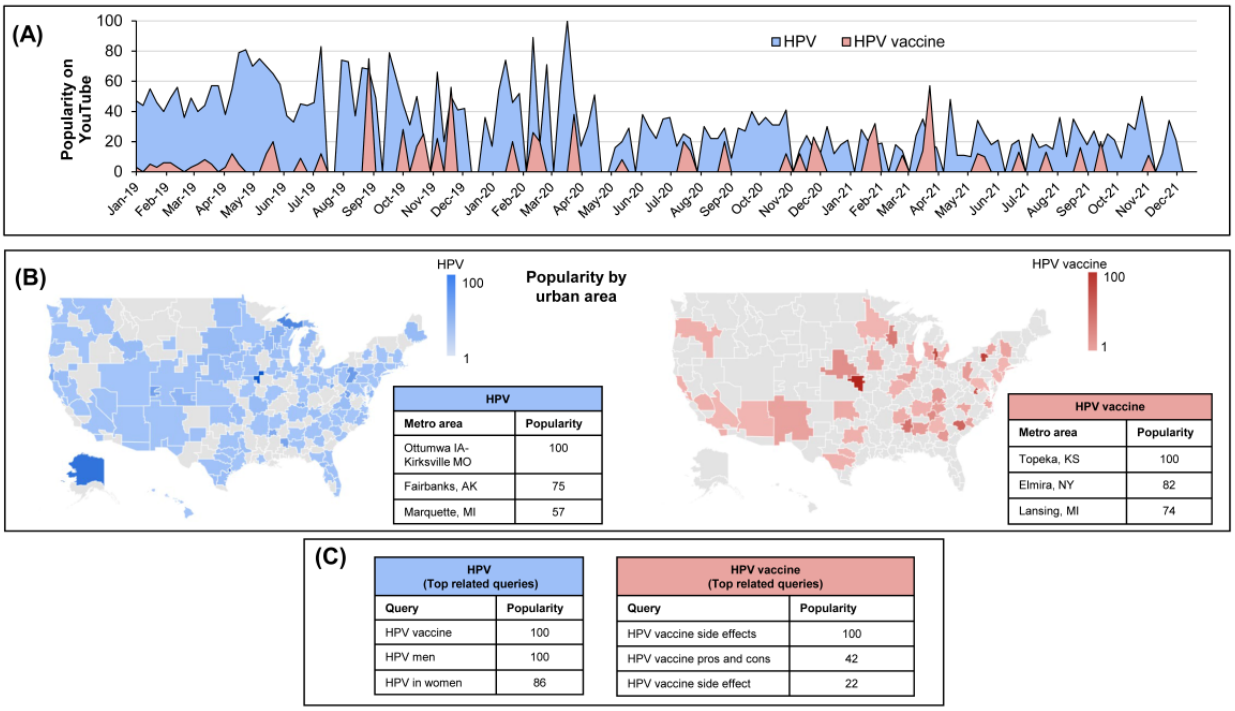

Discussion

This is the first study, to our knowledge, that estimated the prevalence of watching health videos on social media among urban and rural US adults. Overall, more than one-half of American adults in 2022 reported that they watched health-related videos. The prevalence of watching videos was not statistically significantly different between urban (3069/5358, 59.8%) and rural (404/800, 58.0%) residents in the United States. Awareness of HPV topics, including awareness of both the HPV and the HPV vaccine, was statistically significantly higher (all $P<.001$) among participants who watched health-related videos compared with those who did not watch videos, regardless of urbanity or rurality. Our data shows that the number of US individuals watching web-based videos that contain health information increased dramatically during 2017-2022 (100% increase) in the United States, and HPV awareness was relatively higher in individuals who watched health content on web-based platforms compared with individuals who did not watch web-based health-related videos.

Social media platforms have become a prominent source of exposure to health content created by medical experts or organizations as well as nonprofessionals. The number of social media platforms and their users has grown exponentially. YouTube alone is reportedly used by over 84% of urban and 74% of rural US adults [25]. We found that searches for videos pertaining to "HPV men," "HPV women," "HPV vaccine," "HPV vaccine side effects," and "HPV pros and cons" were most commonly on YouTube. These patterns in searches suggest that individuals might be seeking videos to gain an understanding of HPV infection in men (men have relatively lower knowledge about HPV than women) [26] as well as understand the benefits versus harms of getting the HPV vaccine (recent studies report safety concerns are increasingly contributing towards HPV vaccine hesitancy) [27,28]. The millions of views and positive engagement metrics on YouTube illustrate that sharing and seeking information on social media websites have become deeply rooted, regardless of place. The readily available, easily accessible web-based health information ecosystem has altered the ways in which individuals, communities, and societies are exposed to health information on a day-to-day basis.

At an individual level, the nature of information (favorable or unfavorable) can alter perceptions and actions either positively or negatively. For example, exposure to positive information regarding the HPV vaccine on social media increased the likelihood of having favorable views regarding the HPV vaccine in one study [29]. In contrast, encounters with antivaccine content were linked with a decrease in HPV vaccine uptake, mediated through increased vaccine hesitancy [30]. Previous studies report that the majority of HPV vaccination videos on social media tend to be neutral or negative in tone [31,32]. Similar to evidence from individual-level studies, HPV vaccine information exposure at the community or societal level is reportedly positively associated with vaccination rates. In Australia, regional HPV vaccine coverage in females (R=.75, 0.49-0.88) and males (R=.76, 0.51-0.89) was found to be strongly associated with area-level exposure to HPV vaccine tweets [33]. Findings were similar in a US-based study; HPV vaccination information exposure on Twitter (subsequently rebranded X) across the 50 states contributed to 68% of the variance in HPV vaccine coverage among female individuals

XSL•FO
RenderX

and 63% among male individuals [34]. These ecological studies indicate that beyond individual-level interventions, social media can also be exploited for social marketing to increase public awareness of health topics.

Community and societal exemplars of HPV vaccination are also worth mentioning. A few years ago, the country of Denmark saw a sharp decline (reduced to 49.6%) in HPV vaccine coverage due to widespread vaccine misinformation. In response, the Danish Health Authority, the Danish Cancer Society, and the Danish Medical Association launched a strategic national campaign "Stop HPV–stop cervical cancer" (Stop HPV–stop livmoderhalskræft) to disseminate informational videos through social media [35,36]. The campaign was highly successful in combating HPV vaccine hesitancy. Engagement and the number of positive comments regarding the HPV vaccine reached nearly 75% after the campaign launch [35]. The campaign eventually led to the full recovery (100%) of HPV vaccination rates [36]. Ireland faced similar challenges (coverage dropped from 89.7% to 50%) due to widespread misinformation on HPV vaccination and was able to implement a national campaign to tackle plummeting HPV vaccination rates. The Irish National Immunization Office launched media campaigns that included personal narrative videos strongly supporting HPV vaccination alongside a comprehensive training program for health professionals [37]. Multiple studies show that user engagement is highest with informational health content that includes a visual component, such as a photo or video [8-11]. These exemplars of bringing HPV vaccination back on track provide support for communicating health information in video format to a broad audience, including those living in rural areas who are otherwise hard to reach. Targeted outreach through paid promotions and investment in video content creation will be key to maximizing the impact of health communication in rural communities [38]. Pairing with additional resources (eg, provider communication, enhanced access, and reminders) will be necessary to overcome barriers to the uptake of medical intervention in communities.

Indeed, trials as well as a social media campaign study have reported positive and significant health outcomes (increased knowledge, intentions, and actions in individuals) with video-based educational interventions for cancer prevention [12,39]. In a large multicenter trial, an informational HPV vaccination video intervention directed toward parents increased the odds of their child receiving an HPV vaccine dose (odds ratio 3.07, CI 1.47-6.42; P=.003) [39]. Pooled data estimates from a meta-analysis suggest that educational interventions delivered in audio-visual format are effective (risk ratio 1.27, 95% CI 1.15-1.39; 4 studies; 2542 participants) for increasing HPV vaccination [12]. In a social media campaign study that used educational videos for the dissemination of HPV vaccine information, increases in perceived susceptibility (P<.01), perceived severity (P<.05), and self-efficacy (P<.01) were reported in a postcampaign survey [40]. Dissemination of evidence-based and scientifically accurate information through videos will be specifically important for improving HPV topic awareness (overall HPV awareness among urban and rural US adults is reportedly low at 67.2% and 55.8%, respectively) [41] in both urban and rural individuals and might help them make informed decisions regarding HPV vaccination and cancer screenings. This is particularly important in the context of the rapidly rising incidence and burden of cancers caused by HPV in the United States [42-48]. Collectively, in the light of recent evidence highlighting knowledge gaps in the United States regarding susceptibility to HPV and incorrect perceptions regarding the safety and effectiveness of the HPV vaccine [27,28,41,49], this population-based study and previous research provide important evidence in support of using informational videos for increasing health topic awareness in the public.

The limitations of this study should be carefully considered when interpreting the findings. Exposure to health-related videos and HPV awareness were evaluated in general terms ("yes" or "no"), but the nature of the video content (quality, depth, message framing, and tone) and the level of awareness were not available in the survey. Additionally, the awareness questionnaire in the HINTS 6 was limited to HPV; therefore, we could not examine awareness of other health topics in relation to watching health-related videos on social media websites. Nevertheless, HPV awareness was a reasonable outcome because (1) the entire population of US adults (aged 18 years or older and being sexually active) is considered "at risk" of acquiring HPV infection, unlike topics (medication adherence, smoking, pregnancy, etc) that might be relevant to only certain subgroups; and (2) HPV awareness is low in the United States despite being the most common sexually transmitted infection [26]. While the large nationally representative sample allowed us to examine the correlation between consumption of health-related videos on social media and HPV awareness, the cross-sectional nature and residual confounding from variables not captured in the data preclude making causal interpretations. YouTube data illustrating public interest are abstracted from social media analytics platforms and cannot differentiate multiple views from the same users or discern user duplicates. Ancillary data on public interest in health-related videos were from YouTube (the world's largest platform for watching and sharing videos). Survey participants might have watched videos on other social media platforms. However, previous studies examining motivations for using social media platforms have reported "information seeking" as one of their main motivations for watching YouTube content [50,51]. In comparison, in previous studies, the underlying motives for the use of other social platforms were not associated with information seeking (Instagram for "surveillance or knowledge about others;" Facebook and Twitter for "social interaction" and "social connection;" Pinterest for "fashion;" and Reddit for "socializing and community building") [52-56]. Finally, the urban versus rural dichotomy in this study presents a simplified assessment of exposure-outcome. This relationship should be further explored over the urban-rural continuum in future studies.

In conclusion, more than one-half of urban and rural US adults in this nationally representative study watched health-related videos on social media websites. Exposure to videos containing health information was associated with greater HPV topic awareness. Health communication through informational videos should be further researched as a strategy for increasing health topic awareness in clinic and community settings.

XSL•FO
RenderX

Acknowledgments

The research reported in this study was supported by the US National Institute on Minority Health and Health Disparities (K01MD016440), the Clinical and Translational Science Institute (UL1TR001450), and the US National Institutes of Health (R01CA232888).

The funders did not participate in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Conflicts of Interest

KRS and AAD reported receiving a consulting fee from Value Analytics Labs on unrelated projects. All other authors declare no competing interests.

Multimedia Appendix 1
Consumption of health-related videos and HPV awareness: Cross-sectional Analyses of a US national survey and YouTube from the urban-rural context.
[DOCX File , 114 KB - jmir_v26i1e49749_app1.docx]

Multimedia Appendix 2
Prevalence of watching health-related videos on social media in HINTS 5 (2017) versus HINTS 6 (2022).* The figure illustrates the sample size (unweighted N), population size (weighted N), and prevalence for consumption of health-related videos on social media among US adults. Abbreviations: HINTS, Health Information National Trends Survey *P values are for survey design-adjusted Wald F test.
[PNG File , 67 KB - jmir_v26i1e49749_app2.png]

Multimedia Appendix 3
Awareness of HPV and HPV vaccine among urban and rural adults by online health video consumption, 2022 HINTS 6. * The figure illustrates the awareness of HPV and HPV vaccine among urban and rural US adults by consumption by online health video consumption. Abbreviations: HPV, Human papillomavirus *P values are for survey design-adjusted Wald F test.
[PNG File , 37 KB - jmir_v26i1e49749_app3.png]

References

1. Advancing broadband connectivity as a social determinant of health. Federal Communications Commision. URL: https://www.fcc.gov/health/SDOH [accessed 2023-12-14]
2. Poverty and access to Internet, by county. Agency for Healthcare Research and Quality. Rockville, MD; 2023. URL: https://www.ahrq.gov/sdoh/data-analytics/sdoh-tech-poverty.html [accessed 2023-12-14]
3. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and demographic patterns of health-related Internet use among adults in the United States: a secondary data analysis of the health information national trends survey. Int J Environ Res Public Health 2020;17(18):6856 [FREE Full text] [doi: 10.3390/ijerph17186856] [Medline: 32961766]
4. Rutten LJF, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625 [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
5. Prestin A, Vieux SN, Chou WYS. Is online health activity alive and well or flatlining? Findings from 10 years of the health information national trends survey. J Health Commun 2015;20(7):790-798. [doi: 10.1080/10810730.2015.1018590] [Medline: 26042588]
6. Zahnd WE, Bell N, Larson AE. Geographic, racial/ethnic, and socioeconomic inequities in broadband access. J Rural Health 2022;38(3):519-526. [doi: 10.1111/jrh.12635] [Medline: 34792815]
7. Federal Communications Commision. Broadband Deployment Report. 2020. URL: https://openresearch-repository.anu.edu.au/bitstream/1885/277864/1/USTelco_76.pdf [accessed 2023-12-14]
8. Kite J, Foley BC, Grunseit AC, Freeman B. Please like me: Facebook and public health communication. PLoS One 2016;11(9):e0162765 [FREE Full text] [doi: 10.1371/journal.pone.0162765] [Medline: 27632172]
9. Rus HM, Cameron LD. Health communication in social media: message features predicting user engagement on diabetes-related Facebook pages. Ann Behav Med 2016;50(5):678-689 [FREE Full text] [doi: 10.1007/s12160-016-9793-9] [Medline: 27059761]
10. Veale HJ, Sacks-Davis R, Weaver ER, Pedrana AE, Stoové MA, Hellard ME. The use of social networking platforms for sexual health promotion: identifying key strategies for successful user engagement. BMC Public Health 2015;15(1):85 [FREE Full text] [doi: 10.1186/s12889-015-1396-z] [Medline: 25884461]

XSL•FO
RenderX

11. Strekalova YA, Krieger JL. A picture really is worth a thousand words: public engagement with the National Cancer Institute on social media. J Cancer Educ 2017;32(1):155-157 [FREE Full text] [doi: 10.1007/s13187-015-0901-5] [Medline: 26351003]
12. Rodriguez AM, Do TQN, Goodman M, Schmeler KM, Kaul S, Kuo YF. Human papillomavirus vaccine interventions in the U.S.: a systematic review and meta-analysis. Am J Prev Med 2019;56(4):591-602. [doi: 10.1016/j.amepre.2018.10.033] [Medline: 30773231]
13. Acuna N, Vento I, Alzate-Duque L, Valera P. Harnessing digital videos to promote cancer prevention and education: a systematic review of the literature from 2013-2018. J Cancer Educ 2020;35(4):635-642. [doi: 10.1007/s13187-019-01624-0] [Medline: 31707640]
14. Hill AM, McPhail S, Hoffmann T, Hill K, Oliver D, Beer C, et al. A randomized trial comparing digital video disc with written delivery of falls prevention education for older patients in hospital. J Am Geriatr Soc 2009;57(8):1458-1463 [FREE Full text] [doi: 10.1111/j.1532-5415.2009.02346.x] [Medline: 19515102]
15. Hirshfield S, Chiasson MA, Joseph H, Scheinmann R, Johnson WD, Remien RH, et al. An online randomized controlled trial evaluating HIV prevention digital media interventions for men who have sex with men. PLoS One 2012;7(10):e46252 [FREE Full text] [doi: 10.1371/journal.pone.0046252] [Medline: 23071551]
16. Viswanath K, Flynt-Wallington S, Blake KD. Media effects and population health. In: Nabi R, Oliver MB, editors. The SAGE Handbook of Media Processes and Effects. Thousand Oaks, CA: SAGE Publications, Inc; 2009:313-329.
17. Ackerson LK, Viswanath K. The social context of interpersonal communication and health. J Health Commun 2009;14(Suppl 1):5-17. [doi: 10.1080/10810730902806836] [Medline: 19449264]
18. Xu L, Odum M. Cancer awareness and behavioral determinants associated with cancer prevention-a quantitative study among young adults in rural settings. J Cancer Educ 2019;34(3):562-570. [doi: 10.1007/s13187-018-1342-8] [Medline: 29508230]
19. Khapre MP, Mudey A, Goyal RC, Wagh V. Low awareness of diabetes affecting the clinical outcome of patient: a cross-sectional study conducted in rural tertiary care hospital. Int J Biol Med Res 2011;2(3):627-630 [FREE Full text]
20. Chow CK, Teo KK, Rangarajan S, Islam S, Gupta R, Avezum A, et al. Prevalence, awareness, treatment, and control of hypertension in rural and urban communities in high-, middle-, and low-income countries. JAMA 2013;310(9):959-968 [FREE Full text] [doi: 10.1001/jama.2013.184182] [Medline: 24002282]
21. Tummalapalli SL, Vittinghoff E, Crews DC, Cushman M, Gutiérrez OM, Judd SE, et al. Chronic kidney disease awareness and longitudinal health outcomes: results from the reasons for geographic and racial differences in stroke study. Am J Nephrol 2020;51(6):463-472 [FREE Full text] [doi: 10.1159/000507774] [Medline: 32349001]
22. Health Information National Trends Survey (HINTS). National Cancer Institute. URL: https://hints.cancer.gov/ [accessed 2023-12-14]
23. Google Trends. URL: https://trends.google.com/trends/ [accessed 2023-12-14]
24. YouTube Data API. URL: https://developers.google.com/youtube/v3 [accessed 2023-12-14]
25. Social media fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/#panel-3814afe3-3f3c-4623-910b-8a6a37885ab8 [accessed 2023-12-14]
26. Suk R, Montealegre JR, Nemutlu GS, Nyitray AG, Schmeler KM, Sonawane K, et al. Public knowledge of human papillomavirus and receipt of vaccination recommendations. JAMA Pediatr 2019;173(11):1099-1102 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.3105] [Medline: 31524931]
27. Sonawane K, Zhu Y, Montealegre JR, Lairson DR, Bauer C, McGee LU, et al. Parental intent to initiate and complete the human papillomavirus vaccine series in the USA: a nationwide, cross-sectional survey. Lancet Public Health 2020;5(9):e484-e492 [FREE Full text] [doi: 10.1016/S2468-2667(20)30139-0] [Medline: 32707126]
28. Sonawane K, Lin YY, Damgacioglu H, Zhu Y, Fernandez ME, Montealegre JR, et al. Trends in human papillomavirus vaccine safety concerns and adverse event reporting in the United States. JAMA Netw Open 2021;4(9):e2124502 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.24502] [Medline: 34533574]
29. Llavona-Ortiz JY, Spanos KE, Kraschnewski JL, D'Souza G, Myrick JG, Sznajder KK, et al. Associations between human papillomavirus vaccine decisions and exposure to vaccine information in social media. Cancer Control 2022;29:10732748221138404 [FREE Full text] [doi: 10.1177/10732748221138404] [Medline: 36394959]
30. Argyris YA, Kim Y, Roscizewski A, Song W. The mediating role of vaccine hesitancy between maternal engagement with anti- and pro-vaccine social media posts and adolescent HPV-vaccine uptake rates in the US: the perspective of loss aversion in emotion-laden decision circumstances. Soc Sci Med 2021;282:114043 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114043] [Medline: 34147269]
31. Basch CH, Hillyer GC, Berdnik A, Basch CE. YouTube™ videos related to human papillomavirus: the need for professional communication. Int J Adolesc Med Health 2016;30(1). [doi: 10.1515/ijamh-2015-0122] [Medline: 27060736]
32. Ekram S, Debiec KE, Pumper MA, Moreno MA. Content and commentary: HPV vaccine and YouTube. J Pediatr Adolesc Gynecol 2019;32(2):153-157. [doi: 10.1016/j.jpag.2018.11.001] [Medline: 30445163]
33. Dyda A, Shah Z, Surian D, Martin P, Coiera E, Dey A, et al. HPV vaccine coverage in Australia and associations with HPV vaccine information exposure among Australian Twitter users. Hum Vaccin Immunother 2019;15(7-8):1488-1495 [FREE Full text] [doi: 10.1080/21645515.2019.1596712] [Medline: 30978147]

XSL•FO
RenderX

34. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine 2017;35(23):3033-3040 [FREE Full text] [doi: 10.1016/j.vaccine.2017.04.060] [Medline: 28461067]
35. Pedersen EA, Loft LH, Jacobsen SU, Søborg B, Bigaard J. Strategic health communication on social media: insights from a Danish social media campaign to address HPV vaccination hesitancy. Vaccine 2020;38(31):4909-4915 [FREE Full text] [doi: 10.1016/j.vaccine.2020.05.061] [Medline: 32482460]
36. Hansen PR, Schmidtblaicher M, Brewer NT. Resilience of HPV vaccine uptake in Denmark: decline and recovery. Vaccine 2020;38(7):1842-1848. [doi: 10.1016/j.vaccine.2019.12.019] [Medline: 31918860]
37. Corcoran B, Clarke A, Barrett T. Rapid response to HPV vaccination crisis in Ireland. Lancet 2018;391(10135):2103 [FREE Full text] [doi: 10.1016/S0140-6736(18)30854-7] [Medline: 29856339]
38. Crawford D. Catalyzing your communication: HPV communications moving forward. The National HPV vaccination Roundtable. 2023. URL: https://www.youtube.com/watch?v=iv_jjM8ykLE [accessed 2023-12-14]
39. Dixon BE, Zimet GD, Xiao S, Tu W, Lindsay B, Church A, et al. An educational intervention to improve HPV vaccination: a cluster randomized trial. Pediatrics 2019;143(1):e20181457 [FREE Full text] [doi: 10.1542/peds.2018-1457] [Medline: 30530637]
40. Hughes CT, Kirtz S, Ramondetta LM, Lu Q, Cho D, Katzin C, et al. Designing and implementing an educational social media campaign to increase HPV vaccine awareness among men on a large college campus. Am J Health Educ 2020;51(2):87-97. [doi: 10.1080/19325037.2020.1722297]
41. Mohammed KA, Subramaniam DS, Geneus CJ, Henderson ER, Dean CA, Subramaniam DP, et al. Rural-urban differences in human papillomavirus knowledge and awareness among US adults. Prev Med 2018;109:39-43 [FREE Full text] [doi: 10.1016/j.ypmed.2018.01.016] [Medline: 29378268]
42. Deshmukh AA, Suk R, Shiels MS, Sonawane K, Nyitray AG, Liu Y, et al. Recent trends in squamous cell carcinoma of the anus incidence and mortality in the United States, 2001-2015. J Natl Cancer Inst 2020;112(8):829-838 [FREE Full text] [doi: 10.1093/jnci/djz219] [Medline: 31742639]
43. Damgacioglu H, Lin YY, Ortiz AP, Wu CF, Shahmoradi Z, Shyu SS, et al. State variation in squamous cell carcinoma of the anus incidence and mortality, and association with HIV/AIDS and smoking in the United States. J Clin Oncol 2023;41(6):1228-1238 [FREE Full text] [doi: 10.1200/JCO.22.01390] [Medline: 36441987]
44. Shahmoradi Z, Damgacioglu H, Clarke MA, Wentzensen N, Montealegre J, Sonawane K, et al. Cervical cancer incidence among US women, 2001-2019. JAMA 2022;328(22):2267-2269 [FREE Full text] [doi: 10.1001/jama.2022.17806] [Medline: 36409512]
45. Deshmukh AA, Suk R, Shiels MS, Damgacioglu H, Lin YY, Stier EA, et al. Incidence trends and burden of human papillomavirus-associated cancers among women in the United States, 2001-2017. J Natl Cancer Inst 2021;113(6):792-796 [FREE Full text] [doi: 10.1093/jnci/djaa128] [Medline: 32833021]
46. Lin YY, Damgacioglu H, Suk R, Wu CF, Zhu Y, Ortiz AP, et al. Trends in the incidence of human papillomavirus-associated cancers by county-level income and smoking prevalence in the United States, 2000-2018. JNCI Cancer Spectr 2022;6(2):pkac004 [FREE Full text] [doi: 10.1093/jncics/pkac004] [Medline: 35603851]
47. Damgacioglu H, Sonawane K, Chhatwal J, Lairson DR, Clifford GM, Giuliano AR, et al. Long-term impact of HPV vaccination and COVID-19 pandemic on oropharyngeal cancer incidence and burden among men in the USA: a modeling study. Lancet Reg Health Am 2022;8:100143 [FREE Full text] [doi: 10.1016/j.lana.2021.100143] [Medline: 34927126]
48. Damgacioglu H, Sonawane K, Zhu Y, Li R, Balasubramanian BA, Lairson DR, et al. Oropharyngeal cancer incidence and mortality trends in all 50 states in the US, 2001-2017. JAMA Otolaryngol Head Neck Surg 2022;148(2):155-165 [FREE Full text] [doi: 10.1001/jamaoto.2021.3567] [Medline: 34913945]
49. Nguyen KH, Santibanez TA, Stokley S, Lindley MC, Fisher A, Kim D, et al. Parental vaccine hesitancy and its association with adolescent HPV vaccination. Vaccine 2021;39(17):2416-2423 [FREE Full text] [doi: 10.1016/j.vaccine.2021.03.048] [Medline: 33775438]
50. Khan ML. Social media engagement: what motivates user participation and consumption on YouTube? Comput Hum Behav 2017;66:236-247 [FREE Full text] [doi: 10.1016/j.chb.2016.09.024]
51. Park DY, Goering EM. The health-related uses and gratifications of YouTube: motive, cognitive involvement, online activity, and sense of empowerment. J Consum Health Internet 2016;20(1-2):52-70. [doi: 10.1080/15398285.2016.1167580]
52. Sheldon P, Bryant K. Instagram: motives for its use and relationship to narcissism and contextual age. Comput Hum Behav 2016;58:89-97 [FREE Full text] [doi: 10.1016/j.chb.2015.12.059]
53. Kim B, Kim Y. Facebook versus Instagram: how perceived gratifications and technological attributes are related to the change in social media usage. Soc Sci J 2019;56(2):156-167 [FREE Full text] [doi: 10.1016/j.soscij.2018.10.002]
54. Chen GM. Tweet this: a uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. Comput Hum Behav 2011;27(2):755-762 [FREE Full text] [doi: 10.1016/j.chb.2010.10.023]
55. Mull IR, Lee SE. "PIN" pointing the motivational dimensions behind Pinterest. Comput Hum Behav 2014;33:192-200 [FREE Full text] [doi: 10.1016/j.chb.2014.01.011]

XSL•FO
RenderX

56. Moore CM. Redditors Revealed: A Uses and Gratifications Approach Motivational Factors Influencing Users of the Social News Website Reddit.com.: Hawaii Pacific University; 2015. URL: https://www.proquest.com/openview/f62dae286f907155ac9305873e941c83/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2023-12-14]

Abbreviations

API: Application Programming Interface
dPP: difference in predicted probability
HINTS: Health Information National Trends Survey
HPV: human papillomavirus
NCI: National Cancer Institutes
PP: predicted probability
RUCA: Rural-Urban Commuting Area

Edited by A Mavragani; submitted 07.06.23; peer-reviewed by A Marcon, M Rahman; comments to author 12.10.23; revised version received 01.11.23; accepted 10.12.23; published 15.01.24.

Please cite as:
Garg A, Nyitray AG, Roberts JR, Shungu N, Ruggiero KJ, Chandler J, Damgacioglu H, Zhu Y, Brownstein NC, Sterba KR, Deshmukh AA, Sonawane K
Consumption of Health-Related Videos and Human Papillomavirus Awareness: Cross-Sectional Analyses of a US National Survey and YouTube From the Urban-Rural Context
J Med Internet Res 2024;26:e49749
URL: https://www.jmir.org/2024/1/e49749
doi:10.2196/49749
PMID:38224476

©Ashvita Garg, Alan G Nyitray, James R Roberts, Nicholas Shungu, Kenneth J Ruggiero, Jessica Chandler, Haluk Damgacioglu, Yenan Zhu, Naomi C Brownstein, Katherine R Sterba, Ashish A Deshmukh, Kalyani Sonawane. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Depression and Anxiety in Adolescents During the COVID-19 Pandemic in Relation to the Use of Digital Technologies: Longitudinal Cohort Study

Chen Shen[1,2], PhD; Rachel B Smith[1,2,3], PhD; Joel Heller[1,2], MSc; Alexander D V Spiers[1,2], PhD; Rhiannon Thompson[1,2,4], PhD; Helen Ward[5], PhD; Jonathan P Roiser[6], PhD; Dasha Nicholls[7], MD; Mireille B Toledano[1,2,3], PhD

[1]MRC Centre for Environment and Health, School of Public Health, Imperial College London, London, United Kingdom

[2]National Institute for Health Research Health Protection Research Unit in Chemical and Radiation Threats and Hazards, Imperial College London, London, United Kingdom

[3]Mohn Centre for Children's Health and Wellbeing, School of Public Health, Imperial College London, London, United Kingdom

[4]National Institute for Health Research School for Public Health Research, Imperial College London, London, United Kingdom

[5]Patient Experience Research Centre, School of Public Health, Imperial College London, London, United Kingdom

[6]Institute of Cognitive Neuroscience, University College London, London, United Kingdom

[7]Department of Brain Sciences, Division of Psychiatry, Imperial College London, London, United Kingdom

Corresponding Author:
Mireille B Toledano, PhD
MRC Centre for Environment and Health
School of Public Health
Imperial College London
School of Medicine Building
Norfolk Place
London, W21PG
United Kingdom
Phone: 44 02075943298
Email: m.toledano@imperial.ac.uk

Abstract

Background: Adolescents are susceptible to mental illness and have experienced substantial disruption owing to the COVID-19 pandemic. The digital environment is increasingly important in the context of a pandemic when in-person social connection is restricted.

Objective: This study aims to estimate whether depression and anxiety had worsened compared with the prepandemic period and examine potential associations with sociodemographic characteristics and behavioral factors, particularly digital behaviors.

Methods: We analyzed cross-sectional and longitudinal data from a large, representative Greater London adolescent cohort study: the Study of Cognition, Adolescents and Mobile Phones (SCAMP). Participants completed surveys at T1 between November 2016 and July 2018 (N=4978; aged 13 to 15 years) and at T2 between July 2020 and June 2021 (N=1328; aged 16 to 18 years). Depression and anxiety were measured using the Patient Health Questionnaire and Generalized Anxiety Disorder scale, respectively. Information on the duration of total mobile phone use, social network site use, and video gaming was also collected using questionnaires. Multivariable logistic regression was used to assess the cross-sectional and longitudinal associations of sociodemographic characteristics, digital technology use, and sleep duration with clinically significant depression and anxiety.

Results: The proportion of adolescents who had clinical depression and anxiety significantly increased at T2 (depression: 140/421, 33.3%; anxiety: 125/425, 29.4%) compared with the proportion of adolescents at T1 (depression: 57/421, 13.5%; anxiety: 58/425, 13.6%; *P* for 2-proportion *z* test <.001 for both depression and anxiety). Depression and anxiety levels were similar between the summer holiday, school opening, and school closures. Female participants had higher odds of new incident depression (odds ratio [OR] 2.5, 95% CI 1.5-4.18) and anxiety (OR 2.11, 95% CI 1.23-3.61) at T2. A high level of total mobile phone use at T1 was associated with developing depression at T2 (OR 1.89, 95% CI 1.02-3.49). Social network site use was associated with depression and anxiety cross-sectionally at T1 and T2 but did not appear to be associated with developing depression or anxiety longitudinally. Insufficient sleep at T1 was associated with developing depression at T2 (OR 2.26, 95% CI 1.31-3.91).

XSL•FO
RenderX

Conclusions: The mental health of this large sample of adolescents from London deteriorated during the pandemic without noticeable variations relating to public health measures. The deterioration was exacerbated in girls, those with preexisting high total mobile phone use, and those with preexisting disrupted sleep. Our findings suggest the necessity for allocating resources to address these modifiable factors and target high-risk groups.

(J Med Internet Res 2024;26:e45114) doi:10.2196/45114

KEYWORDS

COVID-19; depression; anxiety; public health; adolescence; digital technology use; sleep; mobile phone

Introduction

Background

The global COVID-19 pandemic was an extraordinary public health crisis, involving unprecedented public health measures such as social distancing and school and business closures. On March 23, 2020, the UK government announced a national lockdown and urged the public to stay at home to reduce the spread of COVID-19. Lockdown restrictions were relaxed during the summer (May to August 2020), allowing for some easing of social measures, such as permitting 2 households to meet while adhering to social distancing guidelines. The lockdown was reintroduced in winter 2020 to mitigate the increasing transmission and protect the health care system from being overwhelmed. However, such measures may put people at risk of depression, anxiety, stress, and helplessness [1,2].

Adolescence is a life stage when individuals are susceptible to the onset of mental illness, of which depression and anxiety are most common. Therefore, adolescents may have been more susceptible to the mental health impacts of the public health crisis than adults. A United Kingdom–based survey found that 80% of the respondents believed that the pandemic had worsened their mental health, and 67% of the respondents believed that the pandemic would have a long-term negative effect on their mental health [3]. Adolescence is a critical period when peer relationships become more influential than family relationships on their life [4]. Social isolation from peers and loneliness are related to depressive symptoms, suicidal ideation, and anxiety [5]. In addition, school closures were associated with fear, restlessness, and sadness [6]. However, whether the impacts of the pandemic on mental health were acutely related to public health measures (eg, mental health status improves as restrictions ease) or enduring (eg, persistently poor mental health status during the pandemic) remains unclear. Understanding these impacts remains a policy priority to protect and improve young people's mental health [7].

Use of mobile phones and other wireless devices is ubiquitous among children and adolescents. In the United Kingdom, 97% of the adolescents aged 12 to 15 years and 100% of the adolescents aged 16 to 17 years own a mobile phone, and most of them use a mobile phone, tablet, or laptop to go to the web [8]. Our recent systematic review found suggestive but limited evidence indicating associations between greater technology use and poorer mental health in children and adolescents [9]. More high-quality longitudinal studies with detailed information on digital technology use are needed, particularly in the context of the COVID-19 pandemic when the digital environment is increasingly important for learning, connection, and social support to offset the negative impacts of school closures and social isolation on mental health.

Previous cross-sectional studies found a high prevalence of depression and anxiety in young people during the pandemic [10-12]. Female gender, older age, and lower socioeconomic status (SES; eg, financial strain and living in rural areas) were associated with more mental health problems [13]. However, it is difficult to assess the impacts of the pandemic without information on the prepandemic symptom levels for comparison. In addition, other factors related to mental health including increased digital technology use (particularly the use of social network sites [SNSs]) and more sleep problems were reported during the pandemic [14,15]. Given that the associations between such behaviors and mental health are complex and often bidirectional, longitudinal analysis is necessary to unravel this interrelationship. A few longitudinal studies in adolescents have investigated depression and anxiety before and during the pandemic with mixed results [16-18]. One study reported an increase in depression and anxiety [16]. One study found an increase in depression, but anxiety remained stable both before and during the pandemic [17]. Another survey in school Year 9 students (the third year of secondary school; aged 13 years) found a decrease in anxiety during the UK lockdown [18]. However, older adolescents in critical examination years (General Certificate of Secondary Education [at the end of compulsory secondary education, aged approximately 16 years] and advanced level [for university admissions, aged approximately 18 years] examinations) may experience more stress owing to missing school time, as the potential impact on future university and employment opportunities is more immediate. However, none of these studies examined the longitudinal associations between digital technology use, sleep, and mental health.

Objectives

This study aims to investigate the impact of the COVID-19 pandemic on adolescents' mental health in a large representative (in terms of gender, ethnicity, and SES) longitudinal adolescent cohort across Greater London (the Study of Cognition, Adolescents and Mobile Phones [SCAMP]), for whom there was detailed information on mental and physical health and digital behaviors such as mobile phone use, SNS use, and video gaming, both before and during the pandemic. Thus, we were in a unique position to assess and tease out potential relationships between depression and anxiety, sociodemographic characteristics, digital technology use, sleep, and COVID-19 infection status and ascertain impacts in this important age group specifically related to the pandemic and public health measures.

XSL•FO
RenderX

This will help to identify the groups at greatest risk who may benefit from targeted support.

Methods

Participants

The SCAMP is a prospective adolescent cohort study that was originally set up to investigate cognitive and behavioral outcomes affected by the use of mobile phones and other wireless technologies that emit radio-frequency electromagnetic fields. The details of this study have been reported previously [19]. Between November 2014 and July 2016, baseline data were collected from 6581 participants in Year 7 (aged 11 to 12 years) from 39 secondary schools (26 state and 13 independent) in and around Greater London, the United Kingdom. Participants in all SCAMP schools completed a computer-based assessment using the Psytools software (Delosis Ltd) under examination-like conditions in school. The assessment included a questionnaire on their digital technology behaviors (eg, smartphone use, SNS engagement, and video gaming); a battery of cognitive tests; and physical and mental health, lifestyle, and behavior scales. Data collected from 4978 participants at 31 schools between November 2016 and July 2018 (T1) when they were in Year 9 or 10 (aged 13 to 15 years) were included in this study. Participants who had depression (n=3292) and anxiety measures (n=3350) were defined as the T1 cross-sectional sample.

All SCAMP participants were invited to complete another assessment between July 2020 and June 2021 (T2) when they were aged 16 to 18 years, comprising cognitive tests and questionnaires on mental health, digital technology behaviors, home and outdoor environment, lifestyle, and access to public health information. Participants completed the assessment at home from July to September 2020. The assessment from September 2020 onward was conducted in school, supervised by the SCAMP team onsite or remotely via internet depending on school policies. Data collected from 1328 adolescents at T2 were included in this study. Participants who had depression and anxiety measures (n=968) were defined as the T2 cross-sectional sample.

A subset of the participants had longitudinal depression (n=421) and anxiety (n=425) data, defined as the longitudinal sample. In the longitudinal sample, we excluded participants with clinically significant symptoms at T1 when assessing the longitudinal associations between the exposure variables and new incident depression and anxiety. The remaining participants were defined as the sample for incident case analysis (n=364 for incident depression analysis and n=367 for incident anxiety analysis). The sample for incident case analysis had 80% power to detect odds ratio (OR) of 1.82 and 1.84 when assessing the associations between digital technology use (eg, total mobile phone use) in tertile categories and incident depression and anxiety, respectively. Figure 1 shows the structure of the SCAMP cohort data relevant to this study.

Figure 1. Structure of the Study of Cognition, Adolescents and Mobile Phones (SCAMP) cohort data relevant to this study. GAD-7: 7-item Generalized Anxiety Disorder; PHQ-8: 8-item Patient Health Questionnaire; T1: November 2016 to July 2018; T2: July 2020 to June 2021.

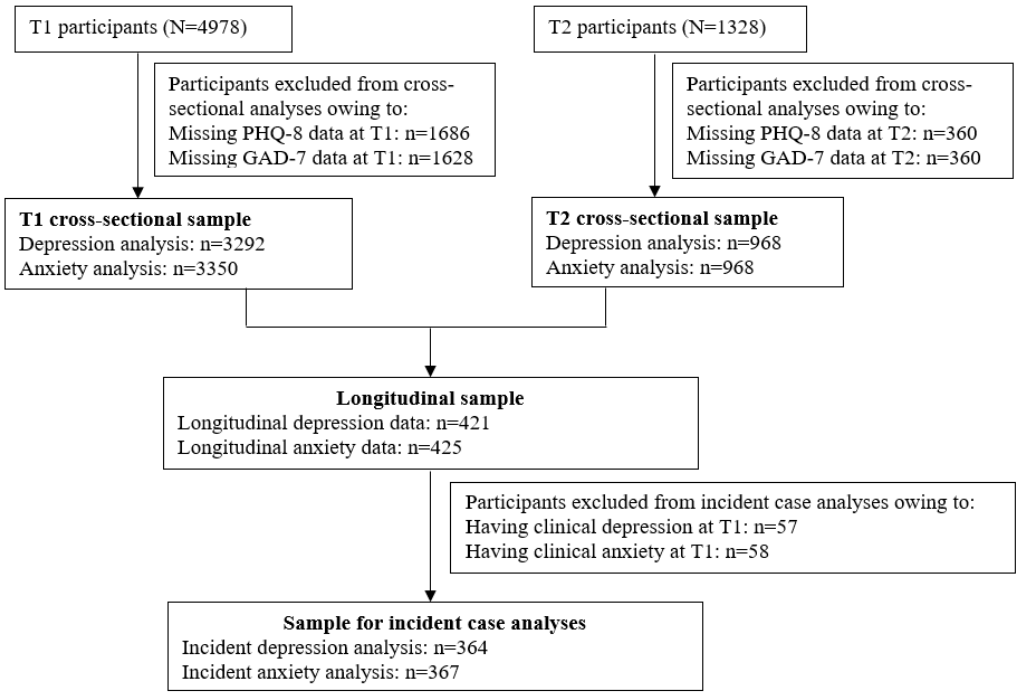

XSL•FO
RenderX

Measures (Self-Reported Questionnaire Data)

Sociodemographic Characteristics

Sociodemographic information including age, gender, ethnicity, and parental occupation was collected. The occupation of both parents was assessed at T1 only. We used the Office for National Statistics classification of occupation, categorizing it as managerial and professional, intermediate, and routine and manual [20]. The participants were assigned the higher parental occupation level. We also assessed the parental job situation since the COVID-19 lockdown at T2. We dichotomized the parental job situation as change during the pandemic and no change. School type (independent or state) was also included in the analyses. Missing data on ethnicity (262/4978, 5.26%), parental occupation (691/4978, 13.88%), and parental job situation (328/1328, 24.7%) were assigned a "missing" category for each variable rather than being excluded from the analyses.

Digital Technology Use (Total Mobile Phone Use, SNS Use, and Video Gaming)

At T1, participants reported the daily duration of phone calls and internet use (eg, web browsing, WhatsApp, Facebook, YouTube, and any other web-based apps) on mobile phones separately for weekdays and weekends. At T1, total mobile phone use was calculated as the combined daily duration of phone calls and internet use on mobile phones, with the average taken over both weekdays and weekends. At T2, participants reported an average daily duration of total mobile phone use for any purpose in the previous week.

At T1, participants reported the daily duration of SNS (eg, Facebook, Instagram, and Twitter) use on mobile phones and other devices separately for weekdays and weekends. At T1, SNS use was calculated as the combined daily duration of SNS use on mobile phones and other devices, with the average taken over both weekdays and weekends. At T2, participants reported the average time spent on specific SNS platforms (regardless of device) per day, including Facebook, Instagram, Twitter, TikTok, Snapchat, and YouTube in the last week. SNS use at T2 was calculated as the combined duration of SNS use on different platforms.

At T1, participants reported the daily duration of playing video games on any device at T1 separately for weekdays and weekends. Video gaming at T1 was the average weekday and weekend daily duration. At T2, participants reported an average daily duration of playing video games on any device in the last week.

Categorical responses were provided for questions on total mobile phone use, SNS use, and video gaming. To enable calculation and combination, we took the midpoints of category intervals, except in the highest category where we used the lowest value. For instance, the category "11-30 minutes per day" was converted to 20.5 minutes per day or 0.34 hours per day, whereas "more than 5 hours per day" was converted to 5 hours per day. Details for response category intervals for each question are shown in Table S1 in Multimedia Appendix 1. Participants were categorized into 3 groups based on the tertiles of daily duration of total mobile phone use, SNS use, and video gaming at each time point.

Sleep

At T1, participants provided information on when they usually got into bed, how long it took them to fall asleep, and what time they usually woke up separately for weekdays and weekends. The sleep duration was derived from these responses. The details have been reported elsewhere [21]. At T2, participants reported an average duration of sleep each night in the past 4 weeks. At T1 and T2, a sleep duration of <7 hours and >10 hours per day was defined as insufficient sleep and oversleep, respectively [22].

COVID-19 Infection Status

At T2, participants were asked if they had had a confirmed (by a positive test) or suspected COVID-19 infection (by a physician or themselves, but not tested).

Mental Health Outcomes—Depression and Anxiety

Depression was assessed using the 8-item Patient Health Questionnaire (PHQ-8). A summary score of ≥10 is considered indicative of clinically significant depression [23]. Cronbach α for the PHQ-8 score was .86 in the T1 cross-sectional sample and .88 in the T2 cross-sectional sample, indicating good internal reliability. To our knowledge, no previous study has examined the validity of the PHQ-8 in adolescents, but the PHQ-9 has demonstrated good validity in adolescents [24]. Anxiety was assessed using the 7-item Generalized Anxiety Disorder (GAD-7) scale. A summary score of ≥10 is considered indicative of clinically significant anxiety [25]. Cronbach α for the GAD-7 score was .89 in the T1 cross-sectional sample and .91 in the T2 cross-sectional sample, indicating good internal reliability. Evidence from current literature has indicated good validity and reliability of the GAD-7 in adolescents [26].

Statistical Analyses

We tested the significance of the difference between the T1 and T2 time-varying factors in participants with data at both time points. We performed multivariable logistic regression in T1 and T2 cross-sectional samples to assess the associations of sociodemographic characteristics, digital technology use, and sleep with clinically significant depression and anxiety at ages 13 to 15 years and 16 to 18 years. In the T2 cross-sectional sample, we also assessed whether mental health differed by public health measures (eg, lockdown, school closures, and school reopening), using time as a proxy. We categorized it as summer holiday (July to August 2020), school opening (September to December 2020 and March to June 2021), and school closures (January to February 2021). To assess whether the associations between COVID-19 infection and mental health were attributed to residual confounding bias, we used depression and anxiety at T1 as the negative control dependent variables. We expected no significant association between COVID-19 infection status at T2 and depression or anxiety at T1.

Multivariable logistic regression was performed to assess longitudinal associations of sociodemographic characteristics, digital technology use, and sleep with new incident depression and anxiety at T2 in the sample for incident case analysis. Multiple logistic regression was used to yield ORs and 95% CIs. All variables were mutually adjusted in the models to examine the associations between sociodemographic factors

XSL•FO
RenderX

and depression and anxiety. Depression and anxiety associated with total mobile phone use, SNS use, video gaming, sleep, and COVID-19 infection status were analyzed separately for each exposure variable (to avoid potential collider bias) by adjusting for all sociodemographic variables.

We conducted sensitivity analyses to examine cross-sectional associations between sociodemographic variables, digital technology use, sleep, COVID-19 infection status, and mental health outcomes at T1 and T2 in the longitudinal sample only. These analyses aimed to ascertain whether the associations observed in the longitudinal sample were substantially different from those observed in T1 and T2 cross-sectional samples. We performed an additional sensitivity analysis to investigate both cross-sectional and longitudinal associations between SNS use on mobile phones and depression and anxiety. This was done to determine if the patterns of association differed from those observed between SNS use on any device and mental health.

We did not use survey weights in our analyses, as the SCAMP is a school-based study that did not use probability sampling in participant recruitment. All analyses in this study were complete case analyses. Multiple imputation is not appropriate in our study, as it is probable that our data are missing not at random (ie, missingness is related to unobserved data). Participation in the T2 assessment might have been self-selected, given that data collection was mostly undertaken in a remote setting [27]. A *P* value of <.05 was considered statistically significant. All analyses were performed using R (version 4.0.3; R Foundation for Statistical Computing) and STATA (version 16.0; StataCorp).

Ethical Considerations

The North-West Haydock Research Ethics Committee approved the SCAMP study protocol and its subsequent amendments (#14/NW/0347). School head teachers consented to participate in the SCAMP. Participants received written information about the study in advance and were given the option to withdraw from the research at any time. This study was conducted in accordance with the Declaration of Helsinki.

Results

Descriptive Statistics of the SCAMP Cohort

Table 1 shows that our sample was diverse in terms of sociodemographic characteristics. Table S2 in Multimedia Appendix 1 shows that the level of total mobile phone and SNS use significantly increased at T2 (*P* values for paired *t* test <.001). The prevalence of insufficient sleep also significantly increased at T2 (*P* value for 2-proportion *z* test <.001). The proportion of depression and anxiety significantly increased from 13.5% (57/421) at T1 to 33.3% (140/421) at T2 and from 13.6% (58/425) at T1 to 29.4% (125/425) at T2 (both *P* values for 2-proportion *z* test <.001), respectively.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics, digital technology use, sleep, COVID-19 infection status, and depression and anxiety in the SCAMP cohort at T1[a] and T2[b].

Characteristics	T1 (n=4978)	T2 (n=1328)
Age (y), median (IQR)	14.02 (13.11-14.06)	17.06 (17.01-17.1)
Gender, n (%)		
Male	2273 (45.66)	548 (41.27)
Female	2705 (54.34)	777 (58.51)
Missing	0 (0)	3 (0.23)
Ethnicity, n (%)		
Asian	1319 (26.5)	434 (32.68)
Black	728 (14.62)	128 (9.64)
White	2153 (43.25)	600 (45.18)
Other	516 (10.37)	166 (12.5)
Missing	262 (5.26)	0 (0)
Parental occupation[c], n (%)		
Managerial and professional	2645 (53.13)	477 (35.92)
Intermediate	991 (19.91)	158 (11.9)
Routine or manual	651 (13.08)	82 (6.17)
Missing	691 (13.88)	611 (46.01)
School type, n (%)		
Independent	1287 (25.85)	459 (34.56)
State	3691 (74.15)	869 (65.44)
Parental job situation, n (%)		
No change since the lockdown	N/A[d]	786 (59.19)
Change since the lockdown	N/A	214 (16.11)
Missing	N/A	328 (24.7)
Time of data collection, n (%)		
Summer holiday	N/A	722 (54.37)
School opening	N/A	451 (33.96)
School closures	N/A	155 (11.67)
Total mobile phone use (hours), mean (SD)	2.52 (2.13)	4.6 (2.43)
SNS[e] use on any device (hours), mean (SD)	2.27 (2.69)	5.06 (3.82)
Video gaming on any device (hours), mean (SD)	1.12 (1.43)	1.15 (1.59)
Sleep, n (%)		
Normal	2990 (60.06)	366 (27.56)
Insufficient	1469 (29.51)	505 (38.03)
Oversleep	265 (5.32)	27 (2.03)
Missing	254 (5.1)	430 (32.38)
COVID-19 infection status, n (%)		
No	N/A	807 (60.77)
Suspected infection	N/A	231 (17.39)
Confirmed diagnosis	N/A	37 (2.79)
Missing	N/A	253 (19.05)

XSL•FO
RenderX

Characteristics	T1 (n=4978)	T2 (n=1328)
Prevalence of depression, % (n/N)	15.31 (504/3292)	30.89 (299/968)
Prevalence of anxiety, % (n/N)	13.22 (443/3350)	25.93 (251/968)

[a]T1: November 2016 to July 2018.

[b]T2: July 2020 to June 2021.

[c]Parental occupation was asked at T1 only; we assumed that parental occupation remained the same at T2.

[d]N/A: not applicable.

[e]SNS: social network site.

Results From Cross-Sectional Analyses

Tables 2 and 3 show analyses in T1 and T2 cross-sectional samples. Table 2 shows that compared with male participants, female participants had higher odds of presenting depression (T1: OR 2.64, 95% CI 2.15-3.25; T2: OR 2.59, 95% CI 1.9-3.52) and anxiety (T1: OR 3.3, 95% CI 2.62-4.15; T2: OR 2.55, 95% CI 1.83-3.54) at both time points. At T1, older participants and those from state schools had higher odds of depression and anxiety than younger participants and those from independent schools. There were no associations between mental health measures and ethnicity or parental occupation at any time point. Depression and anxiety at T2 did not differ significantly between the summer holiday, school opening, and school closures. In sensitivity cross-sectional analyses restricted to the longitudinal sample, although there was some variation in OR values, the cross-sectional associations with gender persisted. However, the associations with age and school type at T1 were no longer statistically significant (Table S3 in Multimedia Appendix 1). It is worth noting that the smaller sample size in the longitudinal sample might have reduced the power to detect associations.

XSL•FO
RenderX

Table 2. The associations between sociodemographic factors[a] and the presence of depression and anxiety in T1[b] and T2[c] cross-sectional samples.

Outcome and exposure	T1, OR[d] (95% CI)	T2, OR (95% CI)
Depression		
Age (per 1-y increase)	1.23 (1.02-1.48)	0.85 (0.66-1.11)
Gender		
Male	Reference	Reference
Female	2.64 (2.15-3.25)	2.59 (1.9-3.52)
Ethnicity		
Asian	0.97 (0.76-1.24)	1.01 (0.71-1.44)
Black	0.91 (0.66-1.27)	0.90 (0.51-1.57)
White	Reference	Reference
Other	1.20 (0.87-1.65)	0.82 (0.5-1.32)
Parental occupation		
Managerial and professional	Reference	Reference
Intermediate	0.95 (0.74-1.22)	1.03 (0.64-1.67)
Routine or manual	0.98 (0.73-1.32)	0.88 (0.44-1.77)
Parental job situation		
No change since the lockdown	N/A[e]	Reference
Change since the lockdown	N/A	1.39 (0.97-1.98)
School type		
Independent	Reference	Reference
State	1.49 (1.15-1.94)	1.41 (0.98-2.03)
Time of data collection		
Summer holiday	N/A	Reference
School opening	N/A	0.74 (0.53-1.04)
School closures	N/A	1.24 (0.79-1.94)
Anxiety		
Age (per 1-y increase)	1.27 (1.04-1.55)	0.92 (0.7-1.2)
Gender		
Male	Reference	Reference
Female	3.3 (2.62-4.15)	2.55 (1.83-3.54)
Ethnicity		
Asian	0.79 (0.61-1.04)	0.93 (0.64-1.35)
Black	0.93 (0.66-1.31)	0.57 (0.30-1.07)
White	Reference	Reference
Other	1.34 (0.97-1.85)	0.83 (0.50-1.38)
Parental occupation		
Managerial and professional	Reference	Reference
Intermediate	0.84 (0.64-1.1)	0.90 (0.54-1.51)
Routine or manual	0.97 (0.71-1.33)	1.32 (0.66-2.62)
Parental job situation		
No change since the lockdown	N/A	Reference
Change since the lockdown	N/A	1.06 (0.72-1.56)

XSL•FO
RenderX

Outcome and exposure	T1, OR[d] (95% CI)	T2, OR (95% CI)
School type		
Independent	Reference	Reference
State	1.38 (1.05-1.81)	1.16 (0.79-1.68)
Time of data collection		
Summer holiday	N/A	Reference
School opening	N/A	0.90 (0.63-1.27)
School closures	N/A	0.74 (0.44-1.22)

[a]All sociodemographic variables were mutually adjusted at T1 and T2.

[b]T1: November 2016 to July 2018.

[c]T2: July 2020 to June 2021.

[d]OR: odds ratio.

[e]N/A: not applicable.

XSL•FO
RenderX

Table 3. The associations[a] between digital technology use, sleep, COVID-19 infection status, and the presence of depression and anxiety in T1[b] and T2[c] cross-sectional samples.

Outcome and exposure	T1, OR[d] (95% CI)	T2, OR (95% CI)
Depression		
Total mobile phone use[e]		
1st tertile	Reference	Reference
2nd tertile	1.12 (0.87-1.45)	1.26 (0.87-1.81)
3rd tertile	2.09 (1.64-2.65)	1.87 (1.3-2.69)
SNS[f] use on any device[g]		
1st tertile	Reference	Reference
2nd tertile	1.34 (1.04-1.74)	1.13 (0.78-1.64)
3rd tertile	1.97 (1.53-2.52)	1.63 (1.12-2.36)
Video gaming on any device[h]		
1st tertile	Reference	Reference
2nd tertile	1.45 (1.14-1.85)	1.12 (0.79-1.6)
3rd tertile	1.82 (1.37-2.43)	1.44 (0.91-2.29)
Sleep		
Normal	Reference	Reference
Insufficient	2.73 (2.22-3.34)	3.6 (2.52-5.14)
Oversleep	1.6 (1.02-2.49)	10.79 (4.12-28.29)
COVID-19 infection status[i]		
No	Reference	Reference
Suspected infection	1.7 (0.9-3.21)	1.79 (1.28-2.5)
Confirmed diagnosis	0.61 (0.08-4.88)	0.59 (0.23-1.5)
Anxiety		
Total mobile phone use[j]		
1st tertile	Reference	Reference
2nd tertile	1.36 (1.03-1.79)	1.16 (0.79-1.71)
3rd tertile	2.33 (1.8-3.03)	1.57 (1.06-2.3)
SNS use on any device[k]		
1st tertile	Reference	Reference
2nd tertile	1.21 (0.92-1.57)	1.23 (0.83-1.83)
3rd tertile	1.57 (1.21-2.04)	1.91 (1.29-2.82)
Video gaming on any device[l]		
1st tertile	Reference	Reference
2nd tertile	1.31 (1.02-1.69)	1.09 (0.75-1.57)
3rd tertile	1.44 (1.06-1.96)	1.09 (0.66-1.78)
Sleep		
Normal	Reference	Reference
Insufficient	2.34 (1.89-2.9)	3.08 (2.12-4.46)
Oversleep	1.16 (0.7-1.92)	3.31 (1.32-8.3)
COVID-19 infection status		

XSL•FO
RenderX

Outcome and exposure	T1, OR[d] (95% CI)	T2, OR (95% CI)
No	Reference	Reference
Suspected infection	1.67 (0.88-3.17)	1.93 (1.36-2.74)
Confirmed diagnosis	1.44 (0.3-6.96)	1.27 (0.54-2.96)

[a]Mental health measures in relation to each exposure were analyzed separately by adjusting for the confounders as follows: T1 analysis: adjusted for age, gender, ethnicity, parental occupation, and school type; T2 analysis: adjusted for age, gender, ethnicity, parental occupation, parental job situation, school type, and time of data collection.

[b]T1: November 2016 to July 2018.

[c]T2: July 2020 to June 2021.

[d]OR: odds ratio.

[e]At T1, *P* for trend <.001; at T2, *P* for trend <.001.

[f]SNS: social network site.

[g]At T1, *P* for trend <.001; at T2, *P* for trend =.01.

[h]At T1, *P* for trend <.001; at T2, *P* for trend =.14.

[i]COVID-19 infection status reflects whether a participant has ever had a confirmed or suspected COVID-19 infection at any time before T2. Depression and anxiety at T1 were negative control outcomes when assessing the association between the COVID-19 infection status and mental health.

[j]At T1, *P* for trend <.001; at T2, *P* for trend =.03.

[k]At T1, *P* for trend <.001; at T2, *P* for trend =.001.

[l]At T1, *P* for trend =.01; at T2, *P* for trend =.69.

Table 3 shows that the total use of mobile phones and SNSs was associated with depression and anxiety at both time points with a dose-response relationship (all *P* values for trend <.05). Video gaming was associated with depression and anxiety at T1 with a dose-response relationship (both *P* values for trend <.05). Insufficient sleep was associated with depression (T1: OR 2.73, 95% CI 2.22-3.34; T2: OR 3.6, 95% CI 2.52-5.14) and anxiety (T1: OR 2.34, 95% CI 1.89-2.9; T2: OR 3.08, 95% CI 2.12-4.46) at both time points. Oversleep was associated with depression at both time points, with a stronger association at T2 (T1: OR 1.6, 95% CI 1.02-2.49; T2: OR 10.79, 95% CI 4.12-28.29). Suspected COVID-19 infection was associated with depression and anxiety at T2. This was not the case for confirmed COVID-19 infection; however, the low numbers of participants in this category (n<40) limit the sensitivity of this analysis. However, no significant associations between suspected or confirmed COVID-19 infection and depression or anxiety were observed at T1, as expected. In general, these cross-sectional associations were similar in the longitudinal sample, although some associations were no longer significant (Table S4 in Multimedia Appendix 1).

Results From Longitudinal Analyses

After excluding participants with preexisting depression or anxiety at T1, female participants had higher odds of new incident depression (OR 2.5, 95% CI 1.5-4.18) and anxiety (OR 2.11, 95% CI 1.23-3.61) at T2 than male participants (Table 4). Compared with White ethnicity, ethnic minority was not associated with greater new incident depression or anxiety at T2. Black ethnicity was associated with a lower incidence of anxiety at T2 (OR 0.33, 95% CI 0.09-1.22), although the association was marginally significant. Longitudinal associations between age and school type and the incidence of depression and anxiety were not marked.

XSL•FO
RenderX

Table 4. The longitudinal associations between sociodemographic factors[a] and the incidence of depression and anxiety at T2[b] in the sample for incident case analysis.

Exposure	Incidence of depression at T2[c], OR[d] (95% CI)	Incidence of anxiety at T2[e], OR (95% CI)
Age at T1[f] (per 1-y increase)	0.79 (0.47-1.33)	0.81 (0.47-1.40)
Age at T2 (per 1-y increase)	0.82 (0.51-1.32)	0.70 (0.42-1.15)
Gender		
Male	Reference	Reference
Female	2.5 (1.5-4.18)	2.11 (1.23-3.61)
Ethnicity		
Asian	0.76 (0.43-1.36)	0.89 (0.49-1.61)
Black	0.57 (0.2-1.64)	0.33 (0.09-1.22)
White	Reference	Reference
Other	0.51 (0.2-1.27)	0.62 (0.23-1.64)
Parental occupation		
Managerial and professional	Reference	Reference
Intermediate	0.99 (0.52-1.89)	0.83 (0.41-1.7)
Routine or manual	0.97 (0.41-2.29)	1.51 (0.64-3.59)
School type		
Independent	Reference	Reference
State	1.43 (0.78-2.65)	1.17 (0.61-2.22)

[a]All sociodemographic variables were mutually adjusted.

[b]T2: July 2020 to June 2021.

[c]Participants with clinically significant depression at T1 were excluded.

[d]OR: odds ratio.

[e]Participants with clinically significant anxiety at T1 were excluded.

[f]T1: November 2016 to July 2018.

Compared with the participants who used mobile phones in the lowest tertile, those in the highest tertile had higher odds of new incident depression at T2 (OR 1.89, 95% CI 1.02-3.49; Table 5). SNS use and video gaming at T1 were not associated with the development of depression or anxiety at T2. Moderate SNS users (ie, those in the 2nd tertile) had slightly lower odds of new incident depression at T2, although this association was not significant. The associations between SNS use on mobile phones and mental health outcomes were generally similar to those found between SNS use on any device and mental health outcomes (Table S5 in Multimedia Appendix 1). Insufficient sleep at T1 was also associated with new incident depression at T2 (OR 2.26, 95% CI 1.31-3.91).

XSL•FO
RenderX

Table 5. The longitudinal associations[a] between digital technology use and sleep at T1[b] and the incidence of depression and anxiety at T2[c] in the sample for incident case analysis.

Exposure at T1	Incidence of depression at T2[d], OR[e] (95% CI)	Incidence of anxiety at T2[f], OR (95% CI)
Total mobile phone use[g]		
1st tertile	Reference	Reference
2nd tertile	1.12 (0.64-1.97)	1.11 (0.62-2)
3rd tertile	1.89 (1.02-3.49)	1.48 (0.77-2.84)
SNS[h] use on any device[i]		
1st tertile	Reference	Reference
2nd tertile	0.72 (0.42-1.25)	0.80 (0.45-1.43)
3rd tertile	0.92 (0.48-1.75)	0.90 (0.47-1.75)
Video gaming on any device[j]		
1st tertile	Reference	Reference
2nd tertile	1.42 (0.79-2.57)	1.55 (0.82-2.91)
3rd tertile	1.32 (0.63-2.76)	1.8 (0.83-3.92)
Sleep		
Normal	Reference	Reference
Insufficient	2.26 (1.31-3.91)	1.14 (0.64-2.05)
Oversleep	1.41 (0.45-4.45)	0.66 (0.17-2.49)

[a]Adjusted for age at T1 and T2, gender, ethnicity, parental occupation, and school type at T1.
[b]T1: November 2016 to July 2018.
[c]T2: July 2020 to June 2021.
[d]Participants with clinically significant depression at T1 were excluded.
[e]OR: odds ratio.
[f]Participants with clinically significant anxiety at T1 were excluded.
[g]For depression, *P* for trend =.05; for anxiety, *P* for trend =.26.
[h]SNS: social network site.
[i]For depression, *P* for trend =.62; for anxiety, *P* for trend =.68.
[j]For depression, *P* for trend =.37; for anxiety, *P* for trend =.11.

Discussion

Principal Findings

To our knowledge, this is the first study to investigate detailed digital technology use in relation to longitudinal changes in mental health from the prepandemic period in a large representative adolescent sample with high sociodemographic diversity. We are also the first to consider mental health associated with public health measures (eg, lockdown, school closures, and school reopening) during the pandemic. We identified clear increases in depression and anxiety symptoms reaching clinical thresholds in adolescents over the course of the COVID-19 pandemic compared with the prepandemic assessments in the same cohort. Female individuals were more likely to develop depression and anxiety during the pandemic than male individuals. Black adolescents were less likely to develop anxiety during the pandemic than White adolescents. Depression and anxiety levels did not differ significantly according to socioeconomic factors or public health measures. High mobile phone use and insufficient sleep in the prepandemic period were associated with new incident depression during the pandemic. SNS use and video gaming were associated with depression and anxiety cross-sectionally only. Suspected COVID-19 infection was associated with increased depression and anxiety.

Comparison With Prior Work

Compared with previous longitudinal studies examining the impacts of the pandemic on adolescent mental health [16-18,28], our sample had greater ethnic and SES diversity, enabling us to examine the associations between ethnicity and SES and mental health. In addition, we investigated mental health related to various aspects of digital technology use and sleep both cross-sectionally and longitudinally to indicate potential risk factors. There were no marked variations in depression and anxiety based on public health measures, suggesting that mental health impacts are not contingent on immediate context but rather arise from broader and potentially more lasting societal and individual influences.

Our findings indicate a marked increase in the prevalence of depression and anxiety during the pandemic. This could be

XSL•FO
RenderX

attributed to factors such as enforced stay-at-home measures and school closures, which hindered peer interactions for school-aged adolescents. Such interactions are crucial for maintaining good mental health [29]. However, our results should be interpreted with caution, as the most recent data collection wave before the pandemic commenced when participants were aged 13 to 15 years (2016 to 2018) and still at a developmental stage when mental health problems begin to manifest. Our findings were comparable with another survey in a nationally representative sample, which also found that the rates of a probable mental health disorder increased from 1 in 10 in 2017 to 1 in 6 in 2021 among adolescents aged 17 to 19 years [28]. Moreover, we found that older age was not associated with new incident depression or anxiety during the pandemic (when participants were aged 16 to 18 years), suggesting the impacts of pandemic may have a more substantial influence on mental health decline than age itself.

In line with the findings of other studies that indicate gender differences in the prevalence of depression and anxiety during adolescence [30,31], we found that female gender was associated with depression and anxiety at both time points. Cognitive vulnerabilities, biological mechanisms (eg, sex hormones), and psychosocial factors may explain these differences [32-34]. We found that compared with male participants, female participants had a higher incidence of depression and anxiety during the pandemic, which echoes a recent longitudinal study [16]. This is perhaps because girls are more prone to rely on social support from their peer networks to cope with stressful life events [35]. Physical isolation and restricted interaction with peers, teachers, and the community posit a negative effect on coping strategies commonly used by girls [36]. However, another large UK population–based adolescent cohort study before the pandemic showed earlier and steeper increases in depressive symptoms in girls than in boys over the secondary school years [37], indicating that both usual gender-specific mental health trajectories and the impacts of the pandemic may underlie our findings.

Our findings on mental health differences by ethnicity are inconsistent with the hypothesized mental health vulnerability of ethnic minority groups in the general population [38]. One possible reason is that ethnic minority groups are clustered in some schools; thus, the potential compounding effects of social marginalization are not seen. In addition, Black ethnicity has been associated with higher psychological resilience during the pandemic [39], possibly explaining the lower incidence of anxiety in Black participants. It is also possible that our measures of anxiety and depression have measurement variance with respect to ethnicity or gender, reflecting limitations in how the items capture symptoms across groups (eg, owing to language) rather than differences in symptoms themselves [40,41].

We found new incident depression associated with total mobile phone use but not with SNS use at T1. The same levels of total mobile phone use and SNS use were associated with depression cross-sectionally at T1 in the T1 cross-sectional sample and longitudinal sample, corroborating our findings that the lack of longitudinal association between SNS use and depression is unlikely to be attributed to reduced power. In addition, we found that moderate SNS use was associated with slightly lower odds of new incident depression, although the association was not significant. Social networks established before the pandemic and moderate use of technology to maintain social connection remotely may be an important means to mitigate social isolation during lockdown. Adolescence is a sensitive period when social connection is particularly vital for brain development and mental health [42]. We did not find an association between video gaming and new incident depression or anxiety at T2. Although video gaming is associated with more depressive symptoms and behavioral difficulties in adolescents, web-based video gaming has a social component, that is, players are connected to each other, which might also mitigate social isolation during lockdown [43,44]. However, high mobile phone use may also carry the risk of displacement of sleep [45], enhancing social comparison [46], driving perfectionism [47], and increasing potential exposure to cyberbullying including social exclusion [48]. In addition, the short-term increase in screen time during the pandemic or the maladaptive technology use driven by poor mental health may explain the cross-sectional associations because of the possible bidirectional relationship between digital technology use and mental health. Our study found that insufficient sleep was associated with depression both cross-sectionally and longitudinally, indicating that insufficient sleep may be both a risk factor and a symptom of depression. It is also possible that individuals with preexisting sleep problems are more susceptible to the impact of adverse life experiences on mental health. Although COVID-19 infection suspected by the participants was associated with depression and anxiety, this requires further investigation with COVID-19 test results to confirm the infection status and its effect on adolescent mental health, as our sample of confirmed positive tests was low.

Strengths and Limitations

Our study has strengths and limitations. Strengths include that this was a large-scale adolescent longitudinal study using an ethnically and socioeconomically diverse cohort and detailed data on mental health measures, lifestyle, and sociodemographic characteristics. The cohort is representative of school-aged adolescents across Greater London. The longitudinal study design with the same mental health measures before and during the pandemic enabled the investigation of changes in depression and anxiety symptoms and new incident depression and anxiety and clearly demonstrated the worsening of mental health over this public health crisis. The limitations of this study include the loss of follow-up from T1 to T2 owing to logistical difficulties in data collection during the pandemic, which reduced our power to undertake more detailed analyses in subgroups. However, the association patterns in the longitudinal sample were generally similar to those observed in T1 and T2 cross-sectional samples, indicating that selection bias owing to loss of follow-up is unlikely to be a major concern. The use of self-reported questionnaires to collect information on digital technology use, sleep, and COVID-19 infection status in our study was subject to recall bias. Questionnaires on digital technology use were different between T1 and T2, making direct comparisons for each question challenging. We only measured the duration of SNS use in our study. However, the mental

XSL•FO
RenderX

health impact of SNSs could stem from the content adolescents interact with, or the type of interactions developed on the web, rather than the duration of SNS use itself. Given that the COVID-19 pandemic was a universal exposure, it is difficult to attribute the increase in mental health symptoms solely to the pandemic.

Public Health Implications

This study has several public health implications. Worsened mental health and a high proportion of clinical depression and anxiety in adolescents during the pandemic require more mental health support beyond clinical services (eg, school-level and digitally delivered approaches) for young people, especially for girls. For instance, there may be some good options for girls to gain social support by connecting with their social networks via moderate use of SNSs to cope with stressful situations. However, schools, parents, and adolescents should be aware of the potential adverse effects of digital technology overuse on mental health and better manage digital activities. In addition, tackling sleep problems might mitigate the impacts of the pandemic on mental health in terms of prevention and early treatment. No significant impacts of contemporaneous public health measures on mental health were observed.

Future Work

Future work should follow the SCAMP participants for more time to investigate the long-term effects on mental health. Objective data on screen time and sleep are needed in the future to minimize social desirability bias and measurement error from self-reported information. Future research should incorporate qualitative methodology to explore the reasons behind the varying mental health experiences among young people—some experience deterioration whereas others display resilience. This approach can provide insights to support those currently in need and mitigate the impact of future crises.

Conclusions

Our findings show longitudinal increases in the prevalence of clinically significant depression and anxiety during the COVID-19 pandemic compared with those measured in the prepandemic period. However, variations in depression and anxiety relating to public health measures were not evident, suggesting that the impacts on adolescent mental health are enduring regardless of the immediate context. Girls, those with preexisting high total mobile phone use, and those with preexisting insufficient sleep had higher risks of developing new depression during the pandemic. Girls also had higher risks of developing new anxiety. Ethnic minority groups were not associated with higher levels of depression or anxiety. Our findings indicate that resources to target these modifiable factors and high-risk groups are needed. It is possible that these adverse effects on mental health will have a lasting impact. Therefore, we recommend continued surveillance of adolescents' mental health to aid adequate provision of mental health services to this susceptible group of the population.

Acknowledgments

The Study of Cognition, Adolescents and Mobile Phones (SCAMP) COVID-19 study is an independent research project funded by the Medical Research Council (MRC; MR/V028472/1). The SCAMP is originally commissioned and funded (March 2014 to December 2021) by the National Institute for Health Research (NIHR) Policy Research Program (Secondary School Cohort Study of Mobile Phone Use and Neurocognitive and Behavioural Outcomes/091/0212) via the Research Initiative on Health and Mobile Telecommunications—a partnership between public funders and the mobile phone industry. This study was partly supported by the MRC Centre for Environment and Health, which is currently funded by the MRC (MR/S019669/1, 2019-2024). This study was also supported by funds from the NIHR Health Protection Research Unit in Chemical and Radiation Threats and Hazards (NIHR200922), based at the Imperial College London, in partnership with the United Kingdom Health Security Agency (UKHSA). Infrastructure support for the Department of Epidemiology and Biostatistics, Imperial College London, was provided by the NIHR Imperial Biomedical Research Centre. MBT's chair, RBS's fellowship, and the work in this paper are supported in part by a donation from Marit Mohn to Imperial College London to support Population Child Health through the Mohn Centre for Children's Health and Wellbeing. DN was supported by the NIHR Applied Research Collaboration program for Northwest London and by the Imperial Biomedical Research Collaboration. The funders of the study had no role in the design or conduct of the study or in the reporting of the SCAMP study results. The views expressed in this paper are those of the authors and not necessarily those of the MRC, NIHR, or UKHSA. The authors would like to thank all the schools, parents, and pupils who participated in the SCAMP study.

Authors' Contributions

CS conceived the analysis plan, analyzed the data (with JH), interpreted the results, and wrote and revised the manuscript. RBS guided the data analysis and contributed to the interpretation of data for this study. JH, ADVS, and RT revised the manuscript. MBT, JPR, DN, and HW obtained the funding and revised the manuscript. MBT conceptualized the study, obtained funding, inputted the data analysis, and revised the manuscript. All authors have approved the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Details of digital technology use variables (Table S1), comparison of T1 (November 2016 to July 2018) and T2 (July 2020 to June 2021) time-varying factors (Table S2), and results from sensitivity analyses (Tables S3 to S5).
[DOCX File , 54 KB - jmir_v26i1e45114_app1.docx]

References

1. Pancani L, Marinucci M, Aureli N, Riva P. Forced social isolation and mental health: a study on 1,006 Italians under COVID-19 lockdown. Front Psychol 2021 May 21;12:663799 [FREE Full text] [doi: 10.3389/fpsyg.2021.663799] [Medline: 34093358]
2. Nkire N, Mrklas K, Hrabok M, Gusnowski A, Vuong W, Surood S, et al. COVID-19 pandemic: demographic predictors of self-isolation or self-quarantine and impact of isolation and quarantine on perceived stress, anxiety, and depression. Front Psychiatry 2021 Feb 1;12:553468 [FREE Full text] [doi: 10.3389/fpsyt.2021.553468] [Medline: 33597900]
3. The impact of Covid-19 on young people with mental health needs. YoungMinds. URL: https://www.youngminds.org.uk/about-us/reports-and-impact/coronavirus-impact-on-young-people-with-mental-health-needs [accessed 2024-01-23]
4. Kilford EJ, Garrett E, Blakemore SJ. The development of social cognition in adolescence: an integrated perspective. Neurosci Biobehav Rev 2016 Nov;70:106-120. [doi: 10.1016/j.neubiorev.2016.08.016] [Medline: 27545755]
5. Loades ME, Chatburn E, Higson-Sweeney N, Reynolds S, Shafran R, Brigden A, et al. Rapid systematic review: the impact of social isolation and loneliness on the mental health of children and adolescents in the context of COVID-19. J Am Acad Child Adolesc Psychiatry 2020 Nov;59(11):1218-39.e3 [FREE Full text] [doi: 10.1016/j.jaac.2020.05.009] [Medline: 32504808]
6. Esposito S, Giannitto N, Squarcia A, Neglia C, Argentiero A, Minichetti P, et al. Development of psychological problems among adolescents during school closures because of the COVID-19 lockdown phase in Italy: a cross-sectional survey. Front Pediatr 2020 Jan 22;8:628072 [FREE Full text] [doi: 10.3389/fped.2020.628072] [Medline: 33553079]
7. Bunn S, Lewis S. Children's mental health and the COVID-19 pandemic. UK Parliament. 2021 Sep 9. URL: https://post.parliament.uk/research-briefings/post-pn-0653/ [accessed 2024-01-23]
8. Children and parents: media use and attitudes report 2022. Ofcom. 2022 Mar 30. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0024/234609/childrens-media-use-and-attitudes-report-2022.pdf [accessed 2024-01-23]
9. Girela-Serrano BM, Spiers AD, Ruotong L, Gangadia S, Toledano MB, Di Simplicio M. Impact of mobile phones and wireless devices use on children and adolescents' mental health: a systematic review. Eur Child Adolesc Psychiatry 2022 Jun 16:1-31 (forthcoming) [FREE Full text] [doi: 10.1007/s00787-022-02012-8] [Medline: 35705765]
10. Al Omari O, Al Sabei S, Al Rawajfah O, Abu Sharour L, Aljohani K, Alomari K, et al. Prevalence and predictors of depression, anxiety, and stress among youth at the time of COVID-19: an online cross-sectional multicountry study. Depress Res Treat 2020 Oct 06;2020:8887727-8887729 [FREE Full text] [doi: 10.1155/2020/8887727] [Medline: 33062331]
11. Chen X, Qi H, Liu R, Feng Y, Li W, Xiang M, et al. Depression, anxiety and associated factors among Chinese adolescents during the COVID-19 outbreak: a comparison of two cross-sectional studies. Transl Psychiatry 2021 Mar 02;11(1):148 [FREE Full text] [doi: 10.1038/s41398-021-01271-4] [Medline: 33654058]
12. Islam MA, Barna SD, Raihan H, Khan MN, Hossain MT. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: a web-based cross-sectional survey. PLoS One 2020 Aug 26;15(8):e0238162 [FREE Full text] [doi: 10.1371/journal.pone.0238162] [Medline: 32845928]
13. Racine N, Cooke JE, Eirich R, Korczak DJ, McArthur B, Madigan S. Child and adolescent mental illness during COVID-19: a rapid review. Psychiatry Res 2020 Oct;292:113307 [FREE Full text] [doi: 10.1016/j.psychres.2020.113307] [Medline: 32707216]
14. Bruni O, Malorgio E, Doria M, Finotti E, Spruyt K, Melegari MG, et al. Changes in sleep patterns and disturbances in children and adolescents in Italy during the Covid-19 outbreak. Sleep Med 2022 Mar;91:166-174 [FREE Full text] [doi: 10.1016/j.sleep.2021.02.003] [Medline: 33618965]
15. Bailey E, Boland A, Bell I, Nicholas J, La Sala L, Robinson J. The mental health and social media use of young Australians during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Jan 19;19(3):1077 [FREE Full text] [doi: 10.3390/ijerph19031077] [Medline: 35162101]
16. Magson NR, Freeman JY, Rapee RM, Richardson CE, Oar EL, Fardouly J. Risk and protective factors for prospective changes in adolescent mental health during the COVID-19 pandemic. J Youth Adolesc 2021 Jan 27;50(1):44-57 [FREE Full text] [doi: 10.1007/s10964-020-01332-9] [Medline: 33108542]
17. Barendse ME, Flannery J, Cavanagh C, Aristizabal M, Becker SP, Berger E, et al. Longitudinal change in adolescent depression and anxiety symptoms from before to during the COVID-19 pandemic: an international collaborative of 12 samples. PsyArXiv Preprints. Preprint posted online February 3, 2021 [FREE Full text] [doi: 10.31234/osf.io/hn7us]
18. Widnall E, Winstone L, Mars B, Haworth C, Kidger J. Young people's mental health during the COVID-19 pandemic: initial findings from a secondary school survey study in South West England. National Institute for Health Research School for Public Health Research. 2020 Aug. URL: https://sphr.nihr.ac.uk/wp-content/uploads/2020/08/Young-Peoples-Mental-Health-during-the-COVID-19-Pandemic-Report-Final.pdf [accessed 2024-01-23]

XSL•FO
RenderX

19. Toledano MB, Mutz J, Röösli M, Thomas MS, Dumontheil I, Elliott P. Cohort profile: the study of cognition, adolescents and mobile phones (SCAMP). Int J Epidemiol 2019 Feb 01;48(1):25-6l [FREE Full text] [doi: 10.1093/ije/dyy192] [Medline: 30325429]
20. The National Statistics Socio-economic classification (NS-SEC). Office for National Statistics. URL: https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010 [accessed 2024-01-23]
21. Shen C, Mireku MO, Di Simplicio M, Dumontheil I, Thomas MS, Röösli M, et al. Bidirectional associations between sleep problems and behavioural difficulties and health-related quality of life in adolescents: evidence from the SCAMP longitudinal cohort study. JCPP Adv 2022 Sep 17;2(3):e12098 [FREE Full text] [doi: 10.1002/jcv2.12098] [Medline: 36941946]
22. Hirshkowitz M, Whiton K, Albert SM, Alessi C, Bruni O, DonCarlos L, et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health 2015 Mar;1(1):40-43. [doi: 10.1016/j.sleh.2014.12.010] [Medline: 29073412]
23. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
24. Richardson LP, McCauley E, Grossman DC, McCarty CA, Richards J, Russo JE, et al. Evaluation of the Patient Health Questionnaire-9 Item for detecting major depression among adolescents. Pediatrics 2010 Dec;126(6):1117-1123 [FREE Full text] [doi: 10.1542/peds.2010-0852] [Medline: 21041282]
25. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
26. Mossman SA, Luft MJ, Schroeder HK, Varney ST, Fleck DE, Barzman DH, et al. The Generalized Anxiety Disorder 7-item scale in adolescents with generalized anxiety disorder: signal detection and validation. Ann Clin Psychiatry 2017 Nov;29(4):227-34A [FREE Full text] [Medline: 29069107]
27. Bethlehem J. Selection bias in web surveys. Int Statistical Rev 2010 Aug 03;78(2):161-188. [doi: 10.1111/j.1751-5823.2010.00112.x]
28. Mental health of children and young people in England 2021 - wave 2 follow up to the 2017 survey. National Health Service Digital. 2021 Sep 30. URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2021-follow-up-to-the-2017-survey [accessed 2024-01-24]
29. La Greca AM, Harrison HM. Adolescent peer relations, friendships, and romantic relationships: do they predict social anxiety and depression? J Clin Child Adolesc Psychol 2005 Mar;34(1):49-61. [doi: 10.1207/s15374424jccp3401_5] [Medline: 15677280]
30. Lewinsohn PM, Gotlib IH, Lewinsohn M, Seeley JR, Allen NB. Gender differences in anxiety disorders and anxiety symptoms in adolescents. J Abnorm Psychol 1998 Feb;107(1):109-117. [doi: 10.1037/0021-843x.107.1.109]
31. Breslau J, Gilman SE, Stein BD, Ruder T, Gmelin T, Miller E. Sex differences in recent first-onset depression in an epidemiological sample of adolescents. Transl Psychiatry 2017 May 30;7(5):e1139 [FREE Full text] [doi: 10.1038/tp.2017.105] [Medline: 28556831]
32. Calvete E, Cardeñoso O. Gender differences in cognitive vulnerability to depression and behavior problems in adolescents. J Abnorm Child Psychol 2005 Apr;33(2):179-192. [doi: 10.1007/s10802-005-1826-y] [Medline: 15839496]
33. Altemus M, Sarvaiya N, Neill Epperson C. Sex differences in anxiety and depression clinical perspectives. Front Neuroendocrinol 2014 Aug;35(3):320-330 [FREE Full text] [doi: 10.1016/j.yfrne.2014.05.004] [Medline: 24887405]
34. Petersen AC, Sarigiani PA, Kennedy RE. Adolescent depression: why more girls? J Youth Adolescence 1991 Apr;20(2):247-271. [doi: 10.1007/bf01537611]
35. Eschenbeck H, Kohlmann CW, Lohaus A. Gender differences in coping strategies in children and adolescents. J Individ Differ 2007 Jan;28(1):18-26. [doi: 10.1027/1614-0001.28.1.18]
36. Ellis WE, Dumas TM, Forbes LM. Physically isolated but socially connected: psychological adjustment and stress among adolescents during the initial COVID-19 crisis. Can J Behav Sci 2020 Jul;52(3):177-187. [doi: 10.1037/cbs0000215]
37. Kwong AS, Manley D, Timpson NJ, Pearson RM, Heron J, Sallis H, et al. Identifying critical points of trajectories of depressive symptoms from childhood to young adulthood. J Youth Adolesc 2019 Apr 22;48(4):815-827 [FREE Full text] [doi: 10.1007/s10964-018-0976-5] [Medline: 30671716]
38. Smith K, Bhui K, Cipriani A. COVID-19, mental health and ethnic minorities. Evid Based Ment Health 2020 Aug 17;23(3):89-90 [FREE Full text] [doi: 10.1136/ebmental-2020-300174] [Medline: 32680834]
39. Riehm KE, Brenneke SG, Adams LB, Gilan D, Lieb K, Kunzler AM, et al. Association between psychological resilience and changes in mental distress during the COVID-19 pandemic. J Affect Disord 2021 Mar 01;282:381-385 [FREE Full text] [doi: 10.1016/j.jad.2020.12.071] [Medline: 33421866]
40. Borgogna NC, Brenner RE, McDermott RC. Sexuality and gender invariance of the PHQ-9 and GAD-7: implications for 16 identity groups. J Affect Disord 2021 Jan 01;278:122-130. [doi: 10.1016/j.jad.2020.09.069] [Medline: 32956961]
41. Parkerson HA, Thibodeau MA, Brandt CP, Zvolensky MJ, Asmundson GJ. Cultural-based biases of the GAD-7. J Anxiety Disord 2015 Apr;31:38-42. [doi: 10.1016/j.janxdis.2015.01.005] [Medline: 25725310]
42. Orben A, Tomova L, Blakemore SJ. The effects of social deprivation on adolescent development and mental health. Lancet Child Adolesc Health 2020 Aug;4(8):634-640 [FREE Full text] [doi: 10.1016/S2352-4642(20)30186-3] [Medline: 32540024]

XSL•FO
RenderX

43. Li X, Vanderloo LM, Keown-Stoneman CD, Cost KT, Charach A, Maguire JL, et al. Screen use and mental health symptoms in Canadian children and youth during the COVID-19 pandemic. JAMA Netw Open 2021 Dec 01;4(12):e2140875 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.40875] [Medline: 34962557]
44. Colder Carras M, Van Rooij AJ, Van de Mheen D, Musci R, Xue QL, Mendelson T. Video gaming in a hyperconnected world: a cross-sectional study of heavy gaming, problematic gaming symptoms, and online socializing in adolescents. Comput Human Behav 2017 Mar;68:472-479 [FREE Full text] [doi: 10.1016/j.chb.2016.11.060] [Medline: 28260834]
45. Cabré-Riera A, Torrent M, Donaire-Gonzalez D, Vrijheid M, Cardis E, Guxens M. Telecommunication devices use, screen time and sleep in adolescents. Environ Res 2019 Apr;171:341-347. [doi: 10.1016/j.envres.2018.10.036] [Medline: 30716511]
46. He D, Shen X, Liu QQ. The relationship between upward social comparison on SNSs and excessive smartphone use: a moderated mediation analysis. Child Youth Serv Rev 2020 Sep;116:105232. [doi: 10.1016/j.childyouth.2020.105232]
47. Long J, Liu TQ, Liao YH, Qi C, He HY, Chen SB, et al. Prevalence and correlates of problematic smartphone use in a large random sample of Chinese undergraduates. BMC Psychiatry 2016 Nov 17;16(1):408 [FREE Full text] [doi: 10.1186/s12888-016-1083-3] [Medline: 27855666]
48. Tsimtsiou Z, Haidich AB, Drontsos A, Dantsi F, Sekeri Z, Drosos E, et al. Pathological internet use, cyberbullying and mobile phone use in adolescence: a school-based study in Greece. Int J Adolesc Med Health 2017 Apr 22;30(6):22. [doi: 10.1515/ijamh-2016-0115] [Medline: 28432846]

Abbreviations

GAD-7: 7-item Generalized Anxiety Disorder
OR: odds ratio
PHQ-8: 8-item Patient Health Questionnaire
SCAMP: Study of Cognition, Adolescents and Mobile Phones
SES: socioeconomic status
SNS: social network site

Edited by A Mavragani; submitted 18.12.22; peer-reviewed by S Darrow, R Plackett; comments to author 03.11.23; revised version received 20.12.23; accepted 19.01.24; published 07.02.24.

Please cite as:
Shen C, Smith RB, Heller J, Spiers ADV, Thompson R, Ward H, Roiser JP, Nicholls D, Toledano MB
Depression and Anxiety in Adolescents During the COVID-19 Pandemic in Relation to the Use of Digital Technologies: Longitudinal Cohort Study
J Med Internet Res 2024;26:e45114
URL: https://www.jmir.org/2024/1/e45114
doi:10.2196/45114
PMID:38324379

©Chen Shen, Rachel B Smith, Joel Heller, Alexander D V Spiers, Rhiannon Thompson, Helen Ward, Jonathan P Roiser, Dasha Nicholls, Mireille B Toledano. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Based Prediction of Suicidality in Adolescents With Allergic Rhinitis: Derivation and Validation in 2 Independent Nationwide Cohorts

Hojae Lee[1,2*], MSc; Joong Ki Cho[3*], MD; Jaeyu Park[1,2*], MSc; Hyeri Lee[1,2*], MSc; Guillaume Fond[4], MD, PhD; Laurent Boyer[4], MD, PhD; Hyeon Jin Kim[1,2], MSc; Seoyoung Park[5], BS; Wonyoung Cho[1], PhD; Hayeon Lee[2,5], PhD; Jinseok Lee[5,6], PhD; Dong Keon Yon[1,2,7], MD, PhD

[1]Department of Regulatory Science, Kyung Hee University, Seoul, Republic of Korea
[2]Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[3]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY, United States
[4]Assistance Publique-Hôpitaux de Marseille, Research Centre on Health Services and Quality of Life, Aix Marseille University, Marseille, France
[5]Department of Biomedical Engineering, Kyung Hee University, Yongin, Republic of Korea
[6]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin, Republic of Korea
[7]Department of Pediatrics, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Dong Keon Yon, MD, PhD
Department of Regulatory Science, Kyung Hee University
23 Kyungheedae-ro, Dongdaemun-gu
Seoul, 02447
Republic of Korea
Phone: 82 2 6935 2476
Fax: 82 504 478 0201
Email: yonkkang@gmail.com

Abstract

Background: Given the additional risk of suicide-related behaviors in adolescents with allergic rhinitis (AR), it is important to use the growing field of machine learning (ML) to evaluate this risk.

Objective: This study aims to evaluate the validity and usefulness of an ML model for predicting suicide risk in patients with AR.

Methods: We used data from 2 independent survey studies, Korea Youth Risk Behavior Web-based Survey (KYRBS; n=299,468) for the original data set and Korea National Health and Nutrition Examination Survey (KNHANES; n=833) for the external validation data set, to predict suicide risks of AR in adolescents aged 13 to 18 years, with 3.45% (10,341/299,468) and 1.4% (12/833) of the patients attempting suicide in the KYRBS and KNHANES studies, respectively. The outcome of interest was the suicide attempt risks. We selected various ML-based models with hyperparameter tuning in the discovery and performed an area under the receiver operating characteristic curve (AUROC) analysis in the train, test, and external validation data.

Results: The study data set included 299,468 (KYRBS; original data set) and 833 (KNHANES; external validation data set) patients with AR recruited between 2005 and 2022. The best-performing ML model was the random forest model with a mean AUROC of 84.12% (95% CI 83.98%-84.27%) in the original data set. Applying this result to the external validation data set revealed the best performance among the models, with an AUROC of 89.87% (sensitivity 83.33%, specificity 82.58%, accuracy 82.59%, and balanced accuracy 82.96%). While looking at feature importance, the 5 most important features in predicting suicide attempts in adolescent patients with AR are depression, stress status, academic achievement, age, and alcohol consumption.

Conclusions: This study emphasizes the potential of ML models in predicting suicide risks in patients with AR, encouraging further application of these models in other conditions to enhance adolescent health and decrease suicide rates.

(J Med Internet Res 2024;26:e51473) doi:10.2196/51473

XSL•FO
RenderX

KEYWORDS

machine learning; allergic rhinitis; prediction; random forest; suicidality

Introduction

Background

Allergic rhinitis (AR) is a common atopic disorder that affects approximately 14% of the global population [1-3]. AR is called allergic rhinoconjunctivitis when the eyes are involved and is an inflammatory condition characterized by at least one of the following symptoms: nasal congestion, rhinorrhea, itching, and sneezing [4,5]. In addition, it has been found to affect quality of life measures such as sleep, physical and social functioning, and learning and memory [6]. Furthermore, AR has been found to be associated with depressive symptoms, including suicidal ideation and suicide attempts [7-13]. Suicide rates in adolescents continue to increase, and suicide is the leading cause of adolescent death in Korea [14]. Adolescents are by nature susceptible to mental health problems owing to the many transitions involved in this period of their life, including changes in school, living situations, pressures of fitting into peer groups, and building their own identity. This can invoke helplessness, insecurity, stress, and a loss of control, possibly accelerating suicide rates in adolescents, a leading cause of adolescent death in South Korea [14]. Owing to this substantially burden in adolescents, it is important to further understand and investigate potential methods of mitigating the risk that AR adds to an already pressing issue. There is a possibility that the suicide rate increased because of the decrease in the quality of life among adolescents with AR [15]. Therefore, we will predict the suicidal attempts among those with AR using machine learning (ML) models.

Suicide prediction is elusive and thus adds to the challenges of suicide prevention worldwide. No practical methods for anticipating individual suicides or stratifying individuals according to risk have been well established [16,17]; however, ML-based models is a potential method for more accurately identifying adolescents at risk of suicide. A systematic review on the prediction of self-injurious thoughts and behaviors with ML determined that despite its limited application, ML has made a significant advancement in suicide prediction [18]. Another review found that ML has the potential to improve suicide predictions compared with traditional suicide prediction models [18]. Such studies illustrate the rapidly growing potential of ML.

Objectives

Given the additional risk of suicide-related behaviors in adolescents with AR, it would be relevant to use the growing field of ML to better evaluate this risk. Using nationwide population data, this study aimed to develop an ML-based model to predict suicide attempts among patients with AR using 2 independent nationwide cohorts in South Korea. We expect that this ML model produced from these data will have a high balanced accuracy and area under the receiver operating characteristic curve (AUROC) and consequently assist in better understanding suicide risk in adolescents with AR.

Methods

Study Design and Participants

This study aimed to develop an ML model to predict suicidality in Korean adolescents aged 13 to 18 years using clinical features extracted from 2 large independent data sets: the Korea Youth Risk Behavior Web-based Survey (KYRBS) and the Korea National Health and Nutrition Examination Survey (KNHANES) [19,20]. Figure 1 shows the workflow diagrams of the KYRBS and KNHANES data sets, which both offer nationally representative samples and estimates of the total adolescent population in South Korea. The original sample size for KYRBS was 1,067,169. However, after excluding patients without totaling 767,701, the final study population in the KYRBS data set was reduced to 299,468. Similarly, the KNHANES data set initially comprised 152,791 participants. However, after excluding 140,724 individuals either aged <13 or >19 years, 869 individuals with missing values on school performance, and 10,365 individuals without AR, the final study population in the KNHANES data set was 833.

XSL•FO
RenderX

Figure 1. Study workflow. AR: allergic rhinitis; KNHANES: Korea National Health and Nutrition Examination Survey; KYRBS: Korea Youth Risk Behavior Web-based Survey.

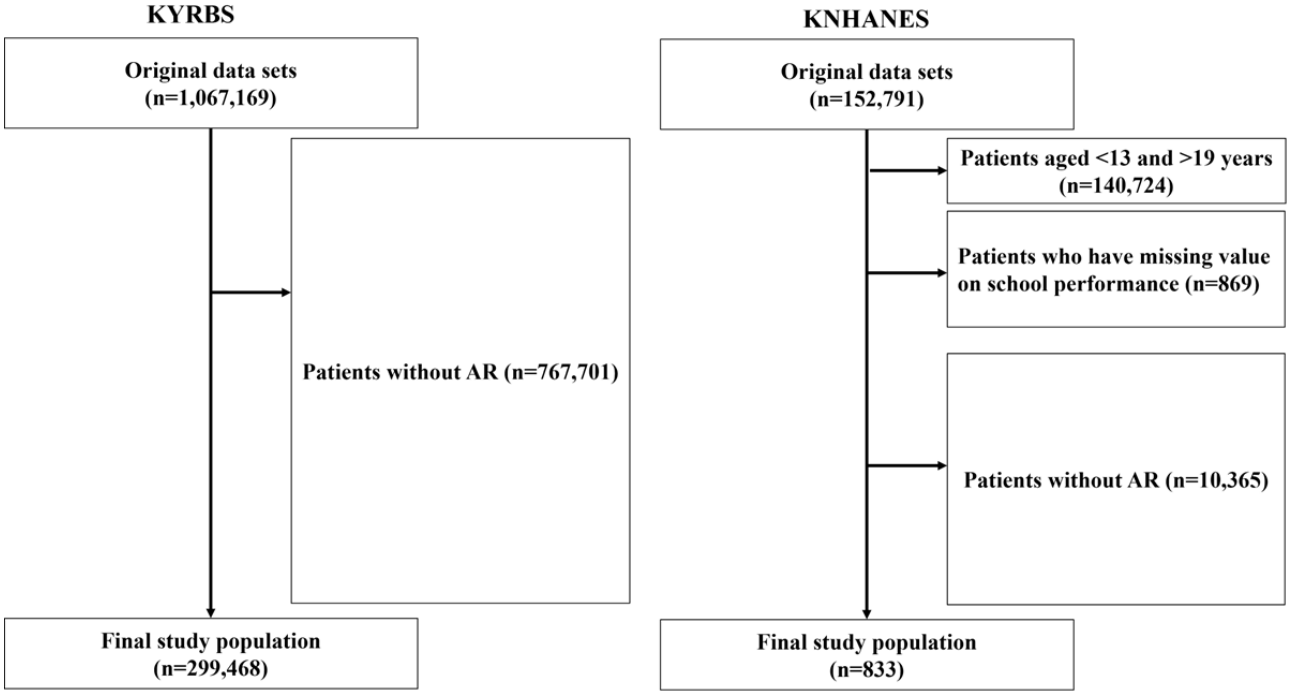

We included Korean adolescents aged between 13 and 18 years who completed the survey between 2005 and 2021 in KYRBS and between 1998 and 2021 in KNHANES. The outcome of suicidality was defined for people who attempted suicide more than once within 1 year [12], and covariates included age, sex, BMI (kg/m^2), residential area, household income, parents' level of education, academic achievement, smoking status, stress status, and feelings of sadness and despair.

We trained, validated, and externally tested the ML model's predictive accuracy and potential clinical efficacy in identifying the presence of mental health conditions using data from adolescents who met the same inclusion and exclusion criteria as those in the KYRBS data set. This study followed the guidelines outlined in the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement (Table S1 in Multimedia Appendix 1).

Ethical Considerations

The study protocol was approved by the institutional review board of the Korean Centers for Disease Control and Prevention Agency and Kyung Hee University (2022-06-042), and all participants provided written informed consent.

Variables and Algorithm Selection

To solve the imbalance issue of our data set, we used the synthetic minority oversampling technique to balance the training data set. The synthetic minority oversampling technique synthesizes new data from existing data using k-nearest neighbors and inserts them into the original data set [21]. The data set was randomly divided into 4:1 at a base training set (462,203/577,854, 79.98%) and a base test set (115,651/577,854, 20.01%) with equal distribution of different classes of patient data. This study aimed to develop a predictive model with a small number of variables and good performance; a model trained with the basic training set is needed to compare with a model with fewer variables than the basic training set. In addition, a corresponding test set was required to evaluate each training set, including the basic training set. Continuous variables were compared using the 2-tailed *t* test or Mann-Whitney *U* test, and categorical variables were compared using the chi-square test [22]. The odds ratios of the variables were determined by logistic regression (method: enter). Data set variables were analyzed using SAS software (version 9.3; SAS Institute Inc).

ML Model

In this study, as depicted in Figure 2, we analyzed the original KYRBS data set. The data set was divided into training and test data sets using a 4:1 ratio, with the training data set being used for model development and the test data set being used for model evaluation. We applied various ML algorithms to the training data set and assessed their performance based on the AUROC scores on the test data set. Models that exhibited high performance were selected for further investigation.

XSL•FO
RenderX

Figure 2. Model architecture. The original Korea Youth Risk Behavior Web-based Survey (KYRBS) data set was partitioned into training and test sets in a 4:1 ratio. The training set was used for model development, with performance assessed using area under the receiver operating characteristic curve (AUROC) scores on the test set. Selected high-performing models were further validated using an external Korea National Health and Nutrition Examination Survey (KNHANES) data set. The train data results were derived from the training data set, the test results were derived from the test data set, and the external results were derived from the additional KNHANES data set. CV: cross validation.

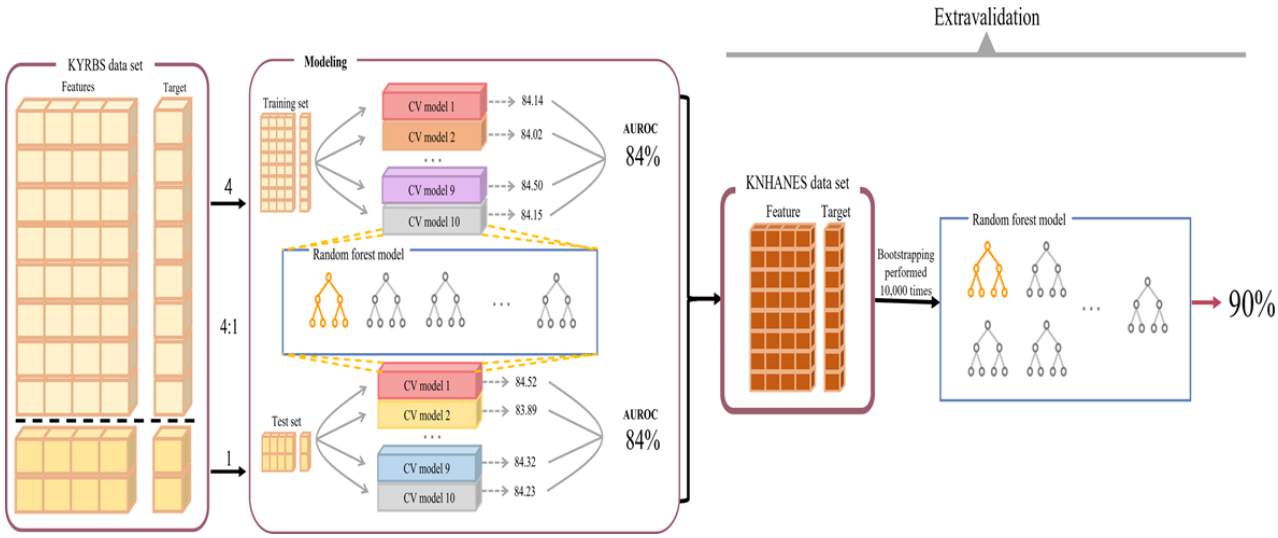

For external validation, we used an additional data set from the KNHANES, which contained the same columns as the original KYRBS data set. To enhance our understanding of the model's performance and its variability, we applied bootstrapping techniques. Bootstrapping was repeated 10,000 times to evaluate the model's performance on the external data set [23]. This process involved creating numerous resamples from the data set, with each sample being used to calculate the model's performance metrics. We calculated the mean and SE of these performance metrics across all the bootstrap samples. This technical approach provided a robust measure of the model's performance, accounting for variability and uncertainty in the external data set. The performance of the selected models on this external data set was evaluated, and their performance metrics were compared with those obtained from the training and test data sets. In summary, the train results were derived from the training data set, the test results were derived from the test data set, and the external results were derived from the external validation data set provided by KNHANES.

As shown in Figure 1, in the data preprocessing phase of this study, we took several steps to clean and prepare the data for efficient analysis. We used SAS software for data processing, which included categorizing and handling the missing values from the data set from the survey. Moreover, to address variables present in KYRBS but absent in KNHANES, we filled the missing values in KNHANES with the median values from KYRBS. These steps included handling missing values, categorizing categorical variables, and scaling numerical features using SAS software. The preprocessing steps aimed to ensure that the data were in a suitable format for the subsequent application of various ML algorithms. It is crucial to preprocess the data effectively, as this can substantially impact the performance and generalizability of the models being developed. Moreover, we used a 10-fold cross-validation approach to assess the performance of the ML models more reliably. This method involves partitioning the original data set into 10 equal-sized subsets, with each subset being used as a test data set once, whereas the remaining subsets serve as the training data set. The process was repeated 10 times, and the performance metrics, such as the AUROC score, sensitivity, specificity, accuracy, balanced accuracy score, precision, and F_1-score, were averaged over these iterations.

To estimate the uncertainty and variability of our results, we calculated the 95% CIs for each performance metric, including the AUROC score, sensitivity, specificity, accuracy, balanced accuracy score, precision, and F_1-score during the 10-fold cross-validation process to the train and test data sets. The 95% CI provides a range of plausible values for the performance metrics and is a useful tool for determining the stability and generalizability of the models. Data processing was performed using SAS software, and ML analysis was performed using Python (version 3.9.16), TensorFlow-gpu (version 2.6.0), Keras (version 2.6.0), NumPy (version 1.23.5), *pandas* (version 1.5.3), *scikit-learn* (version 1.2.2), Matplotlib (version 3.7.1), and *shap* (version 0.42.1). The ML models that were used were tree-based models, which are random forest, XGBoost, AdaBoost, and light gradient boost. We used GridSearch to fine-tune the hyperparameters of the models with the objective of maximizing the AUROC scores. The model hyperparameters were tuned using GridSearch, which uses many combinations of different hyperparameters to obtain the best AUROC score. GridSearch is an exhaustive search method that systematically explores a range of hyperparameter combinations, evaluating the performance of each combination on the given data set. By selecting the optimal set of hyperparameters from multiple variables using GridSearch, we selected from a range of hyperparameters for the random forest model. Finally, we chose number estimators at 100, maximum depth at 6, and maximum features as *sqrt*. This approach aimed to improve the performance and generalizability of our models, ultimately leading to more accurate and reliable predictions.

Subsequently, we focused on a detailed analysis of feature impact within the model. Feature importance was assessed using

XSL•FO
RenderX

the mean decrease in impurity within the random forest model, indicating how each feature contributes to more uniform node splits. In addition, the Seaborn library was used for visualizing this importance, enhancing the interpretability of the results, and aiding in the identification of the most impactful features for the model's predictive accuracy. This approach provides effective feature selection and model optimization.

To interpret and gain insights into the model's predictions, we calculated the Shapley Additive Explanations (SHAP) values from the random forest model. SHAP is a popular model-agnostic, local explanation approach designed to explain any given classifier. Lundberg and Lee [24] proposed the SHAP value as a united approach to explain the output of any ML model. We used the force plot and waterfall plot of the random forest model. This visualizes the contribution of each feature to the model's prediction for a specific instance, showing how each feature pushes the model's output from the base value. In contrast, the waterfall plot provides a detailed, step-by-step breakdown of how each feature contributes to moving the model's output from the expected value to the actual prediction.

Software and Libraries

All computations, model training, and evaluations were executed using Python (version 3.9.16), TensorFlow-gpu (version 2.6.0), Keras (version 2.6.0), NumPy (version 1.23.5), *pandas* (version 1.5.3), *scikit-learn* (version 1.2.2), and Matplotlib (version 3.7.12) for ML tasks and data wrangling. Visualization was facilitated using Matplotlib (version 3.7.2), Seaborn (version 0.12.2), and *shap* (version 0.42.1).

Results

Demographic Characteristics

This study was conducted using nationwide population data from 2 independent cohorts in South Korea to develop and investigate an ML-based model for predicting suicide attempts in patients with AR. The demographic characteristics of the study population were as follows: both cohorts consisted of patients with AR, with the KYRBS cohort including 299,468 patients and the KNHANES cohort comprising 833 adolescents aged 13 to 18 years. Table 1 shows the baseline characteristics of the KYRBS and KNHANES. In the original KYRBS cohort, the sex distribution revealed that among 299,468 patients, 152,789 (51.02%) were male patients and 146,679 (48.98%) were female patients. In the extravalidated KNHANES cohort, comprising 833 patients, 492 (59.06%) were male patients and 341 (40.94%) were female patients. The patient samples in both cohorts encompassed diverse socioeconomic backgrounds, including varying levels of education, income, and occupation. Overall, the study included a total of 300,301 patients with AR from diverse demographic backgrounds, ensuring a representative sample for the development and evaluation of the ML model. By considering these demographic characteristics, this study aimed to provide valuable insights into the risk of suicide attempts among individuals with AR, with a particular focus on the adolescent population.

XSL•FO
RenderX

Table 1. Demographic characteristics of Korea Youth Risk Behavior Web-based Survey (KYRBS) data (2005-2022) and Korea National Health and Nutrition Examination Survey (KNHANES) data cohorts (1998-2022) among excluded and included participants.

Characteristics	KYRBS			KNHANES		
	Total	AR[a]	Non-AR	Total	AR	Non-AR
Total, n	1,067,169	299,468	767,701	6306	833	5473
Age (years), mean (SD)	15.41 (1.69)	15.55 (1.69)	15.35 (1.69)	15.26 (1.65)	15.12 (1.65)	15.29 (1.65)
Sex, n (%)						
Male	548,952 (51.44)	152,789 (51.02)	396,134 (51.60)	3338 (52.93)	492 (59.06)	2846 (52.00)
Female	518,217 (48.56)	146,679 (48.98)	371,567 (48.40)	2968 (47.07)	341 (40.94)	2627 (48.00)
Region, n (%)						
Urban	495,593 (46.44)	141,229 (47.16)	354,418 (46.17)	5172 (82.02)	716 (85.95)	4456 (81.42)
Rural	571,576 (53.56)	158,239 (52.84)	413,283 (53.84)	1134 (17.98)	117 (14.05)	1017 (18.58)
BMI (kg/m^2), n (%)[b]						
Underweight (<18.5)	265,832 (24.91)	69,716 (23.28)	196,148 (25.55)	1392 (22.07)	166 (19.93)	1226 (22.40)
Normal (18.5-23.0)	565,813 (53.02)	159,736 (53.34)	406,036 (52.89)	3199 (50.73)	396 (47.54)	2803 (51.22)
Overweight (23.0-25.0)	120,270 (11.27)	35,368 (11.81)	84,908 (11.06)	755 (11.97)	112 (13.45)	643 (11.75)
Obese (≥25.0)	115,254 (10.80)	34,678 (11.58)	80,609 (10.50)	960 (15.22)	159 (19.09)	801 (14.64)
Smoking status, n (%)						
Nonsmoker	850,534 (79.70)	244,576 (81.67)	605,870 (78.92)	5945 (94.28)	794 (95.32)	5151 (94.12)
Smoker	216,635 (20.30)	54,892 (18.33)	161,831 (21.08)	361 (5.72)	39 (4.68)	322 (5.88)
Alcohol consumption, n (%)						
Nondrinker	868,996 (81.43)	247,540 (82.66)	621,376 (80.94)	4965 (78.73)	678 (81.39)	4287 (78.33)
1-2 days	115,361 (10.81)	30,995 (10.35)	84,294 (10.98)	1009 (16.00)	126 (15.13)	883 (16.13)
3-5 days	36,284 (3.40)	9643 (3.22)	26,716 (3.48)	239 (3.79)	23 (2.76)	216 (3.95)
6-9 days	21,770 (2.04)	5301 (1.77)	16,506 (2.15)	57 (0.90)	6 (0.72)	51 (0.93)
≥10 days	24,758 (2.32)	5989 (2.00)	18,809 (2.45)	36 (0.57)	0 (0.00)	36 (0.66)
Parents' highest educational level, n (%)						
University graduate or higher	485,455 (45.49)	160,545 (53.61)	324,841 (42.32)	N/A[c]	N/A	N/A
High school graduate	283,440 (26.56)	67,919 (22.68)	215,467 (28.07)	N/A	N/A	N/A
Middle school graduate or below	19,209 (1.80)	2995 (1.00)	16,275 (2.12)	N/A	N/A	N/A
Unknown	279,065 (26.15)	68,009 (22.71)	211,118 (27.50)	N/A	N/A	N/A
Academic achievement (percentile), n (%)						
Low (0-19)	99,674 (9.34)	24,586 (8.21)	75,081 (9.78)	N/A	N/A	N/A
Lower middle (20-39)	248,117 (23.25)	64,535 (21.55)	183,511 (23.91)	N/A	N/A	N/A
Middle (40-59)	310,226 (29.07)	83,822 (27.99)	226,442 (29.50)	N/A	N/A	N/A
Upper middle (60-79)	276,397 (25.90)	83,162 (27.77)	193,307 (25.18)	N/A	N/A	N/A
High (80-100)	132,755 (12.44)	43,363 (14.48)	89,360 (11.64)	N/A	N/A	N/A
Household income, n (%)						
Low (0-39)	163,277 (15.30)	43,962 (14.68)	119,301 (15.54)	805 (12.77)	71 (8.52)	734 (13.41)
Middle (40-59)	523,767 (49.08)	143,745 (48.00)	380,012 (49.50)	1573 (24.94)	198 (23.77)	1375 (25.12)
Upper middle (60-79)	292,298 (27.39)	85,977 (28.71)	206,281 (26.87)	2034 (32.26)	282 (33.85)	1752 (32.01)
High (80-100)	87,828 (8.23)	25,784 (8.61)	62,107 (8.09)	1894 (30.03)	282 (33.85)	1612 (29.45)
Asthma, n (%)						

XSL•FO
RenderX

Characteristics	KYRBS			KNHANES		
	Total	AR[a]	Non-AR	Total	AR	Non-AR
No	990,013 (92.77)	255,596 (85.35)	73,4383 (95.66)	6095 (96.65)	763 (91.60)	5332 (97.42)
Yes	77,156 (7.23)	43,872 (14.65)	33,318 (4.34)	211 (3.35)	70 (8.40)	141 (2.58)
Dermatitis, n (%)						
No	856,083 (80.22)	203,758 (68.04)	652,316 (84.97)	5666 (89.85)	639 (76.71)	5027 (91.85)
Yes	211,086 (19.78)	95,710 (31.96)	115,385 (15.03)	640 (10.15)	194 (23.29)	446 (8.15)
Stress status, n (%)[d]						
Mild	189,743 (17.78)	47,166 (15.75)	142,562 (18.57)	956 (15.16)	105 (12.61)	851 (15.55)
Moderate	449,705 (42.14)	123,860 (41.36)	325,862 (42.44)	3629 (57.55)	458 (54.98)	3171 (57.94)
High	312,467 (29.28)	92,745 (30.97)	219,743 (28.62)	1465 (23.23)	237 (28.45)	1228 (22.44)
Severe	115,254 (10.80)	35,697 (11.92)	79,534 (10.36)	256 (4.06)	33 (3.96)	223 (4.07)
Sadness and despair in the past year, n (%)						
No	740,615 (69.40)	202,830 (67.73)	537,775 (70.05)	5861 (92.94)	764 (91.72)	5097 (93.13)
Yes	326,554 (30.60)	96,638 (32.27)	229,926 (29.95)	445 (7.06)	69 (8.28)	376 (6.87)
Suicidal thoughts in the past year, n (%)						
No	898,023 (84.15)	250,265 (83.57)	647,709 (84.37)	5509 (87.36)	788 (94.60)	4721 (86.26)
Yes	169,146 (15.85)	49,203 (16.43)	119,992 (15.63)	797 (12.64)	45 (5.40)	752 (13.74)
Suicide attempt in the past year, n (%)						
No	1,030,779 (96.59)	289,136 (96.55)	741,599 (96.60)	6241 (98.97)	821 (98.56)	5420 (99.03)
Yes	36,390 (3.41)	10,332 (3.45)	26,102 (3.40)	65 (1.03)	12 (1.44)	53 (0.97)

[a]AR: allergic rhinitis.

[b]According to Asia-Pacific guidelines, BMI is divided into 4 groups: underweight (<18.5 kg/m^2), normal (18.5-22.9 kg/m^2), overweight (23.0-24.9 kg/m^2), and obese (≥25.0 kg/m^2).

[c]N/A: not applicable.

[d]Stress was defined by receipt of mental health counseling owing to stress.

ML Model Results

As shown in Figure 3 and Figure S1 in Multimedia Appendix 1, upon conducting extensive model evaluations, it was found that the random forest model was the best model in predicting suicide attempts in patients with AR. The train data results revealed that the random forest model achieved a sensitivity of 76.83 (95% CI 76.31-77.35), a specificity of 75.62 (95% CI 75.04-76.20), an accuracy of 76.22 (95% CI 76.07-76.38), a balanced accuracy of 76.22 (95% CI 76.07-76.38), a precision of 75.91 (95% CI 75.57-76.25), an F_1-score of 76.37 (95% CI 76.19-76.54), and an AUROC of 84.12 (95% CI 83.98-84.27).

XSL•FO
RenderX

Figure 3. The area under the receiver operating characteristic curve (AUROC) of 4 different machine learning algorithms in the train, test, and external validation data set. KNHANES: Korea National Health and Nutrition Examination Survey; KYRBS: Korea Youth Risk Behavior Web-based Survey; ROC: receiver operating characteristic.

Data set		Model	AUROC	Sensitivity	Specificity	Accuracy	Balanced Accuracy
KYRBS	Train	**Random forest**	**84.12 (83.98, 84.27)**	**76.83 (76.31, 77.35)**	**75.62 (75.04, 76.20)**	**76.22 (76.07, 76.38)**	**76.22 (76.07, 76.38)**
		LightGBM	84.07 (83.91, 84.22)	77.05 (76.85, 77.24)	75.64 (75.31, 75.97)	76.34 (76.15, 76.54)	76.34 (76.15, 76.54)
		XGBoost	84.06 (83.89, 84.23)	78.87 (78.66, 79.09)	73.99 (73.62, 74.37)	76.43 (76.21, 76.66)	76.43 (76.21, 76.66)
		AdaBoost	84.42 (84.34, 84.50)	78.29 (78.11, 78.46)	75.02 (74.82, 75.22)	76.65 (76.56, 76.75)	76.65 (76.56, 76.75)
	Test	**Random forest**	**84.08 (83.98, 84.19)**	**79.48 (79.31, 79.65)**	**73.28 (73.05, 73.50)**	**76.38 (76.20, 76.55)**	**76.38 (76.20, 76.55)**
		LightGBM	84.02 (83.81, 84.23)	77.92 (77.53, 78.31)	74.73 (74.37, 75.09)	76.32 (76.06, 76.59)	76.32 (76.06, 76.59)
		XGBoost	84.08 (83.87, 84.28)	78.97 (78.63, 79.32)	74.00 (73.68, 74.32)	76.49 (76.26, 76.71)	76.49 (76.26, 76.71)
		AdaBoost	84.27 (84.06, 84.48)	78.14 (77.95, 78.33)	75.18 (74.78, 75.58)	76.66 (76.40, 76.92)	76.66 (76.40, 76.92)
KNHANES	Validation	**Random forest**	**89.84 (89.78, 89.90)**	**91.72 (91.55, 91.88)**	**77.36 (77.33, 77.39)**	**77.57 (77.54, 77.60)**	**84.54 (84.46, 84.62)**
		LightGBM	89.01 (88.95, 89.07)	75.10 (74.84, 75.35)	80.51 (80.48, 80.54)	80.43 (80.41, 80.46)	77.80 (77.68, 77.93)
		XGBoost	89.80 (89.74, 89.85)	91.77 (91.61, 91.93)	73.58 (73.55, 73.61)	73.84 (73.81, 73.87)	82.67 (82.59, 82.76)
		AdaBoost	89.12 (89.06, 89.18)	75.09 (74.84, 75.34)	82.09 (82.06, 82.11)	81.99 (81.96, 82.01)	78.59 (78.46, 78.72)

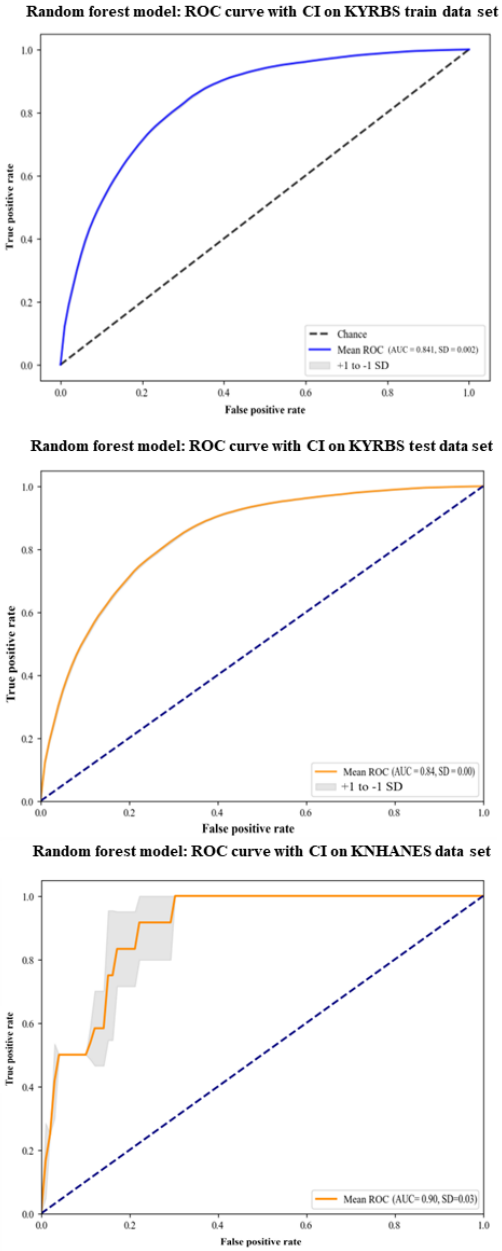

In contrast, the AdaBoost model yielded slightly different results with a sensitivity of 78.29 (95% CI 78.11-78.46), a specificity of 75.02 (95% CI 74.82-75.22), an accuracy of 76.65 (95% CI 76.56-76.75), a precision of 75.81 (95% CI 75.68-75.94), an F_1-score of 77.03 (95% CI 76.93-77.12), and an AUROC of 84.42 (95% CI 84.34-84.50).

However, when these models were evaluated on a separate test set, their performance varied. The random forest model obtained a sensitivity of 77.61 (95% CI 77.43-77.79), a specificity of 75.03 (95% CI 74.83-75.23), an accuracy of 76.32 (95% CI 76.16-76.49), a balanced accuracy of 76.32 (95% CI 76.16-76.49), a precision of 75.66 (95% CI 75.49-75.83), an F_1-score of 76.62 (95% CI 76.46-76.78), and an AUROC of 84.18 (95% CI 84.07-84.28). Conversely, the AdaBoost model showed a sensitivity of 78.14 (95% CI 77.95-78.33), a specificity of 75.18 (95% CI 74.78-75.58), an accuracy of 76.66 (95% CI 76.40-76.92), a balanced accuracy of 76.66 (95% CI 76.40-76.92), a precision of 75.89 (95% CI 75.57-76.21), an F_1-score of 77.00 (95% CI 76.77-77.23), and an AUROC of 84.27 (95% CI 84.06-84.48).

For external validation, an independent data set, KNHANES, was used. The random forest model achieved a sensitivity of 91.72 (95% CI 91.55-91.88), a specificity of 77.36 (95% CI 77.33-77.39), an accuracy of 77.57 (95% CI 77.54-77.60), a balanced accuracy of 84.54 (95% CI 84.46-84.62), and an AUROC of 89.84 (95% CI 89.78-89.90). Meanwhile, the AdaBoost model's external validation results revealed a sensitivity of 75.09 (95% CI 74.84-75.34), a specificity of 82.09 (95% CI 82.06-82.11), an accuracy of 81.99 (95% CI

XSL•FO
RenderX

81.96-82.01), a balanced accuracy of 78.59 (95% CI 78.46-78.72), and an AUROC of 89.12 (95% CI 89.06-89.18).

On the basis of the comprehensive results, the random forest model demonstrated superior performance compared with the AdaBoost model when evaluated on both internal and external data sets. In addition, the area under the precision-recall curve for the random forest model, a measure of model performance under conditions of class imbalance, was 81.98 (95% CI 79.88-84.08), as shown in Figure S1 in Multimedia Appendix 1. This indicates the model's robust ability to maintain precision across various levels of recall.

Feature Importance

Table 2 shows that the random forest model identified sadness and despair (53.1%) as the most influential feature in predicting suicide attempts in patients with AR, followed by stress status (28.35%), academic achievement (5.18%), age (4.08%), alcohol consumption (2.96%), household income (1.65%), sex (1.56%), smoking status (1.33%), BMI (kg/m^2; 0.69%), region (0.47%), parents' highest educational level (0.27%), atopic dermatitis (0.23%), and asthma (0.14%) in descending order of importance.

Table 2. Feature importance of the random forest model.

Feature	Importance, %
Sadness and despair	53.1
Stress status	28.35
Academic achievement	5.18
Age	4.08
Alcohol consumption	2.96
Household income	1.65
Sex	1.56
Smoking status	1.33
BMI (kg/m^2)[a]	0.69
Region	0.47
Parents' highest educational level	0.27
Atopic dermatitis	0.23
Asthma	0.14

[a]According to Asia-Pacific guidelines, BMI is divided into 4 groups: underweight (<18.5 kg/m^2), normal (18.5-22.9 kg/m^2), overweight (23.0-24.9 kg/m^2), and obese (≥25.0 kg/m^2).

SHAP Value

We addressed a deeper visual interpretation of the SHAP values within our ML model. Figure S2 in Multimedia Appendix 1 shows a waterfall plot, distinctively showcasing the cumulative contribution of each feature to a single prediction. We interpreted individual predictions by starting from the initial estimate and sequentially incorporating the influence of each feature to reach the final prediction. E[f(x)] refers to the average predicted output of the model across the entire data set, providing insights into the model's overall prediction tendency. The starting point of the illustration, denoted as E[f(X)]=0.50, represents the model's average prediction for the given data set. Among the variables, sadness and despair stood out, boosting the prediction by 0.16 and ranking as the most influential factor. Conversely, stress status, school performance, and sex reduced the prediction by 0.1, 0.02, and 0.01, respectively. This visualization offers a clear insight into the profound influence each feature wields in predicting adolescent suicidal thinking. Our ML model notably underscores substantial reliance on sadness and despair and stress status features. Moreover, in the force plot, features pushing the prediction higher are usually shown in one red color, whereas those pushing the prediction lower are shown in blue, clearly displaying the push and pull effect of each feature on the model's prediction. This type of visualization will allow us to see the balance of each effect at each individual prediction level, further clarifying the roles of sadness despair, stress status, and other features in assessing the risk of adolescent suicide attempts.

Code Availability

On the basis of the results of the ML model, we established a web-based application for policy makers or health system managers to support their decision-making process for cases involving suicidal attempts in adolescents with AR [25]. An example of a web interface and the results is shown in Figure S3 in Multimedia Appendix 1. Custom code for the website is available on the internet [26].

Discussion

Principal Findings

The study results showed that ML models can predict suicide attempts in patients with AR with relatively high accuracy. The random forest model is the best ML model to predict suicide attempts among Korean adolescents with AR, with an AUROC of 84.12% (original data set) and 89.87% (external validation data set). While looking at feature importance, the 5 most

XSL•FO
RenderX

important features in predicting suicide attempts in adolescent patients with AR are depression, stress status, academic achievement, age, and alcohol consumption.

To our knowledge, this is the first study to use an ML model in the context of patients with AR and suicidality, especially at this population level. These results reinforce the importance for clinicians to pay close attention to atopic conditions such as AR when screening for suicide risk as well as to help identify the most important risk factors in such patients.

Comparison With Previous Studies

In concordance with our findings, studies on suicide-related behavior prediction using ML have shown great promise. A study found that ML for suicide risk prediction in children and adolescents with electronic health records was able to detect 53% to 62% of suicide-positive participants with 90% specificity [27], and a case-control study of first-time suicide attempts with a cohort of >45,000 patients demonstrated accurate and robust first-time suicide attempt prediction [28], with the best predicting model achieving an AUROC of 0.932. A study that used the Korea Welfare Panel Study to develop an ML algorithm determined that >80% of individuals at risk of suicide-related behaviors could be predicted by various mental and socioeconomic characteristics of the respondents [29]. In addition, ML together with in-person screening has been found to result in the best suicide risk prediction [30], illustrating its potential to be used by clinicians in the medical field. These studies, as well as our study, support the continued need to build and improve ML models for predicting suicide risk, especially for at-risk patients.

As discussed, adolescents remain at a high risk for suicide-related behaviors because of their unique social situation. A study identified that significant risk factors for suicide in youth include a history of mental disorders, previous suicide attempts, impulsivity, family structure or environment, interpersonal strain, school problems and academic stress status, etc [31]. This is reflected in the findings of this study that showed adolescent's age, academic achievement, BMI group, and household income as some of the most important contributors to suicide attempt prediction in adolescents with AR. In addition, it has been observed that atopic dermatitis and asthma, although less common than the other risk factors, contribute to this already high burden of risk.

Plausible Mechanism

This study shows the importance of understanding what puts adolescents at risk for suicide, especially in the context of AR. There is a proposed pathogenic mechanism that connects atopy and its associated risk of increased suicidality. Allergic inflammatory mediators, interleukin (IL)-4, IL-5, and IL-13, are released and perpetuated by "allergic" helper T subtype 2 (T_H2) cells [32,33]. Along with atopic dermatitis and asthma, AR is associated with systemic increases in such cytokines [34]. Early life overexposure to IL-4, which can occur because of T_H2 sensitization from allergic disease, has been reported to reduce myelination and lead to cognitive impairment and developmental delays [35], and these effects have been found to be inhibited with IL-4 neutralization [36]. Allergy-mediated cytokines can also lead to aberrations in rapid eye movement (REM) sleep, increased REM latency, increased arousal, and decreased REM duration [37], thereby reducing sleep quality, quality of life, and overall happiness. In addition, T_H2 sensitization has been shown to possibly lead to negative effects on the developing brain, leading to increased attention-deficit/hyperactivity disorder, depression, anxiety, and suicidal ideation [38]. These mechanisms of action point toward a functional correlation between atopy and psychological disorders, including depression. This study helps to elucidate which other risk factors further contribute to such patients' already increased risk.

Strengths and Limitations

This study had several limitations. As the data sets only contained Korean adolescents, this model may not extend to the global adolescent population. South Korea's unique cultural and environmental setting may particularly affect the generalizability of the study. In the validation data set, participants aged <13 and >19 years were treated as missing data, and only patients with AR were analyzed, resulting in a low figure of 1%. This is because the KNHANES data set targets all ages, and not just adolescents; hence, it lacks the specificity for adolescents compared with the KYRBS data set. However, this data set represents South Korea and is used in studies as an external validation for the KYRBS data set [39-41]. In addition, the risk calculator that we produced was purely created for academic purposes, and its application should be limited to that scope. It is to be used as an example of what could be developed in the future with further refinement of ML models and our understanding of AR and suicide risk. The observed results are not intended to guide clinical management at this stage. As for the strengths of our study, to the best of our knowledge, this is the first study to create an ML model to predict suicide attempts in adolescents with AR. It is important to continue to investigate the role of atopic conditions in exacerbating suicide risk and to further understand how such patients may be affected by their disease process in the context of depression and suicide-related behaviors. Our model has the potential to make a significant impact on improving suicide risk assessment, early identification, and effective interventions for patients with AR, and it is important to further investigate the usefulness of ML and suicidality in patients with atopic diseases.

Clinical and Policy Implications

The findings of this study have various implications for clinicians and policy makers. It reveals the importance of screening adolescent patients with allergic diseases such as AR for suicidality. It can be assumed that as symptoms of atopy worsen, the risk for suicide also increases, and thus, it is important to encourage physicians to treat and pay close attention to their patients with allergies. These findings may also encourage psychiatrists to begin screening for allergies in their patients with depression or who are at risk of suicide attempts. Moreover, this ML-derived algorithm can be used by clinicians to independently screen their patients for the risk of suicide attempts. This can be done quickly and efficiently in the outpatient setting and can be incorporated as a tool to improve health outcomes for patients with atopic diseases. The

XSL•FO
RenderX

relatively high accuracy of this model encourages further research into developing similar ML models for other atopic diseases, including atopic dermatitis and asthma. For policy makers, these findings illustrate the importance of raising awareness of the contribution of allergic diseases, especially untreated ones, to increased rates of suicide. As part of this awareness, it is critical to highlight the potential benefits of interventions such as anti-inflammatory diets and fasting on mental health [42,43]. If the general population understands the risks, they can become more diligent in bringing adolescent patients to their clinicians to be screened and treated for their allergies. They will then be able to use an ML model such as this study to uniquely understand each patient's individualized suicide risk based on their various risk factors, including academic achievement and stress status. From this perspective, this ML algorithm holds great potential for improving the lives of adolescents affected by AR and its consequences.

Conclusions

ML models are a new and innovative field of study that show great potential in predicting suicide, a task that has proven difficult for clinicians thus far. This study confirms this potential and demonstrates its accuracy in the context of AR. Being able to identify patients with a specific risk factor and understand their unique risks for suicide is incredibly important and relevant for clinicians of all different specialties. This encourages the development of ML models not only in other atopic conditions such as asthma and atopic dermatitis but also in conditions outside of atopy. Further research should be conducted to investigate the utility of these ML models in the clinical field with the goal of decreasing suicide rates in the already vulnerable group that is adolescents. Although much research remains to be done, this new and exciting field of ML holds promise for improving the health of patients with atopic diseases.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant HV22C0233). The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

Data are available upon reasonable request. Study protocol and statistical code are available from DKY. The data set is available from the Korea Disease Control and Prevention Agency through a data use agreement.

Authors' Contributions

DKY had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All the authors approved the final version of the manuscript before submission. Hojae L, JKC, JP, Hyeri L, Hayeon L, JL, and DKY proposed the study concept and design; Hojae L, JKC, JP, Hyeri L, Hayeon L, JL, and DKY performed the acquisition, analysis, or interpretation of data; Hojae L, JKC, JP, Hyeri L, Hayeon L, JL, and DKY drafted the manuscript; all authors critically revised the manuscript for important intellectual content; Hojae L, JKC, JP, Hyeri L, Hayeon L, JL, and DKY did the statistical analysis; and Hayeon L, JL, and DKY supervised the study. DKY is the guarantor for this study. Hojae L, JKC, JP, and Hojae L contributed equally as joint first authors. Hayeon L, JL, and DKY contributed equally as corresponding authors. Hayeon L is a senior author. The corresponding author (Hayeon L, JL, and DKY) attests that all listed authors met the authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplement material and Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement.
[DOCX File , 465 KB - jmir_v26i1e51473_app1.docx]

References

1. Koo MJ, Kwon R, Lee SW, Choi YS, Shin YH, Rhee SY, et al. National trends in the prevalence of allergic diseases among Korean adolescents before and during COVID-19, 2009-2021: a serial analysis of the national representative study. Allergy 2023 Jun;78(6):1665-1670. [doi: 10.1111/all.15600] [Medline: 36440490]
2. Lee K, Lee H, Kwon R, Shin YH, Yeo SG, Lee YJ, et al. Global burden of vaccine-associated anaphylaxis and their related vaccines, 1967-2023: a comprehensive analysis of the international pharmacovigilance database. Allergy (Forthcoming) 2023 Dec 10. [doi: 10.1111/all.15968] [Medline: 38071735]
3. Shin YH, Hwang J, Kwon R, Lee SW, Kim MS, GBD 2019 Allergic Disorders Collaborators, et al. Global, regional, and national burden of allergic disorders and their risk factors in 204 countries and territories, from 1990 to 2019: a systematic

XSL•FO
RenderX

analysis for the Global Burden of Disease Study 2019. Allergy 2023 Aug;78(8):2232-2254. [doi: 10.1111/all.15807] [Medline: 37431853]
4. Brown T. Diagnosis and management of allergic rhinitis in children. Pediatr Ann 2019 Dec 01;48(12):e485-e488. [doi: 10.3928/19382359-20191111-01] [Medline: 31830288]
5. Yon DK, Hwang S, Lee SW, Jee HM, Sheen YH, Kim JH, et al. Indoor exposure and sensitization to formaldehyde among inner-city children with increased risk for asthma and rhinitis. Am J Respir Crit Care Med 2019 Aug 01;200(3):388-393. [doi: 10.1164/rccm.201810-1980LE] [Medline: 30958969]
6. Noh H, An J, Kim MJ, Sheen YH, Yoon J, Welsh B, et al. Sleep problems increase school accidents related to allergic diseases. Pediatr Allergy Immunol 2020 Jan;31(1):98-103. [doi: 10.1111/pai.13132] [Medline: 31578742]
7. Timonen M, Jokelainen J, Hakko H, Silvennoinen-Kassinen S, Meyer-Rochow VB, Herva A, et al. Atopy and depression: results from the Northern Finland 1966 Birth Cohort Study. Mol Psychiatry 2003 Aug;8(8):738-744. [doi: 10.1038/sj.mp.4001274] [Medline: 12888802]
8. Guzman A, Tonelli LH, Roberts D, Stiller JW, Jackson MA, Soriano JJ, et al. Mood-worsening with high-pollen-counts and seasonality: a preliminary report. J Affect Disord 2007 Aug;101(1-3):269-274 [FREE Full text] [doi: 10.1016/j.jad.2006.11.026] [Medline: 17222915]
9. Timonen M, Jokelainen J, Silvennoinen-Kassinen S, Herva A, Zitting P, Xu B, et al. Association between skin test diagnosed atopy and professionally diagnosed depression: a Northern Finland 1966 Birth Cohort study. Biol Psychiatry 2002 Aug 15;52(4):349-355. [doi: 10.1016/s0006-3223(01)01364-6] [Medline: 12208642]
10. Postolache TT, Stiller JW, Herrell R, Goldstein MA, Shreeram SS, Zebrak R, et al. Tree pollen peaks are associated with increased nonviolent suicide in women. Mol Psychiatry 2005 Mar;10(3):232-235 [FREE Full text] [doi: 10.1038/sj.mp.4001620] [Medline: 15599378]
11. Marshall PS, O'Hara C, Steinberg P. Effects of seasonal allergic rhinitis on fatigue levels and mood. Psychosom Med 2002;64(4):684-691. [doi: 10.1097/01.psy.0000021944.35402.44] [Medline: 12140359]
12. Woo HG, Park S, Yon H, Lee SW, Koyanagi A, Jacob L, et al. National trends in sadness, suicidality, and COVID-19 pandemic-related risk factors among South Korean adolescents from 2005 to 2021. JAMA Netw Open 2023 May 01;6(5):e2314838. [doi: 10.1001/jamanetworkopen.2023.14838] [Medline: 37223902]
13. Amritwar AU, Lowry CA, Brenner LA, Hoisington AJ, Hamilton R, Stiller JW, et al. Mental health in allergic rhinitis: depression and suicidal behavior. Curr Treat Options Allergy 2017 Mar;4(1):71-97 [FREE Full text] [doi: 10.1007/s40521-017-0110-z] [Medline: 28966902]
14. Kim N, Song JY, Yang H, Kim MJ, Lee K, Shin YH, et al. National trends in suicide-related behaviors among youths between 2005-2020, including COVID-19: a Korean representative survey of one million adolescents. Eur Rev Med Pharmacol Sci 2023 Feb;27(3):1192-1202 [FREE Full text] [doi: 10.26355/eurrev_202302_31226] [Medline: 36808368]
15. Lee KH, Yon DK, Suh DI. Prevalence of allergic diseases among Korean adolescents during the COVID-19 pandemic: comparison with pre-COVID-19 11-year trends. Eur Rev Med Pharmacol Sci 2022 Apr;26(7):2556-2568 [FREE Full text] [doi: 10.26355/eurrev_202204_28492] [Medline: 35442470]
16. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
17. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS One 2016 Jun;11(6):e0156322 [FREE Full text] [doi: 10.1371/journal.pone.0156322] [Medline: 27285387]
18. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. J Affect Disord 2019 Feb 15;245:869-884. [doi: 10.1016/j.jad.2018.11.073] [Medline: 30699872]
19. Park S, Yon H, Ban CY, Shin H, Eum S, Lee SW, et al. National trends in alcohol and substance use among adolescents from 2005 to 2021: a Korean serial cross-sectional study of one million adolescents. World J Pediatr 2023 Nov;19(11):1071-1081 [FREE Full text] [doi: 10.1007/s12519-023-00715-9] [Medline: 36977821]
20. Kweon S, Kim Y, Jang MJ, Kim Y, Kim K, Choi S, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). Int J Epidemiol 2014 Feb;43(1):69-77 [FREE Full text] [doi: 10.1093/ije/dyt228] [Medline: 24585853]
21. Ren Y, Wu D, Tong Y, López-DeFede A, Gareau S. Issue of data imbalance on low birthweight baby outcomes prediction and associated risk factors identification: establishment of benchmarking key machine learning models with data rebalancing strategies. J Med Internet Res 2023 May 31;25:e44081 [FREE Full text] [doi: 10.2196/44081] [Medline: 37256674]
22. Lee SW. Methods for testing statistical differences between groups in medical research: statistical standard and guideline of life cycle committee. Life Cycle 2022 Jan 24;2:e1 [FREE Full text] [doi: 10.54724/lc.2022.e1]
23. Koo JH, Park YH, Kang DR. Factors predicting older people's acceptance of a personalized health care service app and the effect of chronic disease: cross-sectional questionnaire study. JMIR Aging 2023 Jun 21;6:e41429 [FREE Full text] [doi: 10.2196/41429] [Medline: 37342076]

XSL•FO
RenderX

24. Lundberg SL, Lee S. A unified approach to interpreting model predictions. arXiv Preprint posted online May 22, 2017 [FREE Full text] [doi: 10.5860/choice.189890]
25. Streamlit. URL: https://predictsuicidality.streamlit.app [accessed 2024-01-26]
26. Machine learning-based prediction of suicidality in adolescents with allergic rhinitis. GitHub. URL: https://github.com/CenterForDH/suicidality [accessed 2024-01-29]
27. Su C, Aseltine R, Doshi R, Chen K, Rogers SC, Wang F. Machine learning for suicide risk prediction in children and adolescents with electronic health records. Transl Psychiatry 2020 Nov 26;10(1):413 [FREE Full text] [doi: 10.1038/s41398-020-01100-0] [Medline: 33243979]
28. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Jan;4(1):ooab011 [FREE Full text] [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
29. Lee J, Pak TY. Machine learning prediction of suicidal ideation, planning, and attempt among Korean adults: a population-based study. SSM Popul Health 2022 Sep;19:101231 [FREE Full text] [doi: 10.1016/j.ssmph.2022.101231] [Medline: 36263295]
30. Wilimitis D, Turer RW, Ripperger M, McCoy AB, Sperry SH, Fielstein EM, et al. Integration of face-to-face screening with real-time machine learning to predict risk of suicide among adults. JAMA Netw Open 2022 May 02;5(5):e2212095 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.12095] [Medline: 35560048]
31. Bilsen J. Suicide and youth: risk factors. Front Psychiatry 2018;9:540 [FREE Full text] [doi: 10.3389/fpsyt.2018.00540] [Medline: 30425663]
32. Akdis SH. Allergy and the immunologic basis of atopic disease. In: Kliegman RM, Stanton BF, St. Geme JW, Schor NF, Behrman RE, editors. Nelson Textbook of Pediatrics: Expert Consult. Philadelphia, PA: Elsevier Academic Press; Aug 01, 2019:388-393.
33. Mukundan TH. Screening for allergic disease in a child with sleep disorder and screening for sleep disturbance in allergic disease. In: Fishbein A, Sheldon SH, editors. Allergy and Sleep: Basic Principles and Clinical Practice. Cham, Switzerland: Springer; Dec 10, 2023:77-85.
34. Kim HJ, Kim YJ, Lee SH, Yu J, Jeong SK, Hong SJ. Effects of Lactobacillus rhamnosus on allergic march model by suppressing Th2, Th17, and TSLP responses via CD4(+)CD25(+)Foxp3(+) Tregs. Clin Immunol 2014 Jul;153(1):178-186. [doi: 10.1016/j.clim.2014.04.008] [Medline: 24769377]
35. Gadani SP, Cronk JC, Norris GT, Kipnis J. IL-4 in the brain: a cytokine to remember. J Immunol 2012 Nov 01;189(9):4213-4219 [FREE Full text] [doi: 10.4049/jimmunol.1202246] [Medline: 23087426]
36. Zanno AE, Romer MA, Fox L, Golden T, Jaeckle-Santos L, Simmons RA, et al. Reducing Th2 inflammation through neutralizing IL-4 antibody rescues myelination in IUGR rat brain. J Neurodev Disord 2019 Dec 16;11(1):34 [FREE Full text] [doi: 10.1186/s11689-019-9297-6] [Medline: 31839002]
37. Thompson A, Sardana N, Craig TJ. Sleep impairment and daytime sleepiness in patients with allergic rhinitis: the role of congestion and inflammation. Ann Allergy Asthma Immunol 2013 Dec;111(6):446-451. [doi: 10.1016/j.anai.2013.05.020] [Medline: 24267356]
38. Jackson-Cowan L, Cole EF, Arbiser JL, Silverberg JI, Lawley LP. TH2 sensitization in the skin-gut-brain axis: how early-life Th2-mediated inflammation may negatively perpetuate developmental and psychologic abnormalities. Pediatr Dermatol 2021 Sep;38(5):1032-1039. [doi: 10.1111/pde.14657] [Medline: 34338364]
39. Lee H, Park J, Lee M, Kim HJ, Kim M, Kwon R, et al. National trends in allergic rhinitis and chronic rhinosinusitis and COVID-19 pandemic-related factors in South Korea, from 1998 to 2021. Int Arch Allergy Immunol 2024 Jan 05:1-7. [doi: 10.1159/000535648] [Medline: 38185118]
40. Kwon R, Lee H, Kim MS, Lee J, Yon DK. Machine learning-based prediction of suicidality in adolescents during the COVID-19 pandemic (2020-2021): derivation and validation in two independent nationwide cohorts. Asian J Psychiatr 2023 Oct;88:103704. [doi: 10.1016/j.ajp.2023.103704] [Medline: 37541104]
41. Kang J, Park J, Lee H, Lee M, Kim S, Koyanagi A, et al. National trends in depression and suicide attempts and COVID-19 pandemic-related factors, 1998-2021: a nationwide study in South Korea. Asian J Psychiatr 2023 Oct;88:103727. [doi: 10.1016/j.ajp.2023.103727] [Medline: 37633158]
42. Berthelot E, Etchecopar-Etchart D, Thellier D, Lancon C, Boyer L, Fond G. Fasting interventions for stress, anxiety and depressive symptoms: a systematic review and meta-analysis. Nutrients 2021 Nov 05;13(11):3947 [FREE Full text] [doi: 10.3390/nu13113947] [Medline: 34836202]
43. Fond G, Young AH, Godin O, Messiaen M, Lançon C, Auquier P, et al. Improving diet for psychiatric patients : high potential benefits and evidence for safety. J Affect Disord 2020 Mar 15;265:567-569. [doi: 10.1016/j.jad.2019.11.092] [Medline: 31757621]

Abbreviations

AR: allergic rhinitis
AUROC: area under the receiver operating characteristic curve

XSL•FO
RenderX

IL: interleukin
KNHANES: Korea National Health and Nutrition Examination Survey
KYRBS: Korea Youth Risk Behavior Web-based Survey
ML: machine learning
REM: rapid eye movement
SHAP: Shapley Additive Explanations
TH2: helper T subtype 2

Edited by Q Jin; submitted 01.08.23; peer-reviewed by HG Woo, V Ruiz, B Montezano; comments to author 31.10.23; revised version received 24.12.23; accepted 16.01.24; published 14.02.24.

Please cite as:
Lee H, Cho JK, Park J, Lee H, Fond G, Boyer L, Kim HJ, Park S, Cho W, Lee H, Lee J, Yon DK
Machine Learning–Based Prediction of Suicidality in Adolescents With Allergic Rhinitis: Derivation and Validation in 2 Independent Nationwide Cohorts
J Med Internet Res 2024;26:e51473
URL: https://www.jmir.org/2024/1/e51473
doi:10.2196/51473
PMID:38354043

©Hojae Lee, Joong Ki Cho, Jaeyu Park, Hyeri Lee, Guillaume Fond, Laurent Boyer, Hyeon Jin Kim, Seoyoung Park, Wonyoung Cho, Hayeon Lee, Jinseok Lee, Dong Keon Yon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Status and Influencing Factors of Social Media Addiction in Chinese Workers: Cross-Sectional Survey Study

Weitao Kong[1,2,3,4], PhD; Yuanyuan Li[5], MS; Aijing Luo[1,2,3,4*], PhD; Wenzhao Xie[2,3,4,6*], PhD

[1]The Second Xiangya Hospital, Central South University, Changsha, China
[2]School of Life Sciences, Central South University, Changsha, China
[3]Key Laboratory of Medical Information Research, Central South University, College of Hunan Province, Changsha, China
[4]Clinical Research Center For Cardiovascular Intelligent Healthcare In Hunan Province, Changsha, China
[5]Department of Geratology, Hunan Provincial People's Hospital, The First Affiliated Hospital of Hunan Normal University, Changsha, China
[6]The Third Xiangya Hospital, Central South University, Changsha, China
[*]these authors contributed equally

Corresponding Author:
Aijing Luo, PhD
The Second Xiangya Hospital
Central South University
139 Renmin Middle Road
Changsha, 410000
China
Phone: 86 0731 85292212
Email: luoaj@csu.edu.cn

Abstract

Background: Social media addiction (SMA) caused by excessive dependence on social media is becoming a global problem. At present, most of the SMA studies recruit college students as research participants, with very few studies involving workers and other age groups, especially in China.

Objective: This study aims to investigate the current status of SMA among Chinese workers and analyze its influencing factors.

Methods: From November 1, 2022, to January 30, 2023, we conducted an anonymous web-based questionnaire survey in mainland China, and a total of 5176 participants completed the questionnaire. The questionnaire included the Social Networking Service Addiction Scale, Maslach Burnout Inventory–General Survey scale, Mindful Attention Awareness Scale, as well as questionnaires regarding participants' social media use habits and demographic information.

Results: Through strict screening, 3468 valid questionnaires were included in this study. The main findings of this study revealed the following: the average SMA score of workers was higher (mean 53.19, SD 12.04), and some of them (393/3468, 11.33%) relied heavily on social media; there were statistically significant differences in SMA scores among workers in different industries ($F_{14,3453}$=3.98; P<.001); single workers (t_{3106}=8.6; P<.001) and workers in a relationship (t_{2749}=5.67; P<.001) had higher SMA scores than married workers, but some married workers (214/3468, 6.17%) were highly dependent on social media; the level of SMA among female workers was higher than that of male workers (t_{3466}=3.65; P<.001), and the SMA score of workers negatively correlated with age (r=−0.22; P<.001) and positively correlated with education level (r=0.12; P<.001); the frequency of using social media for entertainment during work (r=0.33; P<.001) and the frequency of staying up late using social media (r=0.14; P<.001) were positively correlated with the level of SMA in workers; and the level of SMA in workers was significantly positively correlated with their level of burnout (r=0.35; P<.001), whereas it was significantly negatively correlated with their level of mindfulness (r=−0.55; P<.001).

Conclusions: The results of this study suggest that SMA among Chinese workers is relatively serious and that the SMA problem among workers requires more attention from society and academia. In particular, female workers, young workers, unmarried workers, highly educated workers, workers with bad social media habits, workers with high levels of job burnout, and workers with low levels of mindfulness were highly dependent on social media. In addition, occupation is an important influencing factor in SMA. Thus, the government should strengthen the supervision of social media companies. Medical institutions should provide health education on SMA and offer intervention programs for those addicted to social media. Workers should cultivate healthy habits while using social media.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48026) doi:10.2196/48026

KEYWORDS

social media addiction; job burnout; mindfulness; mobile phone; technology addiction; cross-sectional survey

Introduction

Background

In 2022, there were 4.62 billion active social media users worldwide, and each user was reported to spend an average of 2.5 hours a day on social media [1]. On the one hand, social media improves the efficiency of people's communication and work and provides new ways of entertainment. On the other hand, many scholars have pointed out that poor use of social media has become a new addiction [2]. Social media addiction (SMA) is a type of technology addiction that manifests as excessive attention to social media, strong motivation to use social media, and a large amount of energy expended toward the use of social media, which may affect physical health, mental health, interpersonal relationships, study and work efficiency, and happiness in life [3].

SMA is becoming a global problem, and research in many countries has found that some residents are overly dependent on social media [4]. From the perspective of age distribution, teenagers [5], middle-aged people [6], and some older adults [7] may have excessive dependence on social media. Previous studies have found that SMA has many adverse effects on people's physical and mental health. In terms of physical health, SMA is associated with poor eating habits of adolescents [8], making them addicted to food [9] or anorexia [10], which causes more adverse physiological reactions [11], and SMA can also reduce the quality of sleep in young people [12], which affects their executive function [13]. SMA has a negative impact on many aspects of mental health [14]. High levels of SMA can lead to higher levels of anxiety, depression, and stress [15]. People with high levels of SMA often feel anxious about their appearance [16], develop jealousy, and increase their willingness to undergo cosmetic surgery [17]. In addition, SMA can make students less engaged in learning, affect their academic performance [18], distract workers and lead to lower work efficiency, and make them face difficulties in career [19]. At the same time, SMA can have a negative impact on interpersonal relationships, which may lead to young romantic relationships becoming unstable [20]. People with high levels of SMA become introverted, prefer communication on the internet, ignore the importance of real social interaction, which affects their real interpersonal relationships [21].

Previous studies have identified many factors that influence SMA among young people and students. Among young people, psychological factors such as stress, narcissism, and nomophobia [22] as well as thoughts of the fear of missing out [23] have been found to increase their dependence on social media. Studies in the adolescent population have found that family life satisfaction [24], family economic status [25], negative emotions, and peer relationships [26] are predictors of SMA. In addition, some young people's motivations for using social media may lead to higher levels of SMA, such as escaping from reality and seeking a sense of belonging through social media or seeking social satisfaction through social media [27].

However, most previous studies were conducted among students [14]. Researchers have pointed out that SMA studies currently rely heavily on students as research participants [28]. However, little attention has been paid to the SMA of workers and middle-aged and older adults, and there are very few research data for reference. In China, in particular, there are few studies specifically on SMA among workers. Our study on physicians and nurses in China is one of the few SMA studies on Chinese workers. We found that physicians and nurses generally have high SMA levels [6]. In addition, previous research has found that some workers are addicted to the internet, which also indicates that they are at risk of technology addiction [29,30].

On the basis of the limitations of existing research, the main purpose of this study was to investigate the current status of SMA in workers (including middle-aged and older adults) and analyze the influencing factors to provide theoretical support for the prevention and intervention of SMA in future.

The Proposal of Research Hypothesis

Demographic Factors

Existing studies have shown that there are demographic factors, such as sex and age, that influence SMA levels; young people or women have higher levels of SMA than men and older people [31]; and there are differences in SMA levels based on marital status and income satisfaction [6]. Educational level is also a potential influencing factor [7]. Unreasonable social media use also increases the risk of SMA [32]. In existing studies, the definition of the unreasonable use of social media is not uniform [33]. We used the definition of SMA by Andreassen and Pallesen [3], which is recognized by many researchers in academia, that is, whether there will be adverse consequences, to judge whether the use behavior is reasonable. Considering the characteristics of workers' lives, their unreasonable use of social media is mainly reflected in their use of social media for entertainment during work and staying up late. On the basis of the abovementioned findings, we propose hypothesis 1 and hypothesis 2:

- H1: The SMA of Chinese workers is related to demographic factors.
- H2: The SMA of workers is positively correlated with the frequency of using social media for entertainment while working and the frequency of staying up late to use social media.

Burnout

Burnout is a prolonged response to chronic emotional and interpersonal stressors on the job [34]. It manifests as an imbalance between effort put into work and the results achieved [35] or as a response to work stress and work overload [36]. Burnout is a psychosomatic syndrome characterized by

XSL•FO
RenderX

emotional exhaustion, depersonalization, and reduced personal fulfillment [34]. Job burnout is common among Chinese workers, and it is a common psychological problem [37]. Our research on medical workers indicated that SMA is positively correlated with the degree of burnout [6], but there have been few studies on burnout and SMA among workers in other industries. However, it remains to be seen whether this relationship can be observed among the entire workforce. On the basis of the current evidence, we propose the following third hypothesis:

H3: The SMA level of Chinese workers is positively correlated with the level of job burnout.

Mindfulness

Derived from Eastern Buddhist meditation, mindfulness is defined as "awareness brought about by intentional, present-moment, nonjudgmental attention" [38]. Mindfulness can be studied as an intervention method, effect, or trait [39]. Meta-analysis studies have shown that mindfulness interventions are significant in improving physical and mental symptoms and are effective in relieving physical symptoms [40] as well as psychological symptoms [41]. Therefore, mindfulness training has been widely applied to combat addictive behaviors. At present, studies have confirmed a strong negative correlation between mindfulness and technology addiction. For example, a lack of mindfulness can aggravate smartphone addiction in adults, children, and adolescents [42,43], and people with low mindfulness are usually more prone to internet addiction [41]. At present, there are few research data on mindfulness and SMA, and the relationship between mindfulness and SMA in workers remains to be studied. On the basis of the correlation between mindfulness and addictive behaviors, we propose hypothesis 4:

H4: The level of SMA in Chinese workers is negatively correlated with the level of mindfulness.

Methods

Study Design and Procedure

We conducted an anonymous web-based questionnaire survey among Chinese workers to investigate their SMA status and influencing factors. The total questionnaire included the Social Networking Service Addiction Scale (SNSAS), the Maslach Burnout Inventory–General Survey (MBI-GS) scale, the Mindful Attention Awareness Scale (MAAS), a demographic data questionnaire, and a social media use habit questionnaire. We used WJX [44] to create the questionnaire as well as the link and QR code for the questionnaire. WJX is widely used in questionnaire research in China [6], and it is efficient and reliable.

Recruitment

We used random sampling methods to select 2 provinces each from eastern, central, and western China, namely Guangdong, Jiangsu, Hunan, Hubei, Sichuan, and Yunnan. Our research team searched the official websites of the 6 provincial trade unions, obtained their contact information, and contacted them for data collection support. After obtaining the consent of the 6 trade unions, the trade union staff distributed our questionnaire to various work-related group chats on WeChat in the region. They also encouraged participants who completed the questionnaire to further share it within their social media networks. We also publicly released the questionnaire on the official website of the Key Laboratory of Medical Information Research of Central South University, and any participant interested in participating in the survey could fill out the questionnaire through the link. Through the abovementioned efforts, our questionnaire has been widely used on Chinese social media. The formal questionnaire survey was completed from November 1, 2022, to January 30, 2023. On the basis of the principle of voluntary participation, a total of 5176 people ultimately completed the survey questionnaire. After screening, 3468 valid questionnaires were received.

In our questionnaire survey system, we had set all questions to be completed before submitting the questionnaire. Therefore, all submitted questionnaires were complete. The following methods, including specific inclusion and exclusion criteria as well as screening criteria for invalid questionnaires, were used to ensure the quality of the questionnaire in this study. The inclusion criteria were as follows: (1) participants must have a full-time job or be self-employed (participants were asked to fill in their own occupation names), (2) participants who choose "yes" on the question "Do you understand all the questionnaire questions," and (3) participants who choose "yes" on the question of "Are you an adult?" The exclusion criteria were as follows: (1) participants who filled in "student," "retired," or "unemployed" for the question "What is your occupation?" and (2) participants who answered "no" to the question "Do you work full-time?" The screening criteria for invalid questionnaires were as follows: (1) participants who completed the questionnaire in <300 seconds (lower than the normal answering time) and (2) participants who answered polygraph questions incorrectly.

Measures

SNSAS Scale

We measured participants' SMA levels using a scale developed by Chinese researchers [45] based on the Bergen Facebook Addiction Scale [46]. This scale has been successfully applied to workers in mainland China [6]. This scale has 6 dimensions: mood modification, salience, tolerance, withdrawal, conflict, and relapse. The SNSAS has 18 questions with a total score of 90 points, with higher scores indicating a more serious SMA. Each question is assessed on a 5-point Likert scale. Cronbach α coefficient of the scale in this study was 0.92.

MBI-GS Scale

The MBI-GS is a revised version of the MBI scale. It is a widely used tool to measure burnout among different working groups in China [47]. The Chinese version of the MBI-GS scale consists of 15 questions divided into 3 dimensions: emotional exhaustion with 5 questions, depersonalization with 4 questions, and personal accomplishment with 6 questions [48]. Each item was evaluated using a 7-point Likert scale. The higher the score, the higher the level of burnout. In this study, Cronbach α coefficient of the MBI-GS was 0.88.

XSL•FO
RenderX

MAAS Scale

This study used the Chinese version of the MAAS to measure participants' mindfulness levels [49]. MAAS is a common tool used to collect mindfulness data in China [49]. The Chinese version of the MAAS is a 1D scale consisting of 15 questions scored on a 6-point Likert scale, with a higher score indicating a higher level of mindfulness. Cronbach α coefficient of the MAAS in this study was 0.92.

Demographic Variables and Other Variables

We also identified the following 6 demographic variables that may have affected the data through literature analysis and expert consultation: work industry, sex, age, marital status, monthly income, and educational background. We measured respondents' bad social media habits using a single-item question, for example, "How many days a week do you stay up late using social media?" All the questions were scored on a 4 to 7 Likert scale.

Statistical Analysis

SPSS software (version 24.0; IBM Corp) was used for data analysis. The differences in SMA levels between participants of different sex and marital status were analyzed using a 2-tailed *t* test. ANOVA was used to examine the relationship between SMA scores and discontinuous variables such as marital status. Pearson correlation coefficient was used to measure the correlation between continuous variables such as SMA, burnout, and mindfulness, and linear regression analysis was used to test the effect of independent variables on SMA outcomes. Statistical significance was set at $P<.05$.

Ethical Considerations

All research procedures were approved by the Ethics Committee of the School of Life Sciences, Central South University (Number 2022-1-46). Before the survey, each participant was given an web-based informed consent form specifying the purpose of this study. The survey was conducted anonymously.

Results

Sociodemographic Characteristics and Social Media Use Habits of the Participants

Table 1 lists all social demographic characteristics and the corresponding means and SDs of the SMA scores. We used International Standard Industry Classification 4.0, and the occupations of all participants were manually categorized into 15 categories; the largest number of workers was in the education industry, which was dominated by teachers and educational ancillary occupations (992/3468, 28.6%). The lowest number of workers was in the transportation and storage industry, which mainly included workers in the passenger transport industry and logistics industry. The proportion of workers in the other 13 industries ranged from 2.02% (70/3468) to 10.78% (374/3468) of the total number industries that were inconvenient to classify, such as the funeral industry, were classified as "other service activities," accounting for 4.64% (161/3468) of the total number of workers. Of the 3468 workers, 2029 (58.51%) were women. The participants' age ranged from 18 to 69 years, with 31.6% (1096/3468) of the middle-aged workers aged 36 to 45 years. Moreover, 39.73% (1378/3468) of the workers were aged between 18 and 35 years, and workers aged >56 years accounted for only 1.99% (69/3468) of the total workforce sample. Overall, 68.95% (2391/3468) of the participants were married, 20.67% (717/3468) of the participants were single, and 10.38% (360/3468) of the participants were in a relationship. The monthly income of most participants (2736/3468, 78.89%) ranged from RMB 2000 to RMB 10,000 (a currency exchange rate of RMB ¥1=US $0.15 is applicable), with 2.62% (91/3468) of the participants in the highest income group (>RMB 20,000) and 3.2% (111/3468) of the participants in the lowest income group (<RMB 1000). Most of the participants (1527/3468, 44.03%) had a bachelor's degree. Table 2 lists the frequency of undesirable social media use by all participants. Overall, 6.55% (227/3468) of the workers used social media in inopportune situations such as work and meetings, 43.63% (1513/3468) of the workers used it whenever they were free, 42.47% (1473/3468) of the workers used it when they were bored, and only 7.35% (255/3468) rarely used it. Of the 3468 participants, 3374 (97.29%) reported staying up late using social media at least 1 day per week and 849 (24.48%) reported staying up late using social media every day. Only 2.71% (94/3468) of the participants said they never stayed up late using social media.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics and Social Networking Service Addiction Scale (SNSAS) score of the participants (N=3468).

Characteristics	Values, n (%)	SNSAS score, mean (SD)
SNSAS scores of all participants	3468 (100)	53.19 (12.04)
Occupation		
Agriculture, forestry, and fishing	289 (8.33)	50.43 (11.44)
Manufacturing	176 (5.07)	51.15 (11.74)
Construction	374 (10.78)	52.40 (11.60)
Wholesale and retail trade; repair of motor vehicles and motorcycles	263 (7.58)	52.41 (11.96)
Transportation and storage	52 (1.5)	51.27 (11.43)
Accommodation and food service activities	70 (2.02)	53.91 (12.01)
Information and communication	83 (2.39)	50.41 (11.19)
Financial and insurance activities	117 (3.37)	53.74 (10.46)
Professional, scientific, and technical activities	135 (3.89)	53.84 (12.10)
Administrative and support service activities	190 (5.48)	53.72 (11.97)
Public administration and defense; compulsory social security	200 (5.77)	54.28 (12.84)
Education	992 (28.6)	54.81 (12.04)
Human health and social work activities	99 (2.85)	54.67 (12.71)
Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	267 (7.7)	53.25 (12.40)
Other service activities	161 (4.64)	51.31 (12.59)
Sex		
Male	1439 (41.49)	52.31 (12.17)
Female	2029 (58.51)	53.82 (11.91)
Age (y)		
18-25	755 (21.77)	56.91 (11.49)
26-35	623 (17.96)	55.64 (11.86)
36-45	1096 (31.6)	52.08 (12.02)
46-55	925 (26.67)	50.07 (11.58)
56-69	69 (1.99)	49.91 (10.89)
Marital status		
Single (include divorced or widowed)	717 (20.67)	56.26 (12.27)
In a relationship	360 (10.38)	55.67 (11.44)
Married	2391 (68.95)	51.90 (11.83)
Monthly income (RMB ¥[a])		
500-1000	111 (3.2)	52.17 (12.37)
1000-1500	101 (2.91)	52.87 (11.94)
1500-2000	161 (4.64)	51.76 (13.07)
2000-3000	547 (15.77)	53.24 (11.48)
3000-5000	1167 (33.65)	53.93 (11.86)
5000-10,000	1022 (29.47)	53.07 (12.29)
10,000-20,000	268 (7.73)	52.52 (11.76)
>20,000	91 (2.62)	50.80 (12.83)
Educational background		
No education	27 (0.78)	51.04 (12.55)

XSL•FO
RenderX

Characteristics	Values, n (%)	SNSAS score, mean (SD)
Primary school	242 (6.98)	50.14 (11.34)
Junior high school	581 (16.75)	51.69 (11.09)
High school	459 (13.24)	51.57 (11.49)
College degree	345 (9.95)	53.15 (13.22)
Bachelor's degree	1527 (44.03)	54.65 (11.97)
Master's degree	248 (7.15)	54.19 (12.89)
Doctorate degree	39 (1.12)	51.82 (13.48)

[a]RMB ¥1=US $0.15.

Table 2. The frequency of Chinese workers using social media during work and staying up late using social media (N=3468).

Characteristics	Values, n (%)	SNSAS[a] score, mean (SD)
How often do you use social media for entertainment during the workday		
Almost all the time (even during work or meetings)	227 (6.55)	62.06 (11.99)
Frequent (almost all the time outside of work)	1513 (43.63)	56.05 (11.18)
Occasional (use when there is nothing to do)	1473 (42.47)	49.83 (10.96)
Hardly ever (never used unless necessary)	255 (7.35)	47.11 (13.71)
How many days a week do you stay up late using social media?		
0	94 (2.71)	46.85 (14.29)
1-2	286 (8.25)	44.80 (12.13)
3-4	1031 (29.73)	49.52 (10.56)
5-6	1208 (34.83)	54.36 (10.64)
Every day	849 (24.48)	59.51 (11.71)

[a]SNSAS: Social Networking Service Addiction Scale.

Distribution of the SNSAS, MBI-GS, and MAAS Scores of the Participants

The data in Table 3 show that the SMA score of all participants was 53.19 (SD 12.04), with a median score of 54. Among the 6 SNSAS scale dimensions, the highest average score was for mood modification (mean 9.63, SD 2.4 points), followed by salience (mean 9.48, SD 2.41 points). As mentioned by other researchers, the current SMA scale can only measure the degree of SMA in participants, and there are no clear diagnostic criteria [28]. However, the average score of the participants in this study exceeded half of the total score of the scale (45 points), with 76.85% (2665/3468) of the participants scoring SNSAS >45 points and 11.33% (393/3468) of the participants scoring 67.5 points (75% of the total score), indicating that the overall SMA score of the Chinese workers was relatively high. Among the 15 industries, the highest SNSAS score was in the education industry (mean 54.81, SD 12.04), followed by those engaged in human health and social work activities (mean 54.67, SD 12.71); the lowest score was in the information and communications industry (mean 50.41, SD 11.19) and agriculture, forestry, and fishery industry (mean 50.43, SD 11.44); and the rest of the industries scored between 54.28 (12.84) and 51.15 (SD 11.74). The average burnout score of the participants was 48.34 (SD 18.24), which was lower than the MBI-GS burnout standard (score ≥50), but the SD of the data was 18.24, and the burnout scores of the participants varied significantly. More than half of the participants (1863/3468, 53.72%) had scores >50, indicating that burnout is very common among Chinese workers. The average score of the workers was 62.83 (SD 11.66). At present, there is no method to clearly classify the mindfulness score, and as the total score of the MAAS scale was just 90, we believe that Chinese workers have a high mindfulness score overall.

XSL•FO
RenderX

Table 3. Social Networking Service Addiction Scale (SNSAS), Maslach Burnout Inventory–General Survey (MBI-GS), and Mindful Attention Awareness Scale (MAAS) scores of the participants.

Scale	Values, mean (SD)	Values, median (IQR)
SNSAS	53.19 (12.04)	54 (45-61)
Mood modification	9.63 (2.4)	10 (8-11)
Salience	9.48 (2.41)	9 (8-11)
Tolerance	9.21 (2.41)	9 (8-11)
Withdrawal	8.62 (2.97)	9 (6-11)
Conflict	8.25 (2.33)	8 (7-10)
Relapse	8 (2.48)	8 (6-9)
MBI-GS	48.34 (18.24)	52 (37.33-60)
Emotional exhaustion	12.75 (6.42)	11 (9-16)
Depersonalization	8.25 (5.23)	8 (4-11)
Personal accomplishment	15.25 (8.23)	17 (9-21)
MAAS	62.83 (11.66)	63 (57-71)

The Relationship Between Demographic Variables and SMA Scores

In Table 4, we conducted a *t* test for sex, and the total SMA score of female workers was significantly higher than that of male workers (t_{3466}=3.65; *P*<.001). ANOVA showed that there were significant differences in SMA scores among workers with different marital status (F_2=46.04; *P*<.001). Furthermore, we divided marital status into 3 groups and conducted *t* test analysis for each group. The results of the data analysis showed that there was no difference in SMA scores between single and "in a relationship" workers (t_{1075}=0.76; *P*=.45), with single workers having a higher SMA score than married workers (t_{3106}=8.6; *P*<.001) and "in a relationship" workers having a higher SMA score than married workers (t_{2749}=5.67; *P*<.001). The ANOVA results showed that the SMA scores were not the same for workers in different industries (F_{14}=3.98; *P*<.001). Pearson correlation analysis showed that the age of workers was significantly negatively correlated with the SMA score (*r*=−0.22; *P*<.001). There was no correlation between monthly income level and SMA score (*r*=−0.003; *P*=.88), but educational level was significantly positively correlated with SMA score (*r*=0.12; *P*<.001). How often workers used social media unreasonably at work (*r*=0.33; *P*<.001) and stayed up late using social media (*r*=0.14; *P*<.001) was significantly positively associated with the SMA score.

XSL•FO
RenderX

Table 4. The results of *t* tests on the participants' Social Networking Service Addiction Scale (SNSAS) scores and sex, and the results of analysis of variance on marital status and occupation.

Characteristic	SNSAS	Mood modification	Salience	Tolerance	Withdrawal	Conflict	Relapse
Sex (*t* test)							
t test (*df*)	−3.65 (3466)	−7.08 (3466)	−5.48 (3466)	−2.97 (3466)	−4.03 (3466)	0.04 (3466)	2.08 (3466)
P value	<.001	<.001	<.001	.003	<.001	.96	.04
Marital status (ANOVA)							
F test (*df*)	46.04 (2, 3465)	66.52 (2, 3465)	78.75 (2, 3465)	30.65 (2, 3465)	31.19 (2, 3465)	9.21 (2, 3465)	3.85 (2, 3465)
P value	<.001	<.001	<.001	<.001	<.001	<.001	.02
Single or in a relationship (*t* test)							
t test (*df*)	0.76 (1075)	−0.53 (1075)	0.40 (1075)	1.01 (1075)	−0.14 (1075)	1.91 (1075)	1.19 (1075)
P value	.45	.60	.69	.31	.89	.06	.23
Single or married (*t* test)							
t test (*df*)	8.60 (3106)	9.63 (3106)	11.03 (3106)	7.20 (3106)	6.42 (1095)	4.27 (3106)	2.68 (1119)
P value	<.001	<.001	<.001	<.001	<.001	<.001	.008
Married or in a relationship (*t* test)							
t test (*df*)	5.67 (2749)	7.88 (3106)	7.84 (3106)	4.28 (3106)	5.34 (3106)	1.07 (3106)	0.68 (3106)
P value	<.001	<.001	<.001	<.001	<.001	.28	.50
Work industry (ANOVA)							
F test (*df*)	3.98 (14)	4.57 (14)	5.50 (14)	2.86 (14)	3.40 (14)	2.45 (14)	1.61 (14)
P value	<.001	<.001	<.001	<.001	<.001	.002	.07

The Relationship Between MBI-GS Scores, MAAS Scores, and SNSAS Scores

Pearson correlation analysis (Table 5) showed that SMA level was positively correlated with burnout level (r=0.35; P<.001) and negatively correlated with mindfulness level (r=−0.55; P<.001). Only the dimension of personal accomplishment was not correlated with salience, tolerance, and withdrawal of SMA, whereas the other dimensions were positively correlated with SNSAS score and dimension score. The mindfulness score was negatively correlated with the score of each dimension of the SNSAS. Furthermore, we used job burnout and mindfulness as independent variables and SMA as the dependent variable for the linear regression analysis (Table 6). It was found that the model passed the F test ($F_{2,3465}$=780.36; P<.001). The unstandardized coefficient of job burnout was 0.07 (P<.001), indicating that job burnout had a significant positive influence on SMA. The unstandardized coefficient of mindfulness was −0.511 (P<.001), indicating that mindfulness had a significant negative effect on SMA. Comparing the standardized coefficients, the effect of mindfulness on the SNSAS score was greater than that of job burnout. The variance inflation factor values in the linear regression model were all <5, indicating no collinearity problem.

XSL•FO
RenderX

Table 5. The Pearson correlation analysis results between the participants' Social Networking Service Addiction Scale (SNSAS) scores and their Maslach Burnout Inventory–General Survey (MBI-GS) scores, Mindful Attention Awareness Scale (MAAS) scores, monthly income, age, and educational background.

	SNSAS	Mood modification	Salience	Tolerance	Withdrawal	Conflict	Relapse
MBI-GS							
r	0.35	0.31	0.30	0.27	0.25	0.27	0.29
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Emotional exhaustion							
r	0.36	0.35	0.33	0.30	0.27	0.25	0.25
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Depersonalization							
r	0.36	0.32	0.33	0.29	0.26	0.29	0.28
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Personal accomplishment							
r	0.06	0.03	0.03	0.03	0.03	0.07	0.11
P value	<.001	.04	.08	.12	.07	<.001	<.001
MAAS							
r	−0.55	−0.43	−0.45	−0.44	−0.40	−0.46	−0.47
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Monthly income							
r	−0.003	−0.04	0.01	0.00	0.01	0.03	−0.02
P value	.88	.02	.69	.83	.50	.11	.16
Age							
r	−0.22	−0.25	−0.25	−0.18	−0.19	−0.10	−0.08
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Educational background							
r	0.12	0.09	0.16	0.09	0.12	0.08	0.02
P value	<.001	<.001	<.001	<.001	<.001	<.001	.31

Table 6. Linear regression analysis results of Social Networking Service Addiction Scale (SNSAS) scores, Maslach Burnout Inventory–General Survey (MBI-GS) scores, and Mindful Attention Awareness Scale (MAAS) scores for all participants (dependent variable: SNSAS scores).

	Unstandardized coefficients value	Standardized coefficients value	*P* value	95% CI	VIF[a]
MBI-GS score	0.07	0.11	<.001	0.05 to 0.10	1.29
MAAS score	−0.51	−0.50	<.001	−0.54 to −0.48	1.29

[a]VIF: variance inflation factor.

Discussion

Principal Findings

Our study is one of the largest SMA studies conducted among workers. This study investigated the status quo of SMA among Chinese workers and analyzed its influencing factors. The results showed that the average level of SMA among Chinese workers was relatively high, and some workers had serious dependence on social media. Moreover, the results of data analysis also confirm our hypotheses 1 (except for "income"), 2, 3, and 4. The level of SMA among workers varies significantly across many demographic items and is positively correlated with the frequency of negative use of social media. At the same time, the level of SMA among workers was also positively affected by job burnout and negatively affected by mindfulness.

Demographic Differences in SMA Levels Among Workers

Occupation

To the best of our knowledge, this study is the first to simultaneously collect data on the level of SMA among multiple professional workers and compare their SMA levels. Previous studies have pointed out that the lack of cross-comparison owing to differences in sampling and classification is one of the

XSL•FO
RenderX

shortcomings of current research on SMA [50]. A study reviewed SMA research in 32 countries and found that different studies used different SMA measurement tools, and different studies had different standards for grading SMA levels [51]. Therefore, this study makes up for the shortcomings of the existing research. Statistically significant differences in SMA levels among workers in different occupations are one of the important findings of this study. Future research on SMA should consider the impact of occupation on the SMA levels of study participants.

Marital Status

The data showed that unmarried people had higher SMA scores than married people. Previous studies have also shown that unmarried people have a higher risk of SMA than married people [52]. Therefore, we conducted a deeper study on the relationship between marital relationships and personal SMA levels. First, it is worth noting that the data we collected this time showed that there were many married workers with high SMA scores, which accounted for 54.45% (214/3468) of all participants with scores >67.5. This means that although the average SMA score among married groups was not high, they still had a risk of SMA. Previous studies have found that family disharmony may lead to SMA among family members [53]. If married workers have a weak marital relationship, they may need social media to meet their communication and entertainment needs. In addition, this is the first SMA study to divide participants' marital status into 3 categories, namely, "single," "in a relationship," and "married." The results showed that although both love-involved and married workers had their own partners, there was a significant statistical difference in their SMA scores. Workers who are in love may need social media more to stay in touch, and attachment anxiety is a common factor that causes SMA in individuals [54]. Our study further explains the relationship between marital status and SMA level. In future research, compared with whether one is married, the quality of the marital relationship and the degree of attachment anxiety are more worthy considerations as factors influencing SMA.

Sex, Age, and Education Level

The SMA of female workers was more serious than that of male workers, which is similar to the conclusions drawn from previous studies. Women tend to attach more importance to maintaining relationships than men, tend to use social media as a tool for interaction, and are more likely to use social media to cope with feelings of emptiness when social needs are not met, leading to higher SMA [55]. In addition, younger workers tended to have higher SMA scores, which is consistent with previous studies indicating that younger people prefer communication on the internet and entertainment [5]. In addition, the level of education of workers was positively correlated with their level of SMA. Previous studies have reached similar conclusions in mixed samples of students and workers [56]. As the education level increases, the virtual tolerance subfactor increases and the risk of SMA increases [7].

Monthly Income

The SMA of workers is not correlated with their monthly income, which conflicts with our research results in the medical community [6]. Some studies have shown that family income satisfaction affects the SMA of family members [25], whereas other studies have shown no correlation [57]; however, most of these studies were conducted in students who do not have a stable job to earn a stable income. It is worth mentioning that this study and our previous study in the medical care group used absolute income (monthly income) and relative income (income satisfaction) to measure the income level of workers. Existing studies have found that both absolute income and relative income can affect people's positive or negative psychology, and the degree of impact of the 2 is quite different [58]. Therefore, in the study of worker's SMA, the choice of income measurement standard is worthy of further study. In summary, we believe that whether there is a correlation between workers' SMA and income is still debatable, and whether to use income satisfaction or monthly income to measure workers' income level in SMA research also deserves further discussion.

Workers' SMA and Inappropriate Use of Social Media

The SMA of Chinese workers was positively correlated with the frequency of unreasonable use of social media at work and the frequency of staying up late using social media. Previous studies have found that the frequency of using social media itself increases the risk of SMA [32], and if used in inappropriate situations or at inappropriate times, such as students using social media during class or self-study, the risk is further increased [59]. For workers, unreasonable social media use is mainly manifested in using social media during work. Fear of missing out [23] may be one of the reasons why workers cannot help but use social media while at work, as they fear missing out on friends' messages or the latest information on social media platforms. In addition, many workers with too many work tasks may only be able to squeeze in time for social media entertainment because of limited rest time.

Previous studies have shown that using social media before bedtime significantly reduces sleep efficiency [60], and poor sleep leads to burnout, which in turn leads to higher cognitive and emotional arousal, which is a core element of chronic insomnia [61]. Burnout increases the need for social media to alleviate stress for workers. Staying up late, burnout, and SMA thus form a vicious cycle.

SMA and Job Burnout in Workers

This study found that SMA among workers was significantly positively correlated with job burnout, which is consistent with the results of our previous small-scale research [6]. In research conducted among college students, SMA is generally explained as the cause of their poor psychological state or bad habits [59]. However, in the relationship between SMA and job burnout among workers, we preferred to analyze burnout as an influencing factor. We believe that the heavy workload of Chinese workers forced them to choose social media as their main form of entertainment.

At present, the work pressure of Chinese workers is generally high, and job burnout among workers is very common [62]. For

XSL•FO
RenderX

example, the main causes of job burnout among educators are overwork, concerns about students' examination results, and insufficient social support [63]. The reason for job burnout among medical practitioners in China is that they are constantly exposed to high-risk and high-pressure work environments [64]. In addition, in the era before the popularization of social media, many Chinese workers were plagued by job burnout [65]. Therefore, SMA is definitely not the main cause of job burnout among workers. Convenience and speed are the main advantages of using social media for entertainment. It takes only a few seconds to a minute to watch a short video, and this can be done with just a smartphone [66]. Social media allows workers who have insufficient rest time to take a break during work, commuting, and even while walking or eating. Previous studies have found that using social media primarily as a means of entertainment significantly elevates an individual's SMA level. Some studies have suggested that controlling social media use may improve workers' productivity [67], which we believe may not be effective and may even be counterproductive in people with high burnout, depriving them of one of their few forms of entertainment and causing other adverse effects. We believe that reducing workers' workload may be an effective way to reduce SMA among workers.

SMA and Mindfulness in Workers

This study is one of the earliest studies to investigate the relationship between mindfulness and SMA among workers. We found a significant positive correlation between worker mindfulness and SMA. Previous studies on college students have found similar results [68]. In previous studies, mindfulness was found to be negatively correlated with a variety of addictive behaviors, including substance addiction, such as drug addiction [69], and in technology addiction, such as internet addiction [41], mindfulness has a significant restraining effect on addictive behaviors. Our findings provide further evidence of the influence of mindfulness on technology addiction and that this influence persists on SMA in workers. Meanwhile, there was a significant negative correlation between workers' mindfulness scores and burnout scores (r=−0.472; P<.001), which may lead them to believe that they do not need to turn to social media to relieve work stress. On the basis of the findings of this study, we believe that it may be worth further research to improve workers' level of mindfulness through mindfulness training combined with cognitive behavioral therapy to reduce their dependence on social media.

Limitations

First, it is important to note that although this study encompasses a working population across 15 categories of occupations, many industries were not included in the sample, and the data in this paper are only from mainland China, therefore, whether the conclusions in this paper are applicable to all occupational groups or all countries or regions is uncertain. Second, the number of people in various industries collected in these data was not balanced. For example, the number of workers in the education industry was high, whereas the number of workers in the transportation and storage industry was very small. This imbalance may lead to inaccurate measurements of SMA levels in some occupational groups, and larger sample sizes are needed for research in the future. Third, in this study, social media includes most of the social media platforms commonly used by Chinese people, but we did not analyze specific social media or conduct a categorical analysis. In future research, we hope to discuss social media platforms by category or conduct empirical analysis on specific social media platforms, such as Douyin.

Conclusions

This study investigated the current status of SMA among Chinese workers and analyzed the impact of demographic factors, social media use habits, job burnout, and mindfulness on workers' SMA. It was found that the average level of SMA among workers was high, and it was affected by many factors. In particular, the risk of SMA among women, young people, single people, married people but whose family relationships are not harmonious, and highly educated workers deserve our attention. Many workers use social media as a main entertainment tool because of excessive work pressure and use it during work or stay up late, which makes them rely heavily on social media. Therefore, alleviating the work pressure on workers and allowing them to have more time to rest and participate in healthier activities are ways to reduce their dependence on social media. In terms of intervention strategies, the strong influence of mindfulness on workers' SMA makes it a very valuable research direction to develop a mindfulness intervention plan for workers. Social media is an indispensable tool for workers in their daily lives and work, and every worker will inevitably use social media. The government should strengthen the supervision of social media companies as many social media platforms operate with the aim of increasing user stickiness and allowing users to spend more time on social media. Medical institutions should provide popular science on SMA to help workers understand the harm of SMA and set up intervention medical programs for SMA to help them get rid of their dependence on social media. For workers, changing their bad social media habits and choosing healthier recreational activities are effective ways to reduce the risk of SMA.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (number 72204272). WK designed this study. WK, AL, WX, and YL participated in data collection, data analysis, and paper writing. All authors contributed to the manuscript and approved the submitted version.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Mougin F, Hollis KF, Soualmia LF. Inclusive digital health. Yearb Med Inform 2022 Aug 04;31(1):2-6 [FREE Full text] [doi: 10.1055/s-0042-1742540] [Medline: 36463863]
2. Casale S, Rugai L, Fioravanti G. Exploring the role of positive metacognitions in explaining the association between the fear of missing out and social media addiction. Addict Behav 2018 Oct;85:83-87. [doi: 10.1016/j.addbeh.2018.05.020] [Medline: 29864680]
3. Andreassen C, Pallesen S. Social network site addiction - an overview. Curr Pharm Des 2014;20(25):4053-4061. [doi: 10.2174/13816128113199990616] [Medline: 24001298]
4. Arora S, Mehta M. Love it or hate it, but can you ignore social media? - A bibliometric analysis of social media addiction. Comput Human Behav 2023 Oct;147:107831 [FREE Full text] [doi: 10.1016/j.chb.2023.107831]
5. Al-Samarraie H, Bello KA, Alzahrani AI, Smith AP, Emele C. Young users' social media addiction: causes, consequences and preventions. Inf Technol People 2021 Nov 18;35(7):2314-2343 [FREE Full text] [doi: 10.1108/itp-11-2020-0753]
6. Luo A, Kong W, He H, Li Y, Xie W. Status and influencing factors of social media addiction in Chinese medical care professionals: a cross-sectional survey. Front Psychol 2022 Apr 27;13:888714 [FREE Full text] [doi: 10.3389/fpsyg.2022.888714] [Medline: 35572263]
7. Koçak O, Arslan H, Erdoğan A. Social media use across generations: from addiction to engagement. Eur Integr Stud 2021 Sep 16;1(15):63-77 [FREE Full text] [doi: 10.5755/j01.eis.1.15.29080]
8. Gumus D, Sevim S, Kizil M. Social media addiction and adolescents: relationship between social media and eating behaviors during pandemic. ADDICTA 2023 Apr 06;46:S623-S624. [doi: 10.5152/ADDICTA.2023.22083]
9. Huang P, Latner JD, O'Brien KS, Chang Y, Hung C, Chen J, et al. Associations between social media addiction, psychological distress, and food addiction among Taiwanese university students. J Eat Disord 2023 Mar 21;11(1):43 [FREE Full text] [doi: 10.1186/s40337-023-00769-0] [Medline: 36945011]
10. Yılmazel G. Orthorexia tendency and social media addiction among candidate doctors and nurses. Perspect Psychiatr Care 2021 Oct 16;57(4):1846-1852. [doi: 10.1111/ppc.12758] [Medline: 33861475]
11. He Z, Li M, Liu C, Ma X. Common predictive factors of social media addiction and eating disorder symptoms in female college students: state anxiety and the mediating role of cognitive flexibility/sustained attention. Front Psychol 2021 Mar 29;12:647126 [FREE Full text] [doi: 10.3389/fpsyg.2021.647126] [Medline: 35422727]
12. Lin S, Longobardi C, Gastaldi FG, Fabris MA. Social media addiction and aggressive behaviors in early adolescents: the mediating role of nighttime social media use and sleep quality. J Early Adolesc 2023 Feb 27;44(1):41-58 [FREE Full text] [doi: 10.1177/02724316231160142]
13. Zhang K, Li P, Zhao Y, Griffiths MD, Wang J, Zhang MX. Effect of social media addiction on executive functioning among young adults: the mediating roles of emotional disturbance and sleep quality. Psychol Res Behav Manag 2023 May;16:1911-1920 [FREE Full text] [doi: 10.2147/PRBM.S414625] [Medline: 37255996]
14. Pellegrino A, Stasi A, Bhatiasevi V. Research trends in social media addiction and problematic social media use: a bibliometric analysis. Front Psychiatry 2022 Nov 10;13:1017506 [FREE Full text] [doi: 10.3389/fpsyt.2022.1017506] [Medline: 36458122]
15. Peng P, Liao Y. Six addiction components of problematic social media use in relation to depression, anxiety, and stress symptoms: a latent profile analysis and network analysis. BMC Psychiatry 2023 May 08;23(1):321 [FREE Full text] [doi: 10.1186/s12888-023-04837-2] [Medline: 37158854]
16. Üngüren E, Tekin Ö. The effects of social disconnectedness, social media addiction, and social appearance anxiety on tourism students' career intentions: the moderating role of self-efficacy and physical activity. J Hosp Leis Sports Tour Educ 2023 Nov;33:100463 [FREE Full text] [doi: 10.1016/j.jhlste.2023.100463]
17. Öztürk G, Albayrak Y, Beyazyüz M. Social media addiction among individuals who underwent rhinoplasty: myth or reality? Facial Plast Surg Aesthet Med 2020 May 04;22(5):336-341. [doi: 10.1089/fpsam.2020.0101] [Medline: 32366125]
18. Zhuang J, Mou Q, Zheng T, Gao F, Zhong Y, Lu Q, et al. A serial mediation model of social media addiction and college students' academic engagement: the role of sleep quality and fatigue. BMC Psychiatry 2023 May 12;23(1):333 [FREE Full text] [doi: 10.1186/s12888-023-04799-5] [Medline: 37173670]
19. Majid A, Yasir M, Javed A, Ali P. From envy to social anxiety and rumination: how social media site addiction triggers task distraction amongst nurses. J Nurs Manag 2020 Apr 30;28(3):504-513. [doi: 10.1111/jonm.12948] [Medline: 31891441]
20. Wang L, Yan S, Wang Y, Qiu J, Zhang Y. Do mobile social media undermine our romantic relationships? The influence of fear-of-missing-out on young people's romantic relationships. Research Square. Preprint posted online July 7, 2021 2021 [FREE Full text] [doi: 10.21203/rs.3.rs-636141/v1]
21. Bevan JL, Ang PC, Fearns JB. Being unfriended on Facebook: an application of expectancy violation theory. Comput Human Behav 2014 Apr;33:171-178. [doi: 10.1016/j.chb.2014.01.029]

XSL•FO
RenderX

22. Maftei A, Pătrău anu A. Digital reflections: narcissism, stress, social media addiction, and nomophobia. J Psychol (Forthcomong) 2023 Sep 14:1-14. [doi: 10.1080/00223980.2023.2256453] [Medline: 37707786]
23. Zhu X, Xiong Z. Exploring association between social media addiction, fear of missing out, and self-presentation online among university students: a cross-sectional study. Front Psychiatry 2022 May 13;13:896762 [FREE Full text] [doi: 10.3389/fpsyt.2022.896762] [Medline: 35633794]
24. Taş İ. The relationship between social ignore and social media addiction among adolescents: mediator effect of satisfaction with family life. Youth Soc 2021 Oct 27;55(4):708-729 [FREE Full text] [doi: 10.1177/0044118x211055210]
25. He Z, Li M, Ma X, Liu C. Family socioeconomic status and social media addiction in female college students: the mediating role of impulsiveness and inhibitory control. J Genet Psychol 2021 Dec 09;182(1):60-74. [doi: 10.1080/00221325.2020.1853027] [Medline: 33292090]
26. Huang L, Zhang J, Duan W, He L. Peer relationship increasing the risk of social media addiction among Chinese adolescents who have negative emotions. Curr Psychol 2021 Jul 20;42(9):7673-7681 [FREE Full text] [doi: 10.1007/s12144-021-01997-w]
27. Miranda S, Trigo I, Rodrigues R, Duarte M. Addiction to social networking sites: motivations, flow, and sense of belonging at the root of addiction. Technol Forecast Soc Change 2023 Mar;188:122280. [doi: 10.1016/j.techfore.2022.122280]
28. Kuss DJ, Kanjo E, Crook-Rumsey M, Kibowski F, Wang GY, Sumich A. Problematic mobile phone use and addiction across generations: the roles of psychopathological symptoms and smartphone use. J Technol Behav Sci 2018;3(3):141-149 [FREE Full text] [doi: 10.1007/s41347-017-0041-3] [Medline: 30238057]
29. Buneviciene I, Bunevicius A. Prevalence of internet addiction in healthcare professionals: systematic review and meta-analysis. Int J Soc Psychiatry 2021 Aug 22;67(5):483-491. [doi: 10.1177/0020764020959093] [Medline: 32962501]
30. Shrivastava A, Sharma MK, Marimuthu P. Internet addiction at workplace and it implication for workers life style: exploration from Southern India. Asian J Psychiatr 2018 Feb;32:151-155. [doi: 10.1016/j.ajp.2017.11.014] [Medline: 29275219]
31. Stănculescu E, Griffiths MD. Social media addiction profiles and their antecedents using latent profile analysis: the contribution of social anxiety, gender, and age. Telemat Inform 2022 Nov;74:101879. [doi: 10.1016/j.tele.2022.101879]
32. Flynn S, Noone C, Sarma KM. An exploration of the link between adult attachment and problematic Facebook use. BMC Psychol 2018 Aug 10;6(1):34 [FREE Full text] [doi: 10.1186/s40359-018-0245-0] [Medline: 30092833]
33. Sun Y, Zhang Y. A review of theories and models applied in studies of social media addiction and implications for future research. Addict Behav 2021 Mar;114:106699. [doi: 10.1016/j.addbeh.2020.106699] [Medline: 33268185]
34. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001 Feb;52(1):397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
35. Bakker AB, Demerouti E, Euwema MC. Job resources buffer the impact of job demands on burnout. J Occup Health Psychol 2005 Apr;10(2):170-180. [doi: 10.1037/1076-8998.10.2.170] [Medline: 15826226]
36. Carlin M, de los Fayos Ruiz JG. The burnout syndrome: historical evolution since the employment context to the sports field. An Psicol 2010;26:169-180 [FREE Full text]
37. Huang C, Xie X, Cheung SP, Zhou Y, Ying G. Job demands, resources, and burnout in social workers in China: mediation effect of mindfulness. Int J Environ Res Public Health 2021 Oct 07;18(19):10526 [FREE Full text] [doi: 10.3390/ijerph181910526] [Medline: 34639826]
38. Kabat‐Zinn J. Mindfulness‐based interventions in context: past, present, and future. Clin Psychol Sci Pract 2006 May 11;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
39. Garland EL. Mindfulness research in social work: conceptual and methodological recommendations. Soc Work Res 2013 Dec 19;37(4):439-448. [doi: 10.1093/swr/svt038]
40. Abbott RA, Whear R, Rodgers LR, Bethel A, Thompson Coon J, Kuyken W, et al. Effectiveness of mindfulness-based stress reduction and mindfulness based cognitive therapy in vascular disease: a systematic review and meta-analysis of randomised controlled trials. J Psychosom Res 2014 May;76(5):341-351 [FREE Full text] [doi: 10.1016/j.jpsychores.2014.02.012] [Medline: 24745774]
41. Song WJ, Park JW. The influence of stress on internet addiction: mediating effects of self-control and mindfulness. Int J Ment Health Addict 2019 Feb 27;17(4):1063-1075. [doi: 10.1007/s11469-019-0051-9]
42. Kim K, Milne GR, Bahl S. Smart phone addiction and mindfulness: an intergenerational comparison. Int J Pharm Healthc Mark 2018 Apr 03;12(1):25-43 [FREE Full text] [doi: 10.1108/ijphm-08-2016-0044]
43. Yang X, Zhou Z, Liu Q, Fan C. Mobile phone addiction and adolescents' anxiety and depression: the moderating role of mindfulness. J Child Fam Stud 2019 Jan 21;28(3):822-830 [FREE Full text] [doi: 10.1007/s10826-018-01323-2]
44. More than 360 degree assessment. Changsha Ranxing Information Technology Co. URL: https://www.wjx.cn/ [accessed 2024-01-29]
45. Wang X. Developing a mobile SNS addiction scale utilizing factor analysis and the Rasch model. Online Inf Rev 2019 Nov 11;43(7):1284-1301 [FREE Full text] [doi: 10.1108/oir-10-2018-0300]
46. Andreassen CS, Torsheim T, Brunborg GS, Pallesen S. Development of a Facebook addiction scale. Psychol Rep 2012 Apr 01;110(2):501-517. [doi: 10.2466/02.09.18.pr0.110.2.501-517]
47. Yuan Z, Yu D, Zhao H, Wang Y, Jiang W, Chen D, et al. Burnout of healthcare workers based on the effort-reward imbalance model: a cross-sectional study in China. Int J Public Health 2021 Feb 25;66:599831 [FREE Full text] [doi: 10.3389/ijph.2021.599831] [Medline: 34744566]

XSL•FO
RenderX

48. Maslach C, Jackson SE, Leiter MP. The Maslach burnout inventory manual. In: Zalaquett CP, Wood RJ, editors. Evaluating Stress: A Book of Resources. Lanham, MD: Scarecrow Press; 1997:191-218.
49. Deng YQ, Li S, Tang YY, Zhu LH, Ryan R, Brown K. Psychometric properties of the Chinese translation of the Mindful Attention Awareness Scale (MAAS). Mindfulness 2011 Sep 17;3(1):10-14. [doi: 10.1007/s12671-011-0074-1]
50. Kuss DJ, Griffiths MD. Social networking sites and addiction: ten lessons learned. Int J Environ Res Public Health 2017 Mar 17;14(3):311 [FREE Full text] [doi: 10.3390/ijerph14030311] [Medline: 28304359]
51. Hart J, Nailling E, Bizer GY, Collins CK. Attachment theory as a framework for explaining engagement with Facebook. Pers Individ Dif 2015 Apr;77:33-40. [doi: 10.1016/j.paid.2014.12.016]
52. Yang X, Yip BH, Lee EK, Zhang D, Wong SY. The relationship between technology use and problem technology use and potential psychosocial mechanisms: population-based telephone survey in community adults during COVID-19. Front Psychol 2021 Aug 9;12:696271 [FREE Full text] [doi: 10.3389/fpsyg.2021.696271] [Medline: 34434146]
53. Sujarwoto, Saputri RA, Yumarni T. Social media addiction and mental health among university students during the COVID-19 pandemic in Indonesia. Int J Ment Health Addict 2023 Jul 01;21(1):96-110 [FREE Full text] [doi: 10.1007/s11469-021-00582-3] [Medline: 34230820]
54. D'Arienzo MC, Boursier V, Griffiths MD. Addiction to social media and attachment styles: a systematic literature review. Int J Ment Health Addiction 2019 Apr 18;17(4):1094-1118. [doi: 10.1007/s11469-019-00082-5]
55. Su W, Han X, Yu H, Wu Y, Potenza MN. Do men become addicted to internet gaming and women to social media? A meta-analysis examining gender-related differences in specific internet addiction. Comput Human Behav 2020 Dec;113:106480. [doi: 10.1016/j.chb.2020.106480]
56. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719 [FREE Full text] [doi: 10.2196/27719] [Medline: 34081596]
57. Aydin S, Koçak O, Shaw TA, Buber B, Akpinar EZ, Younis MZ. Investigation of the effect of social media addiction on adults with depression. Healthcare (Basel) 2021 Apr 11;9(4):450 [FREE Full text] [doi: 10.3390/healthcare9040450] [Medline: 33920478]
58. Ball R, Chernova K. Absolute income, relative income, and happiness. Soc Indic Res 2007 Nov 27;88(3):497-529. [doi: 10.1007/s11205-007-9217-0]
59. Hamam B, Khandaqji S, Sakr S, Ghaddar A. Social media addiction in university students in Lebanon and its effect on student performance. J Am Coll Health (Forthcoming) 2023 Jan 03:1-7. [doi: 10.1080/07448481.2022.2152690] [Medline: 36595574]
60. Hjetland GJ, Skogen JC, Hysing M, Sivertsen B. The association between self-reported screen time, social media addiction, and sleep among Norwegian university students. Front Public Health 2021 Dec 16;9:794307 [FREE Full text] [doi: 10.3389/fpubh.2021.794307] [Medline: 34976935]
61. Rothe N, Specht M. Schlaf und Burnout: Ein Überblick. Somnologie 2021 Aug 24;25(3):186-196 [FREE Full text] [doi: 10.1007/s11818-021-00318-4]
62. Liu B, Chen H, Yang X, Hou C. Why work overtime? A systematic review on the evolutionary trend and influencing factors of work hours in China. Front Public Health 2019 Nov 15;7:343 [FREE Full text] [doi: 10.3389/fpubh.2019.00343] [Medline: 31803708]
63. Cheng H, Fan Y, Lau H. An integrative review on job burnout among teachers in China: implications for human resource management. Int J Hum Resour Manag 2022 Jun 17;34(3):529-561 [FREE Full text] [doi: 10.1080/09585192.2022.2078991]
64. Xia M, Wang J, Bi D, He C, Mao H, Liu X, et al. Predictors of job burnout among Chinese nurses: a systematic review based on big data analysis. Biotechnol Genet Eng Rev (Forthcoming) 2023 Jan 22:1-24. [doi: 10.1080/02648725.2023.2168910] [Medline: 36683283]
65. Xianyu Y, Lambert VA. Investigation of the relationships among workplace stressors, ways of coping, and the mental health of Chinese head nurses. Nurs Health Sci 2006 Sep 03;8(3):147-155. [doi: 10.1111/j.1442-2018.2006.00281.x] [Medline: 16911174]
66. Qin Y, Omar B, Musetti A. The addiction behavior of short-form video app TikTok: the information quality and system quality perspective. Front Psychol 2022 Sep 6;13:932805 [FREE Full text] [doi: 10.3389/fpsyg.2022.932805] [Medline: 36148123]
67. Majid A, Yasir M, Javed A, Ali P. From envy to social anxiety and rumination: how social media site addiction triggers task distraction amongst nurses. J Nurs Manag 2020 Apr 30;28(3):504-513. [doi: 10.1111/jonm.12948] [Medline: 31891441]
68. You C, Liu Y. The effect of mindfulness on online self-presentation, pressure, and addiction on social media. Front Psychol 2022 Dec 5;13:1034495 [FREE Full text] [doi: 10.3389/fpsyg.2022.1034495] [Medline: 36544442]
69. Nallet A, Briefer JF, Perret I. [Mindfulness in addiction therapy]. Rev Med Suisse 2015 Jun 24;11(480):1407-1409. [Medline: 26267948]

Abbreviations

MAAS: Mindful Attention Awareness Scale

XSL•FO
RenderX

MBI-GS: Maslach Burnout Inventory–General Survey
SMA: social media addiction
SNSAS: Social Networking Service Addiction Scale

Edited by A Mavragani; submitted 10.04.23; peer-reviewed by Y Zhu, Y Li, F Wang; comments to author 24.10.23; revised version received 31.10.23; accepted 07.02.24; published 06.03.24.

Please cite as:
Kong W, Li Y, Luo A, Xie W
Status and Influencing Factors of Social Media Addiction in Chinese Workers: Cross-Sectional Survey Study
J Med Internet Res 2024;26:e48026
URL: https://www.jmir.org/2024/1/e48026
doi:10.2196/48026
PMID:38446542

©Weitao Kong, Yuanyuan Li, Aijing Luo, Wenzhao Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Outcomes and Costs of the Transition From a Paper-Based Immunization System to a Digital Immunization System in Vietnam: Mixed Methods Study

Thi Thanh Huyen Dang[1], MD, PhD; Emily Carnahan[2], MPH; Linh Nguyen[3], MPH; Mercy Mvundura[2], PhD; Sang Dao[3], MPH; Thi Hong Duong[1], MD, PhD; Trung Nguyen[1], MPH; Doan Nguyen[1], MD; Tu Nguyen[3], MSc; Laurie Werner[2], MPA; Tove K Ryman[4], MPH, PhD; Nga Nguyen[3], MD, PhD

[1]National Expanded Program on Immunization, National Institute of Hygiene and Epidemiology, Hanoi, Vietnam

[2]PATH, Seattle, WA, United States

[3]PATH, Hanoi, Vietnam

[4]Bill & Melinda Gates Foundation, Seattle, WA, United States

Corresponding Author:
Nga Nguyen, MD, PhD
PATH
1101, 11th floor, Hanoi Towers
49 Hai Ba Trung Street
Hanoi, 100000
Vietnam
Phone: 84 243936221 ext 130
Email: ntnguyen@path.org

Abstract

Background: The electronic National Immunization Information System (NIIS) was introduced nationwide in Vietnam in 2017. Health workers were expected to use the NIIS alongside the legacy paper-based system. Starting in 2018, Hanoi and Son La provinces transitioned to paperless reporting. Interventions to support this transition included data guidelines and training, internet-based data review meetings, and additional supportive supervision visits.

Objective: This study aims to assess (1) changes in NIIS data quality and use, (2) changes in immunization program outcomes, and (3) the economic costs of using the NIIS versus the traditional paper system.

Methods: This mixed methods study took place in Hanoi and Son La provinces. It aimed to analyses pre- and postintervention data from various sources including the NIIS; household and health facility surveys; and interviews to measure NIIS data quality, data use, and immunization program outcomes. Financial data were collected at the national, provincial, district, and health facility levels through record review and interviews. An activity-based costing approach was conducted from a health system perspective.

Results: NIIS data timeliness significantly improved from pre- to postintervention in both provinces. For example, the mean number of days from birth date to NIIS registration before and after intervention dropped from 18.6 (SD 65.5) to 5.7 (SD 31.4) days in Hanoi (P<.001) and from 36.1 (SD 94.2) to 11.7 (40.1) days in Son La (P<.001). Data from Son La showed that the completeness and accuracy improved, while Hanoi exhibited mixed results, possibly influenced by the COVID-19 pandemic. Data use improved; at postintervention, 100% (667/667) of facilities in both provinces used NIIS data for activities beyond monthly reporting compared with 34.8% (202/580) in Hanoi and 29.4% (55/187) in Son La at preintervention. Across nearly all antigens, the percentage of children who received the vaccine on time was higher in the postintervention cohort compared with the preintervention cohort. Up-front costs associated with developing and deploying the NIIS were estimated at US $0.48 per child in the study provinces. The commune health center level showed cost savings from changing from the paper system to the NIIS, mainly driven by human resource time savings. At the administrative level, incremental costs resulted from changing from the paper system to the NIIS, as some costs increased, such as labor costs for supportive supervision and additional capital costs for equipment associated with the NIIS.

Conclusions: The Hanoi and Son La provinces successfully transitioned to paperless reporting while maintaining or improving NIIS data quality and data use. However, improvements in data quality were not associated with improvements in the immunization

XSL•FO
RenderX

program outcomes in both provinces. The COVID-19 pandemic likely had a negative influence on immunization program outcomes, particularly in Hanoi. These improvements entail up-front financial costs.

(J Med Internet Res 2024;26:e45070) doi:10.2196/45070

KEYWORDS

eHealth; digital health; immunization information system; electronic immunization registry; immunization; data quality; data use; costing

Introduction

Background

Since 2017, the National Immunization Information System (NIIS) in Vietnam has been used nationwide by immunization facilities from the national, provincial, district, and commune levels to capture, store, and access immunization data [1]. The NIIS is a digital system that includes an immunization registry comprising individual-level, longitudinal information on vaccine doses administered and a logistics management system for vaccines and related supplies. As of June 2022, the NIIS has recorded data from approximately 31 million clients and has been used across 15,000 immunization facilities.

Although there is a growing body of evidence on the outcomes associated with digital systems to support vaccine service delivery in low- and middle-income countries [2-9], there are still many questions regarding how they affect data quality, use, and vaccination outcomes in practice. A challenge that has emerged in multiple country contexts is the parallel or dual reporting required when digital systems are introduced and health workers are expected to continue to use the legacy paper-based forms in addition to the new digital system. For example, Tanzanian facilities that had transitioned entirely to using an electronic immunization registry had higher odds of system use compared with those maintaining parallel electronic and paper-based systems [10].

Moreover, there is limited evidence on the costs of development and implementation of the system, recurrent costs of the system, and cost implications of eliminating a parallel paper system at the service delivery level [11,12]. The Vietnamese experience of introducing and scaling digital tools for immunization presents an opportunity to help fill these evidence gaps.

History of the NIIS

Before the introduction of the NIIS, a paper-based immunization registry and vaccine stock management system were used. Health workers captured data on paper forms, which were compiled in a monthly report. The paper-based system created a significant workload for health workers, and the immunization data were often delayed or incomplete, limiting the availability of reliable data at the higher levels of the health system [13].

From 2009 to 2012, the National Expanded Program on Immunization (NEPI), in collaboration with PATH and the World Health Organization, developed and piloted a logistics management tracking system for vaccines and related supplies (VaxTrak) and an immunization registry (ImmReg) at the commune and district levels. After the pilot phase, ImmReg expanded to all districts in the pilot province, and VaxTrak was scaled nationwide. An evaluation of ImmReg in 2015 showed that the system was highly accepted by health workers and improved vaccine coverage and timeliness [6]. In 2014 and 2015, NEPI and PATH integrated ImmReg and VaxTrak into a single comprehensive software system. From 2016 to 2018, the Vietnam General Department of Preventive Medicine and PATH developed the NIIS based on the pilot software and scaled it nationwide [1].

Transition to a Paperless System

Beginning in 2018, the government of Vietnam collaborated with PATH to provide technical support to strengthen NIIS implementation and transition to a paperless system in 2 provinces: Hanoi and Son La. This work was funded by the Bill & Melinda Gates Foundation and was implemented from 2018 to 2022.

We hypothesized that transitioning to a paperless system would improve data quality and data use in the intervention areas. We hypothesized that if data quality and data use improved, these changes could lead to improvements in immunization program outcomes. The objective of this study was to examine the short-term outcomes and costs associated with the NIIS and the transition to paperless reporting, focusing on three main categories:

1. What are the changes in data quality and data use because to the paperless transition interventions?
2. What are the changes in immunization program outcomes (timeliness, dropout rates, and coverage) because of the paperless transition interventions?
3. What are the incremental financial costs associated with developing, deploying, and maintaining the NIIS, including the economic cost implications of transitioning to paperless reporting?

Methods

Overview

This mixed methods study aimed to evaluate the short-term outcomes and costs associated with the transition to a paperless system. Mixed methods were used to quantify the observed changes in outcomes and costs and to qualitatively describe why and how changes occurred. Pre-post analyses were conducted to understand the short-term changes in data quality, data use, and immunization program outcomes. Financial data were extracted from the project and partners' records, and interviews were conducted at various levels of the health system to inform the cost analysis. This study was conducted from July 2019 to November 2021.

XSL•FO
RenderX

Setting

This study was conducted in 2 intervention provinces, Hanoi and Son La. These provinces were selected for the transition to a paperless system because they have a variety of geographic, demographic, and health system characteristics that may influence digital readiness. Hanoi, the capital city, is primarily urban with a high population density, high immigration rate, good infrastructure, and many private-sector and fee-based facilities. In contrast, Son La is a mountainous border province primarily composed of rural districts and has low population density, limited resources, and fee-based facilities.

Each province provides immunizations in public district health centers, hospitals, and commune health centers (CHCs) as well as private fee-based immunization facilities (FIFs). Primary data collection for this study also occurred at the national level, for example, to capture financial costs from the project and partners related to the development of the NIIS.

The transition to a paperless system and overall immunization activities in Vietnam were impacted by the COVID-19 pandemic, starting in 2020. The government mandated social distancing lockdowns multiple times in each province, which meant that individuals were not allowed to leave their home without special authorization, and nonessential businesses were closed. Immunization services were disrupted and, in some cases, unavailable, as health care workers were occupied with the COVID-19 response.

Multimedia Appendix 1 includes an overview of the characteristics of the study provinces (Table S1 in Multimedia Appendix 1) and dates when COVID-19 social distancing was applied in each province (Table S2 in Multimedia Appendix 1).

Interventions

A technical working group composed of the Ministry of Health, NEPI, PATH, and Viettel (the NIIS developer) oversees the implementation of the NIIS. A readiness assessment was conducted from June to July 2019 to provide the technical working group with information about the progress, needs, and challenges of transitioning to a fully paperless immunization system [14]. NEPI and PATH designed interventions (summarized in Textbox 1) to support the transition to paperless reporting based on the readiness assessment results. Interventions included detailed implementation guidelines and standard operating procedures for the transition to paperless reporting, internet-based data review meetings, additional supportive supervision visits, and Zalo (Vietnam's popular social media and chat app) groups for end users to exchange knowledge and experiences in NIIS use.

Textbox 1. Key interventions implemented as part of the transition to paperless reporting.

Key interventions

- *Guidelines and training on the shift to paperless reporting*: implementation guidelines and standard operating procedures for the transition to paperless were implemented through a training of trainers and cascaded training approach for health workers.
- *Data quality and data use guidelines and training* for health workers at the province, district, and commune health center levels.
- *Internet-based data review meetings* at the district level where all communes share progress on paperless transition. Challenges identified through these meetings were used to prioritize areas for support. Initially, these meetings were held monthly but later shifted to quarterly.
- *Additional supportive supervision visits* from the government and PATH at district and commune facilities to support data quality, data use, and the overall transition to paperless. During the COVID-19 pandemic, these shifted to internet-based supportive supervision visits. Internet-based supportive supervision guidelines and trainings were developed for districts and provinces.
- *Zalo groups for end users* (at least 1 National Immunization Information System [NIIS] user per facility) in each district to exchange knowledge and experiences in NIIS use. Zalo is a popular social media and chat app in Vietnam.

Facilities began transitioning to paperless reporting in November 2019 in Son La and in January 2020 in Hanoi. All facilities in both provinces have retired from the paper-based immunization management logbook and have completely transitioned to paperless reporting using the NIIS as of January 2020. Although all sites are officially reporting using the NIIS, some paper-based systems are still used to comply with various inspections, payment procedures, or requirements from other ministries.

In addition to the key interventions (Textbox 1), an e-learning system and e-immunization cards were piloted at a smaller scale. The e-learning system was developed and piloted in 6 districts (in the 2 provinces) to train managers at the national, regional, provincial, and district levels and facility health workers (CHCs, FIFs, and hospitals) on using the NIIS. The e-immunization card, a mobile phone app that allows parents or clients to access their individual demographic information and vaccination data, was also developed and launched in the 2 provinces.

Study Design and Data Collection

Overview

Various methods were used to collect data related to each of the study aims (data quality and use, immunization performance, and costing) in the 2 provinces. This section describes the study design, data sources, and data collection approach for each study objective. Table S3 in Multimedia Appendix 1 summarizes the data collection methods across the 3 study aims.

Data Quality and Use

A pre- or postintervention study design was used to assess changes in data quality and use. Data collection included self-administered facility assessments, household surveys, and facility surveys conducted at a sample of CHCs, FIFs, and hospitals; details of the methodology have been published elsewhere [14]. The same methods were used for pre- and postintervention data collection, with the addition of in-depth

XSL•FO
RenderX

interviews with Expanded Program on Immunization (EPI) officers at the district and commune levels at postintervention.

The self-administered facility assessment was sent via email to all immunization facilities to collect basic information regarding infrastructure, capacity, and NIIS data use.

Facility surveys were conducted on a purposively selected sample of districts, communes, hospitals, and FIFs in each province. Purposive sampling was performed in consultation with NEPI and provincial Centers for Disease Control and Prevention (CDC) to select a mix of facility types (fee based, private, and public), geographies (urban, semiurban, rural, and mountainous), and experiences with NIIS. The study was designed to survey the same facilities at pre- and postintervention. At preintervention, 8 FIFs and 7 hospitals were included; at postintervention, 5 FIFs and 7 hospitals were included (the change was owing to 3 FIFs that had closed by the time of the postintervention survey). In each facility, 20 clients in the paper logbook were randomly selected. The facility surveys captured structured information about the facility, the use of the NIIS, and demographic and immunization information for the 20 sampled clients.

A household survey was conducted in a sample of households with children aged <2 years to capture demographic and immunization information about the children from their home-based immunization cards. The household survey was conducted in the same purposively selected communes as the facility surveys. Within each commune, villages or living quarters were selected for convenience, and all households with children in the defined age range were included.

The NEPI and PATH staff trained data collectors who used KoboToolbox platform for data entry for all data collection forms. More details on the sampling approach and a full list of facilities included in the different evaluations conducted as part of this study are included in Table S4 in Multimedia Appendix 1.

Structured interviews were conducted to understand the factors related to data quality and data use. CDC staff led the interviews with immunization managers at the district level and health workers at the commune level. Interviews involving NEPI and CDC personnel were performed by the PATH staff. Interviews were conducted over the phone and in person at the interviewees' workplaces, lasting around 35 minutes. The interviews were recorded with consent and transcribed.

Immunization Program Outcomes

A pre- or postintervention study design was used to assess changes in immunization program outcomes in the 2 provinces using NIIS data. NIIS data were exported for a pre- and postintervention cohort of children in each province. Due to the differing commencement dates of paperless reporting in 2 project provinces was different, Son La began in November 2019, while Hanoi followed in Jan 2020. As a result, the preintervention cohort group comprises children born between July 1 and September 30, 2018, while the postintervention group comprises those born between July 1 and September 30, 2020. Each child's immunization information was analyzed for their first 12 months of life.

Costing

The costing study used a mixed methods approach including primary data collection using a microcosting approach and secondary data collection from financial record reviews. The costs of all activities were estimated from the perspectives of the implementing organizations (Hanoi and Son La provinces for NIIS systems use and Viettel, PATH, and NEPI) and hence take the health system perspective. No client costs were included.

We captured the costs for the different activities of implementing the NIIS from the software design and development activities to the deployment, accounting for the costs of the different partners engaged in this process. For the partner leading the software design, development, and deployment, we include the costs of the infrastructure, server, bandwidth, technical support, help desk, training, and maintenance and operations of the system. For the NEPI and the subnational levels, the costs included those for development of training materials, conducting staff training, data entry, meetings, and internet setup, at facilities where it was needed.

Data on the incremental financial costs for designing, developing, and deploying the NIIS software were obtained through the NEPI and partner organization expenditure records review. We also obtained information from each EPI administration and health facility in the sample on the expenditures for NIIS-related training and meetings and other deployment costs, including costs for data back entry and internet or phone setup at facilities.

Data on the recurrent financial costs of the NIIS were obtained via interviews with the head of each facility or person in charge of the facility finances in each study facility in the 2 provinces. The NEPI provided records on hardware inventory and repair requests, which were used to estimate replacement rates and maintenance costs for equipment.

Data Analysis

Overview

This section describes the variables and data analysis approach for each study objective.

For qualitative data analysis, we first transcribed all the interviews. Then, a member of the research team, who was trained on qualitative data methods, was assigned to code the transcripts using Microsoft Excel. This coding process followed a content analysis approach and involved a 3-level coding process: initially, open coding was applied to 5 transcripts to identify major themes; subsequently, the research team held discussions to reach a consensus on the major themes and any emergent themes; and finally, a final codebook was created before coding the remaining transcripts. This approach helped to gain a comprehensive understanding of health workers' perspectives on improving data quality and data use from the NIIS as well as to identify the barriers and facilitators of immunization coverage.

In the analysis of quantitative data, we used Stata (version 14; StataCorp) as our statistical tool. For categorical variables, we used the chi-square test, and in cases where the expected cell

XSL•FO
RenderX

counts were <5, Fisher exact test was used. For continuous variables, we first checked the variable distribution, and given the absence of a normal distribution, the Wilcoxon-Mann-Whitney *U* test was used to compare group differences. In addition, to investigate the relationship between immunization outcomes and its determinants, we conducted a multivariable logistic regression analysis. The threshold for statistical significance was set at $P<.05$.

Data Quality and Use

The data use outcome of interest was the percentage of facilities using NIIS data to inform specified routine activities (eg, making monthly vaccination plans).

The data quality outcomes of interest were the quantitative measures of timeliness, completeness, and accuracy. Information from the household and health facility surveys was compared with the NIIS data to assess data quality. Refer to Table S5 in Multimedia Appendix 1 for the definitions of the data quality indicators.

Immunization Program Outcomes

Data exported from the NIIS were cleaned. Duplicate records, unreliable data (eg, vaccination date before birth date), and records with "lost to follow-up" status were excluded. The primary outcome of interest was on-time vaccination, determined by the recommended age for vaccine delivery according to the NEPI vaccination schedule [15]. The secondary immunization program outcomes of interest were dropout rates and full vaccination coverage. Multimedia Appendix 1 provides details on NIIS data cleaning and definitions for the primary and secondary outcomes of interest (Table S6 in Multimedia Appendix 1).

Costing

Quantitative data were analyzed using Stata (version 14). We computed the total incremental costs per health system level associated with the NIIS implementation. We also estimated the cost per child for the NIIS implementation activities. For this analysis, we allocated a proportion of the costs for NIIS implementation to the 2 provinces according to their annual birth cohort size relative to the national birth cohort. The annual birth cohort for Vietnam in 2019 was approximately 1.5 million [16], whereas the 2 study provinces (Hanoi and Son La) had an annual birth cohort of 162,000 and 25,000, respectively, based on data from the NIIS, representing approximately 12% (187,000/1,500,000) of the annual birth cohort of Vietnam. The NIIS implementation costs were spread over 5 birth cohorts in these cost-per-child calculations, as the system implementation was done over the 5-year period.

To estimate the economic costs associated with service delivery and reporting using either the paper-based system, electronic system, or both, we used an activity-based costing approach. Ingredients or components of the activities were quantified for each resource type, including human resource time use for different immunization activities where there would be a change when using the NIIS versus the paper system, capital costs for equipment and supplies, and recurrent costs for internet connectivity and equipment maintenance attributable to using the NIIS or the paper system. The quantification was done by conducting interviews at the study facilities using structured costing questionnaires. The unit cost of each resource was obtained from secondary data sources, and the total costs for each activity by resource type were estimated. To obtain the total costs, we aggregated the costs for each activity by resource type. We estimated the recurrent costs per facility and per child, with the latter based on only 1 birth cohort as these are annual recurrent costs.

Ethical Considerations

This study served as the end-line evaluation activity within the project's work plan. The project was a collaborative effort between the NEPI and PATH. This evaluation constitutes 1 of the project activities outlined in the project documents submitted to the Vietnam Ministry of Health. As per the regulations set forth by the Vietnamese government, the project documents were reviewed and certified by units within the Ministry of Health and other relevant ministries before receiving approval from the Ministry of Health. The study was reviewed, considered as project evaluation, and approved by the Vietnam Ministry of Health (decision 1996/QĐ-BYT), and this study does not require ethics review in accordance with the circular 04/TT-BYT issued by the Vietnam Ministry of Health [17]. This circular [17] regulates the establishment, functions, tasks, and rights of research ethics committees, and it is specified that only research involving human subjects necessitates research ethics committee approval before implementation and supervision during the research process. In addition, the study was reviewed by PATH's US-based Research Determination Committee, which concluded that the activity did not involve "human subjects" as defined in the US Government 45 Code of Federal Regulations 46.102(e) [18] and did not require US ethics review.

In addition, before conducting the interviews, comprehensive information was provided to all participants, encompassing the study's objectives, participant rights, and strict confidentiality measures applied to protect their personal information. Informed consent was diligently obtained from each participant before starting the interviews. Their consent to participate in the study was obtained before proceeding with the interviews. Each qualitative interviewee received VN 150.000 (US $6.50) as payment for their time. In facilities selected for the costing study, health facility staff received VN 400.000 (US $17) for their time participating in structured costing interviews. Interviews were conducted in private settings to ensure confidentiality. All information was coded and only accessible to the study team, and data privacy was emphasized during training of the data collection team. The NEPI provided official permission for the use of the data extracted from the NIIS. All identifying data were coded, and names were eliminated before data analysis.

Results

Data Quality and Use

Data quality and use were measured through the NIIS data export, household surveys, and facility surveys and further explained through qualitative interviews.

XSL•FO
RenderX

Data Quality

The NIIS data quality evaluation considered the attributes of timeliness, completeness, and accuracy.

Timeliness

On the basis of the data exported from the NIIS, timeliness significantly improved from pre- to postintervention for all indicators across all health system levels in both provinces (Table 1). Between pre- and postintervention, there was a significant decrease in the mean number of days from birth date to NIIS registration (Hanoi: 18.6, SD 65.5 to 5.7, SD: 31.4 days; *P*<.001 and Son La: 36.1, SD 94.2 to 11.7, SD 40.1; *P*<.001). Across all health system levels (CHCs, FIFs, and hospitals), there were significant decreases in the mean number of days from the injection date to when the injection was updated in the NIIS. Stock transactions (only assessed at the CHC level) also showed a significant decrease in the mean number of days from stock arrival date to NIIS voucher date (Hanoi: 10.5, SD 36.1 to 5.2, SD: 19.8 days; *P*<.001 and Son La 13.4, SD 38.1 to 6.5, SD 23.5; *P*<.001).

Table 1. Timeliness measures pre- and postintervention, broken down by facility type and province.

Level and indicator	Data source	Hanoi			Son La		
		Preintervention	Postintervention	*P* value	Preintervention	Postintervention	*P* value
Overall							
Registration, n	N/A[a]	33,861	35,622	N/A	6296	5706	N/A
Days from birth date to NIIS[b] registration day[c], mean (SD)	NIIS data export	18.6 (65.5)	5.7 (31.4)	<.001	36.1 (94.2)	11.7 (40.1)	<.001
Commune health center level							
Injections, n	N/A	197,497	148,601	N/A	55,958	48,807	N/A
Days from injection date to injection updated in the NIIS, mean (SD)	NIIS data export	6.9 (49.1)	2.8 (17.5)	<.001	30.6 (101.4)	4.2 (20.6)	<.001
Stock, n	N/A	76,483	79,573	N/A	12,596	12,168	N/A
Days from stock arrival date to NIIS voucher date[c], mean (SD)	NIIS data export	10.5 (36.1)	5.2 (19.8)	<.001	13.4 (38.1)	6. (23.5)	<.001
Fee-based immunization facility level							
Injections, n	N/A	202,889	313,117	N/A	1879	5459	N/A
Days from injection date to injection updated in the NIIS, mean (SD)	NIIS data export	11.5 (64.7)	1.4 (11.7)	<.001	32.3 (98.0)	4.3 (26.0)	<.001
Hospital Level							
Injections, n	N/A	27,988	34,698	N/A	5710	7490	N/A
Days from injection date to injection updated to the NIIS, mean (SD)	NIIS data export	11.6 (66.8)	1.9 (17.1)	<.001	9.0 (58.9)	1.9 (22.9)	<.001

[a]N/A: not applicable.

[b]NIIS: National Immunization Information System.

[c]Dependent variables were not normally distributed; therefore, the Wilcoxon-Mann-Whitney *U* test was used.

Completeness

Completeness was assessed by comparing information from the household and facility surveys with the information from the NIIS (Table 2). At the CHC level, completeness of registration, client information, and injection information captured in the NIIS significantly improved between pre- and postintervention in Hanoi and Son La. At the FIF level in Hanoi, there was a decline in the percentage of clients registered in the NIIS; however, among those registered, there was an increase in the completeness of client information. At the FIF level in Son La, there was increased completeness of registration, client information, and immunization information. At the hospital level, there was an improvement in the percentage of clients registered in the NIIS and a decline in the completeness of client information in both provinces, and the completeness of immunization information remained unchanged at 100% (Hanoi: preintervention n=63, postintervention n=153; Son La: preintervention n=41, postintervention n=74) in both provinces at pre- and postintervention.

XSL•FO
RenderX

Table 2. Completeness measures pre- and postintervention, by health system level and province.

Level and indicator	Data source	Hanoi			Son La		
		Preintervention	Postintervention	*P* value	Preintervention	Postintervention	*P* value
Commune health center level							
Registered clients, n	N/A[a]	229	367	N/A	149	110	N/A
Clients registered in the NIIS[b], n (%)	HH[c] survey	217 (94.8)	362 (98.6)	.007	121 (81.2)	110 (100)	<.001
Client information, n	N/A	217	362	N/A	119	110	N/A
Clients with personal information entered fully in the NIIS, n (%)	HH survey	199 (91.7)	359 (99.2)	<.001	101 (84.9)	110 (100)	<.001
Immunization information, n	N/A	2275	4070	N/A	1322	1106	N/A
Injections with immunization information entered fully in the NIIS, n (%)	HH survey	2230 (98.0)	4018 (98.72)	.04	1223 (92.51)	1105 (99.90)	<.001
Fee-based immunization facility level							
Registered clients, n	N/A	122	80	N/A	30	20	N/A
Clients registered in the NIIS, n (%)	FIF[d] survey	121 (99.2)	72 (90)	.003	27 (90)	(20) 100	.21
Client information, n	N/A	121	72	N/A	27	20	N/A
Clients with personal information entered fully in the NIIS, n (%)	FIF survey	117 (96.7)	71 (98)	.65	20 (74)	20 (100)	.02
Immunization information, n	N/A	121	100	N/A	27	32	N/A
Injections with immunization information entered fully in the NIIS, n (%)	FIF survey	117 (96.7)	97 (97.0)	.65	20 (74)	32 (100)	.02
Hospital level							
Registered clients, n	N/A	71	100	N/A	43	40	N/A
Clients registered in the NIIS, n (%)	Hospital survey	65 (91)	100 (100)	.005	41 (95)	39 (97)	.53
Client information, n	N/A	63	100	N/A	41	39	N/A
Clients with personal information entered fully in the NIIS, n (%)	Hospital survey	63 (100)	99 (99.0)	.30	41 (100)	38 (97)	.99
Immunization information, n	N/A	63	153	N/A	41	74	N/A
Injections with immunization information entered fully in the NIIS, n (%)	Hospital survey	63 (100)	153 (100)	.99	41 (100)	74 (100)	.99

[a]N/A: not applicable.

[b]NIIS: National Immunization Information System.

[c]HH: household.

[d]FIF: fee-based immunization facility.

Accuracy

The accuracy of demographic and immunization information was assessed by comparing information on clients' personal immunization cards (captured through the household survey) with their information entered in the NIIS. The percentage of clients with demographic information matched between the 2 sources significantly increased in Hanoi (199/217, 91.7% to 340/353, 96.3%; *P*=.02) and Son La (101/119, 84.9% to 107/107, 100%; *P*<.001) from pre- to postintervention (Table 3). The percentage of injections with immunization information matched between the 2 sources also significantly increased in Son La (1037/1216, 85.3% to 1097/1097, 100%; *P*<.001) but significantly decreased in Hanoi (2147/2188, 98.1% to 3786/3981, 95.1%; *P*=.01).

XSL•FO
RenderX

Table 3. Accuracy measures pre- and postintervention, by health system level and province.

Level and indicator	Data source	Hanoi			Son La		
		Preintervention	Postintervention	*P* value	Preintervention	Postintervention	*P* value
Overall							
Demographic information, n	N/A[a]	217	353	N/A	119	107	N/A
Clients with demographic information matched between personal immunization card and the NIIS[b], n (%)	HH[c] survey	199 (91.7)	340 (96.3)	.02	101 (84.9)	107 (100)	<.001
Immunization information, n	N/A	2188	3981	N/A	1216	1097	N/A
Injections with immunization information matched between personal immunization card and the NIIS, n (%)	HH survey	2147 (98.12)	3786 (95.10)	.01	1037 (85.27)	1097 (100)	<.001

[a]N/A: not applicable.

[b]NIIS: National Immunization Information System.

[c]HH: household.

The interviews indicated that respondents at all levels (national, provincial, district, and facility levels) and across both provinces had a strong understanding of data quality, defined as timeliness, completeness, and accuracy. Most respondents (13/16, 81%) participated in the intervention training on data quality and data use and indicated that it was useful for their work and that they had applied what they learned:

> *I have participated in the training course on data quality and data use last year. After the training, my knowledge and skill in data quality and data use are better, so I applied to my daily work, I usually check the input data to make them complete and accurate before entering into the system.* [CHC staff]

All respondents rated their facility's data quality in the NIIS as "good" or "very good" in terms of timeliness, completeness, and accuracy. All respondents mentioned human resources as the most important factor associated with data quality, including health workers' knowledge and skills (in data entry, analysis, quality assessment, and use), understanding the importance of data quality, and bandwidth to support the immunization program when working across health areas.

Data Use

Data use was measured through facility assessments asking health workers about the activities that the NIIS data were used to inform. In the preintervention survey, 34.8% (202/580) of the facilities in Hanoi and 29.4% (55/187) of the facilities in Son La indicated that they had used the NIIS data for additional activities beyond monthly reporting. In the postintervention survey, 100% (Hanoi: 468/468 and Son La: 199/199) of the facilities in both provinces indicated using the NIIS data for activities beyond monthly reporting. Table S7 in Multimedia Appendix 1 shows the frequency by activity. At postintervention, the most common uses of data among facilities were to inform monthly vaccination plans, campaign plans, or annual immunization plans. At the management level, the most common use of the NIIS data was to evaluate the performance of health facilities. From the qualitative interviews, the most frequently mentioned obstacle to using the data was the lack of health workers' capacity for data analysis and use.

Immunization Program Outcomes

On-time vaccination was the primary immunization performance outcome of interest. Secondary outcomes were dropout rates and vaccination coverage (refer to the Tables S8-12 in Multimedia Appendix 1).

Study Population Characteristics

Immunization outcomes were assessed by comparing the NIIS data for a cohort of children pre- and postintervention in the 2 provinces. After data cleaning, 81,485 children were included in the sample. Their population characteristics are summarized in Table S8 in Multimedia Appendix 1. In Hanoi, there were small but significant differences in the ethnicity and rural and urban location of children in the pre- and postintervention cohorts. In Son La, there was also a significant difference in the ethnicity of children in the pre- and postintervention cohorts. In both provinces, there were significant differences in the percentage of children vaccinated primarily from FIFs versus CHCs at pre- and postintervention.

On-Time Vaccination by Antigen

Across all antigens, apart from the measles-containing vaccine first dose in Hanoi, the percentage of children who received the vaccine on time increased from pre- to postintervention, and nearly all increases were statistically significant (Table 4).

XSL•FO
RenderX

Table 4. On-time vaccination coverage, by vaccine and province.

Vaccine	Hanoi			Son La		
	Preintervention (n=33,752), n (%)	Postintervention (n=35,611), n (%)	*P* value	Preintervention (n=6233), n (%)	Postintervention (n=5705), n (%)	*P* value
BCG[a]	29,972 (88.8)	32,400 (90.98)	<.001	3760 (60.32)	4410 (77.3)	<.001
Penta[b] 1	24,409 (72.31)	31,780 (89.24)	<.001	1935 (31.04)	2985 (52.32)	<.001
Penta 2	8660 (25.66)	13,540 (38.02)	<.001	1999 (32.07)	2810 (49.26)	<.001
Penta 3	8360 (24.77)	10,640 (29.88)	<.001	2169 (34.8)	2695 (47.24)	<.001
Full polio	19,920 (59.02)	26,930 (75.62)	<.001	1965 (31.53)	1855 (32.51)	.28
MCV1[c]	29,900 (88.59)	29,000 (81.44)	<.001	4413 (70.8)	4856 (85.12)	<.001

[a]BCG: bacillus Calmette-Guérin.

[b]Penta: pentavalent.

[c]MCV1: measles-containing vaccine first dose.

In the multiple logistic regression analysis, children at postintervention were approximately 1.18 times as likely in Hanoi and 1.69 times as likely in Son La to receive timely administration of pentavalent (Penta) 3 compared with those in the preintervention cohort (*P*<.001; Table 5). In Hanoi, there was no significant difference in timely Penta 3 vaccination by gender or ethnicity, but children in urban areas were 1.6 times as likely to receive Penta 3 on time compared with those in rural areas. In Son La, there was also no difference by gender, but Thai children were 1.3 times as likely to receive Penta 3 on time compared with other ethnicities. In both provinces, children who were mostly vaccinated from FIFs were more likely to receive Penta 3 on time compared with children mostly vaccinated from CHCs.

A separate logistic regression analysis for on-time full vaccination is included in Table S12 in Multimedia Appendix 1

Table 5. Factors associated with Penta 3 timeliness.

Characteristics	Hanoi (n=69,301)		Son La (n=11,918)	
	Odds ratio (95% CI)	*P* value	Odds ratio (95% CI)	*P* value
Survey round				
Preintervention	Reference	N/A[a]	Reference	N/A
Postintervention	1.18 (1.13-1.22)	<.001	1.69 (1.56-1.81)	<.001
Child's gender				
Girl	Reference	N/A	Reference	N/A
Boy	1.0 (0.98-1.04)	.49	1.0 (0.96-1.12)	.32
Ethnicity				
Kinh	1.1 (0.83-1.44)	.52	1.1 (0.98-1.20)	.12
Thai	0.7 (0.28-1.78)	.46	1.3 (1.21-1.44)	<.001
Other ethnicities	Reference	N/A	Reference	N/A
Region				
Rural	Reference	N/A	Reference	N/A
Urban	1.6 (1.51-1.62)	<.001	N/A	N/A
Vaccinated mostly from				
Commune health center	Reference	N/A	Reference	N/A
Fee-based facility	2.7 (2.62-2.82)	<.001	1.59 (1.1-2.3)	.01

[a]N/A: not applicable.

On-Time Full Immunization Coverage

Figure 1 shows the full immunization coverage over time for each birth cohort. In Hanoi, 87.8% (29,649/33,752) of the children in the preintervention cohort and 81.9% (29,167/35,611) of the children in the postintervention cohort (*P*<.001) had reached full immunization before their first birthday. In contrast, there was an increase in the percentage of

XSL•FO
RenderX

children reaching full immunization before their first birthday in Son La, from 63.3% (3947/6233) preintervention to 88.3% (5040/5705) postintervention (*P*<.001). The results for immunization coverage by antigen for the pre- and postintervention birth cohorts in each province are included Table S9 in Multimedia Appendix 1.

Figure 1. Full immunization coverage by age for the pre- (2018) and postintervention (2020) birth cohorts, by province.

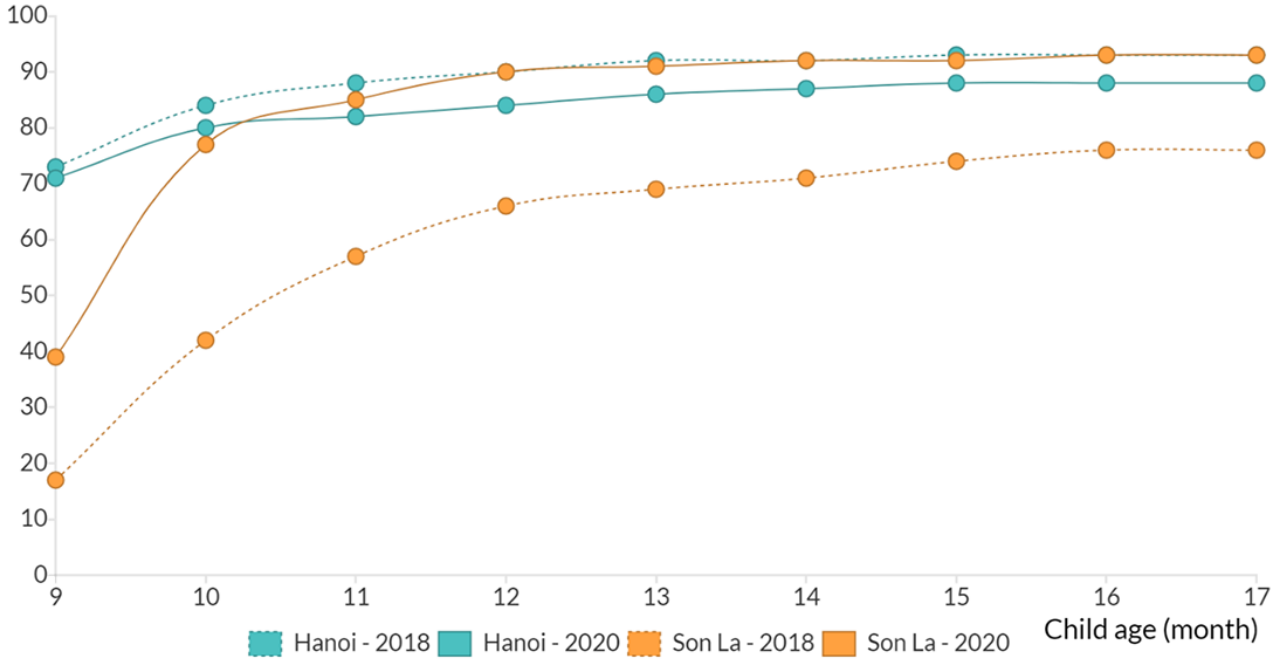

Costs

Up-Front Costs of NIIS Implementation

The up-front NIIS software design, development, and deployment took approximately 2300 person-months of labor from 2015 to 2020, and the partner costs for this labor were estimated at approximately US $1.75 million (Table 6). Most of the software developer costs (US $1,233,152/US $1,745,712, 71%) pertained to the deployment of the system. In addition, there were up-front costs per facility for implementing the NIIS, including costs for back entry of data, internet setup (where needed), training of users, and meeting costs. At the national and provincial levels, the bulk of the up-front costs were spent on training and meetings. As trainings were paid for by the higher administrative levels, training costs are low or 0 at district levels and health facilities. At these levels, the larger cost share was for deploying the NIIS.

XSL•FO
RenderX

Table 6. Estimated up-front financial costs of implementing the NIISa between 2015 and 2020.

Level and indicator	Amount (US $)
National level, total	
Software development	
NIIS design	198,457
NIIS development	314,103
NIIS deployment	1,233,152
Up-front costs	1,745,712
Training and deployment	53,911
Subnational levels, mean (SD)	
Provincial Centers for Diseases Control and Prevention (costs per province; n=2)	
Deployment and training	41,259 (41,830)
District health centers (costs per district; n=12)	
Deployment and training	762 (402)
Commune health centers (costs per commune health center; n=24)	
Deploying NIIS	69 (61.7)

[a]NIIS: National Immunization Information System.

The up-front costs for the NIIS implementation allocated to the 2 study provinces were approximately US $419,000 (Table 7). When these costs are allocated over 5 birth cohorts, the estimated cost per child for the NIIS implementation was estimated at US $0.48.

Table 7. Incremental financial cost for developing and deploying the NIIS[a] allocated to the 2 study provinces.

Cost classification	Amount (US $)
System design, development, and deployment costs by the software developer	217,726
Trainings and deployment costs by NEPI[b] and subnational levels	201,061
Total cost allocated to the 2 study provinces	418,787

[a]NIIS: National Immunization Information System.

[b]NEPI: National Expanded Program on Immunization.

Recurrent Costs for the NIIS

The software developer estimated the annual recurrent costs for the system operation to be US $85,000. At health facilities, the average monthly economic cost for health worker labor for immunization-related activities done using the paper system were estimated to be US $146 (Table 8). In comparison, the monthly cost for labor with the NIIS was US $67, which is less than half of the labor cost when using the paper system. The most significant savings in labor time costs, resulting from the transition from the paper system to the NIIS, occurred through reduced time spent on organizing immunization sessions, data management, and reporting. However, the NIIS also resulted in additional activities for staff, including checking of duplicates and introduction of e-immunization cards. The NIIS also added new facility costs, including recurrent costs for internet, printing, and SMS text messaging reminders and the capital costs of equipment, which amounts to an average cost of US $28 per facility. However, with the NIIS, there were savings in printing costs as registers and ledgers would not be printed, and the average costs of these were US $4.51 per month or US $58 per year. Overall, the total monthly costs per facility with NIIS (US $95) are lower than with the paper system (US $151).

XSL•FO
RenderX

Table 8. Estimated recurrent economic costs per facility for the paper-based system compared with the NIIS[a] at the commune health centers (n=24).

Health facilities	Cost per facility (US $), mean (SD)	Estimated cost (US $), IQR
Monthly cost of the paper system		
Labor cost for health workers	146 (15.1)	23-155
Printing of registers and stock ledgers	4.51 (1.7)	4.51-4.51
Total monthly cost of the paper system	151 (15.2)	28-160
Monthly cost of the NIIS		
Recurrent cost	2 (1.1)	01-03
Equipment cost	26 (4.1)	19-33
Labor cost for health workers	67 (36.9)	15-86
Total monthly cost of the NIIS	95 (35.3)	35-122

[a]NIIS: National Immunization Information System.

Table 9 presents similar costs for the administrative levels. At most levels, except at provincial CDC, labor costs are lower with the NIIS than with the paper system. At the CDC, the labor costs for supportive supervision are the largest share of costs, and these costs increase when using the NIIS compared with when using the paper system. At the district health centers, there is a decline in labor costs for activities such as management and reporting when using the NIIS, and hence, labor costs are lower with the NIIS. As mentioned above, implementing the NIIS incurs additional costs, including recurrent costs for internet, printing, and equipment maintenance and capital costs for equipment, which makes the total monthly costs for NIIS more than the total monthly costs for the paper system. There are incremental costs resulting from the change from the paper system to the NIIS at the administrative levels.

XSL•FO
RenderX

Table 9. Estimated economic costs per district, province, and at the national level for the paper-based system compared with the NIIS[a].

Level and cost	Cost per facility (US $), mean (SD)	Estimated cost (US $), IQR
District health centers (n=12)		
Monthly cost of the paper system		
Labor cost for health workers	74 (35.5)	17-88
Printing of registers and stock ledgers	4.51 (1.2)	4.51-451
Total monthly cost of the paper system	79 (36)	22-93
Monthly cost of the NIIS		
Recurrent cost	296 (141.4)	153-419
Equipment cost	46 (11.5)	32-62
Labor cost for health workers	61 (32.2)	10-63
Total monthly cost of the NIIS	403 (151.4)	195-543
Provincial Centers for Disease Control and Prevention (n=2)		
Monthly cost of the paper system		
Labor cost for health workers	212 (130)	74-350
Printing of registers and stock ledgers	4.51 (1.7)	4.51-451
Total monthly cost of the paper system	217 (130.3)	79-355
Monthly cost of the NIIS		
Recurrent cost	1 (0.09)	1-1
Equipment cost	148 (13)	132-163
Labor cost for health workers	300 (235.5)	73-526
Total monthly cost of the NIIS	449 (224.9)	206-690
National Expanded Program on Immunization		
Monthly cost of the paper system		
Labor cost for health workers	320 (113.1)	N/A[b]
Printing of registers and stock ledgers	4.51 (0.7)	N/A
Total monthly cost of the paper system	325 (113.8)	N/A
Monthly cost of the NIIS system		
Recurrent cost	152 (4.2)	N/A
Equipment cost	15 (0)	N/A
Labor cost for health workers	287 (18.4)	N/A
Total monthly cost of the NIIS	455 (22.6)	N/A

[a]NIIS: National Immunization Information System.

[b]N/A: not applicable.

Although labor costs decrease at most health system levels with the transition to the NIIS, in practice, this may not translate into budget line savings as staff are retained and their time is reallocated. The financial costs associated with the paper system primarily entail printing registers, but these expenses are relatively minor in comparison with the those incurred with the electronic system. The estimated annual financial recurrent costs per child for the NIIS are US $3.17, and these account for the annual costs of the server (with the relevant portion allocated to the study provinces) and the costs for capital equipment and internet. There are incremental financial costs to the health system when moving from the legacy paper system to the electronic system.

Discussion

Principal Findings

Health workers at the province, district, and CHC levels successfully transitioned from the legacy paper-based system to the NIIS for paperless reporting, proving that the transition was possible in 2 very different provincial contexts. Although the transition to paperless reporting results in lower labor costs at the facility, district, and national levels, it requires incremental

XSL•FO
RenderX

financial recurrent costs (estimated at US $3.17 per child per year) to maintain the NIIS.

This study found improvements in data quality in Hanoi and Son La between pre- and postintervention. There were significant improvements in data timeliness at all levels of the health system and in both provinces. It is likely that the interventions contributed to this improvement because the guidelines, trainings, data review meetings, and supportive supervision visits emphasized the importance of timely data. Timeliness indicators performed better in Hanoi than in Son La, which may be because of the different service delivery practices; in Hanoi, vaccines are delivered on a weekly basic, whereas in Son La, they are delivered monthly. In addition, in Hanoi, more clients visit FIFs that deliver immunization daily.

In Son La, there were also improvements across nearly all the completeness and accuracy indicators. However, there were mixed results in terms of data completeness and accuracy in Hanoi, with some statistically significant declines in quality. This may be partly because Hanoi was more affected by the COVID-19 pandemic than Son La; there were 6 rounds of mandated social distancing across all 30 districts in Hanoi. When immunization services were available, facilities may have been understaffed if human resources were reassigned to the COVID-19 response and they may have faced higher demand from a backlog of children who could not be vaccinated during social distancing. Overburdened health workers may prioritize service delivery over data quality.

The observed differences in provincial results may also have been influenced by the way interventions were delivered in each province. The facilities in Son La received timely support through internet-based supervisory assistance. In Hanoi, they maintained in-person supportive supervision, which was less timely, and limited the number of facilities that could be visited. Facilities in Son La also received refresher trainings, which Hanoi did not, as Hanoi had an ongoing focus on COVID-19 vaccination.

There were large increases from pre- to postintervention in the percentage of facilities using the NIIS data to inform their activities. Results from the costing analysis showed that at the CHCs, health workers' time spent on immunization program data entry and reporting declined as the transition to paperless resulted in efficiencies and reduced workload. It is possible that this saved time was dedicated to data use.

For most activities, a higher percentage of facilities in Son La reported using NIIS data compared with Hanoi. These results may have also been influenced by the COVID-19 pandemic or the differences in project interventions. In addition, Hanoi has a more mobile population, with clients frequently moving between facilities. This implies that vaccination plans informed by NIIS data may not be as accurate or comprehensive compared with a location such as Son La, where there is less population movement.

Although only 44.7% (84/188) of facilities in Son La and 27.7% (125/451) of facilities in Hanoi reported using the NIIS to evaluate data quality, previous studies have shown that data use can drive data quality improvements [19]. Through the qualitative interviews, health workers recommended building data quality indicators directly into the NIIS to help users understand their data quality easily.

The study also found that the majority of on-time vaccination rates for various antigens in the 2 project provinces improved compared to those in the preintervention group. With more timely data captured in the NIIS, health workers know which vaccines a child is due for and can send reminders or follow-up to deliver the vaccines on time. For the clients using the e-immunization card, it may have also contributed to timely vaccination by reminding clients of their vaccination schedule and due dates. However, the on-time vaccination rate for the first dose of measles in Hanoi decreased after the intervention compared to before, which may have been owing to the COVID-19 lockdowns from July to September 2021, when the postintervention cohort was aged 9 to 12 months. In Hanoi, many parents prefer waiting to receive the measles-rubella vaccine at 12 months, which is delivered at FIFs, versus the monovalent measles vaccine, which is part of the standard EPI schedule at 9 months.

In Son La, improvements in data quality and on-time vaccination were also associated with improvements in vaccination coverage. Full immunization at 12 months was 88.3% (5040/5705) in the postintervention cohort compared with 63.3% (3947/6233) in the preintervention cohort. A previous evaluation of ImmReg (the earlier version of the NIIS) in another province in Vietnam, Ben Tre, also found improvements in on-time vaccination and vaccination coverage when comparing pre- and postintroduction of the digital immunization registry, and these improvements were sustained or increased 1 year later [6]. However, in Hanoi, despite improvements in on-time vaccination, there were declines in vaccination coverage, again likely because of the COVID-19 pandemic. According to the 2021 annual EPI report for Vietnam, full immunization coverage decreased nationwide from 96.8% in 2020 to 87.1% in 2021 [20].

Introducing these new electronic systems, which have improved data quality and timeliness and could potentially improve immunization outcomes, comes at a cost, as up-front financial investments have to be made for software design, development, and deployment. The cost per child for NIIS development and deployment was estimated at US $0.48 in the 2 study provinces in Vietnam. This is much lower than that estimated in Tanzania and Zambia, the only other low- and middle-income countries with comparable cost estimates for the development of electronic immunization registries [11,12]. In our study, costs were annualized over 5 birth cohorts, whereas in Tanzania and Zambia, costs were annualized over 3 birth cohorts. Even if we were to use 3 birth cohorts for this analysis, our estimated costs for Vietnam would still be lower. The estimated lower costs per child for Vietnam could be in part because most CHCs had existing computers, and hence, there was no mass procurement of equipment required, which is not the case in many other countries. In Vietnam, the equipment and connectivity costs were shared across multiple health programs, further reducing the costs borne by the immunization program, unlike in Tanzania and Zambia, where immunization was the first health area to be digitalized at the health facility level. The availability of

XSL•FO
RenderX

electricity at the health facility level in Vietnam also reduced costs, as there was no need for the procurement of alternative power sources such as solar chargers. In addition, software development in Vietnam was conducted by an in-country telecommunications partner that provided in-kind and pro bono services and support, some of which could not be quantified, resulting in lower costs than if done through an external organization.

Our study found that at the CHC level, there are cost savings for health workers' time use and other resources such as printing with the NIIS compared with the paper system, and at the administrative level, there are incremental costs. The NIIS results in some savings in labor time, although some new activities are added for the staff. There are also recurring costs, such as ongoing technical support and maintenance provided by Viettel, the need for refresher training at the local level for new staff, and expenses associated with equipment maintenance and replacement as well as connectivity provided by local government. These recurrent costs are important to consider for sustainability.

Limitations

Our study has several limitations. First, we conducted the study in 2 provinces and had small sample sizes for the costing results, facility survey results for data quality, and other results based on provider interviews. The small sample size and purposive sampling limit the generalizability of our findings; however, 2 distinct provinces with a mix of facility types were intentionally sampled so that the results can be used to understand the costs and outcomes of the transition to paperless reporting in a range of settings. Second, the time frame for the evaluation was short, which limited the magnitude of the change that could be observed. Third, there were minor differences in the data collection approaches and responses between pre- and postintervention. For example, the urban sampling approach was adjusted for the feasibility of data collection, and the response rate for facility assessment was lower at postintervention. Fourth, for data collected through provider interviews, such as for the costing study, potential recall bias arises due to retrospective nature of the inquires, as interviewees were asked to recall cost from past periods. Future studies should consider including costing data collection at different phases of the project to reduce recall bias. This would also facilitate the availability of costing data available to inform decision-making at various implementation phases. Our study may also have underestimated the costs associated with the paper system, as we did not incorporate expenses related to printing paper home-based records and invitation letters. These materials are printed to assist the caregivers in monitoring their children's vaccination history or reminding them when their child is due for vaccination. We also have not accounted for the saved labor costs for health care workers who would deliver the invitation letters to households. Another limitation is that the COVID-19 pandemic affected the originally planned timelines for data collection and the immunization service delivery. We did not try to account for the impact of the pandemic on findings regarding the data quality and data use presented in this study.

Conclusions

Health workers in the 2 provinces successfully transitioned to paperless reporting while maintaining or improving data quality. We recommend that other provinces in Vietnam transition to paperless reporting by introducing the guidelines and standard operating procedures used in Hanoi and Son La and providing ongoing support through trainings, data review meetings, supportive supervision, and peer networks. Future studies should monitor data quality and immunization outcomes in other provinces as well as the sustainability of the observed changes in Hanoi and Son La.

Introducing these new electronic systems comes with costs—both up-front and recurrent—but there are advantages, as seen in the improvements in data quality and on-time vaccination. In Vietnam, stakeholders should plan and budget for the sustainability of the system at each level of the health system, given the recurrent costs including repairing and replacement of equipment, connectivity, refresher training, software system, and supportive supervision. Other countries planning to implement similar interventions should plan to collect costing data throughout to inform decision-making and budgeting.

Acknowledgments

The authors thank the Bill & Melinda Gates Foundation for providing support for this study and the National Immunization Information System (NIIS) Technical Working Group members for their leadership in designing, developing, and deploying the NIIS. The authors thank the data collectors who participated in this study. Finally, the authors acknowledge the invaluable collaboration of the health workers, managers, and leaders in Hanoi and Son La provinces.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Supplementary tables.
[DOCX File , 400 KB - jmir_v26i1e45070_app1.docx]

References

1. Dang H, Dao S, Carnahan E, Kawakyu N, Duong H, Nguyen T, et al. Determinants of scale-up from a small pilot to a national electronic immunization registry in Vietnam: qualitative evaluation. J Med Internet Res 2020 Sep 22;22(9):e19923 [FREE Full text] [doi: 10.2196/19923] [Medline: 32960184]
2. Yohana E, Mitiku S, Kayumba PC, Swalehe O. Electronic immunization registry in improving vaccine supply chain availability in Tanga City Council, Tanzania. Rwanda J Med Health Sci 2021 Sep 10;4(2):223-236. [doi: 10.4314/rjmhs.v4i2.3]
3. Siddiqi DA, Abdullah S, Dharma VK, Shah MT, Akhter MA, Habib A, et al. Using a low-cost, real-time electronic immunization registry in Pakistan to demonstrate utility of data for immunization programs and evidence-based decision making to achieve SDG-3: insights from analysis of big data on vaccines. Int J Med Inform 2021 May;149:104413. [doi: 10.1016/j.ijmedinf.2021.104413] [Medline: 33652259]
4. Uddin MJ, Shamsuzzaman M, Horng L, Labrique A, Vasudevan L, Zeller K, et al. Use of mobile phones for improving vaccination coverage among children living in rural hard-to-reach areas and urban streets of Bangladesh. Vaccine 2016 Jan 04;34(2):276-283 [FREE Full text] [doi: 10.1016/j.vaccine.2015.11.024] [Medline: 26647290]
5. Chen L, Du X, Zhang L, van Velthoven MH, Wu Q, Yang R, et al. Effectiveness of a smartphone app on improving immunization of children in rural Sichuan Province, China: a cluster randomized controlled trial. BMC Public Health 2016 Aug 31;16(1):909 [FREE Full text] [doi: 10.1186/s12889-016-3549-0] [Medline: 27581655]
6. Nguyen NT, Vu HM, Dao SD, Tran HT, Nguyen TX. Digital immunization registry: evidence for the impact of mHealth on enhancing the immunization system and improving immunization coverage for children under one year old in Vietnam. Mhealth 2017;3:26 [FREE Full text] [doi: 10.21037/mhealth.2017.06.03] [Medline: 28828373]
7. Gilbert SS, Thakare N, Ramanujapuram A, Akkihal A. Assessing stability and performance of a digitally enabled supply chain: retrospective of a pilot in Uttar Pradesh, India. Vaccine 2017 Apr 19;35(17):2203-2208 [FREE Full text] [doi: 10.1016/j.vaccine.2016.11.101] [Medline: 28364932]
8. Gilbert SS, Bulula N, Yohana E, Thompson J, Beylerian E, Werner L, et al. The impact of an integrated electronic immunization registry and logistics management information system (EIR-eLMIS) on vaccine availability in three regions in Tanzania: a pre-post and time-series analysis. Vaccine 2020 Jan 16;38(3):562-569 [FREE Full text] [doi: 10.1016/j.vaccine.2019.10.059] [Medline: 31706808]
9. Fritz J, Herrick T, Gilbert SS. Estimation of health impact from digitalizing last-mile logistics management information systems (LMIS) in Ethiopia, Tanzania, and Mozambique: a lives saved tool (LiST) model analysis. PLoS One 2021;16(10):e0258354 [FREE Full text] [doi: 10.1371/journal.pone.0258354] [Medline: 34695158]
10. Carnahan E, Ferriss E, Beylerian E, Mwansa FD, Bulula N, Lyimo D, et al. Determinants of facility-level use of electronic immunization registries in Tanzania and Zambia: an observational analysis. Glob Health Sci Pract 2020 Sep 30;8(3):488-504 [FREE Full text] [doi: 10.9745/GHSP-D-20-00134] [Medline: 33008860]
11. Mvundura M, Di Giorgio L, Lymo D, Mwansa FD, Ngwegwe B, Werner L. The costs of developing, deploying and maintaining electronic immunisation registries in Tanzania and Zambia. BMJ Glob Health 2019 Nov 25;4(6):e001904 [FREE Full text] [doi: 10.1136/bmjgh-2019-001904] [Medline: 31803511]
12. Mvundura M, Di Giorgio L, Vodicka E, Kindoli R, Zulu C. Assessing the incremental costs and savings of introducing electronic immunization registries and stock management systems: evidence from the better immunization data initiative in Tanzania and Zambia. Pan Afr Med J 2020 Feb 12;35(Suppl 1):11 [FREE Full text] [doi: 10.11604/pamj.supp.2020.35.1.17804] [Medline: 32373262]
13. World Health Organization, PATH. Optimize: Vietnam report. PATH. 2013. URL: https://media.path.org/documents/TS_opt_vietnam_rpt.pdf [accessed 2024-01-29]
14. Duong H, Dao S, Dang H, Nguyen L, Ngo T, Nguyen T, et al. The transition to an entirely digital immunization registry in Ha Noi province and Son La Province, Vietnam: readiness assessment study. JMIR Form Res 2021 Oct 25;5(10):e28096 [FREE Full text] [doi: 10.2196/28096] [Medline: 34694232]
15. Minh An DT, Lee JK, Van Minh HV, Trang NT, Thu Huong NT, Nam YS, et al. Timely immunization completion among children in Vietnam from 2000 to 2011: a multilevel analysis of individual and contextual factors. Glob Health Action 2016 Mar 01;9(1):29189 [FREE Full text] [doi: 10.3402/gha.v9.29189] [Medline: 26950555]
16. Vietnam: country information. Gavi, The Vaccine Alliance. URL: https://www.gavi.org/programmes-impact/country-hub/west-pacific/vietnam [accessed 2022-09-28]
17. 4/TT-BYT on establishment, functions, tasks and rights of research ethics committees, 05 March 2020. Vietnam Ministry of Health. URL: https://thuvienphapluat.vn/van-ban/Bo-may-hanh-chinh/Thong-tu-4-TT-BYT-2020-thanh-lap-chuc-nang-nhiem-vu-Hoi-dong-dao-duc-trong-nghien-cuu-y-sinh-hoc-440446.aspx [accessed 2024-01-25]
18. Home page. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2024-01-25]

XSL•FO
RenderX

19. Werner L, Seymour D, Puta C, Gilbert S. Three waves of data use among health workers: the experience of the better immunization data initiative in Tanzania and Zambia. Glob Health Sci Pract 2019 Sep;7(3):447-456 [FREE Full text] [doi: 10.9745/GHSP-D-19-00024] [Medline: 31558600]
20. National immunization program and vaccine delivery the national expanded program on immunization. The World Bank Document. URL: https://openknowledge.worldbank.org/server/api/core/bitstreams/022f4737-59de-4f41-960f-e82f05093402/content [accessed 2024-02-27]

Abbreviations

CDC: Centers for Disease Control and Prevention
CHC: commune health center
EPI: Expanded Program on Immunization
FIF: fee-based Immunization facility
NEPI: National Expanded Program on Immunization
NIIS: National Immunization Information System
Penta: Pentavalent

Edited by T Leung, T de Azevedo Cardoso; submitted 14.12.22; peer-reviewed by V Horner, N Mejia, M Muhonde; comments to author 07.06.23; revised version received 28.07.23; accepted 26.01.24; published 18.03.24.

Please cite as:
Dang TTH, Carnahan E, Nguyen L, Mvundura M, Dao S, Duong TH, Nguyen T, Nguyen D, Nguyen T, Werner L, Ryman TK, Nguyen N
Outcomes and Costs of the Transition From a Paper-Based Immunization System to a Digital Immunization System in Vietnam: Mixed Methods Study
J Med Internet Res 2024;26:e45070
URL: https://www.jmir.org/2024/1/e45070
doi:10.2196/45070
PMID:38498020

©Thi Thanh Huyen Dang, Emily Carnahan, Linh Nguyen, Mercy Mvundura, Sang Dao, Thi Hong Duong, Trung Nguyen, Doan Nguyen, Tu Nguyen, Laurie Werner, Tove K Ryman, Nga Nguyen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Moving Biosurveillance Beyond Coded Data Using AI for Symptom Detection From Physician Notes: Retrospective Cohort Study

Andrew J McMurry[1,2], PhD; Amy R Zipursky[1,3], MD, MBI; Alon Geva[1,4,5], MD, MPH; Karen L Olson[1,2], PhD; James R Jones[1], MPhil; Vladimir Ignatov[1], MFA; Timothy A Miller[1,2], PhD; Kenneth D Mandl[1,2,6], MD, MPH

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

[2]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

[3]Division of Pediatric Emergency Medicine, Department of Pediatrics, The Hospital for Sick Children, Toronto, ON, Canada

[4]Division of Critical Care Medicine, Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital, Boston, MA, United States

[5]Department of Anaesthesia, Harvard Medical School, Boston, MA, United States

[6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Kenneth D Mandl, MD, MPH
Computational Health Informatics Program
Boston Children's Hospital
Landmark 5506 Mail Stop BCH3187, 401 Park Drive
Boston, MA, 02215
United States
Phone: 1 6173554145
Email: kenneth_mandl@harvard.edu

Abstract

Background: Real-time surveillance of emerging infectious diseases necessitates a dynamically evolving, computable case definition, which frequently incorporates symptom-related criteria. For symptom detection, both population health monitoring platforms and research initiatives primarily depend on structured data extracted from electronic health records.

Objective: This study sought to validate and test an artificial intelligence (AI)–based natural language processing (NLP) pipeline for detecting COVID-19 symptoms from physician notes in pediatric patients. We specifically study patients presenting to the emergency department (ED) who can be sentinel cases in an outbreak.

Methods: Subjects in this retrospective cohort study are patients who are 21 years of age and younger, who presented to a pediatric ED at a large academic children's hospital between March 1, 2020, and May 31, 2022. The ED notes for all patients were processed with an NLP pipeline tuned to detect the mention of 11 COVID-19 symptoms based on Centers for Disease Control and Prevention (CDC) criteria. For a gold standard, 3 subject matter experts labeled 226 ED notes and had strong agreement (F_1-score=0.986; positive predictive value [PPV]=0.972; and sensitivity=1.0). F_1-score, PPV, and sensitivity were used to compare the performance of both NLP and the *International Classification of Diseases, 10th Revision* (ICD-10) coding to the gold standard chart review. As a formative use case, variations in symptom patterns were measured across SARS-CoV-2 variant eras.

Results: There were 85,678 ED encounters during the study period, including 4% (n=3420) with patients with COVID-19. NLP was more accurate at identifying encounters with patients that had any of the COVID-19 symptoms (F_1-score=0.796) than ICD-10 codes (F_1-score =0.451). NLP accuracy was higher for positive symptoms (sensitivity=0.930) than ICD-10 (sensitivity=0.300). However, ICD-10 accuracy was higher for negative symptoms (specificity=0.994) than NLP (specificity=0.917). Congestion or runny nose showed the highest accuracy difference (NLP: F_1-score=0.828 and ICD-10: F_1-score=0.042). For encounters with patients with COVID-19, prevalence estimates of each NLP symptom differed across variant eras. Patients with COVID-19 were more likely to have each NLP symptom detected than patients without this disease. Effect sizes (odds ratios) varied across pandemic eras.

Conclusions: This study establishes the value of AI-based NLP as a highly effective tool for real-time COVID-19 symptom detection in pediatric patients, outperforming traditional ICD-10 methods. It also reveals the evolving nature of symptom prevalence across different virus variants, underscoring the need for dynamic, technology-driven approaches in infectious disease surveillance.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53367) doi:10.2196/53367

KEYWORDS

natural language processing; COVID-19; artificial intelligence; AI; public health, biosurveillance; surveillance; respiratory; infectious; pulmonary; SARS-CoV-2; symptom; symptoms; detect; detection; pipeline; pipelines; clinical note; clinical notes; documentation; emergency; urgent; pediatric; pediatrics; paediatric; paediatrics; child; children; youth; adolescent; adolescents; teen; teens; teenager; teenagers; diagnose; diagnosis; diagnostic; diagnostics

Introduction

Real-time emerging infection surveillance requires a case definition that often involves symptomatology. To detect symptoms, population health monitoring systems and research studies tend to largely rely on structured data from electronic health records, including the *International Classification of Diseases, 10th Revision* (ICD-10) codes [1]. However, symptoms are not diagnoses and, therefore, may not be consistently coded, leading to incorrect estimates of the prevalence of COVID-19 symptoms [2]. Natural language processing (NLP) of unstructured data from electronic health records has proven useful in recognizing COVID-19 symptoms and identifying additional signs and symptoms compared to structured data alone [3,4]. However, surveillance of COVID-19 symptoms is nuanced as symptoms have been shown to differ by variant eras [5,6] and by age, with pediatric patients generally experiencing milder symptoms [7]. For example, while loss of taste or smell was reported with early COVID-19 variants, it was less commonly reported during the Omicron wave and in younger patients who more frequently experience fever and cough [8-11]. Understanding symptom patterns in children during different COVID-19 variant eras is important. Early in the pandemic, the availability of molecular testing was extremely limited. The less severe course of infection and varying presentations may lead to under testing due to mild symptoms [12], potentially underestimating pediatric COVID-19 cases. Additionally, relatively asymptomatic children can still transmit the virus. Tailoring interventions based on age-specific manifestations contribute to effective control of virus transmission within communities.

We sought to validate and test an open-source artificial intelligence (AI)–based NLP pipeline that includes a large language model (LLM) to detect COVID-19 symptoms from physician notes. As a formative use case, we sought to illustrate how this pipeline could detect COVID-19 symptoms and differentiate symptom patterns across SARS-CoV-2 variant eras in pediatric patients. We specifically study patients presenting to the emergency department (ED) who can be sentinel cases in an outbreak.

Methods

Study Design and Setting

This was a retrospective cohort study of all patients up to 21 years of age presenting to the ED of a large, free-standing, university-affiliated, pediatric hospital between March 1, 2020, and May 31, 2022.

Ethical Considerations

The Boston Children's Hospital Committee on Clinical Investigation performed ethical, privacy, and confidentiality reviews of the study and found it to be exempt from human subjects oversight. A waiver of consent was obtained to cover the targeted extraction and secure review of clinical notes by approved study personnel in protected environments within the hospital firewall.

Study Variables

The main dependent variables were a set of 11 COVID-19 symptoms based on Centers for Disease Control and Prevention (CDC) criteria [13]—fever or chills, cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea. We identified these symptoms by both NLP and ICD-10 codes. For the formative use case, the study period was divided into 3 variant eras defined using Massachusetts COVID-19 data from Covariant [14]. The pre-Delta era was from March 1, 2020, to June 20, 2021; the Delta era was from June 21, 2021, to December 19, 2021; and the Omicron era was from December 20, 2021, onward. A diagnosis of COVID-19 was defined as a positive SARS-CoV-2 polymerase chain reaction (PCR) test or the presence of ICD-10 code U07.1 for COVID-19 during the same ED encounter in which symptoms were evaluated.

AI/NLP Pipeline Development

A total of 3 reviewers reached a consensus on a symptom concept dictionary [15] to capture each of the 11 COVID-19 symptoms. They relied on the Unified Medical Language System [16], which has a nearly comprehensive list of symptom descriptors [17], including SNOMED (SNOMED International) coded clinical terms [18], ICD-10 codes for administrative billing, abbreviations, and common language for patients [19]. The open-source and free Apache cTAKES (Apache Software Foundation) NLP pipeline was tuned to recognize and extract coded concepts for positive symptom mentions (based on the dictionary) from physician notes [20]. Apache cTAKES uses a NegEx algorithm which can help address negation [20-23]. To further address negation, we incorporated an LLM, Bidirectional Encoder Representations from Transformers, that was fine-tuned for negation classification on clinical text [24,25].

Gold Standard

A total of 2 reviewers established a gold standard by manually reviewing physician ED notes. After all notes were labeled by the cTAKES pipeline, a test set of 226 ED notes was loaded into Label Studio [26], an open-source application for ground truth labeling. These notes were from patients both with and without COVID-19 and were selected to ensure that each of the

XSL•FO
RenderX

11 symptoms was mentioned in at least 30 ED notes. Some notes mentioned more than 1 symptom. Using an annotation guide (Multimedia Appendix 1), 2 reviewers, who were masked from the terms identified by the NLP pipeline for note selection, each labeled 113 notes for mention of the 11 COVID-19 symptoms. As per the guide, only symptoms relevant to the present illness were considered positive mentions. Symptoms were not considered positive mentions if stated as past medical history, family history, social history, or an indication for a medication unrelated to the encounter.

Interrater Reliability

The F_1-score was used to assess consistency in manual chart review. The F_1-score is the balance of sensitivity and positive predictive value (PPV) [27]. It was computed by comparing the annotations of each of the 2 initial reviewers to those of a third reviewer, who independently labeled a subset (56/226, 25%) of notes annotated by the other reviewers. The choice of F_1-score as the metric for agreement was informed by the observed high frequency of true negative annotations when they were assigned by chance [20,27,28]. Reliability analyses used Python (version 3.10; Python Software Foundation).

AI/NLP and ICD-10 Accuracy

Accuracy measures of the true symptom percentages in the test set for each symptom included F_1-score, PPV, sensitivity, and specificity [29,30].

Formative Use Case

The impact of pandemic variant era on COVID-19 symptomatology was examined. Descriptive statistics were used to characterize patients presenting to the ED during each pandemic era. The percentage of patients in the ED with symptoms of COVID-19 was assessed in separate analyses for each symptom using chi-square analyses of 3×2 tables (pandemic era × symptom presence or absence) with α set at .05. Post hoc chi-square tests were used to compare each pandemic era with all others using a Bonferroni adjusted α of .017. To assess the effect of pandemic era, COVID-19 status, and the interaction of these variables on whether or not a patient had each symptom, logistic regression was used in separate analyses for each symptom. Bonferroni adjusted confidence limits were used for post hoc analyses. If the interaction term was not significant, the main effects of COVID-19 and variant era were reported. Data were analyzed using SAS (version 9.4; SAS Institute Inc).

Results

Study Population

There were 59,173 unique patients with 85,678 ED encounters during the study period. For each ED encounter, there was 1 final physician ED note that aggregated all ED physician documentation. Characteristics of the entire study cohort and variant-specific cohorts are summarized in Table 1. A patient could appear in the cohort more than once if they had multiple ED encounters.

Table 1. Characteristics of patients at emergency department encounters.

Characteristics	Total (n=85,678), n (%)	Pre-Delta (n=38,985), n (%)	Delta (n=24,432), n (%)	Omicron (n=22,261), n (%)
Age range (years)				
<5	36,835 (43.0)	15,403 (39.5)	11,749 (48.1)	9683 (43.5)
≥5	48,843 (57.0)	23,582 (60.5)	12,683 (51.9)	12,578 (56.5)
Sex				
Female	40,250 (47.0)	18,659 (47.9)	11,236 (46.0)	10,355 (46.5)
Male	45,428 (53.0)	20,326 (52.1)	13,196 (54.0)	11,906 (53.5)
Race				
American Indian	147 (0.2)	64 (0.2)	54 (0.2)	29 (0.1)
Asian	3244 (3.8)	1457 (3.7)	949 (3.9)	838 (3.8)
African American	13,354 (15.6)	6007 (15.4)	3943 (16.1)	3404 (15.3)
Pacific Islander	81 (0.1)	28 (0.1)	24 (0.1)	29 (0.1)
White	34,186 (39.9)	16,990 (43.6)	9093 (37.2)	8103 (36.4)
Not identified	34,666 (40.4)	14,439 (37.0)	10,369 (42.4)	9858 (44.2)
COVID-19 classification method				
COVID-19 diagnosis	3420 (4.0)	854 (2.2)	500 (2.0)	2066 (9.3)
PCR[a] positive	2167 (2.5)	518 (1.3)	294 (1.2)	1355 (6.1)
ICD-10[b] code	3305 (3.9)	820 (2.1)	458 (1.9)	2027 (9.1)

[a]PCR: polymerase chain reaction.

[b]ICD-10: International Classification of Diseases, 10th Revision.

XSL•FO
RenderX

Interrater Reliability

High consistency was demonstrated between reviewer 3, who labeled a subset of notes, and both reviewers 1 and 2, who each labeled half of the notes chosen to establish the gold standard. The F_1-scores for the 2 reviewers were 0.988 and 0.984, respectively. The PPV was 0.976 and 0.968 and sensitivity was 1.0 for both.

AI or NLP ICD-10 Accuracy

As shown in Table 2, the F_1-score for NLP was higher and thus more accurate at identifying encounters in the test set with patients that had any of the COVID-19 symptoms than ICD-10. NLP also had higher F_1-score for each individual symptom. In addition, NLP sensitivity of true positive symptoms was higher than ICD-10. However, NLP accuracy of true negative symptoms (specificity) was somewhat lower compared to ICD-10.

Table 2. Accuracy of COVID-19 symptom monitoring using NLP[a] and ICD-10[b] in the test set.

Symptom	F_1-score[c]		PPV[d]		Sensitivity		Specificity	
	NLP, n	ICD-10, n	NLP, n	ICD-10, n	NLP, n	ICD-10, n	NLP, n	ICD-10, n
Any COVID-19 symptom	0.796	0.451	0.696	0.906	0.930	0.300	0.917	0.994
Congestion or runny nose	0.828	0.042	0.788	1.000	0.872	0.021	0.938	1.000
Cough	0.914	0.541	0.841	0.952	1.000	0.377	0.942	0.994
Diarrhea	0.629	0.474	0.489	0.692	0.880	0.360	0.884	0.980
Fatigue	0.817	0.057	0.744	0.333	0.906	0.031	0.948	0.990
Fever or chills	0.864	0.700	0.768	0.977	0.987	0.545	0.844	0.993
Headache	0.744	0.566	0.667	1.000	0.842	0.395	0.914	1.000
Loss of taste or smell	0.667	0.167	0.500	1.000	1.000	0.091	0.948	1.000
Muscle or body aches	0.723	0.211	0.567	1.000	1.000	0.118	0.937	1.000
Nausea or vomiting	0.820	0.535	0.722	0.885	0.950	0.383	0.866	0.982
Shortness of breath or difficulty breathing	0.685	0.400	0.595	0.889	0.806	0.258	0.912	0.995
Sore throat	0.774	0.207	0.649	0.750	0.960	0.120	0.935	0.995

[a]NLP: natural language processing.

[b]ICD-10: International Classification of Diseases, 10th Revision.

[c]F_1-score: accuracy measure balancing PPV and sensitivity.

[d]PPV: positive predictive value.

The 2 most prevalent symptoms, cough and fever, had sensitivity scores for NLP that were among the highest of the symptoms, and much higher than those for ICD-10 codes. The greatest discrepancy between NLP and ICD-10 F_1-scores was for congestion or runny nose. The smallest difference was for diarrhea.

Formative Use Case

Prevalence of Symptoms Over Time

The percentage of ED encounters with patients with COVID-19 who had symptoms was estimated using the NLP pipeline and ICD-10 codes. As shown in Figure 1, during each month of the study, the percentage of encounters with no symptoms detected was much lower using NLP compared to ICD-10. Using NLP, the range was from 0% to 19% of encounters (mean 6%, SD 4%), while with ICD-10, the range was 22% to 52% (mean 38%, SD 7%).

The percentage of encounters with patients with COVID-19 who presented with each symptom by month was higher using NLP than ICD-10 (Multimedia Appendix 2). The 2 most common symptoms, cough and fever, are shown in Figures 2 and 3. On average, cough was identified during 52% (SD 13%) of the encounters each month using NLP, but only 15% (SD 5%) using ICD-10. On average, fever characterized 70% (SD 11%) of encounters using NLP, but 41% (SD 9%) using ICD-10.

XSL•FO
RenderX

Figure 1. The percentage of encounters with patients with COVID-19 presenting to the emergency department each month with no symptoms detected, as measured using NLP and ICD-10. ICD-10: International Classification of Diseases, 10th Revision; NLP: natural language processing.

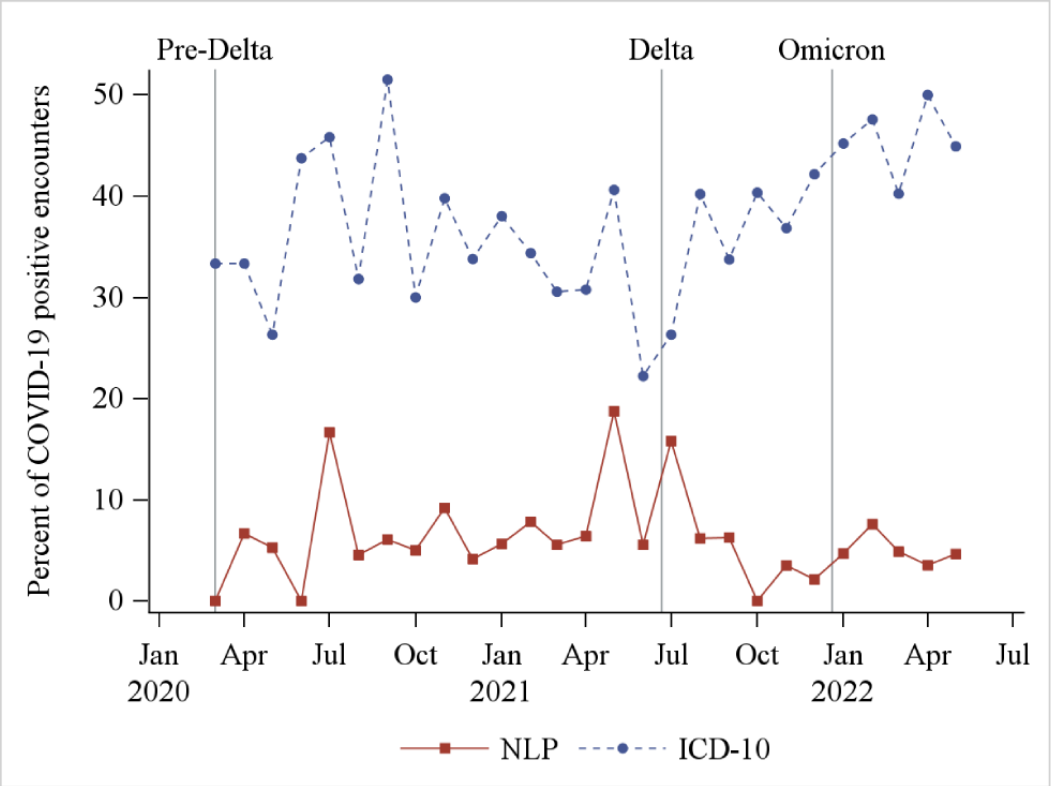

Figure 2. The percentage of encounters with patients with COVID-19 presenting to the emergency department each month with cough, as measured using NLP and ICD-10. ICD-10: International Classification of Diseases, 10th Revision; NLP: natural language processing.

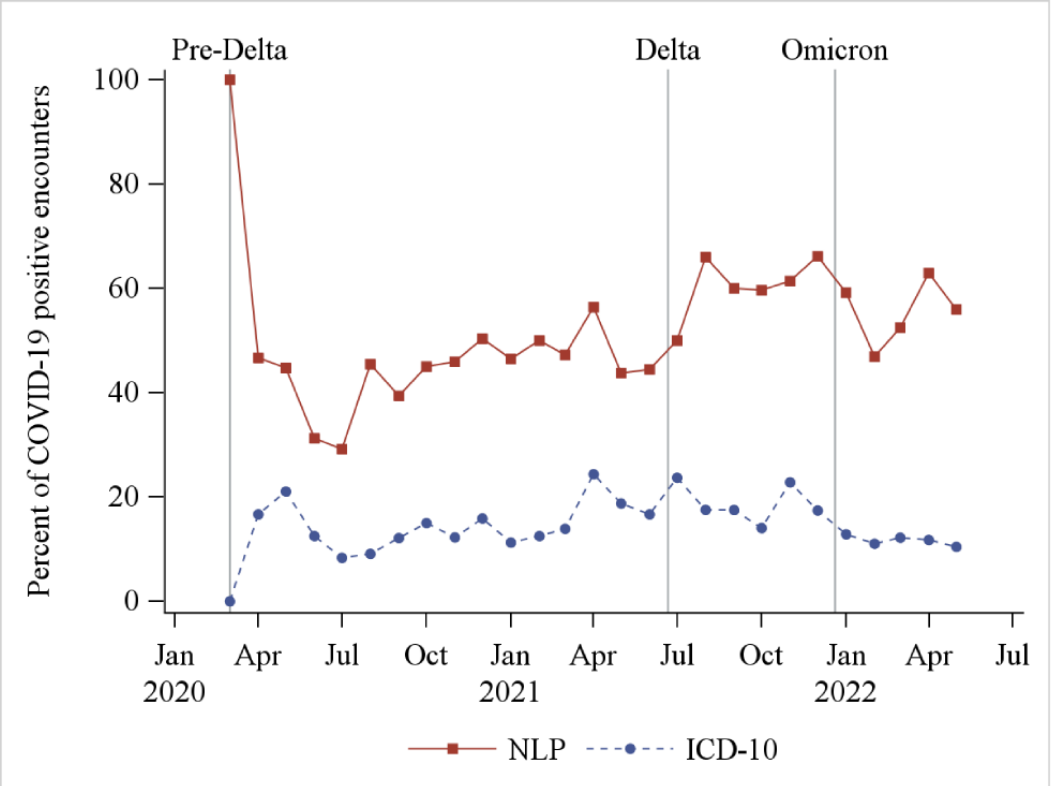

XSL•FO
RenderX

Figure 3. The percentage of encounters with patients with COVID-19 presenting to the emergency department each month with fever, as measured using NLP and ICD-10. ICD-10: International Classification of Diseases, 10th Revision; NLP: natural language processing.

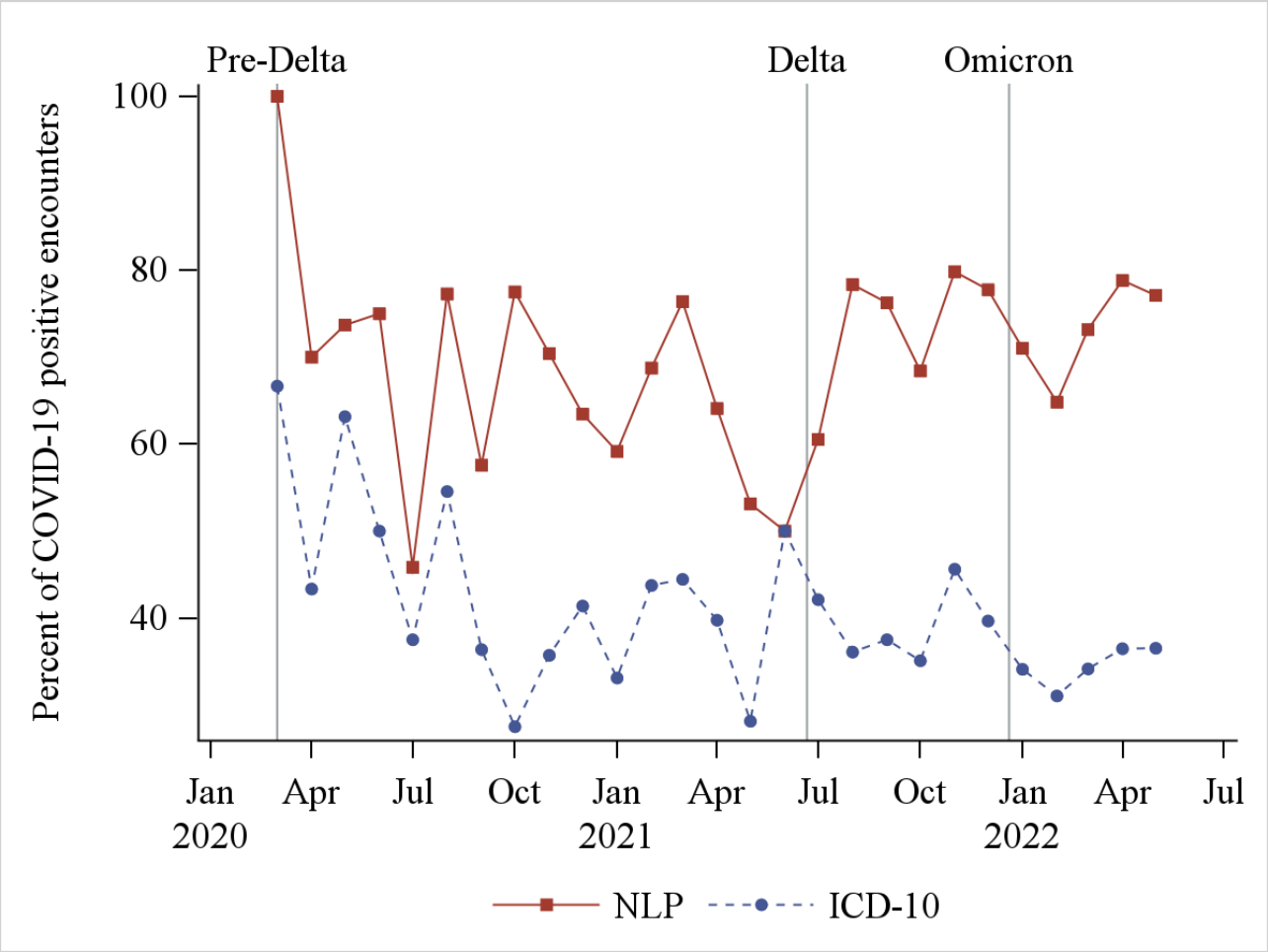

Using ICD-10, there were many months where individual symptoms were not detected. Of the 27 study months, loss of taste or smell was not detected using ICD-10 during 24 months, nor were muscle or body aches during 13 months. A total of 3 more symptoms had at least 3 consecutive months where each was not detected using ICD-10. These were congestion or runny nose (9 total months, not all consecutive), sore throat (8 months), and fatigue (7 months). Sporadic months without detection using ICD-10 were observed for headache (5 months), diarrhea (2 months), cough (1 month), and nausea or vomiting (1 month). Using NLP, sporadic months without detection were observed for just 2 symptoms, loss of taste or smell (6 months) and sore throat (2 months).

Prevalence of Symptoms Across Variant Eras

The prevalence estimates of symptoms across variant eras for encounters with patients with COVID-19 differed for each symptom identified by NLP, except for nausea or vomiting and sore throat (Table 3). Post hoc analyses revealed several patterns. New loss of taste or smell was the only symptom that varied across all 3 eras. It was most common in the pre-Delta era, followed by the Delta era, and then the Omicron era. Congestion or runny nose, cough, and fever or chills were more common during the Delta and Omicron era than during the pre-Delta era, but the Delta era did not differ from the Omicron era. Muscle or body aches were more common during the pre-Delta era than both the Delta and Omicron eras, but the Delta era did not differ from the Omicron era. Diarrhea, fatigue, headache, and shortness of breath were more common during the pre-Delta era than the Omicron era but were not different than the Delta era, and the Delta era did not differ from the Omicron era. Nausea or vomiting and sore throat did not differ by variant era. The chi-square results are in Multimedia Appendix 3.

XSL•FO
RenderX

Table 3. Prevalence estimates of symptoms using natural language processing by variant era for emergency department encounters with patients with COVID-19.

Symptom	Pre-Delta era (n=854), n (%)	Delta era (n=500), n (%)	Omicron era (n=2066), n (%)	*P* value
Congestion or runny nose	250 (29.3)[a]	186 (37.2)[b]	742 (35.9)[b]	.001
Cough	402 (47.1)[a]	309 (61.8)[b]	1223 (59.2)[b]	<.001
Diarrhea	188 (22.0)[a]	92 (18.4)[a,b]	317 (15.4)[b]	<.001
Fatigue	129 (15.1)[a]	72 (14.4)[a,b]	228 (11.0)[b]	.004
Fever or chills	561 (65.7)[a]	376 (75.2)[b]	1525 (73.8)[b]	<.001
Headache	185 (21.7)[a]	92 (18.4)[a,b]	301 (14.6)[b]	<.001
Muscle or body aches	110 (12.9)[a]	39 (7.8)[b]	164 (7.9)[b]	<.001
Nausea or vomiting	297 (34.8)	170 (34.0)	709 (34.3)	.95
New loss of taste or smell	57 (6.7)[a]	9 (1.8)[b]	9 (0.4)[c]	<.001
Shortness of breath or difficulty breathing	182 (21.3)[a]	84 (16.8)[a,b]	311 (15.1)[b]	<.001
Sore throat	125 (14.6)	83 (16.6)	319 (15.4)	.63

[a,b,c]Variant eras with the same superscript across a row did not differ in post hoc analyses.

Symptoms by COVID-19 Status and Variant Era

The interaction of COVID-19 status and variant era on the presence of each symptom is shown in Table 4. However, because the interaction was not significant for 2 symptoms, fever and chills, and sore throat, the main effects for COVID-19 status are shown for both (*P*<.001). The odds ratios (ORs) indicate that patients with COVID-19 were more likely to have each of these 2 symptoms than patients without this disease. These symptoms were also more likely to occur during the Delta and Omicron era than during the pre-Delta era. For the remaining symptoms, the interaction term was significant and the ORs in each variant era are shown in the table. The ORs comparing patients with COVID-19 to those without the disease differed among the variant eras. Several patterns were observed. Patients with COVID-19 were more likely to exhibit each of the symptoms of congestion or runny nose, cough, fatigue, headache, muscle or body aches, new loss of taste or smell, or shortness of breath or difficulty breathing. However, effect sizes (ORs) differed among pandemic eras. For diarrhea, this symptom was more likely for patients with COVID-19 in the pre-Delta and Delta eras, but not during the Omicron era. And nausea was more likely only in the pre-Delta era. Significant ORs ranged in size from 1.3 to 26.7 (mean 4.6, SD 5.3). The logistic regression results are in Multimedia Appendix 4.

XSL•FO
RenderX

Table 4. Effect of COVID-19 status and variant era on the presence of each symptom detected using natural language processing.

Symptom and pandemic variant era	Odds ratio[a] (95% CL[b])	Interaction[c] *P* value[d]
Congestion or runny nose		*<.001*
Pre-Delta	3.62 (3.11-4.21)	
Delta	2.27 (1.89-2.72)	
Omicron	2.46 (2.23-2.71)	
Cough		*<.001*
Pre-Delta	4.84 (4.22-5.55)	
Delta	3.64 (3.03-4.37)	
Omicron	3.54 (3.23-3.88)	
Diarrhea		*<.001*
Pre-Delta	2.23 (1.89-2.63)	
Delta	1.42 (1.13-1.79)	
Omicron	1.05 (0.92-1.19)	
Fatigue		*.01*
Pre-Delta	3.22 (2.65-3.90)	
Delta	3.42 (2.64-4.42)	
Omicron	2.36 (2.03-2.75)	
Fever or chills	4.82 (4.46-5.21)	.66
Headache		*<.001*
Pre-Delta	2.33 (1.98-2.76)	
Delta	2.09 (1.66-2.63)	
Omicron	1.52 (1.33-1.73)	
Muscle or body aches		*.006*
Pre-Delta	5.96 (4.83-7.36)	
Delta	4.75 (3.38-6.67)	
Omicron	3.78 (3.14-4.55)	
Nausea or vomiting		*.006*
Pre-Delta	1.30 (1.13-1.50)	
Delta	1.03 (0.86-1.25)	
Omicron	0.98 (0.89-1.08)	
New loss of taste or smell		*.049*
Pre-Delta	26.66 (19.13-37.14)	
Delta	11.83 (5.68-24.65)	
Omicron	11.04 (4.25-28.64)	
Shortness of breath or difficulty breathing		*<.001*
Pre-Delta	2.62 (2.22-3.10)	
Delta	1.70 (1.34-2.16)	
Omicron	1.57 (1.38-1.79)	
Sore throat	2.45 (2.22-2.70)	.27

[a]Odds ratios compare patients with COVID-19 at an ED encounter to patients without the disease.

[b]CL: Bonferroni adjusted confidence limits in post hoc analyses.

[c]If the interaction term was significant, the effect of COVID-19 during each variant era is shown. Otherwise, the effect for COVID-19 is shown.

[d]Type 3 test of the interaction term (variant era × COVID-19) in a logistic regression analysis.

XSL•FO
RenderX

Discussion

Principal Findings

We find evidence that AI-based NLP of physician notes is a superior method for capturing patient symptoms for real-time biosurveillance than reliance on traditional approaches using ICD-10. NLP was more sensitive than ICD-10 codes in identifying symptoms and some symptoms could only be detected using NLP. As a form of internal validation, the symptoms identified by the CDC as associated with COVID-19 were more common in patients with than without this disease.

Comparison With Prior Work

The study was also able to capture a nuanced picture of symptom prevalence and odds across different SARS-CoV-2 variant eras. Consistent with previous literature, symptom patterns changed over time as new variants emerged. Variants may present with differences in symptomatology as a result of a number of factors including differences in mutations in spike proteins, receptor binding, and ability to escape host antibodies [31]. As has been previously reported [11,32-35], we found that fever or chills were the most common COVID-19 symptom across the variants. In our cohort, shortness of breath was less common during the Omicron era than during the pre-Delta era. The Omicron variant has less of an ability to replicate in the lungs compared to the bronchi, which may explain why this symptom became less common [36]. Studies have reported sore throat as a common symptom in the Omicron era, but we did not observe a significant difference across eras [8,9]. It is possible that we did not see a higher percentage of sore throats in the Omicron era because it may be more challenging for pediatric patients to describe this symptom. One study found that sore throat was observed more often in those of 5-20 years of age compared to those of 0-4 years of age [8]. Similarly, a study reported that sore throat was more common in those greater than or equal to 13 years of age in the Omicron era compared to the Delta era [37]. In our study cohort, approximately half of the patients were younger than 5 years of age. As children this age may not be able to describe their symptoms well, symptoms that are also signs, such as fever or cough, might be more commonly documented in physician notes than symptoms such as sore throat. New loss of taste or smell was most common in the pre-Delta era, followed by the Delta era and then the Omicron era in this study. This symptom has been reported less commonly in the Omicron era [8,9]. Studies have postulated that patients with the Omicron variant are less likely to present with loss of taste or smell as this variant has less penetration of the mucus layer and therefore, may be less likely to infect the olfactory epithelium [38].

Limitations

There were important limitations in our use of NLP. The NLP pipeline was tested with a set of notes where some symptoms were more frequent in the test set (eg, loss of taste or smell) than in the formative use case. This was done to have sufficient data to evaluate the symptom pipeline. The NLP pipeline does not account for vital signs and so fever may not have been detected with the pipeline if it was documented in a patient's vital signs rather than the clinical text. The cTAKES tool in the pipeline lacks the temporal context to ascertain if the mention of a symptom in a note is a new symptom or a prior symptom. We modified our technique because of this but nevertheless may have overestimated the prevalence of symptoms in our study. Future work will involve filtering by note section so that certain components of a note like past medical history are not included. We used 2 techniques to recognize negation, but some negated symptoms (eg, "patient had no cough") were still captured as positive symptom mentions leading to a possible overestimation of symptom prevalence. Finally, this NLP pipeline did involve substantial preprocessing. We plan to evaluate the implementation of Generative Pre-trained Transformer (GPT) for this task. GPT-4 was able to extract COVID-19 symptoms in a recent study [39] and it may limit the need for preprocessing.

Our formative study had some limitations. First, we examined COVID-19 symptoms in patients presenting to a single urban pediatric ED. Patients presenting to outpatient settings, who likely had milder symptoms, were not included and our results may reflect patients with more severe symptoms. And because the setting was a single site, results may not generalize to other EDs. Second, we defined COVID-19 status as positive if a patient had a PCR positive test for COVID-19 or an appropriate ICD-10 code at the ED encounter. Patients who were COVID-19 positive on a test at home or at an outside center may not have been captured by this definition even if they presented to the ED with COVID-19 [40]. Additionally, symptoms may have differed across variant eras as a result of COVID-19 vaccinations or previous infections rather than variant differences. Literature in adults shows that vaccination is associated with a decrease in systemic symptoms [41]. The United States Food and Drug Administration authorized the use of the COVID-19 vaccine in October 2021, during the Delta era and prior to the Omicron era, for children 5-11 years of age [42]. Vaccination rates for pediatric patients vary by age group in Massachusetts, as of April 3, 2023, of those 0-19 years of age, 3% to 57% have received a primary series but have not been boosted, and 3% to 18% have been boosted since September 1, 2022 [43]. As such, some patients in the Delta and Omicron eras may have been vaccinated or had previous COVID-19 infections [44].

Conclusions

In an era where rapid and accurate infectious disease surveillance is crucial, this study underscores the transformative potential of AI-based NLP for real-time symptom detection, significantly outperforming traditional methods such as ICD-10 coding. The dynamic adaptability of NLP technology allows for the nuanced capture of evolving symptomatology across different virus variants, offering a more responsive and precise tool kit for biosurveillance efforts. Its integration into existing health care infrastructure could be a game changer, elevating our capabilities to monitor, understand, and ultimately control the spread of emerging infectious diseases.

XSL•FO
RenderX

Acknowledgments

This study was supported by the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services (HHS) as part of a financial assistance award. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by CDC, HHS, or the US Government. Support was also obtained from the National Center for Advancing Translational Sciences, National Institutes of Health Cooperative Agreement (U01TR002623). ARZ was supported by a training grant from the National Institute of Child Health and Human Development (T32HD040128). Generative artificial intelligence (AI) was not used to design or conduct this study.

Data Availability

All data analyzed during this study for the formative use case are in Multimedia Appendix 5 of this published article.

Authors' Contributions

KDM, AJM, and TAM contributed to the conceptualization. KDM contributed to the funding. AJM, ARZ, AG, and KLO performed the formal analysis. AJM, JRJ, and VI contributed to the software. AJM, ARZ, and KDM contributed to writing original drafts. KLO and AG contributed to writing review and edits.

Conflicts of Interest

TAM is a member of the advisory council for Lavita AI. Others declare no conflicts of interest.

Multimedia Appendix 1
COVID-19 symptoms annotation guide.
[PDF File (Adobe PDF File), 225 KB - jmir_v26i1e53367_app1.pdf]

Multimedia Appendix 2
Detection of COVID-19 symptoms using NLP and ICD-10 by month for emergency department encounters with patients with COVID-19. ICD-10: International Classification of Diseases, 10th Revision; NLP: natural language processing.
[PDF File (Adobe PDF File), 2060 KB - jmir_v26i1e53367_app2.pdf]

Multimedia Appendix 3
The chi-square analysis of COVID-19 symptom prevalence by pandemic variant era for emergency department encounters with patients with COVID-19, symptoms were detected using NLP. NLP: natural language processing.
[XLSX File (Microsoft Excel File), 17 KB - jmir_v26i1e53367_app3.xlsx]

Multimedia Appendix 4
Logistic regression analysis of the effect of COVID-19 status, pandemic variant era, and their interaction on symptom status for ED encounters, symptoms were detected using NLP. ED: emergency department; NLP: natural language processing.
[XLSX File (Microsoft Excel File), 23 KB - jmir_v26i1e53367_app4.xlsx]

Multimedia Appendix 5
Data files for the time series figures, the chi-square analysis of symptom prevalence, and the logistic regression analysis of the effects of COVID-19 status and pandemic variant era on symptom status.
[XLSX File (Microsoft Excel File), 40 KB - jmir_v26i1e53367_app5.xlsx]

References

1. Subramanian A, Nirantharakumar K, Hughes S, Myles P, Williams T, Gokhale KM, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. Nat Med 2022;28(8):1706-1714 [FREE Full text] [doi: 10.1038/s41591-022-01909-w] [Medline: 35879616]
2. Crabb BT, Lyons A, Bale M, Martin V, Berger B, Mann S, et al. Comparison of International Classification of Diseases and Related Health Problems, Tenth Revision codes with electronic medical records among patients with symptoms of coronavirus disease 2019. JAMA Netw Open 2020;3(8):e2017703 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.17703] [Medline: 32797176]
3. Wang J, Abu-El-Rub N, Gray J, Pham HA, Zhou Y, Manion FJ, et al. COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model. J Am Med Inform Assoc 2021;28(6):1275-1283 [FREE Full text] [doi: 10.1093/jamia/ocab015] [Medline: 33674830]

XSL•FO
RenderX

4. Malden DE, Tartof SY, Ackerson BK, Hong V, Skarbinski J, Yau V, et al. Natural language processing for improved characterization of COVID-19 symptoms: observational study of 350,000 patients in a large integrated health care system. JMIR Public Health Surveill 2022;8(12):e41529 [FREE Full text] [doi: 10.2196/41529] [Medline: 36446133]
5. Di Chiara C, Boracchini R, Sturniolo G, Barbieri A, Costenaro P, Cozzani S, et al. Clinical features of COVID-19 in Italian outpatient children and adolescents during parental, Delta, and Omicron waves: a prospective, observational, cohort study. Front Pediatr 2023;11:1193857 [FREE Full text] [doi: 10.3389/fped.2023.1193857] [Medline: 37635788]
6. Sumner MW, Xie J, Zemek R, Winston K, Freire G, Burstein B, et al. Comparison of symptoms associated with SARS-CoV-2 variants among children in Canada. JAMA Netw Open 2023;6(3):e232328 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.2328] [Medline: 36892839]
7. Liguoro I, Pilotto C, Bonanni M, Ferrari ME, Pusiol A, Nocerino A, et al. SARS-COV-2 infection in children and newborns: a systematic review. Eur J Pediatr 2020;179(7):1029-1046 [FREE Full text] [doi: 10.1007/s00431-020-03684-7] [Medline: 32424745]
8. Menni C, Valdes AM, Polidori L, Antonelli M, Penamakuri S, Nogal A, et al. Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of Omicron and Delta variant dominance: a prospective observational study from the ZOE COVID Study. Lancet 2022;399(10335):1618-1624 [FREE Full text] [doi: 10.1016/S0140-6736(22)00327-0] [Medline: 35397851]
9. Akaishi T, Kushimoto S, Katori Y, Sugawara N, Egusa H, Igarashi K, et al. COVID-19-related symptoms during the SARS-CoV-2 Omicron (B.1.1.529) variant surge in Japan. Tohoku J Exp Med 2022;258(2):103-110 [FREE Full text] [doi: 10.1620/tjem.2022.J067] [Medline: 36002251]
10. García-Vera C, Castejón-Ramírez S, Miranda EL, Abadía RH, Ventura MG, Navarro EB, et al. COVID-19 in children: clinical and epidemiological spectrum in the community. Eur J Pediatr 2022;181(3):1235-1242 [FREE Full text] [doi: 10.1007/s00431-021-04235-4] [Medline: 34406504]
11. Viner RM, Ward JL, Hudson LD, Ashe M, Patel SV, Hargreaves D, et al. Systematic review of reviews of symptoms and signs of COVID-19 in children and adolescents. Arch Dis Child 2021;106:802-807 [FREE Full text] [doi: 10.1136/archdischild-2020-320972] [Medline: 33334728]
12. COVID-19 disease in children and adolescents: scientific brief, 29 September 2021. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Children_and_adolescents-2021.1?ssp=1&setlang=en&cc=US [accessed 2024-02-28]
13. Symptoms of COVID-19. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html [accessed 2024-02-28]
14. Hodcroft E. CoVariants. CoVariants. 2021. URL: https://covariants.org/ [accessed 2024-02-28]
15. Machine-learning-for-medical-language / ctakes-client-py. Github. URL: https://github.com/Machine-Learning-for-Medical-Language/ctakes-client-py/blob/main/ctakesclient/resources/covid_symptoms.bsv [accessed 2024-02-28]
16. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
17. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res 2021;49(D1):D1207-D1217 [FREE Full text] [doi: 10.1093/nar/gkaa1043] [Medline: 33264411]
18. SNOMEDCT_US (SNOMED CT, US edition)—synopsis, UMLS vocabularies. Unified Medical Language System (UMLS). URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT_US/index.html [accessed 2024-02-28]
19. CHV (Consumer Health Vocabulary)—synopsis, UMLS vocabularies. Unified Medical Language System (UMLS). URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html [accessed 2024-02-28]
20. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
21. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
22. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009;42(5):839-851 [FREE Full text] [doi: 10.1016/j.jbi.2009.05.002] [Medline: 19435614]
23. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 2013;192:677-681 [FREE Full text] [Medline: 23920642]
24. Machine-learning-for-medical-language. GitHub. URL: https://github.com/Machine-Learning-for-Medical-Language [accessed 2024-02-28]
25. Miller T, Bethard S, Amiri H, Savova G. Unsupervised domain adaptation for clinical negation detection. 2017 Presented at: BioNLP; August 4, 2017; Vancouver, Canada p. 165-170 URL: https://aclanthology.org/W17-2320/ [doi: 10.18653/v1/w17-2320]

XSL•FO
RenderX

26. Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. Label studio: data labeling software. GitHub. 2020. URL: https://github.com/heartexlabs/label-studio [accessed 2024-02-28]
27. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296-298 [FREE Full text] [doi: 10.1197/jamia.M1733] [Medline: 15684123]
28. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
29. Habibzadeh F, Habibzadeh P, Yadollahie M. The apparent prevalence, the true prevalence. Biochem Med (Zagreb) 2022;32(2):020101 [FREE Full text] [doi: 10.11613/BM.2022.020101] [Medline: 35799992]
30. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (Kaunas) 2021;57(5):503 [FREE Full text] [doi: 10.3390/medicina57050503] [Medline: 34065637]
31. Lauring AS, Hodcroft EB. Genetic variants of SARS-CoV-2-what do they mean? JAMA 2021;325(6):529-531 [FREE Full text] [doi: 10.1001/jama.2020.27124] [Medline: 33404586]
32. Götzinger F, Santiago-García B, Noguera-Julián A, Lanaspa M, Lancella L, Carducci FIC, et al. COVID-19 in children and adolescents in Europe: a multinational, multicentre cohort study. Lancet Child Adolesc Health 2020;4(9):653-661 [FREE Full text] [doi: 10.1016/S2352-4642(20)30177-2] [Medline: 32593339]
33. King JA, Whitten TA, Bakal JA, McAlister FA. Symptoms associated with a positive result for a swab for SARS-CoV-2 infection among children in Alberta. CMAJ 2021;193(1):E1-E9 [FREE Full text] [doi: 10.1503/cmaj.202065] [Medline: 33234533]
34. Takács AT, Bukva M, Gavallér G, Kapus K, Rózsa M, Bán-Gagyi B, et al. Epidemiology and clinical features of SARS-CoV-2 infection in hospitalized children across four waves in Hungary: a retrospective, comparative study from march 2020 to december 2021. Health Sci Rep 2022;5(6):e937 [FREE Full text] [doi: 10.1002/hsr2.937] [Medline: 36425898]
35. Kenney PO, Chang AJ, Krabill L, Hicar MD. Decreased clinical severity of pediatric acute COVID-19 and MIS-C and increase of incidental cases during the Omicron wave in comparison to the Delta wave. Viruses 2023;15(1):180 [FREE Full text] [doi: 10.3390/v15010180] [Medline: 36680220]
36. Hui KPY, Ho JCW, Cheung MC, Ng KC, Ching RHH, Lai KL, et al. SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. Nature 2022;603(7902):715-720 [FREE Full text] [doi: 10.1038/s41586-022-04479-6] [Medline: 35104836]
37. Shoji K, Akiyama T, Tsuzuki S, Matsunaga N, Asai Y, Suzuki S, et al. Clinical characteristics of COVID-19 in hospitalized children during the Omicron variant predominant period. J Infect Chemother 2022;28(11):1531-1535 [FREE Full text] [doi: 10.1016/j.jiac.2022.08.004] [Medline: 35963599]
38. Butowt R, Bilińska K, von Bartheld C. Why does the Omicron variant largely spare olfactory function? implications for the pathogenesis of anosmia in coronavirus disease 2019. J Infect Dis 2022;226(8):1304-1308 [FREE Full text] [doi: 10.1093/infdis/jiac113] [Medline: 35467743]
39. Wei WI, Leung CLK, Tang A, McNeil EB, Wong SYS, Kwok KO. Extracting symptoms from free-text responses using ChatGPT among COVID-19 cases in Hong Kong. Clin Microbiol Infect 2024;30(1):142.e1-142.e3 [FREE Full text] [doi: 10.1016/j.cmi.2023.11.002] [Medline: 37949111]
40. Wang L, Zipursky AR, Geva A, McMurry AJ, Mandl KD, Miller TA. A computable case definition for patients with SARS-CoV2 testing that occurred outside the hospital. JAMIA Open 2023;6(3):ooad047 [FREE Full text] [doi: 10.1093/jamiaopen/ooad047] [Medline: 37425487]
41. Bramante CT, Proper JL, Boulware DR, Karger AB, Murray T, Rao V, et al. Vaccination against SARS-CoV-2 is associated with a lower viral load and likelihood of systemic symptoms. Open Forum Infect Dis 2022;9(5):ofac066 [FREE Full text] [doi: 10.1093/ofid/ofac066] [Medline: 35392460]
42. FDA authorizes Pfizer-BioNTech COVID-19 vaccine for emergency use in children 5 through 11 years of age. U.S. Food and Drug Administration. 2021. URL: https://www.fda.gov/news-events/press-announcements/fda-authorizes-pfizer-biontech-covid-19-vaccine-emergency-use-children-5-through-11-years-age [accessed 2024-02-28]
43. Weekly COVID-19 vaccination report (as of April 3, 2023). Massachusetts Department of Public Health. URL: https://www.mass.gov/doc/weekly-covid-19-vaccination-report-april-5-2023/download [accessed 2024-02-28]
44. Bhattacharyya RP, Hanage WP. Challenges in inferring intrinsic severity of the SARS-CoV-2 Omicron variant. N Engl J Med 2022;386(7):e14 [FREE Full text] [doi: 10.1056/NEJMp2119682] [Medline: 35108465]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
ED: emergency department
GPT: Generative Pre-trained Transformer
ICD-10: International Classification of Diseases, 10th Revision
LLM: large language model

XSL•FO
RenderX

NLP: natural language processing
OR: odds ratio
PCR: polymerase chain reaction
PPV: positive predictive value

Edited by T de Azevedo Cardoso; submitted 06.10.23; peer-reviewed by D Liebovitz; comments to author 09.11.23; revised version received 30.11.23; accepted 27.02.24; published 04.04.24.

Please cite as:
McMurry AJ, Zipursky AR, Geva A, Olson KL, Jones JR, Ignatov V, Miller TA, Mandl KD
Moving Biosurveillance Beyond Coded Data Using AI for Symptom Detection From Physician Notes: Retrospective Cohort Study
J Med Internet Res 2024;26:e53367
URL: https://www.jmir.org/2024/1/e53367
doi:10.2196/53367
PMID:38573752

©Andrew J McMurry, Amy R Zipursky, Alon Geva, Karen L Olson, James R Jones, Vladimir Ignatov, Timothy A Miller, Kenneth D Mandl. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Twitter Analysis of Health Care Workers' Sentiment and Discourse Regarding Post–COVID-19 Condition in Children and Young People: Mixed Methods Study

Macarena Chepo[1*], RN, BSN, MPH, PhD; Sam Martin[2,3*], MSc, PhD; Noémie Déom[2], MSc; Ahmad Firas Khalid[4], MD, PhD; Cecilia Vindrola-Padros[2], BA, MA, PhD

[1]School of Nursing, Universidad Andrés Bello, Santiago, Chile

[2]Department of Targeted Intervention, University College London, London, United Kingdom

[3]Oxford Vaccine Group, Churchill Hospital, University of Oxford, Oxford, United Kingdom

[4]Canadian Institutes of Health Research Health System Impact Fellowship, Centre for Implementation Research, Ottawa Hospital Research Institute, Otawa, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Sam Martin, MSc, PhD
Department of Targeted Intervention
University College London
Charles Bell House 43-45
Foley Street
London, W1W 7TY
United Kingdom
Phone: 44 (0)20 3108 3232
Email: sam.martin@ucl.ac.uk

Abstract

Background: The COVID-19 pandemic has had a significant global impact, with millions of cases and deaths. Research highlights the persistence of symptoms over time (post–COVID-19 condition), a situation of particular concern in children and young people with symptoms. Social media such as Twitter (subsequently rebranded as X) could provide valuable information on the impact of the post–COVID-19 condition on this demographic.

Objective: With a social media analysis of the discourse surrounding the prevalence of post–COVID-19 condition in children and young people, we aimed to explore the perceptions of health care workers (HCWs) concerning post–COVID-19 condition in children and young people in the United Kingdom between January 2021 and January 2022. This will allow us to contribute to the emerging knowledge on post–COVID-19 condition and identify critical areas and future directions for researchers and policy makers.

Methods: From a pragmatic paradigm, we used a mixed methods approach. Through discourse, keyword, sentiment, and image analyses, using Pulsar and InfraNodus, we analyzed the discourse about the experience of post–COVID-19 condition in children and young people in the United Kingdom shared on Twitter between January 1, 2021, and January 31, 2022, from a sample of HCWs with Twitter accounts whose biography identifies them as HCWs.

Results: We obtained 300,000 tweets, out of which (after filtering for relevant tweets) we performed an in-depth qualitative sample analysis of 2588 tweets. The HCWs were responsive to announcements issued by the authorities regarding the management of the COVID-19 pandemic in the United Kingdom. The most frequent sentiment expressed was negative. The main themes were uncertainty about the future, policies and regulations, managing and addressing the COVID-19 pandemic and post–COVID-19 condition in children and young people, vaccination, using Twitter to share scientific literature and management strategies, and clinical and personal experiences.

Conclusions: The perceptions described on Twitter by HCWs concerning the presence of the post–COVID-19 condition in children and young people appear to be a relevant and timely issue and responsive to the declarations and guidelines issued by health authorities over time. We recommend further support and training strategies for health workers and school staff regarding the manifestations and treatment of children and young people with post–COVID-19 condition.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50139) doi:10.2196/50139

KEYWORDS

COVID-19; postacute sequelae of SARS-CoV-2; PASC; post–COVID-19 condition; children; vaccines; social media; social network analysis; Twitter

Introduction

Background

More than 3 years after the outbreak of the COVID-19 pandemic [1], the social, political, and economic impact of this phenomenon has been more than significant, considering >700 million worldwide cases and nearly 7 million people's deaths [2]. Given the scale of the phenomenon, it is imperative for all countries to thoroughly examine the lessons gleaned from the pandemic, particularly regarding a matter that has raised significant concern among the populace: the long-term effects experienced by individuals who have had COVID-19, spanning weeks, months, or even years after their initial infection [3]. This phenomenon, referred to as post–COVID-19 condition (or more commonly "long COVID"), warrants careful consideration and analysis [4].

There is increasing information regarding the clinical manifestation of this condition, particularly in the adult population. The worldwide prevalence has been estimated at approximately 50% to 70% in individuals hospitalized during acute COVID-19 infection and 10% to 12% in vaccinated cases [5]. While children and young people have a low likelihood of severe COVID-19 infection [6], the information available to date indicates that the presence of post–COVID-19 condition in this group may be as disabling as in adults, reaching a prevalence rate of 23.4% (range 3.7%-66.5%) [7].

An agreed definition by the World Health Organization indicates that post–COVID-19 condition in children and young people is a condition that occurs "in individuals with a history of confirmed or probable SARS-CoV-2 infection when experiencing symptoms lasting at least two months which initially occurred within three months of acute COVID-19" [8]. Post–COVID-19 condition strongly impacts daily functioning and can develop or continue after COVID-19 infection and may fluctuate or relapse over time [4,8,9].

Among the symptoms most frequently attributable to post–COVID-19 condition in children and young people are fatigue, altered smell or anosmia, and anxiety [8]. However, other symptoms have also been reported, such as sleep disturbances, difficulty in concentrating, abdominal pain, myalgia or arthralgia, earache or ringing in ears, mood swings, persistent chest pain, stomach pain, light sensitivity, diarrhea, heart palpitations, and skin lesions [8,10]. One of England's most significant studies is the Children and Young People With Long COVID study by Stephenson et al [11]. This national research matched longitudinal and cohort studies in adolescent individuals aged 11 to 17 years and found the presence of symptoms in 35.4% of the adolescent individuals who tested positive at baseline and 8.3% who of the adolescent individuals who tested negative at baseline. A total of 3 months after testing, 66.5% of those who tested positive and 53.3% of those who tested negative had any symptoms [11]. However, Stephenson et al [12] recently indicated that in a 6-month follow-up, the prevalence of specific symptoms reported at the time of the polymerase chain reaction testing decreased over time, where, for example, the prevalence of chills, fever, myalgia, cough, and sore throat among those who tested positive decreased from 10% to 25% to <3%.

As research on the symptoms, prevalence, and treatment of post–COVID-19 condition in children and young people continues, it is essential to add to the literature by developing studies that determine the condition's impact on this group, considering that they are experiencing a range of unwanted symptoms that disrupt their quality of life and that of their families.

Considering that listening to the voices of families and health workers could be helpful to broaden the knowledge achieved in post–COVID-19 condition in children and young people, a powerful tool could be social media, such as Twitter (subsequently rebranded as X). With >3729 million daily active users, Twitter has become one of the most important social platforms in the world [13]. People used Twitter during the COVID-19 pandemic for different purposes, such as world leaders communicating with citizens [14,15], organizations monitoring movement [16], scientists studying public discourse around the pandemic [17,18], and researchers performing sentiment analysis [19-21]. In the case of physicians and health care workers (HCWs), Twitter has been used to share and evaluate scientific evidence, guidelines, and technical advice [22-24] and track the course and burden of disease [25].

Objectives

Using the social media monitoring platform Pulsar [26], we aimed to explore HCWs' perceptions concerning post–COVID condition in children and young people in the United Kingdom between January 2021 and January 2022. We aimed to contribute to the emerging knowledge on post–COVID-19 condition in children and young people and identify critical areas and future directions for researchers and policy makers.

Methods

Overview

We considered a mixed methods approach to be a pragmatic research paradigm. We analyzed data by conducting a Collaborative and Digital Analysis of Big Qualitative Data in Time Sensitive Contexts (LISTEN) [27]. This mixed methods analysis consisted of iterative cycles intercalating team discussion and using digital text and discourse analytics tools to analyze related social media data [27]. We used the LISTEN method to perform quantitative and qualitative analyses of Twitter posts, extracted through the Pulsar platform [26], related to the experience of post–COVID-19 condition in children and

XSL•FO
RenderX

young people in the United Kingdom (eg, phrases, words, hashtags, videos, and images), published between January 1, 2021, and January 31, 2022. We created an advanced Boolean search for keywords mentioning "long COVID" and corelated words, hashtags, and symptoms; furthermore, we filtered for user accounts who identified as HCWs in their Twitter biography description (Multimedia Appendix 1).

Quantitative analysis of all tweets included the following: (1) engagement analysis, where we specifically measured reactions to posts, for example, a retweet, a share, or a comment or quote made toward a tweet; (2) sentiment and emotion analysis, where we measured the positive or negative sentiment in the words and tone of each post within the context of post–COVID-19 condition and HCW's roles (Multimedia Appendix 2); (3) emotion analysis, where we measured the emotions expressed in the tweets, classified as sadness, anger, disgust, fear, and joy; (4) frequency analysis, where we observed the frequency of keywords and themes in the data set; (5) segmentation analysis, where we measured the key connections or relationships between keywords and their frequent use in the same context; (6) demographic analysis, where we measured the occupation, gender (man or woman or nonbinary or unknown), and city of origin related to the users posting tweets; and (7) analyses, where we evaluated the most influential accounts and the most mentioned websites.

Big qualitative analysis was carried out through thematic discourse analysis of the data sample, using InfraNodus [28], specifically analyzing the key themes and topics of concern expressed throughout the data set. A codebook was constructed based on the mapping of themes agreed upon by 3 researchers (ND, SM, and MC; Multimedia Appendix 3).

The principal investigators (ND, AFK, SM, and MC) interpreted and analyzed the data collected, following the recommendations for rigorous research provided by Creswell and Poth [29]. Using the LISTEN method [27], we aimed to show that the integration of qualitative insights through thematic analysis with the quantitative backing of topic modeling can offer a comprehensive view of the discourse. This mixed methods approach allows us to capture the richness of qualitative data while leveraging the objectivity of quantitative measures. Our initial data harvest of the larger corpus data from the Pulsar platform captured 300,000 tweets; this data harvest helped to underpin the software's sentiment analysis modeling of this specific data set, providing a robust quantitative foundation. The addition of further qualitative data analyses from a smaller qualitative sample allowed for an in-depth understanding of nuanced conversations, particularly when exploring new or complex phenomena such as post–COVID-19 condition in children and young people, with the provision of insights into the context, subtext, and sentiment behind the tweets offering valuable snapshots of public perception and discourse. We used an iterative mixed methods approach, iterating between team discussions and using digital analytics tools to discern relevant themes from the Twitter data corpus. Specifically, we used InfraNodus for thematic analysis, which incorporates a topic modeling script for analyzing and identifying key topics of concern with a data set and provides a structured and objective interpretation of the data. The coding process involved 3 independent researchers (MC, SM, and ND), each with expertise in health care, social network analysis, and digital global health. When initial coding disagreements arose, we meticulously tagged any queries and discussed the posts in question. These instances led to 3 structured meetings wherein the research team deliberated collaboratively to resolve conflicting interpretations. This approach resulted in an 81.99% (2122/2588) initial intercoder agreement rate for the tweets analyzed. For the remaining instances where consensus was not initially reached, the majority rule was applied to finalize theme codings. To quantify the reliability of our coding procedure, with 81.99% (2122/2588) of the tweets coded identically, we used the Cohen κ score, which provides a measure of interrater agreement adjusted for chance. Including the calculation of all variations, this score was calculated to be approximately κ=0.70, indicating good agreement among the coders.

Ethical Considerations

The study only collected data from publicly accessible social networks that have been anonymized by various means, particularly by replacing all usernames and links with anonymous text and summaries of tweets that have been edited, retaining the original message, avoiding direct quotations being identifiable, and ensuring that no information is provided on the identity of the individuals who posted the content studied on the platform.

Internet research requires researchers to carefully consider guidelines to determine whether ethics approval and informed consent are needed [30]. On the basis of the terms set out by the Research Ethics Committee at the University College London [31], the study was considered exempt from formal ethics approval for the following reasons: (1) study involving information freely available in the public domain, such as published biographies, newspaper accounts of an individual's activities, and published minutes of a meeting, that although is considered personal under the Data Protection Act, would not require ethics review; and (2) study involving anonymized records and data sets in the public domain, such as data sets available through the Office for National Statistics or the UK Data Archive where appropriate permissions have already been obtained and it is not possible to identify individuals from the information provided.

Therefore, we anonymized all records and data sets collected during the study to make identification impossible. We removed social media usernames from the data samples. No direct or easily traceable quotes have been included. These measures align with best practices [32-35]. While this study was beyond the scope of the human ethics committee, we adhered to the principles of ethics: beneficence, nonmaleficence, autonomy, and justice [36]. We collected and analyzed data through secure encrypted servers via the Meltwater and InfraNodus platforms.

Results

Audience Analysis

During the period from January 2021 to January 2022, we obtained 300,000 tweets from 936 accounts. After filtering for relevant posts (refer to inclusion and exclusion criteria in

XSL•FO
RenderX

Multimedia Appendix 1), we analyzed a sample of 2588 tweets using mixed methods analysis. In terms of gender (man, woman, nonbinary, or unknown), 32.88% (851/2588) were female individuals, 23.49% (608/2588) were male individuals, and 43.59% (1128/2588) were unknown. According to the description given in the user's biography, the most frequently self-reported terms were "NHS" (582/2588, 22.49%), "health" (230/2588, 8.89%), "medical" (168/2588, 6.49%), "nurse" (166/2588, 6.41%), "clinical" (160/2588, 6.18%), "mum" (158/2588, 6.11%), "doctor" (145/2588, 5.6%), and "GP" (145/2588, 5.6%). In terms of city, tweets came mainly from London (958/2588, 37.02%), Newcastle upon Tyne (326/2588, 12.6%), Redcar (160/2588, 6.18%), Manchester (140/2588, 5.41%), and Bradford (111/2588, 4.29%).

Profession

Regarding profession described in the user's biography, the most frequently mentioned roles were nurses (176/2588, 6.8%); medical roles, for example, paramedic and nursing assistant (173/2588, 6.68%); clinical roles, for example, surgeon, physiotherapist, and anesthesiologist (160/2588, 6.18%); general practitioners (GPs), for example, hospital GP or local surgery GP (142/2588, 5.49%); and physician (140/2588, 5.41%). The most frequent organization affiliated with was the National Health Service (587/2588, 22.68%).

Most Influential Accounts

One of the accounts that generated the highest number of mentions and, therefore, some of the most influence, as they were the ones that talked the most about post–COVID-19 condition in children and young people, was the account for @longcovidkids (593/2588, 22.91% tweets), related to the most shared website *longcovidkids.org* [37], an international UK-based charity for families and children living with post–COVID-19 condition. Although the account was created in October 2020, it was first mentioned in our data collection timeline on January 1, 2021. It offers web support services, funding, and research participation and represents children and young people living with post–COVID-19 condition in expert forums, research panels, health organizations, and parliamentary groups. The other most shared web pages were theguardian.com (the United Kingdom) [38], bbc.co.uk (the United Kingdom) [39], peoplewith.com (the United States) [40], and ncbi.nlm.nih.gov (the United States) [41]. This shows that in the United Kingdom, there was a mixed influence of UK and US link resources linked to HCW Twitter users in the United Kingdom.

Keyword Analysis

The volume of social media engagement in the discussion about the post–COVID-19 condition experience in children and young people in the United Kingdom reached 1400 posts, 1550 engagements, and 1.9 million impressions. Overall, comments were very responsive to government decisions regarding the vaccination program and school closures (Multimedia Appendix 4). During the first peak of comments in January 2021, the amount of discourse expanded leading up to March 2021, when there were different announcements of school closures, and the guidelines were delivered regarding the priority groups of the vaccination program (frontline HCW and people aged >80 years first). The highest engagement was between June and July 2021, which coincides with the government announcement regarding the availability of vaccines for people aged >18 years. The third peak of comments occurred in September 2021, the same month the authorities announced the extension of the vaccination program to children aged 12 to 15 years.

Top Keywords Analysis

The top words in posts associated with children and young people's experience of post–COVID-19 condition in the United Kingdom were "Children" (352/2588, 13.6%), "kids" (160/2588, 6.18%), "people" (158/2588, 6.11%), "Young" (148/2588, 5.72%), and "schools" (83/2588, 3.21%). The top hashtags were #longcovid (1387/2588, 53.59%), #longcovidkids (448/2588, 17.31%), #covid19 (370/2588, 14.3%), and #covid (176/2588, 6.8%).

Sentiment and Emotions Analysis

According to sentiment analysis, 99.38% (2572/2588) of the posts reflected negative sentiments and 0.62% (16/2588) reflected positive sentiments. Negative sentiments were mainly associated with comments on hospitalization figures related to the COVID-19 pandemic, criticism of pandemic mitigation policies, and vaccination of children and young people. Furthermore, positive sentiments mainly concerned acknowledgments around decreasing numbers of community support groups.

The primary emotions identified were as follows:

- Sadness (1752/2588, 67.7%), such as in the following tweet:

  @[Username] Really upset, after my tough on-call last night. Hospitalisations are still going up, and Gov announcement minismises the effect of long-COVID in adults and children. It's so hard to keep spirits up today. But we'll try and continue doing our best in the NHS.

- Joy (367/2588, 14.18%), such as in the following tweet:

  @[Username] It's been an amazing day! [...] I've been able to share the experience I've gained treating children and adolescents with Long COVID over the last year.

- Fear (233/2588, 9%), as seen in the following tweet:

  @[Username] It's really urgent that young people get the message that they need to get vaccinated. Long COVID is ruining many people's lives! It's not a lie or hypochondria, there are real, physiological changes, please understand!

Segmentation Analysis

This analysis revealed the critical clusters of conversation around the main topics of concern within the discourse network around post–COVID-19 condition. Comments were distributed in 4 key conversation segments as follows:

1. People, schools, and prevention (1734/2588, 67%): Most of the comments related to measures taken in terms of

XSL•FO
RenderX

COVID-19 prevention in schools, concern about the risk of exposure, and sharing experiences of infection in schools.

2. Health, adults, and impact (401/2588, 15.49%): Comments mainly reflected concerns and uncertainty about the long-term effect of post–COVID-19 condition on both children and young people and adults.
3. Cases, virus, and risk (326/2588, 12.6%): Comments reflected worries about the associated risks and long-term consequences attributable to post–COVID-19 condition (in both adults and children and young people) and the constant mutation of the virus, which will create a permanent risk in the population.
4. Months, distress, and symptoms (106/2588, 4.1%): Some HCWs used Twitter to share how children and young people experience post–COVID-19 condition and the extent of these symptoms. Some HCWs exemplified certain typical manifestations, such as fatigue.

Discourse Analysis by Theme

To better understand the topics discussed from the segmentation analysis, we performed a discourse analysis of the key co-occurring themes and topics of concern shared within discussions regarding post–COVID-19 condition in children and young people. The following themes emerged (Textbox 1): concern or uncertainty for the future, school attendance, mask protection from COVID-19, vaccine uptake, infection rates, policy (support or skepticism), understanding and visualizing symptoms, child mental health, access to care, community support, and research (Figures 1 and 2).

XSL•FO
RenderX

Textbox 1. Discourse analysis themes.

- Concern for the future or uncertainty (615/2588, 23.76% tweets): Most comments showed a concern for the future, focusing on shared statistics regarding the rate and spread of infection in children and young people and how this would affect future health outcomes. Furthermore, this group expressed concern regarding political decisions; the presence of illness in loved ones; the eventual overload and response capacity of the health system in the face of an increase in post–COVID-19 condition cases; and the need for training of health care workers (HCWs) to deal with comorbid, potentially long-term symptoms (Figure 1A).
- Schools (460/2588, 17.77% tweets): Comments aimed to promote vaccination policies for schoolchildren and flexible measures regarding teachers' work and attendance, considering cases of people with prolonged symptoms. In addition, several tweets expressed dissatisfaction with school risk mitigation measures, such as the use of face masks and air filters (Figure 1B).
- Vaccine (386/2588, 14.9% tweets): Most tweets from this group showed their disapproval of the constant changes in the government's decisions regarding schools and priority groups for vaccination. Between March and June 2021, the first set of tweets criticized the lack of priority in the vaccination program for schoolchildren and other at-risk groups (such as teachers). Once the authorities announced a vaccination program for schoolchildren aged 12 to 15 years (Multimedia Appendix 4), most comments promoted vaccination for this group. A few comments (78/2588, 3.01%) shared concerns about the vaccine's efficacy for children, based on the experiences of COVID-19 reinfection in adults despite having received the recommended initial doses. However, to a lesser extent (26/2588, 1%), there was a refusal to vaccinate children, citing fear of possible adverse effects. Nonetheless, it is worth noting that the community frequently refuted such comments (Figure 1C).
- Share statistics (334/2588, 12.91% tweets): Frequently, HCWs shared statistical data, such as the number of affected children and young people, the number of post–COVID-19 condition cases, and hospital admissions and deaths. Some of these data were used to validate the existence of the post–COVID-19 phenomenon or to express concern about it (Figure 1D).
- Policy (316/2588, 12.21% tweets): The comments were responsive to the policies emanating from the authorities over time (Multimedia Appendix 4). There were 5 main criticisms, including changes in school closure or opening policies; HCWs question why the authorities ignore the evidence of post–COVID-19 cases in children and young people, leading them to question whether decision makers have sufficient training to control the pandemic adequately; the failure to include teachers and school workers in the COVID-19 vaccination program as well as the younger population; the lack of mitigation measures in schools, such as improvements in ventilation systems and mandatory use of masks; and the herd immunity as a plan in the government's *hidden agenda*, that is, to promote work and activate the economy (Figure 1E).
- "Proof" (280/2588, 10.82% tweets): Most tweets in this group argued regarding the existence of children and young people with post–COVID-19 condition through pictures; statistics; scientific papers; and personal, family, and professional experiences (Figure 1F).
- Signs and symptoms (189/2588, 7.3% tweets): Among the symptoms described, chronic fatigue and exhaustion were the most frequent symptoms, which prevent normal activities. Other symptoms were respiratory: dyspnea, chronic cough, and shortness of breath; gastrointestinal: acute or intense abdominal pain, nausea, bloating, gastroparesis, and change in smell or taste; muscular: severe joint pain, "painful foot" and difficulty with physical activity; mental health: anxiety and low mood; topical: rash, skin rashes, and redness and pain in the eyes; and nonspecific symptoms, such as chest pain, heart palpitations, constant high body temperature, precocious puberty, hormonal changes, and erectile dysfunction (Figure 2A).
- Face masks (119/2588, 4.6% tweets): Face masks were widely promoted, especially in schools, because HCWs considered them as a practical and straightforward strategy to control the pandemic (Figure 2B).
- Skepticism (101/2588, 3.9% tweets): Comments showed reticence toward post–COVID-19 condition in children and young people. Some of the arguments focused on a perceived lack of clarity in the clinical manifestations and stressed the need to better differentiate the post–COVID-19 condition from other related symptomatologies, such as mood disorders (eg, depression and anxiety due to confinement). In contrast, several arguments agreed on the need for more scientific evidence, arguing that post–COVID-19 condition in children and young people are isolated. Other users claimed not to know of such cases instead of calling post–COVID-19 condition in children and young people SMS text message an exaggeration. In addition, several arguments favored releasing restrictions for children and young people, particularly arguments related to the use of masks, because of possible associated risks, for example, hypoxia (Figure 2C).
- Mental health (54/2588, 2.09% tweets): Symptoms attributable to mental health problems in children and young people were also a concern. For instance, HCWs mentioned sadness, fear of infecting their family, anxiety regarding sick parents, stress, night terrors, self-harm, and suicidal ideation. Furthermore, users discussed a perceived lack of specific support for children and young people and their families in situations such as hospitalization; prolonged COVID-19 condition; admission to intensive care; and death of a family member, schoolmate, or teacher, all situations that triggered permanent stress in these groups (Figure 2D).
- Community support or asking for advice (93/2588, 3.59% tweets): Some HCWs used Twitter to ask for guidance on a specific issue or share experiences of having post–COVID-19 condition or caring for children and young people or family members. Furthermore, they shared informative infographics provided by experts regarding post–COVID-19 condition in children and young people (Figure 2E).
- Access to health care or treatment (72/2588, 2.78% tweets): Some HCWs mentioned the lack of specialist (cardiology) support, concerns regarding prolonged National Health Service burnout, and criticisms regarding how follow-up was carried out concerning the relative symptomatology of children and young people with post–COVID-19 condition. At the same time, opening new centers for children and young people with post–COVID-19 condition generated different reactions. On the one hand, some HCWs recognized it as a substantial development, but on the other hand, some HCWs recognized it as proof of the existence of post–COVID-19 condition in children and young people, which raised concerns for the future (Figure 2F).
- Research (52/2588, 2% tweets): Under this theme, tweets largely promoted study on post–COVID-19 condition in children and young people or highlighted the need for further study on the subject (Figure 2G).
-

XSL•FO
RenderX

Images (57/2588, 2.2% tweets): Images shared were primarily from scientific studies, including infographics (from organizations such as National Health Service or @LongCovidKids) and visualization of children and young people's symptoms, such as rashes, COVID-19 toe, and joint pain. Most infographics shared by organizations (and not individuals), such as the organization LongCovidKids, were related to statistics, such as the number of children and young people with post–COVID-19 condition or the quantification of the type of symptoms experienced. Shared photographs tended to show the more "visually recognizable" symptoms of post–COVID-19 condition, such as skin lesions, rashes, or inflammation. The less visible symptoms, such as chronic fatigue and neurological issues, were represented with photographs of children and young people lying, sleeping under blankets, or duvets or on hospital beds (Figure 2H).

Figure 1. Discourse analysis by theme: (A) concern for the future or uncertainty, (B) schools, (C) vaccine, (D) shared statistics, (E) policy, and (F) proof.

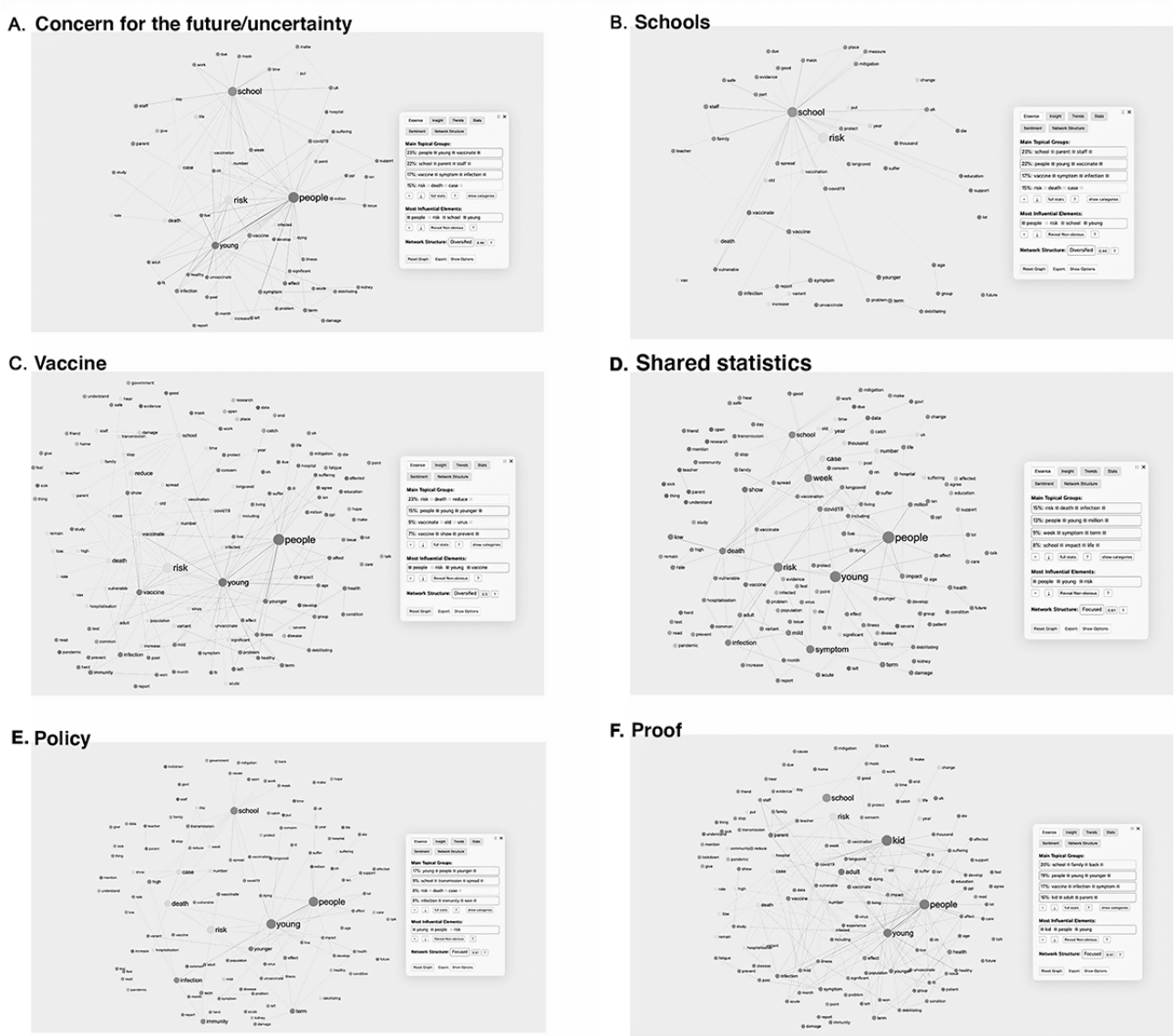

XSL•FO
RenderX

Figure 2. Discourse analysis by theme: (A) signs and symptoms, (B) face masks, (C) skepticism, (D) mental health, (E) community support or asking for advice, (F) access to health care or treatment, (G) research, and (H) images.

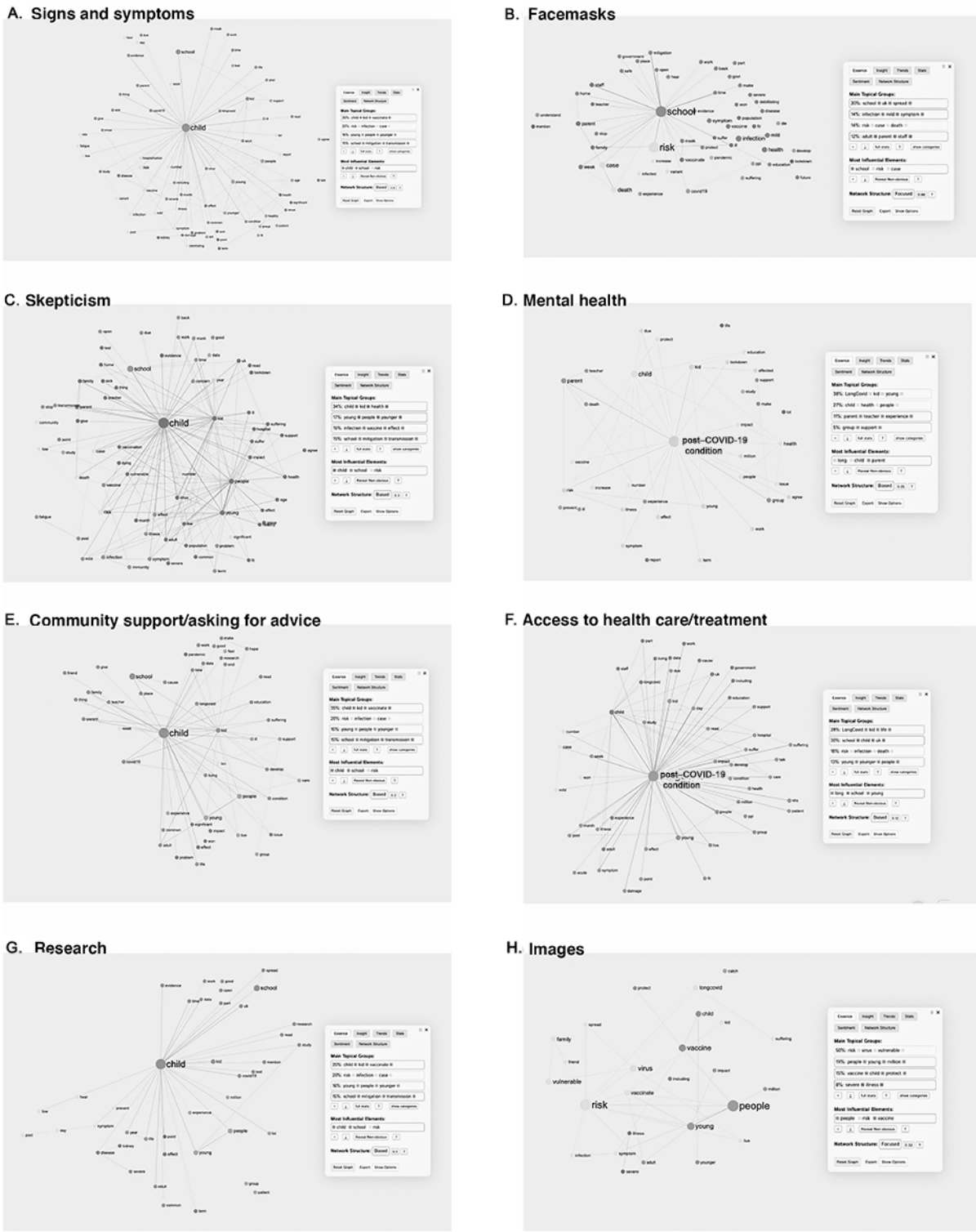

Discussion

Principal Findings

Our primary objective was to explore HCWs' perceptions concerning post–COVID-19 condition in children and young people in the United Kingdom between January 2021 and January 2022. Our findings indicated that comments made by HCWs on Twitter were responsive to announcements issued by authorities regarding the management of the COVID-19 pandemic in the United Kingdom and associated regulations on the operation of schools. The most frequent feelings and emotions were negative, mainly sadness. In turn, we identified relevant themes for HCWs, such as uncertainty or concern about the future; policies; and regulations for the prevention, management, and addressing both COVID-19 and post–COVID-19 condition in children and young people;

XSL•FO
RenderX

vaccination; and the use of Twitter as a strategy to share scientific literature, management strategies, and clinical and personal experiences.

Concern from HCWs regarding the policies for addressing the COVID-19 pandemic in the children and young people in the United Kingdom (including vaccination and schools) was a recurring theme in our findings. Furthermore, concern regarding the side effects of the COVID-19 vaccine and how the vaccine might interact with preexisting physiological symptoms of post–COVID-19 condition in children and young people was a topic of discussion. Similarly, the constant change in policy making in the United Kingdom, as public health bodies and governments have tried to understand and adapt to the emergence of post–COVID-19 condition, have added to the strength of this ongoing debate [42]. The lack of up-to-date evidence on post–COVID-19 condition in children and young people prompted HCWs to rely on Twitter during the pandemic to communicate relevant information. Twitter has a broad audience reach; is used as a communication tool by politicians, health bodies, and other key influences; and facilitates real-time updates [43]. During the pandemic, HCWs, primarily those in frontline roles and local response coordination, have often been challenged to become credible spokespersons for pandemic information [44]. Such credibility directly influences public confidence and decision-making, ultimately determining the success or failure of a public health intervention [43].

Furthermore, failures in risk communication could explain the presence of uncertainty and negative feelings associated with school regulations. When people are upset, distressed, or fearful, they often do not trust the authority, decrease the perceived validity of the communication received, and find information processing difficult [45]. In this regard, Fotheringham et al [46] indicated that during 2020, school leaders in the United Kingdom faced pressures and challenges related to translating and enacting school policies, particularly with the perceived lack of agency shared by the government concerning being able to translate centrally issued guidelines. In turn, Tomson et al [47] reported that the pandemic has negatively impacted the well-being of leaders in all types of schools and across all demographic groups, affecting their ability to think clearly and solve work-related problems. Given that the protection and care of children and young people health during the COVID-19 pandemic ultimately rests with school leaders, the search for support strategies that focus on the needs of these groups becomes an urgent necessity.

Findings in Relation to Other Studies

Using Twitter's information, this is one of the first studies to capture health professionals' perceptions of prolonged COVID-19 in the children and young people in the United Kingdom. However, other studies have addressed post–COVID-19 condition on this social network. Callard and Peregov [48] reviewed how, through social platforms such as Twitter, patients made the persistence and heterogeneity of COVID-19 symptoms visible, thus catapulting the inclusion of post–COVID-19 condition as a relevant phenomenon in clinical and policy debates. In contrast, other authors in the last 2 years have explored on various platforms (including Twitter) the persistence of symptoms and emotional impact after months of suspected and confirmed diagnosis of COVID-19 [49-55], including the period of vaccination. Furthermore, others have explored web discussions regarding this phenomenon [56]. Several of these authors agree on a perceived lack of support and specific resources from governmental bodies, a lack of information or clarity in the instructions given, and the absence of formal mechanisms to allow the voices of patients and the community to be heard. The above point is critical as it highlights the gap between the needs of the population and the response provided by policy makers, which not only translates into a gap in access to health services but also limits citizen participation in decision-making on the issues that affect their own health and increases distrust toward regulations and instructions issued by the government.

Implications for Policy and Practice

Overview

Several policy recommendations and implications are targeted at various stakeholders to consider while implementing future policy guidelines to address post–COVID-19 health care delivery. First, policy makers should consider investing appropriate resources to collect data regarding post–COVID-19 condition in children and young people, specifically on the impact of COVID-19 on the mental health of children and young people. This implies working closely with researchers to streamline data collection and reporting on post–COVID-19 condition. Second, policy makers should consider providing a basic level of psychosocial support with access to quality mental health and psychosocial support services for HCWs, school staff, parents, and children and young people experiencing post–COVID-19 condition. This implies strengthening health systems, community-based programming, and mobilization. Policies must include documenting the impact of mental health and psychosocial support interventions and innovative approaches to be more widely disseminated and scaled up across different contexts and target population groups. Third, to address the criticism around frequent changes in school closure and opening policies, decision makers should develop clear, easy-to-understand school mitigation plans informed by the best available evidence. The plans should incorporate teachers, school workers, and parents to ensure all voices are included in the policy plan. Fourth, policy makers should adopt a shared decision-making approach incorporating HCWs in the decision-making process for managing the COVID-19 pandemic. Finally, government decision makers should set post–COVID-19 pandemic recovery policies informed from a health equity perspective and how this affects children and young people living with post–COVID-19 condition, factoring in childhood, family income, housing, domestic violence, access to health care, and racism.

In terms of the needed clearer road map for recommendations to support training strategies for HCWs and school staff regarding post–COVID-19 condition in children and young people, we have outlined the following 10 steps.

XSL•FO
RenderX

Step 1: Data Collection and Analysis

Our study underlines the critical need for comprehensive data on post–COVID-19 condition's impact on the mental health of children and young people. As a first step, it is recommended that policy makers should allocate resources for the systematic collection and analysis of data on post–COVID-19 condition in children and young people, particularly focusing on mental health outcomes. These data should be used to identify the most prevalent symptoms and the most effective treatment strategies. In this context, it is recommended that experts emphasize the importance of early detection and medical consultation for mental health issues in children and young people diagnosed with post–COVID-19 condition, including mood changes, irritability, social withdrawal, memory problems, difficulty in concentrating, anxiety, depression, posttraumatic stress, school absenteeism, and suicidal ideation [57,58]. This entails working closely with researchers to streamline data collection and reporting on post–COVID-19 condition.

Step 2: Psychosocial Support Framework

It has been noted that globally, programs for managing post–COVID-19 condition in children and young people are heterogeneous, ranging from the use of physiotherapy, pediatric occupational therapy, and psychological support to interventions aimed at lifestyle modifications [59]. This diversity could impact differential outcomes in the treatment, recovery, and timely and effective rehabilitation of children and young people with post–COVID-19 condition. Upon analyzing the wider literature and the social media data in this study, it is recommended that a basic level of psychosocial support should be established. This would involve ensuring access to quality mental health services for HCWs, school staff, parents, and children and young people with post–COVID-19 condition. This framework should be integrated into the health system and community-based programming, emphasizing the mobilization of resources and strengthening of support networks. It is suggested that the psychosocial support framework should facilitate access to quality mental health services and support networks that are robust and responsive. Community engagement gleaned from further Twitter discourse analysis should be a helpful guide in the development of these services to ensure they meet the real and expressed needs of children and young people with post–COVID-19 condition. Practical examples of basic psychosocial support include using web support services; individual or group therapy sessions; school-based emotional support programs; and counseling sessions aimed at parents, family members, or school staff.

Step 3: Educational Mitigation Plans

The frequent policy changes around school closures highlight the necessity for stable and clear educational mitigation plans. It is recommended that these plans should be directly informed by the evidence collected and further analysis of sentiments and emotions surrounding post–COVID-19 condition in schools. Incorporating the viewpoints of teachers, parents, and school staff, as identified in our thematic analysis, will ensure that the mitigation strategies are comprehensive, feasible, and sensitive to the psychosocial impact on children and young people. School staff and policy makers should collaborate to develop clear, evidence-informed educational mitigation plans. These plans should be straightforward and involve teachers, school workers, and parents in their creation, ensuring a unified approach that considers the voices of all stakeholders.

Step 4: Shared Decision-Making in Health Care

In health care settings, the adoption of a shared decision-making model is crucial, enabling HCWs to actively contribute to the formulation of COVID-19 and post–COVID-19 policies. This inclusive approach ensures that frontline workers can provide valuable insights toward policy development. To facilitate this, the establishment of advisory committees composed of representatives from HCWs is recommended. This committee can convene regularly to deliberate on key decisions pertaining to the COVID-19 pandemic management, including prevention measures, resource distribution, and vaccination strategies. Such collaborative groups have demonstrated effectiveness in identifying priority needs within the context of a pandemic [60].

Step 5: Health Equity in Policy Setting

Post–COVID-19 recovery policies should be set with a health equity lens. This means considering factors such as family income, housing, domestic violence, access to health care, and racism and how these factors affect children and young people living with post–COVID-19 condition. Our findings emphasize the importance of framing post–COVID-19 recovery policies through a lens of health equity. The concerns raised by HCWs regarding the socioeconomic impacts, such as family income and access to health care, underline the need for policies that address not just the medical aspects of post–COVID-19 condition but also the social determinants of health. An equitable approach will ensure that children and young people from diverse backgrounds receive appropriate support.

Step 6: Documenting and Disseminating Interventions

It is vital to document the impact of mental health and psychosocial support interventions. In this context, it is crucial to implement innovative strategies to disseminate unbiased information about post–COVID-19 condition among health care professionals and educators working with children and young people, ensuring it reaches different contexts and populations. These strategies may include creating interactive multimedia resources, such as videos and mobile apps; organizing webinars; actively using social media; and forming web support groups. These groups will provide a space where patients, health care professionals, and educators can share their experiences and knowledge regarding post–COVID-19 condition. These actions will not only help reduce isolation and social stigma but also strengthen support for these groups considered vulnerable [61].

Step 7: Developing a Clear Communication Strategy

Policy makers must develop a clear communication strategy to address frequent policy changes and mitigate confusion. This strategy should be informed by the data collected and analysis conducted in Step 1. The data reveal a palpable sense of uncertainty and frustration due to frequent policy shifts, underscoring the need for a clear and consistent communication strategy. This strategy should be grounded in the evidence gathered from the health care community's discourse and aim to minimize confusion by providing timely, transparent, and

XSL•FO
RenderX

reliable information regarding post–COVID-19 policies and support services.

Step 8: Training and Support Strategies

On the basis of the findings of the comprehensive data analysis, specific training and support strategies should be developed for HCWs and school staff. These strategies should be focused on the practical aspects of identifying and managing post–COVID-19 condition in children and young people. For instance, training sessions could include practical workshops on recognizing post–COVID-19 symptoms in children and adolescents, conducting diagnostic assessments, and implementing appropriate treatment and support interventions.

Step 9: Continuous Feedback and Policy Adaptation

The continuous evolution of the post–COVID-19 phenomenon demands an iterative approach to policy making. On the basis of our study, we recommend establishing feedback mechanisms with HCWs and school staff to monitor the reception and effectiveness of implemented policies. This feedback, coupled with ongoing research, should inform policy adaptations to ensure they remain aligned with the evolving landscape of post–COVID-19 condition and its impact on children and young people.

Step 10: Evaluation and Research

Finally, there should be a commitment to ongoing evaluation and research. This will involve not only monitoring the implementation of the abovementioned steps but also supporting new research to fill any remaining gaps in understanding the long-term effects of COVID-19 on children and young people.

This sequence of steps is designed to be iterative and responsive, ensuring that the recommendations from the study are translated into concrete actions that adapt to emerging data and research findings.

Strengths and Limitations

A key strength of this study is that our social media analysis of post–COVID-19 condition contributes toward an emerging understanding of reported experiential, emotional, and practical dimensions of post–COVID-19 condition in children and young people specifically and questions of vaccine hesitancy in children and young people with post–COVID-19 condition. This is one of the few studies to collect HCWs' perceptions regarding post–COVID-19 condition in children and young people in the United Kingdom using information from Twitter. We identify key areas that need considering attention and focus, such as the provision of psychosocial support with access to quality mental health resources to alleviate the impact of post–COVID-19 condition in children and young people and the development of clear post–COVID-19 pandemic recovery guidelines that are informed by health equity perspective, and how this affects children and young people living with post–COVID-19 condition.

One of the limitations this study acknowledges is the definition of post–COVID-19 condition in children and young people. When data were collected, the lack of consensus on the definition of post–COVID-19 condition in children and young people forced us to formulate a definition of post–COVID-19 condition in children and young people based on the available literature. Furthermore, this study is limited to the perceptions of people who used descriptors in their web biography attributable to HCWs; therefore, our results only represent some HCWs in the United Kingdom and those in other countries. In turn, this research collected data from Twitter only; therefore, further inquiry into HCWs' perceptions of post–COVID-19 condition in children and young people required expanding to other data sources or social networks and including languages other than English. We acknowledge that demographic factors, geographic location, and individual daily activities of social media users can significantly influence language use and word choice, introducing potential biases in tweet-based data. Such biases are inherent in any analysis of social media content and can affect the generalizability of findings. For instance, our study relies on Twitter data, which do not encompass the full spectrum of global or the UK public opinion on post–COVID-19 condition in children and young people. While Twitter serves as a valuable platform for capturing real-time sentiments and experiences, it is not fully representative of all demographics and geographic regions. Our results may reflect the perspectives of more vocal or active social media users, which may not correspond to the silent majority or those without access to social media. In addition, the absence of geotagged information for many users limits our ability to conduct a more nuanced spatial analysis of the sentiments expressed.

Furthermore, our study is built upon the recognition that social media data may overrepresent certain demographic groups while underrepresenting others, such as the older population or those without reliable internet access. This skew can influence the apparent prevalence of certain views or experiences of post–COVID-19 condition. Moreover, individuals' patterns of daily life, reflected in their social media use and content, contribute additional layers of complexity and potential bias to the discourse analyzed.

Consistent with scholarly precedents on the subject [62,63], our study acknowledges these biases as intrinsic limitations of social media–based research. Although our analysis did not control for these factors, we recognize their potential impact on our results. Future studies would benefit from incorporating a broader array of data sources, including interviews or focus groups, to provide a more representative and comprehensive understanding of post–COVID-19 condition in children and young people. This approach would complement our Twitter-based findings and help mitigate the biases inherent in social media data.

Conclusions

More than a year after the start of the COVID-19 pandemic, the perceptions described on Twitter by HCWs concerning the presence of post–COVID-19 condition in children and young people appear to be a relevant and timely issue as well as very responsive to the declarations and guidelines issued by the health authorities over time. The most prominent group within the discourse studied was the activist or lobbying organization @LongCovidKids, which shared the most tweets and images over the period studied. We recommend that future research focus on how web health activism is organized and carried out

XSL•FO
RenderX

for children and young people with post–COVID-19 condition. Such a strategy would allow for a better understanding of the scope and impact of this phenomenon and how it can influence decision-making. Furthermore, we suggest different mitigation strategies, support, and training of HCWs and school staff regarding manifestations and treatment of post–COVID-19 condition in children and young people across all demographic areas.

Acknowledgments

The authors would like to thank the Rapid Research Evaluation and Assessment Lab, Department of Targeted Intervention, University College London, London, United Kingdom, whose support has been essential for developing this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Filters used for the search strategy on Twitter.
[DOCX File , 22 KB - jmir_v26i1e50139_app1.docx]

Multimedia Appendix 2
Sentiment analysis framework: attitudes toward post–COVID-19 condition in children and young people.
[DOCX File , 16 KB - jmir_v26i1e50139_app2.docx]

Multimedia Appendix 3
Theme codebook: examples of tweets that fit into main themes tagged for mention of children and young people with post–COVID-19 condition.
[DOCX File , 21 KB - jmir_v26i1e50139_app3.docx]

Multimedia Appendix 4
Timeline of national governmental policies and guidelines regarding children and young people.
[DOCX File , 19 KB - jmir_v26i1e50139_app4.docx]

References

1. World Health Organization. COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum. Global Research Collaboration for Infectious Disease Preparedness. 2020 Feb 12. URL: https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum [accessed 2022-06-06]
2. Johns Hopkins University of Medicine. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Coronavirus Resource Center. 2020 Jan 22. URL: https://coronavirus.jhu.edu/map.html [accessed 2023-06-09]
3. Nature Editorial. Long COVID and kids: more research is urgently needed. Nature 2022 Feb 08;602(7896):183 [FREE Full text] [doi: 10.1038/d41586-022-00334-w] [Medline: 35136225]
4. Long COVID or post-COVID conditions. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html [accessed 2023-06-09]
5. Davis HE, McCorkell L, Vogel JM, Topol E. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol 2023 Mar 13;21(3):133-146 [FREE Full text] [doi: 10.1038/s41579-022-00846-2] [Medline: 36639608]
6. Zimmermann P, Curtis N. Why is COVID-19 less severe in children? A review of the proposed mechanisms underlying the age-related difference in severity of SARS-CoV-2 infections. Arch Dis Child 2020 Dec 01;106(5):429-439 [FREE Full text] [doi: 10.1136/archdischild-2020-320338] [Medline: 33262177]
7. Zheng YB, Zeng N, Yuan K, Tian SS, Yang YB, Gao N, et al. Prevalence and risk factor for long COVID in children and adolescents: a meta-analysis and systematic review. J Infect Public Health 2023 Mar 07;16(5):660-672 [FREE Full text] [doi: 10.1016/j.jiph.2023.03.005] [Medline: 36931142]
8. World Health Organization. A clinical case definition for post COVID-19 condition in children and adolescents by expert consensus. World Health Organization. 2023 Feb 16. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Post-COVID-19-condition-CA-Clinical-case-definition-2023-1 [accessed 2024-03-26]
9. Stephenson T, Allin B, Nugawela MD, Rojas N, Dalrymple E, Pinto Pereira S, CLoCk Consortium, et al. Long COVID (post-COVID-19 condition) in children: a modified Delphi process. Arch Dis Child 2022 Jun 17;107(7):674-680 [FREE Full text] [doi: 10.1136/archdischild-2021-323624] [Medline: 35365499]

XSL•FO
RenderX

10. Esposito S, Principi N, Azzari C, Cardinale F, Di Mauro G, Galli L, et al. Italian intersociety consensus on management of long Covid in children. Ital J Pediatr 2022 Mar 09;48(1):1-9 [FREE Full text] [doi: 10.1186/s13052-022-01233-6] [Medline: 35264214]
11. Stephenson T, Pinto Pereira SM, Shafran R, de Stavola BL, Rojas N, McOwat K, CLoCk Consortium, et al. Physical and mental health 3 months after SARS-CoV-2 infection (long COVID) among adolescents in England (CLoCk): a national matched cohort study. Lancet Child Adolesc Health 2022 Feb 08;6(4):230-239 [FREE Full text] [doi: 10.1016/S2352-4642(22)00022-0] [Medline: 35143770]
12. Stephenson T, Pinto Pereira SM, Nugawela MD, McOwat K, Simmons R, Chalder T, CLoCk Consortium. Long COVID-six months of prospective follow-up of changes in symptom profiles of non-hospitalised children and young people after SARS-CoV-2 testing: a national matched cohort study (The CLoCk) study. PLoS One 2023 Mar 6;18(3):e0277704 [FREE Full text] [doi: 10.1371/journal.pone.0277704] [Medline: 36877677]
13. Twitter users, stats, data and trends. Data Reportal. 2023 May 11. URL: https://datareportal.com/essential-twitter-stats#:~:text=Based [accessed 2023-06-09]
14. Rufai SR, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. J Public Health (Oxf) 2020 Aug 18;42(3):510-516 [FREE Full text] [doi: 10.1093/pubmed/fdaa049] [Medline: 32309854]
15. Haman M. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. Heliyon 2020 Nov;6(11):e05540 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e05540] [Medline: 33294685]
16. Huang X, Li Z, Jiang Y, Li X, Porter D. Twitter reveals human mobility dynamics during the COVID-19 pandemic. PLoS One 2020 Nov 10;15(11):e0241957 [FREE Full text] [doi: 10.1371/journal.pone.0241957] [Medline: 33170889]
17. Wicke P, Bolognesi MM. Framing COVID-19: how we conceptualize and discuss the pandemic on Twitter. PLoS One 2020 Sep 30;15(9):e0240010 [FREE Full text] [doi: 10.1371/journal.pone.0240010] [Medline: 32997720]
18. Chang CH, Monselise M, Yang CC. What are people concerned about during the pandemic? detecting evolving topics about COVID-19 from Twitter. J Healthc Inform Res 2021 Jan 17;5(1):70-97 [FREE Full text] [doi: 10.1007/s41666-020-00083-3] [Medline: 33490856]
19. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550 [FREE Full text] [doi: 10.2196/20550] [Medline: 33119535]
20. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
21. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020 Dec 14;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
22. Rashid MA, Yip SW, Gill D, Arfeen Z. Sharing is caring: an analysis of #FOAMed Twitter posts during the COVID-19 pandemic. Postgrad Med J 2022 Mar 14;98(1157):199-204. [doi: 10.1136/postgradmedj-2020-139267] [Medline: 33318145]
23. Ghosh P, Schwartz G, Narouze S. Twitter as a powerful tool for communication between pain physicians during COVID-19 pandemic. Reg Anesth Pain Med 2021 Feb 21;46(2):187-188. [doi: 10.1136/rapm-2020-101530] [Medline: 32321859]
24. Burgos LM, Gil Ramirez A, Utengen A, Thamman R. Use of Twitter during COVID-19 pandemic: an opportunity for continuing medical education in cardiology. Medicina (B Aires) 2020;80 Suppl 6:122-123 [FREE Full text] [Medline: 33481748]
25. Margus C, Brown N, Hertelendy AJ, Safferman MR, Hart A, Ciottone GR. Emergency physician Twitter use in the COVID-19 pandemic as a potential predictor of impending surge: retrospective observational study. J Med Internet Res 2021 Jul 14;23(7):e28615 [FREE Full text] [doi: 10.2196/28615] [Medline: 34081612]
26. Pulsar TRAC- Audience intelligence platform and social listening tool. Pulsar. URL: https://www.pulsarplatform.com/solutions/pulsar-trac/ [accessed 2022-03-01]
27. Vera San Juan N, Aceituno D, Djellouli N, Sumray K, Regenold N, Syversen A, et al. Mental health and well-being of healthcare workers during the COVID-19 pandemic in the UK: contrasting guidelines with experiences in practice. BJPsych Open 2020 Dec 10;7(1):e15 [FREE Full text] [doi: 10.1192/bjo.2020.148] [Medline: 33298229]
28. Paranyushkin D. InfraNodus: generating insight using text network analysis. In: Proceedings of the World Wide Web Conference. 2019 Presented at: WWW '19; May 13-17, 2019; San Francisco, CA. [doi: 10.1145/3308558.3314123]
29. Creswell JW, Poth CN. Qualitative Inquiry and Research Design: Choosing Among Five Approaches 4th Edition. Thousand Oaks, CA: SAGE Publications, Inc; 2017.
30. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001 Nov 10;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
31. UCL research ethics. University College London. URL: https://www.ucl.ac.uk/research-ethics/do-i-need-ethical-approval#When [accessed 2023-06-12]
32. Martin S, Kilich E, Dada S, Kummervold PE, Denny C, Paterson P, et al. "Vaccines for pregnant women…?! Absurd" - Mapping maternal vaccination discourse and stance on social media over six months. Vaccine 2020 Sep 29;38(42):6627-6637. [doi: 10.1016/j.vaccine.2020.07.072] [Medline: 32788136]

XSL•FO
RenderX

33. Kummervold PE, Martin S, Dada S, Kilich E, Denny C, Paterson P, et al. Categorizing vaccine confidence with a transformer-based machine learning model: analysis of nuances of vaccine sentiment in Twitter discourse. JMIR Med Inform 2021 Oct 08;9(10):e29584 [FREE Full text] [doi: 10.2196/29584] [Medline: 34623312]
34. Ahmed W, Bath PA, Demartini G. Using Twitter as a data source: an overview of ethical, legal, and methodological challenges. In: Woodfield K, editor. The Ethics of Online Research. Bingley, UK: Emerald Publishing Limited; Dec 12, 2017.
35. Ayers JW, Caputi TL, Nebeker C, Dredze M. Don't quote me: reverse identification of research participants in social media studies. NPJ Digit Med 2018 Aug 2;1(1):30 [FREE Full text] [doi: 10.1038/s41746-018-0036-2] [Medline: 31304312]
36. Varkey B. Principles of clinical ethics and their application to practice. Med Princ Pract 2021 Jun 4;30(1):17-28 [FREE Full text] [doi: 10.1159/000509119] [Medline: 32498071]
37. Long Covid Kids homepage. Long Covid Kids. URL: https://es.longcovidkids.org/ [accessed 2023-06-12]
38. The Guardian homepage. The Guardian. URL: https://www.theguardian.com/uk [accessed 2023-06-12]
39. Welcome to the BBC. BBC. URL: https://www.bbc.com/ [accessed 2023-06-12]
40. People with health. PeopleWith. URL: https://peoplewith.com/ [accessed 2023-06-12]
41. Welcome to NCBI. National Institutes of Health National Library of Medicine. URL: https://www.ncbi.nlm.nih.gov/ [accessed 2023-06-12]
42. Edobor M. Trust and public discourse during the Covid‐19 pandemic. IPPR Progress Rev 2021 Feb 15;27(4):354-360. [doi: 10.1111/newe.12232]
43. Nielsen RK, Fletcher R, Kalogeropoulos A, Simon F. Communications in the coronavirus crisis: lessons for the second wave. Reuters Institute for the Study of Journalism. 2020 Oct 27. URL: https://reutersinstitute.politics.ox.ac.uk/communications-coronavirus-crisis-lessons-second-wave [accessed 2024-03-26]
44. Zohar T, Negev M, Sirkin M, Levine H. Trust in COVID-19 policy among public health professionals in Israel during the first wave of the pandemic: a cross-sectional study. Isr J Health Policy Res 2022 Apr 11;11(1):20 [FREE Full text] [doi: 10.1186/s13584-022-00529-6] [Medline: 35410309]
45. Enria L, Waterlow N, Rogers NT, Brindle H, Lal S, Eggo RM, et al. Trust and transparency in times of crisis: results from an online survey during the first wave (April 2020) of the COVID-19 epidemic in the UK. PLoS One 2021 Feb 16;16(2):e0239247 [FREE Full text] [doi: 10.1371/journal.pone.0239247] [Medline: 33591985]
46. Fotheringham P, Harriott T, Healy G, Arenge G, Wilson E. Pressures and influences on school leaders navigating policy development during the COVID‐19 pandemic. Brit Educ Res J 2021 Jul 30;48(2):201-227. [doi: 10.1002/berj.3760]
47. Thomson P, Greany T, Martindale N. The trust deficit in England: emerging research evidence about school leaders and the pandemic. J Educ Admin History 2021 Sep 13;53(3-4):296-300. [doi: 10.1080/00220620.2021.1975366]
48. Callard F, Perego E. How and why patients made long COVID. Soc Sci Med 2021 Jan;268:113426 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113426] [Medline: 33199035]
49. Davis HE, Assaf GS, McCorkell L, Wei H, Low RJ, Re'em Y, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. EClinicalMedicine 2021 Aug;38:101019 [FREE Full text] [doi: 10.1016/j.eclinm.2021.101019] [Medline: 34308300]
50. Miyake E, Martin S. Long Covid: online patient narratives, public health communication and vaccine hesitancy. Digit Health 2021 Nov 29;7:20552076211059649 [FREE Full text] [doi: 10.1177/20552076211059649] [Medline: 34868622]
51. Matharaarachchi S, Domaratzki M, Katz A, Muthukumarana S. Discovering long COVID symptom patterns: association rule mining and sentiment analysis in social media tweets. JMIR Form Res 2022 Sep 07;6(9):e37984 [FREE Full text] [doi: 10.2196/37984] [Medline: 36069846]
52. Bhattacharyya A, Seth A, Rai S. The effects of long COVID-19, its severity, and the need for immediate attention: analysis of clinical trials and Twitter data. Front Big Data 2022 Dec 15;5:1051386 [FREE Full text] [doi: 10.3389/fdata.2022.1051386] [Medline: 36588926]
53. Samannodi M, Alwafi H, Naser AY, Al Qurashi AA, Qedair JT, Salawati E, et al. Determinants of post-COVID-19 conditions among SARS-CoV-2-infected patients in Saudi Arabia: a web-based cross-sectional study. Diseases 2022 Aug 23;10(3):55 [FREE Full text] [doi: 10.3390/diseases10030055] [Medline: 36135212]
54. Santarossa S, Rapp A, Sardinas S, Hussein J, Ramirez A, Cassidy-Bushrow AE, et al. Understanding the #longCOVID and #longhaulers conversation on Twitter: multimethod study. JMIR Infodemiology 2022;2(1):e31259 [FREE Full text] [doi: 10.2196/31259] [Medline: 35229074]
55. Déguilhem A, Malaab J, Talmatkadi M, Renner S, Foulquié P, Fagherazzi G, et al. Identifying profiles and symptoms of patients with long COVID in France: data mining infodemiology study based on social media. JMIR Infodemiology 2022 Nov 22;2(2):e39849 [FREE Full text] [doi: 10.2196/39849] [Medline: 36447795]
56. Awoyemi T, Ebili U, Olusanya A, Ogunniyi KE, Adejumo AV. Twitter sentiment analysis of long COVID syndrome. Cureus 2022 Jun;14(6):e25901 [FREE Full text] [doi: 10.7759/cureus.25901] [Medline: 35844354]
57. Malone LA, Morrow A, Chen Y, Curtis D, de Ferranti SD, Desai M, et al. Multi-disciplinary collaborative consensus guidance statement on the assessment and treatment of postacute sequelae of SARS-CoV-2 infection (PASC) in children and adolescents. PM R 2022 Oct;14(10):1241-1269 [FREE Full text] [doi: 10.1002/pmrj.12890] [Medline: 36169159]

XSL•FO
RenderX

58. Shachar-Lavie I, Shorer M, Segal H, Fennig S, Ashkenazi-Hoffnung L. Mental health among children with long COVID during the COVID-19 pandemic. Eur J Pediatr 2023 Apr 14;182(4):1793-1801 [FREE Full text] [doi: 10.1007/s00431-023-04854-z] [Medline: 36786887]
59. Brackel CL, Noij LC, Vijverberg SJ, Legghe CL, Maitland-van der Zee AH, van Goudoever JB, et al. International care programs for pediatric post-COVID condition (long COVID) and the way forward. Pediatr Res 2024 Jan 29 (forthcoming). [doi: 10.1038/s41390-023-03015-0] [Medline: 38287106]
60. Ripp J, Peccoralo L, Charney D. Attending to the emotional well-being of the health care workforce in a New York City health system during the COVID-19 pandemic. Acad Med 2020 Aug;95(8):1136-1139 [FREE Full text] [doi: 10.1097/ACM.0000000000003414] [Medline: 32282344]
61. Buonsenso D, Camporesi A, Morello R, De Rose C, Fracasso M, Chieffo DP, et al. Social stigma in children with long COVID. Children (Basel) 2023 Sep 07;10(9):1518 [FREE Full text] [doi: 10.3390/children10091518] [Medline: 37761479]
62. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. PLoS One 2013 May 29;8(5):e64417 [FREE Full text] [doi: 10.1371/journal.pone.0064417] [Medline: 23734200]
63. Padilla JJ, Kavak H, Lynch CJ, Gore RJ, Diallo SY. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. PLoS One 2018 Jun 14;13(6):e0198857 [FREE Full text] [doi: 10.1371/journal.pone.0198857] [Medline: 29902270]

Abbreviations

HCW: health care worker
GP: general practitioner
LISTEN: Collaborative and Digital Analysis of Big Qualitative Data in Time Sensitive Contexts

Edited by A Mavragani; submitted 20.06.23; peer-reviewed by R Gore, A Wahbeh; comments to author 02.11.23; revised version received 14.02.24; accepted 08.03.24; published 17.04.24.

Please cite as:
Chepo M, Martin S, Déom N, Khalid AF, Vindrola-Padros C
Twitter Analysis of Health Care Workers' Sentiment and Discourse Regarding Post–COVID-19 Condition in Children and Young People: Mixed Methods Study
J Med Internet Res 2024;26:e50139
URL: https://www.jmir.org/2024/1/e50139
doi:10.2196/50139
PMID:38630514

©Macarena Chepo, Sam Martin, Noémie Déom, Ahmad Firas Khalid, Cecilia Vindrola-Padros. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining the Role of Information Behavior in Linking Cancer Risk Perception and Cancer Worry to Cancer Fatalism in China: Cross-Sectional Survey Study

Lianshan Zhang[1], PhD; Shaohai Jiang[2], PhD

[1]School of Media and Communication, Shanghai Jiao Tong University, Shanghai, China

[2]Department of Communications and New Media, National University of Singapore, Singapore, Singapore

Corresponding Author:
Shaohai Jiang, PhD
Department of Communications and New Media
National University of Singapore
Blk AS6, 11 Computing Drive
Singapore, 117416
Singapore
Phone: 65 6516 2003
Email: cnmjs@nus.edu.sg

Abstract

Background: Reducing cancer fatalism is essential because of its detrimental impact on cancer-related preventive behaviors. However, little is known about factors influencing individuals' cancer fatalism in China.

Objective: With a general basis of the extended parallel process model, this study aims to examine how distinct cancer-related mental conditions (risk perception and worry) and different information behaviors (information seeking vs avoidance) become associated with cancer fatalism, with an additional assessment of the moderating effect of information usefulness.

Methods: Data were drawn from the Health Information National Trends Survey in China, which was conducted in 2017 (N=2358). Structural equation modeling and bootstrapping methods were performed to test a moderated mediation model and hypothesized relationships.

Results: The results showed that cancer risk perception and cancer worry were positively associated with online health information seeking. In addition, cancer worry was positively related to cancer information avoidance. Moreover, online health information seeking was found to reduce cancer fatalism, while cancer information avoidance was positively associated with cancer fatalism. The results also indicated that the perceived usefulness of cancer information moderated this dual-mediation pathway.

Conclusions: The national survey data indicate that cancer mental conditions should not be treated as homogeneous entities, given their varying functions and effects. Apart from disseminating useful cancer information to encourage individuals to adaptively cope with cancer threats, we advocate for health communication programs to reduce cancer information avoidance to alleviate fatalistic beliefs about cancer prevention.

(J Med Internet Res 2024;26:e49383) doi:10.2196/49383

KEYWORDS

cancer fatalism; cancer risk perception; cancer worry; health information seeking; information avoidance

Introduction

Background

Cancer is rapidly becoming a global health burden and is the leading cause of death in >110 countries [1]. In China, the context for this study, cancer incidence and mortality have escalated, with an estimated 4.8 million new cancer cases and 3.2 million new cancer deaths in 2022, approximately 40% higher cancer mortality than in the United States [2]. Despite the crude cancer deaths, studies have found that globally 40% to 50% of cancers are preventable by choosing positive lifestyle factors, such as following a healthy diet, maintaining regular exercise and cancer screenings, and reducing tobacco use and alcohol consumption [3].

To promote cancer prevention, fostering positive coping beliefs is an essential step. However, many people still hold fatalistic

XSL•FO
RenderX

beliefs about cancer, considering it as neither preventable nor curable [4]. Those with fatalism contend that external forces, such as fate and predestination, control the causes and outcomes of cancer and hence deny the need to engage in any other form of coping [5]. Such maladaptive coping modes have been documented in both Eastern and Western societies, despite limited studies in Asia [6]. In China, cancer fatalism has long been prevalent, and it carries a negative connotation (eg, hopelessness and pessimism) that is associated with negative action tendencies in the face of cancer risks [7].

Hence, it is important to reduce cancer fatalism, and health information seeking may play a key role. Past research has documented the benefits of health information seeking, such as lowering health anxiety, managing health-related uncertainties, and increasing health literacy and confidence in fighting cancer [8]. However, people are not always active in searching for health information. Instead, some people intentionally avoid cancer information or prevent exposure to related topics, which is called cancer information avoidance (CIA) [9]. People who consistently avoid cancer information may miss opportunities to be empowered in making informed decisions to take positive coping behaviors. In past studies, CIA has shown to be associated with low levels of perceived behavioral control, cancer knowledge, and delays in seeking help [10,11]. Although the detrimental impacts of CIA on maladaptive coping have been suggested, prior research predominantly concentrated on health information seeking, inadvertently overlooking the simultaneous examination of information seeking and CIA as distinct appraisals within the context of cancer fatalism development from the theoretical lens of the extended parallel process model (EPPM). This narrow focus hinders a comprehensive understanding of various information behaviors and their potentially varying implications on fatalistic beliefs concerning cancer prevention, particularly considering that CIA is more prevalent than avoidance of any other disease-related information given its threatening nature [12]. Thus, it is valuable to investigate how cancer fatalism may be influenced by 2 distinct appraisals—the danger control process related to information seeking and the fear control process related to CIA—which are believed to have apparently contrasting effects on cancer fatalism.

To investigate why some people actively seek cancer information on their own initiative while others choose to avoid it, one must take into account cancer-related affect and cognition, such as cancer worry and cancer risk perception [13]. Noticeably, cancer worry and cancer risk perception are distinct constructs, and they act in different ways in influencing people's information behavior [14]. However, how distinct cancer mental conditions are associated with different information behaviors (seeking vs avoidance), which further become associated with cancer fatalism, has not been addressed. As mentioned earlier, the EPPM provides a guiding theoretical framework for our examination, which demonstrates that in the face of a threat, individuals may engage in different information responses (adaptive vs maladaptive) based on their risk appraisals, which can further make a difference in outcome variables such as one's threat coping tendencies [15]. Considering that cancer fatalism involves individuals who deny their coping or behavioral needs [16-18], it is reasonable to expect that different information behaviors that individuals engaged in would be associated with different levels of fatalism, reflecting individuals' negative coping needs.

Apart from cancer-related mental conditions, people's information behavior is also influenced by information-carrier characteristics, for example, perceived cancer information usefulness, especially in the complex digital information environment [19]. In our study context, individuals would perceive the information to be useful if they deem the information can provide them with useful information or resources to deal with cancers. In this sense, perceived cancer information usefulness can be understood as a manifestation of response efficacy (eg, a belief as to whether a recommended response works in preventing a given threat) from the theoretical perspective of EPPM. Moreover, EPPM articulates that whether individuals engage in adaptive response (eg, information seeking) or maladaptive response (eg, information avoidance) depends on the interplay between threat appraisals (eg, severity and susceptibility) and efficacy appraisals (eg, response efficacy), suggesting the moderating role cancer information usefulness plays in our study.

The EPPM has traditionally been applied to elucidate how perceived self-threat influences an individual's coping tendencies in the context of explicit message persuasion. However, a growing body of research has expanded the EPPM's scope to include contexts beyond message persuasion, such as the incidental influence context [20]. While originally not designed for persuasion purposes, media coverage of health concerns has been found to incidentally influence variables relevant to public health, such as risk perceptions and effective responses [21,22]. This is not surprising, given that individuals need to be made aware of potential threats, and authorities are tasked with providing guidance on how to address them [23]. Consequently, the EPPM has been a valuable framework for application in nonpersuasion contexts to understand why and how people respond to health threats, often influenced by daily exposure to media reports containing persuasive health messages [23]. Hence, building upon the core tenets of the EPPM and drawing from prior empirical studies applying the EPPM to nonpersuasion contexts [24,25], one of the objectives of our study is to examine how individuals' subjective evaluation of a threat (ie, cancer) becomes associated with their coping responses via 2 appraisals in a nonpersuasion context. Within this context, the perceived threat is expected to be shaped by persuasive health messages that individuals encounter daily in the media. In this regard, it is important to note that our study does not seek to examine the effects of the intentionally crafted persuasive message on health outcomes (eg, attitude or behavioral change) or to test all the postulations of the EPPM. Instead, our focus centers on predicting individuals' coping responses through 2 appraisals (ie, danger control and fear control), which are grounded in their subjective evaluations of a threat and efficacy.

In light of the above, this study examines the path from 2 distinct cancer mental conditions (cancer worry and cancer risk perception) to 2 information behaviors (health information seeking and CIA) and further onward to cancer fatalism,

XSL•FO
RenderX

considering the moderating role of perceived cancer information usefulness (Figure 1). The next sections discuss the key concepts of this study and offer evidence for the proposed pathways.

Figure 1. Hypothesized model.

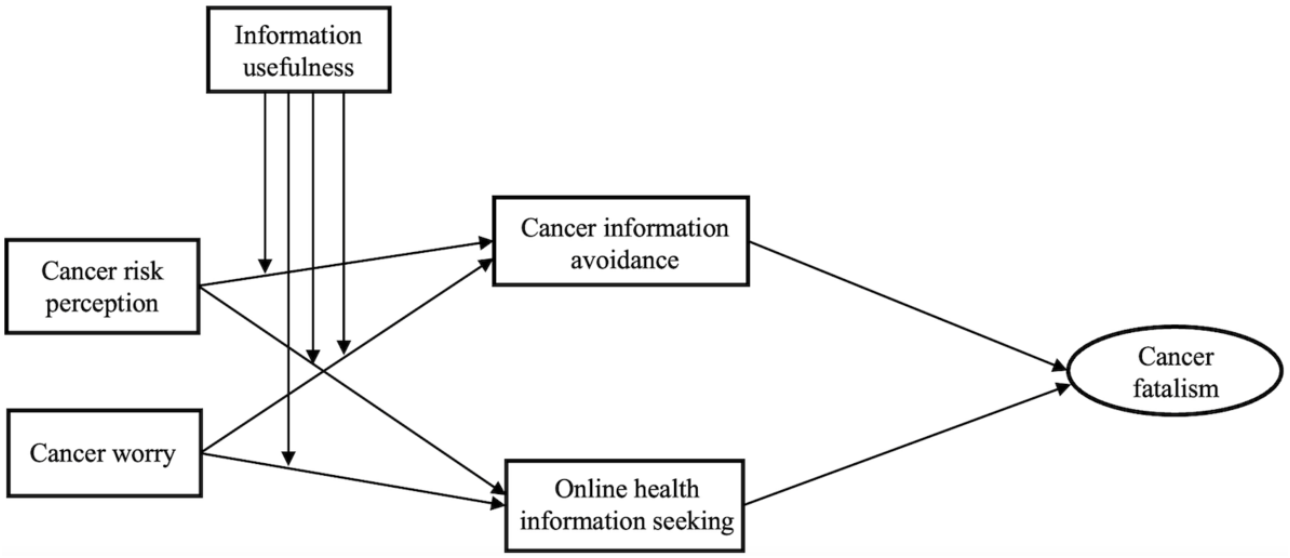

Study Hypotheses

Cancer Risk Perception, Cancer Worry, and Information Behavior

Cancer risk perception and cancer worry are salient cancer-related thoughts and feelings that have been frequently investigated. Specifically, cancer risk perception has been largely conceptualized as one's cognitive evaluation of perceived susceptibility to getting cancer, whereas cancer worry has been primarily regarded as an affective response to cancer threat [26,27]. In particular, Chae [14] developed a cancer-related mental condition model that differentiated cancer worry and cancer risk perception. She contended that cancer worry is a more affective condition compared to cancer risk perception, a more cognitive state. In other words, cancer worry is a mental activity that is closely linked to one's emotions (eg, anxiety and fear) triggered by cancer threats and thus an affective-cognitive condition. Cancer risk perception centers on one's rational judgment of the likelihood of getting cancer, which often involves deliberative and intellectual assessment and thus a cognitive appraisal.

Previous studies have documented ample evidence in linking cancer risk perception and cancer worry to health information seeking. For example, Nan et al [28] found that higher levels of cancer risk perception were associated with a greater likelihood of seeking prostate and breast cancer information. Yoo et al [29] indicated that people who perceived themselves with a high degree of getting cervical cancer were more likely to seek health information on social media. In the same vein, heightened cancer worry has been argued as a motivator for information acquisition. For instance, Griffin et al [30] demonstrated that personal worry prompts one's information needs in coping with health risks. The planned risk information seeking model [31] and its subsequent studies further confirmed the positive association between personal worry and searches for health information. Consistent with previous research, this study has the following hypotheses:

- Hypothesis 1: Cancer risk perception will be positively associated with online health information seeking (OHIS).
- Hypothesis 2: Cancer worry will be positively associated with OHIS.

Despite such motivational triggers, a growing body of research has made a seemingly competing argument, stating that cancer risk perception and cancer worry may lead to more CIA. For instance, Moser et al [32] found that cancer is a substantial threat to many people who consider it a death sentence, increasing their fear and anxiety. Under such circumstances, people refuse to be exposed to cancer-related information to reduce uncomfortable feelings [33]. This inhibiting role of cancer worry and risk perception is also elucidated by the EPPM [15], which demonstrates 2 appraisals people may adopt in dealing with threats. On the one hand, when people perceive a high appraisal of threat (eg, heightened risk perception), they may be activated to take adaptive actions (eg, information seeking) to better equip themselves in coping with the threat, known as the danger control process. On the other hand, people might choose defensive avoidance (eg, information avoidance) to escape the potential of eliciting negative emotions and feelings, known as the fear control process. In line with this notion of the EPPM, some people would engage in CIA in reducing unconformable feelings, especially when they perceive a high degree of cancer threats [34]. Several studies provide empirical evidence for this argument [11]. For example, Case et al [35] demonstrated that people tended to avoid or ignore threatening information to manage emotional states such as anxiety and fear. Vrinten et al [11] also found that CIA significantly increased, as cancer worry escalated. Hence, in light of prior literature, this study postulates the following hypotheses:

- Hypothesis 3: Cancer risk perception will be positively associated with CIA.

XSL•FO
RenderX

- Hypothesis 4: Cancer worry will be positively associated with CIA.

OHIS, CIA, and Fatalistic Beliefs About Cancer Prevention

Fatalism, a deterministic outlook that one's health is controlled by external forces, and therefore, there is no need to engage in positive coping behaviors, has been viewed as a prominent barrier to cancer prevention and screening behaviors [5,36,37]. By definition [5], cancer fatalism can be understood as one's negative behavioral tendency (eg, no need to cope and refusing coping behaviors) in the face of cancer threats [16,18]. Although some studies have approached cancer fatalism as a concept embedded in culture, primarily investigating its influence on information behaviors [12,38], we argue that cancer fatalism is a malleable concept that can be intervened through media learning and health education, such as information and knowledge acquisition from media use. In fact, numerous empirical studies have provided strong evidence of the positive impact of educational attainment and health literacy in reducing cancer fatalism [10,39-42]. These findings suggest that diverse information behaviors (ie, seeking and avoidance), involving varying levels of media exposure and educational opportunities, can make a significant difference in shaping the development of cancer fatalism. Thus, it is both reasonable and essential to examine the relationship from information behaviors to cancer fatalism. It is worth noting that both cancer-specific information seeking and general health information seeking are beneficial. While cancer-specific information seeking aids in gaining cancer-related knowledge, general health information seeking is effective in narrowing disparities in health literacy, thereby reducing cancer fatalism [43]. Particularly in this digital era, the internet offers convenient access to health information. With useful health information, patients have a better understanding of their health conditions, prescription drugs, treatments, and disease management options, which can empower them, reducing cancer fatalism [41,44]. Health information exchange with physicians or peers on the internet may also encourage individuals to take a more active role in preventive behaviors, lowering fatalistic beliefs about cancer [43]. By contrast, if people intentionally avoid cancer information, they might lose opportunities to receive information relevant to them, increasing health uncertainties and cancer fatalism [33].

Our reasoning is well aligned with the theoretical standpoint of EPPM, which demonstrates that the 2 information responses (adaptive vs maladaptive) that individuals adopt driven by their threat appraisals would lead to disparities in outcome variables such as one's threat coping tendency. Contextualized in this study, individuals who take adaptive actions in engaging in health information seeking tend to be well equipped with cancer-related knowledge, which in turn helps eliminate their fatalistic belief about cancer prevention, whereas individuals who choose defensive steps in engaging in information avoidance are more likely to be vulnerable to cancer fatalism due to their refusing coping behaviors [18]. A couple of empirical studies have also illustrated that CIA can lead to fatalistic beliefs about cancer and less frequent cancer screenings [10,17]. Hence, based on prior literature, we proposed the following hypotheses:

- Hypothesis 5: OHIS is negatively associated with fatalistic beliefs about cancer prevention.
- Hypothesis 6: CIA is positively associated with fatalistic beliefs about cancer prevention.

So far, this study reviewed 2 well-established relationships that link 3 elements: cancer mental conditions, information behaviors, and cancer fatalism. Given the established 2-step relationship, an underlying dual pathway between cancer risk perception and cancer fatalism as well as between cancer worry and cancer fatalism is likely to be mediated by OHIS and CIA, which suggests the following hypotheses:

- Hypothesis 7: OHIS will mediate (1) the relationship between cancer risk perception and cancer fatalism and (2) the relationship between cancer worry and cancer fatalism.
- Hypothesis 8: CIA will mediate (1) the relationship between cancer risk perception and cancer fatalism and (2) the relationship between cancer worry and cancer fatalism.

Moderating Role of Perceived Usefulness of Online Cancer Information

Given the dynamic process of information seeking that involves interactions between information seekers and information platforms, we need to consider how information seekers perceive health information. Specifically, we investigated the moderating role of one's perceived information usefulness, a vital information-carrier predictor of individuals' information behavior [45]. Barbour et al [46] demonstrated that if people viewed health information as questionable and unclear, they tended to avoid such information to reduce stress and uncertainties despite their serious illnesses. A review study also concluded that the decision to seek or avoid cancer information was contingent upon situational factors, such as the usefulness of the information [33]. As Johnson [45] posited, information seekers are concerned about the content of the information. They put greater effort into seeking information that is deemed useful in coping with their cancer threats. Conversely, if they consider the information to be less effective, they may have a higher tendency to avoid it.

Moreover, echoing the EPPM [15], engaging in fear (information avoidance) or danger control (information seeking) is a synergistic effect of 2 appraisals: threat (eg, severity and susceptibility) and efficacy (eg, response efficacy and self-efficacy). Specifically, response efficacy refers to the perception of whether the provided information or recommended response works in allaying the threat [34]. Particularly relevant to the information environment, useful cancer information is a typical resource that offers people informational and emotional strategies to cope with threats [47]. Therefore, conceptualizing the usefulness of cancer information as a manifestation of response efficacy, it is expected that the relationship between one's cancer-related mental conditions and information behavior will be moderated by the perceived usefulness of cancer information from the theoretical perspective of EPPM. Accordingly, the following hypothesis is posited:

- Hypothesis 9: The perceived usefulness of online cancer information will moderate (1) the relationship between cancer risk perception and OHIS, (2) the relationship

XSL•FO
RenderX

between cancer risk perception and CIA, (3) the relationship between cancer worry and OHIS, and (4) the relationship between cancer worry and CIA.

Methods

Data and Participants

This study used cross-sectional data from the Health Information National Trends Survey (HINTS) in China (HINTS-China). Similar to the HINTS that has been implemented in the United States since 2003, the current HINTS-China was conducted in May 2017. HINTS-China is an international collaboration involving the National Cancer Institute, the Chinese Ministry of Health, and the Chinese Food and Drug Administration, in conjunction with George Mason University. It was initially established with Renmin University of China and has subsequently collaborated with Beijing Normal University [48]. A multistage stratified random sampling method was adopted, and a door-to-door survey method was used. Specifically, 2 cites were purposely chosen due to their representativeness: Beijing (representing a tier-1 city) and Hefei (representing a tier-2 city). Then, 1 urban district (representing an urban area) and 1 rural county (representing a rural area) were randomly selected in each of the 2 cities. Within each urban district and rural county, 1 subdistrict and township was randomly selected from 3 strata by the level of economic development (high, medium, and low). A total of 4 residential neighborhoods were then randomly selected from each subdistrict and township stratified by sex and age (for detailed sampling methodology, refer to the study by Zhao et al [49]).

A total of 3090 respondents completed the survey. In this study, we only included those who had internet access, as 1 focal variable was OHIS. In addition, patients with cancer were excluded from our sample because 1 key variable, cancer risk perception, measured people's evaluation of the likelihood of getting cancer. Therefore, the final sample size in this study was 2358. The participants' mean age was 33.98 (SD 10.88, range 18 to 60) years. In total, 60.3% (1421/2358) were female. More than half of the participants (1332/2358, 56.49%) obtained some college education or more. Less than a third (705/2358, 29.9%) earned monthly income >CNY 5000 (US$700). Approximately 94.44% (2227/2358) of the respondents had health insurance coverage, and 16% (377/2358) had at least 1 chronic condition. The average self-reported health condition was at the "good" level (mean 3.98, SD 0.76).

Ethical Considerations

The HINTS-China was approved by the institutional review board (IRB) at Beijing Normal University in 2017. Respondents who participated in the survey gave their written consent. The data were deidentified and publicly available [50]. Secondary data analysis using the HINTS-China data set in our study did not need to obtain IRB approval because research involving the study of existing data, if these sources are publicly available or research participants cannot be identified, is in the exemption category of IRB [51]. This is also a common practice in prior research using HINTS-China data [50,52].

Measures

Cancer Risk Perception

Drawing from prior research examining cancer risk perception [14,53], respondents were asked to indicate their judgment of the likelihood of getting cancer on a 5-point Likert scale (1=*very unlikely* and 5=*very likely*): "Compared to the average person of your age and sex, how likely would you rate your chance of developing cancer sometime in your life?" (mean 2.32, SD 0.84).

Cancer Worry

Similar to prior studies using HINTS data [54], this study used a single item to ask participants to indicate to what extent they worried about getting cancer on a 5-point Likert scale (1=*not at all* and 5=*extremely*): "How worried are you about getting cancer?" (mean 2.25, SD 1.00).

Perceived Usefulness of Cancer Information

Adapted from prior research that used a single item by using HINTS data [55], in examining to what extent respondents considered online cancer information to be useful, participants were asked to rate the overall usefulness of online cancer information on a 4-point Likert scale (1=*not at all useful* and 4=*very useful*; mean 2.35, SD 0.68).

OHIS Measure

To investigate the extent to which respondents sought general health information on the internet, we used six items, drawn from previous research [43], that asked participants whether they have carried out the following activities on the internet in the last 12 months (1=*yes* and 0=*no*): (1) looked for a health care provider or information about hospitals, (2) looked for exercise, weight control, or fitness information, (3) looked for information about quitting smoking, (4) looked for health or medical information for someone else, (5) asked and exchanged health-related information and topics, and (6) downloaded health or medical information. A summative scale of these 6 dichotomous items was created (mean 1.03, SD 0.99).

CIA Measure

To estimate the extent to which people intentionally avoid cancer information, participants were asked to report their agreement with the following statement on a 5-point scale (1=*strongly disagree* and 5=*strongly agree*) adapted from prior research [9]: "I avoid being exposed to cancer information" (mean 2.76, SD 0.98).

Fatalistic Beliefs About Cancer Prevention

Drawing from previous studies using HINTS data [37,41], participants were asked to evaluate their agreement with 3 statements about fatalistic beliefs concerning cancer prevention on a 5-point Likert scale (1=*strongly disagree* and 5=*strongly agree*): (1) "There is not much I can do to lower my chances of getting cancer," (2) "It seems that everything causes cancer," and (3) "When I think about cancer, I automatically think about death." (mean 3.16, SD 0.74; Cronbach α=0.74).

Control variables included social demographics such as age, sex, education, and personal monthly income. In addition, as

XSL•FO
RenderX

this study investigated people's cancer-related beliefs, health-related variables were also controlled, including participants' general health status (1=*very poor* and 5=*very good*), chronic disease conditions (1=*yes* and 0=*no*), health insurance coverage (1=*yes* and 0=*no*), and family cancer history (1=*yes* and 0=*no*).

Analytic Approach

To examine the hypothesized model, structural equation modeling was conducted using the lavaan package in R. Maximum likelihood of estimation was adopted. Following the widely used combinational rules and prior research [56], the goodness of fit of the hypothesized model should be (1) Tucker-Lewis index or comparative fit index≥0.95 and standardized root mean square residual (SRMR) ≤0.08, or, alternatively, (2) root mean square error of approximation<0.05 and SRMR <0.06. To assess the mediating effect (ie, hypothesis 7 and hypothesis 8), the bias-corrected bootstrapping method was used to estimate the path CI [57]. A CI that does not include 0 indicates a statistically significant mediating effect at the 95% CI. To examine the moderating effect of the perceived usefulness of cancer information (ie, hypothesis 9), interaction terms between independent variables (ie, cancer risk perception and cancer worry) and the perceived usefulness of cancer information were created, and the 3 variables were centered before forming the interaction terms to reduce multicollinearity problem.

Results

Structural Model and Path Coefficients

Table 1 shows the descriptive statistics and bivariate correlations for measured variables. Controlling for social demographics and health-related variables, the structural model showed an acceptable fit (χ^2_{92}=254.4; *P*<.001; comparative fit index=0.95; Tucker-Lewis index=0.94; root mean square error of approximation=0.05; 90% CI 0.041 to 0.053; SRMR=0.04). As shown in Figure 2, our findings revealed that cancer risk perception positively predicted OHIS (β=.08; *P*=.007). Similarly, cancer worry was positively related to OHIS (β=.10; *P*<.001), supporting hypothesis 1 and hypothesis 2. In addition, cancer worry was positively associated with CIA (β=.11; *P*<.001), whereas the results indicated a nonsignificant relationship between cancer risk perception and CIA (β=–.03; *P*=.23). Hence, hypothesis 4 was supported but not hypothesis 3. Moreover, the results showed that CIA was positively associated with fatalistic beliefs about cancer prevention (β=.29; *P*<.001); conversely, OHIS was negatively related to cancer fatalism (β=–.09; *P*=.003), supporting hypothesis 5 and hypothesis 6. In total, our hypothesized model explained 25.1% of the variance of cancer fatalism among the participants.

XSL•FO
RenderX

Table 1. Descriptive statistics and correlation analysis (Person r and 2-tailed *P* value) among measured variables (N=2358).

Variable	Cancer risk perception	Cancer worry	Perceived usefulness of online cancer information	Online health information seeking	Cancer information avoidance	Fatalistic beliefs about cancer prevention
Cancer risk perception (mean 2.32, SD 0.84)						
r	1	0.51	0.10	0.15	0.05	0.12
P value	—[a]	<.001	<.001	<.001	.01	<.001
Cancer worry (mean 2.25, SD 1.00)						
r	0.51	1	0.12	0.16	0.11	0.15
P value	<.001	—	<.001	<.001	<.001	<.001
Perceived usefulness of online cancer information (mean 2.35, SD 0.68)						
r	0.10	0.12	1	0.18	0.03	0.15
P value	<.001	<.001	—	<.001	.18	<.001
Online health information seeking (mean 1.03, SD 0.99)						
r	0.15	0.16	0.18	1	−0.01	−0.08
P value	<.001	<.001	<.001	—	.63	<.001
Cancer information avoidance (mean 2.76, SD 0.98)						
r	0.05	0.11	0.03	−0.01	1	0.27
P value	.01	<.001	.18	.63	—	<.001
Fatalistic beliefs about cancer prevention (mean 3.15, SD 0.80)						
r	0.12	0.15	0.15	−0.08	0.27	1
P value	<.001	<.001	<.001	<.001	<.001	—
Age (mean 33.98, SD 10.88 years)						
r	0.07	0.08	0.10	0.01	0.20	0.27
P value	.001	<.001	<.001	.50	<.001	<.001
Sex (female=1 and male=0)						
r	0.01	0.04	0.08	0.07	0.04	0.08
P value	.76	.07	<.001	.006	.05	<.001
Education (mean 3.74, SD 1.13)						
r	−0.01	−0.01	−0.04	0.11	−0.10	−0.19
P value	.59	.78	.05	<.001	<.001	<.001
Personal income (mean 5.51, SD 1.94)						
r	0.01	0.03	0.01	0.01	0.13	0.11
P value	.63	.22	.77	.94	<.001	<.001
Health status (mean 3.98, SD 0.76)						
r	−0.33	−0.29	−0.08	−0.10	−0.06	−0.05
P value	<.001	<.001	<.001	<.001	.002	.02
Chronic disease (yes=1 and no=0)						
r	0.12	0.12	0.04	0.07	0.08	0.12
P value	<.001	<.001	.04	.001	<.001	<.001
Family cancer history (yes=1 and no=0)						
r	0.14	0.15	0.12	0.06	0.05	0.10
P value	<.001	<.001	<.001	.001	.02	<.001
Health insurance (yes=1 and no=0)						

XSL•FO
RenderX

Variable	Cancer risk perception	Cancer worry	Perceived usefulness of online cancer information	Online health information seeking	Cancer information avoidance	Fatalistic beliefs about cancer prevention
r	−0.01	0.03	0.03	0.03	0.01	0.06
P value	.49	.19	.18	.23	.92	.001

[a]Not applicable.

Figure 2. Final model with standardized path coefficients. Dotted lines indicate nonsignificant relationships. The covariances between all exogenous factors (eg, controls) and coefficients with control variables are not presented for the purpose of clarity. **P*<.05, ***P*<.01, ****P*<.001.

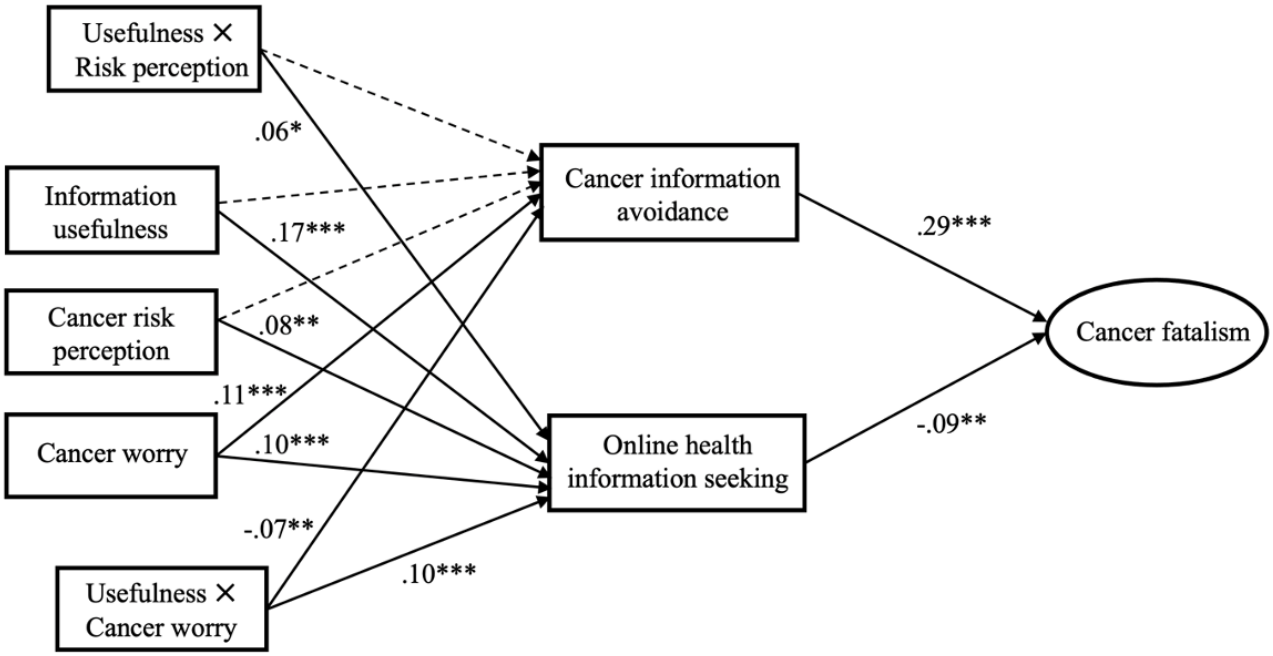

Mediation and Moderated Mediation

To assess the mediating effect, the bias-corrected bootstrapping method was used to estimate the path CI. The results of bootstrapped CIs, with 5000 resamples, showed that cancer risk perception indirectly reduced cancer fatalism through OHIS (95% CI –0.010 to –0.003) but not through CIA (95% CI –0.014 to 0.003). In addition, the results supported an indirect effect of cancer worry on cancer fatalism, as mediated by OHIS (95% CI –0.012 to –0.004]) and CIA (95% CI 0.009 to 0.025). Hence, hypothesis 7 and hypothesis 8b were supported but not hypothesis 8a.

As for the moderating effects, the results revealed that there was a main effect of the perceived usefulness of online cancer information on OHIS (β=.17; *P*<.001) but not for CIA (β=−.01; *P*=.58). More importantly, there was a significant interaction effect between cancer risk perception and perceived usefulness in predicting OHIS (β=.06; *P*=.01). The results revealed that the simple slope of the relationship between cancer risk perception and OHIS differed significantly from 0 when the perceived usefulness of cancer information was 1 SD above the mean (β=0.14; SE=0.05; *P*<.001) but not 1 SD below (β=0.06; SE=0.04; *P*=.11). This indicates that the positive association between cancer risk perception and OHIS was salient only among participants who perceived online cancer information to be of high usefulness but not among those who deemed the information was of low usefulness (Figure 3). However, there was no significant interaction effect between cancer risk perception and perceived usefulness in predicting CIA (β=−.04; *P*=.18).

XSL•FO
RenderX

Figure 3. Moderating effect of perceived usefulness of online cancer information on the relationship between cancer risk perception and online health information seeking; the higher value is 1 SD above the mean, and the lower value is 1 SD below the mean.

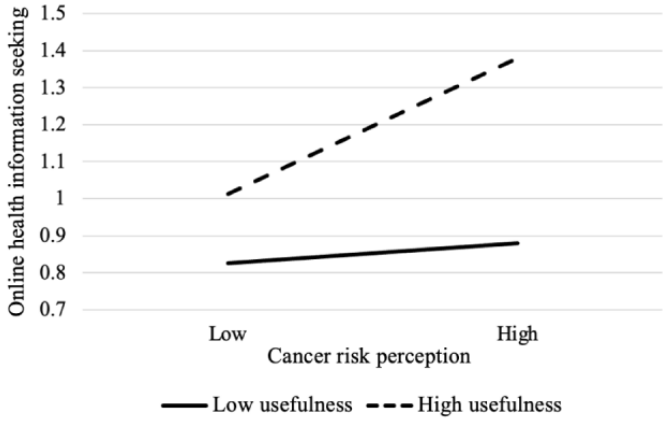

Moreover, a significant interaction between cancer worry and information usefulness was observed in predicting OHIS (β=.10; *P*<.001). The results of the simple slopes revealed that when online cancer information was perceived as of high usefulness, worried participants frequently acquired health information on the web (B=0.18; SE=0.04; *P*<.001). However, this conditional effect of cancer worry was not observed when online cancer information was perceived as of low usefulness (B=0.02; SE=0.04; *P*=.61; Figure 4).

Figure 4. Moderating effect of perceived usefulness of online cancer information on the relationship between cancer worry and online health information seeking; the higher value is 1 SD above the mean, and the lower value is 1 SD below the mean.

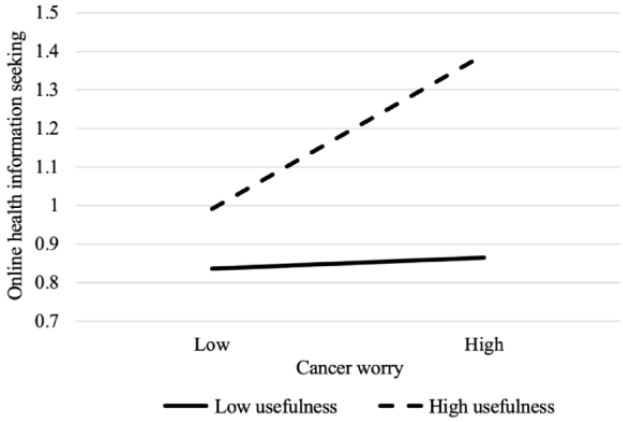

Furthermore, a significant interaction between cancer worry and perceived usefulness was detected in predicting CIA (β=−.07; *P*=.009). Specifically, the positive association between cancer worry and CIA existed only for people who rated the usefulness of online cancer information as low (B=0.16; SE=0.04; *P*<.001) but not for those who scored the usefulness as high (B=0.06; SE=0.04; *P*=.09; Figure 5). In sum, hypotheses 9a, 9c, and 9d were supported but not hypothesis 9b.

Figure 5. Moderating effect of perceived usefulness of online cancer information on the relationship between cancer worry and cancer information avoidance; the higher value is 1 SD above the mean, and the lower value is 1 SD below the mean.

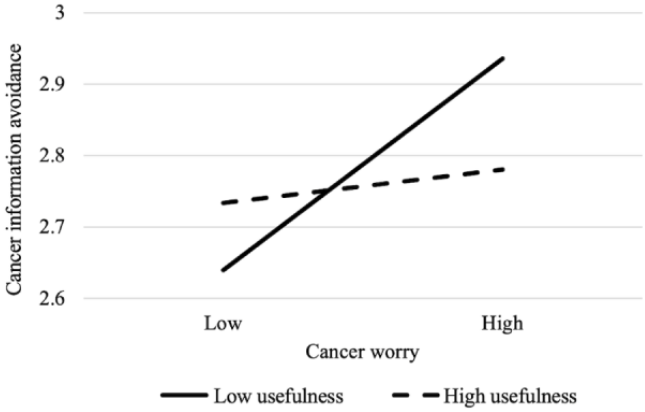

XSL•FO
RenderX

The results also displayed significant moderated mediation effects (Table 2). The perceived high usefulness of cancer information strengthened the indirect negative influence of cancer risk perception and cancer worry on cancer fatalism through OHIS. However, a lower level of perceived usefulness significantly intensified the indirect positive influence of cancer worry on cancer fatalism through CIA.

Table 2. Conditional indirect effects of cancer risk perception and cancer worry on cancer fatalism through online health information seeking (OHIS) and cancer information avoidance (CIA) at different levels of perceived cancer information usefulness. B represents the unstandardized coefficient; italicized values indicate significant effects.

Independent variables (cancer risk perception and cancer worry), mediators (OHIS and CIA), and moderator (low, mean, and high perceptions of cancer information usefulness)	Dependent variable: cancer fatalism		
	B	SE	Bootstrap, 95% CI
Cancer risk perception			
OHIS			
Low (mean–1 SD)	–0.003	0.002	–0.008 to 0.000
Mean	*–0.005*	0.003	*–0.012* to *–0.001*
High (mean+1 SD)	*–0.010*	0.003	*–0.018* to *–0.002*
CIA			
Low (mean–1 SD)	0.011	0.006	–0.002 to 0.023
Mean	0.009	0.005	–0.001 to 0.019
High (mean+1 SD)	–0.004	0.007	–0.017 to 0.013
Cancer worry			
OHIS			
Low (mean–1 SD)	–0.002	0.002	–0.005 to 0.002
Mean	*–0.007*	0.003	*–0.014* to *–0.001*
High (mean+1 SD)	*–0.011*	0.004	*–0.019* to *–0.004*
CIA			
Low (mean–1 SD)	*0.035*	0.007	*0.022* to *0.047*
Mean	*0.021*	0.005	*0.011* to *0.030*
High (mean+1 SD)	0.010	0.006	0.000 to 0.024

Discussion

Principal Findings

This study reveals a dual-mediation pathway linking distinct cancer mental conditions to cancer fatalism, focusing on different information behaviors and considering the moderating role of the perceived usefulness of online cancer information. Findings from the HINTS-China data revealed that cancer risk perception and cancer worry were positively associated with OHIS. Consistent with previous studies [28,30], individuals who perceived a high susceptibility to getting cancer and felt worried about it tended to actively engage in OHIS, such as looking for exercise, weight control, or fitness information and exchanging health-related information on the internet. As such, the results suggest that both affective-cognitive (cancer worry) and cognitive (cancer risk perception) mental cognitions can serve as driving forces for people's self-protective behaviors, such as health information acquisition.

However, the results indicated a different relationship between cancer worry and CIA when compared to risk perception and CIA, such that cancer worry rather than risk perception was positively associated with CIA. This finding might suggest that unlike risk perception, which has been widely noted as a problem-solving mechanism that leads to active information seeking [13,31], cancer worry tends to increase both general health information seeking and cancer-specific information avoidance, with mixed findings in the literature [58-60]. On the one hand, the finding that cancer worry was positively associated with both OHIS and CIA suggests the operation of moderating factors (eg, message characteristics) that facilitate seeking behaviors in some circumstances but avoidance actions in other contexts. On the other hand, psychologically, cancer worry is closely related to negative emotions, such as fear and anxiety. As noted by uncertainty management theory, information avoidance serves as a way of managing uncertainty and providing an escape from negative emotions [61]. This avoidance behavior tends to be more pronounced when confronting threatening and complex cancer information that may bring about more confusion and mental discomfort, even though it might compromise treatment. Hence, the results highlight that cancer risk perception and cancer worry should not be treated as homogeneous entities or used interchangeably because of their varying functions and effects.

XSL•FO
RenderX

Our study also found that OHIS was negatively associated with cancer fatalism, while CIA was positively related to it. In accord with previous studies [8,41], health information seeking, particularly via the internet, offers people diverse formats and depths of information across various health topics, helps specify a diagnosis or treatment plan, and provides clarity about prognoses. All these outcomes contribute to individuals' increased health literacy and cancer knowledge, which are critical in reducing individuals' negative coping needs embedded in cancer fatalism. In addition, OHIS offers people more opportunities to interact with others in social media communities and support groups, providing a broad sense and social proof that many others are active in engaging in self-protective behaviors for cancer prevention [43]. These perceptions help reduce people's cancer fatalism, especially in societies that tend toward collectivism, such as China. In contrast to this study's findings about OHIS, the finding of a positive association between CIA and cancer fatalism implies a detrimental influence of CIA on cancer prevention. Consistent with previous studies [10], people who refused to be exposed to cancer information delayed the discovery of positive information, thus maintaining their biased perceptions of their actual risks and self-agency. This biased belief is closely related to individuals' tendencies to avoid physicians, other forms of help, and preventive screening [60,62]. These behaviors exacerbate individuals' health risks, especially for those who are vulnerable to cancer for whom early detection is quite literally a life-or-death matter.

Another key finding pertains to the moderation effect. The results indicate that people only sought health information on the internet when they perceived it to be useful. If they deemed information to be useless, they tended to avoid it despite their cancer worries. Such results confirm the central postulate of the updated EPPM—the additive model—which suggests that higher levels of each threat and efficacy would lead to a more favorable impact, and the interaction effects between the threat and efficacy are additive [34]. In addition, the mediating effect of people's information behavior on cancer fatalism was found to be contingent upon perceived information usefulness. This finding is consistent with the 3-stage model developed by Street [63], which highlights the vital role that positive experiences play in producing desired health outcomes out of user-media-message interactions. Particularly in the context of China, researchers have long questioned the problematic digital information environment and expressed concern about the negative influences of poor-quality health information, which are exacerbated by low levels of health literacy [64]. Therefore, positive media message characteristics (eg, information usefulness) are particularly important to encourage people to engage with more adaptive information behavior to better reap health benefits and combat cancers. Useless and low-quality cancer information may make people frustrated and overwhelmed, dampening their information seeking and even spurring CIA that leads to cancer fatalism. Hence, the results reinforce a challenging but imperative public health goal of providing more useful, understandable, and high-quality cancer information for people in China, especially in the digital era.

Theoretical and Practical Implications

This study has contributed new insights to inform future research on health-related information behavior and the EPPM. First, in contrast to some previous studies that primarily focused either on information seeking behavior or information avoidance, a strength of this study is that it considers both information-related behaviors, which are of equal importance in understanding the development of fatalistic beliefs about cancer prevention. More to the point, this study broadens the scope of the EPPM by incorporating cancer fatalism, which reflects individuals' negative behavioral coping tendencies, as a fear control response, and exploring its connection with both OHIS and CIA. This expansion helps elucidate the differing implications of these 2 distinct appraisals on fatalistic beliefs concerning cancer prevention. Second, conceptualizing the perceived usefulness of online cancer information as one of the manifestations of response efficacy, this study adds a new perspective to the EPPM and the literature on health-related information management. Third, building upon the cancer-related mental condition model [14], this study has taken a step further to investigate how distinct cancer mental conditions influence disparate information behavior differently, which contributes to the theoretical advancement of the effects of cancer-related affective responses and cognitive thoughts on cancer communication. In addition, Witte [65] has demonstrated that the "danger control processes are primarily cognitive processes," whereas the "fear control processes are mainly emotional processes." By establishing the positive relationship between cancer worry (an affective-cognitive condition) and OHIS (a danger control process) as well as the positive relationship between cancer worry and CIA (a fear control process), this study contributes to the EPPM by highlighting the dual nature of cancer worry in engaging the 2 different appraisals proposed by the EPPM. This paves the way for future EPPM research to thoroughly explore how various cancer-related mental conditions (eg, affective, cognitive, and affective-cognitive) may either motivate or inhibit individuals in safeguarding themselves against threats such as cancer as well as by which conditions. This is particularly significant as the EPPM has traditionally focused on purely emotional appeals (eg, fear) or cognitions (eg, perceived susceptibility and perceived severity).

The findings also provide useful implications for cancer communication and control. First, cancer worry has both positive and negative influences in our model; thus, developers of future health campaigns aimed at increasing people's risk perceptions should be cautious about unintended outcomes. They must be vigilant in avoiding negative affective responses toward cancer threats to avoid eliciting excessive cancer worry that provokes an "ostrich effect." More tailored communication strategies are needed to promote rational thinking about cancer and personal risk, avoid inflating anxiety, and avert CIA and possible anxiety disorders [58]. Second, in clinical settings, it would be useful for physicians to identify and pay special attention to patients with high trait anxiety. Practitioners should help them attenuate unnecessary worries and anxiety through affectionate and personalized counseling. Instead of information avoidance, training in new coping skills (eg, breathing exercises, relaxation

XSL•FO
RenderX

strategies, and mental imagery exercises) should be incorporated into health education and counseling. Given the moderating role of perceived information usefulness and the effectiveness of OHIS in reducing cancer fatalism, health educators are encouraged to disseminate useful, accurate, and feasible information with concrete skill sets that are easy and effective in fighting cancer threats. Furthermore, considering the potential of OHIS to alleviate cancer fatalism, public health practitioners need to make efforts to promote information seeking behavior that informs and empowers people, particularly for certain groups who are vulnerable to cancer fatalism (eg, those with low educational attainment or low health literacy).

Limitations and Suggestions for Future Research

Several limitations are worth examining more closely in future research. First, the use of cross-sectional data precluded the causal inferences in this study. To determine causality, longitudinal studies with panel data are encouraged to affirm the temporal order. Second, due to the use of secondary data, cancer worry, information usefulness, and CIA were assessed with single items. Although these measures have been frequently used in previous studies [28,54,55], future research would ideally use multiple-item scales to enhance content validity. Third, this study did not directly investigate what types of external stimuli can trigger individuals' threats but only examined the relationship between individuals' perceived cancer threats and subsequent coping responses. To expand upon the EPPM, it is essential for future research to use experimental methods to evaluate the message characteristics that can effectively induce adequate levels of risk perceptions, thereby encouraging adaptive actions. This endeavor holds significant promise for advancing our theoretical understanding of how various persuasive messages, including those designed to induce fear, are processed within the theoretical framework of the EPPM.

Moreover, given the specific scope of this study, our research model exclusively examined the relationship between 2 appraisals and a fear control outcome (ie, cancer fatalism), without delving into potential danger control outcomes, such as changes in belief, attitude, and behaviors concerning cancer prevention. Future research is encouraged to incorporate potential manifestations of danger control in providing a more comprehensive understanding of both fear control and danger control outcomes and their relationship with appraisals of threat and efficacy. Finally, cancer fatalism is a multidimensional construct that has been conceptualized differently across the cancer continuum [5]. A direct extension of this study would be to include other aspects of fatalism, such as fatalistic beliefs about cancer treatments among survivors of cancer who are receiving treatments.

Conclusions

Cancer is a threatening health problem and becoming an increasing burden on a global scale. Although a great proportion of cancer cases can be prevented and cured, cancer fatalism is one of the major obstacles for cancer prevention and control. This study provides evidence that OHIS is an effective mechanism for reducing cancer fatalism and minimizing CIA is necessary to allay fatalistic beliefs about cancer prevention. To facilitate healthy behavior, apart from disseminating more useful cancer information that assists people in coping with cancer threats, future endeavors should heighten rational risk perception while being cautious about elevating unnecessary cancer worry that may prompt information avoidance.

Acknowledgments

The authors are grateful to the Health Information National Trends Survey in China research program for sharing the data used for this study. This research was supported by the Shanghai Philosophy and Social Sciences Program (grant 2022ZXW005).

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

LZ and SJ conceptualized the study. LZ conducted data analysis and wrote the original draft. LZ and SJ revised the manuscript. Both authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Xia C, Dong X, Li H, Cao M, Sun D, He S, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl) 2022 Feb 09;135(5):584-590 [FREE Full text] [doi: 10.1097/CM9.0000000000002108] [Medline: 35143424]
3. Soerjomataram I, Bray F. Planning for tomorrow: global cancer incidence and the role of prevention 2020-2070. Nat Rev Clin Oncol 2021 Oct;18(10):663-672. [doi: 10.1038/s41571-021-00514-z] [Medline: 34079102]

XSL•FO
RenderX

4. Niederdeppe J, Fowler EF, Goldstein K, Pribble J. Does local television news coverage cultivate fatalistic beliefs about cancer prevention? J Commun 2010 Jun 01;60(2):230-253 [FREE Full text] [doi: 10.1111/j.1460-2466.2009.01474.x] [Medline: 20563221]
5. Powe BD, Finnie R. Cancer fatalism: the state of the science. Cancer Nurs 2003 Dec;26(6):454-65; quiz 466-7. [doi: 10.1097/00002820-200312000-00005] [Medline: 15022977]
6. Jun J, Oh KM. Asian and Hispanic Americans' cancer fatalism and colon cancer screening. Am J Health Behav 2013 Mar;37(2):145-154. [doi: 10.5993/AJHB.37.2.1] [Medline: 23026095]
7. Cheng H, Sit JW, Twinn SF, Cheng KK, Thorne S. Coping with breast cancer survivorship in Chinese women: the role of fatalism or fatalistic voluntarism. Cancer Nurs 2013;36(3):236-244. [doi: 10.1097/NCC.0b013e31826542b2] [Medline: 22964870]
8. Lee CJ, Niederdeppe J, Freres D. Socioeconomic disparities in fatalistic beliefs about cancer prevention and the internet. J Commun 2012 Dec;62(6):972-990 [FREE Full text] [doi: 10.1111/j.1460-2466.2012.01683.x] [Medline: 25530627]
9. Miles A, Voorwinden S, Chapman S, Wardle J. Psychologic predictors of cancer information avoidance among older adults: the role of cancer fear and fatalism. Cancer Epidemiol Biomarkers Prev 2008 Aug;17(8):1872-1879 [FREE Full text] [doi: 10.1158/1055-9965.EPI-08-0074] [Medline: 18708374]
10. Emanuel AS, Kiviniemi MT, Howell JL, Hay JL, Waters EA, Orom H, et al. Avoiding cancer risk information. Soc Sci Med 2015 Dec;147:113-120 [FREE Full text] [doi: 10.1016/j.socscimed.2015.10.058] [Medline: 26560410]
11. Vrinten C, Boniface D, Lo SH, Kobayashi LC, von Wagner C, Waller J. Does psychosocial stress exacerbate avoidant responses to cancer information in those who are afraid of cancer? A population-based survey among older adults in England. Psychol Health 2018 Jan;33(1):117-129 [FREE Full text] [doi: 10.1080/08870446.2017.1314475] [Medline: 28391710]
12. Chae J, Lee CJ, Kim K. Prevalence, predictors, and psychosocial mechanism of cancer information avoidance: findings from a national survey of U.S. adults. Health Commun 2020 Mar;35(3):322-330. [doi: 10.1080/10410236.2018.1563028] [Medline: 30606065]
13. Griffin RJ, Dunwoody S, Neuwirth K. Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. Environ Res 1999 Feb;80(2 Pt 2):S230-S245. [doi: 10.1006/enrs.1998.3940] [Medline: 10092438]
14. Chae J. A three-factor cancer-related mental condition model and its relationship with cancer information use, cancer information avoidance, and screening intention. J Health Commun 2015;20(10):1133-1142. [doi: 10.1080/10810730.2015.1018633] [Medline: 26161844]
15. Witte K. Putting the fear back into fear appeals: the extended parallel process model. Commun Monogr 1992;59(4):329-349. [doi: 10.1080/03637759209376276]
16. Hong SJ. Linking environmental risks and cancer risks within the framework of genetic-behavioural causal beliefs, cancer fatalism, and macrosocial worry. Health Risk Soc 2020 Nov 29;22(7-8):379-402. [doi: 10.1080/13698575.2020.1852535]
17. Ramondt S, Ramírez AS. Fatalism and exposure to health information from the media: examining the evidence for causal influence. Ann Int Commun Assoc 2017;41(3-4):298-320 [FREE Full text] [doi: 10.1080/23808985.2017.1387502] [Medline: 34307882]
18. Rippetoe PA, Rogers RW. Effects of components of protection-motivation theory on adaptive and maladaptive coping with a health threat. J Pers Soc Psychol 1987;52(3):596-604. [doi: 10.1037/0022-3514.52.3.596]
19. Johnson JD, Case DO. Health Information Seeking. New York, NY: Peter Lang Publishing Group; 2012.
20. Goodall C, Sabo J, Cline R, Egbert N. Threat, efficacy, and uncertainty in the first 5 months of national print and electronic news coverage of the H1N1 virus. J Health Commun 2012;17(3):338-355. [doi: 10.1080/10810730.2011.626499] [Medline: 22188164]
21. Slater MD, Goodall CE, Hayes AF. Self-reported news attention does assess differential processing of media content: an experiment on risk perceptions utilizing a random sample of US local crime and accident news. J Commun 2009 Mar;59(1):117-134. [doi: 10.1111/j.1460-2466.2008.01407.x]
22. Slater MD, Hayes AF, Ford VL. Examining the moderating and mediating roles of news exposure and attention on adolescent judgments of alcohol-related risks. Commun Res 2007;34(4):355-381. [doi: 10.1177/0093650207302783]
23. Goodall CE, Reed P. Threat and efficacy uncertainty in news coverage about bed bugs as unique predictors of information seeking and avoidance: an extension of the EPPM. Health Commun 2013;28(1):63-71. [doi: 10.1080/10410236.2012.689096] [Medline: 23330859]
24. Ivanova A, Kvalem IL. Psychological predictors of intention and avoidance of attending organized mammography screening in Norway: applying the Extended Parallel Process Model. BMC Womens Health 2021 Feb 15;21(1):67 [FREE Full text] [doi: 10.1186/s12905-021-01201-y] [Medline: 33588809]
25. Roberto AJ, Goodall CE. Using the extended parallel process model to explain physicians' decisions to test their patients for kidney disease. J Health Commun 2009 Jun;14(4):400-412. [doi: 10.1080/10810730902873935] [Medline: 19466650]
26. Borkovec TD. The nature, functions, and origins of worry. In: Davey GC, Tallis F, editors. Worrying: Perspectives on Theory, Assessment and Treatment. Hoboken, NJ: John Wiley & Sons; 1994.
27. McQueen A, Vernon SW, Meissner HI, Rakowski W. Risk perceptions and worry about cancer: does gender make a difference? J Health Commun 2008;13(1):56-79. [doi: 10.1080/10810730701807076] [Medline: 18307136]

XSL•FO
RenderX

28. Nan X, Underhill J, Jiang H, Shen H, Kuch B. Risk, efficacy, and seeking of general, breast, and prostate cancer information. J Health Commun 2012;17(2):199-211. [doi: 10.1080/10810730.2011.585690] [Medline: 22026448]
29. Yoo SW, Kim J, Lee Y. The effect of health beliefs, media perceptions, and communicative behaviors on health behavioral intention: an integrated health campaign model on social media. Health Commun 2018 Jan;33(1):32-40. [doi: 10.1080/10410236.2016.1242033] [Medline: 27858470]
30. Griffin RJ, Neuwirth K, Dunwoody S, Giese J. Information sufficiency and risk communication. Media Psychol 2004;6(1):23-61. [doi: 10.1207/s1532785xmep0601_2]
31. Kahlor L. PRISM: a planned risk information seeking model. Health Commun 2010 Jun;25(4):345-356. [doi: 10.1080/10410231003775172] [Medline: 20512716]
32. Moser RP, Arndt J, Han PK, Waters EA, Amsellem M, Hesse BW. Perceptions of cancer as a death sentence: prevalence and consequences. J Health Psychol 2014 Dec;19(12):1518-1524 [FREE Full text] [doi: 10.1177/1359105313494924] [Medline: 23864071]
33. Sweeny K, Melnyk D, Miller W, Shepperd JA. Information avoidance: who, what, when, and why. Rev Gen Psychol 2010;14(4):340-353. [doi: 10.1037/a0021288]
34. Witte K, Allen M. A meta-analysis of fear appeals: implications for effective public health campaigns. Health Educ Behav 2000 Oct;27(5):591-615. [doi: 10.1177/109019810002700506] [Medline: 11009129]
35. Case DO, Andrews JE, Johnson JD, Allard SL. Avoiding versus seeking: the relationship of information seeking to avoidance, blunting, coping, dissonance, and related concepts. J Med Libr Assoc 2005 Jul;93(3):353-362 [FREE Full text] [Medline: 16059425]
36. Miles A, Rainbow S, von Wagner C. Cancer fatalism and poor self-rated health mediate the association between socioeconomic status and uptake of colorectal cancer screening in England. Cancer Epidemiol Biomarkers Prev 2011 Oct;20(10):2132-2140 [FREE Full text] [doi: 10.1158/1055-9965.EPI-11-0453] [Medline: 21953115]
37. Niederdeppe J, Levy AG. Fatalistic beliefs about cancer prevention and three prevention behaviors. Cancer Epidemiol Biomarkers Prev 2007 May;16(5):998-1003. [doi: 10.1158/1055-9965.EPI-06-0608] [Medline: 17507628]
38. Lu L, Liu J, Yuan YC. Cultural differences in cancer information acquisition: cancer risk perceptions, fatalistic beliefs, and worry as predictors of cancer information seeking and avoidance in the U.S. and China. Health Commun 2022 Oct;37(11):1442-1451. [doi: 10.1080/10410236.2021.1901422] [Medline: 33752516]
39. Chung JE, Lee CJ. The impact of cancer information online on cancer fatalism: education and eHealth literacy as moderators. Health Educ Res 2019 Dec 01;34(6):543-555. [doi: 10.1093/her/cyz027] [Medline: 31550372]
40. Fleary SA, Paasche-Orlow MK, Joseph P, Freund KM. The relationship between health literacy, cancer prevention beliefs, and cancer prevention behaviors. J Cancer Educ 2019 Oct;34(5):958-965 [FREE Full text] [doi: 10.1007/s13187-018-1400-2] [Medline: 30022378]
41. Kobayashi LC, Smith SG. Cancer fatalism, literacy, and cancer information seeking in the American public. Health Educ Behav 2016 Aug;43(4):461-470 [FREE Full text] [doi: 10.1177/1090198115604616] [Medline: 26377524]
42. Morris NS, Field TS, Wagner JL, Cutrona SL, Roblin DW, Gaglio B, et al. The association between health literacy and cancer-related attitudes, behaviors, and knowledge. J Health Commun 2013;18 Suppl 1(Suppl 1):223-241 [FREE Full text] [doi: 10.1080/10810730.2013.825667] [Medline: 24093358]
43. Lee CJ, Chae J. An initial look at the associations of a variety of health-related online activities with cancer fatalism. Health Commun 2016 Nov;31(11):1375-1384. [doi: 10.1080/10410236.2015.1072885] [Medline: 27007443]
44. Lee SY, Hwang H, Hawkins R, Pingree S. Interplay of negative emotion and health self-efficacy on the use of health information and its outcomes. Commun Res 2008;35(3):358-381. [doi: 10.1177/0093650208315962]
45. Johnson JD. Cancer-related Information Seeking. New York, NY: Hampton Press; 1997.
46. Barbour JB, Rintamaki LS, Ramsey JA, Brashers DE. Avoiding health information. J Health Commun 2012;17(2):212-229. [doi: 10.1080/10810730.2011.585691] [Medline: 22004015]
47. Brashers DE, Goldsmith DJ, Hsieh E. Information seeking and avoiding in health contexts. Hum Commun Res 2002 Apr;28(2):258-271. [doi: 10.1111/j.1468-2958.2002.tb00807.x]
48. Yang Y, Yu G, Pan J, Kreps GL. Public trust in sources and channels on judgment accuracy in food safety misinformation with the moderation effect of self‐affirmation: evidence from the HINTS‐China database. World Med Health Policy 2022 Aug 24;15(2):148-162. [doi: 10.1002/wmh3.544]
49. Zhao X, Mao Q, Kreps GL, Yu G, Li Y, Chou SW, et al. Cancer information seekers in China: a preliminary profile. J Health Commun 2015;20(5):616-626. [doi: 10.1080/10810730.2015.1012244] [Medline: 25845031]
50. Chang A, Schulz PJ, Jiao W, Yu G, Yang Y. Media source characteristics regarding food fraud misinformation according to the Health Information National Trends Survey (HINTS) in China: comparative study. JMIR Form Res 2022 Mar 16;6(3):e32302 [FREE Full text] [doi: 10.2196/32302] [Medline: 34939565]
51. NUS-IRB'S amended exemption categories for social, behavioural and educational research (SBER). National University of Singapore Institutional Review Board. URL: https://fass.nus.edu.sg/cnm/wp-content/uploads/sites/2/2021/06/SBER-Exemption-Categories-2020-08-25.pdf [accessed 2024-03-22]
52. Zhang L, Jiang S. Linking health information seeking to patient-centered communication and healthy lifestyles: an exploratory study in China. Health Educ Res 2021 Apr 12;36(2):248-260. [doi: 10.1093/her/cyab005] [Medline: 33544831]

XSL•FO
RenderX

53. Chae J, Lee CJ. The psychological mechanism underlying communication effects on behavioral intention: focusing on affect and cognition in the cancer context. Commun Res 2019;46(5):597-618. [doi: 10.1177/0093650216644021]
54. Chen Y, Yang Q. How do cancer risk perception, benefit perception of quitting, and cancer worry influence quitting intention among current smokers: a study using the 2013 HINTS. J Subst Use 2017 Jan 27;22(5):555-560. [doi: 10.1080/14659891.2016.1271033]
55. Rains SA. Perceptions of traditional information sources and use of the world wide web to seek health information: findings from the health information national trends survey. J Health Commun 2007;12(7):667-680. [doi: 10.1080/10810730701619992] [Medline: 17934943]
56. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling Multidiscip J 1999;6(1):1-55. [doi: 10.1080/10705519909540118]
57. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: Guilford Publications; 2013.
58. Borkovec TD, Alcaine OM, Behar E. Avoidance theory of worry and generalized anxiety disorder. In: Heimberg RG, Turk CL, Mennin DS, editors. Generalized Anxiety Disorder: Advances in Research and Practice. New York, NY: Guilford Press; 2004.
59. Ferrer RA, Portnoy DB, Klein WM. Worry and risk perceptions as independent and interacting predictors of health protective behaviors. J Health Commun 2013;18(4):397-409. [doi: 10.1080/10810730.2012.727954] [Medline: 23272708]
60. Persoskie A, Ferrer RA, Klein WM. Association of cancer worry and perceived risk with doctor avoidance: an analysis of information avoidance in a nationally representative US sample. J Behav Med 2014 Oct;37(5):977-987. [doi: 10.1007/s10865-013-9537-2] [Medline: 24072430]
61. Brashers DE. Communication and uncertainty management. J Commun 2001 Sep;51(3):477-497. [doi: 10.1111/j.1460-2466.2001.tb02892.x]
62. Kim HK, Lwin MO. Cultural determinants of cancer fatalism and cancer prevention behaviors among Asians in Singapore. Health Commun 2021 Jul;36(8):940-949. [doi: 10.1080/10410236.2020.1724636] [Medline: 32041434]
63. Street RLJ. Mediated consumer-provider communication in cancer care: the empowering potential of new technologies. Patient Educ Couns 2003 May;50(1):99-104. [doi: 10.1016/s0738-3991(03)00089-2] [Medline: 12767594]
64. Jiang S, Street RL. Pathway linking internet health information seeking to better health: a moderated mediation study. Health Commun 2017 Aug;32(8):1024-1031. [doi: 10.1080/10410236.2016.1196514] [Medline: 27464036]
65. Witte K. Fear control and danger control: a test of the extended parallel process model (EPPM). Commun Monogr 2009 Jun 02;61(2):113-134. [doi: 10.1080/03637759409376328]

Abbreviations

CIA: cancer information avoidance
EPPM: extended parallel process model
HINTS: Health Information National Trends Survey
IRB: institutional review board
OHIS: online health information seeking
SRMR: standardized root mean square residual

Edited by T de Azevedo Cardoso; submitted 26.05.23; peer-reviewed by JS Tham, L Chen; comments to author 07.10.23; revised version received 27.10.23; accepted 25.04.24; published 31.05.24.

Please cite as:
Zhang L, Jiang S
Examining the Role of Information Behavior in Linking Cancer Risk Perception and Cancer Worry to Cancer Fatalism in China: Cross-Sectional Survey Study
J Med Internet Res 2024;26:e49383
URL: https://www.jmir.org/2024/1/e49383
doi:10.2196/49383
PMID:38819919

©Lianshan Zhang, Shaohai Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

COVID-19 Vaccine Effectiveness and Digital Pandemic Surveillance in Germany (eCOV Study): Web Application–Based Prospective Observational Cohort Study

Anna-Lena Lang[1*], MD; Nils Hohmuth[1,2*], MD; Vukašin Višković[1], MA; Stefan Konigorski[3,4,5], PhD; Stefan Scholz[6], PhD; Felix Balzer[2], MD, PhD; Cornelius Remschmidt[1*], MD, PD; Rasmus Leistner[7*], MD, PD

[1]d4l Data4Life gGmbH, Potsdam, Germany
[2]Institute of Medical Informatics, Charité University Medicine Berlin, Berlin, Germany
[3]Digital Health Center, Hasso Plattner Institute for Digital Engineering, Potsdam, Germany
[4]Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[5]Department of Statistics, Harvard University, Cambridge, MA, United States
[6]Health Services Research and Health Economics, Martin Luther University Halle-Wittenberg, Halle Saale, Germany
[7]Department of Gastroenterology, Infectiology and Rheumatology, Charité University Medicine Berlin, Berlin, Germany
[*]these authors contributed equally

Corresponding Author:
Anna-Lena Lang, MD
d4l Data4Life gGmbH
c/o Digital Health Cluster (DHC) im Hasso-Plattner-Institut
Rudolf-Breitscheid-Straße 187
Potsdam, 14482
Germany
Phone: 49 015756025551
Email: annalena.lang.26@gmail.com

Abstract

Background: The COVID-19 pandemic posed significant challenges to global health systems. Efficient public health responses required a rapid and secure collection of health data to improve the understanding of SARS-CoV-2 and examine the vaccine effectiveness (VE) and drug safety of the novel COVID-19 vaccines.

Objective: This study (*COVID-19 study on vaccinated and unvaccinated subjects over 16 years*; eCOV study) aims to (1) evaluate the real-world effectiveness of COVID-19 vaccines through a digital participatory surveillance tool and (2) assess the potential of self-reported data for monitoring key parameters of the COVID-19 pandemic in Germany.

Methods: Using a digital study web application, we collected self-reported data between May 1, 2021, and August 1, 2022, to assess VE, test positivity rates, COVID-19 incidence rates, and adverse events after COVID-19 vaccination. Our primary outcome measure was the VE of SARS-CoV-2 vaccines against laboratory-confirmed SARS-CoV-2 infection. The secondary outcome measures included VE against hospitalization and across different SARS-CoV-2 variants, adverse events after vaccination, and symptoms during infection. Logistic regression models adjusted for confounders were used to estimate VE 4 to 48 weeks after the primary vaccination series and after third-dose vaccination. Unvaccinated participants were compared with age- and gender-matched participants who had received 2 doses of BNT162b2 (Pfizer-BioNTech) and those who had received 3 doses of BNT162b2 and were not infected before the last vaccination. To assess the potential of self-reported digital data, the data were compared with official data from public health authorities.

Results: We enrolled 10,077 participants (aged ≥16 y) who contributed 44,786 tests and 5530 symptoms. In this young, primarily female, and digital-literate cohort, VE against infections of any severity waned from 91.2% (95% CI 70.4%-97.4%) at week 4 to 37.2% (95% CI 23.5%-48.5%) at week 48 after the second dose of BNT162b2. A third dose of BNT162b2 increased VE to 67.6% (95% CI 50.3%-78.8%) after 4 weeks. The low number of reported hospitalizations limited our ability to calculate VE against hospitalization. Adverse events after vaccination were consistent with previously published research. Seven-day incidences and test positivity rates reflected the course of the pandemic in Germany when compared with official numbers from the national infectious disease surveillance system.

XSL•FO
RenderX

Conclusions: Our data indicate that COVID-19 vaccinations are safe and effective, and third-dose vaccinations partially restore protection against SARS-CoV-2 infection. The study showcased the successful use of a digital study web application for COVID-19 surveillance and continuous monitoring of VE in Germany, highlighting its potential to accelerate public health decision-making. Addressing biases in digital data collection is vital to ensure the accuracy and reliability of digital solutions as public health tools.

(J Med Internet Res 2024;26:e47070) doi:10.2196/47070

KEYWORDS

COVID-19; SARS-CoV-2; COVID-19 vaccines; BNT162b2; vaccine effectiveness; participatory disease surveillance; web application; digital public health; vaccination; Germany; effectiveness; data collection; disease surveillance; tool

Introduction

Background

At the onset of the COVID-19 pandemic in early 2020, the world was far from prepared for pandemic surveillance. Manual case reporting from testing centers and medical facilities to public health authorities was common and remains so even today. However, timely access to accurate disease surveillance data has proven essential for informed public health decision-making and the adaptable response of health care systems [1]. Over the course of the pandemic, numerous innovative digital health solutions [1-8] were developed, focusing on real-time monitoring and predicting the dynamics of the COVID-19 pandemic. The broad adoption of smartphones enabled scalable population-wide data collection [9, 10]. However, recruiting a substantial number of participants for meaningful analyses without the risk of participant bias as well as data privacy concerns remained an ongoing challenge in these digital public health efforts [9].

The most prominent of such solutions developed in Europe were the ZOE COVID Study app in the United Kingdom [2,11,12] and the COVID Symptom Study app in Sweden [1]. Both apps were developed to gather self-reported data for COVID-19 syndromic surveillance on a national scale. The app used in Sweden primarily focused on syndromic surveillance, while the ZOE COVID Study app's data were also used to test vaccine effectiveness (VE) and drug safety under real-world conditions, as well as to continuously monitor VE as various SARS-CoV-2 variants emerged. The ZOE COVID Study showcased the benefits of using self-reported data to identify COVID-19 hot spots in a timely and efficient manner. It also illustrated how digital apps can continuously monitor VE, providing data for informed decision-making in public health policies. In comparison to phase 3 vaccine efficacy trials, the data from the ZOE COVID Study app showed similar VE and lower frequencies of expected adverse events (AEs) after vaccination and highlighted the importance of second- and third-dose vaccinations for increased protection [2,11,13]. The study conducted in Sweden further confirmed the value of app-based COVID-19 syndromic surveillance. It highlighted the potential of digital health solutions to monitor pandemics at a population level with limited resources, offering a viable approach for addressing public health challenges. Amid varying national strategies, vaccination rates, and COVID-19 prevalence, it remains crucial to evaluate pandemic parameters nationally [14-16].

Objectives

The objective of this study was to assess the real-world effectiveness of COVID-19 vaccines among the German population, using self-reported data from a web application. The secondary parameters of interest included VE against hospitalization and across different SARS-CoV-2 variants, the frequency of SARS-CoV-2 infections, SARS-CoV-2 testing data over time, and symptom progression. We assessed the validity of the self-reported data gathered in this digital study by comparing these data to official data from public health authorities and nondigital studies. This study highlights the role of digital participatory pandemic surveillance to inform and accelerate efficient public health responses.

Methods

We followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [17] for cohort studies to ensure comprehensive and transparent reporting of our research (Multimedia Appendix 1).

Study Design

This digital prospective observational cohort study was conducted from May 1, 2021, to August 1, 2022. The study was only promoted in Germany, but the web-based setting of the study facilitated international enrollment. Participants were able to self-enroll in the study. Enrollment was first restricted to individuals aged >18 years. As vaccines were subsequently approved for younger age groups, the study was opened to people aged ≥16 years in December 2021. No further exclusion criteria were applied for self-enrollment. Recruitment was carried out on the web through paid and unpaid marketing on social media platforms and websites as well as offline through advertisements in trains and public spaces, at a mass gathering, and in general practitioner offices across Germany. The study's offline recruitment was carried out to reduce selection bias and reach a broader range of participants, including individuals who may not have been active on social media platforms. The wording and pictures in the web-based and offline campaigns were targeted to ensure outreach to diverse target groups, including people from different age groups, genders, and professional backgrounds (Multimedia Appendix 2). Demographic information of the study cohort was continuously evaluated for representativeness of the cohort and influenced the ongoing recruitment efforts. Participants could choose to complete the study in German or English.

Participants were asked to complete different questionnaires provided in the study app, which could be accessed on both

XSL•FO
RenderX

mobile and desktop devices (Figure S1 in Multimedia Appendix 3). Email notifications reminded participants to respond in a timely manner to any pending questionnaires to minimize recall bias. Participant feedback from an in-app feedback function and qualitative interviews showed that weekly email notifications were perceived as too frequent. Therefore, we switched to monthly email notifications starting December 2021. The email notifications were carefully crafted to ensure neutrality and motivate individuals (regardless of their COVID-19 symptoms or test results) to log in to their accounts and complete the pending questionnaires. This was done to reduce response bias to email notifications.

The digital questionnaires collected demographic information and COVID-19–related data on vaccination status, symptoms, testing, and contact behavior (Multimedia Appendix 4). Study enrollment was defined as creating a user account, agreeing to the terms of use of the study web application, and providing informed digital consent for study participation. Upon enrollment, individuals reported their demographic data and vaccination status information as well as prior SARS-CoV-2 infections. Vaccines available in Germany during the study period were AZD1222 (Oxford-AstraZeneca), BNT162b2 (Pfizer-BioNTech), JNJ-78436735 (Janssen), mRNA-1273 (Moderna), and NVX-CoV2373 (Novavax). COVID-19 tests and symptoms could be reported daily through on-demand questionnaires in the web application. Participants could also report data retrospectively, that is, report a symptom they had experienced in the last month. On a monthly basis, they were asked to provide information on COVID-19 exposure and contact behavior. All answers could be edited within 48 hours. Individuals who experienced AEs after vaccination were able to report them through the certified web-based pharmacovigilance platform Medikura [18]. Medikura forwarded all reports on AEs to the EudraVigilance database hosted by the European Medicines Agency [19]. Under a contractual agreement with Medikura, we were provided with data from our study participants for the purposes of this study. By design, the collected data were pseudonymized and not linked to study participants in our data set.

Vaccination Status

On the basis of the self-reported data, study participants were grouped by vaccination status for downstream analysis. We differentiated between the total number of vaccine doses received and the combination of types of vaccines. For our main VE analysis, we focused on 2 specific groups: unvaccinated participants (1768/9287, 19.04%) and individuals who completed the primary vaccination series with 2 doses of BNT162b2 and had no prior infections before vaccination (3463/9287, 37.29%).

Outcomes

Our primary outcome measure for this study was the effectiveness of SARS-CoV-2 vaccines against laboratory-confirmed (polymerase chain reaction [PCR] test) SARS-CoV-2 infection, independent of any current symptomatology. The secondary outcome measures included the effectiveness of SARS-CoV-2 vaccines against symptomatic infection, against severe infection defined by hospitalization, and VE across different SARS-CoV-2 variants. We further evaluated COVID-19 symptoms over time and by variant and investigated self-reported AEs possibly associated with the SARS-CoV-2 vaccinations. To assess the validity of the self-reported data, we compared the data with official surveillance data from public health authorities, focusing on COVID-19 vaccine coverage, test positivity rates, and 7-day incidence.

Case Definitions

Study participants' exposure to COVID-19 was based on results from self-reported PCR tests. An individual was classified as a *confirmed case* if a positive PCR test was reported, following the definition of SARS-CoV-2 infection used by the German public health authorities [20]. Reinfection was only considered if it occurred >60 days after the last infection. An infection was considered symptomatic if any symptoms occurred within 5 days before the first positive test to 10 days after it. By design, participants were not required to get tested and submit test results on a regular basis but rather on demand. Not reporting a SARS-CoV-2 test was thus weighted as equal to not being infected. To reduce reporting bias, monthly email notifications invited individuals to answer the pending questionnaires independent of symptomatology.

Data Source

The *COVID-19 study on vaccinated and unvaccinated subjects over 16 years* (eCOV study) was registered with the German Clinical Trials Register (DRKS00025169). The acronym "*eCOV*" was selected to reflect the electronic ("*e*") or digital nature of this study on COVID-19 ("*COV*"). The study and the web application were developed by the nonprofit organization Data4Life, based in Potsdam, Germany. The study app was a web application that was accessible on both desktop and mobile devices (Figure S1 in Multimedia Appendix 3). Self-reported data were collected via the study app and processed by Data4Life's General Data Protection Regulation–compliant research infrastructure, certified by the German Federal Office for Information Security. Data from two other data sources were collected to monitor and analyze web application log-ins and email notifications: (1) logs generated from the study app; and (2) if the participant had provided consent to place optional cookies, data generated from the General Data Protection Regulation–compliant web analytics tool Matomo [21]. The logs generated from the study app enabled the analysis on daily log-ins and responses to email notifications. Data from Matomo enabled the analysis on device types used for log-ins and participation.

Quality Control

Quality control of the self-reported data included removing individuals from the cohort who reported unlikely dates of vaccination. We thus removed participants if no real date was reported, the date of vaccination was before the first possible vaccination date in Germany in December 2021, participants reported an impossible order of vaccination dates, and if vaccine dates lay within <2 weeks of each other. Participants were further removed from the cohort if the reported vaccine type was *unknown*, *other*, or *Sputnik V*. If the same vaccination

XSL•FO
RenderX

questionnaire was submitted multiple times with differing vaccine dates, the first reported date was considered. Apart from this, whenever questionnaires were answered multiple times, or answers were edited by study participants, the last reported answer was considered for analysis.

Statistical Analysis

All analysis was performed on the Data4Life analytics platform, hosting a JupyterHub notebook using *pandas* (version 1.4.2) [22], *matplotlib* (version 3.5.2) [23], *seaborn* (version 0.11.2) [24], and *numpy* (version 1.22.4) [25]. Descriptive statistics were used to report details about the study cohort and the respondents' reporting behavior. Continuous values were reported as mean and SD, and categorical values were reported as numbers with percentages. The chi-square test of independence was used for a difference in the numbers of symptoms between vaccinated and unvaccinated individuals. The Shapiro-Wilk test was used to test for normal distribution of the data. We used the Mann-Whitney *U* test for not normally distributed data to compare the 7-day incidences of COVID-19 among groups and the number of reported tests.

VE analysis was carried out as a measure of how well vaccination protects against the defined outcome. For VE analysis, we ran an intention-to-treat analysis protocol, including individuals in the analysis independent of their study dropout date. As the questionnaires administered upon enrollment included retrospective questions about prior vaccinations and infections, data from early dropouts were still valuable for VE analysis. VE was calculated for individuals who received 2 doses of any vaccine brand as well as for individuals who received 2 or 3 doses of BNT162b2. We calculated VE for 4 to 48 weeks after the vaccine dose of interest. If participants received >2 vaccine doses, data were only considered until the time point of the next vaccination. In preparation for logistic regression analyses, each vaccinated person was matched with an unvaccinated person of the same age group (−5 y to +5 y) and gender who had submitted data within the same time frame. Subsequently, a binary logistic regression model was used to obtain the odds ratios of SARS-CoV-2 infections in the respective time frames of interest, given the vaccination status. As the primary outcome of interest, we analyzed SARS-CoV-2 infection, regardless of symptomatology. We adjusted for comorbidity (binary variable, with or without comorbidities), health care worker status (binary variable), age, and age-adjusted 7-day incidence in the German population during the time frame of interest. These covariates were added to the logistic regression model following theoretical criteria using backward elimination. For the secondary outcomes—VE against symptomatic infections, VE by variant, and VE against hospitalization—the same covariates were added to the logistic regression model. Next, the logit model was fitted using the *Newton-Raphson* optimization method implemented in the *statsmodels* Python library [26], accounting for matched pairs. Adjusted VE was then calculated as 1–odds ratio. One logistic regression model was run for each time frame of interest. To consider the effect of different SARS-CoV-2 variants on VE, we defined periods of variant dominance. In Germany, the Delta variant was dominant starting calendar week 26, 2021, and the Omicron variant was dominant starting calendar week 2, 2022 [27]. Only individuals vaccinated during the dominance periods of the respective variants were considered and matched with unvaccinated individuals who had contributed data in the same time frame. To assess the comparability of our results and the influence of covariates added to the logistic regression models, we used the Farrington screening method to calculate unadjusted VE and 95% CIs using the same matching principle [28,29].

Ethical Considerations

The study was approved by the ethics committee of the Berlin Chamber of Physicians (Eth-11/22). Registration was open for people aged ≥16 years. All participants provided digital informed consent for study participation. Participation was voluntary, and no incentives were offered. Email address and password as well as 2-factor authentication were required for log-in. Individuals could access all study content through a web application on both desktop and mobile devices. All study data were end-to-end encrypted and pseudonymized. Only authorized researchers were provided access to the data on the Data4Life analytics platform. Data were stored exclusively at Data4Life data centers in Germany. Data4Life is certified by the German Federal Office for Information Security on the basis of IT-Grundschutz (ISO 27001).

Results

Study Setting and Participants

A total of 10,077 participants self-enrolled in this digital cohort study. After thorough quality control (Figure S2 in Multimedia Appendix 3), data from 9287 (92.16%) of the 10,077 participants were available for final analyses presented in this paper (Table 1). Most of the study participants were women (6567/9287, 70.71%), while 0.32% (30/9287) identified as nonbinary. Study participation was not limited to German residents, but because the study was mainly advertised in Germany, only a few people participated from abroad (755/9287, 8.13%). Concerning possible risk factors for infections, slightly more than a quarter (2463/9287, 26.52%) of the participants reported living with allergies, while 5.69% (529/9287) had an immune deficiency, and 19.96% (1854/9287) reported living with other chronic health conditions. Active smokers made up 16.7% (1551/9287) of the study population, and the average BMI was 25 (SD 5.6) kg/m^2. As the study was also advertised at a mass gathering for medical professionals, a large proportion (3610/9287, 38.87%) of the study population hailed from the health care sector, mainly hospital employees (1046/3610, 29%) or medical students (1258/3610, 34.85%). At study end (August 1, 2022), of the 9287 participants, 1768 (19.04%) were unvaccinated, 1801 (19.39%) had reported 1 vaccine dose, 3119 (33.58%) had reported 2 doses, 2433 (26.19%) had reported 3 doses (commonly referred to as the first *booster* vaccination), 164 (1.77%) had reported 4 doses, and 2 (0.02%) had reported 5 doses.

XSL•FO
RenderX

Table 1. Descriptive characteristics of the study cohort (n=9287).

Characteristics	Values
Age (y), mean (SD)	37.3 (14.9)
Gender, n (%)	
Woman	6567 (70.71)
Man	2658 (28.62)
Nonbinary	30 (0.32)
Residency, n (%)	
Germany	8532 (91.87)
Abroad	755 (8.13)
Health status, n (%)	
Immune deficiency	529 (5.69)
Allergies	2463 (26.52)
Other chronic diseases	1854 (19.96)
Active smoker, n (%)	1551 (16.7)
BMI (kg/m^2), mean (SD)	25 (5.6)
Profession, n (%)	
Not health care sector	4647 (50.03)
Health care sector	3610 (38.87)
Physician's office	248 (2.67)
Nursing facility or retirement home	294 (3.16)
Other medical field	764 (8.22)
Medical student	1258 (14.62)
Hospital	1046 (11.26)
Vaccine status at study end, n (%)	
Unvaccinated	1768 (19.04)
1 dose	1801 (19.39)
2 doses	3119 (33.58
3 doses	2433 (26.19)
4 doses	164 (1.77)
5 doses	2 (0.02)

Reporting Behavior and Retention

Over the course of the study, participants reported a total of 44,786 tests and 5530 symptoms. With 4263 (45.9%) of the 9287 participants leaving the study within 1 week after enrollment, the average length of participation was 10 weeks (mean 70.2, SD 105.6 days). Individuals who stayed in the study for >1 week (5024/9287, 54.1%) stayed for an average of 18.4 weeks (mean 129.3, SD 113.0 days). Individuals leaving the study after 1 week were still considered in the analysis because the questionnaires administered upon enrollment included retrospective questions about prior vaccinations and infections. An average of 145.7 (SD 324.5) daily log-ins were observed during the study period, of which an average of 111.6 (SD 312.2) were observed after an email notification containing reminders about study participation (Figure S3 in Multimedia Appendix 3). Daily log-ins were always the highest in the first 48 hours after email notifications (Figure S3 in Multimedia Appendix 3). Most log ins were conducted from mobile devices (18,547/23,865, 77.71%).

VE Against SARS-CoV-2 Infection of Any Severity

At study end, approximately three-fourths of the nationally distributed vaccine doses in Germany were BNT162b2 [30], explaining the low sample size for other vaccine brands in our cohort (Multimedia Appendix 5). Therefore, our primary VE analysis focused on 2 specific groups: unvaccinated participants (1768/9287, 19.04%) and individuals who completed the primary vaccination series with 2 doses of BNT162b2 and had no prior infections before vaccination (3463/9287, 37.29%). The participants' ages—unvaccinated: mean 38.9 (SD 13.3) years and vaccinated: mean 37.2 (SD 15.0) years—did not differ

XSL•FO
RenderX

substantially by vaccination status (Table 2). The majority of the participants in the 2 groups—unvaccinated: 69.62% (1231/1768) and vaccinated: 73.08% (2531/3463)—identified as women. Study participation was not limited to German residents, but because the study was mainly advertised in Germany, only a few participants—unvaccinated: 14.03% (248/1768) and vaccinated: 6.12% (212/3463)—resided outside Germany. Both groups showed similar risk profiles in terms of BMI and comorbidities. More health care workers were among those vaccinated (1428/3463, 41.23%), whereas more active smokers (417/1768, 23.58%) were among those unvaccinated (Table 2). Relatively fewer infections were reported after receiving 2 doses of BNT162b2 (992/3463, 28.64%) compared with unvaccinated individuals during the study period (637/1768, 36.02%; chi-square test, $P<.001$). However, vaccinated individuals were more likely to report symptoms associated with infections (chi-square test, $P=.02$) and reported more tests (mean 5.33, SD 10.6) than unvaccinated individuals (mean 5.02, SD 13.0; Mann-Whitney U test, $P<.001$; Table 2). Response rates to monthly questionnaires showed a substantial difference between vaccinated and unvaccinated individuals: only 38.85% (687/1768) of the unvaccinated individuals completed at least 1 monthly questionnaire asking about exposure to COVID-19 and infections in past weeks, whereas 65.2% (2258/3463) of the vaccinated individuals did so.

We calculated VE against infection across all SARS-CoV-2 variants and for the Delta variant separately using logistic regression models (refer to the *Statistical Analysis* subsection under Methods). Across all vaccine brand combinations, VE against infection of any severity 4 weeks after the second dose was 87.8% (95% CI 64.3%-95.8%), which waned to 42.8% (95% CI 26%-55.8%) after 24 weeks and 37.9% (95% CI 26.5%-47.6%) after 48 weeks (Figure 1; Table 3). Similar numbers were seen for VE after 2 doses of BNT162b2: after 4 weeks, VE was 91.2% (95% CI 70.4%-97.4%), which decreased to 42.3% (95% CI 24.7%-55.8%) after 24 weeks. VE then declined in weeks 28 to 32 after the second dose (Figure 1; Tables 3 and 4) before increasing to 37.2% (95% CI 23.5%-48.5%) after 48 weeks. Similar results were seen using the Farrington screening method [28] to calculate unadjusted VE (Table 4. VE was increased to 67.6% (95% CI 50.3%-78.8%) at 4 weeks after the third dose of BNT162b2 (Figure S4 in Multimedia Appendix 3).

XSL•FO
RenderX

Table 2. Descriptive characteristics of the study population split by vaccination status. Infections among unvaccinated individuals were compared with those among individuals who received at least 2 doses of BNT162b2 (Pfizer-BioNTech) and did not have any prior infections before vaccination.

Characteristics	Unvaccinated individuals (n=1768)	Vaccinated individuals (2 doses of BNT162b2; n=3463)
Age (y), mean (SD)	38.9 (13.3)	37.3 (15.0)
Gender, n (%)		
Woman	1231 (69.63)	2531 (73.1)
Man	514 (29.07)	921 (26.6)
Nonbinary	6 (0.34)	10 (0.28)
Residency, n (%)		
Germany	1520 (85.97)	3252 (93.9)
Abroad	248 (14.03)	212 (6.12)
Austria	162 (9.16)	148 (4.27)
Switzerland	36 (2.03)	23 (0.66)
Other	50 (2.83	41 (1.18)
Any chronic disease, n (%)		
Allergies	418 (23.64)	959 (27.69)
Others	348 (19.68)	721 (20.82
Immune deficiency, n (%)	118 (6.67)	206 (5.95)
Active smoker, n (%)	417 (23.59)	519 (14.99)
Health care worker, n (%)	388 (21.95)	1428 (41.24)
BMI (kg/m^2), mean (SD)	25.4 (6.3)	25 (5.5)
Infections[a], n (%)	637 (36.03)	992 (28.65)[b]
Symptomatic infections, n (%)	318 (17.99)[c]	734 (21.2)
Test reports per person, mean (SD)	5.0 (13.0)	5.3 (10.6)[d]

[a]Infections after vaccination for vaccinated individuals and infections over the study period for unvaccinated individuals.

[b]Chi-square test, $P<.001$.

[c]Chi-square test, $P=.02$.

[d]Mann-Whitney U test, $P<.001$.

XSL•FO
RenderX

Figure 1. Vaccine effectiveness (VE) against infection of any severity in the COVID-19 study on vaccinated and unvaccinated subjects over 16 years (eCOV study) cohort in the weeks after the second dose of BNT162b2. VE (%) was calculated for weeks 4 to 48 after completing the primary vaccination series with 2 doses of BNT162b2 (Pfizer-BioNTech). Shaded areas indicate 95% CIs.

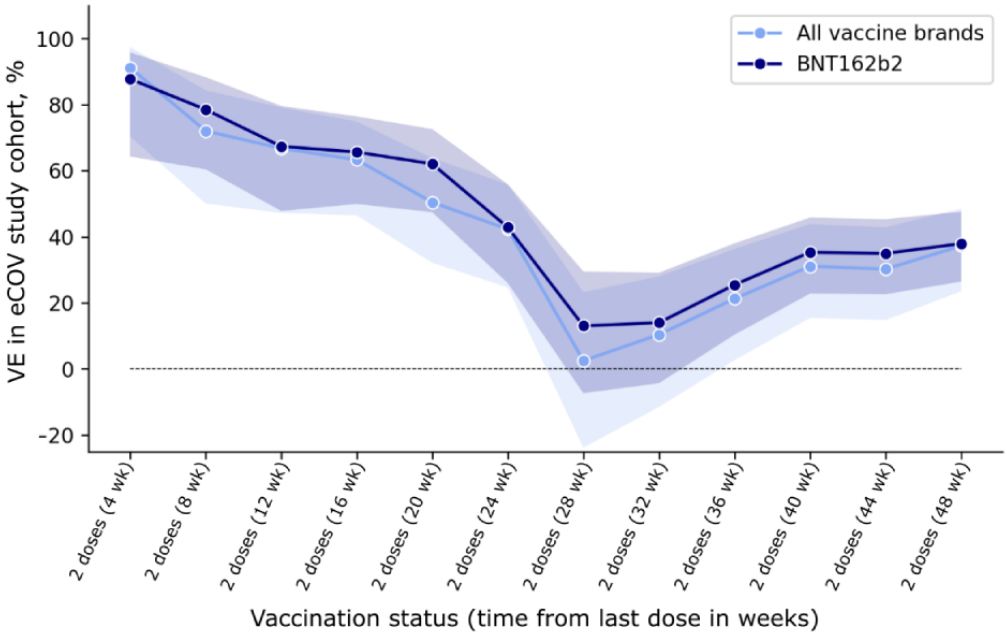

Table 3. Vaccine effectiveness against infection of any severity after the second vaccine dose across all SARS-CoV-2 variants. The numbers only reflect vaccine effectiveness against infection, independent of symptoms.

Weeks after last dose	Second dose, any vaccine brand (%, 95% CI)	Second dose, BNT162b2 (Pfizer-BioNTech; %, 95% CI)
4	87.8 (64.3 to 95.8)	91.2 (70.4 to 97.4)
8	78.5 (60.5 to 88.3)	72 (50.1 to 84.3)
12	67.4 (47.8 to 79.6)	66.8 (47.2 to 79.1)
16	65.6 (49.9 to 76.4)	63.3 (46.5 to 74.8)
20	62.1 (47.5 to 72.6)	50.4 (23.1 to 63.7)
24	42.8 (26 to 55.8)	42.3 (24.7 to 55.8)
28	13 (7.3 to 29.5)	2.5 (−23.8 to 23.1)
32	14 (4.3 to 29.1)	10.4 (−11.5 to 28)
36	25.5 (10.4 to 38)	21.3 (2.7 to 36.3)
40	35.3 (22.8 to 45.8)	31.1 (15.5 to 43.8)
44	34.9 (22.6 to 45.3)	30.2 (14.8 to 42.8)
48	37.9 (26.5 to 47.6)	37.2 (23.5 to 48.5)

XSL•FO
RenderX

Table 4. Vaccine effectiveness (VE) against infection of any severity after the second dose of BNT162b2 (Pfizer-BioNTech) across all SARS-CoV-2 variants. The numbers only reflect VE against infection, independent of symptoms. Adjusted VE was calculated using a logistic regression model with adjustment for age, comorbidity, health worker status, and age-adjusted average incidence in the German population during the time frame of interest. Unadjusted VE was calculated using the Farrington screening method.

Weeks after second vaccine dose	Sample size, n	Cases (positive PCR[a] test), n (%)		Adjusted VE (logistic regression; %, 95% CI)	Unadjusted VE (Farrington method; %, 95% CI)
		Vaccinated individuals	Unvaccinated individuals		
4	1679	3 (0.2)	33 (2)	91.2 (70.4 to 97.4)	90.9 (70.4 to 97.2)
8	1680	17 (1)	55 (3.3)	72 (50.1 to 84.3)	69.1 (46.8 to 82.1)
12	1679	32 (1.9)	79 (4.7)	66.8 (47.2 to 79.1)	59.5 (38.9 to 73.1)
16	1678	49 (2.9)	113 (6.7)	63.3 (46.5 to 74.8)	56.6 (39.4 to 69)
20	1680	78 (4.6)	142 (8.5)	50.4 (23.1 to 63.7)	45.1 (27.6 to 58.3)
24	1658	118 (7.1)	184 (11.1)	42.3 (24.7 to 55.8)	35.9 (19.2 to 49.1)
28	1559	174 (11.2)	187 (12)	2.5 (−23.8 to 23.1)	7 (−14.4 to 24.3)
32	1446	213 (14.7)	241 (16.7)	10.4 (−11.5 to 28)	11.6 (−6.3 to 26.5)
36	1391	219 (15.7)	272 (19.6)	21.3 (2.7 to 36.3)	19.5 (3.8 to 32.6)
40	1355	231 (17)	309 (22.8)	31.1 (15.5 to 43.8)	25.2 (11.3 to 37)
44	1314	238 (18.1)	320 (24.4)	30.2 (14.8 to 42.8)	25.6 (12 to 37.1)
48	1319	233 (17.7)	343 (26)	37.2 (23.5 to 48.5)	32.1 (19.8 to 42.5)

[a]PCR: polymerase chain reaction.

VE by SARS-CoV-2 Variant

To further inspect the influence of the SARS-CoV-2 Delta variant on VE, we filtered the data set for individuals who received their second vaccine dose during the dominance period of the Delta variant in Germany (1363/3463, 39.36%). In this case, 4 weeks after the second dose of BNT162b2 was administered, the VE against infection was 76.8% (95% CI 12.5%-93.8%), which decreased to 40.6% (95% CI 8%-61.76%) at 24 weeks (Figure S5 in Multimedia Appendix 3). The sample size of vaccinated individuals reporting data after week 24 was too low (206/3463, 5.95%). Furthermore, we were underpowered for analysis of VE for the SARS-CoV-2 Omicron variant because only 30 (0.87%) of the 3463 participants vaccinated with BNT162b2 completed their primary immunization series with BNT162b2 during the dominance period of the Omicron variant.

VE Against Symptomatic SARS-CoV-2 Infection

As the number of symptomatic infections in our study was low (1563/2834, 55.15%), the CIs of the estimated VE against symptomatic infection were correspondingly large (Multimedia Appendix 6). VE against symptomatic infection in the eCOV study cohort seemed to be lower after vaccination than VE against infection of any severity (Figure S6 in Multimedia Appendix 3).

Time Between SARS-CoV-2 Infections Was Prolonged in Vaccinated Individuals

Among the 2834 PCR test–confirmed COVID-19 cases, 90 (3.2%) individuals reported ≥2 infections. Among these 90 individuals, 32 (36%) were unvaccinated, and 23 (26%) had been vaccinated twice with any vaccine type at the time of both infections. The median time between the first and second infections for the unvaccinated individuals was 159.0 (IQR 106.25-362.0) days, and for those who had been vaccinated twice, it was 385.5 (IQR 342.0-438.5) days (Mann-Whitney *U* test, $P<.001$).

Pain at the Injection Site Was the Most Reported AE After Vaccination

A total of 16,004 vaccine doses were reported during the study. The question of whether expected or unexpected AEs occurred after vaccination was answered 15,363 times by 95.04% (7147/7519) of those who had been vaccinated at least once. The reported incidence of AEs decreased slightly with the increasing number of vaccinations, from 81.45% (5821/7147) after the first dose to 72.49% (4007/5528) after the second dose and 71.35% (1659/2325) after the third dose of any vaccine brand. In addition to categorical *yes* or *no* responses to suspected AEs, 2364 reports were submitted that included a total of 7373 expected and unexpected AEs after vaccination. As most AEs are expected within 6 weeks after vaccination, we reported only AEs occurring in this time frame (3988/7373, 54.08%). Of these AEs, 48.84% (1948/3988) occurred after the first dose, 31.24% (1246/3988) after the second dose, and 17.3% (690/3988) after the third dose. Across all vaccine brands and regardless of the number of doses received, local pain or tenderness at the injection site was the most common expected AE after vaccination (697/3988, 17.47% of all AEs), followed by headache (665/3988, 16.67%), fatigue or exhaustion (657/3988, 16.47%), body ache (459/3988, 11.5%), and fever (376/3988, 9.42%). Among the local AEs (822/3988, 20.61%), pain or tenderness at the injection site was the most common, (697/822, 84.8%), followed by swelling or pain of the lymph nodes (69/822, 8.4%; Figure S7A in Multimedia Appendix 3).

XSL•FO
RenderX

Headache was the leading systemic AE (665/3166, 21%), followed by fatigue and exhaustion, body ache, and fever (Figure S7B in Multimedia Appendix 3). The distribution of local and systemic AEs did not differ notably by number of doses and vaccine brands (Figures S8 and S9 in Multimedia Appendix 3). Apart from the expected AEs, the following serious unexpected AEs were reported within 6 weeks of vaccination: 1 cerebral sinus thrombosis, 2 reports of anaphylaxis (not further specified), and 3 reports of hemiplegia and paralysis of the face (facial nerve palsy). We do not know whether the reported AEs were attributable to the vaccination because the reporting was anonymous, and we were not involved in follow-up investigations. All AEs collected by our study partner Medikura were reported to the EudraVigilance database hosted by the European Medicines Agency. We cannot comment on AEs associated with different vaccine combinations because this information was not available to us.

Loss of Taste or Smell Was the Most Frequent Symptom During SARS-CoV-2 Infection

Of the 2834 PCR test–confirmed COVID-19 cases reported to our app, 1563 (55.15%) infections were reportedly symptomatic. A total of 3374 symptoms were reported within 5 days before the first positive PCR test to 10 days after it. While 74.1% (734/992) of the infections in individuals who had received 2 doses of BNT162b2 were reportedly symptomatic, only 49.9% (318/637) of the infections in unvaccinated individuals were reportedly symptomatic. The most frequent symptom accompanying infection, regardless of vaccination status and SARS-CoV-2 variant, was the loss of taste or smell, which was reported in 26.4% (413/1563) of all PCR test–confirmed symptomatic infections and, during the dominance period of the Delta variant, in 54.5% (165/303) of all PCR test–confirmed symptomatic infections (Multimedia Appendix 7), followed by fever (343/1563, 21.94%), cough (299/1563, 19.12%), and diarrhea (271/1563, 17.34%). Across all SARS-CoV-2 variants, infected individuals primarily vaccinated with 2 doses of BNT162b2 reported fewer occurrences of loss of taste or smell (177/744, 23.8%) than unvaccinated individuals (114/364, 31.3%; chi-square test, P=.009). This difference was most notable in infections during the dominance period of the Delta variant and disappeared during the dominance period of the Omicron variant, when only 18.29% (217/1186) of all symptomatic infections were associated with loss of taste or smell as a symptom, with similar distribution in vaccinated and unvaccinated individuals (Multimedia Appendix 7). Notably, cough during infection was reported more often by individuals vaccinated with 2 doses of BNT162b2 (163/744, 21.9%) compared to unvaccinated individuals (28/364, 7.7%, chi-square test P<.001). Of the 1563 individuals with symptomatic infections, 397 (25.4%) received medical treatment, and 20 (1.28%) were hospitalized. The hospitalization rate for symptomatic infections was 2.47% (9/364) among unvaccinated individuals and 0.8% (6/744) among individuals who had received 2 doses of BNT162b2 before infection (chi-square test, P=.048; Multimedia Appendix 8). Notably, the average age as a highly influential factor for hospitalization did not differ significantly between the 2 groups.

Vaccine Coverage for Primary Vaccination Series Shows Representative Timeline

The coverage for the first vaccine dose (7496/9287, 80.71%) in the eCOV study cohort showed a representative timeline, with slightly higher coverage at study end compared to the German population (78%) [30]. The percentage of eCOV study participants who received 2 doses of any vaccine increased to 61.58% (5719/9287) until study end in August 2022, which was below the vaccine coverage in the German population (76.2%; Figure 2), although the data from Germany were not matched by age in this comparison [30]. Notably, 71.4% (1286/1801) of the individuals who only reported 1 vaccination until study end dropped out after 1 week, which explains the lower percentage of vaccine coverage for 2 vaccines in our cohort compared to population data from Germany. The highest vaccine coverage for 2 doses in our cohort was seen in the younger age group (20-29 y; 2483/3764, 65.96%) and in individuals aged >60 years (630/930, 67%; Figure S10 in Multimedia Appendix 3). A high percentage (2396/3764, 63.65%) of the younger eCOV study participants were health care workers, which explains early access to vaccination. The early access to vaccination and high vaccine coverage among individuals aged >60 years possibly reflect the prioritized distribution of vaccines among older adults in Germany. At study end, only 27.98% (2599/9287) of the eCOV study cohort reported that they had received at least 3 vaccine doses, compared to the booster coverage of 61.9% in the German population during that time, which can be explained with study dropouts over time (refer to the *Reporting Behavior and Retention* subsection in this section).

Regarding the distribution of vaccine brands in our cohort, the majority of individuals who received 2 vaccine doses were vaccinated with BNT162b2 (3561/5718, 62.27%; Multimedia Appendix 5), while 0.59% (34/5718) of the individuals did not report the brand of their second vaccine dose. Among individuals who received a booster dose, 3 doses of BNT162b2 were again most common (1167/2587, 45.11%; Multimedia Appendix 9). The mean time between administered doses was 50 (SD 37.4) days between the first and second doses, 187 (SD 44.8) days between the second and third does, and 150 (SD 55.8) days between the third and fourth does.

XSL•FO
RenderX

Figure 2. Vaccine coverage over time. Line plot visualizing the vaccine coverage over time within the COVID-19 study on vaccinated and unvaccinated subjects over 16 years (eCOV study) cohort from the first possible COVID-19 vaccination date in Germany (December 26, 2020) to study end (August 1, 2022). Data are split and color coded by the number of vaccines received. The vaccine coverage data for first, second, and third vaccine doses among the German population are provided as a reference as a dashed line.

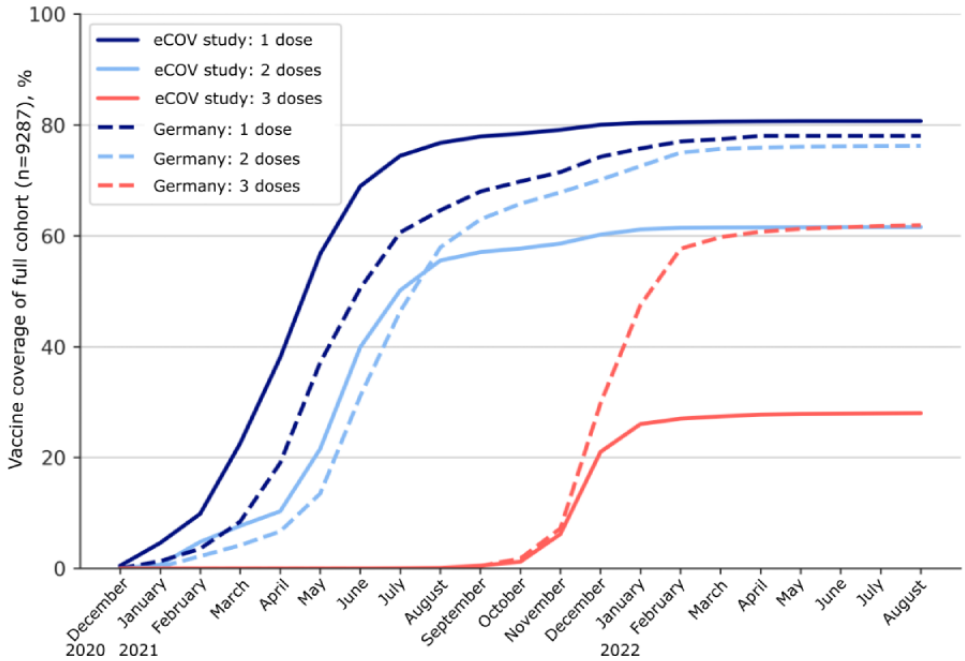

Test Positivity Rate in the eCOV Study Cohort Was Higher Than Germany-Wide Numbers

The test positivity rate for PCR testing in the eCOV study cohort mirrored the time course of the respective data for the test positivity rate in the German population, with peak rates in 2022 around calendar weeks 7 to 13 (dominant variant: Omicron BA.1) and calendar weeks 22 to 25 (dominant variant: Omicron BA.5). The test positivity rate in the eCOV study cohort was mostly higher compared to not–age-matched data from the German population over time (Figure 2) [31]. The peak around calendar week 18 in 2021 can be explained with the low sample size (58/9287, 0.62%) at the beginning of the study. For antigen tests reported in our cohort, the test positivity rate reflected the course of the pandemic but was lower compared with PCR positivity rates (Figure 3).

Figure 3. Test positivity rates in the COVID-19 study on vaccinated and unvaccinated subjects over 16 years (eCOV study) cohort and the German population over time. Calendar week 17, 2021, marks the start of the eCOV study in May 2021. PCR: polymerase chain reaction.

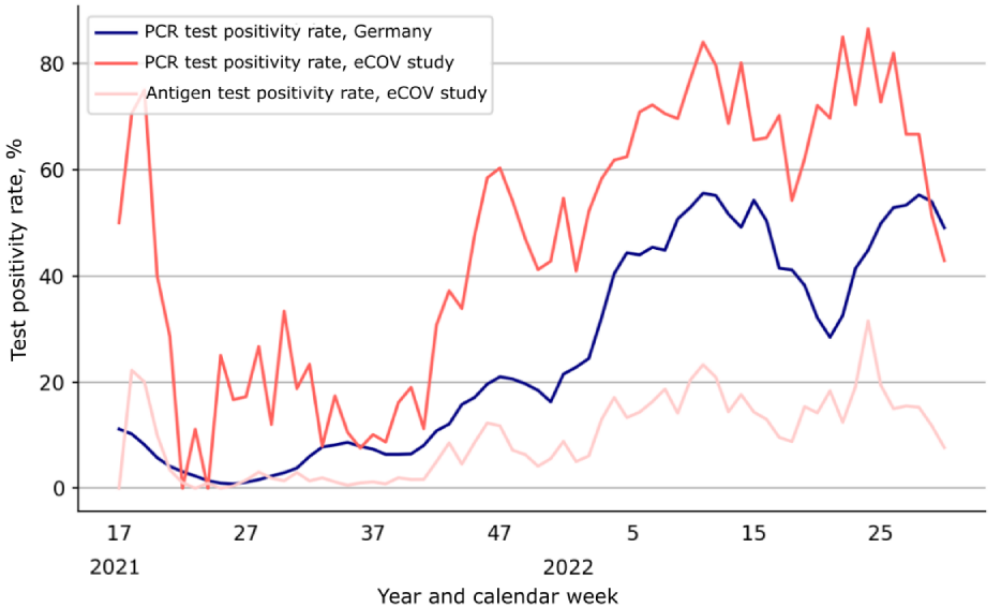

XSL•FO
RenderX

Self-Reported SARS-CoV-2 Infections Reflected the Peak Incidences in the German Population

Over the course of the study, we registered a total of 2834 PCR test–confirmed SARS-CoV-2 infections from 2743 participants. The demographics of these individuals split by COVID-19 case status, independent of vaccine status, can be found in Multimedia Appendix 10. The trajectory of COVID-19 incidence in the eCOV study cohort aligned with the age-matched Germany-wide incidences registered by public health authorities (Figure 3) [32]. Starting in calendar week 38 in 2021, the 7-day incidence in the eCOV study cohort was consistently above the Germany-wide 7-day incidence for the same age groups (Figure 4). Except for a high incidence peak during the Omicron wave at the beginning of 2022, incidences did not differ substantially between health care workers and participants not working in health care (Mann-Whitney *U* test, *P*=.56; Figure S11 in Multimedia Appendix 3). In addition, the number of reported tests did not differ between health care workers and individuals who did not work in the health care sector (Figure S12 in Multimedia Appendix 3). Overall, peak 7-day incidences were the highest during the dominance period of the Omicron variant, with the highest incidences among individuals aged 20 to 59 years (Figure S13 in Multimedia Appendix 3).

Figure 4. Comparison of 7-day incidence per 100,000 people in the COVID-19 study on vaccinated and unvaccinated subjects over 16 years (eCOV study) cohort and the German population during the study period from May 1, 2021, to August 1, 2022. The x-axis shows the timeline in calendar weeks from calendar week 5, 2021, to calendar week 30, 2022. The data for Germany consisted of the official figures reported by the national infectious disease surveillance system (refer to the Methods section).

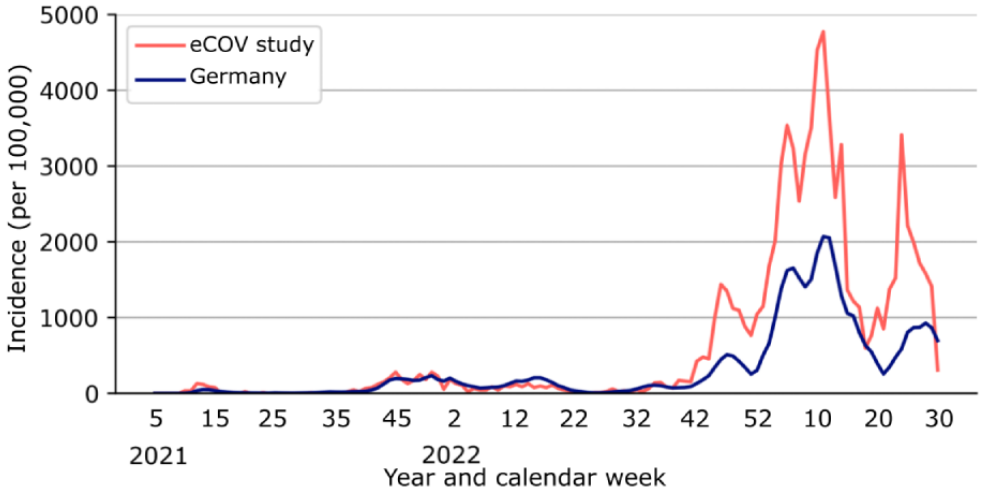

Discussion

Principal Findings

To our knowledge, this is the first digital cohort study to collect population-level data related to the COVID-19 pandemic in Germany, focusing on the real-world effectiveness of COVID-19 vaccines. We observed a gradual decline in VE against infection over time, especially in the context of the SARS-CoV-2 Omicron variant. Furthermore, we saw lower hospitalization rates among vaccinated individuals. Booster vaccinations were found to partially restore VE. The reported expected AEs, both local and systemic, aligned with results from previously published findings. Our self-reported digital data closely mirrored official surveillance trends, indicating the utility of digital apps in tracking pandemic waves. Notably, we identified potential biases, such as differences in reporting behavior between vaccinated and unvaccinated individuals, underscoring the importance of optimal study designs and the meticulous interpretation of self-reported data.

With respect to our primary outcome, the study provides valuable insights into the real-world effectiveness of COVID-19 vaccines. Our data suggest that the immunity of this young, mainly healthy, and primarily female cohort waned substantially 24 weeks after primary vaccination series with any vaccine combination and with BNT162b2. Thus, we saw a faster waning compared to previously published results from the ZOE COVID Study app in the United Kingdom [1], where VE was still hovering at 75.7% at 8 months (approximately 32 weeks) after the second dose of BNT162b2 compared to 10.4% (95% CI −11.5% to 28%) in the eCOV study cohort. The waning of VE over time was comparable to results from Qatar [33], where VE dropped to <20% after 24 weeks, and to Sweden [1], where VE waned from 92% at 4 weeks to 23% at 30 weeks. By contrast, in a multinational, placebo-controlled, observer-blinded, pivotal efficacy trial, VE against infection of any severity stayed at >90% at 24 weeks after the second vaccine dose [34]. The waning effect seen in our cohort could additionally have been influenced by the dominance period of the Omicron variant in Germany starting January 2022 because by then most of the vaccinated participants (4052/5718, 79.86%) had already received their second vaccine dose >24 weeks previously. Notably, VE in our cohort was not steadily decreasing but increasing again starting week 36 after the second dose—an observation that could also have been influenced by reporting behavior in our self-selected cohort. Vaccinations could potentially not have been reported among the unvaccinated participants, or booster effects could have occurred through undetected natural infections. We ran several sensitivity analyses adding different covariates to the logistic regression models to

XSL•FO
RenderX

check for confounders. Health worker status seemed to negatively impact VE results in weeks 28 to 36 (Figure S14 in Multimedia Appendix 3).

With regard to the secondary outcome measures of this study, exploring VE of third-dose (booster) vaccinations, our data suggest that VE against infection was increased in the first weeks after the third vaccine dose, highlighting the importance of booster vaccinations to restore protection [2,35,36,37]. Interestingly, starting week 20 after the third dose, VE dropped below 0%. As most individuals in the eCOV study cohort received their third vaccination in the time frame from November 2021 to January 2022, this again coincides with the spread of the Omicron variant in Germany. Our data therefore possibly reflect the known reduced protection of vaccination against (symptomatic) infection with the Omicron variant [38], but several other forms of bias may have played a role here. By design, we did not require individuals to get tested and submit test results on a regular basis but rather on demand. Not reporting a SARS-CoV-2 test was thus weighted as equal to not being infected. As it can be assumed that testing behavior is different in vaccinated and unvaccinated individuals [39], infections could have stayed undetected in the absence of testing. Even if individuals got tested regularly, the differences in reporting behavior between vaccinated and unvaccinated individuals could have introduced a bias: we saw that monthly questionnaires were less frequently answered by unvaccinated individuals compared to vaccinated individuals, possibly leading to the underreporting of tests, symptoms, and infections in this group (reporting bias). Apart from reporting behavior, it is also possible that risk taking, and thus the risk of becoming infected, was higher among vaccinated individuals than among unvaccinated individuals [40], leading to an increasing number of infections as vaccine protection declined.

In the eCOV study cohort, the calculated VE against symptomatic infection after the second dose of BNT162b2 was generally lower than protection against infection of any severity. This is unexpected and contradicts the current body of evidence from other studies [2,41,42]. This discrepancy in our data could again be due to underreporting among the unvaccinated participants. While 74% (734/992) of the infections in the individuals who had received 2 doses of BNT162b2 were reportedly symptomatic, only 49.9% (318/637) of the infections in unvaccinated individuals were reportedly symptomatic (chi-square test, P=.02). As we know from existing studies with larger sample sizes and different study settings [2,42,34], vaccinated individuals tend to experience fewer symptoms.

Supporting existing observational data that show a protective effect of vaccines against symptomatic infection and hospitalization [2,13,33, 43], we saw lower hospitalization rates among vaccinated individuals.

The top expected AEs, both local and systemic, after vaccination reported by our study participants were minor and aligned with those reported in other studies [2,34,43-45]. Our findings thus support extensive reports from public health authorities worldwide on the safety of COVID-19 vaccines [46-48].

In addition to valuable VE insights, our data also suggest a surveillance capability of our study app in tracking infection waves. The vaccine coverage, test positivity rates, and 7-day incidence estimates in the eCOV study cohort showed comparable trajectories to official surveillance data from public health authorities [31,32]. The test positivity rate for PCR testing was higher in the eCOV study cohort than in the German population. One reason for this might be the greater motivation to report positive tests. Notably, the antigen test positivity rate mirrored the pandemic waves, potentially providing a viable solution for pandemic surveillance at a time when PCR testing is increasingly being replaced by self-swab antigen testing. Evaluating our data on digital pandemic surveillance, we demonstrate that digital study apps can be a feasible and scalable solution for pandemic surveillance. In the future, public health authorities could leverage digital platforms for cost-efficient and fast collection of real-world data, complementing traditional surveillance systems.

In summary, our real-world VE data emphasize the importance of continued vaccination efforts in mitigating the impact of pandemics. Overall, the digital collection of self-reported data proved to be a viable tool for digital pandemic surveillance.

Limitations

It is essential to acknowledge potential biases when interpreting the self-reported data from our self-selected digital cohort. We did not collect data on ethnicity but assumed that the cohort was majority White, similar to the German population, and therefore lacked ethnic diversity. Our cohort was heterogeneous and not representative of the German normal population, with 70.72% (6568/9287) of the participants identifying as women and 38.87% (3610/9287) hailing from the health care sector. Nevertheless, health care workers in our cohort did not report more tests than non–health care workers and did not show increased infection rates. Individuals with a positive COVID-19 test might have had a higher probability to enroll in the study, possibly leading to an overestimation of COVID-19 cases in our cohort [49].

While all our recruitment efforts were aimed at recruiting a representative cohort, we were limited by the nature of a web-based study. Digital literacy bias may have influenced testing behavior because those with higher digital literacy may have been more inclined to use digital platforms for seeking testing and may have been more inclined to report positive tests in the study app. By contrast, individuals with lower digital literacy may have faced barriers in reporting their symptoms or infections through digital means. This bias can skew the data, potentially overestimating the true prevalence of symptoms and infections in the study population. This bias may further have inadvertently excluded marginalized populations and low-income groups, as they are less likely to be digitally literate. Especially for generalizable public health interventions and policies, it is important to assess and acknowledge the needs of these populations.

Moreover, participatory research may attract health-literate users who might behave differently to the general population [50]. Furthermore, health literacy and health-seeking behavior are known to be higher in women than in men [51]. Given that most of the participants were women, our study sample might have been more proactive in seeking information and adhering

XSL•FO
RenderX

to preventive measures, potentially leading to an overestimation of VE in our cohort. To improve the generalizability of future research, efforts should be made to include a more diverse and representative sample as well as participants with varying degrees of health literacy, digital literacy, and socioeconomic backgrounds.

Regarding retention, we observed a significant number of dropouts after week 1 of the study (4263/9287, 45.9%). We also observed that answer rates to questionnaires were the highest within 48 hours of email notification. Given that the majority of participants accessed the study app on their mobile devices, transitioning from a web application to a native app could allow for push notifications, potentially enhancing participation rates and retention. Exploring retention further, people who remained in the study beyond the first week stayed on average 129 (SD 114) days, indicating that retention efforts should be intensified in the early weeks after enrollment. To address potential bias, we ran a per-protocol analysis, removing individuals who had dropped out after week 1. The VE analysis still yielded similar results compared to the intention-to-treat approach, where all individuals were included (Figure S15 in Multimedia Appendix 3). Concerning long-term engagement of study participants, survey fatigue may have additionally led to underreporting of symptoms and tests, therefore leading to an underestimation of SARS-CoV-2 infections [52]. Narrowing the definition of primary outcomes and reducing the number of survey questions could help decrease survey fatigue and increase retention rates [53].

Due to the observational nature of our data, we are unable to draw definitive conclusions regarding VE. The low number of reported hospitalizations further limited our ability to calculate VE against hospitalization. Moreover, reports of AEs after vaccination were purely observational and did not allow for causal interpretation.

Conclusions

We successfully collected self-reported data on the effectiveness of COVID-19 vaccines, COVID-19 testing behavior, symptoms, and AEs after vaccination. The VE in our study cohort was comparable to that in previously published literature. Our data indicate that vaccinations are safe and effective, and third-dose vaccinations partially restore protection against infection. Overall, our data suggest that immunity against SARS-CoV-2 infection in a young, healthy, digital-literate, and predominantly women cohort wanes substantially 24 weeks after primary vaccination series with BNT162b2. Our open study design with self-enrollment led to a nonrepresentative cohort, limiting the generalizability of our study findings. Nevertheless, the data from our cohort showed comparable pandemic trajectories to official surveillance data from public health authorities, highlighting the potential of digital methods to capture the course of the pandemic from real-world data. By applying increased efforts to improve diversity in study cohorts and effectively addressing biases in self-reported data, policy makers could make use of a continuous longitudinal collection of real-world data on pandemic surveillance and VE to inform vaccination strategies. Larger sample sizes and controlled enrollment may be needed to establish causal relationships and estimate VE against specific outcomes such as hospitalization. To improve participation rates, it can be helpful to remind participants about the study through push notifications with short questionnaires that minimize survey fatigue.

Despite its limitations, our study illustrates the potential of easily scalable participatory public health tools. Serving as a novel addition to the public health infrastructure, these tools have the capacity to aid in monitoring real-world VE and enhancing national disease surveillance systems.

Acknowledgments

The authors would like to thank all study participants for contributing to this study through their data donations. The whole Data4Life team has put in great effort to design and run the study, develop the data collection infrastructure and study app, as well as in recruitment and retention. The authors want to especially thank the business intelligence and analytics team for helping with all inquiries concerning business intelligence data. This study is fully funded by Data4Life GmbH.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

NH and CR designed and initiated the *COVID-19 study on vaccinated and unvaccinated subjects over 16 years* (eCOV study). ALL and VV analyzed the data and visualized the results. ALL drafted the first version of the manuscript. SK, SS, RL, and FB advised on data analysis. All authors contributed to writing the manuscript and approved the final version.

Conflicts of Interest

ALL, NH, CR, and VV worked for Data4Life, the company that developed the study and study app as well as the underlying research infrastructure. Outside the submitted work, FB reports grants from the German Federal Ministry of Education and Research, grants from the German Federal Ministry of Health, grants from the Berlin Institute of Health, personal fees from Medtronic and Elsevier Publishing, grants from the Hans Böckler Foundation, and travel support from the Robert Koch Institute. All other authors declare no other conflicts of interest.

XSL•FO
RenderX

Multimedia Appendix 1
The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement (checklist of items that should be included in reports of cohort studies).
[DOCX File , 41 KB - jmir_v26i1e47070_app1.docx]

Multimedia Appendix 2
Examples of different marketing campaigns used to promote the COVID-19 study on vaccinated and unvaccinated subjects over 16 years (eCOV study).
[DOCX File , 8412 KB - jmir_v26i1e47070_app2.docx]

Multimedia Appendix 3
Supplementary figures.
[DOCX File , 3317 KB - jmir_v26i1e47070_app3.docx]

Multimedia Appendix 4
Overview of study questionnaires.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v26i1e47070_app4.xlsx]

Multimedia Appendix 5
Vaccine brand combination in primary vaccination series.
[XLSX File (Microsoft Excel File), 17 KB - jmir_v26i1e47070_app5.xlsx]

Multimedia Appendix 6
Results of vaccine effectiveness against symptomatic infection in the COVID-19 study on vaccinated and unvaccinated subjects over 16 years (eCOV study) cohort.
[XLSX File (Microsoft Excel File), 74 KB - jmir_v26i1e47070_app6.xlsx]

Multimedia Appendix 7
Symptoms reported during symptomatic SARS-CoV-2 infection, split by vaccination status and SARS-CoV-2 variant.
[XLSX File (Microsoft Excel File), 38 KB - jmir_v26i1e47070_app7.xlsx]

Multimedia Appendix 8
Distribution of medical treatment and hospitalization during SARS-CoV-2 infection, split by vaccine status and SARS-CoV-2 variant.
[XLSX File (Microsoft Excel File), 16 KB - jmir_v26i1e47070_app8.xlsx]

Multimedia Appendix 9
Distribution of vaccine brands for respondents who received a third vaccine dose (booster).
[XLSX File (Microsoft Excel File), 31 KB - jmir_v26i1e47070_app9.xlsx]

Multimedia Appendix 10
Descriptive statistics for entire sample, split by infection status.
[XLSX File (Microsoft Excel File), 18 KB - jmir_v26i1e47070_app10.xlsx]

References

1. Kennedy B, Fitipaldi H, Hammar U, Maziarz M, Tsereteli N, Oskolkov N, et al. App-based COVID-19 syndromic surveillance and prediction of hospital admissions in COVID symptom study Sweden. Nat Commun 2022 Apr 21;13(1):2110 [FREE Full text] [doi: 10.1038/s41467-022-29608-7] [Medline: 35449172]
2. Menni C, May A, Polidori L, Louca P, Wolf J, Capdevila J, et al. COVID-19 vaccine waning and effectiveness and side-effects of boosters: a prospective community study from the ZOE COVID Study. Lancet Infect Dis 2022 Jul;22(7):1002-1010 [FREE Full text] [doi: 10.1016/S1473-3099(22)00146-3] [Medline: 35405090]
3. Liang F. COVID-19 and health code: how digital platforms tackle the pandemic in China. Soc Media Soc 2020 Jul;6(3):2056305120947657 [FREE Full text] [doi: 10.1177/2056305120947657] [Medline: 34192023]
4. Corona data donation project. Robert Koch Institute & Research on Complex Systems of Humboldt University of Berlin. URL: https://corona-datenspende.de/science/en [accessed 2022-12-12]

XSL•FO
RenderX

5. Shapiro A, Marinsek N, Clay I, Bradshaw B, Ramirez E, Min J, et al. Characterizing COVID-19 and influenza illnesses in the real world via person-generated health data. Patterns (N Y) 2021 Jan 08;2(1):100188 [FREE Full text] [doi: 10.1016/j.patter.2020.100188] [Medline: 33506230]
6. Gadaleta M, Radin JM, Baca-Motes K, Ramos E, Kheterpal V, Topol EJ, et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. NPJ Digit Med 2021 Dec 08;4(1):166 [FREE Full text] [doi: 10.1038/s41746-021-00533-1] [Medline: 34880366]
7. Timmers T, Janssen L, Stohr J, Murk JL, Berrevoets MA. Using eHealth to support COVID-19 education, self-assessment, and symptom monitoring in the Netherlands: observational study. JMIR Mhealth Uhealth 2020 Jun 23;8(6):e19822 [FREE Full text] [doi: 10.2196/19822] [Medline: 32516750]
8. Rossman H, Keshet A, Shilo S, Gavrieli A, Bauman T, Cohen O, et al. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. Nat Med 2020 May;26(5):634-638 [FREE Full text] [doi: 10.1038/s41591-020-0857-9] [Medline: 32273611]
9. Pandit JA, Radin JM, Quer G, Topol EJ. Smartphone apps in the COVID-19 pandemic. Nat Biotechnol 2022 Jul;40(7):1013-1022. [doi: 10.1038/s41587-022-01350-x] [Medline: 35726090]
10. Bates M. Tracking disease: digital epidemiology offers new promise in predicting outbreaks. IEEE Pulse 2017;8(1):18-22. [doi: 10.1109/MPUL.2016.2627238] [Medline: 28129137]
11. Menni C, Klaser K, May A, Polidori L, Capdevila J, Louca P, et al. Vaccine side-effects and SARS-CoV-2 infection after vaccination in users of the COVID Symptom Study app in the UK: a prospective observational study. Lancet Infect Dis 2021 Jul;21(7):939-949 [FREE Full text] [doi: 10.1016/S1473-3099(21)00224-3] [Medline: 33930320]
12. Landscape of observational studies on the effectiveness of COVID-19 vaccination. World Health Organization. URL: https://www.who.int/publications/m/item/draft-landscape-of-observational-study-designs-on-the-effectiveness-of-covid-19-vaccination [accessed 2022-12-14]
13. Chemaitelly H, Yassine HM, Benslimane FM, Al Khatib HA, Tang P, Hasan MR, et al. mRNA-1273 COVID-19 vaccine effectiveness against the B.1.1.7 and B.1.351 variants and severe COVID-19 disease in Qatar. Nat Med 2021 Sep;27(9):1614-1621. [doi: 10.1038/s41591-021-01446-y] [Medline: 34244681]
14. Alanezi F, Althumairi A, Aljaffary A, Alfayez A, Alsalman D, Alhodaib H, et al. A comparative study of strategies for containing the COVID-19 pandemic in Gulf Cooperation Council countries and the European Union. Inform Med Unlocked 2021;23:100547 [FREE Full text] [doi: 10.1016/j.imu.2021.100547] [Medline: 33754126]
15. Chen H, Shi L, Zhang Y, Wang X, Jiao J, Yang M, et al. Response to the COVID-19 pandemic: comparison of strategies in six countries. Front Public Health 2021;9:708496 [FREE Full text] [doi: 10.3389/fpubh.2021.708496] [Medline: 34660510]
16. Yang J, Shi L, Chen H, Wang X, Jiao J, Yang M, et al. Strategies comparison in response to the two waves of COVID-19 in the United States and India. Int J Equity Health 2022 Apr 29;21(1):57 [FREE Full text] [doi: 10.1186/s12939-022-01666-9] [Medline: 35488277]
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
18. Medikura OP. Medikura XO life. Medikura online platform. 2023. URL: https://xo-life.com/ [accessed 2023-08-18]
19. EudraVigilance. European Medicines Agency. URL: https://www.ema.europa.eu/en/human-regulatory/research-development/pharmacovigilance/eudravigilance [accessed 2023-08-18]
20. Falldefinition Coronavirus-Krankheit-2019 (COVID-19) (SARS-CoV-2). Robert Koch-Institut. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Falldefinition.pdf?__blob=publicationFile [accessed 2023-07-02]
21. Matomo. Analytics platform. Matomo. URL: https://matomo.org/ [accessed 2023-08-18]
22. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SCIPY '10; June 28-30, 2010; Austin, TX p. 56-61. [doi: 10.25080/majora-92bf1922-00a]
23. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/mcse.2007.55]
24. Waskom M. seaborn: statistical data visualization. J Open Source Softw 2021 Apr;6(60):3021. [doi: 10.21105/joss.03021]
25. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 Sep;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
26. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SCIPY '10; June 28-July 3, 2010; Austin, TX p. 92-96 URL: http://conference.scipy.org.s3-website-us-east-1.amazonaws.com/proceedings/scipy2010/pdfs/seabold.pdf [doi: 10.25080/majora-92bf1922-011]
27. RKI Übersicht zu Omikron-Fällen in Deutschland. Robert Koch-Institut. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Omikron-Faelle/Omikron-Faelle.html [accessed 2023-01-06]
28. Farrington CP. Estimation of vaccine effectiveness using the screening method. Int J Epidemiol 1993 Aug;22(4):742-746. [doi: 10.1093/ije/22.4.742] [Medline: 8225751]
29. Guidance on conducting vaccine effectiveness evaluations in the setting of new SARS-CoV-2 variants: interim guidance, 22 July 2021. Addendum to evaluation of COVID-19 vaccine effectiveness. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-vaccine_effectiveness-variants-2021.1 [accessed 2023-01-19]
30. Impfdashboard. Bundesministerium für Gesundheit. 2022. URL: https://impfdashboard.de/ [accessed 2022-12-05]

XSL•FO
RenderX

31. Coronavirus Testzahlen gesamt. Robert Koch-Institut. 2022. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/Testzahlen-gesamt.xlsx?__blob=publicationFile [accessed 2023-01-09]
32. SurvStat@RKI 2. Robert Koch-Institut. 2022. URL: https://www.rki.de/DE/Content/Infekt/SurvStat/survstat_node.html [accessed 2022-12-01]
33. Chemaitelly H, Tang P, Hasan MR, AlMukdad S, Yassine HM, Benslimane FM, et al. Waning of BNT162b2 vaccine protection against SARS-CoV-2 infection in Qatar. N Engl J Med 2021 Dec 09;385(24):e83 [FREE Full text] [doi: 10.1056/NEJMoa2114114] [Medline: 34614327]
34. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA COVID-19 vaccine. N Engl J Med 2020 Dec 31;383(27):2603-2615 [FREE Full text] [doi: 10.1056/NEJMoa2034577] [Medline: 33301246]
35. Mallah N, Pardo-Seco J, López-Pérez LR, González-Pérez JM, Rosón B, Otero-Barrós MT, et al. Effectiveness of COVID-19 vaccine booster in the general population and in subjects with comorbidities. A population-based study in Spain. Environ Res 2022 Dec;215(Pt 2):114252 [FREE Full text] [doi: 10.1016/j.envres.2022.114252] [Medline: 36096168]
36. Intawong K, Chariyalertsak S, Chalom K, Wonghirundecha T, Kowatcharakul W, Thongprachum A, Pimpinan Khammawan, et al. Effectiveness of heterologous third and fourth dose COVID-19 vaccine schedules for SARS-CoV-2 infection during delta and omicron predominance in Thailand: a test-negative, case-control study. Lancet Reg Health Southeast Asia 2023 Mar;10:100121 [FREE Full text] [doi: 10.1016/j.lansea.2022.100121] [Medline: 36465090]
37. Barda N, Dagan N, Cohen C, Hernán MA, Lipsitch M, Kohane IS, et al. Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. Lancet 2021 Dec 04;398(10316):2093-2100 [FREE Full text] [doi: 10.1016/S0140-6736(21)02249-2] [Medline: 34756184]
38. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. COVID-19 vaccine effectiveness against the Omicron (B.1.1.529) variant. N Engl J Med 2022 Apr 21;386(16):1532-1546 [FREE Full text] [doi: 10.1056/NEJMoa2119451] [Medline: 35249272]
39. Glasziou P, McCaffery K, Cvejic E, Batcup C, Ayre J, Pickles K, et al. Testing behaviour may bias observational studies of vaccine effectiveness. J Assoc Med Microbiol Infect Dis Can 2022 Sep;7(3):242-246 [FREE Full text] [doi: 10.3138/jammi-2022-0002] [Medline: 36337606]
40. Takahashi S, Yamada S, Sasaki S, Takahashi N, Nohara M, Kawachi I. A high level of vaccine knowledge increases people's behavioral risks for contracting COVID-19 in Japan. Soc Sci Med 2022 Oct;310:115256 [FREE Full text] [doi: 10.1016/j.socscimed.2022.115256] [Medline: 36029712]
41. Haas EJ, McLaughlin JM, Khan F, Angulo FJ, Anis E, Lipsitch M, et al. Infections, hospitalisations, and deaths averted via a nationwide vaccination campaign using the Pfizer-BioNTech BNT162b2 mRNA COVID-19 vaccine in Israel: a retrospective surveillance study. Lancet Infect Dis 2022 Mar;22(3):357-366 [FREE Full text] [doi: 10.1016/S1473-3099(21)00566-1] [Medline: 34562375]
42. Lopez Bernal J, Andrews N, Gower C, Robertson C, Stowe J, Tessier E, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on COVID-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. BMJ 2021 May 13;373:n1088 [FREE Full text] [doi: 10.1136/bmj.n1088] [Medline: 33985964]
43. Nordström P, Ballin M, Nordström A. Risk of infection, hospitalisation, and death up to 9 months after a second dose of COVID-19 vaccine: a retrospective, total population cohort study in Sweden. Lancet 2022 Feb 26;399(10327):814-823 [FREE Full text] [doi: 10.1016/S0140-6736(22)00089-7] [Medline: 35131043]
44. Thomas SJ, Moreira Jr ED, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA COVID-19 vaccine through 6 months. N Engl J Med 2021 Nov 04;385(19):1761-1773 [FREE Full text] [doi: 10.1056/NEJMoa2110345] [Medline: 34525277]
45. Sadoff J, Gray G, Vandebosch A, Cárdenas V, Shukarev G, Grinsztejn B, et al. Safety and efficacy of single-dose Ad26.COV2.S vaccine against COVID-19. N Engl J Med 2021 Jun 10;384(23):2187-2201 [FREE Full text] [doi: 10.1056/NEJMoa2101544] [Medline: 33882225]
46. COVID-19 vaccines side effects and safety. National Health Service. URL: https://www.nhs.uk/conditions/covid-19/covid-19-vaccination/covid-19-vaccines-side-effects-and-safety/ [accessed 2023-07-27]
47. Virginia R, Guido R. Safety of COVID-19 vaccines. Eur J Intern Med 2023 Jun;112:15-16 [FREE Full text] [doi: 10.1016/j.ejim.2023.04.009] [Medline: 37045709]
48. Dossier coronavirus SARS-CoV-2 and COVID-19 - safety of COVID-19 vaccines. Paul-Ehrlich-Institut. URL: https://www.pei.de/EN/newsroom/dossier/coronavirus/medicine-safety.html [accessed 2023-07-27]
49. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun 2020 Nov 12;11(1):5749 [FREE Full text] [doi: 10.1038/s41467-020-19478-2] [Medline: 33184277]
50. Bonaccorsi G, Gallinoro V, Guida A, Morittu C, Ferro Allodola V, Lastrucci V, et al. Digital health literacy and information-seeking in the era of COVID-19: gender differences emerged from a Florentine university experience. Int J Environ Res Public Health 2023 Jan 31;20(3):2611 [FREE Full text] [doi: 10.3390/ijerph20032611] [Medline: 36767976]

XSL•FO
RenderX

51. Thompson AE, Anisimowicz Y, Miedema B, Hogg W, Wodchis WP, Aubrey-Bassler K. The influence of gender and other patient characteristics on health care-seeking behaviour: a QUALICOPC study. BMC Fam Pract 2016 Mar 31;17:38 [FREE Full text] [doi: 10.1186/s12875-016-0440-0] [Medline: 27036116]
52. Egleston BL, Miller SM, Meropol NJ. The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. Stat Med 2011 Dec 30;30(30):3560-3572 [FREE Full text] [doi: 10.1002/sim.4377] [Medline: 21953305]
53. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health 2011 Dec;14(8):1101-1108 [FREE Full text] [doi: 10.1016/j.jval.2011.06.003] [Medline: 22152180]

Abbreviations

AE: adverse event
eCOV study: COVID-19 study on vaccinated and unvaccinated subjects over 16 years
PCR: polymerase chain reaction
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VE: vaccine effectiveness

Edited by T de Azevedo Cardoso; submitted 09.03.23; peer-reviewed by N Mungoli, I Iyamu; comments to author 06.07.23; revised version received 18.08.23; accepted 22.03.24; published 04.06.24.

Please cite as:
Lang AL, Hohmuth N, Višković V, Konigorski S, Scholz S, Balzer F, Remschmidt C, Leistner R
COVID-19 Vaccine Effectiveness and Digital Pandemic Surveillance in Germany (eCOV Study): Web Application–Based Prospective Observational Cohort Study
J Med Internet Res 2024;26:e47070
URL: https://www.jmir.org/2024/1/e47070
doi:10.2196/47070
PMID:38833299

©Anna-Lena Lang, Nils Hohmuth, Vukašin Višković, Stefan Konigorski, Stefan Scholz, Felix Balzer, Cornelius Remschmidt, Rasmus Leistner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Creation of Standardized Common Data Elements for Diagnostic Tests in Infectious Disease Studies: Semantic and Syntactic Mapping

Caroline Stellmach[1], MSc; Sina Marie Hopff[2], Dr med; Thomas Jaenisch[3], Dr med, PhD; Susana Marina Nunes de Miranda[2], Dr rer nat; Eugenia Rinaldi[1], MSc; The NAPKON, LEOSS, ORCHESTRA, and ReCoDID Working Groups[4]

[1]Berlin Institute of Health, Charité - Universitätsmedizin Berlin, Berlin, Germany

[2]Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Department I of Internal Medicine, University Hospital Cologne and Faculty of Medicine, University of Cologne, Cologne, Germany

[3]Heidelberg Institut für Global Health, Universitätsklinikum Heidelberg, Heidelberg, Germany

[4]See Acknowledgments

Corresponding Author:
Caroline Stellmach, MSc
Berlin Institute of Health
Charité - Universitätsmedizin Berlin
Anna-Louisa-Karsch-Str 2
Berlin, 10178
Germany
Phone: 49 15752614677
Email: caroline.stellmach@charite.de

Abstract

Background: It is necessary to harmonize and standardize data variables used in case report forms (CRFs) of clinical studies to facilitate the merging and sharing of the collected patient data across several clinical studies. This is particularly true for clinical studies that focus on infectious diseases. Public health may be highly dependent on the findings of such studies. Hence, there is an elevated urgency to generate meaningful, reliable insights, ideally based on a high sample number and quality data. The implementation of core data elements and the incorporation of interoperability standards can facilitate the creation of harmonized clinical data sets.

Objective: This study's objective was to compare, harmonize, and standardize variables focused on diagnostic tests used as part of CRFs in 6 international clinical studies of infectious diseases in order to, ultimately, then make available the panstudy common data elements (CDEs) for ongoing and future studies to foster interoperability and comparability of collected data across trials.

Methods: We reviewed and compared the metadata that comprised the CRFs used for data collection in and across all 6 infectious disease studies under consideration in order to identify CDEs. We examined the availability of international semantic standard codes within the Systemized Nomenclature of Medicine - Clinical Terms, the National Cancer Institute Thesaurus, and the Logical Observation Identifiers Names and Codes system for the unambiguous representation of diagnostic testing information that makes up the CDEs. We then proposed 2 data models that incorporate semantic and syntactic standards for the identified CDEs.

Results: Of 216 variables that were considered in the scope of the analysis, we identified 11 CDEs to describe diagnostic tests (in particular, serology and sequencing) for infectious diseases: viral lineage/clade; test date, type, performer, and manufacturer; target gene; quantitative and qualitative results; and specimen identifier, type, and collection date.

Conclusions: The identification of CDEs for infectious diseases is the first step in facilitating the exchange and possible merging of a subset of data across clinical studies (and with that, large research projects) for possible shared analysis to increase the power of findings. The path to harmonization and standardization of clinical study data in the interest of interoperability can be paved in 2 ways. First, a map to standard terminologies ensures that each data element's (variable's) definition is unambiguous and that it has a single, unique interpretation across studies. Second, the exchange of these data is assisted by "wrapping" them in a standard exchange format, such as Fast Health care Interoperability Resources or the Clinical Data Interchange Standards Consortium's Clinical Data Acquisition Standards Harmonization Model.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50049) doi:10.2196/50049

KEYWORDS

core data element; CDE; case report form; CRF; interoperability; semantic standards; infectious disease; diagnostic test; covid19; COVID-19; mpox; ZIKV; patient data; data model; syntactic interoperability; clinical data; FHIR; SNOMED CT; LOINC; virus infection; common element

Introduction

In response to the spread of SARS-CoV-2 starting in late 2019, large-scale observational studies as well as clinical trials have been launched worldwide to gain insights into disease patterns, treatment options, prevention measures, severity, and outcomes [1]. New findings related to the diagnosis, prevention, and treatment of many infectious diseases, including COVID-19, heavily rely on data generated by diagnostic tools and laboratory analysis of the pathogen and host response [2].

Immunological testing has become a cost- and time-efficient way to monitor infections [3]. Hence, a growing number of clinical studies include biosample information as part of their data collection targets, particularly results of analytical tests performed on blood samples [4].

Data from patients enrolled in a study are commonly collected using a case report form (CRF) [5]. The International Conference on Harmonization Guidelines for Good Clinical Practice defines a CRF as a "printed, optical or electronic document designed to record all of the protocol-required information to be reported to the sponsor on each trial subject" [6]. Since the design of a CRF can affect study outcomes, time and resources need to be invested to maximize the quality of the data collected and ensure that good clinical practice guidelines are being followed [7].

The identification of common data elements (CDEs), each comprising 1 or more questions and respective answer value sets, is an approach to standardize data collection instruments (ie, CRFs) across studies [8]. A CDE may also contain standardized ontology concepts directly or include a link to the unique identifier for an appropriate ontology concept [9].

We have previously described [10] how incorporating standard codes into clinical trials metadata can increase their findability, accessibility, interoperability, and reusability (FAIR)ness [11]. The FAIR principles are recognized internationally as important guides to conducting research [12]. Interoperability, in particular, is defined as the ability of several systems to exchange information, as well as read and use the received information without requiring further preprocessing [13]. Although there are several levels of interoperability [14], the focus of this study in the context of health care data was on semantic (use of standard terminologies and classifications) and syntactic (implementation of a standard exchange format) interoperability.

The use of data standards when designing CRFs can serve multiple purposes: in addition to supporting data quality, it facilitates the merging and exchange of data from multiple sources, as well as subsequent analysis [5]. International standards development organizations (SDOs), such as Health Level Seven (HL7) or Integrating the Healthcare Enterprise (IHE), promote and coordinate the use of these standards [15]. HL7 has developed the exchange standard Fast Healthcare Interoperability Resource (FHIR), which allows for the exchange of health-related information based on packaging it into so-called resources. The FHIR can represent a wide range of data, particularly those generated in care settings [16]. In comparison, the Clinical Data Interchange Standards Consortium (CDISC) has published standards for the representation of CRF data used in clinical trials [17].

By mapping study data elements to international semantic standard codes, the included concepts receive an unambiguous definition that is tied to an identifier that makes it machine-readable [18]. Among the widely used terminologies and classifications for health care concepts are the Logical Observation Identifiers Names and Codes (LOINC) and the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). The National Cancer Institute Thesaurus (NCIt) is also available as a reference terminology focused, among others, on translational research and clinical care information [19]. LOINC provides standard codes (each comprising a set of an identifier, a name, and a code) for laboratory observations, documents, and questionnaires [20]. SNOMED CT covers a broad range of health care information, and each of its concepts has a unique identifier and is defined by a description and 1 or more relationships [21].

In this study, we set out to analyze CRF variables from 6 study protocols capturing information about diagnostic testing with the purpose of identifying CDEs specific to infectious diseases. The selected studies investigated 3 different infectious diseases in humans: COVID-19 [1], monkeypox (mpox) [22], and Zika [23]. The CRF variables we included originate from the International Severe Acute Respiratory and emerging Infection Consortium (ISARIC) COVID-19 Core CRF [24], as well as from 3 of the many international research projects focused on gaining new insights into SARS-CoV-2: the ORCHESTRA project [25], the Intersectoral Platform (SUEP) of the National Pandemic Cohort Network (NAPKON SUEP) study [26], and the Lean European Open Survey on SARS-CoV-2 (LEOSS) [27] study. Additionally, we analyzed the World Health Organization (WHO) CRF on the mpox infection [28] and the Zika CRFs of ZIKAlliance [29].

Our goal of proposing standardized paninfectious disease CRF variables for diagnostic testing information for use in CRFs was broken down into 3 subtasks: (1) identification of interstudy CDEs, (2) creation of a preliminary map of the CDEs to semantic standard codes, and (3) development of a proposed mapping of the CDEs to the FHIR syntax standards [30] and the CDISC's standards for data collection [17].

XSL•FO
RenderX

Methods

Ethical Considerations

Since only CRF metadata (meaning definitions of questions and answers used to comprise CRFs) were used and no actual patient data were reviewed in this study, ethics approval was not required.

Study Design

Figure 1 provides a graphical overview of the steps we followed to create a standardized set of variables for use in data collection instruments in infectious disease studies focusing on diagnostic testing.

We examined 6 CRFs provided to us by 4 research consortia, and we downloaded the publicly available CRFs from the ISARIC and WHO websites [24]. We proceeded to extract diagnostic testing variables from each CRF and organized them for analysis and comparison in a Microsoft Excel sheet.

The following CRFs were included:

- ORCHESTRA work package 6 CRF [31]
- Cross-sectoral platform (SUEP) CRF of NAPKON [32]
- LEOSS study [27,33] electronic case report form (eCRF)
- ISARIC-WHO COVID-19 core CRF [24]
- Zika study CRF
- Mpox study CRF [28]

We translated the variables from the NAPKON SUEP study from German into English to harmonize it with the language of the other selected studies (English). The study manager verified the translation.

Figure 1. Flowchart describing the methodology of identifying common, standardized CRF variables reporting on diagnostic testing for use in infectious disease studies. CDE: common data element; CRF: case report form; ISARIC: International Severe Acute Respiratory and emerging Infection Consortium; LEOSS: Lean European Open Survey on SARS-CoV 2; mpox: monkeypox; NAPKON SUEP: Intersectoral Platform (SUEP) of the National Pandemic Cohort Network.

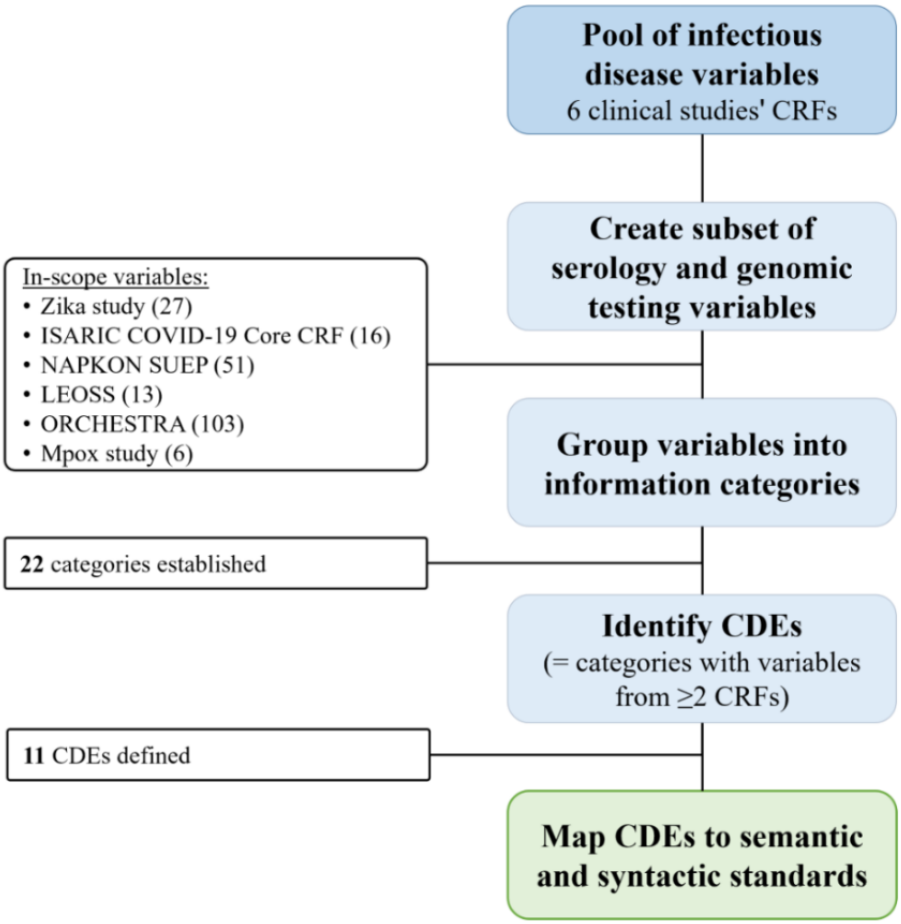

Common Data Elements

In the first step of analyzing the study metadata, we reviewed all CRF variables (questions and answers). Adopting the National Institutes of Health's methodology to derive CDEs [34], we created common categories to group variables based on the key information they contained. We then reviewed the newly organized variables to determine which categories were present in at least 2 (33%) of the 6 CRFs. These common variables then formed the basis as newly identified CDEs for infectious diseases.

For each of these preliminary CDEs, the extensive value set (sum of all unique answers) across all reviewed CRFs was determined. If necessary, we created value set subsets based on informational content and pathogen type.

XSL•FO
RenderX

Mapping to Standards

Each CDE (question and value set) was then mapped to the appropriate semantic standard code(s) and FHIR element(s). We searched for available terminology codes using the NCIt browser (version 23.02d, release date February 27, 2023), the SearchLOINC tool (v2.26), and the SNOMED CT browser (version 2023-03-31). If no semantic standard code was found, we prepared a submission to request the creation of a new code, depending on the informational domain, with NCIt, SNOMED CT, or LOINC.

Results

CRF Analysis

The analysis of the CRFs used in 6 infectious disease studies led to the identification of 216 variables focusing on diagnostic testing, which were in the scope of further analysis: 103 (47.7%) from ORCHESTRA, 51 (23.6%) from NAPKON SUEP, 27 (12.5%) from the Zika study, 16 (7.4%) from the ISARIC CRF, 13 (6%) from the LEOSS survey, and 6 (2.8%) from the mpox study (Table S1 in Multimedia Appendix 1 [25,28,32,35-42]). These diagnostic testing variables could be grouped into 22 newly defined categories, which are shown in Table S2 in Multimedia Appendix 1.

Common Data Elements

Based on the analysis of the 6 CRFs, we identified 11 CDEs, each of which was present in at least 2 (33%) of the 6 reviewed data collection instruments and reflected diagnostic testing information applicable to infectious disease studies. We mapped these CDEs to semantic standard codes and FHIR resources (illustrated in Figure 2), as well as to the CDISC (Multimedia Appendix 2).

Figure 2. List of 11 CDEs identified based on the review of CRF variables from 6 infectious disease studies capturing diagnostic testing information. Also noted are the proposed data type and suggested mapping to the FHIR (version R4). For each CDE, the data type that was most commonly used across the reviewed data collection instruments is shown. CDE: common data element; CRF: case report form; FHIR: Fast Healthcare Interoperability Resources.

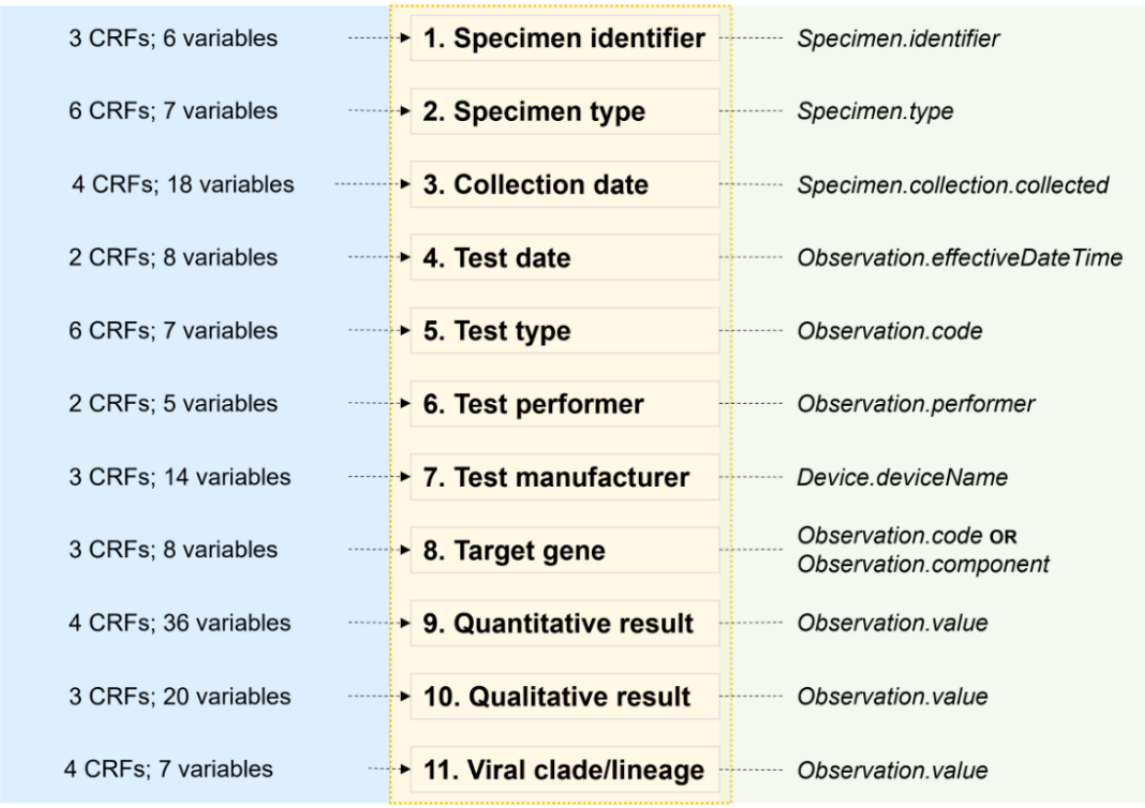

Viral Lineage/Clade

The first CDE was defined as “viral lineage” or “viral clade.” Depending on the virus investigated, its value sets would vary to reflect the applicable clade and lineage details, as exemplified in Figure 3.

Genetic diversity, as described in a phylogenetic tree, is classified by clades. A clade, also called genotype or subtype, comprises a set of lineages that are all descended from only 1 ancestor, common to them [43].

ORCHESTRA and the human mpox study contained 3 (1.4%) variables providing monkeypox virus (MPXV) and SARS-CoV-2 clade details. In addition, viral lineage information was collected from ORCHESTRA, the ISARIC CRF, and NAPKON SUEP across 4 (1.9%) variables.

There is no uniform convention for naming viral clades and lineages. In the case of SARS-CoV-2, the most widely used nomenclatures for subtypes are provided by the Global Initiative on Sharing All Influenza Data [44], Rambault et al [43], and Nextstrain [45], which differ in the position at which clades are differentiated from one another.

XSL•FO
RenderX

Figure 3. Two example variables that were grouped into the "viral lineage" CDE from two CRFs that were reviewed. CDE: common data element; CRF: case report form.

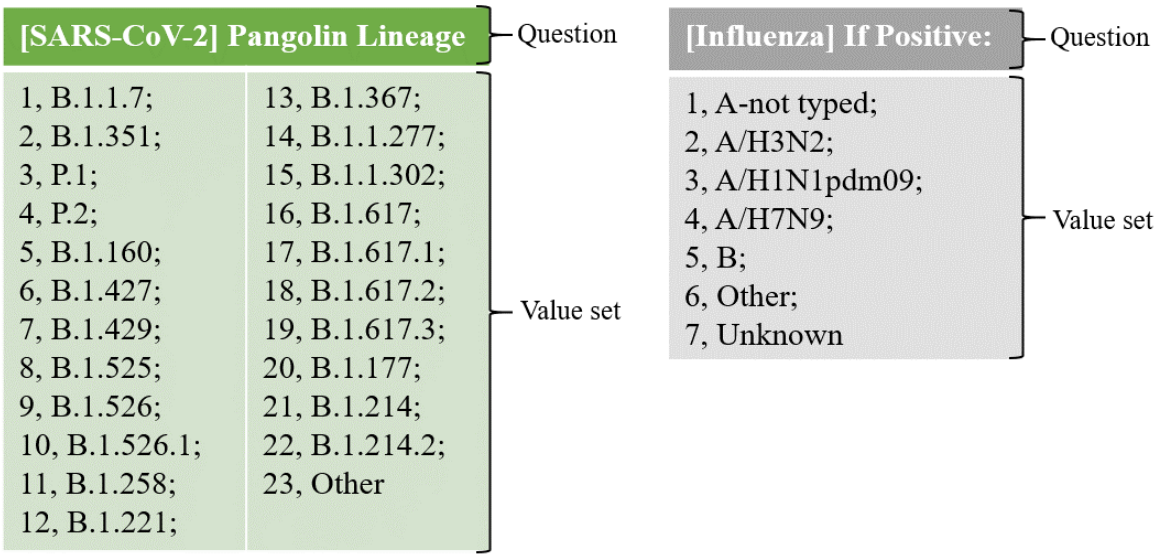

Specimen Identifier, Specimen Collection Date, and Specimen Type

Our analysis led to the identification of "specimen identifier," "specimen collection date," and "specimen type" as additional CDEs across the 6 studies. ORCHESTRA and the Zika and mpox studies included 6 (2.8%) variables that were grouped as "specimen identifiers" and had a free-text format. Any biological specimen (ie, blood, urine, cerebrospinal fluid, feces) used for laboratory analysis must be uniquely identified so that the resulting findings are associated with the right patient. Identifiers might contain a patient's first and last names, birth date, medical facility number, or a unique, randomly generated code [46]. In addition to this internal laboratory-based specimen identifier, a particular specimen might have a second, external identifier that is assigned when results based on the analysis of said specimen are uploaded to a public/restricted databases or to a biobank [47].

Furthermore, 8 (3.7%) variables across the Zika study, NAPKON SUEP, ORCHESTRA, and ISARIC CRFs constituted the data element "specimen collection date," requiring the input of a date format (mm/dd/yyyy). The specimen collection date marks the date on which a specimen was collected from a patient and placed in a specimen container for ensuing processing and analysis.

Details about the kind of specimen collected and used for analysis are provided by the coded "specimen type" CDE. All 6 reviewed studies included the data element "specimen type" in their variables. Our analysis led to the finding that there tended to be 2 axes involved in the value set elements of the specimen type, which covered information about the method used to collect the specimen (ie, swab) and the site of origin (ie, skin lesion). Examples are shown in Multimedia Appendix 3.

Test Date and Test Performer

The CDEs "test date" and "test performer" included variables from the ORCHESTRA and Zika study CRFs and the ORCHESTRA and NAPKON SUEP CRFs, respectively. The test date refers to the calendar date on which a particular laboratory diagnostic test (specified by the CDE "test type") was conducted. The "test performer" CDE captures the full name of the individual(s) executing this diagnostic test in free-text format.

Test Type, Target Gene, and Test Manufacturer

The coded CDE "test type" captures a specific laboratory test, which in this context would fall into 3 main categories: serology, sequencing, and polymerase chain reaction (PCR) analysis. All 6 reviewed CRFs included variables providing details about diagnostic tests. For serology tests, the test type in the analyzed SARS-CoV-2 studies provided details on the method, along with the analyzed target, whereas in the Zika study, only the target was given (Multimedia Appendix 4).

In the context of COVID-19 research, lateral flow testing, immunofluorescence assay (IFA), enzyme-linked immunosorbent assay (ELISA), and chemiluminescence immunoassay (CLIA) are frequently used methods for the diagnosis of infections [48]. The detection of Zika and mpox infections is usually also based on serology, specifically ELISA-based antibody measurements [23].

The coded CDE "target gene" grouped 8 (3.7%) variables across the NAPKON SUEP, LEOSS, and ORCHESTRA CRFs. It refers to the target of a genome-focused diagnostic test, such as PCR or a sequencing method. Using primers that contain bases that are complementary to a conserved sequence within the target gene of a particular virus, this sequence, if present in the biological sample, is amplified and can be detected through PCR [49].

In total, 13 (6%) variables used across the Zika, NAPKON SUEP, and ORCHESTRA CRFs were grouped into the coded CDE "test manufacturer." This data element provides information about the manufacturer of the diagnostic test (ie, kit or testing system). For example, the following PCR systems

XSL•FO
RenderX

(manufacturers) were mentioned in a study variable in the NAPKON SUEP CRF: Seegene (Allplex) [50], altona Diagnostics (RealStar) [51], and Roche Deutschland Holding (cobas) [52].

Qualitative and Quantitative Results

All reported results of diagnostic testing covered by variables in the 6 CRFs we reviewed could be clustered into either qualitative or quantitative results, and thus, they formed the last 2 (18%) of 11 coded CDEs that we identified. A qualitative result details the findings about the presence or absence of a measured observable, such as virus-specific antibody or gene material. In contrast, a quantitative result constitutes numeric measurements (see Table S3 in Multimedia Appendix 1). In the studies that we analyzed, those numeric values were given for the titer, cycle threshold, and concentration of the same observables mentioned before.

Mapping to Standards

Semantic Standards

To facilitate semantic interoperability of the proposed diagnostic testing CDEs, we suggested mapping each CDE and respective value set to the terminology standards SNOMED CT, LOINC, and NCIt. For each CDE, we created a suggested mapping that covers the variable itself and a nonexclusive list of possible value set elements (Table S4 in Multimedia Appendix 1).

The CDEs "viral lineage" and "viral clade" could be mapped to the following NCIt codes (code and description are shown), respectively: "C60792 Lineage" and "C179767 Clade." Depending on the analyzed virus, the value sets (answers) could differ and be represented through mapping to either NCIt or LOINC codes. For example, in the case of detection of the SARS-CoV-2 variant B.1.1.7, the NCIt code "C179573 SARS Coronavirus 2 B.1.1.7" or the LOINC code "LA31705-9 SARS-CoV-2 B.1.1.7 lineage" is available.

The CDE "specimen identifier" could be represented in a standardized way using SNOMED CT, LOINC, and NCIt terms, as shown in Table S4 in Multimedia Appendix 1. Likewise, codes from all 3 standards were available to represent the free-text CDEs "specimen collection date" and "specimen type."

There are semantic standard concepts available to describe the "test date" and "test performer" CDEs. Using SNOMED CT codes from the "procedure" hierarchy or using NCIt terms, diagnostic test types, such as serology assays, sequencing, and PCR, can be described in a standardized manner. Incidentally, there are a few standard codes available to represent the value sets for "target gene" (for the envelope gene in SNOMED CT and a few in the NCIt), although not necessarily specifically meant to map viral pathogens' genes (exception in the NCIt: "C19108 Viral Envelope Gene"). Thus, we prepared a submission to the NCIt for the creation of concepts that cover the prominently analyzed SARS-CoV-2 [53] and Zika virus (ZIKV) genes [54]. We submitted 33 concepts for code creation to the SDOs LOINC and NCIt (Table S5 in Multimedia Appendix 1).

No SNOMED CT codes were available to describe the value set elements for the "test manufacturer" CDE. However, both the NCIt and LOINC provide terms for this purpose; the NCIt has created concepts for specific COVID-19 diagnostic kits, detailing the manufacturer, analytical target, and method. Likewise, LOINC has created codes that bundle several kits into a single term, such as "94558-4 SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay," which represents 4 commercially available kits [55].

There are generic semantic terms from SNOMED CT and the NCIt to describe the "quantitative result" and "qualitative result" CDEs in a standardized manner, which can be used across viral pathogen studies, such as "Laboratory Test Result" or just "Result." However, this would omit the distinction between "qualitative" and "quantitative."

LOINC provides a comprehensive list of terms to describe qualitative results of laboratory diagnostic tests for SARS-CoV-2 and antibody measurements specific to ZIKV.

The use of the SNOMED CT terminology requires a country (or institutional) license. SNOMED International has, however, been releasing its Global Patient Set containing currently around 24,000 concepts, which can be used free of charge [56]. Of the 90 SNOMED CT codes, 33 (37%) that we included in the exemplary value set mappings for our proposed infectious disease diagnostic CDEs are covered by the Global Patient Set.

Syntax Standard

We proposed a preliminary mapping of the diagnostic testing CDEs to FHIR (version R4) elements as a first step toward establishing syntactical interoperability (Figure 2, right). Of the 11 CDEs that we identified, 8 (72.7%) were mapped to the *Observation* resource and the remaining 3 (27.3%) to the *Specimen* resource.

Additionally, we provided a preliminary suggested mapping of the FHIR elements to the CDISC according to the *FHIR to CDISC Joint Mapping Implementation Guide v1.0* [57] (see Multimedia Appendix 2).

Discussion

Principal Findings

Common Data Elements

Resulting from the review of 6 CRFs, we identified 11 panstudy CDEs that capture key diagnostic testing information commonly collected across the reviewed infectious disease studies. These CDEs were purposefully kept generic to enhance the probability that they could be adopted by researchers and integrated into data collection instruments of other infectious disease studies, even if a different pathogen was studied. The pathogen under investigation in a given study would determine the value set elements of CDEs of the coded data.

The CDEs "viral lineage" and "viral clade" provide the means to describe genetic relatedness of viruses, which is critical to pathogen surveillance and relies on the availability of well-defined nomenclature [58]. Currently, no panvirus approach to naming viral clades and lineages exists. The International Committee on Taxonomy of Viruses, founded in 1966, has the goal to develop a taxonomy for viruses and establish names for

XSL•FO
RenderX

viral taxa based on international agreement. However, the International Committee on Taxonomy of Viruses does not address the naming of viral clades and lineages [59]. In the context of ensuring that diagnostic testing results are linked to the right sample (specimen) and patient, the CDEs "specimen identifier," "specimen collection date," and "specimen type" are important parameters. Regarding the diagnostic test itself, documentation of the CDEs "test date" and "test performer" can help identify quality problems retrospectively. Diagnostic testing results can be split into the CDEs "qualitative result" and "quantitative result," which would confirm the presence/absence of signs of a pathogen or numeric values of measured observables, such as antibody titers. The CDEs "test type," "target gene," and "test manufacturer" provide all complementary details to the diagnostic tests conducted. Along with the increasing inclusion of molecular testing variables in the study of infectious diseases, we expect that this number of recurring elements (which would be candidate CDEs) that describe diagnostic tests across different studies will continue to grow.

The power of research findings can be expanded through combining data from several clinical studies for analysis in an effort to create a larger data set. Without considering privacy or legal considerations, the basis for merging data from different sources is that the correct information (ie, data variables) is linked together to ensure accuracy and avoid misinterpretation. Defining standardized CDEs that serve as a common language across clinical studies is one way to approach this challenge [9]. Lin et al [5] described a similar approach of how CRF design can be optimized for data harmonization by creating a pool of reusable CDEs. There are numerous examples for the creation of CDEs for specific medical specialties and use cases, such as stroke trials [60], pregnancy pharmacovigilance [61], and COVID-19 [62]. This includes a set of CDEs on the quality of life in neurological disorders, as well as the PhenX Toolkit to capture key information on phenotypes [8].

Mapping to Standards

To facilitate interoperability of study data in particular, we proposed a mapping of the identified CDEs to semantic and syntactic standards. We also created a table with practical examples of available standard codes to identify value set concepts ambiguously for variables contained in CRFs from studies focused on SARS-CoV-2, ZIKV, and MPXV (see Table S4 in Multimedia Appendix 1).

In the past, we have described how semantic interoperability standard codes can be integrated directly into the study metadata to facilitate merging, sharing, and analysis of patient data that are being collected across several clinical studies and cohort types, where several methods for data storage and collection have been used [10]. Kush et al [9] and Kersloot et al [11], among others, have discussed the advantage of introducing interoperability standards prior to data collection rather than retrospectively with the aim to save time and other resources.

An important aspect of mapping study data to semantic standard concepts is choosing appropriate terminology. Although there is no universal guidance for this process, we can draw instructive conclusions from our attempt to propose a mapping for the CDEs we identified for which we searched within the LOINC, SNOMED CT and NCIt, terminologies.

The selection of semantic standards to represent CDEs and their value sets depends on the way the CDEs (and underlying CRF variables) are phrased with regard to the level of detail and the kind of information that are described. The category of information covered by a CRF variable is the first "filter" for finding the appropriate terminology. The NCIt, which is managed by the National Cancer Institute, focuses on providing a vocabulary for the cancer domain [63]; hence, it comprises many (gen)omics-related terms. Each NCIt term is represented by a code and a name and has several annotations [64].

In contrast, the LOINC coding system, which is published by the Regenstrief Institute, is used by numerous large laboratories and government agencies, such as the Centers for Disease Control and Prevention, to describe laboratory and clinical findings, as well as documents [65]. Although LOINC has a clear focus on representing laboratory terms, SNOMED CT terms have a broader coverage of information and are commonly used to represent clinical information in electronic health records [63]. SNOMED CT and th eNCIt both provide concepts that are suitable to describe variables and value sets if they are kept more generic in their wording. LOINC terms, in contrast, are specific and should only be used to represent questions, not value sets. Contrary to the NCIt, SNOMED CT comprises a limited set of concepts to describe genomic methods and results.

Unlike the use of LOINC and the NCIt, embedding SNOMED CT concepts into the metadata of research data requires a license. In recent years, many countries have purchased a SNOMED CT affiliate license or become a SNOMED CT member, including Germany, Spain, and Portugal [66].

The LOINC coding system includes suitable codes for several of the CDEs we defined. For example, we chose the concept "95609-4 SARS-CoV-2 (COVID-19) S gene [Presence] in Respiratory specimen by Sequencing" as 1 of the available standard terms for coding the "qualitative result" CDE. However, it also covers the "target gene" (S gene), "specimen type" (respiratory specimen), and "test type" (sequencing) CDEs. Another aspect that should be kept in mind, especially concerning selecting standard terms for the "quantitative result" CDE when used in a CRF, is that the units of the result should be clearly defined and match those of the standard term. Although each LOINC term has a defined unit, SNOMED CT concepts do not necessarily implicitly or explicitly define units. The concept "1240461000000109 Measurement of severe acute respiratory syndrome coronavirus 2 antibody (observable entity)" has no unit of measure attached and hence can be used if a CRF variable can be measured using several different units. A standard way to describe units is offered by the Unified Code for Units of Measure [67].

Regarding finding the appropriate standard code for viral lineage, the more general-purpose terminology of SNOMED CT does not include the required level of detail for this CDE, which is captured in the NCIt. However, the list of microorganisms defined as concepts by SNOMED CT under the hierarchy "organism" is detailed and can be used to describe a pathogen. The hierarchical organization of SNOMED CT,

XSL•FO
RenderX

which also includes sublevels of concepts, provides a clear idea of the positioning of any microorganism within the complex classification of organisms overall.

As knowledge rapidly evolves in health care, missing concepts are regularly added to ontologies. The process involves concept creation requests from the public, which are submitted to the SDOs. Zheng et al [63] describe an approach of using formal concept analysis to identify missing concepts in the NCIt and SNOMED CT.

We also proposed a mapping of the 11 diagnostic testing CDEs to the corresponding FHIR (version R4) element. This provides data with a standardized exchange format, which can incorporate standard terminologies. Elements in the *Specimen* [68] and *Observation* [69] (and for the test manufacturer, also *Device* [70]) resources can be used to represent all 11 CDEs.

Limitations

The identified CDEs focus on diagnostic tests used in infectious disease studies. Additional CDEs that would fall into other informational categories (eg, therapeutics or comorbidities) were not considered as they were out of the scope of our study. Furthermore, since the reviewed ORCHESTRA variables include CRF variables from several COVID-19 studies, the selection of protocols might appear unbalanced.

Conclusion

The need to investigate COVID-19 quickly and extensively has made the pool of available variables describing diagnostic tests particularly abundant. Kush et al [9] point out that although the name “CDE” implies that these elements are common, they are not so commonly used. This is due to a lack of mandatory requirements for their use [9]. A necessary step to increase the adoption and value of CDEs would be that funding bodies (eg, the National Institutes of Health or the European Commission) in collaboration with SDOs create and impose mandatory requirements for the implementation of existent CDEs on recipients of project funding.

Acknowledgments

This material contains content from Logical Observation Identifiers Names and Codes (LOINC) [71]. LOINC is the copyright of Regenstrief Institute, Inc, and the LOINC Committee and is available at no cost under the license [72]. LOINC is a registered United States trademark of Regenstrief Institute, Inc.

This material also includes content from the National Cancer Institute Thesaurus, published by the National Cancer Institute [19].

The Systemized Nomenclature of Medicine - Clinical Terms (SNOMED CT) was used by the permission of SNOMED International. SNOMED CT was originally created by the College of American Pathologists. “SNOMED,” “SNOMED CT,” and “SNOMED Clinical Terms” are registered trademarks of SNOMED International [73].

We would like to thank all 3 SDOs for their collaboration and support with new term submissions and SNOMED International, in particular, for granting us permission to display and share the suggested terminology bindings.

The ORCHESTRA project received funding from the European Union’s Horizon 2020 Research and Innovation Program (grant agreement 101016167). The ZIKAlliance project received funding from the European Union’s Horizon 2020 Research and Innovation Program (grant agreement 734548). The ReCODID project received funding from the European Union’s Horizon 2020 Research and Innovation Program (grant agreement 825746). The Lean European Open Survey on SARS-CoV-2 (LEOSS) registry was supported by the German Centre for Infection Research (DZIF) and the Willy Robert Pitzer Foundation. The National Pandemic Cohort Network (NAPKON) is part of the Network University Medicine and was funded by the German Federal Ministry of Education and Research (FKZ: 01KX2021). Parts of the infrastructure of the Würzburg study site were supported by the Bavarian Ministry of Research and Art to support coronavirus research projects. Parts of the NAPKON project suite and study protocols of the cross-sectoral cohort platform are based on projects funded by the DZIF.

The members of the working groups are as follows: NAPKON Working Group: Gabriele Anton, Katharina Appel, Sabine Blaschke, Isabel Bröhl, Johanna Erber, Karin Fiedler, Ramsia Geisler, Peter U. Heuschmann, Thomas Illig, Monika Kraus, Dagmar Krefting, Jens-Peter Reese, Margarete Scherer, Jörg Janne Vehreschild, Maria J.G.T. Vehreschild, and Luise Wolf. ORCHESTRA Working Group: Chiara Dellacasa, Miroslav Puskaric, Thomas Osmo, Elisa Rossi, and Anna Gorska. LEOSS Working Group: Jörg Janne Vehreschild, Carolin E. M. Koll, Margarete Scherer, and Maria J.G.T. Vehreschild. ReCoDID Working Group: Lauren Maxwell, Heather Hufstedler, and Frank Tobian.

Data Availability

The analyzed case report forms are stored and available in Excel format for review on the project’s online data repository [74].

Authors' Contributions

CS and ER created the first draft of the manuscript, together with SMH, SMNdM, and TJ. All authors reviewed the draft, commented on it, and provided revisions, and they have approved the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 47 KB - jmir_v26i1e50049_app1.docx]

Multimedia Appendix 2
Map between the proposed FHIR representation (FHIR resources) and the corresponding CDISC elements for the 11 identified diagnostic testing CDEs for infectious disease studies. CDE: common data element; CDISC: Clinical Data Interchange Standards Consortium; FHIR: Fast Healthcare Interoperability Resources.
[XLSX File (Microsoft Excel File), 32 KB - jmir_v26i1e50049_app2.xlsx]

Multimedia Appendix 3
Example specimen types listed within CRF variables in the analyzed infectious disease studies. CRF: case report form.
[PNG File , 94 KB - jmir_v26i1e50049_app3.png]

Multimedia Appendix 4
Example variables (question and value set) from 2 of the reviewed CRFs. One came from the ZIKV study and the other from the NAPKON SUEP study. CLIA: chemiluminescence immunoassay; CRF: case report form; ELISA: enzyme-linked immunosorbent assay; IFA: immunofluorescence assay; IgA: immunoglobulin A; IgG: immunoglobulin G; IgM: immunoglobulin M; LFT: lateral flow immunoassay; PCR: polymerase chain reaction; SGTF: S gene target failure; VNTR: variable number of tandem repeats; WES: whole exome sequencing; WGS: whole genome sequencing; ZIKV: Zika virus.
[PNG File , 82 KB - jmir_v26i1e50049_app4.png]

References

1. He Z, Erdengasileng A, Luo X, Xing A, Charness N, Bian J. How the clinical research community responded to the COVID-19 pandemic: an analysis of the COVID-19 clinical studies in ClinicalTrials.gov. JAMIA Open 2021 Apr;4(2):ooab032 [FREE Full text] [doi: 10.1093/jamiaopen/ooab032] [Medline: 34056559]
2. Radzikowska U, Ding M, Tan G, Zhakparov D, Peng Y, Wawrzyniak P, et al. Distribution of ACE2, CD147, CD26, and other SARS-CoV-2 associated molecules in tissues and immune cells in health and in asthma, COPD, obesity, hypertension, and COVID-19 risk factors. Allergy 2020 Nov;75(11):2829-2845 [FREE Full text] [doi: 10.1111/all.14429] [Medline: 32496587]
3. Fox T, Geppert J, Dinnes J, Scandrett K, Bigio J, Sulis G, Cochrane COVID-19 Diagnostic Test Accuracy Group. Antibody tests for identification of current and past infection with SARS-CoV-2. Cochrane Database Syst Rev 2022 Nov 17;11(11):CD013652 [FREE Full text] [doi: 10.1002/14651858.CD013652.pub2] [Medline: 36394900]
4. Courtot M, Gupta D, Liyanage I, Xu F, Burdett T. BioSamples database: FAIRer samples metadata to accelerate research data management. Nucleic Acids Res 2022 Jan 07;50(D1):D1500-D1507 [FREE Full text] [doi: 10.1093/nar/gkab1046] [Medline: 34747489]
5. Lin C, Wu N, Liou D. A multi-technique approach to bridge electronic case report form design and data standard adoption. J Biomed Inform 2015 Feb;53:49-57 [FREE Full text] [doi: 10.1016/j.jbi.2014.08.013] [Medline: 25200473]
6. Committee for Human Medicinal Products. Guideline for good clinical practice E6(R2). European Medicines Agency. 2016 Jan 12. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-good-clinical-practice-e6r2-step-5_en.pdf [accessed 2023-02-14]
7. Bellary S, Krishnankutty B, Latha M. Basics of case report form designing in clinical research. Perspect Clin Res 2014 Oct;5(4):159-166 [FREE Full text] [doi: 10.4103/2229-3485.140555] [Medline: 25276625]
8. Cohen MZ, Thompson CB, Yates B, Zimmerman L, Pullen CH. Implementing common data elements across studies to advance research. Nurs Outlook 2015;63(2):181-188 [FREE Full text] [doi: 10.1016/j.outlook.2014.11.006] [Medline: 25771192]
9. Kush R, Warzel D, Kush M, Sherman A, Navarro E, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. J Biomed Inform 2020 Jul;107:103421 [FREE Full text] [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]
10. Rinaldi E, Stellmach C, Rajkumar NMR, Caroccia N, Dellacasa C, Giannella M, et al. Harmonization and standardization of data for a pan-European cohort on SARS- CoV-2 pandemic. NPJ Digit Med 2022 Jun 14;5(1):75 [FREE Full text] [doi: 10.1038/s41746-022-00620-x] [Medline: 35701537]
11. Kersloot MG, Jacobsen A, Groenen KH, Dos Santos Vieira B, Kaliyaperumal R, Abu-Hanna A, et al. De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into

XSL•FO
RenderX

machine-readable data. J Biomed Inform 2021 Oct;122:103897 [FREE Full text] [doi: 10.1016/j.jbi.2021.103897] [Medline: 34454078]
12. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet 2018 Jul;26(7):931-936 [FREE Full text] [doi: 10.1038/s41431-018-0160-0] [Medline: 29777206]
13. IEEE. IEEE standard computer dictionary: a compilation of IEEE standard computer glossaries. IEEE Std 610 1991 Jan 18:1-217. [doi: 10.1109/IEEESTD.1991.106963]
14. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
15. Lee S, Do H. Comparison and Analysis of ISO/IEEE 11073, IHE PCD-01, and HL7 FHIR Messages for Personal Health Devices. Healthc Inform Res 2018 Jan;24(1):46-52 [FREE Full text] [doi: 10.4258/hir.2018.24.1.46] [Medline: 29503752]
16. Benson T, Grieve G. Principles of FHIR. In: Benson T, editor. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. London: Springer-Verlag London; Jul 01, 2016:329-348.
17. Facile R, Muhlbradt EE, Gong M, Li Q, Popat V, Pétavy F, et al. Use of Clinical Data Interchange Standards Consortium (CDISC) Standards for Real-world Data: Expert Perspectives From a Qualitative Delphi Survey. JMIR Med Inform 2022 Jan 27;10(1):e30363 [FREE Full text] [doi: 10.2196/30363] [Medline: 35084343]
18. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
19. NCI thesaurus. National Cancer Institute. URL: https://ncithesaurus.nci.nih.gov/ncitbrowser/ [accessed 2023-10-09]
20. About LOINC. Regenstrief Institute. URL: https://loinc.org/about/ [accessed 2022-11-10]
21. Millar J. The Need for a Global Language - SNOMED CT Introduction. Stud Health Technol Inform 2016;225:683-685. [Medline: 27332304]
22. Jalilian S, Bastani MN. The Mpox, serious menace, or paper tiger? Iran J Microbiol 2022 Dec;14(6):770-777 [FREE Full text] [doi: 10.18502/ijm.v14i6.11251] [Medline: 36721442]
23. Panning M. Zika Virus Serology: More Diagnostic Targets, more Reliable Answers? EBioMedicine 2017 Feb;16:12-13 [FREE Full text] [doi: 10.1016/j.ebiom.2017.01.045] [Medline: 28174133]
24. World Health Organization, ISARIC. COVID-19 core case report form. Acute respiratory infection clinical characterisation data tool. ISARIC. URL: https://isaricdev.wpenginepowered.com/wp-content/uploads/2021/02/ISARIC-WHO-COVID-19-CORE-CRF_EN.pdf [accessed 2023-03-20]
25. Work packages. ORCHESTRA. URL: https://orchestra-cohort.eu/work-packages/ [accessed 2023-05-09]
26. Vehreschild J. Intersectoral platform (SÜP) of the National Pandemic Cohort Network (NAPKON) (SUEP-NAPKON). ClinicalTrials.gov. 2021 Feb 24. URL: https://clinicaltrials.gov/ct2/show/record/NCT04768998 [accessed 2023-03-20]
27. Study protocol: LEOSS; Lean European Open Survey on SARS-CoV-2. Deutsche Gesellschaft für Infektiologie (DGI). URL: https://www.dgi-net.de/wp-content/uploads/2020/03/LEOSS-Protocol-Submission-1-20200316.pdf [accessed 2023-04-04]
28. Mpox case investigation form (CIF). World Health Organization. 2022 Dec 22. URL: https://cdn.who.int/media/docs/default-source/documents/health-topics/monkeypox/mpox_cif-narrative_epi_20221222.pdf?sfvrsn=d52108e5_1 [accessed 2023-04-04]
29. ZIKAlliance and ReCoDID working together to promote data and sample sharing across infectious disease cohort studies. ZIKAlliance. URL: https://zikalliance.tghn.org/articles/zikalliance-and-recodid-working-together-promote-data-and-sample-sharing-across-infectious-disease-cohort-studies/ [accessed 2023-04-04]
30. Benson T, Grieve G. HL7 dynamic model. In: Benson T, editor. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Cham: Springer International Publishing; Jul 01, 2016:303-309.
31. Gupta A, Konnova A, Smet M, Berkell M, Savoldi A, Morra M, mAb ORCHESTRA working group, et al. Host immunological responses facilitate development of SARS-CoV-2 mutations in patients receiving monoclonal antibody treatments. J Clin Invest 2023 Mar 15;133(6):e166032 [FREE Full text] [doi: 10.1172/JCI166032] [Medline: 36727404]
32. Schons M, Pilgram L, Reese J, Stecher M, Anton G, Appel KS, NAPKON Research Group. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. Eur J Epidemiol 2022 Aug;37(8):849-870 [FREE Full text] [doi: 10.1007/s10654-022-00896-z] [Medline: 35904671]
33. FAQ - Lean European Open Survey on SARS-CoV-2 infected patients. LEOSS. URL: https://leoss.net/faq/ [accessed 2023-03-20]
34. Glossary - NINDS common data elements. National Institutes of Health. URL: https://www.commondataelements.ninds.nih.gov/glossary [accessed 2023-10-09]
35. ISARIC Clinical Characterisation Group. The value of open-source clinical science in pandemic response: lessons from ISARIC. Lancet Infect Dis 2021 Dec;21(12):1623-1624 [FREE Full text] [doi: 10.1016/S1473-3099(21)00565-X] [Medline: 34619109]
36. About - REDCap. Vanderbilt. URL: https://projectredcap.org/about/ [accessed 2024-06-07]

XSL•FO
RenderX

37. Pilgram L, Schons M, Jakob CEM, Claßen AY, Franke B, Tscharntke L, et al. [The COVID-19 pandemic as an opportunity and challenge for registries in health services research: lessons learned from the Lean European Open Survey on SARS-CoV-2 infected patients (LEOSS)]. Gesundheitswesen 2021 Nov;83(S 01):S45-S53. [doi: 10.1055/a-1655-8705] [Medline: 34731893]
38. Lean European Open Survey on SARS-CoV-2 infected patients (LEOSS). URL: https://leoss.net/ [accessed 2023-06-07]
39. Avelino-Silva VI, Mayaud P, Tami A, Miranda MC, Rosenberger KD, Alexander N, ZIKAlliance Clinical Study Group. Study protocol for the multicentre cohorts of Zika virus infection in pregnant women, infants, and acute clinical cases in Latin America and the Caribbean: the ZIKAlliance consortium. BMC Infect Dis 2019 Dec 26;19(1):1081 [FREE Full text] [doi: 10.1186/s12879-019-4685-9] [Medline: 31878895]
40. NAPKON. Nationales Pandemie Kohorten Netz. 2023. URL: https://napkon.de/ [accessed 2023-06-05]
41. Follow-up of COVID-19 long term sequelae. ClinicalTrials.gov. 2021 Dec. URL: https://clinicaltrials.gov/ct2/show/NCT05097677 [accessed 2024-05-29]
42. Monitoring COVID-19 vaccination response in fragile populations (ORCHESTRA-4). ClinicalTrials.gov. 2022 Feb. URL: https://clinicaltrials.gov/ct2/show/NCT05222139 [accessed 2023-03-21]
43. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020 Nov;5(11):1403-1407 [FREE Full text] [doi: 10.1038/s41564-020-0770-5] [Medline: 32669681]
44. GISAID. Re3data.org. URL: https://www.re3data.org/repository/r3d100010126 [accessed 2022-10-03]
45. Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months. Nextstrain. 2024 Apr 23. URL: https://nextstrain.org/ncov/gisaid/global/6m [accessed 2023-03-28]
46. Ernst DJ, Martel AM, Astin D, Dew TR, Dietz, Jr RL, Dubrowny N, et al. GP33: accuracy in patient and specimen identification. Clinical and Laboratory Standards Institute. 2019 Apr. URL: https://clsi.org/media/3121/gp33-ed2_sample.pdf [accessed 2023-10-03]
47. Durant TJ, Gong G, Price N, Schulz WL. Bridging the Collaboration Gap: Real-time Identification of Clinical Specimens for Biomedical Research. J Pathol Inform 2020;11:14 [FREE Full text] [doi: 10.4103/jpi.jpi_15_20] [Medline: 32477620]
48. Gong F, Wei HX, Li Q, Liu L, Li B. Evaluation and Comparison of Serological Methods for COVID-19 Diagnosis. Front Mol Biosci 2021;8:682405 [FREE Full text] [doi: 10.3389/fmolb.2021.682405] [Medline: 34368226]
49. Kuchinski KS, Loos KD, Suchan DM, Russell JN, Sies AN, Kumakamba C, et al. Targeted genomic sequencing with probe capture for discovery and surveillance of coronaviruses in bats. Elife 2022 Nov 08;11:e79777 [FREE Full text] [doi: 10.7554/eLife.79777] [Medline: 36346652]
50. AllplexTM 2019-nCoV assay. Seegene. URL: https://www.seegene.com/assays/allplex_2019_ncov_assay [accessed 2023-10-07]
51. RealStar® product range. altona Diagnostics. URL: https://www.altona-diagnostics.com/en/realstar-product-range-a.html [accessed 2023-10-07]
52. cobas® modular platform. Roche Deutschland Holding. URL: https://www.roche.de/diagnostik/produkte-loesungen/systeme/cobas-modular-platform [accessed 2023-10-07]
53. Brant AC, Tian W, Majerciak V, Yang W, Zheng Z. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. Cell Biosci 2021 Jul 19;11(1):136 [FREE Full text] [doi: 10.1186/s13578-021-00643-z] [Medline: 34281608]
54. Lundberg R, Melén K, Westenius V, Jiang M, Österlund P, Khan H, et al. Zika Virus Non-Structural Protein NS5 Inhibits the RIG-I Pathway and Interferon Lambda 1 Promoter Activation by Targeting IKK Epsilon. Viruses 2019 Nov 04;11(11):1024 [FREE Full text] [doi: 10.3390/v11111024] [Medline: 31690057]
55. Center for Surveillance, Epidemiology and Laboratory Services (CSELS). LOINC in vitro diagnostic (LIVD) test code mapping. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/csels/dls/livd-codes.html [accessed 2023-04-03]
56. Global Patient Set. SNOMED International. URL: https://www.snomed.org/gps [accessed 2023-10-10]
57. FHIR to CDISC joint mapping implementation guide 1.0.0 - STU 1. HL7 International. URL: http://hl7.org/fhir/uv/cdisc-mapping/STU1/ [accessed 2023-10-04]
58. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro Surveill 2020 Aug;25(32):2001410 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.32.2001410] [Medline: 32794443]
59. Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Dutilh BE, Harrach B, et al. Binomial nomenclature for virus species: a consultation. Arch Virol 2020 Feb;165(2):519-525 [FREE Full text] [doi: 10.1007/s00705-019-04477-6] [Medline: 31797129]
60. Jabal MS, Ibrahim MK, Thurnham J, Kallmes KM, Kobeissi H, Ghozy S, et al. Common Data Elements Analysis of Mechanical Thrombectomy Clinical Trials for Acute Ischemic Stroke with Large Core Infarct. Clin Neuroradiol 2023 Jun;33(2):307-317. [doi: 10.1007/s00062-022-01239-x] [Medline: 36520186]
61. Richardson JL, Moore A, Bromley RL, Stellfeld M, Geissbühler Y, Bluett-Duncan M, et al. Core Data Elements for Pregnancy Pharmacovigilance Studies Using Primary Source Data Collection Methods: Recommendations from the IMI

XSL•FO
RenderX

ConcePTION Project. Drug Saf 2023 May;46(5):479-491 [FREE Full text] [doi: 10.1007/s40264-023-01291-7] [Medline: 36976447]
62. Weissman A, Cheng A, Mainor A, Gimbel E, Nowak K, Pan H, et al. Development and implementation of the National Heart, Lung, and Blood Institute COVID-19 common data elements. J Clin Transl Sci 2022;6(1):e142 [FREE Full text] [doi: 10.1017/cts.2022.466] [Medline: 36590348]
63. Zheng L, Min H, Chen Y, Keloth V, Geller J, Perl Y, et al. Outlier concepts auditing methodology for a large family of biomedical ontologies. BMC Med Inform Decis Mak 2020 Dec 15;20(Suppl 10):296 [FREE Full text] [doi: 10.1186/s12911-020-01311-x] [Medline: 33319713]
64. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI thesaurus. Comp Funct Genomics 2004;5(8):648-654 [FREE Full text] [doi: 10.1002/cfg.445] [Medline: 18629178]
65. Stram M, Gigliotti T, Hartman D, Pitkus A, Huff SM, Riben M, et al. Logical Observation Identifiers Names and Codes for Laboratorians. Arch Pathol Lab Med 2020 Feb;144(2):229-239 [FREE Full text] [doi: 10.5858/arpa.2018-0477-RA] [Medline: 31219342]
66. Our members. SNOMED International. URL: https://www.snomed.org/members [accessed 2023-04-04]
67. Rychert J. In support of interoperability: A laboratory perspective. Int J Lab Hematol 2023 Aug;45(4):436-441. [doi: 10.1111/ijlh.14113] [Medline: 37337695]
68. Specimen - FHIR v4.0.1. HL7 International. URL: http://hl7.org/fhir/R4/specimen.html [accessed 2023-03-13]
69. Observation - FHIR v4.0.1. HL7 International. URL: http://hl7.org/fhir/R4/observation.html [accessed 2023-04-04]
70. Device - FHIR v4.0.1. HL7 International. URL: http://hl7.org/fhir/R4/device.html [accessed 2023-04-06]
71. Home - LOINC. Regenstrief Institute. URL: https://loinc.org/ [accessed 2023-10-09]
72. Knowledge base - LOINC. Regenstrief Institute. URL: https://loinc.org/kb/ [accessed 2023-10-09]
73. Home - SNOMED International. SNOMED International. URL: https://www.snomed.org [accessed 2023-10-09]
74. Stellmach C, Rinaldi E. Raw study data - CDEs for infectious disease studies. ORCHESTRA Cloud. URL: https://cloud.orchestra-cohort.eu/s/HnG8YzDC4WNFsBY [accessed 2023-10-09]

Abbreviations

CDE: common data element
CDISC: Clinical Data Interchange Standards Consortium
CRF: case report form
ELISA: enzyme-linked immunosorbent assay
FAIR: findability, accessibility, interoperability, and reusability
FHIR: Fast Healthcare Interoperability Resource
HL7: Health Level Seven
IHE: Integrating the Healthcare Enterprise
ISARIC: International Severe Acute Respiratory and emerging Infection Consortium
LEOSS: Lean European Open Survey on SARS-CoV-2
LOINC: Logical Observation Identifiers Names and Codes
mpox: monkeypox
MPXV: monkeypox virus
NAPKON SUEP: Intersectoral Platform (SUEP) of the National Pandemic Cohort Network
NCIt: National Cancer Institute Thesaurus
PCR: polymerase chain reaction
SDO: standards development organization
SNOMED CT: Systematized Nomenclature of Medicine - Clinical Terms
WHO: World Health Organization
ZIKV: Zika virus

Edited by A Mavragani; submitted 19.06.23; peer-reviewed by K Ndlovu, S Hume; comments to author 19.09.23; revised version received 10.10.23; accepted 18.01.24; published 10.06.24.

Please cite as:
Stellmach C, Hopff SM, Jaenisch T, Nunes de Miranda SM, Rinaldi E, The NAPKON, LEOSS, ORCHESTRA, and ReCoDID Working Groups
Creation of Standardized Common Data Elements for Diagnostic Tests in Infectious Disease Studies: Semantic and Syntactic Mapping
J Med Internet Res 2024;26:e50049
URL: https://www.jmir.org/2024/1/e50049
doi:10.2196/50049
PMID:38857066

XSL•FO
RenderX

©Caroline Stellmach, Sina Marie Hopff, Thomas Jaenisch, Susana Marina Nunes de Miranda, Eugenia Rinaldi, The NAPKON, LEOSS, ORCHESTRA, and ReCoDID Working Groups. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Physiological Response to the COVID-19 Vaccine: Insights From a Prospective, Randomized, Single-Blinded, Crossover Trial

Andjela Markovic[1,2,3,4], PhD; Vladimir Kovacevic[4,5], PhD; Timo B Brakenhoff[6], PhD; Duco Veen[7,8], PhD; Paul Klaver[6], MSc; Marianna Mitratza[9], PhD; George S Downward[9], MBChB, PhD; Diederick E Grobbee[9], MD, PhD; Maureen Cronin[4], MD, PhD; Brianna M Goodale[4,6], PhD; COVID-19 Remote Early Detection (COVID-RED) consortium[10]

[1]Department of Psychology, University of Fribourg, Fribourg, Switzerland
[2]Department of Social Neuroscience and Social Psychology, Institute of Psychology, University of Bern, Bern, Switzerland
[3]Department of Pulmonology, University Hospital Zurich, Zurich, Switzerland
[4]Ava Aktiengesellschaft (AG), Zurich, Switzerland
[5]The Institute for Artificial Intelligence Research and Development of Serbia, Belgrade, Serbia
[6]Julius Clinical, Zeist, Netherlands
[7]Department of Methodology & Statistics, Utrecht University, Utrecht, Netherlands
[8]Optentia Research Programme, North-West University, Potchefstroom, South Africa
[9]Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, Netherlands
[10]See Acknowledgments

Corresponding Author:
Andjela Markovic, PhD
Department of Social Neuroscience and Social Psychology
Institute of Psychology
University of Bern
Fabrikstrasse 8
Bern, 3012
Switzerland
Phone: 41 316844724
Email: andjela.markovic@unibe.ch

Abstract

Background: Rapid development and implementation of vaccines constituted a crucial step in containing the COVID-19 pandemic. A comprehensive understanding of physiological responses to these vaccines is important to build trust in medicine.

Objective: This study aims to investigate temporal dynamics before and after COVID-19 vaccination in 4 physiological parameters as well as the duration of menstrual cycle phases.

Methods: In a prospective trial, 17,825 adults in the Netherlands wore a medical device on their wrist for up to 9 months. The device recorded their physiological signals and synchronized with a complementary smartphone app. By means of multilevel quadratic regression, we examined changes in wearable-recorded breathing rate, wrist skin temperature, heart rate, heart rate variability, and objectively assessed the duration of menstrual cycle phases in menstruating participants to assess the effects of COVID-19 vaccination.

Results: The recorded physiological signals demonstrated short-term increases in breathing rate and heart rate after COVID-19 vaccination followed by a prompt rebound to baseline levels likely reflecting biological mechanisms accompanying the immune response to vaccination. No sex differences were evident in the measured physiological responses. In menstruating participants, we found a 0.8% decrease in the duration of the menstrual phase following vaccination.

Conclusions: The observed short-term changes suggest that COVID-19 vaccines are not associated with long-term biophysical issues. Taken together, our work provides valuable insights into continuous fluctuations of physiological responses to vaccination and highlights the importance of digital solutions in health care.

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-021-05241-5

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51120) doi:10.2196/51120

KEYWORDS

wearable technology; biosignals; digital health; SARS-CoV-2; vaccine reactogenicity; menstrual cycle; vaccine; wearables; sex; development; implementation; medical device; breathing rate; heart rate; biological mechanism; immune response; vaccination

Introduction

In December 2020, the first COVID-19 vaccine worldwide was made available to the public under emergency use authorization to contain the COVID-19 pandemic [1]. Subsequently, national COVID-19 vaccination programs were rolled out across the globe. COVID-19 vaccines were shown to significantly reduce the risk of infection [2] as well as reinfection [3]. Furthermore, it was demonstrated that hospitalization due to COVID-19 was less likely in vaccinated individuals, as was a severe disease course with progression to mechanical ventilation or death [4].

Similar to routine vaccines, mild side effects have been reported for COVID-19 vaccines. These side effects most commonly include fever, headache, fatigue, and pain at the injection site [5]. More severe but rare side effects include anaphylaxis [6] and thromboembolic events [7]. A prospective observational study in the United Kingdom examined self-reported systemic (eg, headache and fever) and local (eg, pain and redness at the injection site) side effects. It found that side effects in the community were less prevalent than what was expected from phase 3 clinical trials [2]. They were also more prevalent in females than males, in younger individuals compared to older individuals, and in individuals with a higher BMI compared to those with a lower BMI.

Additionally, side effects in the study by Menni et al [2] were more common in individuals who had had a previous SARS-CoV-2 infection compared to their infection-free peers. The authors proposed that the higher reactogenicity in previously infected individuals may be linked to higher immunogenicity leading to higher antibody titers as previously observed [8]. Postvaccination physical symptoms arise as manifestations of inflammatory events mimicking the response to a natural infection [9]. Such inflammatory events are important for the development of protection against the vaccine-targeted disease.

Whether triggered by a vaccine or the disease itself, inflammatory immune responses are often detectable via changes in physiological signals (eg, high fever). Risch et al [10] documented an increase in heart rate, breathing rate, and wrist skin temperature as well as a decrease in heart rate variability during a SARS-CoV-2 infection. Their findings underscore a broader shift toward personalized medicine, with recent systematic reviews demonstrating how monitoring physiological signals can serve as an important tool in digital health care [11]. In particular, modern wearable technology offers an opportunity to collect real-time physiological data remotely, reliably and continuously in order to inform medical decisions. Recently, physiological signals have been suggested as a proxy for vaccine reactogenicity [12]. In response to COVID-19 vaccination, a short-term increase in heart rate has consistently been measured by means of wearable technology [12-15]. Several confounding factors have been identified including age, vaccine type, and prior SARS-CoV-2 infection [12]. Furthermore, differences between female and male physiological reaction to vaccination have been reported [12], highlighting the importance of considering sex as a biological variable in vaccine-induced immune response.

Further supporting sex-based analyses, physiological signals are strongly impacted by hormones and the menstrual cycle [16-18]. For example, shifts in wrist-skin temperature, heart rate, and breathing rate have been demonstrated and tracked by wearable devices across the menstrual cycle [16]. Accordingly, sex differences in the physiological response to vaccines are important to take into account. Previously, female participants were often underrepresented or entirely excluded in clinical trials, leading to a lack of evidence regarding the response to specific drugs and treatments in female patients [19]. However, different immune responses to non–COVID-19 vaccines have been observed between female and male participants including higher antibody responses and more reported adverse events in female participants [20]. Therefore, it is crucial to stratify reports on vaccine biology by sex [21].

Furthermore, anecdotal reports of COVID-19 vaccine effects on the menstrual cycle have led to discussions in the media and caused concerns among the general public [22,23]. Thus far, only small, short-term decreases have been observed in self-reported menstrual cycle duration following vaccination [24,25]. Moreover, among 39,129 respondents to a web-based survey, increased bleeding after vaccination was reported by 42% of people with regular menstrual cycles [26]. However, such changes in menstrual bleeding intensity do not necessarily indicate changes in fertility following vaccination [27]. Similarly, contrary to initial public concern, COVID-19 mRNA vaccination led to *increases* in sperm concentration and motility as well as semen volume [28]. In contrast, a SARS-CoV-2 infection in male participants may be associated with a short-term decline in fertility [27], thereby suggesting vaccinations may have a potential protective factor for male fertility.

In this study, we sought to quantify the physiological immune response to COVID-19 vaccination by looking for sex-based differences in wearable-measured physiological parameters. We investigated temporal dynamics of 4 physiological parameters: breathing rate, wrist skin temperature, heart rate, and heart rate variability before and after COVID-19 vaccination. Further focusing on sex-specific changes, we objectively assessed the duration of different phases of the menstrual cycle before and after vaccination. Since information about potential vaccine-induced changes to specific menstrual cycle phases is lacking and discrepancies in these biologically distinct periods may affect the interpretation of vaccination effects, this is an important knowledge gap to fill.

XSL•FO
RenderX

Methods

Study Design

In a prospective trial, the COVID-19 Remote Early Detection (COVID-RED) trial [29], we investigated the use of a wrist-worn device for remote early detection of SARS-CoV-2 infections. The trial has been registered at the Netherlands Trial Register (see [29] for details and the study protocol). The study's main objectives aimed to detect infection-based deviations in the measured physiological parameters from their baseline in both male and female participants; the current analysis constitutes an exploratory objective of the main COVID-RED clinical trial.

Over the course of the COVID-RED trial (from February to November 2021), 17,825 participants from the Netherlands wore the CE-certified and Food and Drug Administration–cleared Ava Fertility Tracker (Ava AG) for up to 9 months. Originally built to detect a woman's fertile days in real time, the wearable bracelet tracks biophysical changes in heart rate, breathing rate, skin temperature, and heart rate variability with 3 types of sensors: a photoplethysmograph, an accelerometer and a temperature sensor. In this study, sensor data were recorded every 10 seconds during the user's sleep and synchronized with a complementary smartphone app ("Ava COVID-RED") upon waking. The COVID-RED study procedure has been previously published [29]. Briefly, in addition to wearing the Ava bracelet nightly, participants entered information about their daily health status and behavior including experience of any symptoms and vaccinations in the Ava COVID-RED app. Furthermore, they filled out biweekly surveys allowing us to collect information on health care use, COVID-19 testing, behavioral changes, and COVID-19 vaccination. Finally, they collected a capillary blood sample upon enrolment and every 3 months throughout the trial allowing for a laboratory confirmation of a SARS-CoV-2 infection. Based on these data, we examined the trajectory of physiological parameters in a subset of participants who reported at least one COVID-19 vaccination event across the two sources (ie, Ava COVID-RED app and the biweekly survey).

Statistical Analysis

Only participants with successful synchronizations of bracelet data with the Ava COVID-RED app on 15 of 20 days before as well as after their first COVID-19 vaccination were included in the current analyses. All participants who reported more than one vaccination event within this period (ie, 20 days before and 20 days after the first COVID-19 vaccination) were excluded to facilitate analysis of specific effects associated with the first dose of the COVID-19 vaccine.

Part 1: Physiological Signals Around Vaccination

With a total of 2189 participants from 19 to 80 years old (mean 50.32, SD 13.17 years; 1613 female; see Table 1), regression models were constructed to examine the dynamics of 4 physiological parameters around vaccination (ie, breathing rate in breaths per minute, skin temperature in degree Celsius, heart rate in beats per minute, and heart rate variability). Heart rate variability was quantified as the ratio of heart rate oscillations with low frequency (0.04-0.15 Hz) to those with high frequency (0.15-0.4 Hz), as described by Risch et al [10]. For each of the 4 parameters, 4 models were built and compared with regard to their Akaike Information Criterion, conditional R^2, and root-mean-square error with daily parameter measurements as the outcome variable. The four models included: (1) only a random intercept; (2) sex, menstruation status, age, BMI, type of vaccine (ie, Pfizer-BioNTech [BNT162], Moderna [mRNA-1273], AstraZeneca [AZD1222], and Janssen [Ad26.COV2.S]), existence of laboratory-confirmed SARS-CoV-2 infection prior to vaccination, and sleep duration as fixed factors and number of days from vaccination as a fixed quadratic term; (3) random intercept in addition to the fixed factors from model 2; and (4) interaction of the quadratic term with sex, existence of SARS-CoV-2 infection prior to vaccination, and sleep duration in addition to all terms from model 3. We did not expect a linear change in physiological parameters in response to vaccination; rather, we sought to model a quadratic effect of vaccination-induced changes reflected in a short-term alteration in biophysical parameters before a return to prior baseline values. For this term, we quantified the time from vaccination in number of days with 0 reflecting the day of vaccination, negative values reflecting days preceding vaccination and positive values reflecting days following vaccination. This method adequately models the initial impact and subsequent adjustment postvaccination, better reflecting the complex temporal dynamics than linear models. In line with standard approaches to time-varying effect sizes [30], it reflects the differential impact of COVID-19 vaccination on biophysical parameters before their return to baseline. The quadratic multilevel model provides a nuanced analysis of physiological responses over time, justifying its selection over simpler approaches for our longitudinal data. For models 2-4, we categorized female participant's menstruation status into 3 categories: not menstruating (including both pregnant and postmenopausal participants), perimenopausal, and menstruating. Interaction terms were included in model 4 to capture potential differences in factors previously identified as potentially influencing COVID-19 vaccine reactogenicity [12].

XSL•FO
RenderX

Table 1. Demographic characteristics stratified by vaccine type.

Characteristics	All COVID-19 vaccines		Pfizer-BioNTech (BNT162)		Moderna (mRNA-1273)		AstraZeneca (AZD1222)		Janssen (Ad26.COV2.S)	
	Female (n=1613)	Male (n=576)	Female (n=1142)	Male (n=389)	Female (n=221)	Male (n=78)	Female (n=153)	Male (n=62)	Female (n=28)	Male (n=16)
Age (years), mean (SD)	49.32 (12.8)	52.71 (13.86)	48.6 (12.85)	52.53 (14.5)	44.34 (10.54)	45.27 (9.84)	59.61 (99.9)	61.32 (8.7)	49.5 (10.28)	46.07 (13.05)
BMI (kg/m^2), mean (SD)	26.88 (5.44)	26.27 (4.17)	26.95 (5.46)	26.46 (4.25)	25.76 (4.45)	25.85 (4.38)	27.75 (6.32)	26.55 (3.52)	26.37 (4.91)	24.93 (4.23)
Menstruation status, n										
Not menstruating	223	N/A[a]	169	N/A	27	N/A	17	N/A	4	N/A
Pregnant	2	N/A	2	N/A	0	N/A	0	N/A	0	N/A
Premenopausal	642	N/A	480	N/A	119	N/A	17	N/A	11	N/A
Perimenopausal	168	N/A	122	N/A	28	N/A	6	N/A	2	N/A
Postmenopausal	577	N/A	368	N/A	47	N/A	113	N/A	11	N/A

[a]N/A: not applicable.

Part 2: Menstrual Cycle Duration Around Vaccination

To investigate the differences in menstrual cycle duration following COVID-19 vaccination, phases of the menstrual cycle were determined based on fluctuations in physiological signals and participant-reported menses by means of a proprietary algorithm (Ava AG). Of note, the algorithm's effectiveness is not solely dependent on the precise reporting of the cycle's onset but rather on its ability to learn by aggregating and analyzing biophysical data points across women's repeated menstrual cycles (as captured during the 9-month COVID-RED study). Ava's fertility detection algorithm has been previously validated in clinical trials, with its accuracy on par with urine-based ovulation tests [16,31]. By means of this method, the following phases were detected: menstrual phase (ie, 5 days from the first day of menses), follicular phase (ie, from the first day post menses through 6 days before ovulation), fertile window (ie, 5 days before ovulation and the ovulation day), early luteal phase (ie, from the first day after ovulation through a week after ovulation), and late luteal phase (ie, from the eighth day after ovulation through the day before the onset of subsequent menses). As these algorithms' performance depends on the reliability of the physiological data, we included only those female participants with synchronized data for each day of each menstrual cycle and more than 1 available cycle (n=179 from 19 to 55 years of age; mean age 41.31, SD 8 years). For this analysis, a variable describing the duration of each cycle phase in the current cycle was created to serve as the outcome variable in regression models. As in part 1, for each of the 5 phases, we applied four models including: (1) only a random intercept; (2) age, BMI, daily alcohol use, contraceptive method, type of COVID-19 vaccine, existence of laboratory-confirmed SARS-CoV-2 infection prior to vaccination and sleep duration as fixed factors and number of days from vaccination as a fixed quadratic term; (3) random intercept in addition to the fixed factors from model 2; and (4) interaction of the quadratic term with existence of SARS-CoV-2 infection prior to vaccination and sleep duration in addition to all terms from model 3. In addition to information about COVID-19 vaccination and infection, only factors that have previously been shown to significantly impact menstrual cycle duration were included [32-34].

For all analyses, the reported *P* values were corrected by means of the false discovery rate [35] to maintain a family-wise α rate of .05 across the performed statistical tests (ie, the number of calculated *P* values in the 9 models examining the 4 physiological parameters and the 5 phases of the menstrual cycle as the outcome).

Ethical Considerations

The study was approved by the responsible ethics committee Medisch Ethische Toetsingscommissie Utrecht (SL/nb/21/500101), and electronic informed consent was provided by each participant. Participants were informed that their participation in the study is voluntary and that they could withdraw from the study at any time. No compensation was offered for participation. Participant data were pseudonymized immediately after collection to ensure confidentiality and privacy.

Results

Population characteristics

The 2189 included participants reported receiving 4 different types of COVID-19 vaccines (Table 1). The majority (1531/2189, 70%) received the vaccine from Pfizer-BioNTech (BNT162), followed by 14% (299/2189) reporting vaccination with Moderna (mRNA-1273), 10% (215/2189) AstraZeneca (AZD1222), and 2% (44/2189) Janssen (Ad26.COV2.S).

Part 1: Physiological Signals Around Vaccination

The trajectory of the 4 physiological parameters across 2189 participants showed short-term changes in the first days after vaccination and a rebound to baseline levels measured before vaccination thereafter (Figure 1). Accuracy metrics of the 4 models for each parameter are shown in Table S1 in Multimedia Appendix 1 and indicate the models including fixed factors and

XSL•FO
RenderX

a random intercept as most suitable (ie, model 3). The only exception was observed for heart rate variability with the model additionally including interactions between factors showing the best accuracy (ie, model 4). From this point onward, we will exclusively focus on the results obtained from the best fitting model described above. The short-term increase in breathing rate and heart rate was revealed as statistically significant. This increase was reflected by a negative effect of the quadratic term depicting days since vaccination, indicating that breathing rate and heart rate initially rise but then decrease again as time moves away from the vaccination event (Table 2). Nevertheless, the interaction between sex and days since vaccination remained nonsignificant (Table 2), suggesting no meaningful sex differences in the physiological response to COVID-19 vaccination. Similarly, the nonsignificant interaction with sleep duration (Table 2) indicates no impact of sleep duration on the physiological response to vaccination. Finally, the qualitative short-term decrease in heart rate variability following vaccination was not apparent in individuals who experienced a SARS-CoV-2 infection prior to vaccination (interaction effect in Table 2; Figure S1 in Multimedia Appendix 1).

Regardless of the vaccination event, female participants demonstrated higher wrist skin temperature and heart rate as well as a lower heart rate variability than male participants (Table 2). Furthermore, older participants showed a lower breathing rate as well as a higher heart rate variability as reflected in significant effects of age (Table 2). Participants with a higher BMI exhibited higher breathing rate and heart rate but lower skin temperature and heart rate variability (Table 2). Perimenopausal and menstruating female participants had a higher skin temperature, while menstruating female participants also had a higher breathing rate than nonmenstruating participants (Table 2). Finally, longer sleep duration was associated with lower breathing rate and heart rate as well as higher skin temperature and heart rate variability (Table 2). There were no significant effects of vaccine type (Table 2).

Figure 1. Trajectory of physiological signals around COVID-19 vaccination. The values of each parameter (with standard deviation in gray) were normalized according to each individual's mean baseline measurement (ie, during days before vaccination) and averaged across female (n=1613) and male participants separately (n=576).

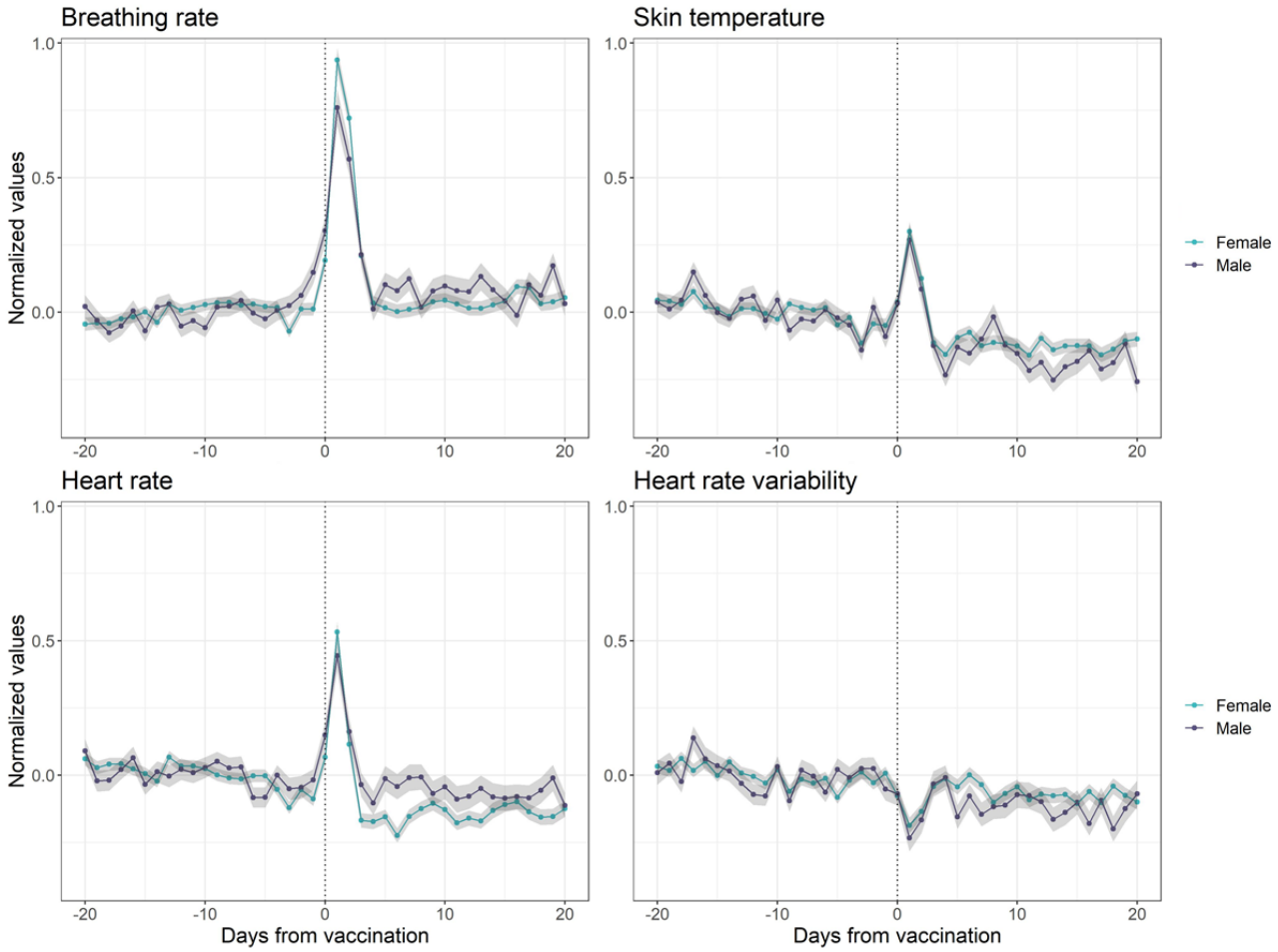

XSL•FO
RenderX

Table 2. Results from multilevel quadratic regression models including physiological signals from 2189 participants (1613 females) around COVID-19 vaccination[a].

	Breathing rate (breaths/min)	Skin temperature (°C)	Heart rate (beats/min)	Heart rate variability
Intercept	*13.41 (<.0001)*	*36.1 (<.0001)*	*43.8 (<.0001)*	*3 (<.0001)*
Main effects				
Days from vaccination	*–0.04 (<.0001)*	–0.004 (.36)	*–0.03 (.04)*	–0.001 (1)
Vaccine type				
AstraZeneca	Reference	Reference	Reference	Reference
Janssen	0.11 (1)	0.12 (.92)	–0.49 (1)	–0.11 (1)
Moderna	0.08 (1)	0.07 (.58)	0.54 (1)	0.07 (1)
Pfizer-BioNTech	0.39 (.19)	0.07 (.36)	0.43 (1)	0.04 (1)
Previous infection	0.56 (.73)	–0.003 (1)	0.02 (1)	0.11 (1)
Sex	0.3 (.19)	*0.18 (<.0001)*	*3.1 (<.0001)*	*–0.72 (<.0001)*
Age	*–0.02 (<.0001)*	–0.002 (.36)	0.03 (.11)	*0.007 (.03)*
BMI	*0.08 (<.0001)*	*–0.02 (<.0001)*	*0.32 (<.0001)*	*–0.04 (<.0001)*
Menstruation status				
Not menstruating	Reference	Reference	Reference	Reference
Perimenopausal	0.37 (.65)	*0.12 (.02)*	0.61 (1)	0.02 (1)
Menstruating	*0.44 (.02)*	*0.16 (<.0001)*	0.76 (1)	–0.05 (1)
Sleep duration	*–0.06 (<.0001)*	*0.04 (<.0001)*	*–0.46(<.0001)*	*0.01 (.004)*
Interactions				
Sex×Days from vaccination	N/A[b]	N/A	N/A	0.008 (.98)
Previous infection×Days from vaccination	N/A	N/A	N/A	*–0.07 (<.0001)*
Sleep duration×Days from vaccination	N/A	N/A	N/A	–0.004 (.87)

[a]Unstandardized β coefficients are presented with *P* values in parentheses and significant effects italicized (*P*<.05). Sex was coded such that positive coefficients represent greater values in females.

[b]N/A: not applicable.

Part 2: Menstrual Cycle Duration Around Vaccination

Our analysis of menstrual cycle duration revealed a significant positive effect of the quadratic term representing time since vaccination on menstrual phase duration. This finding suggests a decrease in menstrual phase duration around the time of vaccination. In contrast, we found a trend-level increase in duration of follicular phase following vaccination (Table 3). The mean within-participant difference between pre- and postvaccine duration of the menstrual cycle revealed that this change amounted to 0.8% for the menstrual phase and 1.4% for the follicular phase (Figure 2). We found an overall decrease in total cycle duration of 0.26 days following vaccination (Table S2 in Multimedia Appendix 1). Accuracy metrics of the 4 models for each cycle phase are shown in Table S3 in Multimedia Appendix 1 and indicate the models including fixed factors and a random intercept as most suitable (ie, model 3).

Regardless of the vaccination event, older females experienced a shorter follicular phase (Table 3). Furthermore, alcohol use was associated with a shorter fertile window (Table 3). There were no significant effects of vaccine type, contraceptive method, or BMI on menstrual cycle duration (Table 3).

XSL•FO
RenderX

Table 3. Results from multilevel quadratic regression models including 642 cycles from 179 female participants who received a COVID-19 vaccine[a].

	Menstrual phase duration	Follicular phase duration	Fertile window duration	Early luteal phase duration	Late luteal phase duration
Intercept	*4.81 (<.0001)*	*6.22 (.02)*	*5.89 (<.0001)*	*7.01 (<.0001)*	*4.04 (.002)*
Main effects					
Days from vaccination	*0.007 (.0003)*	–0.04 (.06)	0.005 (1)	0.009 (.31)	–0.008 (1)
Vaccine type					
AstraZeneca	Reference	Reference	Reference	Reference	Reference
Janssen	0.12 (1)	0.17 (1)	–0.65 (1)	0.5 (1)	–1.24 (1)
Moderna	0.22 (1)	0.69 (1)	–0.41 (1)	0.13 (1)	–0.67 (1)
Pfizer-BioNTech	0.21 (1)	0.25 (1)	–0.22 (1)	0.25 (1)	–0.75 (1)
Previous infection	0.4 (1)	3.16 (1)	–0.19 (1)	–0.04 (1)	1.01 (1)
Age	0.003 (1)	*–0.1 (.02)*	0.01 (.8)	–0.002 (1)	0.02 (1)
BMI	–0.006 (1)	0.11 (.38)	–0.009 (1)	–0.003 (1)	0.02 (1)
Contraceptive method					
None	Reference	Reference	Reference	Reference	Reference
Hormonal	0.41 (1)	0.07 (1)	0.16 (1)	–0.13 (1)	–0.63 (1)
Intrauterine device	0.13 (1)	–0.36 (1)	–0.26 (1)	–0.05 (1)	–0.23 (1)
Other	0.02 (1)	0.81 (1)	0.05 (1)	–0.1 (1)	–0.13 (1)
Alcohol use					
None	Reference	Reference	Reference	Reference	Reference
1-2 drinks	–0.002 (1)	0.005 (1)	*–0.05 (<.0001)*	0.002 (1)	–0.04 (.38)
3-4 drinks	–0.008 (1)	–0.02 (1)	*–0.07 (.02)*	0.02 (1)	–0.1 (.1)
5+ drinks	–0.001 (1)	0.08 (1)	0.01 (1)	–0.07 (.31)	–0.05 (1)
Sleep duration	–0.003 (1)	0.006 (1)	–0.004 (1)	0.004 (1)	–0.005 (1)

[a]Unstandardized β coefficients are presented with *P* values in parentheses and significant effects italicized (*P*<.05).

Figure 2. Mean duration of each cycle phase in days before and after vaccination across 179 female participants with 2 to 8 cycles per subject (mean number of cycles included in analysis 3.59, SD 1.36). The error bars depict the standard deviation.

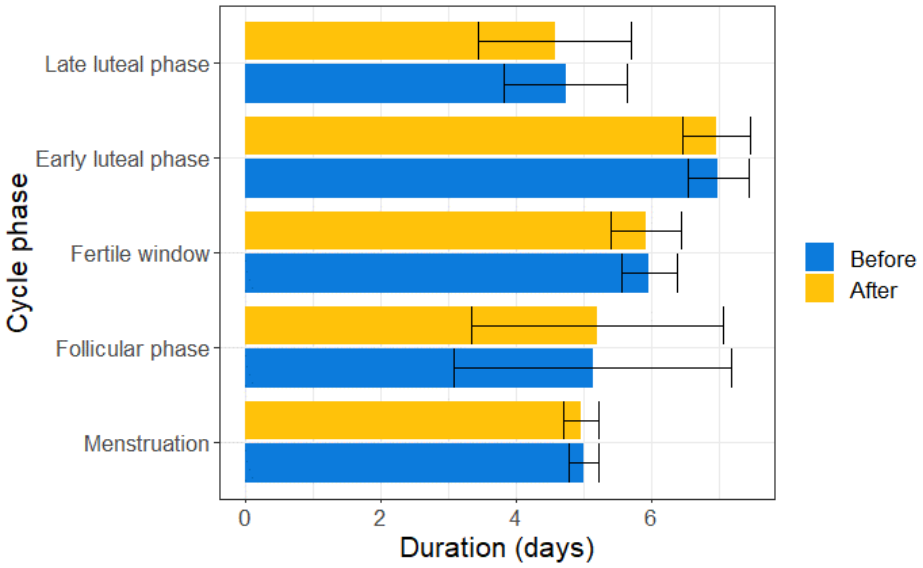

XSL•FO
RenderX

Discussion

Principal Findings

This study examined the effects of COVID-19 vaccination on several physiological signals measured continuously for up to 9 months in 2189 adult participants. We found short-term increases in breathing rate and heart rate following vaccination, which did not differ between the sexes. Furthermore, we observed a 0.8% decrease in menstrual phase duration after vaccination as compared to before. Taken together, our findings indicate that COVID-19 vaccines may cause physiological reactions in line with other routine vaccines [36,37] that typically last only a few days.

Comparison With Prior Work

The observed short-term physiological response parallels the duration of self-reported side effects from COVID-19 vaccination in the community [2] as well as previous reports of physiological response to COVID-19 vaccination [12-14]. During the immune reaction to any vaccination, inflammatory responses may be propagated to the central nervous system causing systemic side effects such as fever and headache [9]. The activated neuronal circuits then activate autonomic responses like peripheral vasoconstriction [9]. Such modulations may cause the observed physiological alterations following vaccination in our study. For example, peripheral vasoconstriction is associated with increases in heart rate [38], which is in turn positively associated with breathing rate [39], supporting our findings.

Additionally, decreased heart rate variability during inflammatory response to natural infection [40] as well as after influenza vaccination [41] has been observed in electrocardiographic measurements. Such observations are not surprising, given that heart rate variability reflects vagal and thereby autonomic nervous system activity [42]. Although, on average, we observed decreases in heart rate variability in the first days after vaccination, these changes did not reach statistical significance. This observation is in line with a previous study on wearable-measured physiological signals during the course of SARS-CoV-2 infection, which reported less pronounced changes in heart rate variability as compared to heart rate and breathing rate [10]. Similarly, while skin temperature did increase in response to vaccination, this increase was not significant in our study. This finding is consistent with the report by Gepner et al [13], showing a relative change of below 1% for skin temperature and above 5% for heart and respiratory rate following COVID-19 vaccination.

We observed no differences between the sexes in the physiological response to vaccination. For the Pfizer-BioNTech and Moderna vaccines, slightly lower (2.3%-2.7%) efficacy has been reported in female participants compared to male participants [43,44]. In contrast, the efficacy of non–COVID-19 routine vaccines in the older participants is greater in female participants than in male participants [45], suggesting an important relation with age [21]. Genetics constitute an additional factor for consideration, as highlighted by twin studies that show a heritability of 36% to 90% for immune response to vaccines [46]. Importantly, a larger increase in breathing rate has recently been measured in female participants as compared to male participants after COVID-19 vaccination [12]. Although our findings appear to suggest a larger initial physiological response to vaccination in female participants for breathing rate, heart rate, and skin temperature (Figure 1), this difference is small and did not reach statistical significance. As the lack of statistical significance does not necessarily imply a lack of clinical significance [47], we recommend that future studies specifically examine the clinical relevance of sex differences in the context of vaccine reactogenicity. Interactions among genetic factors, age, and sex hormones render the biological mechanisms behind sex differences complex, potentially masking the sex effects in our sample. Nonetheless, such effects are important to consider and report in studies to ensure that the needs of male and female subgroups have been adequately addressed.

In addition to sex differences, the COVID-RED data set allowed us to examine changes in duration of different menstrual cycle phases following vaccination. We found that the follicular phase slightly increased in duration, while the menstrual phase was shorter after vaccination. Such dynamics could obscure vaccination effects on the overall duration of the menstrual cycle. Prior work found that the menstrual cycle was less than 1 day longer immediately after as compared to before vaccination [24]. The overall decrease in menstrual cycle duration in our study contradicts this previous finding. The discrepancy may be due to methodological differences, as Edelman et al [24] examined self-reported durations, while we additionally relied on objective fluctuations in physiological signals to determine cycle phases using machine learning. Despite this difference, both investigations showed a small change of less than 1 day in menstrual cycle duration following vaccination. Similarly, irregularities in the menstrual cycle have been reported after non–COVID-19 vaccines including missed, late, and early menstrual bleeding [48,49]. These findings hint at large interindividual differences among menstruating women in response to vaccination, which are also reflected in the wide variance observed in the duration of specific menstrual cycle phases depicted in Figure 2. The mechanisms behind such observations may be linked to alterations in mast cell activation, which is implicated in both inflammatory processes [50] as well as the initiation of menstruation [51]. Further factors known to influence these interactions also likely contribute to interindividual variability and inconsistency between findings. For instance, age at onset of menarche and adiposity have been shown to affect the relationship between inflammatory biomarkers and ovarian function [52,53]. Taken together, we cannot exclude the possibility that vaccination may cause small temporal shifts of the fertile window in some female participants with consequences for their chances to conceive. Finally, as our study design did not allow for an investigation of long-term effects of COVID-19 vaccination on menstrual cycle duration, we cannot draw any conclusions about the long-term regularity of menstrual cycles postvaccination. It is, however, important to note that previous research has reported links between menstrual cycle irregularities and increased risk of cardiovascular disease [54] as well as diabetes [55]. This underscores the necessity for future research to examine vaccine-induced changes in menstrual cycle characteristics over

XSL•FO
RenderX

longer periods of time in order to understand their potential clinical implications.

Across analyses, the 4 examined types of COVID-19 vaccines did not differ with regard to the observed physiological responses. The COVID-19 vaccines by Pfizer-BioNTech and Moderna use the mRNA isolated from the SARS-CoV-2 spike protein to induce synthesis of this protein in host cells upon injection and initiate an immune response [56]. The AstraZeneca vaccine uses a modified chimpanzee DNA adenovirus that does not cause an immune response to the virus but only to the SARS-CoV-2 spike protein produced in host cells after injection [5,57]. Finally, the vaccine developed by Janssen uses an inactivated adenovirus vector that codes for the viral spike protein [57] and thus induces a similar host response as the AstraZeneca vaccine. Regardless of their type, all COVID-19 vaccines ultimately activate the same immune pathways in the host and were similarly reflected in physiological responses measured in this study.

Limitations

Several limitations are important to acknowledge in the context of this study. First, the self-reported nature of the data collected through the Ava COVID-RED app and the biweekly surveys may introduce subjective bias into our findings as they rely on participants' recall. To mitigate this potential risk, we strategically designed the Ava COVID-RED app to ask for daily input of self-reported data. Additionally, participants could only edit their responses for up to a week. Furthermore, we collected information in a redundant manner through multiple sources whenever possible. Therefore, we believe that the key measures of our study (ie, information about menstrual cycle start and COVID-19 vaccination) are reliable. Further, we recognize that there may be additional factors arising from pandemic-induced changes in lifestyle and particularly stress levels that could affect physiological signals. However, the observed effects are specific to the immediate postvaccination period and rely on relative changes in physiological signals from each individual's baseline (measured during the pandemic), while the trajectory of physiological signals remains consistent with baseline levels prior to vaccination. Therefore, we argue that these contextual factors as well as potential acute stress or anxiety anticipating vaccination are unlikely to confound our findings significantly. Moreover, while we acknowledge variability in wearable device accuracy [58,59], the large sample size and the temporal specificity of the observed effects mitigate this concern. Finally, this study's cohort was recruited exclusively from the Netherlands, potentially limiting the generalizability of our findings to other countries and thus calling for replication across diverse populations.

This study examined the physiological response to the first dose of COVID-19 vaccine. Most participants received their first dose during the last trimester of the trial, yielding only a small sample with sufficient data to investigate the trajectory of physiological signals after the second dose. We thus refrained from conducting this additional analysis due to lack of statistical power and recommend that future studies investigate potential impact of further doses as well as long-term effects thereof. It is imperative that such studies continue to consider potential sex-specific responses to vaccination, particularly concerning menstrual cycle changes, and assess the clinical relevance of the minor and transient physiological changes observed in our study. Understanding these differences is vital for tailoring public health policies and recommendations for vaccine development to better serve all segments of the society. Finally, the continuous, real-time data provided by wearable technology offer an unprecedented opportunity to advance such efforts, facilitating a move toward more personalized health care.

Conclusions

This study constitutes the first comprehensive examination of continuous measurements of multiple physiological signals as well as objectively assessed menstrual cycle duration around a vaccination event. We rely on data collected by a wearable medical device in a large-scale clinical trial including contextual information such as the type of COVID-19 vaccine received to ensure higher validity results and find short-term changes in breathing rate and heart rate following vaccination. The prompt rebound to baseline levels measured before vaccination suggests that the COVID-19 vaccines are not associated with any long-term effects on heart rate, breathing rate, and heart rate variability or skin temperature. Similarly, our investigation of changes in the menstrual cycle duration showed minimal deviations following vaccination for several phases of the cycle. Based on these observations, we recommend that female participants relying on the regularity of their menstrual cycle both to prevent conception or conceive be particularly cautious during their menstrual cycles immediately following vaccination. Future investigations should consider whether this phenomenon extends to non–COVID-19 vaccines. We emphasize the importance of ongoing monitoring and further research to comprehensively assess the long-term impacts of these vaccines on physiological parameters and the menstrual cycle. Such extended research is vital to address remaining uncertainties and reinforce confidence in vaccine safety through robust, longitudinal data analysis, leveraging the demonstrated potential of wearable technology. Taken together, our work provides valuable insights into continuous fluctuations of our body's responses to vaccination and may help refute misinformation about vaccines contributing to fertility or biophysical issues.

Acknowledgments

We thank the more than 17,000 participants who enrolled in this extensive study. Additionally, we thank the entire team at Ava AG and Julius Clinical as well as the consortium partners Roche Diagnostics Nederland BV (Almere, the Netherlands), Labormedizinisches Zentrum Dr. Risch (Vaduz, Liechtenstein), Sanquin (Amsterdam, the Netherlands), Vive (Aarhus, Denmark), Takeda Pharmaceuticals USA Inc (Cambridge, MA, USA), and King's College London (London, UK) for their contributions to the COVID-RED project. The COVID-RED project has received funding from the Innovative Medicines Initiative [60] 2 Joint

XSL•FO
RenderX

Undertaking (grant agreement 101005177). This Joint Undertaking receives support from the European Union's Horizon 2020 [61] research and innovation programme and EFPIA [62].

The contributors associated with COVID-19 Remote Early Detection (COVID-RED) consortium are as follows: Maureen Cronin, Vladimir Kovacevic, Andjela Markovic, Maja Rudinac, from Ava AG, Switzerland; Kirsten Grossmann, Lorenz Risch, Martin Risch, Ornella Weideli, from Dr. Risch, Liechtenstein; Billy Franks, Brianna Goodale, Ellen Dutman, Eric Houtman, Glenn Van Wigcheren, Hans Van Dijk, Ishak Elmouhajir, Jon Bouwman, Lotte Smets, Marcel van Willigen, Niki de Vink, Timo Brakenhoff, Titia Leurink, Wendy van Scherpenzeel, Wout Aarts, Pieter van der Meer, Myrna Verhulst, Paul Klaver, Tessa Heikamp, Kai Hage, José Broersen, Jungyeon Choi, Maartje Hoffmann, Marjolein Jansen, Jeffrey Burggraaff, from Julius Clinical, The Netherlands; Alison Kuchta, Christian Simon, Santiago Montes, from Roche, The Netherlands; Theo Rispens, Maurice Steenhuis, Floris Loeff, Sofie Keijzer, Jim Keijser, Olvi Christianawati, Aren Boogaard, Nadine Commandeur, from Sanquin, The Netherlands; Ariel Dowling and Steve Emby, from Takeda, USA; Charisma Hehakaya, Daniel Oberski, George Downward, Duco Veen, Marianna Mitratza, Hans Reitsma, Janneke van de Wijgert, Nathalie Vigot, Patricia Bruijning, Pieter Stolk, Diederick Grobbee, Gulseren Yalvac, from University Medical Center Utrecht, The Netherlands; Johann Fevrier, Amos Folarin, Pablo Fernandez Medina, Richard Dobson, Spiros Denaxas, from University College London, UK; Eskild Fredslund, Serkan Korkmaz, and Jesper Strømstad, from VIVE, Denmark.

Data Availability

The data sets generated and analyzed during this study are available in the DataverseNL repository [63].

Authors' Contributions

AM contributed to data collection and curation, project administration, safety reporting, methodology, statistical planning, formal analysis, visualization, and manuscript preparation and revision. VK contributed to data curation, methodology, formal analysis, validation, and manuscript revision. TBB contributed to data collection and curation, project administration, methodology, statistical planning, validation, and manuscript revision. DV contributed to data curation, methodology, statistical planning, validation, and manuscript revision. PK contributed to data collection, project administration, and manuscript revision. MM contributed to data curation and manuscript revision. GSD contributed to project administration, and manuscript revision. DEG contributed to conceptualization, funding acquisition, project administration, supervision, and manuscript revision. MC contributed to conceptualization, funding acquisition, project administration, supervision, and manuscript revision. BMG contributed to conceptualization, data collection, project administration, safety reporting, methodology, supervision, validation, visualization, and manuscript preparation and revision. All authors critically reviewed and approved the final version of this manuscript and had final responsibility for the decision to submit for publication.

Conflicts of Interest

AM, VK, MC, and BMG are past employees of Ava AG. TBB, DV, PK, and BMG are current or past employees of Julius Clinical Research. MM is a current employee of P95 CVBA. The authors have no financial or personal conflicts of interest to declare.

Multimedia Appendix 1
Accuracy metrics of statistical models; numbers of reported vaccinations for analysis of menstrual cycle duration changes; and interaction between the number of days from vaccination and previous infection in influencing the trajectory of heart rate variability. [PDF File (Adobe PDF File), 196 KB - jmir_v26i1e51120_app1.pdf]

References

1. Vaccine BNT162b2: conditions of authorisation under Regulation 174. Medicines and Healthcare Products Regulatory Agency. 2020. URL: https://www.gov.uk/government/publications/regulatory-approval-of-pfizer-biontech-vaccine-for-covid-19/conditions-of-authorisation-for-pfizerbiontech-covid-19-vaccine [accessed 2022-10-16]
2. Menni C, Klaser K, May A, Polidori L, Capdevila J, Louca P, et al. Vaccine side-effects and SARS-CoV-2 infection after vaccination in users of the COVID symptom study app in the UK: a prospective observational study. Lancet Infect Dis 2021;21(7):939-949 [FREE Full text] [doi: 10.1016/S1473-3099(21)00224-3] [Medline: 33930320]
3. Cavanaugh AM, Spicer KB, Thoroughman D, Glick C, Winter K. Reduced risk of reinfection with SARS-CoV-2 after COVID-19 vaccination—Kentucky, May-June 2021. MMWR Morb Mortal Wkly Rep 2021;70(32):1081-1083 [FREE Full text] [doi: 10.15585/mmwr.mm7032e1] [Medline: 34383732]
4. Tenforde MW, Self WH, Adams K, Gaglani M, Ginde AA, McNeal T, InfluenzaOther Viruses in the Acutely Ill (IVY) Network. Association between mRNA vaccination and COVID-19 hospitalization and disease severity. JAMA 2021;326(20):2043-2054 [FREE Full text] [doi: 10.1001/jama.2021.19499] [Medline: 34734975]
5. Mascellino MT, Di Timoteo F, De Angelis M, Oliva A. Overview of the main anti-SARS-CoV-2 vaccines: mechanism of action, efficacy and safety. Infect Drug Resist 2021;14:3459-3476 [FREE Full text] [doi: 10.2147/IDR.S315727] [Medline: 34511939]

XSL•FO
RenderX

6. Shimabukuro T, Nair N. Allergic reactions including anaphylaxis after receipt of the first dose of Pfizer-BioNTech COVID-19 vaccine. JAMA 2021;325(8):780-781 [FREE Full text] [doi: 10.1001/jama.2021.0600] [Medline: 33475702]
7. Suvvari TK, Rajesh E, D Silva RG, Corriero AC, Kutikuppala LVS. SARS-CoV-2 vaccine-induced prothrombotic immune thrombocytopenia: promoting awareness to improve patient-doctor trust. J Med Virol 2021;93(10):5721-5723 [FREE Full text] [doi: 10.1002/jmv.27176] [Medline: 34232526]
8. Krammer F, Srivastava K, Alshammary H, Amoako AA, Awawda MH, Beach KF, et al. Antibody responses in seropositive persons after a single dose of SARS-CoV-2 mRNA vaccine. N Engl J Med 2021;384(14):1372-1374 [FREE Full text] [doi: 10.1056/NEJMc2101667] [Medline: 33691060]
9. Hervé C, Laupèze B, Del Giudice G, Didierlaurent AM, Tavares Da Silva F. The how's and what's of vaccine reactogenicity. NPJ Vaccines 2019;4:39 [FREE Full text] [doi: 10.1038/s41541-019-0132-6] [Medline: 31583123]
10. Risch M, Grossmann K, Aeschbacher S, Weideli OC, Kovac M, Pereira F, COVID-19 remote early detection (COVID-RED) consortium. Investigation of the use of a sensor bracelet for the presymptomatic detection of changes in physiological parameters related to COVID-19: an interim analysis of a prospective cohort study (COVI-GAPP). BMJ Open 2022;12(6):e058274 [FREE Full text] [doi: 10.1136/bmjopen-2021-058274] [Medline: 35728900]
11. Mitratza M, Goodale BM, Shagadatova A, Kovacevic V, van de Wijgert J, Brakenhoff TB, et al. The performance of wearable sensors in the detection of SARS-CoV-2 infection: a systematic review. Lancet Digit Health 2022;4(5):e370-e383 [FREE Full text] [doi: 10.1016/S2589-7500(22)00019-X] [Medline: 35461692]
12. Quer G, Gadaleta M, Radin JM, Andersen KG, Baca-Motes K, Ramos E, et al. Inter-individual variation in objective measure of reactogenicity following COVID-19 vaccination via smartwatches and fitness bands. NPJ Digit Med 2022;5(1):49 [FREE Full text] [doi: 10.1038/s41746-022-00591-z] [Medline: 35440684]
13. Gepner Y, Mofaz M, Oved S, Yechezkel M, Constantini K, Goldstein N, et al. Utilizing wearable sensors for continuous and highly-sensitive monitoring of reactions to the BNT162b2 mRNA COVID-19 vaccine. Commun Med (Lond) 2022;2:27 [FREE Full text] [doi: 10.1038/s43856-022-00090-y] [Medline: 35603274]
14. Guan G, Mofaz M, Qian G, Patalon T, Shmueli E, Yamin D, et al. Higher sensitivity monitoring of reactions to COVID-19 vaccination using smartwatches. NPJ Digit Med 2022;5(1):140 [FREE Full text] [doi: 10.1038/s41746-022-00683-w] [Medline: 36085312]
15. Mason AE, Kasl P, Hartogensis W, Natale JL, Dilchert S, Dasgupta S, et al. Metrics from wearable devices as candidate predictors of antibody response following vaccination against COVID-19: data from the second TemPredict study. Vaccines (Basel) 2022;10(2):264 [FREE Full text] [doi: 10.3390/vaccines10020264] [Medline: 35214723]
16. Goodale BM, Shilaih M, Falco L, Dammeier F, Hamvas G, Leeners B. Wearable sensors reveal menses-driven changes in physiology and enable prediction of the fertile window: observational study. J Med Internet Res 2019;21(4):e13404 [FREE Full text] [doi: 10.2196/13404] [Medline: 30998226]
17. Shilaih M, Clerck VD, Falco L, Kübler F, Leeners B. Pulse rate measurement during sleep using wearable sensors, and its correlation with the menstrual cycle phases, a prospective observational study. Sci Rep 2017;7(1):1294 [FREE Full text] [doi: 10.1038/s41598-017-01433-9] [Medline: 28465583]
18. Shilaih M, Goodale BM, Falco L, Kübler F, De Clerck V, Leeners B. Modern fertility awareness methods: wrist wearables capture the changes in temperature associated with the menstrual cycle. Biosci Rep 2018;38(6):BSR20171279 [FREE Full text] [doi: 10.1042/BSR20171279] [Medline: 29175999]
19. Regulations, guidance, and reports related to women's health. U.S. Food and Drug Administration. 2020. URL: https://www.fda.gov/science-research/womens-health-research/regulations-guidance-and-reports-related-womens-health [accessed 2022-10-20]
20. Flanagan KL, Fink AL, Plebanski M, Klein SL. Sex and gender differences in the outcomes of vaccination over the life course. Annu Rev Cell Dev Biol 2017;33:577-599. [doi: 10.1146/annurev-cellbio-100616-060718] [Medline: 28992436]
21. Jensen A, Stromme M, Moyassari S, Chadha AS, Tartaglia MC, Szoeke C, et al. COVID-19 vaccines: considering sex differences in efficacy and safety. Contemp Clin Trials 2022;115:106700 [FREE Full text] [doi: 10.1016/j.cct.2022.106700] [Medline: 35149232]
22. Villarreal A. 'No data' linking Covid vaccines to menstrual changes, US experts say? The Guardian 2021;256:258.
23. Cuda A. Is there a link between COVID vaccine and 'funky' menstrual periods? Experts say it's too soon to know. Conn. Post 2021.
24. Edelman A, Boniface ER, Benhar E, Han L, Matteson KA, Favaro C, et al. Association between menstrual cycle length and coronavirus disease 2019 (COVID-19) vaccination: a U.S. cohort. Obstet Gynecol 2022;139(4):481-489 [FREE Full text] [doi: 10.1097/AOG.0000000000004695] [Medline: 34991109]
25. Baena-García L, Aparicio VA, Molina-López A, Aranda P, Cámara-Roca L, Ocón-Hernández O. Premenstrual and menstrual changes reported after COVID-19 vaccination: the EVA project. Womens Health (Lond) 2022;18:17455057221112237 [FREE Full text] [doi: 10.1177/17455057221112237] [Medline: 35833668]
26. Lee KMN, Junkins EJ, Luo C, Fatima UA, Cox ML, Clancy KBH. Investigating trends in those who experience menstrual bleeding changes after SARS-CoV-2 vaccination. Sci Adv 2022;8(28):eabm7201 [FREE Full text] [doi: 10.1126/sciadv.abm7201] [Medline: 35857495]

XSL•FO
RenderX

27. Wesselink AK, Hatch EE, Rothman KJ, Wang TR, Willis MD, Yland J, et al. A prospective cohort study of COVID-19 vaccination, SARS-CoV-2 infection, and fertility. Am J Epidemiol 2022;191(8):1383-1395 [FREE Full text] [doi: 10.1093/aje/kwac011] [Medline: 35051292]
28. Gonzalez DC, Nassau DE, Khodamoradi K, Ibrahim E, Blachman-Braun R, Ory J, et al. Sperm parameters before and after COVID-19 mRNA vaccination. JAMA 2021;326(3):273-274 [FREE Full text] [doi: 10.1001/jama.2021.9976] [Medline: 34137808]
29. Brakenhoff TB, Franks B, Goodale BM, van de Wijgert J, Montes S, Veen D, COVID-RED Consortium. A prospective, randomized, single-blinded, crossover trial to investigate the effect of a wearable device in addition to a daily symptom diary for the remote early detection of SARS-CoV-2 infections (COVID-RED): a structured summary of a study protocol for a randomized controlled trial. Trials 2021;22(1):412 [FREE Full text] [doi: 10.1186/s13063-021-05241-5] [Medline: 34158099]
30. Feingold A. Time-varying effect sizes for quadratic growth models in multilevel and latent growth modeling. Struct Equ Modeling 2019;26(3):418-429 [FREE Full text] [doi: 10.1080/10705511.2018.1547110] [Medline: 31579365]
31. Niggli A, Rothenbühler M, Sachs M, Leeners B. Can wrist-worn medical devices correctly identify ovulation? Sensors (Basel) 2023;23(24):9730 [FREE Full text] [doi: 10.3390/s23249730] [Medline: 38139575]
32. Stubblefield PG. Menstrual impact of contraception. Am J Obstet Gynecol 1994;170(5 Pt 2):1513-1522. [doi: 10.1016/s0002-9378(94)05013-1] [Medline: 8178900]
33. Liu Y, Gold EB, Lasley BL, Johnson WO. Factors affecting menstrual cycle characteristics. Am J Epidemiol 2004;160(2):131-140. [doi: 10.1093/aje/kwh188] [Medline: 15234934]
34. Song S, Choi H, Pang Y, Kim O, Park H. Factors associated with regularity and length of menstrual cycle: korea nurses' health study. BMC Womens Health 2022;22(1):361 [FREE Full text] [doi: 10.1186/s12905-022-01947-z] [Medline: 36050702]
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical powerful approach to multiple testing? J R Stat Soc Ser B Methodol 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
36. Shukla VV, Shah RC. Vaccinations in primary care. Indian J Pediatr 2018;85(12):1118-1127. [doi: 10.1007/s12098-017-2555-2] [Medline: 29306993]
37. Gidengil C, Goetz MB, Newberry S, Maglione M, Hall O, Larkin J, et al. Safety of vaccines used for routine immunization in the United States: an updated systematic review and meta-analysis. Vaccine 2021;39(28):3696-3716. [doi: 10.1016/j.vaccine.2021.03.079] [Medline: 34049735]
38. Toska K. Handgrip contraction induces a linear increase in arterial pressure by peripheral vasoconstriction, increased heart rate and a decrease in stroke volume. Acta Physiol (Oxf) 2010;200(3):211-221. [doi: 10.1111/j.1748-1716.2010.02144.x] [Medline: 20456282]
39. Gąsior JS, Sacha J, Jeleń PJ, Zieliński J, Przybylski J. Heart rate and respiratory rate influence on heart rate variability repeatability: effects of the correction for the prevailing heart rate. Front Physiol 2016;7:356 [FREE Full text] [doi: 10.3389/fphys.2016.00356] [Medline: 27588006]
40. von Känel R, Nelesen RA, Mills PJ, Ziegler MG, Dimsdale JE. Relationship between heart rate variability, interleukin-6, and soluble tissue factor in healthy subjects. Brain Behav Immun 2008;22(4):461-468 [FREE Full text] [doi: 10.1016/j.bbi.2007.09.009] [Medline: 17977694]
41. Perring S, Jones E. Assessment of changes in cardiac autonomic tone resulting from inflammatory response to the influenza vaccination. Clin Physiol Funct Imaging 2012;32(6):437-444. [doi: 10.1111/j.1475-097X.2012.01147.x] [Medline: 23031064]
42. Williams DP, Koenig J, Carnevali L, Sgoifo A, Jarczok MN, Sternberg EM, et al. Heart rate variability and inflammation: a meta-analysis of human studies. Brain Behav Immun 2019;80:219-226. [doi: 10.1016/j.bbi.2019.03.009] [Medline: 30872091]
43. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, COVE Study Group. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. N Engl J Med 2021;384(5):403-416 [FREE Full text] [doi: 10.1056/NEJMoa2035389] [Medline: 33378609]
44. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, C4591001 Clinical Trial Group. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. N Engl J Med 2020;383(27):2603-2615 [FREE Full text] [doi: 10.1056/NEJMoa2034577] [Medline: 33301246]
45. Fink AL, Klein SL. Sex and gender impact immune responses to vaccines among the elderly. Physiology (Bethesda) 2015;30(6):408-416 [FREE Full text] [doi: 10.1152/physiol.00035.2015] [Medline: 26525340]
46. Zimmermann P, Curtis N. Factors that influence the immune response to vaccination. Clin Microbiol Rev 2019;32(2):e00084 [FREE Full text] [doi: 10.1128/CMR.00084-18] [Medline: 30867162]
47. Armijo-Olivo S, Warren S, Fuentes J, Magee DJ. Clinical relevance vs. statistical significance: using neck outcomes in patients with temporomandibular disorders as an example. Man Ther 2011;16(6):563-572. [doi: 10.1016/j.math.2011.05.006] [Medline: 21658987]
48. Suzuki S, Hosono A. No association between HPV vaccine and reported post-vaccination symptoms in Japanese young women: results of the Nagoya study. Papillomavirus Res 2018;5:96-103 [FREE Full text] [doi: 10.1016/j.pvr.2018.02.002] [Medline: 29481964]

XSL•FO
RenderX

49. Shingu T, Uchida T, Nishi M, Hayashida K, Kashiwagi S, Hayashi J, et al. Menstrual abnormalities after hepatitis B vaccine. Kurume Med J 1982;29(3):123-125. [doi: 10.2739/kurumemedj.29.123]
50. Caughey GH. Mast cell proteases as protective and inflammatory mediators. Adv Exp Med Biol 2011;716:212-234. [doi: 10.1007/978-1-4419-9533-9_12] [Medline: 21713659]
51. Jeziorska M, Salamonsen LA, Woolley DE. Mast cell and eosinophil distribution and activation in human endometrium throughout the menstrual cycle. Biol Reprod 1995;53(2):312-320. [doi: 10.1095/biolreprod53.2.312] [Medline: 7492683]
52. Clancy KB, Baerwald AR, Pierson RA. Cycle-phase dependent associations between CRP, leptin, and reproductive hormones in an urban, Canadian sample. Am J Phys Anthropol 2016;160(3):389-396. [doi: 10.1002/ajpa.22976] [Medline: 27009682]
53. Clancy KB, Klein LD, Ziomkiewicz A, Nenko I, Jasienska G, Bribiescas RG. Relationships between biomarkers of inflammation, ovarian steroids, and age at menarche in a rural polish sample. Am J Hum Biol 2013;25(3):389-398. [doi: 10.1002/ajhb.22386] [Medline: 23606228]
54. Solomon CG, Hu FB, Dunaif A, Rich-Edwards JE, Stampfer MJ, Willett WC, et al. Menstrual cycle irregularity and risk for future cardiovascular disease. J Clin Endocrinol Metab 2002;87(5):2013-2017. [doi: 10.1210/jcem.87.5.8471] [Medline: 11994334]
55. Rostami Dovom M, Ramezani Tehrani F, Djalalinia S, Cheraghi L, Behboudi Gandavani S, Azizi F. Menstrual cycle irregularity and metabolic disorders: a population-based prospective study. PLoS One 2016;11(12):e0168402 [FREE Full text] [doi: 10.1371/journal.pone.0168402] [Medline: 27992506]
56. Zhou X, Jiang X, Qu M, Aninwene GE, Jucaud V, Moon JJ, et al. Engineering antiviral vaccines. ACS Nano 2020;14(10):12370-12389 [FREE Full text] [doi: 10.1021/acsnano.0c06109] [Medline: 33001626]
57. Mendonça SA, Lorincz R, Boucher P, Curiel DT. Adenoviral vector vaccine platforms in the SARS-CoV-2 pandemic. NPJ Vaccines 2021;6(1):97 [FREE Full text] [doi: 10.1038/s41541-021-00356-x] [Medline: 34354082]
58. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2020;3:18 [FREE Full text] [doi: 10.1038/s41746-020-0226-6] [Medline: 32047863]
59. Poli A, Cosoli G, Scalise L, Spinsante S. Impact of wearable measurement properties and data quality on ADLs classification accuracy. IEEE Sensors J 2021;21(13):14221-14231. [doi: 10.1109/jsen.2020.3009368]
60. Innovative Medicines Initiative. URL: https://www.imi.europa.eu [accessed 2024-06-24]
61. Horizon 2020. European Commission. URL: https://ec.europa.eu/programmes/horizon2020/ [accessed 2024-06-24]
62. EFPIA. URL: https://www.efpia.eu/ [accessed 2024-06-24]
63. Remote Early Detection of SARS-CoV-2 infections (COVID-RED). Data Archiving and Networked Services (DANS). URL: https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/FW9PO7 [accessed 2024-07-02]

Abbreviations

COVID-RED: COVID-19 Remote Early Detection

Edited by YH Lin; submitted 21.07.23; peer-reviewed by E van der Velde, S Mitra; comments to author 21.02.24; revised version received 05.04.24; accepted 30.04.24; published 31.07.24.

Please cite as:
Markovic A, Kovacevic V, Brakenhoff TB, Veen D, Klaver P, Mitratza M, Downward GS, Grobbee DE, Cronin M, Goodale BM, COVID-19 Remote Early Detection (COVID-RED) consortium
Physiological Response to the COVID-19 Vaccine: Insights From a Prospective, Randomized, Single-Blinded, Crossover Trial
J Med Internet Res 2024;26:e51120
URL: https://www.jmir.org/2024/1/e51120
doi:10.2196/51120
PMID:

©Andjela Markovic, Vladimir Kovacevic, Timo B Brakenhoff, Duco Veen, Paul Klaver, Marianna Mitratza, George S Downward, Diederick E Grobbee, Maureen Cronin, Brianna M Goodale, COVID-19 Remote Early Detection (COVID-RED) consortium. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Linking Opinions Shared on Social Media About COVID-19 Public Health Measures to Adherence: Repeated Cross-Sectional Surveys of Twitter Use in Canada

José Denis-Robichaud[*], DVM, MSc, PhD, DipACVPM; Erin E Rees[1,2,3*], BSc, MSc, PhD; Patrick Daley[4], MSc; Christina Zarowsky[5,6], BSc, MA, MPH, MD, PhD; Assane Diouf[2,7], MD, PhD; Bouchra R Nasri[5,6,8,9], MSc, PhD; Simon de Montigny[1,5], BIng, MSc, PhD; Hélène Carabin[1,3,5,6], DVM, MSc, PhD

[1]Groupe de recherche en e´pide´miologie des zoonoses et sante´ publique, Université de Montréal, Saint-Hyacinthe, QC, Canada
[2]Public Health Risk Sciences Division, National Microbiology Laboratory, Public Health Agency of Canada, Saint-Hyacinthe, QC, Canada
[3]Faculté de médecine vétérinaire, Université de Montréal, Saint-Hyacinthe, QC, Canada
[4]Policy Research Group, Canadian Heritage, Gatineau, QC, Canada
[5]School of Public Health, Université de Montréal, Montreal, QC, Canada
[6]Centre de Recherche en Santé Publique, Université de Montréal, Montreal, QC, Canada
[7]Department of Infectious and Tropical Diseases, Cheikh Anta Diop University, Dakar, Senegal
[8]Centre de recherches mathématiques, Université de Montréal, Montreal, QC, Canada
[9]The Data Informatics Center of Epidemiology, PathCheck Foundation, Cambridge, MA, United States
[*]these authors contributed equally

Corresponding Author:
Hélène Carabin, DVM, MSc, PhD
Faculté de médecine vétérinaire
Université de Montréal
3190, rue Sicotte
Saint-Hyacinthe, QC, J2S 2M2
Canada
Phone: 1 514 343 6111 ext 8569
Email: helene.carabin@umontreal.ca

Abstract

Background: The effectiveness of public health measures (PHMs) depends on population adherence. Social media were suggested as a tool to assess adherence, but representativeness and accuracy issues have been raised.

Objective: The objectives of this repeated cross-sectional study were to compare self-reported PHM adherence and sociodemographic characteristics between people who used Twitter (subsequently rebranded X) and people who did not use Twitter.

Methods: Repeated Canada-wide web-based surveys were conducted every 14 days from September 2020 to March 2022. Weighted proportions were calculated for descriptive variables. Using Bayesian logistic regression models, we investigated associations between Twitter use, as well as opinions in tweets, and self-reported adherence with mask wearing and vaccination.

Results: Data from 40,230 respondents were analyzed. As self-reported, Twitter was used by 20.6% (95% CI 20.1%-21.2%) of Canadians, of whom 29.9% (95% CI 28.6%-31.3%) tweeted about COVID-19. The sociodemographic characteristics differed across categories of Twitter use and opinions. Overall, 11% (95% CI 10.6%-11.3%) of Canadians reported poor adherence to mask-wearing, and 10.8% (95% CI 10.4%-11.2%) to vaccination. Twitter users who tweeted about COVID-19 reported poorer adherence to mask wearing than nonusers, which was modified by the age of the respondents and their geographical region (odds ratio [OR] 0.79, 95% Bayesian credibility interval [BCI] 0.18-1.69 to OR 4.83, 95% BCI 3.13-6.86). The odds of poor adherence to vaccination of Twitter users who tweeted about COVID-19 were greater than those of nonusers (OR 1.76, 95% BCI 1.48-2.07). English- and French-speaking Twitter users who tweeted critically of PHMs were more likely (OR 4.07, 95% BCI 3.38-4.80 and OR 7.31, 95% BCI 4.26-11.03, respectively) to report poor adherence to mask wearing than non–Twitter users, and those who tweeted in support were less likely (OR 0.47, 95% BCI 0.31-0.64 and OR 0.96, 95% BCI 0.18-2.33, respectively) to report poor adherence to mask wearing than non–Twitter users. The OR of poor adherence to vaccination for those tweeting critically about

XSL•FO
RenderX

PHMs and for those tweeting in support of PHMs were 4.10 (95% BCI 3.40-4.85) and 0.20 (95% BCI 0.10-0.32), respectively, compared to non–Twitter users.

Conclusions: Opinions shared on Twitter can be useful to public health authorities, as they are associated with adherence to PHMs. However, the sociodemographics of social media users do not represent the general population, calling for caution when using tweets to assess general population-level behaviors.

(J Med Internet Res 2024;26:e51325) doi:10.2196/51325

KEYWORDS

adherence to mask wearing; adherence to vaccination; social media; sociodemographic characteristics; Twitter; COVID-19; survey data

Introduction

Background

Vaccination and nonpharmaceutical public health measures (PHMs) are used to reduce the spread of SARS-CoV-2 and minimize the occurrence of severe COVID-19, hospitalizations, and deaths. Vaccination provides some immunity against infection, reduces the severity of symptoms, and prevents death [1-3]. Nonpharmaceutical PHMs aim to minimize and prevent infectious contacts and include case detection and isolation, contact tracing and quarantine, travel restrictions, physical distancing, hand washing, and mask wearing. The effectiveness of vaccination against SARS-CoV-2 transmission was repetitively shown to be high, although heterogeneous for different populations and virus variants [4]. The reported effectiveness of nonpharmaceutical PHMs in lowering SARS-CoV-2 transmission is also variable [5,6]. Whether considering vaccination or nonpharmaceutical PHMs, effectiveness is influenced by the levels of coverage and adherence across the target population. During the pandemic, several methods have been used to measure adherence to PHMs. Public opinion research can measure the knowledge, attitudes, and behavior of respondents from panels of people who have agreed to participate [7]. Limitations of this approach include time- and resource-intensiveness, missing individuals reluctant to share their opinions, and errors from respondents who may not truthfully answer or misinterpret the meaning of survey questions.

Another potential source of adherence data is social media [8,9]. While the use of open-access social media data for public health has grown over the past 20 years, there are challenges. In particular, sociodemographic information of social media users is unavailable. Consequently, it is not possible to obtain a representative sample through sampling strategy or weights. Moreover, it is difficult to validate opinions and behaviors. For example, someone may tweet that they do not value COVID-19 vaccination, even if they are vaccinated.

Objectives

The primary objective of this repeated cross-sectional study was to compare self-reported PHM adherence among people who tweeted about COVID-19 in the past 2 weeks, people who tweeted about something else in the past 2 weeks, and non–Twitter users. In addition, we aimed to compare the sociodemographic characteristics among these 3 groups, as well as among Twitter users, and to analyze reported PHM adherence according to the opinion regarding PHMs shared in their tweets.

Methods

Ethical Considerations

This project was approved by the Health Canada and the Public Health Agency of Canada Research Ethics Board (REB 2021-043P) and Comité d'éthique de la recherche en sciences et en santé at the University of Montreal (CERSES-20-056-D). Survey respondents were invited from the Angus Reid Forum, a large-scale national panel whose members receive small monetary compensation and occasional prize draws for completing surveys [10]. Consent was obtained through the Angus Reid platform, and participation to the panel and surveys was voluntary. Angus Reid reported participants could opt out at any time and refuse to answer any question. The data provided by Angus Reid were anonymized and deidentified. Although some information such as age and gender was provided, the identification of participants by a third party is very unlikely. Moreover, the data were kept in a secure server and were password protected.

Recruitment

The repeated cross-sectional surveys were conducted every fortnight (cycles) from September 2020 to March 2022, except during the Canadian federal election (end of July 2021 to mid-November 2021). Moreover, 5 additional cycles were skipped due to financial restrictions (October 7, 2020; November 4, 2020; February 16, 2021; March 16, 2021; and March 30, 2021). While the Angus Reid Forum is not a probabilistic sample of the Canadian population, it includes enough panelists from basic demographic groups and regions to ensure that survey samples are similar to the adult Canadian population and weights are calculated according to demographic and regional data [10]. To minimize survey fatigue, forum respondents were removed from the pool of potential invitees for the 4 cycles after completing the survey. This means that some participants were repeated over cycles, but because respondents were asked to report behavior and social media use during the past 2 weeks, they were treated as unique respondents. The 27 cycles of survey data were divided into 6 periods representing different phases of the virus transmission and vaccination campaign in Canada (Table 1).

XSL•FO
RenderX

Table 1. Period division of the 40,230 respondents from 27 repeated surveys and the corresponding changes in the COVID-19 pandemic in Canada.

Period	Dates	Respondents, n (%)	Pandemic information
1	September 25, 2020, to December 16, 2020	7026 (17.46)	Start of the second wave
2	January 5, 2021, to March 4, 2021	6029 (14.99)	Peak of the second wave and start of the vaccination campaign targeting the older and most susceptible populations
3	April 14, 2021, to June 10, 2021	7550 (18.77)	Third wave, and Canadian adults became eligible for vaccination
4	June 25, 2021, to August 6, 2021	6003 (14.92)	In between waves (few cases occurred)
CFE[a,b]	September 20, 2021	0 (0)	Fourth wave
5	November 18, 2021, to December 23, 2021	6074 (15.1)	Inception of the fifth wave, largely driven by the Omicron variant
6	January 10, 2022, to March 8, 2022	7548 (18.76)	Peak of the fifth wave

[a]CFE: Canadian federal election.

[b]No surveys from the dissolution of Parliament (August 15, 2021) until after the unveiling of the new federal cabinet (October 26, 2021).

Outcome Variables

Respondents were asked how often they wore a mask in public places in the previous fortnight. Adherence was dichotomized as poor (never, rarely, or sometimes) or good (every time or most of the time). Respondents were also asked whether they would be, or were already, vaccinated. The wording of the question changed according to the availability of vaccines (Multimedia Appendix 1). Adherence to vaccination was categorized into poor (would not be vaccinated) or good (was not sure, would be vaccinated, or had at least 1 dose of vaccine). The inclusion of respondents who were unsure about vaccination in the good adherence category was chosen as the vaccines were not approved at the beginning of the study and did not become available to everyone at the same time (Multimedia Appendix 1).

Exposure Variables of Primary Interest

We were mainly interested in the association of 2 variables (referred to as exposure) with the adherence to mask wearing and vaccination (referred to as outcomes): the reported use of Twitter and related tweeted opinion. We used 2 approaches to measure Twitter use and the tweeted opinion about PHMs because the questions varied over the study period (Multimedia Appendix 2). Until August 2021 (periods 1-4), reported Twitter use in the previous fortnight was categorized as one of following: tweeting about COVID-19, not tweeting about COVID-19, or not using Twitter. In periods 5 and 6, question wording was changed such that no information on Twitter users not tweeting about COVID-19 was collected (Multimedia Appendix 2). For all periods, Twitter users who reported having posted about COVID-19 were asked to select all the specific COVID-19–related topics they had tweeted about in the previous fortnight (Multimedia Appendix 2), including PHMs and vaccination. Those selecting these options were asked whether they posted being in support, neutral, or critical of the PHMs. The 3-category opinion variable was combined with Twitter use to create a 5-category variable for regression analyses over all 6 periods: did not use Twitter or did not tweet about COVID-19, tweeted about COVID-19 but not about PHMs, tweeted in support of PHMs, tweeted about PHMs in a neutral manner, and tweeted critically about PHMs.

Other Variables of Interest

Collected data also included survey date and respondent information, namely the province of residence, forward sortation area, age, gender, education level, occupation, household income, used official language, the country of origin, and ethnicity. We used the population size of the respondents' residence by linking the forward sortation area to census data [11]. The daily containment and health index (CHI) for each province provided a measure for the implementation intensity of PHMs [5]. These data were obtained from a group at the Public Health Agency of Canada that used the methodology developed by the Oxford COVID-19 Government Response Tracker [12]. It was summarized as the median CHI per cycle per province and categorized into quartiles.

Statistical Analyses

Statistical analyses were performed using R (version 4.2.2; R Foundation for Statistical Computing) [13]. For all analyses, data were considered cross-sectional due to the study design and the fact that Twitter use was reported over the 2 weeks before the survey. Descriptive statistics, including proportions with 95% CIs, were calculated for all periods using the weights provided by Angus Reid (*srvyr* package) [14]. Weighted proportions are presented in the *Results* section as proportion of Canadians. We considered the sociodemographic characteristics of participants to be similar to those of the overall Canadian population if the CI included the proportion reported by the National Census of Population 2021 [15]. Sociodemographic descriptive statistics by Twitter use categories were run for periods 1 to 4 only. We considered categories to be different when trends were outside the CI.

While there were no missing values in the survey answers, the option *prefer not to say* for income and ethnicity variables was considered as a missing value to minimize misclassification and removed from regression analyses (listwise deletion). Associations were estimated using Bayesian logistic regression models (*brms* package) [16]. Priors for regression coefficients were normal distributions centered on 0 with a precision of 1.

XSL•FO
RenderX

Intercepts were normal distributions centered on 0 with a sigma of 1.

The modifying effect of selected covariables (region, period, age, gender, language, education, and CHI) on the association between the exposure of interest (Twitter use or combined tweeted opinions) and the 2 outcomes (vaccination and mask-wearing adherence) was assessed. Additional covariables were added to the model to adjust for confounding as follows: to avoid including collider variables, the variables of the minimal sufficient adjustment set identified in a directed acyclic graph (Multimedia Appendix 3) were added one at a time and kept for the final model if the beta coefficients changed by >10% [17]. Briefly, the directed acyclic graph included 10 covariables (sociodemographic variables, period, and CHI) and their expected association with Twitter use or opinions (exposures), PHMs' adherence (outcome), and each other. Associations were based on previous findings about social media use [18,19], census information [11,15], and hypotheses from the research team. The final models were generated using 3 chains with a length of 3000, in which the first 1000 iterations were used as burn-in [20]. Convergence was monitored through the visual inspection of trace plots of variance components and density plots and by obtaining effective sample sizes. Goodness of fit was assessed using the correct classification rate and area under the curve [21]. Results are presented as odds ratio (OR) with 95% Bayesian credibility intervals (BCIs).

Results

Respondents

Survey data included 40,230 survey responses. Sociodemographic characteristics of the respondents varied slightly throughout the 6 periods (Multimedia Appendix 4). Table 2 compares the weighted distribution of respondents over the 6 periods to that of the Canadian census [15]. Several differences were noted, for example, for gender, household income, age, and the country of origin (Table 2).

XSL•FO
RenderX

Table 2. Weighted proportion of sociodemographic characteristics by Twitter use among 26,608 respondents to web-based surveys from September 2020 to mid-August 2021 and among all respondents (n=40,230; from September 2020 to March 2022)[a].

Variable	Twitter use (periods 1-4[b]), weighted proportion (95% CI)			All respondents (periods 1-6[b]), weighted proportion (95% CI)	Census 2021 (%)
	Did not use (n=20,902)	Other tweets (n=4023)	COVID-19 tweets (n=1683)		
Region					
British Columbia	13.3 (12.8-13.8)	12.9 (11.8-14.1)	10.6 (8.9-12.3)	13.1 (12.7-13.4)	13.5
Prairies[c]	17.9 (17.3-18.5)	17 (15.7-18.3)	19.5 (17.4-21.6)	17.9 (17.5-18.3)	18.2
Ontario	35.6 (34.9-36.3)	46.4 (44.6-48.1)	44.4 (41.6-47.0)	37.6 (37.2-38.2)	38.5
Quebec	26 (25.3-26.7)	15.3 (14.0-16.6)	20 (17.8-22.3)	24.1 (23.6-24.6)	23
Atlantic[d]	7.2 (6.8-7.6)	8.4 (7.4-9.3)	5.5 (4.3-6.8)	7.3 (7.0-7.6)	6.5
Population size					
Large (≥100,000)	70.7 (70.0-71.4)	76.2 (74.7-77.7)	77.6 (75.3-79.9)	71.5 (70.9-71.9)	NR[e]
Medium (30,000-99,999)	12.1 (11.6-12.6)	10.4 (9.3-11.4)	9.8 (8.2-11.4)	11.9 (11.6-12.3)	NR
Small (<30,000)	17.2 (16.7-17.8)	13.4 (12.2-14.6)	12.6 (10.8-14.4)	16.6 (16.2-17.0)	NR
Age (y)					
18-24	5.5 (5.1-5.8)	7.2 (6.3-8.1)	7.2 (5.9-8.6)	5.3 (5.1-5.6)	7.6[f]
25-34	22.4 (21.8-23.1)	21.8 (20.3-23.2)	24 (21.6-26.4)	22.9 (22.4-23.4)	17.1
35-44	16.2 (15.6-16.7)	18.4 (17.1-19.7)	21.8 (19.5-24.0)	16.7 (16.3-17.1)	17
45-54	16.8 (16.2-17.4)	21.1 (19.7-22.7)	20.9 (18.7-23.2)	17.9 (17.4-18.3)	16.1
55-64	18.1 (17.5-18.7)	17.2 (15.9-18.4)	12.1 (10.5-13.8)	17.2 (16.8-17.6)	18
65-74	16.1 (15.6-16.6)	11.7 (10.7-12.7)	11.7 (10.1-13.2)	15.6 (15.3-16.0)	14
≥75	4.9 (4.6-5.2)	2.6 (2.1-3.1)	2.3 (1.6-3.0)	4.4 (4.2-4.6)	10.2
Gender					
Men	46.8 (46.0-47.5)	52.1 (50.4-53.8)	53.4 (50.7-56.1)	48 (47.4-48.4)	49.3
Women	53.2 (52.5-54.0)	47.9 (46.2-49.6)	46.6 (43.9-49.3)	52 (51.5-52.6)	50.7
Other	__[g]	—	—	<0.1	NR
Education					
High school	34.8 (34.0-35.6)	29.4 (27.6-31.2)	31.6 (28.8-34.4)	33.8 (33.2-34.4)	32.9
College or trade school	33.3 (32.6-34.0)	31.7 (30.2-33.3)	30.7 (28.3-33.1)	32.9 (32.4-33.4)	34.3
University	31.9 (31.3-32.5)	38.9 (37.3-40.4)	37.7 (35.3-40.1)	33.3 (32.8-33.8)	32.9
Household income (CAD $)[h,i]					
<50,000	28.1 (27.4-28.9)	23.8 (22.2-25.4)	29.4 (26.7-32.0)	27.9 (27.4-28.4)	30.9
50,000-74,999	18.9 (18.2-19.5)	17.5 (16.1-18.9)	15.2 (13.2-17.2)	18.7 (18.3-19.2)	16.8[i]
75,000-99,999	17.6 (16.9-18.2)	16.2 (14.8-17.5)	17.3 (15.3-19.4)	17.2 (16.8-17.7)	20.6[j]
≥100,000	35.4 (34.7-36.2)	42.5 (40.8-44.3)	38.1 (35.4-40.8)	36.2 (35.6-36.7)	31.9
Official language					
English	79.7 (79.1-80.4)	89.9 (88.8-91.0)	84 (81.9-86.1)	81.4 (81.0-81.9)	NR
French	20.3 (19.6-20.9)	10.1 (9.0-11.2)	16 (13.9-18.1)	18.6 (18.1-19.0)	NR
Country of origin					
Canada	89.1 (88.7-89.5)	88.2 (87.2-89.3)	87.7 (86.0-89.4)	88.9 (88.6-89.2)	76.4
Abroad	10.9 (10.5-11.3)	11.8 (10.7-12.8)	12.3 (10.6-14.0)	11.1 (10.8-11.4)	23.6

XSL•FO
RenderX

Variable	Twitter use (periods 1-4[b]), weighted proportion (95% CI)			All respondents (periods 1-6[b]), weighted proportion (95% CI)	Census 2021 (%)
	Did not use (n=20,902)	Other tweets (n=4023)	COVID-19 tweets (n=1683)		
Ethnicity[k]					
European ancestry	85.8 (85.3-86.3)	82.7 (81.4-84.0)	80.8 (78.7-82.9)	84.8 (84.4-85.2)	N/A[l]
Indigenous, First Nation, Inuit, or Metis	4.7 (4.3-5.0)	5.1 (4.3-6.0)	5.3 (4.1-6.5)	4.8 (4.6-5.1)	N/A
Other ethnic ancestries	9.5 (9.1-10.0)	12.1 (11.0-13.3)	13.9 (12.1-15.7)	10.4 (10.1-10.8)	N/A

[a]The distribution of the Canadian population (census 2021) is also presented when available [15].

[b]Period 1: mid-September to the end of December 2020; period 2: January to March 2021; period 3: April to mid-June 2021; period 4: mid-June to mid-August 2021; period 5: mid-November to the end of December 2021; and period 6: January to March 2022.

[c]Alberta, Saskatchewan, and Manitoba.

[d]New Brunswick, Nova Scotia, Prince Edward Island, and Newfoundland and Labrador.

[e]NR: not reported.

[f]20 to 24 years of age in the 2021 census.

[g]Weighted estimates were 0 (0-0).

[h]Excluding the respondents who answered, “prefer not to say” (4229/40,230, 10.51%).

[i]According to the Bank of Canada, the annual conversion rates to US $ for the study period were 1.34 (2020), 1.25 (2021), and 1.30 (2022).

[j]CAD $50,000 to $69,999 and CAD $70,000 to $99,999 in the 2021 census.

[k]Excluding the respondents who answered, “Prefer not to say” (581/40,230, 1.44%).

[l]N/A: not available due to the reporting format (reporting combinations of ethnic origins, not the first reported one).

Twitter Use and Tweeted Opinions About PHMs

During periods 1 to 4, 20.6% (95% CI 20.1%-21.2%) of Canadians reported using Twitter in the previous fortnight (Multimedia Appendix 4), of whom 29.9% (95% CI 28.6%-31.3%) tweeted about COVID-19 at least once. Table 2 shows the weighted distribution of respondents by the type of reported Twitter use, compared to that of the whole sampled population and to the census data. For example, compared to the overall study population or Canadian census, Twitter nonusers were overrepresented in Quebec but underrepresented in Ontario. Indeed, the proportion of Twitter users was lower in Quebec (14.3%, 95% CI 13.3%-15.3%) than in other provinces (British Columbia: 19.3%, 95% CI 17.8%-20.7%; Prairies: 20.5%, 95% CI 19.3%-21.8%; Ontario: 25%, 95% CI 24.1%-25.9%; and Atlantic: 21.4%, 95% CI 19.4%-23.4%). There was a greater proportion of Twitter users in large (22%, 95% CI 21.3%-22.6%) than in medium (18%, 95% CI 16.4%-19.5%) and small (16.5%, 95% CI 15.3%-17.8%) population size areas, among younger (aged 18 to 54 years; 23%, 95% CI 22.3%-23.8%) than among older (aged ≥55 years; 16.6%, 95% CI 15.8%-17.3%) age groups, in men (22.6%, 95% CI 21.8%-23.4%) than in women (18.8%, 95% CI 18.1%-19.6%), and in people with university education (23.9%, 95% CI 23.1%-24.7%) than in people with lower levels of education (high school: 18.3%, 95% CI 17.3%-19.4%; college: 19.7%, 95% CI 18.8%-20.5%). Moreover, there was a lower proportion of Twitter users of European ancestry (19.9%, 95% CI 19.4%-20.5%) than Twitter users of Indigenous, First Nation, Inuit, or Metis (22.4%, 95% CI 19.7%-25.1%) or of other ethnic ancestries (25.7%, 95% CI 23.9%-27.5%).

Of the Twitter users who tweeted about PHMs, 39.8% (95% CI 37.4%-42.3%) self-declared that their tweets were in support of PHMs, 49.4% (95% CI 46.9%-51.9%) self-declared that their tweets were critical, and 10.7% (95% CI 9.2%-12.3%) self-declared that their tweets were neutral. Opinions about PHMs varied over the study period, with a greater proportion of tweets supporting PHMs in period 1 and a greater proportion of critical tweets in period 6 (Figure 1). A greater proportion of Quebec, Ontario, and Prairies inhabitants reported tweeting critically about PHMs than those from the Atlantic provinces and British Columbia (Table 3). This was also the case for men, those with an annual household income >CAD $100,000 (according to the Bank of Canada, the annual conversion rate to US $ for the study period were 1.34 in 2020, 1.25 in 2021, and 1.30 in 2022), and those without a university degree (Table 3).

XSL•FO
RenderX

Figure 1. Weighted proportions (95% CIs) of respondents who tweeted in support, critically, or neutrally about public health measures from September 2020 to March 2022 (n=1840).

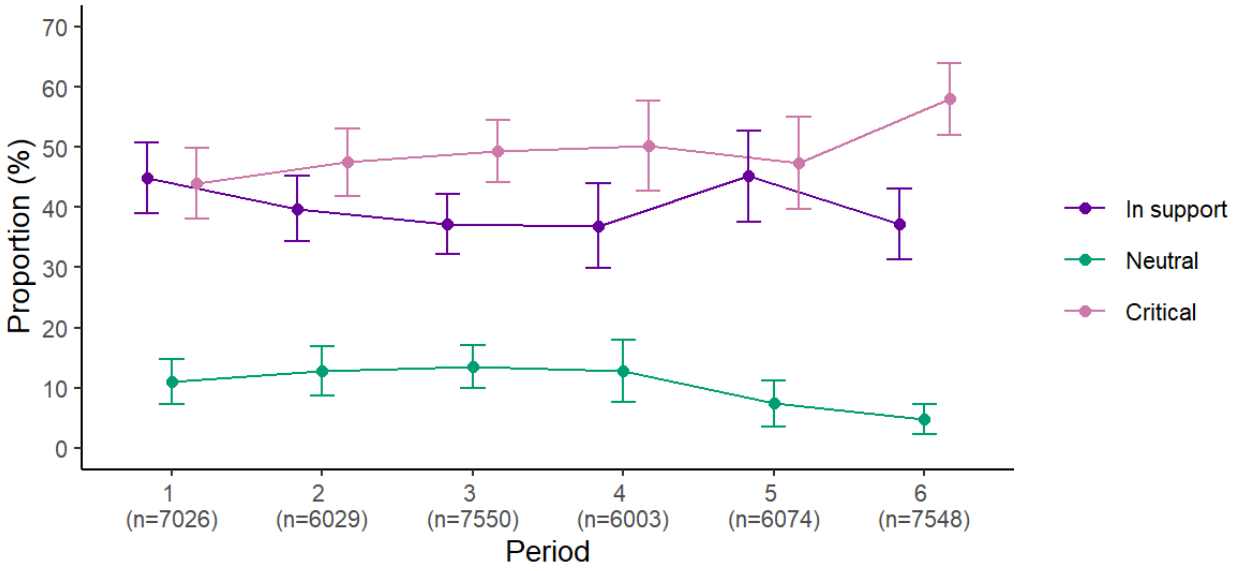

XSL•FO
RenderX

Table 3. Proportion of the sociodemographic characteristics of 1916 respondents to web-based surveys (September 2020 to March 2022) who tweeted in support of, neutrally about, and critically about public health measures.

Variable	In support (n=804), proportion (95% CI)	Neutral (n=194), proportion (95% CI)	Critical (n=918), proportion (95% CI)
Region			
British Columbia	103 (47.7 (40.2-55.1)	26 (13.2 (7.5-18.9)	84 (39.1 (31.9-46.4)
Prairies[a]	166 (32.1 (27.2-37.0)	33 (9.2 (5.8-12.6)	224 (58.7 (53.4-64.1)
Ontario	377 (42.2 (38.6-45.8)	83 (9.7 (7.6-11.8)	422 (48.1 (44.5-51.7)
Quebec	100 (33.5 (27.5-39.5)	39 (14.6 (9.9-19.3)	141 (51.9 (45.4-58.4)
Atlantic[b]	58 (49.7 (39.9-59.5)	13 (9.5 (4.3-14.7)	47 (40.8 (31.1-50.4)
Population size			
Large (≥100,000)	635 (40.3 (37.6-43.1)	156 (11.2 (9.3-13.0)	719 (48.5 (45.7-51.3)
Medium (30,000-99,999)	80 (39.5 (31.6-47.4)	19 (10.1 (5.4-14.9)	80 (50.4 (42.1-58.6)
Small (<30,000)	89 (37.6 (30.7-44.4)	18 (8.4 (4.4-12.4)	118 (54 (46.9-61.1)
Age (y)			
18-24	32 (35.8 (25.5-46.1)	17 (20.6 (11.7-29.5)	40 (43.6 (33.1-54.3)
25-34	164 (40.7 (35.2-46.3)	52 (13.2 (9.3-17.0)	179 (46.1 (40.5-51.7)
35-44	172 (37.2 (32.1-42.4)	35 (8.8 (5.6-11.9)	209 (54 (48.6-59.4)
45-54	181 (38.4 (33.3-43.5)	37 (8.4 (5.5-11.2)	224 (53.2 (47.9-58.5)
55-64	128 (40.1 (34.0-46.3)	29 (12.4 (7.7-17.2)	144 (47.5 (41.1-53.7)
65-74	106 (44.7 (38.1-51.2)	19 (7.8 (4.4-11.3)	108 (47.5 (40.9-54.1)
≥75	21 (52.9 (37.2-68.6)	5 (13.5 (2.3-24.6)	14 (33.6 (18.9-48.4)
Gender			
Men	420 (36.8 (33.6-40.0)	112 (10.8 (8.7-12.9)	538 (52.4 (49.1-55.8)
Women	384 (43.5 (39.8-47.3)	82 (10.7 (8.3-13.1)	380 (45.8 (42.0-49.6)
Education			
High school	113 (35.1 (29.5-40.8)	43 (12.8 (8.9-16.7)	167 (52.1 (46.2-58.0)
College or trade school	229 (37.1 (33.1-41.1)	51 (9.1 (6.6-11.6)	312 (53.9 (49.7-58.0)
University	462 (45.5 (42.3-48.7)	100 (10.6 (8.6-12.7)	439 (43.9 (40.7-47.1)
Household income (CAD $)[c]			
<50,000	164 (44.3 (38.6-50.0)	50 (13.5 (9.7-17.4)	157 (42.2 (36.6-47.8)
50,000-74,999	117 (42.8 (36.4-49.3)	22 (10.7 (6.0-15.4)	134 (46.5 (40.0-53.0)
75,000-99,999	134 (42.2 (36.2-48.3)	36 (11.7 (7.8-15.7)	132 (46 (39.8-52.3)
≥100,000	329 (34.6 (31.2-38.1)	72 (8.7 (6.6-10.8)	429 (56.7 (53.0-60.4)
Official language			
English	734 (41.2 (38.6-43.8)	168 (10 (8.4-11.6)	806 (48.8 (46.2-51.5)
French	70 (30.9 (24.2-37.6)	26 (15.8 (9.7-21.9)	112 (53.3 (45.8-60.8)
Country of origin			
Canada	701 (39.5 (36.9-42.1)	156 (9.8 (8.2-11.4)	826 (50.7 (48.0-53.4)
Abroad	103 (42.4 (35.5-49.4)	38 (18 (12.2-23.8)	92 (39.6 (32.7-46.4)
Ethnicity[d]			
European ancestry	658 (39 (36.4-41.7)	143 (9.8 (8.1-11.5)	790 (51.2 (48.5-54.0)
Indigenous, First Nation, Inuit, or Metis	45 (48.3 (36.4-60.1)	7 (9.4 (2.4-16.5)	30 (42.3 (30.3-54.3)

XSL•FO
RenderX

Variable	In support (n=804), proportion (95% CI)	Neutral (n=194), proportion (95% CI)	Critical (n=918), proportion (95% CI)
Other ethnic ancestry	93 (42.2 (35.0-49.2)	41 (18.8 (13.2-24.5)	85 (39 (32.0-46.0)

[a]Alberta, Saskatchewan, and Manitoba.

[b]New Brunswick, Nova Scotia, Prince Edward Island, and Newfoundland and Labrador.

[c]Excluding the respondents who answered, "prefer not to say" (135/1776, 7.6%) According to the Bank of Canada, the annual conversion rates to US $ for the study period were 1.34 (2020), 1.25 (2021), and 1.30 (2022).

[d]Excluding the respondents who answered, "prefer not to say" (24/1892, 1.27%).

Adherence to PHMs

Overall, 11% (95% CI 10.6%-11.3%) of Canadians reported poor mask-wearing adherence, which varied by period between 5.9% and 16.3%, with the highest value reported in period 4 (Multimedia Appendix 5). Mask-wearing adherence also varied by sociodemographic characteristics, with French-speaking respondents and women wearing a mask more frequently than English-speaking respondents and men, respectively. Mask-wearing frequency increased with age and education level, and decreased with income (Multimedia Appendix 6). Similarly, 10.8% (95% CI 10.4%-11.2%) of Canadians reported poor adherence to vaccination, which also varied by period. The proportion of people who reported poor adherence was 15.1% (95% CI 13.9%-16.3%) in period 1, which decreased to a minimum of 7.4% (95% CI 5.7%-9.2%) in period 4 before increasing again to 9.8% (95% CI 9.0%-10.6%) in period 6 (Multimedia Appendix 5). Vaccination adherence also varied by sociodemographic characteristics, with a greater proportion of Canadians who would not be vaccinated in the Prairies than in other regions, among men than women, among those aged between 18 and 54 years than among the older groups, and in urban areas than in rural areas. The proportion of Canadians in favor of vaccination increased with education level and household income (Multimedia Appendix 7). During periods 2 and 3 (not measured in other periods), 91.1% (95% CI 90.3%-92.0%) of Canadians who were planning to receive vaccination said they would also continue to follow other preventive health measures, including mask wearing.

Association Among Twitter Use, Tweeted Opinion, and Adherence to PHMs

The proportion of Canadians who reported poor adherence to PHMs differed by Twitter use and tweeted opinion. Overall, 19% (95% CI 16.4%-21.6%) of those who tweeted about COVID-19 had poor adherence to vaccination, and the proportion of those who had poor adherence to mask wearing varied by geographical region and age (Tables 4 and 5). The highest proportion of respondents with poor adherence to mask wearing and vaccination was in the subset who tweeted critically about PHMs (29.6% to 33.1%), while the lowest proportion was in the subset who tweeted in support of PHMs (2.2% to 5.2%; Tables 6 and 7).

In comparing non–Twitter users to participants who tweeted about COVID-19, the adjusted OR of poor adherence to mask wearing ranged from 0.79 (95% BCI 0.18-1.69) to 4.83 (95% BCI 3.13-6.86; Table 4). Geographical region and age were effect modifiers: the association was not observed in the Prairies and the Atlantic provinces, and the strength of the association decreased with age. In comparing non–Twitter users to those who tweeted about something else, the adjusted OR of poor adherence to mask wearing ranged from 0.27 (95% BCI 0.04-0.67) to 1.02 (95% BCI 0.59-1.51; Table 4). The adjusted OR of poor vaccination adherence was 1.76 (95% BCI 1.48 to 2.07) for participants who tweeted about COVID-19 and 0.67 (95% BCI 0.57 to 0.78) for those who tweeted about something else compared to non–Twitter users (Table 5).

The adjusted ORs of poor adherence to mask wearing for participants tweeting critically about PHMs were 4.07 (95% BCI 3.38-4.80) for English-speaking participants and 7.31 (95% BCI 4.26-11.03) for French-speaking participants, compared to non–Twitter users and those who did not tweet about COVID-19 (Table 6). For participants tweeting in support of PHMs compared to non–Twitter users and those who did not tweet about COVID-19, the adjusted ORs were 0.47 (95% BCI 0.31-0.64) and 0.96 (95% BCI 0.18-2.33) for English- and French-speaking participants, respectively (Table 6). The adjusted ORs of poor vaccination adherence for those tweeting critically about PHMs and for those tweeting in support of PHMs were 4.10 (95% BCI 3.40-4.85) and 0.20 (95% BCI 0.10-0.32), respectively, compared to non–Twitter users or those not tweeting about COVID-19 (Table 7).

XSL•FO
RenderX

Table 4. Weighted proportions of respondents with poor mask-wearing adherence by Twitter use, and effect modifier categories, between September 2020 and November 2021 (n=21,054)[a].

Variable	Poor adherence, weighted proportion (95% CI)	Unadjusted OR[b] (95% BCI[c])	Adjusted OR[d] (95% BCI)
Did not use Twitter in British Columbia	14.8 (13.2-16.4)	Reference	Reference
Tweeted about something else in British Columbia	10 (6.3-13.8)	0.66 (0.47-0.88)	0.53 (0.34-0.73)
Tweeted about COVID-19 in British Columbia	30.2 (21.5-38.9)	2.25 (1.52-3.13)	2.00 (1.27-2.95)
Did not use Twitter in the Prairies[e]	18.7 (17.1-20.3)	Reference	Reference
Tweeted about something else in the Prairies[e]	13.7 (10.6-16.7)	0.75 (0.58-0.95)	0.57 (0.41-0.77)
Tweeted about COVID-19 in the Prairies[e]	23.8 (18.1-29.4)	1.29 (0.95-1.69)	1.26 (0.85-1.73)
Did not use Twitter in Ontario	8.2 (7.5-9.0)	Reference	Reference
Tweeted about something else in Ontario	6.7 (5.3-8.0)	0.80 (0.64-0.99)	0.66 (0.49-0.87)
Tweeted about COVID-19 in Ontario	18 (14.8-21.2)	2.35 (1.86-2.90)	2.35 (1.71-3.00)
Did not use Twitter in Quebec	5 (4.3-5.8)	Reference	Reference
Tweeted about something else in Quebec	6.2 (3.8-8.5)	1.23 (0.81-1.73)	0.89 (0.53-1.31)
Tweeted about COVID-19 in Quebec	21.8 (16.2-27.3)	5.13 (3.54-6.83)	4.83 (3.13-6.86)
Did not use Twitter in the Atlantic provinces[f]	6.3 (4.8-7.9)	Reference	Reference
Tweeted about something else in the Atlantic provinces[f]	6.4 (3.2-9.6)	0.96 (0.50-1.53)	0.89 (0.45-1.44)
Tweeted about COVID-19 in the Atlantic provinces[f]	4 (0-8.8)	0.88 (0.23-1.88)	0.79 (0.18-1.69)
Did not use Twitter for age 18 to 24 years	10.4 (8.3-12.4)	Reference	Reference
Tweeted about something else for age 18 to 24 years	7.4 (4.0-10.8)	0.72 (0.45-1.07)	0.75 (0.43-1.13)
Tweeted about COVID-19 for age 18 to 24 years	29.5 (20.2-38.8)	2.92 (1.81-4.33)	3.15 (1.82-4.84)
Did not use Twitter for age 25 to 34 years	11.6 (10.4-12.9)	Reference	Reference
Tweeted about something else for age 25 to 34 years	9.6 (7.1-12.1)	0.89 (0.68-1.16)	0.93 (0.65-1.22)
Tweeted about COVID-19 for age 25 to 34 years	26.1 (20.7-31.5)	2.94 (2.21-3.77)	2.75 (1.90-3.77)
Did not use Twitter for age 35 to 44 years	11 (9.7-12.3)	Reference	Reference
Tweeted about something else for age 35 to 44 years	8.7 (6.3-11.2)	0.79 (0.57-1.03)	0.74 (0.52-0.99)
Tweeted about COVID-19 for age 35 to 44 years	22 (16.6-27.5)	2.01 (1.47-2.68)	1.84 (1.20-2.56)
Did not use Twitter for age 45 to 54 years	10.5 (9.3-11.8)	Reference	Reference
Tweeted about something else for age 45 to 54 years	8.3 (5.7-10.9)	0.69 (0.48-0.90)	0.71 (0.48-0.97)
Tweeted about COVID-19 for age 45 to 54 years	17.1 (12.4-21.9)	1.74 (1.24-2.40)	1.69 (1.10-2.41)
Did not use Twitter for age 55 to 64 years	9.8 (8.6-10.9)	Reference	Reference
Tweeted about something else for age 55 to 64 years	7.3 (5.1-9.5)	0.80 (0.56-1.06)	0.78 (0.53-1.09)
Tweeted about COVID-19 for age 55 to 64 years	13.9 (8.3-19.5)	1.48 (0.88-2.16)	1.40 (0.78-2.18)
Did not use Twitter for age 65 to 74 years	6.9 (6.0-7.8)	Reference	Reference
Tweeted about something else for age 65 to 74 years	7.3 (4.7-9.9)	1.05 (0.64-1.49)	1.02 (0.59-1.51)
Tweeted about COVID-19 for age 65 to 74 years	15.1 (9.8-20.4)	2.44 (1.53-3.55)	2.14 (1.18-3.29)
Did not use Twitter for age ≥75 years	8.6 (6.8-10.5)	Reference	Reference
Tweeted about something else for age ≥75 years	4.4 (0.1-8.7)	0.51 (0.14-1.04)	0.27 (0.04-0.67)
Tweeted about COVID-19 for age ≥75 years	5.9 (0-13.8)	1.03 (0.22-2.32)	0.98 (0.16-2.30)

[a]Odds ratio and 95% Bayesian credibility interval values were obtained from Bayesian logistic regression models built using repeated cross-sectional data from surveys of Canadians.

[b]OR: odds ratio.

[c]BCI: Bayesian credibility interval.

XSL•FO
RenderX

[d]Adjusted for gender, education level, language, income, population size, and period.

[e]Alberta, Saskatchewan, and Manitoba.

[f]New Brunswick, Nova Scotia, Prince Edward Island, and Newfoundland and Labrador.

Table 5. Weighted proportions of respondents with poor vaccination adherence by Twitter use between September 2020 and November 2021 (n=18,036)[a].

Variable	Poor adherence, weighted proportion (95% CI)	Unadjusted OR[b] (95% BCI[c])	Adjusted OR[d] (95% BCI)
Did not use Twitter	12.2 (11.5-12.8)	Reference	Reference
Tweeted about something else	8.2 (6.9-9.4)	0.64 (0.54-0.73)	0.67 (0.57-0.78)
Tweeted about COVID-19	19 (16.4-21.6)	1.67 (1.40-1.94)	1.76 (1.48-2.07)

[a]Odds ratio and 95% Bayesian credibility interval values were obtained from Bayesian logistic regression models built using repeated cross-sectional data from surveys of Canadians.

[b]OR: odds ratio.

[c]BCI: Bayesian credibility interval.

[d]Adjusted for age, education level, population size, and period.

Table 6. Weighted proportions of respondents with poor mask-wearing adherence by shared opinions on Twitter, and effect modifier categories, between September 2020 and March 2022 (n=33,296)[a].

Variable	Poor adherence, weighted proportion (95% CI)	Unadjusted OR[b] (95% BCI[c])	Adjusted OR[d] (95% BCI)
Did not tweet about COVID-19 or use Twitter for English-speaking respondents	11.6 (11.2-12.1)	Reference	Reference
Did not tweet about PHMs[e] for English-speaking respondents	13.9 (11.6-16.2)	1.26 (1.05-1.49)	1.03 (0.84-1.24)
Tweeted in support of PHMs for English-speaking respondents	5.0 (3.3-6.7)	0.47 (0.32-0.63)	0.47 (0.31-0.64)
Tweeted neutrally about PHMs for English-speaking respondents	10.9 (5.5-16.3)	1.01 (0.56-1.52)	1.00 (0.53-1.54)
Tweeted critically about PHMs for English-speaking respondents	33.1 (29.3-36.9)	3.90 (3.30-4.54)	4.07 (3.38-4.80)
Did not tweet about COVID-19 or use Twitter for French-speaking respondents	5.3 (4.6-5.9)	Reference	Reference
Did not tweet about PHMs for French-speaking respondents	14.4 (8.7-20.2)	2.82 (1.69-4.16)	2.69 (1.59-4.06)
Tweeted in support of PHMs for French-speaking respondents	5.2 (0-10.5)	0.96 (0.19-2.10)	0.96 (0.18-2.33)
Tweeted neutrally about PHMs for French-speaking respondents	26.3 (2.5-50.0)	2.84 (0.65-6.76)	3.13 (0.60-7.63)
Tweeted critically about PHMs for French-speaking respondents	29.6 (20.0-39.2)	8.43 (5.26-12.31)	7.31 (4.26-11.03)

[a]Odds ratio and 95% Bayesian credibility interval values were obtained from Bayesian logistic regression models built using repeated cross-sectional data from surveys of Canadians.

[b]OR: odds ratio.

[c]BCI: Bayesian credibility interval.

[d]Adjusted for gender, age, education level, income, population size, containment and health index, and period.

[e]PHM: public health measure.

XSL•FO
RenderX

Table 7. Weighted proportions of respondents with poor vaccination adherence by shared opinions on Twitter between September 2020 and March 2022 (n=28,343)[a].

Variable	Poor adherence, weighted proportion (95% CI)	Unadjusted OR[b] (95% BCI[c])	Adjusted OR[d] (95% BCI)
Did not tweet about COVID-19 or use Twitter	10.5 (10.0-10.9)	Reference	Reference
Did not tweet about PHMs[e]	10.8 (8.7-12.9)	0.98 (0.79-1.18)	1.12 (0.90-1.35)
Tweeted in support of PHMs	2.2 (0.7-3.6)	0.19 (0.10-0.30)	0.20 (0.10-0.32)
Tweeted neutrally about PHMs	18.9 (11.1-26.7)	1.75 (1.06-2.66)	1.62 (0.91-2.44)
Tweeted critically about PHMs	31.7 (27.9-35.6)	4.07 (3.47-4.77)	4.10 (3.40-4.85)

[a]Odds ratio and 95% Bayesian credibility interval values were obtained from Bayesian logistic regression models built using repeated cross-sectional data from surveys of Canadians.

[b]OR: odds ratio.

[c]BCI: Bayesian credibility interval.

[d]Adjusted for age, income, containment health index, and period.

[e]PHM: public health measure.

Discussion

Principal Findings

This is the first individual-level repeated cross-sectional study to measure the magnitude of the association, in the context of COVID-19, of self-reported Twitter use and tweeted content with self-reported mask-wearing and vaccination behaviors. Previous studies were all susceptible to ecological bias and confounding linked to the sociodemographic characteristics of social media users [8,9]. Our study lasted nearly 2 years and used web-based survey data from participants weighted to be representative of the Canadian population. Our results underscore the sociodemographic differences among types of Twitter users and non–Twitter users. Similar to other study findings [18,19], more Twitter users were younger men and living in large cities with a higher income or level of education compared to nonusers. Furthermore, our study provides evidence that people who tweet critically about PHMs are also less likely to comply with these measures.

Among all study participants, there was an overall 89% adherence to both mask wearing and vaccination. Among participants tweeting about COVID-19, however, the adherence to mask wearing was between 70% and 96%, and the adherence to vaccination was 81%. This is somewhat comparable to what was found in a worldwide Twitter poll where vaccination intention in February 2021 was 83% [22]. Canadians who tweeted critically about PHMs and those who tweeted in support of PHMs differed drastically in their adherence to mask wearing (67% to 70%, and 95%, respectively) and vaccination (68% and 98%, respectively). These results suggest that sentiments shared on Twitter are linked to adherence to PHMs but that these estimates are not representative of overall adherence. It is possible these discrepancies would be observed for other social media platforms, but this would need to be confirmed. For example, correlations were found between the sentiment of tweets extracted using an algorithm, announcement of nonpharmaceutical interventions from governmental entities, and reported compliance [23]. However, our results highlight the importance of considering sociodemographic and behavioral discrepancies in the general population, given the types of social media use and nonuse, when using health intelligence from social media for health surveillance applications [24].

The association between tweeting about COVID-19 as well as opinions about PHMs and poor adherence to PHMs was confirmed in multivariable analyses adjusting for sociodemographic confounders. Although we did not explore the underlying explanation for the association between Twitter use and PHM adherence, our findings align with a Canadian study that found an association between social media exposure and having more misperceptions and less social distancing adherence, compared to exposure to traditional news media [25]. Other studies have reported a negative association between believing in misinformation and adherence to PHMs [26-28], although not specific to social media exposure. Therefore, our findings may reflect greater misperceptions among some Twitter users [25].

The association between Twitter use and adherence to mask wearing was not uniform across age or geographical regions. The association between Twitter use and adherence to mask wearing weakened with age, except among those aged 65 to 74 years. From the beginning of the pandemic, this age group was a highly susceptible population and generally reported high adherence to PHMs [29,30], which could explain this finding. Regional differences in the association between Twitter use and adherence to mask wearing reflected stronger associations between mask wearing and Twitter use in Quebec and no associations in the Atlantic provinces [31]. There was a greater adherence to mask wearing in both these regions than in the other regions, but the Twitter user population differed. In the Atlantic provinces, the proportion of adherence to PHMs was high across Twitter users and nonusers. This could be due to the aggressive implementation of PHMs early in the pandemic, which allowed a return to economic and social activities [32]. There was a smaller proportion of Twitter users in Quebec, but those who tweeted had generally low support for PHMs. Perhaps the few Twitter users in Quebec were more prone to use this medium to voice their discontent with PHMs than those in other regions. A major difference in Quebec is that French is the

XSL•FO
RenderX

official language and is the primary language spoken. While 79% of the Quebec population speaks French at home, this proportion ranged from <1% (British Columbia, Alberta, Saskatchewan, and Newfoundland and Labrador) to 28% (New Brunswick) in other provinces [15]. While geographical region was not an effect modifier for the model assessing the association between adherence to mask wearing and opinions shared on Twitter, language was. There are French-speaking communities across the country, but this interaction is likely driven by differences in Quebec. These results point to the need to account for regional and cultural differences when using social media data for estimating populations' behaviors.

Social Media and Real-Time Monitoring of PHM Adherence

Open-access social media are among the most timely and voluminous digital information sources for public health surveillance to infer population-level health-related behaviors. Application program interfaces (APIs) enable efficient collection of high volumes of posts. Some APIs allow targeted sample collection, such as sample collection by content (eg, keywords and hashtags), languages, and regions. Associated metadata from collected social media publications can include a user identifier, time stamp of publication, and republication status. Some users make the location of publication or residence public. The publications and associated metadata can then be used to derive geolocated indicators to measure real- or near real–time changes in health-related behaviors. Possible indicators include time-series variables for the volume of publications, volume of publications about a topic, or sentiment of publications or topics. Monitoring social media for trends in PHM adherence could improve agility in adapting PHMs and communication strategies to encourage adherence. Likewise, real-time signals of low adherence could provide advanced warning for areas at a higher risk of disease emergence or re-emergence.

Public health institutes could capitalize on using social media for health intelligence through event-based surveillance systems and other tools that scan open-access data from the internet for evidence of public health threats [33,34]. However, there are ethical and practical challenges to using social media for public health surveillance, which are constantly evolving. People who post publicly may not be aware that their data are being collected or used for purposes beyond their intent, and this raises ethical, legal, and privacy considerations [35]. In addition, the open access to data through APIs can be discontinued at any moment, which happened with Twitter following the ownership change in 2022. Representative sampling from the population of interest is also difficult. Platforms may be used differently over time and according to social demographic factors [36]. Furthermore, social media contain misinformation and disinformation about health, which can influence health-related behaviors [37].

Limitations and Biases

The surveyed population was slightly different from the Canadian population [15], with an overrepresentation of younger people with a high income. This could have driven part of the sociodemographic characteristics we observed for Twitter users and overestimated the proportion of Twitter users in Canada [18,19]. However, it likely had a limited impact on the adjusted associations.

Classification errors were also possible. First, we asked respondents to self-report their adherence to PHMs; answers could have been influenced by social desirability bias [38]. It is possible that Canadians who tend to tweet in support of PHMs would be more susceptible to social desirability bias, while those tweeting critically might overstate their nonadherent behaviors. Hence, we cannot discount the possibility of differential bias when estimating the type of Twitter use. This is less likely to have happened for the measurement of Twitter use. Second, the classification used for good and poor adherence could have resulted in errors: we determined that respondents with poor adherence to vaccination included only respondents who would definitely not get vaccinated and classified respondents who were unsure as having good adherence. We used this approach because of the change in vaccine approval and availability throughout the study period [31] and to allow for the aggregation of answer choices that also changed (Multimedia Appendices 1 and 2). We classified good adherence to mask wearing as wearing a mask in public "most of the time" or "all the time." The classification bias resulting from these decisions would likely underestimate the proportion of Canadians with poor adherence, but the direction or magnitude of this bias on association measures is difficult to predict.

Conclusions

We found that Twitter users who tweeted about COVID-19 had different sociodemographic characteristics and reported lower adherence to PHMs than users who did not tweet about COVID-19 and nonusers. Moreover, the lowest adherence was among Twitter users who tweeted critically about PHMs. Furthermore, the proportion of people tweeting in support of (or critically about) PHMs was not a good indicator of adherence to PHMs in the general population. Study trends were relatively stable over the 2-year study period, supporting the utility of using indicators of health-related behavior from social media over a limited multiyear period. However, because sociodemographic information from social media users is usually unavailable, our results call for caution in using social media information to estimate support for and adherence to PHMs in the general population. Future research should determine whether adjustments using sociodemographic characteristics could mitigate this, improving the ability to predict behavior and adherence and, ultimately, to characterize or forecast trends in disease transmission.

Acknowledgments

The authors thank Suzanne Hindmarch for her feedback for the survey development and her revision of the manuscript. The authors also thank Esther Perez Trejo, who contributed to an earlier version of this manuscript. This project received funding

XSL•FO
RenderX

support from the Canadian Safety and Security Program (CSSP-2018-CP-2334) and the Canadian Institutes Health Research: Canadian 2019 Novel Coronavirus (COVID-19) Rapid Research Funding Opportunity.

Authors' Contributions

HC, EER, PD, AD, BRN, SdM, and CZ conceived the work; JD-R and AD contributed to the literature search; EER, HC, PD, CZ, BRN, and SdM coordinated data collection and quality control; JD-R performed the analysis; EER and PD supported the analysis process; JD-R, EER, SdM, and HC interpreted the data; JD-R wrote the first draft of the manuscript; EER and HC revised the consecutive drafts; and HC, EER, PD, BRN, CZ, and SdM revised the final draft.
The author JD-R is currently not affiliated with any institution but is an Independent Researcher.

Conflicts of Interest

JD-R was hired as an independent researcher for her contribution to the study and manuscript. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Questions and answer options for vaccination status for a repeated cross-sectional study assessing Twitter use and adherence to public health measures.
[DOCX File , 69 KB - jmir_v26i1e51325_app1.docx]

Multimedia Appendix 2
Questions and answer options for Twitter use for a repeated cross-sectional study assessing Twitter use and adherence to public health measures.
[DOCX File , 69 KB - jmir_v26i1e51325_app2.docx]

Multimedia Appendix 3
Directed acyclic graph for the association between Twitter use and adherence to 2 public health measures: mask wearing and vaccination.
[PNG File , 1998 KB - jmir_v26i1e51325_app3.png]

Multimedia Appendix 4
Weighted proportions of the sociodemographic characteristics of 40,230 respondents to web-based surveys over 6 periods.
[PNG File , 4608 KB - jmir_v26i1e51325_app4.png]

Multimedia Appendix 5
Weighted proportions of mask-wearing and vaccination adherence of Canadian respondents from September 2020 to March 2022.
[PNG File , 1003 KB - jmir_v26i1e51325_app5.png]

Multimedia Appendix 6
Weighted proportions of mask-wearing frequency in public places for 37,222 respondents to web-based surveys from September 2020 to March 2022 by period and sociodemographic characteristics.
[DOCX File , 71 KB - jmir_v26i1e51325_app6.docx]

Multimedia Appendix 7
Weighted proportions of the vaccination status of 31,666 respondents to web-based surveys from September 2020 to March 2022 by period and sociodemographic characteristics.
[DOCX File , 71 KB - jmir_v26i1e51325_app7.docx]

References

1. Hodgson SH, Mansatta K, Mallett G, Harris V, Emary KR, Pollard AJ. What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. Lancet Infect Dis 2021 Feb;21(2):e26-e35 [FREE Full text] [doi: 10.1016/S1473-3099(20)30773-8] [Medline: 33125914]
2. Cai C, Peng Y, Shen E, Huang Q, Chen Y, Liu P, et al. A comprehensive analysis of the efficacy and safety of COVID-19 vaccines. Mol Ther 2021 Sep 01;29(9):2794-2805 [FREE Full text] [doi: 10.1016/j.ymthe.2021.08.001] [Medline: 34365034]
3. Development and licensure of vaccines to prevent COVID-19: guidance for industry. US Food and Drug Administration. 2020. URL: https://www.fda.gov/media/139638/download [accessed 2024-04-29]

XSL•FO
RenderX

4. Chenchula S, Karunakaran P, Sharma S, Chavan M. Current evidence on efficacy of COVID-19 booster dose vaccination against the Omicron variant: a systematic review. J Med Virol 2022 Jul;94(7):2969-2976 [FREE Full text] [doi: 10.1002/jmv.27697] [Medline: 35246846]
5. Rees EE, Avery BP, Carabin H, Carson CA, Champredon D, de Montigny S, et al. Effectiveness of non-pharmaceutical interventions to reduce SARS-CoV-2 transmission in Canada and their association with COVID-19 hospitalization rates. Can Commun Dis Rep 2022 Oct 01;48(10):438-448 [FREE Full text] [Medline: 38162959]
6. Spinelli MA, Glidden DV, Gennatas ED, Bielecki M, Beyrer C, Rutherford G, et al. Importance of non-pharmaceutical interventions in lowering the viral inoculum to reduce susceptibility to infection by SARS-CoV-2 and potentially disease severity. Lancet Infect Dis 2021 Sep;21(9):e296-e301 [FREE Full text] [doi: 10.1016/S1473-3099(20)30982-8] [Medline: 33631099]
7. Defining public opinion research. Treasury Board of Canada. 2014. URL: https://www.canada.ca/en/treasury-board-secretariat/services/government-communications/public-opinion-research-government.html [accessed 2022-09-06]
8. Daughton AR, Shelley CD, Barnard M, Gerts D, Watson Ross C, Crooker I, et al. Mining and validating social media data for COVID-19-related human behaviors between January and July 2020: infodemiology study. J Med Internet Res 2021 May 25;23(5):e27059 [FREE Full text] [doi: 10.2196/27059] [Medline: 33882015]
9. Porcher S, Renault T. Social distancing beliefs and human mobility: evidence from Twitter. PLoS One 2021;16(3):e0246949 [FREE Full text] [doi: 10.1371/journal.pone.0246949] [Medline: 33657145]
10. How we poll. Angus Reid Institute. URL: https://www.angusreid.com/how-we-poll/ [accessed 2022-05-18]
11. Geographic attribute file, reference guide, census year 2021. Statistics Canada, Government of Canada. 2022. URL: https://www150.statcan.gc.ca/n1/pub/92-151-g/92-151-g2021001-eng.htm [accessed 2023-05-03]
12. Oxford COVID-19 Government Response Tracker (OxCGRT). Oxford University. URL: https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker#:~:text=The%20Oxford%20Covid%2D19%20Government,2020%2C%202021%2C%20and%202022 [accessed 2022-09-07]
13. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2024-04-29]
14. Ellis GF, Lumley T, Żółtak T, Schneider B, Krivitsky PN. srvyr: 'dplyr'-like syntax for summary statistics of survey data. Cran R. URL: https://cran.r-project.org/web/packages/srvyr/index.html [accessed 2023-06-11]
15. Census of population - data products, 2021 census. Government of Canada, Statistics Canada. 2021. URL: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/index-eng.cfm [accessed 2022-12-29]
16. Bürkner PC. brms: an R package for Bayesian multilevel models using stan. J Stat Soft 2017;80(1):1-28 [FREE Full text] [doi: 10.18637/jss.v080.i01]
17. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993 Dec 01;138(11):923-936. [doi: 10.1093/oxfordjournals.aje.a116813] [Medline: 8256780]
18. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 7. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2022-09-12]
19. Sloan L, Morgan J, Burnap P, Williams M. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. PLoS One 2015;10(3):e0115545 [FREE Full text] [doi: 10.1371/journal.pone.0115545] [Medline: 25729900]
20. Homan MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian monte Carlo. J Mach Learn Res 2014;15(1):1593-1623 [FREE Full text]
21. Fang Q, van de Schoot R. Intro to Bayesian (Multilevel) generalised linear models (GLM) in R with brms. Rens Van Schoot. URL: https://www.rensvandeschoot.com/tutorials/generalised-linear-models-with-brms/ [accessed 2022-06-05]
22. Eibensteiner F, Ritschl V, Nawaz FA, Fazel SS, Tsagkaris C, Kulnik ST, et al. People's willingness to vaccinate against COVID-19 despite their safety concerns: Twitter poll analysis. J Med Internet Res 2021 Apr 29;23(4):e28973 [FREE Full text] [doi: 10.2196/28973] [Medline: 33872185]
23. Niu J, Rees EE, Ng V, Penn G. Statistically evaluating social media sentiment trends towards COVID-19 non-pharmaceutical interventions with event studies. In: Proceedings of the 6th Social Media Mining for Health Workshop. 2021 Presented at: SMM4H '21; June 10, 2021; Mexico City, Mexico p. 1-6 URL: https://aclanthology.org/2021.smm4h-1.1.pdf [doi: 10.18653/v1/2021.smm4h-1.1]
24. Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 2021 Mar;3(3):e175-e194 [FREE Full text] [doi: 10.1016/S2589-7500(20)30315-0] [Medline: 33518503]
25. Bridgman A, Merkley E, Loewen PJ, Owen T, Ruths D, Teichmann L, et al. The causes and consequences of COVID-19 misperceptions: understanding the role of news and social media. Harv Kennedy Sch Misinformation Rev 2020 Jun 18;1:1-18 [FREE Full text] [doi: 10.37016/MR-2020-028]
26. Hornik R, Kikut A, Jesch E, Woko C, Siegel L, Kim K. Association of COVID-19 misinformation with face mask wearing and social distancing in a nationally representative US sample. Health Commun 2021 Jan;36(1):6-14. [doi: 10.1080/10410236.2020.1847437] [Medline: 33225745]

XSL•FO
RenderX

27. Lee JJ, Kang KA, Wang MP, Zhao SZ, Wong JY, O'Connor S, et al. Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. J Med Internet Res 2020 Nov 13;22(11):e22205 [FREE Full text] [doi: 10.2196/22205] [Medline: 33048825]
28. Souliotis K, Giannouchos TV, Peppou LE, Samara MT, Nimatoudis J, Papageorgiou C, et al. "Public health behaviors during the COVID-19 pandemic in Greece and associated factors: a nationwide cross-sectional survey". Inquiry 2021;58:469580211022913 [FREE Full text] [doi: 10.1177/00469580211022913] [Medline: 34053304]
29. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun 2020 May;109:102433 [FREE Full text] [doi: 10.1016/j.jaut.2020.102433] [Medline: 32113704]
30. Gustavsson J, Beckman L. Compliance to recommendations and mental health consequences among elderly in Sweden during the initial phase of the COVID-19 pandemic-a cross sectional online survey. Int J Environ Res Public Health 2020 Jul 26;17(15):e5380 [FREE Full text] [doi: 10.3390/ijerph17155380] [Medline: 32722624]
31. A global database of public health and social measures applied during the COVID-19 pandemic. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm [accessed 2022-07-08]
32. Ogden NH, Turgeon P, Fazil A, Clark J, Gabriele-Rivet V, Tam T, et al. Counterfactuals of effects of vaccination and public health measures on COVID-19 cases in Canada: what could have happened? Can Commun Dis Rep 2022 Jul 07;48(7-8):292-302 [FREE Full text] [doi: 10.14745/ccdr.v48i78a01] [Medline: 37334255]
33. Balajee SA, Salyer SJ, Greene-Cramer B, Sadek M, Mounts AW. The practice of event-based surveillance: concept and methods. Glob Secur Health Sci Policy 2021 Jan 20;6(1):1-9. [doi: 10.1080/23779497.2020.1848444]
34. Espinosa L, Wijermans A, Orchard F, Höhle M, Czernichow T, Coletti P, et al. Epitweetr: early warning of public health threats using Twitter data. Euro Surveill 2022 Sep;27(39):2200177 [FREE Full text] [doi: 10.2807/1560-7917.ES.2022.27.39.2200177] [Medline: 36177867]
35. Gilbert JP, Ng V, Niu J, Rees EE. A call for an ethical framework when using social media data for artificial intelligence applications in public health research. Can Commun Dis Rep 2020 Jun 04;46(6):169-173 [FREE Full text] [doi: 10.14745/ccdr.v46i06a03] [Medline: 32673381]
36. Campos-Castillo C, Laestadius LI. Racial and ethnic digital divides in posting COVID-19 content on social media among US adults: secondary survey analysis. J Med Internet Res 2020 Jul 03;22(7):e20472 [FREE Full text] [doi: 10.2196/20472] [Medline: 32568726]
37. Desai AN, Ruidera D, Steinbrink JM, Granwehr B, Lee DH. Misinformation and disinformation: the potential disadvantages of social media in infectious disease and how to combat them. Clin Infect Dis 2022 May 15;74(Suppl_3):e34-e39 [FREE Full text] [doi: 10.1093/cid/ciac109] [Medline: 35568471]
38. Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. Qual Quant 2011 Nov 19;47(4):2025-2047. [doi: 10.1007/s11135-011-9640-9]

Abbreviations

API: application program interface
BCI: Bayesian credibility interval
CHI: containment and health index
OR: odds ratio
PHM: public health measure

Edited by S Ma; submitted 31.07.23; peer-reviewed by J Li, S Chenchula, N Parikh, S Porcher, K Kim; comments to author 08.12.23; revised version received 31.12.23; accepted 10.06.24; published 13.08.24.

Please cite as:
Denis-Robichaud J, Rees EE, Daley P, Zarowsky C, Diouf A, Nasri BR, de Montigny S, Carabin H
Linking Opinions Shared on Social Media About COVID-19 Public Health Measures to Adherence: Repeated Cross-Sectional Surveys of Twitter Use in Canada
J Med Internet Res 2024;26:e51325
URL: https://www.jmir.org/2024/1/e51325
doi:10.2196/51325
PMID:39137009

©José Denis-Robichaud, Erin E Rees, Patrick Daley, Christina Zarowsky, Assane Diouf, Bouchra R Nasri, Simon de Montigny, Hélène Carabin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

From Tweets to Streets: Observational Study on the Association Between Twitter Sentiment and Anti-Asian Hate Crimes in New York City from 2019 to 2022

Hanxue Wei[1], MSUD, PhD; Yulin Hswen[2], MPH, ScD; Junaid S Merchant[3], PhD; Laura B Drew[3], MPH, PhD; Quynh C Nguyen[3], MSPH, PhD; Xiaohe Yue[3], MS; Heran Mane[3], BS; Thu T Nguyen[3], MSPH, SCD

[1]Department of City and Regional Planning, Cornell University, Ithaca, NY, United States

[2]Department of Epidemiology and Biostatistics, Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, United States

[3]Department of Epidemiology & Biostatistics, University of Maryland, Maryland, MD, United States

Corresponding Author:
Thu T Nguyen, MSPH, SCD
Department of Epidemiology & Biostatistics
University of Maryland
4254 Stadium Drive College Park
Maryland, MD, 20742-2611
United States
Phone: 1 301 405 3575
Email: ttxn@umd.edu

Abstract

Background: Anti-Asian hate crimes escalated during the COVID-19 pandemic; however, limited research has explored the association between social media sentiment and hate crimes toward Asian communities.

Objective: This study aims to investigate the relationship between Twitter (rebranded as X) sentiment data and the occurrence of anti-Asian hate crimes in New York City from 2019 to 2022, a period encompassing both before and during COVID-19 pandemic conditions.

Methods: We used a hate crime dataset from the New York City Police Department. This dataset included detailed information on the occurrence of anti-Asian hate crimes at the police precinct level from 2019 to 2022. We used Twitter's application programming interface for Academic Research to collect a random 1% sample of publicly available Twitter data in New York State, including New York City, that included 1 or more of the selected Asian-related keywords and applied support vector machine to classify sentiment. We measured sentiment toward the Asian community using the rates of negative and positive sentiment expressed in tweets at the monthly level (N=48). We used negative binomial models to explore the associations between sentiment levels and the number of anti-Asian hate crimes in the same month. We further adjusted our models for confounders such as the unemployment rate and the emergence of the COVID-19 pandemic. As sensitivity analyses, we used distributed lag models to capture 1- to 2-month lag times.

Results: A point increase of 1% in negative sentiment rate toward the Asian community in the same month was associated with a 24% increase (incidence rate ratio [IRR] 1.24; 95% CI 1.07-1.44; *P*=.005) in the number of anti-Asian hate crimes. The association was slightly attenuated after adjusting for unemployment and COVID-19 emergence (ie, after March 2020; *P*=.008). The positive sentiment toward Asian tweets with a 0-month lag was associated with a 12% decrease (IRR 0.88; 95% CI 0.79-0.97; *P*=.002) in expected anti-Asian hate crimes in the same month, but the relationship was no longer significant after adjusting for the unemployment rate and the emergence of COVID-19 pandemic (*P*=.11).

Conclusions: A higher negative sentiment level was associated with more hate crimes specifically targeting the Asian community in the same month. The findings highlight the importance of monitoring public sentiment to predict and potentially mitigate hate crimes against Asian individuals.

(J Med Internet Res 2024;26:e53050) doi:10.2196/53050

XSL•FO
RenderX

KEYWORDS

anti-Asian; hate crime; Twitter; racism; social media, machine learning, sentiment analysis

Introduction

In 2020, as COVID-19 infection began spreading globally, the terms “Chinese virus” and “Kung Flu” were frequently used in public discourse and media coverage [1,2]. This terminology coincided with a rise in discrimination and xenophobia toward Asian and Pacific Islander individuals residing in the United States and a subsequent increase in hate incidents that targeted Asian and Pacific Islander individuals [3]. The New York Times reported nationwide anti-Asian sentiment and violence, affecting individuals of all ages, socioeconomic status, and geographical regions [4]. According to a report by Stop AAPI Hate [5], most anti-Asian hate incidents in the United States occurred in public spaces (68%) and businesses (22%), while 10% of incidents took place online. California and New York were the top 2 states with the most reported incidents [5]. In response to the increasing violence toward Asian individuals, the COVID-19 Hate Crimes Act was instituted in 2021, providing supplementary resources for law enforcement and funding for public education [6]. Several cities, including New York City, launched local initiatives such as the “Stop Asian Hate” campaigns to address the issue at the grassroots level [7].

The effects of discrimination and violence against Asian individuals, which escalated during the COVID-19 pandemic, can be extensive and long-lasting, influencing areas such as psychological and physical health, community belonging, cultural identity, and economic security. Experiences of racial discrimination, which are a notable type of psychosocial stressor, can escalate health risks and worsen health status [8]. For Asian communities, there is a significant correlation between racial discrimination and the prevalence of anxiety, depression, suicidal thoughts, and overall psychological distress [9]. Research on racialized disasters suggests that racial discrimination related to the COVID-19 pandemic may have a detrimental impact on the health of Asian Americans and other minority groups [10,11]. A recent survey study found increasing racial discrimination during the COVID-19 pandemic was correlated with worse self-reported mental and physical health among Asian individuals in the United States [12]. In addition to direct potential health effects, discrimination and xenophobia may lead to economic consequences. In 2020, Asian restaurants experienced nearly 20% less traffic than non-Asian restaurants, translating to an estimated loss of US $7.4 billion in revenue [1].

The large-scale public interactions on social media platforms provide an opportunity to track and measure changes in public opinions and sentiment [13], including discrimination and xenophobia toward minority groups [14]. In addition, social media data can capture population-level attitudes while overcoming the limitations of traditional survey methods, including social desirability bias [15]. Social media activities and real-world hate crimes may be interconnected. Social media platforms can be used to facilitate the dissemination of extreme viewpoints, potentially contributing to the propagation of violent crimes [16]. Social media platforms also reflect public sentiment toward real-world issues and events in real time. Using the data collected from social media platforms, the link between social media sentiment and hate crimes against minority individuals has been investigated in numerous studies [17,18]. Previous work has also demonstrated an increase in solidarity toward minority groups on social media data after hate crimes, notably following the Atlanta spa shootings in 2021 [19].

While anti-Asian hate crimes have been increasing in the United States, there has been limited quantitative research linking social media sentiment to anti-Asian hate crimes. Existing studies on anti-Asian sentiment have primarily explored descriptive aspects of social or news media sources and have illustrated a rising trend during the COVID-19 pandemic [2,18,20]. Yet how online sentiment against the Asian community on social media may be an early proxy or contributor to anti-Asian hate crimes has not been studied. Understanding the link between social media sentiment and real-world hate crimes, particularly hate crimes against Asian individuals and other vulnerable communities, is crucial for creating targeted policy and social interventions. In response to this research gap, our study used quantitative methods to explore the association between anti-Asian Twitter (rebranded as X) sentiment and anti-Asian hate crimes in New York from 2019 to 2022. Our study is motivated by 1 primary research question, that is, is Twitter-derived sentiment toward the Asian community associated with the anti-Asian hate crime rate?

Methods

Overview

This research used several data sources, including New York City hate crime data from 2019 to 2022; a total of 868,012 tweets sent from New York State between 2019 and 2022; New York City demographic data in 2020; and New York City unemployment data from 2019 to 2022.

Hate crime data were obtained from the New York City Open Data Portal [21]. Crimes categorized as “anti-Asian” in terms of bias motive were selected for analysis, and we aggregated the counts of anti-Asian hate crimes in New York City at the monthly level from January 1, 2019, to December 31, 2022. Hate crimes are substantially underreported, and, in particular, Asian survivors are less likely to report hate crimes perpetrated against them compared with other survivors [22]. Thus, the actual number of monthly hate crimes was likely an underestimation than what was reported. Furthermore, finer spatial-level visualizations of yearly anti-Asian hate crimes, along with the distribution of Asian communities, were conducted to illustrate the spatial patterns at the police precinct level in New York City. We linked hate crime data with the police precinct boundary data from the New York City Open Data Portal [23]; the police precinct–level Asian population data were calculated by aggregating block-level population data from the 2020 Decennial Census [24]. Based on the existing datasets, we used QGIS and Python via Jupyter Notebook to create all the figures. For Twitter sentiment data, we used

XSL•FO
RenderX

Twitter's application programming interface (API) for Academic Research to collect a random 1% sample of publicly available tweets between January 1, 2019, and December 31, 2022. Our collection was restricted to tweets in English in New York State, including New York City, that included 1 or more of the 10 Asian-related keywords: "Asian," "Asians," "Filipino," "Japanese," "Korean," "Nepal," "Pacific Islander," "Thai," "Vietnamese," and "Chinese." The word list originated from previous research [25,26]. In cases where tweets lacked city or state data, users' locations were used instead. If no city or state information was obtainable, we categorized them as having missing values. We applied a support vector machine, a supervised machine learning algorithm, to classify sentiment. The training dataset includes a total of 19,178 prelabeled tweets, consisting of both public sources from Sentiment140 (n=498), Kaggle (n=7086), and Sanders (n=5113) and data manually labeled by our study team (n=6,481). We created distinct binary classifiers for negative and positive sentiments, respectively. The support vector machine has an accuracy of 0.91, an F_1-score of 0.84, and a precision of 0.89 in the categorization of negative sentiment.

Negative binomial models were used to investigate the relationship between hate crimes and Twitter sentiment. Negative binomial models were better for our count data as compared with the Poisson regression given the overdispersed distribution of the monthly counts of hate crimes. In our study, the dependent variable in the negative binomial model was the monthly count of anti-Asian hate crimes in New York City. We first used negative binomial models to investigate the association between the proportion of tweets referencing the Asian community that expressed negative (or positive) sentiment and the counts of anti-Asian hate crimes in New York City in the same month from 2019 to 2022. In addition to the crude models, we ran adjusted models with an indicator variable for the COVID-19 pandemic (before and after the start date of March 2020) and the monthly unemployment rate in New York City in the same month. March 2020 was when the World Health Organization (WHO) characterized COVID-19 infection as a pandemic and when COVID-19 infection began to be referred to as the "China virus". For our COVID-19 pandemic variable, we used a 0 value for periods before March 2020 and a 1 value for periods from March 2020 onward. This allowed us to determine the social shift that the term "China virus" had on Twitter sentiment and anti-Asian hate crimes. In addition, the unemployment rate is considered a confounder in our study: unemployment could affect psychological well-being [27], which may be reflected in increased frustration and negativity expressed online. A previous study also shows that sentiments of surprise and fear were associated with poverty and unemployment rates since the start of the COVID-19 pandemic [28]. Unemployment may also be linked to the rate of offline hate crimes. Research has shown that periods of high unemployment often coincide with an increase in hate crimes [29,30], as economic distress may exacerbate social tensions and prejudices. Thus, we adjusted for New York City's monthly unemployment rate based on data from the Bureau of Labor Statistics [31]. In addition, it may take time for sentiment expressed online to be reflected in real-world hate crimes—for instance, a previous study connecting online sentiment with daily worry among Hispanic individuals found a significant 1-week lag [12]. In our main analysis, using the same month data would include a time lag within days and up to 4 weeks. We also used distributed lag models as sensitivity analysis, adding in a 1- and 2-month lag in addition to a 0-month lag time to capture longer time lags [32].

Ethical Considerations

All data used in this study were publicly available and did not include any personal identifiers. The Twitter sentiment data were analyzed in aggregate, and individual tweets were not identified in any way. The publicly available hate crime data provided by the New York City Police Department were anonymized and deidentified before our analysis, ensuring the privacy of the individuals involved in these incidents. This study was determined exempt by the University of Maryland College Park Institutional Review Board (1797788-1).

Results

Figure 1 [21] illustrates the number of anti-Asian hate crimes in New York City by month. A noticeable increase can be seen starting from March 2020. The peak occurred in March 2021, and a substantial number of hate crimes continued to be reported. In total, there were 1, 33, 150, and 91 hate crimes reported in 2019, 2020, 2021, and 2022, respectively, indicating a dramatic increase during the early stages of the pandemic [21].

Figures 2 and 3 [21,23,24] reveal a possible association between anti-Asian hate crimes and the locations of high Asian population densities at the police precinct level. Figure 2 provides an annual overview of anti-Asian hate crime distribution across New York City's police precincts. It is noteworthy that the frequency of such crimes experienced a surge since 2020, particularly in the Garment District and Union Square regions where there are significant Asian populations. Furthermore, by 2021, an escalation of anti-Asian hate crimes was observed in the northern region of Queens and the southern part of Brooklyn [21,23]. Figure 3 represents the proportion of the Asian population to the total population at the police precinct level in 2020 (for calculation, refer Multimedia Appendix 1). As illustrated in the figure, areas witnessing higher frequencies of anti-Asian hate crimes aligned with regions housing a significant Asian demographic to some extent. At the precinct level, the Kendall tau correlation was 0.28 and the Pearson correlation coefficient (r) was 0.33, suggesting that there was a moderate positive relationship between the concentration of Asian populations and the total number of anti-Asian hate crimes [23,24].

XSL•FO
RenderX

Figure 1. Monthly count of anti-Asian hate crimes in New York City from 2019 to 2022. Data source: New York City Police Department.

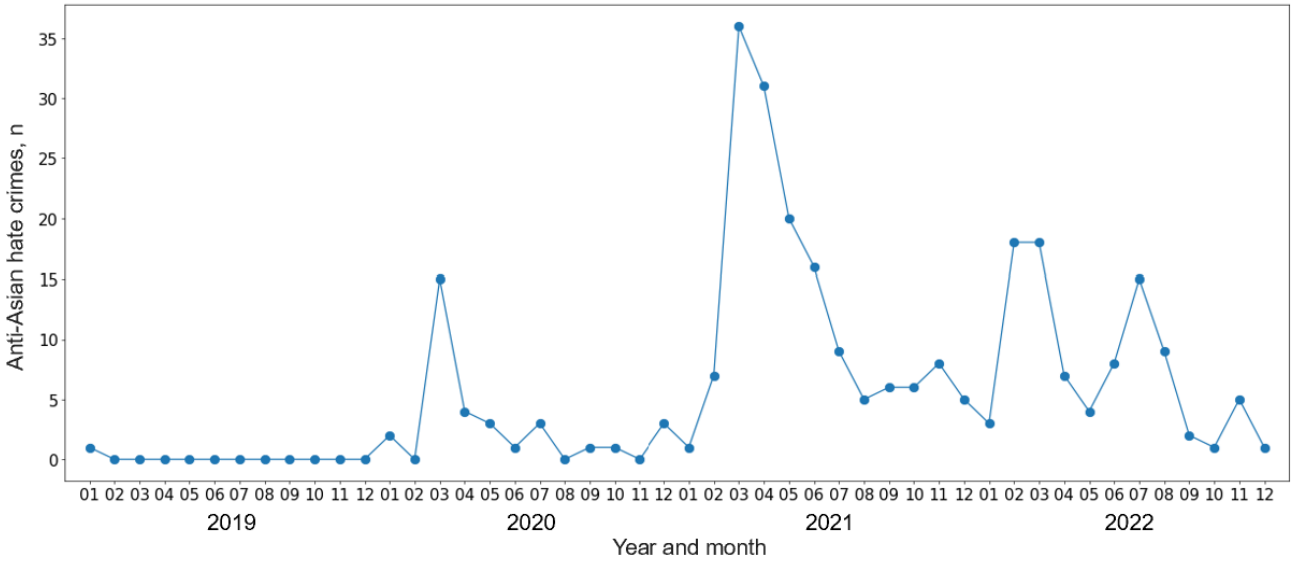

Figure 2. Number of anti-Asian hate crimes by police precincts in New York City from 2019 to 2022. Data source: New York City Police Department and NYC Open Data.

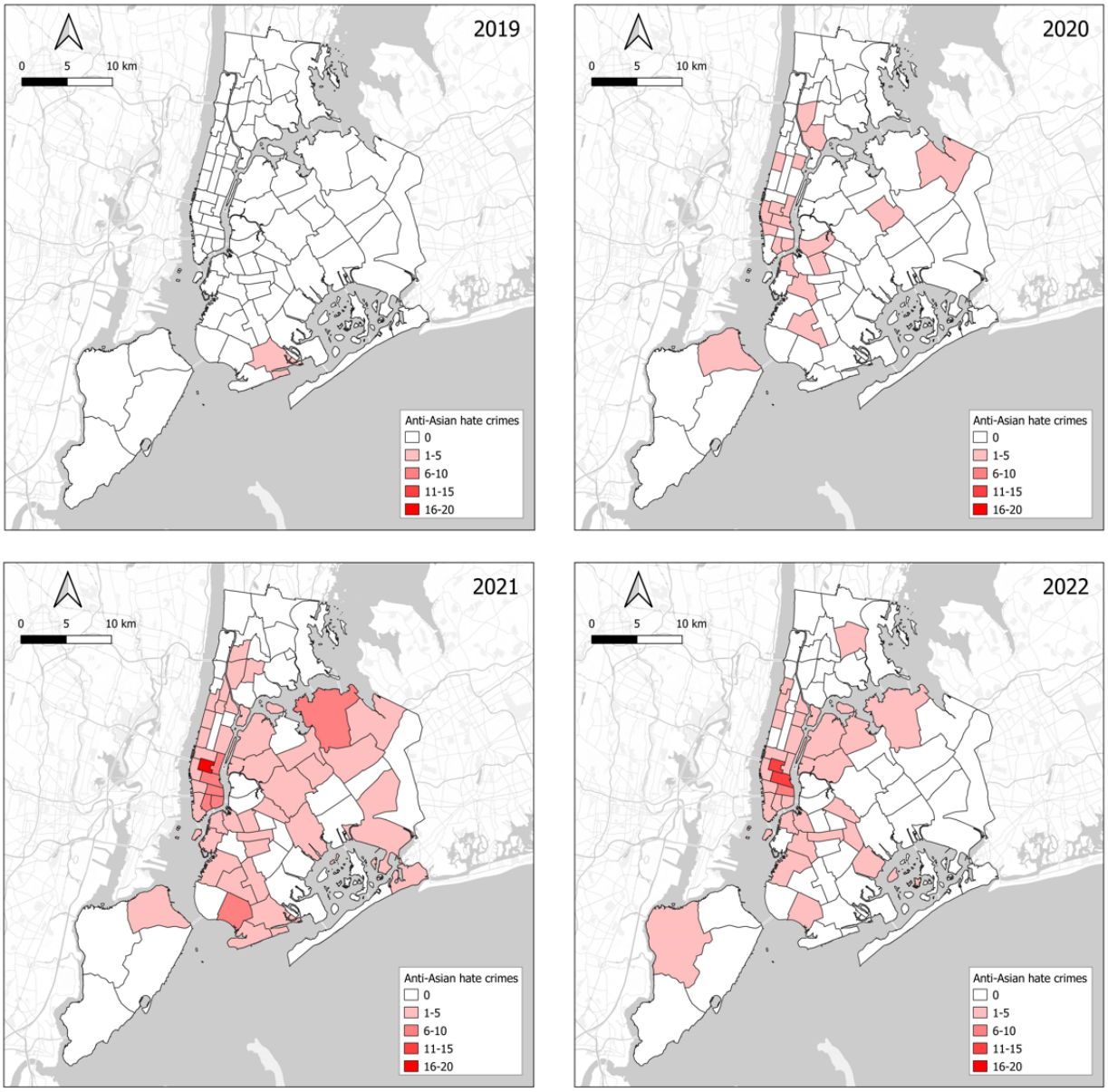

XSL•FO
RenderX

Figure 3. Asian population proportion out of the total population by police precinct in New York City based on 2020 Decennial Census. Data source: NYC Open Data and 2020 Decennial Census.

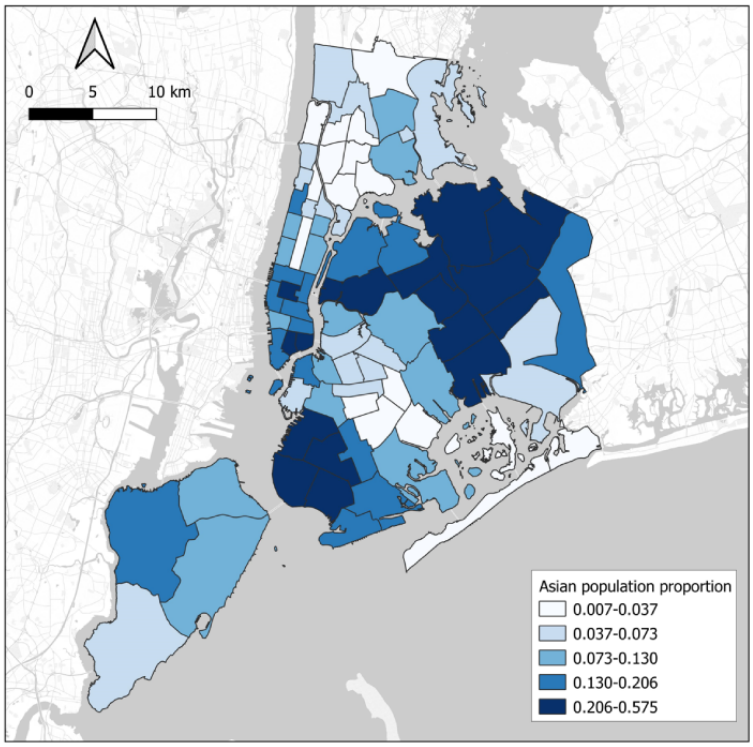

Figure 4 illustrates the temporal fluctuations of positive and negative Twitter sentiments for tweets containing the term "Asian" for New York City on a monthly basis. Despite the observable fluctuations in sentiment trends, there was an overall downtrend in positive sentiment over time. For the negative sentiment, it peaked around March 2020, when COVID-19 infections started to spread nationally in the United States, and in March 2021. At these 2 time points, positive sentiment rates were at their lowest. In addition, the proportion of New York City tweets made up approximately three-fourths of the total New York tweets, and these 2 displayed similar trajectories in terms of both positive and negative sentiment rates (with r=0.97 for both positive and negative sentiment between New York City and New York).

Figure 4. Negative and positive rates of Twitter sentiment toward the Asian community in New York City (NYC) by month from 2019 to 2022. The vertical line in March 2020 marks the point when the COVID-19 pandemic began to spread nationally in the United States.

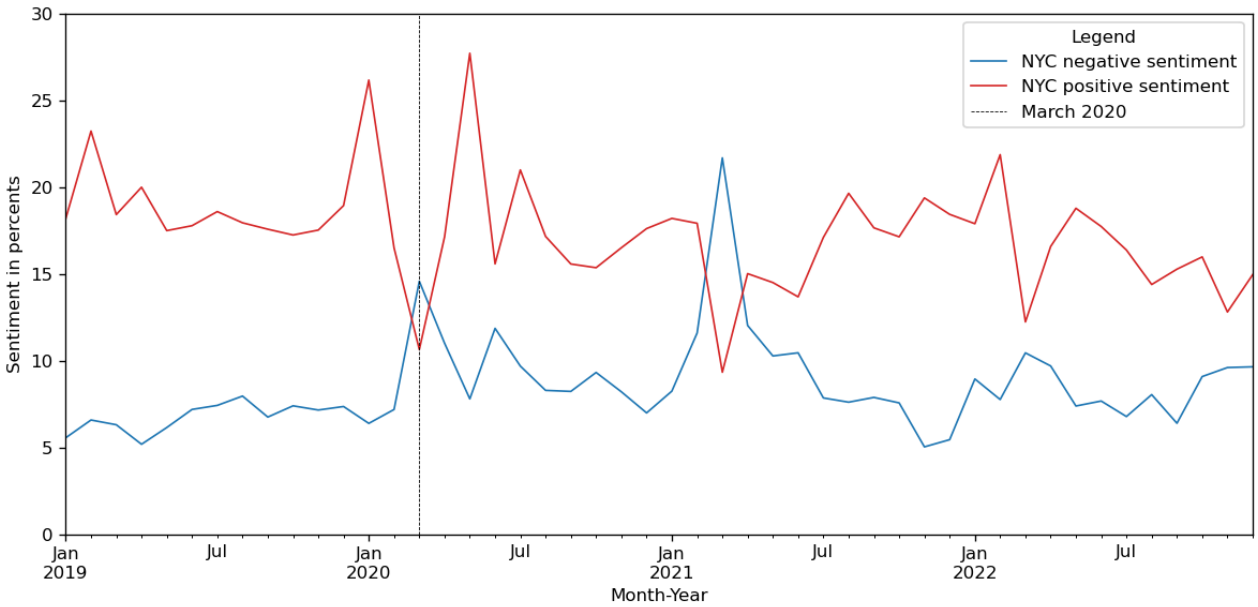

Table 1 shows the model between anti-Asian sentiment and anti-Asian hate crimes in the same month. The crude model 1.0, with a 0-month lag, demonstrates that a 1% point increase in negative sentiment rate toward the Asian community in the same month was associated with a 24% increase (incidence rate ratio [IRR] 1.24; 95% CI 1.07-1.44; P=.005) in the number of

XSL•FO
RenderX

anti-Asian hate crimes. The significant association persisted but was slightly attenuated after adjusting for unemployment and COVID-19 emergence (ie, after March 2020) in adjusted model 1.1, whereby each percentage point increase in negative sentiment was associated with a 16% increase (IRR 1.16; 95% CI 1.06-1.28; *P*=.008) in expected anti-Asian hate crimes. The positive sentiment toward Asian tweets with a 0-month lag was associated with a 12% decrease (IRR 0.88; 95% CI 0.79-0.97; *P*=.008) in expected anti-Asian hate crimes in the same month in crude model 2.0, and the relationship was no longer significant after adjusting for the unemployment rate and the emergence of COVID-19 pandemic, as shown in model 2.1 (*P*=.11). In addition, there are low to moderate correlations among the independent variables, as shown in Multimedia Appendix 2.

Table 1. Results of negative binomial models examining the association between Twitter sentiment rate (negative and positive) and the anti-Asian hate crimes in New York City from 2019 to 2022.

Model[a] (N=48)	Incidence rate ratio (95% CI)	*P* value
Crude model 1.0		
Negative sentiment (0-month lag)	1.24 (1.07-1.44)	.005
Crude model 2.0		
Positive sentiment (0-month lag)	0.88 (0.79-0.97)	.008
Adjusted model 1.1		
Negative sentiment (0-month lag)	1.16 (1.06-1.28)	.002
Adjusted model 2.1		
Positive sentiment (0-month lag)	0.93 (0.86-1.02)	.11

[a]The crude models 1.0 and 2.0 regress the number of anti-Asian hate crimes on negative or positive sentiment with a 0-month lag. The adjusted models 1.1 and 2.1 control for the monthly unemployment rate and whether it was during the COVID-19 pandemic (using a 0 value for periods before March 2020 and a value of 1 from March 2020 onward). All models use monthly data and measure sentiment in percentages.

With the assumption that the sentiment on social media may affect hate crimes over time, Multimedia Appendix 3 presents a cross-correlation table between the number of anti-Asian hate crimes per month and sentiment of Twitter data in New York City regarding the Asian community using Pearson *r*. The results suggest a complex relationship where immediate correlations (eg, 0-month lag) are stronger, but there are also potential delayed associations between Twitter sentiment and the number of anti-Asian hate crimes. We further conducted a sensitivity analysis using distributed lag models (presented in Multimedia Appendix 4) adding in a 1- and 2-month lag in addition to a 0-month lag time. For the model with a 0-, 1-, and 2-month lag, a 1% point increase in negative sentiment toward the Asian community with a 0-month lag was associated with a 15% increase (IRR 1.15; 95% CI 1.01-1.31; *P*=.04) in anti-Asian hate crimes. Positive sentiment toward the Asian community on Twitter was associated with an 11% (IRR 0.89; 95% CI 0.81-0.97; *P*=.01) decrease in anti-Asian hate crimes with a 0-month lag. There was no significant association between negative or positive sentiment toward the Asian community and anti-Asian hate crimes when examining 1- to 2-month lag times (all *P*>.05).

Discussion

Principal Findings

This study investigates how Twitter sentiment toward the Asian community was associated with anti-Asian hate crimes before and during the COVID-19 pandemic in New York City. We found a positive association between the negative Twitter sentiment rate and the counts of anti-Asian hate crimes in the same month. Positive sentiment toward the Asian community was not associated with hate crimes when adjusted for covariates. Our study highlights the potential of using social media data in predicting and addressing societal issues, including hate crimes. The finding that negative sentiment toward the Asian community was associated with a higher number of anti-Asian hate crimes thereafter may be due to several reasons; it is possible that increased anti-Asian negative sentiment on Twitter could serve as an early indicator of the overall societal climate of racism and discrimination, which were shown in various forms, including hate crimes in the following period. In addition, Twitter can spread and amplify existing prejudices and discriminative views, leading to offline actions such as hate crimes that reflect these attitudes. Supporting this, a study by Nguyen et al [18] demonstrated that residing in a state with a 10% higher negative sentiment in tweets referencing Black individuals was linked to lower odds of individuals endorsing that Black-White disparities in jobs, income, and housing stem from discrimination and higher odds of endorsing that disparities are due to lack of willpower, as well as higher explicit and implicit racial biases. Furthermore, increased negative sentiment on Twitter could represent growing feelings of fear, anger, and vulnerability within Asian communities in response to racism and discrimination, serving as another indicator for the subsequent rise in hate crimes.

The pandemic likely exacerbated anti-Asian hate crimes, potentially due to rhetoric from the media blaming Asian communities for the COVID-19 pandemic, thus fostering an environment that may have facilitated anti-Asian discrimination and violence. March 2020 and March 2021 witnessed peaks in negative sentiment and sharp declines in positive sentiment. March 2020 aligns with the initial COVID-19 outbreak and pandemic declaration by the WHO, and there was a surge of

XSL•FO
RenderX

stigmatizing language in the media during this period, including the emergence and use of the phrase "Chinese virus" by public officials and news media [20,33], which likely contributed to the dramatic spike in negative sentiment and drop in positive sentiment as the pandemic emerged. In March 2021, anti-Asian stigma persisted nearly 1 year into the pandemic. In addition, high-profile incidents such as the Atlanta spa shootings in March 2021 coincided with increased awareness of enduring bias and discrimination toward Asian individuals, a sentiment that was reflected in online discussions on Twitter [19].

We are among the first to quantitatively analyze the use of Twitter sentiment data to model anti-Asian hate crimes in the context of the COVID-19 pandemic and to identify the links between online expressions of negative anti-Asian sentiment on social media and offline anti-Asian hate. Although the relationship between social media data and real-world hate crimes has been investigated for some time, it is only recently that social media data have been used in research on anti-Asian hate during the COVID-19 pandemic. Arguing that "racism is a virus," He et al [34] collected anti-Asian hate and counter speech in tweets and measured the probability of a user becoming hateful (ie, tweeting hate speech for the first time) after being exposed to different percentages of hateful versus counterspeech neighbors in the network. The study found that counterspeech messages discouraged users from adopting hateful attitudes.

Our research highlights how societal crises such as pandemics may potentially exacerbate prejudices, racism, and xenophobia—a phenomenon that is not new in history [35]. Crises like the COVID-19 pandemic may lead to more social cleavages, as evidenced by racist sentiments on social media platforms. This could then contribute to hate crimes targeting minority communities. As an important subject of public health, hate crimes are linked to various psychological and physical harm such as depression, anxiety, stress, and alcohol or drug use [36]. We underscore the urgent need for public health policies that address the compound effects of such experiences. It is important for communities, governments, and social media companies to proactively implement targeted interventions, launch education campaigns, and establish regulations aimed at combating prejudice and discrimination and fostering social cohesion.

This study is not without limitations, but it paves the way for future research directions. First, our research relied upon reported hate crimes, which are often underreported, and Asian survivors are less likely to report hate crimes compared with non-Asian survivors [22]. Therefore, the current count of hate crimes may be an underestimate. Second, there were temporal and geographical limitations concerning granularity. The models evaluated hate crimes and Twitter sentiment at the city level, disregarding finer spatial variations due to the limitations of Twitter data. In addition, the time unit was a month, and subsequent studies could use weekly or daily data, which could provide more units for the modeling. Future research could also benefit from incorporating more granular demographic and socioeconomic data if available. Furthermore, our models using a 0-month lag could not determine causal directionality, as they did not establish whether the increase in anti-Asian Twitter sentiment preceded or followed the rise in anti-Asian hate crimes. In addition, while we used dichotomized time variables to indicate the start of the COVID-19 pandemic, future research might benefit from using other indicators, such as the number of COVID-19 pandemic cases, to further explore their potential influence. Future research could also expand upon our exploratory research and delve deeper into analyses that examine causality.

Twitter users are not representative of the general US population, as younger and higher socioeconomic populations are overrepresented on Twitter [37]. While the sentiment of tweets does not necessarily specify racial prejudice, through this approach, we were able to capture a broad signal of public attitudes, providing insights that may be indicative of a proxy of larger societal trends. It is also important to note the role of social bots on social media. Deceptive bots can manipulate discussions and potentially skew social media study results, a factor that warrants further consideration [38]. However, it can be challenging for the public to differentiate between a bot and a human posting a message. Bots can help create the social context on online social media platforms, which is important to capture. Eliminating bots from the dataset would mean not capturing this exposure. In addition, bots may represent only a small portion of overall social media activity. In our race-related dataset of tweets from 2011 to 2021, we previously investigated the presence of highly frequent users, which can indicate the presence of bots. We found that users tweeting more than 1000 times per year represented less than 1% of the data [39].

Studies have also shown a surge in hate speech following Musk's acquisition of Twitter in 2022, which could potentially skew the findings of this study [40-42]. This implies that the sentiments expressed on Twitter, especially negative ones, may not be directly comparable before and after the takeover without certain adjustments. Future research should take into account such significant events that could potentially impact social media sentiment when investigating its correlation with real-world crimes. Furthermore, changes in Twitter's API access and user behavior following Musk's takeover in 2023 [43-45] pose challenges to the replication of this study and the general use of Twitter data for research purposes. However, our study provides a method framework that can be applied to other social media platforms or data sources. The landscape of social media is dynamic and constantly evolving, and researchers must adapt to these changes. Future research should explore alternative data sources, consider the potential biases introduced by changes in social media platforms, and develop methods to ensure the validity and reliability of social media data for research.

Public Health Implications

This study underscores the association between negative Twitter sentiment toward the Asian community and anti-Asian hate crimes in New York City, particularly in the context of the COVID-19 pandemic. Most importantly, we demonstrate that social media data can offer new insights for public health in the face of hate crimes. Early detection of spikes in sentiment could help with timely interventions, reducing the negative impacts on the well-being of communities resulting from hate crimes. Overall, it is crucial for governments, social media companies,

XSL•FO
RenderX

and communities to address both the immediate health consequences of hate crimes and the broader issue of societal discrimination, and social media data can serve as an effective surveillance tool.

Acknowledgments

Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities (R01MD015716 [TTN] and R01MD016037 [QCN]) and the National Library of Medicine (R01LM012849 [QCN]). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets using hate crime data analyzed during this study were provided by the New York City Police Department and are publicly available. The data from X were collected using X's application programming interface (API) for Academic Research. The API for Academic Research was discontinued in May 2023. X data analyzed during this study cannot be shared because X restricts content distribution.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Calculation of precinct-level population proportion.
[DOCX File , 13 KB - jmir_v26i1e53050_app1.docx]

Multimedia Appendix 2
Correlation table for independent variables (without lag).
[PNG File , 126 KB - jmir_v26i1e53050_app2.png]

Multimedia Appendix 3
Cross-correlation between the number of anti-Asian hate crimes per month and sentiment of Twitter data in New York City regarding the Asian community (using Pearson r). (A) Cross-correlation for the number of anti-Asian hate crimes per month and negative Twitter sentiment at different lag times. (B) Cross-correlation for the number of anti-Asian hate crimes per month and positive Twitter sentiment at different lag times.
[PNG File , 72 KB - jmir_v26i1e53050_app3.png]

Multimedia Appendix 4
Model results for New York City negative and positive sentiment rate using distributed lag regression (n=46 months).
[DOCX File , 14 KB - jmir_v26i1e53050_app4.docx]

References

1. Huang JT, Krupenkin M, Rothschild D, Lee Cunningham J. The cost of anti-Asian racism during the COVID-19 pandemic. Nat Hum Behav 2023;7(5):682-695. [doi: 10.1038/s41562-022-01493-6] [Medline: 36658211]
2. Hswen Y, Xu X, Hing A, Hawkins JB, Brownstein JS, Gee GC. Association of "#covid19" versus "#chinesevirus" with anti-Asian sentiments on Twitter: March 9-23, 2020. Am J Public Health 2021;111(5):956-964. [doi: 10.2105/AJPH.2021.306154] [Medline: 33734838]
3. Lantz B, Wenger MR. Anti-Asian xenophobia, hate crime victimization, and fear of victimization during the COVID-19 pandemic. J Interpers Violence 2023;38(1-2):NP1088-NP1116 [FREE Full text] [doi: 10.1177/08862605221086651] [Medline: 35544766]
4. Cai W, Burch ADS, Patel JK. Swelling anti-Asian violence: who is being attacked where.: The New York Times; 2021. URL: https://www.nytimes.com/interactive/2021/04/03/us/anti-asian-attacks.html [accessed 2024-07-29]
5. Two years and thousands of voices: what community-generated data tells us about anti-AAPI hate.: Stop AAPI Hate URL: https://stopaapihate.org/wp-content/uploads/2023/06/22-SAH-NationalReport-July-F.pdf [accessed 2024-07-29]
6. S.937 - COVID-19 Hate Crimes Act.: United States Congress; 2021. URL: https://www.congress.gov/bill/117th-congress/senate-bill/937/text [accessed 2024-07-29]
7. Stop Asian hate.: New York City Human Rights URL: https://www.nyc.gov/site/cchr/community/stop-asian-hate.page [accessed 2024-07-29]
8. Williams DR, Mohammed SA. Racism and health I: pathways and scientific evidence. Am Behav Sci 2013;57(8):1152-1173 [FREE Full text] [doi: 10.1177/0002764213487340] [Medline: 24347666]

XSL•FO
RenderX

9. Lee DL, Ahn S. Racial discrimination and Asian mental health: a meta-analysis. The Counseling Psychologist 2011;39(3):463-489. [doi: 10.1177/0011000010381791]
10. Hswen Y. Online hate: the new virus. Am J Public Health 2022;112(4):545-547. [doi: 10.2105/AJPH.2022.306754] [Medline: 35319954]
11. Chen JA, Zhang E, Liu CH. Potential impact of COVID-19-related racial discrimination on the health of Asian Americans. Am J Public Health 2020;110(11):1624-1627. [doi: 10.2105/AJPH.2020.305858] [Medline: 32941063]
12. Lee S, Waters SF. Asians and Asian Americans' experiences of racial discrimination during the COVID-19 pandemic: impacts on health outcomes and the buffering role of social support. Stigma and Health 2021;6(1):70-78. [doi: 10.1037/sah0000275]
13. Klašnja M, Barberá P, Beauchamp N, Nagler J, Tucker J. Measuring public opinion with social media data. In: The Oxford Handbook of Polling and Polling Methods. New York, NY: Oxford University Press; 2015:555-582.
14. Hswen Y, Qin Q, Williams DR, Viswanath K, Subramanian SV, Brownstein JS. Online negative sentiment towards Mexicans and Hispanics and impact on mental well-being: a time-series analysis of social media data during the 2016 United States Presidential Election. Heliyon 2020;6(9):e04910 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e04910] [Medline: 33005781]
15. Krumpal I. Social desirability bias and context in sensitive surveys. In: Michalos AC, editor. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht, Netherlands: Springer; 2023:6037-6043.
16. Müller K, Schwarz C. Fanning the flames of hate: social media and hate crime. J Eur Econ Assoc 2021;19(4):2131-2167. [doi: 10.1093/jeea/jvaa045]
17. Müller K, Schwarz C. From hashtag to hate crime: Twitter and antiminority sentiment. Am Econ J 2023;15(3):270-312 [FREE Full text]
18. Nguyen TT, Huang D, Michaels EK, Glymour MM, Allen AM, Nguyen QC. Evaluating associations between area-level Twitter-expressed negative racial sentiment, hate crimes, and residents' racial prejudice in the United States. SSM Popul Health 2021;13:100750 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100750] [Medline: 33665332]
19. Criss S, Nguyen TT, Michaels EK, Gee GC, Kiang MV, Nguyen QC, et al. Solidarity and strife after the Atlanta spa shootings: a mixed methods study characterizing Twitter discussions by qualitative analysis and machine learning. Front Public Health 2023;11:952069 [FREE Full text] [doi: 10.3389/fpubh.2023.952069] [Medline: 36825140]
20. Nguyen TT, Criss S, Dwivedi P, Huang D, Keralis J, Hsu E, et al. Exploring U.S. shifts in anti-Asian sentiment with the emergence of COVID-19. Int J Environ Res Public Health 2020;17(19):7032 [FREE Full text] [doi: 10.3390/ijerph17197032] [Medline: 32993005]
21. NYPD Hate Crimes.: NYC Open Data URL: https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78 [accessed 2024-07-29]
22. Lantz B, Wenger MR. Are Asian victims less likely to report hate crime victimization to the police? Implications for research and policy in the wake of the COVID-19 pandemic. Crime & Delinquency 2021;68(8):1292-1319. [doi: 10.1177/00111287211041521]
23. Police precincts.: NYC Open Data URL: https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz?category=Public-Safety&view_name=Police-Precincts [accessed 2024-07-29]
24. 2020 Decennial Census.: US Census Bureau URL: https://data.census.gov/all?q=redistricting [accessed 2024-07-29]
25. Bartlett J, Reffin J, Rumball N, Williamson S. Anti-Social Media. London, United Kingdom: Demos; 2014.
26. The racial slur database. URL: http://www.rsdb.org/ [accessed 2023-02-03]
27. Proserpio D, Counts S, Jain A. The psychology of job loss: using social media data to characterize and predict unemployment. 2016 Presented at: Proceedings of the 8th ACM Conference on Web Science; 2016 May 22; Hannover, Germany p. 223-232.
28. Fan L, Yu H, Yin Z. Stigmatization in social media: documenting and analyzing hate speech for COVID-19 on Twitter. Proc Assoc Inf Sci Technol 2020;57(1):e313 [FREE Full text] [doi: 10.1002/pra2.313] [Medline: 33173820]
29. Green DP, Glaser J, Rich A. From lynching to gay bashing: the elusive connection between economic conditions and hate crime. J Pers Soc Psychol 1998;75(1):82-92. [doi: 10.1037//0022-3514.75.1.82] [Medline: 9686451]
30. Raphael S, Winter‐Ebmer R. Identifying the effect of unemployment on crime. J Law Econ 2001;44(1):259-283 [FREE Full text]
31. Civilian labor force - seasonally adjusted.: U.S. Bureau of Labor Statistics URL: https://www.bls.gov/regions/new-york-new-jersey/data/xg-tables/ro2xglausnyc.htm [accessed 2024-07-29]
32. Hswen Y, Zhang A, Brownstein JS. Estimating the incidence of cocaine use and mortality with music lyrics about cocaine. NPJ Digit Med 2021;4(1):100 [FREE Full text] [doi: 10.1038/s41746-021-00448-x] [Medline: 34193959]
33. Budhwani H, Sun R. Creating COVID-19 stigma by referencing the novel Coronavirus as the "Chinese virus" on Twitter: quantitative analysis of social media data. J Med Internet Res 2020;22(5):e19301 [FREE Full text] [doi: 10.2196/19301] [Medline: 32343669]
34. He B, Ziems C, Soni S, Ramakrishnan N, Yang D, Kumar S. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. 2021 Presented at: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining; 2021 November; Virtual Event, Netherlands p. 90-94.
35. Schaller M, Neuberg SL. Chapter one - Danger, disease, and the nature of prejudice(s). Advances in Experimental Social Psychology 2012;46:1-54 [FREE Full text]

XSL•FO
RenderX

36. Cramer RJ, Fording RC, Gerstenfeld P, Kehn A, Marsden J, Deitle C, et al. Hate-motivated behavior: impacts, risk factors, and interventions. HealthAffairs Health Policy Brief. 2020. URL: https://tinyurl.com/ydravrd6 [accessed 2024-08-05]
37. Smith A, Anderson M. Social media use in 2018.: Pew Research Center: Internet & Technology; 2018. URL: https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/ [accessed 2024-07-29]
38. Kai-Cheng Y, Ferrara E, Menczer F. Botometer 101: social bot practicum for computational social scientists. J Comput Soc Sci 2022;5(2):1511-1528 [FREE Full text] [doi: 10.1007/s42001-022-00177-5] [Medline: 36035522]
39. Nguyen TT, Merchant JS, Criss S, Makres K, Gowda KN, Mane H, et al. Examining Twitter-derived negative racial sentiment as indicators of cultural racism: observational associations with preterm birth and low birth weight among a multiracial sample of mothers, 2011-2021. J Med Internet Res 2023;25:e44990 [FREE Full text] [doi: 10.2196/44990] [Medline: 37115602]
40. Benton B, Choi JA, Luo Y, Green K. Hate speech spikes on Twitter after Elon Musk acquires the platform. In: School of Communication and Media. United States: Montclair State University; 2022.
41. Hickey D, Schmitz M, Fessler D, Smaldino PE, Muric G, Burghardt K. Auditing Elon Musk's impact on hate speech and bots. 2023 Presented at: Proceedings of the 17th International AAAI Conference on Web and Social Media; June 05-08, 2023; Limassol, Cyprus p. 1133-1137 URL: https://ojs.aaai.org/index.php/ICWSM/article/view/22222
42. Jikeli G, Soemer K. The value of manual annotation in assessing trends of hate speech on social media: was antisemitism on the rise during the tumultuous weeks of Elon Musk's Twitter takeover? J Comput Soc Sc 2023;6:943-971 [FREE Full text] [doi: 10.1007/s42001-023-00219-6]
43. Chang CH, Deshmukh NR, Armsworth PR, Masuda YJ. Environmental users abandoned Twitter after Musk takeover. Trends Ecol Evol 2023;38(10):893-895. [doi: 10.1016/j.tree.2023.07.002] [Medline: 37591718]
44. Ledford H. How Musk's takeover might change Twitter: what researchers think. Nature 2022 Nov;611(7936):436-437. [doi: 10.1038/d41586-022-03552-4] [Medline: 36333606]
45. Stokel-Walker C. Twitter may have lost more than a million users since Elon Musk took over.: MIT Technology Review; 2022. URL: https://www.technologyreview.com/2022/11/03/1062752/twitter-may-have-lost-more-than-a-million-users-since-elon-musk-took-over/ [accessed 2024-07-29]

Abbreviations

API: application programming interface
IRR: incidence rate ratio
WHO: World Health Organization

Edited by A Mavragani; submitted 01.10.23; peer-reviewed by A Hohl, T Wangmo; comments to author 28.02.24; revised version received 25.04.24; accepted 20.06.24; published 09.09.24.

Please cite as:
Wei H, Hswen Y, Merchant JS, Drew LB, Nguyen QC, Yue X, Mane H, Nguyen TT
From Tweets to Streets: Observational Study on the Association Between Twitter Sentiment and Anti-Asian Hate Crimes in New York City from 2019 to 2022
J Med Internet Res 2024;26:e53050
URL: https://www.jmir.org/2024/1/e53050
doi:10.2196/53050
PMID:39250221

©Hanxue Wei, Yulin Hswen, Junaid S Merchant, Laura B Drew, Quynh C Nguyen, Xiaohe Yue, Heran Mane, Thu T Nguyen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality, Usability, and Trust Challenges to Effective Data Use in the Deployment and Use of the Bangladesh Nutrition Information System Dashboard: Qualitative Study

Berhaun Fesshaye[1], MSPH; Shivani Pandya[1], MPH; Lena Kan[1], MSPH; Anna Kalbarczyk[1], MPH, DrPH; Kelsey Alland[1], MSPH; SM Mustafizur Rahman[2], MPhil, MPH, MBBS; Md. Mofijul Islam Bulbul[2], MPH, MBBS; Piyali Mustaphi[3], MS; Muhammad Abu Bakr Siddique[3], MS; Md. Imtiaz Alam Tanim[4], MPH, BPharm; Mridul Chowdhury[4], BSc, MPA; Tajkia Rumman[4], MPH, MBBS; Alain B Labrique[5], MHS, MS, PhD

[1]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[2]National Nutrition Services, Ministry of Health & Family Welfare, Dhaka, Bangladesh

[3]Nutrition Section, UNICEF, Bangladesh Country Office, Dhaka, Bangladesh

[4]mPower Social Enterprises Ltd, Dhaka, Bangladesh

[5]Department of Digital Health & Innovation, World Health Organization, Geneva, Switzerland

Corresponding Author:
Berhaun Fesshaye, MSPH
Department of International Health
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
Baltimore, MD, 21205
United States
Phone: 1 (410) 955 3934
Email: bfessha1@jh.edu

Abstract

Background: Evidence-based decision-making is essential to improve public health benefits and resources, especially in low- and middle-income countries (LMICs), but the mechanisms of its implementation remain less straightforward. The availability of high-quality, reliable, and sufficient data in LMICs can be challenging due to issues such as a lack of human resource capacity and weak digital infrastructure, among others. Health information systems (HISs) have been critical for aggregating and integrating health-related data from different sources to support evidence-based decision-making. Nutrition information systems (NISs), which are nutrition-focused HISs, collect and report on nutrition-related indicators to improve issues related to malnutrition and food security—and can assist in improving populations' nutritional statuses and the integration of nutrition programming into routine health services. Data visualization tools (DVTs) such as dashboards have been recommended to support evidence-based decision-making, leveraging data from HISs or NISs. The use of such DVTs to support decision-making has largely been unexplored within LMIC contexts. In Bangladesh, the Mukto dashboard was developed to display and visualize nutrition-related performance indicators at the national and subnational levels. However, despite this effort, the current use of nutrition data to guide priorities and decisions remains relatively nascent and underused.

Objective: The goal of this study is to better understand how Bangladesh's NIS, including the Mukto dashboard, has been used and areas for improvement to facilitate its use for evidence-based decision-making toward ameliorating nutrition-related service delivery and the health status of communities in Bangladesh.

Methods: Primary data collection was conducted through qualitative semistructured interviews with key policy-level stakeholders (n=24). Key informants were identified through purposive sampling and were asked questions about the experiences and challenges with the NIS and related nutrition dashboards.

Results: Main themes such as trust, data usability, personal power, and data use for decision-making emerged from the data. Trust in both data collection and quality was lacking among many stakeholders. Poor data usability stemmed from unstandardized indicators, irregular data collection, and differences between rural and urban data. Insufficient personal power and staff training coupled with infrastructural challenges can negatively affect data at the input stage. While stakeholders understood and expressed

XSL•FO
RenderX

the importance of evidence-based decision-making, ultimately, they noted that the data were not being used to their maximum potential.

Conclusions: Leveraging DVTs can improve the use of data for evidence-based decision-making, but decision makers must trust that the data are believable, credible, timely, and responsive. The results support the significance of a tailored data ecosystem, which has not reached its full potential in Bangladesh. Recommendations to reach this potential include ensuring a clear intended user base and accountable stakeholders are present. Systems should also have the capacity to ensure data credibility and support ongoing personal power requirements.

(J Med Internet Res 2024;26:e48294) doi:10.2196/48294

KEYWORDS

digital health; nutrition; data for decision-making; health information systems; information system; information systems; LMIC; low- and middle-income countries; nutritional; dashboard; experience; experiences; interview; interviews; service; services; delivery; health care management

Introduction

Background

Health information systems (HISs) play a critical role in aggregating and integrating data from health facilities and communities to improve health services' delivery and health systems' functioning [1]. However, the availability of high-quality, reliable, and sufficient data can be a challenge in low- and middle-income countries (LMICs), which are often characterized by fragmented governance structures, a lack of human resource capacity, and weak digital infrastructure and monitoring systems [2]. Nutrition HISs, also known as nutrition information systems (NISs), focuses specifically on nutrition-related indicators to understand and address challenges related to malnutrition and food security [3]. NISs can play a key role in shaping policy and programs for improving the nutritional statuses of populations and can guide nutrition programming integration into the routinely provided health services [3]. However, the bottlenecks that plague HISs broadly also impede the effective use of NISs. Disjointed efforts to improve the availability, accessibility, quality, and use of health data have limited countries' ability to use data for health decision-making and problem-solving [4].

Evidence-based decision-making is integral to improving public health benefits and resources, and for translating quality evidence into action to achieve health impact, especially in LMICs [5-7]. Key components of adopting an evidence-based public health approach involve the systematic use of data and HISs, community engagement in decision-making, and having trust in the data being used for decision-making [8]. Although evidence-informed decision-making is strongly encouraged by key public health decision makers, the mechanisms of its implementation remain less straightforward and untested [7,9-13].

Data visualization tools (DVTs) have been recommended to support decision-making to better understand data and prioritize the next steps; however, their use remains an unexplored field of research in LMICs [10]. Dashboards, a type of DVT, are used to display relevant summary health performance metrics using visualization techniques (eg, graphs), which provide timely and actionable feedback that informs the decision-making of health providers (eg, health managers and supervisors) [11] and ultimately strengthens the overall quality of care delivered [14]. Dashboards enable the regular review of data by key stakeholders and decision makers in identifying data gaps and following up with corrective actions to improve overall health performance; moreover, they can capture information on data quality and completeness [2]. However, such as the use of HISs broadly—their existence alone is not a silver-bullet solution to improving data monitoring and increasing evidence-based decision-making. Furthermore, having more available data does not necessarily translate into more evidence-based programs and policies in global health [10]. Their impact on improving patient care and health outcomes remains unclear, with only a few studies having investigated how dashboards have been integrated into the decision-making process within primary health care systems [14].

Need for Effective Data Use in Bangladesh to Track Nutrition Indicators

Bangladesh developed its NIS in the early 2010s, which is integrated into the national HIS. Bangladesh has one of the world's largest deployments of the open-source District Health Information Software 2 (DHIS2) and has been able to integrate previously fragmented data systems into a central data repository—making Bangladesh one of the more digitally mature LMICs [15]. As Bangladesh's nutritional programming is provided under 2 distinct directorates under the Ministry of Health and Family Welfare—the Directorate General of Health Services and the Directorate General of Family Planning—the NIS plays a critical role in integrating nutrition-related service data into 1 system which provides an overview of national nutrition progress [16]. The Nutrition Information and Planning Unit (NIPU)—a joint effort from the government of Bangladesh and UNICEF (United Nations Children's Fund)—was established to provide dedicated personnel to work with the NIS and support the use of its data for decision-making. Further, 1 of NIPU's developments was the Mukto dashboard—which is a free, open-access platform to view nutrition-related performance indicators at the national, district, and subdistrict levels [16].

However, despite these efforts and advancements, the research agenda around the effective use of nutrition data to guide key priorities and decisions in Bangladesh remains relatively nascent. There is an increasing need to develop and strengthen

XSL•FO
RenderX

the use of high-quality, reliable, and timely data to accelerate progress in achieving nutrition-related goals. This is especially of importance in Bangladesh, where—despite the substantial progress it has made in food access and nutritional status of the population—malnutrition and food insecurity continue to remain high: over 30% of children aged <5 years are stunted [17]; 22% of children aged <5 years are underweight [17]; over 50% of pregnant women are anemic [18]; and only around 35% of children aged 6-23 months are fed a minimum acceptable diet, highlighting inadequate infant and young child feeding practices in the country [17,18]. Addressing these nutritional needs has direct implications on both maternal and infant mortality as well as their quality of life.

Objective of This Study

This study provides further context into the use of Bangladesh's NIS, including the Mukto dashboard, and identifies areas of improvement so that the NIS can be used effectively and efficiently for evidence-based planning and data-driven actions to improve nutrition service delivery and, more broadly, the nutritional status of communities.

Methods

Study Setting and Context

Key informant interviews (KIIs) were conducted with stakeholders involved with both the NIS and the Mukto dashboard at the national level to understand the NIS landscape and challenges with data use. Participants were identified purposively through consultations with Bangladesh's National Nutrition Services (NNS), UNICEF, and NIPU to identify those in both the government and nongovernment sectors that work with nutrition data and were perceived to use that data.

Study Design and Data Collection

A total of 40 individuals were identified and contacted through email for interviews. Of those, 24 responded and were contacted via phone calls to confirm availability. Once confirmed, individuals were emailed once again with an invitation for participation and a brief description of study objectives. In total, 21 semistructured KIIs took place across 24 respondents. Further, 18 interviews were conducted individually and 3 were conducted in pairs. Paired individuals were from the same organization. Individuals interviewed were national stakeholders from both governmental organizations (n=12) and nongovernmental organizations (NGOs, n=12) including UNICEF, the World Health Organization, and the NNS, among others. Stakeholders were questioned on the perceived achievements of NIS, personal and organizational data sources for decision-making, as well as knowledge about Mukto and other nutrition dashboards.

Due to the COVID-19 pandemic, interviews were conducted virtually through Zoom (Zoom Video Communications, Qumu Corporation) between December 2020 and January 2021. All participants provided oral consent for both participation in the KIIs and recording of the interview. Semistructured interview guides created by study team members were used (Multimedia Appendix 1). KIIs were conducted by 2 facilitators (MIAT and TR), both of whom have postgraduate level training and received training on study design and research ethics. Further, 1 facilitator is the lead of mobile health initiatives at mPower, identifying as male, while the other is a research coordinator, identifying as female. Additionally, 1 facilitator led the KII and the second took notes with the 2 maintaining the same roles throughout each interview. KIIs ranged from 1 to 2 hours, were conducted in Bangla or English based on the interviewee's preference, audio recorded, and subsequently transcribed and translated into English. Study team members based in Bangladesh (MIAT, MC, and TR) completed the transcriptions, translations, and verification of the KIIs. Upon review of initial interview transcripts, the study team determined that data saturation was reached, and additional interviews were not necessary.

Analysis

KII data were coded and analyzed using a thematic content analysis approach. Data analysis was conducted by a team of 4 (BF, LK, AK, and SP). An initial read-through of transcripts was performed to identify and discuss emerging themes. The creation of the codebook was guided by the PATH (Program for Appropriate Technology in Health) theory of change for accelerating data use in health systems (Figure 1; Multimedia Appendix 2) [19,20]. The theory of change framework was developed by PATH for visualizing the process of strengthening data systems and increasing data use, to improve health systems, for over 10 years. A formula that encapsulates the theory of change framework is "data production" plus "information use" multiplied by "levers and accelerators," which leads to a culture of data use that allows for better health system performance [20]. Levers are the components such as workforce, leadership, governance, and infrastructure that set up an environment conducive to effective data use with accelerators increasing the impact of each lever. PATH has used the framework to create an investment roadmap for Tanzania to use data for improving the country's health system and reach middle-income status [19]. The framework was chosen as a guide for qualitative analysis of the KIIs as it encapsulates the components of a robust data ecosystem. The codebook was iteratively developed to leverage the 7 components of the World Health Organization–International Telecommunication Union National eHealth Strategy within the PATH theory of change. The codebook was then adapted as the team of researchers read through the transcripts and identified emerging themes to include in the final codebook. Interviews were divided among the team for analysis, with illustrative quotes extracted for themes. Data were coded and managed in Excel (Microsoft Corp).

XSL•FO
RenderX

Figure 1. The PATH (Program for Appropriate Technology in Health) theory of change for accelerating data use in health systems. This theory explains how the data use cycle, made up of information use and data production, is influenced by levers such as infrastructure and workforce. These levers can act as accelerators for the data use process. All of these factors working together can contribute to improved health system performance, which in turn, leads to improved health outcomes. ITU: International Telecommunication Union; WHO: World Health Organization. Source: "Theory of Change for the Data Use Partnership [20] which is published under Creative Commons Attribution 4.0 International License [21] .

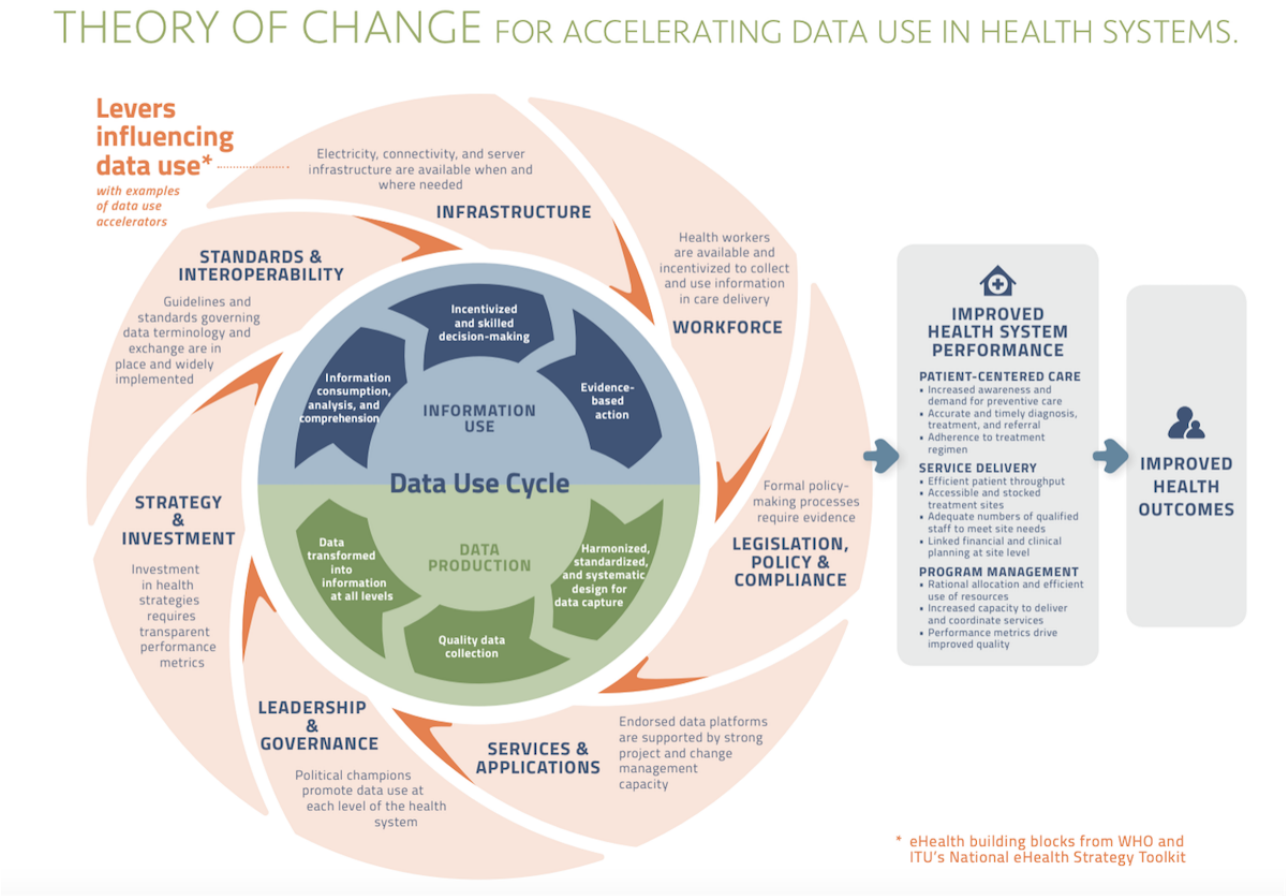

Ethical Considerations

This study received approval from the Johns Hopkins University Bloomberg School of Public Health Institutional Review Board (14637) and received a nonhuman subjects research determination as defined by US Department of Health and Human Services regulations 45 Code of Federal Regulations 46.102. All participants provided oral consent before interviews and were allowed to opt out at any time. The presented data have been deidentified. No compensation was provided to participants.

Results

Overview

The KIIs (n=24) provided important insights into how Bangladesh's NIS is being used, including the Mukto dashboard (Figure 2 [22]), and the challenges that inhibited its use. Following an initial read-through of all interviews, high-level categories, known as levers, from the PATH framework were used to guide the creation of relevant subcategories. The predetermined levers include standards and interoperability, infrastructure, workforce, legislation, policy and guidance, services and applications, leadership and governance, and strategy and investment [19]. From these levers, this study team created subcategories that included indicator standardization, internet connectivity, and capacity development, among others. A summary of results, across all interviews, mapped according to PATH framework categories and our own subcategories are presented (Multimedia Appendix 3 [19]).

Additional cross-cutting themes emerged across these categories and subcategories including trust, data usability, personal power, and data use for decision-making. These themes provide further contextualization of the factors that support or inhibit the use of NIS by key stakeholders, and considerations for improvement to encourage its use and uptake.

XSL•FO
RenderX

Figure 2. Nutritional Information System Mukto dashboard. A web-based platform for easy viewing of nutrition-related performance indicators at the national, district, and subdistrict levels in Bangladesh. FWC: Family welfare centre; IMCI: Integrated management of childhood illness; SAM:Severe acute malnutrition; CCTN: Comprehensive competency training on nutrition; CBT: Cash based transfers; IFA-Iron and Folic acid; IYCF -Infant and young child feeding; MICS-Multiple Indicator Cluster Survey; WHA-World Health Assembly; BDHS-Bangladesh Demographic and Health Survey.

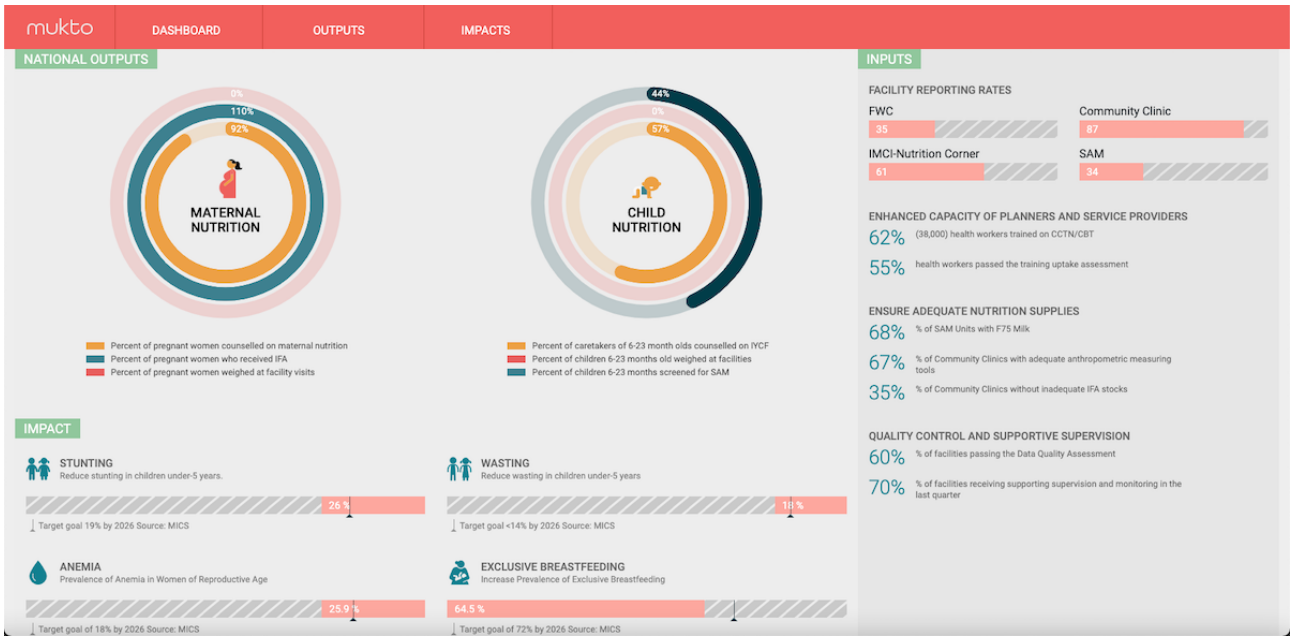

Trust

Trust was perceived as lacking in both the nutrition data sources and quality. Participants saw trust as an integral part of decision-making and without it, efforts at all levels—from data collected in the field to the policy level—contain faults. All interviewed stakeholders, except 1, commented on data quality and its shortcomings. Additionally, 3 government stakeholders and 5 NGO stakeholders expressed that data quality stems from the input stage and that any compromise there carries on throughout the lifecycle of the data. However, data input is not simply the responsibility of community health workers (CHWs), as clear and standard indicators must be well established beforehand. Manual processes for data input and management also affect quality as errors can go unnoticed.

> *The data that is coming from [the clinic level] is very weak, so if you think properly, you will see that the fault happened at the beginning.* [NGO stakeholder]

> *At the root, [data quality has been compromised] where the data is entered. Before that, the right indicators should be there. The design is important. If the design is faulty then every step after that will be faulty.* [NGO stakeholder]

Further, 2 government stakeholders stressed that data must be of high quality for use at the policy level. Without trust in where the data are coming from and what they comprise, there will be limited use of data to inform decision-making. Ambiguous and confusing data sources will only further diminish the credibility and assurance of the data ecosystem and impede data use and flow.

> *We just see the data, but to work on a policy level we have to be careful about the authenticity of the data, because you cannot provide misleading data on a national level.* [Government stakeholder]

However, 1 nongovernment stakeholder remarked that processes for data validation and avoiding duplicates require significant personal power especially with data coming from different sources, as is the case in Bangladesh. Although trust in data quality was lacking for many interviewees, 1 government stakeholder reported that their organization implements intensive data audits to ensure high-quality data:

> *There are monitoring officers…they monitor everything on their visit. They have a lot of mechanisms, and we visit them from the headquarter[s] to ensure quality…. we cross-check the data of the register with the MIS, and then the MIS with entered data. We cross-check the linkages among them. If there is any mistake, the links will not work. So everything is checked through a mechanism.* [Government stakeholder]

Data Usability

Data usability was another issue identified by stakeholders. Indicators are not currently standardized across all organizations, regions, and data sources. For example, 1 nongovernment stakeholder stated that while their organization is aware of 64 standard indicators, they only focus on 25 of them. There is also a lack of standardization within health facilities and between government organizations. Gaps emerge in data collected if 1 health facility focuses on more indicators than another or if government organizations use inconsistent indicator definitions. Stakeholders noted that indicators should be standardized for maximum usability for decision-making, but this is a time-consuming process that requires clear leadership. Further, 1 NGO director noted that this leadership would ideally come from the government, but discussions around the standardization process have not yet occurred. Moreover, simply identifying indicators is not enough as there must be money and action to put those indicators to use.

XSL•FO
RenderX

> *Because it is not about just including indicators; a lot of money is needed to publish them as well. If two indicators are included in a list, it needs to be published, go to the facilities, and training should be provided again. Where will this money come from? So I would not say it is not harmonized at all. It was harmonized but it needs to be reviewed.* [Nongovernment stakeholder]

> *I remember [one] nutrition specific indicator, a small change was needed but we had to keep on telling them to change it for many days.* [Nongovernment stakeholder]

Another issue affecting data usability is the irregularity of data collection, which leads to an unclear picture of the country's progress. Continuous data monitoring and review are critical to track the progress of data collection and implement iterative data quality improvements.

> *The problem is that [the data] is long term. One has to wait 4-5 years. In that case, if a sudden recommendation is required, then that old data has to be used from long ago, for example from 2014.* [Government stakeholder]

> *In many cases, I have observed that if we cannot follow up properly the data will not reach everywhere. So, you will find out that even after updating, a program is being carried out with the old data.* [NGO stakeholder]

Stakeholders also reported that there are usability differences between urban and rural data given that rural NISs are more established than urban systems. Health and nutrition efforts in Bangladesh have historically been focused on rural areas, leaving urban areas with inequitable access to these vital services [23]. Unlike rural areas, urban areas are structurally complicated, with less accountability and government support. While rural nutrition systems have been operating under the guidance of NNS and the Ministry of Health and Family Welfare, urban systems do not do so, with their reporting being voluntary and largely completed with the support of NGOs. Another discrepancy described between urban and rural areas is the lack of indicator standardization. Differences in rural and urban areas were a cause of concern for 13—over half (54%)—of the 24 interviewed stakeholders, from both the government and NGOs. Insufficient data negatively affects policy and investment decisions and diminishes the ability to understand the complete picture of the country's nutrition status. Moreover, as Bangladesh experiences rapid urbanization, lacking sufficient urban data to improve health and nutrition could diminish national progress [23]. Further, 1 of the cited reasons for the lag in urban data is the dependence on NGO support without assistance from government entities and the lack of accountability that comes with this.

> *No one has told [urban areas] yet that the indicators are under the national system, and it is happening in rural, not in urban. It has to be done. It has not been communicated with the urban system on behalf of the government counterpart. So they don't feel obliged to the nationally harmonized indicator-based reporting. So they only do it when someone tells them or backstopping them.* [Nongovernment stakeholder]

Personal Power

Stakeholders reported that data collection, management, analysis, and monitoring are frequently executed manually, requiring significant personal power. Additionally, high staff turnover leads to the constant need for training. Reasons for staff turnover include transfer to other cadres, staff leaving after the completion of a project, and term-based NNS positions that do not set up a permanent workforce.

Many stakeholders also suggested that regular training of all relevant persons, such as CHWs, data analysts, and policy makers, should occur to help mitigate errors during data collection, management, and analysis. Currently, although there are trainers deployed at the field level and national levels—for government staff—there remain some gaps as government and NGOs variably provide such training to their respective staff. Further, 1 nongovernment stakeholder provided an example of a gap, which is that their organization provides less emphasis on nutrition training compared to other areas such as child health services and immunization.

> *Some technical systematic errors and [field workers] have to be motivated with one thing that the data entry is important. [Field workers] think of it as extra work. This thing, this reporting is often seen in a neglecting manner, NIS is not taken much seriously. I am doing a work, but if I do not report that work, it does [not] hold any value. Does it? I have to make a report. So I think we should put more stress and be aware of this reporting system and an information system above that. And those who work at the policy level, if they take this thing more seriously, then I think it is good.* [Government stakeholder]

> *But the way it has been presented, it seems that only you need the specialized kind of training or skill that only the NIPU team can do.* [Nongovernment stakeholder]

Infrastructural challenges, including poor internet connection at the field level, also exacerbate human resource issues. Internet issues were described as a barrier to both staff training and capacity development. Poor internet connection can interfere with data input, leading to poor quality or incorrect data reporting. Further, connectivity issues may affect staff workload and motivation.

> *However, still we face some problems with the network coverage, maybe he/she does not get the network coverage in office. But if he/she gets the network out of the office or at his home. For this issue, they take the register book with them at his home and work there. But he/she should not do this, but though he/she has the responsibility, he/she has to do it at any cost.* [Government stakeholder]

Data for Decision-Making

The topic of data use for decision-making, and the ease of access to that data, was a central theme throughout the interviews.

XSL•FO
RenderX

Stakeholders expressed the importance of evidence-based decisions and the necessity of timely and accurate data. Most participants reported receiving their data for decision-making directly from reports rather than from dashboards. These data were used to inform policy development, provide support, strengthen the understanding of data systems, and assess the impact data has on the government.

> *However, we use these data in many cases for example, for policy advocacy, trend analyzing, giving a better understanding to the government, showing them the difference, providing support to the places that have no support apart from additional supervision. We also use the data to observe if any government impact is being created for the input we are giving in the technical module and for the additional training that we provided.* [Government stakeholder]

The amount of data used for decision-making differs among stakeholders with some reporting that a vast amount of data is available, but not always being used. Available data sources also vary among organizations. For example, 2 government managerial stakeholders and 1 NGO director noted that their organization's managers are aware of dashboards but do not use them as a primary data source. Further, 3 government stakeholders reported that data for decision-making is received directly through reports or their own personal networks, while 2 other government stakeholders receive their data from the national systems including the DHIS2, Directorate General of Family Planning Management Information System, and the Bangladesh Demographic and Health Survey 2014. Respondents expressed that data must be accurate for decision-making, which underscores the importance of data trust and data quality.

> *We can only observe the nutrition related indicators that have been collected on DHIS2 platform by MIS unit whereas they grant us access to those data. But we cannot [access] other nutrition indicators such as adolescent nutrition, elderly nutrition, or emergency nutrition on DHIS2 managed by MIS unit of the Ministry of Health.* [Government stakeholder]

> *If I say straightway about DHIS-2 that I want to see the information there about my division, district and upazila, the result will be shown there. I can get the information of four to five nutrition indicators which are very much important and to run a system or to see what is happening and what is the status of the nutrition in the country, it is easy to understand.* [NGO stakeholder]

Additionally, 1 government stakeholder noted that when making national-level decisions, the local context is not always considered. A nongovernment stakeholder stated that making decisions based on data at every level, such as the district and subdistrict, remains a challenge to be addressed. However, a different government stakeholder expressed that decision-making based on data does indeed occur at the subnational level.

> *I think these few achievements have helped to create a better process where a manager of either an event central level or upazila level can take some decisions based on the information.* [Government stakeholder]

A noted barrier to evidence-based decisions and dashboard use is personal power and staff motivation. Without sufficient human resources and incentives to use dashboards, the introduction of data dashboards may in turn increase the data and amount of work burden among existing CHWs.

> *So nutrition is one of the priorities of my activities and I will discuss about the dashboard with everyone and get solutions for it. If they provide us any decision regarding this, I think we can work so well from the national level. Especially managers, we send the program and data to them first and tell them about the situation. Then they consider it as a burden and think when will they do this pile of work.* [Government stakeholder]

Stakeholders were aware of health and nutrition dashboards, and while most did not use these for their decision-making process, some discussed the value of dashboards. Further, 1 government stakeholder expressed that dashboards allow for comparison between districts and facilities, which in turn facilitates decision-making and behavior change.

> *We almost get the national data from there [DGHS dashboard] or we can learn how other facilities of the health ministry are working. So it helps us to decide what we will do next and how we will do it. This sometimes helps in our decision-making.* [Government stakeholder]

Discussion

Principal Findings

This study sought to identify and understand how Bangladesh's NIS can be improved to foster effective data use for increased coverage of nutrition services. Using PATH's theory for accelerating data use (Figure 1), we organized our results to note gaps and areas for improvement and increase in data use for decision-making. The KIIs conducted with key government and NGO nutrition stakeholders identified several facilitators and challenges toward data use—largely centered around lack of trust in data, concerns around data usability, lack of personal power, and how that ultimately impacts data use for decision-making. We found that trust in the quality of nutrition data is lacking among stakeholders, mainly due to errors occurring during the data collection phase. These findings are similar to those conducted in Kenya and Malawi, which showed that routine data collected at the field level is regularly perceived as poor quality [24-26]. Possible reasons for poor quality data in other settings ranged from unstandardized data collection to lack of training [26], which echo the concerns of stakeholders in Bangladesh.

Perceptions of data usability were influenced by data quality and existing discrepancies between rural and urban data. Bangladesh's subnational discrepancies in data have been perpetuated by the lack of accountability and support for urban areas coupled with a historic focus on nutrition services in rural areas [23]. Another usability issue noted is unstandardized

XSL•FO
RenderX

nutrition indicators which lead to inconsistencies during data collection. Personal power was also a central theme as stakeholders reported that manual data management is hindered by high staff turnover and that staff training should take place regularly. Other studies in LMICs including Bangladesh, have also shown that field-level staff turnover is high while motivation is low, negatively affecting data quality and staff performance [27,28]. Poor internet connection at the field level is a barrier to personal power as it interferes with data collection, staff training, and capacity development. The importance of using data for decision-making was stressed by stakeholders, and despite knowledge of dashboards, such as Mukto, decision makers tend to receive data from their own preferred sources. A study conducted in Mozambique found that dashboards were viewed as an externally (nongovernment) developed strategy and garnered minimal interest from stakeholders, leading to the discontinuation of their use in this setting [25]. While our results did not reveal explicit reasons for the limited use of dashboards in the decision-making process, other settings may reveal possible causes. Addressing these possible causes will allow nutrition indicator dashboards, such as the Mukto dashboard, to reach their full potential.

Data Ecosystems for Effective Decision-Making

Governments globally are increasingly using data to support their functions and decision-making capabilities [29]. Data systems are complex, and decision makers are often far removed from the entire data lifecycle—from data capture, compilation, distillation, and processing to visualization and reporting. While DVTs can be leveraged to improve the use of data for global health decision-making [10], we found that decision makers first have to trust that the data are believable, credible, timely, and responsive in order to use it. Generating this trust starts with having clear governance, leadership, and support for those who are at the genesis of the data ultimately presented in the summary program performance reports and dashboards. Trust also arises from relevant persons having salient roles in the construction and upkeep of the systems' housing data [30]. These roles, contributing to a sense of ownership, are central to the success and longevity of the systems in place [30,31]. Ownership can also contribute to awareness of the data content, allowing for increased monitoring of the setting's progress [30]. Continuous monitoring of progress and areas of improvement requires an ecosystem of reliable, timely, trusted, and actionable data that may arise from the use of digital visualization tools such as dashboards.

Importantly, data ecosystems should be localized and tailored to any setting or context and present key public health decision makers with actionable messages to optimize decision-making and translate knowledge into health practice, with the ultimate goal of improving health outcomes and reducing health inequalities. Ideally, all public health practitioners should incorporate evidence-informed decision-making during program implementation, policy development, and progress evaluation [12]. The data use ecosystem should foster the marriage between evidence-based public health interventions with real-world contexts and settings. The improvement in data quality and use will help facilitate the translation of findings into impactful policy recommendations and contribute to the improvement of overall health and well-being [8,32,33].

Results from this study indicate that the Bangladesh nutrition data ecosystem, which includes the NIS and the Mukto dashboard, has not yet reached its full potential. Gaps exist in nutrition data and decision makers do not feel confident in data collection and quality. Continued improvement in nutrition-related health outcomes will require tailoring of the Bangladesh nutrition data ecosystem along with increased ownership among stakeholders.

Power, Trust, and Worker Motivation

Improving data quality, trust, and use cannot occur without establishing an environment that fosters worker motivation. CHWs, who are the cornerstone of primary health care systems globally by providing health services to hard-to-reach and rural communities, simultaneously play an important role in collecting health services (including nutrition-based services) data [27,34]. Bangladesh's CHW program faces significant levels of attrition and high workloads, with some CHWs reporting discontent with their positions [27]. In our study, stakeholders expressed a lack of trust in data collected by CHWs and reported that field-level staff believed many tasks were extra work. However, the unrealistic workloads regularly placed on field staff [27] are likely the cause of the negative sentiments reported. Negative stakeholder perceptions of CHWs may be detrimental to the overall motivation and productivity of staff, further decreasing the quality of and trust in data. Motivation and satisfaction among CHWs can be undermined by both external factors such as program structure and provision of training, as well as internal factors such as satisfaction with work and CHWs' perception of relevant stakeholders [28,34].

It is critical to also consider how gender and power dynamics impact relationships between field-level and high-level staff. Gender roles can shape CHW experiences and relationships with decision makers [35]. A majority of the global health workforce, which includes CHWs, are women with relatively low socioeconomic status and education, while only 30% of global health leadership roles are occupied by women [35,36]. Our interviewees expressed concern about the high worker turnover, yet specific reasons for this turnover were largely unspecified. Previous research has shown that reasons for attrition among women CHWs in Bangladesh include competing family commitments, disapproval from family members, and insufficient income [35,37]. Given Bangladesh's reliance on CHWs for field-level data collection and service provision, a gender lens is critical for identifying appropriate solutions.

Some solutions that have been shown to increase health worker motivation include reasonable workloads, positive relationships with decision makers, sufficient monetary compensation, and recognition of efforts [27,28]. Without taking worker motivation and satisfaction into account, it is difficult to ensure high-quality data at the field level that decision makers will trust. Necessary improvements in worker motivation should come from those in leadership positions who have the power to affect change in work environments. Responsibility to increase data quality cannot fall solely on the backs of field-level workers as pressure to sustain large workloads and insufficient support from

XSL•FO
RenderX

leadership are factors that decrease motivation and accuracy [27,28].

Innovative Data Systems and System Improvements

The number of interactive dashboards and other DVTs, especially in the nutrition sector, has increased in recent years [38]. This is partly due to increased investments in nutrition that have increased the availability of data for decision makers and researchers [39]. However, the "build it and they will come" approach has been shown not to work [39] vis-a-vis health sector dashboards, leading to a larger discourse around better resource use and user-centered design of program monitoring systems [40]. We found that while government and nongovernment stakeholders were aware of dashboards, most did not use them in their decision-making process. Furthermore, the use of data dashboards does not guarantee improved data use and quality without sufficient user data literacy, important research considerations, and an established environment of trust among key decision makers. Moreover, the selection of key dashboard indicators to assess primary health care performance also depends on the availability of data, which often is lacking in LMICs [11]. Data ecosystems, which include data visualizations and dashboards, must have the capacity for constant improvements and iterative changes. Users must be motivated by the innovativeness and benefit of these tools to confidently and effectively use them in the decision-making process. Only with the effective integration of dashboards in health decision-making along with concurrent supporting policies, infrastructure, trust in data, empowered personnel, and an environment conducive to data use can the momentum of the data revolution be followed.

Limitations

This study is not without limitations, Due to this study's design, results may be influenced by social desirability bias. Additionally, due to the sample size and methods, results are likely not generalizable to other settings.

Conclusion

We believe this study is valuable as it provides insights into stakeholder experience and suggestions for the effective use of nutrition data systems in Bangladesh. Although the creation of data systems and dashboards is important to increase the transparency of data, the mere creation of these systems may not solve Bangladesh's NIS performance monitoring needs, as a clear intended user base, and accountable stakeholders must be present at the onset. Systems should be in place to ensure authenticity and credibility at the point of generation, with steps being taken to correct erroneous data. Data collection, management, and monitoring require robust, permanent human resources and strong infrastructure, as well as strong government and community engagement at each step of the data generation process.

Although these recommendations are not necessarily a prescription for success, increasing the presence of credible data ecosystems can enhance decision-making efforts. The culmination of many factors, such as appropriate infrastructure and trust, along with effective data use are necessary to keep up with the data revolution. While Bangladesh has made considerable progress in the nutrition field, ensuring that all future policy and governance decisions are informed by data is vital for continued improvement.

Acknowledgments

The authors would like to thank the participants for their time and their insights regarding Bangladesh's Nutrition Information System.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

MC, MIAT, TR, ABL, KA, and AK conceptualized this study. MC, MIAT, and TR managed and conducted data collection. BF, LK, SP, and AK conducted data analysis. BF, LK, SP, AK, and ABL wrote the first draft of this paper. All authors edited and reviewed this final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured key informant interview guide.
[DOCX File , 24 KB - jmir_v26i1e48294_app1.docx]

Multimedia Appendix 2
Qualitative codebook.
[DOCX File , 16 KB - jmir_v26i1e48294_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Contextualizing key informant interview results with the PATH (Program for Appropriate Technology in Health) framework.
[DOCX File , 18 KB - jmir_v26i1e48294_app3.docx]

References

1. AbouZahr C, Boerma T. Health information systems: the foundations of public health. Bull World Health Organ 2005;83(8):578-583 [FREE Full text] [Medline: 16184276]
2. Etamesor S, Ottih C, Salihu IN, Okpani AI. Data for decision making: using a dashboard to strengthen routine immunisation in Nigeria. BMJ Glob Health 2018;3(5):e000807 [FREE Full text] [doi: 10.1136/bmjgh-2018-000807] [Medline: 30294456]
3. Salam RA, Das JK, Bhutta ZA. Integrating nutrition into health systems: What the evidence advocates. Matern Child Nutr 2019;15 Suppl 1(Suppl 1):e12738 [FREE Full text] [doi: 10.1111/mcn.12738] [Medline: 30748112]
4. Lippeveld T. Routine health facility and community information systems: creating an information use culture. Glob Health Sci Pract 2017;5(3):338-340 [FREE Full text] [doi: 10.9745/GHSP-D-17-00319] [Medline: 28963169]
5. Aryeetey R, Holdsworth M, Taljaard C, Hounkpatin WA, Colecraft E, Lachat C, et al. Evidence-informed decision making for nutrition: African experiences and way forward. Proc Nutr Soc 2017;76(4):589-596 [FREE Full text] [doi: 10.1017/S0029665117001082] [Medline: 28803565]
6. Majdzadeh R, Yazdizadeh B, Nedjat S, Gholami J, Ahghari S. Strengthening evidence-based decision-making: is it possible without improving health system stewardship? Health Policy Plan 2012;27(6):499-504. [doi: 10.1093/heapol/czr072] [Medline: 22027555]
7. van der Graaf P, Cheetham M, McCabe K, Rushmer R. Localising and tailoring research evidence helps public health decision making. Health Info Libr J 2018;35(3):202-212 [FREE Full text] [doi: 10.1111/hir.12219] [Medline: 29808957]
8. Shafaghat T, Bastani P, Nasab MHI, Bahrami MA, Montazer MRA, Zarchi MKR, et al. A framework of evidence-based decision-making in health system management: a best-fit framework synthesis. Arch Public Health 2022;80(1):96 [FREE Full text] [doi: 10.1186/s13690-022-00843-0] [Medline: 35351210]
9. Orton L, Lloyd-Williams F, Taylor-Robinson D, O'Flaherty M, Capewell S. The use of research evidence in public health decision making processes: systematic review. PLoS One 2011;6(7):e21704 [FREE Full text] [doi: 10.1371/journal.pone.0021704] [Medline: 21818262]
10. Aung T, Niyeha D, Heidkamp R. Leveraging data visualization to improve the use of data for global health decision-making. J Glob Health 2019;9(2):020319 [FREE Full text] [doi: 10.7189/jogh.09.020319] [Medline: 31893028]
11. Veillard J, Cowling K, Bitton A, Ratcliffe H, Kimball M, Barkley S, et al. Better measurement for performance improvement in low- and middle-income countries: the primary health care performance initiative (PHCPI) experience of conceptual framework development and indicator selection. Milbank Q 2017;95(4):836-883 [FREE Full text] [doi: 10.1111/1468-0009.12301] [Medline: 29226448]
12. Brownson RC, Fielding JE, Maylahn CM. Evidence-based public health: a fundamental concept for public health practice. Annu Rev Public Health 2009;30:175-201. [doi: 10.1146/annurev.publhealth.031308.100134] [Medline: 19296775]
13. Bédard PO, Ouimet M. Persistent misunderstandings about evidence-based (sorry: informed!) policy-making. Arch Public Health 2016;74:31 [FREE Full text] [doi: 10.1186/s13690-016-0142-z] [Medline: 27441090]
14. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform 2015;84(2):87-100 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.10.001] [Medline: 25453274]
15. Khan MAH, Cruz VO, Azad AK. Bangladesh's digital health journey: reflections on a decade of quiet revolution. WHO South East Asia J Public Health 2019;8(2):71-76. [doi: 10.4103/2224-3151.264849] [Medline: 31441440]
16. Guidance document on integration of priority nutrition indicators into existing systems and consolidation at national level. Bangladesh National Nutrition Council. 2017. URL: https://bnnc.portal.gov.bd/sites/default/files/files/bnnc.portal.gov.bd/publications/e210466e_2d90_42f3_88c1_e7d4e42587b0/2021-02-08-13-43-39f71014c7ff7e94501768b746f46a68.pdf [accessed 2022-07-29]
17. Bangladesh demographic and health survey 2017-18. The DHS Program.: NIPORT, ICF; 2020. URL: https://microdata.worldbank.org/index.php/catalog/3825 [accessed 2022-07-21]
18. Bangladesh: nutrition profile. Bangladesh: USAID; 2021. URL: https://2012-2017.usaid.gov/what-we-do/global-health/nutrition/countries/bangladesh-nutrition-profile [accessed 2022-07-21]
19. Arenth B, Bennett A, Bernadotte C. Definining and building a data use culture. USA: PATH; 2017. URL: https://media.path.org/documents/PATH_Data_Use_Culture_Whitepaper_R2_DYmiuzp.pdf [accessed 2022-07-26]
20. Data use partnership: theory of change. Seattle: PATH; 2016. URL: https://media.path.org/documents/DHS_DUP_Theory_of_Change_rpt.pdf [accessed 2022-07-28]
21. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-08-02]]
22. Mukto. URL: https://mukto.nnsop.gov.bd/ [accessed 2023-11-20]
23. Govindaraj R, Raju D, Secci F, Chowdhury S, Frere JJ. Health and nutrition in urban Bangladesh: social determinants and health sector governance. Findings on urban health sector governance in Bangladesh.: World Bank Group; 2018. URL:

XSL•FO
RenderX

https://documents1.worldbank.org/curated/en/399461514540007634/pdf/122330-PUB-PUBLIC-PUBDATE-12-28-17.pdf [accessed 2022-07-22]
24. Admon AJ, Bazile J, Makungwa H, Chingoli MA, Hirschhorn LR, Peckarsky M, et al. Assessing and improving data quality from community health workers: a successful intervention in Neno, Malawi. Public Health Action 2013;3(1):56-59 [FREE Full text] [doi: 10.5588/pha.12.0071] [Medline: 25767750]
25. Gimbel S, Mwanza M, Nisingizwe MP, Michel C, Hirschhorn L, AHI PHIT Partnership Collaborative. Improving data quality across 3 sub-Saharan African countries using the consolidated framework for implementation research (CFIR): results from the African health initiative. BMC Health Serv Res 2017;17(Suppl 3):828 [FREE Full text] [doi: 10.1186/s12913-017-2660-y] [Medline: 29297401]
26. Regeru R, Chikaphupha K, Bruce Kumar M, Otiso L, Taegtmeyer M. 'Do you trust those data?'-a mixed-methods study assessing the quality of data reported by community health workers in Kenya and Malawi. Health Policy Plan 2020;35(3):334-345 [FREE Full text] [doi: 10.1093/heapol/czz163] [Medline: 31977014]
27. Roy S, Pandya S, Hossain MI, Abuya T, Warren CE, Mitra P, et al. Beyond institutionalization: planning for sustained investments in training, supervision, and support of community health worker programs in Bangladesh. Glob Health Sci Pract 2021;9(4):765-776 [FREE Full text] [doi: 10.9745/GHSP-D-21-00156] [Medline: 34933974]
28. Olaniran A, Madaj B, Bar-Zeev S, Banke-Thomas A, van den Broek N. Factors influencing motivation and job satisfaction of community health workers in Africa and Asia-a multi-country study. Int J Health Plann Manage 2022;37(1):112-132. [doi: 10.1002/hpm.3319] [Medline: 34476842]
29. Matheus R, Janssen M, Maheshwari D. Data science empowering the public: data-driven dashboards for transparent and accountable decision-making in smart cities. Gov Inf Q 2020;37(3):101284. [doi: 10.1016/j.giq.2018.01.006]
30. Watson-Grant S, Xiong K, Thomas JC. Achieving sustainability in health information systems: a field tested measure of country ownership. Global Health 2017;13(1):36 [FREE Full text] [doi: 10.1186/s12992-017-0258-0] [Medline: 28646878]
31. Watson-Grant S, Xiong K, Thomas JC. Country ownership in international development: toward a working definition. Working Paper. 2016. URL: https://www.measureevaluation.org/resources/publications/wp-16-164/at_download/document#:~:text=Country%20ownership%20in%20development%20aid,lead%20to%20more%20successful%20outcomes [accessed 2022-08-05]
32. Oxman AD, Lavis JN, Lewin S, Fretheim A. Support tools for evidence-informed health policymaking (STP) 1: what is evidence-informed policymaking? Health Res Policy Syst 2009;7 Suppl 1(Suppl 1):S1 [FREE Full text] [doi: 10.1186/1478-4505-7-S1-S1] [Medline: 20018099]
33. Imani-Nasab MH, Yazdizadeh B, Salehi M, Seyedin H, Majdzadeh R. Validity and reliability of the evidence utilisation in policymaking measurement tool (EUPMT). Health Res Policy Syst 2017;15(1):66 [FREE Full text] [doi: 10.1186/s12961-017-0232-6] [Medline: 28778204]
34. Gottert A, McClair TL, Hossain S, Dakouo SP, Abuya T, Kirk K, et al. Development and validation of a multi-dimensional scale to assess community health worker motivation. J Glob Health 2021;11:07008 [FREE Full text] [doi: 10.7189/jogh.11.07008] [Medline: 33763222]
35. Steege R, Taegtmeyer M, McCollum R, Hawkins K, Ormel H, Kok M, et al. How do gender relations affect the working lives of close to community health service providers? empirical research, a review and conceptual framework. Soc Sci Med 2018;209:1-13 [FREE Full text] [doi: 10.1016/j.socscimed.2018.05.002] [Medline: 29777956]
36. Batson A, Gupta G, Barry M. More women must lead in global health: a focus on strategies to empower women leaders and advance gender equality. Ann Glob Health 2021;87(1):67 [FREE Full text] [doi: 10.5334/aogh.3213] [Medline: 34307070]
37. Alam K, Oliveras E. Retention of female volunteer community health workers in Dhaka urban slums: a prospective cohort study. Hum Resour Health 2014;12:29 [FREE Full text] [doi: 10.1186/1478-4491-12-29] [Medline: 24886046]
38. Zhou B, Liang S, Monahan KM, Singh GM, Simpson RB, Reedy J, et al. Food and nutrition systems dashboards: a systematic review. Adv Nutr 2022;13(3):748-757 [FREE Full text] [doi: 10.1093/advances/nmac022] [Medline: 35254406]
39. Manorat R, Becker L, Flory A. Global data visualization tools to empower decision-making in nutrition. DNM. 2019. URL: https://cms.sightandlife.org/wp-content/uploads/2023/03/Celebrating-our-35th-anniversary-sightandlife.pdf#page=114 [accessed 2022-07-18]
40. Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R, et al. Best practices in scaling digital health in low and middle income countries. Global Health 2018;14(1):103 [FREE Full text] [doi: 10.1186/s12992-018-0424-z] [Medline: 30390686]

Abbreviations

CHW: community health worker
DGHS: Directorate General of Health Services
DHIS2: District Health Information Software 2
DVT: data visualization tool
HIS: health information system

XSL•FO
RenderX

KII: key informant interview
LMIC: low- and middle-income country
MIS: management information system
NGO: nongovernmental organization
NIPU: Nutrition Information and Planning Unit
NIS: nutrition information system
NNS: National Nutrition Services
PATH: Program for Appropriate Technology in Health
UNICEF: United Nations Children's Fund

Edited by T de Azevedo Cardoso, S Ma; submitted 09.05.23; peer-reviewed by I Wilson, N Wang, D Chrimes; comments to author 02.11.23; revised version received 21.11.23; accepted 15.05.24; published 30.09.24.

Please cite as:
Fesshaye B, Pandya S, Kan L, Kalbarczyk A, Alland K, Rahman SMM, Bulbul MMI, Mustaphi P, Siddique MAB, Tanim MIA, Chowdhury M, Rumman T, Labrique AB
Quality, Usability, and Trust Challenges to Effective Data Use in the Deployment and Use of the Bangladesh Nutrition Information System Dashboard: Qualitative Study
J Med Internet Res 2024;26:e48294
URL: https://www.jmir.org/2024/1/e48294
doi:10.2196/48294
PMID:39348172

©Berhaun Fesshaye, Shivani Pandya, Lena Kan, Anna Kalbarczyk, Kelsey Alland, SM Mustafizur Rahman, Md. Mofijul Islam Bulbul, Piyali Mustaphi, Muhammad Abu Bakr Siddique, Md. Imtiaz Alam Tanim, Mridul Chowdhury, Tajkia Rumman, Alain B Labrique. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Response to Federal Electronic Cigarette Regulations Analyzed Using Social Media Data Through Natural Language Processing: Topic Modeling Study

Shuo-Yu Lin[1], PhD; Sahithi Kiran Tulabandu[2], MS, PharmD; J Randy Koch[3], PhD; Rashelle Hayes[4], PhD; Andrew Barnes[5], PhD; Hemant Purohit[6], PhD; Songqing Chen[7], PhD; Bo Han[7], PhD; Hong Xue[1], PhD

[1]Department of Health Administration and Policy, George Mason University, Fairfax, VA, United States
[2]WVU Health Affairs Institute, Morgantown, WV, United States
[3]Department of Psychology, Virginia Commonwealth University, Richmond, VA, United States
[4]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, United States
[5]Department of Health Behavior and Policy, Virginia Commonwealth University, Richmond, VA, United States
[6]Department of Information Sciences and Technology, George Mason University, Fairfax, VA, United States
[7]Department of Computer Science, George Mason University, Fairfax, VA, United States

Corresponding Author:
Hong Xue, PhD
Department of Health Administration and Policy
George Mason University
4400 University Dr
Fairfax, VA, 22030
United States
Phone: 1 703 993 9833
Email: hxue4@gmu.edu

Abstract

Background: e-Cigarette (electronic cigarette) use has been a public health issue in the United States. On June 23, 2022, the US Food and Drug Administration (FDA) issued marketing denial orders (MDOs) to Juul Labs Inc for all their products currently marketed in the United States. However, one day later, on June 24, 2022, a federal appeals court granted a temporary reprieve to Juul Labs that allowed it to keep its e-cigarettes on the market. As the conversation around Juul continues to evolve, it is crucial to gain insights into the sentiments and opinions expressed by individuals on social media.

Objective: This study aims to conduct a comprehensive analysis of tweets before and after the ban on Juul, aiming to shed light on public perceptions and sentiments surrounding this contentious topic and to better understand the life cycle of public health–related policy on social media.

Methods: Natural language processing (NLP) techniques were used, including state-of-the-art BERTopic topic modeling and sentiment analysis. A total of 6023 tweets and 22,288 replies or retweets were collected from Twitter (rebranded as X in 2023) between June 2022 and October 2022. The encoded topics were used in time-trend analysis to depict the boom-and-bust cycle. Content analyses of retweets were also performed to better understand public perceptions and sentiments about this contentious topic.

Results: The attention surrounding the FDA's ban on Juul lasted no longer than a week on Twitter. Not only the news (ie, tweets with a YouTube link that directs to the news site) related to the announcement itself, but the surrounding discussions (eg, potential consequences of this ban or block and concerns toward kids or youth health) diminished shortly after June 23, 2022, the date when the ban was officially announced. Although a short rebound was observed on July 4, 2022, which was contributed by the suspension on the following day, discussions dried out in 2 days. Out of the top 50 most retweeted tweets, we observed that, except for neutral (23/45, 51%) sentiment that broadcasted the announcement, posters responded more negatively (19/45, 42%) to the FDA's ban.

Conclusions: We observed a short life cycle for this news announcement, with a preponderance of negative sentiment toward the FDA's ban on Juul. Policy makers could use tactics such as issuing ongoing updates and reminders about the ban, highlighting its impact on public health, and actively engaging with influential social media users who can help maintain the conversation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58919) doi:10.2196/58919

KEYWORDS

social media analysis; data mining; natural language processing; topic modeling; sentiment analysis; e-cigarette regulation; vaping; Twitter analysis; public health related policy; marketing denial orders

Introduction

e-Cigarette use has been a public health issue in the United States. Its prevalence increased among adults from 1.4% in 2014 to 6.4% in 2021, primarily owing to an increase among those who had never smoked cigarettes [1,2]. The upward trend was significantly higher later in the COVID-19 pandemic (April 2021-April 2022) compared with its initial months (March-July 2020) [3]. In addition, e-cigarette (electronic cigarette) use may serve as a behavioral marker of risk for mental health problems among youth, including potential depression and suicidal ideation [4,5], as well as potential pathological changes in their oral and respiratory systems [6].

Juul, which once owned a majority market share of e-cigarette sales, has garnered significant attention and controversy due to the widespread use of its products, especially among young people, and their potential health risks [7,8]. Concerns were raised by state and federal investigators, and it was determined that Juul executives knew their marketing contributed to skyrocketing youth vaping rates nationwide, reversing years of tobacco control efforts [9,10]. According to the US Food and Drug Administration (FDA), nearly 10.7 million young people from 12 to 17 years old have used e-cigarettes or are open to trying them [11]. After reviewing the company's premarket tobacco product applications, the FDA determined that the application lacked sufficient evidence regarding the toxicological profile, raising public health concerns [12]. On June 23, 2022, the FDA issued marketing denial orders (MDOs) to Juul Labs Inc for all of their products currently marketed in the United States [12]. However, one day later, on June 24, 2022, a federal appeals court granted a temporary reprieve to Juul Labs that allowed it to keep its e-cigarettes on the market [13]. On July 5, 2022, the FDA administratively stayed the MDO. This administrative stay temporarily suspends the order during the additional review but does not rescind it [12].

From the policy maker's point of view, as the conversation around Juul continues to evolve, it is crucial to gain insights into the sentiments and opinions expressed by individuals on social media platforms like Twitter (rebranded as X on July 23, 2023). Understanding public opinions, as well as the life cycle of an important policy announcement that moves the needle of public perceptions, can better position the government to facilitate public awareness and curb miscommunication in a resource-constrained environment.

Traditionally, the way to gather responses from the general public relies on large public surveys or small focus group discussions. It was not until recently that copious amounts of unstructured text data on social media could be processed, analyzed, and reasoned, all with the help of natural language processing (NLP) techniques, methods that integrate computer science, artificial intelligence, and computational linguistics. Over the past several years, there have been innovations in NLP research that resonate in public health and social media research [14]. More importantly, recent applications have started to adopt a cutting-edge pretraining of deep Bidirectional Transformers For Language Understanding (BERT) approach that transforms NLP tasks. NLP pipelines have been intensively developed for a variety of text-processing problems, like topic modeling, text summarization, and sentiment analysis [15]. In this study, we use these NLP techniques to delve into the social media data on e-cigarettes, allowing us to understand better the response of the general public to the evolving policy issues and the life cycle of public health-related policy concerns expressed on social media. We conducted a comprehensive analysis of tweets before and after the ban on Juul, aiming to shed light on public perceptions and sentiments surrounding this contentious topic.

Methods

Overview

We collected Twitter posts, responses, and retweets related to the FDA's ban on Juul between June 2022 and October 2022. Keywords used for searching include "FDA," "Ban," "JUUL," and "e-cigarettes," and combinations of these 4. We used Twitter API v2, which has been shown to create almost complete samples of Twitter data [16]. A total of four data sets were obtained from Twitter for further processing: (1) the master file that contains Twitter post text, associated post identifier, and reply; (2) the postmetric file that contains public metrics, including the count of retweets, replies, likes, and quotes; (3) the author-metric file that contains author (tweet poster) metrics, including number of followers, following, and tweets previously posted and listed; and (4) the entity file, which indicates the potential entity that posters are affiliated with. We removed the duplicates and kept each post that had only 1 observation in the master file. We then performed one-to-one merges between the master file and the remaining 3 subsets. In total, we obtained 6023 tweets and 22,288 replies or comments and retweets.

Not only did we collect text and emoji from these posts and responses, but we also collected information regarding the date and time when posts and comments were created, as well as metrics associated with a post (ie, counts of retweets, replies, likes, and quotes) at the time of data collection. In addition, we also collected user metrics such as the number of followers a poster has, the number of accounts a poster follows, the number of tweets an author has posted from the time the account was established, and the number of publicly listed organizations of which the user is a member.

In this study, we conducted a descriptive analysis, topic modeling using the state-of-the-art BERTopic technique, and sentiment analysis.

In the descriptive analysis, we described the number of posts during the observational period to give readers an idea of how

XSL•FO
RenderX

long the life cycle of "FDA's ban on Juul" lasted. We separated the time period into 3 sections: before June 23, 2022; between June 23, 2022, and July 5, 2022; and after July 5, 2022, reflecting the fact that this ban was suspended on July 5, 2022. We counted the total number of tweets and retweets, and we recorded the selected topics over time.

For topic modeling analysis, traditionally, the field has been dominated by the Latent Dirichlet Allocation (LDA) model [17]. A limitation of this model is that, through bag-of-words representations, it disregards semantic relationships among words. As these representations do not account for the context of words in a sentence, the bag-of-words input may fail to represent documents accurately. BERTopic is an emerging topic modeling network that simplifies the topic-building process. There are various elements to model these topics. In this study, 5 procedures are processed:

1. Embedding, where documents are converted into numerical representatives. The sentence transformer used here is all-MiniLM-L6-v2, a pretrained monolingual model that compresses large transformers simply and effectively [18].
2. Dimension reduction of the input embeddings; we use the default Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [19]. UMAP is a fuzzy topological structure-based dimension reduction technique that is used to reduce the complexity of a data set by reducing the number of features while keeping the most important properties of the original data. The default parameters were used (ie, 15 neighbors and 5 components with zero minimum distance).
3. Clustering after the dimension reduction of the input embeddings, that is, similar embeddings were clustered into groups to extract topics. The clustering is key to the accuracy of topic representations. Here, we used the HDBSCAN, a hierarchical clustering algorithm [20], to group the input embeddings into distinct clusters based on a user-specified fine-tuned cluster size (set to 15 as a default). These clusters are a rudimentary representation of potential topics. More default fine-tuning parameters were used (see the 2 repositories for more information [21,22]). The stable clusters were extracted from the condensed tree.
4. c-TF-IDF was used to get an accurate representation of the topic from the bag-of-words that were generated from the previous steps. c-TF-IDF is used, and it takes into account what makes the documents in one cluster different from documents in another cluster, and the importance score per word is calculated. The score is used to determine the presence and order within each topic that a user specifies [22].
5. In sentiment analysis, we analyzed the sentiments toward the ban (positive, negative, and neutral) of the most commonly outlined messages formulated in the tweets based on the topics generated from topic modeling. We constructed sentiment scores using VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithms. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [23]. The score is computed by summing the valence scores of each word in the lexicon, adjusted according to the grammatical and syntactical conventions that humans use when expressing or emphasizing sentiment intensity, and then normalized to be between –1 (most extreme negative) and +1 (most extreme positive) [24]. Following the literature, we set standardized thresholds to classify sentences as either positive (normalized score ≥ 0.05), neutral (–0.05 < normalized score <0.05), or negative (normalized score ≤ –0.05) [23].

We selected the top 50 most-retweeted tweets to perform content analysis and sentiment analysis. The reason we chose the top 50 is because of the skewed distribution. Of all the tweets included, those with 22 or more retweets are in the 99th percentile (mean 1.69, SD 20.34 retweets). The top 50 tweets we selected are those that were most popular, with retweet counts in the 99th percentile. Furthermore, 2 coders manually coded the theme for each post separately, with kappa ranging from 0.25 to 0.89. All the analysis was conducted in Python (version 3.6; Python Software Foundation) [25].

Ethical Considerations

This study was exempted from institutional review board approval because data provided by Twitter academic API are publicly available and deidentified. The data was analyzed with permission from Twitter (rebranded as X).

Results

As shown in Table 1, between June 23 and July 5, the mean (SD) retweet count was 1.47 (SD 17.85), with a higher count before June 23 (5.81, SD 46.41) and a lower count after July 05 (0.75, SD 3.17). More in-depth analysis can be found in Table 2. Similarly, the mean (SD) reply count was highest before June 23 (1.98, SD 12.07) and lowest after July 5 (0.38, SD 1.87), with a mean (SD) of 0.82 (SD 10.39) until the end of the observational period (October 22, 2022). The mean (SD) like count was 9.58 (SD 304.64) during the period of interest, with a higher count before June 23 (37.57, SD 384.17) and a lower count after July 5 (2.35, SD 11.74). Regarding author metrics, the mean (SD) number of followers was highest after July 5 (366,305, SD 2,344,692) and lowest before June 23 (143,546, SD 1,043,550), with a mean (SD) of 229,253 (SD 1,825,307) across the entire sample. The mean (SD) number of accounts followed was highest before June 23 (2707, SD 13,995) and lowest after July 5 (2061, SD 6605), with a mean (SD) of 2508 (SD 14,059) across the entire sample.

XSL•FO
RenderX

Table 1. Summary statistics.

		Timestamps		
	Total (N=6023), mean (SD)	Before June 23, 2022 (N=539), mean (SD)	Between June 23 and July 5, 2022 (N=4109), mean (SD)	After July 5, 2022 (N=1375), mean (SD)
Post metrics				
Retweet count	1.69 (20.34)	5.81 (46.41)	1.47 (17.85)	0.75 (3.17)
Reply count	0.82 (10.39)	1.98 (12.07)	0.82 (11.74)	0.38 (1.87)
Quote count	0.76 (24.60)	3.92 (73.05)	0.56 (13.66)	0.12 (0.67)
Like count	10.44 (276.78)	37.57 (384.17)	9.58 (304.64)	2.35 (11.74)
Author matrices				
Author followers count	22,9253 (1,825,307)	143,546 (1,043,550)	194,634 (1,701,244)	366,305 (2,344,692)
Author following count	2508 (14,059)	2707 (13,995)	2632 (15,793)	2061 (6605)
Author listed count	1575 (10,057)	1115 (6412)	1314 (8877)	2534 (13,800)
Author tweet count	141,171 (269,747)	99,475 (175,178)	139,556 (271,059)	162,345 (293,400)

Table 2. Content and sentiment analysis.

	Expected					Sentiment analysis[a]			Topics
Agreement, %	Agreement, %	Kappa	SE	z	Prob>z	Positive or yes	Negative or no	Neutral	
92	44.72	0.86	0.12	7.03	0	3	19	23	Whether the author supports the FDA's[b] ban
80	44.6	0.64	0.12	5.25	0	2	19	21	The sentiment of Twitter post
66	54.48	0.25	0.14	1.79	0.04	9	24	__[c]	Directly tweeted or feeds from news
76	65.68	0.30	0.14	2.13	0.02	5	32	—	Topic: mentioning or comparing specifically about tobacco, nicotine, cigarettes
88	81.92	0.34	0.14	2.44	0.01	2	40	—	Topic: Concerns about kids or youth
66	52.4	0.29	0.14	2.02	0.02	11	21	—	Topic: the consequence of the FDA's ban or block of Juul
90	74.48	0.61	0.14	4.31	0	5	38	—	Topic: The suspension of the FDA's ban
88	88.	0.00	—	—	—	—	44	—	Topic: FDA's expected move with @youtube news link
96	63.52	0.89	0.14	6.3	0	38	12	—	With URL

[a]Sentiment analysis numbers reflect both coders' agreement.

[b]FDA: US Food and Drug Administration.

[c]Not applicable.

Figures 1 and 2 present the time-trend analysis of all tweets related to the FDA's ban on Juul. Figure 1 plots tweets between June 23 and July 5, 2022, the presuspension period, while Figure 2 plots the period after July 5, 2022, the postsuspension period. The number of tweets and retweets associated with the FDA's ban experienced a significant decline between June 23 and June

XSL•FO
RenderX

26, 2022. On June 26, 2022, the quantity of posts experienced a significant decline, with only about 200 tweets compared with 5000 posts on June 23, 2022. Figure 1 shows the tweets experienced a further decline until July 5, 2022. The peak number of tweets related to the suspension of the FDA ban on Juul (Figure 2), which reached approximately 310, was observed on July 5, 2022. The number of posts dropped subsequently, down to around 25 by July 8, 2022. Tweets and retweets associated with the suspension were not observed from July 9 to July 11, 2022. A slight increase in posts was noticed on July 12, 2022, but then wound down to null until the end of the observation period.

Figure 1. The sharp decline of tweets associated with the US Food and Drug Administration's ban.

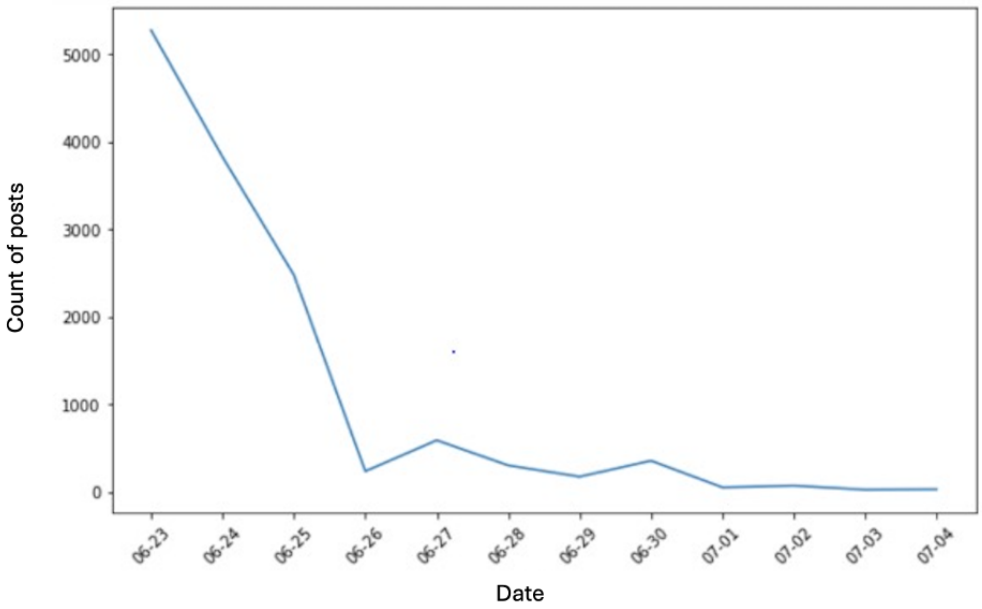

Figure 2. The sharp decline of tweets associated with the suspension of the US Food and Drug Administration's ban.

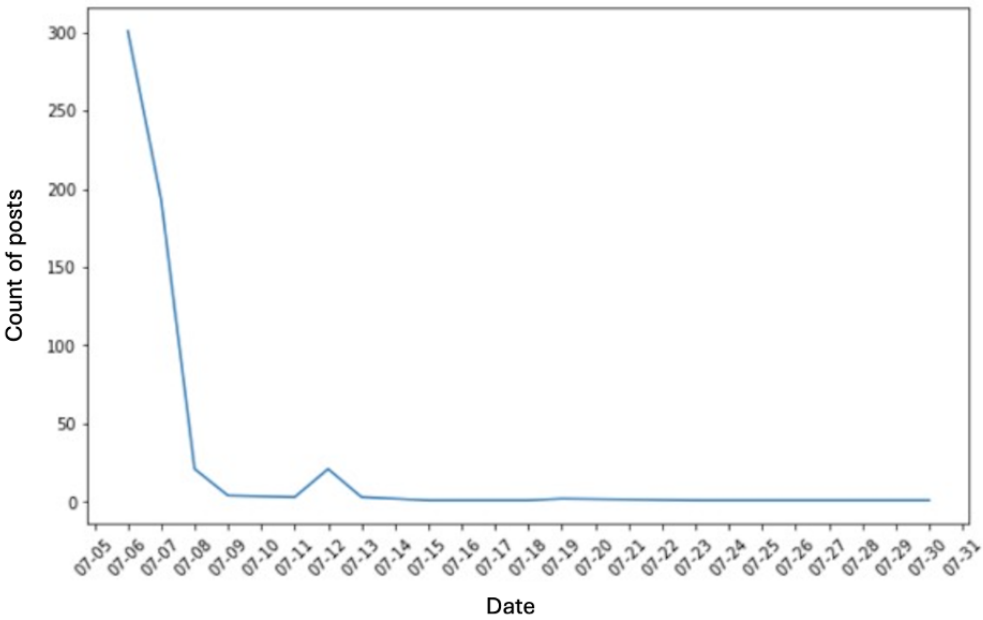

Figure 3 presents the life cycle of the top 6 most popular topics among the tweets related to the FDA's ban on Juul. These 6 topics were chosen because they are most inclusive and concise (Figure S1 in Multimedia Appendix 1 for a detailed inter-topic distance map).

XSL•FO
RenderX

Figure 3. The life cycle of the top 6 topics. FDA: US Food and Drug Administration.

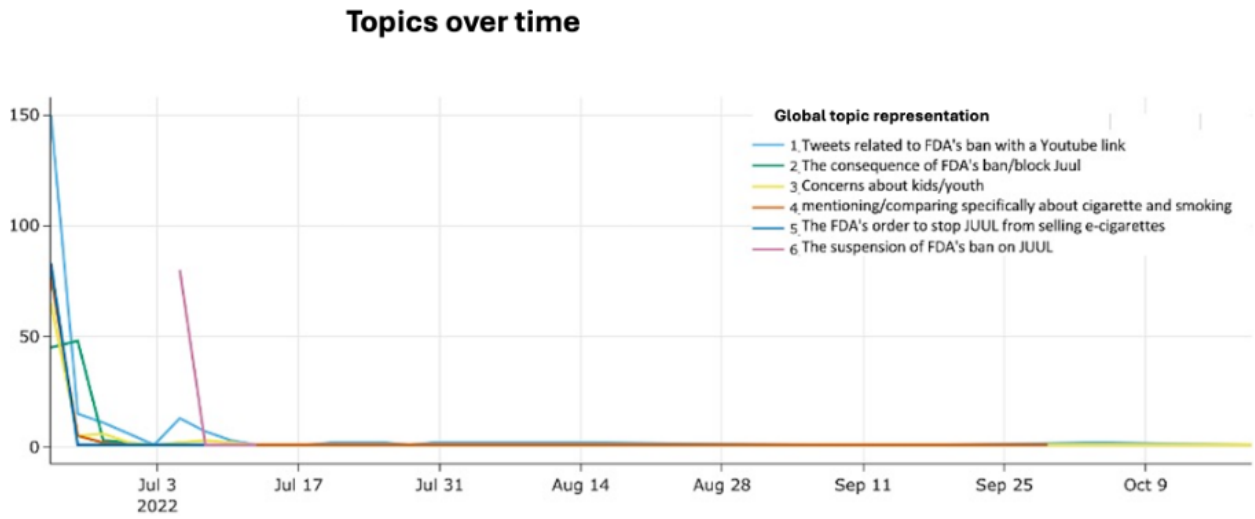

The topics mentioned most frequently in all collected tweets, in rank order, are (1) tweets containing the FDA's ban on Juul with a YouTube link to a news media site or other relevant video content, as in the 2 examples below:

> *Prof. Peter Pitts talks about the reported upcoming FDA ban on Juul with... https://t.co/8fW2Bf31Ct via @YouTube.*

The link above directs viewers to a news show talking about the FDA's ban on Juul (Figure 4).

> *FDA Expected To Ban Juul Products https://t.co/m2MGiGpER2 via @YouTube.*

The link in this quote directs viewers to NBC News (Figure 5).

(2) Tweets talking about the consequence of a continuous ban or block on Juul, (3) tweets related to concerns about kids and youth, (4) tweets mentioning cigarettes and smoking, (5) Tweets that restated the FDA's order to stop Juul from selling e-cigarettes, and (6) tweets associated with the suspension of the FDA's ban on Juul. More concrete examples of word clouds are provided in Figures S2 and S3 in Multimedia Appendix 1.

Figure 4. YouTube link from example post. The screenshot represents external news media sites on YouTube from tweets containing comments on the US Food and Drug Administration's ban on Juul.

XSL•FO
RenderX

Figure 5. YouTube link from example post. The screenshot represents external news media sites on YouTube from Tweets containing comments on the US Food and Drug Administration's ban on Juul.

We observed interesting patterns in the frequency of the 6 different topics between June 23, 2022, and October 9, 2022. Specifically, we found that Topics 1-5 experienced a peak in frequency on June 23, 2022, with a maximum of 150 tweets, followed by a sharp decline. By July 5, 2022, the frequency of topics other than the suspension of the FDA's ban on Juul was down to near null. We observed a short rebound for Topic 1 (tweets related to the FDA's ban with a YouTube news link) on July 4, 2022, right before the announcement of the suspension of the FDA's ban. The rebound in the number of tweets for Topic 1 diminished quickly, followed by a peak for Topic 6 (the suspension of the FDA's ban on Juul). The life cycle of Topic 6, however, was similar to other topics in that it faded out within a day or two.

Table 2 shows the result of the content analysis. In this study, we conducted a conceptual analysis, which determines the existence and frequency of concepts in Twitter text. We developed 2 sentiment concepts, 2 objective facts, and 5 topics. The 2 sentiment concepts are (1) the sentiment of Twitter posts and (2) whether the author of Twitter posts supports the FDA's ban on e-cigarettes. The 2 objective facts are (1) whether the post is directly tweeted or feeds from news and (2) whether the post has a URL or link attached. The 5 topics are (1) related specifically to tobacco, nicotine, cigarettes, or smoking, (2) concerns of kids or youth, (3) the consequence of the FDA's ban or block on Juul, (4) the suspension of the FDA's ban, and (5) the FDA's expected move with a YouTube news link. The raw agreements of these concepts ranged from 66% to 96%, with Kappa agreement statistics being significant ($P<.05$) except for 1 topic. We found, among the top 50 retweeted posts with a 2-interrater agreement of between 66% and 96%, that 23 authors (23/45, 51%) expressed a neutral opinion on the FDA's ban on Juul (most of these retweets were redirected from the FDA's official announcement or a news channel). A total of 19 authors (19/45, 42%) opposed the ban, and 3 (3/45, 6%) authors supported the FDA's ban on Juul. In total, 27% (9/33) of posts were directly tweeted or retweeted from a news agency. The sentiment of Twitter posts (generated by the machine learning models), similar to the authors' supportiveness (measured and verified by 2 independent coders), concentrated toward the neutral (21/42, 50%) and negative (19/42, 45%).

Among the 5 topics generated by topic modeling, 34% (11/32) related to the consequence of the FDA's ban or block on Juul; 13% (5/37) mentioned or compared e-cigarettes with tobacco, nicotine, cigarettes, or smoking; 11% (5/43) talked about the suspension of the FDA's ban; and 4% (4/42) concerned the use of e-cigarettes among kids or youth. Although many other posts among the 6000 tweets were attached to a YouTube link, none of the top 50 posts provided such a link.

Discussion

Principal Findings

In this study, we found that the life cycle of reactions to the FDA's ban on Juul lasted no longer than a week on Twitter. Not only the news related to the announcement itself but the surrounding discussions (the 6 topics presented in the study) diminished shortly after June 23, 2022, the date when the ban

XSL•FO
RenderX

was officially announced. Among the top 50 most-retweeted tweets, we found posters responded more negatively on the corresponding topics.

Our trend analysis findings reveal a pattern commonly observed in the life cycle of internet-based news. Our research shows a sharp increase in Twitter discussions surrounding the FDA's ban on Juul when the announcement was made on June 23, 2022, followed by a rapid decline in 3 days. The patterns for various specific topics were very similar to the general trend. Other Twitter-based studies focusing on the FDA's drug safety communication messaging about Zolpidem (a sedative-hypnotic used for sleeping problems) and antibiotics found that daily mention in Twitter posts spiked at the time of official announcements on these topics, reaching maximal activity within 24 hours and returning to pre-peak basal levels within 48 hours [26,27]. The sharp increase in engagement is not maintained. This pattern aligns with previous literature on the topic, which has documented the short attention span required and the fast-paced nature of internet-based discourse [28,29]. The initial surge of interest and subsequent decline can be attributed to the fleeting nature of trending topics and the constant influx of new information competing for users' attention [30].

Although sharing a common pattern of short life cycles, differences among topics warrant a further dissection. First, concern about kids' and youths' health is one of the most popular topics associated with the FDA's Juul ban. The discussions surrounding the topic echo the growing body of evidence showing the adverse health effects of e-cigarettes on the younger population [31,32]. Second, the potential consequence of the ban, as a categorized topic, sparked a more sustained discussion for 2 days after the announcement (Topic 2, green line shown in Figure 3).

Providing internet-based space for people to reason and debate, rather than simply preaching facts, could be a way to increase user engagement. For example, governments can partner with social media influencers that actively manage and encourage debates about important public health policies. This could not only increase the reach of information but also alter potential miscommunication on social media. Ongoing reminders about the ban could be another way to extend the duration of public engagement. Evidence showed that a reminder system on social media could enhance medical adherence to the treatment of commutable diseases [33]. We observed a short rebound of discussions about the FDA's ban with a YouTube news link on July 4, 2022, right before the announcement of the suspension of the FDA's ban, indicating that timely reminders may be able to prolong user engagement.

Regarding sentiment, we found that of the top 50 retweeted posts, more than half were neutral regarding the ban (where the announcement was directly retweeted or posted), while 42% (19/45) opposed the ban. We observed that Twitter users tended to be more negative toward public health policies. The distribution of sentiments shows great similarity with attitudes toward other US tobacco control laws. For example, a study focused on the Tobacco 21 law found that 42.4% (405/955) of tweets opposed Tobacco 21, and 42.6% (407/955) neither supported nor opposed the law [34]. The proportionally higher opposition is consistent with previous literature focusing on other social media sites. It has been shown that Facebook users are more likely to engage with negatively-framed antitobacco campaign posts [24]. On Twitter, research suggests that positive sentiment has dominated the discourse surrounding e-cigarettes [35]. Such positive sentiments were raised by e-cigarette advocates, along with nudging tactics, to communicate their beliefs. Positive sentiment about the use of Juul suggests that the product is being normalized among young people [36]. The positive attitude toward e-cigarettes forms a feedback loop that could be disseminated to ordinary people, making them more suspicious and more likely to hold negative opinions toward vaping-related policies [35].

In fact, similar public responses were observed in previous e-cigarette–related policies. On January 2, 2020, the FDA released the e-cigarette flavor enforcement policy to prohibit the sale of all flavored cartridge-based e-cigarettes [37]. The proportion of negative sentiment tweets about e-cigarettes significantly increased after the announcement of this FDA policy compared with before the announcement of the policy [37]. Similar negative sentiments were found when New York state banned flavored e-cigarettes [38], as well as in public response to a social media tobacco prevention campaign [39].

When looking into topics more specifically, those Twitter authors who hold negative opinions are mostly individual influencers. For example, the second, fourth, and ninth most retweeted tweets stated:

> *Honestly crazy as that the FDA decided to ban Juul pods before banning actual cigarettes, the agenda obviously isn't just public health.*
>
> *So the FDA wants to ban JUUL e-cigarettes for adults, but approved experimental mRNA vaccines for 6 month old infants at ~0% risk of having any complications from COVID?*
>
> *The WSJ reports that the FDA will ban Juul e-cigarettes tomorrow.*
>
> *Bad idea. E-cigarettes save lives.*

This pattern is similar to what was found in public reactions to other e-cigarette regulations on Twitter. Lazard et al [40] found all 8 top influencers identified were actively against the FDA deeming of e-cigarettes, of which 7 of them were individual consumers and proponents, resulting in miscommunication. Miscommunication of tobacco control policy on social media sites has been documented. For example, many news tweets about the US Federal Tobacco 21 law, a sales law, incorrectly described the law as a purchase law, and some doubted its ability to limit youth access to tobacco products [41]. In addition, our selected example of the most retweeted tweets with negative sentiment showed that people often questioned the identity of the targeted groups, the intention of the authorities, and the effectiveness of the rules. This is consistent with previous research focusing on the FDA's action to prohibit menthol. It suggested that the tweets with a negative attitude questioned the FDA's proposed menthol cigarette rules from several angles, from the effectiveness of the rules and the targeted groups to even the feasibility of their enforcement [42].

XSL•FO
RenderX

Empirical Contributions and Policy Implications

Our findings have important implications for policy makers aiming to prolong the life cycle of discussions and increase their reach. To extend the duration of public engagement, policy makers could use tactics such as ongoing updates and reminders about the ban, highlighting its impact on public health, and actively engaging with influential social media users who can help maintain the conversation. In addition, strategies used in other fields, such as marketing and entertainment, could be adapted to enhance the reach of the information. Leveraging storytelling techniques, creating compelling visuals, and using collaborative formats have proven effective in capturing and maintaining public attention [43,44]. Finally, using social media influencers that actively create cohesive communications could increase the reach of information and alter potential miscommunication on social media. It has been shown that the antitobacco messages had a significantly lower potential reach, received a lower proportion of impressions, and spent a lower proportion of money per message [45]. Using social influencers as message sources is a key factor for message dissemination and sustention. Evidence suggests that campaigns that used social influencers as message sources generated a greater volume of tweets per day and a broader reach per day. More importantly, the oppositional messages diminished over time, which indicates a decrease in miscommunications. Using these tactics and strategies, policy makers can aim to foster sustained engagement and maximize the impact of policy announcements in the age of rapidly evolving internet-based discourse.

Strengths and Limitations

Although we collected comprehensive tweets to disentangle the public response to the FDA's ban on Juul, several limitations are worth noting in this study. First, the opinions we observed might not be generalizable to the entire public since the typical Twitter users are between 25 and 34 years old [46]. Second, our analyses were limited to the data collected using prespecified keywords, which might not be exhaustively comprehensive. As a result, some public responses may not have been covered. Furthermore, the sentiment analysis did not include replies or comments to those tweets, which could result in neglecting some user sentiments. However, the number of responses was low, and the chance of underestimating positive sentiments and overestimating negative sentiments was likely minimal. Future research should develop insights into maintaining the life cycle of discussions and avoiding miscommunication. Specifically, government agencies and organizations could work with influencers to share more focused and nonjudgmental messaging about policy reasoning with fun experiences that resonate with the targeted audiences' interests and values [47]. Researchers should also develop bottom-up agent-based simulation models to develop more insight into how tobacco control policies are disseminated and received by the general public under different social network structures.

Conclusions

In this observational study, we found that individual Twitter users, other than regular news media, hold more negative sentiments toward the FDA's ban on Juul. Furthermore, we observed a short life cycle for this news announcement. To extend the duration of public engagement, policy makers could use tactics such as ongoing updates and reminders about the ban, highlighting its impact on public health, and actively engaging with influential social media users who can help maintain the conversation.

Data Availability

The data sets generated during and/or analyzed during this study are available in the Google Drive repository. The codes used for the analyses are available upon proper request to the authors. No generative AI was used in any portion of the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material.
[DOCX File , 517 KB - jmir_v26i1e58919_app1.docx]

References

1. Bandi P, Cahn Z, Goding Sauer A, Douglas CE, Drope J, Jemal A, et al. Trends in e-cigarette use by age group and combustible cigarette smoking histories, U.S. adults, 2014-2018. Am J Prev Med 2021;60(2):151-158 [FREE Full text] [doi: 10.1016/j.amepre.2020.07.026] [Medline: 33032869]
2. Bandi P, Star J, Minihan AK, Patel M, Nargis N, Jemal A. Changes in E-cigarette use among U.S. adults, 2019-2021. Am J Prev Med 2023;65(2):322-326. [doi: 10.1016/j.amepre.2023.02.026] [Medline: 37479423]
3. Kreslake JM, O'Connor KM, Liu M, Vallone DM, Hair E. A resurgence of e-cigarette use among adolescents and young adults late in the COVID-19 pandemic. PLoS One 2023;18(3):e0282894 [FREE Full text] [doi: 10.1371/journal.pone.0282894] [Medline: 36989261]
4. Livingston JA, Chen CH, Kwon M, Park E. Physical and mental health outcomes associated with adolescent E-cigarette use. J Pediatr Nurs 2022;64:1-17. [doi: 10.1016/j.pedn.2022.01.006] [Medline: 35121206]

XSL•FO
RenderX

5. Cambron C. E-cigarette use is associated with increased psychological distress among youth: a pooled cross-sectional analysis of state-level data from 2019 and 2021. Int J Environ Res Public Health 2022;19(18):11726 [FREE Full text] [doi: 10.3390/ijerph191811726] [Medline: 36142001]
6. Huizink AC. Trends and associated risks in adolescent substance use: E-cigarette use and nitrous oxide use. Curr Opin Psychol 2022;45:101312 [FREE Full text] [doi: 10.1016/j.copsyc.2022.101312] [Medline: 35313182]
7. Alonso F, Rath J, Ramírez AS, Cantrell J, Jordan A, Suarez S, et al. Perceptions of and experiences with cigarette and E-cigarette use among a diverse population of US latino adolescents and young adults. Am J Health Promot 2023;37(5):646-653 [FREE Full text] [doi: 10.1177/08901171221151125] [Medline: 36608694]
8. Gorukanti A, Delucchi K, Ling P, Fisher-Travis R, Halpern-Felsher B. Adolescents' attitudes towards e-cigarette ingredients, safety, addictive properties, social norms, and regulation. Prev Med 2017;94:65-71 [FREE Full text] [doi: 10.1016/j.ypmed.2016.10.019] [Medline: 27773711]
9. Furlow B. Juul settles youth e-cigarette marketing lawsuits. Lancet Respir Med 2023;11(6):e56. [doi: 10.1016/s2213-2600(23)00179-0]
10. Wang Y, Duan Z, Weaver SR, Popova L, Spears CA, Ashley DL, et al. Consumption of JUUL vs. other e-cigarette brands among U.S. e-cigarette users: evidence from wave 5 of the PATH study. Int J Environ Res Public Health 2022;19(17):10837 [FREE Full text] [doi: 10.3390/ijerph191710837] [Medline: 36078551]
11. Chappell B. The FDA orders Juul to pull all of its vaping products from the U.S. market. Washington, DC: NPR; 2023. URL: https://www.npr.org/2022/06/23/1106846310/juul-fda-vaping-e-cigarettes [accessed 2024-08-21]
12. U.S. Food and Drug Administration. FDA denies authorization to market JUUL products. 2022. URL: https://www.fda.gov/news-events/press-announcements/fda-denies-authorization-market-juul-products [accessed 2024-08-21]
13. Richtel M, Hirsch L, Jacobs A. Juul gets temporary reprieve to keep selling its e-cigarettes. New York: The New York Times URL: https://www.nytimes.com/2022/06/24/health/juul-ecigarettes-ban.html [accessed 2024-08-21]
14. Hao T, Huang Z, Liang L, Weng H, Tang B. Health natural language processing: methodology development and applications. JMIR Med Inform 2021;9(10):e23898 [FREE Full text] [doi: 10.2196/23898] [Medline: 34673533]
15. Elbattah M, Arnaud E, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. Healthinf 2021:825-832 [FREE Full text] [doi: 10.5220/0010414508250832]
16. Juergen P, Mooseder A, Lasser L, Hammer L, Stritzel O, Garcia D. This sample seems to be good enough! Assessing coverage and temporal reliability of Twitter's academic API. In: Computer Science > Social and Information Networks. Palo Alto, California: AAAI Press; 2022 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; June 5–8, 2023; Limassol, Cyprus p. 720-729 URL: https://www.researchgate.net/publication/371517429_This_Sample_Seems_to_Be_Good_Enough_Assessing_Coverage_and_Temporal_Reliability_of_Twitter%27s_Academic_API
17. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. In: Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference. Cambridge, Massachusetts: The MIT Press; 2003:993-1022.
18. Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems 2020;33:5776-5788 [FREE Full text]
19. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. Statistics > Machine Learning 2018:1-63 [FREE Full text]
20. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. Springer 2013:160-172. [doi: 10.1007/978-3-642-37456-2_14]
21. McInnes L, Astels JSH. The hdbscan clustering library. 2017. URL: https://hdbscan.readthedocs.io/en/latest/index.html [accessed 2024-08-21]
22. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. ArXiv. Preprint posted online on March 11, 2022 2022.
23. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media 2014;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]
24. Lin SY, Cheng X, Zhang J, Yannam JS, Barnes AJ, Koch JR, et al. Social media data mining of antitobacco campaign messages: machine learning analysis of Facebook posts. J Med Internet Res 2023;25:e42863 [FREE Full text] [doi: 10.2196/42863] [Medline: 36780224]
25. van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
26. Dyar OJ, Castro-Sánchez E, Holmes AH. What makes people talk about antibiotics on social media? A retrospective analysis of Twitter use. J Antimicrob Chemother 2014;69(9):2568-2572 [FREE Full text] [doi: 10.1093/jac/dku165] [Medline: 24862092]
27. Sinha MS, Freifeld CC, Brownstein JS, Donneyong MM, Rausch P, Lappin BM, et al. Social media impact of the food and drug administration's drug safety communication messaging about zolpidem: mixed-methods analysis. JMIR Public Health Surveill 2018;4(1):e1 [FREE Full text] [doi: 10.2196/publichealth.7823] [Medline: 29305342]
28. Bakshy E, Rosenn I, Marlow C, Adamic L. The role of social networks in information diffusion. 2012 Presented at: Proceedings of the 21st International Conference on World Wide Web; April 16-20, 2012; Lyon, France. [doi: 10.1145/2187836.2187907]

XSL•FO
RenderX

29. Lerman K, Ghosh R. Information contagion: an empirical study of the spread of news on digg and Twitter social networks. 2010 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; May 23–26, 2010; Washington, DC. [doi: 10.1609/icwsm.v4i1.14021]
30. Castillo C, El-Haddad M, Pfeffer J, Stempeck M. Characterizing the life cycle of online news stories using social media reactions. 2014 Presented at: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing; February 15-19, 2014; Baltimore, Maryland, USA. [doi: 10.1145/2531602.2531623]
31. Lin SY, Wang L, Zhou W, Kitsantas P, Wen X, Xue H. E-cigarette use during pregnancy and its association with adverse birth outcomes in the US. Prev Med 2023;166:107375 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107375] [Medline: 36481272]
32. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. E-Cigarette Use Among Youth and Young Adults: A Report of the Surgeon General. Atlanta, Georgia: National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health; 2016.
33. Lin B, Liu J, He W, Pan H, Ma Y, Zhong X. Effect of a reminder system on pre-exposure prophylaxis adherence in men who have sex with men: prospective cohort study based on WeChat intervention. J Med Internet Res 2022;24(8):e37936 [FREE Full text] [doi: 10.2196/37936] [Medline: 35969436]
34. Dobbs PD, Colditz JB, Shields S, Meadows A, Primack BA. Policy and behavior: comparisons between Twitter discussions about the US tobacco 21 law and other age-related behaviors. Int J Environ Res Public Health 2022;19(5):2613 [FREE Full text] [doi: 10.3390/ijerph19052613] [Medline: 35270306]
35. McCausland K, Maycock B, Leaver T, Wolf K, Freeman B, Jancey J. E-cigarette advocates on Twitter: content analysis of vaping-related Tweets. JMIR Public Health Surveill 2020;6(4):e17543 [FREE Full text] [doi: 10.2196/17543] [Medline: 33052130]
36. Sidani J, Colditz JB, Barrett EL, Chu KH, James AE, Primack BA. JUUL on Twitter: analyzing Tweets about use of a new nicotine delivery system. J Sch Health 2020;90(2):135-142 [FREE Full text] [doi: 10.1111/josh.12858] [Medline: 31828791]
37. Lu X, Sun L, Xie Z, Li D. Perception of the food and drug administration electronic cigarette flavor enforcement policy on Twitter: observational study. JMIR Public Health Surveill 2022;8(3):e25697 [FREE Full text] [doi: 10.2196/25697] [Medline: 35348461]
38. Gao Y, Xie Z, Li D. Investigating the impact of the New York state flavor ban on e-cigarette-related discussions on Twitter: observational study. JMIR Public Health Surveill 2022;8(7):e34114 [FREE Full text] [doi: 10.2196/34114] [Medline: 35802417]
39. Majmundar A, Le N, Moran MB, Unger JB, Reuter K. Public response to a social media tobacco prevention campaign: content analysis. JMIR Public Health Surveill 2020;6(4):e20649 [FREE Full text] [doi: 10.2196/20649] [Medline: 33284120]
40. Lazard AJ, Wilcox GB, Tuttle HM, Glowacki EM, Pikowski J. Public reactions to e-cigarette regulations on Twitter: a text mining analysis. Tob Control 2017;26(e2):e112-e116. [doi: 10.1136/tobaccocontrol-2016-053295] [Medline: 28341768]
41. Dobbs PD, Schisler E, Colditz JB, Primack BA. Miscommunication about the US federal Tobacco 21 law: a content analysis of Twitter discussions. Tob Control 2023;32(6):696-700. [doi: 10.1136/tobaccocontrol-2021-057099] [Medline: 35173067]
42. Zhou R, Tang Q, Xie Z, Li D. Public perceptions of the food and drug administration's proposed rules prohibiting menthol cigarettes on Twitter: observational study. JMIR Form Res 2023;7:e42706 [FREE Full text] [doi: 10.2196/42706] [Medline: 36763414]
43. Rusbridger A. Breaking news: the remaking of journalism and why it matters now (a review). TLC Journal 2018;2(3):101-104 [FREE Full text] [doi: 10.29366/2018tlc.2.3.8]
44. Keller KL, Richey K. The importance of corporate brand personality traits to a successful 21st century business. Journal of Brand Management 2006;14:74-81. [doi: 10.1057/palgrave.bm.2550055]
45. Majmundar A, Moran MB. For or against tobacco control: sponsored tobacco advocacy messages on Facebook and Instagram. Nicotine Tob Res 2021;23(11):1861-1868. [doi: 10.1093/ntr/ntab111] [Medline: 34049405]
46. Dixon S. Twitter: distribution of global audiences 2021, by age group 2022. URL: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/ [accessed 2024-08-21]
47. Toledo G, McQuoid J, Ling PM. "It's Not Too Aggressive": key features of social branding anti-tobacco interventions for high-risk young adults. Health Promot Pract 2021;22(3):423-432 [FREE Full text] [doi: 10.1177/1524839920910372] [Medline: 32111139]

Abbreviations

BERT: Bidirectional Transformers For Language Understanding
e-Cigarette: electronic cigarette
FDA: US Food and Drug Administration
LDA: Latent Dirichlet Allocation
MDO: marketing denial order
NLP: natural language processing

XSL•FO
RenderX

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by A Mavragani; submitted 27.03.24; peer-reviewed by GK Gupta, M Elbattah; comments to author 08.05.24; revised version received 13.06.24; accepted 25.07.24; published 01.10.24.

Please cite as:
Lin SY, Tulabandu SK, Koch JR, Hayes R, Barnes A, Purohit H, Chen S, Han B, Xue H
Public Response to Federal Electronic Cigarette Regulations Analyzed Using Social Media Data Through Natural Language Processing: Topic Modeling Study
J Med Internet Res 2024;26:e58919
URL: https://www.jmir.org/2024/1/e58919
doi:10.2196/58919
PMID:39352739

©Shuo-Yu Lin, Sahithi Kiran Tulabandu, J Randy Koch, Rashelle Hayes, Andrew Barnes, Hemant Purohit, Songqing Chen, Bo Han, Hong Xue. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of Machine Learning to Detect Influenza Using Wearable Sensor Data and Patient-Reported Symptoms: Cohort Study

Kamran Farooq[1], PhD, SMIEEE; Melody Lim[2], MSc; Lawrence Dennison-Hall[3], BSc; Finn Janson[3], BSc; Aspen Hazel Olszewska[4], High School Diploma; Muhammad Mamduh Ahmad Zabidi[5], PhD; Anna Haratym-Rojek[4], MSc; Karol Narowski[6], MSc; Barry Clinch[3], PhD; Marco Prunotto[7,8], PhD; Devika Chawla[2], PhD; Victoria Hunter[2], MSc; Vincent Ukachukwu[3], MSc, MRCGP, FFPM

[1]Roche Data & Analytics Chapter (Data Science), Kaiseraugst, Switzerland
[2]Genentech, Inc, South San Francisco, CA, United States
[3]Roche Products Ltd, Welwyn Garden City, United Kingdom
[4]Roche Global IT Solution Centre, Warsaw, Poland
[5]Roche Services & Solutions (Asia Pacific) Sdn Bhd, Subang Jaya, Malaysia
[6]Badger Software Sp z oo, Wroclaw, Poland
[7]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland
[8]F Hoffmann-La Roche Ltd, Basel, Switzerland

Corresponding Author:
Kamran Farooq, PhD, SMIEEE
Roche Data & Analytics Chapter (Data Science)
Wurmisweg
Kaiseraugst, 4303
Switzerland
Phone: 41 616881111
Email: kamran.farooq@roche.com

Abstract

Background: Machine learning offers quantitative pattern recognition analysis of wearable device data and has the potential to detect illness onset and monitor influenza-like illness (ILI) in patients who are infected.

Objective: This study aims to evaluate the ability of machine-learning algorithms to distinguish between participants who are influenza positive and influenza negative in a cohort of symptomatic patients with ILI using wearable sensor (activity) data and self-reported symptom data during the latent and early symptomatic periods of ILI.

Methods: This prospective observational cohort study used the extreme gradient boosting (XGBoost) classifier to determine whether a participant was influenza positive or negative based on 3 models using symptom-only data, activity-only data, and combined symptom and activity data. Data were collected from the Home Testing of Respiratory Illness (HTRI) study and FluStudy2020, both conducted between December 2019 and October 2020. The model was developed using the FluStudy2020 data and tested on the HTRI data. Analyses included participants in these studies with an at-home influenza diagnostic test result. Fitbit (Google LLC) devices were used to measure participants' steps, heart rate, and sleep parameters. Participants detailed their ILI symptoms, health care–seeking behaviors, and quality of life. Model performance was assessed by area under the curve (AUC), balanced accuracy, recall (sensitivity), specificity, precision (positive predictive value), negative predictive value, and weighted harmonic mean of precision and recall (F_2) score.

Results: An influenza diagnostic test result was available for 953 and 925 participants in HTRI and FluStudy2020, respectively, of whom 848 (89%) and 840 (90.8%) had activity data. For the training and validation sets, the highest performing model was trained on the combined symptom and activity data (training AUC=0.77; validation AUC=0.74) versus symptom-only (training AUC=0.73; validation AUC=0.72) and activity-only (training AUC=0.68; validation AUC=0.65) data. For the FluStudy2020 test set, the performance of the model trained on combined symptom and activity data was closely aligned with that of the symptom-only model (combined symptom and activity test AUC=0.74; symptom-only test AUC=0.74). These results were validated using independent HTRI data (combined symptom and activity evaluation AUC=0.75; symptom-only evaluation

XSL•FO
RenderX

AUC=0.74). The top features guiding influenza detection were cough; mean resting heart rate during main sleep; fever; total minutes in bed for the combined model; and fever, cough, and sore throat for the symptom-only model.

Conclusions: Machine-learning algorithms had moderate accuracy in detecting influenza, suggesting that previous findings from research-grade sensors tested in highly controlled experimental settings may not easily translate to scalable commercial-grade sensors. In the future, more advanced wearable sensors may improve their performance in the early detection and discrimination of viral respiratory infections.

(J Med Internet Res 2024;26:e47879) doi:10.2196/47879

KEYWORDS

influenza; influenza-like illness; wearable sensor; person-generated health care data; machine learning

Introduction

Background

Between 2010 and 2020, an estimated 9 to 41 million annual illnesses were attributed to influenza infection in the United States [1]. Estimates of US annual hospitalizations ranged from 140,000 to 710,000 and deaths from 12,000 to 52,000 [1]. Early diagnosis and implementation of nonpharmaceutical interventions (eg, quarantine) are critical in preventing onward transmission and reducing the disease burden of influenza-like illnesses (ILIs). Recent data show that wearable devices (eg, fitness trackers and smartwatches) may help detect viral infection before symptoms develop and may provide an early warning system for viral illness [2-12].

In 2019 and 2020, approximately 29% of US adults were reported to use wearable devices [13], which range from fitness trackers that passively record heart rate (HR) and the number of daily steps to more sophisticated devices that can measure parameters such as sleep duration and quality, blood pressure, blood glucose, and oxygen saturation levels [14]. The wide availability and increasing popularity of wearable devices make them convenient, passive tools to record person-generated health data that could be harnessed to improve both individual and public health.

Wearable devices have the potential to detect the onset of illness and monitor disease progression or severity in patients infected with virus. In the future, this may allow people to be alerted to possible infection or the need to seek medical care in the early stages of disease [2-6]. Grzesiak et al [4] further showed that wearable devices may be able to predict the infection severity profile of a patient up to 24 hours before the onset of symptoms following exposure to the influenza virus or rhinovirus. The ability to predict illness severity may provide opportunities to discriminate between respiratory viral infections with more severe clinical presentation that carry a greater risk to public health, such as influenza and COVID-19, and those with a mostly mild clinical presentation, such as the common cold. Our previous analysis objectively characterized the “wearable phenotype” of individuals with ILI as well as those with confirmed influenza infection [15]. We demonstrated that before symptom onset, and throughout an ILI event, individuals experience reduced total daily steps, total active time, and sleep efficiency, as well as increased sleep duration and changes in resting HR (RHR) [15].

Here, we report the development and evaluation of a machine-learning model to detect laboratory-confirmed influenza infection based on wearable sensor and symptom data in the latent and early symptomatic periods (up to 1 day after symptom onset), using an extreme gradient boosting (XGBoost) classifier [16,17] in a cohort of symptomatic patients with ILI.

Machine learning offers a quantitative analysis of the data collected from wearable devices; XGBoost is an optimized distributed gradient boosting library that implements machine-learning algorithms, providing parallel-tree boosting [17]. XGBoost, a supervised machine-learning process, can be used to solve classification tasks, in which one can determine whether an instance is in a particular category by studying the features of that instance [17].

Using commercial wearable sensors (Fitbit [Google LLC]), it has previously been demonstrated that nationwide mobility (measured as total daily steps in a US population) decreased due to ILI symptoms and that ILI burden (determined by the difference in total daily steps) was associated with care-seeking behaviors, the number of workdays missed, and self-reported overall health [18]. Another study showed that abnormalities in RHR and sleep duration, measured by wearable sensors, could be leveraged to predict the real-time incidence of ILI [19]. Recently, wearable sensor data have also been used to assess physiological signs associated with COVID-19 [5,10,12,20-26].

Objective

The objective of our analysis was to determine the ability of an XGBoost model to distinguish between participants who are influenza positive and influenza negative during the latent and early symptomatic periods of ILI (days –4 to +1). Wearable and symptoms data were used, gathered from 2 independent studies, FluStudy2020 and the HTRI study (NCT04245800); the former was used for training, testing, and validation, and the latter was used as a secondary holdout set for evaluation.

Methods

Study Design and Participants

Overview

This prospective observational cohort study evaluated the ability of machine-learning algorithms to distinguish between participants who are influenza positive and influenza negative in a cohort of patients with ILI. Analyses were conducted using wearable sensor (activity) data and self-reported symptom

XSL•FO
RenderX

severity data from participants enrolled in FluStudy2020, with an influenza diagnostic test result from a self-administered kit [15,27]. Data from the HTRI study were used as an independent holdout set. All participants provided written consent.

The XGBoost model was used to classify whether a participant was influenza positive or negative based on 3 models using symptom-only data, activity-only data, and combined symptom and activity data. Other models were not assessed with these data based on previous internal analyses with different data, in which H2O AutoML was used to train and tune various models; XGBoost was found to be the best-performing model. Participant variables, including age, gender, BMI, and month in which the participant conducted an at-home influenza test, were considered before the Boruta feature selection algorithm was applied to the activity-only and combined symptom and activity models. The symptom-only model included all participant variables. The XGBoost model was assessed for its early detection of influenza infection in the FluStudy2020 training, validation, and test sets, as well as the HTRI secondary holdout set, using the following metrics: balanced accuracy, recall (sensitivity), specificity, precision (positive predictive value), negative predictive value, weighted harmonic mean of precision and recall (F_2) score, and area under the receiver operating characteristic curve (AUC ROC). Calibration plots and feature importance plots were generated for each of the 3 models. A model evaluation schematic is shown in Figure S1 in Multimedia Appendix 1.

Data Collection and Preprocessing

The HTRI study and FluStudy2020 were conducted by Evidation Health in adults in the United States between December 2019 and October 2020. Participants in each study were aged ≥18 years, lived in the United States, and owned and were willing to wear a Fitbit device during the day and during sleep for the duration of the study. Full inclusion and exclusion criteria are shown in Table S1 in Multimedia Appendix 1. Steps, HR, and sleep data were collected through continuous passive monitoring via the participants' Fitbit devices. Participants also completed daily surveys of whether they experienced influenza symptoms in the past 24 hours, self-reported ILI symptom severity, health care–seeking behaviors, and quality of life. Biweekly and monthly surveys were used to capture influenza-related complication events and vaccination history. Participants reporting certain ILI symptoms were instructed to perform a self-administered influenza diagnostic test. Samples were returned to the laboratory for the confirmation of influenza by a highly sensitive reverse transcription polymerase chain reaction test. The primary assessment of data from the 2 studies, including the removal of physiologically implausible data or null estimates, has been described previously [15]. Missing data were automatically handled by XGBoost by finding the best split direction when missing data were noted.

Participants with an influenza diagnostic test result were identified, and activity data were assessed for quality and completeness for each participant day. Step data were considered valid if the participant had at least 10 hours of step wear time [28,29] or if they had a valid HR day. HR data were considered valid if they included a minimum of 600 minutes (10 hours) of HR measurements and if a Fitbit-estimated RHR measure was available for that day. Sleep data were considered valid if nonzero and nonmissing total sleep minutes were available for the day. Finally, any day with <10 hours of wear time was considered invalid.

The maximum self-reported severity of 8 symptoms was analyzed: early fever, sore throat, cough, headache, muscle ache, chills, fatigue, and nasal congestion. In total, 41 activity features were analyzed, including RHR, total minutes asleep, the total number of steps, the proportion of the day that the participant spent being physically active (defined as ≥50 steps per minute), the maximum amount of activity the participant was able to complete within a single hour of the day, sleep efficiency score during main sleep, minutes in bed for the main sleep only of the day, the number of naps, total minutes in bed, the percentage of minutes with HR >1.5×RHR for the day, the proportion of minutes with nonzero steps out of the total minutes the device was worn, and mean RHR during main sleep. In addition, 29 HR variability (HRV) features were analyzed, derived from RHR captured during the participant's sleep period.

Model Building and Optimization

Baseline predictors were assessed by including symptom-only features and activity-only features and then combining both features for the latent period to day 1 of ILI (days –4 to +1). This baseline model was built on the XGBoost classifier, which was selected as the machine-learning algorithm trained to detect influenza due to its scalability, regularization, and ability to detect complex nonlinear relationships. Metric calculations were based on a previously published study [5]:

where test data represents ILI days –4 to +1, which encompass the latent period (days –4 to –1; ie, the incubation period for influenza), ILI onset (day 0), and part of the early symptomatic period (day +1), and baseline data represents the participants' healthy baseline data from 2 weeks before the latent period (days –18 to –5). The model was optimized using the Bayesian hyperparameter optimization algorithm, a Bayesian inference, and a Gaussian process to find the maximum value of an unknown function with minimal iterations. AUC ROC was the metric subject to optimization; 100 optimization trials were run per model (symptom-only, activity-only, and combined symptom and activity data; each model included participant variables). The parameters optimized were maximum number of trees: (2, 50); learning rate: (0.0001, 0.2); maximum tree depth: (2, 10); subsample ratio of training instances before growing trees: (0.2, 1.0); column subsample ratio at each level: (0.1, 1.0); column subsample ratio at each tree: (0.1, 1.0); column subsample at each node: (0.1, 1.0); maximum delta step allowed by each leaf output: (0, 10); minimum sum of instance weight needed in a child (subtree: [0.0, 10.0]); L1 regularization: (0.00001, 1); and L2 regularization: (0.00001, 100). To mitigate the imbalanced ratio of participants who were influenza positive to influenza negative, class weights were calculated and applied to the model to give greater weight to the minority influenza-positive class.

Given these differences in study design, the current analysis excluded participants who tested positive for influenza in the

XSL•FO
RenderX

HTRI study but did not meet any of the FluStudy2020 criteria for populations with ILI. There was also a small subset of HTRI participants whose symptoms would have met the influenza test kit criteria for FluStudy2020 but from whom a sample was not collected because they did not meet the HTRI influenza test kit criteria. To combine the HTRI data with the FluStudy2020 data, it was necessary to verify that the HTRI participants' self-reported illness dates (which were provided in the same recovery survey in which a health care visit was reported) aligned with the analysis-derived ILI event dates (which were created during the analysis of the daily survey responses). This permitted verification that only health care visits during the same illness period as the ILI event period were included in the analysis. HTRI participants were categorized as having made or not made a health care visit only if their self-reported ILI event period overlapped with the analysis-derived ILI event period.

Model Validation and Evaluation

Stratified k-fold shuffle cross-validation (k=50) was used to ensure model performance was reliable and robust. Overall, 50 models were trained on different training and validation sets before being evaluated on a single test set and the HTRI data. Of the FluStudy2020 data, 64% were used for k-stratified splits consisting of 50 training sets and 16% were used for validation. The remaining 20% of the data were set aside as the testing set. External validation was performed using the HTRI data. The Boruta feature selection algorithm was applied to reduce the dimensions of the activity features to minimize the impact of noise and reduce overfitting.

Model performance was assessed using the following metrics: balanced accuracy, recall (sensitivity), specificity, precision (positive predictive value), negative predictive value, F_2-score, and AUC ROC. The model performance results consisted of the AUC ROC curves and mean performance across each k-fold along with 95% CIs for the training, validation, and test sets. The distribution of positive and negative predictions for the aggregated performance (based on symptom features, activity features, and these features combined) was described using confusion matrices. The values in each confusion matrix comprised the mean across each fold with their respective 95% CIs. Feature importance analyses were performed for each model, with the most important features summarized in feature importance plots.

Software

Analyses were performed using Python (version 3.7; Python Software Foundation); *xgboost* (version 1.5.2) was used for modeling, and *bayesian-optimization* (version 1.2.0) was used for hyperparameter optimization. Feature importance was determined using XGBoost's built-in feature importance. Data processing and visualization were performed with *pandas* (version 1.3.4), *NumPy* (version 1.21.4), and *Matplotlib* (version 3.5.1) [30-32]. *Kedro* was used to build robust and scalable data pipelines [33]. Feature selection was performed with *Boruta* (version 0.3) [34]. Statistical analysis was performed with *SciPy* (version 1.7.3) [35]. Metric computation, k-fold data splitting, and class weight calculations were performed using *scikit-learn* (version 0.24.2) [36]. The Python package *hrvanalysis* [37] was used to derive HRV features.

Ethical Considerations

The HTRI study and FluStudy2020 were conducted by Evidation Health, Inc. Institutional review board approval was given by WCG Clinical for both the HTRI study (study number: 1271380; tracking number: 20192965) and FluStudy 2020 (study number: 1271500; tracking number: 20193003). Participants were recruited from the Evidation consumer platform, a free application that allows members to earn compensation for completing surveys, sharing health activity data, and reading health articles. Individuals were given the opportunity to enroll into the study once they provided informed consent to participate study activities and for use of their data. Participants earned reward points, redeemable for money, as compensation for completing study activities. Reward points worth up to US $10 were available on completion of enrollment, and a maximum of US $109 could be earned over the course of the study, if all study activities were completed. The data used for analysis were deidentified; each participant enrolled in the study was coded with a unique participant identification number.

Results

Participants

FluStudy2020 had 925 participants, of whom 840 (90.8%) had activity data that met the data density criteria. Of these 840 participants, 639 (76.1%) were influenza negative and 201 (23.9%) were influenza positive (Figure S1 in Multimedia Appendix 1). The HTRI study had 953 participants, and activity data meeting the data density criteria were available for 848 (89%) participants. Of these 848 participants, 657 (77.5%) were influenza negative and 191 (22.5%) were influenza positive. Baseline demographics of participants included in the model evaluation are presented in Table 1. Most participants were female (764/840, 90.9% and 660/848, 77.8%) with mean ages of 37.4 (SD 9.6) years and 37.6 (SD 9.1) years for FluStudy2020 and HTRI, respectively. Distributions of age, BMI, and gender were balanced between the group that was influenza negative and the group that was influenza positive. The maximum self-reported symptom severities and wearable sensor data during ILI days –4 to +1 are shown in Table 2 and Table S2 in Multimedia Appendix 1, respectively.

XSL•FO
RenderX

Table 1. Baseline demographics of participants included in the model evaluation.

Characteristics	FluStudy2020 participants			HTRI[a] participants		
	Overall (n=840)	Influenza negative (n=639)	Influenza positive (n=201)	Overall (n=848)	Influenza negative (n=657)	Influenza positive (n=191)
Age (y), mean (SD)	37.42 (9.59)	37.10 (9.35)	38.45 (10.25)	37.55 (9.10)	37.47 (9.15)	37.83 (8.95)
BMI (kg/m^2), mean (SD)	31.25 (8.16)	31.42 (8.16)	30.72 (8.18)	30.64 (7.51)	30.64 (7.69)	30.61 (6.89)
Region, n (%)						
Midwest	291 (34.6)	206 (32.2)	85 (42.3)	299 (35.3)	216 (32.9)	83 (43.5)
Northeast	139 (16.6)	109 (17.1)	30 (14.9)	134 (15.8)	99 (15.1)	35 (18.3)
South	257 (30.6)	198 (31)	59 (29.4)	239 (28.2)	191 (29.1)	48 (25.1)
West	153 (18.2)	126 (19.7)	27 (13.4)	176 (20.8)	151 (23)	25 (13.1)
Gender, n (%)						
Female	764 (91)	586 (91.7)	178 (88.6)	660 (77.8)	519 (79)	141 (73.8)
Male	71 (8.5)	48 (7.5)	23 (11.4)	186 (21.9)	136 (20.7)	50 (26.2)
Nonbinary	5 (0.6)	5 (0.8)	0 (0)	2 (0.2)	2 (0.3)	0 (0)
Race, n (%)						
Alaska Native, American Indian, Native Hawaiian, or other Pacific Islander	2 (0.2)	2 (0.3)	0 (0)	1 (0.1)	1 (0.2)	0 (0)
Asian	16 (1.9)	13 (2)	3 (1.5)	27 (3.2)	19 (2.9)	8 (4.2)
Black or African American	31 (3.7)	25 (3.9)	6 (3)	21 (2.5)	14 (2.1)	7 (3.7)
White	753 (89.6)	568 (88.9)	185 (92)	763 (90)	593 (90.3)	170 (89)
Multiple races	32 (3.8)	27 (4.2)	5 (2.5)	28 (3.3)	23 (3.5)	5 (2.6)
Other	6 (0.7)	4 (0.6)	2 (1)	8 (0.9)	7 (1.1)	1 (0.5)

[a]HTRI: Home Testing of Respiratory Illness.

XSL•FO
RenderX

Table 2. Frequency of maximum self-reported symptom severity during ILI[a] days –4 to +1.

Characteristics	FluStudy2020 participants			HTRI[b] participants		
	Overall (n=840)	Influenza negative (n=639)	Influenza positive (n=201)	Overall (n=848)	Influenza negative (n=657)	Influenza positive (n=191)
Age (y), mean (SD)	37.42 (9.59)	37.10 (9.35)	38.45 (10.25)	37.55 (9.10)	37.47 (9.15)	37.83 (8.95)
BMI (kg/m^2), mean (SD)	31.25 (8.16)	31.42 (8.16)	30.72 (8.18)	30.64 (7.51)	30.64 (7.69)	30.61 (6.89)
Region, n (%)						
Midwest	291 (34.6)	206 (32.2)	85 (42.3)	299 (35.3)	216 (32.9)	83 (43.5)
Northeast	139 (16.5)	109 (17.1)	30 (14.9)	134 (15.8)	99 (15.1)	35 (18.3)
South	257 (30.6)	198 (31)	59 (29.4)	239 (28.2)	191 (29.1)	48 (25.1)
West	153 (18.2)	126 (19.7)	27 (13.4)	176 (20.8)	151 (23)	25 (13.1)
Gender, n (%)						
Female	764 (91)	586 (91.7)	178 (88.6)	660 (77.8)	519 (79)	141 (73.8)
Male	71 (8.4)	48 (7.5)	23 (11.4)	186 (21.9)	136 (20.7)	50 (26.2)
Nonbinary	5 (0.6)	5 (0.8)	0 (0)	2 (0.2)	2 (0.3)	0 (0)
Race, n (%)						
Alaska Native, American Indian, Native Hawaiian, or other Pacific Islander	2 (0.2)	2 (0.3)	0 (0)	1 (0.1)	1 (0.2)	0 (0)
Asian	16 (1.9)	13 (2)	3 (1.5)	27 (3.2)	19 (2.9)	8 (4.2)
Black or African American	31 (3.7)	25 (3.9)	6 (3)	21 (2.5)	14 (2.1)	7 (3.7)
White	753 (89.6)	568 (88.9)	185 (92)	763 (90)	593 (90.3)	170 (89)
Multiple races	32 (3.8)	27 (4.2)	5 (2.5)	28 (3.3)	23 (3.5)	5 (2.6)
Other	6 (0.7)	4 (0.6)	2 (1)	8 (0.9)	7 (1.1)	1 (0.5)

[a]ILI: influenza-like illness.

[b]HTRI: Home Testing of Respiratory Illness.

Assessment of the XGBoost Model for Influenza Prediction During ILI Days –4 to +1

XGBoost models informed by symptom-only data, activity-only data, or a combination of both symptoms and activity data were evaluated across training, validation, and test sets for FluStudy 2020. ROC curves and stratified k-fold cross-validation analyses for all models are presented in Figure 1, with confusion matrices shown in Figure S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Receiver operating characteristic (ROC) curves for extreme gradient boosting (XGBoost) model discrimination between participants who were influenza positive and participants who were influenza negative for FluStudy2020 and Home Testing of Respiratory Illness (HTRI) data. XGBoost model performance was assessed for symptom-only data, activity-only data, and a combination of symptom and activity data. The mean performance across each k-fold and 95% CI for the training, validation, and test sets are presented. Mean values + or – the margin of error are shown for area under the curve (AUC), balanced accuracy (BA), sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), and weighted harmonic mean of precision and recall (F2). The red line represents random guess, and the blue line represents mean ROC +95% and –95% CI.

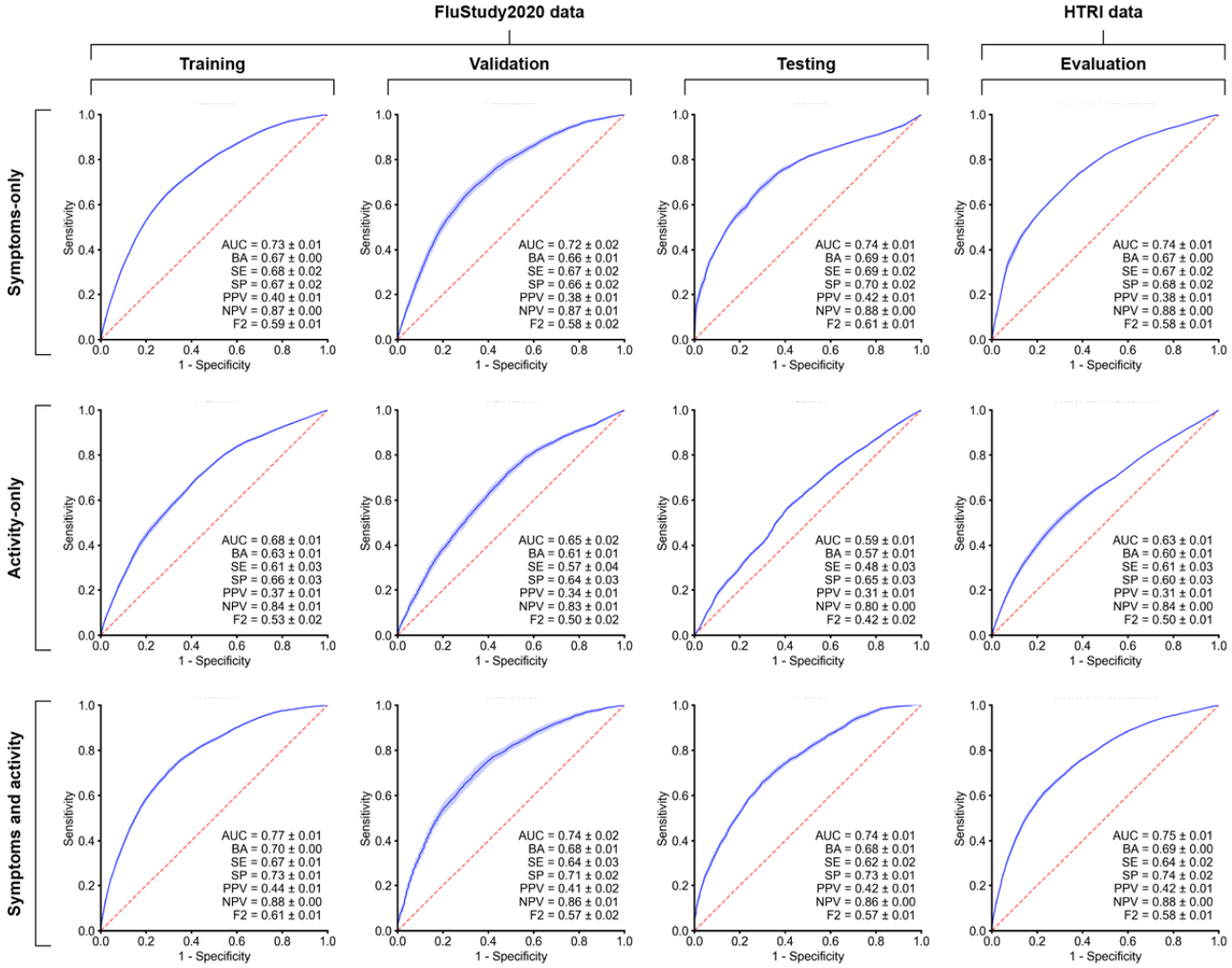

For the training and validation sets, the model trained on the combined symptom and activity data (training area under the curve [AUC]=0.77; validation AUC=0.74) consistently outperformed the models trained on the symptom-only data (training AUC=0.73; validation AUC=0.72) and activity-only data (training AUC=0.68; validation AUC=0.65; Figure 1). When applied to the FluStudy2020 test set, the model performance with the combined symptom and activity data was closely aligned with that of the symptom-only data (combined symptom and activity test AUC=0.74; symptom-only test AUC=0.74). We extended our evaluation to the HTRI study, where the model trained on combined symptom and activity data (evaluation AUC=0.75) outperformed the model trained on the symptom-only data (evaluation AUC=0.74), confirming the results of the FluStudy2020 training and validation sets.

Feature importance plots for each model are presented in Figure 2. For the combined symptom and activity model, cough, mean RHR during main sleep, fever, and total minutes in bed were the most important, with mean feature importance values of 0.21, 0.17, 0.15, and 0.15, respectively. For the symptom-only model, fever, cough, and sore throat were the most important, with mean feature importance values of 0.36, 0.33, and 0.10, respectively. The heart low-frequency/high-frequency ratio, total minutes in bed, mean RHR during main sleep, and heart normalized low-frequency power were the top features influencing activity-only model predictions, with mean feature importance values of 0.34, 0.17, 0.15, and 0.15, respectively (Figure 2). Calibration plots highlighting the degree of correspondence between the estimated probability of influenza-positive cases and observed influenza cases for each model are presented in Figure 3.

XSL•FO
RenderX

Figure 2. Feature importance plots for symptom-only data (A), activity-only data (B), and a combination of symptom and activity data (C). Values are presented as mean (95% CI). HR: heart rate; HRV: heart rate variability; RHR: resting heart rate; RR interval: the time elapsed between 2 successive R-waves of the QRS signal on the electrocardiogram.

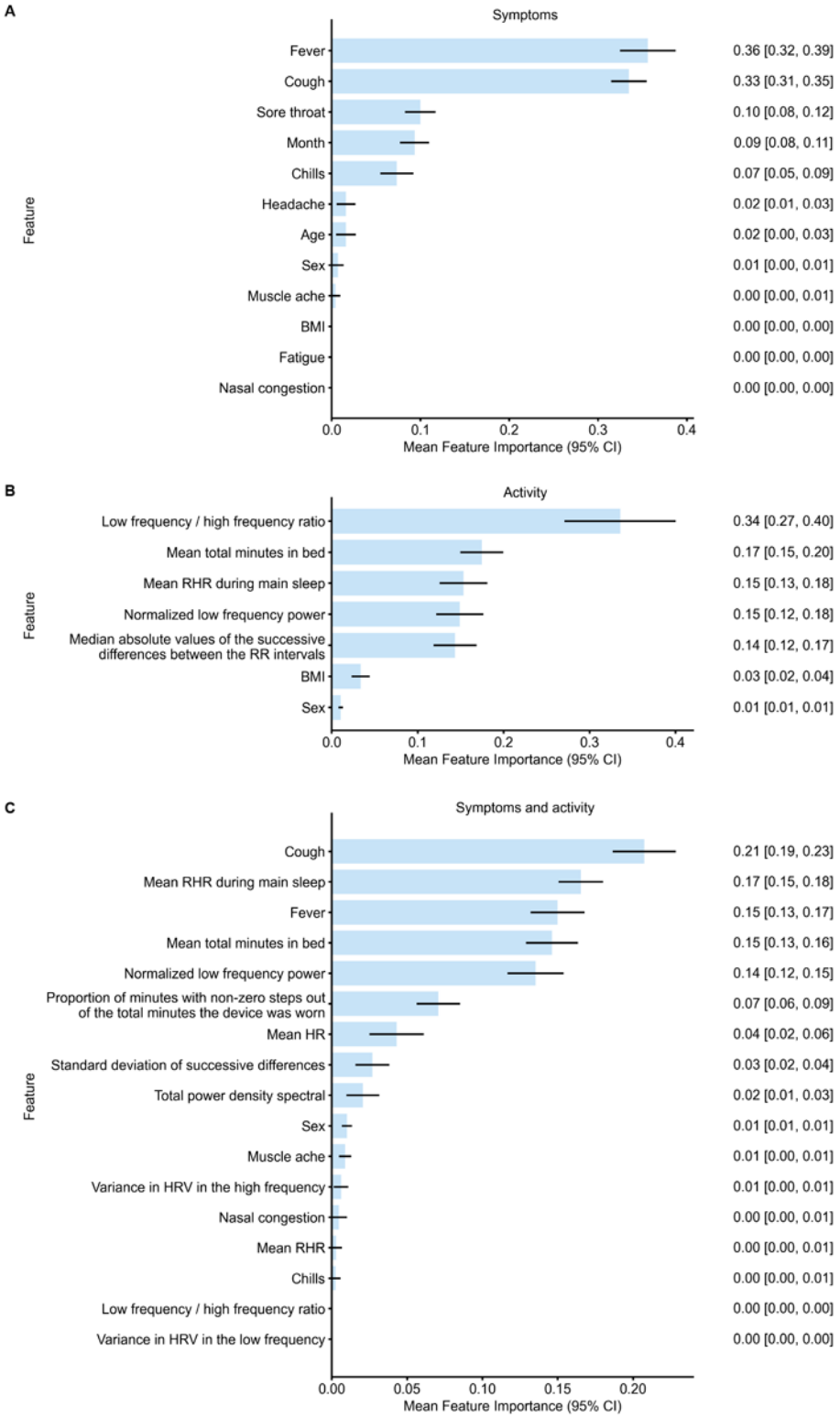

XSL•FO
RenderX

Figure 3. Calibration plots for FluStudy2020 and Home Testing of Respiratory Illness (HTRI) data, as assessed using symptom-only data, activity-only data, and a combination of symptom and activity data. The red dashed line represents perfect calibration, and the blue dots represent mean calibration +95% and –95% CI.

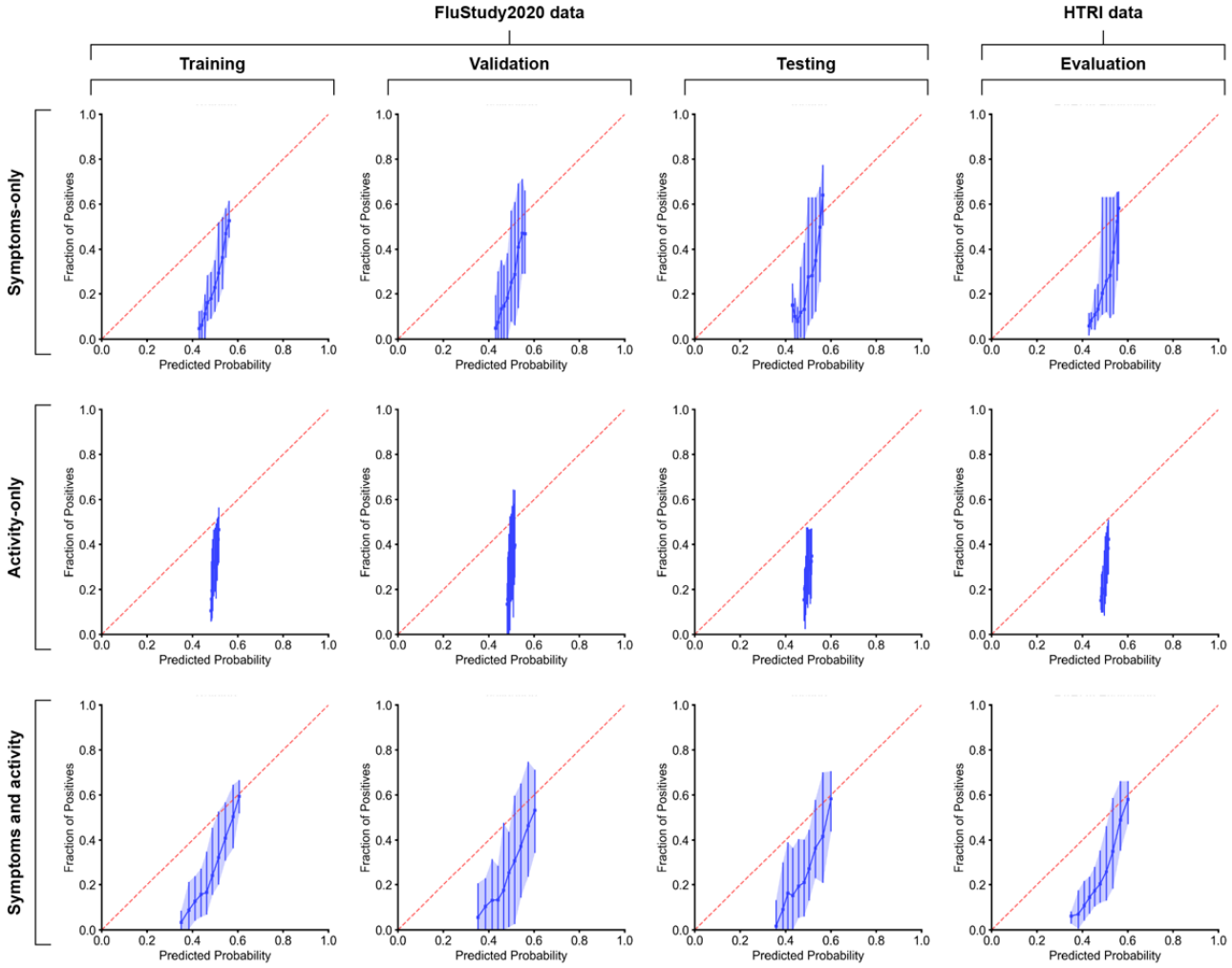

Discussion

Principal Findings

To our knowledge, this is the largest study of commercial wearable sensors for the early detection of influenza incorporating the virological confirmation of influenza infection. The study was specifically designed to test, in the real world, the hypothesis generated in experimental settings that wearable sensor data may predict the onset of viral respiratory infection. For the combined symptom and activity model, the most important variables were cough, mean RHR during main sleep, fever, and total minutes in bed; for the symptom-only model, the most important variables were fever, cough, and sore throat; and for the activity-only model, the most important variables were heart low-frequency/high-frequency ratio, total minutes in bed, mean RHR during main sleep, and heart normalized low-frequency power. The best-performing machine-learning model for influenza detection was trained on the combined symptom and activity data and had a mean training AUC of 0.77. Model performance was validated on an independent data set (HTRI) not used for training, which yielded a mean AUC of 0.75. The accuracy of the combined machine-learning model was further confirmed by calibration plots for combined symptom and activity data, which were well calibrated compared with symptom-only or activity-only plots. Our model performance is significantly lower than that in other studies using machine-learning algorithms to predict influenza infection using wearable sensor data, which achieved accuracies of up to 94% [4,11]. However, these results were from small cohort (n=31 and n=20) challenge studies, where participants used research-grade wearable sensors and remained in controlled environments for up to 10 days following challenge with either influenza A virus subtype H1N1 or influenza A virus subtype H3N2 [4,11]. Grzesiak et al [4] noted that model accuracy was associated with both the knowledge of the timing and dosage of inoculation and the high-fidelity measurements of research-grade sensors. Our results suggest that findings from research-grade sensors tested in a highly controlled experimental setting may not easily translate to scalable low-fidelity commercial-grade sensors deployed in the real world.

With the FluStudy2020 training and validation sets, the best-performing model for influenza detection used a combination of symptom and activity features. However, in the FluStudy2020 test set, model performance was similar between the combined symptom and activity data model and the symptom-only data model. This implies the activity features do not significantly improve the model performance. In contrast,

XSL•FO
RenderX

Quer et al [5], using similar methods to discriminate between symptomatic individuals who were positive or negative for COVID-19, found that a model combining symptom and sensor data performed significantly better than one considering symptoms alone (AUC 0.80, IQR 0.73-0.86 vs AUC 0.71, IQR 0.63-0.79). The different results observed may simply reflect differences between influenza and COVID-19, which have several nonoverlapping symptoms; notably, their model included data from onset day to day 7, whereas our model included data from day –4 to day 1. Our model restricted the data period, as the clinical utility of wearable sensors as an early warning tool for influenza would depend on their ability to detect infection early in its course, when the individual could take action to limit the spread or seek medical attention.

A key strength of our study is the laboratory confirmation of influenza in symptomatic patients using a highly accurate reverse transcription polymerase chain reaction test, which provides an accurate ascertainment of true positives and true negatives. Another strength of our study is that demographic, clinical, and Fitbit device data from a large, real-world population of >800 participants were used in model development. For the combined model (symptom and activity data), cough, total minutes in bed, and mean RHR during main sleep were the top 3 features influencing model predictions. Of the top 17 most important features influencing model predictions, 9 (53%) were HRV metrics. Deviations in HRV metrics have been associated with infection status and the severity of various bacterial and viral illnesses [4,7,10-12,22]. In addition, Hirten et al [22] showed that the mean of the SDs of normal-to-normal interval (the mean amplitude of the circadian pattern of the SD of the interbeat interval of normal sinus beats) was associated with a COVID-19 diagnosis, irrespective of symptomatology. Another study demonstrated that HRV acrophase and HRV midline estimating statistic of rhythm (MESOR) were among the most important predictors of COVID-19 infection, along with age and BMI [8]. Future analyses should consider the impact of biological and lifestyle factors, such as sex, menstrual cycle, and alcohol consumption, on HRV and other physiological features [9,10,38,39].

Limitations

Limitations pertaining to the HTRI and FluStudy2020 study design have been discussed previously [15]. Notably, the very small numbers of African American participants, Asian participants, male participants, and participants aged ≥65 years in this cohort limit the generalizability of the model. The imbalances may be the result of differences in the likelihood of these populations to engage with digital health services; for example, women have been found to be more likely to use a mobile health app than men [40]. Participants were required to own a Fitbit device, which may predispose this cohort to exhibiting increased levels of activity and more health-conscious behaviors than the general population, which could limit the generalizability of the activity-based predictive models. A single device type (Fitbit) was used to minimize measurement errors that could arise from the use of multiple device types in study participants. However, this limits the generalizability of the findings, as several other device types are in widespread use. The self-reporting of symptoms in both studies is subjective and prone to recall bias. However, the results of influenza tests performed as part of the study were not provided to participants, which could have led to a differential recall of symptoms between participants who were influenza positive and influenza negative. Nevertheless, we cannot rule out participants' awareness of their disease status through seeking routine care for their ILI outside of the study. In addition, symptom data were collected daily to minimize the risk of incomplete or inaccurate recall.

The studies included in this analysis were designed before the COVID-19 pandemic but were ongoing until October 2020. COVID-19 mitigation measures such as lockdown procedures may have impacted participants' regular activities and influenza circulation during the period of these studies. Further implications of the COVID-19 pandemic have been discussed previously [15].

While our previous work demonstrates that the amplitude of wearable sensor deviations differs significantly between individuals who are influenza positive and those with ILI symptoms only, the symptom and activity features used in model development in this study are not unique to influenza infection [15]. Strict symptom criteria were used to define the symptomatic population with ILI, which may have led to the selection of a more severe symptomatic population and limited the ability to discriminate between participants who were influenza positive and participants who were influenza negative. Future studies with different study designs and less restrictive symptom eligibility criteria should investigate the ability of machine-learning algorithms to discriminate among a range of other common respiratory viral infections using symptomatic and wearable sensor data.

Finally, the validity and reliability of commercial wearable sensors in the measurement of steps, sleep, and HR have been a subject of debate. A systematic review including >150 publications found that Fitbit HR measurements were variable and tended toward underestimating HR [41]. The wearable devices used by participants in our study measured only steps, sleep, and HR. Future studies should investigate whether more advanced wearable sensors with more accurate accelerometers and including additional physiological measures, such as skin temperature and blood oxygen saturation, could improve the performance of commercial-grade sensors in the early detection and discrimination of respiratory viral infections.

Conclusions

We demonstrate that a machine-learning algorithm combining symptomatic and commercial wearable sensor data during the latent and early symptomatic phases of ILI had moderate accuracy in detecting influenza in a large real-world cohort of symptomatic individuals with ILI, suggesting that previous findings from research-grade sensors tested in highly controlled experimental settings may not easily translate to scalable commercial-grade sensors deployed in the real world. The model maintained consistent performance across 2 distinct studies. The model was initially trained and evaluated on FluStudy2020 data and achieved comparable performance when validated on the HTRI data, affirming its generalizability. If machine-learning algorithms using commercial wearable sensors had strong

XSL•FO
RenderX

predictive power and were validated, they may potentially play a role in public health surveillance and could prompt users to adopt infection-control behavior (eg, self-quarantine) and to seek early medical attention, if necessary. In the future, more advanced wearables measuring additional physiological parameters may improve the performance of wearable sensors in the early detection and discrimination of viral respiratory infections.

Acknowledgments

The FluStudy2020 was supported by F. Hoffmann-La Roche Ltd, Genentech, Inc, and Evidation Health Inc. The Home Testing of Respiratory Illness (HTRI) study was supported by Biomedical Advanced Research and Development Authority (contract number 75A50119C00036), part of the Office of the Assistant Secretary for Preparedness and Response, US Department of Health and Human Services, and Audere, a nonprofit digital health organization. The HTRI study was designed and conducted by Evidation Health Inc. F. Hoffmann-La Roche Ltd was involved in the design and conduct of FluStudy2020; collection, management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; and the decision to submit the manuscript for publication.

These data were contributed by participants as part of the HTRI study developed by Evidation Health, Inc and described in Synapse (ID syn22803188). Support with the acquisition, analysis, and interpretation of the data was provided by Konrad Mysliwiec (Roche Global IT Solution Centre). Support with obtaining funding was provided by authors KF and MP, as well as Ahmed Ansari (F. Hoffmann-La Roche Ltd) and James Harper (Roche Products Ltd). Third-party medical writing assistance, under the direction of the authors, was provided by Stephanie Cumberworth, PhD, and Edmund Harratt, BA, on behalf of Ashfield MedComms, an Inizio company, and funded by F. Hoffmann-La Roche Ltd.

Data Availability

The Home Testing of Respiratory Illness (HTRI) data sets are available in the HTRI study research portal [42]. The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ML and KF had full access to the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. KF, ML, AH-R, AHO, LD-H, FJ, VU, DC, VH, MP, and BC conceived and designed the study. KF, ML, AH-R, AHO, LD-H, FJ, KN, MMAZ, DC, VH, BC, and VU conducted the acquisition, analysis, and interpretation of data. LD-H, ML, AH-R, AHO, and FJ conducted statistical analysis. KF and MP obtained funding for the study. KF and ML provided administrative, technical, and material support. ML, KF, and VU supervised the writing of the manuscript. All authors contributed to the drafting of the manuscript and conducted critical revision for important intellectual content.

Conflicts of Interest

ML, VH, and DC are current or former employees of Genentech, Inc, a member of the Roche Group. LD-H, FJ, BC, and VU are current or former employees of Roche Products Ltd. KF, AHO, AH-R, and MP are employees of F. Hoffmann-La Roche Ltd. MMAZ is an employee of Roche Services (Asia Pacific) Sdn. Bhd. KN is an employee of Badger Software Sp. z o.o. Badger Software Sp. z o.o. received funding from F. Hoffmann-La Roche Ltd for the conduct of this study but was not paid for the development of the manuscript.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 520 KB - jmir_v26i1e47879_app1.docx]

References

1. Disease burden of flu. US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/flu/about/burden/index.html [accessed 2022-11-02]
2. D'Haese PF, Finomore V, Lesnik D, Kornhauser L, Schaefer T, Konrad PE, et al. Prediction of viral symptoms using wearable technology and artificial intelligence: a pilot study in healthcare workers. PLoS One 2021 Oct 14;16(10):e0257997 [FREE Full text] [doi: 10.1371/journal.pone.0257997] [Medline: 34648513]
3. Gadaleta M, Radin JM, Baca-Motes K, Ramos E, Kheterpal V, Topol EJ, et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. NPJ Digit Med 2021 Dec 08;4(1):166 [FREE Full text] [doi: 10.1038/s41746-021-00533-1] [Medline: 34880366]
4. Grzesiak E, Bent B, McClain MT, Woods CW, Tsalik EL, Nicholson BP, et al. Assessment of the feasibility of using noninvasive wearable biometric monitoring sensors to detect influenza and the common cold before symptom onset. JAMA Netw Open 2021 Sep 01;4(9):e2128534 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.28534] [Medline: 34586364]

XSL•FO
RenderX

5. Quer G, Radin JM, Gadaleta M, Baca-Motes K, Ariniello L, Ramos E, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021 Jan;27(1):73-77. [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
6. ACTIV-3/TICO LY-CoV555 Study Group, Lundgren JD, Grund B, Barkauskas CE, Holland TL, Gottlieb RL, et al. A neutralizing monoclonal antibody for hospitalized patients with COVID-19. N Engl J Med 2021 Mar 11;384(10):905-914 [FREE Full text] [doi: 10.1056/NEJMoa2033130] [Medline: 33356051]
7. Ahmad S, Tejuja A, Newman KD, Zarychanski R, Seely AJ. Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection. Crit Care 2009 Nov 24;13(6):232 [FREE Full text] [doi: 10.1186/cc8132] [Medline: 20017889]
8. Hirten RP, Tomalin L, Danieletto M, Golden E, Zweig M, Kaur S, et al. Evaluation of a machine learning approach utilizing wearable data for prediction of SARS-CoV-2 infection in healthcare workers. JAMIA Open 2022 May 18;5(2):ooac041. [doi: 10.1093/jamiaopen/ooac041] [Medline: 35677186]
9. Mitratza M, Goodale BM, Shagadatova A, Kovacevic V, van de Wijgert J, Brakenhoff TB, et al. The performance of wearable sensors in the detection of SARS-CoV-2 infection: a systematic review. Lancet Digit Health 2022 May;4(5):e370-e383 [FREE Full text] [doi: 10.1016/S2589-7500(22)00019-X] [Medline: 35461692]
10. Radin JM, Quer G, Pandit JA, Gadaleta M, Baca-Motes K, Ramos E, et al. Sensor-based surveillance for digitising real-time COVID-19 tracking in the USA (DETECT): a multivariable, population-based, modelling study. Lancet Digit Health 2022 Nov;4(11):e777-e786 [FREE Full text] [doi: 10.1016/S2589-7500(22)00156-X] [Medline: 36154810]
11. Temple DS, Hegarty-Craver M, Furberg RD, Preble EA, Bergstrom E, Gardener Z, et al. Wearable sensor-based detection of influenza in presymptomatic and asymptomatic individuals. J Infect Dis 2023 Apr 12;227(7):864-872 [FREE Full text] [doi: 10.1093/infdis/jiac262] [Medline: 35759279]
12. Shandhi MM, Cho PJ, Roghanizad AR, Singh K, Wang W, Enache OM, et al. A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19. NPJ Digit Med 2022 Sep 01;5(1):130 [FREE Full text] [doi: 10.1038/s41746-022-00672-z] [Medline: 36050372]
13. Dhingra LS, Aminorroaya A, Oikonomou EK, Nargesi AA, Wilson FP, Krumholz HM, et al. Use of wearable devices in individuals with or at risk for cardiovascular disease in the US, 2019 to 2020. JAMA Netw Open 2023 Jun 01;6(6):e2316634 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.16634] [Medline: 37285157]
14. Chandrasekaran R, Katthula V, Moustakas E. Patterns of use and key predictors for the use of wearable health care devices by US adults: insights from a national survey. J Med Internet Res 2020 Oct 16;22(10):e22443 [FREE Full text] [doi: 10.2196/22443] [Medline: 33064083]
15. Hunter V, Shapiro A, Chawla D, Drawnel F, Ramirez E, Phillips E, et al. Characterization of influenza-like illness burden using commercial wearable sensor data and patient-reported outcomes: mixed methods cohort study. J Med Internet Res 2023 Mar 23;25:e41050 [FREE Full text] [doi: 10.2196/41050] [Medline: 36951890]
16. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minin. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA p. 785-794 URL: https://dl.acm.org/doi/10.1145/2939672.2939785 [doi: 10.1145/2939672.2939785]
17. XGBoost documentation. XGBoost developers. URL: https://xgboost.readthedocs.io/en/stable/ [accessed 2022-11-02]
18. Mezlini A, Shapiro A, Daza EJ, Caddigan E, Ramirez E, Althoff T, et al. Estimating the burden of influenza-like illness on daily activity at the population scale using commercial wearable sensors. JAMA Netw Open 2022 May 02;5(5):e2211958 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.11958] [Medline: 35552722]
19. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health 2020 Feb;2(2):e85-e93 [FREE Full text] [doi: 10.1016/S2589-7500(19)30222-5] [Medline: 33334565]
20. Mishra T, Wang M, Metwally AA, Bogu GK, Brooks AW, Bahmani A, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec;4(12):1208-1220 [FREE Full text] [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
21. Natarajan A, Su HW, Heneghan C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. NPJ Digit Med 2020 Nov 30;3(1):156 [FREE Full text] [doi: 10.1038/s41746-020-00363-7] [Medline: 33299095]
22. Hirten RP, Danieletto M, Tomalin L, Choi KH, Zweig M, Golden E, et al. Use of physiological data from a wearable device to identify SARS-CoV-2 infection and symptoms and predict COVID-19 diagnosis: observational study. J Med Internet Res 2021 Feb 22;23(2):e26107 [FREE Full text] [doi: 10.2196/26107] [Medline: 33529156]
23. Ates HC, Yetisen AK, Güder F, Dincer C. Wearable devices for the detection of COVID-19. Nat Electron 2021 Jan 25;4(1):13-14. [doi: 10.1038/s41928-020-00533-1]
24. Cleary JL, Fang Y, Sen S, Wu Z. A caveat to using wearable sensor data for COVID-19 detection: the role of behavioral change after receipt of test results. PLoS One 2022 Dec 30;17(12):e0277350 [FREE Full text] [doi: 10.1371/journal.pone.0277350] [Medline: 36584148]
25. Shapiro A, Marinsek N, Clay I, Bradshaw B, Ramirez E, Min J, et al. Characterizing COVID-19 and influenza illnesses in the real world via person-generated health data. Patterns (N Y) 2021 Jan 08;2(1):100188 [FREE Full text] [doi: 10.1016/j.patter.2020.100188] [Medline: 33506230]

XSL•FO
RenderX

26. Mekhael M, Lim CH, El Hajjar AH, Noujaim C, Pottle C, Makan N, et al. Studying the effect of long COVID-19 infection on sleep quality using wearable health devices: observational study. J Med Internet Res 2022 Jul 05;24(7):e38000 [FREE Full text] [doi: 10.2196/38000] [Medline: 35731968]
27. Kotnik JH, Cooper S, Smedinghoff S, Gade P, Scherer K, Maier M, et al. Flu@home: the comparative accuracy of an at-home influenza rapid diagnostic test using a prepositioned test kit, mobile app, mail-in reference sample, and symptom-based testing trigger. J Clin Microbiol 2022 Mar 16;60(3):e0207021 [FREE Full text] [doi: 10.1128/JCM.02070-21] [Medline: 35107302]
28. Saint-Maurice PF, Troiano RP, Bassett Jr DR, Graubard BI, Carlson SA, Shiroma EJ, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160 [FREE Full text] [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
29. Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, et al. Association of step counts over time with the risk of chronic disease in the All of Us Research Program. Nat Med 2022 Nov;28(11):2301-2308 [FREE Full text] [doi: 10.1038/s41591-022-02012-w] [Medline: 36216933]
30. McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SCIPY '10; June 28-July 3, 2010; Austin, TX p. 56-61 URL: http://conference.scipy.org.s3-website-us-east-1.amazonaws.com/proceedings/scipy2010/pdfs/mckinney.pdf [doi: 10.25080/majora-92bf1922-00a]
31. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 Sep;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
32. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007 Jun 18;9(3):90-95. [doi: 10.1109/mcse.2007.55]
33. Alam S, Chan N, Couto L, Dada Y, Danov I, Datta D. Kedro (version 0.18.12). GitHub. URL: https://github.com/kedro-org/kedro [accessed 2023-09-28]
34. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Soft 2010 Sep 16;36(11):1-13. [doi: 10.18637/jss.v036.i11]
35. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011 Oct;12:2825-2830 [FREE Full text]
37. Champseix R, Ribiere L, Le Couedic C. A Python package for heart rate variability analysis and signal preprocessing. J Open Res Softw 2021 Oct 06;9(1):28. [doi: 10.5334/jors.305]
38. Goodale BM, Shilaih M, Falco L, Dammeier F, Hamvas G, Leeners B. Wearable sensors reveal menses-driven changes in physiology and enable prediction of the fertile window: observational study. J Med Internet Res 2019 Apr 18;21(4):e13404 [FREE Full text] [doi: 10.2196/13404] [Medline: 30998226]
39. Zhu G, Li J, Meng Z, Yu Y, Li Y, Tang X, et al. Learning from large-scale wearable device data for predicting the epidemic trend of COVID-19. Discrete Dyn Nat Soc 2020 May 05:1-8. [doi: 10.1155/2020/6152041]
40. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018 Mar 15;6(3):e65 [FREE Full text] [doi: 10.2196/mhealth.9131] [Medline: 29549071]
41. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
42. Home testing of respiratory illness study research portal. Evidation. 2020. URL: https://www.synapse.org/#!Synapse:syn22803188/wiki/606343 [accessed 2023-09-27]

Abbreviations

AUC ROC: area under the receiver operating characteristic curve
AUC: area under the curve
HR: heart rate
HRV: heart rate variability
HTRI: Home Testing of Respiratory Illness
ILI: influenza-like illness
MESOR: midline estimating statistic of rhythm
RHR: resting heart rate
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 05.04.23; peer-reviewed by J Wang, M Mekhael, K Singh; comments to author 06.07.23; revised version received 01.11.23; accepted 03.07.24; published 04.10.24.

Please cite as:
Farooq K, Lim M, Dennison-Hall L, Janson F, Olszewska AH, Ahmad Zabidi MM, Haratym-Rojek A, Narowski K, Clinch B, Prunotto M, Chawla D, Hunter V, Ukachukwu V
Evaluation of Machine Learning to Detect Influenza Using Wearable Sensor Data and Patient-Reported Symptoms: Cohort Study
J Med Internet Res 2024;26:e47879
URL: https://www.jmir.org/2024/1/e47879
doi:10.2196/47879
PMID:39365646

©Kamran Farooq, Melody Lim, Lawrence Dennison-Hall, Finn Janson, Aspen Hazel Olszewska, Muhammad Mamduh Ahmad Zabidi, Anna Haratym-Rojek, Karol Narowski, Barry Clinch, Marco Prunotto, Devika Chawla, Victoria Hunter, Vincent Ukachukwu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile Phone Syndromic Surveillance for Respiratory Conditions in an Emergency (COVID-19) Context in Colombia: Representative Survey Design

Andres I Vecino-Ortiz[1*], MD, PhD; Deivis Nicolas Guzman-Tordecilla[1*], RN, MPH; Vidhi Maniar[1], MPH; Sandra Agudelo-Londoño[2*], PhD; Oscar Franco-Suarez[2*], MPH; Nathaly Aya Pastrana[3*], PhD; Mariana Rodríguez-Patarroyo[2], JD, PhD; Marino Mejía-Rocha[2*], MSc; Jaime Cardona[4*], MSc; Mariangela Chavez Chamorro[4*], MSc; Dustin Gibson[1*], MPH, PhD

[1]Health Systems Program, International Health Department, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[2]Pontificia Universidad Javeriana, Bogota, Colombia
[3]IMEK, Cali, Colombia
[4]Inter-American Development Bank, Bogota, Colombia
[*]these authors contributed equally

Corresponding Author:
Andres I Vecino-Ortiz, MD, PhD
Health Systems Program
International Health Department
Johns Hopkins Bloomberg School of Public Health
615 N Wolfe Street
Baltimore, MD, 21205
United States
Phone: 1 410 955 3934
Email: andres.vecino@gmail.com

Abstract

Background: Syndromic surveillance for respiratory infections such as COVID-19 is a crucial part of the public health surveillance toolkit as it allows decision makers to detect and prepare for new waves of the disease in advance. However, it is labor-intensive, costly, and increases exposure to survey personnel. This study assesses the feasibility of conducting a mobile phone–based respiratory syndromic surveillance program in a middle-income country during a public health emergency, providing data to support the inclusion of this method in the standard infection control protocols at the population level.

Objective: This study aims to assess the feasibility of a national active syndromic surveillance system for COVID-19 disease in Colombia.

Methods: In total, 2 pilots of syndromic mobile phone surveys (MPSs) were deployed using interactive voice response technology in Colombia (367 complete surveys in March 2022 and 451 complete surveys in April and May 2022). Respondents aged 18 years and older were sampled using random digit dialing, and after obtaining consent, they were sent a 10-minute survey with modules on sociodemographic status, respiratory symptoms, past exposure to COVID-19 infection and vaccination status, preferences about COVID-19 vaccination, and information source for COVID-19. Pilot 1 used a nationally representative sample while pilot 2 used quota sampling to yield representative results at the regional level. In this work, we assessed the performance characteristics of the survey pilots and compared the demographic information collected with a nationally representative household survey.

Results: For both pilots, contact rates were between 1% and 2%, while participation rates were above 80%. The results revealed that younger, female, and higher educated participants were more likely to participate in the syndromic survey. Survey rates as well as demographics, COVID-19 vaccination status, and prevalence of respiratory symptoms are reported for both pilots. We found that respondents of the MPSs are more likely to be younger and female.

Conclusions: In a COVID-19 pandemic setting, using an interactive voice response MPS to conduct syndromic surveillance may be a transformational, low-risk, and feasible method to detect outbreaks. This evaluation expects to provide a path forward to the inclusion of MPSs as a traditional surveillance method.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50184) doi:10.2196/50184

KEYWORDS

mobile phone surveys; syndromic surveillance; COVID-19; public health surveillance; IVR; interactive voice response; survey; surveys; voice response; syndromic; surveillance; respiratory; pandemic; SARS-CoV-2; feasibility; data collection; public health; emergency; outbreak; mobile phone

Introduction

Surveillance systems that provide real-time and accurate information are increasingly critical for decision-making in public health, as demonstrated by the COVID-19 pandemic [1]. Commonly used indicators for monitoring the COVID-19 pandemic were the number of laboratory-confirmed cases, the number of persons hospitalized or in intensive care, and the number of COVID-19 deaths [2,3]. However, these indicators are highly dependent on the health systems' capacity, might fail to represent real-time conditions, are often mediated by other factors such as health care access, and are less likely to detect less severe cases [4]. These challenges become more important in low- and middle-income countries (LMICs) where traditional surveillance tools are scarce. For this reason, new strategies to complement traditional epidemiologic surveillance are needed [5]. The World Health Organization (WHO) has recommended enhancing traditional epidemiologic surveillance systems with other tools such as syndromic surveillance, where individuals can self-report symptoms related to infection [5], improving the timeliness and coverage, at lower costs [6-9].

Mobile phones have the potential to be an efficient tool to perform syndromic surveillance [1,4,10] because of their widespread use. For example, in Colombia, it has been estimated that there are 133 mobile telephone subscriptions for every 100 people [11]. In addition, it has been identified that surveys using mobile phone technology have lower costs than household surveys [9]. Finally, mobile phone syndromic surveillance is easy to deploy in emergency contexts where social distancing is required.

Mobile phone–based syndromic surveillance systems used during the COVID-19 pandemic have been mainly passive and limited to mobile phone apps that require smartphone technology linked to contact-tracing apps [1,4,10,12-14], which could be limited in LMIC where the number of smartphones with internet access is limited [10]. Other studies in LMIC have used the interactive voice response (IVR) as an active surveillance system on behavior, exposure, knowledge, and perception related to COVID-19, but not as a syndromic surveillance system [6,15,16]. To our knowledge, there have been no published experiences of an active syndromic surveillance system on COVID-19 disease or other health emergencies using IVR as a data collection tool. We hypothesize that syndromic surveillance is a feasible option to conduct syndromic surveillance in respiratory emergencies. This study aims to assess the feasibility of a national active syndromic surveillance system for COVID-19 disease in Colombia.

Methods

Data and Surveillance Instrument

IVR surveys were developed and cognitive testing was performed [17]. Participant phone numbers were obtained through random digit dialing [18], which included a prefix ranging from 300 to 323 (which are the prefixes for all mobile phone numbers used in Colombia), followed by 7 digits randomly selected. Random digit dialing is a technique used in previous research to yield a random sample of phone numbers when a previous database does not exist or cannot be accessed for privacy reasons [19]. Respondents who were aged ≥18 years and provided consent, were considered eligible and could participate in a 10-minute survey that comprised the following topical modules: sociodemographic status, respiratory symptoms, previous COVID-19 infection and vaccination status, preferences about COVID-19 vaccination and information on COVID-19. On survey completion, respondents were given an airtime incentive of COL $4000 (around US $1) [20].

Sampling

After testing, we deployed 2 pilots of IVR syndromic surveys with different levels of representativeness. The objective of the first pilot was to assess the feasibility of the IVR system to carry out syndromic surveys, as well as determine population profiles at the national level more likely to respond to this tool. The second pilot was aimed at determining whether quota sampling was effective in improving population representativeness for less populated regions. Following previous work [21-23], the 5 different regions were defined as follows: Caribbean (including Córdoba, Sucre, Bolívar, Magdalena, Atlántico, La Guajira, Cesar, and San Andrés y Providencia) with around 12 million inhabitants, Pacific region (including Chocó, Valle del Cauca, Cauca, and Nariño) with a population around 8.5 million, Amazonas river region (including Amazonas, Putumayo, Caquetá, Vaupés, Guaviare, and Guainía) with about 1.1 million population, Orinoco River region (including Meta, Vichada, Casanare, and Arauca) with 2 million population, and the Central region (including Bogotá, Cundinamarca, Huila, Tolima, Boyacá, Santander, Norte de Santander, Caldas, Risaralda, Antioquia, and Quindío) with around 29 million population [24].

The sample size for the first pilot was calculated using the STEPS (WHO Stepwise Approach to Noncommunicable Disease Risk Factor Surveillance) survey calculator as used in previous research [9,16,25,26]. The STEPS survey is a recommended calculation for repeated cross-sectional, population-based household surveys (see equation 1). The assumptions for the sample size are a 95% CI (z score=1.96), margin of error of 0.05, and baseline prevalence of 0.34 [27],

XSL•FO
RenderX

yielding 345 complete surveys to be deployed to have nationally representative surveys.

The second pilot instead focused on producing values that are representative of 5 regions in Colombia through automated strata sampling [28]. To achieve this, we recalculated the sample size with the prevalence of symptoms obtained from the previous survey (0.06), yielding a total of 87 complete surveys for each of the 5 regions, keeping all other parameters the same.

The deployment of the surveys took place between March 12, 2022, and March 15, 2022, and April 23, 2022, and May 13, 2022, respectively. The second pilot took longer because of the quota sampling, where many calls were deemed ineligible if the quota for that region had been already filled.

Outcomes—Disposition Codes

Disposition codes were assessed using the following nomenclature. I stands for *complete interview*. The respondents were aged ≥18 years, consented, and answered all applicable questions. P stands for incomplete interview. The respondents were aged ≥18 years, consented, answered demographics questions, and some questions from the other modules, but did not finish the information on the COVID-19 module (did not reach the airtime incentive question). B stands for break-off the interview. The respondents were aged ≥18 years, consented, and answered the demographics questions. R stands for refuse after listening to the consent. U stands for unknown as these were phone numbers that did not answer or did answer but did not reach the consent and their eligibility status could not be ascertained. Respondents can be deemed ineligible for three reasons: (1) line not registered, (2) they might report being aged younger than 18 years, or (3) in pilot 2 they might be from a region whose quota sample was already met [29].

Survey rates are defined as follows: (1) contact rate measures the proportion of all cases in which some eligible mobile phone user was reached by the survey (I+P+B+R/I+P+B+R+U). (2) Response rate measures the proportion of both complete and incomplete interviews divided by all known eligible mobile phone users and users with unknown eligibility (I+P/I+P+B+R+U). (3) Refusal rate measures the proportion of mobile phone users that do not consent or break off the survey after consent divided by all known eligible users and users with unknown eligibility (B+R/I+P+B+R+U). (4) Cooperation rate: measures the proportion of complete interviews divided by all known eligible mobile phone users. (I/I+P+B)

An outline for the disposition codes can be seen in Figure 1.

Figure 1. Outline for the disposition codes.

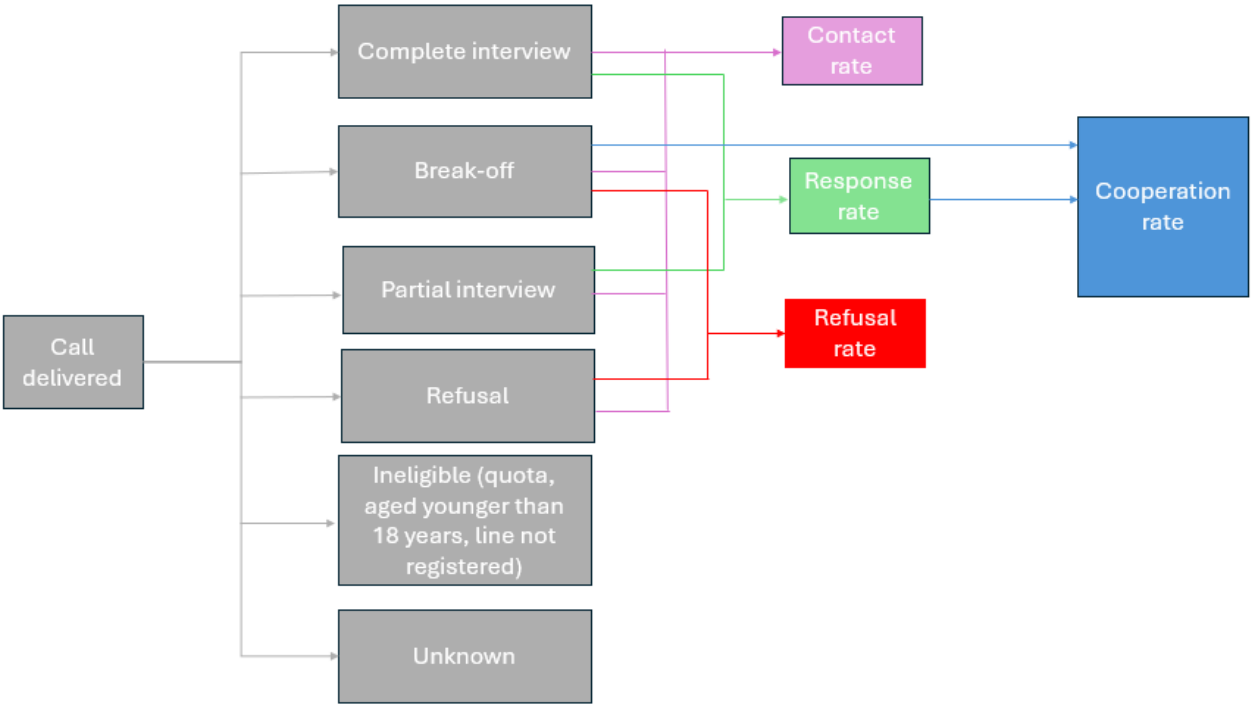

Outcomes—Demographic Characteristics

Demographic characteristics of the respondents were compared to the National Quality of Life Survey 2021 (NQLS). Importantly, we did not use statistical tests to compare the 2 databases given that the sources of data are different in structure (Table 1).

XSL•FO
RenderX

Table 1. Comparison of sociodemographic and vaccination variables for Colombia, pilot 1[a].

Variables	NQLS 2021[b,c,d], % (SE)	Pilot 1 (N=367), n (%, SE)	Point difference between NQLS and pilot 1
Sociodemographic			
Age (years), mean (SD)	41.8 (0.068)	36.3 (0.734)	–5.4
Age groups (years)			
18-29	27 (0.0019)	139 (38, 0.025)	+11
30-39	21 (0.0018)	88 (24, 0.022)	+3
40-49	18 (0.0017)	70 (19, 0.020)	+1
50-59	16 (0.0015)	37 (10, 0.015)	–6
60-78	18 (0.0016)	33 (9, 0.014)	–9
Sex (female)	52 (0.002)	209 (57, 0.025)	+5
Geographical area			
Rural	22 (0.0012)	99 (27, 0.023)	+5
Urban	78 (0.0012)	268 (73, 0.023)	–5
Education level			
Do not have	5 (00007)	26 (7, 0.012)	+2
Elementary school	24 (0.0017)	48 (13, 0.017)	–11
High school	44 (0.0022)	128 (35, 0.024)	–9
Technical	12 (0.0016)	99 (27, 0.023)	+15
Undergraduate or more	15 (0.0019)	66 (18, 0.020)	+3
Vaccinated against COVID-19			
Yes	__[e]	330 (90, 0.015)	—
No	—	37 (10, 0.015)	—

[a]We do not use tests to compare both databases given that both sources of data are different in structure and cannot be merged in the same database.

[b]NQLS: National Quality of Life Survey 2021.

[c]For the National Quality of Life Survey, absolute values are not reported because we estimated weighted percentages.

[d]The percentages of the National Quality of Life Survey 2021 were calculated with sample weights.

[e]Not applicable.

All statistical analyses were conducted using Stata (version 14; StataCorp LLC).

Ethical Considerations

This study was approved by the Institutional Review Board of the Johns Hopkins Bloomberg School of Public Health (17868), and the Ethics Committee of the Public Health Institute at Universidad Javeriana under filing number 4 of the session conducted on August 13, 2021.

Results

Disposition Codes and Survey Rates for Pilots 1 and 2

Pilot 1

In total, 55,000 phone calls were made for the first pilot. The contact rate was 1.64% (900 out of 55,000), the response rate was 0.73% (403 out of 55,000), and the cooperation rate was 80.48% (367 of 456; see Tables 2 and 3).

XSL•FO
RenderX

Table 2. Disposition codes and survey rates for pilots 1 and 2[a].

Call outcomes	Pilot 1 (N=55,000), n (%)	Pilot 2 (N=588,891), n (%)
Complete interview	367 (0.7)	451 (0.08)
Partial interview	36 (0.07)	24 (0.004)
Break-off	53 (0.10)	72 (0.01)
Ineligible: quota met	N/A[b]	3486 (0.59)
Ineligible: <18 years	72 (0.13)	633 (0.11)
Ineligible: line not registered	10,891 (19.80)	119,542 (20.30)
Refusal	444 (0.81)	6230 (1.06)
Unknown	43,137 (78.43)	458,453 (77.85)

[a]The disposition codes of the pilot surveys are according to the American Association for Public Opinion Research [29].

[b]N/A: not applicable.

Table 3. Disposition codes and survey rates for pilots 1 and 2[a].

Survey rate	Pilot 1, n/N (%)	Pilot 2, n/N (%)
Contact rate[b]	900/43,984 (2.05)	6777/465,230 (1.46)
Response rate[c]	403/44,037 (0.92)	475/465,230 (0.10)
Refusal rate[d]	497/44,037 (1.13)	6302/465,230 (1.35)
Cooperation rate[e]	367/456 (80.48)	451/547 (82.45)

[a]The disposition codes of the pilot surveys are according to the American Association for Public Opinion Research [29].

[b]Contact rate: measures the proportion of all cases in which some eligible mobile phone user was reached by the survey.

[c]Response rate: measures the proportion of both complete and incomplete interviews over both all known eligible mobile phone users and users with unknown eligibility.

[d]Refusal rate: measures the proportion of mobile phone users that do not consent or break off the survey after consent over both all known eligible and users with unknown eligibility.

[e]Cooperation rate: measures the proportion of complete interviews over all known eligible mobile phone users.

We met the sample size with 367 surveys for a nationally representative sample. The average age was 36.36 years (SD 14.06 years, range 18 to 78 years); the population with the highest participation in the sample were aged 18 to 29 (n=139, 38%) years and 30 to 39 (n=88, 24%) years, followed by 40 to 49 age group (n=70, 19%). More than half were women (n=209, 57%), had not completed a college degree (n=301, 82%), and lived in an urban area (n=268, 73%).

We found that 6% (22 out of 367) of respondents reported that someone in their household had had any COVID-19–related symptoms in the past 3 days. It was primarily the respondents who had any COVID-19–related symptoms (n=11, 52%). Likewise, 44% (n=17) of the respondents reported having had a confirmed or suspected case of COVID-19 in the past. Further, 90% (n=329) of respondents reported having received at least one dose of the COVID-19 vaccine, 54% (n=177) reported being fully vaccinated but without the booster, and 27% (n=88) stated being fully vaccinated with the booster. Of the 10% (38/367) who had not been vaccinated, they cited concerns about vaccine safety (n=10, 27%). Additionally, the top 3 means by which respondents learned about the COVID-19 vaccine were *other networks or the internet* (n=180, 49%), *Facebook* (n=77, 21%), and *friends or relatives* (n=43, 12%). Finally, most respondents reported that they clearly understood the information on COVID-19 and trusted the information they received (Table 4).

XSL•FO
RenderX

Table 4. COVID-19 variables for pilot 1.

Variables	Values
Symptoms related to COVID-19 (3 days ago), n (%)	
Yes	21 (6)
No	346 (94)
Who has COVID-19, n (%)	
Respondent	11 (52)
Household	5 (24)
Respondent and household	5 (24)
Age of sick household member (years), mean (SD)	26.4 (21.74)
Sex of sick household member (female), n (%)	4 (80)
Community symptoms, n (%) [a]	
Yes	11 (65)
No	6 (35)
Have you ever had COVID-19, n (%)	
Yes, confirmed	69 (19)
Yes, suspected	92 (25)
No	206 (56)
Wanted to receive the COVID-19 vaccine, n (%) [b]	
Yes	276 (84)
No	53 (16)
Would you like to receive the COVID-19 vaccine, n (%)	
Yes	17 (44)
No	21 (55)
Why have you not been vaccinated, n (%)	
Vaccines were not available	6 (16)
Vaccine does not work	2 (5)
Vaccine is not safe	10 (27)
I do not think I need it	5 (13)
Already got COVID-19	2 (5)
I have not been able to get the vaccine	6 (16)
Other reasons	5 (13)
Do not know or do not want to respond	2 (5)
Number of COVID-19 vaccine doses, n (%)	
One dose	61 (19)
Fully, not booster	177 (54)
Fully, plus booster	88 (27)
The brand of the first vaccine dose, n (%)	
CoronaVac	87 (26)
Pfizer	78 (24)
Johnson & Johnson	66 (20)
AstraZeneca	45 (14)
Moderna	44 (13)

XSL•FO
RenderX

Variables	Values
Another vaccine	2 (1)
Do not know	7 (2)
The brand of the second vaccine dose, n (%)	
CoronaVac	75 (28)
Pfizer	76 (29)
Johnson & Johnson	35 (13)
AstraZeneca	33 (12)
Moderna	37 (14)
Do not know	9 (3)
The brand of the booster vaccine dose, n (%)	
CoronaVac	16 (18)
Pfizer	24 (27)
Johnson & Johnson	12 (14)
AstraZeneca	16 (18)
Moderna	17 (19)
Do not know	3 (3)
Vaccination record for COVID-19, n (%)	
Yes, I have it	331 (90)
I do not have it	30 (8)
I do not know it	3 (1)
I prefer not to answer	3 (1)
Main source of information about COVID-19 vaccination, n (%)	
Facebook	77 (21)
WhatsApp	29 (8)
Other networks or internet	180 (49)
Friends or relatives (word of mouth)	43 (12)
Radio	4 (1)
Television	26 (7)
Print media	3 (1)
I have not received any information	5 (1)
Your understanding about COVID-19 vaccination seems, n (%)	
Very clear	244 (66)
It is more or less clear	92 (25)
Little clear	24 (7)
Not clear	7 (2)
Do you trust the information you receive about COVID-19 vaccination, n (%)	
Always	157 (43)
Almost always	135 (37)
Almost never	55 (15)
Never	20 (5)

[a]Community symptoms: do you know about someone outside your household who has had in the last month any of the following symptoms: fever, sore throat, frequent cough, feeling sick, or loss of the sense of smell?

[b]Wanted to receive the COVID-19 vaccine: did you want to receive the COVID-19 vaccine?

XSL•FO
RenderX

Pilot 2

In total, 588,891 phone calls were made for the second pilot. The contact rate was 1.16% (6777 out of 588,891), the response rate was 0.08% (474 out of 588,891), and the cooperation rate was 82.45% (451 out of 547; see Tables 2 and 3).

We met the sample size for the 5 regions in Colombia (Caribbean region: n=90, Pacific region: n=91, Amazon region: n=88, Orinoco River region: n=93, and Central region: n=89). In total, for the 5 regions, a sample of 451 responses was reached. All results are presented in Table 5.

XSL•FO
RenderX

Table 5. Comparison of sociodemographic and vaccination variables for Colombia, pilot 2.

Variables	The Caribbean region			The Pacific region			The Amazon region			The Orinoquia region			The Central region		
	NQLS[a,b,c] 2021, % (SE)	IVR[d] (n=90), % (SE)	Difference between NQLS and pilot 2	NQLS 2021, % (SE)	IVR (n=91), % (SE)	Difference between NQLS and pilot 2	NQLS 2021, % (SE)	IVR (n=88), % (SE)	Difference between NQLS and pilot 2	NQLS 2021, % (SE)	IVR (n=93), % (SE)	Difference between NQLS and pilot 2	NQLS 2021, % (SE)	IVR (n=89), % (SE)	Difference between NQLS and pilot 2
Sociodemographic															
Age (years), mean (SD)	42.04 (0.101)	36.16 (1.623)	–5.88	43.43 (0.167)	34.89 (1.367)	–8.41	40.41 (0.152)	29.72 (1.115)	–10.69	41.01 (0.171)	32.69 (1.286)	–8.32	43.52 (0.111)	34.17 (1.491)	–9.35
Age groups (years)															
18-29	29 (0.0028)	41 (0.052)	12	27 (0.0043)	40 (0.051)	13	32 (0.0045)	52 (0.053)	20	31 (0.0047)	48 (0.052)	17	26 (0.0029	46 (0.053)	20
30-39	21 (0.0025)	24 (0.045)	3	21 (0.0039)	25 (0.045)	4	22 (0.0040)	28 (0.048)	6	22 (0.0042)	26 (0.045)	4	21 (0.0027)	20 (0.042)	–1
40-49	18 (0.0024)	15 (0.037)	–3	17 (0.0036)	23 (0.044)	6	17 (0.0035)	16 (0.039)	–1	18 (0.0039)	13 (0.034)	–5	18 (0.0025)	16 (0.038)	–2
50-59	15 (0.0021)	11 (0.033)	–4	15 (0.0033)	9 (0.029)	–6	14 (0.0031)	2 (0.015)	–12	13 (0.0034)	11 (0.032)	–2	15 (0.0023)	10 (0.032)	–5
60-91	17 (0.0022)	9 (0.030)	–8	20 (0.0038)	3 (0.018)	–17	15 (0.0031)	1 (0.011)	–14	16 (0.0037)	2 (0.015)	–14	20 (0.0025)	8 (0.028)	–12
Sex (female)	52 (0.0030)	59 (0.053)	7	53 (0.0048)	52 (0.052)	–1	51 (0.0047)	51 (0.054)	0	50 (0.0051)	49 (0.052)	–1	53 (0.0033)	64 (0.051)	11
Geographical area															
Rural	26 (0.0023)	19 (0.041)	–7	33 (0.0037)	30 (0.048)	–3	42 (0.0045)	35 (0.051)	–7	28 (0.0037)	25 (0.044)	–3	16 (0.0014)	25 (0.045)	9
Urban	74 (0.0023)	81 (0.041)	7	67 (0.0037)	70 (0.048)	3	58 (0.0045)	65 (0.051)	7	72 (0.0037)	75 (0.044)	3	84 (0.0014)	75 (0.045)	–9
Education level															
Do not have	9 (0.0015)	4 (0.021)	–5	5 (0.0017)	4 (0.021)	–1	7 (0.0022)	6 (0.024)	–1	5 (0.0020)	3 (0.018)	–2	3 (0.0010)	2 (0.015)	–1
Elementary school	23 (0.0025)	8 (0.028)	–15	30 (0.0042)	8 (0.028)	–22	37 (0.0047)	10 (0.032)	–27	29 (0.0045)	12 (0.033)	–17	24 (0.0026)	14 (0.036)	–10
High school	45 (0.0031)	31 (0.049)	–14	44 (0.0049)	45 (0.052)	1	43 (0.0048)	30 (0.048)	–13	45 (0.0052)	34 (0.049)	–11	43 (0.0034)	47 (0.053)	4
Technical	12 (0.0021)	30 (0.048)	18	10 (0.0034)	29 (0.047)	19	7 (0.0023)	37 (0.051)	30	11 (0.0034)	39 (0.050)	28	12 (0.0024)	19 (0.053)	7
Undergraduate or more	11 (0.0021)	27 (0.046)	16	11 (0.0034)	14 (0.036)	3	6 (0.0023)	17 (0.040)	11	10 (0.0036)	12 (0.033)	1	18 (0.0030)	18 (0.041)	0
Vaccination for COVID-19[e]															
Yes	__[f]	96 (0.021)	—	—	82 (0.040)	—	—	78 (0.044)	—	—	90 (0.030)	—	—	94 (0.024)	—
No	—	4 (0.021)	—	—	18 (0.040)	—	—	22 (0.044)	—	—	10 (0.030)	—	—	6 (0.024)	—

[a]NQLS: National Quality of Life Survey 2021.

[b]The percentages of the National Quality of Life Survey 2021 were calculated with sample weights.

XSL•FO
RenderX

[c]For the National Quality of Life Survey, absolute values are not reported because we estimated weighted percentages.

[d]IVR: interactive voice response.

[e]Vaccination data were compared with information published in Our World in Data. There should be at least one dose of vaccination. There is no available data from government sources.

[f]Not applicable.

For the 5 regions, we found that between 6% (5/90) to 12% (11/93) of respondents reported that someone in their household had had any COVID-19–related symptoms in the past 3 days. For the 5 regions, more than 50% (374/451) of the respondents reported that they did not have confirmed or suspected COVID-19 in the past. The COVID-19 vaccination rate varied across regions, from 78% (69/88) in Amazonas to 96% (86/90) in the Caribbean. Most regions reported close to 50% (221/451) full vaccination but without the booster. Of those who had not been vaccinated, the main reason cited was about vaccine safety in all regions. Additionally, the top 3 main sources of information about COVID-19 vaccination were other networks or the internet, Facebook, and friends or relatives (336/451, 74%; Table 6).

XSL•FO
RenderX

Table 6. COVID-19 variables for pilot 2.

Variables	The Caribbean region (n=90)	The Pacific region (n=91)	The Amazon region (n=88)	The Orinoquia region (n=93)	The Central region (n=89)
Symptoms related to COVID-19 (3 days ago), n (%)					
Yes	5 (6)	6 (7)	7 (7)	11 (12)	5 (6)
No	85 (94)	85 (93)	82 (93)	82 (88)	84 (94)
Who has COVID-19, n (%)					
Respondent	1 (20)	1 (17)	3 (50)	5 (46)	2 (40)
Household	4 (80)	5 (83)	3 (50)	3 (27)	2 (40)
Respondent and household	__[a]	—	—	27 (3)	20 (1)
Age of sick household member, mean (SD)	17 (16.51)	34 (30.76)	22.33 (9.07)	32.33 (24)	17.5 (17.67)
Sex of sick household member (female)	4 (100)	3 (60)	1 (33)	2 (67)	1 (50)
Community symptoms [b], n (%)					
Yes	1 (50)	1 (33)	1 (20)	6 (75)	—
No	1 (50)	2 (67)	4 (80)	2 (25)	4 (100)
Have you ever had COVID-19, n (%)					
Yes, confirmed	16 (18)	17 (19)	12 (14)	12 (13)	20 (22)
Yes, suspected	28 (31)	25 (27)	32 (36)	27 (29)	21 (24)
No	46 (51)	49 (54)	44 (50)	54 (58)	48 (54)
Wanted to receive the COVID-19 vaccine [c], n (%)					
Yes	70 (81)	60 (80)	51 (74)	68 (81)	69 (82)
No	16 (19)	15 (20)	18 (26)	16 (19)	15 (18)
Would you like to receive the COVID-19 vaccine, n (%)					
Yes	1 (25)	3 (19)	7 (37)	3 (33)	1 (20)
No	3 (75)	13 (81)	12 (63)	6 (67)	4 (80)
Why have you not been vaccinated, n (%)					
Vaccines were not available	—	12 (2)	—	22 (2)	—
Vaccine does not work	—	7 (1)	11 (2)	—	—
Vaccine is not safe	3 (75)	9 (57)	10 (53)	4 (45)	3 (60)
I do not think I need it	—	2 (12)	—	1 (11)	1 (20)
Already got covid	—	—	1 (5)	—	—
I have not been able to get the vaccine	—	—	5 (26)	2 (22)	—
Other reasons	1 (25)	2 (12)	1 (5)	—	1 (20)
Do not know or do not want to respond	—	—	—	—	—
Number of COVID-19 vaccine doses, n (%)					
One dose	19 (22)	15 (20)	18 (26)	15 (18)	19 (23)
Fully, not booster	42 (50)	49 (49)	35 (51)	50 (62)	45 (54)
Fully, plus booster	24 (28)	23 (31)	16 (23)	16 (20)	20 (24)
The brand of the first vaccine dose, n (%)					
CoronaVac	17 (20)	15 (20)	1 (2)	19 (23)	20 (24)
Pfizer	23 (27)	16 (21)	14 (20)	22 (26)	19 (23)
Johnson & Johnson	18 (21)	25 (33)	16 (23)	11 (13)	22 (26)
Astra Zeneca	14 (16)	7 (10)	16 (23)	16 (19)	7 (8)
Moderna	12 (14)	11 (15)	9 (13)	14 (17)	15 (18)

XSL•FO
RenderX

Variables	The Caribbean region (n=90)	The Pacific region (n=91)	The Amazon region (n=88)	The Orinoquia region (n=93)	The Central region (n=89)
Another vaccine	1 (1)	—	11 (16)	—	—
Do not know	1 (1)	1 (1)	2 (3)	2 (2)	1 (1)
The brand of the second vaccine dose, n (%)					
CoronaVac	12 (18)	11 (19)	10 (20)	3 (4)	17 (26)
Pfizer	23 (35)	15 (25)	13 (25)	20 (30)	17 (26)
Johnson & Johnson	8 (12)	11 (19)	8 (16)	17 (26)	10 (15)
Astra Zeneca	11 (17)	9 (15)	8 (16)	2 (3)	7 (11)
Moderna	10 (15)	12 (20)	11 (21)	13 (20)	12 (19)
Do not know	2 (3)	1 (2)	1 (2)	11 (17)	2 (3)
The brand of the booster vaccine dose, n (%)					
CoronaVac	2 (8)	3 (13)	3 (19)	2 (12)	5 (1)
Pfizer	12 (50)	5 (22)	8 (50)	6 (38)	9 (45)
Johnson & Johnson	4 (18)	4 (17)	2 (12)	2 (12)	2 (10)
Astra Zeneca	2 (8)	1 (4)	1 (7)	3 (20)	4 (20)
Moderna	2 (8)	3 (13)	2 (12)	2 (12)	2 (10)
Do not know	2 (8)	7 (31)	—	1 (6)	1 (5)
Vaccination record for COVID-19, n (%)					
Yes, I have it	80 (89)	71 (78)	66 (75)	80 (86)	82 (92)
I do not have it	8 (9)	14 (16)	16 (18)	11 (12)	7 (8)
I do not know it	2 (2)	3 (3)	4 (5)	2 (2)	—
I prefer not to answer	—	3 (3)	2 (2)	—	—
Main source of information about COVID-19 vaccination, n (%)					
Facebook	16 (18)	19 (21)	18 (20)	20 (21)	17 (19)
WhatsApp	3 (3)	8 (9)	6 (7)	9 (10)	13 (15)
Other networks or internet	49 (55)	39 (43)	35 (40)	35 (38)	45 (51)
Friends or relatives (word of mouth)	9 (10)	8 (9)	9 (10)	10 (11)	7 (8)
Radio	2 (2)	5 (5)	4 (5)	3 (3)	3 (3)
Television	8 (9)	9 (10)	12 (14)	10 (11)	2 (2)
Print media	2 (2)	1 (1)	1 (1)	2 (2)	1 (1)
I have not received any information	1 (1)	2 (2)	3 (3)	4 (4)	1 (1)
Your understanding about COVID-19 vaccination seems, n (%)					
Very clear	63 (70)	61 (67)	52 (59)	59 (63)	58 (65)
It is more or less clear	19 (21)	19 (21)	24 (27)	17 (18)	23 (26)
Little clear	4 (4)	9 (10)	8 (9)	11 (12)	7 (8)
Not clear	4 (4)	2 (2)	4 (5)	6 (7)	1 (1)
Do you trust the information you receive about COVID-19 vaccination, n (%)					
Always	44 (49)	41 (45)	38 (43)	36 (39)	46 (52)
Almost always	30 (33)	31 (34)	26 (30)	38 (41)	18 (20)
Almost never	10 (11)	14 (15)	19 (22)	10 (11)	24 (27)

XSL•FO
RenderX

Variables	The Caribbean region (n=90)	The Pacific region (n=91)	The Amazon region (n=88)	The Orinoquia region (n=93)	The Central region (n=89)
Never	6 (7)	5 (5)	5 (5)	9 (9)	1 (1)

[a]Not applicable.

[b]Community symptoms: Do you know about someone outside your household who has had in the last month any of the following symptoms: fever, sore throat, frequent cough, feeling sick, or loss of the sense of smell?

[c]Wanted to receive the COVID-19 vaccine: Did you want to receive the COVID-19 vaccine?

Comparison of NQLS With Mobile Phone Survey Samples

Pilot 1

We found that demographic variables have a similar distribution between the mobile phone surveys and NQLS. Even though respondents are more likely to be younger and female, differences do not seem to be meaningful (we are not making statements on significance because we cannot test the differences between points estimated in both surveys). Regarding geographical variables, we found that in both IVR and NQLS samples, a quarter of respondents stated living in rural areas. Importantly, IVR respondents tend to be more educated than the general population (Table 5).

Pilot 2

Overall, for the 5 regions, we found that demographic variables have a similar distribution to pilot 1 results. When comparing pilot 2 results to NQLS, we found that in regions with lower access to mobile phone technology, population groups are less likely to be represented in IVR surveys (older, less educated, and rural) than in regions with higher access to mobile phone technology. For example, the central region concentrates around 60% of the population of the country. In this region, the share of respondents reporting living in rural areas was 9 percentage points higher than its share of the population. Meanwhile, in the Orinoco and Amazon river regions (each with around 1.5% of the population), the share of respondents reporting living in urban areas was 3 percentage points higher than its share of the population.

Discussion

In this study, we conducted 2 different pilots to assess the feasibility of a national active syndromic surveillance system for COVID-19 in Colombia. In both pilots, we found that it is possible to deploy a rapid active syndromic surveillance system for respiratory disease at the national level in Colombia with fair contact rates (between 1% and 2%) and excellent cooperation rates (above 80%) when compared to other studies [6]. Evidence from Burkina Faso, a low-income country, has reported similar cooperation rates using IVR [30] in a nonemergency context. In another study, an IVR survey implemented during the COVID-19 pandemic reported refusal rates in Ecuador and Sri Lanka of 5.1% and 2%, respectively, whereas our refusal rate was close to 1% (6799/509,267). Another study conducted in 5 LMICs reported similar contact, response, cooperation, and refusal rates [8].

Sampling and operational characteristics allow public health officials to monitor changes in the prevalence of respiratory symptoms over time to detect and prepare for changes in respiratory disease prevalence. An additional benefit of this technology is that it does not need face-to-face interaction reducing exposure to survey personnel, can be deployed faster in challenging or low-density environments [26], and is generally less costly [9]. However, this system has 2 main limitations. On the demand side, it is clear that respondents are more likely to be younger, female, and more educated than the average population implying that they might be overrepresented, as it has been found in other research using IVR [16,31]. To reduce selection bias, measures must be taken to improve participation from less represented groups [28]. It is possible that over time and as younger cohorts grow, issues around digital literacy and access to cell phone numbers will subside.

On the supply side, it is also clear that connectivity and mobile phone availability might play a role in survey participation [31]. To reduce selection bias associated with this effect, the second pilot included quotas by region, revealing similar patterns to those found in the first pilot (more likely to find younger, female, and more educated respondents), but still providing closer demographic estimates to those found in external surveys. The strategy of using quotas can be further expanded to smaller areas of particular interest such as municipalities, if needed.

This study is limited mainly by the availability of government data to compare results, particularly related to COVID-19 and other cases of respiratory disease, and vaccination status by region. The available information about behavior and COVID-19 in Colombia is reported on the Johns Hopkins COVID Behaviors Dashboard, which is at the national level. For example, in March 2022, the Dashboard indicated that in Colombia, about 40% of the unvaccinated wanted to get a vaccine, and our data at the national level (pilot 1) found that that was about 44% (17/38). Likewise, the Dashboard reported an ever–COVID-19 prevalence of 43%, and our data found an ever-prevalence (confirmed or suspected) of 44% (161/367), proving similar estimates with very different data sources.

This evaluation shows that using IVR as a surveillance system may be useful and feasible for conducting syndromic surveillance amid a health emergency. This study informs new areas of expansion of this work: including developing quotas for demand-side variables that affect response rates including sex, age, and educational level as well as sentinel systems so people can report on the status of their households and neighborhoods to expand the network of surveillance.

XSL•FO
RenderX

Acknowledgments

We thank the anonymous participants who responded to the survey and the Ministry of Health of Colombia and the National Institute of Health for their request to carry out this evaluation, and for their guidance and advice throughout the development of the project. Finally, many thanks to engageSPARK, the mobile phone survey provider, for their outstanding role in delivering the surveys and continuous support. This evaluation was supported by the Inter-American Development Bank (137799). The content is the responsibility of the authors and does not necessarily reflect the views of the Inter-American Development Bank.

Data Availability

The anonymized surveillance data used in this study are available upon written request.

Conflicts of Interest

None declared.

References

1. Kennedy B, Fitipaldi H, Hammar U, Maziarz M, Tsereteli N, Oskolkov N, et al. App-based COVID-19 syndromic surveillance and prediction of hospital admissions in COVID Symptom Study Sweden. Nat Commun 2022;13(1):2110 [FREE Full text] [doi: 10.1038/s41467-022-29608-7] [Medline: 35449172]
2. Mian MS, Razaq L, Khan S, Hussain N, Razaq M. Pathological findings and management of COVID-19 patients: a brief overview of modern-day pandemic. Cureus 2020;12(5):e8136 [FREE Full text] [doi: 10.7759/cureus.8136] [Medline: 32550056]
3. Graham LM. Observations from the COVID-19 pandemic. Pediatr Allergy Immunol Pulmonol 2020;33(2):45-46 [FREE Full text] [doi: 10.1089/ped.2020.1181] [Medline: 35921583]
4. Güemes A, Ray S, Aboumerhi K, Desjardins MR, Kvit A, Corrigan AE, et al. A syndromic surveillance tool to detect anomalous clusters of COVID-19 symptoms in the United States. Sci Rep 2021;11(1):4660 [FREE Full text] [doi: 10.1038/s41598-021-84145-5] [Medline: 33633250]
5. World Health Organization. Public health surveillance for COVID-19: interim guidance, 16 December 2020; Report No.: WHO/2019-nCoV/SurveillanceGuidance/2020.8. World Health Organization. 2020. URL: https://apps.who.int/iris/handle/10665/337897 [accessed 2022-06-28]
6. Phadnis R, Wickramasinghe C, Zevallos JC, Davlin S, Kumarapeli V, Lea V, et al. Leveraging mobile phone surveys during the COVID-19 pandemic in Ecuador and Sri Lanka: methods, timeline and findings. PLoS One 2021;16(4):e0250171 [FREE Full text] [doi: 10.1371/journal.pone.0250171] [Medline: 33857226]
7. Nagpal K, Mathur MR, Biswas A, Fraker A. Who do phone surveys miss, and how to reduce exclusion: recommendations from phone surveys in nine Indian states. BMJ Glob Health 2021;6(Suppl 5):e005610 [FREE Full text] [doi: 10.1136/bmjgh-2021-005610] [Medline: 34380709]
8. Song Y, Phadnis R, Favaloro J, Lee J, Lau CQ, Moreira M, et al. Using mobile phone data collection tool, Surveda, for noncommunicable disease surveillance in five low- and middle-income countries. Online J Public Health Inform 2020;12(2):e13 [FREE Full text] [doi: 10.5210/ojphi.v12i2.10574] [Medline: 33381279]
9. Vecino-Ortiz AI, Nagarajan M, Katumba KR, Akhter S, Tweheyo R, Gibson DG, et al. A cost study for mobile phone health surveys using interactive voice response for assessing risk factors of noncommunicable diseases. Popul Health Metr 2021;19(1):32 [FREE Full text] [doi: 10.1186/s12963-021-00258-z] [Medline: 34183013]
10. Adeniyi EA, Awotunde J, Ogundokun R, Kolawole PO, Abiodun M, Adeniyi AA. Mobile health application and COVID-19: opportunities and challenges. Semantic Scholar. 2020. URL: https://www.semanticscholar.org/paper/MOBILE-HEALTH-APPLICATION-AND-COVID-19%3A-AND-Adeniyi-Awotunde/a29aeaf1b3653feab5830a6f678f697159d5642f [accessed 2022-06-28]
11. The world's richest source of ICT statistics and regulatory information. ITU DataHub. URL: https://datahub.itu.int/ [accessed 2022-06-28]
12. Timmers T, Janssen L, Stohr J, Murk JL, Berrevoets MAH. Using eHealth to support COVID-19 education, self-assessment, and symptom monitoring in the Netherlands: observational study. JMIR mHealth uHealth 2020;8(6):e19822 [FREE Full text] [doi: 10.2196/19822] [Medline: 32516750]
13. Mahmud AS, Chowdhury S, Sojib KH, Chowdhury A, Quader MT, Paul S, et al. Participatory syndromic surveillance as a tool for tracking COVID-19 in Bangladesh. Epidemics 2021;35:100462 [FREE Full text] [doi: 10.1016/j.epidem.2021.100462] [Medline: 33887643]
14. Community based self-administered syndromic surveillance mobile application—a strategic approach to monitor COVID situation at micro level. National Journal of Community Medicine. URL: https://www.njcmindia.com/index.php/file/article/view/197 [accessed 2022-06-28]

XSL•FO
RenderX

15. Greenleaf AR, Gibson DG, Khattar C, Labrique AB, Pariyo GW. Building the evidence base for remote data collection in low- and middle-income countries: comparing reliability and accuracy across survey modalities. J Med Internet Res 2017;19(5):e140 [FREE Full text] [doi: 10.2196/jmir.7331] [Medline: 28476728]
16. Pariyo GW, Greenleaf AR, Gibson DG, Ali J, Selig H, Labrique AB, et al. Does mobile phone survey method matter? Reliability of computer-assisted telephone interviews and interactive voice response non-communicable diseases risk factor surveys in low and middle income countries. PLoS One 2019;14(4):e0214450 [FREE Full text] [doi: 10.1371/journal.pone.0214450] [Medline: 30969975]
17. Gibson DG, Farrenkopf BA, Pereira A, Labrique AB, Pariyo GW. The development of an interactive voice response survey for noncommunicable disease risk factor estimation: technical assessment and cognitive testing. J Med Internet Res 2017;19(5):e112 [FREE Full text] [doi: 10.2196/jmir.7340] [Medline: 28476724]
18. Wolter K, Chowdhury S, Kelly J. Chapter 7—Design, conduct, and analysis of random-digit dialing surveys. In: Rao CR, editor. Handbook of Statistics. Edinburgh, London, Oxford, UK: Elsevier; 2009:125-154.
19. Silva J, Solano D, Fernandez C, Romero L, Vargas VJ. Privacy preserving, protection of personal data, and big data: a review of the Colombia case. Procedia Computer Science. 2019. URL: http://hdl.handle.net/11323/4836 [accessed 2022-09-27]
20. Gibson DG, Kibria GMA, Pariyo GW, Ahmed S, Ali J, Labrique AB, et al. Promised and lottery airtime incentives to improve interactive voice response survey participation among adults in Bangladesh and Uganda: randomized controlled trial. J Med Internet Res 2022;24(5):e36943 [FREE Full text] [doi: 10.2196/36943] [Medline: 35532997]
21. Vecino-Ortiz AI, Bardey D, Castano-Yepes R. Hospital variation in cesarean delivery: a multilevel analysis. Value Health Reg Issues 2015;8:116-121 [FREE Full text] [doi: 10.1016/j.vhri.2015.07.003] [Medline: 29698163]
22. Ministerio de Salud y Protección Social. Encuesta Nacional de Salud Pública. 2007. URL: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/ENCUESTA%20NACIONAL.pdf [accessed 2019-05-18]
23. Vecino-Ortiz AI. Determinants of demand for antenatal care in Colombia. Health Policy 2008;86(2-3):363-372 [FREE Full text] [doi: 10.1016/j.healthpol.2007.12.002] [Medline: 18201794]
24. Proyecciones de población. Departamento Administrativo Nacional de Estadística (DANE). URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion [accessed 2019-05-04]
25. Gibson DG, Pariyo GW, Wosu AC, Greenleaf AR, Ali J, Ahmed S, et al. Evaluation of mechanisms to improve performance of mobile phone surveys in low- and middle-income countries: research protocol. JMIR Res Protoc 2017;6(5):e81 [FREE Full text] [doi: 10.2196/resprot.7534] [Medline: 28476729]
26. Gibson DG, Pereira A, Farrenkopf BA, Labrique AB, Pariyo GW, Hyder AA. Mobile phone surveys for collecting population-level estimates in low- and middle-income countries: a literature review. J Med Internet Res 2017;19(5):e139 [FREE Full text] [doi: 10.2196/jmir.7428] [Medline: 28476725]
27. Goyes AB, Gaona JV, López VB, Niño AS, Díaz GAH. Prevalencia de síntomas respiratorios y riesgo de obstrucción al flujo aéreo en ginebra—valle del cauca. Rev Med 2017;25(2):42-54. [doi: 10.18359/rmed.3207]
28. Labrique A, Blynn E, Ahmed S, Gibson D, Pariyo G, Hyder AA. Health surveys using mobile phones in developing countries: automated active strata monitoring and other statistical considerations for improving precision and reducing biases. J Med Internet Res 2017;19(5):e121 [FREE Full text] [doi: 10.2196/jmir.7329] [Medline: 28476726]
29. Standard definitions final dispositions of case codes and outcome rates for surveys. American Association for Public Opinion Research. 2016. URL: https://aapor.org/standards-and-ethics/standard-definitions/ [accessed 2022-09-28]
30. Greenleaf AR, Gadiaga A, Guiella G, Turke S, Battle N, Ahmed S, et al. Comparability of modern contraceptive use estimates between a face-to-face survey and a cellphone survey among women in Burkina Faso. PLoS One 2020;15(5):e0231819 [FREE Full text] [doi: 10.1371/journal.pone.0231819] [Medline: 32401773]
31. Torres-Quintero A, Vega A, Gibson DG, Rodriguez-Patarroyo M, Puerto S, Pariyo GW, et al. Adaptation of a mobile phone health survey for risk factors for noncommunicable diseases in Colombia: a qualitative study. Glob Health Action 2020 Dec 31;13(1):1809841-1809893 [FREE Full text] [doi: 10.1080/16549716.2020.1809841] [Medline: 32856572]

Abbreviations

IVR: interactive voice response
LMIC: low- and middle-income country
MPS: mobile phone survey
NQLS: National Quality of Life Survey 2021
STEPS: World Health Organization Stepwise Approach to Noncommunicable Disease Risk Factor Surveillance
WHO: World Health Organization

XSL•FO
RenderX

Edited by G Eysenbach; submitted 22.06.23; peer-reviewed by DK Yon; comments to author 22.12.23; revised version received 03.03.24; accepted 08.03.24; published 17.10.24.

Please cite as:
Vecino-Ortiz AI, Guzman-Tordecilla DN, Maniar V, Agudelo-Londoño S, Franco-Suarez O, Aya Pastrana N, Rodríguez-Patarroyo M, Mejía-Rocha M, Cardona J, Chavez Chamorro M, Gibson D
Mobile Phone Syndromic Surveillance for Respiratory Conditions in an Emergency (COVID-19) Context in Colombia: Representative Survey Design
J Med Internet Res 2024;26:e50184
URL: https://www.jmir.org/2024/1/e50184
doi:10.2196/50184
PMID:39418077

©Andres I Vecino-Ortiz, Deivis Nicolas Guzman-Tordecilla, Vidhi Maniar, Sandra Agudelo-Londoño, Oscar Franco-Suarez, Nathaly Aya Pastrana, Mariana Rodríguez-Patarroyo, Marino Mejía-Rocha, Jaime Cardona, Mariangela Chavez Chamorro, Dustin Gibson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association Between Daily Internet Use and Intrinsic Capacity Among Middle-Aged and Older Adults in China: Large Prospective Cohort Study

Xing-Ling Chen[1,2,3,4], MD; Jin Li[1,2,3,4], MD; Shu-Ning Sun[1,2,3,4], PhD; Qiang-Qiang Zhao[1,2], MD; Sheng-Rong Lin[1,2,3,4], MD; Ling-Jun Wang[1,2,3,4*], PhD; Zhong-Qi Yang[1*], PhD; Shi-Hao Ni[1,2,3,4*], PhD; Lu Lu[1,2,3,4*], PhD

[1]State Key Laboratory of Traditional Chinese Medicine Syndrome, The First Affiliated Hospital, Guangzhou University of Chinese Medicine, Guangzhou, China

[2]Lingnan Medical Research Center, Guangzhou University of Chinese Medicine, Guangzhou, China

[3]University Key Laboratory of Traditional Chinese Medicine Prevention and Treatment of Chronic Heart Failure, Guangdong Province, Guangzhou, China

[4]Guangzhou Key Laboratory for Chinese Medicine Prevention and Treatment of Chronic Heart Failure, Guangzhou University of Chinese Medicine, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Lu Lu, PhD
State Key Laboratory of Traditional Chinese Medicine Syndrome
The First Affiliated Hospital
Guangzhou University of Chinese Medicine
No. 12 Airport Road
Baiyun District
Guangzhou, 510405
China
Phone: 86 02036585059
Email: coinland@gzucm.edu.cn

Abstract

Background: Intrinsic capacity (IC), as a comprehensive measure of an individual's functional ability, has gained prominence in the framework for healthy aging introduced by the World Health Organization (WHO). As internet usage continues to integrate into daily life, it is imperative to scrutinize the association between internet use and IC to effectively promote healthy aging among the middle-aged and older population.

Objective: This study aimed to investigate whether daily internet use in middle-aged and older adults delays or accelerates the decline in IC.

Methods: Participants included in the China Health and Retirement Longitudinal Study (CHARLS) comprised individuals aged ≥45 years residing in China. We analyzed 4 years of CHARLS data from the first wave (May 2011-March 2012) to the third wave (July 2015-January 2016). Data from the first and third waves were used for longitudinal studies. Self-reported data encompassed internet use, frequency of use, and demographic baseline characteristics. In addition, the IC evaluation involved physical examination and blood test data. Initially, linear regression was used to assess the relationship between daily internet use and IC, followed by regression splines to explore potential nonlinear associations. Subgroup and sensitivity analyses were used to investigate the heterogeneity of IC in specific conditions and the robustness of our results. Mediation effect analysis was conducted to identify the factors that mediate the relationship between daily internet use and IC, focusing on social participation, physical activity, and health status.

Results: Among the 12,826 participants included in the longitudinal analyses, 12,305 (95.9%) did not use the internet, while 521 (4.1%) reported daily internet use with a mean age of 52.62 (SD 7.67) years. After adjusting for demographic variables, socioeconomic factors, lifestyle behaviors, and health conditions and examining the impact of daily internet use and frequency on changes in IC, our findings indicated important associations. Specifically, daily internet use is significantly linked to a slower decline in IC over time (marginal effect 1.58, 95% CI 1.03-2.12; $P<.001$). Individuals with moderate and regular internet use frequency exhibit higher levels of maintenance in IC (marginal effect 0.74, 95% CI 0.45-1.03, $P<.001$). In addition, the relationship

XSL•FO
RenderX

between IC changes and internet use frequency demonstrated a nonlinear inverted U-shaped curve (nonlinear P=.003). Subgroup analysis further revealed that improvements in IC vary based on age and gender. Furthermore, mediation analysis denoted that more than 28.78% (95% CI 21.24-40.33) of the observed association is mediated by social participation (P<.001).

Conclusions: The findings of our research underscore the potential benefits of consistent and moderate internet use in promoting and preserving IC, particularly in cognitive capacity, sensory, vitality, and locomotion. The observed effects may be related to social participation. These insights offer valuable guidance for crafting strategies aimed at fostering healthy aging within the middle-aged and older adult demographics.

(J Med Internet Res 2024;26:e54200) doi:10.2196/54200

KEYWORDS

daily internet use; intrinsic capacity; IC; middle-aged and older adult; healthy aging; social participation

Introduction

Healthy aging is a multidimensional concept that encompasses the absence of disease and the ability to maintain functional and social well-being in later life. The World Health Organization (WHO) has proposed a capabilities-based approach to measure healthy aging, which focuses on the intrinsic capacity (IC) of individuals as the composite of all the physical and mental capacities that they can draw on at any point in time [1,2]. IC reflects the residual biological potential of the organism rather than its impairments or deficits. It is influenced by both individual and contextual factors that enable people to be and do what they value as they age [3]. IC has been validated and tested as a reliable indicator of healthy aging in different populations and settings, and it has been shown to have predictive value for mortality, disability, and quality of life outcomes [4,5]. IC also provides a holistic and comprehensive framework for assessing and managing the health needs of older people, shifting from a disease-oriented to a function-and-life–oriented approach [6]. However, despite the growing recognition and adoption of IC in clinical and research practice [3,6,7], there is still a lack of understanding of the factors that influence its development and maintenance over time.

One of the potential factors that may affect IC is internet use, which has become increasingly prevalent among older adults in recent years [8]. Internet technology offers various opportunities for enhancing the health and well-being of older people, such as providing access to information, education, entertainment, social support, and health care services [9,10]. For example, a longitudinal study conducted in China found that daily internet use was associated with a lower risk of developing chronic diseases among middle-aged and older adults [11]. However, internet use may also have negative consequences for the health of older people, such as increasing the risk of internet addiction, depression, cognitive decline, sleep disturbance, social isolation, and visual impairment [12-14]. Therefore, the relationship between internet use and healthy aging is complex. It may depend on various factors, such as the frequency, duration, purpose, and quality of internet use [15]. To effectively respond to the challenges of aging, there is a need for a more comprehensive and integrated assessment of the impact of internet use on healthy aging using the IC framework. This would allow us to examine whether internet use can enhance or impair the physical and mental capacities of older people across different domains and dimensions.

This study leveraged data from a representative cohort of individuals aged 45 and older across China to investigate the relationship between daily internet use and IC, a multifaceted indicator of healthy aging. Specifically, the research aims to delineate how varying patterns of internet engagement influence the physical and mental capacities of this demographic, addressing demographic factors and multiple dimensions of activity. By examining these associations, the study seeks to clarify the potential benefits and drawbacks of internet use in preserving or enhancing IC among middle-aged and older adults.

Methods

Data and Study Participants

We used publicly available deidentified data from the China Health and Retirement Longitudinal Study (CHARLS) to measure the IC construct and to examine daily internet use and frequency within it. The longitudinal study aims to collect a high-quality, nationally representative sample of Chinese residents ages 45 and older to serve the needs of scientific research on older adults. The baseline national wave of CHARLS was fielded in 2011 and included about 150 counties/districts and 450 villages/resident committees. The individuals were followed up every 2 years [16]. All living respondents from the first and third waves (2011 and 2015) were invited to participate in the physical examinations, blood tests, and life history surveys, which included a series of questions about health and health care history, residential history, education history, and important life events. We included all individuals who participated in the 2011 and 2013 waves and answered the questions about daily internet use. The study began by identifying individuals who had registered at least one instance of internet use during the baseline period. The inclusion criteria required participants to have follow-up records available for the subsequent data collection wave. Those without follow-up data were subsequently excluded. Furthermore, participants were excluded if they had substantial missing data, defined as exceeding 40% missing information, in the IC evaluations conducted. Detailed processes are shown in Figure S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Daily Internet Use and Frequency

The participants were asked if they had used the internet in the previous month (yes=1, no=0). If the participant's answer was "yes," they were asked about their frequency of using the internet in the past month (almost every day=3, almost every week=2, and not often=1) with higher scores indicating higher frequency. In this study, "internet use" refers to daily internet use, which to some extent conforms to the usage habits of middle-aged and older people, including but not limited to instant chatting, viewing news, watching movies, playing games, managing finances, etc [11,17].

Intrinsic Capacity

The IC used in this cohort study has been proven to have prognostic value in previous studies [3]. IC is an overarching domain containing 5 subdomains, namely locomotion, sensory, vitality, psychological capacity, and cognitive capacity. Each subdomain is assessed by various indicators, such as (1) locomotion: balance, walking speed time, and chair-stand test; (2) sensory: vision and hearing impairments; (3) vitality: forced expiratory volume, grip strength, and hemoglobin; (4) psychological capacity: affect the status, sleep quantity, and quality; (5) cognitive capacity: episodic memory test. The details of the measurement methods are provided in the Supplementary Methods sections 1-10 in Multimedia Appendix 2.

The score of IC was estimated by a structural equation model (SEM) that included a confirmatory factor analysis (CFA). The CFA was based on the subfactors identified from a previous exploratory factor analysis (EFA). A bifactor model (1 general factor and 6 subfactors) was fitted, which is presented in Figure 1. The goodness of fit statistics, such as root-mean-square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI), were used to test the model fit. The IC score was standardized by subtracting the mean and dividing by the SD. A higher score indicates a stronger capacity to perform valued activities.

Figure 1. Confirmatory factor analysis bifactor model for intrinsic capacity, illustrating the structural equation modeling used to assess intrinsic capacity with 1 general factor and 5 subfactors.

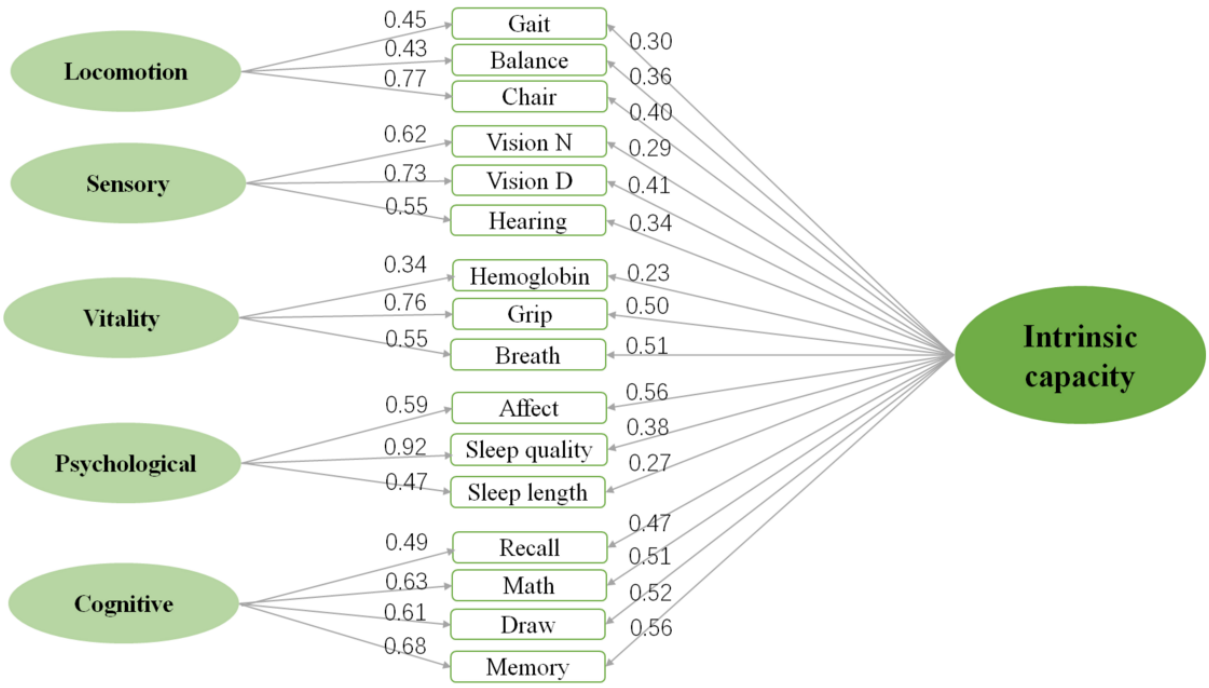

Covariates

The covariates include the basic demographic characteristics, lifestyle behaviors, and health status of the population collected in 2011. Demographic variables were collected, which included age (years), sex (male/female), education level (primary school or lower/middle school/high school or higher), residence (urban/rural), and annual household income. Lifestyle behaviors were obtained from a self-report questionnaire that included the physical activity level (metabolic equivalent physical activities [MET-PA]) [18], social participation monthly frequency (unsocial/occasional/usual), smoking (yes/no), and drinking (yes/no). We explored the potential impact of self-reported health conditions on IC. Specifically, we assessed the correlation between IC scores and 14 self-reported chronic diseases using a Spearman correlation matrix. This approach was chosen under the premise that the accumulation of chronic health conditions may exert an influence on IC. Details of this correlation analysis can be found in Figures S1-S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Health conditions contained 14 medically diagnosed conditions (yes/no), including hypertension, dyslipidemia, diabetes, cancer, chronic lung diseases, liver disease, heart diseases, stroke, kidney diseases, digestive diseases, psychiatric problems, memory-related diseases, arthritis or rheumatism, and asthma. CHARLS medical physical examinations recorded BMI, calculated as measured weight (kg) divided by the square of height (m^2).

Descriptive Statistics

The baseline characteristics of participants were divided into 2 groups based on whether they used the internet daily. Continuous variables are displayed as mean (SD), and categorical variables are displayed as numbers (percentages). The multiple imputation method was used to fill in the missing data, thereby reducing bias and enhancing the efficiency of our analyses.

Linear and Restricted Cubic Spline Regression

Multivariate linear regression models were used to evaluate the association of daily internet use and frequency with IC, adjusting for current healthy conditions, as well as demographic, socioeconomic, and lifestyle information. The association we represent is measured as marginal effects on the average of each covariate. The marginal effect and 95% CIs were determined in 5 subdomains. Our main strategy is to use the comprehensive information in survey data to directly explain as many potential confounding factors as possible, as the estimation relationship between IC and benefit factors may deviate when the outcome variable is associated with the error term [19]. Restricted cubic spline (RCS) regression was used to determine the frequency-response association between the frequency of daily internet use and outcomes among IC.

Subgroup and Sensitivity Analyses

To explore the relationship between daily internet use and IC more comprehensively and reliably, we undertook both subgroup and sensitivity analyses to affirm our findings. The subgroup analysis stratified participants by sex (male/female), age (above/below 60 years), residence (urban/rural), and whether their IC was above or below the median in 2011. Our sensitivity analysis encompassed several steps to ensure robustness: (1) we applied the Mills ratio to account for any increase or decrease in IC total scores from 2011 to 2015; (2) we assessed the consistency of internet usage patterns by designating participants as “constant users” if they reported internet use in both survey waves and as “new users” if they commenced internet use in the 2013 wave after not using it in 2011; (3) data points with extreme IC values were excluded to enhance the stability of our analysis; (4) participants with baseline cognitive impairments, which could potentially affect cognitive subdomain outcomes, such as those with Alzheimer disease, brain atrophy, or Parkinson disease, were omitted from the study; (5) we made adjustments for lifestyle variables using the baseline data from 2015 to account for changes over time; and finally (6) propensity score matching was used to align the baseline characteristics of the study groups, promoting a balanced comparison.

Mediation Effect Analysis

To investigate whether the impact of IC with its 5 subdomains is mediated by various factors, our study explores the mediating effects of social participation, MET-PA, and chronic diseases on the relationship between daily internet use and IC. The analysis involves several steps. First, we calculate the linear relationship between social participation, MET-PA, chronic diseases, and IC. Second, we use a linear model to analyze the time-related results and all the variables. Finally, we determine the mediation proportion using Preacher and Hayes method [20]. To establish the 95% CIs, we use the bootstrap method, repeating the process 5000 times.

Statistics Programs

All analyses were conducted using R (version 3.6.1; R Foundation for Statistical Computing) and Stata MP 17 (StataCorp LLC). Cubic spline analysis was performed with the “rcs” package. Subgroup analysis adopted the “Publish” package. The mediation effect was used in the “mediation” package. Multiple imputation method was performed using the “mice” package. Propensity score matching was performed using the “MatchIt” package. A 2-sided $P<.05$ was considered statistically significant.

Ethical Considerations

CHARLS was a survey approved by the Ethical Review Committee of Peking University (approval number IRB00001052–11015), and the study data were anonymous. Each participant provided signed informed consent at the time of participation. There was no requirement for additional ethics approval for approved data users.

Result

Daily Internet Use and Study Population Characteristics

In our study, a total of 12,826 participants were included, with the vast majority (12,305, 95.9%) not using the internet, while 521 (4.1%) individuals reported daily internet use. Among the 521 internet users, 294 (56.4%) were men, and 227 (43.6%) were women, with a mean age of 52.62 years. Further details on the characteristics of the study population can be found in Table 1. Participants who reported daily internet use were more likely to have higher social participation, attain a higher education level, reside in urban areas, and have higher BMI. In addition, this group tended to be younger, predominantly male, and more likely to come from households with higher economic income. In the EFA, 5 factors show eigenvalues greater than 1 (3.69, 1.69, 1.34, 1.25, 1.11). χ^2_{104}=32730.621, RMSEA=0.055 (90% CI 0.053-0.058), CFI=0.96, and TLI=0.95 for the 5-factor model and χ^2_{75}=347.4, RMSEA=0.102 (90% CI 0.101-0.103), CFI=0.77, and TLI=0.72 for the bifactor model, which suggested the model fits the data well.

XSL•FO
RenderX

Table 1. Characteristics of participants (N=12,826) by internet use.

Characteristics and variables[a]	Internet nonuse (n=12,305)	Internet use (n=521)	*P* value
Locomotion, mean (SD)			
Walking-speed time	1.42 (0.46)	1.32 (0.62)	<.001
Balance	0.26 (0.51)	0.10 (0.36)	<.001
Chair	2.36 (0.50)	2.16 (0.39)	<.001
Sensory, n (%)			
Hearing			<.001
Good	1773 (14.4)	139 (26.7)	
Fair	3709 (30.1)	197 (37.8)	
Poor	6823 (55.4)	185 (35.5)	
Vision distance			<.001
Good	1633 (13.3)	147 (28.2)	
Fair	3032 (24.6)	156 (29.9)	
Poor	7640 (62.1)	218 (41.8)	
Vision near			<.001
Good	1178 (9.5)	95 (18.3)	
Fair	2880 (23.4)	158 (30.3)	
Poor	8247 (67.0)	268 (51.4)	
Vitality, mean (SD)			
Breath	291.68 (125.86)	385.28 (138.89)	<.001
Grip	32.67 (12.30)	39.27 (10.65)	<.001
Hemoglobin	14.36 (2.17)	14.51 (1.92)	.12
Psychological capacity			
Affect, mean (SD)	21.24 (6.48)	24.89 (4.69)	<.001
Sleep length, mean (SD)	6.84 (2.15)	7.22 (1.58)	<.001
Sleep quality, n (%)			<.001
Good	4293 (34.9)	115 (22.1)	
Fair	2081 (16.9)	88 (16.9)	
Poor	5931 (48.2)	318 (61.0)	
Cognitive capacity			
Draw, complete, n (%)	7664 (62.3)	482 (92.5)	<.001
Memory, mean (SD)	3.69 (1.42)	4.75 (0.53)	<.001
Recall, mean (SD)	6.94 (3.36)	9.93 (3.21)	<.001
Math, mean (SD)	2.75 (2.00)	4.04 (1.54)	<.001
Age (years), mean (SD)	58.50 (9.19)	52.62 (7.67)	<.001
Sex[b], n (%)			<.001
Male	5649 (45.9)	294 (56.4)	
Female	6656 (54.1)	227 (43.6)	
Residence, n (%)			<.001
Urban	2059 (16.7)	390 (74.9)	
Rural	10,246 (83.3)	131 (25.1)	
BMI (kg/m^2), mean (SD)	23.48 (3.58)	24.59 (3.40)	<.001

XSL•FO
RenderX

Characteristics and variables[a]	Internet nonuse (n=12,305)	Internet use (n=521)	*P* value
Education level, n (%)			<.001
Primary school or lower	8658 (70.4)	48 (9.2)	
Middle school	2536 (20.6)	154 (29.6)	
High school or higher	1111 (8.9)	319 (61.3)	
Smoking, n (%)			.15
Yes	4706 (38.2)	216 (41.5)	
No	7599 (61.8)	305 (58.5)	
Drinking, n (%)			<.001
Yes	3978 (32.3)	222 (42.6)	
No	8327 (67.7)	299 (57.4)	
Household income (US $[c]), mean (SD)	1064.06 (1250.30)	1231.47 (1209.11)	.003
Physical activities[d], n (%)			
Vigorous activities	4431 (36.0)	51 (9.8)	<.001
Middle activities	7527 (61.2)	280 (53.7)	.001
Leisure activities	10,479 (85.2)	463 (88.9)	.02
Social participation, n (%)			<.001
Unsocial, 0 per/month	6265 (50.9)	157 (30.1)	
Occasional, 1-3 per/month	5979 (48.6)	324 (62.2)	
Usual, 4-8 per/month	61 (0.5)	40 (7.7)	
Chronic disease, n (%)			
Hypertension			.92
Yes	2869 (23.3)	123 (23.6)	
No	9436 (76.7)	398 (76.4)	
Dyslipidemia			<.001
Yes	1096 (8.9)	93 (17.9)	
No	11,209 (91.1)	428 (82.1)	
Diabetes			>.99
Yes	667 (5.4)	28 (5.4)	
No	11,638 (94.6)	493 (94.6)	
Cancer			.48
Yes	121 (1.0)	3 (0.6)	
No	12,184 (99.0)	518 (99.4)	
Chronic lung diseases			.003
Yes	1228 (10.0)	31 (6.0)	
No	11,077 (90.0)	490 (94.0)	
Liver diseases			.82
Yes	483 (3.9)	22 (4.2)	
No	11,822 (96.1)	499 (95.8)	
Heart diseases			.62
Yes	1435 (11.7)	65 (12.5)	
No	10,870 (88.3)	435 (87.5)	
Stroke			>.99

XSL•FO
RenderX

Characteristics and variables[a]	Internet nonuse (n=12,305)	Internet use (n=521)	*P* value
Yes	226 (1.8)	10 (1.9)	
No	12,079 (98.2)	511 (98.1)	
Kidney diseases			.73
Yes	767 (6.2)	30 (5.8)	
No	11,538 (93.8)	491 (94.2)	
Digestive diseases			.01
Yes	2908 (23.6)	95 (18.2)	
No	9397 (76.4)	426 (81.8)	
Psychiatric problems			.17
Yes	139 (1.1)	2 (0.4)	
No	12,166 (98.9)	519 (99.6)	
Memory-related diseases			.17
Yes	138 (1.1)	2 (0.4)	
No	12,167 (98.9)	519 (99.6)	
Arthritis or rheumatism			<.001
Yes	4357 (35.4)	92 (17.7)	
No	7948 (64.6)	429 (82.3)	
Asthma			.05
Yes	419 (3.4)	9 (1.7)	
No	11,886 (96.6)	512 (98.3)	

[a]Inconsistencies arise in some values due to rounding.

[b]The interviewer recorded the respondent's sex as either female or male.

[c]The study used the baseline year (2011 exchange rate of 1 US $=6.4588 CNY for conversion.

[d]Physical activity was defined according to each cohort's questionnaire. Vigorous activities: Activities that significantly increase heart rate and respiration, such as heavy lifting, digging, aerobics, or fast bicycling. Moderate physical activities: Activities that moderately increase heart rate and respiration, including carrying light loads, regular pace bicycling, or mopping. Leisure activities: These include walking for travel or recreation and other forms of walking for sport or exercise, both at work and at home.

Association of Daily Internet Use With IC

In the linear regression analysis, we controlled for demographic variables, socioeconomic information, lifestyle behaviors, and health conditions and examined the effects of daily internet use and frequency on IC change. Figure 2 showed that daily internet use was significantly associated with a slower decline of IC over time (marginal effect 1.58, 95% CI 1.03-2.12; *P*<.001). In addition, an incremental increase in the frequency of Internet use is associated with significant maintenance of IC, as depicted in Figure 3 (marginal effect 0.74, 95% CI 0.45-1.03; *P*<.001).

We further explored the relationship between internet use and the 5 subdomains of IC, including locomotion, sensory, vitality, psychological capacity, and cognitive capacity. In Figure 2, internet use was positively linked to all subdomains except psychological capacity, with the strong effects on locomotion (marginal effect 1.30, 95% CI 0.58-2.03; *P*<.001), sensory (marginal effect 1.32, 95% CI 0.60-2.04; *P*<.001), vitality (marginal effect 1.17, 95% CI 0.65-1.69; *P*<.001) and cognitive capacity (marginal effect 1.56, 95% CI 0.98-2.13; *P*<.001). In Figure 3, we observed similar patterns when analyzing the effects of internet frequency on the subdomains of IC, with positive associations for all subdomains except psychological capacity and the large effects for locomotion (marginal effect 0.68, 95% CI 0.29-1.06; *P*<.001), sensory (marginal effect 0.63, 95% CI 0.24-1.01; *P*<.001), vitality (marginal effect 0.50, 95% CI 0.23-0.78; *P*<.001), and cognitive capacity (marginal effect 0.80, 95% CI 0.50-1.11; *P*<.001).

XSL•FO
RenderX

Figure 2. Associations between daily internet use and changes in intrinsic capacity and its subdomains on China Health and Retirement Longitudinal Study (CHARLS) data from 2011 to 2015. FEV: forced expiratory volume; Hgb: hemoglobin.

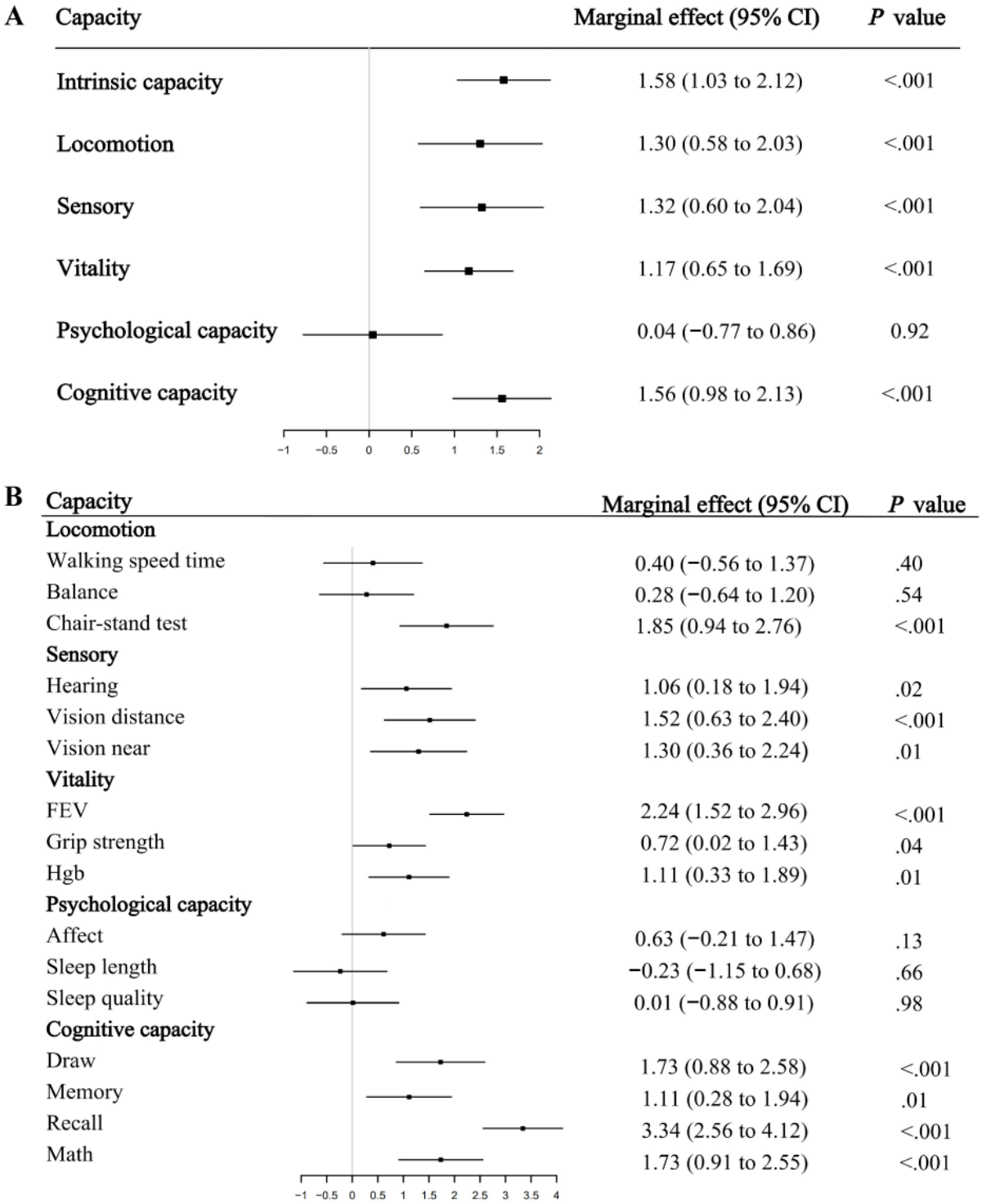

XSL•FO
RenderX

Figure 3. Associations between internet use frequency and changes in intrinsic capacity and its subdomains on China Health and Retirement Longitudinal Study (CHARLS) data from 2011 to 2015. FEV: forced expiratory volume; Hgb: hemoglobin.

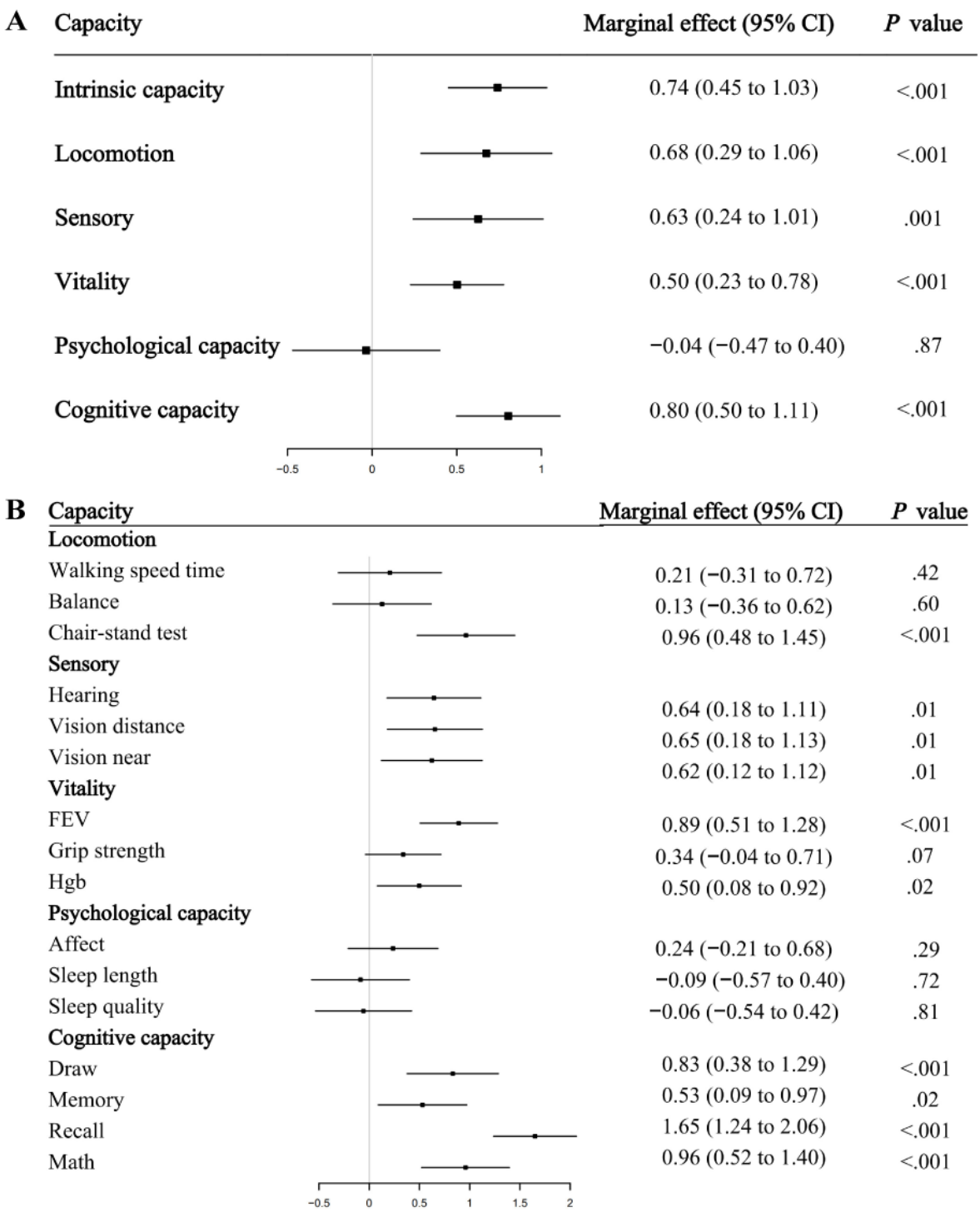

Examination of the Frequency–Response Effect

We used RCS functions to examine the relationship between IC and its 5 subdomains with daily internet use. Figure 4 shows the results of the frequency-response effect analysis on IC and its subdomains. The analysis revealed a non-linear, inverted U-shaped relationship between IC change and internet use frequency (nonlinear P=.003), indicating that there was a positive association within a certain range of internet use frequency but no significant additional benefits beyond that range. Similarly, sensory (nonlinear P=.04), vitality (nonlinear P<.001), and cognitive capacity (nonlinear P=.048) showed positive associations with increasing internet use frequency within a moderate range.

XSL•FO
RenderX

Figure 4. Impact of internet use frequency on intrinsic capacity change (2011-2015). X-axis represents internet usage frequency (almost every day=3, almost every week=2, not often=1, non-use=0), Y-axis represents the mean difference between the 2015 intrinsic capacity score and the 2011 intrinsic capacity score. CS: chi-square; df: degrees of freedom.

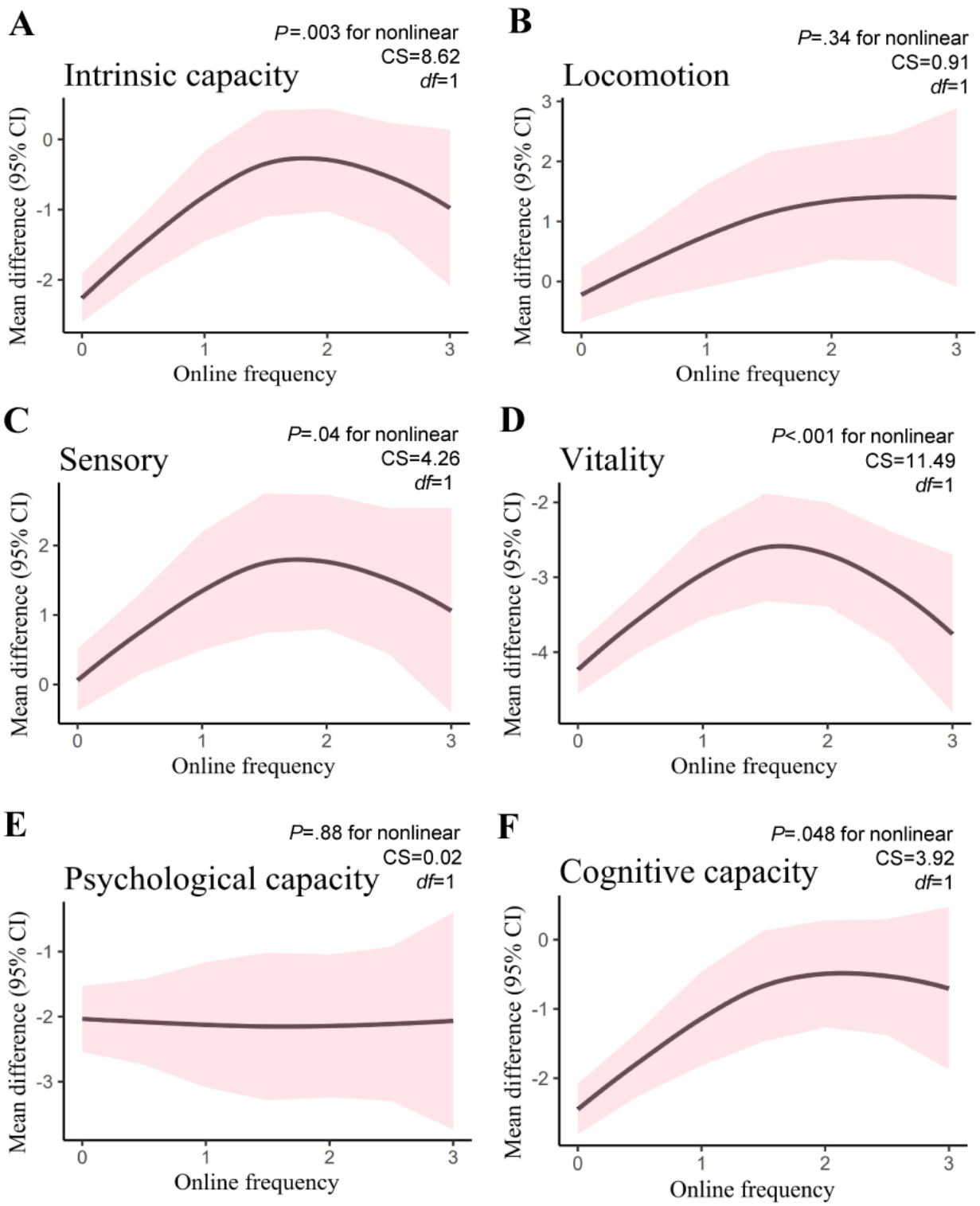

Subgroup Analyses and Sensitivity Analyses

We performed multiple subgroup analyses and sensitivity analyses to test the robustness of our results, which are shown in Figures S7-S12 in Multimedia Appendix 1. We found that internet use had a differential impact on IC across age groups and genders. Specifically, older individuals (aged 60 years and older) seem to benefit more from internet use in terms of locomotion (*P*=.02 for interaction) and vitality (*P*=.01 for interaction) than younger ones (under 60 years of age). Similarly, women show more improved cognitive capacity (*P*=.03 for interaction) while men enhanced more in vitality (*P*<.001 for interaction) through internet use. Our findings were also stable under various sensitivity analyses and did not

XSL•FO
RenderX

materially change our results (Tables S3-S4 in Multimedia Appendix 3)

Mediation Analysis for Social Participation, Physical Activity, and Chronic Diseases

To further explore the mechanisms underlying the relationship between daily internet use and IC, we conducted a mediation analysis to assess the roles of social participation, physical activity, and chronic diseases (Table S5 in Multimedia Appendix 3). The results revealed that social participation was a significant mediator between daily internet use and IC change, explaining 28.78% of the total effect (95% CI 21.24-40.33; $P<.001$), especially for the domains of locomotion, sensory, vitality, and cognitive capacity. We also confirmed some mediation effects from leisure exercise (4.26%, 95% CI 1.33-9.79; $P=.002$) and arthritis/rheumatic diseases (1.42%, 95% CI 0.30-3.02, $P=.01$).

Discussion

Principal Findings

In this prospective cohort study involving middle-aged and older adults, we found an inverted U-shaped association between daily internet use and IC. Specifically, we observed that daily internet use and increased frequency of use were associated with higher levels of cognitive capacity, sensory function, vitality, and locomotion. The results of our study underscore the potential advantages of maintaining a moderate and regular pattern of internet use in promoting and preserving IC.

Comparison to Previous Studies

Our study assessed the association between daily internet use and IC, extending beyond the focus of previous research on frailty, resilience, and aging [21-23]. IC provides a comprehensive measure of overall health status, facilitating early interventions to prevent significant health deteriorations. Unlike other aging concepts, IC emphasizes functional abilities rather than the mere presence or absence of disease [2,24]. It also considers personal attributes, environmental factors, and their interactions, enabling cross-time, cross-country, and cross-cultural comparisons of healthy aging [1,25,26].

The global interest in the association between daily internet use and cognitive capacity has been growing. Functional imaging scans have shown that older adults who were not familiar with the internet exhibited significant increases in brain neural activity when they learned to search online [27]. Certain computer programs and video games have demonstrated the potential to improve memory, multitasking abilities, fluid intelligence, and other cognitive functions [27]. Previous research has also suggested that engaging in brain-challenging activities like online searches may positively affect brain health and possibly delay cognitive decline [28]. The findings of our study supported these results by assessing various cognitive tasks, such as math, drawing, memory, and delayed recall. Furthermore, another study indicates that internet use may act as a protective factor against mild cognitive impairment in Chinese older individuals, potentially mitigating cognitive decline by influencing the volume of the globus pallidus [29]. The association between daily internet use and vitality is also consistent. This might be explained by the use of social media to promote body image management and engage in muscle exercise, which may stimulate physical vitality [30]. Previous hypotheses proposed that vitality is a biological change inherent to aging and serves as a marker of individual physiological resilience [24]. In this context, the daily use of the internet may help to delay age-related biomarkers to some extent [11]. In addition, most individuals who use the internet daily tend to have good hearing and vision, and middle-aged and older individuals who use the internet often experience stimulation through visual and auditory senses [31]. Therefore, the benefits of internet use among this population are receiving increasing attention [8]. The relationship between daily internet use and locomotion is also noteworthy. A study has highlighted the positive impact of internet use on the physical health of middle-aged and older individuals, with exercise frequency serving as a mediating factor between these variables [32]. It is plausible that accessing health-related information through the internet motivates middle-aged and older individuals to actively participate in exercise and improve their physical fitness [33,34].

The lack of a direct positive association between daily internet use and psychological capacity is somewhat unexpected. Previous research has indeed shown that internet use may reduce symptoms of depression [9,15], yet it also may negatively affect sleep quality [15], which is crucial for psychological well-being. In our study, sleep quality and duration emerged as significant factors influencing psychological capacity, particularly among older adults, who often face sleep-related challenges [35]. We hypothesize that while the internet might offer some psychological benefits, its usage can also lead to poor sleep habits, such as staying up late [36]. This, in turn, could negate potential positive impacts on psychological capacity. Therefore, the absence of a significant positive association in our findings does not necessarily contradict previous literature but highlights the complex interplay of factors affecting psychological outcomes in the aging population [37,38]. This nuanced interpretation is supported by studies suggesting that prolonged internet usage or mood fluctuations induced by internet content may lead to detrimental cycles not conducive to the mental health of older adults with preexisting sleep issues [39].

Heterogeneity analysis revealed that the impact of internet use on IC varies based on age and gender. Internet use had a more significant promoting effect on cognitive capacity in older women compared with men. This may be attributed to the generational context of women in this study, who may have a stronger desire for social resources due to historical gender biases [40]. Women are more likely to maintain positive social networks, obtaining more instrumental and emotional resources that support good cognition and active engagement in life [21,41]. On the other hand, middle-aged and older men who use the internet appear more energetic, possibly because they are more inclined to showcase physical attributes like muscle strength on social media [30]. Older individuals (older than 60 years) benefited more from internet use in terms of locomotion and vitality compared to younger individuals, possibly due to poorer physical conditions and fewer acceptable ways to promote health in the older group [25].

The results of the mediation analysis showed that internet use is positively associated with physical exercise behavior and

XSL•FO
RenderX

social participation of older people, as well as arthritis or rheumatic diseases, which is consistent with the results of previous research [11,32,42]. Regarding social participation, some researchers believe that when using the internet to chat with strangers, they tend to stay away from their families and reduce social activities, resulting in dissatisfaction with life [12,43]. A prevailing viewpoint among researchers is that the internet may overcome time and space limitations, thereby increasing opportunities for social connections and promoting meaningful contact with close relatives and friends [44,45]. In addition, internet use may enhance the social adaptability of older adults, helping them to better integrate into a wide range of social groups [46,47].

Strength and Limitations

The study investigates the relationship between daily internet use and IC based on a national large-scale prospective cohort of middle-aged and older people. However, it also has several limitations that may affect the interpretation and generalization of the results. First, as the database sample consists of Chinese subjects only, this may limit the applicability of the research results to other countries or populations. In addition, the average age of internet-using respondents in our study is 52, indicating a comparatively younger subset of the older adult population. Second, the database used in this study did not collect detailed information on the main modes and features of internet use, as well as the motivations for internet use. Therefore, we could not explore how different types of internet use may affect IC differently. Third, there is no precise measure of usage frequency that can provide a reasonable online duration. We only used self-reported categories of daily internet use, which may introduce measurement error and bias. Fourth, the data reflects usage patterns from a period that may not accurately capture current trends due to the rapid evolution and increasing ubiquity of internet technology over time. There are still many other possible factors that may mediate or mitigate the effect of internet use on IC, which are worth further investigation. For example, social support, personality traits, and lifestyle factors may influence both internet use and IC. Although we controlled for some potential confounders in the analysis, we cannot completely rule out the possibility of residual confounding or reverse causation.

Policy Recommendations

According to the 48th Statistical Report on the Development of Chinese Internet, China's Internet penetration rate rose from 45.80% in 2013 to 75.60% in December 2022, and the number of Internet users increased from 618 million in 2013 to 1067 million in 2022 [8]. Notably, older individuals are more likely to use the internet for routine daily activities, such as watching videos, staying informed about current events, and health care delivery [11,48]. Internet use may also provide older people with monitoring of routine functional abilities and depression, representing one way to aid their social integration, increase social participation, and, thus, reduce depressive symptoms [9,42]. The daily use of the internet by older adults has been found to have a significant impact on healthy aging in various aspects. Therefore, the following beneficial policy support can be considered: (1) encouraging moderate internet use: we recommend that older individuals engage in internet activities at a moderate frequency to maximize benefits and avoid the potential drawbacks of excessive use; (2) educational campaigns: it is essential to implement educational and promotional campaigns that inform older adults about healthy internet practices, enhancing their understanding and guiding them towards effective use; and (3) enhancing internet literacy: we propose the expansion of internet literacy training programs designed specifically for older adults to improve their abilities to access information, communicate, and interact online, thus maximizing the benefits of internet connectivity [49]. These policy supports can help older people better enjoy the convenience and benefits brought by the Internet and promote digital inclusion and healthy aging of older adults.

Conclusions

In conclusion, this study aims to investigate the relationship between daily internet usage and IC, an aspect that has not been previously explored. IC serves as a crucial assessment and predictive system for healthy aging, as it reflects individuals' functional abilities and overall well-being. The results of our study underscore the potential advantages of maintaining a moderate and regular pattern of internet use in promoting and preserving IC, particularly in terms of cognitive capacity, sensory capacity, vitality, and locomotion. The effect may be related to social participation. Consequently, our results offer valuable insights for developing interventions to promote healthy aging among the middle-aged and older adult population.

Acknowledgments

This work was supported by a grant from the National Science Foundation of China (82374406), the Young Elite Scientists Sponsorship Program by CAST (CACM-2021-QNRC2-B30), Guangdong Basic and Applied Basic Research Foundation (2023A1515030146), Guangzhou University of Chinese Medicine's Youth Elite Talents Cultivation "List Unveiling and Leadership" Team Project, and the Excellent Doctoral Dissertation Cultivation Project of the First Clinical School of Guangzhou University of Chinese Medicine in 2023 (YB202301), The Funding Project from Guangdong Provincial Department of Education (2024KTSCX114).

Data Availability

The datasets analyzed during this study are available in the CHARLS repository [50].

XSL•FO
RenderX

Authors' Contributions

XL-C and JL conducted the study design, performed data and statistical analysis, and drafted the manuscript. These two authors have contributed equally to this work. S-NS, Q-QZ, and S-RL assisted with literature search and manuscript editing. LL, S-HN, Z-QY, and L-JW contributed to grant submission and manuscript revisions. These 4 authors have contributed equally to this work. All authors reviewed and provided feedback on previous versions of the manuscript. All authors have read and approved the published version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional figures (S1-S13) for the study.
[DOCX File , 2150 KB - jmir_v26i1e54200_app1.docx]

Multimedia Appendix 2
Additional methods for intrinsic capacity measurement in this study.
[DOCX File , 18 KB - jmir_v26i1e54200_app2.docx]

Multimedia Appendix 3
Additional tables (S1-S5) for the study.
[DOCX File , 56 KB - jmir_v26i1e54200_app3.docx]

References

1. WHO. World Report on Ageing and Health. Geneva, Switzerland: World Health Organization; 2015.
2. Beard JR, Officer A, de Carvalho IA, Sadana R, Pot AM, Michel JP, et al. The World report on ageing and health: a policy framework for healthy ageing. Lancet 2016;387(10033):2145-2154. [doi: 10.1016/S0140-6736(15)00516-4] [Medline: 26520231]
3. Beard JR, Si Y, Liu Z, Chenoweth L, Hanewald K. Intrinsic capacity: validation of a new WHO concept for healthy aging in a longitudinal chinese study. J Gerontol A Biol Sci Med Sci 2022;77(1):94-100. [doi: 10.1093/gerona/glab226] [Medline: 34343305]
4. Beard JR, Jotheeswaran AT, Cesari M, Araujo de Carvalho I. The structure and predictive value of intrinsic capacity in a longitudinal study of ageing. BMJ Open 2019;9(11):e026119 [FREE Full text] [doi: 10.1136/bmjopen-2018-026119] [Medline: 31678933]
5. Zeng X, Shen S, Xu L, Wang Y, Yang Y, Chen L, et al. The impact of intrinsic capacity on adverse outcomes in older hospitalized patients: a One-year Follow-up study. Gerontology 2021;67(3):267-275 [FREE Full text] [doi: 10.1159/000512794] [Medline: 33735899]
6. Aliberti MJR, Bertola L, Szlejf C, Oliveira D, Piovezan RD, Cesari M, et al. Validating intrinsic capacity to measure healthy aging in an upper middle-income country: findings from the ELSI-Brazil. Lancet Reg Health Am 2022;12:100284 [FREE Full text] [doi: 10.1016/j.lana.2022.100284] [Medline: 36776430]
7. Ramírez-Vélez R, Iriarte-Fernández M, Santafé G, Malanda A, Beard JR, Garcia-Hermoso A, et al. Association of intrinsic capacity with incidence and mortality of cardiovascular disease: prospective study in UK biobank. J Cachexia Sarcopenia Muscle 2023;14(5):2054-2063 [FREE Full text] [doi: 10.1002/jcsm.13283] [Medline: 37434422]
8. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021;23(2):e26145 [FREE Full text] [doi: 10.2196/26145] [Medline: 33533727]
9. Kouvonen A, Kemppainen L, Ketonen EL, Kemppainen T, Olakivi A, Wrede S. Digital information technology use, Self-rated health, and depression: Population-based analysis of a survey study on older migrants. J Med Internet Res 2021;23(6):e20988 [FREE Full text] [doi: 10.2196/20988] [Medline: 34125069]
10. Falk Erhag H, Ahlner F, Rydberg Sterner T, Skoog I, Bergström A. Internet use and self-rated health among Swedish 70-year-olds: a cross-sectional study. BMC Geriatr 2019;19(1):365 [FREE Full text] [doi: 10.1186/s12877-019-1392-8] [Medline: 31870313]
11. Li P, Zhang C, Gao S, Zhang Y, Liang X, Wang C, et al. Association between daily internet use and incidence of chronic diseases among older adults: prospective cohort study. J Med Internet Res 2023;25:e46298 [FREE Full text] [doi: 10.2196/46298] [Medline: 37459155]
12. Zhang C, Wang Y, Wang J, Liu X. Does internet use promote mental health among middle-aged and older adults in China? Front Psychol 2022;13:999498 [FREE Full text] [doi: 10.3389/fpsyg.2022.999498] [Medline: 36457930]
13. Ma JQ, Sheng L. Internet use time and mental health among rural adolescents in China: a longitudinal study. J Affect Disord 2023;337:18-26. [doi: 10.1016/j.jad.2023.05.054] [Medline: 37236271]

XSL•FO
RenderX

14. Mylona I, Deres ES, Dere GDS, Tsinopoulos I, Glynatsis M. The impact of internet and videogaming addiction on adolescent vision: a review of the literature. Front Public Health 2020;8:63 [FREE Full text] [doi: 10.3389/fpubh.2020.00063] [Medline: 32195219]
15. Lam SSM, Jivraj S, Scholes S. Exploring the relationship between internet use and mental health among older adults in England: longitudinal observational study. J Med Internet Res 2020;22(7):e15683 [FREE Full text] [doi: 10.2196/15683] [Medline: 32718913]
16. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China health and retirement longitudinal study (CHARLS). Int J Epidemiol 2014;43(1):61-68 [FREE Full text] [doi: 10.1093/ije/dys203] [Medline: 23243115]
17. Li C, Mo W, Wei Q. The role of internet use and offline social engagement in the health of chinese older adults: evidence from nationally representative samples. Healthcare (Basel) 2023;11(5):653 [FREE Full text] [doi: 10.3390/healthcare11050653] [Medline: 36900658]
18. Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR, Tudor-Locke C, et al. 2011 compendium of physical activities: a second update of codes and MET values. Med Sci Sports Exerc 2011;43(8):1575-1581. [doi: 10.1249/MSS.0b013e31821ece12] [Medline: 21681120]
19. Violato M, Petrou S, Gray R, Redshaw M. Family income and child cognitive and behavioural development in the United Kingdom: does money matter? Health Econ 2011;20(10):1201-1225. [doi: 10.1002/hec.1665] [Medline: 20945340]
20. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput 2004;36(4):717-731. [doi: 10.3758/bf03206553] [Medline: 15641418]
21. Jiang Y, Yang F. Association between internet use and successful aging of older Chinese women: a cross-sectional study. BMC Geriatr 2022;22(1):536 [FREE Full text] [doi: 10.1186/s12877-022-03199-w] [Medline: 35764930]
22. Chang R, Low H, McDonald A, Park G, Song X. Web-based software applications for frailty assessment in older adults: a scoping review of current status with insights into future development. BMC Geriatr 2021;21(1):723 [FREE Full text] [doi: 10.1186/s12877-021-02660-6] [Medline: 34922466]
23. Enrique A, Mooney O, Salamanca-Sanabria A, Lee CT, Farrell S, Richards D. Assessing the efficacy and acceptability of an internet-delivered intervention for resilience among college students: a pilot randomised control trial protocol. Internet Interv 2019;17:100254 [FREE Full text] [doi: 10.1016/j.invent.2019.100254] [Medline: 31304095]
24. Stolz E, Mayerl H, Freidl W, Roller-Wirnsberger R, Gill TM. Intrinsic capacity predicts negative health outcomes in older adults. J Gerontol A Biol Sci Med Sci 2022;77(1):101-105. [doi: 10.1093/gerona/glab279] [Medline: 34569602]
25. Zhou M, Kuang L, Hu N. The association between physical activity and intrinsic capacity in chinese older adults and its connection to primary care: China health and retirement longitudinal study (CHARLS). Int J Environ Res Public Health 2023;20(7):5361 [FREE Full text] [doi: 10.3390/ijerph20075361] [Medline: 37047975]
26. Si Y, Hanewald K, Chen S, Li B, Bateman H, Beard JR. Life-course inequalities in intrinsic capacity and healthy ageing, China. Bull World Health Organ 2023;101(5):307-316C [FREE Full text] [doi: 10.2471/BLT.22.288888] [Medline: 37131938]
27. Small GW, Lee J, Kaufman A, Jalil J, Siddarth P, Gaddipati H, et al. Brain health consequences of digital technology use . Dialogues Clin Neurosci 2020;22(2):179-187 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
28. Small GW, Silverman DH, Siddarth P, Ercoli LM, Miller KJ, Lavretsky H, et al. Effects of a 14-day healthy longevity lifestyle program on cognition and brain function. Am J Geriatr Psychiatry 2006;14(6):538-545. [doi: 10.1097/01.JGP.0000219279.72210.ca] [Medline: 16731723]
29. Li W, Yue L, Xiao S. Association between internet use, cognitive function, and globus pallidus volumes: a study among the elderly in chinese communities. Front Public Health 2022;10:886974. [doi: 10.3389/fpubh.2022.886974] [Medline: 35646776]
30. Rodgers RF, Slater A, Gordon CS, McLean SA, Jarman HK, Paxton SJ. A biopsychosocial model of social media use and body image concerns, disordered eating, and Muscle-building behaviors among adolescent girls and boys. J Youth Adolesc 2020;49(2):399-409. [doi: 10.1007/s10964-019-01190-0] [Medline: 31907699]
31. Wolffsohn JS, Leteneux-Pantais C, Chiva-Razavi S, Bentley S, Johnson C, Findley A, et al. Social media listening to understand the lived experience of presbyopia: systematic search and content analysis study. J Med Internet Res 2020;22(9):e18306 [FREE Full text] [doi: 10.2196/18306] [Medline: 32955443]
32. Guo E, Li J, Luo L, Gao Y, Wang Z. The effect and mechanism of internet use on the physical health of the older people-empirical analysis based on CFPS. Front Public Health 2022;10:952858. [doi: 10.3389/fpubh.2022.952858] [Medline: 36339140]
33. Heponiemi T, Kainiemi E, Virtanen L, Saukkonen P, Sainio P, Koponen P, et al. Predicting internet use and digital competence among older adults using performance tests of visual, physical, and cognitive functioning: longitudinal Population-based study. J Med Internet Res 2023;25:e42287. [doi: 10.2196/42287] [Medline: 37145836]
34. Müller AM, Khoo S. Non-face-to-face physical activity interventions in older adults: a systematic review. Int J Behav Nutr Phys Act 2014;11(1):35 [FREE Full text] [doi: 10.1186/1479-5868-11-35] [Medline: 24612748]
35. Liu H, Liu F, Ji H, Dai Z, Han W. A bibliometric analysis of sleep in older adults. Front Public Health 2023;11:1055782 [FREE Full text] [doi: 10.3389/fpubh.2023.1055782] [Medline: 36908435]

XSL•FO
RenderX

36. Brautsch LA, Lund L, Andersen MM, Jennum PJ, Folker AP, Andersen S. Digital media use and sleep in late adolescence and young adulthood: a systematic review. Sleep Med Rev 2023;68:101742 [FREE Full text] [doi: 10.1016/j.smrv.2022.101742] [Medline: 36638702]
37. Tam W, Poon SN, Mahendran R, Kua EH, Wu XV. The effectiveness of reminiscence-based intervention on improving psychological well-being in cognitively intact older adults: a systematic review and meta-analysis. Int J Nurs Stud 2021;114:103847. [doi: 10.1016/j.ijnurstu.2020.103847] [Medline: 33352435]
38. Kung CSJ, Steptoe A. Internet use and psychological wellbeing among older adults in England: a difference-in-differences analysis over the COVID-19 pandemic. Psychol Med 2023;53(11):5356-5358 [FREE Full text] [doi: 10.1017/S0033291722003208] [Medline: 36177888]
39. Bertocchi FM, De Oliveira AC, Lucchetti G, Lucchetti ALG. Smartphone use, digital addiction and physical and mental health in Community-dwelling older adults: a population-based survey. J Med Syst 2022;46(8):53. [doi: 10.1007/s10916-022-01839-7] [Medline: 35716194]
40. Shan. The effect of cognitive aging on elderly women's willingness to care for the elderly. Chinese Journal of Gerontology 2019;39(11).
41. Pettus AJ, Mendez-Luck CA, Bergeron CD, Ahn S, Towne SD, Ory MG, et al. Internet-based resources for disease Self-care among Middle-aged and older women with chronic conditions. J Womens Health (Larchmt) 2017;26(3):222-233. [doi: 10.1089/jwh.2016.5843] [Medline: 27779440]
42. Xu Y, Huang Y. Chinese Middle-aged and older adults' internet use and happiness: the mediating roles of loneliness and social engagement. J Appl Gerontol 2021;40(12):1846-1855. [doi: 10.1177/0733464820959168] [Medline: 32954905]
43. Choi NG, Dinitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. J Med Internet Res 2013;15(5):e97 [FREE Full text] [doi: 10.2196/jmir.2333] [Medline: 23681083]
44. Szabo A, Allen J, Stephens C, Alpass F. Longitudinal analysis of the relationship between purposes of internet use and Well-being among older adults. Gerontologist 2019;59(1):58-68. [doi: 10.1093/geront/gny036] [Medline: 29688332]
45. He W, Li P, Gao Y, You J, Chang J, Qu X, et al. Self-reported visual impairment and depression of middle-aged and older adults: the chain-mediating effects of internet use and social participation. Front Public Health 2022;10:957586 [FREE Full text] [doi: 10.3389/fpubh.2022.957586] [Medline: 36466466]
46. Zhou D, Xu Y, Ai P. The effects of online social interactions on life satisfaction of older chinese adults: new insights based on a longitudinal approach. Healthcare (Basel) 2022;10(10):1964 [FREE Full text] [doi: 10.3390/healthcare10101964] [Medline: 36292411]
47. Yang HL, Wu YY, Lin XY, Xie L, Zhang S, Zhang SQ, et al. Internet use, life satisfaction, and subjective well-being among the elderly: evidence from 2017 China general social survey. Front Public Health 2021;9:677643 [FREE Full text] [doi: 10.3389/fpubh.2021.677643] [Medline: 34268289]
48. Arcury TA, Sandberg JC, Melius KP, Quandt SA, Leng X, Latulipe C, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020;39(2):141-150 [FREE Full text] [doi: 10.1177/0733464818807468] [Medline: 30353776]
49. Lee MA, Ferraro KF, Kim G. Digital technology use and depressive symptoms among older adults in Korea: beneficial for those who have fewer social interactions? Aging Ment Health 2021;25(10):1839-1847. [doi: 10.1080/13607863.2020.1839863] [Medline: 33131296]
50. Harmonized CHARLS. CHARLS. 2024. URL: https://charls.charlsdata.com/pages/data/111/zh-cn.html [accessed 2024-10-15]

Abbreviations

CFA: confirmatory factor analysis
CFI: comparative fit index
CHARLS: China Health and Retirement Longitudinal Study
EFA: exploratory factor analysis
IC: intrinsic capacity
MET-PA: metabolic equivalent physical activities
RCS: restricted cubic spline
RMSEA: root-mean-square error of approximation
SEM: structural equation model
TLI: Tucker-Lewis index
WHO: World Health Organization

XSL•FO
RenderX

Edited by N Cahill, T Leung; submitted 02.11.23; peer-reviewed by H Wan, X Qu; comments to author 12.04.24; revised version received 29.04.24; accepted 01.10.24; published 12.11.24.

Please cite as:
Chen XL, Li J, Sun SN, Zhao QQ, Lin SR, Wang LJ, Yang ZQ, Ni SH, Lu L
Association Between Daily Internet Use and Intrinsic Capacity Among Middle-Aged and Older Adults in China: Large Prospective Cohort Study
J Med Internet Res 2024;26:e54200
URL: https://www.jmir.org/2024/1/e54200
doi:10.2196/54200
PMID:

©Xing-Ling Chen, Jin Li, Shu-Ning Sun, Qiang-Qiang Zhao, Sheng-Rong Lin, Ling-Jun Wang, Zhong-Qi Yang, Shi-Hao Ni, Lu Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Differences in Use of a Patient Portal Across Sociodemographic Groups: Observational Study of the NHS App in England

Sukriti KC[1], MSc; Chrysanthi Papoutsi[2], PhD; Claire Reidy[2], PhD; Bernard Gudgin[3], PhD; John Powell[2], PhD; Azeem Majeed[1], MD; Felix Greaves[1*], PhD; Anthony A Laverty[1*], PhD

[1]Imperial College London, School of Public Health, London, United Kingdom

[2]University of Oxford, Nuffield Department of Primary Care Health Sciences, Oxford, United Kingdom

[3]Public Member, Cambridge, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Anthony A Laverty, PhD
Imperial College London
School of Public Health
South Kensington
London, SW7 2BX
United Kingdom
Phone: 44 2075945312
Email: a.laverty@imperial.ac.uk

Abstract

Background: The adoption of patient portals, such as the National Health Service (NHS) App in England, may improve patient engagement in health care. However, concerns remain regarding differences across sociodemographic groups in the uptake and use of various patient portal features, which have not been fully explored. Understanding the use of various functions across diverse populations is essential to ensure any benefits are equally distributed across the population.

Objective: This study aims to explore differences in the use of NHS App features across age, sex, deprivation, ethnicity, long-term health care needs, and general practice (GP) size categories.

Methods: We used weekly NHS App use data from the NHS App dashboard for 6386 GPs in England from March 2020 to June 2022. Negative binomial regression models explored variations in weekly rates of NHS App features used (registrations, log-ins, prescriptions ordered, medical record views, and appointments booked). Outcomes were measured as weekly rates per 1000 GP-registered patients, and we conducted separate models for each outcome. Regression models included all covariates mentioned above and produced incident rate ratios, which we present here as relative percentages for ease of interpretation. GP-level covariate data on sociodemographic variables were used as categorical variables in 5 groups for deprivation (Q1=least deprived practices and Q5=most deprived practices) and 4 groups for all other variables (Q1=least deprived practices and Q4=most deprived practices).

Results: We found variations in the use of different features overall and across sociodemographic categories. Fully adjusted regression models found lower use of features overall in more deprived practices (eg, Q5 vs Q1: registrations=–34%, log-ins=–34.9%, appointments booked=–39.7%, medical record views=–32.3%, and prescriptions ordered=–9.9%; $P<.001$). Practices with greater proportions of male patients also had lower levels of NHS App use (eg, Q4 vs Q1: registration=–7.1%, log-in=–10.4%, and appointments booked=–36.4%; $P<.001$). Larger practices had an overall higher use of some NHS App features (eg, Q4 vs Q1: registration=3.2%, log-ins=11.7%, appointments booked=73.4%, medical record views=23.9%, and prescriptions ordered=20.7%; $P<.001$), as well as those with greater proportions of White patients (eg, Q4 vs Q1: registration=1.9%, log-ins=9.1%, appointments booked=14.1%, medical record views=28.7%, and prescriptions ordered=130.4%; $P<.001$). Use patterns varied for practices with greater proportions of patients with long-term health care needs (eg, Q4 vs Q1: registrations=–3.6%, appointments booked=–20%, and medical record views=6%; $P\leq.001$).

Conclusions: This study highlights that the use of the NHS App features varied across sociodemographic groups. In particular, it is used less by people living in more deprived areas. Tailored interventions and patient support are required to ensure that any benefits from the NHS App are spread equally throughout the population.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56320) doi:10.2196/56320

KEYWORDS

digital health; patient portals; technological health divide; eHealth; inequality; observational; ecological; England; app; patient portal; disparities; deprivation; demographics; long-term health care; negative binomial regression model; intervention; patient support; general practice; digital technology; patient; youth;

Introduction

Digital technologies are rapidly gaining global momentum as drivers for transformative changes in health care [1,2]. In England, strategies to modernize the National Health Service (NHS) in part involved the introduction of "digital-first" options for all patients and key digitalization milestones such as the implementation of comprehensive electronic health care records across all NHS trusts by 2025 [3,4]. As part of these initiatives, the NHS App was introduced in 2019 as a digital route to a range of primary care services for all general practice (GP)–registered patients in England aged 13 years and older [5,6]. It is a patient portal that offers patients access to their health care data and facilitates interaction with their care providers using functions that allow viewing GP records, requesting repeat prescriptions, managing GP appointments, checking symptoms, registering for organ donation, and setting data sharing preferences [6]. Subsequently, additional features have also been introduced that enable accessing vaccination records, receiving messages regarding care, managing secondary care appointments, and viewing hospital waiting times [7]. Prior to 2019, portal use varied by country and was growing in England, albeit without one consistent national portal. From 2015 onward, patient portals had to be offered by GPs, and there were different iterations of national schemes offering services such as appointment booking.

Research on electronic health care record access using patient portals underscores their potential to improve patient satisfaction, enable self-management, and support empowerment by actively engaging patients in their health care decisions [8-10]. There are also reported benefits in several aspects of care, such as disease awareness, medication adherence, and patient safety [8,11]; however, their impacts on health inequalities are less well understood [12,13]. These not only relate to patient portal adoption, which is generally lower in older people [14,15], ethnic minority groups [14,16], and those from a lower socioeconomic status [15], but they are also reflected in sustained engagement trends and the use of different portal functionalities [16]. Research highlights that users may show preferences for specific patient portal features, and their ability to engage with certain functions is influenced by factors such as age, digital literacy, health conditions, and personal health care needs [17-19]. For example, a study looking at patient portal adoption by older adults revealed a notable interest among this group in using functions enabling access to health information and appointment booking. However, there were also reported issues surrounding the usability of the technology and a need for increased user support [20]. In a separate study, younger adults and female individuals exhibited greater use of patient portal features overall. However, older patients used functions associated with education more frequently and individuals experiencing higher poverty engaged with features related to billing and insurance more [19]. Such variations in patient engagement with different patient portal features may indicate an inherent user preference and the influence of the perceived value of the technology based on patient-specific needs.

Some existing research has further explored these issues. Analyses of data from 2015 to 2017 reported that patients with multiple health conditions placed a higher value on patient access portals than other patients [21]. Research focused on appointment booking and medication ordering using patient surveys from 2018 to 2020 has found these services are more likely to be used by patients with long-term health conditions and higher socioeconomic status [22]. Previous analyses by our group that used data until May 2021 focused on the impacts of COVID-19 events, but we also found higher NHS App registration rates among larger-sized, less deprived, and less ethnically diverse GP practices and younger people [23]. This study, however, assessed differences in NHS App registration only and no other app functions.

Such differences across sociodemographic groups in patient engagement need careful consideration as digital innovations may follow an inverse care law, where the availability and use of health care are inversely related to the population who need it most [24,25]. Furthermore, as national platforms, such as the NHS App, continue to evolve and play more prominent roles in health care delivery [4], prioritizing sustained patient engagement and participation beyond initial uptake is essential to achieve meaningful success. Therefore, identifying patterns of use of the various app features and understanding how different population groups are engaging with those functions is important. This study thus aims to explore the patterns of uptake of various NHS App functions and analyze differences in use across different sociodemographic groups.

Methods

Data Sources

We conducted an ecological study using data at the GP practice level exploring differences in the use of NHS App features by the GP-registered population in England. The outcome variables were weekly registrations, log-ins, prescriptions ordered, medical record views, and appointments booked. NHS App use data were available from March 2020 to June 2022, provided by the NHS Digital team as anonymized daily logs from the NHS App dashboard [26].

Data on the sociodemographic characteristics of the GP-registered population and their health care needs at the GP level were used as covariates. Data on the age and sex of all GP-registered populations were obtained from the NHS Digital website, which is a reliable source for the collection and

XSL•FO
RenderX

processing of national data across the health and social care sector in the United Kingdom [27]. We excluded data for those younger than 15 years of age, as the NHS App is not offered to younger age groups [6], and the resultant total for all GP-registered male and female patients was combined to calculate the GP practice size. Population aged 15-34 years was used as the age identifier for analyses, and the percentage of male patients was used as the identifier for sex.

Data on long-term health care needs were extracted from the GP Patient Survey, which is a survey of approximately 2.4 million adult GP-registered patients in the United Kingdom [28]. Population-weighted positive responses to the question "Do you have any long-term physical or mental health conditions, disabilities, or illnesses?" obtained from the GP Patient Survey database were included.

Information on ethnic composition and the Index of Multiple Deprivation (IMD) quintile were obtained from the Fingertips public health profile, which is a web-based repository of a range of health and social care indicators mapped as public health profiles at local and national levels [29]. The identifier for ethnicity included the percentage of the population who reported their ethnicity as "White."

Data obtained from these public health data sources were then ranked into quartiles (4 groups). Quartile 1 (Q1) included practices with the lowest population percentage for the given variable, and quartile 4 (Q4) was the highest. The only exception to this was the IMD split, which followed the Office of National Statistics method of ranking into quintiles (5 groups) [30], where quintile 1 (Q1) included practices with the lowest IMD score (ie, least deprived practices), and quintile 5 (Q5) included practices with the highest (ie, most deprived practices).

For analyses, covariate data were linked to the NHS App data using the GP practice code, which serves as a unique identifier code for each GP practice in England. Practices with incomplete data on practice size or those with missing area codes were removed to avoid errors during analysis (n=126). Practices with fewer than 200 registered patients were also excluded (n=20) to remove practices that service atypical populations such as specific sexual health services or services for homeless people. A comparison of characteristics of included and excluded practices (Table S1 in Multimedia Appendix 1) found that excluded practices had greater proportions of male patients and younger patients, and had substantially lower mean practice sizes. The population covered by our data was similar to the general population in terms of age, sex, and ethnicity (Table S2 in Multimedia Appendix 1). The ranges of categorical variables used for analyses are shown in Table S3 in Multimedia Appendix 1. The percentage of male patients ranged from 28.57% to 48.99% in the lowest quartile 1 compared with from 51.05% to 94.56% in quartile 4. The percentage of White patients ranged from 9.50% to 75.93% in the lowest quartile compared with from 97.30% to 99.56% in quartile 4. The number of registered patients in quartile 1 ranged from 255 to 5394 compared with from 12,033 to 110,443 in quartile 4.

Analyses

We had complete weekly use data for a period between March 23, 2020, and June 27, 2022, for a total of 6386 GP practices in England. Outcomes were measured as weekly rates per 1000 registered GP-registered patients and as cumulative totals at the end of the study period. Negative binomial regression models using incident rate ratios were used to explore subgroup differences in the patterns of NHS App use by patient sociodemographic characteristics, their long-term health care needs, and GP practice size. The regression models compared variations in app use for the different covariates in comparison to the reference group Q1 (ie, for IMD=least deprived practices, and for all other covariates=practices with the lowest population percentage for the given variable). Incident rate ratio values were then used to calculate the percentage change to report the relative difference between the different covariate quintiles. We also present analyses modeling covariates linearly. For these, we still use IMD in 5 categories, but the percentage of the practice population that is male, the percentage in the youngest age group, the percentage of White ethnicity, and the percentage with a long-term health illness or disability were all modeled as linear variables. Practice size coefficients are presented per 1000 increase. All statistical analyses were performed using Stata 17.0 (StataCorp LLC).

Ethical Considerations

Ethical approval was provided by the Imperial College London Research Ethics Committee (reference 21IC7292).

Results

Overview

We present cumulative totals for all NHS App features in Table 1.

Table 1. Cumulative total National Health Service (NHS) App functions used between March 23, 2020, and June 27, 2022.

NHS App feature	Cumulative total use, n	Rate per 1000 patients, mean (SD)
Registrations	24,168,835	3.27 (3.96)
Log-ins	447,976,852	59.30 (65.26)
Appointments booked	1,757,175	0.21 (0.80)
Medical record views	117,962,559	15.24 (15.83)
Prescriptions ordered	21,324,472	2.80 (3.61)
Patient population (week of June 20-27, 2022)	61,099,213	__[a]

[a]Not applicable.

XSL•FO
RenderX

Registrations

Across the whole study period, there were 24,168,835 NHS App registrations (Table 1).

Results for all outcomes from fully adjusted models are shown in Table S2 in Multimedia Appendix 1 and results for NHS App registrations are presented in Figure 1. These show strong associations between IMD and app registration, with lower rates of registration in all quintiles than in Q1 (ie, least deprived; test for trend *P*<.001). For example, registration rates were 34% lower in Q5 than in Q1 (*P*<.001). Registrations were also lower in practices with higher proportions of male patients (eg, 7.1% lower in Q4 than in Q1; *P*<.001) and in practices with greater proportions of patients with long-term health needs (eg, 3.6% lower in Q4 compared with Q1, *P*<.001; test for trend *P*<.001).

Registrations were higher in practices with higher proportions of White patients (eg, 1.9%, lower in Q4 compared with Q1; *P*<.001) and those with more patients from the youngest age group (aged 15-34 years; eg, 3.1%, lower in Q4 compared with Q1; *P*<.001). Registration rates were also higher in larger-sized practices (eg, 3.2% lower in Q4 than in Q1, *P*<.001; test for trend *P*<.001).

Figure 1. Weekly NHS App registration rates per 1000 GP-registered patients for the different covariate quantiles. IMD Q1 (reference group)=least deprived practice and IMD Q5=most deprived practices. For all other covariates, Q1=practices with the lowest population percentage for the given variable and Q4=practices with the highest population percentage for the given variable. GP: general practice; IMD: Index of Multiple Deprivation; NHS: National Health Service.

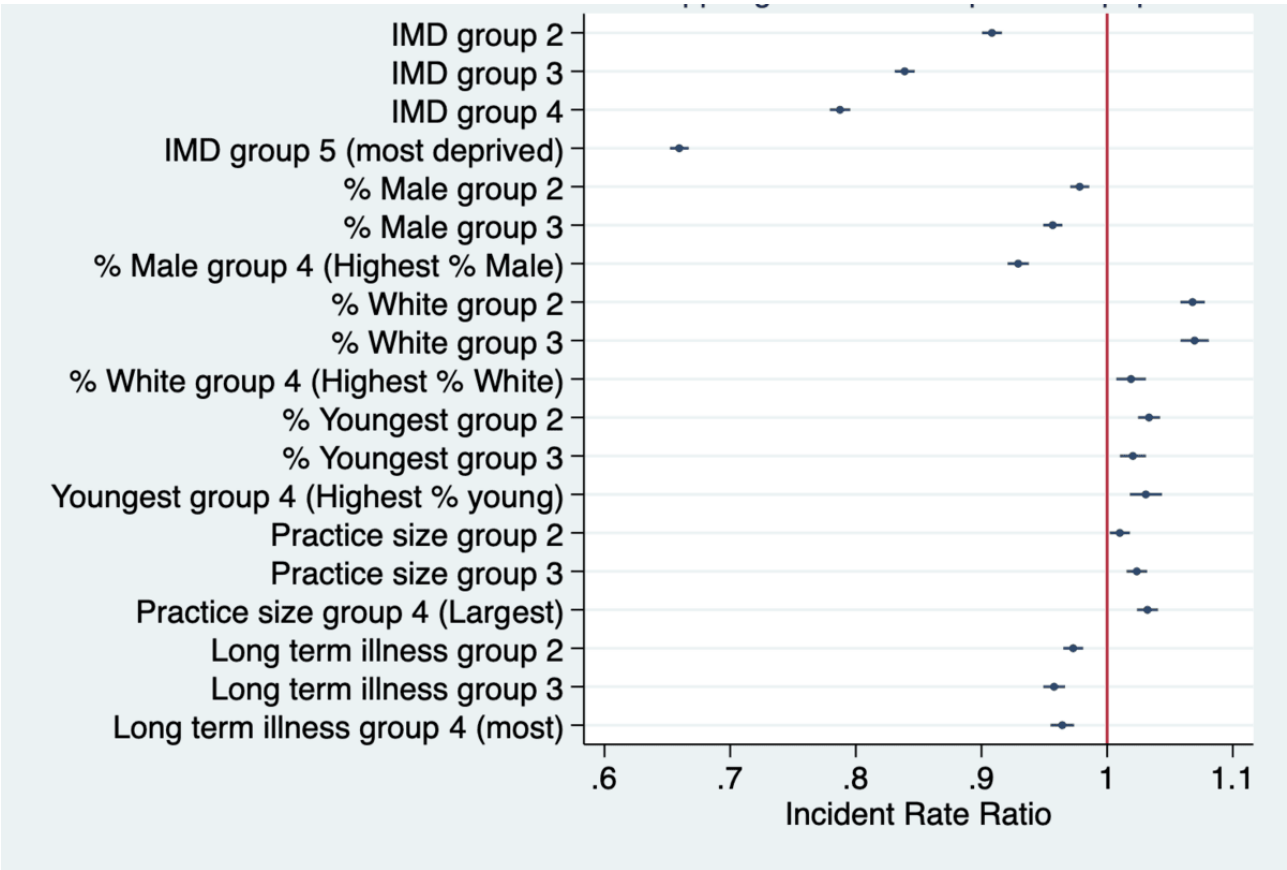

Log-ins

Across the whole study period, there were 447,976,852 NHS App log-ins (Table 1). Results showing subgroup differences in NHS App registrations using the fully adjusted models are presented in Figure 2 and they show strong associations between IMD and app log-ins, with lower log-in rates in all quintiles compared to Q1 (ie, least deprived; test for trend *P*<.001). For example, log-in rates were 34.9% lower in Q5 than in Q1 (*P*<.001). Log-ins were also lower in practices with higher proportions of male patients (eg, 10.4% lower in Q4 than in Q1; *P*<.001) and in practices with greater proportions of patients with long-term health needs (eg, 3.4% lower in Q3 than in Q1 and 1.7% lower in Q4 than in Q1, *P*=.001; test for trend *P*<.001).

App log-ins were generally higher in practices with a higher proportion of White patients (eg, 16.4% higher in Q2 than in Q1 and 9.1% higher in Q4 than in Q1; *P*<.001) and in those with more patients from the youngest age group (eg, 5.6% higher in Q4 than in Q1; *P*<.001). NHS App log-in rates were also higher in the larger-sized practices (eg, 11.7% higher in Q4 than in Q1, *P*<.001; test for trend *P*<.001).

XSL•FO
RenderX

Figure 2. Weekly NHS App log-in rates per 1000 GP-registered patients for the different covariate quantiles. IMD Q1 (reference group)=least deprived practice and IMD Q5=most deprived practices. For all other covariates, Q1=practices with the lowest population percentage for the given variable and Q4=practices with the highest population percentage for the given variable. GP: general practice; IMD: Index of Multiple Deprivation; NHS: National Health Service.

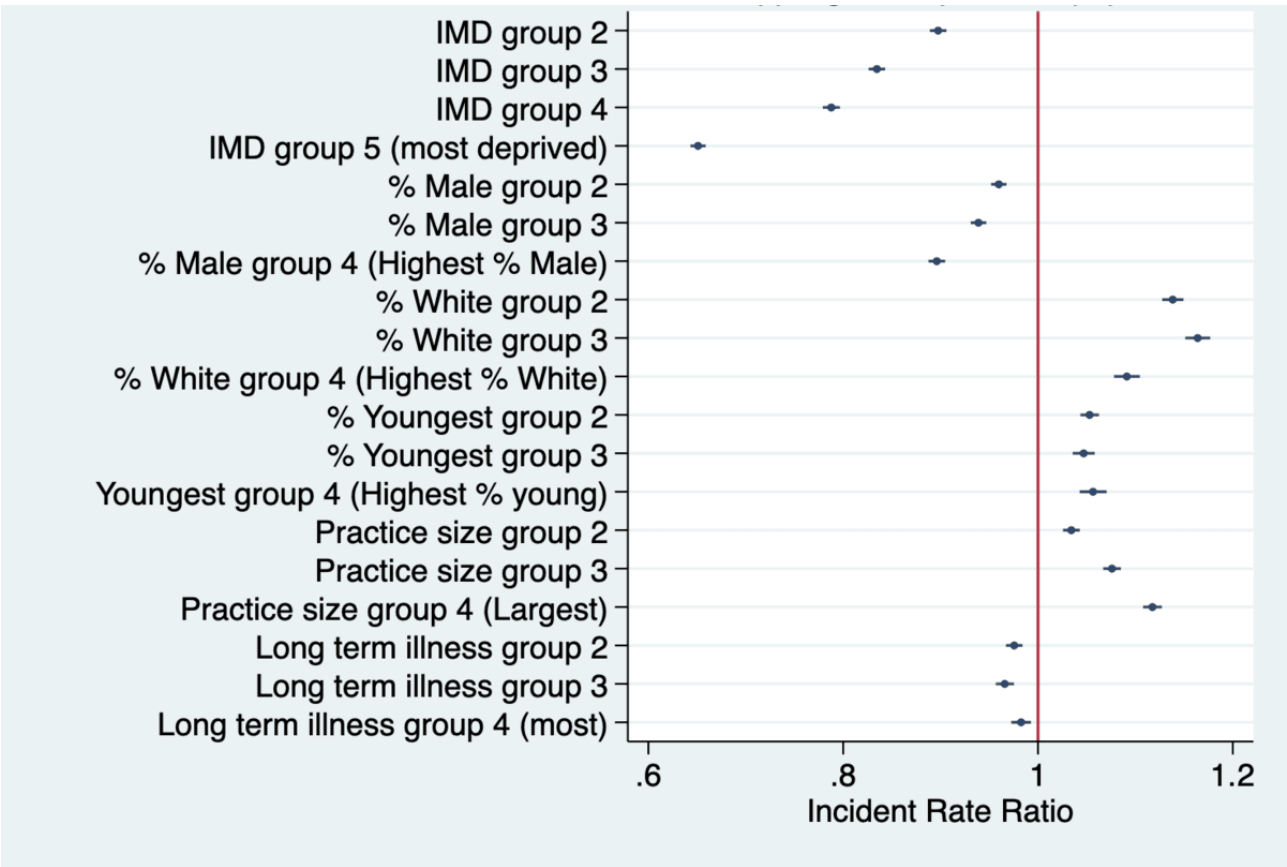

Appointments Booked

Across the whole study period, there were 1,757,975 GP appointments booked using the NHS App. Results showing subgroup differences in the rates of appointment booked using the fully adjusted models are presented in Figure 3 and they show strong associations between IMD and appointments booked, with lower rates of appointment booking rate in all quintiles than in Q1 (ie, least deprived; test for trend $P<.001$). For example, rates of appointments booked were 45.7% lower in Q2 than in Q1 and 39.7% lower in Q5 compared to Q1 ($P<.001$). Appointment booking rates were also lower in practices with higher proportions of male patients (eg, 36.4% lower in Q4 than in Q1; $P<.001$) and in those with more patients with long-term health needs (eg, 20% lower in Q4 than in Q1; $P=.001$; test for trend $P<.001$).

The rates of appointments booked were higher in practices with the highest proportion of White patients (eg, 14.1% higher in Q4 than Q1, $P<.001$) and those with the most patients from the youngest age group (ie, 46.5% higher in Q4 than Q1; $P<.001$). The rates of appointments booked were also higher among the larger-sized practices (eg, 73.4% in Q4 than in Q1, $P<.001$; test for trend $P<.001$).

XSL•FO
RenderX

Figure 3. Weekly rates of appointments booked using the NHS App per 1000 GP-registered patients for the different covariate quantiles. IMD Q1 (reference group)=least deprived practice and IMD Q5=most deprived practices. For all other covariates, Q1=practices with the lowest population percentage for the given variable and Q4=practices with the highest population percentage for the given variable. GP: general practice; IMD: Index of Multiple Deprivation; NHS: National Health Service.

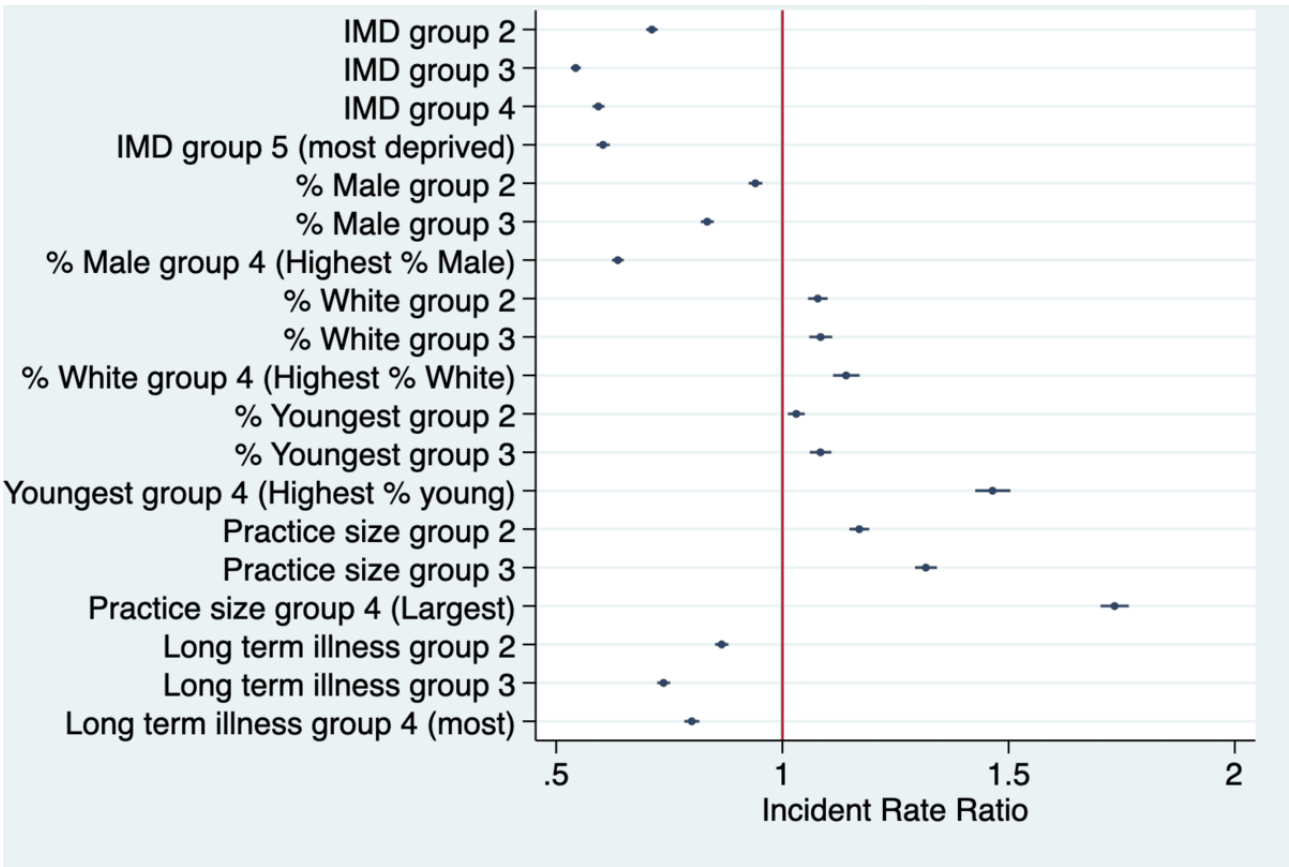

Medical Record Views

Across the whole study period, there were 117,962,559 medical record views through the NHS App. Results showing subgroup differences in the rates of medical records viewed using the fully adjusted models are presented in Figure 4 and they show strong associations between IMD and medical record views, with higher rates in all quintiles than in Q1 (ie, least deprived; test for trend $P<.001$). For example, rates of medical record views were 32.3% lower in Q5 than in Q1 ($P<.001$). They were also lower in practices with higher proportions of male patients (eg, 12% lower in Q4 than in Q1, $P<.001$; test for trend $P<.001$). Rates of medical record views were higher in practices with a higher proportion of White patients (eg, 34% higher in Q2 than in Q1 and 28.7% higher in Q4 than in Q1; $P<.001$) and in those with more patients from the youngest age group (eg, 6.7% higher in Q2 than in Q1, $P<.001$; and 1.7% higher in Q4 than in Q1, $P=.003$). The rates of medical records views were also higher in larger-sized practices (eg, 23.9% higher in Q4 than in Q1, $P<.001$) and in practices with more people with long-term health care needs (eg, 6% in Q4 than in Q1, $P<.001$; test for trend $P<.001$).

XSL•FO
RenderX

Figure 4. Weekly rates of medical record views using the NHS App per 1000 GP-registered patients for the different covariate quantiles. IMD Q1 (reference group)=least deprived practice and IMD Q5=most deprived practices. For all other covariates, Q1=practices with the lowest population percentage for the given variable and Q4=practices with the highest population percentage for the given variable. GP: general practice; IMD: Index of Multiple Deprivation; NHS: National Health Service.

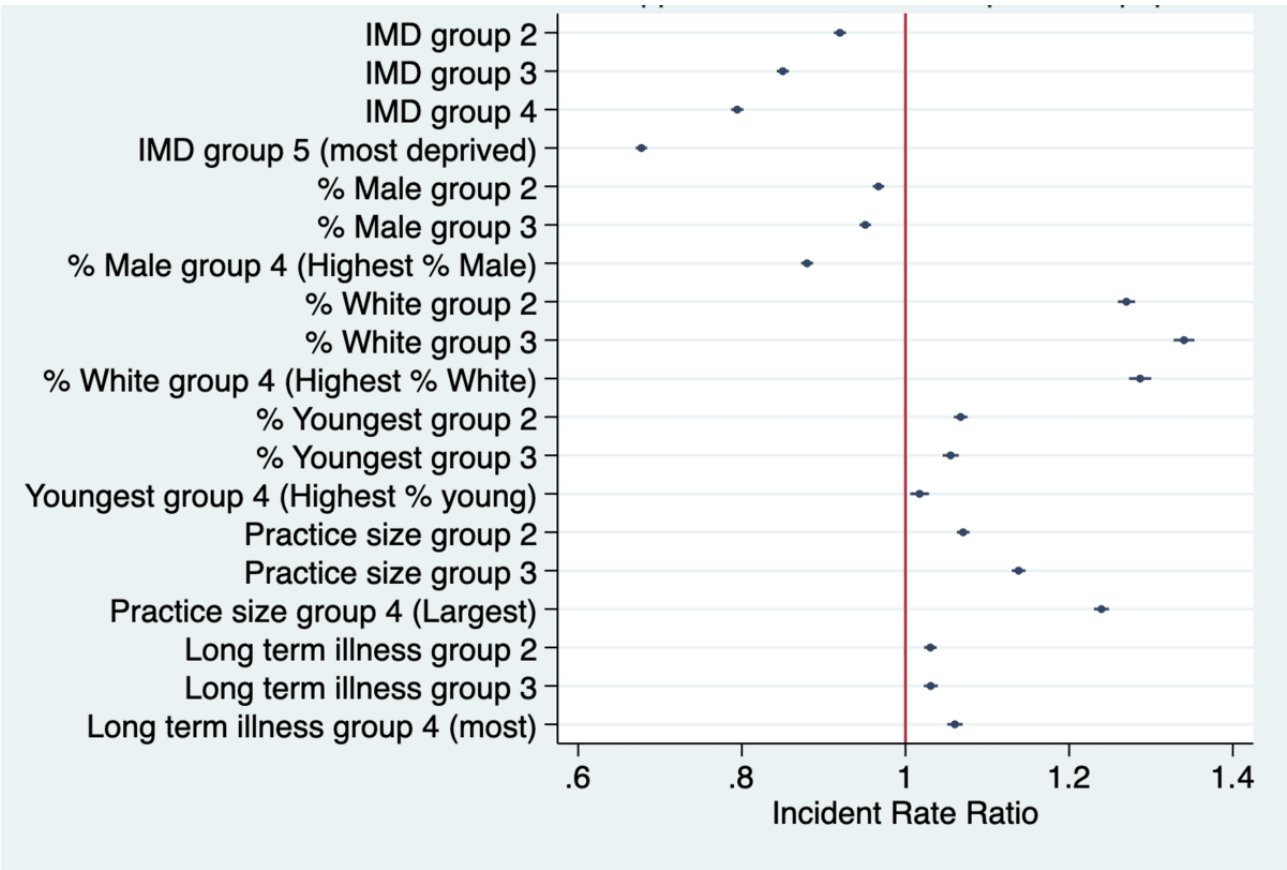

Prescriptions Ordered

Across the whole study period, there were 21,324,472 prescriptions ordered using the NHS App. Results showing subgroup differences in the rates of prescriptions ordered using the fully adjusted models are presented in Figure 5 and they show strong associations between IMD and rates of prescriptions ordered, with lower rates in all quintiles compared to Q1 (ie, least deprived; test for trend $P<.001$). For example, rates of prescriptions ordered were 9.9% lower in Q5 than in Q1 ($P<.001$). Prescription order rates were also lower in practices with higher proportions of male patients (eg, 14. 5% lower in Q4 than in Q1; $P<.001$) and in practices with a higher proportion of people aged 15-34 years (eg, 14.8% lower in Q4 than Q1; $P<.001$).

The rates of prescriptions ordered were significantly higher in practices with a higher proportion of White patients (eg, 119.4% higher in Q3 than in Q1 and 130.40% higher in Q4 than in Q1; $P<.001$) and in larger practice sizes (eg, 20.7% higher in Q4 than in Q1, $P<.001$). The rates of prescriptions ordered were also higher in practices with more people with long-term health care needs (eg, 18.3% higher in Q4 than in Q1, $P<.001$; test for trend $P<.001$).

XSL•FO
RenderX

Figure 5. Weekly rates of prescriptions ordered using the NHS App per 1000 GP-registered patients for the different covariate quantiles. IMD Q1 (reference group)=least deprived practice and IMD Q5=most deprived practices. For all other covariates, Q1=practices with the lowest population percentage for the given variable and Q4=practices with the highest population percentage for the given variable. GP: general practice; IMD: Index of Multiple Deprivation; NHS: National Health Service.

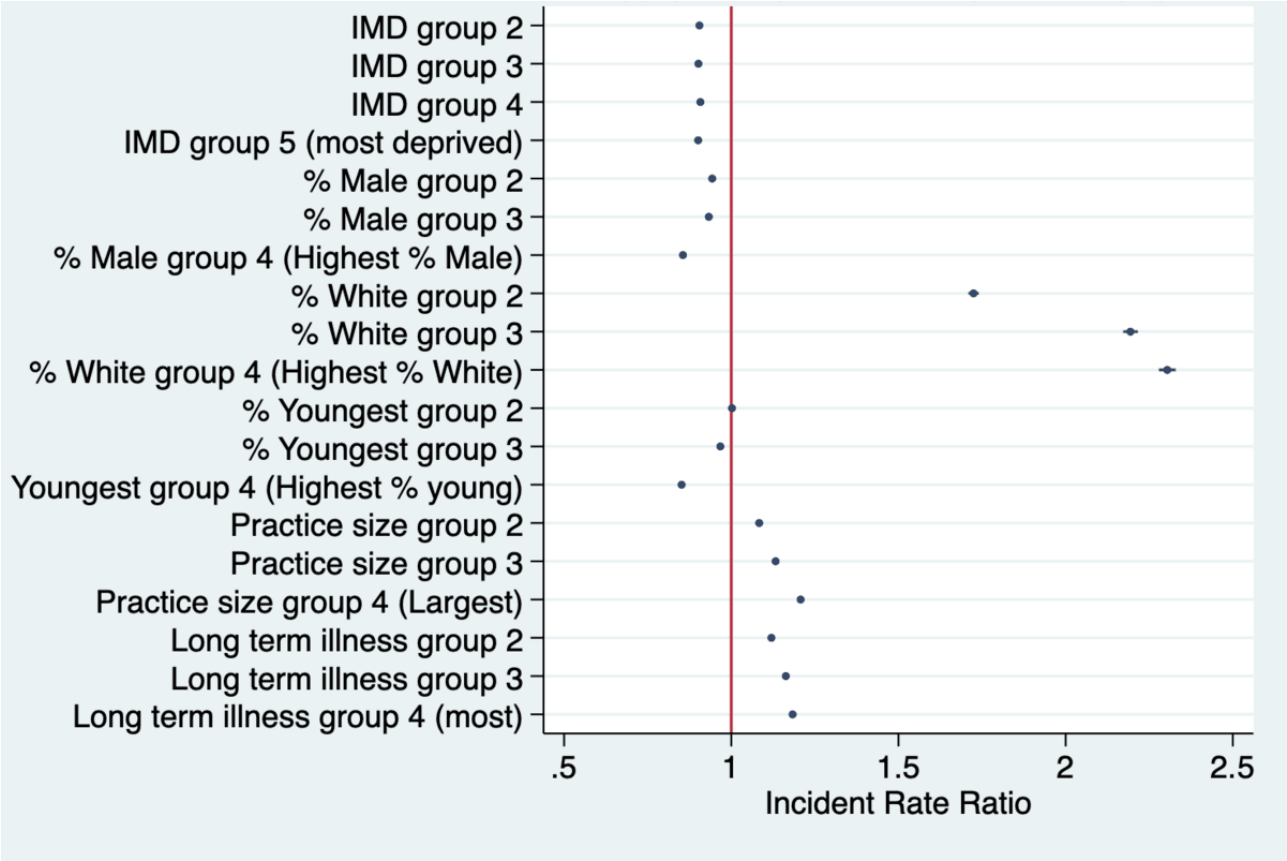

Additional Analyses

Additional analyses using covariates except for IMD as linear variables found similar patterns (Table S4 in Multimedia Appendix 1). These found each one percent increase in the percentage of male patients was associated with lower use of all outcomes (eg, –2.78%; *P*<.001 for prescription ordering). Each 1% increase in the percentage of patients in the youngest age group was associated with a –0.45% (*P*<.001) decrease in prescription ordering but higher levels of other app functions. Increased percentage of patients of White ethnicity and increased practice sizes were associated with greater use of all app functions. Increased percentage of patients with a long-term illness was associated with fewer registrations (–0.28%; *P*<.001), fewer log-ins (–0.28%; *P*<.001), and appointment booking (–0.77%; *P*<.001), but higher rates of medical record views (0.01; *P*<.001) and prescription ordering (0.61; *P*<.001).

We also conducted analyses controlling for season which gave similar results for the sociodemographic factors of interest (Table S5 in Multimedia Appendix 1).

Discussion

Principal Findings

This study explored the NHS App and found there to be high levels of patient engagement, although this differed across sociodemographic categories. The most prominent finding was the clear deprivation gradient with lower rates of use of the various NHS App features among practices in more deprived areas. In contrast, the use of the different features was higher among the largest-sized practices and in those with greater proportions of White patients. There was a 130% difference in the rates of prescriptions ordered among practices with the highest proportion of White patients compared to the lowest. In practices with greater proportions of younger patients, appointment booking rates were higher, but rates of prescription ordered were lower. Similarly, in practices with greater proportions of patients with long-term health care needs, rates of medical record views and prescriptions ordered were higher, but the rates of appointment booking, registrations, and log-ins were generally lower.

Some other results also stand out as particularly striking; for example, with ordering prescriptions digitally—one of the most commonly used features—there is a strong ethnicity gradient, with practices with a higher proportion of White patients using this function more than twice as much as practices with a lower proportion of White patients.

Strengths and Limitations

This is the first study to explore variations in the use of different NHS App features in relation to different sociodemographic groups using data over 27 months. It provides a comprehensive overview of public uptake of the NHS App using a nationally representative sample of the GP-registered population in England and offers a crucial understanding of digital health

XSL•FO
RenderX

engagement in a real-world setting. While these findings suggest an unequal pattern of patient engagement with the NHS App features in relation to patient age, sex, ethnicity, deprivation, and patient health care needs, there are limitations in using population-level data to understand individual-level differences. While the NHS App allows patients to select their data sharing preferences, individual-level data are not yet available for research, which restricts a more direct evaluation of these use patterns [6].

There are challenges in capturing and controlling the effects of the influences of wider confounding factors, such as regulatory changes, competing availabilities of similar commercial digital health platforms, and changes in strategic decisions [31], which have implications for our research findings. We included several sociodemographic variables and their proxies identified in consultation with the wider research team and the Public and Patient Involvement group, but the full range of indicators categories (eg, additional age ranges and ethnic groups) and broader socioeconomic variables were excluded due to a lack of available data. Specifically, we were unable to use data on patients on prescribed medication at each GP practice, and rates of prescriptions ordered use the whole population as a denominator. While some variables are encompassed within existing indicators (eg, income and education are parts of IMD value calculation), these factors may operate in silos to influence how people engage with digital resources [1]. Furthermore, exclusions were applied to remove practices with incomplete data, as well as removing practices serving unconventional populations. This along with the ecological nature of the analysis may affect the generalizability of results and present missed opportunities to capture the demography these practices serve and their distinct challenges.

Comparison With Prior Work on NHS App Use

The results of this study extend the findings of our previous research [23] and indicate differences in patient engagement with the NHS App across sociodemographic groups. NHS Digital highlights that the NHS App currently has over 30 million registrations, with over 65 million patient record views and 22 million repeat prescription orders through the app in 2022 [32]. While our data support these findings by highlighting high use across all app features, evidence of a varied pattern of app function use among diverse population groups is congruent with wider research in the field [21,22]. Most notably, reduced use of the different functions in more deprived and ethnically diverse practices reinforces concerns related to the digital inverse care law [25]. A previous study exploring differences in the uptake of NHS primary care services, including the NHS App, highlighted reduced use of both the NHS App and web-based patient portals among the 4.27 million patients living in the most deprived quintiles in England [33]. As the NHS plans for the app to be a primary way to access services, reduced use among deprived groups risks widening health disparities—that is, avoidable differences in outcomes between groups.

Although the diffusion of innovation theory suggests that some of these differences may decrease as the technology is more widely used by the general population [34], certain patient portal features may continue to have differential adoption based on individual capacity and need [19]. Discrepancies in the use of the various functions, as highlighted by our results, such as younger adults showing increased appointment bookings but decreased prescription requests, underscore these complexities. Younger adults generally have a better capacity for engagement with digital health technology, yet their overall health care use tends to be lower, primarily because their health care needs are generally less pronounced [35]. Therefore, the reduction in prescription orders through the NHS App among this group may reflect a good health status overall. The evidence of increased uptake of certain transactional services, such as appointment booking, could be indicative of their technological literacy and it may have been influenced by their motivation to use these services. While we could not measure these associations directly for the NHS App, they warrant further exploration in future studies.

In congruence with previous studies, our results also highlight notable differences in different app feature use among individuals with long-term health conditions [21,22]. While appointment booking rates were lower among this group, medical record views and prescription orders were higher, signaling heightened patient engagement with features catering to their ongoing health care needs [36,37]. The lower levels of appointment booking could be associated with the configuration of this feature, which is contingent upon the specific provisions implemented by individual practices. It could also be linked to variations in patient self-care ability as studies highlight that individuals with long-term health care needs often depend on their social networks and caregivers for support and they could potentially benefit from delegated digital access for care coordination [12,38]. Although the NHS App offers access to linked profiles for family members and carers, the availability of these proxy access entitlements relies on individual GP practice provisions and this could be examined in future research [39].

Nevertheless, our results show variation in the use of the app overall and specific functions by various patient characteristics, suggesting opportunities for design, policy, and practice change. This might include targeted marketing of the app and its functions toward specific groups, as well as targeted outreach and education for groups such as older people, ethnic minorities, and those with lower incomes. This could include, for example, digital literacy programs to equip people from these groups with the skills needed to use the NHS App more effectively. Other measures to reduce differences in the use of NHS App could include helplines for those who need support or advice; and collaboration with community groups that work with underserved populations. Furthermore, if those with better capacity, interest, and opportunity to engage with technology use digital services more, it may reduce pressures on the health care system and offer cost-saving opportunities to improve care for those who continue to benefit from in-person support [40]. This may be particularly relevant to a patient-facing tool, such as the NHS App, which is available nationally, offering a range of functions for diverse population groups. Therefore, identifying users and nonusers of the NHS App and supporting engagement with features most relevant to their specific needs

XSL•FO
RenderX

may be crucial to creating equal outcomes for all [41] and needs further research. There is also a need for assessment of the uptake trends of the NHS App contingent on continued policy efforts and facilitation strategies to understand its relevance, user satisfaction, and equitable access, contributing to the overall success of such digital health initiatives.

Conclusions

We found ongoing high levels of patient engagement with the various NHS App functions, although these differed across sociodemographic groups. NHS App uptake and use had a strong deprivation gradient as well as influences of patient age, ethnicity, and health care needs, indicating the requirement for dedicated efforts to meaningfully engage different population groups. Further research is also needed to explore how these differences affect health outcomes.

This study contributes to the growing body of evidence on digital health engagement and lays the groundwork for future research and policy making. It underscores the importance of continuous monitoring, evaluation, and adaptation of digital health initiatives to ensure they meet the evolving needs of diverse populations and contribute to equitable health outcomes for all. The NHS App has the potential to be a valuable tool for patient care in the NHS in England, but it is important to ensure that all individuals have the opportunity to use the app effectively.

Acknowledgments

This study was funded by the National Institute of Health Research (NIHR) Health Services Delivery and Research Programme (NIHR128285). In addition, the study was supported by the NIHR Applied Research Collaborations in North West London and Oxford and Thames Valley at Oxford Health NHS Foundation Trust, and the latter fund, JP. The authors express gratitude to the National Health Service (NHS) App teams at NHS Digital and NHSX for enabling access to NHS App use data.

Conflicts of Interest

FG and JP are employed by the National Institute for Health and Care Excellence. FG is also employed by the Department of Health and Social Care, United Kingdom.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 29 KB - jmir_v26i1e56320_app1.docx]

References

1. Equity within digital health technology within the WHO European region: a scoping review. World Health Organization. 2022. URL: https://phw.nhs.wales/publications/publications1/who-scoping-review-final-report/ [accessed 2023-11-11]
2. Digital-in-health: unlocking the value for everyone. World Bank. URL: https://openknowledge.worldbank.org/handle/10986/40212 [accessed 2024-01-10]
3. Transformation directorate: What Good Looks Like framework. NHS England. 2021 Aug 31. URL: https://transform.england.nhs.uk/digitise-connect-transform/what-good-looks-like/what-good-looks-like-publication/ [accessed 2024-11-11]
4. A plan for digital health and social care. Department of Health and Social Care. 2022. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2024-11-10]
5. Donnelly T. Digital first primary care and how the NHS long term plan set a clear direction to mainstream digitally enabled care across the NHS. NHS England. 2019. URL: https://www.england.nhs.uk/blog/digital-first-primary-care-and-how-the-nhs-long-term-plan-set-a-clear-direction/ [accessed 2024-11-10]
6. NHS app features. NHS Digital. URL: https://digital.nhs.uk/services/nhs-app/nhs-app-features [accessed 2024-11-08]
7. Future developments for the NHS app. NHS Digital. URL: https://digital.nhs.uk/services/nhs-app/future-developments [accessed 2024-11-10]
8. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
9. Han HR, Gleason KT, Sun C, Miller HN, Kang SJ, Chow S, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019;6(4):e15038 [FREE Full text] [doi: 10.2196/15038] [Medline: 31855187]
10. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]
11. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
12. Antonio MG, Petrovskaya O, Lau F. Is research on patient portals attuned to health equity? A scoping review. J Am Med Inform Assoc 2019;26(8-9):871-883 [FREE Full text] [doi: 10.1093/jamia/ocz054] [Medline: 31066893]

XSL•FO
RenderX

13. Cresswell K, Rigby M, Georgiou A, Wong ZS, Kukhareva P, Medlock S, et al. The role of formative evaluation in promoting digitally-based health equity and reducing bias for resilient health systems: the case of patient portals. Yearb Med Inform 2022;31(1):33-39 [FREE Full text] [doi: 10.1055/s-0042-1742498] [Medline: 35654424]
14. Walker DM, Hefner JL, Fareed N, Huerta TR, McAlearney AS. Exploring the digital divide: age and race disparities in use of an inpatient portal. Telemed e-Health 2020;26(5):603-613 [FREE Full text] [doi: 10.1089/tmj.2019.0065] [Medline: 31313977]
15. Paccoud I, Baumann M, Le Bihan E, Pétré B, Breinbauer M, Böhme P, et al. Socioeconomic and behavioural factors associated with access to and use of personal health records. BMC Med Inform Decis Mak 2021;21(1):18 [FREE Full text] [doi: 10.1186/s12911-020-01383-9] [Medline: 33435970]
16. Goel MS, Brown TL, Williams A, Hasnain-Wynia R, Thompson JA, Baker DW. Disparities in enrollment and use of an electronic patient portal. J Gen Intern Med 2011;26(10):1112-1116 [FREE Full text] [doi: 10.1007/s11606-011-1728-3] [Medline: 21538166]
17. Taha J, Czaja SJ, Sharit J, Morrow DG. Factors affecting usage of a personal health record (PHR) to manage health. Psychol Aging 2013;28(4):1124-1139 [FREE Full text] [doi: 10.1037/a0033911] [Medline: 24364414]
18. Irizarry T, Shoemake J, Nilsen ML, Czaja S, Beach S, Dabbs AD. Patient portals as a tool for health care engagement: a mixed-method study of older adults with varying levels of health literacy and prior patient portal use. J Med Internet Res 2017;19(3):e7099 [FREE Full text] [doi: 10.2196/jmir.7099] [Medline: 28360022]
19. Tuan WJ, Mellott M, Arndt BG, Jones J, Simpson AN. Disparities in use of patient portals among adults in family medicine. J Am Board Fam Med 2022;35(3):559-569 [FREE Full text] [doi: 10.3122/jabfm.2022.03.210486] [Medline: 35641056]
20. Wildenbos GA, Maasri K, Jaspers M, Peute L. Older adults using a patient portal: registration and experiences, one year after implementation. Digital Health 2018;4:2055207618797883 [FREE Full text] [doi: 10.1177/2055207618797883] [Medline: 30186619]
21. Mohammed MA, Montague J, Faisal M, Lamming L. The value of a patient access portal in primary care: a cross-sectional survey of 62,486 registered users in the UK. Univ Access Inf Soc 2019;19(4):855-872. [doi: 10.1007/s10209-019-00693-8]
22. Alturkistani A, Greenfield G, Beaney T, Norton J, Costelloe CE. Cross-sectional analyses of online appointment booking and repeat prescription ordering user characteristics in general practices of England in the years 2018-2020. BMJ Open 2023;13(10):e068627 [FREE Full text] [doi: 10.1136/bmjopen-2022-068627] [Medline: 37827735]
23. Kc S, Tewolde S, Laverty AA, Costelloe C, Papoutsi C, Reidy C, et al. Uptake and adoption of the NHS app in England: an observational study. Br J Gen Pract 2023;73(737):e932-e940 [FREE Full text] [doi: 10.3399/BJGP.2022.0150] [Medline: 37783512]
24. Marmot M. An inverse care law for our time. BMJ 2018;362:k3216. [doi: 10.1136/bmj.k3216] [Medline: 30065009]
25. Honeyman M, Maguire D, Evans H, Davies A. Digital Technology and Health Inequalities: A Scoping Review. Cardiff, United Kingdom: Public Health Wales NHS Trust; 2020.
26. NHS app dashboard. NHS Digital. URL: https://digital.nhs.uk/services/nhs-app/nhs-app-dashboard [accessed 2024-11-10]
27. Your data. NHS Digital. URL: https://digital.nhs.uk/your-data [accessed 2024-09-12]
28. GP patient survey 2021: technical annex. Ipsos MORI. 2021. URL: https://gp-patient.co.uk/downloads/2021/GPPS_2021_Technical_Annex_PUBLIC.pdf [accessed 2024-09-08]
29. Public health profiles. Office for Health Improvement & Disparities. 2021. URL: https://fingertips.phe.org.uk/ [accessed 2024-09-11]
30. Fingertips guidance. Technical guide—assigning deprivation categories. Office for Health Improvement & Disparities. 2017. URL: https://fingertips.phe.org.uk/profile/guidance/supporting-information/ph-methods/ [accessed 2024-09-08]
31. Majeed A, Maile EJ, Bindman AB. The primary care response to COVID-19 in England's national health service. J R Soc Med 2020;113(6):208-210 [FREE Full text] [doi: 10.1177/0141076820931452] [Medline: 32521196]
32. NHS app hits over 30 million sign-ups. NHS Digital. 2023. URL: https://www.england.nhs.uk/2023/12/nhs-app-reaches-record-users-on-fifth-anniversary/ [accessed 2024-11-08]
33. Zhang J, Gallifant J, Pierce RL, Fordham A, Teo J, Celi L, et al. Quantifying digital health inequality across a national healthcare system. BMJ Health Care Inform 2023;30(1):e100809 [FREE Full text] [doi: 10.1136/bmjhci-2023-100809] [Medline: 38007224]
34. Rogers EM. Diffusion of Innovations. New York, NY: Free Press of Glencoe; 1962.
35. Mason CN, Miller T. International projections of age specific healthcare consumption: 2015–2060. J Econ Ageing 2018;12:202-217. [doi: 10.1016/j.jeoa.2017.04.003]
36. Ancker JS, Barrón Y, Rockoff ML, Hauser D, Pichardo M, Szerencsy A, et al. Use of an electronic patient portal among disadvantaged populations. J Gen Intern Med 2011;26(10):1117-1123 [FREE Full text] [doi: 10.1007/s11606-011-1749-y] [Medline: 21647748]
37. Lafata JE, Miller C, Shires D, Dyer K, Ratliff SM, Schreiber M. Patients' adoption of and feature access within electronic patient portals. Am J Manag Care 2018;24(11):e352-e357 [FREE Full text] [Medline: 30452203]
38. Gross N, Byers V, Geiger S. Digital health's impact on integrated care, carer empowerment and patient-centeredness for persons living with dementia. Health Policy Technol 2021;10(3):100551. [doi: 10.1016/j.hlpt.2021.100551]

XSL•FO
RenderX

39. Proxy access. NHS England. URL: https://www.england.nhs.uk/long-read/proxy-access/#proxy-access-and-the-nhs-app [accessed 2024-11-11]
40. Winward S, Patel T, Al-Saffar M, Noble M. The effect of 24/7, digital-first, NHS primary care on acute hospital spending: retrospective observational analysis. J Med Internet Res 2021;23(7):e24917 [FREE Full text] [doi: 10.2196/24917] [Medline: 34292160]
41. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]

Abbreviations

GP: general practice
IMD: Index of Multiple Deprivation
NHS: National Health Service

Edited by E Sezgin; submitted 12.01.24; peer-reviewed by S de Sequeira, L Wisk; comments to author 22.03.24; revised version received 24.06.24; accepted 16.08.24; published 13.11.24.

Please cite as:
KC S, Papoutsi C, Reidy C, Gudgin B, Powell J, Majeed A, Greaves F, Laverty AA
Differences in Use of a Patient Portal Across Sociodemographic Groups: Observational Study of the NHS App in England
J Med Internet Res 2024;26:e56320
URL: https://www.jmir.org/2024/1/e56320
doi:10.2196/56320
PMID:

©Sukriti KC, Chrysanthi Papoutsi, Claire Reidy, Bernard Gudgin, John Powell, Azeem Majeed, Felix Greaves, Anthony A Laverty. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Performance of a Full-Coverage Cervical Cancer Screening Program Using on an Artificial Intelligence– and Cloud-Based Diagnostic System: Observational Study of an Ultralarge Population

Lu Ji[1], PhD; Yifan Yao[1], PhD; Dandan Yu[2], BD; Wen Chen[1], MD; Shanshan Yin[1], MD; Yun Fu[3], BD; Shangfeng Tang[1], PhD; Lan Yao[1], PhD

[1]School of Medicine and Health Management, Tongji Medical College of Huazhong University of Science and Technology, Wuhan, China
[2]The Fifth Hospital of Wuhan, Wuhan, China
[3]Landing Artificial Intelligence Industry Research Institute, Wuhan, China

Corresponding Author:
Lan Yao, PhD
School of Medicine and Health Management
Tongji Medical College of Huazhong University of Science and Technology
13 Hangkong Road
Wuhan, 430030
China
Phone: 86 027 83692727
Email: ylhuster@163.com

Abstract

Background: The World Health Organization has set a global strategy to eliminate cervical cancer, emphasizing the need for cervical cancer screening coverage to reach 70%. In response, China has developed an action plan to accelerate the elimination of cervical cancer, with Hubei province implementing China's first provincial full-coverage screening program using an artificial intelligence (AI) and cloud-based diagnostic system.

Objective: This study aimed to evaluate the performance of AI technology in this full-coverage screening program. The evaluation indicators included accessibility, screening efficiency, diagnostic quality, and program cost.

Methods: Characteristics of 1,704,461 individuals screened from July 2022 to January 2023 were used to analyze accessibility and AI screening efficiency. A random sample of 220 individuals was used for external diagnostic quality control. The costs of different participating screening institutions were assessed.

Results: Cervical cancer screening services were extended to all administrative districts, especially in rural areas. Rural women had the highest participation rate at 67.54% (1,147,839/1,699,591). Approximately 1.7 million individuals were screened, achieving a cumulative coverage of 13.45% in about 6 months. Full-coverage programs could be achieved by AI technology in approximately 1 year, which was 87.5 times more efficient than the manual reading of slides. The sample compliance rate was as high as 99.1%, and compliance rates for positive, negative, and pathology biopsy reviews exceeded 96%. The cost of this program was CN ¥49 (the average exchange rate in 2022 is as follows: US $1=CN ¥6.7261) per person, with the primary screening institution and the third-party testing institute receiving CN ¥19 and ¥27, respectively.

Conclusions: AI-assisted diagnosis has proven to be accessible, efficient, reliable, and low cost, which could support the implementation of full-coverage screening programs, especially in areas with insufficient health resources. AI technology served as a crucial tool for rapidly and effectively increasing screening coverage, which would accelerate the achievement of the World Health Organization's goals of eliminating cervical cancer.

(J Med Internet Res 2024;26:e51477) doi:10.2196/51477

KEYWORDS

full coverage; cervical cancer screening; artificial intelligence; primary health institutions; accessibility; efficiency

XSL•FO
RenderX

Introduction

Cervical cancer can be largely prevented through early intervention [1]. However, it remains the fourth leading cause of cancer-related deaths in women worldwide. The Global Cancer Statistics for 2020 reported approximately 600,000 new cases of and 340,000 deaths from cervical cancer [2]. In response to this significant health challenge, the World Health Organization (WHO) launched the Global Strategy for the Elimination of Cervical Cancer in 2020, highlighting 3 key interventions: human papillomavirus (HPV) vaccination, cervical cancer screening, and cervical disease treatment [3]. The target for coverage of cervical cancer screening in specific populations is to reach 70% and above. In 2021, WHO updated its cervical cancer screening guidelines, providing recommendations to different countries and regions [4]. However, many countries with the highest burden of disease face limitations in achieving sufficient screening coverage [5]. In low- and middle-income countries (LMICs), only 9%-11% of women have been screened, in contrast to the 84% screening rate in high-income countries [6].

The national cervical cancer screening program implemented by the Chinese government in 2009 has played a crucial role in cervical cancer prevention and control (Multimedia Appendix 1). However, there still exists a considerable gap between the current cervical cancer screening rate and the target set by the WHO [7,8]. The economic burden of cervical cancer disease accounts for 18.2% of the global burden [2], and its incidence and mortality rates are 11.4% and 3.4%, respectively [9], with both being the highest in the central region, including Hubei province. The screening rate for Chinese women aged 35-64 years is approximately 36.8%, with lower rates observed in rural areas compared to urban areas [10]. Affordability and service accessibility are still the key factors influencing Chinese women's use of cervical cancer screening services, followed by education, health insurance, occupation, and family income [11]. Organized free screening can significantly increase their willingness to be screened. While China has already screened 130 million individuals, there are still 300 million people to be screened. Screening technology is a critical factor affecting the increase of screening coverage. Currently, cytology screening is the primary method used in China's national screening program. However, the shortage of pathologists in primary health institutions has limited the widespread adoption of cytology screening [12]. Studies have also indicated higher cervical cancer mortality rates in rural areas compared to urban areas [9,13]. Therefore, the selection of appropriate technology to rapidly enhance screening coverage, particularly in regions with limited health resources, is an urgent issue in China today.

Artificial intelligence (AI) technology has the potential to address the issue of insufficient pathologists in primary health institutions and accelerate the improvement of screening coverage rates. Therefore, this technology has been implemented in the Hubei government's organizational screening for 5 years. Now, Hubei province has become the first in China to implement full-coverage of cervical cancer screening for women of appropriate age by adopting AI technology, namely completing cervical cancer screening for 12.67 million urban and rural women in 3 years [14]. AI technology has a wide range of applications in cancer screening with favorable diagnostic performance [15-19]. A few studies conducted on Chinese populations have explored the use of AI technology in cervical cancer screening and diagnosis [11,20,21], comparing its compliance to manual reading, identifying high-grade lesions, and developing AI algorithms [22-24]. Further, 1 recent study compared AI technology with traditional cytology screening and HPV testing, demonstrating comparable cost-effectiveness of AI technology versus HPV testing [25]. Another study focused on the application of AI technology in population screening, including the roles and costs of health care workers, as well as measures to overcome challenges in online screening [26]. Although previous studies have reported results of internal quality control, such as compliance rates between AI and manual groups, evidence of external quality control is lacking. Furthermore, the screening efficiency of AI for ultralarge populations, the speed of improving coverage, the quality of screening, and the responsibilities and costs of the different screening organizations involved are not known. The performance of HPV testing for screening in the population has been reported [27-29], but the performance of AI technology in real-world screening is lacking and there is an urgent need to understand the performance of AI technology in full-coverage screening.

We described the governing and implementing agencies of the full-coverage screening program and their responsibilities, as well as how AI technology made the program a reality. More importantly, this study aimed to evaluate the performance of an AI and cloud diagnostic system in the first full-coverage cervical cancer screening program in China. The evaluation encompassed 4 key dimensions: accessibility of screening services, screening efficiency, diagnostic quality, and cost of different institutions. The assessment provided real-world evidence of the adoption of AI technology to improve screening coverage for other provinces in China and LMICs.

Methods

Full-Coverage Screening Program

The full-coverage screening program consists of 2 interconnected systems, a management system led by the provincial health committee and an implementation system led by the maternal and child health hospitals at the county and district levels (Figure 1). These 2 systems work collaboratively to ensure the successful implementation and coordination of the screening program. The health commission is responsible for program development, implementation guidance, project quality control, and evaluation. The cervical cancer screening implementation body consists of primary screening institutions, sampling institutions, and third-party testing institutions. In the implementation of the screening program, the key technology that makes the full-coverage plan a reality is the application of the AI and cloud diagnostic platform.

XSL•FO
RenderX

Figure 1. Management and implementation of the full-coverage program.

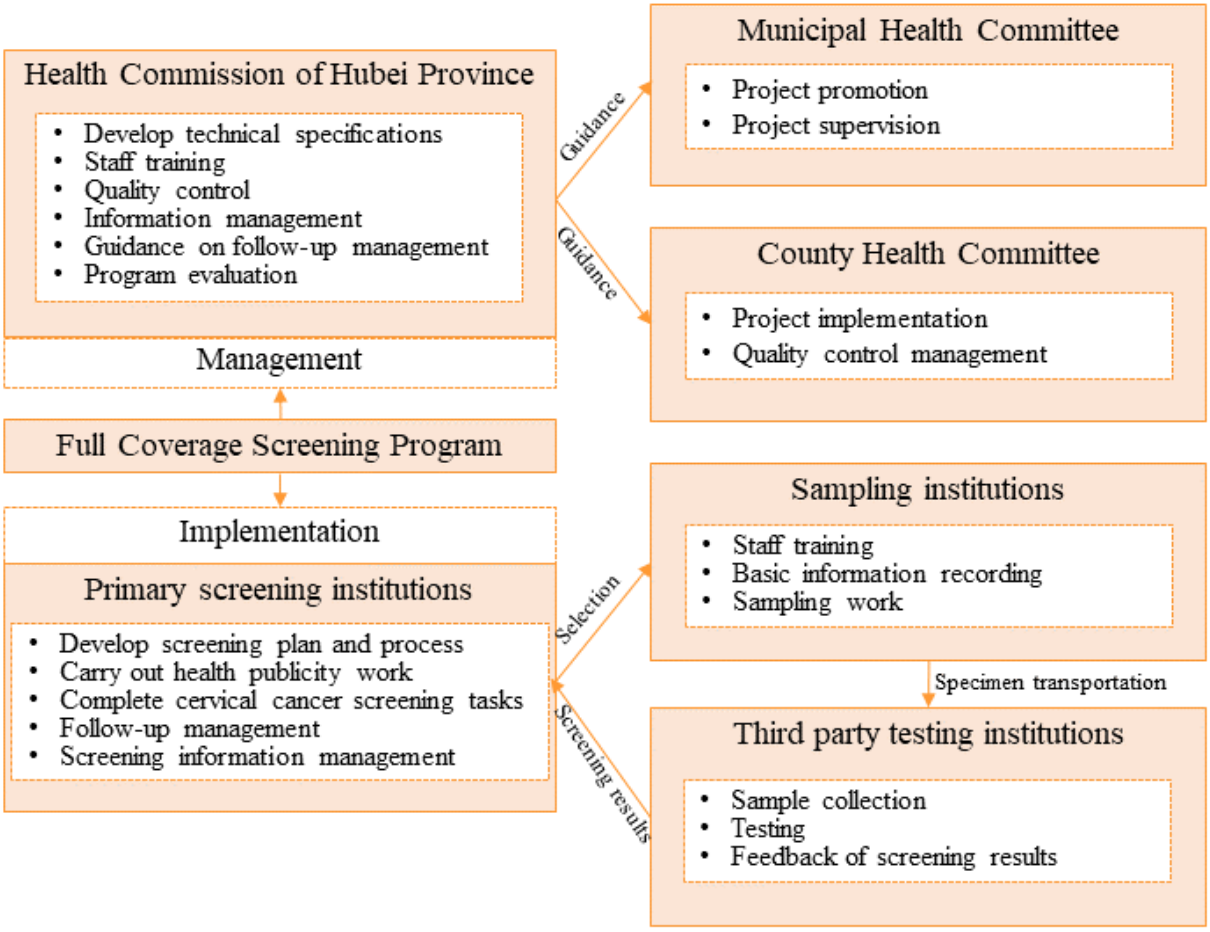

AI and Cloud Diagnostic System

Principles of AI-Assisted Liquid-Based Cytology

The DNA content of cervical epithelial exfoliated cells is correlated with cervical carcinogenesis. In the AI-assisted diagnostic process, images of cell samples are captured, and the cell nuclei within these images are segmented. For accurate diagnosis, normal cells' nuclei within a single sample group are the reference in the AI's algorithm. The segmented nuclei are then quantified based on their cellular DNA content, which ranges from 0 to 255 pixels. Then, by analyzing the pixels, the AI obtains the information on the cell nucleus DNA, and finally, the parameters such as the DNA content of the cell nucleus in this sample are counted to figure out the type of cell lesion. The specific algorithms involved in AI diagnosis are reported in previous studies [24].

Screening Procedure

Step 1: Sample Acquisition

Women can access free cervical cancer screening services at primary health institutions by presenting their ID cards (Figure 2). At the same time, they are required to sign an informed consent form. Labels with QR codes are automatically produced through identification of the ID card. The staff of the primary health institutions uses a special cervical brush to collect cervical exfoliated cells, which are preserved in sample preservation liquids. Cell samples are transported to the Landing laboratory, a government-appointed testing facility. The information of the whole process from sample collection to diagnostic report is available through a QR code.

XSL•FO
RenderX

Figure 2. Screening process and quality control. TBS: The Bethesda System.

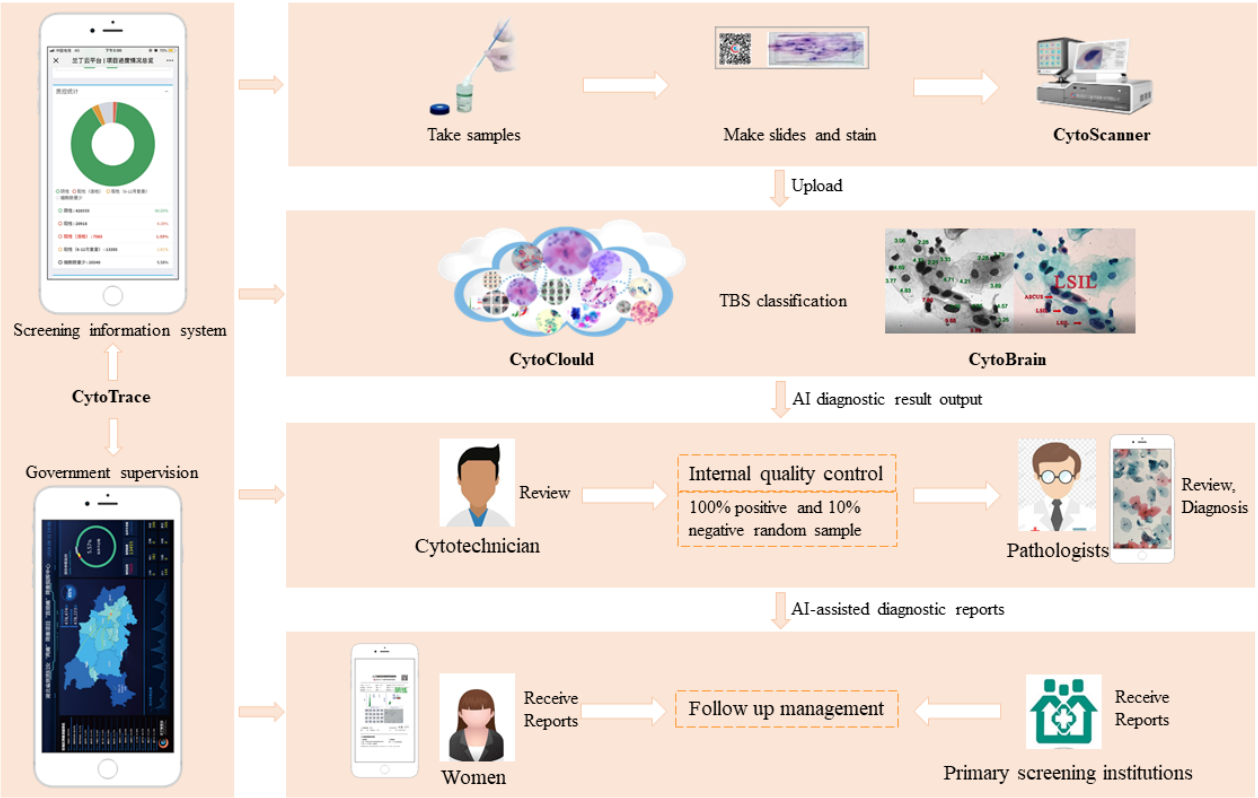

Step 2: AI-Assisted Liquid-Based Cytology Diagnosis

Laboratory technicians make slides of the samples. The slides are scanned and the cell images are uploaded to the cloud platform. Trained and validated algorithms within the AI diagnostic engine analyze the cellular features present in the images. The results of the AI diagnosis are reviewed by cytotechnicians who identify negative cases and retain suspicious or positive cases for further evaluation.

Step 3: Quality Control

To ensure accuracy, all suspicious or positive cases, as well as a randomly selected 10% of negative cases, are mixed. The pathologist then diagnoses these cases and diagnostic reports will be issued. The entire process from AI reading to cytotechnicians review and pathologist review belongs to the internal quality control of our AI-assisted diagnostic system. Importantly, pathologists have the flexibility to work from any location, including their homes or even abroad, as long as they can connect to the cloud platform using their electronic devices.

Step 4: Results Notification

Diagnostic reports with negative screening results are sent to the primary health institutions in a centralized file bag. For positive screening results, the diagnostic reports are sent to the primary health institutions in a separate packet. Facility staff are responsible for notifying examinees. They can also obtain the information by scanning the receipt with a QR code through WeChat, which is the fastest way.

Step 5: Follow-Up Management

Based on the results of cytological screening feedback, primary health institutions follow up on examinees with abnormal initial screening results. Those with normal cytology results will be screened once every 3 years. If the cytology results are abnormal or suspicious, colposcopy is performed directly. If the colposcopy is negative, then a cytology reexamination is needed after 1 year. If colposcopy is positive, a histopathologic biopsy is performed and follow-up treatment is provided for those patients who require therapy.

Quality Control Process

Quality control consists of internal quality control and external quality control (Figure 2). Internal quality control (for details see step 3) is completed by 4 pathologists with senior titles from the landing laboratory, and the frequency of rechecking is daily. Besides, the internal diagnostic quality control is performed using a fully digital process, involving quality control in making slides, staining, and diagnosis. The external quality control is to invite pathologists from tertiary care hospitals to randomly select slides for a recheck regularly, including to recheck the results of cytology diagnosis and pathology diagnosis. That is, external quality control is a reexamination of internal quality control. Quality control adheres to the “Quality Assurance and Quality Control Guidelines for Cervical Cancer Screening” and the “Cervical Cancer Screening Quality Assessment Manual (2022 Edition).”

Implementation Stage

The application of the AI and cloud diagnostic system has progressed through 3 stages.

Phase 1: Pilot Phase

This phase was mainly used in some municipal hospitals to improve the efficiency of institutional diagnosis and save the expenditure on institutional labor costs.

XSL•FO
RenderX

Phase 2: Provincial Organizational Screening Phase

In 2017, it was first applied in the cervical screening program organized by the provincial government, with a cumulative screening quantity of 1.46 million completed in the first 3 years.

Phase 3: Full-Coverage Program Phase

AI-assisted diagnostics was applied in the provincial government's full-coverage screening program. The technology was also being expanded in some countries in the Belt and Road Initiative, such as Pakistan.

Data Collection

We collected data on screening quantity, screening time, diagnostic quality, and cost. The screening data were collected from the initiation of the full-coverage screening program until January 31, 2023, with a total of 1,704,461 individuals and 18,261 histopathology biopsies. The screening program followed a closed-loop management process, encompassing screening, follow-up, and treatment for all examinees. We clarified the time and workload of AI-assisted diagnosis through the Wuhan landing cloud intelligence laboratory. Additionally, the daily workload of pathologists was estimated based on guidelines and literature studies, such as the Guidelines for Construction and Management of Pathology Departments (Trial) and literature studies. To evaluate the diagnostic quality, an external quality assessment was organized. The sample size was 220, including 170 randomly selected cytology samples and 50 histopathology samples. Data on the cost of full-coverage screening were obtained from experts of the Health Commission of Hubei Province.

Ethical Considerations

This study did not expose women's privacy or identifying information, nor did it result in associated economic or social risks. Women participated in cervical cancer screening voluntarily and signed an informed consent form before being screened. The consent included that their screening information could be used for big data research. This study was conducted in compliance with the Declaration of Helsinki principles and approved by the ethics committee of Tongji Medical College, Huazhong University of Science and Technology (number [2023]IEC(A235)).

Results

Accessibility

Cervical cancer screening services were extended to all administrative districts, including 13 prefecture-level administrative regions and 4 directly administered county-level administrative regions in Hubei Province, and a total of 1,704,461 women had been screened (Table 1). The number of county administrative districts covered by the screening program has increased from 83 in 2020 to 103 at present, of which 39 are in urban areas and 64 are in rural areas. The completed screening coverage rate was 13.45% (Multimedia Appendix 2). Among them, Huanggang city had the highest coverage rate of 18.91%.

Table 1. Accessibility of cervical cancer screening services in Hubei province.

Initiated cities or counties	Women screened, n	Women aged 35-64 years, n	Screening coverage, %
Wuhan	223,603	2,731,324	8.19
Huangshi	71,423	541,774	13.18
Shiyan	65,536	704,131	9.31
Yichang	124,646	854,964	14.58
Xiangyang	171,260	1,154,377	14.84
Ezhou	28,879	236,836	12.19
Jingmen	81,348	569,827	14.28
Xiaogan	161,670	937,020	17.25
Jingzhou	185,026	1,147,845	16.12
Huanggang	244,097	1,290,808	18.91
Xianning	95,935	583,297	16.45
Enshi	128,674	758,358	16.97
Suizhou	61,633	449,363	13.72
Xiantao	35,249	248,983	14.16
Tianmen	15,986	254,233	6.29
Qianjiang	6986	194,529	3.59
Shennongjia	2510	14,607	17.18
Total	1,704,461	12,672,278	13.45

The screening project was officially launched in September 2022, with 1.54 million individuals screened in 5 months. Before the launch of the project, there were screening pilots in some areas in July (n=155,103) and August (n=9417). The amount

XSL•FO
RenderX

of screening was limited, except for Wuhan, where a total of 153,747 individuals were screened in July. A screening peak was rapidly formed after the launch, including a screening volume of about 498,909 in November. The COVID-19 outbreak affected accessibility to screening services, most significantly in December (n=295,199) and January (n=87,760).

The highest active participation after the launch of this project was among rural women with 67.54% (n=1,147,839) of the total screening population. This was followed by urban nonunion women with 27.91% (n=474,420; Table 2). In total, 160,376 (94.07%) of these screenings were normal. Further, 199,473 (5.85%) were cytologically positive, indicating a need for referral for colposcopy. The positive rate of AI-assisted cytological screening was 5.84%. In the pathological examination performed in the project, 1863 individuals were examined for precancerous lesions and 93 individuals were examined for cervical cancer.

Table 2. Basic characteristics of examinees.

Characteristics	Examinees, n (%)
Population categories[a]	
Urban nonunion	474,420 (27.91)
Urban unions	65,184 (3.84)
Rural women of appropriate age	1,147,839 (67.54)
Women workers in difficulty	12,148 (0.71)
Cytology screening results	
NILM[b]	1,603,376 (94.07)
ASC-US[c]	68,637 (4.03)
LSIL[d]	17,977 (1.05)
ASC-H[e]	9613 (0.56)
HSIL[f]	2129 (0.12)
AGC[g]	1117 (0.07)
Failure	1612 (0.09)
Histopathological results[h]	
Cervicitis	10,645 (58.29)
CIN1[i]	5660 (31)
CIN2/3[j]	1863 (10.2)
CA[k]	93 (0.51)

[a]At the beginning of project implementation, 4870 women were not classified in the population categories.
[b]NILM: negative for intraepithelial lesion or malignancy.
[c]ASC-US: atypical squamous cells of undetermined significance.
[d]LSIL: low-grade squamous intraepithelial lesion.
[e]ASC-H: atypical squamous cells, cannot exclude a high-grade lesion.
[f]HSIL: high-grade squamous intraepithelial lesion.
[g]AGC: atypical glandular cell.
[h]Some biopsy results were not available because women were screened at program sites.
[i]CIN1: cervical intraepithelial neoplasia grade 1.
[j]CIN2/3: cervical intraepithelial neoplasia grade 2 and 3, including cervical cancer in situ.
[k]CA: cervical cancer.

Screening Efficiency

Based on real-world screening volumes, full-coverage programs could be achieved in approximately 1 year by using AI technology (Table 3). For the number of pathologists involved in full coverage, AI technology was 87.5 (35,000/400) times more efficient than manual reading of slides. Based on the screening capacity of primary health institutions [12], it would take approximately 16 years to complete full coverage of the program's screening population. This would be even longer if repeat screening every 3 years is taken into account.

XSL•FO
RenderX

Table 3. AI[a]-assisted LBC[b] versus manual cytology diagnosis.

Screening Method	Scenes	Screening cases/d	Completed 1.7 million/d[c]	Full coverage of Hubei/d[d]
AI-assisted diagnosis	Real-world	35,000[e]	49	362
Pathologists	Hypothetical scenario[f]	400	4250	31,675
Pathologists	Hypothetical scenario[g]	2100	810	6033
Pathologists	Hypothetical scenario[h]	2200	773	5759

[a]AI: artificial intelligence.

[b]LBC: liquid-based cytology.

[c]Screening volume completed as of January 31, 2023, for the project, approximately 1.7 million.

[d]Cervical cancer full-coverage screening program in Hubei province is about 12.67 million.

[e]The current maximum daily screening volume in the real-world is 35,000 individuals, which could continue to expand.

[f]Workload of each pathologist is 100 individuals per day according to the recommended standards. This artificial intelligence–assisted diagnosis system has 4 pathologists. It is assumed that these 4 pathologists used manual reading to do cytology diagnosis.

[g]The survey of screening capacity in 14 provinces in the national cervical cancer screening program found that it took 1003.64 days for urban areas to test 2,107,648 individuals with cytology screening, which means that the urban screening quantity was about 2100 individuals per day.

[h]The same study as above found that the quantity of cytology screening used in rural areas was 2200 (1,120,136/509.15) individuals per day.

Diagnostic Quality

A total of 220 individuals were reassessed for quality, including 170 cytology slides and 50 histopathology slides (Multimedia Appendix 3). According to the criteria in the cervical cancer quality assessment handbook [30], the specimen satisfaction rate, the compliance rate of negative and positive smear rechecks, and the recheck rate of pathological examination results all met the quality requirements (Table 4). The AI-assisted diagnosis had a 99% compliance rate for positive smears and a 100% compliance rate for negative smears when compared to the cytologist's diagnosis. All negative cases and CIN2+ (cervical intraepithelial neoplasia) cases were exact matches, with a small number of deviations mainly in low-grade lesions. Compared to pathologists, AI diagnostic systems were less likely to miss or underestimate the severity of cervical lesions.

Table 4. Quality assessment indicators and results.

Evaluation indicators	Rate (%)	Related evaluation criteria (%)
Satisfactory rate of specimens[a]	99.1	≥95
Positive smear recheck compliance rate[b]	99	≥85
Negative smear recheck rate	100	≥95
Compliance rate of pathological examination	96	≥95

[a]The assessment method is on-site assessment.

[b]The judgment criteria for compliance is a difference of 2 levels or less.

Cost

The per capita cost of screening in the full-coverage program is CN ¥49 (US $6.89; Table 5), while the price of health services in primary health institutions is about CN ¥160 (US $22.48). The primary health institutions and third-party testing institutions would receive CN ¥19 (US $2.67) and CN ¥27 (US $3.79) per case, respectively. The primary health institutions are responsible for only cytology sampling, colposcopy examination, and histopathology sampling throughout the screening process. Ultimately the primary health institutions can obtain the cost of services 2, 6, and 7 in Table 5. Third-party testing laboratories provide consumables and bear the costs of sample transportation, subsequent diagnostics, management of information systems, etc. The staff of the Women's Union were involved in publicity and advocacy, so they received CN ¥3 (US $0.42) per case.

XSL•FO
RenderX

Table 5. Screening program costs.

Number	Service	Cost (CN ¥)[a]
1	Training cost for sampling workers, marketing publicity, and organizational management cost	1.5
2	Sampling cost for primary screening institutions	11
3	Logistics cost for samples	1
4	Consumables cost	5
5	Slides, staining, and examination cost	15
6	Colposcopy (for those with a positive primary cytology screening)	5
7	Histopathological diagnosis cost (for those with positive colposcopy)	3
8	Development operation and maintenance costs for the screening information management system	4
9	External quality control and evaluation cost	0.5
10	Cost for the organization and promotion of screening participants	3
11	Total	49

[a]The average exchange rate in 2022 is as follows: US $1=CN ¥6.7261.

Discussion

Principal Findings

The diagnostic system based on an AI and cloud platform can achieve coverage of the whole population for cervical cancer screening, especially in areas with insufficient health resources. Cervical cancer screening services were extended to all administrative districts, the number of which has increased from 83 in 2020 to 103 at present. With the voluntary participation of women, a total of 170,461 individuals were screened between July 2022 and January 31, 2023, achieving a coverage rate of 13.45%. Rural women accounted for the largest proportion of screenings at 67.54% (1,147,839/1,699,591). In the screening program, 1,603,376 (94.07%) were diagnosed as normal, 1612 (1612/1,704,461, 0.09%) samples failed, with 1863 (1863/18,261, 10.20%) cases of precancerous lesions and 93 (93/18,261, 0.51%) cases of cervical cancer. The positive rate of AI-assisted cytological screening was 5.84%. Based on the maximum screening quantity of AI-assisted LBC in the real-world, full-coverage programs could be achieved in approximately 1 year, which was 87.5 times more efficient than manual reading of slides. Compared to real-world screening capacity, AI technology could accomplish full-coverage screening amounts at least 15 years ahead of schedule. The review compliance rates of positive and negative smears and pathological biopsies were 96% and above, meeting national quality control requirements. From a public payment perspective, the cost of this program was CN ¥49 (US $6.89) per person, with the primary screening institution and the third-party testing institute receiving CN ¥19 (US $2.67) and CN ¥27 (US $3.79), respectively. In short, AI-assisted diagnosis has proven to be accessible, efficient, reliable, and cost-effective, and plays an important role in facilitating improvements in screening coverage, especially in areas with insufficient health resources. To our knowledge, this study reports the largest sample of AI-assisted cervical cancer screening and is the first to evaluate the performance of AI technology in a full cervical cancer screening program. AI technology plays an important role in promoting screening coverage and opening up new pathways for the elimination of cervical cancer in China and other LMICs.

AI technology has greatly enhanced the accessibility of screening services, especially for rural women. Improving the equity and accessibility of screening services is an important issue for the elimination of cervical cancer [31]. China mainly adopts cytology screening, and the lack of pathologists limits the large-scale application of this technique, especially in remote or rural areas. The full-coverage screening covered all county-level administrative regions in China, with the largest number of rural women screened at 1,147,839 (67.54%), far more than urban women. This is on the one hand because screening services do increase the accessibility of screening services for rural women, and on the other hand, because of the long-term implementation of the screening program in rural areas, rural women have a high level of recognition and willingness to participate in the program. Previous studies have also shown that tissue-based screening can increase participation in screening [32,33]. In addition, according to the geographical characteristics of the county, the implementers adopted the approach of going to the village for screening or organizing women to be screened at the township health center for remote areas to ensure the convenience and accessibility of screening services. AI technology can not only solve the screening problem in remote or poor areas of China but also solve the cervical cancer screening problem in Pakistan by establishing laboratories and exporting equipment abroad. This technology will benefit more Belt and Road Initiative countries in the future.

AI-assisted LBC has greatly enhanced screening efficiency and accelerated the achievement of WHO's screening goals. With a real-world maximum daily diagnostic workload of 35,000 cases, the AI system will achieve full cervical cancer screening coverage in less than a year, 2 years earlier than expected. Further, the daily diagnostic volume can be continuously increased, which can be achieved by adding high-throughput scanners and technical staff. This screening was completed in about 6 months with a coverage of 13.45% (1,704,461/12,672,278) of women in Hubei, which was huge

XSL•FO
RenderX

progress. A survey of the screening capacity of primary health institutions found that some women may not be able to access screening services once in their lifetime [12]. In contrast, AI screening efficiency in the real world was approximately 16 times more efficient than the screening efficiency of primary health institutions. The coverage of screening has a significant impact on cervical cancer incidence and mortality [34-36], so the high efficiency of AI technology also accelerates the process of achieving the goal of cervical cancer elimination.

Diagnostic quality is an important evaluation indicator, including the consistency between AI and manual reading, the failure rate, and the positive cytology screening rate. The national cervical cancer cytology screening positivity rate is 4%, while HPV testing is as high as 10.1% [28]. The cytology screening misdiagnosis rate was lower than that of HPV testing and the leakage rate was higher than that of HPV testing. The positive rate of AI-assisted cytological screening was 5.84%, which was higher than traditional manual reading. This may be attributed to the improved diagnostic performance of AI-assisted cytological screening.

The total number of samples screened this time was about 1.7 million, with only 0.9% (1612/1,704,461) failure samples. Compared to the 3.3% [23] and 2.91% [26] in the previous study, the full-coverage program demonstrated better sample pass rates with the largest sample sizes. This may be due to the continuous training of technicians and strict quality control during the implementation of the full-coverage program. In a previous study with a sample of about 0.7 million, the overall agreement rate between AI and manual reading was 94.7% [23], which belonged to internal quality control. In the census population, AI was significantly more sensitive than manual reading for CIN2+/CIN3+ respectively, while the corresponding specificity was slightly reduced when compared to cytotechnicians. In suspicious populations, AI-assisted reading has similar sensitivity in detecting CIN2+/CIN3+, but is more specific than manual reading [22]. Our study adds another aspect regarding the rate of agreement, namely the consistency of external quality assessments. External quality control was a reevaluation of the outcomes of internal quality control. Additionally, the results of the external quality evaluation showed that the compliance rates for cytology screening and histopathology were 99%-100% and 96%, respectively. This all met the requirements for quality control of cervical screening in China and the WHO.

AI-assisted LBC has a favorable price advantage to support large-scale screening. Price is a very important consideration for the government to purchase this screening service and provide it free to women, especially for the population group. The per capita cost of this full-coverage screening program is CN ¥49 (US $6.89). In comparison, hospitals may charge almost 3 to 4 times (about CN ¥160 [US $22.48]) as much. CN ¥3 (US $0.42) of the project cost is for advocacy and mobilization. However, the cost of the project will continue to decrease over time with the widespread use of social media and the increased willingness to participate. Previous studies have also confirmed that AI-assisted diagnosis is more economical than HPV testing [25]. The affordability of screening services is an important factor affecting the availability of screening services [8,37]. Low cost offers the possibility of achieving screening in large-scale populations, especially in poor areas.

Policy Implication

The full-coverage screening program in Hubei province is the first provincial coverage program and could provide useful lessons for other cities and regions in China, as well as other LMICs, to rapidly increase screening coverage. First, AI-assisted diagnosis technology has a lower cost and reliable quality to reach as many people as possible with limited government funding and to enhance screening coverage. Second, AI technology can address the challenges of primary health institutions, especially the accessibility of screening services in remote or poor areas. Thus, coverage of vulnerable populations will be achieved and equity in screening services will be improved. Finally, AI technology has tremendous capacity advantages that can rapidly improve screening coverage and shorten the process of achieving the target of 70% screening coverage.

Unlike mass population-based screening, precision screening can be implemented in remote, poor, and financially resource-limited areas, even in LMICs. The target population for precision screening may come from areas with high incidence or mortality of cervical cancer, remote and inaccessible areas, or areas with low screening coverage. The cytology screening approach of AI-assisted diagnosis not only addresses screening in areas with insufficient health resources, but more importantly, through the screening program, improves the screening capacity of staff in the area, follow-up management capacity, and health service providers' treatment capacity, forming a virtuous cycle of service provision and capacity enhancement.

Limitation

Women's participation in screening is a voluntary selection process. Thus, patients who were diagnosed with a positive screening result did not necessarily choose a county or district health service again when they needed further screening or treatment. This is mainly reflected in the absence of pathological results of screened women in the AI information system. This part of the population often chooses to go to the best municipal hospitals for further screening or treatment. This to some extent wastes health resources and makes it more difficult to manage the follow-up of screened women. The challenge of inducing a rational order for medical treatment is also a problem that China's health care reform is addressing. However, the AI-assisted diagnosis system has facilitated the follow-up management of women's screening and the overall improvement of women's health, thanks to the powerful information management system of the AI-assisted diagnosis system, which can keep track of the implementation of the screening program in the whole region and the information of all the test results of screened women.

Conclusion

Our AI-assisted diagnostic system can complete cervical cancer screening in a short period with low cost and high quality, and make up for the shortage of pathologists in primary health institutions. It has been proven to perform well in large-scale

XSL•FO
RenderX

screening and is an effective pathway to address the lack of health resources, especially in areas with insufficient pathologists. AI diagnostics will help China achieve its goal of eliminating cervical cancer ahead of schedule, and also provide a reference for other LMICs and WHO screening guidelines.

Acknowledgments

This work was supported by the Fundamental Research Funds for Central Universities (YCJJ202204019) and the Health Commission of Hubei Province (WJ2023Z008).

Authors' Contributions

LJ reviewed the literature, analyzed the data, and wrote this paper, and ST was mainly responsible for the framework design of this paper. LY advised on the final version of this paper. YY, WC, and SY helped review the literature. DY and YF helped with data collection and collation. All authors contributed to this paper and agreed to the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cervical cancer screening policy in China.
[DOCX File , 19 KB - jmir_v26i1e51477_app1.docx]

Multimedia Appendix 2
Number of women and screening coverage.
[DOCX File , 20 KB - jmir_v26i1e51477_app2.docx]

Multimedia Appendix 3
Quality assessment indicators and data.
[DOCX File , 20 KB - jmir_v26i1e51477_app3.docx]

References

1. Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. Lancet 2019;393(10167):169-182. [doi: 10.1016/S0140-6736(18)32470-X] [Medline: 30638582]
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
3. Global strategy to accelerate the elimination of cervical cancer as a public health problem. WHO. 2020. URL: https://www.who.int/publications/i/item/9789240014107 [accessed 2022-10-21]
4. WHO guideline for screening and treatment of cervical pre-cancer lesions for cervical cancer prevention, second edition. 2021. URL: https://www.who.int/publications/i/item/9789240040434 [accessed 2023-04-09]
5. Lemp JM, De Neve JW, Bussmann H, Chen S, Manne-Goehler J, Theilmann M, et al. Lifetime prevalence of cervical cancer screening in 55 low- and middle-income countries. JAMA 2020;324(15):1532-1542 [FREE Full text] [doi: 10.1001/jama.2020.16244] [Medline: 33079153]
6. Bruni L, Serrano B, Roura E, Alemany L, Cowan M, Herrero R, et al. Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis. Lancet Glob Health 2022;10(8):e1115-e1127 [FREE Full text] [doi: 10.1016/S2214-109X(22)00241-8] [Medline: 35839811]
7. GBD 2019 Universal Health Coverage Collaborators. Measuring universal health coverage based on an index of effective coverage of health services in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet 2020 Oct 17;396(10258):1250-1284 [FREE Full text] [doi: 10.1016/S0140-6736(20)30750-9] [Medline: 32861314]
8. Zhao F, Qiao Y. Cervical cancer prevention in China: a key to cancer control. Lancet 2019;393(10175):969-970. [doi: 10.1016/S0140-6736(18)32849-6] [Medline: 30860036]
9. Zheng R, Zhang S, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2016. J Natl Cancer Cent 2022;2(1):1-9 [FREE Full text] [doi: 10.1016/j.jncc.2022.02.002] [Medline: 39035212]
10. Zhang M, Zhong Y, Wang L, Bao H, Huang Z, Zhao Z, et al. Cervical cancer screening coverage—China, 2018-2019. China CDC Wkly 2022 Dec 02;4(48):1077-1082 [FREE Full text] [doi: 10.46234/ccdcw2022.217] [Medline: 36751373]
11. Ji L, Chen M, Yao L. Strategies to eliminate cervical cancer in China. Front Oncol 2023;13:1105468 [FREE Full text] [doi: 10.3389/fonc.2023.1105468] [Medline: 37333817]

XSL•FO
RenderX

12. Li Y, Ma L, Yang C, Chen Z, Zhao Y, Dang L, et al. A study on service capacity of primary medical and health institutions for cervical cancer screening in urban and rural areas in China. Chin J Cancer Res 2019;31(5):838-848 [FREE Full text] [doi: 10.21147/j.issn.1000-9604.2019.05.13] [Medline: 31814687]
13. Sun K, Zheng R, Lei L, Zhang S, Zeng H, Wang S, et al. Trends in incidence rates, mortality rates, and age-period-cohort effects of cervical cancer—China, 2003-2017. China CDC Wkly 2022;4(48):1070-1076 [FREE Full text] [doi: 10.46234/ccdcw2022.216] [Medline: 36751372]
14. 【Hubei Daily】The province will provide free cervical cancer screening for 12.67 million urban and rural women of the appropriate age—Hubei Provincial Health and Wellness Commission. URL: http://wjw.hubei.gov.cn/bmdt/mtjj/mtgz/202201/t20220125_3979707.shtml [accessed 2022-11-09]
15. Wentzensen N, Lahrmann B, Clarke MA, Kinney W, Tokugawa D, Poitras N, et al. Accuracy and efficiency of deep-learning-based automation of dual stain cytology in cervical cancer screening. J Natl Cancer Inst 2021;113(1):72-79 [FREE Full text] [doi: 10.1093/jnci/djaa066] [Medline: 32584382]
16. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. J Natl Cancer Inst 2019;111(9):923-932 [FREE Full text] [doi: 10.1093/jnci/djy225] [Medline: 30629194]
17. Yu K, Hyun N, Fetterman B, Lorey T, Raine-Bennett TR, Zhang H, et al. Automated cervical screening and triage,based on HPV testing and computer-interpreted cytology. J Natl Cancer Inst 2018;110(11):1222-1228 [FREE Full text] [doi: 10.1093/jnci/djy044] [Medline: 29659930]
18. Holmström O, Linder N, Kaingu H, Mbuuko N, Mbete J, Kinyua F, et al. Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. JAMA Netw Open 2021;4(3):e211740 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1740] [Medline: 33729503]
19. Rubin R. Artificial intelligence for cervical precancer screening. JAMA 2019;321(8):734. [doi: 10.1001/jama.2019.0888] [Medline: 30806677]
20. Hou X, Shen G, Zhou L, Li Y, Wang T, Ma X. Artificial intelligence in cervical cancer screening and diagnosis. Front Oncol 2022;12:851367 [FREE Full text] [doi: 10.3389/fonc.2022.851367] [Medline: 35359358]
21. Xue P, Ng MTA, Qiao Y. The challenges of colposcopy for cervical cancer screening in LMICs and solutions by artificial intelligence. BMC Med 2020;18(1):169 [FREE Full text] [doi: 10.1186/s12916-020-01613-x] [Medline: 32493320]
22. Bao H, Bi H, Zhang X, Zhao Y, Dong Y, Luo X, et al. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: a multicenter, clinical-based, observational study. Gynecol Oncol 2020;159(1):171-178. [doi: 10.1016/j.ygyno.2020.07.099] [Medline: 32814641]
23. Bao H, Sun X, Zhang Y, Pang B, Li H, Zhou L, et al. The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: a population-based cohort study of 0.7 million women. Cancer Med 2020;9(18):6896-6906 [FREE Full text] [doi: 10.1002/cam4.3296] [Medline: 32697872]
24. Chen H, Liu J, Wen QM, Zuo ZQ, Liu JS, Feng J, et al. CytoBrain: cervical cancer screening system based on deep learning technology. J Comput Sci Technol 2021;36(2):347-360. [doi: 10.1007/s11390-021-0849-3]
25. Shen M, Zou Z, Bao H, Fairley C, Canfell K, Ong JJ, et al. Cost-effectiveness of artificial intelligence-assisted liquid-based cytology testing for cervical cancer screening in China. Lancet Reg Health West Pac 2023;34:100726 [FREE Full text] [doi: 10.1016/j.lanwpc.2023.100726] [Medline: 37283979]
26. Zhu X, Yao Q, Dai W, Ji L, Yao Y, Pang B, et al. Cervical cancer screening aided by artificial intelligence, China. Bull World Health Organ 2023;101(6):381-390 [FREE Full text] [doi: 10.2471/BLT.22.289061] [Medline: 37265676]
27. Bao H, Ma L, Zhao Y, Song B, Di J, Wang L, et al. Age-specific effectiveness of primary human papillomavirus screening versus cytology in a cervical cancer screening program: a nationwide cross-sectional study. Cancer Commun (Lond) 2022;42(3):191-204 [FREE Full text] [doi: 10.1002/cac2.12256] [Medline: 35142100]
28. Zhao Y, Bao H, Ma L, Song B, Di J, Wang L, et al. Real-world effectiveness of primary screening with high-risk human papillomavirus testing in the cervical cancer screening programme in China: a nationwide, population-based study. BMC Med 2021;19(1):164. [doi: 10.1186/s12916-021-02026-0] [Medline: 34261463]
29. Zhang J, Zhao Y, Dai Y, Dang L, Ma L, Yang C, et al. Effectiveness of high-risk human papillomavirus testing for cervical cancer screening in China: a multicenter, open-label, randomized clinical trial. JAMA Oncol 2021;7(2):263-270 [FREE Full text] [doi: 10.1001/jamaoncol.2020.6575] [Medline: 33377903]
30. Cervical cancer quality assessment handbook. 2022. URL: https://www.chinawch.org.cn/tzgg2021/tzgg2_2021/202210/P020221024556061262306.pdf [accessed 2024-10-16]
31. So WKW, Chan DNS, Law BMH, Rana T, Wong CL. Achieving equitable access to cancer screening services to reduce the cancer burden in the Asia-pacific region:experience from Hong Kong. Lancet Reg Health West Pac 2022;29:100587 [FREE Full text] [doi: 10.1016/j.lanwpc.2022.100587] [Medline: 36605882]
32. Bao H, Wang L, Brown M, Zhang M, Hunt K, Di J, et al. A nationally quasi-experimental study to assess the impact of partial organized breast and cervical cancer screening programme on participation and inequalities. BMC Cancer 2020;20(1):1191 [FREE Full text] [doi: 10.1186/s12885-020-07686-4] [Medline: 33276761]

XSL•FO
RenderX

33. Gu C, Chan CWH, Chow KM, Yang S, Luo Y, Cheng H, et al. Understanding the cervical screening behaviour of chinese women: the role of health care system and health professions. Appl Nurs Res 2018;39:58-64. [doi: 10.1016/j.apnr.2017.09.009] [Medline: 29422178]
34. Brisson M, Kim JJ, Canfell K, Drolet M, Gingras G, Burger EA, et al. Impact of HPV vaccination and cervical screening on cervical cancer elimination: a comparative modelling analysis in 78 low-income and lower-middle-income countries. Lancet 2020;395(10224):575-590 [FREE Full text] [doi: 10.1016/S0140-6736(20)30068-4] [Medline: 32007141]
35. Canfell K, Kim JJ, Brisson M, Keane A, Simms KT, Caruana M, et al. Mortality impact of achieving WHO cervical cancer elimination targets: a comparative modelling analysis in 78 low-income and lower-middle-income countries. Lancet 2020;395(10224):591-603 [FREE Full text] [doi: 10.1016/S0140-6736(20)30157-4] [Medline: 32007142]
36. Simms KT, Steinberg J, Caruana M, Smith MA, Lew JB, Soerjomataram I, et al. Impact of scaled up human papillomavirus vaccination and cervical screening and the potential for global elimination of cervical cancer in 181 countries, 2020-99: a modelling study. Lancet Oncol 2019;20(3):394-407. [doi: 10.1016/S1470-2045(18)30836-2] [Medline: 30795950]
37. Zhao S, Huang L, Basu P, Domingo EJ, Supakarapongkul W, Ling WY, et al. Cervical cancer burden, status of implementation and challenges of cervical cancer screening in Association of Southeast Asian Nations (ASEAN) countries. Cancer Lett 2022;525:22-32. [doi: 10.1016/j.canlet.2021.10.036] [Medline: 34728309]

Abbreviations

AI: artificial intelligence
CIN: cervical intraepithelial neoplasia
HPV: human papillomavirus
LBC: liquid-based cytology
LMIC: low- and middle-income country
WHO: World Health Organization

Edited by A Schwartz; submitted 01.08.23; peer-reviewed by M Schiffman, RG Mishuris; comments to author 06.01.24; revised version received 10.03.24; accepted 09.09.24; published 20.11.24.

Please cite as:
Ji L, Yao Y, Yu D, Chen W, Yin S, Fu Y, Tang S, Yao L
Performance of a Full-Coverage Cervical Cancer Screening Program Using on an Artificial Intelligence– and Cloud-Based Diagnostic System: Observational Study of an Ultralarge Population
J Med Internet Res 2024;26:e51477
URL: https://www.jmir.org/2024/1/e51477
doi:10.2196/51477
PMID:

©Lu Ji, Yifan Yao, Dandan Yu, Wen Chen, Shanshan Yin, Yun Fu, Shangfeng Tang, Lan Yao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Themes in Abortion Forum Discussions in a Restrictive Access Context: Qualitative and Quantitative Analyses

Krzysztof Bartosz Klimiuk[1], MD; Amelia Kot[2]; Ewa Majcherek[3], MD; Katarzyna B Kubiak[1], MD; Łukasz Balwicki[1], MD, PhD, Prof Dr Med

[1]Department of Public Health and Social Medicine, Faculty of Health Sciences, Medical University of Gdańsk, Gdańsk, Poland

[2]Faculty of Medicine, Medical University of Gdańsk, Gdańsk, Poland

[3]Faculty of Medicine, Poznan University of Medical Sciences, Poznań, Poland

Corresponding Author:
Krzysztof Bartosz Klimiuk, MD
Department of Public Health and Social Medicine
Faculty of Health Sciences
Medical University of Gdańsk
Marii Skłodowskiej-Curie 3a
Gdańsk, 80-210
Poland
Phone: 48 583491111
Email: krzysztof_klimiuk@gumed.edu.pl

Abstract

Background: Abortion is one of the most common medical procedures worldwide. Despite this, access to abortion on demand remains restricted in many countries, including Poland. As a result, many women resort to undergoing the procedure without medical supervision, putting themselves at risk of serious health consequences such as drug poisoning, excessive bleeding, and hypovolemia. Unfortunately, some abortions also lead to severe infections.

Objective: This study aims to qualitatively identify key themes in abortion forums to analyze the issues faced by individuals. The forums were then quantitatively analyzed to determine which problems were most prevalent.

Methods: The most popular abortion forums were identified. A preliminary thematic analysis was conducted for the qualitative phase, followed by a manual quantitative analysis. Two independent researchers reviewed forum posts, categorizing them into specific themes.

Results: An analysis of 13,397 responses from 370 threads across 4 forums revealed “Abortion Process Progression” as the most discussed theme, highlighting a strong demand for information, emotional support, and medical guidance. The highest percentage of threads (162/886, 18.3%) focused on the need for mental support and the desire to have someone in contact during the abortion process. Concerns about the effectiveness of the termination also emerged as a significant topic, accounting for 83 of the 886 (9.4%) discussions. “Emotional and Psychological Aspects” and “Medical and Pharmacological Aspects” were also significant, highlighting the need for holistic care. Anxiety and fear related to the process were frequently discussed, accounting for 60 of the 886 (6.8%) responses. The issue of isolation and lack of support was also a common topic, appearing in 30 of the 886 (3.4%) threads. Notably, “Social and Ethical Perception” accounted for only 13 of the 886 (1.5%) responses and appeared in just 13 threads.

Conclusions: This study highlights the critical need for information and support for women navigating abortion, particularly in regions where access is restricted. It emphasizes the importance of addressing the multifaceted challenges women face and calls for policy changes and strengthened support networks to improve the health and rights of women, as well as all those seeking gynecological care in abortion contexts. Further research is encouraged to refine and expand support strategies.

(J Med Internet Res 2024;26:e59544) doi:10.2196/59544

KEYWORDS

abortion; content analysis; reproductive health; online forum; Poland; women; support

XSL•FO
RenderX

Introduction

Numerous countries, including Poland, are changing their laws to tighten access to abortion [1]. International discourse increasingly suggests that abortion should be viewed not just as a health issue but fundamentally as a human right [2]. Notably, the 1994 Conference on Population and Development in Cairo established a cohesive approach to population policies worldwide and recognized the rights of women to access abortion and contraception. The Cairo Action Plan was designed to allow its principles to be adapted to the national laws of each country, ensuring that the diverse religious, cultural, and ethical contexts of different nations were respected and upheld [3].

Limited access to abortion, or its absence, poses a significant challenge to public health [4]. Abortion is among the primary causes of maternal mortality. A study conducted across 115 countries between 2003 and 2009 documented that 7.9% of maternal deaths were attributable to abortion [5]. The actual number of deaths resulting from abortion could be higher due to potential underreporting [6]. Paradoxically, legalizing abortion can help reduce its incidence, as women receive improved postabortion care and access to contraception consultations [7].

The abortion law in Poland has experienced significant fluctuations over the decades, reflecting changes in the country's political landscape [1,8]. Since 1997, the so-called abortion compromise—a middle ground between those advocating for liberalized abortion laws and those supporting a total ban—has allowed the procedure under 3 specific circumstances [9]. However, recent years have seen mounting pressure from antiabortion movements and influential segments of the Catholic Church [10], which holds significant sway in Polish politics. In 2020, the ruling party, Law and Justice (PiS), known for its staunch antiabortion stance, sought to introduce even stricter regulations [10,11]. That year, the Constitutional Court ruled that irreversible damage threatening the life of the fetus is not a prerequisite for terminating pregnancies under the Constitution, significantly restricting abortion rights in the country.

Since then, in Poland, pregnancy termination is legally permitted only in 2 scenarios: (1) when the pregnancy poses a threat to the life or health of the pregnant woman, or (2) when there is a reasonable suspicion that the pregnancy resulted from a prohibited act, such as rape [12,13].

An additional challenge is the right of conscientious objection, which allows any medical doctor to refuse to perform an abortion [14]. This has resulted in situations highlighted by the Commissioner for Human Rights in Poland, showing that in some regions, up to 1.1 million women lack access to abortion under any circumstances [15,16]. Consequently, women in Poland must seek alternative solutions, regardless of legal provisions [1,17].

Currently, there is no legal, official way to obtain an abortion in Poland outside of the specified indications. However, in the first trimester, a pharmacological termination of pregnancy is possible, with medications provided by international third-sector organizations [18]. This method, using Misoprostol, is one of the approaches recommended by the World Health Organization (WHO) [19].

Against the backdrop of evolving regulations in Poland, a country with one of the most restrictive abortion laws in the European Union [20], the authors have identified a gap in the literature regarding current studies on the situation of women who seek abortions outside official or legal facilities. The primary objective of this study is to qualitatively identify the challenges faced by women in Poland who wish to terminate their pregnancies. An attempt will also be made to approach these findings from a quantitative perspective to identify the most pressing public health challenges. In conclusion, these issues will be discussed in the context of current guidelines.

Methods

Study Overview

This observational study aimed to collect and analyze comments from online forums discussing abortion. Comments were gathered by reviewing various forums based on predefined criteria.

Data Collection

Data collection and analysis were conducted from November 11, 2023, to December 13, 2023. The data were collected from the first page of Google (Alphabet Inc.) search results for the term "forum aborcja" (abortion forum). All forums on the first page containing discussions on abortion were included in the study. The data set covered posts from January 2023 to June 2023. Forum comments were categorized based on specific themes and subjects related to abortion discourse. To minimize bias and ensure reliability, 2 authors (KBK and AK) independently analyzed each comment. Any discrepancies were resolved through discussion. The analysis was supported by a spreadsheet documenting the source of each comment, its thematic classification, and relevant criteria based on the study's categorization framework.

Analysis Criteria

A thematic analysis was conducted in July 2023 on forum threads from May and June 2023, following guidelines for thematic analysis [21]. The forum responses were analyzed using qualitative content analysis, based on the framework by Georgsson et al [22] and supplemented by the study of Aiken et al [23]. The manually collected responses were initially reviewed by the author KBK and refined by removing criteria that did not align with the research objectives. The criteria from the referenced studies were expanded to include additional factors based on themes emerging from the forum responses. The process of coding, data collection, and preparation is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. This flowchart illustrates the study's methodology, starting with identifying the largest abortion forum via Google search, collecting posts from May to June 2023 (82 threads, 1,877 posts), followed by coding, reanalysis, and quantitative analysis of data from all forums (370 threads, 13,397 posts) collected from January to June 2023.

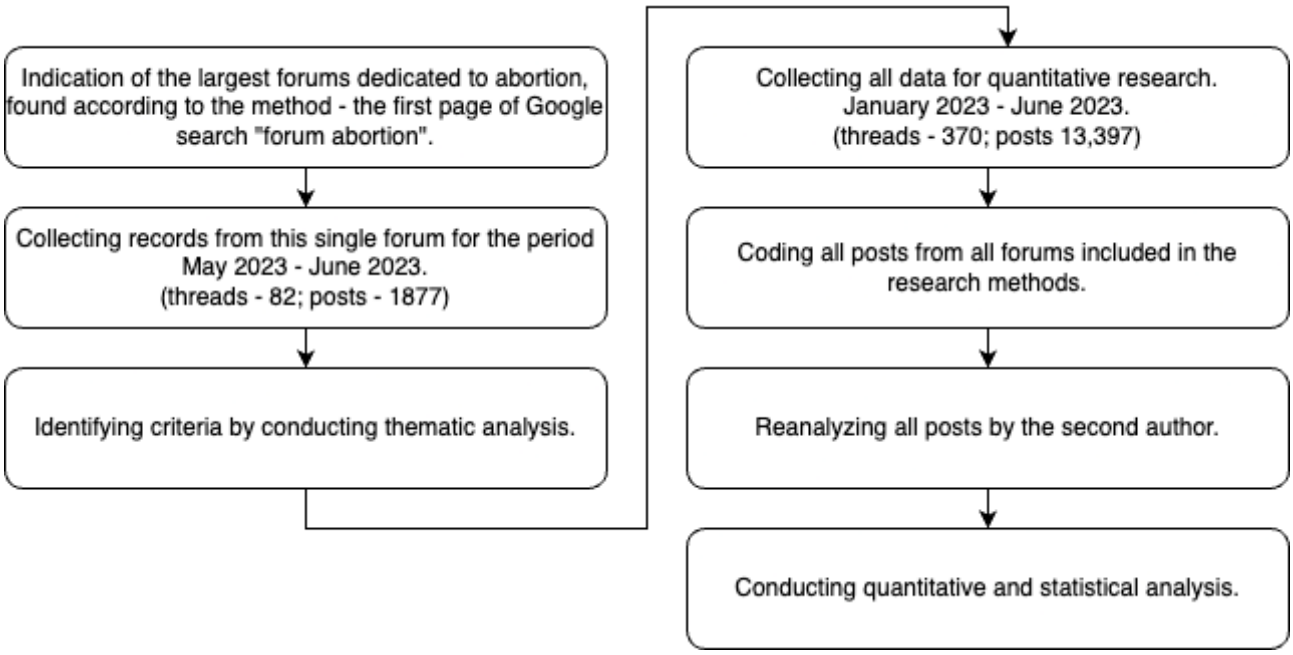

Data Analysis

The data for this part of the analysis were collected from January to June 2023 and analyzed using a spreadsheet. The paper presents the quantitative occurrences of various themes. As multiple subcategories could appear within a single forum thread, the quantitative occurrences of entire categories were not aggregated. An analysis of the interdependence between subcategories was not conducted due to insufficient justification for such an approach.

Statistical Analysis

A chi-square test was conducted to assess the relationship between the day of the week and the themes of forum threads. Additionally, another chi-square test was used to examine which themes cooccurred within threads. A significance level of .05 was applied, and the analysis was performed using R version 4.4.0 and RStudio version 2023.06.0 (R Foundation).

Ethical Considerations

This observational study involved collecting and analyzing publicly available comments from online abortion forums, focusing on public behavior without direct interaction with participants. The data used were either anonymous or deidentified to ensure that no personally identifiable information could be traced to individuals. As all forum content was publicly accessible, formal ethical approval was not required. However, the authors adhered to ethical guidelines for internet research [24], ensuring user privacy by not disclosing any information that could identify individuals.

Results

Overview

A total of 13,397 responses from 370 threads across 4 forums were analyzed. A Google search (Google LLC/Alphabet Inc.) for the term “forum aborcja” (abortion forum) yielded 7 results, of which only 4 were actual forums. One link redirected to one of the mentioned forums, while the other 2 were articles about abortion. The search results and the number of threads collected are presented in Table 1.

XSL•FO
RenderX

Table 1. Categorization of the top search results from Google when querying "forum aborcja" at the beginning of the study[a].

Source	Type	Number of forum threads analyzed
Aborcyjny Dream Team [25]	Redirects to Kobiety w Sieci [26]	0
Wolna Aborcja [27]	Forum	1
WP Parenting [28]	Forum	36
Kobiety w Sieci [26]	Forum	6
ABORCJA [29]	Forum	327
WP abcZdrowie [30]	Article	0
Aborcja Forum [31]	Prolife website with content discouraging the termination of pregnancy	0

[a]The entries are listed in the order they appeared in the search results. Each source is classified as a forum, article, or other type, and the number of forum threads analyzed from each source is indicated. Entries with no analyzed threads were excluded from further study.

Analyzed Themes

Categorization and Analysis of Forum Threads

The collected and analyzed threads were categorized according to established criteria, which are detailed in Table 2. Each criterion was further divided into subcriteria to illustrate the prevalent themes. Table 2 shows the number of forum threads in which each theme appeared. Although a single thread could contain multiple themes, each specific theme within a given thread was counted only once, even if mentioned in multiple posts.

XSL•FO
RenderX

Table 2. Criteria for analyzing themes and their frequency in forum threads (N=886)[a].

Themes and subthemes	Threads in which the theme appeared, n (%)
Abortion Process Progression	
Questions and Description of Termination Process	162 (18.3)
Uncertainty About Procedure Effectiveness	83 (9.4)
Method of Medication Administration and Dosage	75 (8.5)
Alarming Symptoms and Complications	61 (6.9)
Failed Pharmacological Abortion	16 (1.8)
Emotional and Psychological Aspects	
Anxiety Related to the Process	60 (6.8)
Lack of Support and Isolation	30 (3.4)
Impact on Mental Health and Well-Being	19 (2.1)
Medical and Pharmacological Aspects	
Pharmacological Interactions (eg, medications taken, alcohol, tobacco)	28 (3.2)
Impact on Health With Preexisting Conditions	18 (2.0)
Postabortion Contraception Options	18 (2.0)
Long-Term Side Effects	17 (1.9)
Logistical Issues and Specific Situations	
Long Wait for Medication	54 (6.1)
Challenges in Caring for Existing Children	33 (3.7)
Hiding Procedure From Family Or Partner	30 (3.4)
Uncertainty About Pregnancy Presence or Gestational Week	29 (3.3)
Alternative Methods of Termination	18 (2.0)
Logistical and Financial Barriers	13 (1.5)
Concerns About Postprocedure Activities (eg, returning to work)	10 (1.1)
Issues Related to Assistance From Medical Staff	
Doctor's Ability to Identify an Abortion	21 (2.4)
Concerns About Medical Procedures (eg, uterine curettage)	15 (1.7)
Recording Pregnancy in the Medical System	14 (1.6)
Fear of Medical Staff's Reaction	9 (1.0)
Legal and Ethical Issues	
Online Scams and False Offers	19 (2.1)
Social and Ethical Perception	13 (1.5)
Legal Consequences of Abortion	11 (1.2)
Legislation and Restrictions	10 (1.1)

[a]Each theme is broken down into subcriteria to indicate how often it appeared across all threads. Although a single thread could include multiple themes, each theme was counted only once per thread, even if mentioned in several posts.

Abortion Process Progression

Overview

The findings in the "Abortion Process Progression" category provide valuable insights into forum discussions about abortion. They highlight a strong need for information and support on various aspects of the process, including practical advice on medication use and emotional assistance for complications and uncertainties.

Questions and Description of Termination Process

This was the most discussed theme, appearing in 162 of the 886 (18.3%) threads, indicating a significant demand for information and clarity about the abortion process. Discussions likely included questions about the procedure itself, what to expect, and personal experiences shared by those who had undergone it. This criterion included threads where people provided real-time updates on what was happening, as well as threads where individuals shared detailed personal experiences.

XSL•FO
RenderX

Hey, maybe it's because women are somehow embarrassed to write about what happens to their bodies during such a procedure. For me, nothing human is alien, so after the first dose, I had mega chills, diarrhea, and I felt nauseous. After the second dose, my stomach hurt a lot, and after the third dose, the chills and diarrhea stopped, but I started having terrible cramps :/ and such strong stomach pain that I had to use painkillers, otherwise, I don't know if I would have managed :/ I expelled a clot about a third of the size of my palm with a membrane...so I suspect that was it. After that, the pain decreased, and today my bleeding is like during a normal period. I still pass small clots, so as I understand, the uterus is cleaning itself, but besides that, I feel okay. I have an ultrasound scheduled, I will go and check if everything is okay, but yes, I feel that it went successfully.

Uncertainty About Procedure Effectiveness

Doubts regarding the effectiveness of the abortion procedure were discussed in 83 of the 886 (9.4%) threads. These discussions likely included concerns about confirming the procedure's success, such as monitoring human chorionic gonadotropin (hCG) levels, highlighting anxiety, and the need for reassurance about the outcomes.

Hey girls, listen, I started the treatment on Friday. After the first dose, I had a slight stomach ache, after the second one, diarrhea and stomach pain, and after the third, there was spotting and 2 or 3 small clots, about 2 cm, came out. The next day there was only light brown spotting. This is the 5th week. I'll be honest, I don't know if it was successful, I thought maybe something would develop more after 3 days, but nothing. I'll also add that I didn't have the typical bleeding that the girls here describe, I went through it very mildly. Please support me.

Method of Medication Administration and Dosage

This theme appeared in 75 of the 886 (8.5%) threads, indicating notable concern about the practical aspects of medication-based abortions. Discussions likely focused on the correct use of abortion pills, dosage instructions, and handling issues such as a crushed tablet, highlighting the need for accurate medical guidance. Questions included methods of drug administration and dosing details.

Has anyone taken the pills vaginally? Supposedly, they are more effective this way?

Hello girls, today I took the first pill and I have a question: if I took it at 15:00, can I start two hours earlier tomorrow?

As well as more specific situations:

Girls, in the package, a few pills were open, and this one big pill was crushed. What now? Will the big one still work even though it's crushed? Help please.

Alarming Symptoms and Complications

This criterion, appearing in 61 of the 886 (6.9%) forum discussions, was identified when users reported serious side effects of the medications and the abortion procedure, both during and after, extending up to the next menstrual cycle. It included descriptions of intense or persistent bleeding and pain, as well as excessive diarrhea, vomiting, and other atypical symptoms that could pose health risks.

Girls, are you also bleeding so heavily? I've been bleeding non-stop since Monday, and my stomach hurts, is this normal?

I once had an allergic reaction to the medication; if such a reaction starts affecting your décolletage, neck, feet, etc., this is serious, and I advise going to the emergency room. [...] If it gets worse, what to tell the doctor? Nothing...just show your skin, that's enough.

I can't deal with the pain; it's terrible. I don't know if natural childbirth hurt as much as it does now, I'm writhing in pain on the bed, one cramp after another like crazy. There's no bleeding. The diarrhea was severe but now it's somewhat okay. The last dose is at 3 and I can barely function...It's just terrible

Failed Pharmacological Abortion

The least discussed but still notable, this theme appeared in 16 of the 886 (1.8%) threads. It indicates that while less common, there are cases where pharmacological abortions fail, prompting discussions on how to identify failure and what actions to take next.

Girls, unfortunately, it didn't work Yesterday I ordered another set, and I'm starting again next week, it's a nightmare.

Emotional and Psychological Aspects

Overview

Findings within the "Emotional and Psychological Aspects" category reveal that the emotional and psychological components of abortion are a crucial part of the experience, highlighting a clear need for support systems that address these issues. The data emphasize the importance of providing comprehensive care that includes psychological support for individuals undergoing an abortion.

Anxiety Related to the Process

Anxiety was a notable concern, with 60 of the 886 (6.8%) threads reflecting apprehension about the process. Many users expressed uncertainty about what to expect, the severity of pain, and the duration of the procedure. This anxiety could stem from the fear of the unknown, worries about potential complications, or stress related to the decision itself.

Girls, I'm really scared. I'm in the 14th week. It turns out that my son has defects. I can't afford to travel abroad, and in Poland, you know - it may or may not work. That's why I want to try with the pills. I'm just wondering whether A) it's too late B) maybe I should order two sets right away. Did any of you have a

XSL•FO
RenderX

situation where the child survived the treatment? Or did anyone do it this late and it was successful?

Lack of Support and Isolation

The issue of feeling unsupported and isolated was present in 30 of the 886 (3.4%) discussions, highlighting the emotional difficulty some individuals face when lacking a supportive network during the abortion process. This can exacerbate the emotional distress associated with the experience. In some cases, forum members were the only people aware of the user's abortion and the only source with whom they could discuss their concerns and the process.

oh that's nice, I don't have such a doctor. I'm terribly afraid of this treatment and I'll also be alone at home and I have anemia as well.

Impact on Mental Health and Well-Being

The impact on mental health and well-being, though less frequently discussed in 19 of the 886 (2.1%) threads, was still notable. This subcategory encompassed discussions on emotional responses following an abortion, including feelings of relief or regret, and the overall psychological impact of the procedure.

I guess the strongest thing I feel is emptiness and the awareness of loss. I need to somehow sort this out in my head. To go home, tell my 10-year-old, who is waiting for me, that the baby is no more. Then, to return to work, where everyone knew about the pregnancy.

I wonder if it would be a good idea to ask one of my colleagues to inform the rest about my situation, so that no one would ask me how I and the child are feeling.

Medical and Pharmacological Aspects

Overview

The results in the "Medical and Pharmacological Aspects" category highlight the critical need for comprehensive medical information and guidance for individuals considering or undergoing an abortion. This includes managing long-term health, understanding drug interactions, evaluating preexisting conditions, and planning for postabortion contraception.

Pharmacological Interactions (eg, Medications Taken, Alcohol, Tobacco)

This theme, accounting for 28 of the 886 (3.2%) threads, highlights concerns about potential interactions between abortion pills and other medications, as well as alcohol or tobacco use.

The most frequently asked questions in this thread revolved around alcohol and tobacco consumption. Members wanted to know if it was safe to drink alcohol during the procedure, hoping it would help ease their stress. Discussions also included concerns about the use of medications for current and chronic conditions, with members expressing fears about potential drug interactions and whether abortion medications would be effective when combined with other substances.

Hey, does anyone know if taking antibiotics will interfere with the abortion procedure? I was prescribed antibiotics by the dentist today for a tooth infection, and I don't know if I can take them...

And tell me, darling, when can I smoke? I'm so nervous that a cigarette could save me.

Impact on Health With Preexisting Conditions

Representing 18 of the 886 (2.0%) threads, this topic highlights concerns about how undergoing an abortion might impact individuals with preexisting health conditions. These discussions often included worries about potential complications during the procedure or the possibility that it could worsen existing health issues.

For me, it was a nightmare!:/ I was completely exhausted. The bleeding lasted very long, plus I had a tense stomach :/ and I only thought about being active again after 2 months. after a month???

Postabortion Contraception Options

Making up 18 of the 886 (2.0%) threads, this subcategory indicates that participants were seeking information and advice on contraception following an abortion. Forum members questioned whether hormonal contraception could be used immediately after an abortion and were curious about the contraindications of various contraceptive methods post-abortion.

Hey, I'm looking for a clinic/doctor in Germany who will perform a tubal ligation/sterilization procedure. I live near the border (Lubuskie). Unfortunately, I don't speak the language to call around and ask.

Long-Term Side Effects

This subcategory, comprising 17 of the 886 (1.9%) discussions, involved conversations about potential long-term side effects following the abortion procedure. Some participants expressed concerns about future fertility, fearing that they might have difficulty getting pregnant after an abortion. Additionally, there were questions about whether any tests or treatments were necessary after undergoing the procedure.

Hello everyone, I was reading on IG lekarkiproabo about serologic conflict, precisely, "Many people plan to have children in the future after an abortion, so it's important to know your blood type and, if necessary, see a doctor for prophylaxis." I hadn't heard about this before, and I've already had three abortions. I don't know my blood type; I'm going to get it tested tomorrow. I want to have children in the future. Presumably, if it turns out I am Rh-, does the fact that I never received/didn't go for a prophylactic injection after the abortions mean I might have problems in future pregnancies? I read that it can cause miscarriages and generally not good things...:(Request for clarification.

XSL•FO
RenderX

Logistical Issues and Specific Situations

Overview

The results in the "Logistical Issues and Specific Situations" category highlight the diverse and complex logistical challenges faced by individuals considering or undergoing an abortion. This underscores the need for accessible, supportive, and comprehensive abortion care services.

> *I can't go out to get a test because I have sick children at home and no one to leave them with. And I don't want to ask anyone to buy it for me because I don't want anyone to know.*

Long Wait for Medication

A notable concern, discussed in 54 of the 886 (6.1%) threads, was the prolonged waiting period to receive abortion medication. This delay heightened stress and anxiety for forum members, who were worried that their ordered medication might arrive after the 9th week of pregnancy, when pharmacological abortion would no longer be an option. Consequently, some members ordered multiple sets from different sources simultaneously, and others even considered buying pills from strangers on the forum, an action that is illegal in Poland.

> *Hey. I ordered a set two days ago. I'm currently 5 weeks along. The order was confirmed yesterday, but the package has not yet been dispatched. My mind is in turmoil. [...] I'm from a small town in Podlasie, and if this got out, I would be lynched. I just don't know what to do with myself. It's nice that you are here, and I really count on your support :)*

Challenges in Caring for Existing Children

Accounting for 33 of the 886 (3.7%) discussions, this theme highlights the difficulties faced by individuals with children who are considering or undergoing an abortion. Balancing childcare responsibilities during this process poses unique challenges. Many mothers shared that they planned to have an abortion when their children were in kindergarten, school, or during the night while they were asleep. Forum members expressed concerns about their ability to care for their children during the procedure. Additionally, some individuals worried about their capacity to support another child financially and emotionally, which influenced their decision to undergo an abortion.

Hiding Procedure From Family or Partner

This subcategory, present in 30 of the 886 (3.4%) threads, highlights that many individuals felt the need to hide their abortion from family members or partners, reflecting the social stigma and personal conflict associated with the decision. Reasons for concealing the procedure varied across the threads. For example, some individuals were in abusive relationships and feared for the future of their children, while others faced differing opinions about having children within their relationships. When it came to hiding the abortion from family members (eg, mom and dad), religious beliefs and faith were often cited as the primary reasons for maintaining secrecy.

> *Hey girls. I've decided to have an abortion. I'm at the beginning of the 5th week. I told my husband initially that I wanted an abortion because we are in a difficult situation; we live separately, he's abroad, and this won't change any time soon. But now I'm wondering whether to tell him that I'm going to do it, or just blame it on nature, I don't know if he'll be able to handle it psychologically. I'm just waiting for the pills. I live with my parents and have to keep all of this a secret.*

Uncertainty About Pregnancy Presence or Gestational Week

This subcategory, making up 29 of the 886 (3.3%) discussions, highlights the uncertainty that some forum participants had regarding their pregnancy status or the specific week of their pregnancy. Forum members frequently asked for help interpreting beta-hCG test results. This uncertainty can notably influence the decision-making process and the choice of the abortion method.

> *Hey girls. I'm totally new here, reading your threads, and it's easing my stress a bit, but I'm still loaded with emotions...[...] The worst part is that I can't remember when I last had my period. I stopped taking birth control pills about 3 or 4 weeks ago, that's when I had my period, but unfortunately, I didn't note the exact date...[...] On Friday, I did two pregnancy tests, they came out positive. For me, it was a shock because I totally didn't expect it, I had my period just a short while ago...I have no idea when exactly the fertilization could have happened, so I don't even know how to calculate it :(Today I went for a beta test, tomorrow I'm picking up the results to be 100% sure and to figure out approximately which week it will be.*

Alternative Methods of Termination

Present in 18 of the 886 (2.0%) discussions, this topic reflects the exploration or pursuit of alternative methods for abortion, possibly due to the inaccessibility or unavailability of preferred options. Forum members inquired about what the procedure would be like in abortion clinics abroad, its cost, the quality of care, and whether they could communicate effectively with doctors from another country.

> *Girls, I have an issue. I've been through two unsuccessful treatments. Now I'm looking for a reliable clinic. Has any of you had any experience with clinics abroad?? If so, please help me with some verified contact information. There's a lot on the internet but you never know where you might end up.*

Logistical and Financial Barriers

Making up 13 of the 886 (1.5%) discussions, this theme highlights the logistical and financial obstacles encountered in accessing abortion services. These barriers can include travel costs, medical fees, and the complexity of arranging the procedure. Additionally, people living in rural areas expressed concerns that their medication parcels might not arrive on time, and some were worried that if complications arose, they would be unable to reach a hospital for professional help.

XSL•FO
RenderX

Girls, I'm really scared. I'm in the 14th week. It turns out that my son has defects. I can't afford to travel abroad, and in Poland, you know - it may or may not work. That's why I want to try with the pills. I'm just wondering whether A) it's too late B) maybe I should order two sets right away. Did any of you have a situation where the child survived the treatment? Or did anyone do it this late and it was successful?

Concerns About Postprocedure Activities (eg, Returning to Work)

Making up 10 of the 886 (1.1%) threads, concerns about postprocedure activities were prevalent. Forum members were uncertain about the recovery period and the duration of bleeding, which could impact their return to work. Additionally, there were questions about how soon after the abortion it would be safe to have sexual intercourse with a partner.

How long after taking the pills can I return to full physical activity?? I mean doing everything like work, sports, etc? I'm a professional swimmer and I wouldn't want this to negatively affect my body.

Issues Related to Assistance From Medical Staff

Overview

The results under the category "Issues Related to Assistance From Medical Staff" reveal the challenges forum participants faced when interacting with medical professionals during the abortion process. These findings highlight the complexities and anxieties surrounding medical staff interactions and underscore the need for sensitive, nonjudgmental, and supportive care for individuals undergoing an abortion.

Doctor's Ability to Identify an Abortion

Accounting for 21 of the 886 (2.4%) threads, this subcategory reflects concerns about whether doctors can differentiate between a spontaneous miscarriage and an induced abortion. Additionally, threads included questions about when during a medical abortion participants should go to the hospital, believing the pregnancy could no longer be saved. The topic also covered whether gynecologists could identify that a miscarriage was induced and if doctors asked questions to confirm this.

I have a question, girls, If I go through with the procedure on Monday, and I have a gynecologist appointment on Wednesday, is there a possibility that the gynecologist will realize that it wasn't a spontaneous miscarriage?

Concerns About Medical Procedures (eg, Uterine Curettage)

Making up 15 of the 886 (1.7%) threads, this theme highlights fears surrounding the medical procedures associated with abortion, such as uterine curettage. Participants expressed anxiety about whether they would be given an injection and sent home immediately or if they would need to stay for further tests and procedures. There was particular concern about the standard practice of curettage in their region. Additionally, forum members questioned the necessity of using curettage. Many expressed the belief that curettage was unnecessary, as they thought the uterus would naturally "clean itself" after an abortion.

Will they give me the injection right away and send me home? Or will they want to keep me for more tests? An ultrasound? That dreaded curettage, which, at least in my region, seems to be the standard...That's what I'm most afraid of.

Recording Pregnancy in the Medical System

Representing 14 of the 886 (1.6%) discussions, this topic highlights concerns regarding the recording of medical details, such as beta-hCG test results, in medical registries. Participants expressed worries about the privacy and confidentiality of their medical information related to pregnancy and abortion. Many were anxious that, after a positive beta-hCG result, a pregnancy card would be added to their medical records, and if they underwent an induced miscarriage, they would need to explain the circumstances of the miscarriage.

Girls, I have a question, do you know if beta test results are entered into a registry?

Fear of Medical Staff's Reaction

This subcategory, making up 19 of the 886 (1.0%) discussions, reflects apprehension about how medical staff might react upon learning about an abortion or pregnancy. Forum members shared concerns and negative experiences involving judgmental attitudes from health care providers. For instance, some participants recounted instances where doctors made critical or biased remarks, such as reprimanding them for being pregnant at a young age or assuming they would be single mothers, which added to their anxiety and feelings of discomfort during the medical process.

I'll describe how the entire visit went. I arrived and said that I mainly wanted an ultrasound because I suspected I was pregnant. From the beginning, he was unfriendly to me, making comments like "at such an age and already pregnant" (I'm 20 years old this year). When I said I was in the 14th week, I burst into tears out of nerves because I thought it was too late for an abortion, but I knew I couldn't mention anything about it. After the examination, he started asking if I had a partner; I briefly answered that he's abroad. He then started saying that in that case, he must be a bad partner, that I am therefore a single mother, and that I need to think about my actions. I didn't respond at all, just tears were streaming down my face from stress. He said that there's a place for people like me and then he wrote down the number of this home for single mothers XD. He made up the whole narrative while being a complete jerk, while I didn't say a word about whether I want or don't want the child. He asked me if I wanted him to manage my pregnancy; I said I didn't know because I'm in shock and need to think about it. After that answer, he snorted, Well well, it's your decision. That's all, goodbye.

XSL•FO
RenderX

Legal and Ethical Issues

Overview

The findings in the "Legal and Ethical Issues" category highlight the multifaceted concerns and experiences shared by forum participants, illustrating the challenges related to the legal, ethical, and social aspects of abortion. The discussions revealed apprehensions about potential legal repercussions, the fear of being reported, and the ethical dilemmas that come with seeking abortions outside of official channels. The conversations emphasized the importance of creating supportive and secure spaces for individuals to discuss their situations without fear of judgment or legal consequences.

Online Scams and False Offers

Representing 19 of the 886 (2.1%) discussions, this theme involves experiences of being scammed while attempting to obtain abortion pills online. Users in this thread also debated whether the offers found on the internet were fraudulent posts made by the Polish police or by individuals associated with the government. One user shared:

> *I came here on a friend's recommendation, just finished the treatment, and before that, I was scammed twice! I naively transferred the money and saw nothing :/ I can't understand how someone can behave like this towards other people!*

Social and Ethical Perception

Making up 13 of the 886 (1.5%) threads, this subcategory reflects the social and ethical attitudes and emotions surrounding abortion. Some of the most fearful users were those from small communities, concerned that their neighbors would discover their abortion and ostracize them. These threads also included users grappling with guilt following their abortion. One user asked:

> *Hey, I'd like you to share how you felt psychologically a few months...a few years after the abortion? Did any memories, guilt come back?*

Another harsh comment read:

> *All of you undergoing abortions should be locked up in a cage. Trash.*

Legal Consequences of Abortion

Accounting for 11 of the 886 (1.2%) discussions, this topic covers fears and questions regarding the legal implications of purchasing abortion pills in Poland. There were numerous inquiries and concerns about the current abortion laws and the potential consequences for those who have undergone an abortion. Some users also questioned the exact definition of "assistance in having an abortion," which is illegal in Poland. A forum participant expressed:

> *Many people warn me against buying pills in PL. They scare me with the police. How is it? I'm desperate because I need the pills to cleanse myself.. I need them fast. What are your experiences? Aren't you afraid to order?*

Legislation and Restrictions

This subcategory, making up 10 of the 886 (1.1%) discussions, involves concerns about the legal aspects and restrictions related to abortion. Although similar to the previous thread, this one included posts discussing direct contact with the justice system, such as being summoned to the police station for questioning. For example, one participant shared:

> *Hi girls, I haven't been on the forum for a long time. Today, I've been a mother for 4 months, I have a son and I'm very happy although I won't hide that I used to think that it wasn't for me. To the point, did any of you recently receive a summons to the police as a witness? The articles cited in the summons are: 152 § 2 of the Penal Code - punishable assistance in abortion; 127a sec. 2 of the Pharmaceutical Law - mediating in the trade of medicinal products without the required entry or permission; 270 § 1 of the Penal Code - forging documents. It surely isn't about you, I bought pills from wow, but before I learned about women on the internet, I once ordered pills in 2019 from the site 9tygodni.pl (through which I was later blackmailed to tell my family). Did any of you have such an interrogation? What to expect?*

Statistical Analysis

Day of the Week and Themes

There was no significant relationship between the day of the week and the themes discussed in the threads (chi-square test, P=.22). This indicates that the occurrence of themes was independent of the day of the week.

Occurrence of Threads

The theme of Abortion Process Progression was more frequently discussed in conjunction with Emotional and Psychological Aspects and Legal and Ethical Issues compared with other themes. Logistical Issues and Specific Situations were more commonly associated with Emotional and Psychological Aspects and Issues Related to Assistance From Medical Staff. Lastly, Legal and Ethical Issues were frequently discussed alongside Issues Related to Assistance From Medical Staff. The P values from the chi-square test are presented in Table 3.

XSL•FO
RenderX

Table 3. *P* values of the chi-square test to determine which themes occurred together in threads.

Themes	Abortion Process Progression	Emotional and Psychological Aspects	Medical and Pharmacological Aspects	Logistical Issues and Specific Situations	Issues Related to Assistance From Medical Staff	Legal and Ethical Issues
Abortion Process Progression	x[a]	.008	.23	.33	.42	<.001
Emotional and Psychological Aspects	.008	x	.56	<.001	.13	.29
Medical and Pharmacological Aspects	.23	.56	x	.82	.11	.33
Logistical Issues and Specific Situations	.33	<.001	.82	x	.01	.32
Issues Related to Assistance From Medical Staff	.42	.13	.11	.01	x	.001
Legal and Ethical Issues	.001	.29	.33	.32	.001	x

[a]This value was entered when the same category appears in both the row and column.

Discussion

Principal Findings

In response to the 2020 restrictions, which effectively implemented a near-total ban on abortion across the country, women seeking abortion care in Poland often rely on community support from specific organizations [12]. Additionally, they may organize trips abroad, purchase medications on the black market, or import them from other countries. However, there is a lack of an official, fact-based source of information developed by specialists. All necessary information must be obtained from forums, chats, and internet pages, some of which have been analyzed in this work [9,18].

Following the introduction of near-total abortion restrictions, the number of legal abortions in Poland decreased significantly. According to data from the Polish Ministry of Health, there were 1110 legal abortions in Poland in 2019 and 1076 in 2020. This number dropped 10-fold to 107 legal abortions in 2021 and slightly increased to 161 in 2022 [32]. Referring to data from major aid organizations, it is believed that the reported statistics significantly underestimate the actual number of abortions taking place in the country [33,34].

The Abortion Dream Team, the largest organization in Poland providing abortion assistance and uniting other aid organizations, supported 32,000 individuals in accessing abortion services in 2021 and facilitated travel for 1544 individuals to foreign abortion clinics. Between October 22, 2021, and October 17, 2022, the Abortion Dream Team assisted 44,000 individuals in obtaining abortions [34,35]. The Polish Foundation for Women and Family Planning (FEDERA), in the 2 years following the court ruling that restricted abortion access, provided services to thousands of women. This included handling over 20,000 helpline calls, responding to 12,000 emails, offering 1016 consultations for refugees from Ukraine, providing 1200 legal consultations, intervening in 28 hospitals, and filing 30 complaints to the European Court of Human Rights regarding hospital violations of patients' rights [33].

Our analysis indicates that many individuals have sought abortion support independently through unofficial online forums, suggesting significant underestimation and underreporting of abortion statistics (a total of 13,397 responses from 370 threads across 4 forums were analyzed). This highlights the lack of reliable data and comprehensive discussions about abortion in Poland. Thousands of individuals who have undergone the procedure remain in a gray area, with their experiences unrepresented in official statistics.

This research aims to highlight the perspectives of those with firsthand experience. The objective was to gain insight into the decision-making process of individuals who choose to terminate pregnancies, exploring the broader context of their situation, as well as their concerns and obstacles. The results show that women face a range of challenges related to abortion, from technical and clinical aspects—such as selecting the appropriate medication, timing, and dosage—to the actual process of terminating the pregnancy, including how it should unfold and potential complications. Our analysis parallels the work of Georgsson et al [22], highlighting the urgent need for comprehensive abortion care that includes both procedural information and emotional support. Both studies reveal a significant gap in patient knowledge and preparation, emphasizing the importance of providing accessible, accurate information about abortion procedures. A shared focus on the emotional and psychosocial support required to address concerns about physical pain, procedure effectiveness, and the psychological aftermath further underscores the necessity of a holistic approach to care. Additionally, the prominence of online forums in our study aligns with Georgsson et al [22]'s recognition of digital platforms as essential resources for support, suggesting an integrated model of health care that combines both traditional and digital approaches to meet the diverse needs of individuals undergoing an abortion.

Inquiries about the lack of knowledge regarding the abortion process are the most common, accounting for 162 of the 886 (18.3%) forum discussions. These primarily focus on questions and descriptions of the termination process. Alarmingly, the issues raised today strongly resemble those identified after 1993,

XSL•FO
RenderX

when abortion rights in Poland were severely restricted—particularly the removal of termination options based on the difficult life circumstances of the pregnant woman [1,36-38]. During that period, researchers observed a decline in public awareness about pregnancy termination and an increase in the use of underground abortion practices or seeking terminations abroad, which significantly compromised patient safety. This situation not only reflects a repetition of past mistakes in public health policy but also represents a regression that further exacerbates these issues.

Legally, the situation today mirrors past issues, further complicated by the fears faced by both patients and physicians. The antiabortion law has historically been more restrictive in practice than its statutory language suggests. There is evidence showing that women, despite being entitled to a legal abortion and having all the required documentation, were still denied access to the procedure [1,38]. According to our data, 53 of the 886 (6.0%) forum users are currently facing legal and ethical issues related to abortion. This underscores the persistent challenges and barriers within the legal framework that continue to affect access to abortion services.

Patients express apprehensions about the potential detection of abortions by medical staff and concerns regarding the registration of pregnancies—these issues appeared in 21 of the 886 (2.4%) and 14 of the 886 (1.6%) analyzed texts, respectively. Since October 1, 2022, the mandatory recording of pregnancies in the medical events registry has subjected patients to oppressive measures [39]. The legal framework surrounding the “Pregnancy Registry” exacerbates women’s feelings of anxiety and oppression. There is a palpable fear that the registry information could prompt inquiries into the reasons for not proceeding with childbirth, undermining trust in the doctor-patient relationship and compromising the confidentiality of medical data. This situation has notably disturbed patients, further eroding their sense of privacy and confidence in health care services [40].

Ideally, such a registry could serve as a valuable data source for analyzing maternal health realities, identifying potential threats, and enhancing guidelines and preventive measures, similar to practices applied to certain states or medical conditions [41-43]. However, in the current legal context, achieving these objectives seems unattainable without jeopardizing patients’ sense of security and legal safety.

Polish Public Health is responsible for ensuring abortion safety amid unfavorable legal and systemic conditions, making this study one of the few of its kind in Poland. Currently, there is a noticeable lack of official health data that highlight the challenges faced by individuals during abortions. Through this analysis, we are able to outline the problems, concerns, and aspects of the abortion process experienced by Polish women. As a result, it becomes possible to develop targeted support strategies, systemic solutions, and improve patient safety by providing accurate medical information. However, it is clear that more studies of this nature are urgently needed.

Comprehensive abortion care, recognized by the WHO as part of the List of Essential Health Services, requires strategic initiatives [44]. By conducting appropriate educational campaigns, decriminalizing abortion, and ensuring access to medical assistance, we could systematically address 23 out of the 27 issues identified in our study (Table 1). Information on both pharmacological and surgical abortion procedures should be made accessible to patients and health care professionals [45].

It is crucial to focus on providing information and support to marginalized groups, including low-income women and girls, refugees and migrants, youth, lesbians, bisexual cisgender women, transgender or nonbinary individuals, and members of ethnic and religious minorities [46,47]. This is especially important in light of the ongoing armed conflict in Ukraine, which has driven many women to migrate to Poland in search of reproductive health services, including abortion [48].

Patients should have the right to access abortion services within the Polish health care system, supported by doctors, midwives, or appropriately trained staff, and receive comprehensive pre- and postabortion care. The framework for such access suggests that decriminalization of both those seeking and providing abortions could make a significant contribution. Ensuring access to high-quality, evidence-based information could enhance patient safety in reproductive health, minimize health risks, and uphold fundamental human rights [49-51].

Limitations

This research has several limitations. Primarily, as a content-based study, it faces typical constraints associated with this methodology, such as selection bias and limited generalizability, as it relies on data from individuals active in online communities [52]. This approach may not fully capture the diverse experiences of all women seeking abortions in Poland. Additionally, the study’s time frame (January to June 2023) limits the understanding of the evolving context of abortion access, which may affect the long-term relevance of the findings. The qualitative analysis introduces a level of subjectivity through the researchers’ interpretation. Another limitation is the possibility that forum moderators may have removed or edited posts, potentially affecting the data set’s completeness and the depth of the analysis. A sentiment analysis could have provided additional insights; however, due to the varying formats of the forums from which the data were collected and the differing terms of service, it was not conducted. Finally, systemic barriers within the health care system were not fully explored, highlighting an area for further research.

Conclusions

This study is one of the first to explore the abortion experiences of Polish women in such comprehensive detail, shedding light on issues that often escape the attention of policy makers and health care practitioners. Through an in-depth analysis of online forum discussions, it uncovers the nuanced challenges women face under Poland’s restrictive abortion laws, highlighting the gap between legislative frameworks and the actual needs of women. The research highlights critical issues related to access, information, and emotional support—factors that are pivotal in shaping public health strategies and informing policy decisions. In doing so, it not only enriches the academic discourse on

XSL•FO
RenderX

reproductive health but also signals to policy makers the urgent need for revisions that align with women's rights and well-being.

Moreover, the findings highlight a stark contrast between Poland's approach to abortion and the standards set by the European Union regarding women's health care and rights. The study offers actionable recommendations for public health in Poland, including decriminalizing abortion, launching educational campaigns to combat misinformation, and strengthening support systems for women. These suggestions aim to bridge the gap between the current state of abortion services in Poland and the comprehensive care model advocated by international human rights standards, urging policy makers to guarantee women's access to safe and supportive reproductive health care.

Acknowledgments

The authors, as nonnative English speakers, used a chat-based artificial intelligence (AI) tool to assist with language editing and grammatical improvements of the manuscript. No part of the data collection, data processing, or interpretation of the results was influenced by any generative AI tools. The authors take full responsibility for the content of the manuscript, ensuring that the analysis and conclusions are their own, and that AI was used solely for language improvement.

Data Availability

All data used in this study were publicly available at the time of collection. The forums and discussions analyzed are accessible through public internet sources. The combination and presentation of these data in the manuscript are considered exhaustive and comprehensive. Because of the public nature of the data, no additional data sets have been deposited in a separate repository. Any further details can be accessed directly from the sources mentioned in the manuscript.

Authors' Contributions

Conceptualization was led by KBK, while data collection was jointly carried out by KBK and AK. Manuscript preparation was handled by KBK, with EM contributing to the preparation of the "Discussion" section. Supervision of the project was provided by ŁB.

Conflicts of Interest

None declared.

References

1. Zaręba K, Herman K, Kołb-Sielecka E, Jakiel G. Abortion in countries with restrictive abortion laws-possible directions and solutions from the perspective of Poland. Healthcare (Basel) 2021 Nov 20;9(11):1594 [FREE Full text] [doi: 10.3390/healthcare9111594] [Medline: 34828639]
2. Zampas C, Gher JM. Abortion as a human right—international and regional standards. Human Rights Law Review 2008;8(2):249-294. [doi: 10.1093/hrlr/ngn008]
3. World Health Organization (WHO). Safe Abortion: Technical and Policy Guidance for Health Systems (2nd edition). Geneva, Switzerland: WHO; 2012.
4. Fathalla MF. Safe abortion: the public health rationale. Best Pract Res Clin Obstet Gynaecol 2020 Feb;63:2-12. [doi: 10.1016/j.bpobgyn.2019.03.010] [Medline: 31201007]
5. Say L, Chou D, Gemmill A, Tunçalp Ö, Moller AB, Daniels J, et al. Global causes of maternal death: a WHO systematic analysis. Lancet Glob Health 2014 Jun;2(6):e323-e333 [FREE Full text] [doi: 10.1016/S2214-109X(14)70227-X] [Medline: 25103301]
6. Singh S, Wulf D, Hussain R, Bankole A, Sedgh G. Abortion worldwide: a decade of uneven progress. Guttmacher Institute. New York, NY: Guttmacher Institute; 2009. URL: https://www.guttmacher.org/report/abortion-worldwide-decade-uneven-progress [accessed 2024-12-03]
7. Faúndes A. Strategies for the prevention of unsafe abortion. Int J Gynaecol Obstet 2012 Oct;119 Suppl 1:S68-S71. [doi: 10.1016/j.ijgo.2012.03.021] [Medline: 22883917]
8. Szawarski Z. Abortion in Poland. Br J Obstet Gynaecol 1991 Dec;98(12):1202-1204. [doi: 10.1111/j.1471-0528.1991.tb15389.x] [Medline: 1777450]
9. Bucholc M. Abortion law and human rights in Poland: the closing of the jurisprudential horizon. Hague J Rule Law 2022 Feb 16;14(1):73-99. [doi: 10.1007/s40803-022-00167-9]
10. Szelewa D. Killing 'unborn children'? The Catholic church and abortion law in Poland since 1989. Social & Legal Studies 2016;25(6):741-764. [doi: 10.1177/0964663916668247]
11. Wiśniewska P. Brukselski dyskurs wobec polskiej aborcji. Dyskurs & Dialog 2021;III(1):125-144. [doi: 10.5281/zenodo.5040745]

XSL•FO
RenderX

12. Trybunał Konstytucyjny. Wyrok Trybunału Konstytucyjnego z dnia 22 października 2020 r. sygn. akt K 1/20. SEJM. 2020. URL: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20210000175 [accessed 2024-08-13]
13. Ustawa z dnia 7 stycznia 1993 r. o planowaniu rodziny, ochronie płodu ludzkiego i warunkach dopuszczalności przerywania ciąży. SEJM. 1993. URL: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu19930170078 [accessed 2024-08-13]
14. Ustawa z dnia 5 grudnia 1996 r. o zawodach lekarza i lekarza dentysty. SEJM. URL: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu19970280152 [accessed 2024-08-13]
15. Podkarpacki Ośrodek Badań Regionalnych. Ludność Podkarpacia w 2021 Roku. Urząd Statystyczny w Rzeszowie. 2022. URL: https://rzeszow.stat.gov.pl/dla-mediow/informacje-prasowe/ludnosc-podkarpacia-w-2021-roku,299,1.html [accessed 2024-08-13]
16. Podkarpacie – bez zabiegów przerywania ciąży. Rzecznik pyta o to rzeszowski NFZ. Biuletyn Informacji Publicznej RPO. 2019. URL: https://bip.brpo.gov.pl/pl/content/podkarpacie-%E2%80%93-bez-zabiegow-przerwania-ciazy-rpo-pyta-rzeszowski-nfz [accessed 2024-08-13]
17. Aborcja w domu (tabletkami/farmakologiczna). Aborcyjny Dream Team. URL: https://aborcyjnydreamteam.pl/aborcja-w-domu/ [accessed 2024-08-13]
18. Women help women consultation. Women Help Women. URL: https://consult.womenhelp.org/pl/page/378/in-collection/377/medical-abortion [accessed 2024-08-13]
19. World Health Organization (WHO). Abortion Care Guideline. Geneva, Switzerland: WHO; 2022.
20. Center for Reproductive Rights. European Abortion Laws. A Comparative Overview. New York, NY: Center for Reproductive Rights; 2022.
21. Maguire M, Delahunt B. Doing a thematic analysis: a practical, step-by-step guide for learning and teaching scholars. All Ireland Journal of Higher Education 2017 Oct 30;9(3):3351 [FREE Full text] [doi: 10.62707/aishej.v9i3.335]
22. Georgsson S, Krautmeyer S, Sundqvist E, Carlsson T. Abortion-related worries, fears and preparedness: a Swedish Web-based exploratory and retrospective qualitative study. Eur J Contracept Reprod Health Care 2019 Oct;24(5):380-389. [doi: 10.1080/13625187.2019.1647334] [Medline: 31368829]
23. Aiken ARA, Guthrie KA, Schellekens M, Trussell J, Gomperts R. Barriers to accessing abortion services and perspectives on using mifepristone and misoprostol at home in Great Britain. Contraception 2018 Feb;97(2):177-183 [FREE Full text] [doi: 10.1016/j.contraception.2017.09.003] [Medline: 28941978]
24. Franzke AS, Bechmann A, Ess CM, Zimmer M. Internet Research: Ethical Guidelines 3.0. The International Association of Internet Researchers. 2020. URL: https://aoir.org/reports/ethics3.pdf [accessed 2024-12-03]
25. Aborcyjny Dream Team. URL: https://aborcyjnydreamteam.pl/mialam-aborcje/ [accessed 2024-12-03]
26. Kobiety w Sieci. URL: https://www.maszwybor.net/ [accessed 2024-12-03]
27. Wolna Aborcja. URL: https://wolna-aborcja.com/forum/ [accessed 2024-12-03]
28. WP Parenting. URL: https://forum.parenting.pl/topic/4629938-aborcja/ [accessed 2024-12-03]
29. ABORCJA. URL: https://forum.gazeta.pl/forum/f,20823,ABORCJA.html [accessed 2024-12-03]
30. WP abcZdrowie. URL: https://portal.abczdrowie.pl/poczulam-ulge-gdy-juz-bylo-po-agnieszka-opowiada-o-aborcji-farmakologicznej [accessed 2024-12-03]
31. Aborcja Forum. URL: https://aborcjaforum.pl/ [accessed 2024-12-03]
32. Sejm Rzeczypospolitej Polskiej IX kadencja Prezes Rady Ministrów DSP.WPP.4521.18.2022. Druk Nr 2758 Rady Ministrów z Wykonywania Oraz o Skutkach Stosowania w 2021 r. Ustawy z Dnia 7 Stycznia 1993 r. o Planowaniu Rodziny, Ochronie Płodu Ludzkiego i Warunkach Dopuszczalności Przerywania Ciąży. SEJM. 2018. URL: https://orka.sejm.gov.pl/Druki9ka.nsf/0/75710091486B2822C12588EF0034990C/%24File/2758.pdf [accessed 2024-08-13]
33. Dwa lata po wyroku pseudo-tk. Podsumowanie działań federy – Federa fundacja na rzecz kobiet i planowania rodziny. Federa. URL: https://federa.org.pl/2-lata-po-pseudowyroku/ [accessed 2024-08-13]
34. Aborcja bez Granic: najnowsze dane 2021-2022. Aborcyjny Dream Team. URL: https://adt.pl/news/organizacje-aborcyjne/aborcja-bez-granic-najnowsze-dane-z-2021-2022/ [accessed 2024-03-07]
35. Pacewicz P, Chrzczonowicz M. „To liczba hańby". Aborcja wymazana z polskich statystyk. Ale tylko oficjalnie. OKO.press. 2022. URL: https://oko.press/to-liczba-hanby-aborcja-wymazana-z-polskich-statystyk-ale-tylko-oficjalnie-wykres [accessed 2024-08-13]
36. Nowicka W. The effects of the 1993 anti-abortion law in Poland. Entre Nous Cph Den 1996 Dec(34-35):13-15. [Medline: 12222281]
37. Gajewski M. Abortion laws cause problems in Poland. BMJ 1995 Jun 17;310(6994):1556-1557. [doi: 10.1136/bmj.310.6994.1556a] [Medline: 7787640]
38. David H, Titkow A. Abortion and women's rights in Poland, 1994. Studies in Family Planning 1994 Jul;25(4):239-295. [doi: 10.2307/2137907]
39. Rozporządzenie Ministra Zdrowia z dnia 26 czerwca 2020 r. w sprawie szczegółowego zakresu danych zdarzenia medycznego przetwarzanego w systemie informacji oraz sposobu i terminów przekazywania tych danych do Systemu Informacji Medycznej. SEJM. 2020. URL: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20200001253 [accessed 2024-12-03]
40. Filipowska M. Rejestr ciąż – co oznacza, jak działa? wylecz.to. URL: https://wylecz.to/choroby-w-ciazy/rejestr-ciaz/ [accessed 2024-08-13]

XSL•FO
RenderX

41. Bose CL, Bauserman M, Goldenberg RL, Goudar SS, McClure EM, Pasha O, et al. The Global Network Maternal Newborn Health Registry: a multi-national, community-based registry of pregnancy outcomes. Reprod Health 2015;12 Suppl 2(Suppl 2):S1-95 [FREE Full text] [doi: 10.1186/1742-4755-12-S2-S1] [Medline: 26063166]
42. Cohen L, Church T, Freeman M, Gaccione P, Caplin P, Kobylski L, et al. Reproductive safety of lurasidone and quetiapine: update from the National Pregnancy Registry for Psychiatric Medications. Journal of Women's Health 2023 Apr 01;32(4):452-462. [doi: 10.1089/JWH.2022.0310]
43. Jokubaitis VG, Skibina O, Alroughani R, Altintas A, Butzkueven H, Eichau S, et al. The MSBase pregnancy, neonatal outcomes, and women's health registry. Ther Adv Neurol Disord 2021;14:17562864211009104. [doi: 10.1177/17562864211009104] [Medline: 33912245]
44. World Health Organization (WHO). Maintaining Essential Health Services: Operational Guidance for the COVID-19 Context. In: WHO. Geneva, Switzerland: WHO; 2020.
45. Kapp N, Whyte P, Tang J, Jackson E, Brahmi D. A review of evidence for safe abortion care. Contraception 2013 Sep;88(3):350-363. [doi: 10.1016/j.contraception.2012.10.027] [Medline: 23261233]
46. Espinoza C, Samandari G, Andersen K. Abortion knowledge, attitudes and experiences among adolescent girls: a review of the literature. Sex Reprod Health Matters 2020 Dec;28(1):1744225 [FREE Full text] [doi: 10.1080/26410397.2020.1744225] [Medline: 32284012]
47. Munakampe MN, Zulu JM, Michelo C. Contraception and abortion knowledge, attitudes and practices among adolescents from low and middle-income countries: a systematic review. BMC Health Serv Res 2018 Nov 29;18(1):909 [FREE Full text] [doi: 10.1186/s12913-018-3722-5] [Medline: 30497464]
48. Cioffi A, Cecannecchia C, Cioffi F. Violation of the right to abortion at the time of the war in Ukraine. Sex Reprod Healthc 2022 Sep;33:100738. [doi: 10.1016/j.srhc.2022.100738] [Medline: 35640526]
49. Erdman JN. Harm reduction, human rights, and access to information on safer abortion. Int J Gynaecol Obstet 2012 Jul;118(1):83-86. [doi: 10.1016/j.ijgo.2012.04.002] [Medline: 22608022]
50. Lim LM, Singh K. Termination of pregnancy and unsafe abortion. Best Pract Res Clin Obstet Gynaecol 2014 Aug;28(6):859-869. [doi: 10.1016/j.bpobgyn.2014.05.005] [Medline: 24962349]
51. Gambir K, Kim C, Necastro K, Ganatra B, Ngo T. Self-administered versus provider-administered medical abortion. Cochrane Database Syst Rev 2020 Mar 09;3(3):CD013181 [FREE Full text] [doi: 10.1002/14651858.CD013181.pub2] [Medline: 32150279]
52. Sinha A, Porter T, Wilson A. The use of online health forums by patients with chronic cough: qualitative study. J Med Internet Res 2018 Jan 24;20(1):e19 [FREE Full text] [doi: 10.2196/jmir.7975] [Medline: 29367181]

Abbreviations

AI: artificial intelligence
FEDERA: Polish Foundation for Women and Family Planning
hCG: human chorionic gonadotropin
WHO: World Health Organization

Edited by A Mavragani; submitted 15.04.24; peer-reviewed by E Jaafar, NJ Haque; comments to author 11.05.24; revised version received 03.06.24; accepted 28.10.24; published 09.12.24.

Please cite as:
Klimiuk KB, Kot A, Majcherek E, Kubiak KB, Balwicki Ł
Themes in Abortion Forum Discussions in a Restrictive Access Context: Qualitative and Quantitative Analyses
J Med Internet Res 2024;26:e59544
URL: https://www.jmir.org/2024/1/e59544
doi:10.2196/59544
PMID:

©Krzysztof Bartosz Klimiuk, Amelia Kot, Ewa Majcherek, Katarzyna B Kubiak, Łukasz Balwicki. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ecological Momentary Assessment of Mental Health Problems Among University Students: Data Quality Evaluation Study

Ana Portillo-Van Diest[1,2,3], MPH; Philippe Mortier[1,2], MD, PhD; Laura Ballester[1,2], PhD; Franco Amigo[1,2], MPH; Paula Carrasco[2,4,5], PhD; Raquel Falcó[6], PhD; Margalida Gili[7,8], MD, PhD; Glenn Kiekens[9,10,11], PhD; Francisco H Machancoses[4], PhD; Jose A Piqueras[12], PhD; Marisa Rebagliato[2,4,5], MPH, MD, PhD; Miquel Roca[7,8], MD, PhD; Tíscar Rodríguez-Jiménez[13], PhD; Jordi Alonso[1,2,3], MD, PhD; Gemma Vilagut[1,2], PhD

[1]Hospital del Mar Research Institute, Barcelona, Spain

[2]Centro de Investigación Biomédica en Red (CIBER) de Epidemiología y Salud Pública, Instituto de Salud Carlos III, Madrid, Spain

[3]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain

[4]Department of Medicine, Science Health Faculty, Universitat Jaume I, Castelló de la Plana, Spain

[5]Epidemiology and Environmental Health Joint Research Unit, Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (FISABIO)-Universitat Jaume I-Universitat de València, Valencia, Spain

[6]Department of Education Sciences, University of La Rioja, Logroño, Spain

[7]Instituto Universitario de Investigación en Ciencias de la Salud-Instituto de Investigación Sanitaria Illes Balears (IUNICS-IDISBA), University of Balearic Islands, Palma de Mallorca, Spain

[8]Department of Psychology, University of Balearic Islands, Palma de Mallorca, Spain

[9]Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium

[10]Research Unit of Clinical Psychology, KU Leuven, Leuven, Belgium

[11]Department of Medical and Clinical Psychology, Tilburg University, Tilburg, Netherlands

[12]Department of Health Psychology, Miguel Hernandez University of Elche, Elche, Spain

[13]Department of Psychology and Sociology, University of Zaragoza, Teruel, Spain

Corresponding Author:
Philippe Mortier, MD, PhD
Hospital del Mar Research Institute
Carrer Doctor Aiguader, 88
Barcelona, 08023
Spain
Phone: 34 93 316 07 60
Fax: 34 93 316 07 60
Email: pmortier@researchmar.net

Abstract

Background: The use of ecological momentary assessment (EMA) designs has been on the rise in mental health epidemiology. However, there is a lack of knowledge of the determinants of participation in and compliance with EMA studies, reliability of measures, and underreporting of methodological details and data quality indicators.

Objective: This study aims to evaluate the quality of EMA data in a large sample of university students by estimating participation rate and mean compliance, identifying predictors of individual-level participation and compliance, evaluating between- and within-person reliability of measures of negative and positive affect, and identifying potential careless responding.

Methods: A total of 1259 university students were invited to participate in a 15-day EMA study on mental health problems. Logistic and Poisson regressions were used to investigate the associations between sociodemographic factors, lifetime adverse experiences, stressful events in the previous 12 months, and mental disorder screens and EMA participation and compliance. Multilevel reliability and intraclass correlation coefficients were obtained for positive and negative affect measures. Careless responders were identified based on low compliance or individual reliability coefficients.

Results: Of those invited, 62.1% (782/1259) participated in the EMA study, with a mean compliance of 76.9% (SD 27.7%). Participation was higher among female individuals (odds ratio [OR] 1.41, 95% CI 1.06-1.87) and lower among those aged ≥30 years (OR 0.20, 95% CI 0.08-0.43 vs those aged 18-21 years) and those who had experienced the death of a friend or family member in the previous 12 months (OR 0.73, 95% CI 0.57-0.94) or had a suicide attempt in the previous 12 months (OR 0.26,

XSL•FO
RenderX

95% CI 0.10-0.64). Compliance was particularly low among those exposed to sexual abuse before the age of 18 years (exponential of β=0.87) or to sexual assault or rape in the previous year (exponential of β=0.80) and among those with 12-month positive alcohol use disorder screens (exponential of β=0.89). Between-person reliability of negative and positive affect was strong (R_{kRn}>0.97), whereas within-person reliability was fair to moderate (R_{cn}>0.43). Of all answered assessments, 0.86% (291/33,626) were flagged as careless responses because the response time per item was <1 second or the participants gave the same response to all items. Of the participants, 17.5% (137/782) could be considered careless responders due to low compliance (<25/56, 45%) or very low to null individual reliability (raw Cronbach α<0.11) for either negative or positive affect.

Conclusions: Data quality assessments should be carried out in EMA studies in a standardized manner to provide robust conclusions to advance the field. Future EMA research should implement strategies to mitigate nonresponse bias as well as conduct sensitivity analyses to assess possible exclusion of careless responders.

(J Med Internet Res 2024;26:e55712) doi:10.2196/55712

KEYWORDS

experience sampling method; ecological momentary assessment; mental health; university students; participation; compliance; reliability; sensitivity analysis; mobile phone

Introduction

Background

Cross-sectional and traditional longitudinal study designs, which conduct assessments at one point or across several time points (usually 2-5 surveys months to years apart), have been widely used in the field of mental health epidemiology [1,2]. These methods serve well to provide detailed information on a population's mental health at a given time and present important insights into epidemiological patterns (eg, prevalence, incidence, and persistence rates). They are also able to identify who within the population under study is more at risk of having mental health problems by assessing risk and protective factors that are static in the middle to long term, such as sociodemographic characteristics, personality traits, or childhood experiences. However, it is known that mental health outcomes are complex and key symptoms and risk factors such as stress, affect, sleep, and suicidal ideation may fluctuate in a matter of hours or even minutes [3-7]. Thus, it is important to study not just who is at risk but also *when* and under what daily life conditions individuals may be more vulnerable in the short term. This has prompted a diversification in how mental health is studied by introducing high-resolution designs [8]. Ecological momentary assessment (EMA; also referred to as "experience sampling method" [ESM]) is a type of intensive longitudinal study that collects survey data several times a day over multiple days [9], providing information on variables as they fluctuate and unfold in the context of daily life [10]. These data allow for the development of explanatory and predictive models at both an individual and group level [8] and enable individualized and contextualized interventions at critical moments when the risk of mental health problems may be high [3,11,12]. To date, these studies have typically relied on small sample sizes using nonrepresentative sampling designs [7,13], leaving important questions about EMA data representativeness and reliability unaddressed.

Although methodological guidelines have emerged in recent years [7,14,15], there is a large heterogeneity of study designs as well as an underreporting of methodological details and data quality parameters [16,17]. Hence, to avoid methodological anarchy, it is important that studies provide a thorough description of the methodology and basic information about EMA data representativeness in terms of degree and determinants of participation and compliance [16,18]. Providing this information is fundamental for ensuring the validity, reliability, and generalizability of study findings and improving the overall quality of the research. Furthermore, there needs to be an evaluation of the reliability of the measures as well as the participants to assess whether the collected data are robust enough to provide solid results. Assessment of the reliability of the measures in EMA data provides information on how reliable scales are to capture differences between and within individuals [14,19]. Unreliable answers are those that have been provided without paying sufficient attention to the questions, what has been commonly described as "careless responding" [20-22], and can be evaluated in 2 ways. A priori items can be used that either explicitly ask participants whether they have paid attention while answering the assessments or test the participants' level of attention [23,24]. Alternatively, post hoc analyses have been proposed to detect possible careless responding by looking at parameters such as duration of response or variance in responses [20,21,24].

Objectives

It has been well established that the university years are a sensitive period for the onset and persistence of mental health problems [25-27], and there has been a growing interest in the use of EMA among university students [28-30]. To our knowledge, no large-scale epidemiological study has provided an in-depth analysis of the representativeness and reliability of EMA data among university students. In this paper, we present data collected as part of the Promoting Mental Health Among University Students (PROMES-U) project, in which undergraduate students from 5 public universities in Spain were invited to participate in a web-based survey and subsequent 15-day EMA study [31]. This study addresses this gap in the literature by assessing EMA data quality in complementary ways. First, we estimate the participation rate and mean compliance and identify predictors of individual-level participation and compliance. Second, we provide between- and within-person reliability scores for measures of negative affect (NA) and positive affect (PA). Third, we build on previous work to identify potential careless responding [20,21].

XSL•FO
RenderX

Methods

Study Design and Population

The PROMES-U project is a prospective observational multicenter cohort study implementing 2 web-based self-reported surveys (one at baseline and one at the 12-month follow-up) and a 15-day EMA study and is part of the World Health Organization World Mental Health International College Student Initiative [32]. The project's target population consists of undergraduate students from 5 public universities in Spain: University of the Balearic Islands, Jaume I University, Miguel Hernández University, Pompeu Fabra University, and University of Zaragoza. All Spanish-speaking students aged ≥18 years enrolled in any of the participating universities were eligible to participate. The 15-day EMA study immediately followed the baseline web-based survey (April 2022-June 2022) with asynchronous and continuous enrollment and is the focus of this report.

For further details regarding the PROMES-U project, its registration can be consulted on the internet [33].

Recruitment

Recruitment for the baseline web survey was conducted through mass emailing using university administrative lists (census sampling), with 1 to 3 reminder emails (in 3 of the 5 participating universities and in some colleges of a fourth one) as well as on-campus project information sessions, classroom announcements, links on the universities' websites, poster campaigns, and social media posts on the universities' accounts. A total of 2427 eligible students completed the baseline survey (operationalized as having provided basic sociodemographic variables, >50% of the required questions, and the mental health section of the survey; Multimedia Appendix 1).

Recruitment for the 15-day EMA study was done using quota sampling based on 30-day mental disorder screens in the baseline web survey. Students were hierarchically assigned to 1 of 4 mutually exclusive subsamples. The first subsample was composed of students who had reported passive or active suicidal ideation in the previous 30 days on the Columbia-Suicide Severity Rating Scale [34]. The second subsample consisted of students without 30-day suicidal ideation but with a positive screen for harmful alcohol use. This was assessed using core items of the Alcohol Use Disorders Identification Test scale (ie, frequency of consumption, quantity, and frequency of binge drinking episodes), which ranged from 0 (*never*) to 4 (*4 or more times a week*). Those who scored ≥4 in the sum of items screened positive for harmful alcohol use [35]. Students in the third subsample did not report 30-day suicidal ideation or harmful alcohol use but presented core symptoms of depression or anxiety in the previous 30 days through questions from the Composite International Diagnostic Interview screening scales [36]. Core symptoms of depression were "feeling sad or depressed," "feeling discouraged about how things were going in their life," and "taking little or no interest or pleasure in things," which were operationalized as indicating "most of the time" or more. Core anxiety symptoms were "feeling worried or anxious," "worrying about a number of different aspects in your life, such as your work, family, health, or finances," "feeling more worried than other people in your same situation," and "worrying excessively or too much," which were operationalized as indicating "most of the time" or more. Participants in the third subsample needed to present at least one of the aforementioned key symptoms of depression or anxiety. Finally, the fourth subsample (control) was composed of students who did not meet any of the aforementioned criteria.

To be invited to the EMA study, students had to complete the baseline web-based survey and provide answers for the 30-day mental disorder screens required to enable subsample assignation during quota sampling. After providing informed consent, participants were given instructions on how to participate in the study through an explainer video and link to the study's website, as well as information on how to download and register in the ExpiWell smartphone-based app available for Android and iOS.

A total of 1259 eligible students were invited, of whom 782 (62.1%) participated: 167 (21.4%) from the suicidal ideation subsample, 234 (29.9%) from the alcohol use subsample, 206 (26.3%) from the anxiety or depression subsample, and 175 (22.4%) from the control subsample. Multimedia Appendix 1 provides more information regarding the stages of EMA study recruitment.

Almost all participants (717/782, 91.7%) began the EMA study the same day they completed the baseline web-based survey, and 99.2% (776/782) began within a week of completing the baseline survey. The length of the EMA study was suggested by stakeholders in focus groups that were carried out during the design of the project [31] and by students who participated in a pilot study and falls within what is average for EMA studies in youth [18,37].

EMA Study Protocol and Measures

The 15-day EMA study was conducted following the baseline survey and assessed relevant constructs related to stress sensitivity (ie, affect reactivity) and mental health problems among university students. Measures were selected based on an extensive literature review by our research team and were adapted to our study's objectives using semirandom time intervals and recall periods to decrease participant burden. Participants received push notifications on their smartphones at a random time within 2-hour time intervals in the morning (8-10 AM), midday (noon-2 PM), afternoon (4-6 PM), and evening (8-10 PM). A reminder notification was sent after 30 minutes. The median response time was 15 minutes after receiving the prompt. The momentary assessment was marked as missed if the participant did not complete the assessment within the 2-hour window.

On days 1 and 15, the instrument evaluated 2-week major depressive disorder using the 9-item Patient Health Questionnaire [38], 2-week generalized anxiety disorder using the 7-item Generalized Anxiety Disorder scale [39], and 2-week mental well-being using the Short Warwick-Edinburgh Mental Wellbeing Scale [40].

On days 2 to 15, participants were asked to complete 4 momentary assessments each day. Each assessment included twelve 7-point Likert-type scales (from *not at all* to *a lot*) that

XSL•FO
RenderX

consisted of an introductory phrase ("Right now, I feel...") followed by very short descriptions of momentary states. They were used to assess stress levels ("...stress") from the ESM item repository [41] and energy levels ("...tired, with little energy") and perceived concentration capacity ("...having trouble concentrating") from the 9-item Patient Health Questionnaire [38].

Affect was assessed using 9 of the described 7-point Likert-type scales based on the 2-axis circumplex model of affect [42] (ie, high or low, and positive or negative valence). All items assessing affect were taken from validated scales, namely, the PHQ-9 [38], the 7-item Generalized Anxiety Disorder scale [39], the Short Warwick-Edinburgh Mental Wellbeing Scale [43], and the ESM item repository [41]. In total, 5 items were used to measure PA (ie, feeling "happy," "interest in doing things," "relaxed, "optimistic," or "pleasure in doing things") and 4 to measure NA (ie, feeling "nervous," "upset," "depressed," or "worried"). PA and NA scores for each individual momentary assessment were calculated as the person-specific mean of the corresponding items.

Finally, 3 additional multiple-response items taken from the ESM item repository assessed momentary social context (company), physical context (location), and activity [41]. Morning and evening assessments contained additional questions regarding the previous night's sleep duration and quality, as well as alcohol use, type of stressful events, and suicidal ideation within the previous 24 hours [41,44-47]. More information regarding the EMA measures used in this study is provided in Multimedia Appendix 2.

Explanatory Variables for Participation and Compliance

Explanatory variables for partic9ipation and compliance were collected in the baseline web survey, which was developed by the World Mental Health International College Student Initiative consortium and implemented using the Qualtrics survey platform (Qualtrics International Inc) [31,48]. In this survey, we included four groups of explanatory variables: (1) sociodemographic and college-related variables, (2) childhood and adolescence adverse experiences, (3) stressful events in the previous 12 months, and (4) mental disorder screens for the previous year. Multimedia Appendix 3 provides more information regarding the web-based survey measures.

Sociodemographic and college-related variables included sex, nationality, age, and field of study. Childhood and adolescence adverse experiences before the age of 18 years were assessed using 19 items adapted from several questionnaires. Items that assessed parental psychopathology (ie, parent with any serious mental or emotional problems, substance abuse, suicidal thoughts and behaviors or death by suicide, and criminal activities or interpersonal violence) were taken from the Adverse Childhood Experiences Study survey [49]. Items measuring neglect and sexual and emotional abuse were from the Childhood Trauma Questionnaire–Short Form [50]. Physical abuse was reported using items from the Childhood Trauma Questionnaire–Short Form and the Army Study to Assess Risk and Resilience in Servicemembers New Soldier Study survey [51]. Bully victimization (ie, direct verbal or physical bullying as well as indirect bullying and cyberbullying) was measured using items of the Youth Risk Behavior Survey [52]. Finally, dating violence was measured using items from the 2016 to 2017 Healthy Minds Study [53].

Past-year stressful events were assessed using 11 dichotomous items based on the Army Study to Assess Risk and Resilience in Servicemembers survey [54] that included life-threatening illness or injury of a friend or family member; death of a friend or family member; breakup with a romantic partner; cheating of a romantic partner; serious betrayal by someone close; serious ongoing arguments or breakup with friends or family members; and having had a life-threatening accident or illness, been physically assaulted, experienced sexual harassment, been sexually assaulted or raped, or had a serious legal problem (ie, any serious legal problem or problems with the police).

We evaluated major depressive disorder and generalized anxiety in the previous 12 months using the Composite International Diagnostic Interview screening scales [36]. We evaluated alcohol use disorder in the previous 12 months using the Alcohol Use Disorders Identification Test [55]. We assessed suicidal thoughts and behaviors in the previous 12 months using an adapted version of the Columbia-Suicide Severity Rating Scale [34,56].

Statistical Analysis

EMA Study Sample Representativity

First, we assessed the representativity of the EMA sample vis-à-vis the target population by comparing the distribution of sociodemographic and college-related characteristics of the sample during various stages of EMA study recruitment on the one hand with the distribution of these characteristics in the target populations on the other hand. Target populations were defined in two ways: (1) all students in Spanish public universities, with information on key characteristics obtained from the Ministry of Education, Vocational Training, and Sports [57]; and (2) all students in the participating universities, with the information being obtained from the universities themselves.

Predictors of EMA Study Participation and Compliance

EMA study participation rate and mean compliance were calculated for the total sample and stratified by the explanatory variables under study (refer to the Explanatory Variables for Participation and Compliance section). Participation rate was defined as the number of respondents who provided a usable response divided by the total number of initial personal invitations requesting participation (ie, participants who completed at least one momentary assessment divided by those who were invited to participate in the EMA study [58]). Compliance for each participant was defined as the percentage of momentary assessments completed by the participant out of the maximum number of momentary assessments allowed by the design [16]—in our case, a maximum of 56 assessments. Mean compliance was calculated from the individual compliance rates of all EMA study participants. The associations between the explanatory variables and participation were assessed using logistic regression, whereas Poisson regression was used to assess associations between the explanatory variables and compliance. Separate models were constructed for each

XSL•FO
RenderX

explanatory variable, each time adjusting for all sociodemographic and college-related variables. Results are presented as odds ratios (ORs) for logistic regression models and as the exponential of the β coefficient for Poisson regression models, with corresponding 95% CIs.

Variance and Reliability of Measures

The overall variance of PA and NA scores was broken down into between-person and within-person variances using a multilevel model. Between-person variance measures the variance of the measure due to differences observed between participants, whereas within-person variance accounts for the variance due to differences in scores of that measure within a participant. Intraclass correlation coefficients (ICCs) for PA and NA were then calculated by dividing the between-person variance by the sum of the between- and within-person variance [14]. The ICC was calculated for the full sample of EMA study participants (n=782) as well as separately for the control subsample (n=175) and the subsample of participants with mental health problems in the previous 30 days (n=607). The ICC is a useful measure to assess whether the construct to be assessed changes within a person in such a way that its frequent assessment is justified. An ICC of <0.5 means that variance is mostly due to within-person differences (ie, momentary fluctuations). The *ggplot* and *lattice* packages in R (R Foundation for Statistical Computing) were used to visualize within-person variability for PA and NA in 7 random individuals within each subsample (28 in total).

We evaluated between- and within-person reliability of PA and NA measures using a multilevel mixed-effects model in which items and time were treated as random effects, with items nested within assessments that were then nested within persons [59]. Between-person reliability, obtained using the R_{kRn} estimate developed by Shrout and Lane [60], reflects the internal consistency reliability of between-person differences in the set of items that form the PA and NA scale scores. On the other hand, within-person reliability was obtained using the R_{cn} estimate [60], which assesses the internal consistency reliability of the set of PA and NA items across time within individuals [61]. These indicators range from 0 to 1 and can be interpreted using classic test theory conventions (0.00-0.10=practically no reliability; 0.11-0.40=slight reliability; 0.41-0.60=fair reliability; 0.61-0.80=moderate reliability; 0.81-1.00=substantial reliability) [62]. Raw Cronbach α reliabilities were calculated for each participant for the PA and NA items across time. Reliability estimates were obtained using the *mlr* function from the *psych* R package (version 2.2.9) [59].

Careless Responding

We evaluated careless responding by expanding on previous work by Jaso et al [21], which provides a data-driven approach and makes a distinction between careless responses and careless responders. They identify "careless responses" by evaluating how the relationship between psychometric antonyms (ie, items with opposite meaning) is affected at different thresholds of three parameters: (1) an overly fast time to complete the assessment, (2) an overly narrow within-assessment response variance, and (3) a high percentage of items with the same response. The premise of the method is that, when the relationship between the psychometric antonyms becomes neutral or positive, the response can be considered as careless. Although we followed the method by Jaso et al [21], we did not use the R package they developed as it was not applicable due to differences in measurements. For example, the variables in the study by Jaso et al [21] were measured using sliders that ranged from 0 to 100 with a default answer at 50, whereas our instrument consisted of obligatory checkbox questions ranging from 1 to 7 with no default answers.

In line with the work by Jaso et al [21], psychometric antonyms were determined by selecting the 2 items that had the strongest negative correlation (ie, feeling "nervous," "anxious," "on edge" and feeling "relaxed"; Multimedia Appendix 4). The parameter used to evaluate whether an assessment had been completed too fast was the time per item (ie, the average number of seconds per item that it took a participant to complete a specific momentary assessment). This was calculated by dividing the total time spent completing the assessment by the number of items of that assessment. The SD of item responses and percentage of items with the same response was calculated for each momentary assessment for the conjoint set of 12 items common to all momentary assessments, all measured using a 7-point Likert-type scale assessing momentary stress, affect, depression, anxiety, and well-being (refer to the EMA Study Protocol and Measures section). Different thresholds were evaluated for each of the 3 careless responding parameters using graphs that displayed the relationship between the scores for feeling "nervous," "anxious," "on edge" and feeling "relaxed" at each of the thresholds.

Participants were flagged as possible "careless responders" under 2 circumstances. The first was if they had less than the minimum number of noncareless assessments needed to have a stable estimate of the mean and variability of affect [21]. To identify the number of assessments required, we selected a random subset of an increasing number of assessments for each participant (2 to 40) and, for each number of randomly selected assessments, we calculated the 2-way ICC for absolute agreement comparing the mean and SD of the PA and NA scores of each random selection to the participant's gold standard. We did not consider the participant's gold standard to be all 56 assessments given that only 2.9% (23/782) of participants completed all 56 assessments. Instead, we considered 40 assessments to be the gold standard, which allowed us to include 78% (610/782) of the participants in this analysis. We ran the analysis 5 times, thus selecting 5 different random samples of assessments per individual. The minimum number of reliable assessments needed was determined whenever the ICC score stabilized to a value of ≥0.90 for both the mean and SD. In addition, participants were flagged as possible careless responders if they had a within-person reliability score of <0.11 (ie, practically no reliability [62]) for either PA or NA.

All analyses were conducted using R (version 4.1.1). For the reporting of this study, we followed an adapted version of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist specific for reporting EMA studies (Checklist for Reporting EMA Studies) [63]. The completed checklist can be found in Multimedia Appendix 5.

XSL•FO
RenderX

Ethical Considerations

Ethics approval was provided by the Parc de Salut Mar Clinical Research Ethical Committee (protocol 2020/9198/I). The project complies with the principles established by national and international regulations, including the Declaration of Helsinki. Explicit informed consent was obtained at the beginning of the web-based surveys, and a separate consent form was provided at the end of the web-based survey for participation in the EMA study. Data were deidentified before all analyses through pseudoanonymization using encrypted identifiers to ensure privacy, and neither the study investigators nor the individuals conducting the analyses had access to the personal data. Participants who completed at least 80% of the momentary assessments as well as the mental health screener scales of days 1 and 15 received a monetary incentive of 30€ (US $32.03). Participants were informed in advance of the requirements to receive the incentive and were not informed explicitly when they qualified for it, nor did the app show any information on compliance other than which assessments had been completed.

Results

EMA Study Sample Representativity

A total of 1259 baseline survey participants were invited to take part in the EMA study, of whom 1126 (89.4%) provided informed consent, and 782 (62.1% of those invited; 69.4% of those providing informed consent) participated (Multimedia Appendix 1). Table 1 shows the distribution of sociodemographic and college-related characteristics of the target populations and of the sample across the various phases of recruitment for the EMA study up until EMA participation. Compared to the target populations, EMA participants were more likely to be women (619/782, 79.2% vs 39,214/69,805, 56.2%-753,749/1,338,304, 56.3%), younger students (577/782, 73.8% vs 724,853/1,338,304, 54.2%-43,664/69,805, 62.6%), and science (94/782, 12% vs 3633/69,805, 5.2%-84,750/1,338,304, 6.3%) and health sciences (219/782, 28% vs 258,967/1,338,304, 19.4%-14,316/69,805, 20.5%) students.

These observed differences in EMA sample composition compared to the target population are mainly due to a different likelihood of participation in the baseline survey and, to a far lesser extent, to a different likelihood of EMA participation among those invited from the baseline survey.

XSL•FO
RenderX

Table 1. Distribution of key sociodemographic and college-related characteristics of the ecological momentary assessment (EMA) study target student populations and of the EMA study participants across the EMA study recruitment phases.

	EMA study target population		EMA study recruitment phases			
	All public universities in Spain[a] (n=1,338,304), n (%)	PROMES-U[b] universities[c] (n=69,805), n (%)	Completed baseline web survey[d] (n=2427), n (%)	Invited to the EMA study[e] (n=1259), n (%)	Provided IC[f] for EMA study (n=1126), n (%)	Participated in EMA study[g] (n=782), n (%)
Sex						
Male	584,555 (43.7)	30,591 (43.8)	609 (25.1)	288 (22.9)	243 (21.6)	163 (20.8)
Female	753,749 (56.3)	39,214 (56.2)	1818 (74.9)	971 (77.1)	883 (78.4)	619 (79.2)
Nationality						
Spanish	1,253,880 (93.7)	65,444 (93.8)	2236 (92.1)	1156 (91.8)	1027 (91.2)	722 (92.3)
Other	84,424 (6.3)	4361 (6.2)	191 (7.9)	103 (8.2)	99 (8.8)	60 (7.7)
Age (y)						
18-21	724,853 (54.2)	43,664 (62.6)	1722 (71)	894 (71)	800 (71)	577 (73.8)
22-25	320,156 (23.9)	17,701 (25.4)	539 (22.2)	297 (23.6)	267 (23.7)	176 (22.5)
26-29	105,557 (7.9)	3988 (5.7)	73 (3)	36 (2.9)	33 (2.9)	21 (2.7)
≥30	187,738 (14)	4452 (6.4)	93 (3.8)	32 (2.5)	26 (2.3)	8 (1)
Field of study						
Arts and humanities	140,969 (10.5)	6177 (8.8)	269 (11.1)	154 (12.2)	132 (11.7)	88 (11.3)
Sciences	84,750 (6.3)	3633 (5.2)	217 (8.9)	141 (11.2)	121 (10.7)	94 (12)
Health sciences	258,967 (19.4)	14,316 (20.5)	667 (27.5)	335 (26.6)	315 (28)	219 (28)
Social and legal sciences	616,880 (46.1)	32,340 (46.3)	997 (41.1)	470 (37.3)	420 (37.3)	280 (35.8)
Engineering and architecture	236,738 (17.7)	13,339 (19.1)	277 (11.4)	159 (12.6)	138 (12.3)	101 (12.9)

[a]Information retrieved from the Spanish Ministry of Education and Professional Formation [64].

[b]PROMES-U: Promoting Mental Health Among University Students.

[c]Information obtained from each of the 5 participating universities, accessible on demand.

[d]Defined as making it to the end of the baseline web survey, as well as having provided basic sociodemographic and college-related variables (ie, sex, nationality, age, and field of study), having completed the mental health section, and having answered >50% of the required questions.

[e]Defined as having been assigned to a group following the hierarchical quota sampling scheme (refer to the Sampling Procedures subsection in the Methods section).

[f]IC: informed consent.

[g]Defined as having completed at least one momentary assessment in the EMA study.

Predictors of EMA Study Participation and Compliance

The EMA study's participation rate was 62.1% (782/1259). Mean compliance was 76.9% (SD 27.7%; range 1.9%-100%), with a median of 89.3% and a mode of 96.4% (Figure 1). A total of 82.5% (645/782) of the participants completed at least one assessment on the final day.

Multimedia Appendix 6 shows the associations between participation and compliance and sociodemographic and college-related characteristics, childhood and adolescence adverse experiences before the age of 18 years, stressful experiences, mental disorder screens, and suicidal thoughts and behaviors in the previous 12 months.

Female individuals had significantly higher odds of participation (OR 1.41, 95% CI 1.06-1.87), whereas those who were aged ≥30 years (OR 0.20, 95% CI 0.08-1.06), had had a death of a friend or family member in the previous year (OR 0.73, 95% CI 0.57-0.94), or had had a suicide attempt in the previous year (OR 0.26, 95% CI 0.10-0.64) had significantly lower odds of participation.

Compliance was significantly lower among non-Spanish participants (exponential of β=0.94, 95% CI 0.90-0.98), older age groups (exponential of β=0.88, 95% CI 0.82-0.94 for those aged 26-29 years; exponential of β=0.86, 95% CI 0.76-0.96 for those aged ≥30 years), and students of social and juridical sciences (exponential of β=0.92, 95% CI 0.89-0.95) than among arts and humanities students. Compliance was significantly lower among those who had experienced any of the childhood and adolescence adverse experiences under study except for parental psychopathology (exponential of β range 0.87-0.98), especially among those who experienced neglect (exponential of β=0.92, 95% CI 0.89-0.94) or sexual abuse (exponential of

XSL•FO
RenderX

β=0.87, 95% CI 0.82-0.92). In total, 45% (5/11) of the stressful experiences in the previous 12 months under study were negatively associated with compliance (exponential of β range 0.80-0.97), especially sexual assault or rape (exponential of β=0.80, 95% CI 0.74-0.86). Compliance was higher among those with any stressful event in the previous 12 months (exponential of β=1.03, 95% CI 1.01-1.06) and those with exactly one stressful experience in the previous 12 months (exponential of β=1.09, 95% CI 1.05-1.13, compared to none). Regarding adverse mental health, compliance was lowest among those with a positive screen for alcohol use disorder in the previous 12 months (exponential of β=0.89, 95% CI 0.85-0.93) but also significantly lower among those with positive screens for major depression in the previous 12 months (exponential of β=0.95, 95% CI 0.93-0.97), generalized anxiety (exponential of β=0.95, 95% CI 0.92-0.97), and suicidal ideation (exponential of β=0.97, 95% CI 0.95-0.99). Compliance gradually decreased throughout the 15-day EMA study in a linear fashion both in the total sample and in each separate subsample (Multimedia Appendix 7).

Figure 1. Violin plots showing the distribution of compliance (ie, the percentage of momentary assessments completed out of the maximum of 56 allowed by the design) stratified by each subsample of students: suicidal ideation, harmful alcohol use, anxiety or depression, and controls. Alc: harmful alcohol use; AnxDep: anxiety and depression; SI: suicidal ideation.

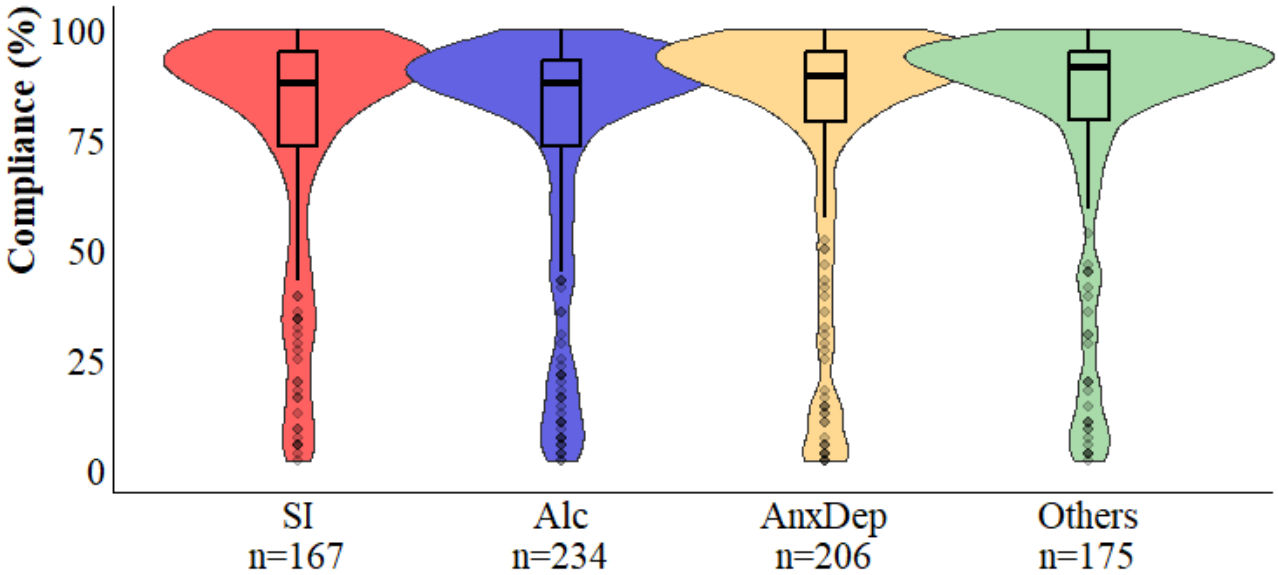

Variance and Reliability of Measures

Figure 2 presents PA and NA scores in all assessments for a random selection of 7 individuals within each subsample. Important differences in PA and NA scores across individuals were observed. For example, when looking at NA, some individuals presented very little intraindividual variability (eg, alc_508, anxdep_597, and cont_524), and others had important fluctuations from one assessment to the next (eg, sui_360 and anxdep_293; Figures 2A and 2B).

The total ICCs for mean PA and NA were 0.54 and 0.52, respectively. The ICC for both PA and NA was 0.50 in the subsample of participants with mental health problems, meaning that, on average, between- and within-person variance were very similar (Table 2), in line with the differences in variability observed in Figure 2. In the control subsample, the ICCs for PA and NA were slightly higher and lower (ICC=0.54 and 0.46, respectively). Overall affect variance (both between and within persons) was greater among those in the adverse mental health subsamples (Table 2).

As seen in Table 3, between-person reliability was high for both PA and NA (ie, random-effects reliability; R_{kRn}=0.99). Within-person reliability was lower but still acceptable (R_{cn}=0.48 and 0.59 for PA and NA, respectively). Within-person reliability was slightly higher on average among those with mental health problems. Within-person reliabilities were consistently higher for NA than for PA. Within-person reliability scores were highest when including all measured items both for PA and NA.

Multimedia Appendix 8 shows individual Cronbach α reliability coefficients by participant, which range from negative values to 0.97 for PA and 0.99 for NA. As expected, individuals that only answered 1 assessment (8/781, 1%) did not provide enough information to have reliability outcomes. Among the rest, there were 3.7% (19/782) of individuals with negative values for individual NA or PA reliability. Of all participants, 86.8% (679/782) and 66.8% (522/782) had moderate to substantial individual reliability (ie, raw Cronbach α>0.60) for PA and NA, respectively.

Participants that appeared to have very low NA variability in Figure 2B (eg, anxdep_597 and cont_524) had the lowest within-person reliability for NA (Multimedia Appendix 8). Furthermore, participant alc_508, who presented practically no variation in their mean level of reported NA, had no value for raw Cronbach α (Multimedia Appendix 8). Generally, participants with very low individual reliability scores presented either very low compliance rates or low variability in their answers.

XSL•FO
RenderX

Figure 2. (A) Positive and (B) negative affect scores across all momentary assessments for 7 random participants from each subsample (n=28) selected among those who completed at least one assessment on the last day of the ecological momentary assessment study (n=645). Red represents the suicidal ideation subsample, blue represents the harmful alcohol use subsample, yellow represents the anxiety and depression subsample, green represents the control subsample.

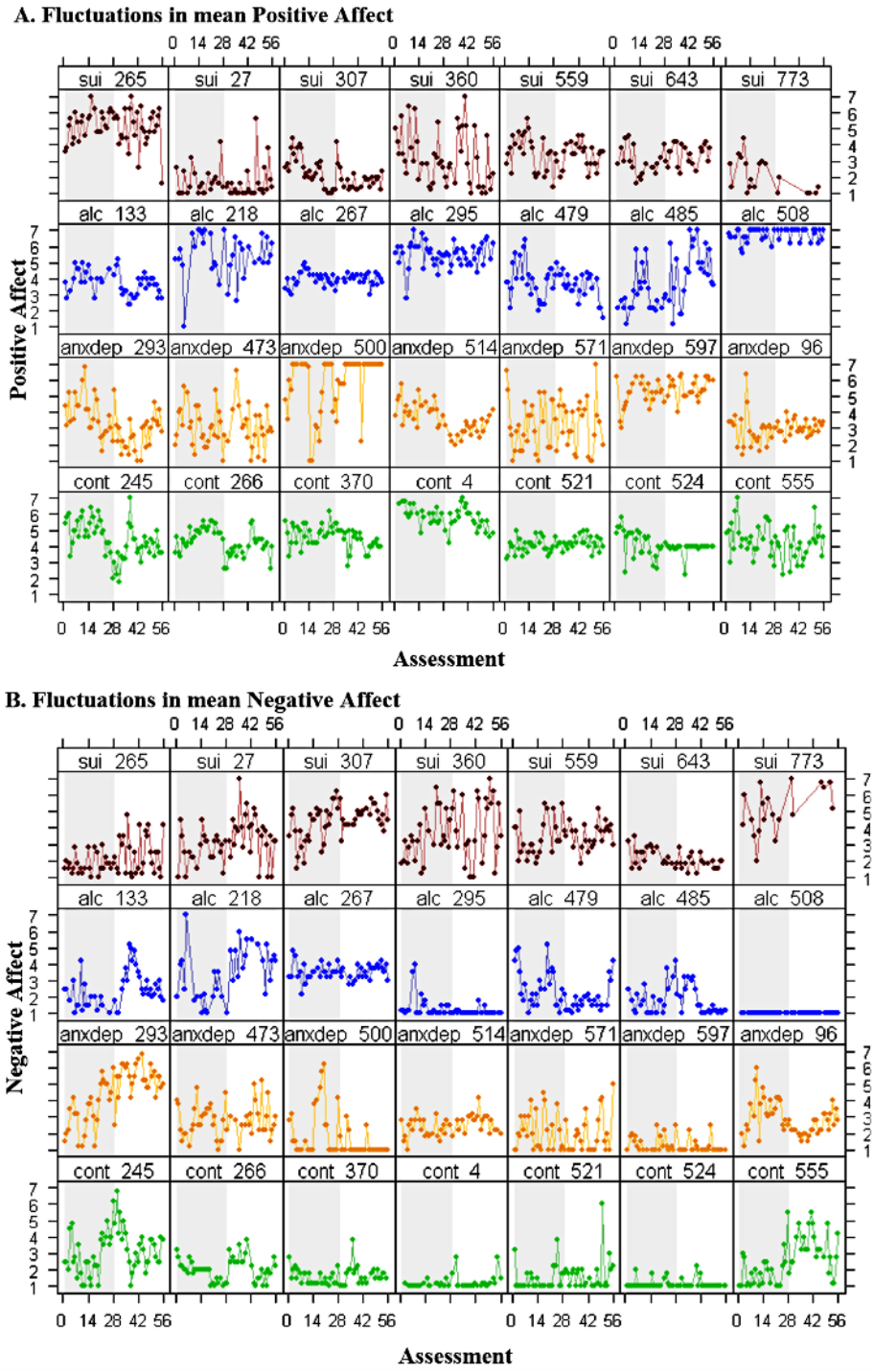

XSL•FO
RenderX

Table 2. Intraclass correlation coefficients (ICCs) of positive and negative affect to show how much of their variability was due to differences between participants as opposed to fluctuations within each participant, calculated among ecological momentary assessment (EMA) participants and stratified by subsamples (N=782)[a].

	EMA participants			Control subsample of EMA participants (n=175)			Mental health problem subsamples of EMA participants (n=607)		
	Between-person variance	Within-person variance	ICC	Between-person variance	Within-person variance	ICC	Between-person variance	Within-person variance	ICC
Positive affect	0.94	0.81	0.54	0.75	0.59	0.56	0.88	0.88	0.50
Negative affect	0.88	0.80	0.52	0.42	0.50	0.46	0.88	0.89	0.50

[a]Interpretation: an ICC of <0.5 means that the variance is mostly due to within-person differences (ie, momentary fluctuations).

Table 3. Multilevel between- and within-person reliability coefficients of positive affect (PA) and negative affect (NA) among ecological momentary assessment (EMA) participants stratified by subsamples (N=782).

	EMA participants		Control subsample of EMA participants (n=175)		Mental health problem subsamples of EMA participants (n=607)	
	PA items	NA items	PA items	NA items	PA items	NA items
Between-person reliability (R_{kRn}[a])	0.99	0.99	0.97	0.98	0.98	0.99
Within-person reliability (R_{cn}[b])	0.48	0.59	0.43	0.52	0.49	0.60

[a]R_{kRn}: generalizability of between-person differences averaged over time; time nested within people.

[b]R_{cn}: generalizability of within-person variations averaged over items; time nested within people [3].

Careless Responses

Different thresholds were evaluated for each of the 3 careless responding parameters (ie, time per item, SD of item responses in each assessment, and percentage of items of each assessment that fell at the mode of that assessment).

The distribution of assessments according to time per item (Figure 3A) was sharply skewed to the right, with a median at 3.9 seconds per item. Figure 4 shows that, in assessments with a time per item of ≤1 second (r=−0.04), the relationship between feeling “anxious” and feeling “relaxed” flattens, thus conveying that responses below this threshold are likely careless. A total of 0.13% (44/33,626) of all completed assessments were below this threshold.

Figure 3B shows that the distribution of the SDs of the item responses in each assessment was normal, with a group of outliers that had an SD of 0 (indicating no variability in item responses). The relationship between psychometric antonyms shifted in those assessments at an SD of ≤1 and became bluntly positively correlated at an SD of ≤0.5 and ≤0.25 (r=0.94 and 1.00, respectively), which comprised 10.1%, 1.2%, and 0.7% of all answered assessments, respectively (Figure 5).

Figure 3C shows the distribution of the percentage of individual responses within a given assessment that fall at the mode of that assessment. This distribution also had a right skew. The relationship between “anxious” and “relaxed” items flattened in assessments with ≥60% of items with answers at the mode (r=0.16) and became sharply positive when ≥80% of the answers fell at the mode (r=0.82), which comprised 4.4% and 1.5% of all answered assessments, respectively (Figure 6).

Although the relationship between the psychometric antonyms flattened when the SD was of ≤1 and when ≥60% of items with answers fell at the mode, most of the answers that were below and above these thresholds were not outliers in the distribution (Figures 3B and 3C) and were theoretically plausible responses. Thus, we decided that only assessments with no variance in responses whatsoever (ie, SD of 0 or 100% of items at the mode) should be considered careless responses. Of all the answered assessments, 0.77% (261/33,626) had no variation in responses.

Considering the aforementioned thresholds, a total of 0.86% (291/33,626) of all completed assessments were flagged as having been responded carelessly. There was little overlap between careless responses identified through no variation and those identified through a time per item of ≤1 second (Figure S3A in Multimedia Appendix 9).

When looking at how careless responses clustered within individuals, 8.8% (69/782) of participants had at least one assessment flagged as careless, 0.8% (6/782) had more than a third of their answers flagged as careless, and 0.26% (2/782) had at least half of their answers flagged as careless (Multimedia Appendix 8). Furthermore, half of all careless responses clustered around 8 participants.

To evaluate whether careless responses could be related to monetary incentives, we looked at the relationship between the number of careless responses and compliance by performing a logistic regression dichotomizing compliance at the incentive threshold (ie, 80%). We found that there were no differences in the percentage of careless responses among those who did and did not receive an incentive (OR 0.99, 95% CI 0.96-1.02; Multimedia Appendix 10).

XSL•FO
RenderX

Figure 3. Distribution of time per item (TPI) of each assessment, SD of item responses in each assessment, and percentage of items in each assessment that fell at the mode among completed assessments (n=33,626) by ecological momentary assessment study participants (n=782).

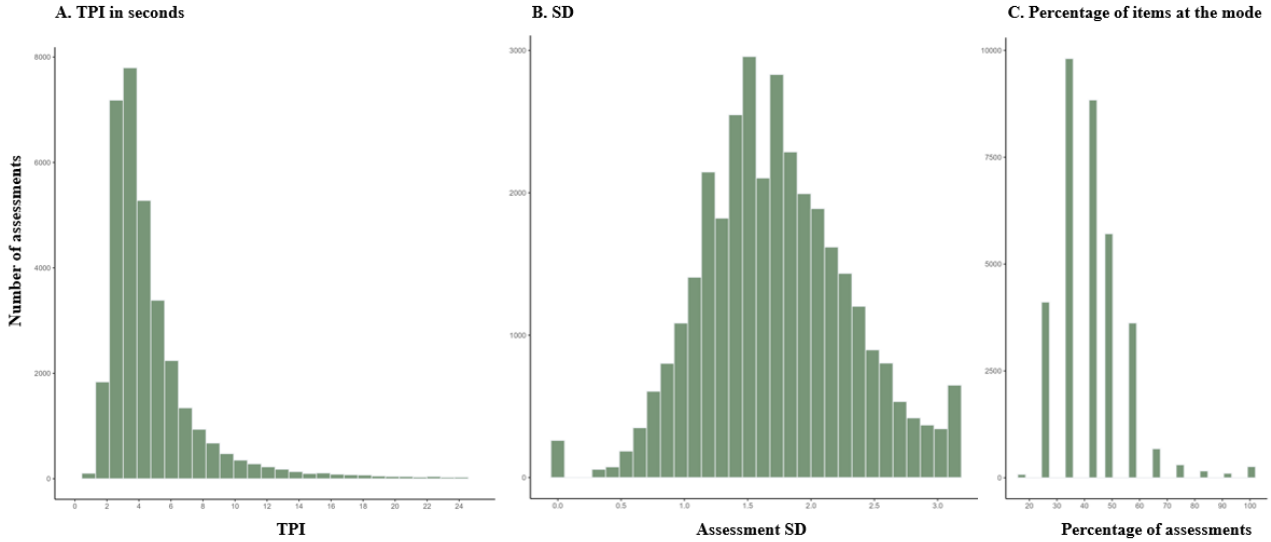

Figure 4. Comparing the relationship between psychometric antonym items (feeling nervous, anxious, on edge and relaxed) at different time per item (TPI) thresholds, measured in seconds, for all completed assessments (n=33,626). Purple bubbles and trendlines indicate assessments below the indicated threshold; salmon-colored bubbles and black lines indicate all answered assessments. Note: bubble sizes are relative to the number of assessments at each combination of responses to the 2 selected items.

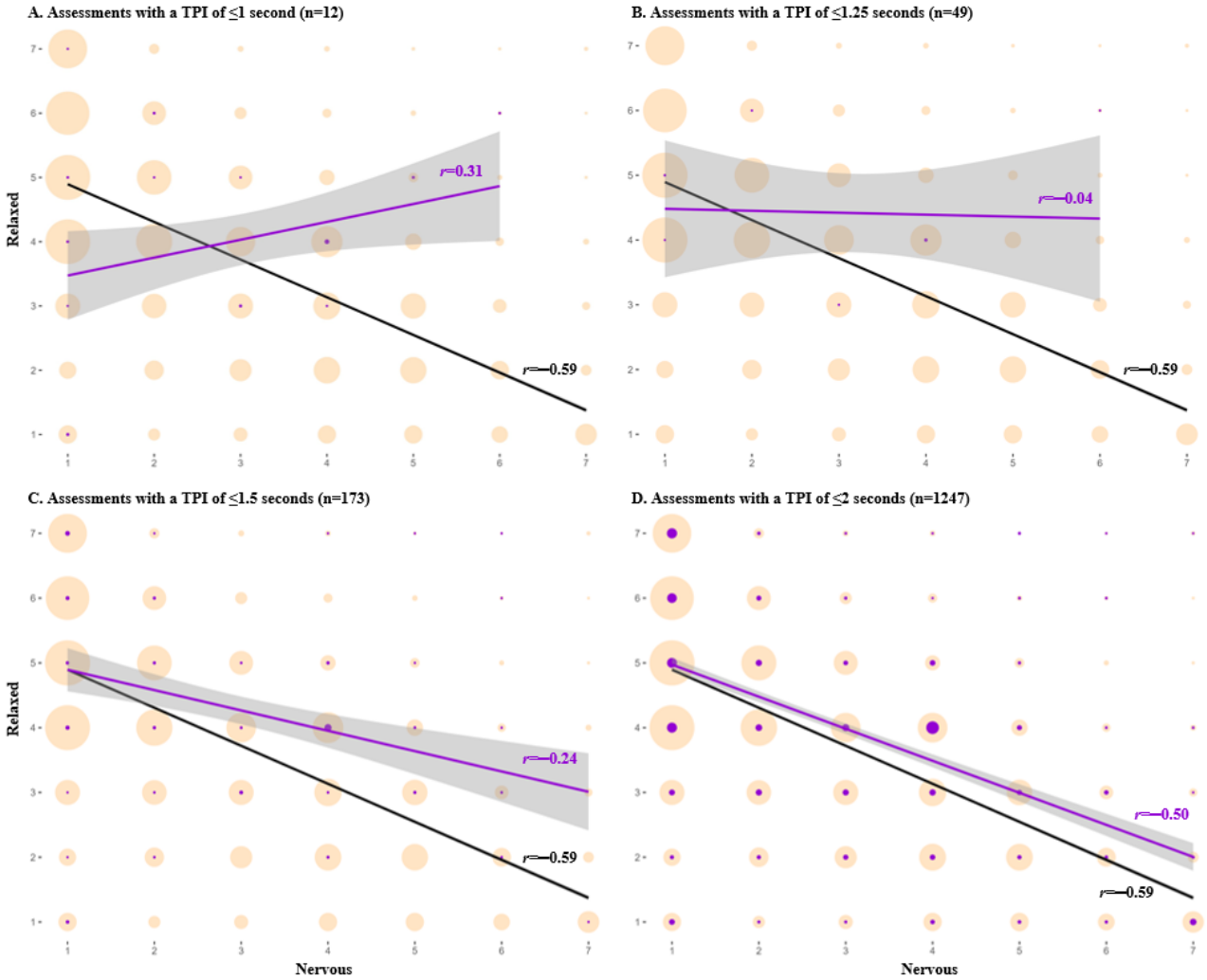

XSL•FO
RenderX

Figure 5. Comparing the relationship between psychometric antonym items (feeling nervous, anxious, on edge and relaxed) at different thresholds for the SD of item responses for all completed assessments (n=33,626). Purple bubbles and trendlines indicate assessments below the indicated threshold; yellow bubbles and black lines indicate all answered assessments. Note: bubble sizes are relative to the number of assessments at each point of correlation.

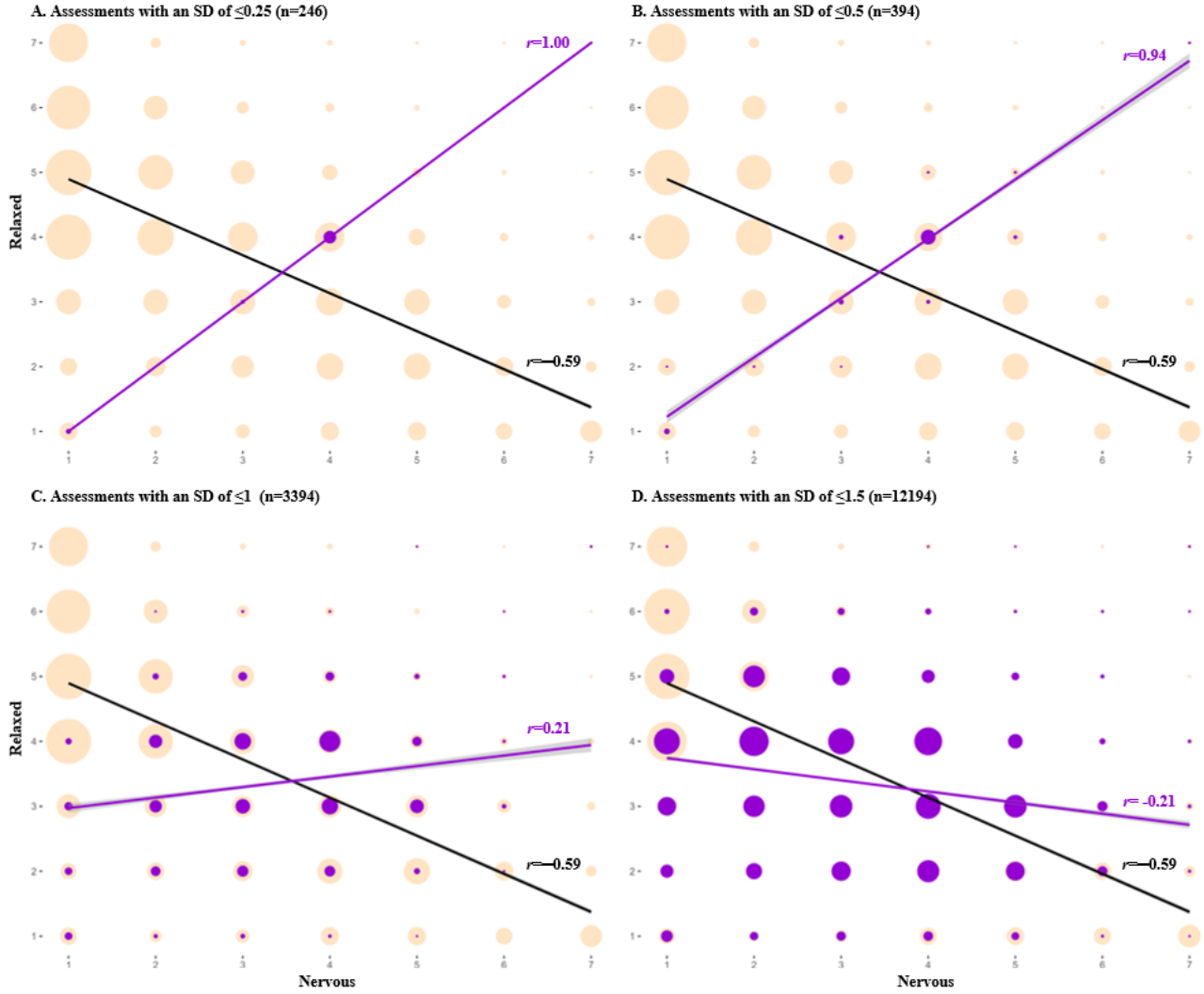

XSL•FO
RenderX

Figure 6. Comparing the relationship between selected psychometric antonym items (feeling nervous, anxious, on edge and relaxed) at different thresholds of percentage of items per assessment at the mode. Purple bubbles and trendlines indicate assessments below the indicated threshold; yellow bubbles and black lines indicate all answered assessments. Note: bubble sizes are relative to the number of assessments at each point of correlation.

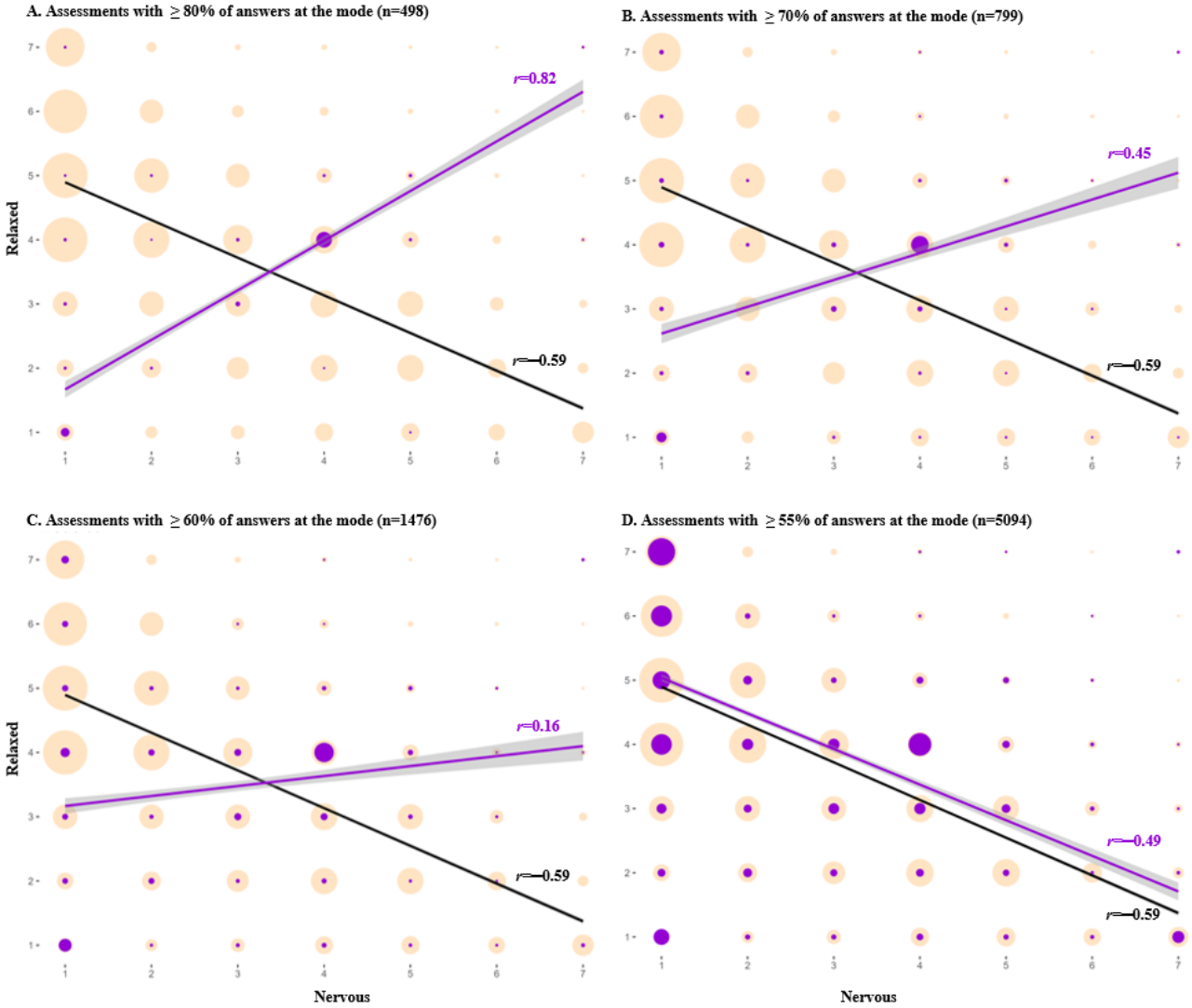

Careless Responders

Figure 7 shows the number of assessments needed to have a stable estimate of the mean and variability of PA and NA scores. The mean PA and NA scores reached an ICC of 0.9 at approximately 4 assessments, clearly stabilizing at approximately 12 assessments, whereas approximately 25 were needed to be able to capture variability well. There was a total of 3.7% (29/782) of the participants who had <25 valid assessments (ie, completed and not flagged as carelessly responded) according to the considered parameters. In total, 6.4% (50/782) of the participants had very low or null individual reliability (raw Cronbach $\alpha<0.11$) for either PA or NA.

In total, 17.5% (137/782) of the participants were identified as potential careless responders (Figure S3B in Multimedia Appendix 9).

XSL•FO
RenderX

Figure 7. Intraclass correlation coefficient (ICC) for positive affect (PA) and negative affect (NA) score mean and SD between an increasing number of random assessments (x-axis) and its value when selecting 40 assessments (considered the gold standard) using data from all participants who completed ≥40 assessments (n=610).

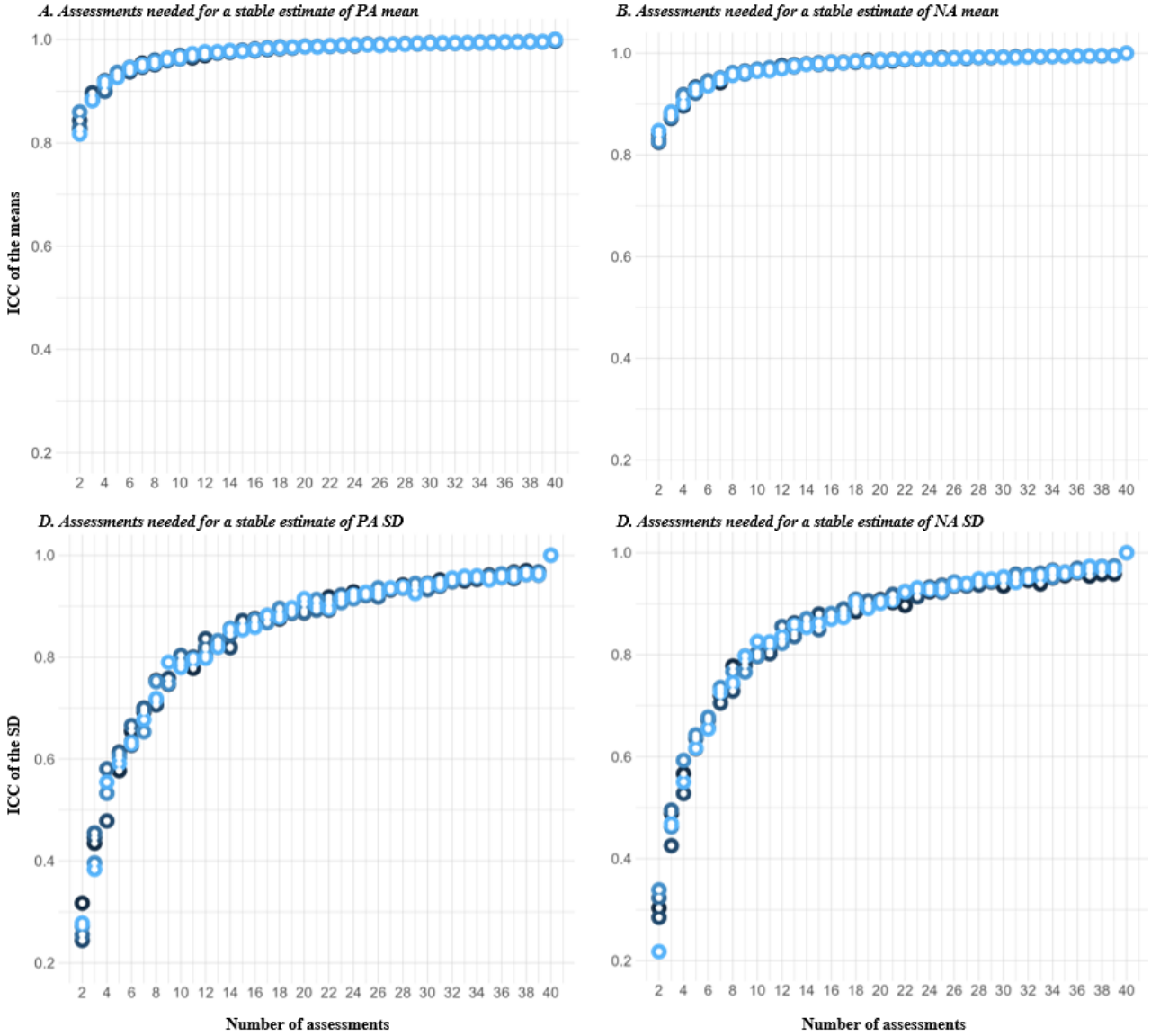

Discussion

Contextualization of Principal Findings

We investigated EMA data quality in a large sample of university students that included participants who screened positive for mental health problems (607/782, 77.6%) as well as students who did not (175/782, 22.4%) by using state-of-the-art data-driven analyses to assess participation, compliance, and reliability indicators. A total of 1259 students were invited to participate in the EMA study, with a participation of 782 (62.1%) and mean compliance of 76.9% (SD 27.7%). We found that participants in our EMA study were representative of the invited population, whereas compliance was slightly different across groups of explanatory variables. The data were highly reliable to account for between-person differences in NA and PA and fair to moderate to study differences within each participant. The findings suggest that 0.86% (291/33,624) of all completed assessments were responded carelessly and that 17.5% (137/782) of the participants were potential careless responders.

Participation and its predictors are rarely investigated in EMA studies given that such analyses require having information on those who refuse to participate. This demands epidemiological study designs that include careful tracking of potential participants across recruitment phases as well as probability sampling techniques. By linking our EMA study to a large cohort study, we were able to investigate and predict EMA participation and compliance, including a wide range of relevant candidate predictor variables for these outcomes. We found that men had significantly lower rates of participation, which is a tendency across scientific studies [65]. Interestingly, the only mental health outcome that was significantly associated with lower participation was having had a suicide attempt in the previous year. This could be due to multiple reasons outside the scope of this study that could affect the desire or capacity to participate in an intensive longitudinal study [66]. It is possible that students that have attempted suicide in the past

XSL•FO
RenderX

year have some with the highest levels of psychological pain and stress. The only stressful event related to lower participation was having had a friend or family member die in the previous year. Mourning is a process that is also highly distressing psychologically [67] and is more common than having recent suicide attempts, thus having potentially a higher impact on overall EMA study participation.

Mean compliance was found to be similar to what other studies have reported [16,68], including a pattern of declining compliance as the days passed [16,24,69,70]. Although we cannot empirically confirm that compliance may have been positively impacted by monetary incentives, there seems to be a relationship between the 2; we did observe an inflection point around the 80% mark of compliance, and the literature has also found that studies with monetary incentives tend to have better compliance rates [16,68]. Concerning predictors of compliance, we found that almost all childhood adversities and a significant number of recent stressful events were associated with lower compliance. However, it is important to stress that effect sizes on compliance were very low and the differences across groups were generally small. Compliance was lowest among participants that had been sexually abused during their childhood or adolescence and those who had been sexually assaulted or raped in the previous year. Childhood and recent sexual abuse have been widely identified as being associated with grave emotional distress [71-73]. Concerning mental health outcomes, positive screening for alcohol use disorder was the strongest mental health predictor for low compliance. This finding aligns with the literature, which argues that activities or behaviors such as drinking alcohol and experiencing stress, which draw attention away from participating, could result in lower compliance [66,74]. Lower compliance rates among those who have lived through extremely stressful events and those who have poor mental health can also be related to how these experiences or states can impact cognition [66].

When testing reliability, we found that between-person reliability for both PA and NA was excellent (R_{kRn}=0.99), whereas within-person reliability was lower (R_{cn}=0.48 and 0.59 for NA and PA, respectively). EMA studies typically do not report reliability coefficients; however, when they do, within-person reliability is lower than between-person reliability, with similar results to ours [19,75,76]. This is to be expected because within-person variability is also generally smaller than variability between individuals [77]. Although it is important to evaluate between- and within-person reliability in each study as they will vary from sample to sample, they are not typically reported in EMA studies [75]. We did find 3 EMA studies that reported affect reliability scores, all of which presented very high between-person reliability scores for PA and NA, ranging from 0.96 to 1.00, and lower within-person reliability scores. This is to be expected because within-person variability is also generally smaller than variability between individuals [77]. In total, 2 of the aforementioned EMA studies calculated within-person reliability in the same way as we did and reported very similar coefficients, ranging from 0.50 to 0.71 [19,76].

Next, we attempted to identify careless responses, finding that <1% of assessments (291/33,624, 0.86%) had been answered carelessly, whereas 1 in every 6 participants in our sample was a potential careless responder. This step, though crucial for future analyses, is typically overlooked in EMA research. Potential careless responses were evaluated building on research in this area [21] (refer to the Statistical Analysis subsection in the Methods section).

When implementing the proposed method, we realized that focusing strictly on shifts in the relationship between psychometric antonyms would lead to discarding many plausible responses. For example, it is reasonable for a participant to provide similar ratings across items measuring affect if those ratings are near the middle of the scale. This issue was also raised in the study by Jaso et al [21]. Thus, we decided to take a more conservative approach, which led to simplifying the analysis from using 3 parameters to just using 2, namely, low duration and no variability in responses (ie, straight-line responding). Interestingly, these are the 2 most common parameters mentioned in the literature [20-22]. Although the overall number of careless responses in our study was negligible, it is still crucial to identify them because they tend to cluster around a small number of participants, which in turn allows us to spot careless responders.

Assessments flagged as being answered carelessly will be considered “missed” in future studies using these data. This, in turn, raises the important question as to how missing data can hamper data quality and interpretation of EMA results [66,70], especially as we have seen that missed assessments do not occur at random and may be influenced by multiple factors. Future analyses could address this issue using full information maximum likelihood [14], imputation techniques, or dynamic structural equation modeling [78].

We believe that the parameters we used to identify potential careless responders (ie, minimum number of completed assessments and individual reliability scores) can provide a stepping stone for future studies as these parameters are data driven and grounded in previous literature [20,21,24]. Determining the minimum number of assessments needed to have a stable estimate of the key measures’ mean and variability may be particularly relevant as it is common practice in EMA studies to consider an arbitrary threshold of minimum responses [14,66]. This method was also adapted from the work by Jaso et al [21]. However, they used rank-order correlation, whereas we used the ICC as we considered it to be more statistically sound given that it also accounts for agreement between the measures. We found that a minimum of 25 valid assessments (ie, 25/56, 45% compliance in the case of our study) were needed to have a stable estimate of the mean and variability of affect.

When looking at individual reliability of PA and NA scores, we observed that most participants had moderate to substantial reliability (ie, raw Cronbach α>0.60) for both PA and NA. However, some individuals had very low or even negative values, especially for NA. This mainly happened for two reasons: (1) they had low or negative interitem covariance that conformed a construct or (2) the responses to these items very seldom varied across the 56 assessments.

XSL•FO
RenderX

When the variability of the items is small, it becomes mathematically challenging to assess reliability as it is difficult to determine whether the responses truly covary (ie, have internal consistency) [19,75,77,79]. This phenomenon has been previously described in the literature [75]. Therefore, the internal consistency reliability of students with lower variability in their affective states may be penalized even if the responses are plausible. For this reason, we decided to set a low threshold of individual reliability for careless responders, flagging participants with a very low individual reliability (<0.11; "virtually no reliability") [62]. These potentially careless responders will not be excluded from the sample but flagged for sensitivity analyses in future studies.

Limitations

This study has several limitations that should be noted. First, although we conducted a pilot study to ensure that the software and recruitment flow were effective, technological issues such as bugs and connectivity problems could have affected participation and compliance, although reported incidences were minimal. Second, our EMA sample is not fully representative of the target population due to relatively low baseline survey participation and due to quota sampling. Nevertheless, we have detailed mental health information and basic sociodemographic data that can help mitigate this lack of representativeness and restore it for key variables through weighting procedures. Third, although we used sound methods to identify careless responders and responses, we cannot externally validate these findings as being truly indicative of careless response behavior. In any case, we will conduct sensitivity analyses in future studies to assess the impact of excluding these participants. Finally, it is important to consider that our conclusions are specific to the context of our sample and study design and may not be generalizable to other contexts.

Conclusions

The increase in the use of EMA in the mental health epidemiology field demands standardized ways to assess data quality to provide robust conclusions to advance the field. Our data are highly reliable to account for between-person differences in affect and fair to moderate study differences within each participant. Careless responses in our sample were minimal, and 1 out of 6 participants was found to be a potential careless responder (Textbox 1).

Textbox 1. Ecological momentary assessment design and quality evaluation recommendations for researchers.

Recommendations for researchers

- Ecological momentary assessment studies should collect basic information from nonresponders to restore representativity post hoc as it can be affected by gender, recent stressful events, and high mental health distress, among other things.
- Researchers should be mindful that compliance is particularly affected by stressful events that have an impact on cognition and by engagement in behaviors that draw attention away from participation, such as alcohol consumption.
- Careless responses can be identified post hoc when the time of response is <1 second per item or when all answers within an assessment are identical if these include psychological antonyms.
- Potential careless responders are those who have no individual reliability (raw Cronbach $\alpha<0.11$) or do not have a minimum number of valid assessments to have a stable estimate of the mean and variability of key measures.

Acknowledgments

Projects PI20/00006 and PI22/00107, funded by the Instituto de Salud Carlos III and cofunded by the European Union. Predoctoral contract Agència de Gestió d'Ajust Universitaris i de Recerca (PFIS) FI23/00004, funded by the Instituto de Salud Carlos III and cofunded by the European Union with the European Social Fund Plus, Miguel Servet (CP21/00078; Instituto de Salud Carlos III and European Social Fund Plus), and Agency for Management of University and Research Grants (AGAUR) 2021 SGR 00624. Consorcio Centro de Investigación Biomédica en Red (Biomedical Research Networking Center; CB06/02/0046).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AP-VD, PM, LB, JA, FA, PC, MG, GK, FHM, JAP, M Rebagliato, M Roca, TR-J, and GV were responsible for study conceptualization and design. AP-VD, PM, and GV wrote the first draft of the manuscript. LB, JA, FA, PC, MG, GK, FHM, JAP, M Rebagliato, M Roca, and TR-J provided critical revisions. All authors read and approved the submitted manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Flowchart of participation in the Promoting Mental Health Among University Students project.
[PNG File , 31 KB - jmir_v26i1e55712_app1.png]

Multimedia Appendix 2
Ecological momentary assessment measures.
[DOCX File , 23 KB - jmir_v26i1e55712_app2.docx]

Multimedia Appendix 3
Web-based survey measures.
[DOCX File , 21 KB - jmir_v26i1e55712_app3.docx]

Multimedia Appendix 4
Correlation coefficients for all items measuring affect using all data from all ecological momentary assessment participants (n=782).
[DOCX File , 25 KB - jmir_v26i1e55712_app4.docx]

Multimedia Appendix 5
Checklist for Reporting Ecological Momentary Assessment Studies.
[PDF File (Adobe PDF File), 22 KB - jmir_v26i1e55712_app5.pdf]

Multimedia Appendix 6
Associations between ecological momentary assessment study participation and mean compliance and sociodemographic characteristics, childhood and adolescence adverse experiences, recent stressful events, and mental disorders in the previous 12 months.
[DOCX File , 34 KB - jmir_v26i1e55712_app6.docx]

Multimedia Appendix 7
Percentages of missed assessments per day for all participants in the ecological momentary assessment study (n=782) and for each subsample.
[PNG File , 120 KB - jmir_v26i1e55712_app7.png]

Multimedia Appendix 8
Within-person reliability scores for positive and negative affect, compliance, and percentage of assessments responded carelessly for all ecological momentary assessment participants (n=782).
[DOCX File , 88 KB - jmir_v26i1e55712_app8.docx]

Multimedia Appendix 9
Overlap between the different parameters to identify careless responses and responders.
[PNG File , 50 KB - jmir_v26i1e55712_app9.png]

Multimedia Appendix 10
Scatterplot presenting the relationship between the percentage of identified careless responses and compliance for all participants in the ecological momentary assessment study (n=782).
[PNG File , 44 KB - jmir_v26i1e55712_app10.png]

References

1. Gordis L. Epidemiology E-Book. Amsterdam, The Netherlands: Elsevier Health Sciences; 2013.
2. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
3. Colombo D, Palacios AG, Alvarez JF, Patané A, Semonella M, Cipresso P, et al. Current state and future directions of technology-based ecological momentary assessments and interventions for major depressive disorder: protocol for a systematic review. Syst Rev 2018 Dec 13;7(1):233 [FREE Full text] [doi: 10.1186/s13643-018-0899-y] [Medline: 30545415]
4. Ferreri F, Bourla A, Mouchabac S, Karila L. e-addictology: an overview of new technologies for assessing and intervening in addictive behaviors. Front Psychiatry 2018 Mar 01;9:51 [FREE Full text] [doi: 10.3389/fpsyt.2018.00051] [Medline: 29545756]

XSL•FO
RenderX

5. Hallensleben N, Spangenberg L, Forkmann T, Rath D, Hegerl U, Kersting A, et al. Investigating the dynamics of suicidal ideation. Crisis 2018 Jan;39(1):65-69. [doi: 10.1027/0227-5910/a000464] [Medline: 28468557]
6. Schoevers RA, van Borkulo CD, Lamers F, Servaas M, Bastiaansen JA, Beekman ATF, et al. Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. Psychol Med 2021 Aug;51(11):1906-1915 [FREE Full text] [doi: 10.1017/S0033291720000689] [Medline: 32234092]
7. Sedano-Capdevila A, Porras-Segovia A, Bello HJ, Baca-García E, Barrigon ML. Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Curr Psychiatry Rep 2021 May 18;23(7):41. [doi: 10.1007/s11920-021-01255-7] [Medline: 34003405]
8. Chevance G, Perski O, Hekler EB. Innovative methods for observing and changing complex health behaviors: four propositions. Transl Behav Med 2021 Mar 16;11(2):676-685 [FREE Full text] [doi: 10.1093/tbm/ibaa026] [Medline: 32421196]
9. Csikszentmihalyi M, Larson R. Validity and reliability of the experience-sampling method. J Nerv Ment Dis 1987 Sep;175(9):526-536. [doi: 10.1097/00005053-198709000-00004] [Medline: 3655778]
10. Myin-Germeys I, Kasanova Z, Vaessen T, Vachon H, Kirtley O, Viechtbauer W, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun;17(2):123-132 [FREE Full text] [doi: 10.1002/wps.20513] [Medline: 29856567]
11. Balaskas A, Schueller SM, Cox AL, Doherty G. Ecological momentary interventions for mental health: a scoping review. PLoS One 2021 Mar 11;16(3):e0248152 [FREE Full text] [doi: 10.1371/journal.pone.0248152] [Medline: 33705457]
12. Bos FM, von Klipstein L, Emerencia AC, Veermans E, Verhage T, Snippe E, et al. A web-based application for personalized ecological momentary assessment in psychiatric care: user-centered development of the PETRA application. JMIR Ment Health 2022 Aug 09;9(8):e36430 [FREE Full text] [doi: 10.2196/36430] [Medline: 35943762]
13. Hall M, Scherner PV, Kreidel Y, Rubel JA. A systematic review of momentary assessment designs for mood and anxiety symptoms. Front Psychol 2021 May 17;12:642044 [FREE Full text] [doi: 10.3389/fpsyg.2021.642044] [Medline: 34079492]
14. Myin-Germeys I, Kuppens P. The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies. Leuven, Belgium: Center for Research on Experience Sampling and Ambulatory Methods Leuven (REAL); 2021.
15. Trull TJ, Ebner-Priemer UW. Ambulatory assessment in psychopathology research: a review of recommended reporting guidelines and current practices. J Abnorm Psychol 2020 Jan;129(1):56-63. [doi: 10.1037/abn0000473] [Medline: 31868388]
16. Vachon H, Viechtbauer W, Rintala A, Myin-Germeys I. Compliance and retention with the experience sampling method over the continuum of severe mental disorders: meta-analysis and recommendations. J Med Internet Res 2019 Dec 06;21(12):e14475 [FREE Full text] [doi: 10.2196/14475] [Medline: 31808748]
17. Thunnissen MR, Aan Het Rot M, van den Hoofdakker BJ, Nauta MH. Youth psychopathology in daily life: systematically reviewed characteristics and potentials of ecological momentary assessment applications. Child Psychiatry Hum Dev 2022 Dec;53(6):1129-1147 [FREE Full text] [doi: 10.1007/s10578-021-01177-8] [Medline: 34075490]
18. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017 Nov 01;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
19. Revelle W, Wilt J. Analyzing dynamic data: a tutorial. Pers Individ Differ 2019 Jan;136:38-51. [doi: 10.1016/j.paid.2017.08.020]
20. Welling J, Fischer RL, Schinkel-Bielefeld N. Is it possible to identify careless responses with post-hoc analysis in EMA studies? In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. 2021 Presented at: UMAP '21; June 21-25, 2021; Utrecht, The Netherlands. [doi: 10.1145/3450614.3462237]
21. Jaso BA, Kraus NI, Heller AS. Identification of careless responding in ecological momentary assessment research: from posthoc analyses to real-time data monitoring. Psychol Methods 2022 Dec;27(6):958-981. [doi: 10.1037/met0000312] [Medline: 34582244]
22. Meade AW, Craig SB. Identifying careless responses in survey data. Psychol Methods 2012 Sep;17(3):437-455. [doi: 10.1037/a0028085] [Medline: 22506584]
23. Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. Detecting and deterring insufficient effort responding to surveys. J Bus Psychol 2011 May 31;27:99-114. [doi: 10.1007/s10869-011-9231-8]
24. Eisele G, Vachon H, Lafit G, Kuppens P, Houben M, Myin-Germeys I, et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment 2022 Mar;29(2):136-151. [doi: 10.1177/1073191120957102] [Medline: 32909448]
25. Alonso J, Mortier P, Auerbach RP, Bruffaerts R, Vilagut G, Cuijpers P, et al. Severe role impairment associated with mental disorders: results of the WHO world mental health surveys international college student project. Depress Anxiety 2018 Sep;35(9):802-814 [FREE Full text] [doi: 10.1002/da.22778] [Medline: 29847006]
26. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. WHO world mental health surveys international college student project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]

XSL•FO
RenderX

27. Ballester L, Alayo I, Vilagut G, Almenara J, Cebrià AI, Echeburúa E, et al. Mental disorders in Spanish university students: prevalence, age-of-onset, severe role impairment and mental health treatment. J Affect Disord 2020 Aug 01;273:604-613. [doi: 10.1016/j.jad.2020.04.050] [Medline: 32560960]
28. Kleiman EM, Yeager AL, Grove JL, Kellerman JK, Kim JS. Real-time mental health impact of the COVID-19 pandemic on college students: ecological momentary assessment study. JMIR Ment Health 2020 Dec 15;7(12):e24815 [FREE Full text] [doi: 10.2196/24815] [Medline: 33207308]
29. Bai Y, Copeland WE, Burns R, Nardone H, Devadanam V, Rettew J, et al. Ecological momentary assessment of physical activity and wellness behaviors in college students throughout a school year: longitudinal naturalistic study. JMIR Public Health Surveill 2022 Jan 04;8(1):e25375 [FREE Full text] [doi: 10.2196/25375] [Medline: 34982721]
30. Bennett BL, Whisenhunt BL, Hudson DL, Wagner AF, Latner JD, Stefano EC, et al. Examining the impact of social media on mood and body dissatisfaction using ecological momentary assessment. J Am Coll Health 2020 Jul;68(5):502-508. [doi: 10.1080/07448481.2019.1583236] [Medline: 30908131]
31. Portillo-Van Diest A, Ballester Coma L, Mortier P, Vilagut G, Amigo F, Puértolas Gracia B, et al. Experience sampling methods for the personalised prediction of mental health problems in Spanish university students: protocol for a survey-based observational study within the PROMES-U project. BMJ Open 2023 Jul 14;13(7):e072641 [FREE Full text] [doi: 10.1136/bmjopen-2023-072641] [Medline: 37451741]
32. Cuijpers P, Auerbach RP, Benjet C, Bruffaerts R, Ebert D, Karyotaki E, et al. The World Health Organization World Mental Health International College Student initiative: an overview. Int J Methods Psychiatr Res 2019 Jun;28(2):e1761 [FREE Full text] [doi: 10.1002/mpr.1761] [Medline: 30614123]
33. Observational study within the PROMES-U (Promoting Mental Health Among University Students) project. OSF Registries. 2023. URL: https://osf.io/p7csq [accessed 2024-11-20]
34. Posner K, Brent D, Lucas C, Gould M, Stanley B, Brown G, et al. Columbia-Suicide Severity Rating Scale (C-SSRS). The Research Foundation for Mental Hygiene. 2008. URL: https://cssrs.columbia.edu/wp-content/uploads/C-SSRS_Pediatric-SLC_11.14.16.pdf [accessed 2022-07-21]
35. Ballester L, Alayo I, Vilagut G, Almenara J, Cebrià AI, Echeburúa E, et al. Validation of an online version of the alcohol use disorders identification test (AUDIT) for alcohol screening in Spanish university students. Int J Environ Res Public Health 2021 May 14;18(10):5213 [FREE Full text] [doi: 10.3390/ijerph18105213] [Medline: 34068945]
36. Kessler RC, Calabrese JR, Farley PA, Gruber MJ, Jewell MA, Katon W, et al. Composite international diagnostic interview screening scales for DSM-IV anxiety and mood disorders. Psychol Med 2013 Aug;43(8):1625-1637. [doi: 10.1017/S0033291712002334] [Medline: 23075829]
37. van Roekel E, Keijsers L, Chung JM. A review of current ambulatory assessment studies in adolescent samples and practical recommendations. J Res Adolesc 2019 Sep;29(3):560-577 [FREE Full text] [doi: 10.1111/jora.12471] [Medline: 31573762]
38. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
39. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
40. Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. Health Qual Life Outcomes 2009 Feb 19;7:15 [FREE Full text] [doi: 10.1186/1477-7525-7-15] [Medline: 19228398]
41. Kirtley O, Eisele G, Kunkels YK, Hiekkaranta AP, Van Heck L, Pihlajamäki M, et al. The Experience Sampling Method (ESM) item repository. Open Science Framework Home. 2018 Dec 5. URL: https://osf.io/kg376/ [accessed 2022-07-20]
42. Russell JA. A circumplex model of affect. J Pers Soc Psychol 1980;39(6):1161-1178. [doi: 10.1037/h0077714]
43. Shah N, Cader M, Andrews B, McCabe R, Stewart-Brown SL. Short Warwick-Edinburgh mental well-being scale (SWEMWBS): performance in a clinical sample in relation to PHQ-9 and GAD-7. Health Qual Life Outcomes 2021 Nov 24;19(1):260 [FREE Full text] [doi: 10.1186/s12955-021-01882-x] [Medline: 34819104]
44. Dalton ED, Hammen CL. Independent and relative effects of stress, depressive symptoms, and affect on college students' daily health behaviors. J Behav Med 2018 Dec;41(6):863-874. [doi: 10.1007/s10865-018-9945-4] [Medline: 29926314]
45. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 01;35(2):287-302 [FREE Full text] [doi: 10.5665/sleep.1642] [Medline: 22294820]
46. Baker MR, Nguyen-Feng VN, Nilakanta H, Frazier PA. Childhood maltreatment predicts daily stressor exposure in college students but not perceived stress or stress reactivity. J Couns Psychol 2020 Jan;67(1):79-89. [doi: 10.1037/cou0000359] [Medline: 31144850]
47. Gratch I, Choo TH, Galfalvy H, Keilp JG, Itzhaky L, Mann JJ, et al. Detecting suicidal thoughts: the power of ecological momentary assessment. Depress Anxiety 2021 Jan;38(1):8-16. [doi: 10.1002/da.23043] [Medline: 32442349]
48. Cuijpers P, Auerbach RP, Benjet C, Bruffaerts R, Ebert D, Karyotaki E, et al. Introduction to the special issue: the WHO World Mental Health International College Student (WMH-ICS) initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1762 [FREE Full text] [doi: 10.1002/mpr.1762] [Medline: 30623516]

XSL•FO
RenderX

49. Dube SR, Anda RF, Felitti VJ, Chapman DP, Williamson DF, Giles WH. Childhood abuse, household dysfunction, and the risk of attempted suicide throughout the life span: findings from the adverse childhood experiences study. JAMA 2001 Dec 26;286(24):3089-3096. [doi: 10.1001/jama.286.24.3089] [Medline: 11754674]
50. Bernstein DP, Stein JA, Newcomb MD, Walker E, Pogge D, Ahluvalia T, et al. Development and validation of a brief screening version of the Childhood Trauma Questionnaire. Child Abuse Negl 2003 Feb;27(2):169-190. [doi: 10.1016/s0145-2134(02)00541-0] [Medline: 12615092]
51. Ursano RJ, Heeringa SG, Stein MB, Jain S, Raman R, Sun X, et al. Prevalence and correlates of suicidal behavior among new soldiers in the U.S. Army: results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Depress Anxiety 2015 Jan;32(1):3-12 [FREE Full text] [doi: 10.1002/da.22317] [Medline: 25338964]
52. Kann L, McManus T, Harris WA, Shanklin SL, Flint KH, Queen B, et al. Youth risk behavior surveillance - United States, 2017. MMWR Surveill Summ 2018 Jun 15;67(8):1-114 [FREE Full text] [doi: 10.15585/mmwr.ss6708a1] [Medline: 29902162]
53. Lipson SK, Zhou S, Abelson S, Heinze J, Jirsa M, Morigney J, et al. Trends in college student mental health and help-seeking by race/ethnicity: findings from the national healthy minds study, 2013-2021. J Affect Disord 2022 Jun 01;306:138-147 [FREE Full text] [doi: 10.1016/j.jad.2022.03.038] [Medline: 35307411]
54. Ursano RJ, Colpe LJ, Heeringa SG, Kessler RC, Schoenbaum M, Stein MB. The Army study to assess risk and resilience in servicemembers (Army STARRS). Psychiatry 2014;77(2):107-119 [FREE Full text] [doi: 10.1521/psyc.2014.77.2.107] [Medline: 24865195]
55. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction 1993 Jun;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
56. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011 Dec;168(12):1266-1277 [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
57. Estudiantes matriculados en Grado y Ciclo. Ministerio de Ciencia, Innovación y Universidades. URL: https://estadisticas.universidades.gob.es/dynPx/inebase/index.htm?type=pcaxis&path=/Universitaria/Alumnado/EEU_2024/GradoCiclo/Matriculados/&file=pcaxis&l=s0 [accessed 2024-11-14]
58. Star SA. American association for public opinion research. Public Opin Q 1980;44(3):420. [doi: 10.1093/poq/44.3.420]
59. Revelle W. Package ‘psych’: procedures for psychological, psychometric, and personality research. The Comprehensive R Archive Network. 2023 Mar 14. URL: https://personality-project.org/r/psych/HowTo/psych_manual.pdf [accessed 2024-11-19]
60. Shrout PE, Lane SP. Psychometrics. In: Mehl MR, Conner TS, editors. Handbook of Research Methods for Studying Daily Life. New York, NY: Guilford Press; 2012.
61. Revelle W, Condon DM. Reliability from α to ω: a tutorial. Psychol Assess 2019 Dec;31(12):1395-1411. [doi: 10.1037/pas0000754] [Medline: 31380696]
62. Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 1998 Sep;7(3):301-317. [doi: 10.1177/096228029800700306] [Medline: 9803527]
63. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
64. Student statistics. Ministry of Science, Innovation and Universities. URL: https://www.universidades.gob.es/estadistica-de-estudiantes/ [accessed 2024-11-13]
65. Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol 2007 Sep;17(9):643-653. [doi: 10.1016/j.annepidem.2007.03.013] [Medline: 17553702]
66. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023 May 09;19:107-131 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]
67. Hamilton IJ. Understanding grief and bereavement. Br J Gen Pract 2016 Oct;66(651):523 [FREE Full text] [doi: 10.3399/bjgp16X687325] [Medline: 27688503]
68. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
69. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017 Mar 15;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]
70. Kivelä L, van der Does WA, Riese H, Antypa N. Don't miss the moment: a systematic review of ecological momentary assessment in suicide research. Front Digit Health 2022 May 6;4:876595 [FREE Full text] [doi: 10.3389/fdgth.2022.876595] [Medline: 35601888]
71. Devries KM, Mak JY, Child JC, Falder G, Bacchus LJ, Astbury J, et al. Childhood sexual abuse and suicidal behavior: a meta-analysis. Pediatrics 2014 May;133(5):e1331-e1344. [doi: 10.1542/peds.2013-2166] [Medline: 24733879]

XSL•FO
RenderX

72. Mortier P, Alonso J, Auerbach RP, Bantjes J, Benjet C, Bruffaerts R, et al. Childhood adversities and suicidal thoughts and behaviors among first-year college students: results from the WMH-ICS initiative. Soc Psychiatry Psychiatr Epidemiol 2022 Aug;57(8):1591-1601 [FREE Full text] [doi: 10.1007/s00127-021-02151-4] [Medline: 34424350]
73. Whiffen VE, Macintosh HB. Mediators of the link between childhood sexual abuse and emotional distress: a critical review. Trauma Violence Abuse 2005 Jan;6(1):24-39. [doi: 10.1177/1524838004272543] [Medline: 15574671]
74. Messiah A, Grondin O, Encrenaz G. Factors associated with missing data in an experience sampling investigation of substance use determinants. Drug Alcohol Depend 2011 Apr 01;114(2-3):153-158. [doi: 10.1016/j.drugalcdep.2010.09.016] [Medline: 21075563]
75. Brose A, Schmiedek F, Gerstorf D, Voelkle MC. The measurement of within-person affect variation. Emotion 2020 Jun;20(4):677-699. [doi: 10.1037/emo0000583] [Medline: 31008618]
76. Haney AM, Fleming MN, Wycoff AM, Griffin SA, Trull TJ. Measuring affect in daily life: a multilevel psychometric evaluation of the PANAS-X across four ecological momentary assessment samples. Psychol Assess 2023 Jun;35(6):469-483 [FREE Full text] [doi: 10.1037/pas0001231] [Medline: 36931821]
77. Fuller-Tyszkiewicz M, Hartley-Clark L, Cummins RA, Tomyn AJ, Weinberg MK, Richardson B. Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. Psychol Assess 2017 Sep;29(9):1120-1128. [doi: 10.1037/pas0000411] [Medline: 27819437]
78. Asparouhov T, Hamaker EL, Muthén B. Dynamic structural equation models. Struct Equ Model Multidiscip J 2017 Dec 27;25(3):359-388. [doi: 10.1080/10705511.2017.1406803]
79. Nezlek JB. A practical guide to understanding reliability in studies of within-person variability. J Res Personal 2017 Aug;69:149-155 [FREE Full text] [doi: 10.1016/j.jrp.2016.06.020]

Abbreviations

EMA: ecological momentary assessment
ESM: experience sampling method
ICC: intraclass correlation coefficient
NA: negative affect
OR: odds ratio
PA: positive affect
PHQ-9: 9-item Patient Health Questionnaire
PROMES-U: Promoting Mental Health Among University Students
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 21.12.23; peer-reviewed by Y Motreff, K Diviak, S Lau, JS Wooldridge, D Rath; comments to author 29.05.24; revised version received 23.07.24; accepted 04.10.24; published 10.12.24.

Please cite as:
Portillo-Van Diest A, Mortier P, Ballester L, Amigo F, Carrasco P, Falcó R, Gili M, Kiekens G, H Machancoses F, Piqueras JA, Rebagliato M, Roca M, Rodríguez-Jiménez T, Alonso J, Vilagut G
Ecological Momentary Assessment of Mental Health Problems Among University Students: Data Quality Evaluation Study
J Med Internet Res 2024;26:e55712
URL: https://www.jmir.org/2024/1/e55712
doi:10.2196/55712
PMID:39657180

©Ana Portillo-Van Diest, Philippe Mortier, Laura Ballester, Franco Amigo, Paula Carrasco, Raquel Falcó, Margalida Gili, Glenn Kiekens, Francisco H Machancoses, Jose A Piqueras, Marisa Rebagliato, Miquel Roca, Tíscar Rodríguez-Jiménez, Jordi Alonso, Gemma Vilagut. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Extensible Evaluation Framework Applied to Clinical Text Deidentification Natural Language Processing Tools: Multisystem and Multicorpus Study

Paul M Heider[1], PhD; Stéphane M Meystre[2], MD, PhD

[1]Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, United States

[2]Institute of Digital Technologies for Personalised Healthcare (MeDiTech), University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland

Corresponding Author:
Paul M Heider, PhD
Biomedical Informatics Center
Medical University of South Carolina
22 WestEdge Street, Suite 200
Charleston, SC, 29403
United States
Phone: 1 843 792 3385
Email: heiderp@musc.edu

Abstract

Background: Clinical natural language processing (NLP) researchers need access to directly comparable evaluation results for applications such as text deidentification across a range of corpus types and the means to easily test new systems or corpora within the same framework. Current systems, reported metrics, and the personally identifiable information (PII) categories evaluated are not easily comparable.

Objective: This study presents an open-source and extensible end-to-end framework for comparing clinical NLP system performance across corpora even when the annotation categories do not align.

Methods: As a use case for this framework, we use 6 off-the-shelf text deidentification systems (ie, CliniDeID, deid from PhysioNet, MITRE Identity Scrubber Toolkit [MIST], NeuroNER, National Library of Medicine [NLM] Scrubber, and Philter) across 3 standard clinical text corpora for the task (2 of which are publicly available) and 1 private corpus (all in English), with annotation categories that are not directly analogous. The framework is built on shell scripts that can be extended to include new systems, corpora, and performance metrics. We present this open tool, multiple means for aligning PII categories during evaluation, and our initial timing and performance metric findings. Code for running this framework with all settings needed to run all pairs are available via Codeberg and GitHub.

Results: From this case study, we found large differences in processing speed between systems. The fastest system (ie, MIST) processed an average of 24.57 (SD 26.23) notes per second, while the slowest (ie, CliniDeID) processed an average of 1.00 notes per second. No system uniformly outperformed the others at identifying PII across corpora and categories. Instead, a rich tapestry of performance trade-offs emerged for PII categories. CliniDeID and Philter prioritize recall over precision (with an average recall 6.9 and 11.2 points higher, respectively, for partially matching spans of text matching any PII category), while the other 4 systems consistently have higher precision (with MIST's precision scoring 20.2 points higher, NLM Scrubber scoring 4.4 points higher, NeuroNER scoring 7.2 points higher, and deid scoring 17.1 points higher). The macroaverage recall across corpora for identifying names, one of the more sensitive PII categories, included deid (48.8%) and MIST (66.9%) at the low end and NeuroNER (84.1%), NLM Scrubber (88.1%), and CliniDeID (95.9%) at the high end. A variety of metrics across categories and corpora are reported with a wider variety (eg, F_2-score) available via the tool.

Conclusions: NLP systems in general and deidentification systems and corpora in our use case tend to be evaluated in stand-alone research articles that only include a limited set of comparators. We hold that a single evaluation pipeline across multiple systems and corpora allows for more nuanced comparisons. Our open pipeline should reduce barriers to evaluation and system advancement.

(J Med Internet Res 2024;26:e55676) doi:10.2196/55676

XSL•FO
RenderX

KEYWORDS

natural language processing; evaluation methodology; deidentification; privacy protection; de-identification; secondary use; patient privacy

Introduction

Background

An ironclad pillar of clinical data reuse is the proper protection of protected health information (PHI). Deidentification is the process of tagging and removing personally identifiable information (PII) to prevent incidental privacy breaches. Unfortunately, manual deidentification is an expensive and error-prone process [1,2], and automated deidentification remains an unsolved challenge [3,4]. Since the first published automated deidentification system [5], a variety of systems using a range of technologies have been released. In tandem, a series of competitions have been organized around shared corpora annotated with PII to further encourage the development of deidentification systems [6-9]. Researchers publishing about a new system tend to release comparative performance metrics against 1 or 2 publicly available systems using 1 or 2 corpora [3,10-12]. These stand-alone research articles that only include a limited set of comparator systems and corpora cannot always even be directly compared to create a single meta-analysis because of the differences in annotation categories evaluated and reported on or how exactly matches are aligned and scored. Replication of the process to confirm results and extension of the process to evaluate new systems or corpora have been stymied by a combination of lack of documented evaluation code, closed evaluation tools, and incompatible PII categories, among other reasons.

As such, a single evaluation pipeline across multiple systems and corpora allows for more nuanced comparisons between systems. A single pipeline allows researchers (and staff scientists responsible for evaluating systems before deployment) to consistently and reproducibly generate scores for a range of systems across a range of corpora using the exact same methods. Researchers with preferences to evaluate deidentification systems at the character level versus token level versus PII mention can run each evaluation in turn. Similarly, researchers can differentiate or collapse PII categories as desired across all systems and corpora at once to get a clear picture of how each system and corpus, respectively, treats different categories. Not all categories of PII are equally sensitive [3]. Patient names are more revealing than provider names, which are, in turn, more revealing than hospital names.

Objectives

To that end, we evaluated PII extraction performance at multiple levels of granularity because these systems should not be judged on a single summary performance metric. For instance, this pipeline provides an easy means for surfacing the false negative rate or recall (also called sensitivity) for patient names as distinct from provider names (or other names). We can also easily compare performance when differences between PII categories are important to maintain, as opposed to when all PII categories are treated interchangeably.

As a case study, to help us understand the performance trade-offs of available deidentification systems that are critical to clinical data reuse and natural language processing (NLP) and to foster building larger repositories of directly comparable evaluation results, we developed a reusable and extensible pipeline for evaluating 6 off-the-shelf deidentification systems across 2 freely available corpora, 1 previously available corpus, and 1 private corpus. The systems and corpora all use English, although the pipeline is language agnostic. As none of the systems or corpora use exactly the same annotation schema, we provided mappings to allow approximately equitable performance metrics across all components. Any deidentification system programmatically runnable from the command line can be added to the set. Similarly, new corpora can be added to the evaluation process with minimal constraints on their format or annotation categories. Furthermore, the evaluation tool used in the pipeline allows for analysis at configurable levels of annotation category granularity and with multiple text annotation matching styles [13,14]. We focused on off-the-shelf systems for 3 primary reasons. First, off-the-shelf systems have the lowest barrier to entry. Not all potential users have the skills, capacity, or annotated corpora available to retrain a model. If a deidentification system developer considers their tool to be unusable without retraining, then that limitation should be made explicit, which brings us to the second reason. Undertaking the controlled experiments required to determine when a retrained system has met a reliable and safe performance threshold requires a test harness exactly like the one we propose here. Third, the potential variants for retraining any given system to optimize its performance for a local site constitutes its own large undertaking and falls outside the scope of this research. Furthermore, comparing the retrained variants to determine optimal performance is best organized using a test harness as described in this study.

Thus, we hold that a single evaluation pipeline across multiple systems and corpora allows for the nuanced comparisons required for safe deployment of NLP systems, in general, and deidentification systems, in particular. We document the use and extensibility of such a tool with deidentification as a use case. In summary, we found an order of magnitude difference between the fastest and slowest systems in terms of processing speed. We found that none of the 6 systems consistently outperformed the others across corpora and PII tag categories. A nuanced comparison of the top performers under slightly different conditions would be much more difficult without a cohesive framework like the one we describe.

Methods

Ethical Considerations

This study was assessed by the Medical University of South Carolina (MUSC) Institutional Review Board for Human Research (IRB) and officially considered Not Human Research. It was therefore not subject to oversight by the MUSC IRB,

XSL•FO
RenderX

since it met the criteria set forth by the Code of Federal Regulations (45CFR46): (1) the data were not collected specifically for the currently proposed research project through an interaction or intervention with living individuals and (2) investigators including collaborators on the proposed research cannot readily ascertain the identity of the individuals to whom the coded private information of specimens pertains.

The other three corpora used in this study (ie, 2006 i2b2 shared task, 2014 i2b2 and University of Texas Health Science Center at Houston (i2b2/UTHealth), and 2016 Centers of Excellence in Genomic Science Neuropsychiatric Genome-Scale and RDOC Individualized Domains [CEGS N-GRID] shared task) were publicly available and already deidentified, and therefore not subject to IRB approval requirements.

PII Categories

In the United States, HIPAA (Health Insurance Portability and Accountability Act) is the primary legal mandate guiding and governing data privacy and security provisions within the health care domain [15]. Other regions of the world have enacted similar privacy laws such as the General Data Protection Regulation in the European Union [16], although they are not all specific to the health care industry, as evidenced by the General Data Protection Regulation. The HIPAA privacy rule specifies sensitive classes of PII that should be removed for the data to be considered deidentified. In this study, we focus on the subset of HIPAA's 18 categories of PII relevant to unstructured clinical notes, as all our data sets were created in the United States under the jurisdiction of HIPAA. We have listed a curated and organized set of these categories in Figure 1 in the 4 left-most columns. The first column indicates the original HIPAA Safe Harbor category names. The following 3 columns are curated classes to help group or simplify the categories. We call these the tier 0 category (which represents the general class of PII), the tier 1 categories (which represent 7 high-level sets of categories grouped by semantic domain and common textual realizations), and tier 2 categories (which represent the finest-grained division of categories approximating the original HIPAA categories). The tier 2 categories include several practical and functional extensions of the strict HIPAA categories that have been treated as PII by deidentification researchers in the clinical domain. For instance, HIPAA considers ages >89 years to be PII but not the not younger ages, while some researchers consider any age to be PII. Additional columns in this figure represent categories annotated in deidentification corpora (columns 5-7) and categories flagged by deidentification systems (columns 8-13). Rows across the figure indicate rough equivalency of categories between tiers, corpora, or systems. Specific details of corpus and system categories are covered in the next two subsections: *Deidentification Shared Tasks* and *Corpora and Deidentification Systems*. Perfect category synchrony between corpora and annotations is impossible. We have included a sample sentence in Figure 2, cross-annotated according to the specifics of each category schema to help highlight the local variation between corpora and deidentification systems.

Figure 1. Corpus and system category alignment. Each column represents the categories relevant to HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor Guidelines (column 1), a specific analysis tier (columns 2-4), corpus (columns 5-7), or system (columns 8-13). Each row represents approximate equivalency between categories. An asterisk by a category label (eg, "Relative*") indicates that the category is attested in the system documentation but not labeled by the system in any corpus. BIO ID: Biometric Identifier; i2b2: Informatics for Integrating Biology and the Bedside; MIST: MITRE Identity Scrubber Toolkit; MUSC: Medical University of South Carolina; NLM: National Library of Medicine; OtherGeo: other geographic subdivision; PII: Personally Identifiable Information; SSN: Social Security Number.

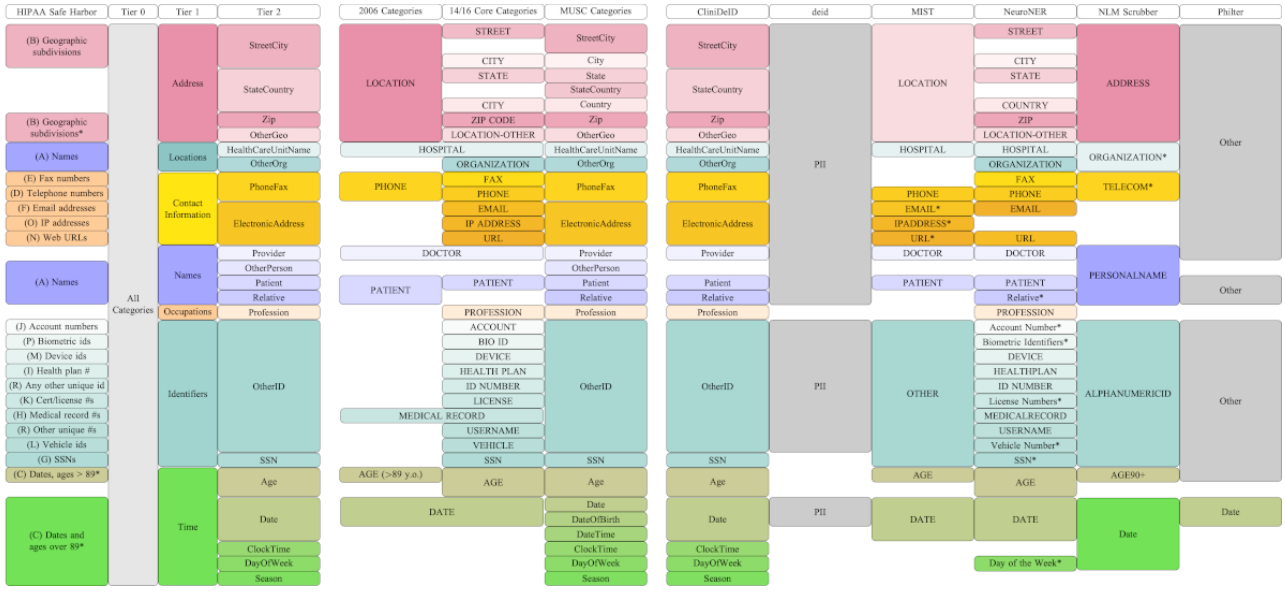

XSL•FO
RenderX

Figure 2. Personally identifiable information terms in sample sentence mapped to categories across all corpora and deidentification systems. Geo. subdiv.: geographic subdivision; HCPName: Health Care Provider Name; HIPAA: Health Insurance Portability and Accountability Act; i2b2: Informatics for Integrating Biology and the Bedside; MIST: MITRE Identity Scrubber Toolkit; MUSC: Medical University of South Carolina; NLM: National Library of Medicine; PTName: patient name.

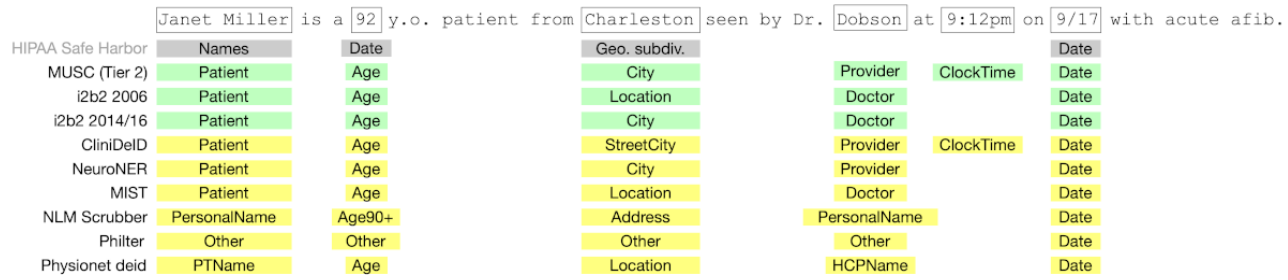

Deidentification Shared Tasks and Corpora

Several shared tasks organized around the goal of evaluating deidentification systems have been organized within the clinical domain. The 2006 Informatics for Integrating Biology and the Bedside (i2b2) shared task was the first such task and focused on a small set of PII categories present in unstructured clinical notes written in English [6]. A similar shared task was organized as part of the 2012 NII Test Beds and Community for Information Access Research Medical Natural Language Processing task for fabricated but realistic medical reports written in Japanese [7]. Two more English language tasks were organized with an array of PII categories that were more representative of the full list of HIPAA categories as part of the 2014 i2b2 and UTHealth shared task [8] and 2016 CEGS N-GRID shared task [9].

We used each of the 3 English language deidentification shared task corpora for our study as they have been used as the standard reference corpora in the domain. The 2006 and 2014 corpora are publicly available with an appropriate data use agreement. The 2016 corpus was publicly available but has since been removed from circulation due to privacy concerns. The fourth corpus, also in English, is called the "Medical University of South Carolina (MUSC)" corpus in this study as it was developed at MUSC by the authors. It is not publicly available as its PII has only been annotated and not redacted or resynthesized to prevent the release of PHI.

We refer to the oldest corpus in our study as "2006" as it was used for the 2006 i2b2 shared task [6]. It consists of 889 discharge summaries from Partners HealthCare. This corpus has the most reduced set of annotation categories with 8 distinct categories: location, hospital, phone, doctor, patient, medical record, age (>89 years), and date, as shown in Figure 1. Of note, the tier 1 category that we call "Address" maps to the 2006 category "Location." "Medical Records" are the only recorded tier 1 "Identifiers." "Age" is confined by the strict HIPAA notion to those >89 years, in contrast to the other 3 corpora that treat all ages as PII.

Internally, we aligned the annotation categories for the 2014 i2b2 and UTHealth [8] and the 2016 CEGS N-GRID shared tasks [9] on deidentification, which we refer to as "2014" and "2016," respectively. In contrast with 2006, the tier 1 Address and Identifiers categories for 2014 and 2016 are split into fine-grained categories. Overall, there are 28 distinct categories, as shown in Figure 1. The 2014 corpus consists of 1304 discharge summaries and correspondences between providers from Partners HealthCare. The 2016 corpus consists of 1000 psychiatric notes from Partners HealthCare.

The MUSC corpus consists of 728 notes split across 8 note types: consults, discharge summaries, history and physicals, nursing, pathology and cytology, patient instructions, plan of care, and progress notes. Similar to the 2014 and 2016 corpora, the MUSC corpus includes a wide range of tier 1 "Identifiers" but groups all of them except social security numbers into 1 class: "Other ID." Due to a change in the annotation guidelines over the course of the corpus annotation, some "Street" annotations and "City" annotations have been merged into a single "StreetCity" annotation. Similarly, some "State" annotations and "Country" annotations have been merged into "StateCountry" annotations.

All 4 corpora have been divided into train and test splits. None of the systems have used the official test splits for training. Therefore, we report all performance metrics with respect to the official test split in the main body of this paper. Multimedia Appendix 1 contains results for both train and test splits and Multimedia Appendix 2 [2,11,17-32] contains de-identification system details. All reported timing results include both train and test splits.

Deidentification Systems

Our initial set of systems was constrained by an implementation science framing of the problem. As researchers invested in reproducible science, we need to understand the best tools for facilitating ethical data sharing, but our primary research area may not be related to deidentification. We need a resource for evaluating scalable, off-the-shelf systems that do not require additional model training or fine-tuning. Concretely, this framing restricts us to freely available systems that can be run programmatically on a local server without any requirements to annotate a local data set. A total of 6 systems meet these requirements: CliniDeID (version 1.6.1) [17,18], deid by Massachusetts Institute of Technology (version 1.1, now available via PhysioNet) [2,19], MITRE Identity Scrubber Toolkit (MIST; version 2.0.4) [20-22], NeuroNER (commit 3817fea on GitHub) [23,24], National Library of Medicine Scrubber (version 19.0403L Linux ×86 64) [25-27], and Philter (commit 780da99 on GitHub) [11,28]. All 6 of these systems run on English clinical notes, read in files from disc, write annotated files indicating identified PII mentions to disc, can be run from the command line on a Linux server, and have been

XSL•FO
RenderX

run using the latest available version. System reporting will be in alphabetical order. Refer to Multimedia Appendix 2 for more details on each system's configuration and use.

Metrics

We evaluated systems in terms of both timing and binary classification performance metrics. For timing purposes (reported in the *Timing Results* subsection), each system processed 1 corpus at a time on the same Red Hat Linux server. This evaluation pipeline was the only program beyond routine background processes running on the machine. We used the real (ie, wall clock) time generated by the command line tool time. For performance metrics (ie, encompassing the 3 *Performance Results* subsections in the *Results* section), we used the Evaluation Tool for Unstructured Data and Extractions (ETUDE) to score each system by corpus pairing. ETUDE uses definitions from a configuration file for each of the reference and system outputs to determine how each "native" annotation category is represented in the corpus. These native categories are then mapped in the same configuration file to "scoring value" categories. Drawing from an example depicted in Figure 1, the 2016 native categories of "Phone" and "Fax" are mapped in the 2016 configuration file to the "PhoneFax" scoring value category for tier 1 entries. The configuration files allow us to separate the logic for annotation extraction from annotation alignment matching.

ETUDE also provides several alignment matching algorithms for determining which set of annotations are considered a match between the reference and system outputs. We focused on 3 of these matching algorithms for this study: "exact," "partial," and "fully contained." Exact matching requires that the character offsets of 2 annotations be the same to count as a match. Partial matching only requires that any part of the 2 annotations overlap to be considered a match. Finally, fully contained matching requires the system output annotation to at least cover the entirety of the reference annotation to count as a match. The system output annotation can include more text before or after the extent of the reference annotation but not less. The intuition relevant to deidentification is that if the system PII output annotation fully contains the reference PII annotation, then we know that no PII is leaking, as is potentially possible for a partial match of annotations.

The final evaluation feature we used in ETUDE was the ability to collapse all patterns into a single category for evaluation purposes. Thus, for some of our evaluations, we tracked both annotation and category. For other evaluations, we only tracked the annotations and ignored the differences in annotation category.

In the end, ETUDE generated counts for true positives, false positives, and false negatives. From these counts, we calculated precision, recall, and F_1-score values (ie, the harmonic mean of precision and recall). Given the sensitive nature of deidentification, we give primacy to recall in our reporting in the main body of this paper but include all 3 values in Multimedia Appendix 1. The F_2-score, which gives more weight to recall than precision, can also be generated by ETUDE.

Results

Processing Pipeline

Expanding on the work of our previous comparison of 3 deidentification systems [33], we developed a larger and more flexible pipeline for using a set of off-the-shelf deidentification systems to process a set of corpora and then scoring all system output at a range of PII category granularities, as shown in the block diagram in Figure 3. The core of this pipeline rests on command line shell scripts with configurable custom functions for processing any given corpus with a given deidentification system. Specifically, the core shell script uses environment variables to set input and output folders for each corpus, running folders and Python environments for each deidentification system, and flags for which set of corpora and systems should run on any given instantiation. The processed output for a system and corpus combination is written to disc in a given folder, allowing for repeated evaluation loops by ETUDE [13,14], a freely available open-source tool developed by the first author. Each evaluation loop uses different configuration settings to highlight different tiers of categories, different alignments of categories between reference and system, and different annotation alignment algorithms. Adding a new deidentification system requires adding a few simple shell commands to preprocess files (as expected by the system), to run the system from the command line with all parameters fully specified, and to postprocess files (if they are not in a format already supported by ETUDE). A new corpus or deidentification system may require creating a new annotation schema mapping file, if the schema is not already covered by those shown in Figure 1. As ETUDE, all shared task corpora, and all deidentification systems are already publicly available, we also released our shell scripts and R-based evaluation scripts on Codeberg and GitHub to allow for near-complete reproducibility of this study [34,35]. System-specific settings, in terms of explicit configuration files or command line flags and settings, are also included in these repositories.

XSL•FO
RenderX

Figure 3. A block diagram showing pseudocode for the logic of running corpus and deidentification system pairs along with the major processes required to be defined for adding new systems.

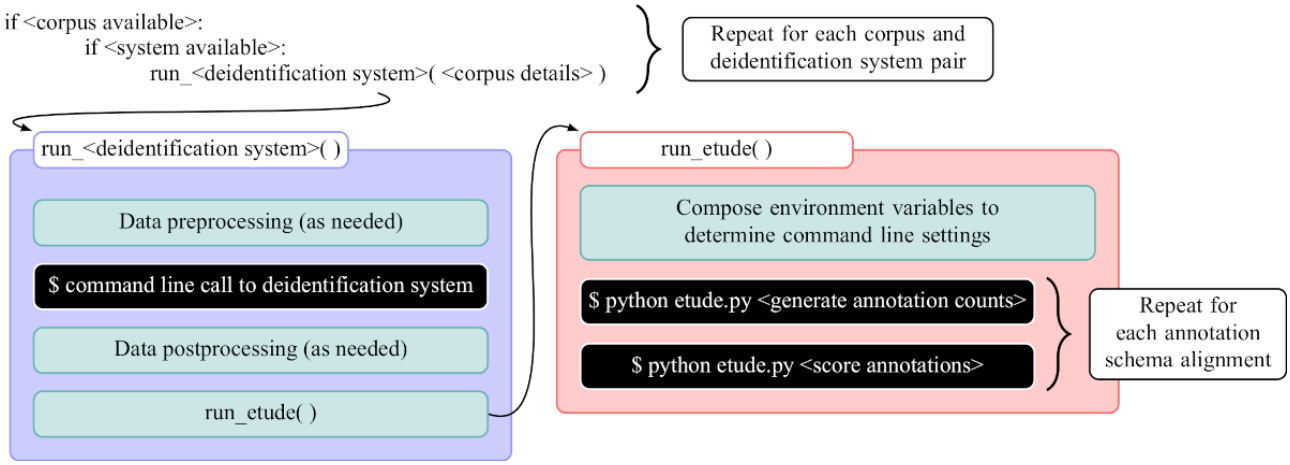

Timing Results

The 8 timed trials for each system (ie, 4 corpora with 1 train and 1 test split) are summarized in Table 1. We report both the seconds per note (for which a lower value is better) and notes per second (for which a higher value is better). MIST is the fastest with an average of 24.57 (SD 26.23) notes per second and is an order of magnitude faster than all systems except Scrubber, which averages 8.56 (SD 3.54) notes per second. CliniDeID is the slowest with an average of 1 (SD 0.38) note per second. Deid, NeuroNER, and Philter are all marginally faster at an average of 1.41 (SD 0.60), 1.28 (SD 0.48), and 1.36 (SD 0.64) notes per second, respectively.

Table 1. Minimum, mean, SD, and maximum processing times for each deidentification system derived from the real time (ie, wall clock) value given by the Linux time utility[a].

System	Seconds per note			Notes per second		
	Values, minimum (best)	Values, mean (SD)	Values, maximum (worst)	Values, maximum (best)	Values, mean (SD)	Values, minimum (worst)
CliniDeID	0.65	1.19 (0.59)	2.14	1.54	1.00 (0.38)	0.47
deid	0.44	0.85 (0.39)	1.41	2.28	1.41 (0.60)	0.71
MITRE Identity Scrubber Toolkit	0.01	0.08 (0.07)	0.19	80.85	24.57 (26.23)	5.37
NeuroNER	0.52	0.91 (0.41)	0.16	1.93	1.28 (0.48)	0.64
National Library of Medicine Scrubber	0.08	0.14 (0.06)	0.23	13.62	8.56 (3.54)	4.44
Philter	0.45	0.97 (0.62)	1.94	2.24	1.36 (0.64)	0.51

[a]Each data point reflects a single corpus split (eg, the 2006 train split vs the 2006 test split).

Performance Results Across All Categories Based on Partial Matching

For our initial performance metric evaluation, we focused on each system's overall ability to identify PII, regardless of the category. In the terminology of ETUDE, we collapsed evaluation across all categories. Figure 4 presents the recall, precision, and F_1-scores for each system against each corpus using the partial match alignment (this is the most generous evaluation possible).

The u-shaped curves for CliniDeID and Philter indicate that both these systems prioritize recall over precision. In contrast, the other 4 systems consistently have higher precision scores than recall scores. CliniDeID and NeuroNER show the highest scores across all corpora. Deid and MIST have the largest variance between corpora.

XSL•FO
RenderX

Figure 4. Recall, precision, and F1-score values for partial match annotation alignment wherein all personally identifiable information categories are collapsed into the broad tier 0 of "All Categories." MIST: MITRE Identity Scrubber Toolkit; MUSC: Medical University of South Carolina; NLM: National Library of Medicine.

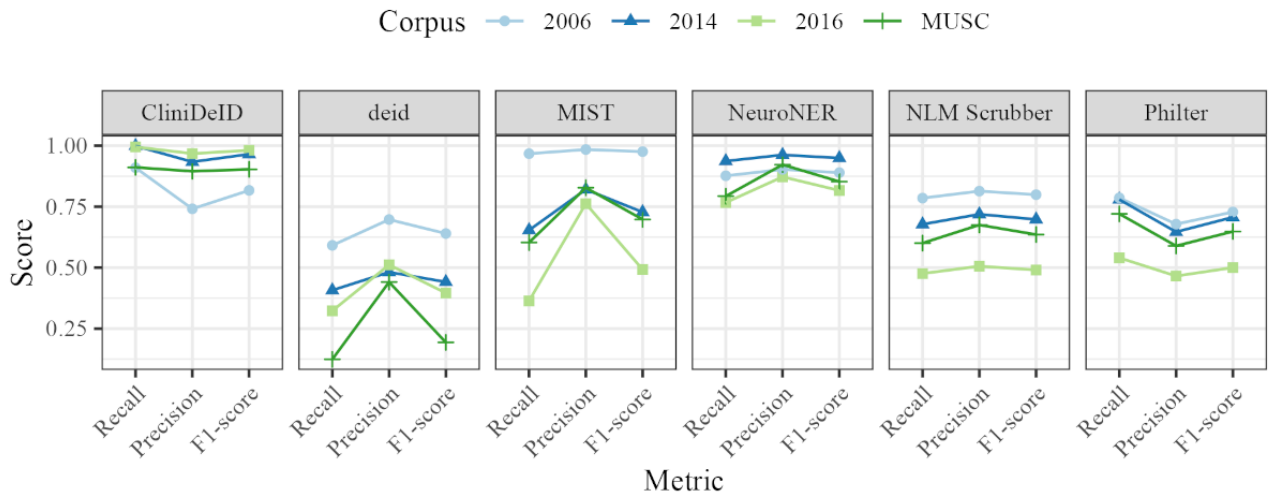

Performance Results for Specific Tier 1 Categories Based on Partial Matching

As different categories of PII are differentially sensitive, we wanted to delve deeper into the specific performance of systems with respect to the 7 tier 1 categories shown in Figure 1. For this analysis, we report the recall for partial annotation matching in Figure 5. Precision and F_1-score values are available in Multimedia Appendix 1. However, the nature of the limited categories generated by deid and Philter means that precision values for these systems are not truly meaningful.

CliniDeID shows more between-corpus variance than between-category variance and overall performs most consistently well. NeuroNER performs the next best across all categories, although "Contact Information," "Names," and "Time" are clearly better identified than the other 4 categories. Philter shows the same facility with "Contact Information" and "Names." Deid shows a large between-corpus variance but is generally consistent across categories. "Occupations" is not included as deid, MIST, and Scrubber do not extract that category. MIST identifies "Address" and "Contact Information" categories at a lower rate than other categories. Similarly, Scrubber identifies "Address" categories at a lower rate and does not extract any "Contact Information" annotations.

Figure 5. Recall values for partial match annotation alignment broken down into 7 bins for each of the major personally identifiable information groupings in the tier 1 categories. MIST: MITRE Identity Scrubber Toolkit; MUSC: Medical University of South Carolina; NLM: National Library of Medicine.

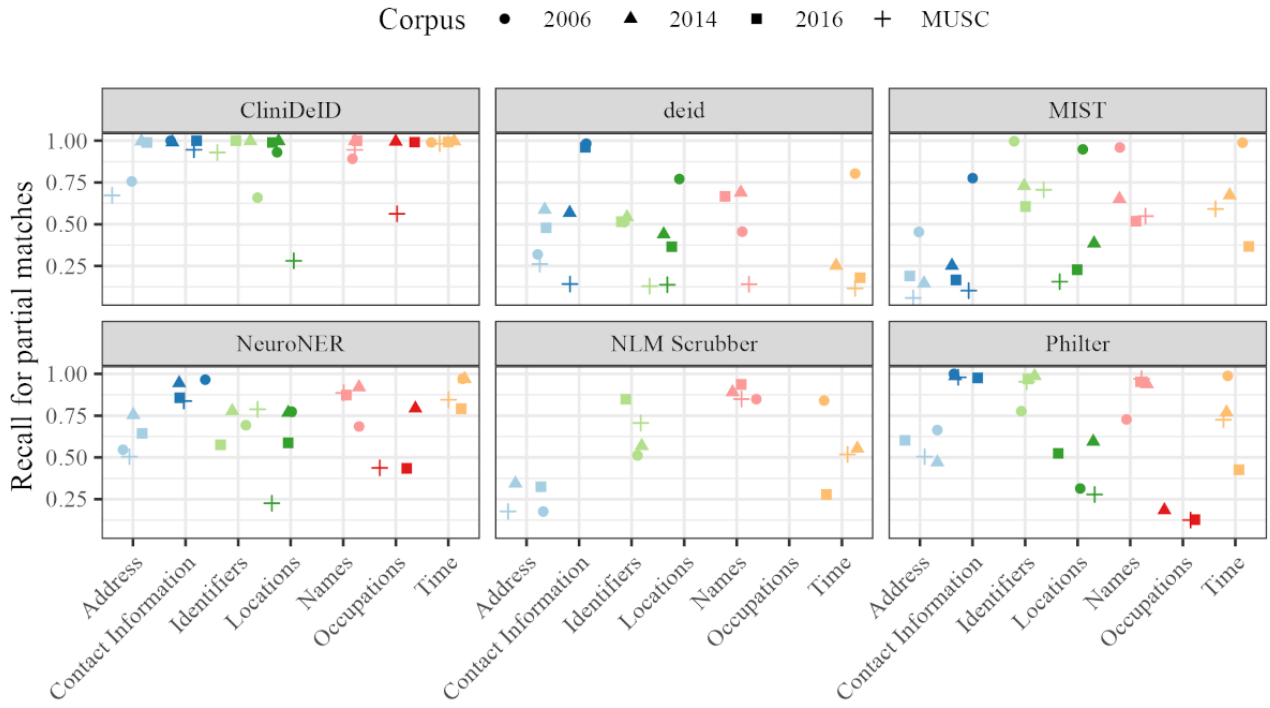

XSL•FO
RenderX

Performance Results for Names Based on Fully Contained Matching

As names are perhaps the most sensitive and unique PII category, we further analyzed the tier 1 "Names" category and the tier 2 categories of "Patient" and "Provider" for those systems that distinguish between these categories. For this analysis, we used the stricter fully contained annotation matching algorithm. Figure 6 summarizes the recall, precision, and F_1-score values across systems and corpora.

No strong corpus-specific trends show up in these results. MIST and NeuroNER have worse recall for identifying patient names than provider names, while CliniDeID does not appear to treat them differently. MIST and NeuroNER also have worse recall than precision for patient names but, similar to CliniDeID, no obvious difference between the 2 metrics for provider names. The gap between recall and precision is reduced for MIST and NeuroNER at the tier 1 "Names" level, which implies that some (but not all) of the tier 2 performance issues are due to patient names being incorrectly flagged as provider names. CliniDeID is consistently high for both recall and precision at tier 1. The deid system performs the worst for this evaluation, although its scores for "Names" are not so low for the partial matching evaluation in the *Performance Results for Specific Tier 1 Categories* subsection. This performance discrepancy indicates that deid does extract names reliably but does not extract large enough annotations to fully identify the relevant PII. Scrubber and Philter show much higher recall than precision for tier 1 "Names." As noted previously, Philter's precision for this evaluation is expected to be very low as the native system category relevant to "Names" includes all PII categories except dates. It is not obvious why Scrubber's precision is low.

Figure 6. Recall, precision, and F1-score values for fully contained match annotation alignment. The tier 1 "Names" category collapses "Patient," "Provider," "Relative," and "Other Person" names. The tier 2 "Patient" and "Provider" evaluations are given for the 3 systems that distinguish between name categories. MIST: MITRE Identity Scrubber Toolkit; MUSC: Medical University of South Carolina; NLM: National Library of Medicine.

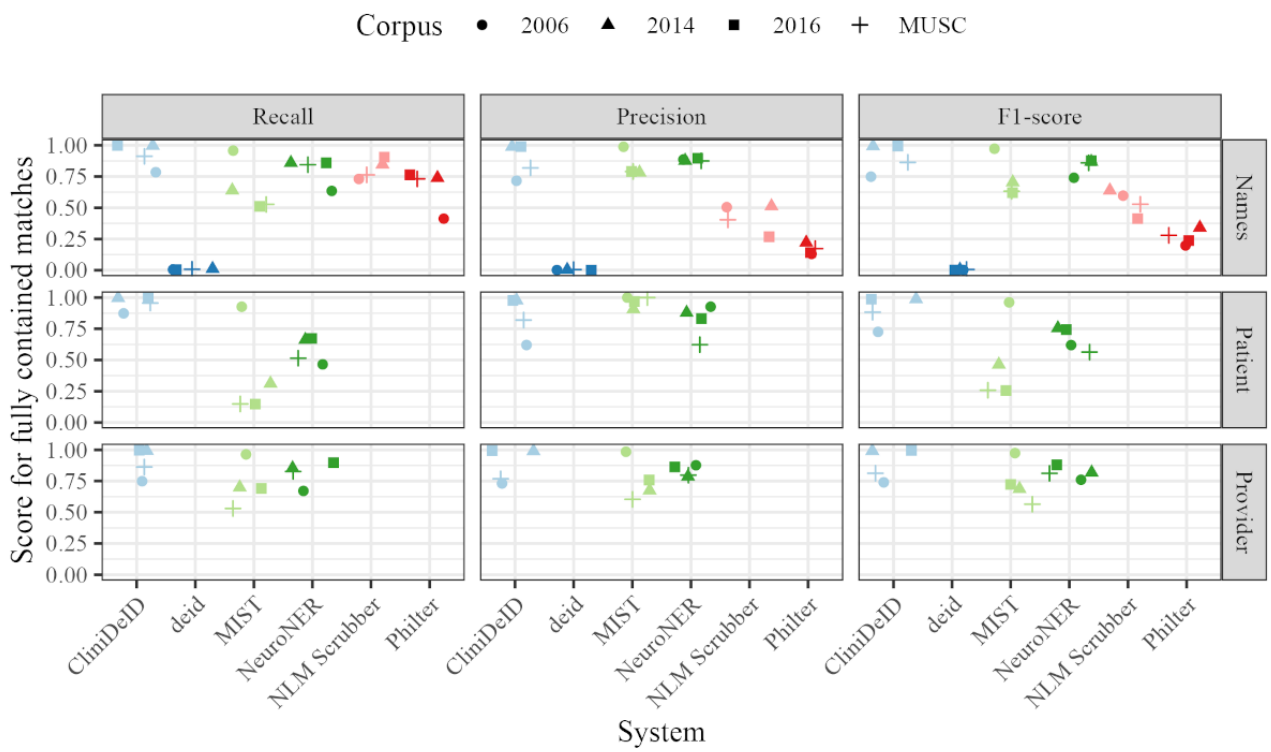

Discussion

Overview

As expected, given the complexity of the deidentification domain and feature variability in deidentification tools, we found that that none of the 6 systems uniformly outperformed the others across corpora and PII tag categories. CliniDeID, NeuroNER, and Philter generally outperformed the older (and not actively maintained) systems, but they also each have their performance drawbacks or limitations. Our evaluation framework was helpful in identifying these strong and weak facets, which, in turn, helped us identify additional aspects to include in future releases of the framework, as discussed in the next subsection.

Principal Findings

The principal findings of this study fall into 3 major classes: the evaluation pipeline implementation, the category mapping for bridging analyses across corpora or deidentification systems, and the evaluation results for the 6 off-the-shelf systems. The evaluation pipeline implementation includes many reusable components, including the shell scripts for passing corpus notes through systems, the ETUDE configuration scripts for scoring system outputs against reference annotations, and R scripts for plotting results of these analyses. The code for replication of the workflow is available via Codeberg and GitHub public code

XSL•FO
RenderX

repositories [34,35]. This code base can be used to increase reproducibility between studies on deidentification systems, lower the barrier to fine-tuning or retraining systems, and increase portability of evaluation across data sets and systems. The ETUDE configuration scripts provide a means for encoding and leveraging the PII category cross-mapping displayed in Figure 1. The cross-mapping can also be used to help compare the coverage of different deidentification systems with an eye to identifying the validity of different deidentification systems for different use cases (eg, when some categories of PII are known a priori to be either absent from or very frequent in a data set). Finally, our analysis of the text deidentification systems themselves yielded mixed results. CliniDeID and Philter had the highest recall for select categories, while CliniDeID and NeuroNER had the highest recall across all categories. We also found large differences in performance when comparing partial matches for PII categories to fully contained matches. The tabular data used to generate all tables and figures in this study have been included in Multimedia Appendix 1 to allow for finer-grained evaluation of the results.

Limitations

As discussed in the Introduction section, we focused our analysis on off-the-shelf models. CliniDeID, MIST, and NeuroNER can be retrained, while National Library of Medicine Scrubber and Philter have dictionaries that can be customized to a local environment. Thus, while our framework provides a means to easily measure the average baseline performance (across multiple corpora), more work is required to ascertain the average potential performance for each system (after retraining). Fortunately, each retrained model would only require a limited pass through the evaluation framework but would not require reprocessing and reevaluating all systems against all corpora. To further facilitate this use case, future iterations of this tool will make it even easier to add new corpora, systems, and evaluation configurations. Specifically, a more robust implementation of the pipeline would leverage reproducible workflow tools such as the targets R package [36]. We have released a related implementation using the targets R package for evaluating algorithmic bias in deidentification systems [37]. Xiao et al [38] provide an alternate approach using 100 synthetic templates imitating realistic contexts for PII as it occurs in unstructured clinical notes. For researchers who do not have the infrastructure or privacy needs of data sets with PII, other reproducible pipelines and shared evaluation frameworks, such as NLP Sandbox, exist [39].

In addition, we have not included all the essential evaluations for deciding between deidentification systems and will continue to expand our result set. For instance, we did not account for system start-up time. A system that takes 1 second per note for 1 or 10 million notes has different usability than a system that takes 0.1 second per note after loading models for 10 minutes. Another evaluation complication is how to equitably compare systems with significantly different input or output features. For instance, CliniDeID resynthesizes annotations by replacing the original PII with similar surrogate values, which likely slows down the overall pipeline compared to a system such as deid that does not track PII category in a manner accessible to a standard user. Similarly, TiDE (Text DEidentification) [40] meets all the aforementioned system requirements but uses prefilled PII details to boost performance. That is, when running a note through TiDE, one can provide the system with the patient's name, known providers, and other relevant PII that are likely to show up in a note. While this seems like a good plan at the system level, we need to develop a more complex evaluation pipeline to accommodate for the given information and to disaggregate results based on whether TiDE has correctly annotated known PII versus unknown PII (eg, a patient's nickname or misspelled name and a relative's name). TiDE was not evaluated and compared in this study for this reason, among others.

Finally, we relied on the mapping in Figure 1 to perform our evaluations, which, in turn, relies on a manual mapping built using experts' knowledge. Kim et al [41] evaluated a successful automated category mapping algorithm on deidentification concepts, which could be used as an alternate means for cross-mapping categories between corpus annotations and system outputs.

Comparison With Prior Work

Previous studies have included many of the same individual components that we present as novel in this study but, to the best of our knowledge, none have included all the components together. For instance, a category mapping between the 2014 and 2016 data sets was released at the time of the canonical paper on the 2016 shared task [9]. That same year, the researchers who developed Philter [42] released a mapping between the HIPAA Safe Harbor categories, the i2b2 2014 corpus annotations, and the categories redacted in the public release of the Multiparameter Intelligent Monitoring in Intensive Care II corpus [43,44]. The mapping we present in Figure 1 includes more corpora and extends the mapping to the unique set of categories used by each deidentification system, which greatly extends the utility of this new mapping for the community as a whole. Furthermore, our evaluation tool (ie, ETUDE), along with the published configuration files, provides the means to easily evaluate different category groupings without reannotating the data sets.

Most prior studies do not explicitly mention details of the evaluation (eg, whether they use token counts and exact match character offsets) unless they are using the standard evaluation script released with the 2016 corpus [9]. While the widespread use and public availability of this script are a boon to reproducible science, the script's limited options, hardcoded categories, and fixed file format are a bane. Most prior studies report the precision, recall, and F_1-score, while a select few report the F_2-score [11,45] or both [46]. We reported the F_1-score, but ETUDE can easily be configured to report the F measure for any arbitrary β value. The relative processing speeds (Table 1) for deidentification systems is also rarely reported [33]. A few prior studies have included specific annotation guidelines, including examples and edge cases, to help clarify the underlying ground truth intended by category labels [8,25,26]. Finally, our own work and many of the other researchers and developers cited in this study have relied on the same standard deidentification data sets released via the i2b2 or CEGS N-GRID shared tasks. For less common but still

XSL•FO
RenderX

publicly available data sets, the developers of NeuroNER evaluated their system against the CoNLL 2003 shared task on named entity recognition [23], and Xiao et al [38] recently released a new smaller set of notes based on MIMIC-IV. We are not aware of any other study with a publicly released evaluation framework that includes all steps from initial corpus processing through plotting evaluation results.

Conclusions

We present a reusable and extensible evaluation framework applied to deidentification systems for clinical unstructured notes. We release the tool for other researchers to reduce the overhead in testing new systems and new corpora, a critical and common task within the clinical NLP community, especially for researchers who wish to ethically share data without leaking PHI.

From this initial case study, we found an order of magnitude difference in terms of processing speed between the fastest and slowest systems and that no single system out of the 6 uniformly outperformed the others across corpora and PII categories. Instead, a richer tapestry of recall and precision trade-offs emerged for different PII categories and groups of PII categories. We hold that a single evaluation pipeline across multiple systems and corpora allows for more nuanced comparisons between systems and serves as a boon to the clinical NLP community.

Acknowledgments

This publication was supported, in part, by the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1 TR001450) and the SmartState Endowment for Translational Biomedical Informatics. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The 2006 and 2014 corpora analyzed during this study are available in the National NLP Clinical Challenges website repository [47]. The 2016 corpus analyzed during this study is not publicly available due to privacy concerns. The Medical University of South Carolina corpus is not publicly available due to privacy concerns, as the data set has not been fully redacted or resynthesized to remove personally identifiable information. Code for replication of the workflow is available via Codeberg and GitHub public code repositories [34,35].

Authors' Contributions

PMH and SMM made substantial contributions to the design and implementation of the research and to the analysis of the experimental results. All authors drafted the work or revised it critically. PMH drafted the initial manuscript. SMM performed critical revision of the manuscript. All authors provided final approval for the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

CliniDeID, one of the systems evaluated in this manuscript, was initially developed and marketed as a commercial solution but is not commercialized anymore and has been released as free and open-source software since 2023.

Multimedia Appendix 1
CSV data files used to generate figures.
[ZIP File (Zip Archive), 47 KB - jmir_v26i1e55676_app1.zip]

Multimedia Appendix 2
Details on each of the 6 deidentification system's release information (eg, version number), configuration, and use.
[DOCX File , 32 KB - jmir_v26i1e55676_app2.docx]

References

1. Douglass MM, Cliffford GD, Reisner A, Long WJ, Moody GB, Mark RG. De-identification algorithm for free-text nursing notes. In: Proceedings of the Computers in Cardiology. 2005 Presented at: CIC 2005; September 25-28, 2005; Lyon, France. [doi: 10.1109/CIC.2005.1588104]
2. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008 Jul 24;8:32 [FREE Full text] [doi: 10.1186/1472-6947-8-32] [Medline: 18652655]
3. Steinkamp JM, Pomeranz T, Adleberg J, Kahn CEJ, Cook TS. Evaluation of automated public de-identification tools on a corpus of radiology reports. Radiol Artif Intell 2020 Oct 14;2(6):e190137 [FREE Full text] [doi: 10.1148/ryai.2020190137] [Medline: 33937843]

XSL•FO
RenderX

4. Chen F, Bokhari SM, Cato K, Gürsoy G, Rossetti SC. Examining the generalizability of pretrained de-identification transformer models on narrative nursing notes. Appl Clin Inform 2024 Mar 06 (forthcoming) [FREE Full text] [doi: 10.1055/a-2282-4340] [Medline: 38447965]
5. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996:333-337 [FREE Full text] [Medline: 8947683]
6. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550-563 [FREE Full text] [doi: 10.1197/jamia.M2444] [Medline: 17600094]
7. Morita M, Kano Y, Ohkuma T, Miyabe M, Aramaki E. Overview of the NTCIR-10 MedNLP task. In: Proceedings of the 10th NTCIR Conference. 2013 Presented at: NTCIR 2013; June 18-21, 2013; Tokyo, Japan.
8. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S11-S19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.007] [Medline: 26225918]
9. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform 2017 Nov;75S:S4-18 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.011] [Medline: 28614702]
10. Murugadoss K, Rajasekharan A, Malin B, Agarwal V, Bade S, Anderson JR, et al. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. Patterns (N Y) 2021 May 12;2(6):100255 [FREE Full text] [doi: 10.1016/j.patter.2021.100255] [Medline: 34179842]
11. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med 2020 Apr 14;3:57 [FREE Full text] [doi: 10.1038/s41746-020-0258-y] [Medline: 32337372]
12. Liu L, Perez-Concha O, Nguyen A, Bennett V, Jorm L. De-identifying Australian hospital discharge summaries: an end-to-end framework using ensemble of deep learning models. J Biomed Inform 2022 Nov;135:104215 [FREE Full text] [doi: 10.1016/j.jbi.2022.104215] [Medline: 36195240]
13. Heider PM, Accetta JK, Meystre SM. ETUDE for easy and efficient NLP application evaluation. In: Proceedings of the AMIA NLP-WG Pre-Symposium. 2018 Presented at: AMIA NLP-WG Pre-Symposium 2018; November 3, 2018; San Francisco, CA.
14. Heider PM, Accetta JK, Meystre SM. ETUDE: demonstrating multiple matching styles and offset calculators within a natural language processing evaluation tool. In: Proceedings of the 17th World Congress on Medical and Health Informatics. 2019 Presented at: MEDINFO 2019; August 25-30, 2019; Lyon, France.
15. 45 CFR parts 160 and 164. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/introduction/index.html [accessed 2024-05-01]
16. General data protection regulation compliance guidelines. EU General Data Protection Regulation. URL: https://gdpr.eu/ [accessed 2024-05-01]
17. Meystre SM, Heider PM, Kim Y, Trice A, Underwood G. Clinical text automatic de-identification to support large scale data reuse and sharing: pilot results. In: Proceedings of the American Medical Informatics Association Annual Symposium. 2018 Presented at: AMIA 2018; November 3-7, 2018; San Francisco, CA.
18. Clinacuity / CliniDeID. GitHub. URL: https://github.com/Clinacuity/CliniDeID [accessed 2024-04-08]
19. De-identification software package. PhysioNet. 2007 Dec 18. URL: https://physionet.org/content/deid/1.1/ [accessed 2024-04-09]
20. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE identification scrubber toolkit: design, training, and assessment. Int J Med Inform 2010 Dec;79(12):849-859. [doi: 10.1016/j.ijmedinf.2010.09.007] [Medline: 20951082]
21. MITRE identification scrubber toolkit files. SourceForge. URL: https://sourceforge.net/projects/mist-deid/files/ [accessed 2024-04-08]
22. MIST: the MITRE identification scrubber toolkit (documentation). SourceForge. URL: https://mist-deid.sourceforge.net/docs_2_0_4/html/index.html [accessed 2024-04-08]
23. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2017 Presented at: EMNLP 2017; September 9-11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/d17-2017]
24. Franck-Dernoncourt / NeuroNER. GitHub. URL: https://github.com/Franck-Dernoncourt/NeuroNER [accessed 2024-04-08]
25. Kayaalp M, Browne AC, Sagan P, McGee T, McDonald CJ. Challenges and insights in using HIPAA privacy rule for clinical text annotation. AMIA Annu Symp Proc 2015 Nov 05;2015:707-716 [FREE Full text] [Medline: 26958206]
26. Kayaalp M, Sagan P, Browne AC, McDonald CJ. Guidelines for annotating personal identifiers in the clinical text repository of the National Institutes of Health (version 6/28/2016). Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health. 2016. URL: https://lhncbc.nlm.nih.gov/scrubber/annotation.html [accessed 2022-12-23]
27. NLM-scrubber downloads. National Institutes of Health National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/scrubber/download.html [accessed 2024-04-08]

XSL•FO
RenderX

28. BCHSI/philter-ucsf. GitHub. URL: https://github.com/BCHSI/philter-ucsf [accessed 2024-04-08]
29. Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc 2013;20(2):342-348 [FREE Full text] [doi: 10.1136/amiajnl-2012-001034] [Medline: 22771529]
30. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: EMNLP 2014; October 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
31. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012 Presented at: EACL 2012; April 23-27, 2012; Avignon, France.
32. MUSC-TBIC/corpus-utils. GitHub. URL: https://github.com/MUSC-TBIC/corpus-utils [accessed 2020-07-27]
33. Heider PM, Obeid JS, Meystre SM. A comparative analysis of speed and accuracy for three off-the-shelf de-identification tools. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:241-250 [FREE Full text] [Medline: 32477643]
34. Heider PM. HeiderLab/article-addenda. Codeberg. URL: https://codeberg.org/HeiderLab/article-addenda [accessed 2023-10-04]
35. MUSC-TBIC/article-addenda. GitHub. URL: https://github.com/MUSC-TBIC/article-addenda [accessed 2023-10-04]
36. Landau WM. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. J Open Source Softw 2021 Jan 15;6(57):2959. [doi: 10.21105/joss.02959]
37. Heider PM. Algorithmic bias in de-identification tools. In: Proceedings of the IEEE 11th International Conference on Healthcare Informatics. 2023 Presented at: ICHI 2023; June 26-29, 2023; Houston, TX. [doi: 10.1109/ichi57859.2023.00129]
38. Xiao Y, Lim S, Pollard TJ, Ghassemi M. In the name of fairness: assessing the bias in clinical record de-identification. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023 Presented at: FAccT '23; June 12-15, 2023; Chicago, IL. [doi: 10.1145/3593013.3593982]
39. NLP sandbox. GitHub. URL: https://github.com/nlpsandbox [accessed 2020-07-27]
40. TiDE text deid. STAnford Medicine Research data Repository. URL: https://starr.stanford.edu/methods/tide-text-deid [accessed 2023-10-02]
41. Kim Y, Heider PM, Meystre SM. Automated category alignment applied to different de-identification annotation schemata. In: Proceedings of the AMIA 2021 Virtual Informatics Summit. 2021 Presented at: AMIA 2021; March 22-25, 2021; Online.
42. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017 May 01;24(3):596-606 [FREE Full text] [doi: 10.1093/jamia/ocw156] [Medline: 28040687]
43. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
44. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, et al. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 2011 May;39(5):952-960 [FREE Full text] [doi: 10.1097/CCM.0b013e31820a92c6] [Medline: 21283005]
45. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol 2012 Jul 27;12:109 [FREE Full text] [doi: 10.1186/1471-2288-12-109] [Medline: 22839356]
46. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013 Jan 01;20(1):77-83 [FREE Full text] [doi: 10.1136/amiajnl-2012-001020] [Medline: 22947391]
47. Data sets. National NLP Clinical Challenges (n2c2). URL: https://n2c2.dbmi.hms.harvard.edu/data-sets [accessed 2022-03-08]

Abbreviations

CEGS N-GRID: Centers of Excellence in Genomic Science Neuropsychiatric Genome-Scale and RDOC Individualized Domains
ETUDE: Evaluation Tool for Unstructured Data and Extractions
HIPAA: Health Insurance Portability and Accountability Act
i2b2: Informatics for Integrating Biology and the Bedside
IRB: institutional review board
MIST: MITRE Identity Scrubber Toolkit
MUSC: Medical University of South Carolina
NLP: natural language processing
PHI: protected health information
PII: personally identifiable information
TiDE: Text DEidentification
UTHealth: University of Texas Health Science Center at Houston

XSL•FO
RenderX

Edited by G Tsafnat; submitted 20.12.23; peer-reviewed by C Taridzo, L Radhakrishnan; comments to author 07.03.24; revised version received 11.04.24; accepted 13.04.24; published 28.05.24.

Please cite as:
Heider PM, Meystre SM
An Extensible Evaluation Framework Applied to Clinical Text Deidentification Natural Language Processing Tools: Multisystem and Multicorpus Study
J Med Internet Res 2024;26:e55676
URL: https://www.jmir.org/2024/1/e55676
doi:10.2196/55676
PMID:38805692

©Paul M Heider, Stéphane M Meystre. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Augmenting K-Means Clustering With Qualitative Data to Discover the Engagement Patterns of Older Adults With Multimorbidity When Using Digital Health Technologies: Proof-of-Concept Trial

Yiyang Sheng[1], MSc; Raymond Bond[2], PhD; Rajesh Jaiswal[3], PhD; John Dinsmore[4], PhD; Julie Doyle[1], PhD

[1]NetwellCASALA, Dundalk Institution of Technology, Dundalk, Ireland
[2]School of Computing, Ulster University, Jordanstown, United Kingdom
[3]School of Enterprise Computing and Digital Transformation, Technological University Dublin, Dublin, Ireland
[4]Trinity Centre for Practice and Healthcare Innovation, School of Nursing and Midwifery, Trinity College Dublin, Dublin, Ireland

Corresponding Author:
Yiyang Sheng, MSc
NetwellCASALA
Dundalk Institution of Technology
Dublin Road, PJ Carrolls Building, Dundalk Institute of Technology
Co.Louth, Ireland
Dundalk, A91 K584
Ireland
Phone: 353 894308214
Email: shengexz@gmail.com

Abstract

Background: Multiple chronic conditions (multimorbidity) are becoming more prevalent among aging populations. Digital health technologies have the potential to assist in the self-management of multimorbidity, improving the awareness and monitoring of health and well-being, supporting a better understanding of the disease, and encouraging behavior change.

Objective: The aim of this study was to analyze how 60 older adults (mean age 74, SD 6.4; range 65-92 years) with multimorbidity engaged with digital symptom and well-being monitoring when using a digital health platform over a period of approximately 12 months.

Methods: Principal component analysis and clustering analysis were used to group participants based on their levels of engagement, and the data analysis focused on characteristics (eg, age, sex, and chronic health conditions), engagement outcomes, and symptom outcomes of the different clusters that were discovered.

Results: Three clusters were identified: the typical user group, the least engaged user group, and the highly engaged user group. Our findings show that age, sex, and the types of chronic health conditions do not influence engagement. The 3 primary factors influencing engagement were whether the same device was used to submit different health and well-being parameters, the number of manual operations required to take a reading, and the daily routine of the participants. The findings also indicate that higher levels of engagement may improve the participants' outcomes (eg, reduce symptom exacerbation and increase physical activity).

Conclusions: The findings indicate potential factors that influence older adult engagement with digital health technologies for home-based multimorbidity self-management. The least engaged user groups showed decreased health and well-being outcomes related to multimorbidity self-management. Addressing the factors highlighted in this study in the design and implementation of home-based digital health technologies may improve symptom management and physical activity outcomes for older adults self-managing multimorbidity.

(J Med Internet Res 2024;26:e46287) doi:10.2196/46287

KEYWORDS

aging; digital health; multimorbidity; chronic disease; engagement; k-means clustering

XSL•FO
RenderX

Introduction

Background

According to the United Nations, the number of people aged ≥65 years is growing faster than all other age groups [1]. The worldwide population of people aged ≥65 years will increase from approximately 550 million in 2000 to 973 million in 2030 [2]. Furthermore, by 2050, approximately 16% of the world's population will be aged >65 years, whereas 426 million people will be aged >80 years [1]. Living longer is a great benefit to today's society. However, this comes with several challenges. Aging can be associated with many health problems, including multimorbidity (ie, the presence of ≥2 chronic conditions) [3]. The prevalence rate of multimorbidity among older adults is estimated to be between 55% and 98%, and the factors associated with multimorbidity are older age, female sex, and low socioeconomic status [4]. In the United States, almost 75% of older adults have multimorbidity [5], and it was estimated that 50 million people in the European Union were living with multimorbidity in 2015 [6]. Likewise, the prevalence rate of multimorbidity is 69.3% among older adults in China [5].

Home-based self-management for chronic health conditions involves actions and behaviors that protect and promote good health care practices comprising the management of physical, emotional, and social care [7]. Engaging in self-management can help older adults understand and manage their health conditions, prevent illness, and promote wellness [7,8]. However, self-management for older adults with multimorbidity is a long-term, complex, and challenging mission [9,10]. There are numerous self-care tasks to engage in, which can be very complicated, especially for people with multiple chronic health conditions. Furthermore, the severity of the disease can negatively impact a person's ability to engage in self-management [10].

Digital home-based health technologies have the potential to support better engagement with self-management interventions, such as the monitoring of symptom and well-being parameters as well as medication adherence [10,11]. Such technologies can help older adults understand their disease or diseases, respond to changes, and communicate with health care providers [12-14]. Furthermore, digital health technologies can be tailored to individual motivations and personal needs [13], which can improve sustained use [15] and result in people feeling supported [16]. Digital self-management can also create better opportunities for adoption and adherence in the long term compared with paper booklet self-management [16]. Moreover, digital health technologies, such as small wearable monitoring devices, can increase the frequency of symptom monitoring for patients with minimal stress compared with symptom monitoring with manual notifications [17].

A large body of research implements data mining and machine learning algorithms using data acquired from home-based health care data sets. Data mining techniques, such as data visualization, clustering, classification, and prediction, to name a few, can help researchers understand users, behaviors, and health care phenomena by identifying novel, interesting patterns. These techniques can also be used to build predictive models [18-21]. In addition, data mining techniques can help in designing health care management systems and tracking the state of a person's chronic disease, resulting in appropriate interventions and a reduction in hospital admissions [18,22]. Vast amounts of data can be generated when users interact with digital health technologies, which provides an opportunity to understand chronic illnesses as well as elucidate how users engage with digital health technologies in the real world. Armstrong et al [23] used the k-means algorithm to identify previously unknown patterns of clinical characteristics in home care rehabilitation services. The authors used k-means cluster analysis to analyze data from 150,253 clients and discovered new insights into the clients' characteristics and their needs, which led to more appropriate rehabilitation services for home care clients. Madigan and Curet [22] used classification and regression trees to investigate a home-based health care data set that comprised 580 patients who had 3 specific conditions: chronic obstructive pulmonary disease (COPD), heart failure (HF), and hip replacement. They found that data mining methods identified the dependencies and interactions that influence the results, thereby improving the accuracy of risk adjustment methods and establishing practical benchmarks [22]. Other research [24] has developed a flow diagram of a proposed platform by using machine learning methods to analyze multiple health care data sets, including medical images as well as diagnostic and voice records. The authors believe that the system could help people in resource-limited areas, which have lower ratios of physicians and hospitals, to diagnose diseases such as breast cancer, heart disease (HD), diabetes, and liver disease at a lower cost and in less time than local hospitals. In the study, the accuracy of disease detection was >95% [24].

There are many different approaches to clustering analysis of health care data sets, such as k-means, density-based spatial clustering of applications with noise, agglomerative hierarchical clustering, self-organizing maps, partitioning around medoids algorithm, hybrid hierarchical clustering, and so on [25-28]. K-means clustering is 1 of the most commonly used clustering or unsupervised machine learning algorithms [19,29], and it is relatively easy to implement and relatively fast [30-32]. In addition, k-means has been used in research studies related to chronic health conditions such as diabetes [33], COPD [34,35], and HF [36]; for example, a cloud-based framework with k-means clustering technique has been used for the diagnosis of diabetes and was found to be more efficient and suitable for handling extensive data sets in cloud computing platforms than hierarchical clustering [32]. Violán et al [37] analyzed data from 408,994 patients aged 45 to 64 years with multimorbidity using k-means clustering to ascertain multimorbidity patterns. The authors stratified the k-means clustering analysis by sex, and 6 multimorbidity patterns were found for each sex. They also suggest that clusters identified by multimorbidity patterns obtained using nonhierarchical clustering analysis (eg, k-means and k-medoids) are more consistent with clinical practice [37].

The majority of data mining studies on chronic health conditions focus on the diseases themselves and their symptoms; there is less exploration of the patterns of engagement of persons with multimorbidity with digital health technologies. However, data mining and machine learning are excellent ways to understand

XSL•FO
RenderX

users' engagement patterns with digital health technologies. A study by McCauley et al [38] compared clustering analysis of the user interaction event log data from a reminiscence mobile app that was designed for people living with dementia. In addition to performing quantitative user interaction log analysis, the authors also gathered data on the qualitative experience of users. The study showed the benefits of using data mining to analyze the user log data with complementary qualitative data analysis [38]. This is a research challenge where both quantitative and qualitative methods can be combined to fully understand users; for example, the quantitative analysis of the user event data can tell us about use patterns, the preferred times of day to use the app, the feature use, and so on, but qualitative data (eg, user interviews) are necessary to understand *why* these use patterns exist.

Objectives

The aim of this study was to analyze how older adults with multimorbidity engage with digital symptom and health monitoring over a period of approximately 12 months using a digital health platform. In this study, user log data of engagement with digital health technology and user interview qualitative data were examined to explore the patterns of engagement. K-means clustering was used to analyze the user log data. The study had four research questions: (1) How do clusters differ in terms of participant characteristics such as age, sex, and health conditions? (2) How do clusters differ in terms of patterns of engagement, such as the number of days a week participants take readings (eg, weight and blood pressure [BP])? (3) How do engagement rates with the different devices correlate with each other (determined by analyzing the weekly submissions of every parameter and the interviews of participants)? and (4) How do engagement rates affect participants' health condition symptoms, such as BP, blood glucose (BG) level, weight, peripheral oxygen saturation (SpO_2) level, and physical activity (PA)?

Methods

Overview

The study was a proof-of-concept trial with an action research design and mixed methods approach. Action research is a period of investigation that "describes, interprets, and explains social situations while executing a change intervention aimed at improvement and involvement" [39]. An action research approach supports the generation of solutions to practical problems while using methods to understand the contexts of care as well as the needs and experiences of participants.

Recruitment and Sample

Although 120 participants consented to take part across Ireland and Belgium, this paper reports on data from 60 Irish older adults with multiple chronic health conditions (≥2 of the following: COPD, HF, HD, and diabetes). Participants were recruited through purposive sampling and from multiple sources, including through health care organizations (general practitioner clinics and specialist clinics), relevant older adult networks, chronic disease support groups, social media, and local newspaper advertising. Recruitment strategies included the use of study flyers and advertisements as well as giving talks and platform demonstrations.

Sources of Data

The data set was collected during the Integrated Technology Systems for Proactive Patient Centred Care (ProACT) project proof-of-concept trial. As the trial was a proof-of-concept of a novel digital health platform, the main goal was to understand how the platform worked or did not work, rather than whether it worked. Thus, to determine sample size, a pragmatic approach was taken in line with two important factors: (1) Is the sample size large enough to provide a reliable analysis of the ecosystem? and (2) Is the sample size small enough to be financially feasible? The literature suggests that overall sample size in proof-of-concept digital health trials is low. A review of 1030 studies on technical interventions for management of chronic disease that focused on HF (436 studies), stroke (422 studies), and COPD (172 studies) suggested that robust sample sizes were 17 for COPD, 19 for HF, and 21 for stroke [40]. Full details on the study protocol can be found in the study by Dinsmore et al [41].

Participants used a suite of sensor devices (ie, BP monitors, weight scales, glucometers, pulse oximeters, and activity watches) and a tablet app to monitor their health conditions and well-being. All participants received a smartwatch to measure PA levels and sleep, a BP monitor to measure BP and pulse rate, and a weight scale. A BG meter was provided to participants with diabetes, and a pulse oximeter was provided to those with COPD to measure SpO_2 levels. In addition, all participants received an iPad with a custom-designed app, the ProACT CareApp, that allowed users to view their data, provide self-report (SR) data on symptoms that could not be easily captured through a sensor (eg, breathlessness and edema) and well-being (eg, mood and satisfaction with social life), receive targeted education based on their current health status, set PA goals, and share their data with others. The ProACT platform was designed and developed following an extensive user-centered design process. This involved interviews, focus groups, co-design sessions (hands-on design activities with participants), and usability testing before the platform's deployment in the trial. A total of 58 people with multimorbidity and 106 care network participants, including informal carers, formal carers, and health care professionals, took part in this process. Findings from the user-centered design process have been published elsewhere [42,43]. More detailed information about the full ProACT platform and the CareApp used by participants can be found in the study by Doyle et al [44].

The study took place between April 1, 2018, and June 30, 2019. Participants in the trial typically participated for 12 months, although some stayed on for 14 months and others for 9 months (in the case of those who entered the trial later). One of the trial objectives was to understand real-world engagement. Therefore, participants were asked to take readings with the devices and provide SR data in the ProACT CareApp whenever they wished (not necessarily daily). As part of the trial, participants were assisted by technical help desk staff who responded to questions about the technology, and home visits were conducted as needed to resolve issues. In addition, a clinical triage service monitored

XSL•FO
RenderX

the participants' readings and contacted them in instances of abnormal parameter values (eg, high BP and low SpO_2 levels) [45]. Participants also received a monthly check-in telephone call from 1 of the triage nurses.

Table 1 outlines the types of health and well-being metrics that were collected, as well as the collection method and the number of participants who collected that type of data. The health and well-being metrics were determined from the interviews and focus groups held with health care professionals during the design of the ProACT platform to determine the most important symptom and well-being parameters to monitor across the health conditions of interest [42]. Off-the-shelf digital devices manufactured by 2 providers, Withings and iHealth, were used during the trial. Data from these providers were extracted into a custom platform called Context-Aware Broker and Inference Engine–Subject Information Management System (CABIE-SIMS), which includes a data aggregator for storing health and well-being data. All devices require the user to interact with them in some way. However, some devices needed more interaction than others (eg, taking a BG reading involved several steps, but PA and sleep only required participants to open the activity watch app to sync the relevant data). The activity watch was supposed to synchronize automatically without user interaction. However, inconsistencies with syncing meant that users were advised to open the Withings app to sync their data. The CABIE-SIMS platform would display the readings in *near* real time, apart from PA data, which were collected at regular intervals throughout the day, whereas sleep data were gathered every morning. Table 1 lists the types of data that were collected and the number of participants who collected them. In addition, semistructured interviews were conducted with all participants at 4 time points throughout the trial to understand their experience of using the ProACT platform. Although a full qualitative thematic analysis was outside the scope of this study and was reported on elsewhere [44], interview transcripts for participants of interest to the analysis presented in this paper were reviewed as part of this study to provide an enhanced understanding of the results.

Table 1. Types of data, collection methods, and number of participants collecting these data (n=60).

Data type	Collection method	Participants (at start of trial), n (%)
Blood pressure	Place device on arm and turn on device, which opens Withings Health Mate app to collect data; press "Start" in app to take reading	60 (100)
Pulse rate	Collected as part of blood pressure measurement	60 (100)
Blood glucose level	Turn on device and open app; prepare lancing device by inserting new lancet and setting puncture depth; wash hands thoroughly; insert test strip into device; take blood sample from the finger; apply blood sample to test strip, and wait for result to display; discard test strip and lancet	34 (57)
SpO_2[a] level	Place device in current orientation on index finger; turn on device and open app to take reading	22 (37)
Weight	Stand on weight scales; reading is automatically transferred via Wi-Fi to app	60 (100) as lifestyle parameter, including 11 (18) as symptom parameter for HF[b]
Physical activity	Participants advised to open Withings Health Mate app at least once per day to ensure syncing of data	60 (100)
Sleep	Participants advised to open Withings Health Mate app at least once per day to ensure syncing of data	60 (100)
Self-report (general well-being, eg, mood, anxiety, satisfaction, and medication adherence)	Answered through ProACT[c] CareApp and automatically pulled into the CABIE-SIMS[d] platform; most questions delivered daily	60 (100)
Self-report (COPD[e] symptoms, eg, breathlessness and sputum)	Answered through ProACT CareApp and automatically pulled into the CABIE-SIMS platform; questions delivered daily	22 (37)
Self-report (HF symptoms, eg, swelling and nighttime breathlessness)	Answered through ProACT CareApp and automatically pulled into the CABIE-SIMS platform; questions delivered daily	10 (17)

[a]SpO_2: peripheral oxygen saturation.

[b]HF: heart failure.

[c]ProACT: Integrated Technology Systems for Proactive Patient Centred Care.

[d]CABIE-SIMS: Context-Aware Broker and Inference Engine–Subject Information Management System.

[e]COPD: chronic obstructive pulmonary disease.

Data Analysis Methods

The original data set in the CABIE-SIMS platform was formatted using the JSON format. As a first step, a JSON-to-CSV file converter was used to make the data set more accessible for data analysis. The main focus was on dealing with duplicate data and missing data during the data cleaning phase. Data duplication might occur when a user uploads their SpO_2 reading 3 times in 2 minutes as a result of *mispressing* the button. In such cases, only 1 record was added to the cleaned

XSL•FO
RenderX

data file. As for missing data, the data set file comprised "N/A" (not available) values for all missing data.

The cleaned data set was preprocessed using Microsoft Excel, the R programming language (R Foundation for Statistical Computing), and RStudio (Posit Software, PBC). The preprocessed data set included participants' details (ID, sex, age, and chronic health conditions) and the number of days of weekly submissions of every parameter (BP, pulse rate, SpO_2 level, BG level, weight, PA, SR data, and sleep). All analyses (including correlation analysis, principal component analysis [PCA], k-means clustering, 2-tailed *t* test, and 1-way ANOVA) were implemented in the R programming language and RStudio.

After performing Shapiro-Wilk normality tests on the data submitted each week, we found that the data were not normally distributed. Therefore, Spearman correlation was used to check the correlation among the parameters. Correlation analysis and PCA were used to determine which portions of the data would be included in the k-means clustering. Correlation analysis determined which characteristics or parameters should be selected, and PCA determined the number of dimensions that should be selected as features for clustering. In the clustering process, the weekly submission of each parameter was considered as an independent variable for the discovery of participant clusters, and the outcome of the clustering was a categorical taxonomy that was used to label the 3 discovered clusters. Similarly, the Shapiro-Wilk test was conducted to check the normality of the variables in each group. It was found that most of the variables in each group were normally distributed, and only the weight data submission records of cluster 3, the PA data submission records of cluster 2, the SR data submission records of cluster 3, and the sleep data submission records of cluster 1 were not normally distributed. Therefore, the 2-tailed *t* test and 1-way ANOVA were used to compare different groups of variables. The 2-tailed *t* test was used to compare 2 groups of variables, whereas 1-way ANOVA was used to compare ≥2 groups of variables. *P* values >.05 indicated that there were no statistically significant differences among the groups of variables [46].

As for the qualitative data from the interviews, we performed keyword searches after a review of the entire interview; for example, when the data analysis was related to BP and weight monitoring, a search with the keywords "blood pressure," "weight," or "scale" was performed to identify relevant information. In addition, when the aim was to understand the impact of digital health care technology, we focused on specific questions in the second interview, such as "Has it had any impact on the management of your health?"

Ethical Considerations

Ethics approval was received from 3 ethics committees: the Health Service Executive North East Area Research Ethics Committee, the School of Health and Science Research Ethics Committee at Dundalk Institute of Technology, and the Faculty of Health Sciences Research Ethics Committee at Trinity College Dublin. All procedures were in line with the European Union's General Data Protection Regulation for research projects, with the platform and trial methods and procedures undergoing data protection impact assessments. Written informed consent was obtained on an individual basis from participants in accordance with legal and ethics guidelines after a careful explanation of the study and the provision of patient information and informed consent forms in plain language. All participants were informed of their right to withdraw from the study at any time without having to provide a reason. Participants were not compensated for their time. Data stored within the CABIE-SIMS platform were identifiable because they were shared (with the participant's consent) with the clinical triage teams and health care professionals. This was clearly outlined in the participant information leaflet and consent form. However, the data set that was extracted for the purpose of the analysis presented in this paper was pseudonymized.

Results

Participants

A total of 60 older adults were enrolled in the study. The average age of participants was 74 (SD 6.4; range 65-92) years; 60% (36) were male individuals, and 40% (24/60) were female individuals. The most common combination of health conditions was diabetes and HD (30/60, 50%), which was followed by COPD and HD (16/60, 27%); HF and HD (7/60, 12%); diabetes and COPD (3/60, 5%); diabetes and HF (1/60, 2%); COPD and HF (1/60, 2%); HF, HD, and COPD (1/60, 2%); and COPD, HD, and diabetes (1/60, 2%). Of the 60 participants, 11 (18%) had HF, 55 (92%) had HD, 22 (37%) had COPD, and 31 (52%) had diabetes. Over the course of the trial, of the 60 participants, 8 (13%) withdrew, and 3 (5%) died. However, this study included data from all participants in the beginning, as long as the participant had at least 1 piece of data. Hence, of the 60 participants, we included 56 (93%) in our analysis, whereas 4 (7%) were excluded because no data were recorded.

Correlation of Submission Parameters

To help determine which *distinct* use characteristics or parameters (such as the weekly frequency of BP data submissions) should be selected as features for clustering, the correlations among the parameters were calculated. Figure 1 shows the correlation matrix for all parameter weekly submissions (days). In this study, a moderate correlation (correlation coefficient between 0.3 to 0.7 and −0.7 to −0.3) [47,48] was chosen as the standard for selecting parameters. First, every participant received a BP monitor to measure BP, and pulse rate was collected as part of the BP measurement. Moreover, the correlation coefficient between BP and pulse rate was 0.93, a strong correlation. In this case, BP was selected for clustering rather than pulse rate. As for the other parameters, the correlations between BP and weight (0.51), PA (0.55), SR data (0.41), and sleep (0.55) were moderate, whereas the correlations between BP and SpO_2 level (0.05) and BG (0.24) were weak. In addition, the correlations between SpO_2 level and weight (−0.25), PA (0.16), SR data (0.29), and sleep (−0.24) were weak. Therefore, SpO_2 level was not selected for clustering. Likewise, the correlations between BG and weight (0.19), PA (0.2), SR data (−0.06), and sleep (0.25) were weak. Therefore, BG was not selected for clustering. Thus, BP, weight, PA, SR data, and sleep were selected for clustering.

XSL•FO
RenderX

Figure 1. Correlation matrix for weekly submissions (days) of all parameters. BG: blood glucose; BP: blood pressure; PA: physical activity; SpO2: peripheral oxygen saturation; SR: self-report.

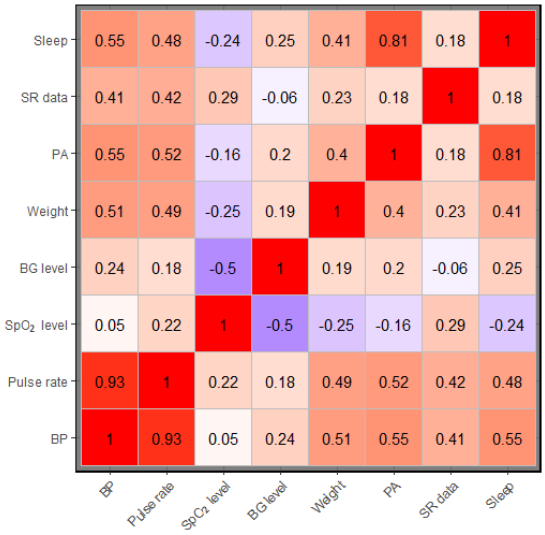

PCA and Clustering

The fundamental question for k-means clustering is this: how many clusters (k) should be discovered? To determine the optimum number of clusters, we further investigated the data through visualization offered by PCA. As can be seen from Figure 2, the first 2 principal components (PCs) explain 73.6% of the variation, which is an acceptably large percentage. However, after a check of individual contributions, we found that there were 3 participants—P038, P016, and P015—who contributed substantially to PC1 and PC2. After a check of the original data set, we found that P038 submitted symptom parameters only on 1 day, and P016 submitted symptom parameters only on 2 days. Conversely, P015 submitted parameters almost every day during the trial. Therefore, P038 and P016 were omitted from clustering.

After removing the outliers (P038 and P016), we found that the first 2 PCs explain 70.5% of the variation (Figure 3), which is an acceptably large percentage.

The clusters were projected into 2 dimensions as shown in Figure 4. Each subpart in Figure 4 shows a different number of clusters (k). When k=2, the data are obviously separated into 2 big clusters. Similarly, when k=3, the clusters are still separated very well into 3 clusters. When k=4, the clusters are well separated, but compared with the subpart with 3 clusters, 2 clusters are similar, whereas cluster 1, which only has 3 participants, is a relatively small cluster. When k=5, there is some overlap between cluster 1 and cluster 2. Likewise, Figure 5 shows the optimal number of clusters using the elbow method. In view of this, we determined that 3 clusters of participants separate the data set best. The 3 clusters can be labeled as the least engaged user group (cluster 1), the highly engaged user group (cluster 2), and the typical user group (cluster 3).

In the remainder of this section, we report on the examination of the clusters with respect to participant characteristics and the weekly submissions (days) of different parameters in a visual manner to reveal potential correlations and insights. Finally, we report on the examination of the correlations among all parameters by PCA.

Figure 2. The scree plot of every dimension by principal component analysis.

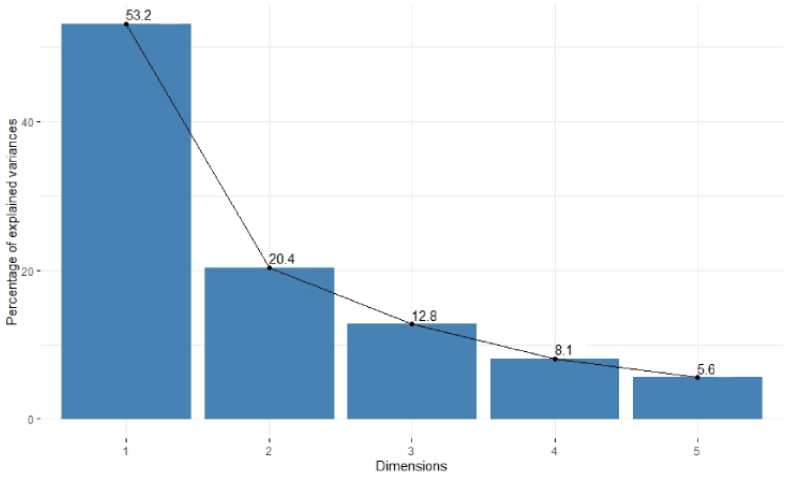

XSL•FO
RenderX

Figure 3. The scree plot of every dimension by principal component analysis (without the outliers).

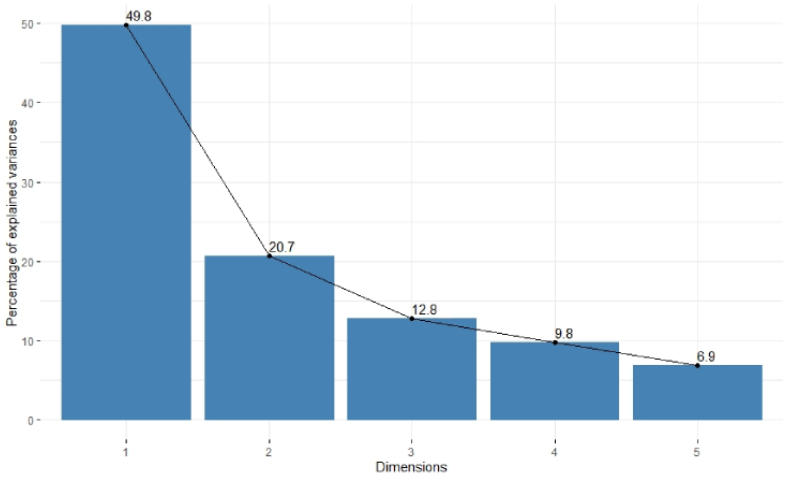

Figure 4. The visualization of clustering with the number of clusters (k) ranging from 2 to 5.

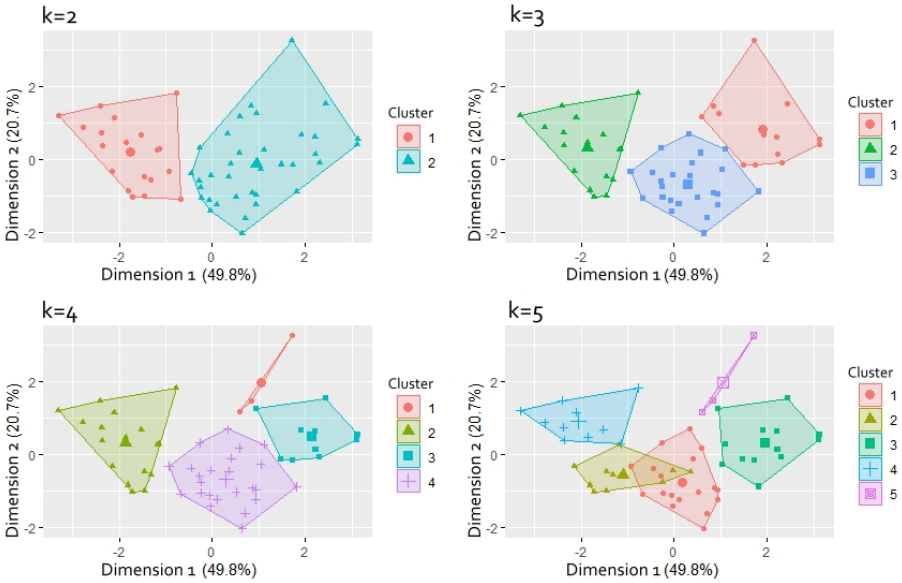

Figure 5. The optimal number of clusters by the elbow method.

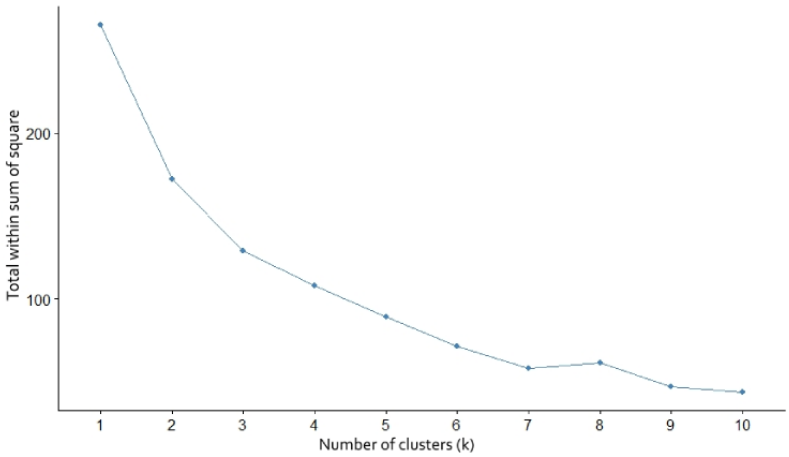

XSL•FO
RenderX

Participant Characteristics

As seen in Figure 6, the distribution of age within the 3 clusters is similar, with the *P* value of the 1-way ANOVA being .93, because all participants in this trial were older adults. However, the median age in the cluster 3 box plot is slightly higher than the median ages in the box plots of the other 2 clusters, and the average age of cluster 2 participants (74.1 years) is lower than that of cluster 1 (74.6 years) and cluster 3 (74.8 years; Table 2) participants. As Table 2 shows, 6 (26%) of the 23 female participants are in cluster 1 compared with 7 (23%) of the 31 male participants. However, the male participants in cluster 2 (10/31, 32%) and cluster 3 (14/31, 45%) represent higher proportions of total male participants compared with female participants in cluster 2 (7/23, 30%) and cluster 3 (10/23, 43%). Figure 7 shows the proportion of the 4 chronic health conditions within the 3 clusters. Cluster 1 has the largest proportion of participants with COPD and the smallest proportion of participants with diabetes. Moreover, cluster 3 has the smallest proportion of participants with HF (3/24, 13%; Table 2).

Figure 6. The variation in age within the 3 clusters based on the weekly submissions.

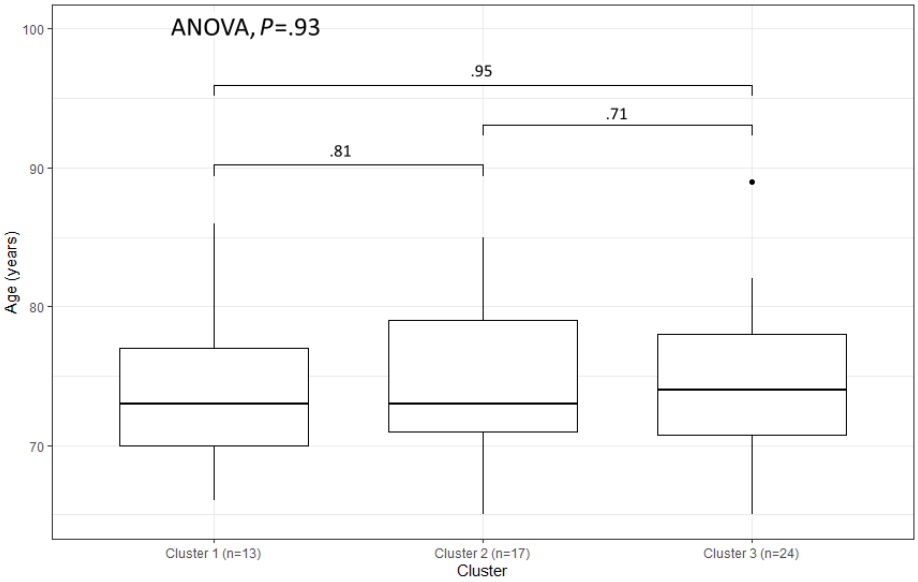

Table 2. Characteristics of the participants in each cluster (n=54).

Characteristics	Cluster 1 (n=13)	Cluster 2 (n=17)	Cluster 3 (n=24)
Age (y), mean (SD; range)	74.6 (6.2; 66-86)	74.1 (5.5; 65-85)	74.8 (5.9; 65-89)
Sex, n (%)			
Male	7 (23)	10 (32)	14 (45)
Female	6 (26)	7 (30)	10 (43)
Chronic health conditions, n (%)			
COPD[a]	9 (69)	4 (24)	9 (38)
Heart disease	11 (85)	16 (94)	22 (92)
HF[b]	4 (31)	4 (24)	3 (13)
Diabetes	4 (31)	11 (65)	14 (58)

[a]COPD: chronic obstructive pulmonary disease.

[b]HF: heart failure.

XSL•FO
RenderX

Figure 7. The variation in chronic health conditions within the 3 clusters. Each bar presents the percentage of each condition out of all conditions in the cluster (bearing in mind that participants can have multiple conditions); for example, there are 13 participants and 28 records under the 4 condition groups in cluster 1. Hence, chronic obstructive pulmonary disease (COPD) represents 32.1% of the conditions in cluster 1; however, of the 13 participants in cluster 1, a total of 9 (69%) have COPD, as presented in Table 2.

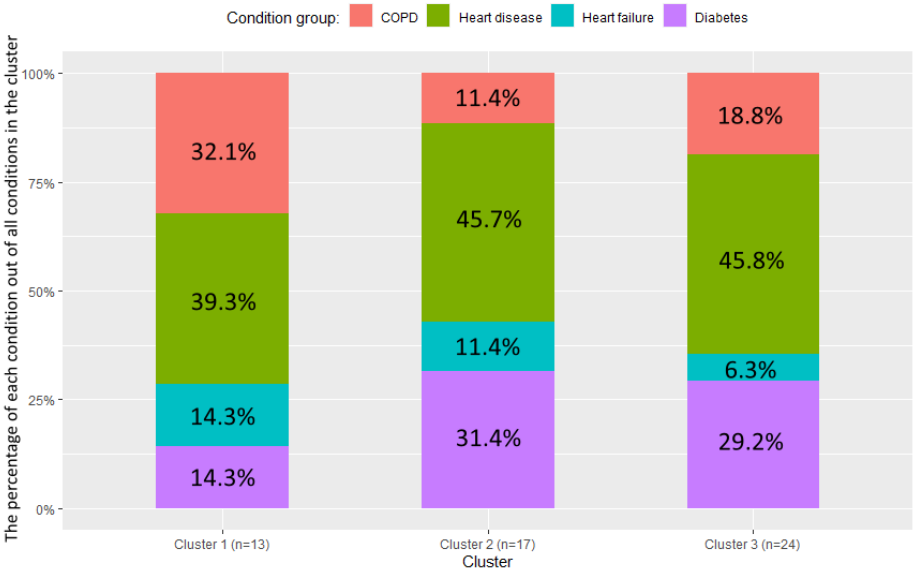

Participant Engagement Outcomes

Cluster 2 has the longest average enrollment time at 352 days compared with cluster 3 at 335 days and cluster 1 at 330 days. As seen in Figure 8, the overall distribution of the BP data weekly submissions is different, with the *P* value of the 1-way ANOVA being 8.4×10^{-9}. The frequency of BP data weekly submissions (days) of cluster 2 exceeds the frequencies of cluster 1 and cluster 3, which means that participants in cluster 2 have a higher frequency of BP data submissions than those in the other 2 clusters. The median and maximum of cluster 3 are higher than those of cluster 1, but the minimum of cluster 3 is lower than that of cluster 1. Likewise, as seen in Table 3, the mean and SD of cluster 1 (mean 2.5, SD 1.4) are smaller than those of cluster 3 (mean 2.9, SD 2.9).

As Figure 9 shows, the overall distribution of the weekly submissions of weight data is different, with the *P* value of the 1-way ANOVA being 1.4×10^{-13}, because the participants in cluster 2 submitted weight parameters more frequently than those in cluster 1 and cluster 3. In addition, similar to the BP data submissions, the median of cluster 3 is higher than that of cluster 1. As seen in Figure 9, there are 3 outliers in cluster 2. The top outlier is P015, who submitted a weight reading almost every day. During the trial, this participant mentioned many times in the interviews that his goal was to lose weight and that he used the scale to check his progress:

> *I've set out to reduce my weight. The doctor has been saying to me you know there's where you are and you should be over here. So, I've been using the weighing thing just to clock, to track reduction of weight.* [P015]

The other 2 outliers are P051 and P053, both of whom mentioned taking their weight measurements as part of their daily routine:

> *Once I get up in the morning the first thing is I weigh myself. That is, the day starts off with the weight, right.* [P053]

Although their frequency of weekly weight data submissions is lower than that of all other participants in cluster 2, it is still higher than that of most of the participants in the other 2 clusters.

In Table 3, it can be observed that the average frequency of weekly submissions of PA and sleep data for every cluster is higher than the frequencies of other variables, and the SDs are relatively low. This is likely because participants only needed to open the Withings app once a day to ensure the syncing of data. However, the overall distributions of PA and sleep data submissions are different in Figure 10 and Figure 11, with the *P* values of the 1-way ANOVA being 1.1×10^{-9} and 3.7×10^{-10}, respectively. Moreover, as Figure 10 and Figure 11 show, there are still some outliers who have a low frequency of submissions, and the box plot of cluster 1 is lower than the box plots of cluster 2 and cluster 3 in both figures. The reasons for the low frequency of submissions can mostly be explained by (1) technical issues, including internet connection issues, devices not syncing, and devices needing to be paired again; (2) participants forgetting to put the watch back on after taking it off; and (3) participants stopping using the devices (eg, some participants do not like wearing the watch while sleeping or when they go on holiday):

> *I was without my watch there for the last month or 3 or 4 weeks [owing to technical issues], and I missed it very badly because everything I look at the watch to tell the time, I was looking at my steps.* [P042]
>
> *I don't wear it, I told them I wouldn't wear the watch at night, I don't like it.* [P030]

Unlike in the case of other variables, the submission of SR data through the ProACT CareApp required participants to reflect on each question and their status before selecting the appropriate answer. Participants had different questions to answer based on

XSL•FO
RenderX

their health conditions; for example, participants with HF and COPD were asked to answer symptom-related questions, whereas those with diabetes were not. All participants were presented with general well-being and mood questions. Therefore, for some participants, self-reporting could possibly take more time than using the health monitoring devices. As shown in Table 3, the frequency of average weekly submissions of SR data within the 3 clusters is relatively small and the SDs are large, which means that the frequency of SR data submissions is lower than that of other variables. Furthermore, there were approximately 5 questions asked daily about general well-being, and some participants would skip the questions if they thought the question was unnecessary or not relevant:

Researcher: And do you answer your daily questions? P027: Yeah, once a week.

Researcher: Once a week, okay. P027: But they're the same.

As Figure 12 shows, the distribution of SR data submissions is different, with the *P* value of the 1-way ANOVA being .001. In Figure 12, the median of cluster 2 is higher than the medians of the other 2 clusters, and compared with other variables, but unlike other parameters, cluster 2 also has some participants who had very low SR data submission rates (close to 0). SR data is the only parameter where cluster 1 has a higher median than cluster 3.

Figure 8. The variation in weekly submissions (days) for blood pressure (BP) data within the 3 clusters.

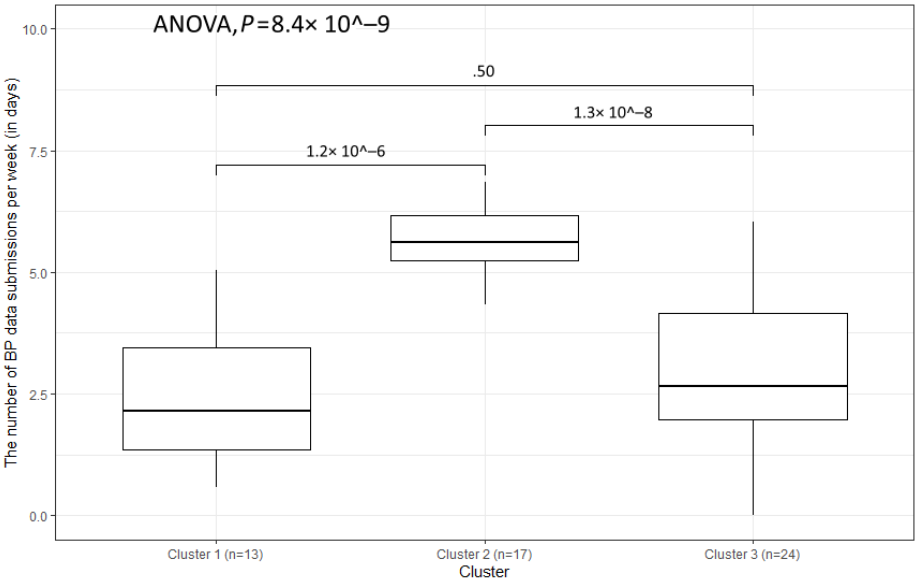

Table 3. Weekly submissions (days) of parameters.

Parameter	Cluster 1 (n=13), mean (SD)	Cluster 2 (n=17)	Cluster 3 (n=24)
Blood pressure	2.5 (1.4)[a]	5.7 (0.7)[b]	2.9 (1.6)
Weight	1.2 (0.9)[a]	5.4 (0.8)[b]	1.8 (1.5)
Physical activity	5.2 (0.7)[a]	6.7 (0.5)[b]	6.5 (0.4)
Self-report data	1.9 (1.4)	3.7 (2.1)[b]	1.6 (1.4)[a]
Sleep	4.2 (1.3)[a]	6.5 (0.4)[b]	6.1 (0.6)

[a]Lowest submission rate across the clusters.

[b]Highest submission rate across the clusters.

XSL•FO
RenderX

Figure 9. The variation in weekly submissions (days) for weight data within the 3 clusters.

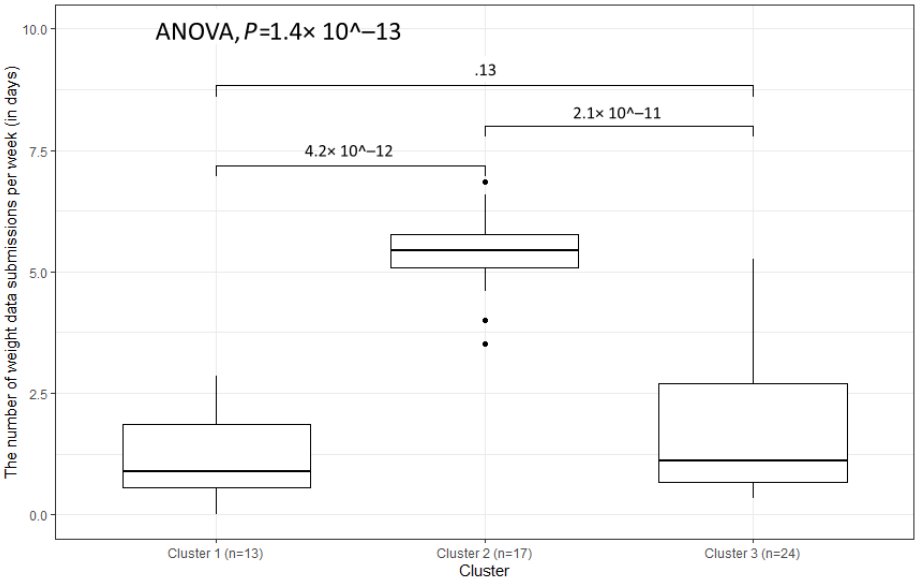

Figure 10. The variation in weekly submissions (days) for physical activity (PA) data within the 3 clusters.

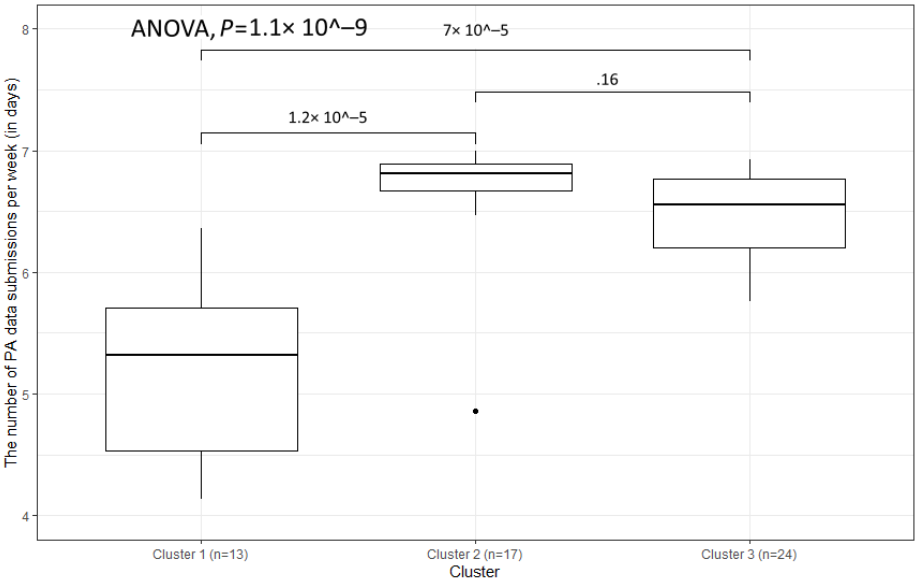

XSL•FO
RenderX

Figure 11. The variation in weekly submissions (days) for sleep data within the 3 clusters.

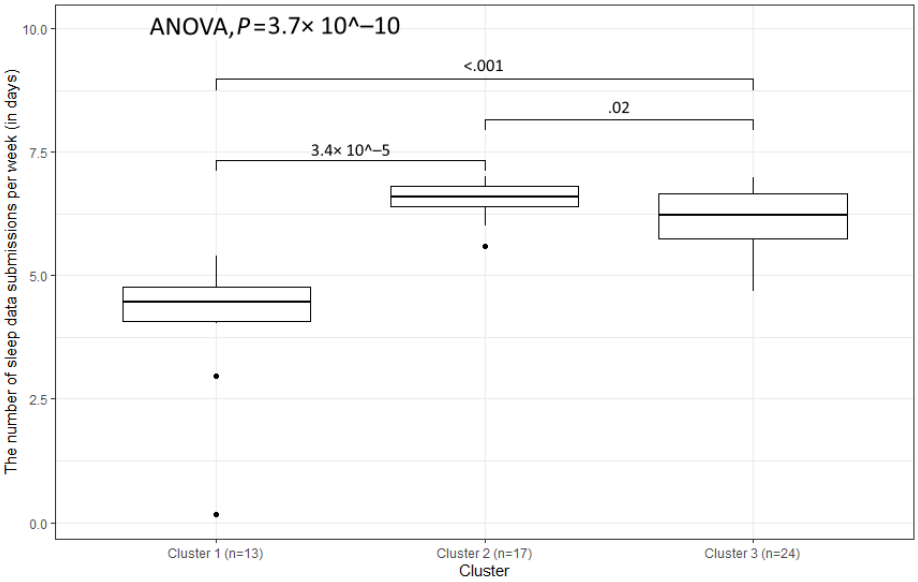

Figure 12. The variation in weekly submissions (days) for self-report (SR) data within the 3 clusters.

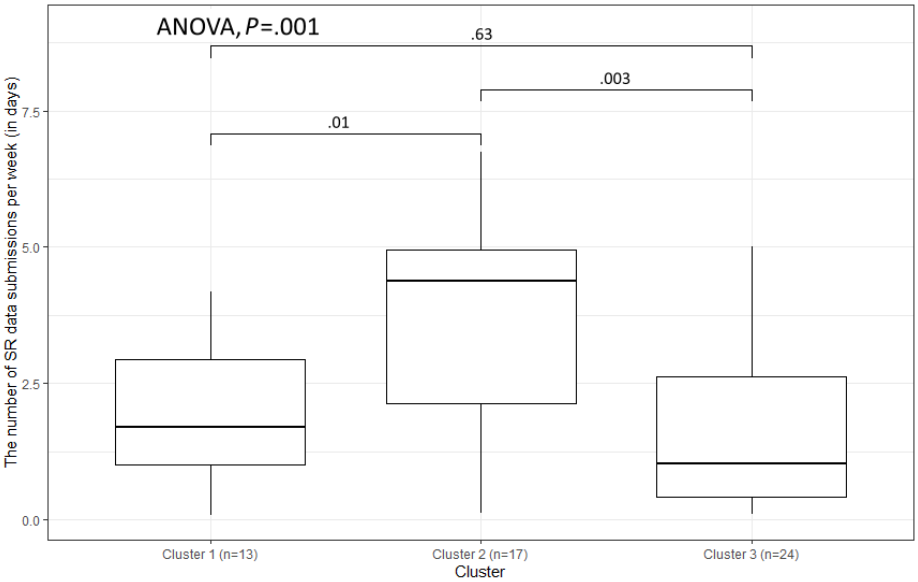

The Correlation Among the Weekly Submissions of Different Parameters

As seen in Figure 13, the arrows of BP and weight point to the same side of the plot, which shows a strong correlation. Likewise, PA and sleep also have a strong correlation. As noted previously, the strong correlation between PA and sleep is because the same device collected these 2 measurements, and participants only needed to sync the data once a day. By contrast, BP and weight were collected by 2 different devices but are strongly correlated. During interviews, many participants mentioned that their daily routine with the ProACT platform involved taking both BP and weight readings:

> *Usually in the morning when I get out of the bed, first, I go into the bathroom, wash my hands and come back, then weigh myself, do my blood pressure, do my bloods.* [P008]
>
> *I now have a routine that I let the system read my watch first thing, then I do my blood pressure thing and then I do the weight.* [P015]
>
> *As I said, it's keeping me in line with my, when I dip my finger, my weight, my blood pressure.* [P040]
>
> *I use it in the morning and at night for putting in the details of blood pressure in the morning and then the blood glucose at night. Yes, there's nothing else, is there? Oh, every morning the [weight] scales.* [P058]

By contrast, as shown in Figure 13, SR data have a weak correlation with other parameters, for reasons noted earlier.

XSL•FO
RenderX

Figure 13. The principal component analysis for variables. BP: blood pressure; PA: physical activity; SR: self-report.

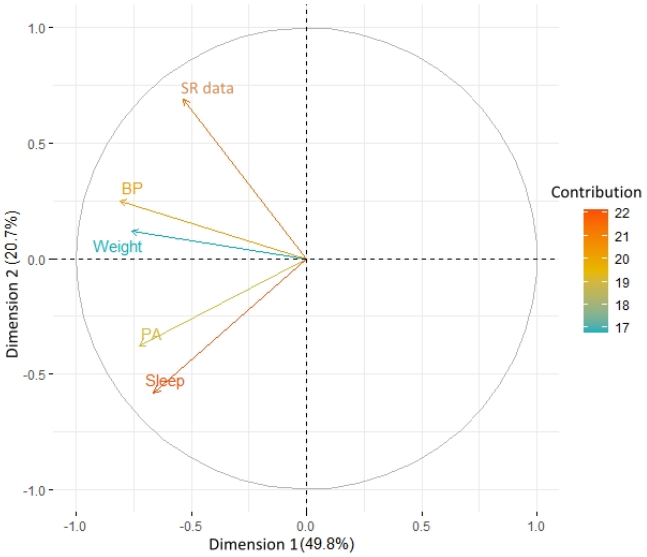

Parameter Variation Over Time

Analysis was conducted to determine any differences among the clusters in terms of symptom and well-being parameter changes over the course of the trial. Table 4 provides a description of each cluster in this regard. As Figure 14 shows, the box plot of cluster 2 is comparatively short in every time period of the trial, and the medians of cluster 2 and cluster 3 are more stable than the median of cluster 1. In addition, the median of cluster 1 is increasing over time, whereas the medians of cluster 2 and cluster 3 are decreasing and within the normal systolic BP of older adults [49] (Figure 14). As can be seen in Table 5, cluster 2 has a *P* value of .51 for systolic BP and a *P* value of .52 for diastolic BP, which are higher than the *P* values of cluster 1 (*P*=.19 and *P*=.16, respectively) and cluster 3 (*P*=.27 and *P*=.35, respectively). Therefore, participants in cluster 2, as highly engaged users, have more stable B*P* values than those in the other 2 clusters. By contrast, participants in cluster 1, as the least engaged users, have the most unstable B*P* values.

As seen in Figure 15, the median of cluster 2 is relatively higher than the medians of the other 2 clusters. The median of cluster 3 is increasing over time. In the second and third time periods of the trial, the box plot of cluster 1 is comparatively short. Normal SpO_2 levels are between 95% and 100%, but older adults may have SpO_2 levels closer to 95% [50]. In addition, for patients with COPD, SpO_2 levels range between 88% and 92% [51]. In this case, there is not much difference in terms of SpO_2 levels, and most of the SpO_2 levels are between 90% and 95% in this study. However, the SpO_2 levels of cluster 1 and cluster 2 were maintained at a relatively high level during the trial. As for cluster 3, the SpO_2 levels were comparatively low but relatively the same as those in the other 2 clusters in the later period of the trial. Therefore, the SpO_2 levels of cluster 3 (*P*=.25) are relatively unstable compared with those of cluster 1 (*P*=.66) and cluster 2 (*P*=.59). As such, there is little correlation between SpO_2 levels and engagement with digital health monitoring.

In relation to BG, Figure 16 shows that the box plot of cluster 2 is relatively lower than the box plots of the other 2 clusters in the second and third time periods. Moreover, the medians of cluster 2 and cluster 3 are lower than those of cluster 1 in the second and third time periods. The BG levels in cluster 2 and cluster 3 decreased at later periods of the trial compared with the beginning of the trial, but those in cluster 1 increased. Cluster 3 (*P*=.25), as the typical user group, had more significant change than cluster 1 (*P*=.50) and cluster 2 (*P*=.41). Overall, participants with a higher engagement rate had better BG control.

In relation to weight, Figure 17 shows that the box plot of cluster 2 is lower than the box plots of the other 2 clusters and comparatively short. As Table 5 shows, the *P* value of cluster 2 weight data is .72, which is higher than the *P* values of cluster 1 (.47) and cluster 3 (.61). Therefore, participants in cluster 2 had a relatively stable weight during the trial. In addition, as seen in Figure 17, the median weight of cluster 1 participants is decreasing, whereas that of cluster 3 participants is increasing. It is well known that there are many factors that can influence body weight, such as PA, diet, environmental factors, and so on. [52]. In this case, engagement with digital health and well-being monitoring may help control weight but the impact is not significant.

As Table 5 shows, the *P* value of cluster 2 PA (.049) is lower than .05, which means that there are significant differences among the 3 time slots in cluster 2. However, the median of cluster 2 PA, as seen in Figure 18, is still higher than the medians of the other 2 clusters. In cluster 2, approximately 50% of daily PA (steps) consists of >2500 steps. Overall, participants with a higher engagement rate also had a higher level of PA.

XSL•FO
RenderX

Table 4. The description of each cluster.

Cluster	Description	Label
Cluster 1	In cluster 1, each feature and submission rate are lower than those in the other 2 clusters, and cluster 1 has the least participants among the clusters. Typically, users have increasing systolic BP[a] over time, decreasing weight over time, and unstable BG[b] levels over time.	Least engaged user
Cluster 2	In cluster 2, every parameter's submission rate is higher than that in the other 2 clusters, the average submission rate is high, and the SDs of the submission rates are low except in the case of SR[c] data. Typically, users have stable BP over time, which is also within the recommended thresholds.	Highly engaged user
Cluster 3	In cluster 3, the submission rates for PA[d] and sleep are high, and the submission rates of the other 3 parameters are lower than those of cluster 2. However, cluster 3, which includes 44% (24/54) of the participants, is the largest cluster. The users' systolic BP usually decreases over time.	Typical user

[a]BP: blood pressure.

[b]BG: blood glucose.

[c]SR: self-report.

[d]PA: physical activity.

Figure 14. (A) The variation in systolic blood pressure in the 3 clusters among different time periods of the trial. (B) The variation in diastolic blood pressure in the 3 clusters among different time periods.

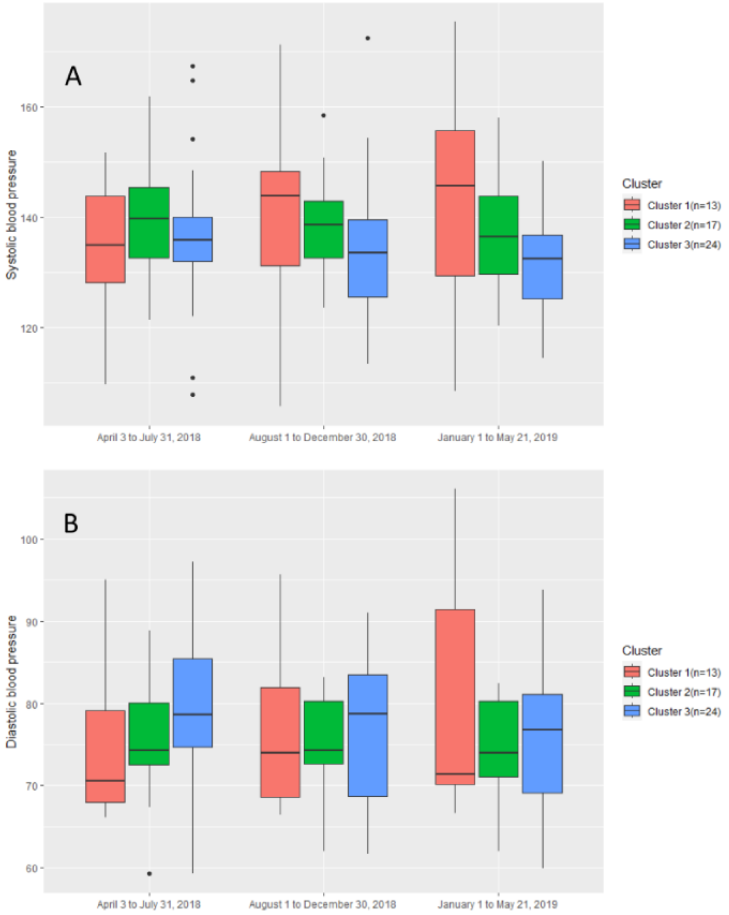

XSL•FO
RenderX

Table 5. The *P* value of each cluster among all time slots by 1-way ANOVA.

Cluster and parameters	*P* value
Cluster 1	
Systolic BP[a]	.19
Diastolic BP	.16
SpO_2[b]	.66
BG[c]	.50
Weight	.47
PA[d]	.68
Cluster 2	
Systolic BP	.51
Diastolic BP	.52
SpO_2	.59
BG	.41
Weight	.72
PA	.049
Cluster 3	
Systolic BP	.27
Diastolic BP	.35
SpO_2	.25
BG	.22
Weight	.61
PA	.86

[a]BP: blood pressure.
[b]SpO_2: peripheral oxygen saturation.
[c]BG: blood glucose.
[d]PA: physical activity.

Figure 15. The variation in peripheral oxygen saturation (SpO2) levels in the 3 clusters among different time periods.

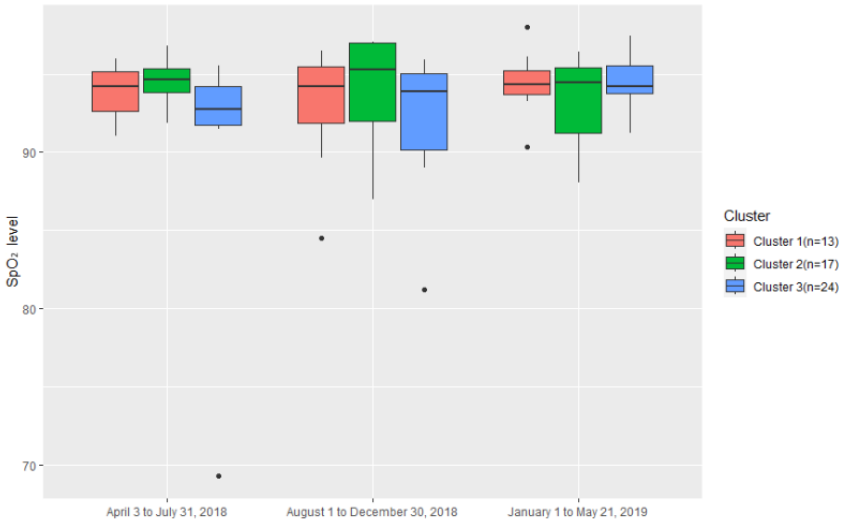

XSL•FO
RenderX

Figure 16. The variation in blood glucose levels in the 3 clusters among different time periods.

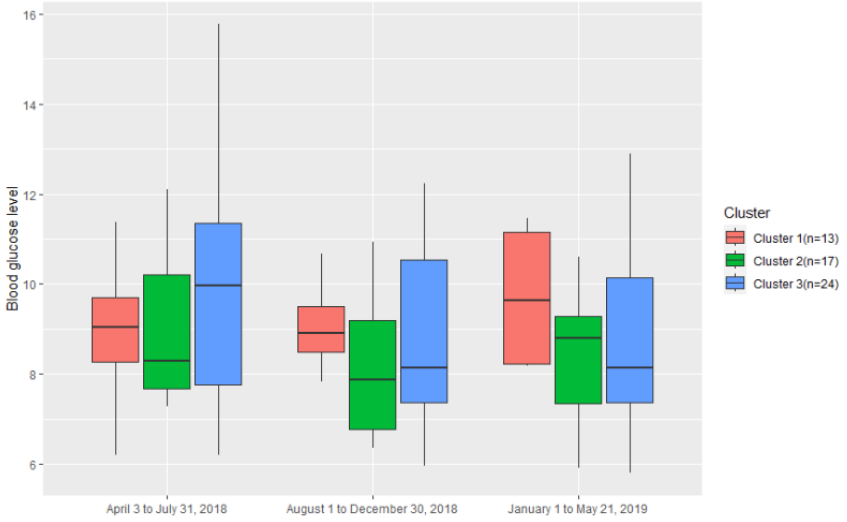

Figure 17. The variation in weight in the 3 clusters among different time periods.

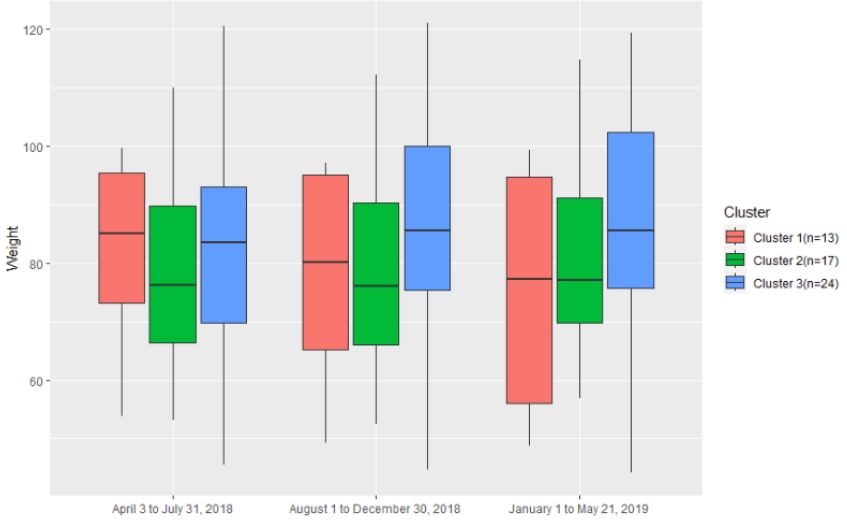

Figure 18. The variation in physical activity in the 3 clusters among different time periods.

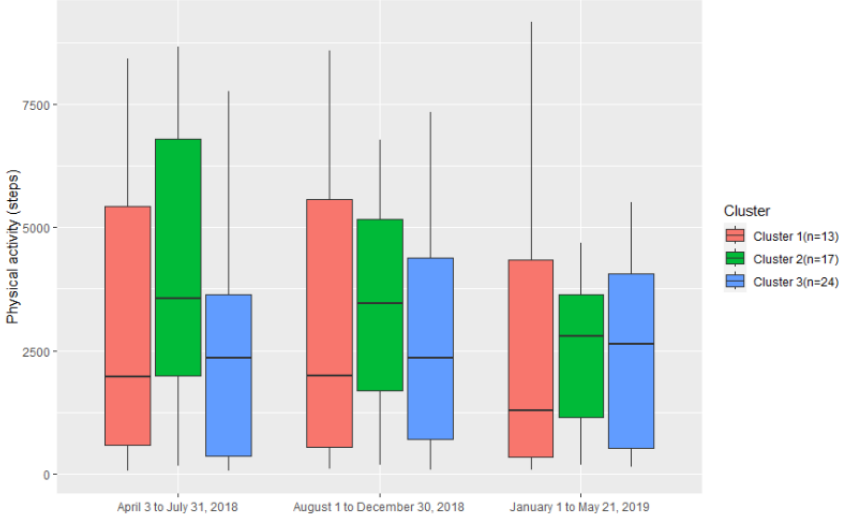

XSL•FO
RenderX

Discussion

Principal Findings

Digital health technologies hold great promise to help older adults with multimorbidity to improve health management and health outcomes. However, such benefits can only be realized if users engage with the technology. The aim of this study was to explore the engagement patterns of older adults with multimorbidity with digital self-management by using data mining to analyze users' weekly submission data. Three clusters were identified: cluster 1 (the least engaged user group), cluster 2 (the highly engaged user group), and cluster 3 (the typical user group). The subsequent analysis focused on how the clusters differ in terms of participant characteristics, patterns of engagement, and stabilization of health condition symptoms and well-being parameters over time, as well as how engagement rates with the different devices correlate with each other.

The key findings from the study are as follows:

- There is no significant difference in participants' characteristics among the clusters in general. The highly engaged group had the lowest average age (Table 4), and there was no significant difference with regard to sex and health conditions among these clusters. The least engaged user group had fewer male participants and participants with diabetes.
- There are 3 main factors influencing the correlations among the submission rates of different parameters. The first concerns whether the same device was used to submit the parameters, the second concerns the number of manual operations required to submit the parameter, and the third concerns the daily routine of the participants.
- Increased engagement with devices may improve the participants' health and well-being outcomes (eg, symptoms and PA levels). However, the difference between the highly engaged user group and the typical user group was relatively minimal compared with the difference between the highly engaged user group and the least engaged user group.

Each of these findings is discussed in further detail in the following subsections.

Although the findings presented in this paper focus on engagement based on the ProACT trial participants' use data, the interviews that were carried out as part of the trial identified additional potential factors of engagement. As reported in the study by Doyle et al [44], participants spoke about how they used the data to support their self-management (eg, taking action based on their data) and experienced various benefits, including increased knowledge of their health conditions and well-being, symptom optimization, reductions in weight, increased PA, and increased confidence to participate in certain activities as a result of health improvements. The peace of mind and encouragement provided by the clinical triage service as well as the technical support available were also identified during the interviews as potential factors positively impacting engagement [44]. In addition, the platform was found to be usable, and it imposed minimal burden on participants (Table 1). These findings supplement the quantitative findings presented in this paper.

Age, Sex, Health Condition Types, and Engagement

In this study, the difference in engagement with health care technologies between the sex was not significant. Of the 23 female participants, 6 (26%) were part of the least engaged user group compared with 7 (23%) of the 31 male participants. Moreover, there were lower proportions of female participants in the highly engaged user group (7/23, 30%) and typical user group (10/23, 43%) compared with male participants (10/31, 32% and 14/31, 45%, respectively). Other research has found that engagement with mobile health technology for BP monitoring was independent of sex [53]. However, there are also some studies that show that female participants are more likely to engage with digital mental health care interventions [54,55]. Therefore, sex cannot be considered as a separate criterion when comparing engagement with health care technologies, and it was not found to have significant impact on engagement in this study. Regarding age, many studies have shown that younger people are more likely to use health care technologies than older adults [56,57]. Although all participants in our study are older adults, the highly engaged user group is the youngest group. However, there was no significant difference in age among the clusters, with some of the oldest users being part of cluster 3, the typical user cluster. Similarly, the health conditions of a participant did not significantly impact their level of engagement. Other research [53] found that participants who were highly engaged with health monitoring had higher rates of hypertension, chronic kidney disease, and hypercholesterolemia than those with lower engagement levels. Our findings indicate that the highly engaged user group had a higher proportion of participants with diabetes, and the least engaged user group had a higher proportion of participants with COPD. Further research is needed to understand why there might be differences in engagement depending on health conditions. In our study, participants with COPD also self-reported on certain symptoms, such as breathlessness, chest tightness, and sputum amount and color. Although engagement with specific questions was not explored, participants in cluster 1, the least engaged user group, self-reported more frequently than those in cluster 3, the typical user group. Our findings also indicate that participants monitoring BG level and BP experienced better symptom stabilization over time than those monitoring SpO_2 level. It has been noted that the expected benefits of technology (eg, increased safety and usefulness) and need for technology (eg, subjective health status and perception of need) are 2 important factors that can influence the acceptance and use of technology by older adults [58]. It is also well understood that engaging in monitoring BG level can help people with diabetes to better self-manage and make decisions about diet, exercise, and medication [59].

Factors Influencing Engagement

Many research studies use *P* values to show the level of similarity or difference among clusters [60-63]. For most of the engagement outcomes in this study, all clusters significantly differed, with 1-way ANOVA $P<.001$, with the exception being SR data ($P=.001$). In addition, the 2-tailed *t* test *P* values showed that cluster 2 was significantly different from cluster 1 and cluster 3 in BP and weight data submission rates, whereas cluster 1 was significantly different from cluster 2 and cluster 3 in PA

XSL•FO
RenderX

and sleep data submission rates. As for SR data submission rates, all 3 two-tailed *t* tests had *P* values >.001, meaning that there were no significant differences between any 2 of these clusters. Therefore, all 5 parameters used for clustering were separated into 3 groups based on the correlations of submission rates: 1 for BP and weight, 1 for PA and sleep, and 1 for SR data. PA and sleep data submission rates have a strong correlation because participants used the same device to record daily PA and sleeping conditions. SR data submission rates have a weak correlation with other parameters' submission rates. Our previous research found that user retention in terms of submitting SR data was poorer than user retention in terms of using digital health devices, possibly because more manual operations are involved in the submission of SR data than other parameters or because the same questions were asked regularly, as noted by P027 in the *Participant Engagement Outcomes* subsection [64].

Other research that analyzed engagement with a diabetes support app found that user engagement was lower when more manual data entry was required [65]. In contrast to the other 2 groups of parameters, BP and weight data are collected using different devices. Whereas measuring BP requires using a BP monitor and manually synchronizing the data, measuring weight simply requires standing on the weight scale, and the data are automatically synchronized. Therefore, the manual operations involved in submitting BP and weight data are slightly different. However, the results showed a strong correlation between BP and weight because many participants preferred to measure both BP and weight together and incorporate taking these measurements into their daily routines. Research has indicated that if the use of a health care device becomes a regular routine, then participants will use it without consciously thinking about it [66]. Likewise, Yuan et al [67] note that integrating health apps into people's daily activities and forming regular habits can increase people's willingness to continue using the apps. However, participants using health care technology for long periods of time might become less receptive to exploring the system compared with using it based on the established methods to which they are accustomed [68]. In this study, many participants bundled their BP measurement with their weight measurement during their morning routine. Therefore, the engagement rates of interacting with these 2 devices were enhanced by each other. Future work could explore how to integrate additional measurements, such as monitoring SpO_2 level as well as self-reporting into this routine (eg, through prompting the user to submit these parameters while they are engaging with monitoring other parameters, such as BP and weight).

Relationship Between Engagement and Health and Well-Being Outcomes

Our third finding indicates that higher levels of engagement with digital health monitoring may result in better outcomes, such as symptom stabilization and increased PA levels. Milani et al [69] found that digital health care interventions can help people achieve BP control and improve hypertension control compared with usual care. In their study, users in the digital intervention group took an average of 4.2 readings a week. Compared with our study, this rate is lower than that of cluster 2 (5.7), the highly engaged user group, but higher than cluster 1 (2.5) and cluster 3 (2.9) rates. In our study, participants with a higher engagement rate experienced more stable BP, and for the majority of these participants (34/41, 83%), levels were maintained within the recommended thresholds of 140/90 mm Hg [70]. Many studies have shown that as engagement in digital diabetes interventions increases, patients will experience greater reductions in BG level compared with those with lower engagement [71,72]. However, in our study, BG levels in both the highly engaged user group (cluster 2) and the least engaged user group (cluster 1) increased in the later stages of the trial. Only the BG levels of the typical user group (cluster 3) decreased over time, which could be because the cluster 3 participants performed more PA in the later stages of the trial than during other time periods, as Figure 18 shows. Cluster 2, the highly engaged user group, maintained a relatively high level of PA during the trial period, although it continued to decline throughout the trial. Other research shows that more PA can also lead to better weight control and management [73,74], which could be 1 of the reasons why cluster 2 participants maintained their weight.

Limitations

There are some limitations to the research presented in this paper. First, although the sample size (n=60) was relatively large for a digital health study, the sample sizes for some parameters were small because not all participants monitored all parameters. Second, the participants were clustered based on weekly submissions of parameters only. If more features were included in clustering, such as submission intervals, participants could be grouped differently. It should also be pointed out that correlation is not a causality with respect to analyzing engagement rates with outcomes.

Conclusions

This study presents findings after the clustering of a data set that was generated from a longitudinal study of older adults using a digital health technology platform (ProACT) to self-manage multiple chronic health conditions. The highly engaged user group cluster (includes 17/54, 31% of users) had the lowest average age and highest frequency of submissions for every parameter. Engagement with digital health care technologies may also influence health and well-being outcomes (eg, symptoms and PA levels). The least engaged user group in our study had relatively poorer outcomes. However, the difference between the outcomes of the highly engaged user group and those of the typical user group is relatively small. There are 3 possible reasons for the correlations between the submission rates of parameters and devices. First, if 2 parameters are collected by the same device, they usually have a strong correlation, and users will engage with both equally. Second, the devices that involve fewer steps and parameters with less manual data entry will have a weak correlation with those devices that require more manual operations and data entry. Finally, participants' daily routines also influence the correlations among devices; for example, in this study, many participants had developed a daily routine to weigh themselves after measuring their BP, which led to a strong correlation

XSL•FO
RenderX

between BP and weight data submission rates. Future work should explore how to integrate the monitoring of additional parameters into a user's routine and whether additional characteristics, such as the severity of disease or technical proficiency, impact engagement.

Acknowledgments

This work was part funded by the Integrated Technology Systems for Proactive Patient Centred Care (ProACT) project and has received funding from the European Union (EU)–funded Horizon 2020 research and innovation program (689996). This work was part funded by the EU's INTERREG VA program, managed by the Special EU Programs Body through the Eastern Corridor Medical Engineering Centre (ECME) project. This work was part funded by the Scaling European Citizen Driven Transferable and Transformative Digital Health (SEURO) project and has received funding from the EU-funded Horizon 2020 research and innovation program (945449). This work was part funded by the COVID-19 Relief for Researchers Scheme set up by Ireland's Higher Education Authority. The authors would like to sincerely thank all the participants of this research for their valuable time.

Conflicts of Interest

None declared.

References

1. Ageing. United Nations. 2020. URL: https://www.un.org/en/global-issues/ageing [accessed 2022-01-13]
2. Centers for Disease Control and Prevention (CDC). Trends in aging--United States and worldwide. MMWR Morb Mortal Wkly Rep 2003 Feb 14;52(6):101-104 [FREE Full text] [Medline: 12645839]
3. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. Ann Fam Med 2009 Jul 13;7(4):357-363 [FREE Full text] [doi: 10.1370/afm.983] [Medline: 19597174]
4. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 2011 Sep;10(4):430-439. [doi: 10.1016/j.arr.2011.03.003] [Medline: 21402176]
5. Zhang L, Ma L, Sun F, Tang Z, Chan P. A multicenter study of multimorbidity in older adult inpatients in China. J Nutr Health Aging 2020 Mar;24(3):269-276 [FREE Full text] [doi: 10.1007/s12603-020-1311-x] [Medline: 32115607]
6. van der Heide I, Snoeijs S, Melchiorre MG, Quattrini S, Boerma W, Schellevis F, et al. Innovating care for people with multiple chronic conditions in Europe. Innovating Care for people with Multiple Chronic Conditions in Europe (ICARE4EU). 2015. URL: http://www.icare4eu.org/pdf/Innovating-care-for-people-with-multiple-chronic-conditions-in-Europe.pdf [accessed 2024-01-29]
7. Bartlett SJ, Lambert SD, McCusker J, Yaffe M, de Raad M, Belzile E, et al. Self-management across chronic diseases: targeting education and support needs. Patient Educ Couns 2020 Feb;103(2):398-404. [doi: 10.1016/j.pec.2019.08.038] [Medline: 31575442]
8. Anekwe TD, Rahkovsky I. Self-management: a comprehensive approach to management of chronic conditions. Am J Public Health 2018 Dec;108(S6):S430-S436. [doi: 10.2105/ajph.2014.302041r]
9. Barlow J, Wright C, Sheasby J, Turner A, Hainsworth J. Self-management approaches for people with chronic conditions: a review. Patient Educ Couns 2002;48(2):177-187. [doi: 10.1016/s0738-3991(02)00032-0] [Medline: 12401421]
10. Setiawan IM, Zhou L, Alfikri Z, Saptono A, Fairman AD, Dicianno BE, et al. An adaptive mobile health system to support self-management for persons with chronic conditions and disabilities: usability and feasibility studie. JMIR Form Res 2019 Apr 25;3(2):e12982 [FREE Full text] [doi: 10.2196/12982] [Medline: 31021324]
11. Alanzi T. mHealth for diabetes self-management in the Kingdom of Saudi Arabia: barriers and solutions. J Multidiscip Healthc 2018;11:535-546 [FREE Full text] [doi: 10.2147/JMDH.S174198] [Medline: 30349285]
12. Nunes F, Verdezoto N, Fitzpatrick G, Kyng M, Grönvall E, Storni C. Self-care technologies in HCI. ACM Trans Comput Hum Interact 2015 Dec 14;22(6):1-45. [doi: 10.1145/2803173]
13. Klasnja P, Kendall L, Pratt W, Blondon K. Long-term engagement with health-management technology: a dynamic process in diabetes. AMIA Annu Symp Proc 2015;2015:756-765 [FREE Full text] [Medline: 26958211]
14. Talboom-Kamp EP, Verdijk NA, Harmans LM, Numans ME, Chavannes NH. An eHealth platform to manage chronic disease in primary care: an innovative approach. Interact J Med Res 2016 Feb 09;5(1):e5 [FREE Full text] [doi: 10.2196/ijmr.4217] [Medline: 26860333]
15. Tighe SA, Ball K, Kensing F, Kayser L, Rawstorn JC, Maddison R. Toward a digital platform for the self-management of noncommunicable disease: systematic review of platform-like interventions. J Med Internet Res 2020 Oct 28;22(10):e16774 [FREE Full text] [doi: 10.2196/16774] [Medline: 33112239]
16. Pettersson B, Wiklund M, Janols R, Lindgren H, Lundin-Olsson L, Skelton DA, et al. 'Managing pieces of a personal puzzle' - older people's experiences of self-management falls prevention exercise guided by a digital program or a booklet. BMC Geriatr 2019 Feb 18;19(1):43 [FREE Full text] [doi: 10.1186/s12877-019-1063-9] [Medline: 30777026]
17. Kario K. Management of hypertension in the digital era: mall wearable monitoring devices for remote blood pressure monitoring. Hypertension 2020 Sep;76(3):640-650. [doi: 10.1161/hypertensionaha.120.14742]

XSL•FO
RenderX

18. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(2):64-72. [Medline: 15869215]
19. Alsayat A, El-Sayed H. Efficient genetic k-means clustering for health care knowledge discovery. In: Proceedings of the 14th International Conference on Software Engineering Research, Management and Applications. 2016 Presented at: SERA '16; June 8-10, 2016; Towson, MD p. 45-52 URL: https://ieeexplore.ieee.org/document/7516127 [doi: 10.1109/sera.2016.7516127]
20. Katsis Y, Balac N, Chapman D, Kapoor M, Block J, Griswold WG, et al. Big data techniques for public health: a case study. In: Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies. 2017 Presented at: CHASE '17; July 17-19, 2017; Philadelphia, PA p. 222-231 URL: https://ieeexplore.ieee.org/document/8010636 [doi: 10.1109/chase.2017.81]
21. Elbattah M, Molloy O. Data-driven patient segmentation using k-means clustering: the case of hip fracture care in Ireland. In: Proceedings of the 2017 Australasian Computer Science Week Multiconference. 2017 Presented at: ACSW '17; January 30- February 3, 2017; Geelong, Australia p. 1-8 URL: https://dl.acm.org/doi/10.1145/3014812.3014874 [doi: 10.1145/3014812.3014874]
22. Madigan EA, Curet OL. A data mining approach in home healthcare: outcomes and service use. BMC Health Serv Res 2006 Feb 24;6(1):18 [FREE Full text] [doi: 10.1186/1472-6963-6-18] [Medline: 16504115]
23. Armstrong JJ, Zhu M, Hirdes JP, Stolee P. K-means cluster analysis of rehabilitation service users in the home health care system of Ontario: examining the heterogeneity of a complex geriatric population. Arch Phys Med Rehabil 2012 Dec;93(12):2198-2205. [doi: 10.1016/j.apmr.2012.05.026] [Medline: 22705468]
24. Islam MS, Liu D, Wang K, Zhou P, Yu L, Wu D. A case study of healthcare platform using big data analytics and machine learning. In: Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference. 2019 Presented at: HPCCT '19; June 22-24, 2019; Guangzhou, China p. 139-146 URL: https://dl.acm.org/doi/10.1145/3341069.3342980 [doi: 10.1145/3341069.3342980]
25. Delias P, Doumpos M, Grigoroudis E, Manolitzas P, Matsatsinis N. Supporting healthcare management decisions via robust clustering of event logs. Knowl Based Syst 2015 Aug;84:203-213. [doi: 10.1016/j.knosys.2015.04.012]
26. Lefèvre T, Rondet C, Parizot I, Chauvin P. Applying multivariate clustering techniques to health data: the 4 types of healthcare utilization in the Paris metropolitan area. PLoS One 2014 Dec 15;9(12):e115064 [FREE Full text] [doi: 10.1371/journal.pone.0115064] [Medline: 25506916]
27. Ahmad P, Qamar S, Qasim Afser Rizvi S. Techniques of data mining in healthcare: a review. Int J Comput Appl 2015 Jun 18;120(15):38-50. [doi: 10.5120/21307-4126]
28. Mahoto NA, Shaikh FK, Ansari AQ. Exploitation of clustering techniques in transactional healthcare data. Mehran Univ Res J Eng Technol 2014;33(1):77-92.
29. Zahi S, Achchab B. Clustering of the population benefiting from health insurance using k-means. In: Proceedings of the 4th International Conference on Smart City Applications. 2019 Presented at: SCA '19; October 2-4, 2019; Casablanca, Morocco p. 1-6 URL: https://dl.acm.org/doi/abs/10.1145/3368756.3369103 [doi: 10.1145/3368756.3369103]
30. Jain AK. Data clustering: 50 years beyond k-means. Pattern Recognit Lett 2010 Jun;31(8):651-666. [doi: 10.1016/j.patrec.2009.09.011]
31. Silitonga P. Clustering of patient disease data by using k-means clustering. Int J Comput Sci Inf Sec 2017;15(7):219-221 [FREE Full text]
32. Shakeel PM, Baskar S, Dhulipala VR, Jaber MM. Cloud based framework for diagnosis of diabetes mellitus using k-means clustering. Health Inf Sci Syst 2018 Dec 24;6(1):16 [FREE Full text] [doi: 10.1007/s13755-018-0054-0] [Medline: 30279986]
33. Berry E, Davies M, Dempster M. Illness perception clusters and relationship quality are associated with diabetes distress in adults with type 2 diabetes. Psychol Health Med 2017 Oct 19;22(9):1118-1126. [doi: 10.1080/13548506.2017.1281976] [Medline: 28103702]
34. Harrison S, Robertson N, Graham C, Williams J, Steiner M, Morgan M, et al. Can we identify patients with different illness schema following an acute exacerbation of COPD: a cluster analysis. Respir Med 2014 Feb;108(2):319-328 [FREE Full text] [doi: 10.1016/j.rmed.2013.10.016] [Medline: 24238773]
35. Lopes AC, Xavier RF, Ac Pereira AC, Stelmach R, Fernandes FL, Harrison SL, et al. Identifying COPD patients at risk for worse symptoms, HRQoL, and self-efficacy: a cluster analysis. Chronic Illn 2019 Jun 17;15(2):138-148. [doi: 10.1177/1742395317753883] [Medline: 29343090]
36. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail 2019 Jan 17;21(1):74-85 [FREE Full text] [doi: 10.1002/ejhf.1333] [Medline: 30328654]
37. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract 2018 Jul 03;19(1):108 [FREE Full text] [doi: 10.1186/s12875-018-0790-x] [Medline: 29969997]
38. McCauley CO, Bond RB, Ryan A, Mulvenna MD, Laird L, Gibson A, et al. Evaluating user engagement with a reminiscence app using cross-comparative analysis of user event logs and qualitative data. Cyberpsychol Behav Soc Netw 2019 Aug;22(8):543-551. [doi: 10.1089/cyber.2019.0076] [Medline: 31403855]

XSL•FO
RenderX

39. Waterman H, Tillen D, Dickson R, de Koning K. Action research: a systematic review and guidance for assessment. Health Technol Assess 2001;5(23):iii-157 [FREE Full text] [Medline: 11785749]
40. Bashshur RL, Shannon GW, Smith BR, Alverson DC, Antoniotti N, Barsan WG, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014 Sep;20(9):769-800 [FREE Full text] [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
41. Dinsmore J, Hannigan C, Smith S, Murphy E, Kuiper JM, O'Byrne E, et al. A digital health platform for integrated and proactive patient-centered multimorbidity self-management and care (ProACT): protocol for an action research proof-of-concept trial. JMIR Res Protoc 2021 Dec 15;10(12):e22125 [FREE Full text] [doi: 10.2196/22125] [Medline: 34914613]
42. Doyle J, Murphy E, Kuiper J, Smith S, Hannigan C, Jacobs A, et al. Managing multimorbidity: identifying design requirements for a digital self-management tool to support older adults with multiple chronic conditions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, Scotland p. 1-14 URL: https://dl.acm.org/doi/10.1145/3290605.3300629 [doi: 10.1145/3290605.3300629]
43. Doyle J, Murphy E, Hannigan C, Smith S, Bettencourt-Silva J, Dinsmore J. Designing digital goal support systems for multimorbidity self-management: insights from older adults and their care network. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: PervasiveHealth '18; May 21-24, 2018; New York, NY p. 168-177 URL: https://dl.acm.org/doi/10.1145/3240925.3240982 [doi: 10.1145/3240925.3240982]
44. Doyle J, Murphy E, Gavin S, Pascale A, Deparis S, Tommasi P, et al. A digital platform to support self-management of multiple chronic conditions (ProACT): findings in relation to engagement during a one-year proof-of-concept trial. J Med Internet Res 2021 Dec 15;23(12):e22672 [FREE Full text] [doi: 10.2196/22672] [Medline: 34914612]
45. Doyle J, McAleer P, van Leeuwen C, Smith S, Murphy E, Sillevis Smitt M, et al. The role of phone-based triage nurses in supporting older adults with multimorbidity to digitally self-manage - findings from the ProACT proof-of-concept study. Digit Health 2022 Oct 09;8:20552076221131140 [FREE Full text] [doi: 10.1177/20552076221131140] [Medline: 36238753]
46. Ross A, Willson VL. One-way anova. In: Ross A, Willson VL, editors. Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures. Cham, Switzerland: Springer; 2017:21-24.
47. Dancey CP, Reidy J. Statistics without Maths for Psychology. Upper Saddle River, NJ: Prentice Hall; 2007.
48. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018 Sep;18(3):91-93 [FREE Full text] [doi: 10.1016/j.tjem.2018.08.001] [Medline: 30191186]
49. Master AM, Dublin LI, Marks HH. The normal blood pressure range and its clinical implications. J Am Med Assoc 1950 Aug 26;143(17):1464-1470. [doi: 10.1001/jama.1950.02910520006004] [Medline: 15428295]
50. Cunha JP. What is a good oxygen rate by age? eMedicineHealth. URL: https://www.emedicinehealth.com/what_is_a_good_oxygen_rate_by_age/article_em.htm [accessed 2024-01-29]
51. Echevarria C, Steer J, Wason J, Bourke S. Oxygen therapy and inpatient mortality in COPD exacerbation. Emerg Med J 2021 Mar 26;38(3):170-177. [doi: 10.1136/emermed-2019-209257] [Medline: 33243839]
52. Atkinson Jr RL, Butterfield G, Dietz W, Fernstrom J, Frank A, Hansen B. Weight Management: State of the Science and Opportunities for Military Programs. Washington, DC: National Academies Press; 2003.
53. Kaplan AL, Cohen ER, Zimlichman E. Improving patient engagement in self-measured blood pressure monitoring using a mobile health technology. Health Inf Sci Syst 2017 Dec 07;5(1):4 [FREE Full text] [doi: 10.1007/s13755-017-0026-9] [Medline: 29081974]
54. Mikolasek M, Witt CM, Barth J. Adherence to a mindfulness and relaxation self-care app for cancer patients: mixed-methods feasibility study. JMIR Mhealth Uhealth 2018 Dec 06;6(12):e11271 [FREE Full text] [doi: 10.2196/11271] [Medline: 30522990]
55. Harjumaa M, Halttu K, Koistinen K, Oinas-Kukkonen H. User experience of mobile coaching for stress-management to tackle prevalent health complaints. In: Proceedings of the 6th Scandinavian Conference on Information Systems. 2015 Presented at: SCIS '15; August 9-12, 2015; Oulu, Finland URL: https://cris.vtt.fi/en/publications/user-experience-of-mobile-coaching-for-stress-management-to-tackl [doi: 10.1007/978-3-319-21783-3_11]
56. Kannisto KA, Korhonen J, Adams CE, Koivunen MH, Vahlberg T, Välimäki MA. Factors associated with dropout during recruitment and follow-up periods of a mHealth-based randomized controlled trial for mobile.net to encourage treatment adherence for people with serious mental health problems. J Med Internet Res 2017 Feb 21;19(2):e46 [FREE Full text] [doi: 10.2196/jmir.6417] [Medline: 28223262]
57. Abel EA, Shimada SL, Wang K, Ramsey C, Skanderson M, Erdos J, et al. Dual use of a patient portal and clinical video telehealth by veterans with mental health diagnoses: retrospective, cross-sectional analysis. J Med Internet Res 2018 Nov 07;20(11):e11350 [FREE Full text] [doi: 10.2196/11350] [Medline: 30404771]
58. Peek ST, Wouters EJ, van Hoof J, Luijkx KG, Boeije HR, Vrijhoef HJ. Factors influencing acceptance of technology for aging in place: a systematic review. Int J Med Inform 2014 Apr;83(4):235-248 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.01.004] [Medline: 24529817]
59. Weinstock RS, Aleppo G, Bailey TS, Bergenstal RM, Fisher WA, Greenwood DA, et al. The role of blood glucose monitoring in diabetes management. Compendia 2022 Oct;2020(3):1-32. [doi: 10.2337/db2020-31] [Medline: 33411424]

XSL•FO
RenderX

60. Rahman QA, Janmohamed T, Pirbaglou M, Ritvo P, Heffernan JM, Clarke H, et al. Patterns of user engagement with the mobile app, manage my pain: results of a data mining investigation. JMIR Mhealth Uhealth 2017 Jul 12;5(7):e96 [FREE Full text] [doi: 10.2196/mhealth.7871] [Medline: 28701291]
61. Booth FG, R Bond R, D Mulvenna M, Cleland B, McGlade K, Rankin D, et al. Discovering and comparing types of general practitioner practices using geolocational features and prescribing behaviours by means of K-means clustering. Sci Rep 2021 Sep 14;11(1):18289 [FREE Full text] [doi: 10.1038/s41598-021-97716-3] [Medline: 34521920]
62. Sulistyono MT, Pane ES, Wibawa AD, Purnomo MH. Analysis of EEG-based stroke severity groups clustering using k-means. In: Proceedings of the 2021 International Seminar on Intelligent Technology and Its Applications. 2021 Presented at: ISITIA '21; July 21-22, 2021; Surabaya, Indonesia p. 67-74 URL: https://ieeexplore.ieee.org/document/9502250 [doi: 10.1109/isitia52817.2021.9502250]
63. Oskooei A, Chau SM, Weiss J, Sridhar A, Martínez MR, Michel B. DeStress: deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (HRV) data. In: Shaban-Nejad A, Michalowski M, Buckeridge DL, editors. Explainability and Interpretability: Keys to Deep Medicine. Cham, Switzerland: Springer; 2020:93-105.
64. Sheng Y, Doyle J, Bond R, Jaiswal R, Gavin S, Dinsmore J. Home-based digital health technologies for older adults to self-manage multiple chronic conditions: a data-informed analysis of user engagement from a longitudinal trial. Digit Health 2022 Sep 22;8:20552076221125957 [FREE Full text] [doi: 10.1177/20552076221125957] [Medline: 36171962]
65. Böhm AK, Jensen ML, Sørensen MR, Stargardt T. Real-world evidence of user engagement with mobile health for diabetes management: longitudinal observational study. JMIR Mhealth Uhealth 2020 Nov 06;8(11):e22212 [FREE Full text] [doi: 10.2196/22212] [Medline: 32975198]
66. Kim SS, Malhotra NK. A longitudinal model of continued is use: an integrative view of four mechanisms underlying postadoption phenomena. Manag Sci 2005 May;51(5):741-755. [doi: 10.1287/mnsc.1040.0326]
67. Yuan S, Ma W, Kanthawala S, Peng W. Keep using my health apps: discover users' perception of health and fitness apps with the UTAUT2 model. Telemed J E Health 2015 Sep;21(9):735-741. [doi: 10.1089/tmj.2014.0148] [Medline: 25919238]
68. O'Connor Y, O'Reilly P, O'Donoghue J. M-health infusion by healthcare practitioners in the national health services (NHS). Health Policy Technol 2013 Mar;2(1):26-35. [doi: 10.1016/j.hlpt.2012.12.002]
69. Milani RV, Lavie CJ, Bober RM, Milani AR, Ventura HO. Improving hypertension control and patient engagement using digital tools. Am J Med 2017 Jan;130(1):14-20. [doi: 10.1016/j.amjmed.2016.07.029] [Medline: 27591179]
70. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, ESC Scientific Document Group. 2018 ESC/ESH guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). Eur Heart J 2018 Sep 01;39(33):3021-3104 [FREE Full text] [doi: 10.1093/eurheartj/ehy339] [Medline: 30165516]
71. Quinn CC, Butler EC, Swasey KK, Shardell MD, Terrin MD, Barr EA, et al. Mobile diabetes intervention study of patient engagement and impact on blood glucose: mixed methods analysis. JMIR Mhealth Uhealth 2018 Feb 02;6(2):e31 [FREE Full text] [doi: 10.2196/mhealth.9265] [Medline: 29396389]
72. Sepah SC, Jiang L, Ellis RJ, McDermott K, Peters AL. Engagement and outcomes in a digital diabetes prevention program: 3-year update. BMJ Open Diabetes Res Care 2017 Sep 07;5(1):e000422 [FREE Full text] [doi: 10.1136/bmjdrc-2017-000422] [Medline: 28948027]
73. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? a secondary data analytics approach. J Med Internet Res 2017 Apr 19;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
74. Demark-Wahnefried W, Schmitz KH, Alfano CM, Bail JR, Goodwin PJ, Thomson CA, et al. Weight management and physical activity throughout the cancer care continuum. CA Cancer J Clin 2018 Jan;68(1):64-89 [FREE Full text] [doi: 10.3322/caac.21441] [Medline: 29165798]

Abbreviations

BG: blood glucose
BP: blood pressure
CABIE-SIMS: Context-Aware Broker and Inference Engine–Subject Information Management System
COPD: chronic obstructive pulmonary disease
HD: heart disease
HF: heart failure
PC: principal component
PCA: principal component analysis
ProACT: Integrated Technology Systems for Proactive Patient Centred Care
SpO2: peripheral oxygen saturation
SR: self-report

XSL•FO
RenderX

Edited by T Leung, T de Azevedo Cardoso; submitted 05.02.23; peer-reviewed by B Chaudhry, M Peeples, A DeVito Dabbs; comments to author 12.09.23; revised version received 25.10.23; accepted 29.01.24; published 28.03.24.

Please cite as:
Sheng Y, Bond R, Jaiswal R, Dinsmore J, Doyle J
Augmenting K-Means Clustering With Qualitative Data to Discover the Engagement Patterns of Older Adults With Multimorbidity When Using Digital Health Technologies: Proof-of-Concept Trial
J Med Internet Res 2024;26:e46287
URL: https://www.jmir.org/2024/1/e46287
doi:10.2196/46287
PMID:38546724

©Yiyang Sheng, Raymond Bond, Rajesh Jaiswal, John Dinsmore, Julie Doyle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Service Quality and Residents' Preferences for Facilitated Self-Service Fundus Disease Screening: Cross-Sectional Study

Senlin Lin[1,2,3*], MSc; Yingyan Ma[1,2,3,4*], PhD; Yanwei Jiang[5*], MPH; Wenwen Li[6], PhD; Yajun Peng[1,2,3], BA; Tao Yu[1,2,3], BA; Yi Xu[1,2,3], MD; Jianfeng Zhu[1,2,3], MD; Lina Lu[1,2,3], MPH; Haidong Zou[1,2,3,4], MD

[1]Shanghai Eye Diseases Prevention &Treatment Center/ Shanghai Eye Hospital, School of Medicine, Tongji University, Shanghai, China

[2]National Clinical Research Center for Eye Diseases, Shanghai, China

[3]Shanghai Engineering Research Center of Precise Diagnosis and Treatment of Eye Diseases, Shanghai, China

[4]Shanghai General Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[5]Shanghai Hongkou Center for Disease Control and Prevention, Shanghai, China

[6]School of Management, Fudan University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Haidong Zou, MD
Shanghai Eye Diseases Prevention &Treatment Center/ Shanghai Eye Hospital
School of Medicine
Tongji University
No 1440, Hongqqiao Road
Shanghai, 200336
China
Phone: 86 02162539696
Email: zouhaidong@sjtu.edu.cn

Abstract

Background: Fundus photography is the most important examination in eye disease screening. A facilitated self-service eye screening pattern based on the fully automatic fundus camera was developed in 2022 in Shanghai, China; it may help solve the problem of insufficient human resources in primary health care institutions. However, the service quality and residents' preference for this new pattern are unclear.

Objective: This study aimed to compare the service quality and residents' preferences between facilitated self-service eye screening and traditional manual screening and to explore the relationships between the screening service's quality and residents' preferences.

Methods: We conducted a cross-sectional study in Shanghai, China. Residents who underwent facilitated self-service fundus disease screening at one of the screening sites were assigned to the exposure group; those who were screened with a traditional fundus camera operated by an optometrist at an adjacent site comprised the control group. The primary outcome was the screening service quality, including effectiveness (image quality and screening efficiency), physiological discomfort, safety, convenience, and trustworthiness. The secondary outcome was the participants' preferences. Differences in service quality and the participants' preferences between the 2 groups were compared using chi-square tests separately. Subgroup analyses for exploring the relationships between the screening service's quality and residents' preference were conducted using generalized logit models.

Results: A total of 358 residents enrolled; among them, 176 (49.16%) were included in the exposure group and the remaining 182 (50.84%) in the control group. Residents' basic characteristics were balanced between the 2 groups. There was no significant difference in service quality between the 2 groups (image quality pass rate: P=.79; average screening time: P=.57; no physiological discomfort rate: P=.92; safety rate: P=.78; convenience rate: P=.95; trustworthiness rate: P=.20). However, the proportion of participants who were willing to use the same technology for their next screening was significantly lower in the exposure group than in the control group (P<.001). Subgroup analyses suggest that distrust in the facilitated self-service eye screening might increase the probability of refusal to undergo screening (P=.02).

Conclusions: This study confirms that the facilitated self-service fundus disease screening pattern could achieve good service quality. However, it was difficult to reverse residents' preferences for manual screening in a short period, especially when the original manual service was already excellent. Therefore, the digital transformation of health care must be cautious. We suggest

XSL•FO
RenderX

that attention be paid to the residents' individual needs. More efficient man-machine collaboration and personalized health management solutions based on large language models are both needed.

(J Med Internet Res 2024;26:e45545) doi:10.2196/45545

KEYWORDS

digital technology; screening; self-service; eye disease; health economics evaluation; health technology assessment; disease screening; artificial intelligence; AI; eye; community; effectiveness; screening efficiency; safety

Introduction

Vision impairment and blindness are caused by a variety of eye diseases, including cataracts, glaucoma, uncorrected refractive error, age-related macular degeneration, diabetic retinopathy, and other eye diseases [1]. They not only reduce economic productivity but also harm the quality of life and increase mortality [2-6]. In 2020, an estimated 43.3 million individuals were blind, and 1.06 billion individuals aged 50 years and older had distance or near vision impairment [7]. With an increase in the aging population, the number of individuals affected by vision loss has increased substantially [1].

High-quality public health care for eye disease prevention, such as effective screening, can assist in eliminating approximately 57% of all blindness cases [8]. Digital technologies, such as telemedicine, 5G telecommunications, the Internet of Things, and artificial intelligence (AI), have provided the potential to improve the accessibility, availability, and productivity of existing resources and the overall efficiency of eye care services [9,10]. The use of digital technology not only reduces the cost of eye disease screening and improves its efficiency, but also assists residents living in remote areas to gain access to eye disease screening [11-13]. Therefore, an increasing number of countries (or regions) are attempting to establish eye screening systems based on digital technology [9].

Fundus photography is the most important examination in eye disease screening because the vast majority of diagnoses of blinding retinal diseases are based on fundus photographs. Diagnoses can be made by human experts or AI software. However, traditional fundus cameras must be operated by optometrists, who are usually in short supply in primary health care institutions when faced with the large demand for screening services.

Fortunately, the fully automatic fundus camera has been developed on the basis of digital technologies including AI, industrial automation, sensors, and voice navigation. It can automatically identify the person's left and right eyes, search for pupils, adjust the lens position and shooting focus, and provide real-time voice feedback during the process, helping the residents to understand the current inspection steps clearly and cooperatively complete the inspection. Therefore, a facilitated self-service eye screening pattern has been newly established in 2022 in Shanghai, China.

However, evidence is inadequate about whether this new screening pattern performs well and whether the residents prefer it. Therefore, this cross-sectional study aims to compare the service quality and residents' preferences of this new screening pattern with that of the traditional screening pattern. We aimed to (1) investigate whether the facilitated self-service eye screening can achieve service quality similar to that of traditional manual screening, (2) compare residents' preferences between the facilitated self-service eye screening and traditional manual screening, and (3) explore the relationship between the screening service quality and residents' preferences.

Methods

Study Setting

This study was conducted in Shanghai, China, in 2022. Since 2010, Shanghai has conducted an active community-based fundus disease telemedicine screening program. After 2018, an AI model was adopted (Figure 1). At the end of 2021, the fully automatic fundus camera was adopted, and the facilitated self-service fundus disease screening pattern was established (Figure 1). Within this new pattern, residents could perform fundus photography by themselves without professionals' assistance (Multimedia Appendix 1). The fundus images were sent to the cloud server center of the AI model, and the screening results were fed back immediately.

XSL•FO
RenderX

Figure 1. Comparison of the facilitated self-service fundus disease screening pattern and the traditional telemedicine screening pattern. (A) Workflow of the traditional telemedicine screening pattern. (B) Workflow of the facilitated self-service eye screening pattern. (C) Actual image of a Vision Health Intelligent Management Center. It is a site built in the communities for residents to receive eye disease screening and health management, with the facilitated self-service eye screening pattern. Three fully automatic self-service ophthalmic examination devices have been equipped. The device on the left in the photo is a fully automatic self-service visual acuity, the one in the middle is a fully automatic self-service computer optometry device, and the one on the right one is a fully automatic self-service nonmydriatic fundus camera. AI: artificial intelligence.

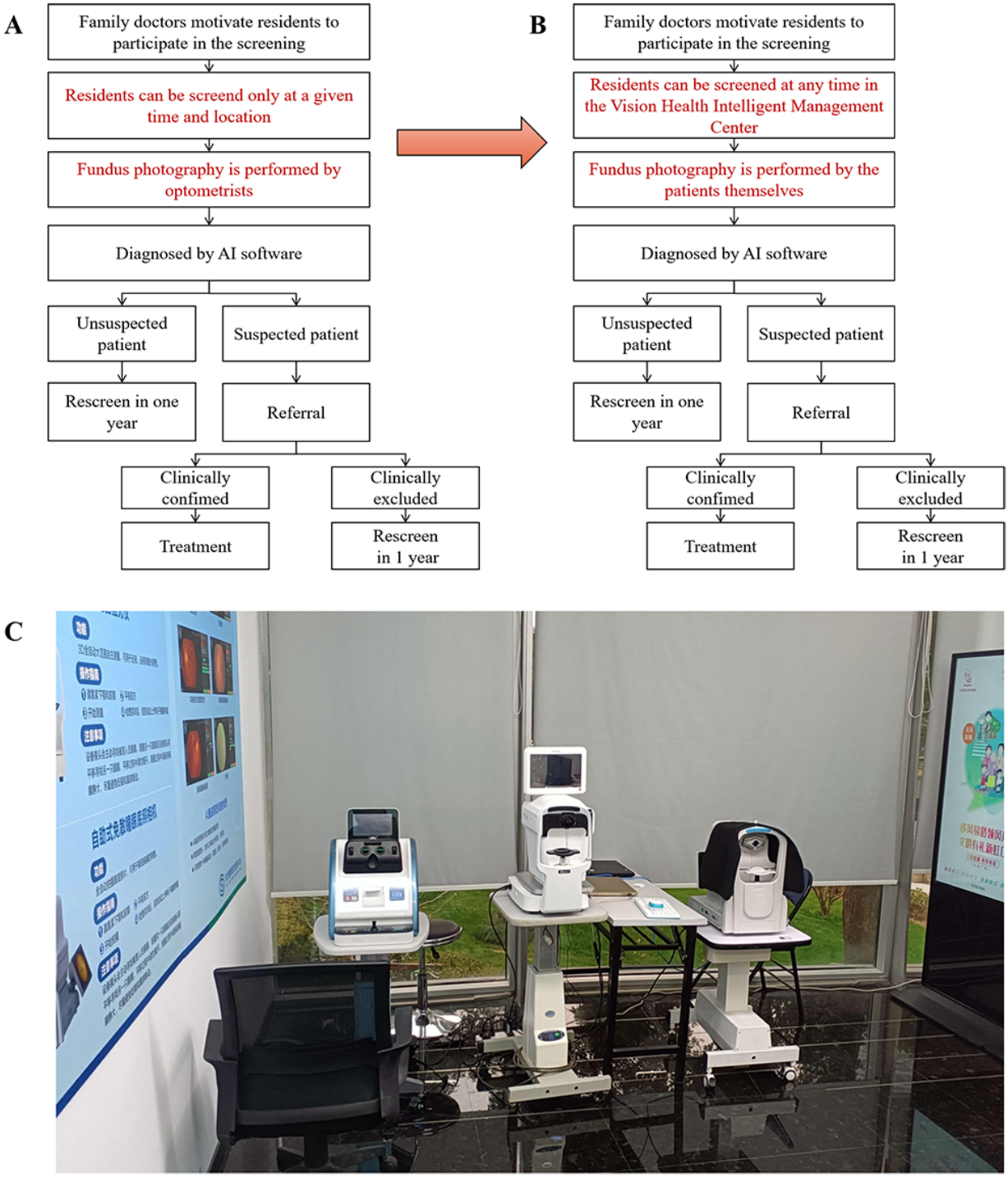

Study Design

We conducted a cross-sectional study at 2 adjacent screening sites. These 2 sites were expected to be very similar in terms of their socioeconomic and educational aspects since they were located next to each other. One site provided facilitated self-service fundus disease screening, and the residents who participated therein comprised the exposure group; the other site provided screening with a traditional fundus camera operated by an optometrist, and the residents who participated therein comprised the control group. All the adult residents could participant in our screening program, but their data were used for analysis only if they signed the informed consent form. Residents could opt out of the study at any time during the screening.

XSL•FO
RenderX

In the exposure group, the residents were assessed using an updated version of the nonmydriatic fundus camera Kestrel 3100m (Shanghai Top View Industrial Co Ltd) with a self-service module. In the process of fundus photography, the residents pressed the "Start" button by themselves. All checking steps (including focusing, shooting, and image quality review) were undertaken automatically by the fundus camera (Figure 2). Screening data were transmitted to the AI algorithm on a cloud-based server center through the telemedicine platform, and the screening results were fed back immediately. Residents were fully informed that the assessment was fully automated and not performed by the optometrist.

Figure 2. Workflow of the fully automatic self-service nonmydriatic fundus camera. AI: artificial intelligence.

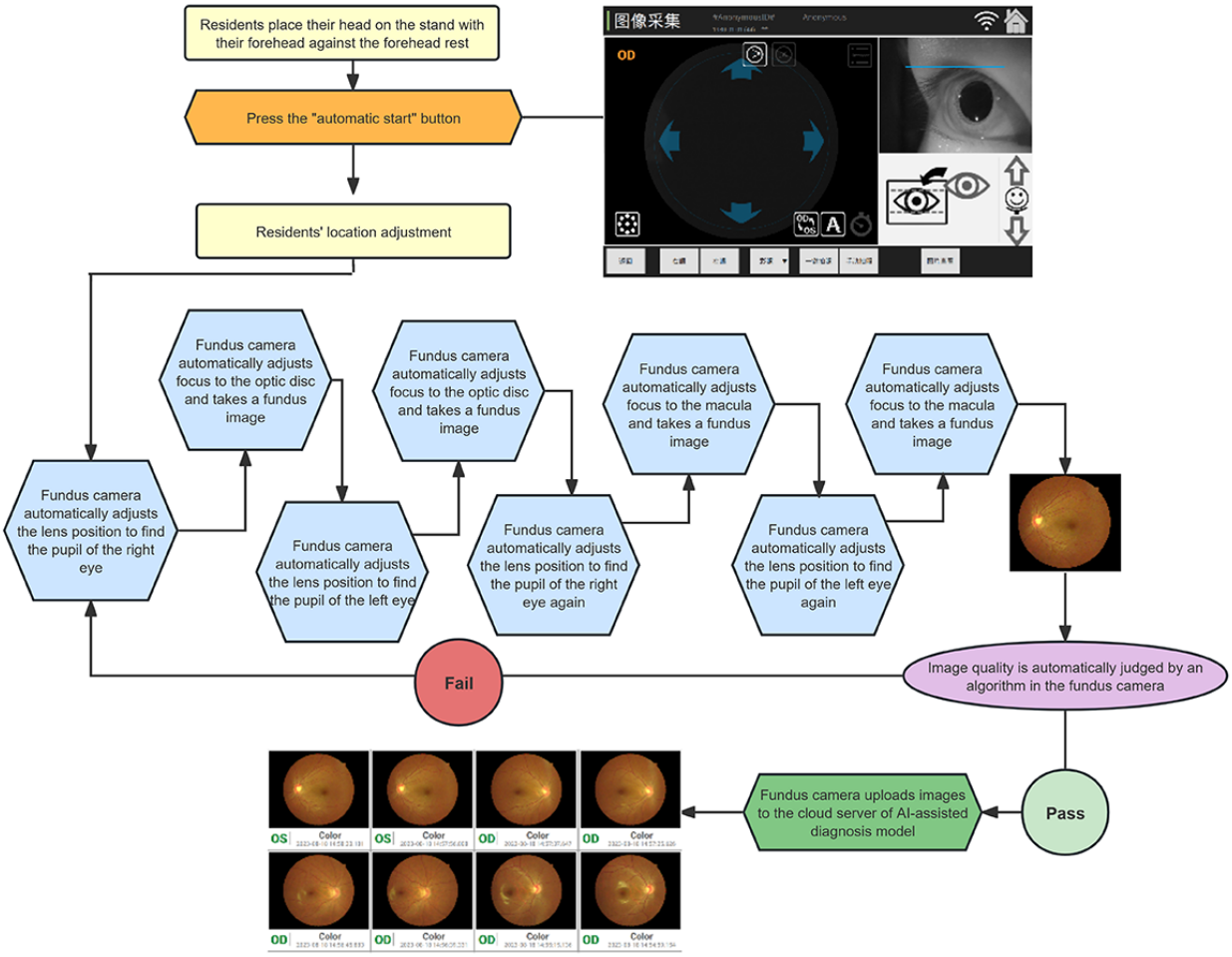

In the control group, the residents were assessed using the basic version of the same nonmydriatic fundus camera. The optical components were identical to those in the exposure group but without the self-service module. In the process of fundus photography, all steps were carried out by the optometrist (including focusing, shooting, and image quality review). Screening data were transmitted to the AI algorithm on a cloud-based server center through the telemedicine platform, and the screening results were fed back immediately. Residents were also fully informed.

Measures and Outcomes

The primary outcome was the screening service's quality. Based on the World Health Organization's recommendations for the evaluation of AI-based medical devices [14] and the European Union's Assessment List for Trustworthy Artificial Intelligence [15], 5 dimensions were selected to reflect the service quality of eye disease screening: effectiveness, physiological discomfort, safety, convenience, and trustworthiness.

Furthermore, effectiveness was based on 2 indicators: image quality and screening efficiency. A staff member recorded the time required for each resident to take fundus photographs (excluding the time taken for diagnosis) at the screening site. Then, a professional ophthalmologist evaluated the quality of each fundus photograph after the on-site experiment. The ophthalmologist was blinded to the grouping of participants. Image quality was assessed on the basis of the image quality pass rate, expressed as the number of eyes with high-quality fundus images per 100 eyes. Screening efficiency was assessed on the basis of the average screening time, expressed as the mean of the time required for each resident to take fundus photographs.

To assess physiological discomfort, safety, convenience, and trustworthiness of screening services, residents were asked to finish a questionnaire just after they received the screening results. A 5-point Likert scale was adopted for each dimension, from the best to the worst, except for the physiological discomfort (Multimedia Appendix 2). A no physiological discomfort rate was expressed as the number of residents who

XSL•FO
RenderX

chose the "There is no physiological discomfort during the screening" per 100 individuals in each group. Safety rate is expressed as the number of residents who chose "The screening is very safe" or "The screening is safe" per 100 individuals in each group. Convenience rate is expressed as the number of residents who chose "The screening is very convenient" or "The screening is convenient" per 100 individuals in each group. The trustworthiness rate is expressed as the number of residents who chose "The screening result is very trustworthy" or "The screening result is trustworthy" per 100 individuals in each group.

The secondary outcome was the preference rate, expressed as the number of residents who were willing to use the same technology for their next screening per 100 individuals. In detail, in the exposure group, the preference rate was expressed as the number of the residents who preferred facilitated self-service eye screening per 100 individuals, while in the control group, it was expressed as the number of residents who preferred traditional manual screening per 100 individuals.

To understand the residents' preference, a video displaying the processes of both facilitated self-service eye screening and traditional manual screening was shown to the residents. Then, the following question was asked: "At your next eye disease screening, you can choose either facilitated self-service eye screening or traditional manual screening. Which one do you prefer?" A total of 4 alternatives were set: "Prefer traditional manual screening," "Prefer facilitated self-service eye screening," "Both are acceptable," and "Neither is acceptable (Refusal of screening)." Each resident could choose only 1 option, which best reflected their preference.

Sample Size

The rule of events per variable was used for sample size estimation. In this study, 2 logit models were established for the 2 groups separately, each containing 8 independent variables. We set 10 events per variable in general. According to a previous study [16], when the decision-making process had high uncertainty, the proportion of individuals who preferred the algorithms was about 50%. This led us to arrive at a sample size of 160 (8 variables multiplied by 10 events each, with 50% of individuals potentially preferring facilitated screening [ie, 50% of 8×10]) for each group.

Analysis

Every dimension of the screening service quality and the preference rate were calculated separately. Chi-square and *t* tests were used to test whether the service quality or the residents' preferences differed between the 2 groups. A total of 7 hypotheses were tested, as shown in Textbox 1.

Textbox 1. Study hypotheses tested.

1. H1: image quality pass rate $_{\text{exposure group}}$≠ image quality pass rate $_{\text{control group}}$

   H0: image quality pass rate $_{\text{exposure group}}$=image quality pass rate $_{\text{control group}}$
2. H1: screening time $_{\text{exposure group}}$≠screening time $_{\text{control group}}$

   H0: screening time $_{\text{exposure group}}$=screening time $_{\text{control group}}$
3. H1: no discomfort rate$_{\text{exposure group}}$≠no discomfort rate $_{\text{control group}}$

   H0: no discomfort rate $_{\text{exposure group}}$ = no discomfort rate $_{\text{control group}}$
4. H1: safety rate$_{\text{exposure group}}$≠safety rate $_{\text{control group}}$

   H0: safety rate $_{\text{exposure group}}$ = safety rate $_{\text{control group}}$
5. H1: convenience rate$_{\text{exposure group}}$≠convenience rate $_{\text{control group}}$

   H0: convenience rate $_{\text{exposure group}}$ = convenience rate $_{\text{control group}}$
6. H1: trustworthiness rate $_{\text{exposure group}}$≠trustworthiness rate $_{\text{control group}}$

   H0: trustworthiness rate $_{\text{exposure group}}$ = trustworthiness rate $_{\text{control group}}$
7. H1: preference rate $_{\text{exposure group}}$≠preference rate $_{\text{control group}}$

   H0: preference rate $_{\text{exposure group}}$ = preference rate $_{\text{control group}}$

If any of the hypotheses among hypotheses 1-6 (Textbox 1) were significant, it indicated that the service quality was different between facilitated self-service eye screening and traditional manual screening. If hypothesis 7 was significant, it meant that the residents' preference for facilitated self-service eye screening was different from that for traditional manual screening.

Additionally, subgroup analyses in the exposure and control groups were conducted to explore the relationships between the screening service quality and the residents' preferences, using generalized logit models. The option "Prefer facilitated self-service eye screening" was used as the reference level for the dependent variable in the models. The independent variables included age, sex, image quality, screening efficiency, physiological discomfort, safety, convenience, and trustworthiness. All statistics were performed using SAS (version 9.4; SAS Institute).

XSL•FO
RenderX

Ethical Considerations

The study adhered to the ethical principles of the Declaration of Helsinki and was approved by the Shanghai General Hospital Ethics Committee (2022SQ272). All participants provided written informed consent before participating in this study. The study data were anonymous, and no identification of individual participants in any images of the manuscript or supplementary material is possible.

Results

Participants' Characteristics

A total of 358 residents enrolled; among them, 176 (49.16%) were in the exposure group and the remaining 182 (50.84%) were in the control group. Residents' basic characteristics were balanced between the 2 groups. The mean age was 65.05 (SD 12.28) years for the exposure group and 63.96 (SD 13.06) years for the control group; however, this difference was nonsignificant (*P*=.42). The proportion of women was 67.05% (n=118) for the exposure group and 62.09% (n=113) for the control group; this difference was also nonsignificant between the 2 groups (*P*=.33).

Screening Service Quality

In the exposure group, high-quality fundus images were obtained for 268 out of 352 eyes (image quality pass rate=76.14%; Figure 3). The average screening time was 81.03 (SD 36.98) seconds (Figure 3). In the control group, high-quality fundus images were obtained for 274 out of 364 eyes (image quality pass rate=75.27%; Figure 3). The average screening time was 78.22 (SD 54.01) seconds (Figure 3). There was no significant difference in the image quality pass rate (χ^2_1=0.07, *P*=.79) and average screening time ($t_{321.01}$=−0.58 [Welch–Satterthwaite–adjusted *df*], *P*=.56) between the 2 groups (Figure 3).

Figure 3. Screening effectiveness and residents' preferences in the exposure and control groups. There was no significant difference between the 2 groups in the measures reflecting service quality. However, the proportion of participants who were willing to use the same technology for their next screening in the exposure group was much lower than in the control group (*P*<.001).

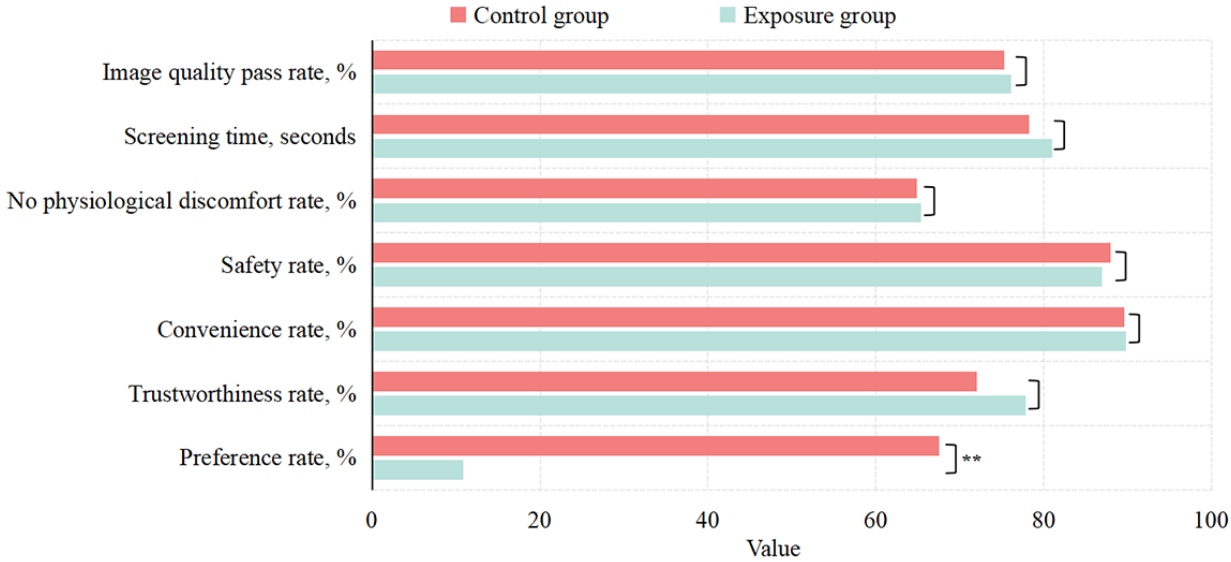

For the other dimensions, detailed information is shown in Figure 3. There were no significant differences between any of these rates between the 2 groups (no physiological discomfort rate: χ^2_1=0.01, *P*=.92; safety rate: χ^2_1=0.08, *P*=.78; convenience rate: χ^2_1=0.004, *P*=.95; trustworthiness rate: χ^2_1=1.63, *P*=.20).

Residents' Preferences

In the exposure group, 120 (68.18%) residents preferred traditional manual screening, 19 (10.80%) preferred facilitated self-service eye screening, 19 (10.80%) preferred both, and the remaining 18 (10.23%) preferred neither. In the control group, 123 (67.58%) residents preferred traditional manual screening, 14 (7.69%) preferred facilitated self-service eye screening, 20 (10.99%) preferred both, and the remaining 25 (13.74%) preferred neither.

The proportion of residents who chose the category "Prefer facilitated self-service eye screening" in the exposure group was significantly lower than that of residents who chose the category "Prefer traditional manual screening" in the control group (χ^2_1=120.57, *P*<.001; Figure 3).

Subgroup Analyses

In the exposure group, 4 generalized logit models were generated (Table 1). Regarding the effectiveness of facilitated self-service eye screening, neither the image quality nor the screening time had an impact on the residents' preferences. Regarding the other dimensions for facilitated self-service eye screening service quality, models 3 and 4 demonstrated that distrust in the results of facilitated self-service eye screening might decrease the probability of preferring this screening service and increase the probability of preferring neither of the 2 screening services.

XSL•FO
RenderX

Table 1. Subgroup analysis for exploring the factors affecting residents' preferences[a] in the exposure group (n=176)[b] (models 1-4).

	Effectiveness scores, mean (SE; *P* value)			
	Model 1	Model 2	Model 3	Model 4
Prefer traditional manual screening				
Male gender	0.35 (0.30; .24)	0.36 (0.30; .23)	0.37 (0.31; .24)	0.37 (0.31; .24)
Age	0.02 (0.02; .34)	0.02 (0.02; .28)	0.02 (0.02; .32)	0.03 (0.02; .23)
Number of high-quality images	__[c]	0.05 (0.37; .89)	—	0.12 (0.37; .75)
Screening time	—	–0.01 (0.01; .43)	—	–0.01 (0.01; .37)
Feel that the facilitated self-service eye screening is uncomfortable	—	—	0.1 (0.27; .71)	0.1 (0.28; .73)
Feel that the facilitated self-service eye screening is unsafe	—	—	6.37 (105.60; .95)	6.78 (147.30; .96)
Feel that the facilitated self-service eye screening is inconvenient	—	—	–1.23 (0.75; .10)	–1.29 (0.75; .09)
Distrust the results of the facilitated self-service eye screening	—	—	–0.18 (0.38; .64)	–0.20 (0.38; .59)
Prefer both traditional and facilitated self-service eye screening				
Male gender	0.53 (0.37; .15)	0.54 (0.37; .14)	0.62 (0.39; .11)	0.65 (0.39; .10)
Age	0.03 (0.03; .32)	0.02 (0.03; .36)	0.03 (0.03; .35)	0.03 (0.03; .37)
Number of high-quality images	—	0.08 (0.48; .87)	—	0.21 (0.51; .68)
Screening time	—	0 (0.01; .66)	—	0 (0.01; .55)
Feel that the facilitated self-service eye screening is uncomfortable	—	—	–0.04 (0.37; .90)	–0.02 (0.38; .96)
Feel that the facilitated self-service eye screening is unsafe	—	—	6.08 (105.60; .95)	6.47 (147.30; .96)
Feel that the facilitated self-service eye screening is inconvenient	—	—	0.66 (0.76; .38)	0.70 (0.78; .37)
Distrust the results of the facilitated self-service eye screening	—	—	–1.57 (0.83; .06)	–1.67 (0.86; .05)
Prefer neither traditional nor facilitated self-service screening (refusal for screening)				
Male gender	0.21 (0.39; .58)	0.22 (0.39; .58)	–0.07 (0.46; .87)	–0.10 (0.47; .84)
Age	0.03 (0.03; .26)	0.04 (0.03; .22)	0.02 (0.03; .48)	0.03 (0.04; .45)
Number of high-quality images	—	0.20 (0.49; .68)	—	0.21 (0.62; .74)
Screening time	—	0 (0.01; .81)	—	0 (0.01; .99)
Feel that the facilitated self-service eye screening is uncomfortable	—	—	0.03 (0.46; .95)	0.01 (0.47; .98)
Feel that the facilitated self-service eye screening is unsafe	—	—	6.38 (105.60; .95)	6.77 (147.30; .96)
Feel that the facilitated self-service eye screening is inconvenient	—	—	0 (0.74; >.99)	–0.06 (0.75; .94)
Distrust the results of the facilitated self-service eye screening	—	—	1.32 (0.55; .02)	1.31 (0.55; .02)

[a]Age and gender were adjusted in model 1. Age, gender, image quality, and screening efficiency were adjusted in model 2. Age, gender, physiological discomfort, safety, convenience, and trustworthiness were adjusted in model 3. Age, gender, image quality, screening efficiency, physiological discomfort, safety, convenience, and trustworthiness were adjusted in model 4.

[b]In the exposure group, distrust in the results of facilitated self-service eye screening might decrease the probability of preferring this screening service and increase the probability of preferring neither the traditional nor the facilitated self-service screening services.

[c]Not available.

XSL•FO
RenderX

In the control group, another 4 generalized logit models were generated (Table 2). Men were more likely to choose a preference both screening services. The probability of preferring manual screening might increase with age, as long as the probability of preferring facilitated self-service eye screening decreased. Regarding the effectiveness of traditional manual screening, neither the image quality pass rate nor the screening time had an impact on the residents' preferences. For the other dimensions of the quality of traditional manual screening, models 7 and 8 showed that if the residents feel unsafe about traditional manual screening, their preference for traditional manual screening might decrease, and they might turn to facilitated self-service eye screening.

XSL•FO
RenderX

Table 2. Subgroup analysis for exploring the factors affecting residents' preferences[a] in the control group (n=182)[b] (models 5-8).

	Effectiveness scores, mean (SE; *P* value)			
	Model 5	Model 6	Model 7	Model 8
Prefer traditional manual screening				
Male gender	0.49 (0.36; .17)	0.51 (0.36; .16)	0.42 (0.37; .27)	0.43 (0.38; .26)
Age	0.07 (0.02; .002)	0.07 (0.02; .002)	0.06 (0.02; .006)	0.07 (0.02; .004)
Number of high-quality images	__[c]	0.19 (0.47; .69)	—	0.26 (0.52; .61)
Screening time	—	0 (0.01; .93)	—	0 (0.01; .86)
Feel that the facilitated self-service eye screening is uncomfortable	—	—	0.40 (0.40; .32)	0.38 (0.40; .34)
Feel that the facilitated self-service eye screening is unsafe	—	—	–1.36 (0.51; .007)	–1.45 (0.53; .006)
Feel that the facilitated self-service eye screening is inconvenient	—	—	6.33 (151.00; .97)	6.37 (130.50; .96)
Distrust the results of the facilitated self-service eye screening	—	—	–0.24 (0.35; .48)	–0.27 (0.35; .44)
Prefer both traditional and facilitated self-service eye screening				
Male gender	0.94 (0.41; .02)	0.92 (0.42; .03)	0.94 (0.44; .03)	0.96 (0.45; .03)
Age	0.03 (0.02; .22)	0.03 (0.03; .27)	0.02 (0.03; .55)	0.02 (0.03; .58)
Number of high-quality images	—	–0.27 (0.54; .62)	—	–0.16 (0.60; .79)
Screening time	—	0 (0.01; .85)	—	0 (0.01; .84)
Feel that the facilitated self-service eye screening is uncomfortable	—	—	0.52 (0.46; .26)	0.61 (0.46; .19)
Feel that the facilitated self-service eye screening is unsafe	—	—	–8.03 (126.40; .95)	–8.84 (79.57; .91)
Feel that the facilitated self-service eye screening is inconvenient	—	—	8.63 (151.00; .95)	9.64 (130.50; .91)
Distrust the results of the facilitated self-service eye screening	—	—	–0.56 (0.52; .29)	–0.46 (0.51; .37)
Prefer neither traditional nor facilitated self-service screening (refusal for screening)				
Male gender	0.48 (0.39; .23)	0.38 (0.41; .36)	0.64 (0.47; .17)	0.46 (0.49; .35)
Age	0.01 (0.02; .72)	0.01 (0.02; .58)	–0.01 (0.03; .80)	0 (0.03; .89)
Number of high-quality images	—	1.21 (0.67; .07)	—	1.07 (0.80; .18)
Screening time	—	0.02 (0.01; .06)	—	0.02 (0.01; .19)
Feel that the facilitated self-service eye screening is uncomfortable	—	—	0.71 (0.51; .16)	0.54 (0.52; .30)
Feel that the facilitated self-service eye screening is unsafe	—	—	–1.32 (0.80; .10)	–0.90 (0.79; .25)
Feel that the facilitated self-service eye screening is inconvenient	—	—	8.21 (151.00; .96)	7.55 (130.50; .95)
Distrust the results of the facilitated self-service eye screening	—	—	0.79 (0.46; .09)	0.62 (0.47; .19)

[a]Age and gender were adjusted in model 5. Age, gender, image quality, and screening efficiency were adjusted in model 6. Age, gender, physiological discomfort, safety, convenience, and trustworthiness were adjusted in model 7. Age, gender, image quality, screening efficiency, physiological discomfort, safety, convenience, and trustworthiness were adjusted in model 8.

[b]In the control group, if the residents feel unsafe about traditional manual screening, their preference for traditional manual screening might decrease, and they might turn to facilitated self-service eye screening.

[c]Not available.

XSL•FO
RenderX

Discussion

A new fundus disease screening pattern was established using the fully automatic fundus camera without any manual intervention. Our findings suggest that facilitated self-service eye screening can achieve a service quality similar to that of traditional manual screening. The study further evaluated the residents' preferences and associated factors for the newly established self-service fundus disease screening. Our study found that the residents' preference for facilitated self-service eye screening is significantly less than that for traditional manual screening. This implies that the association between the service quality of the screening technology and residents' preferences was weak, suggesting that aversion to the algorithm might exist. In addition, the subgroup analyses suggest that even the high quality of facilitated self-service eye screening cannot increase the residents' preference for this new screening pattern. Worse still, distrust in the results of this new pattern may lead to lower usage of eye disease screening services as a whole. To the best of our knowledge, this study is one of the first to evaluate service quality and residents' preferences for facilitated self-service fundus disease screening.

Previous studies have suggested that people significantly prefer manual services to algorithms in the field of medicine [16-18]. Individuals have an aversion to algorithms underlying digital technology, especially when they see errors in the algorithm's functioning [18]. The preference for algorithms does not increase even if the residents are told that the algorithm outperforms human doctors [19,20]. Our results confirm that fundus image quality in the exposure group is similar to that in the control group in our study, and both are similar to or even better than those reported in previous studies [21,22]. However, the preference for facilitated self-service fundus disease screening is significantly less than that for traditional manual screening. One possible explanation is that uniqueness neglect—a concern that algorithm providers are less able than human providers to account for residents' (or patients') unique characteristics and circumstances—drives consumer resistance to digital medical technology [23]. Therefore, personalized health management solutions based on large language models should be developed urgently [24] to meet the residents' individual demands. In addition, a survey of population preferences for medical AI indicated that the most important factor for the public is that physicians are ultimately responsible for diagnosis and treatment planning [25]. As a result, man-machine collaboration, such as human supervision, is still necessary [26], especially in the early stages of digital transformation to help residents understand and accept the digital technologies.

Furthermore, our study suggests that distrust in the results of facilitated self-service fundus disease screening may cause residents to abandon eye disease screening, irrespective of whether it is provided using this new screening pattern or via the traditional manual screening pattern. This is critical to digital transformation in medicine. This implies that if the digital technology does not perform well, residents will not only be averse to the digital technology itself but also be more likely to abandon health care services as a whole. Digital transformation is a fundamental change to the health care delivery system. This implies that it can self-disrupt its ability to question the practices and production models of existing health care services. As a result, it may become incompatible with the existing models, processes, activities, and even cultures [27]. Therefore, it is important to assess whether the adoption of digital technologies contributes to health system objectives in an optimal manner, and this assessment should be carried out at the level of health services but not at the level of digital transformation [28].

The most prominent limitation of our study is that it was conducted only in Shanghai, China. Because of the sound health care system in Shanghai, residents have already received high-quality eye disease screening services before the adoption of the facilitated self-service eye screening pattern. Consequently, residents are bound to demand more from this new pattern. This situation is quite different from that in lower-income regions. Digital technology was adapted in poverty-stricken areas to build an eye care system, but it did not replace the original system that is based on manually delivered services [13]. Therefore, the framing effect may be weak [29], and there is little practical value in comparing digital technology and manual services in these regions. Second, our study is an observational study and blind grouping was not practical due to the special characteristics of fundus examination. However, we have attempted to use blind processing whenever possible. For instance, ophthalmologists' evaluation of image quality was conducted in a blinded manner. Third, the manner in which we inquired about residents' preferences might affect the results. For example, participants in the exposure group generally have experience with manual screening, but those in the control group may not have had enough experience with facilitated screening despite having been shown a video. This might make the participants in the control group more likely to choose manual screening because the new technology was unfamiliar. Finally, individual-level socioeconomic factors or educational level were not recorded, so we cannot rule out the influence of these factors on residents' preferences.

In summary, this study confirms that the facilitated self-service fundus disease screening pattern could achieve high service quality. The preference of the residents for this new mode, however, was not ideal. It was difficult to reverse residents' preference for manual screening in a short period, especially when the original manual service was already excellent. Therefore, the digital transformation of health care must proceed with caution. We suggest that attention be paid to the residents' individual needs. Although more efficient man-machine collaboration is necessary to help the public understand and accept new technologies, personalized health management solutions based on large language models are required.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Shanghai Public Health Three-Year Action Plan (GWVI-11.1-30, GWVI-11.1-22), Science and Technology Commission of Shanghai Municipality (20DZ1100200 and 23ZR1481000), Shanghai Municipal Health Commission (2022HP61, 2022YQ051, and 20234Y0062), Shanghai First People's Hospital featured research projects (CCTR-2022C08) and Medical Research Program of Hongkou District Health Commission (Hongwei2202-07).

Data Availability

Data are available from the corresponding author upon reasonable request.

Authors' Contributions

SL, YM, and YJ contributed to the conceptualization and design of the study. SL, YM, YJ, YP, TY, and YX collected the data. SL and YM analyzed the data. SL, YM, and YJ drafted the manuscript. WL, YX, JZ, LL, and HZ extensively revised the manuscript. All authors read and approved the final manuscript submitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video of the non-mydriatic fundus camera Kestrel-3100m with the self-service module.
[MP4 File (MP4 Video), 14817 KB - jmir_v26i1e45545_app1.mp4]

Multimedia Appendix 2
Questions for screening service quality.
[DOCX File , 12 KB - jmir_v26i1e45545_app2.docx]

References

1. GBD 2019 BlindnessVision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e144-e160 [FREE Full text] [doi: 10.1016/S2214-109X(20)30489-7] [Medline: 33275949]
2. Marques AP, Ramke J, Cairns J, Butt T, Zhang JH, Muirhead D, et al. Global economic productivity losses from vision impairment and blindness. EClinicalMedicine 2021 May;35:100852 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100852] [Medline: 33997744]
3. Jan C, Li S, Kang M, Liu L, Li H, Jin L, et al. Association of visual acuity with educational outcomes: a prospective cohort study. Br J Ophthalmol 2019 Nov 18;103(11):1666-1671. [doi: 10.1136/bjophthalmol-2018-313294] [Medline: 30658989]
4. Chai YX, Gan ATL, Fenwick EK, Sui AY, Tan BKJ, Quek DQY, et al. Relationship between vision impairment and employment. Br J Ophthalmol 2023 Mar 16;107(3):361-366. [doi: 10.1136/bjophthalmol-2021-319655] [Medline: 34656985]
5. Nayeni M, Dang A, Mao AJ, Malvankar-Mehta MS. Quality of life of low vision patients: a systematic review and meta-analysis. Can J Ophthalmol 2021 Jun;56(3):151-157. [doi: 10.1016/j.jcjo.2020.10.014] [Medline: 33189608]
6. Wang L, Zhu Z, Scheetz J, He M. Visual impairment and ten-year mortality: the Liwan Eye Study. Eye (Lond) 2021 Aug 19;35(8):2173-2179 [FREE Full text] [doi: 10.1038/s41433-020-01226-x] [Medline: 33077908]
7. GBD 2019 BlindnessVision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e130-e143 [FREE Full text] [doi: 10.1016/S2214-109X(20)30425-3] [Medline: 33275950]
8. Cheng C, Wang N, Wong TY, Congdon N, He M, Wang YX, Vision Loss Expert Group of the Global Burden of Disease Study. Prevalence and causes of vision loss in East Asia in 2015: magnitude, temporal trends and projections. Br J Ophthalmol 2020 May 28;104(5):616-622. [doi: 10.1136/bjophthalmol-2018-313308] [Medline: 31462416]
9. Li JO, Liu H, Ting DS, Jeon S, Chan RP, Kim JE, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res 2021 May;82:100900 [FREE Full text] [doi: 10.1016/j.preteyeres.2020.100900] [Medline: 32898686]
10. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb 25;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]

XSL•FO
RenderX

11. Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit Health 2020 May;2(5):e240-e249. [doi: 10.1016/s2589-7500(20)30060-1]
12. Tang J, Liang Y, O'Neill C, Kee F, Jiang J, Congdon N. Cost-effectiveness and cost-utility of population-based glaucoma screening in China: a decision-analytic Markov model. Lancet Glob Health 2019 Jul;7(7):e968-e978. [doi: 10.1016/s2214-109x(19)30201-3]
13. Xiao X, Xue L, Ye L, Li H, He Y. Health care cost and benefits of artificial intelligence-assisted population-based glaucoma screening for the elderly in remote areas of China: a cost-offset analysis. BMC Public Health 2021 Jun 04;21(1):1065 [FREE Full text] [doi: 10.1186/s12889-021-11097-w] [Medline: 34088286]
14. Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training Validation and Evaluation. World Health Organization. URL: https://www.who.int/publications/i/item/9789240038462 [accessed 2024-03-27]
15. The Assessment List for Trustworthy Artificial Intelligence. URL: https://altai.insight-centre.org/ [accessed 2024-03-27]
16. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 2020 Oct 11;31(10):1302-1314. [doi: 10.1177/0956797620948841] [Medline: 32916083]
17. DeCamp M, Tilburt JC. Why we cannot trust artificial intelligence in medicine. Lancet Digit Health 2019 Dec;1(8):e390. [doi: 10.1016/s2589-7500(19)30197-9]
18. Frank D, Elbæk CT, Børsting CK, Mitkidis P, Otterbring T, Borau S. Drivers and social implications of artificial intelligence adoption in healthcare during the COVID-19 pandemic. PLoS One 2021 Nov 22;16(11):e0259928 [FREE Full text] [doi: 10.1371/journal.pone.0259928] [Medline: 34807907]
19. Juravle G, Boudouraki A, Terziyska M, Rezlescu C. Trust in artificial intelligence for medical diagnoses. Prog Brain Res 2020;253:263-282. [doi: 10.1016/bs.pbr.2020.06.006] [Medline: 32771128]
20. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/s2589-7500(19)30123-2]
21. Scanlon PH, Foy C, Malhotra R, Aldington SJ. The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening. Diabetes Care 2005 Oct;28(10):2448-2453 [FREE Full text] [doi: 10.2337/diacare.28.10.2448] [Medline: 16186278]
22. Cen L, Ji J, Lin J, Ju S, Lin H, Li T, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nat Commun 2021 Aug 10;12(1):4828 [FREE Full text] [doi: 10.1038/s41467-021-25138-w] [Medline: 34376678]
23. Longoni C, Bonezzi A, Morewedge C. Resistance to medical artificial intelligence. J Consum Res 2019;46:650. [doi: 10.1093/jcr/ucz013]
24. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management. JAMA Ophthalmol 2024 Feb 22 [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.6917] [Medline: 38386351]
25. Ploug T, Sundby A, Moeslund TB, Holm S. Population preferences for performance and explainability of artificial intelligence in health care: choice-based conjoint survey. J Med Internet Res 2021 Dec 13;23(12):e26611 [FREE Full text] [doi: 10.2196/26611] [Medline: 34898454]
26. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021 Sep;3(9):e599-e611. [doi: 10.1016/s2589-7500(21)00132-1]
27. Alami H, Gagnon M, Fortin J. Digital health and the challenge of health systems transformation. Mhealth 2017 Aug 08;3:31-31 [FREE Full text] [doi: 10.21037/mhealth.2017.07.02] [Medline: 28894741]
28. Ricciardi W, Pita Barros P, Bourek A, Brouwer W, Kelsey T, Lehtonen L, Expert Panel on Effective Ways of Investing in Health (EXPH). How to govern the digital transformation of health services. Eur J Public Health 2019 Oct 01;29(Supplement_3):7-12 [FREE Full text] [doi: 10.1093/eurpub/ckz165] [Medline: 31738442]
29. Khan WU, Shachak A, Seto E. Understanding decision-making in the adoption of digital health technology: the role of behavioral economics' prospect theory. J Med Internet Res 2022 Feb 07;24(2):e32714 [FREE Full text] [doi: 10.2196/32714] [Medline: 35129459]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

Edited by A Mavragani; submitted 06.01.23; peer-reviewed by B Li, A Bate, CW Pan; comments to author 13.09.23; revised version received 15.10.23; accepted 12.03.24; published 17.04.24.

Please cite as:
Lin S, Ma Y, Jiang Y, Li W, Peng Y, Yu T, Xu Y, Zhu J, Lu L, Zou H
Service Quality and Residents' Preferences for Facilitated Self-Service Fundus Disease Screening: Cross-Sectional Study
J Med Internet Res 2024;26:e45545
URL: https://www.jmir.org/2024/1/e45545
doi:10.2196/45545
PMID:38630535

©Senlin Lin, Yingyan Ma, Yanwei Jiang, Wenwen Li, Yajun Peng, Tao Yu, Yi Xu, Jianfeng Zhu, Lina Lu, Haidong Zou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Acceptability, Engagement, and Feasibility of Mental Health Apps for Marginalized and Underserved Young People: Systematic Review and Qualitative Study

Holly Alice Bear[1], DPhil; Lara Ayala Nunes[1], DPhil; Giovanni Ramos[2], PhD; Tanya Manchanda[1], MEd; Blossom Fernandes[1], PhD; Sophia Chabursky[3], MSc; Sabine Walper[3], DPhil; Edward Watkins[4], DPhil; Mina Fazel[1], DM

[1]Department of Psychiatry, University of Oxford, Oxford, United Kingdom

[2]Department of Psychological Science, University of California, Irvine, CA, United States

[3]German Youth Institute, Munich, Germany

[4]School of Psychology, University of Exeter, Exeter, United Kingdom

Corresponding Author:
Holly Alice Bear, DPhil
Department of Psychiatry
University of Oxford
Warneford Hospital
Warneford Lane
Oxford, OX3 7JX
United Kingdom
Phone: 44 01865 6182
Email: holly.bear@psych.ox.ac.uk

Abstract

Background: Smartphone apps may provide an opportunity to deliver mental health resources and interventions in a scalable and cost-effective manner. However, young people from marginalized and underserved groups face numerous and unique challenges to accessing, engaging with, and benefiting from these apps.

Objective: This study aims to better understand the acceptability (ie, perceived usefulness and satisfaction with an app) and feasibility (ie, the extent to which an app was successfully used) of mental health apps for underserved young people. A secondary aim was to establish whether adaptations can be made to increase the accessibility and inclusivity of apps for these groups.

Methods: We conducted 2 sequential studies, consisting of a systematic literature review of mental health apps for underserved populations followed by a qualitative study with underserved young male participants (n=20; age: mean 19). Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, an electronic search of 5 databases was conducted in 2021. The search yielded 18,687 results, of which 14 articles met the eligibility criteria.

Results: The included studies comprised a range of groups, including those affected by homelessness, having physical health conditions, living in low- and middle-income countries, and those with sexual and gender minority identities. Establishing and maintaining user engagement was a pervasive challenge across mental health apps and populations, and dropout was a reported problem among nearly all the included studies. Positive subjective reports of usability, satisfaction, and acceptability were insufficient to determine users' objective engagement.

Conclusions: Despite the significant amount of funding directed to the development of mental health apps, juxtaposed with only limited empirical evidence to support their effectiveness, few apps have been deliberately developed or adapted to meet the heterogeneous needs of marginalized and underserved young people. Before mental health apps are scaled up, a greater understanding is needed of the types of services that more at-risk young people and those in limited-resource settings prefer (eg, standard vs digital) followed by more rigorous and consistent demonstrations of acceptability, effectiveness, and cost-effectiveness. Adopting an iterative participatory approach by involving young people in the development and evaluation process is an essential step in enhancing the adoption of any intervention, including apps, in "real-world" settings and will support future implementation and sustainability efforts to ensure that marginalized and underserved groups are reached.

Trial Registration: PROSPERO CRD42021254241; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=254241

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48964) doi:10.2196/48964

KEYWORDS

adolescent mental health; marginalized groups; smartphone apps; engagement; implementation science; mobile app; smartphone; mobile health; mHealth; mental health; challenges; acceptability; young; effectiveness; mobile phone

Introduction

Background

Addressing health inequities is a key challenge for the mental health field, especially when trying to ensure that interventions and services are accessible and acceptable for all populations. Nearly 50% of lifelong mental health disorders begin by the age of 14 years, and by the age of 24 years, 75% of mental health disorders have begun [1]. Given the frequent onset of mental health problems during youth, here defined as the period between 15 and 24 years, special attention must be paid to older adolescents, including those from underserved and marginalized minority groups and socioeconomically deprived backgrounds [2]. In these groups, common barriers to accessing mental health services can be exacerbated (eg, poor mental health literacy, lack of knowledge about where to seek help, negative attitudes toward professional help seeking, embarrassment, preference for self-reliance, fear of stigma, and confidentiality concerns), and additional barriers exist (eg, reliance on informal supports, shame, lack of housing or money, and therapist factors, such as different race and level of experience), creating increased risk of untreated mental health problems and thus poorer mental health outcomes [3-7]. In this study, marginalized and underserved populations are defined as those with higher prevalence of mental health problems and lower rates of help seeking, such as racially and ethnically minoritized individuals, rural and remote communities, financially deprived groups, individuals experiencing homelessness, refugee and migrant populations, and sexual and gender minority groups [8-10] and those with lower inclusion in mental health intervention research than one would expect from population estimates [11], respectively. These groups are exposed to risk factors for poor mental health and experience disparities in mental health care, including lower access to care, poorer treatment quality, and limited engagement in treatment [5,10,12].

Smartphone apps could offer an opportunity to deliver mental health and well-being resources and interventions in a scalable, cost-effective, and potentially personalized manner, particularly for those who experience the greatest barriers to accessing health care [13,14]. Given that smartphone ownership is nearly ubiquitous among young people in high-income nations and increasingly across lower-resource settings, apps have the potential to address some of the accessibility issues in service provision for young people's mental health [15]. Young people are more digitally connected (ie, they are more likely to own smartphones and spend more time on the web) and more likely to seek health information on the web than older generations, meaning that app-based interventions may be particularly well-suited for this population [16,17].

Not surprisingly, the number of mental health apps being developed, both commercially and in academic research programs, has expanded rapidly, outpacing scientific evaluations of their effectiveness [18,19]. Emerging evidence suggests that some apps may produce significant symptom improvement across multiple outcomes, compared with waitlist or control conditions [20-22]. Despite promise, empirical research often fails to translate into meaningful and sustained implementation in "real-world" settings [23,24]. Research has focused primarily on efficacy under ideal "laboratory" conditions rather than effectiveness in real-world settings [25]. Therefore, at present, most apps, especially those available to the public, lack strong empirical support [19]. The acceptability (ie, perception that a given technology is useful, agreeable, palatable, or satisfactory); accessibility (ie, the technology being easy to obtain or use); engagement (ie, initial adoption and sustained interactions with the technology, including the level of app use, intervention adherence, and premature dropout); and feasibility (ie, the actual fit, utility, or suitability and the extent to which the technology can be successfully used or conducted within a given context) of apps for marginalized and underserved groups remain poorly understood [25-28]. Although mental health apps may provide a possible solution, marginalized and underserved groups of young people face unique challenges to engage with and benefit from these interventions (eg, intervention cost, content that is not culturally attuned and lack of reliable access to the internet) and are typically underrepresented in intervention research [25,27]. Although increased access is often seen as a major benefit of digital mental health interventions, issues related to the "digital divide" describe the phenomena that technology is not equally available to all social groups due to economic, social, or cultural inequalities and is a potential ethical concern [29]. Furthermore, underrepresentation in the intervention development process potentially reinforces structural inequalities by limiting the availability of products that are culturally accessible, inclusive, and effective or by skewing the product features to attract young people from more advantaged backgrounds [9,14]. Therefore, considering diversity, equity, and inclusion issues at the outset of health care research as well as within app evaluation is essential to prevent the perpetuation of existing inequities [27].

To date, little has been published on the attempts to create new or adapt existing app interventions to meet the heterogeneous needs of diverse groups of young people [9,14]. Moving forward, careful consideration is needed to ensure optimal leveraging of all mental health intervention research, including that of mental health apps, to increase health equity while also ensuring that innovations do not inadvertently widen the digital divide and exacerbate health inequalities [14]. Although the efficacy of many mental health apps remains unclear, future attempts to translate findings for underserved populations will need to ensure that all apps are developed with enough flexibility to fit a wider range of user needs and preferences. To achieve this goal, research is needed to assess the acceptability and feasibility of mental health apps for underserved young people to ensure that they are not further excluded from research and

XSL•FO
RenderX

to advance toward mental health provision that meets their needs.

This Research

We conducted two sequential studies: (1) a systematic literature review and (2) a qualitative study with a targeted sample of young people who often are underrepresented in research, with limited access to health care and socioeconomic deprivation. The overarching aim of these combined studies was to better understand whether mental health apps are feasible and acceptable to underserved young people. A secondary aim was to determine which adaptations might enable accessibility of and effective engagement with mental health apps for these groups.

The research questions of interest were as follows: (1) On the basis of the existing literature, are mental health apps acceptable, feasible, and engaging for marginalized and underserved young people and how have these constructs been measured? (2) On the basis of the qualitative study, what are young people's experiences of using a mental well-being app, including its acceptability, feasibility, and level of engagement? (3) On the basis of both studies, are apps an acceptable, feasible, and engaging intervention approach to meet the specific needs of underserved young people? What adaptations can be made to ensure that mental health apps are accessible and inclusive for these groups?

Methods

Overview

To fully address our research questions, we adopted a 2-pronged approach. First, we conducted a systematic review of the literature to better understand the acceptability and feasibility of mental health apps for underserved young people. To explore the findings of the systematic review in greater depth and to provide further insights from multiple perspectives, we next conducted a qualitative study with young men not in education, employment, or training (NEET) in the United Kingdom and Spain and asylum seekers and refugees in Germany. The interviews were conducted between August 2021 and February 2022.

Systematic Review

Literature Search and Search Strategy

An electronic literature search was performed in English on the following databases from January 2009 to May 2021: Cochrane Library, Embase, MEDLINE, and PsycINFO. We used key search terms relating to (1) underserved young people; (2) mental health mobile apps; and (3) acceptability, feasibility, and engagement. The search strategy was guided by similar reviews exploring digital mental health interventions for young people [25,30], and the terms for apps were derived from Cochrane reviews [31,32]. An updated search was conducted in September 2023. The full search strategies are available in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Screened articles were included if (1) the study targeted marginalized and underserved young people with a mean age of 15 to 25 years, including individuals who were NEET, apprentices, teenage parents, members of minoritized racial or ethnic groups, members of sexual and gender minoritized groups, residents of low- and middle-income countries (LMICs), experiencing homelessness, socioeconomic deprivation, refugees or asylum seekers, and migrants; individuals with substance use disorders; those under state or statutory care; people with physical disabilities; and individuals involved in the criminal justice system or incarcerated; (2) the intervention was a "native" mobile app (ie, not on a web browser), whose primary aim was to promote well-being, prevent mental health problems, or treat existing mental health problems, delivered as a stand-alone intervention or as an adjunct to therapist-assisted interventions; (3) the primary outcome was a measure of mental health or well-being; and (4) the study reported a measure of user acceptability or feasibility.

Articles were excluded if (1) the mean age of participants was outside of the 15 to 25 years range; (2) the intervention was not a mobile app, that is, other digital interventions, including teletherapy (eg, therapy delivered by phone, SMS text messages, video platforms, or PCs); and (3) there was no measure of acceptability or feasibility. Gray literature was not included in the search.

Study Selection

In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [33], the flowchart presented in Figure 1 provides step-by-step details of the study selection procedure. The PRISMA checklist is provided in Multimedia Appendix 1. The search strategy identified 11,539 citations after deduplication. After an initial screening of the titles, which resulted in the exclusion of 10,061 (87.19%) irrelevant entries, the abstracts of 1478 (12.81%) studies were screened by 4 members of the review team (LAN, HAB, BF, and TM). The identified 176 (11.91%) full texts were then screened by LAN. In this final stage, 11 (6.2%) studies, corresponding to 9 interventions, were identified for inclusion in the review, with 8 (73%) found through the electronic search and 3 (27%) through manual searches of the reference lists of relevant articles. The updated search identified a further 7148 citations after deduplication, of which 3 met inclusion criteria and were included in the review.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process.

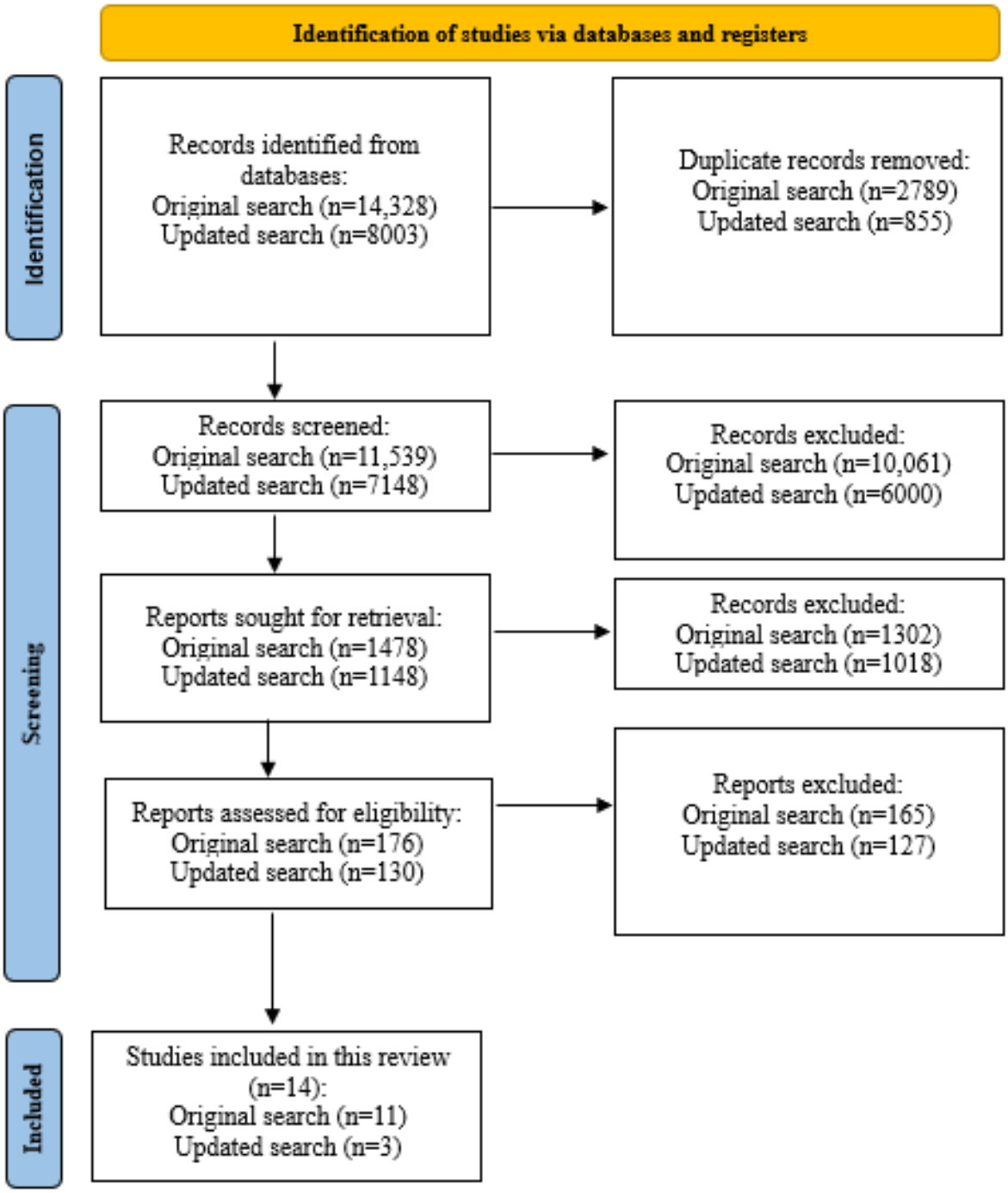

Data Extraction

Data were extracted by 1 reviewer (LAN, HAB, or BF) and reviewed for accuracy and completeness by a second reviewer. After verifying all the extracted data, discrepancies were resolved by discussion or adjudication by another author (MF). Extracted data included information on study characteristics (ie, authors, publication year, country, study design, and study population); intervention characteristics (ie, characteristics of the technology, app name, therapeutic modality, and intervention outcomes); and feasibility and acceptability.

Quality Assessment

We used the Mixed Methods Appraisal Tool (MMAT; version 2018) to assess the methodological quality of the included studies [34]. MMAT was developed by combining the core relevant methodological criteria found in different well-known and widely used qualitative and quantitative critical appraisal tools. The MMAT consists of 2 screening questions applicable to all types of study design and a further 5 questions applicable to specific study designs. Responses were rated on a categorical scale as “no,” “unclear,” or “yes” to any of the methodological

XSL•FO
RenderX

quality criteria. Quality assessments were made by 1 reviewer (TM). We did not exclude any studies based on quality assessment scores.

Data Synthesis and Analysis

The extracted data were collated and summarized to produce a narrative summary of the study; sample characteristics; and acceptability, feasibility, and engagement outcomes. A codebook approach was used to code and synthesize implementation data from all available sources according to the outcome categories [35].

Qualitative Study

Study Context

The Emotional Competence for Well-Being (ECoWeB) cohort multiple randomized controlled trial involved a longitudinal prospective cohort to examine the well-being, mental health, and emotional competence in individuals aged 16 to 22 years across 12 months. The experimental arm was an emotional competence self-help app (ie, MyMoodCoach).

Intervention

MyMoodCoach was designed to test if an app could improve different processes affecting mental well-being, including, but not limited to, improving emotion regulation by reducing maladaptive strategies such as worry and rumination and replacing them with constructive alternatives and problem-solving and enhancing emotional knowledge and perception through psychoeducation and learning tasks. The app was designed for young people broadly and was not targeted at a specific population. Full details of the ECoWeB trial are reported in the study protocol [36].

Participant Recruitment

In parallel to the trial, we additionally recruited young people to understand the views and experiences of those underrepresented in the study sample (and most other app-based studies). As there was an overrecruitment of White, university-educated female participants in the study sample, we decided to only recruit male participants and to focus on 2 specific groups: NEET men and migrant populations (including both voluntary and forced migrants). In the United Kingdom and Spain, recruitment was conducted through a variety of channels, including Twitter, targeted adverts on social media (eg, Instagram and Facebook), newsletters sent to youth and practitioner networks, outreach to third sector organizations and mental health support groups, and advertisements placed on university and charity websites. In Germany, participants were recruited directly from refugee homes, integration courses, migration services, and youth centers.

Inclusion and Exclusion Criteria

Inclusion criteria required the participants to be (1) aged between 16 and 22 years; (2) able to speak and read English, Spanish, or German; (3) male; (4) NEET (United Kingdom), NEET or migrant (Spain), or asylum seeker or refugee (Germany); (5) having access to a smartphone with the minimal technological specifications necessary for the app (ie, iOS 9 or later or Android 8.0 or later); and (6) having access to the internet via mobile data or Wi-Fi. The exclusion criteria were having current suicidal ideation, psychosis, or bipolar disorder.

Procedure

The research team was led by the principal investigator, MF, professor of Adolescent Psychiatry and Consultant in Children's Psychological Medicine, with particular expertise in the mental health needs of refugee populations. EW, professor of Experimental and Applied Clinical Psychology, Chartered Clinical Psychologist, provided additional supervisory input and guidance as a leading expert in the field of child and adolescent mental health research. SW, professor of Education with a focus on youth research and director at German Youth Institute, provided supervisory oversight of the interviews in Germany. HAB and LAN were postdoctoral researchers at the time the research was conducted, both of whom have several years of experience conducting qualitative research with young people and expertise in analyzing qualitative data. SC, researcher at the German Youth Institute, conducted and analyzed the interviews in Germany and has experience conducting qualitative research with refugee populations. The authors had no relationship with any of the participants.

In the United Kingdom and Spain, participation involved (1) short questionnaires about mood and feelings and demographic questions such as age, gender, race, ethnicity, educational attainment, and country of origin; (2) downloading MyMoodCoach and using it for 4 weeks; and (3) completing a follow-up interview. Interviews were conducted by HAB and LAN via MS Teams (Microsoft, Corp) and lasted approximately 45 minutes. A similar procedure was followed in conducting the interviews with refugees in Germany, but given the likely language barriers in navigating the app, an additional, earlier interview was included 2 to 3 weeks after the initial instructions had been sent to explore if features of the app were understood and to clarify questions. After 2 further weeks, a second interview was conducted. One interview was conducted in person, while all the others were conducted over the telephone.

Interviews followed a semistructured schedule based on a taxonomy of implementation constructs [26]. Topics included self-reported app use, satisfaction and feedback about content, usability, and acceptability (Multimedia Appendix 2). In Germany, the interview topic guide was translated and adapted to the target group. As not to stigmatize the young men, the wording "refugee" was avoided, and instead, when referring to the target group, the wording "young man such as yourself" was used.

Ethical Considerations

In the United Kingdom, ethical approval was granted by the University of Exeter Research Ethics Committee (eCLESPsy000048 v10.0); in Spain, by the Jaime I University Research Ethics Committee (CD/93/2021); and in Germany, by the Ethics Committee of the Faculty of Medicine at Ludwig-Maximilians University Munich (PNr 19-0468/19-0315). Cognizant of the ethical and practical implications of conducting research with underserved populations, we conducted the interviews in private, quiet spaces

XSL•FO
RenderX

and in a friendly and reassuring manner to ensure that participants felt comfortable and safe.

With prior consent, all interviews were audio recorded. Participants were reimbursed with a shopping voucher of up to €50 or £50 (US $64) for their time and app use.

Analytic Strategy

Interviews were transcribed verbatim, and the transcripts were assigned a unique pseudonym to anonymize participants. The interviews were analyzed using a combination of theory- and data-driven analysis techniques, consisting primarily of deductive, theory-driven thematic analysis [37]. Analysis of the transcripts in the United Kingdom and Spain was conducted by HAB, using NVivo (version 11; Lumivero). Following a similar procedure, the interviews in Germany were transcribed and analyzed using the coding software MAXQDA (VERBI GmbH) by SC.

Initial familiarization with the data was achieved through the transcription process and iterative rereading of the interviews. Analysis was carried out through a recursive process of open coding, when concepts were named and their properties and dimensions identified, followed by axial coding, when links and associations were drawn between codes. Codes were based on language used by the young people and were applied to each new unit of meaning. Data extracts were multiply coded when appropriate, as were contradictory and minority features of the data. The data set was iteratively reviewed, and codes were systematically applied to the whole data set until a finalized coding manual was established. Codes were organized into potential themes using thematic maps and tables. The development of the coding manual was iteratively reviewed and refined through discussion with all authors throughout the analysis process to ensure the reliability and rigor of the process and results.

Approach to Inquiry

Analysis was conducted from a critical realist perspective to provide a more nuanced understanding and explanation of participants' experiences [38]. This position assumes that although participants' accounts provide important insights about the real world, these accounts are not objective and represent an interpretation of reality [39]. These data require interpretation and explanation by the researcher, who also has their own perspectives on the world, to better understand the underlying mechanisms and processes and, in turn, make recommendations for practice [40].

Results

Systematic Review

Study Characteristics

Characteristics of the 14 included studies, examining 12 interventions, are presented in Table 1. The included studies were published in the United States (n=3), Australia (n=3), Switzerland (n=2), India (n=2), South Korea (n=1) and Germany (n=1). As for the type of intervention, 7 apps were stand-alone, and the rest (n=5) were delivered in combination with other forms of professional support. The interventions (n=12) were targeted at apprentices and the unemployed (n=3), homeless populations (n=2), those with physical health conditions (n=2), sexual and gender minoritized individuals (n=3), those residing in LMICs (n=1) and those with co-occurring autism spectrum disorders (n=1). Most of the interventions had been co-designed with young people (11/12, 92%). The study sample size ranged from 9 to 877, with 4 of the 14 studies having a sample size of >200 participants.

XSL•FO
RenderX

Table 1. Study characteristics.

Study, year	App name	Population	Country	Study design	Intervention focus	Sample size, n	Age (y), mean (SD)	Sex (female, %)
Bohleber et al [41], 2016	Companion App	Employed (apprentices) and unemployed	Switzerland	Mixed methods	Peer mentoring and interactive health content to increase social support and reduce stress	619	Employed: 16.9 (1.73), unemployed: 18.4 (1.96)	Employed first year: 50.2, employed second year: 56.7, unemployed group: 40.4
Deady et al [42], 2020	HeadGear	Apprentices	Australia	Mixed methods	Behavioral activation and mindfulness therapy	54	21.68 (3.62)	4
Fleming et al [43], 2017	TODAY!	Sexual minority male participants	United States	Qualitative	CBT[a] to manage anxiety and depressive symptoms	9	19. (0.71)	0
Francis et al [44], 2020	CyFi Space	Individuals with CF[b]	Australia	Mixed methods	Social connectedness and well-being of young people living with CF	22	12-17	50
Geirhos et al [45], 2022	Minddistrict. Program: youth-COACH$_{CD}$	Individuals with a chronic illness (CF, JIA[c], and T1D[d])	Germany	Pilot RCT[e]	iCBT[f] targeting symptoms of anxiety and depression	30	16.13 (2.34)	73
Glover et al [46], 2019	A suite of 15 apps including Pocket Helper2.0[g]	Individuals experiencing homelessness	United States	Pilot study	Daily coping skills; focused tips and brief CBT	100	20.03 (1.83)	39
Gonsalves et al [47], 2019	POD Adventures	Students	India	Intervention design	Problem-solving for adolescents at risk of anxiety, depression, and conduct difficulties	Students: 118, service providers: 16	14	46
Gonsalves et al [48], 2021	POD Adventures	Students	India	Pilot study	Problem-solving for adolescents at risk of anxiety, depression, and conduct difficulties	230	15.57	50
Haug et al [49], 2017	ready4life	Vocational students	Switzerland	Pilot study	Life skills training: self-management skills, social skills, and substance use resistance	877	17.4 (2.7)	58.3
Leonard et al [50], 2018	Calm Mom	Mothers experiencing homelessness	United States	Pilot study	Emotion regulation strategies	49	18.54	100
Schueller et al [51], 2019	Pocket Helper, Purple Chill, Slumber Time, and IntelliCare (12 mini apps)[g]	Individuals experiencing homelessness	United States	Pilot study	Emotional support and coping skills	35	19.06 (0.85)	65

XSL•FO
RenderX

Study, year	App name	Population	Country	Study design	Intervention focus	Sample size, n	Age (y), mean (SD)	Sex (female, %)
Escobar-Viera et al [52][h], 2023	REALbot	Rural living LGBTQ+ youth	United States	Pilot study	Chatbot deployed on Facebook Messenger and Instagram apps to deliver educational content	20	16.6 (1.5)	65
Torok et al [53][h], 2022	LifeBuoy	Community sample (over 50% of sample LGBQI sexual minority)	Australia	RCT	DBT to treat persistent emotional dysregulation to prevent self-harm and suicidal behaviours	455	21.5 (2.18)	84.6
Yang and Chung [54][h], 2022	HARU ASD	Individuals with ASD	South Korea	RCT	CBT for anxiety and co-occurring intellectual disability	30	20.97 (5.06)	10

[a]CBT: cognitive behavioral therapy.

[b]CF: cystic fibrosis.

[c]JIA: juvenile idiopathic arthritis.

[d]T1D: type 1 diabetes.

[e]RCT: randomized controlled trial.

[f]iCBT: internet-based cognitive behavioral therapy.

[g]Mobile phones were preloaded with several apps designed to promote mental health wellness and provide real-time resources. Pocket Helper was 1 app specifically designed for this study.

[h]Studies identified in the updated search.

Study Quality

The included studies varied in their methodological quality (Multimedia Appendix 3) [41-54]. Most (13/14, 93%) were judged to contain possible limitations in at least 1 criterion. All studies but 1 (13/14, 93%) were clear in their description of study participants or the process of recruiting a sample representative of the population of interest. All studies addressed the research question using collected data and reported in some way on feasibility and acceptability. Most studies (12/14, 86%) effectively used appropriate qualitative, quantitative, or mixed methods to answer their research question. However, many studies (6/14, 43%) did not have a sufficiently large sample to warrant definitive conclusions about the feasibility and acceptability of the intervention studied.

Acceptability and Feasibility

User acceptability was found to be high across all included studies, with participants rating the apps positively and reporting high satisfaction with the content of the interventions (Table 2). In studies where participants were asked to indicate if they would recommend the study to someone else, the vast majority reported they would [42,44-46,51]. It is notable that despite many users across the studies reporting high satisfaction levels and being willing to recommend the apps, they themselves did not intend to continue using the apps (ie, low predicted engagement), as they did not find them useful or relevant for their own circumstances [41,42,51]. It is also important to note that many studies incentivized participation with payments, prize draws, and vouchers, and this may have influenced acceptability ratings and engagement [41-43,46,49-51,53,54]. Furthermore, in the only study that asked if participants would pay to use apps, most were unwilling to do so [42].

XSL•FO
RenderX

Table 2. Acceptability, feasibility, engagement, usefulness, accessibility, and outcomes.

Reference, year	App name	Measurement	Acceptability and feasibility	Engagement	Barriers to engagement	Perceived usefulness	Accessibility and inclusivity	Intervention outcomes
Bohleber et al [41], 2016	Companion App	Questionnaires and semistructured interviews	Adolescents regarded the concept of the app as well conceived, especially the peer-mentoring system. However, the app did not compare well to other available apps.	Engagement decreased markedly after the first 2 weeks. Average daily visits: in the first 2 weeks, 61; after 6 months, 8.	Technical problems, unclear benefits, and lack of time.	Content was judged informative and interesting. However, some reported that the purpose of app was not evident.	Unemployed participants suggested that reminders to use the app would help.	No significant effect on stress or the perception of social support.
Deady et al [42], 2020)	HeadGear	Questionnaires, semistructured interviews, and focus groups	Apprentices rated the app positively, (average 4/5 stars). Participants had no or neutral willingness to pay for the app. Most would widely recommend the app but predicted their own use would be infrequent (3-10 times) over the next 12 months.	Users completed approximately one-third of the app challenges.	Noncompletion of challenges attributed to "forgetting" and choosing not to "catch up." Users wanted to be able to skip challenges and suggested gamification and greater personalization.	87.2% claimed it had at least moderately improved their mental fitness. Moderate impact on awareness, knowledge, attitudes, intention to change, help seeking, and behavior change around mental health and well-being.	Participants emphasized the importance of gamification and greater personalization, for example, through the inclusion of personalized music.	No significant differences between baseline and 3-month follow-up measurements. Engagement (intervention completion) directly related to effectiveness.
Fleming et al [43], 2017	Today!	Semistructured interview	Expressed enthusiasm for a comprehensive mobile phone app designed to treat clinically significant symptoms of anxiety and depression among young sexual minoritized men.	Not assessed in this paper.	Weekly phone calls with the coach was described by participants as a barrier to engagement.	Overall, participants had positive reactions to the app, but each individual found different features to be useful (eg, community resources and mood rater).	Usability testers had a wealth of suggestions for topics they would like to see addressed in this kind of app.	Not assessed in this paper.
Francis et al [44], 2020	CyFi Space	Questionnaire, group, and individual interviews	Acceptability of the app was rated moderate.	Overall, 37% recruitment response rate. 77% participants used the app at least once a week. Some participants indicated that the use of the app declined as the 6-week trial progressed.	40.9% reported watching the entertainment and motivational videos.	Many participants found the app both useful and fun to use and agreed they would. recommend the app to others.	Participants rated the app's usability as high. Age-related accessibility measured.	Not assessed in this paper.

XSL•FO
RenderX

Reference, year	App name	Measurement	Acceptability and feasibility	Engagement	Barriers to engagement	Perceived usefulness	Accessibility and inclusivity	Intervention outcomes
Geirhos et al [45], 2022	Minddistrict. Program: youth-COACH$_{CD}$	Questionnaires	Content was perceived as appropriate. 58% would recommend intervention to a friend, 17% would likely recommend it, 17% would partly recommend it, and 8% would not recommend it.	Intervention adherence=40%, dropout=20%.	Not reported.	Individual tasks perceived as particularly helpful.	Not explicitly reported.	No symptom improvement. Small sample.
Glover et al [46], 2019	Pocket helper 2.0	Questionnaires	73% would recommend the program.	48% of the sample completed the 3-month assessment, while 19% completed the 6-month assessment.	Use and satisfaction with various features reported.	63% of participants at 3 months and 68% of participants at 6 months reported at least moderate benefit from intervention	Designed for youth experiencing homelessness based on the initial input from these youths and was refined based on the feedback received during a previous pilot trial.	Not assessed in this paper.
Gonsalves et al [47] , 2019	POD[a] Adventures	Focus group discussions, co-design workshops, and user testing	Service providers highlighted that self-help was not a culturally congruent concept for most Indian adolescents.	Not reported in this paper.	Following user testing, activities were shortened to be kept <2-minutes to minimize boredom and disengagement.	Problem-solving reported as being a useful and valued skill.	Design was sensitive to cultural context, language, participant media preferences, and digital access helped focus on user needs. Adaptations were made to address widespread literacy difficulties.	Not assessed in this paper.
Gonsalves et al [48], 2021	POD Adventures	Questionnaires and semistructured interviews	Satisfaction scores ranged from good to excellent.	Intervention completion rate was 92%.	App generally considered easy to use, but a few participants identified confusing game components and issues related to typing and difficult log-in passwords.	Most participants felt that the program had positively impacted their prioritized problem.	As reported in [47].	Outcomes at 4 weeks showed significant improvements on all measures.

XSL•FO
RenderX

Reference, year	App name	Measurement	Acceptability and feasibility	Engagement	Barriers to engagement	Perceived usefulness	Accessibility and inclusivity	Intervention outcomes
Haug et al [49], 2017	Ready4life	Questionnaires	Large proportion of invited adolescents participated. Program evaluated as "very good" or "good" by 94.6% of participants.	Follow-up assessments were completed by 49.7% of the participants. Of the 39 program activities, the mean number carried out was 15.5. In total, 15% failed to engage in any activity, and 52% engaged in fewer than half of the activities.	Participation in the program was lower in male participants and among those reporting an immigrant background.	Not explicitly reported.	Not explicitly reported.	Statistically significant increases in targeted life skills, decline in at-risk alcohol use, and stable rates for tobacco and cannabis use.
Leonard et al [50], 2018	Calm Mom	Technology logs, questionnaires, and in-depth semistructured interviews	Participants felt the general content of app was highly relevant. 75% were "very" satisfied, and 18% were somewhat satisfied.	Mean of 14.77 minutes of using the app. Participants used at least one of the elements on the app on average on 44% of days when they had the study phone.	Technology challenges.	Supported their ability to effectively regulate their emotions.	Majority of participants noted that the app was very accessible, and several indicated that they felt less alone and felt genuinely cared for.	Not assessed in this paper.
Schueller et al [51], 2019	Pocket Helper, Purple Chill, and Slumber Time	Questionnaires	Satisfaction was high; 100% of participants would recommend the program, and 52% reported that they were very or extremely satisfied app.	57% of the participants completed all 3 sessions. Mean 2.09 sessions.	Mobile phone loss (through damage, theft, or other loss).	43% reported app as helpful; 48% found the skills they learned to be beneficial; 43% regularly used the skills.	The apps were preinstalled on all mobile phones before distribution to participants.	Participants experienced limited change on clinical outcomes with small effect sizes.
Escobar-Viera et al [52][a], 2023	REALbot	Questionnaires	High user satisfaction. Acceptability rated 5.3/7 but only 25% of participants described the app as exciting or leading edge.	42% of participants interacted with the app for 2 or more days.	Primary challenges were that app felt robotic and not smart enough.	Usability ratings were high on both measures.	Lack of voiceover feature.	Nonsignificant changes in scores of perceived isolation, depressive symptoms and social media self-efficacy.
Torok et al [53][a], 2022	LifeBuoy	Questionnaires	Not reported	71.5% completed 5 or more modules (completers)	Participants who completed first survey had significantly lower baseline anxiety symptoms compared to those who did not complete it.	Not reported	Not reported	Depression, anxiety, distress, and well-being symptoms improved in app group and control.

XSL•FO
RenderX

Reference, year	App name	Measurement	Acceptability and feasibility	Engagement	Barriers to engagement	Perceived usefulness	Accessibility and inclusivity	Intervention outcomes
Yang and Chung [54][a], 2022	HARU ASD	Questionnaires	Acceptable scores in the Satisfaction and Usability Questionnaire.	No participants dropped out.	Not reported	Not reported	Not reported	Significant decrease in anxiety level, an increase in positive affect, and a decline in stereotypic behaviors, hyperactivity, noncompliance, and inappropriate speech.

[a]Studies identified in the updated search.

Regarding co-design strategies used in these studies, early stakeholder consultation and service provider focus groups were conducted in the early development phase of POD Adventures, a gamified intervention for people with or at risk of anxiety, depression, and conduct difficulties in India; the results highlighted that "self-help" was not a culturally congruent concept for most Indian adolescents [47]. This early feedback was important as it revealed the norms around seeking or receiving direct instruction from parents, teachers, and other elders and that support from a counselor might be necessary to ensure acceptability, feasibility, and engagement [47]. The app was therefore designed to incorporate a combination of teaching methods, including direct instruction, modeling, and practice to accommodate different learning styles and to emphasize self-efficacy [47]. Furthermore, user testing also highlighted the need for more direct language, particularly around problem-solving concepts [47]. The iterative study methodology used in this study enabled the participants to guide the development and provide their inputs at each stage to increase acceptability and feasibility.

Engagement

Although, overall, the apps were well received by young people, poor engagement (eg, not engaging at the recommended frequency or complete the full course of the intervention), measured through both self-report, intervention adherence, and data capture was a commonly reported issue. Many studies failed to achieve continued participation, with high rates of attrition [41,42,44]. In addition, app use often decreased markedly after the initial few weeks [41,42,44]. The results of some studies suggested that engagement (in the form of intervention completion) was related to the effectiveness the intervention [41,42]. Although engagement was problematic in many of the stand-alone interventions, engagement and study participation in a school setting seemed more promising [49]. For example, a proactive invitation for study participation in a school enabled 4 out of 5 eligible adolescents to participate in the "ready4life" life skills program [49]. This strategy consisted of individuals who were trained in the program to be delivered, giving arranged sessions lasting 30 minutes in participating vocational schools during regular school lessons reserved for health education. Within this session, the students were informed about and invited to participate in the study, including being informed about the study's aims and assessments, reimbursement, and data protection.

Barriers to Engagement

Qualitative interviews and user feedback provided important insights about relevant barriers to engagement. The most frequently mentioned reasons for not using the app were that participants could not see the obvious benefits of using the app [41], lack of time or forgetting [42], and technical difficulties [41,50,51]. In a life skills training app for vocational students, participation was lower in male adolescents and among those reporting an immigrant background [49], although the reasons behind this poor engagement remained unclear.

What Do Young People Want From Apps?

There was some heterogeneity between studies in terms of the features and content that participants found acceptable and appropriate. For example, findings suggested that young people who experience homelessness tended to prefer both automated and self-help features compared with ones involving more direct human interaction [46]. However, participants in other studies valued both human interaction with professionals either via the app interface or through face-to-face contact and self-help features [47]. Human support was suggested as being helpful in offering both instruction and guidance as well as personalized support when needed. Numerous participants wanted opportunities to interact with peers [43,44,52] and even suggested connecting apps to social media [41]. Others also wanted the design of the apps to be more attractive (eg, improve the layout and create a more intuitive structure) and made suggestions about how gamifying apps could make them more interesting [41,43,47].

Qualitative Study

Overview

We interviewed 13 young men in the United Kingdom (age: mean 18.7, SD 2.5 y), 2 in Spain (age: mean 17, SD 0 y), and 5 in Germany (age: mean 20.2, SD 1.6 y). In the United Kingdom, 62% (8/13) of the participants self-reported as ethnically White, compared with 50% (1/2) in Spain and 20% (1/5) in Germany (Table 3).

XSL•FO
RenderX

Table 3. Participant characteristics.

	United Kingdom (n=13)	Spain (n=2)	Germany (n=5)
Age (y), mean (SD)	18.7 (2.5)	17 (0)	20.2 (1.6)
Ethnicity, n (%)			
Arab or Middle Eastern	0 (0)	0 (0)	4 (80)
Asian	4 (31)	0 (0)	0 (0)
White	8 (62)	1 (50)	1 (20)
Other ethnic group	1 (8)	1 (50)	0 (0)
Refugee or an asylum seeker, n (%)	0 (0)	0 (0)	5 (100)
Chronic medical condition, n (%)	0 (0)	0 (0)	0 (0)
Disability, n (%)	1 (8)	0 (0)	0 (0)
Educational attainment, n (%)			
Lower secondary school	6 (46)	2 (100)	4 (80)
Upper secondary school	4 (31)	0 (0)	1 (20)
Other higher education	1 (8)	0 (0)	0 (0)
Undergraduate degree	1 (8)	0 (0)	0 (0)
Postgraduate degree	1 (8)	0 (0)	0 (0)

In terms of participants' mental health and well-being (Table 4), the mean Patient Health Questionnaire-9 score in the United Kingdom was 9.7 (SD 7.3) compared with 5 (SD 1.4) in Spain.

Table 4. Mental health and well-being.

Measures	United Kingdom (n=13), mean (SD)	Spain (n=2), mean (SD)	Germany (n=5), mean (SD)
WEMWBS[a]	44.2 (7.8)	51.5 (3.5)	—[b]
PHQ-9[c]	9.7 (7.3)	5 (1.4)	—
GAD-7[d]	6.5 (4.4)	8.5 (2.1)	—

[a]WEBWBS: Warwick-Edinburgh Mental Well-Being Scale.

[b]Not available.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]GAD-7: Generalized Anxiety Disorder Assessment.

Acceptability and Feasibility

A key finding was that despite best efforts and financial incentives, recruiting underserved young male participants, especially in Spain and Germany, was challenging. This might suggest that these young people may not deem such an emotional competence app as relevant or useful to them, making recruitment and engagement problematic. We also assessed if the app was deemed acceptable (ie, useful, agreeable, palatable, or satisfactory) and appropriate (ie, relevant, suitable, or compatible). Overall, the app was viewed by participants in the United Kingdom, Spain, and Germany as being appropriate and relevant for young people of different ages and walks of life, as they thought that all young people had a smartphone and were adept at using technology:

> *So, I was able to learn about my feelings, I was able to evaluate how I actually felt today, concerning my feelings, if I was angry or I was sad. I was actually able to write them down in detail.* [Participant in Germany]

Several participants commented that the content of the app was best suited to university and school students. Another common view was that the app was better suited to those struggling with their mental health and that it was less relevant for those for whom things were going well. Many participants perceived the app to be aimed at improving mental health problems, as opposed to being a universal intervention intended to improve well-being, which represented a barrier to engagement. Of those who reported that the app was not relevant to them, they did see it as being of potential use to friends and family members who were stressed, anxious, or going through a difficult time:

> *There will be folks who maybe aren't going through a good time in their lives, and they will need the app to feel... to understand themselves, mostly. And I think it's relevant at any age, because I am lucky that I don't think I need it as much as someone else who feels like that.* [Participant in Spain]

XSL•FO
RenderX

> *Partly it was important, partly it was not. I'll give an example again, for example if a refugee came to Germany from a war zone, it's going to be difficult, very difficult to find a topic that would fit him, for the future I mean, so the version now is already okay if you want all persons to use this app. Partly it's already relevant and partly it's not. If someone has mental problems or bad experiences, you cannot find such a topic in the app.* [Participant in Germany]

Engagement

Although some participants reported using the app regularly during the 4-week study period, a consistent finding was that participants tended to use the app most when they first downloaded it, with a marked reduction in use over time:

> *Uh, I probably used it about three times in the first week. And then not really that much at all I'm afraid.* [Participant in the United Kingdom]

> *I don't know, I just dropped off using a little bit after a couple of weeks, but I've been trying to keep on top of doing that like the daily rating things and everything.... I kind of lost my motivation to use it.* [Participant in the United Kingdom]

Barriers to Engagement

We identified several barriers that hindered participants' engagement and use of the app. These included the following: (1) repetitive and time-consuming app contents, (2) a paucity of new content and personalized or interactive tools (eg, matching mood to tools), (3) unclear instructions, (4) a lack of rationale for the app, (5) perceiving the app as not being relevant, (6) a lack of motivation, and (7) privacy concerns:

> *Yes, for example, I would not like to write in this diary, because I do not know if it would be one hundred percent anonymous and if others might read it. And maybe I have more privacy if I do not write it.* [Participant in Germany]

> *I think by now I would slowly stop using the app. It was nice up to this point, but I think for me I might need a step further now. To really deal with my personal problems and I don't know how much an app like this can help and that rather an expert and therapy is needed.* [Participant in Germany]

For the asylum seekers and refugees in Germany, the language and content of the app was not suited to their needs. The participants would have preferred the app in their native language as some had to use translation programs to help access the content. Furthermore, specific topics of relevance to refugees were missing, such as dealing with asylum uncertainty, whereabouts of family members, and their living situation.

Finally, underserved young people, including asylum seekers and refugees, migrants, and those NEET, are more likely to experience financial deprivations and therefore less likely to pay directly for apps, especially for those that do not address their primary difficulties:

> *If it came to the point that I had to pay for it, I would look for free options.* [Participant in Germany]

Discussion

Summary of Findings

The use of mobile apps in mental health care continues to attract interest and investment; however, research geared toward understanding the needs of marginalized and underserved populations is still nascent. This study, focusing on the implementation of mental health apps in underserved young people, highlighted that little research exists to support the widespread adoption of these apps as a mental health intervention for marginalized and underserved groups. Findings from both our systematic review and qualitative study were largely consistent: markers of acceptability and usability were positive; however, engagement for underserved young people was low, which is notable given the widespread ownership of smartphones [55,56]. To date, research has focused primarily on efficacy studies rather than effectiveness and implementation in "real-world" settings and may have overestimated users' "natural tendency" to adopt smartphone apps for their mental health and well-being [57]. Our findings suggest that despite the rapid proliferation of mobile mental health technology, the uptake and engagement of mental health apps among marginalized young people are low and remain a key implementation challenge.

Our data suggest that establishing and maintaining user commitment and engagement in the content of the intervention as intended is a pervasive challenge across mental health apps and marginalized populations, and premature dropout was prominent in nearly all the included studies. This is consistent with the literature that suggests that the majority of those offered these app-based interventions do not engage at the recommended frequency or complete the full course of treatment [58,59]. In this study, various app components were associated with engagement level, with the most engaging interventions providing young people with some form of associated real-human interaction and those having a more interactive interface. This aligned with other findings that the feedback of personalized information to participants is an especially important aspect of creating engaging and impactful digital tools [60]. Young people tend to quickly disengage if there are technical difficulties or if the app does not specifically target their perceived needs [41,50,51]. Furthermore, recruitment of marginalized groups to app-based studies is difficult. For instance, in this study, the use of advertisements, financial incentives, vouchers, and prize draw incentives seemed to be insufficient to recruit a significant number of participants in Spain and Germany.

Measuring engagement is a challenge that has likely contributed to our lack of knowledge on app components that effectively increase user engagement. Reporting engagement with mental health apps in intervention trials is highly variable, and a number of basic metrics of intervention engagement, such as rate of intervention uptake, weekly use patterns, and number of intervention completers, are available, yet not routinely reported [58,59]. The results of this study highlight the importance of objective engagement measures and that relying on positive subjective self-reports of usability, satisfaction, acceptability,

XSL•FO
RenderX

and feasibility is insufficient to determine actual engagement. Furthermore, the findings suggest that apps involving human interactions with a professional (eg, therapist or counselor) or that are completed in a supervised setting tend to be more acceptable and effective and have higher engagement rates [47,48]. Our research suggests that similar to traditional face-to-face mental health services, app-based programs still face numerous barriers to reach marginalized youth, especially since the mental health apps available to the public do not seem to consider the unique developmental needs of these groups, participants do not seem to perceive an obvious benefit from using them, and some potential users prefer to interact with a professional face to face. Thus, it is also possible that the digital mental health field might be inadvertently contributing to mental health inequities among this population by not engaging marginalized groups sufficiently at the outset of research to ensure that the designed app meets their needs. However, for the studies included in this study that did engage these groups in the co-design of the apps, there was no notable improvement in engagement. Thus, we hope these findings encourage researchers and clinicians to think more critically of the role that mental health apps can truly have in addressing mental health equities among underserved groups.

As in other areas of mental health research, young people from LMICs were underrepresented in these studies, which typically originated from high-income settings, including the United States, Australia, and Canada. There are relatively few app-based interventions that were designed or adapted for young people in LMICs that have been rigorously evaluated or are even available in local languages [47,48]. Many living in LMIC regions, for example, adults in Asian countries, are often faced with apps that are not culturally relevant or in the right language [61]. These inequities are surprising given the high rates of smartphone use in Asia, even in rural regions [62]. Yet, it is still likely that youth in this region faced barriers related to data availability and more limited phone access, which will likely inhibit the broad implementation of apps beyond research studies [16]. Considerable work is required to ensure the availability of mental health apps that fit a wide range of user needs and preferences. It is important to ensure that the acceptability and feasibility of mental health apps for young people residing in LMICs are prioritized so that they are not further excluded from relevant mental health research.

Finally, a significant challenge is the lack of diversity in mental health app research participation, which limits our understanding of “real-world” efficacy and implementation for underserved and marginalized groups. While undoubtedly invaluable, and indeed deemed gold standard when evaluating efficacy of interventions, randomized controlled trial of mental health apps are not without flaws [63,64]. Trial recruitment is often highly selective due to stringent inclusion and exclusion criteria resulting in lower inclusion in research than one would expect from population estimates [65]. In the United Kingdom, the National Institute for Health and Care Research data have revealed that geographies with the highest burden of disease also have the lowest number of patients taking part in research [66]. The postcodes in which research recruitment is low also aligns closely to areas where earnings are the lowest and indexes of deprivation are the highest [66]. There are many reasons why some groups are underrepresented in research: language barriers, culturally inappropriate explanations, poor health literacy and the use of jargon, communication not being suitable for people with special learning needs, requirement to complete many administrative forms, negative financial impact in participating, lack of effective incentives for participation, or lack of clarity around incentives, and specific cultural and religious beliefs [66]. Failing to include a broad range of participants is problematic in that results may not be generalizable to a broad population.

Limitations

Although this research was carefully executed and used a robust methodological approach with an exhaustive search strategy, it is not without limitations. Foremost, although the systematic review attempted to identify and include as many articles as possible, some papers may have been missed because of the inconsistencies in how feasibility and acceptability outcomes are recorded and reported. It was also difficult to ensure that all apps for this age group were identified because those aged between 15 and 25 years are harder to differentiate in adolescent and adult studies, meaning we might have missed some relevant studies where data could not be disaggregated by age. The exclusion of gray literature (eg, institutional reports and websites) may have also made us overlook potentially relevant apps, albeit lacking the quality assurance of peer-reviewed research. It is also likely that commercial organizations, including app companies, collect rich user demographic and engagement data but do not share it publicly, thus limiting our ability to conduct empirical analyses about the “real-world” acceptability, engagement, and implementation for specific populations. We did not analyze the extent to which publication bias may have influenced the results of our search, and, therefore, there may be a much higher number of mental health apps that have been developed with an underserved sample of young people, but due to their lack of efficacy or acceptability, these studies have not been submitted or accepted for publication. The sample sizes of many of the included studies were relatively low, which potentially limits their generalizability. However, we included all study designs so as to ensure that our learning from existing research was maximized. Furthermore, many of the studies included in the systematic review, as well as our qualitative study, had some form of language competency as an inclusion criterion (eg, English speaking), which likely excludes important perspectives from the results. For the qualitative study, we were only able to gather data from those who had used the app at least once and who were therefore somewhat engaged in the app. Despite our best efforts, we were unable to recruit participants who, following consent, had never then downloaded or used the app and so could not explore barriers to engagement for the least engaged young people or understand why the app was not appealing to those who chose not to proceed or take part. Those who did participate in this research were financially incentivized to do so and often highlighted the importance of this incentive in keeping them engaged. Therefore, we were unable to draw conclusions about the naturalistic engagement, feasibility, and acceptability of the app, if it were to be made available without

XSL•FO
RenderX

payment in schools, universities, and health services or to be made commercially available on the app marketplace. It is also possible that social desirability bias (ie, a tendency to present reality to align with what is perceived to be socially acceptable) occurred during the interviews, whereby participants responded to the interview questions in a manner that they believed would be more acceptable to the study team, concealing their true opinions or experiences [67,68]. As previously noted by others, results may be subject to further bias in that findings could be led by more articulate young people, while it is more difficult to hear the voices of those who are less articulate or digitally literate [69]. Finally, it is also possible that the positionality of the research team, including our own experiences, backgrounds, and biases, impacted what information participants disclosed to the research team as well as the interpretation of the qualitative data in this study.

Recommendations of Adaptations to Increase Acceptability, Feasibility, and Engagement

To overcome this complex engagement and implementation challenge, we have taken together our findings with relevant previous literature to generate 3 key suggestions about how to improve the feasibility and potential utility of apps for young people from marginalized and underserved populations.

Increasing Participant Diversity in Mental Health Intervention Research

Studies should aim to prioritize the inclusion of marginalized groups in trials testing the effectiveness of digital interventions by intentionally planning recruitment efforts aimed to reach these communities [70]. First, steps can be taken to build trust, connections, and credibility between the research team and these communities. NHS England [66] suggests involving representatives from those groups during the inception and implementation of recruitment efforts. This approach ensures that the intervention is relevant to the target group by meeting their preferences and needs, incorporating culturally salient factors relevant for recruitment efforts, addressing concerns about community mistrust and participant resource constraints, and establishing partnerships with key community stakeholders that can be gatekeepers in the community [14,71]. These strategies are likely to improve research accessibility, recruitment, and retention. Research teams need to ensure that the findings and any actionable takeaways from the research conducted with the participants are shared with them by asking participants how they would like to receive this information (eg, verbal, written, or via a trusted advocate). Equally important is to explain that the research process can be slow. These steps help create a positive legacy for the research project and build trust between individuals and public institutions, helping future health researchers to further address underrepresentation of marginalized groups in digital research.

Identifying and Addressing Needs and Preferences of Underserved and Marginalized Groups by Using Human-Centered Design Principles

A comprehensive understanding of the needs, challenges, and life circumstances of the target population is a key implementation driver for designing relevant, engaging, and effective mental health apps. This knowledge is particularly important when the app is a stand-alone intervention received during daily life outside of traditional psychotherapy or human support [50]. This goal can be best achieved through a participatory approach, which reflects a growing recognition among intervention researchers and developers that end users need to be involved in the creation of interventions and their future iterations [47,72]. This process may involve a series of stages, including (1) person-centered co-design to ensure that tools are developed to be acceptable to the underserved or marginalized populations as well as meet their specific needs, life circumstances, and cultural norms [47]; (2) iterative testing that incorporates users' feedback on a rolling basis to ensure the relevance of the intervention [43,47,72]; and (3) changes and adaptations needed to meet users' needs in "real-world" settings including consideration of economic viability and implementation [27].

Especially relevant for the underserved and marginalized groups is the need (or lack thereof) to culturally adapt app interventions for specific racial, ethnic, or cultural groups through this person-centered design. In traditional face-to-face interventions, some have argued that all treatments need to be culturally adapted to ensure their validity, relevance, and effectiveness since these interventions are often developed with individuals who can be substantially different from some marginalized populations [73]. Similar to culturally adapted face-to-face interventions [74-76], culturally adapted digital mental health interventions seem to be effective [77,78]. However, there is no evidence that these culturally adapted interventions outperform the original programs [79,80]. Given that culturally adapting digital interventions is a time-consuming and resource-intensive process, this approach may not be sustainable and limit the dissemination and implementation impact of app programs [28]. In lieu of culturally adapting digital interventions without careful consideration, Ramos and Chavira [28] recommend using information gathered through person-centered approaches to integrate culture into the use of already available digital interventions (including apps), using an idiographic, flexible, and personalized approach. This strategy may have a broader implementation and dissemination potential, given that few researchers and clinicians are in a position to develop new apps.

Embedding Apps Within Existing Care Structures

Several systematic reviews and meta-analyses have demonstrated that app-based mental health interventions with a human-support component are more effective and more acceptable than stand-alone, fully automatized, or self-administered apps [13,25,81]. Young people seem to want practical skills and usable tools to apply to their current daily life stressors to improve their well-being and functioning. Intervention engagement is enhanced if the intervention serves an obvious purpose, is relevant, and has a clear rationale and instructions, and embedding these interventions within the systems and structures that are already working with users (eg, clinical services, schools, universities, and community agencies) will likely improve implementation. Considering the broad and highly varied nature of intervention formats and modalities, it may be useful for future research to focus on identifying core

XSL•FO
RenderX

components of app-based interventions (ie, active ingredients of interventions associated with uptake, adherence, and clinical outcomes) that will allow such integration of app interventions into the varied context of care for marginalized youth.

Conclusions

Despite the enthusiasm that has surrounded the potential of digital technologies to revolutionize mental health and health care service delivery, little evidence yet supports the use of mental health apps for marginalized and underserved young people. Despite the substantial financial and human investment directed to the development of mental health apps over several years, only a small proportion have empirical evidence to support their effectiveness, and there have been few attempts to develop or adapt interventions to meet some of the more unique and heterogeneous needs of diverse groups of young people. Although acceptability seems to be good, engagement is poor and attrition is high, particularly if not supported by in-person elements. Given that most interventions are implemented in high-income countries, very little is known about the generalizability of the findings to LMICs and to a range of adolescents and young people with different socioeconomic, cultural, and racial backgrounds. In this paper, we have drawn several insights about the feasibility and acceptability of mental health apps for underserved young people that may be useful to future app-based mental health promotion and treatment projects. However, before the widespread adoption and scaling-up of digital mental health interventions progresses further, especially for more vulnerable and underserved populations and in settings with limited resources, a greater understanding is needed on the unique barriers faced by these groups in accessing treatment and the types of services young people themselves prefer (eg, standard vs digital) followed by more rigorous and consistent demonstrations of feasibility, effectiveness, and cost-effectiveness.

Acknowledgments

This project received funding from the European Union's Horizon 2020 research and innovation program (grant agreement number 754657).

The authors are grateful to the young people who took the time to participate in this research and who shared their insights with us. The authors would also like to thank those who supported this research including professional youth advisor Emily Bampton, research assistant Catherine Reeve, and researchers Dr Alexandra Langmeyer and Simon Weiser. Finally, the authors would like to thank the ECoWeB (Emotional Competence for Well-Being) Consortium for their support and feedback throughout the duration of this research, including, but not limited to, Dr Lexy Newbold, Dr Azucena Garcia Palacios, and Dr Guadalupe Molinari.

Data Availability

The data extracted to support the findings of the systematic review are available from the corresponding author upon reasonable request. Due to the confidential and sensitive nature of the interview transcripts, qualitative data will not be made available.

Authors' Contributions

HAB, LAN, and MF designed the systematic review including the research questions and methods. LAN carried out the database search. HAB, LAN, TM, and BF conducted the study screening and data extraction. TM did the study quality assessments, and HAB did the data synthesis and analysis. MF, SW, EW, and HAB were involved in the conception of the qualitative study. HAB, LAN, and SC conducted the quality study including conducting the qualitative interviews and analysis. HAB wrote the first draft and HAB, LAN, MF, and GR contributed substantially to manuscript drafting. All authors contributed to the manuscript and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 43 KB - jmir_v26i1e48964_app1.docx]

Multimedia Appendix 2
Topic guide.
[DOCX File , 25 KB - jmir_v26i1e48964_app2.docx]

Multimedia Appendix 3
Study quality assessment.
[XLSX File (Microsoft Excel File), 30 KB - jmir_v26i1e48964_app3.xlsx]

XSL•FO
RenderX

References

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
2. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
3. Brown A, Rice SM, Rickwood DJ, Parker AG. Systematic review of barriers and facilitators to accessing and engaging with mental health care among at-risk young people. Asia Pac Psychiatry 2016 Mar 03;8(1):3-22. [doi: 10.1111/appy.12199] [Medline: 26238088]
4. Marrast L, Himmelstein DU, Woolhandler S. Racial and ethnic disparities in mental health care for children and young adults: a national study. Int J Health Serv 2016 Oct 20;46(4):810-824. [doi: 10.1177/0020731416662736] [Medline: 27520100]
5. Cook BL, Trinh NH, Li Z, Hou SS, Progovac AM. Trends in racial-ethnic disparities in access to mental health care, 2004-2012. Psychiatr Serv 2017 Jan 01;68(1):9-16 [FREE Full text] [doi: 10.1176/appi.ps.201500453] [Medline: 27476805]
6. Rimes KA, Shivakumar S, Ussher G, Baker D, Rahman Q, West E. Psychosocial factors associated with suicide attempts, ideation, and future risk in lesbian, gay, and bisexual youth. Crisis 2019 Mar;40(2):83-92. [doi: 10.1027/0227-5910/a000527] [Medline: 29932021]
7. Reiss F. Socioeconomic inequalities and mental health problems in children and adolescents: a systematic review. Soc Sci Med 2013 Aug;90:24-31. [doi: 10.1016/j.socscimed.2013.04.026] [Medline: 23746605]
8. Warnecke RB, Oh A, Breen N, Gehlert S, Paskett E, Tucker KL, et al. Approaching health disparities from a population perspective: the National Institutes of Health Centers for Population Health and Health Disparities. Am J Public Health 2008 Sep;98(9):1608-1615. [doi: 10.2105/ajph.2006.102525]
9. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
10. Introduction: NIH minority health and health disparities strategic plan 2021-2025. National Institutes of Health National Institute on Minority Health and Health Disparities. URL: https://www.nimhd.nih.gov/about/strategic-plan/nih-strategic-plan-directors-foreword.html [accessed 2024-06-29]
11. Improving inclusion of under-served groups in clinical research: Guidance from INCLUDE project. National Institute for Health and Care Excellence. 2020 Aug 7. URL: https://www.nihr.ac.uk/documents/improving-inclusion-of-under-served-groups-in-clinical-research-guidance-from-include-project/25435 [accessed 2024-06-29]
12. Evans-Lacko S, Aguilar-Gaxiola S, Al-Hamzawi A, Alonso J, Benjet C, Bruffaerts R, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. Psychol Med 2017 Nov 27;48(9):1560-1571. [doi: 10.1017/s0033291717003336]
13. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual research review: digital health interventions for children and young people with mental health problems - a systematic and meta-review. J Child Psychol Psychiatry 2017 Apr 10;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
14. Friis-Healy EA, Nagy GA, Kollins SH. It is time to REACT: opportunities for digital mental health apps to reduce mental health disparities in racially and ethnically minoritized groups. JMIR Ment Health 2021 Jan 26;8(1):e25456 [FREE Full text] [doi: 10.2196/25456] [Medline: 33406050]
15. Mobile fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-07-12]
16. Silver L. 2. In emerging economies, smartphone adoption has grown more quickly among younger generations. Pew Research Center. 2019 Feb 5. URL: https://www.pewresearch.org/global/2019/02/05/in-emerging-economies-smartphone-adoption-has-grown-more-quickly-among-younger-generations/ [accessed 2024-07-12]
17. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019 Nov 19;21(11):e13873 [FREE Full text] [doi: 10.2196/13873] [Medline: 31742562]
18. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
19. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
20. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: a systematic review of randomised controlled trials. Child Youth Serv Rev 2021 Aug;127:106073. [doi: 10.1016/j.childyouth.2021.106073]
21. Buttazzoni A, Brar K, Minaker L. Smartphone-based interventions and internalizing disorders in youth: systematic review and meta-analysis. J Med Internet Res 2021 Jan 11;23(1):e16490 [FREE Full text] [doi: 10.2196/16490] [Medline: 33427682]
22. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 08;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]

XSL•FO
RenderX

23. Lipschitz J, Hogan TP, Bauer MS, Mohr DC. Closing the research-to-practice gap in digital psychiatry: the need to integrate implementation science. J Clin Psychiatry 2019 May 14;80(3):18com12659 [FREE Full text] [doi: 10.4088/JCP.18com12659] [Medline: 31091029]
24. Hollis C. Youth mental health: risks and opportunities in the digital world. World Psychiatry 2022 Feb 11;21(1):81-82 [FREE Full text] [doi: 10.1002/wps.20929] [Medline: 35015352]
25. Bear HA, Ayala Nunes L, DeJesus J, Liverpool S, Moltrecht B, Neelakantan L, et al. Determination of markers of successful implementation of mental health apps for young people: systematic review. J Med Internet Res 2022 Nov 09;24(11):e40347 [FREE Full text] [doi: 10.2196/40347] [Medline: 36350704]
26. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar 19;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
27. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021 Dec;147:103990 [FREE Full text] [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
28. Ramos G, Chavira DA. Use of technology to provide mental health care for racial and ethnic minorities: evidence, promise, and challenges. Cognit Behav Pract 2022 Feb;29(1):15-40. [doi: 10.1016/j.cbpra.2019.10.004]
29. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021 Sep 6;3:697072 [FREE Full text] [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
30. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
31. Marcano Belisario JS, Huckvale K, Greenfield G, Car J, Gunn LH. Smartphone and tablet self management apps for asthma. Cochrane Database Syst Rev 2013 Nov 27;2013(11):CD010013 [FREE Full text] [doi: 10.1002/14651858.CD010013.pub2] [Medline: 24282112]
32. Marcano Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison CP, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. Cochrane Database Syst Rev 2015 Jul 27;2015(7):MR000042 [FREE Full text] [doi: 10.1002/14651858.MR000042.pub2] [Medline: 26212714]
33. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
34. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo MP, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/efi-180221]
35. Crabtree BF, Miller WL. A template approach to text analysis: developing and using codebooks. In: Crabtree BF, Miller WL, editors. Doing Qualitative Research. Thousand Oaks, CA: SAGE Publications, Inc; 1992:93-109.
36. Newbold A, Warren FC, Taylor RS, Hulme C, Burnett S, Aas B, et al. Promotion of mental health in young adults via mobile phone app: study protocol of the ECoWeB (emotional competence for well-being in young adults) cohort multiple randomised trials. BMC Psychiatry 2020 Sep 22;20(1):458 [FREE Full text] [doi: 10.1186/s12888-020-02857-w] [Medline: 32962684]
37. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
38. Ellaway RH, Kehoe A, Illing J. Critical realism and realist inquiry in medical education. Acad Med 2020 Jul;95(7):984-988. [doi: 10.1097/ACM.0000000000003232] [Medline: 32101916]
39. Willig C. Perspectives on the epistemological bases for qualitative research. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology, Vol. 1. Foundations, Planning, Measures, and Psychometrics. Washington, DC: American Psychological Association; 2012.
40. Fletcher AJ. Applying critical realism in qualitative research: methodology meets method. Int J Soc Res Methodol 2016 Feb 29;20(2):181-194. [doi: 10.1080/13645579.2016.1144401]
41. Bohleber L, Crameri A, Eich-Stierli B, Telesko R, von Wyl A. Can we foster a culture of peer support and promote mental health in adolescence using a web-based app? A control group study. JMIR Ment Health 2016 Sep 23;3(3):e45 [FREE Full text] [doi: 10.2196/mental.5597] [Medline: 27663691]
42. Deady M, Glozier N, Collins D, Einboden R, Lavender I, Wray A, et al. The utility of a mental health app in apprentice workers: a pilot study. Front Public Health 2020 Sep 4;8:389 [FREE Full text] [doi: 10.3389/fpubh.2020.00389] [Medline: 33014953]
43. Fleming JB, Hill YN, Burns MN. Usability of a culturally informed mHealth intervention for symptoms of anxiety and depression: feedback from young sexual minority men. JMIR Hum Factors 2017 Aug 25;4(3):e22 [FREE Full text] [doi: 10.2196/humanfactors.7392] [Medline: 28842389]
44. Francis J, Cross D, Schultz A, Armstrong D, Nguyen R, Branch-Smith C. Developing a smartphone application to support social connectedness and wellbeing in young people with cystic fibrosis. J Cyst Fibros 2020 Mar;19(2):277-283 [FREE Full text] [doi: 10.1016/j.jcf.2019.12.011] [Medline: 31917112]

XSL•FO
RenderX

45. Geirhos A, Domhardt M, Lunkenheimer F, Temming S, Holl RW, Minden K, et al. Feasibility and potential efficacy of a guided internet- and mobile-based CBT for adolescents and young adults with chronic medical conditions and comorbid depression or anxiety symptoms (youthCOACH): a randomized controlled pilot trial. BMC Pediatr 2022 Jan 29;22(1):69 [FREE Full text] [doi: 10.1186/s12887-022-03134-3] [Medline: 35093047]
46. Glover AC, Schueller SM, Winiarski DA, Smith DL, Karnik NS, Zalta AK. Automated mobile phone-based mental health resource for homeless youth: pilot study assessing feasibility and acceptability. JMIR Ment Health 2019 Oct 11;6(10):e15144 [FREE Full text] [doi: 10.2196/15144] [Medline: 31605516]
47. Gonsalves PP, Hodgson ES, Kumar A, Aurora T, Chandak Y, Sharma R, et al. Design and development of the "POD adventures" smartphone game: a blended problem-solving intervention for adolescent mental health in India. Front Public Health 2019 Aug 23;7:238 [FREE Full text] [doi: 10.3389/fpubh.2019.00238] [Medline: 31508404]
48. Gonsalves PP, Hodgson ES, Bhat B, Sharma R, Jambhale A, Michelson D, et al. App-based guided problem-solving intervention for adolescent mental health: a pilot cohort study in Indian schools. Evid Based Ment Health 2021 Feb 18;24(1):11-18 [FREE Full text] [doi: 10.1136/ebmental-2020-300194] [Medline: 33208507]
49. Haug S, Paz Castro R, Meyer C, Filler A, Kowatsch T, Schaub MP. A mobile phone-based life skills training program for substance use prevention among adolescents: pre-post study on the acceptance and potential effectiveness of the program, Ready4life. JMIR Mhealth Uhealth 2017 Oct 04;5(10):e143 [FREE Full text] [doi: 10.2196/mhealth.8474] [Medline: 28978498]
50. Leonard NR, Casarjian B, Fletcher RR, Praia C, Sherpa D, Kelemen A, et al. Theoretically-based emotion regulation strategies using a mobile app and wearable sensor among homeless adolescent mothers: acceptability and feasibility study. JMIR Pediatr Parent 2018 Mar 01;1(1):e1 [FREE Full text] [doi: 10.2196/pediatrics.9037] [Medline: 30637376]
51. Schueller SM, Glover AC, Rufa AK, Dowdle CL, Gross GD, Karnik NS, et al. A mobile phone-based intervention to improve mental health among homeless young adults: pilot feasibility trial. JMIR Mhealth Uhealth 2019 Jul 02;7(7):e12347 [FREE Full text] [doi: 10.2196/12347] [Medline: 31267980]
52. Escobar-Viera CG, Porta G, Coulter RW, Martina J, Goldbach J, Rollman BL. A chatbot-delivered intervention for optimizing social media use and reducing perceived isolation among rural-living LGBTQ+ youth: development, acceptability, usability, satisfaction, and utility. Internet Interv 2023 Dec;34:100668 [FREE Full text] [doi: 10.1016/j.invent.2023.100668] [Medline: 37746640]
53. Torok M, Han J, McGillivray L, Wong Q, Werner-Seidler A, O'Dea B, et al. The effect of a therapeutic smartphone application on suicidal ideation in young adults: findings from a randomized controlled trial in Australia. PLoS Med 2022 May;19(5):e1003978 [FREE Full text] [doi: 10.1371/journal.pmed.1003978] [Medline: 35639672]
54. Yang YJ, Chung KM. Pilot randomized control trial of an app-based CBT program for reducing anxiety in individuals with ASD without intellectual disability. J Autism Dev Disord 2023 Apr;53(4):1331-1346. [doi: 10.1007/s10803-022-05617-9] [Medline: 35689137]
55. Children and parents: media use and attitudes report 2018. Ofcom. 2019 Mar 5. URL: https://www.ofcom.org.uk/media-use-and-attitudes/media-habits-children/children-and-parents-media-use-and-attitudes-report-2018/ [accessed 2024-07-12]
56. Silver L. Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. 2019 Feb 5. URL: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ [accessed 2024-07-12]
57. Chan A, Kow R, Cheng JK. Adolescents' perceptions on smartphone applications (apps) for health management. JournalMTM 2017 Aug;6(2):47-55. [doi: 10.7309/jmtm.6.2.6]
58. Lipschitz JM, Van Boxtel R, Torous J, Firth J, Lebovitz JG, Burdick KE, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022 Oct 14;24(10):e39204 [FREE Full text] [doi: 10.2196/39204] [Medline: 36240001]
59. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
60. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-health: a scoping review. Telemed J E Health 2022 Jun;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
61. Brian RM, Ben-Zeev D. Mobile health (mHealth) for mental health in Asia: objectives, strategies, and limitations. Asian J Psychiatr 2014 Aug;10:96-100. [doi: 10.1016/j.ajp.2014.04.006] [Medline: 25042960]
62. Li H, Lewis C, Chi H, Singleton G, Williams N. Mobile health applications for mental illnesses: an Asian context. Asian J Psychiatr 2020 Dec;54:102209 [FREE Full text] [doi: 10.1016/j.ajp.2020.102209] [Medline: 32623190]
63. Spielmans GI, Gatlin ET, McFall JP. The efficacy of evidence-based psychotherapies versus usual care for youths: controlling confounds in a meta-reanalysis. Psychother Res 2010 Mar 23;20(2):234-246. [doi: 10.1080/10503300903311293] [Medline: 19856234]
64. Rapport F, Smith J, Hutchinson K, Clay-Williams R, Churruca K, Bierbaum M, et al. Too much theory and not enough practice? The challenge of implementation science application in healthcare practice. J Eval Clin Pract 2022 Dec 15;28(6):991-1002. [doi: 10.1111/jep.13600] [Medline: 34268832]

XSL•FO
RenderX

65. Jensen SA, Corralejo SM. Measurement issues: large effect sizes do not mean most people get better - clinical significance and the importance of individual results. Child Adolesc Ment Health 2017 Sep;22(3):163-166. [doi: 10.1111/camh.12203] [Medline: 32680375]
66. Increasing diversity in research participation: a good practice guide for engaging with underrepresented groups. National Health Service England. 2023 Feb 14. URL: https://www.england.nhs.uk/aac/publication/increasing-diversity-in-research-participation/ [accessed 2023-03-16]
67. Bergen N, Labonté R. "Everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. Qual Health Res 2020 Apr;30(5):783-792. [doi: 10.1177/1049732319889354] [Medline: 31830860]
68. Vogt WP, Johnson RB. Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences. Thousand Oaks, CA: SAGE Publications; 2011.
69. Midgley N, Parkinson S, Holmes J, Stapley E, Eatough V, Target M. "Did I bring it on myself?" An exploratory study of the beliefs that adolescents referred to mental health services have about the causes of their depression. Eur Child Adolesc Psychiatry 2017 Jan 20;26(1):25-34 [FREE Full text] [doi: 10.1007/s00787-016-0868-8] [Medline: 27207089]
70. Willis A, Isaacs T, Khunti K. Improving diversity in research and trial participation: the challenges of language. Lancet Public Health 2021 Jul;6(7):e445-e446. [doi: 10.1016/s2468-2667(21)00100-6]
71. Ellard-Gray A, Jeffrey NK, Choubak M, Crann SE. Finding the hidden participant: solutions for recruiting hidden, hard-to-reach, and vulnerable populations. Int J Qual Method 2015 Dec 17;14(5). [doi: 10.1177/1609406915621420]
72. Moltrecht B, Patalay P, Bear HA, Deighton J, Edbrooke-Childs J. A transdiagnostic, emotion regulation app (Eda) for children: design, development, and lessons learned. JMIR Form Res 2022 Jan 19;6(1):e28300 [FREE Full text] [doi: 10.2196/28300] [Medline: 35044312]
73. Bernal G, Jiménez-Chafey MI, Domenech Rodríguez MM. Cultural adaptation of treatments: a resource for considering culture in evidence-based practice. Prof Psychol Res Pract 2009 Aug;40(4):361-368. [doi: 10.1037/a0016401]
74. Benish SG, Quintana S, Wampold BE. Culturally adapted psychotherapy and the legitimacy of myth: a direct-comparison meta-analysis. J Couns Psychol 2011 Jul;58(3):279-289. [doi: 10.1037/a0023626] [Medline: 21604860]
75. Hall GC, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016 Nov;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
76. Huey SJJ, Polo AJ. Evidence-based psychosocial treatments for ethnic minority youth. J Clin Child Adolesc Psychol 2008 Jan 03;37(1):262-301 [FREE Full text] [doi: 10.1080/15374410701820174] [Medline: 18444061]
77. Ellis DM, Draheim AA, Anderson PL. Culturally adapted digital mental health interventions for ethnic/racial minorities: a systematic review and meta-analysis. J Consult Clin Psychol 2022 Oct;90(10):717-733. [doi: 10.1037/ccp0000759] [Medline: 36227330]
78. Harper Shehadeh M, Heim E, Chowdhary N, Maercker A, Albanese E. Cultural adaptation of minimally guided interventions for common mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016 Sep 26;3(3):e44 [FREE Full text] [doi: 10.2196/mental.5776] [Medline: 27670598]
79. Balci S, Spanhel K, Sander LB, Baumeister H. Culturally adapting internet- and mobile-based health promotion interventions might not be worth the effort: a systematic review and meta-analysis. NPJ Digit Med 2022 Mar 23;5(1):34 [FREE Full text] [doi: 10.1038/s41746-022-00569-x] [Medline: 35322172]
80. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
81. Grist R, Croker A, Denne M, Stallard P. Technology delivered interventions for depression and anxiety in children and adolescents: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2019 Jun 18;22(2):147-171 [FREE Full text] [doi: 10.1007/s10567-018-0271-8] [Medline: 30229343]

Abbreviations

ECoWeB: Emotional Competence for Well-Being
LMIC: low- and middle-income country
MMAT: Mixed Methods Appraisal Tool
NEET: not in education, employment, or training
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, S Ma; submitted 13.05.23; peer-reviewed by P Whelan, I Vainieri, H Bao; comments to author 13.09.23; revised version received 26.09.23; accepted 10.06.24; published 30.07.24.

Please cite as:
Bear HA, Ayala Nunes L, Ramos G, Manchanda T, Fernandes B, Chabursky S, Walper S, Watkins E, Fazel M
The Acceptability, Engagement, and Feasibility of Mental Health Apps for Marginalized and Underserved Young People: Systematic Review and Qualitative Study
J Med Internet Res 2024;26:e48964
URL: https://www.jmir.org/2024/1/e48964
doi:10.2196/48964
PMID:39078699

©Holly Alice Bear, Lara Ayala Nunes, Giovanni Ramos, Tanya Manchanda, Blossom Fernandes, Sophia Chabursky, Sabine Walper, Edward Watkins, Mina Fazel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Experiences of Community-Dwelling Older Adults on Using Wearable Monitoring Devices With Regular Support From Community Health Workers, Nurses, and Social Workers: Qualitative Descriptive Study

Arkers Kwan Ching Wong[1], PhD; Jonathan Bayuo[1], PhD; Jing Jing Su[1], PhD; Karen Kit Sum Chow[2], MSc; Siu Man Wong[2], MSc; Bonnie Po Wong[2], BSc; Athena Yin Lam Lee[3], BSc; Frances Kam Yuet Wong[1], PhD

[1]School of Nursing, The Hong Kong Polytechnic University, Kowloon, China (Hong Kong)

[2]Hong Kong Lutheran Social Service, Kowloon, China (Hong Kong)

[3]Department of Health, The Government of the Hong Kong Special Administrative Region, Hong Kong Island, China (Hong Kong)

Corresponding Author:
Arkers Kwan Ching Wong, PhD
School of Nursing
The Hong Kong Polytechnic University
GH 502
Hung Hom
Kowloon
China (Hong Kong)
Phone: 852 34003805
Email: arkers.wong@polyu.edu.hk

Abstract

Background: The use of wearable monitoring devices (WMDs), such as smartwatches, is advancing support and care for community-dwelling older adults across the globe. Despite existing evidence of the importance of WMDs in preventing problems and promoting health, significant concerns remain about the decline in use after a period of time, which warrant an understanding of how older adults experience the devices.

Objective: This study aims to explore and describe the experiences of community-dwelling older adults after receiving our interventional program, which included the use of a smartwatch with support from a community health workers, nurses, and social workers, including the challenges that they experienced while using the device, the perceived benefits, and strategies to promote their sustained use of the device.

Methods: We used a qualitative descriptive approach in this study. Older adults who had taken part in an interventional study involving the use of smartwatches and who were receiving regular health and social support were invited to participate in focus group discussions at the end of the trial. Purposive sampling was used to recruit potential participants. Older adults who agreed to participate were assigned to focus groups based on their community. The focus group discussions were facilitated and moderated by 2 members of the research team. All discussions were recorded and transcribed verbatim. We used the constant comparison analytical approach to analyze the focus group data.

Results: A total of 22 participants assigned to 6 focus groups participated in the study. The experiences of community-dwelling older adults emerged as (1) challenges associated with the use of WMDs, (2) the perceived benefits of using the WMDs, and (3) strategies to promote the use of WMDs. In addition, the findings also demonstrate a hierarchical pattern of health-seeking behaviors by older adults: seeking assistance first from older adult volunteers, then from social workers, and finally from nurses.

Conclusions: Ongoing use of the WMDs is potentially possible, but it is important to ensure the availability of technical support, maintain active professional follow-ups by nurses and social workers, and include older adult volunteers to support other older adults in such programs.

(J Med Internet Res 2024;26:e49403) doi:10.2196/49403

XSL•FO
RenderX

KEYWORDS

community-dwelling older adults; focus group; wearable monitoring devices; mobile phone

Introduction

Background

Technological advancements have facilitated the self-management of chronic diseases among community-dwelling older adults. Wearable monitoring devices (WMDs), such as smartwatches, are among the common technological tools that assist older adults with health monitoring, physical and cognitive training, medication reminders, and fall prevention [1,2]. The literature shows that WMDs are effective at reducing the risk of developing cardiovascular diseases [3], increasing the physical activity levels [4], and improving the quality of life [5] of older adults. However, despite the benefits and high adoption rate of these wearable devices, there is a lack of studies demonstrating the adherence rate of older adults in maintaining consistent use of WMDs [6-8]. A survey with a sample of >4000 Canadian adults revealed that 33% of the participants did not use WMDs to monitor their health on a regular basis [9]. Similarly, another survey conducted in Australia reported an abandonment rate of 29% for WMDs, without specifying the population [10]. Physical disability, a lack of knowledge about the functions of wearable devices, and technological anxiety were summarized as notable reasons for poor adherence to these devices among older adults [11-13].

Self-determination theory has highlighted that long-term behavioral change is determined by one's intrinsic motivation, which is defined as one's action driven by the enjoyment and interest in the activity itself and is affected by 3 factors: competence, autonomy, and relatedness [14]. When an individual has a sense of competence and autonomy in adopting a new behavior and has someone who is socially and psychologically connected (relatedness) to support the behavior, they are more likely to adhere to, and maintain, the behavior over the long term. Recent studies have focused on providing training sessions to help older adults familiarize themselves with the functions of WMDs and enhance their competence and autonomy. However, the results showed no difference in adherence between the participants who received the training sessions and those who did not [15,16]. Older adults have expressed in a qualitative study that 1 preintervention training session is not sufficient to enhance their knowledge of WMDs or resolve their technological anxiety [17]. It was suggested that nursing or peer support, with the simultaneous provision of social support, might be necessary throughout the health program to increase the intrinsic motivation of older adults to adopt WMDs [15]. However, there have been limited studies on the offering of nursing or peer support for older adults in the use of WMDs. In the study by Farivar et al [11], nursing feedback was provided to older adults when their real-time step counts, which were displayed on the WMDs, were unsatisfactory. The program was found to be feasible and acceptable to older adults, but it encountered challenges such as infrequent updates of the WMDs and low engagement and retention rates. Another study, which designed a similar program that provided nursing support to older adults when abnormal vital signs were detected in the WMDs, demonstrated a high dropout rate of 21% and short-term adherence to the WMDs [18]. Recent studies emphasize the importance of implementing a clear nursing service model, such as a case management model, that encompasses problem identification, goal setting, and regular follow-up. This model aims to enhance the intrinsic motivation of older adults to consistently use new technologies, such as mobile health apps and WMDs [7,19], rather than relying solely on providing training sessions to them or intervening only when abnormalities in vital signs are detected through WMDs by the older adults.

Objectives

Because of the perception that they might be causing trouble to others, older adults tended not to actively seek help from health care professionals and peers even when they faced technical problems or when they did not comprehend the medical jargon displayed on the device [11,13]. They were also concerned about their health data being transferred from WMDs to health care professionals but not receiving regular feedback [20]. In view of this, this study was to have a nurse case manager (NCM) work with the older adults to identify factors that could facilitate or hinder their use of a smartwatch and recommended that the older adults use those features of the smartwatch that are linked with their health and social problems, provided suggestions on the duration and frequency of the use of the smartwatch, and provided instructions on how to use these features in their daily routines during the 3-month intervention period. Older adults had the autonomy to adjust or modify their own schedules to ensure that they could use the features of the smartwatch efficiently and effectively. The NCM also encouraged the family members or primary caregivers of the older adults to participate and provide feedback and support. This paper describes the perceptions and experiences of community-dwelling older adults after receiving our interventional program. More specifically, we explored the challenges that they experienced during their use of the WMD, the benefits of using the WMD, and suggestions on how to promote their sustained use of the WMD. The results may provide useful insights for developing a program that can promote the continued adoption of WMDs and, in turn, improve the long-term benefits of the WMDs on health self-management among older adults.

Methods

Study Design

A qualitative descriptive design was adopted for this study [21]. This approach is not associated with any philosophical or theoretical orientation but draws on naturalistic inquiry to understand and describe how people experience a phenomenon [21]. The qualitative descriptive study is the method of choice when straight descriptions of phenomena are desired, which made it appropriate for this study. This study is reported according to the SRQR (Standards for Reporting Qualitative Research) checklist [22].

XSL•FO
RenderX

Setting and Participants

This study was conducted between June 2022 and March 2023 in collaboration with 5 community centers run by a local nongovernmental organization in Hong Kong. Using a purposive sampling approach, the members of these community centers who were interested in this program were screened and recruited into this study if they (1) were aged ≥60 years [23], (2) owned a smartphone, (3) were able to communicate in Cantonese or Mandarin, and (4) had internet access. They were excluded if they (1) had been diagnosed with cognitive impairment, (2) were bedbound, (3) owned a smartwatch, and (4) were involved in other studies using WMDs.

Recruitment Process

Staff working at the collaborating community centers invited their members to join the program using Facebook Live. Those who were interested provided their name to the staff, and trained research assistants screened them via telephone. Eligible members were invited to meet the research assistants at the community centers to receive an explanation of the details of the study and to give their written consent to take part in it. All participants received a health monitoring package that included a smartwatch with an alarm setting, a prepaid SIM card, a blood pressure monitor, and a pulse oximeter.

Interventional Program

Before the program, a 1-hour web-based training session and a practical test were delivered to all participants to explain the basic operation of the WMD. The number of a telephone manned during office hours by staff of the community center was provided to participants to call should they face any technical problems during use.

The participants were provided with a package that included a WMD (ProVista Care smartwatch), a prepaid SIM card, a blood pressure monitor, and a pulse oximeter. ProVista Care was selected as the WMD for this study due to its validated performance, affordability, and comparable functionality to other similar devices. These functions encompass fall detection; location and activity tracking; blood pressure, pulse, and oxygen saturation monitoring; medication and appointment reminders; and calls to preset numbers and SOS calls. This selection enhances the applicability of the study's findings to real-world implementation. Data collected from ProVista Care can be synchronized and transferred to the server via the ProVista Care app installed on participants' personal smartphones. The WMD was designed to be worn on the wrist, securely fastened with an elastic band. Participants were instructed to wear the WMD as frequently and for as long as possible throughout the study duration.

Ten trained community health workers (CHWs), NCMs, and social workers were the interventionists in this 3-month program. The participants in the intervention group received a home visit by a CHW and the NCM in the first month and biweekly telephone calls by the CHW from the 3rd to the 12th weeks. In the first home visit, using the Omaha System, the NCM explored the features of the smartwatch that each participant might find beneficial. The Omaha System is a comprehensive assessment-intervention-evaluation instrument for community-based practice [24]. There were 21 health and social problems listed in the Omaha System that were relevant to the features of the smartwatch used in this study; for example, the feature of fall detection in the smartwatch might help participants with musculoskeletal problems or lower limb weakness. The NCM empowered the participants to set goals and action plans in the first meeting, while the CHWs followed up, recalling the goals and action plans with the participants and, in subsequent telephone calls, motivating the participants to regularly use the smartwatch. The NCM also monitored the vital signs of the participants that were automatically uploaded by the smartwatches at the backend. When abnormal vital signs were detected by the smartwatch, the NCM would call the participants via telephone and provide appropriate interventions, such as a referral to a social worker, based on the validated protocols.

Data Collection

A total of 6 in-depth focus group discussions with 22 participants were conducted at the end of this program. In-depth focus group interviews are conducted to evaluate participants' experiences after an intervention through group interactions [25]. For studies using focus group discussions, it has been suggested that groups ranging from 2 to 40 may be adequate to attain data saturation depending on the phenomenon under investigation [25]. Thus, 6 groups were considered adequate for this study to attain data saturation. All discussions were conducted with a guide developed by the research team. The focus group interviews were conducted in Cantonese and each session lasted for 25 to 65 minutes. All interviews were audio recorded with the consent of the participants. Interview transcripts were written up by members of the research team. To ensure the consistency of coding and interpretation of data, an audit trail was conducted, and all discrepancies were resolved through discussion and consensus.

Data Analysis

All data collected from the focus group discussions were analyzed inductively using the approach to constant comparison analysis formulated by Maykut and Morehouse [26], who proposed a 4-step approach to the constant comparison of focus group data: inductive categorization, refinement of categories, exploring relationships across the categories, and integration of data [26]. In inductive categorization, AKCW and JB read and reread the interview transcripts in both English (JB) and Cantonese (AKCW) to identify recurring concepts independently. Next, overlapping concepts across the groups were categorized and combined by the 2 independent coders (AKCW and JB) to formulate provisional codes. In the second stage, that is, refining the categories of codes, the provisional list of codes was concurrently examined alongside a review of the interview transcripts. The process of categorization was undertaken through discussion with the wider research group to attain consensus. Subsequently, similar codes were grouped to formulate categories for each group. The emerging categories were then concurrently compared across the groups, with recurring categories further refined and grouped. For the third stage, we further refined the categories by grouping them under a distinct umbrella. Categories with common elements were

XSL•FO
RenderX

grouped to make broader groups or emerging themes. With these themes, we worked through each group and the associated categories to attain a complete understanding and create patterns of meaning from the data.

Methodological Rigor and Trustworthiness

The trustworthiness of this study was evaluated according to four criteria: (1) credibility, (2) dependability, (3) confirmability, and (4) transferability [27]. To enhance credibility and dependability, the summarized results were sent to those participants who had agreed to check them for further clarification and to give feedback on the researchers' interpretation. Audit trails and peer debriefings were also conducted during the analysis of data to ensure the consistency of the interpreted data to achieve confirmability. A thick description was ensured in reporting the study to enhance transferability. To attain analytical rigor, we ensured that analyses were undertaken in both Cantonese and English and compared to ensure consistency. The authors responsible for this section were fluent in Cantonese and English. The iterative approach to analysis also ensured consistent coding, with an audit trail on the decisions on coding and categorization. In addition, the constant comparison approach ensured that our focus was not only on individual-level analyses but also on analyses within and across the groups.

Ethical Considerations

This study was conducted under the principles of the Declaration of Helsinki and approved by the ethics committee of the Hong Kong Polytechnic University (HSEARS20220429001). All eligible participants were given the right to refuse participation and the right to withdraw from the interview at any time. Written informed consent was collected from all participants. To protect the participants' privacy, all data collected from this program were kept confidential and anonymized and were only accessible to the members of the research team.

Results

Sociodemographic Data

A total of 22 community-dwelling older adults were involved in 6 focus group discussions. Of these, 5 (23%) were male, and 17 (77%) were female, with ages ranging from 62 to 78 years. Only 1 (5%) participant had experience in using a smartwatch before inclusion in the study. A total of 17 (77%) had a primary level of education, and 5 (23%) had a secondary level of education or higher. The clinical characteristics of the participants have been reported in a previous study [19].

Emerging Themes and Categories

Overview

Three themes and 7 categories emerged from the focus group data, as shown in Textbox 1.

Textbox 1. Emerging themes and categories.

Challenges associated with the use of the wearable monitoring device (WMD)

- Individual-related challenges
- System-related challenges

Perceived benefits of using the WMD

- Self-monitoring and health promotion
- Convenience

Strategies to promote use of the WMD

- Availability of technical support
- Ongoing follow-up professional support
- Peer and family support

Emerging Theme 1: Challenges Associated With the Use of the WMD

This theme describes challenges and concerns that affected the participants' use of the smartwatch. The emerging categories were (1) individual-related challenges and (2) system-related or technical challenges.

Individual-Related Challenges

Participants across all groups emphasized that they were slow in learning to use the WMD and required a great deal of face-to-face instruction to be fully oriented to the device and its functionalities before being able to use it effectively. This issue particularly resonated with those who were using such a device for the first time. Some of the older adults either could not understand the instructions or needed more time to assimilate them. It took weeks to months for the older adults to become familiar with the device:

> *This is my first time wearing a smartwatch. When you wear the watch for the first time, you will definitely not know how to use all the functions. So, I wanted to ask everyone if they have experienced this situation before.* [Participant 1]

> *Well...at first, it was difficult to use. But after using it for a while, it was basically okay, and we can use it on our own.... Hmm yes, actually, if someone*

XSL•FO
RenderX

> *teaches you face-to-face, you can learn it clearly first.* [Participant 15]
>
> *How much time? I think three to four months to learn to use it well.* [Participant 20]
>
> *For me, at first, it took a long time to use it. Sometimes, I just could not get it to work. But after a while, it became much better. For example, measuring blood oxygen and blood pressure readings became much easier for me with time.* [Participant 4]
>
> *The first time I tried, I did not know how to turn on the device or turn it off. It was difficult at first.* [Participant 7]

The participants also highlighted the issue of forgetfulness, which affected how well they used the device. They noted that with their increasing age, forgetfulness was a common occurrence. Some of the participants mentioned forgetting how to operate the device and the functions available during the initial period of use, although over time they were able to become better at using the device consistently:

> *I'm so dumb sometimes that I forget what I am doing. I sometimes cannot even figure out how to wrap a scarf around my head, not to mention how to use the watch.* [Participant 13]
>
> *I am already in my late years. If you even ask me what I ate yesterday, I cannot remember.* [Participant 2]

System-Related or Technical Challenges

System-related or technical challenges were encountered across all groups. The size of the WMD was considered an issue. Participants described the WMD as big, which made it difficult to wear regularly. Occasionally, the size of the watch was considered a source of ridicule. Despite the potential for being ridiculed, some of the older adults noted that they were more concerned about the functionality and capacity of the watch than its size. In addition, the smooth, glass surface of the watch's touchscreen became slippery and unresponsive when used by the older adults in cold weather, creating usability issues:

> *The watch can measure blood pressure and blood oxygen levels. Your watch looks much better and looks great. Our watches are big, like big turtles, and sometimes people make fun of it.* [Participant 9]
>
> *You can see that your watch is smaller than ours. Our watch is bigger, and it obstructs a lot. But even if it's still bulky and unattractive, I think we can still wear it because it will help us.* [Participant 20]
>
> *It feels really troublesome to use during cold weather. There is no problem in hot weather.* [Participant 7]

The power capacity of the WMD presented a significant challenge for the participants. Participants wanted to use the WMD, but the need for frequent charging made it rather inconvenient to do so. In some cases when they wanted to use it when going out, they noticed that the WMD was low on battery. Coupled with the previously mentioned issue of forgetfulness, this meant that they missed the opportunity to charge it before going out. The participants also reported that the need to frequently charge the WMD prevented them from using it for a longer period throughout the day. This issue deterred some of the older adults from using the WMD altogether on some occasions:

> *Hmmm, if you know everything, the main problem with the watch is the frequent charging. That battery needs to be charged frequently—like every day. If you don't charge it, it will just run down fast. Yes, it is so fast and when there is no electricity, things will become difficult. The need to charge is too frequent for us.* [Participant 5]
>
> *Oh, so you realize that the battery is down when you wear it and then you must put it back to charge for a while. Yes, that's right. It is very troublesome to do this every day before going out. The battery runs out quickly all the time.* [Participant 9]

Another technical issue that was identified was the fact that some of the participants felt that the WMD had several functions they did not know how to use. Interestingly, other older adults still struggled to navigate through even the few functions that they were taught to use, and they occasionally experienced digital fatigue after constant use:

> *There may be some functions we cannot use. The watch seems to have many functions, but we do not know them all and also don't know how to use them all.* [Participant 16]
>
> *But, I realized there are so many functions on that watch that we cannot use them all. Also, some functions that were possible to use before, people found it annoying to continue to use them. That is why we do not use them frequently anymore.* [Participant 9]

All participants were enthusiastic about the ability of the device to count their steps as they walked about. However, the older adults mentioned that the device gave them incorrect step counts. In 1 group, the participants mentioned that the step count function also did not display correctly. Occasionally, they used their mobile phones to obtain correct step counts. In addition to this challenge, some of the participants reported occasional challenges with uploading or transmitting data on their vital parameters:

> *The pedometer was malfunctioning and gave incorrect figures. When you count how many steps you take yourself and then check the watch, it doesn't match at all. The watch and the phone both have incorrect counts all the time.* [Participant 8]
>
> *It shows only a few steps, even though I walked quite a lot. Yes, our watches cannot measure many steps. My phone shows 10,000 steps, but my watch shows 2000 or 3000 steps. To be honest, the watch is not accurate when it comes to the step count. The step count displayed on my phone is not the same as the one on my phone. Yes, that is how it is. There are some differences, yes.* [Participant 1]
>
> *Actually the step count is important, but it is not accurate at all. I often check it myself. Usually, I check how many steps I have taken, especially since*

XSL•FO
RenderX

> *I sit in an office for most of the day. But it is not correct when I check the watch and the phone.* [Participant 16]

> *I tell him about my blood pressure on that day. I tell him about my blood pressure and how many steps I took that day. Sometimes, the watch cannot display the values correctly.* [Participant 9]

Some of the participants also found the device to be extremely sensitive, which occasionally caused discomfort because the alarm went off immediately when it sensed a slight movement:

> *But the watch is too sensitive. Sometimes when I move my hands or feet, it shakes and triggers the alarm. And then it keeps telling me how long it has been and what to do.* [Participant 5]

Emerging Theme 2: Perceived Benefits of Using the WMD

Regardless of the notable challenges, participants highlighted the benefits of using the WMD. These were (1) self-monitoring and health promotion and (2) convenience.

Self-Monitoring and Health Promotion

Participants across all groups stated that the WMDs offered them an opportunity to self-monitor some vital parameters, such as blood pressure and oxygen saturation levels. The older adults found this feature to be particularly helpful because it helped them to record their parameters, track them, and share them with health care professionals and to ascertain whether they were maintaining a good health status. Indeed, the use of the device boosted the confidence of older adults across all groups in their ability to actively participate in self-management, particularly because the NCM actively followed up to enable them to attain their health goals:

> *Um, measure blood pressure and blood oxygen levels at the same time. Well, we know now. We know our blood pressure and blood oxygen levels. It helps us to maintain our health by making us aware of the condition of our body and whether it is normal or not.... At night, I have a blood pressure machine and I can measure my blood pressure every night.* [Participant 2]

> *Also, it gives a different perspective on managing your health with more information available to you. For instance, I know how much I walked today.* [Participant 8]

> *Yes, definitely. Using the watch gave me a lot of confidence. I wear it at home and when I go out. The nurse also reminded me to walk a certain amount of time every day, and even though I forget, I still try my best to walk more. The most important thing is to try and walk more.* [Participant 4]

> *And at home, I don't know how high or low the blood sugar is. If I know, I can control it by myself at home. If it is high, I will eat less. It is good to be clear about the blood sugar. For the nurses, it would be helpful if they could find my place and remind me of something regularly.* [Participant 14]

> *The best function would be to be able to monitor your health and detect any potential illnesses.* [Participant 19]

The participants expressed a desire for more regular follow-ups by the nurses and an option to monitor their blood glucose levels in addition to blood pressure and blood oxygen levels:

> *I just think it would have been helpful if the device can also help you to monitor your blood sugar levels just like it helps to monitor blood pressure and blood oxygen levels.* [Participant 5]

> *Oh, she sometimes follows up on us with home visits and phone interviews. Yes, but what about the rest of the time? If the nurse does not contact you, you won't actively look for her, right? Besides, the nurse does not come to the center every day. The nurse is also busy with her work, so where would she have the time? So, in some instances, if you are not feeling better, you go and see a doctor.* [Participant 13]

Although the step count feature of the WMD was described as inaccurate, the participants felt that it was still helpful to know how many steps they had taken because that motivated them to go out more often rather than stay at home. Being able to compare their step counts with others gave them a sense of accomplishment, especially if they found them to be higher than those of their colleagues:

> *But I don't really care so much about how many steps I take in a day. However, it can still calculate something for you. For example, if the doctor tells you how many steps you need to take in a day, the watch can help you to keep track of it. Maybe we don't really need it because our phones can also count the steps.* [Participant 2]

> *I take so many steps every day. Many people can vouch for me. I am the best here; I take so many steps. After finishing my chores at home, I come down and do some healthy dancing, and walk around the center. According to them, I am the best.* [Participant 14]

> *Because you can show off to others, like the person you are exercising with, and say, look I have burned this much fat, right?* [Participant 9]

Participants also mentioned that the device helped them to not only record their vital parameters but also to view these records regularly. Regarding the promotion of health, the participants noted that the device helped them to participate in regular exercises and to build the confidence they need to meet health-related goals:

> *So, wearing a watch can make you want to do more exercise, right? Because when you wear a watch, you want to see how many steps you have taken, which makes you want to move more.* [Participant 6]

Convenience

For participants across the groups in this study, the WMD offered a sense of convenience in being able to monitor their vital parameters, record the values, track them, and share them with the nurse if required. The notion of convenience was also

XSL•FO
RenderX

noted to be related to the ease with which the older adults navigated the device to inform their self-management strategies. In addition, that they did not have to be in a hospital to assess these basic vital parameters was something the older adults considered very convenient. In fact, they could monitor the basic vital parameters from the comfort of their home and even when moving about in the community:

> *With the watch, there is a guide, and I am afraid to be lazy about moving around and not walking around. But when I think about the watch, I have the confidence to do it. In the past, I just sat at home all day, talking on the mobile phone about how many thousands of steps I have to walk, and now I just go out and do it myself.* [Participant 10]

In addition, the aspect of being able to reach out and interact with a nurse or having a nurse follow up on an older adult whose parameters were outside the normal range was considered to be convenient. This may be related to the fact that the participants felt that they were not only using the device but also being professionally supported by a nurse:

> *Yes, if the nurse thinks the blood oxygen is low, she will remind you to do it before and again. Then if something happens to you, you will know to see a doctor.* [Participant 5]

> *At least, the blood pressure can be seen by the nurse. And the blood oxygen levels can be seen with a press of the finger. However, the step count is not accurate.* [Participant 15]

The social aspect of the watch, such as being able to take pictures and share these with families and friends, was considered helpful and made life more convenient for the older adults. In other words, it added a bit of fun to using the device:

> *I discovered a new function or new feature. It is completely possible to use the watch to take a photo and share. Yes, so it is so much more convenient.* [Participant 3]

> *It is best if there is nothing wrong with it. The best thing to do is to take a photo of that watch and the stick together after we finish the test, and it will be the most accurate. It is comfortable and makes life more convenient, I think.* [Participant 5]

Another source of convenience was noted to be related to the fact that the wearable devices afforded older adults or their families a unique opportunity to track their whereabouts. The older adults found this feature particularly helpful because they considered themselves to be forgetful on occasion, and this feature helped them to retrace their steps to their original location or helped others to know where they were:

> *The best feature of the device is the tracking. Some people have a poor memory, or they may not be able to find their way home. In that case, their family members can locate them using the tracking feature.* [Participant 21]

> *Mr Choi once tracked us. I got lost and could not find my way home. I got scared and started sweating. Mr Choi tracked my watch and found me at the Che Kung Temple.* [Participant 2]

> *This is where technology has advanced. The most useful thing is when someone is lost. If he wears the watch, you can find him and track where he has been. Then you can find him using the tracking function.* [Participant 14]

> *Sometimes when I go somewhere far, I don't know where I am, and I cannot see clearly due to my glaucoma. One time, I had to go to the other side for the lunar new year, but I took the wrong bus and did not know where I was. Luckily, I was able to use the watch to track my way.* [Participant 6]

Emerging Theme 3: Strategies to Promote Use of the WMD

This emerging theme discusses approaches observed in the data that highlight strategies to sustain continual use of the device. The following categories were captured: (1) availability of technical support, (2) ongoing follow-up professional support, and (3) peer and family support.

Availability of Technical Support

The plethora of technical issues emerging from the use of the device warrants ongoing availability of technical support. This great need was mentioned by participants in all groups and was particularly felt when the device developed a fault or broke down, or the participants needed more assistance in navigating through the functions:

> *The watch broke down and we did not know how to fix it. Someone at the center said he knew how to turn it back on. We tried for a while, but it still did not work. So, I said forget it. I did not wear it. I only wore it for ten days before, and just for measuring blood pressure at home.* [Participant 14]

Although some of the participants sought assistance from the social workers, most older adults hesitated to disturb the personnel and therefore avoided seeking assistance altogether, regardless of the technical challenges that they were facing:

> *So, it is changed. Actually, you also changed and regarded it as a planned situation, and I did not dare to worry the nurse or the supervisor. So, if there is a problem with the watch, I must handle it on my own.* [Participant 10]

In addition, the participants mentioned that they needed more technical support to access other functions on the watch because they found it difficult to perform this task:

> *And I don't understand why so many functions need to be locked, except the panic button. I wondered if there was help for us to unlock these functions on the watch.* [Participant 7]

> *We tried to figure it out but could not do it and we needed lots of help. In the end, it suddenly made a sound, and we could not figure out how long it had been, it just happened.* [Participant 16]

XSL•FO
RenderX

> *There are too many things to handle. If you suddenly introduce ten functions for us to use, how can we remember them? You are not teaching a class, you won't be able to remember them either.* [Participant 11]

Ongoing Follow-Up Professional Support

Although the device was helpful in various ways, the older adults still preferred to have nurses follow up actively with them. For the participants across all groups, this form of support was generally limited, and they wished that they had interacted more with the nurses to be able to interpret the values that they obtained and to seek more health-related information. Perhaps the nurse support centered on following up on older adults with abnormal readings. Thus, those who maintained readings within normal ranges had limited contact with the nurses. The participants also felt that the limited support that they received from the nurses might have affected how well they met their health-related goals:

> *They [the nurses] do a good job when they call or visit you. With the watch, you set a goal with the nurse, which motivates you to do more. But they are not always there. It is helpful if they can find my place so that they can remind me of something I don't know.* [Participant 19]

Aside from ongoing professional support from nurses to keep the participants motivated in meeting their health-related goals, support from social workers is equally important to promote their continued use of the devices. Social workers played critical roles in promoting the use of the device by offering troubleshooting support, helping the participants to navigate through the device, and offering encouragement. In fact, it seemed as though the older adults who participated in the study trusted the social workers more than they did the nurses and were willing to always seek assistance from them. The older adults seem to have built a strong relationship with the social workers, which made it easier for them to seek assistance from them if required:

> *They do help us a lot and encourage us. Whenever there is a problem, we always look for him to help us out. He is the most reliable. He is very responsible, and he is always willing to help us.* [Participant 5]

> *I did not even know how to turn off my phone. He said to turn off my phone, do it this way. He really taught us a lot of things.* [Participant 10]

Peer and Family Support

Peer support from the CHWs also emerged as a critical factor to sustain the ongoing use of the WMD. These older adult volunteers or older adult ambassadors often offered encouragement to the participants to continue to use the device, record their values, and work toward meeting their health-related goals. Participants across the groups highlighted that it was occasionally difficult to gain access to a nurse; thus, the older adult volunteers or older adult ambassadors became the first point of call for assistance before reaching out to the social workers:

> *It is not so easy to find or see a nurse on some days. The volunteers have done this before, so we can reach out to them. There are days when you will forget to write the values, and they will remind you to do so.* [Participant 16]

Aside from peer support, family support was also observed to be helpful in encouraging the older adults to use the WMD as required. Thus, older adults who resided alone with limited or no family support found it difficult to monitor and continually use the device to promote health:

> *They said that I fall frequently and have fallen several times before. I must be careful now that I am getting older. If anything happens to me, it would be troublesome because I live alone.* [FG2]

Discussion

Principal Findings

Emerging technologies such as wearable devices are advancing care and support for older adults in communities across the globe. Despite the plethora of literature highlighting the importance of wearable devices, significant concerns remain about the decline in use after a period of time. The world's aging population is booming, but only a limited amount of work has been done to unearth the experiences of older adults regarding the use of wearable devices. This critical gap informed this study, which was part of a large trial program. The findings bring to the fore the challenges experienced by older adults regarding the use of wearable devices, which were identified as individual- and system-related challenges. The findings of the study further highlight the perceived benefits of the devices, particularly in the areas of self-monitoring, health promotion, and convenience. In addition, the study identified a hierarchical pattern of health-seeking behaviors of older adults when using the devices. Put together, the findings highlight that ongoing use of the devices is possible, although there is a critical need to ensure the availability of technical support and ongoing active professional follow-up by the health care team (notably nurses and social workers) and to include older adult volunteers to support other older adults in such programs.

Previous studies have uncovered various technical issues associated with using wearable devices. In a recent study, the authors identified interoperability, battery issues, the bulky nature of the device, a lack of personalization, and a lack of support as key issues that affected the use of the device [28]. Similar to this finding, our study also noted similar technical issues that affected the use of the devices. By contrast, however, it was noted in our study that regardless of these issues, older adults were willing to continue using the device because they believed that doing so was to their benefit. Despite this, we observed that issues related to individuals can also affect the use of wearable devices among older adults. For most of the older adults included in this study, this was the first time they had to use a wearable device, and they needed more time to become acquainted with it. Although issues such as forgetfulness may be considered part of the aging process, these findings suggest that aside from intensive training on how to use the device, there is still a need for ongoing technical support to

XSL•FO
RenderX

boost its use. In addition, instruction manuals need to be more user-friendly and easily comprehensible for older adults. Comprehensibility is essential; we observed in this study that the user manuals were unclear, which may have had an impact on how well the participants made use of the WMD.

The inclusion of professional and peer support in this study is particularly noteworthy. An existing study showed that it might be necessary to provide nursing or peer support throughout the duration of the health program to increase the intrinsic motivation of older adults to adapt to WMDs and to provide social support at the same time [17]. In our study, which combined both professional and peer support, we observed that older adults did not want to disturb the nurses. Rather, they felt more comfortable consulting the older adult volunteers first, before reaching out to the social workers and, last of all, to the nurses if necessary. This hierarchical pattern of health-seeking behaviors may be an indication that older adults viewed the older adult volunteers as peers sharing similar experiences and conditions, which made it easier to relate to them than to the professionals. Nurses were perceived by older adults to be busy professionals. Thus, the participants would rather resort to seeking support from social workers, although they wished they had more interactions with the nurses. Put together, the findings suggest that nurses may need to take an active role in reaching out to older adults and being available when needed, regardless of whether they are using the wearable device. The concept of peer support also needs to be promoted further by engaging other older adults as volunteers to support older adults who are transitioning to using wearable devices. A recent study has shown that peer-to-peer support for community-dwelling older adults has the potential to not only promote adherence to therapeutic regimens but also to improve their quality of life, which warrants further exploration [29].

Furthermore, we observed that the ongoing availability of technical support and family support is also essential to promoting the use of wearable devices. It is possible that technical support may be available but unknown to the older adults or that they may not want to disturb others. Thus, older adults need to be encouraged to seek help if needed and should know where to obtain this help. Regarding family support, it remains a major cornerstone of support for older adults [30]. The absence of this form of critical support may lead to loneliness, which can exacerbate health issues and interfere with therapeutic regimens, including the use of wearable devices [31]. Undoubtedly, expanding on the notion of family support is beyond the scope of this study, but it is possible to recommend that older adults with limited or no family support need to be identified and appropriate strategies devised to assist them.

Moreover, we identified both individual- and system-related issues that can adversely impact the use of WMDs. Individual-related factors such as slow learning patterns and forgetfulness were highlighted by the participants as impacting how they initially navigated the WMD. Undoubtedly, aging is not a disease, although it can be associated with forgetfulness, which can impact activities of daily living [32]. Forgetfulness coupled with the slow learning patterns emphasize the need for continuous, gradual education to enable older adults to use WMDs effectively [33]. System-related challenges such as the size of the WMD and its limited power capacity are concerns that need to be addressed in subsequent design studies. In addition, concerns regarding the WMD generating incorrect readings also emerged as a system-related challenge. Previous studies have reported that a common problem with wearable devices is the likelihood of experiencing automatic loss of synchronization, making it difficult or impossible to update data or resulting in an incorrect report [34,35]. Although loss of synchronization was not examined in this study, it may have potentially contributed to the incorrect readings observed by the older adults in this study.

Strengths and Limitations

The strength of this study lies in the use of a rigorous approach to collecting and analyzing data with a focus on individual, within-group, and across-group variations to attain a thick description of what it means to experience the use of a wearable device. This strength notwithstanding, some limitations are noteworthy. First, the experiences of the participants are related to the use of a particular wearable device with distinct features. Thus, the findings may not necessarily be transferable to all wearable devices, although they may offer a useful resource on how older adults are likely to experience using the devices. Second, the study was undertaken in a region with distinct sociocultural features. The findings should therefore be interpreted taking these unique features into consideration. In addition, the nature of the program was such that the older adults were required to have some technological abilities. Thus, the findings may not be transferable to older adults who find it difficult to use technological applications.

Conclusions

Emerging technologies, such as wearable devices, for supporting community-dwelling older adults warrant more work on how users are experiencing these devices. The findings from this study bring to the fore the barriers and benefits of wearable devices and offer insight into strategies that can be considered to improve use. Because of issues that might emerge, it may be helpful to consider the availability of ongoing technical support, professional follow-up support, peer support, and family support. In fact, a personalized approach is needed to promote the use of wearable devices among older adults.

Acknowledgments

The authors would like to thank Hong Kong Lutheran Social Service for providing the smartwatches and participating in and contributing to this study. The study was funded by the Departmental General Research Fund, The Hong Kong Polytechnic University (G-UAQ2).

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AKCW and FKYW conceptualized the interventional program. AKCW, JB, JJS, FKYW, KKSC, BPW, SMW, and AYLL provided intellectual input on the study design, methodology, and evaluation. AKCW and JB drafted the manuscript. AKCW analyzed the data. All authors contributed to, reviewed, and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 09;8(11):e18907 [FREE Full text] [doi: 10.2196/18907] [Medline: 33164904]
2. Teixeira E, Fonseca H, Diniz-Sousa F, Veras L, Boppre G, Oliveira J, et al. Wearable devices for physical activity and healthcare monitoring in elderly people: a critical review. Geriatrics (Basel) 2021 Apr 07;6(2):38 [FREE Full text] [doi: 10.3390/geriatrics6020038] [Medline: 33917104]
3. Roberts LM, Jaeger BC, Baptista LC, Harper SA, Gardner AK, Jackson EA, et al. Wearable technology to reduce sedentary behavior and CVD risk in older adults: a pilot randomized clinical trial. Clin Interv Aging 2019;14:1817-1828 [FREE Full text] [doi: 10.2147/CIA.S222655] [Medline: 31695350]
4. Brickwood KJ, Ahuja KD, Watson G, O'Brien JA, Williams AD. Effects of activity tracker use with health professional support or telephone counseling on maintenance of physical activity and health outcomes in older adults: randomized controlled trial. JMIR Mhealth Uhealth 2021 Jan 05;9(1):e18686 [FREE Full text] [doi: 10.2196/18686] [Medline: 33399541]
5. Jang IY, Kim HR, Lee E, Jung HW, Park H, Cheon SH, et al. Impact of a wearable device-based walking programs in rural older adults on physical activity and health outcomes: cohort study. JMIR Mhealth Uhealth 2018 Nov 21;6(11):e11335 [FREE Full text] [doi: 10.2196/11335] [Medline: 30463838]
6. Chang RC, Lu HP, Yang P, Luarn P. Reciprocal reinforcement between wearable activity trackers and social network services in influencing physical activity behaviors. JMIR Mhealth Uhealth 2016 Jul 05;4(3):e84 [FREE Full text] [doi: 10.2196/mhealth.5637] [Medline: 27380798]
7. Wong AK, Bayuo J, Wong FK, Chow KK, Wong SM, Lau AC. The synergistic effect of nurse proactive phone calls with an mHealth app program on sustaining app usage: 3-arm randomized controlled trial. J Med Internet Res 2023 May 01;25:e43678 [FREE Full text] [doi: 10.2196/43678] [Medline: 37126378]
8. Shin G, Jarrahi MH, Fei Y, Karami A, Gafinowitz N, Byun A, et al. Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. J Biomed Inform 2019 May;93:103153 [FREE Full text] [doi: 10.1016/j.jbi.2019.103153] [Medline: 30910623]
9. Paré G, Leaver C, Bourget C. Diffusion of the digital health self-tracking movement in Canada: results of a national survey. J Med Internet Res 2018 May 02;20(5):e177 [FREE Full text] [doi: 10.2196/jmir.9388] [Medline: 29720359]
10. Gartner survey shows wearable devices need to be more useful. Gartner. 2016 Dec 7. URL: https://www.gartner.com/en/newsroom/press-releases/2016-12-07-gartner-survey-shows-wearable-devices-need-to-be-more-useful [accessed 2024-07-10]
11. Farivar S, Abouzahra M, Ghasemaghaei M. Wearable device adoption among older adults: a mixed-methods study. Int J Inf Manage 2020 Dec;55:102209 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102209] [Medline: 32834339]
12. Kononova A, Li L, Kamp K, Bowen M, Rikard RV, Cotten S, et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e9832 [FREE Full text] [doi: 10.2196/mhealth.9832] [Medline: 30950807]
13. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
14. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
15. Kamei T, Kanamori T, Yamamoto Y, Edirippulige S. The use of wearable devices in chronic disease management to enhance adherence and improve telehealth outcomes: a systematic review and meta-analysis. J Telemed Telecare 2022 Jun;28(5):342-359. [doi: 10.1177/1357633X20937573] [Medline: 32819184]
16. Makai P, Perry M, Robben SH, Schers HJ, Heinen MM, Olde Rikkert MG, et al. Evaluation of an eHealth intervention in chronic care for frail older people: why adherence is the first target. J Med Internet Res 2014 Jun 23;16(6):e156 [FREE Full text] [doi: 10.2196/jmir.3057] [Medline: 24966146]
17. Fjellså HM, Husebø AM, Storm M. eHealth in care coordination for older adults living at home: scoping review. J Med Internet Res 2022 Oct 18;24(10):e39584 [FREE Full text] [doi: 10.2196/39584] [Medline: 36256831]

XSL•FO
RenderX

18. Boyne JJ, Vrijhoef HJ, Spreeuwenberg M, De Weerd G, Kragten J, Gorgels AP. Effects of tailored telemonitoring on heart failure patients' knowledge, self-care, self-efficacy and adherence: a randomized controlled trial. Eur J Cardiovasc Nurs 2014 Jun;13(3):243-252 [FREE Full text] [doi: 10.1177/1474515113487464] [Medline: 23630403]
19. Wong AK, Wong FK, Chow KK, Wong SM, Bayuo J, Ho AK. Effect of a mobile health application with nurse support on quality of life among community-dwelling older adults in Hong Kong: a randomized clinical trial. JAMA Netw Open 2022 Nov 01;5(11):e2241137 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.41137] [Medline: 36350651]
20. Mercer K, Giangregorio L, Schneider E, Chilana P, Li M, Grindrod K. Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation. JMIR Mhealth Uhealth 2016 Jan 27;4(1):e7 [FREE Full text] [doi: 10.2196/mhealth.4225] [Medline: 26818775]
21. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
22. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
23. Ageing and health. World Health Organization. 2022 Oct 1. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2023-12-18]
24. Martin KS. The Omaha System: A Key to Practice, Documentation, and Information Management. Philadelphia, PA: Elsevier Saunders; 2005.
25. Morgan DL. Focus Groups as Qualitative Research. Thousand Oaks, CA: SAGE Publications, Inc; 1988.
26. Maykut P, Morehouse R. Beginning Qualitative Research: A Philosophical and Practical Guide. Milton Park, UK: Taylor & Francis; 1994.
27. Lincoln YS, Guba EG. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. New Direct Progr Eval 2004 Nov 04;1986(30):73-84. [doi: 10.1002/ev.1427]
28. Olmedo-Aguirre JO, Reyes-Campos J, Alor-Hernández G, Machorro-Cano I, Rodríguez-Mazahua L, Sánchez-Cervantes JL. Remote healthcare for elderly people using wearables: a review. Biosensors (Basel) 2022 Jan 27;12(2):73 [FREE Full text] [doi: 10.3390/bios12020073] [Medline: 35200334]
29. Thombs BD, Carboni-Jiménez A. Peer-to-peer support for older adults-what do we know and where do we go? JAMA Netw Open 2021 Jun 01;4(6):e2113941 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.13941] [Medline: 34129026]
30. Wong AK, Ng NP, Hui VC, Montayre J. Effect of a telecare-based intervention on stress levels in informal caregivers of older adults: protocol for a randomized controlled trial. Front Psychiatry 2023 May 25;14:1167479 [FREE Full text] [doi: 10.3389/fpsyt.2023.1167479] [Medline: 37377468]
31. Doblas JL, Palomares-Linares I, Martínez MS. Chapter 8: Loneliness in older adults: a comparative study of four southern European countries. In: Entrena-Durán F, Soriano-Miras RM, Duque-Calvache R, editors. Social Problems in Southern Europe: A Comparative Assessment. Cheltenham, UK: Edward Elgar Publishing; Aug 7, 2020:87-102.
32. Ballard J. Forgetfulness and older adults: concept analysis. J Adv Nurs 2010 Jun 29;66(6):1409-1419. [doi: 10.1111/j.1365-2648.2010.05279.x] [Medline: 20384638]
33. Imbriano G, Beaudreau SA. Subjective cognitive difficulties may communicate more than forgetfulness. Int Psychogeriatr 2023 Mar 24:1-8. [doi: 10.1017/s1041610223000285]
34. Kaewkannate K, Kim S. A comparison of wearable fitness devices. BMC Public Health 2016 May 24;16(1):433 [FREE Full text] [doi: 10.1186/s12889-016-3059-0] [Medline: 27220855]
35. Cho S, Ensari I, Weng C, Kahn MG, Natarajan K. Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth 2021 Mar 19;9(3):e20738 [FREE Full text] [doi: 10.2196/20738] [Medline: 33739294]

Abbreviations

CHW: community health worker
NCM: nurse case manager
SRQR: Standards for Reporting Qualitative Research
WMD: wearable monitoring device

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 27.05.23; peer-reviewed by M Keivani, I Madujibeya, A AL-Asadi; comments to author 06.12.23; revised version received 14.01.24; accepted 24.05.24; published 07.08.24.

Please cite as:
Wong AKC, Bayuo J, Su JJ, Chow KKS, Wong SM, Wong BP, Lee AYL, Wong FKY
Exploring the Experiences of Community-Dwelling Older Adults on Using Wearable Monitoring Devices With Regular Support From Community Health Workers, Nurses, and Social Workers: Qualitative Descriptive Study
J Med Internet Res 2024;26:e49403
URL: https://www.jmir.org/2024/1/e49403
doi:10.2196/49403
PMID:39110493

©Arkers Kwan Ching Wong, Jonathan Bayuo, Jing Jing Su, Karen Kit Sum Chow, Siu Man Wong, Bonnie Po Wong, Athena Yin Lam Lee, Frances Kam Yuet Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Value of Engagement in Digital Health Technology Research: Evidence Across 6 Unique Cohort Studies

Sarah M Goodday[1,2], MSc, PhD; Emma Karlin[1], MSc; Alexa Brooks[1], MS, RD; Carol Chapman[3], MPH; Christiana Harry[1], MPH; Nelly Lugo[1], BS; Shannon Peabody[1], BA; Shazia Rangwala[4], MPH; Ella Swanson[1], BS; Jonell Tempero[1], BS, MS; Robin Yang[1], MS; Daniel R Karlin[1,5,6], MA, MD; Ron Rabinowicz[7,8], MD; David Malkin[7,9], MD; Simon Travis[10], Prof Dr; Alissa Walsh[10], MD; Robert P Hirten[11], MD; Bruce E Sands[11], MS, MD; Chetan Bettegowda[12], MD, PhD; Matthias Holdhoff[13], MD, PhD; Jessica Wollett[12], MS; Kelly Szajna[12], BSc, RN; Kallan Dirmeyer[12], BS; Anna Dodd[14], MS; Shawn Hutchinson[14], MS; Stephanie Ramotar[14], BSc; Robert C Grant[14], MD, PhD; Adrien Boch[15], MA; Mackenzie Wildman[16], PhD; Stephen H Friend[2,4], MD, PhD

[1]4YouandMe, Seattle, WA, United States
[2]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[3]Crohn's & Colitis Foundation, New York, NY, United States
[4]Section of Urology and Renal Transplantation, Virginia Mason Francisan Health, Seattle, WA, United States
[5]MindMed Inc, New York, NY, United States
[6]Tufts University School of Medicine, Boston, MA, United States
[7]Department of Paediatrics, University of Toronto, Toronto, ON, Canada
[8]Department of Pediatric Hematology/Oncology, Schneider Children's Medical Center of Israel, Petach-Tikva, Israel
[9]Department of Pediatrics, University of Toronto, Toronto, ON, Canada
[10]Gasteroentology Unit, Oxford University Hospitals NHS Foundation Trust and Biomedical Research Centre, Oxford, United Kingdom
[11]The Dr. Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[12]Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[13]The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, United States
[14]Princess Margaret Cancer Center, University Health Network, Toronto, ON, Canada
[15]Evidation Health Inc, Santa Mateo, CA, United States
[16]Sage Bionetworks, Seattle, WA, United States

Corresponding Author:
Sarah M Goodday, MSc, PhD
4YouandMe
2901 3rd Ave
Seattle, WA, 98121
United States
Phone: 1 (206) 928 8243
Email: sarah@4youandme.org

Abstract

Background: Wearable digital health technologies and mobile apps (personal digital health technologies [DHTs]) hold great promise for transforming health research and care. However, engagement in personal DHT research is poor.

Objective: The objective of this paper is to describe how participant engagement techniques and different study designs affect participant adherence, retention, and overall engagement in research involving personal DHTs.

Methods: Quantitative and qualitative analysis of engagement factors are reported across 6 unique personal DHT research studies that adopted aspects of a participant-centric design. Study populations included (1) frontline health care workers; (2) a conception, pregnant, and postpartum population; (3) individuals with Crohn disease; (4) individuals with pancreatic cancer; (5) individuals with central nervous system tumors; and (6) families with a Li-Fraumeni syndrome affected member. All included studies involved the use of a study smartphone app that collected both daily and intermittent passive and active tasks, as well as using multiple wearable devices including smartwatches, smart rings, and smart scales. All studies included a variety of

XSL•FO
RenderX

participant-centric engagement strategies centered on working with participants as co-designers and regular check-in phone calls to provide support over study participation. Overall retention, probability of staying in the study, and median adherence to study activities are reported.

Results: The median proportion of participants retained in the study across the 6 studies was 77.2% (IQR 72.6%-88%). The probability of staying in the study stayed above 80% for all studies during the first month of study participation and stayed above 50% for the entire active study period across all studies. Median adherence to study activities varied by study population. Severely ill cancer populations and postpartum mothers showed the lowest adherence to personal DHT research tasks, largely the result of physical, mental, and situational barriers. Except for the cancer and postpartum populations, median adherences for the Oura smart ring, Garmin, and Apple smartwatches were over 80% and 90%, respectively. Median adherence to the scheduled check-in calls was high across all but one cohort (50%, IQR 20%-75%: low-engagement cohort). Median adherence to study-related activities in this low-engagement cohort was lower than in all other included studies.

Conclusions: Participant-centric engagement strategies aid in participant retention and maintain good adherence in some populations. Primary barriers to engagement were participant burden (task fatigue and inconvenience), physical, mental, and situational barriers (unable to complete tasks), and low perceived benefit (lack of understanding of the value of personal DHTs). More population-specific tailoring of personal DHT designs is needed so that these new tools can be perceived as personally valuable to the end user.

(J Med Internet Res 2024;26:e57827) doi:10.2196/57827

KEYWORDS

wearables; wearable; mHealth; mobile health; app; apps; application; applications; engagement; adherence; retention; participatory medicine; participatory; DHT; digital health technology; DHTs; digital health technologies; digital health; mobile phone

Introduction

Wearable digital health technologies (DHTs) [1,2] and mobile apps facilitate the remote, real-world assessment of health including objective signs of disease that are typically confined to health care visits and health care provider interpretation. These specific categories of DHTs, herein referred to as "personal DHTs," hold promise for transforming health research through the new ability to capture high-resolution, high-frequency, in-the-moment health-related multimodal information in decentralized ways. Through the provision of personal DHTs in clinical care, individuals could be better empowered to navigate their health outside the health care system with greater accessibility, agency, and accuracy than currently possible [1,2]. One of the largest challenges in the future of digital health that involves the use of personal DHTs is end-user engagement. While direct comparisons of engagement in personal DHT research are challenging due to the heterogeneous reporting of retention and adherence factors, and a lack of consensus on a definition of "engagement" [3-6], accumulating evidence supports that so far engagement in the use of personal DHTs has been poor. Specifically, retention in personal DHT research studies and the use of health-related apps is low across diverse populations and applications [7-9]. Further, there is evidence of attrition biases in personal DHT research resulting in insufficient representation of minority populations [7]. In addition to poor retention, personal DHT research studies have low adherence to completing active app-based tasks resulting in large amounts of missing data. This missing data problem results in challenges in artificial intelligence models from insufficient volumes of data to follow individual patterns, and limits app-based context "label" data. This "label" data is crucial for validating passively collected information from personal DHTs, particularly given the early state of the field and as the utility of certain approaches such as knowledge graphs and large language models emerge.

Several personal DHTs health research studies have started to surface [7-12], resulting in the identification of barriers to engagement. These barriers include technical problems with the technology and in collecting the data, usability, privacy concerns, and digital literacy. Many of these barriers point to a need to retain a human element in the research process, and to include an aspect of co-designing with end users. Emerging personal DHT research studies that show better engagement retain some form of "human-in-the-loop" (regular contact with research staff) and co-design or end-user approach [11-15]. Among these studies, retention rates of 80% and higher have been observed, while average adherence to wearable device use and daily app surveys have been shown to be >90% and 70%, respectively [11-15].

The promise of digital health rests on the assumption that end users can be engaged in the long-term use of personal DHTs for health monitoring, yet this remains to be seen among most existing research applications. There have been increasing international calls for the inclusion of patients in the design and conduct of health research [16-18], and this seems particularly relevant for digital health research where the patient is the end user of these new remote tools. In this paper, we report on engagement across 6 unique personal DHT health research studies that adopted different aspects of a participant-centric design, but each with distinct population and design features. The objective is to describe how participant engagement techniques and different personal DHT designs affect participant adherence, retention, and overall engagement in personal DHT health research.

XSL•FO
RenderX

Methods

Study Design

In total, 6 personal DHT research studies are included in this quantitative and qualitative analysis of engagement that span diverse populations including a frontline health care population (the stress and recovery in frontline health care workers study) [11]; a conception, pregnancy, and postpartum population (Better Understanding the Metamorphosis of Pregnancy [BUMP] study) [19]; and populations with different diseases including Crohn disease (stress in Crohn: forecasting symptom transitions study), Li-Fraumeni syndrome (stress and LFS: a feasibility study of wearable technologies to detect stress in families with LFS), and patients with pancreatic and central nervous system (CNS) tumors (help enable real-time observations [HERO] in pancreatic [PANC] and CNS tumors studies) [20].

All of these studies were conducted by 4YouandMe—a US-based nonprofit (charitable) organization. 4YouandMe specializes in open-source research into the application of personal DHTs for health and wellness [20]. 4YouandMe has a particular focus on leveraging personal DHTs to empower the patient in navigating their unique disease or life transitional period. These 6 studies were included in this analysis as they reflect all of the completed studies by 4YouandMe at the time of this analysis. Characteristics of these studies can be found in Table 1 and additional methodological detail can be found in Multimedia Appendix 1. All studies involved the use of a bespoke study smartphone app built by 4YouandMe and the use of the Oura smart ring, the Garmin smartwatch, the Apple smartwatch, an Empatica smartwatch, and the Bodyport Cardiac Scale. Details of these devices can be found in Multimedia Appendix 2).

XSL•FO
RenderX

Table 1. Characteristics of included studies.

Study and population	Sample size	Age (years), median (IQR)	Active study time (months)	Recruitment	Devices	Average (SD) app daily burden	Compensation	Engagement strategy
Stress and recovery								
Frontline health care workers	365	33.0 (28.0-42.0)	4-6	Remote: Social media and health care organization newsletters	• App (iPhone only) • Oura ring • Garmin watch	5 (1.8) minutes	None (participants completing the study kept the wearable devices)	• Biweekly check-in phone calls • Investigator-participant Zoom (Zoom Video Communications, Qumu Corporation) calls • Participant feedback-driven app changes • Coauthorship opportunity
Stress in Crohn								
Patients with Crohn disease	195 (MSSM[a], N=139; Oxford, N=56)	MSSM (median 29, IQR 24-37), Oxford (median 39, IQR 32-50)	6-9	In-clinic: through inflammatory bowel disease clinics	• App (iPhone only) • Oura ring • Empatica Embrace • Bodyport scale	7.7 (1.0) minutes	Yes, participants could keep the ring or receive compensation based on points accumulated	• Biweekly check-in phone calls • Participant feedback-driven app changes
HERO-CNS[b]								
Patients with CNS[c] tumors	12	52 (43-56)	7	In-clinic: through cancer specialty clinics	• App (iPhone or Android) • Oura ring • Garmin	5.3 (2.1) minutes	None (participants completing the study kept the wearable devices)	• Biweekly check-in (phone or in-clinic)
HERO-PANC[d]								
Patients with pancreatic cancer	26	57 (53-65)[e]	1 to 14 months[f]	In-clinic: through cancer specialty clinics	• App (iPhone or Android) • Oura ring • Garmin • Bodyport scale	3.1 (1.9) minutes	None (participants completing the study kept the wearable devices)	• Biweekly check-in (phone or in-clinic)
Stress and LFS[g]								
Affected and unaffected family members of a proband with LFS	49	39.0 (7.9-68.0)	6	In-clinic: through cancer specialty clinics	• App (iPhone only) • Empatica Embrace-Plus	2.3 (0.9) minutes	None	• Biweekly check-in (phone or in-clinic)
BUMP[h]								

XSL•FO
RenderX

Study and population	Sample size	Age (years), median (IQR)	Active study time (months)	Recruitment	Devices	Average (SD) app daily burden	Compensation	Engagement strategy
Pregnant individuals (up to 15 weeks)	524	33.0 (30-36)	Up to 12 months	Remote: through patient-provider portals, social media, and community health clinics	• App (iPhone or Android) • Oura ring • Garmin • Bodyport scale	5.0 (2.3) minutes	Yes, participants received compensation based on study points accumulated	• Biweekly check-in phone calls • Investigator-participant Zoom calls • Participant feedback-driven app changes
BUMP-C[i]								
Individuals actively attempting to get pregnant	273	34.0 (31-36)	Up to 6 months	Remote: through patient-provider portals, social media, and community health clinics	• App (iPhone or Android) • Oura ring	3.8 (2.0) minutes	Yes, participants could keep the ring or receive compensation	• Biweekly check-in phone calls • Investigator-participant Zoom calls • Participant feedback-driven app changes

[a]MSSM: Mount Sinai School of Medicine.

[b]HERO-CNS: help enable real-time observations—central nervous system.

[c]CNS: central nervous system.

[d]HERO-PANC: help enable real-time observations—pancreatic cancer.

[e]n=24, 2 unknown.

[f]Until withdrawal, progression, death, or study completion (October 31, 2022).

[g]LFS: Li-Fraumeni syndrome.

[h]BUMP: Better Understanding the Metamorphosis of Pregnancy.

[i]BUMP-C: Better Understanding the Metamorphosis of Pregnancy—Conception.

Ethical Considerations

All included studies were approved by the local institutional research ethics boards (REB) at their local sites (Multimedia Appendix 1): stress and recovery in frontline health care workers study (institutional review board [IRB], Advarra [4UCOVID1901, Pro00043205]), BUMP study (IRB Advarra Pro00047893), stress in Crohn (Oxford site: Hampshire-A IRAS ID: 269286, Mount Sinai School of Medicine [MSSM] site: IRB of MSSM: GCO 19-1543 | IRB-19-02298), stress and LFS (Sick Kids: REB: 1000072240), HERO-CNS (John Hopkins Medicine IRB IRB00253818), and HERO-PANC (University Hospital Network REB: 20-5211).

Statistical Analysis

Definitions of adherence in digital health research studies are heterogeneous [3-6]. Consistent criteria for adherence across all included studies were attempted. While many different wearable features could be used as the basis for the use of the device, features that were most reliably monitored were selected. For studies using the Oura smart ring, daily adherence was defined as at least one sleep data event present for the prior night. The Oura ring was only expected to be worn at night for many of the included studies, which is why sleep data were used as the indicator for adherence. For studies using the Garmin smartwatch, daily adherence was defined as step data present for that day. For the Empatica smartwatch, daily adherence was defined as at least one data event (worn properly in a day). Adherence to the Bodyport Cardiac Scale was defined as the proportion of days where a weight event was present divided by the total number of expected follow-up days. Adherence to in-app task completion was defined as the proportion of tasks completed when prompted in the app divided by the total number of tasks that should have been completed over study follow-up. For example, all included studies had a daily survey. In a study with a minimum of 4 months of follow-up expected from participants, the total number of expected daily surveys is approximately 120. For a weekly app survey, the total number of expected surveys for a 4-month study follow-up would be 16. Adherence to biweekly check-in calls was defined as the proportion of calls completed divided by the total number of expected calls over study follow-up. Medians and ranges are described since the adherence distributions were nonnormally distributed. All adherence estimations were performed only among retained participants.

Differences in adherence and retention by sociodemographic characteristics were estimated using χ^2, Fisher exact, Mann-Whitney *U*, and ANOVA tests where appropriate among studies that have sufficient sample sizes (stress and recovery, BUMP, and stress in Crohn). Survival probabilities using the Kaplan-Meier approach were calculated to display the probability of retention over the course of each included study. Retention (total proportion of participants completing the study among all enrolled) is also reported. Additional information on

XSL•FO
RenderX

how retention was calculated for each unique study can be found in Multimedia Appendix 3.

Results

Description of Included Studies

Study design characteristics of all studies are described in Table 1. All studies included the use of at least one wearable device plus a study app that involved daily, as well as intermittent surveys (daily question prompts, validated questionnaires) and active tasks (cognitive active or physical function tasks [eg, walk tests], video diaries). In all included studies, participants were required to use their own Android or iPhone smartphone for study activities. Recruitment mechanisms differed across studies with some including remote recruitment through digital advertisements on social media, professional organizations and newsletters, and patient portals (stress and recovery, and BUMP), while others recruited patients in-person through specialty clinics (stress in Crohn, HERO studies, and stress and LFS). The daily burden of app active tasks across studies ranged from 2 to 7 minutes. Study follow-up periods across studies ranged from 4 to 18 months. Across all studies except the stress and LFS study, participants were offered to keep some of the study wearable devices (most often the ring and the watch). Further, 2 studies included the option for modest financial compensation (BUMP and stress in Crohn).

All studies included an engagement strategy that centered around a biweekly phone check-in with a consistent engagement specialist that served the purpose of supporting participants, helping them with onboarding, resolving potential technological problems, and discussing and collecting study experience feedback. Additionally, all included studies implemented different strategies that focused on working with participants as co-designers. These strategies included making app changes that were driven by direct participant feedback during active follow-up, offering a “your data” section in the app that allowed participants to track key symptoms over time, hosting optional investigator-participant Zoom calls where participants could meet the study team, receive study updates, preliminary results, and could offer more feedback, and inviting participants to contribute to and be listed as coauthors on published work.

Adherence by Study Population

Median adherence in engagement phone check-in calls, wearable device use, daily app survey completion, and in-app active tasks can be found in Table 2. Median adherence varied across study populations. The stress in Crohn–MSSM site had a lower adherence on the engagement check-in calls (50%) compared to other studies, many of which had 100% adherence on these calls (Table 2). This study site is herein referred to as the low-engagement cohort. In this low-engagement cohort, median adherence to completing daily app surveys, to wearing the Empatica smartwatch, and to using the Bodyport Cardiac Scale were lower than all other study cohorts that included these studies’ activities (except the BUMP-postpartum cohort). Further, median adherence to using the Oura smart ring was lower in the low-engagement cohort compared to other cohorts except for the postpartum and severely ill cancer populations.

The HERO studies included the most severely ill participants including patients with active diagnoses of CNS and pancreatic tumors. Some HERO participants were undergoing chemotherapy, some had therapy-related complications, some had infections, and some had progressive, life-threatening tumor growth. While the total number of participants in these studies was low, these studies showed low adherence on the daily survey (<55%) and wearable device use (<65% HERO-CNS only). Interestingly, HERO-PANC participants exhibited high wearable device use median adherence (83.3%, IQR 51%-93.2%, Oura and 95.5%, IQR 75.2%-99.2%, Garmin), despite the health status of this population. Further, median adherence to in-app cognitive active tasks was higher among the HERO studies compared to most other studies. Engagement check-in call adherence was also high in the HERO studies. Among the BUMP postpartum cohort, there was consistently lower adherence on all study tasks except for the engagement check-in calls compared to other studies, particularly in comparison to the BUMP prenatal cohort. Specifically, median adherence to the Oura ring, Garmin smartwatch use, and the Bodyport Cardiac Scale in the BUMP-prenatal cohort compared to the BUMP postpartum cohort dropped from 87.2% (IQR 68.7%-96.7%) to 55% (IQR 5.5%-83.7%), 96.7% (IQR 82.9%-100%) to 62.5% (IQR 12.3%-96.4%), and 74.7% (IQR 52%-87.3%) to 33.1% (IQR 8.9%-67.7%), respectively (Table 2).

XSL•FO
RenderX

Table 2. Median adherence to study activities across studies.

	Stress and recovery	BUMP-C[a]	BUMP[b]	BUMP-POST[c]	SINC[d]-MSSM[e]	SINC-Oxford	HERO-CNS[f]	HERO-PANC[g]	Stress in LFS[h]
Participants, n	297	98	379	379	117	54	7	19	45
ES[i] check-ins, median (IQR)	75.0 (57.1-87.5)	100.0 (87.9-100.0)	100.0 (88.4-100.0)	100.0 (100.0-100.0)	50.0 (20.0-75.0)	100.0 (90.9-100.0)	85.7 (78.1-88.2)	100.0 (100.0-100.0)	60.0 (40.0-80.0)
Oura ring, median (IQR)	97.0 (86.0-100.0)	90.6 (76.3-97.7)	87.2 (68.7-96.7)	55.0 (5.5-83.7)	80.5 (37.1-92.4)	98.9 (94.0-99.6)	42.3 (32.0-58.2)	83.3 (51.0-93.2)	__[j]
Garmin watch, median (IQR)	—	—	96.7 (82.9-100.0)	62.4 (12.3-96.4)	—	—	63.3 (54.7-64.3)	95.5 (75.2-99.2)	—
Apple watch, median (IQR)	—	—	98.1 (87.7-100.0)	79.8 (32.4-96.3)	—	—	—	—	—
Empatica watch, median (IQR)	—	—	—	—	26.0 (6.2-64.1)	72.5 (37.1-96.8)	—	—	86.8 (66.7-95.6)
Bodyport scale, median (IQR)	—	—	74.7 (52.0-87.3)	33.1 (8.9-67.7)	38.5 (17.1-64.7)	—	—	79.5 (52.7-88.4)	—
Daily survey, median (IQR)	75.4 (57.2-88.2)	42.4 (24.6-69.7)	60.1 (34.4-81.7)	18.4 (1.0-47.6)	27.9 (10.4-51.9)	70.3 (41.9-84.0)	53.3 (47.8-71.5)	49.1 (20.2-83.4)	62.5 (40.96-82.59)
Reaction rime, median (IQR)	88.9 (75.0-100.0)	43.4 (24.3-72.8)	—	—	30.4 (9.7-50.6)	69.5 (46.6-89.3)	59.0 (50.0-66.7)	62.5 (20.9-86.6)	—
Trail making, median (IQR)	88.9 (71.1-100.0)	46.5 (24.0-73.7)	—	—	28.7 (9.4-50.0)	71.6 (45.0-87.3)	61.5 (52.1-76.5)	38.1 (4.2-76.2)	57.7 (36.8-72.0)
EBT[k], median (IQR)	—	30.1 (16.2-54.1)	44.6 (22.6-73.9)	6.5 (0.0-33.3)	23.1 (9.1-44.4)	32.1 (0.0-58.6)	—	—	—
N-Back, median (IQR)	—	—	51.4 (24.9-76.4)	8.3 (0.0-44.4)	—	—	—	—	—
Gait task, median (IQR)	—	—	25.0 (0.0-60.0)	0.0 (0.0-0.0)	—	—	24.5 (18.8-62.8)	36.0 (2.2-74.0)	—
Walk test, median (IQR)	—	—	14.3 (0.0-40.0)	0.0 (0.0-0.0)	—	—	23.1 (13.9-60.4)	25.0 (7.8-49.5)	—
Video diary, median (IQR)	—	4.3 (0.0-27.7)	8.3 (0.0-50.0)	0.0 (0.0-0.0)	5.6 (0.0-22.2)	9.4 (0.0-35.1)	25.0 (8.7-77.1)	0.0 (0.0-37.5)	—

[a]BUMP-C: Better Understanding the Metamorphosis of Pregnancy—Conception.

[b]BUMP: Better Understanding the Metamorphosis of Pregnancy.

[c]BUMP-POST: Better Understanding the Metamorphosis of Pregnancy—Postpartum.

[d]SINC: stress in Crohn.

[e]MSSM: Mount Sinai School of Medicine.

[f]HERO-CNS: help enable real-time observations—central nervous system.

[g]HERO-PANC: help enable real-time observations—pancreatic cancer.

[h]LFS: Li-Fraumeni syndrome.

[i]ES: engagement specialist.

[j]Not available.

XSL•FO
RenderX

[k]EBT: emotional bias test.

Adherence by Study Activity

There were differences in adherence rates across different study activities. Adherence to wearable device use was consistently higher across studies compared to in-app activities, which is not surprising given the passive nature of these devices. Excluding the postpartum and HERO-CNS study, median adherence to Oura ring use was >80% across all studies, and as high as 99% (IQR 94.9%-99.6%; stress in Crohn-Oxford site; Table 2). There were also differences in adherence across specific wearable devices. Garmin and Apple smartwatch adherence was >95% in BUMP pregnant individuals and HERO-PANC participants, while median adherence for the Empatica Watch was lower among the studies that used this device (stress in Crohn-Oxford, 72.5%, IQR 37.1%-96.8%; stress in Crohn-MSSM, low-engagement cohort, 26%, IQR 6.2%-64.1%; and stress in LFS, 86.8%, IQR 0.7%-0.9%). Median adherence to the Bodyport Cardiac Scale was 74.7% (IQR 52%-87.3%) among BUMP pregnant individuals and 79.5% (IQR 52.7%-88.4%) in HERO-PANC participants (Table 2). Excluding the postpartum and HERO study populations and the low-engagement cohort, in-app daily survey adherence was >60% for all studies (Table 2). Finally, adherence to in-app active tasks was lower in general compared to other activities such as wearable device use or in-app surveys. Tasks that involved walking (gait and walk task) or speaking (video diaries) showed lower adherence compared to other active tasks (eg, cognitive and emotional bias tasks; Table 2).

Adherence by Study Recruitment and Engagement Strategy

There did not appear to be any meaningful difference in median adherence rates across study activities by study recruitment methods (in-clinic vs remote) or follow-up time. Further, 2 studies that included modest financial compensation in addition to engagement strategies showed higher adherence rates compared to some of the other studies (ie, BUMP and stress in Crohn), but the impact of compensation is difficult to disentangle from other study characteristics such as population differences, and these studies did not show superior adherence rates compared to the stress and recovery study that did not offer financial compensation.

Retention

The median proportion of participants retained in the study across the 6 studies was 77.2% (IQR 72.6%-88%; Table 3). The probability of staying in the study stayed above 80% for all studies during the first month of study participation and stayed above 50% for the entire active study period across all studies (Multimedia Appendix 4).

Table 3. Retention across studies.

Study	Proportion retained at study completion, retained/enrolled (%)
Stress and recovery	297/365 (81.4)
BUMP-C[a]	134/187 (72.7)[b]
BUMP[c]	379/524 (72.3)
Stress in Crohn-MSSM[d]	117/139 (84.2)
Stress in Crohn-Oxford	54/56 (96.4)
HERO-CNS[e]	7/12 (58.3)
HERO-PANC[f]	19/26 (73.1)[g]
Stress and LFS[h]	45/49 (91.8)

[a]BUMP-C: Better Understanding the Metamorphosis of Pregnancy—Conception.

[b]Only includes participants who were enrolled in the Better Understanding the Metamorphosis of Pregnancy—Conception-specific app.

[c]BUMP: Better Understanding the Metamorphosis of Pregnancy.

[d]MSSM: Mount Sinai School of Medicine.

[e]HERO-CNS: help enable real-time observations—central nervous system.

[f]HERO-PANC: help enable real-time observations—pancreatic cancer.

[g]Help enable real-time observations—pancreatic cancer has unique factors to consider when interpreting the proportion retained until study completion, since the study aimed to monitor patients until they developed progressive disease or died, or the study end date (October 31, 2022; see Multimedia Appendix 3).

[h]LFS: Li-Fraumeni syndrome.

Adherence and Retention by Participant Sociodemographic Characteristics

Median adherence for the Oura smart ring, a smartwatch (Garmin, Apple, and Empatica), and the Bodyport Cardiac Scale was lower among younger participants compared to older participants across most studies (Multimedia Appendix 5). Specifically, Oura smart ring adherence was significantly lower in those aged 18-25 years compared to those aged ≥26 years in the BUMP study (P=.03) and stress in Crohn-MSSM studies (P=.02), and was lower in the BUMP-C and stress and recover studies, but this difference was not statistically significant at

XSL•FO
RenderX

P=.59 and *P*=.08, respectively. Median adherence for Apple smartwatch use was significantly lower in those aged 18-25 years compared to those aged ≥26 years in the BUMP study (*P*=.02), while median adherence for Garmin smartwatch use was lower but not statistically significant (*P*=.06). Median adherence for the Bodyport Cardiac Scale was significantly lower in those aged 18-25 years compared to those aged ≥26 years in BUMP (*P*<.005) and stress in Crohn-MSSM (*P*<.006).

In the BUMP study, Black or African American ethnicity had significantly higher median adherence to completing the in-app daily survey compared to other race or ethnicity groups (*P*=.01). This trend was observed in the stress and recovery study (*P*=.07) and the stress in Crohn-MSSM study (*P*=.24), although the difference was not statistically significant. In contrast, median adherence to Oura smart ring, smartwatch, and Bodyport Cardiac Scale use was lower among Black or African American individuals compared to other race or ethnicity groups, although these differences were not statistically significant (Multimedia Appendix 5).

Retention did not significantly differ by age group or gender (Multimedia Appendix 6).

Retention likelihood was significantly different by race or ethnicity groups in BUMP-C (*P*<.001) and BUMP (*P*=.001). Specifically, participants of White ethnicity were more likely to stay in the study in both BUMP-C and BUMP, while participants reporting their race or ethnicity as either unknown or not reporting this item were less likely to be retained (Multimedia Appendix 6).

Barriers to Engagement (Qualitative Synthesis of Participant Feedback)

Figure 1 describes key themes that impacted participant retention, adherence, and overall engagement that cut across all included studies. These themes include participant burden and forgetfulness, digital literacy, physical and mental barriers, personal and altruistic benefits, and privacy and confidentiality. Qualitative feedback from participants, research staff, and investigators across these 5 themes is summarized in Multimedia Appendix 7. The top three barriers to engagement in active study tasks were (1) participant burden and in particular fatigue with the repetitiveness of tasks; (2) physical or mental and situational barriers that prevented the ability to complete tasks; and (3) personal and altruistic benefit, namely the perception that the use of the personal DHTs was not personally useful for a health benefit or a lack of understanding as to why and how certain features (eg, heart rate variability) could be useful to track for health benefit. Qualitative feedback from participants in the 2 cohorts demonstrating lower adherence (HERO-PANC and BUMP post partum) suggested that while participants were highly engaged, they were either too ill, distracted, or tired to complete many of the study activities while navigating a serious illness or the early postpartum period.

Figure 1. Five key participant engagement barrier themes.

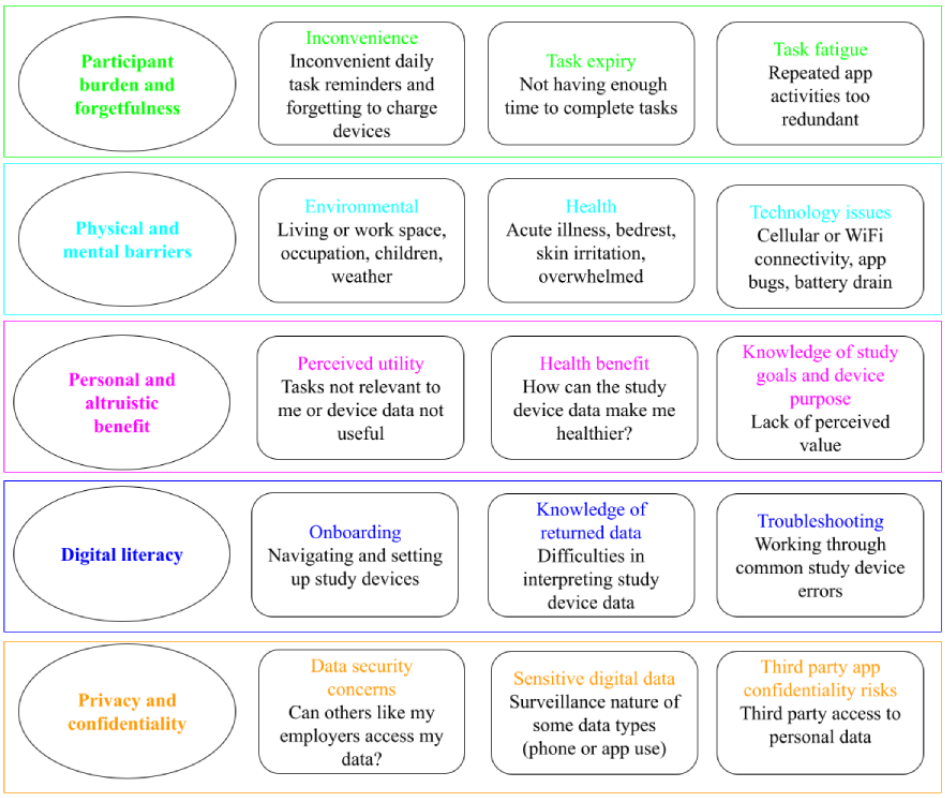

XSL•FO
RenderX

Discussion

Principal Findings

Evidence across 6 unique and diverse studies involving the longitudinal use of personal DHTs supports that participant-centric engagement strategies aid in participant retention and maintaining good adherence in some populations. These strategies centered around (1) human contact with an engagement specialist as often as every 2 weeks, (2) investigator-participant meetings during active study follow-up, (3) offering returned symptom data in the app, (4) inviting participants to contribute as coauthors in published work, and (5) real-time modifications to the study app based on participant feedback.

In the majority of included studies, the probability of staying in the study stayed above 90% for the first month and stayed above 50% for active study periods for all studies. Lower retention or adherence was observed among studies that included a severely ill cancer population and a postpartum population. Barriers to participation in these cohorts were largely the result of physical and situational roadblocks. Excluding studies of a severely ill and postpartum population and the low-engagement cohort in the stress in Crohn study, adherence to Oura smart ring and Garmin smartwatch use was 80% and as high as 99% in some cohorts, while adherence to the Bodyport Cardiac Scale was 75% in a pregnant population. This supports that different populations can successfully be engaged in the use of active app assessments and wearable devices in the long term with adequate support.

Retention and adherence rates observed in these studies are higher than typically reported by other personal DHT research studies [7-9,12,13,21]. For example, a review of 8 large app-based DHT research studies in the United States reported that the probability of staying in the study dropped to or below 50% after the first 4 weeks of participation for all included studies [7]. Further, across the 8 included studies in this review, >50% of participants did not engage with the app for at least 7 days. Another large app-based study in the United States, the Warfighter Analytics Using Smartphones for Health study that collected daily active and passive app data reported a median retention of 45.2% (38/84 days), while the probability of staying in the study hit 50% at approximately 5.5 weeks [10]. A large app-based study in the United Kingdom (cloudy with a chance of pain study) involving daily active app assessments reported that 64% of participants fell into the low engagement or no engagement categories after baseline [12]. The RADAR study [14], a multinational study involving active and passive assessments from an app, and a Fitbit reported comparable retention results among participants with major depression to those reported here. This study reported a retention rate of 54.6% for 43 weeks of study participation; however, the probability of staying in the study stayed above 75% for the first several months of participation (~6 months). While the active app assessments in this study only included assessments every 2 weeks as opposed to daily assessments, this study additionally included aspects of a participant-centric design, which may have contributed to the higher reported retention [15].

Taken together, in comparison to other published personal DHT research studies, the 6 studies included in this paper reflect higher levels of engagement. Importantly, the included studies in this analysis involved high burden designs in comparison to other studies that request, for example, weekly or biweekly active tasks of participants [14] or only involve the use of a smartwatch. Specifically, across the included studies here, participants were expected to complete on average 4.6 (SD 1.62) minutes a day of app activities in addition to continuously using multiple wearable devices.

While different variations of participant-centric strategies were used across the 6 included studies, a key common feature was a biweekly check-in call with an engagement specialist. These calls served the purpose of providing support and building rapport with participants, working through onboarding and technological issues with study devices, tracking adherence, and receiving study-related feedback from participants. Numerous challenges arise in the conduct of remote, personal DHT research, and without frequent check-in and semiregular data monitoring by research staff, knowledge of these issues is a black box. The most significant drop in retention in personal DHT research studies tends to be during the first few weeks of participation [7]. These early onboarding weeks are crucial in working with participants to ensure they can get into a rhythm of participation. The passive sensing nature of personal DHTs has much potential to inform new objective measures of health, however, are not always intuitively understood as personally important for unique diseases (eg, heart rate variability or phone screen time). Personal DHT studies allow for “light touch” research approaches that enable data collection without traditional research coordinator contact, but this may come with a cost that inadvertently creates a less engaging study environment for participants and limits the opportunity to help participants understand the value in their participation. Of the included 6 studies, 1 cohort had much lower engagement on the check-in calls (50% adherence) compared to other included studies and, in turn, consistently demonstrated lower adherence to study-related activities. Still, even with extensive engagement designs, populations that had physical, mental, and situational barriers to study task completion (ie, severely ill, postpartum mothers) showed lower adherence to wearable device use and active smartphone tasks compared to other study populations. Top reported barriers to engagement included participant burden, physical, mental, and situational barriers, and low perceived value of personal DHTs for health care. These engagement barriers have been reported in previous literature [8,9] relating to DHT research and in the use of DHT interventions. However, the conveyed importance of the perceived value of the approach among participants in the current analysis is noteworthy. Given the foreign nature of personal DHTs for many individuals, particularly older populations, further work is needed to co-design and educate end users on the potential value of self-monitoring unique health-related data.

Irrespective of the engagement approach, adherence to in-app surveys and tasks was lower than wearable device use, which is not surprising given the higher burden related to in-app activities. The self-reported information captured from frequent or momentary in-app assessments is extremely valuable as

XSL•FO
RenderX

context information. This context information or "label" data is useful for validating objectively captured information, yet remains the most difficult to capture in sufficient detail. Further, certain in-app activity adherences were consistently lower than others. Namely, activities that required the user to be active (walk in a straight line or complete a video diary) were low across studies. Still, adherence to daily in-app surveys was >60% for all studies excluding the postpartum and HERO study populations.

Limitations

This quantitative and qualitative analysis compared observational data across different digital health studies. However, no true comparison cohort that did not include engagement strategies was included. Therefore, the inferred casualty of participant check-ins with engagement specialists on retention and adherence rates cannot be not concluded. We are formally testing whether the biweekly check-in significantly increases adherence and retention in an ongoing study with an appropriate comparison arm without check-in support (NCT05753605). One of the included studies (stress and recovery) was conducted during the early 2020 COVID-19 pandemic. There is some evidence that engagement in research was higher during the early pandemic time periods [22]. It cannot be ruled out that the higher observed retention and adherence in this study compared to others was not due to this potential time period bias. The stress in the Crohn-Oxford site included a population of patients some of whom were already engaged in the use of web-based monitoring of symptoms. In turn, this could have contributed to the high retention and higher adherence observed at this site compared to the other stress in the Crohn-MSSM site. The results presented on barriers to engagement were primarily qualitative and collected from conversations with participants, research staff, and investigators across studies.

Conclusions

Globally, mobile apps are used for a variety of purposes in everyday life, while the use of smartwatches for activity monitoring is gaining increasing popularity. However, the use of these tools for health remains a challenge. These findings support that human support via phone and other participant-centric engagement strategies centered on giving back to participants and working with them as co-designers can support sufficient retention and adherence in personal DHT research across diverse populations. This has implications for the utility and potential necessity of a digital support worker in digital health care, as highlighted by others [23]. A power of personal DHTs is enabling the patient to be in control of their health through self-monitoring, but this new role comes with a responsibility. This important shift in role from doctor to patient outlines how crucial it is to include patients in the early design phase of personal DHT health research. Further work is needed to inform app designs that support habitual forming activities around task completion so that app-related activities become a part of participants' daily routine and are perceived as personally valuable.

Acknowledgments

The stress and recovery study was supported in part by the Bill & Melinda Gates Foundation (INV-016651). The stress in Crohn study was funded by the Leona M. and Harry B. Helmsley Charitable Trust (1911-03376). The help enable real-time observation (HERO)–central nervus system study was funded by the Mark Foundation for Cancer Research through an ASPIRE award (19-024-ASP). The HERO–pancreatic cancer study was funded by the Mark Foundation for Cancer Research through an ASPIRE award (19-024-ASP), Pancreatic Cancer Canada, the Princess Margaret Cancer Foundation, and 4YouandMe. The Better Understanding the Metamorphosis of Pregnancy (BUMP) study was funded by 4YouandMe and Sema4 along with supplemental in-kind contributions from coalition partners (Evidation Health, Vector Institute, Cambridge Cognition, and Bodyport). The stress and LFS study was funded by in-kind contributions from 4YouandMe, SickKids Hospital, and the Vector Institute.

Conflicts of Interest

CB is a consultant for Depuy Synthes, Bionaut Labs, Galectin Therapeutics, Haystack Oncology, and Privo Technologies. CB is a cofounder of Belay Diagnostics and OrisDx. DRK is an officer, employee, and shareholder of MindMed; a consultant at Tempus, Nightware, and Limitless; and board member of Sonara. RPH is an advisory board member at Bristol Meyers Squibb. MH is an advisory board member for Servier, AnHeart, and Bayer; steering committee member for Novartis; honoraria from Novartis; data safety monitoring committee member for Advarra and Parexel. RG received a graduate scholarship from Pfizer and provided consulting or advisory roles for Astrazeneca, Tempus, Eisai, Incyte, Knight Therapeutics, Guardant Health, and Ipsen. The others declare no conflicts of interest.

Multimedia Appendix 1
Study descriptions.
[DOCX File , 20 KB - jmir_v26i1e57827_app1.docx]

Multimedia Appendix 2
Study wearable devices.
[DOCX File , 16 KB - jmir_v26i1e57827_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Retention calculations.
[DOCX File , 28 KB - jmir_v26i1e57827_app3.docx]

Multimedia Appendix 4
Probability of retaining in the study across studies.
[DOCX File , 134 KB - jmir_v26i1e57827_app4.docx]

Multimedia Appendix 5
Median adherence to study activities stratified by sociodemographic characteristics.
[DOCX File , 28 KB - jmir_v26i1e57827_app5.docx]

Multimedia Appendix 6
Sociodemographic differences in participants who were retained versus not retained.
[DOCX File , 28 KB - jmir_v26i1e57827_app6.docx]

Multimedia Appendix 7
Qualitative feedback from participants, research staff, and investigators surrounding barriers to engagement in digital health research, summarized across 6 unique studies.
[DOCX File , 20 KB - jmir_v26i1e57827_app7.docx]

References

1. Friend SH, Ginsburg GS, Picard RW. Wearable digital health technology. N Engl J Med 2023;389(22):2100-2101. [doi: 10.1056/NEJMe2303219] [Medline: 38048193]
2. Goodday SM, Geddes JR, Friend SH. Disrupting the power balance between doctors and patients in the digital era. Lancet Digit Health 2021;3(3):e142-e143 [FREE Full text] [doi: 10.1016/S2589-7500(21)00004-2] [Medline: 33516653]
3. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
4. White KM, Williamson C, Bergou N, Oetzmann C, de Angel V, Matcham F, et al. A systematic review of engagement reporting in remote measurement studies for health symptom tracking. NPJ Digit Med 2022;5(1):82 [FREE Full text] [doi: 10.1038/s41746-022-00624-7] [Medline: 35768544]
5. Daniore P, Nittas V, von Wyl V. Enrollment and retention of participants in remote digital health studies: scoping review and framework proposal. J Med Internet Res 2022;24(9):e39910 [FREE Full text] [doi: 10.2196/39910] [Medline: 36083626]
6. Chan A, Chan D, Lee H, Ng CC, Yeo AHL. Reporting adherence, validity and physical activity measures of wearable activity trackers in medical research: a systematic review. Int J Med Inform 2022;160:104696 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104696] [Medline: 35121356]
7. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
8. Simblett S, Matcham F, Siddi S, Bulgari V, Barattieri di San Pietro C, Hortas López J, RADAR-CNS Consortium. Barriers to and facilitators of engagement with mHealth technology for remote measurement and management of depression: qualitative analysis. JMIR mHealth uHealth 2019;7(1):e11325 [FREE Full text] [doi: 10.2196/11325] [Medline: 30698535]
9. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair F. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
10. Li SX, Halabi R, Selvarajan R, Woerner M, Fillipo IG, Banerjee S, et al. Recruitment and retention in remote research: learnings from a large, decentralized real-world study. JMIR Form Res 2022;6(11):e40765 [FREE Full text] [doi: 10.2196/40765] [Medline: 36374539]
11. Goodday SM, Karlin E, Alfarano A, Brooks A, Chapman C, Desille R, Stress And Recovery Participants, et al. An alternative to the light touch digital health remote study: the stress and recovery in frontline COVID-19 health care workers study. JMIR Form Res 2021;5(12):e32165 [FREE Full text] [doi: 10.2196/32165] [Medline: 34726607]
12. Druce KL, McBeth J, van der Veer SN, Selby DA, Vidgen B, Georgatzis K, et al. Recruitment and ongoing engagement in a UK smartphone study examining the association between weather and pain: cohort study. JMIR mHealth uHealth 2017;5(11):e168 [FREE Full text] [doi: 10.2196/mhealth.8162] [Medline: 29092810]
13. Reade S, Spencer K, Sergeant JC, Sperrin M, Schultz DM, Ainsworth J, et al. Cloudy with a chance of pain: engagement and subsequent attrition of daily data entry in a smartphone pilot study tracking weather, disease severity, and physical activity in patients with rheumatoid arthritis. JMIR mHealth uHealth 2017;5(3):e37 [FREE Full text] [doi: 10.2196/mhealth.6496] [Medline: 28341616]

XSL•FO
RenderX

14. Matcham F, Leightley D, Siddi S, Lamers F, White KM, Annas P, RADAR-CNS consortium. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022;22(1):136 [FREE Full text] [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
15. Zhang Y, Pratap A, Folarin AA, Sun S, Cummins N, Matcham F, et al. Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. NPJ Digit Med 2023 Feb 17;6(1):25. [doi: 10.1038/s41746-023-00749-3] [Medline: 36806317]
16. How we engage the public. Wellcome. URL: https://wellcome.org/what-we-do/public-engagement [accessed 2024-02-02]
17. Canada's strategy for patient-oriented research. Canadian Institutes of Health Research. URL: https://cihr-irsc.gc.ca/e/44000.html [accessed 2024-02-02]
18. NIHR launches new centre for engagement and dissemination. National Institute of Health and Care Research. URL: https://www.nihr.ac.uk/news/nihr-launches-new-centre-for-engagement-and-dissemination/24576 [accessed 2024-02-02]
19. Goodday SM, Karlin E, Brooks A, Chapman C, Karlin DR, Foschini L, et al. Better Understanding of the Metamorphosis of Pregnancy (BUMP): protocol for a digital feasibility study in women from preconception to postpartum. NPJ Digit Med 2022 Mar 30;5(1):40. [doi: 10.1038/s41746-022-00579-9] [Medline: 35354895]
20. 4YouandMe. URL: http://4youandme.org [accessed 2023-06-07]
21. Nowell WB, Curtis JR, Zhao H, Xie F, Stradford L, Curtis D, et al. Participant engagement and adherence to providing smartwatch and patient-reported outcome data: digital tracking of rheumatoid arthritis longitudinally (DIGITAL) real-world study. JMIR Hum Factors 2023;10:e44034 [FREE Full text] [doi: 10.2196/44034] [Medline: 37934559]
22. Vargo D, Zhu L, Benwell B, Yan Z. Digital technology use during COVID-19 pandemic: a rapid review. Human Behav and Emerg Tech 2020 Dec 28;3(1):13-24. [doi: 10.1002/hbe2.242]
23. Perret S, Alon N, Carpenter-Song E, Myrick K, Thompson K, Li S, et al. Standardising the role of a digital navigator in behavioural health: a systematic review. Lancet Digit Health 2023;5(12):e925-e932 [FREE Full text] [doi: 10.1016/S2589-7500(23)00152-8] [Medline: 38000876]

Abbreviations

BUMP: Better Understanding the Metamorphosis of Pregnancy
CNS: central nervous system
DHT: digital health technology
HERO: help enable real-time observation
IRB: institutional review board
LFS: Li-Fraumeni syndrome
MSSM: Mount Sinai School of Medicine
PANC: pancreatic cancer
REB: research ethics board

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 06.03.24; peer-reviewed by C Godoy Jr; comments to author 05.04.24; revised version received 12.04.24; accepted 29.05.24; published 03.09.24.

Please cite as:
Goodday SM, Karlin E, Brooks A, Chapman C, Harry C, Lugo N, Peabody S, Rangwala S, Swanson E, Tempero J, Yang R, Karlin DR, Rabinowicz R, Malkin D, Travis S, Walsh A, Hirten RP, Sands BE, Bettegowda C, Holdhoff M, Wollett J, Szajna K, Dirmeyer K, Dodd A, Hutchinson S, Ramotar S, Grant RC, Boch A, Wildman M, Friend SH
Value of Engagement in Digital Health Technology Research: Evidence Across 6 Unique Cohort Studies
J Med Internet Res 2024;26:e57827
URL: https://www.jmir.org/2024/1/e57827
doi:10.2196/57827
PMID:

©Sarah M Goodday, Emma Karlin, Alexa Brooks, Carol Chapman, Christiana Harry, Nelly Lugo, Shannon Peabody, Shazia Rangwala, Ella Swanson, Jonell Tempero, Robin Yang, Daniel R Karlin, Ron Rabinowicz, David Malkin, Simon Travis, Alissa Walsh, Robert P Hirten, Bruce E Sands, Chetan Bettegowda, Matthias Holdhoff, Jessica Wollett, Kelly Szajna, Kallan Dirmeyer, Anna Dodd, Shawn Hutchinson, Stephanie Ramotar, Robert C Grant, Adrien Boch, Mackenzie Wildman, Stephen H Friend. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predicting Long-Term Engagement in mHealth Apps: Comparative Study of Engagement Indices

Yae Won Tak[1], MSc; Jong Won Lee[2], MD, PhD; Junetae Kim[3*], PhD; Yura Lee[1*], MD, PhD

[1]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]Division of Breast Surgery, Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[3]Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Yura Lee, MD, PhD
Department of Information Medicine, Asan Medical Center
University of Ulsan College of Medicine
88, Olympic-Ro 43-Gil, Songpa-Gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 1498
Email: haepary@amc.seoul.kr

Abstract

Background: Digital health care apps, including digital therapeutics, have the potential to increase accessibility and improve patient engagement by overcoming the limitations of traditional facility-based medical treatments. However, there are no established tools capable of quantitatively measuring long-term engagement at present.

Objective: This study aimed to evaluate an existing engagement index (EI) in a commercial health management app for long-term use and compare it with a newly developed EI.

Methods: Participants were recruited from cancer survivors enrolled in a randomized controlled trial that evaluated the impact of mobile health apps on recovery. Of these patients, 240 were included in the study and randomly assigned to the Noom app (Noom Inc). The newly developed EI was compared with the existing EI, and a long-term use analysis was conducted. Furthermore, the new EI was evaluated based on adapted measurements from the Web Matrix Visitor Index, focusing on click depth, recency, and loyalty indices.

Results: The newly developed EI model outperformed the existing EI model in terms of predicting EI of a 6- to 9-month period based on the EI of a 3- to 6-month period. The existing model had a mean squared error of 0.096, a root mean squared error of 0.310, and an R^2 of 0.053. Meanwhile, the newly developed EI models showed improved performance, with the best one achieving a mean squared error of 0.025, root mean squared error of 0.157, and R^2 of 0.610. The existing EI exhibited significant associations: the click depth index (hazard ratio [HR] 0.49, 95% CI 0.29-0.84; $P<.001$) and loyalty index (HR 0.17, 95% CI 0.09-0.31; $P<.001$) were significantly associated with improved survival, whereas the recency index exhibited no significant association (HR 1.30, 95% CI 1.70-2.42; $P=.41$). Among the new EI models, the EI with a menu combination of menus available in the app's free version yielded the most promising result. Furthermore, it exhibited significant associations with the loyalty index (HR 0.32, 95% CI 0.16-0.62; $P<.001$) and the recency index (HR 0.47, 95% CI 0.30-0.75; $P<.001$).

Conclusions: The newly developed EI model outperformed the existing model in terms of the prediction of long-term user engagement and compliance in a mobile health app context. We emphasized the importance of log data and suggested avenues for future research to address the subjectivity of the EI and incorporate a broader range of indices for comprehensive evaluation.

(J Med Internet Res 2024;26:e59444) doi:10.2196/59444

KEYWORDS

treatment adherence and compliance; patient compliance; medication adherence; digital therapeutics; engagement index; mobile phone

XSL•FO
RenderX

Introduction

Digital therapeutics (DTx) has the potential to expand accessibility and enhance engagement for patients by addressing the limitations associated with conventional facility-based medical treatments [1]. These interventions have gained considerable attention owing to their effectiveness in addressing various health challenges, which has led to their increasing adoption rate in health care settings [2]. Although DTx offers unique monitoring capabilities, enabling health care providers to remotely track patient progress and tailor interventions, their use remains controversial because of the ambiguity in terms of the DTx's standpoint and effectiveness [3]. A diverse range of DTx, from smartphone apps for mental health support to wearable devices for chronic disease management, are available to meet the evolving needs of patients and health care providers alike with the availability of real-time and continuous log data for further improvements [4,5]. One such technology that has attracted research interest is mobile health (mHealth), which is used to monitor patients.

In recent years, the use of mHealth technologies in cancer care has steadily increased, offering a promising avenue for improving patient outcomes and revolutionizing health care delivery [6]. With the convenience of the high distribution rate of smartphones of over 86.11% globally, mHealth apps have been increasingly integrated into cancer management, providing patients with remote access to personalized care through physical fitness support, weight management, therapy, information provision, and social support [7,8]. Despite the growing adoption of mHealth solutions in cancer care, existing literature reviews have highlighted a significant challenge, that is, the absence of standardized measures for assessing the use of and compliance with these technologies [9].

The vast majority of studies evaluating intervention engagement rely on either postintervention surveys or interviews [10-13]. Furthermore, when assessing the effectiveness of therapeutic education systems, the methodology often involves twice-weekly clinical checkups and self-reports, despite the pioneering nature of the randomized controlled trial (RCT) for internet-delivered interventions [14]. This highlights the need for a more systematic methodology for evaluating mHealth intervention engagement rather than solely relying on subjective interviews.

This study evaluates the engagement index (EI) in commercial health management apps for long-term use by comparing the newly developed EI model with the existing model.

Methods

Study Design

This study aimed to confirm whether a newly developed EI better predicts long-term compliance than an existing EI by using the Web Matrix Visitor Index with modifications, focusing on indices such as click depth, recency, and loyalty based on the Noom app (Noom Inc) usage data. A new menu abundancy index (MI) was introduced, considering the survival time of each menu. In addition, the loyalty index (LI) was enriched by incorporating the final usage week, and the recency index (RI) was refined using permutation entropy to measure the regularity of app usage. This study analyzed data from 233 patients who used the Noom app, part of an RCT involving 960 cancer survivors (breast, colorectal, or lung cancer) aimed at assessing the impact of mHealth apps on recovery. The Noom app, a commercially available weight management tool, was used for its features such as meal logging, step count tracking, weight logging, exercise logging, engagement with health-related content, and messaging.

Study Population

Data obtained from patients who were recruited from an RCT that investigated the impact of mHealth apps on cancer survivors were used; research aimed to facilitate a smoother recovery for patients with breast, colorectal, or lung cancer as they transition back to their daily lives [15-17]. Written informed consent was obtained from all the participants before study participation. Subsequently, the participants were randomly assigned to 1 of 3 mHealth care groups, including the Noom app group. Of the 960 participants, 233 who used Noom were analyzed for this study.

Data Collection

Clinical and pathological information related to demographics were extracted from the electronic medical records of patients at the time of recruitment. The data collection was extended up to 18 months after the final patient enrollment, with follow-up assessments scheduled at 3 months and every 6 months after the initial baseline data collection.

Noom, a commercially available mobile app for weight management, can be downloaded from the Google Play Store and the Apple App Store [18]. Distinguished by its distinctive curriculum and human coaching intervention, Noom is a prominent feature in the realm of health and fitness apps [18]. It strives to be a versatile platform for behavioral change, serving as a potent tool for addressing diverse chronic and nonchronic conditions, with the goal of promoting healthier lifestyles for a wider population [19]. Noom has been shown to be an effective mHealth lifestyle platform, with positive results yielded in various clinical scenarios [20-22]. In this study, various features of Noom were used, including, but not limited to, meal logging, step count tracking, weight logging, exercise logging, engagement in health-related content, and messaging functionalities.

Data Analysis

This study aimed to confirm whether the newly developed EI better predicts long-term compliance than the existing EI. To achieve this, the goal was to predict compliance at 6-9 months or predict survival rates based on that at 3-6 months, with the aim of comparing the performance of the 2 indices. All data analyses were conducted using Python (version 3.8.5; Python Software Foundation).

New Engagement Index

At present, there is no established tool to measure engagement in health care apps, thus, we adopted the Web Matrix Visitor Index [23] to effectively measure engagement in the commercial health management app for cancer patients using 7 indices, that

XSL•FO
RenderX

are click depth, duration, recency, loyalty, brand, feedback, and interaction. Of the 7 indices, we used 3 (click depth, recency, and loyalty) that were applicable; these could be calculated using the app access log data. Click depth measures the impact of page and event views, whereas recency indicates the speed at which visitors return to the website over time. Specifically, click depth is computed by dividing sessions with a reasonable threshold (eg, 4 pages viewed) by all sessions. Loyalty gauges the extent of long-term interaction with the brand, site, or products. Recency is calculated as 1 divided by the number of days since the most recent session, whereas loyalty is derived by subtracting 1 from the number of visitor sessions during the timeframe from 1.

As the app was not able to provide information when accessed each time, we defined the sessions as 1 day with each index ranging between 0 and 1. Click depth was calculated as the number of weeks with at least 2 menus accessed divided by the number of the current week. Loyalty was calculated as the number of accessed weeks divided by the number of this week. Recency was calculated as 1 over the average number of weeks between visits for each period. Finally, EI was calculated as the average (mean) between the click depth, loyalty, and recency indices.

Limitations of Engagement Index

Despite its generalizability, the EI encounters several limitations. First, it may not fully capture all dimensions of user engagement, thus leading to an incomplete representation of user patterns. Second, it is heavily influenced by the natural characteristics of app usage; particularly, over the long term, it can complicate the assessment of long-term app effectiveness. Third, it may fail to account for changes in engagement patterns over time, which limits its applicability in monitoring-maintained user involvement. Finally, its subjective nature could emerge in the metrics when calculating it, thus potentially introducing biases.

Calculation of the New Engagement Index

We aimed to enhance EI and its components based on its original characteristics. As the click depth index failed to account for the number of menus available in the app, we introduced the MI, which was devised from the click depth index. The MI considers the different menus offered by the app. By computing the survival time of each menu, with discontinuation defined as continuous nonusage for 45 days, we constructed a vector for each patient, where each co-ordinate represents the survival time of a menu. We chose 45 days as it represents the 75th percentile value of the nonusage period between usages. Subsequently, we computed the Euclidean distance from the origin for each vector. Consequently, menu abundancy is determined as the Euclidean distance between the patient and the patient with the minimum Euclidean distance, divided by the Euclidean distance between the largest and smallest vectors. The diagram and equation used for these calculations are presented in Figure 1.

Figure 1. Procedure for calculating the menu abundancy index.

When $\boldsymbol{x}$ is considered, let $\boldsymbol{a}$ represent a patient exhibiting maximum Euclidean length, and let $\boldsymbol{b}$ represent a patient exhibiting minimum Euclidean length.

$$\frac{|\boldsymbol{x}-\boldsymbol{b}|}{|\boldsymbol{a}-\boldsymbol{b}|}$$

The figure below shows an example when the Menu has three menus.

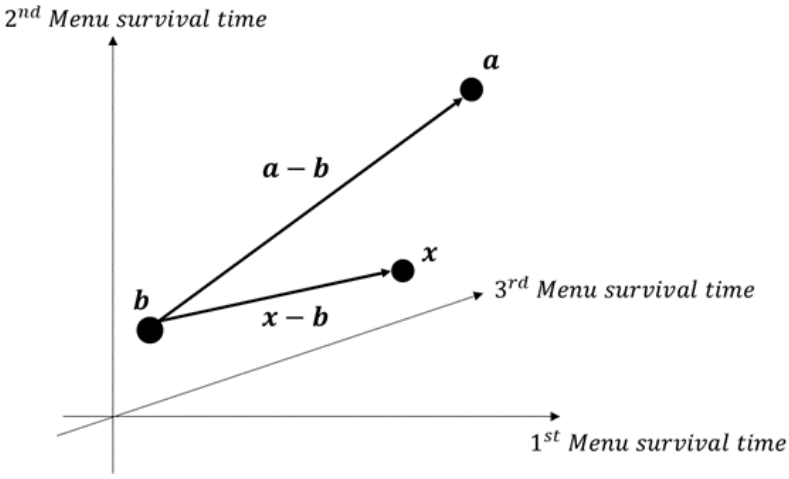

To enrich the existing LI, which solely considers the number of accessed weeks, we deemed it crucial to incorporate the number of the final weeks used. However, we aimed to prevent this addition from disproportionately influencing the overall index. Therefore, we formulated the LI as the sum of the number of accessed weeks and the natural logarithm of the final usage week number, divided by the total number of weeks used in the study. Due to some resulting values exceeding 1, we applied a simple linear transformation to all values. This involved dividing each value by the maximum value of the LI observed in the study, thus ensuring that the new LI ranged from 0 to 1.

The RI calculates the regularity of app usage. To quantify this, we used permutation entropy (PE), which is a robust time series tool. PE quantifies the complexity of a dynamic system by capturing order relations within a time series and deriving a

XSL•FO
RenderX

probability distribution of ordinal patterns [24]. To ensure that PE falls within the range of 0-1, we applied a truncated normal distribution to the values. As lower PE values indicate higher regularity and vice versa, we subtracted the value from 1 to align with the existing RI, where a higher value corresponds to more regular visits. This adjustment maintains consistency with the mathematical representation of the RI.

Assumptions

We measured the effectiveness of the newly developed EI through multiple linear regression and survival analysis. These statistical methods were selected to provide deep insights into the interpretability and statistical significance of predictors.

Multiple linear regression reveals the linear relationships between the dependent variable (ie, EI) and the multiple independent variables. Considering these multiple independent variables as predictors could provide a better understanding of the multidimensional nature of usage engagement influenced by factors. Furthermore, this approach facilitates the identification of the most important drivers of long-term engagement, thereby contributing to the development of a more reliable and tailored EI capable of capturing the nuances of patient behavior and adherence patterns. The multiple linear regression assumptions were evaluated for our model. Linearity between the variables was assessed, with values indicating strong correlations: menu abundancy (0.93), LI (0.93), and RI (0.69), all close to 1. The normality of the residuals was supported by Shapiro-Wilk test results: menu abundancy (0.84), LI (0.83), RI (0.85), and new EI (0.90). However, the independence of residuals, as measured by the Durbin-Watson statistic, showed values of 1.02 (menu abundancy), 1.10 (LI), 0.95 (RI), and 0.88 (new EI2), indicating potential autocorrelation. Given that these indices are derived from the same app usage data, achieving complete independence is inherently challenging.

Multiple linear regression can be expressed in a generalized form, that is, formula 1

$$(EI_{m:m+3} = \beta_0 + \beta_1 \cdot MI_{m:m+3} + \beta_2 \cdot LI_{m:m+3} + \beta_3 \cdot RI_{m:m+3} + \varepsilon) \quad \textbf{(1)}$$

where the dependent variable is defined as formula 2

$$(EI_{m:m+3} := (MI_{m:m+3} + LI_{m:m+3} + RI_{m:m+3})/3) \quad \textbf{(2)}$$

Moreover, β_0 denotes the intercept of the model; β_1 to β_3, the coefficients for the predictors of the MI, LI, and RI, respectively, and ε, the error term, which accounts for the variance in the prediction that cannot be explained by the predictors. For regression models, we set 2-month durations, that is, 3-6 and 6-9 months; hence, the values of *m* for each model are 3 and 6, respectively.

Survival analysis, particularly through Cox regression, shows the discontinuation timing of app usage over a specified period. This method accounts for censored data and provides hazard ratios (HRs), quantifying the effect of different predictors on the likelihood of continued app usage. Thus, this approach helps identify which aspect of patient usage pattern is the most predictive and significant for long-term compliance.

Formula 3,

$$(h(t) = h_0(t)exp(\beta_1 \cdot MI_{3:6} + \beta_2 \cdot LI_{3:6} + \beta_3 \cdot RI_{3:6})) \quad \textbf{(3)}$$

where $h(t)$ denotes the hazard function at time t for predicting the survival days of each app user and $h_0(t)$ denotes the baseline hazard function, which is the hazard for an individual when all the covariates are 0.

Ethical Consideration

This study was approved by the Institutional Review Board of Asan Medical Center, Korea (Institutional Review Board 2021-1631). Stringent measures were in place to protect the privacy and confidentiality of study data, including secure storage within the hospital premises.

Results

Demographic Traits

The demographic characteristics of the study cohort are presented in Table 1. The average age of the patients was 53.55 years, with women constituting approximately two-thirds of the cohort. Each group of cancer type comprised a comparable number of participants, albeit colorectal cancer cases slightly outnumbered the others. Over half of the patients were diagnosed with stage 1, and approximately two-thirds had not undergone chemotherapy. The average BMI of the patients was 23.96 (SD 3.28) kg/m^2.

XSL•FO
RenderX

Table 1. Demographics of participants (N=233).

Demographics	Values
Age (years), mean (SD)	53.55 (10.35)
Gender , n (%)	
Men	80 (34)
Women	153 (66)
Cancer type , n (%)	
Breast	78 (33)
Colorectal	86 (37)
Lung	69 (30)
Cancer stage , n (%)	
0	19 (8)
I	131 (56)
II	47 (20)
III	36 (15)
Chemotherapy , n (%)	
Yes	73 (31)
No	160 (69)
BMI (kg/m^2), mean (SD)	23.96 (3.28)
Living , n (%)	
With family	211 (90)
Alone	20 (9)
Other	2 (1)
Education , n (%)	
Less than high school	22 (10)
High school graduate	82 (35)
College graduate or above	129 (55)
Job , n (%)	
Employed	140 (60)
Unemployed	93 (40)

Evaluation Between the Existing Engagement Index and New Engagement Index

Predicting 6-9 Months Engagement Index Based on the 3-6 Months Engagement Index

To predict the EI of patients with cancer between 6 and 9 months based on their EI between 3 and 6 months, we excluded the initial 0- to 3-month period as the patients were actively under hospital surveillance with ongoing follow-ups. For the existing EI, we observed a mean squared error (MSE) of 0.096, root mean squared error (RMSE) of 0.310, and R^2 of 0.053. For the new EI, we conducted 3 multiple linear regressions to identify the most significant menu combinations. The first combination (new EI1), comprising meal log, exercise log, message sent to the app, reading content, and weight log, exhibited an MSE of 0.036, RMSE of 0.190, and R^2 of 0.511. The second combination (new EI2), involving meal log, exercise log, weight log, and step count login, showed improved performance with an MSE of 0.025, RMSE of 0.157, and R^2 of 0.610. The third combination (new EI3), encompassing meal log, exercise log, message sent to the app, reading content, weight log, and step count login, yielded an MSE of 0.042, RMSE of 0.205, and R^2 of 0.374. The values of the multiple linear regression are presented in Table 2.

XSL•FO
RenderX

Table 2. Multiple linear regression results for the existing engagement index and the new engagement index.

	MSE[a]	RMSE[b]	R^2
Existing EI[c]	0.096	0.310	0.053
New EI1	0.036	0.190	0.511
New EI2	0.025	0.157	0.610
New EI3	0.042	0.205	0.374

[a]MSE: mean squared error.

[b]RMSE: root mean squared error.

[c]EI: engagement index.

Predicting Survival Rate From 3 to 6 Months

When predicting app usage survival using the individual index of the EI from 3 to 6 months through Cox regression, the existing EI exhibited a log rank test result of $P<.05$. The results indicated a significant association between click depth and loyalty indices, while the RI showed no significance. The click depth index exhibited an HR of 0.49 with a P value <.001, which indicates that a higher click depth index is significantly associated with the reduced hazard, thus yielding better outcomes. Similarly, the LI showed an HR of 0.17 and a P value <.001, demonstrating a strong and significant association with reduced hazard. Conversely, the RI showed an HR of 1.30 with a P value of .41, indicating no significant association. All the log rank test results were statistically significant. The values of the existing EI are presented in Table 3.

Table 3. Results of the survival rate prediction using the existing engagement index.

	HR[a] (95% CI)	P value
Click depth index	0.49 (0.29-0.84)	<.001
Loyalty index	0.17 (0.09-0.31)	<.001
Recency index	1.30 (1.70-2.42)	.41

[a]HR: hazard ratio.

For the new EI, we conducted 3 Cox regressions based on the three devised menu combinations. MI 1 incorporates the menus intended for active app users, encompassing those necessitating self-logging. It specifically encompasses meal log, exercise log, messages sent to the app, reading content, and weight log. MI 2 comprises menus available in the app's free version. It consists of a meal log, exercise log, weight log, and step count login. MI 3 includes all available menus, such as meal log, exercise log, message sent to the app, reading content, weight log, and step count login. Hence, 3 new EIs were created (new EI1, new EI2, and new EI3), which includes each MI (MI1, MI2, and MI3).

New EI1 exhibited no significant association with the MI (HR 0.92; $P=.81$). However, it showed a strong and significant association with the LI (HR 0.28; $P<.001$). Furthermore, it showed a significant association with the RI (HR 0.48). Meanwhile, new EI2 exhibited a similar trend to new EI1, showing no significant association with the MI (HR 0.79; $P=.50$). However, it showed a strong and significant association with the LI (HR 0.3). Moreover, it exhibited a significant association with the RI (HR 0.47). Finally, new EI3 showed no significant association with the MI (HR 0.95; $P=.82$). However, it showed a significant and strong association with the LI (HR 0.26; $P<.001$), but it did not exhibit a significant association with the RI (HR 0.74; $P=.23$).

The MI did not exhibit a significant association with any of the new indices, whereas the LI showed a strong and significant association with all 3 indices. The RI was significantly associated with new EI1 and new EI2 but not with new EI3. All the log rank test results were significant for all the new indices. The values of the new EI are presented in Table 4.

XSL•FO
RenderX

Table 4. Survival analysis result for the 3 new engagement indices.

	New EI1	New EI2	New EI3
Menu abundancy index			
HR[a] (95% CI)	0.92 (0.48-1.77)	0.79 (0.40-1.56)	0.95 (0.57-1.58)
P value	.81	.50	.82
Loyalty index			
HR (95% CI)	0.28 (0.14-0.54)	0.31 (0.16-0.62)	0.26 (0.15-.46)
P value	<.001	<.001	<.001
Recency index			
HR (95% CI)	0.48 (0.28-0.81)	0.47 (0.30-0.75)	0.74 (0.45-1.22)
P value	<.001	<.001	.23

[a]HR: hazard ratio.

Discussion

Principal Findings

We evaluated the existing EI in a commercial health management app for long-term use and compared it with the new EI. We evaluated the new EI by first predicting the EI of the 6- to 9-month period based on the EI of the 3- to 6-month period through multiple linear regression and by predicting the survival rate using the EI of the 3- to 6-month period. In both predictions, the new EI exhibited better performance than the existing EI, although the difference was marginal. Moreover, when the RI, the index that best represents the long-term use, was applied in the new EI, a statistically significant difference increased compared with the RI in the existing EI.

Comparison With Previous Work

Retention has been inconsistently measured across studies in the aspect of mHealth. For instance, a previous study [25] defined retention as continuous use of the app for 6 months after the first use, specifically between 150 and 210 days. Another study measured retention based solely on the number of logs [26]. In addition, 1 study [27] measured retention through follow-up interviews conducted 6 months post intervention. These variations highlight the lack of a standardized retention strategy in mHealth research, posing a significant limitation as results may hinge on a single participant's interview response rather than reflecting overall trends and maintained use.

While the use of mHealth has the potential to enhance adherence to chronic disease management, research predominantly focuses on the assessment of the usability, feasibility, and acceptability of such apps rather than the direct measurement of adherence [28]. Similarly, studies addressing patient engagement in mHealth interventions in heart failure cases are often underreported and lacking consistency [29]. Moreover, a pressing need to evaluate user engagement in smartphone apps targeting other significant risk factors for cardiovascular disease, such as dietary behaviors, has been emphasized. Yang et al [30] identified 3 key issues concerning the measurement of adherence in mHealth programs. These include challenges in defining and measuring adherence, a tendency for adherence measurements to be grounded in empirical evidence or established theory, and the recognition that adherence is a multifaceted concept, thus requiring a comprehensive assessment rather than reliance on a 1-dimensional approach [30].

Although existing methodologies for measuring adherence to mHealth are limited, fewer measures of adherence with numerical results. Therefore, measurement using the EI has been considered a methodology that could be generally used and numerically measured. Taki et al [31] conducted a study that used the EI to measure engagement in the mHealth app. They used the click depth, loyalty, interaction, recency, and feedback indices and categorized the results into 3 groups to observe changes in the EI over time. However, they noted that some features were not measured by the EI, which may result in the underestimation of engagement of the participants. Similarly, White et al [32] used the EI to examine the demographic differences among 3 groups formed by the EI and used the reading, loyalty, interaction, recency, and feedback indices. However, they were unable to detect an association between the level of engagement and the duration of exclusive breastfeeding, which was possibly due to the limitations of the EI. Furthermore, Schepens Niemiec et al [33] used the loyalty, interaction, usability, and sentiment feedback indices with semistructured interviews to measure app engagement. They acknowledged that as only 4 indices were used, the statistical norm could not be determined to validate the evaluation of the mHealth apps. Despite its applicability to various programs offered by mHealth apps, EI exhibited similar limitations in each study, thereby raising uncertainties regarding its implications. However, despite the thorough investigation, with its simple characteristics, EI can effectively measure engagement in mHealth apps.

Reliance on postintervention surveys or interviews was common in other previous studies evaluating DTx engagement [10-14]. Alternatively, engagement with DTx was occasionally assessed simplistically, such as by marking the first date of a 28-day period without any data upload or by calculating the percentage of participants who completed follow-up at 8 weeks [34,35]. A review of various literature revealed that a more objective measure was evidently needed to evaluate patient engagement in DTx. Although valuable, manual interviews are difficult to replicate and are time consuming due to their labor-intensive

XSL•FO
RenderX

nature, involving multiple coordinators. Therefore, the proposal and evaluation of an EI for DTx could enhance the quality of research in this field.

Limitations

This study represents the inaugural attempt to evaluate the existing EI. While the effectiveness of the index has not yet been evaluated, we have established its reliability despite the comprehensive evaluation for potential upgrades. Furthermore, we were able to demonstrate the importance of log data from a research viewpoint as well as its objectivity, reproducibility, and potential for use to evaluate adherence to mHealth.

EI has a subjective nature in the metrics that may potentially introduce biases, which cannot be overcome despite the update of the index. Furthermore, although the existing EI comprises 7 indices, this evaluation focused only on 3 indices due to the specific characteristics of the app under scrutiny. Also, while the results may indicate that the newly developed EI outperforms the existing EI, the calculation of the existing EI may be simpler than the newly developed EI. However, we believe that this approach is more effective in predicting and representing long-term use.

Conclusions

This study evaluated the new EI within the commercial health management app by comparing it with the existing EI. Despite thorough evaluation using 2 approaches (forecasting the EI of the 6- to 9-month period based on the EI of the 3- to 6-month period through multiple linear regression and predicting survival rates based on the EI of the 3- to 6-month period), the new EI exhibited a slightly superior performance to the existing EI in both approaches. Although the existing EI appeared too simplistic for evaluating mHealth app adherence, we were able to demonstrate that it effectively reflected adherence without the need for complex calculations, similar to the new EI.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R and D Project through the Korea Health Industry Development Institute and funded by the Ministry of Health & Welfare, Republic of Korea (grant HI20C1058).

Data Availability

The data sets generated or analyzed during this study are not publicly available due to the need for institutional review board approval and privacy considerations but are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the conceptualization and reviewing and editing. YWT, JK, and YL handled the formal analysis. YWT, JK, and YL contributed to writing the original draft. JWL managed the resources.

Conflicts of Interest

None declared.

References

1. Wongvibulsin S, Habeos EE, Huynh PP, Xun H, Shan R, Porosnicu Rodriguez KA, et al. Digital health interventions for cardiac rehabilitation: systematic literature review. J Med Internet Res 2021;23(2):e18773 [FREE Full text] [doi: 10.2196/18773] [Medline: 33555259]
2. Choi E, Yoon E, Park M. Game-based digital therapeutics for children and adolescents: their therapeutic effects on mental health problems, the sustainability of the therapeutic effects and the transfer of cognitive functions. Front Psychiatry 2022;13:986687 [FREE Full text] [doi: 10.3389/fpsyt.2022.986687] [Medline: 36523871]
3. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. NPJ Digit Med 2023 Mar 10;6(1):38. [doi: 10.1038/s41746-023-00777-z] [Medline: 36899073]
4. Huh KY, Oh J, Lee S, Yu K. Clinical evaluation of digital therapeutics: present and future. Healthc Inform Res 2022;28(3):188-197 [FREE Full text] [doi: 10.4258/hir.2022.28.3.188] [Medline: 35982593]
5. Hong J, Wasden C, Han D. Introduction of digital therapeutics. Comput Methods Programs Biomed 2021;209:106319. [doi: 10.1016/j.cmpb.2021.106319] [Medline: 34364181]
6. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mhealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
7. Beckman J. 30 Amazing mobile health technology statistics [2023 Update]. TechReport. URL: https://techreport.com/statistics/mobile-healthcare-technology-statistics/ [accessed 2024-08-07]
8. Buneviciene I, Mekary RA, Smith TR, Onnela J, Bunevicius A. Can mHealth interventions improve quality of life of cancer patients? A systematic review and meta-analysis. Crit Rev Oncol Hematol 2021;157:103123 [FREE Full text] [doi: 10.1016/j.critrevonc.2020.103123] [Medline: 33190065]

XSL•FO
RenderX

9. Nouri R, R Niakan Kalhori S, Ghazisaeedi M, Marchand G, Yasini M. Criteria for assessing the quality of mHealth apps: a systematic review. J Am Med Inform Assoc 2018;25(8):1089-1098 [FREE Full text] [doi: 10.1093/jamia/ocy050] [Medline: 29788283]
10. Adu MD, Malabu UH, Malau-Aduli AE, Drovandi A, Malau-Aduli BS. User retention and engagement with a mobile app intervention to support self-management in Australians with type 1 or type 2 diabetes (my care hub): mixed methods study. JMIR Mhealth Uhealth 2020;8(6):e17802 [FREE Full text] [doi: 10.2196/17802] [Medline: 32525491]
11. Laurence C, Wispelwey E, Flickinger TE, Grabowski M, Waldman AL, Plews-Ogan E, et al. Development of positivelinks: a mobile phone app to promote linkage and retention in care for people with HIV. JMIR Form Res 2019;3(1):e11578 [FREE Full text] [doi: 10.2196/11578] [Medline: 30892269]
12. Alessa T, S Hawley M, Alsulamy N, de Witte L. Using a commercially available app for the self-management of hypertension: acceptance and usability study in Saudi Arabia. JMIR Mhealth Uhealth 2021;9(2):e24177 [FREE Full text] [doi: 10.2196/24177] [Medline: 33560237]
13. Waselewski ME, Flickinger TE, Canan C, Harrington W, Franklin T, Otero KN, et al. A mobile health app to support patients receiving medication-assisted treatment for opioid use disorder: development and feasibility study. JMIR Form Res 2021;5(2):e24561 [FREE Full text] [doi: 10.2196/24561] [Medline: 33620324]
14. Campbell ANC, Nunes EV, Matthews AG, Stitzer M, Miele GM, Polsky D, et al. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. Am J Psychiatry 2014;171(6):683-690 [FREE Full text] [doi: 10.1176/appi.ajp.2014.13081055] [Medline: 24700332]
15. Baek SY, Lee SB, Lee Y, Chung S, Choi C, Lee HJ, et al. Effects of mobile healthcare applications on the lifestyle of patients with breast cancer: a protocol for a randomized clinical trial. J Breast Cancer 2022;25(5):425-435 [FREE Full text] [doi: 10.4048/jbc.2022.25.e42] [Medline: 36314766]
16. Kim Y, Park I, Kim C, Yoon Y, Lim S, Yu C, et al. Lifestyle interventions after colorectal cancer surgery using a mobile digital device: a study protocol for a randomized controlled trial. Medicine (Baltimore) 2022;101(41):e31264 [FREE Full text] [doi: 10.1097/MD.0000000000031264] [Medline: 36254015]
17. Lee JH, Jeong JH, Ji W, Lee HJ, Lee Y, Jo M, et al. Comparative effectiveness of smartphone healthcare applications for improving quality of life in lung cancer patients: study protocol. BMC Pulm Med 2022;22(1):175 [FREE Full text] [doi: 10.1186/s12890-022-01970-8] [Medline: 35501757]
18. Chen J, Cade JE, Allman-Farinelli M. The most popular smartphone apps for weight loss: a quality assessment. JMIR Mhealth Uhealth 2015;3(4):e104 [FREE Full text] [doi: 10.2196/mhealth.4334] [Medline: 26678569]
19. Noom Inc. URL: https://www.noom.com/about-us/.[accessed [accessed 2024-03-18]
20. Michaelides A, Raby C, Wood M, Farr K, Toro-Ramos T. Weight loss efficacy of a novel mobile diabetes prevention program delivery platform with human coaching. BMJ Open Diabetes Res Care 2016;4(1):e000264 [FREE Full text] [doi: 10.1136/bmjdrc-2016-000264] [Medline: 27651911]
21. Mitchell ES, Fabry A, Ho AS, May CN, Baldwin M, Blanco P, et al. The impact of a digital weight loss intervention on health care resource utilization and costs compared between users and nonusers with overweight and obesity: retrospective analysis study. JMIR Mhealth Uhealth 2023;11:e47473 [FREE Full text] [doi: 10.2196/47473] [Medline: 37616049]
22. Keum J, Chung MJ, Kim Y, Ko H, Sung MJ, Jo JH, et al. Usefulness of smartphone apps for improving nutritional status of pancreatic cancer patients: randomized controlled trial. JMIR Mhealth Uhealth 2021;9(8):e21088 [FREE Full text] [doi: 10.2196/21088] [Medline: 34463630]
23. Peterson ET, Carrabis J. Measuring the immeasurable: visitor engagement. Verticalstudio. 2008. URL: https://www.verticalstudio.com/hs-fs/hub/74398/file-15425996-pdf/docs/web_analytics_demystified_and_nextstage_global_-_measuring_the_immeasurable_-_visitor_engagement.pdf.[accessed [accessed 2024-03-06]
24. Henry M, Judge G. Permutation entropy and information recovery in nonlinear dynamic economic time series. Econometrics 2019;7(1):10. [doi: 10.3390/econometrics7010010]
25. Mbotwa CH, Kazaura MR, Moen K, Leshabari MT, Metta E, Mmbaga EJ. Retention in an mHealth app aiming to promote the use of HIV pre-exposure prophylaxis among female sex workers in dar es salaam, Tanzania: prospective cohort study. JMIR Mhealth Uhealth 2023;11:e46853 [FREE Full text] [doi: 10.2196/46853] [Medline: 37855221]
26. Kim M, Yang J, Ahn W, Choi HJ. Machine learning analysis to identify digital behavioral phenotypes for engagement and health outcome efficacy of an mHealth intervention for obesity: randomized controlled trial. J Med Internet Res 2021;23(6):e27218 [FREE Full text] [doi: 10.2196/27218] [Medline: 34184991]
27. Lim JY, Kim Y, Yeo SM, Chae BJ, Yu J, Hwang JH. Feasibility and usability of a personalized mHealth app for self-management in the first year following breast cancer surgery. Health Informatics J 2023;29(1):14604582231156476 [FREE Full text] [doi: 10.1177/14604582231156476] [Medline: 36772832]
28. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015;17(2):e52 [FREE Full text] [doi: 10.2196/jmir.3951] [Medline: 25803266]
29. Madujibeya I, Lennie T, Aroh A, Chung ML, Moser D. Measures of engagement with mHealth interventions in patients with heart failure: scoping review. JMIR Mhealth Uhealth 2022;10(8):e35657 [FREE Full text] [doi: 10.2196/35657] [Medline: 35994345]

XSL•FO
RenderX

30. Yang Y, Boulton E, Todd C. Measurement of adherence to mHealth physical activity interventions and exploration of the factors that affect the adherence: scoping review and proposed framework. J Med Internet Res 2022;24(6):e30817 [FREE Full text] [doi: 10.2196/30817] [Medline: 35675111]
31. Taki S, Lymer S, Russell CG, Campbell K, Laws R, Ong K, et al. Assessing user engagement of an mHealth intervention: development and implementation of the growing healthy app engagement index. JMIR Mhealth Uhealth 2017;5(6):e89 [FREE Full text] [doi: 10.2196/mhealth.7236] [Medline: 28663164]
32. White BK, Burns SK, Giglia RC, Dhaliwal SS, Scott JA. Measuring user engagement with a socially connected, gamified health promotion mobile app. Int J Environ Res Public Health 2022;19(9):5626 [FREE Full text] [doi: 10.3390/ijerph19095626] [Medline: 35565015]
33. Schepens Niemiec SL, Wagas R, Vigen CL, Blanchard J, Barber SJ, Schoenhals A. Preliminary user evaluation of a physical activity smartphone app for older adults. Health Policy Technol 2022;11(3):100639 [FREE Full text] [doi: 10.1016/j.hlpt.2022.100639] [Medline: 36213682]
34. Lin Y, Chen S, Lin P, Tai A, Pan Y, Hsieh C, et al. Assessing user retention of a mobile app: survival analysis. JMIR Mhealth Uhealth 2020;8(11):e16309 [FREE Full text] [doi: 10.2196/16309] [Medline: 33242023]
35. Pagoto S, Tulu B, Agu E, Waring ME, Oleski JL, Jake-Schoffman DE. Using the habit app for weight loss problem solving: development and feasibility study. JMIR Mhealth Uhealth 2018;6(6):e145 [FREE Full text] [doi: 10.2196/mhealth.9801] [Medline: 29925496]

Abbreviations

DTx: digital therapeutics
EI: engagement index
HR: hazard ratio
LI: loyalty index
mHealth: mobile health
MI: menu abundancy index
MSE: mean squared error
PE: permutation entropy
RCT: randomized controlled trial
RI: recency index
RMSE: root mean squared error

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 12.04.24; peer-reviewed by A Wani, D Ghosh; comments to author 09.05.24; revised version received 02.07.24; accepted 30.07.24; published 09.09.24.

Please cite as:
Tak YW, Lee JW, Kim J, Lee Y
Predicting Long-Term Engagement in mHealth Apps: Comparative Study of Engagement Indices
J Med Internet Res 2024;26:e59444
URL: https://www.jmir.org/2024/1/e59444
doi:10.2196/59444
PMID:39250192

©Yae Won Tak, Jong Won Lee, Junetae Kim, Yura Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Impact of a Game (Inner Dragon) on User Engagement Within a Leading Smartphone App for Smoking Cessation: Randomized Controlled Trial

Justin S White[1,2,3], PhD; Séverine Toussaert[4], PhD; Bethany R Raiff[5], PhD; Marie K Salem[2,6], MPH; Amy Yunyu Chiang[2], PhD; David Crane[7], PhD; Edward Warrender[7]; Courtney R Lyles[8], PhD; Lorien C Abroms[9], SCD; J Lee Westmaas[10*], PhD; Johannes Thrul[11,12,13*], PhD

[1]Department of Health Law, Policy and Management, Boston University School of Public Health, Boston, MA, United States

[2]Philip R Lee Institute for Health Policy Studies, University of California, San Francisco, San Francisco, CA, United States

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States

[4]Department of Economics, University of Oxford, Oxford, United Kingdom

[5]Department of Psychology, Rowan University, Glassboro, NJ, United States

[6]Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, United States

[7]23 Limited, London, United Kingdom

[8]Department of Public Health Sciences, University of California, Davis, Davis, CA, United States

[9]Department of Prevention and Community Health, Milken Institute School of Public Health, The George Washington University, Washington, DC, United States

[10]Population Science, American Cancer Society, Atlanta, GA, United States

[11]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[12]Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, United States

[13]Centre for Alcohol Policy Research, La Trobe University, Melbourne, Australia

[*]these authors contributed equally

Corresponding Author:
Justin S White, PhD
Department of Health Law, Policy and Management
Boston University School of Public Health
Talbot Building - 249W
715 Albany Street
Boston, MA, 02118
United States
Phone: 1 617 358 1916
Email: juswhite@bu.edu

Abstract

Background: Smartphone apps are a convenient, low-cost approach to delivering smoking cessation support to large numbers of individuals. Yet, the apps are susceptible to low rates of user engagement and retention.

Objective: This study aims to test the effects of a new game module (called Inner Dragon) integrated into Smoke Free (23 Limited), a leading smoking cessation app with established efficacy. The primary outcomes measured user engagement with the app.

Methods: A 2-arm, parallel-group, randomized controlled trial was conducted in the United States with an 8-week follow-up. Adult individuals who smoked ≥1 cigarettes daily and planned to quit smoking within 7 days were recruited and randomized (N=500), with equal allocation. Both groups received free access to the original Smoke Free app with “core” features of its smoking cessation program (eg, a diary and craving log). The treated group received additional access to the integrated Inner Dragon game that incorporated several game mechanics designed to increase user engagement. User engagement outcomes were the number of unique app sessions, average minutes per session, days with a session, and program adherence. Self-reported and verified smoking abstinence and app satisfaction were also assessed. The main analysis estimated the intention-to-treat effect of

XSL•FO
RenderX

access to Inner Dragon on each outcome. Further analyses assessed effect modification by participant characteristics and the association of intensity of game use with program adherence and abstinence.

Results: Overall, user engagement was greater for treated versus control participants: they had 5.3 more sessions of Smoke Free (mean 29.6, SD 36.5 sessions vs mean 24.3, SD 37.9 sessions; *P*=.06), 0.8 more minutes per session (mean 6.9, SD 5.4 min vs mean 6.1, SD 5.2 min; *P*=.047), and 3.4 more days with a session (mean 14.3, SD 15.3 days vs mean 11.9, SD 14.3 days; *P*=.03). Program adherence, based on the number of times core features of the original Smoke Free app were used, was higher for treated versus control participants (mean 29.4, SD 41.3 times vs mean 22.6, SD 35.6 times; *P*=.03). Self-reported 7-day and 30-day point-prevalence abstinence and verified 7-day point-prevalence abstinence at 8 weeks did not significantly differ by study group. The mean repeated 1-day prevalence of quitting was higher among the treated group versus the control group (mean 17.3%, SD 25.6 vs mean 12.4%, SD 21.3; *P*=.01). App satisfaction and the motivation to (stay) quit did not differ by study group. Higher intensity of game use was associated with increased program adherence and self-reported abstinence.

Conclusions: Findings suggest that the Inner Dragon game increased user engagement and program adherence. Additional refinements to the game design may clarify whether the game increases abstinence rates. Overall, it is feasible to deploy games and gamification to enhance user engagement in existing smoking cessation interventions.

Trial Registration: ClinicalTrials.gov NCT05227027; https://clinicaltrials.gov/study/NCT05227027

(J Med Internet Res 2024;26:e57839) doi:10.2196/57839

KEYWORDS

smoking cessation; mobile app; games for health; gamification; engagement; randomized controlled trial; mobile phone

Introduction

Background

Cigarette smoking remains a leading cause of preventable death and illness in the United States [1]. Although various clinic- and phone-based smoking cessation treatments are effective, the most effective cessation treatment, which combines counseling with medication, is used by fewer than 5% of individuals attempting to quit smoking [1-3]. Most individuals attempting to quit smoking do so without any assistance, resulting in success rates of <5% [4].

The advent of smartphone-delivered smoking cessation support offers a compelling alternative to traditional clinical treatments, given its potential reach and convenience. Smoking cessation apps, in particular, have gained popularity due to their low cost and just-in-time support.

Research shows that English-language smoking cessation apps have been downloaded 33 million times as of 2020 [5]. The widespread ownership of smartphones has contributed to the extensive reach of cessation apps. As of 2023, 90% of US adults, including 84% of Black adults, 91% of Hispanic adults, 79% of adults in low-income households (<US $30,000 per year), and 89% of adults aged between 50 and 64 years, owned a smartphone [6]. Consequently, smartphone-based interventions hold promise to reduce the substantial burden of tobacco-related disease and mortality.

While smoking cessation apps offer the advantage of providing timely assistance, only a small proportion of the numerous apps available in app stores have undergone rigorous testing. A systematic review conducted up until 2019 indicated that 11 smoking cessation apps had been subjected to randomized controlled trials (RCTs), although only 3 were beyond the pilot or feasibility stage [7]. Since then, there have been additional RCTs, although the literature is still emerging. Early-stage evaluations and the full-scale trials have indicated that smartphone cessation apps tend to yield intention-to-treat abstinence rates from 12% to 32% (usually at 1-2 months), with substantial variation due to differences in measures, the inclusion of bioverification, and other design factors [8-20]. However, the translation of these early-efficacy rates into real-world effectiveness remains largely uncertain, particularly regarding app features that enhance user engagement and contribute to sustained abstinence [21].

A major contributor to the modest efficacy rates for many smoking cessation apps is low user engagement and retention [22-24]. Systematic data for smartphone cessation apps are lacking, although evidence suggests a pattern of low retention. Three-quarters of mobile health apps are opened <10 times [25].

To address this gap, researchers have explored the use of serious games and gamification techniques as potential strategies to enhance user engagement and motivation [26-29]. Serious games have a primary purpose other than entertainment, such as health promotion. Gamification is a related motivational tool that uses nonmonetary rewards to make nongame activities fun or challenging. By increasing user engagement, games and gamification may increase exposure and adherence to a smoking cessation program and, in turn, increase the likelihood of quitting [30]. Games may have particular appeal to individuals who smoke, three-quarters of whom play video games, according to 1 survey [31]. Many smokers also express belief in the motivating potential of game-based approaches for smoking cessation [31,32].

Several stand-alone smartphone-based games for smoking cessation, such as Cigbreak, Crush the Crave, Inspired, QuitIT, Quittr, and Tobbstop, have been developed [12,19,20,33-39]. A few of these games have also been tested in RCTs. While users generally reported satisfaction with the gamified apps, most studies have not observed significant increases in smoking abstinence [12,19,20,38]. The variability in user engagement and retention across these apps further complicates the

XSL•FO
RenderX

assessment of their overall effectiveness, with each study adopting different measurement approaches.

This Study

We conducted an RCT to test the efficacy of a novel game intervention called Inner Dragon. Whereas prior games have been stand-alone products; Inner Dragon was integrated into the Smoke Free app (23 Limited), providing a unique opportunity to estimate the incremental benefit of embedding a game in a high-quality smoking cessation app. Smoke Free has been found to increase continuous abstinence at 6 months (12.7% intervention vs 7% comparator) on a per-protocol basis in a pragmatic RCT of 3000 participants [15]. Smoke Free is also one of the most downloaded smoking cessation apps in the Apple (Apple Inc) and Android (Google LLC) stores, with more than 800,000 downloads per year and 7 million downloads to date [14,40].

Inner Dragon is a multifaceted game iteratively developed by our team that uses virtual pet retention mechanics with some social features and customization options to promote user engagement and retention [41]. In the Inner Dragon game, users care for an evolving, customizable pet dragon, whose growth reflects the user's own progress toward quitting smoking. Points are provided for engaging with game elements and nongame educational content. The game includes elements aimed at coping with cravings and offers asynchronous interaction with other users' dragons. In a single-arm feasibility trial, Inner Dragon was found to have high user satisfaction and generally high user engagement [41]. This study aimed to assess whether the novel Inner Dragon game integration increased user engagement among users of a leading smartphone app for smoking cessation.

Methods

Trial Design

We conducted a parallel-group RCT with equal allocation to 2 groups. The treated group received free access to the Smoke Free app with the integrated Inner Dragon game. The control group received free access to the Smoke Free app without the game embedded in the app. The study protocol is provided in the supporting information with a CONSORT (Consolidated Standards of Reporting Trials) guideline checklist (Multimedia Appendix 1 [6,42-62] and Multimedia Appendix 2) [63]. The protocol was preregistered at ClinicalTrials.gov (NCT05227027). No changes to the prespecified trial design and methods were made after trial commencement. Multimedia Appendix 1 provides the study protocol.

Recruitment and Participants

Individuals were recruited from the US-based population of general users of the Smoke Free smartphone app who downloaded the app from the Apple or Android app store. Users were invited to participate in the trial during the initial onboarding screens within the Smoke Free app via an on-screen recruitment message, "Interested in a research study to test the latest version of Smoke Free?" If a user clicked "Yes, learn more," they were taken to a screening questionnaire to assess eligibility and a consent form in Qualtrics (Qualtrics International Inc).

Inclusion criteria were individuals aged ≥18 years living in the United States who currently smoked at least 1 cigarette per day; had downloaded the Smoke Free app; planned to quit smoking within 7 days; and were able to speak, read, and write in English. A quota of at least 100 individuals aged ≥50 years was programmed into Qualtrics to ensure the study was representative with respect to age and to enable an assessment of whether older individuals, who have a higher smoking prevalence and may be less proficient with technology or less drawn to games, had a differential response to the intervention [31,64,65]. There were no exclusion criteria.

Randomization and Masking

Eligible, consenting participants were randomized within strata with equal allocation to one of two groups: (1) a control group that received free access to the original Smoke Free app or (2) a treated group that received free access to the Smoke Free app with the integrated Inner Dragon game. Allocation, concealed from investigators, was performed through the Qualtrics randomizer and transmitted to the Smoke Free database using the Qualtrics application programming interface. Participants were then automatically given access to the assigned intervention within the Smoke Free app. Randomization was stratified by age (<50 and ≥50 years) and smoking intensity (<5 and ≥5 cigarettes per day, on average), to facilitate the assessment of effect modification.

The assessor of the salivary cotinine test was blinded to the participants' group assignment; other aspects of the trial were not blinded.

Procedures

Overview

As part of the screening questionnaire, participants selected an initial quit date within the next 7 days. Participants were invited to use the Smoke Free app in whatever way they liked for 56 days after their planned quit date. Participants were asked via email to complete a baseline questionnaire in Qualtrics after consenting to participate in the study. The baseline questionnaire included additional questions about the person's demographic and smoking history. A follow-up questionnaire in Qualtrics was performed 8 study weeks (56 days) after the person's initial quit date (56-63 days after screening and consent depending on each participant's selected quit date). An invitation to complete the follow-up questionnaire was sent by SMS text messaging. Those who reported having abstained for 7 days in the follow-up questionnaire were asked for a mailing address within the follow-up questionnaire. We then mailed them a salivary cotinine test kit and written and pictorial test instructions via 2-day mail. Following completion of the follow-up questionnaire, an initial SMS text message explained that the saliva test would be shipped shortly, and a second message sent 2 days later provided a Qualtrics link with which participants were asked to upload photos of the cotinine test results. Participants received 3 reminders via SMS text messaging and 2 via email to complete the follow-up questionnaire, and they received 3 reminders via SMS text messaging to complete the

XSL•FO
RenderX

saliva testing. In-app reminders were disabled so as not to conflict with the study procedures.

Control Group Intervention

Participants in the control group received free access to an educational intervention consisting of the original Smoke Free app, which has previous per-protocol trial evidence of continuous abstinence at 6 months [15]. The app leverages behavior change techniques that are effective in face-to-face behavioral support programs [66]. "Core" features of the original Smoke Free app include the following: (1) a calculator that tracks the total amount of money saved by not smoking; (2) a calendar that tracks the time elapsed since the user quit smoking; (3) a scoreboard that awards badges to users for not smoking; (4) progress indicators that inform users about health improvements that the user can expect because they started their quit attempt; (5) daily missions that assign evidence-based tasks to help users avoid and resist urges to smoke (eg, noting situations when the person smokes and plans for handling those situations without smoking); (6) a diary and craving log that track the frequency, strength, and location of cravings to smoke; and (7) a text-based chatbot that delivers quitting guidance in a friendly, conversational tone. In previous randomized trials, a full version of the app with a chatbot was shown to increase user engagement compared with a reduced version of the app [67], and the daily missions were shown to increase user retention and self-reported smoking abstinence at 3 months [14]. The current core version of the Smoke Free app does not contain any additional gamification features.

Treated Group Intervention

Participants in the treated group received free access to core features of the original Smoke Free app as well as the integrated Inner Dragon game module. Inner Dragon uses traditional virtual pet retention with some social features and many options for customization and personalization. Virtual pets have been popular with consumers because they can foster bonding and companionship [68].

More specifically, Inner Dragon includes several game mechanics designed to increase user engagement and retention. First, the user maintains a pet dragon that hatches on the user's quit day and evolves every 7 days to unlock new attributes and powers. The dragon acts as a virtual avatar that represents the user's progress. Second, the user earns experience points by engaging in selected activities, including those in the game (eg, playing a minigame or feeding the dragon) and nongame features in the original Smoke Free app (eg, completing a mission or logging a diary entry). Rewards for using core app features in the original app were designed to improve adherence to the smoking cessation program. The experience points unlock gifts and cosmetics for the dragon, directly rewarding frequent and consistent use of the app. The user can customize their dragon (eg, wing shape or clothing accessories) throughout the quit experience by steadily unlocking features. Third, the Inner Dragon home screen has "care meters" that users must work on to keep from falling very low: calmness, nutrition, hygiene, and energy. Engaging with the dragon in various ways increases the readings on meters. For example, petting the dragon increases calmness, and feeding the dragon increases nutrition. Caring for and interacting with a virtual pet through the care meters can foster a bond with the pet and motivate users to return to the game regularly [68]. Fourth, the game provides tools for the user to better cope with cravings: a dragon-led breathing exercise to provide calmness and a memory minigame as a distraction. Fifth, a user can asynchronously interact with other users' dragons in a "dragon park" by (1) viewing their profile and progress and (2) sending and receiving motivational messages and emojis from a preset menu. We hypothesize that these game mechanics increase users' engagement with the app and, subsequently, their chance of quitting successfully.

Inner Dragon's game design was informed by principles from the fields of psychology and behavioral economics [41]. The avatar provided salient, visual feedback with endogenous value (tied closely to the user's motivation to quit) that may sustain and enhance motivation to quit. Self-determination theory predicts that this type of feedback is highly intrinsically motivating [69]. Furthermore, the user may identify with the digital pet as an avatar, and this may cultivate a digital therapeutic alliance with the game and app, for example, by creating a bond with the dragon and increasing the user's confidence to succeed [42,70]. The use of frequent, salient in-game rewards was designed to counter the behavioral economic constructs of present bias and inattention to app use [71,72]. The design further included evidence-based practices from contingency management, such as the use of escalating in-game rewards for abstinence, with a reset point for lapses and sustained abstinence (by harnessing regret and loss aversion) [73]. The use of surprise gifts provided a variable reward structure designed to boost engagement and novelty [74-76]. The asynchronous interactions with other users in the dragon park provided opportunities for limited social support and social comparisons that may motivate the user to exert more effort in the quit attempt [77,78].

The various game mechanics and elements were designed to appeal to users with different motivations. For example, Yee [79] identified 3 main components of player motivation: achievement (advancement and competition), social (socializing and relationships), and immersion (discovery, customization, and escapism). The experience-point system may appeal to achievers; the interaction with other players and a sense of connection with the dragon may appeal to socializers, and the rich opportunities for dragon customization and distraction games may appeal to players seeking immersion.

A more thorough description of the design and development of Inner Dragon is described elsewhere [41].

Measures

Outcome Measures

The study focused on outcomes within the domains of user engagement, abstinence from smoking, and user satisfaction and motivation.

User engagement measures included the two *primary outcome measures* of (1) the total number of unique app sessions from enrollment through 8 weeks after the user's initial quit date, measured as the number of times the app was opened; and (2) the mean duration of app sessions, in minutes, through 8 weeks

XSL•FO
RenderX

after the quit date; The *secondary outcome measures* of (3) the number of unique days with ≥1 app session, and (4) the proportion entering Inner Dragon, and the *tertiary outcome measures* of (5) program adherence, measured as the total number of times selected "core" features of the original Smoke Free app were used (reported a craving, recorded a diary entry, completed a mission, read a tip, and used the chatbot); and (6) the number of times using selected game features (opening gifts, breathing exercise, cleaning the dragon, feeding the dragon, memory minigame, using the customization menu to change appearance, and reading the dragon instruction guide). All engagement outcomes were passively collected by the app. A new session was defined as opening the app after at least 30 minutes of inactivity.

Smoking abstinence measures included the *secondary outcome measures* of (1) the proportion of participants who reported abstaining during the past 7 days at the 2-month follow-up assessment (self-reported 7-day point-prevalence abstinence), (2) the proportion of participants who reported abstaining during the past 30 days at the 2-month follow-up assessment (self-reported 30-day point-prevalence abstinence), and the *tertiary outcome measures* of (3) biochemically verified 7-day point-prevalence abstinence at the 2-month follow-up assessment, obtained from uploaded results from a self-administered salivary cotinine test (Alere iScreen Oral Fluid Device; Abbott Laboratories); and (4) repeated 1-day point-prevalence smoking abstinence, measured as the mean proportion of days each participant self-reported having abstained in the last 24 hours, collected via a pop-up box that appeared the first time each day a participant opened the app throughout the 8-week, postquit date period. Self-reported 7-day point-prevalence abstinence followed the Russell Standard, allowing for <5 slips [80]. Those who reported using nicotine replacement therapy (NRT) or electronic nicotine delivery systems (ENDSs) within the prior 7 days were coded in the main analysis as abstinent for the verified measures if no cigarettes were used [80,81].

User satisfaction and motivation measures included the *secondary outcome measures* of (1) satisfaction with the Smoke Free app ("I liked using the Smoke Free app"), reported on a 5-point Likert scale ranging from "not at all" (score=1) to "extremely" (score=5); (2) satisfaction with the Inner Dragon game ("I liked the dragon game"), reported on the same 5-point Likert scale, and the *tertiary outcome measures* of (3) rating of whether the person would recommend the assigned app to a friend on a 5-point Likert scale; (4) motivation to (stay) quit at the 8-week follow-up, reported on a 10-point scale from 0 (not at all motivated) to 10 (very motivated); and (5) digital therapeutic alliance, measured from the bonding and confidence subscales of the Mobile Agnew Relationship Measure, reported as a 16-point ordinal measure [42,70].

Covariates

The screening questionnaire collected demographic information on age, gender (woman, man, non-binary or other), race and ethnicity (recoded to non-Hispanic White, non-Hispanic Black, Hispanic, or other), household income category (US $0-US $19,999, US $20,000-US $39,999, US $40,000-US $59,999, US $60,000-US $79,999, US $80,000-US $99,999, and US $100,000 and more), and educational attainment (high-school diploma and less, some college or technical school, bachelor's or associate degree, and graduate degree). Smoking characteristics at the screening included mean cigarettes per day, Fagerström Test for Nicotine Dependence (score of 0-10, where 10 is highly nicotine dependent) [82], number of past quit attempts, years since initiated, use of ENDS in last 30 days (yes or no), and use of NRT in last 30 days (yes or no). Baseline information also included frequency of video game use (not at all, less than once a month, at least monthly but not weekly, at least weekly but not every day, and every day).

Sample Size

Power calculations were based on an anticipated sample size of 500, with an assumed 70% retention (as observed in other app-based smoking cessation trials) [12,20]. We further anticipated a mean number of app sessions of 30 (SD 45.0) per participant in the control group, based on data from past Smoke Free users provided to the study team by Smoke Free. A sample size of 500 participants with 70% retention was estimated to detect a between-group difference in app sessions, one of the primary outcomes, of 13.5 at 80% power (α=.05).

Data Analysis

Main Analysis

The main analyses were conducted on an intention-to-treat basis using all randomized participants. For those outcomes measured continuously, including the 2 primary outcomes, we calculated the crude (unadjusted) difference in outcomes between each study group, using 1-tailed *t* tests of the difference in means to assess statistical significance at the 5% level. For those outcomes measured categorically, we calculated the mean difference in proportions corresponding to the risk difference and used a similar approach to assess statistical significance. For the abstinence outcomes, the intention-to-treat approach implied that those lost to follow-up were assumed to have resumed smoking ("missing=smoking").

Outcome Trends

Next, we assessed visual changes over time in the number of users with a Smoke Free session, a key engagement outcome, and self-reported 1-day point-prevalence abstinence on an intention-to-treat basis (missing=smoking) and using complete cases. For each outcome, we plotted the mean for each study group by study day.

Intensity of Game Use

We assessed whether the intensity of game use was associated with more distal outcomes: total user engagement (total number of sessions), program adherence, and self-reported abstinence (at follow-up and repeated daily, assuming missing=smoking). This analysis, conducted among participants in the treated group only, entailed plotting estimates from local linear regressions of the number of sessions with game use on each of these outcomes, following the nonparametric approach proposed by Calonico et al [83]. The estimates included mean squared error-optimal bandwidth, 30 evaluation points. and a

XSL•FO
RenderX

heteroskedasticity-robust nearest neighbor variance estimator with 3 neighbors.

Sensitivity and Exploratory Analyses

We examined subgroup differences across study groups for the primary engagement outcomes, in which we evaluated heterogeneity by estimating a stratified model for each covariate.

We conducted covariate-adjusted regression analyses to assess the robustness of our findings. We estimated the differences in each outcome by study group using linear regressions for continuous outcomes and logistic regressions for binary outcomes. The models adjusted for a range of prespecified potential confounders (described in the Covariates section earlier), including characteristics of participants' sociodemographic background, smoking history, and frequency of video game use. Coefficients from the logistic regressions were expressed as risk differences, with an interpretation as a change in percentage points.

We repeated our analysis of verified abstinence—coding users of NRT or ENDS at follow-up as having not abstained. We also repeated our abstinence analysis using a complete case analysis that included only those participants who were successfully followed up. Whereas missing=smoking can lead to downward bias if loss to follow-up occurs for reasons other than relapse to smoking, complete case analysis may lead to upward bias if participants refuse to engage with follow-up because they have resumed smoking. As a further sensitivity analysis of the abstinence outcome, we estimated the treatment effect for self-reported 7-day point-prevalence abstinence from a pattern-mixture model that varied the informative missingness odds ratio between the outcome and an indicator for missingness, adjusting for covariates [84,85].

Adverse Events

We tabulated the number of serious adverse events by type for each study group, as reported in an open-ended text field in the follow-up questionnaire.

Analyses were conducted in Stata (version 18.0; StataCorp).

Ethical Considerations

Ethics approval was obtained from the University of California, San Francisco institutional review board (19-29335). All participants were informed about the study details and procedures before providing their consent electronically. Participation was strictly voluntary, and participants were informed about their right to withdraw from the study at any point without any consequences. The collected data were coded and deidentified. Participants' identities were kept confidential and stored securely in password-protected files. Coded identifiers were given to participants after obtaining informed consent. No participants' personal identifiers were used in analyses or report writing. Participants in both groups received US $20 for completing the baseline questionnaire, US $40 for completing the follow-up questionnaire, and US $40 for completing the saliva test, if requested. All payments were provided, via email link, as a choice of a Visa (Visa Inc) prepaid card or one of hundreds of gift cards through the Tango digital payment platform (Tango Card Inc).

Results

Sample Characteristics

Figure 1 depicts the study flow diagram. We screened 1181 new Smoke Free users from February 9, 2022, through March 16, 2022, to reach our goal of enrolling 500 eligible, consenting individuals for the RCT. Of the randomized participants, 95.8% (479/500; treated: 238/251, 94.8%, and control: 241/249, 96.8%) opened Smoke Free after completing the screening questionnaire, thereby completing the onboarding process within Smoke Free. Individuals who did not finish onboarding did not return to the app after the screening survey and did not learn their allocation group. The denominator used for the primary analyses was 479.

XSL•FO
RenderX

Figure 1. Participant flow diagram.

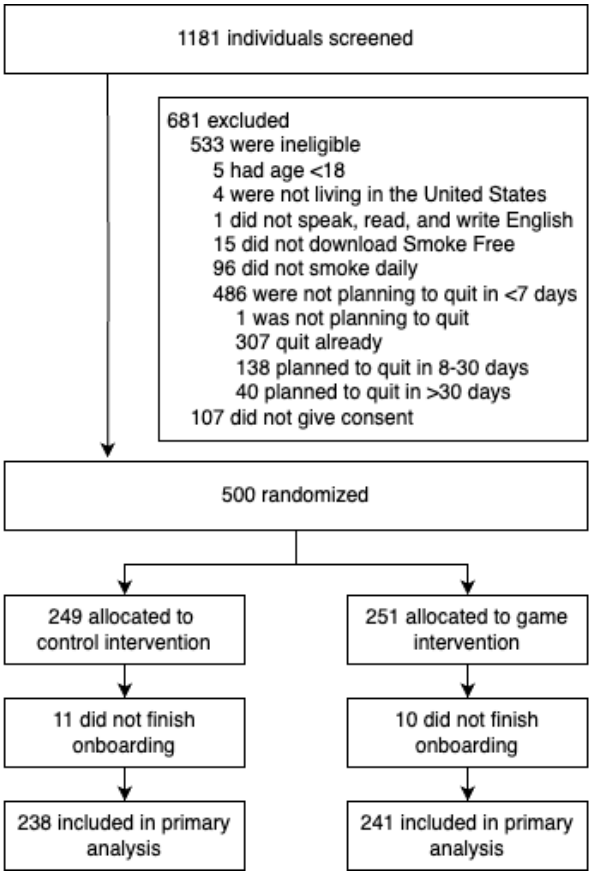

There was a 61% (292/479) response rate for the 8-week follow-up questionnaire (treated: 154/241, 63.9% and control: 138/238, 58%), and a 71% (137/193) response rate for the saliva test (among those completing the follow-up questionnaire and reporting having quit within the prior 7 days), including 74% (73/99) of invited control participants and 68% (64/94) of invited treated participants. Nonresponse for the follow-up questionnaire and saliva test was unrelated to the study group assignment, according to *t* tests of the difference in means. However, response rates for the questionnaire and saliva test increased with the number of app sessions, and the response rate for the saliva test was higher for invited non-Hispanic Black participants (23/28, 82%) and lower for invited non-Hispanic White participants (93/136, 68.4%).

Demographic and smoking characteristics were balanced across study groups (Table 1; Table S1 in Multimedia Appendix 3 [7,8,10,12-20,86-95]). Participants were drawn from 48 states across the United States (Midwest: 126/479, 26.3%, Northeast: 90/479, 18.8%, South: 160/479, 33.4%, and West: 103/479, 21.5%). The median age of participants was 36 (IQR 29-47) years, with 19.8% (95/479) of participants aged ≥50 years due to the recruitment quota. Three-quarters (358/479, 74.7%) of participants identified as women. Most (368/479, 76.8%) were non-Hispanic White, with 10% (48/479) non-Hispanic Black or African American and 5.8% (28/479) Hispanic. Nearly 62.6% (300/479) had a household income below US $60,000, and 41.3% (198/479) had a household income <150% of the federal poverty level at the time. Most (277/479, 57.8%) did not hold a college degree. For smoking and gaming characteristics, participants were mostly moderate-to-heavy smokers (median of 15.0, IQR 10-20, cigarettes per day), had previously attempted to quit (median of 4, IQR 2-9, past attempts), and varied in terms of video game use during the prior 30 days (not at all: 116/479, 24.2% and every day: 162/479, 33.8%).

XSL•FO
RenderX

Table 1. Characteristics at screening of participants enrolled in the Inner Dragon trial, overall and by study group (N=479).

	Control (n=238)	Treated (n=241)	Total (N=479)
Panel A: demographics			
Age (y), median (IQR)	35 (28-47)	38 (29-47)	36 (29-47)
Gender, n (%)			
Women	181 (76.1)	177 (73.4)	358 (74.7)
Men	55 (23.1)	59 (24.5)	114 (23.8)
Nonbinary or other	2 (0.8)	5 (2.1)	7 (1.5)
Race and ethnicity, n (%)			
Hispanic	9 (3.8)	19 (7.9)	28 (5.8)
Non-Hispanic Black	26 (10.9)	22 (9.1)	48 (10)
Non-Hispanic White	182 (76.5)	186 (77.2)	368 (76.8)
Other	21 (8.8)	14 (5.8)	35 (7.3)
Household income (US $), n (%)			
<20,000	35 (14.7)	35 (14.5)	70 (14.6)
20,000-39,999	65 (27.3)	63 (26.1)	128 (26.7)
40,000-59,999	51 (21.4)	51 (21.2)	102 (21.3)
60,000-79,999	24 (10,1)	32 (13.3)	56 (11.7)
80,000-99,999	18 (7.6)	21 (8.7)	39 (8.1)
≥100,000	45 (18.9)	39 (16.2)	84 (17.5)
Education, n (%)			
High school diploma or less	55 (23.1)	45 (18.8)	100 (20.9)
Some college or technical school	84 (35.3)	93 (38.6)	177 (37)
Bachelor's or associate degree	68 (28.6)	83 (34.4)	151 (31.5)
Graduate degree	31 (13)	20 (8.3)	51 (10.6)
Panel B: smoking characteristics			
Cigarettes per day, median (IQR)	15 (10-20)	15 (10-20)	15 (10-20)
Fagerström Test, median (IQR)	5 (3-7)	5 (4-7)	5 (4-7)
Past quit attempts, median (IQR)	4 (2-8)	4 (2-9)	4 (2-9)
Years since initiated, median (IQR)	16 (10-28)	19 (11-29)	18 (10-29)
Used ENDS[a] in last 30 days, n (%)	61 (25.6)	57 (23.7)	118 (24.6)
Used NRT[b] in last 30 days, n (%)	43 (18.1)	56 (23.2)	99 (20.7)
Panel C: other			
Frequency of playing video games, n (%)			
Not at all	59 (24.8)	57 (23.7)	116 (24.2)
Less than once a month	25 (10.5)	36 (14.9)	61 (12.7)
At least monthly but not weekly	25 (10.5)	26 (10.8)	51 (10.6)
At least weekly but not every day	40 (16.8)	49 (20.3)	89 (18.6)
Every day	89 (37.4)	73 (30.3)	162 (33.8)

[a]ENDS: electronic nicotine delivery system.

[b]NRT: nicotine replacement therapy.

XSL•FO
RenderX

Main Intervention Effects

User Engagement

Compared with control participants, treated participants had 5.3 more sessions of Smoke Free app use during the 8 study weeks, although this difference in the primary outcome was not statistically significant (mean 29.6, SD 36.5 sessions vs mean 24.3, SD 37.9 sessions; *P*=.06; Table 2). Treated participants used the app for more minutes per session than control participants (mean 6.9, SD 5.4 min vs 6.2, SD 5.2 min; *P*=.047 for the primary outcome) and had more days with an app session (mean 14.3, SD 15.3 days vs 13.3, SD 11.9 days; *P*=.03 for the secondary outcome). Overall, treated participants used the Inner Dragon game for an average of 7.1 (SD 14.9) days, with at least some use of each of the game features (Table S2 in Multimedia Appendix 3).

Program adherence, as measured by the number of times using certain core features in the original Smoke Free app, was greater for treated participants than control participants (mean 29.4, SD 41.3 times vs mean 22.6, SD 35.6 times; *P*=.03 for tertiary outcome). The largest increase in core features for the treated group occurred for the number of diary entries completed (mean 12.1, SD 15.9 vs mean 9.2, SD 13.2; *P*=.01 for tertiary outcome), although use of all core features increased (Table 2).

Smoking Abstinence

Self-reported 7-day point-prevalence abstinence on an intention-to-treat basis (assuming missing=smoking) was 39% (94/241) for the treated group and 42.4% (101/238) for the control group (difference –3.4 percentage points; *P*=.45 for secondary outcome). The difference widened when considering complete cases only, such that 61% (94/154) abstained in the treated group and 73.2% (101/138) in the control group (difference –12.1 percentage points; *P*=.03).

Verified 7-day point-prevalence abstinence at 2 months similarly showed no advantage for the treated group, such that the intention-to-treat estimates were 19.5% (47/241) in the treated group and 22.7% (54/238) in the control group (difference –3.2 percentage points; *P*=.39). This included 8.3% (20/241) of participants in the treated group and 8% (19/238) of participants in the control group who were coded abstinent because they reported having used NRT or ENDS within the prior 7 days at the 2-month follow-up. Self-reported 30-day point-prevalence abstinence at 2 months was similar among the treated group (54/241, 22.4%) and the control group (50/238, 21%) on an intention-to-treat basis (difference 1.4 percentage points; *P*=.71 for secondary outcome).

Self-reported mean repeated 1-day abstinence was higher for the treated group (mean 17.3%, SD 25.6) than for the control group (mean 12.4%, SD 21.3) on an intention-to-treat basis (difference 4.9 percentage points; *P*=.01) and among complete cases (difference 10 percentage points; *P*=.01, tertiary outcome).

Satisfaction and Motivation

Satisfaction with the Smoke Free app on a 5-point Likert scale (1=not at all to 5=extremely) did not differ by study group among those who completed the follow-up questionnaire (mean 3.7, SD 1.1 treated group vs mean 3.7, SD 1.0 control group; *P*=.91 for secondary outcome). Additional questions regarding satisfaction with the app were also similar between study groups (Figure S2 in Multimedia Appendix 3). Satisfaction with the Inner Dragon game (mean 3.3, SD 1.4) was a bit lower in the treated group than for the app overall.

Motivation to (stay) quit at follow-up, measured on a 10-point scale (0=not at all motivated to 10=very motivated), did not differ significantly (mean 9.3, SD 1.3 treated group vs mean 9.3, SD 1.3 control group; *P*=.87 for tertiary outcome); however, a Qualtrics programming error led to missing data for 51% (148/290) of respondents. The digital therapeutic alliance also did not differ by the study group at follow-up (mean 12.7, SD 2.2 treated group vs mean 12.9, SD 2.4 control group; *P*=.75).

XSL•FO
RenderX

Table 2. Comparison of means for user engagement, smoking abstinence, and user satisfaction according to study group assignment.

	Control	Treated	Difference in means	*P* value[a]	Observations, n[b]
Panel A: user engagement					
App sessions[c], mean (SD)	24.3 (37.9)	29.6 (36.5)	5.3	.06	479
Minutes per session[c], mean (SD)	6.1 (5.2)	6.9 (5.4)	0.9	.047	479
Days with session, mean (SD)	11.9 (13.3)	14.3 (15.3)	2.5	.03	479
Game sessions, mean (SD)	__[d]	7.1 (14.9)	—	—	241
Index of core feature use, mean (SD)	22.6 (35.6)	29.4 (41.3)	6.8	.03	479
Use of core features, mean (SD)					
Cravings reported	1.9 (4.1)	2.7 (5.0)	0.8	.02	479
Diary entries	9.2 (13.2)	12.1 (15.9)	3.0	.01	479
Missions completed	3.8 (8.8)	4.9 (10.2)	1.0	.12	479
Chatbot sessions	4.0 (9.9)	5.4 (11.5)	1.4	.08	479
Index of game feature use, mean (SD)	—	112.9 (258.6)	—	—	241
Panel B: point-prevalence abstinence					
7-day abstinence at 2 months, % (n/N)					
Self-reported, missing=smoking	42.4 (101/238)	39 (94/241)	–3.4	.45	479
Self-reported, complete cases	73.2 (101/138)	61 (94/154)	–12.1	.03	292
Verified[e], missing=smoking	22.7 (54/238)	19.5 (47/241)	–3.2	.39	479
Verified[e], complete cases	74 (54/73)	73.4 (47/64)	–0.5	.94	137
30-day abstinence at 2 months, % (n/N)					
Self-reported, missing=smoking	21 (50/238)	22.4 (54/241)	1.4	.71	479
Self-reported, complete cases	36.2 (50/138)	35.1 (54/154)	–1.2	.84	292
Mean repeated 1-day abstinence, mean (SD)					
Self-reported, missing=smoking	12.4 (21.3)	17.3 (25.6)	4.9	.01	479
Self-reported, complete cases	55 (42.8)	65 (39)	10.0	.01	401
Panel C: satisfaction and motivation					
Satisfaction with app[f], mean (SD)	3.7 (1.1)	3.7 (1.0)	0	.91	271
Satisfaction with game[g], mean (SD)	—	3.3 (1.4)	—	—	84
Recommend app to friends[h], mean (SD)	4.0 (1.1)	3.8 (1.1)	–0.2	.20	271
Motivation to (stay) quit, mean (SD)	9.4 (1.5)	9.3 (1.3)	0	.87	144
Digital therapeutic alliance, mean (SD)	12.9 (2.4)	12.7 (2.2)	–0.2	.75	268

[a]The *P* value is derived from a *t* test of the crude (unadjusted) difference in means between the treated and control groups.

[b]The number of observations (rightmost column) varies by outcome based on whether it was collected passively or in the follow-up survey, as well as whether it applies to all participants or treated participants only.

[c]Denotes a primary outcome measure.

[d]Not applicable. Applies to the treated group only.

[e]Codes as abstinent those who self-reported abstinence but reported use of nicotine replacement therapy or vaping products within the last 7 days.

[f]"I liked using the Smoke Free app" on a Likert scale ranging from "not at all" (1) to "extremely" (5).

[g]"I liked Inner Dragon" on a Likert scale ranging from "not at all" (1) to "extremely" (5).

[h]"I would recommend the Smoke Free App to a friend who wants to quit" on a Likert scale ranging from "not at all" (1) to "extremely" (5).

Outcome Trends

Days with any app sessions decreased for both study groups over time, from roughly 80% on the person's quit day to roughly 20% on day 56 (Figure 2). The treated group retained slightly increased use throughout this period.

XSL•FO
RenderX

Figure 2. Percentage of participants with any sessions of Smoke Free app use by day in the game intervention (treated) group versus the control group (N=479).

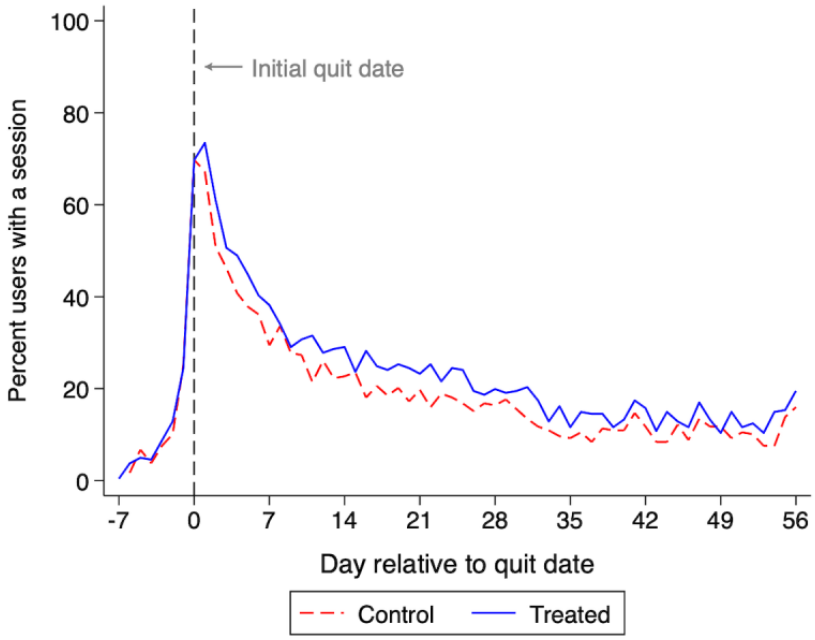

While the share of participants who reported having abstained the prior day decreased for both study groups, the treated group retained its advantage over the control group throughout the study period when analyzing the data on an intention-to-treat basis (Figure 3A). When analyzing complete cases only, we observe that the percentage of participants who reported having abstained remained high throughout the study period, and the magnitude of the difference between study groups was not as great as when assuming missing=smoking (Figure 3B).

Figure 3. Percentage of participants with self-reported repeated 1-day point-prevalence abstinence by day in the game intervention (treated) group versus control group. (A) This panel assumes that missing reports are current smokers. (B) This panel uses complete cases only.

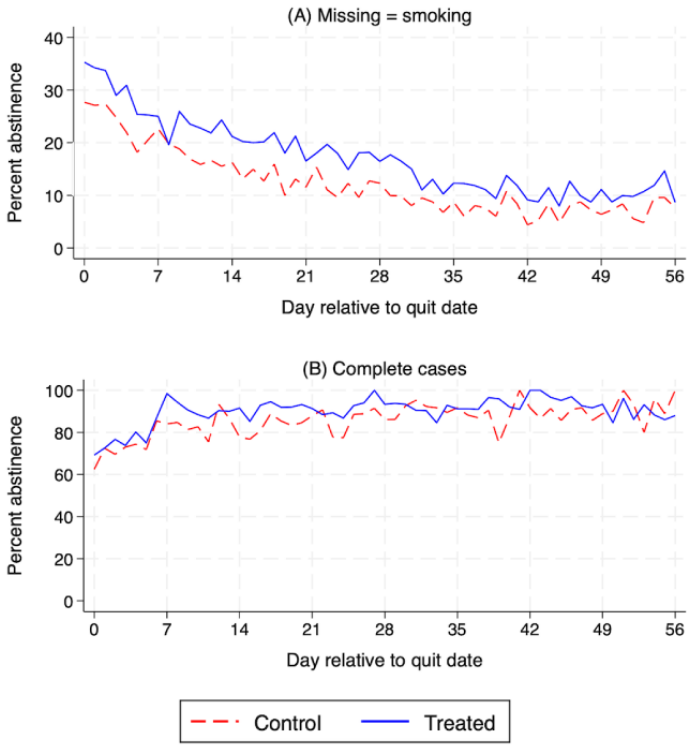

XSL•FO
RenderX

Intensity of Game Use

Treated participants who had more sessions of game use had more app sessions overall, which is a mechanical relationship. More sessions of game use was also associated with increased program adherence (Figure 4B), increased mean repeated 1-day abstinence (Figure 4C), and 7-day abstinence at follow-up (Figure 4D). A 1 SD in sessions of game use (from 0-15 game sessions) was associated with a 4-fold increase in program adherence and an increase in self-reported 7-day point-prevalence abstinence at follow-up of 12.9 percentage points (from 33.8%, 95% CI 24.5%-43.2%, to 46.7%, 95% CI 25.2%-57.5%).

Figure 4. Association of intensity of game use with selected outcomes. The shaded area denotes the 95% CI.

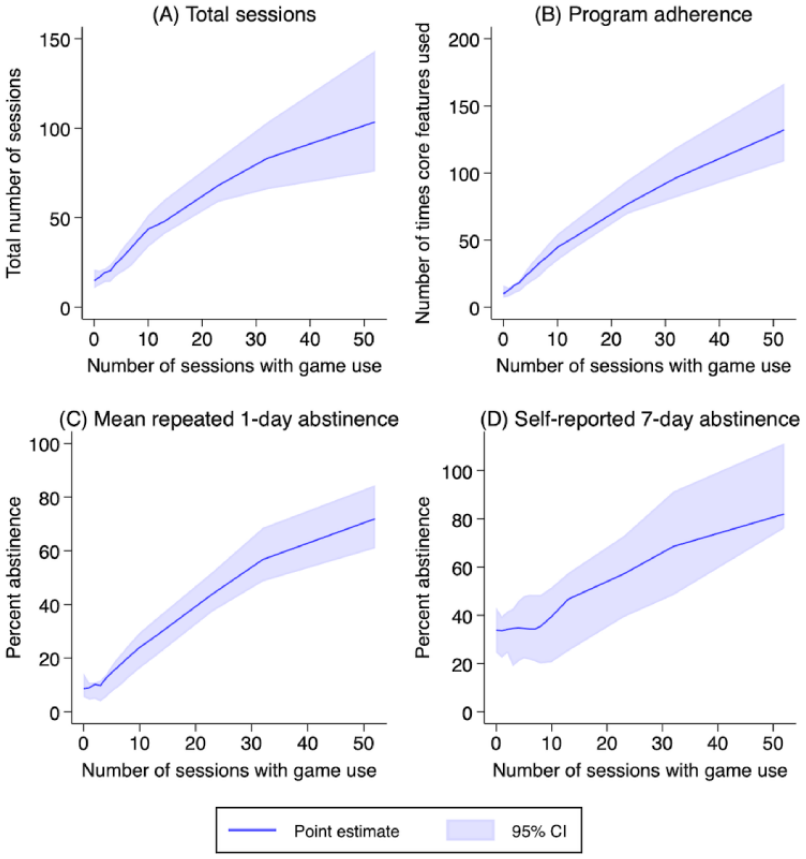

Sensitivity Analyses

Subgroup Analyses

The effects of Inner Dragon on user engagement did not significantly vary with participant characteristics, according to subgroup analyses, although certain patterns emerged (Figure S6 in Multimedia Appendix 3). In unadjusted analyses, the treatment effect on the total number of app sessions was greater for younger participants (aged <50 years) and non-Hispanic White participants compared with non-Hispanic Black or Hispanic participants and decreased with household income. Furthermore, the effects were greater for those with low nicotine dependence, fewer (<4) past quit attempts, and use of e-cigarettes in the last 30 days. Intervention effects did not vary by video gaming history.

Covariate-Adjusted Analyses

Adjusting for participant characteristics (Table 1) as covariates, the differences in user engagement outcomes by study group were similar to, but slightly smaller in magnitude than, the crude differences for number of app sessions (coefficient 4.6, 95% CI –2.2 to 11.3; *P*=.18), average minutes per session (coefficient 0.8, 95% CI –0.2 to 1.9; *P*=.11), days with a session (coefficient 2.2, 95% CI –0.4 to 4.7; *P*=.10), and program adherence (coefficient 6.6, 95% CI –0.4 to 13.6; *P*=.06; Table S3 in Multimedia Appendix 3). The effect sizes for self-reported and verified 7-day and self-reported 30-day point-prevalence abstinence remained similar to the crude differences in magnitude. Estimates, on an intention-to-treat basis, were –2.4 percentage points (95% CI –11.1 to 6.4; *P*=.60) for self-reported 7-day abstinence, –2.7 percentage points (95% CI –10.0 to 4.6; *P*=.47) for verified 7-day abstinence, and 1.4 percentage points (95% CI –6.0 to 8.9; *P*=.70) for self-reported 30-day abstinence. Mean repeated 1-day abstinence remained higher for the treated group than the control group (difference 4.5 percentage points, 95% CI 0.2 to 8.7; *P*=.04) on an intention-to-treat basis (Table S3 in Multimedia Appendix 3).

Sensitivity of Abstinence Estimates

Coding NRT and ENDS users as not abstinent, verified 7-day point-prevalence estimates were similar to the main results (difference 4.3 percentage points, 95% CI –1.8 to 10.5; *P*=.16). Results from the pattern-mixture model indicated that the treatment effect estimates for abstinence were highly sensitive

XSL•FO
RenderX

to assumptions about the nature of the missing outcomes (Figure S1 in Multimedia Appendix 3). Assuming missingness is random (equivalent to missing=smoking), there would be no difference in abstinence between groups, whereas assuming informative missingness, there would be reduced abstinence in the treated group compared with the control group.

Adverse Events

There were no serious adverse events reported during the trial. In total, 1.2% (3/241) reported adverse events in the treated group and 2.9% (7/239) in the control group. Adverse events were feelings of withdrawal, craving, or irritability (1 report in the treated group and 4 in the control group); being anxious or depressed (1 report in the treated group and 3 in the control group); and constipation (1 report in the treated group and 0 in the control group).

Discussion

Principal Findings

Our RCT assessed the efficacy of the novel Inner Dragon game module integrated into the Smoke Free app for promoting user engagement. The study of 479 participants revealed that treated individuals exhibited some increased user engagement metrics, including a (nonsignificant) 21.8% increase in the number of app sessions and a significant 20.2% increase in days of app use. Engaging with Inner Dragon led users in the treated group to use core features of the original Smoke Free app significantly (30.1%) more often, and this increase in smoking cessation program adherence implies that game use complemented (“crowded in”) rather than replaced (“crowded out”) use of the nongame content. User engagement is a fundamental concern for mobile health apps [22-24], and our findings support the role of gamification as a key driver of user engagement.

Smoking abstinence outcomes showed a nuanced picture, with self-reported 7-day point-prevalence abstinence favoring the control group and repeated 1-day abstinence favoring the treated group. In exploratory analyses, we found that this difference was driven in part by treated participants with missing point-prevalence abstinence data having much higher mean repeated abstinence than control participants with missing point-prevalence abstinence data (mean 54.5% SD 5.5 vs mean 35.4% SD 4.9, difference 19.1 percentage points; P=.01). Nevertheless, improving user engagement and program adherence metrics did not definitively produce higher smoking abstinence rates in this trial. One possibility is that gamifying the quit process may have undermined participants’ focus on the higher purpose of quitting smoking. Further research into features of the game, user characteristics, and the original app may be warranted to understand the relationship between user engagement and smoking abstinence.

Finally, the study found a positive association between the intensity of game use and both program adherence and smoking abstinence, suggesting that game-driven engagement may support downstream outcomes.

Comparison With Previous Studies

In the context of the existing literature on smartphone apps for smoking cessation, our findings align with studies emphasizing the importance of user engagement in digital interventions [22-24]. Notably, the increased engagement observed in our study was associated with positive outcomes such as greater program adherence in the form of higher core feature use. To our knowledge, the link between user engagement and program adherence has not been explored with regard to games for smoking cessation, in part because previously evaluated games have been stand-alone products [96].

We found some evidence of effect modification, such that user engagement was greater for younger and lower-income users. Our results echo the call for personalized interventions, consistent with research suggesting the impact of individual characteristics on engagement and outcomes in digital health interventions [97]. Furthermore, our exploration of gamification elements in the Inner Dragon game aligns with literature highlighting the potential of gaming features to enhance engagement and promote behavior change in smoking cessation apps [96]. Prior studies of smartphone-based games for smoking cessation have found mixed evidence regarding smoking abstinence [19,20,38,98], although study quality in the broader literature has generally been deemed fair or poor with important methodological limitations [98]. We found that the bundled game intervention as a whole contributed to the effects of Inner Dragon on increased user engagement. We were not able to isolate the effect of any individual feature of Inner Dragon; exploratory analyses revealed that participants’ use of most key game features was moderately or strongly positively associated with the total number of app sessions (range: 0.39-0.74), and each game feature was strongly or very strongly positively associated with other game features (range: 0.53-1.00; Figure S5 in Multimedia Appendix 3). A future study might use a factorial design to identify the most important game features for stimulating user engagement and abstinence.

Our study supports the integration of personalized strategies, such as games in digital interventions, tailored to individual characteristics and broadly appealing to users with different motivations (eg, achievement, socializing, and immersion), to optimize user engagement and intervention effectiveness. Emerging capabilities of artificial intelligence may facilitate this process [99]. Relatedly, self-paced initiatives such as gamified smoking cessation apps constitute an alternative to traditional smoking cessation interventions that often depend on advanced scheduling. Many researchers have sought ways to leverage technological advancements that enable the delivery of just-in-time, tailored support for behavior change, although large-scale studies on the effectiveness of smoking cessation are lacking [100].

Comparing our abstinence rates to previous studies of smoking cessation apps is difficult due to important differences in study design, such as intervention delivery, comparator, follow-up period, abstinence measure, and the use of bioverification. Focusing on 14 RCTs with follow-up of 1 to 3 months, abstinence ranged widely from 1.6% (234/14,228) to 81.8% (9/11) in the treated groups and 0.9% (124/13,884) to 66.2%

XSL•FO
RenderX

(190/287) in the control groups, and intention-to-treat intervention effects ranged from –0.9 to 11.5 percentage points (Table S4 in Multimedia Appendix 3). We observed abstinence within these ranges, although abstinence in our control group was relatively high (101/238, 42.4%), suggesting that the original Smoke Free app was a particularly effective comparator group on its own. Our study is also distinguished from other app-based studies for having collected high-frequency (daily) abstinence data as surveillance within a smoking cessation intervention; however, there is, related literature on just-in-time adaptive interventions that collect high-frequency data with the aim of identifying smoking triggers and delivering real-time feedback [100].

Strengths and Limitations

This study has important strengths. First, this study is one of the largest randomized trials of a smartphone-based intervention involving games or gamification for smoking cessation. Second, to our knowledge, it is the first study to assess a game that is integrated into an existing smoking cessation app. By integrating Inner Dragon into one of the most downloaded smoking cessation apps, which also has established effectiveness, we were able to evaluate the added benefit of the game module for user engagement. The game intervention can be readily scaled to the 800,000 users who download Smoke Free each year; Inner Dragon is planned to become available to general users of Smoke Free starting in 2024. Third, passive collection of use data ensured that our primary outcome data were complete, accurate, and able to be tracked longitudinally. Fourth, the availability of high-frequency, longitudinal abstinence data was unique for studies of smoking cessation games and uncommon for studies of smoking cessation interventions more generally [101].

However, this study has several limitations. First, the short-term follow-up limits our ability to assess sustained abstinence outcomes. Second, the generalizability of our findings may be constrained by the demographic composition of the app-recruited study population. It is noteworthy that our study drew from a general population of users of the Smoke Free app, a diverse user base in a real-world context. Yet, the use of the Smoke Free app may not be representative of all smoking cessation apps, and the rapidly evolving landscape of digital interventions may influence the generalizability of our findings over time. Third, our game intervention includes a bundle of interlocking features, and we cannot disentangle their individual effects. Fourth, while our study was adequately powered for our primary outcomes, we may have been underpowered for some secondary outcomes, such as smoking abstinence, as well as for subgroup analyses.

Conclusions

In conclusion, our study contributes valuable insights to the growing body of literature on digital smoking cessation interventions involving serious games, emphasizing the role of user engagement. The positive associations between game use with program adherence and smoking cessation outcomes underscore the potential of gamification interventions in promoting behavior change. Future research should extend follow-up durations, refine personalized interventions, evaluate effectiveness in a real-world setting, and address methodological challenges, such as missing data, to guide the development of evidence-based digital health interventions for smoking cessation.

Acknowledgments

The authors would like to thank Keith McCullough and his team at Workinman Interactive, LLC, who contributed to the game design and provided artwork for Inner Dragon. Adam MacLean, Elena McGahey, Giselle Edwards, Pallavi Chaudhary, and Janardan Devkota provided valuable research assistance. This study was supported by funding from the National Cancer Institute (R21 CA238301); the National Institute on Aging (P30 AG012839); the Center on the Economics and Demography of Aging at the University of California, Berkeley; and the Hellman Fellows Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the US government.

Data Availability

The survey datasets generated and analyzed during this study are publicly available [102]. The app use data generated and analyzed during this study are not publicly available due to a third-party agreement between the University of California San Francisco and 23 Limited (the developer of Smoke Free) that prohibits data sharing.

Conflicts of Interest

BRR is the founder of Re-Connect Health, Inc, a company that develops mobile apps for health behavior change. CRL was a visiting researcher at Google LLC during the preparation of this manuscript. DC is the founder and chief executive officer and EW is the chief technology officer of the Smoke Free app, and both derive income from it. LCA receives royalties for the sale of Text2Quit, a quit smoking program. JT reports membership on the scientific advisory board of MindCotine, Inc, which offers a smoking cessation program. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflicts of interest policies. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

XSL•FO
RenderX

Trial protocol.
[PDF File (Adobe PDF File), 787 KB - jmir_v26i1e57839_app1.pdf]

Multimedia Appendix 2
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 82 KB - jmir_v26i1e57839_app2.pdf]

Multimedia Appendix 3
Supplemental tables and figures.
[DOCX File , 1227 KB - jmir_v26i1e57839_app3.docx]

References

1. US Department of Health and Human Services. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta, GA: Centers for Disease Control and Prevention; 2014.
2. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults — United States, 2000–2015. MMWR Morb Mortal Wkly Rep 2017 Jan 06;65(52):1457-1464. [doi: 10.15585/mmwr.mm6552a1]
3. U.S. National Cancer Institute and World Health Organization. The economics of tobacco and tobacco control. National Cancer Institute. 2016. URL: https://cancercontrol.cancer.gov/brp/tcrb/monographs/monograph-21 [accessed 2023-10-01]
4. Rafful C, García-Rodríguez O, Wang S, Secades-Villa R, Martínez-Ortega JM, Blanco C. Predictors of quit attempts and successful quit attempts in a nationally representative sample of smokers. Addict Behav 2013 Apr;38(4):1920-1923. [doi: 10.1016/j.addbeh.2012.12.019]
5. Abroms LC, Lee Westmaas J, Bontemps-Jones J, Ramani R, Mellerson J. A content analysis of popular smartphone apps for smoking cessation. Am J Prev Med 2013 Dec;45(6):732-736. [doi: 10.1016/j.amepre.2013.07.008]
6. Mobile fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-04-29]
7. Chu KH, Matheny SJ, Escobar-Viera CG, Wessel C, Notier AE, Davis EM. Smartphone health apps for tobacco cessation: a systematic review. Addict Behav 2021 Jan;112:106616. [doi: 10.1016/j.addbeh.2020.106616] [Medline: 32932102]
8. Bricker JB, Mull KE, Kientz JA, Vilardaga R, Mercer LD, Akioka KJ, et al. Randomized, controlled pilot trial of a smartphone app for smoking cessation using acceptance and commitment therapy. Drug Alcohol Dependence 2014 Oct 01;143:87-94. [doi: 10.1016/j.drugalcdep.2014.07.006]
9. Iacoviello BM, Steinerman JR, Klein DB, Silver TL, Berger AG, Luo SX, et al. Clickotine, a personalized smartphone app for smoking cessation: initial evaluation. JMIR Mhealth Uhealth 2017 Apr 25;5(4):e56 [FREE Full text] [doi: 10.2196/mhealth.7226] [Medline: 28442453]
10. BinDhim NF, McGeechan K, Trevena L. Smartphone smoking cessation application (SSC App) trial: a multicountry double-blind automated randomised controlled trial of a smoking cessation decision-aid 'app'. BMJ Open 2018 Jan 21;8(1):e017105. [doi: 10.1136/bmjopen-2017-017105] [Medline: 29358418]
11. Marler JD, Fujii CA, Utley DS, Tesfamariam LJ, Galanko JA, Patrick H. Initial assessment of a comprehensive digital smoking cessation program that incorporates a mobile app, breath sensor, and coaching: cohort study. JMIR Mhealth Uhealth 2019 Feb 04;7(2):e12609. [doi: 10.2196/12609]
12. Krebs P, Burkhalter J, Fiske J, Snow H, Schofield E, Iocolano M, et al. The QuitIT coping skills game for promoting tobacco cessation among smokers diagnosed with cancer: pilot randomized controlled trial. JMIR Mhealth Uhealth 2019 Jan 10;7(1):e10071. [doi: 10.2196/10071] [Medline: 30632971]
13. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. JAMA Intern Med 2020 Nov 01;180(11):1472-1480. [doi: 10.1001/jamainternmed.2020.4055] [Medline: 32955554]
14. Crane D, Ubhi HK, Brown J, West R. Relative effectiveness of a full versus reduced version of the 'smoke free' mobile application for smoking cessation: an exploratory randomised controlled trial. F1000Res 2018 Sep 21;7:1524. [doi: 10.12688/f1000research.16148.2] [Medline: 30728950]
15. Jackson SE, Kale D, Beard E, Perski O, West R, Brown J. Effectiveness of the offer of the smoke free smartphone application compared with no intervention for smoking cessation: a pragmatic randomised controlled trial. medRxiv. Preprint posted online on January 12, 2023 . [doi: 10.1101/2023.01.12.23284463]
16. Krishnan N, Elf JL, Chon S, Golub JE. COach2Quit: a pilot randomized controlled trial of a personal carbon monoxide monitor for smoking cessation. Nicotine Tob Res 2019 Oct 26;21(11):1573-1577. [doi: 10.1093/ntr/nty182] [Medline: 30169740]
17. Goldenhersch E, Thrul J, Ungaretti J, Rosencovich N, Waitman C, Ceberio MR. Virtual reality smartphone-based intervention for smoking cessation: pilot randomized controlled trial on initial clinical efficacy and adherence. J Med Internet Res 2020 Jul 29;22(7):e17571 [FREE Full text] [doi: 10.2196/17571] [Medline: 32723722]

XSL•FO
RenderX

18. Herbec A, Brown J, Shahab L, West R, Raupach T. Pragmatic randomised trial of a smartphone app (NRT2Quit) to improve effectiveness of nicotine replacement therapy in a quit attempt by improving medication adherence: results of a prematurely terminated study. Trials 2019 Sep 02;20(1):547. [doi: 10.1186/s13063-019-3645-4] [Medline: 31477166]
19. Pallejà-Millán M, Rey-Reñones C, Barrera Uriarte ML, Granado-Font E, Basora J, Flores-Mateo G, et al. Evaluation of the Tobbstop mobile app for smoking cessation: cluster randomized controlled clinical trial. JMIR Mhealth Uhealth 2020 Jun 26;8(6):e15951. [doi: 10.2196/15951] [Medline: 32589153]
20. Baskerville NB, Struik LL, Guindon GE, Norman CD, Whittaker R, Burns C, et al. Effect of a mobile phone intervention on quitting smoking in a young adult population of smokers: randomized controlled trial. JMIR Mhealth Uhealth 2018 Oct 23;6(10):e10893. [doi: 10.2196/10893] [Medline: 30355563]
21. Paige SR, Alber JM, Stellefson ML, Krieger JL. Missing the mark for patient engagement: mHealth literacy strategies and behavior change processes in smoking cessation apps. Patient Educ Couns 2018 May;101(5):951-955 [FREE Full text] [doi: 10.1016/j.pec.2017.11.006] [Medline: 29153592]
22. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022 Apr 26;24(4):e35120. [doi: 10.2196/35120] [Medline: 35471414]
23. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
24. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
25. Bowman D. Motivating patients to use smartphone health apps. Fierce Healthcare. 2011 May 3. URL: https://www.fiercehealthcare.com/healthcare/motivating-patients-to-use-smartphone-health-apps [accessed 2017-05-20]
26. Cafazzo JA, Casselman M, Hamming N, Katzman DK, Palmert MR. Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study. J Med Internet Res 2012 May 08;14(3):e70. [doi: 10.2196/jmir.2058] [Medline: 22564332]
27. Primack BA, Carroll MV, McNamara M, Klem ML, King B, Rich M, et al. Role of video games in improving health-related outcomes: a systematic review. Am J Prev Med 2012 Jun;42(6):630-638. [doi: 10.1016/j.amepre.2012.02.023] [Medline: 22608382]
28. King D, Greaves F, Exeter C, Darzi A. 'Gamification': influencing health behaviours with games. J R Soc Med 2013 Mar;106(3):76-78 [FREE Full text] [doi: 10.1177/0141076813480996] [Medline: 23481424]
29. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PLoS One 2017;12(3):e0173403 [FREE Full text] [doi: 10.1371/journal.pone.0173403] [Medline: 28362821]
30. Richardson A, Graham AL, Cobb N, Xiao H, Mushro A, Abrams D, et al. Engagement promotes abstinence in a web-based cessation intervention: cohort study. J Med Internet Res 2013 Jan 28;15(1):e14 [FREE Full text] [doi: 10.2196/jmir.2277] [Medline: 23353649]
31. Raiff BR, Jarvis BP, Rapoza D. Prevalence of video game use, cigarette smoking, and acceptability of a video game-based smoking cessation intervention among online adults. Nicotine Tob Res 2012 Dec 15;14(12):1453-1457 [FREE Full text] [doi: 10.1093/ntr/nts079] [Medline: 22422929]
32. Oliver JA, Hallyburton MB, Pacek LR, Mitchell JT, Vilardaga R, Fuemmeler BF, et al. What do smokers want in a smartphone-based cessation application? Nicotine Tob Res 2018 Nov 15;20(12):1507-1514 [FREE Full text] [doi: 10.1093/ntr/ntx171] [Medline: 29065202]
33. Edwards EA, Caton H, Lumsden J, Rivas C, Steed L, Pirunsarn Y, et al. Creating a theoretically grounded, gamified health app: lessons from developing the Cigbreak smoking cessation mobile phone game. JMIR Serious Games 2018 Nov 29;6(4):e10252 [FREE Full text] [doi: 10.2196/10252] [Medline: 30497994]
34. Baskerville NB, Struik LL, Dash D. Crush the crave: development and formative evaluation of a smartphone app for smoking cessation. JMIR Mhealth Uhealth 2018 Mar 02;6(3):e52 [FREE Full text] [doi: 10.2196/mhealth.9011] [Medline: 29500157]
35. Raiff BR, Fortugno N, Scherlis DR, Rapoza D. A mobile game to support smoking cessation: prototype assessment. JMIR Serious Games 2018 Jun 07;6(2):e11 [FREE Full text] [doi: 10.2196/games.9599] [Medline: 29880466]
36. Krebs P, Burkhalter JE, Snow B, Fiske J, Ostroff JS. Development and alpha testing of QuitIT: an interactive video game to enhance skills for coping with smoking urges. JMIR Res Protoc 2013 Sep 11;2(2):e35 [FREE Full text] [doi: 10.2196/resprot.2416] [Medline: 24025236]
37. Bindoff I, de Salas K, Peterson G, Ling T, Lewis I, Wells L, et al. Quittr: the design of a video game to support smoking cessation. JMIR Serious Games 2016 Dec 01;4(2):e19 [FREE Full text] [doi: 10.2196/games.6258] [Medline: 27908844]
38. Bindoff I, Ling TR, Gee P, Geelan B, Ferguson SG, Peterson GM. Effects of a mobile app called Quittr, which utilizes premium currency and games features, on improving engagement with smoking cessation intervention: pilot randomized controlled trial. JMIR Serious Games 2020 Dec 14;8(4):e23734 [FREE Full text] [doi: 10.2196/23734] [Medline: 33315016]

XSL•FO
RenderX

39. Marin-Gomez FX, Garcia-Moreno Marchán R, Mayos-Fernandez A, Flores-Mateo G, Granado-Font E, Barrera Uriarte ML, et al. Exploring efficacy of a serious game (Tobbstop) for smoking cessation during pregnancy: randomized controlled trial. JMIR Serious Games 2019 Mar 27;7(1):e12835 [FREE Full text] [doi: 10.2196/12835] [Medline: 30916655]
40. Jackson SE, Perski O, Crane D, Michie S, West R, Brown J. Effectiveness of an offer of the smoke free smartphone application for smoking cessation: protocol for a randomized controlled trial. Addiction 2019 Nov;114(11):2078-2086. [doi: 10.1111/add.14652] [Medline: 31083767]
41. White JS, Salem MK, Toussaert S, Westmaas JL, Raiff BR, Crane D, et al. Developing a game (inner dragon) within a leading smartphone app for smoking cessation: design and feasibility evaluation study. JMIR Serious Games 2023 Aug 11;11:e46602. [doi: 10.2196/46602] [Medline: 37566442]
42. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895. [doi: 10.2196/21895] [Medline: 33372897]
43. Henley SJ, Thomas CC, Sharapova SR, Momin B, Massetti GM, Winn DM, et al. Vital signs: disparities in tobacco-related cancer incidence and mortality - United States, 2004-2013. MMWR Morb Mortal Wkly Rep 2016 Nov 11;65(44):1212-1218 [FREE Full text] [doi: 10.15585/mmwr.mm6544a3] [Medline: 27832048]
44. Irvin Vidrine J, Reitzel LR, Wetter DW. The role of tobacco in cancer health disparities. Curr Oncol Rep 2009 Nov;11(6):475-481 [FREE Full text] [doi: 10.1007/s11912-009-0064-9] [Medline: 19840525]
45. Fiscella K, Franks P, Doescher MP, Saver BG. Disparities in health care by race, ethnicity, and language among the insured: findings from a national sample. Med Care 2002 Jan;40(1):52-59. [doi: 10.1097/00005650-200201000-00007]
46. Hargraves JL, Hadley J. The contribution of insurance coverage and community resources to reducing racial/ethnic disparities in access to care. Health Serv Res 2003 Jun;38(3):809-829 [FREE Full text] [doi: 10.1111/1475-6773.00148] [Medline: 12822914]
47. Cokkinides VE, Halpern MT, Barbeau EM, Ward E, Thun MJ. Racial and ethnic disparities in smoking-cessation interventions: analysis of the 2005 national health interview survey. Am J Prev Med 2008 May;34(5):404-412. [doi: 10.1016/j.amepre.2008.02.003] [Medline: 18407007]
48. Franks P, Fiscella K, Meldrum S. Racial disparities in the content of primary care office visits. J Gen Intern Med 2005 Jul;20:599-603. [doi: 10.1007/s11606-005-0106-4]
49. Ubhi HK, Michie S, Kotz D, Wong WC, West R. A mobile app to aid smoking cessation: preliminary evaluation of SmokeFree28. J Med Internet Res 2015 Jan 16;17(1):e17 [FREE Full text] [doi: 10.2196/jmir.3479] [Medline: 25596170]
50. Bricker JB, Copeland W, Mull KE, Zeng EY, Watson NL, Akioka KJ, et al. Single-arm trial of the second version of an acceptance and commitment therapy smartphone application for smoking cessation. Drug Alcohol Depend 2017 Jan 01;170:37-42 [FREE Full text] [doi: 10.1016/j.drugalcdep.2016.10.029] [Medline: 27870987]
51. van den Brand FA, Nagelhout GE, Winkens B, Chavannes NH, van Schayck OC. Effect of a workplace-based group training programme combined with financial incentives on smoking cessation: a cluster-randomised controlled trial. Lancet Public Health 2018 Nov;3(11):e536-e544. [doi: 10.1016/s2468-2667(18)30185-3]
52. Augustson E, Cole-Lewis H, Sanders A, Schwarz M, Geng Y, Coa K, et al. Analysing user-reported data for enhancement of SmokefreeTXT: a national text message smoking cessation intervention. Tob Control 2017 Nov;26(6):683-689. [doi: 10.1136/tobaccocontrol-2016-052945] [Medline: 27852892]
53. Theng YL, Lee JW, Patinadan PV, Foo SS. The use of videogames, gamification, and virtual environments in the self-management of diabetes: a systematic review of evidence. Games Health J 2015 Oct;4(5):352-361. [doi: 10.1089/g4h.2014.0114] [Medline: 26287926]
54. Patel MS, Benjamin EJ, Volpp KG, Fox CS, Small DS, Massaro JM, et al. Effect of a game-based intervention designed to enhance social incentives to increase physical activity among families: the BE FIT randomized clinical trial. JAMA Intern Med 2017 Nov 01;177(11):1586-1593 [FREE Full text] [doi: 10.1001/jamainternmed.2017.3458] [Medline: 28973115]
55. Roepke AM, Jaffee SR, Riffle OM, McGonigal J, Broome R, Maxwell B. Randomized controlled trial of SuperBetter, a smartphone-based/internet-based self-help tool to reduce depressive symptoms. Games Health J 2015 Jun;4(3):235-246. [doi: 10.1089/g4h.2014.0046] [Medline: 26182069]
56. El-Hilly AA, Iqbal SS, Ahmed M, Sherwani Y, Muntasir M, Siddiqui S, et al. Game on? Smoking cessation through the gamification of mHealth: a longitudinal qualitative study. JMIR Serious Games 2016 Oct 24;4(2):e18 [FREE Full text] [doi: 10.2196/games.5678] [Medline: 27777216]
57. Hall SM, Delucchi KL, Velicer WF, Kahler CW, Ranger-Moore J, Hedeker D, et al. Statistical analysis of randomized trials in tobacco treatment: longitudinal designs with dichotomous outcome. Nicotine Tob Res 2001 Aug;3(3):193-202. [doi: 10.1080/14622200110050411] [Medline: 11506764]
58. Hedeker D, Mermelstein RJ, Demirtas H. Analysis of binary outcomes with missing data: missing = smoking, last observation carried forward, and a little multiple imputation. Addiction 2007 Oct;102(10):1564-1573. [doi: 10.1111/j.1360-0443.2007.01946.x] [Medline: 17854333]
59. Nelson DB, Partin MR, Fu SS, Joseph AM, An LC. Why assigning ongoing tobacco use is not necessarily a conservative approach to handling missing tobacco cessation outcomes. Nicotine Tob Res 2009 Jan;11(1):77-83. [doi: 10.1093/ntr/ntn013] [Medline: 19246444]

XSL•FO
RenderX

60. Blankers M, Smit ES, van der Pol P, de Vries H, Hoving C, van Laar M. The missing=smoking assumption: a fallacy in internet-based smoking cessation trials? Nicotine Tob Res 2016 Jan;18(1):25-33. [doi: 10.1093/ntr/ntv055] [Medline: 25744969]
61. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodol 2001 Jun;27(1):85-96.
62. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007 Jun;16(3):219-242. [doi: 10.1177/0962280206074463] [Medline: 17621469]
63. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239]
64. Upton CR, Nastasi JA, Raiff BR. Identifying video game preferences among adults interested in quitting smoking cigarettes: survey study. JMIR Serious Games 2022 Mar 24;10(1):e30949 [FREE Full text] [doi: 10.2196/30949] [Medline: 35323116]
65. Cornelius ME, Loretan CG, Jamal A, Davis Lynn BC, Mayer M, Alcantara IC, et al. Tobacco product use among adults – United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 05;72(18):475-483. [doi: 10.15585/mmwr.mm7218a1]
66. Ubhi HK, Michie S, Kotz D, van Schayck OC, Selladurai A, West R. Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features. Behav Med Pract Policy Res 2015 Nov 23;6(3):410-417. [doi: 10.1007/s13142-015-0352-x]
67. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019 Sep 30;5:2055207619880676. [doi: 10.1177/2055207619880676] [Medline: 31620306]
68. Bylieva D, Almazova N, Lobatyuk V, Rubtsova A. Virtual pet: trends of development. In: Proceedings of the 2019 International Conference on Digital Science. 2019 Presented at: DSIC 2019; October 11-13, 2019; Limassol, Cyprus. [doi: 10.1007/978-3-030-37737-3_47]
69. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
70. Berry K, Salter A, Morris R, James S, Bucci S. Assessing therapeutic alliance in the context of mHealth interventions for mental health problems: development of the mobile Agnew relationship measure (mARM) questionnaire. J Med Internet Res 2018 Apr 19;20(4):e90. [doi: 10.2196/jmir.8252] [Medline: 29674307]
71. Chaloupka FJ, Levy MR, White JS. Estimating biases in smoking cessation: evidence from a field experiment (article no. 26522). National Bureau of Economic Research. 2019. URL: https://www.nber.org/system/files/working_papers/w26522/w26522.pdf [accessed 2019-12-21]
72. Goldin J, Homonoff T. Smoke gets in your eyes: cigarette tax salience and regressivity. Am Econ J Econ Policy 2013 Feb;5(1):302-336. [doi: 10.1257/pol.5.1.302]
73. Roll JM, Higgins ST. A within-subject comparison of three different schedules of reinforcement of drug abstinence using cigarette smoking as an exemplar. Drug Alcohol Depend 2000 Feb 01;58(1-2):103-109. [doi: 10.1016/s0376-8716(99)00073-3] [Medline: 10669060]
74. Shen L, Fishbach A, Hsee CK. The motivating-uncertainty effect: uncertainty increases resource investment in the process of reward pursuit. J Consum Res 2015 Feb;41(5):1301-1315. [doi: 10.1086/679418]
75. Golman R, Loewenstein G, Molnar A, Saccardo S. The demand for, and avoidance of, information. Manag Sci 2022 Sep;68(9):6454-6476. [doi: 10.1287/mnsc.2021.4244]
76. Buechel EC, Li R. Mysterious consumption: preference for horizontal (vs. vertical) uncertainty and the role of surprise. J Consum Res 2023 Apr;49(6):987-1013. [doi: 10.1093/jcr/ucac039]
77. Meeker D, Linder JA, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, et al. Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: a randomized clinical trial. JAMA 2016 Feb 09;315(6):562-570. [doi: 10.1001/jama.2016.0275] [Medline: 26864410]
78. Allcott H, Rogers T. The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation. Am Econ Rev 2014 Oct;104(10):3003-3037. [doi: 10.1257/aer.104.10.3003]
79. Yee N. Motivations for play in online games. Cyberpsychol Behav 2006 Dec;9(6):772-775. [doi: 10.1089/cpb.2006.9.772] [Medline: 17201605]
80. West R, Hajek P, Stead L, Stapleton J. Outcome criteria in smoking cessation trials: proposal for a common standard. Addiction 2005 Mar;100(3):299-303. [doi: 10.1111/j.1360-0443.2004.00995.x] [Medline: 15733243]
81. Benowitz NL, Bernert JT, Foulds J, Hecht SS, Jacob P, Jarvis MJ, et al. Biochemical verification of tobacco use and abstinence: 2019 update. Nicotine Tob Res 2020 Jun 12;22(7):1086-1097. [doi: 10.1093/ntr/ntz132] [Medline: 31570931]
82. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström test for nicotine dependence: a revision of the Fagerström Tolerance Questionnaire. Br J Addict 1991 Sep;86(9):1119-1127. [doi: 10.1111/j.1360-0443.1991.tb01879.x] [Medline: 1932883]
83. Calonico S, Cattaneo MD, Farrell MH. On the effect of bias estimation on coverage accuracy in nonparametric inference. J Am Stat Assoc 2018 Mar 07;113(522):767-779. [doi: 10.1080/01621459.2017.1285776]
84. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ 2011 Feb 07;342:d40 [FREE Full text] [doi: 10.1136/bmj.d40] [Medline: 21300711]

XSL•FO
RenderX

85. White IR, Carpenter J, Horton NJ. A mean score method for sensitivity analysis to departures from the missing at random assumption in randomised trials. Stat Sin 2018 Oct;28(4):1985-2003 [FREE Full text] [doi: 10.5705/ss.202016.0308] [Medline: 30283213]
86. Hertzberg JS, Carpenter VL, Kirby AC, Calhoun PS, Moore SD, Dennis MF, et al. Mobile contingency management as an adjunctive smoking cessation treatment for smokers with posttraumatic stress disorder. Nicotine Tob Res 2013 Nov;15(11):1934-1938 [FREE Full text] [doi: 10.1093/ntr/ntt060] [Medline: 23645606]
87. Masaki K, Tateno H, Nomura A, Muto T, Suzuki S, Satake K, et al. A randomized controlled trial of a smoking cessation smartphone application with a carbon monoxide checker. NPJ Digit Med 2020 Mar 12;3:35 [FREE Full text] [doi: 10.1038/s41746-020-0243-5] [Medline: 32195370]
88. Webb J, Peerbux S, Smittenaar P, Siddiqui S, Sherwani Y, Ahmed M, et al. Preliminary outcomes of a digital therapeutic intervention for smoking cessation in adult smokers: randomized controlled trial. JMIR Ment Health 2020 Oct 06;7(10):e22833 [FREE Full text] [doi: 10.2196/22833] [Medline: 33021488]
89. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. Cochrane Database Syst Rev 2019 Oct 22;10(10):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub5] [Medline: 31638271]
90. Barnett A, Ding H, Hay KE, Yang IA, Bowman RV, Fong KM, et al. The effectiveness of smartphone applications to aid smoking cessation: a meta-analysis. Clin eHealth 2020;3:69-81 [FREE Full text] [doi: 10.1016/j.ceh.2020.09.001]
91. Cobos-Campos R, de Lafuente AS, Apiñaniz A, Parraza N, Llanos IP, Orive G. Effectiveness of mobile applications to quit smoking: systematic review and meta-analysis. Tob Prev Cessat 2020 Nov 10;6:62 [FREE Full text] [doi: 10.18332/tpc/127770] [Medline: 33241162]
92. Haskins BL, Lesperance D, Gibbons P, Boudreaux ED. A systematic review of smartphone applications for smoking cessation. Transl Behav Med 2017 Jun;7(2):292-299 [FREE Full text] [doi: 10.1007/s13142-017-0492-2] [Medline: 28527027]
93. Regmi K, Kassim N, Ahmad N, Tuah NA. Effectiveness of mobile apps for smoking cessation: a review. Tob Prev Cessat 2017 Apr 12;3:12 [FREE Full text] [doi: 10.18332/tpc/70088] [Medline: 32432186]
94. Staiger PK, O'Donnell R, Liknaitzky P, Bush R, Milward J. Mobile apps to reduce tobacco, alcohol, and illicit drug use: systematic review of the first decade. J Med Internet Res 2020 Nov 24;22(11):e17156 [FREE Full text] [doi: 10.2196/17156] [Medline: 33231555]
95. Ybarra ML, Jiang Y, Free C, Abroms LC, Whittaker R. Participant-level meta-analysis of mobile phone-based interventions for smoking cessation across different countries. Prev Med 2016 Aug;89:90-97 [FREE Full text] [doi: 10.1016/j.ypmed.2016.05.002] [Medline: 27154349]
96. Derksen ME, van Strijp S, Kunst AE, Daams JG, Jaspers MW, Fransen MP. Serious games for smoking prevention and cessation: a systematic review of game elements and game effects. J Am Med Inform Assoc 2020 May 01;27(5):818-833 [FREE Full text] [doi: 10.1093/jamia/ocaa013] [Medline: 32330255]
97. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
98. Rajani NB, Bustamante L, Weth D, Romo L, Mastellos N, Filippidis FT. Engagement with gamification elements in a smoking cessation app and short-term smoking abstinence: quantitative assessment. JMIR Serious Games 2023 Feb 01;11:e39975 [FREE Full text] [doi: 10.2196/39975] [Medline: 36724003]
99. Hassoon A, Baig Y, Naiman DQ, Celentano DD, Lansey D, Stearns V, et al. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. NPJ Digit Med 2021 Dec 09;4(1):168 [FREE Full text] [doi: 10.1038/s41746-021-00539-9] [Medline: 34887491]
100. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241 [FREE Full text] [doi: 10.1111/add.15687] [Medline: 34514668]
101. Businelle MS, Ma P, Kendzor DE, Frank SG, Wetter DW, Vidrine DJ. Using intensive longitudinal data collected via mobile phone to detect imminent lapse in smokers undergoing a scheduled quit attempt. J Med Internet Res 2016 Oct 17;18(10):e275 [FREE Full text] [doi: 10.2196/jmir.6307] [Medline: 27751985]
102. Evaluating the impact of a game (Inner Dragon) on user engagement within a leading smartphone app for smoking cessation: a randomized controlled trial. OSF Home. 2024 Mar 25. URL: https://osf.io/d8zgc/ [accessed 2024-10-04]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
ENDS: electronic nicotine delivery system
NRT: nicotine replacement therapy
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by C Martínez Martínez, A Hassan; comments to author 24.03.24; revised version received 30.04.24; accepted 23.08.24; published 30.10.24.

Please cite as:
White JS, Toussaert S, Raiff BR, Salem MK, Chiang AY, Crane D, Warrender E, Lyles CR, Abroms LC, Westmaas JL, Thrul J Evaluating the Impact of a Game (Inner Dragon) on User Engagement Within a Leading Smartphone App for Smoking Cessation: Randomized Controlled Trial
J Med Internet Res 2024;26:e57839
URL: https://www.jmir.org/2024/1/e57839
doi:10.2196/57839
PMID:

©Justin S White, Séverine Toussaert, Bethany R Raiff, Marie K Salem, Amy Yunyu Chiang, David Crane, Edward Warrender, Courtney R Lyles, Lorien C Abroms, J Lee Westmaas, Johannes Thrul. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Support From Community Health Workers and Health Care Professionals on the Sustained Use of Wearable Monitoring Devices Among Community-Dwelling Older Adults: Feasibility Randomized Controlled Trial

Arkers Kwan Ching Wong[1], BSc, MSc, PhD; Jonathan Bayuo[1], PhD; Jing Jing Su[1], PhD; Frances Kam Yuet Wong[1], PhD; Karen Kit Sum Chow[2], MSc; Bonnie Po Wong[2], BSc; Siu Man Wong[2], BSc; Vivian Hui[1], PhD

[1]School of Nursing, Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[2]Hong Kong Lutheran Social Service, Homantin, China (Hong Kong)

Corresponding Author:
Arkers Kwan Ching Wong, BSc, MSc, PhD
School of Nursing
Hong Kong Polytechnic University
GH 502, Hung Hom
Hong Kong
China (Hong Kong)
Phone: 852 34003805
Email: arkers.wong@polyu.edu.hk

Abstract

Background: The wearable monitoring device (WMD) is emerging as a promising tool for community-dwelling older adults to monitor personal health, enhance awareness of their activities, and promote healthy behaviors. However, the sustained use of WMDs among this population remains a significant challenge.

Objective: This study aims to implement an interventional program that promotes and motivates the continued use of WMDs among older adults through a peer and professional support approach. This program will facilitate the integration of WMDs into their daily lives.

Methods: This feasibility trial examined the following: (1) the usability of the WMD from the users' perspectives; (2) the feasibility of the Live With Wearable Monitoring Device program; and (3) the effectiveness of the Live With Wearable Monitoring Device program among community-dwelling older adults. The intervention, based on Self-Determination Theory, involved using the Live With Wearable Monitoring Device program over a 3-month period, with ongoing professional and peer support provided by community health workers, aided by a nurse and social workers. This support included 1 home visit and biweekly communication via WhatsApp. Data were collected at baseline and at 1, 3, and 6 months.

Results: A total of 39 participants were enrolled in the intervention group, while 37 participants were in the control group. The recruitment rate was high (76/89, 85%), and the attrition rate was low (8/76, 11%), indicating that the program is feasible for older adults. Participants in the intervention group exhibited higher self-efficacy, lower anxiety levels, and used the smartwatch more frequently, in terms of both days and hours, compared with the control group. A between-group difference was observed in self-efficacy between the intervention and control groups (β=3.31, 95% CI 0.36-6.25, P=.03), with statistically significant higher mean values recorded at all 4 time points.

Conclusions: It is clear that merely providing a WMD to older adults does not guarantee its usage, particularly for those unfamiliar with how to utilize its health-related functions in their daily routines. This study implemented a theory-based program aimed at enhancing the ongoing use of WMDs among older adults, suggesting that continuous professional and peer support may significantly influence WMD usage.

Trial Registration: ClinicalTrials.gov NCT05269303; https://clinicaltrials.gov/ct2/show/NCT05269303

(J Med Internet Res 2024;26:e52435) doi:10.2196/52435

XSL•FO
RenderX

KEYWORDS

wearable monitoring device; lay worker; smartwatch; older adult; nurse; engagement; attrition; engagement; wearable; user experience

Introduction

Background

Given the ever-growing aging population, strategies such as health education and self-care management to promote older adults' health are becoming increasingly important [1]. With digital technology, older adults can engage more actively in these strategies, supporting the prevention, early detection, and management of chronic diseases [1,2]. Wearable monitoring devices (WMDs) are among the most popular electronic tools today for encouraging older adults to select or adjust suitable types and frequencies of health-promoting activities based on real-time personal health data. By becoming more active in their own care, older adults can improve their physical fitness and work toward their goal of remaining in their community for as long as possible [2].

A WMD is defined as body-worn technology that enables continuous monitoring of individual health-related activities without restricting or interrupting movement [2]. Although a wide range of WMDs are available on the market, these devices generally allow users to only capture key personal health indicators, such as blood pressure, pulse, physical activity levels, posture, and body weight. Data collected by commercially available WMDs are typically transmitted to a cloud platform or the manufacturer's server for processing. Users then receive automated, general feedback or alerts based on these processed data, guiding them in taking further actions [3]. Some of the latest WMDs can even provide users with high-touch, personalized feedback when the server is directly connected to the smartphones of supporting professionals, such as doctors or nurses [4]. These supporting professionals use users' real-time health parameters to provide meaningful, timely feedback and establish a communication channel, positively impacting users' health and well-being [4]. Most prior research supports the view that WMDs enhance older adults' sense of security and quality of life [5], reduce sedentary behaviors and health care service utilization [6], improve gait balance [1], and lead to reductions in blood pressure and body mass index [7]. Some studies also suggest that WMDs can assist both lay caregivers and health care professionals in tracking care, identifying issues promptly, and providing immediate support to help older family members or clients maintain their independence [8]. Despite the positive impact of WMDs, it is important to note that certain issues and limitations exist. Concerns about privacy and data sharing, limited battery capacity, and accuracy in screening and prediction highlight challenges regarding the reliability and validity of wearables as detection and prediction tools [9].

Given the potential benefits of WMDs in supporting health self-management for older adults, commercial studies report that WMD sales have risen sharply in recent years and are expected to continue growing annually over the next 5 years [10]. However, despite widespread adoption, reports indicate that older adults often discontinue use within a few months, with an abandonment rate exceeding 30% [11-13]. Another longitudinal study echoed this finding, showing that approximately 50% of new WMD users stop using the device within 2 weeks [14]. Possible explanations for this low adherence rate are issues with how the WMDs were implemented and challenges users face in integrating the device into their daily routines, despite its usefulness.

Often, the motivation to continue using WMDs depends on the extent to which users feel the device fulfills their needs, aligns with their goals, and meets their expectations [15,16]. However, while commercially available WMDs—such as those produced by Apple, Fitbit, and Garmin—include goal-setting features, they lack built-in guidance on maximizing the effectiveness of these features. They also offer limited support for action planning and for identifying facilitators and barriers to their use in daily activities. These issues can discourage individuals from using the device, even if they find it useful. Additionally, previous research has found that, beyond design, features, perceived usefulness, and ease of use, the extent to which the device empowers users as responsible for their own health decisions is a crucial factor motivating older adults to sustain its use [17]. As health care professionals can provide instant support whenever a WMD detects abnormal health findings, older adults may feel less compelled to take ownership of their health and engage in self-monitoring. This situation leaves little room for their involvement in planning their health management, which may hinder their determination and motivation to comply with WMD usage.

Regarding the implementation of WMDs, the typical approach has been to provide participants with the device, based on the assumption that they will use it because they understand the benefits [14]. The problem with this approach is that it is passive and does not consider that people may become demotivated to use the device over time. Therefore, more innovative strategies are needed to sustain usage. Given that challenges are likely to arise during device use, an active approach that offers ongoing professional or peer support may be beneficial. Although in a different context, a recent study on the sustained use of an mHealth app found that ongoing follow-up support was effective in equipping community-dwelling older adults with the skills needed to navigate the app while addressing technical issues and limited social support [18]. This suggests that the availability of ongoing professional and peer support provides a platform for older adults to resolve the issues they encounter, thereby potentially helping to sustain their interest in the app.

The World Health Organization's global strategy for digital health emphasizes the importance of empowering users to integrate technology into their daily routines [19]. Community health workers (CHWs), who share similar backgrounds and come from the same community, are highly regarded as peers who can engage older adults in using electronic tools for self-care management [20]. Taylor and his colleagues [21] noted that CHWs possess high health literacy and strong interpersonal

XSL•FO
RenderX

communication skills, enabling them to serve as role models for their peers and improve engagement in self-care and healthy behaviors. Previous programs involving CHWs in key support roles have been effective in improving health outcomes [22], reducing health care costs [23], decreasing the use of health care services [24], and enhancing the quality of care [25] among older adults. With the support of CHWs, older adults can familiarize themselves with the technical design and functionality of the devices, track and manage their health parameters, overcome barriers to using WMDs in their daily lives, codevelop action plans to achieve health goals, and receive personalized feedback on their performance. This support may ultimately facilitate and motivate them to continue using the devices.

Motivating older adults to sustain the use of WMDs is an emerging area of research, although it currently has a limited evidence base. Much of the existing research focuses on exploring the motivational factors that influence WMD use among older adults, the development process of these devices, and their effects on various populations [26]. However, there are only a few interventional studies aimed at empowering and engaging community-dwelling older adults in the sustainable use of WMDs in their daily lives. A recent 1-group pretest-posttest design study implemented a 3-month program that included weekly reminders and monthly phone calls from a research team member to promote the use of a Fitbit device among older adults with osteoarthritis. The results showed that 96% of participants remained in the study, and Fitbit wear time was high, although no health data were captured [27]. Another study provided a group of community-dwelling individuals with WMDs along with support services from a dietitian, demonstrating an improvement in their continued use of the devices [28]. Our study is the first to use a peer support approach to engage community-dwelling older adults in the continued use of WMDs in their daily routines. It aims to assess the usability of the WMD, the program's feasibility, and its effectiveness from the perspective of older adults. If proven successful, the program could offer a solution to motivate older adults to consistently use WMDs for self-care management, thereby improving their health and quality of life in the long term.

Conceptual Framework

This study was guided by Self-Determination Theory (SDT) [29], which posits that individuals are more likely to adhere to a particular behavior when they have intrinsic motivation to do so.

In this study, we aimed to sustain the intrinsic motivation of community-dwelling older adults to facilitate their continued use of WMDs in their daily routines. According to SDT, intrinsic motivation can be determined and enhanced by 3 factors: competence, autonomy, and relatedness.

Competence refers to the sense of ability to overcome challenges associated with an activity or behavior. To facilitate the achievement of this component, the study used a home visit approach by the CHW and the nurse case manager to demonstrate the use of the WMD to participants. This 1-on-1 approach helped identify and resolve barriers inherent in the participants' environments, equipping them with the necessary knowledge and skills for using the WMD effectively. Autonomy refers to a sense of control over an activity or behavior. Through a structured follow-up approach that included biweekly calls from the nurse case manager, emerging concerns regarding the use of the WMD were identified and addressed, facilitating independent use while still providing ongoing support as needed. Relatedness refers to a sense of connection and interaction with others who share an interest in an activity or behavior. To increase the adherence rate of community-dwelling older adults in using WMDs in their daily lives, the present program included components designed to teach them the knowledge and skills needed to effectively use the WMD and overcome technical difficulties. It also provided a clear rationale for suggestions and offered various options for incorporating WMDs into their routines. Additionally, the program aimed to strengthen interpersonal connections and support from family members and peers, facilitating the continued use of the WMD (Figure 1).

Figure 1. Conceptual framework. WMD: wearable monitoring device.

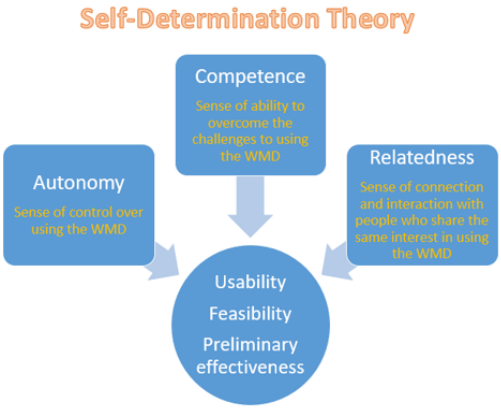

XSL•FO
RenderX

Methods

Study Design, Participants, and Recruitment

This study utilized a feasibility randomized controlled trial design. The study was conducted at 5 community centers affiliated with a local nongovernmental organization. Participants who were members and expressed interest in the program were screened and recruited based on the following criteria: (1) aged 60 or older, (2) owning a smartphone, (3) able to communicate in Cantonese or Mandarin, and (4) having internet access at home or elsewhere. Potential participants were excluded based on the following criteria: (1) a confirmed diagnosis of cognitive impairment, (2) being bedbound, (3) currently owning a WMD, and (4) having previously participated in other wearable device studies.

Staff at the community centers introduced the program to their members via Facebook Live (Meta Platforms, Inc.). Interested members were then contacted and screened by our trained research assistant through telephone calls. For those eligible to participate in the program, the research assistant provided a detailed explanation of the study, obtained consent, and collected baseline data at the community centers.

Randomization

The participants were randomly assigned to either the intervention or control group using the Research Randomizer software (Social Psychology Network) to generate the group assignments. These assignments were sealed and opened sequentially by the principal investigator at the time of randomization. To achieve the highest possible degree of double-blinding, participants were informed only that the intervention aimed to promote the use of WMDs; they were not informed of their specific group assignment (intervention or control). In addition, the research assistant who collected the data was blinded to group allocation; however, the providers, including community center staff, were not.

Intervention

Device Setup and Participant Instructions

After completing baseline data collection and before randomization, participants were provided with a package containing a WMD (ProVista Care; Provista Group), a prepaid SIM card, a blood pressure monitor, and a pulse oximeter. ProVista Care, a commercially available device, was selected as the WMD for this study due to its validation, affordability, and ability to perform various functions comparable to other WMDs. These functions include fall detection, location and activity tracking, blood pressure, pulse, and oxygen saturation monitoring, as well as medication and appointment reminders and telephone capabilities, enhancing the relevance of study findings for real-world implementation. The data collected from ProVista Care could be synchronized and transferred to the server after participants installed the ProVista Care app on their smartphones. The WMD was designed to be worn on the wrist and secured by an elastic band. All participants were encouraged to wear the WMD as frequently and for as long as possible throughout the study period.

The research team developed a computerized decision support model based on the guidelines of the American Heart Association, Hospital Authority, and Department of Health to guide an alert system [30]. When abnormal readings or signs, such as low oxygen levels, irregular pulse rhythm, or falls, are detected by the WMD, a health alert is sent to the smartphone of a nurse case manager for prompt assessment and management. The case manager followed an algorithm developed according to the guidelines from the National Institute for Health and Care Excellence to manage participants and, if necessary, refer them to a community center or a higher level of care, such as the emergency department. For instance, if a participant's heart rate exceeded 100 and they experienced sweating and nausea, the case manager would refer them to the hospital. When deemed necessary, the nurse case manager would refer participants to social workers based on referral guidelines collaboratively developed by the nurse case managers and social workers. Examples of referral reasons include the need for homemaking services due to a high fall risk detected by the WMD, the need for transportation services for those frequently missing medical appointments due to mobility restrictions, and requests for information on fitness classes organized by the community center for participants identified as prehypertensive through daily WMD monitoring.

A 1-hour training session was provided to all participants via Zoom (Zoom Video Communications), covering how to use the WMD, blood pressure monitor, and pulse oximeter; create an account for the ProVista Care app; and synchronize data with their smartphones. Participants were required to pass a practical test demonstrating their ability to operate the devices before they could take the WMD from the community centers. The questions primarily focused on the steps for using the device. Participants were also provided with a telephone number for technical support, available during office hours from community center staff throughout the entire program period.

Intervention Group

Each participant in the intervention group received a 3-month Live With Wearable Monitoring Device program, delivered by CHWs under the supervision of a nurse case manager. Before the program began, 10 CHWs from 5 community centers of a nongovernmental organization, who expressed interest in the program, attended a 1-day training workshop. The training workshop was conducted by the project's principal investigator and a nurse case manager, both with extensive experience in health promotion and older adults' community care. The workshop consisted of 3 sessions: theoretical instruction, role-play, and testing. It combined theoretical knowledge with practical application, incorporating principles of health promotion and adult learning theories to equip CHWs with the competencies needed to effectively support older adults in the community. The first session concentrated on theoretical input, covering essential topics such as the use of the WMD and related apps, an introduction to the assessment tool (the Omaha System), and the application of intervention protocols and referral guidelines. Through interactive presentations and discussions, CHWs gained a comprehensive understanding of these concepts and their practical implications for delivering effective care.

XSL•FO
RenderX

The second session included role-playing activities, where CHWs engaged in simulated case scenarios to apply the knowledge and skills acquired in the theoretical session. This hands-on approach enabled CHWs to practice problem-solving, decision-making, and communication skills in a realistic context, with guidance and feedback from the facilitators.

Finally, all 10 CHWs underwent a competency assessment, consisting of a 10-item multiple-choice test, to ensure they were knowledgeable and skilled enough to provide care within the scope of the program. This assessment validated their learning outcomes and readiness to effectively undertake their roles as CHWs. The questions covered aspects such as understanding the WMD and related apps, familiarity with the Omaha System, application of intervention protocols, and practical implications for delivering effective care.

Participants in the intervention group received a home visit from a CHW and a nurse case manager during the first month, followed by biweekly telephone calls from the CHW from the third to the twelfth week. This arrangement has been validated as feasible for both participants and providers, demonstrating effectiveness in our previous studies [31,32].

The intervention components of this program explicitly focused on facilitating the 3 motivational factors outlined by SDT: autonomy, competence, and relatedness. During the first and only home visit, the CHWs and the nurse case manager visited participants at their homes to provide in-person demonstrations on how to use the devices correctly and to identify any environmental factors that might facilitate or hinder their use of the WMD. During this visit, the nurse case manager engaged participants in a collaborative discussion to explore the features of the WMD that each individual might find beneficial, utilizing the Omaha System. The Omaha System is a comprehensive assessment, intervention, and evaluation tool for community-based practice, covering 42 health and social problems across the domains of environmental, psychological, physiological, and health-related behaviors [33]. After discussions among the research team, the nurse case manager, and the social workers, 21 health and social problems were identified as relevant to, or potentially preventable by, the features of the WMD in this study. For instance, the medication reminder feature could assist participants struggling with medication adherence. The nurse case manager encouraged participants to explore and utilize the features of the WMD that were linked to their identified health and social problems, providing suggestions on usage duration and frequency, as well as instructions for integrating these features into their daily routines. Participants were empowered to adjust or modify their own schedules to ensure efficient and effective utilization of the WMD, with support from the nurse case manager and CHWs. Family members or primary caregivers were also encouraged to participate in discussions, providing feedback and support to enhance the sense of relatedness and social support. During the home visit, CHWs assisted the nurse case manager with health assessments and engaged in collaborative discussions with participants to formulate codeveloped goals and action plans. The CHWs supported, encouraged, and motivated participants to regularly use the WMD, fostering intrinsic motivation and ownership of their health management journey. Subsequent telephone calls were used for follow-ups, progress tracking, and discussions on goal achievement between the CHWs and participants. When necessary, goals and action plans were collaboratively modified with input from participants. Additionally, participants were encouraged to contact the CHWs with any problems or concerns related to the use of the WMD, promoting open communication and ongoing support.

Control Group

Similar to the intervention group participants, those in the control group could utilize the features available on the WMD. When abnormalities, such as high blood pressure, were detected, the nurse case manager followed the intervention protocol and called the participants for treatment and follow-up. However, the services provided were one-off, with no regular or continuous support from CHWs, the nurse case manager, or social workers.

Outcome Measures

Overview

This study measured 3 dimensions of outcomes: usability of the WMD, feasibility, and effectiveness of the program. A description of each measure is presented in the following sections.

Usability of the WMD

Participants were required to complete a questionnaire assessing their attitudes toward using the WMD, perceived usefulness, perceived ease of use, self-efficacy in using the device, anxiety level, and facilitating conditions for using the device. The questionnaire was adapted from Chen and Chan [33]. This 10-point Likert scale ranges from 1 (strongly disagree) to 10 (strongly agree), with higher scores indicating a better outcome, except for the anxiety level, where higher scores reflect a greater level of anxiety associated with using the WMD. The questionnaire has been widely used and validated in prior empirical studies [34,35].

Feasibility of the Program

Feasibility refers to an assessment of the practicality of the study [36]. In this program, feasibility was operationalized through the evaluation of the recruitment rate, attrition rate, incidence of reported adverse events (eg, discomfort, injuries), and incidence of reported technical difficulties (eg, synchronizing data from the WMD to the smartphone app). The recruitment rate was calculated by dividing the number of participants who were recruited and randomized by the number of eligible participants. The attrition rate refers to the number of participants who withdraw or are lost to follow-up. Recruitment and attrition rates, as well as the incidence of reported adverse events and technical difficulties, were recorded by designated staff at each community center and by the nurse.

Effectiveness of the Program

Effectiveness outcomes included continued use intention, adherence rate, quality of life, self-efficacy, and health service utilization. Continued use intention, the primary subjective outcome of this study, was measured using a 3-item, 7-point Likert scale adapted from Bhattacherjee [36] and Windasari et

XSL•FO
RenderX

al [28]. The 7-point scale ranges from 1 (strongly agree) to 5 (strongly disagree), with lower scores indicating a higher likelihood of continuing to use the WMD. This outcome was included to assess the sustained use of the device beyond the study period.

The adherence rate, the objective primary outcome, was measured by recording the number of days participants wore their device and the average time it was worn per day in our database. This information was automatically uploaded to the system whenever participants wore the device. Quality of life was assessed using the Hong Kong version of the EQ-5D-5L [37,38]. The EQ-5D-5L consists of 5 dimensions: mobility, self-care, usual activities, pain or discomfort, and anxiety or depression. The instrument has been widely used globally. Self-efficacy was measured using the Chinese version of the General Self-Efficacy Scale (CGSE) [39]. This scale uses a 4-point Likert format, with higher scores indicating greater levels of self-efficacy in using the WMD. The scale has been extensively validated and is recognized for its high reliability [32]. Health service utilization includes the number of visits to a general practitioner's office, emergency department, hospital, and general outpatient clinic. Participants reported these data, which were then confirmed using medical and attendance certificates, demonstrating good reliability [31].

Background demographics, including age, gender, marital status, educational level, occupation, living status, primary caregiver, frequency of caretaking support, medical history, and eHealth literacy, were collected at baseline to account for potential group differences.

Data Collection

Data collection was conducted at baseline (T0), 1 month (T1), 3 months (T2), and 6 months (T3), with T2 representing the immediate postintervention assessment and T3 measuring the sustained effects of the program. A trained research assistant, who was blinded to group allocation and not involved in either the intervention or control group, called the participants by telephone to collect the data.

Statistical Analysis

Statistical tests were conducted using SPSS version 26 (IBM Corp.). Descriptive analyses were performed to summarize the baseline demographic data, which were presented as mean and SD for continuous variables, median and interquartile range for nonnormally distributed continuous variables, and percentage and frequency for categorical variables. The study used generalized estimating equations to assess differences or changes between the intervention and control groups (between-group effects), within-group (time) effects, and interaction effects (group × time). All outcomes were analyzed using a first-order autoregressive structure. Intention-to-treat analysis was implemented as the primary method for addressing missing data. A result was considered significant when the *P* value (level of significance) was less than 0.05 for a 2-tailed test.

Sample Size

We conducted a power analysis to determine the sample size required for this study. Considering an α value of .05, a power of 80%, and an effect size of 0.72 derived from previous studies using the same primary outcome [40], we concluded that each group should comprise 32 participants. To accommodate a potential dropout rate of 10%, the total sample size for this study was set at 70, with 35 participants assigned to each group.

Ethical Consideration

The ethical application was approved by the Human Ethics Subcommittee of the Hong Kong Polytechnic University (HSEARS20220429001) before the program's commencement. The study is registered at ClinicalTrials.gov (NCT05269303). All eligible participants who joined the program signed a consent form. All documents, including the SPSS file, were encrypted, and only members of the research team, including the research assistant, had access to the password.

Results

Participant Flow

A total of 89 older adults expressed interest and were assessed for eligibility, leading to the recruitment of 76 participants from 5 community centers. They were randomly assigned to either the intervention group (n=39) or the control group (n=37). During the 6-month program, 4 participants from each group dropped out for various reasons, including loss of interest (n=3), dislike of the smartwatch design (n=4), and feeling pressure during the study (n=1; Figure 2 and Multimedia Appendix 1; see also [41]).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

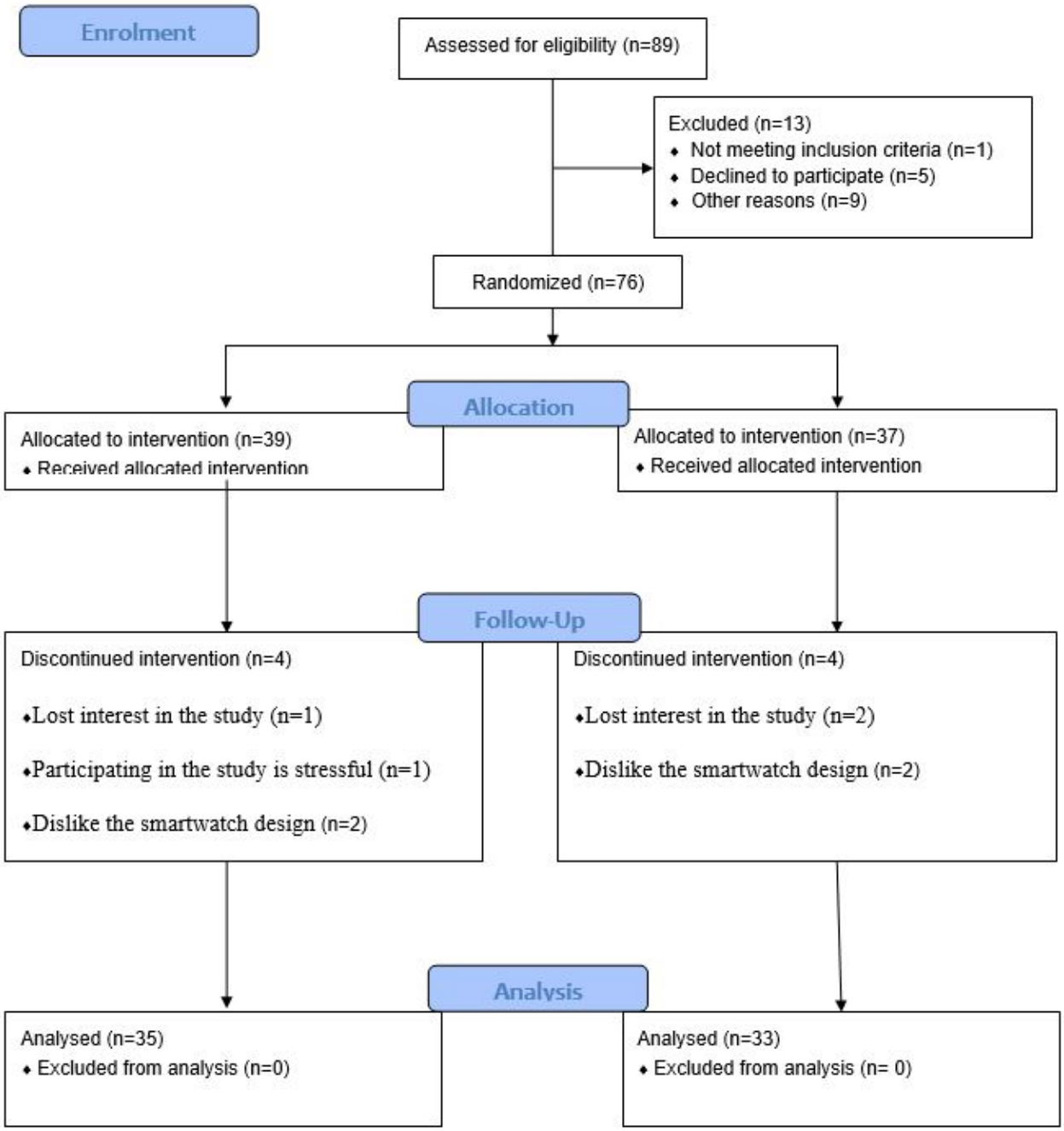

Baseline Characteristics

Baseline demographic characteristics were balanced across the 2 groups (Table 1). Participants (N=76) had an average age of 74 (SD 7.6) years, with only 6 (8%) having no formal education. All but 1 participant were retired, and 30 (39%) lived alone. The majority (69/76, 91%) reported having adequate or more than adequate financial resources, primarily from the old age living allowance (28/76, 37%), family support (26/76, 34%), and personal savings (25/76, 33%). Hypertension (38/76, 50%) and chronic pain (38/76, 50%) were the most prevalent chronic diseases among the participants. Their eHealth literacy level was moderate, with an average score of 25.0 out of 40 on the Chinese version of the eHealth Literacy Scale [42].

XSL•FO
RenderX

Table 1. Demographic characteristics of the participants (N=76).

Demographic characteristics	Total (N=76)	Intervention group (n=39)	Control group (n=37)	*P* value
Gender, n (%)				.88
Male	17 (22)	9 (23)	8 (22)	
Female	59 (78)	30 (77)	29 (78)	
Age (years), mean (SD)	73.74 (7.62)	74.74 (8.16)	72.68 (6.95)	.36
Education level, n (%)				.30
No formal education	6 (8)	2 (5)	4 (11)	
Primary	31 (41)	18 (46)	13 (35)	
Secondary	34 (45)	15 (38)	19 (51)	
Tertiary or above	5 (7)	4 (10)	1 (3)	
Employment status, n (%)				>.99
Part-time	1 (1)	1 (3)	0 (0)	
Retired	75 (99)	38 (97)	37 (100)	
Living status, n (%)				.55
Alone	30 (39)	14 (36)	16 (43)	
With spouse	23 (30)	14 (36)	9 (24)	
With family	23 (30)	11 (28)	12 (32)	
Financial status, n (%)				.91
More than adequate	20 (26)	9 (23)	11 (30)	
Adequate	49 (64)	26 (67)	23 (62)	
Inadequate	5 (7)	3 (8)	2 (5)	
Very inadequate	2 (3)	1 (3)	1 (3)	
Income, n (%)				
Wage	1 (1)	1 (3)	0 (0)	>.99
Family	26 (34)	15 (38)	11 (30)	.42
Savings	25 (33)	14 (36)	11 (30)	.57
Retirement pension	16 (21)	6 (15)	10 (27)	.21
Comprehensive Social Security Assistance	7 (9)	5 (13)	2 (5)	.43
Old age living allowance	28 (37)	16 (41)	12 (32)	.44
Normal disability allowance	1 (1)	0 (0)	1 (3)	.49
Higher disability allowance	2 (3)	0 (0)	2 (5)	.23
Comorbidity, n (%)				
Chronic pain	38 (50)	17 (44)	21 (57)	.25
Chronic obstructive pulmonary disease	5 (7)	3 (8)	2 (5)	>.99
Hypertension	38 (50)	22 (56)	16 (43)	.25
Diabetes	22 (29)	14 (36)	8 (22)	.17
Genital disease	5 (7)	2 (5)	3 (8)	.67
Stroke	3 (4)	2 (5)	1 (3)	>.99
Cancer	8 (11)	4 (10)	4 (11)	>.99
Arthritis	14 (18)	8 (21)	6 (16)	.63
Depression	2 (3)	2 (5)	0 (0)	.49
Cardiac disease	8 (11)	4 (10)	4 (11)	>.99

XSL•FO
RenderX

Demographic characteristics	Total (N=76)	Intervention group (n=39)	Control group (n=37)	*P* value
Cataract	5 (7)	2 (5)	3 (8)	.67
eHealth literacy, mean (SD)	24.96 (7.34)	24.51 (7.71)	25.43 (7.01)	.82

Usability of the WMD

Although the mean values of attitudes toward using the WMD in the intervention group were higher than those in the control group at all 4 time points (Table 2), the between-group differences were not statistically significant (*P*=.42). A significant within-group effect was observed at T2 (β=–1.82, 95% CI –3.21 to –0.43, *P*=.01; Table 3). Similarly, while no statistically significant (*P*=.40) between-group differences were found at any of the 4 time points regarding perceived usefulness, significant within-group effects were demonstrated at T1 (β=–4.70, 95% CI –7.51 to –1.89, *P*=.001) and T2 (β=–4.42, 95% CI –6.98 to –1.85, *P*=.001; Table 3). The mean values of perceived ease of use improved from T0 to T3 in the intervention group (Figure 3). A significant within-group effect was observed at T1 (β=–2.05, 95% CI –3.85 to –0.26, *P*=.02), along with an interaction effect on perceived ease of use (β=2.87, 95% CI 0.56-5.18, *P*=.01) for the intervention group (Table 3). The intervention group recorded the highest mean self-efficacy scores at T3, while the control group achieved its highest mean scores at T0 (Table 2). The Mann-Whitney *U* test indicated that the intervention group had a significantly greater self-efficacy level than the control group at T1 (*z*=–2.05, *P*=.002; Figure 4). Additionally, there was a significant within-group effect on self-efficacy at T1 (β=–2.43, 95% CI –3.78 to –1.07, *P*<.001) and significant interaction effects at T1 (β=3.71, 95% CI 1.76-5.66, *P*<.001) and T3 (β=2.61, 95% CI 0.22-5.00, *P*=.03) for the intervention group (Table 2). The intervention group demonstrated a statistically significant lower anxiety level associated with using the WMD at T3 compared with the control group (*z*=2.10, *P*=.003). The mean anxiety levels in the intervention group also decreased from T0 to T3 (Figure 5). Significant within-group differences in anxiety levels were observed at T1 (β=–2.00, 95% CI –3.67 to –0.32, *P*=.02) and T2 (β=–3.01, 95% CI –4.74 to –1.28, *P*<.001; Table 3). The mean values for facilitating conditions related to the use of the WMD improved in the intervention group; however, no statistically significant (*P*=.16) between-group effect was observed. Significant within-group effects on facilitating conditions were identified at T1 (β=–4.47, 95% CI –7.58 to –1.36, *P*=.005), along with interaction effects at T1 (β=5.44, 95% CI 1.04-9.85, *P*=.01) for the intervention group (Table 3).

XSL•FO
RenderX

Table 2. Mean values of the usability and effectiveness outcomes for the 2 groups at different time points.

Outcomes and groups	Mean	SE	95% Wald CI
Attitude toward using WMD[a]			
Control group			
T3	13.70	0.66	12.39-15.00
T2	12.21	0.77	10.69-13.72
T1	12.63	0.76	11.13-14.12
T0	14.03	0.55	12.94-15.11
Intervention group			
T3	14.09	0.76	12.60-15.57
T2	13.81	0.77	12.30-15.31
T1	14.19	0.68	12.85-15.53
T0	14.64	0.53	13.61-15.67
Perceived usefulness			
Control group			
T3	19.73	1.06	17.64-21.81
T2	17.85	1.13	15.64-20.07
T1	17.57	1.29	15.04-20.10
T0	22.27	0.89	20.53-24.01
Intervention group			
T3	19.97	1.29	17.44-22.50
T2	19.67	1.35	17.03-22.31
T1	18.97	1.13	16.75-21.19
T0	21.28	0.77	19.78-22.78
Perceived ease of use			
Control group			
T3	13.64	0.73	12.21-15.07
T2	14.26	0.64	13.00-15.53
T1	11.51	0.88	9.79-13.24
T0	13.57	0.50	12.59-14.54
Intervention group			
T3	14.26	0.89	12.50-16.01
T2	14.33	0.78	12.80-15.86
T1	13.49	0.69	12.14-14.83
T0	12.67	0.64	11.41-13.92
Anxiety			
Control group			
T3	10.30	0.85	8.64-11.96
T2	7.59	0.80	6.02-9.16
T1	8.60	0.69	7.25-9.95
T0	10.59	0.62	9.38-11.81
Intervention group			
T3	6.86	0.76	5.37-8.34
T2	7.61	0.70	6.24-8.98

XSL•FO
RenderX

Outcomes and groups	Mean	SE	95% Wald CI
T1	7.70	0.66	6.41-8.99
T0	9.64	0.66	8.34-10.94
Self-efficacy of using the device			
Control group			
T3	13.21	0.71	11.83-14.59
T2	13.38	0.56	12.29-14.48
T1	12.11	0.61	10.91-13.31
T0	14.54	0.54	13.48-15.60
Intervention group			
T3	14.49	0.81	12.90-16.07
T2	13.36	0.71	11.97-14.75
T1	14.49	0.60	13.31-15.66
T0	13.21	0.61	12.00-14.41
Facilitating conditions			
Control group			
T3	32.97	1.49	30.05-35.89
T2	31.85	1.41	29.10-34.61
T1	29.77	1.49	26.84-32.70
T0	34.24	1.32	31.65-36.83
Intervention group			
T3	32.71	1.82	29.15-36.28
T2	32.53	1.72	29.16-35.90
T1	32.46	1.44	29.64-35.28
T0	31.49	1.46	28.62-34.35
Continued use intention			
Control group			
T3	15.18	0.70	13.80-16.56
T2	13.71	0.90	11.95-15.46
T1	13.31	0.92	11.52-15.11
Intervention group			
T3	16.23	0.79	14.69-17.77
T2	14.44	0.92	12.64-16.25
T1	14.32	0.74	12.87-15.78
Adherence rate (number of days/week)			
Control group			
T3	3.24	0.498	2.27-4.22
T2	3.91	0.480	2.97-4.85
T1	5.40	0.366	4.68-6.12
Intervention group			
T3	4.86	0.495	3.89-5.83
T2	6.25	0.292	5.68-6.82
T1	6.38	0.266	5.86-6.90
Adherence rate (number of hours/day)			

XSL•FO
RenderX

Outcomes and groups	Mean	SE	95% Wald CI
Control group			
T3	4.48	0.74	3.03-5.94
T2	5.06	0.79	3.52-6.60
T1	6.67	0.92	4.87-8.48
Intervention group			
T3	5.79	0.88	4.06-7.51
T2	9.29	0.82	7.68-10.90
T1	9.68	0.96	7.80-11.55
EQ-5D-5L index value			
Control group			
T3	0.74	0.03	0.69-0.79
T2	0.75	0.02	0.71-0.79
T1	0.72	0.01	0.69-0.75
T0	0.72	0.02	0.69-0.75
Intervention group			
T3	0.79	0.03	0.74-0.84
T2	0.79	0.02	0.75-0.83
T1	0.78	0.02	0.74-0.82
T0	0.75	0.03	0.69-0.81
EQ-5D-5L visual analog scale			
Control group			
T3	75.00	2.033	71.02-78.98
T2	72.79	2.213	68.46-77.13
T1	72.17	2.699	66.88-77.46
T0	68.51	2.563	63.49-73.54
Intervention group			
T3	76.86	2.921	71.13-82.58
T2	74.86	2.668	69.63-80.09
T1	72.27	3.016	66.36-78.18
T0	73.59	2.384	68.92-78.26
The Chinese version of the General Self-Efficacy Scale			
Control group			
T3	24.70	1.19	22.36-27.03
T2	24.74	1.07	22.63-26.84
T1	23.94	0.99	22.00-25.89
T0	24.51	1.04	22.48-26.55
Intervention group			
T3	28.71	0.91	26.93-30.50
T2	27.44	1.06	25.38-29.51
T1	26.30	1.02	24.30-28.30
T0	27.82	1.09	25.69-29.95
Health service utilization			
Control group			

XSL•FO
RenderX

Outcomes and groups	Mean	SE	95% Wald CI
T3	0.48	0.15	0.19-0.78
T2	0.68	0.19	0.30-1.06
T1	0.51	0.16	0.20-0.82
T0	0.57	0.24	0.10-1.04
Intervention group			
T3	0.57	0.19	0.19-0.95
T2	0.53	0.19	0.16-0.89
T1	0.86	0.26	0.36-1.37
T0	1.10	0.36	0.40-1.81

[a]WMD: wearable monitoring device.

XSL•FO
RenderX

Table 3. Parameter estimates of outcomes.

Parameter estimates	β	SE	95% CI	Wald χ^2 (*df*)	*P* value[a]
Attitude toward using WMD[b]					
Intercept	14.027	0.5532	12.943 to 15.111	642.908 (1)	<.001
Intervention group	0.614	0.7643	–0.884 to 2.112	0.645 (1)	.44
Time=3	–0.330	0.7988	–1.896 to 1.236	0.171 (1)	.68
Time=2	–1.821	0.7104	–3.213 to –0.429	6.572 (1)	.01
Time=1	–1.398	0.9149	–3.192 to 0.395	2.336 (1)	.13
Intervention group × time=3	–0.225	1.1350	–2.450 to 1.999	0.039 (1)	.84
Intervention group × time=2	0.986	1.0725	–1.116 to 3.088	0.845 (1)	.36
Intervention group × time=1	0.947	1.1654	–1.338 to 3.231	0.660 (1)	.42
Perceived usefulness					
Intercept	22.270	0.8879	20.530 to 24.011	629.091 (1)	<.001
Intervention group	–0.988	1.1725	–3.286 to 1.310	0.710 (1)	.40
Time=3	–2.543	1.4385	–5.362 to 0.276	3.125 (1)	.08
Time=2	–4.417	1.3092	–6.983 to –1.851	11.384 (1)	<.001
Time=1	–4.699	1.4354	–7.512 to –1.885	10.716 (1)	.001
Intervention group × time=3	1.232	2.1029	–2.889 to 5.354	0.343 (1)	.56
Intervention group × time=2	2.802	1.8610	–0.845 to 6.449	2.267 (1)	.13
Intervention group × time=1	2.390	1.8589	–1.254 to 6.033	1.653 (1)	.20
Ease of use					
Intercept	13.568	0.4977	12.592 to 14.543	743.197 (1)	<.001
Intervention group	–0.901	0.8114	–2.491 to 0.689	1.233 (1)	.27
Time=3	0.069	0.7383	–1.378 to 1.516	0.009 (1)	.93
Time=2	0.697	0.6013	–0.481 to 1.876	1.344 (1)	.25
Time=1	–2.053	0.9168	–3.850 to –0.256	5.016 (1)	.02
Intervention group × time=3	1.522	1.1784	–0.788 to 3.831	1.667 (1)	.20
Intervention group × time=2	0.970	0.9902	–0.971 to 2.910	0.959 (1)	.33
Intervention group × time=1	2.873	1.1789	0.563 to 5.184	5.940 (1)	.01
Self-efficacy of using the device					
Intercept	14.541	0.5386	13.485 to 15.596	728.731 (1)	<.001
Intervention group	–1.335	0.8164	–2.935 to 0.265	2.676 (1)	.10
Time=3	–1.328	0.8231	–2.942 to 0.285	2.605 (1)	.10
Time=2	–1.158	0.7177	–2.565 to 0.249	2.604 (1)	.10
Time=1	–2.426	0.6930	–3.784 to –1.068	12.259 (1)	<.001
Intervention group × time=3	2.609	1.2178	0.222 to 4.996	4.590 (1)	.03
Intervention group × time=2	1.314	1.0201	–0.685 to 3.313	1.660 (1)	.20
Intervention group × time=1	3.708	0.9949	1.758 to 5.658	13.888 (1)	<.001
Anxiety					
Intercept	10.595	0.6204	9.379 to 11.811	291.627 (1)	<.001
Intervention group	–0.954	0.9089	–2.735 to 0.828	1.101 (1)	.29
Time=3	–0.292	0.9682	–2.189 to 1.606	0.091 (1)	.76
Time=2	–3.006	0.8830	–4.737 to –1.276	11.592 (1)	<.001

XSL•FO
RenderX

Parameter estimates	β	SE	95% CI	Wald χ^2 (*df*)	*P* value[a]
Time=1	–1.995	0.8555	–3.671 to –0.318	5.436 (1)	*.02*
Intervention group × time=3	–2.492	1.3337	–5.106 to 0.122	3.492 (1)	.06
Intervention group × time=2	0.976	1.3199	–1.611 to 3.563	0.547 (1)	.46
Intervention group × time=1	0.056	1.2547	–2.403 to 2.515	0.002 (1)	.96
Facilitating conditions					
Intercept	34.243	1.3215	31.653 to 36.833	671.441 (1)	*<.001*
Intervention group	–2.756	1.9695	–6.616 to 1.104	1.958 (1)	.16
Time=3	–1.274	1.6801	–4.566 to 2.019	0.575 (1)	.45
Time=2	–2.390	1.5979	–5.522 to 0.742	2.238 (1)	.13
Time=1	–4.472	1.5865	–7.581 to –1.362	7.945 (1)	*.005*
Intervention group × time=3	2.501	2.6279	–2.650 to 7.651	0.905 (1)	.34
Intervention group × time=2	3.431	2.4679	–1.406 to 8.268	1.933 (1)	.16
Intervention group × time=1	5.444	2.2479	1.038 to 9.850	5.865 (1)	*.01*
Continued use intention					
Intercept	13.314	0.9168	11.517 to 15.111	210.925 (1)	*<.001*
Intervention group	1.010	1.1789	–1.301 to 3.321	0.734 (1)	.39
Time=3	1.868	0.9286	0.048 to 3.687	4.045 (1)	*.04*
Time=2	0.392	0.7277	–1.035 to 1.818	0.290 (1)	.59
Intervention group × time=3	0.037	1.2013	–2.318 to 2.391	0.001 (1)	.98
Intervention group × time=2	–0.271	0.9793	–2.191 to 1.648	0.077 (1)	.78
Adherence rate (number of days/week)					
Intercept	5.400	0.3663	4.682 to 6.118	217.281 (1)	*<.001*
Intervention group	0.978	0.4526	0.091 to 1.865	4.673 (1)	*.03*
Time=3	–2.158	0.5265	–3.190 to –1.126	16.790 (1)	*<.001*
Time=2	–1.488	0.4353	–2.341 to –0.635	11.690 (1)	*<.001*
Intervention group × time=3	0.636	0.7397	–0.813 to 2.086	0.740 (1)	.39
Intervention group × time=2	1.360	0.5964	0.191 to 2.529	5.199 (1)	*.02*
Adherence rate (number of hours/day)					
Intercept	6.671	0.9205	4.867 to 8.476	52.528 (1)	*<.001*
Intervention group	3.004	1.3286	0.400 to 5.608	5.113 (1)	*.02*
Time=3	–2.187	0.9280	–4.005 to –0.368	5.552 (1)	*.02*
Time=2	–1.613	0.7805	–3.142 to –0.083	4.269 (1)	*.04*
Intervention group × time=3	–1.703	1.4599	–4.565 to 1.158	1.361 (1)	.24
Intervention group × time=2	1.229	1.0430	–0.816 to 3.273	1.388 (1)	.24
EQ-5D-5L index value					
Intercept	0.719	0.0162	0.687 to 0.751	1959.643 (1)	*<.001*
Intervention group	0.029	0.0345	–0.038 to 0.097	0.715 (1)	.40
Time=3	0.018	0.0190	–0.019 to 0.055	0.898 (1)	.34
Time=2	0.029	0.0192	–0.008 to 0.067	2.316 (1)	.13
Time=1	0.004	0.0121	–0.020 to 0.027	0.098 (1)	.75
Intervention group × time=3	0.028	0.0354	–0.042 to 0.097	0.605 (1)	.44
Intervention group × time=2	0.014	0.0338	–0.052 to 0.080	0.175 (1)	.68

XSL•FO
RenderX

Parameter estimates	β	SE	95% CI	Wald χ^2 (*df*)	*P* value[a]
Intervention group × time=1	0.028	0.0289	–0.028 to 0.085	0.972 (1)	.32
EQ-5D-5L visual analog scale					
Intercept	68.514	2.5630	63.490 to 73.537	714.561 (1)	*<.001*
Intervention group	5.076	3.5007	–1.785 to 11.937	2.103 (1)	.15
Time=3	6.486	2.5302	1.527 to 11.445	6.572 (1)	*.01*
Time=2	4.281	2.0782	0.207 to 8.354	4.243 (1)	*.04*
Time=1	3.658	3.2388	–2.690 to 10.006	1.276 (1)	.26
Intervention group × time=3	–3.219	3.8121	–10.691 to 4.253	0.713 (1)	.40
Intervention group × time=2	–3.009	3.0391	–8.966 to 2.947	0.980 (1)	.32
Intervention group × time=1	–4.977	3.7929	–12.411 to 2.457	1.722 (1)	.19
The Chinese version of the General Self-Efficacy Scale					
Intercept	24.514	1.0395	22.476 to 26.551	556.128 (1)	*<.001*
Intervention group	3.307	1.5033	0.361 to 6.253	4.839 (1)	*.03*
Time=3	0.183	1.0463	–1.867 to 2.234	0.031 (1)	.86
Time=2	0.222	0.9018	–1.546 to 1.989	0.060 (1)	.81
Time=1	–0.571	0.9197	–2.373 to 1.232	0.385 (1)	.53
Intervention group × time=3	0.710	1.5155	–2.260 to 3.681	0.220 (1)	.64
Intervention group × time=2	–0.598	1.4507	–3.441 to 2.245	0.170 (1)	.68
Intervention group × time=1	–0.953	1.3783	–3.654 to 1.749	0.478 (1)	.49
Health service utilization					
Intercept	0.568	0.2403	0.097 to 1.039	5.577 (1)	*.02*
Intervention group	0.535	0.4317	–0.311 to 1.381	1.536 (1)	.21
Time=3	–0.083	0.2838	–0.639 to 0.473	0.085 (1)	.77
Time=2	0.109	0.2282	–0.338 to 0.556	0.228 (1)	.63
Time=1	–0.053	0.2369	–0.518 to 0.411	0.051 (1)	.82
Intervention group × time=3	–0.448	0.4665	–1.363 to 0.466	0.924 (1)	.34
Intervention group × time=2	–0.684	0.4027	–1.473 to 0.106	2.882 (1)	.09
Intervention group × time=1	–0.184	0.4097	–0.987 to 0.619	0.203 (1)	.65

[a]Italicized values are significant at $P<.05$.

[b]WMD: wearable monitoring device.

XSL•FO
RenderX

Figure 3. Perceived ease of use scores of the 2 groups over time.

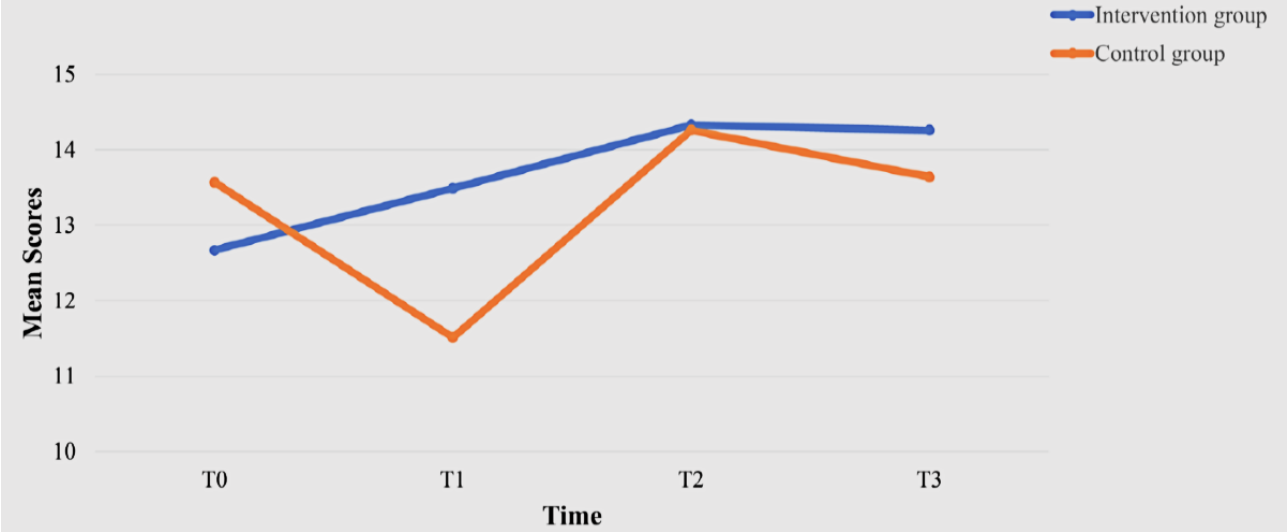

Figure 4. Self-efficacy scores for device use in the 2 groups over time.

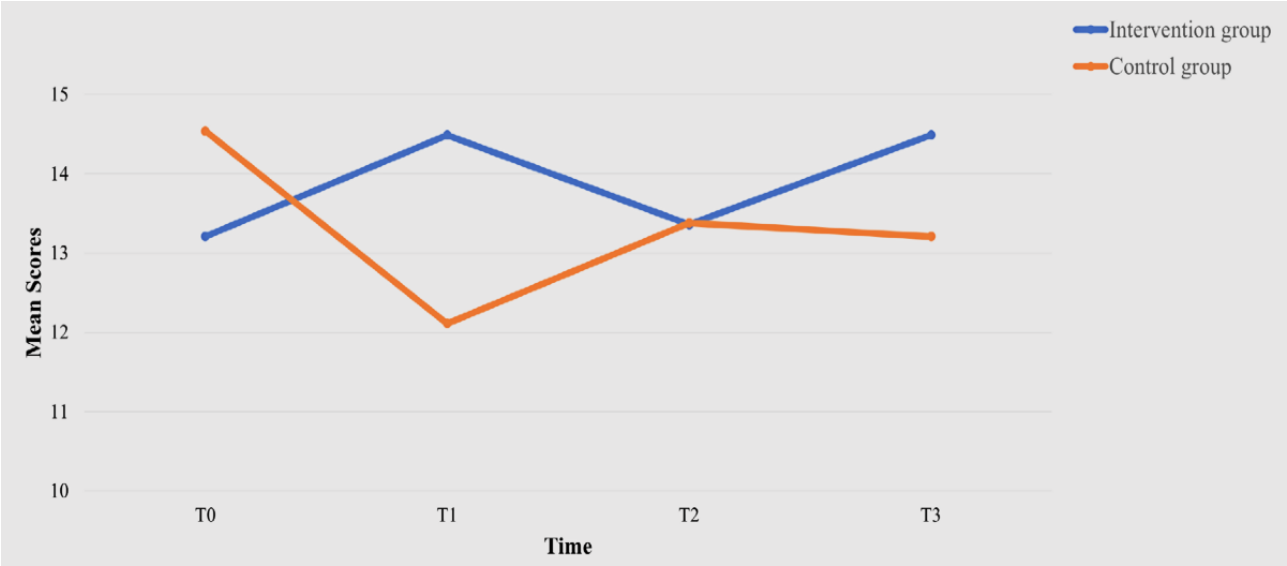

Figure 5. Anxiety scores of the 2 groups over time.

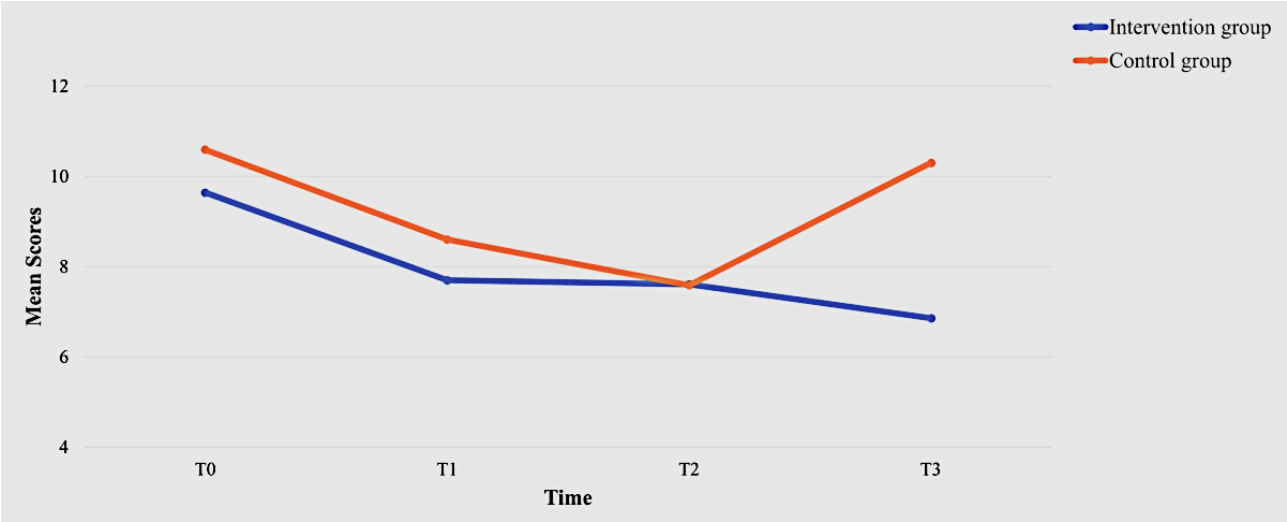

XSL•FO
RenderX

Feasibility of the Program

The recruitment rate for this program was high, with only 13 potential participants (N=89, 15%) deemed ineligible. Reasons for ineligibility included not meeting the inclusion criteria (n=1), declining to participate (n=5), and citing other reasons such as eczema or unavailability (n=7). The attrition rate was low, with only 11% (8/76) of participants withdrawing from the study. Throughout the 6-month program, only 1 instance of a skin problem was reported. The 3 most frequently reported technical problems by participants were smartwatch malfunctions (15/96, 16%), inability to upload vital signs to the mobile app (13/96, 14%), and pace counter malfunctions (10/96, 10%).

Effectiveness of the Program

Both groups showed an improvement in continued intention to use the WMD from T1 to T3 (Table 2), resulting in no statistically significant (*P*=.39) between-group differences. However, a within-group effect was observed at T3 (β=1.87, 95% CI of β 0.05-3.69, *P*=.04; Table 3). Regarding adherence rates, the intervention group demonstrated more frequent use of the smartwatch in terms of both days and hours. Statistically significant between-group differences were noted at all time points: T1 (*P*=.01), T2 (*P*<.001), and T3 (*P*=.02; Figure 6). Although the mean EQ-5D-5L index value of the intervention group improved from T0 to T3 and was higher than that of the control group at each time point (Table 2), no statistically significant between-group (*P*=.40), within-group (*P*=.75 for T1-T2, *P*=.12 for T1-T3, and *P*=.34 for T1-T4), or interaction effects (*P*=.32 for intervention × time=2; *P*=.68 for intervention × time=3; *P*=.44 for intervention × time=4) were observed. However, within-group effects on the EQ-5D-5L visual analog scale were noted at T2 (β=4.28, 95% CI of β 0.21-8.35, *P*=.04) and T3 (β=6.49, 95% CI of β 1.53-11.45, *P*=.01; Table 3). A statistically significant between-group effect was observed in self-efficacy, with the intervention group demonstrating a higher CGSE score (β=3.31, 95% CI of β 0.36-6.25, *P*=.03; Figure 7). However, no significant between-group (*P*=.21), within-group (*P*=.82 for T1-T2, *P*=.63 for T1-T3, and *P*=.77 for T1-T4), or interaction effects (*P*=.65 for intervention × time=2; *P*=.09 for intervention × time=3; *P*=.34 for intervention × time=4) were found regarding health service utilization.

Figure 6. Adherence rate of the 2 groups over time.

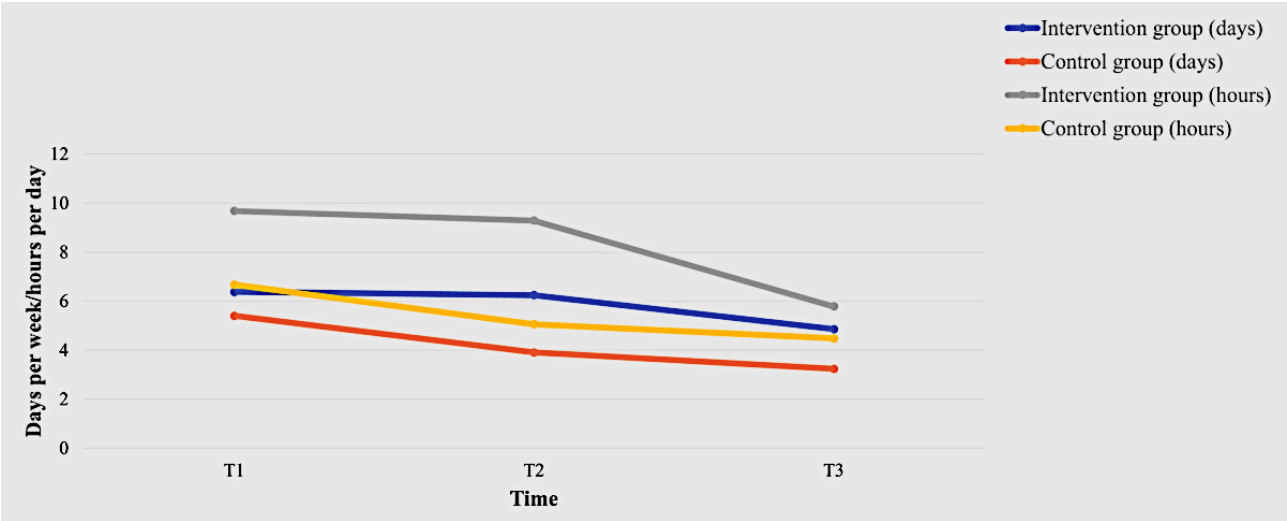

Figure 7. The Chinese version of the General Self-Efficacy Scale (CGSE) scores of the 2 groups over time.

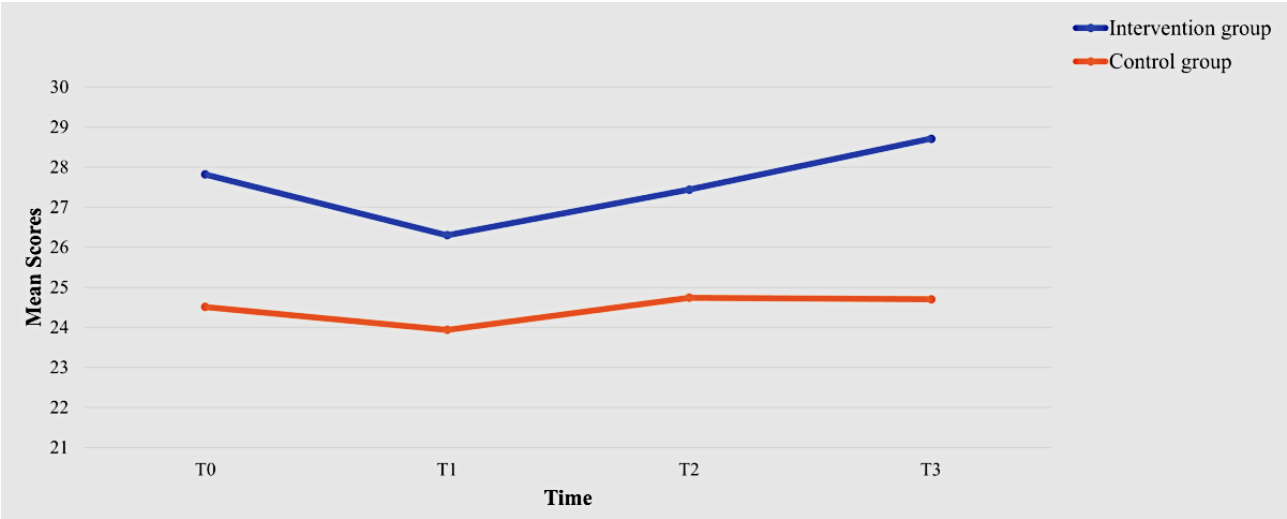

XSL•FO
RenderX

Discussion

Principal Findings and Comparison With Previous Work

WMDs are increasingly recognized as valuable tools for ongoing health monitoring and the remote delivery of health care services. Despite their potential to support self-management across diverse populations, concerns about their sustained use persist. Therefore, this novel study investigates whether ongoing support from CHWs and other health care providers can enhance the usability, feasibility, and effectiveness of WMDs among community-dwelling older adults. Over a 6-month period, the study findings indicate that participants in the intervention group exhibited better attitudes toward the WMD, improved perceived ease of use, higher self-efficacy, and lower anxiety levels compared with the control group. Despite these positive outcomes, it was noted that the intention to continue using the WMD improved in both groups, with participants in the intervention group using the device more frequently. Additionally, while the quality-of-life index improved in the intervention group, the changes were not statistically significant, with only between-group effects observed regarding self-efficacy. Nonsignificant findings were noted for health service utilization across both groups. These results suggest that supplementary professional support may enhance certain outcomes related to the use of WMDs among older adults. Therefore, maintaining long-term access to such support could be essential for ensuring the ongoing use of WMDs.

Usability often reflects the ease and accessibility with which a user can interact with an app or device, serving as a significant determinant for the adoption of WMDs [43]. In this study, participants in the intervention group exhibited generally higher scores in attitude, perceived ease of use, and self-efficacy compared with those in the control group. Additionally, participants in the intervention group reported lower anxiety scores compared with those in the control group. These findings suggest that the ongoing support provided by CHWs may have empowered participants in the intervention group to develop the skills necessary for effectively navigating the WMDs and the monitoring system. This support likely contributed to achieving competency, as outlined in SDT [44]. In fact, the sustained support provided an ongoing learning opportunity for older adults as they engaged with a new device. This support may have facilitated their transition from novices to more proficient users, enhancing their ability to navigate the system and troubleshoot when necessary. This notion reflects an increase in autonomy, aligning with the principles of SDT [45]. Recent studies investigating the factors influencing the use of activity-tracking wearable devices among older adults have suggested that social influence plays a critical role within their social context [46,47]. Moreover, as participants received support and education to use the wearable devices, they began to view them more positively. This shift occurred as they not only learned to navigate the system but also experienced the devices' benefits, integrating them into their daily lives [46,48]. Yildirim and Ali-Eldin [48] observed that the perceived usefulness of a wearable device is the strongest motivator for its continued use. Over time, participants' attitudes and perceptions toward these devices evolved as their ease of use gradually increased [49,50]. This insight emphasizes the importance of providing ongoing support to sustain the use of wearable devices, especially during the initial stages of implementation. As noted in several existing studies, limited engagement during this early phase can lead to frustration and, ultimately, the abandonment of the devices altogether [48,50].

In terms of the intervention's effectiveness, our study found that the intention to continue using the WMDs improved from T1 to T3 in both groups. While this is a positive outcome, it is important to note that intention to use may not always translate into actual usage [51]. This distinction is particularly relevant in this study, as although both groups showed an increased intention to use the WMDs, participants in the intervention group utilized the smartwatches more frequently than those in the control group. With the continued presence of social influence from CHWs, older adults in the intervention group may have been motivated to persist in using the WMDs [43]. This underscores the essential need for ongoing education to support older adults in their use of wearable devices. Therefore, it is insufficient to merely provide access to WMDs; ongoing support is necessary to ensure their effective utilization.

Despite the comprehensive nature of the intervention, which included ongoing support, no statistically significant differences were found concerning quality of life and health service utilization. Nonetheless, an improvement in the quality-of-life index value was observed. Additionally, only between-group effects were noted for self-efficacy, with the intervention group demonstrating higher CGSE scores. Undoubtedly, digital-based monitoring and cost-efficient smart wearable technologies for physical activity tracking represent an innovative platform for health promotion, allowing for the proactive identification and resolution of emerging biopsychosocial issues [52]. These technologies also facilitate the promotion of a healthy lifestyle among end users. In this study, the use of the Omaha System proved particularly beneficial for identifying, managing, and evaluating biopsychosocial and environmental problems, thereby promoting comprehensive health. Thus, the nonsignificant findings regarding quality of life present an intriguing aspect of the study. The inclusion of older adults with varying underlying chronic illnesses may warrant further consideration. As indicated in the "Results" section, many participants were dealing with chronic conditions such as pain, hypertension, diabetes, and arthritis, which can significantly affect quality of life. Factors contributing to a diminished quality of life often include a reduced ability to engage in practical and social activities, with chronic pain being a notable influence [53]. This highlights the possibility that some participants may require a higher dose or intensity of the intervention than others, warranting further investigation. Future trials should consider adjusting the intervention dose to meet the unique needs of participants, ensuring a personalized approach to implementing WMDs. The statistically nonsignificant findings regarding health service utilization may indicate comparable levels of utilization across both the intervention and control groups. This could be attributed to the WMD's design, which includes a monitoring platform intended to detect and respond to health-related issues.

XSL•FO
RenderX

The feasibility outcomes, characterized by high recruitment, retention, and low attrition rates, are noteworthy in this study, especially given the 6-month duration. While it remains unclear what happens to WMD usage after the intervention concludes, it can be inferred that most participants utilized the WMD throughout the entire study period. Many individuals tend to abandon WMDs shortly after starting to use them, highlighting the significance of this finding [54,55]. In fact, most users tend to abandon WMDs within 3-6 months following their purchase [55-57]. In this study, participants in both the intervention and control groups received the same initial support; however, those in the intervention group benefited from ongoing assistance. Moreover, any abnormalities noted in the monitoring records of participants in the control group were actively followed up by the nurse case manager, which may have fostered a social effect by reassuring participants that support was available. This social effect and initial engagement may have stimulated further interest in the WMD and its potential benefits, thereby sustaining participants' interest regardless of their assigned group [43]. Despite these assertions, a qualitative exploration may be warranted to capture participants' experiences and gain an in-depth understanding of why they continue to use the WMD. Such a study could complement the findings from this trial, shedding light on the factors that motivate older adults to adopt and maintain their use of WMDs.

Limitations

Although the study provides valuable insights into the utilization of WMDs among older adults, several limitations warrant attention. First, the study utilized a specific WMD designed for ongoing monitoring. Given the diverse range of WMDs available on the market, exploring other types of devices may yield different outcomes worth investigating further. Second, while the intention to continue using the app was assessed, it remains uncertain whether participants will actually use the WMD beyond the 6-month study period. Future studies should consider extending the duration of research to better understand the long-term actual usage of WMDs among older adults. Applying an implementation science approach grounded in a critical realist perspective may help uncover the generative mechanisms that explain the sustained usage of WMDs in this population. Additionally, the study did not involve older adults in the development process, which may limit how well the intervention aligns with their perspectives and preferences.

Conclusion

WMDs have the potential to support self-management among community-dwelling older adults; however, they may not be sufficient on their own, necessitating ongoing support from health care professionals. This ongoing support is essential for building competence and autonomy, as well as fostering relatedness among older adults, in accordance with SDT. Future studies should consider customizing the dose or intensity of the intervention to meet the individual needs of patients, rather than adopting a one-size-fits-all approach.

Acknowledgments

The study was funded by the Departmental General Research Fund, The Hong Kong Polytechnic University (grant G-UAQ2).

Data Availability

The data presented in this study are available upon reasonable request from the corresponding author.

Authors' Contributions

AKCW and FKYW conceptualized the program. AKCW, JJS, FKYW, VH, KKSC, BPW, and SMW provided intellectual input on the study design, methodology, and evaluation. AKCW and JB drafted the manuscript. All authors contributed to, reviewed, and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[DOC File , 220 KB - jmir_v26i1e52435_app1.doc]

References

1. Howcroft J, Kofman J, Lemaire ED, McIlroy WE. Analysis of dual-task elderly gait in fallers and non-fallers using wearable sensors. J Biomech 2016 May 03;49(7):992-1001. [doi: 10.1016/j.jbiomech.2016.01.015] [Medline: 26994786]
2. Gao W, Emaminejad S, Nyein HYY, Challa S, Chen K, Peck A, et al. Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. Nature 2016 Jan 28;529(7587):509-514 [FREE Full text] [doi: 10.1038/nature16521] [Medline: 26819044]

XSL•FO
RenderX

3. Baig MM, GholamHosseini H, Moqeem AA, Mirza F, Lindén M. A systematic review of wearable patient monitoring systems - current challenges and opportunities for clinical adoption. J Med Syst 2017 Jul 19;41(7):115-115. [doi: 10.1007/s10916-017-0760-1] [Medline: 28631139]
4. Sprogis SK, Currey J, Considine J. Patient acceptability of wearable vital sign monitoring technologies in the acute care setting: a systematic review. Journal of Clinical Nursing 2019 May 09;28(15-16):2732-2744. [doi: 10.1111/jocn.14893] [Medline: 31017338]
5. Rogerson L, Burr J, Tyson S. The feasibility and acceptability of smart home technology using the Howz system for people with stroke. Disabil Rehabil Assist Technol 2020 Feb 21;15(2):148-152. [doi: 10.1080/17483107.2018.1541103] [Medline: 30663921]
6. Bonacaro A, Rubbi I, Sookhoo D. The use of wearable devices in preventing hospital readmission and in improving the quality of life of chronic patients in the homecare setting: a narrative literature review. Prof Inferm 2019;72(2):143-151. [Medline: 31550431]
7. Lee J, Kim D, Ryoo H, Shin B. Sustainable wearables: wearable technology for enhancing the quality of human life. Sustainability 2016 May 11;8(5):466. [doi: 10.3390/su8050466]
8. Bove LA. Increasing patient engagement through the use of wearable technology. The Journal for Nurse Practitioners 2019 Sep;15(8):535-539. [doi: 10.1016/j.nurpra.2019.03.018]
9. Sifaoui A, Eastin M. “Whispers from the wrist”: wearable health monitoring devices and privacy regulations in the U.S.: the loopholes, the challenges, and the opportunities. Cryptography 2024 Jun 19;8(2):26. [doi: 10.3390/cryptography8020026]
10. Wearable fitness trackers market (2013-2023). Report no: LS11415. 2018. URL: https://www.psmarketresearch.com/market-analysis/wearable-fitness-trackers-market [accessed 2023-03-18]
11. Attig C, Franke T. Abandonment of personal quantification: a review and empirical study investigating reasons for wearable activity tracking attrition. Computers in Human Behavior 2020 Jan;102:223-237. [doi: 10.1016/j.chb.2019.08.025]
12. Deranek K, Hewitt B, Gudi A, McLeod A. The impact of exercise motives on adolescents’ sustained use of wearable technology. Behaviour & Information Technology 2020 Feb 01;40(7):691-705. [doi: 10.1080/0144929x.2020.1720295]
13. Gartner survey shows wearable device need to be more useful. Gartner. 2016. URL: https://www.gartner.com/en/newsroom/press-releases/2016-12-07-gartner-survey-shows-wearable-devices-need-to-be-more-useful [accessed 2022-02-01]
14. Stragier J, Vanden Abeele M, Mechant P, De Marez L. Understanding persistence in the use of online fitness communities: comparing novice and experienced users. Computers in Human Behavior 2016 Nov;64:34-42. [doi: 10.1016/j.chb.2016.06.013]
15. Gulotta R, Forlizzi J, Yang R, Newman M. Fostering engagement with personal informatics systems. In: Proceedings of the 2016 ACM Conference on Designing Interactive Systems. New York, NY: ACM; 2016 Presented at: 2016 ACM Conference on Designing Interactive Systems; June 4-8, 2016; Brisbane, QLD, Australia p. 286-300. [doi: 10.1145/2901790.2901803]
16. Shin D. How do technological properties influence user affordance of wearable technologies? IS 2019 Oct 7;20(2):307-338. [doi: 10.1075/is.16024.shi]
17. Shin D, Biocca F. Impact of social influence and users' perception of coolness on smartwatch behavior. Soc Behav Pers 2018 Jun 05;46(6):881-890. [doi: 10.2224/sbp.5134]
18. Wong AKC, Bayuo J, Wong FKY, Chow KKS, Wong SM, Wong BB, et al. Experiences of receiving an mHealth application with proactive nursing support among community-dwelling older adults: a mixed-methods study. BMC Nurs 2024 Apr 07;23(1):232-238 [FREE Full text] [doi: 10.1186/s12912-024-01909-w] [Medline: 38584270]
19. World Health Organization (WHO). Daft Global Strategy on Digital Health. Geneva, Switzerland: World Health Organization; 2021.
20. Durant RW, Brown QL, Cherrington AL, Andreae LJ, Hardy CM, Scarinci IC. Social support among African Americans with heart failure: is there a role for community health advisors? Heart Lung 2013 Jan;42(1):19-25 [FREE Full text] [doi: 10.1016/j.hrtlng.2012.06.007] [Medline: 22920609]
21. Taylor B, Mathers J, Parry J. A conceptual framework for understanding the mechanism of action of community health workers services: the centrality of social support. J Public Health (Oxf) 2019 Mar 01;41(1):138-148. [doi: 10.1093/pubmed/fdx161] [Medline: 29228321]
22. Collinsworth AW, Vulimiri M, Schmidt KL, Snead CA. Effectiveness of a community health worker-led diabetes self-management education program and implications for CHW involvement in care coordination strategies. Diabetes Educ 2013 Sep 19;39(6):792-799. [doi: 10.1177/0145721713504470] [Medline: 24052203]
23. Turner BJ, Liang Y, Ramachandran A, Poursani R. Telephone or visit-based community health worker care management for uncontrolled diabetes mellitus: a longitudinal study. J Community Health 2020 Dec 29;45(6):1123-1131 [FREE Full text] [doi: 10.1007/s10900-020-00849-1] [Medline: 32472457]
24. Enard KR, Ganelin DM. Reducing preventable emergency department utilization and costs by using community health workers as patient navigators. Journal of Healthcare Management 2013;58(6):412-427. [doi: 10.1097/00115514-201311000-00007]

XSL•FO
RenderX

25. Kangovi S, Mitra N, Grande D, Huo H, Smith RA, Long JA. Community health worker support for disadvantaged patients with multiple chronic diseases: a randomized clinical trial. Am J Public Health 2017 Oct;107(10):1660-1667. [doi: 10.2105/ajph.2017.303985]
26. Lin F, Windasari NA. Continued use of wearables for wellbeing with a cultural probe. The Service Industries Journal 2018 Aug 06;39(15-16):1140-1166. [doi: 10.1080/02642069.2018.1504924]
27. Zaslavsky O, Thompson HJ, McCurry SM, Landis CA, Kitsiou S, Ward TM, et al. Use of a wearable technology and motivational interviews to improve sleep in older adults with osteoarthritis and sleep disturbance: a pilot study. Res Gerontol Nurs 2019 Jul 01;12(4):167-173 [FREE Full text] [doi: 10.3928/19404921-20190319-02] [Medline: 30901479]
28. Windasari NA, Lin F, Kato-Lin Y. Continued use of wearable fitness technology: a value co-creation perspective. International Journal of Information Management 2021 Apr;57:102292. [doi: 10.1016/j.ijinfomgt.2020.102292]
29. Ryan R, Deci E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78 [FREE Full text] [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
30. Wong A, Wong F, Chang K. A proactive mobile health application program for promoting self-care health management among older adults in the community: study protocol of a three-arm randomized controlled trial. Gerontology 2020;66(5):506-513. [doi: 10.1159/000509129] [Medline: 32772019]
31. Wong AKC, Wong FKY, Chow KKS, Wong SM, Lee PH. Effect of a telecare case management program for older adults who are homebound during the COVID-19 pandemic: a pilot randomized clinical trial. JAMA Netw Open 2021 Sep 01;4(9):e2123453 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.23453] [Medline: 34499135]
32. Martin K. The Omaha System: A Key to Practice, Documentation, and Information Management. Omaha, NE: Health Connections Press; 2005.
33. Chen K, Chan AHS. Gerontechnology acceptance by elderly Hong Kong Chinese: a senior technology acceptance model (STAM). Ergonomics 2014 Mar 24;57(5):635-652. [doi: 10.1080/00140139.2014.895855] [Medline: 24655221]
34. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
35. McLeod S. Feasibility studies for novel and complex projects: principles synthesised through an integrative review. Project Leadership and Society 2021 Dec;2:100022. [doi: 10.1016/j.plas.2021.100022]
36. Bhattacherjee A. Understanding information systems continuance: an expectation-confirmation model. MIS Quarterly 2001 Sep;25(3):351. [doi: 10.2307/3250921]
37. Wong EL, Cheung AW, Wong AY, Xu RH, Ramos-Goñi JM, Rivero-Arias O. Normative profile of health-related quality of life for Hong Kong general population using preference-based instrument EQ-5D-5L. Value Health 2019 Aug;22(8):916-924 [FREE Full text] [doi: 10.1016/j.jval.2019.02.014] [Medline: 31426933]
38. Leung DY, Leung AY. Factor structure and gender invariance of the Chinese General Self-Efficacy Scale among soon-to-be-aged adults. J Adv Nurs 2011 Jun 05;67(6):1383-1392. [doi: 10.1111/j.1365-2648.2010.05529.x] [Medline: 21129011]
39. Liao J, Xiao H, Li X, Sun S, Liu S, Yang Y, et al. A social group-based information-motivation-behavior skill intervention to promote acceptability and adoption of wearable activity trackers among middle-aged and older adults: cluster randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 09;8(4):e14969 [FREE Full text] [doi: 10.2196/14969] [Medline: 32271151]
40. Keogh A, Dorn JF, Walsh L, Calvo F, Caulfield B. Comparing the usability and acceptability of wearable sensors among older Irish adults in a real-world context: observational study. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e15704 [FREE Full text] [doi: 10.2196/15704] [Medline: 32310149]
41. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010 Mar 24;8:18 [FREE Full text] [doi: 10.1186/1741-7015-8-18] [Medline: 20334633]
42. Ma Z, Wu M. The psychometric properties of the Chinese eHealth Literacy Scale (C-eHEALS) in a Chinese rural population: cross-sectional validation study. J Med Internet Res 2019 Oct 22;21(10):e15720 [FREE Full text] [doi: 10.2196/15720] [Medline: 31642811]
43. Deci EL, Ryan RM. Self-determination theory. In: Van Lange PAW, Kruglanski AW, Higgins ET, editors. Handbook of Theories of Social Psychology. Los Angeles, CA: Sage Publications; 2012:416-436.
44. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY: The Guilford Press; 2017.
45. Abouzahra M, Ghasemaghaei M. The antecedents and results of seniors' use of activity tracking wearable devices. Health Policy and Technology 2020 Jun;9(2):213-217 [FREE Full text] [doi: 10.1016/j.hlpt.2019.11.002]
46. Gopinath K, Selvam G, Narayanamurthy G. Determinants of the adoption of wearable devices for health and fitness: a meta-analytical study. CAIS 2022;50(1):445-450. [doi: 10.17705/1cais.05019]
47. Gupta A, Dhiman N, Yousaf A, Arora N. Social comparison and continuance intention of smart fitness wearables: an extended expectation confirmation theory perspective. Behaviour & Information Technology 2020 Apr 03;40(13):1341-1354. [doi: 10.1080/0144929x.2020.1748715]
48. Yildirim H, Ali-Eldin AM. A model for predicting user intention to use wearable IoT devices at the workplace. Journal of King Saud University - Computer and Information Sciences 2019 Oct;31(4):497-505. [doi: 10.1016/j.jksuci.2018.03.001]

XSL•FO
RenderX

49. Canhoto AI, Arp S. Exploring the factors that support adoption and sustained use of health and fitness wearables. Journal of Marketing Management 2016 Oct 26;33(1-2):32-60. [doi: 10.1080/0267257X.2016.1234505]
50. Kekade S, Hseieh C, Islam MM, Atique S, Mohammed Khalfan A, Li Y, et al. The usefulness and actual use of wearable devices among the elderly population. Comput Methods Programs Biomed 2018 Jan;153:137-159. [doi: 10.1016/j.cmpb.2017.10.008] [Medline: 29157447]
51. Yen H. Smart wearable devices as a psychological intervention for healthy lifestyle and quality of life: a randomized controlled trial. Qual Life Res 2021 Mar 26;30(3):791-802. [doi: 10.1007/s11136-020-02680-6] [Medline: 33104939]
52. Chen H, Chen C. Factors associated with quality of life among older adults with chronic disease in Taiwan. International Journal of Gerontology 2017 Mar;11(1):12-15 [FREE Full text] [doi: 10.1016/j.ijge.2016.07.002]
53. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
54. Peng W, Li L, Kononova A, Cotten S, Kamp K, Bowen M. Habit formation in wearable activity tracker use among older adults: qualitative study. JMIR Mhealth Uhealth 2021 Jan 19;9(1):e22488 [FREE Full text] [doi: 10.2196/22488] [Medline: 33464216]
55. Cordeiro F, Epstein D, Thomaz E, Bales E, Jagannathan A, Abowd G. Barriers and negative nudges: exploring challenges in food journaling. New York, NY: Association for Computing Machinery; 2015 Apr 18 Presented at: CHI '15: CHI Conference on Human Factors in Computing Systems; April 18-23, 2015; Seoul, Republic of Korea p. 1159-1162. [doi: 10.1145/2702123.2702155]
56. Karapanos E, Gouveia R, Hassenzahl M, Forlizzi J. Wellbeing in the making: peoples' experiences with wearable activity trackers. Psychol Well Being 2016;6:4 [FREE Full text] [doi: 10.1186/s13612-016-0042-6] [Medline: 27376017]
57. Kononova A, Li L, Kamp K, Bowen M, Rikard RV, Cotten S, et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e9832 [FREE Full text] [doi: 10.2196/mhealth.9832] [Medline: 30950807]

Abbreviations

CGSE: the Chinese version of the General Self-Efficacy Scale
CHW: community health worker
LWMD: Live With Wearable Monitoring Device
SDT: Self-Determination Theory
WMD: Wearable Monitoring Device

Edited by A Mavragani; submitted 03.09.23; peer-reviewed by A Keogh, LJ Wilde; comments to author 29.02.24; revised version received 23.04.24; accepted 27.09.24; published 18.11.24.

Please cite as:
Wong AKC, Bayuo J, Su JJ, Wong FKY, Chow KKS, Wong BP, Wong SM, Hui V
Effectiveness of the Support From Community Health Workers and Health Care Professionals on the Sustained Use of Wearable Monitoring Devices Among Community-Dwelling Older Adults: Feasibility Randomized Controlled Trial
J Med Internet Res 2024;26:e52435
URL: https://www.jmir.org/2024/1/e52435
doi:10.2196/52435
PMID:

©Arkers Kwan Ching Wong, Jonathan Bayuo, Jing Jing Su, Frances Kam Yuet Wong, Karen Kit Sum Chow, Bonnie Po Wong, Siu Man Wong, Vivian Hui. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Discussions About Virtual Reality on Twitter to Inform Brain Injury Rehabilitation: Content and Network Analysis

Sophie Brassel[1], BAppSc, PhD; Melissa Brunner[1], BAppSc, MHlthSc, PhD; Andrew Campbell[2], BA, MAppSc, PhD; Emma Power[3], BAppSc, PhD; Leanne Togher[1,4], BAppSc, PhD

[1]Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

[2]School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

[3]Graduate School of Health, University of Technology Sydney, Sydney, Australia

[4]Western Sydney Local Health District, Sydney, Australia

Corresponding Author:
Sophie Brassel, BAppSc, PhD
Sydney School of Health Sciences
Faculty of Medicine and Health
The University of Sydney
Western Avenue
Sydney, 2006
Australia
Phone: 61 286279541
Email: sophie.brassel@sydney.edu.au

Abstract

Background: Virtual reality (VR) use in brain injury rehabilitation is emerging. Recommendations for VR development in this field encourage end user engagement to determine the benefits and challenges of VR use; however, existing literature on this topic is limited. Data from social networking sites such as Twitter may further inform development and clinical practice related to the use of VR in brain injury rehabilitation.

Objective: This study collected and analyzed VR-related tweets to (1) explore the VR tweeting community to determine topics of conversation and network connections, (2) understand user opinions and experiences of VR, and (3) identify tweets related to VR use in health care and brain injury rehabilitation.

Methods: Publicly available tweets containing the hashtags #virtualreality and #VR were collected up to twice weekly during a 6-week period from July 2020 to August 2020 using NCapture (QSR International). The included tweets were analyzed using mixed methods. All tweets were coded using inductive content analysis. Relevant tweets (ie, coded as “VR in health care” or “talking about VR”) were further analyzed using Dann’s content coding. The biographies of users who sent relevant tweets were examined descriptively. Tweet data networks were visualized using Gephi computational analysis.

Results: A total of 260,715 tweets were collected, and 70,051 (26.87%) were analyzed following eligibility screening. The sample comprised 33.68% (23,596/70,051) original tweets and 66.32% (46,455/70,051) retweets. Content analysis generated 10 main categories of original tweets related to VR (ie, advertising and promotion, VR content, talking about VR, VR news, general technology, VR industry, VR live streams, VR in health care, VR events, and VR community). Approximately 4.48% (1056/23,596) of original tweets were related to VR use in health care, whereas 0.19% (45/23,596) referred to VR in brain injury rehabilitation. In total, 14.86% (3506/23,596) of original tweets featured commentary on user opinions and experiences of VR applications, equipment, and software. The VR tweeting community comprised a large network of 26,001 unique Twitter users. Users that posted tweets related to “VR in health care” (2124/26,001, 8.17%) did not form an interconnected VR network, whereas many users “talking about VR” (3752/26,001, 14.43%) were connected within a central network.

Conclusions: This study provides valuable data on community-based experiences and opinions related to VR. Tweets showcased various VR applications, including in health care, and identified important user-based considerations that can be used to inform VR use in brain injury rehabilitation (eg, technical design, accessibility, and VR sickness). Limited discussions and small user networks related to VR in brain injury rehabilitation reflect the paucity of literature on this topic and the potential underuse of this technology. These findings emphasize that further research is required to understand the specific needs and perspectives of people with brain injuries and clinicians regarding VR use in rehabilitation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e45168) doi:10.2196/45168

KEYWORDS

virtual reality; Twitter; brain injury; rehabilitation; cognitive communication; social networks; social media; brain injury rehabilitation; engagement; development; clinical practice; injury; users

Introduction

Background

Acquired brain injuries are a global public health concern [1] and refer to brain injuries sustained after birth, including stroke and traumatic brain injury (TBI) [2,3]. Brain injuries result in substantial socioeconomic burdens [4] and considerable impacts on quality of life [5]. In addition to physical, cognitive, and communication impairments, people with brain injuries can experience challenges in vocational [6,7], social, and community participation [5].

Resources for brain injury rehabilitation may be limited and increasingly challenged owing to the rising incidence and associated long-term care requirements [1,8]. The use of digital health and related technologies has been expanding in health care and may be of benefit [9]. The importance of the role of technology in health care was highlighted by the COVID-19 pandemic with the shift to delivering services via telehealth [10]. Virtual reality (VR) is a digital technology that has become more affordable and accessible, with several studies investigating its use in various health care disciplines and conditions such as posttraumatic stress disorder [11], health professional training [12], and as a distraction tool for patients undergoing chemotherapy [13]. VR is also being explored as a tool for brain injury rehabilitation [14,15] as it may produce benefits over traditional rehabilitation methods [16], such as the ability to control task difficulty [17] and the chance to practice skills in lifelike environments [18], as well as to provide services to patients who experience barriers to accessing care such as cost and geographical isolation [19].

Social Media in Health Care

In addition to technology, there has been an increasing presence of social networking sites in the health care industry. Uses include disseminating health information [20], enhancing professional networking [21], and collecting data [22]. Twitter (which is now known as X) is one of these social networking sites shown to be a valuable source of public discourse on health care [22]. Established in 2006, Twitter has an estimated 486 million users [23] who send >500 million tweets (ie, messages of up to 280 characters) each day [24]. People “tweet” for various reasons, such as expressing thoughts and opinions, broadcasting news, sharing information, interacting with brands and celebrities, or raising awareness of specific issues [25].

Twitter has been used in health research to better understand behaviors and attitudes toward public health issues to inform health promotion and product development or supplement survey data [22,26-28]. The platform has been used as a data source to determine opinions on organ donation [29]; analyze intentions behind drinking alcohol [30]; communicate public health information [31]; study the tweet content of health professionals [32]; explore the tweets of people with communication disabilities [33,34]; and, more recently, understand the sentiments and experiences related to the COVID-19 pandemic [35].

There are reported benefits of using social media for health-related data collection. Twitter provides access to a large, publicly available data set with the potential representation of geographical and demographic groups that may not usually participate in research [26,36]. Other social media platforms do not offer this same access to data as reciprocal relationships are often needed to view posts and users do not often use hashtags to direct conversations as they do on Twitter. Tweets can also contain unfiltered feedback and user-centric content (eg, personal interests, attitudes, and beliefs), which may not be collected through traditional research methods such as focus groups or surveys because of researchers being present [26,28,37]. In addition, companies are turning to social media as a communication strategy, with Twitter becoming prominent in several contexts (eg, corporate communications and IT marketing) [38]. Information about VR innovation and research is likely to be discussed and released on Twitter given that the platform is popular with IT companies [38], government organizations [20], and think tanks [39]. For these reasons, using Twitter as a data source may offer insights into the uses of VR in health care and brain injury rehabilitation that have not been explored because of VR’s global popularity. Digital discourse and the use of social media increased during the COVID-19 pandemic, with Twitter becoming a source of discussions and information exchange by consumers and thought leaders as scientific conferences [40] and government announcements became digital [20]. Therefore, seeking information about VR user opinions and experiences on Twitter could inform VR development and implementation for brain injury rehabilitation.

VR Social Media Research

Research on VR using social media platforms is limited. Twitter posts have been explored with regard to VR development [41] and barriers to VR adoption [42]. These studies did not focus on VR in health care but provided useful information about VR trends and potential barriers to use from the business and marketing fields. Shen et al [41] used a concept decomposition approach to analyze Twitter posts about VR in 2015 to 2016. The concept decomposition involved identifying relevant keywords from VR-related tweets and determining the strength of the connections between related terms [41]. The findings demonstrated that tweet content shifted over time, from a focus on VR development to consumer use and applications. Laurell et al [42] used machine learning methods to develop an algorithm that interpreted tweet content related to VR technology, networks, price, and value. The analysis found that Swedish Twitter users viewed VR as becoming more affordable and accessible in 2016 to 2017 yet perceived that VR offered little value for price. The authors concluded that improvements in VR technology may reduce barriers to uptake [42].

XSL•FO
RenderX

Public perceptions of VR in health care have been investigated via the analysis of Facebook user sentiments toward a video depicting VR use in a hospital setting [37]. A total of 1614 public Facebook comments were analyzed from 1 day in March 2016. Most comments were positive in sentiment, and Facebook users offered suggestions for using VR in health care (eg, distraction from pain and decreasing stress). However, some Facebook users expressed potential concerns such as cost, accessibility, VR sickness, and the need for further research [37]. Although this study offered insights into opinions and potential issues with using VR in health care, it was limited in scope.

User-Based Opinions on and Experiences of VR

Seeking user opinions is essential for successful VR design [43] and achieving effective implementation in clinical practice [37]. End users of VR for brain injury rehabilitation have reported benefits or positive aspects (eg, it being interesting and simple to use) as well as challenges (eg, hardware issues) [44,45]. However, there is an overall paucity of literature related to user opinions and experiences [15]. To address this gap, an alternative, novel approach to collecting data is to interrogate a publicly available data source such as Twitter. Collecting and analyzing Twitter content presents an opportunity to explore user-based discussions on this topic at a large scale and in real time [26]. Unsolicited and unfiltered Twitter conversations could yield potentially untapped demographics, user-centric content, and in-the-moment user experiences or complaints related to VR. Exploring Twitter user networks may help identify key VR end user groups (eg, @users with brain injuries, health professionals, and VR stakeholders), their connections and interactions, and Twitter community characteristics. Exploration of Twitter conversations and users may further inform opinions and experiences related to VR in brain injury rehabilitation and determine whether user experiences in Twitter conversations confirm or differ from those reported in research-based experiences.

Objectives

Twitter offers a data source that can supplement traditional forms of academic inquiry, such as systematic reviews, particularly in a field that is in its infancy. A hashtag study of VR-related tweets to explore uses and perceptions, particularly in health care and brain injury rehabilitation, has not yet been conducted. Analysis of VR-related tweets may inform design considerations, identify current or potential future uses of VR in health care and brain injury rehabilitation, and assist those interested in the topic in navigating relevant information and networks on Twitter. Therefore, this study aimed to collect and analyze tweets containing the hashtags #VR and #VirtualReality to (1) explore the VR tweeting community to determine topics of conversation and network connections, (2) understand user opinions on and experiences of VR, and (3) identify tweets related to VR use in health care and brain injury rehabilitation.

Methods

Ethical Considerations

Ethics approval for this study was granted by the University of Sydney Human Research Ethics Committee (2020/425). Tweet content was paraphrased during reporting of the results to ensure that the content could not be traced to specific Twitter @users owing to the capacity of Twitter and internet search engines to locate matching text [46,47].

Data Collection

VR-related tweets were identified using a systematic Twitter search process [33,34,48-53]. Tweets were identified using the Twitter search bar and included those posted by @users who set their accounts to public, meaning that their tweets are available to anyone who accesses Twitter [54]. The Twitter search period and tweet capture ran from July 10, 2020, to August 18, 2020 (before the platform was rebranded as X). New tweets containing the hashtag #VirtualReality were sourced once per week, and tweets containing the hashtag #VR were sourced twice per week during this time frame to capture the most relevant tweets. The authors did not post any tweets with the hashtags #VR or #VirtualReality during the data collection period. Including additional hashtags such as #BrainInjury, #TBI, #ABI, or #health would have limited the scope for gathering tweets related to user opinions and experiences that would be valuable for consideration in VR development and implementation. A broad approach was also undertaken so as to understand VR-related activity across the platform and provide context for its use.

There are no current recommendations for minimum data sets for a Twitter content analysis [55], although hashtag studies in health care with <3000 tweets have produced valuable results [48]. The time frame and frequency of data collection for a meaningful sample of tweets (ie, related to user experiences and opinions and VR in health care and brain injury) were set a priori based on exploratory Twitter searches by the authors to ensure a sample of at least 3000 meaningful tweets.

Tweets were captured using the NCapture program (QSR International) [56]. Data were imported to the NVivo software (version 12; QSR International) [57] and then exported to an Excel (Microsoft Corp) spreadsheet [58] for analysis. Tweets in the data set were reviewed and excluded based on the following criteria: (1) outside the specified date range, (2) duplicate tweets, (3) not written in English, (4) not related directly to VR (eg, tweets specifically referring to artificial intelligence, augmented reality, or mixed reality; #VR referring to other acronyms such as *voluntary redundancy* or *vocational rehabilitation*; or not referring to immersive VR [59]), (5) tweets identified as having no content (eg, a series of hashtags without meaning), (6) tweets posted by bot @users (ie, @users identified as bots via their username or biographical details), and (7) spam tweets (eg, the same tweet posted by the same @user multiple times in 1 week or identical tweets posted by different @users).

Data Analysis

This study used established mixed methods for analyzing tweet and Twitter network data [48-50,60].

XSL•FO
RenderX

Quantitative Analysis

Quantitative analysis examined the (1) number of tweets, (2) tweet type (ie, original or retweet), and (3) number of unique individuals tweeting about VR. The results are presented descriptively.

Content Analysis

All original tweets were analyzed using inductive content analysis [48,51]. The first author assigned codes to each tweet to broadly describe the content that was shared and discussed among the VR tweeting community. If tweets contained insufficient information, the @user's biographical information and any annotated media or pass-along content were reviewed to provide context. The first and second authors met several times to reach a consensus on coding categories and definitions to enhance trustworthiness [61]. All the authors were involved in discussions related to the final coding categories. Paraphrased tweets were included in the presentation of the results to provide context and clarity for the identified categories.

Dann's Content Classification

Dann's content classification [50] is a framework for categorizing tweets in a consistent manner and provides insights into how individuals and groups use Twitter for social media communication according to five broad categories: (1) conversational (ie, tweets directed toward other Twitter @users), (2) news (eg, news events, current affairs, or events in progress), (3) pass-along (ie, sharing content or information via internet links or retweets), (4) social presence (ie, greeting or connecting with other @users), or (5) status broadcast (ie, expression of thoughts, experiences, and feelings) [50]. Owing to the large number of tweets collected in this study, only those coded as "VR in health care" and "talking about VR" were analyzed by the first author using Dann's content classification [50]. The second author independently coded 19.99% (912/4562) of randomly selected tweets from this sample for coding reliability. The point-by-point reliability was 97.6%, and the discrepancies were resolved through consensus discussion. Tweets coded as "VR in health care" and "talking about VR" were also reviewed by the first author for additional content (eg, images, videos, and external URLs) to further describe tweet content and the sharing of information about VR.

@User Analysis

The biographical statements of @users tweeting about "VR in health care" and "talking about VR" were examined to characterize @users and provide information about VR Twitter networks. The results are presented descriptively.

Computational Analysis

Computational analysis was used to supplement @user and network data. The Gephi software [62] provided a visual representation of the Twitter network data to demonstrate network dimensions and connections between @users who tweeted about VR [51]. Through the use of data visualization, Gephi analysis makes the relationships and communication between Twitter @users visible. Tweet data can be used to generate a graphic representing the network, in which the Twitter @user is at the center (referred to as a "node") and the communication paths of tweets to and from that @user are represented by the curved lines between 2 nodes (referred to as an "edge") [49]. Therefore, Gephi visualizations show communication paths between @users (nodes), and they also display tweet communication paths that travel to "the world" (ie, tweets undirected to another @user). A thin line indicates limited interaction between @users, whereas a thicker line represents more frequent communication. A Fruchterman-Reingold layout from Gephi was used given the large network data sets involved [63], with results presented visually and descriptively.

Methodological Rigor

Reporting Framework

This study was guided by the Good Reporting of a Mixed Methods Study [64] (Multimedia Appendix 1 [64]) guidelines.

Positionality Statement

At the time of this study, the first author was a Doctor of Philosophy candidate and qualified speech-language pathologist with clinical and research experience in brain injury rehabilitation. The remaining authors were academic researchers with a background in speech-language pathology or psychology. All authors formed part of a research team with expertise in communication disorders following brain injury and applying VR and other technologies to support rehabilitation. The authors MB, EP, AC, and LT also had previous experience in social media research.

Results

Overview

The Twitter search process captured 260,715 tweets containing the hashtags #VR (n=66,974, 25.69%) or #VirtualReality (n=193,741, 74.31%) that were posted between July 10, 2020, and August 18, 2020. Following data collection, all tweets were reviewed for inclusion (shown in Figure 1). A total of 70,051 tweets met the inclusion criteria and were analyzed, which comprised 23,596 (33.68%) original tweets (including quote tweets, in which @users add to another @user's original tweet) and 46,455 (66.32%) retweets. There were 26,001 unique Twitter @users who posted the 70,051 included tweets.

XSL•FO
RenderX

Figure 1. Flowchart of the Twitter search process.

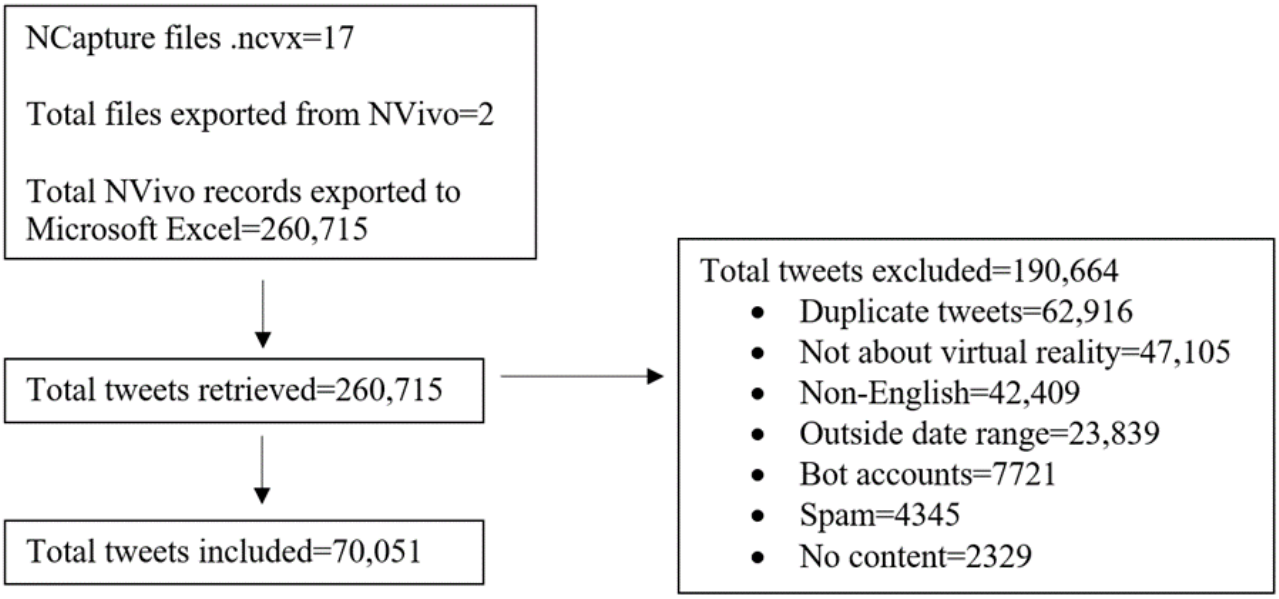

Content Analysis

Overview

Content analysis of the original tweets (23,596/70,051, 33.68%) revealed that @users tweeted about a broad range of uses, content, news, ideas, and opinions related to VR (Table 1). Most of the included tweets advertised VR products or companies (4176/23,596, 17.7%), followed by tweets with VR content such as screen captures or video playthroughs (3571/23,596, 15.13%). Other tweets were coded as talking about VR (3506/23,596, 14.86%), VR news (3254/23,596, 13.79%), general technology (2628/23,596, 11.14%), VR in industry (2569/23,596, 10.89%), VR live streams (1849/23,596, 7.84%), VR in health care (1056/23,596, 4.48%), VR events (700/23,596, 2.97%), and VR community (287/23,596, 1.22%).

Table 1. Content categories of original tweets (N=23,596).

Category	Tweets, n (%)
Advertising and promotion	4176 (17.7)
VR[a] content	3571 (15.13)
Talking about VR	3506 (14.86)
VR news	3254 (13.79)
General technology	2628 (11.14)
VR in industry	2569 (10.89)
VR live streams	1849 (7.84)
VR in health care	1056 (4.48)
VR events	700 (2.97)
VR community	287 (1.22)

[a]VR: virtual reality.

Tweets coded as “VR in health care” and “talking about VR” were examined in further detail to provide insights into VR that may be useful for VR development and the context in which VR was being applied in health care and brain injury rehabilitation. Additional details related to the original tweet content categories are presented in Multimedia Appendix 2.

Content Analysis of “VR in Health Care” Tweets

Overview

Original tweets coded as “VR in health care” (1056/23,596, 4.48%) were further classified using content analysis [48,51] to explore applications of VR in the industry. Most tweets were related to uses or potential uses of VR in various disciplines and for various health conditions.

The use of VR simulations in health education and training was mentioned in 272 tweets, with the most frequent uses or potential applications being in medical and surgical training and staff upskilling (n=219, 80.5%), followed by nursing (n=40, 14.7%), dentistry (n=11, 4%), and allied health (n=2, 0.7%). A total of 23.48% (248/1056) of tweets included information on VR use in the areas of psychology, mental health, and cognition. Many of these tweets were related to the use or potential use of VR

XSL•FO
RenderX

to treat conditions such as anxiety, posttraumatic stress disorder, and phobias. There were 17.52% (185/1056) of "VR in health care" tweets related to the general use of VR in the health care industry. Some tweets provided information on VR for pain management, distraction from medical procedures, and end-of-life care (108/1056, 10.2%). Other tweets were coded as brain injury rehabilitation (45/1056, 4.3%), neurodegenerative conditions (30/1056, 2.8%), general rehabilitation (29/1056, 2.7%), 3D modeling and medical image viewing (26/1056, 2.5%), visual impairments (24/1056, 2.3%), senior care (23/1056, 2.2%), and other (66/1056, 6.3%). Further information about tweets related to VR in health care is presented in Multimedia Appendix 3.

@User Experiences of VR in Health Care

A total of 5.3% (56/1056) of "VR in health care" tweets provided details about @user experiences, commentary, and opinions related to VR in health care. Of these 56 tweets, there were 12 (21%) related to direct @user experiences, including educating medical, nursing, or allied health students (n=7, 58%); distraction from medical procedures (n=2, 17%); pain management (n=2, 17%); and learning about infection control (n=1, 8%). Some @users tweeted to share positive opinions on VR use in health care (19/56, 34%), discuss the use or potential uses of VR in health care (12/56, 21%), network with others using VR in related fields (8/56, 14%), share plans to use VR in health research (1/56, 2%), and express that VR design in health needs user engagement (1/56, 2%). In total, 3 @users expressed challenges with VR: programming difficulties with developing a VR nursing platform, nausea when using a VR exposure therapy platform, and hand tremors affecting VR use by causing shaking and reduced accuracy of gameplay. None of these @users identified as having a brain injury.

Sharing Information and Resources

Considering the large proportion of conversational tweets, there were not many discussions between @users about VR in health care. Many conversational tweets contained references to @users or referred content to other @users by tagging them in a tweet. The 1056 original tweets about "VR in health care" were further examined to determine what content was being shared. Most tweets (450/1056, 42.61%) provided a link to a web article or blog that provided information about the use or potential use of VR in health care. Others shared links related to web-based events such as webinars or conferences (117/1056, 11.08%), research articles (111/1056, 10.51%), images or photos (93/1056, 8.81%), videos that demonstrated the use of VR or VR screen recordings (89/1056, 8.43%), websites (36/1056, 3.41%), podcasts (11/1056, 1.04%), research participation links (10/1056, 0.95%), VR app downloads (7/1056, 0.66%), web-based reports (3/1056, 0.28%), a product brochure (1/1056, 0.09%), and a web-based message board (1/1056, 0.09%). There were 6.34% (67/1056) of quote tweets and 5.68% (60/1056) of tweets that contained a broken URL or no additional content.

VR in Brain Injury Rehabilitation

A total of 4.26% (45/1056) of "VR in health care" tweets provided insights into VR applications for brain injury rehabilitation. Of these 45 tweets, there were 17 (38%) containing information about VR use specific to stroke rehabilitation (eg, gait, vision, and upper limb retraining). Some tweets (11/45, 24%) provided news updates about companies that make VR products for neurorehabilitation. A VR educational platform for concussion was released during the tweet capture period and was the topic of 20% (9/45) of the tweets. In total, 4% (2/45) of the tweets referenced VR rehabilitation programs for brain injuries caused by various etiologies, including for TBIs. Other tweets included details about VR research for acquired communication disorders (2/45, 4%), web-based events (2/45, 4%), testing a VR neurorehabilitation platform (1/45, 2%), and a VR platform for eye tracking in athletes with brain injuries (1/45, 2%). Most tweets in this subcategory (42/45, 93%) shared links to articles or blog posts (27/42, 64%), journal articles (5/42, 12%), videos (4/42, 10%), quote tweets (4/42, 10%), web-based event registrations (1/42, 2%), and research participation information (1/42, 2%).

Content Analysis of "Talking About VR" Tweets

Overview

Original tweets coded as "talking about VR" (3506/23,596, 14.86%) included commentary, discussions, opinions, and questions related to @users' experiences with VR. These tweets were further coded into three subcategories: (1) general conversations about VR (2904/3506, 82.83%), (2) seeking advice or opinions (439/3506, 12.52%), and (3) @user preferences and feedback about VR experiences (163/3506, 4.65%). Additional information about these subcategories is provided in Multimedia Appendix 4.

Almost half of these tweets (1704/3506, 48.6%) shared content related to VR: images or photos (930/1704, 54.58%), videos (427/1704, 25.06%), web-based articles or blogs (177/1704, 10.39%), web-based message boards (118/1704, 6.92%), websites (21/1704, 1.23%), VR live streams (12/1704, 0.7%), VR app store links (10/1704, 0.59%), research articles (4/1704, 0.23%), web-based events (3/1704, 0.18%), and podcasts (2/1704, 0.12%). Content included links to other social media platforms such as YouTube, Reddit, and Instagram. There were 38.36% (1345/3506) of these tweets with no links or media, 11.32% (397/3506) of quote tweets, and 1.71% (60/3506) of broken URLs that were inaccessible at the time of data analysis.

General Conversations About VR

There were 82.83% (2904/3506) of "talking about VR" tweets in this subcategory. Many of these tweets (1229/2904, 42.32%) contained details, commentary, and @user views on VR experiences and purchases. @Users also commented on VR news (eg, game releases or updates and VR equipment) and content (554/2904, 19.08%). Tweets also included general commentary and discussions on VR (520/2904, 17.91%). Some @users discussed ideas for VR games or experiences and the potential of VR and its future (295/2904, 10.16%). These discussions focused on updates to existing VR applications, discussions about cross-platform content such as sports and concert viewing, and VR uptake in the context of the COVID-19 pandemic (eg, applications in industries such as education or remote work). Other @users provided insights into VR development processes and shared related images and videos

XSL•FO
RenderX

(220/2904, 7.58%). There were also conversations about VR industry challenges (86/2904, 2.96%), such as obstacles to realizing the potential of VR and barriers to entering the industry (eg, cost, usefulness, and access).

Seeking Advice and Opinions on VR

These tweets (439/3506, 12.52%) contained questions through which @users sought advice or opinions related to aspects of VR, including suggestions or opinions on VR games and head-mounted displays (HMDs; 147/439, 33.5%). Some @users asked about VR features and development, such as hand tracking, recording, haptics, multiplayer experiences, graphics cards, VR setup, and VR design software (103/439, 23.5%). @Users also requested advice to address VR issues (63/439, 14.4%). Others sought opinions or suggestions related to the VR industry (50/439, 11.4%) and potential uses of VR (24/439, 5.5%).

@User Preferences and Feedback on VR Experiences

These tweets (163/3506, 4.65%) provided insights into @users' VR experiences and preferences that could be useful for informing VR development and associated considerations (eg, design aspects, hardware and software issues, adverse effects, and accessibility). Of these 163 tweets, 55 (33.7%) were related to the potential adverse effects of VR use, such as VR sickness and discomfort when wearing HMDs (eg, nausea, headache, dizziness, and sore eyes). Some @users provided information about ways to minimize adverse effects (eg, starting with seated VR experiences or using HMD comfort modifications), whereas others described a lack of adverse effects. Other tweets (50/163, 30.7%) contained information or solutions related to issues that @users experienced with VR (eg, hardware issues, audio or graphic glitches, connectivity issues, and gameplay error messages). The tweets also contained @user preferences and opinions regarding aspects of VR design (32/163, 19.6%), including discussions about hardware, field of view, latency, scaling, and wireless capabilities. VR accessibility was discussed in 16% (26/163) of tweets, in which @users tweeted about the need to make VR accessible to people who have visual impairments or physical limitations or who are deaf or hard of hearing. Some @users described VR accessibility options such as closed captions, single controller use, and prescription lenses.

Dann's Content Classification

Overview

Tweets within the categories "VR in health care" and "talking about VR" were coded based on Dann's content classification [50] (Table 2).

Table 2. Dann's content classification [50] of original tweets within the "VR[a] in health care" and "talking about VR" categories (N=4562).

	"VR in health care" tweets (n=1056), n (%)	"Talking about VR" tweets (n=3506), n (%)
Conversational	434 (41.1)	1361 (38.82)
Pass-along	561 (53.13)	805 (22.96)
News	29 (2.75)	9 (0.26)
Status broadcast	25 (2.37)	942 (26.87)
Social presence	7 (0.66)	389 (11.1)

[a]VR: virtual reality.

"VR in Health Care" Tweets

Most tweets (561/1056, 53.13%) in this category were coded as pass-along, followed by conversational (434/1056, 41.1%), news (29/1056, 2.75%), status broadcast (25/1056, 2.37%), and social presence (7/1056, 0.66%) tweets. In the subcategory "brain injury rehabilitation" (45/1056, 4.26% of tweets), of the 45 tweets, there were 27 (60%) pass-along tweets and 18 (40%) conversational tweets.

"Talking About VR" Tweets

Original tweets in this category were classified as conversational (1361/3506, 38.82%), followed by status broadcast (942/3506, 26.87%), pass-along (805/3506, 22.96%), social presence (389/3506, 11.1%), and news (9/3506, 0.26%).

Twitter @User Analysis

Overview

The biographical details of @users who tweeted within the categories of "VR in health care" and "talking about VR" were examined to inform @user networks (Table 3).

XSL•FO
RenderX

Table 3. @User groups for original tweets within the "VR[a] in health care" and "talking about VR" categories (N=2679).

User group	@Users posting "VR in health care" original tweets (n=622), n (%)	@Users posting "Talking about VR" original tweets (n=2057), n (%)
VR, technology, or gaming companies	123 (19.8)	191 (9.3)
Health care or health technology professionals and researchers	118 (19)	19 (0.9)
@Users with an interest or expertise in VR, gaming, or technology	105 (16.9)	1195 (58.1)
Health technology companies	73 (11.7)	4 (0.2)
Health organizations	64 (10.3)	0 (0)
Health, technology, or gaming news outlets	33 (5.3)	18 (0.9)
No or insufficient biographical data	32 (5.1)	260 (12.6)
Organizations and companies unrelated to health care or VR	25 (4)	37 (1.8)
General public	21 (3.4)	329 (16)
Twitter accounts for conferences or expositions	12 (1.9)	1 (0)
News outlets unrelated to VR or health	9 (1.4)	3 (0.1)
Academic journals	7 (1.1)	0 (0)

[a]VR: virtual reality.

@Users Tweeting About "VR in Health Care"

There were 2124 unique @users who posted an original tweet, retweet, or both about VR in health care: 622 unique @users posted the 1056 original tweets, and 1599 @users posted 2733 retweets. The biographical statements and usernames of the @users who posted original tweets were examined and coded. The largest group of @users who posted original tweets about "VR in health care" were VR, technology, or gaming companies (123/622, 19.8%). The next largest group included health care or health technology professionals and researchers in the fields of medicine, nursing, and allied health (118/622, 19%). The remaining categories of @users included people with an interest or expertise in VR, gaming, or technology (105/622, 16.9%); health technology companies (73/622, 11.7%); health organizations (64/622, 10.3%); health, technology, or gaming news outlets (33/622, 5.3%); no or insufficient biographical data (32/622, 5.1%); organizations and companies unrelated to health care or VR (25/622, 4%); the general public (21/622, 3.4%); Twitter accounts for conferences or expositions (12/622, 1.9%); news outlets unrelated to VR or health (9/622, 1.4%); and academic journals (7/622, 1.1%). Of the 622 @users tweeting about VR in health care, 43 (6.9%) posted original tweets about VR use in brain injury rehabilitation. Health care professionals accounted for 23% (10/43) of these @users.

@Users Tweeting About "Talking About VR"

A total of 3752 unique @users posted an original tweet, retweet, or both within the category "talking about VR": 2057 @users posted the 3506 original tweets, and 1857 @users posted 2782 retweets. Most @users in this category who posted original tweets (1195/2057, 58.09%) identified as having an interest or expertise in VR, gaming, or technology. Other @users were categorized as the general public (329/2057, 15.99%); no or insufficient biographical data (260/2057, 12.64%); VR, technology, or gaming companies (191/2057, 9.29%); organizations and companies unrelated to VR or technology (37/2057, 1.8%); health care or health technology professionals (19/2057, 0.92%); VR or technology news outlets (18/2057, 0.88%); health technology companies (4/2057, 0.19%); news outlets unrelated to VR or technology (3/2057, 0.15%); and a Twitter account for a conference (1/2057, 0.05%).

Computational Analysis

A Gephi visualization of the 26,001 @users who tweeted or retweeted posts containing the hashtags #VR and #VirtualReality is shown in Figure 2. The visualization shows that (1) a large, dense network of @users discussed VR, with some tweeting more than others (ie, evidenced by the thicker, darker lines between nodes); (2) most of the tweets were undirected (ie, not sent to specific @users but sent out to the Twitter void); and (3) many @users were not connected to one another within a central network.

XSL•FO
RenderX

Figure 2. Gephi network visualization of the virtual reality tweeting community.

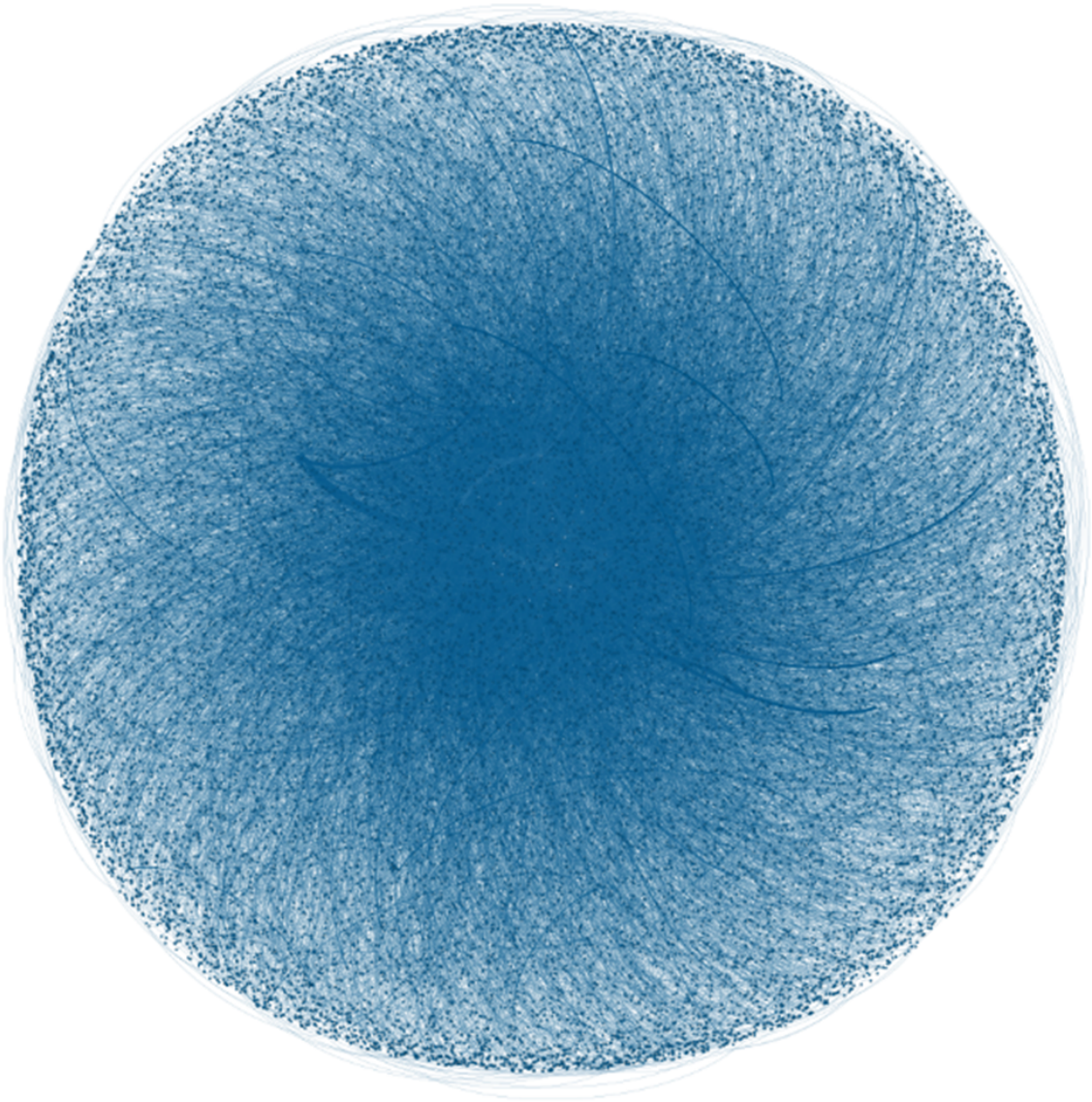

Figure 3 shows the network of 2124 @users who posted tweets coded as "VR in health care." This visualization demonstrates that (1) overall, there were many individual @users sending tweets who were not necessarily connected with others in the network; (2) some individuals were connected within smaller central networks, but they were not highly interconnected; and (3) a small number of @users were connected to others on the periphery and were separate from the more central @users and small networks.

XSL•FO
RenderX

Figure 3. Gephi network visualization of the tweeting community for the "VR in health care" category.

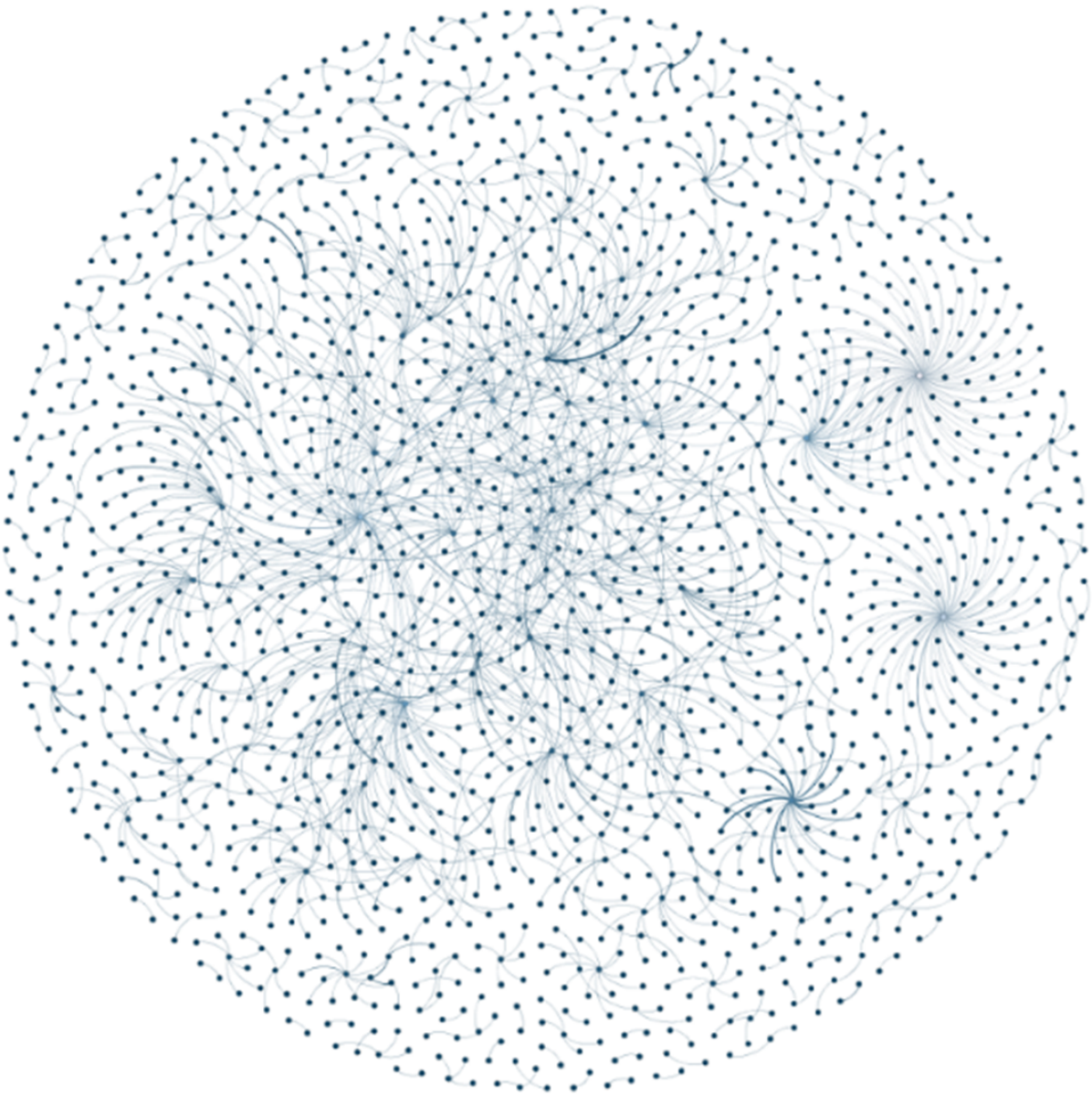

A visualization of the network tweeting within the category "talking about VR" is presented in Figure 4 (n=3752 @users). Within this network, there were (1) many @users connected within a central community, (2) numerous outliers that were not connected to the central community, and (3) no distinct small VR tweeting networks.

XSL•FO
RenderX

Figure 4. Gephi network visualization of the tweeting community for the "talking about VR" category.

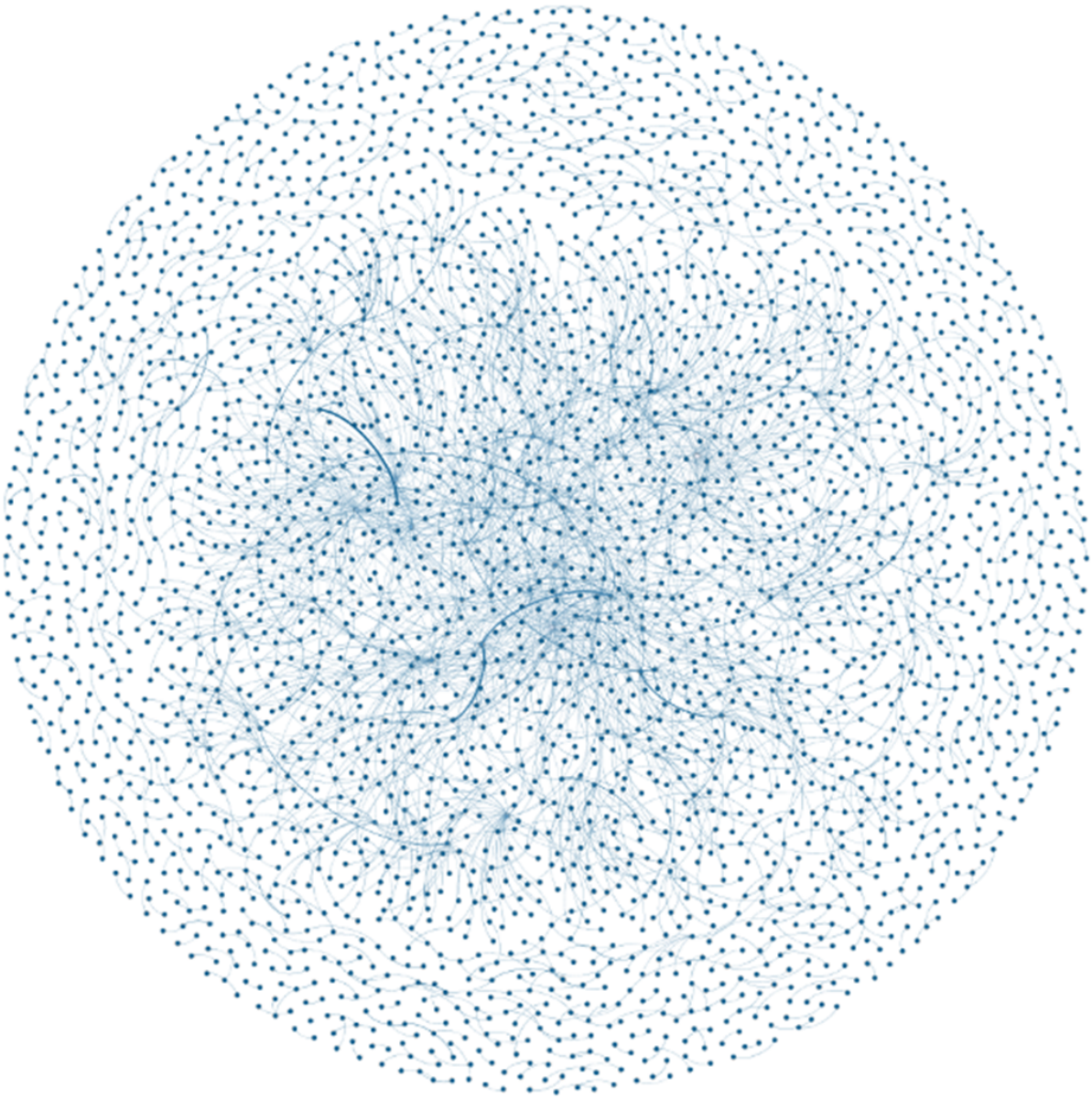

Discussion

Principal Findings

Overview

This study explored tweets about VR that were posted over a 6-week period from July 2020 to August 2020. The analysis of tweets containing the hashtags #VR and #VirtualReality provided an overview of VR-related topics, conversations, and @user networks on Twitter. Applications of VR in health care and brain injury rehabilitation were also explored.

A total of 10 main content themes were identified via a content analysis of 23,596 original tweets. In this tweet sample, @users predominantly tweeted to advertise VR-related content and products such as applications and hardware. This finding was not surprising given the increased use of social media platforms for driving digital engagement with brands [65]. Twitter is also evidently a social networking site that people with an interest in VR use to share VR-related content, experiences, and news. This finding reflects general Twitter use trends as news is frequently distributed via the platform in addition to it being a platform for conversing and sharing ideas [66,67]. Tweet content also reflected VR applications in different industries, including education [68], real estate [69], tourism [70], and architecture [71].

Tweets were typically related to events or news that occurred at the time of data collection. The COVID-19 pandemic may have led to a focus on VR use in industries, discussion of VR ideas and potential owing to stay-at-home directives, and the necessity of technology to keep people entertained and connected. Similar trends have been noted in tweets related to the use of other technologies and telepractices during the pandemic [72,73]. It is important to note that 260,715 tweets

XSL•FO
RenderX

containing the hashtags #VR and #VirtualReality were collected, yet only 70,051 (26.87%) were eligible for analysis. Previous research has suggested that using Twitter as a data source may harvest irrelevant data, especially as tweets are typically unprompted and not created to address research purposes [73]. This was confirmed in this data set, in which >150,000 tweets included irrelevant content, were posted by bot accounts, or included spam content. As such, navigating content on Twitter may be challenging for users to easily find relevant information about VR experiences or best practices for rehabilitation.

The VR Tweeting Community

The Gephi visualizations (Figures 2-4) provided an understanding of the communities engaging in VR discussions. A large network of @users formed the overall VR tweeting community, although many were not interconnected and there were no clear central networks. The visualization (Figure 2) also showed that, although a proportion of @users tweeted several times, most sent one-off tweets within the collected sample. Within the "VR in health care" subset, a small number of @users were connected (Figure 3). However, these connections were not demonstrated across the network and included smaller groups of @users. In comparison, the "talking about VR" network had a larger and more connected central community. This finding indicates that, although there may be smaller networks of health care professionals with an interest in VR, a larger interconnected Twitter community is yet to be established. An interconnected community could offer a platform to exchange ideas, seek support, and establish networks [74].

Examining the biographical data of @users posting original tweets in the "VR in health care" and "talking about VR" content categories provided insights into the VR tweeting community (Table 3). During the data collection period, Twitter handles of health care professionals; VR companies or products; and those with an interest or expertise in VR, technology, or gaming accounted for a similar number of tweets related to VR use in health care (16%-20% each). However, health professionals made up a small proportion of the overall number of individual Twitter @users identified in this study via user analysis (Table 3).

No @user who tweeted using the hashtags #VirtualReality or #VR during the 6-week data collection period identified as having a brain injury. Previous research suggests that people with TBI use Twitter for various reasons, yet few of their tweets refer to rehabilitation experiences [33,48]. People with brain injuries are underrepresented on Twitter [48], which may be reflected in the findings of this study. If individuals with brain injuries are interested in VR, Twitter could provide a platform for accessing relevant information and examples as well as opportunities to engage with others who are also interested in VR (eg, the potential to connect and establish VR-based social interactions).

@Users with an interest or expertise in VR, gaming, or technology accounted for >50% (1195/2057, 58.09%) of @users within the content category "talking about VR," and the network analysis (Figure 4) showed a connected group of @users interacting. This finding suggests that Twitter is an important platform for the VR community to share experiences and opinions, comment on news, disseminate events, and seek advice within a supportive community. This could also explain the fact that many tweets had content or additional hashtags referring to technology or gaming terminology that may not be immediately familiar to the broader public who use Twitter (eg, #IoT, #metaverse, #blockchain, #crypto, and #GameDev; Multimedia Appendices 2 and 4). This content might not be accessible to people with brain injuries and associated communication disabilities as they may have difficulty comprehending these terms.

VR in Health Care and Brain Injury

A small community tweeted about VR use in health care, with tweets related to this topic accounting for 4.48% (1056/23,596) of the analyzed tweets. Over half (561/1056, 53.13%) of the tweets within this subcategory comprised pass-along tweets, in which @users shared information through web-based news articles, research publications, and web-based events. The analyzed tweets demonstrated that VR is being used in various health care disciplines. Education for medical professionals and psychology were the most represented fields, accounting for almost half (520/1056, 49.24%) of the tweets in the "VR in health care" category. This finding reflects major VR research themes in the literature [75]. The use of VR could be lagging in other fields given these findings, or those using or researching VR in these areas may not be using Twitter to promote their work. Alternatively, the data were a snapshot of tweets collected over a 6-week period, so the included tweets only accounted for news, events, experiences, and publications within or close to this time frame that were shared on Twitter. Expanding the time frame of data collection may have provided further insights into tweets related to VR in health care.

Although conversational tweets accounted for 41.1% (434/1056) of tweets in the "VR in health care" category, @users did not have robust conversations about using VR in the field. For example, @users in this study were tagged in tweets to promote VR applications or share related articles or events, with only 5.3% (56/1056) of tweets referencing a direct experience of or opinion on VR related to health care. Similar studies exploring health-related hashtags on Twitter also found minimal discussion about topics among Twitter networks [76]. This differs from non–health care @users in this study, who shared how they used VR and showcased different VR examples and experiences. In addition, only 0.28% (3/1056) of tweets in this category provided insights into the challenges that people using VR in health care may face, and none of these tweets were posted by @users with a brain injury.

The use of VR in brain injury rehabilitation was identified in only 0.19% (45/23,596) of original tweets. Many of these tweets discussed the use of VR for physical or cognitive impairments rather than for acquired communication disorders and for stroke rather than for TBI rehabilitation. This finding is consistent with the published literature [15] in that the use of VR in these fields is yet to be established and evaluated, in addition to there being limited guidance for its use in clinical practice. There was little research dissemination, reflecting hashtag studies for brain injury that have found a lack of research engagement via Twitter

XSL•FO
RenderX

[48]. To date, Twitter research related to brain injury appears to be limited to TBI-related hashtags [48], the use of Twitter by people with TBI [33], a network and content analysis of aphasia-related tweets [76], and tweets about concussion [77-79].

@User Opinions and Experiences Related to VR

This study demonstrated that Twitter is an important platform for people and companies with an interest in VR to share experiences, opinions, and content related to VR. Most tweets coded as "talking about VR" were conversational, with the purpose of engaging in discussions or drawing the attention of other @users. The platform was also used to seek advice from other @users with an interest in VR or share opinions and experiences related to VR that were almost always positive. However, these opinions and experiences should be considered in the context of those who use Twitter as many @users had a specific interest or expertise in VR and the platform does not typically represent the views of the general population [36]. Many of these tweets contained links to content hosted on other social media platforms (eg, YouTube, Twitch, Instagram, or Facebook) or web-based articles and blogs. This content could be used by clinicians and those interested in VR to learn about how it works or view examples of VR applications (eg, viewing VR content on YouTube).

It is important to note that some @users highlighted concerns with VR, either commenting directly on issues affecting their VR experience (eg, VR sickness, hardware and software issues, and accessibility) or the VR industry (eg, cost and access). Previous social media [37,42] and end user research [80-83] related to VR barriers has also found that VR sickness, cost, and accessibility were featured in discussions. However, these tweets formed a small proportion of the overall tweet sample in this study (55/23,596, 0.23% of analyzed tweets). Although these issues were not described by people identifying as having a brain injury, the experiences of >2600 Twitter @users provided valuable insights into potential challenges to be aware of and consider when using VR in health care settings. Considerations not previously described in recommendations for VR development for brain injury rehabilitation [15] were identified in this study, which were mainly related to technical aspects of VR (eg, interaction mechanisms, game engines, latency, and image resolution). These design considerations are similar to those for developing VR-based exposure therapy [84].

Limitations

Although this study analyzed a meaningful sample of tweets, not all tweets containing #VR or #VirtualReality posted during the 6-week sample period were captured and analyzed. This limitation is due to the inherent constraints of the Twitter application programming interface search algorithm, which limits the number of tweets collected at one time. In addition, not all tweets related to brain injury or health and the use of VR were captured and analyzed within this time frame. Tweets sent from private Twitter accounts and those written in languages other than English were also not analyzed. Furthermore, Dann's [50] content coding and @user analysis were limited to tweets relevant to the aims of this study (ie, coded as "VR in health care" and "talking about VR") and content coding focused on original tweets (ie, excluded retweets or content in quote tweets). The analysis of more tweets from the data set using these methods may have provided additional information about the VR tweeting community and how the platform is used to disseminate related information and content.

Comparison With Prior Work

The findings of this study support evidence from the literature that the use of VR in brain injury rehabilitation, particularly for TBI [15], is limited. Important user considerations for VR implementation in health care and brain injury rehabilitation were identified in Twitter @users' experiences (eg, usability, accessibility, and VR sickness) [15,43], highlighting the need to engage end users in VR design and feasibility testing [43,85] as well as take technical aspects into account. This study showed that valuable insights can be drawn from Twitter, which constituted a novel data source of >2600 unique "voices" that discussed their own experiences of and opinions on VR, demonstrated an interest in VR health applications, and reflected related social media research on VR in health care [37].

Clinical Implications and Future Directions

Twitter has been identified as a platform that can be used in health care to determine public perceptions, recruit research participants, and share and advance research [20,22,86,87]. There is potential to disseminate VR news, research, and events in this field via social media platforms. However, given the large number of tweets containing the hashtags #VR and #VirtualReality, it may be challenging to navigate and find relevant information on VR use in health care. Including targeted hashtags (Multimedia Appendices 2-4) in Twitter searches may identify more relevant content if clinicians are interested in obtaining information about VR. Exploring VR-related hashtags on other social networking sites could further inform how social media is used to disseminate information about VR in health care and whether different platforms yield different findings.

Researchers and clinicians who use VR could promote their work via Twitter [48,51] or other social networking sites, which may increase interest in and awareness of VR, promote networking opportunities, and showcase potential uses to patients and clinicians. Sharing examples and engaging in discussions could demonstrate how VR could be implemented in clinical practice or ease uncertainties related to its use (eg, VR sickness or suitability for specific patient groups). @Users with a brain injury could also find it challenging to navigate VR-related topics on Twitter as many can experience cognitive and communication difficulties. For example, people with TBI have reported that they tweet for different purposes, including to share their experiences of living with a TBI, support others, and seek information [33]. Therefore, clinicians could support people with brain injuries and associated communication disorders to navigate topics on Twitter [48,76], particularly given the potential concerns related to the use of social media by people with brain injuries (eg, internet safety and confidentiality) [88].

Some tweets provided insights into challenges that VR users may experience, which could apply to using VR in health care settings. Health professionals may anticipate these potential

XSL•FO
RenderX

issues and identify ways to overcome them. However, it is important to consider that the reported issues were mostly from @users who did not identify as having a health condition or disability, meaning that their experiences cannot be relied on alone to inform VR design in health care. Exploring the functionality of VR with different patient groups is necessary to determine VR usability [83], including any barriers to using VR in clinical settings or for leisure. This is particularly important for people with a brain injury as they can experience physical, cognitive, and communication impairments that may affect their use of VR [15]. It will be necessary to explore the perspectives of people with brain injuries and their clinicians in relation to VR use as it is recommended that end users be involved in developing VR for health care [43] and brain injury rehabilitation [15]. Interviews or focus groups could be conducted with these key end users to establish their views on using VR in rehabilitation, including perceived barriers and facilitators. Once these barriers and facilitators are established, user-based testing of VR with people with brain injuries and their clinicians should be implemented to ensure that VR applications are targeted to their needs.

Conclusions

This study explored VR-related tweets and networks with a focus on user experience and applications in health care and brain injury rehabilitation. Content analysis revealed that the VR tweeting community used Twitter for various purposes, including promoting VR-related products, sharing content, disseminating news and events, and talking about experiences and opinions. The platform provided insights into the use of VR in health care, with @users predominantly disseminating research and information on clinical VR applications rather than experiences of or opinions on VR use in this field. Tweets referring to VR in brain injury rehabilitation comprised a small proportion and did not provide an in-depth insight into the use of VR by this population.

Twitter has the potential to showcase VR use in health care and brain injury rehabilitation to promote uptake or disseminate research findings. However, the vast amount of information and potential for unrelated content suggest that related hashtags should be considered when searching for VR content on Twitter. Those involved in developing or using VR in brain injury rehabilitation should also consider the reported issues and challenges identified in the tweet sample, including cost, accessibility, VR sickness, and technical design aspects. Few tweets provided insights into user experiences of and opinions on VR in health settings, with none provided by people with brain injuries. Further research is needed to determine the VR needs and experiences of people with brain injuries and their clinicians to guide the design of VR applications for effective rehabilitation.

Acknowledgments

At the time of this study, the first author (SB) was supported by an Australian Government Research Training Program scholarship and a Merit Award scholarship (the University of Sydney).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors were involved in the design of this study. SB collected the data, completed the data analysis (except for the Gephi visualizations), and wrote the drafts of the manuscript. MB completed the Gephi visualizations. All authors contributed to manuscript editing and revision. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GRAMMS (Good Reporting of a Mixed Methods Study) checklist.
[PDF File (Adobe PDF File), 127 KB - jmir_v26i1e45168_app1.pdf]

Multimedia Appendix 2
Details on original tweet content categories.
[PDF File (Adobe PDF File), 274 KB - jmir_v26i1e45168_app2.pdf]

Multimedia Appendix 3
Details on original tweets related to “VR in health care.”.
[PDF File (Adobe PDF File), 227 KB - jmir_v26i1e45168_app3.pdf]

Multimedia Appendix 4

XSL•FO
RenderX

Details on original tweets related to “talking about VR.”.
[PDF File (Adobe PDF File), 285 KB - jmir_v26i1e45168_app4.pdf]

References

1. Menon DK, Bryant C. Time for change in acquired brain injury. Lancet Neurol 2019 Jan;18(1):28. [doi: 10.1016/S1474-4422(18)30463-0] [Medline: 30563679]
2. Teasell R, Bayona N, Marshall S, Cullen N, Bayley M, Chundamala J, et al. A systematic review of the rehabilitation of moderate to severe acquired brain injuries. Brain Inj 2007 Feb;21(2):107-112. [doi: 10.1080/02699050701201524] [Medline: 17364527]
3. What is the difference between acquired brain injury and traumatic brain injury? Brain Injury Association of America. URL: https://www.biausa.org/brain-injury/about-brain-injury/nbiic/what-is-the-difference-between-an-acquired-brain-injury-and-a-traumatic-brain-injury [accessed 2020-06-08]
4. Maas AIR, Menon DK, Adelson PD, Andelic N, Bell MJ, Belli A, InTBIR Participants and Investigators. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol 2017 Dec;16(12):987-1048. [doi: 10.1016/S1474-4422(17)30371-X] [Medline: 29122524]
5. Goverover Y, Genova H, Smith A, Chiaravalloti N, Lengenfelder J. Changes in activity participation following traumatic brain injury. Neuropsychol Rehabil 2017 Jun 04;27(4):472-485. [doi: 10.1080/09602011.2016.1168746] [Medline: 27043964]
6. Alaszewski A, Alaszewski H, Potter J, Penhale B. Working after a stroke: survivors' experiences and perceptions of barriers to and facilitators of the return to paid employment. Disabil Rehabil 2007 Dec 30;29(24):1858-1869. [doi: 10.1080/09638280601143356] [Medline: 17852252]
7. Shames J, Treger I, Ring H, Giaquinto S. Return to work following traumatic brain injury: trends and challenges. Disabil Rehabil 2007 Sep 15;29(17):1387-1395. [doi: 10.1080/09638280701315011] [Medline: 17729084]
8. Jackson D, McCrone P, Mosweu I, Siegert R, Turner-Stokes L. Service use and costs for people with long-term neurological conditions in the first year following discharge from in-patient neuro-rehabilitation: a longitudinal cohort study. PLoS One 2014;9(11):e113056 [FREE Full text] [doi: 10.1371/journal.pone.0113056] [Medline: 25401992]
9. Bower KJ, Verdonck M, Hamilton A, Williams G, Tan D, Clark RA. What factors influence clinicians' use of technology in neurorehabilitation? A multisite qualitative study. Phys Ther 2021 May 04;101(5):pzab031. [doi: 10.1093/ptj/pzab031] [Medline: 33522582]
10. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
11. Kothgassner OD, Goreis A, Kafka JX, Van Eickels RL, Plener PL, Felnhofer A. Virtual reality exposure therapy for posttraumatic stress disorder (PTSD): a meta-analysis. Eur J Psychotraumatol 2019;10(1):1654782 [FREE Full text] [doi: 10.1080/20008198.2019.1654782] [Medline: 31489138]
12. Moore N, Yoo S, Poronnik P, Brown M, Ahmadpour N. Exploring user needs in the development of a virtual reality-based advanced life support training platform: exploratory usability study. JMIR Serious Games 2020 Aug 07;8(3):e20797 [FREE Full text] [doi: 10.2196/20797] [Medline: 32763877]
13. Janssen A, Fletcher J, Keep M, Ahmadpour N, Rouf A, Marthick M, et al. Experiences of patients undergoing chemotherapy with virtual reality: mixed methods feasibility study. JMIR Serious Games 2022 Feb 21;10(1):e29579 [FREE Full text] [doi: 10.2196/29579] [Medline: 35188474]
14. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2017 Nov 20;11(11):CD008349 [FREE Full text] [doi: 10.1002/14651858.CD008349.pub4] [Medline: 29156493]
15. Brassel S, Power E, Campbell A, Brunner M, Togher L. Recommendations for the design and implementation of virtual reality for acquired brain injury rehabilitation: systematic review. J Med Internet Res 2021 Jul 30;23(7):e26344 [FREE Full text] [doi: 10.2196/26344] [Medline: 34328434]
16. Massetti T, da Silva TD, Crocetta TB, Guarnieri R, de Freitas BL, Bianchi Lopes P, et al. The clinical utility of virtual reality in neurorehabilitation: a systematic review. J Cent Nerv Syst Dis 2021 May 04;10:1179573518813541. [doi: 10.1177/1179573518813541] [Medline: 30515028]
17. Schultheis MT, Himelstein J, Rizzo AA. Virtual reality and neuropsychology: upgrading the current tools. J Head Trauma Rehabil 2002 Oct;17(5):378-394. [doi: 10.1097/00001199-200210000-00002] [Medline: 12802250]
18. Rizzo A, Kim GJ. A SWOT analysis of the field of virtual reality rehabilitation and therapy. Presence 2005 Apr;14(2):119-146 [FREE Full text] [doi: 10.1162/1054746053967094]
19. Pietrzak E, Pullman S, McGuire A. Using virtual reality and videogames for traumatic brain injury rehabilitation: a structured literature review. Games Health J 2014 Aug;3(4):202-214. [doi: 10.1089/g4h.2014.0013] [Medline: 26192369]
20. Tsao S, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 2021 Mar;3(3):e175-e194 [FREE Full text] [doi: 10.1016/S2589-7500(20)30315-0] [Medline: 33518503]

XSL•FO
RenderX

21. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
22. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
23. The global state of digital trends 2022. Hootsuite. 2022. URL: https://www.hootsuite.com/resources/digital-trends [accessed 2022-11-11]
24. Sayce D. The number of tweets per day in 2020. Paper Gecko Ltd. URL: https://www.dsayce.com/social-media/tweets-day/ [accessed 2021-02-25]
25. Bruns A, Moe H. Structural layers of communication on Twitter. In: Weller AB K, Burgess J, Mahrt M, Puschmann C, editors. Twitter and Society. New York City, NY: Peter Lang; 2014:15-28.
26. Myslín M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013 Aug 29;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
27. Doan S, Ritchart A, Perry N, Chaparro JD, Conway M. How do you #relax when you're #stressed? A content analysis and infodemiology study of stress-related Tweets. JMIR Public Health Surveill 2017 Jun 13;3(2):e35 [FREE Full text] [doi: 10.2196/publichealth.5939] [Medline: 28611016]
28. Pomputius A. Can you hear me now? Social listening as a strategy for understanding user needs. Med Ref Serv Q 2019;38(2):181-186. [doi: 10.1080/02763869.2019.1588042] [Medline: 31173567]
29. Jiang X, Jiang W, Cai J, Su Q, Zhou Z, He L, et al. Characterizing media content and effects of organ donation on a social media platform: content analysis. J Med Internet Res 2019 Mar 12;21(3):e13058 [FREE Full text] [doi: 10.2196/13058] [Medline: 30860489]
30. Riordan BC, Merrill JE, Ward RM. "Can't wait to blackout tonight": an analysis of the motives to drink to blackout expressed on Twitter. Alcohol Clin Exp Res 2019 Aug 10;43(8):1769-1776 [FREE Full text] [doi: 10.1111/acer.14132] [Medline: 31373703]
31. Thackeray R, Neiger BL, Burton SH, Thackeray CR. Analysis of the purpose of state health departments' tweets: information sharing, engagement, and action. J Med Internet Res 2013 Nov 11;15(11):e255 [FREE Full text] [doi: 10.2196/jmir.3002] [Medline: 24217361]
32. Lee JL, DeCamp M, Dredze M, Chisolm MS, Berger ZD. What are health-related users tweeting? A qualitative content analysis of health-related users and their messages on twitter. J Med Internet Res 2014 Oct 15;16(10):e237 [FREE Full text] [doi: 10.2196/jmir.3765] [Medline: 25591063]
33. Brunner M, Palmer S, Togher L, Dann S, Hemsley B. “If I knew what I was doing on Twitter then I would use it more”: Twitter experiences and networks of people with traumatic brain injury (TBI). Brain Impairment 2019 Jun 04;21(1):1-18. [doi: 10.1017/BrImp.2019.12]
34. Hemsley B, Dann S, Palmer S, Allan M, Balandin S. "We definitely need an audience": experiences of Twitter, Twitter networks and tweet content in adults with severe communication disabilities who use augmentative and alternative communication (AAC). Disabil Rehabil 2015 Jun 10;37(17):1531-1542 [FREE Full text] [doi: 10.3109/09638288.2015.1045990] [Medline: 26030339]
35. Lwin MO, Lu J, Sheldenkar A, Schulz PJ, Shin W, Gupta R, et al. Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. JMIR Public Health Surveill 2020 May 22;6(2):e19447 [FREE Full text] [doi: 10.2196/19447] [Medline: 32412418]
36. Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, et al. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res 2015 Oct 27;17(10):e243 [FREE Full text] [doi: 10.2196/jmir.4969] [Medline: 26508089]
37. Keller MS, Park HJ, Cunningham ME, Fouladian JE, Chen M, Spiegel BM. Public perceptions regarding use of virtual reality in health care: a social media content analysis using Facebook. J Med Internet Res 2017 Dec 19;19(12):e419 [FREE Full text] [doi: 10.2196/jmir.7467] [Medline: 29258975]
38. Zhang S, Gosselt JF, de Jong MD. How large information technology companies use Twitter: arrangement of corporate accounts and characteristics of Tweets. J Bus Tech Commun 2020 Jun 22;34(4):364-392 [FREE Full text] [doi: 10.1177/1050651920932191]
39. Muller A. Think tanks on Twitter: what the data tells us. Cast From Clay. URL: https://castfromclay.co.uk/models-research/think-tanks-on-twitter-what-the-data-tells-us/ [accessed 2022-02-23]
40. Banerjee R, Kelkar AH, Logan AC, Majhail NS, Pemmaraju N. The democratization of scientific conferences: Twitter in the era of COVID-19 and beyond. Curr Hematol Malig Rep 2021 Apr 31;16(2):132-139 [FREE Full text] [doi: 10.1007/s11899-021-00620-w] [Medline: 33788125]
41. Shen C, Ho J, Ma H. Temporal trend analysis on virtual reality using social media mining. In: Visvizi A, Lytras MD, editors. Research & Innovation Forum 2019: Technology, Innovation, Education, and their Social Impact. Cham, Switzerland: Springer; 2019:189-198.

XSL•FO
RenderX

42. Laurell C, Sandström C, Berthold A, Larsson D. Exploring barriers to adoption of virtual reality through social media analytics and machine learning – an assessment of technology, network, price and trialability. Journal of Business Research 2019 Jul;100:469-474 [FREE Full text] [doi: 10.1016/j.jbusres.2019.01.017]
43. Birckhead B, Khalil C, Liu X, Conovitz S, Rizzo A, Danovitch I, et al. Recommendations for methodology of virtual reality clinical trials in health care by an international working group: iterative study. JMIR Ment Health 2019 Jan 31;6(1):e11973 [FREE Full text] [doi: 10.2196/11973] [Medline: 30702436]
44. Cikajlo I, Cizman Staba U, Vrhovac S, Larkin F, Roddy M. A cloud-based virtual reality app for a novel telemindfulness service: rationale, design and feasibility evaluation. JMIR Res Protoc 2017 Jun 05;6(6):e108 [FREE Full text] [doi: 10.2196/resprot.6849] [Medline: 28583904]
45. Pennington DL, Reavis JV, Cano MT, Walker E, Batki SL. The impact of exercise and virtual reality executive function training on cognition among heavy drinking veterans with traumatic brain injury: a pilot feasibility study. Front Behav Neurosci 2022 Mar 22;16:802711 [FREE Full text] [doi: 10.3389/fnbeh.2022.802711] [Medline: 35391785]
46. McKee R. Ethical issues in using social media for health and health care research. Health Policy 2013 May;110(2-3):298-301. [doi: 10.1016/j.healthpol.2013.02.006] [Medline: 23477806]
47. Conway M. Ethical issues in using Twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature. J Med Internet Res 2014 Dec 22;16(12):e290 [FREE Full text] [doi: 10.2196/jmir.3617] [Medline: 25533619]
48. Brunner M, Hemsley B, Dann S, Togher L, Palmer S. Hashtag #TBI: a content and network data analysis of tweets about Traumatic Brain Injury. Brain Inj 2018 Dec 08;32(1):49-63. [doi: 10.1080/02699052.2017.1403047] [Medline: 29219635]
49. Palmer S. Characterizing Twitter communication – a case study of international engineering academic units. J Mark High Educ 2014 May 12;24(2):257-273. [doi: 10.1080/08841241.2014.907220]
50. Dann S. Benchmarking micro-blog performance: Twitter content classification framework. In: Burkhalter JN, Wood NT, editors. Maximizing Commerce and Marketing Strategies Through Micro-Blogging. Hershey, PA: IGI Global; 2015:313-332.
51. Hemsley B, Palmer S, Goonan W, Dann S. Motor neurone disease (MND) and amyotrophic lateral sclerosis (ALS): social media communication on selected #MND and #ALS tagged tweets. In: Proceedings of the 50th Hawaii International Conference on System Sciences. 2017 Presented at: HICSS '17; January 4-7, 2017; Maui, HI. [doi: 10.24251/hicss.2017.455]
52. Bruns A, Stieglitz S. Towards more systematic Twitter analysis: metrics for tweeting activities. Int J Soc Res Methodol 2013 Mar;16(2):91-108 [FREE Full text] [doi: 10.1080/13645579.2012.756095]
53. Hemsley B, Palmer S. Two studies on twitter networks and tweet content in relation to amyotrophic lateral sclerosis (ALS): conversation, information, and 'diary of a daily life'. Stud Health Technol Inform 2016;227:41-47 [FREE Full text] [Medline: 27440287]
54. Twitter privacy policy. Twitter Inc. URL: https://twitter.com/en/privacy [accessed 2022-07-01]
55. Kim H, Jang SM, Kim S, Wan A. Evaluating sampling methods for content analysis of Twitter data. Soc Media Soc 2018 May 02;4(2):205630511877283 [FREE Full text] [doi: 10.1177/2056305118772836]
56. NCapture (Version 1.0.290.0). QSR International. 2011. URL: https://chrome.google.com/webstore/detail/ncapture/lgomjifbpjfhpodjhihemafahhmegbek?hl=en [accessed 2023-11-07]
57. NVIVO (Version 12). QSR International. 2018. URL: https://www.qsrinternational.com/ [accessed 2023-11-07]
58. Microsoft Excel (Version 1908). Microsoft. URL: https://www.microsoft.com/en-in/microsoft-365/excel [accessed 2023-11-07]
59. Kalawsky R. Exploiting virtual reality techniques in education and training: technological issues. The Advisory Group on Computer Graphics. 2011. URL: http://www.agocg.ac.uk/reports/virtual/vrtech/title.htm [accessed 2019-08-02]
60. Hemsley B, Palmer S, Balandin S. Tweet reach: a research protocol for using Twitter to increase information exchange in people with communication disabilities. Dev Neurorehabil 2014 Apr 31;17(2):84-89. [doi: 10.3109/17518423.2013.861529] [Medline: 24380574]
61. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018 Dec;24(1):120-124 [FREE Full text] [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]
62. Gephi. The Gephi Consortium. URL: https://gephi.org/ [accessed 2023-11-07]
63. Fruchterman TM, Reingold EM. Graph drawing by force‐directed placement. Softw Pract Exp 2006 Oct 30;21(11):1129-1164. [doi: 10.1002/spe.4380211102]
64. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
65. Voorveld HA, van Noort G, Muntinga DG, Bronner F. Engagement with social media and social media advertising: the differentiating role of platform type. J Advert 2018 Feb 13;47(1):38-54. [doi: 10.1080/00913367.2017.1405754]
66. Brahm L. Ten reasons why we use Twitter. The Communications Network. URL: https://www.comnetwork.org/insights/why-do-we-tweet-here-are-10-reasons/#:~:text=1.,itself%20part%20of%20the%20story [accessed 2022-03-24]
67. Rosenstiel T, Sonderman J, Loker K, Ivancin M, Kjarval N. How people use Twitter in general. American Press Institute. URL: https://www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-in-general/ [accessed 2022-03-24]

XSL•FO
RenderX

68. Mahmoud K, Harris I, Yassin H, Hurkxkens T, Matar O, Bhatia N. Does immersive VR increase learning gain when compared to a non-immersive VR learning experience? In: Proceedings of the 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020 on Learning and Collaboration Technologies, Human and Technology Ecosystems. 2020 Presented at: HCII '20; July 19-24, 2020; Copenhagen, Denmark p. 480-498 URL: https://link.springer.com/chapter/10.1007/978-3-030-50506-6_33 [doi: 10.1007/978-3-030-50506-6_33]
69. Deaky BA, Parv AL. Virtual reality for real estate – a case study. IOP Conf Ser Mater Sci Eng 2018 Sep 19;399:012013 [FREE Full text] [doi: 10.1088/1757-899x/399/1/012013]
70. Beck J, Rainoldi M, Egger R. Virtual reality in tourism: a state-of-the-art review. Tour Rev 2019 Jun 12;74(3):586-612 [FREE Full text] [doi: 10.1108/tr-03-2017-0049]
71. Portman M, Natapov A, Fisher-Gewirtzman D. To go where no man has gone before: virtual reality in architecture, landscape architecture and environmental planning. Comput Environ Urban Syst 2015 Nov;54:376-384. [doi: 10.1016/j.compenvurbsys.2015.05.001]
72. Weidner K, Lowman J, Fleischer A, Kosik K, Goodbread P, Chen B, et al. Twitter, telepractice, and the COVID-19 pandemic: a social media content analysis. Am J Speech Lang Pathol 2021 Nov 04;30(6):2561-2571. [doi: 10.1044/2021_ajslp-21-00034]
73. Ainley E, Witwicki C, Tallett A, Graham C. Using Twitter Comments to Understand People's Experiences of UK health care during the COVID-19 pandemic: thematic and sentiment analysis. J Med Internet Res 2021 Oct 25;23(10):e31101 [FREE Full text] [doi: 10.2196/31101] [Medline: 34469327]
74. Smith M, Lee R, Shneiderman B, Hiamelbiom I. Mapping Twitter topic networks: From polarized crowds to community clusters. Pew Research Center. URL: https://www.pewresearch.org/internet/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/ [accessed 2022-07-18]
75. Yeung AW, Tosevska A, Klager E, Eibensteiner F, Laxar D, Stoyanov J, et al. Virtual and augmented reality applications in medicine: analysis of the scientific literature. J Med Internet Res 2021 Feb 10;23(2):e25499 [FREE Full text] [doi: 10.2196/25499] [Medline: 33565986]
76. Bryant L, Brunner M, Power E, Hemsley B. Content and network analysis of tweets tagged with #aphasia: an emergent community of practice. Aphasiology 2020 May 31;35(8):1084-1102. [doi: 10.1080/02687038.2020.1770195]
77. Tirdad K, Dela Cruz A, Sadeghian A, Cusimano M. A deep neural network approach for sentiment analysis of medically related texts: an analysis of tweets related to concussions in sports. Brain Inform 2021 Jul 01;8(1):12 [FREE Full text] [doi: 10.1186/s40708-021-00134-4] [Medline: 34212268]
78. Workewych AM, Ciuffetelli Muzzi M, Jing R, Zhang S, Topolovec-Vranic J, Cusimano MD. Twitter and traumatic brain injury: a content and sentiment analysis of tweets pertaining to sport-related brain injury. SAGE Open Med 2017;5:2050312117720057 [FREE Full text] [doi: 10.1177/2050312117720057] [Medline: 28890783]
79. Sullivan SJ, Schneiders AG, Cheang C, Kitto E, Lee H, Redhead J, et al. 'What's happening?' A content analysis of concussion-related traffic on Twitter. Br J Sports Med 2012 Mar;46(4):258-263. [doi: 10.1136/bjsm.2010.080341] [Medline: 21406451]
80. Vaezipour A, Aldridge D, Koenig S, Theodoros D, Russell T. a mixed-method investigation of clinician acceptance, barriers and enablers of virtual reality technology in communication rehabilitation. Disabil Rehabil 2022 Jul 14;44(15):3946-3958. [doi: 10.1080/09638288.2021.1895333] [Medline: 33715566]
81. Glegg SM, Levac DE. Barriers, facilitators and interventions to support virtual reality implementation in rehabilitation: a scoping review. PM R 2018 Nov;10(11):1237-51.e1 [FREE Full text] [doi: 10.1016/j.pmrj.2018.07.004] [Medline: 30503231]
82. Brassel S, Brunner M, Power E, Campbell A, Togher L. Speech-language pathologists' views of using virtual reality for managing cognitive-communication disorders following traumatic brain injury. Am J Speech Lang Pathol 2023 Mar 23;32(2S):907-923. [doi: 10.1044/2022_AJSLP-22-00077] [Medline: 36580534]
83. Bryant L, Sedlarevic N, Stubbs P, Bailey B, Nguyen V, Bluff A, et al. Collaborative co-design and evaluation of an immersive virtual reality application prototype for communication rehabilitation (DISCOVR prototype). Disabil Rehabil Assist Technol (Forthcoming) 2022 Apr 20:1-10. [doi: 10.1080/17483107.2022.2063423] [Medline: 35442823]
84. Lindner P, Miloff A, Hamilton W, Reuterskiöld L, Andersson G, Powers MB, et al. Creating state of the art, next-generation virtual reality exposure therapies for anxiety disorders using consumer hardware platforms: design considerations and future directions. Cogn Behav Ther 2017 Sep;46(5):404-420. [doi: 10.1080/16506073.2017.1280843] [Medline: 28270059]
85. Proffitt R, Glegg S, Levac D, Lange B. End-user involvement in rehabilitation virtual reality implementation research. J Enabling Technol 2019;13(2):92-100 [FREE Full text] [doi: 10.1108/JET-10-2018-0050] [Medline: 31663039]
86. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018 May 28;7(6):121 [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
87. Erskine N, Hendricks S. The use of Twitter by medical journals: systematic review of the literature. J Med Internet Res 2021 Jul 28;23(7):e26378 [FREE Full text] [doi: 10.2196/26378] [Medline: 34319238]
88. Brunner M, Hemsley B, Palmer S, Dann S, Togher L. Review of the literature on the use of social media by people with traumatic brain injury (TBI). Disabil Rehabil 2015 May 19;37(17):1511-1521. [doi: 10.3109/09638288.2015.1045992] [Medline: 25985717]

XSL•FO
RenderX

Abbreviations

HMD: head-mounted display
TBI: traumatic brain injury
VR: virtual reality

Edited by T Leung; submitted 19.12.22; peer-reviewed by J Hoepner, J O'Neil; comments to author 21.03.23; revised version received 06.07.23; accepted 29.10.23; published 19.01.24.

Please cite as:
Brassel S, Brunner M, Campbell A, Power E, Togher L
Exploring Discussions About Virtual Reality on Twitter to Inform Brain Injury Rehabilitation: Content and Network Analysis
J Med Internet Res 2024;26:e45168
URL: https://www.jmir.org/2024/1/e45168
doi:10.2196/45168
PMID:38241072

©Sophie Brassel, Melissa Brunner, Andrew Campbell, Emma Power, Leanne Togher. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Impact of Virtual Reality on the Behavioral and Psychological Symptoms of Dementia and Quality of Life of Inpatients With Dementia in Acute Care: Randomized Controlled Trial (VRCT)

Lora Appel[1,2,3], PhD; Eva Appel[2], MA; Erika Kisonas[2], BHSc; Samantha Lewis-Fung[2], MHSc; Susanna Pardini[4], MSc, PsyD; Jarred Rosenberg[3], MD; Julian Appel[5], MBA, BScEE; Christopher Smith[3], MD

[1]School of Health Policy and Management, Faculty of Health, York Universtiy, Toronto, ON, Canada
[2]OpenLab, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada
[3]Michael Garron Hospital, Toronto, ON, Canada
[4]University of Padua, Padua, Italy
[5]Ted Rogers School of Management, Toronto Metropolitan University, Toronto, ON, Canada

Corresponding Author:
Lora Appel, PhD
School of Health Policy and Management
Faculty of Health
York Universtiy
4700 Keele Street
426 Health, Nursing and Environmental Studies Building
Toronto, ON, M3J 1P3
Canada
Phone: 1 416 736 2100 ext 22653
Email: lora.appel@yorku.ca

Abstract

Background: Virtual reality (VR) is increasingly considered a valuable therapeutic tool for people with dementia. However, rigorous studies are still needed to evaluate its impact on behavioral and psychological symptoms of dementia (BPSDs) and quality of life (QoL) across care settings.

Objective: The primary aim of this study was to evaluate the impact of VR therapy on managing BPSDs, falls, length of stay, and QoL in inpatients with dementia admitted to an acute care hospital. The secondary aim was to evaluate the intervention's feasibility in terms of acceptability, safety, and patient experience.

Methods: A prospective, open-label, mixed methods, randomized controlled clinical trial was conducted between April 2019 and March 2020. A total of 69 participants (aged ≥65 years with a diagnosis of dementia and who did not meet the exclusion criteria) were randomly assigned to either the control (n=35, 51%) or VR (n=34, 49%) arm. Participants in the experimental (VR) arm were visited by a researcher and watched 360° VR films on a head-mounted display for up to 20 minutes every 1 to 3 days, whereas individuals in the control arm received standard of care. Instances of daily BPSDs and falls were collected from nurses' daily notes. QoL was measured through semistructured interviews and the Quality of Life in Late-Stage Dementia scale. Structured observations and semistructured interviews were used to measure treatment feasibility. The primary outcomes were analyzed at a 95% significance level based on the intention-to-treat method.

Results: VR therapy had a statistically significant effect on reducing aggressiveness (ie, physical aggression and loud vociferation; P=.01). Substantial impact of VR therapy was not found for other BPSDs (eg, apathy), falls, length of stay, or QoL as measured using the Quality of Life in Late-Stage Dementia scale. The average VR therapy session lasted 6.8 (SD 6.6; range 0-20) minutes, and the intervention was overall an acceptable and enjoyable experience for participants. No adverse events occurred as a result of VR therapy.

Conclusions: Immersive VR therapy appears to have an effect on aggressive behaviors in patients with dementia in acute care. Although the randomized controlled trial was stopped before reaching the intended sample size owing to COVID-19 restrictions,

XSL•FO
RenderX

trends in the results are promising. We suggest conducting future trials with larger samples and, in some cases, more sensitive data collection instruments.

Trial Registration: ClinicalTrials.gov NCT03941119; https://clinicaltrials.gov/study/NCT03941119

International Registered Report Identifier (IRRID): RR2-10.2196/22406

(J Med Internet Res 2024;26:e51758) doi:10.2196/51758

KEYWORDS

digital therapies; nonpharmacological; cognitive impairment; behavioral and psychological symptoms of dementia; BPSDs; randomized controlled trial; virtual reality

Introduction

Managing Behavioral and Psychological Symptoms of Dementia in Acute Care

Behavioral and psychological symptoms of dementia (BPSDs) constitute a heterogeneous cluster of signs and symptoms [1,2] that are present in up to 90% of people with dementia [3] and are difficult to manage [4]. In acute care settings, approximately 75% of inpatients with dementia display BPSDs during their stay [5] as stress and hyperstimulation from being in a hospital environment are known exacerbating factors [6].

BPSDs are linked to key acute care outcomes, including falls [7] and longer length of stay (LoS) [8-10], fueling a cycle of decline. Longer stays can result in negative consequences such as declines in cognition and function [11], which in turn are associated with a significantly higher risk of hospitalization and institutionalization [12-19] and increased health care spending [8,20,21]. Moreover, aggression and agitation, the most frequently displayed BPSDs in this setting [9], mediate symptoms of caregiver burden and decrease quality of life (QoL) and well-being in people with dementia and caregivers alike [8,21].

The current means to manage BPSDs in acute care are insufficient and, at times, pose ethical challenges. Pharmacological therapy, often in the form of atypical antipsychotics, remains the most common treatment for BPSDs [22]. Even with newer drugs in this class, clinical outcomes are poor, and undesired side effects such as cognitive worsening, confusion [22], and extrapyramidal symptoms are prevalent [23]. Other common alternatives for managing BPSDs include using physical barriers such as wrist or ankle restraints, alarms, locks, Buxton chairs, and tethers. Unfortunately, these approaches have a wide range of negative consequences for psychological and physical functioning in people with dementia (ie, lethargy, cognitive decline, anxiety, distress, aggression, increased risk of falls, and pressure sores) [24,25].

For some time, guidelines for managing BPSDs have recommended using person-centered [26-28] psychosocial nonpharmacological interventions as a first-line approach [29,30], and qualitative studies exploring the in-hospital experiences of people with dementia have echoed the need for improved ward-based activity services [31,32]. Some examples considered to be gold standards include the use of physical presence and therapeutic touch, recreational or social activities, behavioral interventions, individual and group psychological or psychosocial interventions, environmental interventions, daily routines, music-based therapies [33,34], or combinations of these [35,36]. Notably, a recent review concluded that nonpharmacological interventions are a safer alternative to pharmacological interventions and can effectively reduce certain BPSDs [37]. Exposure to natural stimuli (eg, views of trees, flowers, or outdoor environments) is a nonpharmacological intervention that has been consistently shown to (1) enhance well-being and QoL; (2) reduce depression, anxiety, agitation or aggression, and the use of analgesics; and (3) decrease hospital LoS for inpatients [38-41]. A landmark study by Ulrich [38] showed that patients who had a window with a view of nature rather than a brick wall had reduced use of analgesics and a shortened LoS in hospital. Since then, several studies have shown that exposure to nature reduces agitation or aggression and promotes feelings of joy and relaxation in people with dementia [39-41]. Even the presence of an ornamental plant in the hospital room has significant positive effects on stress, anxiety, fatigue, and systolic blood pressure in patients recovering from surgery [42].

However, providing people with dementia with exposure to nature is challenging in acute care settings. A key difficulty is that, when individuals with dementia are admitted to hospital, they leave their beds infrequently because of existing mobility challenges [43] or hospital-acquired complications [44]. Although assistance with ambulation is both beneficial for maintaining preadmission functional status [44] and perceived by staff as important for safety [31,45], patients' basic needs often go unmet because of staff workload [31].

Virtual Reality Interventions

One way in which people who are unable to engage directly with a natural environment can still be exposed to nature is by using virtual reality (VR) technology. VR uses a head-mounted display (HMD) to generate simulated immersive experiences that create a sense of presence. The sensation of being present in a different physical place is achieved by simultaneously stimulating multiple senses (typically vision and hearing), obscuring the real environment and replacing it with an artificial "virtual" environment [46].

There has been a recent increase in interest in the therapeutic use of VR with older adults in general [47-50], with most of the studies investigating the impact of VR therapy on well-being and BPSDs in people with dementia published since 2018 [51]. VR interventions for people with dementia constitute a cost-effective, noninvasive, and ethically acceptable alternative for managing BPSDs by redirecting the patient's attention within

XSL•FO
RenderX

a physically safe environment [52,53]. To date, these interventions have typically used music and sensory and mental stimulation with the aim of improving cognitive functions (attention, executive function, and visual and verbal memory), psychological symptoms (agitation, depression, anxiety, and apathy) [46,50,54], and the retrieval of autobiographical memories [51,52,55]. It has been suggested that VR therapy can be deployed in place of pharmacological therapy to improve well-being and QoL [56]; however, further research is necessary to establish its effectiveness for this purpose [51].

Our prior work has shown that it is feasible for frail older adults with varying degrees of physical mobility and cognitive impairment to wear a VR HMD and engage with natural VR environments in a community care setting [52,57]. These findings are consistent with those of other research showing that exposure to a VR forest resulted in increased feelings of pleasure and greater levels of alertness in people with dementia [58]. A growing number of studies have also shown significant positive results in using VR to manage acute pain in hospitalized patients without cognitive impairment [59]. To summarize, although VR offers an apparent solution for providing access to natural scenery to inpatients with limited mobility, its use with inpatients with dementia remains largely unexplored.

The methodology and aims of this randomized controlled trial (RCT) were directly informed by our previous studies: (1) a multisite feasibility study in 2 rehabilitation facilities (a long-term care facility and a dementia day program center) and (2) a pilot study conducted in the hospitalist medicine department of Michael Garron Hospital (MGH). The feasibility study recruited 66 outpatients with cognitive, sensory, and mobility impairments and demonstrated that it is safe and feasible to administer VR therapy in this population using an HMD [57]. The pilot study recruited 10 inpatients with dementia to validate the proposed RCT study protocol (processes, methods, recruitment strategy, resource requirements, and timelines) [52,53]. In both studies, participants reported on average more positive than negative feelings after experiencing nature scenarios, tolerated the VR hardware, and were able to physically explore the 360° virtual environments by turning their heads.

Aims and Hypothesis

To the best of our knowledge, an in-depth analysis of VR's impact on people with dementia has not yet been carried out in this setting. The main objectives of this RCT were to evaluate the impact of VR therapy on BPSDs and associated patient outcomes during the hospital stay. Our secondary objectives were to report on the acceptability, safety, experience, and satisfaction with VR therapy among inpatients with dementia. Specifically, we predicted that, for older adults with dementia admitted to an acute care hospital, (1) VR therapy would help manage BPSDs (ie, decrease anxiety, aggression, depression, violent behaviors, and incidents of wandering) and may decrease the number of incidents that require restraints, LoS, and the number of falls; (2) VR therapy would improve QoL (operationalized through validated instruments that measure QoL for people with dementia); and (3) VR therapy is safe and feasible (in terms of acceptability and patient experience) to administer.

Methods

Study Population

Inpatients aged >65 years with a documented diagnosis of dementia were recruited from MGH in the hospitalist or internal medicine ward. Data collection occurred between April 29, 2019, and March 13, 2020. The study was temporarily stopped under an MGH Research Ethics Board directive on March 26, 2020, when the first wave of COVID-19 peaked in Canada. Subsequently, recruitment was permanently ceased owing to the ongoing pandemic and restrictions on research, specifically with this vulnerable population.

Study Design

This study was an open-label RCT conducted at MGH, a community teaching hospital located in Toronto, Ontario, in collaboration with OpenLab, an innovation center at the University Health Network. The clinical trial protocol was registered at ClinicalTrials.gov (NCT03941119).

Individuals were randomly assigned to either the VR arm or the control arm. During their hospital stay, both groups received a visit from the research coordinator (RC) every 24 to 72 hours to complete a QoL assessment. Patients in the VR arm then participated in a VR therapy session for up to 20 minutes. This dosage regimen was informed by previous studies in which VR was used to promote well-being and manage BPSDs in people with dementia [60], as well as by other nonpharmacological interventions in acute care [51]. Those in the control arm were not exposed to VR therapy and otherwise followed the standard of care (ie, medical treatment for the presenting issue and management of any psychological and behavioral symptoms based on standard guidelines for hospitals in Ontario).

Ethical Considerations

Approval for research involving human participants was received from the MGH Research Ethics Board (782-1812-Mis-332) on January 14, 2019.

The consent process followed Ontario's legislation [61], and the hospital provided a brochure [62] to help patients and families understand the different roles involved in decision-making. Any previously documented capacity assessments were respected. For potential participants with no record of capacity assessment, the capacity to provide informed consent was obtained as per the study site protocol. If the patient had a documented capacity assessment determining that they were able to provide informed consent, the RC approached them in person to ask about their interest in the study. In cases in which a capacity assessment was required or a capacity assessment determined that the patient was unable to provide informed consent, the substitute decision-maker (SDM) was contacted as early into the patient's stay as possible. To ensure patient assent in all cases possible, informed SDM consent was provided with the patient present. The RC provided a copy of the informed consent or assent form and an informational study brochure (Multimedia Appendices 1-3) for review to all

XSL•FO
RenderX

interested participants and their SDMs. For participants in the VR arm, the RC obtained verbal or physical assent (body language indicating that the person agreed with the procedure or activity) from the participant immediately preceding any VR therapy session.

Data were anonymized and deidentified at the time of collection (ie, linked only to the study ID). Participants were not offered an honorarium or any form of compensation for taking part in the study. The patients provided written informed consent to allow their image to be published.

Procedures

Screening

All new patient admissions to the hospitalist or internal medicine department at MGH were screened daily, excluding weekends and statutory holidays, using the hospital's electronic medical record (EMR) system. Inclusion criteria required that participants be aged ≥65 years, have a documented diagnosis of dementia, and be admitted as inpatients at MGH. Patients were excluded if they had open facial wounds (sutured lacerations exempted), a history of seizures or epilepsy, a pacemaker, head trauma or stroke leading to their current admission, cervical conditions that would make use of a VR headset unsafe, or alcohol-related dementia or Wernicke-Korsakoff syndrome or were admitted to the intensive care unit or adult mental health inpatient service. Patients were also excluded if they had participated in the current RCT during a previous hospitalization or were readmitted <30 days after discharge. Finally, patients with no contactable SDM or who had a Public Guardian and Trustee as SDM were also excluded. Before recruitment, potential participants were unknown to the RC who administered the experimental treatment.

Randomization

After obtaining informed consent, the participants were randomized at a 1:1 allocation ratio to either the VR arm or the control arm. Random allocation to 1 of the 2 study arms was achieved using a digital tool performing Maximally-Tolerated-Imbalance randomization, a method designed to mitigate selection biases, simultaneously lowering allocation predictability and protecting against chronological bias [63]. After randomization, the participants, caregivers, and research personnel were not blind to treatment allocation given the nature of the intervention. The ward staff were blinded to treatment allocation to the greatest extent possible. Participants and caregivers were asked not to discuss their arm allocation with the ward staff. The sessions were scheduled to accommodate the daily activities of patients and their care teams with respect to their acute care needs and were facilitated by a research team member who was trained in dementia care.

Sample Size

With no previous experiments from which to obtain expected significant differences or sample SDs, we calculated our sample size based on our feasibility study [57], which collected pretest-posttest changes in emotional state using a modified version of the State-Trait Anxiety Inventory tool. Our feasibility study captured pretest-posttest State-Trait Anxiety Inventory results by category (called "Emotional Modality"). We computed pooled sample SDs using the aggregated results to obtain the required sample sizes for a significance level of 95% and a power of 80%. The results for the categories considered "positive" (eg, calm, relaxed, and content) had a mean estimated sample size of 181. Thus, at an assumed attrition rate of 20%, it was determined that 225 participants would need to be recruited to obtain a final sample of 180 patients.

Intervention

Hardware

The Oculus Go (Meta Reality Labs), a stand-alone mobile VR HMD system (discontinued on June 23, 2020), was selected based on its relative affordability and portability [64] and reduced likelihood of inducing simulator sickness owing to its low motion latency. This HMD requires no external hardware (ie, sensors to track head movement). It has built-in headphones, eliminating a number of difficulties raised by using external headphones (ie, cumbersome to mount and fit during VR therapy and increased time fulfilling hospital infection prevention and control requirements). External headphones (Sennheiser HD 221) were available for use if participants chose to use them or if the RC determined that they would be particularly helpful in certain situations (ie, when a participant was very hard of hearing or in a noisy environment). Figure 1 depicts an older patient wearing a VR HMD without external headphones (informed consent was provided by the patient for the publication of this photograph).

XSL•FO
RenderX

Figure 1. Photo of a patient wearing a virtual reality head-mounted display (without external headphones).

Software

A library of short 360° films was assembled based on previous studies and expert consultation. The VR films were intentionally chosen to be calming in distinctive ways as supported by the literature on nature visualization and well-being. By varying the (1) length of the films (1-, 3-, or 5-min clips), (2) types of natural elements (greenery, sky, and water), (3) distance of view (close-ups and vistas), (4) motion (flowing water and wind in trees), (5) sounds (leaves rustling and music from a live band), and (6) presence of different elements (eg, people in social settings, remote nature scenery, and urban sites), the content provided a wide range of features that could be collaboratively selected by the participant, caregiver, or health care provider for each individual in the VR arm. The scenes were designed specifically for people with dementia. First-person movement was limited to reduce the potential for simulator sickness, as was the use of excessive sound, light, or movement to avoid overstimulation. Figure 2 displays a 2D screen capture of 2 of the 7 VR scenes shown during the study (scene 2, “Open Field with Foliage,” and scene 5, “Aquamarine Beach”). The films were accessed during the study sessions by the researcher from the VR HMD’s internal storage.

Figure 2. 2D screen capture of nature scenes from 2 of the 360° films shown to patients.

Session Protocol

This study used a mixed methods design. Measures included baseline demographics (diagnoses, cognitive impairment, delirium, and comorbidities) and factors related to the participants’ hospital care experience (falls with or without injury, LoS, and nurse documentation of daily BPSDs) collected from the hospital’s EMR. During their hospital stay, recruited patients were visited by the RC every 24 to 72 hours (1-3 days), who documented findings using a validated QoL scale, a standardized observational tool, semistructured interviews with participants and their caregivers, and qualitative observations. Any episodes of adverse events (such as vertigo, dizziness, nausea, neck pain, and headaches) were documented. Table 1 provides a comprehensive list of the outcomes of interest, their evaluation methods, and the frequency of collection.

Before any study session, depending on participant allocation (VR or control arm), the RC consulted ward staff to ensure that the participant was in a stable condition for VR therapy and had

XSL•FO
RenderX

no scheduled treatment or tests in the following hour (see Multimedia Appendix 4 for the presession checklist). As per the study geriatrician's recommendations, if the participant was asleep upon arrival, the RC attempted to gently wake up the patient by saying their name a few times or returned later in the day if the patient remained asleep. If able and willing to sit, the participant was then encouraged to sit upright. Otherwise, they participated in the session lying down with the bed tilted to an upright position. If a participant required assistance to move to an appropriate position, the RC contacted a nurse for assistance.

During the first visit, the RC conducted a baseline demographic survey with the participant and their informed caregiver (if present). At the beginning of each study session, the RC engaged the participant in at least 5 minutes of unstructured conversation followed by a brief semistructured interview regarding their well-being during their hospital stay and then completed the modified Quality of Life in Late-Stage Dementia (QUALID) scale (Multimedia Appendix 5). Participants in the control arm then continued per the current standard of care (ie, the study visit was complete at this juncture, with no exposure to VR therapy). For VR participants, the RC placed the VR headset on the participant's head, assisted with adjusting the straps for a comfortable fit, and then observed while they watched 360° VR films for up to 20 minutes. During the intervention, the RC recorded observational notes pertaining to facial expressions, body language, and verbal feedback based on a standardized VR observational tool that the research team designed and used in previous VR studies (Multimedia Appendix 6; this tool has since been further refined [65]). After the VR therapy intervention, the RC conducted a semistructured interview with the participant and informal caregiver (if present) to gather their feedback on their experience and aspects related to the acceptability of VR therapy and confirmed the participant's willingness to take part in additional sessions during their hospital stay.

XSL•FO
RenderX

Table 1. Evaluation tools and methods by outcome of interest.

	Outcomes measured	Method or data collection tool
Primary objectives: VR[a] and control arms		
Baseline demographics[b]	Gender, age, primary language, highest level of education, current living state, relationship status, major visual and auditory conditions, use of glasses and hearing aids, head mobility, body mobility, mobility aids, dementia diagnosis, delirium severity, Charlson Comorbidity Index, MMSE[c], and MoCA[d]	1. EMR[e] 2. Baseline demographic survey with participant and caregiver (if present)
Changes in BPSDs[f,g]	NPI-10[h] cluster (delusions, hallucinations, agitation, depression, anxiety, apathy, irritability, euphoria, disinhibition, and aberrant motor behavior), aggression cluster (aggression, screaming or loud vocalization, restraints applied, security intervened, and bedside sitter), and wandering cluster	1. Key terms from daily notes in the EMR on patient symptoms were categorized into 3 clusters for analysis: 1. NPI-10–like 2. Violence 3. Wandering
Falls[g]	Fall precautions, falls with injury, and falls without injury	1. EMR
Length of stay[i]	Length of stay (in min)	1. EMR
Quality of life[j]	Interacts with others, smiles, appears sad, cries, demonstrates or verbalizes discomfort, enjoys touching or being touched, appears emotionally calm and comfortable, appears physically uncomfortable, enjoys eating, and sleeps well	1. In-Hospital Quality of Life Observational Scale—adapted from the QUALID[k] scale (includes semistructured interview)
Secondary objectives: VR group only		
VR acceptability[l]	Length of time viewing VR; pain, weight of device, comfort, and ability to wear glasses with device; and quality of visuals and sound	1. Structured observations 2. Semistructured interview
VR safety[l]	Nervous or anxious, confused or disoriented, nauseous, medical device interaction (eg, glasses and hearing aids), and adverse events	1. Structured observations 2. Semistructured interview
VR experience[l]	Enjoyment, relaxation, vocalizations, engagement, and reminiscence; opinions or preferences regarding content; and willingness to try VR again and recommend VR to a friend	1. Standardized VR observation tool 2. Semistructured interview

[a]VR: virtual reality.
[b]Collected during first study session.
[c]MMSE: Mini-Mental State Examination.
[d]MoCA: Montreal Cognitive Assessment.
[e]EMR: electronic medical record.
[f]BPSD: behavioral and psychological symptom of dementia.
[g]Reported daily (data set extracted once at discharge).
[h]NPI-10: Neuropsychiatric Inventory.
[i]Reported once at discharge.
[j]Conducted or assessed at each study session.
[k]QUALID: Quality of Life in Late-Stage Dementia.
[l]Conducted or assessed at each VR therapy session.

Analysis

Overview

Statistical analyses were conducted using the Minitab statistical package (Minitab, LLC) [66]. The Anderson-Darling test of normality was run for all variables to determine whether to use statistical regression or a nonparametric test. Linear regression was used for outcomes with normally distributed values (Anderson-Darling $P>.05$), with arm assignment (VR vs control) as the main factor. To determine whether differences could be attributed to the treatment effect rather than to individual characteristics, other factors included in the regression analyses were whether the participant had had VR therapy sessions, age group, gender, cognitive level, and dementia type (see Multimedia Appendix 7 for a complete description of the participants' baseline characteristics). Only data from the QUALID showed a normal distribution. All other measures were found to not be normally distributed (Anderson-Darling $P<.05$). The nonparametric Mood median test was chosen to measure differences in the distribution of the outcomes between the control and VR arms for these variables.

Primary Outcome Measures

Overview

The primary outcomes in this study pertained to the effects of VR therapy on BPSDs and other critical acute care outcomes.

XSL•FO
RenderX

Data were analyzed at a 95% significance level based on the intention-to-treat (ITT) and per-protocol analysis methods [67-69]. Outcomes that were collected regularly throughout the study (BPSDs, falls, and QoL) were measured as the difference ("after" minus "before") between the means of outcomes after and before the treatment, whereby we defined these as follows:

1. Before: the mean number of events or mean score up to (and including) the first session with the RC
2. After: the mean number of events or mean score after the first session with the RC, up to (and including) the last interaction with the RC or discharge

BPSD Outcome Measure

Data on BPSDs were collected from nurses' notes in the EMR that reported daily patient symptoms. Key terms from the nurses' notes were then categorized and analyzed based on the following predetermined clusters: (1) the Neuropsychiatric Inventory (NPI-10)–like cluster [70,71] (delusions, hallucinations, agitation, depression, anxiety, apathy, irritability, euphoria, disinhibition, and aberrant motor behavior), (2) the aggression cluster (aggression, screaming or loud vocalization, restraints applied, security intervened, and bedside sitter), and (3) the wandering cluster. Each cluster was analyzed separately. The categorization and analysis were completed by 2 coders and reviewed by a third, all of whom were blinded to arm assignment. It should be noted that, for the NPI-10–like cluster, the scale itself was not administered. In practice, the NPI-10 is typically conducted as a periodic measurement at wide time intervals. Thus, the possibility of administering the instrument was deemed impractical in the pilot study that informed this research [53]. Instead, the measurements were taken daily by focusing on keywords in the nurses' notes that corresponded to items described by the NPI-10 instrument.

Falls

Outcomes concerning falls were operationalized as the frequency of additional fall precautions applied, falls with injury, and falls without injury during the hospital stay as recorded in the EMR.

LoS Outcome Measure

The length of acute hospital stays was calculated as the difference in minutes between the discharge date/time and admission date/time as recorded in the EMR.

QoL Outcome Measure

QoL was measured using an adapted version of the QUALID scale [72]. This adapted scale (Multimedia Appendix 5) was administered by the RC at the beginning of each study session. The QUALID contains 11 items with total scores ranging from −22 to 22, with a higher score indicating a better in-hospital QoL. The scale was modified to include a brief semistructured interview in which participants were asked open-ended questions related to their in-hospital QoL and well-being. A 5-point visual scale was used to assist participants in answering the semistructured interview question ("How are you feeling today?").

RC notes detailing participants' responses to the semistructured interview questions were manually coded for overall sentiment (ie, indicating positive, negative, or neutral QoL or well-being at that moment) by 2 independent coders (SL-F and JA) who were blinded to arm assignment. Any discrepancies in coding were discussed by the 2 raters until an agreement was reached. Responses that were judged to be positive (eg, "Not bad. Everyone was pretty nice today"), neutral (eg, "So-so,"), and negative (eg, "Not too well.... Tough night, improving marginally") were scored as 1, 0, and −1, respectively. Unrelated and off-topic responses in which the participant's overall sentiment could not be clearly discerned (eg, "Yes" and "Is there someone there?") were excluded from the analysis.

Secondary Outcome Measures

Overview

Through the secondary outcomes, we sought to investigate the safety, acceptability, and feasibility of administering VR therapy to inpatients with dementia in an acute care setting. The secondary outcomes were analyzed by session and only for patients in the VR arm who received VR therapy. Quantitative data such as session length, observational tool scores, and responses to closed-ended interview questions were analyzed using descriptive statistics. Inferential methods were not used to examine secondary outcomes because of the small sample size. Observational data from the first sessions versus any subsequent sessions were compared manually to determine whether there were any clear trends that would warrant analysis by participant versus by session. Descriptive statistics revealed no clear patterns of differences that would warrant separate analysis by participant as opposed to by session.

Responses to open-ended semistructured interview questions, RC observations, and participant and caregiver comments during the session were analyzed thematically using the NVivo qualitative data analysis computer software package (Lumivero) by EK, SL-F, and LA [73]. The constant comparative method [74,75] was used to systematically explore the qualitative data and corroborate and expand upon the results from quantitative measures. This method involved using an open coding process over several iterations, with the authors meeting at each stage to discuss and compare emerging themes and contradictions within different sections of the data until core categories emerged [75].

Acceptability of VR Therapy

Acceptability was operationalized through the total time spent viewing VR, reported and observed aspects of comfort, and adequacy of the stimuli. Participants in the VR arm were asked whether the VR HMD was painful, too heavy, or otherwise uncomfortable and whether the visual and auditory stimuli were sufficiently provided (ie, good resolution and good sound). Structured observations of acceptability were recorded during sessions by the RC, including reasons for stopping sessions early.

Safety of VR Therapy

Safety was operationalized as any adverse events reported in the EMR as well as any feelings of nervousness or anxiety, confusion or disorientation, or medical device interaction (ie, buzzing from hearing aids). Data were collected from the EMR and through RC observations of VR therapy sessions and postsession semistructured interviews with the participants.

XSL•FO
RenderX

Participant Experience During VR Therapy

Participants' experiences during and immediately after VR therapy were assessed at each study session and operationalized through observations of their level of enjoyment, relaxation, engagement, and vocalizations, as well as the presence of reminiscence. The RC observed and rated participants' reactions to VR therapy using the standardized VR observation tool (Multimedia Appendix 6) during VR therapy and recorded any other behavior suggesting a pleasant or unpleasant experience. A semistructured interview immediately after VR therapy was conducted that included open-ended questions on experience as described previously as well as willingness to try VR therapy again, willingness to recommend VR to a friend, and interest in purchasing a VR device. Opinions and preferences regarding the themes or content of the VR films were also collected.

Results

Demographics

Of the 3514 new admissions to acute care during the study period, 300 (8.54%) met the inclusion criteria. After contacting the SDMs of eligible participants and excluding patients who met the exclusion criteria (ie, discharged or deceased) or refused to use the VR equipment, a total of 77 patients with dementia were consented and enrolled in the study. Participants were randomized to the control arm with no VR therapy exposure (39/77, 51%) and the intervention arm receiving VR therapy once every 1 to 3 days (38/77, 49%).

From each of the 2 arms, 4 participants were excluded (4/39, 10% from the control arm and 4/38, 11% from the intervention arm) as there was no encounter with the participants to allow for meaningful comparison (ie, the only interaction occurred on the last day of the hospital stay), leaving 69 participants in total, with 35 (51%) in the control arm and 34 (49%) in the experimental VR therapy arm. In the VR arm, 85% (29/34) of the participants completed at least one session of VR therapy, whereas 15% (5/34) refused the VR and completed no VR therapy sessions. Figure 3 shows a CONSORT (Consolidated Standards of Reporting Trials) diagram of patient flow based on ITT.

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) diagram. Patient flow based on intention-to-treat method. VR: virtual reality.

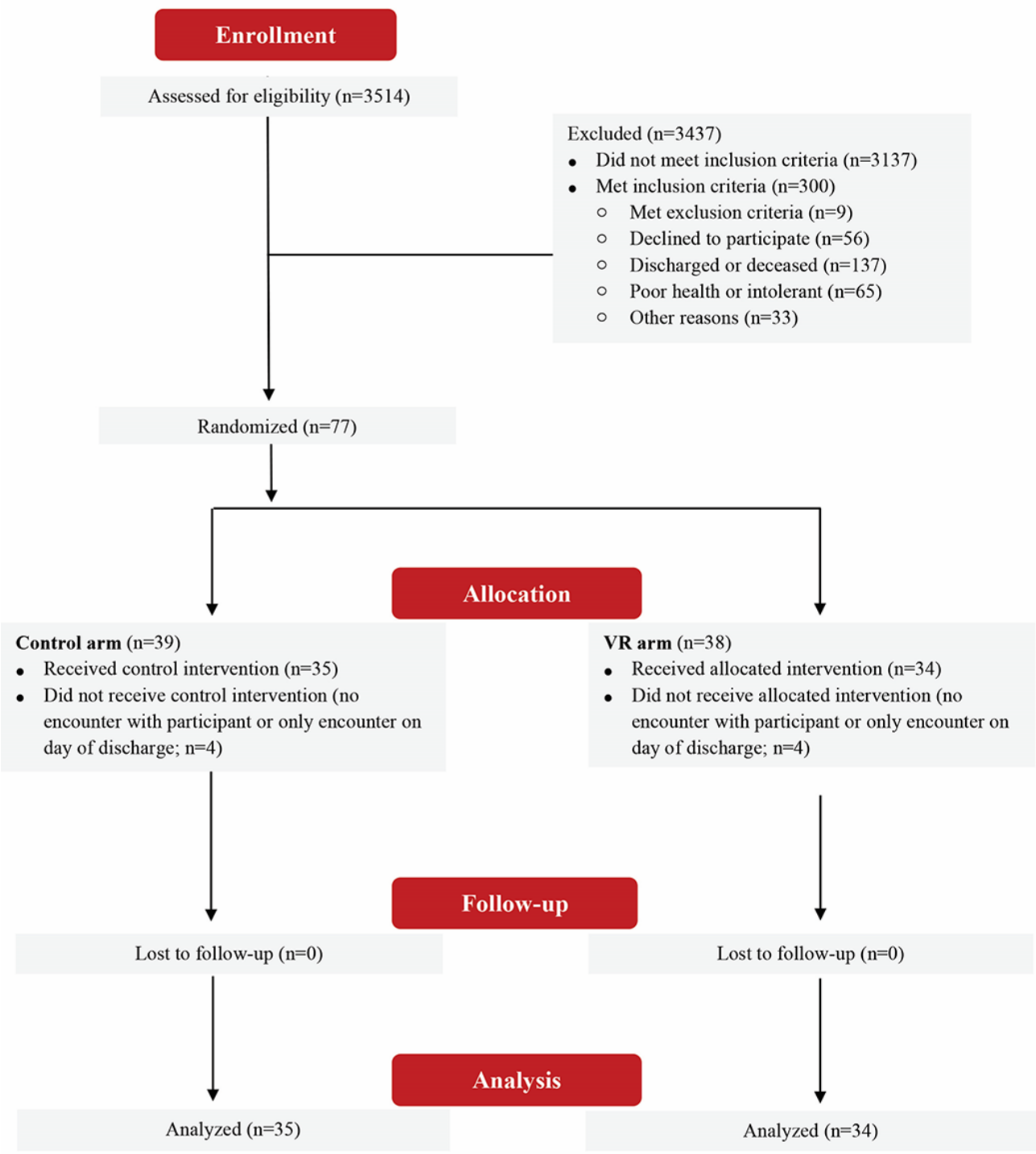

Multimedia Appendix 7 describes the baseline demographics of all study participants (n=69) as reported in clinical documentation based on 3 groups: the control arm and 2 VR arms, "VRx arm: With VRx-therapy" and "VRx arm: No VRx-therapy." Across arms, 65% (45/69) of the participants were female and belonged to the following age groups: 28% (19/69) were aged 65 to 79 years, 33% (23/69) were aged 80 to 89 years, and 39% (27/69) were aged 90 to 100 years. Over half of the participants had either a moderate (23/69, 33%) or severe (11/69, 16%) level of cognitive impairment, and nearly half either required assistance (7/69, 10%) or were unable (22/69, 32%) to sit. Vision and hearing aids were used by 62% (43/69) and 23% (16/69) of the participants, respectively.

To determine the possibility of a priori bias in the random assignment, a chi-square test for association was run on the pairwise categorical variables (arm assignment, whether the participant had had VR therapy, age group, gender, dementia diagnosis, and cognitive level). There was no statistically significant association found between any pairs in any distribution method (P>.05). Owing to the small sample of available data for dementia diagnosis, the test was also run with

XSL•FO
RenderX

a consolidated list (Alzheimer dementia vs other types of dementia), and no association was found (*P*>.05).

Results from the primary outcomes are described based on the ITT analysis method (n=69 total participants; 34/69, 49% in the VR arm and 35/69, 51% in the control arm). This more conservative method was chosen to minimize bias and provide a closer representation of the effectiveness of VR therapy in clinical practice [68]. The secondary outcomes were evaluated only for patients in the VR arm who received the intervention (29/34, 85%) and are reported per session (n=47 sessions) unless otherwise indicated.

Primary Outcome Measures

BPSD Outcome Measure

Arm assignment was found to have a statistically significant effect on the aggression cluster of BPSDs (Mood median test *P*=.01). Specifically, when comparing the before and after differences (overall median difference=0), participants in the VR arm had a significant decrease in physically aggressive behaviors and loud vocalizations compared with participants in the control arm (Mood median test *P*=.01; Table 2). The Mood median test revealed no statistically significant differences in the NPI-10–like cluster (*P*=.28), wandering cluster (*P*=.70), number of disruptions (*P*=.66), or number of medications refused (*P*=.54).

Table 2. Distribution of difference scores (mean number of aggressive events recorded in the electronic medical record after minus before treatment).

Difference[a] in the mean number of aggressive events	VR[b] arm (n=34), n (%)	Control arm (n=35), n (%)
More than overall median[c]	4 (12)	13 (37)
Equal to the overall median	19 (56)	17 (49)
Less than the overall median[c]	11 (32)	5 (14)

[a]Difference=after minus before treatment. Difference scores of less than the overall median indicate a relative decrease in aggressive events after virtual reality (VR) therapy. The VR arm had significantly fewer aggressive behaviors after treatment compared with the control arm (mood median test *P*=.01).

[b]VR: virtual reality.

[c]Overall median difference score=0.

Falls

Falls were found to have no significant dependence on arm assignment or any other associated factors (Mood median test *P*=.59).

Length of Hospital Stay

A Box-Cox logarithmic transformation was conducted on LoS to allow for regression requirements, and a linear regression was run for logarithmic LoS. The model satisfied the linearity independence, equal variance, and normality condition and the low variance inflation factor condition. The regression model showed an overall low explanation for the variance (adjusted R^2=11.5%). LoS showed no statistically significant dependence on arm assignment (*P*=.94).

QoL Outcome Measure

Regression analysis found no statistically significant dependence of arm assignment on QoL scores (QUALID instrument) or on any other factor. The regression model satisfied the linearity independence, equal variance, and normality condition and the low variance inflation factor condition but showed no explanation for the variance (adjusted R^2=0%).

Responses from 61% (42/69) of the participants (22/34, 65% from the VR arm and 20/35, 57% from the control arm) were included in the sentiment analysis of the semistructured QoL interview after excluding records in which insufficient data were collected to calculate a difference score (ie, cases with missing data because of excluded responses or in which the semistructured interview was not completed at least once in the before and after periods). When comparing the difference scores for the VR and control arms, the results from the Mood median test approached statistical significance (*P*=.07). Overall, participants in the VR arm had higher difference scores, meaning that, after treatment, they expressed relatively more positive and less negative sentiments when asked questions related to their current well-being and QoL (Table 3).

Table 3. Change in overall sentiment during quality of life (QoL) interview after minus before treatment—distribution of difference scores.

Difference in sentiment score[a]	VR[b] arm (n=34), n (%)	Control arm (n=35), n (%)
Positive	8 (24)	3 (9)
No change	8 (24)	5 (14)
Negative	6 (18)	12 (34)
Unable to assess	12 (35)	15 (43)

[a]Difference=after minus before treatment, where overall positive, neutral, and negative sentiment during the QoL interview were scored as 1, 0, and –1, respectively. Positive difference scores indicate an increase in positive sentiment after treatment.

[b]VR: virtual reality. The VR arm was overall more positive and less negative in their responses compared with the control arm, though this difference was not found to be statistically significant (mood median test *P*=.07).

XSL•FO
RenderX

Secondary Outcome Measures

Participants completed a mean of 1.6 sessions (SD 0.8) during their stay. Over half (17/29, 59%) of the participants completed 1 session, whereas approximately one-fifth completed 2 (6/29, 21%) and 3 (6/29, 21%) VR therapy sessions.

Acceptability of VR

Participants spent an average of 6.8 (SD 6.6; range 0-20) minutes viewing VR films during the sessions. Approximately half (15/29, 52%) of the participants took part in at least one VR session that lasted >5 minutes, whereas 38% (11/29) completed only shorter sessions lasting 1 to 5 minutes. A minority of the participants (3/29, 10%) took part in only a very short VR session or sessions lasting <1 minute, including one individual who opted to listen to music through headphones for 20 minutes instead of using VR because of blindness. For participants who completed more than one session (12/29, 41%), the mean duration of the first VR therapy session was 6.4 (SD 6.0; range 0-20) minutes, and the mean duration of subsequent sessions was 6.3 (SD 6.8; range 0-20) minutes. For participants who completed multiple sessions, the mean difference between the patients' shortest and longest session was 4.9 (SD 5.7; range 0-16) minutes.

Sessions that ended early were most often because of the participant's choice (ie, the patient removed the headset with no distress). The reasons for stopping treatment before the end of each 20-minute session are summarized in Table 4.

In some sessions (5/47, 11%), participants showed reluctance or apprehension or anxiety regarding trying VR. For example, one participant initially pushed the headset away, and another one held the HMD throughout the session and appeared reluctant to let go. When present in these cases, informal caregivers (family members) appeared to be helpful in encouraging participation. Additional measures of acceptability (audio and visual stimulus quality, VR HMD comfort, and issues encountered during the sessions) are presented in Table 5. Although most participants indicated that the stimulus quality and headset comfort were adequate, the concern most commonly reported by them was discomfort with the weight of the headset. Observed session issues were most often because of delays related to the acute care setting and unrelated to the VR device (eg, nurses completing a stability check).

Table 4. Reasons for stopping virtual reality (VR) therapy sessions (29 participants; n=47 sessions).

	VR sessions, n (%)
Full VR therapy session (20 min)	8 (17)
Stopped early (<20 min)	39 (83)
Participant's choice (no distress)	31 (66)
Participant appeared scared, confused, or dizzy (no distress)	3 (6)
Headset fit related	2 (4)
Participant choice (low interest)	1 (2)
Participant experiencing side effects (eye discomfort)	1 (2)
Participant had a test or treatment	1 (2)

XSL•FO
RenderX

Table 5. Acceptability of virtual reality (VR) therapy (29 participants; n=47 sessions).

	VR sessions, n (%)
Participant responses to postsession questions	
Stimulus quality	
Sound was clear	35 (74)
Image was in focus	33 (70)
The VR HMD[a]	
Was comfortable	35 (74)
Was heavy	11 (23)
Caused feelings of pressure on the nose or face	6 (13)
Felt tight	2 (4)
Was uncomfortable	2 (4)
Caused eye discomfort (too bright or watery eyes)	2 (4)
Observed session issues	
Challenges with HMD fit (difficulty adjusting the straps, following instructions [language barrier], and fitting glasses under the HMD)	5 (11)
Hardware or software technical difficulties (video on pause or headset and controller out of sync)	2 (4)
Delays unrelated to VR device (locating a nurse for a clinical stability check, patient leaving to use the washroom, patient eating lunch, needing to set up a language line call, confirming assent or dissent procedures, or patient anxious upon RC[b] arrival)	8 (17)

[a]HMD: head-mounted display.

[b]RC: research coordinator.

Safety

No serious side effects or adverse events related to having received VR therapy were observed or reported during the trial. Patients who experienced a serious adverse event during their hospital stay were withdrawn from the study. In the VR arm, adverse events (n=2) that occurred after the first encounter with the participant included stroke (n=1) and death (n=1). These events were determined to be unrelated to having received VR therapy. In the control arm, adverse events (n=5) that occurred after the first encounter with the participant included transfer to the intensive care unit (n=1), seizures (n=2), and death (n=2).

The side effects reported and observed during the VR therapy sessions were mild. During the postsession interview, participants reported feeling at least “a bit” nervous or anxious in 4% (2/47) of the sessions, at least “a bit” confused or disoriented in 4% (2/47) of the sessions, and at least “a bit” nauseous in 2% (1/47) of the sessions (Figure 4). The presence of buzzing sounds from hearing aids could not be evaluated as no participant was wearing a hearing aid during the VR sessions.

XSL•FO
RenderX

Figure 4. Patient report—virtual reality (VR) acceptability, safety, and experience evaluated from structured postsession feedback interviews (n=47 sessions).

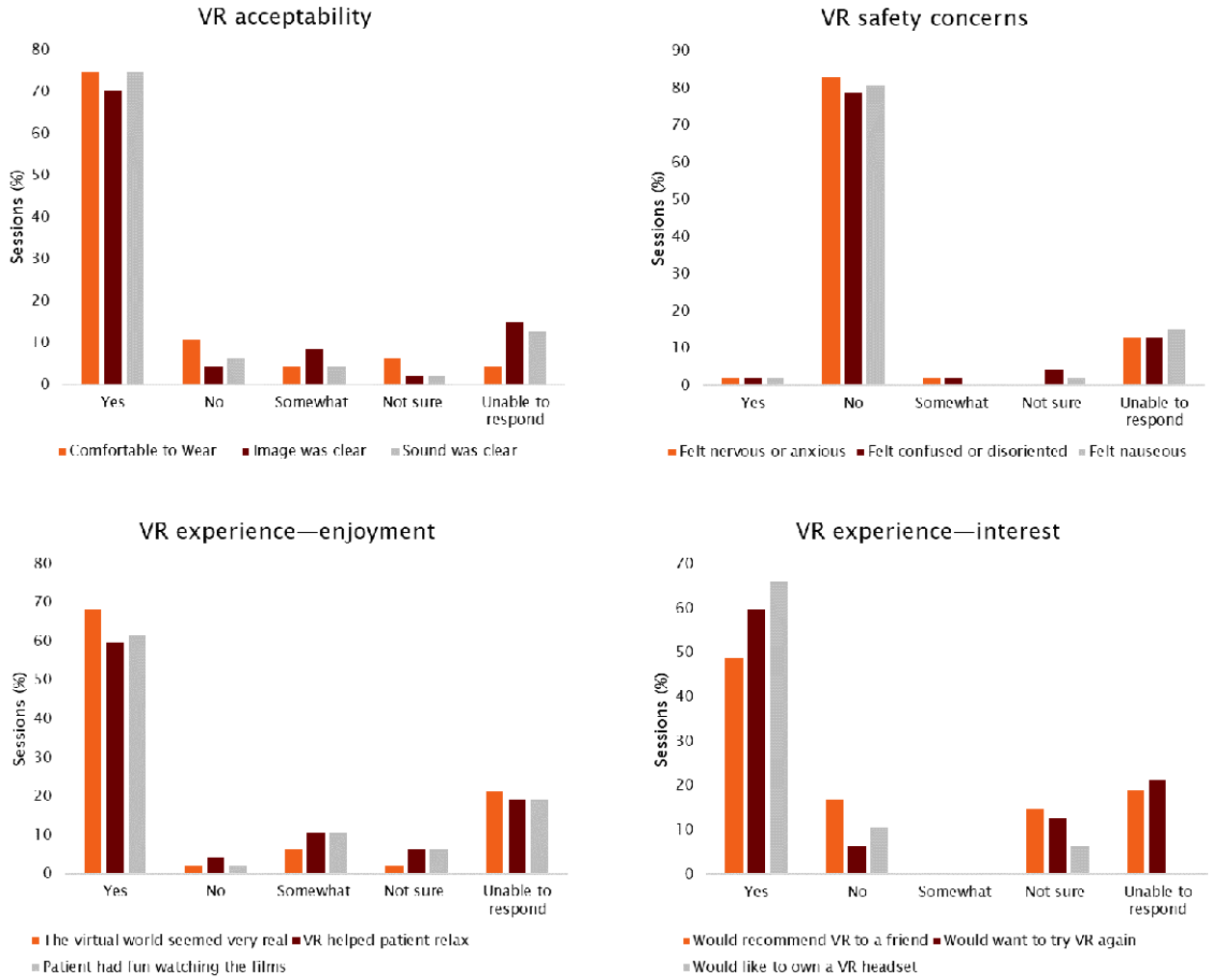

Participant Experience and Satisfaction With VR Therapy

Overview

Structured RC session observations and postsession semistructured interviews with participants indicated that most participants had a positive experience and were satisfied with the VR therapy. Figure 4 summarizes the results of the semistructured interviews. Figure 5 summarizes the results of structured observations using the standardized VR observation tool (vocalizations, enjoyment, relaxation, engagement, and reminiscence; Multimedia Appendix 6).

XSL•FO
RenderX

Figure 5. Standardized virtual reality (VR) observation tool results—observed reactions to VR therapy (n=47 sessions; *n=46 sessions).

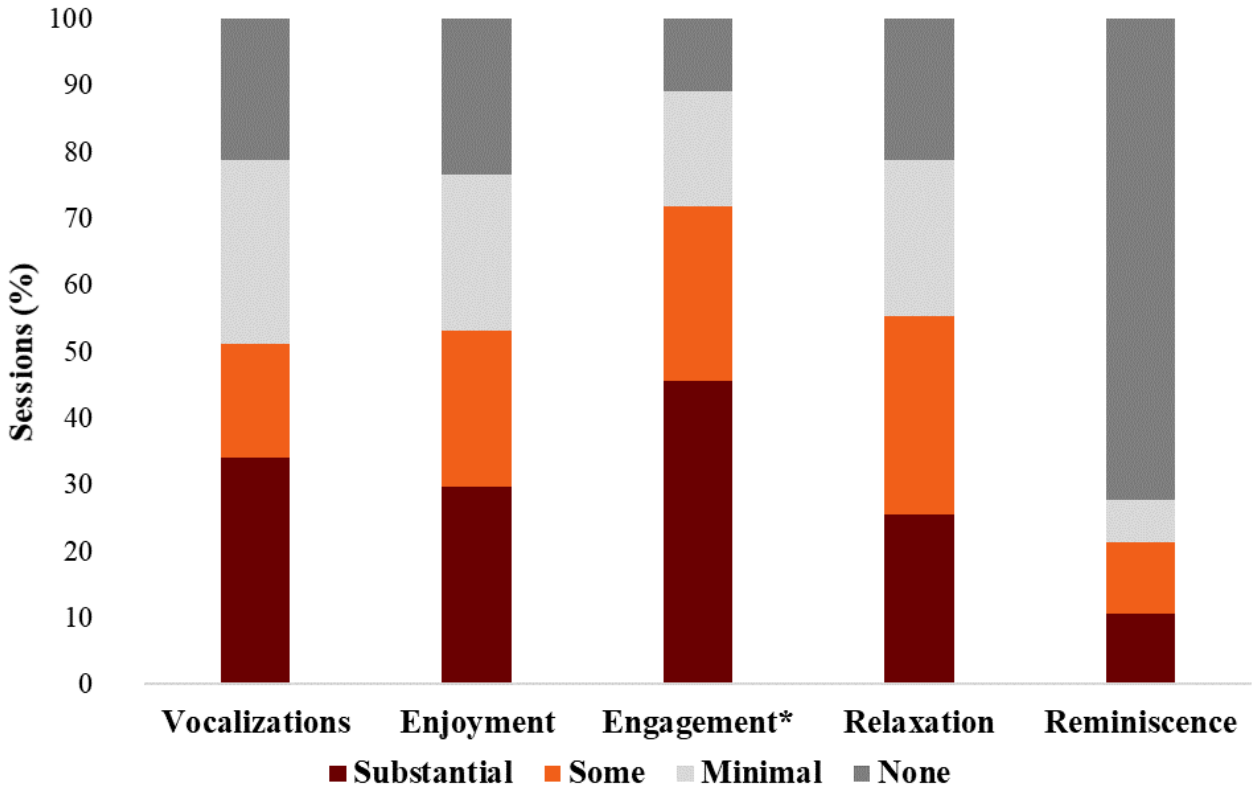

Engagement and Vocalizations

In most sessions, participants demonstrated some or substantial engagement (eg, commenting on or reacting to aspects of the scene) and agreed that the virtual world seemed very real to them. Participants who were more highly engaged interacted with the VR therapy with behaviors such as looking in all directions, pointing at objects or waving at people in the scene, and reacting to changes in scenes with vocalizations such as "Ooh" and "Ah" and moving along with the music. Many participants commented on or described aspects of the scene; for example, "Big trees stretching their hands...near me there is greenery again." Others who appeared less engaged benefited from prompting and encouragement by the RC or a family caregiver to look around within the scene.

Some participants were more engaged when their caregiver was present, and many appeared happy while chatting with family members present during sessions, often initiating conversations based on content observed in the VR (eg, "Look at this!"). Notably, one participant was observed calling his daughter a pet name that she reported that she had not heard in years.

Enjoyment and Reminiscence

Consistent with the RC observations, in most sessions, participants reported that they had fun watching the VR films. Participants who appeared to enjoy their experience were observed laughing and smiling. One participant exclaimed the following while watching the VR:

> *Wow, I am in the sky! I bless the nurses, they put me in the sky! I have a story to tell when I go home!*

One participant began reminiscing about his home country, Tanzania, when he saw water scenes; another described their experience as "Like how we used to watch the water edge [in Sri Lanka]." Another participant said the following:

> *It's like in the plane to Jamaica looking down [clapping hands].*

Sadness was observed in relatively few sessions (3/47, 6%) and was related to reminiscence in 2 cases.

In the postsession interviews, when asked, "How did you like that," participants in most of the sessions (32/47, 68%) responded positively and described their experience using words such as "very good," "very nice," "lovely," "beautiful," and "really enjoyed." One participant noted that they felt that VR was "good for here in the hospital," and one described their experience as "Fair enough, a free trip!" Participants in some sessions (6/47, 13%) were more neutral toward their experience or unable to respond to the question (6/47, 13%). Participants in relatively fewer sessions (3/47, 6%) were more negative when describing their experience and cited the weight of the HMD, discomfort, or having a bad day (eg, feeling shaky).

Relaxation

Also consistent with RC observations, in over half (28/47, 60%) of the sessions, participants reported that VR therapy helped them relax. In 6% (3/47) of the sessions, participants appeared notably calmer during VR therapy compared with before (ie, stopped restless behaviors such as leg movements or humming and grunting sounds), including one participant who was initially hesitant to try VR.

XSL•FO
RenderX

Interest in Trying VR Again

In approximately half (23/47, 49%) of the sessions, participants reported that they would like to own a VR headset, would recommend VR to a friend, and would like to try VR therapy again. Notably, in 13% (6/47) of the sessions, the participants or their caregivers spontaneously mentioned purchasing a VR headset to use at home. For example, one participant said that "I want you to buy for me, Costco maybe," and another said that "I have to get one of these machines." In total, 3 caregivers inquired about funding, affordable options, or lighter options for their loved ones with dementia to have access to VR at home.

VR Therapy Film Content Preferences

In most cases (37/47, 79%), participants chose to watch multiple mixed scenes during their VR sessions. Participants frequently commented on or reacted to aspects of water or nature scenes (eg, one participant said, "Ah-ha!" each time a water scene came on) or animals and interacted with the music in the videos. Some participants were also observed to fix their gaze in the direction of people in the videos (eg, musicians on a stage and children throwing rocks).

When asked what they wanted to see more of in the VR films, their responses varied. The most frequent suggestions included cultural topics (14/50, 28% of suggestions; eg, art exhibits, varied movie genres, scenic tourist locations, and films with religious locations or themes) and nature scenery (11/50, 22% of suggestions). Films with various animals and types of music were also suggested, as were films incorporating people or personally relevant topics (eg, a participant's dog). One individual suggested making VR a shared experience (ie, having 2 people involved with 2 HMDs).

Discussion

Principal Findings

The aims of this study were to investigate the effects of VR therapy on BPSDs and other related outcomes in people with dementia admitted to an acute care hospital. On the basis of encouraging evidence from recent studies [57], we hypothesized that, by administering VR therapy with natural virtual environments to people with dementia admitted to an acute care hospital, we could help manage BPSDs and improve other process measures, including falls, QoL, and LoS.

In line with our predictions, we found that VR therapy significantly reduced aggressive or violent behavior and loud vocalizations. This is a particularly promising finding as the effective management of aggression has been identified as one of the highest priorities in dementia care [76]. Agitation and its sequelae are common, persistent, and distressing symptoms among people with moderate and severe dementia. In particular, patients with agitation and aggression have difficulty spending time in the hospital [77-79]. Conversely, we found no significant differences between the intervention and control groups for the other BPSDs measured, including the NPI-10–like cluster and wandering. Similarly, we found no significant differences between groups in terms of falls, LoS, or QoL as measured using the QUALID tool.

One possible explanation as to why we observed a positive impact on aggressive behaviors and not on other outcomes relates to VR's capability to modify known environmental triggers. In this case, patients may have been experiencing pain from acute medical illness; a lack of meaningful stimulation from being in a hospital room; or, similarly, hyperstimulation related to sensory stimuli in the hospital environment. Indeed, VR environments allow individuals to experience a sense of presence and immersion, offering an escape from painful or irritating stimuli [76]. Agitation and aggression are known to respond to environmental interventions that address individuals' underlying needs, for example, by providing social contact, offering stimulating and relaxing activities, reducing stressful stimuli [80-82], or providing a sense of freedom from confinement [83]. In line with this explanation, we might expect that the short-term effects of VR therapy would be more noticeable for symptoms of agitation provoked by environmental triggers. In contrast, agitation provoked by nonenvironmental factors such as delusional, depressive, and trait anxiety symptoms may require other, more targeted, or extended cognitive and behavioral interventions.

Nevertheless, a critical hindrance to our study and the possible reason for not reaching statistically significant outcomes was the need to cease participant recruitment before reaching the target sample size during the COVID-19 pandemic. Thus, it remains unclear whether this result was due to a lack of statistical power, limited effectiveness of the treatment, or other factors. It is certainly possible that the short intervention period during a typical acute care hospital stay (average LoS in our sample=10.4 days) or number of therapy sessions (mean 1.6 sessions) may have proved insufficient to affect these outcomes in a measurable way. Furthermore, as ours was the first RCT looking at VR therapy for individuals with dementia admitted to an acute care hospital, we were unable to directly compare our findings with those of other similar studies. Nevertheless, we did not observe any clear negative impacts of VR therapy, and we were able to describe factors related to our methodology that may have contributed to the insignificant results.

For measuring BPSDs, analyzing nurses' notes was determined by the research team to be more feasible than administering the NPI-10 instrument during the pilot that informed this trial [57]. However, it is possible that our positive result for aggression and not for other BPSDs was because aggressive behaviors are more overt and more readily captured in nurses' notes. Unlike aggressive behaviors, which are externalized, readily visible, and known to respond to environmental interventions within a short period [84], NPI-10–like BPSDs related to trait anxious and depressive symptoms are characterized by internalized features (eg, feelings of dysphoria, euphoria, and delusional beliefs). It is also important to point out that some of these BPSDs, such as anxiety and depression, can be a trigger for aggressive behavior [84,85] and, thus, any improvement may have been captured as a reduction in aggression in our study.

When considering the acceptability, safety, and patient experience with VR therapy, our observational results are overall promising. No serious adverse events occurred that were related to the intervention. Participants reported that audio and visual stimuli were clear in most sessions, and the few reported

XSL•FO
RenderX

challenges were due to apprehension or anxiety or mild side effects or discomfort from the HMD. The headset weight was the most commonly reported source of discomfort, which is important to consider given that Oculus Go remains relatively lighter than many HMDs on the market [86]. To reduce discomfort, especially for the frail older population, we recommend considering the HMD weight and adding supports that make the device's weight distribution more balanced.

Most participants were willing to try VR and were able to tolerate at least a brief session. The average length of exposure was approximately 7 minutes, which is substantially shorter than the median length of treatment protocol of 20 minutes described in other studies using VR to manage BPSDs and promote QoL in people with dementia [65]. This shorter average session time is likely due to participants being acutely ill, although it is unclear whether this constitutes a potential limitation to VR's effectiveness in people with dementia in acute care as the optimal dosage has not yet been established [65]. Although traditional technologies such as television or music may be easier to administer and better entertain and distract people with dementia in acute care for longer periods, short VR exposures may still be beneficial and arguably easier to administer when compared with longer VR sessions. VR's unique immersive capabilities make it a particularly powerful tool for eliciting strong positive emotions, including awe, joy, and excitement [87], and brief (6-min) exposure to 360° VR nature videos has been shown to have comparable benefits with those of exposure to the outdoors [88]. Nevertheless, a substantial minority of participants in our study (8/34, 24%) either refused VR or stopped their sessions after <1 minute. We also observed variability in session length within participants; as one participant described, "Today [is] not a good day. Maybe I'd enjoy tomorrow. Just a bit shaky now." These findings suggest a fluctuating ability to participate and the importance of providing multiple opportunities for patients to try the intervention—and allocating adequate resources to do so, which may pose a challenge in acute care [89,90]. This challenge would not necessarily be unique to VR therapy; although their effectiveness is supported by long-standing clinical experience, nonpharmacological interventions are seldom implemented in practice because of a lack of time and resources [33,91,92] as well as attitudinal barriers [93-96].

In terms of the impact on QoL or well-being, people with dementia in the VR arm consistently provided more positive responses when asked conversational questions about how they were doing, whereas those in the control arm became more negative in their responses after the first study session. Some caregivers also reported a positive impact of VR therapy on their loved ones' QoL and well-being in general. For example, one caregiver noted the "positive effect" of VR on her mother, and another described VR as "Something to look forward to trying...it explains why he goes out to look at the trees at night." The primary nurse of one patient observed part of a session and reported that she was impressed with the participant's positive reaction as he had been frequently irritable. Most participants exposed to VR therapy described positive feelings and satisfaction with their experience, with many demonstrating engagement, reminiscence, and relaxation during the sessions. As engaging in relaxation therapy in a natural context may be less practical for this population, especially when in an acute care hospital, we align with other studies in support of providing virtual natural environments to promote well-being and positive experiences [97,98].

In this study, participants were offered the opportunity to choose from a variety of video options and provide feedback on the types of VR content that they would like to see more of. Preferences were varied, and the most frequent suggestions for VR content were cultural topics, followed by nature, animals, and music. Although much remains unknown about the impact of VR content on BPSDs and QoL in people with dementia [51], our findings align with those of other studies that have highlighted the importance of customization of VR contexts in enhancing engagement and allowing for a state of relaxation [99].

Limitations and Future Directions

This study was subject to several limitations with potential impact on our ability to answer the research questions as well as on the ability to generalize the findings. The first and most critical limitation was the early termination of the study because of the COVID-19 pandemic, with a final recruitment of less than half of the target sample size. Follow-up studies with the sufficient sample size required for statistical power will be necessary to determine whether VR therapy has a significant effect on LoS, falls, QoL, and BPSDs and further establish its effectiveness in reducing aggressive behaviors.

Second, as less than half (12/29, 41%) of the participants completed more than one VR therapy session, it is difficult to draw conclusions regarding treatment effectiveness. Although this limitation is common among similar studies involving VR therapy in people with dementia, it made it more difficult to evaluate the outcomes in general, and it was not possible to measure the impact of treatment dosage or whether there is a wash-out period after repeated VR exposures. As it is not yet understood how the number, frequency, and length of VR therapy sessions correlate with effectiveness [51], it will be important to design future studies that are able to evaluate dose regimens.

Third, we encountered a number of challenges that could be considered inherent to conducting nonpharmacological intervention trials with this population and in this setting. Although the ward and research staff were blinded where possible, neither patients and caregivers nor the RC who administered the VR therapy and collected QoL measures were blinded to treatment allocation, introducing potential for bias. In addition, during recruitment, nearly half (137/300, 45.7%) of the eligible patients screened were unable to participate because they were discharged or deceased by the time their SDM was reached for informed consent. Given this challenge, and considering a typical length of acute care stay of approximately 1 week [100], we recommend using measures that are more sensitive to short-term changes in this setting, particularly for outcomes for which it may take longer to detect any changes as a result of the intervention, such as QoL.

XSL•FO
RenderX

Furthermore, collecting accurate self-report (and proxy-report) data was a challenge, which may have led to insignificant results within our primary outcomes. Through analysis of nurses' notes, we were able to measure changes in aggression but may have missed capturing less overt outcomes with internalized features (ie, other BPSDs and QoL). Unfortunately, existing self-report measures are difficult to use or inadequate for gathering information from people with dementia [90], and this limited our ability to gather information directly from participants. In our sample up to one-fifth of participants (6/47, 13% to 10/47, 21%) were unable to provide any verbal feedback on their well-being or experience across the various outcome measures that relied on self-report (for which we used supportive communication strategies such as short questions and visual aids). As we anticipated this challenge, we incorporated a standardized VR observation tool to triangulate the self-report findings for the secondary outcomes. Nevertheless, we acknowledge that this methodology presents the possibility of researcher bias, which we took measures to lessen by carefully designing and structuring the tool and providing special training to the researchers involved in data collection. As a future research direction, we recommend designing and validating alternative structured methods that allow for better accessibility for people with dementia [91].

Given these challenges, we suggest focusing on future trials in other care settings (eg, at home, day treatment, or long-term care) to measure effectiveness and continue to refine the intervention with individuals who are less acutely ill. In these cases, changes can be measured over longer periods in a setting in which the individual's environment is more constant, with more reliable access to family caregivers to help administer VR and provide proxy reports. We are in the process of applying findings from this trial to studies that explore the impact of technological properties on outcomes for both people with dementia and their family caregivers at home [101,102]. Finally, it is worth emphasizing that geriatricians repeatedly report reluctantly administering drugs as they are left with no effective alternatives when dealing with responsive behaviors. An important direction for future studies to explore is whether managing BPSDs using VR therapy might reduce the need for applying restraints or using sedative (antipsychotic and benzodiazepine) medications.

Conclusions

To the best of our knowledge, this is the first RCT to investigate the impact of VR therapy on BPSDs and well-being in people with dementia in an acute care setting. Although the necessary sample size to observe significance at the desired confidence level was not reached owing to COVID-19 restrictions, we found that immersive VR therapy was effective in reducing aggressive behaviors. In addition, our results suggest nonsignificant but promising trends in terms of VR therapy promoting QoL and add to emerging evidence of VR therapy being a safe and acceptable treatment option for people with dementia. In particular, this is one of the first trials to demonstrate that individuals with moderate to severe dementia and complex medical issues requiring hospitalization can tolerate and enjoy VR for brief periods.

Considering the importance of introducing ethically acceptable and effective techniques to care for people with dementia, rigorous follow-up studies with carefully designed measures are warranted to understand VR therapy's impact on BPSDs and related outcomes. Given the challenges we encountered in acute care, we recommend first establishing effectiveness and further tailoring the intervention to this population in care settings where individuals are less acutely ill and changes may be more easily detected.

Acknowledgments

This study was funded by the Centre for Aging + Brain Health Innovation. The funding bodies had no role in the design of the study; collection, analysis, and interpretation of the data; or writing of the manuscript. The authors would like to thank Zain Pasat, Suad Ali, and Bryce Chen for their support with data collection. They are grateful to Codrin Talaba and Vlad Luchnikov for working with the authors' research team to film, edit, and provide the virtual reality experiences for this study.

Data Availability

The data sets supporting the conclusions of this study are included within the paper and its additional files; for example, primary outcome data and secondary outcome data can be found in Multimedia Appendices 8 and 9, respectively.

Authors' Contributions

LA, EA, EK, CS, and JR contributed to the conception and design of the study. CS and JR organized the on-site data collection. LA, EA, EK, JA, and SL-F contributed to the design of statistical methodology and qualitative coding structure. JA, EK, and SL-F conducted data cleaning and qualitative coding. JA conducted the statistical analysis. LA, SL-F, EK, and SP wrote the first draft of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Assent form.
[PDF File (Adobe PDF File), 319 KB - jmir_v26i1e51758_app1.pdf]

Multimedia Appendix 2
Informed consent form.
[PDF File (Adobe PDF File), 447 KB - jmir_v26i1e51758_app2.pdf]

Multimedia Appendix 3
Informational study brochure.
[PDF File (Adobe PDF File), 877 KB - jmir_v26i1e51758_app3.pdf]

Multimedia Appendix 4
Presession checklist.
[PDF File (Adobe PDF File), 163 KB - jmir_v26i1e51758_app4.pdf]

Multimedia Appendix 5
Modified Quality of Life in Late-Stage Dementia scale.
[PDF File (Adobe PDF File), 200 KB - jmir_v26i1e51758_app5.pdf]

Multimedia Appendix 6
Standardized virtual reality observation tool.
[PDF File (Adobe PDF File), 94 KB - jmir_v26i1e51758_app6.pdf]

Multimedia Appendix 7
Baseline demographics table.
[DOCX File , 21 KB - jmir_v26i1e51758_app7.docx]

Multimedia Appendix 8
Primary outcome data.
[XLSX File (Microsoft Excel File), 558 KB - jmir_v26i1e51758_app8.xlsx]

Multimedia Appendix 9
Secondary outcome data.
[XLSX File (Microsoft Excel File), 68 KB - jmir_v26i1e51758_app9.xlsx]

Multimedia Appendix 10
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [103].
[PDF File (Adobe PDF File), 2665 KB - jmir_v26i1e51758_app10.pdf]

References

1. Pérez Romero A, González Garrido S. The importance of behavioural and pyschological symptoms in Alzheimer's disease. Neurologia (Engl Ed) 2016 Apr 26;33(6):30011-1 [FREE Full text] [doi: 10.1016/j.nrl.2016.02.024] [Medline: 27130309]
2. Savaskan E, Bopp-Kistler I, Buerge M, Fischlin R, Georgescu D, Giardini U, et al. [Recommendations for diagnosis and therapy of behavioral and psychological symptoms in dementia (BPSD)]. Praxis (Bern 1994) 2014 Jan 29;103(3):135-148. [doi: 10.1024/1661-8157/a001547] [Medline: 24468453]
3. Devshi R, Shaw S, Elliott-King J, Hogervorst E, Hiremath A, Velayudhan L, et al. Prevalence of behavioural and psychological symptoms of dementia in individuals with learning disabilities. Diagnostics (Basel) 2015 Dec 02;5(4):564-576 [FREE Full text] [doi: 10.3390/diagnostics5040564] [Medline: 26854171]
4. Tible OP, Riese F, Savaskan E, von Gunten A. Best practice in the management of behavioural and psychological symptoms of dementia. Ther Adv Neurol Disord 2017 Aug;10(8):297-309 [FREE Full text] [doi: 10.1177/1756285617712979] [Medline: 28781611]
5. Sampson EL, White N, Leurent B, Scott S, Lord K, Round J, et al. Behavioural and psychiatric symptoms in people with dementia admitted to the acute hospital: prospective cohort study. Br J Psychiatry 2014 Sep 02;205(3):189-196 [FREE Full text] [doi: 10.1192/bjp.bp.113.130948] [Medline: 25061120]

XSL•FO
RenderX

6. Feast A, Orrell M, Russell I, Charlesworth G, Moniz-Cook E. The contribution of caregiver psychosocial factors to distress associated with behavioural and psychological symptoms in dementia. Int J Geriatr Psychiatry 2017 Jan 18;32(1):76-85. [doi: 10.1002/gps.4447] [Medline: 26891463]
7. Härlein J, Dassen T, Halfens RJ, Heinze C. Fall risk factors in older people with dementia or cognitive impairment: a systematic review. J Adv Nurs 2009 May;65(5):922-933. [doi: 10.1111/j.1365-2648.2008.04950.x] [Medline: 19291191]
8. Feast A, Moniz-Cook E, Stoner C, Charlesworth G, Orrell M. A systematic review of the relationship between behavioral and psychological symptoms (BPSD) and caregiver well-being. Int Psychogeriatr 2016 Nov;28(11):1761-1774. [doi: 10.1017/S1041610216000922] [Medline: 27345942]
9. Bhagat G, Raj R, Sidhu BS, Sidhu AK. A cross-sectional study on behavioural and psychological symptoms of dementia in elderly and its impact on quality of life in a tertiary care hospital. Int J Adv Med 2018 May 22;5(3):614. [doi: 10.18203/2349-3933.ijam20182112]
10. Kang Y, Hur Y. Nurses' experience of nursing workload-related issues during caring patients with dementia: a qualitative meta-synthesis. Int J Environ Res Public Health 2021 Oct 04;18(19):10448 [FREE Full text] [doi: 10.3390/ijerph181910448] [Medline: 34639748]
11. George J, Long S, Vincent C. How can we keep patients with dementia safe in our acute hospitals? A review of challenges and solutions. J R Soc Med 2013 Sep;106(9):355-361 [FREE Full text] [doi: 10.1177/0141076813476497] [Medline: 23759885]
12. Canevelli M, Adali N, Cantet C, Andrieu S, Bruno G, Cesari M, et al. Impact of behavioral subsyndromes on cognitive decline in Alzheimer's disease: data from the ICTUS study. J Neurol 2013 Jul 17;260(7):1859-1865. [doi: 10.1007/s00415-013-6893-3] [Medline: 23504051]
13. Nourhashémi F, Andrieu S, Sastres N, Ducassé JL, Lauque D, Sinclair AJ, et al. Descriptive analysis of emergency hospital admissions of patients with Alzheimer disease. Alzheimer Dis Assoc Disord 2001;15(1):21-25. [doi: 10.1097/00002093-200101000-00003] [Medline: 11236821]
14. Toot S, Swinson T, Devine M, Challis D, Orrell M. Causes of nursing home placement for older people with dementia: a systematic review and meta-analysis. Int Psychogeriatr 2017 Feb;29(2):195-208. [doi: 10.1017/S1041610216001654] [Medline: 27806743]
15. Yaffe K, Fox P, Newcomer R, Sands L, Lindquist K, Dane K, et al. Patient and caregiver characteristics and nursing home placement in patients with dementia. JAMA 2002 Apr 24;287(16):2090-2097. [doi: 10.1001/jama.287.16.2090] [Medline: 11966383]
16. Currie L. Fall and injury prevention. In: Hughes RG, editor. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Rockville, MD: Agency for Healthcare Research and Quality (US); 2008.
17. Falls. World Health Organization. 2021 Apr 26. URL: https://www.who.int/en/news-room/fact-sheets/detail/falls [accessed 2022-06-15]
18. Centers for Disease Control and Prevention (CDC). Fatalities and injuries from falls among older adults--United States, 1993-2003 and 2001-2005. MMWR Morb Mortal Wkly Rep 2006 Nov 17;55(45):1221-1224 [FREE Full text] [Medline: 17108890]
19. Assessment and management of behaviours and psychological symptoms associated with dementia (BPSD). The Royal Australian and New Zealand College of Psychiatrists. URL: https://www.ranzcp.org/files/resources/reports/a-handbook-for-nsw-health-clinicians-bpsd_june13_w.aspx [accessed 2022-07-27]
20. Cerejeira J, Lagarto L, Mukaetova-Ladinska EB. Behavioral and psychological symptoms of dementia. Front Neurol 2012 May 7;3:73 [FREE Full text] [doi: 10.3389/fneur.2012.00073] [Medline: 22586419]
21. Zhu CW, Scarmeas N, Ornstein K, Albert M, Brandt J, Blacker D, et al. Health-care use and cost in dementia caregivers: longitudinal results from the Predictors Caregiver Study. Alzheimers Dement 2015 Apr;11(4):444-454 [FREE Full text] [doi: 10.1016/j.jalz.2013.12.018] [Medline: 24637299]
22. Ballard CG, Gauthier S, Cummings JL, Brodaty H, Grossberg GT, Robert P, et al. Management of agitation and aggression associated with Alzheimer disease. Nat Rev Neurol 2009 May;5(5):245-255. [doi: 10.1038/nrneurol.2009.39] [Medline: 19488082]
23. Ohno Y, Kunisawa N, Shimizu S. Antipsychotic treatment of behavioral and psychological symptoms of dementia (BPSD): management of extrapyramidal side effects. Front Pharmacol 2019 Sep 17;10:1045 [FREE Full text] [doi: 10.3389/fphar.2019.01045] [Medline: 31607910]
24. White PC, Wyatt J, Chalfont G, Bland JM, Neale C, Trepel D, et al. Exposure to nature gardens has time-dependent associations with mood improvements for people with mid- and late-stage dementia: innovative practice. Dementia (London) 2018 Jul;17(5):627-634 [FREE Full text] [doi: 10.1177/1471301217723772] [Medline: 28835119]
25. Braun A, Trivedi DP, Dickinson A, Hamilton L, Goodman C, Gage H, et al. Managing behavioural and psychological symptoms in community dwelling older people with dementia: 2. A systematic review of qualitative studies. Dementia (London) 2019;18(7-8):2950-2970 [FREE Full text] [doi: 10.1177/1471301218762856] [Medline: 29557193]
26. Chenoweth L, King MT, Jeon YH, Brodaty H, Stein-Parbury J, Norman R, et al. Caring for aged dementia care resident study (CADRES) of person-centred care, dementia-care mapping, and usual care in dementia: a cluster-randomised trial. Lancet Neurol 2009 Apr;8(4):317-325. [doi: 10.1016/S1474-4422(09)70045-6] [Medline: 19282246]

XSL•FO
RenderX

27. Sloane PD, Hoeffer B, Mitchell CM, McKenzie DA, Barrick AL, Rader J, et al. Effect of person-centered showering and the towel bath on bathing-associated aggression, agitation, and discomfort in nursing home residents with dementia: a randomized, controlled trial. J Am Geriatr Soc 2004 Nov;52(11):1795-1804 [FREE Full text] [doi: 10.1111/j.1532-5415.2004.52501.x] [Medline: 15507054]
28. Fossey J, Ballard C, Juszczak E, James I, Alder N, Jacoby R, et al. Effect of enhanced psychosocial care on antipsychotic use in nursing home residents with severe dementia: cluster randomised trial. BMJ 2006 Apr 01;332(7544):756-761 [FREE Full text] [doi: 10.1136/bmj.38782.575868.7C] [Medline: 16543297]
29. Savaskan E, Bopp-Kistler I, Buerge M, Fischlin R, Georgescu D, Giardini U, et al. [Recommendations for diagnosis and therapy of behavioral and psychological symptoms in dementia (BPSD)]. Praxis (Bern 1994) 2014 Jan 29;103(3):135-148 [FREE Full text] [doi: 10.1024/1661-8157/a001547] [Medline: 24468453]
30. Akrour R, Courret-Gilgen C, Perrenoud B. Prevention and management of behavioural and psychological symptoms in patients with dementia in acute care: a best practice implementation project. JBI Evid Implement 2022 Dec 01;20(4):289-300 [FREE Full text] [doi: 10.1097/XEB.0000000000000329] [Medline: 36375022]
31. Scerri A, Scerri C, Innes A. The perceived and observed needs of patients with dementia admitted to acute medical wards. Dementia (London) 2020 Aug 27;19(6):1997-2017. [doi: 10.1177/1471301218814383] [Medline: 30482091]
32. Prato L, Lindley L, Boyles M, Robinson L, Abley C. Empowerment, environment and person-centred care: a qualitative study exploring the hospital experience for adults with cognitive impairment. Dementia (London) 2019 Feb 07;18(7-8):2710-2730. [doi: 10.1177/1471301218755878] [Medline: 29411662]
33. Dementia guide. National Health Service. URL: https://www.nhs.uk/conditions/dementia/ [accessed 2022-07-27]
34. Lin FR, Metter EJ, O'Brien RJ, Resnick SM, Zonderman AB, Ferrucci L. Hearing loss and incident dementia. Arch Neurol 2011 Feb 14;68(2):214-220 [FREE Full text] [doi: 10.1001/archneurol.2010.362] [Medline: 21320988]
35. Cohen-Mansfield J, Marx MS, Dakheel-Ali M, Regier NG, Thein K, Freedman L. Can agitated behavior of nursing home residents with dementia be prevented with the use of standardized stimuli? J Am Geriatr Soc 2010 Aug;58(8):1459-1464 [FREE Full text] [doi: 10.1111/j.1532-5415.2010.02951.x] [Medline: 20579167]
36. Cohen-Mansfield J, Dakheel-Ali M, Marx MS, Thein K, Regier NG. Which unmet needs contribute to behavior problems in persons with advanced dementia? Psychiatry Res 2015 Jul 30;228(1):59-64 [FREE Full text] [doi: 10.1016/j.psychres.2015.03.043] [Medline: 25933478]
37. de Oliveira AM, Radanovic M, de Mello PC, Buchain PC, Vizzotto AD, Celestino DL, et al. Nonpharmacological interventions to reduce behavioral and psychological symptoms of dementia: a systematic review. Biomed Res Int 2015;2015:218980 [FREE Full text] [doi: 10.1155/2015/218980] [Medline: 26693477]
38. Ulrich RS. View through a window may influence recovery from surgery. Science 1984 Apr 27;224(4647):420-421. [doi: 10.1126/science.6143402] [Medline: 6143402]
39. Detweiler MB, Warf C. Dementia wander garden aids post cerebrovascular stroke restorative therapy: a case study. Altern Ther Health Med 2005;11(4):54-58. [Medline: 16053122]
40. Li HC, Wang HH, Lu CY, Chen TB, Lin YH, Lee I. The effect of music therapy on reducing depression in people with dementia: a systematic review and meta-analysis. Geriatr Nurs 2019;40(5):510-516 [FREE Full text] [doi: 10.1016/j.gerinurse.2019.03.017] [Medline: 31056209]
41. Mooney P, Nicell PL. The importance of exterior environment for Alzheimer residents: effective care and risk management. Healthc Manage Forum 1992;5(2):23-29. [doi: 10.1016/S0840-4704(10)61202-1] [Medline: 10171068]
42. Park SH, Mattson RH. Therapeutic influences of plants in hospital rooms on surgical recovery. HortScience 2009 Feb;44(1):102-105 [FREE Full text] [doi: 10.21273/hortsci.44.1.102]
43. Hersh M, Nazarian MA. Palliative care can help reduce incidence of medicare's never-events for dementia patients. J Hosp Palliat Nurs 2010 Jul;12(4):247-254. [doi: 10.1097/njh.0b013e3181df5fc5]
44. Zonsius MC, Cothran FA, Miller JM. CE: acute care for patients with dementia. Am J Nurs 2020 Apr;120(4):34-42. [doi: 10.1097/01.NAJ.0000660024.45260.1a] [Medline: 32218045]
45. Lee DC, Meyer C, Burton E, Kitchen S, Williams C, Hunter SW, et al. A survey of nurses, physiotherapists and occupational therapists in mobility care and gait aid use for hospital patients with dementia. Geriatr Nurs 2022 Mar;44:221-228. [doi: 10.1016/j.gerinurse.2022.02.017] [Medline: 35240401]
46. D'Cunha NM, Nguyen D, Naumovski N, McKune AJ, Kellett J, Georgousopoulou EN, et al. A mini-review of virtual reality-based interventions to promote well-being for people living with dementia and mild cognitive impairment. Gerontology 2019;65(4):430-440 [FREE Full text] [doi: 10.1159/000500040] [Medline: 31108489]
47. Dermody G, Whitehead L, Wilson G, Glass C. The role of virtual reality in improving health outcomes for community-dwelling older adults: systematic review. J Med Internet Res 2020 Jun 01;22(6):e17331 [FREE Full text] [doi: 10.2196/17331] [Medline: 32478662]
48. Hughes S, Warren-Norton K, Spadafora P, Tsotsos LE. Supporting optimal aging through the innovative use of virtual reality technology. Multimodal Technol Interact 2017 Sep 28;1(4):23. [doi: 10.3390/mti1040023]
49. Montana JI, Matamala-Gomez M, Maisto M, Mavrodiev PA, Cavalera CM, Diana B, et al. The benefits of emotion regulation interventions in virtual reality for the improvement of wellbeing in adults and older adults: a systematic review. J Clin Med 2020 Feb 12;9(2):500 [FREE Full text] [doi: 10.3390/jcm9020500] [Medline: 32059514]

XSL•FO
RenderX

50. Strong J. Immersive virtual reality and persons with dementia: a literature review. J Gerontol Soc Work 2020 Apr 24;63(3):209-226. [doi: 10.1080/01634372.2020.1733726] [Medline: 32091323]
51. Appel L, Ali S, Narag T, Mozeson K, Pasat Z, Orchanian-Cheff A, et al. Virtual reality to promote wellbeing in persons with dementia: a scoping review. J Rehabil Assist Technol Eng 2021 Dec 21;8:20556683211053952 [FREE Full text] [doi: 10.1177/20556683211053952] [Medline: 35024166]
52. Appel L, Kisonas E, Appel E, Klein J, Bartlett D, Rosenberg J, et al. Administering virtual reality therapy to manage behavioral and psychological symptoms in patients with dementia admitted to an acute care hospital: results of a pilot study. JMIR Form Res 2021 Feb 03;5(2):e22406 [FREE Full text] [doi: 10.2196/22406] [Medline: 33533720]
53. Appel L, Kisonas E, Appel E, Klein J, Bartlett D, Rosenberg J, et al. Introducing virtual reality therapy for inpatients with dementia admitted to an acute care hospital: learnings from a pilot to pave the way to a randomized controlled trial. Pilot Feasibility Stud 2020 Oct 31;6(1):166 [FREE Full text] [doi: 10.1186/s40814-020-00708-9] [Medline: 33292729]
54. Parsons TD, McPherson S, Interrante V. Enhancing neurocognitive assessment using immersive virtual reality. In: Proceedings of the 1st Workshop on Virtual and Augmented Assistive Technology (VAAT). 2013 Presented at: 1st Workshop on Virtual and Augmented Assistive Technology (VAAT); March 17, 2013; Lake Buena Vista, FL. [doi: 10.1109/vaat.2013.6786190]
55. Cutler C, Hicks B, Innes A. Does digital gaming enable healthy aging for community-dwelling people with dementia? Games and Culture 2015 Aug 30;11(1-2):104-129 [FREE Full text] [doi: 10.1177/1555412015600580]
56. Benoit M, Guerchouche R, Petit PD, Chapoulie E, Manera V, Chaurasia G, et al. Is it possible to use highly realistic virtual reality in the elderly? A feasibility study with image-based rendering. Neuropsychiatr Dis Treat 2015 Mar 3;11:557-563 [FREE Full text] [doi: 10.2147/NDT.S73179] [Medline: 25834437]
57. Appel L, Appel E, Bogler O, Wiseman M, Cohen L, Ein N, et al. Older adults with cognitive and/or physical impairments can benefit from immersive virtual reality experiences: a feasibility study. Front Med (Lausanne) 2019 Jan 15;6:329 [FREE Full text] [doi: 10.3389/fmed.2019.00329] [Medline: 32010701]
58. Moyle W, Jones C, Dwan T, Petrovich T. Effectiveness of a virtual reality forest on people with dementia: a mixed methods pilot study. Gerontologist 2018 May 08;58(3):478-487. [doi: 10.1093/geront/gnw270] [Medline: 28329867]
59. Smith V, Warty RR, Sursas JA, Payne O, Nair A, Krishnan S, et al. The effectiveness of virtual reality in managing acute pain and anxiety for medical inpatients: systematic review. J Med Internet Res 2020 Nov 02;22(11):e17980 [FREE Full text] [doi: 10.2196/17980] [Medline: 33136055]
60. Carr C, Odell-Miller H, Priebe S. A systematic review of music therapy practice and outcomes with acute adult psychiatric in-patients. PLoS One 2013 Aug 2;8(8):e70252 [FREE Full text] [doi: 10.1371/journal.pone.0070252] [Medline: 23936399]
61. Health Care Consent Act, 1996 S.O. 1996, Chapter 2 Schedule A. Ontario Government. URL: https://www.ontario.ca/laws/statute/96h02 [accessed 2020-08-08]
62. My Substitute DecisionMaker (SDM). Michael Garron Hospital. 2016 Feb. URL: https://www.tehn.ca/sites/default/files/file-browser/sdm_brochure_mgh_final_feb_2016.pdf [accessed 2022-07-27]
63. Learn about randomization. National Institutes of Health National Cancer Institute. URL: https://ctrandomization.cancer.gov/about/ [accessed 2023-08-08]
64. Health and safety warnings. Meta. URL: https://www.oculus.com/legal/health-and-safety-warnings/ [accessed 2020-02-27]
65. Appel L, Kisonas E, Appel E. The development of ObsRVR: an observational instrument to measure reactions of people with dementia experiencing virtual reality. In: Proceedings of the 14th Conference Clinical Trials on Alzheimer's Disease. 2021 Presented at: 14th Conference Clinical Trials on Alzheimer's Disease; November 9-12, 2021; Boston, MA URL: https://www.ctad-alzheimer.com/files/files/CTAD21%20Posters.pdf
66. Minitab homepage. Minitab. URL: https://www.minitab.com [accessed 2023-12-30]
67. McCoy CE. Understanding the intention-to-treat principle in randomized controlled trials. West J Emerg Med 2017 Oct 18;18(6):1075-1078 [FREE Full text] [doi: 10.5811/westjem.2017.8.35985] [Medline: 29085540]
68. Shah PB. Intention-to-treat and per-protocol analysis. CMAJ 2011 Apr 05;183(6):696; author reply 696 [FREE Full text] [doi: 10.1503/cmaj.111-2033] [Medline: 21464181]
69. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: intention-to-treat versus per-protocol analysis. Perspect Clin Res 2016;7(3):144-146 [FREE Full text] [doi: 10.4103/2229-3485.184823] [Medline: 27453832]
70. Cummings JL. The neuropsychiatric inventory: assessing psychopathology in dementia patients. Neurology 1997 May;48(5 Suppl 6):S10-S16 [FREE Full text] [doi: 10.1212/wnl.48.5_suppl_6.10s] [Medline: 9153155]
71. Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia. Neurology 1994 Dec;44(12):2308-2314 [FREE Full text] [doi: 10.1212/wnl.44.12.2308] [Medline: 7991117]
72. Weiner MF, Martin-Cook K, Svetlik DA, Saine K, Foster B, Fontaine CS. The quality of life in late-stage dementia (QUALID) scale. J Am Med Dir Assoc 2000;1(3):114-116. [Medline: 12818023]
73. Jackson K, Bazeley P. Qualitative Data Analysis with NVivo. 3rd Edition. Thousand Oaks, CA: SAGE Publications; May 2019.
74. Hallberg LR. The "core category" of grounded theory: making constant comparisons. Int J Qual Stud Health Well-being 2009 Jul 12;1(3):141-148 [FREE Full text] [doi: 10.1080/17482620600858399]

XSL•FO
RenderX

75. Fram SM. The constant comparative analysis method outside of grounded theory. Qual Report 2015 Jan 14;18(1):1-25 [FREE Full text] [doi: 10.46743/2160-3715/2013.1569]
76. Baird CL, Murawski MM, Wu J. Efficacy of guided imagery with relaxation for osteoarthritis symptoms and medication intake. Pain Manag Nurs 2010 Mar;11(1):56-65. [doi: 10.1016/j.pmn.2009.04.002] [Medline: 20207328]
77. Carrarini C, Russo M, Dono F, Barbone F, Rispoli MG, Ferri L, et al. Agitation and dementia: prevention and treatment strategies in acute and chronic conditions. Front Neurol 2021 Apr 16;12:644317 [FREE Full text] [doi: 10.3389/fneur.2021.644317] [Medline: 33935943]
78. Ryu SH, Katona C, Rive B, Livingston G. Persistence of and changes in neuropsychiatric symptoms in Alzheimer disease over 6 months. Am J Geriatr Psychiatry 2005 Nov;13(11):976-983. [doi: 10.1097/00019442-200511000-00008]
79. Livingston G, Johnston K, Katona C, Paton J, Lyketsos CG, Old Age Task Force of the World Federation of Biological Psychiatry. Systematic review of psychological approaches to the management of neuropsychiatric symptoms of dementia. Am J Psychiatry 2005 Nov;162(11):1996-2021 [FREE Full text] [doi: 10.1176/appi.ajp.162.11.1996] [Medline: 16263837]
80. Camp CJ, Cohen-Mansfield J, Capezuti EA. Use of nonpharmacologic interventions among nursing home residents with dementia. Psychiatr Serv 2002 Nov;53(11):1397-1401 [FREE Full text] [doi: 10.1176/appi.ps.53.11.1397] [Medline: 12407266]
81. Cohen-Mansfield J. Nonpharmacologic interventions for inappropriate behaviors in dementia: a review, summary, and critique. Am J Geriatr Psychiatry 2001;9(4):361-381. [Medline: 11739063]
82. Namazi K, Johnson B. Pertinent autonomy for residents with dementias: modification of the physical environment to enhance independence. Am J Alzheimer's Care Relate Disord Res 2016 Sep 04;7(1):16-21 [FREE Full text] [doi: 10.1177/153331759200700105]
83. McMinn BG, Hinton L. Confined to barracks: the effects of indoor confinement on aggressive behavior among inpatients of an acute psychogeriatric unit. Am J Alzheimers Dis 2016 Sep 04;15(1):36-41 [FREE Full text] [doi: 10.1177/153331750001500106]
84. Kunik ME, Snow AL, Davila JA, Steele AB, Balasubramanyam V, Doody RS, et al. Causes of aggressive behavior in patients with dementia. J Clin Psychiatry 2010 Mar 09;71(09):1145-1152. [doi: 10.4088/jcp.08m04703oli]
85. Cipriani G, Vedovello M, Nuti A, Di Fiorino M. Aggressive behavior in patients with dementia: correlates and management. Geriatr Gerontol Int 2011 Oct;11(4):408-413 [FREE Full text] [doi: 10.1111/j.1447-0594.2011.00730.x] [Medline: 21843201]
86. Alsop T. Comparison of virtual reality (VR) headsets worldwide in 2023, by weight. Statista. 2023. URL: https://www.statista.com/statistics/1337114/vr-headset-comparison-by-weight/ [accessed 2023-08-09]
87. Chirico A, Yaden DB, Riva G, Gaggioli A. The potential of virtual reality for the investigation of awe. Front Psychol 2016 Nov 09;7:1766 [FREE Full text] [doi: 10.3389/fpsyg.2016.01766] [Medline: 27881970]
88. Browning M, Mimnaugh KJ, van Riper CJ, Laurent HK, LaValle SM. Can simulated nature support mental health? Comparing short, single-doses of 360-degree nature videos in virtual reality with the outdoors. Front Psychol 2019;10:2667 [FREE Full text] [doi: 10.3389/fpsyg.2019.02667] [Medline: 32010003]
89. Zucchella C, Sinforiani E, Tamburin S, Federico A, Mantovani E, Bernini S, et al. The multidisciplinary approach to Alzheimer's disease and dementia. A narrative review of non-pharmacological treatment. Front Neurol 2018 Dec 13;9:1058 [FREE Full text] [doi: 10.3389/fneur.2018.01058] [Medline: 30619031]
90. Cohen-Mansfield J, Thein K, Marx MS, Dakheel-Ali M. What are the barriers to performing nonpharmacological interventions for behavioral symptoms in the nursing home? J Am Med Dir Assoc 2012 May;13(4):400-405 [FREE Full text] [doi: 10.1016/j.jamda.2011.07.006] [Medline: 21872537]
91. Røsvik J, Rokstad AM. What are the needs of people with dementia in acute hospital settings, and what interventions are made to meet these needs? A systematic integrative review of the literature. BMC Health Serv Res 2020 Aug 07;20(1):723 [FREE Full text] [doi: 10.1186/s12913-020-05618-3] [Medline: 32767987]
92. Yous ML, Ploeg J, Kaasalainen S, Martin LS. Nurses' experiences in caring for older adults with responsive behaviors of dementia in acute care. SAGE Open Nurs 2019 Apr 24;5:2377960819834127 [FREE Full text] [doi: 10.1177/2377960819834127] [Medline: 33415227]
93. Górska S, Forsyth K, Irvine L, Maciver D, Prior S, Whitehead J, et al. Service-related needs of older people with dementia: perspectives of service users and their unpaid carers. Int Psychogeriatr 2013 Mar 27;25(7):1107-1114 [FREE Full text] [doi: 10.1017/s1041610213000343]
94. Oyebode JR, Bradley P, Allen JL. Relatives' experiences of frontal-variant frontotemporal dementia. Qual Health Res 2013 Feb;23(2):156-166 [FREE Full text] [doi: 10.1177/1049732312466294] [Medline: 23150527]
95. Phillipson L, Jones SC. Use of day centers for respite by help-seeking caregivers of individuals with dementia. J Gerontol Nurs 2012 Apr;38(4):24-34 [FREE Full text] [doi: 10.3928/00989134-20120301-50]
96. Stokes LA, Combes H, Stokes G. Understanding the dementia diagnosis: the impact on the caregiving experience. Dementia (London) 2014 Jan;13(1):59-78 [FREE Full text] [doi: 10.1177/1471301212447157] [Medline: 24381039]
97. Torrisi M, De Cola MC, Marra A, De Luca R, Bramanti P, Calabrò RS. Neuropsychiatric symptoms in dementia may predict caregiver burden: a Sicilian exploratory study. Psychogeriatrics 2017 Mar;17(2):103-107 [FREE Full text] [doi: 10.1111/psyg.12197] [Medline: 27411501]

XSL•FO
RenderX

98. Counting the cost: caring for people with dementia on hospital wards. Alzheimer's Society. 2009. URL: https://www.alzheimers.org.uk/sites/default/files/2018-05/Counting_the_cost_report.pdf [accessed 2023-12-11]
99. Pardini S, Gabrielli S, Dianti M, Novara C, Zucco GM, Mich O, et al. The role of personalization in the user experience, preferences and engagement with virtual reality environments for relaxation. Int J Environ Res Public Health 2022 Jun 13;19(12):7237 [FREE Full text] [doi: 10.3390/ijerph19127237] [Medline: 35742483]
100. Hospital stays in Canada. Canadian Institute for Health Information. 2023 Feb 23. URL: https://www.cihi.ca/en/hospital-stays-in-canada [accessed 2023-07-30]
101. VRx@Home: study to evaluate VR-therapy for PwD living at home. Veeva Systems. URL: https://ctv.veeva.com/study/vrx-home-study-to-evaluate-vr-therapy-for-pwd-living-at-home [accessed 2022-07-22]
102. Veeva Systems. URL: https://ctv.veeva.com/study/vr-r-providing-caregiver-respite-by-managing-bpsds-and-improving-qol-in-people-with-dementia-using [accessed 2023-08-07]
103. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]

Abbreviations

BPSD: behavioral and psychological symptom of dementia
CONSORT: Consolidated Standards of Reporting Trials
EMR: electronic medical record
HMD: head-mounted display
ITT: intention to treat
LoS: length of stay
MGH: Michael Garron Hospital
NPI-10: Neuropsychiatric Inventory
QoL: quality of life
QUALID: Quality of Life in Late-Stage Dementia
RC: research coordinator
RCT: randomized controlled trial
SDM: substitute decision-maker
VR: virtual reality

Edited by T de Azevedo Cardoso; submitted 11.08.23; peer-reviewed by M Reber, S Isbel; comments to author 25.10.23; revised version received 07.11.23; accepted 27.11.23; published 30.01.24.

Please cite as:
Appel L, Appel E, Kisonas E, Lewis-Fung S, Pardini S, Rosenberg J, Appel J, Smith C
Evaluating the Impact of Virtual Reality on the Behavioral and Psychological Symptoms of Dementia and Quality of Life of Inpatients With Dementia in Acute Care: Randomized Controlled Trial (VRCT)
J Med Internet Res 2024;26:e51758
URL: https://www.jmir.org/2024/1/e51758
doi:10.2196/51758
PMID:38289666

©Lora Appel, Eva Appel, Erika Kisonas, Samantha Lewis-Fung, Susanna Pardini, Jarred Rosenberg, Julian Appel, Christopher Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Clinical Efficacy of a Virtual Reality Tool for the Treatment of Obesity: Randomized Controlled Trial

Dimitra Anastasiadou[1,2], PhD; Pol Herrero[2], MSc; Paula Garcia-Royo[2], MSc; Julia Vázquez-De Sebastián[2,3], MSc; Mel Slater[4,5,6], DSC; Bernhard Spanlang[4], PhD; Elena Álvarez de la Campa[4], PhD; Andreea Ciudin[7,8,9], MD, PhD; Marta Comas[7,8,9], PhD; Josep Antoni Ramos-Quiroga[2,10,11,12], MD, PhD; Pilar Lusilla-Palacios[2,10,11,12], MD, PhD

[1]Department of Clinical and Health Psychology, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
[2]Psychiatry, Mental Health and Addictions Research Group, Vall d´Hebron Research Institute, Barcelona, Spain
[3]RE-FiT Barcelona Research Group, Vall d'Hebron Research Institute & Parc Sanitari Pere Virgili, Barcelona, Spain
[4]Virtual Bodyworks S.L., Barcelona, Spain
[5]The Institute of Neurosciences, Universitat de Barcelona, Barcelona, Spain
[6]Department of Clinical Psychology and Psychobiology, Universitat de Barcelona, Barcelona, Spain
[7]Endocrinology and Nutrition Department, Vall d'Hebron University Hospital, Barcelona, Spain
[8]Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona, Barcelona, Spain
[9]Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas, Instituto de Salud Carlos III, Madrid, Spain
[10]Psychiatry Department, Vall d'Hebron University Hospital, Barcelona, Spain
[11]Department of Psychiatry and Legal Medicine, Universitat Autònoma de Barcelona, Barcelona, Spain
[12]Biomedical Network Research Centre on Mental Health (CIBERSAM), Barcelona, Spain

Corresponding Author:
Dimitra Anastasiadou, PhD
Department of Clinical and Health Psychology, Universitat Autònoma de Barcelona
Cerdanyola del Vallès, Barcelona
Bellaterra, Barcelona, 08193
Spain
Phone: 34 935813854
Email: dimitra.anastasiadou@uab.cat

Abstract

Background: Virtual reality (VR) interventions, based on cognitive behavioral therapy principles, have been proven effective as complementary tools in managing obesity and have been associated with promoting healthy behaviors and addressing body image concerns. However, they have not fully addressed certain underlying causes of obesity, such as a lack of motivation to change, low self-efficacy, and the impact of weight stigma interiorization, which often impede treatment adherence and long-term lifestyle habit changes. To tackle these concerns, this study introduces the VR self-counseling paradigm, which incorporates embodiment and body-swapping techniques, along with motivational strategies, to help people living with obesity effectively address some of the root causes of their condition.

Objective: This study aims to assess the clinical efficacy of ConVRself (Virtual Reality self-talk), a VR platform that allows participants to engage in motivational self-conversations.

Methods: A randomized controlled trial was conducted with 68 participants from the bariatric surgery waiting list from the obesity unit of the Vall d'Hebron University Hospital in Barcelona, Spain. Participants were assigned to 1 of 3 groups: a control group (CG), which only received treatment as usual from the obesity unit; experimental group 1 (EG1), which, after intensive motivational interviewing training, engaged in 4 sessions of VR-based self-conversations with ConVRself, and underwent embodiment and body-swapping techniques; and experimental group 2 (EG2), which engaged in 4 VR-based sessions led by a virtual counselor with a prerecorded discourse, and only underwent the embodiment technique. In the case of both EG1 and EG2, the VR interventions were assisted by a clinical researcher. Readiness to change habits, eating habits, and psychological variables, as well as adherence and satisfaction with ConVRself were measured at baseline, after the intervention, 1 week after the intervention, and 4 weeks after the intervention.

Results: Regarding the primary outcomes, EG1 (24/68, 35%) and EG2 (22/68, 32%) showed significant improvements in *confidence to lose weight* compared to the CG (22/68, 32%) at all assessment points ($\beta=-.16$; $P=.02$). Similarly, EG1 demonstrated

XSL•FO
RenderX

a significant increase after the intervention in *readiness to exercise more* compared to the CG (β=−.17; *P*=.03). Regarding the secondary outcomes, EG1 participants showed a significant reduction in *uncontrolled eating* (β=.71; *P*=.01) and *emotional eating* (β=.29; *P*=.03) compared to the CG participants, as well as in their *anxiety* levels compared to EG2 and CG participants (β=.65; *P*=.01). In addition, participants from the experimental groups reported high adherence and satisfaction with the VR platform (EG1: mean 59.82, SD 4.00; EG2: mean 58.43, SD 5.22; *d*=0.30, 95% CI −0.30 to 0.89).

Conclusions: This study revealed that using VR self-conversations, based on motivational interviewing principles, may have benefits in helping people with obesity to enhance their readiness to change habits and self-efficacy, as well as reduce dysfunctional eating behaviors and anxiety.

Trial Registration: ClinicalTrials.gov NCT05094557; https://www.clinicaltrials.gov/study/NCT05094557

(J Med Internet Res 2024;26:e51558) doi:10.2196/51558

KEYWORDS

obesity; virtual reality; psychological treatment; embodiment; motivational interviewing; self-conversation

Introduction

Background

Obesity is defined as a severe, chronic, complex, and multifactorial disease with detrimental effects on the individual's physical and psychological health [1]. Various treatment options are currently available for obesity, including psychological interventions, behavioral interventions for lifestyle modification, pharmacotherapy, and bariatric surgery (BS) [2,3]. All these treatments are taken into consideration in a process of shared decision-making to generate a patient-centered plan [4].

According to the National Institute for Health and Care Excellence [5], BS is a treatment option available for people living with obesity with BMI >40 kg/m^2 (or BMI between 35 kg/m^2 and 40 kg/m^2 if relevant comorbidity is present that could be improved with weight loss). People living with obesity who undergo BS require rigorous and comprehensive preoperative and postoperative monitoring and support. This support should emphasize the adoption and maintenance of a healthy lifestyle, as well as the identification of any clinical or psychological barriers that may hinder adherence to postoperative treatment [5]. However, when assessing the long-term maintenance of weight loss in obesity management, including through BS procedures, achieving lasting and sustainable changes in body composition remains highly challenging [6]. Therefore, for more effective obesity management, it is beneficial to incorporate multicomponent psychological interventions that aim to improve health as a whole, foster self-efficacy and self-esteem, and prioritize sustainable goals chosen by the individual [2].

In this context, motivational interviewing (MI) has gained considerable recognition as an effective approach to enhance treatment adherence [7] and has been included in the most recent psychological and behavioral recommendations for obesity management [2]. Specifically, MI is a counseling technique based on a person-centered approach that aims to help individuals to identify the discrepancies between their goals and current circumstances and empower them to explore new alternatives toward behavior change, thereby enhancing motivation and resolving ambivalence [8].

Furthermore, in recent years, telemental health care has emerged as a rapid and efficient means of establishing different communication channels between patients and mental health professionals and has been proven capable of transforming the availability, accessibility, and efficacy of psychological treatments [9,10], particularly during the COVID-19 pandemic and beyond [11]. Recent research findings provide support for the acceptability, feasibility, and preliminary effectiveness of using new technologies in the treatment of eating and weight disorders [12-14]. Specifically, the integration of virtual reality (VR) into psychological interventions for these conditions holds great promise in addressing some of the factors associated with the development or maintenance of the disorders. Through the provision of multisensory experiences, extrinsic feedback, and opportunities for embodiment, psychological VR interventions, which are rooted primarily in cognitive behavioral therapy principles, have demonstrated effectiveness as complementary tools in managing obesity and have been associated with promoting behavior changes and weight reduction, reducing binge eating episodes, and addressing body image concerns [12,14-18]. While these interventions focus on correcting specific behaviors and provide patients with a safe context to practice eating, emotional, and relational management, they do not adequately address some of the root causes of obesity, such as a lack of motivation to change, low self-efficacy, and the impact of weight stigma interiorization, which are factors that often hinder treatment adherence and long-term maintenance of lifestyle habits [19,20].

The SOCRATES project (Self Conversation in Virtual Reality Embodiment to Enhance Healthier Lifestyles Among People with Obesity), funded by the European Commission (951930), has introduced the VR self-counseling paradigm, referred to as ConVRself (Virtual Reality self-talk), which uses embodiment and body-swapping techniques. On the one hand, embodiment enables participants to experience the perceptual illusion that the virtual body is their own [21,22]. On the other hand, body swapping allows embodying alternatingly between 2 virtual bodies (or *avatars*, eg, one representing the self and the other a counselor) and maintaining a conversation between these 2 embodied perspectives [23-25]. In this study, ConVRself is used as a solution to help people living with obesity explore some of the root causes of their condition. Drawing on MI principles, this approach allows participants to engage in self-counseling through motivational conversations with themselves. In particular, participants are immersed in a VR

XSL•FO
RenderX

environment that resembles a counselor's office. First, they embody an avatar that looks like themselves (a *look-alike*), and from this perspective, they explain their problem, goals, and/or aspirations to a counselor, seated in front of them across a table. Once they have stated their problem, participants switch to embody the counselor's avatar. In this new perspective, they listen to a playback of their own words and watch the related body gestures made during their speech. After having listened, they can respond from the counselor's embodied perspective. In this way, participants engage in a self-conversation by adopting 2 different embodied perspectives and maintaining a conversation between them. To ensure that these conversations remain motivational, participants undergo intensive training in MI before engaging in the VR self-counseling experience. The objective of these virtual self-conversations has been to address the following challenges that people living with obesity face: (1) to raise awareness of their actual condition, (2) to better understand and address the impact of weight bias interiorization, and (3) to increase their self-efficacy by setting realistic goals in line with their values.

Objectives

In this randomized controlled trial (RCT) based on the protocol described in the study by Anastasiadou et al [26], our primary objective was to evaluate the clinical efficacy of the ConVRself platform under 3 different conditions: one that uses the embodiment and body-swapping techniques together with the MI training, one only using the embodiment technique, and a group receiving treatment as usual (TAU).

We hypothesized that participants who used ConVRself with embodiment and body-swapping techniques, along with the MI training, would show greater improvement in the primary outcomes (motivation to lose weight and exercise more) as well as the secondary outcomes (lifestyle habits and psychological well-being) from baseline (T0) to the 3 postintervention assessments compared to participants who used ConVRself with only the embodiment element or the TAU group.

Methods

Ethical Considerations

This study was approved by the Clinical Research Ethics Committee and Research Projects Committee of the Vall d'Hebron University Hospital (VHUH). The study protocol was preregistered at ClinicalTrials.gov (NCT05094557) and published at *BMJ Open* [26]. Before enrollment in the study, written informed consent was obtained from all participants. To maintain confidentiality, each participant was assigned a numerical code. No compensation was provided for participating in the study.

Recruitment

Participants from the BS waiting list from the obesity unit of the VHUH within the national health system were assessed for eligibility between December 2021 and April 2023. To be eligible for inclusion, participants had to meet the following criteria: aged between 18 and 65 years; BMI ≥30 kg/m^2; receiving ambulatory treatment at the VHUH; not undergoing any other concurrent treatment specifically related to their obesity condition from other centers; possessing minimal digital skills, which means being able to use a digital device (smartphone, tablet, or computer) to make telephone calls and have video conversations via the internet, send or receive emails, and search for information about products and (health) services; demonstrable oral and written understanding of the Spanish language; and willingness to provide informed consent to participate. As stated in the usability study conducted by Anastasiadou et al [27], the BMI criterion was revised to include only a minimum threshold of ≥30 kg/m^2 for eligibility to participate in the study, contrary to the BMI criteria originally set in the study protocol (BMI between ≥30 kg/m^2 and ≤55 kg/m^2) [26]. Participants were not eligible if they met ≥1 of the following exclusion criteria: presence of an eating disorder in the last 2 years, nonstabilized severe mental disorder that could interfere with the successful implementation of the research protocol (ie, psychosis, depression with suicidal risk, alcohol or drug abuse, and psychotic or manic symptoms), intellectual disability or any major illness seriously affecting cognitive performance (ie, neurological disorders), and personal history of epilepsy (to avoid the potential risk of triggering seizures in this population). Of the 94 participants assessed for eligibility, 68 (72%) were recruited to participate in the study (Figure 1). Details regarding the sample size calculation are available in the protocol (32).

XSL•FO
RenderX

Figure 1. Flow of participants through each stage of the randomized controlled trial. CG: control group; EG1: experimental group 1; EG2: experimental group 2; T0: baseline; T1: after the intervention; T2: 1 week after the intervention; T3: 4 weeks after the intervention.

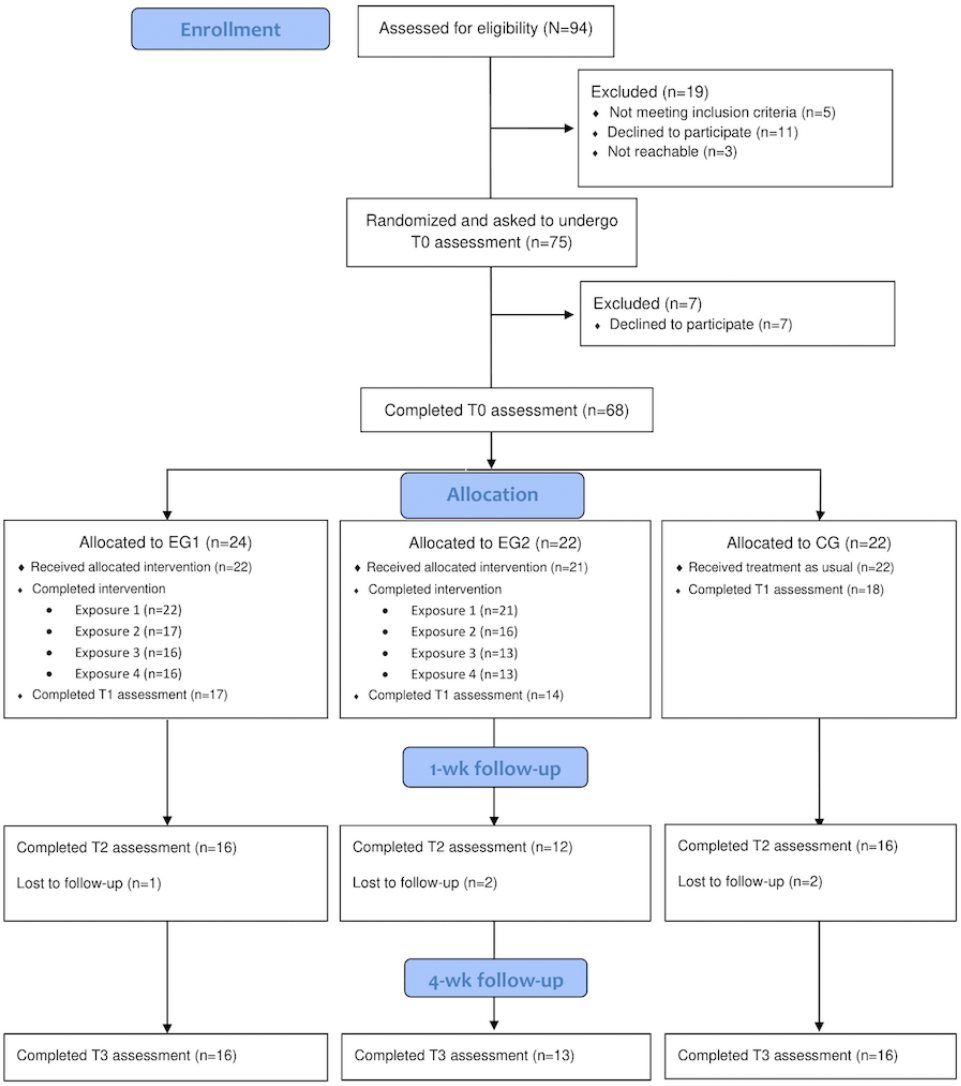

Design

We conducted an RCT with 3 parallel groups (experimental group 1 [EG1], experimental group 2 [EG2], and a control group [CG]) with a 1:1:1 allocation ratio. Measurements were carried out at 4 time points: T0, after the intervention (T1), 1 week after the intervention (T2), and 4 weeks after the intervention (T3).

Procedure

Participant data were managed and automatically distributed using REDCap (Research Electronic Data Capture; Vanderbilt University) [28,29], hosted at the Vall d'Hebron Research Institute in Barcelona, Spain. The data collected at T0, T1, T2, and T3 were obtained through web-based self-report instruments completed by participants using REDCap. In addition, the research team members measured participants' weight and height to calculate their BMI.

Potential eligible participants for the project were referred to the research team by health care professionals from the VHUH obesity unit. These participants received TAU provided by the obesity unit, which included regular medical, nutritional, and psychiatric follow-ups conducted by specialists at the hospital. Eligible participants were contacted via telephone by a research team member who provided information about the study. If they agreed to participate in the study, an appointment was scheduled at the hospital. During this appointment, a clinical interview was conducted by the team member to confirm the participants' eligibility. Sociodemographic and clinical variables were also collected, and weight and height measurements were taken. Once the necessary data were collected, participants were randomly distributed in 1 of the 3 groups (EG1, EG2, and CG) using REDCap and were then informed about the outcome of the randomization. Next, they were asked to complete the web-based T0 assessment within the following week, which was managed by the automatic email distribution facilitated by REDCap. In addition, participants were asked to watch a 30-minute video with psychoeducational advice created by the research team. This video provided information about obesity and the promotion of healthy lifestyle habits. It was shared to ensure that all participants had a similar background concerning the concept of obesity and some knowledge about healthy lifestyle habits.

After the completion of the T0 assessment, participants received their assigned interventions based on their respective groups

XSL•FO
RenderX

for a period of 10 weeks. The timeline of the assessments is depicted in Multimedia Appendix 1. The T1 assessment took place during week 6, T2 assessment during week 7, and T3 assessment during week 10 using the same automatic email distribution as with the T0 assessment. Furthermore, REDCap implemented an automatic alert system to remind participants to complete the questionnaires if they had not done so at the scheduled time. Finally, participants who underwent BS during the course of the 4 exposures (for details, refer to the Experimental Groups subsection) were still requested to respond to the T1 assessment, even if they did not complete the full intervention.

Interventions

CG Participants

After completing the T0 assessment, CG participants did not receive any intervention besides the TAU provided by the VHUH obesity unit and the psychoeducational video.

Experimental Groups

Interventions Overview

After completing the T0 assessment, EG1 participants underwent a 1-day MI training at the hospital facilities. The training lasted 4 hours and focused on developing basic MI skills. In addition, participants had an individual coaching session via telephone 1 week after the training. The initial in-person training session was led by an expert in MI (author PLP), while the follow-up sessions were carried out by the other members of the research team. At the end of the initial session, 3 photographs (2 from the front and 1 from the side) were taken of each participant to create their look-alike avatars, while their counselor's avatar was designed according to each participant's preferences concerning sex, age, and body shape. EG2 participants did not receive any MI training. Instead, they were invited to the hospital facilities where a research team member took photographs for the creation of their avatars.

Two weeks later, EG1 and EG2 participants engaged in 4 VR scenarios at the hospital facilities, assisted by a team clinical researcher. Scenarios were distributed in weekly sessions, each lasting 30 minutes, over a period of 4 weeks. After each exposure, satisfaction and adherence to the VR experiments were assessed for both groups using a semistructured interview designed by the research team. In addition, several self-report questionnaires were administered to the participants, including the readiness ruler (RR), the Suitability Evaluation Questionnaire (SEQ), and the Body Ownership Questionnaire (BOQ). For more detailed information, refer to the Measures subsection.

Specifically, the exposures of the 2 groups are detailed in the following subsections.

EG1 Participants

Overview

In each of the 4 scenarios, participants had a self-conversation using the embodiment and body-swapping techniques. Specifically, for exposures 1, 2, and 4, participants alternated between their look-alike avatar and the avatar of a counselor. During exposure 3, they alternated between their own avatar and an avatar representing their future self—a representation depicting their future self after adopting a healthier lifestyle 5 years from the present. When embodying the counselor and the future-self avatars, participants applied the MI techniques they had learned during the intensive training.

Exposure 1: Embodied Discussion About Problems and Solutions

The purpose of this scenario was to facilitate motivational self-conversations between the participant and their counselor about the lifestyle changes they planned to achieve in terms of eating healthier and being more physically active.

Exposure 2: Overcoming Self-Stigmatization

The objective of this intervention was to explore and address the participants' subjective weight stigma experiences and their interiorization through motivational self-conversations between the participant and their counselor.

Exposure 3: Illustrating the Possibility of Autonomy

The objective of this intervention was to explore, through motivational self-conversations with participants' future selves, how these future selves successfully achieve the goals that participants set in the present and to identify any barriers encountered during the process.

Exposure 4: Summing Up

Participants started their self-conversations by sharing the insights they gained from the previous exposures. In addition, they reflected on how these insights could be effectively implemented in their daily lives.

EG2 Participants

Overview

EG2 participants received a traditional counseling approach in a virtual setting. In all 4 exposures, participants were only embodied in their own look-alike avatars. First, participants engaged in a prerecorded discourse, conducted by a virtual counselor, that posed open-ended questions to which the participants responded. Second, the virtual counselor provided general and prerecorded advice that could be beneficial for the participants in promoting a healthier lifestyle.

Exposure 1: Embodied Discussion About Problems and Solutions

For this exposure, the virtual counselor asked about the perceived barriers that participants faced when trying to adopt a healthier lifestyle and then provided practical recommendations to help overcome these barriers and facilitate the adoption of a healthier lifestyle. Examples of these recommendations are as follows:

> *Thank you for sharing your goals with me. Next, I'm going to give you some simple tips that can help improve your lifestyle and overall well-being. First of all, try to avoid miracle diets as none of them work in the medium and long term...Second, make healthy choices regarding your diet. Certain foods should be prioritized, others limited, and some replaced with healthier alternatives.*

XSL•FO
RenderX

Exposure 2: Overcoming Self-Stigmatization

In this exposure, the participants shared their subjective experiences of body size discrimination, and the virtual counselor offered practical advice about how to deal with them. Some examples of the advice given are as follows:

> *Despite the enticing advertisements encouraging you to believe that image is everything, never forget that your appearance is just one aspect of who you are. Try to develop your sense of identity based on all the things you can do and the person you are deep within, despite inhabiting a larger body...Secondly, appreciate and take care of your body—a body that, when healthy, can accomplish many things.*

Exposure 3: Illustrating the Possibility of Autonomy

In this exposure, the conversation focused on discussing the potential positive effects that people living with obesity may experience when they adopt healthier behaviors that prioritize their overall health, rather than solely focusing on weight. An example of such a conversation is as follows:

> *Given that many factors influence your health status, some of which are beyond your control, one important step you can take to promote good health while living in a larger body is to adopt healthy eating and exercise habits, along with activities that foster social support, without solely focusing on weight loss.*

Exposure 4: Summing Up

Participants engaged in a conversation during which the virtual counselor provided a general summary of the main concepts explained in the previous exposures.

Technical Features of the VR System

The VR system used in the study consisted of both hardware and software components. The VR hardware used was the Meta Quest 2, which is a stand-alone headset developed by Reality Labs (Meta Platforms, Inc). The main part of the hardware consists of a head-mounted display worn by the participants. This head-mounted display has a vision- and inertia-based inside-out tracking system that allows precise tracking of the user's head movements. In addition, the Meta Quest 2 comes equipped with hand controllers that enable interaction within the virtual environment. Finally, the Meta Quest 2 also incorporates a built-in processor that generates stereoscopic images and spatialized audio. The headset runs the ConVRself application, which was developed using the Unity 3D development environment (Unity Technologies).

ConVRself Software

The VR software ConVRself, developed by Virtual Bodyworks SL, displays 3D scenarios and virtual human representations. This VR application enables participants to have self-conversations by embodying their look-alike avatar and another avatar alternatingly. The software generates scenarios that have 3 stages: calibration, tutorial, and experience. In the first stage, calibration, participants wore a VR headset and held VR controllers. The system used this setup to calculate an internal human representation that synchronized the movements of the embodied avatar with the participants' own movements. During the next 2 stages, tutorial and experience, participants were immersed in the virtual environment, embodying their look-alike avatar from a first-person perspective. To enhance the sense of embodiment, they could see themselves reflected in a virtual mirror on their left. To watch a video showing how ConVRself works, please refer to the supplementary materials in the study by Anastasiadou et al [26]. Specifically in the tutorial stage, participants had to follow detailed audio instructions provided by the application to get used to the platform before the actual experience. In addition, during this stage, participants underwent the embodiment technique to foster the illusion that the virtual body represented their own. Finally, in the experience stage, participants engaged in different exposures.

To minimize the risks associated with COVID-19, the research team followed a safety protocol that included wearing masks, carrying out regular hand disinfection, and using the CX1 decontamination system (Cleanbox Technology) to clean the Meta Quest 2 headsets and controllers used in the study.

Measures

Primary Outcomes

The primary outcomes (motivation to lose weight and exercise more) were assessed using 2 measurement tools: the RR [8] and the Stages of Change Questionnaire for Weight Management (S-Weight) and Processes of Change Questionnaire for Weight Management (P-Weight) [30]. The RR is a visual analog scale ranging from 1 to 10 that assessed participants' readiness for, confidence about, and perception of the importance of changing behavior with regard to 2 specific areas: (1) achieving a healthy weight and (2) exercising more. The S-Weight questionnaire consists of 5 mutually exclusive items that aim to allocate participants to 1 of the 5 stages of change in weight management according to the transtheoretical model: precontemplation, contemplation, preparation, action, and maintenance. The P-Weight questionnaire is a 5-point Likert scale (ranging from *strongly disagree* to *strongly agree*) consisting of 34 items developed to assess 4 processes of change for weight management: (1) emotional re-evaluation (13 items), (2) weight management actions (7 items), (3) environmental restructuring (5 items), and (4) weight consequences evaluation (9 items). The scores of each subscale were summed to obtain a total score and were then transformed on a new scale ranging from 0 to 100. The Spanish version of P-Weight showed an adequate internal consistency (Cronbach α coefficients ranged from 0.78 to 0.96 in both individuals with normal weight and individuals with overweight and obesity) [30]. The Cronbach α values in this study for the P-Weight subscales were 0.74 for emotional re-evaluation, 0.74 for weight management actions, 0.76 for environmental restructuring, and 0.76 for weight consequences evaluation.

Secondary Outcomes

Eating Habits

The Three-Factor Eating Questionnaire–Revised 18 items (TFEQ-R18) [31] is a self-report questionnaire designed to measure 3 aspects of eating behavior: (1) cognitive restraint (CR; 6 items), (2) uncontrolled eating (UE; 9 items), and (3)

XSL•FO
RenderX

emotional eating (EE; 3 items). Participants responded to each item on a 4-point Likert scale ranging from *definitely true* to *definitely false*. The total scores of each subscale were obtained by summing the scores of individual items. The Spanish version of the TFEQ-R18 [32] showed good internal consistency (Cronbach α coefficients ranged from 0.75 to 0.87) in a sample of young and healthy adults. The Cronbach α values in this study for each subscale were 0.59 for CR, 0.87 for UE, and 0.78 for EE.

The Eating Habits Questionnaire [33] is a self-report questionnaire with 37 items, each rated using a 5-point Likert scale ranging from *never* to *always*. This questionnaire measures eating habits across 8 different spheres: (1) sugar intake (4 items), (2) healthy eating (9 items), (3) physical activity (3 items), (4) diet caloric intake (5 items), (5) psychological well-being (3 items), (6) types of aliments (5 items), (7) knowledge and control (5 items), and (8) alcohol intake (2 items). A total score was obtained as the average of the scores from the 8 spheres. The Cronbach α coefficient for the complete questionnaire was 0.87 and ranged from 0.58 to 0.94 for the different spheres in a Spanish sample of adult participants living with overweight and obesity [33]. In this study, the Cronbach α value was 0.88.

Psychological Variables

Psychological functioning was estimated with the Hospital Anxiety and Depression Scale (HADS) [34]. The HADS is a self-report 14-item questionnaire (7 items for anxiety and 7 items for depression). Participants rated each item on a 4-point Likert scale to indicate the presence and severity of anxiety and depression symptoms. The total scores of each factor were obtained by summing the scores of individual items. The Spanish version showed high internal consistency, with a Cronbach α value of 0.86 for the 2 factors in a sample of patients and healthy controls [35]. The Cronbach α values in this study were 0.77 for the depression subscale and 0.79 for the anxiety subscale.

Body satisfaction was measured using the 10-item validated Spanish version of the Body Shape Questionnaire [36]. This self-report scale is rated using a 6-point Likert scale ranging from *never* to *always*. A total score was obtained by summing the scores of individual items. In this study, the Cronbach α value for this questionnaire was 0.89.

The Modified Weight Bias Internalization Scale (WBIS-M) [37] was used to assess weight bias interiorization. The WBIS-M is a self-report 11-item unidimensional scale rated using a 7-point scale ranging from *strongly disagree* to *strongly agree*. A total score was obtained as the sum of the scores of individual items. The Cronbach α coefficient for the complete questionnaire ranged from 0.93 to 0.94 in a sample of Spanish adults [37]. The Cronbach α value for this questionnaire in this study was 0.86.

The Cognitive Reserve Questionnaire [38] was used to measure participants' cognitive reserve. This self-report questionnaire consists of 8 items that evaluate aspects generally related to cognitive reserve, such as educational status (own and parental), occupational status, completion of training courses, musical training, and language proficiency. The total score was obtained by summing the item scores. The Cronbach α value for this questionnaire in this study was 0.63.

Adherence and Satisfaction Regarding VR Experiments

To measure satisfaction, acceptance, and security regarding the use of the ConVRself platform, we used the SEQ [39]. The SEQ is a 14-item questionnaire, with 13 items rated on a 5-point Likert scale ranging from *not at all* to *very much*, as well as a last open-ended question where participants can provide suggestions and additional feedback. For the specific purposes of this study, the word "rehabilitation" in item 11 was replaced by "obesity treatment." The total score was obtained by summing the scores of the first 13 items. Validation studies of the SEQ showed an acceptable internal consistency, with a Cronbach α value of 0.70 in samples of individuals with different physical pathologies. The Cronbach α value of the SEQ in this study was 0.61.

The BOQ evaluates the subjective illusion of body ownership in a VR context through a 7-point Likert scale ranging from *strongly disagree* to *strongly agree*. Specifically, the 4 questions of this scale were obtained from a previous study evaluating ConVRself [25]. The questionnaire assesses body ownership when (1) looking down at the virtual body, (2) observing oneself in a virtual mirror, (3) perceiving body movements, and (4) recognizing oneself. These questions were asked for the participants' own avatar as well as the counselor's avatar (except for the question regarding the self-recognition item, which was only asked for the participants' own avatar). For EG2 participants, the questions were asked exclusively for their own avatar because they did not experience the body-swapping technique. For more information about particular items, please refer to the protocol published in the study by Anastasiadou et al [26].

Along with the questionnaires, a brief interview was conducted after each exposure to assess participants' satisfaction with the VR experience and acceptability of ConVRself.

Statistical Analysis

Initial analyses involved comparisons of EG1, EG2, and CG participants on sociodemographic and clinical characteristics, adherence (dropout analysis), and assessment variables at the T0 level. Subsequent assessments examined those participants who completed the long-term follow-up assessment and those who did not on the same sociodemographic and clinical characteristics and outcome variables at T0. Depending on variable types or objective, various statistical tests were used: the Shapiro-Wilk test for assessing the normality of the distribution, a 1-way ANOVA with group as a factor for normally distributed variables, the Friedman test for variables with non-normal distributions, and the chi-square test for qualitative variables.

To handle missing data within questionnaires, passive multiple imputation was used. This approach updates total scores based on recent imputed values at the item level, thereby ensuring complete data for analysis [40,41].

XSL•FO
RenderX

Analyses for our primary and secondary outcomes were tested with 2-level hierarchical linear models (HLMs). These models were implemented with group (EG1, EG2, and CG) and time (T0, T1, T2, and T3 for the RR; T0, T1, and T3 for other variables) as fixed factors and participants nested within time as a random factor. For model adjustment, we used restricted maximum likelihood as the estimation method and the scaled identity as the error covariance structure. Potential moderators (such as age, sex, BMI at T0, time of treatment at the VHUH obesity unit, the presence of physical comorbidities, current mental illness, and Cognitive Reserve Questionnaire scores at T0) were examined. All covariables were grand mean centered.

An intent-to-treat (ITT) analysis was conducted using the available data of all participants for outcome estimation. This analysis, leveraging the ability of HLMs to integrate missing data, offers a more realistic, unbiased analysis compared to traditional methods [42]. To ensure the robustness of the results obtained from the primary and secondary ITT analyses, we also conducted a per-protocol analysis. This analysis included only participants who completed the RCT. Results from the per-protocol analysis were reported if they differed from the results of the aforementioned ITT analyses.

Effect sizes were calculated and reported using R^2 marginal (variance explained by the fixed effects) and R^2 conditional (variance explained by the entire model, both fixed and random effects) [43,44] for HLMs and Cohen *d* for comparison between the experimental groups in VR technique. Their magnitude was interpreted according to the Cohen guidelines where R^2=0.01 or $d \leq 0.2$ represents a small effect, R^2=0.06 or d=0.5 represents a medium effect, and $R^2 \geq 0.14$ or $d \geq 0.8$ represents a large effect [45].

Analyses were conducted using SPSS (version 29.0; IBM Corp) and RStudio (version 2022.12.0; Posit Software, PBC), with the *mice* package from RStudio [46] for passive multiple imputation. A 2-tailed significance level of .05 was applied to all statistical tests.

Results

Sample Description

As shown in Figure 1, of the initial 94 participants assessed for the study, 19 (20%) were excluded before the randomization, resulting in 75 (80%) participants being enrolled and randomized. However, after receiving the T0 assessment, of the 75 participants, 7 (9%) decided not to continue participating; therefore, 68 (91%) participants completed the T0 assessment and were allocated to 1 of the 3 groups: EG1 (n=24, 35%), EG2 (n=22, 32%), and CG (n=22, 32%).

In EG1, of the 24 participants, 22 (92%) received the allocated intervention, of whom 16 (73%) completed the 4 exposures. At T1, 77% (17/22) responded; and at T2 and T3, 73% (16/22) responded. In EG2, of the 22 participants, 21 (95%) received the allocated intervention, of whom 13 (62%) completed the 4 exposures. At T1, 67% (14/21) responded; at T2, 57% (12/21); and at T3, 62% (13/21). Regarding the CG, of the 22 participants, at T1, 18 (82%) responded; at T2, 16 (73%); and at T3, 16 (73%). The adherence analysis did not reveal statistical differences among the groups (χ^2_6=4.6; *P*=.60).

The sociodemographic and clinical characteristics of the sample are presented in Multimedia Appendix 2. Participants had a mean age of 44.22 (SD 10.30) years, a mean BMI of 43.58 (SD 5.96) kg/m^2, and had been receiving treatment at the VHUH obesity unit for an average of 21.62 (SD 11.18) months. Most of the participants were female individuals (54/68, 79%), Spanish citizens (51/68, 75%), employed either part time or full time (39/68, 57%), and lived with their family (50/68, 74%). Regarding clinical data, most of the participants had physical comorbidities (58/68, 85%), with pain and cardiovascular problems being the most prevalent, and no current mental illness (53/68, 78%). When comparing all groups on sociodemographic variables, significant differences were found regarding sex (χ^2_2=8.4; *P*=.02), the presence of current mental illness (χ^2_2=8.4; *P*=.02), and physical comorbidities (χ^2_2=6.0; *P*=.05).

Furthermore, there were no differences among the groups in any assessment variable at the T0 level (*P*>.05). Descriptive results of the primary and secondary outcomes of all participants, and separately for each group, are presented in Multimedia Appendices 3 and 4.

When comparing participants who completed the long-term follow-up assessments and those who did not, significant results were found regarding the following variables: (1) age (t_{65}=−2.68; *P*=.009), (2) TFEQ-R18 CR (t_{66}=−2.74; *P*=.008), (3) Eating Habits Questionnaire total score (t_{64}=−3.23; *P*=.002), and (4) S-Weight (χ^2_3=16.3; *P*=.001). More precisely, participants with a higher mean age (46.4, SD 10.3) were more likely to complete the follow-up assessments than younger participants (mean age 39.7, SD 8.8). In addition, participants who completed all assessment points had significantly higher means in the TFEQ-R18 CR subscale and Eating Habits Questionnaire total score, and a higher percentage of them were at the maintenance stage (S-Weight), compared to those who did not complete all assessments.

Primary Outcomes

Overview

The results of the analysis examining the effects of group condition and time are shown in Table 1. The HLMs revealed significant results for some RR scales, while no effects were observed for group versus time and time for P-Weight and S-Weight.

XSL•FO
RenderX

Table 1. Results from the hierarchical linear models[a] for each variable[b].

Variable	β (SE)	*t* test (*df*)	*P* value	R^2 marginal[c]	R^2 conditional[d]
RR[e]: importance of losing weight				0.099	0.583
Time	.17 (.11)	1.61 (182)	.11		
Time×group	−.05 (.05)	−1.00 (186)	.34		
RR : importance of exercising more				0.122	0.523
Time	.33 (.17)	1.95 (188)	.05		
Time×group	−.07 (.08)	−0.96 (190)	.34		
RR : confidence to lose weight				0.151	0.727
Time	.52 (.15)	3.47 (169)	<.001		
Time×group	−.16 (.07)	−2.34 (172)	.02		
RR : confidence to exercise more				0.124	0.623
Time	.55 (.18)	3.01 (179)	.003		
Time×group	−.16 (.08)	−1.89 (183)	.06		
RR : readiness to lose weight				0.126	0.778
Time	.36 (.15)	2.48 (160)	.01		
Time×group	−.10 (.07)	−1.56 (163)	.12		
RR : readiness to exercise more				0.119	0.649
Time	.60 (.17)	3.54 (177)	.001		
Time×group	−.17 (.08)	−2.21 (181)	.03		
P-Weight[f]: emotional re-evaluation				0.165	0.712
Time	−1.05 (1.03)	0.45 (120)	.31		
Time×group	.21 (.47)	−0.36 (124)	.66		
P-Weight: weight management actions				0.154	0.667
Time	−.06 (.71)	−0.09 (124)	.93		
Time×group	.21 (.32)	0.66 (128)	.51		
P-Weight: environmental restructuring				0.136	0.750
Time	1.07 (.62)	1.72 (116)	.09		
Time×group	−.46 (.28)	−1.64 (119)	.10		
P-Weight: weight consequences evaluation				0.212	0.717
Time	.38 (0.86)	0.44 (122)	.66		
Time×group	−.16 (.39)	−0.42 (125)	.67		
S-Weight[g]				0.084	0.559
Time	−.22 (.11)	−1.92 (121)	.06		
Time×group	.10 (.05)	1.92 (125)	.06		
TFEQ-R18[h] : cognitive restraint				0.106	0.610
Time	.39 (.47)	0.82 (126)	.41		
Time×group	.08 (.21)	0.39 (130)	.70		
TFEQ-R18: uncontrolled eating				0.068	0.801
Time	−1.82 (.62)	−2.93 (107)	.004		
Time×group	.71 (.28)	2.53 (109)	.01		
TFEQ-R18: emotional eating				0.070	0.772
Time	−.83 (.30)	−2.80 (110)	.006		

XSL•FO
RenderX

Variable	β (SE)	*t* test (*df*)	*P* value	R^2 marginal[c]	R^2 conditional[d]
Time×group	.29 (.13)	2.15 (112)	.03		
Eating Habits Questionnaire total score				0.104	0.838
Time	.09 (.05)	1.88 (101)	.06		
Time×group	−.003 (.02)	−0.14 (103)	.90		
HADS[i] : anxiety				0.209	0.735
Time	−1.24 (.57)	−2.18 (117)	.03		
Time×group	.65 (.30)	2.55 (120)	.01		
HADS : depression				0.239	0.735
Time	−1.07 (.51)	−2.09 (118)	.04		
Time×group	.44 (.23)	1.92 (121)	.06		
BSQ-10[j]				0.170	0.844
Time	−.91 (1.29)	−0.70 (105)	.50		
Time×group	.46 (.58)	0.79 (107)	.43		
WBIS-M[k]				0.165	0.805
Time	−.004 (.15)	−0.03 (110)	.98		
Time×group	−.02 (.07)	−0.34 (112)	.73		

[a]All hierarchical linear models were estimated with time and group as fixed effects; participant nested time as random effects; and age, sex, BMI at baseline, time of treatment at the obesity unit of the hospital, the presence of physical comorbidities, current mental illness, and Cognitive Reserve Questionnaire scores at baseline as covariables.

[b]Numbers presented are estimated values.

[c]R^2 marginal refers to the amount of variance explained by the fixed effects.

[d]R^2 conditional refers to the amount of variance explained by the entire model, both fixed and random effects.

[e]RR: readiness ruler.

[f]P-Weight: Processes of Change Questionnaire for Weight Management.

[g]S-Weight: Stages of Change Questionnaire for Weight Management.

[h]TFEQ-R18: Three-Factor Eating Questionnaire–Revised 18 items.

[i]HADS: Hospital Anxiety and Depression Scale.

[j]BSQ-10: Body Shape Questionnaire, 10-item version.

[k]WBIS-M: Modified Weight Bias Internalization Scale.

RR Analysis

First, regarding *confidence to lose weight*, the HLM revealed a significant group versus time effect (β=−.16; *P*=.02). Post hoc comparisons revealed that both EG1 and EG2 showed significant differences compared to the CG at T0 versus T1, T0 versus T2, and T0 versus T3. This notable increase in *confidence to lose weight* for both groups can be seen in Figure 2. Second, a significant group versus time effect for *readiness to exercise more* (β=−.17; *P*=.03) was found, with post hoc comparisons showing a significant increase for EG1 compared to CG at T0 versus T2. The significant interaction effect is represented graphically in Figure 2, where we can also see how the different group conditions evolve through the different time measures. Finally, participants from all groups had a significant improvement in their *confidence to exercise more* between T0 and T1 and between T0 and T3 (β=.55; *P*=.003) and in their *readiness to lose weight* between T0 and T1, T0 and T2, and T0 and T3 (β=.36; *P*=.01).

XSL•FO
RenderX

Figure 2. Estimated means of primary outcomes *confidence to lose weight* and *readiness to exercise more* in group versus time effect. The values are based on estimated means derived from hierarchical linear models with intent-to-treat analysis. Within each group, the central line represents the estimated mean outcome, while the surrounding error areas, delimited by 2 outer lines, illustrate the upper and lower bounds of the 95% CI around the estimated mean. CG: control group; EG1: experimental group 1; EG2: experimental group 2; RR: readiness ruler; T0: baseline; T1: after the intervention; T2: 1 week after the intervention; T3: 4 weeks after the intervention.

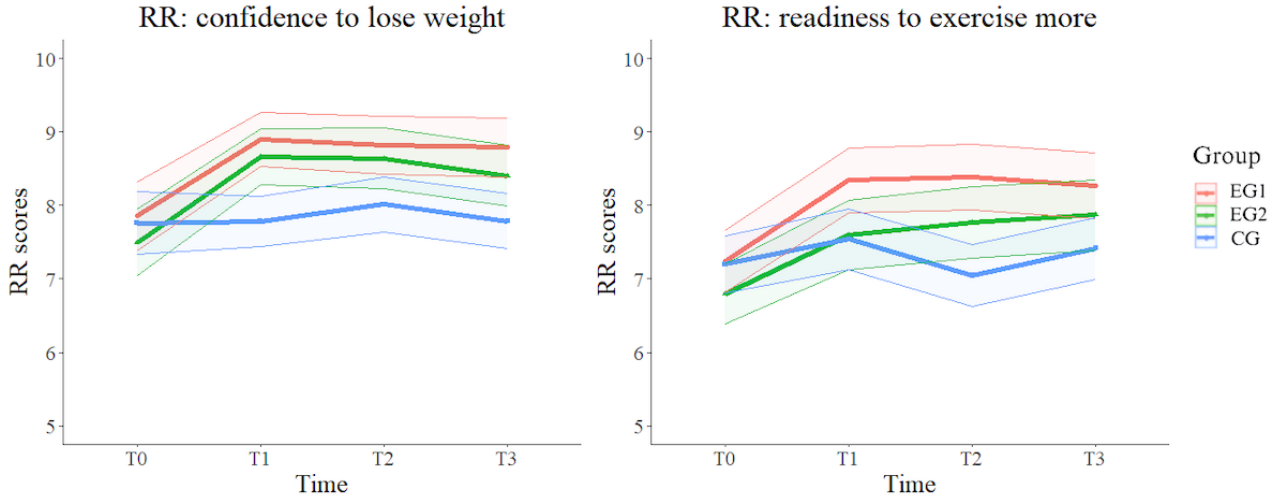

Secondary Outcomes

Overview

The HLMs revealed significant results for some TFEQ-R18 and HADS subscales, while no significant group versus time and time effects were found for the Eating Habits Questionnaire, Body Shape Questionnaire, and WBIS-M. The significant interaction effect for the TFEQ-R18 and HADS subscales is represented graphically in Multimedia Appendix 5.

TFEQ-R18 Analysis

The HLMs revealed a significant time×group effect for the UE (β=.71; *P*=.01) and EE subscales (β=.29; *P*=.03). Post hoc comparisons revealed consistently lower levels of UE for EG1 versus CG across all time measures, and for EE, a reduction for EG1 versus CG at T0 versus T3.

HADS Analysis

A significant group versus time effect was found for the anxiety subscale (β=.65; *P*=.01). In particular, post hoc comparisons revealed greater reductions in anxiety levels between T0 and T1 for EG1 compared to EG2 and CG. For the depression subscale, only a significant time effect was found (β=1.07; *P*=.04), indicating a decrease in depression levels across all groups over time.

The HLM per-protocol analysis revealed consistent results with the ITT analysis, showing no significant differences between the 2 approaches.

Adherence and Satisfaction Regarding VR Experiments (SEQ and BOQ)

For the SEQ, the experimental groups demonstrated high suitability scores with the VR platform (EG1: mean 59.82, SD 4.00; EG2: mean 58.43, SD 5.22; *d*=0.30, 95% CI −0.30 to 0.89).

Regarding the BOQ, both groups displayed average positive agreement scores, showing higher mean in EG1 than in EG2. Specifically, from the perspective of their look-alike avatar, for item 1 (looking down at the virtual body), EG1 had a mean of 2.09 (SD 1.17), while EG2 had a mean of 0.05 (SD 2.12; *d*=1.18, 95% CI 0.53-1.81); for item 2 (observing oneself in a virtual mirror), EG1 scored a mean of 2.34 (SD 0.70), and EG2 scored a mean of 1.53 (SD 1.58; *d*=0.66, 95% CI 0.05-1.26); for item 3 (perceiving body movements), EG1 scored a mean of 2.56 (SD 0.45), and EG2 scored a mean of 2.14 (SD 0.97; *d*=0.56, 95% CI −0.54 to 1.14); and for item 4 (recognizing oneself), EG1 scored a mean of 1.53 (SD 1.31), and EG2 scored a mean of 1.36 (SD 1.57; *d*=0.12, 95% CI −0.47 to 0.71). From the counselor's perspective, EG1 also demonstrated average agreement scores: (1) looking down at the virtual body: mean 1.72 (SD 1.25); (2) observing oneself in a virtual mirror: mean 1.99 (SD 0.81); and (3) perceiving body movements: mean 2.25 (SD 0.64).

Discussion

Principal Findings

This study is focused on assessing the clinical efficacy of the ConVRself platform in tackling some of the root causes of obesity. The findings of this study confirmed our hypothesis, indicating that ConVRself with embodiment and body-swapping elements, along with MI training, significantly enhanced participants' confidence to lose weight and readiness to perform physical exercise. In addition, this intervention proved effective in reducing dysfunctional eating behaviors and anxiety compared to the other groups. Overall, these findings are consistent with previous studies on the beneficial effects of positive, instructional, and motivational self-talk for performance [47].

The adherence analyses revealed that participants who had higher adherence to the treatment were those who reported, at T0, higher CR in relation to their eating habits, had healthier eating habits, and were in the *maintenance* stage of their change process, indicating a sustained commitment to lifestyle changes [8]. In line with these findings, a previous study showed that

XSL•FO
RenderX

patients with obesity before BS who were more ready to limit food intake and were actively engaged in physical activity were more likely to adhere to dietary and physical activity recommendations after BS [48]. The dropout ratio of the study was 33.8% and was similar among the groups. This finding is aligned with another RCT that used VR psychological treatments for weight management [49], in which no differences between treatment conditions were found in the dropout rates.

As regards the sociodemographic information at T0, the average age closely aligns with findings from various studies involving patients who were on the BS waiting list or who had recently undergone surgery [50], as well as studies involving VR in the psychological management of obesity [51]. In addition, the uneven sex distribution in the study sample—79% (54/68) of the participants were female individuals—is common in studies with patients who have undergone BS [50,52,53], and also in VR interventions for weight management [49,51]. In fact, a recent systematic review that included 24 studies that used VR to treat obesity [15] reported that in 8 trials with both men and women, 93% of the sample were women. Several studies have explored this phenomenon [54,55], showing that women tend to experience higher body image dissatisfaction, psychological disturbances, and a greater desire to lose weight than men. This may explain why they are more frequently represented in clinical research studies.

Regarding clinical variables, consistent with previous literature [56], a high proportion of our sample (58/68, 85%) exhibited physical comorbidities, with pain, endocrine disorders, breathing problems, and cardiovascular problems being the most prevalent. These findings, together with the high prevalence of participants in our sample who were unemployed or on sick leave and with severe obesity (BMI: mean 43.58, SD 5.96 kg/m^2), further confirm the debilitating nature of the disease. In addition, the most prevalent mental disorders observed were anxiety (8/68, 12%) and depression (6/68, 9%). However, neither disorder reached significant levels of morbidity based on the recommended cutoff points of the HADS [57]. Interestingly, our results indicate lower levels of anxiety and depression compared to previous research on patients who have undergone BS [58,59].

Regarding the assessment of motivation to change, measured along 3 dimensions (importance, confidence, and readiness) over time, participants initially reported a high importance placed on losing weight or exercising more. This posed challenges in detecting significant changes over time, as they were predominantly situated in the action and maintenance phases, as indicated by the S-Weight scores. However, all 3 groups improved over time in confidence to exercise more and readiness to lose weight. This was likely influenced by their positive expectations related to the BS or their mere involvement in the study. The participants' expectation of undergoing BS, in alignment with recent literature [53], potentially served as an additional motivator in their personal journey toward change [60,61]. Notably, most participants had already consulted with the nutritionist and endocrinologist of the obesity unit on multiple occasions and had received instructions on how to implement modifications to their diet and exercise routines in preparation for surgery. The adherence to these recommendations provided valuable insights into their likelihood of being eligible to undergo surgery soon. Consequently, the findings regarding changes in motivation to change over time remain inconclusive, making it difficult to draw specific conclusions regarding postintervention improvements in these aspects.

Regarding the group versus time effect on motivation to change, the results indicated that participants who used ConVRself reported a significant increase in their confidence to lose weight and a higher readiness to engage in exercise compared to the CG. These positive effects were maintained for a duration of between 1 and 4 weeks after the exposures. This finding is particularly encouraging for behavior change, given that, for a person to initiate and maintain a change process, they must believe that the change is important and have confidence in their ability to achieve it [8]. Notably, increased confidence is often associated with a greater propensity to adopt self-regulation skills, including control over eating behaviors and improvement in physical activity [62]. In this sense, our results derived from the TFEQ-R18 suggested improved control over eating behaviors over time among the ConVRself group, as evidenced by lower tendencies in EE and UE behaviors, compared to the CG. Previous studies employing VR interventions with people living with obesity have yielded similar results, consistently demonstrating increased self-efficacy and readiness to initiate behavior change [63-65].

Apart from the aforementioned improvement in eating control in EG1, the same group also experienced a notable reduction in anxiety from T0 to T1 compared to EG2 and CG. These findings align with the results of 2 RCTs conducted by Manzoni et al [64,65]. In these studies, it was demonstrated that a relaxation treatment augmented by VR was more effective in reducing anxiety and EE behaviors than traditional meditation interventions at 2-week and 3-month follow-ups among women living with obesity. It is worth mentioning that anxiety scores in EG1 decreased more after the intervention compared to EG2 and CG; however, this reduction was not sustained over time. We believe that this result can be attributed to the absence of ongoing psychological support after the intervention period with ConVRself. This finding emphasizes how patients can be highly sensitive to changes in the treatment process during the preoperative period and underscores the constant need for psychological support in all pre- and postoperative treatment phases [5]. In this context, similar to the successful application of VR in addressing various mental disorders, including anxiety [66], ConVRself has the potential to provide an additional benefit by enabling patients to cope with anxiety and develop strategies aligned with their values.

In line with our previous usability study [27], high SEQ scores on ConVRself indicated high usability and acceptance of the platform by people living with obesity, which means that the platform was well adapted to this population. In addition, as regards the body ownership of the avatars, the results obtained are similar to those reported in previous literature [25,27], which indicates that participants experienced a strong sense of body ownership over the virtual avatars. However, inconsistent outcomes were found with EG2 for the *looking down at the*

XSL•FO
RenderX

virtual body item compared to the other body ownership evaluations. The likely reason for this anomaly is a technological problem. Whenever participants looked down while embodied in their look-alike avatars, they could only see their knees. This limited visibility of their legs was primarily due to their anatomy (body size and shape), which caused most of their legs to be out of view. EG2 participants, particularly, failed to infer that seeing their knees implied seeing their entire legs. Despite the avatar's anatomy being the same for all participants, we believe that EG1 participants were more focused on engaging in self-conversation, while EG2 participants placed greater emphasis on the physical appearance of their avatars.

Limitations and Strengths

This study has several limitations. First, the high presence of physical comorbidities in the participants may have been influenced by the impact of delayed medical care and the exacerbation of conditions due to the COVID-19 pandemic [56]. This could potentially affect the generalizability of our results to the broader population. Second, a high dropout rate was observed, which made it challenging to achieve the expected adherence rate as stated in the study protocol [26]. This high dropout rate and the resulting smaller sample size may have impeded the detection of medium or small effects within the sample, particularly the potential differences among the groups. Third, the study design did not allow for a clear separation of the effects of ConVRself and the motivational training on the primary and secondary variables. The interpretations derived from our results could be attributed to either the effects of the motivational training or the virtual self-conversations, or, more likely, a combination of both. Fourth, the short follow-up period may limit the generalizability of the results. It becomes necessary to conduct a longer-term follow-up of the patients to observe whether the changes obtained with the ConVRself platform are sustained. Unfortunately, this was not feasible in this study due to time constraints imposed by the European SOCRATES project and the specific characteristics of our sample (patients on the BS waiting list). Furthermore, despite intensive basic training in MI, real competence in it demands constant and prolonged practice, potentially influencing the results. Finally, the high expectation of improvement from BS may have influenced the positive outcomes; therefore, it would be necessary to corroborate the results in a population with morbid obesity but without BS expectations.

The strengths of this study are its experimental design, specifically a study with 3 experimental groups and 4 assessment points, and the well-balanced distribution of the sample across groups. Moreover, we conducted a comprehensive evaluation of the participants' health, including a clinical interview that considered both physical comorbidities and mental illness at T0. Furthermore, EG1 participants received intensive training before the VR intervention, which was led by an MI expert. In terms of statistical analysis, HLMs were used in conjunction with ITT analyses, enabling the inclusion of all available data from the study, including information from participants who dropped out. Finally, there are no previous studies on MI training aimed at patients rather than therapists, either in obesity or other medical fields. This opens up possibilities to train patients as experts, making them self-aware about their own condition and capable of self-motivation.

Conclusions

In conclusion, using VR self-conversations to address the root causes of obesity has demonstrated important benefits and can be safely applied, with no side effects, among this population. In particular, the VR self-conversation with novel techniques of embodiment and body-swapping was well received by EG1 participants and was effective in enhancing self-efficacy and readiness to change, as well as in reducing dysfunctional eating behaviors and anxiety, compared to the other groups. Despite the apparent complexity of the procedures (self-conversation with embodiment and body-swapping), participants were able to complete the exposures, and they engaged in meaningful self-conversations about their obesity-related challenges and potential solutions. In this regard, a future study will provide qualitative data (currently under analysis and subject to another publication) on the unfolding of the motivational self-conversation process.

As for future perspectives, our findings underscore the importance of incorporating innovative psychological interventions to promote overall well-being and facilitate improvements in eating behaviors and lifestyle beyond mere weight loss. Such integrated interventions are crucial not only during the preoperative phase but also for the long-term maintenance of positive outcomes after BS. Future research should be conducted with ConVRself as a treatment not only for people living with obesity but also for patients with mental disorders or addictive behaviors. The potential of enriching virtual self-conversation during moments of blockage in patients with artificial intelligence techniques presents an exciting future research line.

Acknowledgments

The authors would like to thank the rest of the SOCRATES consortium (Self Conversation in Virtual Reality Embodiment to Enhance Healthier Lifestyles Among People with Obesity) for their contributions to the project. In addition, they would like to thank the staff of the obesity unit and psychiatry department at the Vall d'Hebron University Hospital for their help during the recruitment process. Finally, the authors would like to thank all participants who contributed to this study. This study was funded by the European Union's Horizon 2020 Research and Innovation Programme (951930). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was also financially supported by the Serra Húnter Programme in the form of a grant awarded to DA.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DA, MS, JARQ, and PLP conceived the study. DA, along with PH and PGR, wrote the manuscript. DA, JVDS, PH, and PGR conducted the recruitment during data collection. PLP led the motivational interviewing training for experimental group 1 participants. PGR conducted all statistical analyses and, with the support of PH, created all tables and figures for the study. JVDS and PLP provided valuable revisions during the manuscript writing process. BS, EAdlC, and MS adapted the virtual reality platform to the needs of people living with obesity and offered technical support during the study. AC provided support during the initial phases of the study, and MC contributed to the sample collection. Finally, BS and MS provided valuable feedback regarding previous revisions of the paper.

Conflicts of Interest

MS and BS are founders of Virtual Bodyworks SL, a spin-off company of the Universitat de Barcelona. EAdlC was employed by Virtual Bodyworks SL. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Procedure and timeline of the randomized controlled trial.
[PNG File , 304 KB - jmir_v26i1e51558_app1.png]

Multimedia Appendix 2
Sociodemographic and clinical characteristics of all participants, and separately for each group.
[DOCX File , 23 KB - jmir_v26i1e51558_app2.docx]

Multimedia Appendix 3
Descriptive statistics (mean and SD) on scales and subscales of the study divided into groups and time measures.
[DOCX File , 32 KB - jmir_v26i1e51558_app3.docx]

Multimedia Appendix 4
Stages of Change Questionnaires for Weight Management frequencies and percentages divided into time and groups.
[DOCX File , 19 KB - jmir_v26i1e51558_app4.docx]

Multimedia Appendix 5
Estimated means of secondary outcomes in group versus time effect (intent-to-treat analysis).
[PNG File , 150 KB - jmir_v26i1e51558_app5.png]

Multimedia Appendix 6
CONSORT checklist.
[PDF File (Adobe PDF File), 313 KB - jmir_v26i1e51558_app6.pdf]

References

1. Obesity and overweight. World Health Organization. 2021 Jun 9. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2024-03-06]
2. Vallis M, Macklin D, Russell-Mayhew S. Effective psychological and behavioural interventions in obesity management. Adult Obesity Clinical Practice Guidelines. 2020. URL: https://obesitycanada.ca/wp-content/uploads/2020/08/10-Psych-Interventions-2-v3-with-links_FINAL.pdf [accessed 2024-03-06]
3. Wharton S, Lau DC, Vallis M, Sharma AM, Biertho L, Campbell-Scherer D, et al. Obesity in adults: a clinical practice guideline. CMAJ 2020 Aug 04;192(31):E875-E891 [FREE Full text] [doi: 10.1503/cmaj.191707] [Medline: 32753461]
4. Breen C, O'Connell J, Geoghegan J, O'Shea D, Birney S, Tully L, et al. Obesity in adults: a 2022 adapted clinical practice guideline for Ireland. Obes Facts 2022 Oct 24;15(6):736-752 [FREE Full text] [doi: 10.1159/000527131] [Medline: 36279848]
5. Obesity: identification, assessment and management. National Institute for Health and Care Excellence. 2014 Nov 27. URL: https://www.nice.org.uk/guidance/cg189/chapter/Recommendations#behavioural-interventions [accessed 2024-03-06]
6. Tylka TL, Annunziato RA, Burgard D, Daníelsdóttir S, Shuman E, Davis C, et al. The weight-inclusive versus weight-normative approach to health: evaluating the evidence for prioritizing well-being over weight loss. J Obes 2014;2014:983495-983418 [FREE Full text] [doi: 10.1155/2014/983495] [Medline: 25147734]

XSL•FO
RenderX

7. Barrett S, Begg S, O'Halloran P, Kingsley M. Integrated motivational interviewing and cognitive behaviour therapy for lifestyle mediators of overweight and obesity in community-dwelling adults: a systematic review and meta-analyses. BMC Public Health 2018 Oct 05;18(1):1160 [FREE Full text] [doi: 10.1186/s12889-018-6062-9] [Medline: 30290793]
8. Miller WR, Rollnick S. Motivational Interviewing: Preparing People for Change Second Edition. New York, NY: Guilford Publications; 2002.
9. Holmes EA, Ghaderi A, Harmer CJ, Ramchandani PG, Cuijpers P, Morrison AP, et al. The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. Lancet Psychiatry 2018 Mar;5(3):237-286. [doi: 10.1016/S2215-0366(17)30513-8] [Medline: 29482764]
10. Kazdin AE. Technology-based interventions and reducing the burdens of mental illness: perspectives and comments on the special series. Cognit Behav Pract 2015 Aug;22(3):359-366. [doi: 10.1016/j.cbpra.2015.04.004]
11. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the 'black swan' for mental health care and a turning point for e-health. Internet Interv 2020 Apr;20:100317 [FREE Full text] [doi: 10.1016/j.invent.2020.100317] [Medline: 32289019]
12. Gutiérrez-Maldonado J, Wiederhold BK, Riva G. Future directions: how virtual reality can further improve the assessment and treatment of eating disorders and obesity. Cyberpsychol Behav Soc Netw 2016 Feb;19(2):148-153. [doi: 10.1089/cyber.2015.0412] [Medline: 26378982]
13. Castelnuovo G, Simpson S. Ebesity - e-health for obesity - new technologies for the treatment of obesity in clinical psychology and medicine. Clin Pract Epidemiol Ment Health 2011 Mar 04;7:5-8 [FREE Full text] [doi: 10.2174/1745017901107010005] [Medline: 21559215]
14. Riva G, Gutiérrez-Maldonado J, Wiederhold BK. Virtual worlds versus real body: virtual reality meets eating and weight disorders. Cyberpsychol Behav Soc Netw 2016 Feb;19(2):63-66 [FREE Full text] [doi: 10.1089/cyber.2016.29025.gri] [Medline: 26882322]
15. Al-Rasheed A, Alabdulkreem E, Alduailij M, Alduailij M, Alhalabi W, Alharbi S, et al. Virtual reality in the treatment of patients with overweight and obesity: a systematic review. Sustainability 2022 Mar 11;14(6):3324. [doi: 10.3390/su14063324]
16. Manzoni GM, Cesa GL, Bacchetta M, Castelnuovo G, Conti S, Gaggioli A, et al. Virtual reality-enhanced cognitive-behavioral therapy for morbid obesity: a randomized controlled study with 1 year follow-up. Cyberpsychol Behav Soc Netw 2016 Feb;19(2):134-140 [FREE Full text] [doi: 10.1089/cyber.2015.0208] [Medline: 26430819]
17. Navarro J, Cebolla A, Llorens R, Borrego A, Baños RM. Manipulating self-avatar body dimensions in virtual worlds to complement an internet-delivered intervention to increase physical activity in overweight women. Int J Environ Res Public Health 2020 Jun 05;17(11):4045 [FREE Full text] [doi: 10.3390/ijerph17114045] [Medline: 32517117]
18. Thomas JG, Goldstein CM, Bond DS, Hadley W, Tuerk PW. Web-based virtual reality to enhance behavioural skills training and weight loss in a commercial online weight management programme: the experience success randomized trial. Obes Sci Pract 2020 Aug 27;6(6):587-595 [FREE Full text] [doi: 10.1002/osp4.451] [Medline: 33354337]
19. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. Obes Rev 2015 Apr;16(4):319-326 [FREE Full text] [doi: 10.1111/obr.12266] [Medline: 25752756]
20. Annesi JJ, Powell SM. The role of change in self-efficacy in maintaining exercise-associated improvements in mood beyond the initial 6 months of expected weight loss in women with obesity. Int J Behav Med 2024 Feb;31(1):156-162. [doi: 10.1007/s12529-023-10164-3] [Medline: 36797550]
21. Botvinick M, Cohen J. Rubber hands 'feel' touch that eyes see. Nature 1998 Feb 19;391(6669):756. [doi: 10.1038/35784] [Medline: 9486643]
22. Slater M, Spanlang B, Sanchez-Vives MV, Blanke O. First person experience of body transfer in virtual reality. PLoS One 2010 May 12;5(5):e10564 [FREE Full text] [doi: 10.1371/journal.pone.0010564] [Medline: 20485681]
23. Petkova VI, Ehrsson HH. If I were you: perceptual illusion of body swapping. PLoS One 2008;3(12):e3832 [FREE Full text] [doi: 10.1371/journal.pone.0003832] [Medline: 19050755]
24. Osimo SA, Pizarro R, Spanlang B, Slater M. Conversations between self and self as Sigmund Freud--a virtual body ownership paradigm for self counselling. Sci Rep 2015 Sep 10;5:13899 [FREE Full text] [doi: 10.1038/srep13899] [Medline: 26354311]
25. Slater M, Neyret S, Johnston T, Iruretagoyena G, Crespo M, Alabèrnia-Segura M, et al. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. Sci Rep 2019 Jul 29;9(1):10903 [FREE Full text] [doi: 10.1038/s41598-019-46877-3] [Medline: 31358846]
26. Anastasiadou D, Slater M, Spanlang B, Cano Porras D, Comas M, Ciudin A, et al. Clinical efficacy of a virtual reality tool for the treatment of obesity: study protocol of a randomised controlled trial. BMJ Open 2022 Jun 22;12(6):e060822 [FREE Full text] [doi: 10.1136/bmjopen-2022-060822] [Medline: 35732390]
27. Anastasiadou D, Herrero P, Vázquez-De Sebastián J, Garcia-Royo P, Spanlang B, Álvarez de la Campa E, et al. Virtual self-conversation using motivational interviewing techniques to promote healthy eating and physical activity: a usability study. Front Psychiatry 2023 Apr 19;14:999656 [FREE Full text] [doi: 10.3389/fpsyt.2023.999656] [Medline: 37151962]
28. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]

XSL•FO
RenderX

29. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
30. Andrés A, Saldaña C, Gómez-Benito J. The transtheoretical model in weight management: validation of the processes of change questionnaire. Obes Facts 2011;4(6):433-442 [FREE Full text] [doi: 10.1159/000335135] [Medline: 22248993]
31. Karlsson J, Persson LO, Sjöström L, Sullivan M. Psychometric properties and factor structure of the Three-Factor Eating Questionnaire (TFEQ) in obese men and women. Results from the Swedish Obese Subjects (SOS) study. Int J Obes Relat Metab Disord 2000 Dec;24(12):1715-1725. [doi: 10.1038/sj.ijo.0801442] [Medline: 11126230]
32. Jáuregui-Lobera I, García-Cruz P, Carbonero-Carreño R, Magallares A, Ruiz-Prieto I. Psychometric properties of Spanish version of the Three-Factor Eating Questionnaire-R18 (Tfeq-Sp) and its relationship with some eating- and body image-related variables. Nutrients 2014 Dec 04;6(12):5619-5635 [FREE Full text] [doi: 10.3390/nu6125619] [Medline: 25486370]
33. Castro Rodríguez P, Bellido Guerrero D, Pertega Díaz S, Grupo Colaborativo del Estudio. [Design and validation of a new dietary habits questionnaire for the overweight and obese]. Endocrinol Nutr 2010 Apr;57(4):130-139. [doi: 10.1016/j.endonu.2010.02.006] [Medline: 20395182]
34. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
35. Quintana JM, Padierna A, Esteban C, Arostegui I, Bilbao A, Ruiz I. Evaluation of the psychometric characteristics of the Spanish version of the hospital anxiety and depression scale. Acta Psychiatr Scand 2003 Mar;107(3):216-221. [doi: 10.1034/j.1600-0447.2003.00062.x] [Medline: 12580829]
36. Warren CS, Cepeda-Benito A, Gleaves DH, Moreno S, Rodriguez S, Fernandez MC, et al. English and Spanish versions of the body shape questionnaire: measurement equivalence across ethnicity and clinical status. Int J Eat Disord 2008 Apr;41(3):265-272. [doi: 10.1002/eat.20492] [Medline: 18076088]
37. Macho S, Andrés A, Saldaña C. Validation of the modified weight bias internalization scale in a Spanish adult population. Clin Obes 2021 Aug;11(4):e12454. [doi: 10.1111/cob.12454] [Medline: 33821573]
38. Rami L, Valls-Pedret C, Bartrés-Faz D, Caprile C, Solé-Padullés C, Castellvi M, et al. [Cognitive reserve questionnaire. Scores obtained in a healthy elderly population and in one with Alzheimer's disease]. Rev Neurol 2011 Feb 16;52(4):195-201 [FREE Full text] [Medline: 21312165]
39. Gil-Gómez JA, Manzano-Hernández P, Albiol-Pérez S, Aula-Valero C, Gil-Gómez H, Lozano-Quilis JA. USEQ: a short questionnaire for satisfaction evaluation of virtual rehabilitation systems. Sensors (Basel) 2017 Jul 07;17(7):1589 [FREE Full text] [doi: 10.3390/s17071589] [Medline: 28686174]
40. Mainzer R, Apajee J, Nguyen CD, Carlin JB, Lee KJ. A comparison of multiple imputation strategies for handling missing data in multi-item scales: guidance for longitudinal studies. Stat Med 2021 Sep 20;40(21):4660-4674. [doi: 10.1002/sim.9088] [Medline: 34102709]
41. Eekhout I, de Vet HC, de Boer MR, Twisk JW, Heymans MW. Passive imputation and parcel summaries are both valid to handle missing items in studies with many multi-item scales. Stat Methods Med Res 2018 Apr;27(4):1128-1140. [doi: 10.1177/0962280216654511] [Medline: 27334917]
42. Zettle RD, Rains JC, Hayes SC. Processes of change in acceptance and commitment therapy and cognitive therapy for depression: a mediation reanalysis of Zettle and Rains. Behav Modif 2011 May;35(3):265-283. [doi: 10.1177/0145445511398344] [Medline: 21362745]
43. Nakagawa S, Johnson PC, Schielzeth H. The coefficient of determination and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J R Soc Interface 2017 Sep;14(134):20170213 [FREE Full text] [doi: 10.1098/rsif.2017.0213] [Medline: 28904005]
44. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol 2012 Dec 03;4(2):133-142 [FREE Full text] [doi: 10.1111/j.2041-210x.2012.00261.x]
45. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Milton Park, UK: Routledge; 1988.
46. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1-67. [doi: 10.18637/jss.v045.i03]
47. Tod D, Hardy J, Oliver E. Effects of self-talk: a systematic review. J Sport Exerc Psychol 2011 Oct;33(5):666-687. [doi: 10.1123/jsep.33.5.666] [Medline: 21984641]
48. Bergh I, Lundin Kvalem I, Risstad H, Sniehotta FF. Preoperative predictors of adherence to dietary and physical activity recommendations and weight loss one year after surgery. Surg Obes Relat Dis 2016 May;12(4):910-918. [doi: 10.1016/j.soard.2015.11.009] [Medline: 26965158]
49. Cesa GL, Manzoni GM, Bacchetta M, Castelnuovo G, Conti S, Gaggioli A, et al. Virtual reality for enhancing the cognitive behavioral treatment of obesity with binge eating disorder: randomized controlled study with one-year follow-up. J Med Internet Res 2013 Jun 12;15(6):e113 [FREE Full text] [doi: 10.2196/jmir.2441] [Medline: 23759286]
50. Pyykkö JE, Aydin Ö, Gerdes VE, Acherman YI, Groen AK, van de Laar AW, et al. Psychological functioning and well-being before and after bariatric surgery; what is the benefit of being self-compassionate? Br J Health Psychol 2022 Feb;27(1):96-115 [FREE Full text] [doi: 10.1111/bjhp.12532] [Medline: 33982392]

XSL•FO
RenderX

51. Johnston JD, Massey AP, Devaneaux CA. Innovation in weight loss programs: a 3-dimensional virtual-world approach. J Med Internet Res 2012 Sep 20;14(5):e120 [FREE Full text] [doi: 10.2196/jmir.2254] [Medline: 22995535]
52. Andreu A, Flores L, Molero J, Mestre C, Obach A, Torres F, et al. Patients undergoing bariatric surgery: a special risk group for lifestyle, emotional and behavioral adaptations during the COVID-19 lockdown. Lessons from the first wave. Obes Surg 2022 Feb;32(2):441-449 [FREE Full text] [doi: 10.1007/s11695-021-05792-1] [Medline: 34791617]
53. Lecube A, Sánchez E, Andrés A, Saldaña C, Morales MJ, Calañas A, et al. Assessing motivational stages and processes of change for weight management around bariatric surgery: a multicenter study. Obes Surg 2019 Oct;29(10):3348-3356 [FREE Full text] [doi: 10.1007/s11695-019-04001-4] [Medline: 31209830]
54. Wee CC, Huskey KW, Bolcic-Jankovic D, Colten ME, Davis RB, Hamel M. Sex, race, and consideration of bariatric surgery among primary care patients with moderate to severe obesity. J Gen Intern Med 2014 Jan;29(1):68-75 [FREE Full text] [doi: 10.1007/s11606-013-2603-1] [Medline: 24048655]
55. Mousapour P, Tasdighi E, Khalaj A, Mahdavi M, Valizadeh M, Taheri H, et al. Sex disparity in laparoscopic bariatric surgery outcomes: a matched-pair cohort analysis. Sci Rep 2021 Jun 17;11(1):12809 [FREE Full text] [doi: 10.1038/s41598-021-92254-4] [Medline: 34140595]
56. Walędziak M, Różańska-Walędziak A, Pędziwiatr M, Szeliga J, Proczko-Stepaniak M, Wysocki M, et al. Bariatric surgery during COVID-19 pandemic from patients' point of view-the results of a national survey. J Clin Med 2020 Jun 02;9(6):1697 [FREE Full text] [doi: 10.3390/jcm9061697] [Medline: 32498298]
57. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the hospital anxiety and depression scale. An updated literature review. J Psychosom Res 2002 Feb;52(2):69-77. [doi: 10.1016/s0022-3999(01)00296-3] [Medline: 11832252]
58. Barbuti M, Brancati GE, Calderone A, Fierabracci P, Salvetti G, Weiss F, et al. Prevalence of mood, panic and eating disorders in obese patients referred to bariatric surgery: patterns of comorbidity and relationship with body mass index. Eat Weight Disord 2022 Apr;27(3):1021-1027 [FREE Full text] [doi: 10.1007/s40519-021-01236-y] [Medline: 34137006]
59. Usubini AG, Cattivelli R, Villa V, Varallo G, Granese V, Pietrabissa G, et al. Psychological considerations for bariatric surgery. In: Saiz-Sapena N, Oviedo JM, editors. Bariatric Surgery - From the Non-Surgical Approach to the Post-Surgery Individual Care. London, UK: IntechOpen; 2020.
60. Cohn I, Raman J, Sui Z. Patient motivations and expectations prior to bariatric surgery: a qualitative systematic review. Obes Rev 2019 Nov;20(11):1608-1618. [doi: 10.1111/obr.12919] [Medline: 31419379]
61. Ahlich E, Verzijl CL, Cunning A, Wright E, Rancourt D. Patient motivations and goals for bariatric surgery: a mixed methods study. Surg Obes Relat Dis 2021 Sep;17(9):1591-1602. [doi: 10.1016/j.soard.2021.05.017] [Medline: 34134941]
62. Annesi JJ, Gorjala S. Relations of self-regulation and self-efficacy for exercise and eating and BMI change: a field investigation. Biopsychosoc Med 2010 Sep 03;4:10 [FREE Full text] [doi: 10.1186/1751-0759-4-10] [Medline: 20815891]
63. Riva G, Bacchetta M, Baruffi M, Molinari E. Virtual reality-based multidimensional therapy for the treatment of body image disturbances in obesity: a controlled study. Cyberpsychol Behav 2001 Aug;4(4):511-526. [doi: 10.1089/109493101750527079] [Medline: 11708731]
64. Manzoni GM, Pagnini F, Gorini A, Preziosa A, Castelnuovo G, Molinari E, et al. Can relaxation training reduce emotional eating in women with obesity? An exploratory study with 3 months of follow-up. J Am Diet Assoc 2009 Aug;109(8):1427-1432. [doi: 10.1016/j.jada.2009.05.004] [Medline: 19631051]
65. Manzoni GM, Gorini A, Preziosa A, Pagnini F, Castelnuovo G, Molinari E, et al. New technologies and relaxation: an explorative study on obese patients with emotional eating. J Cyberther Rehabil 2008;1(2):182-192 [FREE Full text]
66. Freeman D, Reeve S, Robinson A, Ehlers A, Clark D, Spanlang B, et al. Virtual reality in the assessment, understanding, and treatment of mental health disorders. Psychol Med 2017 Oct;47(14):2393-2400 [FREE Full text] [doi: 10.1017/S003329171700040X] [Medline: 28325167]

Abbreviations

BOQ: Body Ownership Questionnaire
BS: bariatric surgery
CG: control group
ConVRself: Virtual Reality self-talk
CR: cognitive restraint
EE: emotional eating
EG1: experimental group 1
EG2: experimental group 2
HADS: Hospital Anxiety and Depression Scale
HLM: hierarchical linear model
ITT: intent-to-treat
MI: motivational interviewing
P-Weight: Processes of Change Questionnaire for Weight Management
RCT: randomized controlled trial

XSL•FO
RenderX

REDCap: Research Electronic Data Capture
RR: readiness ruler
SEQ: Suitability Evaluation Questionnaire
S-Weight: Stages of Change Questionnaire for Weight Management
T0: baseline
T1: after the intervention
T2: 1 week after the intervention
T3: 4 weeks after the intervention
TAU: treatment as usual
TFEQ-R18: Three-Factor Eating Questionnaire–Revised 18 items
UE: uncontrolled eating
VHUH: Vall d'Hebron University Hospital
VR: virtual reality
WBIS-M: Modified Weight Bias Internalization Scale

Edited by A Mavragani; submitted 03.08.23; peer-reviewed by N Farič, V Girishan Prabhu; comments to author 22.12.23; revised version received 12.01.24; accepted 30.01.24; published 05.04.24.

Please cite as:
Anastasiadou D, Herrero P, Garcia-Royo P, Vázquez-De Sebastián J, Slater M, Spanlang B, Álvarez de la Campa E, Ciudin A, Comas M, Ramos-Quiroga JA, Lusilla-Palacios P
Assessing the Clinical Efficacy of a Virtual Reality Tool for the Treatment of Obesity: Randomized Controlled Trial
J Med Internet Res 2024;26:e51558
URL: https://www.jmir.org/2024/1/e51558
doi:10.2196/51558
PMID:38578667

©Dimitra Anastasiadou, Pol Herrero, Paula Garcia-Royo, Julia Vázquez-De Sebastián, Mel Slater, Bernhard Spanlang, Elena Álvarez de la Campa, Andreea Ciudin, Marta Comas, Josep Antoni Ramos-Quiroga, Pilar Lusilla-Palacios. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integrating Biomarkers From Virtual Reality and Magnetic Resonance Imaging for the Early Detection of Mild Cognitive Impairment Using a Multimodal Learning Approach: Validation Study

Bogyeom Park[1*], BSc; Yuwon Kim[1*], BSc; Jinseok Park[2], MD, PhD; Hojin Choi[2], MD, PhD; Seong-Eun Kim[1], PhD; Hokyoung Ryu[3], PhD; Kyoungwon Seo[1], PhD

[1]Department of Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul, Republic of Korea

[2]Department of Neurology, College of Medicine, Hanyang University, Seoul, Republic of Korea

[3]Graduate School of Technology and Innovation Management, Hanyang University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Kyoungwon Seo, PhD
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Sangsang Hall, 4th Fl
Gongneung-ro, Gongneung-dong, Nowon-gu
Seoul, 01811
Republic of Korea
Phone: 82 010 5668 8660
Email: kwseo@seoultech.ac.kr

Abstract

Background: Early detection of mild cognitive impairment (MCI), a transitional stage between normal aging and Alzheimer disease, is crucial for preventing the progression of dementia. Virtual reality (VR) biomarkers have proven to be effective in capturing behaviors associated with subtle deficits in instrumental activities of daily living, such as challenges in using a food-ordering kiosk, for early detection of MCI. On the other hand, magnetic resonance imaging (MRI) biomarkers have demonstrated their efficacy in quantifying observable structural brain changes that can aid in early MCI detection. Nevertheless, the relationship between VR-derived and MRI biomarkers remains an open question. In this context, we explored the integration of VR-derived and MRI biomarkers to enhance early MCI detection through a multimodal learning approach.

Objective: We aimed to evaluate and compare the efficacy of VR-derived and MRI biomarkers in the classification of MCI while also examining the strengths and weaknesses of each approach. Furthermore, we focused on improving early MCI detection by leveraging multimodal learning to integrate VR-derived and MRI biomarkers.

Methods: The study encompassed a total of 54 participants, comprising 22 (41%) healthy controls and 32 (59%) patients with MCI. Participants completed a virtual kiosk test to collect 4 VR-derived biomarkers (hand movement speed, scanpath length, time to completion, and the number of errors), and T_1-weighted MRI scans were performed to collect 22 MRI biomarkers from both hemispheres. Analyses of covariance were used to compare these biomarkers between healthy controls and patients with MCI, with age considered as a covariate. Subsequently, the biomarkers that exhibited significant differences between the 2 groups were used to train and validate a multimodal learning model aimed at early screening for patients with MCI among healthy controls.

Results: The support vector machine (SVM) using only VR-derived biomarkers achieved a sensitivity of 87.5% and specificity of 90%, whereas the MRI biomarkers showed a sensitivity of 90.9% and specificity of 71.4%. Moreover, a correlation analysis revealed a significant association between MRI-observed brain atrophy and impaired performance in instrumental activities of daily living in the VR environment. Notably, the integration of both VR-derived and MRI biomarkers into a multimodal SVM model yielded superior results compared to unimodal SVM models, achieving higher accuracy (94.4%), sensitivity (100%), specificity (90.9%), precision (87.5%), and F_1-score (93.3%).

XSL•FO
RenderX

Conclusions: The results indicate that VR-derived biomarkers, characterized by their high specificity, can be valuable as a robust, early screening tool for MCI in a broader older adult population. On the other hand, MRI biomarkers, known for their high sensitivity, excel at confirming the presence of MCI. Moreover, the multimodal learning approach introduced in our study provides valuable insights into the improvement of early MCI detection by integrating a diverse set of biomarkers.

(J Med Internet Res 2024;26:e54538) doi:10.2196/54538

KEYWORDS

magnetic resonance imaging; MRI; virtual reality; VR; early detection; mild cognitive impairment; multimodal learning; hand movement; eye movement

Introduction

Background

Mild cognitive impairment (MCI) represents a transitional stage in cognitive decline, positioned between normal aging and Alzheimer disease (AD). This stage is marked by challenges such as memory loss and difficulties in executing complex daily activities [1-3]. Once MCI deteriorates into AD, cognitive function cannot be restored to normal levels [4,5], leading to a significant reduction in the ability of patients with AD to independently perform daily activities [6]. This makes MCI a critical intervention point to potentially slow down cognitive decline. Therefore, the early detection of MCI plays an essential role, not only in preventing its progression to AD but also in allowing for timely interventions aimed at restoring cognitive function to levels associated with normal aging [7].

Conventionally, biomarkers such as neuropsychological tests and magnetic resonance imaging (MRI) have been deployed for early detection of MCI, facilitating the evaluation of cognitive functions and the tracking of brain changes [8]. For instance, neuropsychological tests aim to quantitatively assess multiple cognitive domains, including memory, to identify patients with MCI [9]. However, despite their widespread clinical use, these tests face challenges related to reproducibility, largely owing to elements such as response bias and the ceiling effect [10]. In contrast, MRI scans identify MCI by examining structural brain changes, particularly in memory-associated regions such as the hippocampus and entorhinal cortex [11,12]. However, the use of MRI scans is restricted due to their limited feasibility for regular and repeated assessments, given their lengthy procedures and high costs [13,14]. Thus, there is an increasing need for novel biomarkers that can effectively and economically detect MCI by leveraging reproducible behavioral characteristics observed in daily activities [15-18].

In recent studies, virtual reality (VR) technology has been used to collect behavioral data related to instrumental activities of daily living (IADLs), which are then analyzed using machine learning techniques to enhance the early detection of MCI [17-20]. As an example, Kim et al [21] devised the virtual kiosk test, wherein participants interact within a virtual environment to order food using a kiosk. Throughout this test, behavioral data related to hand movements, eye movements, and performance were collected. Using these behavioral data, the machine learning model successfully distinguished patients with MCI from healthy controls with an accuracy rate of 80.2%. Although VR-derived biomarkers have shown promise in early MCI detection, there is still an ongoing question regarding the interpretation of these biomarkers in relation to measurable structural brain changes observed through MRI, which is of critical clinical significance.

The integration of behavioral characteristics measured through VR-derived biomarkers with the brain's structural characteristics obtained from MRI biomarkers is gaining increasing attention. For example, Castegnaro et al [22] examined object location memory using VR performance data and found that individuals with damage to the hippocampus and entorhinal cortex demonstrated poorer performance, suggesting a correlation between impaired performance and damage to these brain regions. Similarly, Howett et al [23] conducted a VR navigation task and observed that participants with entorhinal cortex damage exhibited inferior performance in the task. Although these studies established significant positive correlations between performance assessed using VR-derived biomarkers and atrophy in the hippocampus and entorhinal cortex identified using MRI biomarkers, the results were confined to correlation analysis. To address this limitation, Cavedoni et al [19] emphasized the importance of further exploring the multimodal integration of VR-derived and MRI biomarkers, as it would not only establish the clinical validity of VR but also provide valuable behavioral and structural information about MCI. Consistent with these objectives, this study aimed to leverage a multimodal learning approach by integrating VR-derived and MRI biomarkers to enhance clinical feasibility and enable early and accurate detection of MCI.

Objectives

This study had 2 objectives. First, the study compared the MCI classification performance of VR-derived and MRI biomarkers with that of neuropsychological tests, which are considered the gold standard for MCI classification. This comparative analysis aimed to provide a deeper understanding of the advantages and limitations of each approach, namely VR, MRI, and neuropsychological tests. Second, the study introduced and validated a multimodal learning model that effectively improves the early detection of MCI by integrating the unique attributes of VR-derived and MRI biomarkers. This integrated approach harnesses the strengths of both modalities for more accurate and reliable MCI detection. The findings of this study proposed a novel clinical application approach that incorporates VR-derived and MRI biomarkers sequentially, offering a promising framework for clinicians to enhance their diagnostic capabilities in evaluating MCI.

XSL•FO
RenderX

Methods

Ethical Considerations

This study received ethics approval from the institutional review board of Hanyang University Hospital, Republic of Korea, in accordance with the principles outlined in the Declaration of Helsinki (HYUH-2021-08-020-004). Before conducting the experiment, we ensured that all participants received a comprehensive explanation of the study, and we obtained written informed consent from each individual. In addition, all participants provided consent for the use of their de-identified data for scientific research purposes, and only such data were used in the subsequent analysis. As a token of appreciation for their participation, participants were offered compensation in the form of detailed reports summarizing the results from neuropsychological tests, VR tasks, and MRI.

Participants

This study involved a total of 54 participants recruited from Hanyang University Hospitals in Seoul and Guri between January 2022 and July 2023. The 54 participants included 27 (50%) individuals from Seoul (n=11, 41% healthy controls and n=16, 59% patients with MCI) and 27 (50%) individuals from Guri (n=11, 41% healthy controls and n=16, 59% patients with MCI). Participants were randomly recruited from both a volunteer pool and outpatient clinics at the hospitals. The 2 neurologists reached a consensus on the diagnosis of MCI based on the cutoff score of the Seoul Neuropsychological Screening Battery–Core (SNSB-C), a gold-standard neuropsychological assessment tool specifically standardized for the Korean population [24]. The study only included participants who were capable of interacting with the VR environment and who possessed normal sensory perception, specifically in response to visual and auditory stimuli. We excluded individuals with a history of dementia, neurodegenerative disorders, or psychiatric conditions or those who had undergone brain surgeries. It is important to note that one volunteer (1/27, 4%) from Guri was excluded from recruitment due to a history of brain surgery for hydrocephalus. None of the patients with MCI included in the study were diagnosed with dementia.

Neuropsychological Tests

This study used the SNSB-C, a neuropsychological test specifically designed and standardized for the Korean population. The SNSB-C serves as a reliable and clinically validated alternative to the Mini - Mental State Examination for evaluating cognitive functions [24,25]. It assesses five distinct cognitive domains through separate assessments, including the following: (1) attention—measured using the Digit Span Test–Backward, (2) language function—evaluated using the Short Form of the Korean-Boston Naming Test, (3) visuospatial function—assessed using the Rey Complex Figure Test (RCFT), (4) memory—measured using the Seoul Verbal Learning Test–Elderly's Version–Delayed Recall, and (5) frontal and executive function—evaluated using the Digit Symbol Coding. A professional psychological evaluator with 15 years of experience conducted all the SNSB-C evaluations in this study.

VR-Derived Biomarkers

This study used the virtual kiosk test, previously developed in our research [26], as a source of VR-derived biomarkers. VR-derived biomarkers consist of features used to assess cognitive impairment by analyzing behavioral data, including hand and eye movements, collected from VR environments [17]. The virtual kiosk test aims to detect early indications of MCI by assessing behavioral data collected while participants undertake a cognitively complex IADL task, specifically ordering menu items at a virtual kiosk. The experimental setup, as shown in Figure 1, was arranged in a room containing a laptop equipped with an Intel i7-12700H processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3080 graphics card to execute Unity and VIVEPORT software, which is necessary to run the VR program. To facilitate a fully immersive VR experience, participants wore a head-mounted display with integrated eye-tracking capabilities (HTC VIVE Pro Eye) and used a hand controller in their dominant hand to select and order menu items from the virtual kiosk. Overall, 2 base stations tracked participants' movements during the test. The behavioral data recorded during this task, encompassing hand and eye movements and performance data, can provide insightful VR-derived biomarkers. For safety, participants remained seated throughout the test.

The virtual kiosk test, depicted in Figure 2, followed six steps, not including "Start" and "End" stages: (1) selecting a place to eat, (2) choosing a burger item, (3) selecting a side item, (4) choosing a drink item, (5) selecting a payment method, and (6) entering a 4-digit payment password. Before the test, participants received the following instructions verbally: "The place to eat is a restaurant. Please use the kiosk to order a shrimp burger, cheese sticks, and a Coca-Cola. Use a credit card as the payment method, and the card payment password is 6289."

Then, the behavioral data collected during the test were converted into 4 VR-derived biomarkers: hand movement speed, scanpath length, time to completion, and the number of errors (Figure 3). Hand movement speed, the first VR-derived biomarker, was determined by dividing the total distance of hand movements by the total test time, which correlates with cognitive abilities such as recognition and processing speed [21,27-29]. Scanpath length, the second VR-derived biomarker, reflects the distance traveled by participants' gaze during the test, indicating overall cognitive ability [30], information processing efficiency [31], and comprehension level [32]. Time to completion [33], the third VR-derived biomarker, represents the duration required for participants to complete all 6 steps of the test. The number of errors [26], the final VR-derived biomarker, records the total number of incorrect actions during the test, such as incorrect choices for place to eat, burger, drink, side item, or payment method or incorrect password entry.

XSL•FO
RenderX

Figure 1. Experimental setup for the virtual kiosk test. The virtual kiosk test is operated via a laptop. Participants sit and wear a head-mounted display and interact with the virtual environment using a hand controller. Their hand movements, eye movements, and performance data are tracked via base stations.

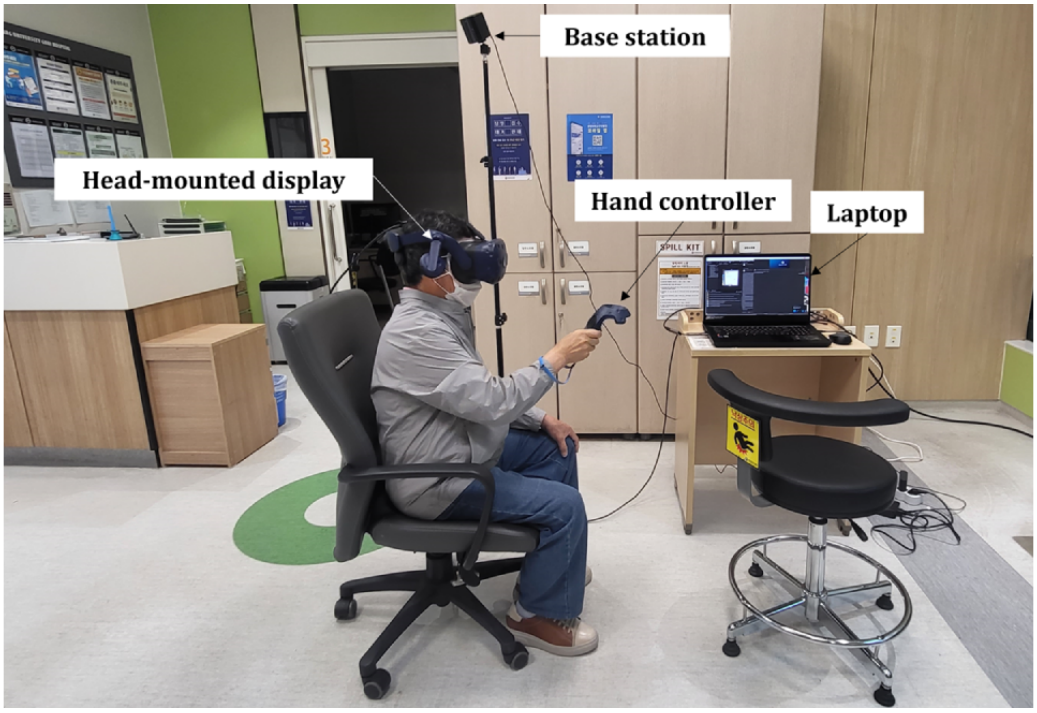

Figure 2. The 6 sequential steps of the virtual kiosk test. In step 1, participants selected a place to eat. In step 2, they chose a burger item. Step 3 involved selecting a side item, and in step 4, participants chose a drink item. Step 5 required them to select a payment method, and finally, in step 6, they entered the payment password.

XSL•FO
RenderX

Figure 3. Extraction of 4 virtual reality (VR)–derived biomarkers from behavioral data in the virtual kiosk test. Hand movement speed is calculated using the hand movement data collected from the virtual kiosk test. Scanpath length is derived from the eye movement data. The time to completion and the number of errors are calculated based on the performance data.

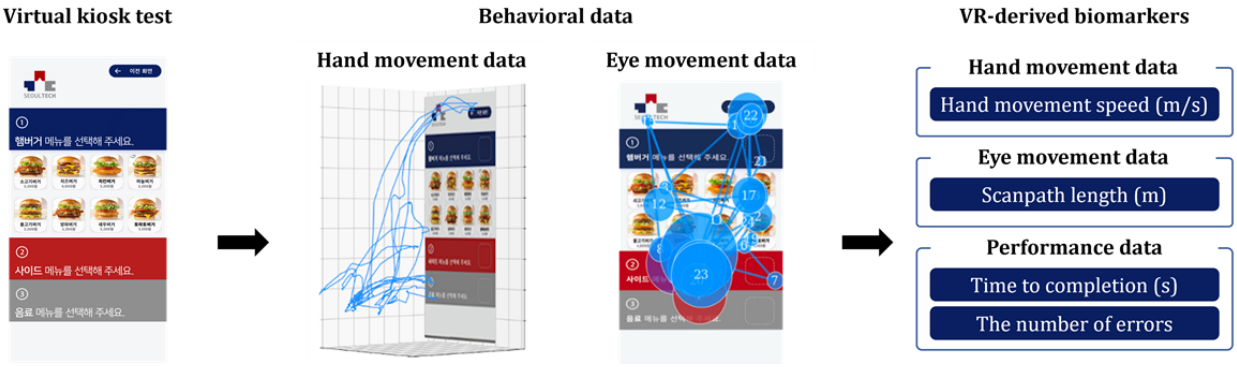

MRI Biomarkers

MRI scans were performed at Hanyang University Hospital in both Guri and Seoul locations using a Philips Ingenia CX 3T scanner. The scans used the 3D T_1-weighted magnetization-prepared rapid gradient echo technique, with specific protocols set for each hospital. The hospital at Guri adhered to the following parameters: echo time/repetition time=2.9 ms/6.3 ms, flip angle=9°, field of view=256×256 mm, with 211 slices and voxel size of 1×1×1 mm^3. The hospital at Seoul followed a slightly different protocol: echo time/repetition time=4.1 ms/6.9 ms, flip angle=8°, field of view=300×299 mm, with 170 slices and voxel size of 0.8×0.8×1 mm^3.

We used AQUA 3.0 software (Neurophet Inc) to process the acquired T_1-weighted MRIs, a method that aids in precise identification and outlining of the region of interest in the brain as depicted in Figure 4 [34,35]. The AQUA 3.0 software uses the Split-Attention U-Net deep learning architecture, which combines elements from ResNeSt and U-Net++. The architecture of Split-Attention U-Net demonstrates robustness to neuroanatomical variability through its encoder and decoder structure, skip pathways, and split-attention module. It also improves segmentation accuracy, particularly for small subcortical regions, by using EvoNorm-based convolution layers and 3D ResNeSt blocks. This methodological approach provides valuable information about brain volume from MRIs, which can be used as MRI biomarkers for neuroimaging analysis. Furthermore, the software accounts for cerebral hemispheric asymmetry by separately analyzing the left and right hemispheres, thus enabling a more holistic evaluation.

In this study, we selected specific MRI biomarkers associated with brain regions known to exhibit early atrophy in AD and play crucial roles in various cognitive functions [12,36]. The chosen MRI biomarkers, as described in Table 1, encompass cerebral white matter [37]; cerebral gray matter [38]; ventricles [39]; amygdala [40]; hippocampus [41,42]; entorhinal cortex [41,43]; parahippocampal gyrus [43]; fusiform gyrus [44]; and the superior, middle, and inferior temporal gyri [45]. These MRI biomarkers were derived from both hemispheres, resulting in a total of 22 MRI biomarkers used in this study. Furthermore, we incorporated intracranial volume (ICV), a measure encapsulating total white matter, gray matter, and cerebrospinal fluid [46], to account for overall brain size. Then, we normalized all the MRI biomarkers to the ICV to mitigate differences in brain volume due to factors such as age, sex, and head circumference [14].

Figure 4. Extraction of 22 magnetic resonance imaging (MRI) biomarkers from both hemispheres of the brain using the Split-Attention U-Net architecture. Following multilabel segmentation of the region of interest in the brain, each brain region's volume is quantified as an MRI biomarker. Each hemisphere has 11 biomarkers including the cerebral white matter; cerebral gray matter; ventricles; amygdala; hippocampus; entorhinal cortex; parahippocampal gyrus; fusiform gyrus; and superior, middle, and inferior temporal gyrus.

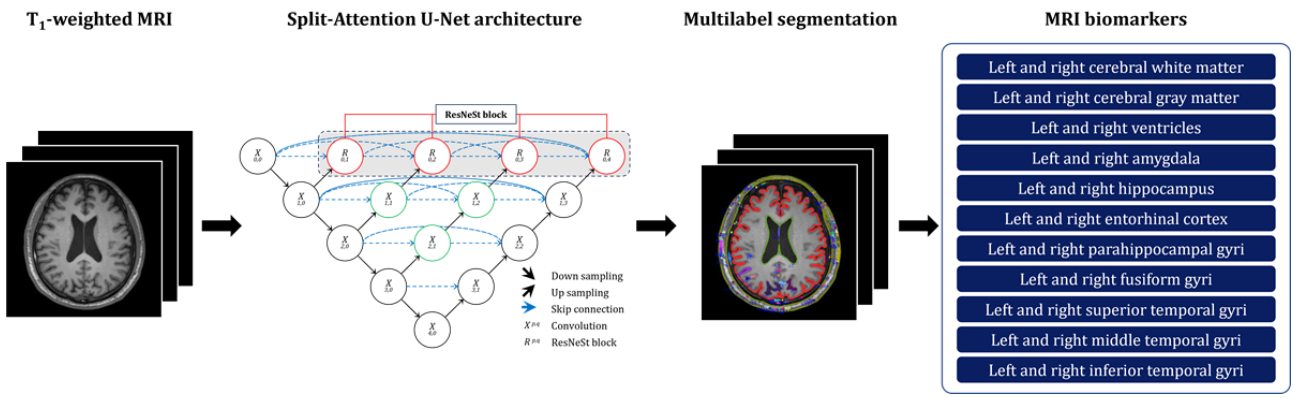

XSL•FO
RenderX

Table 1. The 11 magnetic resonance imaging biomarkers for each hemisphere (ie, a total of 22 magnetic resonance imaging biomarkers from both hemispheres) and their descriptions.

Biomarkers	Description
Cerebral white matter	Cerebral white matter is associated with brain functions such as learning, memory, and transmitting neural information between brain regions.
Cerebral gray matter	Cerebral gray matter is associated with cognitive functions, encompassing information processing, decision-making, and sensory perception.
Ventricles	Ventricles are associated with cerebrospinal fluid circulation, maintaining optimal conditions for cognitive processes in the brain.
Amygdala	The amygdala is associated with emotional processing, and the modulation of memory consolidation.
Hippocampus	The hippocampus is associated with memory formation, spatial memory, and emotional processing.
Entorhinal cortex	The entorhinal cortex is associated with functioning as a key network for spatial memory and the perception of time.
Parahippocampal gyrus	The parahippocampal gyrus is associated with cognitive processes such as spatial and episodic memory.
Fusiform gyrus	The fusiform gyrus is associated with high-level vision and multisensory perception such as object and word recognition.
Superior temporal gyrus	The superior temporal gyrus is associated with the analysis of audio-visual social information such as verbal and nonverbal communication.
Middle temporal gyrus	The middle temporal gyrus is associated with language-related tasks and integration of audio-visual emotional processing.
Inferior temporal gyrus	The inferior temporal gyrus is associated with object recognition, such as recognizing objects based on prior experiences.

Procedures

Participants initially underwent an assessment using the SNSB-C, administered by a professional psychological evaluator with 15 years of experience. MCI diagnosis was subsequently performed by 2 experienced neurologists with 18 years and 22 years of clinical expertise, adhering to the criteria established by Albert et al [47], which considered the results of the SNSB-C. Then, participants completed both the virtual kiosk test and T_1-weighted MRI scans in counterbalanced order. The virtual kiosk test was administered by the same neurologists responsible for the MCI diagnosis, whereas T_1-weighted MRI scans were conducted by a radiologist with 16 years of experience. To ensure participants' comfort and familiarity with the VR setup, they underwent 2 practice sessions before the virtual kiosk test, allowing them to get accustomed to the VR equipment and virtual environment. Throughout the experiment, measures were in place allowing participants to take breaks or halt the procedure if they experienced discomfort or dizziness due to either the VR environment or MRI scans. It is important to note that all participants completed the entire experiment successfully in an average time of 54.32 (SD 4.27) minutes, without requiring any breaks.

Analysis

We conducted statistical analysis using SPSS Statistics (version 27; IBM Corp). First, we compared of the study participants, encompassing aspects such as sex, age, and education level, were examined using a chi-square test and independent sample *t* tests to identify any statistically significant differences between the healthy control group and patients with MCI group, using a χ^2 test for a categorical variable (ie, sex) and independent sample 2-tailed *t* tests for continuous variables. Second, we performed analyses of covariance (ANCOVA) to assess the differences in neuropsychological characteristics, VR-derived biomarkers, and MRI biomarkers between the healthy controls and patients with MCI, while considering age as a covariate in this analysis. This allowed us to discern distinct features within each biomarker. In addition, we examined the relationship between VR-derived and MRI biomarkers by conducting a Pearson correlation analysis. Notably, as our data exhibited a normal distribution and homoscedasticity, all analyses were conducted using parametric tests.

Multimodal Learning

This study used multimodal learning using Python 3, with the ultimate objective of amalgamating statistically significant VR-derived and MRI biomarkers to bolster early MCI detection. The support vector machine (SVM) algorithm, with a track record of extensive use and demonstrated effectiveness in analogous tasks [4,48-50], was chosen as the machine learning model for this study. Hyperparameters for the SVM algorithm were identified through grid search, resulting in the choice of the radial basis function kernel with the regularization parameter (cost) set at 1 and kernel coefficient (γ) set at 0.1. For external validation and to prevent overfitting, we used a train and test split with a ratio of 7:3, where 70% (38/54) of the participants were assigned to the *train* subcohort and the remaining 30% (16/54) of the participants were allocated to the *test* subcohort. During the biomarker integration process, the comparative performance of models using individual VR-derived and MRI biomarkers was examined, revealing the unique characteristics of each modality. The performance evaluation of these models was conducted using several metrics, including accuracy,

XSL•FO
RenderX

sensitivity, specificity, precision, and F_1-score. In addition, the area under the receiver operating characteristic curve (AUC) was used to evaluate the performance of the binary classifier (healthy controls vs patients with MCI) in our models.

Results

Demographic and Neuropsychological Characteristics

The demographic characteristics of the study participants, encompassing aspects such as sex, age, and education level, were examined using a chi-square test and independent sample *t* tests to identify any statistically significant differences between the healthy control group and patients with MCI group. As per the results outlined in Table 2, no significant differences were found in terms of demographic characteristics between the 2 groups. Subsequently, neuropsychological characteristics were assessed using ANCOVA, with age as a covariate. The results showed a distinct contrast between the healthy control group and patients with MCI group. Patients with MCI exhibited significant impairment across all 5 evaluated cognitive functions—attention ($F_{1,51}$=24.181; P<.001; η_p^2=0.322), language function ($F_{1,51}$=14.993; P<.001; η_p^2=0.227), visuospatial function ($F_{1,51}$=19.115; P<.001; η_p^2=0.273), memory ($F_{1,51}$=32.542; P<.001; η_p^2=0.390), and frontal and executive function ($F_{1,51}$=20.584; P<.001; η_p^2=0.288)—when compared to the healthy control group, thus underscoring a marked cognitive decline in individuals with MCI.

Table 2. Comparison of basic demographic characteristics and neuropsychological test results between healthy controls and patients with mild cognitive impairment (MCI).

Characteristics	Group		*P* value
	Healthy controls (n=22)	Patients with MCI (n=32)	
Basic demographic characteristics			
Sex (female), n (%)	14 (64)	14 (44)	.15
Age (y), mean (SD)	69.86 (6.72)	73.47 (8.39)	.07
Educational level (years), mean (SD)	12.09 (4.46)	9.47 (5.12)	.06
Neuropsychological tests, mean (SD)			
DST–B[a] (number of correct answers)	4.27 (0.83)	3.00 (0.88)	<.001[b]
S-K–BNT[c] (number of correct answers)	12.91 (1.44)	10.78 (2)	<.001[b]
RCFT[d] (score)	33.34 (2.31)	26.95 (5.93)	<.001[b]
SVLT-E–DR[e] (number of correct answers)	6.82 (2.63)	2.50 (2.55)	<.001[b]
DSC[f] (number of correct answers)	60.95 (15.83)	37.97 (17.24)	<.001[b]

[a]DST–B: Digit Span Test–Backward.
[b]Analyses of covariance, with age as a covariate.
[c]S-K–BNT: Short Form of the Korean-Boston Naming Test.
[d]RCFT: Rey Complex Figure Test.
[e]SVLT-E–DR: Seoul Verbal Learning Test–Elderly's version–Delayed Recall.
[f]DSC: Digital Symbol Coding.

Differences in VR-Derived Biomarkers Between Healthy Controls and Patients With MCI

As illustrated in Table 3, significant differences were observed in VR-derived biomarkers between the healthy controls and patients with MCI when assessed through ANCOVA, with age as a covariate. Specifically, patients with MCI demonstrated slower hand movement speed ($F_{1,51}$=13.426; P=.001; η_p^2=0), longer scanpath length ($F_{1,51}$=7.108; P=.01; η_p^2=0.122), prolonged time to completion ($F_{1,51}$=9.447; P=.003; η_p^2=0.156), and a greater number of errors ($F_{1,51}$=9.438; P=.003; η_p^2=0.156) during the execution of the virtual kiosk test, as compared to healthy controls.

XSL•FO
RenderX

Table 3. Comparison of virtual reality (VR)–derived biomarkers between healthy controls and patients with mild cognitive impairment (MCI).

VR-derived biomarkers	Group, mean (SD)		*P* value[a]
	Healthy controls (n=22)	Patients with MCI (n=32)	
Hand movement features			
Hand movement speed (m/s)	0.23 (0.06)	0.17 (0.06)	.001
Eye movement feature			
Scanpath length (m)	23.66 (14.29)	60.36 (54.58)	.01
Performance features			
Time to completion (s)	39.48 (18.96)	105.39 (86.35)	.003
Number of errors	1.73 (1.61)	4 (2.81)	.003

[a]Analyses of covariance, with age as a covariate.

Differences in MRI Biomarkers Between Healthy Controls and Patients With MCI

ANCOVA was performed with age as a covariate to scrutinize the discrepancies in ICV and the proportion of ICV between the healthy control group and the patients with MCI group. As presented in Table 4, although patients with MCI exhibited higher ICV compared to healthy controls, this difference was not statistically significant. While higher ICV is typically associated with age and sex effects [51-53], our statistical analysis did not identify significant variations in age and sex between healthy controls and patients with MCI. This suggests that factors other than age and sex may be contributing to the observed higher ICV among our patients with MCI. However, patients with MCI demonstrated notable atrophy in the proportion of ICV. Specifically, significant differences were discerned in the left entorhinal cortex ($F_{1,51}$=7.821; P=.007; η_p^2=0.133), right entorhinal cortex ($F_{1,51}$=11.103; P=.002; η_p^2=0.179), left hippocampus ($F_{1,51}$=11.926; P=.001; η_p^2=0.190), right hippocampus ($F_{1,51}$=8.244; P=.006; η_p^2=0.139), left amygdala ($F_{1,51}$=7.979; P=.007; η_p^2=0.135), and right amygdala ($F_{1,51}$=6.618; P=.01; η_p^2=0.115). Refer to Multimedia Appendix 1 for detailed raw volumes of MRI biomarkers comparing healthy controls and patients with MCI.

XSL•FO
RenderX

Table 4. Comparison of magnetic resonance imaging (MRI) biomarkers between healthy controls and patients with mild cognitive impairment (MCI).

MRI biomarkers	Group, mean (SD)		*P* value[a]
	Healthy controls (n=22)	Patients with MCI (n=32)	
Raw volume (cc)			
ICV[b]	1490.96 (127.40)	1511.49 (128.89)	.56
The proportion of ICV (%)			
Left cerebral white matter	14.40 (0.73)	14.06 (0.91)	.62
Right cerebral white matter	14.51 (0.70)	14.13 (0.93)	.48
Left cerebral gray matter	15.52 (1.01)	14.98 (0.91)	.13
Right cerebral gray matter	15.46 (1.10)	14.77 (1.01)	.06
Left ventricles	1.28 (0.70)	1.57 (0.84)	.48
Right ventricles	1.16 (0.80)	1.41 (0.76)	.52
Left amygdala	0.11 (0.01)	0.10 (0.01)	.007
Right amygdala	0.12 (0.01)	0.11 (0.01)	.002
Left hippocampus	0.24 (0.02)	0.21 (0.03)	.001
Right hippocampus	0.24 (0.02)	0.22 (0.03)	.006
Left entorhinal cortex	0.15 (0.02)	0.13 (0.02)	.007
Right entorhinal cortex	0.13 (0.01)	0.11 (0.02)	.01
Left parahippocampal gyrus	0.13 (0.02)	0.12 (0.02)	.23
Right parahippocampal gyrus	0.12 (0.02)	0.12 (0.02)	.26
Left fusiform gyrus	0.60 (0.06)	0.58 (0.06)	.39
Right fusiform gyrus	0.57 (0.07)	0.57 (0.06)	.78
Left superior temporal gyrus	0.72 (0.07)	0.69 (0.06)	.17
Right superior temporal gyrus	0.69 (0.06)	0.67 (0.06)	.64
Left middle temporal gyrus	0.72 (0.08)	0.66 (0.08)	.11
Right middle temporal gyrus	0.69 (0.08)	0.65 (0.08)	.09
Left inferior temporal gyrus	0.76 (0.06)	0.74 (0.06)	.37
Right inferior temporal gyrus	0.72 (0.08)	0.68 (0.06)	.06

[a]Analyses of covariance, with age as a covariate.

[b]ICV: intracranial volume.

Correlation Between VR-Derived and MRI Biomarkers

We conducted a Pearson correlation analysis to explore the relationships among statistically significant biomarkers, including VR-derived biomarkers (hand movement speed, scanpath length, time to completion, and the number of errors) and MRI biomarkers from the amygdala, hippocampus, and entorhinal cortex. The aim was to uncover any significant associations among these biomarkers and elucidate their potential implications within the context of this study. The analysis revealed several significant correlations. Hand movement speed correlated with the right amygdala (r=0.31; P=.02), left hippocampus (r=0.40; P=.003), right hippocampus (r=0.43; P=.002), left entorhinal cortex (r=0.32; P=.02), and right entorhinal cortex (r=0.35; P=.009). Scanpath length correlated significantly with the left hippocampus (r=−0.28; P=.04). Meanwhile, time to completion showed significant correlations with the left amygdala (r=−0.31; P=.02), left hippocampus (r=−0.38; P=.005), right hippocampus (r=−0.28; P=.04), left entorhinal cortex (r=−0.29; P=.04), and right entorhinal cortex (r=−0.29; P=.04). Furthermore, the number of errors demonstrated significant correlations with the left amygdala (r=−0.34; P=.01), right amygdala (r=−0.38; P=.005), left hippocampus (r=−0.39; P=.003), right hippocampus (r=−0.38; P=.005), left entorhinal cortex (r=−0.29; P=.03), and right entorhinal cortex (r=−0.34; P=.01). Notably, the left hippocampus exhibited significant correlations with all 4 VR-derived biomarkers (Figure 5).

XSL•FO
RenderX

Figure 5. Significant correlations between 4 virtual reality–derived biomarkers (hand movement speed, scanpath length, time to completion, and the number of errors) and the magnetic resonance imaging biomarker, specifically, the left hippocampus. HC: healthy controls; MCI: mild cognitive impairment.

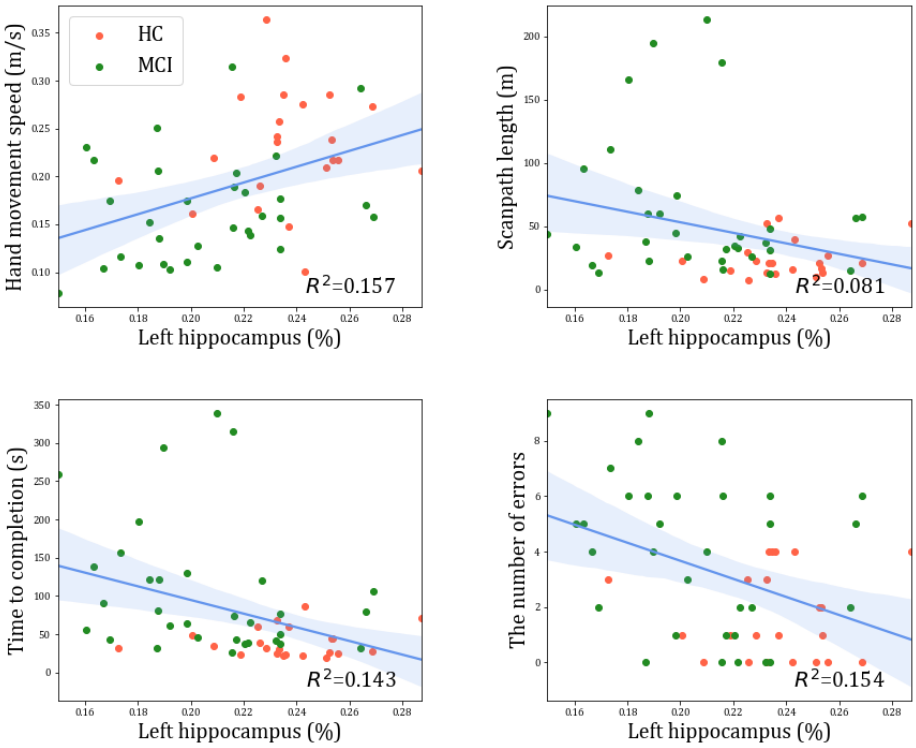

Multimodal Learning Performance Using Both VR-Derived and MRI Biomarkers

The comparative performance of multimodal learning is presented in Table 5 and Figure 6. Initially, the SVM model trained with SNSB-C data (RCFT and SVLT-E–DR), considered as a gold standard for MCI classification, exhibited strong performance, achieving an accuracy of 94.4%, a sensitivity of 100%, a specificity of 85.7%, a precision of 91.7%, an F_1-score of 95.7%, and an AUC of 0.93. The results shown in Table 5 indicate that combining SNSB-C results with either VR-derived or MRI biomarkers improves MCI classification performance compared to using VR or MRI alone. The best-performing SVM model using only VR-derived biomarkers, including hand movement speed, scanpath length, and the number of errors, achieved an accuracy of 88.9%, a sensitivity of 87.5%, a specificity of 90%, a precision of 87.5%, an F_1-score of 87.5%, and an AUC of 0.84 (refer to Multimedia Appendix 2 for more details). In the unimodal SVM model relying solely on MRI biomarkers, the combination of the left hippocampus and left entorhinal cortex led to the best performance, with an accuracy of 83.3%, a sensitivity of 90.9%, a specificity of 71.4%, a precision of 83.3%, an F_1-score of 87%, and an AUC of 0.79 (refer to Multimedia Appendix 3 for more details). Remarkably, the highest performance was achieved when both VR-derived and MRI biomarkers—specifically, hand movement speed, scanpath length, the number of errors, left entorhinal cortex, and left hippocampus—were integrated. This multimodal approach yielded an accuracy of 94.4%, a sensitivity of 100%, a specificity of 90.9%, a precision of 87.5%, an F_1-score of 93.3%, and an AUC of 0.89, with the corresponding code accessible on GitHub [54]. These results suggest that a combined approach using both VR-derived and MRI biomarkers offers the most promising outcomes, closely resembling the gold standard represented by SNSB-C. In addition, the use of VR-derived biomarkers alone showed promising results, whereas the performance while using MRI biomarkers alone was relatively lower.

Table 5. Comparative performance of the multimodal learning approach using Seoul Neuropsychological Screening Battery–Core (SNSB-C), virtual reality (VR)–derived biomarkers, and magnetic resonance imaging (MRI) biomarkers used in the support vector machine model.

Biomarker data	Accuracy, %	Sensitivity, %	Specificity, %	Precision, %	F_1-score, %
SNSB-C	94.4	100	85.7	91.7	95.7
SNSB-C+VR	94.4	92.3	100	100	96
SNSB-C+MRI	94.4	100	83.3	92.3	96
VR+MRI	94.4	100	90.9	87.5	93.3
VR	88.9	87.5	90	87.5	87.5
MRI	83.3	90.9	71.4	83.3	87

XSL•FO
RenderX

Figure 6. Comparison of receiver operating characteristic curves and the area under the receiver operating characteristic curve (AUC). The best classification performance was obtained when the support vector machine was trained using a combination of virtual reality (VR)–derived biomarkers (hand movement speed, scanpath length, and the number of errors) and magnetic resonance imaging (MRI) biomarkers (left entorhinal cortex and left hippocampus). The gold-standard Seoul Neuropsychological Screening Battery–Core (SNSB-C; Rey Complex Figure Test and Seoul Verbal Learning Test–Elderly's Version–Delayed Recall) was omitted from this comparison.

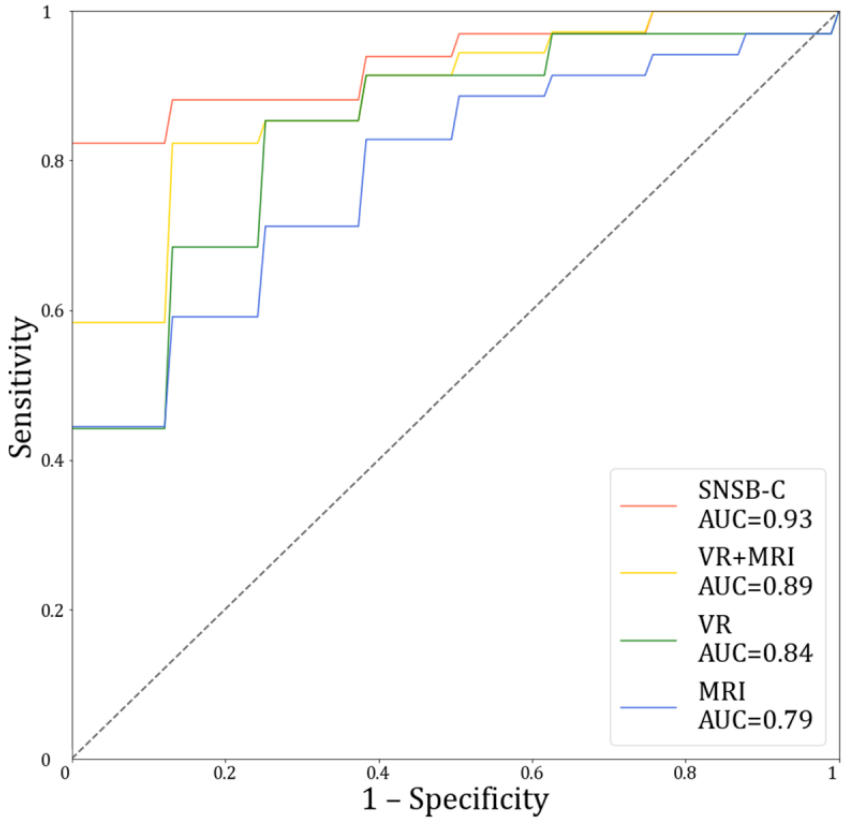

Discussion

Principal Findings

The primary objective of this study was to probe the effectiveness of a multimodal learning approach, integrating both VR-derived and MRI biomarkers, in augmenting early MCI detection. The obtained results showed significant differences in both VR-derived and MRI biomarkers when comparing patients with MCI and healthy controls. Specifically, patients with MCI displayed considerably slower hand movement speed, lengthier scanpath length, prolonged time to completion, and a greater number of errors in the virtual kiosk test compared to their healthy counterparts. The MRI biomarkers indicated noteworthy cerebral atrophy in the bilateral amygdala, hippocampus, and entorhinal cortex among patients with MCI. A remarkable observation was the superior performance of the multimodal learning approach, which incorporated both VR-derived and MRI biomarkers, in the prediction of MCI. This integrated approach achieved an accuracy of 94.4%, a sensitivity of 100%, a specificity of 90.9%, a precision of 87.5%, an F_1-score of 93.3%, and an AUC of 0.89. It outperformed models based solely on either VR-derived or MRI biomarkers and showed performance comparable to that of SNSB-C, the gold standard assessment tool for MCI diagnosis in our study. Importantly, the combination of VR-derived and MRI biomarkers allowed for faster MCI detection, significantly reducing the time required compared to neuropsychological tests such as SNSB-C, which typically take approximately 2 hours [55]. These findings provide substantial evidence highlighting the potential advantages of using a combination of VR-derived and MRI biomarkers for enhancing the detection of MCI.

Our findings from the multimodal learning approach demonstrated the distinct advantages of VR-derived and MRI biomarkers in identifying patients with MCI. VR-derived biomarkers exhibited remarkable specificity of 90%, indicating their effectiveness in accurately classifying healthy controls. This finding aligns with previous studies demonstrating the high specificity of VR-derived biomarkers obtained through IADL tasks in VR, such as financial activities or public transportation tasks [16] and shopping in virtual supermarkets [56,57]. On the other hand, MRI biomarkers displayed superior sensitivity of 90.9%, showcasing their proficiency in correctly detecting patients with MCI. Interestingly, our results showed relatively higher sensitivity compared to recent MRI biomarker studies reviewed by Lombardi et al [58]. Recent MRI studies focused on optimizing a wide range of MRI features [59] and investigated subfields within brain regions, such as the hippocampal tail [60,61]. Meanwhile, our approach incorporated recent techniques for MCI screening. First, we separately measured the volumes of the left and right hemispheres to account for hemisphere asymmetry, as suggested by Mabrouk et al [62]. Subsequently, we used the hippocampus and entorhinal cortex as MRI biomarkers, which are well recognized for their sensitivity in MCI classification, based on the findings

XSL•FO
RenderX

of Park et al [63]. Finally, to mitigate the potential influences of individual characteristics on brain volume, we applied the recently acclaimed ICV normalization technique [64,65]. We believe that these approaches may have contributed to the enhanced sensitivity in MCI detection compared to traditional methodologies. However, despite their strengths, each modality also has identifiable limitations. VR-derived biomarkers may sporadically misclassify patients with MCI as healthy controls, whereas MRI biomarkers might occasionally misidentify healthy controls as patients with MCI. By fusing the strengths of both VR-derived and MRI biomarkers, we managed to circumvent these restrictions, thereby achieving impressive rates of specificity (90.9%) and sensitivity (100%). These results underscore the potential benefits of multimodal learning, harnessing the complementarity of VR-derived and MRI biomarkers to significantly improve the early detection of MCI.

The distinctive advantages of VR-derived and MRI biomarkers suggest their potential use as sequential diagnostic tools in a 2-stage diagnostic process, as proposed by Galvin et al [66]: the "detection" and "assessment and differentiation" phases. In the detection phase, a swift screening tool can be used to identify the risk of MCI among a wider population of older adults. VR-derived biomarkers align well with the requirements for a rapid screening tool in this phase, offering a brief testing duration (<5 min) with high specificity. Thus, they are effective in distinguishing healthy controls and identifying potential patients with MCI who may require further evaluation. For MCI care, these VR-derived biomarkers can be implemented in local dementia care centers. When VR-derived biomarkers are used to identify the risk of MCI among residents in these centers, individuals can be referred to local hospitals for more specific dementia assessments. Subsequently, during the assessment and differentiation phase, individuals who are flagged as potential MCI cases in the detection phase can undergo a more thorough diagnostic process to confirm MCI. In this study, MRI biomarkers prove to be an apt choice, given their high sensitivity in diagnosing patients with MCI and their ability to offer clinical evidence regarding changes in brain structure. For MCI care, the results from MRI biomarkers can be used to accurately diagnose patients with MCI and tailor treatment plans. Furthermore, the integration of both VR-derived and MRI biomarkers demonstrated the highest MCI classification performance. Overall, integrating VR-derived biomarkers as a rapid screening tool into the detection phase and MRI biomarkers as a diagnostic tool into the assessment and differentiation phase can significantly enhance the early detection of and care process for MCI. This approach not only reduces time and cost burdens on individuals but also provides invaluable support to clinicians in making accurate MCI diagnoses for the local older adult population.

Our study's findings highlight a notable correlation between behavioral characteristics derived from VR-derived biomarkers and observed alterations or damage in the brain, as identified through MRI biomarkers. The correlation analysis unveiled a positive relationship between VR-derived biomarkers and the left hippocampus. This implies that participants with a larger left hippocampus tended to exhibit faster hand movement speed, shorter scanpath length, reduced time to completion, and fewer errors during the virtual kiosk test. Considering the pivotal role of the hippocampus in memory formation and learning, this observation is congruent with previous studies [67,68] suggesting that hippocampal damage may lead to cognitive deficits and impact behaviors, including hand and eye movements during everyday activities. Consequently, individuals with hippocampal damage may display complex hand and eye movements when performing tasks such as those presented in the virtual kiosk test (as illustrated in Figure 7). An intriguing finding is that eye movement features exhibited significant correlations with MRI biomarkers (ie, the left hippocampus) and SNSB-C measures (ie, S-K–BNT, RCFT, and DSC), which aligns with recent studies emphasizing the potential importance of eye movements in identifying MCI during complex daily tasks [69,70]. Simultaneously, negative correlation was observed between the number of errors and the right amygdala, implying that participants with a smaller right amygdala tended to make more errors during the virtual kiosk test. The amygdala also plays a crucial role in memory, consistent with the findings of earlier studies [71,72] linking volume reduction to cognitive impairment. Consequently, individuals with amygdala damage may be more prone to errors during task performance. In this context, the use of VR-derived biomarkers obtained from augmented reality or VR tests involving daily activities represents a pioneering approach for the early detection of MCI [73,74], bridging the gap between the identification of behavioral abnormalities and the underlying structural changes in the brain [75,76].

This study had certain limitations that should be acknowledged. First, the MCI classification model in this study did not encompass a diverse range of individuals with varying racial characteristics [77] or include those with neurodegenerative diseases such as Parkinson disease [78] and Lewy body diseases [79], which could potentially impact cognitive impairment in older adults. Future studies investigating these aspects may aid in the development of screening tools for neurodegenerative diseases among older adults, expanding the potential user pool to include different racial groups. Second, while this study used handcrafted features extracted from VR-derived biomarkers, training deep learning models directly on the input data could allow for a more comprehensive analysis of patterns and the exploration of novel features. Finally, although SVM was used as the multimodal learning approach based on previous studies [4,48-50], alternative multimodal learning approaches should be explored and compared to potentially improve classification performance. For instance, Contrastive Language-Image Pretraining [80] is one of the multimodal integration models using text and images. Similarly, future multimodal research with VR-derived and MRI biomarkers is necessary to delve deeper into their relationship.

Despite the limitations acknowledged, our study makes a noteworthy contribution to the field compared to previous studies [21,23]. This is achieved by attaining superior MCI detection performance through the integration of VR-derived and MRI biomarkers via multimodal learning, surpassing the performance using VR-derived or MRI biomarkers individually. This consolidated approach garnered remarkable results, such as an accuracy of 94.4%, a sensitivity of 100%, a specificity of

XSL•FO
RenderX

90.9%, a precision of 87.5%, an F_1-score of 93.3%, and an AUC of 0.89. These findings underscore the synergistic benefits of integrating VR-derived and MRI biomarkers. Moreover, our correlation analysis between VR-derived and MRI biomarkers illustrated how structural brain changes can translate into behavioral modifications in daily life. Finally, we proposed an innovative clinical application strategy, wherein VR-derived biomarkers are used in the preliminary detection stage, followed by MRI biomarkers in the assessment and differentiation stage. With a high specificity (90%) being ideal for initial screening, VR-derived biomarkers are perfect for the first stage, whereas MRI biomarkers, with their high sensitivity (90.9%), are optimal for the subsequent stage. This sequential approach could potentially reduce the time and cost burden for individuals and provide clinicians valuable assistance in making accurate diagnoses. To summarize, our study demonstrated the advantages of VR-derived and MRI biomarkers' integration using multimodal learning to enhance early MCI detection.

Figure 7. Comparison of hand movements, eye movements, and T1-weighted magnetic resonance imaging (MRI) between healthy controls and patients with mild cognitive impairment (MCI). (A) 3D coordinates of hand movements (depicted in blue). (B) Participant focus points indicated by dots, with red, blue, and purple representing the start, middle, and end of gaze, respectively—dot size corresponds to fixation duration. (C) Patients with MCI exhibiting statistically significant atrophy in the amygdala, hippocampus, and entorhinal cortex compared to healthy controls.

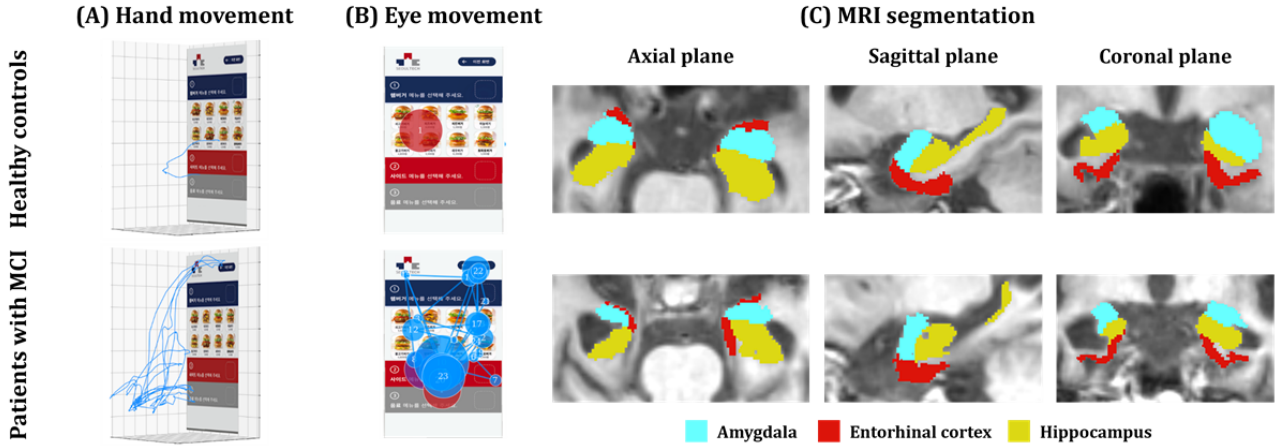

Conclusions

This study determined the unique characteristics of VR-derived and MRI biomarkers while highlighting the significance of integrating both biomarkers for early detection of MCI. The results imply that selecting the appropriate biomarker at different stages is beneficial. Specifically, VR-derived biomarkers with high specificity can be used as an early screening tool for MCI, whereas MRI biomarkers with high sensitivity are more suitable for confirming MCI. However, the most valid approach is integrating both VR-derived and MRI biomarkers, as the SVM trained using both biomarkers outperformed other models that used a single biomarker. These compelling results provide strong evidence for the potential of multimodal learning to enhance overall diagnostic performance in the early detection of MCI. Furthermore, our study contributes to understanding how structural brain changes can manifest as behavior changes by showing the relationship between VR-derived and MRI biomarkers. Our study's multimodal learning approach offers valuable insights into enhancing early MCI detection by integrating diverse biomarkers.

Acknowledgments

The support of patients was important in developing this paper. This study received support from the National Research Foundation of Korea, funded by the Korean government, Ministry of Science and Information and communication Technology (MSIT; NRF-2021R1C1C1005688 and NRF-2021R1A4A5033480). The funding sources had no role in the study conceptualization, methodology, investigation, curation, analysis of data, writing of the paper, or decision to submit the paper for publication.

Data Availability

The data sets produced and examined in this study are not publicly accessible due to concerns related to the privacy and consent of research participants. All data and materials related to the study are securely stored at the Department of Applied Artificial Intelligence, Seoul National University of Science and Technology, and can be accessed upon request.

Authors' Contributions

BP and YK contributed equally to the study conceptualization, methodology, investigation, and data curation and wrote the original draft. JP and HC contributed to the study methodology and investigation and reviewed and edited the paper. SEK reviewed and edited the paper. HR contributed to the study conceptualization, methodology, investigation, project administration, and

XSL•FO
RenderX

funding acquisition and reviewed and edited the paper. KS contributed to the study conceptualization, methodology, investigation, visualization, supervision, project administration, and funding acquisition and wrote, reviewed, and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of raw volumes of magnetic resonance imaging (MRI) biomarkers between healthy controls and patients with mild cognitive impairment (MCI).
[DOCX File , 19 KB - jmir_v26i1e54538_app1.docx]

Multimedia Appendix 2
Comparative performance of different virtual reality (VR)–derived feature combinations utilized in the Support Vector Machine (SVM) model.
[DOCX File , 16 KB - jmir_v26i1e54538_app2.docx]

Multimedia Appendix 3
Comparative performance of different magnetic resonance imaging feature combinations used in the support vector machine model.
[DOCX File , 16 KB - jmir_v26i1e54538_app3.docx]

References

1. Opwonya J, Doan DN, Kim SG, Kim JI, Ku B, Kim S, et al. Saccadic eye movement in mild cognitive impairment and Alzheimer's disease: a systematic review and meta-analysis. Neuropsychol Rev 2022 Jun 06;32(2):193-227 [FREE Full text] [doi: 10.1007/s11065-021-09495-3] [Medline: 33959887]
2. Ryu H, Seo K. The illusion of having a large virtual body biases action-specific perception in patients with mild cognitive impairment. Sci Rep 2021 Dec 15;11(1):24058 [FREE Full text] [doi: 10.1038/s41598-021-03571-7] [Medline: 34912018]
3. Scheltens P, De Strooper B, Kivipelto M, Holstege H, Chételat G, Teunissen CE, et al. Alzheimer's disease. Lancet 2021 Apr;397(10284):1577-1590. [doi: 10.1016/s0140-6736(20)32205-4]
4. Rawtaer I, Mahendran R, Kua EH, Tan HP, Tan HX, Lee TS, et al. Early detection of mild cognitive impairment with in-home sensors to monitor behavior patterns in community-dwelling senior citizens in Singapore: cross-sectional feasibility study. J Med Internet Res 2020 May 05;22(5):e16854 [FREE Full text] [doi: 10.2196/16854] [Medline: 32369031]
5. Ma J, Zhang J, Wang Z. Multimodality Alzheimer's disease analysis in deep Riemannian manifold. Inf Process Manage 2022 Jul;59(4):102965. [doi: 10.1016/j.ipm.2022.102965]
6. Bringas S, Salomón S, Duque R, Lage C, Montaña JL. Alzheimer's Disease stage identification using deep learning models. J Biomed Inform 2020 Sep;109:103514 [FREE Full text] [doi: 10.1016/j.jbi.2020.103514] [Medline: 32711124]
7. Porsteinsson AP, Isaacson RS, Knox S, Sabbagh MN, Rubino I. Diagnosis of early Alzheimer's disease: clinical practice in 2021. J Prev Alzheimers Dis 2021;8(3):371-386. [doi: 10.14283/jpad.2021.23] [Medline: 34101796]
8. Sabbagh MN, Boada M, Borson S, Chilukuri M, Dubois B, Ingram J, et al. Early detection of mild cognitive impairment (MCI) in primary care. J Prev Alzheimers Dis 2020;7(3):165-170. [doi: 10.14283/jpad.2020.21] [Medline: 32463069]
9. Harvey PD. Clinical applications of neuropsychological assessment. Dialogues Clin Neurosci 2022 Apr 01;14(1):91-99. [doi: 10.31887/dcns.2012.14.1/pharvey]
10. Howieson D. Current limitations of neuropsychological tests and assessment procedures. Clin Neuropsychol 2019 Feb 04;33(2):200-208. [doi: 10.1080/13854046.2018.1552762] [Medline: 30608020]
11. Huang H, Zheng S, Yang Z, Wu Y, Li Y, Qiu J, et al. Voxel-based morphometry and a deep learning model for the diagnosis of early Alzheimer's disease based on cerebral gray matter changes. Cereb Cortex 2023 Jan 05;33(3):754-763 [FREE Full text] [doi: 10.1093/cercor/bhac099] [Medline: 35301516]
12. He P, Qu H, Cai M, Liu W, Gu X, Ma Q. Structural alteration of medial temporal lobe subfield in the amnestic mild cognitive impairment stage of Alzheimer's disease. Neural Plast 2022 Jan 24;2022:8461235-8461211 [FREE Full text] [doi: 10.1155/2022/8461235] [Medline: 35111220]
13. Carapelle E, Mundi C, Cassano T, Avolio C. Interaction between cognitive reserve and biomarkers in Alzheimer disease. Int J Mol Sci 2020 Aug 30;21(17):6279 [FREE Full text] [doi: 10.3390/ijms21176279] [Medline: 32872643]
14. Mortimer JA, Borenstein AR, Gosche KM, Snowdon DA. Very early detection of Alzheimer neuropathology and the role of brain reserve in modifying its clinical expression. J Geriatr Psychiatry Neurol 2005 Dec 29;18(4):218-223 [FREE Full text] [doi: 10.1177/0891988705281869] [Medline: 16306243]
15. Jekel K, Damian M, Wattmo C, Hausner L, Bullock R, Connelly PJ, et al. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. Alzheimers Res Ther 2015 Mar 18;7(1):17 [FREE Full text] [doi: 10.1186/s13195-015-0099-0] [Medline: 25815063]

XSL•FO
RenderX

16. Seo K, Kim JK, Oh DH, Ryu H, Choi H. Virtual daily living test to screen for mild cognitive impairment using kinematic movement analysis. PLoS One 2017;12(7):e0181883 [FREE Full text] [doi: 10.1371/journal.pone.0181883] [Medline: 28738088]
17. Kim SY, Park J, Choi H, Loeser M, Ryu H, Seo K. Digital marker for early screening of mild cognitive impairment through hand and eye movement analysis in virtual reality using machine learning: first validation study. J Med Internet Res 2023 Oct 20;25:e48093 [FREE Full text] [doi: 10.2196/48093] [Medline: 37862101]
18. Piau A, Wild K, Mattek N, Kaye J. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild Alzheimer disease and implications for clinical care: systematic review. J Med Internet Res 2019 Aug 30;21(8):e12785. [doi: 10.2196/12785] [Medline: 31471958]
19. Cavedoni S, Chirico A, Pedroli E, Cipresso P, Riva G. Digital biomarkers for the early detection of mild cognitive impairment: artificial intelligence meets virtual reality. Front Hum Neurosci 2020 Jul 24;14:245 [FREE Full text] [doi: 10.3389/fnhum.2020.00245] [Medline: 32848660]
20. Seo K, Lee A, Kim J, Ryu H, Choi H. Measuring the kinematics of daily living movements with motion capture systems in virtual reality. J Vis Exp 2018 Apr 05(134). [doi: 10.3791/57284-V]
21. Kim SY, Park JS, Choi H, Ryu H, Seo K. Virtual kiosk test for early screening of mild cognitive impairment through eye tracking data analysis. In: Proceedings of the 2023 HCI Korea Academic Conference Presentation Paper Collection. 2023 Presented at: ACPPC '23; April 21-26, 2018; Virtual event p. 878-885 URL: https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE11229766&googleIPSandBox=false&mark=0&ipRange=false&accessgl=Y&language=ko_KR&hasTopBanner=true
22. Castegnaro A, Howett D, Li A, Harding E, Chan D, Burgess N, et al. Assessing mild cognitive impairment using object-location memory in immersive virtual environments. Hippocampus 2022 Sep 02;32(9):660-678 [FREE Full text] [doi: 10.1002/hipo.23458] [Medline: 35916343]
23. Howett D, Castegnaro A, Krzywicka K, Hagman J, Marchment D, Henson R, et al. Differentiation of mild cognitive impairment using an entorhinal cortex-based test of virtual reality navigation. Brain 2019 Jun 01;142(6):1751-1766 [FREE Full text] [doi: 10.1093/brain/awz116] [Medline: 31121601]
24. Jahng S, Na DL, Kang Y. Constructing a composite score for the Seoul neuropsychological screening battery-core. Dement Neurocong Disord 2015;14(4):137. [doi: 10.12779/dnd.2015.14.4.137]
25. Lee AY, Lee J, Oh E, Yoon SJ, Yoon B, Yu SD. Clinical utility of Seoul neuropsychological screening battery-core for dementia management project in the community. J Korean Neurol Assoc 2019 Aug 01;37(3):277-283. [doi: 10.17340/jkna.2019.3.5]
26. Kim SY, Park H, Kim H, Kim J, Seo K. Technostress causes cognitive overload in high-stress people: eye tracking analysis in a virtual kiosk test. Inf Process Manag 2022 Nov;59(6):103093 [FREE Full text] [doi: 10.1016/j.ipm.2022.103093] [Medline: 36119755]
27. Chiu HM, Hsu MC, Ouyang WC. Effects of incorporating virtual reality training intervention into health care on cognitive function and wellbeing in older adults with cognitive impairment: a randomized controlled trial. Int J Hum Comput Stud 2023 Feb;170:102957. [doi: 10.1016/j.ijhcs.2022.102957]
28. Ilardi CR, Iavarone A, La Marra M, Iachini T, Chieffi S. Hand movements in mild cognitive impairment: clinical implications and insights for future research. J Integr Neurosci 2022 Mar 23;21(2):67 [FREE Full text] [doi: 10.31083/j.jin2102067] [Medline: 35364655]
29. Park J, Seo K, Kim SE, Ryu H, Choi H. Early screening of mild cognitive impairment through hand movement analysis in virtual reality based on machine learning: screening of MCI through hand movement in VR. J Cogn Interv Digit Health 2022;1(1):1. [doi: 10.58558/jcd.2022.1.1.1]
30. Oyama A, Takeda S, Ito Y, Nakajima T, Takami Y, Takeya Y, et al. Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. Sci Rep 2019 Sep 10;9(1):12932 [FREE Full text] [doi: 10.1038/s41598-019-49275-x] [Medline: 31506486]
31. Tokushige SI, Matsumoto H, Matsuda SI, Inomata-Terada S, Kotsuki N, Hamada M, et al. Early detection of cognitive decline in Alzheimer's disease using eye tracking. Front Aging Neurosci 2023 Mar 21;15:1123456 [FREE Full text] [doi: 10.3389/fnagi.2023.1123456] [Medline: 37025964]
32. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. Educ Psychol Rev 2011 Jul 26;23(4):523-552. [doi: 10.1007/s10648-011-9174-7]
33. Chua SI, Tan NC, Wong WT, Allen Jr JC, Quah JH, Malhotra R, et al. Virtual reality for screening of cognitive function in older persons: comparative study. J Med Internet Res 2019 Aug 01;21(8):e14821 [FREE Full text] [doi: 10.2196/14821] [Medline: 31373274]
34. Kim RE, Lee M, Kang DW, Wang SM, Kim NY, Lee MK, et al. Deep learning-based segmentation to establish East Asian normative volumes using multisite structural MRI. Diagnostics (Basel) 2020 Dec 23;11(1):13 [FREE Full text] [doi: 10.3390/diagnostics11010013] [Medline: 33374745]
35. Lee M, Kim J, Ey Kim R, Kim HG, Oh SW, Lee MK, et al. Split-attention U-net: a fully convolutional network for robust multi-label segmentation from brain MRI. Brain Sci 2020 Dec 11;10(12):974 [FREE Full text] [doi: 10.3390/brainsci10120974] [Medline: 33322640]

XSL•FO
RenderX

36. de Flores R, Das SR, Xie L, Wisse LE, Lyu X, Shah P, et al. Medial temporal lobe networks in Alzheimer's disease: structural and molecular vulnerabilities. J Neurosci 2022 Jan 27;42(10):2131-2141. [doi: 10.1523/jneurosci.0949-21.2021]
37. Buyanova IS, Arsalidou M. Cerebral white matter myelination and relations to age, gender, and cognition: a selective review. Front Hum Neurosci 2021 Jul 6;15:662031 [FREE Full text] [doi: 10.3389/fnhum.2021.662031] [Medline: 34295229]
38. Yaxu Y, Ren Z, Ward J, Jiang Q. Atypical brain structures as a function of gray matter volume (GMV) and gray matter density (GMD) in young adults relating to autism spectrum traits. Front Psychol 2020 Apr 8;11:523 [FREE Full text] [doi: 10.3389/fpsyg.2020.00523] [Medline: 32322224]
39. Shao M, Han S, Carass A, Li X, Blitz AM, Shin J, et al. Brain ventricle parcellation using a deep neural network: application to patients with ventriculomegaly. Neuroimage Clin 2019;23:101871 [FREE Full text] [doi: 10.1016/j.nicl.2019.101871] [Medline: 31174103]
40. Galton CJ, Patterson K, Graham K, Lambon-Ralph MA, Williams G, Antoun N, et al. Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. Neurology 2001 Jul 24;57(2):216-225. [doi: 10.1212/wnl.57.2.216] [Medline: 11468305]
41. Du AT, Schuff N, Amend D, Laakso MP, Hsu YY, Jagust WJ, et al. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J Neurol Neurosurg Psychiatry 2001 Oct;71(4):441-447 [FREE Full text] [doi: 10.1136/jnnp.71.4.441] [Medline: 11561025]
42. Son SJ, Kim J, Park H. Structural and functional connectional fingerprints in mild cognitive impairment and Alzheimer's disease patients. PLoS One 2017 Mar 23;12(3):e0173426 [FREE Full text] [doi: 10.1371/journal.pone.0173426] [Medline: 28333946]
43. Tee B, Watson Pereira C, Lukic S, Bajorek LP, Allen IE, Miller ZA, et al. Neuroanatomical correlations of visuospatial processing in primary progressive aphasia. Brain Commun 2022;4(2):fcac060 [FREE Full text] [doi: 10.1093/braincomms/fcac060] [Medline: 35386217]
44. Weiner KS, Zilles K. The anatomical and functional specialization of the fusiform gyrus. Neuropsychologia 2016 Mar;83:48-62 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2015.06.033] [Medline: 26119921]
45. Singh V, Chertkow H, Lerch JP, Evans AC, Dorr AE, Kabani NJ. Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. Brain 2006 Nov 29;129(Pt 11):2885-2893. [doi: 10.1093/brain/awl256] [Medline: 17008332]
46. Sanfilipo MP, Benedict RH, Sharma J, Weinstock-Guttman B, Bakshi R. The relationship between whole brain volume and disability in multiple sclerosis: a comparison of normalized gray vs. white matter with misclassification correction. Neuroimage 2005 Jul 15;26(4):1068-1077. [doi: 10.1016/j.neuroimage.2005.03.008] [Medline: 15961046]
47. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May;7(3):270-279 [FREE Full text] [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
48. Swati S, Kumar M, Namasudra S. Early prediction of cognitive impairments using physiological signal for enhanced socioeconomic status. Inf Process Manage 2022 Mar;59(2):102845. [doi: 10.1016/j.ipm.2021.102845]
49. Tuladhar A, Gill S, Ismail Z, Forkert ND, Alzheimer's Disease Neuroimaging Initiative. Building machine learning models without sharing patient data: a simulation-based analysis of distributed learning by ensembling. J Biomed Inform 2020 Jun;106:103424 [FREE Full text] [doi: 10.1016/j.jbi.2020.103424] [Medline: 32335226]
50. Wei Q, Cao H, Shi Y, Xu X, Li T. Machine learning based on eye-tracking data to identify autism spectrum disorder: a systematic review and meta-analysis. J Biomed Inform 2023 Jan;137:104254 [FREE Full text] [doi: 10.1016/j.jbi.2022.104254] [Medline: 36509416]
51. Kijonka M, Borys D, Psiuk-Maksymowicz K, Gorczewski K, Wojcieszek P, Kossowski B, et al. Whole brain and cranial size adjustments in volumetric brain analyses of sex- and age-related trends. Front Neurosci 2020 Apr 3;14:278 [FREE Full text] [doi: 10.3389/fnins.2020.00278] [Medline: 32317915]
52. Sanchis-Segura C, Ibañez-Gual MV, Adrián-Ventura J, Aguirre N, Gómez-Cruz Á, Avila C, et al. Sex differences in gray matter volume: how many and how large are they really? Biol Sex Differ 2019 Jul 01;10(1):32 [FREE Full text] [doi: 10.1186/s13293-019-0245-7] [Medline: 31262342]
53. Barnes J, Ridgway GR, Bartlett J, Henley SM, Lehmann M, Hobbs N, et al. Head size, age and gender adjustment in MRI studies: a necessary nuisance? Neuroimage 2010 Dec;53(4):1244-1255. [doi: 10.1016/j.neuroimage.2010.06.025] [Medline: 20600995]
54. seoultech-HAILAB/multimodal-learning-performance-using-both-VR-derived-and-MRI-biomarkers. GitHub. 2023. URL: https://github.com/seoultech-HAILAB/Multimodal-learning-performance-using-both-VR-derived-and-MRI-biomarkers [accessed 2023-12-29]
55. Ryu HJ, Yang DW. The Seoul Neuropsychological Screening Battery (SNSB) for comprehensive neuropsychological assessment. Dement Neurocogn Disord 2023 Jan;22(1):1-15 [FREE Full text] [doi: 10.12779/dnd.2023.22.1.1] [Medline: 36814700]

XSL•FO
RenderX

56. Zygouris S, Giakoumis D, Votis K, Doumpoulakis S, Ntovas K, Segkouli S, et al. Can a virtual reality cognitive training application fulfill a dual role? Using the virtual supermarket cognitive training application as a screening tool for mild cognitive impairment. J Alzheimers Dis 2015;44(4):1333-1347. [doi: 10.3233/JAD-141260] [Medline: 25428251]
57. Eraslan Boz H, Limoncu H, Zygouris S, Tsolaki M, Giakoumis D, Votis K, et al. A new tool to assess amnestic mild cognitive impairment in Turkish older adults: virtual supermarket (VSM). Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 2020 Sep 04;27(5):639-653. [doi: 10.1080/13825585.2019.1663146] [Medline: 31482749]
58. Lombardi G, Crescioli G, Cavedo E, Lucenteforte E, Casazza G, Bellatorre AG, et al. Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. Cochrane Database Syst Rev 2020 Mar 02;3(3):CD009628 [FREE Full text] [doi: 10.1002/14651858.CD009628.pub2] [Medline: 32119112]
59. Zamani J, Sadr A, Javadi AH. Diagnosis of early mild cognitive impairment using a multiobjective optimization algorithm based on T1-MRI data. Sci Rep 2022 Jan 19;12(1):1020 [FREE Full text] [doi: 10.1038/s41598-022-04943-3] [Medline: 35046444]
60. Qu H, Ge H, Wang L, Wang W, Hu C. Volume changes of hippocampal and amygdala subfields in patients with mild cognitive impairment and Alzheimer's disease. Acta Neurol Belg 2023 Aug 12;123(4):1381-1393. [doi: 10.1007/s13760-023-02235-9] [Medline: 37043115]
61. Feng Q, Wang L, Tang X, Ge X, Hu H, Liao Z, et al. Machine learning classifiers and associations of cognitive performance with hippocampal subfields in amnestic mild cognitive impairment. Front Aging Neurosci 2023 Nov 30;15:1273658 [FREE Full text] [doi: 10.3389/fnagi.2023.1273658] [Medline: 38099266]
62. Mabrouk B, BenHamida A, Drissi N, Bouzidi N, Mhiri C. Contribution of brain regions asymmetry scores combined with random forest classifier in the diagnosis of Alzheimer’s disease in his earlier stage. J Med Biol Eng 2023 Jan 28;43(1):74-82. [doi: 10.1007/s40846-023-00775-2]
63. Park HY, Suh CH, Heo H, Shim WH, Kim SJ. Diagnostic performance of hippocampal volumetry in Alzheimer's disease or mild cognitive impairment: a meta-analysis. Eur Radiol 2022 Oct 04;32(10):6979-6991. [doi: 10.1007/s00330-022-08838-9] [Medline: 35507052]
64. Wu Y, Wang X, Fang Y. Predicting mild cognitive impairment in older adults: a machine learning analysis of the Alzheimer's Disease Neuroimaging Initiative. Geriatr Gerontol Int 2024 Mar 21;24 Suppl 1:96-101. [doi: 10.1111/ggi.14670] [Medline: 37734954]
65. Chawalparit O, Wontaneeporn N, Muangpaisan W, Thientunyakit T, Charnchaowanish P, Ngamsombat C. MRI hippocampal subfield volume analysis: comparison between Alzheimer’s disease, mild cognitive impairment, and normal aging subjects in an amyloid PET project. J Alzheimers Dis Parkinsonism 2019;09(01). [doi: 10.4172/2161-0460.1000459]
66. Galvin JE, Aisen P, Langbaum JB, Rodriguez E, Sabbagh M, Stefanacci R, et al. Early stages of Alzheimer's disease: evolving the care team for optimal patient management. Front Neurol 2020 Jan 22;11:592302 [FREE Full text] [doi: 10.3389/fneur.2020.592302] [Medline: 33551954]
67. Horvath AA, Berente DB, Vertes B, Farkas D, Csukly G, Werber T, et al. Differentiation of patients with mild cognitive impairment and healthy controls based on computer assisted hand movement analysis: a proof-of-concept study. Sci Rep 2022 Nov 09;12(1):19128 [FREE Full text] [doi: 10.1038/s41598-022-21445-4] [Medline: 36352038]
68. Whitlock JR, Sutherland RJ, Witter MP, Moser MB, Moser EI. Navigating from hippocampus to parietal cortex. Proc Natl Acad Sci U S A 2008 Sep 30;105(39):14755-14762 [FREE Full text] [doi: 10.1073/pnas.0804216105] [Medline: 18812502]
69. Opwonya J, Ku B, Lee KH, Kim JI, Kim JU. Eye movement changes as an indicator of mild cognitive impairment. Front Neurosci 2023 Jun 15;17:1171417 [FREE Full text] [doi: 10.3389/fnins.2023.1171417] [Medline: 37397453]
70. Wolf A, Tripanpitak K, Umeda S, Otake-Matsuura M. Eye-tracking paradigms for the assessment of mild cognitive impairment: a systematic review. Front Psychol 2023 Jul 20;14:1197567 [FREE Full text] [doi: 10.3389/fpsyg.2023.1197567] [Medline: 37546488]
71. Cuénod CA, Denys A, Michot JL, Jehenson P, Forette F, Kaplan D, et al. Amygdala atrophy in Alzheimer's disease: an in vivo magnetic resonance imaging study. Arch Neurol 1993 Sep 01;50(9):941-945. [doi: 10.1001/archneur.1993.00540090046009] [Medline: 8363448]
72. Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC, Alzheimer's Disease Neuroimaging Initiative. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res 2011 Oct 31;194(1):7-13 [FREE Full text] [doi: 10.1016/j.pscychresns.2011.06.014] [Medline: 21920712]
73. Muurling M, de Boer C, Vairavan S, Harms RL, Chadha AS, Tarnanas I, et al. Augmented reality versus standard tests to assess cognition and function in early Alzheimer's disease. NPJ Digit Med 2023 Dec 18;6(1):234 [FREE Full text] [doi: 10.1038/s41746-023-00978-6] [Medline: 38110486]
74. Liu Q, Song H, Yan M, Ding Y, Wang Y, Chen L, et al. Virtual reality technology in the detection of mild cognitive impairment: a systematic review and meta-analysis. Ageing Res Rev 2023 Jun;87:101889. [doi: 10.1016/j.arr.2023.101889] [Medline: 36806377]
75. Apostolova LG, Cummings JL. Neuropsychiatric manifestations in mild cognitive impairment: a systematic review of the literature. Dement Geriatr Cogn Disord 2008 Dec 14;25(2):115-126 [FREE Full text] [doi: 10.1159/000112509] [Medline: 18087152]

XSL•FO
RenderX

76. Igarashi KM. Entorhinal cortex dysfunction in Alzheimer's disease. Trends Neurosci 2023 Feb;46(2):124-136 [FREE Full text] [doi: 10.1016/j.tins.2022.11.006] [Medline: 36513524]
77. Wright CB, DeRosa JT, Moon MP, Strobino K, DeCarli C, Cheung YK, et al. Race/ethnic disparities in mild cognitive impairment and dementia: the Northern Manhattan study. J Alzheimers Dis 2021;80(3):1129-1138 [FREE Full text] [doi: 10.3233/JAD-201370] [Medline: 33646162]
78. Saredakis D, Collins-Praino LE, Gutteridge DS, Stephan BC, Keage HA. Conversion to MCI and dementia in Parkinson's disease: a systematic review and meta-analysis. Parkinsonism Relat Disord 2019 Aug;65:20-31. [doi: 10.1016/j.parkreldis.2019.04.020] [Medline: 31109727]
79. Ciafone J, Little B, Thomas AJ, Gallagher P. The neuropsychological profile of mild cognitive impairment in Lewy body dementias. J Int Neuropsychol Soc 2019 Dec 12;26(2):210-225. [doi: 10.1017/s1355617719001103]
80. Radford A, Kim JK, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38 th International Conference on Machine Learning. 2021 Presented at: ICML '21; July 18-24, 2021; Virtual event p. 8748-8763 URL: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

Abbreviations

AD: Alzheimer disease
ANCOVA: analysis of covariance
AUC: area under the receiver operating characteristic curve
IADL: instrumental activities of daily living
ICV: intracranial volume
MCI: mild cognitive impairment
MRI: magnetic resonance imaging
RCFT: Rey Complex Figure Test
SNSB-C: Seoul Neuropsychological Screening Battery–Core
SVM: support vector machine
VR: virtual reality

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 15.11.23; peer-reviewed by Y Wang, A Mastropietro; comments to author 11.12.23; revised version received 29.12.23; accepted 09.03.24; published 17.04.24.

Please cite as:
Park B, Kim Y, Park J, Choi H, Kim SE, Ryu H, Seo K
Integrating Biomarkers From Virtual Reality and Magnetic Resonance Imaging for the Early Detection of Mild Cognitive Impairment Using a Multimodal Learning Approach: Validation Study
J Med Internet Res 2024;26:e54538
URL: https://www.jmir.org/2024/1/e54538
doi:10.2196/54538
PMID:38631021

©Bogyeom Park, Yuwon Kim, Jinseok Park, Hojin Choi, Seong-Eun Kim, Hokyoung Ryu, Kyoungwon Seo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Relations Between Unique Patient Characteristics and Virtual Reality Immersion Level on Anxiety and Pain in Patients Undergoing Venipuncture: Secondary Analysis of a Randomized Control Trial

Jeffrey I Gold[1,2], PhD; Krystal M Akbar[3], PhD; Sandra Avila[3], PsyD; Nhat H Ngo[4], BA, BS; Margaret J Klein[2], MS

[1]Departments of Anesthesiology, Pediatrics, and Psychiatry & Behavioral Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

[2]Department of Anesthesiology Critical Care Medicine, The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, CA, United States

[3]The Saban Research Institute at Children's Hospital Los Angeles, Los Angeles, CA, United States

[4]Department of Anesthesiology Critical Care Medicine, The Biobehavioral Pain Lab, The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Jeffrey I Gold, PhD
Department of Anesthesiology Critical Care Medicine
The Saban Research Institute
Children's Hospital Los Angeles
4650 Sunset Boulevard MS#12
Los Angeles, CA, 90027
United States
Phone: 1 3233616341
Fax: 1 3233611022
Email: jgold@chla.usc.edu

Abstract

Background: Virtual reality (VR) is a well-researched digital intervention that has been used for managing acute pain and anxiety in pediatric patients undergoing various medical procedures. This study focuses on investigating the role of unique patient characteristics and VR immersion level on the effectiveness of VR for managing pediatric pain and anxiety during venipuncture.

Objective: The purpose of this study is to determine how specific patient characteristics and level of immersion during a VR intervention impact anxiety and pain levels for pediatric patients undergoing venipuncture procedures.

Methods: This study is a secondary data analysis of 2 combined, previously published randomized control trials on 252 pediatric patients aged 10-21 years observed at Children's Hospital Los Angeles from April 12, 2017, to July 24, 2019. One randomized clinical trial was conducted in 3 clinical environments examining peripheral intravenous catheter placement (radiology and an infusion center) and blood draw (phlebotomy). Conditional process analysis was used to conduct moderation and mediation analyses to assess the impact of immersion level during the VR intervention.

Results: Significant moderation was found between the level of immersion and anxiety sensitivity when predicting postprocedural anxiety (P=.01). Patients exhibiting the highest anxiety sensitivity within the standard of care yielded a 1.9 (95% CI 0.9-2.8; P<.001)-point elevation in postprocedural anxiety relative to individuals with high immersion levels. No other significant factors were found to mediate or moderate the effect of immersion on either postprocedural anxiety or pain.

Conclusions: VR is most effective for patients with higher anxiety sensitivity who report feeling highly immersed. Age, location of the procedure, and gender of the patient were not found to significantly impact VR's success in managing levels of postprocedural pain or anxiety, suggesting that immersive VR may be a beneficial intervention for a broad pediatric population.

Trial Registration: ClinicalTrials.gov NCT04268901; https://clinicaltrials.gov/study/NCT04268901

(J Med Internet Res 2024;26:e53196) doi:10.2196/53196

XSL•FO
RenderX

KEYWORDS

pediatrics; virtual reality; VR; immersion; anxiety; pain management; routine medical procedures; venipuncture; secondary data analysis; mediation; moderation; pain; acute pain; pediatric pain; anxiety sensitivity; pain management

Introduction

The subjective experience of pain results from complex interactions among biological, psychological, and social factors and is largely informed by early life experiences [1]. Thus, what a child remembers about initial painful experiences is a strong predictor of their response to subsequent painful experiences [2-5]. Anxiety has been shown to moderate children's memory for procedural pain, increasing the likelihood of remembering more pain than they initially reported [6]. In addition to the unique interplay of pain and anxiety associated with routine medical procedures, the literature reflects that pain and anxiety associated with routine medical care can lead to adverse consequences that can affect lifelong health, such as a negative impact on one's perception of health care, attempts to escape the distressing medical procedure, poor recovery outcomes, avoidance of preventative health care, and the risk for medical trauma [7-9].

Decreasing pain and anxiety during pediatric medical care is critical to ensuring optimal health and health care experiences for individuals across the lifespan. Best practice guidelines for the treatment of procedural pain and anxiety in pediatric populations have historically been cited as a combination of pharmacological and nonpharmacological interventions [10-13]. While pharmacological interventions (eg, sedatives and opioids) have become increasingly common, these analgesics have been linked to higher mortality risk and longer hospital admissions [14]. Additionally, these medication interventions have been associated with high tolerance, dependence, and unfavorable side effects [15,16]. In fact, when a study compared the efficacy of virtual reality (VR) and opioid therapy as pain management tools during thermal pain stimulation, results indicated that there was no significant difference in pain reduction between the 2 treatment groups [17]. This suggests that VR may be just as effective as routine pharmacological interventions in reducing the experience of pain and anxiety while mitigating the adverse impacts of tolerance, dependence, and other unfavorable side effects.

VR is an immersive and interactive computer-generated environment that has gained traction as a pain and anxiety management strategy within the medical field over the past few decades. In the past 5 years alone, studies have cited VR's success in reducing pain and anxiety during dental procedures [18], cancer-related treatments [19,20], and burn wound care [21]. However, many children and adolescents view venipuncture—a medical procedure in which a needle is used to draw blood from a vein—as one of the most distressing aspects of attending hospital visits [22,23] due to a combination of fear, anxiety, and pain [24]. This finding holds important considerations for children with acute or chronic illnesses, as they are exposed to needle procedures at a much higher rate than similar-aged peers who do not have medical conditions. Further, children may experience these procedures as more distressing than adults due to receptive language, expressive language, and emotional regulation limitations that impact their understanding of the routine nature of medical procedures, their ability to communicate their experiences of pain, and their ability to cope with painful experiences, respectively [25-27].

Extensive research on the use of VR has established it as a safe, feasible, effective, and efficacious intervention for reducing pain and anxiety associated with routine painful medical procedures in pediatric populations compared with standard of care (SOC) [28-31]. Furthermore, pediatric studies have indicated that children report lower pain, distress, and anxiety scores when using VR, an immersive pain management tool, –than when using iPads (Apple Inc), a passive coping tool [32-36]. However, information regarding the impact of specific unique individual characteristics that affect VR's level of efficacy has been limited to preliminary studies with outdated technologies. For example, researchers have found that a higher perception of feeling present in the VR simulation is associated with better outcomes [37-39], and recent studies that have explored the effects of age and sex characteristics across a range of pediatric specialty clinics have neither found any significant correlations with pain nor anxiety levels [40,41]. This suggests a need for further investigation into the demographic, medical, and psychological variables that may impact VR's efficacy to further appreciate "who benefits" from VR interventions and why.

This study aims to identify individual factors that impact the degree of VR's effectiveness in reducing pain and anxiety during routine painful medical venipuncture procedures (phlebotomy and peripheral intravenous catheter [PIVC] placement) in pediatric patients.

Methods

Study Design and Population

This study is a secondary data analysis of 2 combined, previously published randomized control trials on 252 pediatric patients aged 10-21 years observed at an urban pediatric academic medical center (Children's Hospital Los Angeles) from April 12, 2017, to July 24, 2019 [29,30]. Both randomized control trials implemented the same protocol. A total of 3 clinical environments were used to examine PIVC placement (radiology and an infusion center) [30] and blood draw (phlebotomy) [29]. In total, 125 dyads (patient and caregiver) were randomized to receive the VR intervention while 125 received standard of care (SOC). Patients randomized to the VR group played a multisensory (visual and auditory) VR game, BearBlast, where users traveled on a preset path through a colorful and immersive 3D environment filled with animated landscapes, buildings, and clouds, during which the user's gaze controlled the direction of a cannon fired to knock down teddy bears. The VR game is equipped with a head-tracking system, enabling the player to look around the virtual environment (VE), controlling the game with only the movement of their head

XSL•FO
RenderX

(Oculus Gear VR). Patients and participants were English or Spanish-speaking.

Measures

Caregivers provided demographic information for patients younger than 18 years of age, while patients older than 18 years completed self-reported demographic questionnaires focused on age, gender, race or ethnicity, and relevant medical history.

Pain

The Faces Pain Scale-Revised (FPS-R) [42] was used to measure patient pain before and during the PIVC procedure, and uses a horizontal series of 6 faces displaying a range of facial expressions, from no pain (0 points) to significant pain (10 points). Patients and caregivers pointed to the face that indicated the patient's level of pain. Across multiple studies, the FPS-R has been found to be both a reliable and valid measure of patient pain for children between the ages of 4 and 16 years of age [42].

Anxiety

A visual analog scale (VAS) was provided to patients and caregivers to measure patient anxiety prior to and after the PIVC procedure. The VAS provides the patient with a vertical image of a thermometer that shifts in color from yellow (accompanied by an image of a neutral face) at the bottom to dark red (accompanied by an image of a distressed face) at the top. Patients and caregivers were asked to point to the specific part of the thermometer that rated the patient's level of anxiety, with the neutral face scoring as 0 points and the distressed face scoring 10 points. Many studies have evaluated the effectiveness of a VAS and concluded that this type of scale is subject to less bias when compared to categorical scales [43].

Anxiety Sensitivity

The Childhood Anxiety Sensitivity Index (CASI) was used to measure the patient's anxiety sensitivity, and it is comprised of a 3-point Likert scale that assesses the patient's belief that their anxiety will result in a negative consequence, such as sickness, embarrassment, or loss of control. On the CASI, 1 indicates no negative consequences, 2 indicates some negative consequences, and 3 indicates a lot of negative consequences (range 18-54) [30]. Studies have found that the CASI has high internal consistency (α=.87) and good test-retest reliability in both clinical (r=0.79) and nonclinical (r=0.76) samples [44]. Patients in our study who were missing 1 of the 18 items had their total CASI sum imputed by adding the mean of their 17 complete items to their 17-item sum. Patients missing more than 1 item of the CASI were excluded from CASI analyses.

VR Immersion

Patients in the VR group completed the Gold-Rizzo Immersion and Presence (GRIP) inventory, which is a 16-item measure that asks the patients to indicate their degree of immersion in the game, with 0 indicating no immersion, 1 indicating little immersion, and 2 indicating a lot of immersion. This measure is comprised of 3 domains—sense of involvement, perceived realism of the VR game, and sense of transportation into the experience [30]. The scores on the GRIP inventory range from 0 to 32 points, with higher scores indicating higher levels of immersion. Patients in the SOC group were given a score of 0 for the GRIP immersion score since they never experienced the VR intervention. Patients in the VR group who were missing 1 or 2 of the 16-item GRIP had their missing items imputed with their nonmissing median to create a total 16-item score. These final scores were categorized into no immersion (score=0), low (score 1-19), medium (20-25), and high immersion (>25) based on our sample's distribution.

Data Analysis

Patient demographic and preprocedural characteristics were summarized using the median with IQR for continuous variables while frequency and percentage were used for categorical variables. Group differences were tested using the chi-square or Mann-Whitney U test.

Conditional process analysis was used for the moderation and mediation analyses [45]. Separate models were run for the proposed mediation and moderation effects of child characteristics (age, gender, procedure location, and anxiety sensitivity) and level of immersion on the outcomes of postprocedural anxiety VAS and pain FPS-R. All models were controlled for preprocedural pain or anxiety.

When pairwise comparisons were analyzed for significant differences, the Bonferroni multiple comparison adjustment was applied and simultaneous 95% CIs were presented. All P values were assessed at the α level of .05. Statistical analyses and data visualization were carried out with SAS (version 9.4; SAS Institute) for Windows and SPSS Statistics (version 28.0; IBM).

Ethical Considerations

All activities and procedures were approved by the local institutional review board at Children's Hospital Los Angeles (CHLA-15-00549). All patients older than 18 years provided written informed consent. Caregivers of patients younger than 18 years provided patient assent and written informed consent. All activities and procedures were approved by the local institutional review board at Children's Hospital Los Angeles.

Results

A total of 250 patients with complete data were included in the current secondary data analysis. Of the participants, 46% (n=115) were female with a median age of 15 (IQR 13-17.3) years and 51.2% (n=128) were Hispanic or Latinx. A total of 56% (n=141) of the participants underwent phlebotomy procedures, while 17.6% (n=44) and 26% (n=65) were recruited from the radiology and infusion departments, respectively. There were no statistically significant differences in these demographic variables between the SOC and VR groups (all P>.05; Table 1). Preprocedural pain FPS-R, anxiety VAS scores, and anxiety sensitivity CASI scores did not differ between groups (Table 1).

XSL•FO
RenderX

Table 1. Demographics and baseline characteristics.

Demographics and baseline characteristics	Child condition		All (N=250)	*P* value[a]
	SOC[b] (n=125)	VR[c] (n=125)		
Female, n (%)	63 (50.40)	52 (41.6)	115 (46)	.16
Race or ethnicity, n (%)				.09
Hispanic or Latinx	59 (47.2)	69 (55.2)	128 (51.2)	
White or non-Hispanic	32 (25.6)	18 (14.4)	50 (20)	
Other	34 (27.2)	38 (30.4)	72 (28.8)	
Location, n (%)				.99
Radiology	22 (17.6)	22 (17.6)	44 (17.6)	
Infusion	32 (25.6)	33 (26.4)	65(26)	
Phlebotomy	71 (56.8)	70 (56)	141 (56.4)	
Age, median (IQR)	15.00 (13.0-17.0)	15.00 (13.00-17.99)	15.00 (13-17.31)	.75
Preprocedural FPS-R[d] (n=247), median (IQR)	1.00 (0.00-1.00)	1.00 (0.00-1.00)	1.00 (0.00-1.00)	.61
Preprocedural anxiety VAS[e] (n=247), median (IQR)	1.51 (0.33-3.41)	1.33 (0.28-3.41)	1.38 (0.32-3.15)	.90
CASI[f] score (n=246), median (IQR)	28.0 (24.4-32.0)	27.0 (24.0-32.0)	28.0 (24.0-32.0)	.80

[a]*P* values from the chi-square test for categorical variables and the Mann-Whitney *U* test for continuous variables.

[b]SOC: Standard of care.

[c]VR: virtual reality.

[d]FPS-R: Faces Pain Scale-Revised.

[e]VAS: visual analog scale.

[f]CASI: Childhood Anxiety Sensitivity Index.

Conceptual moderation and mediation models for both postprocedural pain and anxiety are shown in Figure 1. Among the 238 participants in the models, the CASI levels were determined using the 25th, 50th, and 75th percentile of our sample, low CASI (score=24), medium CASI (score=28), and high CASI (score=32). Significant moderation was found between the level of immersion and anxiety sensitivity when predicting postprocedural anxiety (*P*=.01). At a low level of anxiety sensitivity (CASI=24; our sample's 25th percentile), the no immersion group (SOC) had the highest adjusted mean postprocedural anxiety (2.4, 95% CI 2-2.8), but was not significantly different from the low, medium, or high immersion groups after adjusting for multiple comparisons. At the median level of anxiety sensitivity (CASI=28), significant decreases in postprocedural anxiety were found between high and medium immersion >groups versus no immersion (all *P*<.05; Table 2). These differences were sustained when anxiety sensitivity was high (75th percentile; CASI=32). No significant differences were found when comparing the low, medium, and high levels of immersion to each other (Table 2, Figure 2).

XSL•FO
RenderX

Figure 1. (A) Moderation model. (B) Mediation model.

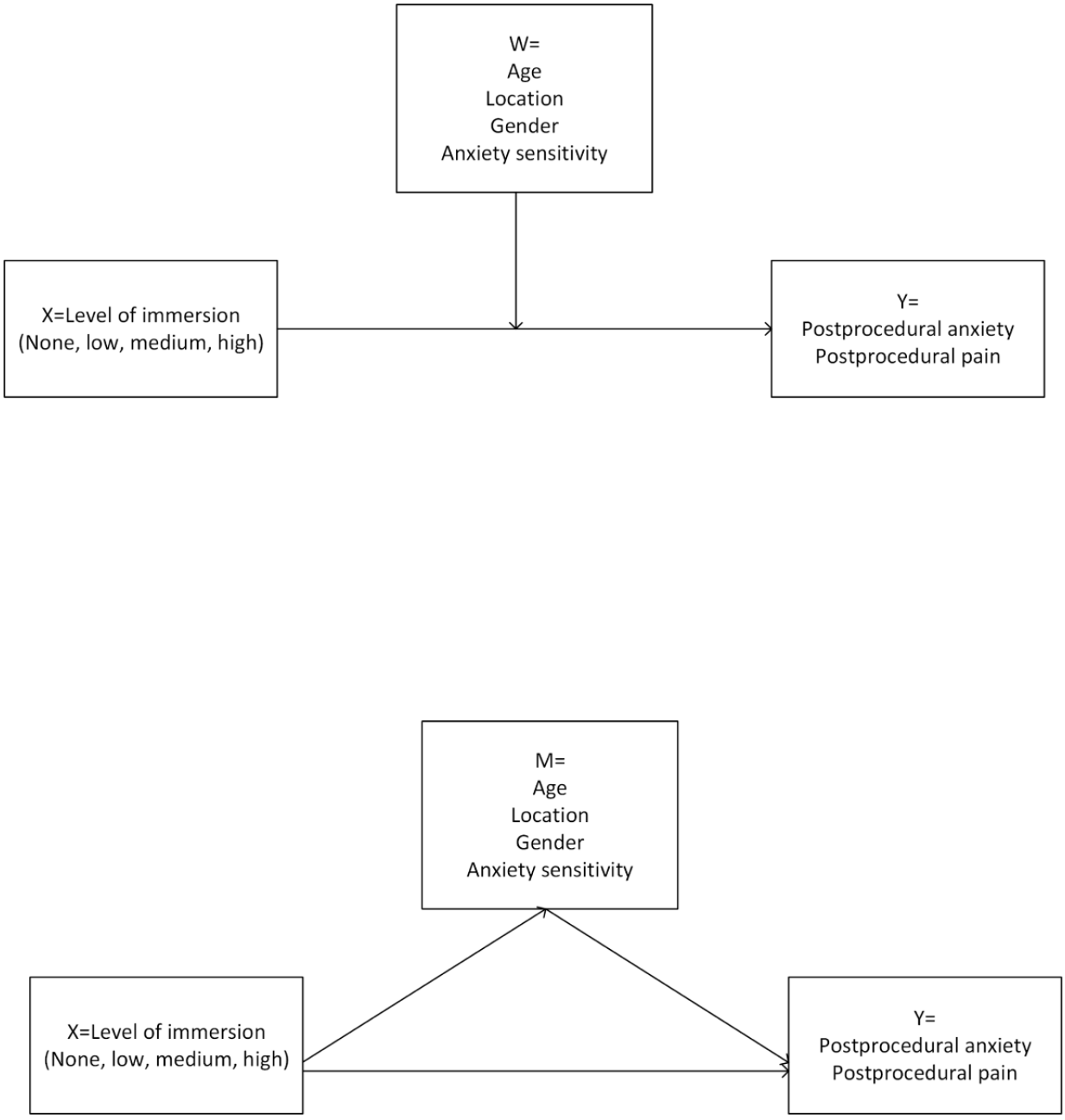

XSL•FO
RenderX

Table 2. Postanxiety moderation model between CASI[a] and level of immersion (N=238)[b].

CASI level[c]	SOC[d] (n=123)	Low immersion (n=31)	Medium immersion (n=41)	High immersion (n=43)
Low CASI (score=24)				
LS mean[e], (95% CI)	2.4 (2.0 to 2.8)	1.9 (1.1 to 2.7)	2.0 (1.3 to 2.8)	1.5 (0.9 to 2.1)
Pairwise differences[f]				
Versus low immersion	0.5 (–0.7 to 1.7)	N/A[g]	N/A	N/A
Versus medium immersion	0.4 (–0.7 to 1.5)	–0.1 (–1.6 to 1.4)	N/A	N/A
Versus high immersion	0.9 (–0.1 to 2.0)	0.4 (–1.0 to 1.8)	0.5 (–0.8 to 1.9)	N/A
Medium CASI (score=28)				
LS mean, (95% CI)	2.8 (2.5 to 3.1)	2.0 (1.4 to 2.7)	1.8 (1.3 to 2.4)	1.4 (0.9 to 2.0)
Pairwise differences				
Versus low immersion	0.8 (–0.2 to 1.7)	N/A	N/A	N/A
Versus medium immersion	1.0 (0.1 to 1.9)[h]	0.2 (–0.9 to 1.4)	N/A	N/A
Versus high immersion	1.4 (0.6 to 2.2)[i]	0.6 (–0.5 to 1.8)	0.4 (–0.6 to 1.4)	N/A
High CASI (score=32)				
LS mean, (95% CI)	3.2 (2.9 to 3.6)	2.2 (1.5 to 2.9)	1.6 (1.0 to 2.2)	1.4 (0.8 to 2.0)
Pairwise differences				
Versus low immersion	1.0 (–0.05 to 2.1)	N/A	N/A	N/A
Versus medium immersion	1.6 (0.7 to 2.5)[i]	0.6 (–0.6 to 1.8)	N/A	N/A
Versus high immersion	1.9 (0.9 to 2.8)[i]	0.9 (–0.4 to 2.1)	0.3 (–0.9 to 1.4)	N/A

[a]CASI: Childhood Anxiety Sensitivity Index.

[b]Resulting from a moderation model controlling for preprocedural anxiety held at the mean of 2.1.

[c]CASI levels were determined using the 25th, 50th, and 75th percentile of our sample.

[d]SOC: standard of care.

[e]LS mean: least squared mean (adjusted mean).

[f]Adjusted pairwise least squared differences with simultaneous 95% CI values. *P* values are adjusted for multiple comparisons using the Bonferroni adjustment.

[g]N/A: not applicable.

[h]$P<.05$.

[i]$P<.001$.

XSL•FO
RenderX

Figure 2. Fit plot with 95% CIs for predicted postprocedural anxiety by anxiety sensitivity grouped by level of immersion. Preprocedural anxiety held constant at 2.107. SOC: standard of care.

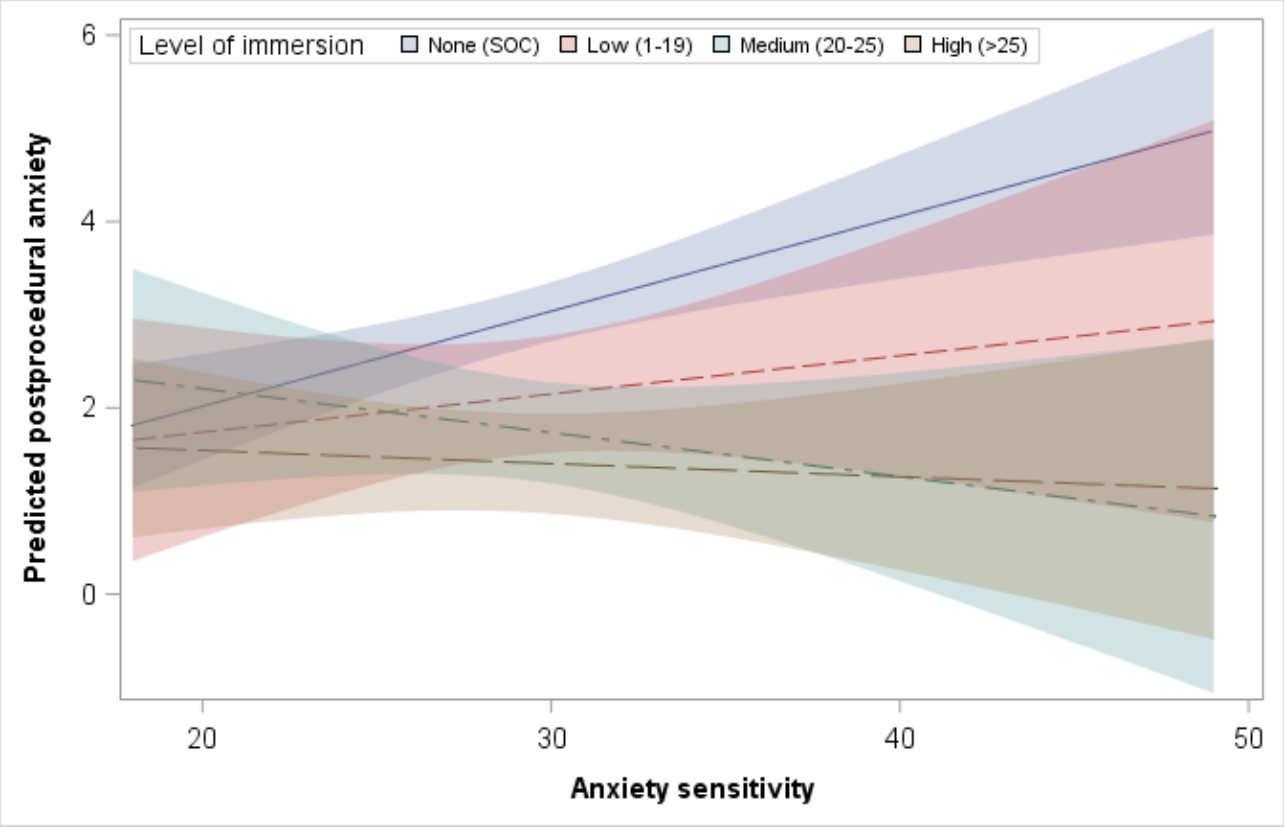

Age, location of procedure, and gender of the patient were not significant moderators of immersion on postprocedural anxiety scores and no significant mediation models were found in our sample. No significant mediation or moderation effects of child characteristics (age, gender, procedure location, and anxiety sensitivity) with level of immersion were found for the outcome of postprocedural pain FPS-R (all $P>.05$).

Discussion

Principal Results

The current findings suggest that a higher level of immersion during a VR intervention compared to no immersion decreases postprocedure anxiety among patients undergoing venipuncture procedures. For the purpose of this study, comparisons were made between each level of immersion. Overall, the immersion effect is stronger for patients who reported a higher anxiety sensitivity score, indicating that VR interventions work better to reduce anxiety in individuals who have a higher level of anxiety sensitivity. Patient characteristics such as age, gender, race or ethnicity, significant medical history, location of venipuncture, and type of procedure were not significant moderators of immersion on postprocedural scores, suggesting that VR intervention may be more universal in its application.

Comparisons With Prior Work

As previously discussed, various studies have concluded that VR reduces patient-reported pain and anxiety during pediatric venipuncture procedures more effectively when compared to the SOC [28-32]; however, this is the first study to analyze patient characteristics and immersion level as predictors of "who benefits" most optimally from the VR intervention.

The current findings support the idea that VR may lend itself to greater benefit for patients undergoing routine painful medical procedures, especially regarding anxiety management. This suggests that VEs need to be highly immersive, especially in the areas of (1) a sense of involvement, (2) perceived realism of the VR game, and (3) a sense of transportation into the experience as measured by the GRIP inventory. As the level of immersion is understood to be a critical element of the VR experience, this study begins to better understand which patients benefit most given their level of immersion. Distraction alone, as is often discussed, may not be the critical element in "why" or "how" VR works, but rather the level or degree of immersion. Previous research has alluded to the fact that deeper levels of engagement or immersion would contribute toward greater VR benefit; however, those ideas were mostly theoretical.

Future Directions

Over the years, it has been postulated that the greater the number of senses involved in the VR experience, the deeper the sense of immersion. In this study, the VR experience primarily harnessed the patient's visual and auditory senses, and patients who were highly immersed benefited the most from the intervention. With this finding, it is important to consider the question "Would VR would be more effective if more senses were involved (eg, olfactory, tactile), and whether medium and low levels of immersion would reduce pain and/or anxiety for these patients?"

XSL•FO
RenderX

Future studies investigating unique patient characteristics that drive VR effectiveness may benefit from increasing the number of senses included in their VR experience. VEs designed to engage all 6 senses (vision, hearing, touch, taste, smell, and proprioception) may transform the VR experience beyond what we currently know. Research groups currently engaging in the kinesthetic aspect of VR, olfactory, and tactile vibration could significantly increase the immersion level for all participants, thus enhancing the effects of VR in line with the current findings.

Limitations

It is important to note that for the purpose of this study, only 1 VE was evaluated (ie, BearBlast), and the current findings are limited to the mentioned procedures and a single virtual experience. However, as the field moves toward a VR pharmacy that allows for choice and customization, the immersive nature may change and be more effective, given personal choice, selection, and preference. Additionally, the participants of this study were primarily Latinx individuals; therefore, it would be beneficial to replicate this study with patients of varying ethnic and racial backgrounds.

Conclusions

As previously mentioned, factors such as age, gender, and location of venipuncture were not significant. These findings are encouraging as they suggest that the impact of VR is more about the VR experience and less about specific patient and location characteristics. Thus, the use of VR can be implemented in various settings and should be more readily available and accessible for diverse groups of pediatric patients. It was previously discussed that children will remember past painful medical procedures [2-5], which may negatively inform their future perception of potentially painful procedures [7-9]. The use of VR with a high level of immersion can decrease future negative expectations, as well as fear and anxiety about medical procedures. Future studies and interventions should focus on VE and activities with high levels of immersion and interaction in order to best support patient care. Additionally, the use of VR interventions across patient age groups and medical settings may decrease or eliminate the need for pharmacological interventions, the associated negative side-effect profiles, and the negative impact of routine painful medical procedures on patients' mental health, patients' medical experiences, and ultimately reduce the fear and anxiety that may impact medical adherence in patients with the critical need for routine and complex chronic medical care.

Acknowledgments

This project was completed in partial fulfillment of the requirements of KMA's and SA's participation in the Psychology Postdoctoral Fellowship at Children's Hospital Los Angeles, in the Leadership Education in Adolescent Health Training Program, and in the California Leadership Education in Neurodevelopmental Disabilities Interdisciplinary Training Program. This work was supported by laboratory funding from Beatrice and Paul Bennett (JIG), research funding from the Tower Cancer Foundation: Cancer Free Generation, and hardware and software donations from AppliedVR.

Conflicts of Interest

None declared.

References

1. Nordgård R, Låg T. The effects of virtual reality on procedural pain and anxiety in pediatrics: a systematic review and meta-analysis. Front Virtual Real 2021 Jul 15;2:1-33 [FREE Full text] [doi: 10.3389/frvir.2021.699383]
2. Noel M, Chambers CT, McGrath PJ, Klein RM, Stewart SH. The influence of children's pain memories on subsequent pain experience. Pain 2012 Aug;153(8):1563-1572. [doi: 10.1016/j.pain.2012.02.020] [Medline: 22560288]
3. Noel M, Rabbitts JA, Fales J, Chorney J, Palermo TM. The influence of pain memories on children's and adolescents' post-surgical pain experience: a longitudinal dyadic analysis. Health Psychol 2017 Oct;36(10):987-995 [FREE Full text] [doi: 10.1037/hea0000530] [Medline: 28726472]
4. von Baeyer CL, Marche TA, Rocha EM, Salmon K. Children's memory for pain: overview and implications for practice. J Pain 2004 Jun;5(5):241-249 [FREE Full text] [doi: 10.1016/j.jpain.2004.05.001] [Medline: 15219255]
5. Gedney JJ, Logan H. Pain related recall predicts future pain report. Pain 2006 Mar;121(1-2):69-76. [doi: 10.1016/j.pain.2005.12.005] [Medline: 16480824]
6. Rocha EM, Marche TA, von Baeyer CL. Anxiety influences children's memory for procedural pain. Pain Res Manag 2009;14(3):233-237 [FREE Full text] [doi: 10.1155/2009/535941] [Medline: 19547763]
7. Chorney JM, Kain ZN. Behavioral analysis of children's response to induction of anesthesia. Anesth Analg 2009 Nov;109(5):1434-1440 [FREE Full text] [doi: 10.1213/ane.0b013e3181b412cf] [Medline: 19713262]
8. El-Housseiny AA, Alamoudi NM, Farsi NM, El Derwi DA. Characteristics of dental fear among Arabic-speaking children: a descriptive study. BMC Oral Health 2014 Sep 22;14:118 [FREE Full text] [doi: 10.1186/1472-6831-14-118] [Medline: 25245109]

XSL•FO
RenderX

9. Kain ZN, Mayes LC, Caldwell-Andrews AA, Karas DE, McClain BC. Preoperative anxiety, postoperative pain, and behavioral recovery in young children undergoing surgery. Pediatrics 2006 Aug;118(2):651-658. [doi: 10.1542/peds.2005-2920] [Medline: 16882820]
10. Birnie KA, Noel M, Parker JA, Chambers CT, Uman LS, Kisely SR, et al. Systematic review and meta-analysis of distraction and hypnosis for needle-related pain and distress in children and adolescents. J Pediatr Psychol 2014 Sep;39(8):783-808 [FREE Full text] [doi: 10.1093/jpepsy/jsu029] [Medline: 24891439]
11. Trottier ED, Doré-Bergeron MJ, Chauvin-Kimoff L, Baerg K, Ali S. Managing pain and distress in children undergoing brief diagnostic and therapeutic procedures. Paediatr Child Health 2019 Dec;24(8):509-535 [FREE Full text] [doi: 10.1093/pch/pxz026] [Medline: 31844394]
12. Best practice guidelines for acute pain management in trauma patients. ACS Trauma Quality Programs. 2020 Nov. URL: https://www.facs.org/media/exob3dwk/acute_pain_guidelines.pdf [accessed 2024-04-10]
13. Shiferaw A, Mola S, Gashaw A, Sintayehu A. Evidence-based practical guideline for procedural pain management and sedation for burn pediatrics patients undergoing wound care procedures. Ann Med Surg (Lond) 2022 Nov;83:104756 [FREE Full text] [doi: 10.1016/j.amsu.2022.104756] [Medline: 36389198]
14. Stephens RJ, Dettmer MR, Roberts BW, Ablordeppey E, Fowler SA, Kollef MH, et al. Practice patterns and outcomes associated with early sedation depth in mechanically ventilated patients: a systematic review and meta-analysis. Crit Care Med 2018 Mar;46(3):471-479 [FREE Full text] [doi: 10.1097/CCM.0000000000002885] [Medline: 29227367]
15. Faber AW, Patterson DR, Bremer M. Repeated use of immersive virtual reality therapy to control pain during wound dressing changes in pediatric and adult burn patients. J Burn Care Res 2013;34(5):563-568 [FREE Full text] [doi: 10.1097/BCR.0b013e3182777904] [Medline: 23970314]
16. Hoffman HG, Patterson DR, Seibel E, Soltani M, Jewett-Leahy L, Sharar SR. Virtual reality pain control during burn wound debridement in the hydrotank. Clin J Pain 2008 May;24(4):299-304. [doi: 10.1097/AJP.0b013e318164d2cc] [Medline: 18427228]
17. Hoffman HG, Richards TL, Van Oostrom T, Coda BA, Jensen MP, Blough DK, et al. The analgesic effects of opioids and immersive virtual reality distraction: evidence from subjective and functional brain imaging assessments. Anesth Analg 2007 Dec;105(6):1776-83, table of contents. [doi: 10.1213/01.ane.0000270205.45146.db] [Medline: 18042882]
18. Shetty V, Suresh LR, Hegde AM. Effect of virtual reality distraction on pain and anxiety during dental treatment in 5 to 8 year old children. J Clin Pediatr Dent 2019;43(2):97-102 [FREE Full text] [doi: 10.17796/1053-4625-43.2.5] [Medline: 30730798]
19. Chow H, Hon J, Chua W, Chuan A. Effect of virtual reality therapy in reducing pain and anxiety for cancer-related medical procedures: a systematic narrative review. J Pain Symptom Manage 2021 Feb;61(2):384-394 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.08.016] [Medline: 32822755]
20. Mohammad EB, Ahmad M. Virtual reality as a distraction technique for pain and anxiety among patients with breast cancer: a randomized control trial. Palliat Support Care 2019;17(1):29-34. [doi: 10.1017/S1478951518000639] [Medline: 30198451]
21. Khadra C, Ballard A, Déry J, Paquin D, Fortin JS, Perreault I, et al. Projector-based virtual reality dome environment for procedural pain and anxiety in young children with burn injuries: a pilot study. J Pain Res 2018;11:343-353 [FREE Full text] [doi: 10.2147/JPR.S151084] [Medline: 29491717]
22. Humphrey GB, Boon CMJ, van Linden van den Heuvell GFEC, van de Wiel HBM. The occurrence of high levels of acute behavioral distress in children and adolescents undergoing routine venipunctures. Pediatrics 1992 Jul;90(1 Pt 1):87-91. [Medline: 1614786]
23. Schechter NL, Blankson V, Pachter LM, Sullivan CM, Costa L. The ouchless place: no pain, children's gain. Pediatrics 1997 Jun;99(6):890-894. [doi: 10.1542/peds.99.6.890] [Medline: 9164787]
24. Duff AJA. Incorporating psychological approaches into routine paediatric venepuncture. Arch Dis Child 2003 Oct;88(10):931-937 [FREE Full text] [doi: 10.1136/adc.88.10.931] [Medline: 14500318]
25. Cohen LL, MacLaren JE, Lim CS. Pain and pain management. In: Steele RG, Elkin TD, Roberts MC, editors. Handbook of Evidence-Based Therapies for Children and Adolescents: Bridging Science and Practice. Boston, MA: Springer US; Dec 7, 2008:281-295.
26. Slifer KJ. A Clinician's Guide to Helping Children Cope and Cooperate with Medical Care: An Applied Behavioral Approach. Baltimore, MD: The Johns Hopkins University Press; Oct 30, 2013.
27. McMurtry CM, Pillai Riddell R, Taddio A, Racine N, Asmundson GJG, Noel M, et al. Far from "just a poke": common painful needle procedures and the development of needle fear. Clin J Pain 2015 Oct;31(10 Suppl):S3-11 [FREE Full text] [doi: 10.1097/AJP.0000000000000272] [Medline: 26352920]
28. Arane K, Behboudi A, Goldman RD. Virtual reality for pain and anxiety management in children. Can Fam Physician 2017 Dec;63(12):932-934 [FREE Full text] [Medline: 29237632]
29. Gold JI, Mahrer NE. Is virtual reality ready for prime time in the medical space? A randomized control trial of pediatric virtual reality for acute procedural pain management. J Pediatr Psychol 2018 Apr 01;43(3):266-275. [doi: 10.1093/jpepsy/jsx129] [Medline: 29053848]

XSL•FO
RenderX

30. Gold JI, SooHoo M, Laikin AM, Lane AS, Klein MJ. Effect of an immersive virtual reality intervention on pain and anxiety associated with peripheral intravenous catheter placement in the pediatric setting: a randomized clinical trial. JAMA Netw Open 2021 Aug 02;4(8):e2122569 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.22569] [Medline: 34432011]
31. Wong CL, Lui MMW, Choi KC. Effects of immersive virtual reality intervention on pain and anxiety among pediatric patients undergoing venipuncture: a study protocol for a randomized controlled trial. Trials 2019 Jun 20;20(1):369 [FREE Full text] [doi: 10.1186/s13063-019-3443-z] [Medline: 31221208]
32. Özsoy F, Ulus B. Comparison of two different methods in reducing pain and fear due to dressing change in 7-10 years old children. J Pediatr Res 2022 Mar 8;9(1):66-75 [FREE Full text] [doi: 10.4274/jpr.galenos.2021.68335]
33. Hundert AS, Birnie KA, Abla O, Positano K, Cassiani C, Lloyd S, et al. A pilot randomized controlled trial of virtual reality distraction to reduce procedural pain during subcutaneous port access in children and adolescents with cancer. Clin J Pain 2021 Dec 30;38(3):189-196 [FREE Full text] [doi: 10.1097/AJP.0000000000001017] [Medline: 34974512]
34. İnangil D, Şendir M, Büyükyılmaz F. Efficacy of cartoon viewing devices during phlebotomy in children: a randomized controlled trial. J Perianesth Nurs 2020 Aug;35(4):407-412. [doi: 10.1016/j.jopan.2020.01.008] [Medline: 32360128]
35. Tennant M, Youssef GJ, McGillivray J, Clark T, McMillan L, McCarthy MC. Exploring the use of immersive virtual reality to enhance psychological well-being in pediatric oncology: a pilot randomized controlled trial. Eur J Oncol Nurs 2020 Oct;48:101804. [doi: 10.1016/j.ejon.2020.101804] [Medline: 32949941]
36. Kumari S, Bahuguna R, Garg N, Yeluri R. Immersive and non-immersive virtual reality distraction on pain perception to intraoral injections. J Clin Pediatr Dent 2021 Dec 01;45(6):389-394 [FREE Full text] [doi: 10.17796/1053-4625-45.6.5] [Medline: 34996111]
37. Miller S, Reid D. Doing play: competency, control, and expression. Cyberpsychol Behav 2003 Dec;6(6):623-632. [doi: 10.1089/109493103322725397] [Medline: 14756927]
38. Hoffman HG, Sharar SR, Coda B, Everett JJ, Ciol M, Richards T, et al. Manipulating presence influences the magnitude of virtual reality analgesia. Pain 2004 Sep;111(1-2):162-168. [doi: 10.1016/j.pain.2004.06.013] [Medline: 15327820]
39. Gutierrez-Martinez O, Gutierrez-Maldonado J, Cabas-Hoyos K, Loreto D. The illusion of presence influences VR distraction: effects on cold-pressor pain. Stud Health Technol Inform 2010;154:155-159. [doi: 10.3233/978-1-60750-561-7-155] [Medline: 20543289]
40. Piskorz J, Czub M. Effectiveness of a virtual reality intervention to minimize pediatric stress and pain intensity during venipuncture. J Spec Pediatr Nurs 2018 Jan 04;23(1):e12201. [doi: 10.1111/jspn.12201] [Medline: 29155488]
41. Atzori B, Hoffman HG, Vagnoli L, Patterson DR, Alhalabi W, Messeri A, et al. Virtual reality analgesia during venipuncture in pediatric patients with onco-hematological diseases. Front Psychol 2018;9:2508 [FREE Full text] [doi: 10.3389/fpsyg.2018.02508] [Medline: 30618938]
42. Hicks CL, von Baeyer CL, Spafford PA, van Korlaar I, Goodenough B. The Faces Pain Scale-Revised: toward a common metric in pediatric pain measurement. Pain 2001 Aug;93(2):173-183. [doi: 10.1016/S0304-3959(01)00314-1] [Medline: 11427329]
43. Klimek L, Bergmann KC, Biedermann T, Bousquet J, Hellings P, Jung K, et al. Visual analogue scales (VAS): measuring instruments for the documentation of symptoms and therapy monitoring in cases of allergic rhinitis in everyday health care. Allergo J Int 2017;26(1):16-24 [FREE Full text] [doi: 10.1007/s40629-016-0006-7] [Medline: 28217433]
44. Silverman WK, Fleisig W, Rabian B, Peterson RA. Childhood Anxiety Sensitivity Index. J Clin Child Psychol 1991 Jun;20(2):162-168. [doi: 10.1207/s15374424jccp2002_7]
45. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach, Third Edition. New York, NY: The Guilford Press; 2022.

Abbreviations

CASI: Childhood Anxiety Sensitivity Index
FPS-R: Faces Pain Scale-Revised
GRIP: Gold-Rizzo Immersion and Presence
PIVC: peripheral intravenous catheter
SOC: standard of care
VAS: visual analog scale
VE: virtual environment
VR: virtual reality

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 28.09.23; peer-reviewed by H Li, T Chang; comments to author 02.02.24; revised version received 07.03.24; accepted 20.03.24; published 01.07.24.

Please cite as:
Gold JI, Akbar KM, Avila S, Ngo NH, Klein MJ
Exploring Relations Between Unique Patient Characteristics and Virtual Reality Immersion Level on Anxiety and Pain in Patients Undergoing Venipuncture: Secondary Analysis of a Randomized Control Trial
J Med Internet Res 2024;26:e53196
URL: https://www.jmir.org/2024/1/e53196
doi:10.2196/53196
PMID:38949862

©Jeffrey I Gold, Krystal M Akbar, Sandra Avila, Nhat H Ngo, Margaret J Klein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Influence of Billboard-Based Tobacco Prevention Posters on Memorization, Attitudes, and Craving: Immersive Virtual Reality Study

Solenne Bonneterre[1,2], PhD; Oulmann Zerhouni[1,3], Prof Dr; Marilisa Boffo[4], PhD

[1]Université Paris Nanterre, Nanterre, France
[2]Université Clermont Auvergne, Clermont-Ferrand, France
[3]Université Rouen Normandie, Rouen, France
[4]Erasmus University of Rotterdam, Rotterdam, Netherlands

Corresponding Author:
Solenne Bonneterre, PhD
Université Paris Nanterre
200 Avenue de la République
Nanterre, 92000
France
Phone: 33 06161484588
Email: solenne.bonneterre@gmail.com

Abstract

Background: Health prevention campaigns often face challenges in reaching their target audience and achieving the desired impact on health behaviors. These campaigns, particularly those aimed at reducing tobacco use, require rigorous evaluation methods to assess their effectiveness.

Objective: This study aims to use immersive virtual reality (iVR) to systematically evaluate recall, attitudinal, and craving responses to antitobacco prevention messages when presented in a realistic virtual environment, thereby exploring the potential of iVR as a novel tool to improve the effectiveness of public health campaigns.

Methods: A total of 121 undergraduate students (mean age 19.6, SD 3.7 years), mostly female (n=99, 82.5%), were invited to take a guided walk in the virtual environment, where they were randomly exposed to a different ratio of prevention and general advertising posters (80/20 or 20/80) depending on the experimental condition. Participants' gaze was tracked throughout the procedure, and outcomes were assessed after the iVR exposure.

Results: Incidental exposure to antitobacco prevention and general advertising posters did not significantly alter attitudes toward tobacco. Memorization of prevention posters was unexpectedly better in the condition where advertising was more frequent (β=–6.15; P<.001), and high contrast between poster types led to a better memorization of the less frequent type. Despite a nonsignificant trend, directing attention to prevention posters slightly improved their memorization (β=.02; P=.07). In addition, the duration of exposure to prevention posters relative to advertisements negatively affected memorization of advertising posters (β=–2.30; P=.01).

Conclusions: Although this study did not find significant changes in attitudes toward tobacco after exposure to prevention campaigns using iVR, the technology does show promise as an evaluation tool. To fully evaluate the use of iVR in public health prevention strategies, future research should examine different types of content, longer exposure durations, and different contexts.

Trial Registration: Open Science Framework E3YK7; https://osf.io/e3yk7

(J Med Internet Res 2024;26:e49344) doi:10.2196/49344

KEYWORDS

tobacco; smoking; health promotion; health prevention campaigns; immersive virtual reality; incidental exposure; advertising

XSL•FO
RenderX

Introduction

Background

Tobacco consumption, including both traditional combustible cigarettes and e-cigarettes (vaping), is a major public health concern [1]. Cigarettes are responsible for various health issues such as cancer and cardiovascular diseases [2]. Public health campaigns aim to address the risks associated with both smoking and vaping, with various channels used to broadcast these messages (eg, videos and social media advertising [3]), with billboard prevention posters as a widely used strategy that can display health promotion campaigns in different locations in daily life [4]. However, the effectiveness of these campaigns may vary depending on the type of tobacco product targeted. This study explores the impact of antitobacco prevention messages in a realistic virtual environment (VE), focusing on both vaping and combustible cigarettes. We aim to understand whether there are differences in attitudes and responses to prevention messages based on the type of tobacco product.

Limitations of Actual Evaluations of Universal Prevention Campaigns

Even if largely used for universal prevention, evidence shows that preventive posters might not be as effective as intended [5]. Several meta-analyses and reviews [3,4,6-9] concluded on a lack of efficacy for most prevention campaigns, which can be explained by unstandardized evaluation, the use of arbitrary outcomes to declare a campaign effective, and sometimes deficits in the design of the campaign itself and its evaluation. Altogether, both laboratory-based and field evaluations of prevention campaigns are considered unsatisfactory [4,7,10], notably because exposure to prevention content in real life is more likely to be incidental (ie, individuals are exposed to prevention messages without being consciously or directly aware of them or without actively processing their content [5,10]) and evaluation methods used so far do not respect this condition of exposure (eg, use of forced exposure paradigms, where individuals are directly exposed to preventive content or stimulus as the primary task). Meanwhile, prevention campaigns lack rigor in their evaluation, as it is difficult to assess to what extent individuals have been exposed to the posters (ie, whether individuals saw or read them), how many times, and so on. Altogether, this points to the need for a systematic and controlled way to recreate ecological situations of exposure to prevention messages, enabling a more effective evaluation of preventive campaigns.

What Is Incidental Exposure?

Incidental exposure has been mainly explored in the brand sponsorship field [11-14], where individuals are exposed to advertising banners while watching a sporting event. Results from these studies indicate that the more individuals spent time being incidentally exposed to brands, the more positive their attitudes toward the brand, and the more they were able to recall the brand. Similar results were found when recreating incidental exposure to a brand while being exposed to advertising in video games, where advertising was present in the environment participants played in [15].

Concerning tobacco prevention, a few studies have explored the effect of incidental exposure on tobacco evaluation or memorization. Earp et al [16,17] showed that individuals exposed to “no-smoking” signs tended to activate tobacco concepts in memory, leading to more approach tendencies toward tobacco. When facing tobacco prevention in an incidental manner, individuals might process the information but in a less elaborated, biased way [18,19], potentially leading to counterproductive effects (ie, increasing the likeability of tobacco). Two studies used eye-tracking technologies to study incidental exposure to tobacco cues in real life [20,21]. While wearing a portable eye-tracking device, a sample of smokers and nonsmokers were asked to buy either a candy bar or a candy bar and a pack of cigarettes at a real-life store where tobacco advertising and tobacco products were displayed behind the cashier. For participants in the candy bar condition, where no instructions were given about tobacco, recall of either a specific part of the wall displaying information about tobacco products behind the cashier or a promotional offer for tobacco was higher. This implies that even when incidentally exposed to tobacco promotion cues, both smokers’ and nonsmokers’ memory was impacted. Similar results have been found in a virtual environment regarding the processing of tobacco products [22-24].

Altogether, evidence is lacking regarding the effects of prevention posters on memorization and attitudes toward tobacco while being incidentally exposed to tobacco prevention. Moreover, to our knowledge, no study has combined both tobacco prevention messages and general advertising in the same environment—similar to real-life daily environments—to investigate whether being exposed to general advertising will interfere with the processing of tobacco prevention campaigns. As such, this underlines the need to investigate the efficacy of prevention under realistic, incidental exposure recreating an ecological situation where individuals’ attention may be attracted to multiple stimuli (eg, prevention, advertising, surroundings, and distractors). As individuals have limited resources to allocate to their environment [4,19,25], we tested whether this scenario results in a diminished efficacy of prevention content to reach participants’ perception, therefore reducing their impact on attitudes and memorization.

Immersive Virtual Reality as a Solution to Systematically Evaluate Prevention Campaigns

Immersive virtual reality (iVR) refers to a type of digital technology including a computer-generated VE, in which individuals are immersed and can interact with as they would in real life. iVR usually uses sight as its main sensory channel by using a head-mounted display (HMD) but may also include other sensory feedback (eg, smell and sounds), deepening the immersive experience and enabling interaction with the VE the closest to real life.

The high degree of immersion and presence associated with iVR technology makes it a powerful tool for research to mimic an authentic experience in the laboratory. It is the ideal tool to create an ecological setting in which prevention posters are displayed, including the sources of information and distractions (eg, signs, advertising, and ongoing tasks) that are present in

XSL•FO
RenderX

real-life exposure to billboard preventive posters [21]. Additionally, iVR includes eye-tracking measurement through the goggles, which allows us to assess participants' attention allocated to stimuli of interest (ie, posters) via the time spent looking at them [20,25,26].

Objectives and Hypotheses

We aimed to explore to what extent being exposed to prevention posters about tobacco in a virtual urban street is effective when in competition with advertising posters displayed in the same VE. We evaluated the impact of incidental exposure to prevention messages on memorization, attitudes, and cravings toward tobacco. We used an 80/20 ratio for the 2 types of posters to expose participants to either mainly prevention or mainly advertising in order to create a clear distinction between the 2 experimental conditions and maximize internal validity (ie, maximizing chances to detect an effect if that exists).

We hypothesized that (hypothesis 1) participants' attitudes toward tobacco would be more likely impacted (ie, less positive) by prevention posters when those are predominant in the environment and (hypothesis 2) tobacco prevention posters would be better memorized when they are displayed predominantly in the environment compared to when advertisements are predominant. These hypotheses are based on the idea that more repetition of a message would enhance its effectiveness [7,10]. Tobacco cravings were also explored both as an outcome and as a potential predictor of attitudes. In the prevention condition, cravings toward tobacco were expected to be higher than in the advertising condition (hypothesis 3). Variables associated with iVR (namely, presence, immersion, and cybersickness) were explored as covariates for each hypothesis. All hypotheses, procedures, material, and analyses were preregistered.

Methods

Participants

We conducted an a priori power analysis [27] assuming a medium effect size (f^2=0.15) for a linear multiple regression (fixed model, R^2 increase with 4 key predictors), which recommended a minimum of 108 participants to ensure 90% power. The mean effect size used in our sample size calculation is based on empirical evidence from previous studies that reported similar effect sizes under analogous conditions. In Zerhouni et al [13], a medium effect size was observed in a study with a comparable intervention and demographics, which is consistent with the scope of our research. Furthermore, Zerhouni et al [14] reinforced these findings in a subsequent study that not only replicated the medium effect size but also extended the research to a similar population, providing additional validation of our assumptions.

Participants were 121 undergraduate students (mean age 19.6, SD 3.7 years; mostly women, n=99, 82.5%) from Université Paris Nanterre, recruited through the web-based research participant recruitment platform for psychology studies. To register, participants needed to complete a preregistration questionnaire containing eligibility screening questions (ie, dichotomic questions asking whether the participant had epilepsy, impaired vision, or memory impairment) and information about the smoking status (detailed below in the "Preregistration Questionnaire" section) alongside a description of the study and the consent form to participate. Participants who did not meet the eligibility criteria (eg, having a normal or corrected vision and no memory impairment) were kindly declined from further subscription to the study. Participants were asked whether they identified themselves as nonsmokers, former smokers, occasional smokers, or regular smokers and the total length of time they had smoked in their lifetime. Most participants were nonsmokers (n=93, 77.5%). Among smoking participants, 9.1% (n=11) were regular smokers, followed by 8.3% (n=10) occasional, and 5% (n=6) former smokers. The data show variability in smoking duration in years within the smoker's group ($mean_{occasional_smoker}$ 1.98, $SD_{occasional_smoker}$ 1.69 and $mean_{regular_smoker}$ 4.3, $SD_{regular_smoker}$ 4.05), ranging up to 5 years or more, suggesting considerable variation in smoking experience between participants. Despite this variability, test statistics show no significant difference in smoking duration between different smoking statuses (χ^2_2=35.2; P=.85). Scores on the Fagerstrom Test for Nicotine Dependence (FTND) [28] indicated that 66.7% (n=14) of the smokers were nondependent, 23.8% (n=5) lightly dependent, and 9.5% (n=2) moderately dependent. Concerning the use of iVR, 54.2% (n=65) of participants indicated that they already used an iVR device in the past, and only 3.1% (n=2) declared feeling sick when using it.

In total, 60 participants were randomly assigned to the prevention condition (ie, n=10, 80% tobacco prevention posters in the VE) and 61 to the advertising condition (ie, n=10, 80% general advertising posters in the VE). No baseline difference was observed between the groups concerning gender (P=.07) and FTND scores (P=.17). However, participants in the prevention condition were slightly younger (mean 18.95, SD 1.57 years) than participants in the advertising condition (mean age 20.33, SD 4.89 years; $t_{welch72.6}$=–2.09; P=.04). Age was then included as a covariate across all analyses.

Measures and Materials

iVR Device, Virtual Environment, and Posters

The iVR device was a Sensiks VR pod, in which participants can sit and wear an Oculus Rift 2. Participants moved around the VE using 2 handheld controllers. The VE was designed by the authors in collaboration with a 3D visual design company (Wonderment by Design) and depicted a Parisian suburban area displaying billboards standing on the sidewalk or wall-mounted display panels (Figure 1). Posters were randomly picked from an original pool and randomly displayed depending on the participants' experimental condition. In both conditions, 15 billboard locations were available, and only 10 of the 15 available billboard locations were randomly picked to display posters. Among the list of 20 possible posters (10 tobacco prevention posters and 10 general advertising posters), 10 were picked randomly (depending on the experimental condition) to be displayed randomly on the 10 selected billboards. The 5 remaining billboards were filled with a blank picture. The randomization of posters and billboard locations was done for

XSL•FO
RenderX

every participant. Billboard's locations in which posters were displayed were randomized among participants. The prevention condition VE included 8 prevention posters and 2 general advertising posters. The advertising condition VE contained 8 advertising posters and 2 prevention posters. Prevention posters (n=10) were real antitobacco prevention posters indicating the health risks of smoking. General advertising posters (n=10) were real advertisements for different commercial products (cars, food, beverage, fragrances, and sports gear), each represented by 2 distinct brands but with similarly looking posters (eg, the 2 car brands depicted a similar type of car, in the same shades of color).

Figure 1. Screenshot of the virtual environment depicting both general advertising and antitobacco preventive posters.

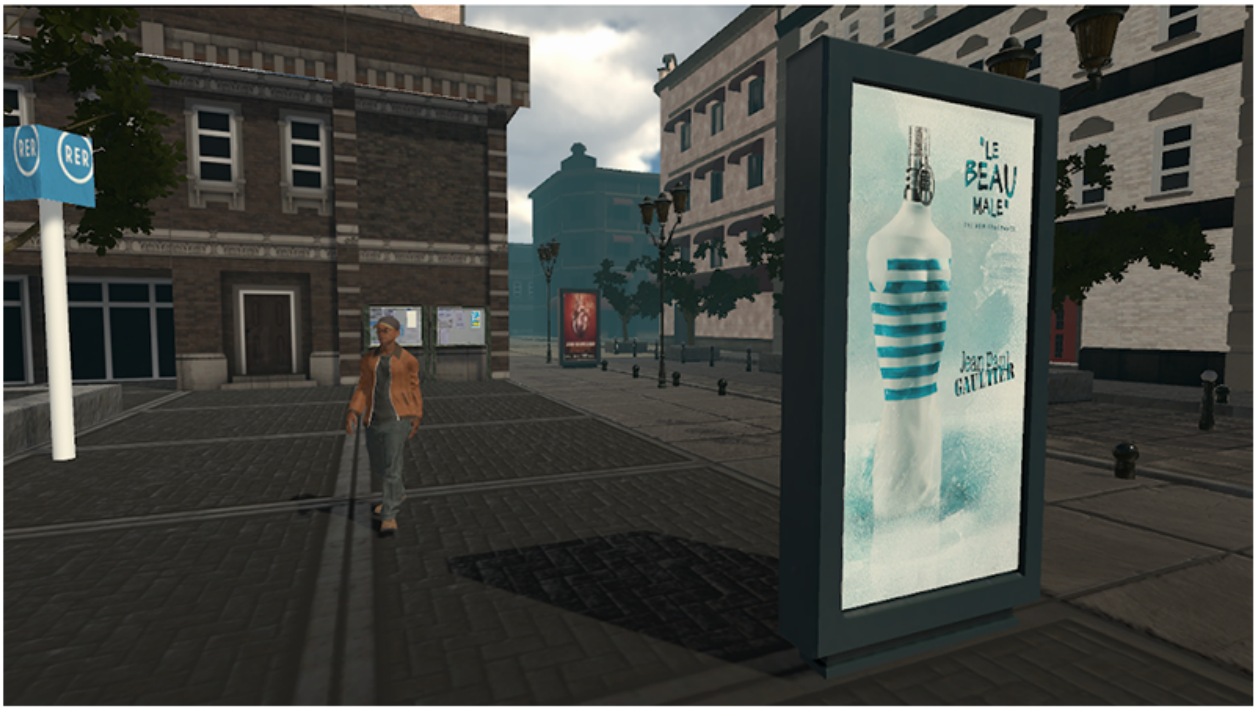

Preregistration Questionnaire

Eligibility screening included questions regarding participants' physical ability to participate and three questions on smoking status: (1) whether the participant used to be a smoker or is currently a smoker, and (2) if so, if they are a regular or occasional smoker, and (3) for how long have they smoked or used to smoke. For smokers, participants completed the FTND [28] to assess the severity of dependence on tobacco. For every item, a number of points are attributed to participants' answers (from 0 to 3 depending on the item), associated with the severity of the dependence on tobacco, and the total score was calculated to obtain 4 levels of dependence: nondependent (from 0 to 2), low dependent (3 to 4), mildly dependent (5 to 6), and highly dependent (over 7).

Attitudes Toward Tobacco

A semantic differential measure assessed attitudes [29] with 4 adjective pairs (risky or safe, not enjoyable or enjoyable, dislike or like, and bad or good) rated on a 7-point scale.

Cravings for Tobacco

The French version of the Tobacco Cravings Questionnaire [30] was used. This scale contains 12 items, rated on a Likert scale from 1=strongly disagree to 7=strongly agree, with items such as "I would do almost anything for a cigarette now." The Tobacco Cravings Questionnaire scale showed high internal consistency within this sample (α=.868; ω=0.896; mean 1.729, SD 1.017).

Explicit Recollection of Posters

We measured explicit recognition of both prevention and advertising posters in consecutive order. Participants were first exposed to a list of 30 tobacco prevention posters followed by 10 advertising posters and were asked to indicate whether they had seen them in the VE. Posters used in this study resulted from a web-based survey including 388 participants who rated antitobacco preventive posters based on their perceived persuasiveness and relevance. A score for correct recognition (ie, correctly indicating that a poster was present in the VE) and a score for incorrect recognition (ie, indicating viewing a poster that was not presented or the other way around) was calculated for prevention and advertising posters separately. In total, 2 global scores were computed (difference between correct minus incorrect recognition), 1 for each type of poster (prevention poster recognition and advertising poster recognition). The higher the recognition score, the more correctly posters were recognized (ie, the higher the memorization score).

Eye-Tracking Measures

Measures such as gaze duration, fixation count, and saccadic movement toward antitobacco posters are proxies for the depth of cognitive processing and engagement that are integral to the effectiveness of health communication strategies. By quantifying these interactions, we aim to dissect the subtleties of how visual

XSL•FO
RenderX

attention correlates with message retention and attitude change, key variables in the field of public health. The choice of these dependent variables aims at bridging the gap between attentional engagement in virtual simulations and potential behavioral outcomes in the real world. Investigating these dynamics in VR could potentially revolutionize our approach to health campaign design, tailoring it to how individuals naturally navigate and process information in their environment. To this end, participants' gaze was continuously assessed via eye-tracking technology embedded in the HMD while being in the VE. We computed 2 scores: exposure duration and directed gaze. Exposure duration refers to the total duration (in milliseconds) participants had any posters in their field of view, irrespective of the direction of their gaze (ie, the posters could be in the peripheral view for instance). Directed gaze was automatically computed by the eye-tracking system, which calculated the percentage of the area a billboard took up on the screen, transformed it to a score from 0 to 20, and summed it up every 0.2 seconds as long as the poster was still on the screen. Consequently, directed gaze is an indicator of more sustained attention to the posters. For each poster presented in the environment, an exposure duration and a directed gaze score were calculated. Hence, exposure duration is a general indicator of how long a poster was anywhere in the participant's field of view even when the gaze was not directed at it, whereas directed gaze is the average duration a poster was directly looked at. Exposure duration scores were summed up for each participant within prevention and advertisement posters separately to obtain their total exposure duration. Directed gaze scores were averaged out across prevention and advertisement posters separately. Additionally, a relative score for each eye-tracking measure was calculated by calculating the difference between the respective scores for preventive posters minus the score for advertising posters. The larger the difference, the longer participants were exposed to prevention posters (relative exposure duration) and the more they directly looked at them (relative directed gaze) compared to advertisements.

Presence

Presence was assessed using the French version of the iGroup Presence Questionnaire (IPQ) [30]. The IPQ contains 14 questions, with answers to indicate on a scale from –3=not at all to +3=completely agree, such as "Somehow I felt like the virtual world surrounded me." The IPQ showed excellent reliability in this sample (α=.835; ω=0.849; mean 0.526, SD 0.908).

Immersion

Immersion was assessed by asking 2 questions about the iVR device characteristics complementing the IPQ scale: "I was able to move into the virtual environment without thinking about how to use handheld controllers" and "I needed to or should have adjusted the helmet during the procedure." The questions were answered on a 7-point scale with 1=not at all and 7=completely agree.

Cybersickness

We used the Virtual Reality Sickness Questionnaire [31] to measure cybersickness. This 9-item questionnaire asked whether participants felt symptoms of cybersickness (eg, general discomfort and blurred vision) during the VE exposure on a 4-point scale (1=none, 2=slight, 3=moderate, or 4=severe). The cybersickness scale had good reliability (α=.736; ω=.755; mean 1.273, SD 0.330).

Procedure

The study was described as an experiment on movement and space perception in iVR. We used this cover story to limit demand bias from participants [32] and ensure that exposure to posters was incidental during the procedure. At the end of the experiment, participants were debriefed and informed about the real goals of the study.

On the experimental day, participants came to the laboratory and were randomly assigned to 1 of the 2 conditions. The procedure began with a short tutorial on how to use the iVR device, particularly the 2 handheld controllers to move around the VE. After that, participants wore the HMD and were immersed in the VE. The experimenter started by verifying that participants could use the controllers properly and stayed near the participants to guide them in the VE. Participants were asked to do the following tasks at their own pace: walking to the bus stop, reaching the end of the main street, seeking a red-headed person (ie, virtual agent), and walking the street crossing the VE from right to left, then from the top of the main street to the beginning point. This took about 5 minutes to complete. After that, the experimenter invited the participants to move freely in the VE if they wanted to. Participants then completed all the postintervention measures, starting with measures concerning iVR and the VE (ie, presence, immersion, and cybersickness), then the tobacco attitudes scale, cravings questionnaire, and the poster recognition task.

Data Analysis

We first screened covariates to evaluate the impact of the iVR on presence, immersion, and cybersickness as an intervention manipulation check. Additionally, a mixed ANOVA was conducted for each eye-tracking measure with poster type as a within-subject factor and group as a between-subject factor to verify that participants saw on average more prevention posters in the prevention condition and more advertisements in the advertising condition (ie, an interaction effect between condition and poster type). In the ANOVA, the untransformed exposure duration scores for the 2 poster types were used.

The main analysis included hierarchical multiple regression models for each main outcome (ie, attitude scores, memorization score for prevention posters, and cravings), controlling first for age, gender, FTND scores, presence, immersion, and cybersickness and successively testing the inclusion of experimental condition, the relative eye-tracking scores, and their interactions with condition as key predictors. The memorization score for advertising posters was added as an additional covariate in the model for the memorization score for prevention posters. Finally, relative exposure duration scores were centered using a *z* score to reduce multicollinearity, which was not necessary for the relative directed gaze scores, as these were already transformed scores (whereas exposure duration was a sum of raw data).

XSL•FO
RenderX

Ethical Considerations

This study falls under the category of social and human sciences research and does not aim to develop biological or medical knowledge. As such, according to the French Public Health Code (Law 2012-300 of March 5, 2012, known as the Jardé Law), this research did not require the ethical approval of a Committee for the Protection of Individuals. The study was conducted in full compliance with ethical standards and the General Data Protection Regulation. Participation was voluntary, informed consent was obtained from all participants, and the study was designed to ensure anonymity, with no personally identifiable information collected. Prior to participation, all participants were given detailed information about the study and signed an informed consent form to ensure they were fully aware of the nature of the study and their rights. Participants were given the opportunity to ask questions and were informed of their right to withdraw from the study at any time without penalty. To protect participants' privacy and confidentiality, all data collected during the study were anonymized. Personal identifiers were removed, and data were stored in a secure, password-protected database accessible only to the research team. Participants were compensated for their time with course credits. Compensation was designed to be fair and consistent with the guidelines of the university's research ethics committee.

Results

Experimental Condition Validation

Overview

Participants felt moderately present in the VE (mean 0.53, SD 0.91). Immersion was also moderate (mean 4.81, SD 1.29), where the use of the controllers was considered a bit unintuitive (mean 2.97, SD 1.76), whereas the usability of the HMD was highly satisfying (mean 6.21, SD 1.56). Participants did not report cybersickness on average (mean 1.27, SD 0.33). No group difference emerged on any of these variables (see Table 1 for exact *P* values).

The ANOVA for exposure duration did not show any main effect of condition ($F_{1,116}$=0.01; P=.92) or poster type ($F_{1,116}$=0.01; P=.93), but it showed a significant interaction between the 2 ($F_{1,116}$=1015.33; P<.001). Namely, advertising posters were viewed the longer in the advertising condition ($mean_{ads}$ 449,375, SD 11,356 and $mean_{prev}$ 107,395, SD 13,316) and the other way around for the prevention condition ($mean_{prev}$ 449,735, SD 13093 and $mean_{ads}$ 109,738, SD 11,165).

The same appeared for directed gaze, with no main effect of condition ($F_{1,116}$=0.26; P=.61) nor of poster type ($F_{1,116}$=2.52; P=.12), while their interaction was significant ($F_{1,116}$=4.73; P=.03). In the advertising condition, individuals gazed more toward advertising posters ($mean_{ads}$ 9.72, SD 1.51 and $mean_{prev}$ 8.47, SD 2.76) than prevention ones and the other way around in the prevention condition ($mean_{prev}$ 11.95, SD 2.71; $mean_{ads}$ 3.94, SD 1.48).

XSL•FO
RenderX

Table 1. Linear regression models.

Predictor variables	β (95% CI)	*t* (*df*=102)	*P* value
Outcome: attitudes			
Model 1: R^2=0.74; AIC[a]=227; BIC[b]=251; $F_{7,107}$=44.1; *P*<.001			
Presence	–.10 (–0.24 to 0.03)	–1.48	.14
Immersion	.04 (–0.06 to 0.14)	0.84	.40
Cybersickness	.01 (–0.37 to 0.38)	0.02	.99
Cravings[c]	1.07 (0.91 to 1.22)	13.87	<.001
Age	.01 (–0.02 to 0.05)	0.79	.43
Gender[c]	.38 (0.05 to 0.72)	2.30	.02
FTND[c,d]	–.22 (–0.38 to –0.06)	–2.68	.008
Model 2: R^2=0.74; AIC=236; BIC=274; $F_{12,102}$=24.8; *P*<.001 and ΔR^2=0.00; $F_{5,102}$=0.17; *P*=.97			
Presence	–.11 (–0.25 to 0.03)	–1.51	.13
Immersion	.04 (–0.07 to 0.14)	0.67	.50
Cybersickness	.02 (–0.37 to 0.42)	0.12	.91
Cravings[c]	1.07 (0.91 to 1.23)	13.50	<.001
Age	.01 (–0.02 to 0.05)	0.86	.39
Gender[c]	.36 (0.01 to 0.71)	2.05	.04
FTND[c]	–.22 (–0.39 to –0.06)	–2.69	.01
Condition (prevention vs advertising)	.00 (–0.74 to 0.74)	0.00	.99
Relative exposure duration	.09 (–0.45 to 0.63)	0.33	.74
Relative directed gaze	–.00 (–0.01 to 0.01)	–0.17	.87
Condition×relative exposure duration[e]	–.15 (–0.93 to 0.62)	–0.39	.70
Condition×relative directed gaze[f]	–.00 (–0.02 to 0.01)	–0.55	.58
Outcome: recognition score for prevention posters			
Model 1: R^2=0.15; AIC=572; BIC=594; $F_{6,108}$=3.09; *P*=.008			
Presence	–.21 (–0.82 to 0.41)	–0.67	.51
Immersion	.01 (–0.44 to 0.46)	0.04	.97
Cybersickness[c]	1.98 (0.30 to 3.66)	2.34	.02
Age[c]	.14 (–0.00 to 0.29)	1.95	.05
Gender[c]	1.92 (0.45 to 3.39)	2.59	.01
FTND	–.10 (–0.65 to 0.45)	–0.35	.73
Model 2: R^2=0.53; AIC=514; BIC=553; $F_{12,102}$=9.74; *P*<.001 and ΔR^2=0.38; $F_{6,102}$=14.1; *P*<.001			
Presence	–.23 (–0.71 to 0.25)	–0.96	.34
Immersion	–.19 (–0.55 to 0.17)	–1.07	.29
Cybersickness[c]	1.43 (0.12 to 2.75)	2.17	.03
Age	.05 (–0.07 to 0.16)	0.83	.41
Gender[c]	1.53 (0.37 to 2.69)	2.62	.01
FTND	–.11 (–0.53 to 0.31)	–0.51	.61
Recognition advertisement posters	.13 (–0.07 to 0.33)	1.33	.19

XSL•FO
RenderX

Predictor variables	β (95% CI)	*t* (*df*=102)	*P* value
Condition (prevention vs advertising)[c]	–6.15 (–8.90 to –3.40)	–4.44	<.001
Relative exposure duration	.79 (–1.07 to 2.66)	0.84	.40
Relative directed gaze	.02 (–0.00 to 0.04)	1.84	.07
Condition×relative exposure duration	.40 (–2.23 to 3.04)	0.30	.76
Condition×relative directed gaze	.01 (–0.03 to 0.05)	0.44	.67
Outcome: cravings			
Model 1: R^2=0.46; AIC=277; BIC=299; $F_{6,108}$=15.30; *P*<.001			
Presence	.08 (–0.09 to 0.25)	0.95	.34
Immersion[c]	–.13 (–0.25 to –0.00)	–1.99	.05
Cybersickness	.22 (–0.25 to 0.68)	0.92	.36
Age	–.00 (–0.04 to 0.04)	–0.12	.91
Gender	.39 (–0.01 to 0.80)	1.92	.06
FTND[c]	.67 (0.52 to 0.83)	8.72	<.001
Model 2: R^2=0.47; AIC=285; BIC=321; $F_{11,103}$=8.27; *P*<.001 and ΔR^2=0.01; $F_{5,103}$=0.38; *P*=.87			
Presence	.07 (–0.11 to 0.25)	0.79	.43
Immersion[c]	–.13 (–0.26 to –0.00)	–1.99	.05
Cybersickness	.20 (–0.29 to 0.68)	0.80	.42
Age	–.00 (–0.04 to 0.04)	–0.05	.96
Gender	.36 (–0.07 to 0.78)	1.64	.10
FTND[c]	.66 (0.51 to 0.82)	8.38	<.001
Condition (prevention vs advertising)	–.42 (–1.34 to 0.50)	–0.91	.36
Relative exposure duration	.36 (–0.31 to 1.03)	1.06	.29
Relative directed gaze	–.00 (–0.01 to 0.01)	–0.49	.63
Condition×relative exposure duration	–.34 (–1.30 to 0.62)	–0.71	.48
Condition×relative directed gaze	.01 (–0.01 to 0.02)	0.82	.41

[a]AIC: Akaike information criterion.

[b]BIC: Bayesian information criterion.

[c]Significant predictors.

[d]FTND: Fagerstrom Test for Nicotine Dependence.

[e]Exposure duration: total time in milliseconds the posters were in the field of view.

[f]Directed gaze: averaged running sum of the percentage of area billboards took up of the screen.

Hypothesis Testing

Results for all outcomes are displayed in Table 1. We did not find any significant effect concerning attitudes or cravings (see Table 1 for exact *P* values), whereas the inclusion of the key predictors explained an additional 38% of the variance in the memorization of prevention posters measure (ΔR^2=0.38; $F_{6,102}$=14.1; *P*<.001). However, contrary to what was expected, participants in the advertising condition better recognized prevention posters than participants in the prevention condition (β=–6.15; *P*<.001). Additionally, relative directed gaze positively predicted recognition of prevention posters (β=.02; *P*=.07), albeit this effect was not statistically significant. Independently of condition, the more participants directly looked at prevention posters relative to advertisements, the better their memorization.

Post Hoc Exploratory Analysis

As a post hoc analysis, we evaluated whether participants in the advertising condition would also better recognize advertisement posters rather than prevention posters by running the same hierarchical regression model with the recognition score for advertising posters as the outcome (Table 2). The inclusion of the key predictors significantly improved the model which explained an additional 49% of the outcome variance (*P*<.001). Not surprisingly, the longer participants had prevention posters in their field of view relative to advertisements (ie, relative exposure duration), the worse the recognition of advertising posters (β=–2.30; *P*=.01). However,

XSL•FO
RenderX

despite being exposed to mostly prevention posters, participants in the prevention condition had significantly higher recognition scores for advertising posters compared to participants in the advertising condition who were actually exposed to mostly advertisements (β=6.63; *P*<.001). This effect appeared to be further moderated by the total duration of exposure to prevention posters relative to advertisements, as shown by the interaction condition×relative exposure duration (β=2.38; *P*=.06), albeit only marginally statistically significant. Compared to the advertising condition, in the prevention condition, the longer participants had prevention posters anywhere in their field of view relative to advertisements, the better their recognition of advertising posters.

Table 2. Linear regression model for post hoc analysis.

Predictor variables	β (95% CI)	*t* (*df*=102)	*P* value
Outcome: recognition score for advertising posters			
Model 1: R^2=0.06; AIC[a]=580; BIC[b]=602; $F_{6,108}$=1.17; *P*=.33			
Immersion	.15 (–0.31 to 0.61)	0.64	.52
Cybersickness	–.25 (–1.99 to 1.48)	–0.29	.77
Age[c]	–.15 (–0.30 to –0.00)	–2.01	.05
Gender	–.38 (–1.89 to 1.14)	–0.49	.63
FTND[d]	–.11 (–0.68 to 0.46)	–0.38	.71
Model 2: R^2=0.55; AIC=508; BIC=547; $F_{12,102}$=10.21; *P*<.001 and ΔR^2=0.49; $F_{6,102}$=18.2; *P*<.001			
Presence	.34 (–0.12 to 0.81)	1.47	.14
Immersion[c]	.42 (0.08 to 0.76)	2.44	.02
Cybersickness	.02 (–1.28 to 1.32)	0.03	.97
Age	–.07 (–0.18 to 0.04)	–1.19	.24
Gender	.23 (–0.93 to 1.39)	0.38	.70
FTND	.00 (–0.41 to 0.42)	0.02	.99
Memorization prevention posters	.13 (–0.06 to 0.32)	1.33	0.19
Condition (prevention vs advertising)[c]	6.63 (4.01 to 9.24)	5.03	<.001
Relative exposure duration[c]	–2.30 (–4.07 to –0.54)	–2.59	.01
Relative directed gaze	–.01 (–0.03 to 0.01)	–0.80	.43
Condition×relative exposure duration[e]	2.38 (–0.14 to 4.90)	1.87	.06
Condition×relative directed gaze[f]	.00 (–0.03 to 0.04)	0.25	.80

[a]AIC: Akaike information criterion.

[b]BIC: Bayesian information criterion.

[c]Significant predictors.

[d]FTND: Fagerstrom Test for Nicotine Dependence.

[e]Exposure duration: total time in milliseconds the posters were in the field of view.

[f]Directed gaze: averaged running sum of the percentage of area billboards took up of the screen.

Discussion

Principal Findings

As the literature on incidental exposure and evaluation of preventive campaigns pointed out [3,4,6-9], the use of tobacco prevention posters might not be as effective as previously thought. This study aimed to assess the perception and recall of tobacco prevention posters in a VE containing both advertising and prevention messages. Results indicated that exposure to either prevention or advertising did not significantly alter attitudes toward tobacco, suggesting that either brief incidental exposure is insufficient to influence attitudes or that simultaneous exposure to prevention and advertising may lead to information saturation. Interestingly, in the VE dominated by tobacco prevention content, prevention posters were less memorable than when advertising dominated. Conversely, advertisements were better remembered in a prevention-saturated environment. This counterintuitive finding suggests that content that contrasts with the dominant message type of the environment is more memorable. The findings call for further research on factors that influence the memorization of health prevention messages, such as content tone and relative exposure.

XSL•FO
RenderX

First, we found that attitudes toward tobacco were not significantly different when individuals were exposed to mainly tobacco prevention or mainly advertising in their environment. This might be interpreted in two ways: (1) being incidentally exposed to billboards displaying prevention posters for a few minutes may not be long enough to impact attitudes toward tobacco or (2) being exposed to both prevention and advertising interfere with each other, potentially saturating individuals' perception or processing of information [18,19], leading to a lack of impact on attitudes toward tobacco. We think that both interpretations are valid and nonmutually exclusive, notably because the literature on the evaluation of universal prevention campaigns pointed out that preventive content is usually unseen or passively walked by [5,10], which results in a lack of efficacy [3,4,6,8]. Moreover, a few studies indicate that promotional content can be more attention-catching than prevention content [20,25], and when exposed incidentally to content, individuals might just preferentially process some information over others [18,19]. Finally, tobacco prevention messages may be perceived as threatening [25]. Therefore, some participants might have just avoided looking at or processing them. Alternatively, habituation to tobacco prevention might occur as individuals get used to being exposed to such messages, and again, they would stop looking at them or processing them [25].

Second, concerning poster memorization, the experimental condition was predictive of memorization of preventive and advertising posters but not as expected. In the VE displaying mainly tobacco prevention content, memorization scores of prevention posters were lower than when mainly advertising posters were displayed. We explored this effect also for the memorization of advertising posters, and the opposite occurred in the condition including mainly advertisements. Further, while controlling for the effect of the other variables, we found that the longer the prevention posters were present anywhere in the participant's field of view relative to advertisements, the lower the memorization of advertisements; and the more the directed gaze (ie, sustained attention) toward prevention posters relative to advertisements, the better their memorization (although this latter effect just missed the cutoff for statistical significance, $P>.10$ for all).

Altogether, the results seem to indicate that being incidentally exposed to billboard posters in the street may be enough to leave a trace in memory. This is in line with brand sponsorship studies, which found that repetitive incidental exposure to posters enhances memorization [12]. This further highlights the importance of strategically locating billboards and working on their size and look to catch the eye [11,12]. However, the pattern of results related to the memorization of the posters seems to suggest that depending on the proportion of prevention and advertising posters in the environment, the posters presented the least (ie, advertisements in the prevention condition and prevention in the advertising condition) are better recognized, contrary to the expectations. In an environment almost saturated with billboards displaying similar content—as is the case in real life where advertising is dominating our urban environment—a qualitatively and quantitatively different type of content is likely to stand out and be better retained in memory. Of note, prevention posters all highlighted the negative consequences of smoking and may have likely been perceived as very different from the general positive "tone" of advertising. Hence, it is unknown if it is the general negative tone of prevention posters, the relative quantity proportion, or both, that matters in their memorization, calling for further investigation into this effect.

An alternative explanation relates to the experimental setup of our study and the assessment of poster recognition. As participants in the advertising condition were only exposed to 2 preventive posters (and the other way around in the prevention condition), it is quite logical that preventive posters can be better retained in this condition (and vice versa) as the likelihood of having a false recognition is lower. Hence, the amount of posters displayed in the VE is an additional factor to manipulate to clarify these results.

Limitations

The first limitation of this study is that the sample is composed of students, with only a few smokers mostly nondependent on tobacco. As the focus is on universal primary prevention campaigns (ie, prevention addressed to anyone), having 27 smokers in our sample respond to the prevalence of smokers in real-life conditions [1,33]. In addition, as universal prevention is also designed for nonsmokers to prevent smoking initiation, it is interesting to assess what happens when this population is exposed to preventive poster campaigns to ensure they do not provide counterproductive effects. Only a few male participants were included, and we found that male participants had both better memorization scores for prevention posters (but not for advertising posters) and more positive attitudes toward tobacco. It might be likely that individuals with more positive attitudes toward tobacco have a bias when it comes to retaining tobacco information, such as smokers who tend to give more attention to smoking-related cues [34]. However, it might also be an effect of the low number of male participants; therefore, replication and explicit recruitment of demographic subgroups based on gender and smoking status are needed to assess the robustness of our findings. It is also important to consider the age range of our participants, which was predominantly undergraduate students with a mean age of 19.6 (SD 3.7) years. While the age cohort is relevant for the investigation of attitudes toward tobacco, it must be acknowledged that younger participants, particularly adolescents, may respond differently to prevention messages. Adolescents are generally more susceptible to advertising and may exhibit different patterns of attention and memorization. Future research should aim to replicate this study with a younger demographic. In particular, it would be beneficial to target adolescents. Adolescents represent a vulnerable population, particularly in relation to the formation of attitudes toward tobacco use. This population may be more susceptible to advertising. Therefore, it would be valuable to examine how adolescents respond to prevention messages in immersive virtual environments. Such insights could provide a deeper understanding of the effectiveness of prevention messages and inform the development of more impactful strategies.

It may be possible that the recognition task was maybe not the most relevant to fit experimental needs (ie, a large number of distractors). We suggest assessing both implicit and explicit

XSL•FO
RenderX

memorization when studying incidental exposure [35-37]. In addition, even though we found that the condition impacted memorization, this result calls for a replication including a control condition. That way, we could address whether it is the comparison of the ratio between posters or the mere fact of being exposed to posters that impacted memorization in an ecological setting.

A last limitation stands in the external validity of our results: to contrast the 2 conditions, we chose a ratio of 80/20 for the 2 poster types in each condition. However, it is rare, if not nonexistent, to have more prevention posters displayed in real-life urban environments than advertising ones. The ratio of posters is therefore not ecological, but it was a methodological choice to more effectively investigate the impact of prevention posters when displayed with general advertising (ie, maximize internal validity). This study is the first to explore this situation; its aim was to give insights on what is more likely to happen in this kind of environment in a controlled VE setting rather than recreating a real-life ratio. Replication should be conducted by also using multiple ratios (eg, 50/50, 20/80, 10/90, and 5/95) to explore the impact of the ratio while mimicking proportions closer to real life to enhance external validity. In addition, tobacco can be displayed to some extent as a form of advertising content in real life (eg, advertising in vaping stores and movie posters, as can be the case for alcohol [38]), but we chose not to include this kind of advertising in our environment, as explicit smoking advertisements are legally forbidden.

iVR offers a novel approach to evaluating the effectiveness of antitobacco campaigns by simulating the environments in which individuals can make health-related decisions. However, it is important to recognize the limitations associated with the use of this technology. For example, cybersickness is a nontrivial issue that may influence participant responses within the iVR environment and, consequently, their responses to the antitobacco messages, introducing a potential confound. In addition, current technology may not fully capture the visceral and emotional cues that influence behavior in real-world settings. This could potentially limit the extent to which findings can be extrapolated to actual settings in which tobacco-related decisions are made. Participant variability in iVR navigation skills and experience may also introduce additional sources of variability into study results. Familiarity with the technology may affect how individuals interact with the VE and respond to the prevention campaigns, thus affecting the internal and external validity of the study findings. The ecological validity of using iVR to assess campaign impact is another factor to consider. While iVR can create compelling and realistic scenarios, there is an inherent artificiality to virtual experiences. The cognitive and emotional responses elicited in these environments may not fully translate to the complexities of real-world behaviors and attitudes. Indeed, the total exposure to the VE was about 5 minutes, and outcome measures were taken immediately after. An effect on attitudes is likely to need repeated exposure to prevention messages over a longer period to eventually impact the subsequent smoking behavior [4,10]. In addition, participants were sitting during the experiment and moving within the VE using handheld controllers, which might have impacted presence and immersion, therefore also possibly lowering ecological validity.

Conclusions

To our knowledge, this study is the first to use iVR to evaluate tobacco prevention posters efficacy in an ecological environment (ie, containing both prevention and general advertising). It highlights the role of iVR in enhancing ecological validity and engaging participants in a life-like manner, providing a dynamic methodological alternative to traditional surveys for behavioral research. In addition, our findings challenge conventional health communication strategies by demonstrating that negative antitobacco messages, while more memorable, do not directly influence attitudes, a key determinant of behavior, suggesting the need for innovative approaches to the design and delivery of preventive health campaigns.

Although this study did not find significant changes in attitudes toward tobacco following exposure to prevention campaigns or advertising through iVR, the potential of iVR as an evaluation tool should not be dismissed. Our findings underscore the necessity for further research to elucidate the circumstances under which iVR may be efficacious in modifying attitudes and behaviors. Future studies should consider using a wider range of content, extending the exposure duration, and using a greater variety of contexts in order to fully assess the use of iVR in public health prevention strategies. Altogether, this study indicates that an environment containing both prevention and general advertising messages does not impact individuals' attitudes toward tobacco as some studies have pointed out [3,4,6-9] while showing a complex pattern of memorization of preventive posters relative to advertisements. Poster memorization does not only depend on the attention given to the posters (eg, gaze allocated to the posters) but on the external environmental context they are embedded in. It is therefore likely that presenting a lower ratio of prevention in an environment saturated with advertising might appear salient by contrast and therefore be better recognized than advertising posters. However, replication is needed to corroborate this result.

Acknowledgments

This study has been funded by an IresP-INCa grant. The authors would like to thank Sensiks for their invaluable support in programming and setting up the virtual reality environments. The authors would also like to thank Wonderment by Design for their assistance in creating the virtual environments. The authors confirm that no generative artificial intelligence, including ChatGPT or other language models, was used in the creation of any part of this manuscript.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study can be accessed from the Open Science Framework [39].

Authors' Contributions

All authors contributed to the study conceptualization and methodology. SB and OZ prepared study materials and experimental tasks and carried out data analysis. SB managed and executed data collection. OZ and MB supervised the project. SB drafted the first manuscript, and MB and OZ reviewed and edited the final draft.

Conflicts of Interest

None declared.

References

1. Tobacco. World Health Organization (WHO). URL: https://www.who.int/news-room/fact-sheets/detail/tobacco [accessed 2022-05-24]
2. Bonaldi C, Boussac M, Nguyen-Thanh V. Estimation du nombre de décès attribuables au tabagisme, en France de 2000 à 2015 [Estimated number of smoking-related deaths in France from 2000 to 2015]. Bull Epidémiol Hebd 2019;15:278-284. [doi: 10.13140/RG.2.2.31002.49601]
3. Chan L, O'Hara B, Phongsavan P, Bauman A, Freeman B. Review of evaluation metrics used in digital and traditional tobacco control campaigns. J Med Internet Res 2020;22(8):e17432 [FREE Full text] [doi: 10.2196/17432] [Medline: 32348272]
4. Durkin S, Brennan E, Wakefield M. Mass media campaigns to promote smoking cessation among adults: an integrative review. Tob Control 2012;21(2):127-138 [FREE Full text] [doi: 10.1136/tobaccocontrol-2011-050345] [Medline: 22345235]
5. Wakefield MA, Loken B, Hornik RC. Use of mass media campaigns to change health behaviour. Lancet 2010;376(9748):1261-1271 [FREE Full text] [doi: 10.1016/s0140-6736(10)60809-4] [Medline: 20933263]
6. Anker AE, Feeley TH, McCracken B, Lagoe CA. Measuring the effectiveness of mass-mediated health campaigns through meta-analysis. J Health Commun 2016;21(4):439-456 [FREE Full text] [doi: 10.1080/10810730.2015.1095820] [Medline: 26953782]
7. Glanz K, Bishop DB. The role of behavioral science theory in development and implementation of public health interventions. Annu Rev Public Health 2010;31(1):399-418 [FREE Full text] [doi: 10.1146/annurev.publhealth.012809.103604] [Medline: 20070207]
8. Noar SM, Palmgreen P, Zimmerman RS. Reflections on evaluating health communication campaigns. Commun Methods Meas 2009;3(1-2):105-114 [FREE Full text] [doi: 10.1080/19312450902809730]
9. Snyder L, Hamilton MA, Huedo-Medina T. Does evaluation design impact communication campaign effect size? A meta-analysis. Commun Methods Meas 2009;3(1-2):84-104 [FREE Full text] [doi: 10.1080/19312450902809722] [Medline: 31268434]
10. Atkin CK, Rice RE. Advances in public communication campaigns. In: Valdivia AN, editor. The International Encyclopedia of Media Studies. USA: Blackwell Publishing Ltd; 2012.
11. Biscaia R, Correia A, Ross S, Rosado AFB. Sponsorship effectiveness in professional sport: an examination of recall and recognition among football fans. Int J Sports Mark Sponsorship 2014;16(1):2-18 [FREE Full text] [doi: 10.1108/IJSMS-16-01-2014-B002]
12. Breuer C, Rumpf C. The viewer's reception and processing of sponsorship information in sport telecasts. J Sport Manage 2012;26:521-531 [FREE Full text] [doi: 10.1123/jsm.26.6.521]
13. Zerhouni O, Bègue L, Duke AA, Flaudias V. Dynamic exposure to alcohol advertising in a sports context influences implicit attitudes. Alcohol Clin Exp Res 2016;40(2):422-428 [FREE Full text] [doi: 10.1111/acer.12966] [Medline: 26842261]
14. Zerhouni O, Bègue L, O'Brien KS. How alcohol advertising and sponsorship works: effects through indirect measures. Drug Alcohol Rev 2019;38(4):391-398 [FREE Full text] [doi: 10.1111/dar.12929]
15. Chung S, Sparks JV. Motivated processing of peripheral advertising information in video games. Commun Res 2015;43(4):518-541 [FREE Full text] [doi: 10.1177/0093650214566623]
16. Earp BD, Dill B, Harris JL, Ackerman JM, Bargh JA. Incidental exposure to no-smoking signs primes craving for cigarettes: an ironic effect of unconscious semantic processing. Yale Rev Undergrad Res Psychol 2011;2(1):12-23 [FREE Full text] [doi: 10.1037/e525772013-002]
17. Earp BD, Dill B, Harris JL, Ackerman JM, Bargh JA. No sign of quitting: incidental exposure to no-smoking signs ironically boosts cigarette-approach tendencies in smokers. J Appl Social Pyschol 2013;43(10):2158-2162 [FREE Full text] [doi: 10.1111/jasp.12202]
18. Hommel B, Wiers RW. Towards a unitary approach to human action control. Trends Cogn Sci 2017;21(12):940-949 [FREE Full text] [doi: 10.1016/j.tics.2017.09.009] [Medline: 29150000]
19. Korteling JE, Brouwer AM, Toet A. A neural network framework for cognitive bias. Front Psychol 2018;9:1561 [FREE Full text] [doi: 10.3389/fpsyg.2018.01561]

XSL•FO
RenderX

20. Bansal-Travers M, Adkison SE, O'Connor RJ, Thrasher JF. Attention and recall of point-of-sale tobacco marketing: a mobile eye-tracking pilot study. AIMS Public Health 2016;3(1):13-24 [FREE Full text] [doi: 10.3934/publichealth.2016.1.13] [Medline: 29546142]
21. Schmälzle R, Lim S, Cho HJ, Wu J, Bente G. Examining the exposure-reception-retention link in realistic communication environments via VR and eye-tracking: the VR billboard paradigm. PLoS One 2023;18(11):e0291924 [FREE Full text] [doi: 10.1371/journal.pone.0291924]
22. Kim AE, Loomis BR, Busey AH, Farrelly MC, Willett JG, Juster HR. Influence of retail cigarette advertising, price promotions, and retailer compliance on youth smoking-related attitudes and behaviors. J Public Health Manag Pract 2013;19(6):E1-E9. [doi: 10.1097/PHH.0b013e3182980c47] [Medline: 23715218]
23. Kim AE, Nonnemaker JM, Loomis BR, Baig A, Hill E, Holloway JW, et al. Influence of tobacco displays and ads on youth: a virtual store experiment. Pediatrics 2013;131(1):e88-e95 [FREE Full text] [doi: 10.1542/peds.2012-0197] [Medline: 23209108]
24. Dutra LM, Nonnemaker J, Taylor N, Feld A, Bradfield B, Holloway J, et al. Visual Attention to Tobacco-Related Stimuli in a 3D Virtual Store. USA: RTI Press; 2005.
25. Liu J, Phua J, Krugman D, Xu L, Nowak G, Popova L. Do young adults attend to health warnings in the first IQOS advertisement in the U.S.? An eye-tracking approach. Nicotine Tob Res 2021;23(5):815-822 [FREE Full text] [doi: 10.1093/ntr/ntaa243]
26. Valsecchi M, Codispoti M. Eye tracking applied to tobacco smoking: current directions and future perspectives. J Eye Mov Res 2022;15(1):10. [doi: 10.16910/jemr.15.1.2]
27. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191 [FREE Full text] [doi: 10.3758/bf03193146]
28. Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. Br J Addict 2010;86(9):1119-1127 [FREE Full text] [doi: 10.1111/j.1360-0443.1991.tb01879.x]
29. Stark E, Borgida E, Kim A, Pickens B. Understanding public attitudes toward tobacco harm reduction: the role of attitude structure. J Appl Social Pyschol 2008;38(10):2615-2635 [FREE Full text] [doi: 10.1111/j.1559-1816.2008.00406.x]
30. Berlin I, Singleton EG, Heishman SJ. Validity of the 12-item French version of the Tobacco Craving Questionnaire in treatment-seeking smokers. Nicotine Tob Res 2010;12(5):500-507 [FREE Full text] [doi: 10.1093/ntr/ntq039] [Medline: 20335281]
31. Kim HK, Park J, Choi Y, Choe M. Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl Ergon 2018;69:66-73 [FREE Full text] [doi: 10.1016/j.apergo.2017.12.016]
32. Corneille O, Lush P. Sixty years after Orne's American Psychologist article: a conceptual framework for subjective experiences elicited by demand characteristics. Pers Soc Psychol Rev 2023;27(1):83-101 [FREE Full text] [doi: 10.31234/osf.io/jqyvx] [Medline: 35801624]
33. Management and governance matters: strengthening country capacity for measurable impact (No SEA/RC75/19). World Health Organization. 2022. URL: https://iris.who.int/bitstream/handle/10665/361249/sea-rc75-19-eng.pdf?isAllowed=y&sequence=3 [accessed 2024-06-22]
34. Baschnagel JS. Using mobile eye-tracking to assess attention to smoking cues in a naturalized environment. Addict Behav 2013;38(12):2837-2840 [FREE Full text] [doi: 10.1016/j.addbeh.2013.08.005] [Medline: 24018227]
35. Herrmann JL, Walliser B, Kacha M. Consumer consideration of sponsor brands they do not remember: taking a wider look at the memorisation effects of sponsorship. Int J Advert 2015;30(2):259-281 [FREE Full text] [doi: 10.2501/ija-30-2-259-281]
36. Shapiro S, Krishnan HS. Memory-based measures for assessing advertising effects: a comparison of explicit and implicit memory effects. J Advert 2001;30(3):1-13 [FREE Full text] [doi: 10.1080/00913367.2001.10673641]
37. Sanchez-Vives MV, Slater M. From presence to consciousness through virtual reality. Nat Rev Neurosci 2005;6(4):332-339 [FREE Full text] [doi: 10.1038/nrn1651] [Medline: 15803164]
38. Basch CH, Berger K, Basch CE. Incidental alcohol appearances in advertisements on city buses in Manhattan, New York City: a descriptive study. J Community Health 2020;45(3):550-553 [FREE Full text] [doi: 10.1007/s10900-019-00774-y]
39. Open Science Framework. URL: https://osf.io/67hng/ [accessed 2024-06-25]

Abbreviations

FTND: Fagerstrom Test for Nicotine Dependence
HMD: head-mounted display
IPQ: iGroup Presence Questionnaire
iVR: immersive virtual reality
VE: virtual environment

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 26.05.23; peer-reviewed by W Shadel, Y Yang, K Champion; comments to author 27.02.24; revised version received 22.04.24; accepted 30.05.24; published 09.07.24.

Please cite as:
Bonneterre S, Zerhouni O, Boffo M
The Influence of Billboard-Based Tobacco Prevention Posters on Memorization, Attitudes, and Craving: Immersive Virtual Reality Study
J Med Internet Res 2024;26:e49344
URL: https://www.jmir.org/2024/1/e49344
doi:10.2196/49344
PMID:38980707

©Solenne Bonneterre, Oulmann Zerhouni, Marilisa Boffo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Associated With Continuous Use of a Cancer Education Metaverse Platform: Mixed Methods Study

Sunghak Kim[1], PhD; Timothy Jung[2], PhD; Dae Kyung Sohn[3], MD, PhD; Mina Suh[4], MD, PhD; Yoon Jung Chang[1], MD, PhD

[1]National Cancer Survivorship Center, National Cancer Control Institute, National Cancer Center, Goyang, Republic of Korea

[2]Faculty of Business and Law, Manchester Metropolitan University, Manchester, United Kingdom

[3]Center for Colorectal Cancer, National Cancer Center, Goyang, Republic of Korea

[4]Division of Cancer Early Detection, National Cancer Control Institute, National Cancer Center, Goyang, Republic of Korea

Corresponding Author:
Yoon Jung Chang, MD, PhD
National Cancer Survivorship Center
National Cancer Control Institute
National Cancer Center
323 Ilsan-ro
Ilsandong-gu
Goyang, 10408
Republic of Korea
Phone: 82 10 8729 5835
Email: eunice.ncc@gmail.com

Abstract

Background: Early detection of cancer and provision of appropriate treatment can increase the cancer cure rate and reduce cancer-related deaths. Early detection requires improving the cancer screening quality of each medical institution and enhancing the capabilities of health professionals through tailored education in each field. However, during the COVID-19 pandemic, regional disparities in educational infrastructure emerged, and educational accessibility was restricted. The demand for remote cancer education services to address these issues has increased, and in this study, we considered medical metaverses as a potential means of meeting these needs. In 2022, we used Metaverse Educational Center, developed for the virtual training of health professionals, to train radiologic technologists remotely in mammography positioning.

Objective: This study aims to investigate the user experience of the Metaverse Educational Center subplatform and the factors associated with the intention for continuous use by focusing on cases of using the subplatform in a remote mammography positioning training project.

Methods: We conducted a multicenter, cross-sectional survey between July and December 2022. We performed a descriptive analysis to examine the Metaverse Educational Center user experience and a logistic regression analysis to clarify factors closely related to the intention to use the subplatform continuously. In addition, a supplementary open-ended question was used to obtain feedback from users to improve Metaverse Educational Center.

Results: Responses from 192 Korean participants (male participants: n=16, 8.3%; female participants: n=176, 91.7%) were analyzed. Most participants were satisfied with Metaverse Educational Center (178/192, 92.7%) and wanted to continue using the subplatform in the future (157/192, 81.8%). Less than half of the participants (85/192, 44.3%) had no difficulty in wearing the device. Logistic regression analysis results showed that intention for continuous use was associated with satisfaction (adjusted odds ratio 3.542, 95% CI 1.037-12.097; *P*=.04), immersion (adjusted odds ratio 2.803, 95% CI 1.201-6.539; *P*=.02), and no difficulty in wearing the device (adjusted odds ratio 2.020, 95% CI 1.004-4.062; *P*=.049). However, intention for continuous use was not associated with interest (adjusted odds ratio 0.736, 95% CI 0.303-1.789; *P*=.50) or perceived ease of use (adjusted odds ratio 1.284, 95% CI 0.614-2.685; *P*=.51). According to the qualitative feedback, Metaverse Educational Center was useful in cancer education, but the experience of wearing the device and the types and qualities of the content still need to be improved.

Conclusions: Our results demonstrate the positive user experience of Metaverse Educational Center by focusing on cases of using the subplatform in a remote mammography positioning training project. Our results also suggest that improving users' satisfaction and immersion and ensuring the lack of difficulty in wearing the device may enhance their intention for continuous use of the subplatform.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57762) doi:10.2196/57762

KEYWORDS

metaverse; virtual reality; cancer education; cancer detection; digital health

Introduction

Background

Breast cancer is one of the most common cancers affecting women. Survival rates can increase with early detection, which is linked to proper treatment [1,2]. Mammography is a widely accepted method for early detection of breast cancer that enables health professionals to detect tumors in patients with no signs or symptoms of the disease [3,4]. High-quality mammograms are essential for the successful early detection of breast cancer, and mammography positioning is one of the key elements in securing high-quality mammograms. Thus, mammography positioning training for radiologic technologists is crucial for improving the feasibility of successful early detection of breast cancer [5]. However, mammography education is challenging because of the lack of material resources [6]. One possible reason for the lack of material resources is that it is difficult to observe other people using mammography directly or to obtain realistic training resources. While other cancer screening education fields, such as endoscope cleaning and disinfection method training, are relatively easy to demonstrate externally, mammography positioning training is relatively difficult to demonstrate externally because of its association with personal privacy issues and sensitivity. In addition, the outbreak of infectious diseases such as COVID-19 and the disparity in cancer screening and diagnosis educational systems among medical institutions have led to a decline in access to excellent cancer education [7,8]. Overcoming these problems requires efficient digital health care services that enable comprehensive, systematic, and continuous remote cancer education. In particular, producing material resources for mammography positioning training using metaverse technology may help alleviate the challenges of the lack of material resources in the real world. Therefore, this study selected remote mammography positioning training in the metaverse subplatform as its research topic.

National Cancer Center (NCC) Korea's Division of Cancer Early Detection runs a national cancer screening support project aimed at increasing the cure rate and reducing cancer-related mortality by inducing early cancer detection and treatment. The division leads regionally driven quality improvement activities of national cancer screening to enhance its accuracy and reliability and to perform regular training in each region to reduce regional disparities in the quality of national cancer screening [9,10]. The division provides various remote educational programs and content for training purposes, including non–face-to-face virtual reality (VR) educational videos for mammography positioning training. However, the previous way of providing VR educational videos to each regional cancer center (RCC) required significant manual effort during the delivery process. The division recognized the need for a digital platform to facilitate content delivery and management and continuous non–face-to-face education.

The need for a digital platform for the early detection of cancer aligned well with the Dr. Meta metaverse platform's need for content variety. We developed a multipurpose metaverse digital cancer care platform, Dr. Meta, at NCC Korea in 2021, supported by the Republic of Korea's Ministry of Science and ICT [11]. The Dr. Meta metaverse platform showed potential for successful cancer control; however, according to participants' feedback, its weakness was the lack of content diversity [11]. In 2022, supported by the Republic of Korea's Ministry of Health and Welfare, we upgraded the metaverse platform to improve the Dr. Meta user experience by embedding new educational VR content in the Metaverse Educational Center subplatform and adding a new VR Healing Theater subplatform under the Dr. Meta platform [12]. Especially, the Metaverse Educational Center subplatform, which helps users overcome physical constraints (eg, time and space) in education and conveniently receive training on demand, was developed to actively respond to the demand for non–face-to-face services in the post–COVID-19 era. The subplatform is equipped with multimedia so that users can upload and interact with various medical training materials, ranging from typical documents, images, videos, presentation slides, and webpages to 3D object data and VR content created using 360° video technology. In this interactive environment, users can directly control and interact with the subplatform's interface, virtual objects, and spatial elements and be more immersed with the subplatform. Users can also look in any direction while playing the educational VR content and have immersive experiences in this subplatform [11,12]. As the previous studies show that immersion and interaction are associated with the intention to continuously use the VR services, it is anticipated that such improved immersion may positively affect the intention for continuous use of the subplatform [13,14]. However, improving the diversity of content in the subplatform was a challenge [11]. Consequently, we collaborated with the Division of Cancer Early Detection of NCC Korea in 2022 using Metaverse Educational Center to virtually educate radiologic technologists working at each RCC on mammography positioning.

Previous studies have shown that the metaverse can serve as a promising tool in cancer education [15-17]. However, to effectively use the tool, it is insufficient to merely know the positive relationship between the metaverse and cancer education. Identifying the factors that may affect the use of metaverse services is important for creating successful strategies for developing, upgrading, and operating them and maximizing their effectiveness. Numerous studies have explored possible factors that can impact the use of metaverse technology, frequently applying the technology acceptance model (TAM) to identify meaningful factors [18-20]. On the basis of the TAM, perceived usefulness and perceived ease of use of technology impact satisfaction and, subsequently, intention to use [21]. In this study, we considered TAM-based factors when developing a metaverse platform and evaluating the user experience. In the context of remote mammography positioning training for

XSL•FO
RenderX

radiologic technologists, this study goes beyond the simple usability test of Metaverse Educational Center and examines the factors and relationships associated with using the subplatform for an advanced understanding of its use.

Study Objectives

This study investigated the user experience of the Metaverse Educational Center subplatform and the factors associated with the intention for continuous use by focusing on cases of using the subplatform in a remote mammography positioning training project. We believe that Metaverse Educational Center is an outstanding digital tool for cancer education. As there are other educational VR contents being produced for home-based, hospice, non–face-to-face practice training and endoscope cleaning and disinfection method training in the Metaverse Educational Center subplatform, this study can serve as evidence of this subplatform's wide applicability across various training topics and target groups if positive outcomes are found. Furthermore, we expected that study outcomes suggest practical strategies for improving the quality of Metaverse Educational Center and facilitating its successful dissemination and implementation.

Methods

Study Design

After demonstrating the Metaverse Educational Center subplatform of the Dr. Meta metaverse platform to participants, we conducted a multicenter, cross-sectional survey to examine their experiences with the subplatform. Quantitative data were collected through Likert-scale survey questions, and qualitative data were collected through an open-ended question.

Procedure

In 2022, we disseminated Dr. Meta to 12 RCCs (ie, Ajou University Hospital, Chonnam National University Hwasun Hospital, Chungbuk National University Hospital, Chungnam National University Hospital, Gachon University Gil Medical Center, Gyeongsang National University Hospital, Jeju National University Hospital, Jeonbuk National University Hospital, Kangwon National University Hospital, Kyungpook National University Chilgok Hospital, Pusan National University Hospital, and Ulsan University Hospital). On the basis of goal of positively contributing to the efficiency of national cancer screening through VR education, we collaborated with NCC Korea's Division of Cancer Early Detection and installed the Division of Cancer Early Detection VR educational videos for mammography positioning training within our Metaverse Educational Center subplatform. After preparing for the launch of the remote mammography positioning training program, we conducted virtual sessions for radiologic technologists through the Metaverse Educational Center subplatform in cooperation with the Division of Cancer Early Detection. The survey was conducted after the completion of virtual sessions.

Study Participants

Metaverse Educational Center usability tests were conducted at 13 cancer centers (the NCC and 12 RCCs). Participants aged >19 years were recruited from each cancer center between July and December 2022. Inclusion criteria for the study were having experience in using Metaverse Educational Center and being aged >19 years. Exclusion criteria for the study were having experience in using the subplatform but being aged ≤19 years, having a severe physical or mental condition that prevented them from participating in the study, being unable to complete the self-reported survey, not consenting to the survey, or withdrawing consent during the study. The remote mammography positioning training program was designed to educate radiologic technologists, so most study participants were radiologic technologists. Doctors, nurses, and hospital administrative staff also participated in this study. Figure 1 shows images of participants using Metaverse Educational Center.

XSL•FO
RenderX

Figure 1. Pictures of participants using Metaverse Educational Center.

Ethical Considerations

The study was approved by the institutional review board of NCC Korea (IRB: NCC2022-0209). Only those who read the research information and voluntarily consented to participate in the survey were able to participate. Data used in this study were anonymous because participants were not asked for identifying information (eg, name and social security number). Users who participated in the study received an everyday item worth approximately ₩7000 (US $5) as a reward.

Measures

In the survey preparation phase, we asked platform users about their impressions of the initially prepared questionnaire. Their responses conveyed a preference to mitigate perceived burdens when participating in the survey after having professional virtual training through new metaverse technology. Therefore, we used theory-based but easy and short survey questions to facilitate participants' understanding of the questions among diverse participant groups. The survey questions were adapted from previous studies examining the TAM in the context of metaverse services [11,12,22], translated into Korean, abbreviated, and modified to suit the purpose and context of the study. In addition, we adopted the viewpoint of separating the experience of technology use into hardware and software aspects [22,23]. Researchers with expertise in communication, metaverse, and health translated survey questions from English into Korean, followed by consensus to determine the most suitable translation. Subsequently, the questionnaire was reviewed by an expert proficient in both Korean and English, who was separate from the research team, to complete the final version. First, participants were asked to provide their sociodemographic information (age, sex, and position in the hospital). Next, they were asked to assess their individual differences (2 items: "I am usually interested in using new technologies or devices" and "I tend to not experience dizziness or motion sickness easily") and user experiences (8 items: "I was generally satisfied with using this subplatform," "This subplatform made me feel interested," "This subplatform made me feel immersed," "This subplatform was easy to use," "There was no discomfort (eg, dizziness and nausea) in using this subplatform," "There was no difficulty in wearing this subplatform device (ie, head-mounted display and controllers)," "There was no difficulty in operating this subplatform device (ie, head-mounted display and controllers)," and "I want to continue using this subplatform in the future") relevant to the use of Metaverse Educational Center. Each item was scored on a 4-point Likert scale (1="strongly disagree" to 4="strongly agree"). Finally, participants were asked to provide their feedback to improve Metaverse Educational Center by answering the open-ended question ("If you have any suggestions or improvements, please write them in"). We used these 10 quantitative questions and the open-ended qualitative question to evaluate participants' experiences with Metaverse Educational Center. The questionnaire used in this study is presented in Multimedia Appendix 1.

XSL•FO
RenderX

Statistical Analysis

First, we conducted a descriptive analysis to examine how participants perceived their experiences using Metaverse Educational Center. The user experience was rated by obtaining the frequency and proportion of participants with positive opinions for each survey question. The number of participants with positive opinions after using Metaverse Educational Center was calculated by adding those who answered "agree" and "strongly agree." Next, we performed a logistic regression analysis to identify the variables associated with the intention to continue using the subplatform. To create a binary dependent variable, we categorized participants' reaction values to a survey question relevant to intention for continuous use into 2 groups ("strongly disagree"+"disagree" vs "agree"+"strongly agree"). Univariable logistic regression analyses were performed for each variable to obtain odds ratios with 95% CIs of intention for continuous use. A multiple logistic regression analysis was performed to identify factors related to this dichotomous dependent variable using only variables with statistical significance from the univariable logistic regression analyses. Finally, we qualitatively reviewed the feedback acquired from the open-ended question on the improvement of the subplatform. To identify and interpret the main themes and patterns within the feedback data with the researchers' subjectivity, we conducted a reflexive thematic analysis by following the 6 phases of analysis (familiarizing yourself with the data set; coding; generating initial themes; developing and reviewing themes; refining, defining, and naming themes; and writing up) [24-26]. We read and reread the data to become familiar with its content. After that, we captured important data content and features relevant to the user experience of the subplatform and generated them as several codes. On the basis of collated codes, we generated and refined each theme in a recursive process and finalized its name to tell the best story of the data. We developed such a story into a discussion of the advantages and improvements of the subplatform and suggested future directions for upgrading the subplatform. The research team has multidisciplinary academic backgrounds in communication, metaverse, cancer prevention, cancer screening, and cancer treatment and has various experiences conducting digital health research. This researcher reflexivity may have influenced data analysis. In addition, we applied a content analysis approach to quantify the qualitative feedback data and assist the discussion derived from the reflexive thematic analysis. Data analysis was conducted using IBM SPSS Statistics V22.0 (IBM Corp).

Results

Participant Characteristics

Overall, 196 individuals participated in the survey after using the Metaverse Educational Center subplatform, and data from 192 participants were analyzed. Among the 192 participants, 74 (38.5%), 69 (35.9%), 38 (19.8%), and 11 (5.7%) were aged 20 to 29, 30 to 39, 40 to 49, and 50 to 59 years, respectively. In addition, 16 (8.3%) participants were male, and 176 (91.7%) were female. Regarding position in the hospital, most participants (160/192, 83.3%) fell into the "Health professionals (other)" category. These results can be attributed to mammography positioning training for radiologic technologists using Metaverse Educational Center in collaboration with the Division of Cancer Early Detection of NCC Korea in 2022. Table 1 presents participants' demographic characteristics.

Table 1. Demographic information of participants (N=192).

Characteristics	Participants, n (%)[a]
Age group (y)	
20-29	74 (38.5)
30-39	69 (35.9)
40-49	38 (19.8)
50-59	11 (5.7)
Sex	
Male	16 (8.3)
Female	176 (91.7)
Position in the hospital	
Health professionals (doctors)	10 (5.2)
Health professionals (nurses)	19 (9.9)
Health professionals (other)[b]	160 (83.3)
Other[c]	3 (1.6)

[a]Percentages may not add up to 100% due to rounding.

[b]"Health professionals (other)" includes radiologic technologists, physical therapists, and pharmacists in this study.

[c]"Other" includes hospital administrative staff, educators, social workers, and researchers in this study.

XSL•FO
RenderX

Participants' Experiences Using the Metaverse Educational Center Subplatform

We conducted a quantitative descriptive analysis to examine how participants perceived their experiences using Metaverse Educational Center. Our results showed that most participants (178/192, 92.7%) were satisfied with their user experience of Metaverse Educational Center. Most participants responded that Metaverse Educational Center was an interesting (184/192, 95.8%) and immersive (174/192, 90.6%) subplatform. A total of 138 (71.9%) participants reported that the subplatform was easy to use. Most participants (157/192, 81.8%) wanted to continue using the subplatform in the future. Table 2 presents the overall positive experiences of Metaverse Educational Center's participants.

Table 2. Participants' positive experiences after using Metaverse Educational Center (N=192).

Items[a]	Participants, n (%)[b]
Individual differences	
I am usually interested in using new technologies or devices.	167 (87)
I tend to not experience dizziness or motion sickness easily.	99 (51.6)
User experiences	
I was generally satisfied with using this subplatform.	178 (92.7)
This subplatform made me feel interested.	184 (95.8)
This subplatform made me feel immersed.	174 (90.6)
This subplatform was easy to use.	138 (71.9)
There was no discomfort (eg, dizziness and nausea) in using this subplatform.	103 (53.6)
There was no difficulty in wearing this subplatform device (ie, head-mounted display and controllers).	85 (44.3)
There was no difficulty in operating this subplatform device (ie, head-mounted display and controllers).	110 (57.3)
I want to continue using this subplatform in the future.	157 (81.8)

[a]Items are based on a 4-point scale (ranging from 1="strongly disagree" to 4="strongly agree"). The number of participants who had positive opinions after using Metaverse Educational Center was calculated by adding the number of participants who answered "agree" and "strongly agree."

[b]Percentages may not add up to 100% due to rounding.

Factors Associated With Intention for Continuous Use of the Metaverse Educational Center Subplatform

We conducted a quantitative logistic regression analysis to identify the variables associated with the intention to continue using the subplatform. In univariable logistic regression analyses, a usual interest in using new technologies or devices (odds ratio 1.998, 95% CI 1.141-3.497; P=.02), satisfaction (odds ratio 8.079, 95% CI 3.284-19.875; P<.001), interest (odds ratio 3.000, 95% CI 1.570-5.731; P=.001), immersion (odds ratio 5.753, 95% CI 2.985-11.088; P<.001), perceived ease of use (odds ratio 2.973, 95% CI 1.741-5.077; P<.001), no discomfort in use (odds ratio 2.021, 95% CI 1.287-3.173; P=.002), and no difficulty in wearing the device (odds ratio 2.494, 95% CI 1.448-4.296; P=.001) were statistically different between participants not having intention for continuous use and participants having intention for continuous use. In the multiple logistic regression analysis, participants who were generally satisfied with using Metaverse Educational Center were more likely to intend to continue use the subplatform than those who were not (adjusted odds ratio 3.542, 95% CI 1.037-12.097; P=.04). When immersion increased by one unit, users were 2.803 times more likely to have intention for continuous use (adjusted odds ratio 2.803, 95% CI 1.201-6.539; P=.02). Having no difficulty in wearing the device was positively associated with intention for continuous use (adjusted odds ratio 2.020, 95% CI 1.004-4.062; P=.049). Table 3 presents the results of the univariable and multiple logistic regression analyses for each variable.

XSL•FO
RenderX

Table 3. Factors associated with intention for continuous use of Metaverse Educational Center.

Items	Intention for continuous use			
	Univariable		Multiple	
	Odds ratio (95% CI)	*P* value	Adjusted odds ratio (95% CI)	*P* value
Sociodemographics				
Age group (y)	0.976 (0.651-1.466)	.91	—[a]	—
Sex	1.559 (0.471-5.157)	.47	—	—
Position in the hospital	1.741 (0.980-3.093)	.06	—	—
Individual differences				
I am usually interested in using new technologies or devices.	1.998 (1.141-3.497)	.02	1.823 (0.949-3.502)	.07
I tend to not experience dizziness or motion sickness easily.	1.282 (0.833-1.973)	.26	—	—
User experiences				
I was generally satisfied with using this subplatform.	8.079 (3.284-19.875)	<.001	3.542 (1.037-12.097)	.04
This subplatform made me feel interested.	3.000 (1.570-5.731)	.001	0.736 (0.303-1.789)	.50
This subplatform made me feel immersed.	5.753 (2.985-11.088)	<.001	2.803 (1.201-6.539)	.02
This subplatform was easy to use.	2.973 (1.741-5.077)	<.001	1.284 (0.614-2.685)	.51
There was no discomfort (eg, dizziness and nausea) in using this subplatform.	2.021 (1.287-3.173)	.002	1.212 (0.682-2.153)	.51
There was no difficulty in wearing this subplatform device (ie, head-mounted display and controllers).	2.494 (1.448-4.296)	.001	2.020 (1.004-4.062)	.049
There was no difficulty in operating this subplatform device (ie, head-mounted display and controllers).	1.475 (0.928-2.346)	.10	—	—

[a]Not applicable.

Feedback on the Metaverse Educational Center Subplatform

We conducted a qualitative feedback review to explore participants' diverse opinions regarding the use of the subplatform. Participants answered the open-ended question "If you have any suggestions or improvements, please write them in." We broke down the 145 participants' responses at the sentence-unit level and collected 276 sentences. We then sorted these sentences into groups with homogeneous attributes and labeled them accordingly. Participants provided a wide range of feedback on the advantages and improvements of Metaverse Educational Center. Through the reflexive thematic analysis, we identified several important themes related to the subplatform's advantages (eg, "immersion and concentration," "usefulness," and "satisfaction") and understood how they are manifested in the data. We also identified various themes related to the subplatform's improvements in hardware (eg, "heaviness of the device," "difficulties in wearing the device," and "difficulties in operating the device") and software (eg, "deficiency of interactivity in the content," "deficiency of content variety," and "poor qualities of the images") aspects and understood how they relate to one another. In addition, the frequency of each theme was calculated by applying the content analysis approach to complement the outcomes of the reflexive thematic analysis. Regarding the subplatform's advantages (21 sentences), the most commonly reported label was "usefulness." Regarding the subplatform's improvements (255 sentences), the most commonly reported one was "deficiency of interactivity in the content." "Deficiency of content variety" and "heaviness of the device" were other representative improvements mentioned by participants. Table 4 presents more information about the feedback on the participants' experiences with Metaverse Educational Center.

XSL•FO
RenderX

Table 4. Feedback details of participants' experiences of Metaverse Educational Center.

Labels	Examples	Sentences, n (%)[a]
Advantages (n=21)		
No inconvenience	• "There was no inconvenience in using Metaverse Educational Center because I was well-instructed on how to wear and operate the device."	2 (9.5)
Immersion and concentration	• "This training method is likely effective because it raises immersion and concentration."	2 (9.5)
Usefulness	• "It was good to obtain in-depth knowledge through this educational metaverse platform." • "If a few more improvements are made, this subplatform can be used for educational purposes in more areas."	6 (28.6)
Ease of understanding	• "The explanations of the topics and content of the training were easy to understand."	2 (9.5)
Vividness and sense of presence	• "The educational materials were vividly delivered and made me feel like I was there."	4 (19)
Interest	• "Unlike viewing the educational materials in a fixed position on a personal computer, I could move and interact with them, so it was fun and not boring."	1 (4.8)
Satisfaction	• "Overall, I was satisfied with the experience."	4 (19)
Improvements (n=255)		
Difficulties in using the software	• "I wish this subplatform changed to be more accessible and more straightforward." • "It was difficult to operate the cancer education metaverse platform, and there were too many options that made me confused."	10 (3.9)
Poor settings of the screen focus	• "It was difficult to freely set the screen's focus, angle, and height in the direction I wanted or as if I was looking from the front." • "I was forced to move my head or body continuously while watching the screen as its perspective was not smoothly switched, and this experience was a bit uncomfortable and painful."	26 (10.2)
Poor qualities of the images	• "The qualities of images need to be improved." • "I would like to see educational materials with more stereoscopic and clearer visuals."	27 (10.6)
Dizziness, fatigue, and distraction	• "I felt dizzy and had eye strain at the end of the training." • "I could not concentrate well in the middle of watching the video."	19 (7.5)
Issues of noise	• "I could not pay attention to the training because this educational platform was vulnerable to external noise, and reversely, the platform sounds echoing outward made me realize the need for earphones or a headset."	7 (2.7)
Lengthy videos	• "The video was too long, preventing me from adequately closing my eyes or taking a break while watching."	2 (0.8)
Poor internet connections	• "I experienced freezing while playing videos or navigating the platform space."	2 (0.8)
Difficulties in wearing the device	• "Wearing the device of this subplatform was frustrating and uncomfortable for me because the device was pressing my head and nose and its fit was not seamless." • "The device of this subplatform did not fit properly and was too loose and fogged up."	24 (9.4)

XSL•FO
RenderX

Labels	Examples	Sentences, n (%)[a]
Heaviness of the device	• "The device of this subplatform was heavy." • "As Metaverse Educational Center's device was heavy, I would like to see improvements in making it lighter or adding instruments that support wearing it stably." • "Metaverse Educational Center's device was heavy and uncomfortable to wear for long periods."	33 (12.9)
Difficulties in operating the device	• "Operating the device of this subplatform was too complicated and needed to be simplified." • "Sometimes, errors occurred during the operation of this subplatform's device, hindering the smooth progress of training."	20 (7.8)
Physical constraints	• "It was uncomfortable to be restricted to the particular location and equipment requirements to experience this service."	1 (0.4)
Deficiency of content variety	• "I want to get training through this educational subplatform not only for mammography positioning but also for other topics." • "I hope different mammography positioning cases and methods for different patients will be handled on educational materials as each patient has different characteristics." • "I would like to see cancer education programs related not only to screening but also to treatment."	34 (13.3)
Deficiency of interactivity in the content	• "I want to control and interact more with the digital content of this subplatform, not just watch it." • "It would be great to have a realistic participatory simulation feature in this subplatform so users can practice mammography positioning." • "As the educational program is created based on metaverse technologies, seeing more immersive and engaging content that allows users to amplify the advantages and uniqueness of metaverse by manipulating the device would be nice."	42 (16.5)
Requests for function enhancement	• "I want this subplatform to work both online and offline, and I want its virtual environment to be closer to the real world along with containing more controlling functions."	7 (2.7)
Necessity of different training methods	• "It would be nice to have other training methods in case people have difficulties using Metaverse Educational Center, as it is different for everyone to feel motion sickness after using this educational metaverse platform."	1 (0.4)

[a]Percentages may not add up to 100% due to rounding.

Discussion

Principal Findings

This study investigated the user experience of the Metaverse Educational Center subplatform and the factors associated with intention for continuous use by focusing on cases of using the subplatform in a remote mammography positioning training project. As securing a sufficient number of participants was difficult in the first study using the Dr. Meta metaverse platform [11], we were able to examine this metaverse platform's user experience but were unable to reveal more detailed relationships among these factors. However, in this study, we secured more participants than in the first study of Dr. Meta use [11] by inviting new RCCs to our project and recruiting participants for an extended period. We also installed specific purpose-driven VR content produced for use in an actual cancer education project in cooperation with the Division of Cancer Early Detection of NCC Korea. On the basis of the increased sample size and new metaverse cancer education content, this study targeted the subplatform in an obvious context, complemented the methodology, and addressed its user experience in various ways.

Our results, focusing on cases of using Metaverse Educational Center in a remote mammography positioning training project, showed the positive user experience of the subplatform. Most users considered the subplatform satisfactory, interesting, and immersive. More than half of the users considered the subplatform easy to use. Although less than half of the users indicated positive opinions regarding the lack of difficulty in wearing this subplatform device, more than half of the users expressed positive opinions on the lack of discomfort in using the subplatform and lack of difficulty in operating its device. Finally, most users wanted to use the subplatform continuously. The positive user experience of the subplatform indicates that it can also be successfully used in another project of the Division of Cancer Early Detection (endoscope cleaning and disinfection method training). We intend to enhance the scalability of the subplatform and its use in the future by developing metaverse

XSL•FO
RenderX

content covering various cancer education topics and mounting them into the subplatform.

Our study's outcomes demonstrated factors related to the intention to continue using Metaverse Educational Center. According to the logistic regression analysis results, the associations between intention for continuous use of Metaverse Educational Center and satisfaction, immersion, and lack of difficulty in wearing the device were statistically significant. In previous studies, satisfaction, immersion, and lack of difficulty in wearing the device have been identified as factors directly or indirectly related to the intention to continuously use metaverse or other interactive media services [13,27,28]. Our results are consistent with those of previous studies. A higher level of satisfaction may increase the likelihood of intention for continuous use [22,27]. Users immersed in metaverse content may lose their self-consciousness, concentrate intensely on the content, and become satisfied with it [13,29]. Users may pursue an easy and relaxed experience of wearing the device, and, therefore, they are more likely to use the device if it is not difficult to wear continuously [28,30,31].

Conversely, the associations between the intention for continuous use of Metaverse Educational Center and interest, perceived ease of use, and lack of discomfort in use were not statistically significant. No associations of intention for continuous use with interest, perceived ease of use, or discomfort in use were unexpected because their associations (whether direct or indirect) have often been observed in previous studies [32-34]. These discrepancies may be due to different contextual characteristics. This study provided remote mammography positioning training to health professionals. Therefore, users may have had a strong motivation to learn skills, and the content may have been produced to deliver professional knowledge as a top priority. Entertainment, ease of use, and comfort may have weak relevance to the strong purposefulness of training; interest, perceived ease of use, and lack of discomfort in use may not be associated with the intention for continuous use of the educational metaverse platform. In addition, many younger individuals known to be familiar with the metaverse [35] and many radiologic technologists known to be familiar with complex medical devices [36] participated in this study. We assume that they tend to use emerging technologies to accomplish goals frequently; that is, they have high technology readiness [37]. However, contrary to the expectation of positive associations [38,39], neither age group nor position in the hospital showed a significant association with the intention for continuous use of the Metaverse Educational Center subplatform. Although younger participants may be familiar with metaverse services for purposes such as entertainment, they may not have been familiar with metaverse services for medical training. Furthermore, radiologic technologists may be familiar with the medical devices they usually handle but may be unfamiliar with metaverse devices, which can be considered relatively cutting-edge and perceived differently from other medical devices. This discussion implies that scrutinizing external variables such as the characteristics of the target period, target population group, and other situational contexts is vital when exploring factors associated with the intention for continuous use of metaverse services, including the Metaverse Educational Center subplatform. Various media theories, including the TAM, have evolved to include external variables for better explanatory power [40]. Therefore, subsequent research is needed to delve into the mixed results regarding the associations of intention for continuous use contingent on external variables to understand the reasons behind them.

Implications

Discovering the factors associated with the intention to continuously use the Metaverse Educational Center subplatform has both theoretical and practical implications. With regard to theoretical implications, knowledge of these statistically supported associations may trigger research into the mechanisms or conditions of associations. Applying mixed methods approaches directly ask participants deeper about quantitative results regarding which factors are associated, in what direction, and why it can be one of the future research directions. Such attempts to understand the patterns of association may lead to the development of a systematic theoretical framework. Regarding practical implications, information about factors related to the intention for continuous use of the subplatform can guide the development of more effective intervention strategies that encourage its continuous use. It will be feasible to save time and cost for upgrading the subplatform and promoting its user experience by prioritizing factors that are likely to affect intention for continuous use.

Furthermore, the feedback on Metaverse Educational Center–shaped discourse pointed to future directions for upgrading the subplatform. Regarding the subplatform's advantages, the total number of relevant sentences was smaller than that of sentences related to the subplatform's improvements because the question did not directly ask about the advantages. However, the usefulness of the subplatform was remarkable. As the usefulness of the Metaverse Educational Center subplatform is proven, it will be possible to increase the quantity and quality of educational content helpful in gaining knowledge with varying topics and formats and expand the range of use of the subplatform. Such upgrading strategies can facilitate dynamic virtual training and activate the use of the subplatform among various user groups. Regarding the subplatform's improvements, the deficiency of interactivity in the content, deficiency of content variety, and heaviness of the device were the main weaknesses. Poor screen focus and image quality, which lead to negative visual experiences, were also notable weaknesses. Despite upgrading the Dr. Meta metaverse platform based on the results of the first study to use the platform [11], the results of this study indicate that further quantitative and qualitative enhancements are needed for the content and devices of the subplatform. To improve the usability of the subplatform, it is crucial to address these weaknesses by securing diverse interactive metaverse content, solving the device's weight problem, and improving functions related to screen focus control and image quality.

In addition, the feedback substantially supported the results of the usability test. The outcomes of the descriptive and logistic regression analyses were similar to those of the responses to the feedback question. More than half of the users thought that the Metaverse Educational Center subplatform was satisfactory,

XSL•FO
RenderX

interesting, immersive, and easy to use. Less than half of the users believed that there was no difficulty in wearing the subplatform device. The former was addressed in the feedback as an advantage, whereas the latter was mentioned as an improvement. Factors associated with the intention for continuous use of the subplatform (ie, satisfaction, immersion, and lack of difficulty in wearing the device) also appeared in the feedback. Beyond emphasizing quantitative results, the feedback also clarifies new elements that received little attention but might be related to the user experience with the subplatform. For example, improving the audio system, reducing the runtime of content, and optimizing the operation of content would attract users' motivation to experience the subplatform.

The feedback also suggests that certain advantages and improvements may be closely related. For example, some participants found it easier to understand the educational content through immersion (eg, "The content was easily understandable due to high immersion" [immersion and concentration - ease of understanding]), and some were overall satisfied with the subplatform because the educational content was beneficial to them (eg, "I was satisfied with this subplatform because I could get helpful information" [usefulness - satisfaction]). Some participants complained of dizziness, fatigue, and distraction due to the lack of breaks, lengthy videos, and blurry image quality when receiving training through the subplatform (eg, "The educational video was too long, so it was unable to rest for a long time. Even the image quality was not sharp. I was dizzy, tired, and had decreased concentration after finishing the training" [poor qualities of the images; lengthy videos - dizziness, fatigue, and distraction]). Some also expressed that wearing the device was challenging due to its heavy weight (eg, "The device was heavy and easy to slip off, making it difficult to wear" [heaviness of the device - difficulties in wearing the device]). Some connected the difficulties in wearing the device with wearing glasses (eg, "It was hard to wear the device to the extent that I had to take off my glasses during the education session. To make matters worse, the poor quality of the educational content's image disrupted my learning" [difficulties in wearing the device; there was no specific categorized label about wearing glasses]). Such feedback can serve as indirect evidence for inferring new potential factors, their relationships, and directions. Determining whether these ideas are statistically valid could also be a topic for future research. By probing the feedback, we could reconfirm the results obtained from statistical methods and deal with more profound levels of user experience statements that are seldom obtained from quantitative approaches.

Future Perspective

This study used a mixed methods design to comprehensively understand the usability of the Metaverse Educational Center subplatform while embracing diverse user experiences. Through quantitative approaches, this study suggests that the upgrade direction of improving factors associated with the intention for continuous use of the subplatform, that is, immersion and lack of difficulty in wearing the device, may effectively enhance the user experience and lead to future successful effects. In addition, through qualitative approaches, this study identified ideas for materializing this upgrade direction.

First, enhancing users' perceived interactivity is an option to increase immersion. Studies have shown that perceived interactivity of new media, including metaverse and other interactive media, is positively associated with immersion [41-43]. In this study, although most participants thought that the subplatform was immersive, there was also considerable feedback indicating that participants wished for the educational content to be more interactive. Adding content scenarios and features that allow users to directly manipulate the 3D equipment needed for mammography or correct patients' mammography positioning in a virtual environment through the switch buttons on the subplatform device would help increase perceived interactivity. The incorporation of participatory content that induces user engagement and sensory expansion through hands-on practice in a virtual space after theoretical education would also be beneficial. By improving interactive content in which users can directly control virtual spaces and digital objects relevant to mammography positioning training, it is possible to enhance the immersion of the subplatform beyond its current level and sustainably increase intention for continuous use.

Next, reducing the device weight is an option to decrease the difficulty in wearing the device. Previous studies have asserted that the heavy weight of a head-mounted display makes it difficult to wear and serves as a barrier preventing the activation of VR service [28,31]. In this study, less than half of the participants thought there was no difficulty in wearing the subplatform device, and there was considerable feedback claiming difficulties in wearing the device. Some feedback indicated that it was uncomfortable to wear the device because of its heavy weight, allowing us to infer the relationship between difficulties in wearing the device and its heaviness. Other feedback also mentioned that the subplatform device pressed on the nose or placed strain on the neck with fatigue, which seemed to be caused indirectly by the heaviness of the device. The difficulties in wearing the device and its heaviness issues were already acknowledged in the first Dr. Meta study [11]. Some improvements had been made using accessories to lighten the device weight; however, the results of this study revealed that further improvements are needed. If the difficulties in wearing the device cannot be adequately solved by improving the head-mounted displays, accessories, or peripheral devices, replacing them with entirely new devices could also be an option. If the disadvantages of a head-mounted display, such as difficulties in wearing it [28], outweigh its advantages, such as greater immersion by blocking additional visual inputs and other external stimuli [44], it may be more efficient to mitigate the disadvantages, even at the expense of some advantages. Overall engagement may slightly decrease if the subplatform is experienced on existing devices, such as desktops, laptops, cellphones, or tablets. However, the engagement inherent in the characteristics of metaverse content will be maintained, and difficulties in wearing the device will largely disappear. Most importantly, if users are given the choice of device to use for the subplatform by themselves, they can weigh the pros and cons and select it according to their needs and preferences. As a self-tailored experience, this may increase user satisfaction [45]. As difficulties in wearing the device may impair immersion [28], it would be beneficial to develop web and app modes and compatible modes to allow users to use the subplatform on

XSL•FO
RenderX

familiar or mobile devices. Development of a new use model that can be run on different portable devices can alleviate difficulties in wearing the device, improve accessibility to the educational content, and complement the existing use model in diffusing the subplatform.

In this study, although not addressed by quantitative approaches, we obtained specific feedback related to the deficiency of content variety, following the first Dr. Meta study [11]. Participants wanted to receive not only mammography positioning training but also other mammography-related education through the Metaverse Educational Center subplatform and showed interest in cancer education on different topics; meeting these demands requires diverse scenarios, digital objects, and virtual environments. Applying artificial intelligence (AI) technology can effectively solve deficiencies in content variety. Scholars argue that embedding generative AI that creates images and 3D digital objects into the metaverse can save the time, energy, and costs associated with content development [46,47]. Moreover, using AI to collect and analyze metaverse use data can further advance service automation and personalization [48,49]. If AI combined with a metaverse captures and analyzes user experience and choice data, it can dynamically change virtual backgrounds and digital objects in real time to match user needs. Nonetheless, many hurdles remain in successfully integrating the metaverse and AI. The controversy related to copyright issues may limit the use of AI for content development in the metaverse. Legal deliberations and discussions are imperative to determine the ownership of copyrights concerning medical education content generated by AI, whether they reside in the final product or the source materials used for AI training, alongside their respective proportions. Data privacy violation is also a potential challenge for using AI to collect and analyze metaverse use data. Defining the scope of data use and establishing data protection methods are indispensable throughout the data collection, processing, and application phases [49]. It is essential to establish policies and ethics related to the use of metaverses and AI, and data security must be ensured. If these prerequisites are met, the industry will rapidly accelerate toward building data-driven metaverse systems using AI in the future. A successful combination of metaverses and AI may also appeal to participants to continuously use the subplatform.

Finally, the positive study outcomes in the user experience of the Metaverse Educational Center subplatform prove the extension of using this subplatform widely in not only various training topics and target groups but also different geographical locations. At the beginning stage of developing the Dr. Meta metaverse platform, we needed to consider regional disparities in medical resources and information and communication technologies to facilitate a multicenter cancer education environment using metaverse technology. However, we conducted research to demonstrate an example where multiple medical institutions could more actively use the metaverse by overcoming such physical constraints. We have established the Dr. Meta metaverse platform's infrastructure and maintained its setting to disseminate and implement it in local communities. We aim to construct a more robust remote cancer education network with each RCC and other hospitals in multiple regions of the Republic of Korea by using the Metaverse Educational Center subplatform. In addition, we plan to develop an English version of Dr. Meta and intend to promote collaboration with foreign institutions. We anticipate that these future research directions will contribute to enhancing the generalizability of our study findings.

Limitations

Although this study presented positive assessments of the cancer education metaverse platform and outlined prospects for its future development, several limitations exist. First, our results may be specific to a particular context, topic, or population, making it challenging to generalize the results to the overall level of education through a medical metaverse. In this study, we used the subplatform only for the remote mammography positioning training of health professionals. Hence, it is difficult to anticipate whether contextual differences affect our results. For instance, if the purpose of cancer education is not to train health professionals but to teach cancer care methods to patients or cancer prevention methods to the general public, the factors associated with intention for continuous use might differ. Demographic information, such as sex and age distribution ratios, might also vary, potentially affecting research outcomes. Further studies are needed to confirm whether the results of this study can be reproduced in the general population or other specific population groups.

Second, as this study used cross-sectional data, not all observed relationships among variables are actual causalities but associations; therefore, careful interpretation of causal assumptions is necessary. Although mixed methods can supplement explanations of the possibilities of some causal relationships, some aspects still need to be addressed. In particular, with the current research design, it is difficult to provide valid reasons for the relationships showing inconsistent results compared with other existing studies, such as those regarding interest or perceived ease of use. To explain the relationships among factors and their mechanisms concretely, an experimental study or other study using different research methods capable of providing evidence of causality should follow.

Third, because this study mainly tested the usability of the cancer education metaverse platform, its effects remain unknown. By expanding the scope of research to encompass variables related to the educational effectiveness of the subplatform, an inclusive understanding of its impacts can be achieved. However, developing a topic-focused questionnaire rather than a universal usability questionnaire is required to accomplish this goal. Efforts to develop survey methods that can overcome physical constraints and can be easily used by each cancer center and to explore a systematically validated questionnaire suitable for educational topics are required. We will consider conducting this subplatform effects study by seeking research methods specialized in a target content topic for metaverse cancer education.

Conclusions

This study introduced a nationwide project in Korea using the Metaverse Educational Center subplatform of the Dr. Meta

XSL•FO
RenderX

multidomain metaverse cancer care digital platform. The results of this study demonstrated the positive user experience of the subplatform. This study also identified variables that may influence the intention to use the subplatform continuously. We comprehensively described how to make effective plans for improving the user experience with the subplatform by considering how to enhance the evaluation of these variables. Although this study showing the subplatform's positive user experience may fall short of demonstrating the full educational effectiveness of the subplatform, it may serve as evidence to expand the use of the subplatform and a catalyst for conducting future effectiveness research. Using this study as a starting point, if future research verifies the efficacy of the subplatform, it will be possible to argue the potential of the subplatform for successful remote cancer education more significantly. Moving forward, beyond mammography positioning training for health professionals, expanding the scope of this cancer education metaverse platform to include various topics and target populations may enable the establishment of a valuable system for successfully conducting remote cancer education. When operating educational intervention programs for at-risk patient populations with limited access to medical information or difficulty in hospital visits, the subplatform may play a key role in sharing knowledge. Moreover, it may provide a communication space that offers enjoyment and psychological support to younger patients who are relatively familiar with virtual spaces and new media (eg, adolescent and young adult patients). To bolster metaverse cancer education programs and amplify their effects in the future, it is necessary to explore the ways to secure their various contents and establish systems to promote their broad use.

Acknowledgments

The authors thank Ajou University Hospital, Chonnam National University Hwasun Hospital, Chungbuk National University Hospital, Chungnam National University Hospital, Gachon University Gil Medical Center, Gyeongsang National University Hospital, Jeju National University Hospital, Jeonbuk National University Hospital, Kangwon National University Hospital, Kyungpook National University Chilgok Hospital, Pusan National University Hospital, and Ulsan University Hospital for their collaboration in developing Dr. Meta and conducting the usability tests. This research was supported by grants from the Ministry of Health and Welfare of Korea (NCC 2261080-1) and National Cancer Center Korea (NCC-23F1620-2).

Data Availability

The data sets generated or analyzed in this study are available from the corresponding author on reasonable request.

Authors' Contributions

SK contributed to the conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing of the original draft, and review and editing of the paper. TJ contributed to the supervision and review and editing of the paper. DKS contributed to the funding acquisition, investigation, and supervision. MS contributed to the resources and review and editing of the paper. YJC contributed to the conceptualization, funding acquisition, investigation, methodology, resources, supervision, project administration, writing of the original draft, and review and editing of the paper. All authors approved the final version of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The questionnaire used to evaluate participants' experiences using Metaverse Educational Center.
[PDF File (Adobe PDF File), 54 KB - jmir_v26i1e57762_app1.pdf]

References

1. Breast cancer. Cleveland Clinic. URL: https://my.clevelandclinic.org/health/diseases/3986-breast-cancer [accessed 2023-12-11]
2. Breast cancer. World Health Organization. 2024 Mar 13. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [accessed 2024-06-26]
3. Breast cancer. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/breast-cancer/index.html [accessed 2023-12-11]
4. Mammograms. National Institutes of Health National Cancer Institute. URL: https://www.cancer.gov/types/breast/mammograms-fact-sheet#how-are-screening-and-diagnostic-mammograms-different [accessed 2023-12-11]
5. Zujić PV, Božanić A, Jurković S, Šegota D, Dujmić EG, Čandrlić B, et al. The role of self-evaluation and education of radiographers involved in a breast cancer screening program at Clinical Hospital Center Rijeka. Radiography (Lond) 2021 Nov;27(4):1162-1165. [doi: 10.1016/j.radi.2021.06.007] [Medline: 34217604]

XSL•FO
RenderX

6. Strøm B, Pires Jorge JA, Richli Meystre N, Henner A, Kukkes T, Metsälä E, et al. Challenges in mammography education and training today: the perspectives of radiography teachers/mentors and students in five European countries. Radiography (Lond) 2018 Feb;24(1):41-46 [FREE Full text] [doi: 10.1016/j.radi.2017.08.008] [Medline: 29306374]
7. Frenk J, Chen LC, Chandran L, Groff EO, King R, Meleis A, et al. Challenges and opportunities for educating health professionals after the COVID-19 pandemic. Lancet 2022 Oct 29;400(10362):1539-1556 [FREE Full text] [doi: 10.1016/S0140-6736(22)02092-X] [Medline: 36522209]
8. Kenamond MC, Mourad WF, Randall ME, Kaushal A. No oncology patient left behind: challenges and solutions in rural radiation oncology. Lancet Reg Health Am 2022 Sep;13:100289 [FREE Full text] [doi: 10.1016/j.lana.2022.100289] [Medline: 35692288]
9. Han KT, Jun JK, Im JS. National cancer control plan of the Korea: current status and the fourth plan (2021-2025). J Prev Med Public Health 2023 May;56(3):205-211 [FREE Full text] [doi: 10.3961/jpmph.23.115] [Medline: 37287197]
10. Shin HY, Jang DH, Jun JK. A nationwide survey on the effectiveness of training on endoscope reprocessing within the national cancer screening program in Korea. Am J Infect Control 2021 Aug;49(8):1031-1035 [FREE Full text] [doi: 10.1016/j.ajic.2021.02.005] [Medline: 33582222]
11. Kim S, Jung T, Sohn DK, Chae Y, Kim YA, Kang SH, et al. The multidomain metaverse cancer care digital platform: development and usability study. JMIR Serious Games 2023 Nov 30;11:e46242 [FREE Full text] [doi: 10.2196/46242] [Medline: 38032697]
12. Kim S, Jung T, Sohn DK, Chang YJ. Enhancing cancer care services with the multidomain metaverse cancer care digital platform. In: Proceedings of the 8th International XR-Metaverse Conference 2023. 2023 Presented at: XR-Metaverse 2023; June 28-30, 2023; Las Vegas, NV. [doi: 10.1007/978-3-031-50559-1_21]
13. Xie T, Zheng L, Liu G, Liu L. Exploring structural relations among computer self-efficacy, perceived immersion, and intention to use virtual reality training systems. Virtual Real 2022;26(4):1725-1744 [FREE Full text] [doi: 10.1007/s10055-022-00656-0] [Medline: 35730035]
14. Barrett AJ, Pack A, Quaid ED. Understanding learners' acceptance of high-immersion virtual reality systems: insights from confirmatory and exploratory PLS-SEM analyses. Comput Educ 2021 Aug;169:104214. [doi: 10.1016/j.compedu.2021.104214]
15. Kok DL, Dushyanthen S, Peters G, Sapkaroski D, Barrett M, Sim J, et al. Virtual reality and augmented reality in radiation oncology education - a review and expert commentary. Tech Innov Patient Support Radiat Oncol 2022 Dec;24:25-31 [FREE Full text] [doi: 10.1016/j.tipsro.2022.08.007] [Medline: 36164438]
16. van der Kruk SR, Zielinski R, MacDougall H, Hughes-Barton D, Gunn KM. Virtual reality as a patient education tool in healthcare: a scoping review. Patient Educ Couns 2022 Jul;105(7):1928-1942. [doi: 10.1016/j.pec.2022.02.005] [Medline: 35168856]
17. Wang LJ, Casto B, Luh JY, Wang SJ. Virtual reality-based education for patients undergoing radiation therapy. J Cancer Educ 2022 Jun 24;37(3):694-700 [FREE Full text] [doi: 10.1007/s13187-020-01870-7] [Medline: 32970303]
18. Cabero-Almenara J, Llorente-Cejudo C, Palacios-Rodríguez A, Gallego-Pérez Ó. Degree of acceptance of virtual reality by health sciences students. Int J Environ Res Public Health 2023 Apr 18;20(8):5571 [FREE Full text] [doi: 10.3390/ijerph20085571] [Medline: 37107853]
19. Tennant M, McGillivray J, Youssef GJ, McCarthy MC, Clark TJ. Feasibility, acceptability, and clinical implementation of an immersive virtual reality intervention to address psychological well-being in children and adolescents with cancer. J Pediatr Oncol Nurs 2020;37(4):265-277. [doi: 10.1177/1043454220917859] [Medline: 32536320]
20. Wang EY, Kennedy KM, Zhang L, Qian D, Forbes T, Zuniga-Hernandez M, et al. Predicting pediatric healthcare provider use of virtual reality using a technology acceptance model. JAMIA Open 2023 Oct;6(3):ooad076 [FREE Full text] [doi: 10.1093/jamiaopen/ooad076] [Medline: 37693368]
21. Han JH, Sa HJ. Acceptance of and satisfaction with online educational classes through the technology acceptance model (TAM): the COVID-19 situation in Korea. Asia Pacific Educ Rev 2021 Sep 15;23(3):403-415. [doi: 10.1007/s12564-021-09716-7]
22. Jang HJ, Kim KH. Study on the influence of VR characteristics on user satisfaction and intention to use continuously -focusing on VR presence, user characteristics, and VR sickness-. J Korea Contents Assoc 2018;18(5):420-431 [FREE Full text] [doi: 10.5392/JKCA.2018.18.05.420]
23. Lin S, Zimmer JC, Lee V. Decoupling software from hardware in technology acceptance research. J Comput Inf Syst 2015 Dec 10;54(2):77-86. [doi: 10.1080/08874417.2014.11645688]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2022 Oct 25;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
26. Doing reflexive TA. The University of Auckland. URL: https://www.thematicanalysis.net/doing-reflexive-ta/ [accessed 2024-06-09]

XSL•FO
RenderX

27. Tran HT, Kim M. Factors influencing the continued intent to use virtual interactive platforms in Korean small- and medium-sized enterprises for remote and hybrid work. Sustainability 2023 Jun 22;15(13):9972. [doi: 10.3390/su15139972]
28. Choi J, Lee KK, Choi J. Determinants of user satisfaction with mobile VR headsets: the human factors approach by the user reviews analysis and product lab testing. Int J Content 2019;15(1):1-9. [doi: 10.5392/IJoC.2019.15.1.001]
29. Guerra-Tamez CR. The impact of immersion through virtual reality in the learning experiences of art and design students: the mediating effect of the flow experience. Educ Sci 2023 Feb 09;13(2):185. [doi: 10.3390/educsci13020185]
30. Herz M, Rauschnabel PA. Understanding the diffusion of virtual reality glasses: the role of media, fashion and technology. Technol Forecast Soc Change 2019 Jan;138:228-242. [doi: 10.1016/j.techfore.2018.09.008]
31. Yan Y, Chen K, Xie Y, Song Y, Liu Y. The effects of weight on comfort of virtual reality devices. In: Proceedings of the AHFE 2018 International Conference on Ergonomics in Design. 2018 Presented at: AHFE 2018; July 21-25, 2018; Orlando, FL. [doi: 10.1007/978-3-319-94706-8_27]
32. Choi J, Moon H, Park M. What factors influence continuous usage intention of head-mounted display-based virtual reality content?: a cross-sectional survey. Korean J Women Health Nurs 2023 Sep;29(3):208-218 [FREE Full text] [doi: 10.4069/kjwhn.2023.09.11.02] [Medline: 37813664]
33. Syed-Abdul S, Malwade S, Nursetyo AA, Sood M, Bhatia M, Barsasella D, et al. Virtual reality among the elderly: a usefulness and acceptance study from Taiwan. BMC Geriatr 2019 Aug 19;19(1):223 [FREE Full text] [doi: 10.1186/s12877-019-1218-8] [Medline: 31426766]
34. Garrido LE, Frías-Hiciano M, Moreno-Jiménez M, Cruz GN, García-Batista ZE, Guerra-Peña K, et al. Focusing on cybersickness: pervasiveness, latent trajectories, susceptibility, and effects on the virtual reality experience. Virtual Real 2022 Mar 01;26(4):1347-1371 [FREE Full text] [doi: 10.1007/s10055-022-00636-4] [Medline: 35250349]
35. Liu Q, Wang Y, Tang Q, Liu Z. Do you feel the same as I do? differences in virtual reality technology experience and acceptance between elderly adults and college students. Front Psychol 2020 Sep 25;11:573673 [FREE Full text] [doi: 10.3389/fpsyg.2020.573673] [Medline: 33101144]
36. Radiologic technologist. Cleveland Clinic. URL: https://my.clevelandclinic.org/health/articles/24667-radiologic-technologist [accessed 2024-02-07]
37. Lin CH, Shih HY, Sher PJ. Integrating technology readiness into technology acceptance: the TRAM model. Psychol Market 2007 May 24;24(7):641-657. [doi: 10.1002/mar.20177]
38. Liu Q, Wang Y, Yao MZ, Tang Q, Yang Y. The effects of viewing an uplifting 360-degree video on emotional well-being among elderly adults and college students under immersive virtual reality and smartphone conditions. Cyberpsychol Behav Soc Netw 2020 Mar;23(3):157-164. [doi: 10.1089/cyber.2019.0273] [Medline: 31663773]
39. Adelana OP, Ayanwale MA, Ishola AM, Oladejo AI, Adewuyi HO. Exploring pre-service teachers' intention to use virtual reality: a mixed method approach. Comput Educ X Reality 2023 Dec;3:100045. [doi: 10.1016/j.cexr.2023.100045]
40. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2014 Feb 16;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
41. Bae S, Jung TH, Moorhouse N, Suh M, Kwon O. The influence of mixed reality on satisfaction and brand loyalty in cultural heritage attractions: a brand equity perspective. Sustainability 2020 Apr 08;12(7):2956. [doi: 10.3390/su12072956]
42. Joo E, Yang J. How perceived interactivity affects consumers' shopping intentions in live stream commerce: roles of immersion, user gratification and product involvement. J Res Interact Market 2023 Feb 09;17(5):754-772. [doi: 10.1108/jrim-02-2022-0037]
43. Rodríguez-Ardura I, Meseguer-Artola A. E-learning continuance: the impact of interactivity and the mediating role of imagery, presence and flow. Inf Manag 2016 Jun;53(4):504-516. [doi: 10.1016/j.im.2015.11.005]
44. Skjoldborg NM, Bender PK, Jensen de López KM. The efficacy of head-mounted-display virtual reality intervention to improve life skills of individuals with autism spectrum disorders: a systematic review. Neuropsychiatr Dis Treat 2022 Oct;Volume 18:2295-2310. [doi: 10.2147/ndt.s331990]
45. Nguyen MH, Smets EM, Bol N, Loos EF, van Weert JC. How tailoring the mode of information presentation influences younger and older adults' satisfaction with health websites. J Health Commun 2018;23(2):170-180. [doi: 10.1080/10810730.2017.1421729] [Medline: 29345531]
46. Lv Z. Generative artificial intelligence in the metaverse era. Cognit Robot 2023;3:208-217. [doi: 10.1016/j.cogr.2023.06.001]
47. Liu Y, Siau KL. Generative artificial intelligence and metaverse: future of work, future of society, and future of humanity. In: Proceedings of the First International Conference on AI-Generated Content. 2023 Presented at: AIGC 2023; August 25-26, 2023; Shanghai, China. [doi: 10.1007/978-981-99-7587-7_10]
48. Huynh-The T, Pham QV, Pham XQ, Nguyen TT, Han Z, Kim DS. Artificial intelligence for the metaverse: a survey. Eng Appl Artif Intell 2023 Jan;117:105581. [doi: 10.1016/j.engappai.2022.105581]
49. Soliman MM, Ahmed E, Darwish A, Hassanien AE. Artificial intelligence powered Metaverse: analysis, challenges and future perspectives. Artif Intell Rev 2024 Feb 05;57:36. [doi: 10.1007/s10462-023-10641-x]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

NCC: National Cancer Center
RCC: regional cancer center
TAM: technology acceptance model
VR: virtual reality

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 26.02.24; peer-reviewed by S Yamane, T Thang; comments to author 15.03.24; revised version received 27.04.24; accepted 18.06.24; published 15.07.24.

Please cite as:
Kim S, Jung T, Sohn DK, Suh M, Chang YJ
Factors Associated With Continuous Use of a Cancer Education Metaverse Platform: Mixed Methods Study
J Med Internet Res 2024;26:e57762
URL: https://www.jmir.org/2024/1/e57762
doi:10.2196/57762
PMID:

©Sunghak Kim, Timothy Jung, Dae Kyung Sohn, Mina Suh, Yoon Jung Chang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Clinical Effectiveness of an Exergame-Based Exercise Training Program Using Ring Fit Adventure to Prevent and Postpone Frailty and Sarcopenia Among Older Adults in Rural Long-Term Care Facilities: Randomized Controlled Trial

Sheng-Hui Tuan[1,2,3], MD; Lin-Hui Chang[1,4], Prof Dr; Shu-Fen Sun[5,6], MD; Chien-Hui Li[2], MS; Guan-Bo Chen[7], MD; Yi-Ju Tsai[1,8], Prof Dr

[1]Institute of Allied Health Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[2]Department of Rehabilitation Medicine, Cishan Hospital, Ministry of Health and Welfare, Kaohsiung, Taiwan
[3]Department of Physical Medicine and Rehabilitation, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[4]Department of Occupational Therapy, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[5]Department of Physical Medicine and Rehabilitation, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan
[6]School of Medicine, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[7]Department of Internal Medicine, Kaohsiung Armed Forces General Hospital, National Defense Medical Center, Kaohsiung, Taiwan
[8]Department of Physical Therapy, College of Medicine, National Cheng Kung University, Tainan, Taiwan

Corresponding Author:
Yi-Ju Tsai, Prof Dr
Institute of Allied Health Sciences
College of Medicine
National Cheng Kung University
No 1, University Road
Tainan, 701
Taiwan
Phone: 886 6 235 3535 ext 5720
Email: yjtsaincku2024@gmail.com

Abstract

Background: Frailty and sarcopenia are geriatric syndromes of increasing concern and are associated with adverse health outcomes. They are more prevalent among long-term care facility (LTCF) users than among community dwellers. Exercise, especially multicomponent and progressive resistance training, is essential for managing these conditions. However, LTCFs, particularly in rural areas, face challenges in implementing structured exercise programs due to health care professional shortages. Moreover, older adults often become bored with repetitive exercise training and may lose interest over time. The Nintendo Switch Ring Fit Adventure (RFA) exergame is a novel exergame that combines resistance, aerobic, and balance exercises and offers a potential solution by boosting motivation in an immersive manner and reducing staff intervention needs.

Objective: We aimed to evaluate the clinical effectiveness of an exergame-based exercise training program delivered via RFA (exergame-RFA) in improving muscle mass and functional performance among older adult LTCF users.

Methods: This was a randomized controlled trial conducted from August 2022 to September 2023 and involved older adult LTCF users (aged ≥60 y) in rural southern Taiwan. Participants were randomized into an intervention group (exergame-RFA plus standard care) or a control group (standard care alone). The intervention, conducted seated with *arm fit skills* and trunk control exercises using the RFA, lasted 30 minutes twice weekly over 12 weeks. The primary outcomes measured were the Study of Osteoporotic Fractures index (serving as an indicator of frailty status) and the diagnostic criteria for sarcopenia (appendicular skeletal muscle mass index, handgrip strength, and gait speed). The secondary outcomes included functional performance (box and block test as well as maximum voluntary isometric contraction of the dominant upper extremity), muscle condition (muscle thickness measured using ultrasonography), activities of daily living (Kihon checklist), health-related quality of life (Short Form Health Survey-36), and cognitive function (brain health test). We used an intention-to-treat analysis, incorporating a simple imputation technique in statistical analysis. A mixed ANOVA, with time as a within-participant factor and intervention as a between-participant factor, was used to compare the training effects on outcomes.

XSL•FO
RenderX

Results: We recruited 96 individuals, of whom 60 (62%) underwent randomization. Of these 60 participants, 55 (92%) completed the study. Significant group×time interactions were observed in the intervention group in all primary outcomes (all $P<.001$, except $P=.01$ for handgrip strength) and most secondary outcomes, including maximum voluntary isometric contraction of the biceps ($P=.004$) and triceps brachii ($P<.001$) muscles, biceps muscle thickness measured using ultrasonography ($P<.001$), box and block test ($P<.001$), Kihon checklist (physical function: $P=.01$, mood status: $P=.003$, and total: $P=.003$), and brain health test ($P<.001$).

Conclusions: The exergame-RFA intervention significantly improved muscle mass, strength, and functional performance among older adult users of rural LTCFs, offering a novel approach to addressing frailty and sarcopenia.

Trial Registration: ClinicalTrials.gov NCT05360667; https://clinicaltrials.gov/study/NCT05360667

International Registered Report Identifier (IRRID): RR2-10.3389/fmed.2022.1071409

(J Med Internet Res 2024;26:e59468) doi:10.2196/59468

KEYWORDS

exergame; Ring Fit Adventure; sarcopenia; frailty; long-term care; multicomponent training

Introduction

Background

Global aging is intensifying due to rapid socioeconomic advances and longer lifespans. It is predicted that, by 2025, Taiwan will evolve into a super-aged society, with >20% of its population being aged ≥65 years [1]. This increase in the older adult population brings multiple health challenges, including the effects of aging, chronic diseases, cognitive decline, malnutrition, diminished physical fitness, and psychosocial issues. Aging and inactivity lead to decreases in muscle mass, structure, and strength [2], with individuals aged >50 years losing muscle mass at a rate of 1% to 2% annually [3]. This contributes to sarcopenia, a gradual decline in muscle mass, strength, and physical performance that occurs as individuals age [4], as well as frailty, a state of heightened vulnerability due to the decline in the reserve and function of multiple physiological systems [4].

Frailty and sarcopenia are significant concerns globally, profoundly influencing older adults' physical capabilities, health risks, quality of life, and longevity [5,6]. In Taiwan, frailty and sarcopenia affect 17.6% [7] and from 4.1% to 9.3% [8] of community-dwelling older adults, respectively. Notably, these conditions are more prevalent among individuals who have been institutionalized at long-term care facilities (LTCFs). Frailty rates can reach 79.4% in nursing homes [9]. Sarcopenia rates vary from 17.7% to 73.3% in long-term nursing facilities [10] and from 22% to 87.1% in day care centers [11]. The higher prevalence of these conditions among older adults in LTCFs is linked to their advanced age, more severe health problems, and increased dependence on assistance for activities of daily living (ADLs) [11].

The current gold standard for managing frailty and sarcopenia focuses on preserving skeletal muscle mass and maintaining muscle strength [12]. Resistance training (RT), particularly progressive RT (PRT), combined with multicomponent exercises, is supported by moderate- to high-quality evidence [13,14]. This evidence highlights their effectiveness in increasing muscle mass, strength, and physical performance in older adults with frailty and sarcopenia [14,15]. When it comes to exercise training, it should be consistently performed at a slightly challenging level of intensity. However, resource limitations, particularly in staffing, present considerable challenges to systematically implementing regular exercise programs in LTCFs [16], especially those in rural areas. Exergames, which are video games necessitating whole-body player engagement, could address the aforementioned issues. They provide real-time interactivity [17] as well as engaging, multisensory environments, facilitating immersive experiences through extensive body movements [18]. The gamification and enticing environments of exergames have been shown to motivate older adults, enhancing their engagement with physical exercises [19]. As a result, the incorporation of exergames reduces the necessity for extensive staff involvement in interventions, encourages patients to engage in more vigorous activities, and boosts their motivation levels.

Objectives

While the clinical effectiveness of using exergames among community-dwelling older adults is recognized [20,21], their efficacy within LTCFs remains underexplored. Most research focuses on improvements in health-related quality of life (HRQoL), cognitive functions, and overall functional status [22,23]. The literature mainly discusses exergames designed for aerobic and balance training, with limited incorporation of PRT principles. The introduction of Nintendo's Ring Fit Adventure (RFA) for the Nintendo Switch marks a significant evolution in exergaming, combining aerobic, strength, and balance training tailored to users' individual needs [24]. Despite its apparent potential, detailed empirical studies to fully ascertain its benefits are still forthcoming. Our team proposed and initiated the Using an Exergame to Prevent and Postpone the Loss of Muscle Mass, Muscle Strength, and Functional Performance in Rural Elders (EXPPLORE) trial [25], aiming to investigate the feasibility and clinical application of RFA for older adults in rural LTCFs. This study presents the outcomes of this protocol.

Methods

Study Design and Participants

This study was a prospective randomized controlled trial conducted from August 2022 to September 2023. The trial was designed in compliance with the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) 2013

XSL•FO
RenderX

statement for randomized controlled trials and the CONSORT (Consolidated Standards of Reporting Trials) guidelines and has been registered at ClinicalTrials.gov (NCT05360667) (Multimedia Appendix 1; [26]).

Participants were recruited from LTCFs, including day care centers and nursing homes, in rural areas of Kaohsiung City, Taiwan. The eligibility criteria included (1) being aged ≥60 years, (2) residing in LTCFs or attending LTCFs for at least 1 month, (3) having the ability to understand and communicate in Chinese or Taiwanese, (4) possessing adequate cognitive abilities as determined by the researchers to provide informed consent, and (5) actively participating in the exergame-based exercise and data collection process. In addition, participants needed to be capable of remaining seated for >50 minutes during the training session and able to complete the gait speed measurement. We excluded individuals (1) diagnosed with significant cardiopulmonary diseases; (2) on regular oxygen supplementation; (3) with unmanaged hypertension; and (4) with recent infections, fractures, or other conditions that would contraindicate exercise participation according to the American College of Sports Medicine (ACSM) guidelines [27].

Participants were randomly assigned to the intervention and control groups in a 1:1 ratio. The intervention group received standard care combined with RFA for 12 weeks, whereas the control group continued with the usual care typically provided in LTCFs. In the intervention group, traditional sedentary activities common in LTCFs, such as singing, tabletop games, and gardening, were replaced with RFA training sessions. The duration allocated for activities was kept consistent across both groups.

A flowchart detailing the participant inclusion and data collection processes is presented in Figure 1. This study was conducted under the published protocol of the EXPPLORE trial [25].

Figure 1. Flowchart of the participant inclusion and data collection processes. ASMMI: appendicular skeletal muscle mass index; BBT: box and block test; BHT: brain health test; HGS: handgrip strength; KC-T: Kihon checklist, Taiwanese version; LTCF: long-term care facility; SF-36: Short Form Health Survey-36; SNOSE: sequentially numbered, opaque, sealed envelope; SOF: Study of Osteoporotic Fractures.

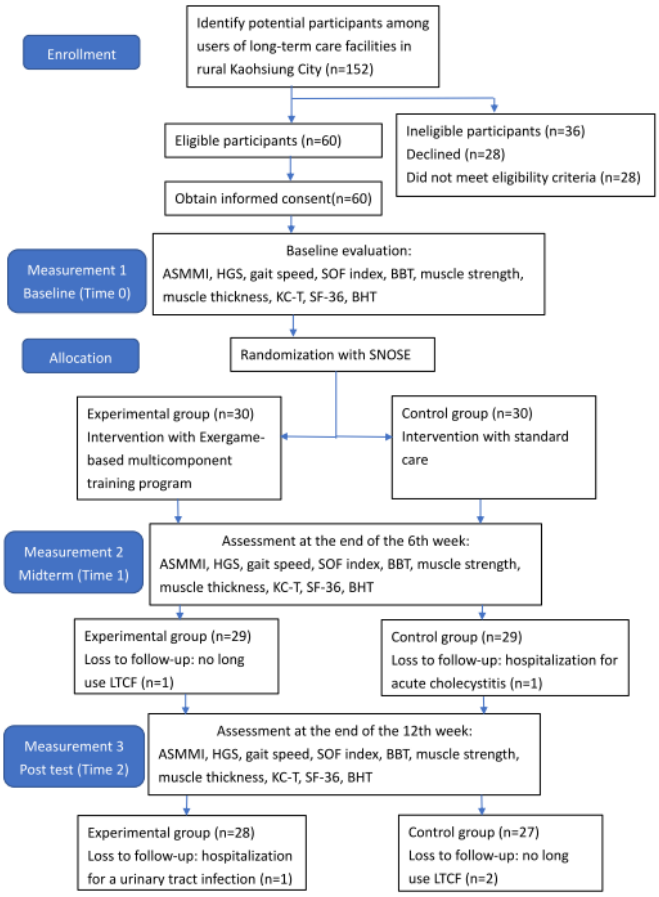

Randomization and Blinding

Randomization was achieved using sequentially numbered, opaque, sealed envelopes that contained the group assignment numbers. These envelopes were prepared by an independent individual, and the group assignment numbers were generated by another individual who was not involved in the clinical aspects of the study to ensure blinding to the research proceedings. After verifying eligibility based on the inclusion and exclusion criteria, participants were randomized equally between the intervention and control groups. Given the nature of the study and the distinct characteristics of the interventions, blinding health professionals in the LTCFs and participants to the treatment conditions was not possible. However, to maintain an objective evaluation and minimize bias, the 2 assessors conducting the outcome measurements and the data analyst interpreting the results were blinded throughout the duration of the study.

XSL•FO
RenderX

Intervention Group Training Program

The training program (exergame-based exercise training program delivered via RFA [exergame-RFA]) focused on PRT and functional movements, primarily targeting the upper extremities. The RFA system requires a Nintendo Switch console, a Ring-Con (a Pilates ring held by the user), Joy-Con wireless controllers (right Joy-Con, attached to the Ring-Con; and left Joy-Con, affixed to the leg strap accessory wrapped around the thigh of the player), and a display screen (Figure 2).

Figure 2. Components of the Nintendo Switch Ring Fit Adventure exergame. (A) Ring-Con, a Pilates ring held by the user; and Joy-Con wireless controllers: (B) right Joy-Con (attached to the Ring-Con) and (C) left Joy-Con (affixed to the leg strap accessory wrapped around the thigh of the player).

In RFA's adventure mode, players move through the game by performing physical activities such as jogging or squatting. Given the increased fall risk among LTCF residents [28], we used RFA's knee assist mode, making the game more accessible. Enabling knee assist mode provides support when users perform exercises requiring stepping, kneeling, or squatting, making these activities more manageable for players with physical limitations. By adjusting exercise intensity and providing support to the knees, knee assist mode enhances accessibility and inclusivity, allowing a wider range of players to enjoy the game's fitness benefits. In RFA's battle mode, players engage in aerobic activities, RT, and yoga exercises involving the full body to advance in the game. Players are rewarded with in-game currency, character customization, and level progression based on exercise volume, encouraging skill improvement.

The exergame-RFA intervention was scheduled twice weekly, with sessions spaced 48 hours apart. Each session lasted 50 minutes: 10 minutes each for warm-up and cool-down exercises and 30 minutes for the main exercise. This was maintained over a 12-week period, with each session overseen by a qualified therapist. Six *fit skills*, which derived from the game's built-in movement data set and focused on upper extremity and trunk training, were included, with participants using 3 skills during battle mode to defeat opponents. The exergame-based exercises, with targeted muscle groups and joint movements highlighted, are presented in Table S1 in Multimedia Appendix 2 [29-40]. The target intensity of the training was set at a level of 13 (*somewhat hard*) on the Borg rating of perceived exertion (RPE) scale, with progression automatically adjusted by RFA based on each player's performance (details on RFA content, gameplay steps, and auto-adjustment mechanisms are presented in Table S2 in Multimedia Appendix 2, which outlines the exergame-RFA exercise prescription using the frequency, intensity, time, type, volume, and progression principle formulated by the ACSM [27]).

Control Group Standard Care Program

For the control group, standard care was provided in accordance with the routine practices of LTCFs. This included group-based activities such as calisthenics (modified for seated positions), horticultural therapy, and sedentary group activities (eg, tabletop games). These sessions, led by a therapist, occurred twice a week, with each session lasting from 30 to 60 minutes, depending on the activity.

Outcomes Measured

All participants underwent 3 distinct assessments. The initial assessment was conducted at baseline (T0) before randomization. Subsequent evaluations took place at the end of week 6 (midstudy; T1) and at the end of week 12 (after the intervention; T2). The selected primary outcomes for this study were based on the diagnostic criteria for sarcopenia as proposed by the Asian Working Group for Sarcopenia. These included the appendicular skeletal muscle mass index (ASMMI), dominant handgrip strength (HGS), and customary gait speed.

XSL•FO
RenderX

In addition, the Study of Osteoporotic Fractures index, serving as an indicator of frailty status, was also designated as a primary outcome.

The secondary outcomes for this study encompassed several dimensions, including muscle strength (indicated by the maximum voluntary isometric contraction [MVIC] of the biceps and triceps brachii muscles, measured using the microFET3 [Hoggan Scientific, LLC]) of the dominant side, muscle morphology (indicated by the thickness of the biceps brachii, quadriceps, and gastrocnemius muscles, measured using a portable LOGIQ e ultrasound device [GE HealthCare] operated by the same operator with gentle pressure in standard positions), manual dexterity of the dominant hand (measured using the box and block test), ADLs (measured using the Kihon checklist), HRQoL [measured using the Short Form Health Survey-36 [SF-36]), and cognitive function (measured using the brain health test). Details of the RFA training process and each outcome measured are available in the published protocol of the EXPPLORE trial [25], and Table S3 in Multimedia Appendix 2 summarizes the instruments and measures implemented for data collection in this study.

Statistical Analysis

Quantitative Data

Statistical analyses were conducted using SPSS for Windows (version 21.0; IBM Corp). Continuous variables were represented as means and SDs, while categorical data were denoted in numbers and percentages. Preliminary checks for data normality and homoscedasticity were performed before executing each analysis. For comparing demographic data between the experimental and control cohorts, appropriate statistical tests, including the chi-square test, 2-tailed *t* test, and Mann-Whitney *U* test, were used, depending on the data distribution characteristics.

To address potential attrition, an intention-to-treat analysis, incorporating a simple imputation technique, was used. Specifically, for any missing data, the last-observation-carried-forward method was used [41]. To assess the effects of the intervention on outcomes over time, a mixed ANOVA was used, designating time as a within-participant variable and the nature of the intervention as a between-participant variable. For any significant interactions or main effects, post hoc assessments were carried out using the Bonferroni test.

Sample Size Estimation

Given the absence of a high-quality study evaluating the effects of a comparable intervention involving using RFA on muscle strength, physical activity, and functionality at the time our manuscript was drafted, we based our primary outcome measure (ASMMI) on the findings from a single study that used an augmented reality–based exercise [42]. The effect size observed in the study for the augmented reality–based intervention was 0.71, which is notably substantial [42]. The calculations using G*Power for Windows (version 3.1.9.2; Heinrich Heine University), indicated a requirement for a minimum of 18 participants in each group to detect a difference of 2 SDs in the ASMMI score between the groups, assuming a power of 80% and an α level of 5%. The effect size, situated within the *F* test family, was set at 0.4 [43]. Anticipating a potential dropout rate of 20%, especially considering that older adult LTCF users often have multiple comorbidities, it was deemed prudent to increase the number of participants in each group to 22. Consequently, the recruitment target for the study was set at a minimum of 44 participants.

Ethical Considerations

This study, conducted in accordance with the Declaration of Helsinki, received approval from the institutional review board of National Cheng Kung University Hospital (B-ER-111-058). Written informed consent was obtained from all participants. The study data were anonymized and deidentified. No compensation was provided to participants.

Results

Baseline Characteristics and Adherence to the Study Protocol

From August 2022 to September 2023, a total of 96 older adults from nursing homes and day care centers in rural Kaohsiung City, Taiwan, were recruited, of whom 28 (29%) declined to participate, and 8 (8%) did not meet the eligibility criteria. The remaining 60 participants consented and were randomized: 30 (50%) to the intervention group and 30 (50%) to the control group. Of these 60 participants, 55 (92%) completed the study. Of the 30 participants in the intervention group, 2 (7%) dropped out (n=1, 50% before T1 because they no longer used the LTCFs; and n=1, 50% before T2 because they were hospitalized for a urinary tract infection). Of the 30 participants in the control group, 3 (10%) withdrew (n=1, 33% before T1 because they were hospitalized for acute cholecystitis; and n=2, 67% before T2 because they no longer used the LTCFs). The dropout rates were not significantly different between the groups (*P*=.64).

Table 1 shows the baseline characteristics of the participants. Of the 30 participants in the intervention group, in terms of the frailty and sarcopenia status distribution, 4 (13%) were categorized as *prefrail*, 13 (43%) as *frail*, 16 (53%) as *possible sarcopenia*, 2 (7%) as *confirmed sarcopenia*, and 3 (10%) as *severe sarcopenia*. Of the 30 participants in the control group, 8 (27%) were categorized as *prefrail*, 7 (23%) as *frail*, 24 (80%) as *possible sarcopenia*, 1 (3%) as *confirmed sarcopenia*, and 2 (7%) as *severe sarcopenia*. The average long-term care case-mix system levels were 4.20 (SD 0.79) for the intervention group and 4.40 (SD 0.82) for the control group. A case-mix system categorizes individuals based on their requirements for assistance with ADLs. These levels play a crucial role in determining the suitable level of care and support needed by individuals in LTCFs, ranging from minimal assistance to intensive nursing care [44]. This result indicated that both groups required moderate daily activity assistance. There were no significant baseline differences between the groups, except for the fact that the participants in the intervention group had a lower average HGS.

XSL•FO
RenderX

Table 1. Demographic characteristics of the participants at baseline.

Characteristics	Intervention group (n=30)	Control group (n=30)	*P* value[a]
Sex, n (%)			.79
Male	10 (33)	11 (37)	
Female	20 (67)	19 (63)	
Age (y), mean (SD)	78.83 (7.71)	78.73 (6.82)	.96
Height (cm), mean (SD)	152.89 (7.99)	153.88 (8.84)	.67
Weight (kg), mean (SD)	57.00 (10.05)	61.44 (11.07)	.12
BMI (kg/m^2), mean (SD)	24.33 (3.57)	25.86 (3.47)	.12
Body fat (%), mean (SD)	29.16 (7.82)	32.86 (6.92)	.08
BMI group, n (%)			.88
Underweight	1 (3)	1 (3)	
Normal	13 (43)	11 (37)	
Overweight	13 (43)	16 (53)	
Obese	3 (10)	2 (7)	
Case-mix system levels, mean (SD)	4.20 (0.79)	4.40 (0.82)	.53
Participants below cutoff scores, n (%)			
ASMMI[b]	5 (17)	4 (13)	.72
Dominant HGS[c]	21 (70)	28 (93)	.02[d]
Gait speed	28 (93)	29 (97)	.55
Frailty status, n (%)			.19
Robust	13 (43)	15 (50)	
Prefrail	4 (13)	8 (27)	
Frail	13 (43)	7 (23)	
Sarcopenia status, n (%)			.16[e]
No	9 (30)	3 (10)	
Possible	16 (53)	24 (80)	
Confirmed	2 (7)	1 (3)	
Severe	3 (10)	2 (7)	

[a]Mann-Whitney *U* test (continuous variables) or chi-square test (categorical variables).

[b]ASMMI: appendicular skeletal muscle mass index.

[c]HGS: handgrip strength.

[d]The *P* value met the threshold for significance (*P*<.05).

[e]Fisher exact test.

Comparisons of Primary Outcomes Between the Intervention and Control Groups

A significant group×time interaction effect was observed in frailty status as measured by the Study of Osteoporotic Fractures index ($F_{2,116}$=5.594; *P*=.007), with a small to medium effect size (η^2=0.088) [45], indicating that the intervention group had a more significant reduction in frailty scores over time (Table 2). In terms of the sarcopenia diagnostic criteria, (1) body components showed notable improvements in appendicular skeletal muscle mass over time ($F_{2,58}$=7.934; *P*=.005), with a significant group×time interaction ($F_{2,116}$=9.895; *P*=.002), indicating larger gains in the intervention group. This trend was reflected in the ASMMI scores, revealing significant time ($F_{2,58}$=7.339; *P*=.005) and group×time effects ($F_{2,116}$=8.408; *P*=.003). Post hoc analysis indicated significant improvements from baseline to after the intervention and from midstudy to after the intervention in both appendicular skeletal muscle mass and ASMMI scores. (2) There was a significant increase in dominant HGS scores in the intervention group, with a notable interaction effect ($F_{2,116}$=5.326; *P*=.01) and a small to medium effect size (η^2=0.087). Post hoc analysis revealed significant improvements from baseline to after the intervention and from midstudy to after the intervention, exceeding the minimal

XSL•FO
RenderX

clinically important difference (MCID). Lastly, (3) functional performance exhibited significant time effects in gait speed (ms; $F_{2,58}$=6.664; P=.005), with a large effect size for the group×time interaction (η^2=0.314), indicating that the intervention group experienced significantly greater improvements in walking speed. Post hoc analysis showed significant improvements from baseline to after the intervention, exceeding the MCID. The MCID signifies the smallest change in a clinical outcome measure perceived as meaningful by patients or clinicians [46]. The improvements in both HGS scores and walking speed exceeded the MCID, which is an indication of not only statistical significance but also clinical relevance after RFA training.

XSL•FO
RenderX

Table 2. Data of primary outcomes at baseline, midstudy (week 6), and after the intervention (week 12).

Primary outcomes	Baseline	Midstudy	After the intervention
SOF[a] index			
Intervention group, mean (SD)	1.07 (1.05)	1.00 (1.01)	0.80 (0.85)
Control group, mean (SD)	0.73 (0.83)	0.80 (0.85)	0.80 (0.81)
2-way ANOVA			
Time			
F test (*df*)	__[b]	—	2.650 (2,58)
P value	—	—	.08
Group×time			
F test (*df*)	—	—	5.594 (2,116)
P value	—	—	.007[c,d]
Effect size (η^2)	—	—	0.088
Appendicular skeletal muscle mass (kg)			
Intervention group, mean (SD)	17.29 (3.81)	18.63 (4.31)	19.13 (4.58)
Control group, mean (SD)	17.05 (2.95)	16.94 (2.81)	16.96 (2.40)
2-way ANOVA			
Time			
F test (*df*)	—	—	7.934 (2,58)
P value	—	—	.005[c]
Group×time			
F test (*df*)	—	—	9.895 (2,116)
P value	—	—	.002[c,d]; .008[e]; .02[f]
Effect size (η^2)	—	—	0.165
Appendicular skeletal muscle mass index (kg/m^2)			
Intervention group, mean (SD)	7.37 (1.25)	7.92 (1.72)	8.15 (1.77)
Control group, mean (SD)	7.17 (0.82)	7.12 (0.86)	7.15 (0.84)
2-way ANOVA			
Time			
F test (*df*)	—	—	7.339 (2,58)
P value	—	—	.005[c]
Group×time			
F test (*df*)	—	—	8.408 (2,116)
P value	—	—	.003[c,d]; .009[e]; .03[f]
Effect size (η^2)	—	—	0.144
Fat mass index (kg/m^2)			
Intervention group, mean (SD)	7.41 (2.73)	6.98 (2.71)	6.62 (2.60)
Control group, mean (SD)	9.00 (2.94)	8.71 (3.03)	8.38 (2.95)
2-way ANOVA			
Time			
F test (*df*)	—	—	16.233 (2,58)
P value	—	—	<.001[c]

XSL•FO
RenderX

Primary outcomes	Baseline	Midstudy	After the intervention
Group×time			
F test (*df*)	—	—	0.241 (2,116)
P value	—	—	.71
Effect size (η^2)	—	—	0.005
Fat-free mass index (kg/m^2)			
Intervention group, mean (SD)	16.96 (1.76)	17.10 (2.00)	17.67 (1.99)
Control group, mean (SD)	16.85 (1.00)	16.86 (1.52)	17.26 (1.80)
2-way ANOVA			
Time			
F test (*df*)	—	—	7.583 (2,58)
P value	—	—	.002[c]
Group×time			
F test (*df*)	—	—	0.480 (2,116)
P value	—	—	.57
Effect size (η^2)	—	—	0.010
Dominant handgrip strength (kg; MCID[g]: 2.44-2.69 kg)			
Intervention group, mean (SD)[h]	14.51 (6.42)	17.60 (6.49)	20.80 (7.59)
Control group, mean (SD)	13.40 (5.51)	15.16 (5.73)	16.20 (6.04)
2-way ANOVA			
Time			
F test (*df*)	—	—	35.698 (2,58)
P value	—	—	<.001[c]
Group×time			
F test (*df*)	—	—	5.326 (2,116)
P value	—	—	.01[c,d]; <.001[e,i]; .001[f]
Effect size (η^2)	—	—	0.087
Walking speed (ms; MCID: 0.1-0.2 ms)			
Intervention group, mean (SD)[h]	0.47 (0.25)	0.51 (0.31)	0.53 (0.33)
Control group, mean (SD)	0.48 (0.23)	0.58 (0.30)	0.53 (0.27)
2-way ANOVA			
Time			
F test (*df*)	—	—	6.664 (2,58)
P value	—	—	.005[c]
Group×time			
F test (*df*)	—	—	25.173 (2,116)
P value	—	—	<.001; .009[a]

XSL•FO
RenderX

Primary outcomes	Baseline	Midstudy	After the intervention
Effect size (η^2)	—	—	0.314

[a]SOF: Study of Osteoporotic Fractures.

[b]Not applicable.

[c]*P* values that met the threshold for significance were analyzed using mixed repeated measures ANOVA with time as a within-participant factor and group as a between-participant factor.

[d]A post hoc Bonferroni test was performed to compare baseline, midstudy, and postintervention scores in the intervention group.

[e]Post hoc analysis showed significant differences between baseline and postintervention scores.

[f]Post hoc analysis showed significant differences between midstudy and postintervention scores.

[g]MCID: minimal clinically important difference.

[h]The improvements in the scores from baseline to after the intervention exceeded the MCID.

[i]Post hoc analysis showed significant differences between baseline and midstudy scores.

Comparisons of Secondary Outcomes Between the Intervention and Control Groups

The secondary outcomes included muscle strength, morphology, hand dexterity, ADLs, HRQoL, and cognitive function. The intervention group had significant increases in the MVIC of the biceps and triceps brachii muscles, with time effects (biceps: $F_{2,58}$=12.879; P<.001; triceps: $F_{2,58}$=32.684; P<.001) and group×time interactions (biceps: $F_{2,116}$=6.532; P=.004; triceps: $F_{2,116}$=11.797; P<.001; Multimedia Appendix 3). Muscle thickness of the biceps brachii (measured using ultrasonography) also increased ($F_{2,58}$=10.520; P<.001), with significant group×time interaction ($F_{2,116}$=10.029; P<.001) in the intervention group. Hand dexterity improvements in the intervention group were significant over time ($F_{2,58}$=16.848; P<.001), with a notable group×time interaction ($F_{2,116}$=12.374; P<.001). ADLs, as measured by the Kihon checklist, showed no significant changes over time within the groups for physical function ($F_{2,58}$=1.005; P=.34), but there were significant group×time interactions for both physical function ($F_{2,116}$=5.599; P=.01) and mood function ($F_{2,116}$=7.376; P=.003), indicating impactful intervention on ADLs. HRQoL, as assessed by SF-36 scores, demonstrated no significant changes over time in physical function ($F_{2,58}$=0.667; P=.47) or in the interaction effect between group and time ($F_{2,116}$=0.516; P=.54). However, improvements in mental function were observed ($F_{2,58}$=6.571; P=.009) without significant differences between the groups.

In summary, the exergame-RFA intervention resulted in significant improvements in muscle strength (medium to large effect sizes), hand dexterity (large effect size), ADLs (medium effect size), and cognitive function (large effect) compared to the control group. However, the exergame-RFA intervention did not significantly enhance HRQoL or the thickness of the quadriceps and gastrocnemius muscles measured using ultrasonography compared to the control group. Summary of the significant group×time interactions of the outcome measures of this study are presented in Multimedia Appendix 4.

Discussion

Principal Findings

To our knowledge, this was the first study to evaluate the clinical effectiveness of RFA, an exergame that uniquely integrates aerobic and RT with real-time feedback in an engaging manner. Considering the multifaceted nature of frailty and sarcopenia and their varied definitions [12], our focus was on muscle and functional performance parameters. Our findings revealed that the exergame-RFA intervention significantly improved muscle mass, strength, and functional performance among older adult LTCF users compared to standard care. This underscores its potential as a novel therapeutic strategy to partially prevent and address frailty and sarcopenia in older adults.

This project stands out because it delivered an exergame-based exercise program using the RFA system. Similar to other exergames, RFA uses a gamified approach and immersive scenarios for player motivation through role-playing [47]. It offers visual feedback and sensory feedback from the screen and the Ring-Con, respectively. Unlike other exergames that mostly require fingertip operation, RFA, with its unique Ring-Con, demands significantly higher energy expenditure [48] and encompasses a multicomponent exercise regimen, blending PRT with aerobic activities for enhanced strength, balance, and muscle flexibility. Previous studies have confirmed the efficacy of PRT in improving physical function and strength among community-dwelling older adults due to its accessibility, low cost, and effectiveness [49,50]. Consistent with these findings, our results demonstrate that the exergame-RFA intervention can significantly enhance ASMMI, HGS, MVIC of the dominant upper extremity, and overall functional performance.

Another distinctive aspect of delivering exergames through RFA is its capability to determine the initial volume of exercise required at each stage, adjusting it progressively and automatically based on the player's performance. At the start of the exergame-RFA intervention, participants were allowed to select their exercise intensity based on age, physical fitness, and preferred exercise intensity. This intensity setting in RFA is determined by a manufacturer-developed algorithm and follows a specific procedure: initially, users input their age and sex (step 1), followed by their exercise habits and desired exercise intensity (load; step 2). From these inputs, the load

XSL•FO
RenderX

level (1-30) is automatically set on a scale from 1 (weakest) to 30 (strongest). As the load level increases, users need to apply more force on the Ring-Con to defeat enemies in the game, facilitating personalized exercise progression. This feature saves time and manpower by being machine adjustable [51]. However, for older adults (aged ≥65 y), the intensity options are limited, with the system defaulting to a maximum load setting of ≤15. While a preliminary study on older adults suggests that they can generally achieve moderate or greater intensity levels in RFA as assessed by the Karvonen heart rate reserve formula, it also recommended using established and validated exercise intensity indices (eg, heart rate and Borg Rating of Perceived Exertion Scale) alongside considering RFA's intensity settings [51]. Furthermore, the ACSM recommends increasing intensity over time to maintain exercises at moderate levels (41%-60% of 1 repetition maximum for resistance exercise and Borg RPE 12-14 for aerobic exercise) [52]. Consequently, we set the target intensity of our exergame-based exercise at Borg RPE 13, leaving the weight intensity to the discretion of the Nintendo Switch. Although this approach is practical in real-world settings, and it is safer for older adults to perform PRT by squeezing or stretching the Ring-Con, a limitation remains in not being able to verify whether the Ring-Con's weight intensity aligns with moderate levels.

Although they have the capacity to walk short distances for gait assessments, many older adults in LTCFs rely on wheelchairs for mobility (attributed to poor muscle endurance) [28]; hence, our RFA program incorporated a knee assist mode directly within the RFA system. This feature allowed us to focus on exercises for strengthening the upper extremities and trunk. This approach enabled participants to train in a seated position, thereby minimizing the risk of falls. Indeed, no participant in the intervention group experienced serious adverse effects. Consequently, it was logical that most outcomes of this study, including HGS, MVIC of the biceps and triceps brachii muscles, thickness of the biceps muscle measured using ultrasonography, and the box and block test, demonstrated significant group×time training effects.

In this study, we assessed both MVIC and muscle thickness measured using ultrasonography to determine whether early morphological changes in the trained muscles could be detected prematurely. This approach was taken because aging can dampen the hypertrophic response of muscle groups to RT [53]. Unlike in healthy young adults, where neural factors play a significant role in the initial increase in strength, and muscle hypertrophy becomes predominant after the first 3 to 5 weeks, the effects of muscle training in older adults may rely entirely on neuromuscular adaptation after an 8-week training course [54]. Interestingly, we observed significant improvements in biceps brachii muscle thickness measured using ultrasonography, rather than in MVIC, at the midstudy assessment. By the end of the intervention, both MVIC and biceps brachii muscle thickness measured using ultrasonography showed significant enhancements. However, no improvement in the thickness of lower leg muscles measured using ultrasonography was observed, despite a significant increase in walking speed after the intervention. This suggests that improvements in muscle parameters do not directly translate into functional performance enhancements. The relationship between muscle parameters and functional performance is complex and likely influenced by multiple factors.

Although the exergame-RFA intervention was focused on the training of upper extremities and the trunk, we found significant improvement in gait speed, one of the diagnostic criteria of sarcopenia. Walking encompasses complex movements requiring several functional tasks [55]. Participants had to maintain balance in various positions, such as leaning forward, reaching forward, and lateral shifting, to perform the exergame-RFA exercise prescription. Although our protocol did not permit training of the lower extremities, the movements in *arm fit skills* exercises likely engaged the trunk, hips, and knees. Granacher et al [56] demonstrated that the ability of older adults to rise from a chair, ambulate, and make turns improved after 9 weeks of core muscle strength training. Park et al [57] observed an increase in walking speed after a sitting boxing program for 6 weeks. Both studies showed that strengthening programs can induce adaptive processes, particularly in the neuromuscular system. These processes, in turn, enhance balance performance and functional mobility. Therefore, the observed increase in gait speed in our study was reasonable.

We observed significant improvements in ADLs, both in physical (Kihon checklist items 1-20) and mood (Kihon checklist items 21 to 25) aspects. Nevertheless, after the exergame-RFA intervention, we only noted a significant improvement in mental function, not in physical function, as measured by the SF-36. A systematic review involving 15 eligible randomized controlled trials indicated that exercise positively affects physical outcomes and enhances QoL and ADLs in older adults [58]. A large cohort study showed that community-dwelling older adults engaging in moderate to vigorous physical activity had a lower risk of ADL dependence [59]. Another study found a significant association between better physical function and higher levels of physical function as measured by the SF-36 [60]. Our results align with these findings to some extent. However, whether there is a strong correlation between muscle strength, muscle mass, functional performance, and these questionnaires still requires larger studies with longer follow-up periods to elucidate the association and even the causal relationship.

Exergames have been shown to have therapeutic effects on cognitive function among older adults [61,62]. Exergame environments enhance spatial awareness, challenging players with tasks that involve visual and auditory stimuli, cues, and feedback. Immersion into these virtual environments aids in attention restoration, stress reduction, and cognitive rehabilitation [62]. In addition, exergames necessitate decision-making during gameplay [61]. Therefore, it was not surprising to observe significant improvements in the brain health test score after the exergame-RFA intervention. Identifying a cutoff value for cognitive measurements to determine suitability for exergame participation among LTCF users would enhance the application of these findings.

Limitations

Our study includes several limitations. First, while the sample size exceeded the minimum requirements for statistical analysis,

XSL•FO
RenderX

it remained relatively small. Second, participants were recruited from nursing homes and day care centers in rural southern Taiwan, leading to a lower attrition rate than anticipated [63], potentially due to the unique characteristics of rural regions where older adults have limited health care options and may value services more highly. As a result, our findings may only be generalizable to similar populations. Third, the exergame-RFA intervention and evaluation period spanned only 3 months. Although improvements in muscle mass, strength, muscle thickness measured using ultrasonography, and certain questionnaire aspects were observed, this duration may have been too brief to detect significant increases in functional performance among the older adults. Fourth, there was no established protocol for training with RFA. Although the exercise prescription was based on ACSM guidelines for older adults, we cannot confirm the idealness of this exergame-based exercise protocol. The individualized setting and progression of exercises, a hallmark of using RFA for older adult training, were assessed solely through the Borg RPE scale, leaving uncertainties about the precise resistance offered by the Ring-Con during each session. Fourth, there was no established protocol for training with RFA. Although the exercise prescription was based on ACSM guidelines for older adults, we cannot confirm the idealness of this exergame-based exercise protocol. Furthermore, although the ability to automatically customize exercise settings and progression is a key feature of using RFA for training older adults, in this project, exercise intensity was only measured using the subjective Borg RPE scale. We were unable to verify the precise resistance provided by the Ring-Con during each training session. Finally, due to the small sample size, we did not perform a subgroup analysis to determine whether specific subgroups would benefit more from the exergame-RFA intervention. Future studies should aim to recruit larger samples to explore how older adults with varying degrees of frailty and sarcopenia respond to the exergame-RFA intervention. We also recommend further investigation into the optimal volume and long-term effects of the exergame-RFA intervention on older adults through larger randomized trials with control groups and extended follow-up periods, drawing from this study as a reference.

Conclusions

In summary, our study contributes valuable insights into the practical application of exergame-based exercises within LTCF settings, affirming the potential of such interventions to improve muscle parameters; functional performance; and, by extension, HRQoL among older adults. These findings not only echo the sentiments expressed in the prestudy expectations but also pave the way for future research to refine and expand upon these interventions, ensuring that they meet the holistic needs of the aging population. Further studies could explore the integration of lower extremity exercises and more personalized exergame programs, aiming to encompass the full spectrum of sarcopenia criteria and ADL requirements.

Acknowledgments

This study received funding from the Ministry of Health and Welfare, Taiwan (MOHW-11237), and the Ci-Shan Hospital Research and Development Grant (CHIS-112001). The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SHT planned the project, developed the research design, measured the outcomes, applied for funding, and wrote the first draft. LHC was responsible for reviewing the literature, evaluating the appropriate measurement tools, and revising the manuscript. SFS calculated the sample size and was responsible for the revisions of the manuscript. CHL was responsible for reviewing the literature and evaluating the appropriate measurement tools. GBC measured the outcomes and performed the statistical analysis. YJT planned the project, developed the research design, and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT checklist.
[PDF File (Adobe PDF File), 147 KB - jmir_v26i1e59468_app1.pdf]

Multimedia Appendix 2
Flowchart of the exergame-based multicomponent training program; exercise prescription of an exergame using Ring Fit Adventure based on the frequency, intensity, time, type, volume, and progression principle; and instruments and measures implemented for data collection in this study.

XSL•FO
RenderX

[DOCX File , 166 KB - jmir_v26i1e59468_app2.docx]

Multimedia Appendix 3
Data of secondary outcomes at baseline, midstudy (week 6), and after the intervention (week 12).
[OCX File , 30 KB - jmir_v26i1e59468_app3.ocx]

Multimedia Appendix 4
Summary of significant group×time effects on outcome measures.
[DOCX File , 16 KB - jmir_v26i1e59468_app4.docx]

References

1. Population projections for the Republic of China (Taiwan): 2020-2070. Department of Human Resources Development National Development Council. 2020. URL: https://tinyurl.com/3ex7x343 [accessed 2024-04-29]
2. Seene T, Kaasik P, Riso EM. Review on aging, unloading and reloading: changes in skeletal muscle quantity and quality. Arch Gerontol Geriatr 2012 Mar;54(2):374-380. [doi: 10.1016/j.archger.2011.05.002] [Medline: 21632125]
3. Marzetti E, Calvani R, Bernabei R, Leeuwenburgh C. Apoptosis in skeletal myocytes: a potential target for interventions against sarcopenia and physical frailty - a mini-review. Gerontology 2012;58(2):99-106 [FREE Full text] [doi: 10.1159/000330064] [Medline: 21952604]
4. Morley JE. Frailty and sarcopenia: the new geriatric giants. Rev Invest Clin 2016;68(2):59-67 [FREE Full text] [Medline: 27103041]
5. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, Writing Group for the European Working Group on Sarcopenia in Older People 2 (EWGSOP2),the Extended Group for EWGSOP2. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jan 01;48(1):16-31 [FREE Full text] [doi: 10.1093/ageing/afy169] [Medline: 30312372]
6. Woolford SJ, Sohan O, Dennison EM, Cooper C, Patel HP. Approaches to the diagnosis and prevention of frailty. Aging Clin Exp Res 2020 Sep 30;32(9):1629-1637 [FREE Full text] [doi: 10.1007/s40520-020-01559-3] [Medline: 32356135]
7. Huang CY, Lee WJ, Lin HP, Chen RC, Lin CH, Peng LN, et al. Epidemiology of frailty and associated factors among older adults living in rural communities in Taiwan. Arch Gerontol Geriatr 2020;87:103986. [doi: 10.1016/j.archger.2019.103986] [Medline: 31901458]
8. Kuo YH, Wang TF, Liu LK, Lee WJ, Peng LN, Chen LK. Epidemiology of sarcopenia and factors associated with it among community-dwelling older adults in Taiwan. Am J Med Sci 2019 Feb;357(2):124-133. [doi: 10.1016/j.amjms.2018.11.008] [Medline: 30665493]
9. Peng TC, Chen WL, Wu LW, Chang YW, Kao TW. The prevalence of frailty by the FRAIL-NH scale in Taiwan nursing home residents. J Nutr Health Aging 2020;24(5):507-511 [FREE Full text] [doi: 10.1007/s12603-020-1350-3] [Medline: 32346689]
10. Liu CJ, Latham NK. Progressive resistance strength training for improving physical function in older adults. Cochrane Database Syst Rev 2009 Jul 08;2009(3):CD002759 [FREE Full text] [doi: 10.1002/14651858.CD002759.pub2] [Medline: 19588334]
11. Chang CF, Yeh YL, Chang HY, Tsai SH, Wang JY. Prevalence and risk factors of sarcopenia among older adults aged ≥65 years admitted to daycare centers of Taiwan: using AWGS 2019 guidelines. Int J Environ Res Public Health 2021 Aug 05;18(16):8299 [FREE Full text] [doi: 10.3390/ijerph18168299] [Medline: 34444047]
12. Dodds R, Sayer AA. Sarcopenia and frailty: new challenges for clinical practice. Clin Med (Lond) 2016 Oct 03;16(5):455-458 [FREE Full text] [doi: 10.7861/clinmedicine.16-5-455] [Medline: 27697810]
13. Papa EV, Dong X, Hassan M. Resistance training for activity limitations in older adults with skeletal muscle function deficits: a systematic review. Clin Interv Aging 2017;12:955-961 [FREE Full text] [doi: 10.2147/CIA.S104674] [Medline: 28670114]
14. Dent E, Morley JE, Cruz-Jentoft AJ, Woodhouse L, Rodríguez-Mañas L, Fried LP, et al. Physical frailty: ICFSR international clinical practice guidelines for identification and management. J Nutr Health Aging 2019 Oct 03;23(9):771-787 [FREE Full text] [doi: 10.1007/s12603-019-1273-z] [Medline: 31641726]
15. Beckwée D, Delaere A, Aelbrecht S, Baert V, Beaudart C, Bruyere O, et al. Exercise interventions for the prevention and treatment of sarcopenia. A systematic umbrella review. J Nutr Health Aging 2019;23(6):494-502 [FREE Full text] [doi: 10.1007/s12603-019-1196-8] [Medline: 31233069]
16. Chen KM, Li CH, Huang HT, Cheng YY. Feasible modalities and long-term effects of elastic band exercises in nursing home older adults in wheelchairs: a cluster randomized controlled trial. Int J Nurs Stud 2016 Mar;55:4-14. [doi: 10.1016/j.ijnurstu.2015.11.004] [Medline: 26655368]
17. Meekes W, Stanmore EK. Motivational determinants of exergame participation for older people in assisted living facilities: mixed-methods study. J Med Internet Res 2017 Jul 06;19(7):e238 [FREE Full text] [doi: 10.2196/jmir.6841] [Medline: 28684385]

XSL•FO
RenderX

18. Adcock M, Thalmann M, Schättin A, Gennaro F, de Bruin ED. A pilot study of an in-home multicomponent exergame training for older adults: feasibility, usability and pre-post evaluation. Front Aging Neurosci 2019 Nov 22;11:304 [FREE Full text] [doi: 10.3389/fnagi.2019.00304] [Medline: 31824295]
19. Jahouh M, González-Bernal JJ, González-Santos J, Fernández-Lázaro D, Soto-Cámara R, Mielgo-Ayuso J. Impact of an intervention with Wii video games on the autonomy of activities of daily living and psychological-cognitive components in the institutionalized elderly. Int J Environ Res Public Health 2021 Feb 07;18(4):1570 [FREE Full text] [doi: 10.3390/ijerph18041570] [Medline: 33562249]
20. Ismail NA, Hashim HA, Ahmad Yusof H. Physical activity and exergames among older adults: a scoping review. Games Health J 2022 Feb 01;11(1):1-17. [doi: 10.1089/g4h.2021.0104] [Medline: 34851732]
21. Pacheco TB, de Medeiros CS, de Oliveira VH, Vieira ER, de Cavalcanti FA. Effectiveness of exergames for improving mobility and balance in older adults: a systematic review and meta-analysis. Syst Rev 2020 Jul 18;9(1):163 [FREE Full text] [doi: 10.1186/s13643-020-01421-7] [Medline: 32682439]
22. Chu CH, Biss RK, Cooper L, Quan AM, Matulis H. Exergaming platform for older adults residing in long-term care homes: user-centered design, development, and usability study. JMIR Serious Games 2021 Mar 09;9(1):e22370 [FREE Full text] [doi: 10.2196/22370] [Medline: 33687337]
23. Gunst M, de Meyere I, Willems H, Schoenmakers B. Effect of exergaming on wellbeing of residents in a nursing home: a single blinded intervention study. Aging Clin Exp Res 2022 Jan 22;34(1):151-157 [FREE Full text] [doi: 10.1007/s40520-021-01903-1] [Medline: 34156650]
24. Lu C, Hassan L, Buruk OO, Nummenmaa T, Peltonen J. “Switch” up your exercise: an empirical analysis of online user discussion of the ring fit adventure exergame. In: Proceedings of the 5th International GamiFIN Conference. 2021 Presented at: GamiFIN '21; April 7-9, 2021; Levi, Finland p. e2021 URL: https://www.researchgate.net/publication/352682130_Switch_Up_Your_Exercise_An_Empirical_Analysis_of_Online_User_Discussion_of_the_Ring_Fit_Adventure_Exergame
25. Tuan SH, Chang LH, Sun SF, Lin KL, Tsai YJ. Using exergame-based exercise to prevent and postpone the loss of muscle mass, muscle strength, cognition, and functional performance among elders in rural long-term care facilities: a protocol for a randomized controlled trial. Front Med (Lausanne) 2022;9:1071409 [FREE Full text] [doi: 10.3389/fmed.2022.1071409] [Medline: 36582297]
26. Using exergame-based program in elders of rural LTCF. ClinicalTrials. 2022. URL: https://clinicaltrials.gov/study/NCT05360667 [accessed 2024-02-20]
27. Garber CE, Blissmer B, Deschenes MR, Franklin BA, Lamonte MJ, Lee IM, American College of Sports Medicine. American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. Med Sci Sports Exerc 2011 Jul;43(7):1334-1359. [doi: 10.1249/MSS.0b013e318213fefb] [Medline: 21694556]
28. Chen GB, Lin CW, Huang HY, Wu YJ, Su HT, Sun SF, et al. Using virtual reality-based rehabilitation in sarcopenic older adults in rural health care facilities-a quasi-experimental study. J Aging Phys Act 2021 Oct 01;29(5):866-877 [FREE Full text] [doi: 10.1123/japa.2020-0222] [Medline: 33596540]
29. Chen LK, Liu LK, Woo J, Assantachai P, Auyeung TW, Bahyah KS, et al. Sarcopenia in Asia: consensus report of the Asian working group for Sarcopenia. J Am Med Dir Assoc 2014 Feb;15(2):95-101. [doi: 10.1016/j.jamda.2013.11.025] [Medline: 24461239]
30. Trampisch US, Franke J, Jedamzik N, Hinrichs T, Platen P. Optimal Jamar dynamometer handle position to assess maximal isometric hand grip strength in epidemiological studies. J Hand Surg Am 2012 Nov;37(11):2368-2373. [doi: 10.1016/j.jhsa.2012.08.014] [Medline: 23101534]
31. Ensrud KE, Ewing SK, Cawthon PM, Fink HA, Taylor BC, Cauley JA, Osteoporotic Fractures in Men Research Group. A comparison of frailty indexes for the prediction of falls, disability, fractures, and mortality in older men. J Am Geriatr Soc 2009 Mar 03;57(3):492-498 [FREE Full text] [doi: 10.1111/j.1532-5415.2009.02137.x] [Medline: 19245414]
32. Desrosiers J, Bravo G, Hébert R, Dutil É, Mercier L. Validation of the box and block test as a measure of dexterity of elderly people: reliability, validity, and norms studies. Arch Phys Med Rehabil 1994 Jul;75(7):751-755. [doi: 10.1016/0003-9993(94)90130-9]
33. Stark T, Walker B, Phillips JK, Fejer R, Beck R. Hand-held dynamometry correlation with the gold standard isokinetic dynamometry: a systematic review. PM R 2011 May 11;3(5):472-479. [doi: 10.1016/j.pmrj.2010.10.025] [Medline: 21570036]
34. Sanz-Paris A, González-Fernandez M, Hueso-Del Río LE, Ferrer-Lahuerta E, Monge-Vazquez A, Losfablos-Callau F, et al. Muscle thickness and echogenicity measured by ultrasound could detect local sarcopenia and malnutrition in older patients hospitalized for hip fracture. Nutrients 2021 Jul 13;13(7):2401 [FREE Full text] [doi: 10.3390/nu13072401] [Medline: 34371911]
35. Koca I, Boyaci A, Tutoglu A, Boyaci N, Ozkur A. The relationship between quadriceps thickness, radiological staging, and clinical parameters in knee osteoarthritis. J Phys Ther Sci 2014 Jun;26(6):931-936 [FREE Full text] [doi: 10.1589/jpts.26.931] [Medline: 25013299]
36. Wang J, Hu Y, Tian G. Ultrasound measurements of gastrocnemius muscle thickness in older people with sarcopenia. Clin Interv Aging 2018;13:2193-2199 [FREE Full text] [doi: 10.2147/CIA.S179445] [Medline: 30464428]

XSL•FO
RenderX

37. Norkin CC, White DJ. Measurement of Joint Motion: A Guide to Goniometry. 5th edition. Philadelphia, PA: FA Davis; 2016.
38. Satake S, Senda K, Hong YJ, Miura H, Endo H, Sakurai T, et al. Validity of the Kihon checklist for assessing frailty status. Geriatr Gerontol Int 2016 Jun 14;16(6):709-715. [doi: 10.1111/ggi.12543] [Medline: 26171645]
39. Tseng HM, Lu JF, Gandek B. Cultural issues in using the SF-36 health survey in Asia: results from Taiwan. Health Qual Life Outcomes 2003 Nov 26;1:72 [FREE Full text] [doi: 10.1186/1477-7525-1-72] [Medline: 14641915]
40. Tsai PH, Liu JL, Lin KN, Chang CC, Pai MC, Wang WF, et al. Development and validation of a dementia screening tool for primary care in Taiwan: brain health test. PLoS One 2018 Apr 25;13(4):e0196214 [FREE Full text] [doi: 10.1371/journal.pone.0196214] [Medline: 29694392]
41. Shah PB. Intention-to-treat and per-protocol analysis. CMAJ 2011 Apr 05;183(6):696 [FREE Full text] [doi: 10.1503/cmaj.111-2033] [Medline: 21464181]
42. Jeon S, Kim J. Effects of augmented-reality-based exercise on muscle parameters, physical performance, and exercise self-efficacy for older adults. Int J Environ Res Public Health 2020 May 07;17(9):3260 [FREE Full text] [doi: 10.3390/ijerph17093260] [Medline: 32392833]
43. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
44. Chen CF, Fu TH. Policies and transformation of long-term care system in Taiwan. Ann Geriatr Med Res 2020 Sep;24(3):187-194 [FREE Full text] [doi: 10.4235/agmr.20.0038] [Medline: 33012140]
45. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
46. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes 2006 Sep 27;4:70 [FREE Full text] [doi: 10.1186/1477-7525-4-70] [Medline: 17005038]
47. Schättin A, Pickles J, Flagmeier D, Schärer B, Riederer Y, Niedecken S, et al. Development of a novel home-based exergame with on-body feedback: usability study. JMIR Serious Games 2022 Dec 06;10(4):e38703 [FREE Full text] [doi: 10.2196/38703] [Medline: 36472900]
48. Suire KB, Spring K, Jones CS, Winkler DW, Carroll A, Merritt M, et al. Obesity and diabetes research get fit with ring fit adventure: how does a novel active video game compare to a traditional exercise protocol? Obes Diabetes Res 2022:1-3 [FREE Full text] [doi: 10.47275/2692-0964-117]
49. O'Bryan SJ, Giuliano C, Woessner MN, Vogrin S, Smith C, Duque G, et al. Progressive resistance training for concomitant increases in muscle strength and bone mineral density in older adults: a systematic review and meta-analysis. Sports Med 2022 Aug 24;52(8):1939-1960 [FREE Full text] [doi: 10.1007/s40279-022-01675-2] [Medline: 35608815]
50. Farrell E, Fox S. 149 power-agers: the effect of progressive resistance training (PRT) on sarcopenia status of community dwelling older adults. Age Ageing 2023;52(Supplement_3). [doi: 10.1093/ageing/afad156.013]
51. Onishi T, Fujiwara D, Abe T, Miyazaki R. A proposal concerning exercise intensity with the Nintendo Ring Fit Adventure Exergame among older adults: a preliminary study. J Digit Life 2023:3. [doi: 10.51015/jdl.2023.3.12]
52. McDermott AY, Mernitz H. Exercise and older patients: prescribing guidelines. Am Fam Physician 2006 Aug 01;74(3):437-444 [FREE Full text] [Medline: 16913163]
53. Brook MS, Wilkinson DJ, Phillips BE, Perez-Schindler J, Philp A, Smith K, et al. Skeletal muscle homeostasis and plasticity in youth and ageing: impact of nutrition and exercise. Acta Physiol (Oxf) 2016 Jan 21;216(1):15-41 [FREE Full text] [doi: 10.1111/apha.12532] [Medline: 26010896]
54. Moritani T, deVries HA. Potential for gross muscle hypertrophy in older men. J Gerontol 1980 Sep 01;35(5):672-682. [doi: 10.1093/geronj/35.5.672] [Medline: 7430562]
55. Clark DJ. Automaticity of walking: functional significance, mechanisms, measurement and rehabilitation strategies. Front Hum Neurosci 2015;9:246 [FREE Full text] [doi: 10.3389/fnhum.2015.00246] [Medline: 25999838]
56. Granacher U, Lacroix A, Muehlbauer T, Roettger K, Gollhofer A. Effects of core instability strength training on trunk muscle strength, spinal mobility, dynamic balance and functional mobility in older adults. Gerontology 2013 Oct 24;59(2):105-113. [doi: 10.1159/000343152] [Medline: 23108436]
57. Park J, Gong J, Yim J. Effects of a sitting boxing program on upper limb function, balance, gait, and quality of life in stroke patients. NeuroRehabilitation 2017;40(1):77-86. [doi: 10.3233/NRE-161392] [Medline: 27792020]
58. Campbell E, Petermann-Rocha F, Welsh P, Celis-Morales C, Pell JP, Ho FK, et al. The effect of exercise on quality of life and activities of daily life in frail older adults: a systematic review of randomised control trials. Exp Gerontol 2021 May;147:111287 [FREE Full text] [doi: 10.1016/j.exger.2021.111287] [Medline: 33609689]
59. Martínez-Hernández BM, Rosas-Carrasco O, López-Teros M, González-Rocha A, Muñoz-Aguirre P, Palazuelos-González R, et al. Association between physical activity and physical and functional performance in non-institutionalized Mexican older adults: a cohort study. BMC Geriatr 2022 May 03;22(1):388 [FREE Full text] [doi: 10.1186/s12877-022-03083-7] [Medline: 35505279]

XSL•FO
RenderX

60. Sunde S, Hesseberg K, Skelton DA, Ranhoff AH, Pripp AH, Aarønæs M, et al. Associations between health-related quality of life and physical function in older adults with or at risk of mobility disability after discharge from the hospital. Eur Geriatr Med 2021 Dec;12(6):1247-1256 [FREE Full text] [doi: 10.1007/s41999-021-00525-0] [Medline: 34106446]
61. Liao YY, Chen IH, Lin YJ, Chen Y, Hsu WC. Effects of virtual reality-based physical and cognitive training on executive function and dual-task gait performance in older adults with mild cognitive impairment: a randomized control trial. Front Aging Neurosci 2019 Jul 16;11:162 [FREE Full text] [doi: 10.3389/fnagi.2019.00162] [Medline: 31379553]
62. Yen HY, Chiu HL. Virtual reality exergames for improving older adults' cognition and depression: a systematic review and meta-analysis of randomized control trials. J Am Med Dir Assoc 2021 May;22(5):995-1002. [doi: 10.1016/j.jamda.2021.03.009] [Medline: 33812843]
63. Lee WJ, Peng LN, Lin CH, Chen RC, Lin SZ, Loh CH, Taiwan Integrated Geriatric Care Study Group. Effects of incorporating multidomain interventions into integrated primary care on quality of life: a randomised controlled trial. Lancet Healthy Longev 2021 Nov;2(11):e712-e723 [FREE Full text] [doi: 10.1016/S2666-7568(21)00248-8] [Medline: 36098028]

Abbreviations

ACSM: American College of Sports Medicine
ADL: activity of daily living
ASMMI: appendicular skeletal muscle mass index
CONSORT: Consolidated Standards of Reporting Trials
exergame-RFA: exergame-based exercise training program delivered via Ring Fit Adventure
EXPPLORE: Using an Exergame to Prevent and Postpone the Loss of Muscle Mass, Muscle Strength, and Functional Performance in Rural Elders
HGS: handgrip strength
HRQoL: health-related quality of life
LTCF: long-term care facility
MCID: minimal clinically important difference
MVIC: maximum voluntary isometric contraction
PRT: progressive resistance training
RFA: Ring Fit Adventure
RPE: rating of perceived exertion
RT: resistance training
SF-36: Short Form Health Survey-36
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials

Edited by S Ma; submitted 12.04.24; peer-reviewed by KL Lin, IH Liou; comments to author 26.04.24; revised version received 03.05.24; accepted 24.05.24; published 18.07.24.

Please cite as:
Tuan SH, Chang LH, Sun SF, Li CH, Chen GB, Tsai YJ
Assessing the Clinical Effectiveness of an Exergame-Based Exercise Training Program Using Ring Fit Adventure to Prevent and Postpone Frailty and Sarcopenia Among Older Adults in Rural Long-Term Care Facilities: Randomized Controlled Trial
J Med Internet Res 2024;26:e59468
URL: https://www.jmir.org/2024/1/e59468
doi:10.2196/59468
PMID:39024000

©Sheng-Hui Tuan, Lin-Hui Chang, Shu-Fen Sun, Chien-Hui Li, Guan-Bo Chen, Yi-Ju Tsai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Immersive Technology–Based Education for Undergraduate Nursing Students: Systematic Review and Meta-Analysis Using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) Approach

Subin Park[1], MSN; Hui Ju Shin[1,2], BSN; Hyoeun Kwak[1], MSN; Hyun Joo Lee[3,4], PhD

[1]College of Nursing, Yonsei University, Seoul, Republic of Korea

[2]Severance Hospital, Yonsei University Health System, Seoul, Republic of Korea

[3]Mo-Im Kim Nursing Research Institute, College of Nursing, Seoul, Republic of Korea

[4]Yonsei Evidence-Based Nursing Centre of Korea: A Joanna Briggs Institute-Affiliated Group, Seoul, Republic of Korea

Corresponding Author:
Hyun Joo Lee, PhD
Mo-Im Kim Nursing Research Institute
College of Nursing
Seodaemun-gu
50-1, Yonsei-ro
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 3282
Fax: 82 2 2227 8303
Email: leehj1509@yuhs.ac

Abstract

Background: The adoption of immersive technology in simulation-based nursing education has grown significantly, offering a solution to resource limitations and enabling safe access to clinical environments. Despite its advantages, there are still diverse reports regarding the effectiveness of immersive technology. It is crucial to verify the effectiveness of immersive technology in nursing education to inform future educational programs.

Objective: This systematic review aimed to identify the contents of immersive technology–based education for undergraduate nursing students and evaluate the effectiveness of immersive technology compared to traditional teaching methods.

Methods: A literature search was performed using 4 databases: PubMed, CINAHL, Embase, and Web of Science; the latest search was completed on January 19, 2023. The inclusion criteria were as follows: participants were undergraduate nursing students; studies were published in Korean or English; designs included randomized controlled trials (RCTs) or nonrandomized studies; and interventions involved virtual reality (VR), augmented reality (AR), mixed reality, or extended reality. Quality assessment was conducted using Cochrane Risk-of-Bias Tool version 2 for RCTs and the Risk-of-Bias Assessment Tool for Nonrandomized Studies. The main outcomes of the included studies were classified according to the New World Kirkpatrick Model (NWKM), ranging from level 1 (reaction) to level 4 (results). Meta-analysis was conducted using RevMan 5.4 software, and subgroup analysis was conducted due to heterogeneity of the results of the meta-analysis. The Grading of Recommendations, Assessment, Development, and Evaluation approach was adopted for assessing certainty and synthesizing results of the relevant literature.

Results: A total of 23 studies were included, with participant numbers ranging from 33 to 289. Of these, 19 (82.6%) studies adopted VR to simulate various nursing scenarios, including disaster training, resuscitation, health assessments, and home health care; 4 (17.4%) studies used AR technologies; and 15 (65.2%) studies involved virtual patients in their scenarios. Based on the NWKM, the main outcome variables were classified as level 1 (usability and satisfaction), level 2 (knowledge, motivation, confidence, performance, attitude, and self-efficacy), and level 3 (clinical reasoning); level 4 outcomes were not found in the selected studies. Results of the subgroup analysis showed that immersive technology–based nursing education is more effective than traditional education in knowledge attainment (standard mean difference [SMD]=0.59, 95% CI 0.28-0.90, $P<.001$, $I^2=49\%$).

XSL•FO
RenderX

Additionally, there were significant difference differences between the experimental and control group in confidence (SMD=0.70, 95% CI 0.05-1.35, P=.03, I^2=82%) and self-efficacy (SMD=0.86, 95% CI 0.42-1.30, P<.001, I^2=63%).

Conclusions: These findings support the effectiveness of immersive technology–based education for undergraduate nursing students, despite heterogeneity in methods and interventions. We suggest that long-term cohort studies be conducted to evaluate the effects of immersive technology–based nursing education on NWKM level 4.

(J Med Internet Res 2024;26:e57566) doi:10.2196/57566

KEYWORDS

nursing education; nursing students; immersive technology; systematic review; meta-analysis; virtual reality; augmented reality; extended reality; simulation-based learning; medical education

Introduction

Background

Immersive technology is widely recognized to improve learning in nursing education [1,2]. The idea of immersive technology emerged 6 decades ago with the human-computer prototype known as the "Man-Machine Graphical Communication System" [3]. Immersive technology is derived from the reality-virtuality continuum concept and encompasses virtual reality (VR), augmented reality (AR), and mixed reality (MR) [2]. "Extended reality" (XR), an umbrella term that includes all 3 technologies, is recognized as a type of immersive technology [4]. VR refers to technology that enables users to immerse themselves in virtual worlds and interact with them, while AR involves adding virtual elements to the real environment to merge reality with virtuality. MR integrates VR and AR to provide an experience where the real and virtual environments interact, and the degree of immersion is related to the stimulated senses, interactions, and similarities between reality and virtuality [5]. According to Cipresso et al [5], VR is classified into 3 technology types based on the degree of immersion: nonimmersive technology involves using desktop computers to reproduce images of virtual worlds, semi-immersive technology uses perspective projection to provide stereo images of 3D scenes viewed on a monitor, and immersive technology represents the highest level of technical immersion, providing users with a sense of presence in virtual environments [5].

A significant goal of nursing education is the transfer of theoretical knowledge to clinical practice [6]. However, limited clinical resources impact students' opportunities to gain hands-on experience with patients, and the lack of hands-on experience in clinical practice may pose a risk of patient safety when the students face challenges in a real-world clinical environment in the future as health professionals [7,8]. Therefore, nursing educators should provide students with sufficient alternative clinical experiences [7]. To ensure the quality of nursing education in clinical practice, educators have incorporated various educational strategies, including simulations [9]. Various clinical simulation methods have been developed [10]. In the nursing field, there is a growing interest in using immersive technology as an effective educational tool for simulation-based programs to enhance students' knowledge and skill acquisition [4,11].

Immersive technology reduces the limitations of the resources required for 2D simulation–based learning [12]. Education programs adapted to immersive technology enable students to access clinical practice with ease and develop their skills within a secure setting, minimizing risks to patient safety [12,13]. According to Foronda et al [14], 98% of participants expressed a preference for incorporating virtual learning environments. There have been efforts to further the leverage of immersive technology, especially with the increased significance of remote classes due to the outbreak of COVID-19 [15,16]. Additionally, because immersive technology improves the interaction between students and instructors by facilitating discussions, it is frequently used in simulation-based learning [17]. Student-instructor interaction helps derive successful outcomes when properly supported with high-fidelity simulations [18].

The New World Kirkpatrick Model

The Kirkpatrick Model, developed by Donald L Kirkpatrick in 1959 and expanded in 1967, is a widely used framework for evaluating the effectiveness of educational programs. This model categorizes program outcomes into 4 levels [19]. Level 1 encompasses participant reactions, assessing how favorable, engaging, and relevant they find the training to be in relation to their jobs. Level 2 includes the learning outcomes; at this level, the focus is on the knowledge, skills, attitude, confidence, and commitment acquired by learners because of training. Level 3 evaluation is related to changes in the participants' behavior based on the simulation experience. Critical behavior must have a few key actions that are performed consistently by the primary group to bring about the targeted outcome. Level 4 is the final outcome evaluation, which indicates the actual changes in the output or results due to the training. In 2010, the New World Kirkpatrick Model (NWKM) emerged, presenting a framework with 4 levels of evaluation that is more effectively applicable to the current changing circumstances [19]. It modifies the direction of the result levels in reverse, in the order of levels 4-1 [19,20]. The NWKM proposes planning eventual program outcomes in the planning stage. Some outcomes have been added to each level of evaluation, and parts of the definitions have been revised. Both quantitative and qualitative methods can be used to evaluate each level, and this model has been widely used to evaluate the outcomes of education programs in the nursing field [21-23].

Study Objective

Although there are evident advantages to using immersive technology in nursing education, there are claims suggesting that it may not be notably effective compared to traditional teaching methods, such as didactic lectures, use of audiovisual

XSL•FO
RenderX

materials, and students' practice following the instructor's demonstration [9,24-28]. In addition, many studies have verified the effectiveness of VR methods in nursing education [9,29,30]; however, there remains a shortage of studies that comprehensively assess the effectiveness of immersive technology encompassing all concepts of VR, AR, MR, and XR. Therefore, this study aimed to identify the contents of immersive technology–based education programs for undergraduate nursing students and evaluate the effectiveness of the interventions.

Methods

Reporting Guidelines

This study adhered to the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines [31]. The study protocol was preregistered in the International Prospective Register of Systematic Reviews (PROSPERO; registration number CRD42023400085).

Eligibility Criteria

Eligibility criteria were established based on the Population, Intervention, Comparison, and Outcome (PICO) framework. The target population was undergraduate nursing students. The search for studies included randomized controlled trials (RCTs) and nonrandomized studies that used VR, AR, XR, and MR technologies. Regarding the immersion of VR technology, the search specifically focused on studies that used head-mounted devices (HMDs), including glasses, goggles, and helmets—the most immersive and extensively used visual devices in VR technology [5]. Other immersive technologies, such as AR, XR, and MR, encompassed all devices such as smartphones and smart glasses [5]. The outcome variables were not restricted to the search and were categorized according to the NWKM [19]. Theses and dissertations, along with studies not published in either English or Korean, and those designed as pilot studies or case studies were excluded.

Search Strategy

A thorough search was conducted across 4 databases: PubMed, CINAHL, Embase, and Web of Science. Search terms were selected judiciously, adhering to the principles of the Medical Subject Headings (MeSH), with specific terms customized for each database (Multimedia Appendix 1). Additionally, consultation with a librarian at the medical library informed and refined the search strategy. The search was conducted on January 19, 2023. Search records were imported into the reference management tools EndNote (Clarivate) and Covidence, a specialized program for systematic reviews. Following the application of Covidence's artificial intelligence (AI) function to automatically identify and remove duplicate studies, manual confirmation was performed. After eliminating duplicates, the remaining studies underwent eligibility screening by 2 independent reviewers (authors SP and HJS) according to predefined inclusion and exclusion criteria. In cases of discrepancies between the 2 reviewers, a third reviewer (author HK) was consulted to reach a consensus. Finally, all researchers agreed on the final literature to be included in the analysis. The initial screening process involved the assessment of titles and abstracts for relevance. Subsequently, full-text screening was performed, and the rationale for exclusion was documented in the PRISMA flowchart.

Quality Assessment

The quality assessment of all RCTs was performed based on Cochrane Risk-of-Bias Tool version 2 (RoB 2) [32]. RoB 2 consists of 5 key domains that evaluate potential biases in study design and conduct. These domains are related to the randomization process, intended intervention, absence of outcome data, and selective reporting of results. Individual domains were assessed for their potential influence on the validity of the findings. The quality assessment of nonrandomized studies was based on the Risk-of-Bias Assessment Tool for Nonrandomized Studies (RoBANS) [33]. The domains encompass issues such as bias stemming from participant selection, confounding variables, measurement, blinding, incomplete outcome data, and selective reporting of results. Two reviewers (SP and HJS) independently assessed the quality of each of the 23 included studies. For the assessment results of 22 (95.7%) studies, there was agreement between the 2 reviewers; however, 1 (4.3%) study on which agreement was not reached between the 2 reviewers was reassessed by a third reviewer (HK).

Data Extraction

An exclusive data extraction template was used to collect pertinent details from each study: author, publication year, country, research design, participants, and sample size. Regarding immersive technology interventions, the extracted items included the type of technology, content characteristics, the length and duration of the intervention, facilitator details, the presence of prebriefing and debriefing sessions, scenarios, and the VR content development company. The data extraction template was filled in independently by 2 reviewers (SP and HJS). For disagreements, the third reviewer (HK) reevaluated the papers and facilitated consensus among the researchers based on clear evidence. Finally, a fourth reviewer (author HJL) reviewed and confirmed the overall content. For outcome variables, the extracted information included the measurement timing, evaluated variables, measurement tools, and classification based on the NWKM, and mean (SD) values were extracted for subsequent meta-analysis. Outcome variables were classified into the 4 levels of the NWKM: level 4 (whether the organization exists to perform, deliver, or contribute to its customers or society at a high level), level 3 (critical behaviors, required drivers, and on-the-job learning), level 2 (confidence and commitment highlighted to close the gap between learning and behavior, along with intended knowledge, skills, and attitude), and level 1 (measures such as satisfaction and usability to assess the extent to which participants perceived the education as positive, engaging, and relevant) [19].

Statistical Analysis

RevMan 5.4 software was used to synthesize the data [34]. The overall effect size was calculated using the SMD, along with the 95% CI, as the studies used different measurements for each outcome. For pre- and posttests, if changes in the measurement variables were not reported, the correlation between the 2

XSL•FO
RenderX

covariances was assumed to be 0.5 [35], and the changes and measurement variances of the variables were reported accordingly. The effect size was classified into small (0.2), medium (0.5), large (0.8), or very large (1.2) based on Cohen's guidelines. To analyze the overall effect, Z-statistics were applied at a significance level of $P<.05$. Heterogeneity was estimated using the Higgins I^2 statistic, which provided insight into the degree of variation among the included studies. Heterogeneity can be interpreted as nonobserved (0%), low (0%-25%), moderate (25%-50%), or high (>50%) [36]. The random effects model was used for data analysis due to the presence of heterogeneity in the mediation process across various scenarios and measurement variables [37]. The results of the meta-analysis were presented as forest plots.

Assessment of Certainty of Evidence

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach was used to assess the quality of evidence. This involved evaluating the study design, risk of bias, inconsistency, indirectness, imprecision, and other relevant factors. Based on these evaluations, the quality of evidence was rated on a 4-point scale: high, moderate, low, or very low [38].

Results

Study Selection

Figure 1 shows the process of study selection based on the PRISMA 2020 flow diagram. A total of 3204 studies were identified by searching the 4 databases. The AI function of Covidence automatically removed 1534 (47.9%) duplicates. In addition, 18 (0.6%) studies were manually identified as duplicates. Of the remaining 1652 (51.6%) papers, 1546 (93.6%) were excluded after a thorough review of their titles and abstracts. The full text of the remaining 106 (6.4%) papers was screened, and 83 (78.3%) papers were excluded for the following reasons: not related to immersive technology (eg, VR not using an HMD; n=46, 55.4%), incorrect study design (n=22, 26.5%), unavailable full text (n=10, 12%), not including undergraduate nursing students (n=2, 2.4%), not in Korean or English (n=1, 1.2%), and theses or dissertations (n=2, 2.4%). Finally, 23 studies (n=22, 95.7%, in English and n=1, 4.2%, in Korean) were selected.

Figure 1. PRISMA flow diagram of the study selection. HMD: head-mounted device; PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis.

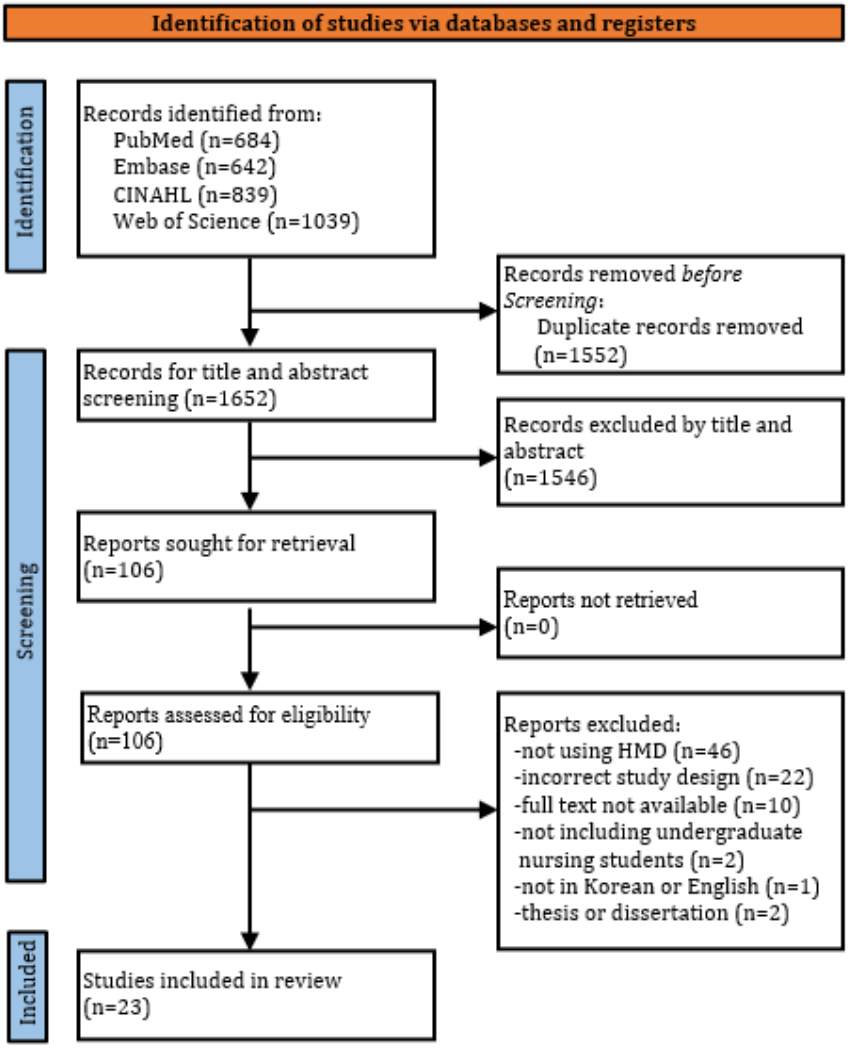

XSL•FO
RenderX

Study Characteristics

Table 1 presents the study characteristics. The studies included in this research were conducted in 8 countries: China [39], Finland [40], Norway [41,42], South Korea [43-48], Spain [49,50], Taiwan [51-55], Turkey [56,57], and the United States [58-61]. All studies were published within the past 5 years, except for 1 (4.3%) study [60] published in 2018; notably, there was a significant increase in the number of studies published in 2021 (n=10, 43.5%). The number of participants ranged from a minimum of 33 to a maximum of 289. Of the 23 studies, 21 (91.3%) were conducted with nursing students as the primary participants, whereas the remaining 2 (8.7%) specifically targeted both medical and nursing students. In addition, 6 (26.1%) studies were designed as RCTs, while 17 (73.9%) were nonrandomized studies. Furthermore, 4 (17.4%) studies [40,44,51,54] used AR, whereas the remaining 19 (82.6%) used VR in their educational programs. The experimental group participated in VR or AR simulation programs, whereas the control group received traditional education or no intervention.

XSL•FO
RenderX

Table 1. General characteristics of included studies (N=23).

Author	Country	Type of immersive technology	Study design	Participants	Total sample size, N
Shujuan et al [39]	China	VR[a]	RCT[b]	Second-year nursing students of a tertiary program, registered for the disaster nursing course	101
Havola et al [40]	Finland	VR	Nonrandomized study	Graduating nursing students from a single university	40
Berg and Steinsbekk [41]	Norway	VR	RCT	First-year medical/nursing students	289
Berg and Steinsbekk [42]	Norway	VR	RCT	First-year medical/nursing students	289
Ahn and Lee [43]	South Korea	VR	Nonrandomized study	Third-year nursing students	84
Jung and Park [44]	South Korea	VR	Nonrandomized study	Second-, third-, and fourth-year nursing students	60
Lee and Han [45]	South Korea	VR	Nonrandomized study	Fourth-year nursing students	60
Yang and Oh [46]	South Korea	VR	Nonrandomized study	Prelicensure nursing students	83
Yu et al [47]	South Korea	VR	Nonrandomized study	Senior nursing students	50
Yu and Yang [48]	South Korea	VR	Nonrandomized study	Third- and fourth-year nursing students	50
Rodríguez-Abad et al [49]	Spain	AR	Nonrandomized study	Second-year nursing students	137
Mayor Silva et al [50]	Spain	VR	RCT	First-year students from the Faculty of Nursing	100
Chang et al [51]	Taiwan	VR	Nonrandomized study	Nursing students from a single nursing college	64
Chao et al [52]	Taiwan	VR	RCT	Nursing students	45
Chen et al [53]	Taiwan	VR	Nonrandomized study	Third-year nursing students	79
Chen and Liou [54]	Taiwan	AR[c]	RCT	Fourth-year nursing students	95
Wu et al [55]	Taiwan	VR	Nonrandomized study	Third-year nursing students from a single university, 9 pediatric classes	105
Kurt and Öztürk [56]	Turkey	AR	Nonrandomized study	First-year nursing students	122
Sen et al [57]	Turkey	VR	Nonrandomized study	Second-year nursing students taking the operating room nursing course	40
Dang et al [58]	United States	VR	Nonrandomized study	Prelicensure, baccalaureate nursing students in their first medical-surgical course	160
Herbert et al [59]	United States	AR	Nonrandomized study	Second-semester junior nursing students	33
Smith et al [60]	United States	VR	Nonrandomized study	Senior baccalaureate nursing students, recruited from 4 different Midwest university campuses	172
Smith et al [61]	United States	VR	Nonrandomized study	Senior nursing students in the final semester of a baccalaureate nursing program	121

[a]VR: virtual reality.

[b]RCT: randomized controlled trial.

[c]AR: augmented reality.

XSL•FO
RenderX

Intervention Characteristics

Scenario Features

The intervention scenarios covered a wide range of nursing situations, including disaster training [5,48,61], resuscitation [40,46,54], nursing education [44,45,47,50-52,55,57,58], health assessments [53], and home health care nursing [43]. The virtual locations for the intervention scenarios varied, including settings such as the patient's home [43] and clinical environments, such as general wards [52,55], the emergency room [60], intensive care units (ICUs) [40,45], neonatal intensive care units (NICUs) [46,47], delivery rooms [51], angiography rooms [44], operating rooms [57], and isolation units [48]. Of the 23 studies, 15 (65.2%) [39-48,53-55,60,61] mentioned that they featured virtual patients. Of these, 7 (46.7%) studies [39,41-43,54,55,58] allowed for interaction between the virtual patient and the learner. This interaction involved assessing the virtual patient's health status through the airway, breathing, circulation, disability, and exposure (ABCDE) approach [41,42] or providing nursing interventions following the assessment of the patient's condition [39,43,54,55]. However, in 1 (14.3%) study [58], it was unclear whether there was any interaction between the virtual patient and the learner.

Implementation of Immersive Technology

Various devices were used to operate the immersive technology. Of 19 (82.6%) studies that used VR technology, 11 (57.9%) [39,40,44-46,50-52,55,57,58] operated the scenario using only HMDs (ie, VR goggles, glasses, headsets, and helmets) with embedded controllers, while 9 (47.4%) studies [41-43,46-48,53,60,61] used haptic devices and motion trackers for controlling and tracking their motions in a virtual environment. The remaining 4 (17.4%) studies [49,54,56,59] used AR technology operated with smartphones or tablets so that they could augment fidelity via lenses and screens.

Administration of Immersive Technology–Based Education

Regarding learning methods, 1 (4.3%) study [53] used immersive technology in a 15-week classroom lecture, and 1 (4.3%) study [42] used a team-based approach. The other studies conducted interventions independently. The length of the scenarios ranged from 8 to 110 minutes, with some studies not specifying a time limit or providing explicit information regarding the scenario length. In addition, 10 (43.5%) studies [43-48,50,55,58,60] included both prebriefing and debriefing sessions, 7 (30.4%) studies [39,40,52-54,59,61] included only prebriefing sessions, 2 (8.7%) studies [49,56] included only debriefing sessions, and 1 (4.3%) study [51] included neither prebriefing nor debriefing. Furthermore, 10 (43.5%) studies [43,44,50-54,56,58,59] did not describe the role of the instructor. In the remaining studies, the instructor provided minimal intervention, offering only technical support when learners engaged with immersive technology.

Outcome Variables

The outcome variables of each study were classified using the NWKM [19], as shown in Table S1 in Multimedia Appendix 2. The outcomes included satisfaction [41,42,44,45,47,48,51,52,59-61], usability [41,42,58], a sense of realism [46,58], anxiety [46], knowledge [39,41-48,50-60], confidence [39,43,52,54,57,61], self-efficacy [43,45,47,48,61], performance [39,43,45,46,48-50,53,54,56,57,60,61], attitude [44,51], motivation [44,46,49,51], critical thinking [51], and clinical reasoning [40,46]. The outcomes were categorized into NWKM levels 1-3; no outcome variable corresponded to level 4.

Risk of Bias

Figures 2 and 3 show the results of risk-of-bias assessment. RoB 2 was used to appraise 6 RCTs [39,41,42,50,51,53], of which 4 (66.7%) [41,42,50,52] confirmed all components to have low risk, while for the other 2 (33.3%) studies, the risk of bias was unclear for the items of missing outcome data [54] and deviation from the intended intervention [39]. RoBANS was used to assess the quality of 17 (73.9%) nonrandomized studies [40,43-49,51,53,55-61], of which 15 (88.2%) [40,43-49,51,55-60] had a high risk of bias in the measurement of the intervention. They used self-reported methods to assess outcome variables. In the incomplete outcome data category, 15 (88.2%) studies reported a low dropout rate, making them suitable for classification under a low risk of bias, while 2 (11.8%) studies [43,59] were categorized as having an unclear risk of bias. One nonrandomized study [51] was reevaluated to resolve discrepancies in quality assessment, ultimately reaching a consensus.

XSL•FO
RenderX

Figure 2. Risk-of-bias summary of RCTs. RCT: randomized controlled trial.

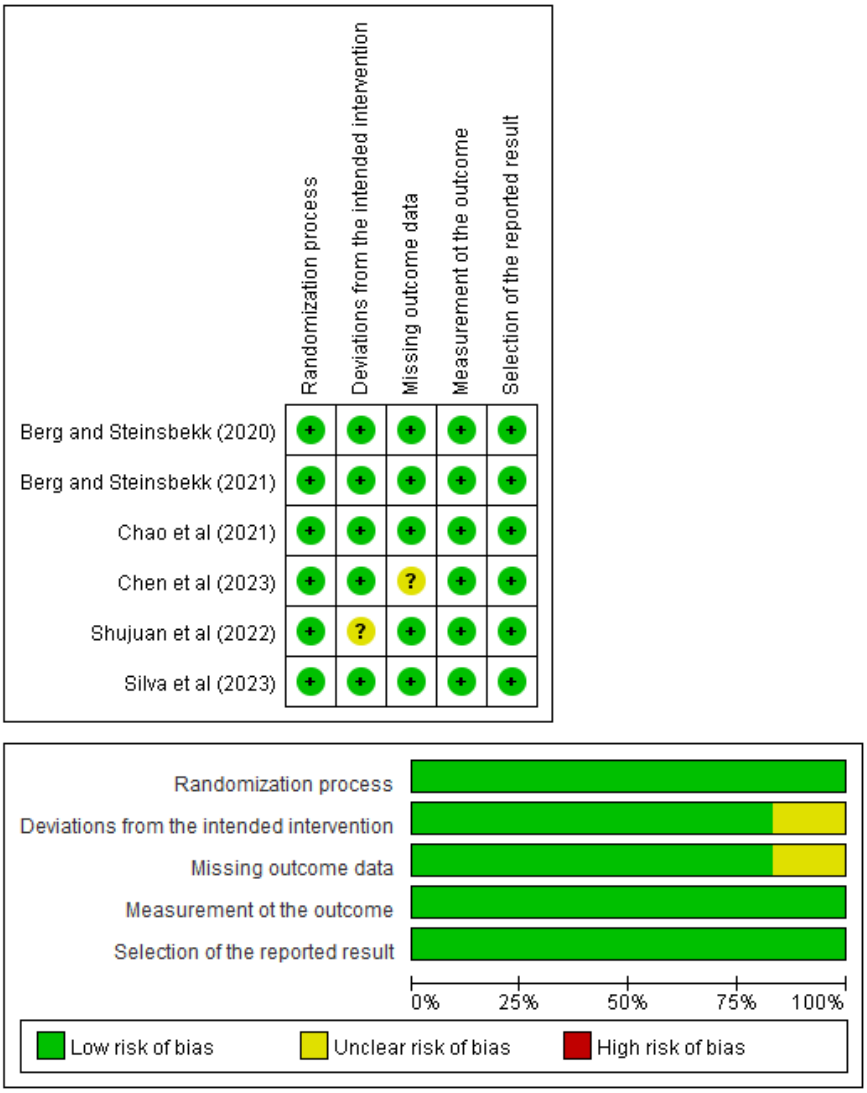

XSL•FO
RenderX

Figure 3. Risk-of-bias summary of nonrandomized studies.

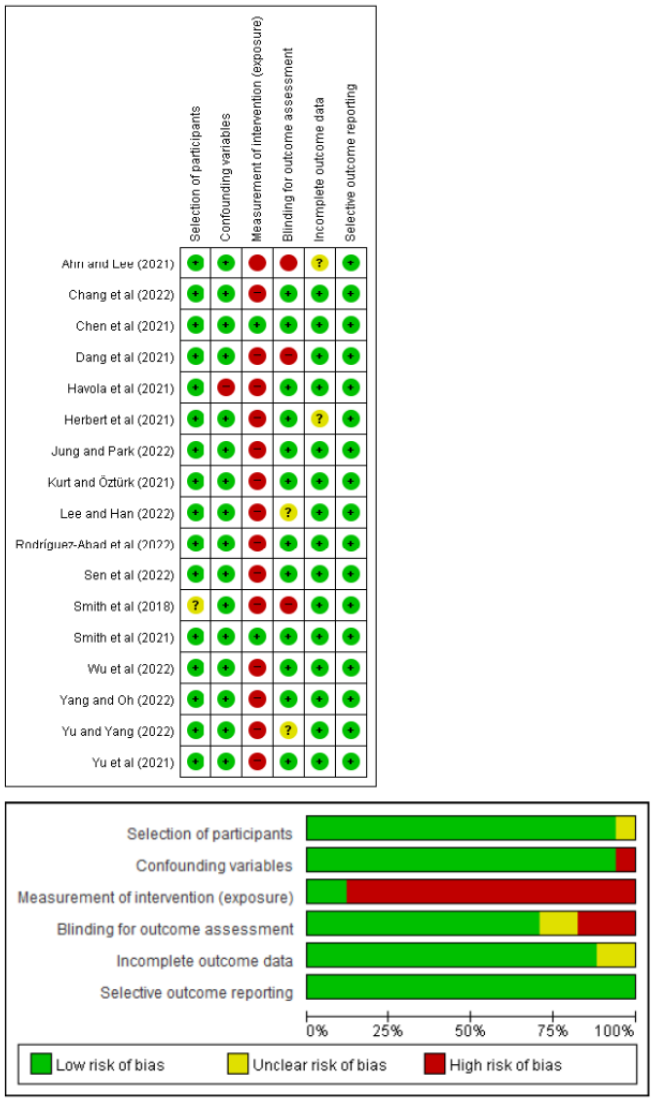

Meta-Analysis Findings

Effects of Immersive Technology–Based Education on Students' Knowledge Attainment

Of the 23 studies, 14 (60.9%) studies [39,43-48,50,52-55,58,60] involving 553 students were analyzed to evaluate the effects of immersive technology–based education on students' knowledge attainment. Of the 14 studies, 10 (71.4%) [39,40,43,45-48,54,55,58] used nonrandomized design and the other 4 (28.6%) [39,50,52,53] used RCTs. The findings revealed that the experimental group significantly enhanced students' knowledge (SMD=0.71, 95% CI 0.37-1.06, *P*<.001); however, significant heterogeneity (I^2=87%, *P*<.001) was observed between the 14 (60.9%) studies. To address this heterogeneity, subgroup analysis was conducted based on study design to distinguish between RCTs and nonrandomized studies. In the RCTs, a reduction in heterogeneity was observed (SMD=0.59, 95% CI 0.28-0.90, *P*<.001, I^2=49%, *P*=.12). Further subgroup analysis was performed on the nonrandomized studies based on the type of equipment used. The results showed that studies using only HMDs did not significantly enhanced students' knowledge (SMD=0.46, 95% CI –0.06 to 0.99, *P*=.09, I^2=82%, *P*<.001), while studies using both HMDs and controllers significantly enhanced students' knowledge (SMD=0.99, 95% CI 0.27-1.71, *P*=.007, I^2=92%, *P*<.001), as shown in Figure 4. Sensitivity tests were additionally carried out for nonrandomized studies using HMDs and controllers, except 1 (4.3%) study [60] published in 2018. The results are shown in Figure 5, and I^2 reduced to 73%. However, the test for subgroup differences indicated no statistically significant subgroup effect (*P*=.84), implying that the study design does not modify the effect of knowledge attainment. Nevertheless, a fewer number of trials provided data for the RCT subgroup compared to the nonrandomized study subgroup, potentially limiting the ability of the analysis to detect subgroup disparities (Figure 5).

XSL•FO
RenderX

Figure 4. Forest plot of immersive technology–based education on students' knowledge acquisition. HMD: head-mounted device.

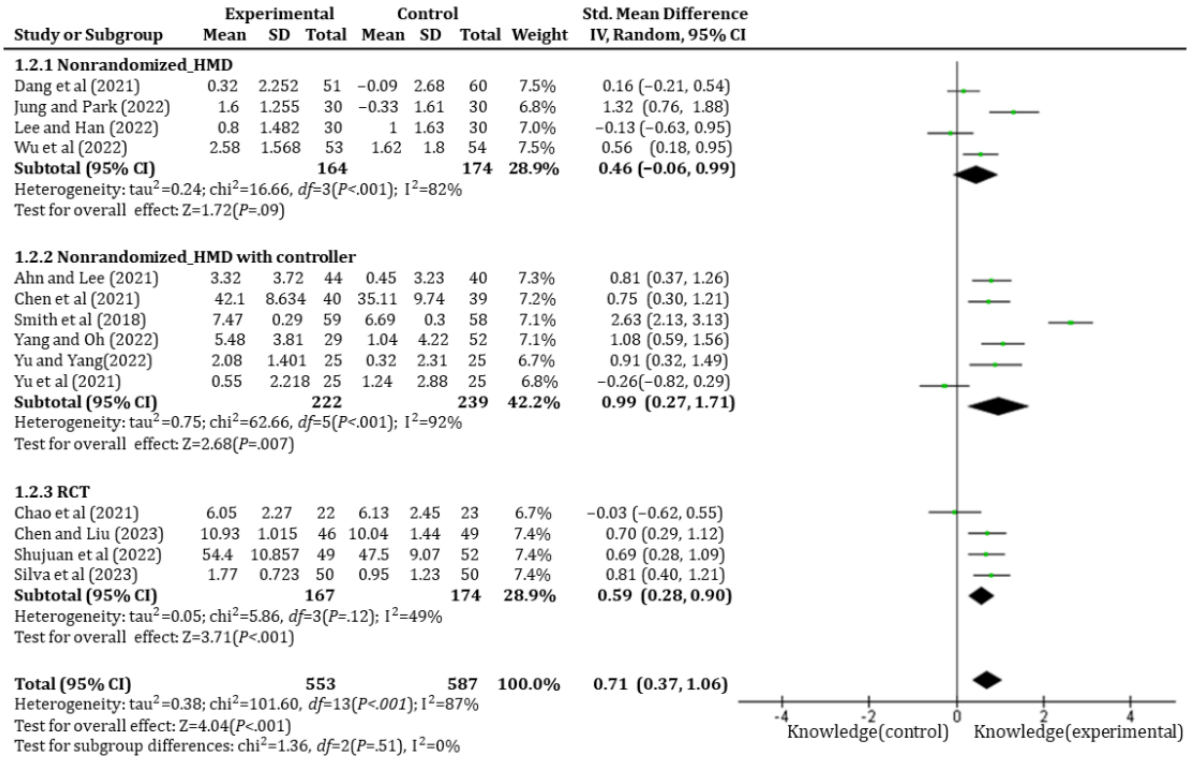

Figure 5. Sensitivity analysis results of in nonrandomized studies using HMDs with controllers. HMD: head-mounted device.

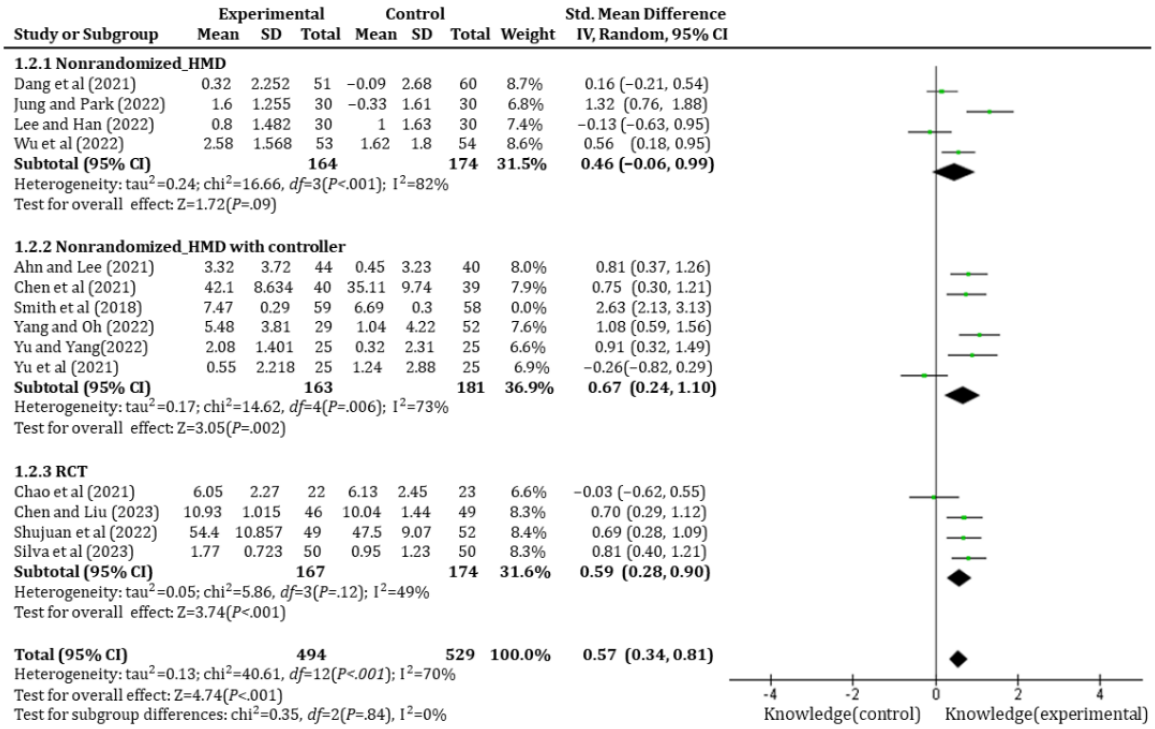

Effects of Immersive Technology–Based Education on Students' Confidence

Of the 23 studies, 3 (13%) studies [39,52,53] involving 117 students were analyzed to evaluate the effects of immersive technology–based education on students' confidence. All 3 (13%) studies were designed as RCTs and used VR. The findings revealed that compared to control conditions, the interventions for the experimental group significantly enhanced students' confidence with a medium effect size (SMD=0.70, 95% CI 0.05-1.35, *P*=.03). The heterogeneity test showed a high level of heterogeneity across the studies (I^2=82%, *P*<.001), as shown in Figure 6.

XSL•FO
RenderX

Figure 6. Forest plot of immersive technology–based education on students' confidence.

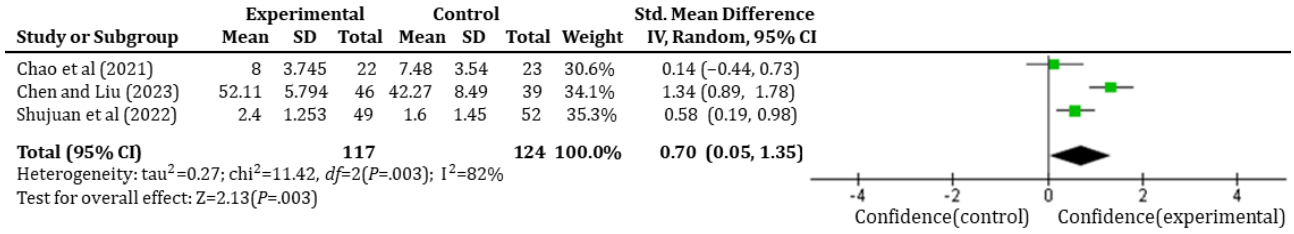

Effects of Immersive Technology–Based Education on Students' Self-Efficacy

Of the 23 studies, 4 (17.4%) studies [43,45,47,48] involving 120 students were analyzed to evaluate the effects of immersive technology–based education on students' self-efficacy. Only nonrandomized studies were included, with immersive technology using VR. The findings revealed that compared to control conditions, the interventions significantly enhanced students' self-efficacy with a large effect size (SMD=0.86, 95% CI 0.42-1.31, P<.001). The heterogeneity test showed a high level of heterogeneity across the studies (I^2=63%, P=.04), as shown in Figure 7.

Figure 7. Forest plot of immersive technology–based education on students' self-efficacy.

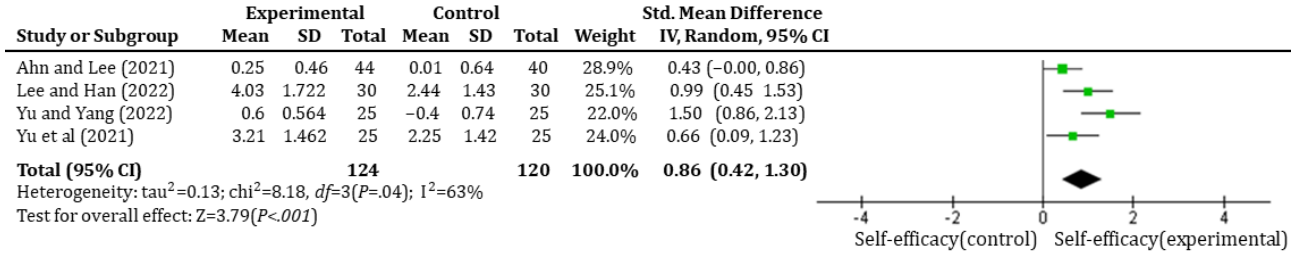

Quality of the Evidence

The results of GRADE assessment are shown in Table 2. The evidence for knowledge outcomes in RCTs was rated as high, while that for confidence outcomes in RCTs was rated as low. Conversely, the evidence for knowledge outcomes in the nonrandomized study design subgroup, which included the use of HMDs or additional controllers, was rated as very low and low, respectively. Similarly, the evidence for self-efficacy was also rated as low.

Table 2. Summary of findings using GRADE[a].

Study design	Outcomes	Participants, N; studies (N=23), n (%)	SMD[b] (95% CI)	Heterogeneity	Quality of evidence
RCT[c]	Knowledge	341; 4 (17.4)	0.59 (0.29 to 1.27)	I^2=49%, P=.12	High (imprecision, large magnitude of effect)
RCT	Confidence	241; 3 (13.0)	0.70 (0.05 to 1.35)	I^2=82%, P<.001	Low (inconsistency, imprecision, large magnitude of effect)
Nonrandomized	Knowledge (HMD[d])	338; 4 (17.4)	0.46 (–0.06 to 0.99)	I^2=82%, P<.001	Very low (risk of bias, inconsistency, imprecision)
Nonrandomized	Knowledge (HMD[d] with controller)	344; 5 (21.7)	0.67 (0.24 to 1.10)	I^2=73%, P=.006	Low (risk of bias, inconsistency, imprecision, large magnitude of effect)
Nonrandomized	Self-efficacy	244; 4 (17.4)	0.86 (0.42 to 1.31)	I^2=63%, P=.04	Low (risk of bias, inconsistency, imprecision, large magnitude of effect)

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]SMD: standard mean difference.

[c]RCT: randomized controlled trial.

[d]HMD: head-mounted device.

XSL•FO
RenderX

Discussion

Principal Findings

This systematic review identified the effectiveness of immersive technology in nursing education and assessed the quality of evidence according to the GRADE approach. Of the 23 studies selected, 19 used VR and 4 used AR, with 22 of the 23 studies published within the past 5 years. This underscores the contemporary relevance of immersive technology–based education in current teaching and learning methods. In addition, 18 studies implemented scenario-based interventions, which covered a diverse range of health care scenarios from clinical settings to home health care nursing, while 15 studies incorporated virtual patients; notably, in 7 of these studies, learners interacted with the virtual patients, allowing them to practice nursing care similar to that provided to actual patients but within a safe environment. Scenario-based learning, which encompasses diverse patient populations in various settings, has become essential in nursing education. In this context, immersive technology that implements computer-generated virtual environments has proven to be an effective approach in enhancing the effects of scenario-based education [2,9]. Furthermore, 4 studies indicated the educational effectiveness of observation in simulation settings. Observation is an advanced learning method in nursing practice, which can be advantageous if it is planned appropriately with pedagogical theories and resources in simulation-based education [62]. Observers can acquire new knowledge through objective perspectives. Immersive technology has the potential to develop a method of observing in the field of nursing education, which is presently restricted by limitations of resources.

Advanced technological equipment for implementing immersive technologies continues to be developed. Immersive technologies were facilitated by the use of equipment such as controller-embedded HMDs, haptic devices, and motion trackers in 19 studies, reflecting the latest trends in technology used in education. In addition, the 4 studies using AR technologies also used smartphones and tablets already owned by learners, enhancing accessibility and cost-effectiveness by capturing the real environment and overlaying digital images. The continuous development of various advanced devices necessitates that educators select equipment that effectively supports the achievement of learning objectives.

The outcome variables of the studies comprised 3 dimensions aligned with the NWKM, and the findings align with the current challenge in assessing the long-term effects of nursing education on professional nursing practice. To overcome this challenge, it is necessary to first plan how to assess the long-term achievement of educational programs prior to undertaking the programs. To facilitate the evaluation of the long-term outcomes at level 4 of the NWKM, it is imperative to undertake follow-up research, which can provide the ascertained efficacy of immersive technology–based nursing education [63].

According to the results of the meta-analysis in this study, compared with traditional learning methods, immersive technology–based nursing education is effective in improving undergraduate nursing students' knowledge attainment, confidence, and self-efficacy. Additionally, heterogeneity was observed among the studies measuring the effects on knowledge acquisition. A subgroup analysis based on research design and the equipment used moderately reduced the heterogeneity, and significant effects on knowledge acquisition were reported in RCTs. Nonrandomized studies using HMDs showed no significant effect on knowledge acquisition, while those that combined HMDs with controllers showed positive effects. The level of evidence through GRADE for knowledge acquisition in RCTs was high based on the study design, a low risk of bias, moderate heterogeneity, and a large effect size, while that in nonrandomized studies (HMDs with controllers) was downgraded due to the study design, a high risk of bias, high heterogeneity, and imprecision but upgraded for a large effect size, resulting in an overall rating of low. Therefore, the results should be interpreted carefully as future research might yield divergent findings. Moderate-to-high heterogeneity across studies underscores the need for the development of standardized guidelines to design immersive technology–based education and gold-standard tools to measure the efficacy of educational programs.

Limitations and Future Research

This study has several limitations. The major limitation was the large heterogeneity between studies, which requires careful interpretation of the research findings. Various types of software and equipment were used for interventions, and the results from each study varied owing to differences in technical functions. To ensure the quality of the studies, we included only those published in peer-reviewed journals. However, published research often emphasizes only significant results, posing a risk of reporting bias. Furthermore, the majority of the identified studies were published within the past 5 years, and additional studies may have been published since the completion of the review in January 2023.

Despite these limitations, this review comprehensively analyzed the characteristics of immersive technology–based education, providing valuable insights for educators and researchers aiming to implement such technologies in their teaching. We recommend the use of interactive virtual patients with scenario-based learning and the selection of devices that enhance interaction, such as HMDs or haptic devices, while considering learning objectives and practicality, including cost-effectiveness. Additionally, RCTs show that groups receiving immersive technology–based education are significantly effective in acquiring knowledge compared to traditional education groups, and also report significant effects on enhancing nursing students' confidence and self-efficacy. Although there was high heterogeneity among the studies for confidence and self-efficacy, the findings indicate the potential and feasibility of immersive technology–based education to improve learning outcomes in various aspects compared to traditional teaching methods, such as lectures and demonstrations.

Comparison With Prior Work

With the growing prevalence of immersive technology, scholars have extensively assessed its effectiveness through systematic reviews and meta-analyses, particularly focusing on VR in educational settings [64-67]. However, prior to this study,

XSL•FO
RenderX

comprehensive literature reviews considering immersive technologies beyond VR were scarce, with only 1 study addressing the learning outcomes associated with immersive technologies [2]. Therefore, it is essential to understand the educational components and effects of immersive technologies, including VR, AR, MR, and XR, given their expanding role in instructional practices.

Prior studies have often overlooked the categorization of VR based on immersion levels, despite the distinction between immersive, semi-immersive, and nonimmersive VR proposed by Cipresso et al [5]. Only a few studies have conducted literature reviews that consider the distinction of VR based on immersion levels [12,65]. Given the variability in educational effects based on the degree of immersion [68], it is crucial to classify VR according to the level of immersion to ascertain its educational impact. Therefore, this study specifically focused on VR, which uses HMDs, and investigated holistic educational effects without restrictions on scenarios or outcomes, contributing to a comprehensive understanding of the influence of immersive technology.

The results of interventions or research can be validated through the application of a theoretical framework, and it is more useful to determine distinctions under the same standard. Previous studies on nursing education, including several review studies, have assessed program effectiveness based on the NWKM [48,69-72]. Corresponding to the results of our study, the outcomes pertinent to levels 1 and 2 have been most frequently evaluated. Although the NWKM emphasizes the importance of level 4, few studies have explored level 4 outcomes [71,72]. Two studies conducted meta-analysis based on the NWKM: Delisle et al [62] compared the learning effectiveness of observers with active participants in health care simulation, finding no statistically significant differences in the subgroup analysis of NWKM level 2 outcomes. Piot et al [70] compared the learning effectiveness of various simulation types and reported that simulation effects on skills and attitudes, categorized as level 2, are more effective than comparisons, while the impact on knowledge does not reach statistical significance. However, this meta-analysis showed a significant effect on knowledge acquisition with immersive technology–based education compared to traditional teaching approach in RCTs. This finding suggests that among various types of simulations, particularly those using immersive technologies would significantly enhance learners' knowledge acquisition. Consequently, educators may find this evidence useful for integrating immersive technologies into their educational curricula. To the best of our knowledge, this study is the first to compare the effectiveness of traditional and immersive technologies.

Conclusion

The results of this meta-analysis offer insights into the current application of immersive technology in nursing education among college nursing students. Additionally, the results show that immersive technology can contribute to improving knowledge acquisition, confidence, and self-efficacy. Notably, the outcome variables predominantly aligned with levels 1- 3 of the NWKM. Nevertheless, given the moderate-to-high heterogeneity among the studies included in the meta-analysis and the varying levels of evidence according to GRADE for each outcome, it is recommended that future well-designed RCTs be conducted and subsequent research be carried out.

Acknowledgments

This work was supported by the Yonsei University College of Nursing (grant number 6-2022-0037) and the National Research Foundation of Korea grant funded by the Korean government (grant number RS-2022-00166500).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search term.
[DOCX File , 14 KB - jmir_v26i1e57566_app1.docx]

Multimedia Appendix 2
Intervention characteristics of included studies.
[DOCX File , 25 KB - jmir_v26i1e57566_app2.docx]

Multimedia Appendix 3
Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) checklist.
[PDF File (Adobe PDF File), 151 KB - jmir_v26i1e57566_app3.pdf]

References

1. Fealy S, Jones D, Hutton A, Graham K, McNeill L, Sweet L, et al. The integration of immersive virtual reality in tertiary nursing and midwifery education: a scoping review. Nurse Educ Today 2019 Aug;79:14-19. [doi: 10.1016/j.nedt.2019.05.002] [Medline: 31078869]

XSL•FO
RenderX

2. Ryan GV, Callaghan S, Rafferty A, Higgins MF, Mangina E, McAuliffe F. Learning outcomes of immersive technologies in health care student education: systematic review of the literature. J Med Internet Res 2022 Feb 01;24(2):e30082 [FREE Full text] [doi: 10.2196/30082] [Medline: 35103607]
3. Sutherland IE. Sketch pad: a man-machine graphical communication system. 1964 Presented at: DAC '64: SHARE Design Automation Workshop; May 6-7, 1964; Cambridge, MA. [doi: 10.1145/800265.810742]
4. Tang YM, Chau KY, Kwok APK, Zhu T, Ma X. A systematic review of immersive technology applications for medical practice and education - trends, application areas, recipients, teaching contents, evaluation methods, and performance. Educ Res Rev 2022 Feb;35:100429. [doi: 10.1016/j.edurev.2021.100429]
5. Cipresso P, Giglioli IAC, Raya MA, Riva G. The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. Front Psychol 2018 Nov 6;9:2086 [FREE Full text] [doi: 10.3389/fpsyg.2018.02086] [Medline: 30459681]
6. Bowers B, McCarthy D. Developing analytic thinking skills in early undergraduate education. J Nurs Educ 1993 Mar;32(3):107-114. [doi: 10.3928/0148-4834-19930301-05] [Medline: 8388926]
7. Shin S, Park J, Kim J. Effectiveness of patient simulation in nursing education: meta-analysis. Nurse Educ Today 2015 Jan;35(1):176-182. [doi: 10.1016/j.nedt.2014.09.009] [Medline: 25459172]
8. Uysal N. Improvement of nursing students' learning outcomes through scenario-based skills training. Rev Lat Am Enfermagem 2016 Aug 08;24:e2790 [FREE Full text] [doi: 10.1590/1518-8345.1310.2790] [Medline: 27508922]
9. Chen F, Leng Y, Ge J, Wang D, Li C, Chen B, et al. Effectiveness of virtual reality in nursing education: meta-analysis. J Med Internet Res 2020 Sep 15;22(9):e18290 [FREE Full text] [doi: 10.2196/18290] [Medline: 32930664]
10. Jeffries PR, Bambini D, Hensel D, Moorman M, Washburn J. Constructing maternal-child learning experiences using clinical simulations. J Obstet Gynecol Neonatal Nurs 2009 Sep;38(5):613-623. [doi: 10.1111/j.1552-6909.2009.01060.x] [Medline: 19883484]
11. Saab MM, Hegarty J, Murphy D, Landers M. Incorporating virtual reality in nurse education: a qualitative study of nursing students' perspectives. Nurse Educ Today 2021 Oct;105:105045 [FREE Full text] [doi: 10.1016/j.nedt.2021.105045] [Medline: 34245956]
12. Choi J, Thompson CE, Choi J, Waddill CB, Choi S. Effectiveness of immersive virtual reality in nursing education: systematic review. Nurse Educ 2021 Oct 12;47(3):E57-E61. [doi: 10.1097/nne.0000000000001117]
13. Cobbett S, Snelgrove-Clarke E. Virtual versus face-to-face clinical simulation in relation to student knowledge, anxiety, and self-confidence in maternal-newborn nursing: a randomized controlled trial. Nurse Educ Today 2016 Oct;45:179-184. [doi: 10.1016/j.nedt.2016.08.004] [Medline: 27537670]
14. Foronda C, Gattamorta K, Snowden K, Bauman EB. Use of virtual clinical simulation to improve communication skills of baccalaureate nursing students: a pilot study. Nurse Educ Today 2014 Jun;34(6):e53-e57. [doi: 10.1016/j.nedt.2013.10.007] [Medline: 24231637]
15. Rim D, Shin H. Effective instructional design template for virtual simulations in nursing education. Nurse Educ Today 2021 Jan;96:104624. [doi: 10.1016/j.nedt.2020.104624] [Medline: 33099091]
16. Ganji J, Shirvani MA, Motahari-Tabari N, Tayebi T. Design, implementation and evaluation of a virtual clinical training protocol for midwifery internship in a gynecology course during COVID-19 pandemic: a semi-experimental study. Nurse Educ Today 2022 Apr;111:105293 [FREE Full text] [doi: 10.1016/j.nedt.2022.105293] [Medline: 35134637]
17. JENSON CE, FORSYTH DM. Virtual Reality SimulationVirtual reality simulation: using three-dimensional technology to teach nursing students. Comput Inform Nurs 2012;30(6):312-318. [doi: 10.1097/nxn.0b013e31824af6ae]
18. Jeffries PR, Rodgers B, Adamson K. NLN Jeffries simulation theory: brief narrative description. Nurs Educ Perspect 2015 Sep;36(5):292-293. [doi: 10.5480/1536-5026-36.5.292] [Medline: 26521496]
19. Kirkpatrick J, Kirkpatrick W. Kirkpatrick's Four Levels of Training Evaluation. Alexandria: ATD Press; 2016.
20. Liao S, Hsu S. Evaluating a continuing medical education program: New World Kirkpatrick Model approach. Int J Manag Econ Soc Sci 2019;8(4):266-279. [doi: 10.32327/IJMESS/8.4.2019.17]
21. Huang X, Wang R, Chen J, Gao C, Wang B, Dong Y, et al. Kirkpatrick's evaluation of the effect of a nursing innovation team training for clinical nurses. J Nurs Manag 2022 Oct 18;30(7):2165-2175. [doi: 10.1111/jonm.13504] [Medline: 34747090]
22. Johnston S, Fox A. Kirkpatrick's evaluation of teaching and learning approaches of workplace violence education programs for undergraduate nursing students: a systematic review. J Nurs Educ 2020 Aug 01;59(8):439-447. [doi: 10.3928/01484834-20200723-04] [Medline: 32757007]
23. Lee H, Song Y. Kirkpatrick model evaluation of accelerated second-degree nursing programs: a scoping review. J Nurs Educ 2021 May;60(5):265-271. [doi: 10.3928/01484834-20210420-05] [Medline: 34039135]
24. Bryant R, Miller CL, Henderson D. Virtual clinical simulations in an online advanced health appraisal course. Clin Simulat Nurs 2015 Oct;11(10):437-444. [doi: 10.1016/j.ecns.2015.08.002]
25. Farra S, Smith S, Gillespie G. Decontamination training: with and without virtual reality simulation. Adv Emerg Nurs J 2015;37(2):125-133. [doi: 10.1097/tme.0000000000000059]

XSL•FO
RenderX

26. Liu JYW, Yin Y, Kor PPK, Cheung DSK, Zhao IY, Wang S, et al. The effects of immersive virtual reality applications on enhancing the learning outcomes of undergraduate health care students: systematic review with meta-synthesis. J Med Internet Res 2023 Mar 06;25:e39989 [FREE Full text] [doi: 10.2196/39989] [Medline: 36877550]
27. Du L, Zhao L, Xu T, Wang Y, Zu W, Huang X, et al. Blended learning vs traditional teaching: the potential of a novel teaching strategy in nursing education - a systematic review and meta-analysis. Nurse Educ Pract 2022 Aug;63:103354. [doi: 10.1016/j.nepr.2022.103354] [Medline: 35580368]
28. Alamrani MH, Alammar KA, Alqahtani SS, Salem OA. Comparing the effects of simulation-based and traditional teaching methods on the critical thinking abilities and self-confidence of nursing students. J Nurs Res 2018 Jun;26(3):152-157. [doi: 10.1097/jnr.0000000000000231] [Medline: 29016466]
29. Liu K, Zhang W, Li W, Wang T, Zheng Y. Effectiveness of virtual reality in nursing education: a systematic review and meta-analysis. BMC Med Educ 2023 Sep 28;23(1):710 [FREE Full text] [doi: 10.1186/s12909-023-04662-x] [Medline: 37770884]
30. Hong C, Wang L. Virtual reality technology in nursing professional skills training: bibliometric analysis. JMIR Serious Games 2023 Aug 21;11:e44766 [FREE Full text] [doi: 10.2196/44766] [Medline: 37603389]
31. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
32. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
33. Kim SY, Park JE, Lee YJ, Seo H, Sheen S, Hahn S, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol 2013 Apr;66(4):408-414. [doi: 10.1016/j.jclinepi.2012.09.016] [Medline: 23337781]
34. The Nordic Cochrane Centre. Review Manager (RevMan). Copenhagen: The Cochrane Collaboration; 2020.
35. Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol 1992 Jul;45(7):769-773. [doi: 10.1016/0895-4356(92)90054-q] [Medline: 1619456]
36. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
37. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. BMJ 1997 Dec 06;315(7121):1533-1537 [FREE Full text] [doi: 10.1136/bmj.315.7121.1533] [Medline: 9432252]
38. Granholm A, Alhazzani W, Møller MH. Use of the GRADE approach in systematic reviews and guidelines. Br J Anaesth 2019 Nov;123(5):554-559 [FREE Full text] [doi: 10.1016/j.bja.2019.08.015] [Medline: 31558313]
39. Shujuan L, Mawpin T, Meichan C, Weijun X, Jing W, Biru L. The use of virtual reality to improve disaster preparedness among nursing students: a randomized study. J Nurs Educ 2022 Feb;61(2):93-96. [doi: 10.3928/01484834-20211213-05] [Medline: 35112954]
40. Havola S, Haavisto E, Makinen H. The effects of computer-based simulation game and virtual reality simulation in nursing students' self-evaluated clinical reasoning skills. Comput Inform Nurs 2021;39(11):725-735. [doi: 10.1097/cin.0000000000000748]
41. Berg H, Steinsbekk A. Is individual practice in an immersive and interactive virtual reality application non-inferior to practicing with traditional equipment in learning systematic clinical observation? A randomized controlled trial. BMC Med Educ 2020 Apr 22;20(1):123 [FREE Full text] [doi: 10.1186/s12909-020-02030-7] [Medline: 32326948]
42. Berg H, Steinsbekk A. The effect of self-practicing systematic clinical observations in a multiplayer, immersive, interactive virtual reality application versus physical equipment: a randomized controlled trial. Adv Health Sci Educ Theory Pract 2021 May 28;26(2):667-682 [FREE Full text] [doi: 10.1007/s10459-020-10019-6] [Medline: 33511505]
43. Ahn MK, Lee CM. [Development and effects of head-mounted display-based home-visits virtual reality simulation program for nursing students. J Korean Acad Nurs 2021 Aug;51(4):465-477. [doi: 10.4040/jkan.21051] [Medline: 34497255]
44. Jung A, Park E. The effectiveness of learning to use HMD-based VR technologies on nursing students: chemoport insertion surgery. Int J Environ Res Public Health 2022 Apr 15;19(8):4823 [FREE Full text] [doi: 10.3390/ijerph19084823] [Medline: 35457689]
45. Lee H, Han J. Development and evaluation of a virtual reality mechanical ventilation education program for nursing students. BMC Med Educ 2022 Nov 10;22(1):775 [FREE Full text] [doi: 10.1186/s12909-022-03834-5] [Medline: 36357886]
46. Yang S, Oh Y. The effects of neonatal resuscitation gamification program using immersive virtual reality: a quasi-experimental study. Nurse Educ Today 2022 Oct;117:105464 [FREE Full text] [doi: 10.1016/j.nedt.2022.105464] [Medline: 35914345]
47. Yu M, Yang M, Ku B, Mann JS. Effects of virtual reality simulation program regarding high-risk neonatal infection control on nursing students. Asian Nurs Res (Korean Soc Nurs Sci) 2021 Aug;15(3):189-196 [FREE Full text] [doi: 10.1016/j.anr.2021.03.002] [Medline: 33894407]
48. Yu M, Yang MR. Effectiveness and utility of virtual reality infection control simulation for children with COVID-19: quasi-experimental study. JMIR Serious Games 2022 May 27;10(2):e36707 [FREE Full text] [doi: 10.2196/36707] [Medline: 35575818]

XSL•FO
RenderX

49. Rodríguez-Abad C, Rodríguez-González R, Martínez-Santos AE, Fernández-de-la-Iglesia JDC. Effectiveness of augmented reality in learning about leg ulcer care: a quasi-experimental study in nursing students. Nurse Educ Today 2022 Dec;119:105565 [FREE Full text] [doi: 10.1016/j.nedt.2022.105565] [Medline: 36155210]
50. Mayor Silva LI, de la Calle RC, Cuevas-Budhart MA. Development of communication skills through virtual reality on nursing school students: clinical trial. Comput Inform Nurs 2023;41(1):24-30. [doi: 10.1097/cin.0000000000000866]
51. Chang C, Sung H, Guo J, Chang B, Kuo F. Effects of spherical video-based virtual reality on nursing students' learning performance in childbirth education training. Interact Learn Environ 2019 Sep 27;30(3):400-416. [doi: 10.1080/10494820.2019.1661854]
52. Chao Y, Hu SH, Chiu H, Huang P, Tsai H, Chuang Y. The effects of an immersive 3d interactive video program on improving student nurses' nursing skill competence: a randomized controlled trial study. Nurse Educ Today 2021 Aug;103:104979. [doi: 10.1016/j.nedt.2021.104979] [Medline: 34049120]
53. Chen C, Chen Y, Lee M, Wang C, Sung H. Effects of three-dimensional holograms on the academic performance of nursing students in a health assessment and practice course: a pretest-intervention-posttest study. Nurse Educ Today 2021 Nov;106:105081. [doi: 10.1016/j.nedt.2021.105081] [Medline: 34418588]
54. Chen P, Liou W. The effects of an augmented reality application developed for paediatric first aid training on the knowledge and skill levels of nursing students: an experimental controlled study. Nurse Educ Today 2023 Jan;120:105629. [doi: 10.1016/j.nedt.2022.105629] [Medline: 36375385]
55. Wu M, Chao L, Xiao X. A pediatric seizure management virtual reality simulator for nursing students: a quasi-experimental design. Nurse Educ Today 2022 Dec;119:105550. [doi: 10.1016/j.nedt.2022.105550] [Medline: 36198247]
56. Kurt Y, Öztürk H. The effect of mobile augmented reality application developed for injections on the knowledge and skill levels of nursing students: an experimental controlled study. Nurse Educ Today 2021 Aug;103:104955. [doi: 10.1016/j.nedt.2021.104955] [Medline: 34051543]
57. Sen S, Usta E, Bozdemir H. The effect of mobile virtual reality on operating room nursing education. Teach Learn Nurs 2022 Apr;17(2):199-202. [doi: 10.1016/j.teln.2022.01.008]
58. Dang BK, Johnson BK, Butzlaff A, Gilbert GE, Schleicher M, Palicte J, et al. A multi-site evaluation of observer roles and telepresence technology in simulation with prelicensure nursing students. Clin Simulat Nurs 2021 Jun;55:86-94. [doi: 10.1016/j.ecns.2021.03.007]
59. Herbert VM, Perry RJ, LeBlanc CA, Haase KN, Corey RR, Giudice NA, et al. Developing a smartphone app with augmented reality to support virtual learning of nursing students on heart failure. Clin Simulat Nurs 2021 May;54:77-85. [doi: 10.1016/j.ecns.2021.02.003]
60. Smith SJ, Farra SL, Ulrich DL, Hodgson E, Nicely S, Mickle A. Effectiveness of two varying levels of virtual reality simulation. Nurs Educ Perspect 2018;39(6):E10-E15. [doi: 10.1097/01.nep.0000000000000369]
61. Smith S, Farra SL, Hodgson E. Evaluation of two simulation methods for teaching a disaster skill. BMJ Simul Technol Enhanc Learn 2021 May 18;7(2):92-96 [FREE Full text] [doi: 10.1136/bmjstel-2019-000572] [Medline: 35520385]
62. Delisle M, Ward MAR, Pradarelli JC, Panda N, Howard JD, Hannenberg AA. Comparing the learning effectiveness of healthcare simulation in the observer versus active role: systematic review and meta-analysis. Sim Healthc 2019 May 24;14(5):318-332. [doi: 10.1097/sih.0000000000000377]
63. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. J Thorac Dis 2015 Nov;7(11):E537-E540 [FREE Full text] [doi: 10.3978/j.issn.2072-1439.2015.10.63] [Medline: 26716051]
64. Rourke S. How does virtual reality simulation compare to simulated practice in the acquisition of clinical psychomotor skills for pre-registration student nurses? A systematic review. Int J Nurs Stud 2020 Feb;102:103466. [doi: 10.1016/j.ijnurstu.2019.103466] [Medline: 31783192]
65. Plotzky C, Lindwedel U, Sorber M, Loessl B, König P, Kunze C, et al. Virtual reality simulations in nurse education: a systematic mapping review. Nurse Educ Today 2021 Jun;101:104868. [doi: 10.1016/j.nedt.2021.104868] [Medline: 33798987]
66. Woon APN, Mok WQ, Chieng YJS, Zhang HM, Ramos P, Mustadi HB, et al. Effectiveness of virtual reality training in improving knowledge among nursing students: a systematic review, meta-analysis and meta-regression. Nurse Educ Today 2021 Mar;98:104655. [doi: 10.1016/j.nedt.2020.104655] [Medline: 33303246]
67. Jallad ST, Işık B. The effectiveness of virtual reality simulation as learning strategy in the acquisition of medical skills in nursing education: a systematic review. Ir J Med Sci 2022 Jun 05;191(3):1407-1426. [doi: 10.1007/s11845-021-02695-z] [Medline: 34227032]
68. Lønne TF, Karlsen HR, Langvik E, Saksvik-Lehouillier I. The effect of immersion on sense of presence and affect when experiencing an educational scenario in virtual reality: a randomized controlled study. Heliyon 2023 Jun;9(6):e17196 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e17196] [Medline: 37360072]
69. Griffiths M, Creedy D, Carter A, Donnellan-Fernandez R. Systematic review of interventions to enhance preceptors' role in undergraduate health student clinical learning. Nurse Educ Pract 2022 Jul;62:103349. [doi: 10.1016/j.nepr.2022.103349] [Medline: 35427893]

XSL•FO
RenderX

70. Piot M, Dechartres A, Attoe C, Romeo M, Jollant F, Billon G, et al. Effectiveness of simulation in psychiatry for nursing students, nurses and nurse practitioners: a systematic review and meta-analysis. J Adv Nurs 2022 Feb 11;78(2):332-347. [doi: 10.1111/jan.14986] [Medline: 34378236]
71. Lee E, De Gagne JC, Randall PS, Kim H, Tuttle B. Effectiveness of speak-up training programs for clinical nurses: a scoping review. Int J Nurs Stud 2022 Dec;136:104375. [doi: 10.1016/j.ijnurstu.2022.104375] [Medline: 36327680]
72. Vinette B, Lapierre A, Lavoie A, Leclerc-Loiselle J, Charette M, Deschênes MF. Educational strategies used in master's and doctoral nursing education: a scoping review. J Prof Nurs 2023 Sep;48:84-92. [doi: 10.1016/j.profnurs.2023.06.006] [Medline: 37775246]

Abbreviations

AI: artificial intelligence
AR: augmented reality
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HMD: head-mounted device
MR: mixed reality
NWKM: New World Kirkpatrick Model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
RCT: randomized controlled trial
RoB 2: Risk-of-Bias Tool version 2
RoBANS: Risk-of-Bias Assessment Tool for Nonrandomized Studies
SMD: standardized mean difference
VR: virtual reality
XR: extended reality

Edited by N Cahill; submitted 20.02.24; peer-reviewed by M Mahmic Kaknjo, S Nadolny; comments to author 14.03.24; revised version received 03.04.24; accepted 08.07.24; published 24.07.24.

Please cite as:
Park S, Shin HJ, Kwak H, Lee HJ
Effects of Immersive Technology–Based Education for Undergraduate Nursing Students: Systematic Review and Meta-Analysis Using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) Approach
J Med Internet Res 2024;26:e57566
URL: https://www.jmir.org/2024/1/e57566
doi:10.2196/57566
PMID:38978483

©Subin Park, Hui Ju Shin, Hyoeun Kwak, Hyun Joo Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effectiveness of Therapeutic Exercise Interventions With Virtual Reality on Balance and Walking Among Persons With Chronic Stroke: Systematic Review, Meta-Analysis, and Meta-Regression of Randomized Controlled Trials

Maria Krohn[1], MSc; Aki Rintala[2], MSc, PhD; Jaakko Immonen[3], MSc; Tuulikki Sjögren[1], MSc (Sport Sciences), MSc (Education), PhD

[1]Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

[2]Physical Activity and Functional Capacity Research Group, Faculty of Health Care and Social Services, LAB University of Applied Sciences, Lahti, Finland

[3]Department of Medicine, Wellbeing Services County of Central Finland, Jyväskylä, Finland

Corresponding Author:
Maria Krohn, MSc
Faculty of Sport and Health Sciences
University of Jyväskylä
PL 35
Keskussairaalantie 4
Jyväskylä, FI-40014
Finland
Phone: 358 405493755
Email: maria_krohn@icloud.com

Abstract

Background: Well-targeted balance, walking, and weight-shift training can improve balance capabilities in the chronic phase of stroke. There is an urgent need for a long-term approach to rehabilitation that extends beyond the acute and subacute phases, supporting participation without increasing the demand for health care staff.

Objective: This study aims to evaluate the effectiveness of therapeutic exercise interventions with virtual reality (VR) training on balance and walking at the activity and participation levels in individuals with chronic stroke, compared with control groups receiving no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without VR.

Methods: Studies were searched across 6 databases. The inclusion criteria were as follows: Adults aged 18 years or older with a stroke diagnosis for at least 6 months (population). Therapeutic exercises within a VR environment, using VR glasses or interactive games (intervention). Control groups without the use of VR (including no treatment, conventional physical therapy, specific training, similar treatment without VR, or identical treatment without the additional use of VR; comparison). We evaluated the Berg Balance Scale score, Functional Reach Test performance, Activities-specific Balance Confidence Scale score, Six-minute Walk Test, Two-minute Walk Test, 10-meter Walk Test results, and cadence (outcome measures). We investigated randomized controlled trials (study design). A meta-analysis and a meta-regression analysis were conducted to evaluate whether the content of VR interventions or control groups, as well as the level of VR immersion used, was related to balance or walking outcomes.

Results: A total of 43 randomized controlled trials involving 1136 participants were included in this review. The use of VR training in therapeutic exercise interventions had a large effect on balance (standardized mean difference 0.51, 95% CI 0.29-0.72; P<.001) and a moderate effect on walking (standardized mean difference 0.31, 95% CI 0.09-0.53; P=.006) in individuals with chronic stroke, compared with pooled control groups (no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without the use of VR). According to the meta-regression findings, the content of VR interventions (P=.52), the type of control groups (P=.79), and the level of VR immersion (P=.82) were not significantly related to the pooled balance or walking outcomes. The GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) was moderate for balance and low for walking.

Conclusions: Therapeutic exercise training with VR had a positive, albeit moderate, effect on balance and a low impact on walking at the level of activity (capacity), even in the chronic phase of stroke, without serious side effects. The results are applicable

XSL•FO
RenderX

to working-aged stroke rehabilitees who are able to walk without assistance. Further research is needed with defined VR methods and outcomes that assess performance at the level of real-life participation.

(J Med Internet Res 2024;26:e59136) doi:10.2196/59136

KEYWORDS

stroke; chronic; virtual reality; physiotherapy; therapeutic exercise; balance; walking

Introduction

A stroke can be classified as either an infarction or a hemorrhage in the brain [1]. Causes of ischemic stroke include emboli from the heart, artery-to-artery emboli, and small-vessel disease [2]. The most common causes of hemorrhagic stroke are hypertension, cerebral amyloid angiopathy, and anticoagulation [3]. Stroke incidence, case fatality, and mortality rates vary even among countries with similar demographic and socioeconomic conditions [4]. Globally, age-adjusted stroke incidence ranges from 76 to 119 per 100,000 population per year [4]. Although age-adjusted stroke mortality has decreased worldwide, the absolute number of stroke survivors and the overall burden of stroke—including incidence, prevalence, deaths, and disability-adjusted life years—remain high and have increased over the years in low- and middle-income countries [5]. Therefore, there is an increasing need to develop therapeutic exercise interventions in physiotherapy (hereafter referred to as “therapeutic exercises”) that incorporate task-specific physical training without requiring an increase in health care staffing [6].

Rehabilitation research should place greater emphasis on individuals with chronic stroke. Increasing evidence shows that people with chronic stroke have slower walking speeds and reduced walking endurance compared with the healthy older age adult population [7]. Walking activity is lowest among individuals with stroke who experience the greatest balance impairments, highlighting the need to study the often-neglected chronic stroke population separately [8]. Research highlights the need for a long-term perspective on rehabilitation that extends beyond the acute and subacute phases to support participation in social and leisure activities [9]. Well-targeted balance, walking, and weight-shift training have been shown to improve balance capacities even in the chronic phase [10]. Incorporating virtual reality (VR) into physiotherapy could be an effective solution to enhance, maintain, and promote functional capacity and performance among individuals with chronic stroke. The advantage of VR is that the user feels present and interacts in real time with the virtual environment [11-13]. The degree of immersion in VR primarily depends on how much the user is surrounded by the virtual environment [14-16]. The highest level of immersion is achieved with head-mounted displays (HMDs), while moderate immersion uses wide curved or projection screens. Low immersion relies on traditional monitors or televisions, resulting in a significantly weaker sensory connection [17-20]. In all levels of immersion, the experience can be enhanced using a combination of visual, auditory, tactile, and motion-sensory inputs [16,21].

In recent decades, technology-assisted interventions have demonstrated benefits in stroke rehabilitation [22-24]. For example, a systematic review and meta-analysis by Cheok et al [22] reported that Nintendo Wii training, when combined with conventional rehabilitation, led to improvements in functional mobility for individuals with chronic stroke. However, other systematic reviews in the field have focused on all stages of stroke (ie, acute, subacute, chronic) without specifically targeting chronic stroke alone [22,23,25-31]. For chronic stroke, previous systematic reviews have primarily explored the effectiveness of VR interventions, focusing on outcomes such as walking [28], upper limb function and activity [6,32], or VR methods, without conducting meta-analysis or meta-regression [25,27]. Only a few meta-analyses have focused on chronic stroke, reporting the benefits of VR interventions on balance and physical function [33,34]. Additionally, a meta-regression indicated that the duration and frequency of VR training were not associated with its effects on physical function [34].

According to previous systematic reviews, the effectiveness of VR training as part of therapeutic exercise treatment has not been extensively studied in individuals with chronic stroke, and prior studies have often included various stages (acute, subacute, chronic) of stroke and different age groups in their analyses [23-34]. Additionally, small sample sizes, heterogeneity, and insufficient statistical power have been noted in these reviews, which lower the level of evidence [23,24,29,33]. There is a lack of information on balance and walking outcomes at the levels of activity and participation according to the International Classification of Functioning, Disability, and Health (ICF) framework [27,31,35]. A deeper understanding of how intervention-related factors influence rehabilitation outcomes is needed [33,34]. VR training in therapeutic exercise may offer new ways to practice movements in environments that would be more time-consuming and physically demanding in real life [36].

This systematic review with meta-analysis aims to investigate the effectiveness of VR as part of therapeutic exercise interventions on balance and walking at the ICF levels of activity and participation in individuals with chronic stroke. We addressed the following research question: What is the effectiveness of therapeutic exercise interventions incorporating VR training on balance and walking at the levels of activity and participation, compared with control groups receiving no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without the use of VR? In addition, meta-regression was used to investigate the components of interventions in both experimental and control groups, as well as balance and walking outcomes, and the effects of 3 levels of immersion in VR.

XSL•FO
RenderX

Methods

Identification and Selection of Studies

The protocol for our systematic review was published in PROSPERO's International Registry of Impact Research (CRD42020184572) [37]. The search strategy was developed by members of the research team (TS, AR, Benjamin Waller, and Arja Piirainen). Specific search terms for each database were created in collaboration with a librarian (Anitta Pälvimäki) from the university's health science department. The original search strategy is available in Multimedia Appendix 1. Searches were conducted using the following 6 databases: the National Library of Medicine's Database (Medline), the Cumulative Index to Nursing and Allied Health Literature (CINAHL), the Excerpta Medica Database (Embase), the Cochrane Database of Systematic Reviews, the Physiotherapy Evidence Database (PEDro), and the Cochrane Controlled Trials Register (CENTRAL). The original systematic literature search covered studies published between January 2000 and May 2017. An updated search was conducted for studies published between June 2017 and September 2023. In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 2 [38]) guidelines [39], 2 reviewers (Benjamin Waller and Matti Munukka) independently screened all titles and abstracts of the studies using the Covidence program [40]. After the title and abstract screening phase, potentially relevant studies were independently evaluated for full-text assessment by pairs within the research team (Benjamin Waller, Matti Munukka, Susanne Aalto, Heidi Niemi, Heidi Nousiainen, or MK). The update was conducted by 2 researchers (MK and TS). In case of disagreement, a third reviewer (AR) was consulted. Additionally, a supplementary manual search was conducted on the accepted studies using their reference lists.

The PICOS (Population, Intervention, Comparison, Outcomes, Study) strategy [41] was used to define the research topic, formulate the research question, and establish the inclusion/exclusion criteria. All adults (over 18 years of age) with a stroke diagnosis for more than 6 months were included as the study population. The chronic stage of stroke is generally defined as more than 24 weeks (6 months) poststroke [42-44]. Research articles that did not clearly define cerebrovascular chronicity were excluded from the meta-analysis. Studies involving individuals with stroke identified as acute or subacute were also excluded (*population*). The *intervention* considered was therapeutic exercise (physiotherapy) sessions in a VR environment, using VR glasses or interactive games. We defined therapeutic exercise interventions as any movements systematically performed to enhance patients' activity and participation or to prevent impairments, with the aim of reducing risks, optimizing health, and reinforcing fitness and well-being [45]. A list of accepted interventions was created before study selection (MK and TS). The definition of VR adhered to the principle established by Schultheis and Rizzo [46]: "An advanced form of human-computer interface that allows the user to interact naturally with a computer-assisted environment." In addition to VR glasses, VR can be created through projection on a wall or through interactive gamification. The highest degree of immersion is experienced with various HMDs. Moderate immersion in VR is achieved using a wide curved screen [17] or by projecting screens onto walls [18]. In low immersion, the technology is typically limited to the use of a traditional monitor or television screen, resulting in a significantly weaker sensory connection compared with high-immersion experiences [19,20]. In all degrees of immersion, sensory experiences can be enhanced using a variety of visual, auditory, tactile, and motion stimuli [16-21]. For *comparison*, we chose no treatment, conventional physical therapy, specific training, similar treatment without VR, or identical treatment without the additional use of VR, pooled into a control group consisting of the aforementioned 5 control groups. The *outcomes* included balance or walking outcomes (see Multimedia Appendix 3 for the priority listing of accepted outcomes) at the ICF levels of activity (capacity) and participation (performance) (see Figure 1 for the categorization of walking, changing, and maintaining body position according to the ICF) [35,47]. Only originally published randomized controlled trials (RCTs) written in English were accepted (*study type*).

XSL•FO
RenderX

Figure 1. Categorization of walking, changing, and maintaining body position according to the ICF.

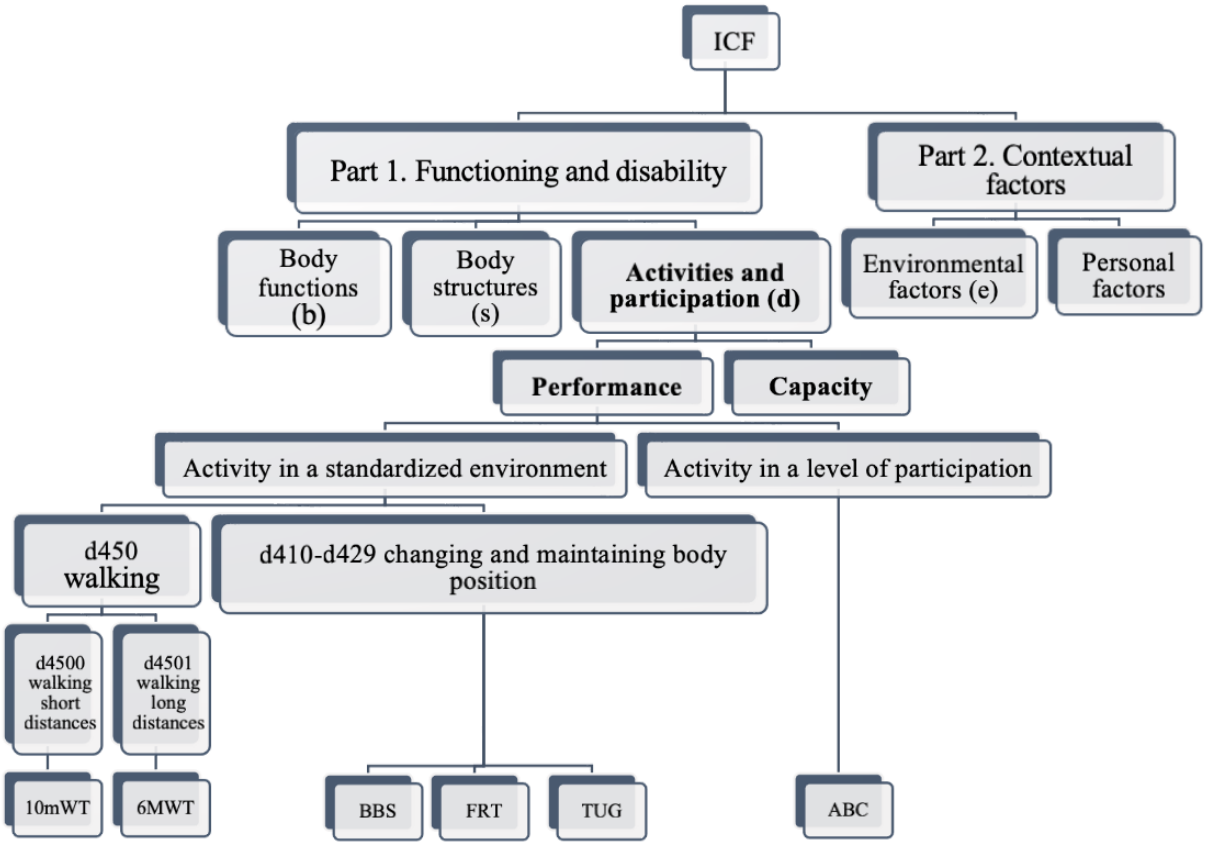

Data Extraction

First, 2 members of the research team (Matti Munukka and Benjamin Waller) independently evaluated studies published between January 2008 and May 2017 based on the inclusion criteria using the Covidence program (Covidence 2020). In case of disagreement between the 2 screeners, a third reviewer (TS) resolved the issue by evaluating the study in question. Second, the updated search (January 2017-December 2019) was screened by 2 reviewers (MK and TS), with results confirmed by another member of the research team (AR). In addition, a manual search was conducted in September 2023. The research team (MK, AR, JI, and TS) collaboratively carried out the data extraction from the included RCT studies, which included details such as the main author, country, year of publication, population characteristics (eg, age, gender, affected side, and type of cerebrovascular accident), study duration, description of the VR intervention, description of the comparison, therapy frequency, measures used, and the main results.

Risk of Bias

The methodological quality of the included articles was evaluated using the Cochrane Risk of Bias (RoB) tool [48]. The risk of bias was assessed by 1 researcher (MK) following an agreed-upon strategy among the research team. In case of disagreement, the consensus was reached by consulting a third reviewer (Susanne Aalto, Matti Munukka, Heidi Niemi, Heidi Nousiainen, Benjamin Waller, or TS). During data updates, the risk of bias assessment was conducted by 2 independent researchers (MK and TS). In case of disagreement, a third reviewer (AR) was consulted. Each article was categorized into groups of low, high, or unclear risk of bias based on the following criteria: random sequence generation (selection bias), allocation concealment (selection bias), blinding of participants (performance bias), blinding of personnel (performance bias), blinding of outcome assessment (detection bias), incomplete outcome data (attrition bias), selective reporting (reporting bias), and other sources of bias (eg, whether intervention and control groups were comparable in terms of demographic information or baseline measures). A high-quality study had to meet 5 weighted rating criteria, and the remaining points could not carry a high risk of error. The weighted rating domains included the randomization method, allocation concealment, blinding of evaluators, description of dropouts, and other potential sources of bias (eg, differences between experimental and control groups at baseline). A valid study required at least four low-risk ratings on the weighted criteria, and no valid study could have high-risk ratings on the other criteria.

Statistical Analysis

Meta-analyses were performed separately for balance and walking outcomes using a predefined list of measurements for both outcomes (see Multimedia Appendix 3 for the priority listing of accepted outcomes). Studies that did not report adequate posttreatment values (mean and SD, or mean change and SD) were excluded from the meta-analysis. All meta-analyses were conducted using a random-effects model with standardized mean differences (SMDs). The SMD between the groups was classified as large (>0.5), moderate (0.3-0.5), small (0.1-0.2), or insubstantial (<0.1) [49]. Pooled effect estimates, representing a combination of single effects from the RCTs, were analyzed using the Review Manager 5.3 statistical

XSL•FO
RenderX

software (Cochrane Collaboration) [50]. The percentage of variation across studies due to statistical heterogeneity, rather than chance, was assessed using the I^2 statistic [51]. A percentage value close to 0 indicates a low level of variation across the studies [51]. The level of variation due to statistical heterogeneity was defined at the following thresholds: <40% was considered low, 30%-60% moderate, 50%-90% substantial, and >75% considerable [51]. Potential publication bias was evaluated using funnel plots [52] from the main analysis.

Sensitivity analyses were performed for both outcomes. The inclusion criteria for the sensitivity analysis were based on the Cochrane Risk of Bias tool [48]. Studies were excluded from the sensitivity analysis if they were pilot studies or studies had clearly inconsistent reporting of results.

Meta-regression, based on meta-analytic and linear regression principles, was used to explore the heterogeneity of the meta-analysis. The use of meta-regression offers better opportunities to investigate the clinical and methodological diversity, as well as the statistical heterogeneity often observed in meta-analyses [53]. The aim was to determine whether any statistical heterogeneity could be explained by performing analysis of variance–like analyses based on 4 categorical covariates. The first covariate related to the content of the experimental intervention with VR was as follows: 1=VR training alone; 2=specific physiotherapy (eg, progressive balance training or treadmill training) with VR training as an addition to that specific physiotherapy; and 3=nonspecific conventional physiotherapy with VR training as an addition. The second covariate related to the content of the control groups without VR was as follows: 1=no training, no physiotherapy, or placebo training; 2=conventional physiotherapy (traditional); 3=specific physiotherapy training, which was not comparable to the content of the experimental group; 4=comparable physiotherapy training in terms of content and duration; and 5=identical treatment as in the experimental group, but without the additional use of VR. The third covariate related to the degree of VR immersion was categorized as follows: low=monitor or television screen; medium=wide curved screen; and high=HMDs. The fourth covariate was the response outcome variables themselves: balance and walking. Three categorical variables were considered: control, intervention, and outcome variable (balance/walking). Each of these variables was fitted separately. R software (R Foundation) [54], specifically the "metafor" package [55], was used to fit the models. The analysis of categorical variables (in this case, balance and walking) is a special case of a linear model, which indicates level differences.

Grading of Recommendations, Assessment, Development, and Evaluations (GRADE)

GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) is a framework used to rate the certainty of evidence. The health benefits and harms of therapy are crucial when developing new guidelines. The GRADE approach categorizes the quality of evidence into 4 levels: high, moderate, low, and very low: "We are very confident that the true effect lies close to that of the estimate of the effect" (*high*); "We are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different" (*moderate*); "Our confidence in the effect estimate is limited. The true effect may be substantially different from the estimate of the effect." (*low*), and "We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect." (*very low*). The following factors can reduce the quality of evidence: limitations in study design or execution, inconsistency of results, indirectness of evidence, imprecision, and publication bias. However, a large effect size, dose-response gradient, and the impact of plausible residual confounding can increase the quality of evidence [56]. Triangulation by 3 researchers (MK, TS, and AR) was used to assess the quality of the evidence.

The quality of the RCT studies was evaluated using the kappa coefficient, which indicates the consistency of evaluation between 2 independent researchers (MK and TS). The consistency of categorization was classified as follows: over 0.75 indicates excellent agreement, 0.40-0.75 indicates fair to good agreement, and below 0.40 indicates poor agreement [57].

Results

Study Selection Process and Detailed Study Information

The flowchart of study selection is presented in Figure 2. Detailed information on the interventions, comparisons, outcome variables, and inclusion criteria for the accepted studies is provided in Multimedia Appendix 4 (see also [49,58-99]).

XSL•FO
RenderX

Figure 2. Flowchart of study selection.

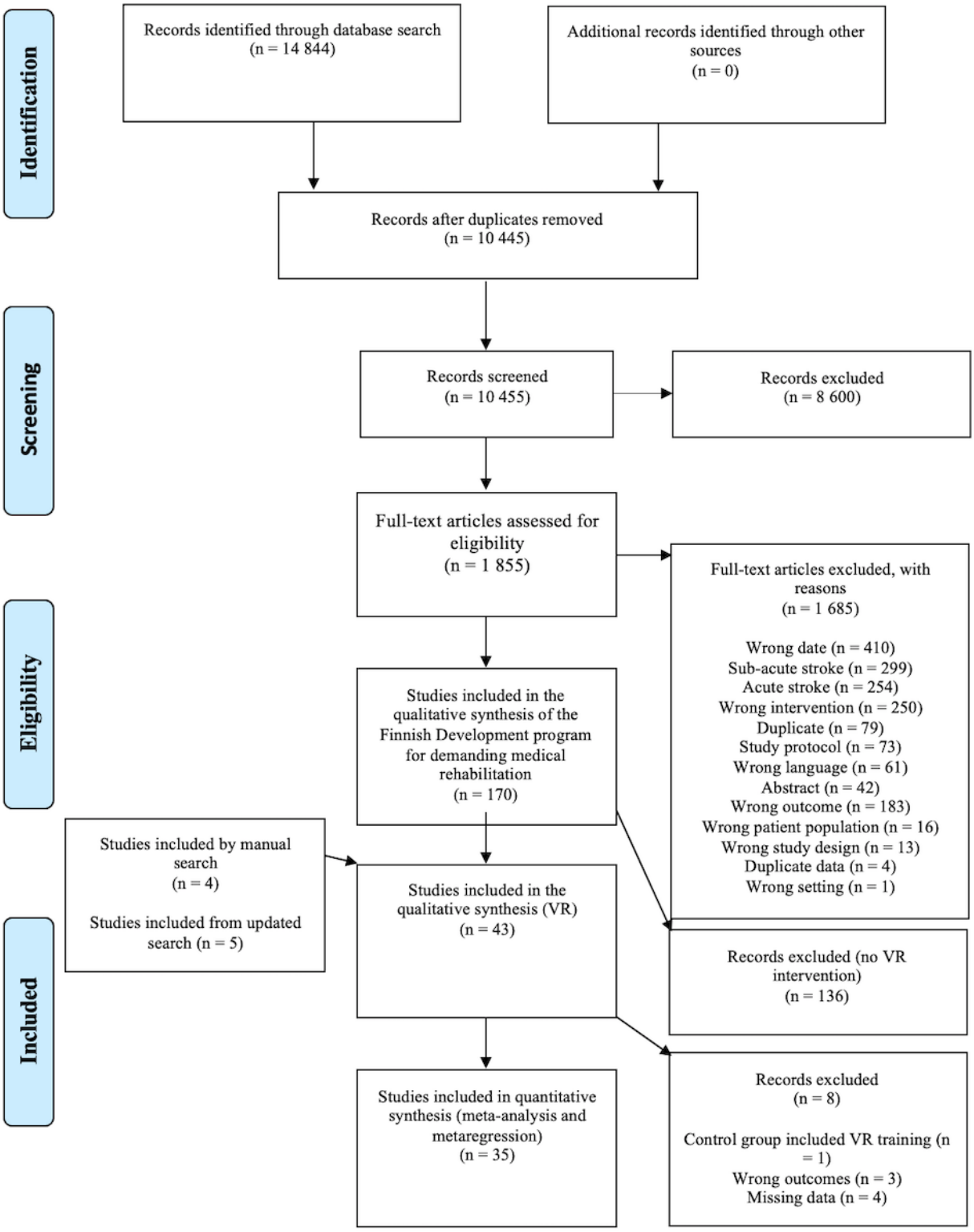

Description of the Participants

A total of 43 RCT studies (1136 participants) were included in this review, with VR utilized independently in 16 studies (246 participants) or as part of the therapeutic exercise intervention in 27 studies (343 participants). More than half of the participants were male, with an average age of 59 (SD 10.5) years. On average, stroke was diagnosed 31.1 (SD 11.9) months ago in participants. Based on the mean of 13 included studies, 214 out of 333 (64.3%) participants experienced ischemic cerebrovascular events, while 119 out of 333 (35.7%) experienced hemorrhagic events. Ten studies reported the type of cerebrovascular event, with infarction occurring in 176 of 282 (62.4%) cases and hemorrhage in 106 out of 282 (37.6%) cases. Twenty research articles did not specify the classification of the cerebrovascular accident; 454 of 862 (52.7%) participants were affected on the right side of the body, while 11 studies did not report the affected side.

XSL•FO
RenderX

The main inclusion criteria for the 43 RCT studies were adequate functioning in balance and walking ability (n=39, 91%), cognitive function and the ability to understand communication and instructions (n=36, 84%), as well as sight, hearing, and vestibular function (n=14, 33%). Additionally, participants were required to have no severe comorbidities (n=12, 28%) that could affect the therapeutic exercise. Cognitive function was generally defined as normal (Mini-Mental State Examination [MMSE] ≥24 points) [58-74] or as mild cognitive impairment (MMSE 19-23 points) [75-81]. In terms of physical functioning, there was more variation. Independent balance was mentioned in 14 studies (33%). Ten studies used a time-bound minimum requirement for standing position (minimum 30 seconds, maximum 30 minutes) [49,61,82-86], or sitting position (minimum 10 seconds, maximum 30 minutes) [49,71,87]. One study used the Berg Balance Scale definition (<56 sum score) [88]. In 3 studies, independent balance was not defined in more detail [79,92,95]. Independent walking was mentioned in 25 studies (58%). The minimum requirement was reported either in distance (10 or 15 m) [59,60,63,64,68-70,75,77,78,81,89,90] or in feet (10 feet=3.048 m) [82], or as a time-bound minimum (minimum 1 minute, maximum 30 minutes) [62,65,71,83,84,91]. Ten studies (23%) included both independent walkers and those with walking assistance [59,60,68,71,75,78,81,82,89,91]. More detailed inclusion and exclusion criteria of the original RCT studies are provided in Multimedia Appendix 4.

Description of the Therapeutic Exercise Interventions and VR Technologies

In 13 studies [49,60,65,66,71,74,83,84,90-94], VR training was integrated with therapeutic exercises. The average duration of both therapeutic exercise and VR training was 5.6 (SD 2.4) weeks, with a frequency of 3.5 sessions per week. The average duration of each therapy session was 61.3 (SD 26.6) minutes. For VR training specifically, the average duration was 5.6 weeks (SD 2.4 weeks; range 3-12 weeks), with sessions occurring 3 times per week. The average duration of a VR session was 31.8 (SD 10.0) minutes. VR technologies varied across studies, and the specific VR technologies used are detailed in Multimedia Appendix 4. Dynamic balance training was the most common VR intervention, implemented in 25 of the 43 studies (58%) [60,61,63,68,69,72-74,78,80-88,90,92,93,95-98]. The second most common intervention was walking exercises, used in 11 studies (26%) [58,59,62,64-66,75,77,89,94,99], typically performed as treadmill exercises in a VR environment. In 9 studies [60,68,74,82,92,93,95,96,98], VR was implemented using the Nintendo Wii game console with a compatible balance board, which detects weight shifts and distributions. Four studies combined balance and stepping exercises [69,72,84,86]. Five interventions used the Nintendo Wii Sports with Wii Remote controllers, excluding the balance board [61,75,79,88,97], while 5 interventions utilized the Microsoft Kinect technology for VR training [72,75,81,82,85]. Additionally, other methods were used to create VR, including head-mounted VR glasses [49,66,70], the BioRescue platform [69], IREX (Immersive Rehabilitation and Exercise) technology [84], optic flow [77], real-world recordings [58,59], and the Logitech 29 unit with a monitor [67]. VR interventions provided real-time visual feedback on participants' movements. In studies combining treadmill training with VR, the landscape change was adjusted to the participant's walking speed. Most studies enhanced the VR experience with auditory recordings or game sounds.

The content of the experimental intervention with VR was classified as follows: VR training alone (n=14, 33%) [62,63,69,72,73,75,77,82,87-89,95,96,99], VR training as an addition to specific physiotherapy (n=2, 5%) [94,98], and nonspecific conventional physiotherapy with additional VR training (n=26, 60%) [49,58-61,64-68,70,71,74,76,78,79,81,83-86,90-93,97].

The level of immersion of VR interventions used in the RCT studies was mostly low (n=30, 70%), with methods of moderate and high immersion used less frequently (n=12, 28%). The technology used included monitors or television screens in 30 studies [60,61,63,67-69,71,72,74-76,78-83,85,87-98], wide curved screens or projected screens in 9 studies [58,59,62,64,65,73,77,84,99], and various HMDs in 3 studies [49,66,70]. The degree of immersion could not be classified in 1 study [86].

Description of the Interventions in Control Groups

Control groups mostly consisted of therapeutic exercises aimed at improving balance [59-61,63,68,74,80,81,83-88,90,92,93,96-98] or walking [58,59,62,64,66,67,89,95,99] without VR technology. The content of the control intervention without VR was classified as follows: no training, no physiotherapy, or placebo training (3/43, 7%) [73,76,82]; conventional physiotherapy (8/43, 19%) [75,77,81,85,86,88,96,97]; specific physiotherapy training (8/43, 19%) [63,66-69,72,89,95], which was not comparable to the experimental group in terms of content and duration; and identical treatment to the control group (13/43, 30%) [49,65,70,71,74,79,83,84,90-94]. Conventional physiotherapy in the control groups typically involved progressive exercises using neurodevelopmental treatment, Bobath (the Bobath concept; neurodevelopmental treatment), and proprioceptive neuromuscular facilitation–based techniques. The control group generally received a combination of 2 or more treatments, such as a standard rehabilitation program and treadmill training. In 9 of 43 studies (21%) [58,59,62,64,66,67,89,95,99], the control groups received treadmill training (for a detailed description of the contents in control groups, see Multimedia Appendix 4).

Description of the Used Outcome Measurements

Accepted outcome measures for balance included the BBS, Functional Reach Test (FRT), Timed "Up and Go" (TUG), and Activities-Specific Balance Confidence Scale (ABC). Among these balance measures, BBS, FRT, and TUG primarily assess short-term balance performance rather than participation, as they involve changing and maintaining body position in a controlled, standardized environment. The ABC was the only balance measure that assessed capacity, as it evaluates activity at the level of participation through questions about perceived confidence in various activities in nonstandardized environments. Walking under standardized conditions corresponds to the performance level of the ICF framework, while walking under changing (nonstandardized) conditions is related to the ICF level of participation [100]. See Figure 1 for

XSL•FO
RenderX

the categorization of walking, changing, and maintaining body position according to the ICF. Accepted walking outcomes included the 10-meter Walk Test (10mWT), 2-minute Walk Test (2MWT), 6-minute Walk Test (6MWT), and cadence. All of these are performed in standardized environments, which means they do not capture potential changes at the level of participation. Because of the absence of outcomes that measure the level of participation, not all research questions can be addressed.

Methodological Quality and the Risk of Bias

The overall methodological quality of the studies was unclear. As many as 17 out of the 43 studies were rated as having an unclear risk of bias [58,60,62,63,67-70,73-75,78,80,82-84,97]. One study demonstrated high-quality methodology [59], while 9 studies met acceptable quality standards [49,71,76,81,85,88,92,93,99]. The main methodological weaknesses contributing to the increased risk of bias were nonblinded treatment of participants, nonblinded caregivers, potential reporting bias, and insufficiently described randomization methods, which were not described in an acceptable manner. According to the quality assessment, the reliability of the meta-analysis and meta-regression was compromised due to an unclear risk of bias related to the blinding of participants and personnel in 36 studies (see Figure 3 for the risk of bias in individual studies; also see [49,58-99]). The kappa factor, which measures the consistency of the methodological quality evaluation between 2 independent researchers, was 0.64, indicating a good level of agreement [57].

XSL•FO
RenderX

Figure 3. Risk of bias of the individual studies.

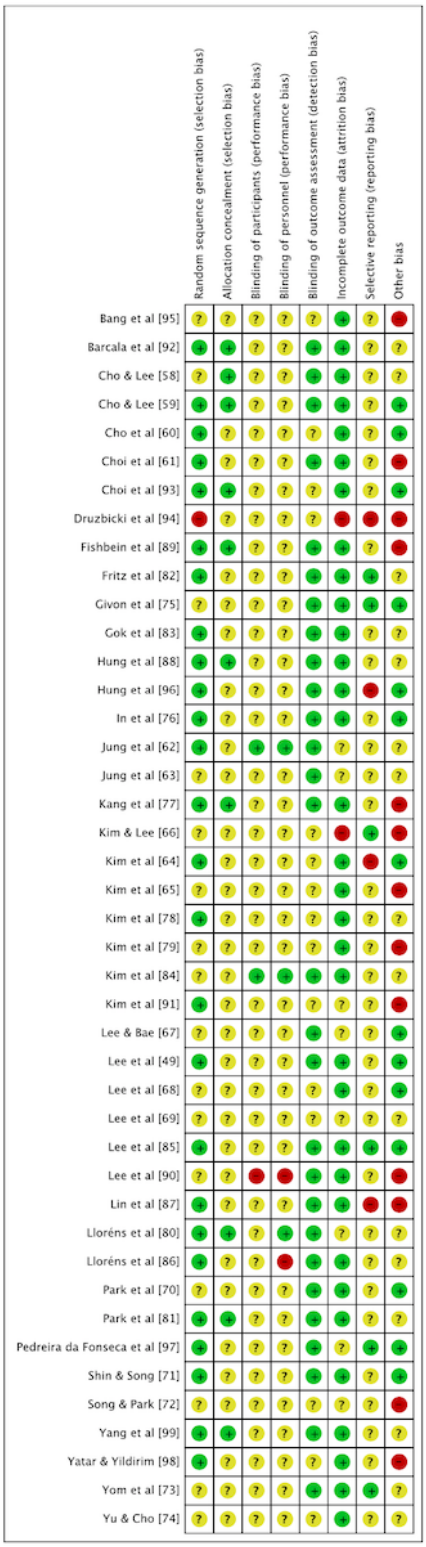

Effectiveness of VR Training on Balance: Meta-Analysis

A meta-analysis was conducted on 29 studies [49,58-60,62,63,66-69,71-74,76-78,81,82,84-87,90,92,93,96,98,99] (n=717) to assess balance outcomes. Therapeutic exercise interventions combined with VR training demonstrated a large effect on balance compared with control groups, which included no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without the use of VR (SMD 0.51, 95% CI 0.29-0.72; $P<.001$; Figure 4; also see [49,58-60,62,63,66-69,71-74,76-78,81,82,84-87,90,92,93,96,98,99]).

XSL•FO
RenderX

Figure 4. Standardized mean difference (SMD) in the effectiveness of VR interventions on balance compared to no treatment, conventional physical therapy, specific training, similar treatment, and same treatment without the use of VR.

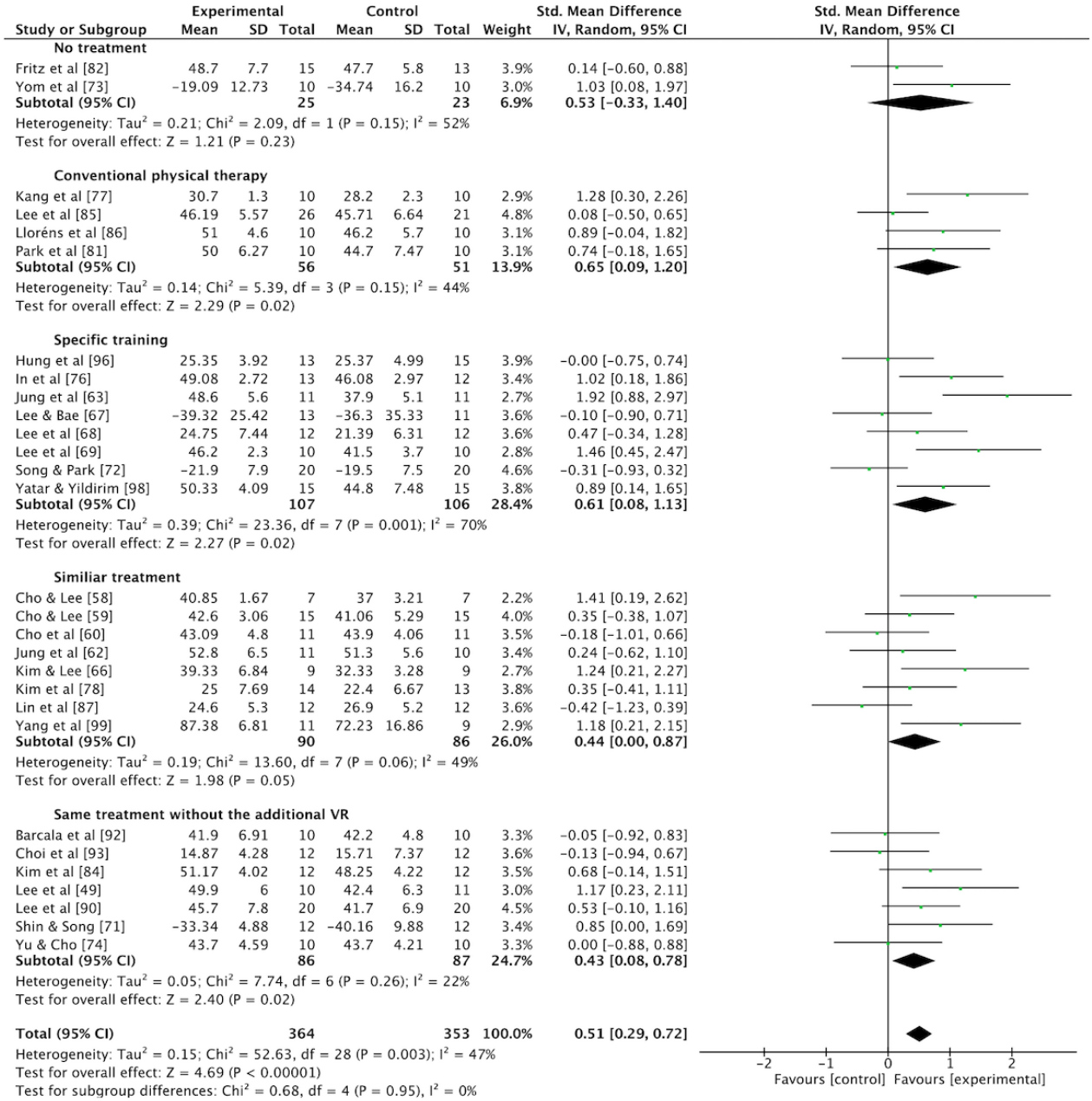

The results of the meta-analysis indicated moderate heterogeneity among the included studies (I^2=47%). A funnel plot for the SMD analysis, which may suggest possible reporting bias, is presented in Multimedia Appendix 5.

In the sensitivity analysis, 2 studies were excluded: 1 due to being a pilot study [87] and the other because of inconsistent reporting regarding the TUG test results [72]. The sensitivity analysis of 27 studies [49,58-60,62,63,66-69,71,73,74,76-78,81,82,84-86,90,92,93,96,98,99] (n=653) revealed that the statistically significant difference between the experimental and control groups remained (SMD 0.57, 95% CI 0.37 to 0.78; P<.001; Multimedia Appendix 6 (see also [49,58-60,62,63,66-69,71-74,76-78,81,82,84-87,90,92,93,96,98,99]). After the sensitivity test, the statistical heterogeneity, as measured by I^2, was reduced to 37%.

Effectiveness of VR Training on Walking: Meta-Analysis

A meta-analysis was conducted on 21 studies [49,58,59,63,64,70,72,73,75-78,81,82,84,86,90,91,94,95,99] (n=558) for walking outcomes. Therapeutic exercise interventions with VR training showed a moderate effect on walking compared with the pooled control group (which included no treatment, conventional physical therapy, specific training, similar treatment without VR, or identical treatment without the addition of VR) (SMD 0.31, 95% CI 0.09-0.53; P=.006; Figure 5; also see [49,58,59,63,64,70,72,73,75-78,81,82,84,86,90,91,94,95,99]). A moderate level of statistical heterogeneity was observed (I^2=37%). The funnel plot showing possible reporting bias is presented in Multimedia Appendix 5.

XSL•FO
RenderX

In a sensitivity analysis of walking outcomes (Multimedia Appendix 7; see also [49,58,59,63,64,70,72,73,75-78,81,82,84,86,90,91,94,95,99]), 1 study was excluded due to inconsistent reporting of the 10mWT test results [72]. The sensitivity analysis of 20 studies [49,58,59,63,64,70,73,75-78,81,82,84,86,90,91,94,95,99] (n=518) yielded similar results to the main analysis, with a slight increase in the SMD value (SMD 0.34, 95% CI 0.12-0.56; *P*=.002). The level of statistical heterogeneity remained moderate (I^2=34%).

Figure 5. Standardized mean difference (SMD) of the effectiveness of VR interventions on walking compared to no treatment, conventional physical therapy, specific training, similiar treatment and same treatment without the additional VR.

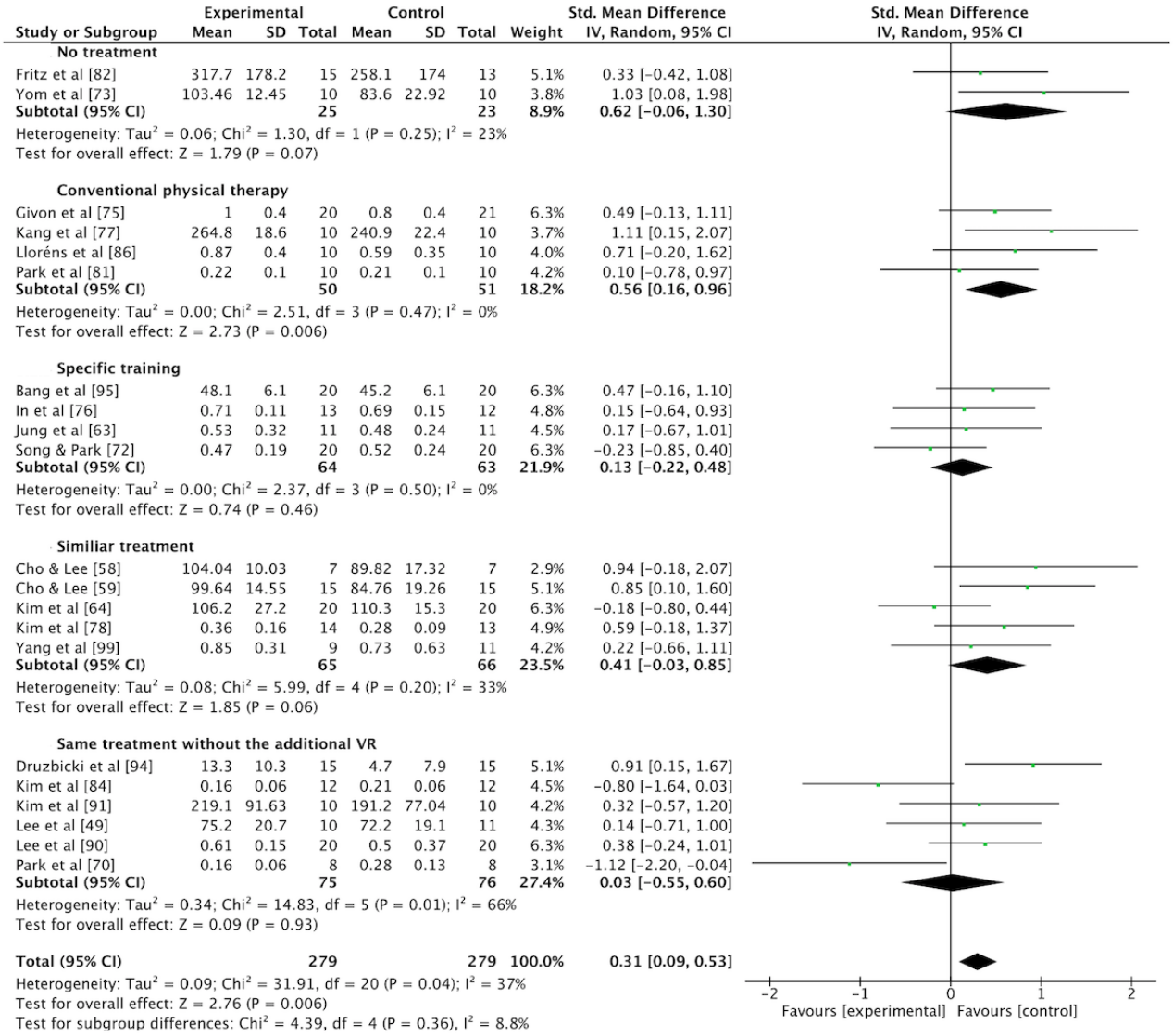

Exploring the Methodological Diversity on Walking and Balance: Meta-Regression

A meta-regression was performed on 35 studies (n=938) for the pooled group of balance and walking outcomes. There were no statistically significant differences between intervention groups (*P*=.52), control groups (*P*=.79), degree of immersion (*P*=.82), or balance and walking outcomes (*P*=.74; see Table 1 for meta-regression covariates). The wide CIs suggest a large amount of uncertainty.

XSL•FO
RenderX

Table 1. Meta-regression covariates of therapeutic exercise interventions with VR[a] on balance and walking among persons with chronic stroke.

Covariates	Estimate	SE	*Z* value	*P* value (95% CI)	Omnibus *P* value
Content of the experimental intervention					.52
VR alone (intercept)	0.449	0.163	2.751	N/A[b]	
Specific physiotherapy + VR	0.453	0.420	1.079	.28 (–0.370 to 1.276)	
Conventional physiotherapy + VR	–0.009	0.207	–0.045	.96 (–0.414 to 0.396)	
Content of the control intervention without VR training					.79
No physiotherapy or training (intercept)	0.697	0.349	1.997	N/A	
Conventional physiotherapy	–0.203	0.423	–0.480	.63 (–1.031 to 0.625)	
Noncomparable specific physiotherapy	–0.079	0.417	–0.188	.85 (–0.895 to 0.738)	
Comparable physiotherapy (content and time)	–0.409	0.403	–1.016	.31 (–1.198 to 0.380)	
Similar physiotherapy	–0.232	0.399	–0.581	.56 (–1.013 to 0.550)	
Degree of VR immersion					.82
Low; monitor or television screen (intercept)	0.425	0.119	3.563	N/A	
Medium; wide curved screen	0.155	0.245	0.636	.52 (–0.324 to 0.635)	
High; head-mounted displays	0.059	0.395	0.148	.88 (–0.715 to 0.832)	
Response outcome					N/A
Balance (intercept)	0.493	0.116	4.257	N/A	
Walking	–0.072	0.217	–0.330	.74 (–0.496 to 0.353)	

[a]VR: virtual reality.

[b]N/A: not applicable.

Strength of Research Evidence According to GRADE Evaluation

GRADE evaluation was conducted using the results from the meta-analysis and descriptive analyses (see Table 2 for factors contributing to the grade of evidence). The GRADE evaluation was downgraded for balance outcomes from high to moderate and for walking outcomes from moderate to low, primarily due to the methodological quality of the studies (risk of bias), clinical heterogeneity (inconsistency), and the low number of participants included in the meta-analysis (imprecision). Similar observations were made for other outcomes, but these were based on descriptive analysis since meta-analyses could not be performed due to the lack of specific data.

XSL•FO
RenderX

Table 2. Grade of evidence of VR[a] training in balance and walking among chronic stroke rehabilitees.[b,c,d]

Outcomes and number of studies	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Quality of the evidence (GRADE[e])	Comments
Balance outcomes (29 RCT[f] studies)	• Evaluated study quality using the Cochrane Risk of Bias tool (unclear in 14/29). • The primary methodological faults that increased the risk of bias were nonblinded treatment of the patients, nonblinded caregivers, possible reporting bias, and randomization of the studies, which was not described in an acceptable manner.	• Technology varied between Wii Fit, Wii Fit Balance Board, real-world video recordings, VR reflection therapy, optic flow, and Microsoft Kinect training. • The control group received traditional physical therapy, no intervention, treadmill walking, weight-shift exercises, proprioceptive neuromuscular facilitation training, or progressive balance training. • Moderate statistical heterogeneity (I2=47%).	• Chronic stroke survivors (>6 months)	• A meta-analysis of 29 studies with sample sizes ranging from 7 to 21 participants (n=717).	The funnel plot indicated possible reporting bias.	⊕⊕⊕ Therapeutic exercise interventions with VR training had a large effect on balance compared with control groups, which consisted of a combination of no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without the use of VR (standardized mean difference 0.51, 95% CI 0.29-0.72; *P*<.001).	• Methodological quality indicated a somewhat risk of bias. • Clinical heterogeneity was observed in the use of technology and in the treatments in control groups.
Walking outcomes (21 RCTs)	• Evaluated study quality using the Cochrane Risk of Bias tool (unclear in 9/21)	• The technology used in the experimental groups was treadmill training in a VR environment. • The control group consisted of moderate statistical heterogeneity (I2=37%).	• Chronic stroke survivors (>6 months)	• A meta-analysis of 21 studies with a sample size ranging from 7 to 21 participants (n=558).	Funnel plot indicated possible reporting bias.	⊕⊕ Therapeutic exercise interventions with VR training had a moderate effect on walking compared with the control group consisting of no treatment, conventional physical therapy, specific training, similar treatment with no VR, or identical treatment without the additional use of VR (standardized mean difference 0.31, 95% CI 0.09-0.53; *P*=.006).	• Clinical heterogeneity observed in the compared treatments of control groups.

[a]VR: virtual reality.

[b]Patient or population: persons with chronic stroke receiving VR rehabilitation.

[c]Settings: home or rehabilitation care facilities.

[d]Intervention: VR-based therapeutic exercise.

[e]GRADE: Grading of Recommendations, Assessment, Development, and Evaluations

[f]RCT: randomized controlled trial.

XSL•FO
RenderX

Discussion

Principal Findings

The results of this review provide updated insights into the benefits of VR training on balance and walking, viewed through the ICF framework, and enhance understanding of the effectiveness of therapeutic exercises incorporating VR. Our meta-analysis revealed a statistically significant large effect on balance within the ICF domains of activity and participation, and a moderate effect on walking, for therapeutic exercise interventions that included VR training compared with the pooled control group (which consisted of no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without the use of VR). The findings confirm the small [24] to moderate [32] effects on lower limb functioning and overall functioning. At the same time, our results highlight the challenges identified by previous systematic reviews on VR-assisted physiotherapy: small sample sizes, heterogeneity, insufficient statistical power, and low levels of evidence [24,33,34]. Furthermore, clinical importance [33] or the quality of evidence (eg, GRADE) [34] has not been adequately addressed in previous reviews.

Compared with previous reviews [33,34], our review provides a deeper focus on the outcomes of balance and walking within the ICF components of activities and participation [35]. Additionally, we attempted to investigate the heterogeneity associated with balance and walking outcomes by examining several covariates. According to our findings, neither the content of VR interventions nor the level of VR immersion used in the interventions was significantly related to the outcomes of balance or walking. Furthermore, the outcome variables used (balance and walking) were not associated with the heterogeneity of the results. Previous systematic reviews with meta-analysis and meta-regression have observed that the heterogeneity in VR intervention duration or weekly frequency was not significantly related to physical function outcomes [34].

This review of the used interventions (covariates 1-3) and outcome-based covariates (covariate 4) was insufficient to investigate the heterogeneity of results, as no associations were found between the covariates and pooled outcomes (balance and walking). Potential reasons for the lack of significant associations between the covariates may include the homogeneity of the research material and interventions. First, the selected RCT studies' inclusion and exclusion criteria focused on normal cognitive function and adequate physical functioning, with an emphasis on independent walking, with or without assistive devices. The only clear exception was the study by Lee et al [49], in which cognitive function was defined as below normal on the MMSE scale (<24). Second, the differences in the contents of the experimental and control groups, both within and between groups, were relatively minor. The contents of the experimental interventions in groups 1-3, as well as the control interventions in groups 2-5, are largely based on the basic principles and practices of physiotherapy science (eg, Core Competence of Physiotherapy 2016 [101]). The only exceptions were 3 control groups, where the content consisted of no treatment, including no training, no physiotherapy, or placebo training (n=3) [73,76,82]. However, the control group with no treatment was underrepresented (3/43, 7%) in this review. Third, a low level of immersion in VR was most commonly used (n=30, 70%), while HMDs, which were classified as a high level of immersion, were rarely used [49,66,70]. It is worth noting that a study-level meta-analysis does not allow for examining individual-level relationships between effect size and covariates. Study-level associations are generally weaker, making them harder to identify.

Alternative covariates that could be tested in future research on balance and walking are biopsychosocial functioning status at baseline, support for human dignity and a sense of meaningful life in therapeutic exercise with VR, VR technology as an enabling tool for motivation and commitment, dose-response relationships related to therapeutic exercise training and physiotherapist support, and broader intervention content according to the ICF framework (functioning at the level of the body, the individual, and as a member of society). The use of more detailed covariates requires careful planning, implementation, and reporting in VR RCT studies. The level of reporting in the RCTs selected for this systematic review is not yet sufficient to incorporate these covariates into a more detailed analysis of the study impact (eg, using them as covariates in meta-regression).

Meta-regression is a sophisticated tool used to explore heterogeneity by examining whether a linear relationship exists between selected outcome measures and 1 or more covariates. However, associations found in a meta-regression should be considered hypothesis-generating rather than evidence of causality. Many clinicians, such as physiotherapists, as well as policy decision makers in physiotherapy and rehabilitation, may still be unfamiliar with the principles and assumptions underlying meta-regression [53].

In the following paragraphs, we discuss the results related to balance and walking in greater detail, in relation to the findings from previous systematic literature reviews.

Effectiveness of VR Interventions on Balance in Persons With Chronic Stroke

Our study focused on balance outcomes because poor dynamic balance is strongly associated with an increased risk of falls in patients with stroke, which can hinder participation in social activities [102]. Dynamic balance tests, such as TUG and BBS, are often better predictors of falls compared with static balance tests [102]. Our meta-analysis demonstrated a statistically significant effect on balance in therapeutic exercise interventions with VR training compared with the pooled control group, which included no treatment, conventional physical therapy, specific training, similar treatment, or identical treatment without the use of VR.

Our study provides updated and more detailed knowledge on the effects of therapeutic exercises on balance in individuals with chronic stroke, as our systematic review includes 7 newly published RCTs [61,71,74-76,93,95] that were not part of the latest systematic literature review by Iruthayarajah et al [33]. Our findings align with those of Iruthayarajah et al [33], where the authors reported significantly higher BBS scores in the VR

XSL•FO
RenderX

balance training group (including Nintendo Wii Fit Balance Board training, VR combined with treadmill training, or postural VR training) compared with the control group, which received alternative rehabilitation therapy [33].

In addition, using the ICF as a framework revealed a lack of outcomes related to participation (performance) in previous RCTs on physiotherapy (see Multimedia Appendix 4 for outcome variables of the included studies). As a result, the research question concerning balance and walking outcomes can only be answered at the ICF performance level (activity in a standardized environment). This highlights the need for more research on the effects of VR training at the ICF participation level.

Effectiveness of VR Interventions on Walking in Persons With Chronic Stroke

Our findings indicated a moderate effect on walking in therapeutic exercise interventions with VR training compared with control groups without the use of VR for individuals with chronic stroke. Our research design clearly differs from previous systematic literature reviews, offering new insights into the effectiveness of VR on walking in individuals with chronic stroke. In previous reviews examining the effects of VR on walking, no meta-analysis was performed combining walking outcomes or prioritizing long-distance walking [6,24,26,28]. Our study provides new insights into the effects of therapeutic exercises on walking in individuals with chronic stroke, as previous reviews focused on stroke populations with varying disease durations [6,24,26,28]. In the review by Rodrigues-Baroni et al [28], a significant difference was observed in favor of VR practitioners regarding walking, as assessed by comfortable gait speed, compared with a placebo or no intervention. Additionally, in contrast to other systematic literature reviews, our study evaluated walking using both short-distance walking tests (2MWT) and long-distance walking tests (6MWT). Our findings on walking outcomes can only be interpreted from the ICF performance level, as all tests were conducted in standardized environments, leaving capacity (activity at the level of participation) unexamined. In our study, therapeutic exercise interventions were not limited to walking exercises alone; therefore, the effects were less pronounced than those observed in the study by Rodrigues-Baroni et al [28], although the results remain consistent.

Risk of Bias

In this review, the overall risk of bias in studies included in the meta-analysis was assessed using the RoB with weighted rating criteria. Of the 43 studies, the risk of bias was found to be low in 10 (23%), high in 16 (37%), and raised some concerns in 17 (40%) RCTs included. These findings are somewhat aligned with those of Hohenschurz-Schmidt et al [103], who reported a risk of bias of 17% low, 44% high, and 38% with some concerns, as assessed using the updated Cochrane RoB tool (RoB 2). The comparison article [103] focused on physical, psychological, and self-management interventions for pain, providing general insights into risk assessment levels in physiotherapy, rehabilitation, and self-care studies. It also highlights potential differences between the Cochrane Risk of Bias tools (RoB and RoB 2), suggesting the need for further investigation. Since 2024, it has been recommended that the Cochrane Risk of Bias tool RoB 2 be used for all randomized trials in Cochrane Reviews [104].

In our research, the weighted quality criteria included blinding of rehabilitees and therapists, which lowered the overall quality assessment. Blinding the assessor and reporting it within the study would be an effective way to reduce the risk of unclear bias. A common issue in the original RCTs included in our systematic review—and in therapeutic exercise studies in physiotherapy more broadly [103]—was the lack of blinding among participants and those delivering the interventions to the intervention groups. This challenge arises because participants play an active role during interventions. Future research should explore solutions to the difficulties of implementing blinding in physiotherapy studies or consider addressing the issue from the perspective of the control group. Hohenschurz-Schmidt et al [103] emphasized how control design features—such as the number of treatment sessions, application methods, individualization of interventions, patient participation, fidelity monitoring, and treatment environment—can all influence outcomes. Similarly, in our review, we evaluated the diversity of VR interventions by classifying the content of the experimental and control groups and assessing how these factors were related to balance or walking outcomes.

Commonly used practices, such as standardizing registration and preregistration of RCTs (eg, ClinicalTrials.gov or ISRCTN Registry) and systematic reviews (eg, PROSPERO), could enhance research reliability. Accurately reporting dropout rates and consistently presenting all outcome variable results could further improve the quality of research articles [105,106]. If the quality of the RCTs included in a systematic review is not clearly defined, it can undermine the credibility of the results, the generalizability of findings, and the clinical implications [107]. In this review, the presence of an unclear risk of bias in many studies reduced the level of research evidence according to the GRADE evaluation. To achieve greater internal reliability, VR RCT studies should be reported more clearly and systematically, with a more accurate evaluation of bias and the use of internationally standardized tools and methods. Additionally, new analytical approaches should be developed to account for the multidimensional factors related to rehabilitee functioning, as well as individual and environmental factors, which may contribute to result heterogeneity in meta-analyses [107].

Study Strengths and Limitations

The phenomenon of VR training has not yet been sufficiently investigated, but there are reasons to believe that VR-assisted therapeutic exercise contributes to improving physical functioning (balance and walking) in individuals with chronic stroke. The strengths of our study include the careful planning of the systematic review (including review registration), accurate thematic classification of the content of the experimental and control groups, levels of VR immersion, the outcomes used, and sophisticated meta-regression analyses to explore heterogeneity and the consistency of results. Furthermore, we gained updated insights into balance and walking among chronic

XSL•FO
RenderX

stroke rehabilitees and conducted a thorough GRADE evaluation.

One limitation, related to study selection bias, is that our search strategy was not extended to national and regional databases, reference indices, the Scopus database, Google Scholar, Microsoft Academic search engine, doctoral dissertations, and thesis databases, as well as gray literature, such as reports, conference summaries, and unpublished or nonbibliographic sources [108]. Other clear limitations of the study include shortcomings in the quality of the original RCTs, for which no clarification was sought from the authors, and the inability to control for heterogeneity in balance and walking outcomes. The intervention-related covariates we used did not provide additional insights into the associations related to the heterogeneity of balance and walking results. Additionally, an older version of the RoB was used to evaluate the original studies.

The unclear risk of bias in many studies highlights its potential impact on validity. This potential bias was exacerbated by the nonblinding of participants and caregivers to treatment, reporting biases, and inadequate or unclear randomization procedures. Considering the strengths and weaknesses of this systematic review's results, the GRADE of evidence [56,109], and the unclear risk of bias in methodological validity, the evidence is downgraded to moderate for balance and low for walking.

Future Recommendations

Qualitative studies on the experience of VR technology among individuals with chronic stroke should be recommended, as VR technology may alter the nature of communication between the therapist and the participant. For instance, Sjögren and Korpi [110] used thematic analysis and inductive synthesis of qualitative studies among stroke and multiple sclerosis rehabilitees to identify key elements of (digital) technology–assisted physiotherapy. These themes included support for motivation and commitment, enablement of social interactions and relationships, design of safe and variable training environments, flexibility in choosing relevant and meaningful activities, identification of rehabilitation needs and goals, and support for rehabilitees to understand their current functioning status and the appropriate paths to improvement [110].

We observed a lack of studies using VR technology to measure the level of participation according to the ICF. We argue that there is still insufficient evidence regarding the effectiveness of VR technology among chronic stroke populations, and changes in activity and participation are not clearly defined at the ICF levels. For example, the outcomes in this meta-analysis were primarily short performance tests conducted in a clinical setting, which may not necessarily reflect changes in activities of daily living functions in participants' everyday environments.

The use of VR methods in a home environment as part of self-rehabilitation should be explored more specifically, as many VR methods provide real-time feedback on body movements, aiding in relearning. A comparative study of VR methods with standardized training volumes and durations should be conducted to determine which VR methods are most effective. To enable independent VR training in the home environment, VR should also be studied without the guidance of a physiotherapist or other health professionals. In such research designs, safety factors and careful selection of the appropriate target group should be prioritized.

In this review, we found that physiotherapy still primarily uses commercial devices that provide less immersion, rather than VR technology, which offers greater immersion (30/43, 70%). A similar trend has been observed in previous rehabilitation studies (eg, Saposnik et al [111] and Laver et al [6]). One reason for this may be the high cost, limited availability, or nausea associated with using more advanced VR headsets [112,113], as well as the fact that few devices have been specifically developed for physiotherapy and physical rehabilitation. To date, our research evidence does not show a statistically significant advantage for deeper immersion technologies compared with lower levels of immersion. However, commercial VR methods may still be useful as part of rehabilitation, increasing the overall amount of training, and they are more accessible to caregivers.

Lohse et al [31] criticized the small number of participants in RCT studies, as well as the insufficient determination of conventional therapy, adherence effects, and motivational components in the original studies. In addition, our research highlights the need for future studies to take a closer look at VR technology interventions from the perspective of physiotherapy, specifically examining what kind of functional capacity training prerequisites VR enables and promotes, and how this activity is connected to the rehabilitee's meaningful everyday life. For example, it is important to deepen our understanding of how exercises with VR glasses or lower extremity platforms differ, taking into account biopsychosocial functioning, cognitive requirements, and the degree of immersion. Additionally, it is crucial to determine whether the observed changes have significance for the rehabilitee's functioning, such as changes in body function, body structures, capacity, and performance.

To enhance knowledge and understanding, there is an urgent need for RCT studies that also consider functional performance across varying operating environments. The RCT studies included in our review primarily focused on balance or walking outcomes at the performance level. In the future, greater attention should be given to the quality of RCT research. It would also be important to more accurately investigate the potential connections between motivation and outcome variables. Rohrbach et al [114] also emphasized the need for standardized terminology and outcome measures to better understand how factors such as enjoyment, engagement, motivation, immersion, and presence contribute individually or interactively to VR intervention effectiveness. For example, rehabilitee motivation could not be used as a covariate, as none of the original studies included it as an outcome.

Generalizability and Clinical Implications

Chronic stroke rehabilitation is a comprehensive field that includes participants with varying limitations and levels of function. Therefore, individuals with chronic stroke should not be viewed as a homogeneous group, but rather as persons with

XSL•FO
RenderX

specific, individualized needs. In this review, the results cannot be generalized to the entire population, as the volunteers selected for the RCT studies do not fully represent the target group. The mean age of the participants included in our review (59 years) was significantly lower than the average age for men (69 years) and women (73 years) with their first cerebrovascular disease in Western Europe [115,116]. Additionally, based on the inclusion and exclusion criteria of our study, the participants in the selected RCTs had physical and cognitive functioning at a level that allowed them to engage in demanding balance and walking training with VR without assistance. Furthermore, the participants did not have severe comorbidities. These factors limit the applicability of our study results to older populations or rehabilitees with more significant functional impairments or multimorbidities. Future VR studies should include categorized subgroup analyses to provide more comprehensive clinical recommendations.

Despite the limitations previously mentioned, our findings may be useful for health care professionals, especially physiotherapists, working with individuals with chronic stroke. They could consider integrating VR technology into their clinical practice as an adjunct to conventional therapy to increase the amount of therapeutic activity.

In terms of clinical implications, it is important to consider not only the groups of rehabilitees in one's practice but also potential adverse effects and factors related to defining clinically significant results. While our results indicated a large effect on balance and a moderate effect on walking, the clinical significance of these findings remains partly unclear. VR training did not improve BBS scores to a clinically significant level compared with rehabilitees who did not participate in VR training [117]. However, according to the review by Downs et al [118], a mean difference of more than 3 points may be considered clinically significant if the baseline scores are between 20 and 50 points.

This systematic review of the included VR studies reported some minor adverse effects. Park et al [70] noted dizziness caused by VR glasses and a time delay between real-time and reference motion, which resulted in visual feedback not occurring on schedule. Hung et al [96] reported a possible increase in spasticity, leading to a higher risk of falling as an adverse effect in their study. The authors highlighted that the safety of VR training could be improved by ensuring the physiotherapist and, if necessary, relatives are familiar with VR technology, as well as by providing a barrier-free and quiet training environment. They also argued that, for the most part, VR training related to balance and walking is not suitable for wheelchair users [96].

More research is needed with outcome measures at the participation level, such as long-distance walking or the ABC test. Original RCT studies utilizing deeper levels of immersion (eg, head-mouth devices) in the VR world are also required. For better implementation of the research findings, accurate definitions of chronic stroke rehabilitators' entry levels are necessary, along with solutions for studying rehabilitees who are not able to walk independently without assistance. Considering the inclusion and exclusion criteria of the original studies, rehabilitees with limitations in vision, hearing, cognition, sense of touch, or physical function are often excluded. Current research settings tend to focus on chronic stroke rehabilitators with fewer disadvantages. This raises the question of whether VR can be effective if the baseline level is low.

Conclusions

The use of VR training in therapeutic exercise interventions may improve balance and walking in persons with chronic stroke. However, this review suggests that no significant association was found between the content of VR interventions, the control group, or the level of immersion used on balance and walking outcomes. The evidence was of moderate certainty for balance and low certainty for walking.

For clinical implementation, statistically significant results were achieved on average with the following factors: VR training using commercially available technology in addition to conventional training, 3 times a week, with each session lasting 30 minutes. This approach was applied to individuals with chronic stroke who had sitting and standing balance, no assistive devices, and the ability to walk independently. High-quality RCTs are needed to evaluate VR interventions for chronic stroke, specifically focusing on balance and walking at the ICF participation level. Future studies should also include participants of varying ages, functional abilities, and comorbidities to help generalize the results to a broader population of individuals with chronic stroke.

Acknowledgments

The research was managed by the Department of Physiotherapy at the Faculty of Sports Science, University of Jyväskylä, Jyväskylä, Finland. We thank all the researchers involved in the broader VAKU project: Benjamin Waller, PhD, Matti Munukka, PhD, Susanne Aalto, MSc, Heidi Niemi, MSc, Heidi Nousiainen, MSc, Riikka Köngäs, MSc, Juho Korpi, MSc, Hilkka Korpi, PhD, Arja Piirainen, PhD, and Ari Heinonen, PhD from the University of Jyväskylä, Jyväskylä, Finland; and Jaana Paltamaa, PhD, from Jyväskylä University of Applied Sciences. This study received funding from the Finnish Social Insurance Institution.

Data Availability

The data generated during or analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
The original search strategy.
[DOCX File , 19 KB - jmir_v26i1e59136_app1.docx]

Multimedia Appendix 2
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 128 KB - jmir_v26i1e59136_app2.pdf]

Multimedia Appendix 3
Prioritization of balance and walking outcomes in the meta-analysis and meta regression.
[DOCX File , 23 KB - jmir_v26i1e59136_app3.docx]

Multimedia Appendix 4
Description of interventions, comparisons, outcome variables and inclusion criteria of included studies.
[DOCX File , 48 KB - jmir_v26i1e59136_app4.docx]

Multimedia Appendix 5
Funnel plot of standardized mean difference (SMD) analysis with balance and walking outcomes.
[DOCX File , 853 KB - jmir_v26i1e59136_app5.docx]

Multimedia Appendix 6
Sensitivity test for standardized mean difference (SMD) of the effectiveness of therapeutic exercise interventions utilizing VR technology on balance.
[DOCX File , 545 KB - jmir_v26i1e59136_app6.docx]

Multimedia Appendix 7
Sensitivity test for standardized mean difference (SMD) of the effectiveness of therapeutic exercise interventions on walking utilizing VR technology.
[DOCX File , 496 KB - jmir_v26i1e59136_app7.docx]

References

1. Hankey GJ. Stroke. The Lancet 2017 Feb;389(10069):641-654. [doi: 10.1016/s0140-6736(16)30962-x]
2. Amarenco P, Bogousslavsky J, Caplan L, Donnan G, Wolf M, Hennerici M. The ASCOD phenotyping of ischemic stroke (updated ASCO phenotyping). Cerebrovasc Dis 2013 Jul 30;36(1):1-5 [FREE Full text] [doi: 10.1159/000352050] [Medline: 23899749]
3. Timsit S. Stroke at the beginning of the XXIst century. Presse Med 2016 Dec;45(12 Pt 2):e389-e390. [doi: 10.1016/j.lpm.2016.11.007] [Medline: 27932077]
4. Thrift AG, Thayabaranathan T, Howard G, Howard VJ, Rothwell PM, Feigin VL, et al. Global stroke statistics. Int J Stroke 2017 Jan 28;12(1):13-32. [doi: 10.1177/1747493016676285] [Medline: 27794138]
5. Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, Global Burden of Diseases, Injuries,Risk Factors Study 2010 (GBD 2010)the GBD Stroke Experts Group. Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet 2014 Jan 18;383(9913):245-254 [FREE Full text] [doi: 10.1016/s0140-6736(13)61953-4] [Medline: 24449944]
6. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2017 Nov 20;11(11):CD008349 [FREE Full text] [doi: 10.1002/14651858.CD008349.pub4] [Medline: 29156493]
7. Ng SS, Hui-Chan CW. Arch Phys Med Rehabil 2005 Aug;86(8):1641-1647. [doi: 10.1016/j.apmr.2005.01.011] [Medline: 16084820]
8. Michael KM, Allen JK, Macko RF. Reduced ambulatory activity after stroke: the role of balance, gait, and cardiovascular fitness. Arch Phys Med Rehabil 2005 Aug;86(8):1552-1556. [doi: 10.1016/j.apmr.2004.12.026] [Medline: 16084807]
9. Norlander A, Iwarsson S, Jönsson AC, Lindgren A, Månsson Lexell E. Participation in social and leisure activities while re-constructing the self: understanding strategies used by stroke survivors from a long-term perspective. Disabil Rehabil 2022 Aug 28;44(16):4284-4292. [doi: 10.1080/09638288.2021.1900418] [Medline: 33779458]

XSL•FO
RenderX

10. van Duijnhoven HJR, Heeren A, Peters M, Veerbeek J, Kwakkel G, Geurts, et al. Effects of exercise therapy on balance capacity in chronic stroke: systematic review and meta-analysis. Stroke 2016 Oct;47(10):2603-2610. [doi: 10.1161/STROKEAHA.116.013839] [Medline: 27633021]
11. Wilson PN, Foreman N, Stanton D. Virtual reality, disability and rehabilitation. Disability and Rehabilitation 2009 Jul 28;19(6):213-220. [doi: 10.3109/09638289709166530] [Medline: 9195138]
12. Imam B, Jarus T. Virtual reality rehabilitation from social cognitive and motor learning theoretical perspectives in stroke population. Rehabil Res Pract 2014;2014:594540-594511. [doi: 10.1155/2014/594540] [Medline: 24523967]
13. Kaplan AD, Cruit J, Endsley M, Beers SM, Sawyer BD, Hancock PA. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: a meta-analysis. Hum Factors 2021 Jun;63(4):706-726. [doi: 10.1177/0018720820904229] [Medline: 32091937]
14. Mujber T, Szecsi T, Hashmi M. Virtual reality applications in manufacturing process simulation. Journal of Materials Processing Technology 2004 Nov;155-156:1834-1838. [doi: 10.1016/j.jmatprotec.2004.04.401]
15. Henderson A, Korner-Bitensky N, Levin M. Virtual reality in stroke rehabilitation: a systematic review of its effectiveness for upper limb motor recovery. Top Stroke Rehabil 2007 Dec 18;14(2):52-61. [doi: 10.1310/tsr1402-52] [Medline: 17517575]
16. Rizzo A, Koenig ST. Is clinical virtual reality ready for primetime? Neuropsychology 2017 Nov;31(8):877-899. [doi: 10.1037/neu0000405] [Medline: 29376669]
17. Kalron A, Fonkatz I, Frid L, Baransi H, Achiron A. The effect of balance training on postural control in people with multiple sclerosis using the CAREN virtual reality system: a pilot randomized controlled trial. J Neuroeng Rehabil 2016 Mar 01;13(1):13 [FREE Full text] [doi: 10.1186/s12984-016-0124-y] [Medline: 26925955]
18. Cruz-Neira C, Sandin D, DeFanti T. Surround-screen projection-based virtual reality. In: Proceedings of the 20th Annual Conference and Exhibition on Computer Graphics and Interactive Techniques. 1993 Aug 6 Presented at: 20th Annual Conference and Exhibition on Computer Graphics and Interactive Techniques; August 2-6, 1993; Anaheim, CA p. 135-142. [doi: 10.1145/166117.166134]
19. Ma M, Zheng H. Virtual reality and serious games in healthcare. In: Brahnam S, Jain LC, editors. Advanced Computational Intelligence Paradigms in Healthcare 6. Virtual Reality in Psychotherapy, Rehabilitation, and Assessment. Berlin, Heidelberg, Germany: Springer; 2011:169-192.
20. Mirelman A, Rochester L, Maidan I, Del Din S, Alcock L, Nieuwhof F, et al. Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial. The Lancet 2016 Sep;388(10050):1170-1182. [doi: 10.1016/s0140-6736(16)31325-3]
21. Galvin J, Levac D. Facilitating clinical decision-making about the use of virtual reality within paediatric motor rehabilitation: describing and classifying virtual reality systems. Dev Neurorehabil 2011 Mar 17;14(2):112-122. [doi: 10.3109/17518423.2010.535805] [Medline: 21410403]
22. Cheok G, Tan D, Low A, Hewitt J. Is Nintendo Wii an effective intervention for individuals with stroke? A systematic review and meta-analysis. J Am Med Dir Assoc 2015 Nov 01;16(11):923-932. [doi: 10.1016/j.jamda.2015.06.010] [Medline: 26253322]
23. Rintala A, Hakala S, Sjögren T. Etäteknologian Vaikuttavuus Liikunnallisessa Kuntoutuksessa: Järjestelmällinen Kirjallisuuskatsaus Ja Meta-Analyysi. Helsinki, Finland: Kela; 2017.
24. Rintala A, Päivärinne V, Hakala S, Paltamaa J, Heinonen A, Karvanen J, et al. Effectiveness of technology-based distance physical rehabilitation interventions for improving physical functioning in stroke: a systematic review and meta-analysis of randomized controlled trials. Arch Phys Med Rehabil 2019 Jul;100(7):1339-1358. [doi: 10.1016/j.apmr.2018.11.007] [Medline: 30529323]
25. Chen L, Lo WLA, Mao YR, Ding MH, Lin Q, Li H, et al. Effect of virtual reality on postural and balance control in patients with stroke: a systematic literature review. Biomed Res Int 2016;2016:7309272-7309278 [FREE Full text] [doi: 10.1155/2016/7309272] [Medline: 28053988]
26. Gibbons EM, Thomson AN, de Noronha M, Joseph S. Are virtual reality technologies effective in improving lower limb outcomes for patients following stroke - a systematic review with meta-analysis. Top Stroke Rehabil 2016 Dec 30;23(6):440-457. [doi: 10.1080/10749357.2016.1183349] [Medline: 27237336]
27. Palma GCDS, Freitas TB, Bonuzzi GMG, Soares MAA, Leite PHW, Mazzini NA, et al. Effects of virtual reality for stroke individuals based on the International Classification of Functioning and Health: a systematic review. Top Stroke Rehabil 2017 May;24(4):269-278. [doi: 10.1080/10749357.2016.1250373] [Medline: 27796177]
28. Rodrigues-Baroni JM, Nascimento LR, Ada L, Teixeira-Salmela LF. Walking training associated with virtual reality-based training increases walking speed of individuals with chronic stroke: systematic review with meta-analysis. Braz J Phys Ther 2014;18(6):502-512 [FREE Full text] [doi: 10.1590/bjpt-rbf.2014.0062] [Medline: 25590442]
29. Corbetta D, Imeri F, Gatti R. Rehabilitation that incorporates virtual reality is more effective than standard rehabilitation for improving walking speed, balance and mobility after stroke: a systematic review. J Physiother 2015 Jul;61(3):117-124 [FREE Full text] [doi: 10.1016/j.jphys.2015.05.017] [Medline: 26093805]
30. Li Z, Han X, Sheng J, Ma S. Virtual reality for improving balance in patients after stroke: a systematic review and meta-analysis. Clin Rehabil 2016 May;30(5):432-440. [doi: 10.1177/0269215515593611] [Medline: 26141808]

XSL•FO
RenderX

31. Lohse KR, Hilderman CGE, Cheung KL, Tatla S, Van der Loos HFM. Virtual reality therapy for adults post-stroke: a systematic review and meta-analysis exploring virtual environments and commercial games in therapy. PLoS One 2014;9(3):e93318 [FREE Full text] [doi: 10.1371/journal.pone.0093318] [Medline: 24681826]
32. Chen J, Or CK, Chen T. Effectiveness of using virtual reality-supported exercise therapy for upper extremity motor rehabilitation in patients with stroke: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Jun 20;24(6):e24111 [FREE Full text] [doi: 10.2196/24111] [Medline: 35723907]
33. Iruthayarajah J, McIntyre A, Cotoi A, Macaluso S, Teasell R. The use of virtual reality for balance among individuals with chronic stroke: a systematic review and meta-analysis. Top Stroke Rehabil 2017 Jan 16;24(1):68-79. [doi: 10.1080/10749357.2016.1192361] [Medline: 27309680]
34. Lee H, Park Y, Park S. The effects of virtual reality training on function in chronic stroke patients: a systematic review and meta-analysis. Biomed Res Int 2019 Jun 18;2019:7595639-7595n/a [FREE Full text] [doi: 10.1155/2019/7595639] [Medline: 31317037]
35. World Health Organization (WHO). International classification of functioning, disability and health: ICF. WHO. 2001. URL: https://apps.who.int/iris/handle/10665/42407 [accessed 2024-09-11]
36. Kern F, Winter C, Gall D, Käthner I, Pauli P, Latoschik M. Immersive virtual reality and gamifigation within procedurally generated environments to increase motivation during gait rehabilitation. 2019 Mar 27 Presented at: 2019 IEEE Conference on virtual reality and 3D user interfaces (VR); March 23-27, 2019; Osaka, Japan. [doi: 10.1109/vr.2019.8797828]
37. Sjögren T, Waller B, Rintala A, Paltamaa J, Krohn M, Aalto S, et al. The effectiveness of therapeutic exercise on activity and participation with chronic stroke: a systematic review with meta-analysis. PROSPERO. 2020. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020184572 [accessed 2024-09-11]
38. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Moher D, Liberati A, Tetzlaff J, Altman D, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
40. Systematic Review Tool. Covidence. URL: https://www.covidence.org [accessed 2020-01-05]
41. O'Connor D, Green. Chapter 5: Defining the review question and developing criteria for including studies. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions (version 5.1.0). London, UK: The Cochrane Collaboration; Mar 2011.
42. Rehme AK, Eickhoff SB, Rottschy C, Fink GR, Grefkes C. Activation likelihood estimation meta-analysis of motor-related neural activity after stroke. Neuroimage 2012 Feb 01;59(3):2771-2782. [doi: 10.1016/j.neuroimage.2011.10.023] [Medline: 22023742]
43. Rintala A, Kossi O, Bonnechère B, Evers L, Printemps E, Feys P. Mobile health applications for improving physical function, physical activity, and quality of life in stroke survivors: a systematic review. Disabil Rehabil 2023 Dec 02;45(24):4001-4015. [doi: 10.1080/09638288.2022.2140844] [Medline: 36325613]
44. Wu P, Zeng F, Li Y, Yu B, Qiu L, Qin W, et al. Changes of resting cerebral activities in subacute ischemic stroke patients. Neural Regen Res 2015 May;10(5):760-765 [FREE Full text] [doi: 10.4103/1673-5374.156977] [Medline: 26109950]
45. Guide to physical therapist practice. American Physical Therapy Association. 2014. URL: http://guidetoptpractice.apta.org [accessed 2024-11-18]
46. Schultheis MT, Rizzo AA. The application of virtual reality technology in rehabilitation. Rehabilitation Psychology 2001 Aug;46(3):296-311. [doi: 10.1037/0090-5550.46.3.296]
47. World Health Organization (WHO). How to use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for comment. WHO. 2013 Oct. URL: https://cdn.who.int/media/docs/default-source/classification/icf/drafticfpracticalmanual2.pdf?sfvrsn=8a214b01_4&download=true [accessed 2024-11-18]
48. Higgins J, Altman D, Sterne J. Assessing risk of bias. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions (version 5.1.0). London, UK: The Cochrane Collaboration; 2011.
49. Lee C, Kim Y, Lee B. Augmented reality-based postural control training improves gait function in patients with stroke: randomized controlled trial. Hong Kong Physiotherapy Journal 2014 Dec;32(2):51-57. [doi: 10.1016/j.hkpj.2014.04.002]
50. The Cochrane Collaboration. URL: https://training.cochrane.org/online-learning/core-software-cochrane-reviews/revman [accessed 2020-06-13]
51. Deeks J, Higgins J, Altman D. Analysing dataundertaking meta-analyses. In: Higgins J, Green S, editors. Cochrane Handbook of Systematic Reviews of Interventions (version 5.1.0). London, UK: The Cochrane Collaboration; 2011.
52. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002-d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
53. Baker W, White CM, Cappelleri J, Kluger J, Coleman C, Health Outcomes, Policy,Economics (HOPE) Collaborative Group. Understanding heterogeneity in meta-analysis: the role of meta-regression. Int J Clin Pract 2009 Oct 14;63(10):1426-1434. [doi: 10.1111/j.1742-1241.2009.02168.x] [Medline: 19769699]

XSL•FO
RenderX

54. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. URL: https://www.R-project.org/ [accessed 2023-10-24]
55. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Soft 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
56. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE Handbook. GRADEpro. 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html#h.g2dqzi9je57e [accessed 2019-09-29]
57. Fleiss JL. Statistical Methods for Rates and Proportions (2nd ed.). New York, NY: John Wiley and Sons; 1981.
58. Cho. Virtual walking training program using a real-world video recording for patients with chronic stroke: a pilot study. Am J Phys Med Rehab 2013 May;92(5):371-380. [doi: 10.1097/phm.0b013e31828cd5d3]
59. Cho KH, Lee WH. Effect of treadmill training based real-world video recording on balance and gait in chronic stroke patients: a randomized controlled trial. Gait Posture 2014 Jan;39(1):523-528. [doi: 10.1016/j.gaitpost.2013.09.003] [Medline: 24091250]
60. Cho KH, Lee KJ, Song CH. Virtual-reality balance training with a video-game system improves dynamic balance in chronic stroke patients. Tohoku J Exp Med 2012 Sep;228(1):69-74 [FREE Full text] [doi: 10.1620/tjem.228.69] [Medline: 22976384]
61. Choi D, Choi W, Lee S. Influence of Nintendo Wii Fit balance game on visual perception, postural balance, and walking in stroke survivors: a pilot randomized clinical trial. Games for Health Journal 2018 Dec;7(6):377-384. [doi: 10.1089/g4h.2017.0126]
62. Jung J, Yu J, Kang H. Effects of virtual reality treadmill training on balance and balance self-efficacy in stroke patients with a history of falling. J Phys Ther Sci 2012;24(11):1133-1136. [doi: 10.1589/jpts.24.1133]
63. Jung J, Goo B, Lee D, Roh H. Effects of 3D visual feedback exercise on the balance and walking abilities of hemiplegic patients. J Phys Ther Sci 2011;23(6):859-862. [doi: 10.1589/jpts.23.859]
64. Kim H, Choi W, Lee, Song C. Virtual dual-task treadmill training using video recording for gait of chronic stroke survivors: a randomized controlled trial. J Phys Ther Sci 2015 Dec 27;27(12):3693-3697 [FREE Full text] [doi: 10.1589/jpts.27.3693] [Medline: 26834334]
65. Kim N, Park Y, Lee B. Effects of community-based virtual reality treadmill training on balance ability in patients with chronic stroke. J Phys Ther Sci 2015 Mar;27(3):655-658 [FREE Full text] [doi: 10.1589/jpts.27.655] [Medline: 25931701]
66. Kim I, Lee B. Effects of augmented reality with functional electric stimulation on muscle strength, balance and gait of stroke patients. J Phys Ther Sci 2012;24(8):755-762. [doi: 10.1589/jpts.24.755]
67. Lee D, Bae Y. The effectiveness of driving game on trunk control and gait ability in stroke. J Mot Behav 2020 Feb 22;52(1):33-40. [doi: 10.1080/00222895.2019.1574259] [Medline: 30794093]
68. Lee HY, Kim YL, Lee SM. Effects of virtual reality-based training and task-oriented training on balance performance in stroke patients. J Phys Ther Sci 2015 Jun;27(6):1883-1888. [doi: 10.1589/jpts.27.1883] [Medline: 26180341]
69. Lee I, Kim Y, Lee D. Effect of a virtual reality exercise program accompanied by cognitive tasks on the balance and gait of stroke patients. J Phys Ther Sci 2015 Jul;27(7):2175-2177 [FREE Full text] [doi: 10.1589/jpts.27.2175] [Medline: 26311949]
70. Park Y, Lee C, Lee B. Clinical usefulness of the virtual reality-based postural control training on the gait ability in patients with stroke. J Exerc Rehabil 2013 Oct 31;9(5):489-494 [FREE Full text] [doi: 10.12965/jer.130066] [Medline: 24282810]
71. Shin DC, Song CH. Smartphone-based visual feedback trunk control training using a gyroscope and mirroring technology for stroke patients: single-blinded, randomized clinical trial of efficacy and feasibility. Am J Phys Med Rehabil 2016 May;95(5):319-329. [doi: 10.1097/PHM.0000000000000447] [Medline: 26829087]
72. Song GB, Park EC. Effect of virtual reality games on stroke patients' balance, gait, depression, and interpersonal relationships. J Phys Ther Sci 2015 Jul;27(7):2057-2060 [FREE Full text] [doi: 10.1589/jpts.27.2057] [Medline: 26311925]
73. Yom C, Cho H, Lee B. Effects of virtual reality-based ankle exercise on the dynamic balance, muscle tone, and gait of stroke patients. J Phys Ther Sci 2015 Mar;27(3):845-849 [FREE Full text] [doi: 10.1589/jpts.27.845] [Medline: 25931744]
74. Yu JH, Cho KH. Effectiveness of virtual reality game on functional movement and activities of daily living in hemiparetic stroke patients. Journal of Nanoelectronics and Optoelectronics 2016 Feb 01;11(1):98-102. [doi: 10.1166/jno.2016.1875]
75. Givon N, Zeilig G, Weingarden H, Rand D. Video-games used in a group setting is feasible and effective to improve indicators of physical activity in individuals with chronic stroke: a randomized controlled trial. Clin Rehabil 2016 Apr 07;30(4):383-392. [doi: 10.1177/0269215515584382] [Medline: 25952586]
76. In T, Lee K, Song C. Virtual reality reflection therapy improves balance and gait in patients with chronic stroke: randomized controlled trials. Med Sci Monit 2016 Oct 28;22:4046-4053 [FREE Full text] [doi: 10.12659/msm.898157] [Medline: 27791207]
77. Kang H, Kim Y, Chung Y, Hwang S. Effects of treadmill training with optic flow on balance and gait in individuals following stroke: randomized controlled trials. Clin Rehabil 2012 Mar 04;26(3):246-255. [doi: 10.1177/0269215511419383] [Medline: 21971754]
78. Kim S, Cho H, Kim K, Lee S. Effects of ankle biofeedback training on strength, balance, and gait in patients with stroke. J Phys Ther Sci 2016 Sep;28(9):2596-2600 [FREE Full text] [doi: 10.1589/jpts.28.2596] [Medline: 27799701]
79. Kim EK, Kang JH, Park JS, Jung BH. Clinical feasibility of interactive commercial Nintendo gaming for chronic stroke rehabilitation. J Phys Ther Sci 2012;24(9):901-903. [doi: 10.1589/jpts.24.901]

XSL•FO
RenderX

80. Lloréns R, Noé E, Colomer C, Alcañiz M. Effectiveness, usability, and cost-benefit of a virtual reality-based telerehabilitation program for balance recovery after stroke: a randomized controlled trial. Arch Phys Med Rehabil 2015 Mar;96(3):418-425.e2. [doi: 10.1016/j.apmr.2014.10.019] [Medline: 25448245]
81. Park D, Lee D, Lee K, Lee G. Effects of virtual reality training using Xbox Kinect on motor function in stroke survivors: a preliminary study. J Stroke Cerebrovasc Dis 2017 Oct;26(10):2313-2319. [doi: 10.1016/j.jstrokecerebrovasdis.2017.05.019] [Medline: 28606661]
82. Fritz SL, Peters DM, Merlo AM, Donley J. Active video-gaming effects on balance and mobility in individuals with chronic stroke: a randomized controlled trial. Top Stroke Rehabil 2013 Jan 05;20(3):218-225. [doi: 10.1310/tsr2003-218] [Medline: 23841969]
83. Alptekin N, Gok H, Geler-Kulcu D, Dincer G. Efficacy of treatment with a kinaesthetic ability training device on balance and mobility after stroke: a randomized controlled study. Clin Rehabil 2008 Oct 01;22(10-11):922-930. [doi: 10.1177/0269215508090673] [Medline: 18955424]
84. Kim JH, Jang SH, Kim CS, Jung JH, You JH. Use of virtual reality to enhance balance and ambulation in chronic stroke: a double-blind, randomized controlled study. Am J Phys Med Rehabil 2009 Sep;88(9):693-701. [doi: 10.1097/PHM.0b013e3181b33350] [Medline: 19692788]
85. Lee H, Huang C, Ho S, Sung W. The effect of a virtual reality game intervention on balance for patients with stroke: a randomized controlled trial. Games Health J 2017 Oct;6(5):303-311. [doi: 10.1089/g4h.2016.0109] [Medline: 28771379]
86. Lloréns R, Gil-Gómez JA, Alcañiz M, Colomer C, Noé E. Improvement in balance using a virtual reality-based stepping exercise: a randomized controlled trial involving individuals with chronic stroke. Clin Rehabil 2015 Mar 23;29(3):261-268. [doi: 10.1177/0269215514543333] [Medline: 25056999]
87. Lin K, Chen C, Chen Y, Huang W, Lai J, Yu S, et al. Bidirectional and multi-user telerehabilitation system: clinical effect on balance, functional activity, and satisfaction in patients with chronic stroke living in long-term care facilities. Sensors 2014 Jul 11;14(7):12451-12466 [FREE Full text] [doi: 10.3390/s140712451] [Medline: 25019632]
88. Hung J, Chou C, Chang H, Wu W, Hsieh Y, Chen P, et al. Cognitive effects of weight-shifting controlled exergames in patients with chronic stroke: a pilot randomized comparison trial. Eur J Phys Rehabil Med 2017 Oct;53(5):694-702. [doi: 10.23736/s1973-9087.17.04516-6]
89. Fishbein P, Hutzler Y, Ratmansky M, Treger I, Dunsky A. A preliminary study of dual-task training using virtual reality: influence on walking and balance in chronic poststroke survivors. J Stroke Cerebrovasc Dis 2019 Nov;28(11):104343. [doi: 10.1016/j.jstrokecerebrovasdis.2019.104343] [Medline: 31495668]
90. Lee SH, Byun SD, Kim CH, Go JY, Nam HU, Huh JS, et al. Feasibility and effects of newly developed balance control trainer for mobility and balance in chronic stroke patients: a randomized controlled trial. Ann Rehabil Med 2012 Aug;36(4):521-529. [doi: 10.5535/arm.2012.36.4.521] [Medline: 22977778]
91. Yim J. Effects of respiratory muscle and endurance training using an individualized training device on the pulmonary function and exercise capacity in stroke patients. Med Sci Monit 2014;20:2543-2549. [doi: 10.12659/msm.891112]
92. Barcala L, Grecco LAC, Colella F, Lucareli PRG, Salgado ASI, Oliveira CS. Visual biofeedback balance training using wii fit after stroke: a randomized controlled trial. J Phys Ther Sci 2013 Aug;25(8):1027-1032 [FREE Full text] [doi: 10.1589/jpts.25.1027] [Medline: 24259909]
93. Choi H, Shin W, Bang D, Choi S. Effects of game-based constraint-induced movement therapy on balance in patients with stroke. American Journal of Physical Medicine & Rehabilitation 2017;96(3):184-190. [doi: 10.1097/phm.0000000000000567]
94. Drużbicki M, Kwolek A, Depa A, Przysada G. The use of a treadmill with biofeedback function in assessment of relearning walking skills in post-stroke hemiplegic patients – a preliminary report. Neurologia i Neurochirurgia Polska 2010;44(6):567-573. [doi: 10.1016/s0028-3843(14)60154-7] [Medline: 21225519]
95. Bang Y, Son KH, Kim HJ. Effects of virtual reality training using Nintendo Wii and treadmill walking exercise on balance and walking for stroke patients. J Phys Ther Sci 2016 Nov;28(11):3112-3115 [FREE Full text] [doi: 10.1589/jpts.28.3112] [Medline: 27942130]
96. Hung J, Chou C, Hsieh Y, Wu W, Yu M, Chen P, et al. Randomized comparison trial of balance training by using exergaming and conventional weight-shift therapy in patients with chronic stroke. Arch Phys Med Rehabil 2014 Sep;95(9):1629-1637. [doi: 10.1016/j.apmr.2014.04.029] [Medline: 24862764]
97. Pedreira da Fonseca E, da Silva Ribeiro NM, Pinto EB. Therapeutic effect of virtual reality on post-stroke patients: randomized clinical trial. J Stroke Cerebrovasc Dis 2017 Jan;26(1):94-100. [doi: 10.1016/j.jstrokecerebrovasdis.2016.08.035] [Medline: 27693404]
98. Yatar GI, Yildirim SA. Wii Fit balance training or progressive balance training in patients with chronic stroke: a randomised controlled trial. J Phys Ther Sci 2015 Apr;27(4):1145-1151 [FREE Full text] [doi: 10.1589/jpts.27.1145] [Medline: 25995576]
99. Yang Y, Tsai M, Chuang T, Sung W, Wang R. Virtual reality-based training improves community ambulation in individuals with stroke: a randomized controlled trial. Gait Posture 2008 Aug;28(2):201-206. [doi: 10.1016/j.gaitpost.2007.11.007] [Medline: 18358724]
100. Tetzlaff B, Barzel A, Stark A, Ketels G, Scherer M. To what extent does therapy of chronic stroke patients address participation? A content analysis of ambulatory physical and occupational therapy based on the International Classification

XSL•FO
RenderX

of Functioning, Disability, and Health framework. Disabil Rehabil 2020 Feb 18;42(4):545-551. [doi: 10.1080/09638288.2018.1503732] [Medline: 30451030]
101. Finnish Association of Physiotherapists. The Core Competences of a Physiotherapist. Helsinki, Finland: Finnish Association of Physiotherapists; 2016:1-32.
102. Bower K, Thilarajah S, Pua Y, Williams G, Tan D, Mentiplay B, et al. Dynamic balance and instrumented gait variables are independent predictors of falls following stroke. J Neuroeng Rehabil 2019 Jan 07;16(1):3 [FREE Full text] [doi: 10.1186/s12984-018-0478-4] [Medline: 30612584]
103. Hohenschurz-Schmidt D, Draper-Rodi J, Vase L, Scott W, McGregor A, Soliman N, et al. Blinding and sham control methods in trials of physical, psychological, and self-management interventions for pain (article II): a meta-analysis relating methods to trial results. Pain 2023 Mar 01;164(3):509-533 [FREE Full text] [doi: 10.1097/j.pain.0000000000002730] [Medline: 36271798]
104. Higgins J, Savović J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventionsversion (version 6.5). London, UK: The Cochrane Collaboration; 2024.
105. Chan A, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004 May 26;291(20):2457-2465. [doi: 10.1001/jama.291.20.2457] [Medline: 15161896]
106. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 2008 Aug 28;3(8):e3081 [FREE Full text] [doi: 10.1371/journal.pone.0003081] [Medline: 18769481]
107. Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions (Version 6.5). London, UK: The Cochrane Collaboration; 2024.
108. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf MI, et al. Chapter 4: Searching for and selecting studies [last updated September 2024]. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions (version 6.5). London, UK: The Cochrane Collaboration; 2024.
109. Komulainen J, Jousimaa J. Hoitosuositusryhmien käsikirja. Terveysportti. 2019. URL: https://www.terveysportti.fi/dtk/khk/khk00007 [accessed 2024-08-19]
110. Sjögren T, Korpi H. Supporting sense of meaningful life and human dignity in digitally assisted physiotherapy environment – qualitative secondary research with thematic analyses and inductive synthesis. In: Särestöniemi M, Keikhosrokiani P, Singh D, Harjula E, Tiulpin A, Jansson M, et al, editors. Digital Health and Wireless Solutions. NCDHWS 2024. Cham, Switzerland: Springer; 2024:371-386.
111. Saposnik G, Cohen LG, Mamdani M, Pooyania S, Ploughman M, Cheung D, et al. Efficacy and safety of non-immersive virtual reality exercising in stroke rehabilitation (EVREST): a randomised, multicentre, single-blind, controlled trial. The Lancet Neurology 2016 Sep;15(10):1019-1027. [doi: 10.1016/s1474-4422(16)30121-1]
112. Holden MK. Virtual environments for motor rehabilitation: review. Cyberpsychol Behav 2005 Jun;8(3):187-211; discussion 212. [doi: 10.1089/cpb.2005.8.187] [Medline: 15971970]
113. Clifton J, Palmisano S. Effects of steering locomotion and teleporting on cybersickness and presence in HMD-based virtual reality. Virtual Reality 2019 Nov 02;24(3):453-468. [doi: 10.1007/s10055-019-00407-8]
114. Rohrbach N, Chicklis E, Levac DE. What is the impact of user affect on motor learning in virtual environments after stroke? A scoping review. J Neuroeng Rehabil 2019 Jun 27;16(1):79 [FREE Full text] [doi: 10.1186/s12984-019-0546-4] [Medline: 31248439]
115. Aivoverenkierronhäiriöt (AVH) lukuina. Aivoliitto. 2013. URL: http://aivoliitto.fi [accessed 2020-05-07]
116. Appelros P, Stegmayr B, Tere´nt A. Sex differences in stroke epidemiology. Stroke 2009 Apr;40(4):1082-1090. [doi: 10.1161/strokeaha.108.540781]
117. Donoghue D, Physiotherapy ResearchOlder People (PROP) group, Stokes EK. How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. J Rehabil Med 2009 Apr;41(5):343-346 [FREE Full text] [doi: 10.2340/16501977-0337] [Medline: 19363567]
118. Downs S, Marquez J, Chiarelli P. The Berg Balance Scale has high intra- and inter-rater reliability but absolute reliability varies across the scale: a systematic review. Journal of Physiotherapy 2013 Jun;59(2):93-99. [doi: 10.1016/s1836-9553(13)70161-9]

Abbreviations

10mWT: 10-meter Walk Test
2MWT: 2-minute Walk Test
6MWT: 6-minute Walk Test

XSL•FO
RenderX

ABC: Activities-Specific Balance Confidence Scale
BBS: Berg Balance Scale
CENTRAL: Cochrane Controlled Trials Register
CINAHL: Cumulative Index to Nursing and Allied Health Literature
Embase: Excerpta Medica Database
FRT: Functional Reach Test
GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
HMD: head-mounted display
ICF: International Classification of Functioning, Disability, and Health
MMSE: Mini-Mental State Examination
PEDro: Physiotherapy Evidence Database
PICOS: Population, Intervention, Comparison, Outcomes, Study
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference
TUG: Timed "Up and Go"
VR: virtual reality

Edited by A Coristine; submitted 04.04.24; peer-reviewed by A Azizan, G Rodriguez-Fuentes; comments to author 08.08.24; revised version received 29.09.24; accepted 19.10.24; published 02.12.24.

Please cite as:
Krohn M, Rintala A, Immonen J, Sjögren T
The Effectiveness of Therapeutic Exercise Interventions With Virtual Reality on Balance and Walking Among Persons With Chronic Stroke: Systematic Review, Meta-Analysis, and Meta-Regression of Randomized Controlled Trials
J Med Internet Res 2024;26:e59136
URL: https://www.jmir.org/2024/1/e59136
doi:10.2196/59136
PMID:39621381

©Maria Krohn, Aki Rintala, Jaakko Immonen, Tuulikki Sjögren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Nurses' Knowledge and Skills After Use of an Augmented Reality App for Advanced Cardiac Life Support Training: Randomized Controlled Trial

Wan-Na Sun[1,2,3], MS; Min-Chai Hsieh[4], PhD; Wei-Fang Wang[1,2], PhD

[1]Department of Nursing, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[2]Department of Nursing, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[3]College of Nursing, Chung Hwa University of Medical Technology, Tainan, Taiwan
[4]Tainan University of Technology, Tainan, Taiwan

Corresponding Author:
Wei-Fang Wang, PhD
Department of Nursing
National Cheng Kung University Hospital
College of Medicine, National Cheng Kung University
No. 138, Sheng Li Road
Tainan, 704
Taiwan
Phone: 886 62353535 ext 4676
Fax: 886 62377550
Email: 820hosp@gmail.com

Abstract

Background: Advanced cardiac life support (ACLS) skills are essential for nurses. During the COVID-19 pandemic, augmented reality (AR) technologies were incorporated into medical education to increase learning motivation and accessibility.

Objective: This study aims to determine whether AR for educational applications can significantly improve crash cart learning, learning motivation, cognitive load, and system usability. It focused on a subgroup of nurses with less than 2 years of experience.

Methods: This randomized controlled trial study was conducted in a medical center in southern Taiwan. An ACLS cart training course was developed using AR technologies in the first stage. Additionally, the efficacy of the developed ACLS training course was evaluated. The AR group used a crash cart learning system developed with AR technology, while the control group received traditional lecture-based instruction. Both groups were evaluated immediately after the course. Performance was assessed through learning outcomes related to overall ACLS and crash cart use. The Instructional Materials Motivation Survey, System Usability Scale, and Cognitive Load Theory Questionnaire were also used to assess secondary outcomes in the AR group. Subgroup analyses were performed for nurses with less than 2 years of experience.

Results: All 102 nurses completed the course, with 43 nurses in the AR group and 59 nurses in the control group. The AR group outperformed the control group regarding overall ACLS outcomes and crash cart learning outcomes (P=.002; P=.01). The improvement rate was the largest for new staff regardless of the overall learning effect and the crash cart effect. Subgroup analysis revealed that nurses with less than 2 years of experience in the AR group showed more significant improvements in both overall learning (P<.001) and crash cart outcomes (P<.001) compared to their counterparts in the control group. For nurses with more than 2 years of experience, no significant differences were found between the AR and control groups in posttraining learning outcomes for the crash cart (P=.32). The AR group demonstrated high scores for motivation (Instructional Materials Motivation Survey mean score 141.65, SD 19.25) and system usability (System Usability Scale mean score 90.47, SD 11.91), as well as a low score for cognitive load (Cognitive Load Theory Questionnaire mean score 15.42, SD 5.76).

Conclusions: AR-based learning significantly improves ACLS knowledge and skills, especially for nurses with less experience, compared to traditional methods. The high usability and motivational benefits of AR suggest its potential for broader applications in nursing education.

Trial Registration: ClinicalTrials.gov NCT06057285; https://clinicaltrials.gov/ct2/show/NCT06057285

(J Med Internet Res 2024;26:e57327) doi:10.2196/57327

XSL•FO
RenderX

KEYWORDS

augmented reality; technology intervention; randomized controlled trial; advanced cardiac life support; nursing education

Introduction

Challenges in Advanced Cardiac Life Support Training and Nurse Preparedness

Advanced cardiac life support (ACLS) is an essential and indispensable nursing skill. A nurse with adequate ACLS skills can provide a patient with comprehensive first aid treatment in an emergency, avoiding missing the "golden hour" of treatment [1]. In-hospital cardiac arrest (IHCA) affected 292,000 patients in the United States between 2008 and 2017 each year, equivalent to IHCA occurring in nearly 10 out of every 1000 inpatients [2]. The IHCA incidence rate increased sharply during the COVID-19 pandemic, with a 30-day survival rate after first aid of merely 25% [2,3]. In Taiwan, 957 incidents of unexpected IHCA occurred in 2020, accounting for 1.2% of all patient safety incidents that year and ranking first among acute life-threatening events for patients (51.1%) [4]. During an IHCA, nurses are generally responsible for determining and providing appropriate first aid measures. Therefore, training and enhancing ACLS skills among nurses is crucial. A lack of experience, knowledge, and familiarity with ACLS cart–related skills may have contributed to the underperformance of nurses delivering first aid [5,6]. In first aid scenarios, nurses who lack clinical experience generally experience immense stress due to the immediacy and imminence of patient rescue. Chen and Tzeng [7] surveyed 164 nurses with less than 2 years of experience in terms of their knowledge of ACLS medications and perceived stress during the delivery of ACLS and found that 94% of the nurses experienced immense stress delivering first aid. The stress was more significant for those with less knowledge of ACLS medications (P=.002). Moreover, Chen and Tzeng [7] reported that only 6% of the nurses were adequately trained through ACLS training programs, and the best time to implement relevant training programs was deemed by 41.8% of nurses between 3 and 6 months after starting work.

Challenges in Traditional ACLS Training

Current ACLS training programs in Taiwan are generally conducted through experience-sharing, case discussions, or bedside teaching, which is challenging for nurses with insufficient clinical experience and does not give every nurse a chance to practice [1]. Furthermore, traditional teaching methods were suspended or hindered in 2019 due to the COVID-19 pandemic. Clinical ACLS education should focus on practical learning and teaching material accessibility. The scientific statement titled "Resuscitation Education Science: Educational Strategies to Improve Outcomes from Cardiac Arrest," published by the American Heart Association, mentions that immersive technologies and gamified learning effectively enhance learning outcomes [8]. The attention, relevance, confidence, and satisfaction model of learning motivation is a teaching strategy model proposed by Keller [9,10] in 1987. An instructional process based on this model first draws the learner's attention, demonstrates relevance with them through personal or relevant experience, and elevates his or her confidence during learning, thereby achieving satisfaction. For Keller [9,10], maintaining attention is crucial. The development of medical technology has trended toward integrating artificial intelligence into medical systems. The COVID-19 pandemic further acted as a catalyst for the incorporation of augmented reality (AR), virtual reality (VR), and even mixed reality into medical education, in which relevance is demonstrated to the learner through actual and virtual reality simulated scenarios, enabling them to learn independently without the constraints of location, time, and other factors, thereby enhancing their confidence. Learner satisfaction is increased through repeatable practice beyond time and location constraints [11].

Trends in the Adoption of AR in Health Care

The AR industries are not industries of emerging technologies, and their flourishing development in recent years into various domains has been catalyzed through their progressive technological maturation and the COVID-19 pandemic. Application of AR technologies in health care has increased by 25% over the past 3 years, and in 2028, the annual economic output value was estimated to exceed US $9.5 billion [12]. AR involves superimposing virtual information onto an actual environment through a mobile device, with the user viewing the virtual image and data in real-life scenarios shown on the mobile device's screen; AR is characterized by a combination of interaction between the real world and virtual information [13,14]. AR is more valuable than VR for health care education. VR involves computer-based simulation of real-life scenarios in which the user interacts with and views the simulation through a device. Therefore, VR requires a headset and another person's assistance while immersed in a virtual environment. Furthermore, VR has been associated with physical discomforts, such as dizziness, nausea, vomiting, and sore eyes, that can affect or interrupt the learning process and decrease learning motivation [14]. AR systems are more portable than VR systems. AR can be operated using a simple mobile device, such as a cellphone or a tablet, increasing the accessibility of teaching materials and equipment and allowing drills to be conducted through real-life scenarios and images. While mobile-based AR offers the convenience of accessibility through smartphones or tablets, it is essential to note that AR technology is not limited to mobile devices. AR can also be performed using head-mounted displays (HMDs), which provide immersive experiences but are typically not considered "mobile" in the same sense as smartphones or AR glasses. AR systems can be delivered through various devices, including mobile-based platforms such as smartphones or AR glasses, which are generally more portable than VR systems. However, AR can also be performed using HMDs, which offer a more immersive experience but are typically less portable than mobile devices. Mobile-based AR and HMDs offer unique advantages depending on the training context. Mobile-based AR and HMD-based AR offer unique advantages depending on the context of their use, with mobile-based AR being more accessible for everyday learning environments and HMDs offering enhanced immersion for more specialized training [15]. In health care education,

XSL•FO
RenderX

drills tend to be conducted based on the operation of related technologies. Many studies have indicated that learning materials developed using AR technology can effectively enhance participants' performance and improve focus. This is especially important in clinical work settings, where simulations are required. AR technology allows learners to study independently, significantly improving learning outcomes [11].

The studies have demonstrated that education on ACLS nursing skills is an important issue, and an appropriate model of learning motivation should first trigger the learner's motivation and enable them to identify the relevance of the content learned. This study used AR technologies to develop a first-aid cart learning system. The accessibility of mobile devices enables learners to surpass location and time constraints and learn by relating to crash cart–related actual clinical images and sites, thereby enhancing their confidence and increasing their satisfaction with acquiring ACLS knowledge and skills.

Methods

Study Design, Participants, and Setting

This interventional and experimental study was registered on ClinicalTrials.gov (NCT06057285). A single-blinded research method was adopted, and participants were recruited through convenience sampling. The research setting was a medical center in southern Taiwan. All nurses were required to have a university degree as part of their work at the institution. This study was divided into 2 stages. In the first stage, a crash cart learning system was developed using AR technologies, and to enhance learning accessibility, necessary teaching materials for ACLS drills (ie, the structure and content of a crash cart) were digitized to enable the operation of this system on mobile devices (Figure 1).

The crash cart learning system uses AR technology to transform a physical crash cart into a virtual teaching tool comprising five parts. The first part includes 13 emergency medications, detailing their effects, side effects, and dosages. The second part covers infusions, such as normal saline, Dextrose 5%, lactated Ringer's solution, and blood draw supplies. The third part features intubation equipment, including various sizes of endotracheal tubes. The fourth part covers tracheostomy tubes, laryngeal mask airways, and central venous catheters. The fifth part covers Ambu bags and suction devices. In addition to simulating a physical crash cart, the system provides instructions for each item.

In the second stage, the developed crash cart learning system was applied to ACLS drills in a regular ACLS training program, and an AR group used this learning system to assist in their learning. The AR group scanned the dedicated AR marker using a mobile device (Figure 2) to learn about emergency medications, intubation equipment, and other tools. In contrast, the control group received instruction through lectures. Regardless of the group, the learning content included using emergency drugs, side effects and dosage, intubation, infusion, and ACLS procedures. The same lecturer and teaching assistant conducted it, and each class lasted about one hour. After the class, the learning effectiveness was evaluated immediately.

Figure 1. Actual operation process.

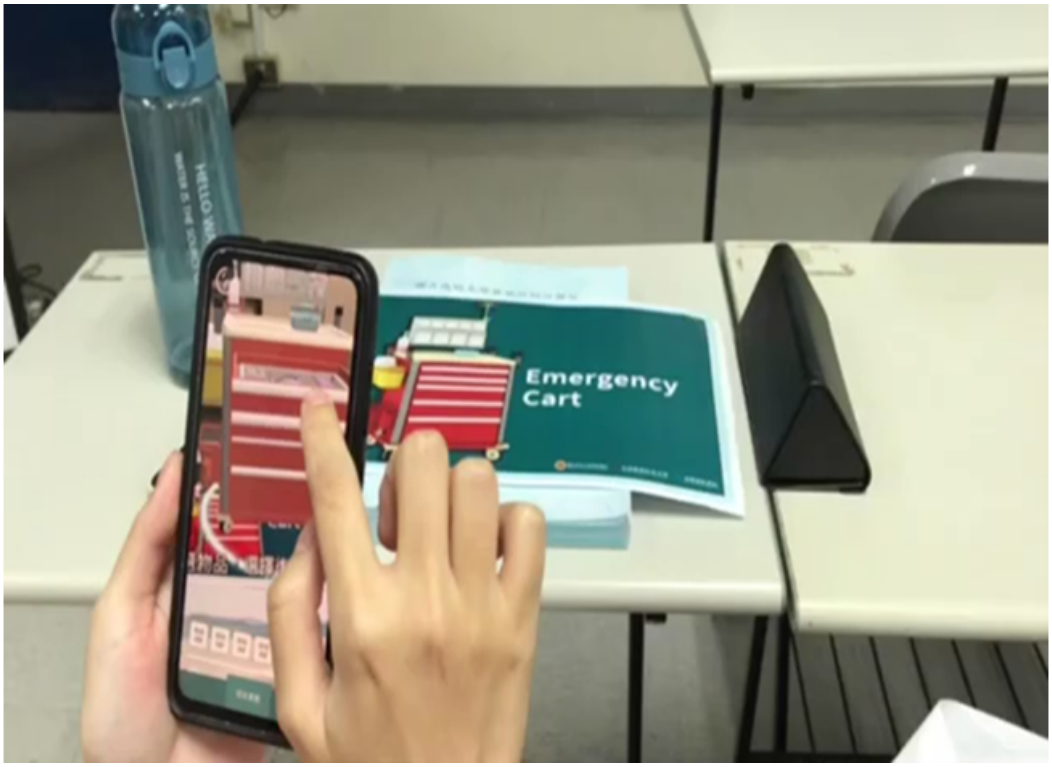

XSL•FO
RenderX

Figure 2. AR marker of the crash cart. The AR marker image opens the crash cart learning system when scanned with a mobile device.

Participants

Nurses aged 20 years or older, who could speak and read Mandarin; worked in intensive care units, emergency rooms, or general wards; and participated in resuscitation processes were recruited. All participants completed the questionnaire after the research objective was explained. Nurses working in pediatric departments, as well as nursing supervisors or part-time nurses, were excluded.

A priori power analysis was conducted using G*Power (Heinrich-Heine-Universität Düsseldorf). The sample size (n=82) was obtained using a power level of 0.8, an α level set at .05, and a small to medium effect size of 0.3 [16,17]. The estimated dropout rate was 20%, and the estimated number of persons to be recruited was 99.

Instrumentation

The questionnaire used in this study was rated by 3 experts (Bo-Shen Chen, Jhen-Yu Ji, and Shu-Jhen Su), all of whom have at least 10 years of experience in their professional fields. The experts were a clinical teacher, a departmental teaching nurse, and a senior cardiopulmonary resuscitation (CPR) instructor rated as excellent by the education center for this study in 2019. The experts assessed the appropriateness, relevancy, and clarity of the questionnaire content, giving scores of 1 (inappropriate and to be deleted), 2 (appropriate but requiring vast amendment), 3 (appropriate but requiring slight adjustment), and 4 (extremely appropriate). After the assessment, items with a score of ≥3 were included in the calculation of the content validity index (CVI), and items with a score of <3 were amended or deleted.

Learning Outcomes Questionnaire

To measure the outcomes of the ACLS drills using AR-based teaching aids, this study developed a learning outcome questionnaire. The questionnaire was formulated according to the latest version of the *Advanced Cardiovascular Life Support Provider Manual*. To increase its reliability and validity, an open call was conducted to choose 3 experienced CPR instructors with a minimum of 5 years of experience. The items were classified as easy (20%), intermediate (50%), and difficult (30%). The CPR instructors created 25 questions, including 10 items that were based on the content of the crash cart learning system. The total score ranged from 0 to 100, with a higher score indicating better learning effectiveness. The appropriateness of the finalized learning outcome test was evaluated by 2 experienced CPR instructors other than those who created the items. This study demonstrated content validity with an item-CVI ranging from 0.88 to 0.95.

Instructional Materials Motivation Survey

The Instructional Materials Motivation Survey (IMMS) was developed, derived from the attention, relevance, confidence, and satisfaction model proposed by Keller [9,10]. It comprises 36 items, including 26 positively and 10 negatively worded items. It consists of 4 subscales, namely the attention subscale (12 items), relevance subscale (9 items), confidence subscale (9 items), and satisfaction subscale (6 items). Items are rated on a 5-point Likert scale with end points ranging from 1 (strongly disagree) to 5 (strongly agree), and the total score ranges between 36 and 180. Higher scores indicate stronger learning motivation. The scale exhibited good internal reliability, as evidenced by Cronbach α coefficients ranging from 0.81 to 0.96, and it also demonstrated content validity with a CVI of 0.85 [9].

XSL•FO
RenderX

System Usability Scale

Brooke used the System Usability Scale (SUS) in 1986 to test the overall usability of a product. The SUS measures a user's perception of use and comprises ten items, including five positively worded items and five negatively worded items, which are rated on a 5-point Likert scale with endpoints ranging from 1 (strongly disagree) to 5 (strongly agree). The total score was multiplied by 2.5 to obtain a final score, which ranged between 25 and 100, with higher total scores indicating higher satisfaction with the system and total scores <68 indicating inferior system usability [18]. The scale has exhibited good internal reliability, as evidenced by Cronbach α coefficient of 0.84, and it has also demonstrated content validity with a CVI of 0.97 [19,20].

Cognitive Load Theory Questionnaire

The cognitive load theory (CLT) proposed by Sweller et al [21] comprises endogenous and exogenous factors. Endogenous factors refer to the learner per se and are associated with their effort and mental ability. In contrast, exogenous factors refer to the external environment, such as teaching materials and environments. The Cognitive Load Theory Questionnaire contains 8 items rated on a 5-point Likert scale with endpoints ranging from 1 (strongly disagree) to 5 (strongly agree) and total scores ranging between 8 and 40. Higher scores indicate a more significant cognitive load perceived by the learner for the item. The scale has exhibited good internal reliability, as evidenced by Cronbach α coefficients ranging from 0.88 to 0.93, and it has also demonstrated content validity with a CVI of 0.96 [21,22].

Ethical Considerations

Before the second stage, this study was approved by the Institutional Review Board of the National Cheng Kung University Hospital (approval A-ER-109-383). Before the study commenced, the research plan was presented to the unit supervisors, and the study procedures were clearly explained to all participants. Participation in the study was entirely voluntary, and nonparticipation did not affect the participants' job duties or educational opportunities. Informed consent was obtained from all participants before enrolling in the study, and each participant signed a consent form. To ensure the privacy and confidentiality of the participants, all collected data were deidentified using numerical codes, and no personally identifiable information was linked to the data. The deidentified data were securely stored in a locked drawer accessible only to the research team. No monetary or other forms of compensation were provided to the participants for their involvement in the study.

Data Collection

Participants were recruited from 4 units with similar numbers of CPR deliveries, selected through the hospital's provided database and assigned using convenience sampling. Additionally, participants were grouped using convenience sampling in 2 consecutive ACLS training courses organized by the hospital for new nurses. This is different from the general ACLS course. It is a training course for all new nurses in the hospital. The drugs and tools used are all from the hospital, which increases the new nurses' familiarity with the hospital. The ACLS course was conducted after the research objective and course outline were presented to related executives and the participants and after the participants provided written consent.

The participants of both groups completed a learning outcome test before and after learning immediately by trained research assistants who were not involved in the training sessions. The participants of the AR group completed the IMMS, SUS, and Cognitive Load Theory Questionnaire the day after the course, which took approximately 10 minutes to complete.

Data Analysis

Database construction and statistical analysis were performed in SPSS (version 20.0; IBM Corp). A P value of <.05 was considered statistically significant. Descriptive statistics (numbers, percentages, means, and SD) were used to summarize the participants' demographic characteristics and item scores. The participants' essential attributes were tested for homogeneity using the independent 2-tailed t test and chi-square test. Moreover, descriptive statistics and 2-tailed t tests were conducted to investigate the effects of the crash cart learning system for ACLS training on nurses' performance in terms of learning effectiveness, motivation, system usability, and cognitive load. The government has implemented the nurse postgraduate year program in Taiwan to help nursing students transition after graduation. This program focuses on intensive training during the first 2 years of postgraduate study [23], aiding their adaptation to clinical work. Therefore, this study includes a specific subgroup analysis of nurses with less than 2 years of experience.

Results

Demographic Characteristics

Data were collected between August and December 2021. No participants dropped out. In total, 102 nurses completed the course, with 43 nurses in the AR group and 59 nurses in the control group. For an overview of participant flow, see the CONSORT (Consolidated Standards of Reporting Trials) diagram (Figure 3). The average age of the nurses was 26.18 (SD 5.88) years. Women formed the majority (n=95, 93.1%). Among the departments nurses worked in, nurses working in internal medical wards comprised the majority (n=59, 57.8%). A total of 63.7% (n=65) of the nurses had less than 2 years of work experience. Most nurses were ranked N-level (n=66, 64.7%; the nursing ladder in Taiwan is divided into 5 levels: N, N1, N2, N3, and N4, with higher levels indicating greater proficiency) and had an ACLS certificate (n=71, 69.6%). No significant differences in demographic characteristics were observed between the 2 groups (Table 1).

XSL•FO
RenderX

Figure 3. Study enrollment flowchart. AR: augmented reality; CLT: Cognitive Load Theory; IMMS: Instructional Materials Motivation Survey; SUS: System Usability Scale.

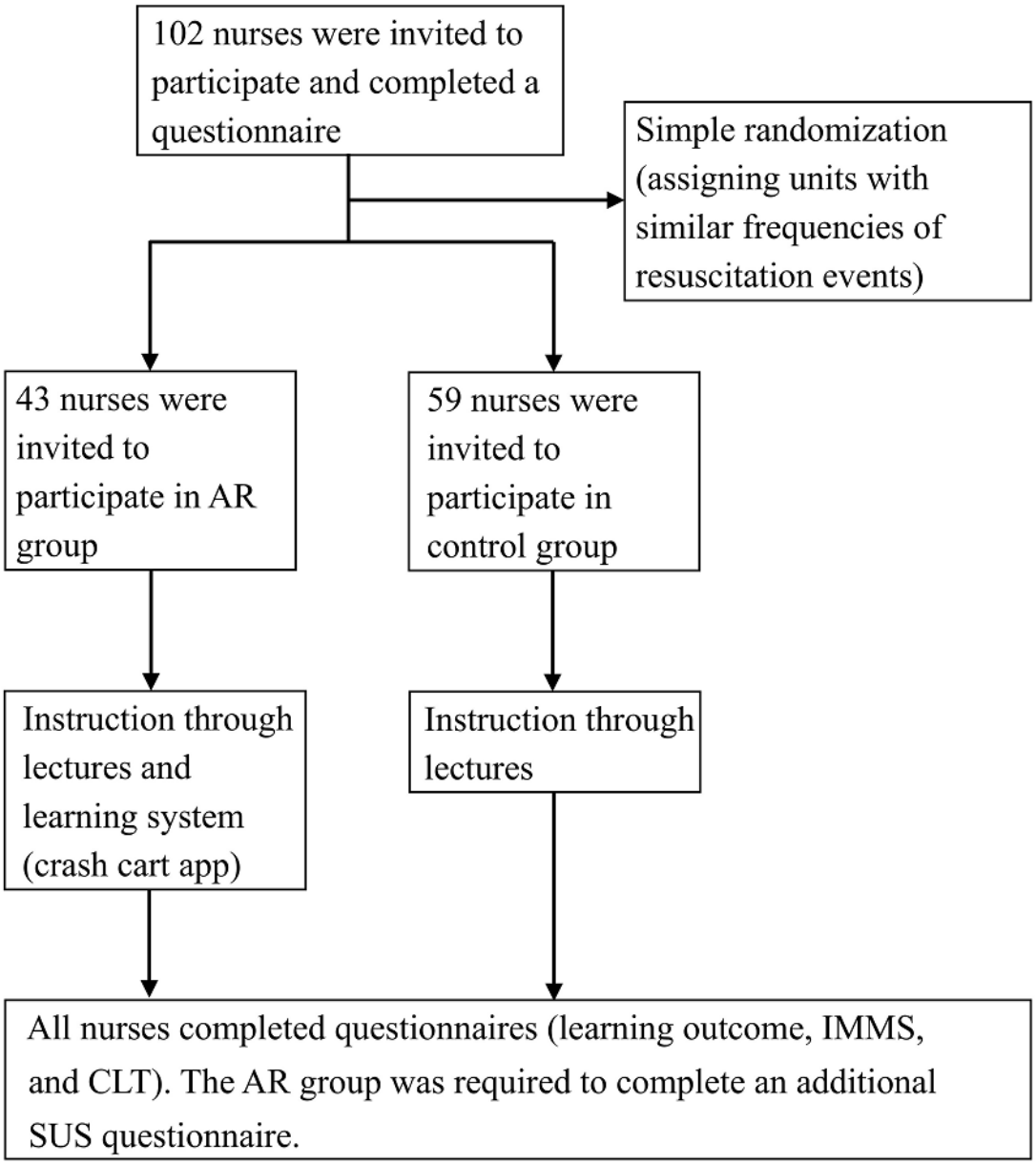

XSL•FO
RenderX

Table 1. Demographic data of the sample.

	All (n=102)	AR[a] group (n=43)	Control group (n=59)	*P* value
Age (years), mean (SD)	26.18 (5.88)	27.12 (6.63)	25.49 (5.21)	
Gender, n (%)				.45
Men	7 (6.9)	4 (9)	3 (5)	
Women	95 (93.1)	39 (91)	56 (95)	
Department, n (%)				.40
Internal medical	59 (57.8)	19 (44)	40 (68)	
Surgery	13 (12.7)	5 (12)	8 (14)	
Integrated branch	15 (14.7)	11 (26)	4 (7)	
Emergency	15 (14.7)	8 (19)	7 (12)	
Work experience, n (%)				.68
< 2 years	65 (63.7)	26 (61)	39 (66)	
2-6 years	24 (23.6)	11 (26)	13 (22)	
>6-8 years	13 (12.7)	6 (14)	7 (12)	
Nursing ladder[b], n (%)				.51
N	66 (64.6)	24 (56)	42 (71)	
N1-N2	27 (26.5)	13 (30)	14 (24)	
>N3	9 (8.9)	6 (14)	3 (5)	
ACLS[c] licenses, n (%)				.67
Yes	71 (69.6)	31 (72)	40 (38)	
No	31 (30.4)	12 (28)	19 (32)	

[a]AR: augmented reality.

[b]The nursing ladder in Taiwan is divided into 5 levels: N, N1, N2, N3, and N4. Nurses are promoted based on their caregiving, teaching, research, and administrative abilities, with higher levels indicating greater proficiency.

[c]ACLS: advanced cardiac life support.

Learning Effectiveness

For the overall learning outcomes in the AR group, the mean score was 61.12 (SD 14.72) before learning and 74.23 (SD 13.22) after learning, and the difference between overall outcomes before and after learning was significant (*P*<.001). For the control group, the mean score was 58.24 (SD 14.31) before learning, and 64.54 (SD 16.21) after learning, and the difference in the overall learning outcomes before and after learning was also significant (*P*<.001). The difference in learning outcomes between the 2 groups after learning was compared. The AR group outperformed the control group, and the difference was significant (*P*=.002; Table 2).

In Taiwan, new nurses undergo 2 years of training. Therefore, a subgroup analysis was conducted in which nurses were divided according to whether they had 2 years of experience. The mean score of the overall learning outcomes of the nurses of the AR group who had less than 2 years of experience was 52.77 (SD 10.25) before learning, and 68.31 (SD 12.41) after learning, and the difference before and after learning was significant (*P*<.001). For the overall learning outcomes of the nurses of the control group who had less than 2 years of experience, the mean score was 52.56 (SD 11.89) before learning, and 56.92 (SD 11.73) after learning, and the difference between learning outcomes before and after learning was significant (*P*=.01). The difference in learning outcomes after learning between the 2 groups of nurses with less than 2 years of experience was evaluated. The AR group outperformed the control group, and the difference was significant (*P*<.001).

XSL•FO
RenderX

Table 2. Learning effectiveness of the 2 groups.

	AR[a] group			Control group			*P* value[b]
	Pretest, mean (SD)	After test, mean (SD)	*P* value[c]	Pretest, mean (SD)	After test, mean (SD)	*P* value[c]	
All nurses[d]							
Learning outcome	61.12 (14.72)	74.23 (13.22)	<.001	58.24 (14.31)	64.54 (16.21)	<.001	.002
Learning outcome of crash cart	60.00 (22.99)	75.12 (20.75)	<.001	54.75 (23.88)	59.49 (27.32)	.02	.01
New nurses[e,f]							
Learning outcome	52.77 (10.25)	68.31 (12.41)	<.001	52.56 (11.89)	56.92 (11.73)	.01	<.001
Learning outcome of crash cart	46.54 (16.48)	68.46 (21.11)	<.001	43.59 (18.57)	46.15 (21.96)	.34	<.001
Experienced nurses[g,h]							
Learning outcome	73.88 (10.87)	83.29 (8.51)	.001	69.30 (12.14)	79.40 (13.25)	<.001	.80
Learning outcome of crash cart	80.59 (14.78)	85.29 (15.86)	.20	76.50 (17.25)	85.50 (15.38)	.001	.32

[a]AR: augmented reality.
[b]Between group *P* value.
[c]Within-group *P* value.
[d]For all nurses, n=43 for the AR group and n=59 for the control group.
[e]For new nurses, n=26 for the AR group and n=39 for the control group.
[f]Work experience of fewer than 2 years.
[g]For nurses, n=17 for the AR group and n=20 for the control group.
[h]Work experience of more than 2 years.

Among nurses with less than 2 years of experience, in the AR group, the mean score was 52.77 (SD 10.25) before learning and 68.31 (SD 12.41) after learning, and the difference between overall outcomes before and after learning was significant (*P*<.001). For the control group, the mean score was 52.56 (SD 11.89) before learning, and 56.92 (SD 11.73) after learning, and the difference in the overall learning outcomes before and after learning was also significant (*P*=.01). The difference in learning outcomes between the 2 groups after learning was compared. The AR group outperformed the control group, and the difference was significant (*P*<.001; Table 2). Among nurses with more than 2 years of experience, in the AR group, the mean overall learning outcomes score was 73.88 (SD 10.87) before learning and 83.29 (SD 8.51) after learning, and the difference was significant (*P*=.001). In the control group, the mean score was 69.30 (SD 12.14) before learning and 79.40 (SD 13.25) after learning, and the difference was significant (*P*<.001). The difference in learning outcomes after the course between the 2 groups of nurses with more than 2 years of experience was evaluated, and there was no significant difference (*P*=.80; Table 2).

Learning Effectiveness for the Crash Cart

Of the 25 questions in the learning outcomes test, 10 items addressed the content of a crash cart. For the overall learning outcomes of a crash cart in the AR group, the mean score was 60.00 (SD 22.99) before learning and 75.12 (SD 20.75) after learning, and the difference between learning outcomes for the crash cart before and after learning was significant (*P*<.001). For the control group, the mean score was 54.75 (SD 23.88) before learning and 59.49 (SD 27.32) after learning, and the difference in the overall crash cart learning outcomes before and after learning was also significant (*P*=.02). The difference between the 2 groups in content learning outcome for the crash cart scores after learning was significant (*P*=.01; Table 2).

For nurses with under 2 years of experience, for the AR group, the average crash cart learning outcomes score increased from 46.54 (SD 16.48) before learning to 68.46 (SD 21.11) after learning, showing a significant improvement (*P*<.001). In the control group, the mean score rose from 43.59 (SD 18.57) to 47.15 (SD 21.96) after learning. Still, there was no significant difference (*P*=.34). A comparison of postlearning crash cart outcomes between the 2 groups of nurses with less than 2 years of experience revealed a significant difference (*P*<.001; Table 2). Among nurses who have more than 2 years of experience, In the AR group, the average crash cart learning outcomes score was 80.59 (SD 14.78) before learning and 85.29 (SD 15.86) after learning, with no significant difference (*P*=.20). In contrast, the control group had a mean score of 76.50 (SD 17.25) before learning and 85.50 (SD 15.35) after learning, showing a

XSL•FO
RenderX

significant improvement (*P*=.001). There was no significant difference when comparing the post learning crash cart outcomes between the 2 groups of nurses with over 2 years of experience (*P*=.32; Table 2).

Motivation, Usability, and Cognitive Load Assessment

For the AR group, the mean scores of the IMMS, SUS, and Cognitive Load Theory Questionnaire were 141.65 (SD 19.25), 90.47 (SD 11.91), and 15.42 (SD 5.76), respectively. Specifically, for the IMMS, the mean scores for the attention, relevance, confidence, and satisfaction subscales were 47.84 (SD 6.94), 35.21(SD 5.39), 35.81 (SD 5.09), and 22.79 (SD 3.20), respectively (Table 3).

Table 3. The mean score distribution in IMMS[a], SUS[b], and CLTQ[c] in the AR[d] group.

	Total	Mean (SD)
IMMS	36-180	141.65 (19.25)
Attention	12-60	47.84 (6.94)
Relevance	9-45	35.21 (5.39)
Confidence	9-45	35.81 (5.09)
Satisfaction	6-30	22.79 (3.20)
SUS	25-100	90.47 (11.91)
CLT	8-40	15.42 (5.76)

[a]IMMS: Instructional Materials Motivation Survey.

[b]SUS: System Usability Scale.

[c]CLTQ: Cognitive Load Theory Questionnaire.

[d]AR: augmented reality.

Discussion

Principal Findings

One of the strengths of this study is that all participants were university graduates, and there were no significant demographic differences in work experience or ACLS certification between the control and intervention groups. This homogeneity minimized potential confounding factors, allowing us to attribute the observed performance improvements in the AR group primarily to the intervention rather than differences in participants' prior knowledge or experience.

Effectiveness of AR-Based Learning for ACLS Training

This study found that all nurses significantly improved after completing the ACLS training course (*P*=.002). However, the crash cart learning system developed using AR technology demonstrated significantly better learning outcomes than traditional teaching methods. Chen and Liou [24] developed a teaching intervention that involved AR technologies. A total of 95 junior nursing students who were learning how to handle airway obstruction during the delivery of ACLS participated in the study. The results revealed that AR technologies improved learning outcomes, confidence, and satisfaction more than traditional teaching methods [24]. Qualitative interviews revealed an emergency airway obstruction training system using AR technology, and focus group interviews were conducted with 82 nursing students. The results showed that AR technology helped them understand real clinical scenarios better and reduced their learning pressure. Unlike traditional teaching methods, which require following the pace of others, AR allowed them to practice independently and repeatedly after the course, based on their individual needs [1]. However, the crash cart learning system developed using AR technology can enhance participants' performance. It can be inferred that because the AR system allows for practicing with virtual objects in real-world scenarios, it overcomes the limitations of traditional teaching methods, which lack the ability to simulate emergency clinical situations. By using AR to simulate emergency clinical scenarios, learners experience reduced stress when faced with real-life emergencies [25]. Additionally, this study uses AR marker technology on mobile devices, allowing learners to access the crash cart learning system by simply scanning with their phones. This eliminates the need for specific devices or designated spaces for practice, thereby increasing the accessibility of learning.

Impact on Nurses With Different Levels of Experience

This study found that nurses with work experience of fewer than 2 years showed significantly improved learning outcomes (*P*<.001) when using AR technology for virtual simulation learning, particularly in crash cart learning. Traditional teaching methods, even using pictures, descriptions, or videos, made it difficult for less experienced nurses to visualize the scenarios. Interestingly, the study also found that for more experienced nurses, there was no significant difference in learning outcomes before and after using the AR-based crash cart learning system (*P*=.20). In contrast, traditional teaching methods yielded significant differences in both ACLS (*P*<.001) and crash cart learning (*P*=.001) for work experience of more than 2 years. Balian et al [8] developed a CPR training system that integrated AR technologies. Their study involved 51 hospital employees. Nurses with an average of 5 years of experience achieved more favorable learning outcomes in complicated and changing ACLS scenarios than participants with less experience. In addition to another study involving 30 cardiopulmonary cerebral

XSL•FO
RenderX

resuscitation instructors, it was found that learning through AR techniques was more appealing to inexperienced participants, helping them stay focused during their learning. AR also mitigated the adverse effects of their lack of experience, which often hindered their learning performance, by providing virtual objects overlaid onto real-world environments. This feature reduced inexperienced participants' challenges, allowing them to grasp concepts better and improve their learning effectiveness [26]. On the other hand, more experienced learners were not as accustomed to using mobile devices for learning. Instead, those with extensive work experience preferred traditional teaching methods, favoring interaction and discussion with instructors to analyze complex clinical scenarios [26]. This prompts us to reflect that learning through AR technology cannot fully address all participants' needs. In addition to continuously optimizing the variability of simulated scenarios, the learning process should incorporate discussions with instructors to cater to the diverse needs of participants.

Enhancing Motivation and Reducing Cognitive Load Through AR

The AR technologies in this study did not burden the nurses in terms of system operation and cognitive load. The AR technologies stimulated and maintained learning motivation through mobile devices that were readily accessible to the nurses. A systematic review of students' attitudes, effectiveness, and motivation in using AR for learning analyzed 28 studies from 2016 to 2023. The results showed that students had more positive learning attitudes when using AR ($P<.001$), and their learning outcomes were better compared to students who did not use AR technology ($P<.01$). However, AR-assisted learning did not improve learning motivation ($P=.12$). The study noted that AR-assisted learning systems are often disrupted by issues such as learning time, internet connectivity, and device problems, which can interrupt the learning process [1].

Regarding system usability, the average score in this study was 90.47 (SD 11.91), indicating that the crash cart learning system is highly usable. According to Choi et al [18], to better quantify the score range, the scale was divided into 5 levels: a score of 90-100 is an A, indicating extremely useful; 80-89 is a B, indicating very useful; 70-79 is a C, indicating good; 60-69 is a D, indicating acceptable; and 0-59 is an F, indicating poor. Based on this study's score, the system falls into the A level, indicating that participants found the crash cart learning system extremely easy to use.

Limitation

This study's limitation is that the teaching course based on the developed crash cart learning system focused on the functions and operation of the crash cart without providing opportunities for nurses to engage in emergency clinical scenarios. This lack of integration with actual ACLS scenarios limits the overall learning experience in first aid training. Future system updates could address this by adding emergency scenarios to enhance the connection between the system and practical emergency response training.

Conclusions

This study was conducted in a medical center where all nurses must be university graduates and have completed ACLS certifications during their employment. These stringent requirements may have contributed to an artificially higher learning effectiveness, as the participants may have had a more robust foundational knowledge and prior experience. As a result, the findings may not be fully generalizable to settings where nurses have lower requirements.

Acknowledgments

We appreciate all participants for their generous involvement. We also thank the National Cheng Kung University Hospital (grant NCKUH-11008019) and Taiwan's Ministry of Science and Technology (grant MOST-109-2511-H-165-002) for their support.

Data Availability

The data sets generated during this study are not publicly available to protect student privacy but are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.2).
[PDF File (Adobe PDF File), 83 KB - jmir_v26i1e57327_app1.pdf]

References

1. Chen PJ, Liou WK. Development and application of AR-based assessment system for infant airway obstruction first aid training. Children (Basel) 2022;9(11):1622 [FREE Full text] [doi: 10.3390/children9111622] [Medline: 36360350]
2. Andersen LW, Holmberg MJ, Berg KM, Donnino MW, Granfeldt A. In-hospital cardiac arrest: a review. JAMA 2019;321(12):1200-1210 [FREE Full text] [doi: 10.1001/jama.2019.1696] [Medline: 30912843]
3. Penketh J, Nolan JP. In-hospital cardiac arrest: the state of the art. Crit Care 2022;26(1):376 [FREE Full text] [doi: 10.1186/s13054-022-04247-y] [Medline: 36474215]

XSL•FO
RenderX

4. Taiwan patient-safety reporting system annual report 2020. Joint Commision of Taiwan. URL: https://www.jct.org.tw/cp-1352-8371-a819a-2.html [accessed 2024-11-02]
5. Chen Y, Hsu S, Liu H. Combination of online teaching and high-fidelity simulation team training to improve the first aid effectiveness of nurses in a medical center ward. Natl Taiwan Univ Hosp J Nurs 2022;18(1):104-118. [doi: 10.6740/ntuhjn.202201_18(1).0010]
6. Lee SN, Ho CL, Lin YJ, Wu LM, Wang CH. Using situational simulations to raise the first aid competency rate of hemodialysis nurses. J Taiwan Nephrol Nurses Assoc 2021;19(2):49-65. [doi: 10.3966/172674042020121902004]
7. Chen M, Tzeng W. Association of knowledge on resuscitation medications and the stress perception in resuscitation event in newly graduated nurses. Yuan-Yuan Nurs 2017;11(1):25-34. [doi: 10.6530/yyn/2017.1.1]
8. Balian S, McGovern SK, Abella BS, Blewer AL, Leary M. Feasibility of an augmented reality cardiopulmonary resuscitation training system for health care providers. Heliyon 2019;5(8):e02205 [FREE Full text] [doi: 10.1016/j.heliyon.2019.e02205] [Medline: 31406943]
9. Keller JM. Development and use of the ARCS model of instructional design. J Instr Dev 1987;10(3):2-10 [FREE Full text] [doi: 10.1007/bf02905780]
10. Keller JM. Strategies for stimulating the motivation to learn. Perform Instr 1987;26(8):1-7 [FREE Full text] [doi: 10.1002/pfi.4160260802]
11. Ryan GV, Callaghan S, Rafferty A, Higgins MF, Mangina E, McAuliffe F. Learning outcomes of immersive technologies in health care student education: systematic review of the literature. J Med Internet Res 2022;24(2):e30082 [FREE Full text] [doi: 10.2196/30082] [Medline: 35103607]
12. IndustryARC. 2023. URL: https://www.industryarc.com/Report/78/augmented-reality-and-virtual-reality.html [accessed 2024-11-02]
13. Hsieh MC, Lin YH. VR and AR applications in medical practice and education. Hu Li Za Zhi 2017;64(6):12-18. [doi: 10.6224/JN.000078] [Medline: 29164542]
14. Mendez KJW, Piasecki RJ, Hudson K, Renda S, Mollenkopf N, Nettles BS, et al. Virtual and augmented reality: implications for the future of nursing education. Nurse Educ Today 2020;93:104531 [FREE Full text] [doi: 10.1016/j.nedt.2020.104531] [Medline: 32711132]
15. Oun A, Hagerdorn N, Scheideger C, Cheng X. Mobile devices or head-mounted displays: a comparative review and analysis of augmented reality in healthcare. IEEE Access 2024;12:21825-21839. [doi: 10.1109/access.2024.3361833]
16. Polit DF, Sherman RE. Statistical power in nursing research. Nurs Res 1990;39(6):365-369. [Medline: 2092311]
17. Chiou H. Quantitative Research Statistical Analysis: Analysis of SPSSR Data Analysis Examples. Taiwan: Wu-Nan Book Inc; 2019.
18. Choi J, Baker E, Nalawade S, Lee H. Steps to develop a mobile app for pain assessment of cancer patients: a usability study. Comput Inform Nurs 2020;38(2):80-87. [doi: 10.1097/CIN.0000000000000581] [Medline: 31850936]
19. Aaron B, Miller J, Kortum P. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123 [FREE Full text]
20. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum-Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
21. Sweller J, Van Merrienboer JJG, Paas FG. Cognitive architecture and instructional design. Educ Psychol Rev 1998;10(3):251-296. [doi: 10.1023/a:1022193728205]
22. Krieglstein F, Beege M, Rey GD, Sanchez-Stockhammer C, Schneider S. Development and validation of a theory-based questionnaire to measure different types of cognitive load. Educ Psychol Rev 2023;35(1):9. [doi: 10.1007/s10648-023-09738-0]
23. Hsiao P, Lin C, Han C, Chen L, Wang L, Su C. Role transition of newly graduated nurses: a qualitative study. Contemp Nurse 2021;57(6):450-461. [doi: 10.1080/10376178.2022.2029519] [Medline: 35021962]
24. Chen PJ, Liou WK. The effects of an augmented reality application developed for paediatric first aid training on the knowledge and skill levels of nursing students: an experimental controlled study. Nurse Educ Today 2023;120:105629. [doi: 10.1016/j.nedt.2022.105629] [Medline: 36375385]
25. Mendez KJW, Piasecki RJ, Hudson K, Renda S, Mollenkopf N, Nettles BS, et al. Virtual and augmented reality: implications for the future of nursing education. Nurse Educ Today 2020;93:104531 [FREE Full text] [doi: 10.1016/j.nedt.2020.104531] [Medline: 32711132]
26. Wong MAME, Chue S, Jong M, Benny HWK, Zary N. Clinical instructors' perceptions of virtual reality in health professionals' cardiopulmonary resuscitation education. SAGE Open Med 2018;6:2050312118799602 [FREE Full text] [doi: 10.1177/2050312118799602] [Medline: 30245815]

Abbreviations

ACLS: advanced cardiac life support
AR: augmented reality
CLT: cognitive load theory

XSL•FO
RenderX

CPR: cardiopulmonary resuscitation
CVI: content validity index
HMD: head-mounted display
IHCA: in-hospital cardiac arrest
IMMS: Instructional Materials Motivation Survey
SUS: System Usability Scale
VR: virtual reality

Edited by A Coristine; submitted 12.02.24; peer-reviewed by A Hassan, R Bidkar, HF Hsieh, X Cheng; comments to author 17.09.24; revised version received 22.09.24; accepted 20.10.24; published 05.12.24.

Please cite as:
Sun WN, Hsieh MC, Wang WF
Nurses' Knowledge and Skills After Use of an Augmented Reality App for Advanced Cardiac Life Support Training: Randomized Controlled Trial
J Med Internet Res 2024;26:e57327
URL: https://www.jmir.org/2024/1/e57327
doi:10.2196/57327
PMID:

©Wan-Na Sun, Min-Chai Hsieh, Wei-Fang Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining the Use of Virtual Reality to Support Mindfulness Skills Practice in Mood and Anxiety Disorders: Mixed Methods Study

Rebecca Blackmore[1*], PhD; Claudia Giles[1*], MPsych; Hailey Tremain[1], PhD; Ryan Kelly[2], PhD; Fiona Foley[1], BSc; Kathryn Fletcher[1], PhD; Maja Nedeljkovic[1], PhD; Greg Wadley[3], PhD; Elizabeth Seabrook[1*], PhD; Neil Thomas[1*], PhD

[1]Centre for Mental Health and Brain Sciences, Swinburne University of Technology, Hawthorn, VIC, Australia

[2]School of Computing Technologies, RMIT University, Melbourne, Australia

[3]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

[*]these authors contributed equally

Corresponding Author:
Neil Thomas, PhD
Centre for Mental Health and Brain Sciences
Swinburne University of Technology
John Street
Hawthorn, VIC, 3122
Australia
Phone: 61 392148444
Email: neilthomas@swin.edu.au

Abstract

Background: Virtual reality (VR) has been proposed as a technology to support mindfulness practice through promoting increased engagement and presence. The proposed benefits of this technology have been largely unexamined with clinical populations. Further research is required to understand its clinical potential and utility in improving and managing mental health symptoms.

Objective: This study aims to investigate the proximal impacts of a single, brief, VR-supported mindfulness practice for individuals with a mood or anxiety disorder and to understand user experiences, which may affect the acceptability and efficacy of VR mindfulness for this population.

Methods: This mixed methods study recruited 28 participants with a primary diagnosis of major depressive disorder, bipolar disorder, or anxiety disorder. Participants completed a mindfulness practice wearing a VR headset that was presenting an omnidirectional video of a forest scene, which was overlaid with a guided audio voiceover. Before and after the practice, measures were completed assessing state mindfulness (Toronto Mindfulness Scale), affect (Positive and Negative Affect Schedule), and anxiety (State-Trait Anxiety Inventory Y-1; n=27). Semistructured interviews were then held inquiring about the user experience and were analyzed using thematic analysis (n=24).

Results: After completing the VR-supported mindfulness practice, both measures of state mindfulness on the Toronto Mindfulness Scale, mean curiosity and decentering, increased significantly (Cohen d=1.3 and 1.51, respectively; P<.001). Negative affect on the Positive and Negative Affect Schedule (Cohen d=0.62; P=.003) and State-Trait Anxiety Inventory Y-1 state anxiety (Cohen d=0.84; P<.001) significantly reduced. There was no significant change in positive affect (Cohen d=0.29; P=.08). Qualitative analysis of interviews identified 14 themes across 5 primary theme categories. The results suggested that being mindful during the use of the app was experienced as relatively effortless because of the visual and immersive elements. It was also experienced as convenient and safe, including when compared with prior traditional experiences of mindfulness. Participants also identified the uses for VR-supported mindfulness in managing emotions and symptoms of mental illness.

Conclusions: The results provide preliminary evidence that VR-supported mindfulness can improve emotional states and manage mental health symptoms for those with mood or anxiety disorders. It offers some potential clinical applications for those with mood or anxiety disorders for exploration within future research.

(J Med Internet Res 2024;26:e45640) doi:10.2196/45640

XSL•FO
RenderX

KEYWORDS

virtual reality; mindfulness; mood disorders; anxiety disorders; depression; bipolar disorder

Introduction

Background

Mindfulness has been described as a purposeful quality of attention toward the present moment involving an attitude of nonjudgment and curiosity [1,2]. Mindfulness-based approaches have increasingly been used as psychological interventions, with demonstrated benefits for a range of mental health outcomes, including depression, anxiety, and stress [3]. Several mechanisms have been proposed to explain how mindfulness skills lead to symptom reduction and behavior change. Although conceptualizations vary, mindfulness practice has been posited to affect change via modifying emotion regulation (reappraisal of emotions, exposure, and extinction), attention regulation, body awareness, and perspective of the self [4,5].

Mindfulness-based approaches are promising interventions for clinical populations. Mindfulness-based approaches can reduce depressive and anxiety symptoms both in primary anxiety and depressive disorders [6] and other mental health conditions, such as bipolar disorder (BD) [7,8]. The mechanisms through which mindfulness operates may be of particular relevance for clinical populations. For example, a qualitative metasynthesis of clinical studies [9] found that following mindfulness, participants reported observing and responding to early warning signs more quickly, acknowledging and accepting their extreme emotional states, which may prevent spiraling: observing fewer negative impacts of episodes (such as shame) and improved perspectives of themselves (eg, as distinct from their symptoms and challenges).

Psychological mindfulness-based interventions (MBIs), such as Mindfulness-Based Stress Reduction [2] and Mindfulness-Based Cognitive Therapy [10], primarily focus on learning mindfulness via regular practice in meditation. However, some studies suggest that learning mindfulness via traditional meditations may be demanding and aversive, and this may be particularly salient for clinical populations [11,12]. A metasynthesis of qualitative studies of group MBIs in clinical populations identified challenges, including being confronted with one's difficulties with fresh awareness, and, in some cases, experiencing more distress than before commencing, and practical concerns, such as finding the time and physical space to practice and difficulty grasping the core concepts of mindfulness [9]. Difficulties with maintaining stillness and concentration have also been reported, especially in individuals with BD [13].

Although traditionally being learned via formal meditation practices, mindfulness may be cultivated using any number of techniques and short exercises that support attention orientation to the present moment and acceptance [1,14]. This more flexible definition of mindfulness coincides with an increasing use of brief MBIs as alternative ways of developing mindfulness skills [15]. Brief MBIs, including one-off and stand-alone practices of as little as 5 minutes, have demonstrated benefits for outcomes, including depression, anxiety, negative affect and mindfulness, as well as altered neural activity [4,15,16]. Briefer, more accessible mindfulness practices may therefore provide the benefits of mindfulness while addressing challenges associated with other practices.

Digital technologies, such as virtual reality (VR) apps, can support mindfulness practice. Interactive and immersive technologies, such as VR apps, may support components of mindfulness, such as emotion regulation and attention [17]. VR technology typically encompasses a 360° head-mounted display, with hand controllers or natural gestures that enable the user to interact with a computer-generated virtual environment (VE). VR apps can be created using either computer-generated imagery [18] or omnidirectional video footage [19].

The features of VR may provide accessible mindfulness practices, which address some of the barriers associated with mindfulness. VEs provide the user with the sensation of being located within the environment, contributing to the psychological sense of presence or "being there" [20]. VEs are highly engaging, immersive, and interactive, which may facilitate concentration during mindfulness practice [21,22]. Furthermore, access to natural environments during mindfulness practice has been found to offset the attentional effort of developing mindfulness skills [23] and increase adherence [24]. These potential benefits may be of particular relevance for individuals with anxiety and mood disorders, for whom distractibility and difficulties sustaining attention and concentration are hallmarks [25]. Participants are free to choose their own anchors and interaction with the environment, allowing a personalized experience [21] and potentially minimizing the aversion associated with some prior practices. Indeed, compared with non–VR mindfulness, reported benefits have included improved adherence [26,27] and mindfulness [27].

Several studies have piloted the use of VR for mindfulness skills practice in nonclinical populations, showing promising findings with regard to improving state mindfulness and positive affect as well as general acceptability of the apps [21,22,28]. For example, a mixed methods study by Seabrook et al [21] studied the utility of VR for supporting mindfulness practice using omnidirectional video of a natural forest environment, which was overlaid with a guided audio voiceover. Nonclinical participants reported feelings of peace, a sense of presence in the VE, having present moment awareness, and being able to use the natural VE to direct or control attention.

Several pilot studies have investigated VR for mindfulness skills practice for mental health conditions, such as depression and anxiety disorders [26,29,30]. Navarro-Haro et al [26] found that MBI groups with and without VR mindfulness skills training showed similar improvements in symptoms of anxiety, depression, and emotion regulation, but the VR group was more adherent to the program. Tarrant et al [30] compared a brief, nature-based mindfulness experience to a resting control group with individuals with self-reported anxiety symptoms. The results showed that both the VR experience and the quiet rest

XSL•FO
RenderX

control condition reduced anxiety. However, the VR experience was very brief, 5 minutes in total, and no measure of state mindfulness was administered; therefore, it was unclear how either of the conditions contributed to mindful attention. Further research is required to understand potential clinical applications of VR technology and whether it has the capability to improve the ease of mindfulness practice, support mindful attention, improve affect, or reduce symptoms of anxiety among those with mental health conditions [31]. It is also important to explore the suitability and safety of this technology, particularly in its application to clinical groups and individuals experiencing mental health disorders. The use of novel immersive technology for mental health interventions may carry specific risks that are yet not well understood. Thus, the safety of using VR for mindfulness practice with mental health disorders should be evaluated through pilot studies before long-term use in large trials.

This Study

Using a mixed methods design, this study aimed to investigate the proximal effects of a single-session, VR-supported mindfulness practice [21] on state mindfulness, affect, and state anxiety with a sample of individuals with mood or anxiety disorders. This study also explored the potential uses and clinical suitability of VR-supported mindfulness as a mental health self-management tool. Qualitative assessment explored the participants' experience of completing the mindfulness practice, comparison with their previous experiences of mindfulness practice, and the potential uses participants envisaged for VR-supported mindfulness as a tool in supporting their mental health self-management.

Methods

Ethical Considerations

All procedures received ethics approval from the Swinburne University of Technology Human Research Ethics Committee (SHR 2018/256) and the University of Melbourne Human Research Ethics Committee (ID# 1852613.2). All participants gave full informed consent, data were deidentified, and participants received AUD $80 (USD $50) reimbursement for their participation.

Participants

Participants were recruited via promotion on Facebook and emails to mental health websites and services. Interested participants completed a web-based screening survey assessing their eligibility. Inclusion criteria were aged 18 to 65 years; fluent in the English language; reported normal or corrected-to-normal vision and hearing; diagnosis of BD, major depressive disorder (MDD), or anxiety disorder; and willingness to engage in a mindfulness session. Participants were excluded if they reported a diagnosis of photosensitive epilepsy; had previous experience of severe cybersickness; or were currently experiencing a mood episode, active suicidality, or psychosis. Eligibility was confirmed by telephone using the Mini International Neuropsychiatric Interview [32] and the Columbia-Suicide Severity Rating Scale [33]. Of 55 participants assessed for eligibility, 34 met the criteria, and a final sample of 28 took part (4 did not respond to further contact and 2 were canceled because of the COVID-19 pandemic), of which 24 additionally took part in the qualitative interviews.

Measures

Baseline Variables

To characterize the sample, a demographic questionnaire asked participants to identify information related to their sex, age, education level, history of mental illness, mental health treatment, and use of prescribed medications. Participants were also asked about their previous experience with mindfulness and confidence and experience in using VR technology.

Five Facet Mindfulness Questionnaire Short Form

The Five Facet Mindfulness Questionnaire Short Form (FFMQ-SF) [34] total score was used to assess baseline trait mindfulness. The FFMQ-SF asks respondents to rate statements, indicating their general tendency to be mindful in daily life. Statements are rated on a response scale from 1=never or very rarely true to 5=very often or always true. The measure shows good internal consistency and convergent validity with the full version of the questionnaire [35]. In the current sample, Cronbach α scores ranged between 0.62 and 0.83 for the 5 scales, which indicated adequate reliability.

Pre-Post Intervention Measures

Toronto Mindfulness Scale

The Toronto Mindfulness Scale (TMS) consists of 13 items asking participants to consider the degree to which they were mindful in the previous 15 minutes [36]. The measure consists of 2 distinct subscales, curiosity (6 items) and decentering (7 items). Response options are rated on a 5-point Likert scale ranging from 1 (not at all) to 5 (very much). Higher scores indicate greater association with each construct. A strong internal consistency of 0.88 for curiosity and 0.84 for decentering has been previously reported [36]. This study found an acceptable Cronbach α of 0.88 and 0.90 for curiosity before and after viewing the VR-supported mindfulness practice, respectively. The results for the decentering scale were acceptable (α=0.68 before and α=0.72 after the practice).

Positive and Negative Affect Schedule

The Positive and Negative Affect Schedule (PANAS) [37] was used to measure state positive and negative affect. This is a self-report measure comprising two 10-item scales measuring positive and negative affect. Items are rated from 1 (*very slightly or not at all*) to 5 (*extremely*), producing positive and negative affect scores ranging from 10 to 50. Cronbach α for the positive affect scale was 0.89 (before) and 0.90 (after), and the negative affects scale was 0.88 (before) and 0.72 (after).

State-Trait Anxiety Inventory Y-1

The State-Trait Anxiety Inventory Y-1 [38] was used to measure state anxiety**.** This self-report measure of trait and state anxiety includes 20 state anxiety items that were used in this study. Response options are rated on a 4-point scale. Higher scores indicate higher self-perceived state anxiety. Normative adult sample data have shown strong internal consistency coefficients ranging from 0.86 to 0.95 [38,39]. In this study, the resultant

XSL•FO
RenderX

Cronbach α of 0.93 both before and after the VR-supported mindfulness practice indicated good internal consistency.

Semistructured Interviews

A semistructured interview guide was developed for this study. Questions were arranged into topics, including feedback and comments on the VR experience, the use of mindfulness in VR for participants both with and without prior mindfulness experience, potential impact of mental health symptoms on use of the program, and potential uses for the app. A copy of the interview guide is available from the authors on request.

VR Mindfulness App

The app, called Place, was used to deliver a 15-minute program of guided mindfulness within a VE [21,40]. The app was developed through a user-centered design process and has been evaluated with a general population sample [19]. Crucially, the app has been shown to create positive changes in mindful awareness as well as being safe and not producing any related simulator sickness. The VE consisted of 2 forest scenes created from omnidirectional footage: first, a clearing next to a river and, second, at the edge of a river with a waterfall in view. The audio guidance used was tailored to the VE and invited the user to attend to physical sensations, the breath, different parts of the VE, and prompting the user to redirect their attention if they notice their mind wandering. The voiceover also allowed space for unguided practice for up to 70 seconds at a time. A recent article articulated some of the challenges in designing and evaluating digital interventions for mental health, providing a framework for classifying such interventions [41]. Within this framework, the Place app was designed for *experiential* (as opposed to *didactic*) learning and as an *on-the-spot* (as opposed to *offline training*) tool; that is, it is intended that users practice mindfulness while using the app and that, ultimately, the app is used when and as needed within users' real contexts. Although there is evidence that even brief mindfulness practices may lead to altered neural activity and longer-term or "downstream" improvements [4,15,16,42] (within this framework classified as *offline training*), this is not the focus of the present investigation.

Procedure

After completing a telephone assessment, participants attended an in-person session in research rooms in a university setting. Participants completed the questionnaires electronically via the survey software Qualtrics, followed by the 15-minute VR mindfulness practice completed with an Oculus Go headset, with 2 additional minutes of instructions and safety information provided verbally by research staff. Participants were seated in a soundproof booth with headphones and a swivel chair, which allowed for exploration of the 360° environment. Participants then completed quantitative postexercise measures, followed by the optional qualitative semistructured interviews. Interviews were audio recorded and transcribed by NVivo. Transcripts were checked, and any errors were corrected. Interviews were completed by 4 postgraduate-trained research assistants, who received training from the investigators, which included practice interviews.

Data Analysis

Descriptive statistics were processed and analyzed in IBM SPSS Statistics (version 26.0). Normality and skewness checks were conducted, and Levene test for equality of variances showed no violations of assumptions. Paired sample *t* tests (2-tailed) were conducted to compare pre- and posttreatment scores across measures as the primary analysis, with repeated measures ANOVAs with a Cronbach α level of 0.05 conducted to investigate differences between clinical groups across measures and timepoints.

The qualitative data were analyzed using an inductive approach to thematic analysis [43]. Transcripts were open coded to attribute nominal codes. A secondary coder was allocated 3 transcripts at random to code, and 1 theme was altered as a result. As a participant checking procedure, all participants who took part in the interviews were emailed the initial themes and invited to provide feedback: 4 participants opted to provide feedback, all of whom endorsed the themes.

Results

Overview

Of the 28 participants recruited for the study, 1 was excluded because of an IT issue impacting their time 2 data. As a result, 27 participants were included in the final analysis: 17 female participants and 10 male participants. The participants had a mean age of 40 (SD 11.15; range 20-63) years. All participants endorsed prior experience of counseling or psychotherapy, with 15 (56%) participants receiving therapy in the past month. In total, 20 (74%) participants indicated they were currently taking medication as part of their current treatment. Table 1 presents the demographic characteristics of the sample (n=27).

Primary diagnoses included MDD (n=9), BD (n=10), and anxiety disorders (n=8), which included generalized anxiety, social anxiety, and panic disorder. A total of 14 participants had secondary diagnoses: 8 (57%) of which were anxiety disorders and 6 (43%) were MDD. We noted no statistically significant baseline differences between clinical groups (MDD, BD, and anxiety) on trait mindfulness (FFMQ-SF total score: $F_{2,24}$=0.25, P=.78) or any of the measures used in the pre-post analyses (TMS curiosity: $F_{2,24}$=0.48, P=.63; TMS decentering $F_{2,24}$=0.30, P=.74; PANAS positive affect $F_{2,24}$=0.20, P=.82; PANAS negative affect $F_{2,24}$=0.39, P=.68; State-Trait Anxiety Inventory Y-1: $F_{2,24}$=0.15, P=.86).

Most participants (24/27, 89%) had experience of mindfulness practice, although 3 (11%) reported not having tried it before but knew of it. Most participants (20/27, 74%) reported feeling very to extremely confident in using technology, and 63% (17/27) of participants had previous experience with VR. The mean total score of the FFMQ-SF was 44.3 (SD 7.6).

XSL•FO
RenderX

Table 1. Demographic characteristics of the sample (n=27).

Sample characteristics	Values, n (%)
Education	
Completed secondary school	1 (4)
Some tertiary education	7 (26)
Completed tertiary education	8 (30)
Some postgraduate education	3 (11)
Completed postgraduate education	8 (30)
Relationship status	
Single	9 (33)
Partnered	4 (15)
De facto	3 (11)
Married	6 (22)
Separated	2 (7)
Divorced	3 (11)
Primary diagnosis	
Bipolar disorder	10 (37)
Major depressive disorder	9 (33)
Generalized anxiety disorder	5 (19)
Social anxiety disorder	2 (7)
Panic disorder	1 (4)
Previous mindfulness experience	
None, but have heard of it	3 (11)
Tried it once	3 (11)
Tried it a few times (2-5)	13 (48)
Do it monthly	1 (4)
Do it weekly	4 (15)
Do it daily	3 (11)
Previous virtual reality experience	
None, but have heard of it	10 (37)
Tried it once	8 (30)
Tried it a few times (2-5)	8 (30)
Do it monthly	1 (4)
Confidence with technology	
Slightly	1 (4)
Moderately	6 (22)
Very	13 (48)
Extremely	7 (26)

Pre- and Postintervention Measures

The mean effort participants rated it took to be mindful in the program was 37.6 out of 100 (SD 26.7), with 70% (19/27) of participants rating this less than 50. Table 2 presents changes in measures from before to after experiencing the VR-supported mindfulness practice. Scores on both subscales of the TMS increased significantly (P<.001), with large effect sizes. Both state negative affect and state anxiety reduced, also with large effect sizes. There was no change in state positive affect. Repeated measures ANOVAs indicated no main effect of diagnostic group (MDD, BD, and anxiety) or time×group interaction, and similarly, no interactions were observed for recent receipt of therapy or whether taking medication.

XSL•FO
RenderX

Table 2. Mean scores on pre- and postintervention measures (n=27).

Measures	Preintervention scores, mean (SD)	Postintervention scores, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
TMS[a] curiosity	15.81 (5.76)	22.03 (5.04)	6.93 (26)	<.001	1.30
TMS decentering	16.93 (4.53)	24.00 (4.48)	8.07 (26)	<.001	1.51
PANAS[b] positive affect	28.26 (6.65)	30.30 (7.03)	1.83 (26)	.08	0.29
PANAS negative affect	14.90 (4.83)	11.56 (2.04)	3.33 (26)	.003	0.62
State-Trait Anxiety Scale Short Form state anxiety	41.96 (9.70)	32.41 (8.42)	4.49 (26)	<.001	0.83

[a]TMS: Trait Mindfulness Scale.

[b]PANAS: Positive and Negative Affect Scale.

Semistructured Interviews

Overview

In total, 24 of the participants agreed to participate in a semistructured interview to gain a better understanding of the user experience of the VR-supported mindfulness practice with a clinical population. A total of 14 themes were developed across 5 primary theme categories: ease of present moment awareness, multilayered engagement in mindfulness, comparisons to practicing mindfulness outdoors or in public settings, helping develop mindfulness practice, and supporting and interacting with mental health (Table 3).

Within many of the identified themes, participants reflected on their experiences with the Place app in comparison to their previous mindfulness experiences, which may illuminate the *unique* benefits and challenges of VR mindfulness. Where such comparative statements contributed the theme (in contrast to highlighting characteristics or experiences of the intervention more broadly), these have been identified within their relative themes.

XSL•FO
RenderX

Table 3. Qualitative results.

Theme category and themes	Illustrative quotes
Ease of present moment awareness	
Reduced external distractions	"I liked that. It was focused so that I wasn't distracted by surrounding noises because I didn't know what was happening in the real world or real environment." [P24]
Positive emotions are expected and experienced	"...my meditation videos...has to be about an hour that I feel like I'm like, relax...but this one [the app] like in 15 minutes, I was very calm" [P9]
Embedding parts of regular mindfulness practice	"...when I was watching the water, I was trying it out and saying if you watch the river move, then that's kind of the same thing as radical acceptance because you can only accept what's going on in the second that it's happening." [P9]
A multilayered engagement in mindfulness	
Being able to see things adds to feeling involved	"...I've done a lot of auditory [mindfulness practice]...I was unaware how much I actually really need both things [auditory and visual] to be stimulated for actually get me really in a calm, mindful place" [P9]
Not focusing on internal elements was both helpful and confusing	"I get frustrated with mindfulness things where it's all about breathing...I just get bored with that. So I liked having something else to focus on..." [P16]
Comparisons to practicing mindfulness outdoors or in public settings	
A convenient way to access mindfulness in nature, indoors	"...regardless of...your mood or your anxiety...just chuck it on...and you can immediately start work on the mindfulness side of things without the 10 hour round trip to do it." [P9]
A sense of safety in the virtual environment	"just more awareness of what's around you that's safe and guided. Because often I try to practice mindfulness outside in different environments that I may not have control over...but making sure that this has been overseen and this is a practice thing. Just made me feel a lot more safe" [P1]
Helping to develop mindfulness practice	
Easing into different forms of mindfulness practice	"[the app] helps to engage you in the moment, which I think is probably one of the hardest things about mindfulness. And once you've gotten to that point, then you can take that next step to reflecting and exploring." [P7]
Encouragement to start mindfulness again	"I think that I'd get the benefits of it, which would mean that I'd probably do it more often..." [P24]
Supporting mental health	
Use for increasing present moment awareness	"...when I'm depressed, I feel like I'm just dissociating and losing time. So doing that [the app] sort of brings me back to the present mind." [P17]
Changing emotions and getting away	"If there was some kind of a trigger that had happened, if I'd had an argument, whatever the trigger was, if I was feeling that way, I would probably gravitate towards it [the app] in order to feel calm, and then be able to investigate and associate the emotions and feelings that I had..." [P23]
Disorder-specific uses	"I think it could be a way of again, kind of stepping back...stepping down some of the intensity of feeling [when mood is elevated] and just some moments of calm in the rush..." [P16]
Use for prevention	"...depends if you notice you're in an elevated mood state because often you don't or don't until it's too late. So if you catch it in time really early, then yes, I think this would be helpful..." [P9]
Barriers to using virtual reality during a mental health episode	"...one of the associations with depression is you don't feel like doing anything so...I: So that might be something that would get in the way..." [P2]

Theme Category 1: Ease of Present Moment Awareness

Participants reported that being mindful, or having present moment awareness, while using the app was relatively easy.

Reduced External Distractions

External distractions in previous mindfulness practice were noted, in contrast to a sense of reduced distractions while using the app:

> *You need that personal, alone [time], which is really hard to get in life...It's a bit hard to lock yourself in a room and have no noise...whereas this [the app] just allows that.* [P24]

Within this theme, participants drew comparisons between experiences with the app and previous mindfulness experiences. The headset was a contributor to accessing this space:

> *You put the headset on...And it makes it easier. The extra sensory stimulation lends itself to the sense of being in a "bubble."* [P24]
>
> *I feel like [the app is] a more all-encompassing sensory experience...And I felt like if you do a guided meditation at home, you're very aware you're at home in your lounge room. But this was...I feel like I was in a bubble...It really helped a lot.*

However, one participant found the images in the VE distracting in comparison to previous mindfulness experiences.

XSL•FO
RenderX

> *Like it'd take an effort for me, a bit more than usual, to do the mindfulness exercise [using the app] because of the fact that I was distracted by the images.* [P15]

Positive Emotions Are Expected and Experienced

Positive emotional experiences were accessed sooner using the app, such as feeling peaceful:

> *...I didn't achieve that sense of peace until I'd been practicing an hour a day, or more calm and relaxed.* [P20]

Participants also drew comparisons with prior experiences of mindfulness within this theme. For participant 20, practicing mindfulness with VR contrasted to their prior experiences of mindfulness in a Mindfulness-Based Stress Reduction course, where there had been a period of emotional discomfort:

> *...[the app] kind of skipped over the part where you're sitting there going ok I'm supposed to just be focusing on my breath... [and that causes] it's not even anxiety, but aggravation?*

Embedding Parts of Regular Mindfulness Practice

Participants integrated their existing mindfulness practice to enhance the experience of the app:

> *Looking at the trees and imagining their roots growing down into the soil and thinking of that as a grounding kind of exercise...I thought that was a bit you could run with and think about for an extended period of time.* [P16]

Theme Category 2: A Multilayered Engagement in Mindfulness

A perceived importance of having multiple sources of sensory input in practicing mindfulness was further enhanced by the VR app.

Being Able to See Things Adds to Feeling Involved

Again within this theme, participants contrasted their experience with other mindfulness practices. For example, the visual element of the app was important in providing a level of stimulation and engagement compared to previous audio-only mindfulness practices:

> *I think it really draws you in because it's visual and it is 360°...so it lands you quicker in the experience rather than if you're just imagining it in your head.* [P19]

For those participants who had previously practiced mindfulness involving guided imagery, such as imagining leaves floating down a stream, the app provided the image so that there was less attentional fatigue:

> *I can visualise things really well, but it requires quite a lot of effort...So with this [the app], having the visual there was quite relaxing.* [P18]

Participants identified that they may gravitate toward different mindfulness practices when they are in different states. The visual aspect of the app may be more useful for calming down rather than a more introspective mindfulness practice:

> *they feel different [the app and audio guided mindfulness] and I think if I was wanting to feel more calm, I would automatically go for the VR version. If I was in a place where I was willing to be more reflective, I personally would find it easier to do it with audio only.* [P7]

Not Focusing on Internal Elements Was Both Helpful and Confusing

This theme also involved comparisons to other mindfulness modalities. For some participants, the focus on the VE resulted in a reduced focus on internal elements such as the breath, emotions, and thoughts. This was perceived as helpful in staying in the present moment. Participant 7 found the visual focus improved present moment awareness:

> *I didn't have sort of that enquiring or curiosity...about the thoughts that I was having [in the app]...I have done other audio mindfulness in the past. I think the difference being that when you have something visual to take up your attention, it's quite in-depth and you've got a lot of focus on...it kept me present better than when I've done it with just audio.*

Other participants indicated that focusing on external stimuli created confusion in whether their focus should be on the self, the environment, or both:

> *when he was talking about the breathing...I was closing my eyes and then I thought oh but if I'm closing my eyes then I can't see the forest. So that was a bit confusing as to what I should do.* [P22]

Theme Category 3: Comparisons to Practicing Mindfulness Outdoors or in Public Settings

The app was viewed as a convenient and safe way to practice mindfulness in public settings.

A Convenient Way to Access Mindfulness in Nature: Indoors

Participants described the app as convenient and an alternative to practicing outdoors. The app was not seen as a replacement but could be used when outdoor spaces were unavailable.

A Sense of Safety in the VE

Participants felt safe using the VE and again drew on relative experiences within other modalities within this theme, "even though it's virtual reality, sometimes it's like that is safer than the real world," and this was related to "...the real world [being] very confronting, people especially..." (P8).

There was a sense the VE was more predictable and a trust that nothing threatening would appear as part of the experience.

Theme Category 4: Helping to Develop Mindfulness Practice

This theme identifies some ways the VR app could be used both for initiating and restarting regular mindfulness, encompassing a longer-range view that the app may be used to develop mindfulness skills over time.

XSL•FO
RenderX

Easing Into Different Forms of Mindfulness Practice

The app could benefit beginners to mindfulness practice. Participant 16 noted this was important for those with *histories of trauma*:

> *...[because] it's really hard to get in touch with our bodies...[and] having a way...to maybe ease into it [mindfulness practice] through an external pathway could be really effective.*

Encouragement to Start Mindfulness Again

The app prompted a reconsideration to initiate a regular mindfulness practice for those participants who had been unsuccessful or found barriers. Participants stated they would be encouraged to practice more regularly.

Theme Category 5: Supporting Mental Health

Participants described the app as providing short-term "outcomes," such as improving state mindfulness, mood, and mental health symptoms.

Use for Increasing Present Moment Awareness

The app was used to achieve a state of mindfulness or a way of directing attention back to the present.

Changing Emotions and Getting Away

The VR app could be used to reduce heightened emotions during a period of ill mental health. Participants described the notion of "getting away" as the desire to take a break from these emotions or stressors:

> *It would help with trying to slow down some of those thoughts that come with the depressive episode.* [P7]

Some participants reported using the program to "let go" or sort through thoughts that were bothering them:

> *So part of it is negative thoughts that you can't quite get away from. Having mindfulness in general and the VR, as part of that, to let you get into a space where you're letting go of those feelings, those thoughts. That's where it could be very helpful.* [P2]

Related to this was the notion of a "reset," where participants used the program to take a break before continuing a busy or stressful day:

> *So it would be something that I would like to use, even if it's a 5 or 10 minute exercise in the middle of a crazy day, just to regroup and re-energize.* [P19]

Disorder-Specific Uses

The app could improve motivation or energy:

> *With depression you can often feel like really flat and lethargic and lack of motivation. So I think that [the app] can help with feeling energised too.* [P11]

The potential for this to be translated into taking action and "doing something" was noted:

> *Whether it's sparking my interest in my environment or just distracting myself from my mood or wanting to go out and do something afterwards, I think that would be a good exercise to improve my mood.* [P1]

For participants with an anxiety disorder, the program was viewed as a way of reducing anxiety before stressful activities:

> *Knowing the shift I felt in fifteen minutes from the beginning to the end...if I had something that I was anxious about doing I would totally do that first...And then go do...the activity...* [P20]

Participants with a diagnosis of BD discussed using the program to manage elevated mood or as a way of "stepping back" from this.

Use for Prevention

The program may be useful as a preventive mental health strategy. With regard to depressed mood:

> *Ironically, I would probably use it more in that early part...because I know that it helps...and at that point in the early part of a depressive episode, I do want to help myself and I think I would reach for something like that.* [P7]

Participant 15 spoke about breaking the cycle for worrying:

> *...probably being especially for people who get that circular thinking, being a bit of a circuit breaker before we kind of spiral.*

Similar to the outcome focused uses, such uses were deemed as reacting to the signs of developing mental health symptoms in the hope to stop it at that moment rather than developing skills over time.

Barriers to Using VR During a Mental Health Episode

During a period of ill mental health, there may be barriers, such as a lack of awareness or insight, which may limit the usefulness of the app:

> *I just know me. I don't think I would. And I feel like a different person in that state [manic or hypomanic]. And I don't do really anything that's a good decision at that point so, I would find it unlikely.* [P7]

For participants with a diagnosis of MDD, low motivation to engage in daily activities could prevent use of the app:

> *Yeah, but that that wouldn't be an objection to the to the VR…* [P2]

Discussion

Principal Findings

This study aimed to investigate the effects of a single, brief, VR-supported mindfulness practice on individuals with a mood or anxiety disorder and well as to explore the potential uses of this technology as a mental health management tool. The results from this study suggest that for individuals with a mood or anxiety disorder, VR apps, such as Place, have potential to improve symptoms in the short term as well as offering a safe and acceptable tool for the self-management of mental health symptoms.

Baseline FFMQ scores reflect levels of trait mindfulness, and our FFMQ mean scores were slightly lower than average scores found in the general population [21,44]. Following the VR-supported mindfulness experience, there was a significant

XSL•FO
RenderX

increase in state mindfulness, suggesting that the VR app was successful in enhancing and supporting mindfulness practice. This finding is consistent with other research in the field [22,26] as well as a previous study, which investigated the use of this VR app with a general population sample [21].

There was a significant reduction in state anxiety following the VR-supported mindfulness experience. These results contribute to a growing body of evidence for the use of VR in improving anxiety symptoms [30,45,46]. There was no significant change in positive affect following the VR experience; however, there was a significant reduction in participants' negative affect. These results differ to those of Seabrook et al [21] who found that positive affect increased, but there was no change to negative affect following the VR experience in a nonclinical population. Furthermore, there have been associations between mindfulness and positive affect found in previous studies [47]. Seabrook et al [21] used single-item measures of positive and negative affect, and it is not possible to directly compare results to this study. Nonetheless, there was a trend toward increased positive affect after the intervention, and it is likely there was insufficient power to detect small magnitude changes in this study. Furthermore, while this study had a clinical sample, positive affect scores were comparable to the general population (while negative affect scores reflected those from clinical samples), potentially resulting in a ceiling effect for positive affect [48]. The aim of mindfulness practice is not to directly alter mood states, yet mindfulness targets emotion regulation [49]. Therefore, the benefits of mindfulness for positive affect may be more evident for those experiencing low positive affect or affect variability (ie, dysregulation) than participants in this study. Further research with larger samples and time series data is needed to elucidate these relationships.

A plausible alternative explanation for the changes participants reported to anxiety and negative affect following engagement with the app is that participants were engaging in distraction (disengagement from their present moment experiences), rather than mindfulness. For example, 1 participant observed that the app allows "you [to] have something visual to take up your attention, it's quite in-depth and you've got a lot of focus on...it kept me present better than when I've done it with just audio." However, state mindfulness improved significantly after the intervention, suggesting that, at the group level, mindfulness was effectively targeted. This aligned with participants' self-assessments of having experienced mindfulness while using the app. Furthermore, prior research has suggested that more immersive, multisensory mindfulness practices and connectedness to nature both confer benefits, including increased mindfulness [27,50]. Nonetheless, it is possible that some individuals used the VE for mindfulness and others for distraction. Furthermore, central to mindfulness is noticing distraction and bringing attention back to the present. It is also possible that reducing external distractors with the VE inhibited this central aspect of mindfulness. Therefore, the potential role of distraction in the observed benefits of the VR mindfulness warrants further investigation.

Our key qualitative findings were that participants found being mindful during the VR app less effortful than previous experiences of mindfulness, specifically around attentional effort. Participants reported integrating their existing mindfulness practice skills to the app to enhance the experience. In this study, participants made direct contrasts to previous experiences, including assisting with maintaining attention, improving participants' ability to focus on their present moment experience and immersion in the practice, inducing a state of calm and reducing discomfort previously associated with mindfulness, and increasing sense of safety relative to practicing mindfulness outdoors in real-world environments. External distractions were less noticeable because of the multisensory experience of the app, leading to the feeling of being in "a bubble." Given the barriers some clinical populations face in commencing formal mindfulness practice, the app contributes to an ease of present moment awareness because of both the high level of presence offered by VEs [31] and the possible benefits of viewing a natural environment in reducing attentional demands [23,51].

Among the novel findings within this study were those relating to the utility of the app for clinical populations. This study lends support to VR mindfulness as a safe and acceptable tool for those with mood and anxiety disorders, with the potential to improve negative affect and anxiety. This finding adds to the limited, yet growing, body of literature investigating VR mindfulness in clinical populations [26,29,30]. In addition, the findings suggest that there may be unique benefits associated with VR mindfulness for different clinical groups. Specifically, participants with depression highlighted that the app may help improve motivation, interest, and energy as well as reducing the likelihood of dissociation. Those with anxiety disorders speculated that it may help them reduce their anxiety in preparation for a stressful event. Those experiencing BD identified the app as having potential to help them manage elevated moods by stepping back. These findings are speculative yet suggest that there may be unique benefits of VR mindfulness for different clinical groups, providing novel directions to be further explored within future research.

Conversely, some participants identified that their mental health symptoms may interfere with their use of the app, with the perception that they are less likely to use, or benefit from, the app when experiencing more severe symptoms. In line with this, the severity of mental health symptoms has been identified as a common barrier for accessing technology-supported mental health interventions [52]. Therefore, the role of symptoms in engagement is an essential consideration for the use of VR mindfulness in clinical populations. One possibility that emerged from qualitative findings within this study is the potential for VR mindfulness to be used as a preventive measure, which could provide the additional benefit of circumventing this barrier. Mindfulness as a preventive intervention is not novel [53,54]; however, the potential benefits of VR mindfulness as a preventive tool for clinical populations is, to the authors' knowledge, yet to be explored.

Focus on the visual stimuli was generally seen as engagement and present moment awareness. However, for 1 participant, it was not seen as compatible with their understanding of prior mindfulness practice in which they closed their eyes. Previous focus group evaluations of the app reported that participants expressed a desire to close their eyes because of attentional

XSL•FO
RenderX

fatigue and to match previous mindfulness experiences [40]. Visual stimuli are attentionally favored when both visual and auditory stimuli are presented [55,56]. This may be an important consideration for future development of the audio voiceover used for the app, with directions for participants to open and close their eyes offered as part of the introduction to the exercise.

Participants' perspectives suggested that the app might be used as a starting point in learning mindfulness skills. It was noted that for participants previously finding it difficult to connect with their bodies, the app enabled a more externally focused practice. While participants may see the value in focusing inward, as is commonly the approach in MBIs, they may not be ready to do this straight away. An approach of focusing out before focusing in may support participants who may otherwise be confronted with an increased awareness of internal state, with this experience of confrontation or awareness of unpleasant internal experiences being echoed in prior qualitative research [9,12].

Participants indicated perceiving multiple potential future uses of the app in supporting their mental health self-management including the goals of altering emotional state and increasing mindful awareness. Some identified benefits were perceived as addressing challenges participants had faced with previous mindfulness practices, and others were characteristics of the present intervention. For example, participants suggested the app could function as both a mindfulness and emotion regulation tool, depending on the goals of their mental health treatment. Our results indicate that the use of this VR app is associated with reduced state anxiety and negative affect, making the present moment more appealing. In a study by Riva et al [57], emotional valence and felt presence (as opposed to present moment awareness) were found to be related in VEs, such that environments with more emotionally prominent features (eg, experimentally manipulated relaxing or anxious environments) evoked higher felt presence than emotionally neutral environments. It is possible that participants who found the environment relaxing in this study were more able to feel involved and engaged in the environment, in turn supporting present moment awareness. In addition to participants reporting a feeling of safety within the VE, there was also a lack of reported adverse events associated with the study.

Given the dearth of literature investigating VR mindfulness in clinical populations, there is currently inconclusive support for benefits of VR mindfulness over and above mindfulness interventions more broadly [58]. This study identified some potential unique benefits of the app relative to other mindfulness practices based on participants' qualitative experiences. In line with these findings, a recent systematic review found that more immersive VR-based mindfulness training was more effective than either less-immersive VR (eg, 2D computer-generated scenes) or conventional mindfulness training for improving engagement and adherence and for improving mindfulness levels in clinical and nonclinical populations [27]. While participants in this study offered reflections based on their prior experiences with mindfulness, we did not directly compare the VR app with other mindfulness practices. Firm conclusions cannot be drawn based on these data; however, they offer some potentially informative directions warranting further exploration.

Some of the barriers to conventional mindfulness practice may translate to VR mindfulness as well. Participants reported a potential lack of motivation to practice mindfulness during depressive episodes, which was consistent with previous research with individuals with BD [13]. In this study, barriers relating to using the VR app to improve mental health were almost solely identified by those with a diagnosis of MDD or BD, suggesting these groups may be at greater risk of difficulties in engagement compared to those with an anxiety disorder. Given the limited context for using the app this study, it was not possible to corroborate the opposing perspective of some participants that the app may help improve energy and motivation. However, as noted, a recent review found VR mindfulness was associated with better engagement and adherence than non–VR mindfulness [27]. Future research could investigate how to increase the likelihood that participants will engage with the app: this might include enhancing the design of the VR experience. There is potential for integration with other areas of developing digital technologies, such as ecological momentary intervention.

Important limitations of this study were its sample size and context, both limiting the generalizability of findings. A single app was used, using 360° video environments and with audio guidance, so findings may not generalize to other types of VR-supported mindfulness apps. The testing session was a one-off experience in a laboratory setting, and the results may not generalize to real-world environments, where distractions and ease of focus may differ. The study sample was small, heterogenous, and familiar with mindfulness practice, with no control group. In addition, there was no measurement of how well participants engaged in the mindfulness task; therefore, it is possible the changes to state anxiety, affect, and state and trait mindfulness may have been attributable to other processes, such as "resting" in a virtual nature environment or distraction, irrespective of the mindfulness instructions. We sought participants' perspectives on whether the Place app supports mindfulness practice, with responses indicating that the app was perceived to have this potential benefit. However, it is important to contextualize these findings; participants were not recruited based on expertise in mindfulness practice and varied in mindfulness experience. They could only reflect on their own experiences and use of the Place app in the given context. Furthermore, the qualitative interviews were semistructured, and the questions asked are likely to have influenced the data generated, while more open-ended prompts may allow themes to emerge more organically. Similarly, we asked participants to speculate about how they might use Place in future, generating some promising directions; however, these need to be validated via observation of app use within real-world contexts. Our participants' perspectives provide helpful avenues for further investigation, and further research is essential to form any conclusions.

Future directions include assessment of the VR app in more ecologically valid settings, over a more prolonged course of mindfulness practice, and with a broader and more diverse sample, reflecting potential end users. Future research should

XSL•FO
RenderX

also investigate how to increase the likelihood that participants will engage with the app: this might include enhancing the design of the VR experience. While this study provided some indications of potential future uses for the Place app, investigating end users' engagement and interactions with the app in real-world contexts is an important next step. As noted, Place was designed as an experiential learning and on-the-spot self-management tool [41]. However, short-term and stand-alone mindfulness practices have been associated with altered neural activity as well as "downstream," longer-term improvements [4,15,16,42], which may indicate generalization of skills in accordance with "offline training" [41]. This study used a one-off experience of the app in the laboratory context, and as such further research, supported by tailored measurement (eg, immediate vs generalized benefits) is needed to better understand the contexts in which the app may be beneficial.

Further investigation is needed to assess whether the VE is facilitating mindfulness or potentially other processes, such as distraction or relaxation. In addition, one possible use generated by participants in this study is VR mindfulness as a preventive measure. Furthermore, this study provided some preliminary signals of unique benefits of the app for different clinical groups, which warrant investigation within future research. A further unanswered question requiring confirmation is whether VR mindfulness confers additional benefits to traditional mindfulness practices for clinical groups. Promising avenues for future research therefore include rigorous clinical trials comparing outcomes for different clinical groups as well as comparing VR mindfulness with other mindfulness practices, evaluating the use of the app for prevention in both research and real-world contexts, assessing whether results can be generalized to VR mindfulness practices other than Place, and integration with other developing digital technologies, such as ecological momentary assessment and intervention.

Conclusions

This study provides preliminary empirical support for VR-supported mindfulness having potential utility in those with mood or anxiety disorders. Evidence of proximal effects and perceived benefits support the potential for VR apps having a role in supporting mindfulness practice. While based on a single app, this study generated some evidentiary signals relating to potential future uses of VR mindfulness within clinical populations.

Acknowledgments

The authors would like to acknowledge the contributions of PHORIA in the development of the virtual reality mindfulness app, Place.

Conflicts of Interest

None declared.

References

1. Bishop SR, Lau M, Shapiro S, Carlson L, Anderson ND, Carmody J, et al. Mindfulness: a proposed operational definition. Clin Psychol Sci Pract 2004 Aug 01;11(3):230-241. [doi: 10.1093/clipsy/bph077]
2. Kabat-Zinn J. Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness. New York, NY: Delta Trade Paperbacks; 1990.
3. Zhang D, Lee EK, Mak EC, Ho CY, Wong SY. Mindfulness-based interventions: an overall review. Br Med Bull 2021 Jun 10;138(1):41-57 [FREE Full text] [doi: 10.1093/bmb/ldab005] [Medline: 33884400]
4. Chiesa A, Serretti A, Jakobsen JC. Mindfulness: top-down or bottom-up emotion regulation strategy? Clin Psychol Rev 2013 Feb;33(1):82-96. [doi: 10.1016/j.cpr.2012.10.006] [Medline: 23142788]
5. Hölzel BK, Lazar SW, Gard T, Schuman-Olivier Z, Vago DR, Ott U. How does mindfulness meditation work? proposing mechanisms of action from a conceptual and neural perspective. Perspect Psychol Sci 2011 Nov;6(6):537-559. [doi: 10.1177/1745691611419671] [Medline: 26168376]
6. Hofmann SG, Sawyer AT, Witt AA, Oh D. The effect of mindfulness-based therapy on anxiety and depression: a meta-analytic review. J Consult Clin Psychol 2010 Apr;78(2):169-183 [FREE Full text] [doi: 10.1037/a0018555] [Medline: 20350028]
7. Chu CS, Stubbs B, Chen TY, Tang CH, Li DJ, Yang WC, et al. The effectiveness of adjunct mindfulness-based intervention in treatment of bipolar disorder: a systematic review and meta-analysis. J Affect Disord 2018 Jan 01;225:234-245. [doi: 10.1016/j.jad.2017.08.025] [Medline: 28841486]
8. Xuan R, Li X, Qiao Y, Guo Q, Liu X, Deng W, et al. Mindfulness-based cognitive therapy for bipolar disorder: a systematic review and meta-analysis. Psychiatry Res 2020 Aug;290:113116. [doi: 10.1016/j.psychres.2020.113116] [Medline: 32480120]
9. Wyatt C, Harper B, Weatherhead S. The experience of group mindfulness-based interventions for individuals with mental health difficulties: a meta-synthesis. Psychother Res 2014 Dec 07;24(2):214-228. [doi: 10.1080/10503307.2013.864788] [Medline: 24313300]
10. Segal ZV, Williams JM, Teasdale JD. Mindfulness-Based Cognitive Therapy for Depression: A New Approach to Preventing Relapse. New York, NY: The Guilford Press; 2002.
11. Anderson T, Suresh M, Farb NA. Meditation benefits and drawbacks: empirical codebook and implications for teaching. J Cogn Enhanc 2019 Jan 14;3(2):207-220. [doi: 10.1007/s41465-018-00119-y]

XSL•FO
RenderX

12. Lomas T, Cartwright T, Edginton T, Ridge D. A qualitative analysis of experiential challenges associated with meditation practice. Mindfulness 2014 Aug 10;6(4):848-860. [doi: 10.1007/S12671-014-0329-8]
13. Hanssen I, van der Horst N, Boele M, Lochmann van Bennekom M, Regeer E, Speckens A. The feasibility of mindfulness-based cognitive therapy for people with bipolar disorder: a qualitative study. Int J Bipolar Disord 2020 Nov 11;8(1):33 [FREE Full text] [doi: 10.1186/s40345-020-00197-y] [Medline: 33175338]
14. Hayes SC, Shenk C. Operationalizing mindfulness without unnecessary attachments. Clin Psychol Sci Pract 2004;11(3):249-254. [doi: 10.1093/clipsy.bph079]
15. Howarth A, Smith JG, Perkins-Porras L, Ussher M. Effects of brief mindfulness-based interventions on health-related outcomes: a systematic review. Mindfulness 2019 May 31;10(10):1957-1968. [doi: 10.1007/s12671-019-01163-1]
16. Blanck P, Perleth S, Heidenreich T, Kröger P, Ditzen B, Bents H, et al. Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: systematic review and meta-analysis. Behav Res Ther 2018 Mar;102:25-35. [doi: 10.1016/j.brat.2017.12.002] [Medline: 29291584]
17. Sliwinski J, Katsikitis M, Jones CM. A review of interactive technologies as support tools for the cultivation of mindfulness. Mindfulness 2017 Mar 9;8(5):1150-1159. [doi: 10.1007/s12671-017-0698-x]
18. Rothbaum BO. Using virtual reality to help our patients in the real world. Depress Anxiety 2009 Mar;26(3):209-211. [doi: 10.1002/da.20556] [Medline: 19253416]
19. MacQuarrie A, Steed A. Cinematic virtual reality: evaluating the effect of display type on the viewing experience for panoramic video. In: Proceedings of the 2017 IEEE Annual International Symposium Virtual Reality. 2017 Presented at: VR '17; March 18-22, 2017; Los Angeles, CA p. 45-54. [doi: 10.1109/vr.2017.7892230]
20. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557 [FREE Full text] [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]
21. Seabrook E, Kelly R, Foley F, Theiler S, Thomas N, Wadley G, et al. Understanding how virtual reality can support mindfulness practice: mixed methods study. J Med Internet Res 2020 Mar 18;22(3):e16106 [FREE Full text] [doi: 10.2196/16106] [Medline: 32186519]
22. Navarro-Haro MV, López-Del-Hoyo Y, Campos D, Linehan MM, Hoffman HG, García-Palacios A, et al. Meditation experts try virtual reality mindfulness: a pilot study evaluation of the feasibility and acceptability of virtual reality to facilitate mindfulness practice in people attending a mindfulness conference. PLoS One 2017 Nov 22;12(11):e0187777 [FREE Full text] [doi: 10.1371/journal.pone.0187777] [Medline: 29166665]
23. Lymeus F, Lundgren T, Hartig T. Attentional effort of beginning mindfulness training is offset with practice directed toward images of natural scenery. Environment and Behavior 2016 Jul 12;49(5):536-559. [doi: 10.1177/0013916516657390]
24. Lymeus F, Lindberg P, Hartig T. A natural meditation setting improves compliance with mindfulness training. J Environ Psychol 2019 Aug;64:98-106. [doi: 10.1016/j.jenvp.2019.05.008]
25. American Psychiatric Association. Diagnostic And Statistical Manual Of Mental Disorders, Text Revision (DSM-5-TR). 5th edition. Washington, DC: American Psychiatric Association Publishing; 2022.
26. Navarro-Haro MV, Modrego-Alarcón M, Hoffman HG, López-Montoyo A, Navarro-Gil M, Montero-Marin J, et al. Evaluation of a mindfulness-based intervention with and without virtual reality dialectical behavior therapy mindfulness skills training for the treatment of generalized anxiety disorder in primary care: a pilot study. Front Psychol 2019 Jan 28;10:55 [FREE Full text] [doi: 10.3389/fpsyg.2019.00055] [Medline: 30745888]
27. Ma J, Zhao D, Xu N, Yang J. The effectiveness of immersive virtual reality (VR) based mindfulness training on improvement mental-health in adults: a narrative systematic review. Explore (NY) 2023 May;19(3):310-318 [FREE Full text] [doi: 10.1016/j.explore.2022.08.001] [Medline: 36002363]
28. Chandrasiri A, Collett J, Fassbender E, de Foe A. A virtual reality approach to mindfulness skills training. Virtual Real 2019 Feb 18;24(1):143-149. [doi: 10.1007/s10055-019-00380-2]
29. Seol E, Min S, Seo S, Jung S, Lee Y, Lee J, et al. "Drop the beat": virtual reality based mindfulness and cognitive behavioral therapy for panic disorder --- a pilot study. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology. 2017 Presented at: VRST '17; November 8-10, 2017; Gothenburg, Sweden p. 1-3 URL: https://dl.acm.org/doi/abs/10.1145/3139131.3141199 [doi: 10.1145/3139131.3141199]
30. Tarrant J, Viczko J, Cope H. Virtual reality for anxiety reduction demonstrated by quantitative EEG: a pilot study. Front Psychol 2018 Jul 24;9:1280 [FREE Full text] [doi: 10.3389/fpsyg.2018.01280] [Medline: 30087642]
31. Kitson A, Prpa M, Riecke BE. Immersive interactive technologies for positive change: a scoping review and design considerations. Front Psychol 2018;9:1354 [FREE Full text] [doi: 10.3389/fpsyg.2018.01354] [Medline: 30123161]
32. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-34. [Medline: 9881538]
33. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011 Dec;168(12):1266-1277 [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
34. Baer RA, Smith GT, Hopkins J, Krietemeyer J, Toney L. Using self-report assessment methods to explore facets of mindfulness. Assessment 2006 Mar;13(1):27-45. [doi: 10.1177/1073191105283504] [Medline: 16443717]

XSL•FO
RenderX

35. Gu J, Strauss C, Bond R, Cavanagh K. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? a systematic review and meta-analysis of mediation studies. Clin Psychol Rev 2015 Apr;37:1-12. [doi: 10.1016/j.cpr.2015.01.006] [Medline: 25689576]
36. Lau MA, Bishop SR, Segal ZV, Buis T, Anderson ND, Carlson L, et al. The Toronto mindfulness scale: development and validation. J Clin Psychol 2006 Dec;62(12):1445-1467. [doi: 10.1002/jclp.20326] [Medline: 17019673]
37. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 1988;54(6):1063-1070. [doi: 10.1037/0022-3514.54.6.1063]
38. Spielberger CD. State-trait anxiety inventory for adults (STAI-AD). APA PsycTests. 1983. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft06496-000 [accessed 2024-04-05]
39. Bradley A. Trait & State Anxiety: Assessment, Predictors & Outcomes. New York, NY: Nova Publishers; 2016.
40. Kelly RM, Seabrook EM, Foley F, Thomas N, Nedeljkovic M, Wadley G. Design considerations for supporting mindfulness in virtual reality. Front Virtual Real 2022 Jan 21;2:155. [doi: 10.3389/frvir.2021.672556]
41. Slovak P, Antle A, Theofanopoulou N, Daudén Roquet C, Gross J, Isbister K. Designing for emotion regulation interventions: an agenda for HCI theory and research. ACM Trans Comput Hum Interact 2023 Mar 18;30(1):1-51. [doi: 10.1145/3569898]
42. Kral TR, Schuyler BS, Mumford JA, Rosenkranz MA, Lutz A, Davidson RJ. Impact of short- and long-term mindfulness meditation training on amygdala reactivity to emotional stimuli. Neuroimage 2018 Nov 01;181:301-313 [FREE Full text] [doi: 10.1016/j.neuroimage.2018.07.013] [Medline: 29990584]
43. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: Sage Publications; Mar 01, 2016.
44. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
45. Maples-Keller JL, Bunnell BE, Kim SJ, Rothbaum BO. The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. Harv Rev Psychiatry 2017;25(3):103-113 [FREE Full text] [doi: 10.1097/HRP.0000000000000138] [Medline: 28475502]
46. Parsons TD, Rizzo AA. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: a meta-analysis. J Behav Ther Exp Psychiatry 2008 Sep;39(3):250-261. [doi: 10.1016/j.jbtep.2007.07.007] [Medline: 17720136]
47. Giluk TL. Mindfulness, big five personality, and affect: a meta-analysis. Pers Individ Dif 2009 Dec;47(8):805-811. [doi: 10.1016/j.paid.2009.06.026]
48. Crawford JR, Henry JD. The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. Br J Clin Psychol 2004 Sep;43(Pt 3):245-265. [doi: 10.1348/0144665031752934] [Medline: 15333231]
49. Guendelman S, Medeiros S, Rampes H. Mindfulness and emotion regulation: insights from neurobiological, psychological, and clinical studies. Front Psychol 2017;8:220 [FREE Full text] [doi: 10.3389/fpsyg.2017.00220] [Medline: 28321194]
50. Howell AJ, Dopko RL, Passmore HA, Buro K. Nature connectedness: associations with well-being and mindfulness. Pers Individ Dif 2011 Jul;51(2):166-171. [doi: 10.1016/j.paid.2011.03.037]
51. Lymeus F, Lindberg P, Hartig T. Building mindfulness bottom-up: meditation in natural settings supports open monitoring and attention restoration. Conscious Cogn 2018 Mar;59:40-56. [doi: 10.1016/j.concog.2018.01.008] [Medline: 29438869]
52. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
53. McCartney M, Nevitt S, Lloyd A, Hill R, White R, Duarte R. Mindfulness-based cognitive therapy for prevention and time to depressive relapse: systematic review and network meta-analysis. Acta Psychiatr Scand 2021 Jan 27;143(1):6-21. [doi: 10.1111/acps.13242] [Medline: 33035356]
54. Tang YY, Leve LD. A translational neuroscience perspective on mindfulness meditation as a prevention strategy. Transl Behav Med 2016 Mar;6(1):63-72 [FREE Full text] [doi: 10.1007/s13142-015-0360-x] [Medline: 27012254]
55. Colavita FB. Human sensory dominance. Percept Psychophys 1974 Mar;16(2):409-412. [doi: 10.3758/bf03203962]
56. Hirst RJ, Cragg L, Allen HA. Vision dominates audition in adults but not children: a meta-analysis of the Colavita effect. Neurosci Biobehav Rev 2018 Nov;94:286-301 [FREE Full text] [doi: 10.1016/j.neubiorev.2018.07.012] [Medline: 30048672]
57. Riva G, Mantovani F, Capideville CS, Preziosa A, Morganti F, Villani D, et al. Affective interactions using virtual reality: the link between presence and emotions. Cyberpsychol Behav 2007 Feb;10(1):45-56. [doi: 10.1089/cpb.2006.9993] [Medline: 17305448]
58. Zhang S, Chen M, Yang N, Lu S, Ni S. Effectiveness of VR based mindfulness on psychological and physiological health: a systematic review. Curr Psychol 2021 May 19;42(6):5033-5045. [doi: 10.1007/s12144-021-01777-6]

Abbreviations

BD: bipolar disorder
FFMQ-SF: Five Facet Mindfulness Questionnaire Short Form
MBI: mindfulness-based intervention

XSL•FO
RenderX

MDD: major depressive disorder
PANAS: Positive and Negative Affect Schedule
TMS: Toronto Mindfulness Scale
VE: virtual environment
VR: virtual reality

Edited by T McCall; submitted 10.01.23; peer-reviewed by S Sun, A Kitson, I Ezawa; comments to author 22.03.23; revised version received 15.12.23; accepted 12.02.24; published 06.12.24.

Please cite as:
Blackmore R, Giles C, Tremain H, Kelly R, Foley F, Fletcher K, Nedeljkovic M, Wadley G, Seabrook E, Thomas N
Examining the Use of Virtual Reality to Support Mindfulness Skills Practice in Mood and Anxiety Disorders: Mixed Methods Study
J Med Internet Res 2024;26:e45640
URL: https://www.jmir.org/2024/1/e45640
doi:10.2196/45640
PMID:39641990

©Rebecca Blackmore, Claudia Giles, Hailey Tremain, Ryan Kelly, Fiona Foley, Kathryn Fletcher, Maja Nedeljkovic, Greg Wadley, Elizabeth Seabrook, Neil Thomas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis

Nuša Farič[1,2], BSc, MSc, PhD; Henry WW Potts[3], BA, MSc, PhD; James M Heilman[4], BSc, MD

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

[2]Advanced Care Research Center, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

[3]University College London, Institute of Health Informatics, London, United Kingdom

[4]Department of Emergency Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Henry WW Potts, BA, MSc, PhD
University College London
Institute of Health Informatics
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 0207679200
Email: h.potts@ucl.ac.uk

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e66590

Abstract

Background: Wikipedia is the largest free online encyclopedia and the seventh most visited website worldwide, containing >45,000 freely accessible English-language medical articles accessed nearly 1.6 billion times annually. Concerns have been expressed about the balance of content related to biological sex on Wikipedia.

Objective: This study aims to categorize the top 1000 most-read (most popular) English-language Wikipedia health articles for June 2019 according to the relevance of the article topic to each sex and quality.

Methods: In the first step, Wikipedia articles were identified using WikiProject Medicine Popular Pages. These were analyzed on 13 factors, including total views, article quality, and total number of references. In the second step, 2 general medical textbooks were used as comparators to assess whether Wikipedia's spread of articles was typical compared to the general medical coverage. According to the article's content, we proposed criteria with 5 categories: 1="exclusively female," 2="predominantly female but can also affect male individuals," 3="not sex specific or neutral," 4=predominantly male but can affect female individuals," and 5="exclusively male."

Results: Of the 1000 Wikipedia health articles, 933 (93.3%) were not sex specific and 67 (6.7%) were sex specific. There was no statistically significant difference in the number of reads per month between the sex-specific and non–sex-specific articles (P=.29). Coverage of female topics was higher (50/1000, 5%) than male topics (17/1000, 1.7%; this difference was also observed for the 2 medical textbooks, in which 90.2% (2330/2584) of content was not sex specific, female topics accounted for 8.1% (209/2584), and male topics for accounted for 1.7% (45/2584; statistically significant difference; Fisher exact test P=.03). Female-category articles were ranked higher on the Wikipedia medical topic importance list (top, high, or mid importance) than male-category articles (borderline statistical significance; Fisher exact test P=.05). Female articles had a higher number of total and unique references; a slightly higher number of page watchers, pictures, and available languages; and lower number of edits than male articles (all were statistically nonsignificant).

Conclusions: Across several metrics, a sample of popular Wikipedia health-related articles for both sexes had comparable quality. Wikipedia had a lower number of female articles and a higher number of neutral articles relative to the 2 medical textbooks. These differences were small, but statistically significant. Higher exclusively female coverage, compared to exclusively male coverage, in Wikipedia articles was similar to the 2 medical textbooks and can be explained by inclusion of sections on obstetrics and gynecology. This is unlike the imbalance seen among biographies of living people, in which approximately 77.6% pertain

XSL•FO
RenderX

to male individuals. Although this study included a small sample of articles, the spread of Wikipedia articles may reflect the readership and the population's content consumption at a given time. Further study of a larger sample of Wikipedia articles would be valuable.

(J Med Internet Res 2024;26:e47562) doi:10.2196/47562

KEYWORDS

Wikipedia; wikis; writing; internet; health information; sex; sex bias; consumer health information; health communication; public education; social media

Introduction

Background

Wikipedia is the largest, free, multilingual online encyclopedia. It is the seventh most-visited website worldwide and is available in 329 languages, with 6,829,103 articles in the English language as of May 30, 2024 [1-3]. It is also a source for a new generation of artificial intelligence technologies, such as ChatGPT [4]. Wikipedia articles can be edited and created by anyone from around the world.

Wikipedia includes >45,000 freely accessible English articles related to medical topics, which are accessed nearly 1.8 billion times per year [5]. The editors of health articles range from physicians, researchers, policy makers, medical students, and anyone else interested in human health (eg, patients) [6,7]. Medical articles include topics about human health, health organizations, and notable people in medicine and health care. On Wikipedia, these articles are ranked according to their importance to the field of medicine, and they are also ranked in order of popularity in terms of number of views (traffic to the page). These articles are also assessed for quality within 6 categories ranging from "stub" (an article with limited information, not considered enough as an encyclopedic coverage) to "start," "C-class," "B-class," "good article" (GA), and "featured article" (FA; with an in-depth examination and peer review) [8]. Quality can also be assessed by a language-agnostic quality prediction tool that can predict the quality of the Wikipedia article using a single model based on length, references, images, categories, links, and sections [9].

Before COVID-19, estimates of English medical monthly page views on Wikipedia were around 200 million, with the early part of the pandemic in March 2020 seeing a rise to 400 million hits [10]. This number decreased to around 150 million in January 2022 [10]. Estimates in December 2021 found that Wikipedia was the fourth most-visited website for health content (approximately 260 million page views across all languages), surpassing WebMD (approximately 240 million page views), the Mayo Clinic (approximately 230 million page views), the World Health Organization (approximately 80 million page views), and UpToDate (approximately 50 million page views), but below Healthline (which had the highest level of hits; approximately 1100 million page views), the US National Institutes of Health (approximately 600 million page views), and the United Kingdom National Health Service (approximately 460 million page views) [10]. Nonetheless, Wikipedia remains among the most widespread resources for accessing health or medical information.

Online Health-Seeking Behavior

The most preferred way of obtaining health information is seeking it on the web because of fewer barriers to accessing content, affordability, coverage of information, interactivity, privacy, and confidentiality [11,12]. Information seeking is often initiated as a response to uncertainty regarding one's health, such as considering treatment options or as preparation for a consultation with a health care provider [13,14]. Among other factors, sex has been shown to play a considerable role in information-seeking behavior in some studies, suggesting that female individuals are more active seekers of general health-related information than male individuals, even across disparities in health literacy [15]. However, other studies have not confirmed any difference in health information seeking between the sexes (eg, [13]) and instead argue that this is likely because of sociodemographic factors and individual differences. Health information is often also subject to scrutiny from both health care professionals and laypeople [16]. Part of health behavior, either by default or as a rule, involves seeking and evaluating the quality and reliability of information presented to us [16].

Representativeness of Male or Female Wikipedia Content

Wikipedia editors have been reported to be predominantly male for medical and health articles [7], and studies have found similar trends for other topics on Wikipedia in different languages, with about 85% to 90% of editors identifying as male, 10% to 15% as female, and 1% to 2% as transgender, even after a slow increase in female contributors from year to year [17-21], a difference attributed to the enduring gender imbalance in computer-related fields [22]. This difference in editing numbers among the self-identified sexes is believed to be due to psychosocial factors [23,24]. Some editorial behavior is influenced by Wikipedia culture across the whole encyclopedia and the tensions between different editors (eg, core contributors vs peripheral contributors) [25]. If this study finds obvious unequal coverage for either sex, it is a possibility that it may partly be explained by the creators of the content [26].

Disparity in gender representation and participation in technology-related fields, particularly in open-source communities, has been studied since the 1980s [22]. There is sufficient evidence that gender imbalance carries into content generation via social contexts because of various factors, such as women taking up fewer technology-related jobs and sociocultural factors surrounding women (eg, gender stereotypes, gender expectations, women being averse to

XSL•FO
RenderX

conflict, lack of time, other life responsibilities such as raising children, lack of user-friendliness, women steering away from competitive environments, women fearing that their content would be deleted or reverted ["impostor syndrome"], and personal choice) [24,27]. The evidence for sexism on Wikipedia is mixed. One article states, "WP [Wikipedia] is male dominated. That doesn't mean it is sexist" [28]. Most Wikipedia editor profiles are anonymous (ie, use a username rather than own identity), and it would be difficult to enact direct sexism; however, some reports showed Wikipedia being hostile to women (eg, female editors being harassed by male colleagues) [29]. Studies have revealed that Wikipedia is biased in how men and women are characterized in their biographies [30], including how women's biographies suitable for inclusion on Wikipedia are twice as frequently categorized into nonnotable and are more frequently nominated for deletion (25%) than men's biographies (17%) [31]. This is true even after international coverage of the "Donna Strickland effect"—a female Nobel Prize winner in physics in 2018 whose Wikipedia profile was deleted [31]. This points to multilayered gender discrimination and may carry over into content production [32]. Considering this, no studies have been performed exclusively on health and medical content related to sex and gender.

Despite fewer women than men editing Wikipedia, recent research found a minimal amount of evidence of biological sex bias in terms of content [33,34], and 1 study found evidence of bias against men, with articles on male chief executive officer (CEO) profiles generally having a longer profile life span but less reliable sources, being less edited, and receiving less attention than female profiles [35]. Notably, for coverage of biographies of living persons, out of 1.7 million biographies on Wikipedia, as of May 5, 2019, a total of 77.6% were of men and 22.7% were of women, pointing to possible inequalities in digital coverage of biographies of people [36]. Other researchers reported the proportion of female biographies to be at 16% [33].

Several studies have also looked into the gender distribution gap in the representation of actual social groups [37]. For example, a study of actively publishing economists on Wikipedia found that women are half as likely to have a biographical entry on the English-language Wikipedia than men [38]. Halfaker [39] found that for English-language Wikipedia, the average quality of articles on women scientists was lower than that of male scientists before 2014 but exceeded the average quality after that. A study on women mathematicians found biases toward men [40], whereas a study including Fortune's top 1000 CEO profiles found a bias toward women [41]. It is important to note that these disparities likely reflect the traditional sex roles that existed in society before the turn of the century.

There are recent efforts by Wikipedia to remove this gap (#WikiGap and #ProjectRewrite, particularly by Wikimedia Sweden) and by WikiProject Women in Red to address sex inequality on Wikipedia [42-44]. Despite the still existing gaps in the coverage of biographies of living persons, it is unclear if the same disparity exists within other topic areas of Wikipedia, for example, the medical and health pages or regarding content coverage of female or male health topics [45,46]. Because anyone can benefit from access to medical and health information and because health and medical information is available in hundreds of languages, we decided to investigate this area for clarification. We are not investigating the number of male and female editors or readers but are specifically interested in the coverage of male and female health content on Wikipedia. There is a lack of research about the coverage of sex-based health and medical topics in the English-language Wikipedia. We also wanted to compare the Wikipedia articles to the coverage of the topics in general medical textbooks, which, in this study, will provide our baseline.

Access to free, reliable, and up-to-date health information is essential because of the size of global traffic to medical and health pages, including Wikipedia. If we learn that there is an obvious disparity in the coverage of medical and health topics, we may be able to investigate the possible contributing factors. We would also like to identify Wikipedia article types that may require improvement in terms of quality.

Objectives

Because of Wikipedia's significance as a leading global resource for health information, this study aims to explore potential disparities in the coverage of health articles related to biological sex. The objectives of the study are as follows: (1) to analyze Wikipedia's top 1000 most-read health articles and determine the coverage of sex-specific and non–sex-specific topics, and (2) to assess the quality of male versus female health content on English-language Wikipedia, using both Wikipedia's classification system and other markers of completeness.

Methods

Ethical Considerations

The data that we extracted from Wikipedia is freely available. It did not include any personally identifiable data or any other forms of data that may cause a breach of any ethical or regulatory rules and regulations. No ethics approval was needed.

Design

This study used quantitative research methods. We have concentrated on the top 1000 Wikipedia health articles as the most popular Wikipedia pages by traffic.

Search Strategy

First, we defined what we meant by "male" and "female" health topics. We focused on the medical, biological, or physical characteristics of the human body and not gender (social perception and individual choice of identity). We defined male articles in which the Wikipedia topic in question included at least 80% of people with male biological sex, female articles in which the Wikipedia topic in question included at least 80% of people with female biological sex, and not sex specific, in which the Wikipedia topic in question could relate in equal proportions to all biological sexes. Therefore, we developed criteria for sex-specific and non–sex-specific health articles based on the degree to which they may correspond to sex specificity: 1=exclusively female, 2=predominantly female but can affect male individuals, 3=not sex specific or neutral, 4=predominantly male but can also affect female individuals, 5=exclusively male. Second, a list of the 1000 most-read medical articles on Wikipedia for June 2019 was identified using the

XSL•FO
RenderX

"WikiProject Medicine Popular pages" list, which is a list of the top 1000 pages ordered by the number of views as of July 16, 2019 [5] (Wikipedia provides an overview of top articles monthly; hence, we collected data for June 2019 in July 2019).

Eligibility Criteria

Our research was not guided by any framework, although similar categories have been analyzed in previous Wikipedia literature [14]. Topics related to female anatomy and childbirth and conditions that affect female individuals were deemed "exclusively female." Articles related to male anatomy were deemed to be "exclusively male." Non–sex-specific articles included topics such as famous people in medicine (both physicians and patients alike, eg, Louis Pasteur, Florence Nightingale, and Phineas Gage), topics ranging from epidemiology or political events (eg, rape statistics and domestic violence), lifestyles related to health (eg, Mediterranean diet, noise pollution, and surrogacy), medical procedures (eg, Rorschach test, hygiene, and x-ray), medical drugs and chemical compounds (eg, buprenorphine, vitamin E, and cannabis), poisonous plants (eg, Datura), medical literature (eg, PubMed and the *Diagnostic and Statistical Manual of Mental Disorders*), medical conditions (eg, hernia, gastritis, and polio), medical terminology (eg, VO_2 max, biosafety level, and reflex syncope), medical events (eg, 2019 India doctor's strike and National Doctor's Day), and medical professions (eg, pathology and nursing).

Screening Process

Our methodology included several steps. After extracting a list of the 1000 most-read medical articles on Wikipedia for June 2019, the authors extracted only sex-specific articles (n=67, 6.7%) and placed them in a spreadsheet (Multimedia Appendix 1). The remaining articles concerned health or medical issues previously mentioned in various neutral categories and were therefore classified as "not sex specific." The list of medical articles was reviewed manually by 2 authors (JMH and NF) to classify the article as primarily or exclusively related to health care in either sex. Disagreements were resolved via discussion and by searching peer-reviewed medical publications for existing data to resolve unclear cases.

Assessment of the Quality of Articles

We needed to create a comparator to conclude whether the coverage of medical topics on Wikipedia was typical or deviated from a gold standard for the general medical field. Therefore, in step 2, we followed the same method of classification as in step 1, but we replaced the source with medical textbooks. We chose 2 medical textbooks and evaluated topics in the same way as in step 1. Medical textbooks that are standardly used in medical or medicine studies in the United Kingdom and Canada, where the authors are based, are the *Oxford Handbook of Clinical Medicine* [47] and *Toronto Notes: Comprehensive Medical Reference and Review for Medical Council of Canada Qualifying Examination and US Medical License Examination* [48]. The counts for these were kept in the Excel spreadsheet for each separate resource (Tables 1 and 2). Unlike in step 1, we did not assess the quality of medical content in the medical textbooks.

Table 1. Breakdown of the results for sex-specific health articles on Wikipedia in terms of article assessment.

Number of Wikipedia articles based on article assessment	FA[a], n (%)	GA[b], n (%)	B[c], n (%)	C[d], n (%)	Start[e], n (%)	Total, n (%)
Male	0 (0)	1 (6)	7 (41)	6 (35)	3 (18)	17 (100)
Female	2 (4)	2 (4)	32 (64)	10 (20)	4 (8)	50 (100)

[a]FA: featured article status.

[b]GA: good article status.

[c]B: B-class article status.

[d]C: C-class article status.

[e]Start: "start" article status.

Table 2. Breakdown of the results for sex-specific health articles on Wikipedia in terms of article status.

Importance rating of male vs female health-related Wikipedia articles	Top, n (%)	High, n (%)	Mid, n (%)	Low, n (%)	Total, n (%)
Male	1 (6)	3 (18)	10 (59)	3 (18)	17 (100)
Female	7 (14)	23 (46)	17 (34)	3 (6)	50 (100)

Categories Generated

We developed 5 categories: 1="exclusively female," 2="predominantly female but can also affect male individuals," 3="not sex specific or neutral," 4=predominantly male but can affect female individuals," and 5="exclusively male." Category 2, for example, covers urinary tract infection, breast cancer, human papillomavirus, anorexia, and bulimia nervosa, which all primarily affect the female population, but they can also affect male individuals. Similarly, there are conditions or topics concerning health that primarily affect male individuals but can be found in female individuals too, such as testosterone or inguinal hernia [49], which make up category 4. Our cutoff point for "predominantly" categories was 80% of cases (eg, inguinal hernia appears in an 8:1 male:female ratio) [49].

XSL•FO
RenderX

Statistical Analyses

To assess statistically significant differences for probability samples, we performed standard nonparametric tests, including the Mann-Whitney test and Fisher exact tests. Mann-Whitney tests were chosen as our outcome variables typically do not follow a normal distribution, and Mann-Whitney tests make few distributional assumptions. Fisher exact tests are preferred for contingency tables as they provide an exact test, especially because smaller sample sizes can lead to low expected cell counts [50].

Results

Overview

Concerning the first aim, 67 sex-specific Wikipedia articles were extracted and are displayed in the Multimedia Appendix 1. For each article, the table includes information on the 13 items: rank by number of page views, title of the page, number of page views that month, WikiProject Medicine's quality assessment, WikiProject Medicine's assessment of importance to medicine, classification, bytes of text, unique references, bytes of text per reference, number of pictures, number of sections, total edits, and number of page watchers (Wikipedia editors who have the page on their watchlist).

Breakdown of Health Articles

Of the 67 sex-specific medical articles on Wikipedia, 17 (25%) concerned male individuals or predominantly male individuals (exclusively male: n=9, 53%; predominantly male: n=8, 47%), while 50 (75%) concerned female individuals or predominantly female individuals (exclusively female: n=37, 74%; predominantly female: n=13, 26%). The remaining 93.3% (933/1000) of health and medical articles were either unrelated to diseases and medications or were classified as neutral or not sex specific. For quality assessments, male and female articles were grouped to include the predominant and exclusive categories (ie, 1 and 2 vs 4 and 5). When examining Wikipedia articles by type of article, we recognized that 8% (80/1000) of the articles were about political events, people, or medical terminology (eg, the black death, the scientific method, and Theranos) that are generally not covered in medical textbooks.

Assessment of Wikipedia Articles or Their Article Status

Overview

We looked at 1000 out of 50,000 health-related Wikipedia articles which represent 2% of all health content on the English-language Wikipedia, however, these articles made up 43% of total page views (65/155 million) in June 2019. Of the 67 sex-specific articles, the mean average daily views for all sex-specific health articles were 1968 (SD 975.1) views per day, and the mean of total views was 58,877.9 (SD 29,251). Of the 67 sex-specific articles, 2 (3%) were classified as FA status, 3 (5%) as GA, 39 (58%) as B-class, 13 (19%) as C-class, and 7 (10%) as "start." In terms of article rating, slightly more female articles than male articles had FA status (0/67, 0% male articles; 2/67, 4% female articles) and a B-class status (7/67, 41% male articles; 32/67, 64% female articles); however, the number of male articles within GA status (1/67, 6% male articles; 2/67, 4% female articles), C-class (6/67, 35% male articles; 10/67, 20% female articles), and "start" (3/67, 18% male articles; 4/67, 8% female articles) was slightly higher than that for female articles (Table 1). This was not a statistically significant difference (Fisher exact P=.29). FA was only found for exclusively female articles, which included "female genital mutilation" and "menstrual cycle." The GA class in exclusively male or predominantly male Wikipedia articles included "circumcision," whereas for the exclusively female and predominantly female articles, GA articles included "birth control" and "urinary tract infection." The start category included "priapism," "minoxidil," and "penis enlargement" for exclusively male or predominantly male articles and "gynecology," "Braxton Hicks contractions," "menstrual cup," and "triple X syndrome" for exclusively female or predominantly female articles.

Wikipedia Importance Rating

Rank place" is the order of the article based on the number of visits and reads per month, with higher viewed articles being higher on the list. The sex-specific articles ranged from rank place 30 to rank place 929 on the list of the 1000 most visited medical articles on Wikipedia. They do not differ significantly by rank (Mann-Whitney test, Z=0.25; P=.81).

Proportionally, a higher number of female health articles were classified as top priority (1/67, 6% male articles; 7/67, 14% female articles) and high priority (3/67,18% male articles; 23/67, 46% female articles) but not mid (10/67, 59% male articles; 17/67, 34% female articles) and low priority (3/67, 18% male articles; 3/67, 6% female articles; Table 2). This result was borderline for statistical significance (Fisher exact P=.05).

Wikipedia Total References

The number of total references for the whole sample was 10,968 references. For female category articles, the total number of references was 8788, whereas for male category articles, this was 1918. This amounts to a mean of 175.8 references per article for the exclusively female or predominantly female category and 128.2 references per article for the exclusively male or predominantly male category. This difference is not statistically significant (Z=1.7; P=.08).

Wikipedia Unique References

The number of unique references for the whole sample was 6715. For female category articles, this was 5308, and for male category articles, unique references totaled 1407. This amounts to a mean of 106.2 unique references per article for the exclusively female or predominantly female category and 82.7 unique references per article for exclusively male or predominantly male categories. This is not a statistically significant difference (Z=1.4; P=.18).

Wikipedia Number of Edits

The total edit number for all 67 sex-specific Wikipedia health articles was 188,541, of which 130,522 (69.2%) were for female category articles and 58,019 (30.7%) were for male category articles. This amounts to a mean of 2610 for female-category

XSL•FO
RenderX

articles and 3413 for male-category articles. This is not a statistically significant difference (Z=1.4; P=.18).

Wikipedia Watchers of Health Articles

The total number of page watchers for all Wikipedia health articles was 16,426. For female category articles, the total number of page watchers was 11,692, whereas for male category articles, the total number of page watchers was 4735. This is a mean of 234 watchers per female category pages and 279 watchers for male category pages. This is not a statistically significant difference (Z=1.1; P=.27).

Wikipedia Article Pictures

The total number of pictures in all Wikipedia health articles was 799. For female category articles, the total number of pictures was 613, whereas for male category articles, the total number of pictures was 186. This average is 12.3 pictures per article for female category articles and 10.9 pictures per article for male category articles. This is not a statistically significant difference (Z=.34; P=.73).

Language Count

We have assessed the differences regarding the number of languages in which sex-specific Wikipedia articles were available. The Mann-Whitney test showed no statistical significance between the sex-specific article categories (P=.36). Language count was slightly higher for female category articles (mean 55.9 languages) than male category articles (mean 48.4 languages).

Breakdown of the 2 Medical Textbooks

The comparator table (Table 3) includes the breakdown of typical or general coverage of the 2 medical textbooks as comparators to Wikipedia. The coverage was similar. For the *Oxford Handbook of Clinical Medicine* [47], the coverage was 1.2% (7/560) exclusively female, 4.2% (24/560) predominantly female, 93.4% (522/560) neutral or not sex specific, 0.7% (4/560) predominantly male, and 0.5% (3/560) exclusively male. For *Toronto Notes Medical Council of Canada Qualifying Examination and United States Medical License Examination II* [48], the coverage was 6.9% (139/2024) exclusively female, 1.9% (39/2024) predominantly female, 89.3% (1808/2024) neural or not sex specific, 0.8% (16/2024) predominantly male, and 1.1%% (22/2024) exclusively male. There was a statistically significant difference when comparing Wikipedia to the combined counts across both textbooks (Fisher exact P=.03) with Wikipedia having a lower number of female health articles but a higher number of neutral articles than the 2 medical textbooks. However, the absolute differences were not large. Table 4 shows a comparison of final results by each of the 3 sources (Wikipedia, book 1, and book 2), with an additional row for combined book 1 and book 2, for those who may be interested in comparing Wikipedia to the 2 textbooks together.

XSL•FO
RenderX

Table 3. Breakdown of sex-specific topics for the 2 medical textbooks as comparators to Wikipedia health articles' coverage.

Books and topics listed within the book	Number of topics[a]				
	1	2	3	4	5
Oxford Handbook of Clinical Medicine **[47] (n=560)**					
Thinking about medicine	1	0	16	0	0
History and examination	0	0	23	0	0
Cardiovascular medicine	0	0	30	0	0
Chest medicine	0	0	22	0	0
Endocrinology	2	0	23	1	1
Gastroenterology	0	2	34	1	0
Renal medicine	1	1	12	0	0
Hematology	0	3	30	0	0
Infectious diseases	0	0	33	0	0
Neurology	0	2	38	1	0
Oncology and palliative care	1	11	6	1	0
Rheumatology	0	1	16	0	0
Surgery	1	3	52	0	2
Clinical chemistry	1	1	24	0	0
Eponymous syndromes	0	0	90	0	0
Radiology	0	0	16	0	0
Reference intervals, etc	0	0	5	0	0
Practical procedures	0	0	10	0	0
Emergencies	0	0	42	0	0
Total, n (%)	7 (1.2)	24 (4.2)	522 (93.4)	4 (0.7)	3 (0.5)
Toronto Notes: Comprehensive Medical Reference and Review for Medical Council of Canada Qualifying Examination and US Medical License Examination II **[48] (n=2024)**					
Ethical, legal, and organizational medicine	0	1	20	0	0
Anesthesia and perioperative medicine	0	0	58	0	0
Cardiology and cardiac surgery	0	1	62	1	0
Clinical pharmacology	0	0	24	0	0
Dermatology	0	0	77	0	0
Emergency medicine	0	1	71	0	0
Endocrinology	1	3	70	2	0
Family medicine	0	5	59	1	6
Gastroenterology	1	4	86	1	0
General surgery and thoracic surgery	0	2	105	2	0
Geriatric medicine	0	4	20	3	0
Gynecology	53	0	1	0	0
Hematology	0	1	80	0	0
Infectious diseases	0	0	90	0	0
Medical genetics	0	1	9	0	0
Medical imaging	0	0	52	0	0
Nephrology	0	0	55	0	0
Neurology	0	1	99	0	0

XSL•FO
RenderX

Books and topics listed within the book	Number of topics[a]				
	1	2	3	4	5
Neurosurgery	0	0	65	1	0
Obstetrics	82	1	0	0	0
Ophthalmology	0	1	107	0	0
Orthopedic surgery	0	0	96	0	0
Otolaryngology	0	4	87	0	0
Pediatrics	0	2	108	1	0
Plastic surgery	0	1	58	0	0
Psychiatry	0	0	87	0	0
Public health and preventive medicine	1	2	44	0	0
Respirology	1	1	47	0	0
Rheumatology	0	2	37	0	0
Urology	0	0	23	4	16
Vascular surgery	0	1	11	0	0
Total, n (%)	139 (6.9)	39 (1.9)	1808 (89.3)	16 (0.8)	22 (1.1)

[a]Classification category: 1="exclusively female," 2="predominantly female but can also affect male individuals," 3="not sex specific or neutral," 4=predominantly male but can affect female individuals," and 5="exclusively male."

Table 4. Comparator table for topics for all 3 sources (Wikipedia and the 2 medical textbooks).

Source	Total of all articles, n (%)	Sex-specific total, n (%)	Number of topics per classification[a], n (%)				
			1	2	3	4	5
Wikipedia	1000 (100)	67 (6.7)	37 (3.7)	13 (1.3)	933 (93.3)	8 (0.8)	9 (0.9)
Book 1: *Oxford Handbook* [47]	560 (100)	28 (5)	7 (1.2)	24 (4.2)	522 (93.4)	4 (0.7)	3 (0.5)
Book 2: *Toronto Notes* [48]	2024 (100)	216 (10.7)	139 (6.9)	39 (1.9)	1808 (89.3)	16 (0.8)	22 (1.1)
Book 1 and book 2 combined	2584 (100)	244 (9.4)	146 (5.6)	63 (2.4)	2330 (90.2)	20 (0.8)	25 (1)
Total for all 3	3584 (100)	311 (8.6)	183 (5.1)	76 (2.1)	3263 (91.1)	28 (0.8)	34 (0.9)

[a]Classification category: 1="exclusively female," 2="predominantly female but can also affect male individuals," 3="not sex specific or neutral," 4=predominantly male but can affect female individuals," and 5="exclusively male."

Discussion

Principal Findings

In the top 1000 health Wikipedia articles in our selected timeframe (up to June 2019), there are almost 3 times as many exclusively female or predominantly female articles as exclusively male or predominantly male health articles. However, many of the articles contain sex ratios such as 4:1 and 5:1 (eg, "human papillomavirus" and "anabolic steroid") as they were classified in the predominantly categories. This study uncovered that exclusively female and predominantly female articles rank higher on the priority list on Wikipedia, which could be explained by the inclusion of topics related to childbirth and gynecology. There are many more topics for the female reproductive system than for the male reproductive system. This is just the nature of biology, with the menstrual cycle, pregnancy, maternity, gynecology, and perinatal concerns alone having many more subcategories than those concerning the male reproductive system. Wikipedia health content was similar to the 2 medical textbooks, but there was a statistically significant difference, with the 2 medical textbooks having a higher number of female articles but fewer neutral articles. It is important to note that the 2 medical textbooks do not include the same degree of topics about psychology, people in medicine, or political events related to health. When examining Wikipedia articles by type of article, we recognized that 8% (80/1000) were about political events or people.

Our analysis has shown that exclusively female or predominantly female articles in this sample contained more references and more unique references than exclusively male or predominantly male health articles but with no statistically significant difference. Both categories of articles had a comparable number of pictures per article. Slightly more edits, on average, were made to exclusively male and predominantly male articles compared to the female topic categories. Both sets of articles include a single controversial article that attracted a much higher number of edits: "circumcision" and "abortion." These 2 articles also have much higher numbers of page

XSL•FO
RenderX

watchers. However, with or without these 2 articles, the number of edits and page watchers was not statistically significantly different. The controversial nature of a topic affects the number of edits to an article, making it an unreliable marker of quality. This cannot be used as a measure against a bias toward either sex but as an indication of the controversy within the subject matter.

One of the other important facts is that we did not assess readership, as it would be difficult to assess the sex of the readers, and therefore, we cannot draw any conclusions regarding the sexes of readers of specific health articles. It is possible that readership is equal between the sexes or that it varies depending on different countries, cultures, or times of the year (eg, medical school examinations).

Comparison With Prior Work

Differences between the coverage and quality of content on Wikipedia are not new, and certainly, there are particular gaps that begin with the fact that only about 16% of Wikipedia contributors are women [17,18]. Wikipedia includes a list of all up-to-date research on the subject of sex, gender, gender gap, and women on Wikimedia projects by individuals or organizations associated with the Wikimedia Foundation (WWC2023) and external researchers [51]. The list on Gender Bias On Wikipedia categorizes types of gender bias into "content," "participation," and "consumption," and includes research spanning from 2011 to 2023. This list currently includes 42 research projects for "content," 17 items for "participation," and 3 items for "consumption." This study will likely fit under the "content" gender bias type. Of these 42 studies, only 1 (2%; on measuring what matters in gender diversity) mentions health content on Wikipedia [52]. This study cautions us on using reader attention as a proxy for the importance of the content, given the consistent gender gap in readership and self-focus bias (eg, men reading more biographies of men compared to women), further recognizing that health articles are not biographies and have clear gender dimensions [52].

A study in 2020 on metrics for quantifying the gender content gap on Wikipedia has shown that in some instances, such as in health care, gender division is important because it allows Wikipedia to organize efforts to improve the gender gap [51]. We found that on Wikipedia female and male sex categories were comparable in quality, but not in spread. Female category articles were more numerous and also had a higher (albeit nonstatistically significant) number of references and unique references and a higher number of articles in FA status. Wikipedia also contained less content on female health than the 2 medical textbooks. Our results are somewhat in line with a study conducted in 2015 that compared the mean length of the articles on men and women for English-language Wikipedia, in which articles on women were longer (statistically significant) than those on men [30]. A study in 2016 also found that articles on female CEOs had more references, which were more diverse [41]. This contrasts with the findings of Gray [36] and Konieczny and Klein [53], who, across 25 different language versions of Wikipedia, found that articles about women were consistently 10% shorter on average than those about men.

Although our articles ranked similarly on quality for male-specific and female-specific topics, this does not mean there is no room for improvement. For example, all articles, particularly those in health care and medicine, have the potential to reach FA status by improving on several quality dimensions. However, we should remain mindful that the completeness of content may reflect global social or demographic aspects. Articles on certain topics could be needed more because of sociodemographic reasons and not because of favorability or exclusion of one sex over the other.

Our study should also not downplay the fact that there could still be other types of inequalities within the content that we could not assess (eg, language structure). For example, a 2019 study found that women in physics, economics, and philosophy are considerably less likely than men to be recognized on Wikipedia across all levels of achievement [54,55]. Differences have been observed in metadata, language, and network structure when comparing male and female Wikipedia biographies, and these differences have been attributed to gender bias [30]. We did not assess the representativeness or achievements of either sex or gender, but this could potentially add a dimension to our analysis in the future.

Implications of Our Results for the Real World

To answer the question of whether there is a difference between male and female articles on Wikipedia in terms of quality and coverage, we can say there are several differences; however, we must also consider the categorization and the basics of how and when a Wikipedia article is created and why. For example, WikiProject Medicine volunteers who work through a priorities list of missing articles, student initiatives to fill gaps on Wikipedia, individuals who edit Wikipedia with a particular interest in a field, professionals who work in their fields and volunteer on Wikipedia as editors or translators, and others, have determined, based on population's needs, that these articles were most important. Wikipedia may, therefore, be a mirror of what is required by the general population worldwide rather than a decision of select groups of people or an inherent view. A study conducted in 2021 also found similar results when examining a selected set of professions in terms of gender distribution. It concluded that Wikidata was no more biased than the real world, with men and women being included in similar percentages [56]. Our study has shown that Wikipedia spread of health-related and medical content for female and male categories was similar to the 2 widely used medical textbooks, but with some relative underrepresentation of female centric topics.

More female individuals than males individuals may be seeking health information online. Although the results of this study may be misinterpreted to be more female centric because, it appears that in our sample, there are simply more conditions that have been classified as predominantly female, such as "labiaplasty," or those exclusively female, such as the "triple X syndrome," all of which can be due to societal trends or even a rise in medical students who require this information, regardless of their sex. Our sample included conditions such as lung cancer, which accounts for 42% of total deaths in women and 58% of deaths in men in the United States [57,58]. Because

XSL•FO
RenderX

of the small difference, this was classified as not sex specific, but on a population level, it is more tilted to male individuals. Therefore, our data must not be viewed in a clear-cut male-female split but contextually.

In the United States, there are 21,570 obstetricians and gynecologists (ob-gyns; 85.2% female individuals; 14.8% male individuals) [59] and 13,780 urologists (69.5% female individuals; 30.5% male individuals) [60], whereby urology treats both sexes [61]. However, it is fair to say that (in the United States in this case) every woman will require at least 1 ob-gyn at some point in her life, but the needs for a urologist are more individual. Ob-gyns specialize in a wide spectrum of situations, which include fertility and hormone disorders, which alone are more numerous than those in urology.

There are several major implications of our research: (1) this study indicates that there may be a difference but not a large disparity between male and female health articles on English-language Wikipedia; (2) this study could be repeated for Wikipedia in other languages, and a comparative analysis could be performed; (3) health information–seeking behavior may be the driving force for the topmost viewed 1000 Wikipedia articles; (4) these data should be interpreted contextually and not be viewed as a simplistic female-male split, especially for non–sex-specific articles in which there may still be disparities in presentation, and understanding the nuances at the population-level data is essential for accurate interpretation; and (5) some health care specializations may be more represented than others on the English-language Wikipedia within the top 1000 most-read health articles, but a wider sample of articles is needed to draw any conclusions.

Limitations

There is no direct prior work on sex coverage within health Wikipedia articles. We only analyzed a subset of the 1000 most-viewed Wikipedia articles about medicine and health in a snapshot (June 2019) rather than all 50,000 medical articles. This limits the power of our results. The number of male-specific articles within these 1000 articles is low, which is an important finding, but also limits what conclusions can be drawn given the sample size. Non-significant results on some comparisons may reflect insufficient power rather than the absence of any differences. Future research could be conducted using data from the most recent month to compare results and with a larger sample of Wikipedia articles. When viewing the top 1000 articles for January 2023, it was clear that several articles included in our sample had changed their status. For example, the status of many articles has changed from start to class C, and several have risen or fallen in importance or number of views, which can be partly explained by the changes in the COVID-19 pandemic and people's lifestyles. In addition, our analysis was performed before the COVID-19 pandemic; therefore, it is likely that the coverage of some topics was different from the time after the COVID-19 pandemic.

We have used a methodology previously used in other studies involving Wikipedia articles (eg, [36,48]). One study showed that the quality of Wikipedia articles is affected by the editor's previous number of edited articles, but not the volume of edits [40]. We could not assess these metrics, which may be interesting to correlate in future studies. Unlike research conducted in 2017 [32], we did not assess the articles using supervised machine learning and, therefore, could not consider metrics that may have introduced other types of biases among contributors and the content itself, such as article and author perspectives, author characteristics, and balance of nonneutral edits. These measurements might have uncovered supplementary information, indicating that both an author's viewpoints and the article's prior perspectives played a role in predicting the stance of the resulting edits, with the caveat that these predictors also influence each other [32].

We may not be able to answer the question of whether Wikipedia is biased, but more if the society is biased. For example, with some pages such as monarchs, you would find many more male individuals covered than female individuals, but that is because female individuals were not allowed to have power or these positions. This raises the following question: Is Wikipedia merely reflecting societal trends, and is the Wikipedia community adequately covering what is needed? We could also not assess the degree of risk particular medical conditions pose and the need or urgency of the information for these. Therefore, the strength and limitation of this study is that it captures Wikipedia activity at a specific point in time, but the results are fluid and mirror societal needs. Our analysis identified some gaps and room for improvement in Wikipedia health articles.

Biographies of living persons may not be a fair representation of issues within the core community of Wikipedians (people who edit or contribute to Wikipedia) prioritizing the creation of articles for one sex over the other. Many biographies of living persons end up within Wikipedia because of undisclosed paid editing. An analysis of now-deleted articles created by a collection of accounts from various paid editing companies found that most of the created articles were about male individuals. This leads to the assumption that male individuals and their representatives are more likely to pay money to try to buy an article on Wikipedia than female individuals and their representatives.

Conclusions

This was the first study to examine the coverage of medical and health content on Wikipedia related to biological sex. In our sample, there were more female health topics than male health topics in the top 1000 most-read health Wikipedia pages in June 2019, but there are relatively fewer female health topics covered than what is seen in medical textbooks. Despite the higher volume of articles on female health topics, the average quality based on various metrics was similar. It is vital to recognize that the needs of people are individual and that both sexes may require or access any of Wikipedia articles. Future endeavors could include prioritizing content to reach GA and FA status as soon as possible. Future research should strive to identify where there may be sex-based differences in needs for health information and whether behavior around health information seeking on Wikipedia differs among the sexes.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Wikipedians for all their efforts and contributions to WikiProject Medicine. The research team would like to thank Dr Jennifer Dawson from CHEO Research Institute in Ottawa, Canada, for helping with Wikipedia health articles screening; Netha Hussain for extracting language count; and Aljoša Vertot, a software engineering and programming student (Maribor, Slovenia), for helping with data cleaning.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JMH conceived the study and collected data. NF and HWWP analyzed the data. NF wrote the manuscript, but all authors contributed to various drafts. No generative artificial intelligence tools were used to generate any portion of the manuscript.

Conflicts of Interest

HWWP and JMH are long-term Wikipedia editors. JMH was previously on the board of the Wikimedia Foundation, the not-for-profit that hosts the site, and is currently the chair of Wiki Project Med Foundation, a movement thematic organization that promotes the development of open-source health care content, including that on Wikipedia.

Multimedia Appendix 1
A list of 67 sex-specific medical Wikipedia articles and their associated data from views (January 2019) from a list of 1000 most-viewed health Wikipedia articles.
[DOCX File , 69 KB - jmir_v26i1e47562_app1.docx]

References

1. Wikipedia: size of Wikipedia. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia [accessed 2024-05-30]
2. List of Wikipedias. Wikipedia. URL: https://en.wikipedia.org/wiki/List_of_Wikipedias [accessed 2023-03-19]
3. List of most-visited websites. Wikipedia. URL: https://en.wikipedia.org/wiki/List_of_most-visited_websites [accessed 2023-03-19]
4. McDowell ZJ, Vetter MA. The realienation of the commons: Wikidata and the ethics of "free" data. Int J Commun 2024 Jan;18:590-608.
5. Wikipedia:WikiProject medicine/popular pages. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages [accessed 2023-03-19]
6. Shafee T, Masukume G, Kipersztok L, Das D, Häggström M, Heilman J. Evolution of Wikipedia's medical content: past, present and future. J Epidemiol Community Health 2017 Nov;71(11):1122-1129 [FREE Full text] [doi: 10.1136/jech-2016-208601] [Medline: 28847845]
7. Farič N, Potts HW. Motivations for contributing to health-related articles on Wikipedia: an interview study. J Med Internet Res 2014 Dec 03;16(12):e260 [FREE Full text] [doi: 10.2196/jmir.3569] [Medline: 25498308]
8. Wikipedia: content assessment. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:Content_assessment [accessed 2023-03-19]
9. Item quality evaluator. Toolhub. URL: https://toolhub.wikimedia.org/tools/toolforge-item-quality-evaluator [accessed 2024-08-13]
10. WikiProject medicine. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine [accessed 2023-03-19]
11. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
12. Osei Asibey B, Agyemang S, Boakye Dankwah A. The internet use for health information seeking among Ghanaian university students: a cross-sectional study. Int J Telemed Appl 2017;2017:1756473 [FREE Full text] [doi: 10.1155/2017/1756473] [Medline: 29225620]
13. Lagoe C, Atkin D. Health anxiety in the digital age: an exploration of psychological determinants of online health information seeking. Comput Hum Behav 2015 Nov;52:484-491. [doi: 10.1016/j.chb.2015.06.003]
14. Anker AE, Reinhart AM, Feeley TH. Health information seeking: a review of measures and methods. Patient Educ Couns 2011 Mar;82(3):346-354. [doi: 10.1016/j.pec.2010.12.008] [Medline: 21239134]
15. Covolo L, Guana M, Bonaccorsi G, Brunelli L, Castaldi S, De Donno A, et al. Exploring the online health information-seeking behavior in a sample of Italian women: the "SEI Donna" study. Int J Environ Res Public Health 2022 Apr 14;19(8):4745 [FREE Full text] [doi: 10.3390/ijerph19084745] [Medline: 35457610]

XSL•FO
RenderX

16. Bromme R, Thomm E, Wolf V. From understanding to deference: laypersons' and medical students' views on conflicts within medicine. Int J Sci Educ Part B 2013 Nov 07;5(1):68-91. [doi: 10.1080/21548455.2013.849017]
17. Wikipedia: gender bias and editing on Wikipedia. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:Gender_bias_and_editing_on_Wikipedia [accessed 2023-03-19]
18. Research: Wikipedia editors survey 2011 April. Wikimedia. URL: https://meta.wikimedia.org/wiki/Research:Wikipedia_Editors_Survey_2011_April [accessed 2024-07-23]
19. Lam SK, Uduwage A, Dong Z, Sen S, Musicant DR, Terveen L, et al. WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In: Proceedings of the 7th International Symposium on Wikis and Open Collaboration. 2011 Presented at: WikiSym '11; October 3-5, 2011; Mountain View, CA. [doi: 10.1145/2038558.2038560]
20. Antin J, Yee R, Cheshire C, Nov O. Gender differences in Wikipedia editing. In: Proceedings of the 7th International Symposium on Wikis and Open Collaboration. 2011 Presented at: WikiSym '11; October 3-5, 2011; Mountain View, CA. [doi: 10.1145/2038558.2038561]
21. Cohen N. Wikipedia ponders its gender-skewed contributions. The New York Times. 2011. URL: https://scholar.google.com/scholar?q=Wikipedia+ponders+its+gender-skewed+contributions+Cohen+2011 [accessed 2024-07-23]
22. Murciano-Goroff R. Missing women in tech: the labor market for highly skilled software engineers. Manag Sci 2022 May 01;68(5):3262-3281. [doi: 10.1287/mnsc.2021.4077]
23. Hill BM, Shaw A. The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. PLoS One 2013 Jun 26;8(6):e65782 [FREE Full text] [doi: 10.1371/journal.pone.0065782] [Medline: 23840366]
24. Gardner S. Nine reasons women don't edit Wikipedia (in their own words). WordPress. 2011 Feb 19. URL: https://suegardner.org/2011/02/19/nine-reasons-why-women-dont-edit-wikipedia-in-their-own-words/#:~:text=3)%20Some%20women%20don't,a%20lot%20of%20self%2Dconfidence.&text=4)%20Some%20women%20don't,to%20suggest%20this%20is%20true [accessed 2023-03-19]
25. Young AG, Wigdor AD, Kane GC. The gender bias tug-of-war in a co-creation community: core-periphery tension on Wikipedia. J Manag Inf Syst 2020 Dec 01;37(4):1047-1072. [doi: 10.1080/07421222.2020.1831773]
26. Jirschitzka J, Kimmerle J, Halatchliyski I, Hancke J, Meurers D, Cress U. A productive clash of perspectives? The interplay between articles' and authors' perspectives and their impact on Wikipedia edits in a controversial domain. PLoS One 2017 Jun 2;12(6):e0178985 [FREE Full text] [doi: 10.1371/journal.pone.0178985] [Medline: 28575077]
27. Vasilescu B, Capiluppi A, Serebrenik A. Gender, representation and online participation: a quantitative study. Interact Comput 2013 Sep 27;26(5):488-511. [doi: 10.1093/iwc/iwt047]
28. Mac Donald H. Wikipedia is male-dominated. That doesn't mean it's sexist. The Slate Group. 2011 Feb 9. URL: https://slate.com/human-interest/2011/02/wikipedia-sexist-how-the-new-york-times-misinterprets-the-site-s-gender-gap.html [accessed 2024-06-04]
29. Paling E. Wikipedia's hostility to women. The Atlantic. 2015 Oct 21. URL: https://www.theatlantic.com/technology/archive/2015/10/how-wikipedia-is-hostile-to-women/411619/ [accessed 2024-08-13]
30. Graells-Garrido E, Lalmas M, Menczer F. First women, second sex: gender bias in Wikipedia. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media. 2015 Presented at: HT '15; September 1-4, 2015; Güzelyurt, Northern Cyprus. [doi: 10.1145/2700171.2791036]
31. Tripodi F. Ms. categorized: gender, notability, and inequality on Wikipedia. New Media Soc 2021 Jun 27;25(7):1687-1707. [doi: 10.1177/14614448211023772]
32. Menking A, Erickson I. The heart work of Wikipedia: gendered, emotional labor in the world's largest online encyclopedia. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15; April 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702514]
33. Reagle J, Rhue L. Gender bias in Wikipedia and Britannica. Int J Commun 2011 Jan;5:1138-1158.
34. Wagner C, Garcia D, Jadidi M, Strohmaier M. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;9(1):454-463. [doi: 10.1609/icwsm.v9i1.14628]
35. Wikipedia: writing about women. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:Writing_about_women [accessed 2023-03-19]
36. Gray A. Gender and deletion on Wikipedia. Generalist. URL: https://www.generalist.org.uk/blog/2019/gender-and-deletion-on-wikipedia/ [accessed 2024-01-28]
37. Adams J, Brückner H, Naslund C. Who counts as a notable sociologist on Wikipedia? Gender, race, and the "professor test". Socius 2019 Feb 15;5. [doi: 10.1177/2378023118823946]
38. Venus N. The representation of female economists on Wikipedia. SSRN J 2024 Feb 18. [doi: 10.2139/ssrn.4540744]
39. Halfaker A. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In: Proceedings of the 13th International Symposium on Open Collaboration. 2017 Presented at: OpenSym '17; August 23-25, 2017; Galway, Ireland. [doi: 10.1145/3125433.3125475]
40. Vitulli MA. Writing women in mathematics into Wikipedia. Notices Amer Math Soc 2018 Mar;65(03):330-334. [doi: 10.1090/noti1650]

XSL•FO
RenderX

41. Young A, Wigdor AD, Kane G. It's not what you think: gender bias in information about Fortune 1000 CEOs on Wikipedia. In: Proceedings of the 37th International Conference on Information System. 2016 Presented at: ICIS 2016; December 11-14, 2016; Dublin, Ireland.
42. Rodriguez L. Only 20% of Wikipedia's biographies are about women — #Wikigap wants to change that. Global Citizen. 2021 Mar 09. URL: https://www.globalcitizen.org/en/content/wikimedia-wikigap-international-womens-day/ [accessed 2024-01-28]
43. Langrock I, González-Bailón S. The gender divide in Wikipedia: quantifying and assessing the impact of two feminist interventions. J Commun 2022 Feb 16;72(3):297-321. [doi: 10.1093/joc/jqac004]
44. Wikipedia: WikiProject women in red. Wikipedia. URL: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red [accessed 2024-01-28]
45. Abancens M, Bustos V, Harvey H, McBryan J, Harvey BJ. Sexual dimorphism in colon cancer. Front Oncol 2020 Dec 9;10:607909 [FREE Full text] [doi: 10.3389/fonc.2020.607909] [Medline: 33363037]
46. Holtz P, Fetahu B, Kimmerle J. Effects of contributor experience on the quality of health-related Wikipedia articles. J Med Internet Res 2018 May 10;20(5):e171 [FREE Full text] [doi: 10.2196/jmir.9683] [Medline: 29748161]
47. Wilkinson IB, Raine T, Wiles K, Goodhart A, Hall C, O'Neill H. Oxford Handbook of Clinical Medicine, Tenth Edition. Oxford, UK: Oxford University Press; 2017.
48. Toronto Notes 2020: Comprehensive Medical Reference and Review for the Medical Council of Canada Qualifying Exam (MCCQE) Part I and the United States Medical Licensing Exam (USMLE) Step II. Toronto, ON: Toronto Notes for Medical Students, Incorporated; 2020.
49. Burcharth J, Pedersen M, Bisgaard T, Pedersen C, Rosenberg J. Nationwide prevalence of groin hernia repair. PLoS One 2013 Jan 14;8(1):e54367 [FREE Full text] [doi: 10.1371/journal.pone.0054367] [Medline: 23342139]
50. McDonald JH. Handbook of Biological Statistics, Third Edition. Baltimore, MD: Sparky House; 2014.
51. Research: metrics for quantifying the gender content gap. Wikimedia. URL: https://meta.wikimedia.org/wiki/Research:Metrics_for_quantifying_the_gender_content_gap [accessed 2024-01-28]
52. WWC2023/Research. Wikimedia. URL: https://meta.wikimedia.org/wiki/WWC2023/Research [accessed 2024-01-28]
53. Konieczny P, Klein M. Gender gap through time and space: a journey through Wikipedia biographies via the Wikidata Human Gender Indicator. New Media Soc 2018 Jun 18;20(12):4608-4633. [doi: 10.1177/1461444818779080]
54. Shellekens MH, Holstege F, Yasseri T. Female scholars need to achieve more for equal public recognition. arXiv Preprint posted online April 12, 2019 [FREE Full text] [doi: 10.48550/arXiv.1904.06310]
55. Ross MB, Glennon BM, Murciano-Goroff R, Berkes EG, Weinberg BA, Lane JI. Women are credited less in science than men. Nature 2022 Aug 22;608(7921):135-145 [FREE Full text] [doi: 10.1038/s41586-022-04966-w] [Medline: 35732238]
56. Abián D, Meroño-Peñuela A, Simperl E. An analysis of content gaps versus user needs in the Wikidata knowledge graph. In: Proceedings of the 21st International Semantic Web Conference. 2022 Presented at: ISWC 2022; October 23-27, 2022; Virtual Event. [doi: 10.1007/978-3-031-19433-7_21]
57. Roser M, Ritchie H. Cancer. Our World in Data. URL: https://ourworldindata.org/cancer [accessed 2023-03-19]
58. Factsheets. International Agency for Research on Cancer. URL: https://gco.iarc.fr/today/en [accessed 2024-08-23]
59. Ob/Gyn demographics and statistics in the US. Zippia. URL: https://www.zippia.com/ob-gyn-jobs/demographics/ [accessed 2023-03-19]
60. Urologist demographics and statistics in the US. Zippia. URL: https://www.zippia.com/urologist-jobs/demographics/ [accessed 2024-08-13]
61. Obstetricians and gynecologists. U.S. Bureau of Labour Statistics. URL: https://bls.gov/oes/2019/may/oes291218.htm [accessed 2023-03-19]

Abbreviations

CEO: chief executive officer
FA: featured article
Ob-gyn: obstetrics and gynecology
WP: Wikipedia

XSL•FO
RenderX

Edited by A Mavragani; submitted 24.03.23; peer-reviewed by J Kimmerle, DTC Uhawenimana; comments to author 29.06.23; revised version received 31.08.23; accepted 11.06.24; published 12.09.24.

Please cite as:
Farič N, Potts HWW, Heilman JM
Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis
J Med Internet Res 2024;26:e47562
URL: https://www.jmir.org/2024/1/e47562
doi:10.2196/47562
PMID:

©Nuša Farič, Henry WW Potts, James M Heilman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Added Value of Medical Subject Headings Terms in Search Strategies of Systematic Reviews: Comparative Study

Victor Leblanc[1], MD; Aghiles Hamroun[1], MD, PhD; Raphaël Bentegeac[1], MD; Bastien Le Guellec[1], MD; Rémi Lenain[1], MD; Emmanuel Chazard[1,2], MD, PhD

[1]Public Health Department, CHU Lille, Université de Lille, Lille, France

[2]ULR 2694 Metrics, CERIM, Université de Lille, Lille, France

Corresponding Author:
Victor Leblanc, MD
Public Health Department
CHU Lille
Université de Lille
42 Rue Paul Duez
Lille, 59000
France
Phone: 33 637004971
Email: leblancvictor59@gmail.com

Abstract

Background: The massive increase in the number of published scientific articles enhances knowledge but makes it more complicated to summarize results. The Medical Subject Headings (MeSH) thesaurus was created in the mid-20th century with the aim of systematizing article indexing and facilitating their retrieval. Despite the advent of search engines, few studies have questioned the relevance of the MeSH thesaurus, and none have done so systematically.

Objective: The objective of this study was to estimate the added value of using MeSH terms in PubMed queries for systematic reviews (SRs).

Methods: SRs published in 4 high-impact medical journals in general medicine over the past 10 years were selected. Only SRs for which a PubMed query was provided were included. Each query was transformed to obtain 3 versions: the original query (V1), the query with free-text terms only (V2), and the query with MeSH terms only (V3). These 3 queries were compared with each other based on their sensitivity and positive predictive values.

Results: In total, 59 SRs were included. The suppression of MeSH terms had an impact on the number of relevant articles retrieved for 24 (41%) out of 59 SRs. The median (IQR) sensitivities of queries V1 and V2 were 77.8% (62.1%-95.2%) and 71.4% (42.6%-90%), respectively. V1 queries provided an average of 2.62 additional relevant papers per SR compared with V2 queries. However, an additional 820.29 papers had to be screened. The cost of screening an additional collected paper was therefore 313.09, which was slightly more than triple the mean reading cost associated with V2 queries (88.67).

Conclusions: Our results revealed that removing MeSH terms from a query decreases sensitivity while slightly increasing the positive predictive value. Queries containing both MeSH and free-text terms yielded more relevant articles but required screening many additional papers. Despite this additional workload, MeSH terms remain indispensable for SRs.

(J Med Internet Res 2024;26:e53781) doi:10.2196/53781

KEYWORDS

Medical Subject Headings; MeSH; MeSH thesaurus; systematic review; PubMed; search strategy; comparative analysis; literature review; positive predictive value; PPV; review; scientific knowledge; medical knowledge; utility; systematic literature review

Introduction

The number of articles published in scientific and medical journals have been increasing exponentially since the late 20th century. In 2021 alone, over 1,700,000 indexed, full-text articles were included in the PubMed database. In response to the massive production of scientific knowledge, the need for access to synthetic scientific data has been driven by the emergence of evidence-based medicine [1] and the establishment of national

XSL•FO
RenderX

regulatory bodies, medical associations, and learned societies that provide guidelines on best practice.

In this context, systematic reviews (SRs; a type of analysis developed in the 1970s) are becoming more important. Given that quality of an SR depends largely on the research methodology, building search queries is a crucial part of the review process. The challenge of constructing a query for a SR lies in the absolute necessity of being as sensitive as possible, despite the fact that this query will return at most a few tens of thousands of articles among the hundreds of millions that make up the scientific literature [2].

In the mid-20th century, researchers started to develop a common vocabulary that facilitated article indexing and retrieval and helped to avoid misunderstandings [3-5]. These efforts led to the creation of the Medical Subject Headings (MeSH) thesaurus in the 1960s by the US National Library of Medicine (NLM) [6]. PubMed (the NLM's search engine), which is one of the most widely used search engines [7], heavily relies on MeSH terms to assist users in their literature searches. The MeSH thesaurus is intended to facilitate literature searches by limiting term permutations [8,9]. In other words, it assigns a unique term to a concept—regardless of the language used or the time period concerned.

Subsequent improvements in search engine performance have enabled researchers to query databases with simple free-text terms, rather than MeSH terms. Furthermore, the massive influx of publications and the emergence of many new scientific and medical topics have led to delays in MeSH indexing and difficulties in updating the thesaurus [10]. In addition, although frequently recommended [11-13], the value of using the MeSH thesaurus in queries for literature reviews has never been systematically assessed. The few studies to have tested the utility of MeSH terms in SRs have limitations, such as a small sample size or a lack of generalizability [14-20]. Finally, some studies simply compared the numbers of results retrieved for a given query but did not evaluate the results' relevance [21].

To the best of our knowledge, only 1 study has extensively explored the relevance of MeSH terms with regard to the results of SRs [22]. The study concluded that the use of queries based on free-text alone (ie, free-text terms) appeared to decrease the retrieval of articles of interest, relative to queries based on both free-text terms and MeSH terms. However, this study included SRs from a single research center, which limited the generalizability of the findings. Furthermore, the MEDLINE database was queried with the Ovid search engine, rather than PubMed. We therefore decided to evaluate this question in more detail. The objective of this work was to estimate the added value of using MeSH terms in PubMed queries for SRs.

Methods

Paper Selection

We first selected the top 6 journals in the "Medicine. General & Internal" Journal Citation Reports category, according to the impact factors computed by Clarivate [23,24]. Next, we selected all the PubMed-indexed SRs published in the 6 journals between 2012 and 2021 and for which the free full text was available on PubMed Central. The time period was chosen arbitrarily, with the objective of obtaining at least 60 SRs. The following PubMed query was used: '("The New England Journal of Medicine"[Journal] OR "Lancet London England"[Journal] OR "JAMA"[Journal] OR "Nature Reviews Disease Primers"[Journal] OR "BMJ Clinical Research Ed"[Journal] OR "Annals of Internal Medicine"[Journal]) AND "loattrfree full text"[Filter] AND 2012/01/01:2021/12/31[Date - Publication] AND systematic review[Filter]'.

The exclusion criteria were as follows: (1) articles other than an SR, (2) the absence of a published search query, (3) the use of queries in multiple parts that had to be assembled, (4) the absence of a query specifically built for PubMed, (5) a query that did not return any results, (6) a query that returned more than 100,000 results, and (7) a query with only MeSH terms or without MeSH terms. The sorting was carried out by a single researcher (VL).

Analysis of the PubMed Results

The query was extracted from each included SR and inserted into the PubMed search bar. PubMed has a feature called automatic term mapping (ATM) [25]; when terms not enclosed in quotation marks are inserted in the search bar, they are automatically transformed into a query segment that contains several descriptors, such as [MeSH terms], [tiab], and [all fields]. To ensure greater reproducibility, we checked for the automatic transformation of queries. This step was important because PubMed's ATM feature might add MeSH terms to query initially considered to be free of such terms. Hence, we always retrieved the query formatted by PubMed's ATM (henceforth referred to as V1; Figure 1).

XSL•FO
RenderX

Figure 1. Example of an automated transformation of queries. MeSH: Medical Subject Headings.

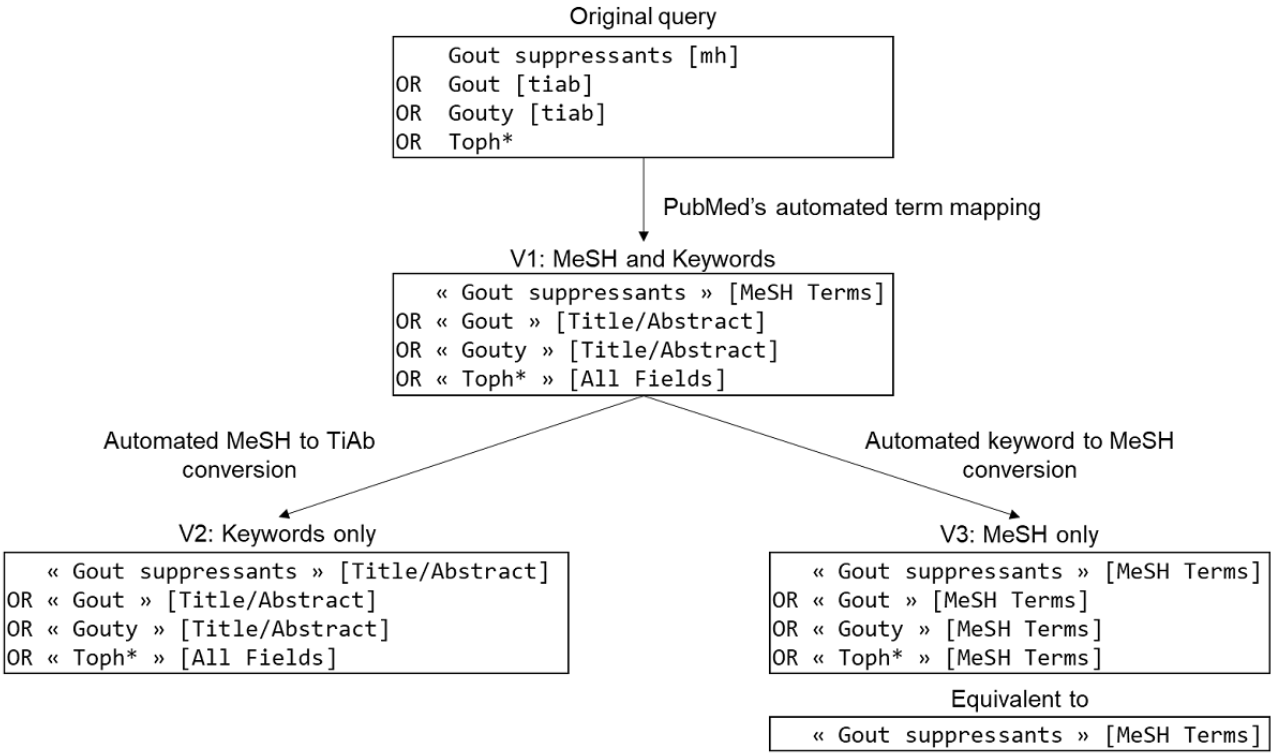

For each SR, V1 was transformed into a V2 query by replacing each MeSH term in the query with a free-text term that had to be present in the title or in the abstract. To do this, we simply replaced the [MeSH] tag with a [Title/Abstract] tag. Hence, the resulting V2 did not contain any explicit [MeSH] tags (Figure 1). Lastly, the V3 (MeSH-only) query was obtained by transforming all free-text terms into MeSH terms. It should be noted that terms stated as MeSH terms in the query but that do not actually exist in the MeSH thesaurus are ignored by the PubMed engine; this is equivalent to deleting the terms (Figure 1 [26]).

The transformations from V1 to V2 and V3 were the same for all queries, regardless of whether they contained MeSH terms only or free-text terms only. However, we noted that some PubMed filters are based on MeSH terms [27]. It would therefore not be relevant to convert these terms into free-text terms. We drew up a list of these terms so that they were not transformed and were still able to serve as filters. Those 14 terms are "80 and over," "adolescent," "adult," "aged," "animals," "child," "female," "humans," "infant," "male," "middle aged," "newborn," "preschool," and "young adult."

Hence, each SR had a query written by the SR's authors (a combination of MeSH and free-text terms; V1), a free-text-only query (V2), and a MeSH-only query (V3). Therefore, we intend to interpret the comparison of V2 with V1 as the added value of MeSH terms, and we intend to interpret the comparison of V3 with V1 as the added value of free-text terms.

Each query was submitted to the PubMed search engine, and the results were retrieved and sorted by the "Best Match" option. If there were more than 10,000 results, only the first 10,000 results were retained; in fact, PubMed does not allow more than 10,000 results to be extracted. The results were identified by their PubMed Identifier (PMID).

For each SR, the "gold standard" (GS) consisted of the articles selected by the authors of the SR. Each SR was read in order to extract the list of PMIDs selected by the authors. This work was done "by hand" by 4 researchers (VL, RB, BLG, and AH). Publications cited in the SR but not indexed in MEDLINE were not analyzed. If the reference section did not contain the items selected in the SR, data extraction from supplementary files allowed for the completion of the GS.

Data Analysis

For each SR, we obtained 4 lists of PMIDs: the GS, those retrieved by V1 (MeSH and free-text terms), those retrieved by V2 (free-text terms only), and those retrieved by V3 (MeSH terms only). For each list, we computed the sensitivity (also referred to as "recall") and the positive predictive value (PPV; also referred to as "precision") with respect to the GS. We then computed the F_1-score, which is the harmonic mean of the sensitivity and the PPV.

For each query i (V1, V2, and V3), the odds for the PPV was defined as the ratio between 2 numbers:

Next, for a given SR and using the same GS, the odds ratio (OR) of query_2 to query_1 for the PPV was defined as:

Likewise, the odds for the sensitivity of each query i (V1, V2, and V3) was defined as the ratio between 2 numbers:

XSL•FO
RenderX

Hence, for a given SR and using the same GS, the OR for $query_2$ versus $query_1$ with regard to sensitivity was:

We computed the respective ORs for V2 versus V1 and V3 versus V1 for the PPV and the sensitivity:

An OR of 1 means that the queries have the same level of performance with regard to the chosen indicator. An OR<1 denotes worse performance, and an OR>1 denotes better performance.

Statistical Analysis

Qualitative variables, binary variables, or discrete variables with very few modalities were expressed as the frequency (percentage). Quantitative variables were expressed as the mean (SD) when symmetrically distributed and the median (IQR) when not. The independence of 2 qualitative variables was probed in a chi-square test.

All statistical tests were 2-sided. The threshold for statistical significance was set to $P<.05$. The 95% CI of a proportion was calculated using the Wald method. Statistical analyses were performed with R software (R Core Team), RStudio software (Posit PBC), and the R *metafor* package [28-30].

Ethical Considerations

The research was performed using publicly available documents. It did not involve individuals or personal data. Approval by an institutional review board was not required.

Results

Flowchart

The SRs used to compile the set of queries were selected by a single researcher (VL; Figure 2).

Figure 2. Flowchart for the selection of systematic reviews. MeSH: Medical Subject Heading; SR: systematic review.

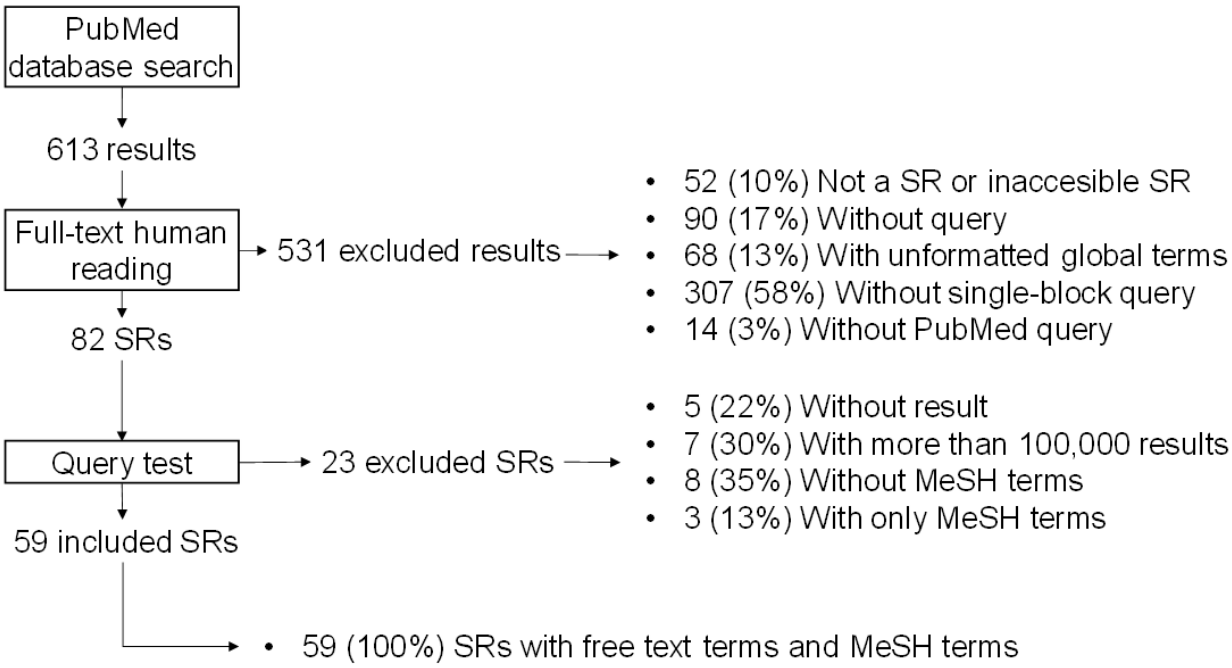

Description of the Included Systematic Reviews

A total of 59 SRs were selected for analysis, which contained both MeSH terms and free-text terms (Table 1 and Multimedia Appendices 1 and 2 [26,31-88]).

XSL•FO
RenderX

Table 1. General description of the attributes for each included systematic review.

SR's[a] PMID[b]	Items, n				Number of GS found			V1[c]			V2[d]			V3[e]		
	V1	V2	V3	GS[f]	V1	V2	V3	Se[g]	PPV[h]	*F*-Sc[i]	Se	PPV	*F*-Sc	Se	PPV	F-Sc
33472813 [46]	297	294	147	30	27	27	24	0.900	0.091	0.165	0.900	0.092	0.167	0.800	0.163	0.271
33441384 [50]	4408	2265	901	16	14	13	11	0.875	0.003	0.006	0.812	0.006	0.011	0.688	0.012	0.024
33186535 [49]	145	153	1	6	4	4	1	0.667	0.028	0.053	0.667	0.026	0.050	0.167	1.000	0.286
33148618 [40]	9024	10000	3706	66	16	18	3	0.242	0.002	0.004	0.273	0.002	0.004	0.045	0.001	0.002
32909814 [59]	349	349	40	9	1	1	0	0.111	0.003	0.006	0.111	0.003	0.006	0.000	0.000	0.000
32496521 [47]	94	162	0	24	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
32459529 [54]	1950	812	1441	9	7	2	7	0.778	0.004	0.007	0.222	0.002	0.005	0.778	0.005	0.010
32442035 [73]	10000	10000	842	15	8	7	2	0.533	0.001	0.002	0.467	0.001	0.001	0.133	0.002	0.005
32371466 [64]	4185	3320	324	50	49	48	21	0.980	0.012	0.023	0.960	0.014	0.028	0.420	0.065	0.112
32199484 [70]	6347	6111	1231	128	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31255301 [43]	1164	1023	456	61	41	35	34	0.672	0.035	0.067	0.574	0.034	0.065	0.557	0.075	0.132
30884526 [33]	770	716	30	79	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30617123 [75]	1086	788	292	29	28	26	14	0.966	0.026	0.050	0.897	0.033	0.064	0.483	0.048	0.087
30326495 [77]	4360	2212	951	158	123	117	62	0.778	0.028	0.054	0.741	0.053	0.099	0.392	0.065	0.112
30158148 [41]	10000	2673	395	45	33	24	28	0.733	0.003	0.007	0.533	0.009	0.018	0.622	0.071	0.127
29049756 [80]	3155	2858	913	20	16	16	2	0.800	0.005	0.010	0.800	0.006	0.011	0.100	0.002	0.004
28903922 [35]	687	610	95	24	23	23	16	0.958	0.033	0.065	0.958	0.038	0.073	0.667	0.168	0.269
27893131 [60]	10000	10000	2439	48	34	33	27	0.708	0.003	0.007	0.688	0.003	0.007	0.562	0.011	0.022
27802505 [74]	2299	2227	195	21	19	19	14	0.905	0.008	0.016	0.905	0.009	0.017	0.667	0.072	0.130
27802478 [26]	3847	3525	2526	89	45	43	43	0.506	0.012	0.023	0.483	0.012	0.024	0.483	0.017	0.033
27548070 [63]	1634	910	626	26	20	10	9	0.769	0.012	0.024	0.385	0.011	0.021	0.346	0.014	0.028
27142267 [78]	10000	10000	10000	10	7	9	5	0.700	0.001	0.001	0.900	0.001	0.002	0.500	0.000	0.001
26903336 [81]	903	414	115	92	47	47	0	0.511	0.052	0.094	0.511	0.114	0.186	0.000	0.000	0.000
26349907 [53]	2675	265	1728	8	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
26199070 [58]	554	303	0	20	17	15	0	0.850	0.031	0.059	0.750	0.050	0.093	0.000	0.000	0.000
26109551 [66]	298	297	14	14	9	9	5	0.643	0.030	0.058	0.643	0.030	0.058	0.357	0.357	0.357
25770113 [42]	3046	2761	1956	7	7	5	7	1.000	0.002	0.005	0.714	0.002	0.004	1.000	0.004	0.007
25569206 [39]	746	691	0	49	49	49	0	1.000	0.066	0.123	1.000	0.071	0.132	0.000	0.000	0.000
25556126 [67]	212	206	50	9	8	8	4	0.889	0.038	0.072	0.889	0.039	0.074	0.444	0.080	0.136
25006006 [52]	834	395	407	25	24	23	20	0.960	0.029	0.056	0.920	0.058	0.110	0.800	0.049	0.093
24727842 [62]	2046	1989	0	69	66	66	0	0.957	0.032	0.062	0.957	0.033	0.064	0.000	0.000	0.000
24157497 [87]	10000	81	8636	61	61	0	61	1.000	0.006	0.012	0.000	0.000	0.000	1.000	0.007	0.014
24046285 [48]	978	978	812	12	12	12	12	1.000	0.012	0.024	1.000	0.012	0.024	1.000	0.015	0.029
23935058 [69]	628	20	887	5	3	0	1	0.600	0.005	0.009	0.000	0.000	0.000	0.200	0.001	0.002
23900314 [51]	499	300	195	6	5	5	4	0.833	0.010	0.020	0.833	0.017	0.033	0.667	0.021	0.040
23529983 [65]	283	160	0	8	8	2	0	1.000	0.028	0.055	0.250	0.013	0.024	0.000	0.000	0.000
23420235 [37]	6848	2305	3434	27	25	20	10	0.926	0.004	0.007	0.741	0.009	0.017	0.370	0.003	0.006
23033409 [86]	434	258	0	16	11	11	0	0.688	0.025	0.049	0.688	0.043	0.080	0.000	0.000	0.000
22986378 [76]	8633	7169	3884	40	26	38	19	0.650	0.003	0.006	0.950	0.005	0.011	0.475	0.005	0.010
22422870 [55]	1010	980	0	4	3	3	0	0.750	0.003	0.006	0.750	0.003	0.006	0.000	0.000	0.000

XSL•FO
RenderX

SR's[a] PMID[b]	Items, n				Number of GS found			V1[c]			V2[d]			V3[e]		
	V1	V2	V3	GS[f]	V1	V2	V3	Se[g]	PPV[h]	*F*-Sc[i]	Se	PPV	*F*-Sc	Se	PPV	F-Sc
22323502 [32]	10000	10000	9784	5	3	3	2	0.600	0.000	0.001	0.600	0.000	0.001	0.400	0.000	0.000
22226047 [44]	1911	1239	1807	49	38	32	35	0.776	0.020	0.039	0.653	0.026	0.050	0.714	0.019	0.038
33176180 [84]	87	87	0	4	3	3	0	0.750	0.034	0.066	0.750	0.034	0.066	0.000	0.000	0.000
32479176 [68]	303	266	64	29	28	26	2	1.000	0.092	0.169	0.931	0.098	0.177	0.069	0.031	0.043
32427305 [57]	2035	2004	0	13	12	12	0	0.923	0.006	0.012	0.923	0.006	0.012	0.000	0.000	0.000
31727627 [56]	1884	1878	610	132	127	127	29	0.985	0.067	0.126	0.985	0.068	0.127	0.227	0.048	0.079
31585960 [36]	10000	10000	9514	227	186	206	139	0.819	0.019	0.036	0.907	0.021	0.040	0.612	0.015	0.029
30383109 [72]	4723	4716	570	38	36	36	2	0.947	0.008	0.015	0.947	0.008	0.015	0.053	0.004	0.007
28348110 [79]	27	27	0	24	1	1	0	0.042	0.037	0.039	0.042	0.037	0.039	0.000	0.000	0.000
28114600 [71]	85	35	25	68	9	3	0	0.132	0.106	0.118	0.044	0.086	0.058	0.000	0.000	0.000
26868137 [34]	10000	10000	10000	32	2	7	2	0.062	0.000	0.000	0.219	0.001	0.001	0.062	0.000	0.000
26830221 [82]	6167	1166	5102	76	74	28	72	0.974	0.012	0.024	0.368	0.024	0.045	0.947	0.014	0.028
26830055 [45]	6167	1166	5102	29	29	19	27	1.000	0.005	0.009	0.655	0.016	0.032	0.931	0.005	0.011
26420598 [83]	8405	8387	2421	57	47	47	30	0.825	0.006	0.011	0.825	0.006	0.011	0.526	0.012	0.024
26420387 [38]	8405	8387	2421	78	59	59	30	0.756	0.007	0.014	0.756	0.007	0.014	0.385	0.012	0.024
25059938 [61]	1524	1503	33	21	14	14	1	0.667	0.009	0.018	0.667	0.009	0.018	0.048	0.030	0.037
24592495 [31]	873	798	100	16	9	9	0	0.562	0.010	0.020	0.562	0.011	0.022	0.000	0.000	0.000
23460092 [85]	2426	2411	0	20	18	18	0	0.900	0.007	0.015	0.900	0.007	0.015	0.000	0.000	0.000
22777524 [88]	4645	3048	2838	37	37	35	31	1.000	0.008	0.016	0.946	0.011	0.023	0.838	0.011	0.022

[a]SR: systematic review.

[b]PMID: PubMed Identifier.

[c]V1: original query.

[d]V2: query with free-text terms only.

[e]V3: query with Medical Subject Headings terms only.

[f]GS: gold standard.

[g]Se: sensitivity.

[h]PPV: positive predictive value.

[i]*F*-Sc: F_1-score.

Of the 59 selected SRs, 29 (49%) came from *The BMJ*, 19 (32%) came from the *Annals of Internal Medicine*, 6 (10%) came from *The Lancet*, and 5 (9%) came from the *Journal of the American Medical Association*. The publication dates were evenly distributed; the mean publication year and the median publication year were both 2016.

The countries of origin of the first authors were available for 49 (83%) SRs. The 3 most frequent countries of origin were the United States (21/49, 43%), the United Kingdom (5/49, 10%), and Canada (5/49, 10%).

Quantification of the Utility of Medical Subject Headings Terms

The queries contained a median (IQR) of 43 (17.0-98) terms. The median (IQR) number of MeSH terms in the V1 queries was 6.0 (3.0-19.5). The median (IQR) proportion of MeSH terms relative to all terms in queries was 18.5% (13.7-25.5).

The V1 queries returned a total of 206,095 items, of which 1628 (0.79%) were included in the GS (Table 1). The V2 queries returned a total of 157,698 items, of which 1473 (0.93%) were included in the GS. In other words, an average of 820.29 additional articles per SR had to be screened for V1, relative to V2. Furthermore, V1 retrieved an average of 2.62 additional relevant articles, when compared with V2.

The median (Q1-Q3) sensitivities of queries V1 and V2 were 77.8% (62.1%-95.2%) and 71.4% (42.6%-90%), respectively (Table 2). The median (Q1-Q3) PPV of queries V1 and V2 were 0.9% (0.3%-2.8%) and 1.1% (0.3%-3.4%), respectively. The median (Q1-Q3) F_1-scores of queries V1 and V2 were 1.8% (0.7%-5.4%) and 2.2% (0.7%-6.1%), respectively. A graphic visualization of the sensitivity and PPV per SR showed that the addition of MeSH terms to a query typically increased the sensitivity but decreased the PPV (Figure 3). Furthermore, it can be seen that the transition from V2 to V1 had no effect for many SRs.

XSL•FO
RenderX

Table 2. Comparison of the performance levels of queries V1, V2, and V3.

Query	Sensitivity, median (IQR)	PPV[a], median (IQR)	F_1-score, median (IQR)	Number of results, median (IQR)	Number of GS[b] items found, median (IQR)	Number of results per GS item found, median (IQR)
Query V1 (MeSH[c] and FTTs[d])	77.8 (62.1-95.2)	0.9 (0.3-2.8)	1.8 (0.7-5.4)	1950 (657.50-6167.00)	17 (7.00-36.50)	108.857 (35.062-298.574)
Query V2 (FTTs only)	71.4 (42.6-90)	1.1 (0.3-3.4)	2.2 (0.7-6.1)	1166 (301.50-2953.00)	15 (4.50-32.50)	88.667 (29.682-314.848)
Query V3 (MeSH only)	35.7 (0-61.7)	0.5 (0-2.6)	1 (0-3.9)	456 (31.50-2188.50)	4 (0-22.50)	81.305 (20.99-564.125)

[a]PPV: positive predictive value.

[b]GS: gold standard.

[c]MeSH: Medical Subject Headings.

[d]FTT: free-text terms (n=59).

Figure 3. Contribution of Medical Subject Headings (MeSH) terms to the queries. The orange circles correspond to V2 (free-text terms only), and the blue dots correspond to V1 (free-text terms and MeSH terms). PPV: positive predictive value.

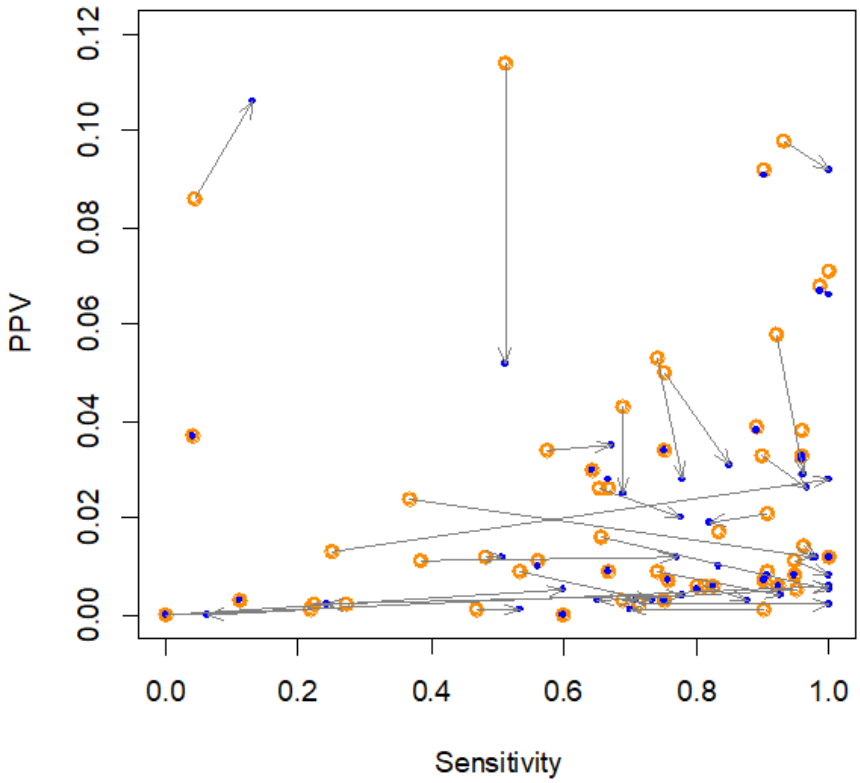

Overall, V1 provided 8.49% more of the GS's items than V2 and 35.55% more of the GS's items than V3. V2 provided 27.06% more of the GS's items than V3. The ratio between the number of GS references retrieved by V1 and the number retrieved by V2 was within the interval (0-1.05) in 66% (39/59) cases (Figure 4). In 59% (35/59) of cases, the ratio was 1 or less. In other words, the transition from V1 to V2 did not have a marked effect on the number of relevant articles retrieved for more than half of the SRs.

We also calculated the ORs for the number of relevant articles retrieved by V2 relative to V1 (Figure 5 [26,31-88]). Overall, the OR (95% CI) for V2 versus V1 was 0.55 (0.38-0.78) for sensitivity and 1.26 (1.03-1.54) for the PPV (Figure 5). The OR (95% CI) for V3 versus V1 was 0.31 (0.23-0.41) for sensitivity and 3.11 (2.15-4.48) for the PPV (Multimedia Appendix 3 [26,31-88]).

XSL•FO
RenderX

Figure 4. Distribution of the ratio between the number of relevant articles found by V1 and the number found by V2. The pink bar corresponds to the interval (1-1.05). For 2 cases, the ratio corresponded to the division of 0 by 0, and we considered that the result was 1. For other 2 cases, the result of the ratio corresponded to infinity (division by 0).

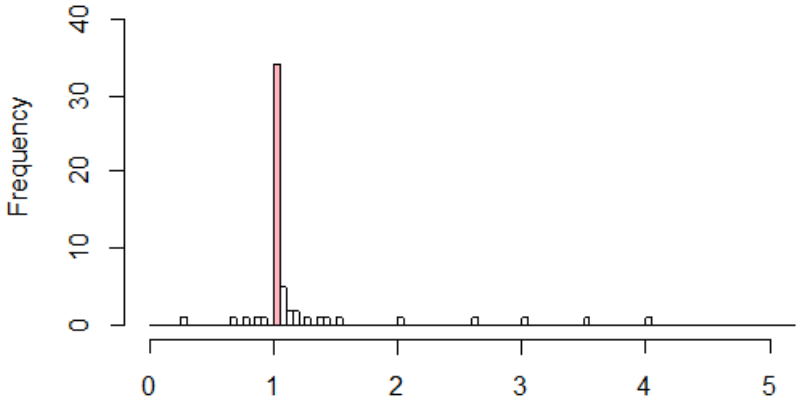

Figure 5. Forest plot of the odds ratio (OR) for V2 versus V1. An OR>1 means that V2 was better than V1 and so that inclusion of the Medical Subject Headings (MeSH) terms was harmful. An OR<1 means that V2 was worse than V1 and so that MeSH terms were useful. FPV1: false positive V1; FPV2: false positive V2; GS: gold standard; PMID: PubMed Identifier; PPV: positive predictive value; TPV1: true positive V1; TPV2: true positive V2.

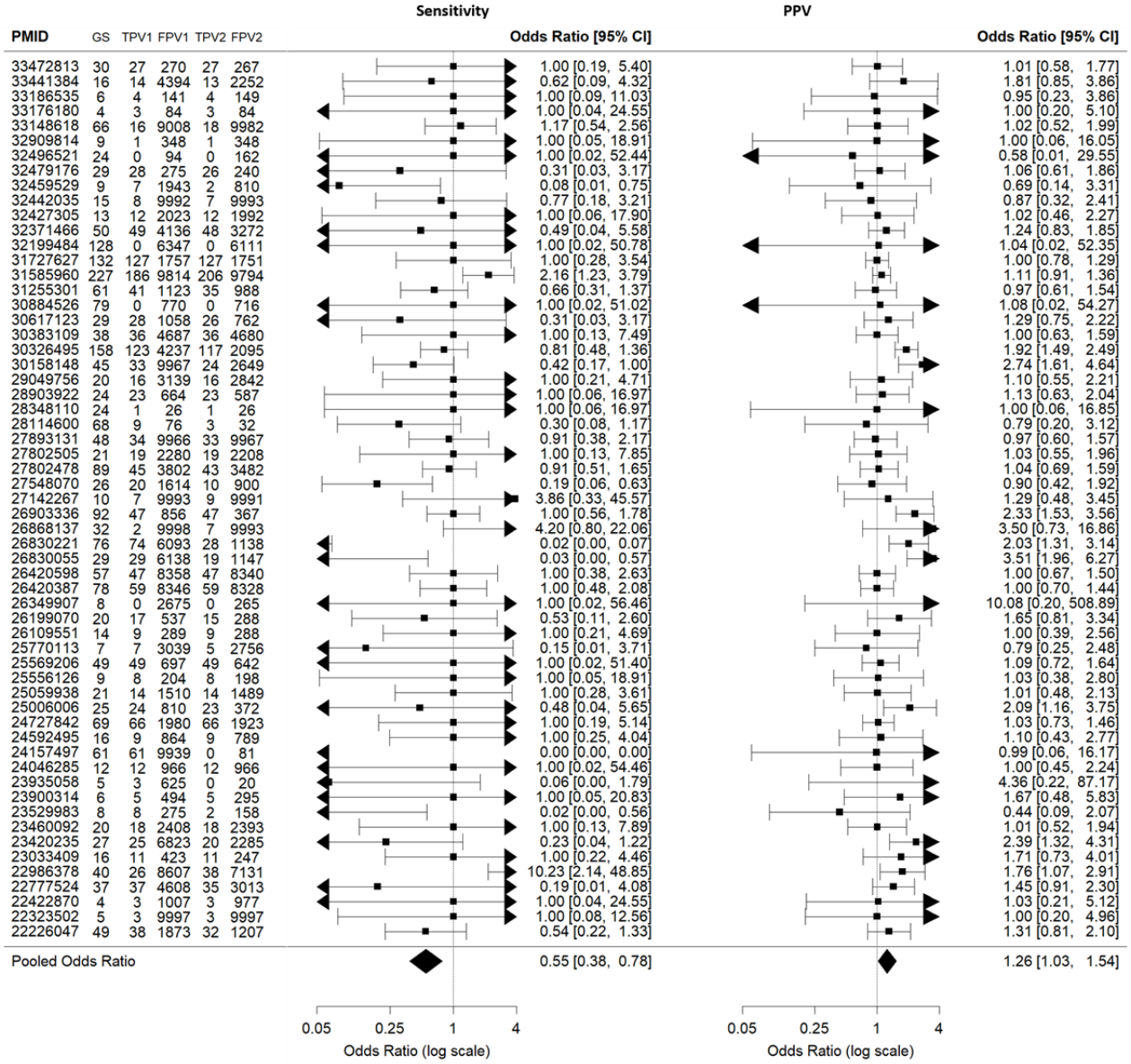

XSL•FO
RenderX

Discussion

Key Results

The objective of this work was to quantify the utility of MeSH terms in SR queries. To this end, we retrieved the queries drafted by the authors of 59 SRs published in 4 prestigious medical journals. We then modified the V1 query to give a free-text terms only query and a MeSH-only query. Finally, we calculated the 3 queries' sensitivities, PPVs, and F_1-scores.

Our first key observation was that MeSH terms typically accounted for a nonnegligible proportion (on average, 20.4%) of the terms in the query. Second, the removal of MeSH terms from SR queries decreased the sensitivity (by 6.4%, on the median) and increased the PPV (by 0.2%, on the median). In other words, queries containing both MeSH terms and free-text terms yield an average of 2.62 additional relevant papers per SR, necessitating the screening of an additional 820.29 papers. The cost of screening an additional collected paper was therefore 313.09, which was slightly more than triple the mean reading cost associated with free-text terms only queries (88.67). Third, our results indicated that the deletion of MeSH terms had no effect on the number of relevant articles retrieved for 35 (59%) of the 59 reviews.

Discussion of the Literature Data

The results of a previous study were similar to those found here; 95% of the relevant articles were retrieved in 67% (49/73) of the analyzed SRs when the query contained free-text terms alone (relative to the V1 query with a mixture of MeSH terms and free-text terms) [22]. Another study with a similar objective gave significantly different results; the free-text terms–only query was 25% less sensitive than MeSH-only query [15]. However, it should be noted that (1) the latter findings were based on a single query, and (2) the MeSH terms were converted to free-text terms manually, with a relatively limited set of synonyms used in the free-text terms strategy.

Furthermore, 3 messages should be highlighted. First, MeSH terms remain an indispensable tool for SRs despite the significant advancements in free-text search engines, especially in an era where the quality of SRs is declining [89]. Second, free-text terms appear to contribute more effectively to the retrieval of relevant articles compared with MeSH terms. Third, mixed queries (combining free-text and MeSH terms) exhibit poor PPV; for rapid literature reviews, it is preferable to use either MeSH terms or free-text terms exclusively.

Our study involved queries developed by experienced researchers; choosing free-text terms can be challenging and requires expertise. It is possible that clinicians with limited experience in literature searching struggle to choose free-text terms effectively, and yet, bibliographic research among clinicians is essential [90]. MeSH terms offer a distinct advantage over free-text terms by covering a broad range of vocabulary, which can be particularly beneficial for clinicians, early-career researchers, or nonnative English speakers. In such cases, incorporating MeSH terms can help clinicians construct more comprehensive and effective queries.

Discussion of the Method

The GS comprised solely MEDLINE-indexed documents with a PMID. This choice was restrictive but technically essential, given that the 3 queries were submitted to the PubMed search engine. However, our restriction to documents with a PMID increased the queries' sensitivities and decreased their PPVs. We expect this bias to be nondifferential, insofar as it should affect the 3 types of queries in the same way.

The publications with PMIDs 26420387 [38] and 26420598 [83] were written by the same authors and were based on the same search query. This was also the case for PMIDs 26830055 [45] and 26830221 [82]. However, we considered these publications to be independent SRs, insofar as the corresponding GSs were different.

Strengths and Weaknesses

Strengths

One strength of our study is that we used queries from a number of different researchers and research centers; this should mean that our results are more representative of currently used search strategies. Furthermore, the automatic transformation of V1 to V2 probably helped us to avoid any bias associated with the differences in an individual's knowledge of the MeSH thesaurus.

Weaknesses

Interpreting the results of V3 is delicate because the authors' queries are not designed to remain viable when ignoring all [tiab] and [all fields], etc. Indeed, after transformation to V3, a total of 11 queries become nonviable and return zero items.

In addition, it is important to note that the use of MeSH terms by the authors of the included SRs may be suboptimal and depends on each author's level of expertise. We assessed the quality of the MeSH selected by the authors of the included SR, not the actual utility of the MeSH as a feature. Finally, we are not able to measure the free-text terms retrieved from initial PubMed searches using only MeSH terms. However, the initial queries using MeSH terms alone may have enriched the search by helping to identify relevant free-text terms. It represents a potentially valuable contribution of MeSH terms that we do not measure here.

Perspectives

Our results and the literature data provide quantitative information on the use and value of MeSH terms in the queries used for SRs. MeSH terms still appear to be important for achieving a comprehensive SR. Our results also emphasized how difficult it is to build a query for an SR and highlighted the significant variability in the results obtained; the search strategies are a matter of concern for researchers [91-93]. With a view to gaining insights into the possible benefits of MeSH terms for use by less experienced researchers, it would be interesting to conduct a similar study of literature searches performed by clinicians. Finally, our study also highlights that any bibliographic research involves a tedious process of sifting through articles, akin to finding a needle in a haystack. While the authors of SRs perform this task efficiently, inexperienced clinicians might find it discouraging to search for scientific

XSL•FO
RenderX

articles. New tools based on network analysis [94] could help these clinicians find relevant articles more quickly.

Conclusion

The objective of this study was to estimate the utility of MeSH terms, selected by authors, in SR queries by analyzing the queries from 59 SRs published in 4 high-impact medical journals in general medicine. Our results revealed that removing MeSH terms from a query decreases sensitivity while slightly increasing the PPV. Queries containing both MeSH and free-text terms yielded more relevant articles but required screening many additional papers. Despite this additional workload, MeSH terms remain indispensable for SRs and can be particularly beneficial for inexperienced clinicians or nonnative English speakers, aiding in constructing more comprehensive queries. However, mixed queries combining MeSH and free-text terms show poor PPV, suggesting the exclusive use of either MeSH terms or free-text terms for rapid reviews.

Acknowledgments

This research did not receive any specific funding from agencies or organizations in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of literature reviews analyzed.
[DOCX File , 22 KB - jmir_v26i1e53781_app1.docx]

Multimedia Appendix 2
All systematic reviews' queries: V1, V2, and V3.
[XLSX File (Microsoft Excel File), 100 KB - jmir_v26i1e53781_app2.xlsx]

Multimedia Appendix 3
Forest plot of the OR for V3 versus V1. An OR>1 means that V3 was better than V1. FPV1: false positive V1; FPV2: false positive V2; GS: gold standard; PMID: PubMed Identifier; OR: odds ratio; PPV: positive predictive value; TPV1: true positive V1; TPV2: true positive V2.
[PNG File , 803 KB - jmir_v26i1e53781_app3.png]

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71-72 [FREE Full text] [doi: 10.1136/bmj.312.7023.71] [Medline: 8555924]
2. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc 2018;106(4):531-541 [FREE Full text] [doi: 10.5195/jmla.2018.283] [Medline: 30271302]
3. Brodman E, Field HG. Librarians' symposia and problems in Medical Subject Headings. Bull Med Libr Assoc 1947;35(4):287-295 [FREE Full text] [Medline: 16016776]
4. Brodman E. Practical or service aspects of Medical Subject Headings. Bull Med Libr Assoc 1948;36(2):102-107 [FREE Full text] [Medline: 16016805]
5. Larkey SV. Introduction to the problem of medical subject headings. Bull Med Libr Assoc 1948;36(2):69-81 [FREE Full text] [Medline: 16016801]
6. Rogers FB. Medical subject headings. Bull Med Libr Assoc 1963;51(1):114-116 [FREE Full text] [Medline: 13982385]
7. Doherty C, Joorabchi A, Megyesi P, Flynn A, Caulfield B. Physiotherapists' use of web-based information resources to fulfill their information needs during a theoretical examination: randomized crossover trial. J Med Internet Res 2020;22(12):e19747 [FREE Full text] [doi: 10.2196/19747] [Medline: 33331826]
8. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc 2000;88(3):265-266 [FREE Full text] [Medline: 10928714]
9. Rennesson M, Georget M, Paillard C, Perrin O, Pigeotte H, Tête C. Le thésaurus, un vocabulaire contrôlé pour parler le même langage. Médecine Palliative 2020;19(1):15-23. [doi: 10.1016/j.medpal.2019.09.003]
10. Salgado TM, Fernandez-Llimos F. Missing pharmacy-specific medical subject headings (MeSH) terms: problems and solutions. Res Social Adm Pharm 2019;15(9):1189-1190. [doi: 10.1016/j.sapharm.2019.04.008] [Medline: 31000335]
11. Clarke M, Greaves L, James S. MeSH terms must be used in Medline searches. BMJ 1997;314(7088):1203 [FREE Full text] [doi: 10.1136/bmj.314.7088.1203a] [Medline: 9146422]

XSL•FO
RenderX

12. Richter R, Austin T. Using MeSH (Medical Subject Headings) to enhance PubMed search strategies for evidence-based practice in physical therapy. Phys Ther 2012;92(1):124-132. [doi: 10.2522/ptj.20100178] [Medline: 21979271]
13. Klerings I, Robalino S, Booth A, Escobar-Liquitay CM, Sommer I, Gartlehner G, Cochrane Rapid Reviews Methods Group. Rapid reviews methods series: guidance on literature search. BMJ Evid Based Med 2023;28(6):412-417 [FREE Full text] [doi: 10.1136/bmjebm-2022-112079] [Medline: 37076268]
14. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. J Med Libr Assoc 2004;92(3):349-353 [FREE Full text] [Medline: 15243641]
15. DeMars MM, Perruso C. MeSH and text-word search strategies: precision, recall, and their implications for library instruction. J Med Libr Assoc 2022;110(1):23-33 [FREE Full text] [doi: 10.5195/jmla.2022.1283] [Medline: 35210959]
16. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309(6964):1286-1291 [FREE Full text] [doi: 10.1136/bmj.309.6964.1286] [Medline: 7718048]
17. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1(6):447-458 [FREE Full text] [doi: 10.1136/jamia.1994.95153434] [Medline: 7850570]
18. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc 2002;9(6):653-658 [FREE Full text] [doi: 10.1197/jamia.m1124] [Medline: 12386115]
19. Kassaï B, Sonié S, Shah NR, Boissel J. Literature search parameters marginally improved the pooled estimate accuracy for ultrasound in detecting deep venous thrombosis. J Clin Epidemiol 2006;59(7):710-714 [FREE Full text] [doi: 10.1016/j.jclinepi.2005.09.013] [Medline: 16765274]
20. Golder S, McIntosh HM, Duffy S, Glanville J, Centre for Reviews and Dissemination and UK Cochrane Centre Search Filters Design Group. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J 2006;23(1):3-12 [FREE Full text] [doi: 10.1111/j.1471-1842.2006.00634.x] [Medline: 16466494]
21. Chang AA, Heskett KM, Davidson TM. Searching the literature using medical subject headings versus text word with PubMed. Laryngoscope 2006;116(2):336-340. [doi: 10.1097/01.mlg.0000195371.72887.a2] [Medline: 16467730]
22. Bramer WM, Giustini D, Kleijnen J, Franco OH. Searching embase and MEDLINE by using only major descriptors or title and abstract fields: a prospective exploratory study. Syst Rev 2018;7(1):200 [FREE Full text] [doi: 10.1186/s13643-018-0864-9] [Medline: 30458825]
23. An introduction to journal impact factor. Clarivate. 2023. URL: https://clarivate.com/products/scientific-and-academic-research/research-analytics-evaluation-and-management-solutions/journal-citation-reports/publishers/first-time-publishers/ [accessed 2023-08-24]
24. Garfield E. Journal impact factor: a brief review. CMAJ 1999;161(8):979-980 [FREE Full text] [Medline: 10551195]
25. How PubMed works: automatic term mapping (ATM). PubMed User Guide. 2023. URL: https://pubmed.ncbi.nlm.nih.gov/help/#automatic-term-mapping [accessed 2023-06-01]
26. Shekelle PG, Newberry SJ, FitzGerald JD, Motala A, O'Hanlon CE, Tariq A, et al. Management of gout: a systematic review in support of an American College of Physicians Clinical Practice Guideline. Ann Intern Med 2017;166(1):37-51 [FREE Full text] [doi: 10.7326/M16-0461] [Medline: 27802478]
27. McKeever L, Nguyen V, Peterson SJ, Gomez-Perez S, Braunschweig C. Demystifying the search button: a comprehensive PubMed search strategy for performing an exhaustive literature review. JPEN J Parenter Enteral Nutr 2015;39(6):622-635 [FREE Full text] [doi: 10.1177/0148607115593791] [Medline: 26129895]
28. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
29. RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC; 2020.
30. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J. Stat. Soft 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
31. Anglemyer A, Horvath T, Rutherford G. The accessibility of firearms and risk for suicide and homicide victimization among household members: a systematic review and meta-analysis. Ann Intern Med 2014;160(2):101-110 [FREE Full text] [doi: 10.7326/M13-1301] [Medline: 24592495]
32. Asbridge M, Hayden JA, Cartwright JL. Acute cannabis consumption and motor vehicle collision risk: systematic review of observational studies and meta-analysis. BMJ 2012;344:e536 [FREE Full text] [doi: 10.1136/bmj.e536] [Medline: 22323502]
33. Balk EM, Rofeberg VN, Adam GP, Kimmel HJ, Trikalinos TA, Jeppson PC. Pharmacologic and nonpharmacologic treatments for urinary incontinence in women: a systematic review and network meta-analysis of clinical outcomes. Ann Intern Med 2019;170(7):465-479 [FREE Full text] [doi: 10.7326/M18-3227] [Medline: 30884526]
34. Bangalore S, Fakheri R, Toklu B, Messerli FH. Diabetes mellitus as a compelling indication for use of renin angiotensin system blockers: systematic review and meta-analysis of randomized trials. BMJ 2016;352:i438 [FREE Full text] [doi: 10.1136/bmj.i438] [Medline: 26868137]
35. Batelaan NM, Bosman RC, Muntingh A, Scholten WD, Huijbregts KM, van Balkom AJLM. Risk of relapse after antidepressant discontinuation in anxiety disorders, obsessive-compulsive disorder, and post-traumatic stress disorder:

XSL•FO
RenderX

systematic review and meta-analysis of relapse prevention trials. BMJ 2017;358:j3927 [FREE Full text] [doi: 10.1136/bmj.j3927] [Medline: 28903922]
36. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ 2019;367:l5358 [FREE Full text] [doi: 10.1136/bmj.l5358] [Medline: 31585960]
37. Boland MV, Ervin AM, Friedman DS, Jampel HD, Hawkins BS, Vollenweider D, et al. Comparative effectiveness of treatments for open-angle glaucoma: a systematic review for the U.S. preventive services task force. Ann Intern Med 2013;158(4):271-279 [FREE Full text] [doi: 10.7326/0003-4819-158-4-201302190-00008] [Medline: 23420235]
38. Bolland MJ, Leung W, Tai V, Bastin S, Gamble GD, Grey A, et al. Calcium intake and risk of fracture: systematic review. BMJ 2015;351:h4580 [FREE Full text] [doi: 10.1136/bmj.h4580] [Medline: 26420387]
39. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. BMJ 2015;350:g7773 [FREE Full text] [doi: 10.1136/bmj.g7773] [Medline: 25569206]
40. Chersich MF, Pham MD, Areal A, Haghighi MM, Manyuchi A, Swift CP, Climate ChangeHeat-Health Study Group. Associations between high temperatures in pregnancy and risk of preterm birth, low birth weight, and stillbirths: systematic review and meta-analysis. BMJ 2020;371:m3811 [FREE Full text] [doi: 10.1136/bmj.m3811] [Medline: 33148618]
41. Chowdhury R, Ramond A, O'Keeffe LM, Shahzad S, Kunutsor SK, Muka T, et al. Environmental toxic metal contaminants and risk of cardiovascular disease: systematic review and meta-analysis. BMJ 2018;362:k3310 [FREE Full text] [doi: 10.1136/bmj.k3310] [Medline: 30158148]
42. Conway R, Low C, Coughlan RJ, O'Donnell MJ, Carey JJ. Methotrexate use and risk of lung disease in psoriasis, psoriatic arthritis, and inflammatory bowel disease: systematic literature review and meta-analysis of randomised controlled trials. BMJ 2015;350:h1269 [FREE Full text] [doi: 10.1136/bmj.h1269] [Medline: 25770113]
43. Drolet M, Bénard É, Pérez N, Brisson M, HPV Vaccination Impact Study Group. Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: updated systematic review and meta-analysis. Lancet 2019;394(10197):497-509 [FREE Full text] [doi: 10.1016/S0140-6736(19)30298-3] [Medline: 31255301]
44. Edmond KM, Kortsalioudaki C, Scott S, Schrag SJ, Zaidi AKM, Cousens S, et al. Group B streptococcal disease in infants aged younger than 3 months: systematic review and meta-analysis. Lancet 2012;379(9815):547-556 [FREE Full text] [doi: 10.1016/S0140-6736(11)61651-6] [Medline: 22226047]
45. Eng J, Wilson RF, Subramaniam RM, Zhang A, Suarez-Cuervo C, Turban S, et al. Comparative effect of contrast media type on the incidence of contrast-induced nephropathy: a systematic review and meta-analysis. Ann Intern Med 2016;164(6):417-424 [FREE Full text] [doi: 10.7326/M15-1402] [Medline: 26830055]
46. Ferreira GE, McLachlan AJ, Lin CC, Zadro JR, Abdel-Shaheed C, O'Keeffe M, et al. Efficacy and safety of antidepressants for the treatment of back pain and osteoarthritis: systematic review and meta-analysis. BMJ 2021;372:m4825 [FREE Full text] [doi: 10.1136/bmj.m4825] [Medline: 33472813]
47. Ferreyro BL, Angriman F, Munshi L, Del Sorbo L, Ferguson ND, Rochwerg B, et al. Association of noninvasive oxygenation strategies with all-cause mortality in adults with acute hypoxemic respiratory failure: a systematic review and meta-analysis. JAMA 2020;324(1):57-67 [FREE Full text] [doi: 10.1001/jama.2020.9524] [Medline: 32496521]
48. Gagne JJ, Bykov K, Choudhry NK, Toomey TJ, Connolly JG, Avorn J. Effect of smoking on comparative efficacy of antiplatelet agents: systematic review, meta-analysis, and indirect comparison. BMJ 2013;347:f5307 [FREE Full text] [doi: 10.1136/bmj.f5307] [Medline: 24046285]
49. Gencer B, Marston NA, Im K, Cannon CP, Sever P, Keech A, et al. Efficacy and safety of lowering LDL cholesterol in older patients: a systematic review and meta-analysis of randomised controlled trials. Lancet 2020;396(10263):1637-1643 [FREE Full text] [doi: 10.1016/S0140-6736(20)32332-1] [Medline: 33186535]
50. Goldenberg JZ, Day A, Brinkworth GD, Sato J, Yamada S, Jönsson T, et al. Efficacy and safety of low and very low carbohydrate diets for type 2 diabetes remission: systematic review and meta-analysis of published and unpublished randomized trial data. BMJ 2021;372:m4743 [FREE Full text] [doi: 10.1136/bmj.m4743] [Medline: 33441384]
51. Goto A, Arah OA, Goto M, Terauchi Y, Noda M. Severe hypoglycaemia and cardiovascular disease: systematic review and meta-analysis with bias analysis. BMJ 2013;347:f4533 [FREE Full text] [doi: 10.1136/bmj.f4533] [Medline: 23900314]
52. Haycock PC, Heydon EE, Kaptoge S, Butterworth AS, Thompson A, Willeit P. Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. BMJ 2014;349:g4227 [FREE Full text] [doi: 10.1136/bmj.g4227] [Medline: 25006006]
53. Hendriksen JMT, Geersing GJ, Lucassen WAM, Erkens PMG, Stoffers HEJH, van Weert HCPM, et al. Diagnostic prediction models for suspected pulmonary embolism: systematic review and independent external validation in primary care. BMJ 2015;351:h4438 [FREE Full text] [doi: 10.1136/bmj.h4438] [Medline: 26349907]
54. Hernandez AV, Roman YM, Pasupuleti V, Barboza JJ, White CM. Hydroxychloroquine or chloroquine for treatment or prophylaxis of COVID-19: a living systematic review. Ann Intern Med 2020;173(4):287-296 [FREE Full text] [doi: 10.7326/M20-2496] [Medline: 32459529]
55. Hu EA, Pan A, Malik V, Sun Q. White rice consumption and risk of type 2 diabetes: meta-analysis and systematic review. BMJ 2012;344:e1454 [FREE Full text] [doi: 10.1136/bmj.e1454] [Medline: 22422870]

XSL•FO
RenderX

56. Huang SW, Tsai CY, Tseng CS, Shih MC, Yeh YC, Chien KL, et al. Comparative efficacy and safety of new surgical treatments for benign prostatic hyperplasia: systematic review and network meta-analysis. BMJ 2019;367:l5919 [FREE Full text] [doi: 10.1136/bmj.l5919] [Medline: 31727627]
57. Hughes D, Judge C, Murphy R, Loughlin E, Costello M, Whiteley W, et al. Association of blood pressure lowering with incident dementia or cognitive impairment: a systematic review and meta-analysis. JAMA 2020;323(19):1934-1944 [FREE Full text] [doi: 10.1001/jama.2020.4249] [Medline: 32427305]
58. Imamura F, O'Connor L, Ye Z, Mursu J, Hayashino Y, Bhupathiraju SN, et al. Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction. BMJ 2015;351:h3576 [FREE Full text] [doi: 10.1136/bmj.h3576] [Medline: 26199070]
59. Juraschek SP, Hu JR, Cluett JL, Ishak A, Mita C, Lipsitz LA, et al. Effects of intensive blood pressure treatment on orthostatic hypotension : a systematic review and individual participant-based meta-analysis. Ann Intern Med 2021;174(1):58-68 [FREE Full text] [doi: 10.7326/M20-4298] [Medline: 32909814]
60. Kavalieratos D, Corbelli J, Zhang D, Dionne-Odom JN, Ernecoff NC, Hanmer J, et al. Association between palliative care and patient and caregiver outcomes: a systematic review and meta-analysis. JAMA 2016;316(20):2104-2114 [FREE Full text] [doi: 10.1001/jama.2016.16840] [Medline: 27893131]
61. Kerrigan D, Kennedy CE, Morgan-Thomas R, Reza-Paul S, Mwangi P, Win KT, et al. A community empowerment approach to the HIV response among sex workers: effectiveness, challenges, and considerations for implementation and scale-up. Lancet 2015;385(9963):172-185 [FREE Full text] [doi: 10.1016/S0140-6736(14)60973-9] [Medline: 25059938]
62. Kim CA, Rasania SP, Afilalo J, Popma JJ, Lipsitz LA, Kim DH. Functional status and quality of life after transcatheter aortic valve replacement: a systematic review. Ann Intern Med 2014;160(4):243-254 [FREE Full text] [doi: 10.7326/M13-1316] [Medline: 24727842]
63. Kim DH, Kim CA, Placide S, Lipsitz LA, Marcantonio ER. Preoperative frailty assessment and outcomes at 6 months or later in older adults undergoing cardiac surgical procedures: a systematic review. Ann Intern Med 2016;165(9):650-660 [FREE Full text] [doi: 10.7326/M16-0652] [Medline: 27548070]
64. Kisely S, Warren N, McMahon L, Dalais C, Henry I, Siskind D. Occurrence, prevention, and management of the psychological effects of emerging virus outbreaks on healthcare workers: rapid review and meta-analysis. BMJ 2020;369:m1642 [FREE Full text] [doi: 10.1136/bmj.m1642] [Medline: 32371466]
65. Kramer CK, Zinman B, Gross JL, Canani LH, Rodrigues TC, Azevedo MJ, et al. Coronary artery calcium score prediction of all cause mortality and cardiovascular events in people with type 2 diabetes: systematic review and meta-analysis. BMJ 2013;346:f1654 [FREE Full text] [doi: 10.1136/bmj.f1654] [Medline: 23529983]
66. Lamont K, Scott NW, Jones GT, Bhattacharya S. Risk of recurrent stillbirth: systematic review and meta-analysis. BMJ 2015;350:h3080 [FREE Full text] [doi: 10.1136/bmj.h3080] [Medline: 26109551]
67. Liao WC, Tu YK, Wu MS, Lin JT, Wang HP, Chien KL. Blood glucose concentration and risk of pancreatic cancer: systematic review and dose-response meta-analysis. BMJ 2015;350:g7371 [FREE Full text] [doi: 10.1136/bmj.g7371] [Medline: 25556126]
68. Lyles CR, Nelson EC, Frampton S, Dykes PC, Cemballi AG, Sarkar U. Using electronic health record portals to improve patient engagement: research priorities and best practices. Ann Intern Med 2020;172(11 Suppl):S123-S129 [FREE Full text] [doi: 10.7326/M19-0876] [Medline: 32479176]
69. Långström N, Enebrink P, Laurén EM, Lindblom J, Werkö S, Hanson RK. Preventing sexual abusers of children from reoffending: systematic review of medical and psychological interventions. BMJ 2013;347:f4630 [FREE Full text] [doi: 10.1136/bmj.f4630] [Medline: 23935058]
70. Martinez L, Cords O, Horsburgh CR, Andrews JR, Pediatric TB Contact Studies Consortium. The risk of tuberculosis in children after close exposure: a systematic review and individual-participant meta-analysis. Lancet 2020;395(10228):973-984 [FREE Full text] [doi: 10.1016/S0140-6736(20)30166-5] [Medline: 32199484]
71. Mendelson A, Kondo K, Damberg C, Low A, Motúapuaka M, Freeman M, et al. The effects of pay-for-performance programs on health, health care use, and processes of care: a systematic review. Ann Intern Med 2017;166(5):341-353 [FREE Full text] [doi: 10.7326/M16-1881] [Medline: 28114600]
72. Naik RP, Smith-Whitley K, Hassell KL, Umeh NI, de Montalembert M, Sahota P, et al. Clinical outcomes associated with sickle cell trait: a systematic review. Ann Intern Med 2018;169(9):619-627 [FREE Full text] [doi: 10.7326/M18-1161] [Medline: 30383109]
73. Schünemann HJ, Khabsa J, Solo K, Khamis AM, Brignardello-Petersen R, El-Harakeh A, et al. Ventilation techniques and risk for transmission of coronavirus disease, including covid-19: a living systematic review of multiple streams of evidence. Ann Intern Med 2020;173(3):204-216 [FREE Full text] [doi: 10.7326/M20-2306] [Medline: 32442035]
74. Newberry SJ, FitzGerald JD, Motala A, Booth M, Maglione MA, Han D, et al. Diagnosis of gout: a systematic review in support of an American College of Physicians Clinical Practice Guideline. Ann Intern Med 2017;166(1):27-36 [FREE Full text] [doi: 10.7326/M16-0462] [Medline: 27802505]

XSL•FO
RenderX

75. Ochen Y, Beks RB, van Heijl M, Hietbrink F, Leenen LPH, van der Velde D, et al. Operative treatment versus nonoperative treatment of achilles tendon ruptures: systematic review and meta-analysis. BMJ 2019;364:k5120 [FREE Full text] [doi: 10.1136/bmj.k5120] [Medline: 30617123]
76. Prvu Bettger J, Alexander KP, Dolor RJ, Olson DM, Kendrick AS, Wing L, et al. Transitional care after hospitalization for acute stroke or myocardial infarction: a systematic review. Ann Intern Med 2012;157(6):407-416 [FREE Full text] [doi: 10.7326/0003-4819-157-6-201209180-00004] [Medline: 22986378]
77. Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S, et al. Prevalence of burnout among physicians: a systematic review. JAMA 2018;320(11):1131-1150 [FREE Full text] [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
78. Salvo F, Moore N, Arnaud M, Robinson P, Raschi E, de Ponti F, et al. Addition of dipeptidyl peptidase-4 inhibitors to sulphonylureas and risk of hypoglycaemia: systematic review and meta-analysis. BMJ 2016;353:i2231 [FREE Full text] [doi: 10.1136/bmj.i2231] [Medline: 27142267]
79. Schandelmaier S, Kaushal A, Lytvyn L, Heels-Ansdell D, Siemieniuk RAC, Agoritsas T, et al. Low intensity pulsed ultrasound for bone healing: systematic review of randomized controlled trials. BMJ 2017;356:j656 [FREE Full text] [doi: 10.1136/bmj.j656] [Medline: 28348110]
80. Selby K, Baumgartner C, Levin TR, Doubeni CA, Zauber AG, Schottinger J, et al. Interventions to improve follow-up of positive results on fecal blood tests: a systematic review. Ann Intern Med 2017;167(8):565-575 [FREE Full text] [doi: 10.7326/M17-1361] [Medline: 29049756]
81. Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, Sepsis Definitions Task Force. Developing a new definition and assessing new clinical criteria for septic shock: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):775-787 [FREE Full text] [doi: 10.1001/jama.2016.0289] [Medline: 26903336]
82. Subramaniam RM, Suarez-Cuervo C, Wilson RF, Turban S, Zhang A, Sherrod C, et al. Effectiveness of prevention strategies for contrast-induced nephropathy: a systematic review and meta-analysis. Ann Intern Med 2016;164(6):406-416 [FREE Full text] [doi: 10.7326/M15-1456] [Medline: 26830221]
83. Tai V, Leung W, Grey A, Reid IR, Bolland MJ. Calcium intake and bone mineral density: systematic review and meta-analysis. BMJ 2015;351:h4183 [FREE Full text] [doi: 10.1136/bmj.h4183] [Medline: 26420598]
84. Thomalla G, Boutitie F, Ma H, Koga M, Ringleb P, Schwamm LH, et al. Intravenous alteplase for stroke with unknown time of onset guided by advanced imaging: systematic review and meta-analysis of individual patient data. Lancet 2020;396(10262):1574-1584 [FREE Full text] [doi: 10.1016/S0140-6736(20)32163-2] [Medline: 33176180]
85. Weaver SJ, Lubomksi LH, Wilson RF, Pfoh ER, Martinez KA, Dy SM. Promoting a culture of safety as a patient safety strategy: a systematic review. Ann Intern Med 2013;158(5 Pt 2):369-374 [FREE Full text] [doi: 10.7326/0003-4819-158-5-201303051-00002] [Medline: 23460092]
86. Wehner MR, Shive ML, Chren MM, Han J, Qureshi AA, Linos E. Indoor tanning and non-melanoma skin cancer: systematic review and meta-analysis. BMJ 2012;345:e5909. [doi: 10.1136/bmj.e5909] [Medline: 23033409]
87. Wu HY, Huang JW, Lin HJ, Liao WC, Peng YS, Hung KY, et al. Comparative effectiveness of renin-angiotensin system blockers and other antihypertensive drugs in patients with diabetes: systematic review and bayesian network meta-analysis. BMJ 2013;347:f6008 [FREE Full text] [doi: 10.1136/bmj.f6008] [Medline: 24157497]
88. Yeh HC, Brown TT, Maruthur N, Ranasinghe P, Berger Z, Suh YD, et al. Comparative effectiveness and safety of methods of insulin delivery and glucose monitoring for diabetes mellitus: a systematic review and meta-analysis. Ann Intern Med 2012;157(5):336-347 [FREE Full text] [doi: 10.7326/0003-4819-157-5-201209040-00508] [Medline: 22777524]
89. Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016;94(3):485-514 [FREE Full text] [doi: 10.1111/1468-0009.12210] [Medline: 27620683]
90. McKibbon KA, Lokker C, Keepanasseril A, Wilczynski NL, Haynes RB. Net improvement of correct answers to therapy questions after PubMed searches: pre/post comparison. J Med Internet Res 2013;15(11):e243 [FREE Full text] [doi: 10.2196/jmir.2572] [Medline: 24217329]
91. Lazarus JV, Palayew A, Rasmussen LN, Andersen TH, Nicholson J, Norgaard O. Searching PubMed to retrieve publications on the COVID-19 pandemic: comparative analysis of search strings. J Med Internet Res 2020;22(11):e23449 [FREE Full text] [doi: 10.2196/23449] [Medline: 33197230]
92. Agoritsas T, Merglen A, Courvoisier DS, Combescure C, Garin N, Perrier A, et al. Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews. J Med Internet Res 2012;14(3):e85 [FREE Full text] [doi: 10.2196/jmir.2021] [Medline: 22693047]
93. Kastner M, Wilczynski NL, Walker-Dilks C, McKibbon KA, Haynes B. Age-specific search strategies for medline. J Med Internet Res 2006;8(4):e25 [FREE Full text] [doi: 10.2196/jmir.8.4.e25] [Medline: 17213044]
94. BibliZap: an open-source and non-profit tool for reference mining that helps find similar articles. BibliZap Team. URL: https://biblizap.org/ [accessed 2024-09-18]

Abbreviations

ATM: automatic term mapping

XSL•FO
RenderX

GS: gold standard
MeSH: Medical Subject Headings
NLM: US National Library of Medicine
OR: odds ratio
PMID: PubMed Identifier
PPV: positive predictive value
SR: systematic review
V1: original query
V2: query with free-text terms only
V3: query with MeSH terms only

Edited by G Tsafnat; submitted 18.10.23; peer-reviewed by M Rethlefsen, M Arab-Zozani, F Fernandez-Limós; comments to author 06.04.24; revised version received 17.04.24; accepted 07.07.24; published 19.11.24.

Please cite as:
Leblanc V, Hamroun A, Bentegeac R, Le Guellec B, Lenain R, Chazard E
Added Value of Medical Subject Headings Terms in Search Strategies of Systematic Reviews: Comparative Study
J Med Internet Res 2024;26:e53781
URL: https://www.jmir.org/2024/1/e53781
doi:10.2196/53781
PMID:

©Victor Leblanc, Aghiles Hamroun, Raphaël Bentegeac, Bastien Le Guellec, Rémi Lenain, Emmanuel Chazard. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Generation Z's Health Information Avoidance Behavior: Insights From Focus Group Discussions

Chenjin Jia[1*], MA; Pengcheng Li[1*], MA

School of Communication, Universiti Sains Malaysia, Gelugor, Malaysia

[*]all authors contributed equally

Corresponding Author:
Pengcheng Li, MA
School of Communication
Universiti Sains Malaysia
11800 USM, Penang
Gelugor
Malaysia
Phone: 60 174564739
Email: DaPeng0605@outlook.com

Abstract

Background: Younger generations actively use social media to access health information. However, research shows that they also avoid obtaining health information online at times when confronted with uncertainty.

Objective: This study aims to examine the phenomenon of health information avoidance among Generation Z, a representative cohort of active web users in this era.

Methods: Drawing on the planned risk information avoidance model, we adopted a qualitative approach to explore the factors related to information avoidance within the context of health and risk communication. The researchers recruited 38 participants aged 16 to 25 years for the focus group discussion sessions.

Results: In this study, we sought to perform a deductive qualitative analysis of the focus group interview content with open, focused, and theoretical coding. Our findings support several key components of the planned risk information avoidance model while highlighting the underlying influence of cognition on emotions. Specifically, socioculturally, group identity and social norms among peers lead some to avoid health information. Cognitively, mixed levels of risk perception, conflicting values, information overload, and low credibility of information sources elicited their information avoidance behaviors. Affectively, negative emotions such as anxiety, frustration, and the desire to stay positive contributed to avoidance.

Conclusions: This study has implications for understanding young users' information avoidance behaviors in both academia and practice.

(J Med Internet Res 2024;26:e54107) doi:10.2196/54107

KEYWORDS

information avoidance; health information; Generation Z; information overload; planned risk information avoidance model

Introduction

Background

As digital natives, Generation Z (born between 1997 and 2012) actively uses a variety of sources for health information. However, they have limited life experience and lower cognitive ability to process large amounts of information compared with older generations [1,2]. As indicated by Robb and Shellenbarger [3], college students are able to retrieve health information on their own, but they are less confident for individual decision-making in health choices. When health risks appear, young people are often fearful and stressed, which can lead to stronger reactions such as avoiding information about the risk [1] or even withdrawing from social media [4]. The studies found that young adults, when perceiving a potential threat and lacking the ability to effectively cope, either seek health information to fulfill their information needs or avoid health information that conflicts with their cognitive beliefs or elicits unpleasant emotions [5,6].

Information avoidance refers to *any behavior designed to prevent or delay the acquisition of available but potentially unwanted information* [6]. Unlike the state of not actively

XSL•FO
RenderX

seeking information, information avoidance is a deliberate act of avoiding or delaying available but unwanted information [7]. This behavior manifests in various ways, such as selective access to information from intermediary sources [8], diverting attention [9], biased interpretation of information, and selective forgetting [10]. Although information avoidance is considered the least active state of information seeking [11], these 2 behaviors are fundamentally distinct. Information seeking entails users having information needs and making efforts to fulfill those needs [12]. On the contrary, information avoidance represents a motivated decision to steer clear of information that threatens how individuals wish to think, feel, or behave [7].

Information avoidance in the health context can occur in various ways. For example, Persoskie et al [13] found that 29.4% of participants above the age of 50 years who knew they should have obtained information from their physicians still avoided seeking medical care. Similarly, a national survey in the United States showed that 31.1% of adults said they would not know their chances of developing cancer [14]. In addition, it encompasses other information-related psychological phenomena, including avoiding exposure to information that is inconsistent with one's attitudes [15], ignoring information about a disease after it has been diagnosed [16], and other forms of defense and avoidance [17]. During a crisis, the avoidance of health information can reduce personal anxiety. However, avoiding health information may negatively impact individuals. It prevents people from digesting valuable information to improve health decisions, leading to delayed or missed disease screenings and medical visits [10].

Exploring how and why individuals seek health information has long been a focus of psychology and communication research. However, relatively little attention has been paid to the phenomenon of information avoidance, especially among specific populations. The purpose of this study is to explore how Generation Z perceives the factors that influence their information avoidance behavior, drawing on the planned risk information avoidance (PRIA) model as a theoretical guide to reveal the underlying motivations for this behavior. In the *PRIA Model*, *Sociocultural Factors*, *Cognitive Factors*, and *Affective Factors* sections, we first review the factors related to information avoidance to provide information for our qualitative investigation. We then describe the research methods and results. Finally, we review our main findings, theoretical and practical implications, limitations, and future research directions.

PRIA Model

As information avoidance behavior has gradually gained attention from scholars, a wide range of theories and frameworks, for example, the theory of planned behavior, the stimulus-organism-response model, and the risk information seeking and processing (RISP) model, have been applied to the study of information avoidance by researchers from psychology, communication, informatics, and medical backgrounds. Some scholars have attempted to understand information avoidance from the perspective of risk communication using a cognitive model [18,19]. Of these, the RISP model is one of the most widely used models for understanding individuals' information-seeking and -processing behaviors across a variety of environmental or health risks [20]. It assumes that the importance of perceived information insufficiency in determining the importance of proactive risk information seeking, including cognitive and psychosocial motivational drivers such as perceived hazard characteristics, affective responses, relevant channel beliefs, and subjective norms of information, which in turn influence perceived information insufficiency. Although the model has been applied to explain and predict information avoidance behavior in different crisis contexts [21-23], the shortcoming is that the correlates of information avoidance are different from information seeking, and it has been shown that the direct relationship between risk perception, insufficient information, and information avoidance has not been supported [24].

To clarify the predictors of information avoidance behavior, Deline and Kahlor [25] proposed the PRIA model as a further complement to the extended iteration of RISP, as shown in Figure 1 [25]. They reviewed the variables in previous risk information–seeking studies and identified potential factors from the literature that may contribute to information avoidance behavior, integrating and proposing models that are more appropriate for studying information avoidance behavior. The model emphasizes the behavioral as-intended nature of risk information avoidance and explains the sociocultural, cognitive, and affective factors that underpin the behavior. Notably, Deline and Kahlor [25] pointed out that PRIA is a "pre-theoretical conceptual representation of a phenomenon...that can be used to elaborate theory, integrate research fields and test variable relationships." Therefore, in this study, we followed the PRIA model as a theoretical framework and coding justification for exploring health information avoidance behaviors in Generation Z.

XSL•FO
RenderX

Figure 1. Planned risk information avoidance (PRIA) model (adapted from Deline and Kahlor). Bold text refers to the factors involved in this study.

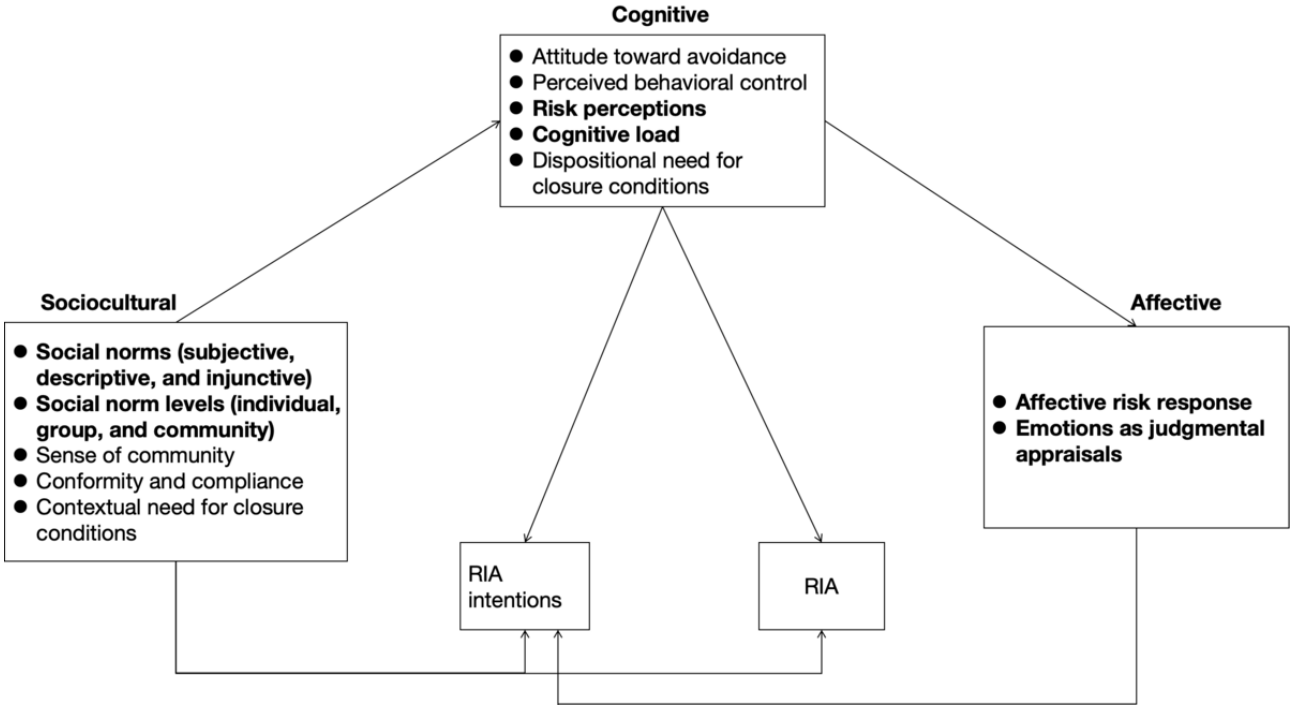

Sociocultural Factors

The social factors encompass the investigation of how an individual's social environment network, including the perceived pressures and expectations from loved ones and peers, influences their evaluation of risk information and subsequent behavioral responses [20]. This study focused on social norms as one of the most relevant factors guiding individual health behaviors, including health information behaviors [9,26]. Social norms can be described as "rules and standards that guide and/or constrain social behavior as understood by group members and without the force of law" [27]. The PRIA model posits a significant association between information avoidance behavior and both descriptive norms and injunctive norms [25]. This proposition is substantiated by Qu et al [28], who contend that individuals who perceive a higher degree of involvement of their social relationships in information avoidance or the approval of such behaviors are more prone to engaging in such behaviors themselves.

Indeed, health behaviors are often influenced by family, friends, and other social relationships, such as health information behaviors [23]. Link et al [9] argued that individuals would avoid health information more frequently if information avoidance was perceived to be endorsed by relevant others (injunctive avoidance norms) or was perceived to be a frequently performed behavior (descriptive avoidance norms). In contrast, Heck and Meyer [29] found that people who avoid genetic health information are evaluated more negatively and that people tend to believe that learning relevant information is the right choice.

Cognitive Factors

Cognitive factors are related to how people perceive risk, how much they know about it, and how these factors drive information avoidance [25]. In the first place, risk perception refers to the individual's perception of the hazardous circumstances associated with a specific risk they encounter. It is determined through subjective evaluation of the severity and susceptibility of potential risks [19,30]. Dunwoody and Griffin [31] posited that heightened risk perception may serve as a catalyst for individuals to engage in information avoidance, particularly when the information is construed as excessively menacing. However, in instances where individuals perceive a minimal likelihood of risk occurrence, they similarly exhibit an active inclination to avoid information regarding the subject matter [25].

It should be noted that risk perception may be altered by cognitive load. Cognitive load is the pressure or burden on an individual's information-processing resources while performing a cognitive task. With limited cognitive capacity, it is difficult for people to effectively handle more than one cognitive work task at a time [32]. Information overload transpires when individuals perceive that they must process an excessive amount of information that surpasses their capacity to effectively use it, thereby resulting in an excessively burdensome cognitive load [1]. The extensive use of the internet coupled with the abundance of intricate information surpasses the information-processing capabilities of Generation Z users, frequently subjecting them to heightened levels of information overload [1]. When information overload exists, individuals strive to alleviate their cognitive load by exhibiting a proclivity for actively pursuing and embracing information that aligns with their preexisting beliefs while avoiding information that is inconsistent with prior convictions [25,33].

Affective Factors

The avoidance of health information is an emotionally driven, maladaptive defense response [6,34]. Following the PRIA model, we focused on the valence of the emotional response (ie, positive or negative). Yang and Kahlor [35] argued that the more negative the emotional risk responses an individual

XSL•FO
RenderX

perceives, the less likely they are to avoid information. Nevertheless, recent research has provided evidence that both negative and positive emotional responses can serve as motivators for information avoidance [20]. Specifically, when individuals receive health information that is associated with negative emotions (eg, getting a cancer diagnosis elicits fear) or requires them to take an unexpected behavior (eg, undergoing surgery), they choose to avoid the information [6,36]. Similarly, Dai et al [37] stated that when users feel frustrated by information overload, they may adopt emotional or behavioral coping strategies, including interrupting or avoiding information, to alleviate or prevent this unpleasant feeling.

There is no doubt that positive emotions such as hope, optimism, and excitement can also influence an individual's behavior. Yang and Kahlor [35] posited that the more positively a person feels about risky information, the more inclined they are to avoid the information. It has been shown that information avoidance occurs when individuals have a higher degree of confidence in their own level of knowledge [22,23]. Generation Z may be inclined to avoid health information when they believe they have sufficient knowledge of it, especially if they do not find it credible or relevant to their situation [38].

Methods

Data Collection

We conducted 7 online focus groups between February 2023 and May 2023 consisting of 38 participants. We recruited samples using WeChat snowball sampling, online community and volunteer advertising sites, and social media advertisements (Weibo promote). Eligible participants must (1) be aged between 16 and 25 years, (2) have a high frequency of social media use (at least once a day), and (3) be experienced with health information avoidance behaviors, among other things. Before formal data collection, all participants signed a consent document agreeing that their conversations would be recorded and used for research purposes. Specifically, the written informed consent was obtained from the parents of 4 of the underage participants. The document emphasized that all records would be handled anonymously.

Data were collected using online focus groups because they may be an effective way to solicit public opinion on health and medical issues, especially from diverse and geographically dispersed participants [39]. The focus groups were conducted through Voov Meeting (Tencent Ltd), the most used online meeting platform in mainland China, which allows for instant communication and enables the moderator to coordinate the entire discussion process appropriately [40].

The first author was the moderator of the focus group discussion. The online focus groups were conducted in a semistructured discussion with participants joining in using audio. Participants were informed of the purpose of the study and introduced to the core concepts discussed (eg, What is information avoidance?). Then, the formal discussion began. The discussion was based on a practical guide to focus groups prepared by Krueger and Casey [41]. The topic guide was developed using the existing literature on attitudes and experiences of information avoidance and was tested and refined in a pilot focus group. The online focus group interview guide is presented in Multimedia Appendix 1. The group sizes ranged from 4 to 8 participants.

Analysis

All discussions were audio recorded and transcribed verbatim with the consent of all participants. We recorded a total of 10 hours and 45 minutes of audio data from the focus group interviews. The duration of each individual interview session ranged from 75 minutes to 106 minutes, resulting in an average total length of 92 minutes per focus group. Participants' names were replaced with P1-P38, all documents and transcripts related to the study were deidentified, and all audio files were destroyed after the transcription was complete. Theme extraction was conducted through deductive coding, whereby predefined coding categories were applied to the qualitative data based on the PRIA model. This was used by 2 trained coresearchers with extensive experience in qualitative content analysis, and they considered generational nuances as well as emerging themes that may not have been captured in the initial coding framework. Themes were extracted and coded over multiple iterations until no new themes could be generated. Coders independently coded all transcripts and discussed and compared their coding results with each other. Disagreements were resolved by checking their coding. Both coders provided coding rationalizations and jointly determined the best theme for the response. Final themes were reviewed and finalized by the coders and the first author.

Ethical Considerations

At the time of the study, the authors' working institution had no policy of ethical clearance for internet-based studies such this one. Therefore, formal ethical approval from the institutional review board was not obtained. Nevertheless, the study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards, including existing laws and regulations on personal data, privacy, and research data management. Informed consent was obtained from all participants before data collection. The collected data was strictly utilized for scientific research purposes and was not stored or used for the identification of any individual participant. At the end of the final focus group, each participant was compensated with CN ¥80 (approximately US $11) e-commerce vouchers.

Results

Overview

A total of 38 participants were recruited, of which 55% (21/38) were male participants. The participants were aged 16 to 25 (mean 20.9, SD 2.52) years. Moreover, 66% (25/38) of the participants were undergraduate students or had a degree. The participants had used social media for an average of 6.89 years. Detailed background characteristics of the participants are shown in Multimedia Appendix 2.

Participants agreed that health information is important, but that the information environment presented significant challenges to their understanding and processing of health information and

XSL•FO
RenderX

that it was easier for them to avoid this information at times. We organized the findings in the *Irrelevant Information Topics*, *The Effects of Group Identity*, *The Conflict of Values*, *Risk Perception Discrepancy*, *Information Overload*, and *Affective Responses to Cognition* sections according to the 3 factors discussed in the literature review (sociocultural, cognitive, and affective). Multimedia Appendix 3 lists all the factors and provides citations for each subtheme.

Irrelevant Information Topics

Participants received information and stimuli from all aspects of internet use on a daily basis. For personal information consumption, they may pay more attention to topics directly related to their daily lives, hobbies, or social relationships, and they may spend time browsing their social media circles of friends or paying attention to content related to their interests while ranking health information as secondary. This preference for the allocation of attention may cause them to ignore or avoid accessing and concerning about health information. One participant said the following:

> *Fashion, entertainment, and study related information takes up my free time. Health information? Nah, not really my thing...[what I'm into are] looking good, having fun, and keeping up with my academic game. Health stuff just doesn't grab my attention, you know?* [P33, aged 20 years, male participant]

For social interactions, if participants consider health information to be irrelevant to their current topic of conversation, they may find it out of place and difficult to integrate with the topic or social interaction they are discussing. Many younger participants felt that there were underlying rules about the topics discussed in their social circles, which made them consciously choose to be exposed to certain information and avoid others. One participant stated the following:

> *Sometimes we just avoid health info because it's not cool to talk about certain stuff. Like, Youth social circle has these unwritten rules about what's okay to discuss and what's off-limits.* [P16, aged 18 years, male participant]

For them, health information was perceived as boring and dull content, as one participant disclosed the following:

> *It seems like everyone agrees that health information is totally unfashionable and nothing worth talking about.* [P5, aged 18 years, male participant]

The Effects of Group Identity

In the lives of young adults, the formation of their identity and integration into social groups represent a crucial developmental task. Some participants felt that if other young people in a WeChat group or other social media community preferred to avoid health risk information, then a sense of belonging as a member of the community led them to follow these avoidance norms, as one participant stated the following:

> *Every group has underlying norms and group rules, and since I'm a member of the WeChat group, I should follow them. It's all about fitting in and following the group's unspoken rules.* [P18, aged 18 years, male participant]

Participants with a strong sense of belonging also indicated a greater tendency to persist in the behavior if positive feedback from peers reinforced avoidance of health messages, as one participant said the following:

> *If all my friends think that leaving those health messages alone is what young folks should be doing, I'll stick with it. I gotta fit in and be consistent with them.* [P10, aged 21 years, female participant]

Young people are often concerned about their image and identity in their social circles. They may be concerned about being seen as different or being excluded from social circles if they focus on or discuss health issues. One participant said the following:

> *When no one else is paying attention to health information, the fear of being judged or ridiculed by my peers is a concern, so I simply avoid it all together as well.* [P34, aged 18 years, female participant]

Another participant shared similar perspectives, saying the following:

> *I worry about being labeled and considered abnormal. So simply choose to avoid it and save us the trouble.* [P38, aged 21 years, male participant]

In this case, participants expressed that although individuals may be aware of the importance of health information, based on the desire for social expectations and group identity, they wanted to align themselves with others. One participant explained the following:

> *Honestly, I understand the significance of health info. But when I see the lack of attention or importance people generally give it in our community or society, I feel this need to conform. It's like I wanna be part of the collective vibe.* [P19, aged 22 years, male participant]

The Conflict of Values

The younger generation has easy access to a wide range of health information, but the quality and credibility of the information varies. They may face conflict over different health advice. Avoiding health information can preserve self-identity in the face of information that challenges their existing values. One participant said the following:

> *I totally believe in scientific health information released by official or medical professionals, while some information always promotes folk remedies and even recommends that cancer can be cured by brewing some unknown plant with water, which is unacceptable to me.* [P29, aged 23 years, male participant]

Some participants in this study reported that this conflict of values also manifested in the way they questioned the source of a particular piece of health information or perceived the source to be potentially biased. For example, one participant said the following:

XSL•FO
RenderX

> *You know that clickbait online? Those accounts don't even publish decent health information; they are actually advertisements.* [P20, aged 21 years, male participant]

Another participant explained as follows:

> *I never read health information published by a self-media outlet, they must have published it for commercial interests. I also avoid all those accounts that promote natural remedies, energy healing, and so on, because I firmly choose modern medicine, and I will consult a doctor if I have health issues.* [P27, aged 23 years, female participant]

Sometimes this conflict of values comes from family members. Some participants expressed that this could make them feel helpless as older family members often share unscientific health information. One participant complained as follows:

> *I know my parents were concerned about our health, but they were often attracted to unscientific or deceptive health information and shared it with our family groups.* [P2, aged 19 years, male participant]

Younger generations perceive older generations as being poor at recognizing the credibility of health information and have a strong bias against their health literacy, with many participants expressing similar views:

> *Maybe it's bias, but I avoid all health information whenever it is shared by relatives and parents.* [P13, aged 21 years, female participant]

As pointing out mistakes may result in family conflict, these young participants hoped to maintain a harmonious family atmosphere and chose to simply avoid health information.

In addition, a culture of self-presentation and comparison is prevalent on social media. If social media emphasizes aspects such as appearance, body shape, and quick slimming, it may cause some self-esteem damage to young users, who may feel that they are not healthy or perfect enough compared with others. To avoid conflicting with the values on social media, they chose to avoid health-related information and comparisons with others. One participant stated the following:

> *A lot of health information preaches perfect, healthy living, as if everyone should have a perfect body or adhere to a specific diet. In real life, I don't stick to these norms, so sometimes I just avoid the info to keep things simpler for myself.* [P17, aged 25 years, female participant]

Risk Perception Discrepancy

Risk perception refers to an individual's ability to perceive and recognize potential risks and possible negative consequences. There were differences in the participants' perceptions of health risks. Some participants stated that the possible consequences of health information can make them uncomfortable, especially when they see some information about serious diseases. One participant said the following:

> *Every time I come across health-related content, it's like my imagination goes wild, conjuring up all these horrifying scenarios. It really gets under my skin, you know? It's just so unsettling and makes me feel really uncomfortable.* [P4, aged 22 years, female participant]

Some individuals were more sensitive to the risks mentioned in the health information, which made them likely to be more concerned about health issues and potential risk factors and may have adopted more cautious and precautionary behaviors:

> *The health information tends to make me self-referential, and I prefer not to hear about potential health issues or underlying diseases that may apply to me. Ignorance is bliss, avoiding them keeps me from having to face possible risks and challenges head-on.* [P12, aged 23 years, male participant]

Health issues are often associated with negative concepts such as disease, pain, and death. As a result, young people may be more willing to focus on the positive aspects and avoid information about health risks and concerns. In contrast, some participants were less sensitive and concerned about potential risks and possible negative consequences. They may perceive themselves to be in relatively good health, with no obvious risk factors, and therefore see themselves as less susceptible to health threats. A participant declared the following:

> *I don't really see health risks everywhere. When it comes to health threats, they just don't bother me at all. I've got this invincible mindset, like nothing can touch me, so I don't even care these info.* [P22, aged 16 years, male participant]

This perception makes individuals less interested in information about their health conditions and disease risk. In addition, there were participants who felt confident in their health knowledge or had sufficient information about the risks, and this confident judgment made them avoid the health information they thought they knew. One participant claimed the following:

> *I know all about those health topics, it's just a matter of repeating that.* [P8, aged 19 years, female participant]

Information Overload

The internet and social media are inundated with a plethora of health information. When individuals are exposed to an information overload that surpasses their capacity to assimilate it, information overload occurs. Most participants expressed that they were exposed to a significantly greater volume of information than what they require. One participant said the following:

> *As soon as I turn on my mobile phone every day, all these apps start flooding me with tons of information. There's just too much health stuff out there, it's like trying to drink from a fire hose.* [P15, aged 23 years, female participant]

The variety of information and socialization needs can be greater than young users can handle, and many participants reported being exhausted after expending too much energy to deal with these needs. One participant said the following:

XSL•FO
RenderX

> *It's like a never-ending stream of articles, videos, and opinions. I'm so tired and afraid of being misinformed that I'd rather avoid them.* [P33, aged 20 years, male participant]

Meanwhile, participants expressed that during a health crisis they perceived a higher level of information overload, especially health-related overuse of the internet:

> *During COVID-19, it felt like everyone and even their grandparents were sharing health stuff left and right, you know? It didn't matter where it came from, it was like everyone suddenly became a doctor or something. And let me tell you, it was a freakin' info overload. True or false, all that health info was flying' around, and I was just getting drowned in it.* [P25, aged 25 years, male participant]

Participants reported that at the beginning of a crisis, they would seek health information to get help because of lack of information. As more information became available, however, more than they could process, they turned to avoiding health information. One participant said the following:

> *I do know that it's useful to know about health information, but it takes more effort to always check out that information and to assess the quality of the content of the information. I'm already swamped with work and don't have the time and energy to delve into these things.* [P24, aged 21 years, female participant]

Affective Responses to Cognition

Cognitive factors relate to participants' understanding and processing of information, whereas affective factors relate to participants' emotional responses and emotional experiences with health information. For participants, different cognitive assessments may lead to different affective responses. For example, the abundance of diverse health information from various sources could lead individuals to experience an influx of different perspectives. Under such circumstances, some participants reported that they may encounter confusion, contradiction, and a sense of being lost in navigating health information:

> *It's like conflicting info everywhere, and we end up feeling confused and unsure about what to believe. Sometimes I just avoid it altogether to save my sanity.* [P26, aged 20 years, male participant]

Furthermore, participants felt that the excessive cognitive load owing to information overload sometimes caused them to feel overly stressed and anxious, and to minimize emotional tension, participants intentionally stayed away from stressful situations. One participant said the following:

> *Sometimes it feels like you gotta absorb every piece of info to stay healthy, but it's impossible. It's like this constant fear of missing out or not doing enough, and it can give you major anxiety.* [P28, aged 21 years, male participant]

Sometimes participants may show stronger emotional response when confronted with information about health risks. Participants with higher risk perceptions said that the possible consequences of the risk could cause them fear, especially when they saw some information about serious diseases:

> *Those anti-smoking ads on cigarette packs are always there...[those pictures] just make me cringe. I mean, I do smoke once in a while, but my best friends smoke like a chimney. So, to save myself from feeling grossed out and uncomfortable, I just steer clear of any health info about lungs.* [P7, aged 25 years, female participant]

When faced with discomforting or anxiety- and fear-inducing health information, people tend to adopt emotional defense mechanisms to avoid negative emotions. Participants indicated that their avoidance of health information reduces emotional upset and stress. One participant said the following:

> *Diving into a bunch of health info just makes me go off on a tangent. I'm all about staying positive and living a healthy life, you know? Happiness is what matters most, right? So, no need to get hung up on all that health stuff.* [P35, aged 23 years, male participant]

Discussion

Principal Findings

The results of this qualitative study shed light on how Generation Z perceives the social, cognitive, and affective factors in their health information avoidance and provides insights for understanding and addressing these behaviors. Two main types of health information avoidance behaviors were observed among participants, regardless of education or age: passive and active avoidance. Passive avoidance occurs mainly under social influence, whereas active avoidance is a stress-coping mechanism used to temporarily avoid confusion when exposed to excessive information. Our findings also showed that the level of risk perception had an inverted U-shaped relationship with information avoidance in participants, with both high- and low-risk perception leading to avoidance. When faced with excessive or complex information, Generation Z experiences increased cognitive load and psychological discomfort, such as stress and anxiety, pushing them to avoid information. In the *Comparison With Prior Work* section, we consider the implications of these findings in relation to existing literature and discuss their significance.

Comparison With Prior Work

This study provides an insightful perspective on the health information behaviors of the younger generation. This indicates that health information avoidance behaviors should not be viewed as an isolated way of information processing but rather as part of the development of healthy information behaviors among young internet users.

Specifically, our findings are consistent with previous research showing that social norms continue to provide strong socially and culturally negotiated, contextually relevant, group identity-based behavioral norms in Generation Z socialization [9]. Young generations commonly select and adapt their information behaviors based on the behavior and acceptance of

XSL•FO
RenderX

their social relationships [28]. This study further points to the finding that, as group interaction is one of the mechanisms through which norms emerge and develop, when peers are not interested in health information, they may be concerned that they will be seen as different or excluded from their social circle if they focus on or discuss health issues. Thus, even if at the individual level they believe that they should be concerned about health information, they may perceive that the dominant norms of the community do not value health information. At this point, health information is no longer cognitively processed, resulting in a passive, involuntary behavior. This supports previous research [25], where subjective avoidance norms cause someone to believe that their friends expect relevant health information possibilities to be avoided in conversations, forcing them to change their personal level norms to be more in line with what they perceive as social norms. Meanwhile, we found no mention in the previous literature that social media widely disseminates information about health views such as the perfect body and particular diets, which may have an influence on young people's self-identity and behavior. When participants realize that they are not aligned with these social norms, they may choose to avoid this information to minimize conflict or discomfort with social norms.

Earlier studies indicated that risk perception is negatively related to information avoidance [22]. In this study, we found that the level of risk perception of the younger generation was more differentiated from their information avoidance behavior, with both high- and low-risk perceptions leading to information avoidance. Therefore, we suggest that there is an inverted U-shaped relationship between the level of risk perception and Generation Z's health information avoidance behavior, which is consistent with the proposition of Deline and Kahlor [25]. Moreover, older participants had higher levels of risk perception, whereas younger participants generally showed lower risk perceptions. This difference may be because higher risk perceptions motivate individuals to avoid information if they are perceived as threatening [31]. As mentioned by Narayan et al [42], active avoidance is a short-term rejection of information and is more of a stress-coping mechanism to temporarily avoid confusion to cope with the risk information. Faced with possible health risks, some young people may experience concerns, and as a defense mechanism, they may choose to avoid information related to health risks to reduce feelings of unease. However, other younger people may lack sufficient knowledge and experience regarding health risks, leading to a diminished sensitivity in risk perception. They are more inclined to prioritize immediate pleasures and adventures, potentially resulting in insufficient attention being given to long-term health risks. In addition, they may exhibit a certain degree of optimistic bias, wherein they hold excessively optimistic views regarding the risks of negative events befalling themselves. They may perceive a lower risk of developing severe illnesses and consequently engage in information avoidance to maintain a lower perceived threat level.

In addition, we found that both the quantity and quality of information can increase the cognitive load of Generation Z, pushing them to avoid information. On the one hand, in line with previous studies, the information overload perceived by the young people in this study made it difficult for them to process and use information effectively [24,43]. When millennials are faced with a plethora of information and the pressure to filter it, there may be a stronger tendency to avoid information that is perceived as less important or irrelevant. On the other hand, our findings show that young people, when faced with information and sources prone to value conflicts, tended to avoid health information to avoid value and family relationship conflicts. This supports a previous study [33], as people often use heuristic decision-making and information filtering strategies during information processing, young people tend to assess the credibility of health information on the basis of their prior knowledge and avoid all information from sources that they perceive as low credibility. Both nonprofessional and nonofficial media were identified as low credibility sources, including information sharing from older family members.

Finally, we found that whether during a health crisis or daily health information dissemination, Generation Z experiences psychological discomfort and negative affective stress, such as stress, anxiety, and frustration, in situations of high cognitive load. Song et al [44] found that perceived information overload significantly predicted the level of anxiety in people's information behavior and positively influenced health information avoidance behavior. Moreover, when young people need to control their fear of risk, they may deny the existence of a health threat by avoiding health information as a way to reduce anxiety or remain hopeful and optimistic.

Implications

This study provides several important implications. To the best of our knowledge, this is one of the first studies guided by the PRIA model to explore Generation Z's health information avoidance behavior. Most studies have focused on Generation Z's seeking of health information rather than their avoidance behavior, for example, Basch et al [45] and Zhou and Roberto [46]. Our study provides new insights into the health information–processing behaviors of the Generation Z. More precisely, it breaks through the stereotype of Generation Z as digital natives in complete control of social media, whereas sometimes cognitive and emotional stresses cause them to avoid information that could cause psychological distress or to stay away from social media temporarily [1]. We explored the relationship between social norms, cognitive load, emotion, and health information avoidance behaviors from the perspective of Generation Z, providing empirical support for the propositions of Deline and Kahlor [25], which complement existing research in communication psychology. As one of the few qualitative studies examining information avoidance behavior, our study has potential value beyond its contextual limitations. Future research could be generalized to other generations, or to information avoidance behavior in one specific disease.

Furthermore, our study has practical implications for social media users and platforms. Short-term avoidance of health information may help reduce feelings of being overwhelmed owing to high cognitive stress and information overload. In the long term, avoiding health information may have adverse effects on health and well-being [16]. For Generation Z users, we suggest engaging in self-monitoring and self-assessing their

XSL•FO
RenderX

psychological state related to health information, thereby promoting healthy information use habits and maintaining psychological well-being. For social media platforms, we offered new insights into the underlying reasons behind Generation Z users' avoidance of health information. It is crucial for platforms to provide personalized services to Generation Z users by developing content filtering and credibility feedback functions that cater to their information consumption needs and preferences.

Finally, for health education development, it is important to improve health literacy among all health system stakeholders so that they are able to filter the required information [21]. Information overload is a serious barrier to health literacy, reducing people's understanding of health information and limiting their ability to evaluate and act on it [47]. Therefore, we recommend that health literacy training be part of the health education program for university students, whereas for young employees, regular health seminars can be arranged to improve health literacy. Exposure to high volumes of health information causes cognitive and emotional stress in young people, and they should be provided with normalization training and guidance during health crises to help them better process health information.

Conclusions

This study explores the negative side of the use of health information in Generation Z and provides empirical evidence on the negative processing of risk information among young people. Using the PRIA model, this study revealed how young people perceive the factors that influence their avoidance of health information. We found that social norms can influence Generation Z's health information behavior through normative pressure and group influence. Moreover, there is an inverted U-shaped relationship between risk perception and the information avoidance behavior of Generation Z, which means that either high- or low-risk perception could lead to information avoidance. Finally, information overload and low credibility increase the cognitive load of young people, causing them to experience negative emotions and therefore choose to avoid information to relieve stress.

However, this study has some limitations. We mainly focused on participants' perceptions and experiences of health information avoidance, rather than directly observing their actual behaviors. Although perceptions provide valuable insights, they may not always align with participants' realistic behaviors. Future research could incorporate behavioral measures to supplement self-reported data to provide a more comprehensive understanding of health information avoidance behaviors. In addition, rapid advances in technology and information environments may affect changes in health information behaviors. Given the dynamic nature of Generation Z's information consumption behaviors, future research could expand this study to explore other health information behavioral similarities or differences in underserved populations.

Acknowledgments

The authors would like to sincerely appreciate the participants who gave their time and effort to this study. The authors would also like to thank the reviewers for their careful review and valuable comments on this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial sample semistructured focus group guideline.
[DOCX File , 16 KB - jmir_v26i1e54107_app1.docx]

Multimedia Appendix 2
Focus group participants' background characteristics.
[DOCX File , 17 KB - jmir_v26i1e54107_app2.docx]

Multimedia Appendix 3
Coding scheme for participants' health information avoidance behavior.
[DOCX File , 17 KB - jmir_v26i1e54107_app3.docx]

References

1. Liu H, Liu W, Yoganathan V, Osburg VS. COVID-19 information overload and generation Z's social media discontinuance intention during the pandemic lockdown. Technol Forecast Soc Change 2021 May;166:120600 [FREE Full text] [doi: 10.1016/j.techfore.2021.120600] [Medline: 34876758]
2. Priporas CV, Stylos N, Fotiadis AK. Generation Z consumers' expectations of interactions in smart retailing: a future agenda. Comput Hum Behav 2017 Dec;77:374-381. [doi: 10.1016/j.chb.2017.01.058]
3. Robb M, Shellenbarger T. Influential factors and perceptions of eHealth literacy among undergraduate college students. Online J Nurs Inform 2014;18(3) [FREE Full text]

XSL•FO
RenderX

4. Liu Y, He J. "Why are you running away from social media?" analysis of the factors influencing social media fatigue: an empirical data study based on Chinese youth. Front Psychol 2021;12:674641 [FREE Full text] [doi: 10.3389/fpsyg.2021.674641] [Medline: 34621208]
5. Savolainen R. Cognitive barriers to information seeking: a conceptual analysis. J Inf Sci 2015 May 28;41(5):613-623. [doi: 10.1177/0165551515587850]
6. Sweeny K, Melnyk D, Miller W, Shepperd JA. Information avoidance: who, what, when, and why. Rev Gen Psychol 2010 Dec 01;14(4):340-353. [doi: 10.1037/a0021288]
7. Howell JL, Lipsey NP, Shepperd JA. Health information avoidance. In: Sweeny K, Robbins ML, Cohen LM, editors. The Wiley Encyclopedia of Health Psychology. Hoboken, NJ: John Wiley & Sons; Sep 2, 2020.
8. Chae J. Who avoids cancer information? Examining a psychological process leading to cancer information avoidance. J Health Commun 2016 Jul 23;21(7):837-844. [doi: 10.1080/10810730.2016.1177144] [Medline: 27337343]
9. Link E, Baumann E, Leuppert R. Clarifying the relationship between information seeking and avoidance: longitudinal analyses of information behaviors and their normative influences. Health Commun 2023 Apr 18:1-15. [doi: 10.1080/10410236.2023.2201736] [Medline: 37072691]
10. Golman R, Hagmann D, Loewenstein G. Information avoidance. J Econ Lit 2017 Mar 01;55(1):96-135. [doi: 10.1257/jel.20151245]
11. Costello KL, Veinot TC. A spectrum of approaches to health information interaction: from avoidance to verification. J Assoc Inf Sci Technol 2019 Oct 26;71(8):871-886. [doi: 10.1002/asi.24310]
12. Pian W, Song S, Zhang Y. Consumer health information needs: a systematic review of measures. Inf Process Manag 2020 Mar;57(2):102077. [doi: 10.1016/j.ipm.2019.102077]
13. Persoskie A, Ferrer RA, Klein WM. Association of cancer worry and perceived risk with doctor avoidance: an analysis of information avoidance in a nationally representative US sample. J Behav Med 2014 Oct 27;37(5):977-987. [doi: 10.1007/s10865-013-9537-2] [Medline: 24072430]
14. St. Jean B, Jindal G, Liao Y. Is ignorance really bliss?: exploring the interrelationships among information avoidance, health literacy and health justice. Proc Assoc Inf Sci Technol 2017 Oct 24;54(1):394-404. [doi: 10.1002/pra2.2017.14505401043]
15. Li K, Li Y, Zhang P. Selective exposure to COVID-19 vaccination information: the influence of prior attitude, perceived threat level and information limit. Library Hi Tech 2021 Nov 30;40(2):323-339. [doi: 10.1108/lht-03-2021-0117]
16. Hua J, Howell JL. Coping self-efficacy influences health information avoidance. J Health Psychol 2022 Mar 21;27(3):713-725. [doi: 10.1177/1359105320965664] [Medline: 33086899]
17. Howell JL, Ratliff KA, Shepperd JA. Automatic attitudes and health information avoidance. Health Psychol 2016 Aug;35(8):816-823. [doi: 10.1037/hea0000330] [Medline: 27505200]
18. Choo CW. Seeking and avoiding information in a risky world. Inf Res Int Electron J 2017 Sep;22(3) [FREE Full text]
19. Griffin RJ, Dunwoody S, Neuwirth K. Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. Environ Res 1999 Feb;80(2 Pt 2):S230-S245. [doi: 10.1006/enrs.1998.3940] [Medline: 10092438]
20. Li J. Information avoidance in the age of COVID-19: a meta-analysis. Inf Process Manag 2023 Jan;60(1):103163 [FREE Full text] [doi: 10.1016/j.ipm.2022.103163] [Medline: 36405670]
21. Soroya SH, Farooq A, Mahmood K, Isoaho J, Zara SE. From information seeking to information avoidance: understanding the health information behavior during a global health crisis. Inf Process Manag 2021 Mar;58(2):102440 [FREE Full text] [doi: 10.1016/j.ipm.2020.102440] [Medline: 33281273]
22. Kim HK, Ahn J, Atkinson L, Kahlor LA. Effects of COVID-19 misinformation on information seeking, avoidance, and processing: a multicountry comparative study. Sci Commun 2020 Sep 13;42(5):586-615. [doi: 10.1177/1075547020959670]
23. Kahlor LA, Olson HC, Markman AB, Wang W. Avoiding trouble: exploring environmental risk information avoidance intentions. Environ Behav 2018 Sep 11;52(2):187-218. [doi: 10.1177/0013916518799149]
24. Link E. Information avoidance during health crises: predictors of avoiding information about the COVID-19 pandemic among german news consumers. Inf Process Manag 2021 Nov;58(6):102714 [FREE Full text] [doi: 10.1016/j.ipm.2021.102714] [Medline: 34539039]
25. Deline MB, Kahlor LA. Planned risk information avoidance: a proposed theoretical model. Commun Theor 2019;29(3):260-282 [FREE Full text] [doi: 10.1093/ct/qty035]
26. McEachan RR, Conner M, Taylor NJ, Lawton RJ. Prospective prediction of health-related behaviours with the theory of planned behaviour: a meta-analysis. Health Psychol Rev 2011 Sep;5(2):97-144. [doi: 10.1080/17437199.2010.521684]
27. Cialdini RB, Trost MR. Social influence: social norms, conformity and compliance. In: Gilbert DT, Fiske ST, Lindzey G, editors. The Handbook of Social Psychology. New York, NY: McGraw-Hill Education; 1998:151-192.
28. Qu Y, Saffer AJ, Austin L. What drives people away from COVID-19 information?: uncovering the influences of personal networks on information avoidance. Health Commun 2023 Feb 30;38(2):216-227. [doi: 10.1080/10410236.2021.1944457] [Medline: 34187260]
29. Heck PR, Meyer MN. Information avoidance in genetic health: perceptions, norms, and preferences. Soc Cognit 2019 Jun;37(3):266-293. [doi: 10.1521/soco.2019.37.3.266]

XSL•FO
RenderX

30. Yang ZJ, Aloe AM, Feeley TH. Risk information seeking and processing model: a meta-analysis. J Commun 2014 Jan 07;64(1):20-41. [doi: 10.1111/jcom.12071]
31. Dunwoody S, Griffin R. Risk information seeking and processing model. In: Cho H, McComas KA, Reimer T, editors. The SAGE Handbook of Risk Communication. Thousand Oaks, CA: SAGE Publications; 2014:102-116.
32. Lang A. Using the limited capacity model of motivated mediated message processing to design effective cancer communication messages. J Commun 2006 Aug;56(suppl_1):S57-S80 [FREE Full text] [doi: 10.1111/j.1460-2466.2006.00283.x]
33. Wu S, Zhang J, Du L. "I do not trust health information shared by my parents": credibility judgement of health (mis)information on social media in China. Health Commun 2024 Apr 22;39(1):96-106. [doi: 10.1080/10410236.2022.2159143] [Medline: 36548158]
34. Howell JL, Shepperd JA. Reducing health-information avoidance through contemplation. Psychol Sci 2013 Sep 10;24(9):1696-1703. [doi: 10.1177/0956797613478616] [Medline: 23842956]
35. Yang ZJ, Kahlor L. What, me worry? the role of affect in information seeking and avoidance. Sci Commun 2012 May 24;35(2):189-212. [doi: 10.1177/1075547012441873]
36. Siebenhaar KU, Köther AK, Alpers GW. Dealing with the COVID-19 infodemic: distress by information, information avoidance, and compliance with preventive measures. Front Psychol 2020;11:567905 [FREE Full text] [doi: 10.3389/fpsyg.2020.567905] [Medline: 33224060]
37. Dai B, Ali A, Wang H. Exploring information avoidance intention of social media users: a cognition–affect–conation perspective. Internet Res 2020 Jul 14;30(5):1455-1478. [doi: 10.1108/intr-06-2019-0225]
38. Guo Y, Lu Z, Kuang H, Wang C. Information avoidance behavior on social network sites: information irrelevance, overload, and the moderating role of time pressure. Int J Inf Manag 2020 Jun;52:102067. [doi: 10.1016/j.ijinfomgt.2020.102067]
39. Williams SN, Armitage CJ, Tampe T, Dienes K. Public perceptions and experiences of social distancing and social isolation during the COVID-19 pandemic: a UK-based focus group study. BMJ Open 2020 Jul 20;10(7):e039334 [FREE Full text] [doi: 10.1136/bmjopen-2020-039334] [Medline: 32690752]
40. Chen SC, Cheng HL, Wang DD, Wang S, Yin YH, Suen LK, et al. Experience of parents in delivering pediatric tuina to children with symptoms of attention deficit hyperactivity disorder during the COVID-19 pandemic: qualitative findings from focus group interviews. BMC Complement Med Ther 2023 Feb 18;23(1):53 [FREE Full text] [doi: 10.1186/s12906-023-03891-3] [Medline: 36804041]
41. Krueger RA, Casey MA. Focus Groups: A Practical Guide for Applied Research. Thousand Oaks, CA: SAGE Publications; 2014.
42. Narayan B, Case DO, Edwards SL. The role of information avoidance in everyday-life information behaviors. Proc Am Soc Inf Sci Technol 2012 Jan 11;48(1):1-9. [doi: 10.1002/meet.2011.14504801085]
43. Ford JL, Douglas M, Barrett AK. The role of pandemic fatigue in seeking and avoiding information on COVID-19 among young adults. Health Commun 2023 Oct 05;38(11):2336-2349. [doi: 10.1080/10410236.2022.2069211] [Medline: 35514105]
44. Song S, Yao X, Wen N. What motivates Chinese consumers to avoid information about the COVID-19 pandemic?: the perspective of the stimulus-organism-response model. Inf Process Manag 2021 Jan;58(1):102407 [FREE Full text] [doi: 10.1016/j.ipm.2020.102407] [Medline: 33041437]
45. Basch CH, MacLean SA, Romero RA, Ethan D. Health information seeking behavior among college students. J Community Health 2018 Dec 19;43(6):1094-1099. [doi: 10.1007/s10900-018-0526-9] [Medline: 29779074]
46. Zhou X, Roberto AJ. An application of the risk information seeking and processing model in understanding college students' COVID-19 vaccination information seeking and behavior. Sci Commun 2022 Aug 24;44(4):446-474. [doi: 10.1177/10755470221120415]
47. Swar B, Hameed T, Reychav I. Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. Comput Hum Behav 2017 May;70:416-425. [doi: 10.1016/j.chb.2016.12.068]

Abbreviations

PRIA: planned risk information avoidance
RISP: risk information seeking and processing

XSL•FO
RenderX

Edited by Z Yin; submitted 30.10.23; peer-reviewed by K Keeling, AM Huhta; comments to author 22.01.24; revised version received 02.02.24; accepted 03.02.24; published 08.03.24.

Please cite as:
Jia C, Li P
Generation Z's Health Information Avoidance Behavior: Insights From Focus Group Discussions
J Med Internet Res 2024;26:e54107
URL: https://www.jmir.org/2024/1/e54107
doi:10.2196/54107
PMID:38457223

©Chenjin Jia, Pengcheng Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patients' Experiences With Digitalization in the Health Care System: Qualitative Interview Study

Christian Gybel Jensen[1*], MA; Frederik Gybel Jensen[1*], MA; Mia Ingerslev Loft[1,2*], MSc, PhD

[1]Department of Neurology, Rigshospitalet, Copenhagen, Denmark

[2]Institute for People and Technology, Roskilde University, Roskilde, Denmark

[*]all authors contributed equally

Corresponding Author:
Mia Ingerslev Loft, MSc, PhD
Department of Neurology
Rigshospitalet
Inge Lehmanns Vej 8
Copenhagen
Denmark
Phone: 45 35457076
Email: belle.mia.ingerslev.loft@regionh.dk

Abstract

Background: The digitalization of public and health sectors worldwide is fundamentally changing health systems. With the implementation of digital health services in health institutions, a focus on digital health literacy and the use of digital health services have become more evident. In Denmark, public institutions use digital tools for different purposes, aiming to create a universal public digital sector for everyone. However, this digitalization risks reducing equity in health and further marginalizing citizens who are disadvantaged. Therefore, more knowledge is needed regarding patients' digital practices and experiences with digital health services.

Objective: This study aims to examine digital practices and experiences with public digital health services and digital tools from the perspective of patients in the neurology field and address the following research questions: (1) How do patients use digital services and digital tools? (2) How do they experience them?

Methods: We used a qualitative design with a hermeneutic approach. We conducted 31 semistructured interviews with patients who were hospitalized or formerly hospitalized at the department of neurology in a hospital in Denmark. The interviews were audio recorded and subsequently transcribed. The text from each transcribed interview was analyzed using manifest content analysis.

Results: The analysis provided insights into 4 different categories regarding digital practices and experiences of using digital tools and services in health care systems: social resources as a digital lifeline, possessing the necessary capabilities, big feelings as facilitators or barriers, and life without digital tools. Our findings show that digital tools were experienced differently, and specific conditions were important for the possibility of engaging in digital practices, including having access to social resources; possessing physical, cognitive, and communicative capabilities; and feeling motivated, secure, and comfortable. These prerequisites were necessary for participants to have positive experiences using digital tools in the health care system. Those who did not have these prerequisites experienced challenges and, in some cases, felt left out.

Conclusions: Experiences with digital practices and digital health services are complex and multifaceted. Engagement in digital practices for the examined population requires access to continuous assistance from their social network. If patients do not meet requirements, digital health services can be experienced as exclusionary and a source of concern. Physical, cognitive, and communicative difficulties might make it impossible to use digital tools or create more challenges. To ensure that digitalization does not create inequities in health, it is necessary for developers and institutions to be aware of the differences in digital health literacy, focus on simplifying communication with patients and next of kin, and find flexible solutions for citizens who are disadvantaged.

(J Med Internet Res 2024;26:e47278) doi:10.2196/47278

XSL•FO
RenderX

KEYWORDS

digitalization; digital health; eHealth; digital health literacy; digital practices; patient experiences; digital health services; inequity in health; qualitative research; interview; implementation; tool; neurology; digital tool; communication; mobile phone

Introduction

Background

In 2022, the fourth most googled question in Denmark was, "Why does MitID not work?" [1]. MitID (My ID) is a digital access tool that Danes use to enter several different private and public digital services, from bank accounts to mail from their municipality or the state. MitID is a part of many Danish citizens' everyday lives because the public sector in Denmark is digitalized in many areas. In recent decades, digitalization has changed how governments and people interact and has demonstrated the potential to change the core functions of public sectors and delivery of public policies and services [2]. When public sectors worldwide become increasingly digitalized, this transformation extends to the public health sectors as well, and some studies argue that we are moving toward a "digital public health era" that is already impacting the health systems and will fundamentally change the future of health systems [3]. While health systems are becoming more digitalized, it is important that both patients and digitalized systems adapt to changes in accordance with each other. Digital practices of people can be understood as what people do with and through digital technologies and how people relate to technology [4]. Therefore, it is relevant to investigate digital practices and how patients perceive and experience their own use of digital tools and services, especially in relation to existing digital health services. In our study, we highlight a broad perspective on experiences with digital practices and particularly add insight into the challenges with digital practices faced by patients who have acute or chronic illness, with some of them also experiencing physical, communicative, or cognitive difficulties.

An international Organization for Economic Cooperation and Development report indicates that countries are digitalized to different extents and in different ways; however, this does not mean that countries do not share common challenges and insights into the implementation of digital services [2].

In its global Digital Government Index, Denmark is presented as one of the leading countries when it comes to public digitalization [2]. Recent statistics indicate that approximately 97% of Danish families have access to the internet at home [5]. The Danish health sector already offers many different digital services, including web-based delivery of medicine, e-consultations, patient-related outcome questionnaires, and seeking one's own health journal or getting test results through; "Sundhed" [6] (the national health portal) and "Sundhedsjournalen" (the electronic patient record); or the apps "Medicinkortet" (the shared medication record), "Minlæge" (My Doctor, consisting of, eg, communication with the general practitioner), or "MinSP" (My Health Platform, consisting of, eg, communication with health care staff in hospitals) [6-8].

The Danish Digital Health Strategy from 2018 aims to create a coherent and user-friendly digital public sector for everyone [9], but statistics indicate that certain groups in society are not as digitalized as others. In particular, the older population uses digital services the least, with 5% of people aged 65 to 75 years and 18% of those aged 75 to 89 years having never used the internet in 2020 [5]. In parts of the literature, it has been problematized how the digitalization of the welfare state is related to the marginalization of older citizens who are socially disadvantaged [10]. However, statistics also indicate that the probability of using digital tools increases significantly as a person's experience of using digital tools increases, regardless of their age or education level [5].

Understanding the digital practices of patients is important because they can use digital tools to engage with the health system and follow their own health course. Researching experiences with digital practices can be a way to better understand potential possibilities and barriers when patients use digital health services. With patients becoming more involved in their own health course and treatment, the importance of patients' health literacy is being increasingly recognized [11]. The World Health Organization defines health literacy as the "achievement of a level of knowledge, personal skills and confidence to take action to improve personal and community health by changing personal lifestyles and living conditions" [12]. Furthermore, health literacy can be described as "a person's knowledge and competencies to meet complex demands of health in modern society," and it is viewed as a critical step toward patient empowerment [11,12]. In a digitalized health care system, this also includes the knowledge, capabilities, and resources that individuals require to use and benefit from eHealth services, that is, "digital health literacy (eHealth literacy)" [13]. An eHealth literacy framework created by Norgaard et al [13] identified that different aspects, for example, the ability to process information and actively engage with digital services, can be viewed as important facets of digital health literacy. This argument is supported by studies that demonstrate how patients with cognitive and communicative challenges experience barriers to the use of digital tools and require different approaches in the design of digital solutions in the health sector [14,15]. Access to digital services and digital literacy is becoming increasingly important determinants of health, as people with digital literacy and access to digital services can facilitate improvement of health and involvement in their own health course [16].

Objectives

The need for a better understanding of eHealth literacy and patients' capabilities to meet public digital services' demands as well as engage in their own health calls for a deeper investigation into digital practices and the use of digital tools and services from the perspective of patients with varying digital capabilities. Important focus areas to better understand digital practices and related challenges have already been highlighted in various studies. They indicate that social support, assessment of value in digital services, and systemic assessment of digital capabilities are important in the use and implementation of digital tools, and they call for better insight into complex

XSL•FO
RenderX

experiences with digital services [13,17,18]. Therefore, we aimed to examine digital practices and experiences with public digital health services and digital tools from the perspective of patients, addressing the following research questions: how do patients use digital services and digital tools, and how do they experience them?

Methods

Design

We aimed to investigate digital practices and experiences with digital health services and digital tools; therefore, we used a qualitative design and adopted a hermeneutic approach as the point of departure, which means including preexisting knowledge of digital practices but also providing room for new comprehension [19]. Our interpretive approach is underpinned by the philosophical hermeneutic approach by Gadamer et al [19], in which they described the interpretation process as a "hermeneutic circle," where the researcher enters the interpretation process with an open mind and historical awareness of a phenomenon (preknowledge). We conducted semistructured interviews using an interview guide. This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [20].

Setting and Participants

To gain a broad understanding of experiences with public digital health services, a purposive sampling strategy was used. All 31 participants were hospitalized or formerly hospitalized patients in a large neurological department in the capital of Denmark (Table 1). We assessed whether including patients from the neurological field would give us a broad insight into the experiences of digital practices from different perspectives. The department consisted of, among others, 8 inpatient units covering, for example, acute neurology and stroke units, from which the patients were recruited. Patients admitted to a neurological department can have both acute and transient neurological diseases, such as infections in the brain, stroke, or blood clot in the brain from which they can recover completely or have persistent physical and mental difficulties, or experience chronic neurological and progressive disorders such as Parkinson disease and dementia. Some patients hospitalized in neurological care will have communicative and cognitive difficulties because of their neurological disorders. Nursing staff from the respective units helped the researchers (CGJ, FGJ, and MIL) identify patients who differed in terms of gender, age, and severity of neurological illness. Some patients (6/31, 19%) had language difficulties; however, a speech therapist assessed them as suitable participants. We excluded patients with severe cognitive difficulties and those who were not able to speak the Danish language. Including patients from the field of neurology provided an opportunity to study the experience of digital health practice from various perspectives. Hence, the sampling strategy enabled the identification and selection of information-rich participants relevant to this study [21], which is the aim of qualitative research. The participants were invited to participate by either the first (CGJ) or last author (MIL), and all invited participants (31/31, 100%) chose to participate.

All 31 participants were aged between 40 to 99 years, with an average age of 71.75 years (Table 1). Out of the 31 participants, 10 (32%) had physical disabilities or had cognitive or communicative difficulties due to sequela in relation to neurological illness or other physical conditions.

Table 1. Participant demographics (N=31).

Demographics	Participant, n (%)
Gender	
Men	15 (48)
Women	16 (52)
Age (years)	
40-49	4 (13)
50-59	2 (6)
60-69	3 (10)
70-79	14 (45)
80-89	6 (19)
90-99	2 (6)
Hospitalized or formerly hospitalized in the neurology department	
Interviewed only during admission	20 (65)
Interviewed after being discharged	11 (35)
Patients with physical, communicative, or cognitive difficulties	10 (32)

Data Collection

The 31 patient interviews were conducted over a 2-month period between September and November 2022. Of the 31 patients, 20 (65%) were interviewed face-to-face at the hospital in their patient room upon admission and 11 (35%) were interviewed on the phone after being discharged. The interviews had a mean length of 20.48 minutes.

XSL•FO
RenderX

We developed a semistructured interview guide (Table 2). The interview questions were developed based on the research aim, findings from our preliminary covering of literature in the field presented in the Introduction section, and identified gaps that we needed to elaborate on to be able to answer our research question [22]. The semistructured interview guide was designed to support the development of a trusting relationship and ensure the relevance of the interviews' content [22]. The questions served as a prompt for the participants and were further supported by questions such as "please tell me more" and "please elaborate" throughout the interview, both to heighten the level of detail and to verify our understanding of the issues at play. If the participant had cognitive or communicative difficulties, communication was supported using a method called Supported Communication for Adults with Aphasia [23] during the interview.

The interviews were performed by all authors (CGJ, FGJ, and MIL individually), who were skilled in conducting interviews and qualitative research. The interviewers are not part of daily clinical practice but are employed in the department of neurology from where the patients were recruited. All interviews were audio recorded and subsequently transcribed verbatim by all 3 authors individually.

XSL•FO
RenderX

Table 2. Interview guide.

Main questions	Supporting questions, if needed
Background questions	
Please tell me a little about yourself.	• What is your gender? What is your age? • What is your level of education and degree? • Why were you admitted to the hospital (if they were), and have you been admitted to a hospital before?
Questions regarding digital practices	
Please tell me about your digital devices.	• Do you own your own computer or smartphone or a similar device? • How do you use these devices?
What is your experience with using digital devices?	• Can you elaborate and give examples? • Would you consider computer or smartphones a part of your everyday life? If so, in which way? • Have you ever used these devices for work?
What does it mean to you, to be able/not able to use these devices?	• Can you elaborate and give examples? • Have your ever wished you were better at using these devices? • Is using these devices something you would like to learn? (if they have not) • Could you imagine using any of these devices 1 day in the future? Why or why not?
Please tell me how you use digital services from the public sector.	• For example, do you read mail digitally? • Do you access your bank account on the web?
What do you think about the use of digital devices in the public sector?	• Not available
Please tell me how you have experienced your use of the digital services in the public sector.	• Can you elaborate and give examples? • Did you face any challenges? • How did you handle them? (if they have any) • What do you think was the cause of these challenges? • How did it feel when you experienced these challenges? (if they have any)
How did you learn to use digital devices?	• Can you elaborate and give examples? • Have you ever had any assistance in the use of digital devices? If so, what type of assistance? • What would you have done if you could not get that assistance? Have you been offered any help to overcome these challenges? • Have you ever found it difficult to talk with someone about challenges in the use of digital devices? • If you got the opportunity to improve your digital skills, is this something you would consider? Why or why not? • Have you ever been offered courses or classes in the use of digital devices?
How is digital practice part of your everyday life?	• Have you ever experienced having trouble doing something digitally, you thought could easily do? • Have you ever experienced learning or doing something on a digital device you thought you never would? • Which type of emotions would you describe as connected to the use of digital devices?
Questions regarding digital practices in the health system	
Please tell me how you experience the use of digital technology in the health system.	• Have you heard of sundhed.dk, and if so, do you use it and for what purpose? • Have you heard of MinSP (app) and, if so, do you use it and for what purpose? • Do you think these services could be made more accessible or relevant to you and, if so, how?
With what purpose do you use digital services in the health system? (if they do)	• For example, video consulting, answering PRO[a], and ordering vaccinations

XSL•FO
RenderX

Main questions	Supporting questions, if needed
How do you communicate with institutions in the health system?	• Have you ever had to respond to something from the hospital digitally? • Do you use digital devices to communicate in any form with any part of the health system? • Do you read results and invitations from the hospital digitally on a computer or smartphone or tablet? • Do you use any other digital services in the public health system?
How do you experience the accessibility of these digital Health services?	• Do you think these services could be made more accessible or relevant to you and, if so, how? • Do you have any advice for people who do not find it easy to use digital devices? (if they are experienced)

[a]PRO: patient-related outcome.

Data Analysis

The text from each transcribed interview was analyzed using manifest content analysis, as described by Graneheim and Lundman [24]. Content analysis is a method of analyzing written, verbal, and visual communication in a systematic way [25]. Qualitative content analysis is a structured but nonlinear process that requires researchers to move back and forth between the original text and parts of the text during the analysis. Manifest analysis is the descriptive level at which the surface structure of the text central to the phenomenon and the research question is described. The analysis was conducted as a collaborative effort between the first (CGJ) and last authors (MIL); hence, in this inductive circular process, to achieve consistency in the interpretation of the text, there was continued discussion and reflection between the researchers. The transcriptions were initially read several times to gain a sense of the whole context, and we analyzed each interview. The text was initially divided into domains that reflected the lowest degree of interpretation, as a rough structure was created in which the text had a specific area in common. The structure roughly reflected the interview guide's themes, as guided by Graneheim and Lundman [24]. Thereafter, the text was divided into meaning units, condensed into text-near descriptions, and then abstracted and labeled further with codes. The codes were categorized based on similarities and differences. During this process, we discussed the findings to reach a consensus on the content, resulting in the final 4 categories presented in this paper.

Ethical Considerations

The interviewees received oral and written information about the study and its voluntary nature before the interviews. Written informed consent was obtained from all participants. Participants were able to opt of the study at any time. Data were anonymized and stored electronically on locked and secured servers. The Ethics Committee of the Capitol Region in Denmark was contacted before the start of the study. This study was registered and approved by the ethics committee and registered under the Danish Data Protection Agency (number P2021-839). Furthermore, the ethical principles of the Declaration of Helsinki were followed for this study.

Results

The analysis provided insights into 4 different categories regarding digital practices and experiences of using digital tools and services in health care systems: social resources as a digital lifeline, possessing the necessary capabilities, big feelings as facilitators or barriers, and life without digital tools.

Social Resources as a Digital Lifeline

Throughout the analysis, it became evident that access to both material and social resources was of great importance when using digital tools. Most participants already possessed and had easy access to a computer, smartphone, or tablet. The few participants who did not own the necessary digital tools told us that they did not have the skills needed to use these tools. For these participants, the lack of material resources was tied particularly to a lack of knowledge and know-how, as they expressed that they would not know where to start after buying a computer—how to set it up, connect it to the internet, and use its many systems.

However, possessing the necessary material resources did not mean that the participants possessed the knowledge and skill to use digital tools. Furthermore, access to material resources was also a question of having access to assistance when needed. Some participants who had access to a computer, smartphone, and tablet and knew how to use these tools still had to obtain help when setting up hardware, updating software, or getting a new device. These participants were confident in their own ability to use digital devices but also relied on family, friends, and neighbors in their everyday use of these tools. Certain participants were explicitly aware of their own use of social resources when expressing their thoughts on digital services in health care systems:

> *I think it is a blessing and a curse. I think it is both. I would say that if I did not have someone around me in my family who was almost born into the digital world, then I think I would be in trouble. But I feel sorry for those who do not have that opportunity, and I know quite a few who do not. They get upset, and it's really frustrating.* [Woman, age 82 years]

The participants' use of social resources indicates that learning skills and using digital tools are not solely individual tasks but rather continuously involve engagement with other people, particularly whenever a new unforeseen problem arises or when

XSL•FO
RenderX

the participants want a deeper understanding of the tools they are using:

> *If tomorrow I have to get a new ipad...and it was like that when I got this one, then I had to get XXX to come and help me move stuff and he was sweet to help with all the practical stuff. I think I would have cursed a couple of times (if he hadn't been there), but he is always helpful, but at the same time he is also pedagogic so I hope that next time he showed me something I will be able to do it.* [Man, age 71 years]

For some participants, obtaining assistance from a more experienced family member was experienced as an opportunity to learn, whereas for other participants, their use of public digital services was even tied directly to assistance from a spouse or family member:

> *My wife, she has access to mine, so if something comes up, she can just go in and read, and we can talk about it afterwards what (it is).* [Man, age 85 years]

The participants used social resources to navigate digital systems and understand and interpret communication from the health care system through digital devices. Another example of this was the participants who needed assistance to find, answer, and understand questionnaires from the health care department. Furthermore, social resources were viewed as a support system that made participants feel more comfortable and safer when operating digital tools. The social resources were particularly important when overcoming unforeseen and new challenges and when learning new skills related to the use of digital tools. Participants with physical, cognitive, and communicative challenges also explained how social resources were of great importance in their ability to use digital tools.

Possessing the Necessary Capabilities

The findings indicated that possessing the desire and knowing how to use digital tools are not always enough to engage with digital services successfully. Different health issues can carry consequences for motor skills and mobility. Some of these consequences were visibly affecting how our participants interacted with digital devices, and these challenges were somewhat easy to discover. However, our participants revealed hidden challenges that posed difficulties. In some specific cases, cognitive and communicative inabilities can make it difficult to use digital tools, and this might not always be clear until the individual tries to use a device's more complex functions. An example of this is that some participants found it easy to turn on a computer and use it to write but difficult to go through security measures on digital services or interpret and understand digital language. Remembering passwords and logging on to systems created challenges, particularly for those experiencing health issues that directly affect memory and cognitive abilities, who expressed concerns about what they were able to do through digital tools:

> *I think it is very challenging because I would like to use it how I used to before my stroke; (I) wish that everything (digital skills) was transferred, but it just isn't.* [Man, age 80 years]

Despite these challenges, the participants demonstrated great interest in using digital tools, particularly regarding health care services and their own well-being. However, sometimes, the challenges that they experienced could not be conquered merely by motivation and good intentions. Another aspect of these challenges was the amount of extra time and energy that the participants had to spend on digital services. A patient diagnosed with Parkinson disease described how her symptoms created challenges that changed her digital practices:

> *Well it could for example be something like following a line in the device. And right now it is very limited what I can do with this (iPhone). Now I am almost only using it as a phone, and that is a little sad because I also like to text and stuff, but I also find that difficult (...) I think it is difficult to get an overview.* [Woman, age 62 years]

Some participants said that after they were discharged from the hospital, they did not use the computer anymore because it was too *difficult* and too *exhausting*, which contributed to them *giving up*. Using digital tools already demanded a certain amount of concentration and awareness, and some diseases and health conditions affected these abilities further.

Big Feelings as Facilitators or Barriers

The findings revealed a wide range of digital practices in which digital tools were used as a communication device, as an entertainment device, and as a practical and informative tool for ordering medicine, booking consultations, asking health-related questions, or receiving email from public institutions. Despite these different digital practices, repeating patterns and arguments appeared when the participants were asked why they learned to use digital tools or wanted to improve their skills. A repeating argument was that they wanted to "follow the times," or as a participant who was still not satisfied with her digital skills stated:

> *We should not go against the future.* [Woman, age 89 years]

The participants expressed a positive view of the technological developments and possibilities that digital devices offered, and they wanted to improve their knowledge and skills related to digital practice. For some participants, this was challenging, and they expressed frustration over how technological developments "moved too fast," but some participants interpreted these challenges as a way to "keep their mind sharp."

Another recurring pattern was that the participants expressed great interest in using digital services related to the health care system and other public institutions. The importance of being able to navigate digital services was explicitly clear when talking about finding test answers, written electronic messages, and questionnaires from the hospital or other public institutions. Keeping up with developments, communicating with public institutions, and taking an interest in their own health and well-being were described as good reasons to learn to use digital tools.

However, other aspects also affected these learning facilitators. Some participants felt alienated while using digital tools and described the practice as something related to feelings of

XSL•FO
RenderX

anxiety, fear, and stupidity as well as something that demanded "a certain amount of courage." Some participants felt frustrated with the digital challenges they experienced, especially when the challenges were difficult to overcome because of their physical conditions:

> *I get sad because of it (digital challenges) and I get very frustrated and it takes a lot of time because I have difficulty seeing when I look away from the computer and have to turn back again to find out where I was and continue there (...) It pains me that I have to use so much time on it.* [Man, age 71 years]

Fear of making mistakes, particularly when communicating with public institutions, for example, the health care system, was a common pattern. Another pattern was the fear of misinterpreting the sender and the need to ensure that the written electronic messages were actually from the described sender. Some participants felt that they were forced to learn about digital tools because they cared a lot about the services. Furthermore, fears of digital services replacing human interaction were a recurring concern among the participants. Despite these initial and recurring feelings, some participants learned how to navigate the digital services that they deemed relevant. Another recurring pattern in this learning process was repetition, the practice of digital skills, and consistent assistance from other people. One participant expressed the need to use the services often to remember the necessary skills:

> *Now I can figure it out because now I've had it shown 10 times. But then three months still pass... and then I think...how was it now? Then I get sweat on my forehead (feel nervous) and think; I'm not an idiot.* [Woman, age 82 years]

For some participants, learning how to use digital tools demanded time and patience, as challenges had to be overcome more than once because they reappeared until the use of digital tools was more automatized into their everyday lives. Using digital tools and health services was viewed as easier and less stressful when part of everyday routines.

Life Without Digital Tools: Not a Free Choice

Even though some participants used digital tools daily, other participants expressed that it was "too late for them." These participants did not view it as a free choice but as something they had to accept that they could not do. They wished that they could have learned it earlier in life but did not view it as a possibility in the future. Furthermore, they saw potential in digital services, including digital health care services, but they did not know exactly what services they were missing out on. Despite this lack of knowledge, they still felt sad about the position they were in. One participant expressed what she thought regarding the use of digital tools in public institutions:

> *Well, I feel alright about it, but it is very, very difficult for those of us who do not have it. Sometimes you can feel left out—outside of society. And when you do not have one of those (computers)...A reference is always made to w and w (www.) and then you can read on. But you cannot do that.* [Woman, age 94 years]

The feeling of being left out of society was consistent among the participants who did not use digital tools. To them, digital systems seemed to provide unfair treatment based on something outside of their own power. Participants who were heavily affected by their medical conditions and could not use digital services also felt left out because they saw the advantages of using digital tools. Furthermore, a participant described the feelings connected to the use of digital tools in public institutions:

> *It is more annoying that it does not seem to work out in my favour.* [Woman, age 62 years]

These statements indicated that it is possible for individuals to want to use digital tools and simultaneously find them too challenging. These participants were aware that there are consequences of not using digital tools, and that saddens them, as they feel like they are not receiving the same treatment as other people in society and the health care system.

Discussion

Principal Findings

The insights from our findings demonstrated that our participants had different digital practices and different experiences with digital tools and services; however, the analysis also highlighted patterns related to how digital services and tools were used. Specific conditions were important for the possibility of digital practice, including *having access to social resources; possessing the necessary capabilities;* and *feeling motivated, secure, and comfortable*. These prerequisites were necessary to have positive experiences using digital tools in the health care system, although some participants who lived up to these prerequisites were still skeptical toward digital solutions. Others who did not live up to these prerequisites experienced challenges and even though they were aware of opportunities, this awareness made them feel left out. A few participants even viewed the digital tools as a threat to their participation in society. This supports the notion of Norgaard et al [13] that the attention paid to digital capability demands from eHealth systems is very important. Furthermore, our findings supported the argument of Hjeltholt and Papazu [17] that it is important to better understand experiences related to digital services. In our study, we accommodate this request and bring forth a broad perspective on experiences with digital practices; we particularly add insight into the challenges with digital practices for patients who also have acute or chronic illness, with some of them also experiencing physical, communicative, and cognitive difficulties. To our knowledge, there is limited existing literature focusing on digital practices that do not have a limited scope, for example, a focus on perspectives on eHealth literacy in the use of apps [26] or intervention studies with a focus on experiences with digital solutions, for example, telemedicine during the COVID-19 pandemic [27]. As mentioned by Hjeltholt et al [10], certain citizens are dependent on their own social networks in the process of using and learning digital tools. Rasi et al [28] and Airola et al [29] argued that digital health literacy is situated and should include the capabilities of the individual's social network. Our findings support these arguments that access to social resources is an important condition; however, the

XSL•FO
RenderX

findings also highlight that these resources can be particularly crucial in the use of digital health services, for example, when interpreting and understanding digital and written electronic messages related to one's own health course or when dealing with physical, cognitive, and communicative disadvantages. Therefore, we argue that the awareness of the disadvantages is important if we want to understand patients' digital capabilities, and the inclusion of the next of kin can be evident in unveiling challenges that are unknown and not easily visible or when trying to reach patients with digital challenges through digital means.

Studies by Kayser et al [30] and Kanoe et al [31] indicated that patients' abilities to interpret and understand digital health–related services and their benefits are important for the successful implementation of eHealth services—an argument that our findings support. Health literacy in both digital and physical contexts is important if we want to understand how to better design and implement services. Our participants' statements support the argument that communication through digital means cannot be viewed as similar to face-to-face communication and that an emphasis on digital health literacy demonstrates how health systems are demanding different capabilities from the patients [13]. We argue that it is important to communicate the purposes of digital services so that both the patient and their next of kin know why they participate and how it can benefit them. Therefore, it is important to make it as clear as possible that digital health services can benefit the patient and that these services are developed to support information, communication, and dialogue between patients and health professionals. However, our findings suggest that even after interpreting and understanding the purposes of digital health services, some patients may still experience challenges when using digital tools.

Therefore, it is important to understand how and why patients learn digital skills, particularly because both experience with digital devices and estimation of the value of digital tools have been highlighted as key factors for digital practices [5,18]. Our findings indicate that a combination of these factors is important, as recognizing the value of digital tools was not enough to facilitate the necessary learning process for some of our participants. Instead, our participants described the use of digital tools as complex and continuous processes in which automation of skills, assistance from others, and time to relearn forgotten knowledge were necessary and important facilitators for learning and understanding digital tools as well as becoming more comfortable and confident in the use of digital health services. This was particularly important, as it was more encouraging for our participants to learn digital tools when they felt secure, instead of feeling afraid and anxious, a point that Bailey et al [18] also highlighted. The value of digital solutions and the will to learn were greater when challenges were viewed as something to overcome and learn from instead of something that created a feeling of being stupid. This calls for attention on how to simplify and explain digital tools and services so that users do not feel alienated. Our findings also support the argument that digital health literacy should take into account emotional well-being related to digital practice [32].

The various perspectives that our participants provided regarding the use of digital tools in the health care system indicate that patients are affected by the use of digital health services and their own capabilities to use digital tools. Murray et al [33] argued that the use of digital tools in health sectors has the potential to improve health and health delivery by improving efficacy, efficiency, accessibility, safety, and personalization, and our participants also highlighted these positive aspects. However, different studies found that some patients, particularly older adults considered socially vulnerable, have lower digital health literacy [10,34,35], which is an important determinant of health and may widen disparities and inequity in health care [16]. Studies on older adult populations' adaptation to information and communication technology show that engaging with this technology can be limited by the usability of technology, feelings of anxiety and concern, self-perception of technology use, and the need for assistance and inclusive design [36]. Our participants' experiences with digital practices support the importance of these focus areas, especially when primarily older patients are admitted to hospitals. Furthermore, our findings indicate that some older patients who used to view themselves as being engaged in their own health care felt more distanced from the health care system because of digital services, and some who did not have the capabilities to use digital tools felt that they were treated differently compared to the rest of society. They did not necessarily view themselves as vulnerable but felt vulnerable in the specific experience of trying to use digital services because they wished that they were more capable. Moreover, this was the case for patients with physical and cognitive difficulties, as they were not necessarily aware of the challenges before experiencing them. Drawing on the phenomenological and feministic approach by Ahmed [37], these challenges that make patients feel vulnerable are not necessarily visible to others but can instead be viewed as invisible institutional "walls" that do not present themselves before the patient runs into them. Some participants had to experience how their physical, cognitive, or communicative difficulties affected their digital practice to realize that they were not as digitally capable as they once were or as others in society. Furthermore, viewed from this perspective, our findings could be used to argue that digital capabilities should be viewed as a privilege tied to users' physical bodies and that digital services in the health care system are indirectly making patients without this privilege vulnerable. This calls for more attention to the inequities that digital tools and services create in health care systems and awareness that those who do not use digital tools are not necessarily indifferent about the consequences. Particularly, in a context such as the Danish one, in which the digital strategy is to create an intertwined and user-friendly public digital sector for everyone, it needs to be understood that patients have different digital capabilities and needs. Although some have not yet had a challenging experience that made them feel vulnerable, others are very aware that they receive different treatment and feel that they are on their own or that the rest of the society does not care about them. Inequities in digital health care, such as these, can and should be mitigated or prevented, and our investigation into the experiences with digital practices can help to show that we are creating standards and infrastructures that deliberately exclude the perspectives of those

XSL•FO
RenderX

who are most in need of the services offered by the digital health care system [8]. Therefore, our findings support the notions that flexibility is important in the implementation of universal public digital services [17]; that it is important to adjust systems in accordance with patients' eHealth literacy and not only improve the capabilities of individuals [38]; and that the development and improvement of digital health literacy are not solely an individual responsibility but are also tied to ways in which institutions organize, design, and implement digital tools and services [39].

Limitations

This qualitative study provided novel insights into the experiences with public digital health services from the perspective of patients in the Danish context, enabling a deeper understanding of how digital health services and digital tools are experienced and used. This helps build a solid foundation for future interventions aimed at digital health literacy and digital health interventions. However, this study has some limitations. First, the study was conducted in a country where digitalization is progressing quickly, and people, therefore, are accustomed to this pace. Therefore, readers must be aware of this. Second, the study included patients with different neurological conditions; some of their digital challenges were caused or worsened by these neurological conditions and are, therefore, not applicable to all patients in the health system. However, the findings provided insights into the patients' digital practices before their conditions and other challenges not connected to neurological conditions shared by patients. Third, the study was broad, and although a large number of informants was included, from a qualitative research perspective, we would recommend additional research in this field to develop interventions that target digital health literacy and the use of digital health services.

Conclusions

Experiences with digital tools and digital health services are complex and multifaceted. The advantages in communication, finding information, or navigating through one's own health course work as facilitators for engaging with digital tools and digital health services. However, this is not enough on its own. Furthermore, feeling secure and motivated and having time to relearn and practice skills are important facilitators. Engagement in digital practices for the examined population requires access to continuous assistance from their social network. If patients do not meet requirements, digital health services can be experienced as exclusionary and a source of concern. Physical, cognitive, and communicative difficulties might make it impossible to use digital tools or create more challenges that require assistance. Digitalization of the health care system means that patients do not have the choice to opt out of using digital services without having consequences, resulting in them receiving a different treatment than others. To ensure digitalization does not create inequities in health, it is necessary for developers and the health institutions that create, design, and implement digital services to be aware of differences in digital health literacy and to focus on simplifying communication with patients and next of kin through and about digital services. It is important to focus on helping individuals meet the necessary conditions and finding flexible solutions for those who do not have the same privileges as others if the public digital sector is to work for everyone.

Acknowledgments

The authors would like to thank all the people who gave their time to be interviewed for the study, the clinical nurse specialists who facilitated interviewing patients, and the other nurses on shift who assisted in recruiting participants.

Conflicts of Interest

None declared.

References

1. Year in search 2022. Google Trends. URL: https://trends.google.com/trends/yis/2022/DK/ [accessed 2024-04-02]
2. Digital government index: 2019. Organisation for Economic Cooperation and Development. URL: https://www.oecd-ilibrary.org/content/paper/4de9f5bb-en [accessed 2024-04-02]
3. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019 Oct 01;29(Supplement_3):13-17 [FREE Full text] [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
4. Digital practices. Umeå University. URL: https://www.umu.se/en/humlab/research/digital-practice/ [accessed 2024-04-02]
5. It-anvendelse i befolkningen 2020. Danmarks Statistik. URL: https://www.dst.dk/da/Statistik/nyheder-analyser-publ/Publikationer/VisPub?cid=29450 [accessed 2024-04-02]
6. Sundhed.dk homepage. Sundhed.dk. URL: https://www.sundhed.dk/borger/ [accessed 2024-04-02]
7. Nøhr C, Bertelsen P, Vingtoft S, Andersen SK. Digitalisering af Det Danske Sundhedsvæsen. Odense, Denmark: Syddansk Universitetsforlag; 2019.
8. Eriksen J, Ebbesen M, Eriksen KT, Hjermitslev C, Knudsen C, Bertelsen P, et al. Equity in digital healthcare - the case of Denmark. Front Public Health 2023 Sep 6;11:1225222 [FREE Full text] [doi: 10.3389/fpubh.2023.1225222] [Medline: 37744503]
9. Digital health strategy. Sundhedsdatastyrelsen. URL: https://sundhedsdatastyrelsen.dk/da/english/digital_health_solutions/digital_health_strategy [accessed 2024-04-02]

XSL•FO
RenderX

10. Hjelholt M, Schou J, Bojsen LB, Yndigegn SL. Digital marginalisering af udsatte ældre: arbejdsrapport 2. IT-Universitetet i København. 2018. URL: https://egv.dk/images/Projekter/Projekter_2018/EGV_arbejdsrapport_2.pdf [accessed 2024-04-02]
11. Sørensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12(1):80 [FREE Full text] [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
12. Improving health literacy. World Health Organization. URL: https://www.who.int/activities/improving-health-literacy [accessed 2024-04-02]
13. Norgaard O, Furstrand D, Klokker L, Karnoe KA, Batterham R, Kayser L, et al. The e-health literacy framework: a conceptual framework for characterizing e-health users and their interaction with e-health systems. Knowl Manag E Learn 2015;7(4). [doi: 10.34105/j.kmel.2015.07.035]
14. Kramer JM, Schwartz A. Reducing barriers to patient-reported outcome measures for people with cognitive impairments. Arch Phys Med Rehabil 2017 Aug;98(8):1705-1715. [doi: 10.1016/j.apmr.2017.03.011] [Medline: 28400180]
15. Menger F, Morris J, Salis C. Aphasia in an internet age: wider perspectives on digital inclusion. Aphasiology 2016;30(2-3):112-132. [doi: 10.1080/02687038.2015.1109050]
16. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
17. Hjelholt M, Papazu I. "De har fået NemID, men det er ikke nemt for mig" - Digital rum(me)lighed i den danske velfærdsstat. Social Kritik 2021;2021-2(163) [FREE Full text]
18. Bailey C, Sheehan C. Technology, older persons' perspectives and the anthropological ethnographic lens. Alter 2009;3(2):96-109. [doi: 10.1016/j.alter.2009.01.002]
19. Gadamer HG, Weinsheimer HG, Marshall DG. Truth and Method. New York, NY: Crossroad Publishing Company; 1991.
20. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec 16;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
21. Polit DF, Beck CT. Nursing Research: Generating and Assessing Evidence for Nursing Practice. Philadelphia, PA: Lippincott Williams & Wilkins, Inc; 2012.
22. Kvale S, Brinkmann S. InterViews: Learning the Craft of Qualitative Research Interviewing. Thousand Oaks, CA: SAGE Publications; 2009.
23. Kagan A. Supported conversation for adults with aphasia: methods and resources for training conversation partners. Aphasiology 1998 Sep;12(9):816-830. [doi: 10.1080/02687039808249575]
24. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
25. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: SAGE Publications; 1980.
26. Klösch M, Sari-Kundt F, Reibnitz C, Osterbrink J. Patients' attitudes toward their health literacy and the use of digital apps in health and disease management. Br J Nurs 2021 Nov 25;30(21):1242-1249. [doi: 10.12968/bjon.2021.30.21.1242] [Medline: 34839691]
27. Datta P, Eiland L, Samson K, Donovan A, Anzalone AJ, McAdam-Marx C. Telemedicine and health access inequalities during the COVID-19 pandemic. J Glob Health 2022 Dec 03;12:05051 [FREE Full text] [doi: 10.7189/jogh.12.05051] [Medline: 36462207]
28. Rasi P, Lindberg J, Airola E. Older service users' experiences of learning to use eHealth applications in sparsely populated healthcare settings in Northern Sweden and Finland. Educ Gerontol 2020 Nov 24;47(1):25-35. [doi: 10.1080/03601277.2020.1851861]
29. Airola E, Rasi P, Outila M. Older people as users and non-users of a video conferencing service for promoting social connectedness and well-being – a case study from Finnish Lapland. Educ Gerontol 2020 Mar 29;46(5):258-269. [doi: 10.1080/03601277.2020.1743008]
30. Kayser L, Kushniruk A, Osborne RH, Norgaard O, Turner P. Enhancing the effectiveness of consumer-focused health information technology systems through eHealth literacy: a framework for understanding users' needs. JMIR Hum Factors 2015 May 20;2(1):e9 [FREE Full text] [doi: 10.2196/humanfactors.3696] [Medline: 27025228]
31. Karnoe A, Furstrand D, Christensen KB, Norgaard O, Kayser L. Assessing competencies needed to engage with digital health services: development of the ehealth literacy assessment toolkit. J Med Internet Res 2018 May 10;20(5):e178 [FREE Full text] [doi: 10.2196/jmir.8347] [Medline: 29748163]
32. Nielsen AS, Hanna L, Larsen BF, Appel CW, Osborne RH, Kayser L. Readiness, acceptance and use of digital patient reported outcome in an outpatient clinic. Health Informatics J 2022 Jun 03;28(2):14604582221106000 [FREE Full text] [doi: 10.1177/14604582221106000] [Medline: 35658693]
33. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]

XSL•FO
RenderX

34. Chesser A, Burke A, Reyes J, Rohrberg T. Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. Inform Health Soc Care 2016 Feb 24;41(1):1-19. [doi: 10.3109/17538157.2014.948171] [Medline: 25710808]
35. Chesser AK, Keene Woods N, Smothers K, Rogers N. Health literacy and older adults: a systematic review. Gerontol Geriatr Med 2016 Mar 15;2:2333721416630492 [FREE Full text] [doi: 10.1177/2333721416630492] [Medline: 28138488]
36. Mitra S, Singh A, Rajendran Deepam S, Asthana MK. Information and communication technology adoption among the older people: a qualitative approach. Health Soc Care Community 2022 Nov 21;30(6):e6428-e6437. [doi: 10.1111/hsc.14085] [Medline: 36269139]
37. Ahmed S. How not to do things with words. Wagadu. 2016. URL: https://sites.cortland.edu/wagadu/wp-content/uploads/sites/3/2017/02/v16-how-not-to-do-ahmed.pdf [accessed 2024-04-02]
38. Monkman H, Kushniruk AW. eHealth literacy issues, constructs, models, and methods for health information technology design and evaluation. Knowl Manag E Learn 2015;7(4). [doi: 10.34105/j.kmel.2015.07.036]
39. Brørs G, Norman CD, Norekvål TM. Accelerated importance of eHealth literacy in the COVID-19 outbreak and beyond. Eur J Cardiovasc Nurs 2020 Aug 15;19(6):458-461 [FREE Full text] [doi: 10.1177/1474515120941307] [Medline: 32667217]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by A Mavragani; submitted 14.03.23; peer-reviewed by G Myreteg, J Eriksen, M Siermann; comments to author 18.09.23; revised version received 09.10.23; accepted 27.02.24; published 11.04.24.

Please cite as:
Gybel Jensen C, Gybel Jensen F, Loft MI
Patients' Experiences With Digitalization in the Health Care System: Qualitative Interview Study
J Med Internet Res 2024;26:e47278
URL: https://www.jmir.org/2024/1/e47278
doi:10.2196/47278
PMID:38602748

©Christian Gybel Jensen, Frederik Gybel Jensen, Mia Ingerslev Loft. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Serbian Version of the eHealth Literacy Questionnaire (eHLQ): Translation, Cultural Adaptation, and Validation Study Among Primary Health Care Users

Branko Vujkovic[1*], MD, MHM; Voin Brkovic[2*], MD, ASS; Ana Pajičić[1*], MD; Vedrana Pavlovic[2*], MD, PhD; Dejana Stanisavljevic[2*], Prof Dr Med; Dušanka Krajnović[3*], Prof Dr; Aleksandra Jovic Vranes[2*], Prof Dr Med

[1]Institute of Public Health of Sabac, Sabac

[2]Medical Faculty, University of Belgrade, Belgrade

[3]Faculty of Pharmacy, University of Belgrade, Belgrade

[*]all authors contributed equally

Corresponding Author:
Branko Vujkovic, MD, MHM
Institute of Public Health of Sabac
1 Jovana Cvijica
Sabac, 15000
Phone: 381 648623647
Fax: 381 15343606
Email: vujkovicb@gmail.com

Abstract

Background: As digital health services are increasingly developing and becoming more interactive in Serbia, a comprehensive instrument for measuring eHealth literacy (EHL) is needed.

Objective: This study aimed to translate, culturally adapt, and investigate the psychometric properties of the Serbian version of the eHealth Literacy Questionnaire (eHLQ); to evaluate EHL in the population of primary health care (PHC) users in Serbia; and to explore factors associated with their EHL.

Methods: The validation study was conducted in 8 PHC centers in the territory of the Mačva district in Western Serbia. A stratified sampling method was used to obtain a representative sample. The Translation Integrity Procedure was followed to adapt the questionnaire to the Serbian language. The psychometric properties of the Serbian version of the eHLQ were analyzed through the examination of factorial structure, internal consistency, and test-retest reliability. Descriptive statistics were calculated to determine participant characteristics. Differences between groups were tested by the 2-tailed Students *t* test and ANOVA. Univariable and multivariable linear regression analyses were used to determine factors related to EHL.

Results: A total of 475 PHC users were enrolled. The mean age was 51.0 (SD 17.3; range 19-94) years, and most participants were female (328/475, 69.1%). Confirmatory factor analysis validated the 7-factor structure of the questionnaire. Values for incremental fit index (0.96) and comparative fit index (0.95) were above the cutoff of ≥0.95. The root mean square error of approximation value of 0.05 was below the suggested value of ≤0.06. Cronbach α of the entire scale was 0.95, indicating excellent scale reliability, with Cronbach α ranging from 0.81 to 0.90 for domains. The intraclass correlation coefficient ranged from 0.63 to 0.82, indicating moderate to good test-retest reliability. The highest EHL mean scores were obtained for the *understanding of health concepts and language* (mean 2.86, SD 0.32) and *feel safe and in control* (mean 2.89, SD 0.33) domains. Statistically significant differences (all $P<.05$) for all 7 eHLQ scores were observed for age, education, perceived material status, perceived health status, searching for health information on the internet, and occupation (except domain 4). In multivariable regression models, searching for health information on the internet and being aged younger than 65 years were associated with higher values of all domain scores except the domain *feel safe and in control* for variable age.

Conclusions: This study demonstrates that the Serbian version of the eHLQ can be a useful tool in the measurement of EHL and in the planning of digital health interventions at the population and individual level due to its strong psychometric properties in the Serbian context.

(J Med Internet Res 2024;26:e57963) doi:10.2196/57963

XSL•FO
RenderX

KEYWORDS

eHealth; digital health; eHLQ; eHealth Literacy Questionnaire; digital health literacy; primary healthcare; Serbia; questionnaire; technology; communication

Introduction

Background

The increasing digitization of the health system requires the users of health services to have sufficient competence in the use of digital health (DH) technologies. The use of DH technologies for health, also called eHealth, has brought a revolution in the way we diagnose and treat patients and take care of health. eHealth is also defined as "the use of information and communication technologies in support of health and health-related fields" [1].

eHealth is often presented as a possible solution for numerous challenges in the health system, including a lack of health personnel, an aging population, and the comorbidities of several chronic diseases [2].

Digital technologies can strengthen health systems, improve health financing and public health, and also increase the availability of health services to vulnerable groups. eHealth and the data obtained in this way are particularly useful for the prevention and control of chronic noncommunicable diseases and care for the young, as well as for older adults and for preparing a better response to future emergencies and health risks caused by climate change [3]. Digital technologies can provide a new tool for educating patients and improving their health literacy [4]. Instead of being a passive participant in health care, eHealth solutions allow individuals to be active participants [5].

It is estimated that about half of all patients search for information about their health problems on the internet before seeking professional advice [6]. It is believed that 6 out of 10 Europeans use the internet to solve health concerns, and 9 out of 10 trust the information they find [7]. This information can be wrong or incomplete, and if not critically evaluated, can lead to the adoption of behaviors that can potentially harm our health [8].

Today, digital spaces have become a central environment for communication and engagement, learning, and work, as well as for disease prevention and health promotion. This has led to the development of a new dimension of health literacy: the emergence of eHealth literacy (EHL). For the adequate implementation of eHealth services, the users of these services must be eHealth literate [9].

EHL can be defined as "the ability to search, find, understand, and evaluate health information from electronic sources and apply the acquired knowledge to approaching or solving health problems" [10]. Good EHL is directly related to improving health outcomes, reducing health care costs, increasing users' motivation to seek health information, having better knowledge about chronic diseases, adopting preventive health behaviors, and having better self-perception and care for one's health [6-8].

As a superdeterminant of health, EHL has added significant complexity to the way users of health system services, health workers, and digital technologies interact. Health portals and telehealth systems have enabled service users to communicate remotely with service providers; cloud-based electronic health records have enabled patients to manage diagnostic data with clinicians; and wearable devices and apps have enabled users to self-manage their conditions. The increased complexity of interacting in the digital world requires additional skills and competencies from people using eHealth. However, with the increased complexity of the DH landscape, scholars have called for a more comprehensive view and have included elements related to users' cognitive skills, communication elements, social and cultural context, or system-level attributes [11-13]. Some researchers talk about DH literacy as an evolved concept of EHL and its impact on health, but if we take into account the existence of analog computing, EHL is definitely a more correct term, and as a relatively new area, it needs further research [14].

The COVID-19 pandemic has made it clear how important it is to have access to digital platforms. Digital technologies, such as mobile phones, have made it possible to quickly trace contacts, check symptoms, seek advice, obtain necessary information, and engage in public communication and education. Access to digital spaces is particularly important when mobility is limited or when people live in rural or remote areas. Compared to traditional communication strategies, digital spaces support "accessibility and expansion of access to health information to diverse population groups, regardless of age, education, race or ethnicity, and location" and may encourage further development of health literacy [15,16]. On the other hand, a meta-analysis by Estrela et al [17] showed that EHL is significantly related to sociodemographic, economic, and cultural factors.

The rapid transition to online health services and digital communication with health professionals due to the COVID-19 pandemic has accelerated the level of digital exclusion in certain population groups, such as those with low levels of electronic literacy or insufficient access to digital devices [18].

The benefits of digitization in health care may be unused due to unequal opportunities to use digital resources [19]. Recent studies have shown that EHL interventions have a positive effect on the health and health care of older adults, and therefore it is important to provide support and guidance to older adults to narrow the aging technology gap [20,21].

Previous Work

Regardless of the population of interest, the need for reliable internet navigation is particularly important for health issues where the consequences of using poor quality, misleading, or false information are high [22]. By providing tools and resources to evaluate online health information and critically evaluate eHealth resources, we offer an opportunity to protect consumers from harm while empowering them [12,13]. To provide relevant

XSL•FO
RenderX

tools to help users navigate eHealth, an understanding of what health care users have at the outset, or their EHL, is required.

EHL is generally lower among older adults with chronic noncommunicable diseases and is characterized by reluctance to change their usual health care routines and concerns that mobile health apps will replace doctors' visits. On the other hand, a study by Kouri et al [22] showed that mobile health innovations have the potential to help the older adult population manage chronic diseases more effectively.

In the era of digitization of health systems, EHL is a significant predictor of an individual's health condition. As a determinant of health, EHL is important in the analysis of the health system of each country. Previous research has shown that a satisfactory level of EHL is needed for citizens to actively participate in making correct health decisions and participating in health care [23].

Study in the field of EHL has attracted the attention of a significant number of researchers, both in the field of health care and in other areas of public life. A low level of EHL exposes both the individual and an entire society to loss (health, economic, and social loss). Strategies to strengthen EHL should be developed as part of lifelong learning skills, and health care professionals should embrace improving EHL as part of regular patient care activities [24].

To develop and implement strategies to strengthen EHL, an adequate instrument for measuring EHL is needed, which also allows evaluation of the effects of implementing those strategies on increasing EHL at the individual and population levels [25]. The Serbian government prioritizes promoting health care and citizen health through digitization of services, as stated in the "digitalization program in the health care system of the Republic of Serbia" [26]. This includes connecting facilities with advanced software for secure data exchange and enabling consumers to use information and communication technologies for health support, but it is also necessary to empower users of health services to "use information and communication technologies in support of health and health-related fields" [1,26]. As eHealth services are increasingly developing and becoming more interactive in Serbia, and everywhere in the world, a comprehensive instrument for measuring EHL is necessary [27].

The eHealth Literacy Questionnaire (eHLQ) is currently licensed for use in more than 12 countries, and its ongoing translations and cultural adaptations indicate that the instrument is robust across various contexts [28]. However, the instrument has not yet been translated into Serbian.

Goal of This Study

This study aimed to translate, culturally adapt, and investigate the psychometric properties of a Serbian version of the eHLQ; to evaluate EHL in the population of primary health care (PHC) users in Serbia; and to explore factors associated with EHL.

Methods

Study Design

This was a cross-sectional study conducted from April 1 to April 30, 2023, in 8 state-owned PHC centers in the territory of the Mačva district in Western Serbia. Both qualitative and quantitative research methods were used in this study.

Data Collection

The participants of this study were PHC consumers who were recruited from PHC centers waiting rooms in 8 local municipalities of the Mačva administrative district in Western Serbia, were aged older than 19 years, and had adequate cognitive abilities to fill out the questionnaire independently. We used a stratified sampling method. In the first step, the variable for stratification was the type of settlement. By dividing into urban and rural settlements, 2 strata were obtained. In the second step, the variable for stratification was the age of the respondents, and by dividing them into young, middle-aged, and older respondents, 6 strata were obtained. The selection of respondents was carried out proportionally to the size of the stratum. The required number of respondents to examine the psychometric properties of the EHL eHLQ was 475, as defined by the questionnaire's creator [28]. The criteria for exclusion from the research were respondents aged younger than 19 years, a health condition that prevented a respondent from filling out the questionnaire, if the Serbian language was not a respondent's first language, illiterate persons, and refusal of participation by a respondent.

Further, 3 trained examiners delivered paper-based questionnaires to all consumers of PHC who met the criteria for inclusion in this study at the PHC centers' waiting rooms. Participants had to fill out the questionnaire at the PHC center. Since the topic of our study included digital use and literacy competencies, we chose to consistently administer paper-based questionnaires.

The data collection consisted of administering the Serbian eHLQ along with asking general questions about owning a digital device and seeking health information on the internet.

We also collected demographic data on participants' age, sex, education, work situation, and health status.

Ethical Considerations

Ethical approval was granted by the Ethical Review Board at the Medical Faculty, University of Belgrade (17/IX-5) and the Ethics Committees of PHC institutions. Participation was voluntary and anonymous. All participants gave informed consent to participate in this study, and all data were anonymized. No compensation was offered to the respondents. All steps were managed according to the ethical principles described in the Helsinki Declaration [29].

About eHLQ

The eHLQ contains 35 items on 7 scales representing the EHL framework domains: using technology to process health information, understanding health concepts and language, ability to actively engage with digital services, feeling safe and in

XSL•FO
RenderX

control, being motivated to engage with digital services, access to digital services that work, and digital services that suit individual needs. Each scale consists of 4 to 6 items on a 4-point Likert scale ranging from "strongly disagree" to "strongly agree." Scale scores range from 1 to 4, calculated on an index by averaging item scores within each scale with equal weighting. Each scale is presented separately, and no overall eHLQ score is calculated. Higher scores indicate higher abilities [28].

Translation, Cultural Adaptation, and Pretesting

The license to translate the English version of the eHLQ to Serbian (TE2203IG) was obtained from Swinburne University, Australia. As required by the eHLQ developers, we used the Translation Integrity Procedure (TIP) to maintain equivalence between the original (English) and translated (Serbian) versions of the instrument while ensuring the linguistic and cultural appropriateness of the Serbian version. The process was further facilitated by using clear "item intent" descriptions [30,31].

One of the eHLQ's developers (Lars Kayser) chaired the TIP process. The translation and adaptation team included 2 native Serbian forward translators (AM and RJ), 1 native English back translator (DZ), a cognitive interviewer team, and academic professionals (BV, VB, AJV, and AP), all fluent in Serbian and English. All 3 translators had excellent English and Serbian language skills. The translation and cultural adaptation process involved 3 steps (the first 2 steps were forward translation and item intent description, and the third was back translation). (1) The original English eHLQ questionnaire was translated independently into Serbian by the 2 forward translators. (2) The translators then used the item intent descriptions, which thoroughly explained the intent of each item and scale, as a guide when synthesizing their translations. During the following team discussion, the best statements for each item were chosen and combined to form the first version of the Serbian eHLQ. (3) The first version of the Serbian eHLQ was back-translated by a native English-speaking translator who had never seen the original version of the instrument. The Serbian-to-English back-translation was then compared with the English version of the eHLQ, and the items were discussed by the team to achieve consensus on the preferred version of the Serbian eHLQ. The preferred Serbian eHLQ was then tested using cognitive interviews.

Cognitive interviewing is valuable for ensuring accurate interpretations of items when translating and validating a questionnaire in another language and culture. The process of cognitive interviewing allows researchers to discover and correct items that are not interpreted as intended, thereby avoiding the future collection of inaccurate data. Cognitive interviewing does not require a large sample size, but the sample should represent demographic variety [31,32].

Cognitive interviews were conducted with 20 adults (12 female individuals) aged 27 to 63 (median 50, IQR 37-59) years, with varying educational backgrounds. The respondents were given a printed version of the questionnaire and were carefully observed while answering the items. The interviewer (BV or AP) then went through each item with the individual respondents, focusing on items the respondents appeared to find difficult. The main questions were as follows: "what were you thinking about when you were answering that item?" and "can you tell me why you selected that answer?" Participants were encouraged to elaborate on their interpretations of the items. A protocol was used for making notes during the interviews, which were also recorded, transcribed, and analyzed using a text summary [31].

Results from the cognitive interviews revealed that although most items were understood as intended, minor revisions were needed to clarify a few items and instructions. The following corrections were made throughout the Serbian eHLQ:

- The Serbian term "organizujem informacije o svom zdravlju" (organize my health information) was replaced by "upravljam informacijama o svom zdravlju" (manage information about my health) in item 25.
- The Serbian term "u sisteme zdravstvenih tehnologija" (into health technology systems) was replaced by "u digitalne zdravstvene sisteme" (into eHealth systems) in item 8.

When we reached an agreement on all formulations, the final version of the Serbian eHLQ was considered ready to be psychometrically tested.

Statistical Analysis

Descriptive statistics were calculated to determine participant characteristics. The psychometric properties of the Serbian version of the eHLQ were analyzed through the examination of factorial structure and internal consistency (reliability). Confirmatory factor analysis (CFA) was performed to confirm the original 7D structure of the eHLQ. The absolute goodness-of-fit of the 7D model was evaluated using the χ^2 test (P values that are <.05 signify that a model may be a bad fit for the data, whereas values >.05 may render the model a good fit) and 3 additional fit measures: the comparative fit index (CFI), the incremental fit index (IFI), and the root mean square error of approximation (RMSEA). Values of CFI and IFI above 0.95 were considered adequate, whereas the RMSEA value of 0.05 was below the suggested value of ≤0.06 indicating an acceptable model fit. The CFA was conducted using Amos 21 (IBM SPSS Inc). To measure reliability, we assessed internal consistency and test-retest reliability. The internal consistency of the eHLQ was evaluated using the Cronbach α coefficient (ranges from 0 to 1, with the latter meaning perfect reliability). Test-retest reliability was evaluated using the intraclass correlation coefficient (ICC). Differences between groups were tested by the 2-tailed Students t test and ANOVA. Univariable and multivariable linear regression analyses were used to determine factors related to EHL. The results were expressed as linear regression coefficients (β) and their 95% CIs. All tests were 2-tailed. P<.05 was considered statistically significant. The IBM SPSS (version 21) package was used for these analyses.

Results

Demographic Characteristics of Participants

The Serbian version of the eHLQ questionnaire was completed by 475 PHC users. The mean age of the participants was 51.0 (SD 17.3; range 19-94) years, and most participants were female (328/475, 69.1%). The majority of the participants were married

XSL•FO
RenderX

(339/475, 71.4%) and had completed secondary education or higher (409/475, 86.1%). More than half of the participants (278/474, 58.6%) were employees, and 57.1% (266/466) of the sample did not have any longstanding illnesses. The perceived material status was bad in 10.5% (49/466), regular in 50.4% (235/466), and good in 39.1% (182/466). One-half (235/466, 50.4%) of the participants perceived their health as good to excellent, 35.2% (164/466) as regular, and 14.4% (67/466) as bad. A high percentage of participants owned digital devices (431/475, 90.7%), but 28.5% (133/466) did not search for any web-based health information. The participants' demographics are summarized in Table 1.

Descriptive statistics of eHLQ domain scores are presented in Table 2.

The highest EHL mean scores were obtained for the *understanding of health concepts and language* (mean 2.86, SD 0.32) and *feel safe and in control* (mean 2.89, SD 0.33) domains, while the lowest values were for *using technology to process health information* (mean 2.51, SD 0.33), *digital services that suit individual needs* (mean 2.55, SD 0.27), and *access to digital services that work* (mean 2.57, SD 0.45) domains.

XSL•FO
RenderX

Table 1. Participant demographics (N=475).

Variables	Participants, n (%)
Age (years)	
<65	343 (72.2)
≥65	132 (27.8)
Sex	
Male	147 (30.9)
Female	328 (69.1)
Place of living	
City	313 (65.9)
Other	162 (34.1)
Education	
Incomplete secondary education	66 (13.9)
Complete secondary education	263 (55.4)
Higher education	146 (30.7)
Marital status	
Married	339 (71.4)
Single, separated, or widowed	136 (28.6)
Occupation	
Employed	278 (58.6)
Unemployed	196 (41.4)
Perceived material status	
Very bad or bad	49 (10.5)
Neutral	235 (50.4)
Good or very good	182 (39.1)
Chronic disease	
Yes	200 (42.9)
No	266 (57.1)
Perceived health status	
Very bad or bad	67 (14.4)
Neutral	164 (35.2)
Good or very good	235 (50.4)
Digital device	
Yes	431 (90.7)
No	44 (9.3)
Searching for health information on the internet	
Never	133 (28.5)
Rarely	271 (58.2)
Frequently	62 (13.3)

XSL•FO
RenderX

Table 2. eHLQ[a] domain scores, internal consistency, and test-retest reliability.

Domains	Score, mean (SD)	Cronbach α	ICC[b] (95% CI)
1. *Using technology* to process health information	2.51 (0.33)	0.90	0.63 (0.22-0.82)
2. Understanding of *health concepts* and language	2.86 (0.32)	0.81	0.79 (0.57-0.90)
3. Ability to actively *engage* with digital services	2.71 (0.35)	0.90	0.82 (0.62-0.91)
4. *Feel safe* and in control	2.89 (0.33)	0.82	0.81 (0.61-0.91)
5. *Motivated* to engage with digital services	2.59 (0.41)	0.85	0.65 (0.26-0.83)
6. *Access* to digital services that work	2.57 (0.45)	0.83	0.73 (0.43-0.87)
7. Digital services that *suit* individual *needs*	2.55 (0.27)	0.89	0.81 (0.60-0.91)

[a]eHLQ: eHealth Literacy Questionnaire.

[b]ICC: intraclass correlation coefficient.

Psychometric Properties

The 7-factor structure of the eHLQ has been validated with maximum likelihood confirmatory analysis, and the results demonstrated a good fit of the data to the hypothesized 7-factor model. The χ^2 test rejected the 7D model (χ^2=1001.9, P<.001), as we expected, due to the large sample size. Values for IFI (0.96) and CFI (0.95) were above the cutoff of ≥0.95. The RMSEA value of 0.05 was below the suggested value of ≤0.06. All standardized factor loadings were statistically significant (all P<.05) and ranged from 0.43 to 0.86 (see Figure 1).

Analysis of the internal consistency of the Serbian version of the eHLQ showed that the Cronbach α of the entire scale (items 1-35) was 0.95, indicating excellent scale reliability. The α coefficients of the 7 domains were estimated to be 0.90 for domain 1, estimated to be 0.81 for domain 2, estimated to be 0.90 for domain 3, estimated to be 0.82 for domain 4, estimated to be 0.85 for domain 5, estimated to be 0.83 for domain 6, and estimated to be 0.89 for domain 7. For the test-retest, 30 participants completed the retest, and the ICC ranged from 0.63 to 0.82, indicating moderate to good test-retest reliability. Test-retest reliability showed moderate agreement for 2 domains (ICC 0.63 and 0.65) and good agreement for 5 domains (ICC from 0.73 to 0.82; Table 2).

Statistically significant differences (all P<.05) for all 7 eHLQ scores were observed for age, education, perceived material status, perceived health status, searching for health information on the internet, and occupation (except domain 4). Participants aged younger than 65 years scored higher in every eHLQ domain when compared with those aged older than 65 years. Female participants had higher scores than male participants in *using technology to process health information* and the *ability to actively engage with digital service* domains. No significant difference was obtained concerning marital status (all P>.05). Participants with completed secondary education or higher showed a higher level of EHL compared to those with incomplete secondary education. Employed participants scored higher in every eHLQ domain when compared with the unemployed, except for domain 4. Participants with better-perceived material and health status showed a higher level of EHL (see Table 3).

In multivariable regression models, searching for health information on the internet and aged younger than 65 years were associated with higher values of all domain scores except the domain *feel safe and in control* for variable age (Table 4). Secondary education or higher was positively associated with the domains of *understanding of health concepts and language* and *the ability to actively engage with digital services.* Domains of *ability to actively engage with digital services,motivation to engage with digital services, access to digital services that work,* and *digital services that suit individual needs* were all negatively affected by chronic disease. Having a material status that was considered to be good or very good was associated with higher values of domains of *understanding of health concepts and language, feeling safe and in control, access to digital services that work,* and *digital services that suit individual needs.* Place of living and marital status were found to be negatively correlated to the *feel safe and in control* domain.

XSL•FO
RenderX

Figure 1. Standardized factor loadings of the 7-factor structure of the eHLQ. eHLQ: eHealth Literacy Questionnaire.

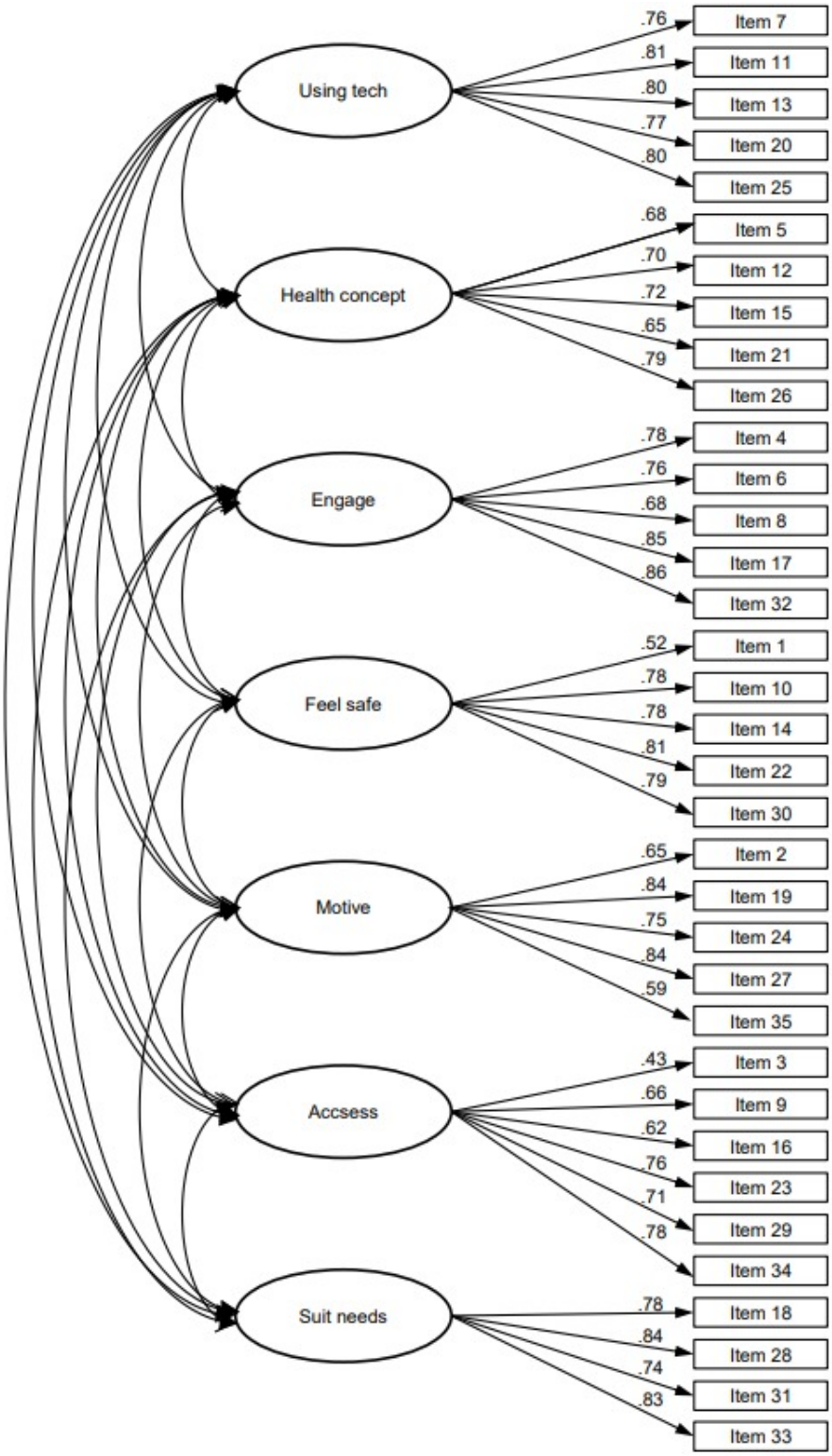

XSL•FO
RenderX

Table 3. Associations among sociodemographic factors and eHLQ[a] domain scores.

Variables	Using tech	Health concepts	Engage	Feel safe	Motive	Access	Suit needs
Age (years), mean (SD)							
<65	2.66 (0.58)	2.95 (0.44)	2.96 (0.52)	2.93 (0.51)	2.70 (0.56)	2.70 (0.51)	2.71 (0.57)
≥65	2.12 (0.61)	2.61 (0.44)	2.08 (0.57)	2.78 (0.49)	2.31 (0.50)	2.24 (0.46)	2.11 (0.50)
P value (*t* test)	*<.001*[b]	*<.001*	*<.001*	*.003*	*<.001*	*<.001*	*<.001*
Sex, mean (SD)							
Male	2.42 (0.64)	2.82 (0.43)	2.55 (0.67)	2.88 (0.55)	2.54 (0.58)	2.50 (0.54)	2.50 (0.61)
Female	2.55 (0.63)	2.88 (0.48)	2.79 (0.65)	2.89 (0.48)	2.61 (0.57)	2.60 (0.53)	2.58 (0.61)
P value (*t* test)	*.04*	.21	*<.001*	.79	.27	.07	.10
Place of living, mean (SD)							
City	2.55 (0.63)	2.88 (0.45)	2.77 (0.65)	2.85 (0.49)	2.58 (0.57)	2.56 (0.55)	2.58 (0.62)
Other	2.43 (0.65)	2.81 (0.48)	2.60 (0.68)	2.96 (0.53)	2.59 (0.57)	2.59 (0.62)	2.49 (0.60)
P value (*t* test)	.051	.12	*.007*	*.03*	.85	.52	.13
Education, mean (SD)							
Incomplete secondary education	2.45 (0.68)	2.60 (0.53)	2.22 (0.71)	2.77 (0.47)	2.38 (0.55)	2.37 (0.55)	2.22 (0.69)
Complete secondary education	2.55 (0.64)	2.89 (0.46)	2.71 (0.64)	2.94 (0.52)	2.64 (0.59)	2.61 (0.54)	2.59 (0.60)
Higher education	2.56 (0.57)	2.91 (0.40)	2.93 (0.56)	2.85 (0.49)	2.60 (0.53)	2.59 (0.50)	2.62 (0.55)
P value (ANOVA)	*.001*	*<.001*	*<.001*	*.03*	*.005*	*.004*	*<.001*
Marital status, mean (SD)							
Married	2.53 (0.61)	2.85 (0.47)	2.74 (0.64)	2.86 (0.51)	2.58 (0.57)	2.57 (0.53)	2.54 (0.61)
Single, separated, or widowed	2.47 (0.69)	2.87 (0.45)	2.66 (0.73)	2.96 (0.49)	2.61 (0.59)	2.58 (0.56)	2.55 (0.64)
P value (*t* test)	.36	.67	.24	.06	.61	.76	.82
Occupation , mean (SD)							
Employed	2.64 (0.56)	2.94 (0.44)	2.97 (0.49)	2.92 (0.52)	2.68 (0.56)	2.70 (0.50)	2.72 (0.55)
Unemployed	2.33 (0.69)	2.74 (0.47)	2.35 (0.72)	2.85 (0.49)	2.46 (0.57)	2.40 (0.53)	2.31 (0.63)
P value (*t* test)	*<.001*	*<.001*	*<.001*	.12	*<.001*	*<.001*	*<.001*
Perceived material status, mean (SD)							
Very bad or bad	2.20 (0.62)	2.53 (0.49)	2.22 (0.59)	2.59 (0.57)	2.34 (0.60)	2.23 (0.48)	2.15 (0.55)
Neutral	2.44 (0.63)	2.82 (0.45)	2.68 (0.66)	2.88 (0.47)	2.51 (0.57)	2.54 (0.54)	2.50 (0.63)
Good or very good	2.65 (0.62)	2.30 (0.43)	2.89 (0.64)	2.98 (0.52)	2.73 (0.54)	2.69 (0.52)	2.69 (0.57)
P value (ANOVA)	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*
Perceived health status, mean (SD)							
Very bad or bad	2.31 (0.71)	2.70 (0.52)	2.26 (0.69)	2.81 (0.51)	2.43 (0.57)	2.36 (0.51)	2.31 (0.58)
Neutral	2.40 (0.64)	2.76 (0.50)	2.57 (0.64)	2.83 (0.52)	2.49 (0.61)	2.46 (0.55)	2.38 (0.61)
Good or very good	2.63 (0.59)	2.98 (0.39)	2.94 (0.59)	2.95 (0.49)	2.70 (0.53)	2.70 (0.50)	2.72 (0.59)
P value (ANOVA)	*<.001*	*<.001*	*<.001*	*.02*	*<.001*	*<.001*	*<.001*
Searching for health information on the internet, mean (SD)							
Never	1.96 (0.56)	2.60 (0.48)	2.14 (0.67)	2.78 (0.53)	2.23 (0.55)	2.28 (0.53)	2.14 (0.62)
Rarely	2.67 (0.52)	2.96 (0.40)	2.94 (0.51)	2.93 (0.50)	2.70 (0.53)	2.67 (0.48)	2.69 (0.54)
Frequently	2.92 (0.54)	2.95 (0.51)	2.96 (0.54)	2.96 (0.50)	2.82 (0.49)	2.72 (0.54)	2.76 (0.57)
P value (ANOVA)	*<.001*	*<.001*	*<.001*	*.008*	*<.001*	*<.001*	*<.001*

XSL•FO
RenderX

[a]eHLQ: eHealth Literacy Questionnaire.
[b]Significant *P* values are italicized.

Table 4. Regression models of variables associated with eHLQ[a] domain scores.

eHLQ domains	β (95% CI)	*P* value
Using technology		
Searching for health information on the internet	0.46 (0.37 to 0.54)	*<.001*[b]
Age (years)	0.29 (0.17 to 0.40)	*<.001*
Health concepts		
Age (years)	0.21 (0.11 to 0.30)	*<.001*
Searching for health information on the internet	0.13 (0.06 to 0.20)	*<.001*
Material status	0.19 (0.06 to 0.33)	*.006*
Education	0.15 (0.02 to 0.27)	*.02*
Engage		
Age (years)	0.61 (0.49 to 0.73)	*<.001*
Searching for health information on the internet	0.29 (0.21 to 0.37)	*<.001*
Chronic disease	–0.17 (–0.27 to –0.06)	*.001*
Education	0.20 (0.06 to 0.35)	*.005*
Feel safe		
Material status	0.33 (0.18 to 0.48)	*<.001*
Place of living	–0.15 (–0.24 to –0.05)	*.003*
Searching for health information on the internet	0.10 (0.03 to 0.18)	*.005*
Marital status	–0.13 (–0.23 to –0.03)	*.01*
Motive		
Searching for health information on the internet	0.28 (0.20 to 0.36)	*<.001*
Chronic disease	–0.14 (–0.25 to –0.03)	*.01*
Age (years)	0.16 (0.04 to 0.29)	*.01*
Access		
Age (years)	0.28 (0.16 to 0.40)	*<.001*
Searching for health information on the internet	0.16 (0.08 to 0.24)	*<.001*
Chronic disease	–0.11 (–0.21 to –0.01)	*.04*
Material status	0.16 (0.01 to 0.31)	*.04*
Suit needs		
Age (years)	0.39 (0.26 to 0.52)	*<.001*
Searching for health information on the internet	0.23 (0.15 to 0.32)	*<.001*
Chronic disease	–0.13 (–0.24 to –0.02)	*.02*
Material status	0.17 (0.003 to 0.34)	*.046*

[a]eHLQ: eHealth Literacy Questionnaire.
[b]Significant *P* values are italicized.

Discussion

Principal Findings

One of the primary goals of the digitalization program in the Republic of Serbia's health care system is to establish and support the development of a health system that places the patient at the center (also known as a "patient-centered approach"). More than 80 patient associations, organizations, and associations were contacted during the preparation of this document; they acknowledged the benefits of digitization, but they also identified several needs and difficult aspects, such as the low levels of the EHL of their members, as well as all patients in Serbia [26].

XSL•FO
RenderX

In this study, we used a well-defined translation and cultural adaptation process to reproduce the original instrument's concepts and meanings. We evaluated the psychometric properties of the Serbian version of the eHLQ in the population of PHC users and explored factors associated with EHL. Our data from a sample of PHC users from urban and rural municipalities demonstrated that the Serbian eHLQ has strong psychometric properties and is in line with the psychometric outcomes of the versions in English and other languages [1,27,33,34].

Based on previous recommendations for examining the validity of a translated instrument, this study used qualitative and quantitative approaches. During the first phase, we adopted the TIP, which includes a multistep translation and review process and detailed item intent descriptions [30,31]. The results of the cognitive interviews and several review board meetings resulted in our revising a few words that were considered problematic in a Serbian context; however, most items on the Serbian eHLQ were understood as intended, and their equivalence to the original and translated versions was maintained [28].

The overall findings of this study provide evidence for the good validity and reliability of the Serbian eHLQ. CFA validated the 7-factor structure of the questionnaire. According to the preestablished CFI, IFI, and RMSEA thresholds, the original 7D model can be considered acceptable. Only the χ^2 test revealed a bad fit for the 7D model analyzed due to the large sample size. All standardized factor loadings were statistically significant and ranged from 0.43 to 0.86. These findings are per previous validation studies conducted in other populations that confirmed the 7-factor structure of the questionnaire [1,27,28,34].

Analysis of the internal consistency of the Serbian version of the eHLQ showed that the Cronbach α of the entire scale was 0.95, indicating excellent scale reliability. All scales demonstrated good internal consistency, with a Cronbach α of >0.80, and ICC ranged from 0.63 to 0.82, indicating moderate to good test-retest reliability. The lowest value was for domain 2, *understanding of health concepts and language* [27,28,33], which was consistent with Danish, Australian, Taiwanese, and Swedish eHLQ validation studies.

The majority of the participants in our study were married (339/475, 71.4%) and had completed secondary education or higher (409/475, 86.1%). More than half of the participants were employed (278/474, 58.6%) and did not have any longstanding illnesses (266/466, 57.1%). Every 10th PHC user (49/466, 10.5%) in our sample had a bad perceived material status, while half (235/466, 50.4%) of the participants considered their health to be good to excellent. A high percentage (431/475, 90.7%) of participants owned digital devices, but more than a quarter (133/466, 28.5%) of the sample did not search for any web-based health information, showing that patients with limited use of eHealth were well represented in this study's sample.

The highest EHL mean scores were obtained for domain 2, *understanding of health concepts and language*, and domain 4, *feel safe and in control*, while the lowest values were for domain 1, *using technology to process health information;* domain 7, *digital services that suit individual needs;* and domain 6, *access to digital services that work*. The lowest scores were obtained for domains that depend mostly on interaction with DH services and the accessibility of technology. This finding concurred with the results of a study by García-García et al [8].

Previous studies have already shown that EHL significantly depends on sociodemographic factors [7,8,17]. In our study, statistically significant differences for all 7 eHLQ scores were observed for age, education, perceived material status, perceived health status, searching for health information on the internet, and occupation (except domain 4, *feel safe and in control*). Female participants had higher scores than male participants in domain 1, *using technology to process health information*, and domain 3, *ability to actively engage with digital services.* Literature data on sex influencing EHL was inconclusive because some studies have positioned female sex as a protective factor while other studies have reported higher scores in male participants [8]. Based on previous research on EHL in Serbia, the female sex has been consistently associated with better EHL across age-specific populations, and female individuals tend to use more primary and specialized care compared to men [35].

Participants aged younger than 65 years scored higher in every eHLQ domain when compared with those aged older than 65 years. Age-specific results are similar in other studies, which have stated that older age is associated with decreased adoption and use of health care technologies and holds the most prejudice against them [1,8,36].

Participants with completed secondary education or higher showed a higher level of EHL compared to those with incomplete secondary education, as in other studies [1,7,33,36]. The population with completed secondary education or higher in our study had statistically significantly higher eHLQ scores, except for domain 4, *feel safe and in control*; domain 5, *motivated to engage with digital services*; and domain 6, *access to digital services that work*. People with lower education used eHealth less often [1,34,36]. However, having more education did not mean that the patients felt safer or had better access to eHealth. These results are consistent with some literature data, but the higher score could also be a result of a difference in interpretation between these 2 groups [1,8,33]. The Serbian version of the eHLQ could be a promising tool for understanding digital access at different educational levels, as in other contexts [34].

No significant difference was obtained concerning marital status despite the study of García-García et al [8], where those who were single, separated, or widowed scored significantly lower for domain 1, *using technology to process health information*; domain 4, *feel safe and in control*; domain 5, *motivated to engage with digital services*; and domain 6, *access to digital services that work*, indicating that people who are "alone" might face challenges in these areas.

Employed participants scored higher in every eHLQ domain when compared with the unemployed, except for domain 4. The majority of jobs in contemporary society require good digital skills [34].

XSL•FO
RenderX

In our study, searching for health information on the internet and those aged younger than 65 years were associated with higher values of all domain scores except domain 4, *feel safe and in control*, for variable age. Participants with better-perceived material and health status were found to have a positive association with better EHL, except for domain 4, *feel safe and in control*. Domain 3, *ability to actively engage with digital services*; domain 5, *motivated to engage with digital services*; domain 6, *access to digital services that work*; and domain 7, *digital services that suit individual needs* were all negatively affected by chronic disease. In line with the data from the literature, health status is one of the crucial determinants of health care technology adoption [8].

Although previous instruments, such as the Serbian version of the eHealth Literacy Scale, have focused on individuals' competencies, the eHLQ has the added perspective of interaction between the individual and the eHealth systems in Serbia [27,35].

Limitations

Considering that this was a cross-sectional study, it is impossible to determine causality. Only associations can be interpreted from this data. Collected data may also be subject to errors due to subjective reporting or the selective memory of respondents. This study included only PHC patients from the Mačva district of Western Serbia. Future testing of the Serbian eHLQ in different contexts in Serbia may strengthen the validity of the instrument.

Conclusions

This study provided evidence for the appropriate metric properties of the Serbian version of eHLQ. Searching for health information on the internet and age were factors influencing almost all scale domains. This study demonstrates that the Serbian version of the eHLQ can be a useful instrument in measuring EHL and in planning eHealth interventions at the population and individual levels. It is a useful tool for understanding the socioeconomic determinants of digital access inequity. The Serbian eHLQ can represent a basis for further research, and its results could establish complex connections with the way users of health system services, health workers, and digital technologies interact, which will help policymakers evaluate and implement new eHealth interventions.

Acknowledgments

The authors would like to thank all PHC patients and health care workers in the Mačva district who participated. The authors also thank Ana Milanko, Romina Jovanovic, and Djordje Zegarac for their work on the translation team, and Lars Kayser for chairing the research group meetings and providing key guidelines in the process of translation and cultural adaptation of the Serbian version of the eHealth Literacy Questionnaire (eHLQ). This study was financed by the Institute of Public Health of Sabac and supported by the Faculty of Medicine, Belgrade University (451-03-66/2024-03/200110).

Data Availability

About the Questionnaire Licence Agreement, Swinburne University of Technology is the owner of the license for use of the Serbian version of the eHealth Literacy Questionnaire (eHLQ), and if you would like to access this tool, please contact Ms Kerrie Paulger at kpaulger@swin.edu.au. All other data sets generated or analyzed during this study are not publicly available due to the Institute of Public Health of Sabac's policy but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Poot CC, Meijer E, Fokkema M, Chavannes NH, Osborne RH, Kayser L. Translation, cultural adaptation and validity assessment of the Dutch version of the eHealth Literacy Questionnaire: a mixed-method approach. BMC Public Health 2023;23(1):1006 [FREE Full text] [doi: 10.1186/s12889-023-15869-4] [Medline: 37254148]
2. From Innovation to Implementation: eHealth in the WHO European Region. Geneva, Switzerland: World Health Organization; 2016. URL: https://www.digitalhealthnews.eu/images/stories/pdf/From-Innovation-to-Implementation-eHealth-Report-EU.pdf [accessed 2024-04-20]
3. Prestidge M, Cafagna G, Dalton CW, de Oliviera T, Feliciano JC, Feng SC, et al. Digital in health: unlocking the value for everyone. The World Bank. 2023. URL: https://policycommons.net/artifacts/4776903/digital-in-health/5613181/ [accessed 2024-04-04]
4. Davaris MT, Bunzli S, Trieu J, Dowsey MM, Choong PF. The role of digital health interventions to improve health literacy in surgical patients: a narrative review in arthroplasty. ANZ J Surg 2022;92(10):2474-2486. [doi: 10.1111/ans.17931] [Medline: 35924880]
5. Busse TS, Nitsche J, Kernebeck S, Jux C, Weitz J, Ehlers JP, et al. Approaches to improvement of digital health literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022;19(14):8309 [FREE Full text] [doi: 10.3390/ijerph19148309] [Medline: 35886158]

XSL•FO
RenderX

6. Conard S. Best practices in digital health literacy. Int J Cardiol 2019;292:277-279. [doi: 10.1016/j.ijcard.2019.05.070] [Medline: 31230937]
7. Lwin MO, Panchapakesan C, Sheldenkar A, Calvert GA, Lim LKS, Lu J. Determinants of eHealth literacy among adults in China. J Health Commun 2020;25(5):385-393. [doi: 10.1080/10810730.2020.1776422] [Medline: 32552607]
8. García-García D, Bazán MJA, Pérez-Rivas FJ. Factors influencing eHealth literacy among Spanish primary healthcare users: cross-sectional study. Int J Environ Res Public Health 2022;19(23):15497 [FREE Full text] [doi: 10.3390/ijerph192315497] [Medline: 36497572]
9. Campanozzi LL, Gibelli F, Bailo P, Nittari G, Sirignano A, Ricci G. The role of digital literacy in achieving health equity in the third millennium society: a literature review. Front Public Health 2023;11:1109323 [FREE Full text] [doi: 10.3389/fpubh.2023.1109323] [Medline: 36891330]
10. Monkman H, Kushniruk AW. eHealth literacy issues, constructs, models, and methods for health information technology design and evaluation. Knowl Manag E-Learn 2015;7(4):541-549 [FREE Full text] [doi: 10.34105/j.kmel.2015.07.036]
11. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002;113(9):763-765 [FREE Full text] [doi: 10.1016/s0002-9343(02)01473-0] [Medline: 12517369]
12. Eysenbach G. Rating information on the internet can empower users to make informed decisions. BMJ 1999;319(7206):385-386 [FREE Full text] [doi: 10.1136/bmj.319.7206.385b] [Medline: 10435980]
13. Eysenbach G, Jadad AR. Evidence-based patient choice and consumer health informatics in the internet age. J Med Internet Res 2001;3(2):E19 [FREE Full text] [doi: 10.2196/jmir.3.2.e19] [Medline: 11720961]
14. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and e-health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023;20(2):1472 [FREE Full text] [doi: 10.3390/ijerph20021472] [Medline: 36674237]
15. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
16. Blandford A, Wesson J, Amalberti R, AlHazme R, Allwihan R. Opportunities and challenges for telehealth within, and beyond, a pandemic. Lancet Glob Health 2020;8(11):e1364-e1365 [FREE Full text] [doi: 10.1016/S2214-109X(20)30362-4] [Medline: 32791119]
17. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023;177:105124 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
18. Mee P, Gussy M, Huntley P. Digital exclusion as a barrier to accessing healthcare: a summary composite indicator and online tool to explore and quantify local differences in levels of exclusion. medRxiv Preprint posted online on August 8, 2023. [FREE Full text] [doi: 10.1101/2023.07.12.23292547]
19. Sundell E, Wångdahl J, Grauman Å. Health literacy and digital health information-seeking behavior—a cross-sectional study among highly educated Swedes. BMC Public Health 2022;22(1):2278 [FREE Full text] [doi: 10.1186/s12889-022-14751-z] [Medline: 36471284]
20. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023;25:e48166 [FREE Full text] [doi: 10.2196/48166] [Medline: 37379077]
21. Kim S, Chow BC, Park S, Liu H. The usage of digital health technology among older adults in Hong Kong and the role of technology readiness and eHealth literacy: path analysis. J Med Internet Res 2023;25:e41915 [FREE Full text] [doi: 10.2196/41915] [Medline: 37043274]
22. Kouri A, Gupta S, Straus SE, Sale JEM. Exploring the perspectives and experiences of older adults with asthma and chronic obstructive pulmonary disease toward mobile health: qualitative study. J Med Internet Res 2023;25:e45955 [FREE Full text] [doi: 10.2196/45955] [Medline: 37606961]
23. Abel T, McQueen D. Critical health literacy and the COVID-19 crisis. Health Promot Int 2020;35(6):1612-1613 [FREE Full text] [doi: 10.1093/heapro/daaa040] [Medline: 32239213]
24. Cui GH, Li SJ, Yin YT, Chen LJ, Li JQ, Liang FY, et al. The relationship among social capital, eHealth literacy and health behaviours in Chinese elderly people: a cross-sectional study. BMC Public Health 2021;21(1):45 [FREE Full text] [doi: 10.1186/s12889-020-10037-4] [Medline: 33407275]
25. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
26. Digitalization program in the healthcare system of the Republic of Serbia for the period 2022-2026. Government of Republic of Serbia. 2022 Feb 10. URL: https://tinyurl.com/3pzb7t5n [accessed 2022-10-07]
27. Sjöström AE, Hajdarevic S, Hörnsten Å, Kristjánsdóttir Ó, Castor C, Isaksson U. The Swedish version of the eHealth Literacy Questionnaire: translation, cultural adaptation, and validation study. J Med Internet Res 2023;25:e43267 [FREE Full text] [doi: 10.2196/43267] [Medline: 37043268]
28. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]

XSL•FO
RenderX

29. Rademakers J, Waverijn G, Rijken M, Osborne R, Heijmans M. Towards a comprehensive, person-centred assessment of health literacy: translation, cultural adaptation and psychometric test of the Dutch Health Literacy Questionnaire. BMC Public Health 2020;20(1):1850 [FREE Full text] [doi: 10.1186/s12889-020-09963-0] [Medline: 33267834]
30. Hawkins M, Cheng C, Elsworth GR, Osborne RH. Translation method is validity evidence for construct equivalence: analysis of secondary data routinely collected during translations of the Health Literacy Questionnaire (HLQ). BMC Med Res Methodol 2020;20(1):130 [FREE Full text] [doi: 10.1186/s12874-020-00962-8] [Medline: 32456680]
31. Beatty PC, Willis GB. Research synthesis: the practice of cognitive interviewing. Public Opin Q 2007;71(2):287-311 [FREE Full text] [doi: 10.1093/poq/nfm006]
32. World Medical Association. World Medical Association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194 [FREE Full text] [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
33. Cheng C, Elsworth GR, Osborne RH. Validity evidence of the eHealth Literacy Questionnaire (eHLQ) part 2: mixed methods approach to evaluate test content, response process, and internal structure in the Australian Community Health setting. J Med Internet Res 2022;24(3):e32777 [FREE Full text] [doi: 10.2196/32777] [Medline: 35258475]
34. Chen YC, Cheng C, Osborne RH, Kayser L, Liu CY, Chang LC. Validity testing and cultural adaptation of the eHealth Literacy Questionnaire (eHLQ) among people with chronic diseases in Taiwan: mixed methods study. J Med Internet Res 2022;24(1):e32855 [FREE Full text] [doi: 10.2196/32855] [Medline: 35044310]
35. Gazibara T, Cakic J, Cakic M, Pekmezovic T, Grgurevic A. eHealth and adolescents in Serbia: psychometric properties of eHeals questionnaire and contributing factors to better online health literacy. Health Promot Int 2019;34(4):770-778. [doi: 10.1093/heapro/day028] [Medline: 29800141]
36. Xesfingi S, Vozikis A. eHealth literacy: in the quest of the contributing factors. Interact J Med Res 2016;5(2):e16 [FREE Full text] [doi: 10.2196/ijmr.4749] [Medline: 27226146]

Abbreviations

CFA: confirmatory factor analysis
CFI: comparative fit index
DH: digital health
EHL: eHealth literacy
eHLQ: eHealth Literacy Questionnaire
ICC: intraclass correlation coefficient
IFI: incremental fit index
PHC: primary health care
RMSEA: root mean square error of approximation
TIP: Translation Integrity Procedure

Edited by G Eysenbach, S Ma; submitted 01.03.24; peer-reviewed by A Hosny, E Hernandez-Encuentra; comments to author 23.03.24; revised version received 03.04.24; accepted 09.04.24; published 09.05.24.

Please cite as:
Vujkovic B, Brkovic V, Pajičić A, Pavlovic V, Stanisavljevic D, Krajnović D, Jovic Vranes A
Serbian Version of the eHealth Literacy Questionnaire (eHLQ): Translation, Cultural Adaptation, and Validation Study Among Primary Health Care Users
J Med Internet Res 2024;26:e57963
URL: https://www.jmir.org/2024/1/e57963
doi:10.2196/57963
PMID:38722675

©Branko Vujkovic, Voin Brkovic, Ana Pajičić, Vedrana Pavlovic, Dejana Stanisavljevic, Dušanka Krajnović, Aleksandra Jovic Vranes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Spanish and Catalan Versions of the eHealth Literacy Questionnaire: Translation, Cross-Cultural Adaptation, and Validation Study

Eulàlia Hernández Encuentra[1,2], PhD; Noemí Robles[1], PhD; Ariadna Angulo-Brunet[2], PhD; David Cullen[3], MA; Ignacio del Arco[1], PhD

[1]eHealth Center, Universitat Oberta de Catalunya, Barcelona, Spain
[2]Faculty of Psychology and Education, Universitat Oberta de Catalunya, Barcelona, Spain
[3]Language Service, Universitat Oberta de Catalunya, Barcelona, Spain

Corresponding Author:
Eulàlia Hernández Encuentra, PhD
Faculty of Psychology and Education
Universitat Oberta de Catalunya
Campus UOC. Rambla Poblenou, 156
Barcelona, 08018
Spain
Phone: 34 933263898
Email: ehernandez@uoc.edu

Abstract

Background: The rise of digital health services, especially following the outbreak of COVID-19, has led to a need for health literacy policies that respond to people's needs. Spain is a country with a highly developed digital health infrastructure, but there are currently no tools available to measure digital health literacy fully. A well-thought-through questionnaire with strong psychometric properties such as the eHealth Literacy Questionnaire (eHLQ) is important to assess people's eHealth literacy levels, especially in the context of a fast-growing field such as digital health.

Objective: This study aims to adapt the eHLQ and gather evidence of its psychometric quality in 2 of Spain's official languages: Spanish and Catalan.

Methods: A systematic cultural adaptation process was followed. Data from Spanish-speaking (n=400) and Catalan-speaking (n=400) people were collected. Confirmatory factor analysis was used to confirm the previously established factor structure. For reliability, the Cronbach α and categorical ω were obtained for every subscale. Evidence of convergent and discriminant validity was provided through the correlation with the total score of the eHealth Literacy Scale. Evidence based on relations to other variables was evaluated by examining extreme values for educational level, socioeconomic level, and use of technology variables.

Results: Regarding the confirmatory factor analysis, the 7-factor correlated model and the 7 one-factor models had adequate goodness-of-fit indexes for both Spanish and Catalan. Moreover, measurement invariance was established between the Spanish and Catalan versions. Reliability estimates were considered adequate as all the scales in both versions had values of >0.80. For convergent and discriminant validity evidence, the eHealth Literacy Scale showed moderate correlation with eHLQ scales in both versions (Spanish: range 0.57-0.76 and $P<.001$; Catalan: range 0.41-0.64 and $P<.001$). According to the relationship with external variables, all the eHLQ scales in both languages could discriminate between the maximum and minimum categories in level of education, socioeconomic level, and level of technology use.

Conclusions: The Spanish and Catalan versions of the eHLQ appear to be psychometrically sound questionnaires for assessing digital health literacy. They could both be useful tools in Spain and Catalonia for researchers, policy makers, and health service managers to explore people's needs, skills, and competencies and provide interesting insights into their interactions and engagement regarding their own experiences with digital health services, especially in the context of digital health growth in Spain.

(J Med Internet Res 2024;26:e49227) doi:10.2196/49227

KEYWORDS

eHealth literacy; eHealth; digital health; health literacy; questionnaire; eHealth Literacy Questionnaire; eHLQ; validation

XSL•FO
RenderX

Introduction

Background

The COVID-19 pandemic has accelerated the development of digital health infrastructures, devices, and services. Different digital health solutions were used in different settings worldwide with different approaches and purposes regarding, for example, the dissemination of public information for prevention, the understanding of the disease, the generation of diagnostic tools, or the sharing of real-time data for surveillance and control of the spread of the infection [1]. This scenario provided an opportunity to develop and test the capacity of digital technologies to increase the quality of health services, which required a similar improvement in citizens' digital skills. This is reflected in the World Health Organization global strategy on digital health 2020-2025 [2], one of whose 4 strategic objectives is to advocate for people-centered digital health systems. To this end, it is necessary to promote adequate health literacy with regard to digital health services among users and health care professionals to encourage the adoption and active use of these services.

In line with the World Health Organization strategy, and as other countries such as Ireland, New Zealand, the United Kingdom, and the United States had previously done [3-6], the Spanish Ministry of Health established, as part of its Digital Strategy 2021-2026, that one of the strategic objectives would be the empowerment and involvement of people in their health care and disease control by facilitating the relationship with health services [7]. Within this framework, people's trust in the health system is established as the key element for the implementation of a digitalized and efficient health system [8,9].

In general terms, data from recent reports and surveys show that Spain can be considered one of the leading countries in the use of digital technologies related to health. This is also the case in Catalonia, a region in northeastern Spain with almost 10 million Catalan speakers [10]. According to Eurostat [11], almost 7 out of 10 Spanish citizens used the internet to seek health information in 2021 (European Union average 56%), which is close to the leading countries in the area (Finland: 80%; the Netherlands: 77%; Norway: 77%; Denmark: 75%). In fact, digital clinical appointments were already available in 40% of Spain in 2019 (51.7% in Catalonia), in second position behind Finland (53%). In addition, in 2021, digital medical records and electronic prescriptions were widely implemented (with different levels of access and services depending on the region). A European survey on digital health across Europe [8] showed that Spain outperforms the average of the countries surveyed in indicators such as digital clinical consultations, the use of mobile health apps, access to electronic medical records, and remote supervision for monitoring or sharing symptoms with health professionals, with only 25% of Spanish people not using any digital technology to manage their health in the last year (a proportion that rises to 36% in the rest of the world) [9]. However, the use of wearables for health monitoring together with digital access for mental health in Spain show lower indicators than those of the rest of the countries studied. These promising data are complemented by the positive attitude of Spanish citizens toward the adoption of digital health even among those who have not yet used it [8].

The development of digital health services must be accompanied by digital health literacy, skills, confidence, and a positive attitude toward the relevance of these services [12-15] so that they can respond to people's needs in an equitable and inclusive way. Understanding these needs has been difficult due to the lack of a rigorous theoretical framework for measuring digital health literacy [16]. In 2006, Norman and Skinner [12] pioneered the development of their eHealth Literacy Scale (eHEALS) to assess individuals' self-perceived ability to seek, find, understand, and evaluate health information from electronic sources and apply the knowledge gained to address a problem. Currently, there are different instruments available to measure digital health literacy [17]. However, the only tools currently available in Spanish are the eHEALS questionnaire [12] and the European Health Literacy Survey Questionnaire (HLS-EU-Q) [18], an instrument developed by the European Health Literacy Consortium [19]. However, the HLS-EU-Q is designed primarily to assess health literacy and not digital health literacy. The short version of the HLS-EU-Q (16 items) includes 3 items that assess the ability to evaluate reliability, protect oneself, and understand the health information available in the media (including on the internet). However, this is clearly insufficient in a digital health context. Similarly, as with the 3 items in the short version of the HLS-EU-Q, the eHEALS questionnaire does not capture the breadth of services and aspects that digital health now covers [17]. Moreover, neither instrument is available in Catalan.

To address the shortcomings of the eHEALS in moving beyond Health 1.0, the Digital Health Literacy Instrument was developed, a performance-based tool aiming to include the interactivity of the web [20]. However, it did not comprehensively capture the interactions that digital health entails, including aspects linked to individuals (information and knowledge about one's own health), the intersection between users and technologies (the feeling of being safe and in control and their motivation), and users' experience of the systems (that they work and are accessible and that they respond to users' needs). To cover all these aspects, the eHealth Literacy Questionnaire (eHLQ) was developed [21,22]. The eHLQ is an instrument with 35 items distributed in 7 scales that each assess a dimension of digital health literacy (Textbox 1).

XSL•FO
RenderX

Textbox 1. eHealth Literacy Questionnaire scales, description, and number of items.

Questionnaire overview

- Scale 1 (using technology to process health information; 5 items): being able to use technologies to read, write, remember, appraise, and apply health information
- Scale 2 (understanding of health concepts and language; 5 items): being able to know and understand how the body functions, the current health status, and risks
- Scale 3 (ability to actively engage with digital services; 5 items): being comfortable using digital services for handling information
- Scale 4 (feel safe and in control; 5 items): feeling confident about safety, access, and storage of personal health data
- Scale 5 (motivated to engage with digital services; 5 items): feeling that engaging in the use of digital services will be useful for health management
- Scale 6 (access to digital services that work; 6 items): having access to digital services that work when the users need them and in the way they expect them to work
- Scale 7 (digital services that suit individual needs; 4 items): having access to responsive and adaptable services so that they suit the users' needs

Each scale has between 4 and 6 items with 4 response options (*strongly disagree*, *disagree*, *agree*, and *strongly agree*) with an assigned value of 1 to 4. With the aim of describing the person's relationship with digital health services in a broad way, the questionnaire does not have a total score. This is because health literacy is a relational concept [23], and it is firmly rooted in a specific context, so it depends on the characteristics of each health system. Furthermore, to make it user-friendly, the questionnaire is designed to allow for modular use (by using isolated scales) depending on the aspects that are of interest to analyze in each context (eg, people's ability to use health services or the availability of services that respond to their needs). It is because of this descriptive intent that, rather than quantitative scores, the eHLQ analysis results in user profiles from which vignettes (scenarios) can be defined that exemplify people's needs, skills, and competencies and provide insights into their interactions and engagement regarding their own experiences with digital health services. These can be used by policy makers and health service managers to design and offer improved services.

The eHLQ was constructed simultaneously in Danish and English to minimize cultural references [22]. It has shown good internal consistency reliability, and beyond these 2 original versions, as far as the authors of this study are aware, there are 3 more published versions available today that have shown positive evidence of psychometric quality ([24] in Mandarin [25], Swedish [26], and Dutch), although other adaptations are in progress across Europe. The eHLQ has been used to assess health literacy in different groups [27-29] both in Denmark and Australia [27,30], and further extensions of it have been developed for application in specific groups, for example, for health professionals [31] or older people, to assess their readiness to use digital health services [32]. Therefore, it is constantly evolving and a suitable tool to be used in a context of digital health growth such as that in Spain.

Objectives

The main objective of this study was to undertake the translation and cultural adaptation of the eHLQ and gather evidence of psychometric quality of the eHLQ scores in a large sample of the general population in Spain, creating a simultaneous process for 2 of the official languages spoken in the country: Spanish and Catalan.

Methods

Study Design

This study involved the translation and cross-cultural adaptation of the eHLQ into Spanish and Catalan. A cross-sectional design was used to gather evidence of the psychometric quality of the Spanish and Catalan versions of the eHLQ (Figure 1).

XSL•FO
RenderX

Figure 1. Schematic overview of the study design.

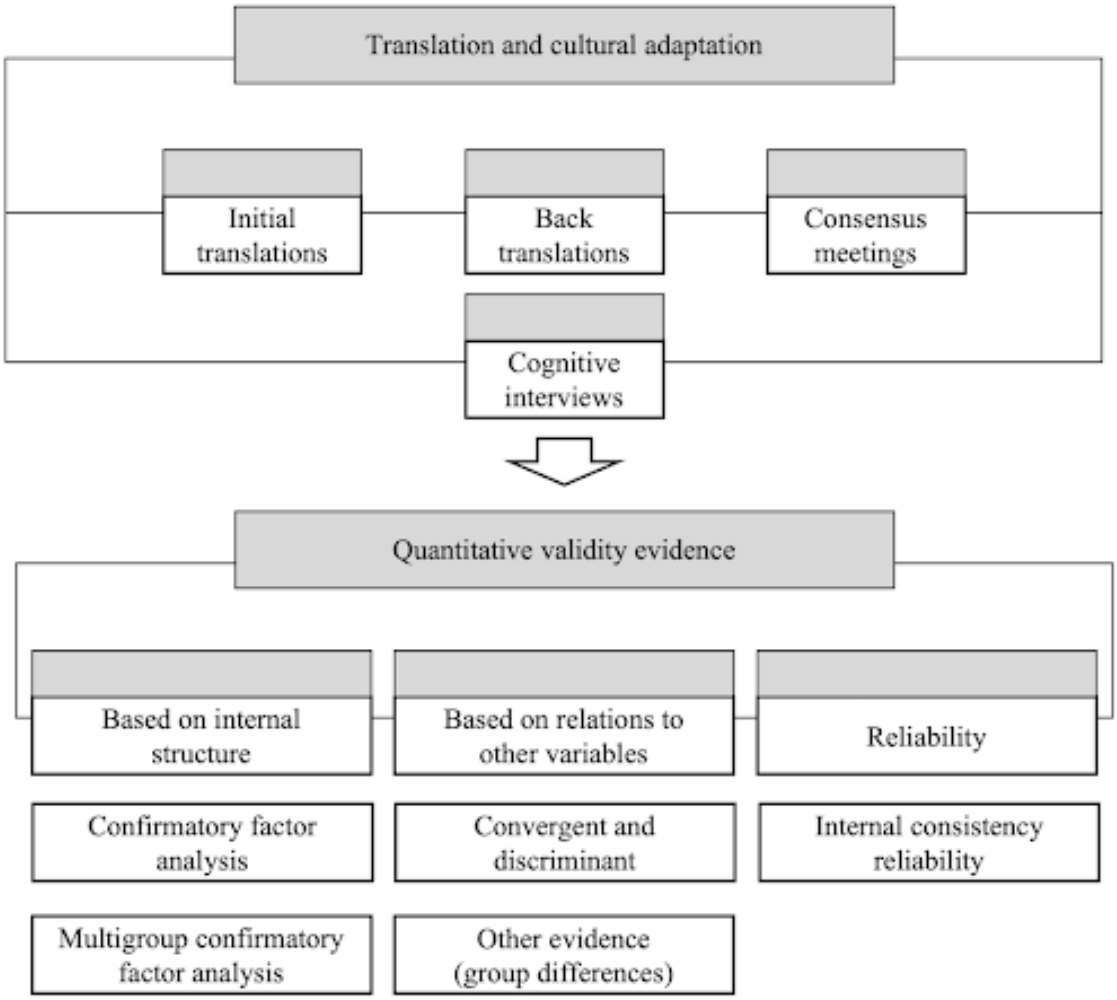

Translation and Cultural Adaptation

The translation and cultural adaptation process followed version 5 of the Translation Integrity Procedure [33], which provides systematic translation documentation and helps adaptation teams maximize measurement equivalence [34]. The same process was used for both the translation into Spanish and into Catalan and consisted of three main steps:

1. A professional translator produced an initial translation, which was revised by a second professional translator. In each case, the translators were fluent in English and native speakers of Spanish and Catalan. The translators were provided with the eHLQ in English and detailed information on the intent of each of the items.
2. In total, 2 professional translators (fluent in Spanish and Catalan and native speakers of English) then produced blinded back translations.
3. A series of consensus meetings were then held to critically review the proposed translations. These meetings involved the forward translators, the authors of this paper, and the authors of the eHLQ. These meetings analyzed the translations (with the help of the back translations for those who were not fluent in Catalan or Spanish), the cultural adaptation, and the equivalence of items with the original English version. Each item was examined in turn to confirm that (1) the intent of the translation matched that of the original; (2) the translations were appropriate across gender and education categories, dialects, and levels of digital health literacy; and (3) people would find it equally easy to respond (*strongly disagree*, *disagree*, *agree*, and *strongly agree*) to the translations as they would to the questions in English.

Cognitive interviews were carried out with 8 Spanish speakers and 8 Catalan speakers (12/16, 75% female; aged 22 to 81 years; 3/16, 19% with a primary school education; 5/16, 31% with a high school education; 8/16, 50% with a university education; 3/16, 19% with a rare or occasional use of technology; 6/16, 38% with a medium use of technology; and 7/16, 44% with habitual or continuous use of technology) to field-test the intelligibility and equivalence of each translation. Following a thinking-aloud approach, respondents were asked to answer each item after silently reading it. Scripted probes were used to explore difficulties and items that needed rephrasing or further explanation. Spontaneous probes were also used, for example, when the respondent hesitated or took a long time to answer an item. Changes suggested were introduced and discussed during the final consensus meeting to obtain the definitive versions of the questionnaire.

XSL•FO
RenderX

Quantitative Validity Evidence

Sample

All data were collected anonymously. The sample was stratified by geographical area, gender, and age of the Spanish and Catalan populations. Inclusion criteria were being aged ≥18 years, being a Spanish or Catalan speaker (depending on the version), and being able to read and understand a text in one of those languages. The final sample comprised 800 adult respondents for the Spanish and Catalan versions (n=400, 50% for each version). Demographic data included age, gender, educational level, socioeconomic status, frequency of technology use, and self-reported health status.

Recruitment and Data Collection

Between May 2021 and May 2022, participants were recruited through an independent partner company that specializes in data capture and processing, statistics, and market research. With >50 years of experience, the company [35] offers high-quality data collection and assessment services and has access to panels of respondents sourced through different methods to reach a wider variety of profiles (eg, loyalty programs in the travel, leisure, and retail sectors; banners, bloggers, influencers, and in-app advertising; and social networks and web-based communities) who had previously registered to participate as volunteers in different studies. The recruitment company targeted a potential 1,080,000 participants eligible for the Spanish sample and 140,000 participants eligible for the Catalan sample who were likely to qualify based on the demographic characteristics indicated in their user profiles. Gender, age group, and place of residence were used to ensure stratified samples mirroring the Spanish and Catalan populations. The company randomly contacted potential participants directly, with a detailed explanation of the project, the informed consent document, and an invitation to participate in a computer-aided web interview including a self-administered version of the eHLQ. Responses were accepted until balanced quotas were reached. To reach 300 responses for each language, the company contacted 350 potential participants for the Spanish sample and 352 for the Catalan sample. High response rates were achieved in both samples (300/350, 85.7% and 300/352, 85.2% for the Spanish and Catalan samples, respectively); thus, no reminders were sent, and no further potential participants were contacted.

As the digital panel was not able to ensure representativity of older people or people with low educational levels or digital skills, a further 200 responses (n=100, 50% for each language) were added to the samples. These respondents were recruited through direct approach in a wide range of settings in the community (health and social services and parks). For this subset of the sample, trained recruiters explained the objectives of the research, and after checking the basic requirements of gender and age group, they provided support to the participants to fill out a version of the same questionnaires on a tablet.

All the instruments were digitalized using COMPLET (ODEC), which is proprietary software [35]. It is used for in-person and digital interviews and surveys, and saves the data on their own servers.

Filling out the questionnaires took approximately 15 minutes. The questionnaires were self-administered or, in the subset of the sample recruited in community settings and where necessary, trained recruiters provided support.

The authors monitored both the recruitment and data gathering processes carried out by the partner company by providing the definition of the pool and sampling criteria, reviewing the digitalization process, and giving specific training instructions to recruit and survey the subset sample reached in community settings. Finally, an initial attention check (ie, an initial survey item that required respondents to provide a specific response) and a speeding check (ie, respondents whose survey duration was less than one-third of the mean survey duration) were asked to be included to guarantee data quality. Those who did not pass either quality control round were excluded from the final samples.

Instruments

The Spanish and Catalan versions were created from the eHLQ questionnaire [22] (refer to the Translation and Cultural Adaptation section). For convergent validity evidence, as with previous studies [36], a positive correlation was assumed between the first scale of the eHLQ and the eHEALS. Evidence of discriminant validity with the rest of the eHLQ subscales was also provided. As no other instrument was available in Spanish or Catalan for digital health literacy, the Spanish version [37] of the eHEALS questionnaire [12] was used as it has shown positive evidence of psychometric quality.

Self-categorization was also used, with a single self-reported question for technology use [38] and a single self-reported question about health status.

Educational level was collected in one self-reported question, and socioeconomic status was calculated according to a reviewed index based on a set of questions that take into account not only income but also education; job; housing; and financial aid for the family unit, not only for the respondent [39].

Statistical Analysis

Stata (version 17.0; StataCorp) and the R software (R Foundation for Statistical Computing) were used for the statistical analysis. Regarding the descriptive analysis, means and SDs were calculated for quantitative variables, and frequencies were calculated for categorical variables. To test differences for the Spanish and Catalan samples, we used 2-tailed *t* tests for quantitative variables and the chi-square test for categorical variables.

For the quantitative evidence, we partially followed the methodology used by previous researchers [15]. Following the recommendations of Doval et al [40], and considering that a 4-point Likert scale was used, the weighted least squares mean and variance corrected estimator was used for the confirmatory factor analysis (CFA). To assess goodness-of-fit indexes (GOFIs), the criteria by Hu and Bentler [41] were followed, which consider adequate a comparative fit model (comparative fit index [CFI]), a Tucker-Lewis index (TLI) of >0.90, and a

XSL•FO
RenderX

root mean square error of approximation (RMSEA) and standardized root mean square residual of <0.05. CFA was used to confirm the established factor structure. Seven 1-factor models and a 7-factor model were fitted. Metric invariance (same factor loadings) was assessed, and thresholds were also set to be equal between both groups. To do so, a change of <0.01 in CFI and RMSEA was also considered [42].

As for internal consistency reliability, the Cronbach α and categorical ω were provided for every subscale [40]. The categorical ω coefficient was reported from the 1-factor models. To assess reliability, a value of >0.80 was considered adequate for research purposes, and a value of 0.90 was considered adequate for individual decisions [43].

Given the approximately normal distribution of test scores (skew and kurtosis of <1 point in the Spanish and Catalan samples), the correlation among the 7 scales of the eHLQ was assessed through a Pearson correlation matrix. Convergent and discriminant validity evidence was assessed using the eHEALS and was evaluated by correlating the total score of the eHEALS with the score of each eHLQ scale.

Finally, the relationship with external variables was evaluated through the association between the eHLQ scores in each scale with the educational and socioeconomic variables and the use of digital technology variables. However, only the minimum and maximum categories were considered: primary school vs university for educational level, a high level corresponding to A1 or A2 vs a low level corresponding to E1 or E2 for socioeconomic level, and never to occasionally vs continuous use for the use of digital technology). The *t* test or a nonparametric test was used for the comparison according to the normal distribution of each eHLQ domain. Measurement invariance across groups could not be assessed due to sample size.

For the calculation of the initial sample, it was deemed necessary to have a minimum of 10 participants per item to conduct the factorial analyses (35 × 10 = 350). It was decided to slightly surpass that number and reach 400 participants. Finally, a post hoc analysis was conducted using *semPower* [44]. For a sample size of 400, an α error of .05, and expecting an RMSEA of 0.05, there was a power of >0.999.

Ethical Considerations

Ethics approval was obtained from the Ethics Committee of the Universitat Oberta de Catalunya (20201013_ehernandez_Traduccio), and the license for translating the eHLQ was obtained from Swinburne University of Technology, Australia (November 11, 2020).

All participants in the validation process gave informed consent before data collection and received a gift card for digital marketplaces as compensation. Data were gathered anonymously to ensure participants' confidentiality.

Results

Translation and Cultural Adaptation

During the series of consensus meetings, which involved the forward translators, the authors of this paper, and 2 of the authors of the original versions of the eHLQ, all the items for both the Spanish and Catalan translations were examined one by one. The first meeting served to highlight the most relevant points and pressing issues, which included the following agreements, all of which were applicable to both versions: (1) "health technology systems" and "eHealth systems" could be translated together as "digital health services"; (2) the "I find that..." expression at the beginning of different items was equivalent to "I think that..." but could not be removed and substituted by a more factual and objective form as it was essential to keep the point of view of the respondent; (3) "measurements" had a broader meaning than results from medical tests as it included other data sources, such as fitness-tracking devices; (4) the "works together" expression implied really interoperable systems, not just connected systems; (5) "find" could be translated as "search" (in relation to health information searching); and (6) it was not possible to combine ≥2 words together (eg, "data/measurements") as it was key to the questionnaire that there be no ambiguity in terms of the intent of each item.

Throughout the systematic process of translation and cultural adaptation, the issues discussed were generally the same for both the Spanish and Catalan versions. In addition to the aforementioned agreements, a number of key wording points were addressed and revised, for example, "perfectly" versus "very well," "being certain" versus "being sure," "available" versus "at the disposal of," "skills" versus "capabilities," or simply word order according to linguistic and cultural uses. A few specific issues also arose from cognitive interviews that required rephrasing both in the Spanish version (Table 1) and the Catalan version (Table 2). In almost all cases, the issues were the same. Scale 4 (*Feel safe and in control*) did not need any changes after the cognitive interviews in either the Spanish or Catalan version.

Cognitive interviews raised the need to reinforce the information initially included in the original English version used. As well as adding a few more examples to the existing ones for health professionals, health technology, and digital health services, further information was included after issues were raised regarding items eHLQ26, eHLQ5, and eHLQ12.

Item eHLQ26 was particularly problematic as it was hard to express everything encapsulated in the word "measurements" with just one word in Spanish or Catalan. The possibility of using more than one word was ruled out as it is key to the questionnaire that there be no ambiguity in terms of the intent of each item. Adding "data" or "results" alongside "measurements" would open the door to respondents stating that they use one but not the other. In the end, the decision was made to translate "measurements about my body" as "data on how my body works." In the instruction page, a new specific section was included to explain what "data on how my body works" meant by providing a few examples.

Similarly, items eHLQ5 and eHLQ12 raised the need to include a new communication and sharing information section on the instructions page, reinforcing the idea that having conversations or sharing information could be both informal and with professionals.

XSL•FO
RenderX

Table 1. Spanish version of the eHealth Literacy Questionnaire (eHLQ) and single-item changes after cognitive interviews.

Scale and item	First translation and first back translation	Final translation and back translation	Comments to changes
Scale 1			
eHLQ11	• Suelo usar la tecnología para entender problemas de salud. • I'm used to using...	• A menudo uso la tecnología para entender problemas de salud. • I usually use...	Emphasizing frequency and avoiding any possible misunderstanding.
Scale 2			
eHLQ5	• Los conocimientos que tengo me permiten participar en conversaciones sobre temas de salud. • ...conversations about health issues.	• Mis conocimientos me ayudan a participar en conversaciones sobre temas de salud con otras personas. • ...conversations about health issues with other people.	Including whom the conversations are with. Emphasizing that conversations are not only with health professionals but also with other people.
eHLQ12	• Tengo la información que necesito para hablar sobre mi salud. • I have the information I need...	• Dispongo de suficiente información para hablar sobre mi salud con otras personas. • I have enough information...	Emphasizing whom the conversations are with.
eHLQ26	• Uso mis resultados sanitarios y otras mediciones sobre mi cuerpo para entender mi salud. • I use my health results and other measurements...	• Uso los datos sobre el funcionamiento de mi cuerpo para entender mi salud. • I use data about how my body works...	Avoiding lists of items (results and measurements). Clarifying that health measurements (original) include professional and informal measurements. "Data" was the final option but specifying which data the item is referring to.
Scale 3			
eHLQ32	• Aprendo rápido a usar las novedades en tecnología para la salud. • I learn quickly...	• Aprendo fácilmente a usar las novedades en tecnología para la salud. • I learn easily...	Sounding more natural (adverb change from "quickly" to "easily"). Promoting the idea of an ability as a continuous behavior.
Scale 5			
eHLQ2	• La tecnología me hace sentir que participo activamente en mi salud. • Technology makes me feel that I am participating...	• La tecnología me hace sentir que me implico en el cuidado de mi salud. • Technology makes me feel involved...	Culturally sounding more natural (from actively participating to involvement).
Scale 6			
eHLQ9	• Los profesionales de la salud me ofrecen servicios a los que puedo acceder mediante la tecnología. • Health professionals offer me...	• Los profesionales de la salud que me atienden me ofrecen servicios a los que puedo atender mediante la tecnología. • Health professionals who look after me offer me...	Emphasizing who the health professionals are.
eHLQ16	• Mis datos de salud están a mi disposición desde cualquier lugar. • ...available to me from anywhere.	• Mis datos de salud están a mi disposición dondequiera que yo esté. • ...available to me from wherever I am.	Emphasizing the point of view of the respondent instead of data location.
eHLQ34	• Tengo a mi disposición tecnología para la salud que me funciona bien. • I have access to health technology...	• Tengo a mi disposicion servicios de salud digital que siempre funcionan. • I have access to digital health services...	Clarifying what health technology really means in a broader sense ("digital health services"). Emphasizing that it always works as intended.

XSL•FO
RenderX

Scale and item	First translation and first back translation	Final translation and back translation	Comments to changes
Scale 7			
eHLQ28	• Los servicios de salud digital parece que se adaptan a mis necesidades particulares. • I think eHealth services seem to adapt to my particular...	• Creo que los servicios de salud digital se adaptan a mis necesidades. • I think that digital health services adapt to my...	Keeping the point of view of the respondent. Simplifying. No need to add the adjective “particular” to the needs. Unifying the expression “digital health services” used throughout the questionnaire.
eHLQ31	• Creo que me ofrecen servicios de salud digital que me resultan útiles. • ...services that I find useful.	• Creo que la forma en la que se me ofrecen los servicios de salud se adapta a mi. • ...services are offered to me in a way that suits me.	Beyond usefulness and personalization, focusing not only on the health needs but also on the capabilities. Unifying the expression “digital health services” used throughout the questionnaire.

XSL•FO
RenderX

Table 2. Catalan version of eHealth Literacy Questionnaire (eHLQ) and single-item changes after cognitive interviews.

Scale and item	First translation and first back translation	Final translation and back translation	Comments to changes
Scale 1			
eHLQ13	• La tecnologia m'ajuda a decidir quina atenció mèdica em convé més. • ...medical care is best for me.	• La tecnologia m'ajuda a decidir quina atenció sanitaria em convé més. • ...health care is best for me.	Clarifying the meaning as "health care" includes care from any health professional (not only medical doctors).
Scale 2			
eHLQ5	• Els coneixements que tinc em permeten participar en converses sobre temes de salut. • ...conversations about health matters.	• Els meus coneixements m'ajuden a participar en converses sobre temes de salut amb altres persones. • ...conversations about health issues with other people.	Including whom the conversations are with. Emphasizing that conversations are not only with health professionals but also with other people.
eHLQ26	• Faig servir els meus resultats sanitaris I altres mesures sobre el meu cos per entendre la meva salut. • I use my health results and other measurements...	• Faig servir les dades sobre com funciona el meu cos per entendre la meva salut. • I use data about how my body works...	Avoiding lists of items (results and measurements). Clarifying that health measurements (original) include professional and informal measurements. "Data" was the final option but specifying which data the item is referring to.
Scale 3			
eHLQ32	• Aprenc ràpidament a fer server les novetats en tecnologia per a la salut. • I learn quickly...	• Aprenc fàcilment a fer server les technologies per a la salut. • I learn easily...	Sounding more natural (adverb change from "quickly" to "easily"). Promoting the idea of an ability as a continuous behavior.
Scale 5			
eHLQ2	• La tecnologia em fa sentir que participo activament en la meva salut. • ...feel that I am participating actively...	• La tecnologia em fa sentir que m'implico en la cura de la meva salut. • ...feel involved...	Culturally sounding more natural (from actively participating to involvement).
Scale 6			
eHLQ9	• Els professionals de la salut m'ofereixen serveis als quals puc accedir per mitjà de la tecnologia. • Health professionals offer me...	• Els professionals de la salut que m'atenen m'ofereixen servis de salut digital. • Health professionals who look after me offer me...	Emphasizing who the health professionals are.
eHLQ16	• Les meves dades de salut estan a la meva disposició des de qualsevol lloc. • ...available to me from anywhere.	• Les meves dades de salut estan a la meva disposició allà on jo sigui. • ...available to me from wherever I am.	Emphasizing the point of view of the respondent instead of data location.
eHLQ34	• Tinc a la meva disposició tecnologia per a la salut que funciona. • ...health technology that works.	• Tinc al meu abast serveis de salut digital que sempre funcionen. • ...digital health services that always work.	Clarifying what health technology really means in a broader sense ("digital health services"). Emphasizing that it always works as intended.
Scale 7			
eHLQ28	• Trobo que els serveis de salut digital s'adapten a les meves necessitats particulars. • I think eHealth services seem to adapt to my particular...	• Crec que els serveis de salut digital s'adapten a les meves necessitats. • I think that digital health services adapt to my...	Simplifying. No need to add the adjective "particular" to the needs. Unifying the expression "digital health services" used throughout the questionnaire.

Quantitative Validity Evidence

Sample Description

Both samples had similar patterns of distribution (Table 3). Briefly, 51% (408/800) of the sample were women, and most of the participants were aged between 50 and 64 years (285/800, 35.6%), had a high school educational level (322/800, 40.3%) and a high-medium socioeconomic level (617/800, 77.1%), and lived in municipalities of >200,000 inhabitants (428/800, 53.5%). Regarding their health, they reported good health status, and they considered themselves habitual technology users.

XSL•FO
RenderX

Table 3. Demographic characteristics of the Spanish and Catalan samples (N=800).

	Spanish version (n=400)	Catalan version (n=400)	*P* value
Gender, n (%)[a]			>.99
Woman	204 (51)	204 (51)	
Man	196 (49)	196 (49)	
I prefer not to answer	0 (0)	0 (0)	
Age (y), n (%)[a]			>.99
18-34	81 (20.3)	81 (20.3)	
35-49	103 (25.8)	103 (25.8)	
50-64	142 (35.5)	143 (35.8)	
≥65	74 (18.5)	73 (18.3)	
Age (y), mean (SD; range)[b]	49.22 (15.48; 18-79)	49.05 (15.80; 18-80)	>.99
Educational level, n (%)[a]			.43
Primary school	97 (24.3)	90 (25.5)	
High school	152 (38)	170 (42.5)	
University	151 (37.8)	140 (35)	
Socioeconomic index, n (%)[a]			.83
High (A1, A2, or B)	161 (40.3)	154 (38.5)	
Medium (C or D)	147 (36.8)	155 (38.8)	
Low (E1 or E2)	92 (23)	91 (22.8)	
Inhabitants per municipality, n (%)[a]			.96
Up to 5000	26 (6.5)	24 (6)	
5001-200,000	161 (40.3)	161 (40.3)	
>200,000	213 (53.3)	215 (53.8)	
Technology use, n (%)[a]			.10
Never or rarely	23 (5.8)	14 (3.5)	
Occasionally	63 (15.8)	43 (10.8)	
Medium	93 (23.3)	95 (23.8)	
Habitual	131 (32.8)	153 (38.3)	
Continuous	90 (22.5)	95 (23.8)	
Health status, n (%)[a]			.30
Very bad	2 (0.5)	7 (1.8)	
Bad	18 (4.5)	26 (6.5)	
Regular	102 (25.5)	100 (25)	
Good	229 (57.3)	214 (53.5)	
Very good	49 (12.3)	53 (13.3)	
eHealth literacy (eHEALS[c]), mean (SD; range)[a]	27.35 (8.40; 8-40)	27.80 (7.08; 8-40)	.41

[a]Quantitative variable.

[b]Categorical variable.

[c]eHEALS: eHealth Literacy Scale.

XSL•FO
RenderX

CFA and Measurement Invariance

Table 4 shows the GOFIs for the proposed models. Table 5 shows standardized factor loadings for the proposed models.

The 7-factor correlated model showed adequate GOFIs (Table 4) for the Spanish and Catalan versions. Thus, in both cases, the fitting values were considered adequate. As can be seen, factor loadings were high and homogeneous (>0.60).

In turn, the seven 1-factor models had adequate GOFIs according to the CFI and TLI values in both versions of the questionnaire. RMSEA values were >0.08 for factor 5 in the Catalan version and for factors 1, 3, 5, and 6 in the Spanish version. This could be explained by the fact that factor loadings are high in those factors.

The results of multigroup CFA showed that there was configural invariance between both languages (χ^2_{1078}=2707.1; CFI=0.98; TLI=0.98; RMSEA=0.07), equal factor loadings (ΔCFI=0.001; ΔRMSEA=0.004), and equal factor loadings and thresholds (ΔCFI=0.007; ΔRMSEA=0.01). Thus, measurement invariance can be assumed for factor loadings but not for thresholds.

Table 4. Goodness-of-fit indexes for the proposed models.

Model	Chi-square (*df*)	*P* value	CFI[a]	TLI[b]	RMSEA[c] (90% CI)	SRMR[d]
Spanish						
7-factor model	1631.2 (539)	<.001	0.98	0.98	0.07 (0.07-0.08)	0.04
Scale 1	103.9 (5)	<.001	0.99	0.98	0.23 (0.19-0.27)	0.02
Scale 2	18.8 (5)	<.001	1.00	0.99	0.08 (0.05-0.13)	0.01
Scale 3	23.7 (5)	<.001	1.00	1.00	0.10 (0.06-0.14)	0.01
Scale 4	29.6 (5)	<.001	1.00	0.99	0.11 (0.08-0.15)	0.01
Scale 5	40.3 (5)	<.001	1.00	0.99	0.14 (0.10-0.18)	0.01
Scale 6	54.8 (9)	<.001	0.99	0.99	0.12 (0.09-0.15)	0.02
Scale 7	3.9 (2)	.14	1.00	1.00	0.05 (0.00-0.12)	0.01
Catalan						
7-factor model	1380.1 (539)	<.001	0.98	0.98	0.07 (0.06-0.07)	0.04
Scale 1	32.9 (5)	<.001	0.99	0.99	0.12 (0.08-0.16)	0.02
Scale 2	12.7 (5)	.03	1.00	0.99	0.06 (0.02-0.11)	0.02
Scale 3	17.2 (5)	<.001	1.00	1.00	0.08 (0.04-0.12)	0.01
Scale 4	17.3 (5)	<.001	1.00	1.00	0.08 (0.04-0.12)	0.01
Scale 5	27.9 (5)	<.001	1.00	0.99	0.11 (0.07-0.15)	0.02
Scale 6	19.9 (9)	<.001	1.00	1.00	0.06 (0.02-0.09)	0.01
Scale 7	6.5 (2)	.04	1.00	1.00	0.08 (0.01-0.15)	0.01

[a]CFI: comparative fit index.

[b]TLI: Tucker-Lewis index.

[c]RMSEA: root mean square error of approximation.

[d]SRMR: standardized root mean square residual.

XSL•FO
RenderX

Table 5. Standardized factor loadings for the models of the Spanish and Catalan versions of the eHealth Literacy Questionnaire (eHLQ).

	7 factors	7 factors	1 factor	1 factor
	Spanish	Catalan	Spanish	Catalan
	Factor loading (95% CI)	Factor loading (95% CI)	Factor loading (95% CI)	Factor loading (95% CI)
Scale 1: Using technology to process health information				
eHLQ7	0.88 (0.85-0.91)	0.89 (0.86-0.91)	0.90 (0.87-0.93)	0.86 (0.83-0.90)
eHLQ11	0.82 (0.78-0.85)	0.89 (0.86-0.91)	0.92 (0.90-0.94)	0.86 (0.83-0.89)
eHLQ13	0.90 (0.87-0.92)	0.92 (0.90-0.94)	0.87 (0.85-0.90)	0.89 (0.86-0.92)
eHLQ20	0.85 (0.82-0.88)	0.87 (0.84-0.89)	0.88 (0.85-0.90)	0.85 (0.82-0.88)
eHLQ25	0.91 (0.89-0.93)	0.91 (0.89-0.93)	0.88 (0.86-0.91)	0.86 (0.83-0.90)
Scale 2: Understanding of health concepts and language				
eHLQ5	0.78 (0.74-0.83)	0.82 (0.78-0.86)	0.83 (0.79-0.87)	0.81 (0.76-0.87)
eHLQ12	0.75 (0.70-0.80)	0.86 (0.83-0.88)	0.90 (0.87-0.92)	0.79 (0.74-0.84)
eHLQ15	0.75 (0.70-0.80)	0.82 (0.78-0.86)	0.82 (0.78-0.85)	0.77 (0.71-0.82)
eHLQ21	0.71 (0.65-0.76)	0.78 (0.73-0.82)	0.79 (0.75-0.83)	0.72 (0.67-0.78)
eHLQ26	0.92 (0.89-0.95)	0.93 (0.91-0.96)	0.83 (0.80-0.87)	0.80 (0.76-0.85)
Scale 3: Ability to actively engage with digital services				
eHLQ4	0.88 (0.85-0.91)	0.9 (0.88-0.93)	0.83 (0.79-0.87)	0.84 (0.81-0.88)
eHLQ6	0.92 (0.89-0.94)	0.92 (0.90-0.94)	0.90 (0.87-0.92)	0.93 (0.90-0.95)
eHLQ8	0.85 (0.81-0.88)	0.86 (0.84-0.89)	0.82 (0.78-0.85)	0.86 (0.83-0.90)
eHLQ17	0.90 (0.87-0.92)	0.92 (0.89-0.94)	0.79 (0.75-0.83)	0.91 (0.87-0.94)
eHLQ32	0.92 (0.89-0.94)	0.93 (0.91-0.95)	0.83 (0.80-0.87)	0.90 (0.90-0.92)
Scale 4: Feel safe and in control				
eHLQ1	0.83 (0.78-0.87)	0.82 (0.78-0.86)	0.87 (0.84-0.90)	0.87 (0.84-0.90)
eHLQ10	0.89 (0.86-0.93)	0.92 (0.89-0.94)	0.91 (0.89-0.93)	0.90 (0.88-0.93)
eHLQ14	0.93 (0.89-0.97)	0.91 (0.87-0.94)	0.78 (0.74-0.82)	0.75 (0.70-0.80)
eHLQ22	0.85 (0.81-0.89)	0.87 (0.84-0.90)	0.88 (0.85-0.91)	0.90 (0.87-0.92)
eHLQ30	0.88 (0.84-0.91)	0.88 (0.85-0.92)	0.88 (0.85-0.91)	0.85 (0.81-0.88)
Scale 5: Motivated to engage with digital services				
eHLQ2	0.85 (0.81-0.88)	0.88 (0.86-0.91)	0.87 (0.84-0.90)	0.85 (0.81-0.88)
eHLQ19	0.90 (0.87-0.92)	0.92 (0.91-0.94)	0.92 (0.90-0.94)	0.90 (0.87-0.92)
eHLQ24	0.82 (0.78-0.85)	0.88 (0.85-0.90)	0.88 (0.85-0.90)	0.82 (0.78-0.85)
eHLQ27	0.85 (0.82-0.89)	0.90 (0.88-0.92)	0.90 (0.88-0.93)	0.84 (0.81-0.88)
eHLQ35	0.87 (0.83-0.90)	0.91 (0.88-0.93)	0.89 (0.87-0.92)	0.86 (0.82-0.89)
Scale 6: Access to digital services that work				
eHLQ3	0.78 (0.74-0.83)	0.81 (0.77-0.85)	0.76 (0.72-0.81)	0.78 (0.73-0.83)
eHLQ9	0.72 (0.67-0.78)	0.77 (0.73-0.81)	0.78 (0.73-0.82)	0.72 (0.66-0.77)
eHLQ16	0.83 (0.80-0.87)	0.78 (0.74-0.81)	0.78 (0.74-0.82)	0.80 (0.76-0.84)
eHLQ23	0.79 (0.75-0.83)	0.88 (0.85-0.91)	0.88 (0.85-0.92)	0.80 (0.76-0.84)
eHLQ29	0.84 (0.81-0.88)	0.85 (0.82-0.88)	0.84 (0.80-0.87)	0.84 (0.80-0.88)
eHLQ34	0.87 (0.84-0.90)	0.90 (0.88-0.92)	0.91 (0.89-0.94)	0.86 (0.83-0.90)

XSL•FO
RenderX

	7 factors Spanish Factor loading (95% CI)	7 factors Catalan Factor loading (95% CI)	1 factor Spanish Factor loading (95% CI)	1 factor Catalan Factor loading (95% CI)
Scale 7: Digital services that suit individual needs				
eHLQ18	0.94 (0.92-0.96)	0.91 (0.89-0.93)	0.87 (0.84-0.89)	0.85 (0.81-0.88)
eHLQ28	0.89 (0.87-0.92)	0.93 (0.91-0.95)	0.92 (0.90-0.94)	0.90 (0.88-0.92)
eHLQ31	0.94 (0.92-0.95)	0.90 (0.88-0.92)	0.92 (0.90-0.94)	0.94 (0.93-0.96)
eHLQ33	0.91 (0.89-0.93)	0.90 (0.88-0.92)	0.90 (0.88-0.92)	0.91 (0.89-0.93)

Mean Scores and Reliability

As can be seen in Table 6, all scales had a mean around 2.5 points, indicating that, on average, there was a good digital health literacy. All subscales had estimates of >0.80, which can be considered adequate for research purposes (Table 7). Moreover, except for scales 2 and 6, all scales had ω coefficients of >0.90, which is the minimum desirable threshold for making individual decisions. The values in the Spanish version, in general, were slightly higher than those in the Catalan version. As expected, considering that 1D factors were being assessed with high factor loadings, α and ω values were similar. Thus, all the internal consistency values were considered acceptable.

Table 6. Mean scores and internal consistency reliability for the Spanish and Catalan versions of the eHealth Literacy Questionnaire.

Scale	Spanish version			Catalan version		
	Values, mean (SD)	Cronbach α (95% CI)	ω (95% CI)	Values, mean (SD)	Cronbach α (95% CI)	ω (95% CI)
Using technology to process health information	2.41 (0.99)	0.90 (0.90-0.91)	0.92 (0.90-0.94)	2.55 (0.81)	0.91 (0.90-0.92)	0.91 (0.89-0.92)
Understanding of health concepts and language	2.60 (0.76)	0.85 (0.82-0.87)	0.87 (0.85-0.89)	2.65 (0.66)	0.85 (0.82-0.87)	0.85 (0.82-0.87)
Ability to actively engage with digital services	2.61 (0.97)	0.92 (0.91-0.93)	0.92 (0.90-0.93)	2.78 (0.79)	0.92 (0.91-0.93)	0.92 (0.90-0.9)
Feel safe and in control	2.65 (0.76)	0.90 (0.88-0.91)	0.90 (0.88-0.92)	2.72 (0.71)	0.90 (0.88-0.91)	0.90 (0.88-0.92)
Motivated to engage with digital services	2.46 (0.96)	0.90 (0.89-0.92)	0.92 (0.90-0.93)	2.54 (0.8)	0.90 (0.89-0.92)	0.91 (0.88-0.92)
Access to digital services that work	2.41 (0.82)	0.88 (0.87-0.89)	0.89 (0.87-0.91)	2.57 (0.71)	0.88 (0.87-0.90)	0.89 (0.86-0.90)
Digital services that suit individual needs	2.46 (0.97)	0.92 (0.91-0.93)	0.90 (0.88-0.92)	2.56 (0.83)	0.92 (0.91-0.93)	0.91 (0.88-0.93)

Table 7. Pearson correlation matrix between the Spanish (below the diagonal) and Catalan (above the diagonal) eHealth Literacy Questionnaire (eHLQ) scales and between the eHLQ scales and the eHealth Literacy Scale (eHEALS) score.

	eHEALS	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Scale 6	Scale 7
eHEALS	1.00	0.63	0.64	0.60	0.41	0.60	0.59	0.61
Scale 1	0.76	1.00	0.86	0.84	0.66	0.89	0.83	0.83
Scale 2	0.71	0.83	1.00	0.83	0.65	0.82	0.79	0.79
Scale 3	0.76	0.89	0.82	1.00	0.67	0.82	0.80	0.81
Scale 4	0.57	0.68	0.74	0.69	1.00	0.71	0.78	0.71
Scale 5	0.74	0.91	0.82	0.87	0.73	1.00	0.85	0.89
Scale 6	0.70	0.86	0.80	0.84	0.76	0.88	1.00	0.88
Scale 7	0.76	0.90	0.82	0.89	0.73	0.93	0.89	1.00

XSL•FO
RenderX

Correlation Matrix

Concerning the correlations among the eHLQ scales, they ranged from 0.69 to 0.93 in the Spanish version and from 0.67 to 0.90 in the Catalan version (Table 7). In both versions, the scale with the lowest correlations with the other dimensions was scale 4 (*Feel safe and in control*). In the Spanish version, the highest correlation was between scales 5 and 7 (*Motivated to engage with digital services* and *Digital services that suit individual needs*; 0.93), and in the Catalan version, it was between scales 1 and 5 (*Using technology to process health information* and *Motivated to engage with digital services*; 0.90).

Convergent and Discriminant Validity Evidence

For gathering further evidence of discriminant and convergent validity, the eHEALS was administered. A high correlation was expected with the scale "Using technology to process health information," and lower correlations were expected with the rest (discriminant validity evidence). Once analyzed, for the eHEALS, the mean value for the Spanish version was 27.35 (SD 8.40), and for the Catalan version, the mean was 27.80 (SD 7.08); in both versions, the range was from 8 to 40. The analysis of the eHEALS and the different eHLQ scales (Table 7) showed a moderate correlation of the global eHEALS score with the eHLQ scales of the Spanish version, with the correlations with scale 1 (*Using technology to process health information*; r=0.76), scale 3 (*Ability to actively engage with digital services*; r=0.76), and scale 7 (*Digital services that suit individual needs*; r=0.76) being the highest and the correlation with scale 4 (*Feel safe and in control*; r=0.57) being the lowest. Regarding the Catalan version, the values were slightly lower than those for the Spanish version. Again, scale 4 had the lowest correlation with the eHEALS (*Feel safe and in control*; r=0.41), whereas scale 2 had the highest correlation (*Understanding of health concepts and language*; r=0.64). Thus, while positive convergent evidence was achieved with scale 1, no positive discriminant evidence was found in the rest of the scales. This could be explained by the high correlation that all the scales of the eHEALS have.

Relationship With External Variables

The scores for all the eHLQ scales were lower in the group with a primary school education in comparison to the group with a university education both in the Spanish and Catalan versions. As measurement invariance could not be assessed due to sample size limitations, it is not possible to assess whether these differences are due to measurement invariance or a real impact. These differences were statistically significant in all the scales (Table 8); thus, both versions discriminated between these groups in all the scales.

The scores for all the eHLQ scales were lower in the group with a low socioeconomic level in comparison to the group with a high socioeconomic level both in the Spanish and Catalan versions. These differences were significant in all the scales (Table 9); thus, both versions discriminated between these groups in all the scales.

The scores for all the eHLQ scales were lower in those who never used technology or used it occasionally in comparison to those who reported a continuous use both in the Spanish and Catalan versions. These differences were significant in all the scales (Table 10); thus, both versions discriminated between these groups in all the scales.

Table 8. Comparison of the 7 scales of the Spanish and Catalan versions of the eHealth Literacy Scale across educational levels.

	Spanish version			Catalan version		
	Primary school education (n=97), mean (SD)	University education (n=151), mean (SD)	*P* value	Primary school education (n=90), mean (SD)	University education (n=140), mean (SD)	*P* value
Using technology to process health information	1.88 (0.56)	2.93 (0.47)	<.001	1.55 (0.57)	2.93 (0.57)	<.001
Understanding of health concepts and language	2.11 (0.38)	3.03 (0.44)	<.001	1.85 (0.49)	2.97 (0.47)	<.001
Ability to actively engage with digital services	2.00 (0.61)	3.12 (0.45)	<.001	1.83 (0.65)	3.15 (0.52)	<.001
Feel safe and in control	2.22 (0.42)	2.99 (0.52)	<.001	2.06 (0.57)	2.94 (0.61)	<.001
Motivated to engage with digital services	1.89 (0.47)	2.99 (0.47)	<.001	1.56 (1.55)	2.91 (0.57)	<.001
Access to digital services that work	2.04 (0.42)	2.83 (0.47)	<.001	1.73 (0.50)	2.89 (0.51)	<.001
Digital services that suit individual needs	1.91 (0.53)	3.00 (0.51)	<.001	1.62 (0.63)	2.92 (0.59)	<.001

XSL•FO
RenderX

Table 9. Comparison of the 7 scales of the Spanish and Catalan versions of the eHealth Literacy Questionnaire across socioeconomic levels.

	Spanish version			Catalan version		
	Low level (n=92), mean (SD)	High level (n=89), mean (SD)	*P* value	Low level (n=91), mean (SD)	High level (n=99), mean (SD)	*P* value
Using technology to process health information	2.20 (0.70)	3.00 (0.45)	<.001	2.19 (0.74)	2.87 (0.68)	<.001
Understanding of health concepts and language	2.37 (0.55)	3.02 (0.41)	<.001	2.35 (0.67)	2.93 (0.59)	<.001
Ability to actively engage with digital services	2.32 (0.72)	3.17 (0.44)	<.001	2.45 (0.83)	3.12 (0.61)	<.001
Feel safe and in control	2.42 (0.56)	2.99 (0.50)	<.001	2.37 (0.76)	2.90 (0.66)	<.001
Motivated to engage with digital services	2.21 (0.64)	3.07 (0.39)	<.001	2.15 (0.80)	2.89 (0.68)	<.001
Access to digital services that work	2.25 (0.54)	2.91 (0.42)	<.001	2.21 (0.68)	2.81 (0.64)	<.001
Digital services that suit individual needs	2.17 (0.66)	3.06 (0.42)	<.001	2.14 (0.84)	2.90 (0.67)	<.001

Table 10. Comparison of the 7 scales of the Spanish and Catalan versions of the eHealth Literacy Questionnaire across levels of use of technology.

	Spanish version			Catalan version		
	Never or occasionally (n=86), mean (SD)	Continuous (n=90), mean (SD)	*P* value	Never or occasionally (n=57), mean (SD)	Continuous (n=95), mean (SD)	*P* value
Using technology to process health information	1.94 (0.63)	3.06 (0.52)	<.001	1.93 (0.84)	3.07 (0.61)	<.001
Understanding of health concepts and language	2.23 (0.47)	3.04 (0.52)	<.001	2.13 (0.74)	3.01 (0.57)	<.001
Ability to actively engage with digital services	2.06 (0.65)	3.31 (0.54)	<.001	2.00 (0.89)	3.23 (0.63)	<.001
Feel safe and in control	2.35 (0.55)	3.03 (0.59)	<.001	2.29 (0.69)	3.00 (0.73)	<.001
Motivated to engage with digital services	1.99 (0.58)	3.08 (0.58)	<.001	1.89 (0.81)	3.00 (0.64)	<.001
Access to digital services that work	2.09 (0.48)	2.93 (0.55)	<.001	1.99 (0.72)	2.95 (0.62)	<.001
Digital services that suit individual needs	1.97 (0.61)	3.08 (0.56)	<.001	1.88 (0.84)	3.07 (0.67)	<.001

Discussion

Principal Findings

The main objective of this study was to undertake the translation and cultural adaptation of the eHLQ and gather evidence of the psychometric quality of the eHLQ in a large stratified sample of the general population in Spain, creating a parallel process for 2 of Spain's official languages: Spanish and Catalan.

Overall, the new versions of the eHLQ in Spanish and Catalan were found to be highly coherent with the original validation of the English and Danish versions [22]. As shown by the simultaneous process carried out for the original versions to maximize comparability over linguistic differences, the Spanish and Catalan versions present similar patterns that were confirmed through measurement invariance in 2 different language versions within the same sociocultural context.

Translation and Cultural Adaptation

The Translation Integrity Procedure [33] provided common ground and facilitated iterative discussions among the professional translators, the authors of the original questionnaire, and the authors of this study, leading to a final consensus that ensured that the translated versions maintained the same semantic and conceptual orientations as the original version [45]. Furthermore, this process generated some interesting insights and qualitative results regarding the conceptualization of the instrument and its equivalence in a context of different languages [34].

The cognitive interviews, which provided validity evidence based on the test content and the response process, showed that some of the items could be misunderstood in the new translated versions, and it was difficult to reach a consensus to overcome the cultural adaptation. This was particularly relevant for item eHLQ26 (*I use measurements about my body...*) as the term

XSL•FO
RenderX

"measurements" in this setting is mainly associated with height and weight in both Spanish and Catalan. Thus, the wording of this item had to be changed slightly to "I use data on how my body works," and some specific examples of this concept were included in the instructions page with the terms used in the questionnaire. Although the difficulties with the concept of measurements (including professional and fitness-tracking devices) can be related with the low indicator for Spain in a European survey [7], this item has also been problematic in previous studies [15,24,26]. For example, in the Australian and Taiwanese populations, this item had low factor loadings, and in the Dutch population, it had residual correlation issues. In this study, this item presented high factor loadings in both versions.

Along the same lines, other groups of examples of terms related to "communication and sharing of information" (both with professionals and other people) were also included to further clarify the intent of items eHLQ20 (*I use technology to share information...*) and eHLQ27 (*Technology improves my communication...*). In fact, although recent reports and surveys show that Spain can be considered one of the leading countries in the use of digital technologies related to health [8,10], it seems that the concepts of "eHealth," "digital health," and "digital health services" are not well understood by the general population. Despite the fact that citizens are increasingly using them, there is little familiarity with these terms outside professional circles, highlighting the need for an instrument such as the eHLQ to assess the readiness of the population and help increase the adoption and use of digital health services.

Reinforcing the instructions page with more information forms part of the cultural adaptation process, and the eHLQ counts on this. The Spanish and Catalan versions of the eHLQ took advantage of this possibility to add more information. However, it did not provide further recommendations on their use, administration, or interpretation, as do other versions [26].

Quantitative Validity Evidence

Results for both the Spanish and Catalan versions showed means around 2.5 points, with the 7 scales showing a similar deviation pattern. The values were in a similar range to that of the values obtained in the studies in Denmark (between 2.42 and 2.97) [22], Sweden (between 2.58 and 3.04) [25], and Taiwan (between 2.37 and 3.08). In all 3 cases, scale 2 (*Understanding of health concepts and language*) had the highest score, and scales 7 (*Digital services that suit individual needs*; for the Danish and Swedish versions) and 6 (*Access to digital services that work*; for the Mandarin version) had the lowest mean.

The 7-factor correlated model presented adequate GOFIs both in the Spanish and Catalan versions. Moreover, measurement invariance was found between both forms for the factor loadings. The CFA 1-factor model for each scale also fit well considering CFI and TLI, and all the items presented a strong factor loading within their respective scales. In this model, the RMSEA values had good results for some of the scales in the Spanish version, although this index can be problematic for assessing models with small df, as in this study [24].

According to the factor loadings, all the items were around the constraint value of each scale. In both versions, the scale with the highest loadings was scale 4 (*Feeling safe and in control*), as mentioned in recent surveys in Spain as the key element for the implementation of a digitalized and efficient health system [8,9]. For the Spanish version, almost all items had factor loadings of >0.80 except for 2 items of scale 6: eHLQ9 (r=0.77; *My health care providers deliver services that I can access through technology*) and eHLQ16 (r=0.78; *My health data are available to me wherever I am*). Item eHLQ9 was also the one with the lowest factor loading in the Catalan version. This cannot be considered a low factor loading as these results are in line with those of the original version, where these items showed the lowest factor loadings (0.44 and 0.43, respectively) [22]. Ensuring access and universal coverage for digital health systems, represented by scale 6, is still a major challenge for individuals to process and understand.

There was a high correlation between the dimensions, especially for factors 1 and 5 (*Using technology to process health information* and *Being motivated to engage with digital services*), both in the Spanish and Catalan versions (r=0.90 in both cases). These values resemble the findings for the original version, which the authors attributed to the result of being located along the same causal path [22]. Both versions showed other high correlations between factors 5 and 7 (r=0.92 in Spanish and r=0.89 in Catalan), perhaps suggesting that motivation and engagement regarding using digital health services are related to the perception that those services suit the users' specific needs and preferences.

To explore convergent and discriminant validity evidence, the eHEALS [12] was also administered, which is a good measure of some elements of digital health literacy. It has been adapted and has shown evidence of psychometric quality in Spanish in a sample of university students [36]. Mean scores were similar in the Spanish (27.35, SD 8.40) and Catalan (27.80, SD 7.08) versions and slightly lower than those in a study on the use of the internet and digital literacy among Spanish nurses, with a mean of 32.44 [46]. Moderate to low correlations of the eHEALS were found with the different eHLQ scales, as expected because of their different theoretical models [12,21]. This was an expected result given that, even though the eHEALS is still a widely used tool, it does not provide sufficient understanding of the individual's interaction with digital services and technology or the user's experience of engaging with the system, which are all covered by the eHLQ's dimensions. For both translated versions of the eHLQ, the lowest correlation with the eHEALS was found for scale 4 (*Feel safe and in control*; r=0.56 in Spanish and r=0.43 in Catalan), probably because all the items within this scale relate to the ownership of personal data and their secure access, which are elements of a dimension not included at all in the eHEALS. Lower correlations were expected between the eHEALS and scales 3, 4, 5, 6, and 7, and discriminant evidence was searched for. The high correlation among the other subscales could be explained by the low discrimination among the subscales of the eHLQ, as was also the case in previous studies and as has been described previously.

XSL•FO
RenderX

As for internal consistency reliability, adequate values were found for all the scales even considering the minimal number of items (4-6) per scale, with Cronbach α and categorical ω values of >0.80 (and almost all >0.90) in all the scales in both the Spanish and Catalan versions. These values confirm that the selected items are consistent in each scale in both translated versions and are in line with the original results of the Danish validity testing study [22], with internal consistency values ranging from 0.77 to 0.86, and with the translation and adaptation to the Mandarin language [24], with values ranging from 0.75 to 0.95. In both studies, the lowest internal consistency values were also found for scale 2. Although the Swedish version [25] shows higher values (ranging from 0.82 to 0.92), scale 2 (together with scale 6) also shows the lowest internal consistency values.

Moreover, reliability was considered good as all the scales in both versions had α and ω values of >0.80, with values close to 0.90.

In both the Spanish and Catalan versions, all the scales showed the sensitivity to discriminate among known groups according to their educational and socioeconomic level and their self-categorization as technology users. However, results must be interpreted with caution as measurement invariance across the groups could not be studied due to sample size limitations. This is a relevant issue as a previous study in a Dutch sample could not establish full invariance among the different groups [26]. The differences were statistically significant for all the scales in these 3 known groups. Thus, people with a primary school education, lower socioeconomic status, or occasional use of technology scored lower than those with university studies, higher socioeconomic status, or continuous use of technology. This was apparent in all the scales and in both the Spanish and Catalan versions, with similar patterns of dispersion. Although the difference among the groups reveals the discrimination power of both versions of the questionnaire, this finding has to be taken with caution as it may also be an artifact due to the large samples and the dispersion shown by the scores.

Finally, a tendency to mark “disagree” rather than “strongly disagree,” especially for those groups with lower socioeconomic status, was revealed across the scales. This might reflect some respondents’ attempts to hide a lack of comprehension or the inappropriateness of the questions to their particular situation, prompting them to provide more socially desirable answers. The possibility was discussed of including some previous instructions to encourage people to provide an answer for all the items and answer “strongly disagree” in cases in which the question was not clearly understood or did not apply to their own situation. However, as answers should not be suggested to respondents under any circumstances, a general recommendation designed to encourage people to respond as well as they could was included, informing them that there were no correct or incorrect answers. In relation to that, the questionnaire raises aspects linked to digital health (eg, services and devices) that the Spanish and Catalan populations, even as educated and digital users, do not recognize in their current daily lives. Considering the rapid development of digital health in Spain, this will no longer be the case in the near future. For this reason, future testing studies of the eHLQ will probably not require such an extensive introduction to the questionnaire as the one that was included as a key element in the cultural adaptation of the instrument in this study.

Strengths and Limitations

The translation and cultural adaptation of an instrument into 2 languages simultaneously is a strength of this study because they provide a tool with psychometric guarantees to assess health literacy. In addition, the stratified sample of the Spanish- and Catalan-speaking populations in terms of gender, age, and place of residence is another important strength that makes the eHLQ a good instrument to use in community and population surveys.

However, all the interviews were conducted using a digitally administered questionnaire, and therefore, all the respondents needed some basic level of digital skills, which may have excluded some people with no internet access or no basic skills from the sample. To overcome this limitation, the sample was balanced by conducting 200 (n=100, 50% for each version) face-to-face interviews in community settings targeted at people with primary school studies and of a low socioeconomic level, and 80% (60/200) of them were aged >60 years. In this sense, it is also important to note that, according to the self-categorization scale regarding technology use, the total sample included respondents who reported that they never or occasionally used technology (54/400, 13.5% in the Spanish version and 27/400, 6.7% in the Catalan version).

Regarding future studies, although the sample in this study was large and covered a wide range of sociodemographic, academic, and technology use profiles, it was not free of voluntary reporting and self-reporting biases. Further studies are suggested involving participants with different health conditions to provide evidence of the psychometric quality of the instrument in different contexts of use.

Future studies with larger sample sizes should consider testing for measurement invariance across groups to test whether group differences are due to measurement issues or real differences. Finally, it would be interesting to consider the development of further versions covering Basque and Galician, which are also spoken in Spain.

Conclusions

This study demonstrates that the translated and adapted versions of the eHLQ in Spanish and Catalan are compatible with the original’s construct and item intent and maintain proper psychometric properties. The eHLQ provides a tool in Spain for researchers, policy makers, and health service managers to explore people’s needs, skills, and competencies and obtain interesting insights into their interactions and engagement regarding their own experiences with digital health services, especially in the context of digital health growth in Spain.

XSL•FO
RenderX

Acknowledgments

The authors thank all the participants both in the cognitive interviews and the quantitative survey.

Funding was received from the eHealth Center of the Universitat Oberta de Catalunya. Generative artificial intelligence was not used in any part of manuscript writing.

Data Availability

The Catalan and Spanish versions of the eHealth Literacy Questionnaire are available upon request. Please contact Centre for Global Health and Equity at the Swinburne University of Technology [47] for further information. The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization was handled by EHE and IdA. Methodology was handled by EHE, DC, and IdA. Data curation was handled by NR. Formal analysis was handled by NR and AAB. Writing of the original draft, review, and editing were handled by all the authors. All authors have read and agreed to the submitted version of this manuscript.

Conflicts of Interest

None declared.

References

1. Tilahun B, Gashu KD, Mekonnen ZA, Endehabtu BF, Angaw DA. Mapping the role of digital health technologies in prevention and control of COVID-19 pandemic: review of the literature. Yearb Med Inform 2021 Aug;30(1):26-37 [FREE Full text] [doi: 10.1055/s-0041-1726505] [Medline: 34479378]
2. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://apps.who.int/iris/handle/10665/344249 [accessed 2023-04-15]
3. eHealth strategy for Ireland. Department of Health Government of Ireland. 2013 Jun 7. URL: https://www.gov.ie/en/publication/6b7909-ehealth-strategy-for-ireland/ [accessed 2023-04-15]
4. Digital health. Health of New Zealand. 2019. URL: https://www.health.govt.nz/our-work/digital-health [accessed 2023-04-15]
5. Chapter 5: digitally-enabled care will go mainstream across the NHS. National Health Service. 2019. URL: https://www.longtermplan.nhs.uk/online-version/chapter-5-digitally-enabled-care-will-go-mainstream-across-the-nhs/ [accessed 2023-04-15]
6. Digital health center of excellence. U.S. Food & Drug Administration. 2020. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence [accessed 2023-04-15]
7. Estrategia de salud digital - Sistema Nacional de Salud. Secretaría General de Salud Digital, Información e Innovación para el SNS. 2021. URL: https://www.sanidad.gob.es/areas/saludDigital/doc/Estrategia_de_Salud_Digital_del_SNS.pdf [accessed 2023-04-15]
8. Sabaté N, Bautista L. Digitalizar la atención sanitaria sin perder el toque humano. Accenture.: Accenture; 2021 Dec 17. URL: https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-171/accenture-digital-health-pharmaceutical-sector-survey.pdf#zoom=40 [accessed 2023-04-15]
9. Informe sobre Transformación Digital en Salud en España. Asociación Salud Digital.: Asociación Salud Digital; 2021. URL: https://salud-digital.es/wp-content/uploads/2021/03/Informe-sobre-Transformacion-Digital-en-Salud-Actualizacion-2021.pdf [accessed 2023-04-15]
10. Pérez Sust P, Solans O, Fajardo JC, Medina Peralta M, Rodenas P, Gabaldà J, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 04;6(2):e19106 [FREE Full text] [doi: 10.2196/19106] [Medline: 32339998]
11. Individuals - internet activities. Eurostat. 2019. URL: https://ec.europa.eu/eurostat/databrowser/view/ISOC_CI_AC_I__custom_2406483/bookmark/table?lang=en [accessed 2019-10-19]
12. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
13. Kayser L, Kushniruk A, Osborne RH, Norgaard O, Turner P. Enhancing the effectiveness of consumer-focused health information technology systems through eHealth literacy: a framework for understanding users' needs. JMIR Hum Factors 2015 May 20;2(1):e9 [FREE Full text] [doi: 10.2196/humanfactors.3696] [Medline: 27025228]
14. Monkman H, Kushniruk AW. The consumer health information system adoption model. Stud Health Technol Inform 2015;218:26-31. [Medline: 26262522]
15. Cheng C, Elsworth G, Osborne RH. Validity evidence based on relations to other variables of the eHealth literacy questionnaire (eHLQ): Bayesian approach to test for known-groups validity. J Med Internet Res 2021 Oct 14;23(10):e30243 [FREE Full text] [doi: 10.2196/30243] [Medline: 34647897]

XSL•FO
RenderX

16. Griebel L, Enwald H, Gilstad H, Pohl AL, Moreland J, Sedlmayr M. eHealth literacy research-Quo vadis? Inform Health Soc Care 2018 Dec;43(4):427-442. [doi: 10.1080/17538157.2017.1364247] [Medline: 29045164]
17. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021 Nov 15;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
18. Nolasco A, Barona C, Tamayo-Fonseca N, Irles M, Más R, Tuells J, et al. [Health literacy: psychometric behaviour of the HLS-EU-Q16 questionnaire]. Gac Sanit 2020 Jul;34(4):399-402 [FREE Full text] [doi: 10.1016/j.gaceta.2018.08.006] [Medline: 30473252]
19. Sørensen K, Van den Broucke S, Pelikan JM, Fullam J, Doyle G, Slonska Z, et al. Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). BMC Public Health 2013 Oct 10;13:948 [FREE Full text] [doi: 10.1186/1471-2458-13-948] [Medline: 24112855]
20. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
21. Norgaard O, Furstrand D, Klokker L, Karnoe A, Batterham R, Kayser L, et al. The e-health literacy framework: a conceptual framework for characterizing e-health users and their interaction with e-health systems. Knowl Manag E-Learn 2015;7(4). [doi: 10.34105/j.kmel.2015.07.035]
22. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
23. Parker R, Ratzan SC. Health literacy: a second decade of distinction for Americans. J Health Commun 2010 Aug 31;15 Suppl 2(sup2):20-33. [doi: 10.1080/10810730.2010.501094] [Medline: 20845190]
24. Cheng C, Elsworth GR, Osborne RH. Validity evidence of the eHealth literacy questionnaire (eHLQ) part 2: mixed methods approach to evaluate test content, response process, and internal structure in the Australian community health setting. J Med Internet Res 2022 Mar 08;24(3):e32777 [FREE Full text] [doi: 10.2196/32777] [Medline: 35258475]
25. Sjöström AE, Hajdarevic S, Hörnsten Å, Kristjánsdóttir Ó, Castor C, Isaksson U. The Swedish version of the eHealth literacy questionnaire: translation, cultural adaptation, and validation study. J Med Internet Res 2023 Apr 12;25:e43267 [FREE Full text] [doi: 10.2196/43267] [Medline: 37043268]
26. Poot CC, Meijer E, Fokkema M, Chavannes NH, Osborne RH, Kayser L. Translation, cultural adaptation and validity assessment of the Dutch version of the eHealth Literacy Questionnaire: a mixed-method approach. BMC Public Health 2023 May 30;23(1):1006 [FREE Full text] [doi: 10.1186/s12889-023-15869-4] [Medline: 37254148]
27. Holt KA, Karnoe A, Overgaard D, Nielsen SE, Kayser L, Røder ME, et al. Differences in the level of electronic health literacy between users and nonusers of digital health services: an exploratory survey of a group of medical outpatients. Interact J Med Res 2019 Apr 05;8(2):e8423 [FREE Full text] [doi: 10.2196/ijmr.8423] [Medline: 30950809]
28. Holt KA, Overgaard D, Engel LV, Kayser L. Health literacy, digital literacy and eHealth literacy in Danish nursing students at entry and graduate level: a cross sectional study. BMC Nurs 2020;19:22 [FREE Full text] [doi: 10.1186/s12912-020-00418-w] [Medline: 32308559]
29. Villadsen SF, Hadi H, Ismail I, Osborne RH, Ekstrøm CT, Kayser L. eHealth literacy and health literacy among immigrants and their descendants compared with women of Danish origin: a cross-sectional study using a multidimensional approach among pregnant women. BMJ Open 2020 May 07;10(5):e037076 [FREE Full text] [doi: 10.1136/bmjopen-2020-037076] [Medline: 32385065]
30. Foley K, Freeman T, Ward P, Lawler A, Osborne R, Fisher M. Exploring access to, use of and benefits from population-oriented digital health services in Australia. Health Promot Int 2021 Aug 30;36(4):1105-1115. [doi: 10.1093/heapro/daaa145] [Medline: 33367568]
31. Kayser L, Karnoe A, Duminski E, Jakobsen S, Terp R, Dansholm S, et al. Health professionals' eHealth literacy and system experience before and 3 months after the implementation of an electronic health record system: longitudinal study. JMIR Hum Factors 2022 Apr 29;9(2):e29780 [FREE Full text] [doi: 10.2196/29780] [Medline: 35486414]
32. Terp R, Kayser L, Lindhardt T. "It is not rocket science.." - Older peoples' understanding of nutrition - A qualitative study. Appetite 2021 Jan 01;156:104854. [doi: 10.1016/j.appet.2020.104854] [Medline: 32871203]
33. Hawkins M, Osborne RH. Translation Integrity Procedure (TIP). Version 5. Swinburne University of Technology. 2019. URL: http://pstorage-npg-968563215.s3.amazonaws.com/22806608/12874_2020_962_MOESM1_ESM.pdf [accessed 2023-04-15]
34. Hawkins M, Cheng C, Elsworth GR, Osborne RH. Translation method is validity evidence for construct equivalence: analysis of secondary data routinely collected during translations of the Health Literacy Questionnaire (HLQ). BMC Med Res Methodol 2020 May 26;20(1):130 [FREE Full text] [doi: 10.1186/s12874-020-00962-8] [Medline: 32456680]
35. ODEC homepage. ODEC. URL: https://www.odec.es/ [accessed 2024-02-10]
36. Marsall M, Engelmann G, Skoda EM, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a revised German version of the eHealth literacy scale. J Med Internet Res 2022 Feb 02;24(2):e28252 [FREE Full text] [doi: 10.2196/28252] [Medline: 35107437]

XSL•FO
RenderX

37. Paramio Pérez G, Almagro BJ, Hernando Gómez Á, Aguaded Gómez JI. [Validation of the eHealth Literacy Scale (eHEALS) in Spanish University Students]. Rev Esp Salud Publica 2015;89(3):329-338 [FREE Full text] [doi: 10.4321/S1135-57272015000300010] [Medline: 26388346]
38. Casas F, Madorell L, Figuer C, González M, Malo S, García M, et al. Preferències i expectatives dels adolescents relatives a la televisió a Catalunya. Consell de l'Audiovisual de Catalunya. 2017. URL: https://www.cac.cat/sites/default/files/migrate/study/preferencies_i_expectatives.pdf [accessed 2023-04-05]
39. Nuevo sistema de clasificación socioeconómica en el EGM. Guía de Referencia EGM.: Asociación Para la Investigación de Medios de Comunicación; 2018. URL: https://www.aimc.es/a1mc-c0nt3nt/uploads/2018/05/info_clas_socioeconomica.pdf [accessed 2023-04-05]
40. Doval E, Viladrich C, Angulo-Brunet A. Coefficient alpha: the resistance of a classic. Psicothema 2023 Feb;35(1):5-20. [doi: 10.7334/psicothema2022.321] [Medline: 36695846]
41. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999;6(1):1-55. [doi: 10.1080/10705519909540118]
42. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Modeling 2002;9(2):233-255. [doi: 10.1207/S15328007SEM0902_5]
43. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd Edition. New York, NY: McGraw Hill Education; 2017.
44. Moshagen M, Erdfelder E. A new strategy for testing structural equation models. Struct Equ Modeling 2015 Apr 22;23(1):54-60. [doi: 10.1080/10705511.2014.950896]
45. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract 2011 Apr;17(2):268-274. [doi: 10.1111/j.1365-2753.2010.01434.x] [Medline: 20874835]
46. Ros-Navarret R. Internet use and digital literacy among Spanish nurses. Index Enferm 2021;30(1-2):147-152 [FREE Full text]
47. Centre for global health and equity. Swinburne. URL: https://www.swinburne.edu.au/research/global-health-equity/ [accessed 2024-04-25]

Abbreviations

CFA: confirmatory factor analysis
CFI: comparative fit index
eHEALS: eHealth Literacy Scale
eHLQ: eHealth Literacy Questionnaire
GOFI: goodness-of-fit index
HLS-EU-Q: European Health Literacy Survey Questionnaire
RMSEA: root mean square error of approximation
TLI: Tucker-Lewis index

Edited by T de Azevedo Cardoso; submitted 22.05.23; peer-reviewed by K Yang, S Nolte, C Poot, JR Bautista; comments to author 02.08.23; revised version received 31.10.23; accepted 01.03.24; published 10.05.24.

Please cite as:
Hernández Encuentra E, Robles N, Angulo-Brunet A, Cullen D, del Arco I
Spanish and Catalan Versions of the eHealth Literacy Questionnaire: Translation, Cross-Cultural Adaptation, and Validation Study
J Med Internet Res 2024;26:e49227
URL: https://www.jmir.org/2024/1/e49227
doi:10.2196/49227
PMID:38728072

©Eulàlia Hernández Encuentra, Noemí Robles, Ariadna Angulo-Brunet, David Cullen, Ignacio del Arco. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Electronic Health Literacy Scale-Web3.0 for Older Adults with Noncommunicable Diseases: Validation Study

Wenfei Cai[1], MEd; Wei Liang[1], PhD; Huaxuan Liu[2], PhD; Rundong Zhou[1], MEd; Jie Zhang[1], PhD; Lin Zhou[3], PhD; Ning Su[1], PhD; Hanxiao Zhu[1], PhD; Yide Yang[4], PhD

[1]School of Physical Education, Shenzhen University, Shenzhen, China

[2]School of Physical Education and Sport Science, Fujian Normal University, Fuzhou, China

[3]School of Physical Education, Hebei Normal University, Shijiazhuang, China

[4]School of Medicine, Hunan Normal University, Changsha, China

Corresponding Author:
Wei Liang, PhD
School of Physical Education
Shenzhen University
3688 Nanhai Road, Nanshan District
Shenzhen, 518060
China
Phone: 86 15217940540
Email: wliang1020@szu.edu.cn

Abstract

Background: In the current digital era, eHealth literacy plays an indispensable role in health care and self-management among older adults with noncommunicable diseases (NCDs). Measuring eHealth literacy appropriately and accurately ensures the successful implementation and evaluation of pertinent research and interventions. However, existing eHealth literacy measures focus mainly on individuals' abilities of accessing and comprehending eHealth information (Web1.0), whereas the capabilities for web-based interaction (Web2.0) and using eHealth information (Web3.0) have not been adequately evaluated.

Objective: This study aimed to examine the reliability, validity, and measurement invariance of the eHealth Literacy Scale-Web3.0 (eHLS-Web3.0) among older adults with NCDs.

Methods: A total of 642 Chinese older adults with NCDs (mean age 65.78, SD 3.91 years; 55.8% female) were recruited in the baseline assessment, of whom 134 (mean age 65.63, SD 3.99 years; 58.2% female) completed the 1-month follow-up assessment. Baseline measures included the Chinese version of the 24-item 3D eHLS-Web3.0, the Chinese version of the 8-item unidimensional eHealth Literacy Scale (eHEALS), and demographic information. Follow-up measures included the 24-item eHLS-Web3.0 and accelerometer-measured physical activity and sedentary behavior. A series of statistical analyses, for example, Cronbach α, composite reliability coefficient (CR), confirmatory factor analysis (CFA), and multigroup CFA, were performed to examine the internal consistency and test-retest reliabilities, as well as the construct, concurrent, convergent, discriminant, and predictive validities, and the measurement invariance of the eHLS-Web3.0 across gender, education level, and residence.

Results: Cronbach α and CR were within acceptable ranges of 0.89-0.94 and 0.90-0.97, respectively, indicating adequate internal consistency of the eHLS-Web3.0 and its subscales. The eHLS-Web3.0 also demonstrated cross-time stability, with baseline and follow-up measures showing a significant intraclass correlation of 0.81-0.91. The construct validity of the 3D structure model of the eHLS-Web3.0 was supported by confirmatory factor analyses. The eHLS-Web3.0 exhibited convergent validity with an average variance extracted value of 0.58 and a CR value of 0.97. Discriminant validity was supported by CFA results for a proposed 4-factor model integrating the 3 eHLS-Web3.0 subscales and eHEALS. The predictive validity of the eHLS-Web3.0 for health behaviors was supported by significant associations of the eHLS-Web3.0 with light physical activity (β=.36, P=.004), moderate to vigorous physical activity (β=.49, P<.001), and sedentary behavior (β=–.26, P=.002). Finally, the measurement invariance of the eHLS-Web3.0 across gender, education level, and residence was supported by the establishment of configural, metric, strong, and strict invariances.

Conclusions: The present study provides timely empirical evidence on the reliability, validity, and measurement invariance of the eHLS-Web3.0, suggesting that the 24-item 3D eHLS-Web3.0 is an appropriate and valid tool for measuring eHealth literacy among older adults with NCDs within the Web3.0 sphere.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52457) doi:10.2196/52457

KEYWORDS

eHealth literacy; measurement; Web 3.0; psychometric properties; NCD; older adults; noncommunicable diseases; Electronic Health Literacy; health literacy; eHealth; reliability; validity

Introduction

Background

Noncommunicable diseases (NCDs), known as chronic diseases, result in the mortality of 41 million people annually, equivalent to approximately 74% of all global deaths [1]. Characterized by high morbidity, high mortality, low control rates, and limited awareness, NCDs impose a considerable financial burden on individuals, their families, and society as a whole, particularly among older patients [2]. In China, the prevalence rate of NCDs among older adults aged 60 years and older was 50%-75%, as reported in recent epidemiological studies [3-5]. Therefore, NCDs in older adults are a vital public health concern, and their management is a global challenge.

Previous evidence has demonstrated that empowering and educating patients with NCDs to focus on self-management and health promotion is essential [2,6]. Enabling patients to inquire about their medical status, comply with medication instructions, enhance their engagement and compliance in the health care process, adopt healthier lifestyles, and ultimately reduce reliance on constant supervision from health care professionals is a challenging task [6,7]. Nevertheless, facilitating patient self-care is a critical step toward improving the overall health status and alleviating the burden on health care facilities, especially within low- and middle-income countries [7,8].

With the rapid advancement of technology, the internet has become the quickest and most easily accessible resource for obtaining and delivering health information, offering ample opportunities for self-management and health promotion [6,9]. Recent review studies have consistently shown that internet-based health interventions for individuals with NCDs can have a substantial impact on enhancing self-management and patient engagement and compliance with their health care [10,11]. Despite the potential of the internet to improve health care services for NCDs, older adults encounter significant challenges in using digital health technologies [12]. In particular, the information found on the internet originates from numerous providers and sources that are difficult to regulate, thereby leading to potential problems in terms of accuracy and the potential dissemination of prejudiced content that aligns with the interests and objectives of certain parties involved [13]. Previous research has highlighted the considerable difficulties faced by older adults in accessing reliable and high-quality health information that addresses their specific health needs [14,15]. Furthermore, studies have revealed that a noteworthy proportion of older internet users lack confidence in their capability to execute basic tasks on the internet [15]. The challenges mentioned above not only impede older adults from harnessing the internet's full potential for health care purposes but also exacerbate the digital divide and health disparities [12]. In such a scenario, eHealth literacy is emphasized in numerous studies as a critical skill that older adults with NCDs must acquire in the digital era of disease management and health care [2,6,16].

eHealth literacy, first proposed by Norman and Skinner in 2006 [17], refers to "individual's abilities to seek, find, understand, and appraise health information from electronic resources and apply that knowledge to solve a health problem or make a health-related decision." The concept of eHealth literacy is founded on social cognitive theory, consisting of 6 essential skills or literacies: traditional literacy, health literacy, information literacy, scientific literacy, media literacy, and computer literacy [17,18]. To provide a general assessment of this concept that can assist in clinical decision-making and health promotion planning for individuals or specific samples, Norman and Skinner [17] developed an 8-item unidimensional eHealth Literacy Scale (eHEALS). The eHEALS is the most well-known and extensively used instrument for assessing eHealth literacy to date [19]. The reliability and validity of the eHEALS have been extensively examined in diverse cultural contexts, including English [17,20], German [21], Spanish [22], Dutch [23], Italian [24], Portuguese [25], Japanese [26], and Chinese [27], providing compelling evidence of its efficacy across multiple languages and cultures.

However, as technology continues to advance, there has been an increasing acknowledgment of the necessity to update the content of eHealth literacy to ensure optimal synchronization with the evolving internet landscape [23,28,29]. Several studies have raised concerns regarding the unidimensional nature of the eHEALS, as well as its inadequate performance in psychometric evaluations, particularly when using it to measure the usage of novel technologies in seeking and assessing health information [23,27-30]. For example, previous research has shown a weak association between eHEALS and eHealth behaviors beyond web-based information-researching skills, indicating the requirement to update the tool [23]. Furthermore, a recent systematic review indicated that the structure of the eHEALS varied across multiple studies, where a 2-factor or 3-factor structure was also identified in certain studies [19]. There has also been some questioning of the variability of the items, even though the eHEALS had the same factor construct [19]. Alongside the influence of cultural contexts, a primary reason for the inconsistencies of the factor structures and corresponding items may be that the eHEALS is outdated for use in evaluating eHealth literacy in the new digital age [19].

Indeed, the arguments outlined above are reasonable. As per the widely acknowledged generation divisions of internet evolution, the present internet landscape has progressed through 3 distinct phases, starting with Web1.0 (a read-only web) to Web2.0 (ie, a read-write mode that provides a participatory social web with increased collaboration and interaction among consumers, programmers, service providers, and organization) and to current Web3.0 (ie, a read-write-execute mode that provides digital, personalized, and intelligent services; also

XSL•FO
RenderX

known as semantic web) [23,30,31]. The eHEALS was developed 15 years ago for measuring individuals' capability related to reading and viewing within a Web1.0 context, and therefore, it is necessary to update it to effectively scale current eHealth usage.

To fill this gap, numerous new measurement tools of eHealth literacy have been developed. One example is the 20-item extended version of the eHEALS (eHEALS-E) created by Petrič et al [32], which is designed to better encompass the complicated factors contributing to eHealth literacy. However, the eHEALS-E is based on the same definition as the original eHEALS, and therefore, it may also have limitations in measuring only a narrow aspect of eHealth literacy [19]. Furthermore, second-generation instruments of eHealth literacy have been developed (eg, eHealth Literacy Scale [eHLS], Digital Health Literacy Instrument [DHLI], Transactional eHealth Literacy Instrument, eHealth Literacy Assessment Toolkit, and Chinese version of the electronic eHealth Literacy Scale [C-eHEALS]) to assess a broader spectrum of eHealth literacy concepts, ensuring their relevance in the age of social media and eHealth [19,28]. These measures have provided novel approaches for evaluating eHealth literacy, with some of them specifically designed to measure web communication capabilities. However, their coverage is limited to Web2.0 skills, and evaluation of eHealth literacy content relevant to Web3.0 technologies remains absent [33]. Recently, Liu et al [33] developed a 24-item eHealth Literacy Scale-Web3.0 (eHLS-Web3.0) to measure eHealth skills in the context of Web3.0. Compared with previous eHEALS and second-generation eHealth literacy scales, the eHLS-Web3.0 is an improvement consisting of 3 distinct dimensions (ie, acquisition, verification, and application) that evaluates the abilities covering the entire spectrum of Web1.0 (eg, searching, understanding, and identifying), 2.0 (eg, sharing and interactive communication), and 3.0 (eg, recording, self-managing, applying, and adjusting) [33]. The reliabilities, validities, and measurement invariance of the eHLS-Web3.0 across gender and region have been supported in a previous study with young adults [33], whereas its psychometric properties remain unexplored in older adults, especially those living with NCDs. Therefore, further research is needed to investigate the applicability and psychometric properties of the eHLS-Web3.0 in older adults with NCDs, which may inform the development of effective interventions to improve eHealth literacy and promote better health care outcomes in this population.

This Study

Given the above, the purpose of this study was to examine the reliability, validity, and measure invariance of the eHLS-Web3.0 in a sample of older adults with NCDs. In particular, this study has 3 main objectives. First, the internal consistency and test-retest reliabilities of the eHLS-Web3.0 would be examined. Based on previous studies [12,17,20,34-36], a conventional 1-month time frame for evaluating the test-retest reliability was used in this study. Second, the construct, convergent, concurrent, discriminant, and predictive validities of the 24-item 3D eHLS-Web3.0 would be ascertained. Because the eHEALS has been proven to be a reliable tool for measuring eHealth literacy among older adults with NCDs in previous studies [2,17], this study would use the eHEALS as the criteria scale for the examination of the concurrent validity of the eHLS-Web3.0. Furthermore, considering the content distinction between the eHLS-Web3.0 and the eHEALS, the discriminant validity of the eHLS-Web3.0 would be examined by comparing the eHLS-Web3.0 subscales and the eHEALS. Additionally, previous studies have established a strong association between eHealth literacy and various health behaviors [13,37]. Specifically, eHealth literacy has been shown to positively correlate with health-promoting behaviors (eg, physical activity) and negatively correlate with risk behaviors (eg, sedentary behavior) among young and older adults [38,39]. Therefore, this study would investigate the predictive validity of the eHLS-Web3.0 for 2 specific health behaviors (ie, physical activity and sedentary behavior), given their crucial impact on the physical and mental well-being of older adults with NCDs [40]. By identifying the predictive validity of the eHLS-Web3.0, this study is expected to make a noteworthy contribution to future research in this field. Finally, considering that gender, education level, and residence are potential correlates of eHealth literacy [6,33,41], the measurement invariance of the eHEALS would be examined at the configural, metric, strong, and strict levels across gender, education level, and residence.

Methods

Design, Participants, and Procedure

This study applied a 2-wave prospective design. Considering an item-to-response ratio of 1:10 and the recommendation for a minimum sample size of 200 in confirmatory factor analysis (CFA) [42], 240 participants were required to ensure a robust statistical estimation. With an approximate response rate of 85% and a prior estimated prevalence rate of NCDs of 50% in older adults [4], a minimum of 564 participants were required to be contacted at the initial recruitment stage. Eligible participants for the study were older adults who met the following inclusion criteria: (1) aged 60 years or older, (2) experiencing at least 1 type of NCD (eg, cardiovascular diseases, cancer, type 2 diabetes, and obesity), (3) no physical mobility restrictions, (4) no cognitive disorders, (5) proficient in reading and understanding Chinese, and (6) having access to a smartphone or laptop.

Participants were recruited from the outpatient departments of 6 hospitals from 3 cities (Shiyan, Wuhan, and Suizhou) of Hubei Province (Central China) using a convenience sampling approach. The survey was implemented using the SOJUMP web-based survey platform (Changsha Ranxing Information Technology Co, Ltd). Two health care professionals undertook an initial review of the scale items to ensure that the wording was appropriate for older adults with NCDs. Subsequently, 6 older adults with NCDs (3 female and 3 male) were invited to complete a pilot assessment aimed at (1) optimizing the design of the electronic questionnaires (eg, using the large font and highlighting the key information) and (2) refining the language and eliminating any errors to ensure that the scale items were easily comprehensible for the target population.

In the main study, participants were provided with a QR code through nurses to gain access to the web-based survey. Before

XSL•FO
RenderX

answering the questionnaires, participants were required to sign an informed consent form on the first page of the survey. The web-based survey lasted approximately 20 minutes. To ensure a robust evaluation for the scale test-retest reliability and predictive validity, a minimum of 100 participants were needed [43]. Accounting for a potential 30% attrition rate (eg, no response and invalid or missing data) [4], a total of 142 participants were required for the second-wave data collection. Invitations were sent out randomly via SMS text messages to those who had completed the first round of data collection until enough participants agreed to participate in the follow-up survey, scheduled for 1 month later. Participants who agreed to participate in the second round of investigation were requested to revisit the hospital, where 2 qualified assistants guided them to complete the follow-up web-based survey and provided detailed instructions on the use of the accelerometer for data collection.

Ethical Considerations

This study adhered to the principles outlined in the Declaration of Helsinki by the World Medical Association. The Medical Ethics Committee of the Faculty of Medicine at Shenzhen University reviewed and approved this study (PN-202300066). All participants provided signed informed consent for both the primary study and the sensitivity analyses. The data were anonymized to protect participant privacy, and participation in the study was entirely voluntary. As a token of appreciation, participants received a participation fee of 5 RMB (US $0.7) on completing the data collection.

Measures

eHealth Literacy Scale-Web3.0

The 24-item eHLS-Web3.0 was originally developed by Liu et al [33] for the Chinese adult population. This scale comprises 3 dimensions: acquisition (items 1-4 and 11-14), verification (items 5-10), and application (items 15-24). Responses were indicated on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). The total score of the scale ranges from 24 to 120, with a higher score indicating a greater level of eHealth literacy. The reliability and validity of the eHLS-Web3.0 have been fully supported by previous research with Chinese young adults (Cronbach α=0.91-0.97).

eHEALS

The 8-item eHEALS was developed by Norman and Skinner [17] for use among Canadian adolescents. The original scale is unidimensional and has been validated in various countries across diverse populations. The Chinese version of the 8-item eHEALS has been examined in previous studies among older adults with NCDs, where the reliability and validity of the scale have been fully supported (Cronbach α=0.95-0.98).

Health Behaviors

Physical activity and sedentary behavior were measured using the ActiGraph GT3X+ (ActiGraph) on the right side of the waist for 7 consecutive days, with the exception of swimming, bathing, and sleeping time. The accelerometer sampling interval was set at 60-second epochs with a sampling frequency of 30 Hz. Nonwear time was defined by an interval of 60 consecutive minutes of 0 counts per minute, allowing for 2 minutes of nonzero count interruptions. Participants with at least 3 valid days of accelerometer use (2 weekdays and 1 weekend day) and a minimum wear time of 10 hours per day were eligible for inclusion in the data analysis. The Freedson cutoff point was used for categorizing light physical activity (100-1951 counts/minute), moderate to vigorous physical activity (>1951 counts/minute), and sedentary behavior (<100 counts/minute) [44].

Demographics

The demographic information included age, sex, marital status, education level, residence, monthly income, living situation, and BMI.

Statistical Analyses

The data analyses were performed using IBM SPSS Statistics (version 28.0; IBM Corp) and Mplus 8 (Muthén & Muthén). Data screening and diagnosis tests of data distribution (eg, mean, SD, skewness, and kurtosis) and missing patterns were performed before the descriptive analysis and scale validation. To ensure a reliable estimation for the multidimensional scale, both Cronbach α and composite reliability (CR) coefficients were calculated to evaluate the internal consistency reliability of the eHLS-Web3.0. Additionally, the test-retest reliability was estimated using the intraclass correlation coefficient of pre- and 1-month follow-up data.

The construct validity of the eHLS-Web3.0 was evaluated using CFAs with maximum likelihood estimation. Several goodness-of-fit indices were computed, including robust chi-square (χ^2_R), robust chi-square to degrees of freedom ratio (χ^2_R/df), comparative-fit index (CFI), Tucker-Lewis index (TLI), root-mean-square error of approximation (RMSEA) and its 90% CI, and standardized root-mean-square residual (SRMR). The following criteria were considered for a satisfactory model goodness of fit: ≤3 for χ^2_R/df, ≥0.9 for CFI and TLI, and ≤0.08 for RMSEA and SRMR [45].

Convergent validity was assessed by examining the average variance extracted (AVE) and CR for each subscale, with AVE >0.5 and CR >0.7 indicating satisfactory convergent validity for the scale. Concurrent validity was assessed by calculating the zero-order correlations of the eHLS-Web3.0 and its subscales with the eHEALS, adjusted for all demographic confounders. For the discriminant validity, a presumptive 4-factor model incorporating 3 eHLS-Web3.0 subscales and unidimensional eHEALS was estimated in the CFA. The discriminant validity of the subscales was confirmed if the 95% CI of the association between these subscales did encompass the value of 0 and if the Wald chi-square test demonstrated a significant change in model fit after removing a constraint that fixed the factor correlation to zero [45]. Additionally, structural equation models were performed to assess the predictive validity of the scale by estimating its association with health behaviors, including physical activity and sedentary behavior.

With a sequential model testing approach, multigroup CFA was used to examine the measurement invariance of the eHLS-Web3.0 across gender, education, and residence. Four

XSL•FO
RenderX

distinctive levels of measurement invariance were examined by progressively constraining the parameter estimates of the models to be equivalent across the samples: (1) configural invariance, where no parameter estimates were restricted to equality; (2) metric invariance, where factor loadings were constrained to equality; (3) strong invariance, where both factor loadings and item intercepts were constrained to equality; and (4) structural and strict invariance, where all factor loadings, item intercepts, and factor variance and covariance were restricted to equality. The measure invariance was supported if the change in the value of CFI and RMSEA was ≤0.01 and ≤0.015, respectively [45,46].

Results

Descriptive Information of the Study Sample

As outlined in Figure 1, a total of 642 eligible participants (mean 65.78, SD 3.91 years; 55.8% female) were included in the data analysis, of whom 134 (mean 65.63, SD 3.99 years; 58.2% female) provided valid data at the follow-up assessment. From the diagnostic evaluation, there were no missing data for eHLS-Web3.0 and eHEALS items in the study sample. All the scale items adhered to the normality distribution with absolute values of skewness and kurtosis <1. Descriptive information of the study sample is shown in Table 1.

Figure 1. Study process diagram.

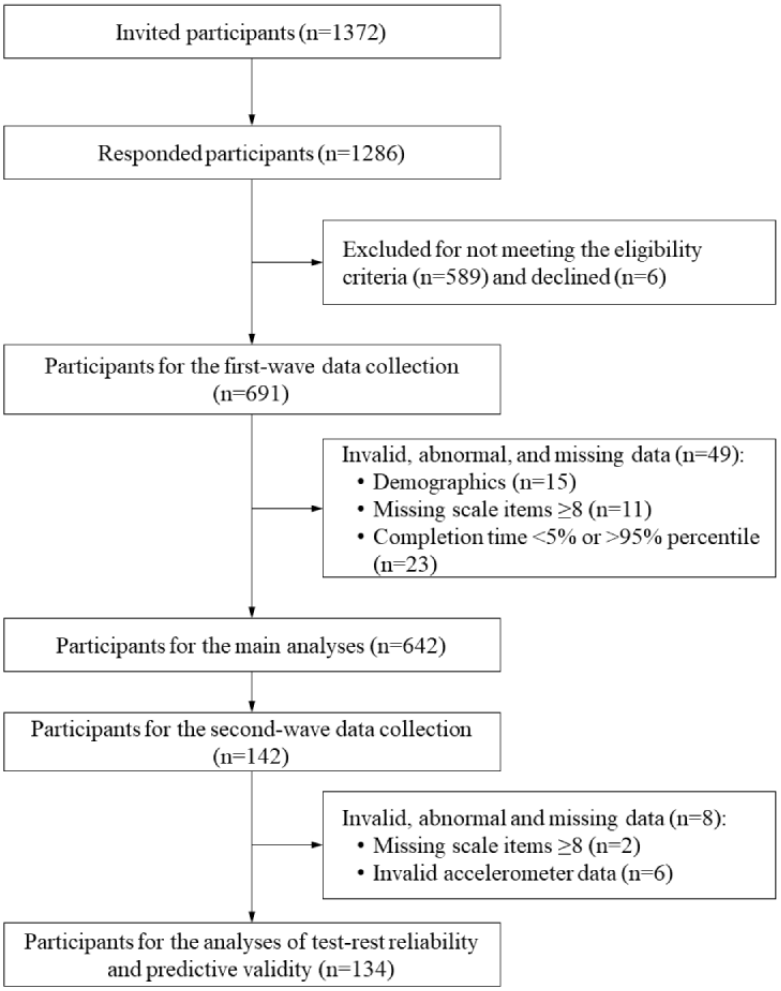

XSL•FO
RenderX

Table 1. Descriptive characteristics of the study sample.

Demographic information	Main study (n=642)	Test-retest (n=134)
Age (years), mean (SD)	65.78 (3.91)	65.63 (3.99)
Age range (years)	60-74	60-74
Gender, n (%)		
Female	358 (55.8)	73 (54.5)
Male	284 (44.2)	61 (45.5)
Marital status, n (%)		
Married	597 (93)	123 (91.8)
Single, divorced, or widowed	45 (7)	11 (8.2)
Education level, n (%)		
Primary	121 (18.8)	26 (19.4)
Middle and high school	134 (20.9)	21 (15.7)
College and above	387 (60.3)	87 (64.9)
Residence, n (%)		
Urban	394 (61.4)	82 (61.2)
Rural	248 (38.6)	52 (38.8)
Monthly income (RMB), n (%)		
≤1000 (US $138.1)	68 (10.6)	15 (11.2)
1001-3000 (US $138.2-$414.2)	149 (23.2)	29 (21.6)
3001-5000 (US $414.3-$690.3)	166 (25.9)	37 (27.6)
5001-7000 (US $690.4-$966.4)	186 (29.0)	35 (26.1)
≥7001 (US $966.5)	73 (11.4)	18 (13.4)
Living situation, n (%)		
Living with others	575 (89.6)	117 (87.3)
Living alone	67 (10.4)	17 (12.7)
Number of NCDs[a]		
1	531 (82.7)	108 (80.6)
2-3	93 (14.5)	17 (12.7)
>3	18 (2.8)	9 (6.7)
BMI (kg/m^2), mean (SD)	22.95 (2.82)	22.46 (2.41)
BMI category, n (%)		
Underweight	36 (5.6)	6 (4.5)
Normal weight	410 (63.9)	87 (64.9)
Overweight and obese	196 (30.5)	41 (30.6)
Health behavior, mean (SD)		
LPA[b] (minutes/day), range=45-420	N/A[c]	145.93 (89.42)
MVPA[d] (minutes/day), range=4-98	N/A	26.82 (24.40)
Sedentary behavior (minutes/day), range=175-865	N/A	573.49 (135.46)

[a]NCD: noncommunicable disease.
[b]LPA: light physical activity.
[c]N/A: not applicable.
[d]MVPA: moderate to vigorous physical activity.

XSL•FO
RenderX

Reliabilities of the eHLS-Web3.0 in the Study Sample

Table 2 shows the mean value, SD, score range, and internal consistency and test-retest reliabilities of the eHLS-Web3.0 in the study sample. Regarding the internal consistency reliability, the eHLS-Web3.0 and its 3 subscales exhibited adequate Cronbach α values (range=0.89-0.94) and CR coefficients (range=0.90-0.97). Regarding the test-retest reliability, 2 time-point measures showed a strong intraclass correlation for the eHLS-Web3.0 and 3 subscales of the eHLS-Web3.0 (*r*=0.81-0.91).

Table 2. Mean (SD), score range, Cronbach α, composite (ρ), and test-retest (*r*) reliabilities of the 24-item eHealth Literacy Scale-Web3.0 (eHLS-Web3.0) in the study sample (n=642).

Scale	Mean (SD)	Range	α	ρ	*r* (n=134)
eHLS-Web3.0	78.61 (11.1)	12-40	0.94	0.97	0.91
Acquisition	27.14 (4.26)	9-30	0.92	0.91	0.87
Verification	19.89 (3.67)	20-50	0.89	0.90	0.81
Application	31.57 (5.08)	42-115	0.93	0.93	0.90

Validities of the eHLS-Web3.0 in the Study Sample

For the constructive validity, the results of the CFA showed that the 24-item 3D eHLS-Web3.0 achieved the criteria for good model fit indices in the study sample, with χ^2_R=674.4, *df*=248, χ^2_R/df=2.72 (<3), CFI=0.952 (>0.9), TLI=0.946 (>0.9), RMSEA=0.052 (90% CI 0.047-0.056; <0.08), and SRMR=0.034 (<0.08). The standardized factor loadings of the eHLS-Web3.0 items ranged from 0.658 to 0.819 (see Table S1 in Multimedia Appendix 1).

For concurrent validity, the zero-order correlations between the eHLS-Web3.0 subscales and eHEALS were significant (*r*=0.47-0.76), indicating a satisfactory result, as outlined in Table 3. The AVE and CR were calculated as 0.58 and 0.97, respectively, using the standardized factor loadings of the items, showing a satisfactory convergent validity of the eHLS-Web3.0.

The latent factor correlations in a proposed 4-factor CFA model (ie, 3 eHLS-Web3.0 subscales and eHEALS) were calculated to evaluate the discriminant validity of the eHLS-Web3.0 subscales and eHEALS in the study sample. The goodness-of-fit indices of the 4-factor model were inferior to those of the original 3-factor eHLS-Web3.0 model, with χ^2_R=2079.8, *df*=461, χ^2_R/df=4.51, CFI=0.874, TLI=0.865, RMSEA=0.074 (90% CI 0.071-0.077), and SRMR=0.078. Table 4 presents the statistical results of the discriminant validity analysis. Wald chi-square findings were statistically significant in the study sample (all *P*<.001), and the CIs for each correlation did not contain the value of 0, demonstrating a satisfactory discriminant validity of the eHLS-Web3.0.

In addition, the eHLS-Web3.0 significantly predicted light physical activity (β=.36, 95% CI 0.19-0.53; *P*=.004), moderate to vigorous physical activity (β=.49, 95% CI 0.35-0.62; *P*<.001), and sedentary behavior (β=–.26, 95% CI–0.40 to –0.12; *P*=.002), supporting the predictive validity of the scale for health behaviors. The goodness-of-fit indices indicated a satisfactory result for the 3 models, with χ^2_R=398.3-403.8, *df*=271, χ^2_R/df=1.47-1.49, CFI=0.949-0.951, TLI=0.944-0.946, RMSEA=0.059-0.060, and SRMR=0.043-0.044.

Table 3. Zero-order correlation between the subscales of the 24-item eHealth Literacy Scale-Web3.0 (eHLS-Web3.0) and 8-item eHealth Literacy Scale (eHEALS), n=642.

Scale	eHEALS	eHLS-Web3.0	Acquisition	Verification	Application
eHEALS	1				
eHLS-Web3.0	0.69[a]	1			
Acquisition	0.76[a]	0.87[a]	1		
Verification	0.48[a]	0.79[a]	0.56[a]	1	
Application	0.57[a]	0.88[a]	0.64[a]	0.52[a]	1

[a]*P*<.001.

XSL•FO
RenderX

Table 4. Latent interfactor correlations and discriminant validity statistics for the 24-item eHealth Literacy Scale-Web3.0 in the study sample (n=642).

Scale	Interfactor correlation	95% CI	Wald chi-square test[a]
Acquisition↔Verification	0.86	0.84-0.88	197.4[b]
Acquisition↔Application	0.91	0.89-0.92	155.8[b]
Verification↔Application	0.84	0.82-0.86	159.0[b]
Acquisition↔eHEALS[c]	0.94	0.93-0.95	62.3[b]
Verification↔eHEALS	0.82	0.79-0.84	151.7[b]
Application↔eHEALS	0.85	0.83-0.87	113.6[b]

[a]Wald chi-square test: Wald chi-square test constraining the values of the latent interfactor correlations to zero.

[b]$P<.001$.

[c]eHEALS: 8-item eHealth Literacy Scale.

Measurement Invariance of the eHLS-Web3.0 in the Study Sample

Table 5 presents the results of the examination regarding the measurement invariance of the eHLS-Web3.0 across gender, education level, and residence. The configural, metric, strong, and strict models were all shown to have a satisfactory fit to the data for all 3 pairs of subsamples, with ΔCFI <0.01 and ΔRMSEA <0.015. These indices provide support for the invariance of the factorial construct, factor loadings, intercepts, and residual variance of the eHLS-Web3.0 across gender, education level, and residence.

Table 5. Measurement invariance of the 24-item eHealth Literacy Scale-Web3.0 across gender, education level, and residence.

Model	Chi-square[a] (*df*)	CFI[b]	ΔCFI[c]	RMSEA[d]	RMSEA 90% CI	ΔRMSEA[e]
Female (n=358)—Male (n=254)						
M0[f]	989.0 (496)	0.944	N/A[g]	0.056	0.051-0.061	N/A
M1[h]	1011.6 (517)	0.944	0	0.055	0.050-0.060	–0.001
M2[i]	1068.1 (541)	0.940	–0.004	0.055	0.050-0.060	0
M3[j]	1072.2 (547)	0.940	0	0.055	0.050-0.050	0
Below college (n=255)—College and above (n=387)						
M0	966.9 (496)	0.948	N/A	0.054	0.049-0.059	N/A
M1	993.0 (517)	0.947	–0.001	0.054	0.049-0.059	0
M2	1017.6 (541)	0.947	0	0.052	0.047-0.057	–0.002
M3	1026.8 (547)	0.946	–0.001	0.052	0.047-0.057	0
Rural (n=248)—Urban (n=394)						
M0	955.3 (496)	0.947	N/A	0.054	0.049-0.059	N/A
M1	978.5 (517)	0.947	0	0.053	0.048-0.058	0.001
M2	1027.5 (541)	0.944	–0.003	0.053	0.048-0.058	0
M3	1056.4 (547)	0.941	–0.003	0.054	0.049-0.059	–0.001

[a]Chi-square: robust chi-square.

[b]CFI: comparative fit index.

[c]ΔCFI: change in the CFI.

[d]RMSEA: root-mean-square error of approximation.

[e]ΔRMSEA: change in the RMSEA.

[f]M0: baseline configural invariance model.

[g]N/A: not applicable.

[h]M1: metric invariance model.

[i]M2: strong invariance model.

[j]M3: strict invariance model.

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to evaluate the reliability, validity, and measurement invariance of the eHLS-Web3.0 for use with older adults who are living with NCDs. In particular, this study examined the internal consistency and test-retest reliabilities, as well as the construct, concurrent, convergent, discriminant, and predictive validities, and the measurement invariance of the eHLS-Web3.0 across gender, education level, and residence. Overall, the results from this study suggest that the eHLS-Web3.0 is a reliable and valid tool for measuring eHealth literacy in Chinese older adults with NCDs.

Regarding the reliabilities, analyses of the Cronbach α and CR coefficients indicated adequate internal consistency reliability for both the eHLS-Web3.0 and its 3 subscales. These findings are consistent with previous research on the use of the eHLS-Web3.0 among Chinese young adults [33]. It is worth noting that while previous studies have generally supported the reliability of the eHEALS and other eHealth literacy assessments (eg, DHLI and C-eHEALS) among older adults or those with NCDs [2,27,28], the novel eHEALS-Web3.0 tool has not yet been evaluated for reliability in older populations. This study is the first to investigate the reliability of the eHEALS-Web3.0 among older adults with NCDs. Additionally, previous psychometric analyses of eHealth literacy measures have primarily focused on internal consistency reliability, with test-retest reliability often overlooked [19,27,28]. In contrast, this study further examined the test-retest reliability of the eHLS-Web3.0 and its subscales, and the findings demonstrated a strong cross-time stability for the scale, as evidenced by a significant correlation between baseline and 1-month follow-up measures.

Regarding the construct validity, the CFA results provided support for the 3D model structure of the eHLS-Web3.0 among Chinese older adults with NCDs. The acquisition and verification subscales of the eHEALS-Web3.0 assess individuals' eHealth abilities in Web1.0 and Web2.0 contexts, similar to the eHEALS and second-general eHealth literacy measures [12,19,27]. However, the eHEALS-Web3.0 stands out by also evaluating individuals' proficiency in applying eHealth information to evolving health needs in the Web3.0 era (ie, the application subscale). As the digital landscape advances, individuals have more opportunities and options to use eHealth information. For instance, they can use eHealth information to make informed health decisions or resolve health-related problems, create their own health data, monitor their health status, interact with others, exchange information, and provide health advice to other health information seekers [31,33]. The 3D eHLS-Web3.0 provides a comprehensive assessment of eHealth literacy, catering to the present digital circumstances.

For other validities of the eHLS-Web3.0, the concurrent validity was confirmed by a significant correlation between the eHLS-Web3.0 and its subscales with the eHEALS, while the AVE and CR supported the convergent validity of the scale. In addition, as the 3 eHLS-Web3.0 subscales and eHEALS differ in conceptual content, a 4-factor model integrating the acquisition, verification, and application subscales and unidimensional eHEALS was established to confirm the discriminant validity of the scale. Although the Wald chi-square test results supported the discriminant validity of the eHLS-Web3.0, a high correlation was observed among the latent factors in the 4-factor model, possibly due to measurement errors [47]. To validate the earlier findings, zero-order correlations were calculated using composite (averaged) scales. Fortunately, the overall results confirmed the discriminant validity of the scale. Finally, the predictive validity of the eHLS-Web3.0 was supported by a significant positive association between the eHLS-Web3.0 and physical activity, as well as a negative association with sedentary behavior. Previous studies have demonstrated a positive association between eHealth literacy and health-promoting behaviors (eg, physical activity) among diverse populations [13,37,48]. However, there is a lack of evidence on the relationship between eHealth literacy and risk behaviors (eg, sedentary behavior). Our findings underline the potential of including eHealth literacy as a modifiable factor in future eHealth interventions to facilitate health behaviors and improve health outcomes among older adults with NCDs.

For measurement invariance, the establishment of configural, metric, strong, and strict invariances demonstrated that the eHLS-Web3.0 is a psychometrically sound instrument for measuring eHealth literacy among Chinese older adults with NCDs, regardless of their gender, education level, and residence. These invariances provide a solid foundation for making appropriate and meaningful transgroup comparisons in future studies.

Limitations and Implications

Some limitations should be noted. First, the nonrandom sampling used in this study may have limited the representativeness of the study findings. Therefore, a stratified random sampling approach is warranted in future studies. Second, given that the study findings are based only on the sample of Chinese older adults with NCDs, one should be cautious when generalizing these results to other samples. Future studies should examine the psychometric properties of the scale across different populations and diverse cultural contexts. Moreover, self-reported measures may result in some response biases (eg, recall bias and social desirability); therefore, the inclusion of objective means for assessing eHealth literacy should be considered in the future. Additionally, it is worth exploring the prediction of eHealth literacy on other health outcomes and examining its underlying mechanisms. Finally, from a pragmatic perspective, it may be beneficial to develop and validate a brief version of the eHLS-Web3.0, particularly with regard to older populations who may struggle with completing lengthy self-reported scales.

Despite the aforementioned limitations, this study addresses a significant gap in the literature by validating and applying the eHLS-Web3.0, a specific measure of eHealth literacy used for Chinese older adults with NCDs in the Web3.0 landscape. Previous reviews have revealed a wide range of influential factors of eHealth literacy as well as a positive correlation between higher eHealth literacy and better health behaviors,

XSL•FO
RenderX

knowledge, and attitudes in older adults [38,49-51]. These findings indicate the potential for developing eHealth literacy interventions to promote positive health behaviors in the future while considering various socioeconomic and cultural variables. However, previous studies have yielded conflicting results regarding certain physical and psychosocial outcomes [38], underlining the need for more high-quality research. It is important to note that the success of these efforts largely depends on a reliable and accurate assessment of eHealth literacy [52].

The findings of this study provide robust support for the reliability, validity, and measurement invariance of the eHLS-Web3.0, indicating that this up-to-date tool can be widely used in future research endeavors to appropriately and accurately assess older adults' abilities to search for, retrieve, evaluate, and use web-based health resources. This advancement has the potential to significantly contribute to both the field of eHealth literacy research and the development of targeted health promotion programs in the future. As digital technology increasingly infiltrates the health care sector, promoting eHealth literacy among older adults is more critical than ever [52,53]. The development and validation of the eHLS-Web3.0 marks a significant milestone in the field of eHealth literacy research, serving as a necessary foundation for future empirical investigations and targeted interventions aimed at improving eHealth literacy among older adults.

Conclusions

To the best of our knowledge, this is the first study to examine the psychometric properties and measurement invariance of the eHLS-Web3.0 among Chinese older adults with NCDs. This study provides evidence for internal consistency and test-retest reliabilities, construct, concurrent, convergent, discriminant, and predictive validities, and the measurement invariance of the 24-item 3D eHLS-Web3.0 for use with Chinese older adults with NCDs. The eHLS-Web3.0 can serve as a psychometrically sound instrument for assessing eHealth literacy in the Chinese context.

Acknowledgments

This research was funded by the Humanities and Social Science Fund of Ministry of Education of China (23YJCZH121), as well as the Humanities and Social Sciences Revitalization Grant of Shenzhen University (WKZX0312). The funding organizations had no role in the study design, study implementation, manuscript preparation, or publication decision. This work is the responsibility of the authors.

Authors' Contributions

WL, WC, and HL conceived and designed the study. HZ, JZ, RZ, and LZ contributed to the preparation of the study materials. WL, WC, NS, HZ, YY, and LZ collected the data. WL screened and analyzed the data. WL and WC drafted and revised the manuscript. All authors have reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standardized factor loadings of the 24-item eHealth Literacy Scale-Web3.0.
[PDF File (Adobe PDF File), 189 KB - jmir_v26i1e52457_app1.pdf]

References

1. Noncommunicable diseases. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2023-08-10]
2. Wu Y, Wen J, Wang X, Wang Q, Wang W, Wang X, et al. Associations between e-health literacy and chronic disease self-management in older Chinese patients with chronic non-communicable diseases: a mediation analysis. BMC Public Health 2022;22(1):2226 [FREE Full text] [doi: 10.1186/s12889-022-14695-4] [Medline: 36447176]
3. Liu J, Yu W, Zhou J, Yang Y, Chen S, Wu S. Relationship between the number of noncommunicable diseases and health-related quality of life in Chinese older adults: a cross-sectional survey. Int J Environ Res Public Health 2020;17(14):5150 [FREE Full text] [doi: 10.3390/ijerph17145150] [Medline: 32708844]
4. Duan Y, Shang B, Liang W, Lin Z, Hu C, Baker JS, et al. Predicting hand washing, mask wearing and social distancing behaviors among older adults during the covid-19 pandemic: an integrated social cognition model. BMC Geriatr 2022;22(1):91 [FREE Full text] [doi: 10.1186/s12877-022-02785-2] [Medline: 35109798]
5. Gong JB, Yu XW, Yi XR, Wang CH, Tuo XP. Epidemiology of chronic noncommunicable diseases and evaluation of life quality in elderly. Aging Med (Milton) 2018;1(1):64-66 [FREE Full text] [doi: 10.1002/agm2.12009] [Medline: 31942482]
6. Shiferaw KB, Tilahun BC, Endehabtu BF, Gullslett MK, Mengiste SA. E-health literacy and associated factors among chronic patients in a low-income country: a cross-sectional survey. BMC Med Inform Decis Mak 2020;20(1):181 [FREE Full text] [doi: 10.1186/s12911-020-01202-1] [Medline: 32762745]

XSL•FO
RenderX

7. Powers MA, Bardsley J, Cypress M, Duker P, Funnell MM, Fischl AH, et al. Diabetes self-management education and support in type 2 diabetes. Diabetes Educ 2017;43(1):40-53. [doi: 10.1177/0145721716689694] [Medline: 28118121]
8. Dwarswaard J, Bakker EJM, van Staa AL, Boeije HR. Self-management support from the perspective of patients with a chronic condition: a thematic synthesis of qualitative studies. Health Expect 2016;19(2):194-208 [FREE Full text] [doi: 10.1111/hex.12346] [Medline: 25619975]
9. Bujnowska-Fedak M, Waligóra J, Mastalerz-Migas A. The internet as a source of health information and services. Adv Exp Med Biol 2019;1211:1-16. [doi: 10.1007/5584_2019_396] [Medline: 31273574]
10. Duff OM, Walsh DM, Furlong BA, O'Connor NE, Moran KA, Woods CB. Behavior change techniques in physical activity eHealth interventions for people with cardiovascular disease: systematic review. J Med Internet Res 2017;19(8):e281 [FREE Full text] [doi: 10.2196/jmir.7782] [Medline: 28768610]
11. Duan Y, Shang B, Liang W, Du G, Yang M, Rhodes RE. Effects of eHealth-based multiple health behavior change interventions on physical activity, healthy diet, and weight in people with noncommunicable diseases: systematic review and meta-analysis. J Med Internet Res 2021;23(2):e23786 [FREE Full text] [doi: 10.2196/23786] [Medline: 33616534]
12. Xie L, Mo PKH. Comparison of eHealth Literacy Scale (eHEALS) and Digital Health Literacy Instrument (DHLI) in assessing electronic health literacy in Chinese older adults: a mixed-methods approach. Int J Environ Res Public Health 2023;20(4):3293 [FREE Full text] [doi: 10.3390/ijerph20043293] [Medline: 36833987]
13. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023;25:e40778 [FREE Full text] [doi: 10.2196/40778] [Medline: 36716080]
14. Stellefson M, Chaney B, Barry AE, Chavarria E, Tennant B, Walsh-Childers K, et al. Web 2.0 chronic disease self-management for older adults: a systematic review. J Med Internet Res 2013;15(2):e35 [FREE Full text] [doi: 10.2196/jmir.2439] [Medline: 23410671]
15. Stellefson M, Paige SR, Tennant B, Alber JM, Chaney BH, Chaney D, et al. Reliability and validity of the telephone-based eHealth Literacy Scale among older adults: cross-sectional survey. J Med Internet Res 2017;19(10):e362 [FREE Full text] [doi: 10.2196/jmir.8481] [Medline: 29074471]
16. Richtering SS, Hyun K, Neubeck L, Coorey G, Chalmers J, Usherwood T, et al. eHealth literacy: predictors in a population with moderate-to-high cardiovascular risk. JMIR Hum Factors 2017;4(1):e4 [FREE Full text] [doi: 10.2196/humanfactors.6217] [Medline: 28130203]
17. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
18. Bandura A. Self-Efficacy: the Exercise of Control. New York: W. H. Freeman; 1997.
19. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
20. Chung SY, Nahm ES. Testing reliability and validity of the eHealth Literacy Scale (eHEALS) for older adults recruited online. Comput Inform Nurs 2015;33(4):150-156 [FREE Full text] [doi: 10.1097/CIN.0000000000000146] [Medline: 25783223]
21. Marsall M, Engelmann G, Skoda EM, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a revised German version of the eHealth Literacy Scale. J Med Internet Res 2022;24(2):e28252 [FREE Full text] [doi: 10.2196/28252] [Medline: 35107437]
22. Pérez GP, Almagro BJ, Gómez Á, Gómez JIA. Validación de la escala eHealth Literacy (eHEALS) en población universitaria española [Validation of the eHealth Literacy Scale (eHEALS) in Spanish University Students]. Rev Esp Salud Publica 2015;89(3):329-338 [FREE Full text] [doi: 10.4321/S1135-57272015000300010] [Medline: 26388346]
23. van der Vaart R, van Deursen AJ, Drossaert CH, Taal E, van Dijk JA, van de Laar MA. Does the eHealth Literacy Scale (eHEALS) measure what it intends to measure? Validation of a Dutch version of the eHEALS in two adult populations. J Med Internet Res 2011;13(4):e86 [FREE Full text] [doi: 10.2196/jmir.1840] [Medline: 22071338]
24. Diviani N, Dima AL, Schulz PJ. A psychometric analysis of the Italian version of the eHealth Literacy Scale using item response and classical test theory methods. J Med Internet Res 2017;19(4):e114 [FREE Full text] [doi: 10.2196/jmir.6749] [Medline: 28400356]
25. Mialhe FL, Moraes KL, de Carvalho Sampaio HA, Brasil VV, da Silva Carvalho Vila V, Soares GH, et al. Evaluating the psychometric properties of the eHealth Literacy Scale in Brazilian adults. Rev Bras Enferm 2021;75(1):e20201320 [FREE Full text] [doi: 10.1590/0034-7167-2020-1320] [Medline: 34495134]
26. Mitsutake S, Shibata A, Ishii K, Okazaki K, Oka K. [Developing Japanese version of the eHealth Literacy Scale (eHEALS)]. Nihon Koshu Eisei Zasshi 2011;58(5):361-371. [Medline: 21905612]
27. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the simplified Chinese eHealth Literacy Scale: cross-sectional study. J Med Internet Res 2020;22(12):e18613 [FREE Full text] [doi: 10.2196/18613] [Medline: 33284123]
28. Chang A, Schulz PJ. The measurements and an elaborated understanding of Chinese eHealth Literacy (C-eHEALS) in chronic patients in China. Int J Environ Res Public Health 2018;15(7):1553 [FREE Full text] [doi: 10.3390/ijerph15071553] [Medline: 30041420]

XSL•FO
RenderX

29. Hyde LL, Boyes AW, Evans TJ, Mackenzie LJ, Sanson-Fisher R. Three-factor structure of the eHealth Literacy Scale among magnetic resonance imaging and computed tomography outpatients: a confirmatory factor analysis. JMIR Hum Factors 2018;5(1):e6 [FREE Full text] [doi: 10.2196/humanfactors.9039] [Medline: 29459356]
30. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
31. Rudman R, Bruwer R. Defining Web 3.0: opportunities and challenges. Electron Libr 2016;34(1):132-154. [doi: 10.1108/el-08-2014-0140]
32. Petrič G, Atanasova S, Kamin T. Ill literates or illiterates? Investigating the eHealth literacy of users of online health communities. J Med Internet Res 2017;19(10):e331 [FREE Full text] [doi: 10.2196/jmir.7372] [Medline: 28978496]
33. Liu HX, Chow BC, Liang W, Hassel H, Huang YW. Measuring a broad spectrum of eHealth skills in the Web 3.0 context using an eHealth Literacy Scale: development and validation study. J Med Internet Res 2021;23(9):e31627 [FREE Full text] [doi: 10.2196/31627] [Medline: 34554098]
34. Zrubka Z, Hajdu O, Rencz F, Baji P, Gulácsi L, Péntek M. Psychometric properties of the Hungarian version of the eHealth Literacy Scale. Eur J Health Econ 2019;20(Suppl 1):57-69 [FREE Full text] [doi: 10.1007/s10198-019-01062-1] [Medline: 31098883]
35. Kim H, Yang E, Ryu H, Kim HJ, Jang SJ, Chang SJ. Psychometric comparisons of measures of eHealth literacy using a sample of Korean older adults. Int J Older People Nurs 2021;16(3):e12369. [doi: 10.1111/opn.12369] [Medline: 33527701]
36. Collins LM. Research design and methods. In: Birren JE, editor. Encyclopedia of Gerontology (Second Edition). Amsterdam: Elsevier/Academic Press; 2007:433-442.
37. Cui GH, Li SJ, Yin YT, Chen LJ, Li JQ, Liang FY, et al. The relationship among social capital, eHealth literacy and health behaviours in Chinese elderly people: a cross-sectional study. BMC Public Health 2021;21(1):45 [FREE Full text] [doi: 10.1186/s12889-020-10037-4] [Medline: 33407275]
38. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
39. Hutchesson MJ, Gough C, Müller AM, Short CE, Whatnall MC, Ahmed M, et al. eHealth interventions targeting nutrition, physical activity, sedentary behavior, or obesity in adults: a scoping review of systematic reviews. Obes Rev 2021;22(10):e13295. [doi: 10.1111/obr.13295] [Medline: 34159684]
40. Li Y, Fan X, Wei L, Yang K, Jiao M. The impact of high-risk lifestyle factors on all-cause mortality in the US non-communicable disease population. BMC Public Health 2023;23(1):422 [FREE Full text] [doi: 10.1186/s12889-023-15319-1] [Medline: 36864408]
41. Lee J, Tak SH. Factors associated with eHealth literacy focusing on digital literacy components: a cross-sectional study of middle-aged adults in South Korea. Digit Health 2022;8:20552076221102765 [FREE Full text] [doi: 10.1177/20552076221102765] [Medline: 35615270]
42. Byrne BM. Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Second Edition. New York: Routledge, Taylor & Francis Group; 2013.
43. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17(1):101-110. [doi: 10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e] [Medline: 9463853]
44. Peters TM, Moore SC, Xiang YB, Yang G, Shu XO, Ekelund U, et al. Accelerometer-measured physical activity in Chinese adults. Am J Prev Med 2010;38(6):583-591 [FREE Full text] [doi: 10.1016/j.amepre.2010.02.012] [Medline: 20494234]
45. Liang W, Wang DD, Shang BR, Zhang CQ, Duan YP, Si GY. Further examination of the psychometric properties of the Brief Self-Control Scale: evidence from Chinese athletes and students. Int J Sport Exerc Psychol 2020;20(1):16-35. [doi: 10.1080/1612197x.2020.1827000]
46. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Model Multidiscip J 2002;9(2):233-255. [doi: 10.1207/S15328007SEM0902_5]
47. Hagger MS, Hankonen N, Kangro EM, Lintunen T, Pagaduan J, Polet J, et al. Trait self-control, social cognition constructs, and intentions: correlational evidence for mediation and moderation effects in diverse health behaviours. Appl Psychol Health Well Being 2019;11(3):407-437. [doi: 10.1111/aphw.12153] [Medline: 30724028]
48. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet users. J Med Internet Res 2016;18(7):e192 [FREE Full text] [doi: 10.2196/jmir.5413] [Medline: 27432783]
49. Watkins I, Xie B. eHealth literacy interventions for older adults: a systematic review of the literature. J Med Internet Res 2014;16(11):e225 [FREE Full text] [doi: 10.2196/jmir.3318] [Medline: 25386719]
50. Shi Y, Ma D, Zhang J, Chen B. In the digital age: a systematic literature review of the e-health literacy and influencing factors among Chinese older adults. Z Gesundh Wiss 2023;31(5):679-687 [FREE Full text] [doi: 10.1007/s10389-021-01604-z] [Medline: 34104627]
51. Chang SJ, Jang SJ, Lee H, Kim H. Building on evidence to improve eHealth literacy in older adults: a systematic review. Comput Inform Nurs 2021;39(5):241-247. [doi: 10.1097/CIN.0000000000000674] [Medline: 32925185]
52. Huang YQ, Liu L, Goodarzi Z, Watt JA. Diagnostic accuracy of eHealth literacy measurement tools in older adults: a systematic review. BMC Geriatr 2023;23(1):181 [FREE Full text] [doi: 10.1186/s12877-023-03899-x] [Medline: 36978033]

XSL•FO
RenderX

53. Arcury TA, Sandberg JC, Melius KP, Quandt SA, Leng X, Latulipe C, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020;39(2):141-150 [FREE Full text] [doi: 10.1177/0733464818807468] [Medline: 30353776]

Abbreviations

AVE: average variance extracted
C-eHEALS: Chinese version of the electronic eHealth Literacy Scale
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
DHLI: Digital Health Literacy Instrument
eHEALS: eHealth Literacy Scale
eHEALS-E: 20-item extended version of the eHEALS
eHLS-Web3.0: eHealth Literacy Scale-Web3.0
NCD: noncommunicable disease
RMSEA: root-mean-square error of approximation
SRMR: standardized root-mean-square residual
TLI: Tucker-Lewis index

Edited by T de Azevedo Cardoso; submitted 04.09.23; peer-reviewed by W Yu, S Jiang, B Shang; comments to author 21.11.23; revised version received 11.12.23; accepted 19.04.24; published 03.06.24.

Please cite as:
Cai W, Liang W, Liu H, Zhou R, Zhang J, Zhou L, Su N, Zhu H, Yang Y
Electronic Health Literacy Scale-Web3.0 for Older Adults with Noncommunicable Diseases: Validation Study
J Med Internet Res 2024;26:e52457
URL: https://www.jmir.org/2024/1/e52457
doi:10.2196/52457
PMID:38830207

©Wenfei Cai, Wei Liang, Huaxuan Liu, Rundong Zhou, Jie Zhang, Lin Zhou, Ning Su, Hanxiao Zhu, Yide Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

eHealth Literacy and Web-Based Health Information–Seeking Behaviors on COVID-19 in Japan: Internet-Based Mixed Methods Study

Seigo Mitsutake[1,2*], PT, PhD; Koichiro Oka[3], PhD; Orkan Okan[4], PhD; Kevin Dadaczynski[5,6], MSc, PhD; Tatsuro Ishizaki[7], MPH, MD, PhD; Takeo Nakayama[8], MD, PhD; Yoshimitsu Takahashi[8*], DrPH

[1]Human Care Research Team, Tokyo Metropolitan Institute for Geriatrics and Gerontology, Tokyo, Japan

[2]Australian Institute of Health Innovation, Macquarie University, NSW, Australia

[3]Faculty of Sport Sciences, Waseda University, Saitama, Japan

[4]WHO Collaborating Centre for Health Literacy, Department of Health and Sport Sciences, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

[5]Department of Health Science, Fulda University of Applied Science, Fulda, Germany

[6]Center for Applied Health Science, Leuphana University of Lueneburg, Lueneburg, Germany

[7]Human Care Research Team, Tokyo Metropolitan Institute for Geriatrics and Gerontology, Itabashi-ku, Japan

[8]Department of Health Informatics, Kyoto University School of Public Health, Kyoto, Japan

[*]these authors contributed equally

Corresponding Author:
Seigo Mitsutake, PT, PhD
Human Care Research Team
Tokyo Metropolitan Institute for Geriatrics and Gerontology
35-2 Sakae-cho, Itabahshi-ku
Tokyo, 173-0015
Japan
Phone: 81 3 3964 3241
Email: mitsu@tmig.or.jp

Abstract

Background: During the COVID-19 pandemic, much misinformation and disinformation emerged and spread rapidly via the internet, posing a severe public health challenge. While the need for eHealth literacy (eHL) has been emphasized, few studies have compared the difficulties involved in seeking and using COVID-19 information between adult internet users with low or high eHL.

Objective: This study examines the association between eHL and web-based health information–seeking behaviors among adult Japanese internet users. Moreover, this study qualitatively shed light on the difficulties encountered in seeking and using this information and examined its relationship with eHL.

Methods: This cross-sectional internet-based survey (October 2021) collected data from 6000 adult internet users who were equally divided into sample groups by gender, age, and income. We used the Japanese version of the eHL Scale (eHEALS). We also used a Digital Health Literacy Instrument (DHLI) adapted to the COVID-19 pandemic to assess eHL after we translated it to Japanese. Web-based health information–seeking behaviors were assessed by using a 10-item list of web sources and evaluating 10 topics participants searched for regarding COVID-19. Sociodemographic and other factors (eg, health-related behavior) were selected as covariates. Furthermore, we qualitatively explored the difficulties in information seeking and using. The descriptive contents of the responses regarding difficulties in seeking and using COVID-19 information were analyzed using an inductive qualitative content analysis approach.

Results: Participants with high eHEALS and DHLI scores on information searching, adding self-generated information, evaluating reliability, determining relevance, and operational skills were more likely to use all web sources of information about COVID-19 than those with low scores. However, there were negative associations between navigation skills and privacy protection scores when using several information sources, such as YouTube (Google LLC), to search for COVID-19 information. While half of the participants reported no difficulty seeking and using COVID-19 information, participants who reported any difficulties, including *information discernment*, *incomprehensible information*, *information overload*, and *disinformation*, had lower DHLI

XSL•FO
RenderX

score. Participants expressed significant concerns regarding "information quality and credibility," "abundance and shortage of relevant information," "public trust and skepticism," and "credibility of COVID-19–related information." Additionally, they disclosed more specific concerns, including "privacy and security concerns," "information retrieval challenges," "anxieties and panic," and "movement restriction."

Conclusions: Although Japanese internet users with higher eHEALS and total DHLI scores were more actively using various web sources for COVID-19 information, those with high navigation skills and privacy protection used web-based information about COVID-19 cautiously compared with those with lower proficiency. The study also highlighted an increased need for information discernment when using social networking sites in the "Health 2.0" era. The identified categories and themes from the qualitative content analysis, such as "information quality and credibility," suggest a framework for addressing the myriad challenges anticipated in future infodemics.

(J Med Internet Res 2024;26:e57842) doi:10.2196/57842

KEYWORDS

COVID-19; infectious; public health; SARS-COV-2; respiratory; eHealth; health communication; web-based information; DHLI; eHEALS; internet; mixed methods study; adult population; Asia; Asian; cross sectional; survey; surveys; questionnaire; questionnaires; Japan; Japanese; information seeking; information behavior; information behavior; health literacy; eHealth literacy; digital health literacy

Introduction

Background

The internet is a powerful source of information on health behavior, health, and medical care. Most of the general adult population uses the internet in Japan, as in other high-income countries [1-3]. Approximately 73% of Japanese internet users have searched for health information in the past 12 months [4]. However, many websites providing health information are unreliable and may be more linked to promoting commercial goods or private health services [5-7]. Misinformation (ie, false information distributed without the intention to cause harm) and disinformation (ie, false information shared deliberately to cause harm) may negatively affect people's physical and mental health, increase stigmatization, and threaten precious health gains, which lead to poor observance of public health measures [8,9]. Therefore, eHealth literacy (eHL), defined as the ability to seek, find, understand, and appraise health information on the internet to address or solve a health problem, is essential for accessing and using reliable health information via the internet.

During the COVID-19 pandemic, an "infodemic"—an epidemic of misinformation or disinformation—emerged and spread rapidly via the internet, posing a severe public health problem [10]. The COVID-19 infodemic has highlighted that poor eHL is a major challenge in using COVID-19 information on the internet [11]. People with poor health literacy are more likely to be confused by COVID-19 information on the internet [12]. Therefore, there is a need to improve health communication strategies for people with poor eHL to access reliable COVID-19 information on the internet easily.

Understanding the COVID-19 information–seeking behavior and identifying the difficulties internet users with low eHL are confronted with when dealing with this information are essential for improving communication strategies on COVID-19 and other health crises. The COVID-19 health literacy (COVID-HL) network surveyed digital health literacy (DHL), defined to have the same meaning as eHL [13]. Studies of the COVID-HL network revealed that university students with low eHL were more likely to use social media but less likely to use search engines and websites of official institutions than those with high eHL using quantitative data [14-18]. However, few studies have compared the difficulties individuals encounter when seeking and using COVID-19 information identified by qualitative content analysis between internet users with low and high eHL as estimated by an assessment tool. Mixed methods analyses, which integrate both quantitative and qualitative data, could increase our understanding of these difficulties and inform the development of strategies to enhance eHL for all individuals and improve the quality of web-based content. In addition, a limitation of these prior studies was that they included only college students [14-18] or physicians [19]. Examining the associations of eHL with health information–seeking behavior among other age groups is needed because the internet is used by not only younger adults but also different age groups, and older adults are reported to have barriers to seeking health information from the internet [1,2].

Objective

Comparing the subjective difficulties in seeking information between internet users with high and low eHL would help improve the strategies promoting access to reliable COVID-19 information. Therefore, this study aimed to examine the association between eHL and web-based health information–seeking behaviors using a mixed methods strategy. In addition, this study aimed to qualitatively shed light on the difficulties encountered in seeking and using this information, and to examine its relationship with eHL.

Methods

Study Design and Setting

This study used data from a cross-sectional internet-based survey that was conducted in Japan in October 2021. The study participants were recruited from the registrants of a Japanese internet research company (MyVoice Communication, Inc), who were asked to respond to the survey. This research company has approximately 553,719 registrants who could respond to

XSL•FO
RenderX

this survey and obtained detailed sociodemographic data from each participant upon registration in 2021.

Study Participants

This study aimed to collect data from 6000 men and women aged 20 to 79 years. The participants were equally divided into 132 sample groups categorized by gender (men and women), age (6 categories: 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79 years), and income (11 categories: <1, 1-<2, 2-<3, 3-<4, 4-<5, 5-<6, 6-<7, 7-<8, 8-<9, 9-<10, ≥10 million Yen [1 Yen=US $0.0088]; October 2021), with 45 participants in each group. The internet research service company randomly chose 250 potential respondents to include 45 participants in each group from the registered participants in accordance with the company's response rate data. Potential respondents could log into a protected site area using a unique ID and password. After the desired number of participants voluntarily signed a web-based informed consent form and completed a sociodemographic information form, further participants were no longer accepted.

Ethical Considerations

The Ethics Committees of the Tokyo Metropolitan Institute for Geriatrics and Gerontology (R21-055) and Kyoto University (R3191) approved the study protocol. All procedures followed the Ethical Guidelines of the Medical and Biological Research Involving Human Subjects established by the Japanese government. Data for analysis was provided by the research company after deidentification. Finally, we obtained informed consent from participants before the survey. Reward points valued at 130 Yen were provided as incentives for participation.

Measures

Exposure: eHL

The Japanese version of the eHL Scale (J-eHEALS) was used to assess eHL for using health information on the internet as a 1-way communication channel (Health 1.0) among participants [20-22]. We selected eHL Scale (eHEALS) because it is the most widely used DHL scale in the world and is easy to answer for participants [23]. The J-eHEALS used a 5-point Likert scale to measure perceived eHL (from 1 [strongly disagree] to 5 [strongly agree]; score range=8-40). To validate the J-eHEALS, a confirmatory factor analysis was conducted using data from the survey [20]. We divided the J-eHEALS scores into 2 categories (high and low) relative to the median score.

Moreover, the Digital Health Literacy Instrument (DHLI) adapted to the COVID-19 pandemic was also used to evaluate eHL levels, including literacy for using social networking sites (SNSs)—such as Facebook (Meta Platforms, Inc) and Twitter (Twitter, Inc)—referred to as "Health 2.0" [15]. The DHLI was designed to assess eHL for Health 1.0 and Health 2.0 and is widely used throughout the world [14-18,22]. We used J-eHEALS and DHLI to evaluate eHL levels for both Health 1.0 and Health 2.0. The DHLI contains 7 subscales: information searching, adding self-generated content, evaluating reliability, determining relevance, operational skills, navigation skills, and protecting privacy. Each subscale included 3 items to be answered on a 4-point Likert scale (ranging from 1 [very difficult] to 4 [very easy]). The COVID-HL network used DHLI adapted to COVID-19 and did not use the subscales of operational and navigation DHLI skills adapted to COVID-19 [13]. However, we included these subscales because they were crucial to accessing health information and navigating the internet. Moreover, although a recent study developed the DHLI [24], data on these skills adapted to COVID-19 among adult internet users in Japan were lacking. We divided each subscale and the total score of the DHLI into 1 of 2 categories (high or low) relative to the median score based on previous studies [14-18]. We translated DHLI adapted to the COVID-19 to Japanese, back translated it, and then confirmed their authors (Multimedia Appendix 1).

Outcomes: Web-Based Health Information–Seeking Behavior

The measures of web-based health information–seeking behaviors on the COVID-19 pandemic were assessed using a list of 10 different web sources: search engines (such as Google [Google LLC], Bing [Microsoft Corp], and Yahoo! [Yahoo Inc]), websites of public authorities (such as Ministry of Health, Labour and Welfare and the Japan Medical Association), Wikipedia, web-based encyclopedias, SNSs (such as Facebook, Instagram, and Twitter), YouTube, blogs providing medicine- and health-related information, medicine- and health-related question and answer sites (such as Yahoo! Answers), medicine- and health-related information portals, websites run by physicians or medical facilities, and news portal sites (including information gathered from newspapers and TV stations). These items were answered using a 5-point scale (0=do not know, 1=never, 2=often, 3=rarely, 4=sometimes, and 5=often). They were then assigned to either a "do not know–rarely" or "sometimes–often" category.

Moreover, we asked the participants to indicate from a list of 10 topics what they were searching for regarding COVID-19: the prevalence (such as number of people infected), infection route, symptoms, preventive measures (including disinfection and handwashing), rules and behavior (such as disinfection), assessment of its current status (such as declarations, measures, and stages), recommendations (including information from the Ministry of Health, Labour and Welfare and municipal governments), refraining from specific actions (such as eating out, traveling, and commuting to work), the economic and social effects, dealing with the psychological stress it causes, and information concerning the vaccine (effectiveness, side effects, and vaccination status). Participants answered "yes" or "no" to these items.

Sociodemographic and Other Variables

Sociodemographic and other variables were included as covariates in this regression model used by prior studies that examined the factors associated with eHL (gender, age groups, equivalent income, education status, marital status, cigarette smoking, alcohol consumption, physical exercise habits, and conditions that could likely lead to severe COVID-19 illness) [20-22]. Equivalent income was estimated by dividing annual income by the square root of the number of families [25]. We divided the equivalent income into 12 categories (<1-≥10 million Yen and "not answered"). Education status was divided into 4

XSL•FO
RenderX

categories (≤high school graduate, 2-year college or career college, higher university education, and "not answered"). Regarding marital status, the participants who answered "married" were categorized as "married." The participants who answered "never married," "widowed," or "divorced" were categorized as "not married." Concerning health behaviors, we assessed 3 items related to smoking, alcohol consumption, and physical exercise. Regarding smoking status, responses such as "never" or "quit" were categorized as "no smoking" and "smoking" or "sometimes smoking" as "smoking." Alcohol consumption was determined using "yes" or "no" responses and the quantity of alcohol consumed. The participants who answered "no" or "quit" were categorized as "no." Participants who responded with an alcohol intake of <20 g at once were categorized as "<20 g/once," and those who drank alcohol ≥20 g at once were categorized as "≥20 g/once." The physical exercise of participants was assessed subjectively based on whether they performed a 30-minute physical exercise ≥2 times a week for a year or longer ("yes" or "no"). We selected 6 conditions (hypertension, diabetes, chronic obstructive pulmonary diseases, heart diseases, and chronic kidney diseases; BMI ≥30) to determine the possibility of becoming severely ill with COVID-19. "Yes" responses to ≥1 questions concerning the prevalence of conditions that were likely to cause severe illness with COVID-19 were categorized as "Yes."

Difficulties in Seeking and Using COVID-19 Information

We asked the participants the descriptive open-ended question, "What difficulties did you have in seeking and using COVID-19-related information on the internet?" The item was in the required field and thus could not be left unanswered.

Analysis Using a Mixed Methods Strategy

Overview

We used the concurrent triangulation design of mixed methods strategy to analyze both quantitative and qualitative data in the internet-based survey [26]. In mixed methods analyses, the use of complementary methods integrating quantitative and qualitative approaches to address a complex question can generate deeper insights than using either approach alone or both approaches separately [27]. Mixed methods research enables a more comprehensive understanding of the phenomenon under investigation by integrating both quantitative and qualitative data. Furthermore, findings can be validated across different data sets by using both quantitative and qualitative methods. The triangulation of data from multiple methods enhances the credibility and reliability of a study's findings. By adopting a mixed methods approach, it is possible to attain a broader understanding of the association between eHL and web-based health information–seeking behaviors, as well as the difficulties encountered in seeking and using this information.

Qualitative Content Analysis of Qualitative Data

Descriptive responses regarding difficulties in seeking and using COVID-19 information were analyzed using the inductive qualitative content analysis approach [28-30]. The contents were inductively organized into codes and categories to achieve trustworthiness [29]. YT and SM performed the analysis. All the responses were read and interpreted repeatedly. After discussing the meanings of the responses, phrases or sentences were coded for the analysis. The coding frame was changed when new codes emerged, and sentences were reread using the new structure. This constant comparison process was also used to develop conceptualize the responses into broad categories after further discussion. We finally aggregated categories into themes. We used the MAXQDA Analytics Pro 2022 (version 22.4.1; VERBI Software GmbH) for qualitative content analysis.

Statistical Analysis of Quantitative Data

First, a chi-square test was performed to compare the proportion of participants with low and high eHL by assessing the eHEALS and subscales of the DHLI. The internal consistencies of the subscales and the total scale were assessed using Cronbach α. We then examined the association of eHL levels with using web sources of COVID-19 information by using a multivariable logistic regression model that adjusted for all covariates. In addition, the associations of eHL levels with searching for specific COVID-19 topics were examined using a multivariable logistic regression model that adjusted for all covariates. Adjusted odds ratios (AORs) and 95% CIs were estimated. We explored the relationship between eHL and categories of difficulties more thoroughly. eHEALS and DHLI total scores were classified into quartiles to observe variations in dose response, followed by the performance of the Cochrane-Armitage test for trend analysis. Two-tailed *P* values <.05 were considered significant. All analyses were conducted using SPSS (version 28.0; IBM Corp).

Results

Study Participant Selection

Figure 1 illustrates this study's participant selection process. The research company chose 18,493 potential respondents in October 2021, and 6000 responses were obtained from respondents who provided complete information for the study variables (response rate: 32.4%).

XSL•FO
RenderX

Figure 1. Flowchart of patient selection in this study.

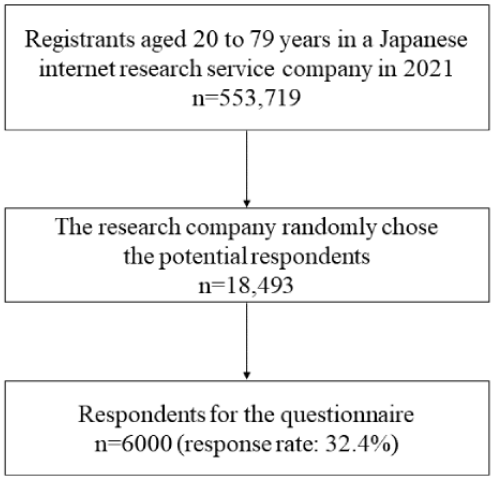

Characteristics of Study Participants

The proportions of each gender and age group were identical (Table 1). The proportion of participants whose equivalent income was 3 to 4 million Yen was 19.98% (1199/6000). About 48.45% (2907/6000) of the participants had graduated from university or had higher education, and 54.67% (3280/6000) were married. Approximately 16.13% (968/6000) of the participants reported a cigarette smoking habit, 61.45% (3687/6000) consumed alcohol, and 32.78% (1967/6000) exercised regularly. Moreover, 23.02% (1381/6000) of the respondents had ≥1 health conditions likely to lead to severe COVID-19 illness.

XSL•FO
RenderX

Table 1. Characteristics of participants.

Characteristics	Participants (N=6000), n (%)
Gender	
Men	3000 (50)
Women	3000 (50)
Age groups (y)	
20-29	1000 (16.67)
30-39	1000 (16.67)
40-49	1000 (16.67)
50-59	1000 (16.67)
60-69	1000 (16.67)
≥70	1000 (16.67)
Equivalent income (million Y en ; Y en 1=US $ 0.0088)	
<1	413 (6.88)
1-<2	878 (14.63)
2-<3	972 (16.2)
3-<4	1199 (19.98)
4-<5	852 (14.2)
5-<6	486 (8.1)
6-<7	430 (7.17)
7-<8	165 (2.75)
8-<9	70 (1.17)
9-<10	72 (1.2)
≥10	107 (1.78)
Not answered	356 (5.93)
Education status	
≤High school	1768 (29.47)
2-year college or career college	1298 (21.63)
University or higher education	2907 (48.45)
No answer	27 (0.45)
Marital status	
No	2683 (44.72)
Yes	3280 (54.67)
Not answered	37 (0.62)
Cigarette smoking	
No	5032 (83.87)
Yes	968 (16.13)
Alcohol consumption	
No	2313 (38.55)
<20 g/once	2053 (34.22)
≥20 g/once	1634 (27.23)
Physical exercise habit	
No	4033 (67.22)
Yes	1967 (32.78)

XSL•FO
RenderX

Characteristics	Participants (N=6000), n (%)
Conditions that could likely lead to severe COVID-19 illness	
No	4619 (76.98)
Yes	1381 (23.02)

Scores and Internal Consistencies of the DHLI Among This Study's Participants

Table 2 presents the DHLI scores and internal consistencies among the study participants. The 7 subscales' internal consistencies (Cronbach α) ranged from acceptable to good (0.83-0.94). Moreover, the mean DHLI total score was 3.08 (SD 0.49), and Cronbach α of the complete scale was 0.92.

Table 2. Scores and internal consistencies of the DHLI[a,b].

Subscales and total score of the DHLI	Values, mean (SD)	Values, median (IQR)	Cronbach α
Information search	3.01 (0.61)	3.0 (2.7-3.3)	0.91
Adding self-generated information	2.73 (0.73)	3.0 (2.0-3.0)	0.94
Evaluating reliability	2.66 (0.65)	2.7 (2.0-3.0)	0.88
Determining relevance	2.87 (0.59)	3.0 (2.7-3.0)	0.90
Operational skills	3.31 (0.62)	3.0 (3.0-4.0)	0.88
Navigation skills	3.59 (0.73)	4.0 (3.3-4.0)	0.83
Privacy protection	3.42 (0.91)	4.0 (3.0-4.0)	0.87
Total score	3.08 (0.49)	3.1 (2.8-3.4)	0.92

[a]DHLI: Digital Health Literacy Instrument.

[b]The subscale scores and total DHLI score range from 0 to 4.

Differences of Characteristics by eHL From eHEALS and DHLI Subscales

Compared to those with low eHEALS, participants with high eHEALS were more likely to be older (P<.001), have higher income (P<.001) and education levels (P<.001), and be married (P=.007; Tables 3 and 4). Moreover, they were more likely to consume alcohol (P=.02) and have physical exercise habits (P<.001) and conditions leading to severe COVID-19 illness (P=.004) than those with low eHEALS.

Among the participants with high total DHLI scores, there were higher proportions of men (P=.002), those aged 20 to 39 years (P<.001), those with higher equivalent income (P<.001), and those with higher education status (P<.001). They were more likely to consume alcohol (P=.02) and have physical exercise habits (P<.001) and less likely to have conditions that could lead to severe COVID-19 illness (P<.001). In addition, participants with higher subscores of DHLI generally consisted of higher proportions of men, had higher equivalent income and higher education status, and were more likely to be married. They were more likely to consume alcohol and less likely to have conditions that could lead to severe COVID-19 illness. Moreover, participants with higher scores on information searching, adding self-generated content, evaluating reliability, determining relevance, and operational skills were more likely to have an exercise habit. However, participants with higher scores on navigation skills and privacy protection were less likely to have an exercise habit.

XSL•FO
RenderX

Table 3. Differences of characteristics based on the scores of eHEALS[a], DHLI[b], and its subscales.

Characteristic	eHEALS			Information searching			Adding self-generated information			Evaluation reliability		
	Low (n=2228), n (%)	High (n=3772), n (%)	*P* value[c]	Low (n=1613), n (%)	High (n=4387), n (%)	*P* value[c]	Low (n=2702), n (%)	High (n=3298), n (%)	*P* value[c]	Low (n=2475), n (%)	High (n=3525), n (%)	*P* value[c]
Gender			.24			.47			.04			<.001
Men	1136 (50.99)	1864 (49.42)		794 (49.23)	2206 (50.28)		1311 (48.52)	1689 (51.21)		1135 (45.86)	1865 (52.91)	
Women	1092 (49.01)	1908 (50.58)		819 (50.77)	2181 (49.72)		1391 (51.48)	1609 (48.79)		1340 (54.14)	1660 (47.09)	
Age groups (y)			<.001			.83			.17			<.001
20-29	421 (18.9)	579 (15.35)		272 (16.86)	728 (16.59)		420 (15.54)	580 (17.59)		362 (14.63)	638 (18.1)	
30-39	376 (16.88)	624 (16.54)		268 (16.62)	732 (16.69)		453 (16.77)	547 (16.59)		389 (15.72)	611 (17.33)	
40-49	374 (16.79)	626 (16.6)		271 (16.8)	729 (16.62)		480 (17.76)	520 (15.77)		405 (16.36)	595 (16.88)	
50-59	387 (17.37)	613 (16.25)		256 (15.87)	744 (16.96)		442 (16.36)	558 (16.92)		424 (17.13)	576 (16.34)	
60-69	355 (15.93)	645 (17.1)		263 (16.31)	737 (16.8)		456 (16.88)	544 (16.49)		422 (17.05)	578 (16.4)	
≥70	315 (14.14)	685 (18.16)		283 (17.54)	717 (16.34)		451 (16.69)	549 (16.65)		473 (19.11)	527 (14.95)	
Equivalent income (million Yen; Yen 1=US $ 0.0088)			<.001			<.001			<.001			<.001
<1	177 (7.94)	236 (6.26)		131 (8.12)	282 (6.43)		202 (7.48)	211 (6.4)		182 (7.35)	231 (6.55)	
1-<2	375 (16.83)	503 (13.34)		279 (17.3)	599 (13.65)		468 (17.32)	410 (12.43)		433 (17.49)	445 (12.62)	
2-<3	394 (17.68)	578 (15.32)		270 (16.74)	702 (16)		468 (17.32)	504 (15.28)		439 (17.74)	533 (15.12)	
3-<4	437 (19.61)	762 (20.2)		320 (19.84)	879 (20.04)		537 (19.87)	662 (20.07)		491 (19.84)	708 (20.09)	
4-<5	300 (13.46)	552 (14.63)		208 (12.9)	644 (14.68)		366 (13.55)	486 (14.74)		336 (13.58)	516 (14.64)	
5-<6	166 (7.45)	320 (8.48)		110 (6.82)	376 (8.57)		189 (6.99)	297 (9.01)		177 (7.15)	309 (8.77)	
6-<7	130 (5.83)	300 (7.95)		93 (5.77)	337 (7.68)		141 (5.22)	289 (8.76)		145 (5.86)	285 (8.09)	
7-<8	45 (2.02)	120 (3.18)		33 (2.05)	132 (3.01)		58 (2.15)	107 (3.24)		48 (1.94)	117 (3.32)	
8-<9	24 (1.08)	46 (1.22)		15 (0.93)	55 (1.25)		28 (1.04)	42 (1.27)		18 (0.73)	52 (1.48)	
9-<10	24 (1.08)	48 (1.27)		14 (0.87)	58 (1.32)		16 (0.59)	56 (1.7)		14 (0.57)	58 (1.65)	
≥10	17 (0.76)	90 (2.39)		12 (0.74)	95 (2.17)		27 (1)	80 (2.43)		18 (0.73)	89 (2.52)	
Not answered	139 (6.24)	217 (5.75)		128 (7.94)	228 (5.2)		202 (7.48)	154 (4.67)		174 (7.03)	182 (5.16)	
Education status			<.001			.001			<.001			<.001
≤High school	725 (32.54)	1043 (27.65)		528 (32.73)	1240 (28.27)		879 (32.53)	889 (26.96)		852 (34.42)	916 (25.99)	

XSL•FO
RenderX

Characteristic	eHEALS			Information searching			Adding self-generated information			Evaluation reliability		
	Low (n=2228), n (%)	High (n=3772), n (%)	*P* value[c]	Low (n=1613), n (%)	High (n=4387), n (%)	*P* value[c]	Low (n=2702), n (%)	High (n=3298), n (%)	*P* value[c]	Low (n=2475), n (%)	High (n=3525), n (%)	*P* value[c]
2-year college or career college	443 (19.88)	855 (22.67)		321 (19.9)	977 (22.27)		590 (21.84)	708 (21.47)		521 (21.05)	777 (22.04)	
≥University	1053 (47.26)	1854 (49.15)		752 (46.62)	2155 (49.12)		1216 (45)	1691 (51.27)		1087 (43.92)	1820 (51.63)	
Not answered	7 (0.31)	20 (0.53)		12 (0.74)	15 (0.34)		17 (0.63)	10 (0.3)		15 (0.61)	12 (0.34)	
Marital status			.007			.002			.03			.14
No	1049 (47.08)	1634 (43.32)		774 (47.99)	1909 (43.51)		1253 (46.37)	1430 (43.36)		1094 (44.2)	1589 (45.08)	
Yes	1170 (52.51)	2110 (55.94)		825 (51.15)	2455 (55.96)		1429 (52.89)	1851 (56.12)		1360 (54.95)	1920 (54.47)	
Not answered	9 (0.4)	28 (0.74)		14 (0.87)	23 (0.52)		20 (0.74)	17 (0.52)		21 (0.85)	16 (0.45)	
Cigarette smoking			.63			.13			<.001			<.001
No	1862 (83.57)	3170 (84.04)		1372 (85.06)	3660 (83.43)		2329 (86.2)	2703 (81.96)		2129 (86.02)	2903 (82.35)	
Yes	366 (16.43)	602 (15.96)		241 (14.94)	727 (16.57)		373 (13.8)	595 (18.04)		346 (13.98)	622 (17.65)	
Alcohol consumption			.02			<.001			<.001			<.001
No	911 (40.89)	1402 (37.17)		680 (42.16)	1633 (37.22)		1117 (41.34)	1196 (36.26)		1031 (41.66)	1282 (36.37)	
<20 g/once	731 (32.81)	1322 (35.05)		555 (34.41)	1498 (34.15)		935 (34.6)	1118 (33.9)		856 (34.59)	1197 (33.96)	
≥20 g/once	586 (26.3)	1048 (27.78)		378 (23.43)	1256 (28.63)		650 (24.06)	984 (29.84)		588 (23.76)	1046 (29.67)	
Physical exercise habit			<.001			<.001			<.001			<.001
No	1683 (75.54)	2350 (62.3)		1157 (71.73)	2876 (65.56)		1957 (72.43)	2076 (62.95)		1781 (71.96)	2252 (63.89)	
Yes	545 (24.46)	1422 (37.7)		456 (28.27)	1511 (34.44)		745 (27.57)	1222 (37.05)		694 (28.04)	1273 (36.11)	
Conditions leading to severe COVID-19 illness			.004			.004			.24			.16
No	1745 (78.32)	2874 (76.19)		1200 (74.4)	3419 (77.93)		2061 (76.28)	2558 (77.56)		1883 (76.08)	2736 (77.62)	
Yes	483 (21.68)	898 (23.81)		413 (25.6)	968 (22.07)		641 (23.72)	740 (22.44)		592 (23.92)	789 (22.38)	

[a]eHEALS: eHealth Literacy Scale.

[b]DHLI: Digital Health Literacy Instrument.

[c]The chi-square test.

XSL•FO
RenderX

Table 4. Differences of characteristics based on the scores of eHEALS[a], DHLI[b], and its subscales (continued).

Characteristic	Determining relevance			Operational skills			Navigation skills			Privacy protection			Total score		
	Low (n=2051), n (%)	High (n=3949), n (%)	*P* value[c]	Low (n=1076), n (%)	High (n=4924), n (%)	*P* value[c]	Low (n=1905), n (%)	High (n=4095), n (%)	*P* value[c]	Low (n=2206), n (%)	High (n=3794), n (%)	*P* value[c]	Low (n=3264), n (%)	High (n=2736), n (%)	*P* value[c]
Gender			.17			<.001			.002			.75			.002
Men	1051 (51.24)	1949 (49.35)		455 (42.29)	2545 (51.69)		896 (47.03)	2104 (51.38)		1109 (50.27)	1891 (49.84)		1572 (48.16)	1428 (52.19)	
Women	1000 (48.76)	2000 (50.65)		621 (57.71)	2379 (48.31)		1009 (52.97)	1991 (48.62)		1097 (49.73)	1903 (50.16)		1692 (51.84)	1308 (47.81)	
Age groups (y)			.73			<.001			<.001			<.001			<.001
20-29	339 (16.53)	661 (16.74)		159 (14.78)	841 (17.08)		313 (16.43)	687 (16.78)		378 (17.14)	622 (16.39)		500 (15.32)	500 (18.27)	
30-39	334 (16.28)	666 (16.87)		163 (15.15)	837 (17)		307 (16.12)	693 (16.92)		349 (15.82)	651 (17.16)		514 (15.75)	486 (17.76)	
40-49	359 (17.5)	641 (16.23)		147 (13.66)	853 (17.32)		269 (14.12)	731 (17.85)		323 (14.64)	677 (17.84)		527 (16.15)	473 (17.29)	
50-59	330 (16.09)	670 (16.97)		168 (15.61)	832 (16.9)		298 (15.64)	702 (17.14)		340 (15.41)	660 (17.4)		535 (16.39)	465 (17)	
60-69	337 (16.43)	663 (16.79)		204 (18.96)	796 (16.17)		303 (15.91)	697 (17.02)		356 (16.14)	644 (16.97)		559 (17.13)	441 (16.12)	
≥70	352 (17.16)	648 (16.41)		235 (21.84)	765 (15.54)		415 (21.78)	585 (14.29)		460 (20.85)	540 (14.23)		629 (19.27)	371 (13.56)	
Equivalent income (million Yen; Yen 1=US $0.0088)			<.001			<.001			.005			.01			<.001
<1	159 (7.75)	254 (6.43)		98 (9.11)	315 (6.4)		149 (7.82)	264 (6.45)		155 (7.03)	258 (6.8)		246 (7.54)	167 (6.1)	
1-<2	340 (16.58)	538 (13.62)		208 (19.33)	670 (13.61)		296 (15.54)	582 (14.21)		339 (15.37)	539 (14.21)		542 (16.61)	336 (12.28)	
2-<3	340 (16.58)	632 (16)		187 (17.38)	785 (15.94)		312 (16.38)	660 (16.12)		382 (17.32)	590 (15.55)		559 (17.13)	413 (15.1)	
3-<4	418 (20.38)	781 (19.78)		201 (18.68)	998 (20.27)		393 (20.63)	806 (19.68)		470 (21.31)	729 (19.21)		660 (20.22)	539 (19.7)	
4-<5	293 (14.29)	559 (14.16)		136 (12.64)	716 (14.54)		280 (14.7)	572 (13.97)		316 (14.32)	536 (14.13)		458 (14.03)	394 (14.4)	
5-<6	137 (6.68)	349 (8.84)		70 (6.51)	416 (8.45)		141 (7.4)	345 (8.42)		154 (6.98)	332 (8.75)		224 (6.86)	262 (9.58)	
6-<7	127 (6.19)	303 (7.67)		47 (4.37)	383 (7.78)		117 (6.14)	313 (7.64)		141 (6.39)	289 (7.62)		198 (6.07)	232 (8.48)	
7-<8	42 (2.05)	123 (3.11)		16 (1.49)	149 (3.03)		39 (2.05)	126 (3.08)		50 (2.27)	115 (3.03)		65 (1.99)	100 (3.65)	
8-<9	18 (0.88)	52 (1.32)		8 (0.74)	62 (1.26)		22 (1.15)	48 (1.17)		26 (1.18)	44 (1.16)		35 (1.07)	35 (1.28)	
9-<10	8 (0.39)	64 (1.62)		4 (0.37)	68 (1.38)		15 (0.79)	57 (1.39)		21 (0.95)	51 (1.34)		25 (0.77)	47 (1.72)	
≥10	20 (0.98)	87 (2.2)		9 (0.84)	98 (1.99)		23 (1.21)	84 (2.05)		38 (1.72)	69 (1.82)		30 (0.92)	77 (2.81)	
Not answered	149 (7.26)	207 (5.24)		92 (8.55)	264 (5.36)		118 (6.19)	238 (5.81)		114 (5.17)	242 (6.38)		222 (6.8)	134 (4.9)	
Education status			.007			<.001			.002			.06			<.001

XSL•FO
RenderX

Characteristic	Determining relevance			Operational skills			Navigation skills			Privacy protection			Total score		
	Low (n=2051), n (%)	High (n=3949), n (%)	*P* value[c]	Low (n=1076), n (%)	High (n=4924), n (%)	*P* value[c]	Low (n=1905), n (%)	High (n=4095), n (%)	*P* value[c]	Low (n=2206), n (%)	High (n=3794), n (%)	*P* value[c]	Low (n=3264), n (%)	High (n=2736), n (%)	*P* value[c]
≤High school	655 (31.94)	1113 (28.18)		451 (41.91)	1317 (26.75)		619 (32.49)	1149 (28.06)		618 (28.01)	1150 (30.31)		1053 (32.26)	715 (26.13)	
2-year college or career college	422 (20.58)	876 (22.18)		248 (23.05)	1050 (21.32)		410 (21.52)	888 (21.68)		460 (20.85)	838 (22.09)		707 (21.66)	591 (21.6)	
≥University	961 (46.86)	1946 (49.28)		364 (33.83)	2543 (51.65)		871 (45.72)	2036 (49.72)		1116 (50.59)	1791 (47.21)		1487 (45.56)	1420 (51.9)	
Not answered	13 (0.63)	14 (0.35)		13 (1.21)	14 (0.28)		5 (0.26)	22 (0.54)		12 (0.54)	15 (0.4)		17 (0.52)	10 (0.37)	
Marital status			.04			<.001			.99			.58			.44
No	913 (44.51)	1770 (44.82)		454 (42.19)	2229 (45.27)		850 (44.62)	1833 (44.76)		967 (43.83)	1716 (45.23)		1460 (44.73)	1223 (44.7)	
Yes	1118 (54.51)	2162 (54.75)		600 (55.76)	2680 (54.43)		1043 (54.75)	2237 (54.63)		1225 (55.53)	2055 (54.16)		1780 (54.53)	1500 (54.82)	
Not answered	20 (0.98)	17 (0.43)		22 (2.04)	15 (0.3)		12 (0.63)	25 (0.61)		14 (0.63)	23 (0.61)		24 (0.74)	13 (0.48)	
Cigarette smoking			.61			.18			.06			.56			.15
No	1727 (84.2)	3305 (83.69)		917 (85.22)	4115 (83.57)		1573 (82.57)	3459 (84.47)		1842 (83.5)	3190 (84.08)		2758 (84.5)	2274 (83.11)	
Yes	324 (15.8)	644 (16.31)		159 (14.78)	809 (16.43)		332 (17.43)	636 (15.53)		364 (16.5)	604 (15.92)		506 (15.5)	462 (16.89)	
Alcohol consumption			.008			<.001			.05			.02			.008
No	844 (41.15)	1469 (37.2)		507 (47.12)	1806 (36.68)		721 (37.85)	1592 (38.88)		801 (36.31)	1512 (39.85)		1303 (39.92)	1010 (36.92)	
<20 g/once	685 (33.4)	1368 (34.64)		321 (29.83)	1732 (35.17)		627 (32.91)	1426 (34.82)		785 (35.58)	1268 (33.42)		1122 (34.38)	931 (34.03)	
≥20 g/once	522 (25.45)	1112 (28.16)		248 (23.05)	1386 (28.15)		557 (29.24)	1077 (26.3)		620 (28.11)	1014 (26.73)		839 (25.7)	795 (29.06)	
Physical exercise habit			<.001			.02			.01			<.001			<.001
No	1494 (72.84)	2539 (64.29)		755 (70.17)	3278 (66.57)		1238 (64.99)	2795 (68.25)		1413 (64.05)	2620 (69.06)		2276 (69.73)	1757 (64.22)	
Yes	557 (27.16)	1410 (35.71)		321 (29.83)	1646 (33.43)		667 (35.01)	1300 (31.75)		793 (35.95)	1174 (30.94)		988 (30.27)	979 (35.78)	
Conditions that could lead to severe COVID-19 illness			.02			.01			<.001			<.001			<.001
No	1544 (75.28)	3075 (77.87)		797 (74.07)	3822 (77.62)		1397 (73.33)	3222 (78.68)		1603 (72.67)	3016 (79.49)		2424 (74.26)	2195 (80.23)	

XSL•FO
RenderX

Characteristic	Determining relevance			Operational skills			Navigation skills			Privacy protection			Total score		
	Low (n=2051), n (%)	High (n=3949), n (%)	*P* value[c]	Low (n=1076), n (%)	High (n=4924), n (%)	*P* value[c]	Low (n=1905), n (%)	High (n=4095), n (%)	*P* value[c]	Low (n=2206), n (%)	High (n=3794), n (%)	*P* value[c]	Low (n=3264), n (%)	High (n=2736), n (%)	*P* value[c]
Yes	507 (24.72)	874 (22.13)		279 (25.93)	1102 (22.38)		508 (26.67)	873 (21.32)		603 (27.33)	778 (20.51)		840 (25.74)	541 (19.77)	

[a]eHEALS: eHealth Literacy Scale.

[b]DHLI: Digital Health Literacy Instrument.

[c]The chi-square test.

Associations of eHL With Using Web Sources for Finding COVID-19 Information

Table 5 illustrates the proportion of “sometimes” or “often” responses to questions on using each web source. The most common web sources were search engines (4614/6000, 76.9%), followed by news portal sites (3350/6000, 55.83%). Participants with high eHEALS were more likely to use all web sources of information about COVID-19 than those with low eHEALS (Tables 6 and 7). The participants with high scores on the DHLI subscales information searching, adding self-generated information, evaluating reliability, determining relevance, and operational skills were also more likely to search for COVID-19 information using all web sources than participants with low scores on these subscales. Participants with high navigation skill scores were more likely to use search engines but less likely to use YouTube to search for COVID-19 information (AOR 0.88, 95% CI 0.79-0.99). Moreover, participants with high privacy protection scores were less likely to use websites of public authorities (AOR 0.80, 95% CI 0.72-0.89), Wikipedia (AOR 0.82, 95% CI 0.74-0.92), SNSs (AOR 0.74, 95% CI 0.66-0.83), YouTube (AOR 0.84, 95% CI 0.75-0.94), blogs providing medicine- and health-related information (AOR 0.81, 95% CI 0.72-0.92), question and answer sites (AOR 0.75, 95% CI 0.67-0.85), medicine- and health-related information portals (AOR 0.75, 95% CI 0.67-0.85), and websites run by physicians or medical facilities (AOR 0.72, 95% CI 0.64-0.81) for finding COVID-19 information. In addition, participants with high total DHLI scores were more likely to use all web sources of COVID-19 information than those with low total scores.

Table 5. The proportion of “sometimes” or “often” responses to questions on using each web source.

Web source	Participants (N=6000), n (%)
Search engines	4614 (76.9)
News portal sites	3350 (55.83)
Websites of public authorities	2652 (44.2)
Wikipedia, web-based encyclopedias	2232 (37.2)
Social media sites	2088 (34.8)
YouTube	2080 (34.67)
Websites run by physicians or medical facilities	1718 (28.63)
Question and answer sites related to medicine and health	1717 (28.62)
Medicine and health-related information portals	1623 (27.05)
Blogs providing medicine and health-related information	1379 (22.98)

XSL•FO
RenderX

Table 6. Associations of eHL[a] levels with using web sources for finding COVID-19 information.

eHL	Search engines		Websites of public authorities		Wikipedia and web-based encyclopedias		SNSs[b]		YouTube	
	Value, n (%)	AOR[c] (95% CI)[d]	Value, n (%)	AOR (95% CI)[d]	Value, n (%)	AOR (95% CI)[d]	Value, n (%)	AOR (95% CI)[d]	Value, n (%)	AOR (95% CI)[d]
eHEALS[e]										
Low (n=2228)	1583 (71.05)	1.00 (reference)	748 (33.57)	1.00 (reference)	641 (28.77)	1.00 (reference)	640 (28.73)	1.00 (reference)	638 (28.64)	1.00 (reference)
High (n=3772)	3031 (80.36)	1.57 (1.38-1.78)	1904 (50.48)	1.88 (1.68-2.11)	1591 (42.18)	1.70 (1.51-1.90)	1448 (38.39)	1.61 (1.43-1.81)	1442 (38.23)	1.50 (1.34-1.69)
Information search										
Low (n=1613)	1133 (70.24)	1.00 (reference)	591 (36.64)	1.00 (reference)	454 (28.15)	1.00 (reference)	466 (28.89)	1.00 (reference)	452 (28.02)	1.00 (reference)
High (n=4387)	3481 (79.35)	1.54 (1.35-1.76)	2061 (46.98)	1.44 (1.27-1.62)	1778 (40.53)	1.66 (1.46-1.89)	1622 (36.97)	1.41 (1.24-1.60)	1628 (37.11)	1.47 (1.30-1.67)
Adding self-generated information										
Low (n=2702)	1992 (73.72)	1.00 (reference)	1034 (38.27)	1.00 (reference)	813 (30.09)	1.00 (reference)	786 (29.09)	1.00 (reference)	771 (28.53)	1.00 (reference)
High (n=3298)	2622 (79.5)	1.29 (1.14-1.46)	1618 (49.06)	1.42 (1.27-1.58)	1419 (43.03)	1.64 (1.47-1.84)	1302 (39.48)	1.51 (1.35-1.70)	1309 (39.69)	1.58 (1.41-1.76)
Evaluating reliability										
Low (n=2475)	1849 (74.71)	1.00 (reference)	934 (37.74)	1.00 (reference)	747 (30.18)	1.00 (reference)	722 (29.17)	1.00 (reference)	718 (29.01)	1.00 (reference)
High (n=3525)	2765 (78.44)	1.16 (1.03-1.32)	1718 (48.74)	1.45 (1.30-1.61)	1485 (42.13)	1.56 (1.39-1.74)	1366 (38.75)	1.39 (1.24-1.56)	1362 (38.64)	1.44 (1.29-1.62)
Determining relevance										
Low (n=2051)	1447 (70.55)	1.00 (reference)	760 (37.06)	1.00 (reference)	617 (30.08)	1.00 (reference)	603 (29.4)	1.00 (reference)	599 (29.21)	1.00 (reference)
High (n=3949)	3167 (80.2)	1.60 (1.41-1.82)	1892 (47.91)	1.46 (1.30-1.63)	1615 (40.9)	1.53 (1.37-1.72)	1485 (37.6)	1.39 (1.23-1.57)	1481 (37.5)	1.40 (1.25-1.58)
Operational skills										
Low (n=1076)	658 (61.15)	1.00 (reference)	310 (28.81)	1.00 (reference)	263 (24.44)	1.00 (reference)	260 (24.16)	1.00 (reference)	322 (29.93)	1.00 (reference)
High (n=4924)	3956 (80.34)	2.43 (2.09-2.82)	2342 (47.56)	2.07 (1.78-2.40)	1969 (39.99)	1.90 (1.62-2.21)	1828 (37.12)	1.67 (1.42-1.96)	1758 (35.7)	1.22 (1.05-1.41)
Navigation skills										
Low (n=1905)	1436 (75.38)	1.00 (reference)	845 (44.36)	1.00 (reference)	692 (36.33)	1.00 (reference)	643 (33.75)	1.00 (reference)	698 (36.64)	1.00 (reference)
High (n=4095)	3178 (77.61)	1.15 (1.01-1.31)	1807 (44.13)	1.00 (0.89-1.12)	1540 (37.61)	1.05 (0.94-1.18)	1445 (35.29)	1.06 (0.94-1.20)	1382 (33.75)	0.88 (0.79-0.99)
Protecting privacy										
Low (n=2206)	1706 (77.33)	1.00 (reference)	1061 (48.1)	1.00 (reference)	891 (40.39)	1.00 (reference)	856 (38.8)	1.00 (reference)	826 (37.44)	1.00 (reference)
High (n=3794)	2908 (76.65)	1.01 (0.88-1.14)	1591 (41.93)	0.80 (0.72-0.89)	1341 (35.35)	0.82 (0.74-0.92)	1232 (32.47)	0.74 (0.66-0.83)	1254 (33.05)	0.84 (0.75-0.94)
Total score of DHLI[f]										
Low (n=3264)	2426 (74.33)	1.00 (reference)	1302 (39.89)	1.00 (reference)	1065 (32.63)	1.00 (reference)	997 (30.55)	1.00 (reference)	1040 (31.86)	1.00 (reference)
High (n=2736)	2188 (79.97)	1.34 (1.18-1.52)	1350 (49.34)	1.38 (1.24-1.54)	1167 (42.65)	1.46 (1.31-1.63)	1091 (39.88)	1.39 (1.24-1.56)	1040 (38.01)	1.25 (1.12-1.40)

XSL•FO
RenderX

[a]eHL: eHealth literacy.

[b]SNS: social networking site.

[c]AOR: adjusted odds ratio.

[d]Multivariable logistic regression analysis adjusted for all covariates (ie, gender, age groups, equivalent income, education status, marital status, cigarette smoking, alcohol consumption, physical exercise habit, and conditions leading to severe illness due to COVID-19).

[e]eHEALS: eHealth Literacy Scale.

[f]DHLI: Digital Health Literacy Instrument.

XSL•FO
RenderX

Table 7. Associations of eHL[a] levels with using web sources for finding COVID-19 information (continued).

eHL	Blogs providing medicine- and health-related information		Medicine- and health-related question and answer sites		Medicine- and health-related information portals		Websites run by physicians or medical facilities		News portal sites	
	Value, n (%)	AOR[b] (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]
eHEALS[d]										
Low (n=2228)	292 (13.1)	1.00 (reference)	417 (18.72)	1.00 (reference)	336 (15.08)	1.00 (reference)	375 (16.83)	1.00 (reference)	1045 (46.9)	1.00 (reference)
High (n=3772)	1087 (28.82)	2.46 (2.13-2.84)	1300 (34.46)	2.08 (1.83-2.36)	1287 (34.12)	2.65 (2.31-3.04)	1343 (35.6)	2.51 (2.20-2.87)	2305 (61.11)	1.65 (1.48-1.85)
Information search										
Low (n=1613)	232 (14.38)	1.00 (reference)	332 (20.58)	1.00 (reference)	305 (18.91)	1.00 (reference)	336 (20.83)	1.00 (reference)	789 (48.92)	1.00 (reference)
High (n=4387)	1147 (26.15)	2.00 (1.71-2.35)	1385 (31.57)	1.71 (1.49-1.97)	1318 (30.04)	1.73 (1.50-2.00)	1382 (31.5)	1.65 (1.44-1.90)	2561 (58.38)	1.40 (1.24-1.57)
Adding self-generated information										
Low (n=2702)	444 (16.43)	1.00 (reference)	610 (22.58)	1.00 (reference)	546 (20.21)	1.00 (reference)	591 (21.87)	1.00 (reference)	1419 (52.52)	1.00 (reference)
High (n=3298)	935 (28.35)	1.87 (1.64-2.13)	1107 (33.57)	1.64 (1.45-1.85)	1077 (32.66)	1.75 (1.55-1.98)	1127 (34.17)	1.71 (1.52-1.93)	1931 (58.55)	1.19 (1.07-1.33)
Evaluating reliability										
Low (n=2475)	378 (15.27)	1.00 (reference)	550 (22.22)	1.00 (reference)	481 (19.43)	1.00 (reference)	527 (21.29)	1.00 (reference)	1268 (51.23)	1.00 (reference)
High (n=3525)	1001 (28.4)	2.07 (1.81-2.37)	1167 (33.11)	1.70 (1.50-1.92)	1142 (32.4)	1.85 (1.63-2.10)	1191 (33.79)	1.75 (1.55-1.98)	2082 (59.06)	1.33 (1.20-1.49)
Determining relevance										
Low (n=2051)	336 (16.38)	1.00 (reference)	466 (22.72)	1.00 (reference)	407 (19.84)	1.00 (reference)	447 (21.79)	1.00 (reference)	1023 (49.88)	1.00 (reference)
High (n=3949)	1043 (26.41)	1.71 (1.49-1.97)	1251 (31.68)	1.48 (1.30-1.68)	1216 (30.79)	1.66 (1.46-1.90)	1271 (32.19)	1.58 (1.39-1.80)	2327 (58.93)	1.37 (1.23-1.53)
Operational skills										
Low (n=1076)	162 (15.06)	1.00 (reference)	222 (20.63)	1.00 (reference)	186 (17.29)	1.00 (reference)	204 (18.96)	1.00 (reference)	439 (40.8)	1.00 (reference)
High (n=4924)	1217 (24.72)	1.78 (1.48-2.14)	1495 (30.36)	1.67 (1.41-1.97)	1437 (29.18)	1.86 (1.56-2.21)	1514 (30.75)	1.76 (1.49-2.09)	2911 (59.12)	2.01 (1.75-2.32)
Navigation skills										
Low (n=1905)	430 (22.57)	1.00 (reference)	569 (29.87)	1.00 (reference)	530 (27.82)	1.00 (reference)	527 (27.66)	1.00 (reference)	1049 (55.07)	1.00 (reference)
High (n=4095)	949 (23.17)	1.06 (0.93-1.21)	1148 (28.03)	0.95 (0.84-1.07)	1093 (26.69)	0.96 (0.85-1.09)	1191 (29.08)	1.07 (0.95-1.21)	2301 (56.19)	1.06 (0.94-1.19)
Protecting privacy										
Low (n=2206)	569 (25.79)	1.00 (reference)	727 (32.96)	1.00 (reference)	687 (31.14)	1.00 (reference)	732 (33.18)	1.00 (reference)	1281 (58.07)	1.00 (reference)
High (n=3794)	810 (21.35)	0.81 (0.72-0.92)	990 (26.09)	0.75 (0.67-0.85)	936 (24.67)	0.75 (0.67-0.85)	986 (25.99)	0.72 (0.64-0.81)	2069 (54.53)	0.90 (0.81-1.01)
Total DHLI[e] score										
Low (n=3264)	606 (18.57)	1.00 (reference)	830 (25.43)	1.00 (reference)	738 (22.61)	1.00 (reference)	799 (24.48)	1.00 (reference)	1691 (51.81)	1.00 (reference)
High (n=2736)	773 (28.25)	1.66 (1.47-1.89)	887 (32.42)	1.39 (1.24-1.57)	885 (32.35)	1.56 (1.39-1.76)	919 (33.59)	1.47 (1.31-1.65)	1659 (60.64)	1.43 (1.29-1.60)

XSL•FO
RenderX

[a]eHL: eHealth literacy.

[b]AOR: adjusted odds ratio.

[c]Multivariable logistic regression analysis adjusted for all covariates (ie, gender, age groups, equivalent income, education status, marital status, cigarette smoking, alcohol consumption, physical exercise habit, and conditions leading to severe illness due to COVID-19).

[d]eHEALS: eHealth Literacy Scale.

[e]DHLI: Digital Health Literacy Instrument.

Associations of eHL Levels With Searching Specific COVID-19 Topics

The most commonly searched specific COVID-19 topics were infectivity (4015/6000, 66.92%), followed by information about vaccine (3650/6000, 60.83%; Table 8). Participants with high eHEALS were more likely to search for all COVID-19-related topics than participants with low eHEALS (Tables 9 and 10). Moreover, participants with high total DHLI scores were more likely to search for information concerning infectivity and economic and social effects. In addition, participants with higher subscores of DHLI generally were more likely to search for information on the route of infection, assessment, economic and social effects, dealing with psychological stress, and the vaccine. However, the odds of searching for information on the route of infection and refraining from specific behaviors among participants with high navigation skills scores were 0.77 times (95% CI 0.67-0.89) and 0.88 times (95% CI 0.78-0.99) lower, respectively, than those among participants with lower scores. In addition, participants with high privacy protection scores were less likely to search for information on the route of infection (AOR 0.71, 95% CI 0.63-0.82), symptoms (AOR 0.81, 95% CI 0.73-0.90), preventive measures (AOR 0.74, 95% CI 0.66-0.83), rules and behaviors (AOR 0.87, 95% CI 0.77-0.99), assessment (AOR 0.84, 95% CI 0.75-0.96), refraining from specific behaviors (AOR 0.78, 95% CI 0.70-0.88), economic and social effects (AOR 0.83, 95% CI 0.72-0.94), and dealing with psychological stress (AOR 0.77, 95% CI 0.66-0.90).

Table 8. The proportion of "yes" responses to questions on searching each topic about COVID-19.

Topic about COVID-19	Participants (N=6000), n (%)
The infectivity of the novel coronavirus	4015 (66.92)
Information about the novel coronavirus vaccine	3650 (60.83)
Symptoms of the novel coronavirus	2494 (41.57)
Things individuals can do to prevent novel coronavirus infection	1920 (32)
Refraining from certain behaviors	1915 (31.92)
Rules and behavior regarding novel coronavirus infection prevention	1530 (25.5)
Assessment of the current novel coronavirus infection status and recommendations	1445 (24.08)
The economic and social effects of the novel coronavirus	1260 (21)
Route of infection of the novel coronavirus	1172 (19.53)
How to deal with the psychological stress caused by the novel coronavirus	775 (12.92)

XSL•FO
RenderX

Table 9. Associations of eHL[a] levels with searching specific COVID-19 topics.

eHL	The infectivity		Route of infection		Symptoms		Preventive measures		Rules and behaviors	
	Value, n (%)	AOR[b] (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]
eHEALS[d]										
Low (n=2228)	1401 (62.88)	1.00 (reference)	319 (14.32)	1.00 (reference)	755 (33.89)	1.00 (reference)	553 (24.82)	1.00 (reference)	400 (17.95)	1.00 (reference)
High (n=3772)	2614 (69.3)	1.25 (1.11-1.40)	853 (22.61)	1.63 (1.41-1.89)	1739 (46.1)	1.58 (1.41-1.77)	1367 (36.24)	1.58 (1.40-1.79)	1130 (29.96)	1.78 (1.56-2.03)
Information search										
Low (n=1613)	1010 (62.62)	1.00 (reference)	273 (16.92)	1.00 (reference)	646 (40.05)	1.00 (reference)	491 (30.44)	1.00 (reference)	388 (24.05)	1.00 (reference)
High (n=4387)	3005 (68.5)	1.25 (1.11-1.41)	899 (20.49)	1.21 (1.04-1.41)	1848 (42.12)	1.06 (0.94-1.19)	1429 (32.57)	1.07 (0.94-1.22)	1142 (26.03)	1.08 (0.94-1.24)
Adding self-generated information										
Low (n=2702)	1771 (65.54)	1.00 (reference)	453 (16.77)	1.00 (reference)	1124 (41.6)	1.00 (reference)	824 (30.5)	1.00 (reference)	653 (24.17)	1.00 (reference)
High (n=3298)	2244 (68.04)	1.05 (0.94-1.18)	719 (21.8)	1.29 (1.12-1.47)	1370 (41.54)	0.96 (0.86-1.07)	1096 (33.23)	1.09 (0.98-1.23)	877 (26.59)	1.09 (0.97-1.23)
Evaluating reliability										
Low (n=2475)	1655 (66.87)	1.00 (reference)	403 (16.28)	1.00 (reference)	1022 (41.29)	1.00 (reference)	772 (31.19)	1.00 (reference)	611 (24.69)	1.00 (reference)
High (n=3525)	2360 (66.95)	0.96 (0.86-1.08)	769 (21.82)	1.33 (1.16-1.53)	1472 (41.76)	1.02 (0.92-1.14)	1148 (32.57)	1.09 (0.97-1.22)	919 (26.07)	1.10 (0.97-1.24)
Determining relevance										
Low (n=2051)	1350 (65.82)	1.00 (reference)	344 (16.77)	1.00 (reference)	820 (39.98)	1.00 (reference)	594 (28.96)	1.00 (reference)	473 (23.06)	1.00 (reference)
High (n=3949)	2665 (67.49)	1.03 (0.92-1.16)	828 (20.97)	1.23 (1.07-1.42)	1674 (42.39)	1.05 (0.94-1.18)	1326 (33.58)	1.18 (1.05-1.34)	1057 (26.77)	1.16 (1.02-1.32)
Operational skills										
Low (n=1076)	625 (58.09)	1.00 (reference)	187 (17.38)	1.00 (reference)	385 (35.78)	1.00 (reference)	284 (26.39)	1.00 (reference)	234 (21.75)	1.00 (reference)
High (n=4924)	3390 (68.85)	1.59 (1.38-1.83)	985 (20)	1.10 (0.92-1.31)	2109 (42.83)	1.37 (1.19-1.58)	1636 (33.23)	1.48 (1.27-1.73)	1296 (26.32)	1.39 (1.18-1.64)
Navigation skills										
Low (n=1905)	1273 (66.82)	1.00 (reference)	433 (22.73)	1.00 (reference)	822 (43.15)	1.00 (reference)	639 (33.54)	1.00 (reference)	506 (26.56)	1.00 (reference)
High (n=4095)	2742 (66.96)	1.06 (0.94-1.19)	739 (18.05)	0.77 (0.67-0.89)	1672 (40.83)	0.96 (0.86-1.08)	1281 (31.28)	0.98 (0.86-1.10)	1024 (25.01)	1.01 (0.89-1.16)
Privacy protection										
Low (n=2206)	1490 (67.54)	1.00 (reference)	515 (23.35)	1.00 (reference)	1002 (45.42)	1.00 (reference)	816 (36.99)	1.00 (reference)	620 (28.11)	1.00 (reference)
High (n=3794)	2525 (66.55)	1.02 (0.91-1.15)	657 (17.32)	0.71 (0.63-0.82)	1492 (39.33)	0.81 (0.73-0.90)	1104 (29.1)	0.74 (0.66-0.83)	910 (23.99)	0.87 (0.77-0.99)
Total DHLI [e] score										
Low (n=3264)	2142 (65.63)	1.00 (reference)	614 (18.81)	1.00 (reference)	1378 (42.22)	1.00 (reference)	1053 (32.26)	1.00 (reference)	833 (25.52)	1.00 (reference)
High (n=2736)	1873 (68.46)	1.14 (1.02-1.28)	558 (20.39)	1.06 (0.93-1.21)	1116 (40.79)	0.95 (0.86-1.06)	867 (31.69)	1.01 (0.90-1.13)	697 (25.48)	1.04 (0.92-1.18)

[a]eHL: eHealth literacy.

XSL•FO
RenderX

[b]AOR: adjusted odds ratio.

[c]The multivariable logistic regression model that adjusted for all covariates (ie, gender, age groups, equivalent income, education status, marital status, cigarette smoking, alcohol consumption, physical exercise habit, and conditions leading to severe illness due to COVID-19).

[d]eHEALS: eHealth Literacy Scale.

[e]DHLI: Digital Health Literacy Instrument.

XSL•FO
RenderX

Table 10. Associations of eHL[a] levels with searching the specific COVID-19 topics (continued).

eHL	Assessment of the current novel coronavirus infection status		Refraining from specific behaviors		The economic and social effects of the novel coronavirus		Dealing with the psychological stress caused by the novel coronavirus		Information about the novel coronavirus vaccine	
	Value, n (%)	AOR[b] (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]
eHEALS[d]										
Low (n=2228)	371 (16.65)	1.00 (reference)	579 (25.99)	1.00 (reference)	322 (14.45)	1.00 (reference)	153 (6.87)	1.00 (reference)	1229 (55.16)	1.00 (reference)
High (n=3772)	1074 (28.47)	1.81 (1.58-2.07)	1336 (35.42)	1.43 (1.27-1.61)	938 (24.87)	1.83 (1.59-2.11)	622 (16.49)	2.47 (2.04-2.98)	2421 (64.18)	1.43 (1.28-1.60)
Information search										
Low (n=1613)	333 (20.64)	1.00 (reference)	498 (30.87)	1.00 (reference)	299 (18.54)	1.00 (reference)	167 (10.35)	1.00 (reference)	956 (59.27)	1.00 (reference)
High (n=4387)	1112 (25.35)	1.24 (1.08-1.43)	1417 (32.3)	1.02 (0.90-1.16)	961 (21.91)	1.19 (1.02-1.38)	608 (13.86)	1.35 (1.12-1.63)	2694 (61.41)	1.13 (1.01-1.28)
Adding self-generated information										
Low (n=2702)	614 (22.72)	1.00 (reference)	859 (31.79)	1.00 (reference)	514 (19.02)	1.00 (reference)	306 (11.32)	1.00 (reference)	1692 (62.62)	1.00 (reference)
High (n=3298)	831 (25.2)	1.05 (0.92-1.19)	1056 (32.02)	0.94 (0.84-1.06)	746 (22.62)	1.18 (1.04-1.34)	469 (14.22)	1.23 (1.05-1.44)	1958 (59.37)	0.91 (0.81-1.01)
Evaluating reliability										
Low (n=2475)	544 (21.98)	1.00 (reference)	789 (31.88)	1.00 (reference)	462 (18.67)	1.00 (reference)	290 (11.72)	1.00 (reference)	1536 (62.06)	1.00 (reference)
High (n=3525)	901 (25.56)	1.17 (1.03-1.32)	1126 (31.94)	0.99 (0.89-1.12)	798 (22.64)	1.24 (1.09-1.42)	485 (13.76)	1.17 (1.00-1.38)	2114 (59.97)	1.00 (0.90-1.12)
Determining relevance										
Low (n=2051)	429 (20.92)	1.00 (reference)	615 (29.99)	1.00 (reference)	391 (19.06)	1.00 (reference)	230 (11.21)	1.00 (reference)	1222 (59.58)	1.00 (reference)
High (n=3949)	1016 (25.73)	1.22 (1.07-1.40)	1300 (32.92)	1.09 (0.97-1.23)	869 (22.01)	1.14 (1.00-1.31)	545 (13.8)	1.19 (1.01-1.41)	2428 (61.48)	1.10 (0.98-1.23)
Operational skills										
Low (n=1076)	177 (16.45)	1.00 (reference)	292 (27.14)	1.00 (reference)	164 (15.24)	1.00 (reference)	127 (11.8)	1.00 (reference)	582 (54.09)	1.00 (reference)
High (n=4924)	1268 (25.75)	1.73 (1.44-2.07)	1623 (32.96)	1.34 (1.15-1.57)	1096 (22.26)	1.57 (1.30-1.88)	648 (13.16)	1.15 (0.93-1.42)	3068 (62.31)	1.60 (1.39-1.84)
Navigation skills										
Low (n=1905)	454 (23.83)	1.00 (reference)	661 (34.7)	1.00 (reference)	421 (22.1)	1.00 (reference)	277 (14.54)	1.00 (reference)	1093 (57.38)	1.00 (reference)
High (n=4095)	991 (24.2)	1.07 (0.94-1.22)	1254 (30.62)	0.88 (0.78-0.99)	839 (20.49)	0.94 (0.82-1.08)	498 (12.16)	0.86 (0.73-1.01)	2557 (62.44)	1.30 (1.16-1.45)
Privacy protection										
Low (n=2206)	589 (26.7)	1.00 (reference)	799 (36.22)	1.00 (reference)	521 (23.62)	1.00 (reference)	334 (15.14)	1.00 (reference)	1305 (59.16)	1.00 (reference)
High (n=3794)	856 (22.56)	0.84 (0.75-0.96)	1116 (29.41)	0.78 (0.70-0.88)	739 (19.48)	0.83 (0.72-0.94)	441 (11.62)	0.77 (0.66-0.90)	2345 (61.81)	1.16 (1.04-1.29)
Total DHLI[e] score										
Low (n=3264)	751 (23.01)	1.00 (reference)	1054 (32.29)	1.00 (reference)	645 (19.76)	1.00 (reference)	404 (12.38)	1.00 (reference)	1980 (60.66)	1.00 (reference)

XSL•FO
RenderX

eHL	Assessment of the current novel coronavirus infection status		Refraining from specific behaviors		The economic and social effects of the novel coronavirus		Dealing with the psychological stress caused by the novel coronavirus		Information about the novel coronavirus vaccine	
	Value, n (%)	AOR[b] (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]	Value, n (%)	AOR (95% CI)[c]
High (n=2736)	694 (25.37)	1.12 (0.99-1.27)	861 (31.47)	0.98 (0.87-1.09)	615 (22.48)	1.18 (1.03-1.34)	371 (13.56)	1.11 (0.95-1.29)	1670 (61.04)	1.10 (0.99-1.23)

[a]eHL: eHealth literacy.

[b]AOR: adjusted odds ratio.

[c]The multivariable logistic regression model that adjusted for all covariates (gender, age groups, equivalent income, education status, marital status, cigarette smoking, alcohol consumption, physical exercise habit, and conditions leading to severe illness due to COVID-19).

[d]eHEALS: eHealth Literacy Scale.

[e]DHLI: Digital Health Literacy Instrument.

Difficulties in Seeking and Using COVID-19 Information

Difficulties in seeking and using COVID-19 information were examined using a qualitative content analysis of 6000 valid answers to open-ended questions. Excluding 3151 (52.52%) participants who responded as perceiving no difficulties, we have listed the top 50 categories and themes (Table 11). “Information quality and credibility,” as theme I, included *information discernment* and *disinformation.* “Abundance and shortage of relevant information,” as theme II, included *incomprehensible information* and *information overload*. “Public trust and skepticism,” as theme III, included *doubting (local) governments* and *doubting specialists and doctors*. “Credibility of COVID-19–related information,” as theme IV, included *vaccination information*. These themes, including top 10 categories, cover common difficulties among people. “Privacy and security concerns,” as theme V, included *protecting personal information*. “Information retrieval challenges,” as theme VI, included *time-consuming information search*. “Anxieties and panic,” as theme VII, included *anxiety and panic*. “Movement restriction,” as theme VIII, included *time-consuming information search*. The number of categories in themes V to VIII was fewer than that in themes I to IV, indicating that the latter themes were related with relatively more specific difficulties.

Moreover, we analyzed the association between eHL and difficulties in seeking and using COVID-19 information (Table 12). The participants with higher total DHLI scores were more likely not to respond and be disinformed and less likely to answer questions on *information discernment*, *incomprehensible information*, and *information overload*. Half of the participants (3151/6000, 52.52%) reported no difficulty seeking and using COVID-19 information. Participants reporting *no* difficulties (*P* for trend=.01) and *incomprehensible information* (*P*<.001) demonstrated lower eHEALS scores. Regarding the DHLI, participants reporting *no* difficulties (*P*<.001) demonstrated a higher total DHLI score, while those reporting *information discernment* (*P*<.001), *incomprehensible information* (*P*<.001), *information overload* (*P*=.003), and *disinformation* (*P*=.02) had a lower score.

XSL•FO
RenderX

Table 11. Top 50 categories and themes of difficulties in seeking and using COVID-19 information.

Categories	Themes								
	I[a]	II[b]	III[c]	IV[d]	V[e]	VI[f]	VII[g]	VIII[h]	IX[i]
1. Information discernment	✓								
2. Incomprehensible information		✓							
3. Information overload		✓							
4. Vaccination information				✓					
5. Disinformation	✓								
6. Lack of information meeting their needs		✓							
7. Information without evidence	✓								
8. Information without credibility or trust	✓								
9. Lack of detailed patient information		✓							
10. Doubting (local) governments			✓						
11. Lack of information concerning their local area		✓							
12. Not seeking information									✓
13. Conflicting information	✓								
14. Lack of up-to-date information		✓							
15. Anxiety and panic						`	✓		
16. Rabble-rousing information	✓								
17. Insufficient aggregated information of patients		✓							
18. Doubting specialists and doctors			✓						
19. Doubting the media			✓						
20. Lack of information after infection		✓							
21. Misinformation	✓								
22. Information control and manipulation	✓								
23. Information resources	✓								
24. Time-consuming information search						✓			
25. Lack of information on prospects		✓							
26. Technical terms and jargon		✓							
27. No answers to unknown virus				✓					
28. Lack of information about other countries		✓							
29. Redundant or repetitive information		✓							
30. Information on infection risk and prevention				✓					
31. Lack of information on COVID-19 testing				✓					
32. Doubting the social media			✓						
33. Lack of information on the availability of essential services		✓							
34. Antivaccination and antigovernment	✓								
35. Lack of comprehensive information		✓							
36. Regulation and self-restraint								✓	
37. Operating PCs and smartphones						✓			
38. Protecting personal information					✓				
39. Lack of high-quality information		✓							
40. The early stage of COVID-19				✓					

XSL•FO
RenderX

Categories	Themes								
	I[a]	II[b]	III[c]	IV[d]	V[e]	VI[f]	VII[g]	VIII[h]	IX[i]
41. Doubting various authorities that lack cooperation			✓						
42. How to deal with information	✓								
43. Information on advertisement		✓							
44. Information on SARS-CoV-2				✓					
45. Differentiating COVID-19 from a cold				✓					
46. Lack of information suitable for oneself		✓							
47. Imbalance in information toward metropolitan areas		✓							
48. Lack of information for close contacts		✓							
49. Financial hardship									✓
50. Trust in authorities			✓						

[a]Theme I: information quality and credibility.
[b]Theme II: abundance and shortage of relevant information.
[c]Theme III: public trust and skepticism.
[d]Theme IV: credibility of COVID-19–related information.
[e]Theme V: privacy and security concerns.
[f]Theme VI: information retrieval challenges.
[g]Theme VII: anxieties and panic.
[h]Theme VIII: movement restriction.
[i]Theme IX: others.

Table 12. Associations of eHL[a] with difficulties in seeking and using COVID-19 information (none and the top 5 difficulties).

Difficulties	Total (n=6000), n (%)	eHEALS[b] (quartile)				*P* for trend	Total score of DHLI[c] (quartile)				*P* for trend
		Q1 (Low) (n=1504), n (%)	Q2 (n=1531), n (%)	Q3 (n=1429), n (%)	Q4 (High) (n=1536), n (%)		Q1 (Low) (n=1437), n (%)	Q2 (n=1511), n (%)	Q3 (n=1531), n (%)	Q4 (High) (n=1521), n (%)	
None	3151 (52.52)	779 (51.8)	899 (58.72)	707 (49.48)	766 (49.87)	.01	628 (43.7)	738 (48.84)	845 (55.19)	940 (61.8)	<.001
Information discernment	409 (6.82)	103 (6.85)	94 (6.14)	115 (8.05)	97 (6.32)	.94	116 (8.07)	135 (8.93)	86 (5.62)	72 (4.73)	<.001
Incomprehensible information	348 (5.8)	140 (9.31)	87 (5.68)	66 (4.62)	55 (3.58)	<.001	139 (9.67)	95 (6.29)	67 (4.38)	47 (3.09)	<.001
Information overload	272 (4.53)	64 (4.26)	69 (4.51)	70 (4.9)	69 (4.49)	.65	68 (4.73)	89 (5.89)	73 (4.77)	42 (2.76)	.003
Vaccination information	261 (4.35)	60 (3.99)	54 (3.53)	75 (5.25)	72 (4.69)	.11	61 (4.24)	73 (4.83)	70 (4.57)	57 (3.75)	.45
Disinformation	209 (3.48)	44 (2.93)	56 (3.66)	51 (3.57)	58 (3.78)	.24	41 (2.85)	47 (3.11)	55 (3.59)	66 (4.34)	.02

[a]eHL: eHealth literacy.
[b]eHEALS: eHealth Literacy Scale.
[c]DHLI: Digital Health Literacy Instrument.

Discussion

Principal Findings

This study using mixed methods is the first to examine the associations between eHL and web-based health information–seeking behaviors and to identify the difficulties in using health information on the internet and the its relationship with eHL among adult internet users. Internet users with high eHEALS and DHLI scores on information searching, adding self-generated content, evaluating reliability, determining relevance, and operational skills were more likely to use all web

XSL•FO
RenderX

sources of information about COVID-19 than those with low eHEALS and DHLI scores. However, there were negative associations between navigation skill scores and privacy protection when using several information sources, such as YouTube, to search for COVID-19 information. In addition, participants with high eHEALS and DHLI scores on information searching, adding self-generated information, evaluating reliability, and determining relevance were more likely to search for information about COVID-19 than those with low eHEALS or DHLI scores. However, some participants with high navigation skills and privacy protection skills were less likely to search for information on COVID-19. Furthermore, this study shed light on the difficulties seeking and using COVID-19 information qualitatively. While half of the participants reported no difficulty seeking and using COVID-19 information, participants who reported any difficulties, including *information discernment*, *incomprehensible information*, *information overload*, and *disinformation*, had lower DHLI score. Finally, participants expressed significant concerns regarding "information quality and credibility," "abundance and shortage of relevant information," "public trust and skepticism," and "credibility of COVID-19–related information." In addition, they disclosed more specific concerns, including "privacy and security concerns," "information retrieval challenges," "anxieties and panic," and "movement restriction."

The study results suggest that internet users with higher eHEALS and total DHLI scores were more likely to use a reliable information source, consistent with prior studies [14,20,31]. Considering the subscale of DHLI, the ability to determine relevance of and evaluate the reliability of information is reportedly positively associated with the search for COVID-19 information through a traditional 1-way communication channel known as "Health 1.0," involving public institution websites [15-18,22,32]. In addition, a previous study showed positive associations of higher skills in information searching and adding self-generated content with using public institution websites [15]. However, to our knowledge, no study has examined the association between operational skills and web-based health information–seeking behavior because most studies have focused on university students and have not assessed operational information skills [14-18]. Operational skills, which are basic skills required to use computers, are vital for searching web-based health information, with implications among people unfamiliar with computers or smartphones, such as older adults. A Japanese government survey indicated that deficiencies in the basic skills required to use computers or smartphones were barriers to internet access among older adults [1]. Therefore, this finding suggests that operational skills are critical for using web-based health information among individuals who are generally less familiar with the internet.

There were negative associations between navigation and protecting privacy skills and using several web-based interactive health-related communications channels via SNSs, such as YouTube. Several studies have reported that university students with low DHL scores are more likely to use Health 2.0 channels for health information than university students with high DHLI scores [14,17]. In addition, the study findings showed that participants with high navigation and protecting privacy skills were less likely to search for information on, for example, the route of infection or on refraining from certain behaviors. Participants with high navigation and protecting privacy skills used web-based information about COVID-19 cautiously compared to those with lower navigation and proficiency skills. However, there was no negative association between eHEALS scores and the use of Health 2.0 communication channels. This result could be explained by the reason that eHEALS scores did not encompass the skills required to use Health 2.0 [33]. The eHEALS would need to be improved for adaptation to Health 2.0 communication channels.

The leading 50 categories related to difficulties seeking and using COVID-19 information were identified using a qualitative approach. Our findings indicate that approximately half of the participants experienced difficulties. Information discernment was the most common issue. Health literacy encompasses functional, interactive, and critical literacies [34]. Information discernment is a crucial aspect of literacy. It concerns an individual's ability to discriminate misinformation from accurate information. It has been assessed by calculating the difference in scores related to discerning accurate information from misinformation [35,36]. Managing the volume of available information and assessing its quality and reliability are essential DHL skills [37]. Our results revealed that information discernment was not linked to proficiency in terms of eHEALS scores but rather to DHLI scores. This finding underscores a pivotal shift in the Health 2.0 era when basic knowledge of Health 1.0 health literacy is insufficient. Our study highlights the need for an enhanced level of health literacy tailored to facilitate the navigation of the complexities and nuances of information in the Health 2.0 landscape.

We qualitatively presented themes related to difficulties in seeking and using COVID-19 information. Themes should be evaluated against the backdrop of previous studies. Several tools or instruments for assessing the quality of health information have been used extensively, such as the *Journal of the American Medical Association* benchmarks, Sandvik's general quality criteria, DISCERN, HONcode by the Health on the Net Foundation, and quality evaluation scoring tools [38]. Denniss et al [39] recently developed the 13 Principles for Health-Related Information on Social Media (PRHISM). The US National Academy of Medicine has proposed 3 foundational principles to guide the identification of credible sources of health information on social media, namely, that they are science based, objective, transparent, and accountable [40]. "Information quality and credibility" and "credibility of COVID-19–related information" were difficulties identified in relation to science-based principles, and "relevant information" included science-based information. "Skepticism" could arise from lacking objectives, transparency, and accountability. These aspects generally align with the PRHISM. In addition, our findings highlight the significance of people-centered or narrative information. Relevant information includes detailed patient information, information suitable for patients, and information concerning the local area. Simultaneously, the demand for narrative information could lead to "privacy and security concerns," emphasizing the need to balance this aspect with personal information protection. The categories and themes

XSL•FO
RenderX

could be used to develop a comprehensive list of challenges likely to be faced during future infodemics.

It is widely acknowledged that the COVID-19 pandemic has brought to light the impact of disparities on health outcomes, thereby highlighting the imperative to tackle these inequalities. Alongside health literacy and other social determinants of health [34], the pandemic has underscored the significance of information as an independent determinant of health [41]. Limited access to high-quality information can exacerbate disparities, particularly in education and economic stability. Addressing such disparities requires collaborative efforts among stakeholders worldwide. Establishing meaningful partnerships between governmental, nongovernmental, and private-sector organizations is crucial to the creation of governance frameworks to counter information-related threats. Given the ongoing pivotal role of information in shaping health outcomes, collective action is essential to mitigate the adverse impact of misinformation. To achieve this, a fundamental framework is needed to universally enhance DHL and improve web-based content. Our insights will provide a valuable contribution to the creation of such a framework.

Limitations

This study has some limitations. First, participants were recruited from a single Japanese internet research service company. These participants were suitable for an internet-based survey because internet users need to have adequate eHL, and the participants were equally divided by gender, age, and income and then recruited by a research company. However, the results may have been biased, as our participants could have a disproportionately higher educational status and higher eHL skill levels than general internet users [42-44]. Therefore, the eHEALS scores, DHLI scores, and proportion of those searching for COVID-19 information identified in this study may have been higher than those among general internet users in Japan. Second, the study findings may not be directly generalizable to other countries because of inherent differences in website and SNS environments. Third, although previous studies have objectively evaluated the eHL of participants using performance tests [22,45], this study did not objectively assess the eHL and DHLI dimensions. Therefore, there may have been inaccuracies in estimating participants' eHL levels. Finally, this study's cross-sectional design means that causality remains unknown.

Conclusions

This study revealed that Japanese individuals with higher eHEALS and DHLI scores were more engaged in using various web sources when seeking COVID-19 information. However, proficiency in terms of the eHEALS may not encompass the skills required to use Health 2.0. Higher scores on navigation and privacy protection skills, not included in the eHEALS, correlated with less use of specific sources, such as YouTube. Participants with high navigation and privacy protection skills used web-based information about COVID-19 cautiously compared with those with lower proficiency in these skills. This study also highlights an increased need for information discernment in the Health 2.0 era. The identified categories and themes, such as "information quality and credibility," "abundance and shortage of relevant information," "public trust and skepticism," and "privacy and security concerns," suggest a framework that could be used to address the myriad challenges anticipated in future infodemics. In the future, we aim to compile a comprehensive list of such challenges. To strengthen the public's resilience against misinformation, a fundamental framework should be established to enhance DHL for all individuals and improve the quality of web-based content.

Acknowledgments

The authors wish to thank Rosalie van der Vaart (Leiden University) for her contribution to translating the items on operational and navigation skills on the Digital Health Literacy Instrument (DHLI) into Japanese. The authors used ChatGPT-4 (OpenAI) to name the categories and themes in English.

This work was supported by a Grant-in-Aid for Scientific Research (B; 20H01594) of the Japan Society for the Promotion of Science (JSPS).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital Health Literacy Instrument adapted to the COVID-19 pandemic and web-based information–seeking behavior on COVID-19 in Japanese.
[DOCX File , 25 KB - jmir_v26i1e57842_app1.docx]

References

1. Results of the communication usage trend survey of 2022. Ministry of Internal Affairs and Communications, Japan. 2023 May 29. URL: https://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/pressrelease/2023/5/29_1.html [accessed 2024-01-26]
2. National Cancer Institute. HINTS 5, cycle 4 (2020) survey dataset. Health Information National Trends Survey. 2021. URL: https://hints.cancer.gov/data/download-data.aspx [accessed 2024-06-24]
3. Digital economy and society statistics - households and individuals. Eurostat. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals [accessed 2024-01-26]

XSL•FO
RenderX

4. Mitsutake S, Takahashi Y, Otsuki A, Umezawa J, Yaguchi-Saito A, Saito J, et al. Chronic diseases and sociodemographic characteristics associated with online health information seeking and using social networking sites: nationally representative cross-sectional survey in Japan. J Med Internet Res 2023 Mar 02;25:e44741 [FREE Full text] [doi: 10.2196/44741] [Medline: 36862482]
5. Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, et al. Health information on the internet: accessibility, quality, and readability in English and Spanish. JAMA 2001 May 23;285(20):2612-2621 [FREE Full text] [doi: 10.1001/jama.285.20.2612] [Medline: 11368735]
6. Charow R, Snow M, Fathima S, Giuliani ME, McEwan K, Winegust J, et al. Evaluation of the scope, quality, and health literacy demand of internet-based anal cancer information. J Med Libr Assoc 2019 Oct 01;107(4):527-537 [FREE Full text] [doi: 10.5195/jmla.2019.393] [Medline: 31607810]
7. Fefer M, Lamb CC, Shen AH, Clardy P, Muralidhar V, Devlin PM, et al. Multilingual analysis of the quality and readability of online health information on the adverse effects of breast cancer treatments. JAMA Surg 2020 Aug 01;155(8):781-784 [FREE Full text] [doi: 10.1001/jamasurg.2020.1668] [Medline: 32520317]
8. Adebesin F, Smuts H, Mawela T, Maramba G, Hattingh M. The role of social media in health misinformation and disinformation during the COVID-19 pandemic: bibliometric analysis. JMIR Infodemiology 2023 Sep 20;3:e48620 [FREE Full text] [doi: 10.2196/48620] [Medline: 37728981]
9. Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation: joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC. World Health Organization. 2020 Sep 23. URL: https://www.who.int/news/item/23-09-2020-managing-the-Covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation [accessed 2024-01-26]
10. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676 [FREE Full text] [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
11. Paakkari L, Okan O. COVID-19: health literacy is an underestimated problem. Lancet Public Health 2020 May;5(5):e249-e250 [FREE Full text] [doi: 10.1016/S2468-2667(20)30086-4] [Medline: 32302535]
12. Okan O, Bollweg TM, Berens EM, Hurrelmann K, Bauer U, Schaeffer D. Coronavirus-related health literacy: a cross-sectional study in adults during the COVID-19 infodemic in Germany. Int J Environ Res Public Health 2020 Jul 30;17(15):5503 [FREE Full text] [doi: 10.3390/ijerph17155503] [Medline: 32751484]
13. The COVID-HL University students survey. COVID-HL Network. URL: https://covid-hl.eu/research/survey/ [accessed 2024-01-26]
14. Bak CK, Krammer J, Dadaczynski K, Orkan O, von Seelen J, Prinds C, et al. Digital health literacy and information-seeking behavior among university college students during the COVID-19 pandemic: a cross-sectional study from Denmark. Int J Environ Res Public Health 2022 Mar 19;19(6):3676 [FREE Full text] [doi: 10.3390/ijerph19063676] [Medline: 35329363]
15. Dadaczynski K, Okan O, Messer M, Leung AY, Rosário R, Darlington E, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Jan 15;23(1):e24097 [FREE Full text] [doi: 10.2196/24097] [Medline: 33395396]
16. Rosário R, Martins MR, Augusto C, Silva MJ, Martins S, Duarte A, et al. Associations between COVID-19-related digital health literacy and online information-seeking behavior among Portuguese university students. Int J Environ Res Public Health 2020 Dec 02;17(23):8987 [FREE Full text] [doi: 10.3390/ijerph17238987] [Medline: 33276647]
17. Vrdelja M, Vrbovšek S, Klopčič V, Dadaczynski K, Okan O. Facing the growing COVID-19 infodemic: digital health literacy and information-seeking behaviour of university students in Slovenia. Int J Environ Res Public Health 2021 Aug 12;18(16):8507 [FREE Full text] [doi: 10.3390/ijerph18168507] [Medline: 34444255]
18. Htay MN, Parial LL, Tolabing MC, Dadaczynski K, Okan O, Leung AY, et al. Digital health literacy, online information-seeking behaviour, and satisfaction of Covid-19 information among the university students of East and South-East Asia. PLoS One 2022;17(4):e0266276 [FREE Full text] [doi: 10.1371/journal.pone.0266276] [Medline: 35417478]
19. Assaye BT, Kassa M, Belachew M, Birhanu S, Worku A. Association of digital health literacy and information-seeking behaviors among physicians during COVID-19 in Ethiopia: a cross-sectional study. Digit Health 2023 Jun 04;9:20552076231180436 [FREE Full text] [doi: 10.1177/20552076231180436] [Medline: 37312956]
20. Mitsutake S, Shibata A, Ishii K, Okazaki K, Oka K. [Developing Japanese version of the eHealth literacy scale (eHEALS)]. Nihon Koshu Eisei Zasshi 2011 May;58(5):361-371. [Medline: 21905612]
21. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
22. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
23. Arias López MD, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJ, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]

XSL•FO
RenderX

24. Miyawaki R, Kato M, Kawamura Y, Ishikawa H, Oka K. [Developing a Japanese version of the Digital Health Literacy Instrument]. Nihon Koshu Eisei Zasshi 2024 Jan 26;71(1):3-14 [FREE Full text] [doi: 10.11236/jph.23-021] [Medline: 37673595]
25. Overview of the 2019 comprehensive survey of living conditions. Ministry of Health, Labour and Welfare, Japan. URL: https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa19/index.html [accessed 2024-01-26]
26. Creswell JW, Plano Clark VL, Gutmann ML, Hanson WE. Advanced mixed methods research designs. In: Tashakkori A, Teddlie C, editors. Handbook of Mixed Methods in Social and Behavioral Research. Thousand Oaks, CA: SAGE Publications; 2003:209-240.
27. Curry L, Nunez-Smith M. Mixed Methods in Health Sciences Research: A Practical Primer. Thousand Oaks, CA: SAGE Publications; 2015.
28. Takahashi Y, Uchida C, Miyaki K, Sakai M, Shimbo T, Nakayama T. Potential benefits and harms of a peer support social network service on the internet for people with depressive tendencies: qualitative content analysis and social network analysis. J Med Internet Res 2009 Jul 23;11(3):e29 [FREE Full text] [doi: 10.2196/jmir.1142] [Medline: 19632979]
29. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
30. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
31. Inoue M, Shimoura K, Nagai-Tanima M, Aoyama T. The relationship between information sources, health literacy, and COVID-19 knowledge in the COVID-19 infodemic: cross-sectional online study in Japan. J Med Internet Res 2022 Jul 22;24(7):e38332 [FREE Full text] [doi: 10.2196/38332] [Medline: 35839380]
32. Van De Belt TH, Engelen LJ, Berben SA, Schoonhoven L. Definition of Health 2.0 and Medicine 2.0: a systematic review. J Med Internet Res 2010 Jun 11;12(2):e18 [FREE Full text] [doi: 10.2196/jmir.1350] [Medline: 20542857]
33. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec 23;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
34. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 01;42:159-173 [FREE Full text] [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
35. Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S. Psychological inoculation improves resilience against misinformation on social media. Sci Adv 2022 Aug 26;8(34):eabo6254 [FREE Full text] [doi: 10.1126/sciadv.abo6254] [Medline: 36001675]
36. Lu C, Hu B, Li Q, Bi C, Ju XD. Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: systematic review and meta-analysis. J Med Internet Res 2023 Aug 29;25:e49255 [FREE Full text] [doi: 10.2196/49255] [Medline: 37560816]
37. Nutbeam D. From health education to digital health literacy - building on the past to shape the future. Glob Health Promot 2021 Dec;28(4):51-55. [doi: 10.1177/17579759211044079] [Medline: 34719292]
38. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018 Oct 19;18(1):87 [FREE Full text] [doi: 10.1186/s12911-018-0668-9] [Medline: 30340488]
39. Denniss E, Lindberg R, McNaughton SA. Development of principles for health-related information on social media: Delphi study. J Med Internet Res 2022 Sep 08;24(9):e37337 [FREE Full text] [doi: 10.2196/37337] [Medline: 36074544]
40. Kington RS, Arnesen S, Chou WY, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021:10.31478/202107a [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
41. Graham G, Goren N, Sounderajah V, DeSalvo K. Information is a determinant of health. Nat Med 2024 Apr;30(4):927-928. [doi: 10.1038/s41591-023-02792-9] [Medline: 38351186]
42. Eysenbach G, Wyatt J. Using the internet for surveys and health research. J Med Internet Res 2002;4(2):E13 [FREE Full text] [doi: 10.2196/jmir.4.2.e13] [Medline: 12554560]
43. Mitsutake S, Shibata A, Ishii K, Oka K. Association of eHealth literacy with colorectal cancer knowledge and screening practice among internet users in Japan. J Med Internet Res 2012 Nov 13;14(6):e153 [FREE Full text] [doi: 10.2196/jmir.1927] [Medline: 23149453]
44. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet. J Med Internet Res 2016 Jul 18;18(7):e192 [FREE Full text] [doi: 10.2196/jmir.5413] [Medline: 27432783]
45. Neter E, Brainin E. Perceived and performed eHealth literacy: survey and simulated performance test. JMIR Hum Factors 2017 Jan 17;4(1):e2 [FREE Full text] [doi: 10.2196/humanfactors.6523] [Medline: 28096068]

Abbreviations

AOR: adjusted odds ratio
COVID-HL: COVID-19 health literacy
DHL: digital health literacy

XSL•FO
RenderX

DHLI: Digital Health Literacy Instrument
eHEALS: eHealth Literacy Scale
eHL: eHealth literacy
J-eHEALS: Japanese version of the eHealth Literacy Scale
PRHISM: 13 Principles for Health-Related Information on Social Media
SNS: social networking site

Edited by G Eysenbach, N Cahill; submitted 27.02.24; peer-reviewed by B Assaye; comments to author 29.03.24; revised version received 02.05.24; accepted 29.05.24; published 11.07.24.

Please cite as:
Mitsutake S, Oka K, Okan O, Dadaczynski K, Ishizaki T, Nakayama T, Takahashi Y
eHealth Literacy and Web-Based Health Information–Seeking Behaviors on COVID-19 in Japan: Internet-Based Mixed Methods Study
J Med Internet Res 2024;26:e57842
URL: https://www.jmir.org/2024/1/e57842
doi:10.2196/57842
PMID:38990625

©Seigo Mitsutake, Koichiro Oka, Orkan Okan, Kevin Dadaczynski, Tatsuro Ishizaki, Takeo Nakayama, Yoshimitsu Takahashi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Age Differences in Electronic Mental Health Literacy: Qualitative Study

Richard Huan Xu[1], PhD; Lidan Tian[1], MA; Liling Zhu[2], MA; Yuan Cao[3], PhD; Sherry Kit-wa Chan[4], MD; Dong Dong[5], PhD; Wai-ling Annie Cheung[5], MPhil; Eliza Lai-yi Wong[5], PhD

[1]Department of Rehabilitaion Sciences, Faculty of Health and Social Sciences, Hong Kong Polytechnic University, Kowloon, China (Hong Kong)
[2]Department of Applied Social Science, Faculty of Health and Social Sciences, Hong Kong Polytechnic University, Kowloon, China (Hong Kong)
[3]Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, China (Hong Kong)
[4]Department of Psychiatry, The University of Hong Kong, Hong Kong, China (Hong Kong)
[5]JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, China (Hong Kong)

Corresponding Author:
Richard Huan Xu, PhD
Department of Rehabilitaion Sciences
Faculty of Health and Social Sciences
Hong Kong Polytechnic University
11 Yuk Choi Rd
Hung Hom
Kowloon
China (Hong Kong)
Phone: 852 27664199
Email: richard.xu@polyu.edu.hk

Abstract

Background: Electronic mental health literacy (eMHL) is critical for accessing and effectively using digital mental health resources. However, there is a paucity of research on how eMHL varies across age groups.

Objective: This study aimed to investigate differences in eMHL among young, middle-aged, and older adults; provide insights into the needs, behaviors, and attitudes of different age groups in relation to digital mental health resources; and ultimately, inform the improvement of mental health services.

Methods: A qualitative investigation was conducted to examine the differences in eMHL across different age demographics in the Chinese population in 2023. The study sample comprised 3 distinct age groups: 18-34 years, 35-64 years, and 65 years and older. Participants were recruited through purposive sampling to ensure a diverse representation of the population. Data were collected through semistructured one-on-one interviews, which allowed for in-depth exploration of individual experiences and perceptions. The gathered data were subsequently subjected to rigorous thematic analysis to enable the identification and interpretation of recurring patterns and themes.

Results: The principal outcomes derived from these interviews were synthesized into 5 distinct dimensions: emotional needs, use of digital mental health resources, assessment of digital mental health information, engagement with social media to regulate emotions, and coping strategies. These dimensions were uniformly observed across the 3 age groups.

Conclusions: We identified differences in knowledge, skills, and attitudes regarding the use of web-based information for managing mental health problems between the 3 age groups. The findings highlight the importance of age-specific strategies for improving eMHL.

(J Med Internet Res 2024;26:e59131) doi:10.2196/59131

KEYWORDS

eHealth literacy; mental health; mental health literacy; age-related difference; electronic mental health literacy; eMHL

XSL•FO
RenderX

Introduction

Traditional mental health services frequently operate at their maximum capacity due to constrained resources. Consequently, some individuals are deprived of the care that they require. Digital mental health platforms, encompassing both mobile and web-based systems, are a potential solution to this challenge [1]. These platforms, which include services such as digital counseling, forums, and chat groups, offer an alternative avenue for individuals who are reluctant to engage in face-to-face therapy or feel alienated by traditional services [2]. Studies have reported age differences in the use of digital mental health platforms. For example, a study in Ireland reported that about half of the respondents aged 18 to 25 years used social media and apps for mental health support [3]. Another study in the United Kingdom showed that nearly two-thirds of young people with a history of mental health needs had accessed digital mental health tools [4]. In contrast, a study in the United States found that only 14.2% of middle-aged individuals used digital mental health tools to cope with stress related to the COVID-19 pandemic [5]. Older adults exhibit a lower acceptance of digital mental health resources than middle-aged and younger adults [6]. This could be due to their general lack of digital skills and discomfort with new technologies or because digital interventions are typically designed for advanced users [7-9].

Research has found that age differences in the use of the internet to manage mental health can be attributed to various factors. The health information demands of older persons differ from those of younger adults [10], including those regarding chronic diseases, social isolation [11], and physical and cognitive limitations due to the natural aging process [12,13]. Studies have also highlighted generational differences in information-seeking behavior in nonhealth contexts. For children and young people, moderate internet use can have a small positive impact on their well-being. Young people use the internet to manage emotional problems, although this use could become a maladaptive coping strategy. They may use the internet to distract themselves from negative emotions; escape from reality; or regulate feelings of loneliness, depression, or anxiety [14]. However, social media platforms can exacerbate feelings of anxiety, isolation, and depression, especially among young people who spend time on the internet at the expense of in-person interaction [15,16]. In summary, younger and older adults require different approaches to health education, treatment options, and digital support systems to effectively address their unique challenges and promote their overall well-being.

Electronic mental health literacy (eMHL) is an emerging concept that extends mental health literacy into the digital era. It reflects the ability of individuals to access, understand, and use digital mental health information and services [17], an ability that is essential for the recognition, management, and prevention of mental health issues.

Studies on eMHL have primarily focused on younger populations. For instance, Cormier et al [18] found a close relationship between low eMHL and an increased risk of mental health problems. Another study on young adults found that those with low eMHL are at high risk of experiencing a mental health disorder [19]. However, mental health issues affect individuals across all age groups, from children and adolescents to adults and older individuals. It is crucial to understand the prevalence and impact of these issues across different age brackets to develop effective treatment strategies. The internet plays an increasingly important role in understanding and addressing the mental health problems of individuals across various age groups. It offers a wealth of resources and information and facilitates access to communities for mental health support. Currently, there is a lack of qualitative evidence on the eMHL of people across different age groups. Research on this issue is necessary to understand how different age groups perceive and use eHealth resources. This knowledge could inform the design and implementation of more effective and personalized digital health interventions. Furthermore, it could aid in the development of strategies to enhance digital health literacy, thereby improving health outcomes for individuals across all age brackets. Therefore, this study investigated age differences in eMHL between young, middle-aged, and older adults. The findings offer insights into the needs, behaviors, and attitudes of individuals of different age groups in relation to digital health, thus informing efforts to improve mental health services.

Methods

Design and Sampling

We conducted a qualitative study to understand individuals' experiences of seeking and using information on the internet to manage their mental health. The study involved in-depth, semistructured interviews with participants from the 3 adult age groups suggested in the Hong Kong Census guidelines: young adults (18-34 years), middle-aged adults (35-64 years), and older adults (65 years and older). The study was conducted from May to September 2023. Participants were recruited via research flyers, which were shared on university notice boards and distributed via the university's electronic mass mail system to all university visitors, students, staff, collaborators, and alumni. The flyers included a detailed study description, participation criteria, information about the offline interviews (each approximately 60 minutes long), the impact of the study, and the researchers' contact information. Interested individuals filled out a web-based questionnaire in which they provided their background information. The research team then screened the prospective participants based on age, gender, and their experiences with counseling. Participants had to fulfill the following criteria: (1) be citizens and residents of Hong Kong, (2) be Cantonese speakers, (3) be aged between 18 and 70 years, and (4) have experienced mild to severe mental health problems (as determined by the Depression, Anxiety and Stress Scale-21 Item [DASS-21] questionnaire [20]).

Data Collection Procedure

As there is no gold standard for sample sizes in qualitative studies, we followed Creswell's [21] recommendation and aimed to recruit 6-8 participants from each age group, totaling 18-24 participants across the 3 groups. However, the final sample size was determined based on data saturation, the point at which no new themes emerged from the participants' experiences. In total, 292 individuals registered for the survey by scanning the

XSL•FO
RenderX

QR code on the flyers. Following the screening of participants' mental health status, we identified 109 individuals with mild or severe mental health issues, as measured by DASS-21, as suitable for inclusion in the study. We then divided these eligible participants into 3 age groups. Each of these individuals was allocated a random number from 1 to 109. A research assistant then contacted them in ascending order of their numbers to organize interviews with them. Once their willingness to be interviewed in person and availability were confirmed, they were invited for the interview session in a university conference room. We stopped this process when saturation was reached, after which no further participants were interviewed.

Before the interview, the participants were given a consent form to review. Any doubts that they had were clarified by the research team. The form covered permissions for recording the interview, assurances of anonymity, and data retention periods. The semistructured interview was conducted after they had signed the informed consent form. In the interview, topics, such as the participants' background, daily mental health experiences, and their knowledge about and interaction with digital mental health resources, were discussed. An interviewer experienced in qualitative research and fluent in Cantonese led the sessions with the help of a research assistant who took notes to capture important information.

Ethical Considerations

The Research Ethics Committee at the Hong Kong Polytechnic University approved the study protocol and the procedure for obtaining informed consent (HSEARS20211130001). All of the participants provided written informed consent. The data generated in the study were anonymized. Each participant received a HK $100 (equal to US $12.8) local supermarket coupon after the interview as a token of appreciation.

Data Analysis

The interview recordings were automatically transcribed using digital software (iFLYREC) and manually proofread. Thematic analyses were used for data synthesis to understand the participants' perspectives on digital mental health resources and the approach they adopted when seeking professional help. We conducted a reflexive thematic analysis using a realist framework and integrated the inductive and deductive approaches. The analysis comprised six steps [22]: (1) data familiarization, (2) coding, (3) initial theme generation, (4) theme review, (5) theme definition or naming, and (6) writing of results. The 2 researchers (RHX and LZ) were responsible for familiarizing themselves with the data and generating codes and initial themes. This process involved line-by-line open coding and discussions between the 2 researchers for the first 3 transcripts (1 for each age group). Afterward, 2 researchers (RHX and LZ) reviewed and revised the codes and themes through focused coding of the first 3 transcripts and 3 additional transcripts (1 for each age group). Interviews were conducted in a preset order: 2 from the young adult group, followed by 2 from the middle-aged adult group, and then 2 from the older adult group, continuing in this sequence until the end. After each interview round, the authors reviewed the data for emerging themes or information. We checked for patterns, concepts, and categories. If new or relevant themes were discovered, we continued the interviews until they started yielding redundant information, indicating data saturation. To confirm saturation, we conducted 1 or 2 additional interviews in each age group. If these offered no new insights, we concluded that saturation was achieved and data collection was ended.

Thematic analysis was conducted in a stepwise manner. First, data related to the different age groups were analyzed individually. This required a thorough exploration of the data from each group to identify unique patterns and trends. The analysis began with data from the young adult group, followed by the middle-aged adult group, and finally, the older adult group. After the initial analysis, the findings for the different age groups were then synthesized. The synthesis allowed for the generation of overarching themes and patterns that spanned the age groups, providing a comprehensive understanding of the data as a whole. This process was conducted in a recursive and reflexive manner. All of the data were analyzed in Chinese, the language in which the interviews were conducted, to ensure accuracy. For the preparation of this paper, key findings and selected quotes were translated into English with the help of a translator from a professional translation company. The translator was a native English speaker with a good understanding of Chinese. All of the translations were jointly confirmed by the research team and the translator. Dedoose was used for coding and thematic analysis.

Results

Overview

We interviewed 30 participants across the 3 age groups, at which point our data reached saturation. Table 1 summarizes the backgrounds of the participants. The primary findings were synthesized into a thematic map (Figure 1). We organized them into 5 categories: emotional needs, use of digital mental health resources, assessment of digital mental health information, engagement with social media for emotional management, and coping strategies. These categories were applicable to all 3 age groups. In the selected quotes presented in this paper, the numbers referring to the participants are prefixed with Y, M, or O to indicate whether they belonged to the young adult, middle-aged adult, or older adult group, respectively.

XSL•FO
RenderX

Table 1. Demographics of participants of in-depth interview (N=30).

Characteristics	Values, n (%)
Sex	
Male	14 (47)
Female	16 (53)
Age groups (years)	
18-34	10 (33)
35-64	10 (33)
≥65	10 (33)
Marital status	
Unmarried	13 (43)
Married	16 (53)
Divorced or separated or widowed	1 (3)
Employment	
Full-time employed	18 (60)
Retired	5 (17)
Student	3 (10)
Part-time employed	1 (0.3)
Unemployed	1 (0.3)
Housewife	1 (0.3)
Others	1 (0.3)
Government assistance received	
No	30 (100)
Educational level	
Secondary or below	3 (10)
Postsecondary	5 (17)
Tertiary or above	22 (73)
Have seen a psychiatrist or received psychological counseling in the past year	
Yes	7 (23)
No	23 (77)
DASS-21[a]	
Depression subscale	
Mild	1 (3)
Moderate	17 (57)
Severe	3 (10)
Extremely severe	9 (30)
Anxiety subscale	
Mild	1 (3)
Moderate	9 (30)
Severe	7 (23)
Extremely severe	13 (43)
Stress subscale	
Mild	7 (23)
Moderate	10 (33)

XSL•FO
RenderX

Characteristics	Values, n (%)
Severe	6 (21)
Extremely severe	7 (23)

[a]DASS-21: Depression, Anxiety and Stress Scale-21 Item.

Figure 1. Thematic map of the findings.

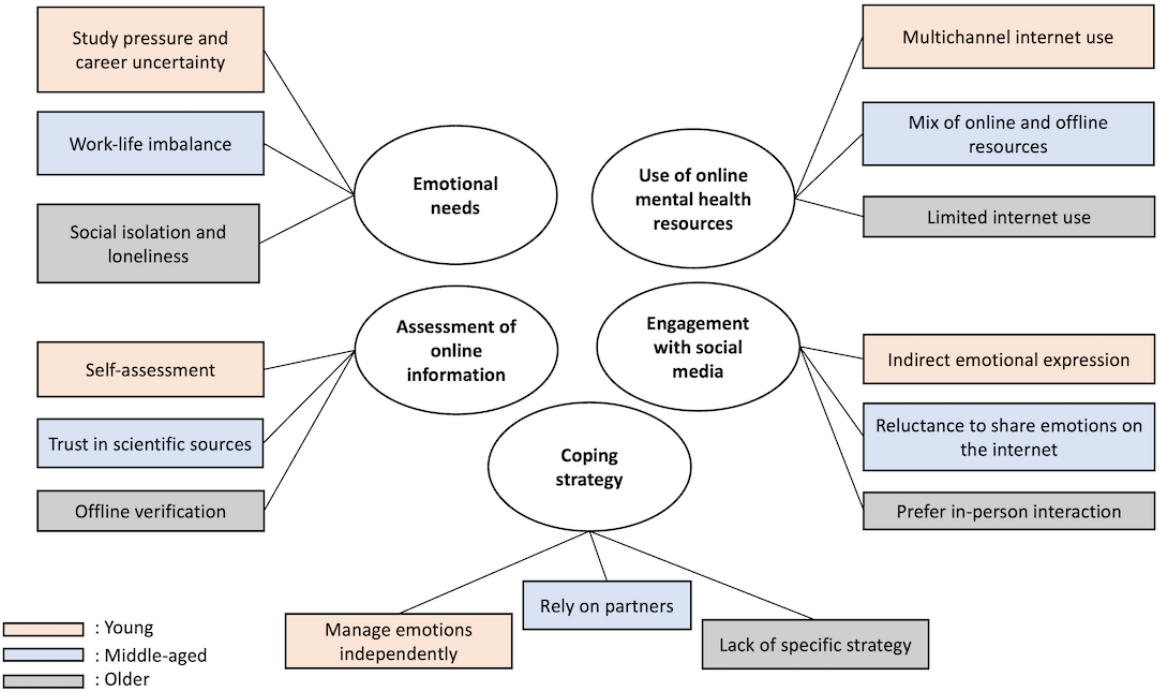

Impact of Emotional Needs on Web-Based Behavior

Study Pressure and Career Uncertainty Among the Young Adults

The pressure of academia and entry into the workforce motivated the young participants to seek digital sources to deal with their emotions. A common theme among this group was uncertainty about education and job prospects, especially for those seeking internships or jobs. Some also mentioned that relationships caused them anxiety and depression.

> *I experienced a lot of stress while job hunting after graduation, so I often browsed online forums and videos for tips and motivation.* [Y 012]

Work-Life Imbalance Among the Middle-Aged Individuals

The middle-aged participants experienced noticeable anxiety and depression. A significant majority of them were employed and faced the delicate task of balancing work commitments with family responsibilities, driving their need to find flexible web-based support.

> *Given my demanding workload, I really need some easily accessible online resources that I can use whenever I have time, such as mental health chatbots.* [M 013]

> *Because I often feel anxious about my work, I go online to look for articles and resources on how to relieve my emotions.* [M 026]

Social Isolation and Loneliness in the Older Adults

Navigating life after retirement posed significant challenges for the older adult participants. The older adults found it crucial to address how to structure their daily lives following retirement to maintain their mental well-being. This necessity drove some to seek support online and offline.

> *After retirement and my father's passing, I lived alone and felt extremely lonely and bored. I realised how isolating it could be for someone unable to go out and interact with others. I then sought help from social workers and a psychiatrist. I was also recommended to go online to find local community groups and activities to join.* [O 010]

Use of Digital Mental Health Resources

Multichannel Internet Use Among the Young Adults

The young individuals, being adept at using the internet, leveraged various digital platforms for emotional relief. They engaged with autonomous sensory meridian response videos, virtual city tours, and resources providing insights into professional life, gaming, music, and short videos. While not directly addressing emotions, these activities provided effective relief in times of emotional difficulty. These participants also

XSL•FO
RenderX

showed interest in mental health services for relaxation and personal growth.

> *I look for resources on many social media platforms, such as official accounts on the Little Red Book, Bilibili, and WeChat. I follow those that I resonate with.* [Y012]

> *I try watching motivational videos on YouTube to see if they help with my emotions.* [Y 016]

Use of a Mix of Online and Offline Resources Among the Middle-Aged Individuals

The middle-aged participants used a mix of traditional and new resources for emotional well-being. The digital resources that they used included platforms, such as YouTube and Facebook; popular science channels; and mental health services, such as "Shall We Talk" and "Know Yourself." They sought professional help on the internet when needed and also explored indirect channels, such as religion and travel to self-manage their emotional health.

> *I learn emotional coping skills by watching instructional videos on YouTube, and I also pay attention to some innovative counselling services like Shall We Talk or Know Yourself.* [M 013]

> *I think some YouTube shorts on releasing stress or Facebook groups for anxiety issues could be helpful. But for serious problems I'd still need guidance from an online psychiatrist.* [M 030]

Restricted Internet Use Among the Older Adults

The older adults struggled with adopting digital resources because of unfamiliarity with the internet and the fear of scams, which are prevalent in Hong Kong. They searched for web-based information but preferred offline engagement with activities, such as lectures and social services. They were unfamiliar with platforms, such as Zoom (Zoom Video Communications) and ChatGPT (OpenAI), and complex web-based processes were challenging to them, particularly those with lower digital literacy.

> *I feel that information on the internet is not easy to find. What I've found are just some Youtubers who call themselves therapists, but I don't really trust them.* [O 003]

> *I sometimes use the internet to find a mental health seminar or activity nearby. But I always end up going there in-person to talk to the social workers and psychologists face-to-face.* [O 007]

Assessment of Digital Information in Mental Health

Self-Assessment by the Young Adults

The younger generation prioritized resources that helped them with their mental health issues, rather than focusing on their reliability. They trusted their own judgment over professional advice, as the latter often did not align with their unique circumstances. However, they were open to digital mental health services due to their cost-effectiveness, convenience, and potential to complement offline care.

> *If emotional support services are provided by non-governmental organisations themselves, their credibility is generally higher. But if it is provided by the government, I would have more reservations about using them.* [Y 017]

> *I have read posts on online forums or social media about mental health, but don't fully trust them because none of them are actually verified as being accurate or inaccurate.* [Y 018]

Evidence-Based Assessment by the Middle-Aged Adults

The middle-aged individuals trusted science over government, actively engaged with digital content, and valued professional advice. They filtered the bulk of digital information using education and common sense, assessing reliability with a rational approach and considering factors, such as information quality, expert sources, and whether the information was shared by their peers.

> *When searching for mental health information online, I tend to trust scientific sources. However, there is too much contradictory information online, making it difficult to discern what is true or false. I also do not entirely trust professionals' advice, I want to see the evidence they provide [to back their advice].* [M 030]

The middle-aged individuals often held conservative views on digital mental health services, believing that face-to-face interactions facilitate personal care and emotional connection in a manner that web-based platforms cannot. However, they acknowledged the value of digital services for emergencies or when schedules are too tight for in-person sessions. Notably, only a few of them were willing to invest in expensive web-based care.

> *Online services may allow some people to express themselves more boldly, because the counsellor only hears your voice and knows your name but doesn't know what you look like.* [M 007]

> *During the COVID-19 pandemic, I was unable to go out. Online services were of some help during this exceptional time.* [M 026]

Offline Verification by the Older Adults

The older adults tended to prefer digital mental health services that were simple to navigate, trustworthy, and connected to offline nonprofit platforms. They often valued government, university, and nongovernmental organization services and trusted the firsthand experiences of others with mental health challenges.

> *I go online with a purpose. If there is some information that I need to check, I search for NGOs in places like Tsuen Wan or Kwai Chung, and then contact them for advice.* [O 009]

Web-based counseling, however, was not their preferred means of coping with mental health difficulties. Instead, they sought local resources—such as social workers—to alleviate their anxiety and depression. Those who had received a formal

XSL•FO
RenderX

diagnosis often sought professional support from counselors or psychologists.

> *Not all online information is reliable, and online counselling only allows text responses. So I prefer to consult social workers or [mental health] professionals.* [O 002]

Engagement With Social Media to Manage Emotions

Indirect Emotional Expression by the Young Adults

The young adults were the group that most frequently used social media to express their emotional needs to their digital social circles. They relied on social media and unique methods of expression to indirectly vent their emotions.

> *During my primary and secondary school years, I used to openly express my emotions online.* [Y 012]

Some of them found it immature to share unhappy feelings on social media. Others found solace in expressing those feelings on social media platforms in which they had fewer contacts, as if these platforms were a virtual tree hole (eg, a virtual tree hole in which one can safely reveal their difficult emotions). However, they typically reserved their main social media channels for sharing positive experiences.

> *I share some of my emotions on Facebook, as it is full of unknown people since I deleted some of my friends.* [Y 017]
>
> *I didn't have many followers on Weibo. I used to post emotional things there.* [Y 012]

The interviews indicated, however, that as young people mature, they naturally gravitate toward their immediate social circles. Social media, in turn, becomes a space primarily for observing and connecting with moments in the lives of those in their social circle.

> *I have IG, Facebook, WeChat and so on, to see what my friends are posting, but I rarely post things myself.* [Y 024]

Reluctance to Share on the Internet Among the Middle-Aged Individuals

The middle-aged adult group exhibited a lack of trust in the internet due to concerns about privacy. They prioritized the security of their personal and family information on the internet. Because their negative emotions often stemmed from work or family issues, they typically found relief for those emotions offline. From their perspective, there was no need to share personal matters with their social circles on the internet, as doing so could lead to gossip.

> *I hardly share or repost anything online. I know there's no need to act or voice my anger, even if I feel that something is unfair. I may wonder at the time, "Is this wrong?," but I eventually let it go.* [M 028]

Preference for In-Person Interaction Among the Older Adults

The older adult group preferred to use the internet or mobile apps to connect with friends, but they also valued offline interactions when they were necessary. They tended to favor traditional phone calls over social media for communication. They typically did not share their emotions on the internet, as they perceived providing feedback via the internet to be complex.

> *If I want to talk to my classmates, we use a phone call. We don't write things out in text online.* [O 009]

Coping Strategies

There were no specific coping mechanisms that each group uniquely exhibited. Nevertheless, many of the interviewees, especially the young adults who needed time and space to heal from depression and anxiety, tried to manage their emotions independently. They did so to avoid receiving unsolicited advice from those who did not understand their perspective, as that often resulted in confusion and increased pressure.

> *I believe that my life should be under my own control, and what I post on social media is [in] my own space and I don't care if others read it or not.* [Y 020]
>
> *There are times when I feel certain things cannot be said out loud. The more one tries to express them, the more annoyed people become. I feel that there needs to be a space for rational discussion, where everyone can truly open up and discuss freely.* [Y 027]

The middle-aged participants often relied on partners for emotional support, perhaps because their emotional challenges often came from work and family issues. The effectiveness of this support was dependent on the participants' expectations of their confidant.

> *Sometimes I share with my wife, especially about matters regarding our kids. If something happens, like they throw a tantrum or cry, we may not know how to handle it. There may also be some minor conflicts when my wife and I have different opinions on family matters.* [M 028]

The interviews indicated, interestingly, that as individuals age, their inclination to share their emotional difficulties with friends or colleagues diminishes. This reluctance arises from a fear of burdening others with one's emotional struggles.

> *As I age, I find it more difficult to express my emotions to those around me, because I worry that I'll be a burden to them or that I'll owe them favours I can't repay.* [O 001]
>
> *I've seen some counsellors who just listen to you but don't give advice. They just provide you with certain materials, and it is up to you to determine how to solve the problem on your own.* [O 004]

Discussion

Principal Findings

This study revealed significant variations in eMHL across different age groups. These disparities included differences in access to digital resources, assessment of digital information, and willingness to express emotions on the internet. The disparities highlight the need for multifaceted strategies for

XSL•FO
RenderX

enhancing eMHL that consider generational preferences and trust levels [23,24]. We also discovered that individuals' eMHL influences their behaviors and strategies for managing emotions. This finding underscores the need for and value of interventions to improve eMHL. The advent of digital technology presents increasing opportunities for digital mental health support and facilitates access to resources for mental health [25]. Our findings provide empirical evidence that individuals' attitudes toward using information on mental health on the web and their skills in navigating the internet to find that information affect their access to mental health services. Therefore, it is essential to develop interventions that cater to the specific mental health needs and preferences of different age groups.

Comparisons With Previous Studies

Our findings suggest that young people excel at gathering information from various sources and using multiple resources concurrently. This is consistent with the findings of Tirocchi et al [26], which highlight Generation Z's (eg, those born between the mid-1990s and early 2010s) strong digital awareness and tendency to use several digital tools simultaneously, perhaps due to their extensive experience with the internet and their open-mindedness. Contrarily, middle-aged individuals often look for authoritative information and depend on their own judgment when seeking digital mental health resources. They generally favor a mix of traditional methods and new digital tools, valuing both professional advice and personal experience. Their approach may reflect a pragmatic and balanced method of seeking and evaluating information, including that on mental health. A study in China revealed that internet-based activities, such as reading news, watching videos, and playing games, effectively improved the mental health status of middle-aged individuals [27]. In contrast, older adults depend more on offline channels than digital resources, possibly due to their limited experience with and proficiency in using web-based platforms. Wei et al [28] found that older adults prefer face-to-face interactions and personal ties, which are better conducted through offline channels. Another study indicated that physical and cognitive impairments impede older people's use of digital resources [29]. In summary, older adults are currently less able to fully leverage digital mental health resources. It is therefore essential to improve the eMHL of older adults to enable them to identify and use digital tools and training resources for managing their mental health. This would not only benefit the older adults themselves but also reduce the strain on caregivers and the overall health care system.

This study revealed that younger individuals often rely on their own judgment and the limited information available to them for decision-making. Influenced by social media and peer networks, they prioritize experiential and intuitive knowledge, especially in the use of digital information [30]. This could lead to a preference for personal experiences and opinions over traditional authoritative sources, for example, "making them susceptible to misinformation." To tackle this challenge, promoting critical thinking and digital literacy is crucial. Numerous studies have emphasized how digital literacy impacts students' ability to discern the veracity of digital information. For instance, a study on Generation Z students found that various aspects of digital literacy, particularly those concerning critical thinking and information literacy, influenced their readiness for web-based learning, particularly concerning critical thinking and information literacy [31].

Middle-aged individuals tend to trust information from authoritative and professional sources, often considering the information provider's professional background. This finding is consistent with the results of a previous study [32]. Other research has also shown that middle-aged adults often rely on general practitioners for health information and that health volunteers play a significant role in enhancing their health literacy [32,33]. In contrast, older adults judge an information provider based on factors, such as accessibility, trust, and familiarity with the provider through prior offline consultation support. Some older adults prefer offline means of obtaining information, such as face-to-face medical consultations, due to the personal interactions that they enable and the relationships that they help to establish with health care providers.

Older individuals generally use social media less frequently than younger and middle-aged people to express their emotions. Prior research has also indicated that social media use varies by age [34]. Younger individuals often indirectly express their emotions on social media, preferring to communicate to smaller circles of connections to avoid negative content from being publicly shared [35]. This behavior may be a result of their concern for personal privacy and maintaining a positive image. However, a study of Nigerian adolescents revealed that cultural and socioeconomic factors influenced their expression and control of anger on social media [36], emphasizing the need to foster digital literacy, emotional intelligence, empathy, and respectful digital interactions.

Middle-aged adults usually avoid expressing negative emotions publicly on social media as doing so may lead to personal or professional repercussions. They are more likely to seek support and resolve issues privately instead of publicizing personal emotional struggles, which could affect their family or work. The factors that influence their choices in this regard include privacy concerns, maturity, and the need for support [37,38]. Research has shown that middle-aged and older adults may have more resources and skills to regulate emotions than younger adults, making them less likely to publicly express negative emotions [39].

Older adults rarely use social media to express emotions, primarily due to unfamiliarity with the technology. The "sociodigital divide" within families can result in the social exclusion of older members who are unfamiliar with newer technologies [40], straining intergenerational communication and leading to feelings of alienation. Older adults' hesitation to engage in typical social media practices, such as chatting and sharing selfies, is often because of a perceived mismatch between social media cultures and the values, interests, or needs of older individuals. This mismatch acts as a social barrier to their digital participation and inclusion [41].

Implication of the Study and Future Research Direction

To the best of our knowledge, this is the first study spanning multiple age groups to investigate eMHL and the various

XSL•FO
RenderX

mechanisms that people use to manage their mental health problems through digital resources. Numerous studies have investigated the eHealth literacy of young people. However, this study focuses on the perspectives of middle-aged and older individuals with mental health issues, comparing them to those of young individuals. It explores their knowledge, skills, and ability to manage these issues, identifying areas where the current mental health system could improve to better serve them. Additionally, these findings are important for improving access to clinical mental health care. They provide empirical evidence that can help clinical professionals design effective interventions to address mental health issues across various age groups. In the future, a quantitative study should be conducted to understand the relationships between eMHL and various socioeconomic factors across different age groups.

Limitations

This study has some limitations that future work should address. The main limitation of the study is the use of the self-reported DASS-21 to recruit individuals with mental health issues. Although this is a valid instrument to assess mental health, our recruitment process may not have been as reliable as we did not obtain a formal confirmation of the participants' mental health status via a proof of clinical diagnosis. Another limitation of the study is that we did not interview individuals without mental health problems, who may have been able to provide different perspectives on managing mental health issues.

Conclusions

This study explored differences in eMHL between 3 age groups in the Chinese population. We identified differences in knowledge, skills, and attitudes regarding the use of digital information to manage mental health problems between young, middle-aged, and older individuals. Younger individuals tend to use a range of digital platforms to manage stress or anxiety, whereas middle-aged individuals tend to rely on information from a few reliable digital sources. In contrast, older adults generally prefer offline consultations with the health care providers that they find via the internet. Our findings highlight the need to develop age group–specific strategies to enhance eMHL. This tailored approach could improve individuals' ability to identify and access reliable digital resources, thus promoting the effective management of their mental health.

Acknowledgments

This study was supported by the Hong Kong Health and Medical Research Fund (19200331).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RHX contributed to developing the study concept and design, data analysis and interpretation, software, writing the original draft, and review and editing. LT contributed to data collection, software, and review and editing. LZ contributed to data collection, software, and review and editing. YC, SKC, WAC, and ELW contributed to data collection, visualization, and review and editing. All authors contributed to the paper and approved the submitted version.

Conflicts of Interest

None declared.

References

1. Xiao H, Carney DM, Youn SJ, Janis RA, Castonguay LG, Hayes JA, et al. Are we in crisis? national mental health and treatment trends in college counseling centers. Psychol Serv 2017;14(4):407-415. [doi: 10.1037/ser0000130] [Medline: 29120199]
2. Lipson SK, Lattie EG, Eisenberg D. Increased rates of mental health service utilization by U.S. college students: 10-year population-level trends (2007-2017). Psychiatr Serv 2019;70(1):60-63 [FREE Full text] [doi: 10.1176/appi.ps.201800332] [Medline: 30394183]
3. Pretorius C, Coyle D. Young people's use of digital tools to support their mental health during Covid-19 restrictions. Front Digit Health 2021;3:763876 [FREE Full text] [doi: 10.3389/fdgth.2021.763876] [Medline: 34927133]
4. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072 [FREE Full text] [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
5. Mata-Greve F, Johnson M, Pullmann MD, Friedman EC, Griffith Fillipo I, Comtois KA, et al. Mental health and the perceived usability of digital mental health tools among essential workers and people unemployed due to COVID-19: cross-sectional survey study. JMIR Ment Health 2021;8(8):e28360 [FREE Full text] [doi: 10.2196/28360] [Medline: 34081592]
6. Woerner M, Sams N, Rivera Nales C, Gorstein T, Johnson M, Mosser BA, et al. Generational perspectives on technology's role in mental health care: a survey of adults with lived mental health experience. Front Digit Health 2022;4:840169 [FREE Full text] [doi: 10.3389/fdgth.2022.840169] [Medline: 35224537]

XSL•FO
RenderX

7. Borghouts J, Eikey EV, De Leon C, Schueller SM, Schneider M, Stadnick NA, et al. Understanding the role of support in digital mental health programs with older adults: users' perspective and mixed methods study. JMIR Form Res 2022;6(12):e43192 [FREE Full text] [doi: 10.2196/43192] [Medline: 36512387]
8. Seifert A, Reinwand DA, Schlomann A. Designing and using digital mental health interventions for older adults: being aware of digital inequality. Front Psychiatry 2019;10:568. [doi: 10.3389/fpsyt.2019.00568]
9. Boucher E, Honomichl R, Ward H, Powell T, Stoeckl SE, Parks A. The effects of a digital well-being intervention on older adults: retrospective analysis of real-world user data. JMIR Aging 2022;5(3):e39851 [FREE Full text] [doi: 10.2196/39851] [Medline: 36053569]
10. Milne RA, Puts MTE, Papadakos J, Le LW, Milne VC, Hope AJ, et al. Predictors of high eHealth literacy in primary lung cancer survivors. J Cancer Educ 2015;30(4):685-692. [doi: 10.1007/s13187-014-0744-5] [Medline: 25355524]
11. Kaye L. Older adults, rural living, and the escalating risk of social isolation. Public Policy Aging Rep 2017;27:44. [doi: 10.1093/ppar/prx029]
12. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]
13. Stellefson M, Chaney B, Barry AE, Chavarria E, Tennant B, Walsh-Childers K, et al. Web 2.0 chronic disease self-management for older adults: a systematic review. J Med Internet Res 2013;15(2):e35 [FREE Full text] [doi: 10.2196/jmir.2439] [Medline: 23410671]
14. Gioia F, Rega V, Boursier V. Problematic internet use and emotional dysregulation among young people: a literature review. Clin Neuropsychiatry 2021;18(1):41-54 [FREE Full text] [doi: 10.36131/cnfioritieditore20210104] [Medline: 34909019]
15. Priyana Y. The effect of social media use and environment on mental health among young people in Sukabumi. WSIS 2023;1(03):27-33. [doi: 10.58812/wsis.v1i03.52]
16. Goldfus TB. The impact of social media use on depression, anxiety, and well-being for teens/young people: using hypnosis to build a strong sense of self. Am J Clin Hypn 2024;66(2):140-156. [doi: 10.1080/00029157.2023.2240863] [Medline: 37586041]
17. Lal S. E-mental health: promising advancements in policy, research, and practice. Healthc Manage Forum 2019;32(2):56-62 [FREE Full text] [doi: 10.1177/0840470418818583] [Medline: 30739487]
18. Cormier E, Park H, Schluck G. College students' eMental health literacy and risk of diagnosis with mental health disorders. Healthcare (Basel) 2022;10(12):2406 [FREE Full text] [doi: 10.3390/healthcare10122406] [Medline: 36553930]
19. Cormier E, Park H, Schluck G. eMental health literacy and knowledge of common child mental health disorders among parents of preschoolers. Issues Ment Health Nurs 2020;41(6):540-551. [doi: 10.1080/01612840.2020.1719247] [Medline: 32400237]
20. Gong X, Xie X, Xu R. Psychometric properties of the Chinese versions of DASS-21 in Chinese college students. J Clin Psychol 2010;18:443-446 [FREE Full text]
21. Creswell JW. Qualitative Inquiry and Research Design: Choosing Among Five Traditions. London: SAGE Publications; 1998.
22. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
23. King C, Cianfrone M, Korf-Uzan K. Youth engagement in eMental health literacy. Knowl Manage E-Learn Int J 2015;7:1-12. [doi: 10.34105/j.kmel.2015.07.042]
24. Watkins I, Xie B. eHealth literacy interventions for older adults: a systematic review of the literature. J Med Internet Res 2014;16(11):e225 [FREE Full text] [doi: 10.2196/jmir.3318] [Medline: 25386719]
25. Bevan Jones R, Hussain F, Agha SS, Weavers B, Lucassen M, Merry S, et al. Digital technologies to support adolescents with depression and anxiety: review. BJPsych Adv 2022;29(4):239-253. [doi: 10.1192/bja.2022.3] [Medline: 37521105]
26. Tirocchi S, Scocco M, Crespi I. Generation Z and cyberviolence: between digital platforms use and risk awareness. Int Rev Sociol 2022;32(3):443-462. [doi: 10.1080/03906701.2022.2133408]
27. Fan S, Yang Y. How does internet use improve mental health among middle-aged and elderly people in rural areas in China? a quasi-natural experiment based on the China Health and Retirement Longitudinal Study (CHARLS). Int J Environ Res Public Health 2022;19(20):13332 [FREE Full text] [doi: 10.3390/ijerph192013332] [Medline: 36293913]
28. Wei Y, Yan X, Zhang T. Patients' selection behaviour of healthcare service channels in the context of online and offline integration. 2022 Presented at: Wuhan International Conference on E-Business; June 29, 2024; Wuhan URL: https://tinyurl.com/4y557jsd
29. Guo S, Dang Y, Vogel D, She B. The effect of offline medical resource distribution on online physician-patient interaction: empirical study with online and offline data. JMIR Form Res 2023;7:e43533 [FREE Full text] [doi: 10.2196/43533] [Medline: 36626204]
30. Khulwa CA, Luthfia A. Generation Z students' digital literacy on online learning readiness. 2023 Presented at: International Conference Innovation Engineering and Technology; March 18-20, 2023; Fujisawa, Japan p. 360-364 URL: https://ieeexplore.ieee.org/document/10111186

XSL•FO
RenderX

31. Kelleher MM, Dunngalvin A, Sheikh A, Cullinane C, Fitzsimons J, Hourihane JO. Twenty four-hour helpline access to expert management advice for food-allergy-triggered anaphylaxis in infants, children and young people: a pragmatic, randomized controlled trial. Allergy 2013;68(12):1598-1604. [doi: 10.1111/all.12310] [Medline: 24410783]
32. Williams SL, Ames K, Lawson C. Preferences and trust in traditional and non-traditional sources of health information—a study of middle to older aged Australian adults. J Commun Healthcare 2019;12(2):134-142. [doi: 10.1080/17538068.2019.1642050]
33. Buawangpong N, Sirikul W, Anukhro C, Seesen M, La-Up A, Siviroj P. Health information sources influencing health literacy in different social contexts across age groups in Northern Thailand citizens. Int J Environ Res Public Health 2022;19(10):6051 [FREE Full text] [doi: 10.3390/ijerph19106051] [Medline: 35627585]
34. Wiesbrock S. Age differences with presentation on social media. Mod Psychol Stud 2021;26(2):1-12 [FREE Full text]
35. Chang PF, Choi YH, Bazarova NN, Löckenhoff CE. Age differences in online social networking: extending socioemotional selectivity theory to social network sites. J Broadcast Electron Media 2015;59(2):221-239 [FREE Full text] [doi: 10.1080/08838151.2015.1029126] [Medline: 31223198]
36. Omumu F. A study of social media‘s role in influencing anger expression and management among young people. Zamfara Int J Humanit 2023;2(01):99-116. [doi: 10.36349/zamijoh.2023.v02i01.008]
37. Park C, Russell B, Fendrich M. Emotion regulation resources explain middle-aged and older adults' COVID-19-related distress. Innov Aging 2021;5(Suppl 1):130-131. [doi: 10.1093/geroni/igab046.501]
38. Park C, Kim D, Russell B. Wisdom-related emotion regulation resources explain middle-aged and older adults' COVID-19-related distress. Innov Aging 2023;7:121. [doi: 10.1093/geroni/igad104.0394]
39. Heiden SM, Holden RJ, Alder CA, Bodke K, Boustani M. Human factors in mental healthcare: a work system analysis of a community-based program for older adults with depression and dementia. Appl Ergon 2017;64:27-40 [FREE Full text] [doi: 10.1016/j.apergo.2017.05.002] [Medline: 28610811]
40. Sun Lim S, Ling Tan Y. Old people and new media in wired societies. Media Asia 2016;30(2):95-102. [doi: 10.1080/01296612.2003.11726709]
41. Bossio D, McCosker A. Reluctant selfies: older people, social media sharing and digital inclusion. Continuum 2021;35(4):634-647. [doi: 10.1080/10304312.2021.1937941]

Abbreviations

eMHL: electronic mental health literacy
DASS-21: Depression, Anxiety and Stress Scale-21 Item

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 03.04.24; peer-reviewed by L McCann; comments to author 01.05.24; revised version received 11.05.24; accepted 27.06.24; published 16.07.24.

Please cite as:
Xu RH, Tian L, Zhu L, Cao Y, Chan SKW, Dong D, Cheung WLA, Wong ELY
Age Differences in Electronic Mental Health Literacy: Qualitative Study
J Med Internet Res 2024;26:e59131
URL: https://www.jmir.org/2024/1/e59131
doi:10.2196/59131
PMID:39012686

©Richard Huan Xu, Lidan Tian, Liling Zhu, Yuan Cao, Sherry Kit-wa Chan, Dong Dong, Wai-ling Annie Cheung, Eliza Lai-yi Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Associations Between Sociodemographic Characteristics, eHealth Literacy, and Health-Promoting Lifestyle Among University Students in Taipei: Cross-Sectional Validation Study of the Chinese Version of the eHealth Literacy Scale

Dan-Ping Chao[1], PhD

Department of Tourism and Leisure Management, China University of Technology, Taipei, Taiwan

Corresponding Author:
Dan-Ping Chao, PhD
Department of Tourism and Leisure Management
China University of Technology
3 Xinglong Road, Wunshan District
Taipei, 11695
Taiwan
Phone: 886 2 2931 3416 ext 2264
Email: chaodp39@cute.edu.tw

Abstract

Background: The popularization of the internet and rapid development of mobile devices have led to an increased inclination and opportunities to obtain health-related information online. The eHealth Literacy Scale (eHEALS), widely used for measuring eHealth literacy, assesses an individual's ability to search, understand, appraise, and use eHealth information. However, the Chinese version of the eHEALS multiple-factor model remains to be validated, and the correlation between eHEALS and the health-promoting lifestyle profile (HPLP) among university students is rarely explored in Taiwan.

Objective: This study aimed to examine the fit, validity, and reliability of the Chinese eHEALS multiple-factor model and to clarify the predictive effects of eHEALS on the HPLP among university students.

Methods: University students in Taipei, the capital of Taiwan, were recruited, and 406 valid questionnaires including sociodemographic characteristics, eHEALS, and HPLP responses were collected. Confirmatory factor analysis was performed to validate the Chinese eHEALS. Independent sample *t* test, 1-way ANOVA, and multiple linear regression analyses were conducted to examine the relationship between sociodemographic variables and the HPLP. Pearson product-moment correlation and binary logistic regression analyses were performed to ascertain the predictive effects of eHEALS on the HPLP.

Results: The Chinese eHEALS exhibited an optimal fit when delineated into the search, usage, and evaluation 3-factor model (comparative fit index=0.991, Tucker-Lewis index=0.984, root mean square error of approximation=0.062), and its validity and reliability were confirmed. The mean eHEALS score of university students was 3.17/4.00 (SD 0.48) points, and the score for the evaluation subscale was the lowest (mean 3.08, SD 0.56 points). Furthermore, there were significant sex, institution orientation, daily reading time, daily screen time, primary information channel, and perceived health status differences in the HPLP: male participants (t_{404}=2.346, P=.02), participants attending general university (t_{404}=2.564, P=.01), those reading ≥1 hour daily ($F_{2,403}$=17.618, P<.001), those spending <3 hours on mobile devices or computers daily ($F_{2,403}$=7.148, P<.001), those acquiring information from others (t_{404}=3.892, P<.001), and those with a good perceived health status ($F_{2,403}$=24.366, P<.001) had a significantly higher score. After adjusting for sociodemographic variables, the eHEALS score remained an independent predictor of the HPLP. Compared to students with relatively high eHEALS scores, those with relatively low eHEALS scores had a 3.37 times risk of a negative HPLP (adjusted odds ratio [OR]=3.37, 95% CI 1.49-7.61), which could explain 14.7%-24.4% of the variance (Cox-Snell R^2=0.147, Nagelkerke R^2=0.244, P=.004).

Conclusions: There is room for improvement in eHealth literacy among university students in Taipei. eHEALS may be used to screen students who require HPLP improvement, thereby providing appropriate eHealth literacy training programs, particularly those targeting evaluation literacy. Additionally, the 3-factor model of the Chinese eHEALS used in this study results in more definite scale content, thus increasing the practicality and applicability of this scale in health-promoting studies.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52314) doi:10.2196/52314

KEYWORDS

college students; online health information; eHealth literacy; search literacy; usage literacy; evaluation literacy; health-promoting behavior

Introduction

Background

In recent years, the accessibility and convenience of the internet have increased. This has allowed the public to use it more frequently for communication, education, work, or recreation. A Taiwan Network Information Center (TWNIC) survey revealed that the percentage of individuals with internet access has remained above 80% since 2015 and reached 84.3% in 2022 [1]. Notably, generation Z (aged ≤25 years) had an internet access rate of 100%, and 20.39% of the general population consistently maintained an active online status [1]. Nevertheless, there exists an opportunity for enhancing public digital literacy [1]. Recent studies have highlighted that 81% of adults in the United States possess the ability to search the internet, with 72% using online sources for health-related information [2]. The rapid and discreet nature of the internet considerably increases the public's inclination to use the internet for accessing health-related information [3,4]. However, online health information may contain complex, inaccurate, or even misleading content, resulting in low comprehensibility and reliability [5]. Individuals may inadvertently jeopardize their well-being if they lack the ability to comprehend and critically evaluate online health information [6]. Consequently, the concept of eHealth literacy has gradually attracted attention.

Health literacy refers to the ability of an individual to engage with health information. The World Health Organization defines health literacy as "the cognitive and social skills which determine the motivation and ability of individuals to gain access to, understand and use information in ways which promote and maintain good health" [7,8]. This concept can be further divided into 3 levels, namely basic/functional, communicative/interactive, and critical [9]. Studies have indicated that individuals with high health literacy tend to effectively comprehend medical information and frequently engage in health-promoting behaviors, thereby cultivating a healthier lifestyle [10]. eHealth literacy refers to the aptitude for sourcing, comprehending, assessing, and applying health information from the internet to address health problems. Scholars have used the lily model to delineate the 6 fundamental competencies in eHealth literacy [11]. In addition to the aforementioned health literacy, these 6 competencies in eHealth literacy extend to traditional, information, scientific, media, and computer literacies [11]. eHealth literacy has been positively correlated with health literacy [12]. Studies have highlighted that eHealth literacy may potentially affect the intention and behavior of an individual to use online health information [13]. Individuals with high eHealth literacy tend to actively search and review health information online, leveraging it to enhance their health behaviors and self-manage their health care needs [14].

The eHealth Literacy Scale (eHEALS) was the first lily model–based tool developed for measuring eHealth literacy [15]. Exploratory factor analysis (EFA) confirms that eHEALS consists of a single factor with 8 questions [15]. The scale is designed to be user friendly, demonstrating strong validity, reliability, and stability [15]. In systematic reviews, eHEALS is the most commonly used tool for eHealth literacy evaluation beyond its initial publication [16]. Both the original and translated eHEALS versions are widely used across different countries and populations [16]. However, limitations exist. Studies have highlighted that eHEALS only assesses 1 dimension, rendering it difficult to effectively evaluate diverse eHealth literacy aspects [17]. Another study noted that the rapid spread of social media and mobile devices in recent years could potentially render eHEALS inadequate in completely capturing the contemporary eHealth literacy of individuals [18]. Nonetheless, although EFA is extremely useful for reducing many questions to a manageable amount, only confirmatory factor analysis (CFA) of a multiple-factor model can rigorously evaluate the one-dimensionality of the scale [19]. Recent studies using CFA have found that eHEALS is not a unidimensional concept, and the fit of the 2-factor model is better than that of the 1-factor model but still not good enough [20]. Therefore, researchers have further recommended using a 3-factor model of the original eHEALS as it has a better fit and can effectively measure an individual's current eHealth information skills and comfort [21]. An attempt was made to divide the Chinese eHEALS into 4 factors for discussion; however, some factors only cover 1 question, and the method for dividing the factors and the model fit have not been determined [22]. In summary, as eHEALS has only 8 questions, it is more suitable to divide it into 2 or 3 factors. However, further study is required to determine the fit, validity, and reliability of the Chinese eHEALS with 2-factor or 3-factor models.

Researchers have emphasized that eHealth literacy is not a static trait but evolves in response to changes in individual circumstances, societal dynamics, and environmental factors [11]. Several studies have identified variations in eHealth literacy across different sexes, educational backgrounds, income levels, health status, degree of health concern, and frequency of health-related discussions [3,15,17,18,23,24]. eHealth literacy is considered to be positively correlated with many health behaviors adopted by an individual. Recent studies have highlighted that individuals with high eHealth literacy have positive social relationships, a balanced diet, and safe sex practice [13]. Researchers have found that people with high eHealth literacy exercise and eat breakfast regularly [18]. Studies have also proved that high eHealth literacy predicts balanced eating, regular exercise, and good sleep behaviors [17,23]. Additionally, cross-sectional and longitudinal studies have indicated that people with high eHealth literacy can successfully cultivate a health-promoting lifestyle that includes health

XSL•FO
RenderX

responsibility, exercise, nutrition, self-actualization, stress management, and interpersonal support [4,24,25].

In Taiwan, adults aged 18-29 years, often called digital natives, have grown up with the internet. The TWNIC survey revealed that less than 1% of this demographic has never used the internet [1]. Notably, permanent online engagement is particularly pronounced in this population, inherently amplifying opportunities for internet-based information retrieval and usage [1]. Entering university is an important stage when adolescents transition into adulthood. Studies have shown that when presented with substantial internet-based health information encompassing both accurate and misleading content, university students may encounter challenges in accessing dependable sources and using effective evaluation methods, underscoring the need for continued strengthening of their eHealth literacy [3,18,26]. Moreover, although in good health, university students tend to exhibit risky health behaviors [27,28]. However, limited Taiwanese studies have explored the correlation between eHEALS and the health-promoting lifestyle profile (HPLP) among university students, highlighting an urgent need to ascertain and address their eHealth literacy educational requirement.

Objective

This study aimed to evaluate the fit, validity, and reliability of Chinese eHEALS 2- and 3-factor models. Moreover, the relationship between eHealth literacy and the HPLP among university students was explored. Specifically, this study sought to uncover sociodemographic factors capable of confounding a health-promoting lifestyle among university students and to ascertain the predictive effects of eHealth literacy on adopting a health-promoting lifestyle by excluding the influence of sociodemographic confounders. Finally, the study proposed health education advice that aligns with the current trends, addressing the specific requirements of individuals with lower eHealth literacy who need prompt intervention.

Methods

Study Design and Participants

This cross-sectional quantitative study was conducted among university students in Taipei, the capital of Taiwan. Two rounds of testing (pretest and formal) were performed, and 2-stage sampling was used in both rounds and the subjects were not repeated. The pretest was conducted to determine the reliability of the Chinese versions of eHEALS and the HPLP in the study population and to conduct EFA to extract the 2- and 3-factor models of the Chinese eHEALS. Next, the formal test was conducted to conduct CFA to further determine the fit, validity, and reliability of the Chinese eHEALS multiple-factor models and to perform inferential statistics on the predictive effects of eHEALS on the HPLP.

In the first stage of the pretest, stratification was conducted based on the types of universities in Taipei. The Taiwan Ministry of Education has classified 24 universities in Taipei based on ownership and educational goals into 6 (25%) public general universities, 5 (21%) public vocational colleges, 8 (33%) private general universities, and 5 (21%) private vocational colleges [29]. One school was randomly selected from each stratum for testing. Subsequently, in stage 2, convenience sampling was conducted in the 4 schools. Responses from 205 subjects aged 18-22 years were collected from September to October 2020, resulting in 201 valid questionnaires being completed and returned, with an effective recovery rate of 98%.

In stage 1 of the formal test, the same method was used to divide universities in Taipei into 4 strata, and 9 (38%) of 24 schools were randomly selected from the 24 universities based on proportional stratification, including 2 (22%) public general universities, 2 (22%) public vocational colleges, 3 (33%) private general universities, and 2 (22%) private vocational colleges. Before stage 2, this study required participation from at least 384 respondents. This need was calculated using the following formula for determining the sample size [30]: where χ^2, P, and d are known in the reference and represent the value of the chi-square for 1 degree of freedom at the desired confidence level (3.841), the population proportion (assumed to be .5, as this would provide the maximum sample size), and the degree of accuracy expressed as a proportion (0.05), respectively [30]. N represents the population size. The total number of university students in Taipei in 2021 (N=250,939) according to the Taiwan Ministry of Education [31], was substituted into in the following formula:

Considering the possibility of 5%-7% of the questionnaires being invalid, the number of respondents in the formal test was expected to increase to 410. Additionally, the number of respondents for each school was calculated based on the ratio of the total students in each school to those in Taipei. Subsequently, stage 2 was performed from March to April 2021, with 39-50 university students randomly selected based on their student ID numbers obtained from each school. Invitations to participate in this study were sent using the students' school email addresses. Students who considered themselves physically and mentally stable and who accepted the invitation were included in this study. They completed the questionnaires at their respective schools at the agreed time. If the initially selected students declined to participate, respondents were drawn again as substitutes. Ultimately, 411 formal questionnaires were collected, resulting in 406 valid completed and returned questionnaires, for an effective recovery rate of 98.8%.

Ethical Considerations

This study was reviewed and approved by the Behavioral and Social Science Research Ethics Committee of National Taiwan University (approval number 202004ES028). The ethical principles of the Declaration of Helsinki were adhered to during the entire study. Interviewers provided participants with questionnaires and explained the study's objectives, procedures, benefits, and potential risks. The self-administered questionnaires were anonymously completed by participants in both rounds of testing after providing signed informed consent. Participants received a small gift—a pen worth New Taiwan dollar (NT $) 38 (US $1.2)—upon questionnaire completion as a token of appreciation.

XSL•FO
RenderX

Measurements

The structured questionnaire with closed-ended items encompassed sociodemographic characteristics, eHealth literacy, and health-promoting lifestyle. Sociodemographic variables considered as potential confounders in this study included sex, institution ownership, institution orientation, living status, parental education level, religious affiliation, monthly disposable amount, daily reading time, daily screen time (on mobile devices and computers), primary information channel, and perceived health status. eHealth literacy and health-promoting lifestyle were the study's main predictor and outcome.

The Chinese eHEALS, a translation of the original eHEALS by Norman and Skinner [15] conducted by Cheng et al [22], was used to assess eHealth literacy (Multimedia Appendix 1). As the translation process of the Chinese eHEALS was not mentioned by Cheng et al [22], 2 language teachers from the Center for General Education of the China University of Technology were invited to inspect and confirm the translation accuracy and fluency of the scale. Six experts on health science and health education reviewed the scale and found that the content validity was good (mean item content validity index=1.00, SD 0.02). This scale was initially constructed with a single factor and comprised 8 questions. A 4-point Likert scale was used for scoring in this study, ranging from 1 (strongly disagree) to 4 (strongly agree). A higher mean score reflected better self-perceived online health information skills and comfort. In binary logistic regression, mean scores for eHEALS of 1.00-3.16 and 3.17-4.00 denoted relatively low and high eHealth literacy, respectively, to facilitate subsequent explanation and application. The scale exhibited good reliability with a Cronbach α of .94. Recent studies have pointed out that the German eHEALS 2-factor model and the English eHEALS 3-factor model have a better fit than the original 1-factor model and are more meaningful [20,21]. Therefore, EFA was performed on pretest data to extract 2- and 3-factor structures from the Chinese eHEALS. The 3-factor model included search (3 questions), usage (2 questions), and evaluation (3 questions), with factor loadings ranging from 0.777-0.829, 0.753-0.833, and 0.717-0.840, respectively. The cumulative explained variance was 85.9%. The 2-factor model included search/usage (5 questions) and evaluation (3 questions), with factor loadings ranging from 0.773-0.826 and 0.758-0.853, respectively. The cumulative explained variance was 79.6%. Subsequently, CFA was performed on the formal test data to determine the fit, validity, and reliability of the 1-, 2-, and 3-factor models.

The HPLP, initially developed by Walker et al [32], was translated into Chinese by Huang and Chiou [33] and further adapted for simplification by Wei and Lu [34]. The simplified version of the Chinese HPLP was used in this study to assess a health-promoting lifestyle (Multimedia Appendix 2). The Chinese HPLP has undergone language revision and content validity review [33], is considered to be faithful to the original scale, and does not have excessive additional content [34]. The scale included 6 subscales, namely self-actualization, health responsibility, exercise, nutrition, interpersonal support, and stress management, each containing 4 items. A 5-point Likert scale was used for scoring, ranging from 1 (never) to 5 (always). A higher mean score indicated a more favorable health-promoting lifestyle. In binary logistic regression, mean scores above or equal to the middle (3/5) denoted relatively positive responses [35,36], indicating active adherence to a health-promoting lifestyle to facilitate subsequent explanation and application. The Cronbach α of the pretest data for the scale was .94, demonstrating its robust reliability.

Statistical Analysis

CFA was conducted using SPSS AMOS 28.0 (IBM Corporation). Recent studies have shown that a model exhibits a good fit when $\chi^2/df<3$; the comparative fit index (CFI), Tucker-Lewis index (TLI), and relative fit index (RFI)>0.95; the goodness-of-fit index (GFI), normed fit index (NFI), and incremental fit index (IFI)>0.9; and the root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR)<0.08 [37-40]. In addition, the scale is considered to have convergent validity if the standardization factor loadings of various questions are >0.7 and the average variance extracted (AVE) of various factors are >0.5 [40,41]. The correlation coefficient of 2 factors is lower than the square root of the AVE of various factors and is considered to have discriminant validity [40]. If the Cronbach α and composite reliability (CR) of the various factors are >0.7, this indicates that the reliability is good [40,42].

In this study, SPSS 23.0 was used for other inferential statistics. An independent sample 2-tailed *t* test or 1-way ANOVA combined with Scheffé post hoc analysis was conducted to present the relationship between sociodemographic characteristics and the HPLP. Multiple linear regression was performed to examine the sociodemographic variables that may affect the HPLP. The total HPLP score was used as the dependent variable, while the 12 sociodemographic variables were transformed into 16 dummy variables to serve as independent variables. Sex (reference=female), institution ownership (reference=public), institution orientation (reference=vocational colleges), living status (reference=alone), father's education level (reference=high school or lower), mother's education level (reference=high school or lower), religious affiliation (reference=with), and primary information channel (reference=self-searching) were each transformed into 1 dummy variable, while monthly disposable amount (reference=NT $15,001 [US $462] or more), daily reading time (reference=less than 1 hour), daily screen time (reference=6 hours or more), and perceived health status (reference=good) were each transformed into 2 dummy variables. Stepwise regression analysis used an inclusion criterion of $P<.05$ and an exclusion criterion of $P>.10$. Tolerance>0.1 and variance inflation factor (VIF)<10 were deemed free from collinearity between independent variables. Subsequently, Pearson product-moment correlation analysis was conducted to depict the correlation between eHEALS and the HPLP. Binary logistic regression was performed to determine the predictive effects of relatively low (1.00-3.16) and relatively high (3.17-4.00) eHEALS scores on both positive (3-5) and negative (1.00-2.99) HPLPs, while accounting for relevant sociodemographic factors associated with the HPLP (ie, sex, institution orientation, daily reading time, daily screen time, primary information channel, and perceived health status). A nonsignificant

XSL•FO
RenderX

Hosmer-Lemeshow test indicated a good fit for the logistic regression model. $P<.05$ was considered statistically significant.

Results

Confirmatory Factor Analysis of eHEALS

In this study, CFA was performed to determine the fit of the Chinese eHEALS. The 1-, 2-, and 3-factor models proposed in the study were evaluated for fit (Table 1). Results showed that the fit of the 3-factor model was significantly better than that of the other models, and the diverse indicators satisfied the recommended fit indicators (χ^2/df=2.574, CFI=0.991, TLI=0.984, RFI=0.975, GFI=0.975, NFI=0.985, IFI=0.991, RMSEA=0.062, SRMR=0.018). Furthermore, the fit of the 3-factor model used for empirical data in this study outperformed the 3-factor model proposed by Sudbury-Riley et al [21].

Table 2 shows the factor loadings, AVE, Cronbach α, and CR of various subscales in the eHEALS 3-factor model, which all met the ideal criteria for convergent validity and reliability. In the search subscale of eHEALS, the questions had factor loadings=0.886-0.950, AVE=0.827, Cronbach α=.93, and CR=0.935. In the usage subscale, the questions had factor loadings=0.857-0.893, AVE=0.766, Cronbach α=.87, and CR=0.867. In the evaluation subscale, the questions had factor loadings=0.785-0.900, AVE=0.710, Cronbach α=.88, and CR=0.880. Table 3 shows the correlation coefficients between various eHEALS factors. The correlation coefficients of 2 factors were lower than the square roots of the AVE of various factors, which suggests that the eHEALS 3-factor model has good discriminant validity.

Table 1. Model fit comparison of CFA[a] for the Chinese eHEALS[b] (N=406).

Model	χ^2 (*df*)[c]	χ^2/df	$\Delta\chi^{2d}$	Δdf^d	*P* value[d]	CFI[e]	TLI[f]	RMSEA[g]	SRMR[h]
3-factor structure[i]	43.762 (17)	2.574	__j	—	—	0.991	0.984	0.062	0.018
2-factor structure[k]	136.094 (19)	7.163	92.332	2	<.001	0.959	0.939	0.123	0.036
1-factor structure	333.727 (20)	16.686	289.965	3	<.001	0.889	0.845	0.197	0.062
3-factor structure by Sudbury-Riley et al [21]	118.985 (17)	6.999	—	—	—	0.964	0.941	0.122	0.031

[a]CFA: confirmatory factor analysis.

[b]eHEALS: eHealth Literacy Scale.

[c]All χ^2 of 4 models are statistically significant.

[d]Difference with 3-factor structure proposed in this study.

[e]CFI: comparative fit index.

[f]TLI: Tucker-Lewis index.

[g]RMSEA: root mean square error of approximation.

[h]SRMR: standardized root mean square residual.

[i]Search (3 questions), usage (2 questions), and evaluation (3 questions) factors extracted using exploratory factor analysis (EFA) of pretest data.

[j]Not applicable.

[k]Search/usage (5 questions) and evaluation (3 questions) factors extracted using EFA of pretest data.

XSL•FO
RenderX

Table 2. Reliability and convergent validity for the Chinese eHEALS[a] 3-factor model (N=406).

Subscale and questions	Standardization factor loading[b]	Cronbach α	CR[c]	AVE[d]
Search		.93	0.935	0.827
Q1: I know what health resources are available on the internet.	0.886	—[e]	—	—
Q2: I know where to find helpful health resources on the internet.	0.950	—	—	—
Q3: I know how to find helpful health resources on the internet.	0.890	—	—	—
Usage		.87	0.867	0.766
Q4: I know how to use the internet to answer my health questions.	0.893	—	—	—
Q5: I know how to use the health information I find on the internet to help me.	0.857	—	—	—
Evaluation		.88	0.880	0.710
Q6: I have the skills I need to evaluate the health resources I find on the internet.	0.785	—	—	—
Q7: I can tell high-quality from low-quality health resources on the internet.	0.839	—	—	—
Q8: I feel confident in using information from the internet to make health decisions.	0.900	—	—	—

[a]eHEALS: eHealth Literacy Scale.

[b]All standardization factor loadings were positive and statistically significant.

[c]CR: composite reliability.

[d]AVE: average variance extracted.

[e]Not applicable.

Table 3. Correlation and discriminant validity for the Chinese eHEALS[a] 3-factor model (N=406).

Factors	Search	Usage	Evaluation
Search	0.909[b]	—[c]	—
Usage	0.864	0.875[b]	—
Evaluation	0.792	0.823	0.843[b]

[a]eHEALS: eHealth Literacy Scale.

[b]Square root of average variance extracted (AVE) for each factor.

[c]Not applicable.

Sociodemographic Characteristics

Table 4 shows the sociodemographic variables in this study. In total, 406 students were enrolled in this study. Overall, 252 (62.1%) of the 406 participants were female, 224 (55.2%) lived with family members or friends, and 269 (66.3%) did not have specific religious beliefs. Regarding institution ownership and educational goals, the sample ratio was close to the distribution ratio of various universities in Taipei. More than half of the students were enrolled in private universities (n=226, 55.7%) than in public universities. Furthermore, the ratio of students attending general universities (n=227, 55.9%) was higher than that of students attending vocational colleges. Regarding the parental education level, 237 (58.4%) of the participants had fathers with a university degree or higher, while 252 (62.1%) had mothers with a university education level or higher. Most participants had a monthly disposable amount of NT $10,000 (US $308) or less (n=182, 44.8%). Additionally, a significant proportion of the participants spent less than 1 hour reading per day (n=198, 48.8%), while the majority spent 6 hours or more on mobile devices and computers daily (n=225, 55.4%). The participants indicated that their primary information-acquiring channel was self-searching (n=361, 88.9%), with only a minority (n=45, 11.1%) relying on asking others. Notably, 253 (62.3%) of the participants reported having a good perceived health status.

XSL•FO
RenderX

Table 4. Sociodemographic characteristics of university students (N=406).

Characteristics	Participants, n (%)
Sex	
Male	154 (37.9)
Female	252 (62.1)
Institution ownership	
Public	180 (44.3)
Private	226 (55.7)
Institution orientation	
General university	227 (55.9)
Vocational college	179 (44.1)
Living status	
With others	224 (55.2)
Alone	182 (44.8)
Father's education level	
High school or lower	169 (41.6)
University or higher	237 (58.4)
Mother's education level	
High school or lower	154 (37.9)
University or higher	252 (62.1)
Religious affiliation	
Without	269 (66.3)
With	137 (33.7)
Monthly disposable amount (NT $)[a]	
≤10,000 (≤US $308)	182 (44.8)
10,001-15,000 (US $308-$462)	137 (33.8)
≥15,001 (≥US $462)	87 (21.4)
Daily reading time (hours)	
<1	198 (48.8)
1.0-2.9	159 (39.1)
≥3	49 (12.1)
Daily screen time (hours)	
<3	44 (10.8)
3.0-5.9	137 (33.8)
≥6	225 (55.4)
Primary information channel	
Consulting others	45 (11.1)
Self-searching	361 (88.9)
Perceived health status	
Good	253 (62.3)
Average	134 (33.0)
Poor	19 (4.7)

[a]An exchange rate of NT $1=US $0.03 was used.

XSL•FO
RenderX

Current eHEALS and the HPLP

In eHEALS, the mean total score was 3.17 (SD 0.48). The score of the usage subscale was the highest (mean 3.25, SD 0.50), followed by the search subscale (mean 3.20, SD 0.52) and the evaluation subscale (mean 3.08, SD 0.56). In the HPLP, the mean total score was 3.55 (SD 0.62). The score of the interpersonal support subscale was the highest (mean 3.87, SD 0.70), followed by the self-actualization subscale (mean 3.85, SD 0.74), the stress management subscale (mean 3.74, SD 0.74), the nutrition subscale (mean 3.41, SD 0.79), and the health responsibility subscale (mean 3.26, SD 0.87), with the exercise subscale (mean 3.18, SD 0.90) being the lowest.

Relationship Between Sociodemographic Variables and the HPLP

As shown in Table 5, the total HPLP score showed significant differences between sexes (t_{404}=2.346, P=.02), institution orientation (t_{404}=2.564, P=.01), daily reading time ($F_{2,403}$=17.618, P<.001), daily screen time ($F_{2,403}$=7.148, P<.001), primary information channel (t_{404}=3.892, P<.001), and perceived health status ($F_{2,403}$=24.366, P<.001). Specifically, the HPLP score was higher for male participants (mean 3.65, SD 0.71) than female ones (mean 3.49, SD 0.55). Participants attending general university (mean 3.62, SD 0.59) had a higher HPLP score than those attending vocational college (mean 3.46, SD 0.65). Regarding daily reading time, participants who read for 1.0-2.9 (mean 3.67, SD 0.58) and ≥3 hours (mean 3.87, SD 0.49) had higher HPLP scores than those who read for <1 hour (mean 3.38, SD 0.64). Regarding daily screen time, participants who spent <3 hours (mean 3.70, SD 0.61) had higher HPLP scores than those who spent ≥6 hours (mean 3.45, SD 0.60). Additionally, the HPLP score was higher for participants who acquired information from others (mean 3.89, SD 0.59) than those who acquired information by themselves (mean 3.51, SD 0.62). Participants with a good perceived health status (mean 3.71, SD 0.62) had higher HPLP scores than those with an average (mean 3.32, SD 0.55) or a poor (mean 3.09, SD 0.50) perceived health status.

Stepwise multiple linear regression was performed to analyze sociodemographic variables that affected the HPLP of participants (Table 6). Results showed that sex, institution orientation, daily reading time, primary information channel, and perceived health status are confounders of the overall HPLP. In particular, male participants, participants attending general university, those who read for ≥1 hour daily, those who acquired information from others, and those with a good perceived health status had a better HPLP. Collinearity was absent between the independent variables (tolerance=0.825-0.969, VIF=1.032-1.212), and the factors explained 19.8% of the variance (adjusted R^2=0.198, $F_{7,398}$=15.290, P<.001).

XSL•FO
RenderX

Table 5. Associations between sociodemographic characteristics and the HPLP[a] (N=406).

Characteristics	HPLP, mean (SD)
Sex (*P*=.02[b])	
Male	3.65 (0.71)
Female	3.49 (0.55)
Institution orientation (*P*=.01)[b]	
General university	3.62 (0.59)
Vocational college	3.46 (0.65)
Daily reading time (*P*<.001)[c]	
<1 hour	3.38[d] (0.64)
1.0-2.9 hours	3.67[e] (0.58)
≥3 hours	3.87[e] (0.49)
Daily screen time (*P*<.001)[c]	
<3 hours	3.70[e] (0.61)
3.0-5.9 hours	3.67[d,e] (0.64)
≥6 hours	3.45[d] (0.60)
Primary information channel (*P*<.001)[b]	
Consulting others	3.89 (0.59)
Self-searching	3.51 (0.62)
Perceived health status (*P*<.001)[c]	
Good	3.71[e] (0.62)
Average	3.32[d] (0.55)
Poor	3.09[d] (0.50)

[a]HPLP: health-promoting lifestyle profile.

[b]From an independent sample 2-tailed *t* test for comparing dichotomized variables.

[c]From 1-way ANOVA combined with the Scheffé post hoc test for comparing variables with more than 2 categories.

[d,e]Values with different superscript letters in variables with more than 2 categories indicate significant differences by Scheffé post hoc test.

XSL•FO
RenderX

Table 6. Stepwise multiple linear regression for factors associated with the HPLP[a] (N=406).

Factors	B	β	*P* value	Tolerance	VIF[b]
Sex					
Male	0.159	0.124	.007	0.958	1.044
Institution orientation					
General university	0.126	0.101	.03	0.940	1.064
Daily reading time (hours)					
1.0-2.9	0.249	0.195	<.001	0.905	1.104
≥3	0.319	0.167	<.001	0.825	1.212
Primary information channel					
Consulting others	0.279	0.141	.002	0.932	1.073
Perceived health status					
Average	–0.345	–0.261	<.001	0.969	1.032
Poor	–0.548	–0.186	<.001	0.967	1.034

[a]HPLP: health-promoting lifestyle profile.

[b]VIF: variance inflation factor.

Relationship Between eHEALS and the HPLP

Pearson product-moment correlation was performed to analyze the correlation between eHEALS and the HPLP (Table 7). The overall eHEALS showed a significantly moderate positive correlation with the overall HPLP among participants (r=0.512, P<.001). Furthermore, different eHEALS dimensions showed a significantly low-to-moderate positive correlation with the various HPLP dimensions (r=0.291-0.522, P<.001).

Binary logistic regression was performed to analyze the predictive effects of the overall eHEALS and its various dimensions on the overall HPLP among participants (Table 8). After adjusting for sociodemographic variables, compared with participants with relatively high overall eHEALS scores, those with relatively low eHEALS scores had 3.37 times the risk of a negative HPLP (adjusted odds ratio [OR]=3.37, 95% CI 1.49-7.61). The model exhibited a good fit (Hosmer-Lemeshow χ^2_8=2.128, P=.98), could explain 14.7%-24.4% of the variance (Cox-Snell R^2=0.147, Nagelkerke R^2=0.244), and had an accurate classification rate of 83.3%.

Table 7. Correlation between eHEALS[a] and the HPLP[b] (N=406).

HPLP items	eHEALS							
	Overall scale		Search subscale		Usage subscale		Evaluation subscale	
	r	*P* value	r	*P* value	r	*P* value	r	*P* value
Overall	0.512	<.001	0.442	<.001	0.406	<.001	0.521	<.001
Self-actualization	0.395	<.001	0.375	<.001	0.328	<.001	0.363	<.001
Health responsibility	0.481	<.001	0.410	<.001	0.336	<.001	0.522	<.001
Exercise	0.375	<.001	0.311	<.001	0.291	<.001	0.397	<.001
Nutrition	0.411	<.001	0.359	<.001	0.326	<.001	0.416	<.001
Interpersonal support	0.348	<.001	0.299	<.001	0.306	<.001	0.338	<.001
Stress management	0.397	<.001	0.327	<.001	0.336	<.001	0.406	<.001

[a]eHEALS: eHealth Literacy Scale.

[b]HPLP: health-promoting lifestyle profile.

XSL•FO
RenderX

Table 8. Binary logistic regression for association of eHEALS[a] with the HPLP[b] (N=406).

eHEALS items	HPLP		Unadjusted		Adjusted[c]		
	Positive	Negative	OR[d] (95% CI)	*P* value	OR (95% CI)	*P* value	Model *P* value
Overall scale							
Relatively low	218	62	4.20 (1.94-9.06)	<.001	3.37 (1.49-7.61)	.004	<.001
Relatively high (reference)	118	8	__[e]	—	—	—	—
Search subscale							
Relatively low	216	61	3.77 (1.81-7.85)	<.001	3.38 (1.54-7.42)	.002	<.001
Relatively high (reference)	120	9	—	—	—	—	—
Usage subscale							
Relatively low	216	58	2.69 (1.39-5.20)	.003	2.25 (1.11-4.59)	.025	<.001
Relatively high (reference)	120	12	—	—	—	—	—
Evaluation subscale							
Relatively low	233	63	3.98 (1.76-8.98)	<.001	3.20 (1.35-7.59)	.008	<.001
Relatively high (reference)	103	7	—	—	—	—	—

[a]eHEALS: eHealth Literacy Scale.

[b]HPLP: health-promoting lifestyle profile.

[c]Adjusted for sex, institution orientation, daily reading time, daily screen time, primary information channel, and perceived health status.

[d]OR: odds ratio.

[e]Not applicable.

Compared with participants with relatively high eHEALS search subscale scores, those with relatively low search abilities had 3.38 times the risk of a negative overall HPLP (adjusted OR=3.38, 95% CI 1.54-7.42). The model exhibited a good fit (Hosmer-Lemeshow χ^2_8=3.052, *P*=.93), could explain 14.8%-24.7% of the variance (Cox-Snell R^2=0.148, Nagelkerke R^2=0.247), and had an accurate classification rate of 83.3%. Compared with participants with relatively high eHEALS usage subscale scores, those with relatively low usage abilities had 2.25 times the risk of a negative HPLP (adjusted OR=2.25, 95% CI 1.11-4.59). The model exhibited a good fit (Hosmer-Lemeshow χ^2_8=10.538, *P*=.23), could explain 13.7%-22.8% of the variance (Cox-Snell R^2=0.137, Nagelkerke R^2=0.228), and had an accurate classification rate of 82.8%. Moreover, compared with participants with relatively high eHEALS evaluation subscale scores, those with relatively low evaluation abilities had 3.20 times the risk of a negative HPLP (adjusted OR=3.20, 95% CI 1.35-7.59). The model exhibited a good fit (Hosmer-Lemeshow χ^2_8=2.916, *P*=.94), could explain 14.3%-23.8% of the variance (Cox-Snell R^2=0.143, Nagelkerke R^2=0.238), and had an accurate classification rate of 83.7%.

Further analysis of the prediction results of eHEALS on various dimensions of the HPLP was conducted (Multimedia Appendix 3). Results showed that compared with participants with relatively high overall eHEALS scores, those with relatively low eHEALS scores had 2.74 times the risk of negative health responsibility (adjusted OR=2.74, 95% CI 1.55-4.84), 2.41 times the risk of negative exercise (adjusted OR=2.41, 95% CI 1.43-4.07), and 1.86 times the risk of negative nutrition (adjusted OR=1.86, 95% CI 1.07-3.22).

Compared with participants with relatively high eHEALS subscales scores, those with relatively low search, usage, and evaluation abilities, respectively, had 2.66 (adjusted OR=2.66, 95% CI 1.52-4.62), 2.00 (adjusted OR=2.00, 95% CI 1.18-3.37), and 3.01 (adjusted OR=3.01, 95% CI 1.63-5.55) times the risk of negative health responsibility; 2.02 (adjusted OR=2.02, 95% CI 1.22-3.35), 2.12 (adjusted OR=2.12, 95% CI 1.29-3.50), and 2.71 (adjusted OR=2.71, 95% CI 1.54-4.76) times the risk of negative exercise; and 2.08 (adjusted OR=2.08, 95% CI 1.12-3.86), 1.83 (adjusted OR=1.83, 95% CI 1.08-3.11), and 2.08 (adjusted OR=2.08, 95% CI 1.07-4.06) times the risk of negative nutrition. In addition, compared with participants with relatively high eHEALS evaluation subscale scores, those with relatively low evaluation abilities had 2.06 times the risk of negative stress management (adjusted OR=2.06, 95% CI 1.01-4.22).

Discussion

Principal Findings

Comparison of the Chinese eHEALS 3-Factor Model With Previous Studies

Norman and Skinner [15] developed eHEALS and highlighted that men's eHEALS scores are significantly higher than those of women, which could be used as an a priori hypothesis. Similar results were obtained by using the Chinese eHEALS in this study; in other words, significant differences were observed in eHEALS scores between sexes (t_{404}=2.708, *P*=.007), with males

XSL•FO
RenderX

having higher scores (mean 3.25, SD 0.51) than females (mean 3.12, SD 0.46). This shows that the Chinese eHEALS has known-groups validity. Moreover, in this study, the original 8 eHEALS questions were classified into 3 factors, namely search (questions 1-3), usage (questions 4 and 5), and evaluation (questions 6-8). Compared with the initial 1-factor model [15], CFA showed that the 3-factor model exhibits a better fit and good validity and reliability. The findings were similar to those of a recent study on the Chinese eHEALS multifactorial model [22]; however, this study showed more robust evidence of fit, validity, and reliability. In contrast to Sudbury-Riley et al [21], who used a 3-factor eHEALS model and defined question 3 as "I know how to find helpful health resources and information on the internet" and questions 4 and 5 as the ability to acquire and use health resources and information, this study defined questions 1-3 as the ability to search for health resources on the internet and questions 4 and 5 as the ability to use online health information. Results revealed that differences in the delineation of questions lead to variations in the model fit. Notably, empirical data showed that the fit of the 3-factor model in this study is superior to that of Sudbury-Riley et al's [21] model. This can be attributed to 2 potential explanations. First, EFA was performed in the pretest to delineate the 3 factors, which differed from Sudbury-Riley et al's [21] method, who carefully reviewed and partitioned the factors based on social cognitive and self-efficacy theories. Second, minor differences in participants' perceptions of the translated scale may have contributed to these disparities [43]. In the English eHEALS, questions 3-5 start with "I know how to," which may have caused participants to perceive them as belonging to the same factor [21]. In the Chinese eHEALS, participants tended to consider questions 1-3 as search factors due to words such as "what," "where," and "find," while the word "use" in questions 4 and 5 led participants to classify it as a usage factor. Nonetheless, the 3-factor model used in this study complies with the foundational theories of the eHEALS lily model (ie, social cognitive theory and self-efficacy theory) [11,21]. This model may be more suitable for regions where the Chinese eHEALS is used in eHealth literacy studies.

eHealth Literacy Level

In this study, the overall eHEALS score of the university students was moderate or higher, and the search and usage dimensions had higher scores. In contrast, the evaluation dimension had a lower score. This reveals that students perceive themselves to have good search and usage capabilities of eHealth information; however, they possess low confidence in evaluating such information and using it for decision-making. In recent studies, the mean scores for eHEALS questions 6-8 were lower than those for questions 1-3 and questions 4 and 5 [3,21,44,45], similar to scores obtained in this study. A Taiwanese study used a self-formulated scale to evaluate the eHealth literacy of university students and divided the questions into functional, interactive, and critical literacies [17]. Interactive literacy encompasses the ability to select, comprehend, and use online health information, which was similar to the search and usage dimensions in this study. Critical literacy refers to the ability to analyze, criticize, and respond to online health information, which was similar to the evaluation dimension in this study. The score for critical literacy was visibly lower than that for interactive literacy in the previous study [17], which was similar to this study. Researchers found that although most university students mentioned that they can understand the general idea of online health information, they have a vague understanding of the jargon, foreign languages, and data [6]. In addition, some university students lack confidence in the quality of online health information and express difficulty in determining the quality of such information [3]. Therefore, in the contemporary landscape characterized by the unlimited accumulation and dissemination of internet-based health information of uncertain veracity, imparting fundamental health knowledge to Taiwanese university students is imperative. This includes fostering a sense of caution toward eHealth information among students and equipping them with the ability to critically assess and validate uncertainties.

Association Between Sociodemographic Variables and the HPLP

The overall HPLP of university students in this study was moderate or higher, wherein interpersonal support and self-actualization scores were the highest, while nutrition, exercise, and health responsibility scores were the lowest, similar to those of the most recent studies [46-49]. Among sociodemographic variables, stepwise multiple linear regression showed that female students, students attending vocational colleges, those with a daily reading time of <1 hour, those who acquired information by themselves, and those with an average or a poor perceived health status were confounders of a poor overall HPLP. This was consistent with the significant differences in the overall HPLP in these sociodemographic variables. Recent studies have found that sex affects the HPLP and health behaviors, such as exercise and sleep [17,48,49]. The frequency of discussions of health problems with others has been highlighted to positively affect the dietary behavior of university students [23]. Individuals with a good perceived health status or great concern for health have a better HPLP and show several health behaviors, such as eating, exercise, and sleep [17,23,24,47,49].

In addition, this study found that a daily reading time of ≥1 hour is a confounder of a good HPLP among university students. This may be because information in books, newspapers, and magazines usually undergoes review and proofreading, and reading more accurate and reliable hardcopy information may lead to a tendency to adopt a positive lifestyle profile. Some studies have highlighted that reading hardcopy materials can promote better comprehension results than reading from screens [50,51]. However, the increased screen time on digital media today has greatly decreased the reading time in print. In this study, only 208 (51.2%) of 406 university students read for ≥1 hour per day, but 362 (89.2%) spent ≥3 hours on mobile devices or computers daily. Recent studies have shown that newspapers and magazines are the media that Taiwanese university students spend the least time on, far below the time spent on mobile devices and computers [22]. Furthermore, mobile devices and computers have diverse online functions. The TWNIC survey found that the most commonly used internet functions among generation Z are real-time messaging, social networks, free videos, online news, online games, and ecommerce; however,

XSL•FO
RenderX

online learning is not their priority [1]. In the ANOVA in this study, the HPLP score of participants with a daily screen time of ≥6 hours was significantly low. However, multiple linear regression excluded the daily screen time from the HPLP confounders. It is believed that frequent usage of mobile devices and computers by university students consumes the time spent on reading. Therefore, the effect of screen time in the multiple linear regression may be explained by the reading time factor. In the binary logistic regression in this study, daily screen time was still considered a confounder of HPLP scores.

Association of eHEALS With the HPLP

After adjusting for sociodemographic factors that may affect the HPLP, this study revealed that eHEALS consistently and significantly affects the HPLP of university students. Compared with students with relatively high overall eHEALS scores, those with relatively low eHEALS scores had a higher probability of a negative overall HPLP, similar to the results of studies in other regions [4,25]. Other researchers have used their own created scales to measure eHealth literacy and proved that it predicts multiple HPLP dimensions in university students [24]. Many studies have found that eHealth literacy has positive effects on exercise, diet, and sleep behaviors [13,17,18,23], or even safe sex practice [13] and COVID-19 prevention [46] among university students. This study found that among the various HPLP dimensions, compared with a relatively high overall eHEALS score, a relatively low eHEALS score is associated with negative health responsibility, exercise, and nutrition. University students in this study had low HPLP health responsibility, exercise, and nutrition scores—dimensions that require improvements. At the same time, many recent studies have found that these health behaviors are poor in university students [46-49]. Additionally, among the 3 eHEALS dimensions in this study, compared with participants with relatively high search, usage, and evaluation literacies, those with relatively low scores had a higher probability of a negative overall HPLP and its health responsibility, exercise, and nutrition dimensions, similar to the overall eHEALS results. Regarding evaluation literacy, this study found that in addition to predicting the negative health responsibility, exercise, and nutrition dimensions of the HPLP, a relatively low eHEALS evaluation score can also reflect poor stress management. This result is similar to that of a recent study indicating that critical eHealth literacy can predict more HPLP dimensions [24].

Limitations and Strengths

This study has certain limitations, which can provide a reference for future studies. First, this study included only university students without major diseases from the capital of Taiwan, and the results can only be generalized to the eHealth literacy and health-promoting lifestyle of this population. It is recommended that future studies extend to other regions in Taiwan or university students with other health statuses. Second, some variables may be related to eHealth literacy and a health-promoting lifestyle, such as majors and health risk behaviors, and it is recommended that future studies expand to include these variables. In addition, although participants were advised that the entire process was anonymized, the self-administered questionnaire may have caused their answers to be exposed to memory recall errors, environmental effects, and social desirability bias. Lastly, a cross-sectional study design was used in this study, and the causal relationship between eHealth literacy and a health-promoting lifestyle, as well as changes in these 2 factors with time, could not be confirmed. Hence, further repeated-measures or longitudinal studies are required for clarification.

Nonetheless, this study confirmed the feasibility of using the Chinese version of the eHEALS 3-factor model to examine eHealth literacy and highlighted that eHealth literacy affects and predicts the HPLP in university students. In the contemporary world where internet use is widespread and portable mobile devices are rapidly advancing, using the internet, mobile phones, tablets, or computers as aids in daily life has become an unstoppable trend. If university students can cultivate the online learning habit early on and establish the concept of consulting to acquire information and reading in print, actively nurturing their skills to search, use, and access internet-based health information, it will undoubtedly positively impact their health-promoting behavior and lifestyle.

Conclusion

This study is the first to validate the Chinese eHEALS 3-factor model, encompassing search, usage, and evaluation dimensions. Notably, eHEALS is the first eHealth literacy measurement tool to be developed and is the most widely used. This 3-factor model results in more definite eHEALS content and undoubtedly increases the practicality and applicability of the scale to satisfy the eHealth literacy evaluation needs of health promoting–related studies, particularly in Chinese-speaking regions.

Higher education represents the most significant and final opportunity for behavioral development and learning in young people. Behavioral health during this period impacts lifetime health outcomes. This study found that alongside specific sociodemographic characteristics, the overall eHEALS and its dimensions are independent predictors of the HPLP. Compared to university students with relatively high overall eHEALS and various dimension scores, those with relatively low scores had a negative overall HPLP and HPLP health responsibility, exercise, and nutrition. University students with relatively low eHEALS evaluation scores compared to those with relatively high evaluation scores also had negative stress management. These findings can be used to screen university students who require HPLP improvement so that health education suitable for their needs can be provided.

In addition, there is room for improving overall eHEALS scores among university students, with particular attention to improving evaluation literacy. It is recommended that the centers for general education, digital learning, and health of the universities and colleges in Taipei, as well as targeting populations with relatively low eHealth literacy (eHEALS score<3.17), be integrated to provide appropriate health education and programs. Courses should be conducted to educate students on identifying objective, credible, and understandable online health information platforms, while cultivating vigilance and critical judgment in evaluating eHealth information. Additionally, fostering a supportive and user-friendly online health information

XSL•FO
RenderX

environment is essential. It is recommended that universities and colleges further establish good campus eHealth literacy learning and support channels. For example, good health information online platforms could be recommended on school websites, and in-person or virtual health information consultation could be applied within schools. These measures would collectively contribute to improving university students' eHealth literacy, thereby encouraging their adoption of health-promoting lifestyles.

Acknowledgments

The author would like to thank Prof Silvia Wen-Yu Lee and Prof Shyh-Hsiang Lin for providing consultation and answering questions during the research process. The author would also like to thank Prof Yuan-Chin Chang and Prof Yi-Jia Lin for inspecting and confirming the translation accuracy and fluency of the Chinese eHealth Literacy Scale (eHEALS). In addition, the author thanks 6 experts and scholars, namely Dr Wen-Ting Huang, Dr Wei-Chen Li, Dr Chin-Yu Tsai, Dr Fang-Min Liu, Dr Chuan-Chun Stella Kuo, and Mr Hsueh-Wei Chen, for reviewing the content validity of eHEALS. The author also thanks all participating university students and 4 interviewers, namely Zi-Yu Peng, Tzu-Yin Lai, Ching-Sin Cheng, and You-Hsuan Yu, for their contribution to the study. Moreover, the author would like to thank Editage for English language editing. The paper-processing fee of publication was funded by PROZ Biotech CO, LTD. The study was not supported by other external funding.

Data Availability

The data sets generated and analyzed during this study are not publicly available but are available from the corresponding author upon reasonable request.

Authors' Contributions

D-PC conceptualized and administered the study, analyzed and interpreted the results, and drafted and revised the manuscript. D-PC also reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The Chinese version of eHEALS. eHEALS: eHealth Literacy Scale.
[DOCX File , 17 KB - jmir_v26i1e52314_app1.docx]

Multimedia Appendix 2
The simplified version of Chinese HPLP. HPLP: health-promoting lifestyle profile.
[DOCX File , 20 KB - jmir_v26i1e52314_app2.docx]

Multimedia Appendix 3
Binary logistic regression for association of eHEALS with the HPLP dimensions (N=406). eHEALS: eHealth Literacy Scale; HPLP: health-promoting lifestyle profile.
[DOCX File , 21 KB - jmir_v26i1e52314_app3.docx]

References

1. 2022 Taiwan internet report. Taiwan Network Information Center. 2022 Jul. URL: https://report.twnic.tw/2022/assets/download/TWNIC_TaiwanInternetReport_2022_CH.pdf [accessed 2023-07-26]
2. Fox S, Duggan M. Health online 2013. Pew Research Center. 2013 Jan 15. URL: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ [accessed 2023-07-26]
3. Dashti S, Peyman N, Tajfard M, Esmaeeli H. E-health literacy of medical and health sciences university students in Mashhad, Iran in 2016: a pilot study. Electron Physician 2017 Mar;9(3):3966-3973 [FREE Full text] [doi: 10.19082/3966] [Medline: 28461871]
4. Yoğurtcu H, Ozturk Haney M. The relationship between e-health literacy and health-promoting behaviors of Turkish hospital nurses. Glob Health Promot 2022 Jun 21;1:17579759221093389. [doi: 10.1177/17579759221093389] [Medline: 35730122]
5. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002 Mar 09;324(7337):573-577 [FREE Full text] [doi: 10.1136/bmj.324.7337.573] [Medline: 11884321]

XSL•FO
RenderX

6. Hsu WC, Chen SF, Ho CJ. Experience of using web health information among college students: an analysis from the health literacy perspective. J Health Promot Health Educ 2011;35:1-22. [doi: 10.7022/JHPHE.201106.0001]
7. Health promotion glossary. World Health Organization. 1998 Jan. URL: http://apps.who.int/iris/bitstream/10665/64546/1/WHO_HPR_HEP_98.1.pdf [accessed 2023-07-26]
8. A practical guidebook for health literate organization. Ministry of Health and Welfare, Taiwan. 2018 Dec. URL: https://www.hpa.gov.tw/Pages/ashx/File.ashx?FilePath=~/File/Attach/10167/File_9540.pdf [accessed 2023-07-26]
9. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
10. Yokokawa H, Yuasa M, Sanada H, Hisaoka T, Fukuda H. Age- and sex-specific impact of health literacy on healthy lifestyle characteristics among Japanese residents in a rural community. Health 2015;7(6):679-688. [doi: 10.4236/health.2015.76081]
11. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
12. Nakayama K, Osaka W, Togari T, Ishikawa H, Yonekura Y, Sekido A, et al. Comprehensive health literacy in Japan is lower than in Europe: a validated Japanese-language assessment of health literacy. BMC Public Health 2015 May 23;15(1):505 [FREE Full text] [doi: 10.1186/s12889-015-1835-x] [Medline: 26001385]
13. Britt RK, Collins WB, Wilson K, Linnemeier G, Englebert AM. eHealth literacy and health behaviors affecting modern college students: a pilot study of issues identified by the American College Health Association. J Med Internet Res 2017 Dec 19;19(12):e392 [FREE Full text] [doi: 10.2196/jmir.3100] [Medline: 29258979]
14. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
15. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
16. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowl Manag E-Learn: Int J 2015;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
17. Chiang CH, Yang SC, Hsu WC. Development and validation of the e-health literacy scale and investigation of the relationships between e-health literacy and healthy behavior among undergraduate students in Taiwan. Formosa J Mental Health 2015;28(3):389-420. [doi: 10.30074/FJMH.201509_28(3).0002]
18. Tsukahara S, Yamaguchi S, Igarashi F, Uruma R, Ikuina N, Iwakura K, et al. Association of eHealth literacy with lifestyle behaviors in university students: questionnaire-based cross-sectional study. J Med Internet Res 2020 Jun 24;22(6):e18155 [FREE Full text] [doi: 10.2196/18155] [Medline: 32579126]
19. Gerbing DW, Anderson JC. An updated paradigm for scale development incorporating unidimensionality and its assessment. J Mark Res 1988 May;25(2):186-192. [doi: 10.2307/3172650]
20. Soellner R, Huber S, Reder M. The concept of eHealth literacy and its measurement. J Media Psychol 2014 Jan;26(1):29-38. [doi: 10.1027/1864-1105/a000104]
21. Sudbury-Riley L, FitzPatrick M, Schulz PJ. Exploring the measurement properties of the eHealth Literacy Scale (eHEALS) among baby boomers: a multinational test of measurement invariance. J Med Internet Res 2017 Feb 27;19(2):e53 [FREE Full text] [doi: 10.2196/jmir.5998] [Medline: 28242590]
22. Cheng SY, Chang FC, Li JM. eHealth literacy and related factors among junior high school students in Taipei City. J Health Promot Health Educ 2014;41:1-24. [doi: 10.3966/207010632014060041001]
23. Hsu W, Chiang C, Yang S. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014 Dec 12;16(12):e287 [FREE Full text] [doi: 10.2196/jmir.3542] [Medline: 25499086]
24. Yang SC, Luo YF, Chiang CH. The associations among individual factors, eHealth literacy, and health-promoting lifestyles among college students. J Med Internet Res 2017 Jan 10;19(1):e15 [FREE Full text] [doi: 10.2196/jmir.5964] [Medline: 28073739]
25. Li S, Cui G, Zhou F, Liu S, Guo Y, Yin Y, et al. The longitudinal relationship between eHealth literacy, health-promoting lifestyles, and health-related quality of life among college students: a cross-lagged analysis. Front Public Health 2022;10:868279 [FREE Full text] [doi: 10.3389/fpubh.2022.868279] [Medline: 35875048]
26. Stellefson M, Hanik B, Chaney B, Chaney D, Tennant B, Chavarria EA. eHealth literacy among college students: a systematic review with implications for eHealth education. J Med Internet Res 2011 Dec 01;13(4):e102 [FREE Full text] [doi: 10.2196/jmir.1703] [Medline: 22155629]
27. Jang HJ. Comparative study of health promoting lifestyle and subjective happiness on nursing students and non-nursing students. Adv Sci Technol Lett 2016;128:78-82. [doi: 10.14257/astl.2016.128.16]
28. Musaiger AO, Awadhalla MS, Al-Mannai M, AlSawad M, Asokan GV. Dietary habits and sedentary behaviors among health science university students in Bahrain. Int J Adolesc Med Health 2017 Apr 01;29(2):20150038. [doi: 10.1515/ijamh-2015-0038] [Medline: 26251982]
29. List of colleges and universities in 2020. Ministry of Education, Taiwan. 2020 Oct 15. URL: https://ulist.moe.gov.tw/Browse/UniversityList [accessed 2021-11-19]

XSL•FO
RenderX

30. Krejcie RV, Morgan DW. Determining sample size for research activities. Educ Psychol Measur 1970;30(3):607-610. [doi: 10.1177/001316447003000308]
31. Number of students in schools at all levels - by county and city. Ministry of Education, Taiwan. 2020 Oct 15. URL: https://depart.moe.edu.tw/ED4500/cp.aspx?n=1AC243AF6EF5E5DD&s=EDC4A4E717ED32CF [accessed 2021-11-19]
32. Walker SN, Sechrist KR, Pender NJ. The health-promoting lifestyle profile: development and psychometric characteristics. Nurs Res 1987;36(2):76-81. [Medline: 3644262]
33. Huang YH, Chiou CJ. Predictors contributing to health-promoting lifestyles among college students in Kaohsiung area. Chin J Public Health 1997;16(1):24-36. [doi: 10.6288/CJPH1997-16-01-03]
34. Wei MH, Lu CM. Development of the short-form Chinese health-promoting lifestyle profile. J Health Educ 2005;24:25-45. [doi: 10.7022/JHE.200512.0025]
35. Al-Kandari F, Vidal VL, Thomas D. Health-promoting lifestyle and body mass index among College of Nursing students in Kuwait: a correlational study. Nurs Health Sci 2008 Mar;10(1):43-50. [doi: 10.1111/j.1442-2018.2007.00370.x] [Medline: 18257831]
36. Lolokote S, Hidru TH, Li X. Do socio-cultural factors influence college students' self-rated health status and health-promoting lifestyles? A cross-sectional multicenter study in Dalian, China. BMC Public Health 2017 May 19;17(1):478 [FREE Full text] [doi: 10.1186/s12889-017-4411-8] [Medline: 28525995]
37. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull 1980 Nov;88(3):588-606. [doi: 10.1037/0033-2909.88.3.588]
38. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
39. McDonald RP, Ho MR. Principles and practice in reporting structural equation analyses. Psychol Methods 2002 Mar;7(1):64-82 [FREE Full text] [doi: 10.1037/1082-989x.7.1.64] [Medline: 11928891]
40. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis: A Global Perspective. 7th Edition. Upper Saddle River, NJ: Pearson Prentice Hall; 2009.
41. Ping RA. On assuring valid measures for theoretical models using survey data. J Bus Res 2004 Feb;57(2):125-141. [doi: 10.1016/S0148-2963(01)00297-1]
42. Nunnally JC. Psychometric Theory. 2nd Edition. New York: McGraw-Hill series in psychology; 1978.
43. Rode N. Translation of measurement instruments and their reliability: an example of job-related affective well-being scale. Metodološki zvezki 2005 Jan 01;2(1):15-26. [doi: 10.51936/vyfe8385]
44. Chung S, Park BK, Nahm ES. The Korean eHealth Literacy Scale (K-eHEALS): reliability and validity testing in younger adults recruited online. J Med Internet Res 2018 Apr 20;20(4):e138 [FREE Full text] [doi: 10.2196/jmir.8759] [Medline: 29678800]
45. Del Giudice P, Bravo G, Poletto M, De Odorico A, Conte A, Brunelli L, et al. Correlation between eHealth literacy and health literacy using the eHealth Literacy Scale and real-life experiences in the health sector as a proxy measure of functional health literacy: cross-sectional web-based survey. J Med Internet Res 2018 Oct 31;20(10):e281 [FREE Full text] [doi: 10.2196/jmir.9401] [Medline: 30381283]
46. Kim KA, Hyun MS, De Gagne JC, Ahn JA. A cross-sectional study of nursing students' eHealth literacy and COVID-19 preventive behaviours. Nurs Open 2023 Feb;10(2):544-551 [FREE Full text] [doi: 10.1002/nop2.1320] [Medline: 36631729]
47. Chu-Ko F, Chong ML, Chung CJ, Chang CC, Liu HY, Huang LC. Exploring the factors related to adolescent health literacy, health-promoting lifestyle profile, and health status. BMC Public Health 2021 Dec 01;21(1):2196 [FREE Full text] [doi: 10.1186/s12889-021-12239-w] [Medline: 34852793]
48. Núñez-Rocha GM, López-Botello CK, Salinas-Martínez AM, Arroyo-Acevedo HV, Martínez-Villarreal RT, Ávila-Ortiz MN. Lifestyle, quality of life, and health promotion needs in Mexican university students: important differences by sex and academic discipline. IJERPH 2020 Oct 31;17(21):8024. [doi: 10.3390/ijerph17218024]
49. Chao DP. Health-promoting lifestyle and its predictors among health-related and non-health-related university students in Taiwan: a cross-sectional quantitative study. BMC Public Health 2023 May 05;23(1):827 [FREE Full text] [doi: 10.1186/s12889-023-15760-2] [Medline: 37147650]
50. Mangen A, Walgermo BR, Brønnick K. Reading linear texts on paper versus computer screen: effects on reading comprehension. Int J Educ Res 2013 Jan;58:61-68. [doi: 10.1016/j.ijer.2012.12.002]
51. Chen G, Cheng W, Chang TW, Zheng X, Huang R. A comparison of reading comprehension across paper, computer screens, and tablets: does tablet familiarity matter? J Comput Educ 2014 Aug 6;1(2-3):213-225. [doi: 10.1007/s40692-014-0012-z]

Abbreviations

AVE: average variance extracted
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
EFA: exploratory factor analysis

XSL•FO
RenderX

eHEALS: eHealth Literacy Scale
GFI: goodness-of-fit index
HPLP: health-promoting lifestyle profile
IFI: incremental fit index
NFI: normed fit index
OR: odds ratio
RFI: relative fit index
RMSEA: root mean square error of approximation
SRMR: standardized root mean square residual
TLI: Tucker-Lewis index
TWNIC: Taiwan Network Information Center
VIF: variance inflation factor

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 30.08.23; peer-reviewed by R Moser; comments to author 03.11.23; revised version received 21.11.23; accepted 28.06.24; published 18.07.24.

Please cite as:
Chao DP
Associations Between Sociodemographic Characteristics, eHealth Literacy, and Health-Promoting Lifestyle Among University Students in Taipei: Cross-Sectional Validation Study of the Chinese Version of the eHealth Literacy Scale
J Med Internet Res 2024;26:e52314
URL: https://www.jmir.org/2024/1/e52314
doi:10.2196/52314
PMID:39024006

©Dan-Ping Chao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing an Educational Resource Aimed at Improving Adolescent Digital Health Literacy: Using Co-Design as Research Methodology

Callum C Lewis[1], BClinSc, MD; Melody Taba[1,2], BSc(Hons), MPH; Tiffany B Allen[1], BSc(Hons); Patrina HY Caldwell[1,3], BMed, PhD; S Rachel Skinner[1,3], MBBS, PhD; Melissa Kang[4], MBBS, MCH, PhD; Hamish Henderson[5], BScIT(Hons); Liam Bray[5], BA; Madeleine Borthwick[5], MSc(Hons); Philippa Collin[6], BA(Hons), PhD; Kirsten McCaffery[2], BSc(Hons), PhD; Karen M Scott[1], BEd, MA, PhD

[1]Specialty of Child and Adolescent Health, Sydney Medical School, Faculty of Medicine and Health, The University of Sydney, Westmead, Australia
[2]Sydney Health Literacy Lab, School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[3]The Children's Hospital at Westmead, Westmead, Australia
[4]General Practice Clinical School, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[5]Sydney School of Architecture, Design and Planning, The University of Sydney, Sydney, Australia
[6]Institute for Culture and Society, Western Sydney University, Parramatta, Australia

Corresponding Author:
Karen M Scott, BEd, MA, PhD
Specialty of Child and Adolescent Health, Sydney Medical School, Faculty of Medicine and Health
The University of Sydney
The Children's Hospital at Westmead, Clinical School
Locked Bag 4001
Westmead, 2145
Australia
Phone: 61 298453385
Fax: 61 298453389
Email: karen.scott@sydney.edu.au

Abstract

Background: Adolescence is a key developmental period that affects lifelong health and is impacted by adolescents regularly engaging with digital health information. Adolescents need digital health literacy (DHL) to effectively evaluate the quality and credibility of such information, and to navigate an increasingly complex digital health environment. Few educational resources exist to improve DHL, and few have involved adolescents during design. The co-design approach may hold utility through developing interventions with participants as design partners.

Objective: This project aimed to explore the co-design approach in developing an educational resource to improve adolescents' DHL.

Methods: Adolescents (12-17 years old) attended 4 interactive co-design workshops (June 2021-April 2022). Participant perspectives were gathered on DHL and the design of educational resources to improve it. Data generated were analyzed through content analysis to inform educational resource development.

Results: In total, 27 participants from diverse backgrounds attended the workshops. Insight was gained into participants' relationship with digital health information, including acceptance of its benefits and relevance, coupled with awareness of misinformation issues, revealing areas of DHL need. Participants provided suggestions for educational resource development that incorporated the most useful aspects of digital formats to develop skills across these domains. The following 4 themes were derived from participant perspectives: ease of access to digital health information, personal and social factors that impacted use, impacts of the plethora of digital information, and anonymity offered by digital sources. Initial participant evaluation of the developed educational resource was largely positive, including useful suggestions for improvement.

Conclusions: Co-design elicited and translated authentic adolescent perspectives and design ideas into a functional educational resource. Insight into adolescents' DHL needs generated targeted educational resource content, with engaging formats, designs,

XSL•FO
RenderX

and storylines. Co-design holds promise as an important and empowering tool for developing interventions to improve adolescents' DHL.

(J Med Internet Res 2024;26:e49453) doi:10.2196/49453

KEYWORDS

Adolescent health; digital health literacy; adolescents; online health information; co-design; health education; eHealth literacy; social media

Introduction

Adolescence is a foundational developmental period, influencing lifelong well-being through the establishment of lifestyle and behavioral choices [1]. Adolescents may be exposed to many changes and potential risks, including first encounters with sexual behaviors, substance use, and mental illness, and are, hence, a priority group for health promotion [2]. The digital age has transformed adolescence, representing a new landscape for lived experience and challenges in navigating health concerns [3]. Adolescents frequently seek digital health information [4,5], with a vast majority engaging with it on websites and social media [6], often using mobile devices [7].

Digital formats provide a private space for health information seeking, which is particularly beneficial for sensitive health domains, including sexual health [8]. The quantity of digital health information available [9] alleviates information poverty, which was previously encountered by adolescents wanting to access targeted health information [10]. The currency of information [9], alongside the ability to share personal perspectives [11], further promotes uptake.

Despite the frequent use of digital health information by adolescents, awareness of its highly variable quality has engendered a lack of trust [12]. Heightened risks to adolescent health exist due to insufficient regulation of digital health information [13], particularly given poor quality information is frequently disseminated, often driven by commercial interests [14]. This has been accentuated in the current era where misinformation, which adolescents often struggle to identify [15], is frequently published on the internet, including social media, as exemplified by the COVID-19 "infodemic" [16].

Digital health information can impact adolescents' engagement with health care providers, including decisions about how and when to attend health facilities [17,18], or decisions to self-manage care [19]. Risks due to poor quality information and delayed health care presentations can potentially cause harm [20]. Conversely, opportunity exists to support health engagement if information is accurate and appropriate [9,21]. Given the importance of digital health information, adolescents' ability to assess its quality needs to be optimized [22].

Digital health literacy (DHL), "the ability to find, evaluate, appraise, integrate, and apply health information from online environments" [23] denotes the multifaceted skills required to assess credibility of digital health information [24]. Adolescents often lack these important digital health literacy skills [25]. In particular, adolescents frequently use ineffective heuristics [26], rather than objective criteria [27], to appraise digital health information, often prioritizing relevance and accessibility over credibility [28,29].

Social cognitive theory (SCT) offers a useful lens through which to analyze DHL. It describes how learning and behavior is influenced by a dynamic interrelationship between a person, their behavior, and their surrounding multifactorial context [30]. Self-efficacy, the notion of individuals trusting their own ability, is a core SCT concept that is understood to be key for effective learning [31], including for building and sustaining DHL skills [32]. However, evidence suggests adolescents frequently overestimate their abilities, perpetuating the impacts of misinformation encountered [27-29]. Education to improve DHL is, therefore, vital, given that higher DHL is associated with positive health behaviors [33], health care engagement, and overall well-being [34,35].

While adolescents' DHL deficits are increasingly acknowledged, there are few resources that target its improvement, and the literature on resource development is limited. A 2012 trial of a digital classroom resource, which featured instructional DHL content around nutrition and exercise, recommended that intervention designers work more actively with adolescents to ensure relevance, engagement, and efficacy [36]. Self-directed educational resources that are accessible on demand have been deemed beneficial [37], as have digitized formats, due to the engagement and accessibility afforded by these platforms [38].

Co-design may offer an effective means to produce such educational resources [29]. Avoiding traditional "top-down" [39] approaches, or mere stakeholder consultations, interventions are instead built with the target population [39] as design partners throughout all phases [40]. As a novel method to close "the translational research gap" [41], co-design may improve the relevance and impact of educational resources by ensuring they are consumer driven and consumer centered [42]. Executed using various formats, past trials have used interactive workshops to involve adolescents [43] and articulate research questions into functional products [39], including most recently for mobile app–based mental health interventions [44,45]. Activities can generate creative solutions, through physical "artefacts," to inform educational resource generation. Evaluation workshops between participants and designers can then ascertain whether the solution effectively responds to participant needs and perspectives, closing the co-design cycle (Figure 1) [42].

Having grown up in the digital age, most adolescents are proficient with common digital technologies, such as the internet and social media, though such proficiency may not be transferred to all forms of technology [26]. Nevertheless, most adolescents are able to effectively express their perspectives on

XSL•FO
RenderX

DHL issues and corresponding educational needs [26]. As such, co-design with adolescents holds promise for development of web-based educational resources, despite limited existing research with this participant group [39]. Our study aimed to explore the co-design approach, based on Hagen et al [42] (Figure 1) to develop an adolescent DHL educational resource through a series of interactive workshops. Insight was sought into the influence of co-design on educational resource generation by understanding adolescents' experience with digital health information and DHL needs, and their design preferences.

Figure 1. Framework utilized for the co-design of a digital health literacy education resource (after Hagen et al) [42]. DHL: digital health literacy.

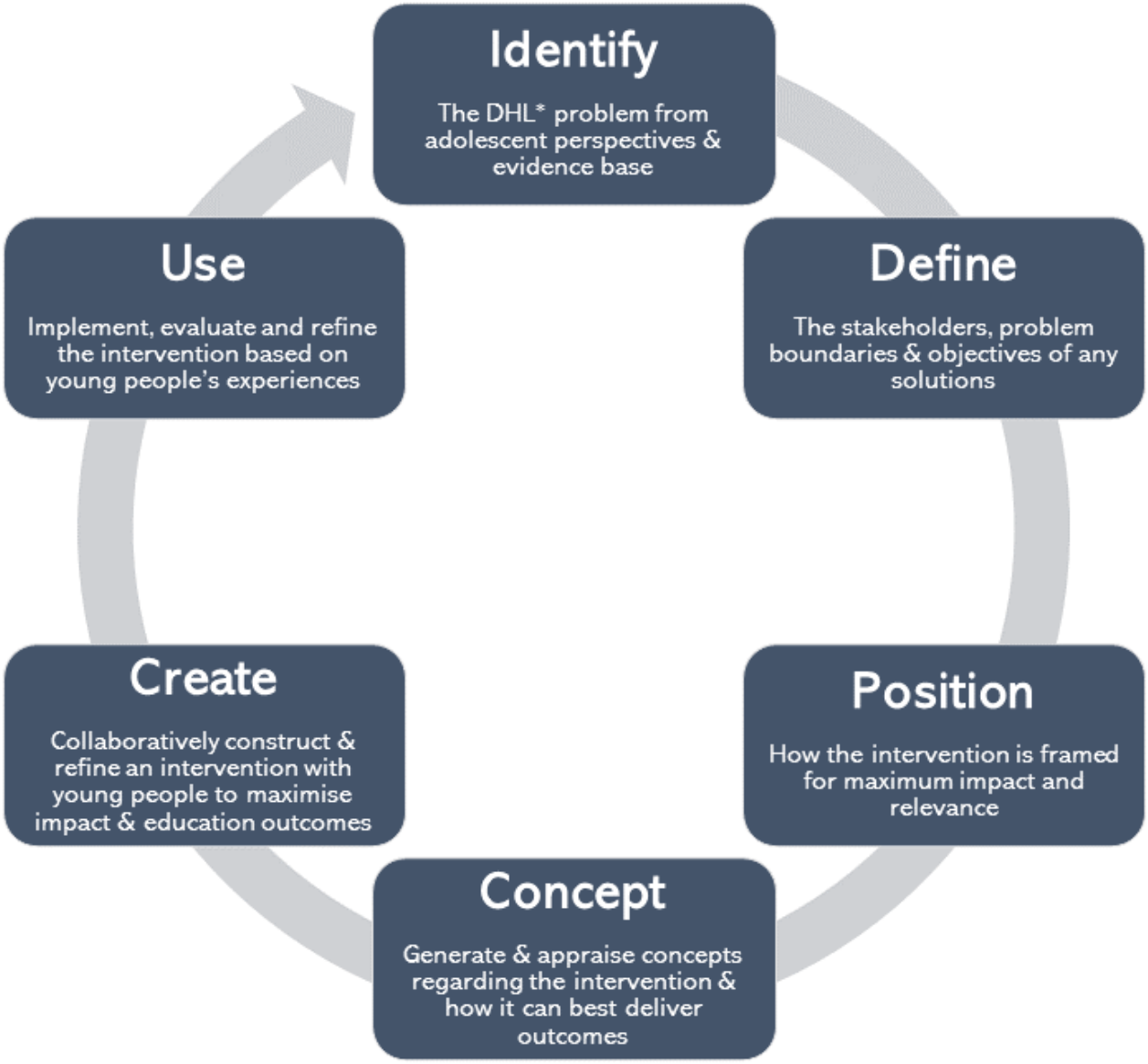

Methods

Overview

This study involved a qualitative co-design methodology, whereby 4 interactive workshops were held with adolescent participants from June 2021 to April 2022. Data generated were analyzed through inductive content analysis [46] to inform the development of an educational resource with the aim of improving users' DHL (Figure 2).

XSL•FO
RenderX

Figure 2. Research design flowchart reflecting the multistage co-design development process of the educational resource. DHL: digital health literacy.

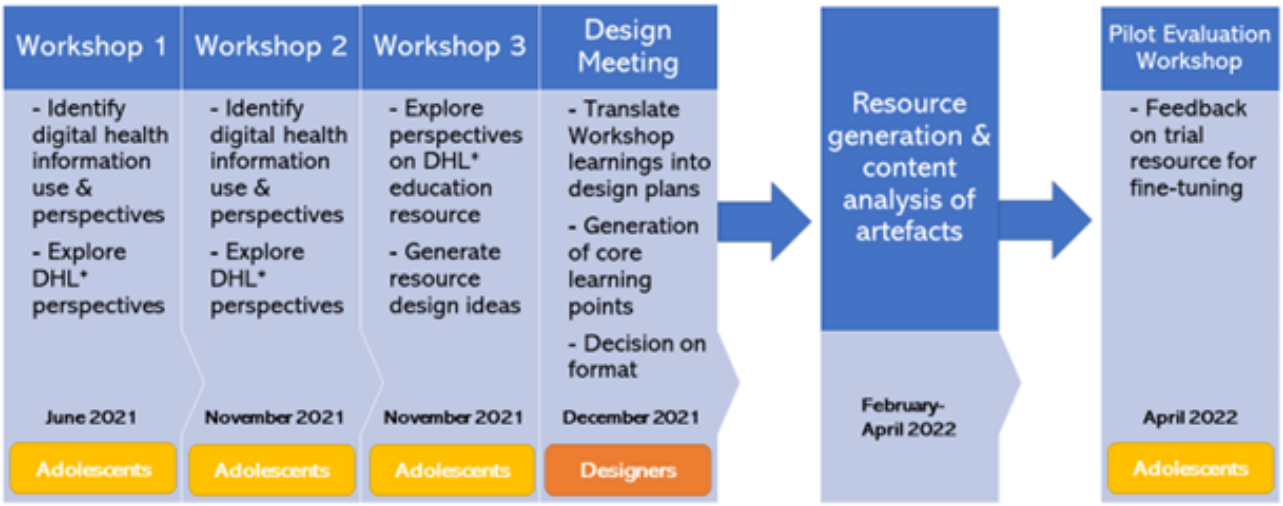

Ethical Considerations

This research project was approved by Sydney Children's Hospitals Network Human Research Ethics Committee (2020/ETH00578).

The participants, who were minors, gave their informed assent, and their parent or guardian gave their informed consent, to participate in this research study. To protect participant privacy, all study data were deidentified prior to analysis.

Participants were each provided with an AUS $50 gift card (approximately US $35 at the time of the study) for use at a range of stores as compensation for their time.

Context

To understand adolescent participants' experiences with digital health information, perspectives on DHL issues, and to ascertain their educational needs, 3 co-design workshops were organized. The workshops also sought to generate design solutions for an educational resource that could develop user skills in appraising digital health information. Interactive activities during workshops generated physical artefacts, such as design maps and sketches, which were collected for data analysis (Table 1) [43].

These activities were designed in line with SCT to foster participant collaboration in a safe, peer-based environment [30] in which interactive activities enable learning through the dynamic interrelationship of personal attitudes, behavior, and the environment [30]. The activities align with prior literature calling for the development of DHL interventions through co-design, which encompasses use of practical activities to explore and improve adolescents' self-efficacy and behavior related to DHL [29]. A design meeting was held among the researchers and the software design team developing the resource to translate workshop findings into educational resource designs. An evaluation workshop with adolescent participants was then conducted to obtain participant feedback on a pilot educational resource. The design meeting and evaluation workshop align with the final phases of the co-design framework, which seek to develop, evaluate, and refine interventions with adolescents (Figure 1) [42].

Co-design facilitators were multidisciplinary academics and research students: a medical educator with experience in eLearning and literacy (KMS), a pediatric clinical academic and child and adolescent health researcher (PHYC), a social scientist focusing on the digital and young people (PC), public health and medical student researchers (MT, CCL, and TBA) and design computing academics and software designers (LB, HH, and MB).

XSL•FO
RenderX

Table 1. Methods employed during co-design events, including activities used, artefacts generated and alignment with social cognitive theory.

Co-design event details and purpose and activities	Artefacts generated
Workshop 1 (June 2021); face-to-face: identified adolescents' perspectives on their use of digital health information, both web-based and on social media, as well as the concept of digital health literacy	
• Introduction: adolescent participants annotated a worksheet asking them about the digital health information they use, where they use such resources and reasons for use • Social Cognitive Theory (SCT) alignment: Interaction between personal, behavioral, and environmental factors	Annotated worksheets
• Enablers and barriers: participants collaboratively identified enablers and barriers of digital health information use, wrote them on sticky notes and placed these on a wall. These were then categorized and collated by idea type • SCT alignment: Influence of environment on behavior and personal factors	Affinity diagram
• Credibility rankings: on A4 posters, participants ranked (in order of credibility) 5 digital health information websites • SCT alignment: Influence of personal factors on behavior and environment	Annotated or ranked examples
• Assess and rank: participants assessed the 5 websites using a set of fixed criteria (looks professional, is understandable, I can cross check information) and criteria they generated, and ranked them based on this • SCT alignment: Influence of personal factors on behavior and environment	Annotated table
Workshop 2 (November 2021); face-to-face: identified adolescents' perspectives on their use of digital health information, both web based and on social media, as well as the concept of DHL	
• Misinformation personal stories: participants completed a worksheet with personal examples of people believing misinformation and what made it believable • SCT alignment: Influence of environment on behavioral and personal factors	Annotated worksheets
• Social spaces: participants added color-coded stickers to posters of social media platforms they use regularly, use for fun and use to find information. A set of posters of common platforms (eg, Facebook, TikTok) were preprinted, with participants adding a new poster for any additional platforms they use that were not present • SCT alignment: Influence of behavior on personal factors and environment	Sticker-annotated poster
• Field notes: researchers noted key points of discussion during all exercises	Digital written notes
Workshop 3 (November 2021); face-to-face: ascertain perspectives on what an educational resource seeking to improve adolescents' DHL should include and how it should work to be effective	
• Brainstorming design ideas: participants identified educational design features and content they believed the design solution should have, wrote them on sticky notes, then added them to a wall of annotated sticky notes. These were then categorized and collated by idea type • SCT alignment: Interaction between personal, behavioral, and environmental factors	Affinity diagram
• Resource design maps: in groups, participants identified the features about form and content that they felt a DHL educational resource should include, then sketched these on posters and labeled them • SCT alignment: Interaction between personal, behavioral, and environmental factors	Annotated posters
• Resource design sketches: participants sketched layout, appearance, format, content, and storylines that they felt would make an effective DHL educational resource • SCT alignment: Interaction between personal, behavioral, and environmental factors	Drawings
• Field notes: researchers noted key points of discussion during all exercises	Digital written notes
Design meeting (December 2021); face-to-face: translate workshop findings into practical design plans through synthesis with the researchers' educational requirements of the educational resource	
• Priority learning points: learning points from the workshops were written on palm cards. Participants (researchers and designers) added a sticker to those they felt were the most important to include in the educational resource. The cards with the most stickers were then prioritized as key learning points • SCT alignment: Interaction between personal, behavioral, and environmental factors	Sticker-annotated cards

XSL•FO
RenderX

Co-design event details and purpose and activities	Artefacts generated
• Storyline flowcharts: in groups, participants collaboratively created storylines for the educational resource that they felt would be effective in communicating learning points. These were then drawn as flowcharts on a poster • SCT alignment: Interaction between personal, behavioral, and environmental factors	Flowcharts
• Field notes: researchers noted key points of discussion during all exercises	Digital written notes
Pilot Evaluation Workshop (April 2022); virtual: gather participants' feedback on a trial version of the educational resource, which the design team will use to undertake fine-tuning	

• Resource trial and field notes: participants were guided through the educational resource app, "*mis-Adventures*," over a video meeting, enabling them to view and engage with its interactive components. The following question prompts were used to capture perspectives:
  • What did you think about *mis-Adventures*?
  • story, videos
  • decisions or choices made along the way
  • quizzes, explanations
  • expert advice at the end
  • design or look and feel
  • What did you like best about *mis-Adventures*?
  • What did you like least about *mis-Adventures*?
  • How can we improve *mis-Adventures*?
  • Please tell us about any problems you had using *mis-Adventures*.
• Researchers noted down key responses and points of discussion during the exercise
• *SCT alignment: Interaction between personal, behavioral, and environmental factors*
	Digital written notes

Participants

Participants were adolescents aged 12-17 years, representing the high school aged population the educational resource was targeted toward. Recruitment involved advertising on flyers, social media, and through word-of-mouth dissemination. The purposive recruitment strategy sought to capture a heterogenous sample despite a small overall size: we aimed to recruit participants who were of differing ages, genders and cultural backgrounds to enhance the diversity of the perspectives obtained [47], and the broad appeal and relevance of the educational resource. For each workshop, recruitment continued until a quota was met, which was set to facilitate small group sizes to enable interactive and collaborative activities. Participants were free to attend as many workshops as they wished: they were reinvited to any subsequent workshops to increase the continuity of the co-design participation cycle [42], while new participants were also invited to ensure diversity of perspectives.

Data Collection

Artefacts collected from the workshops (Table 1) were physical documents, featuring writing and drawings (examples in Figure 3). Artefacts were collected at the completion of each workshop, and subsequently scanned for digitization and analysis.

Figure 3. Examples of artefacts obtained from workshops: (A) educational resource design sketches and (B) affinity diagram.

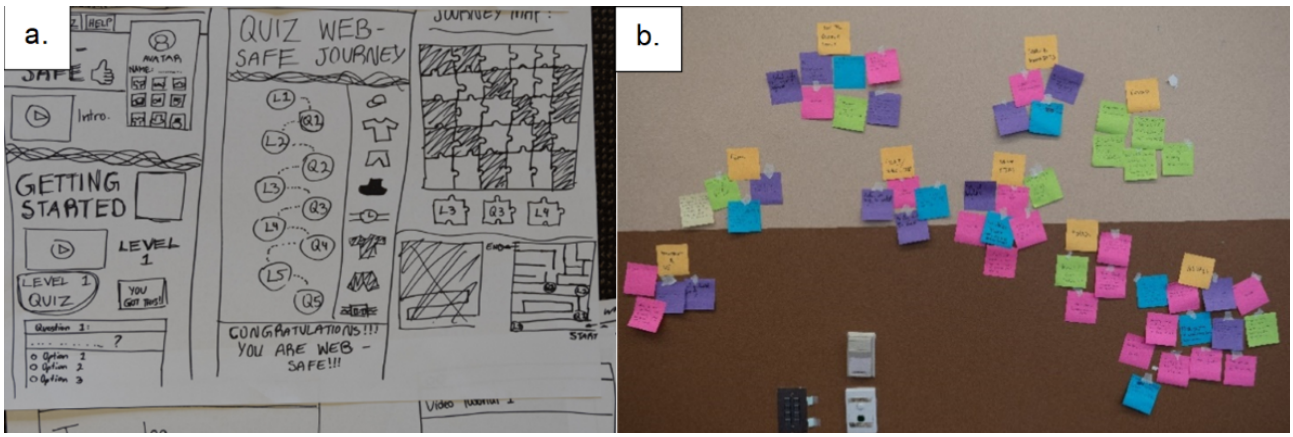

Data Analysis

Data analysis involved content analysis of the artefacts generated using NVivo (Release 1.6, QSR International) data management software [48]. This included review and familiarization with the data before iterative coding of the data features in scanned artefacts by CCL using the NVivo pictorial coding tool. An inductive approach was used [46], whereby analysis of the data set informed the codes generated, which comprised recurrent features and ideas in the artefacts. This resulted in the iterative

XSL•FO
RenderX

generation of themes, ensuring their relevance to the data [49]. This was completed through an initial categorization of recurrent codes by CCL, with researcher triangulation through analysis meetings with KMS and MT in the first instance, later joined by PHYC, which offered multiple perspectives on coding decisions, enhancing credibility [50]. Through discussion and researcher consensus, similar codes were grouped into concepts and broader categories, which became the themes and subthemes in the data, prior to in-depth analysis and interpretation. Credibility was also developed through obtaining data from multiple workshops with a range of participants.

Results are displayed through a conceptual mind map, generated through *diagrams.net* (JGraph Ltd) software [51].

Results

In total, 27 adolescents participated in the face-to-face co-design workshops: 7 at workshop 1 and 15 at workshops 2 and 3 (which were held on one day [morning then afternoon] with the same participants), and 5 at workshop 4, the pilot evaluation, which was conducted via web-based platforms because of COVID-19 social distancing restrictions. Participants were of diverse backgrounds, with a mix of genders, ages (12-16 years), and ethnicities (Anglo-Australian, Chinese, Egyptian, Italian, Latin American, Lebanese, and Taiwanese). Seven participants attended multiple workshop days, including 1 participant who attended all 4 workshops.

Participants' relationship with digital health information was one of duality—indicating an acceptance of its benefits and relevance in their lives, while having an awareness of its dangers in perpetuating misinformation. This revealed areas of DHL educational need, particularly surrounding appraisal of digital health information quality (Figure 4). Participant-derived educational solutions targeting these issues (Figure 5) highlighted the need for an accessible, self-directed digital resource aligning with participants' priorities for educational flexibility and autonomy. The educational resource designs that were generated incorporated the most useful aspects of the digital domain to enhance adolescents' engagement, alongside content that targeted credibility issues at the core of DHL (Figure 6). Through the analysis of these issues and design solutions, we identified 4 key themes: ease of access to digital health information, users' personal and social factors impacting use of DHL, impacts of the plethora of digital health information available, and anonymity offered by digital sources. These themes are detailed below, each with an analysis of participants' perspectives and needs, and corresponding design solutions, followed by participant feedback on the pilot educational resource that was created.

Figure 4. Digital Health Literacy Issues: themes and subthemes representing adolescents' experiences and issues with digital health literacy.

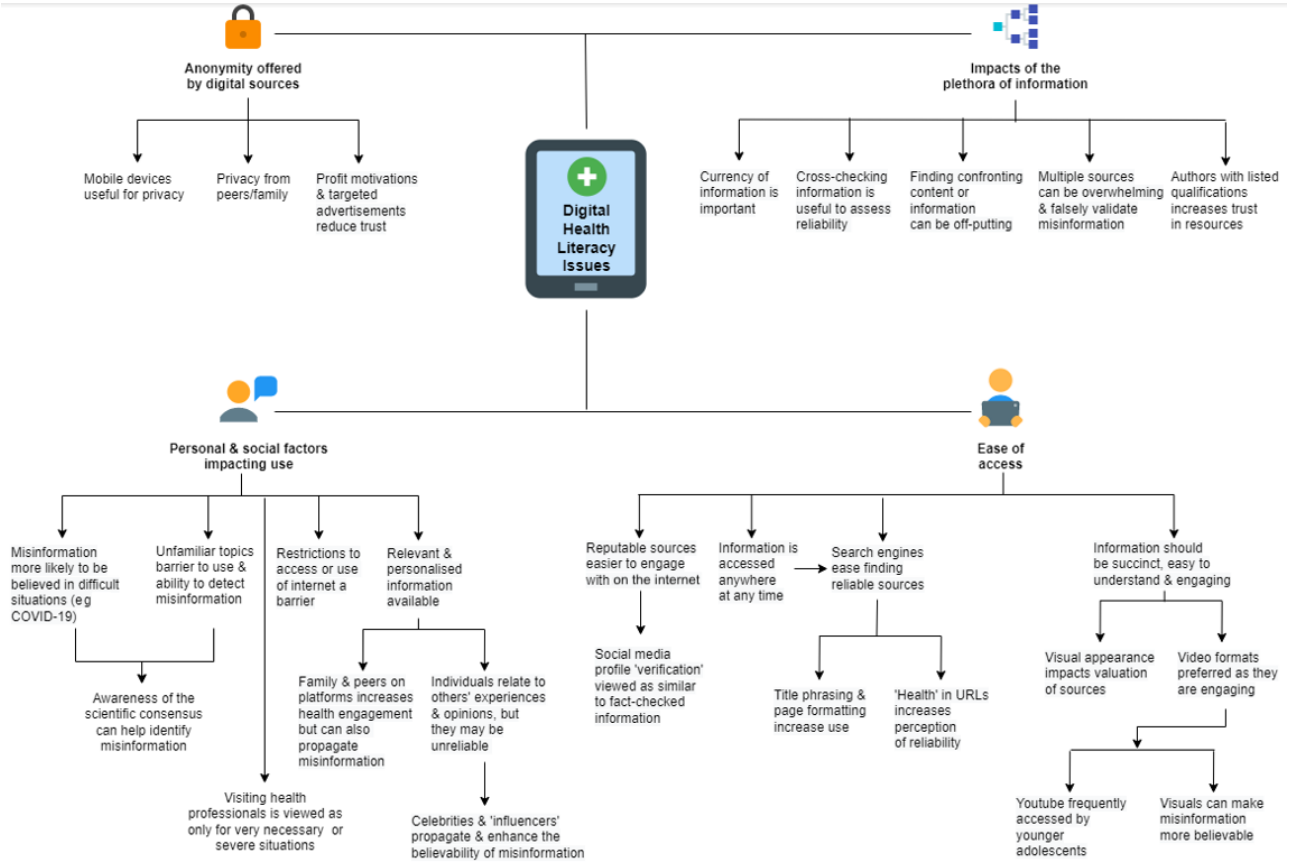

XSL•FO
RenderX

Figure 5. Digital Health Literacy (DHL) Resource Design Solutions: themes and subthemes representing design solutions generated in response to identified DHL needs. SCHN: Sydney Children's Hospitals Network.

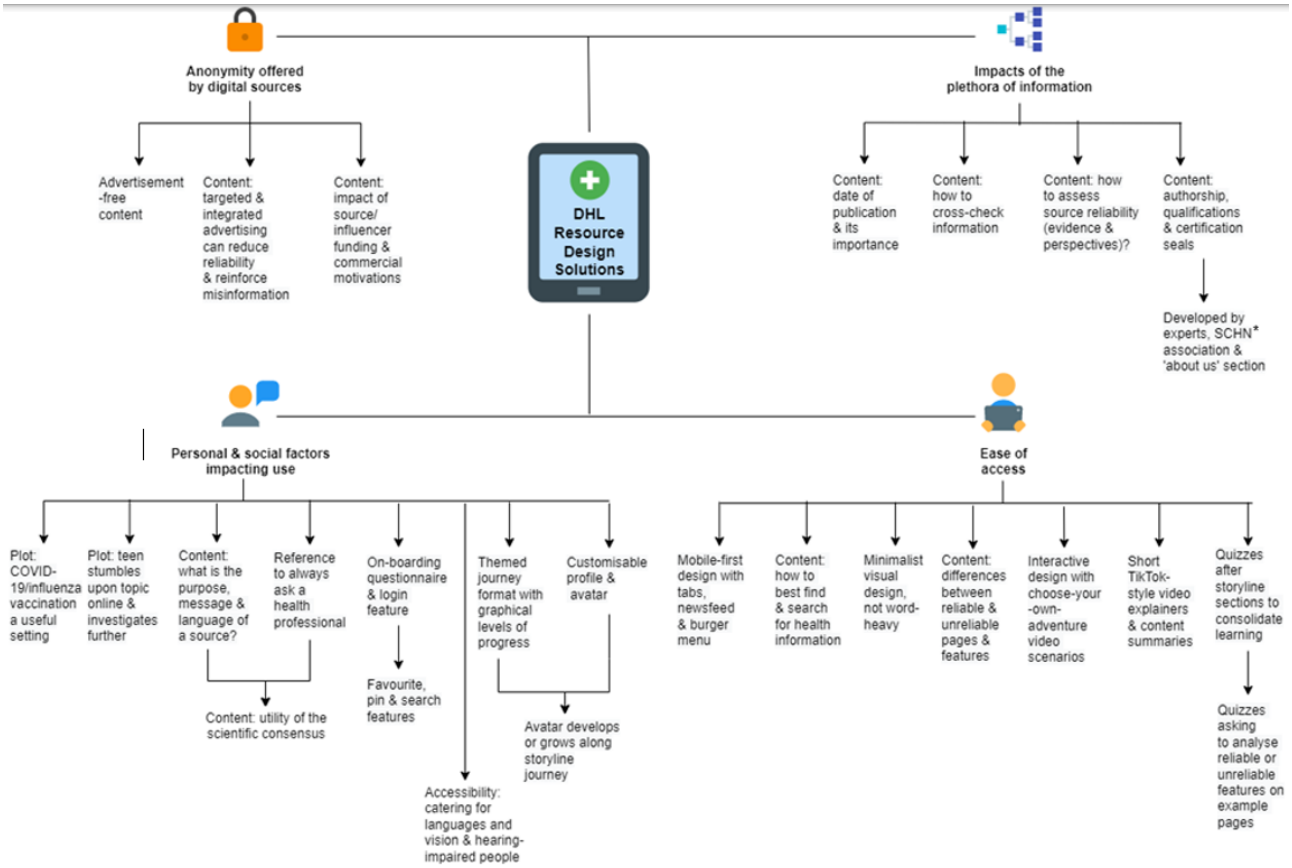

Figure 6. Example screenshots from the digital health literacy resource generated through co-design, featuring health professional explainer videos, realistic digital health information use scenarios, and simulated engagement with digital health information across multiple platforms. The individuals featured in these images have provided their written consent for their image to be published.

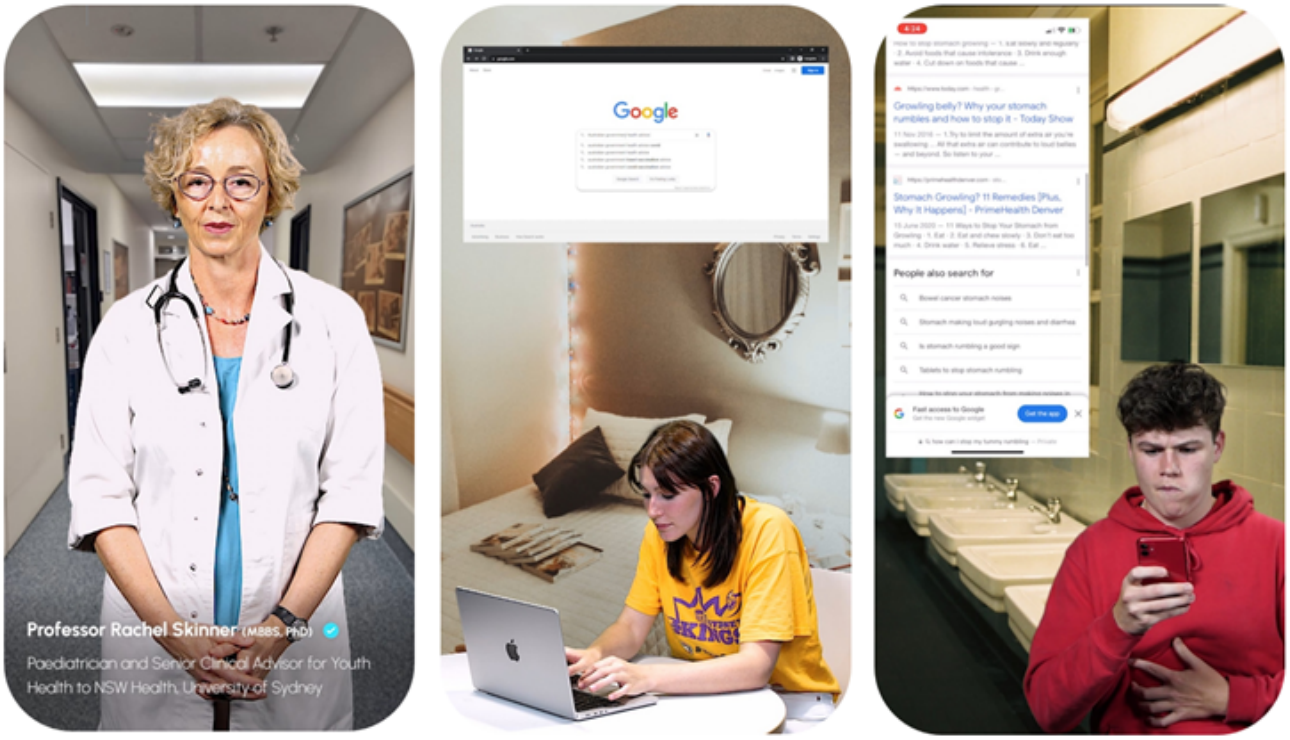

XSL•FO
RenderX

Ease of Access to Digital Health Information

User Perspective and Needs

Participants perceived ease of access to digital technology as underpinning their frequent use of digital health information, which satisfied their need for succinct, understandable, and engaging information. They preferred visual, and particularly video, formats due to their engaging nature, with younger adolescents especially accessing video platforms such as *YouTube* and *TikTok* as their first engagement with a social media network. Participants appreciated animated and image-based content for easily communicating concepts. Some perceived that video features increased credibility, even in cases of misinformation, and the aesthetic appearance of platforms was important in determining quality. Participants identified quizzes as an engaging means of understanding and retaining information. Most participants identified mobile devices as a key driver of use, allowing engagement with information across many times and locations, with search engines offering easy access to perceived credible information. Participants indicated that they appraised search results using simple heuristics, including use of website title format and presence of keywords such as "health" in URLs to predict engagement and perception of credibility. They believed the digital domain offered easier engagement with credible sources. This included "verified" profiles (indicated by a blue tick) on social media platforms, which some believed meant content was fact-checked, when at the time of the study it instead signified that the profile belonged to an authentic identity.

Co-Designed Educational Solutions

In line with their desires for engaging, highly interactive and concise content, participants envisaged an educational platform of "choose-your-own-adventure-style" stories, featuring branching storylines and plot decision points. This was implemented with content comprised of "*TikTok*-style" videos, which adopted the form's archetypal features, such as short duration (<3 min), fast pace and satirical tone, adding to the educational resource's cultural appeal. Videos also featured topic experts explaining DHL concepts and strategies for assessing credibility. As participants desired summaries to consolidate learning, this was enacted using a final video in each storyline, encompassing the educational resource's cardinal learning points. Catering to participants' device usage, the user interface has a minimalist "mobile-first" design, featuring a scrollable content carousel and smooth page transitions. Shortcut tabs and drop-down menus linking to specific content afford learning flexibility, with an emphasis on visual features to improve accessibility. Quizzes offer points of revision following videos and include identification of features reflecting credibility on emulated health websites. Automated answers with feedback are provided to enhance their educational benefit. Content includes explanations about the differences between credible and noncredible information, and how to screen search results for credibility.

Personal and Social Factors Impacting Digital Health Information Use

User Perspective and Needs

Participants identified a range of personal and social aspects of adolescent users that affect how they engage with digital health information, including psychological, situational, and peer circumstances. The web and social media–based offering of interactive, personalized information was highly appealing, such as through customization of profiles and content. This included recommended content in newsfeeds, and participants relating to others' experiences through content sharing and facilities such as forums, messaging features, and comments. This was despite participants often finding such content noncredible, identifying social media "influencers" as a pertinent example of how unqualified individuals can propagate misinformation. Platforms that cater to specific interests and cultural and linguistic groups, such as *WeChat*, which caters for a Chinese population, were popular, along with a preference for educational resources that offer accessibility for the visually and hearing impaired. Participants indicated that family and peers on digital platforms can increase engagement with health information, but also share misinformation. Some found situational factors, such as restrictions to access or content, as a barrier to digital health information use. Conversely, the availability of web and social media–based information meant many only visited health professionals when necessitated by serious situations. Participants indicated that a lack of knowledge and familiarity with specific health topics reduced their engagement and ability to detect incorrect content. For example, they were more likely to trust hyperbolized and inaccurate health content in difficult situations, such as the COVID-19 pandemic. However, participants found that having existing knowledge of the scientific consensus on an issue can aid identification of misinformation.

Co-Designed Educational Solutions

The educational resource design was based on storylines featuring a themed journey layout (eg, health preparations before a music festival) to frame users' progress through it. A live action video format was adopted as a low-resource alternative for participants' suggestion of an animated avatar with customizable features representing the user. A plot device involved encountering recommended health content on search engine results that leads the character to investigate further, reflecting participants' experience of often finding health information coincidentally, especially on social media. Captions were developed for video sections to assist users with hearing impairment.

Three storylines, derived from participants, were developed and included in the educational resource. These include a storyline about finding ways to ease gastrointestinal symptoms, based on participant feedback that stories about sensitive issues are compelling. A plotline about deciding whether to get vaccinated against influenza (as an alternative infectious diseases scenario given participants' interest in COVID-19) was also featured due to participants finding contextually relevant stories engaging. A third storyline involved searching for ways to ease exercise injuries, which was included due to sports injuries

XSL•FO
RenderX

being a common target for health products, and its appeal to male demographics.

Educational design solutions targeted difficulties with appraising information about these technical topics through simple explanations about identifying a source's purpose, message, language, author, date, and provider or funding. A disclaimer—to always seek advice from qualified health professionals—was included in the homepage and throughout the resource due to the potentially significant personal impacts of using digital health information.

Impacts of the Plethora of Digital Information Available

User Perspective and Needs

Participants indicated that the variety and magnitude of health information available on the web and social media impacted engagement. This abundance overwhelmed some or resulted in discouragement if the information was hyperbolic or inflammatory. While accessing information through multiple sources enabled cross-checking, which was believed to be universally important in assessing credibility, it sometimes led to participants falsely verifying sources due to the republishing of misinformation. This included mainstream media reporting on misinformation. Having reputable sources directly discredit such stories was regarded as effective by participants. Participants perceived the presence of author qualifications increased trust in a source, while the currency of information was important for its utility. Transparent referencing and links to further credible information were appreciated.

Co-Designed Educational Solutions

Content in the educational resource was based on research into areas of DHL educational need undertaken by the research team, aligning with participant perspectives that authorship by health professionals and educators with listed qualifications and associations with health authorities builds trust. This was coupled with participants' contributions at co-design workshops, which included their learning priorities and identified areas of DHL deficit, to generate relevant learning content.

Educational design ideas focused on content explaining how to systematically assess source credibility. This included ascertaining whether supporting scientific evidence and multiple perspectives or treatment options are included, the importance of a recent publication date, and how to effectively cross-check information with credible sources. Information about indicators of credible content, such as relevant author health qualifications, was also included. Guided by participants' suggestions to build trust, the educational resource was developed by experts with listed qualifications and association with well-known health and educational authorities.

Anonymity Offered by Digital Sources

User Perspective and Needs

Participants perceived anonymity when engaging with digital health information to be a core benefit of the digital domain because health topics they viewed were often seen as sensitive or embarrassing. They preferred to use mobile devices because they offered physical privacy from those around them, including peers and family. Participants considered the commercial motivations of web and social media–based publishers as detrimental to credibility, disliking targeted advertisements based on search history as they breached privacy.

Co-Designed Educational Solutions

Accordingly, an advertisement-free platform to build trust was implemented. Content about how to identify digital marketing was integrated due to participant interest, including the algorithms and mechanisms behind targeted advertising, and how it exploits users' confirmation bias to accept false ideas. Content was also included about how integrated health advertising on health information sites can reduce credibility due to conflicts of interest. Content was generated identifying a need to recognize source funding and its impact on information quality, including "influencers" and the commercial partnerships that may underlie their personal stories.

Educational Resource Feedback

During the pilot evaluation, many participants recognized their designs and stories in the educational resource, which they appreciated, and indicated it made the resource age-appropriate. They found the user interface to be sleek, visually appealing and easy-to-navigate, responding to their desire for a mobile-first platform. Participants also appreciated the fast pace of the video storylines, finding them concise and engaging, enjoying the "choose-your-own-adventure-style" format. For participants, this interactivity was enhanced by the ability to re-explore storylines and alternate subplots by making different decisions. The narrative and character-driven format were enjoyed, especially through the inclusion of fictitious but believable web and social media–based health personalities.

Some users would have liked the educational resource to be compatible with desktop formats, along with the inclusion of accessibility features, such as catering for sensory impairment and featuring different languages. Despite finding the educational resource age appropriate, some participants indicated the quizzes were too easy, suggesting additional levels of complexity for advanced users. This mirrored their proposals for more nuanced storylines that could explore further content with a less predictable plot. Additionally, participants desired more dramatic and engaging videos to convey the plot, including more apparent consequences for making poor decisions based on health information. Some suggested that content summary videos, which feature at the end of each storyline, could be personalized to the user's in-game decision-making to enhance relevance.

Discussion

Principal Findings

The stepwise approach of this study, from initial exploratory workshops to educational design generation, implementation, and evaluation, has enabled the application of the co-design model to the DHL context, yielding a functional and youth-centered product. Co-design elicited authentic adolescent perspectives, encapsulated by the 4 themes regarding digital health information use: ease of access to digital health

XSL•FO
RenderX

information, users' personal and social factors, the plethora of information, and anonymity offered by digital sources. These perspectives encompassed participants' DHL needs and were addressed through the generation of teaching areas in the educational resource. Targeted educational solutions regarding the most effective modalities and designs were also generated and implemented through integration of participant, researcher, and designer perspectives. The approach enabled the generation of storylines, examples of credible and noncredible sources, and content tone that are highly relevant to an adolescent audience, enhancing potential engagement with and impact of the educational resource.

The results reflect the benefit of co-design in capturing adolescents' perspectives on their DHL needs, aligning with research into adolescents' proficiency in this area, despite limited prior engagement in intervention development [26]. Co-design transformed adolescents' viewpoints into functional educational resource content. In keeping with the importance of social context in Social Cognitive Theory [30] and to influence DHL learning and self-efficacy [32], adolescents' perspectives were successfully elicited through interactive workshop activities that prompted collaborative social engagement. Ease of access to digital health information, a well-discussed driver of adolescent use [52], underpinned the educational resource's teaching points, which included identifying noncredible features of information quickly and effectively. Through these teaching points, the educational resource seeks to improve DHL skillsets by targeting the popular but often ineffective heuristics used to assess credibility, such as website names or professional appearance [26,28]. Identification of the personal and social factors affecting digital health information use, and the impact of the diverse information available, generated content that teaches objective appraisal skills that can be universally used [27]. Storylines about sensitive health issues, such as gastrointestinal symptoms, were implemented as they reflect participants' preference for the anonymity offered by digital platforms, aligning with prior findings [8].

The co-design approach also generated effective modalities and designs for the educational resource for engagement with the adolescent target audience. Ease of access underpinned the DHL needs identified and educational solutions adopted, with strong preference for video as a digital health information format [25]. The educational resource design used these findings as a framework, with short "*TikTok*-style" videos acting as core content within a minimalist mobile-first design, responding to adolescents' preferred engagement with mobile platforms [7]. Quizzes were incorporated in response to participants suggesting "gamification" [53] features to enhance engagement. In line with Social Cognitive Theory, they may also enhance learning by providing positive feedback to increase user self-efficacy [31]. Similarly, although not implemented, log-in and favorite features, and a themed journey with a customizable avatar, was recurrently suggested by participants, exemplifying how personalization and targeted information brings salience to resources [10].

Co-design enabled the generation of specific storylines and content tone that are highly relevant to the adolescent audience, enhancing potential engagement and impact of the educational resource. The interactive video "choose-your-own-adventure-style" format was key to enabling ease of use, as was the integration of authentic cultural references, such as "viral" social media trends, which were found to be important social factors underpinning engagement. This is mirrored by Paek and Hove's [36] prior calls for more effective integration of adolescents' perspectives to enhance the personal relevance and efficacy of educational interventions. Incorporated plot devices, such as a teenager coincidentally encountering a health topic on a social media feed, which leads them to undertake a search, were found by participants to be very common occurrences, paralleling their frequent use of social media to engage with health information [6]. Similarly, subplots regarding "influencers" and sponsored posts were identified and included in the educational resource to add relevance, given their prominence as a DHL issue [14,52]. The capture and implementation of these nuanced references builds upon the use of co-design in different social and cultural contexts, reflecting its versatility [41].

The design meeting, which featured both the research and design teams, was an important aspect of this study's structure, translating workshop findings into formulated content and development plans through collating and filtering participant ideas. For example, in line with recent literature, COVID-19 was a key suggested topic that exemplified DHL needs [15,23] and was featured in multiple suggested storylines to enhance realism. However, the research and resource development teams implemented an alternative influenza vaccine setting as they believed relying on the current COVID-19 context may reduce educational resource longevity. Although this illustrates a disparity between participant and academic ideology [43], the multistage nature of co-design [39] in this instance generated many other implementable features that embody adolescents' ideas. The use of interactive collaborative activities to structure the design meeting were in line with the Social Cognitive Theoretical framework and co-design approach [43]. The meeting aligned with prior workshop findings that adolescents perceived reputable sources to be effective at debunking misinformation and wanted expert consensus to shape the educational resource. While this stage of the methods may be incongruous with purist co-design theory, which prioritizes participant involvement during every development phase [42], the impact of COVID-19 lockdowns and project time and resource limitations prevented scheduling of an additional workshop to engage with adolescents at this content development stage.

The feedback obtained through the pilot evaluation workshop was useful in offering an initial indication as to whether the educational resource had adequately responded to the participants' needs and perspectives. The workshop was largely positive and constructive comments about improvements to the design, plot, and content of the resource led to fine-tuning before finalization. However, time and budget limitations resulted in some of these suggestions being noted but not implemented, presenting areas for further development in subsequent projects. Participants expressed enjoyment in being included in the

XSL•FO
RenderX

co-design process, with the final feedback stage important in closing the cycle of participant engagement [42].

Implications and Recommendations

This study's co-design approach demonstrates significant translational benefit through the articulation of adolescents' DHL perspectives and needs (many of which are in line with previous findings [9,26,29,54]) into implemented educational resource solutions [41]. The qualitative methodology held great utility in authentically exploring participants' ideas in a comfortable, peer-based environment. Mixed methods evaluation of the educational resource is needed to determine the effectiveness of the co-design approach [55].

Limitations

Although small, the participant sample was heterogeneous. Nevertheless, the sample contained some similar demographics; in particular, all participants were from metropolitan areas as planned recruitment in rural and remote regions was canceled because of social restrictions due to the COVID-19 pandemic [56]. Time limitations due to delays in participant recruitment during COVID-19 lockdowns also prevented participant involvement in storyline plot generation. However, as noted by participants in the pilot evaluation workshop, participants' design ideas were largely adopted and transformed into a functional product. Time and financial constraints limited the educational resource from featuring avatars, animations, and multiple language options, despite participant suggestions. Proposed social networking features, including sharing, forums, and messaging features, were also excluded as their ongoing moderation was beyond logistical constraints. These may reveal potential directions for future development.

Conclusions

This study offers novel insights into the implementation and impact of the co-design approach with adolescents to generate DHL educational resources, responding to prior calls for investigation in this area. The workshop format was useful in eliciting authentic adolescent perspectives and enabling their translation to underpin an engaging and functional educational resource that responds to adolescents' needs. Given these positive initial findings, co-design holds promise as an important tool for developing interventions that could improve DHL in adolescents.

Acknowledgments

Sincere thanks to Ms Charlotte Frazer, Ms Mia Rowe and Mr Toby Rowe for their time and acting contributions for the *mis-Adventures* app. Funding for this study was provided by the Australian Government, through the 2020 Australian Research Council Discovery Project "*Enhancing adolescents' health through digital health literacy.*"

Authors' Contributions

Conceptualization of this study was by KMS and PHYC, who also received input from SRS, MK and KMC. Funding acquisition was undertaken by KMS in conjunction with PHYC, SRS, MK, and KMC. Project administration was undertaken by MT and KMS, with CCL also supporting. KMS, with PHYC, played key supervision roles. Methodology was developed by KMS and PHYC, with input from SRS, MK, KMC, MT, HH, and PC. MT oversaw resources in this study, with HH, LB, MB, KMS, and CCL also assisting. Investigation was completed by HH, LB, MB, PC, MT, KMS, CCL, PHYC, and TBA. Data curation was completed by CCL, with assistance from MT and KMS. Formal analysis was undertaken by CCL, in collaboration with KMS, MT, and PHYC, who were essential in data validation. The development of *mis-Adventures* was undertaken by HH, LB, and MB. Visualization of data (Figures 4 and 5) was completed by CCL, with input from KMS, PHYC, and MT. This manuscript was drafted by CCL, in collaboration with KMS, and with critical review and editing from KMS, PHYC, MT, and CCL.
The authors attest that generative artificial intelligence has not been used in any portion of this manuscript.

Conflicts of Interest

None declared.

References

1. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health 2018;2(3):223-228. [doi: 10.1016/S2352-4642(18)30022-1] [Medline: 30169257]
2. Brindis CD, Hair EC, Cochran S, Cleveland K, Valderrama LT, Park MJ. Increasing access to program information: a strategy for improving adolescent health. Matern Child Health J 2007;11(1):27-35. [doi: 10.1007/s10995-006-0143-6] [Medline: 17031582]
3. Ruan S, Raeside R, Singleton A, Redfern J, Partridge SR. Limited engaging and interactive online health information for adolescents: a systematic review of Australian websites. Health Commun 2021;36(6):764-773. [doi: 10.1080/10410236.2020.1712522] [Medline: 31964190]
4. Liira H, Saarelma O, Callaghan M, Harbour R, Jousimaa J, Kunnamo I, et al. Patients, health information, and guidelines: A focus-group study. Scand J Prim Health Care 2015;33(3):212-219 [FREE Full text] [doi: 10.3109/02813432.2015.1067517] [Medline: 26205344]
5. Ettel G, Nathanson I, Ettel D, Wilson C, Meola P. How do adolescents access health information? And do they ask their physicians? Perm J 2012;16(1):35-38 [FREE Full text] [doi: 10.7812/TPP/11-125] [Medline: 22529757]

XSL•FO
RenderX

6. Armstrong M, Halim NK, Raeside R, Jia SS, Hyun K, Boroumand F, et al. How helpful and what is the quality of digital sources of healthy lifestyle information used by Australian adolescents? a mixed methods study. Int J Environ Res Public Health 2021;18(23):12844 [FREE Full text] [doi: 10.3390/ijerph182312844] [Medline: 34886569]
7. Devine KA, Viola AS, Coups EJ, Wu YP. Digital health interventions for adolescent and young adult cancer survivors. JCO Clin Cancer Inform 2018;2:1-15 [FREE Full text] [doi: 10.1200/CCI.17.00138] [Medline: 30652583]
8. Waldman L, Ahmed T, Scott N, Akter S, Standing H, Rasheed S. 'We have the internet in our hands': Bangladeshi college students' use of ICTs for health information. Global Health 2018;14(1):31 [FREE Full text] [doi: 10.1186/s12992-018-0349-6] [Medline: 29554929]
9. Freeman JL, Caldwell PH, Scott KM. The role of trust when adolescents search for and appraise online health information. J Pediatr 2020;221:215-223.e5. [doi: 10.1016/j.jpeds.2020.02.074] [Medline: 32446485]
10. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. Health information-seeking behaviour in adolescence: the place of the internet. Soc Sci Med 2005;60(7):1467-1478. [doi: 10.1016/j.socscimed.2004.08.010] [Medline: 15652680]
11. O'Neill B, Ziebland S, Valderas J, Lupiáñez-Villanueva F. User-generated online health content: a survey of Internet users in the United Kingdom. J Med Internet Res 2014;16(4):e118 [FREE Full text] [doi: 10.2196/jmir.3187] [Medline: 24784798]
12. Henderson EM, Eccleston C. An online adolescent message board discussion about the internet: Use for pain. J Child Health Care 2015;19(3):412-418. [doi: 10.1177/1367493513509420] [Medline: 24270989]
13. Ettel DL, Lamanno LR, Neyra SA, Ettel WJ, Ettel GL, Mitchell MK. Teens and technology transforming acne treatment. Perm J 2017;21:16-192 [FREE Full text] [doi: 10.7812/TPP/16-192] [Medline: 28898196]
14. Qutteina Y, Hallez L, Mennes N, De Backer C, Smits T. What do adolescents see on social media? a diary study of food marketing images on social media. Front Psychol 2019;10:2637 [FREE Full text] [doi: 10.3389/fpsyg.2019.02637] [Medline: 31824391]
15. Hamilton JL, Nesi J, Choukas-Bradley S. Reexamining social media and socioemotional well-being among adolescents through the lens of the COVID-19 pandemic: a theoretical review and directions for future research. Perspect Psychol Sci 2022;17(3):662-679 [FREE Full text] [doi: 10.1177/17456916211014189] [Medline: 34756118]
16. Zarocostas J. How to fight an infodemic. Lancet 2020;395(10225):676 [FREE Full text] [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
17. Gazibara T, Cakic M, Cakic J, Grgurevic A, Pekmezovic T. Patterns of online health information seeking after visiting a physician: perceptions of adolescents from high schools in central Belgrade, Serbia. Fam Pract 2021;38(3):231-237. [doi: 10.1093/fampra/cmaa118] [Medline: 33096547]
18. Asch JM, Asch DA, Klinger EV, Marks J, Sadek N, Merchant RM. Google search histories of patients presenting to an emergency department: an observational study. BMJ Open 2019;9(2):e024791 [FREE Full text] [doi: 10.1136/bmjopen-2018-024791] [Medline: 30787088]
19. Gazibara T, Cakic J, Cakic M, Grgurevic A, Pekmezovic T. Searching for online health information instead of seeing a physician: a cross-sectional study among high school students in Belgrade, Serbia. Int J Public Health 2020;65(8):1269-1278. [doi: 10.1007/s00038-020-01471-7] [Medline: 32876767]
20. Jain A, Bickham D. Adolescent health literacy and the Internet: challenges and opportunities. Curr Opin Pediatr 2014;26(4):435-439. [doi: 10.1097/MOP.0000000000000119] [Medline: 24886952]
21. Coyne I, Prizeman G, Sheehan A, Malone H, While A. An e-health intervention to support the transition of young people with long-term illnesses to adult healthcare services: Design and early use. Patient Educ Couns 2016;99(9):1496-1504. [doi: 10.1016/j.pec.2016.06.005] [Medline: 27372524]
22. Scott K, Caldwell P, Kang M, McCaffery K, Rachel Skinner S. Adolescents' use of dr google: help or hindrance? J Paediatr Child Health 2018;54(11):1282-1283. [doi: 10.1111/jpc.14203] [Medline: 30387259]
23. Patil U, Kostareva U, Hadley M, Manganello JA, Okan O, Dadaczynski K, et al. Health literacy, digital health literacy, and COVID-19 pandemic attitudes and behaviors in U.S. college students: implications for interventions. Int J Environ Res Public Health 2021;18(6):3301 [FREE Full text] [doi: 10.3390/ijerph18063301] [Medline: 33806763]
24. Nutbeam D. Defining, measuring and improving health literacy. Health Evaluation and Promotion 2015;42(4):450-455.
25. Holmberg C, Berg C, Dahlgren J, Lissner L, Chaplin JE. Health literacy in a complex digital media landscape: Pediatric obesity patients' experiences with online weight, food, and health information. Health Informatics J 2019;25(4):1343-1357 [FREE Full text] [doi: 10.1177/1460458218759699] [Medline: 29499615]
26. Freeman JL, Caldwell PH, Bennett PA, Scott KM. How adolescents search for and appraise online health information: a systematic review. J Pediatr 2018;195:244-255.e1. [doi: 10.1016/j.jpeds.2017.11.031] [Medline: 29398062]
27. Maitz E, Maitz K, Sendlhofer G, Wolfsberger C, Mautner S, Kamolz L, et al. Internet-based health information-seeking behavior of students aged 12 to 14 years: mixed methods study. J Med Internet Res 2020;22(5):e16281 [FREE Full text] [doi: 10.2196/16281] [Medline: 32209532]
28. McKinnon KA, H Y Caldwell P, Scott KM. How adolescent patients search for and appraise online health information: A pilot study. J Paediatr Child Health 2020;56(8):1270-1276. [doi: 10.1111/jpc.14918] [Medline: 32479676]
29. Taba M, Allen TB, Caldwell PH, Skinner SR, Kang M, McCaffery K, et al. Adolescents' self-efficacy and digital health literacy: a cross-sectional mixed methods study. BMC Public Health 2022;22(1):1223 [FREE Full text] [doi: 10.1186/s12889-022-13599-7] [Medline: 35725389]

XSL•FO
RenderX

30. Bandura A. Social cognitive theory: an agentic perspective. Asian J of Social Psycho 2002;2(1):21-41. [doi: 10.1111/1467-839X.00024]
31. Zimmerman B. Self-efficacy and educational development. In: Bandura A, editor. Self-efficacy in changing societies. Cambridge: Cambridge University Press; 1995:202-231.
32. Holch P, Marwood JR. Ehealth literacy in UK teenagers and young adults: exploration of predictors and factor structure of the ehealth literacy scale (eHEALS). JMIR Form Res 2020;4(9):e14450 [FREE Full text] [doi: 10.2196/14450] [Medline: 32897230]
33. Li S, Cui G, Kaminga AC, Cheng S, Xu H. Associations between health literacy, ehealth literacy, and COVID-19-related health behaviors among chinese college students: cross-sectional online study. J Med Internet Res 2021;23(5):e25600 [FREE Full text] [doi: 10.2196/25600] [Medline: 33822734]
34. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 Skills. J Med Internet Res 2017;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
35. Rivadeneira MF, Salvador C, Araujo L, Caicedo-Gallardo JD, Cóndor J, Torres-Castillo AL, et al. Digital health literacy and subjective wellbeing in the context of COVID-19: a cross-sectional study among university students in Ecuador. Front Public Health 2022;10:1052423 [FREE Full text] [doi: 10.3389/fpubh.2022.1052423] [Medline: 36711373]
36. Paek H, Hove T. Social cognitive factors and perceived social influences that improve adolescent eHealth literacy. Health Commun 2012;27(8):727-737. [doi: 10.1080/10410236.2011.616627] [Medline: 22452551]
37. Schweder S, Raufelder D. Needs satisfaction and motivation among adolescent boys and girls during self-directed learning intervention. J Adolesc 2021;88:1-13. [doi: 10.1016/j.adolescence.2021.01.007] [Medline: 33571891]
38. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. The Internet: a window on adolescent health literacy. J Adolesc Health 2005;37(3):243. [doi: 10.1016/j.jadohealth.2004.08.023] [Medline: 16109345]
39. Collin P, Swist T. From products to publics? The potential of participatory design for research on youth, safety and well-being. Journal of Youth Studies 2016;19(3):305-318. [doi: 10.1080/13676261.2015.1098774]
40. Janiak E, Rhodes E, Foster AM. Translating access into utilization: lessons from the design and evaluation of a health insurance Web site to promote reproductive health care for young women in Massachusetts. Contraception 2013;88(6):684-690. [doi: 10.1016/j.contraception.2013.09.004] [Medline: 24120249]
41. Ospina-Pinillos L, Davenport TA, Navarro-Mancilla AA, Cheng VWS, Cardozo Alarcón AC, Rangel AM, et al. Involving end users in adapting a spanish version of a web-based mental health clinic for young people in Colombia: exploratory study using participatory design methodologies. JMIR Ment Health 2020;7(2):e15914 [FREE Full text] [doi: 10.2196/15914] [Medline: 32027313]
42. Hagen P, Collin P, Metcalf A, Nicholas M, Rahilly K, Swainston N. Participatory Design of evidence-based online youth mental health promotion, intervention and treatment. Abbotsford, Victoria: Young and Well Cooperative Research Centre; 2012.
43. Bagnoli A, Clark A. Focus groups with young people: a participatory approach to research planning. Journal of Youth Studies 2010;13(1):101-119. [doi: 10.1080/13676260903173504]
44. Li S, Achilles M, Spanos S, Habak S, Werner-Seidler A, O'Dea B. A cognitive behavioural therapy smartphone app for adolescent depression and anxiety: co-design of ClearlyMe. tCBT 2022;15:1-32. [doi: 10.1017/s1754470x22000095]
45. Hugh-Jones S, Pert K, Kendal S, Eltringham S, Skelton C, Yaziji N, et al. Adolescents accept digital mental health support in schools: A co-design and feasibility study of a school-based app for UK adolescents. Mental Health & Prevention 2022 Sep;27:200241. [doi: 10.1016/j.mhp.2022.200241]
46. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
47. Patton MQ. Qualitative evaluation and research methods, 2nd ed. California: SAGE Publications, Inc; 1990.
48. NVivo. Release 1.6 ed. Denver: QSR International Pty Ltd; 2020. URL: https://help-nv.qsrinternational.com/ [accessed 2024-06-05]
49. Vears DF, Gillam L. Inductive content analysis: A guide for beginning qualitative researchers. Focus on Health Professional Education: A Multi-Professional Journal 2022;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
50. Jonsen K, Jehn K. Using triangulation to validate themes in qualitative studies. Qualitative research in organizations and management: an international journal 2009;4(2):123-150.
51. diagrams.net. Northampton, England: JGraph Ltd; 2022. URL: https://diagrams.net [accessed 2024-06-05]
52. Lupton D. Young people's use of digital health technologies in the global north: narrative review. J Med Internet Res 2021;23(1):e18286 [FREE Full text] [doi: 10.2196/18286] [Medline: 33427684]
53. Liverpool S, Mota CP, Sales CMD, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
54. Freeman JL, Caldwell PH, Scott KM. How adolescents trust health information on social media: a systematic review. Acad Pediatr 2023;23(4):703-719 [FREE Full text] [doi: 10.1016/j.acap.2022.12.011] [Medline: 36581098]
55. Morgan DL. Integrating Qualitative and Quantitative Methods: A Pragmatic Approach. United States: SAGE Publications, Inc; 2014:A.

XSL•FO
RenderX

56. Hagen P, Collin P, Gesling K, Herrod L, Metcalf A, Blanchard M, et al. Enabling Participation: A companion document to the Young and Well Cooperative Research Centre Innovative Methodologies Guide Participatory Design of Evidence-Based Online Youth Mental Health Promotion, Intervention and Treatment. Abbotsford, Victoria: Young and Well Cooperative Research Centre; May 2016:1-9.

Abbreviations

DHL: digital health literacy
SCT: Social cognitive theory

Edited by T de Azevedo Cardoso, S Ma; submitted 06.06.23; peer-reviewed by PCI Pang, JR Bautista; comments to author 25.10.23; revised version received 24.11.23; accepted 16.05.24; published 07.08.24.

Please cite as:
Lewis CC, Taba M, Allen TB, Caldwell PHY, Skinner SR, Kang M, Henderson H, Bray L, Borthwick M, Collin P, McCaffery K, Scott KM
Developing an Educational Resource Aimed at Improving Adolescent Digital Health Literacy: Using Co-Design as Research Methodology
J Med Internet Res 2024;26:e49453
URL: https://www.jmir.org/2024/1/e49453
doi:10.2196/49453
PMID:

©Callum C Lewis, Melody Taba, Tiffany B Allen, Patrina HY Caldwell, S Rachel Skinner, Melissa Kang, Hamish Henderson, Liam Bray, Madeleine Borthwick, Philippa Collin, Kirsten McCaffery, Karen M Scott. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Mobile-Centered Digital Health Readiness Scale (mDiHERS): Health Literacy and Equity Scale

Hana Kim[1], PhD; Rebecca Schnall[2], PhD; Nagyeom Yoon[3], BSN; Seong-Joon Koh[4], PhD; Jisan Lee[3*], PhD; Jae Hee Cheon[5*], PhD

[1]Department of Nursing, Hoseo University, Asan, Republic of Korea
[2]School of Nursing, Columbia University, New York, NY, United States
[3]Department of Nursing, Gangneung-Wonju National University, Wonju, Republic of Korea
[4]Department of Internal Medicine and Liver Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea
[5]Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Jisan Lee, PhD
Department of Nursing
Gangneung-Wonju National University
Department of Nursing, Gangneung-Wonju National University
Wonju, 26403
Republic of Korea
Phone: 82 337608646
Email: saan2mari@gmail.com

Abstract

Background: There has been a rapid expansion of digital health care services, making the need for measuring and improving digital health readiness a priority. In response, our study team developed the *Mobile-Centered Digital Health Readiness: Health Literacy and Equity Scale* (mDiHERS) to measure digital health readiness.

Objective: We aim to develop and validate a scale that assesses digital health readiness, encompassing literacy and equity, and to ensure the effective use of mobile-centered digital health services.

Methods: This study was conducted from October 2021 to October 2022 to develop and validate the mDiHERS. Participants included patients with inflammatory bowel disease, which is a chronic condition requiring continuous management, and experts in medical and nursing informatics. The scale development involved a literature review, focus group interviews, and content validity evaluations. A total of 440 patients with inflammatory bowel disease were recruited for the validation phase, with 403 completing the survey. The scale's validity and reliability were assessed through exploratory factor analysis and Cronbach α. The scale was translated into English by translators and bilingual and native researchers, ensuring its applicability in diverse settings.

Results: The mDiHERS consists of 36 items across 6 domains, with a 5-point Likert scale for responses. The validation process confirmed the scale's construct validity, with 4 factors explaining 65.05% of the total variance. The scale's reliability was established with Cronbach α values ranging from 0.84 to 0.91. The scale's development considered the technical proficiency necessary for engaging with health mobile apps and devices, reflecting the importance of subjective confidence and objective skills in digital health literacy.

Conclusions: The mDiHERS is a validated tool for measuring patients' readiness and ability to use digital health services. The mDiHERS assesses user characteristics, digital accessibility, literacy, and equity to contribute to the effective use of digital health services and improve accessibility. The development and validation of the mDiHERS emphasize the importance of confidence and competence in managing health digitally. Continuous improvements are necessary to ensure that all patients can benefit from digital health care.

(J Med Internet Res 2024;26:e58497) doi:10.2196/58497

XSL•FO
RenderX

KEYWORDS

digital health; health literacy; health equity; inflammatory bowel diseases; telemedicine; patient participation; validation; validate; IBD; bowel; inflammatory; inflammation; gastrointestinal; GI; internal medicine; gastroenterology; scale; readiness; adoption; measure; measures; measurement; measurements; assessment; assessments; scales; eHealth; e-health; literacy; mHealth; mobile health; chronic; mobile phone

Introduction

The digital health care service market has experienced significant expansion. Digital health, as defined by the US Food and Drug Administration, represents the convergence of people, information, technology, and connectivity, all of which collaboratively enhance health care delivery and outcomes. This extensive field includes various components such as mobile health (mHealth), health information technology, wearable devices, telehealth and telemedicine, and personalized medicine, each playing a crucial role in advancing the functionality and reach of contemporary medical practices [1].

The evolution of digital health is further propelled by patients who, as informed health consumers, are increasingly harnessing digital devices for health management [2]. Digital health services, as defined in this study, refer to the use of digital health to manage and monitor patient health outcomes, complementing the accessible, efficient, and patient-centric delivery of digital health [3]. A notable advancement in this domain is remote health managing and monitoring, which empowers patients to access health care services within the comfort of their homes, thereby fostering a sense of independence [3,4].

Increased use of digital health services has the potential to induce disparities in accessibility, proficiency, and the degree of health information and technology use among user groups, necessitating careful consideration of vulnerable populations. In response to these challenges, there is a growing emphasis on devising strategies to gauge the digital divide and implement system-wide solutions that champion digital inclusion, ensuring that marginalized groups are not left behind in the digital health landscape [5].

However, to support equitable provision of digital health services, prioritization should be given to measuring the digital divide rather than merely focusing on the means of delivery [6]. “Digital readiness” encompasses digital access, use, literacy, and the competency to engage with digital health services [7]. It emerges as a pertinent metric to assess disparities in digital health service use [7].

Existing digital health literacy scales have primarily focused on specific aspects. For instance, the Digital Health Literacy Instrument evaluates internet usage skills related to health on the web and computers but does not consider aspects related to mobile usage [6]. Similarly, the eHealth Literacy Scale assesses the ability to find, evaluate, and apply electronic health information to health issues but does not evaluate the importance of mHealth apps and devices and the ability to use mobile services required for daily life [8]. These limitations may result in an inaccurate reflection of users’ digital health readiness. Additionally, because existing health literacy scales rely on subjective measurements, it is impossible to know how an individual’s responses relate to actual skill level, whereas objective tests can directly measure an individual’s skills [9,10]. Motivated by this, our study endeavors to construct and validate a scale measuring digital health readiness consisting of subjective and objective questions.

One of the core principles of Healthy People 2020 is to eliminate health disparities and achieve health equity to attain health and well-being [11]. As the health care sector becomes digitized, digital access, such as through mobile apps, is now recognized as a social determinant of health [12]. Low digital access due to low digital literacy undermines health equity [13]. Despite the development and research of many tools to improve and measure equity, most of these tools only measure equity in health care settings and public health domains [14-16]. Specifically, the Health Equity Assessment Toolkit was developed for use primarily by public health professionals, researchers, and those with basic skills in health information systems and interpreting health-related data, rather than by the target population directly [17]. Therefore, our study developed measures to assess health equity based on digital literacy.

Inflammatory bowel disease (IBD), encompassing Crohn disease and ulcerative colitis, is characterized by chronic digestive tract inflammation. This condition leads to a range of symptoms that significantly impact the quality of life [18]. The persistent symptoms and activity of IBD, even with the best medical or surgical interventions, highlight the critical need for ongoing surveillance and management [19]. Given the complex, chronic nature of IBD, which significantly impairs quality of life, there have been various efforts to leverage technologies to modify behaviors and assist in the self-management of patients with IBD [20,21]. As of 2023, while over 40 free English-language IBD mobile apps have been found to meet acceptable quality criteria, there is a crucial need for enhanced design features to improve user interest and engagement [22]. Recognizing the importance of continuous and personalized interventions, which focus on the immediate management of IBD symptoms [23], it becomes essential to evaluate patients’ access to and understanding of mobile technology about the available apps. This study, therefore, seeks to create an instrument for evaluating digital health preparedness, with an initial focus on individuals with IBD. This demographic is not only in dire need of digital health solutions but also exhibits a higher demand for such services. While first tested for specific conditions, the assessment tool is designed to be flexible and applicable across a spectrum of health scenarios [24]. Consequently, this instrument could be applied to a broad range of patient groups beyond IBD in subsequent studies, aiming to bridge the digital health divide.

XSL•FO
RenderX

Methods

Study Design

Overview of Study Design

This study was conducted from October 8, 2021, to October 7, 2022. This study was developed through a 4-step process. Step 1 involved a literature review to derive the initial scale items related to digital health literacy and digital health equity. Step 2 aimed to gain qualitative insights into the initial scale items by conducting focus group interviews (FGIs) with 6 patients with IBD and 6 experts, followed by an evaluation of the content validity of the initial scale items by 8 experts. Step 3 evaluated the validity and reliability of the final scale items quantitatively with 440 patients with IBD. This included survey research and statistical analysis. Step 4 translated the final scale items into both English and Korean simultaneously.

Participants and Inclusion Criteria

In all stages, patient participants were required to meet the following criteria: (1) have a confirmed diagnosis of ulcerative colitis or Crohn disease, (2) possess the ability to use a smartphone proficiently, (3) be adults aged 18 years or older, and (4) be willing and able to provide informed consent. Additionally, participants needed to be capable of completing FGIs or surveys as required in each stage.

Recruitment

Participant recruitment was conducted from November 30, 2021, to September 30, 2022. Patients with IBD were recruited in steps 2-1, 2-3, and 3-1. Recruitment was conducted in the outpatient clinic of the gastroenterology department at S Hospital in Seoul and through the IBD online community in the social network service (eg, KakaoTalk; Kakao Corp). Posters introducing this study, approved by the institutional review board and each institution, were posted in the outpatient clinic and online community. Interested participants were provided with a study consent form and an explanatory document, and this study was initiated after this study's details were explained to them. Since the interviews or surveys of this study were conducted online, an exemption from written consent was obtained.

In steps 2-1 and 2-3, respectively, 6 patients with IBD were recruited [25] to conduct FGIs to evaluate the qualitative appropriateness of the draft version of the scale. Participants in the FGIs were compensated with approximately US $77 (₩100,000) each. In step 3-1, the target sample size had been determined as 440, based on a calculation of 10 times the number of items in the fourth draft version of the scale, while accounting for an anticipated dropout rate of 10% [26].

Experts were recruited in steps 2-1 and 2-2 through snowball sampling to evaluate FGI and content validity. A total of 6 experts were recruited in step 2-1, including 5 nursing informatics experts and 1 medical informatics expert. Experts who participated in content validity verification received approximately US $154 (₩200,000) each. In step 2-2, eight experts were recruited, including 4 nursing informatics experts, 1 medical informatics expert, 1 educational expert, 1 gastroenterology physician, and 1 user experience designer. Patients who participated in the online survey in step 3-1 were compensated with the equivalent of approximately US $23 (₩30,000).

Development of a Digital Health Readiness Scale

The primary objective of this study was to develop and validate a comprehensive scale for evaluating an individual's readiness to engage with digital health services. The instrument focuses on assessing digital health literacy and equity among users. The overall process of scale development and the results of each step are shown in Figure 1.

XSL•FO
RenderX

Figure 1. Study process and results of each step: development and validation of the Mobile-Centered Digital Health Readiness Scale for patients with inflammatory bowel disease. The names of the items reflect those of the initial version of the tool, which may differ from those of the final tool. FGI: focus group interview; IBD: inflammatory bowel disease; UX: user experience.

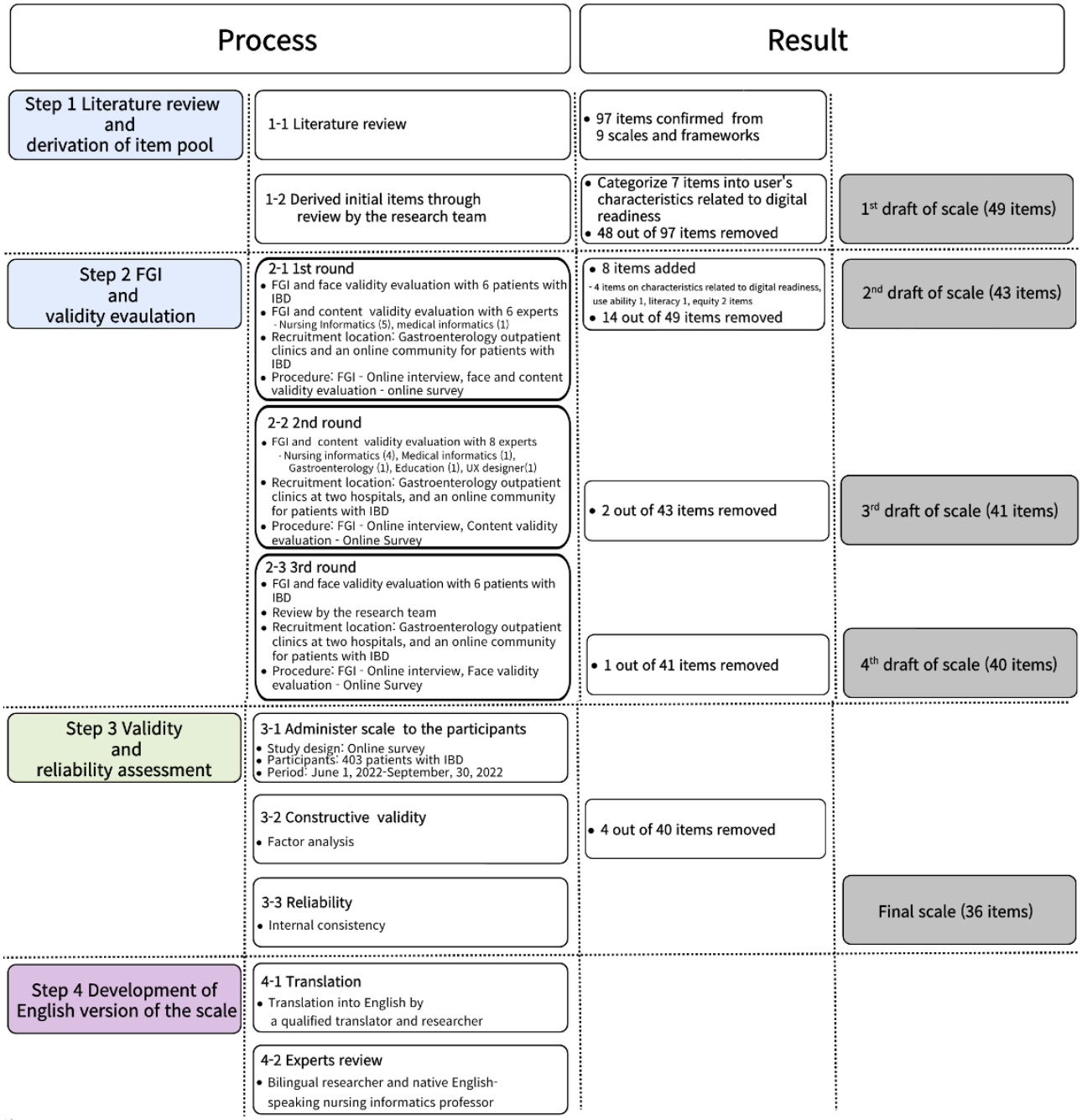

Step 1: Literature Review and Derivation of an Initial Item Pool

Step 1-1: Literature Review

In our literature review process, we conducted a comprehensive search using both English and Korean languages, leveraging databases such as the Web of Science and the Korean database Research Information Sharing Service. Our search strategy focused on identifying literature and frameworks related to digital health literacy, digital health equity, and other relevant terms to ensure inclusivity and comprehensiveness, particularly concerning mHealth services. The timeframe for our literature review spanned from October 2021.

Step 1-2: Derived Initial Items Through Review by a Research Team

To derive items, the research team reviewed the subdomains identified in the literature review to determine the subdomains of this tool. Accordingly, the items derived in the previous stage were classified into each subdomain, and inappropriate items were excluded after reviewing their appropriateness and validity.

Step 2: FGI and Validity Assessment

Step 2-1: First Round

FGI was conducted with 6 patients with IBD to assess the face validity of the initial scale items and to gather qualitative insights regarding their relevance. These interviews were

XSL•FO
RenderX

conducted online and spanned approximately 2 hours. The methodology involved structured open-ended questions, probing for understanding and appropriateness of the scale items. The process continued until data saturation was reached, indicated by the absence of new emerging data [27]. The questions were "Do you comprehend the items presented in the scale?" "In cases of lack of understanding, could you specify which aspects are unclear?" "Do you find the items within the scale to be suitable and relevant?" "If you perceive any items as unsuitable, could you elaborate on the reasons?" "Are there any elements or items that you believe should be added to enhance the scale?" and "Overall, what are your impressions or thoughts regarding the preliminary version of the scale?" In addition, FGI and content validity evaluation were conducted on 6 experts (5 in nursing informatics and 1 in medical informatics). Based on the above results, the second draft version was completed.

Step 2-2: Second Round

In this stage, 8 experts conducted FGI and content validity evaluation on the draft derived from the previous stage. The 8 experts comprised 4 in nursing informatics, 1 in medical informatics, 1 gastroenterology professor, 1 education doctorate, and 1 user experience designer. Then, the third draft of the scale reflecting experts' opinions was derived.

Step 2-3: Third Round

In steps 2-3, FGI and face validity evaluation were conducted for 6 patients with IBD using the draft scale derived in the previous step using the same method as step 2-1. The research team reviewed the revised scale reflecting patient opinions and completed the fourth draft.

Step 3: Validity and Reliability Assessment

Step 3-1: Administer Scale to the Participants

An online survey was administered to 440 patients with IBD using the second draft of the scale derived in the earlier stage.

Step 3-2: Construct Validity

The construct validity was ascertained by examining the correlation between individual items and the overall scale score. A confirmatory factor analysis followed this.

Step 3-3: Reliability Analysis

The Cronbach α value was used to confirm reliability: if it was 0.75 or higher, it was evaluated as satisfactory, and if it was 0.6 or higher, it was evaluated as acceptable [26].

Step 4: Development of an English Version of the Scale

Step 4-1: Translation

After the reliability assessment, the final scale was concurrently developed in English and Korean to facilitate future translations. The Korean version was initially crafted, referenced by a translation into English adhering to part of World Health Organization (WHO) translation guidelines [28].

Step 4-2: Experts Review

Bilingual researchers specializing in medical informatics and nursing informatics reviewed the translated version of steps 4-1. Finally, a native English-speaking nursing informatics professor conducted a thorough review and revision, culminating in the finalized English scale version.

Statistical Analysis

The scale development involved conducting a literature review, classifying the collected content, and deriving categories through analysis of the scale developed from the literature. The scale was evaluated for face validity targeting patients with IBD, and its appropriateness was verified through expert content validity index (CVI) evaluation. The final developed scale was conducted through a survey, and the survey results were analyzed using SPSS (version 21.0; IBM Corp) as follows: first, the demographic characteristics of the participants were analyzed by frequency, percentage, average, and SD. Second, for the construct validity of the developed scale, item analysis was performed by calculating the correlation coefficient between individual items and the overall total score, and exploratory factor analysis was performed. Exploratory factor analysis was verified using principal component analysis and varimax rotation. Convergent validity was judged using the values derived through exploratory factor analysis, and the final questions were confirmed. Third, the Cronbach α was used to confirm the internal consistency of the developed scale.

Ethical Considerations

This study was conducted after obtaining approval (H-2108-238-1251) from the Seoul National University Hospital Bioethics Review Committee before starting this study to ensure the ethical protection of the research participants. All study participants received an explanation of this study and completed written consent. All data collected during this study period were protected through appropriate safeguards. Additionally, compensation was provided to all participants.

Results

Step 1: Literature Review and Derivation of an Initial Item Pool

Step 1-1: Literature Review

A thorough literature review was conducted to identify existing scales relevant to developing a digital health readiness evaluation scale for patients. The focus was primarily on mHealth literacy and digital health equity scales. A total of 6 scales and 3 frameworks on digital health readiness deemed suitable for this study were selected [8,29-36], and 97 items were confirmed from these scales and frameworks.

Step 1-2: Derived Initial Items Through Review by a Research Team

At this stage, we reviewed the subdomains derived from the literature review and determined 4 subdomains: familiarity, importance, equity, and usability (literacy). After classifying the 97 items derived from the previous stage into these 4 subdomains, we reviewed the appropriateness and validity of the items, excluding inappropriate ones. Additionally, we classified 7 user characteristics identified as important in the digital readiness framework into one additional subdomain and decided to include them in the final scale.

XSL•FO
RenderX

Step 2: FGI and Validity Assessment

Step 2-1: First Round

Table 1 presents the results of face validity evaluation and FGI performed on 6 patients with IBD. This process was instrumental in identifying items irrelevant to digital health readiness or requiring further clarification. The draft scale's CVI was rigorously assessed by 6 experts, with a threshold of 0.8 indicating high validity [37]. In addition, FGI was used to identify essential domains of digital health readiness not included in this scale and to evaluate whether the readability and language used were appropriate for the participants or data collectors [38]. As a result of the expert CVI evaluation, 20 out of 78 items were confirmed to be less than 0.8. These items were partially revised and supplemented based on expert opinions. Comments on modifying items identified through FGI included providing commonly used terms and examples to aid respondents' understanding. Moreover, duplicate items and items that did not fit the current medical environment were deleted. As a result, 14 of the existing 49 items were deleted, and 8 items (1 item on usability, 1 item on literacy, 2 items on equity, and 4 items on characteristics related to digital readiness) were added, completing the second draft with a total of 43 items. It was decided to classify 4 digital-related characteristics items added at the expert's suggestion and 1 item among the existing scale into the digital-related characteristics domain and include them in the final scale.

Table 1. Analysis of focus group interview and face validity for initial scale items: evaluations by patients with inflammatory bowel disease and experts on the Mobile-Centered Digital Health Readiness: Health Literacy and Equity Scalea.

Number	Item description	Participant feedback
2	You can use mobile devices to process administrative tasks and use electronic civil service services through public institution websites.	Item relevance: Questioned the significance of differentiating equity based on the usage of electronic civil services.
12	You can contact health-related people and send files through Social Network Services such as mobile KakaoTalk and Facebook (Meta).	Item relevance: It was noted that while proficient mobile users generally use apps effectively, there may be exceptions where skilled mobile users do not engage with social networking services.
18	Compare sources of health information and confirm whether the information is true.	Item relevance: Raised doubts about the relevance of questions about the evaluation of comments or information sources.
34	For digital health equity, health care providers (doctors, nurses, etc) need to receive related education.	Emphasized the need for equal opportunities for those unable to use digital devices, but highlighted the challenge in assessing this through questionnaire items.
35	For digital health equity, medical consumers (patients) need to receive related education.	Emphasized the need for equal opportunities for those unable to use digital devices, but highlighted the challenge in assessing this through questionnaire items.
36	For digital health equity, those involved in medical service development (health app designers, mobile medical device developers) need to receive related education.	Emphasized the need for equal opportunities for those unable to use digital devices, but highlighted the challenge in assessing this through questionnaire items.

[a]Overall feedback: Understanding of items: Participants found the terminology in the items complex and suggested modifications for easier comprehension by a broader audience. Item relevance: Expressed uncertainty about the significance of some items concerning digital literacy and equity.

Step 2-2: Second Round

In the expert CVI evaluation, 2 out of 78 items were confirmed to be less than 0.8 and were deleted. There were opinions that the items related to digital health equity in FGI were ambiguous, so some items were modified. Through this stage, the third draft with a total of 41 items was completed.

Step 2-3: Third Round

Face validity and FGI were performed on 6 patients with IBD. We collected opinions on the third draft derived from the previous stage and modified the scale by reflecting opinions on adding additional explanations to some terms and changing them with more accessible terms. Afterward, through review by the research team, one question inappropriate for measuring digital readiness was deleted. As a result, the fourth draft of 40 items was completed.

Step 3: Validity and Reliability Assessment

Step 3-1: Administer Scale to the Participants

The survey, designed to assess the validity and reliability of the developed scale, encompassed a total of 403 participants. The demographic breakdown revealed that 68.5% (n=276) of the participants were under 40 years of age. Regarding gender distribution, 55.1% (n=222) were male. Educational background indicated that a significant majority, 75.2% (n=303), possessed at least a college degree. Detailed demographic characteristics are presented in Table 2.

XSL•FO
RenderX

Table 2. General characteristics of this study population (n=403): demographic and clinical information of patients with IBD[a]. Digital readiness-related characteristics had a mean of 9.98 (SD 2.74).

Variables and categories	Value, n (%)
Age group (years)[b]	
≤20	131 (32.5)
21-39	145 (36)
40-49	69 (17.1)
50-59	40 (9.9)
≥60	18 (4.5)
Gender	
Male	222 (55.1)
Female	181 (44.9)
Residence area	
Metropolitan area	329 (81.6)
The other	74 (18.4)
Education	
≤High school	100 (24.8)
College	254 (63)
≥Graduate school	49 (12.2)
Disease excluding IBD	
Digestive disease	266 (66)
Cardiovascular disease	15 (3.7)
Hypertension	16 (4)
Diabetes	10 (2.5)
Hyperlipidemia	16 (4)
Musculoskeletal disorders	12 (3)
Kidney disease	8 (2)
Respiratory diseases	14 (3.5)
Others	117 (29)
Number of concurrent diseases excluding IBD[c]	
None	44 (10.9)
1	281 (69.7)
2	67 (16.6)
3	6 (1.5)
≥4	5 (1.2)

[a]IBD: inflammatory bowel disease.

[b]Mean 36.24 (SD 11.75).

[c]Mean 1.13 (SD 0.71).

Step 3-2: Constructive Validity

Construct validity was verified with the fourth draft of 40 questions, and at this time, 12 questions regarding user characteristics and digital readiness-related characteristics that were decided to be included in the final questions were excluded from the analysis. Initially, the correlation coefficient between each item and the overall score was analyzed. Further, 2 items (familiarity 1 and usability 1) with correlation coefficients below 0.3 were excluded [39]. The Kaiser-Meyer-Olkin measure yielded a high value of 0.92, and the Bartlett test of sphericity was statistically significant (χ^2_{378}=6940.63, P<.001), indicating suitability for factor analysis. All factor communalities were above 0.3. In the pattern matrix, items with factor loadings above 0.5 on 2 factors and those uniquely loaded on one factor

XSL•FO
RenderX

were identified. Factor analysis was iteratively performed, removing one item at a time [40]. Ultimately, 24 items loaded on 4 factors. The Kaiser-Meyer-Olkin value for the final items was 0.91, and the Bartlett test of sphericity remained significant (χ^2_{276}= 6114.77, P<.001). The commonalities of the items were above 0.38, except for one item in factor 1. These 4 factors explained 65.05% of the total variance (Figure 2). Despite low communality, one item from factor 1 was retained after consideration by the research team due to its relevance to the scale. The domain of the digital health readiness scale was named factor 1 as "capability to use mobile services," factor 2 as "mHealth literacy," factor 3 as "digital health equity," and factor 4 as "perception of the importance of mHealth apps and devices" according to the content and characteristics of the loaded items.

Figure 2. Confirmatory factor analysis and reliability assessment: evaluation of the Mobile-Centered Digital Health Readiness Scale for patients with inflammatory bowel disease.

Items	Factor roading				Communality
	1	2	3	4	
F 2 (public services available on mobile)	0.76	0.24	0.06	0.01	0.63
F 3 (economic activities available on mobile)	0.84	0.13	0.09	0.04	0.74
F 4 (online maps available on mobile)	0.85	0.11	0.06	0.06	0.75
F 6 (bluetooth available on mobile)	0.78	0.20	0.01	0.12	0.66
U 2 (can communicate with medical staff via mobile)	0.51	0.48	0.08	0.16	0.53
U 3 (can understand health information from mobile)	0.61	0.43	0.13	0.26	0.64
U 8 (can access video content on mobile devices)	0.79	0.23	0.12	0.12	0.7
U 11 (understanding of health app information)	0.52	0.06	0.04	-0.11	0.29
E 1 (freedom to use mobile devices)	0.70	0.33	0.23	0.13	0.66
E 2 (stable internet connectivity)	0.76	0.19	0.21	0.12	0.67
U 4 (ability to discern accurate health information from mobile)	0.35	0.70	0.06	0.14	0.63
U 5 (awareness of health information or app availability via mobile)	0.14	0.83	0.04	0.19	0.74
U 6 (know how to use health information or apps on mobile)	0.24	0.80	0.11	0.17	0.74
U 9 (can assess the quality of health information or apps on mobile)	0.39	0.67	0.11	0.16	0.63
U 10 (can use mobile device information for health decisions)	0.47	0.57	0.06	0.23	0.59
E 5 (can aid in health management and information acquisition)	0.09	0.62	0.24	-0.09	0.46
E 8 (expand support for fair mobile access for vulnerable groups)	0.08	0.01	0.57	0.23	0.38
E 10 (importance of mobile health design reflecting user needs)	0.08	0.16	0.74	0.26	0.65
E 11 (importance of educating health care providers for equity)	0.10	0.11	0.84	0.04	0.72
E 12 (importance of educating consumers for equity)	0.07	0.18	0.80	0.14	0.69
E 13 (importance of educating health care developers for equity)	0.18	0.06	0.85	0.05	0.76
I 1 (importance of mobile health management functions)	0.15	0.19	0.20	0.81	0.76
I 2 (importance of new mobile health information)	0.04	0.09	0.24	0.85	0.8
I 3 (belief in the benefits of mobile health management)	0.08	0.22	0.21	0.83	0.79
Initial eigen value	9.19	3.09	1.91	1.42	
Cumulative variance (%)	38.27	51.15	59.12	65.05	
Cronbach α	0.91	0.86	0.87	0.84	

Step 3-3: Reliability

The internal consistency of the final scale was evaluated using Cronbach α, as shown in Figure 2. Cronbach α for each factor ranged from 0.84 to 0.91, indicating established reliability and acceptability of the newly developed scale [26].

This scale consists of 36 items, including 24 confirmed through reliability and validity verification, and 12 items (user characteristics, characteristics related to digital readiness) that were retained in the final scale during the scale development process (Textbox 1 and Multimedia Appendix 1). Excluding the 12 items measuring user characteristics and characteristics related to digital readiness, the remaining 24 items were structured on a 5-point Likert scale. Responses range from 1 ("not at all") to 5 ("very much"). Additionally, 5 items dedicated to evaluating digital health readiness focus on aspects of digital health accessibility. A total of 3 items, presented as multiple-choice questions, inquire about health management, information acquisition methods, and familiarity with mobile devices used for digital health services. Further, 2 items probe the willingness to pay for mHealth care services or purchase an mHealth care device. Moreover, 7 items gather data on user demographics, including age, gender, residence, occupation, education, subjective health status, and diagnosed diseases. The scale encompasses 6 domains. The total score is computed as the mean of the scores across 4 domains, excluding the user's characteristics and characteristics related to digital readiness. Scores range from 1 to 5, with higher scores indicating greater readiness for digital health use.

XSL•FO
RenderX

Textbox 1. Domains and summary of the Mobile-Centered Digital Health Readiness: Health Literacy and Equity Scale.

Mobile services capability (10 items)

- This domain evaluates the respondent's knowledge and proficiency in mobile device use and the extent of their integration into daily life.

Mobile health (mHealth) literacy understand and use mHealth apps and devices (6 items)

- This section assesses the respondent's ability to comprehend and use information acquired through mobile health apps and devices.

Perception of the importance of mHealth apps and devices (3 items)

- This domain gauges the respondent's perceived significance of mobile health apps and devices in health care management.

Digital health equity (5 items)

- This domain focuses on the environmental and resource factors influencing digital health accessibility and competency. It evaluates the respondent's access to and capability with digital health care resources.

Characteristics related to digital readiness (5 items)

- This domain identifies critical characteristics associated with the respondent's digital readiness. It includes methods of acquiring health information, familiarity with mobile devices for health information, experience in using digital health care services, and willingness to invest in and pay for digital health services. This domain assesses the respondent's accessibility to digital health resources.

User's characteristics (7 items)

- This section captures the respondent's general demographic and socioeconomic characteristics pertinent to digital health readiness. It covered age, gender, area of residence, occupation, education, subjective health status, and diagnosed diseases.

Step 4: Development of an English Version of the Scale

We developed an English version of our questionnaire to widen this study's reach and accessibility. As a global language, English allows us to gather data on Digital Health Readiness from diverse populations, enhancing the applicability of our findings. Offering the questionnaire in Korean and English enables researchers worldwide to use it effectively, strengthening the global understanding of digital health readiness. This decision ensures our research's comprehensive and international relevance, aligning with our aim to adapt the tool for broader use beyond patients with IBD.

Step 4-1: Translation

The final scale completed in the previous step was translated into English by a qualified translator and researcher who majored in English and nursing. After being translated independently, any differences in translation were agreed upon through online communication.

Step 4-2: Experts Review

Bilingual researchers in English and Korean reviewed the translation of steps 4-1. In the process, we ensured that the translation applied to English speakers and corrected any expressions or cultural differences that might convey a different meaning. The English version of the scale was finally completed after a thorough review and revision by a nursing informatics professor in the United States whose native language is English (Multimedia Appendix 2).

Discussion

Principal Findings

In the rapidly evolving field of digital health care, accessibility to digital health services is an essential determinant of health outcomes [41]. Consequently, the concept of digital inclusion, especially for vulnerable groups, is gaining prominence in the digital health care landscape, necessitating a thorough assessment and integration of these considerations into the design and delivery of digital health services [6]. The development of the *Mobile-Centered Digital Health Readiness: Health Literacy and Equity Scale* (mDiHERS) for patients is a pioneering effort to quantify the readiness and capability of patients to engage with digital health services. This scale is particularly relevant given the chronic nature of IBD, which necessitates ongoing and continuous management and the potential for digital tools to enhance patient autonomy and care significantly. The mDiHERS addresses a critical gap in digital health literature by providing a validated tool that can assess patients' digital access, literacy, and equity, which are essential for the effective use of digital health services.

In the mDiHERS, some items assess users' ability to navigate and interpret digital interfaces. This ability is increasingly essential to affording opportunities to increase reach and engagement in the digital health care service [42]. Including items requiring users to interact with actual digital device screens—such as smartphones and wearable devices—addresses a vital component of digital health literacy: the technical proficiency necessary for engaging with health apps and platforms. This approach means that the mDiHERS not only captures the subjective confidence of users in their digital capabilities but also provides an objective measure of their practical skills. In other words, the mDiHERS acknowledges this by ensuring its assessment criteria encompass users' perceived and actual abilities to manage their health digitally. This supports prior research that objective tests may be better suited to assessing an individual's skills [10]. This is particularly pertinent for patients who rely on digital health monitoring and management tools, as it empowers them to become active,

XSL•FO
RenderX

informed participants in their health care journeys. By emphasizing both confidence and competence, the mDiHERS aligns with the goals of digital health initiatives which aim to enhance patient autonomy and improve health outcomes through technology. This focus reflects the broader objectives within the digital health ecosystem, where patient empowerment and the democratization of health information are paramount.

The findings of this study underscore the importance of considering patient-specific factors when assessing digital readiness. The mDiHERS evaluates an individual's ability and familiarity with using digital health services and their level of digital readiness, including the concept of equity, which has recently become necessary with the emergence of digital health services [7]. The tailored approach of the mDiHERS, focusing on patients with IBD, allows for a nuanced understanding of the challenges and opportunities within this group. It is evident that while digital health services offer immense potential, their benefits are not uniformly accessible. The mDiHERS can thus serve as a diagnostic tool to identify areas where interventions are needed to improve digital health engagement and patient outcomes. If digital health services regularly monitor digital readiness, they can easily identify users who need additional support. Furthermore, it can also effectively evaluate interventions' effectiveness at all research stages. Therefore, the needs of digital health service users will be appropriately addressed.

Furthermore, assessing digital device usage skills should focus on more than just the technical aspects. As items in sections E and F of the mDiHERS exemplify, factors such as a patient's cultural and educational background can influence their ability to use digital devices, necessitating a comprehensive approach incorporating these variables. Educational support and interventions tailored to patients from cultural and educational environments with potentially lower digital device usage skills will significantly enhance digital health literacy, accessibility, and, ultimately, digital health equity [43]. The mDiHERS underscores the need for an assessment that evaluates technical skills while also considering the patient's overall background and circumstances. This approach contributes to developing more inclusive and customized support strategies for effectively using digital health services, ensuring all patients can benefit from advancements in digital health care.

The mDiHERS has the potential to be adapted for use in other patient populations and health conditions. Its application could lead to more personalized health care, where digital tools are used to their full potential to support health consumer's care. However, the scale also highlights the need for health care systems to address the digital divide and ensure that all patients, particularly those with chronic conditions, have the necessary skills and resources to benefit from digital health innovations. The evolution of digital health services will likely present new challenges and opportunities, making the continuous refinement and application of tools such as the mDiHERS essential for achieving equitable and effective health care delivery.

Limitations and Recommendation

This study used online surveys and interviews. Participants with low digital literacy may have found participating difficult, possibly excluding them from the survey. Consequently, individuals with higher digital literacy might be overrepresented, leading to self-selection bias. This can hinder generalizing the findings to the entire population and can affect study results. To minimize this bias in future research, incorporating online educational interventions and paper-based surveys can help include those with low digital literacy. Promoting the study on various platforms can also improve accessibility. Recruiting a balanced number of participants across different age groups can enhance accuracy. IBD primarily affects individuals in their teens to thirties, making it challenging to have evenly distributed participants. Future research should aim to recruit a similar number of participants across different age groups to validate the tool thoroughly and improve the representativeness of the results.

Further, it is important to note that while we aimed to gather a broad range of literature and frameworks, our approach did not use a systematic review method, which may be considered a limitation. We acknowledge this limitation in our paper as it affects the comprehensiveness of our findings. Furthermore, due to the resource constraints of our study, it was only feasible to adhere partially to the WHO translation guidelines for tools. However, we suggest that future translations of the tool into other languages should follow the WHO's translation guidelines for tools. Lastly, future studies are proposed to measure patient or health consumer digital readiness using the mDiHERS developed in this study to advance this research field. Subsequently, they should design and evaluate digital health services considering these results. This would facilitate the validation of the correlation between outcomes derived from the tool and the actual usage and effectiveness of digital health services.

Conclusions

The mDiHERS developed for this study measures patients' readiness and ability to use digital health services. It is particularly useful for individuals and groups requiring continuous health management, such as IBD. mDiHERS assesses digital accessibility, literacy, and equity factors, contributing to the effective use of digital health services to enhance accessibility. The development and validation of the mDiHERS highlight the importance of patients' confidence and competence in managing their health digitally. Continuous improvements are necessary to ensure that all patients can benefit from digital health care.

Acknowledgments

We would like to express our gratitude to Seung Gon Kim for his assistance in our research. This research was supported by the National Research Foundation of Korea (NRF-2021R1C1C1004505). The funding for this project was secured by the corresponding author, JL. The text, tables, and figures of this paper were entirely generated by the authors without any primary or significant

XSL•FO
RenderX

contributions from generative artificial intelligence. However, ChatGPT was used to identify and correct grammatical errors in the English text written by the authors.

Data Availability

Data from this study are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mobile-Centered Digital Health Readiness: Digital Health Literacy and Equity Scale (mDiHERS) Korean version.
[PDF File (Adobe PDF File), 455 KB - jmir_v26i1e58497_app1.pdf]

Multimedia Appendix 2
Mobile-Centered Digital Health Readiness: Digital Health Literacy and Equity Scale (mDiHERS) English version.
[PDF File (Adobe PDF File), 440 KB - jmir_v26i1e58497_app2.pdf]

References

1. What is digital health? U.S. Food and Drug Administration. 2022 Sep 22. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2024-06-28]
2. Istepanian RSH, AlAnzi T. Mobile health (m-health): evidence-based progress or scientific retrogression. Biomed Inform Technol 2019:717-733.
3. Maier E, Reimer U, Wickramasinghe N. Digital healthcare services. Electron Mark 2021;31(4):743-746 [FREE Full text] [doi: 10.1007/s12525-021-00513-z] [Medline: 35602115]
4. Condry M, Quan XI. Digital health innovation, informatics opportunity, and challenges. IEEE Eng Manag Rev 2021;49(2):81-88. [doi: 10.1109/emr.2021.3054330]
5. Rodriguez JA, Shachar C, Bates DW. Digital inclusion as health care—supporting health care equity with digital-infrastructure initiatives. N Engl J Med 2022;386(12):1101-1103. [doi: 10.1056/NEJMp2115646] [Medline: 35302722]
6. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
7. Scherrenberg M, Falter M, Kaihara T, Xu L, van Leunen M, Kemps H, et al. Development and internal validation of the digital health readiness questionnaire: prospective single-center survey study. J Med Internet Res 2023;25:e41615 [FREE Full text] [doi: 10.2196/41615] [Medline: 36897627]
8. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
9. Lee SD, Tsai T, Tsai Y. Accuracy in self-reported health literacy screening: a difference between men and women in Taiwan. BMJ Open 2013;3(11):e002928 [FREE Full text] [doi: 10.1136/bmjopen-2013-002928] [Medline: 24259386]
10. Nguyen TH, Paasche-Orlow MK, McCormack LA. The state of the science of health literacy measurement. Stud Health Technol Inf 2017;240:17-33 [FREE Full text] [Medline: 28972507]
11. Jackson DN, Trivedi N, Baur C. Re-Prioritizing Digital Health and Health Literacy in to Affect Health Equity. Health Commun 2021 Sep;36(10):1155-1162 [FREE Full text] [doi: 10.1080/10410236.2020.1748828] [Medline: 32354233]
12. Fridsma DB. AMIA response to FCC notice on accelerating broadband health tech availability. American Medical Informatics Association. 2017 May 24. URL: https://brand.amia.org/m/3a04808bf2cf6591/original/AMIA-Responds-to-FCC-Notice-on-Broadband-Enabled-Health-Technology.pdf
13. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
14. Lobo A, Mateus S. Validity and reliability of an equity in health care scale. SAGE Open 2013;3(4):2158244013506716. [doi: 10.1177/2158244013506716]
15. Pauly B, Martin W, Perkin K, van Roode T, Kwan A, Patterson T, et al. Critical considerations for the practical utility of health equity tools: a concept mapping study. Int J Equity Health 2018;17(1):48 [FREE Full text] [doi: 10.1186/s12939-018-0764-6] [Medline: 29688855]
16. Zimmerman FJ. A robust health equity metric. Public Health 2019;175:68-78. [doi: 10.1016/j.puhe.2019.06.008] [Medline: 31404717]
17. Hosseinpoor AR, Nambiar D, Schlotheuber A, Reidpath D, Ross Z. Health Equity Assessment Toolkit (HEAT): software for exploring and comparing health inequalities in countries. BMC Med Res Methodol 2016;16(1):141 [FREE Full text] [doi: 10.1186/s12874-016-0229-9] [Medline: 27760520]

XSL•FO
RenderX

18. Knowles SR, Graff LA, Wilding H, Hewitt C, Keefer L, Mikocka-Walus A. Quality of life in inflammatory bowel disease: a systematic review and meta-analyses—part I. Inflamm Bowel Dis 2018;24(4):742-751. [doi: 10.1093/ibd/izx100] [Medline: 29562277]
19. Parigi TL, D'Amico F, Abreu MT, Dignass A, Dotan I, Magro F, et al. Difficult-to-treat inflammatory bowel disease: results from an international consensus meeting. Lancet Gastroenterol Hepatol 2023;8(9):853-859. [doi: 10.1016/S2468-1253(23)00154-1] [Medline: 37423233]
20. Noser AE, Lancaster BD, Hommel KA, Roberts CM, King JA, Alt E, et al. Use of behavior change techniques and quality of commercially available inflammatory bowel disease apps. Dig Dis Sci 2023;68(7):2908-2920. [doi: 10.1007/s10620-023-07884-7] [Medline: 36933116]
21. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, Becx MC, Maljaars JP, Cilissen M, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. Lancet 2017;390(10098):959-968. [doi: 10.1016/S0140-6736(17)31327-2] [Medline: 28716313]
22. Jannati N, Salehinejad S, Kuenzig M, Peña-Sánchez JN. Review and content analysis of mobile apps for inflammatory bowel disease management using the Mobile Application Rating Scale (MARS): systematic search in app stores. Int J Med Inform 2023;180:105249. [doi: 10.1016/j.ijmedinf.2023.105249] [Medline: 37857167]
23. Iizawa M, Hirose L, Nunotani M, Nakashoji M, Tairaka A, Fernandez J. A systematic review of self-management interventions for patients with inflammatory bowel disease. Inflamm Intest Dis 2023;8(1):1-12 [FREE Full text] [doi: 10.1159/000530021] [Medline: 37404383]
24. Lee J, Schnall R. Validity and reliability of the Korean version of the health information technology usability evaluation scale: psychometric evaluation. JMIR Med Inform 2022;10(1):e28621 [FREE Full text] [doi: 10.2196/28621] [Medline: 35072630]
25. Schwartz BD, Kim HS. An expansion and elaboration of the hybrid model of concept development. Concept Dev Nurs Found Tech Appl 2000;2:161-192.
26. Nunnally J. Psychometric Theory. 3rd ed. New York, NY: McGraw-Hill; 1994.
27. Kerr C, Nixon A, Wild D. Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Rev Pharmacoecon Outcomes Res 2010;10(3):269-281. [doi: 10.1586/erp.10.30] [Medline: 20545592]
28. Process of translation and adaptation of instruments. WHO. 2018. URL: https://www.coursehero.com/file/30372721/WHO-Process-of-translation-and-adaptation-of-instrumentspdf/ [accessed 2023-12-14]
29. Kang JM, Song H, Kim HS. Development and application of the diagnostic instruments for measuring digital literacy in the smart society. J Korean Assoc Reg Inf Soc 2014;17(3):143-173.
30. Choi Y, Yun H, Lee CC. Developing indices for smartphone usage: smartphone technology quotient. Entrue J Inf Technol 2013;12(3):87-104.
31. Kim DM, Lee TW. A study on development of the digital competence measurement tool for university student. J Korea Soc Comput Inf 2019;24(12):191-199.
32. Rodríguez-de-Dios I, Igartua J, González-Vázquez A. Development and validation of a digital literacy scale for teenagers. 2016 Presented at: In: Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality; 2016 November 02; New York, NY, United States p. 323-330.
33. Muehlensiepen F, Marquardt W, Knitza J, Hueber A, Welcker M. AB1348-HPR telemedicine in rheumatology: acceptance by patients and medical practitioners in Germany (telerheumabb)—interim analysis. Ann Rheum Dis 2020;79(Suppl 1):1962. [doi: 10.1136/annrheumdis-2020-eular.1865]
34. Antonio M, Petrovskaya O. Towards developing an ehealth equity conceptual framework. Stud Health Technol Inf 2019;257:24-30. [Medline: 30741167]
35. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]
36. Were M, Sinha C, Catalani C. A systematic approach to equity assessment for digital health interventions: case example of mobile personal health records. J Am Med Inform Assoc 2019;26(8-9):884-890 [FREE Full text] [doi: 10.1093/jamia/ocz071] [Medline: 31188438]
37. Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35(6):382-385. [Medline: 3640358]
38. Beck CT. Qualitative research: the evaluation of its credibility, fittingness, and auditability. West J Nurs Res 1993;15(2):263-266. [doi: 10.1177/019394599301500212] [Medline: 8470379]
39. Snizek WE. Hall's Professionalism Scale: an empirical reassessment. Am Sociol Rev 1972;37(1):109-114. [doi: 10.2307/2093498]
40. Tabachnick BG, Fidell LS, Ullman JB. Using Multivariate Statistics. 7th ed. Boston, MA: Pearson; 2019.
41. Lyles CR, Wachter RM, Sarkar U. Focusing on digital health equity. JAMA 2021;326(18):1795-1796. [doi: 10.1001/jama.2021.18459] [Medline: 34677577]
42. Kemp E, Trigg J, Beatty L, Christensen C, Dhillon HM, Maeder A, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Austr 2021;32 Suppl 1:104-114. [doi: 10.1002/hpja.387] [Medline: 32681656]

XSL•FO
RenderX

43. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]

Abbreviations

CVI: content validity index
FGI: focus group interview
IBD: inflammatory bowel disease
mDiHERS: Mobile-Centered Digital Health Readiness: Health Literacy and Equity Scale
mHealth: mobile health
WHO: World Health Organization

Edited by A Mavragani; submitted 17.03.24; peer-reviewed by D Yoon, KL Mauco; comments to author 10.04.24; revised version received 01.05.24; accepted 27.06.24; published 13.08.24.

Please cite as:
Kim H, Schnall R, Yoon N, Koh SJ, Lee J, Cheon JH
Development and Validation of a Mobile-Centered Digital Health Readiness Scale (mDiHERS): Health Literacy and Equity Scale
J Med Internet Res 2024;26:e58497
URL: https://www.jmir.org/2024/1/e58497
doi:10.2196/58497
PMID:

©Hana Kim, Rebecca Schnall, Nagyeom Yoon, Seong-Joon Koh, Jisan Lee, Jae Hee Cheon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating Expert-Layperson Agreement in Identifying Jargon Terms in Electronic Health Record Notes: Observational Study

John P Lalor[1], MSc, PhD; David A Levy[2], MD, MPP; Harmon S Jordan[3], MSc, ScD; Wen Hu[2,4], MS; Jenni Kim Smirnova[5], MSc; Hong Yu[2,4,6], MA, MS, PhD

[1]Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, Notre Dame, IN, United States

[2]Center for Biomedical and Health Research in Data Sciences, Miner School of Computer and Information Sciences, University of Massachusetts Lowell, Lowell, MA, United States

[3]Tufts University School of Medicine, Boston, MA, United States

[4]Center for Healthcare Organization & Implementation Research, Veterans Affairs Bedford Healthcare System, Bedford, MA, United States

[5]Precision for Medicine, Flemington, NJ, United States

[6]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, United States

Corresponding Author:
Hong Yu, MA, MS, PhD
Center for Biomedical and Health Research in Data Sciences
Miner School of Computer and Information Sciences
University of Massachusetts Lowell
1 University Ave
Lowell, MA, 01854
United States
Phone: 1 978 934 3620
Email: Hong_Yu@uml.edu

Abstract

Background: Studies have shown that patients have difficulty understanding medical jargon in electronic health record (EHR) notes, particularly patients with low health literacy. In creating the NoteAid dictionary of medical jargon for patients, a panel of medical experts selected terms they perceived as needing definitions for patients.

Objective: This study aims to determine whether experts and laypeople agree on what constitutes medical jargon.

Methods: Using an observational study design, we compared the ability of medical experts and laypeople to identify medical jargon in EHR notes. The laypeople were recruited from Amazon Mechanical Turk. Participants were shown 20 sentences from EHR notes, which contained 325 potential jargon terms as identified by the medical experts. We collected demographic information about the laypeople's age, sex, race or ethnicity, education, native language, and health literacy. Health literacy was measured with the Single Item Literacy Screener. Our evaluation metrics were the proportion of terms rated as jargon, sensitivity, specificity, Fleiss κ for agreement among medical experts and among laypeople, and the Kendall rank correlation statistic between the medical experts and laypeople. We performed subgroup analyses by layperson characteristics. We fit a beta regression model with a logit link to examine the association between layperson characteristics and whether a term was classified as jargon.

Results: The average proportion of terms identified as jargon by the medical experts was 59% (1150/1950, 95% CI 56.1%-61.8%), and the average proportion of terms identified as jargon by the laypeople overall was 25.6% (22,480/87,750, 95% CI 25%-26.2%). There was good agreement among medical experts (Fleiss κ=0.781, 95% CI 0.753-0.809) and fair agreement among laypeople (Fleiss κ=0.590, 95% CI 0.589-0.591). The beta regression model had a pseudo-R^2 of 0.071, indicating that demographic characteristics explained very little of the variability in the proportion of terms identified as jargon by laypeople. Using laypeople's identification of jargon as the gold standard, the medical experts had high sensitivity (91.7%, 95% CI 90.1%-93.3%) and specificity (88.2%, 95% CI 86%-90.5%) in identifying jargon terms.

Conclusions: To ensure coverage of possible jargon terms, the medical experts were loose in selecting terms for inclusion. Fair agreement among laypersons shows that this is needed, as there is a variety of opinions among laypersons about what is considered jargon. We showed that medical experts could accurately identify jargon terms for annotation that would be useful for laypeople.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e49704) doi:10.2196/49704

KEYWORDS

expert-layperson agreement; medical jargon; jargon identification; EHR; electronic health record notes; crowdsourcing; clinical notes

Introduction

In recent years, federal regulations such as the Cures Act and the 2022 Office of the National Coordinator for Health Information's Cures Act Final Rule [1] have mandated that health care providers grant patients full access to their electronic health records (EHRs) through patient portals [2]. Patient access to their EHRs represents a new communication channel between doctors and patients that can facilitate faster communication of test results, medication lists, and other information [3]. These regulations include the EHR notes as part of the data that must be provided [4]. Access to EHRs can enhance patients' understanding of their disease [5], improve communication between patients and their care providers [6,7], improve medication adherence [8], and reduce health care costs [4,9-16]. However, EHRs contain medical jargon that patients, especially those with low health literacy [17], may not understand [18]. The following example is typical of the text that patients see in their EHR notes: "From a GI standpoint, we recommend to proceed with bariatric surgery. However, he will need to continue daily PPI administration to maximize acid reduction." Without significant medical knowledge, it is challenging to understand the meanings of jargon terms such as "GI" (gastrointestinal), "bariatric," and "PPI" (proton pump inhibitor).

Innovations are needed to support patients' use of EHR notes and translate them into language that is easier for them to understand [19]. We found in our previous work that defining jargon and the readability of those definitions were positively associated with improved EHR note comprehension [20,21].

Jargon in EHR Notes

According to the Cambridge Dictionary, jargon is defined as "language used by a particular group of people, especially in their work, and which most other people do not understand" [22]. Technical jargon occurs across disciplines and reflects the amount of specialist knowledge in a field [23]. Jargon can aid in communication by succinctly describing complex concepts. However, it can also impede communication with and comprehension by those unfamiliar with the language [24-26]. For accurate comprehension, a reader must be familiar with 95%-98% of the words in a text [27-30].

Medical records contain large amounts of jargon and abbreviations, and recent work has shown that patients cannot consistently comprehend the meaning of jargon terms, including abbreviations and acronyms [31-33]. Recent studies involving over 10,000 patients showed that allowing patients to read the clinical notes in their medical records confused some, especially those in vulnerable groups such as patients with lower literacy or lower income [4,34]. Similar work on patient portals shows that patients struggle to make use of their access [35-37].

There has been work in the literature on identifying and defining jargon in various fields, from medical informatics to scientific communication [23,30,38,39]. A significant gap in the previous work is that jargon identification is examined from an expert point of view. In medicine, consumer health vocabularies bridge the gap between patient and physician terminology. Consumer health vocabularies are typically collected by analyzing user search queries [40,41], web-based community forums [42-46], and patient-provider communications [41,47]. However, there is little work analyzing what patients recognize as jargon in the context of EHR notes. Furthermore, the literature has not examined if what patients recognize as jargon differs with demographic groups.

Recently, Pitt and Hendrickson [48] proposed a classification system for 7 types of medical jargon: technical terminology, alphabet soup, medical vernacular, medicalized English, unnecessary synonyms, euphemisms, and judgmental jargon. This categorization is broad enough to cover domain-specific technical terms (eg, "myocardial infarction" or "ambulating") as well as more common words with specific medical meanings (eg, "progressing" or "positive") [31,48].

NoteAid

To assist patients with understanding medical notes, we are developing NoteAid [49], a natural language processing system that links medical jargon in EHR notes to definitions and addresses context-dependent meanings (eg, the abbreviation MAP could refer to muscle action potential or mean arterial pressure). A team of interdisciplinary experts identifies jargon from a corpus of EHRs and writes definitions for them. Specifically, NoteAid definitions of identified terms are written for a fourth- to seventh-grade reading level [50]. For example, in the EHR sentence referenced above, NoteAid would provide the following definition: bariatric surgery: Surgery on the stomach and intestines for weight loss. We have written definitions for approximately 30,000 jargon terms.

Our operational use of the term "medical jargon" in this study is patient-focused, not clinician-focused. NoteAid uses a 2-stage process for identifying jargon. First, the software generates a word frequency list from the EHR corpus. Starting from the most frequent word on the list, it presents sentences that contain the word to the definition writer. The definition writer reviews the sentence and decides which terms are jargon, following a set of guidelines in making this decision (more details in Multimedia Appendix 1). Examples of the guidelines include a medical term that would not be recognized by a layperson with a fourth-grade education (eg, duodenum); a word that has a different meaning in the medical context than in the lay context (eg, accommodate: when the eye changes focus from far to near). Still, the determination of what is and is not jargon is uncertain, as patients differ widely in their education, general level of health literacy, and experience with medical conditions.

In this observational study, we examined how well the NoteAid definition writers agreed with each other in identifying jargon

XSL•FO
RenderX

and how well they agreed with laypeople. We used Amazon Mechanical Turk (MTurk) workers as proxies for laypeople. We also investigated how jargon identification varied across different demographic subgroups. We expected that demographic subgroups typically associated with lower health literacy would identify more terms as jargon than other subgroups. These groups include older adults, certain race or ethnic groups, those with a high school education or less, those whose native language is not English, and those who score low on a health literacy screener [51].

Methods

We conducted an observational study to examine the agreement between the NoteAid definition writers and laypeople on what is considered medical jargon.

Data Source

The NoteAid dictionary used medical notes from the PittEHR database of deidentified inpatient medical records [52]. The records consist of emergency department notes, progress notes, consult notes, operative reports, radiology reports, pathology reports, and discharge summaries, all written by physicians. We randomly selected 20 sentences from the database. Sentences that contained only administrative data, contained fewer than 10 words, or were substantially similar to another selected sentence were not included. The NoteAid definition writers had not previously seen these sentences.

Identifying Terms for Annotation Task

The 20 sentences contained a total of 904 words. So as not to inflate the calculated agreement, we excluded from the analysis common words, which we defined as all conjunctions, pronouns, prepositions, numerals, articles, contractions, names of months, punctuation, and the 25 most common nouns, verbs, adjectives, and adverbs, including their plural forms. Terms that were repeated in a sentence were only counted once. Multiword terms were analyzed as single terms to avoid double counting. We considered multiword terms to be adjacent words that represented a distinct medical entity (examples include PR interval, internal capsule, and acute intermittent porphyria), terms that were routinely used together (examples include hemodynamically stable, status post, and past medical history), or terms that were modified by a minor word (examples include trace perihepatic fluid, mild mitral regurgitation, rare positive cells, and deep pelvis). The grouping of multiword terms was determined by 2 members of the research team after reaching a consensus. There were 325 potential jargon terms in the final analysis. We performed a second analysis in which only the common words were excluded, and there was no grouping of multiword terms. This process resulted in 549 potential jargon words in the analysis.

Data Collection

Data collection took place between August 2020 and April 2021. NoteAid definition writers and MTurk workers were shown the 20 sentences. MTurk workers were asked to identify those terms for which they did not know the definition. In this paper, we refer to these identified terms as “jargon.” The NoteAid definition writers were asked to identify terms that they considered to be jargon, that is, terms for which laypeople would not know the definition. MTurk workers were instructed not to consult any sources of information during the task. Interspersed among the 20 sentences were 3 attention-check questions to test whether the participant was paying attention. If a participant answered 2 of the 3 attention checks incorrectly, the participant’s responses were discarded, and the participant was replaced (however, the participant was not excluded from reentering the study).

Participant Recruitment

We recruited adult MTurk workers and collected demographic information about the workers’ age, sex, race or ethnicity, education, native language, and health literacy. We performed subgroup analyses based on MTurk worker characteristics. To evaluate health literacy, MTurk workers were screened with the Single Item Literacy Screener [51]. MTurk workers who worked in the health care field were excluded from the study. We also excluded MTurk workers with a previous approval rating below 95% in the MTurk platform.

MTurk workers were oversampled to obtain equal numbers of MTurk workers in each of the education subgroups and an equal number of non-White and White participants. We sampled 270 MTurk workers and 6 definition writers to complete the study instrument. The 6 definition writers were all experienced biomedical annotators with advanced degrees in medicine, nursing, biostatistics, and biomedical research.

Evaluation

Our evaluation metrics were the proportion of terms rated as jargon, sensitivity, specificity, and Fleiss κ for agreement among NoteAid definition writers and among MTurk workers. Wald CIs were calculated at 95%. We analyzed NoteAid definition writers individually and as a group.

Sensitivity and specificity measured the NoteAid definition writers’ ability to correctly discriminate between jargon and nonjargon, using the MTurk workers’ responses as the gold standard. Since all 270 MTurk workers did not agree on which terms were jargon, the cutoff number of MTurk workers for defining a term as jargon was chosen using the Youden index [53]. Youden index calculates sensitivities and specificities for all possible thresholds (ie, between jargon and nonjargon). The cutoff where the sum of sensitivity and specificity was highest was selected.

To determine whether the definition writers’ jargon selection was systematically different from the MTurk workers, we calculated the Kendall rank correlation statistic between the definition writers and MTurk workers [54]. We also analyzed the results by MTurk worker characteristics to determine if specific subpopulations differed from the definition writers in the terms they identified as jargon.

We fit a beta regression model with a logit link to examine the association between MTurk worker characteristics and whether a term was classified as jargon. We used beta regression because it does not assume that individual terms have the same probability of being rated as jargon (eg, joint vs articulation). Here, the proportion of terms an MTurk worker identified as

XSL•FO
RenderX

jargon served as the dependent variable, and the MTurk worker characteristics (sex, age group, race or ethnicity, education level, native language, and health literacy score) served as the independent variables. We assumed a linear relationship between the predictor and dependent variables, which we confirmed by checking the residuals. To check for the possibility of interactions among the independent variables, we explored adding different combinations of 2-way interaction terms. Models were evaluated using pseudo-R^2 [55] and the Akaike information criterion (AIC) values [56].

Ethical Considerations

We conducted this study with approval from the institutional review board at the University of Massachusetts Lowell (H00010621). Informed consent was obtained from MTurk workers, and they had the option to leave the task at any time. MTurk worker data were anonymized, and EHR data was deidentified. MTurk workers were paid US $3 for the task, which took an average of 20 minutes to complete.

Results

Table 1 shows the characteristics of the 270 MTurk workers. The average proportion of terms identified as jargon by the MTurk workers overall was 25.6% (95% CI 25%-26.2%, Table 1). This proportion compares to 59% for the NoteAid definition writers (95% CI 56.1%-61.8%). Among MTurk worker subgroups, the average proportion of terms identified as jargon ranged from 17.7% to 30.9% (Table 1). Participants with the lowest health literacy score (5) identified fewer terms as jargon (P=.15, n=2), while participants with the second-lowest health literacy score (4) identified more terms as jargon (P=.03, n=10). Both sample sizes were small.

XSL•FO
RenderX

Table 1. Amazon Mechanical Turk worker demographics and the average proportion of terms identified as jargon.

Layperson characteristic	Values, n (%)	Proportion of terms identified as jargon[a], %	95% CI (Wald)
Overall	270 (100)	25.6	25-26.2
Sex			
Female	149 (55.2)	25.6	24.8-26.4
Male	120 (44.4)	25.6	24.8-26.5
Other	1 (0.4)	23.7	14.2-33.2
Age (years)			
18-24	16 (5.9)	25.4	23-27.7
25-34	74 (27.4)	25.6	24.5-26.7
35-44	65 (24.1)	24.9	23.7-26.1
45-54	27 (10)	26.2	24.4-28
55-64	59 (21.9)	25.4	24.2-26.6
65-74	27 (10)	27.4	25.6-29.1
75 and over	2 (0.7)	27.5	21-34.1
Education			
High school degree or less	90 (33.3)	25.3	24.3-26.3
Bachelor's or associate's degree	90 (33.3)	25.3	24.3-26.3
Master's degree or higher	90 (33.3)	26.3	25.3-27.3
Race or ethnicity			
White	135 (50)	25.3	24.5-26.1
Asian	57 (21.1)	25.9	24.7-27.2
Black	46 (17)	25.8	24.4-27.2
Hispanic	15 (5.6)	27.3	24.9-29.7
Other	16 (5.9)	25	22.7-27.3
Native language			
English or bilingual	266 (98.5)	25.7	25.1-26.3
Non-English	4 (1.5)	21.6	16.8-26.4
Health literacy score[b]			
1 - Never	133 (49.3)	25.9	25.1-26.7
2 - Rarely	95 (35.2)	25	24.0-25.9
3 - Sometimes	30 (11.1)	25.3	23.6-27
4 - Often	10 (3.7)	30.9	28.1-33.8
5 - Always	2 (0.7)	17.7	10.7-24.7

[a]Out of 325 potential jargon terms.

[b]Response to the question: "How often do you need to have someone help you when you read instructions, pamphlets, or other written material from your doctor or pharmacy?" [50].

Model

The beta regression model had a pseudo-R^2 of 0.071 (Table 2), indicating that the amount of variability in the proportion of terms identified as jargon explained by the MTurk worker characteristics was very low. The only significant MTurk worker characteristic was being a nonnative English speaker (P=.02; see Table 2). Compared with native English speakers, nonnative English speakers had a 0.535 odds ratio (95% CI 0.321-0.892) of identifying terms as jargon, controlling for sex, age, race or ethnicity, education level, and health literacy score. However, there were only 4 nonnative English speakers in the sample. Of the various 2-way interactions examined, only the addition of an interaction between race or ethnicity and education yielded a good model fit, and the interaction term was not statistically significant. The addition of this interaction slightly increased the proportion of variability explained (pseudo-R^2=0.089) without meaningfully changing the AIC.

XSL•FO
RenderX

Table 2. Beta regression of proportion of Amazon Mechanical Turk workers marking terms as jargon.

Predictors	Participants, n	Coefficient	95% CI	*P* value
(Intercept)	__[a]	0.320	0.252-0.405	<.001
Sex				
Female (reference group)	149	—	—	—
Male	120	0.985	0.883-1.099	.79
Other	1	0.997	0.426-2.332	>.99
Age (years)				
18-24 (reference group)	16	—	—	—
25-34	74	1.067	0.844-1.348	.59
35-44	65	1.042	0.820-1.324	.74
45-54	27	1.162	0.887-1.522	.28
55-64	59	1.080	0.849-1.374	.53
65-74	27	1.187	0.908-1.551	.21
75 and older	2	1.262	0.684-2.325	.46
Education				
High school degree or less (reference group)	90	—	—	—
Bachelor's or associate's degree	90	0.968	0.852-1.100	.62
Master's degree or higher	90	1.027	0.902-1.169	.69
Race or ethnicity				
White (reference group)	135	—	—	—
Asian	57	1.07	0.933-1.226	.33
Black	46	1.015	0.8781.174	.84
Hispanic	15	1.157	0.922-1.451	.21
Other	16	0.923	0.727-1.172	.51
Native language				
English or bilingual (reference group)	266	—	—	—
Non-English	4	0.535	0.321-0.892	.02
Health literacy score[b]				
1 - Never (reference group)	133	—	—	—
2 – Rarely	95	0.949	0.846-1.064	.37
3 – Sometimes	30	1.002	0.845-1.188	.98
4 – Often	10	1.252	0.963-1.627	.09
5 – Always	2	0.571	0.294-1.113	.10
Observations	270	—	—	—
R^2	0.071	—	—	—

[a]Not applicable.

[b]Response to the question: "How often do you need to have someone help you when you read instructions, pamphlets, or other written material from your doctor or pharmacy?" [50].

Agreement Among MTurk Workers and Among Definition Writers

The proportion of terms identified as jargon by NoteAid definition writers ranged from 48.3% to 64.9%. The agreement among NoteAid definition writers was good (Fleiss κ=0.781, 95% CI 0.753-0.809, Table 3), with all agreeing on the categorization (jargon or not jargon) for 74.5% of terms. The proportion of terms identified as jargon by individual MTurk workers ranged from 1.2% to 57.5%. Agreement among MTurk workers was fair (Fleiss κ=0.590, 95% CI 0.589-0.591, Table

XSL•FO
RenderX

3). For 61.9% of terms, at least 90% of MTurk workers agreed on the categorization (jargon or not jargon).

Table 3. Agreement among NoteAid definition writers and among Amazon Mechanical Turk workers.

	Participants, n	Fleiss κ	95% CI
NoteAid definition writers	6	0.781	0.753-0.809
MTurk workers	270	0.590	0.589-0.591

Agreement Between Definition Writers and MTurk Workers

Our main measures of agreement between definition writers and MTurk workers were sensitivity and specificity. Using the Youden index, the highest combined sensitivity and specificity corresponded to at least 3 out of the 270 MTurk workers identifying a term as jargon. Using this cutoff, the mean sensitivity for the NoteAid definition writers was 91.7% (95% CI 90.1-93.3%), and the mean specificity was 88.2% (95% CI 86-90.5%; Table 4). These correspond to a false negative rate of 8.3% and a false positive rate of 11.8%, respectively. Among the individual NoteAid definition writers, sensitivity ranged from 78.1% to 95.8%, and specificity ranged from 79.7% to 94.7%.

Using the same threshold of 3 MTurk workers for classifying a term as jargon, we found that 59.1% of the terms would be classified as jargon, which is remarkably close to the average of 59% identified by definition writers.

The Kendall rank order correlation statistic was consistently high across all the MTurk worker characteristics, indicating no systematic differences in jargon identification between definition writers and different subpopulations of MTurk workers (Table 5).

All of the above analyses were repeated using single-word terms rather than multiword terms as the unit of analysis, and the results were not substantively different.

Table 4. Sensitivity and specificity of NoteAid definition writers.

Parameter	Value
Sensitivity[a], % (95% CI)	91.7 (90.1-93.3)
Specificity[b], % (95% CI)	88.2 (86.0-90.5)

[a]Sensitivity: Probability that definition writers identified a term that at least 3 MTurk workers had identified as jargon.

[b]Specificity: Probability that definition writers did not identify a term that less than 3 MTurk workers had identified as jargon.

XSL•FO
RenderX

Table 5. Kendall rank correlation coefficients comparing definition writer jargon ranks with Amazon Mechanical Turk worker jargon ranks.

Group	Participants, n	Kendall τ-b
All MTurk workers	270	0.743
Sex		
Female	149	0.763
Male	120	0.762
Other	1	0.609
Age (years)		
18-24	16	0.754
25-34	74	0.768
35-44	65	0.758
45-54	27	0.759
55-64	59	0.754
65-74	27	0.759
75 and older	2	0.689
Education		
High school degree or less	90	0.759
Bachelor's or associate's degree	90	0.763
Master's degree or higher	90	0.762
Race or ethnicity		
White	135	0.749
Asian	57	0.768
Black	46	0.772
Hispanic	15	0.782
Other	16	0.726
Health literacy score		
1	133	0.758
2	95	0.756
3	30	0.776
4	10	0.822
5	2	0.555
Primary language		
English	249	0.76
English, Arabic	1	0.753
English, Bengali	1	0.312
English, Chinese	8	0.757
English, Other	1	0.701
English, Spanish	5	0.746
English, Tamil	1	0.544
Other	1	0.704
Tagalog	2	0.713
Tamil	1	0.125

XSL•FO
RenderX

Discussion

Principal Findings

Jargon identification depends on the target audience. Patients differ widely in their education, general level of health literacy, and experience with medical conditions [21,57]. What should or should not be considered jargon is not often clear, as evidenced by recent attempts to formalize the notion in a classification system [48]. Using sensitivity and specificity as measures of agreement, we found good agreement between definition writers and MTurk workers. The high level of sensitivity indicates that definition writers were providing definitions for the terms that laypeople identify as jargon. Correspondingly, the high specificity indicates that the definition writers would not be expending time writing definitions for terms that laypeople do not identify as jargon. These findings validate one goal of the NoteAid project, which is to assist patients in understanding their EHR notes, even if they have limited health literacy.

The calculation of sensitivity and specificity for the NoteAid definition writers required a gold standard as to which terms are and are not jargon. However, since all 270 MTurk workers did not agree on which terms were jargon, we used the Youden index to determine the cutoff number of MTurk workers for defining a word as jargon. In this method, sensitivities and specificities were calculated for all possible cutoffs, and the cutoff whose summed sensitivity and specificity were highest was used. This technique gave the best balance between sensitivity and specificity, though it treats the cost of false positives and false negatives as the same. On average, we found that MTurk workers identified 25.6% (22,480/87,750) of terms as jargon, compared with 59% (1150/1950) of terms identified as jargon by the definition writers. However, this is not necessarily undesirable. The definition writers were identifying jargon terms for inclusion in the NoteAid dictionary; broad coverage in terms of inclusion is preferable in this context. Further, MTurk workers differed considerably as to which terms they considered jargon (Fleiss κ=0.590), so simply matching their average proportion of terms identified as jargon would exclude terms that some laypersons consider to be jargon. Therefore, we also evaluated definition writers on their sensitivity and specificity in identifying MTurk workers' jargon terms and found high agreement. These results suggest that personalized technologies such as NoteAid are needed, where specific results are identified from a wider database; a general consensus on what is or is not jargon may lead to the exclusion of terms that require definitions for a subset of the population.

Also, based on the MTurk worker health literacy scores, the average MTurk worker's reading level was likely higher than fourth grade, the target level for NoteAid, which is consistent with previous work [21]. Therefore, the 25.6% jargon proportion among MTurk workers likely underestimates the prevalence of jargon terms for NoteAid's target population.

Our beta regression model did not find differences between demographic subgroups in the proportion of terms identified as jargon. The only group that was significantly less likely to identify a term as jargon was the group for which English was not their native language. However, the small sample for this subgroup makes interpretation difficult. For the subgroups of adequate sample size, these results suggest that the selection of jargon by the NoteAid definition writers is sufficient.

In this work, MTurk workers and definition writers selected jargon terms from actual deidentified EHR notes. Most existing work looks at identifying jargon more generally, using data from web search logs or web forums. By using EHR notes for our task, the jargon identified should be more relevant for the downstream task of presenting definitions to patients looking at their own EHR notes.

Limitations and Future Work

There are several limitations to this work that can inform future research. First, we examined a relatively small number of passages in this experiment. A different selection of passages could have produced different jargon identifications among the MTurk workers and definition writers.

A second limitation is that we did not examine context dependency in this study. For example, the term "tips" can be either nonjargon (suggestions) or jargon (transjugular intrahepatic portosystemic shunt). The NoteAid system considers context-dependent meanings, and future studies could address this.

Another limitation concerns our lay population. While MTurk is often used for crowdsourced data collection [58,59], the demographic characteristics of the collected sample are not representative of the broader US population [21]. In particular, we are interested in jargon identification behaviors for individuals with low health literacy, while the MTurk workers in our study generally had high health literacy scores. Of note is that the proportion of terms identified as jargon by MTurk workers in the lowest health literacy groups diverged from that of the higher health literacy groups. However, these sample sizes were very small, making interpretation difficult. Future work should attempt to replicate the results in actual patients, as in our previous work [21].

Using our demographic and other variables, the beta regression model only explained a small proportion of the variability in jargon identification among the MTurk workers. It is possible that there are unmeasured variables that would account for additional variability, such as income, occupation, personal experience with health issues, or interest in health topics.

The NoteAid definition-writing process is a distributed task. Each definition writer works on separate notes to identify and define jargon terms. Therefore, a consistent understanding of what is or is not jargon is important to ensure consistent coverage across the notes. Future development of NoteAid can investigate automatic jargon identification for definition writers through natural language processing tools, using the corpus of human-identified jargon as training data may lead to a more effective automated system if those data are consistent. In particular, with the growing impact of large language models [60-64], there is an opportunity to leverage large language models to improve patient understanding of notes. Future work can also use user information such as health literacy level and demographic information to identify the most relevant jargon

XSL•FO
RenderX

terms and definitions, making the system even more personalized. Lastly, updating annotator instructions to be in line with established jargon classification frameworks can enforce consistency in labeling [48].

There are other applications for jargon identification, such as clinical trial regulations requiring plain language summaries [65]. This work can also inform jargon identification in other fields, such as law. A NoteAid-like tool for jargon identification and definition could define technical legal terms for lay individuals as they encounter them on the Web (eg, when reading a contract or terms of service agreement).

Conclusion

In this work, we have shown that trained definition writers could consistently select jargon terms for which laypeople need definitions. These results are encouraging for the continued development of NoteAid, and they have implications for other fields.

Acknowledgments

We are grateful to Weisong Liu, Brian E. Corner, Pamela Aselton, Liz Catherine S. Cory, Heather Keating, Jong Soo, and Fouzi Takelait for their invaluable contributions to this study. We thank the reviewers for their constructive feedback. This work was supported in part by a grant (R01LM012817) from the National Institutes of Health and by internal start-up funding from the University of Massachusetts Lowell. The content is solely the responsibility of the authors and does not represent the views of the National Institutes of Health or the University of Massachusetts Lowell.

Data Availability

The datasets generated during and/or analyzed during this study are available in a GitHub repository [66].

Authors' Contributions

HY conceptualized, supervised, and provided funding acquisition for the study. DAL and HSJ developed the methodologies. WH and JPL designed and implemented the data collection program. JKS, JPL, DAL, and HSJ performed data analysis. JPL, DAL, and HSJ wrote, reviewed, and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
NoteAid guidelines for identifying medical jargon.
[DOCX File , 20 KB - jmir_v26i1e49704_app1.docx]

References

1. Jeffcoat H. The 21st Century Cures Act final rule. American College of Surgeons. 2021. URL: https://www.facs.org/for-medical-professionals/news-publications/news-and-articles/bulletin/2021/09/the-21st-century-cures-act-final-rule/ [accessed 2024-03-07]
2. Blease C, Walker J, DesRoches CM, Delbanco T. New U.S. law mandates access to clinical notes: implications for patients and clinicians. Ann Intern Med 2021;174(1):101-102. [doi: 10.7326/M20-5370] [Medline: 33045176]
3. Steitz BD, Turer RW, Lin CT, MacDonald S, Salmi L, Wright A, et al. Perspectives of patients about immediate access to test results through an online patient portal. JAMA Netw Open 2023;6(3):e233572 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.3572] [Medline: 36939703]
4. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
5. Bronson DL, O'Meara K. The impact of shared medical records on smoking awareness and behavior in ambulatory care. J Gen Intern Med 1986;1(1):34-37. [doi: 10.1007/BF02596322] [Medline: 3772566]
6. Homer CS, Davis GK, Everitt LS. The introduction of a woman-held record into a hospital antenatal clinic: the bring your own records study. Aust N Z J Obstet Gynaecol 1999;39(1):54-57. [doi: 10.1111/j.1479-828x.1999.tb03445.x] [Medline: 10099751]
7. Elbourne D, Richardson M, Chalmers I, Waterhouse I, Holt E. The Newbury Maternity Care Study: a randomized controlled trial to assess a policy of women holding their own obstetric records. Br J Obstet Gynaecol 1987;94(7):612-619. [doi: 10.1111/j.1471-0528.1987.tb03165.x] [Medline: 3304403]
8. Assiri G. The impact of patient access to their electronic health record on medication management safety: a narrative review. Saudi Pharm J 2022;30(3):185-194 [FREE Full text] [doi: 10.1016/j.jsps.2022.01.001] [Medline: 35498224]

XSL•FO
RenderX

9. Baldry M, Cheal C, Fisher B, Gillett M, Huet V. Giving patients their own records in general practice: experience of patients and staff. Br Med J (Clin Res Ed) 1986;292(6520):596-598 [FREE Full text] [doi: 10.1136/bmj.292.6520.596] [Medline: 3081187]
10. Cimino JJ, Patel VL, Kushniruk AW. The Patient Clinical Information System (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records. Int J Med Inform 2002;68(1-3):113-127. [doi: 10.1016/s1386-5056(02)00070-9] [Medline: 12467796]
11. DeWalt DA, Malone RM, Bryant ME, Kosnar MC, Corr KE, Rothman RL, et al. A heart failure self-management program for patients of all literacy levels: a randomized, controlled trial [ISRCTN11535170]. BMC Health Serv Res 2006;6:30 [FREE Full text] [doi: 10.1186/1472-6963-6-30] [Medline: 16533388]
12. Hibbard JH, Greene J, Overton V. Patients with lower activation associated with higher costs; delivery systems should know their patients' 'scores'. Health Aff (Millwood) 2013;32(2):216-222. [doi: 10.1377/hlthaff.2012.1064] [Medline: 23381513]
13. Meltsner M. A patient's view of OpenNotes. Ann Intern Med 2012;157(7):523-524 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00012] [Medline: 23027321]
14. Schillinger D, Handley M, Wang F, Hammer H. Effects of self-management support on structure, process, and outcomes among vulnerable patients with diabetes: a three-arm practical clinical trial. Diabetes Care 2009;32(4):559-566 [FREE Full text] [doi: 10.2337/dc08-0787] [Medline: 19131469]
15. Seitz JF, Ward A, Dobbs WH. Granting patients access to records: the impact of the Privacy Act at a federal hospital. Hosp Community Psychiatry 1978;29(5):288-289. [doi: 10.1176/ps.29.5.288] [Medline: 640590]
16. Winkelman WJ, Leonard KJ, Rossos PG. Patient-perceived usefulness of online electronic medical records: employing grounded theory in the development of information and communication technologies for use by patients living with chronic illness. J Am Med Inform Assoc 2005;12(3):306-314 [FREE Full text] [doi: 10.1197/jamia.M1712] [Medline: 15684128]
17. Mahadevan R. Health literacy fact sheets.: Center for Health Care Strategy; 2013. URL: https://www.chcs.org/resource/health-literacy-fact-sheets/ [accessed 2024-01-10]
18. Sheikh H, Brezar A, Dzwonek A, Yau L, Calder LA. Patient understanding of discharge instructions in the emergency department: do different patients need different approaches? Int J Emerg Med 2018;11(1):5 [FREE Full text] [doi: 10.1186/s12245-018-0164-0] [Medline: 29423767]
19. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc 2022;29(4):585-591 [FREE Full text] [doi: 10.1093/jamia/ocac006] [Medline: 35190824]
20. Lalor JP, Woolf B, Yu H. Improving electronic health record note comprehension with NoteAid: randomized trial of electronic health record note comprehension interventions with crowdsourced workers. J Med Internet Res 2019;21(1):e10793 [FREE Full text] [doi: 10.2196/10793] [Medline: 30664453]
21. Lalor JP, Hu W, Tran M, Wu H, Mazor KM, Yu H. Evaluating the effectiveness of NoteAid in a community hospital setting: randomized trial of electronic health record note comprehension interventions with patients. J Med Internet Res 2021;23(5):e26354 [FREE Full text] [doi: 10.2196/26354] [Medline: 33983124]
22. Jargon. Cambridge Dictionary. URL: https://dictionary.cambridge.org/us/dictionary/english/jargon [accessed 2024-07-02]
23. Chung TM, Nation P. Identifying technical vocabulary. System 2004;32(2):251-263. [doi: 10.1016/j.system.2003.11.008]
24. Baram-Tsabari A, Lewenstein BV. An instrument for assessing scientists' written skills in public communication of science. Sci Commun 2012;35(1):56-85. [doi: 10.1177/1075547012440634]
25. Allen KA, Charpentier V, Hendrickson MA, Kessler M, Gotlieb R, Marmet J, et al. Jargon be gone - patient preference in doctor communication. J Patient Exp 2023;10:23743735231158942 [FREE Full text] [doi: 10.1177/23743735231158942] [Medline: 36873911]
26. Miller AN, Bharathan A, Duvuuri VNS, Navas V, Luceno L, Zraick R, et al. Use of seven types of medical jargon by male and female primary care providers at a university health center. Patient Educ Couns 2022;105(5):1261-1267. [doi: 10.1016/j.pec.2021.08.018] [Medline: 34489148]
27. Hsueh-Chao MH, Nation P. Unknown vocabulary density and reading comprehension. Read Foreign Lang 2000;13:403-430. [doi: 10.26686/wgtn.12560354]
28. Laufer B, Ravenhorst-Kalovski G. Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. Read Foreign Lang 2010;22(1):15-30 [FREE Full text]
29. Wen J, Yi L. Comparing lay summaries to scientific abstracts for readability and jargon use: a case report. Scientometrics 2023;128(10):5791-5800. [doi: 10.1007/s11192-023-04807-1]
30. Rakedzon T, Segev E, Chapnik N, Yosef R, Baram-Tsabari A. Automatic jargon identifier for scientists engaging with the public and science communication educators. PLoS One 2017;12(8):e0181742 [FREE Full text] [doi: 10.1371/journal.pone.0181742] [Medline: 28792945]
31. Gotlieb R, Praska C, Hendrickson MA, Marmet J, Charpentier V, Hause E, et al. Accuracy in patient understanding of common medical phrases. JAMA Netw Open 2022;5(11):e2242972 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.42972] [Medline: 36449293]
32. Praska C, Pitt MB, Marmet J, Gotlieb R, Charpentier V, Hause E, et al. Laypeople's (mis)understanding of common medical acronyms. Hosp Pediatr 2023;13(10):e269-e273. [doi: 10.1542/hpeds.2023-007282] [Medline: 37743804]

XSL•FO
RenderX

33. Grossman Liu L, Russell D, Reading Turchioe M, Myers AC, Vawdrey DK, Masterson Creber RM. Effect of expansion of abbreviations and acronyms on patient comprehension of their health records: a randomized clinical trial. JAMA Netw Open 2022;5(5):e2212320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.12320] [Medline: 35560053]
34. Fisher KA, Kennedy K, Bloomstone S, Fukunaga MI, Bell SK, Mazor KM. Can sharing clinic notes improve communication and promote self-management? A qualitative study of patients with COPD. Patient Educ Couns 2022;105(3):726-733 [FREE Full text] [doi: 10.1016/j.pec.2021.06.004] [Medline: 34175167]
35. Lam BD, Bourgeois F, Dong ZJ, Bell SK. Speaking up about patient-perceived serious visit note errors: patient and family experiences and recommendations. J Am Med Inform Assoc 2021;28(4):685-694 [FREE Full text] [doi: 10.1093/jamia/ocaa293] [Medline: 33367831]
36. Goedhart NS, Zuiderent-Jerak T, Woudstra J, Broerse JEW, Betten AW, Dedding C. Persistent inequitable design and implementation of patient portals for users at the margins. J Am Med Inform Assoc 2021;28(2):276-283 [FREE Full text] [doi: 10.1093/jamia/ocaa273] [Medline: 33463691]
37. Antonio MG, Petrovskaya O, Lau F. Is research on patient portals attuned to health equity? A scoping review. J Am Med Inform Assoc 2019;26(8-9):871-883 [FREE Full text] [doi: 10.1093/jamia/ocz054] [Medline: 31066893]
38. Chen J, Jagannatha AN, Fodeh SJ, Yu H. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: adapted distant supervision approach. JMIR Med Inform 2017;5(4):e42 [FREE Full text] [doi: 10.2196/medinform.8531] [Medline: 29089288]
39. Wood M, Gupta A. Identifying and classifying medical jargon through analysis of recorded standardized patient encounters. Patient Educ Couns 2021;104(8):2122-2125. [doi: 10.1016/j.pec.2021.01.026] [Medline: 33581969]
40. McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to web-based medical information. Proc AMIA Symp 1999:107-111 [FREE Full text] [Medline: 10566330]
41. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24-29 [FREE Full text] [doi: 10.1197/jamia.M1761] [Medline: 16221948]
42. Smith CA. The Ten Thousand Questions Project: towards a better understanding of consumer health vocabulary. J Consum Health Internet 2007;11(1):33-47. [doi: 10.1300/j381v11n01_03]
43. Vydiswaran VGV, Mei Q, Hanauer DA, Zheng K. Mining consumer health vocabulary from community-generated text. AMIA Annu Symp Proc 2014;2014:1150-1159 [FREE Full text] [Medline: 25954426]
44. Gu G, Zhang X, Zhu X, Jian Z, Chen K, Wen D, et al. Development of a consumer health vocabulary by mining health forum texts based on word embedding: semiautomatic approach. JMIR Med Inform 2019;7(2):e12704. [doi: 10.2196/12704] [Medline: 31124461]
45. He Z, Chen Z, Oh S, Hou J, Bian J. Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach. J Biomed Inform 2017;69:75-85 [FREE Full text] [doi: 10.1016/j.jbi.2017.03.016] [Medline: 28359728]
46. Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res 2011;13(2):e37 [FREE Full text] [doi: 10.2196/jmir.1636] [Medline: 21586386]
47. Smith CA, Stavri PZ, Chapman WW. In their own words? A terminological analysis of e-mail to a cancer information service. Proc AMIA Symp 2002:697-701 [FREE Full text] [Medline: 12463914]
48. Pitt MB, Hendrickson MA. Eradicating jargon-oblivion - a proposed classification system of medical jargon. J Gen Intern Med 2020;35(6):1861-1864 [FREE Full text] [doi: 10.1007/s11606-019-05526-1] [Medline: 31713041]
49. Polepalli Ramesh B, Houston T, Brandt C, Fang H, Yu H. Improving patients' electronic health record comprehension with NoteAid. Stud Health Technol Inform 2013;192:714-718. [Medline: 23920650]
50. Chen J, Druhl E, Polepalli Ramesh B, Houston TK, Brandt CA, Zulman DM, et al. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. J Med Internet Res 2018;20(1):e26 [FREE Full text] [doi: 10.2196/jmir.8669] [Medline: 29358159]
51. Morris NS, MacLean CD, Chew LD, Littenberg B. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006;7:1-7 [FREE Full text] [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
52. Voorhees EM. The TREC medical records track. Proc. Int. Conf. Bioinforma. Comput. Biol. Biomed. Inform 2013:239-246. [doi: 10.1145/2506583.2506624]
53. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. Biom J 2005;47(4):458-472. [doi: 10.1002/bimj.200410135] [Medline: 16161804]
54. Abdi H. The Kendall rank correlation coefficient. In: Salkind NJ, editor. Encyclopedia of Measurements and Statistics. New York, United States: Sage Publications, Inc; 2007:508-510.
55. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat 2004;31(7):799-815. [doi: 10.1080/0266476042000214501]
56. Akaike H. A new look at the statistical model identification. IEEE Trans. Automat. Contr 1974;19(6):716-723. [doi: 10.1109/TAC.1974.1100705]
57. Kutner M, Greenburg E, Jin Y, Paulson C. The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy: National Center for Education Statistics; 2006.

XSL•FO
RenderX

58. Difallah D, Filatova E, Ipeirotis P. Demographics and dynamics of Mechanical Turk Workers. 2018 Presented at: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining; February 5-9, 2018; Marina Del Ray, California p. 135-143. [doi: 10.1145/3159652.3159661]
59. Snow R, O'Connor B, Jurafsky D, Ng A. Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. Proc Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics 2008:254-263. [doi: 10.3115/1613715.1613751]
60. Yao Z, Cao Y, Yang Z, Deshpande V, Yu H. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. AMIA Annu Symp Proc 2022;2022:1188-1197 [FREE Full text] [Medline: 37128373]
61. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nat Commun 2023;14(1):7857 [FREE Full text] [doi: 10.1038/s41467-023-43715-z] [Medline: 38030638]
62. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
63. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
64. Kuckelman IJ, Yi PH, Bui M, Onuh I, Anderson JA, Ross AB. Assessing AI-powered patient education: a case study in radiology. Acad Radiol 2024;31(1):338-342. [doi: 10.1016/j.acra.2023.08.020] [Medline: 37709612]
65. Clinical trials - Regulation EU No 536/2014. European Commission. URL: https://health.ec.europa.eu/medicinal-products/clinical-trials/clinical-trials-regulation-eu-no-5362014_en [accessed 2024-07-02]
66. GitHub. nd-ball / jargon-identification. URL: https://github.com/nd-ball/jargon-identification [accessed 2024-09-20]

Abbreviations

AIC: Akaike information criterion
EHR: electronic health record
MTurk: Amazon Mechanical Turk

Edited by A Mavragani; submitted 06.06.23; peer-reviewed by KH Kim, N Hardiker, S Goyal, Y Chu, C Smith; comments to author 20.01.24; revised version received 13.03.24; accepted 20.08.24; published 15.10.24.

Please cite as:
Lalor JP, Levy DA, Jordan HS, Hu W, Smirnova JK, Yu H
Evaluating Expert-Layperson Agreement in Identifying Jargon Terms in Electronic Health Record Notes: Observational Study
J Med Internet Res 2024;26:e49704
URL: https://www.jmir.org/2024/1/e49704
doi:10.2196/49704
PMID:39405109

©John P Lalor, David A Levy, Harmon S Jordan, Wen Hu, Jenni Kim Smirnova, Hong Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Physicians' Perceptions of Digital Health's Impact on the Patient-Physician Relationship in the Primary Health Care Setting: Qualitative Descriptive Study

Kai Ping Sze[1], MBBS, MMed (FM), FCFP; Qi Wei Fong[1], MBBS, MMed (FM); Jacqueline Giovanna De Roza[1], BHsN, MN, MHS; Eng Sing Lee[1,2], MBChB, MMed (FM), FCFP, PhD; Shu Yun Tan[1,2], MBBS, MMed (FM), MCI, FCFP

[1]National Healthcare Group Polyclinics, Singapore, Singapore

[2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

Corresponding Author:
Kai Ping Sze, MBBS, MMed (FM), FCFP
National Healthcare Group Polyclinics
3 Fusionopolis Link, #03-08
Singapore, 138543
Singapore
Phone: 65 63553000
Email: kai_ping_sze@nhgp.com.sg

Abstract

Background: Digital health has become essential for effective clinical practice. However, the successful adoption of digital health is dependent on the strength of the patient-physician relationship. The patient-physician relationship shapes the quality of care and impacts health care outcomes, especially in primary care. However, the impact of the increasing use of digital health on the patient-physician relationship is uncertain.

Objective: This study aims to explore the types of digital health primary care physicians use and understand their impact on the patient-physician relationship from their perspective.

Methods: This exploratory qualitative descriptive study used individual in-depth interviews guided by a semistructured topic guide. We purposively sampled physicians from 6 general primary care clinics in Singapore and used thematic analysis to identify emergent themes.

Results: We conducted 12 interviews. We found that primary care physicians in Singapore had minimal exposure to digital health beyond the scope of institutional implementation. The three key themes that emerged were as follows: (1) evolving roles of both physicians and patients; (2) impact on trust, knowledge acquisition, and longitudinal care; and (3) adoption and use factors of digital health impacting patient-physician relationships. The adoption and use factors comprised "social and personal," "technical and material," and "organization and policy" factors.

Conclusions: The study identified that, while primary care physicians held mostly positive views on adopting digital health in improving the patient-physician relationship, they were concerned that digital health might erode trust, hinder proper knowledge acquisition, and reduce humanistic interaction. These concerns called for a nuanced approach to ensure that digital health would not compromise the patient-physician relationship. This could be achieved by ensuring that physicians possess the necessary skills, knowledge, and positive attitude, while health care organizations would provide robust IT capabilities and support. We recommend that education be refined and government policies on digital health adoption and use be revised to align with the goal of strengthening the patient-physician relationship.

(J Med Internet Res 2024;26:e53705) doi:10.2196/53705

KEYWORDS

patient-physician relationship; patient communication; trust; primary care medicine; digital health; primary care; longitudinal care; policy; implementation

XSL•FO
RenderX

Introduction

Background

The emergence of digital health has become essential for effective clinical practice in primary care. According to the World Health Organization, digital health refers to the knowledge and practices associated with the development and use of digital technologies to improve health. Digital health expands the concept of eHealth to include digital consumers, with a wider range of smart devices and connected equipment. It also encompasses other uses of digital technologies for health such as the Internet of Things, artificial intelligence, big data, and robotics [1-5].

In Singapore, the adoption of digital health in health care accelerated during the COVID-19 pandemic, including telemedicine, video consultations, and a new national database that unified the public and private health care sectors through the COVID-19 Test Repository system [6]. Although both public institution polyclinics and private general practitioner (GP) practices have adopted digital health in Singapore, the former have been more uniform in their adoption of technologies such as electronic health records (EHRs), patient portals, telemedicine, and health analytics, while the latter had more variability in their use.

The term "primary care" is used synonymously with family medicine or general practice [7]. Primary care provides personal, primary, and preventive care to patient care needs and comprehensive, continuing, and coordinated care in managing the patient and his or her family. The primary care system in Singapore is characterized by a dual structure comprising public and private providers. The public sector includes 26 polyclinics offering a wide range of services such as chronic disease management, preventive care, and maternal and child health services. The private sector consists of numerous GP clinics which deliver personalized and continuous care to patients. A polyclinic may have up to 60 clinicians, whereas a GP clinic is typically run by one or a few GPs. Overall, Singapore's primary care system emphasizes a patient-centric approach and plays a pivotal role in maintaining the population's health, managing chronic conditions, and reducing the burden on secondary and tertiary health care facilities.

The digital health tools of particular relevance in Singapore's primary care for public institution care includes video consultation using Zoom (Zoom Video Communications, Inc) interface, clinical decision support using corporate system software, point of care tools, shared electronic clinical data, a secured smartphone app for physicians to communicate with specialists or for patients to have one-stop access to personal medical records, and remote monitoring of blood pressure using smart versions of clinical devices.

The patient-physician relationship is crucial in delivering high-quality health care outcomes, quality indicators, and health equity [8-13]. It is widely recognized as the most potent element in medicine as it shapes the quality of care and impacts a range of health care outcomes [8], especially in primary care, where patient-centeredness is highly valued by patients [9,10], and positive patient perceptions of the consultations have been linked to improved health outcomes [11-13].

The proliferation of digital health has shifted tasks traditionally performed in health care facilities into the patient's home. These tools allow patients to record their own biological parameters, such as smartwatches monitoring their heart rates and rhythms. The interplay of patient-physician communication in face-to-face environments and relationship factors (eg, patient trust and patient satisfaction) could exert significant effects in promoting eHealth adoption [14]. In addition, Balato et al [15] found that patient-physician communication improved in a group receiving SMS text message interventions, suggesting that digital interventions could enhance this relationship. This democratization of monitoring and care contributes to an equal level of the patient-physician relationship as argued by Meskó et al [16]. In addition, the increased reliance on digital health has promoted the transition from the traditional guidance-cooperation model of care, in which the physician makes decisions for the patient, to a mutual participation partnership model, in which the physician and patient work together to achieve the patient's goals [17,18]. The new model bestows equal power to both parties and they are mutually interdependent.

While digital health holds great potential to revolutionize health care by providing innovative solutions to address health-related issues for both physicians and patients, successful adoption of these digital products has to be contingent upon the strength of the patient-physician relationship. A positive patient-physician relationship is crucial to ensuring satisfactory health care encounters and effective disease management in the primary care setting. Conversely, a weak patient-physician relationship can lead to increased medicolegal issues, clinician attrition, and ultimately poorer health outcomes for patients.

This Study

While previous studies examining the impact of digital health on primary care performance and quality have already shown that digital health holds promise in improving access, efficiency, and patient empowerment [19,20], no studies have so far explored the experiences of primary care physicians on their use of digital health and its impact on the patient-physician relationship. One reason might be due to the medicalization of clinical medicine leading to the deprioritization of humanistic medicine [21]. Another reason might be the relatively recent emergence and rapid evolution of digital health [22]. With the proliferation of digital health, it is uncertain how it would affect the patient-physician relationship. Therefore, this study aimed to fill the knowledge gap by finding out about the types of digital health that primary care physicians in Singapore use and understanding the impact of digital health on the patient-physician relationship from their perspectives.

Methods

Study Design and Participant Selection

The qualitative descriptive research method [23] as a study design involving individual in-depth interviews was used for this study.

XSL•FO
RenderX

Recruitment and data collection were completed within an 8-month period from June 2022 to February 2023. A total of 8 months were taken as data analysis was conducted following the data collection from each interview, and there were quality considerations in selecting interviewees capable of giving meaningful responses. We purposively identified primary care physicians to include individuals of both male and female genders with different levels of primary care experience from the professional networks of the study team. These primary care physicians were selected from 6 polyclinics which served the population in the central and northern parts of Singapore. The 6 polyclinics belong to the National Healthcare Group Polyclinics institution and hence leverage similar digital technologies. They provide a wide range of primary care services, including general medical consultations, chronic disease management, maternal and child health services, immunizations, health screenings, and diagnostic services. The polyclinics receive approximately a daily total of 5000 patients for physician consultations. The inclusion criterion was to have at least 3 years of work experience in a primary care setting, as registered medical practitioners need to have at least 3 years of work experience and relevant and recognized postgraduate qualifications to qualify as a family physician in Singapore.

Conceptual Framework

The frameworks proposed by Ridd et al [24] (elements of the patient-physician relationship) and Jacob et al [25] (adoption and use factors of digital health) were used to explore the interaction of the patient-physician relationship in the context of digital health adoption and use in an iterative way. The elements in a patient-physician relationship described by Ridd et al [24] included trust, loyalty, regard, knowledge, and consultation experiences. Jacob et al [25] described the adoption and use factors as social and personal factors, technical and material factors, and organization and policy factors [25]. By combining these 2 frameworks (Figure 1), we were able to cover issues related to physician," "patient," "digital health," and "trust," as well as interdomain issues such as "patient-physician relationship," "positive and negative experiences," and "technology compatibility."

Figure 1. Combined conceptual framework.

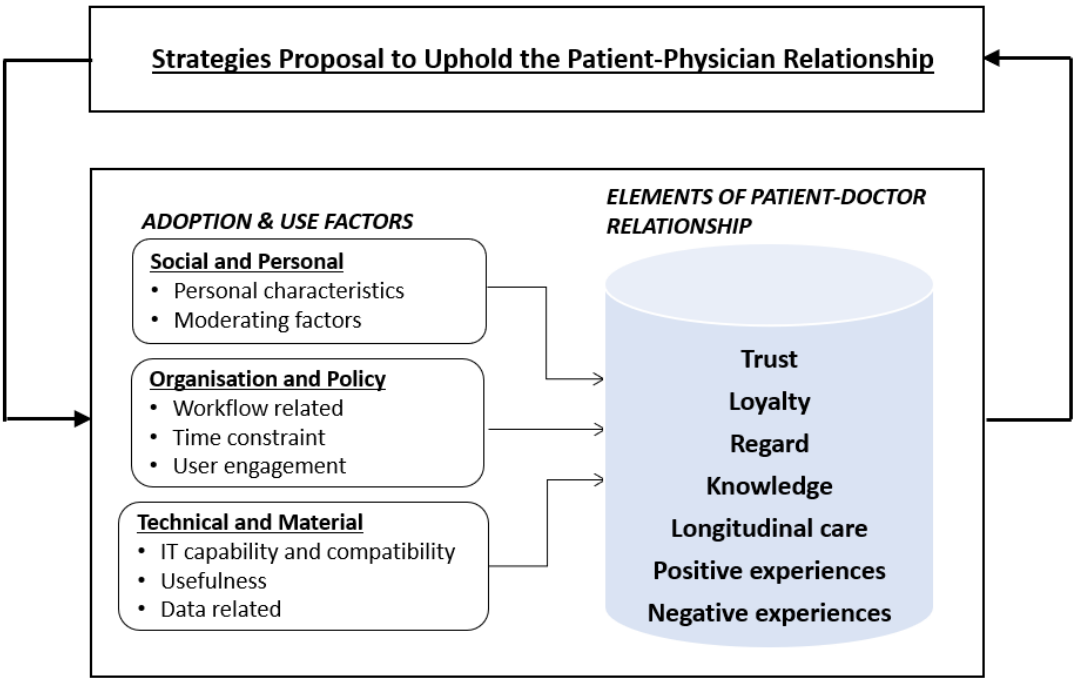

Interview Guide

We formulated a topic guide to direct the semistructured interviews (Multimedia Appendix 1) based on the conceptual framework. The topic guide comprised nondirective, open-ended questions on the following topics: intentions and experience of technological changes, present role and adoptability of digital health in primary care clinical practice, and digital health's impact on the patient-physician relationship. The interview questions were pilot-tested and improved upon before being used. After each subsequent interview, each interviewer completed a standardized debrief with the research team and included suggested revisions to further improve the interview guide.

Data Collection

Each in-depth interview lasted between 45 and 60 minutes and was conducted over Zoom in English and audio recorded. The interviewers were KPS and QWF, both male family physicians with approximately 8 and 15 years of clinical experience in family medicine with a master of medicine postgraduate family medicine qualification, who had no prior relationship with the participants. Another coinvestigator, either SYT (female family physician) or JGDR (advanced practice nurse), was present as an observer in every interview to capture relevant field notes and nonverbal cues. There were no other nonparticipants present. Recruitment for individual interviews was stopped upon thematic saturation. This was attained when themes and

XSL•FO
RenderX

subcategories in the data became repetitive and redundant such that no new information could be gathered by further data collection [26]. No repeat interviews were required for further clarification.

Data Analysis

Each audio-recorded interview was transcribed verbatim. Transcripts were returned to the participants for comment. Reflexive thematic analysis of the data was conducted according to the method described by Clarke and Braun [27]. Data from deidentified transcripts were analyzed alongside ongoing data collection after each interview. Coding was performed manually and data were managed using Microsoft Excel software. We began with reading and familiarization of transcripts for preanalytical understanding, followed by systematic line-by-line coding of transcripts. We independently (KPS, QWF, JGDR, and SYT) coded each transcript before clarification and agreements on the coding framework. Potential themes were then jointly developed based on the codes through consensus. Any disagreements that arose during coding and theme development were discussed among the researchers until a consensus was reached [28]. All coders met regularly to further develop the analysis and provide a check on coding consistency. Conscious attempts were made to be open to unexpected findings. The team concurred that there was no new emergence of data by the 10th interview and the point of thematic saturation [26] was reached. This was confirmed by 2 further more interviews that did not contribute to the development of new themes. Reflective memos were used throughout the data collection and analysis process to ensure the richness of data. Repeat interviews were not required.

Ethical Considerations

Ethics approval was obtained from the National Healthcare Group Domain Specific Review Board (2021/01036), and the study was performed in accordance with the relevant guidelines and regulations laid by it. We used the COREQ (Consolidated Criteria for Reporting Qualitative Research) as our reporting framework. Written informed consent followed by sociodemographic information was obtained physically from the participants before commencing the interviews over Zoom. All participants were reimbursed with grocery store vouchers of 20 Singapore dollars (estimated US $15) as token of appreciation.

Results

A total of 13 primary care physicians were approached and 12 (92%) of them consented to participate in the research. However, 1 (8%) declined participation over concerns with audio recording.

Participants' Characteristics

The participants included primary care physicians holding different designations in the public institution setting with varying years of work experience. Their ages ranged from 28 to 51 years. Their characteristics are represented in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of study participants (N=12).

Characteristic	Participants, n (%)
Gender	
Men	6 (50)
Women	6 (50)
Age (y)	
25-34	4 (33)
35-44	5 (42)
≥45	3 (25)
Ethnicity	
Arab	1 (8)
Chinese	10 (83)
Malay	1 (8)
Number of years working in a primary care setting	
3-5	3 (25)
6-10	3 (25)
>10	6 (50)
Highest postgraduate training attained	
MBBS[a] or MD[b]	1 (8)
Graduate diploma in family medicine	1 (8)
Master of medicine (family medicine)	8 (67)
Fellow of the College of Family Physicians, Singapore	2 (17)
Designation	
Family medicine resident[c]	1 (8)
Family physician[d]	2 (17)
Family physician senior staff[e]	1 (8)
Associate consultant[f]	6 (50)
Consultant[f]	1 (8)
Senior consultant[f]	1 (8)

[a]Bachelor of medicine and bachelor of surgery.

[b]Doctor of medicine.

[c]Qualified physicians granted conditional or full medical registration by the Singapore Medical Council undergoing family medicine residency postgraduate training program.

[d]Registered medical practitioners with relevant and recognized postgraduate qualifications (graduate diploma in family medicine or master of medicine ([family medicine]) and with at least 3 years of relevant clinical experience.

[e]Registered medical practitioners with relevant and recognized postgraduate qualifications (graduate diploma in family medicine or membership in the Royal College of General Practitioners) and with at least 5 years of clinical experience.

[f]Registered medical practitioners with relevant and recognized postgraduate qualifications (master in family medicine, or a fellow of the College of Family Physicians, Singapore).

Participants' Digital Health Use Characteristics

The digital health tools that all our 12 participants used in their clinical practice were summarized in Table 2.

XSL•FO
RenderX

Table 2. Digital health use characteristics of study participants (N=12).

Digital health use	Participants, n (%)
Epic [29]	12 (100)
NEHR[a]	12 (100)
PTEC[b] Home Blood Pressure Monitoring Programme [30]	12 (100)
Zoom	12 (100)
TigerText	12 (100)
HealthHub	12 (100)
Freestyle Libre capillary glucose monitoring device	4 (33)
Wearable devices	2 (17)
UpToDate smart app	1 (8)

[a]NEHR: national electronic health record.

[b]PTEC: Primary Tech-Enhanced Care.

Emerging Themes

The results were grouped into 3 themes (Figure 2) based on the applied conceptual framework for digital health's impact on patient-physician relationship adapted from Ridd et al [24] and Jacob et al [25].

Figure 2. The interdependent themes and subthemes on the impact of digital tools on the patient-physician relationship from primary care physicians' perspectives.

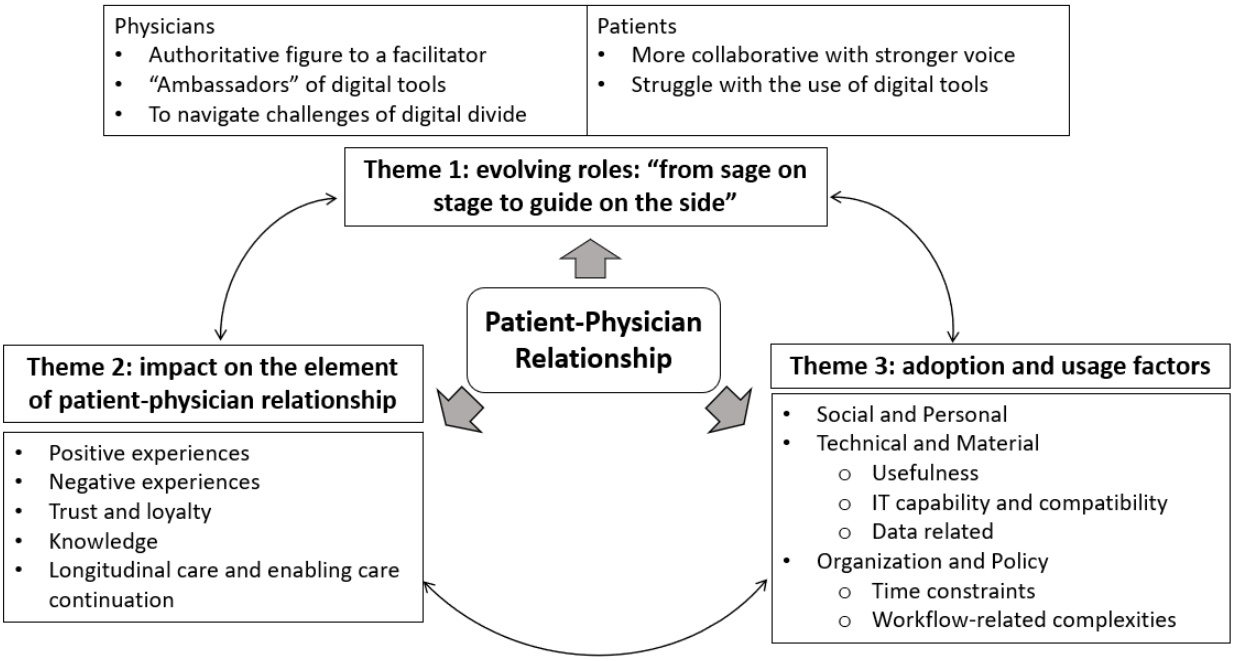

Theme 1: Evolving Roles ("From Sage on Stage to Guide on the Side")

Nowadays, patients generally prefer a more collaborative approach to a paternalistic approach, with patients playing a more active role in their own health care decisions by getting more involved in discussion and incorporating their beliefs and preferences. Primary care physicians would gather information from the patient and combine it with their own medical knowledge to develop safe and reasonable management plans. This collaborative approach allowed patients to have a more significant voice in the final plan:

In the past, the decision is very simple right? What I know is the best for you is going to be what I recommend to you, but now it is different, now I know what is best for you but you give me this other information from Google, and you said that you trust it a lot so how are we going to integrate it so that you don't disagree.... By telling the patient the pros and cons, which is [part of] the decision-making process for the doctor, we eventually come to a conclusion with the patient's voice being a big part of the final plan. [PC5, female family physician]

XSL•FO
RenderX

The democratization of digital information reduced information asymmetry and shifted the role of physicians from an authoritative figure to being a facilitator of patient education and behavior change. As patients had preexisting knowledge, primary care physicians needed to adapt their approach and act as advocates for their patients' health, guiding and nudging them toward desirable behaviors and correcting any misconceptions they may have. Overall, the patient-physician relationship has become more analytical and collaborative in recent years, with physicians taking on a more facilitative role in promoting patient education and behavior change:

> *In this digital day and age, patients would have already come in with a certain knowledge about their conditions, certain opinions, and you try and advocate and nudge them in a way that will promote desirable behaviour, [and] perhaps correct misconceptions.* [PC8, male senior consultant]

Primary care physicians also had to act as "ambassadors" for digital health to educate patients on their potential benefits for managing clinical conditions. By assuming an active role in promoting digital literacy and facilitating access to appropriate resources, physicians could empower patients to take a more active role in their own care. One participant introduced a national health application platform to an older patient successfully, fostering a sense of self-worth in him. By empowering patients to use digital health, the physician demonstrated respect for his patient's capabilities and autonomy, thereby strengthening their relationship:

> *But because he is elderly and I introduced him to HealthHub...I think it was dignity that "Oh, I'm eighty years old, I can still use phone and HealthHub, you know." And I taught him how to do it.* [PC12, female associate consultant]

One participant cautioned primary care physicians to be mindful of the health care inequality. Specifically, physicians who insisted on using certain digital health tools, such as glucometers or telemonitoring of blood pressure, might inadvertently create unintended barriers to care for patients who could not afford them. Patients might feel discriminated against being given optimal care because they did not know, could not afford, or did not have support at home, leading to a breakdown in trust:

> *If I keep insisting on using the glucometer or tele-BP, for safety reason, when the patient cannot afford it, patient might feel that they are not given optimum care.... They might also feel left behind because of lack of knowledge and they really cannot understand what is HealthHub, how to download and all that. Sometimes... they might feel that they are being discriminated against.* [PC12, female associate consultant]

Participants reported on their experiences with older patients who struggled to use digital devices. Furthermore, the study found that for digital health to be effective for older patients, these tools needed to be matched to the level of technological proficiency of the individual patient. Failure to consider these factors could lead to patients having difficulties in adoption and suboptimal use, thus affecting the quality of care and the patient-physician relationship. While the participants remained committed to promoting the use of digital health, they were mindful that some patients might face challenges due to the digital divide. Primary care physicians would therefore need to navigate such situations with tact and resourcefulness:

> *Of course there is the patients' tech-savviness. It can be a very good tool, but if the patient is unable to utilise it because they don't understand or they don't know how to use it, then it's a useless tool to the patient.* [PC5, female family physician]

Nevertheless, some patients would still expect a more prescriptive approach from their primary care physicians, which could lead to misunderstandings and a patient's perception of incompetence on the part of the physician. To maintain a strong patient-physician relationship, physicians must strike a nuanced balance between the 2 approaches and adjust their communication style based on the needs and preferences of each patient:

> *I made the mistake of trying to offer too many options and was viewed as being an incompetent doctor n because the patient had expected a more paternalistic and prescriptive kind of style. On one hand, in the past, it used to be legacy effect that we are prescriptive or paternalistic, but nowadays, it has changed, so it's a balance of both.* [PC10, female associate consultant]

Theme 2: Impact on Trust, Knowledge Acquisition, and Longitudinal Care

Positive Experiences

The overall impact on the participants was positive when digital health was applied to the background of an existing patient-physician relationship:

> *Patient-doctor relationship improves when patients' medical conditions are well controlled [due to digital health] leading to greater satisfaction from both parties.* [PC7, male senior family physician]

Physicians were able to efficiently synthesize clinical information of patients and provide holistic care through a consolidated view of medical history, telemonitoring data charts, laboratory results, medications, and opinions of different care providers. For instance, EHR systems increased the visibility and the clarity of information for both physicians and patients, which in turn increased their confidence in technology. This, in turn, allowed for more effective communication between physicians and patients, as physicians could provide a more holistic picture of the patient's health status and treatment plan:

> *Integrated system enhances the patient experience because now you can access the information from everywhere, you can piece the story together and come out with a better, more consolidated care plan for the patient. You can update them about things that they have done elsewhere, and that actually increases that rapport and gives you that holistic approach to the patient.* [PC3, female consultant]

XSL•FO
RenderX

Negative Experiences

However, most participants felt that clinical documentation systems were time-consuming and reduced eye contact and physical interaction with patients. Simultaneously, patients would perceive physicians as unempathetic and paying less serious attention to their complaints when the latter spent more time on the EHR system. This could upset the dynamics of the consultation when patients perceived physicians as being more interested in completing a checklist than in attending to their needs:

> *Doctor is not necessarily doing less for the patient but purely because more time is spent gleaning all the info from the computer, the patient might feel that the doctor is not paying as much attention to them and not taking them seriously in terms of their complaints. I think that will affect the dynamics of the consultation.* [PC5, female family physician]

Trust

Digital health had a dual effect on the trust shared between physicians and patients, which could be either beneficial or detrimental.

With telemonitoring tools, physicians could adjust medications better and improve patient involvement in their care based on shared clinical information on common platforms, such as home blood pressure and capillary glucose readings. Caregivers were also more involved in the care process and able to provide valuable feedback to the physicians, which could improve treatment outcomes. Technology has enhanced the patient-physician relationship through fostering a shared sense of responsibility for the patient's health and promoting collaborative decision-making. This should ideally result in increased patient satisfaction and better health outcomes:

> *The positive thing is that with the BP machines at home, SMBG [self-monitoring blood glucose], I can better titrate my patients' medication. Patient can also be able to be more involved in their care, as are their caregivers.... I could also show them that "Hey, actually this works, you just trust us.... we are keeping a close watch on you.*" [PC12, female associate consultant]

Patients who read the same web-based information as provided by their physicians were more likely to trust the information validated by their physicians and exhibit increased confidence in their physicians' clinical expertise. There was also increased partnership when both physicians and patients had equal access to clinical knowledge and data, which could help to build a patient-physician relationship based on trust, communication, and accountability:

> *Patients are more knowledgeable so whatever you tell them, they are more convinced because they are also reading the same thing on the internet. So in that way that helps a little bit with their confidence there when delivering this kind of information.* [PC7, male senior family physician]

However, most (9/12, 75%) study participants expressed concerns that the misuse of internet could lead to patient distrust, conflicts with clinicians, and deterioration of the patient-physician relationship. Digital health could contribute to misinformation when patients use search engines and social media to research their health issues without distinguishing reliable from unreliable sources. Physicians might worry that patients trust web-based information more than the advice they receive from them. This could undermine trust and lead to a breakdown in the patient-physician relationship. Patients who relied heavily on digital health to self-diagnose or research their health issues might feel hesitant to discuss their concerns with their physician, leading to reduced communications with their primary care physicians. Similarly, physicians struggled to determine what their patients were reading and where they obtained their sources of information:

> *Things like google can be double edged sword. There's a lot of information out there, [but if] the patient has no insight on the right information then it can lead to a lot more detriment down the road. By then it will create a disparate or a distrust with the clinician.* [PC3, female consultant]

Unfamiliar technology initiated by patients also created a sense of uncertainty and unease among physicians. The quote mentioned subsequently also underscored the importance of physicians being proficient in the use of digital health when recommending them to patients. If they were not knowledgeable with the tools, it could undermine patient trust and confidence in their care and hence negatively impact the patient-physician relationship:

> *If you ask the patient to adopt this digital tool and you don't know how to use the tool, it reflects quite badly on you as a physician because it doesn't instil confidence in the patient and I would hate to say that it even causes the patient to mistrust your advice because you just don't seem to know your stuff.* [PC5, female family physician]

Establishing trust in a new relationship and maintaining trust in an existing patient-physician relationship using digital health could be challenging. One participant highlighted a potential challenge faced by primary care physicians in building trust with newer patients, particularly when digital health tools such as teleconsultation were used as part of the care process. One potential explanation for this phenomenon is that the use of teleconsultation might interfere with the interpersonal aspects of care that are critical to building trust and rapport between patients and providers. Physicians might also be perceived as more distant or impersonal when communicating through digital channels, which could create barriers to developing a meaningful relationship with patients. Another reason could be that patients might be more skeptical of health care providers who relied heavily on digital health, particularly if they perceived these tools as substitutes for traditional forms of care:

> *For newer patients when I'm still trying to build that relationship, I do find that sometimes it does either cause it to take a bit longer or it can become a little bit uh I wouldn't really call it contrived, but generally they will take a little longer time to trust me...* [PC3, female consultant]

XSL•FO
RenderX

> *If I say in a two-year timeframe, I've seen you physically once, and I've done six video consults over time, do I think that trust will be eroded? I think to some extent it will be, because there's only that much we can do over Zoom.* [PC8, male senior consultant]

Using Digital Health to Acquire Knowledge

Digital health was also considered convenient for both patients and physicians. It provided fast access to clinical questions when needed so that physicians would have the means to quickly find clinical answers. Paradoxically, primary care physicians did not experience a reduction in patients' trust when they searched for medical information on the web. Instead, they might feel a lack of trust from patients if they are not as proficient as their patients in using digital health for medical information. In other words, physicians were expected to have extensive medical knowledge, but patients also expected them to be adept in digital resources to enhance their diagnoses and treatment plans:

> *When having a question that someone asks which you don't have the answer immediately, having a tool in your hands to search for the answers quickly will be helpful.* [PC1, male associate consultant]

The need for physicians to be more familiar with these technologies than their patients highlighted the importance of providing patients with clear instructions and support in using them, which could further strengthen the patient-physician relationship:

> *You need to know the tool better than your patients in order to use it as a physician, right, so I think that is one major issue as well.* [PC5, female family physician]

Most participants (10/12, 83%) agreed that the rise of the internet has made primary care physicians' jobs more challenging. Some contended that patients, nowadays, challenge physicians' information with the evidence they find on the internet. Patients often came to their appointments with strong convictions based on compelling information they had read or watched on the web, which made it difficult for physicians to persuade them to change their minds. Consequently, physicians often needed to find creative ways to persuade patients and communicate differently by getting more objective facts and trustworthy sources to explain to patients. Physicians would also need to fight against misinformation and synthesize the information patients find on the internet to give the best advice:

> *With the current advent of the internet and the availability of info out there, there is a lot more misinformation that we have to fight... Our role now becomes more than just first level information providers, we are now someone who needs to integrate whatever info that they find on the internet with what we know and then give our best practice advice.* [PC5, female family physician]

> *I think in this modern age we have seen so many times when patients bring up their emails or their articles and say, "see doctor, cholesterol medicine is harmful for my body." You gotta think of more creative ways to counter those arguments, to convince them to take their medicine. I don't contest the fact that they have found something. I will find a different way to persuade them.* [PC4, male family physician]

Longitudinal Care and Enabling Care Continuation

One participant highlighted that a strong patient-physician relationship and increased interaction were crucial for improving patient outcomes. With the use of digital health, these interactions could occur more often and effectively, leading to improved patient engagements and better long-term health outcomes. Effective communication between physicians and patients was important in addressing patient concerns picked up via digital health:

> *It can be potentially better if there are more touchpoint with the healthcare provider, or there's greater degree of interactions.* [PC2, male family physician]

Digital health has revolutionized care continuation and augmented patient care beyond the physical boundaries of the clinic through virtual consultation and increased touchpoints:

> *If I see a patient in my clinic and then I say, "Ok we do a Zoom call the next time round... It provides you more convenience, [and] we're able to monitor you outside the confines of the clinic setting." So there is an advantage where technology can augment patient care.* [PC8, male senior consultant]

However, increased use of digital health might disrupt traditional patient care-seeking behavior as patients become overly reliant on these technologies and fail to seek care when necessary:

> *Patients may then assume that a lot of things can be done on their own, when they still need to seek care at certain points. [That] sometimes destroys the fine line between too much doctoring vs not being able to intervene because the patient is too reliant on the digital and doesn't want to come back when they need to.* [PC3, female consultant]

Theme 3: Adoption and Use Factors Impacting the Patient-Physician Relationship

Social and Personal

The participants' prior experiences with pen-and-paper documentation and subsequent shift to EHR systems exemplified the challenges of adapting to new technology in the primary health care landscape. The transition to new digital health was described as a steep learning curve by some participants that initially impeded their ability to effectively use the tools until familiarity was achieved. Nevertheless, almost all of the participants (11/12, 92%) displayed motivation and positive attitude to overcome the steep learning curve and to avoid being obsolete, and once they attained competency in digital health's mastery, they were able to reach out to their patients to help them overcome their technophobia:

> *During the initial period, there's a steep learning curve. It will impede us at the start and only after familiar usage, then you will be able to gain, [and]*

XSL•FO
RenderX

> *reap the actual benefits of using these tools.* [PC11, female associate consultant]

> *There's a lot tech phobia among especially all the older generations. They are very fearful how to use this, they are not very sure or not confident to use or whether they can even use it.* [PC7, male senior family physician]

Furthermore, 1 participant emphasized the importance of physicians' familiarity with digital health tools in promoting their use in patient care. Primary care physicians needed to grasp a good understanding of the digital health they used to effectively communicate their benefits to patients and integrate them into their health care process:

> *[U]nless you are familiar with the tools that you are using, sometimes it is difficult to advocate a tool like HealthHub because you don't use it yourself.* [PC1, male associate consultant]

Technical and Material

IT Capability and Compatibility

Some participants (3/12, 25%) expressed difficulty in finding solutions and technical support as a potential negative impact of using digital products. An example given by a participant was waiting for IT support for 2 months after being denied access to national electronics health records. This emphasized the potential impact on the patient-physician relationship, as delayed access to patient information could lead to delayed diagnosis and treatment, reduced patient satisfaction, and possibly a breakdown of trust. The use of digital health could enhance patient care and convenience, but it would be crucial to ensure that technical issues were addressed promptly to avoid potential negative consequences on the patient-physician relationship:

> *In NEHR, I don't have access to sensitive information now, previously I did but now I don't and I don't know why and I'm still waiting for the IT [support] to help me out for the past 2 months.* [PC1, male associate consultant]

Nevertheless, an integrated health record system enabled access to more patient information, which facilitated the delivery of holistic care. Patients would be assured that their physicians were invested in their care and that ample time was devoted to discussing the clinical details. Overall, an integrated system could contribute significantly to improving the quality of care and patient satisfaction, thereby strengthening the patient-physician relationship:

> *When using the [Epic] system...I can visually show the patients how they are doing and, show them the charts. It also gives them that same visual reinforcement of how they are doing. It's also very clear to them that I know what's happening. I think that helps to reinforce the trust not just in them but in the technology that we are both investing the time in. For that, I think it has improved the patient-doctor relationship and the consultation.* [PC3, female consultant]

Data Related

Digital health records helped to keep consultation sessions fruitful by enabling the seamless sharing of information among health care providers involved in a patient's care. This would help primary care physicians understand a patient's medical history and ongoing treatments, leading to more convenient and personalized care and increased patient satisfaction:

> *When your patient doesn't know what happened to himself or herself at first discharge and didn't bring their memo, then at least digital records will keep the consultation fruitful.* [PC5, female family physician]

The ability to countercheck patient-reported medical records using electronic systems could enhance the patient-physician relationship. By ensuring the accuracy and reliability of the data, patients would feel confident that their physicians had the most up-to-date information. If discrepancies or inaccuracies were identified, physicians could engage in discussions with patients to clarify the missing details, that would promote shared decision-making and stronger bonding between them:

> *If the patient was not clear about his or her medical records or for some reason was not accurate in reporting, there is always the counter check you can do in the electronic system.* [PC5, female family physician]

However, the risk of data breach in sharing patient information through digital systems posed a threat to patient confidentiality and privacy. This could lead to a negative impact on the patient-physician relationship, as patients might feel that their personal information was not being handled with adequate care and might question the security of their medical records. A high-profile data breach incident in the past had raised awareness of the potential risks associated with sharing patient information through digital systems [31]:

> *If there's a data breach, then patient confidentiality and patient private personal data is lost. So that's the negative thing.* [PC12, female associate consultant]

Organization and Policy

Time Constraints

Lack of time was often a common limiting factor quoted by many participants. While the integration of digital health in health care could potentially benefit patients, it also required additional time and effort to educate patients on their use. It might take more time initially to explain the use of digital health to patients. However, once patients were comfortable using the tools, it could improve the patient experience and facilitate remote care:

> *Initially, it might take up more time to tell the patient about Primary Tech Enhanced Care [PTEC] Home Blood Pressure Monitoring Programme, about video consult, to teach them how to use them.* [PC11, female associate consultant]

Our study participants had concerns about the number of clicks and screens slowing down the process of seeing patients and increasing the time required for documentation. This could potentially lead to physicians missing important information

XSL•FO
RenderX

and causing harm to patients. Technical issues and difficulty in navigating the tools also hindered patient care and could also create a sense of frustration and burnout among physicians, which would further strain the patient-physician relationship:

> *Epic has a lot more buttons to click to see one patient. It will definitely slow down the overall process of seeing patients.* [PC5, female family physician]

> *Sometimes if the tools are a bit cumbersome, like they take a long time to log in, or carry many technical errors.* [PC2, male family physician]

Workflow-Related Complexities

Some of the primary care physicians interviewed struggled with the complexities in workflow, including navigating complex computer systems and troubleshooting technical issues. Physicians needed to allocate more time and effort toward managing these technical aspects of health care, which might impact the quality of their interactions with patients:

> *20 years ago, all you do is just write S.O.A.P [Subjective, Objective, Assessment and Plan], prescription and send off. Nowadays, you have more things to do. You have to navigate through the Epic system more, and you need to know where to click. Of course, a lot of times the computer system also got issues. Disk crash, system hang, printer cannot print, so that increases the complexity compared to like 20 years ago.* [PC7, male senior family physician]

Discussion

Principal Findings

This qualitative study explored the types of digital health used in primary care and its impact on the patient-physician relationship in Singapore's primary care with results based on primary care physicians' perspectives. Although digital health encompassed a broad scope, our participants' exposure to it was limited to mostly what was offered by their institution in their everyday work, with few venturing to explore other technologies such as capillary glucose monitoring systems and wearable devices.

By using applied conceptual frameworks, we linked key elements of the patient-physician relationship to the adoption and use factors of digital health. We found that primary care physicians held mostly positive views on adopting digital health in improving the patient-physician relationship. These included views that patients would become more empowered, and better holistic care could be provided as the roles of physicians and patients evolved with time.

Despite the positive impact of digital health, the physicians interviewed still viewed it as a double-edged sword in the patient-physician relationship, because digital health undermined the relationship through misinformation, decreased humanistic interactions, and aggravated the digital divide. While an integrated health system could provide holistic care, the use of digital health also requires a capable IT system and implicit skills, knowledge, and positive attitude from physicians to navigate the workflow complexities and reach out to their patients within the consultation time constraints.

Finally, some study participants emphasized the potential impact of the digital divide on patients. Older patients, who are most in need of health care services, might end up losing out in the digital race. With the aging population, health inequalities and inequities in access to health care, digital health, and health outcomes area likely increase [32].

Comparison With Prior Work

The findings resonated with a web-based survey [33], which found a global consensus among primary care physicians on the role of digital health in driving greater patient empowerment. Most of the participants started adopting the bio-psycho-socio-digital paradigm approach to patient care, as they highlighted the attitudes, skills, and knowledge necessary for this approach. The findings aligned with those of Győrffy et al [34], who indicated that modern technology and the spread of digital information had accelerated the shift from a paternalistic model of the patient-physician relationship to a more collaborative and cooperative model. In this new model, digitally proficient physicians view themselves as guides who take on protective and informational roles, adeptly managing the description, collection, and dissemination of reliable content on the web [34]. This perspective is echoed by health care professionals in the study by Macdonald et al [35], who embraced the idea of patients as "partners" to help improve outcomes by educating themselves and conscientiously monitoring their condition and behavior [35]. In addition, a survey [36] found that Singapore had the second-highest percentage of physicians who believed digital health had a positive impact on the quality of treatment decisions and patients' health outcomes; our study participants generally viewed that digital health could provide better holistic care.

As the adage went, "there is only one cardinal rule: one must always listen to the patient" [37]. Similarly, our findings reminded primary care physicians to listen to their patients to be cognizant of their care preferences so that they could act as guides for patients, particularly in navigating through the digital health landscape. Our findings mirrored the findings of a recent mixed methods study [38] that identified preparing with intention, listening intently and completely, and exploring emotional cues as 3 of the 5 practices that have the potential to enhance the patient-physician relationship during a clinic encounter. In addition, doctors must prioritize their primary goal as treating patients, with technology as adjuncts. By adopting a bio-psycho-socio-digital paradigm shifts in patient engagement and emphasizing on developing the soft skills of patient communication and the technical skills of digital health in primary care education, a n ideal patient-physician relationship in this digital age could be preserved.

Both our study and the study by Kludacz-Alessandri et al [39] emphasized patient satisfaction as a crucial outcome of teleconsultations. Our findings further substantiated that an integrated health records system enhances patient satisfaction by enabling more holistic and personalized care, mirroring the convenience highlighted in the study by Kludacz-Alessandri et al [39]. In addition, both studies acknowledged challenges with

XSL•FO
RenderX

digital tools. Our study identified issues such as, the complexity of digital systems and potential data breaches, while the study by Kludacz-Alessandri et al [39] underscored the digital divide and technical difficulties as barriers to effective teleconsultations.

Our research also uncovered an intriguing "tension" that suggested a disparity in how older and younger adults perceived the reliability of physicians and the internet as sources of health information. According to Low et al [40], older adults only seemed to trust health care professionals for health information, while the internet was used to supplement existing health knowledge or for general health information. In contrast, most of the physicians (8/12, 67%) we interviewed expressed that they felt undermined by patients' increasing trust in internet for information and the need to reconvince them of evidence-based treatment.

Implications for Practice

First, physicians could consider tailoring their approach to a patient's level of digital health literacy, as well as being comfortable and competent in enhancing the relationship with digital tools that patients are familiar with. Second, exposure to digital health and the importance of relational ethics [41] during medical education could improve the patient-physician interaction in the bio-psycho-socio-digital paradigm [42]. Third, robust IT capabilities and support should be provided by health care organizations to ensure that physicians leverage digital tools effectively. Finally, government policy might be needed to address challenges and ensure equity in access to digital health [43], and community partnerships and health coaching could reduce costs and help enhance the patient-physician relationships. The fragile and evolving patient-physician relationship needs to be incubated by a broader meso- or macro-organizational, cultural, and geopolitical environment. Supportive actions such as extending consultation time and maximizing enablers to uplift digitally disadvantaged patients could strengthen the patient-physician relationship.

Implications for Research

The digital divide has emerged as a significant social determinant of health. Despite high internet and mobile device penetrations in countries such as Singapore [44], possessing such technology does not guarantee that users have the skills or motivation to use them. Therefore, accessing and engaging with digital primary care becomes a complex and multifactorial process. Given the complexity, our completed research analysis provides a comprehensive understanding of the impact of using digital health among a clinically diverse group of patients in primary care on patient-physician relationship building, with a focus on how the digital divide affects the relationship. Our next study will further explore qualitative data to capture patient perspectives in greater depth.

Strengths and Limitations

The major strength of our work is the novelty of this study, conducted in Singapore and exploring physicians' perspectives on digital health's impact on the patient-physician relationship in primary care services. In addition, we used a conceptual framework to provide a theoretical basis for the study.

Nevertheless, this study also has limitations. The qualitative research was exploratory in nature and the accounts from the participants were within the unique context of a public health care institution in Singapore's health care system, and thus, we advise caution about transferring our findings to other health care systems or local GPs from the private practice. Although purposive sampling was done, all participants interviewed were exposed to similar types of digital health implemented by their institution, with none of them discussing technologies such as artificial intelligence, big data, and robotics.

Conclusions

This study identified that all the primary care physicians interviewed had largely similar exposure to digital health and held mostly positive views on adopting digital health in improving patient-physician relationship, as it improved patient empowerment and collaborative care among health care professionals and aided clinical decision-making processes. However, they were also concerned that digital health might erode trust between physicians and patients, hinder proper knowledge acquisition, and reduce humanistic interactions that are essential for building a strong rapport. These concerns indicated a need for a nuanced approach to ensure digital health does not compromise the patient-physician relationship.

To fully use digital health in enhancing the patient-physician relationship, physicians should possess skills, knowledge, and a positive attitude to navigate the workflow complexities and connect with their patients. This should be supported with robust IT capabilities to leverage digital tools effectively. As primary care physicians continue to adapt to the full potential of digital health, we recommended that primary care education be refined and government policies on digital health adoption and use be revised to align with the goal of strengthening the patient-physician relationship.

Acknowledgments

The authors would like to thank their administrative colleagues at the Clinical Research Unit, National Healthcare Group Polyclinics, in particular Dr Sabrina Lee Poay Sian for their assistance in seeking the institutional review board approval and in obtaining the transcripts. The authors are grateful to all participating primary care physicians who have generously contributed their views and perspectives on digital health's impact on the patient-physician relationship in a local setting.

This study was supported by a funding grant under the National Healthcare Group and Nanyang Technological University's Lee Kong Chian School of Medicine (NHG-LKCMedicine) Clinician Scientist Preparatory Programme (CSPP-21-02-06).

XSL•FO
RenderX

Data Availability

The datasets analyzed during the study are available from the corresponding author on reasonable request.

Authors' Contributions

KPS contributed to conceptualization, methodology, funding acquisition, data gathering and analysis, transcribing, draft preparation, reviewing, and editing the manuscript, and journal submission. SYT contributed to conceptualization, supervision, data analysis, and reviewing, editing, and finalizing the manuscript. QWF contributed to data gathering, data analysis, facilitating interviews, and reviewing the manuscript. JGDR contributed to data gathering, data analysis, and facilitating interviews. ESL contributed to supervision and editing and finalizing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topic guide for interview.
[DOCX File , 19 KB - jmir_v26i1e53705_app1.docx]

References

1. Global strategy on digital health 2020-2025. World Health Organization. 2021 Aug 18. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-09-16]
2. Agenda item 12.4: mHealth: use of appropriate digital technologies for public health. NCD Alliance. 2018. URL: https://tinyurl.com/2xhyp7s2 [accessed 2022-11-21]
3. Guiding optimal development and use of digital health towards improved health outcomes. World Health Organization. URL: https://www.who.int/westernpacific/activities/guiding-optimal-development-and-use-of-digital-health-towards-improved-health-outcomes [accessed 2022-11-21]
4. Gutierrez G. Medicare, the internet, and the future of telemedicine. Crit Care Med 2001 Aug;29(8 Suppl):N144-N150. [doi: 10.1097/00003246-200108001-00002] [Medline: 11496036]
5. Whittaker R. Issues in mHealth: findings from key informant interviews. J Med Internet Res 2012 Oct 02;14(5):e129 [FREE Full text] [doi: 10.2196/jmir.1989] [Medline: 23032424]
6. Lai SH, Tang CQ. Telemedicine and COVID-19: beyond just virtual consultations – the Singapore experience. Bone Joint Open 2020 Jun;1(6):203-204. [doi: 10.1302/2046-3758.16.bjo-2020-0042.r1]
7. Primary health care EURO. World Health Organization. 2024. URL: https://www.who.int/europe/health-topics/primary-health-care#tab=tab_1 [accessed 2024-09-27]
8. Timmermans S. The engaged patient: the relevance of patient-physician communication for twenty-first-century health. J Health Soc Behav 2020 Sep 29;61(3):259-273 [FREE Full text] [doi: 10.1177/0022146520943514] [Medline: 32723112]
9. Little P, Everitt H, Williamson I, Warner G, Moore M, Gould C, et al. Observational study of effect of patient centredness and positive approach on outcomes of general practice consultations. BMJ 2001 Oct 20;323(7318):908-911 [FREE Full text] [doi: 10.1136/bmj.323.7318.908] [Medline: 11668137]
10. Little P, Everitt H, Williamson I, Warner G, Moore M, Gould C, et al. Preferences of patients for patient centred approach to consultation in primary care: observational study. BMJ 2001 Feb 24;322(7284):468-472 [FREE Full text] [doi: 10.1136/bmj.322.7284.468] [Medline: 11222423]
11. Stewart M, Brown JB, Donner A, McWhinney IR, Oates J, Weston WW, et al. The impact of patient-centered care on outcomes. J Fam Pract 2000 Sep;49(9):796-804. [Medline: 11032203]
12. Gordon C, Beresin EV. The doctor-patient relationship. In: Stern TA, Rosenbaum JF, Fava M, Biederman J, Rauch SL, editors. Massachusetts General Hospital Comprehensive Clinical Psychiatry. St. Louis, MO: Mosby, Inc; 2008.
13. Ong LM, de Haes JC, Hoos AM, Lammes FB. Doctor-patient communication: a review of the literature. Soc Sci Med 1995 Apr;40(7):903-918. [doi: 10.1016/0277-9536(94)00155-m] [Medline: 7792630]
14. Jiang S. The relationship between face-to-face and online patient-provider communication: examining the moderating roles of patient trust and patient satisfaction. Health Commun 2020 Mar 01;35(3):341-349. [doi: 10.1080/10410236.2018.1563030] [Medline: 30821506]
15. Balato N, Megna M, Di Costanzo L, Balato A, Ayala F. Educational and motivational support service: a pilot study for mobile-phone-based interventions in patients with psoriasis. Br J Dermatol 2013 Jan;168(1):201-205. [doi: 10.1111/j.1365-2133.2012.11205.x] [Medline: 23240729]
16. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017 Sep;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
17. Asch DA, Nicholson S, Berger ML. Toward facilitated self-service in health care. N Engl J Med 2019 May 16;380(20):1891-1893 [FREE Full text] [doi: 10.1056/nejmp1817104]

XSL•FO
RenderX

18. Szasz TS, Hollender MH. A contribution to the philosophy of medicine; the basic models of the doctor-patient relationship. AMA Arch Intern Med 1956 May 01;97(5):585-592. [doi: 10.1001/archinte.1956.00250230079008] [Medline: 13312700]
19. Erku D, Khatri R, Endalamaw A, Wolka E, Nigatu F, Zewdie A, et al. Digital health interventions to improve access to and quality of primary health care services: a scoping review. Int J Environ Res Public Health 2023 Sep 28;20(19):6854 [FREE Full text] [doi: 10.3390/ijerph20196854] [Medline: 37835125]
20. Neve G, Fyfe M, Hayhoe B, Kumar S. Digital health in primary care: risks and recommendations. Br J Gen Pract 2020 Nov 26;70(701):609-610. [doi: 10.3399/bjgp20x713837]
21. Greenhalgh T, Howick J, Maskrey N, Evidence Based Medicine Renaissance Group. Evidence based medicine: a movement in crisis? BMJ 2014 Jun 13;348(jun13 4):g3725 [FREE Full text] [doi: 10.1136/bmj.g3725] [Medline: 24927763]
22. Lluch M. Healthcare professionals' organisational barriers to health information technologies-a literature review. Int J Med Inform 2011 Dec;80(12):849-862. [doi: 10.1016/j.ijmedinf.2011.09.005] [Medline: 22000677]
23. Sandelowski M, Barroso J. Classifying the findings in qualitative studies. Qual Health Res 2003 Sep;13(7):905-923. [doi: 10.1177/1049732303253488] [Medline: 14502957]
24. Ridd M, Shaw A, Lewis G, Salisbury C. The patient-doctor relationship: a synthesis of the qualitative literature on patients' perspectives. Br J Gen Pract 2009 Apr;59(561):e116-e133 [FREE Full text] [doi: 10.3399/bjgp09X420248] [Medline: 19341547]
25. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding clinicians' adoption of mobile health tools: a qualitative review of the most used frameworks. JMIR Mhealth Uhealth 2020 Jul 06;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
26. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018 Sep 14;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
27. Clarke V, Braun V. Using thematic analysis in counselling and psychotherapy research: a critical reflection. Couns Psychother Res 2018 Mar;18(2):107-110 [FREE Full text] [doi: 10.1002/capr.12165]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Epic. URL: https://www.epic.com/ [accessed 2023-05-25]
30. Teo SH, Chew EA, Ng DW, Tang WE, Koh GC, Teo VH. Implementation and use of technology-enabled blood pressure monitoring and teleconsultation in Singapore's primary care: a qualitative evaluation using the socio-technical systems approach. BMC Prim Care 2023 Mar 16;24(1):71 [FREE Full text] [doi: 10.1186/s12875-023-02014-8] [Medline: 36927496]
31. SingHealth cyber attack: how it unfolded. The Strait Times. 2018 Jul 20. URL: https://graphics.straitstimes.com/STI/STIMEDIA/Interactives/2018/07/sg-cyber-breach/index.html [accessed 2023-04-28]
32. World population ageing 2015 highlights (ST/ESA/SER.A/368). United Nations, Department of Economic and Social Affairs, Population Division. 2015. URL: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Highlights.pdf [accessed 2024-09-17]
33. Digital doctor 2023. Ipsos. 2023. URL: https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-04/Ipsos%20Digital%20Doctor_Apr2023.pdf [accessed 2023-03-23]
34. Győrffy Z, Radó N, Mesko B. Digitally engaged physicians about the digital health transition. PLoS One 2020 Sep 28;15(9):e0238658 [FREE Full text] [doi: 10.1371/journal.pone.0238658] [Medline: 32986733]
35. Macdonald GG, Townsend AF, Adam P, Li LC, Kerr S, McDonald M, et al. eHealth technologies, multimorbidity, and the office visit: qualitative interview study on the perspectives of physicians and nurses. J Med Internet Res 2018 Jan 26;20(1):e31 [FREE Full text] [doi: 10.2196/jmir.8983] [Medline: 29374004]
36. Doctors survey 2015: six-country report. Accenture. 2015. URL: https://drive.google.com/file/d/146pzT-8oOUqIftL48L4Tkk6kvKykbfBa/view?usp=sharing [accessed 2024-09-27]
37. Sacks O. Migraine. New York, NY: Vintage Books; 1999.
38. Zulman DM, Haverfield MC, Shaw JG, Brown-Johnson CG, Schwartz R, Tierney AA, et al. Practices to foster physician presence and connection with patients in the clinical encounter. JAMA 2020 Jan 07;323(1):70-81. [doi: 10.1001/jama.2019.19003] [Medline: 31910284]
39. Kludacz-Alessandri M, Hawrysz L, Korneta P, Gierszewska G, Pomaranik W, Walczak R. The impact of medical teleconsultations on general practitioner-patient communication during COVID- 19: a case study from Poland. PLoS One 2021 Jul 16;16(7):e0254960 [FREE Full text] [doi: 10.1371/journal.pone.0254960] [Medline: 34270587]
40. Low ST, Sakhardande PG, Lai YF, Long AD, Kaur-Gill S. Attitudes and perceptions toward healthcare technology adoption among older adults in Singapore: a qualitative study. Front Public Health 2021 Feb 15;9:588590 [FREE Full text] [doi: 10.3389/fpubh.2021.588590] [Medline: 33659232]
41. Faux-Nightingale A, Philp F, Chadwick D, Singh B, Pandyan A. Available tools to evaluate digital health literacy and engagement with eHealth resources: a scoping review. Heliyon 2022 Aug;8(8):e10380 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10380] [Medline: 36090207]

XSL•FO
RenderX

42. Ahmadvand A, Gatchel R, Brownstein J, Nissen L. The biopsychosocial-digital approach to health and disease: call for a paradigm expansion. J Med Internet Res 2018 May 18;20(5):e189 [FREE Full text] [doi: 10.2196/jmir.9732] [Medline: 29776900]
43. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post-COVID era. Milbank Q 2021 Jun;99(2):340-368 [FREE Full text] [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
44. Kemp S. Digital 2021: Singapore. Data Reportal. 2021 Feb 9. URL: https://datareportal.com/reports/digital-2021-singapore [accessed 2021-12-20]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
GP: general practitioner

Edited by T de Azevedo Cardoso; submitted 31.10.23; peer-reviewed by W Wang, E Ray Chaudhuri, GDTY Yong; comments to author 22.04.24; revised version received 22.05.24; accepted 27.08.24; published 15.10.24.

Please cite as:
Sze KP, Fong QW, De Roza JG, Lee ES, Tan SY
Exploring Physicians' Perceptions of Digital Health's Impact on the Patient-Physician Relationship in the Primary Health Care Setting: Qualitative Descriptive Study
J Med Internet Res 2024;26:e53705
URL: https://www.jmir.org/2024/1/e53705
doi:10.2196/53705
PMID:

©Kai Ping Sze, Qi Wei Fong, Jacqueline Giovanna De Roza, Eng Sing Lee, Shu Yun Tan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Health Information Seeking, eHealth Literacy, and Health Behaviors Among Chinese Internet Users: Cross-Sectional Survey Study

Diyi Liu[1], MA; Shuhang Yang[2], MA; Calvin Yixiang Cheng[1], MA; Lin Cai[2], PhD; Jing Su[3], PhD

[1]Oxford Internet Institute, University of Oxford, Oxford, United Kingdom
[2]Chinese Academy of Cyberspace Studies, Beijing, China
[3]School of Humanities, Tsinghua University, Beijing, China

Corresponding Author:
Jing Su, PhD
School of Humanities
Tsinghua University
314A7 Mingli Building
Beijing, 100084
China
Phone: 86 13520625867
Email: sujingxw@tsinghua.edu.cn

Abstract

Background: The internet has become an increasingly vital platform for health-related information, especially in upper-middle-income countries such as China. While previous research has suggested that online health information seeking (OHIS) can significantly impact individuals' engagement in health behaviors, most research focused on patient-centered health communication.

Objective: This study aims to examine how OHIS influences health behavior engagement among Chinese internet users, focusing on the role of eHealth literacy and perceived information quality in influencing relationships.

Methods: An online cross-sectional survey was conducted in November 2021 among 10,000 Chinese internet users, using quota sampling based on sex, age, and urban and rural residence, in line with the 48th Statistical Report on Internet Development of China. Nonparametric tests were used to examine the differences in eHealth literacy across sociodemographic groups. Partial correlation analysis and stepwise linear regression were conducted to test the associations between key variables. Confirmatory factor analysis and structural equation modeling were conducted to test the hypotheses.

Results: Our study identified significant disparities in functional and critical eHealth literacy between urban and rural residents across age groups, income levels, education backgrounds, and health conditions (all $P<.001$). In terms of sex and regional differences, we found higher functional literacy among female users than male users, and critical literacy varied significantly across different regions. The proposed structural model showed excellent fit ($\chi^2_{404}=4183.6$, $\chi^2_{404}=10.4$, $P<.001$; root mean square error of approximation value of 0.031, 95% CI 0.030-.031; standardized root mean square residual value of 0.029; and comparative fit index value of 0.955), highlighting reciprocal associations between 2 types of eHealth literacy and OHIS. Participants' functional eHealth literacy, critical eHealth literacy, and OHIS have positive impacts on their health behavioral engagement. Perceived information quality was found to mediate the influence of OHIS on health behavior (b=0.003, 95% CI 0.002-0.003; $P<.001$).

Conclusions: The study revealed the pathways linking sociodemographic factors, eHealth literacy, OHIS, and perceived information quality and how they together influenced health outcomes. The findings underscore the significance of enhancing eHealth literacy and improving information quality to promote better health outcomes among Chinese internet users.

(J Med Internet Res 2024;26:e54135) doi:10.2196/54135

KEYWORDS

eHealth literacy; online health information seeking; health behavior; perceived information quality; health promotion; China; mobile phone

XSL•FO
RenderX

Introduction

Background

The internet has become crucial for health information dissemination in the digital era [1,2]. Online health information seeking (OHIS) has gained popularity due to its accessibility, wide information coverage, ease of use, affordability, and anonymity [3,4], especially in upper-middle-income countries such as China, where health care resources and in-person medical appointments are limited [5,6]. As of December 2022, China's internet penetration rate has reached 75.6%, with 1.067 billion internet users accessing health information on the web [7]. The shift to digital health is essential for improving health knowledge, increasing confidence in managing health issues, and promoting healthy behaviors [8,9]. Nevertheless, previous research has noted that information alone might not be sufficient to affect optimal health-related well-being [10]. The notion of health literacy, first introduced in China in 2005 and widely acknowledged as a cost-effective measure for improving public health [11], has evolved into eHealth literacy, which is a foundational skill set that underpins the use of information and communication technologies for health [12-14]. While research has examined digital health literacy and health information seeking among specific user segments within China, such as community-dwelling older population [9], rural residents [15], or active social media users [16], and some studies have leveraged nationwide survey samples [6,17], a comprehensive analysis of the broader population's engagement with online health resources is needed.

OHIS and Health Behaviors

Health behavior, distinct from medical treatment, concerns how health interventions and societal norms affect the health of individuals' lifestyles [18]. While previous research has extensively focused on information seeking and literacy in patient-centered health communication [19,20], OHIS is not limited to individuals facing health threats [1]. More commonly, the general public turns to the internet to find information on leading healthier lifestyles, which warrants examination into how the wider population uses the internet for health-related decisions [21]. By providing access to more health-relevant information, OHIS might enable a more accurate assessment of health status, disease outbreak severity, and the need for health protection measures [8]. It could contribute to individuals' perceived control over health threats and reduce negative emotions associated with uncertainty [22]. As most existing literature has acknowledged the potential of OHIS to contribute to better health outcomes, we first examine the direct path.

- Hypothesis 1: OHIS is positively associated with individuals' engagement in health behaviors.

Linking eHealth Literacy With OHIS and Health Behaviors

Among existing scholarship on internet-facilitated health communication and promotion, the integrative model of eHealth use underscores that macrolevel social disparities, often known as the digital divide in the realm of health communication [23,24], manifest as microlevel individual variations in the orientation and efficacy of OHIS, influencing people's internet use for health-related purposes and ultimately their health outcomes [25]. For instance, previous research found that female, older, married, and better-educated Chinese internet users engage in OHIS more frequently [6] and that people's health status could alter their health behavioral engagement [26].

As the integrative model of eHealth use suggests, online health resources can only improve health outcomes if the public has adequate eHealth literacy and avoids low-quality and harmful content [27]. eHealth literacy refers to a person's perceived ability to (1) have access to health information on the web and (2) understand the health information accessed [13]. It not only affects individuals' OHIS behaviors [28-30] but also directly and indirectly engages with sequential health outcomes [26,31]. Centering on the cognitive mechanisms linking OHIS and health behaviors, the study delved into how 2 interrelated dimensions of transactional eHealth literacy influence the process. *Functional eHealth literacy* refers to individual users' perceived ability to acquire online health information, while *critical eHealth literacy* involves more advanced cognitive processes related to information appraisal, including evaluating the reliability, validity, credibility, and applicability of health information [32]. We first examine the impact of Chinese internet users' sociodemographic factors and health status on their eHealth literacy.

- Research question 1: How are Chinese netizens' sociodemographic factors and health status associated with their (1) functional and (2) critical eHealth literacy, respectively?

Higher eHealth literacy might encourage users to resort to the internet, leading to a stronger likelihood of OHIS under health motivation [25,33]. With better functional and critical literacy, individuals would have more ability to access health information and make efficient use of online searching tools and technological devices [28,34]. Conversely, people with less knowledge and confidence in using digital devices are found less likely to conduct OHIS [35,36]. While an association between eHealth literacy and actual eHealth use is recognized, previous research has lent mixed support for the direction of the relationship, and we cannot rule out the possibility of a bidirectional association. Moreover, individuals' personal traits may be associated with both eHealth literacy as well as OHIS [26,37], thus further confounding the process. In this study, we deployed structural equation modeling (SEM) to assess the possible bidirectional relationship between the 2 types of eHealth literacy and OHIS. High critical literacy might contribute to better well-being and engagement in healthy behaviors [38,39], as it can increase levels of self-engagement, initiative, and control over health concerns and self-care management [19]. Therefore, we propose 2 hypotheses.

- Hypothesis 2: Chinese internet users' (1) functional and (2) critical eHealth literacy have reciprocal associations with their OHIS.
- Hypothesis 3: Chinese internet users' (1) functional and (2) critical eHealth literacy are positively associated with their engagement in health behaviors.

XSL•FO
RenderX

Mediator: Perceived Quality of Online Health Information

Apart from eHealth literacy, users' evaluation and satisfaction with obtained health information or experiences using different sources play an important role in health outcomes [40]. Extant theoretical models suggest that the effectiveness of interactive media use for health information is conditioned on users' information processing while engaged with the media and content, which may or may not lead to optimal health outcomes. For instance, the 3-stage model of health promotion using interactive technology proposed that the use of interactive online health resources functions through the interplay of the characteristics of user, media, and health message, potentially contributing to users' health maintenance and improvement [41,42]. Our study specifically focuses on users' attitudes toward the quality of health sources and the information itself [43], as a previous study denoted that content-related indicators and criteria were used the most in credibility evaluation across different information sources compared with the other functionalities [44]. Specifically, we measure perceptions of online health information in terms of scientific rigor, timeliness, accuracy, objectiveness, credibility, applicability, as well as potential harm. We propose that the psychological mechanism underlying health information processing is contingent on individuals' evaluation of the quality of online health information.

- Hypothesis 4: Chinese internet users' perceived quality of online health information will mediate the relationship between OHIS and health behavioral engagement.

Objectives

This study aimed to examine the underlying mechanism by which OHIS influences the engagement in health behaviors of Chinese internet users across various sociodemographic groups. Our conceptual framework proposes a structural model highlighting the interrelationships among users' OHIS behavior, their self-assessed ability to acquire information on the web (functional literacy), and their ability to critically appraise acquired health information (critical literacy), which potentially affect their health outcomes. In addition, the study accounts for the mediating effect of users' perceptions of online health information quality on the pathways.

Methods

Study Design

A cross-sectional online survey was conducted from November 1, 2021, to November 26, 2021, among Chinese internet users accessing health information on the web. The reporting follows the Checklist for Reporting Results of Internet e-Surveys (CHERRIES; Multimedia Appendix 1 [45,46]).

Our questionnaire was developed with input from media, communication, and public health experts to ensure content relevance and clarity. A group of subject matter experts reviewed the original questionnaire and the adapted scales for eHealth literacy and health behavior. The sample size was determined to balance statistical power with budgetary constraints. We targeted a 95% CI with a 1% margin of error for China's internet population (1.011 billion) as of June 2021. In addition, we followed the rule of thumb for the minimum sample size for SEM based on our proposed model [47]. We used quota sampling, considering national representativeness based on sex, age groups, as well as urban and rural residence. Quotas were determined using data from the 48th Statistical Report on Internet Development of China conducted by the China Internet Network Information Center [48]. The survey was distributed through a professional Chinese online survey platform, IDiaoYan (Zhongyan Technology) [45]. We conducted a closed survey accessible exclusively to registered panelists.

Our study initially reached 96,335 registered users of the platform, among which 17.41% (n=16,774) users clicked into the first page of the survey. In a prescreening question, we asked whether respondents used the internet for health information, and 27.64% (4637/16,774) of the participants who answered "no" to this question were screened out. Of the 12,137 remaining responses, 7.34% (n=891) were incomplete and 8.36% (n=1015) participants were unable to submit responses due to quota sampling limitations, where response collection ceases once the quota for a specific group is reached. We also excluded invalid responses (231/12,137, 1.9%) based on the following criteria: (1) completion time <3 minutes, (2) failed logic checks, or (3) filled-in demographic information misaligned with the registered profile in the panel. Consequently, the final sample included 10,000 complete and valid responses. The survey development is illustrated in the flowchart of Figure 1, and the questionnaire in English and original language (Chinese) are included in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. Flowchart illustrating survey development and data screening process.

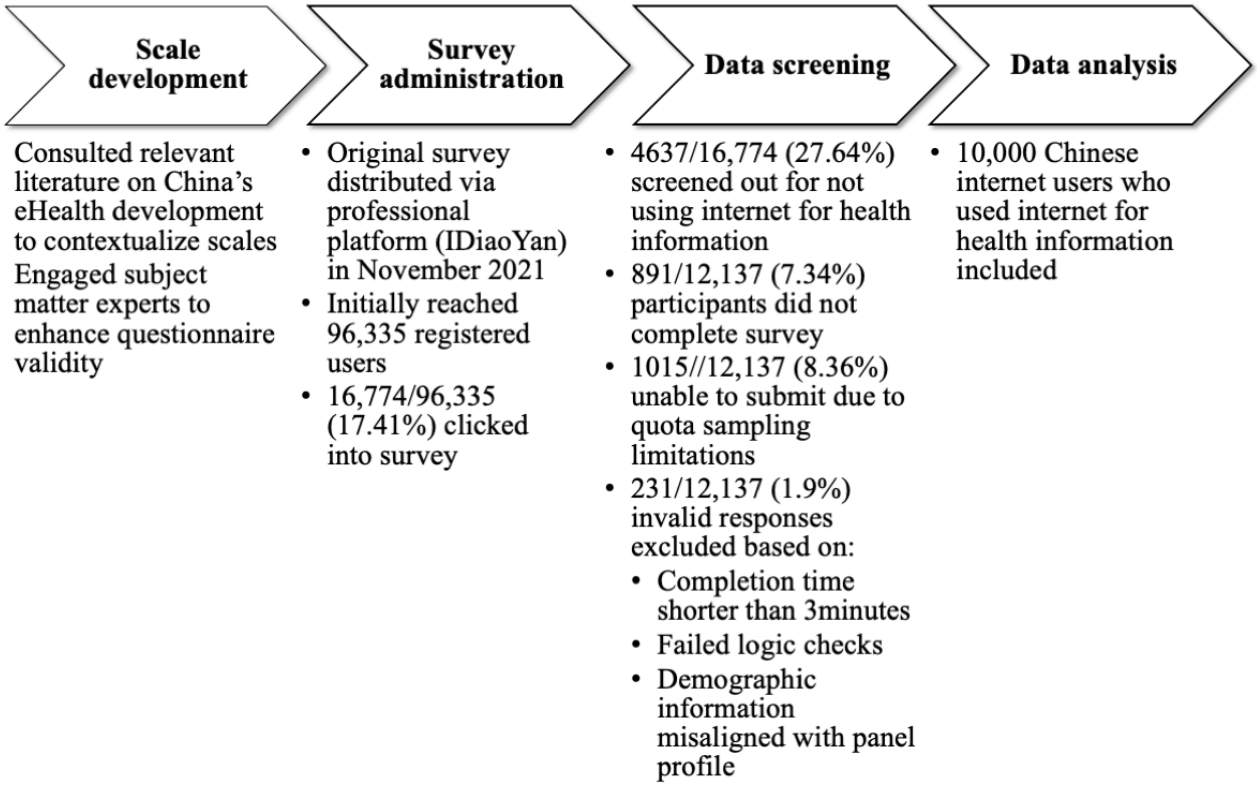

Ethical Considerations

The research protocol for this study was reviewed and approved by the School of Humanities of Tsinghua University and Chinese Academy of Cyberspace Studies. At the beginning of the survey, participants were presented with information about the purpose and procedures of the study, as well as how the data would be handled. They were only allowed to proceed with the study after reading the information and providing informed consent. For participants aged <18 years, parents or guardians provided consent on their behalf, and these participants were required to complete the survey under parental guidance. Participation was on a voluntary basis. All collected data were anonymized, removing any identifiers that could directly or indirectly link the data to individual participants. The data collection and storage protocols were in full compliance with the Personal Information Protection Law of China. Respondents who submitted valid responses were rewarded with 5000 bonus points, equivalent to CNY 5 (US $0.8) through the survey platform's loyalty points program.

Measurements

Frequency of OHIS

Drawing on previous literature [41,49], we operationalized individuals' OHIS as the frequency of participants' engagement in health information seeking through different online channels. Participants were asked to indicate how often they sought health information from (1) mainstream media, (2) professional health media, (3) aggregator news platforms, (4) web portals, (5) open forums, (6) online support forums, (7) search engines, and (8) individual social media accounts. Examples of each information source were provided, as shown in Figure 2 and Multimedia Appendix 2. Responses ranged from "1=never" to "5=always," indicating the frequency of OHIS through each channel. Our intended latent construct of OHIS showed good reliability (Cronbach α=0.85; mean 3.30, SD 0.77).

XSL•FO
RenderX

Figure 2. Frequency of use and perceived credibility of different health information channels (N=10,000).

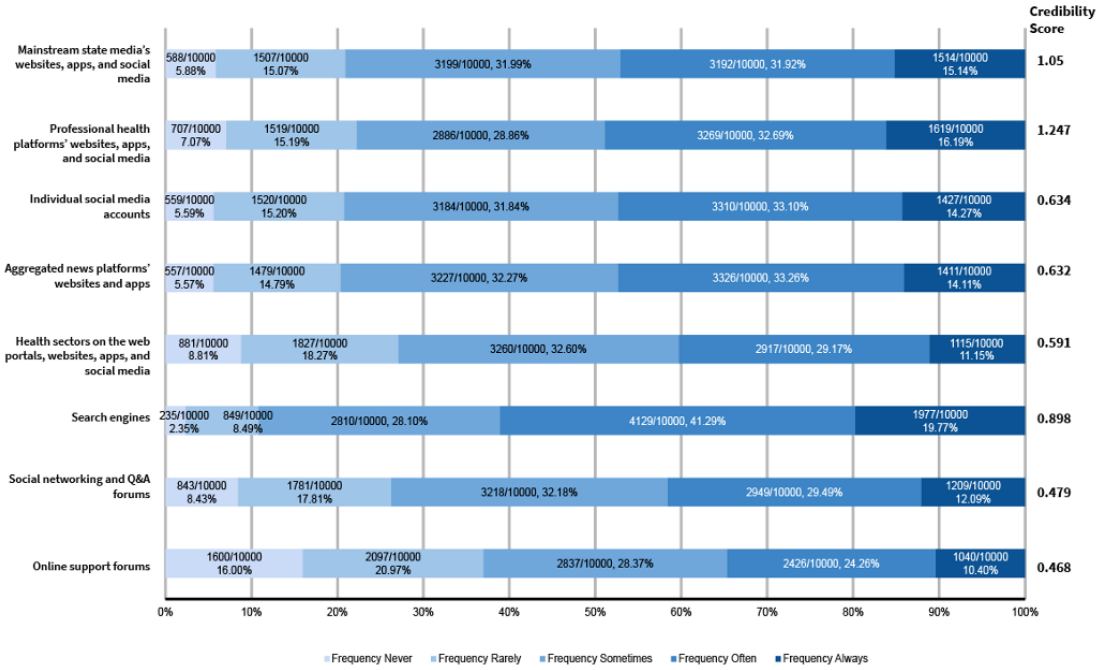

Credibility of Information Source

We calculated a credibility score for each information channel by reverse coding the credibility ranks given by participants. Specifically, participants ranked their top 3 information sources in order of perceived credibility. We later assigned a weight of 3 to the most credible, 2 to the second most credible, and 1 to the third most credible source. We then averaged these weighted rankings across participants to examine the credibility score for each of the 8 sources assessed. The resulting credibility scores ranged from 0 to 3, where a higher score signified greater perceived credibility.

eHealth Literacy

eHealth literacy was assessed using a modified version of the eHealth Literacy scale, which had been translated into Chinese and demonstrated good reliability and validity [13,50]. Our study introduced a 2-factor model based on functional and critical literacy. Specifically, we assessed *functional literacy* by asking participants to what extent they were capable of obtaining and accessing health information in the Web 2.0 context [13,21]. Sample items include "I know what health resources are available on the Internet" and "I am capable of using mobile devices to search for health information." *Critical literacy* focused on how participants assessed and evaluated health information obtained from the internet to make health decisions [14,51], using items such as "I can tell high-quality from low-quality health resources on the Internet." Response options for both variables included a 5-point Likert scale ranging from "1=totally disagree" to "5=totally agree." The scores were then averaged to generate 2 indices of users' functional literacy (Cronbach α=0.804; mean 3.84, SD 0.63) and critical literacy (Cronbach α=0.73; mean 3.55, SD 0.68), showing good reliabilities of the adapted eHealth literacy scales.

Health Behavior

Our study measured health maintenance behaviors using 4 self-reported items that represent people's commitment to reduce negative health outcomes and facilitate psychological and behavioral well-being [52]. The scale was revised based on previous literature [26]. Examples of the items include mental health management as well as good hygiene habits. Participants rated these 4 items on a 5-point scale ranging from "1=not at all" to "5=extremely well." Cronbach α for this variable was reasonably reliable (Cronbach α=0.74; mean 3.77, SD 0.69).

Perceived Quality of Online Health Information

We measured participants' perceived quality of online health information using 5 items assessing the reliability, accuracy, and applicability of health-related online content [28,43]. Respondents were asked to evaluate the quality of internet health information based on to what degree they think that online health information (1) is supported by reliable scientific evidence; (2) is credible and reliable; (3) does not pose a risk to their personal health and well-being; (4) aligns with the latest advancements and consensus within the medical science community; and (5) is actionable and applicable in real-life scenarios. Responses were scored on a 5-point scale from "1=strongly disagree" to "5=strongly agree" (Cronbach α=0.77; mean 3.51, SD 0.65).

Control Variables

We collected participants' sex, age groups, education levels, monthly income, residential areas, and provinces. In addition, participants self-evaluated health status was also measured using scales validated by a previous study [17].

Data Analysis

In our study, we first conducted univariate and multivariate analyses, partial correlation, and regression analysis using SPSS software (version 29.0; SPSS Inc). To assess the normality of eHealth literacy, OHIS, health behavior, and perceived quality, we applied the Kolmogorov-Smirnov test. Non-normal variables

XSL•FO
RenderX

were reported as mean (SD) as well as median (IQR). Categorical variables were reported as frequency and percentages. Differences between sex (female and male) and residence types (rural and urban) were examined using the Mann-Whitney *U* test, while the Kruskal-Wallis test was used for comparison across various age groups, educational levels, income brackets, regions, and health status. We considered *P* values <.05 (2-sided) to be statistically significant. Partial correlations were examined among functional and critical eHealth literacy, OHIS, evaluation, and health behavior while controlling age, sex, education, income, health status, residence, and regional distribution. We then conducted stepwise linear regression with OHIS and health behavior as dependent variables, respectively, and categorical user characteristics, eHealth literacy, and perceived health information quality as independent variables. Using a stepwise method ($P<.05$ as the criterion for entry and $P>.10$ as the criterion for exclusion), all possible combinations of variables were first tested, and the best combinations were selected based on model fit and significance. In addition, we assessed multicollinearity in our regression by examining the variance inflation factor.

To test our hypotheses, we used a 2-step approach following previous research [49,53]. First, to assess the reliability and validity of the latent variables in our model, we performed a confirmatory factor analysis (CFA). This step evaluated the proposed measurement model. Given that we have a relatively large sample size (N=10,000) that strongly influences the result of the chi-square test, we did not refer to the cutoff value 3 of normalized chi-square but the other absolute and incremental model fit indices (eg, root mean square error of approximation [RMSEA], comparative fit index [CFI], and standardized root mean square residual [SRMR]). After the CFA, we applied SEM to examine the pathways among our key variables while controlling for sex, age groups, residence, educational levels, income, and health status. We used a Bootstrap analysis with a 95% CI to estimate the parameters and their associated SEs. Both the CFA and SEM analyses were conducted using the *Lavaan* package [54] in R software (version 4.2.2; R Foundation for Statistical Computing).

Results

Preliminary Analysis

Participants Characteristics and OHIS Engagement

Our study sample included 10,000 participants from 31 provinces in mainland China. The details of participants' demographic distribution and self-reported health status are presented in Table 1. Table S1 in Multimedia Appendix 3 contains detailed descriptive statistics of eHealth literacy across groups, and Table S2 of Multimedia Appendix 3 contains provincial distribution of participants' eHealth literacy. The sampling followed the 48th Statistical Report on Internet Development of China, where most Chinese internet users resided in urban areas (7060/10,000, 70.6%), were aged between 20 and 49 years (5640/10,000, 56.4%), and were relatively evenly distributed between female (4880/10,000, 48.8%) and male (5120/10,000, 51.2%) users. It should be noted that our respondents were mostly well educated, with about 60% holding a bachelor (3393/10,000, 33.9%) or associated degree (2588/10,000, 25.9%). Furthermore, more than half (5628/10,000, 56.3%) of the participants were from China's eastern provinces. Users' health status varied, with most self-reporting good health condition (3777/10,000, 37.8%) and 26.6% (2661/10,000) experiencing subhealth symptoms, such as fatigue and poor appetite. Severe health conditions, such as cancer, were relatively rare among all participants (46/10,000, 0.5%).

XSL•FO
RenderX

Table 1. Demographic characteristics, health status, and distribution of eHealth literacy (N=10,000).

Demographics	Participants, n (%)	Test statistics for functional literacy[a]	*P* value	Test statistics for critical literacy	*P* value
Sex		3.121	.002	–0.973	.33
Female	4880 (48.8)				
Male	5120 (51.2)				
Age group (y)		242.323	<.001	116.489	<.001
<19	1560 (15.6)				
20 to 29	1740 (17.4)				
30 to 39	2030 (20.3)				
40 to 49	1870 (18.7)				
50 to 59	1590 (15.9)				
>60	1210 (12.1)				
Residential area		–10.293	<.001	–3.740	<.001
Urban	7060 (70.6)				
Rural	2940 (29.4)				
Education level		192.303	<.001	62.955	<.001
Primary school or less	174 (1.74)				
Middle school	1066 (10.66)				
High school or secondary vocational school	2499 (24.99)				
Associate degree	2588 (25.88)				
Bachelor degree	3393 (33.93)				
Master and above	280 (2.8)				
Income level [b]		128.293	<.001	226.919	<.001
<¥1500	1267 (12.67)				
¥1500 to 3000	1286 (12.86)				
¥3001 to 5000	2294 (22.94)				
¥5001 to 8000	2632 (26.32)				
¥8001 to 12,000	1629 (16.29)				
¥12,001 to 20,000	693 (6.93)				
>¥20,000	199 (1.99)				
Health status		74.46	<.001	17.458	<.001
Experiencing a severe disease	46 (0.46)				
Experiencing chronic diseases	1397 (13.97)				
Subhealth symptoms	2661 (26.61)				
Not bad	2119 (21.19)				
Good	3777 (37.77)				
Region		4.57	.03	49.240	<.001
East	5628 (56.28)				
Central	1998 (19.98)				
West	1630 (16.3)				
Northeast	744 (7.44)				

[a]Test statistics were reported for dichotomous variables (*z* score) and multicategorical variables (χ^2) alongside *P* value.

[b]A currency exchanged rate of 1¥=US $0.16 is applicable.

XSL•FO
RenderX

We measured OHIS by examining the frequency of use of online sources. As shown in Figure 2, search engines, such as Baidu, emerged as primary tools for health-related inquiries ("often": 4130/10,000, 41.3% and "always": 1980/10,000, 19.8%). Conversely, emerging specialized online support forums, such as the bulletin board system forum for patients with diabetes, saw less frequent engagement, with 16% (1600/10,000) of the participants indicating that they "never" have used such platforms. Regarding source credibility, among the 8 channels assessed, professional health media outlets were perceived as the most credible. State media outlets, such as Xinhua News and People's Daily, were also rated as relatively credible. However, social networking (question and answer) forums and online support communities were regarded as less credible.

Univariate Analysis of eHealth Literacy Across Different Sociodemographic Groups

Our analyses revealed differences in both functional and critical eHealth literacy across various demographic, socioeconomic, and health-related variables. The detailed descriptive statistics across groups can be found in Table S1 in Multimedia Appendix 3. The Mann-Whitney *U* test indicated that female users have a significantly higher functional literacy than their male counterparts (*z* score=3.12; *P*=.002). No significant difference was found in critical literacy (*z* score=–0.97; *P*=.33). A digital divide was observed between urban and rural residents, evident in both functional literacy (*z* score=–10.29; *P*<.001) and critical literacy (*z* score=–3.74; *P*<.001). Multicategorical group comparison showed variations in literacy levels across age cohorts in functional literacy (χ^2_5=242.3; *P*<.001) and critical literacy (χ^2_5=116.5; *P*<.001). Internet users with different income levels demonstrated variations in both functional literacy (χ^2_6=128.3; *P*<.001) and critical literacy (χ^2_6=226.9; *P*<.001). Similarly, respondents with different educational backgrounds also showed varied levels of functional literacy (χ^2_5=192.3; *P*<.001) and critical literacy (χ^2_5=62.9; *P*<.001). Furthermore, there were marked differences based on participants' health conditions in functional literacy (χ^2_4=74.5; *P*<.001) and critical literacy (χ^2_4=17.5; *P*<.001). Notably, we did not observe significant regional differences in internet users' functional literacy (χ^2_3=3.6; *P*=.21); however, users' critical literacy demonstrated significant regional variations across East, Central, West, and Northeast China (χ^2_3=49.2; *P*<.001).

Multivariate Analysis on Social and Individual Differences of OHIS and Health Behavior

Partial correlation analysis, as detailed in Table 2, revealed significant positive associations among the 5 key variables: functional literacy, critical literacy, OHIS, perceived information quality, and health behavior, controlling for covariates (all *P*<.001).

Table 2. Partial correlation analysis (Pearson *r* and 2-tailed *P* value) among key variables[a].

	Functional literacy	Critical literacy	Health behavior	Perceived quality	OHIS[b]
Functional literacy					
r	1	0.606	0.331	0.416	0.136
P value	__[c]	<.001	<.001	<.001	<.001
Critical literacy					
r	0.606	1	0.320	0.540	0.255
P value	<.001	—	<.001	<.001	<.001
Health behavior					
r	0.331	0.320	1	0.290	0.257
P value	<.001	<.001	—	<.001	<.001
Perceived quality					
r	0.416	0.540	0.290	1	0.276
P value	<.001	<.001	<.001	—	<.001
OHIS					
r	0.136	0.255	0.257	0.276	1
P value	<.001	<.001	<.001	<.001	—

[a]Partial correlation coefficients were calculated with sex, age groups, residence, region, education, income, and health status as control variables.

[b]OHIS: online health information seeking.

[c]Not applicable.

The results of stepwise linear regression investigated the associations between user-oriented characteristics and their OHIS and health behavioral engagement. Users' sex and regional differences were excluded in the final model predicting OHIS (Table 3), while users' residence types and regional differences were excluded in the final model predicting health behavior (Table 4).

XSL•FO
RenderX

Aside from the excluded factors, all considered sociodemographic factors, eHealth literacy, and perceived information quality showed statistically significant main effects on both OHIS and health behaviors (all P<.001). Furthermore, the variance inflation factor values were <2 across the model, suggesting no collinearity issues and thereby affirming the reliability of the regression analyses.

Table 3. Stepwise linear regression predicting online health information seeking (OHIS)[a].

Included predictors	b[b] (SE)	β	*t* test (*df*)	*P* value	VIF[c]
Income	0.10 (0.01)	.20	17.38 (9991)	<.001	1.60
Age	0.04 (0.01)	.07	6.13 (9991)	<.001	1.73
Health condition	–0.04 (0.01)	–.06	–6.11 (9991)	<.001	1.35
Residence	0.11 (0.02)	.06	5.97 (9991)	<.001	1.35
Education level	0.05 (0.01)	.07	5.85 (9991)	<.001	1.73
Functional literacy	–0.07 (0.01)	–.06	–4.86 (9991)	<.001	1.66
Critical literacy	0.20 (0.01)	.17	13.85 (9991)	<.001	1.92
Perceived quality	0.23 (0.01)	.20	17.75 (9991)	<.001	1.51

[a]The result shows the final model in stepwise regression predicting OHIS: R^2=0.187, adjusted R^2=0.186, $F_{1,9991}$=315.21 (P<.001). Sex and regional distribution were excluded.

[b]b: unstandardized coefficient.

[c]VIF: variance inflation factor.

Table 4. Stepwise linear regression predicting health behavior[a].

Included predictors	b[b] (SE)	β	*t* test (*df*)	*P* value	VIF[c]
Income	0.03 (0.01)	.06	5.62 (9991)	<.001	1.54
Age	0.06 (0.01)	.15	12.75 (9991)	<.001	1.71
Education level	0.04 (0.01)	.07	6.07 (9991)	<.001	1.62
Health condition	0.05 (0.01)	.08	7.22 (9991)	<.001	1.35
Sex	0.06 (0.01)	.05	5.08 (9991)	<.001	1.03
Functional literacy	0.22 (0.01)	.20	16.84 (9991)	<.001	1.66
Critical literacy	0.12 (0.01)	.12	9.81 (9991)	<.001	1.92
Perceived quality	0.15 (0.01)	.14	12.69 (9991)	<.001	1.51

[a]The result shows the final model in stepwise regression predicting health behavior: R^2=0.187, adjusted R^2=0.187, $F_{1,9991}$=160.99 (P<.001). Residence and regional distribution were excluded.

[b]b: unstandardized coefficient.

[c]VIF: variance inflation factor.

Hypothesis Testing

Our research model posited that participants' functional and critical eHealth levels exhibited reciprocal relationships with OHIS. This, in turn, was posited to enhance their perceptions of the quality of online health information, thereby increasing engagement in health-promoting behaviors. The model mapped all direct and indirect pathways linking 5 key variables, with the literacy-OHIS pathways modeled as bidirectional. We also estimated the covariance between these 5 latent variables in our model, considering their correlations. To control for confounding effects, we included sex, age group, residential area, educational background, income, and health status as covariates in our structural model. Notably, based on the results of stepwise linear regression analyses, internet users' regional difference was not incorporated as a covariate in the final model.

Measurement Model

Above all, our measurement model (Table 5) demonstrated excellent model fit (χ^2_{289}=2889.03, χ^2/df=10, P<.001; RMSEA=0.030, 95% CI 0.029-0.031; SRMR=0.029; CFI=0.968), underscoring the structural integrity of the 5 key constructs. One item from OHIS (ie, seeking information via search engine) was dropped due to low factor loading. All remaining factor loadings were above the recommended threshold of 0.5, showing acceptable indicator reliability. The constructs also exhibited satisfactory content reliability, with Cronbach α coefficients ranging from 0.73 to 0.86 and composite reliability values spanning 0.73 to 0.86. While our average variance extracted (AVE) fell slightly below the standard 0.5 threshold, Fornell and Larcker [55] have discussed that given that AVE is a more stringent measurement,

XSL•FO
RenderX

researchers might still conclude the establishment of convergent validity with satisfactory composite reliability. This notion is particularly relevant to our study's tailored adaptation of existing scales. Given these considerations and supported by supplementary research [56], we concluded that the measurement model has established convergent validity based on (1) AVE values marginally <0.5; (2) factor loadings all >0.5, showing strong item-to-construct relationships; and (3) composite reliability values >0.7 across all constructs, and thus collectively affirming the convergent validity of the model.

XSL•FO
RenderX

Table 5. Statistical outcomes of confirmatory factor analysis[a].

Constructs	Standardized factor loading	*z* score	AVE[b]	CR[c]
OHIS[d] (Cronbach α=0.86)			0.467	0.86
C1: mainstream media	0.674	—[e]		
C2: professional health media	0.663	54.046		
C3: aggregator platforms	0.647	59.950		
C4: web portals	0.779	65.443		
C5: open Q & A[f] forums	0.666	58.870		
C6: online support forums	0.757	63.242		
C7: individual social media accounts	0.542	46.222		
Functional literacy (Cronbach α=0.80)			0.406	0.80
FL1: I know how to find helpful health resources on the internet.	0.651	—		
FL2: I know what health resources are available on the internet.	0.633	51.207		
FL3: I am capable of using mobile devices (eg, smartphones and tablets) to search for health information online.	0.635	50.271		
FL4: I can effectively use relevant keywords and logical search operators when querying or retrieving health information online.	0.615	49.369		
FL5: I have the skills to open and navigate different web pages and websites to access health information across the internet.	0.652	49.608		
FL6: I know how to bookmark or save useful health information from online sources.	0.634	49.341		
Critical literacy (Cronbach α=0.73)			0.398	0.73
CL1: I have the skills I need to evaluate the health resources I find on the internet.	0.621	—		
CL2: I can tell high quality from low-quality health resources on the internet.	0.622	46.044		
CL3: I can distinguish between different sources of health information, such as authoritative sources and primary sources (eg, medical records).	0.636	49.964		
CL4: I feel confident in using information from the internet to make health decisions.	0.642	45.410		
Perceived quality (Cronbach α=0.77)			0.401	0.77
E1: The content of online health information is supported by reliable scientific evidence.	0.666	—		
E2: The sources of online health information are credible and reliable.	0.676	52.905		
E3: Accessing online health information does not pose a risk to my personal health and well-being.	0.583	45.871		
E4: Online health information aligns with the latest advancements and consensus within the medical science community.	0.621	50.624		
E5: The advice contained in online health information is actionable and applicable in real-life scenarios.	0.624	49.412		
Health behaviors (Cronbach α=0.74)			0.421	0.74
HB1: dietary balance	0.685	—		
HB2: active exercise	0.664	50.063		
HB3: mental health maintenance	0.656	44.445		
HB4: hygiene behavior	0.584	43.064		

[a]Model fit: χ^2_{289}=2889.03, χ^2/289=10, *P*<.001; root mean square error of approximation of 0.030; standardized root mean square residual of 0.029; comparative fit index value of 0.968.

[b]AVE: average variance extracted.

[c]CR: composite reliability.

[d]OHIS: online health information seeking.

[e]Items constrained for identification purposes.

[f]Q and A: question and answer.

XSL•FO
RenderX

Structural Model

The result of the SEM is presented in Figure 3 and Table 6 (both unstandardized and standardized coefficients are reported). Several model fit measures suggest that our final model was found to be a good fit to the data with χ^2_{404}=4183.6, $\chi 2/df$=10.36 (*P*<.001), RMSEA=0.031 (95% CI 0.030-0.031), SRMR=0.029, and CFI=0.955.

Figure 3. Path analysis results of structural equation model. Arrows depict the direction of effect. Reported values are standardized coefficients (β). OHIS: online health information seeking.

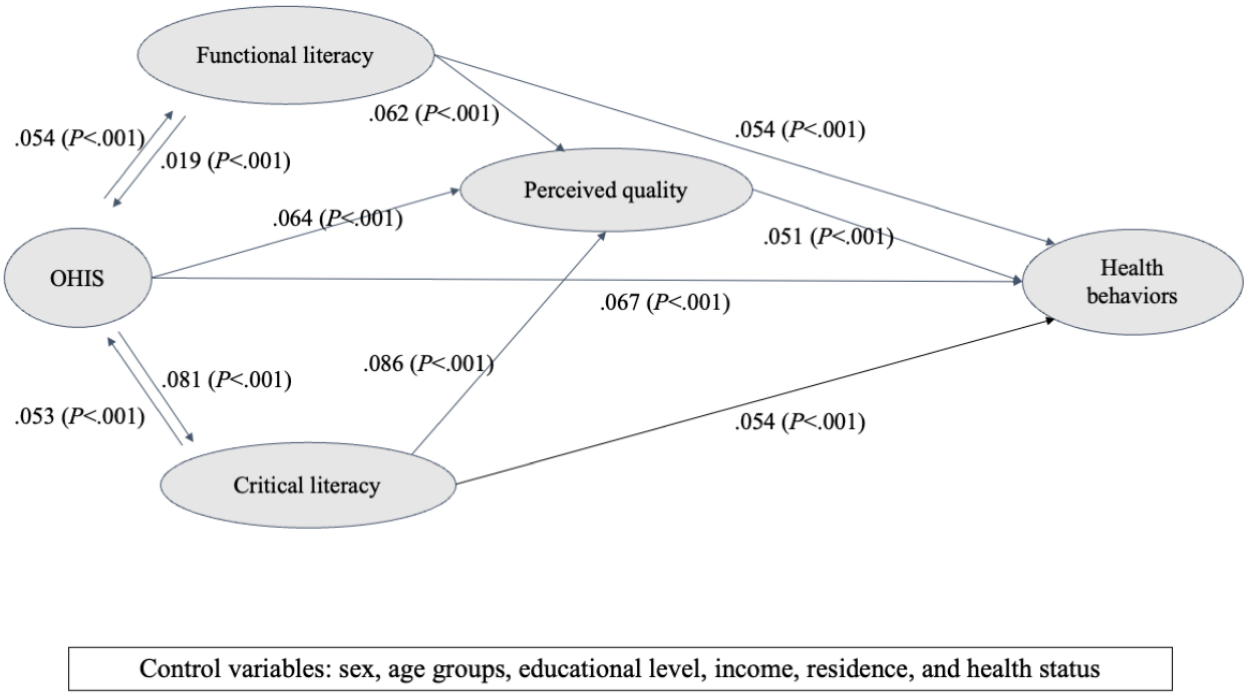

Table 6. Hypotheses testing for direct effects[a].

Hypothesized direct paths	b[b] (SE)	*P* value	β	Results
Hypothesis 1: OHIS[c] to health behavior	0.057 (0.003)	<.001	.067	√
Hypothesis 2a				
Functional literacy to OHIS	0.025 (0.002)	<.001	.019	√
OHIS to functional literacy	0.042 (0.003)	<.001	.054	√
Hypothesis 2b				
Critical literacy to OHIS	0.068 (0.003)	<.001	.053	√
OHIS to critical literacy	0.063 (0.003)	<.001	.081	√
Hypothesis 3a: functional literacy to health behavior	0.060 (0.003)	<.001	.054	√
Hypothesis 3b: critical literacy to health behavior	0.059 (0.003)	<.001	.054	√
Hypothesis 4a: OHIS to perceived quality	0.051 (0.003)	<.001	.064	√
Hypothesis 4b: perceived quality to health behavior	0.054 (0.002)	<.001	.051	√

[a]χ^2_{404}=4183.6, χ^2/404=10.36 (*P*<.001), root mean square error of approximation of 0.031 (95% CI 0.030-0.031), standardized root mean square residual of 0.029, and comparative fit index value of 0.955.

[b]b=unstandardized coefficient. Bootstrap sample=5000 with replacement.

[c]OHIS: online health information seeking.

Specifically, for hypothesis testing, our result showed that OHIS is positively associated with individuals' engagement in health behaviors (b=0.057, SE 0.003, *P*<.001, β=.067). Hypothesis 2 aims to test the relationships between functional literacy, critical eHealth literacy, and OHIS. Considering the statistical significance and relative magnitudes of the standardized regression coefficients of the 4 paths, we conclude that there are bidirectional relationships between functional or critical eHealth literacy and OHIS. Specifically, participants' functional literacy positively predicts the frequency of OHIS (b=0.025,

XSL•FO
RenderX

SE 0.002, $P<.001$, β=.019), while the opposite path is also significant (b=0.042, SE 0.003, $P<.001$, β=.054). Furthermore, respondents' critical literacy positively predicts their OHIS (b=0.068, SE 0.003, $P<.001$, β=.053), while more frequent information seeking also has a significant effect on critical literacy (b=0.063, SE 0.004, $P<.001$, β=.081). Hypothesis 2 was supported.

Hypothesis 3 tackles the direct relationships between both functional and critical eHealth literacy and health behaviors. Results showed that internet users' functional eHealth literacy is positively associated with their engagement in healthier behaviors (b=0.060, SE 0.003, $P<.001$, β=.054). Critical eHealth literacy also positively predicts health behavioral engagement (b=0.059, SE 0.003, $P<.001$, β=.054). Hypotheses 3a and 3b were supported. OHIS is associated with individuals' evaluation of online health information quality (b=0.051, SE 0.003, $P<.001$, β=.064), and the direct path between perceived information quality and health behaviors is also significantly positive (b=0.054, SE 0.002, $P<.001$, β=.051). To further examine the mediating effect of perceived information quality between OHIS and health outcomes, an unstandardized indirect effect was computed for each of the 5000 bootstrapped samples as well as the 95% CI. The indirect effect was statistically significant that users' evaluation of online information positively mediates the relationship between OHIS and health behaviors (b=0.003, 95% CI 0.002-0.003; $P<.001$). This lent support to the mediating effect of perceived OHIS quality, and thus, we accepted hypothesis 4.

Discussion

Principal Findings

While previous studies have identified various factors that are associated with OHIS or health behavioral intentions in the Chinese context [6,12,16,17], few have illustrated the full pathways from development to application of individual eHealth literacy, information seeking, and subsequent health wellness. Building upon existing theoretical models of eHealth use for health promotion [25,42], this paper examined how Chinese internet users' perceptual abilities in leveraging eHealth information were constituted within macrolevel socioeconomic structures and influenced by individual-level health concerns. We found that eHealth literacy interrelates with use of online health information and ultimately affects users' health behavioral engagement. Furthermore, user perceptions of online health information quality positively mediate the relationships.

Social Divides Influencing Functional and Critical eHealth Literacy

Above all, our results reveal how Chinese netizens' sociodemographic background and personal health status are manifested in individuals' varying levels of health information efficacy. In line with previous studies [9,16], gender differences, education gaps, and income inequality are widely observed among Chinese internet users' eHealth literacy [36,57,58]. Notably, our results showed that the younger generation is not necessarily more capable of obtaining and effectively using online health information tailored to their specific needs compared with previous studies [16,21]. Moreover, the results indicated that subhealthy groups tend to have higher levels of functional and critical literacy. Our findings validated established eHealth use models [59,60] in the Chinese contexts, demonstrating that socioeconomic divides in relation to age, education, and income, together with individual health status, are further factored into people's actual engagement with eHealth and final health outcomes. Using post hoc pairwise comparisons, we revealed the difference in main effects on OHIS and health behaviors between various user groups. Consistent with previous research [6], older, moderately high-income, and more well-educated groups show a significantly higher engagement with OHIS and health maintenance than their corresponding counterparts.

The findings also highlighted how eHealth literacy served as a crucial enabling factor on an individual level. High functional literacy would facilitate users' OHIS by leveraging eHealth resources; thus, users might be more likely to adopt health-promoting behaviors [61,62]. Furthermore, better critical literacy might contribute to better well-being and engagement in health behaviors [38,39], as it can increase levels of self-engagement, initiative, and control over health concerns and self-care management [19]. It is particularly important to individuals for mitigating the effect of online misinformation, especially during public health crises, such as the COVID-19 pandemic, when health misinformation is widespread [63]. Notably, we found that people who are dealing with chronic diseases are more inclined to search for professional medical information through the internet, nevertheless with relatively low eHealth literacy. This observation emphasizes the special needs of the vulnerable demographic, who are particularly sensitive to heightened health risks and susceptible to encountering low-quality eHealth content.

OHIS Contributing to Both eHealth Literacy and Health Behaviors

Our structural model complements existing theoretical frameworks with regard to digital health literacy and OHIS, revealing the reciprocal associations existing between both functional or critical literacy and users' OHIS. People who are more confident in navigating online resources are more inspired to devote sufficient effort to seek and use health resources in effective ways [64]. Furthermore, users' ability to evaluate and apply health information could be better achieved by their actual engagement with the online environment [30], while frequency and diversity of internet use could also contribute to higher eHealth literacy [65]. While individuals with limited eHealth literacy might encounter more perceived barriers in using accurate health information [66], frequent OHIS might contribute to wider access to different health information where individuals can evaluate, compare, and decide how it will facilitate their health decision-making and consequently build individual eHealth literacy [67,68]. Therefore, we underscore the importance of improving general users' functional and critical literacy, and the benefit will be amplified with more OHIS.

XSL•FO
RenderX

Intermediary Roles of User Perceptions of Online Health Information

Last but not least, the way users perceive their ability to access and assess health information through the internet significantly influences their attitudes toward health content. Despite concerns raised by previous studies about cognitive biases in highly literate individuals favoring health information aligning with their existing beliefs, particularly regarding vaccinations [69], our results suggest that highly literate users tend to view general health information when not viewpoint-specific as more accurate, objective, relevant, applicable, and skillfully navigating around low-quality content. The quality assessment of health information affects their willingness to integrate such information into their daily lives [40,70]. For instance, high-quality, well-designed health information websites can foster positive attitudes toward OHIS, offering relevant and effective content that boosts health knowledge and outcomes [40,43]. Conversely, misleading or inaccurate information can detract from OHIS's effectiveness [71].

Limitations

There are limitations in our study that point to future prospects of research. First, our sampling approach may not fully represent the regional, cultural, and economic diversity of the Chinese internet user base. Second, our reliance on self-reported survey data to measure eHealth literacy and health status, despite efforts to encompass its complex nature, may not be as reliable as qualitative observations or experimental measures used in other studies [44,71,72]. Third, the cross-sectional design inhibits our ability to draw causal relationships between variables and is subject to social desirability bias. Future research could benefit from longitudinal, experimental, or observational studies to address these limitations [62]. Finally, the influence of some other sociopsychological determinants, such as perceived health threat and health anxiety, in influencing OHIS and health outcomes needs to be further examined [73]. Integrating these determinants into future studies could offer a more rounded understanding of how OHIS affects health outcomes.

Theoretical and Practical Implications

Despite the abovementioned limitations, our study shifts the scholarly focus of OHIS from patient-centered health communication to the well-being of the general population. We presented a structural model that has expanded existing theoretical frameworks of eHealth use for health promotion and revealed the pathways linking sociodemographic determinants to eHealth literacy, the reciprocal associations between eHealth literacy and OHIS, the mediating mechanisms of perceived information quality, and how they ultimately contribute to health outcomes. Crucially, our study highlights the importance of user abilities to locate relevant and high-quality information sources in influencing health wellness [3,74]. While the lack of a precise definition and shared understanding of eHealth literacy may have impeded the progress of eHealth studies [75], we suggest that a nuanced framework of eHealth literacy could encourage more professional discussions and theory-driven research. Moreover, while prior research often focused on health literacy from a public health standpoint, our study bridges interdisciplinary discussion from public health, communication, as well as media psychology. We examined the mechanisms linking macrolevel social dynamics and cognitive processing of online health information. The potential implications for reducing health disparities linked to socioeconomic factors could be extrapolated to diverse populations and cultural contexts, thus broadening the applicability of our research, especially in the context of non–high-income countries [59].

In practical terms, our study offers actionable insights for policy makers and health practitioners aiming to devise targeted health promotion interventions for the wider public. These efforts should not only focus on enhancing the quality of online health information from the supply side but also aim to shift user attitudes, increase the experience of eHealth use, and foster health behaviors. This is especially important for individuals with limited eHealth literacy, who are at greater risk of encountering poor-quality eHealth information and facing health disparities. In addition, we emphasize the necessity of integrating critical eHealth literacy into health promotion initiatives. The design, implementation, and evaluation of health education and promotion should consider the intricacy between individuals' orientation toward health information, their use of eHealth resources, and the broader structural disparities identified in our research framework. Particularly in non–high-income countries such as China, well-structured information campaigns are pivotal in reducing inequalities in literacy and enhancing the accessibility and utility of health information sources. Our results suggest a need for improvements that cater to the diverse literacy levels within the population, indicating a path forward for reducing health disparities and fostering equitable health outcomes.

Acknowledgments

The study received financial support from the National Social Science Fund of China (grant 22BXW069).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

The study was conceptualized by DL, SY, LC, and JS, and the data were curated by DL, CYC, and JS. Formal analysis was performed by DL. Funding acquisition involved SY, LC, and JS. The methodology was devised by DL, SY, LC, and JS. Supervision

XSL•FO
RenderX

was done by SY, LC, and JS, and validation was done by DL, SY, LC, and JS. The original draft was written by DL and CYC, and the paper was reviewed and edited by DL, CYC, and JS.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist for Reporting Results of Internet e-Surveys (CHERRIES).
[DOCX File , 26 KB - jmir_v26i1e54135_app1.docx]

Multimedia Appendix 2
Survey questionnaire in English and Chinese.
[DOCX File , 40 KB - jmir_v26i1e54135_app2.docx]

Multimedia Appendix 3
Descriptive statistics for participants' eHealth literacy.
[DOCX File , 22 KB - jmir_v26i1e54135_app3.docx]

References

1. Anker AE, Reinhart AM, Feeley TH. Health information seeking: a review of measures and methods. Patient Educ Couns 2011 Mar;82(3):346-354. [doi: 10.1016/j.pec.2010.12.008] [Medline: 21239134]
2. Lambert SD, Loiselle CG. Health information seeking behavior. Qual Health Res 2007 Oct;17(8):1006-1019. [doi: 10.1177/1049732307305199] [Medline: 17928475]
3. Chen X, Hay JL, Waters EA, Kiviniemi MT, Biddle C, Schofield E, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734 [FREE Full text] [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
4. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
5. Hao H. The development of online doctor reviews in China: an analysis of the largest online doctor review website in China. J Med Internet Res 2015 Jun 01;17(6):e134 [FREE Full text] [doi: 10.2196/jmir.4365] [Medline: 26032933]
6. Xiong Z, Zhang L, Li Z, Xu W, Zhang Y, Ye T. Frequency of online health information seeking and types of information sought among the general Chinese population: cross-sectional study. J Med Internet Res 2021 Dec 02;23(12):e30855 [FREE Full text] [doi: 10.2196/30855] [Medline: 34860676]
7. The 51st statistical report on China's internet development. China Internet Network Information Center (CNNIC). 2023 Mar. URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202307/P020230707514088128694.pdf [accessed 2023-08-04]
8. Bennett GG, Glasgow RE. The delivery of public health interventions via the Internet: actualizing their potential. Annu Rev Public Health 2009 Apr;30:273-292. [doi: 10.1146/annurev.publhealth.031308.100235] [Medline: 19296777]
9. Liu S, Zhao H, Fu J, Kong D, Zhong Z, Hong Y, et al. Current status and influencing factors of digital health literacy among community-dwelling older adults in Southwest China: a cross-sectional study. BMC Public Health 2022 May 17;22(1):996 [FREE Full text] [doi: 10.1186/s12889-022-13378-4] [Medline: 35581565]
10. Dupas P. Health behavior in developing countries. Annu Rev Econ 2011 Sep 01;3(1):425-449. [doi: 10.1146/annurev-economics-111809-125029]
11. Li Y, Lv X, Liang J, Dong H, Chen C. The development and progress of health literacy in China. Front Public Health 2022 Nov 7;10:1034907 [FREE Full text] [doi: 10.3389/fpubh.2022.1034907] [Medline: 36419995]
12. Cao W, Zhang X, Xu K, Wang Y. Modeling online health information-seeking behavior in China: the roles of source characteristics, reward assessment, and internet self-efficacy. Health Commun 2016 Sep 09;31(9):1105-1114. [doi: 10.1080/10410236.2015.1045236] [Medline: 26861963]
13. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
14. Paige SR, Stellefson M, Krieger JL, Miller MD, Cheong J, Anderson-Lewis C. Transactional eHealth literacy: developing and testing a multi-dimensional instrument. J Health Commun 2019;24(10):737-748 [FREE Full text] [doi: 10.1080/10810730.2019.1666940] [Medline: 31583963]
15. Qiu Y, Ren W, Liu Y, Yin P, Ren J. Online health information in a rural residential population in Zhejiang Province, China: a cross-sectional study. BMJ Open 2019 May 05;9(5):e026202 [FREE Full text] [doi: 10.1136/bmjopen-2018-026202] [Medline: 31061032]
16. Niu Z, Willoughby J, Zhou R. Associations of health literacy, social media use, and self-efficacy with health information-seeking intentions among social media users in China: cross-sectional survey. J Med Internet Res 2021 Feb 25;23(2):e19134 [FREE Full text] [doi: 10.2196/19134] [Medline: 33629955]

XSL•FO
RenderX

17. Li X, Liu Q. Social media use, eHealth literacy, disease knowledge, and preventive behaviors in the COVID-19 pandemic: cross-sectional study on Chinese netizens. J Med Internet Res 2020 Oct 09;22(10):e19684 [FREE Full text] [doi: 10.2196/19684] [Medline: 33006940]
18. Kasl SV, Cobb S. Health behavior, illness behavior, and sick role behavior: I. Health and illness behavior. Arch Environ Health 1966 Feb;12(2):246-266. [doi: 10.1080/00039896.1966.10664365] [Medline: 5322534]
19. Ishikawa H, Yano E. Patient health literacy and participation in the health-care process. Health Expect 2008 Jun;11(2):113-122 [FREE Full text] [doi: 10.1111/j.1369-7625.2008.00497.x] [Medline: 18494956]
20. Jiang S, Liu J. Examining the relationship between internet health information seeking and patient-centered communication in China: taking into account self-efficacy in medical decision-making. Complement Ther Med 2020 May 25;13(4):407-424. [doi: 10.1080/17544750.2020.1769700]
21. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
22. Shiloh S, Drori E, Peleg S, Banai S, Finkelstein A. Mediation and moderation of the effects of watching the angiography screen on patients. Psychol Health Med 2016 Oct 07;21(7):806-818. [doi: 10.1080/13548506.2015.1131997] [Medline: 26740003]
23. Medero K, Merrill Jr K, Ross MQ. Modeling access across the digital divide for intersectional groups seeking online health information: national survey. J Med Internet Res 2022 Mar 15;24(3):e32678 [FREE Full text] [doi: 10.2196/32678] [Medline: 35289761]
24. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
25. Bodie GD, Dutta MJ, Basu A. The integrative model of e-Health use. In: Lazakidou AA, Siassiakos KM, editors. Handbook of Research on Distributed Medical Informatics and E-Health. New York, NY: IGI Global; 2009:104-116.
26. Hsu W, Chiang C, Yang S. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014 Dec;16(12):e287 [FREE Full text] [doi: 10.2196/jmir.3542] [Medline: 25499086]
27. Sørensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, (HLS-EU) Consortium Health Literacy Project European. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80 [FREE Full text] [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
28. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
29. Hassan S, Masoud O. Online health information seeking and health literacy among non-medical college students: gender differences. J Public Health (Berl.) 2020 Mar 09;29(6):1267-1273. [doi: 10.1007/S10389-020-01243-W]
30. Rosenbaum JE, Johnson BK, Deane AE. Health literacy and digital media use: assessing the health literacy skills instrument - short form and its correlates among African American college students. Digit Health 2018 May 14;4:2055207618770765 [FREE Full text] [doi: 10.1177/2055207618770765] [Medline: 29942630]
31. Wolf MS, Davis TC, Osborn CY, Skripkauskas S, Bennett CL, Makoul G. Literacy, self-efficacy, and HIV medication adherence. Patient Educ Couns 2007 Feb;65(2):253-260. [doi: 10.1016/j.pec.2006.08.006] [Medline: 17118617]
32. McLaughlin M, DeVoogd G. Critical literacy as comprehension: expanding reader response. J Adolesc Adult Lit 2004;48(1):52-62. [doi: 10.1598/jaal.48.1.5]
33. Bandura A. Reflections on self-efficacy. Adv Behav Res Ther 1978 Jan;1(4):237-269. [doi: 10.1016/0146-6402(78)90012-7]
34. Lee HY, Jin SW, Henning-Smith C, Lee J, Lee J. Role of health literacy in health-related information-seeking behavior online: cross-sectional study. J Med Internet Res 2021 Jan 27;23(1):e14088 [FREE Full text] [doi: 10.2196/14088] [Medline: 33502332]
35. Jeong SH, Kim HK. Health literacy and barriers to health information seeking: a nationwide survey in South Korea. Patient Educ Couns 2016 Nov;99(11):1880-1887. [doi: 10.1016/j.pec.2016.06.015] [Medline: 27373962]
36. Tang C, Wu X, Chen X, Pan B, Yang X. Examining income-related inequality in health literacy and health-information seeking among urban population in China. BMC Public Health 2019 Feb 21;19(1):221 [FREE Full text] [doi: 10.1186/s12889-019-6538-2] [Medline: 30791882]
37. Lee CJ, Niederdeppe J, Freres D. Socioeconomic disparities in fatalistic beliefs about cancer prevention and the internet. J Commun 2012 Dec 20;62(6):972-990 [FREE Full text] [doi: 10.1111/j.1460-2466.2012.01683.x] [Medline: 25530627]
38. Chinn D. Critical health literacy: a review and critical analysis. Soc Sci Med 2011 Jul;73(1):60-67. [doi: 10.1016/j.socscimed.2011.04.004] [Medline: 21640456]
39. Sykes S, Wills J, Rowlands G, Popple K. Understanding critical health literacy: a concept analysis. BMC Public Health 2013 Feb 18;13:150 [FREE Full text] [doi: 10.1186/1471-2458-13-150] [Medline: 23419015]
40. Shim M, Jo HS. What quality factors matter in enhancing the perceived benefits of online health information sites? Application of the updated DeLone and McLean Information Systems Success Model. Int J Med Inform 2020 May;137:104093. [doi: 10.1016/j.ijmedinf.2020.104093] [Medline: 32078918]

XSL•FO
RenderX

41. Jiang S, Street RL. Pathway linking internet health information seeking to better health: a moderated mediation study. Health Commun 2017 Dec;32(8):1024-1031. [doi: 10.1080/10410236.2016.1196514] [Medline: 27464036]
42. Street RL. Mediated consumer-provider communication in cancer care: the empowering potential of new technologies. Patient Educ Couns 2003 May;50(1):99-104. [doi: 10.1016/s0738-3991(03)00089-2] [Medline: 12767594]
43. Stvilia B, Mon L, Yi YJ. A model for online consumer health information quality. J Am Soc Inf Sci Technol 2009 May 29;60(9):1781-1791. [doi: 10.1002/asi.21115]
44. Chang YS, Zhang Y, Gwizdka J. The effects of information source and eHealth literacy on consumer health information credibility evaluation behavior. Comput Human Behav 2021 Feb;115:106629. [doi: 10.1016/j.chb.2020.106629]
45. 16 years of online community operation quickly reaching your target audiences. IDiaoYan. URL: https://biz.idiaoyan.com/en-GB/sampleService.html [accessed 2024-04-29]
46. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet E-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
47. Hair JF, Hult GT, Ringle CM, Sarstedt M, Danks N, Ray S. Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook. Cham, Switzerland: Springer; 2021.
48. The 48th statistical report on China's internet development. China Internet Network Information Center. 2021 Aug. URL: https://www.cnnic.net.cn/NMediaFile/old_attach/P020210915523670981527.pdf [accessed 2021-11-01]
49. Pang H, Liu Y. Untangling the effect of cognitive trust and perceived value on health-related information seeking, sharing and psychological well-being: motivations sought perspective. Telemat Inform 2023 Apr;79:101964. [doi: 10.1016/j.tele.2023.101964]
50. Chen YC, Cheng C, Osborne RH, Kayser L, Liu CY, Chang LC. Validity testing and cultural adaptation of the eHealth literacy questionnaire (eHLQ) among people with chronic diseases in Taiwan: mixed methods study. J Med Internet Res 2022 Jan 19;24(1):e32855 [FREE Full text] [doi: 10.2196/32855] [Medline: 35044310]
51. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008 Dec;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050] [Medline: 18952344]
52. McDowell I. Measures of self-perceived well-being. J Psychosom Res 2010 Jul;69(1):69-79. [doi: 10.1016/j.jpsychores.2009.07.002] [Medline: 20630265]
53. Koufteros X, Marcoulides GA. Product development practices and performance: a structural equation modeling-based multi-group analysis. Int J Prod Econ 2006 Sep;103(1):286-307. [doi: 10.1016/j.ijpe.2005.08.004]
54. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Soft 2012;48(2):1-36 [FREE Full text] [doi: 10.18637/jss.v048.i02]
55. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39. [doi: 10.2307/3151312]
56. Cheung GW, Wang C. Current approaches for assessing convergent and discriminant validity with SEM: issues and solutions. Proceedings 2017 Aug;2017(1):12706. [doi: 10.5465/ambpp.2017.12706abstract]
57. Amo L. Education-based gaps in eHealth: a weighted logistic regression approach. J Med Internet Res 2016 Oct 12;18(10):e267 [FREE Full text] [doi: 10.2196/jmir.5188] [Medline: 27733329]
58. Schillinger D, Barton LR, Karter AJ, Wang F, Adler N. Does literacy mediate the relationship between education and health outcomes? A study of a low-income population with diabetes. Public Health Rep 2006 Aug 02;121(3):245-254 [FREE Full text] [doi: 10.1177/003335490612100305] [Medline: 16640146]
59. Islam MM, Touray M, Yang HC, Poly TN, Nguyen P, Li YJ, et al. E-health literacy and health information seeking behavior among university students in Bangladesh. Stud Health Technol Inform 2017;245:122-125. [Medline: 29295065]
60. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet users. J Med Internet Res 2016 Jul 18;18(7):e192 [FREE Full text] [doi: 10.2196/jmir.5413] [Medline: 27432783]
61. Shi D, Li J, Wang Y, Wang S, Liu K, Shi R, et al. Association between health literacy and hypertension management in a Chinese community: a retrospective cohort study. Intern Emerg Med 2017 Sep 16;12(6):765-776. [doi: 10.1007/s11739-017-1651-7] [Medline: 28303441]
62. Wang C, Wu X, Qi H. A comprehensive analysis of e-Health literacy research focuses and trends. Healthcare (Basel) 2021 Dec 30;10(1):66 [FREE Full text] [doi: 10.3390/healthcare10010066] [Medline: 35052230]
63. Abel T, McQueen D. Critical health literacy and the COVID-19 crisis. Health Promot Int 2020 Dec 01;35(6):1612-1613 [FREE Full text] [doi: 10.1093/heapro/daaa040] [Medline: 32239213]
64. Shieh C, Mays R, McDaniel A, Yu J. Health literacy and its association with the use of information sources and with barriers to information seeking in clinic-based pregnant women. Health Care Women Int 2009 Nov 05;30(11):971-988. [doi: 10.1080/07399330903052152] [Medline: 19809901]
65. Wong DK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019 Mar 27;21(3):e10831 [FREE Full text] [doi: 10.2196/10831] [Medline: 30916666]
66. Squiers L, Peinado S, Berkman N, Boudewyns V, McCormack L. The health literacy skills framework. J Health Commun 2012 Oct;17 Suppl 3(sup3):30-54. [doi: 10.1080/10810730.2012.713442] [Medline: 23030560]

XSL•FO
RenderX

67. Ghaddar SF, Valerio MA, Garcia CM, Hansen L. Adolescent health literacy: the importance of credible sources for online health information. J Sch Health 2012 Jan;82(1):28-36. [doi: 10.1111/j.1746-1561.2011.00664.x] [Medline: 22142172]
68. Jiang S, Beaudoin CE. Health literacy and the internet: an exploratory study on the 2013 HINTS survey. Comput Human Behav 2016;58:240-248 [FREE Full text] [doi: 10.1016/j.chb.2016.01.007]
69. Meppelink CS, Smit EG, Fransen ML, Diviani N. "I was right about vaccination": confirmation bias and health literacy in online health information seeking. J Health Commun 2019 Mar 21;24(2):129-140. [doi: 10.1080/10810730.2019.1583701] [Medline: 30895889]
70. Sun W, Luo A, Bian Z, Zhao B, Liu P, Wang K, et al. Assessing the quality of online health information about breast cancer from Chinese language websites: quality assessment survey. JMIR Cancer 2021 Nov 18;7(4):e25783 [FREE Full text] [doi: 10.2196/25783] [Medline: 34792471]
71. Feufel MA, Stahl SF. What do web-use skill differences imply for online health information searches? J Med Internet Res 2012 Jun;14(3):e87 [FREE Full text] [doi: 10.2196/jmir.2051] [Medline: 22695686]
72. Quinn S, Bond R, Nugent C. Quantifying health literacy and eHealth literacy using existing instruments and browser-based software for tracking online health information seeking behavior. Comput Human Behav 2017 Apr;69:256-267. [doi: 10.1016/j.chb.2016.12.032]
73. Lagoe C, Atkin D. Health anxiety in the digital age: an exploration of psychological determinants of online health information seeking. Comput Human Behav 2015 Nov;52:484-491. [doi: 10.1016/j.chb.2015.06.003]
74. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778 [FREE Full text] [doi: 10.2196/40778] [Medline: 36716080]
75. Mackert M, Champlin SE, Holton A, Muñoz II, Damásio MJ. eHealth and health literacy: a research methodology review. J Comput Mediat Comm 2014 Apr 12;19(3):516-528. [doi: 10.1111/jcc4.12044]

Abbreviations

AVE: average variance extracted
CFA: confirmatory factor analysis
CFI: comparative fit index
RMSEA: root mean square error of approximation
SEM: structural equation modeling
SRMR: standardized root mean square residual
OHIS: online health information seeking

Edited by T de Azevedo Cardoso; submitted 31.10.23; peer-reviewed by X Gong, B Schmitz, D Dinh; comments to author 26.02.24; revised version received 05.04.24; accepted 30.08.24; published 18.10.24.

Please cite as:
Liu D, Yang S, Cheng CY, Cai L, Su J
Online Health Information Seeking, eHealth Literacy, and Health Behaviors Among Chinese Internet Users: Cross-Sectional Survey Study
J Med Internet Res 2024;26:e54135
URL: https://www.jmir.org/2024/1/e54135
doi:10.2196/54135
PMID:

©Diyi Liu, Shuhang Yang, Calvin Yixiang Cheng, Lin Cai, Jing Su. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

How Do Scholars Conceptualize and Conduct Health and Digital Health Literacy Research? Survey of Federally Funded Scholars

Mayank Sakhuja[1,2], PhD; Brooks Yelton[3], MSW; Simone Kavarana[3], BA; Lauren Schaurer[3]; Jancham Rachel Rumthao[3], MSc; Samuel Noblet[4], MPH; Michelle A Arent[5], MPH; Mark M Macauda[3,6], PhD; Lorie Donelle[7], PhD; Daniela B Friedman[3], PhD

[1]UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[3]Department of Health Promotion, Education, and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States

[4]Envera Health, Richmond, VA, United States

[5]Department of Athletics, University of South Carolina, Columbia, SC, United States

[6]Center for Applied Research and Evaluation, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States

[7]Department of Biobehavioral Health and Nursing Science, College of Nursing, University of South Carolina, Columbia, SC, United States

Corresponding Author:
Mayank Sakhuja, PhD
UNC Lineberger Comprehensive Cancer Center
University of North Carolina at Chapel Hill
323A Rosenau Hall
Chapel Hill, NC, 27599-7295
United States
Phone: 1 8034468063
Email: msakhuja@unc.edu

Abstract

Background: The concept of health literacy (HL) is constantly evolving, and social determinants of health (SDoH) have been receiving considerable attention in public health scholarship. Since a 1-size-fits-all approach for HL fails to account for multiple contextual factors and as a result poses challenges in improving literacy levels, there is a need to develop a deeper understanding of the current state of HL and digital health literacy (DHL) research.

Objective: This study examined scholars' conceptualization and scope of work focused on HL and DHL.

Methods: Using a search string, investigators (N=2042) focusing on HL, DHL, or both were identified from the grantee websites of the National Institutes of Health RePORTER (RePORT Expenditures and Results) and the Canadian Institutes of Health Research. The investigators were emailed a survey via Qualtrics. Survey questions examined the focus of work; whether the investigators studied HL/DHL in combination with other SDoH; the frameworks, definitions, and approaches used; and research settings. We analyzed survey data using SPSS Statistics version 28 and descriptive analysis, including frequencies and percentages, was conducted. Chi-square tests were performed to explore the association between the focus of work, settings, and age groups included in the investigators' research.

Results: A total of 193 (9.5%) of 2042 investigators responded to the online survey. Most investigators (76/153, 49.7%) were from public health, 83/193 (43%) reported their research focused on HL alone, 46/193 (23.8%) mentioned DHL, and 64/193 (33.2%) mentioned both. The majority (133/153, 86.9%) studied HL/DHL in combination with other SDoH, 106/135 (78.5%) conducted HL/DHL work in a community setting, and 100/156 (64.1%) reported not using any specific definition to guide their work. Digital tools (89/135, 65.9%), plain-language materials (82/135, 60.7%), and visual guides (56/135, 41.5%) were the top 3 approaches used. Most worked with adults (131/139, 94.2%) and all races and ethnicities (47/121, 38.8%).

Conclusions: HL and DHL research largely considered SDoH. Multiple HL tools and approaches were used that support the examination and improvement of literacy and communication surrounding health care issues.

(J Med Internet Res 2024;26:e57040) doi:10.2196/57040

XSL•FO
RenderX

KEYWORDS

health literacy; digital health literacy; eHealth literacy; social determinants of health; SDoH; research scholarship; health care; public health research; digital health tools; community health

Introduction

Background

The concept of health literacy (HL) is constantly evolving. The Healthy People 2030 initiative defines HL as "the degree to which individuals have the ability to find, understand and use information and services to inform health-related decisions and actions for themselves and others." [1]. It refers to individuals' using their cognitive and social skills for gaining access to, understanding, and using information for promoting and maintaining good health for themselves and for others around them [2]. According to the United Nations Educational, Scientific and Cultural Institution, about 388 million adults in South and Southwest Asian countries continue to grapple with illiteracy, coupled with a lack of basic HL [3]. Even though many high-income nations, such as the United States, Canada, the United Kingdom, and Australia, have included HL in their national agenda [4], more than one-third of individuals in the United States still have limited HL, which consequently results in misunderstanding of medical information, increased health care–related costs, and poor health outcomes [5,6]. Individuals who face difficulty in understanding health-related materials may experience shame and discomfort in interacting or sharing their medical concerns and questions with health professionals [7].

Health Literacy as a Social Determinant of Health

HL functions as a critical component within the context of other social determinants of health (SDoH), such as education, socioeconomic status, access to quality health care, neighborhood and built environments, and social support networks [8,9]. Low literacy levels impact an individual's ability to access and use health care services by making it challenging for them to navigate health systems, understand health insurance information, and adhere to treatment plans resulting in delayed or inadequate care [10,11]. Individuals with a low socioeconomic status often have low HL levels, leading to difficulties in understanding providers' instructions, managing illness, and making informed health care choices for themselves [12]. Although studies indicate education may be associated with HL, the data cannot speak to a direct cause and effect tying higher education levels to higher HL [13]. Education is not a reliable proxy for determining literacy, as individuals with higher education may still have difficulty in understanding information pertaining to their health or the health of their families.

Navigating Health Care and Health Information

HL plays an important role in reducing health disparities, empowering individuals, and improving patient-provider communication and health outcomes. Individuals who possess high levels of HL are in a better position to understand medical instructions provided by their doctors, actively participate in decisions related to their health, and navigate the complex health care system. Changes within the health care environment, specifically the increased use of digital communication platforms, have transformed health care delivery and highlighted additional concerns around optimal patient-provider engagement. Digital health literacy (DHL) has become a significant component of HL with the increasing use of digital technologies in health care, including telehealth platforms, wearable devices, and mobile health apps and websites [14]. Although DHL shares core conceptual elements of HL, it differs as it requires an individual to possess additional skills to use technological devices, such as mobile phones and computers, and skills to use those devices to search and evaluate health-related information available online through multiple sources [14]. It is therefore essential for individuals to possess DHL skills to navigate digital technologies, access accurate and reliable online health information, and engage with telehealth platforms.

New Challenges and Directions

Despite the recognition of HL and DHL as essential factors of health care quality and outcomes, challenges exist in achieving optimal levels of literacy skills nationally and internationally [15]. Additionally, the increasing focus on DHL interventions poses new challenges in improving DHL, such as possessing skills to use new digital health technologies and identifying credible sources of information online [16]. Furthermore, a 1-size-fits-all approach for HL and DHL fails to account for cultural and linguistic diversity, age and generational differences, multiple health conditions and contexts, unequal access to digital resources, and diverse HL levels [17]. To develop a deeper understanding of the current state of HL and DHL research, we reported the results of a survey of HL and DHL researchers in this study. The aim was to comprehensively examine researchers' conceptualization of and approaches to HL and DHL work, identify strategies, and highlight areas for further investigation and intervention. This research addressed the following research questions:

- What concepts, definitions, and frameworks guide researchers' HL and DHL research?
- What specific settings and methods are used by HL and DHL researchers?
- Which SDoH are addressed in HL and DHL research?
- Who are the participants in HL and DHL research?

Methods

Survey Development and Measures

We used an iterative approach to develop the survey as per CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [18]. Two team members drafted the initial list of questions, and all team members, including experts in HL and DHL, weighed in on multiple drafts. To ensure the survey's reliability, we consulted content experts to design survey questions and response options that were appropriately worded and meaningful to study participants. Since we did not aim to

XSL•FO
RenderX

measure any specific constructs, we did not conduct the type of survey validation that we would have done otherwise. Survey questions examined researchers' specific focus (HL, DHL, or both); funding sources for HL and DHL research; and definitions, conceptual frameworks, and approaches used to study HL and DHL. We were also interested in understanding whether respondents studied HL/DHL in combination with other SDoH, as well as the research settings and geographic locations of their research. We asked about their main discipline/affiliation, number of years studying HL or DHL, educational qualification, and place of employment. Finally, we inquired about the data collection methods used in their research, whether they had ever received mentorship in HL/DHL work, or whether they provided mentorship to others. Respondents could select more than 1 response option for several of the survey questions.

Ethics Approval

This study was considered exempt by the University of South Carolina (USC) Office of Research Compliance under approval number Pro00124306.

Participant Recruitment and Survey Dissemination

We applied a search string ("literacy" or "health literacy" or "digital health literacy" or "e literacy" or "e health" or "m health" or "e health literacy") to project titles, project terms, and project abstracts on the grantee website of the National Institutes of Health (NIH) RePORTER (RePORT Expenditures and Results) and the Canadian Institutes of Health Research in November 2022. Project investigators (PIs) of studies matching those parameters were included in the sample. We obtained the email addresses of the PIs directly from the NIH RePORTER website and manually searched for email addresses of the PIs selected from the Canadian Institutes of Health Research grantee website. We excluded PIs for whom email addresses were unavailable on the NIH RePORTER grantee website, as well as PIs from the Canadian Institutes of Health Research whose email addresses could not be located through manual searches. The final sample included 2042 individual investigators. We also distributed the survey link via 2 organizational listservs consisting of members focused on health communication and health promotion research and practice.

We programmed the survey questions in Qualtrics [19]. Next, we prepared separate email lists for NIH and Canadian Institutes of Health Research PIs in Qualtrics, and invitations to participate in the survey were sent on January 23, 2023. The invitation email included a brief description and purpose of the study and a link to the Qualtrics survey and informed participants that survey completion would take approximately 15-20 minutes. Five email reminders were sent from Qualtrics at regular intervals to encourage participation and completion of the survey before its scheduled closure on April 23, 2023. All survey respondents were entered into a random drawing for 1 of 5 US $25 Amazon gift cards.

Data Analysis

Data were downloaded into Microsoft Excel. IBM SPSS Statistics version 28 [20] was used to conduct descriptive analysis including frequencies and percentages, and chi-square tests with a significance level of $P<.05$ were performed to explore the association between respondents' research focus (HL, DHL, both) and the settings and age groups with whom they conducted their research. We included responses from all 193 respondents and did not eliminate surveys that had missing responses. Analysis was conducted in a way that the results for each survey question contained only the valid responses for that question with the relevant denominator.

Results

Survey Respondent Characteristics

A total of 193 respondents answered the survey questions. Almost half (76/153, 49.7%) reported public health as their main discipline, 41/153 (26.8%) reported medicine, 38/153 (24.8%) reported health services/policy, 21/153 (13.7%) were from education, and 17/153 (11.1%) identified nursing as their academic discipline. Most had a PhD (98/132, 74.2%), 16/132 (12.1%) were MDs, and 13/132 (9.8%) had a master's degree as their highest educational qualification. Employment settings ranged from university (93/132, 70.5%), hospital system (12/132, 9.1%), or research institution (10/132, 7.6%). Most (106/132, 80.3%) of the respondents had never received any formal mentorship in HL or DHL. Those who had received formal mentorship reported receiving training from an academic mentor, professional organization seminars, or professional training. Just over half (73/142, 51.4%) reported that they had been studying HL or DHL for 1-5 years, and 37/142 (26.1%) indicated more than 10 years of experience studying HL or DHL.

Focus of Research

Regarding the scope of work, 83 (43%) of 193 respondents reported that most of their work focused on HL, while 46 (23.8%) reported DHL, and the remaining 64 (33.2%) indicated a dual focus including both HL and DHL. When asked about the specific topics of focus, 87/139 (62.6%) respondents reported that they had a general focus on HL/health communication, 61/139 (43.9%) focused on topics around social justice/health equity, 48/139 (34.5%) focused on face-to-face patient-provider communication, 41/139 (29.5%) focused on misinformation/disinformation, 37/139 (26.6%) focused on virtual patient-provider communication, 36/139 (25.9%) focused on peer-to-peer health conversations, and 27/139 (19.4%) focused on policy development. Respondents had almost equal focus on specific health behaviors, including healthful eating (36/139, 25.9%), physical activity (36/139, 25.9%), and cancer screening (33/139, 23.7%).

Concepts and Frameworks Used in HL or DHL Research

About two-thirds (100/156, 64.1%) of the respondents reported not using any specific definition of HL or DHL to guide their work. Of those who did follow a specific definition (n=56, 35.9%), roughly 19 (33.9%) indicated using the Institute of Medicine's [21] definition. Other definitions cited were by Nutbeam [22], the World Health Organization [23], U.S. Department of Health and Human Services [1,24], among others, as described in Table 1. Research by about one-third (46/143,

XSL•FO
RenderX

32.2%) of the respondents was guided by the Healthy People 2030 framework [1]; 35/143 (24.5%) used the Nutbeam framework that describes HL as functional, interactive, and critical [22]; 27/143 (18.9%) used the HL skills framework by Squiers et al [25]; and 18/143 (12.6%) used the HL systems capacity framework by Sorenson et al [26]. Regarding DHL research, 17/143 (11.9%) respondents used the electronic health literacy (eHL) framework by Bautista [27], 12/143 (8.4%) used the eHL definition and the Lily model by Norman and Skinner [14], 12/143 (8.4%) used the eHL framework by Norgaard et al [28], and 10/143 (7%) used the consumer eHL taxonomy by Chan et al [29]. Some respondents (6/143, 4.2%) did not follow any framework, whereas others mentioned they used dissemination and implementation science–based, collaborator-informed socioecological models and environmental HL frameworks. Frameworks used in HL and DHL research are presented in Table 2.

XSL•FO
RenderX

Table 1. Definitions used in HL[a] and DHL[b] research provided by respondents.

Definition guiding HL and DHL research	Respondents (n=56), n (%)	Source
The individuals' capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions.	19 (33.9)	Institute of Medicine [21]
HL represents the personal knowledge and competencies that accumulate through daily activities and social interactions across generations. Personal knowledge and competencies are mediated by organizational structures and availability of resources that enable people to access, understand, appraise, and use information and services in ways that promote and maintain good health and well-being for themselves and those around them.	3 (5.4)	World Health Organization [23]
The degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions.	3 (5.4)	U.S. Department of Health and Human Services [24]
Personal HL is the degree to which individuals have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others. Organizational HL is the degree to which organizations equitably enable individuals to find, understand, and use information and services to inform health-related decisions and actions for themselves and others.	3 (5.4)	U.S. Department of Health and Human Services [1]
The degree to which individuals and groups can obtain process, understand, evaluate, and act upon information needed to make public health decisions that benefit the community.	3 (5.4)	Freedman et al [30]
The personal, cognitive, and social skills that determine the ability of individuals to gain access to, understand, and use information to promote and maintain good health.	2 (3.6)	Nutbeam [22]
The cognitive and social skills that determine the motivation and ability of individuals to gain access to, understand, and use information in ways that promote and maintain good health.	2 (3.6)	Nutbeam [31]
The knowledge, skills, and abilities that pertain to interactions with the health care system.	2 (3.6)	Ishikawa and Yano [32]
The ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem.	2 (3.6)	Norman and Skinner [14]
DHL involves the context and ability to use digital technologies to access, understand, and use health information in a timely manner to improve health outcomes.	1 (1.8)	Norman and Skinner [14]
DHL is defined as the ability to appraise health information from electronic sources and apply the knowledge gained to address or solve a health-related problem and as such has emerged as an important component of greater HL. Although DHL shares core aspects of HL, DHL is distinguished by additional skills: computer literacy, the ability to use computers and related technology efficiently to accomplish tasks, media literacy to use search engines, and information literacy to evaluate a wide variety of sources.	1 (1.8)	Smith and Magnani [16]
The ability to understand and interpret the meaning of health information in written, spoken, or digital form and how this motivates people to embrace or disregard actions relating to health.	1 (1.8)	Adams et al [33]
The capacity to obtain, interpret, and understand basic health information and services and the competence to use such information to enhance health.	1 (1.8)	Pavlekovic [34]
At its most basic, environmental HL has been described as an ability to make connections between environmental exposures and human health. Representations of environmental HL tend to start with individual understanding of specific risks and then lead to broader understanding, including strategies that empower people to reduce or eliminate environmental exposures that can harm health.	1 (1.8)	Finn and O'Fallon [35]
HL is defined as the ability to access, understand, and use information to make health decisions. It is well known that low HL is associated with poorer health outcomes.	1 (1.8)	Berkman et al [36]
The knowledge and skills required to understand and use information relating to health issues, such as drugs and alcohol, disease prevention and treatment, safety, accident prevention, first aid, emergencies, and staying healthy.	1 (1.8)	Australian Bureau of Statistics [37]
Knowledge and beliefs about mental disorders that aid their recognition, management, or prevention. It is a multidimensional concept that includes (1) an ability to recognize specific disorders or types of psychological distress, (2) knowledge and beliefs about risk factors and prevention, (3) knowledge and attitudes that facilitate help seeking, and (4) knowledge and beliefs regarding formal and informal intervention approaches. DHL focuses more on knowledge and skills related to technology-based approaches to digital health.	1 (1.8)	Jorm [38]
The ability to access, understand, evaluate, and communicate information as a way to promote, maintain, and improve health in a variety of settings across the life course.	1 (1.8)	Rootman and Gordon-El-Bihbety [39]

XSL•FO
RenderX

Definition guiding HL and DHL research	Respondents (n=56), n (%)	Source
The ability to make sound health decision(s) in the context of everyday life: at home, in the community, at the workplace, in the health care system, in the marketplace and in the political arena. It is a critical empowerment strategy to increase people's control over their health, their ability to seek out information, and their ability to take responsibility.	1 (1.8)	Kickbusch et al [40]
The wide range of skills and competencies that people develop to seek out, comprehend, evaluate, and use health information and concepts to make informed choices, reduce health risks, and increase the quality of life.	1 (1.8)	Zarcadoolas et al [41]
An emerging field in the intersection of medical informatics, public health, and business, referring to health services and information delivered or enhanced through the internet and related technologies. It is also a state of mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology.	1 (1.8)	Eysenbach [42]
We define mental HL as the extent to which individuals are knowledgeable about mental health problems and their treatment.	1 (1.8)	N/A[c]
Individuals' ability to apply for/maintain health coverage and navigate the health care system to have their health needs met.	1 (1.8)	N/A
Functional HL is a person's ability to comprehend patient education materials to make an informed decision about their health.	1 (1.8)	N/A
We used an internal definition (unpublished) focused on literacy related to health, research, medicine, and insurance to guide our work.	1 (1.8)	N/A

[a]HL: health literacy.

[b]DHL: digital health literacy.

[c]N/A: not applicable.

Table 2. Frameworks used in HL[a] and DHL[b] research.

Conceptual framework guiding HL and DHL research	Respondents (n=143), n (%)[c]
Healthy People 2030 (U.S. Department of Health and Human Services [1])	46 (32.2)
HL as functional, interactive, and critical (Nutbeam [22])	35 (24.5)
HL skills framework (Squiers et al [25])	27 (18.9)
HL systems capacity (Sorenson et al [26])	18 (12.6)
eHL (Bautista [27])	17 (11.9)
HL as an asset model (Nutbeam [43])	15 (10.5)
eHL and Lily model (Norman and Skinner [14])	12 (8.4)
eHL framework (Norgaard et al [28])	12 (8.4)
Consumer eHL (Chan et al [29])	10 (7.0)
eHL framework (Kayser et al [44])	8 (5.6)
Integrated model for cancer screening (Best et al [45])	7 (4.9)
Other frameworks	34 (23.8)
None	6 (4.2)

[a]HL: health literacy.

[b]DHL: digital health literacy.

[c]Participants could select more than 1 response option, so the total percentage may be more than 100%.

Methods and Settings Used in HL or DHL Research

Table 3 presents the settings in which respondents conducted their research. Most conducted their HL/DHL work in a community setting (106/135, 78.5%); 49/135 (36.3%) reported a primary care setting; 44/135 (32.6%) reported a specialty care setting; 29/135 (21.5%) reported the virtual health care space, such as patient portals/provider portals; and 20/135 (14.8%) reported government-/policy-level settings. Chi-square tests indicated that of those (n=20, 14.8%) who conducted research in a government-/policy-level setting, 14 (70%) were focused on HL research, while 13 (44.8%) of the respondents who conducted research in the virtual health care space were focused on DHL. Only significant (P<.05) chi-square test results are

XSL•FO
RenderX

presented in Table 4. When asked about the specific approaches that they used in their work, 89 (65.9%) respondents mentioned using digital tools, such as apps, websites, patient portals, virtual visits, and wearable devices; 82/135 (60.7%) reported using plain-language materials; 56/135 (41.5%) used visual guides, such as graphical decision tools; 39/135 (28.9%) used teach-back strategies; and 36/135 (26.7%) made use of a partnership with patient navigators. Most respondents used mixed methods (109/134, 81.3%) for their work. For data collection, 112/134 (83.6%) respondents reported that they collected data using interviews, 118/134 (88.1%) conducted surveys, 89/134 (66.4%) conducted focus groups, 50/134 (37.3%) conducted secondary data analysis, and 44/134 (32.8%) completed scoping and systematic reviews and meta-analysis. Other types of data collection methods included content analysis (33/134, 24.6%), comprehension testing (23/134, 17.2%), readability testing (25/134, 18.7%), and observational data (42/134, 31.3%).

Table 3. Settings in which HL[a] and DHL[b] research is conducted.

Research settings	Respondents (n=135), n (%)[c]
Clinical setting	108 (80.0)
Primary care	49 (36.3)
Specialty care	44 (32.6)
Other	15 (11.1)
Community setting	106 (78.5)
Government-/policy-level setting	20 (14.8)
Virtual health care space[d]	29 (21.5)
Other settings[e]	16 (11.9)
Academia	7 (5.2)
County hospital	1 (0.7)
Higher education level	1 (0.7)
Hospital	1 (0.7)
Interactive internet to classrooms	1 (0.7)
Medical center	1 (0.7)
Mobile health	1 (0.7)
School	1 (0.7)
Smoking quitline	1 (0.7)
Tribal health care organization	1 (0.7)

[a]HL: health literacy.
[b]DHL: digital health literacy.
[c]Participants could select more than 1 response option, so the total percentage may be more than 100%.
[d]Patient or provider portals.
[e]Respondent write-in.

Table 4. Chi-square tests between focus of research and research setting.

Focus	Government-/policy-level setting (n=20; *P*=.034), n (%)	Virtual health care space (n=29; *P*=.006), n (%)	Specialty care (n=44; *P*=.014), n (%)
DHL[a]	2 (10.0)	13 (44.8)	16 (36.4)
HL[b]	14 (70.0)	6 (20.7)	11 (25.0)
Both	4 (20.0)	10 (34.5)	17 (38.6)

[a]DHL: digital health literacy.
[b]HL: health literacy.

Consideration of Social Determinants of Health

Most respondents (133/153, 86.9%) reported that they studied HL/DHL in combination with other SDoH. Health care access and quality were studied with HL/DHL most often (93/133, 69.9%), followed by social and community contexts (80/133, 60.2%), education access and quality (57/133, 42.9%), neighborhood and built environment (44/133, 33.1%), and

XSL•FO
RenderX

economic stability (40/133, 30.1%). Other SDoH, as reported by respondents, included equity, gender and caregiving support, insurance and employment status, language equity, mental health stigma, poverty, culture, environmental exposures, and other measures of socioeconomic status.

Profile of Participants With Whom Respondents Conducted Research

Respondents conducted research mostly with adult populations (131/139, 94.2%), 42/139 (30.2%) conducted research with adolescents (12-17 years old), and 23/139 (16.5%) conducted research with youth (<12 years old). There were no significant chi-square test results between the focus of research and the age categories of respondents' research participants. Some respondents reported conducting research with all/diverse races (47/121, 38.8%), while others indicated a more limited approach. One-third (40/121, 33.1%) respondents conducted research with only Black, African American, or both populations; 31/121 (25.6%) reported research with White/Caucasian/European Americans; 32/121 (26.4%) reported research with Hispanic or LatinX populations; and 11/121 (9.1%) reported research with Asian, South Asian, and East Asian populations. Most respondents (83/137, 60.6%) reported that they conducted research with the general public, 43/137 (31.4%) reported research with general patients, and 58/137 (42.3%) reported research with patients of a specific health/medical specialty. About one-third (44/137, 32.1%) of the respondents conducted research with general clinical providers; 43/137 (31.4%) included physicians, nurse practitioners, and PAs as their research participants; and 27/137 (19.7%) worked with professionals who were from nursing and related clinical/health sciences backgrounds. Respondents also conducted research with specific communities who had a specific culture (33/137, 24.1%); the lesbian, gay, bisexual, transgender, queer, questioning, intersex, asexual, and other (LGBTQIA+) community (15/137, 10.9%); individuals without a shelter or stable housing (15/137, 10.9%); and individuals with disabilities (10/137, 7.3%).

Discussion

Principal Findings

This study was a comprehensive examination of researchers' conceptualization of and strategies for conducting HL and DHL research. Understanding principles and methods that guide the work in this area is critical for advancing the field, pursuing needed interventions in both clinical and community settings, and studying literacy in the context of additional social determinants that influence health behaviors and outcomes.

Conceptualizing Health and Digital Health Literacy

Most respondents reported not using any specific definition of HL or DHL to guide their work. Others suggested a variety of definitions and seminal papers from the field. Without the use of common definitions and measures, understanding a discipline will be limited to knowledge obtained from individual cases; however, it may not advance the field holistically so that we can create culturally and linguistically appropriate communication strategies to promote evidence-based programs, interventions, and policies [46,47].

Most respondents reported that their research was generally focused on HL or health communication and on social justice or health equity, demonstrating the importance of studying HL and DHL within the context of other social factors. Over 85% of the respondents investigated HL/DHL in relation to other SDoH. Researchers' attention to HL/DHL and SDoH is consistent with van Kessel et al [48], who claimed that the development of HL, and increasingly DHL, is a pre-requisite for enhanced health and well-being. Others have also highlighted the growing importance of DHL and contend that DHL and digital inclusion constitute the "super–SDoH," especially given our widespread reliance on digital tools for health [49].

Life contexts (eg, age, gender, income, employment, education, HL/DHL skills, social support, residential environments, health care access) create conditions for (in)equitable health and well-being [48,50]. Researchers in this study focused predominantly on the intersection of HL/DHL and health service system issues. One-third to half of the researchers addressed HL/DHL skills in relation to education, the neighborhood context, and economic stability. With agreement that HL and DHL constitute fundamental skills in support of equitable health and well-being, our findings suggest important opportunities to enhance research activity for greater insight into the impact of HL/DHL skills at the intersection of culturally diverse groups of individuals, those living in poverty, individuals with poor mental health, individuals without stable housing, individuals with disabilities, and across gender and sexual identities [51].

Intended Audiences

Almost 80% of the respondents conducted research in a community setting. Community-engaged HL and DHL research is critical to understanding diverse stakeholders' needs, preferences, and pre-existing beliefs [52]. Effective education programs, interventions, and health communication materials are tailored and culturally relevant to different audiences, centering community-held values, imagery, and voices for both trust in the message/messenger and adoption of associated content and behaviors [53,54]. Codeveloping research, programs, and communication products with community members ensures stakeholder priorities are fully accounted for and increases member buy-in [55]. Although clinical interventions are necessary to improve health and digital literacy at the organizational level, design and delivery must balance pragmatic approaches to implementation with the utility and receptibility of a diverse patient population [56]. For improved prevention, it is also imperative that researchers meet stakeholders where they are, particularly among medically underserved populations that may only engage with the health care system in emergency situations.

Despite the racial and ethnic variety of respondents' research audiences, limited research was reportedly conducted with Asian, South Asian, or East Asian populations. This may be due, in part, to perceptions that health outcomes tend to be favorable among individuals of Asian descent in the United States and Canada [57], leading researchers to focus their efforts on communities perceived as being at higher risk for negative

XSL•FO
RenderX

health disparities. However, researchers can inadvertently disadvantage population subgroups by aggregating groups of people into broad categories, masking heterogeneity and limiting research capacity, health education, and funding for groups that may have different risk profiles or poorer health outcomes than other groups within the same broad category [58,59]. In addition, researchers using a health equity lens must make deliberate strides to improve research participation among many population groups by centering aims around each population's interests and goals, repairing and building trust, providing adequate compensation and incentive, and tailoring communication to meet the intended population's language, source, and channel preferences [60]. Regardless, the tenets of value-based health promotion support the need to codevelop health and DHL with individuals and populations across the health-illness continuum, as HL can improve both knowledge of health risks and trust in health care. A smaller portion of respondents conducted research with communities who shared a specific culture, members of the LGBTQIA+ community, individuals without a shelter or stable housing, and individuals with disabilities. These populations are at higher risk for health care bias and other SDoH [61]; thus, it is imperative that they be engaged and appropriately consulted in the field of HL.

It was not surprising that most of our survey respondents conducted research with adult populations (94.2%), whereas a smaller proportion worked with adolescents (30.2%) and the youth (16.5%). Additionally, none of the respondents mentioned using conceptual frameworks adapted for use with younger age groups. This is likely due to the fact that parental involvement and consent are required in adolescent health care and participation in research or other programs. However, it is important to understand and address health and DHL from a life course perspective, as early life experiences and cumulative disadvantages across the lifespan can lead to accumulated stress and uptake of unhealthy behavior, leading to poor health outcomes in adulthood [62]. The limited focus on HL among lower age groups has remained a significant gap requiring more research on HL instruments and interventions among younger age groups [63]. The limited literature focused on HL among adolescents and the youth impedes our understanding of the needs and preferences of HL and DHL in the context of younger age groups.

Implications for Health Care Professionals

Approximately three-quarters of the respondents reported public health and medicine as their academic and professional disciplines. Although the nursing profession is proportionately greater than the discipline of public health or medicine, only 11% of the respondents conducting HL research had a nursing background. Worldwide, nurses serve as critical partners in the health care encounter as they provide education and facilitate communication exchange, which can aid patient HL [64]. However, research demonstrates that nurses in the United States cannot identify patient communication behaviors that point toward low HL levels [65]. Engaging nursing professionals in HL and DHL research is important because nurses interact directly with patients, which makes them well positioned to develop and implement strategies to fully understand the needs and challenges of a diverse patient population. More involvement of nursing professionals would help in developing informed and tailored communication strategies for patient populations and help determine the most effective ways for communicating health information with patients of varying levels of literacy. HL research by nursing professionals would help in reducing disparities as understanding the challenges by patients from diverse backgrounds would allow nursing professionals to deliver more equitable health care.

Limitations

Our survey was distributed primarily via lists of researchers who had NIH funding in the United States or funding through the Canadian Institutes of Health Research; thus, our sample primarily reflects researchers who have been awarded through these funding streams. There are likely others in the field who have carried out their work using other funding streams. Since research agendas are, to some extent, directed by the priorities of funding agencies, this survey may have missed individuals conducting different kinds of HL work. Additionally, since we primarily searched the grantee websites of the NIH and the Canadian Institutes of Health Research, our sample of researchers was primarily based in high-income countries. As we could not reach out to researchers in low- and middle-income countries, our study findings cannot be generalized to other countries. Despite keeping the survey open for a period of 3 months and sending 5 reminder emails for completing the survey, the low response rate of our survey further restricts the generalizability of our study findings. As we did not conduct cognitive testing of our survey questions, there is a possibility that our questions did not address the most meaningful information. Finally, the majority of our questions were closed ended, which may fail to capture the complexity of the working definitions used by respondents, as well as the variety of work they engage in. Even though this study has limitations, we strongly believe that the study findings still provide important insights into the scope of and how researchers conceptualize HL and DHL research, which can inform future research directions. Understanding what funded researchers are doing in this area of research and with the growth in the need for research on SDoH, this study can help inform more comprehensive examination of this field of research.

Conclusion

The results of our survey indicate there is significant variation in how HL is researched and understood. Although a variety of approaches may add richness to the field, it also makes comparisons across studies and generalizability of findings more difficult. Further, few respondents reported receiving any formal mentorship in their study of HL or DHL, which may perpetuate inconsistencies in how HL is defined, since knowledge transfer from one generation of researchers to another may be limited. It is encouraging that the majority of survey respondents investigated HL along with other SDoH, given the inequality that is often pervasive within health systems and the importance of going beyond studies of readability to more contextual factors. However, the lack of work in the clinical setting is problematic, since it has been recognized that HL is a structural issue that the health sector must address. Given the somewhat limited consistency in how the concept is currently defined and

XSL•FO
RenderX

the evolution of the field from the study of readability of documents to the exploration of literacy within the context of social factors and health equity, opportunities exist to leverage conferences, scientific meetings, and other professional spaces to learn from each other and create mentorship opportunities for junior scholars and those who may be moving into this area. This would facilitate continued discussion among practitioners and researchers to create common definitions and goals around HL, DHL, and their relationship to SDoH.

Conflicts of Interest

None declared.

References

1. Health literacy in healthy people 2030. U.S. Department of Health and Human Services. URL: https://health.gov/healthypeople/priority-areas/health-literacy-healthy-people-2030 [accessed 2023-09-04]
2. Health promotion. World Health Organization. URL: https://www.who.int/teams/health-promotion/enhanced-wellbeing/ninth-global-conference/health-literacy [accessed 2023-08-09]
3. Health literacy and sustainable development. United Nations. URL: https://www.un.org/en/chronicle/article/health-literacy-and-sustainable-development [accessed 2023-08-09]
4. Qi S, Hua F, Xu S, Zhou Z, Liu F. Trends of global health literacy research (1995-2020): analysis of mapping knowledge domains based on citation data mining. PLoS One 2021 Aug 9;16(8):e0254988 [FREE Full text] [doi: 10.1371/journal.pone.0254988] [Medline: 34370749]
5. Friedman DB, Arent MA, Yelton B, Sakhuja M, Haynes VE, Noblet S, et al. Development of a clinical-academic-community collaboration to improve health literacy. J Prim Care Community Health 2020 Sep 10;11:215013272095744. [doi: 10.1177/2150132720957440]
6. Feinberg I, Greenberg D, Frijters J. Understanding health information seeking behaviors of adults with low literacy, numeracy, and problem solving skills: results from the 2012 US PIAAC study. Squarespace. 2015. URL: https://static1.squarespace.com/static/51bb74b8e4b0139570ddf020/t/54da7889e4b004762df45b52/1423603849970/Feinberg_Greenberg_Frijters_PIAAC.pdf [accessed 2024-10-07]
7. Mackert M, Donovan EE, Mabry A, Guadagno M, Stout PA. Stigma and health literacy: an agenda for advancing research and practice. Am J Health Behav 2014 Sep 01;38(5):690-698. [doi: 10.5993/ajhb.38.5.6]
8. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007 Jul 01;31(1):19-26. [doi: 10.5993/ajhb.31.s1.4]
9. Rikard RV, Thompson MS, McKinney J, Beauchamp A. Examining health literacy disparities in the United States: a third look at the National Assessment of Adult Literacy (NAAL). BMC Public Health 2016 Sep 13;16(1):975 [FREE Full text] [doi: 10.1186/s12889-016-3621-9] [Medline: 27624540]
10. MacLeod S, Musich S, Gulyas S, Cheng Y, Tkatch R, Cempellin D, et al. The impact of inadequate health literacy on patient satisfaction, healthcare utilization, and expenditures among older adults. Geriatr Nurs 2017 Jul;38(4):334-341. [doi: 10.1016/j.gerinurse.2016.12.003]
11. Quinlan P, Price KO, Magid SK, Lyman S, Mandl LA, Stone PW. The relationship among health literacy, health knowledge, and adherence to treatment in patients with rheumatoid arthritis. HSS J: Musculoskel J Hosp Special Surg 2013 Feb 01;9(1):42-49. [doi: 10.1007/s11420-012-9308-6]
12. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. Health Promot Int 2019 Oct 01;34(5):e1-e17. [doi: 10.1093/heapro/day062] [Medline: 30107564]
13. Beauchamp A, Buchbinder R, Dodson S, Batterham RW, Elsworth GR, McPhee C, et al. Distribution of health literacy strengths and weaknesses across socio-demographic groups: a cross-sectional survey using the Health Literacy Questionnaire (HLQ). BMC Public Health 2015 Jul 21;15(1):678 [FREE Full text] [doi: 10.1186/s12889-015-2056-z] [Medline: 26194350]
14. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9,e506 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
15. Allen-Meares P, Lowry B, Estrella ML, Mansuri S. Health literacy barriers in the health care system: barriers and opportunities for the profession. Health Soc Work 2020 Jan 07;45(1):62-64 [FREE Full text] [doi: 10.1093/hsw/hlz034] [Medline: 31993624]
16. Smith B, Magnani JW. New technologies, new disparities: the intersection of electronic health and digital health literacy. Int J Cardiol 2019 Oct 01;292:280-282 [FREE Full text] [doi: 10.1016/j.ijcard.2019.05.066] [Medline: 31171391]
17. Dias S. Health literacy promotion in complex systems: a paramount paradigm when “one-size-fits-all” is not enough. Port J Public Health 2022 Dec 13;40(3):131-133. [doi: 10.1159/000527722]
18. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
19. Qualtrics AI + the XM platform. Qualtrics. URL: https://www.qualtrics.com/ [accessed 2024-10-07]

XSL•FO
RenderX

20. IBM SPSS Statistics. IBM Corp. URL: https://www.ibm.com/products/spss-statistics [accessed 2021-10-08]
21. Nielsen-Bohlman L, Panzer AM, Kindig DA. Health Literacy: A Prescription to End Confusion. Washington, DC: National Academies Press; 2004.
22. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
23. Health promotion glossary of terms 2021. World Health Organization. 2021 Dec 6. URL: https://www.who.int/publications/i/item/9789240038349 [accessed 2024-10-07]
24. Healthy people 2020. U.S. Department of Health and Human Services. URL: https://www.cdc.gov/nchs/healthy_people/hp2020.htm [accessed 2024-10-15]
25. Squiers L, Peinado S, Berkman N, Boudewyns V, McCormack L. The health literacy skills framework. J Health Commun 2012 Oct;17(Suppl 3):30-54. [doi: 10.1080/10810730.2012.713442] [Medline: 23030560]
26. Sørensen K, Levin-Zamir D, Duong TV, Okan O, Brasil VV, Nutbeam D. Building health literacy system capacity: a framework for health literate systems. Health Promot Int 2021 Dec 13;36(Supplement_1):i13-i23 [FREE Full text] [doi: 10.1093/heapro/daab153] [Medline: 34897445]
27. Bautista JR. From solving a health problem to achieving quality of life: redefining eHealth literacy. J Lit Technol 2015;16(2):33-54.
28. Norgaard O. The e-health literacy framework: a conceptual framework for characterizing e-health users and their interaction with e-health systems. Knowl Manag E-Learn 2015;7(4):522. [doi: 10.34105/j.kmel.2015.07.035]
29. Chan CV, Matthews LA, Kaufman DR. A taxonomy characterizing complexity of consumer eHealth literacy. AMIA Annu Symp Proc 2009 Nov 14;2009:86-90. [Medline: 20351828]
30. Freedman DA, Bess KD, Tucker HA, Boyd DL, Tuchman AM, Wallston KA. Public health literacy defined. Am J Prev Med 2009 May;36(5):446-451. [doi: 10.1016/j.amepre.2009.02.001] [Medline: 19362698]
31. Nutbeam D. Health promotion glossary. Health Promot Int 1998;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
32. Ishikawa H, Yano E. Patient health literacy and participation in the health-care process. Health Expect 2008 Jun;11(2):113-122. [doi: 10.1111/j.1369-7625.2008.00497.x] [Medline: 18494956]
33. Adams RJ, Stocks NP, Wilson DH, Hill CL, Gravier S, Kickbusch I, et al. Health literacy--a new concept for general practice? Aust Fam Physician 2009 Mar;38(3):144-147. [Medline: 19283256]
34. Pavlekovic G. Health literacy. Programmes for training on research in public health for South Eastern Europe. In: Health Promotion and Disease Prevention: A Handbook for Teachers, Researchers, Health Professionals and Decision Makers. Lage, Germany: Hans Jacobs Publishing Company; 2007.
35. Finn S, O'Fallon L. The emergence of environmental health literacy—from its roots to its future potential. Environ Health Perspect 2017 Apr;125(4):495-501 [FREE Full text] [doi: 10.1289/ehp.1409337] [Medline: 26126293]
36. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
37. Australian Bureau of Statistics. Adult Literacy and Life Skills Survey. Summary Results. Canberra, Australia: Australian Bureau of Statistics; 2008:88.
38. Jorm AF. Mental health literacy. Public knowledge and beliefs about mental disorders. Br J Psychiatry 2000 Nov;177:396-401. [doi: 10.1192/bjp.177.5.396] [Medline: 11059991]
39. Rootman I, Gordon-El-Bihbety D. A Vision for a Health Literate Canada: Report of the Expert Panel on Health Literacy. Ottawa, Canada: Canadian Public Health Association; 2008.
40. Kickbusch I, Wait S, Maag D, Banks I. Navigating Health: The Role of Health Literacy. UK: Alliance for Health and the Future, International Longevity Centre; 2006.
41. Zarcadoolas C, Pleasant A, Greer DS. Understanding health literacy: an expanded model. Health Promot Int 2005 Jun;20(2):195-203. [doi: 10.1093/heapro/dah609] [Medline: 15788526]
42. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
43. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008 Dec;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050]
44. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
45. Best AL, Vamos C, Choi SK, Thompson EL, Daley E, Friedman DB. Increasing routine cancer screening among underserved populations through effective communication strategies: application of a health literacy framework. J Cancer Educ 2017 Jun 8;32(2):213-217 [FREE Full text] [doi: 10.1007/s13187-017-1194-7] [Medline: 28275965]
46. Friedman DB, Becofsky K, Anderson LA, Bryant LL, Hunter RH, Ivey SL, et al. Public perceptions about risk and protective factors for cognitive health and impairment: a review of the literature. Int Psychogeriatr 2015 Jan 16;27(8):1263-1275. [doi: 10.1017/s1041610214002877]

XSL•FO
RenderX

47. Koskan AM, Friedman DB, Brandt HM, Walsemann KM, Messias DKH. Preparing promotoras to deliver health programs for Hispanic communities: training processes and curricula. Health Promot Pract 2013 May 14;14(3):390-399 [FREE Full text] [doi: 10.1177/1524839912457176] [Medline: 22982761]
48. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022 Mar;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
49. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52 [FREE Full text] [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
50. Social determinants of health. World Health Organization. URL: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1 [accessed 2023-10-01]
51. Schillinger D. Social determinants, health literacy, and disparities: intersections and controversies. Health Lit Res Pract 2021 Jul;5(3):e234-e243 [FREE Full text] [doi: 10.3928/24748307-20210712-01] [Medline: 34379549]
52. Ross L, Johnson J, Smallwood SW, Luque JS, Tedders SH, Airhihenbuwa CO, MAN UP Prostate Cancer Advocates, et al. Using CBPR to extend prostate cancer education, counseling, and screening opportunities to urban-dwelling African-Americans. J Cancer Educ 2016 Dec 8;31(4):702-708 [FREE Full text] [doi: 10.1007/s13187-015-0849-5] [Medline: 25948412]
53. Abbas MZ. Public understanding and voluntary responsibility to mitigate COVID-19 pandemic: role of local community organizations in information sharing and health literacy. Asia Pac J Public Health 2021 May 17;33(4):451-452. [doi: 10.1177/10105395211002656] [Medline: 33729019]
54. Haidar A, Khoei A, Alex SE, Blick C, Lopez E, Wendt S, et al. Community-academic partnerships to promote health literacy and address social needs among low-income families during COVID-19. J Nutr Educ Behav 2021 Jan;53(1):75-78 [FREE Full text] [doi: 10.1016/j.jneb.2020.10.003] [Medline: 33187874]
55. Berge JM, Mendenhall TJ, Doherty WJ. Using community-based participatory research (CBPR) to target health disparities in families. Fam Relat 2009 Oct 01;58(4):475-488 [FREE Full text] [doi: 10.1111/j.1741-3729.2009.00567.x] [Medline: 20625444]
56. Nutbeam D, McGill B, Premkumar P. Improving health literacy in community populations: a review of progress. Health Promot Int 2018 Oct 01;33(5):901-911. [doi: 10.1093/heapro/dax015] [Medline: 28369557]
57. Singh GK, Daus GP, Allender M, Ramey CT, Martin EK, Perry C, et al. Social determinants of health in the United States: addressing major health inequality trends for the nation, 1935-2016. Int J MCH AIDS 2017 Jan 15;6(2):139-164 [FREE Full text] [doi: 10.21106/ijma.236] [Medline: 29367890]
58. Yom S, Lor M. Advancing health disparities research: the need to include Asian American subgroup populations. J Racial Ethn Health Disparities 2022 Dec 17;9(6):2248-2282 [FREE Full text] [doi: 10.1007/s40615-021-01164-8] [Medline: 34791615]
59. Gordon NP, Lin TY, Rau J, Lo JC. Aggregation of Asian-American subgroups masks meaningful differences in health and health risks among Asian ethnicities: an electronic health record based cohort study. BMC Public Health 2019 Nov 25;19(1):1551. [doi: 10.1186/s12889-019-7683-3]
60. Liu Y, Elliott A, Strelnick H, Aguilar-Gaxiola S, Cottler LB. Asian Americans are less willing than other racial groups to participate in health research. J Clin Transl Sci 2019 Jun 28;3(2-3):90-96 [FREE Full text] [doi: 10.1017/cts.2019.372] [Medline: 31660231]
61. Casanova-Perez R, Apodaca C, Bascom E, Mohanraj D, Lane C, Vidyarthi D, et al. Broken down by bias: healthcare biases experienced by BIPOC and LGBTQ+ patients. AMIA Annu Symp Proc 2021;2021:275-284 [FREE Full text] [Medline: 35308990]
62. Maindal HT, Aagaard-Hansen J. Health literacy meets the life-course perspective: towards a conceptual framework. Glob Health Action 2020 Dec 31;13(1):1775063 [FREE Full text] [doi: 10.1080/16549716.2020.1775063] [Medline: 32588781]
63. Perry EL. Health literacy in adolescents: an integrative review. J Spec Pediatr Nurs 2014 Jul 11;19(3):210-218. [doi: 10.1111/jspn.12072] [Medline: 24612548]
64. Wilandika A, Pandin MGR, Yusuf A. The roles of nurses in supporting health literacy: a scoping review. Front Public Health 2023 Aug 16;11:1022803 [FREE Full text] [doi: 10.3389/fpubh.2023.1022803] [Medline: 37663836]
65. Wittenberg E, Ferrell B, Kanter E, Buller H. Nurse communication challenges with health literacy support. Clin J Oncol Nurs 2018 Feb 01;22(1):53-61 [FREE Full text] [doi: 10.1188/18.CJON.53-61] [Medline: 29350714]

Abbreviations

DHL: digital health literacy
eHL: electronic health literacy
HL: health literacy
LGBTQIA+: lesbian, gay, bisexual, transgender, queer, questioning, intersex, asexual, and other
NIH: National Institutes of Health
PI: project investigator

XSL•FO
RenderX

RePORTER: RePORT Expenditures and Results
SDoH: social determinants of health

Edited by KJ(Craig; submitted 03.02.24; peer-reviewed by L Chinthala, R Marshall, YM Huang; comments to author 27.03.24; revised version received 22.05.24; accepted 17.09.24; published 31.10.24.

Please cite as:
Sakhuja M, Yelton B, Kavarana S, Schaurer L, Rumthao JR, Noblet S, Arent MA, Macauda MM, Donelle L, Friedman DB
How Do Scholars Conceptualize and Conduct Health and Digital Health Literacy Research? Survey of Federally Funded Scholars
J Med Internet Res 2024;26:e57040
URL: https://www.jmir.org/2024/1/e57040
doi:10.2196/57040
PMID:

©Mayank Sakhuja, Brooks Yelton, Simone Kavarana, Lauren Schaurer, Jancham Rachel Rumthao, Samuel Noblet, Michelle A Arent, Mark M Macauda, Lorie Donelle, Daniela B Friedman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Relationships Among eHealth Literacy, Physical Literacy, and Physical Activity in Chinese University Students: Cross-Sectional Study

Shan Jiang[1], MSc; Johan Y Y Ng[1], PhD; Siu Ming Choi[2], PhD; Amy S Ha[1], PhD

[1]Department of Sport Science and Physical Education, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
[2]Faculty of Education, University of Macau, Macau, China

Corresponding Author:
Amy S Ha, PhD
Department of Sport Science and Physical Education
The Chinese University of Hong Kong
G05 Kwok Sports Building, Shatin, N.T.
Hong Kong
China (Hong Kong)
Phone: 852 39436083
Email: sauchingha@cuhk.edu.hk

Abstract

Background: eHealth literacy is critical for evaluating abilities in locating, accessing, and applying digital health information to enhance one's understanding, skills, and attitudes toward a healthy lifestyle. Prior research indicates that enhancing eHealth literacy can improve health behaviors such as physical activity (PA). Physical literacy (PL) refers to the ability to develop sustainable PA habits, taking into account various aspects of an individual. Notably, university students have shown a decline in PA and possess low PL levels. However, the connection between eHealth literacy and PL in this demographic has not been extensively studied, and it remains uncertain whether PA acts as a mediator between eHealth literacy and PL.

Objective: This study examines the extent to which PA mediates the link between eHealth literacy and PL in Chinese university students and explores gender differences in these variables.

Methods: In February 2022, a cross-sectional survey was administered to 1210 students across 3 universities in China. The instruments used were the Perceived PL Instrument, the International Physical Activity Questionnaire, and the Chinese version of the eHealth Literacy Scale. Correlations between eHealth literacy, PA, and PL were analyzed using Pearson product-moment correlation and multiple linear regression, while mediation models helped elucidate the interactions among the 3 variables.

Results: The response rate for the study was 92.9% (1124/1210). In the mediation analysis, eHealth literacy showed a significant direct effect on PL, with a coefficient of 0.78 (β .75, SE 0.02; $P<.001$). Moderate to vigorous physical activity (MVPA) accounted for 2.16% of the total effect, suggesting that MVPA partially mediates the relationship between eHealth literacy and PL. Additionally, male students outperformed female students in terms of MVPA ($t_{636}=4.94$; $P<.001$) and PL ($t_{636}=3.18$; $P<.001$), but no significant differences were found in eHealth literacy ($t_{636}=1.23$; $P=.22$).

Conclusions: The findings indicate that MVPA serves as a mediator in the link between eHealth literacy and PL among university students. Students with low eHealth literacy or limited PA are less likely to be physically literate. Thus, eHealth literacy plays a crucial role in enhancing PL and PA, especially when interventions targeting PL are implemented. Our results also suggest a need for targeted health education interventions aimed at improving MVPA and PL among female students, while also recognizing that eHealth literacy is comparable across genders at universities.

(J Med Internet Res 2024;26:e56386) doi:10.2196/56386

KEYWORDS

eHealth literacy; physical literacy; physical activity; university students; health behavior

XSL•FO
RenderX

Introduction

Physical Activity and Health

In recent times, health challenges have escalated into a global issue. Physical activity (PA) is consistently recognized for its benefits to individual health and as a preventative measure for prevalent noncommunicable diseases. The World Health Organization (WHO) advocates for adults to partake in a minimum of 150 minutes per week of moderate physical activity (MPA) [1,2]. Nonetheless, it is reported that about 24% of young adults have not met these PA guidelines and are in urgent need to address the health risks associated with inadequate PA [3]. A meta-analysis indicates that physical inactivity affects 40% to 50% of university students [4]. Particularly in mainland China, a significant decline in PA has been observed among university students. Studies have identified that the prevalence of engaging in less than 1 hour of PA is highest among Chinese men aged 18 years and women aged 21 years, with rates of 82.5% and 89.8%, respectively [5]. Furthermore, the proliferation of the internet and increased use of electronic devices, such as video games and smartphones, have significantly decreased PA time and frequency [6-8]. Consequently, enhancing PA levels among university students, effectively using web-based resources, and boosting their health awareness are imperative in promoting a culture of wellness, reducing sedentary behaviors, and fostering lifelong habits of healthy living. In addition, moderate to vigorous physical activity (MVPA) plays a crucial role in the overall health and well-being of young adults. Research consistently highlights the significance of MVPA in this demographic due to its profound impact on cardiovascular health [9,10], cognitive function [11,12], and mental health [13]. Engaging in MVPA not only helps to maintain a healthy weight and reduce the risk of chronic diseases but also enhances overall physical fitness and endurance [9]. Moreover, MVPA has been linked to improved emotional regulation, stress reduction, and better sleep patterns, all of which are vital aspects of mental health in young adults and their late life [13,14].

eHealth Literacy

Electronic resources have become integral in disseminating health knowledge and play a pivotal role in public health [15]. Coined by Norman and Skinner [16], “eHealth literacy” refers to the capacity of individuals to search, comprehend, and assess basic health information from the internet, and to apply such knowledge in solving health-related issues [17]. University students are key disseminators of health knowledge and practitioners of health behaviors [18]. They are also primary consumers of digital information, making their understanding of eHealth crucial for societal impact. Numerous studies indicate that eHealth literacy among university students is shaped by factors such as educational environment, gender, and familial background [19,20].

Concerning differences in educational background, medical students typically exhibit higher eHealth literacy and more accessible electronic health information than nonmedical students [18]. Recent findings suggest that men display higher eHealth literacy compared to women [21,22]. Moreover, eHealth literacy has been associated with better health status and healthier behaviors including regular consumption of healthy foods and consistent PA [23-26]. Those with enhanced eHealth literacy are more inclined to participate regularly in PA [27]. However, the exploration of eHealth literacy in China remains nascent. Few studies have investigated the link between eHealth literacy and PA among university students [22,28] including the variations in eHealth literacy among students in health-related fields such as medicine and sports science.

Physical Literacy

Physical literacy (PL) encompasses the skills and attributes demonstrated by individuals through PA and movement throughout their life span [29]. It is both a process and outcome pursued through an amalgamation of physical, emotional, social, and cognitive interactions [30,31]. These interconnected domains foster the comprehensive development of PL, aiding individuals of all ages to lead active, healthy, and fulfilling lives. Additionally, the WHO has emphasized the integration of PL with health promotion and the enhancement of PA in national public health strategies [32]. Previous research underscores that PL is fundamental to PA, and individuals with low PL are less likely to engage in substantial PA [33-35]. Moreover, advancements in internet technologies, such as active video games and social media, have positively influenced students’ PL [36-38]. Thus, it is crucial to improve PL among university students and focus on the interrelation between eHealth literacy and PL. Notably, the link between eHealth literacy and PL has yet to be fully explored.

The Mediating Role of PA in the Relationship Between eHealth Literacy and PL

Delving into the complex relationships among eHealth literacy, PL, and PA is essential. PA acts as a crucial behavioral mediator through which eHealth literacy influences PL. Elevated eHealth literacy enables individuals to access a variety of digital resources, apps, and wearable technologies that encourage PA [39] participation. Web-based platforms also provide instructional content that educates users on proper exercise techniques, the benefits of PA, and methods to integrate movement into daily routines. This information aids individuals in acquiring the necessary skills and knowledge of diverse physical activities, thereby boosting their motivation and self-efficacy for PA engagement, which enhances overall PL [40].

Participation in PA yields cognitive advantages, such as improved memory, attention, and executive functions, enhancing the ability of individuals to understand and apply health-related information from eHealth platforms [41,42]. This synergy between eHealth literacy and PL is further strengthened by the social support found in PA activities, characterized by community interaction and teamwork. Such social engagement is especially beneficial for university students, promoting consistent PA participation. With higher eHealth literacy, individuals gain better access to eHealth resources including social functionalities that enable connection with peers, group participation, and involvement in community challenges [43]. This social integration fosters a sense of community and

XSL•FO
RenderX

accountability, encouraging individuals to remain active and enhance their PL through collective support and motivation.

Gender Differences in PA, eHealth Literacy, and PL

Exploring the disparities in health behaviors and literacy across genders is crucial [15,44]. Prior research indicates that men typically engage in higher levels of PA than women among adolescents and young adults, influenced by societal norms, resource availability, and personal inclinations [45]. Additionally, variations in eHealth literacy have been observed between genders, with some studies suggesting that women may exhibit higher eHealth literacy than men [22,39,46]. Concerning PL, prior studies have explored potential gender-based disparities in its development, pinpointing unique strengths and challenges for men and women [47-49]. These findings underscore the importance of tailored educational interventions to bolster PL within distinct gender cohorts [40,50]. Understanding these gender differences is pivotal for crafting targeted strategies to promote overall health and wellness among both men and women.

Thus, the primary objective of this study is to examine the mediating role of PA between eHealth literacy and PL among Chinese university students. We also aimed to assess how individual factors influence these variables such as gender. MVPA, as a key component of PA, plays a crucial role in promoting overall health development [1,3]. Thus, we hypothesized that eHealth literacy is positively associated with university students' MVPA (H1); MVPA will mediate the relationship between eHealth literacy and PL (H2); and gender differences may exist in eHealth literacy, MVPA, and PL among Chinese university students (H3).

Methods

Recruitment

A cross-sectional survey was conducted using Wenjuanxing, a popular Chinese web-based questionnaire platform, from January to February 2022. Initially, a pilot test was undertaken with 89 sophomore students from Beijing Normal University to assess the readability and item wording of the questionnaire. Feedback was gathered, and after a preliminary analysis, the questionnaire was revised accordingly.

Participants were enlisted through a snowball sampling approach. A recruitment notice was circulated via WeChat, a widely used social media platform in mainland China. Six university students, who were active members of their university student associations, were purposively chosen from 3 prominent Chinese universities—Beijing Normal University, Beijing Jiaotong University, and Beijing University of Chinese Medicine—following a rigorous selection process. These students were selected due to their regular involvement in university activities, which provided them with the capacity to reach various academic groups within their institutions. Following their anonymous participation in the survey, these students invited their peers and other university students to partake in the study through WeChat in Chinese. Importantly, there was no overlap of participants between the pilot and main studies to minimize bias.

Using parameters such as a total student body of 40,000 at each university, a 5% margin of error, a 95% CI, and a 50% response rate assumption, the necessary sample size was calculated using Raosoft [51], resulting in an optimal sample size of 381 participants.

Participants

If participants agreed to participate, they could click the "Start the Survey" button to proceed with the questionnaire; those who chose not to participate could simply click the "Close" button. Eligible participants included Chinese-speaking, full-time undergraduates or postgraduates aged 18-28 years at the specified universities. Although previous studies have indicated varying age ranges for Chinese university students [28,52-55], we have reviewed the current situation in Chinese universities and found that the majority of undergraduate and postgraduate students complete their studies between the ages of 18 and 28, which is the typical university age. This age range is crucial for understanding the dynamics of eHealth literacy, PL, and PA, as these factors are often influenced by the educational and developmental milestones typical for this demographic. The exclusion criteria were (1) individuals with missing or incomplete survey data, (2) respondents who provided inconsistent or implausible answers, (3) those who did not comply with the survey completion guidelines, and (4) individuals with medical conditions or physical constraints that might affect participation.

Ethical Considerations

This study received approval from the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong (EDU2022-069). Prospective participants accessed the survey via a link and were initially presented with an informed consent form. The purpose of both the survey and the informed consent form was clearly stated on the first page of the questionnaire. We ensured that participants' information would be kept confidential and anonymous. Participation was voluntary, and no financial incentives were offered.

Instruments and Measurements

eHealth Literacy

The main measure of this study was the Chinese version of the eHealth Literacy Scale (eHEALS). It comprises 8 items, with responses ranging from 1 (strongly disagree) to 5 (strongly agree), totaling a possible 40 points [56,57]. The scale was adapted following WHO guidelines, and its reliability was confirmed (Cronbach α=0.85) [56].

PL Level

Perceived PL among Chinese university students was evaluated using the Perceived Physical Literacy Instrument (simplified Chinese version) [35]. This instrument includes 8 items distributed across 3 dimensions: confidence and physical competence (eg, "I am physically fit, in accordance with my age"), motivation (eg, "I aspire to know the current sports trend"), and interaction with the environment (eg, "I have strong social skills"). To be deemed physically literate, individuals must exhibit confidence and capability, maintain a positive attitude toward PA [58,59], and effectively interact with their

XSL•FO
RenderX

environment through adept self-expression and communication [29]. Responses were gauged on a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Confirmatory factor analysis validated the questionnaire, with factor loadings ranging from 0.60 to 0.92 [60]. All 3 domains demonstrated reliability (Cronbach α=0.87, 0.83, and 0.86).

PA and Sedentary Behavior

PA levels were assessed using items from the International Physical Activity Questionnaire-Short Form, which measures the intensity and duration of vigorous PA, MPA, and walking over the past 7 days [61,62]. Metabolic equivalent (MET) task scores were calculated based on the International Physical Activity Questionnaire scoring protocol [63], using a weighted total (vigorous=8.0, moderate=4.0, walking=3.3) to estimate the MET of the task (min/wk). The higher the minutes, the greater the PA level. This instrument assesses 3 levels of PA based on METs, and sedentary behavior (ie, sitting time) was also recorded.

Participants reported the frequency and duration of each PA level, with activities lasting at least 10 minutes being counted. Total weekly PA was calculated using the following formula: *total MET minutes per week= walking PA (METs × min × days) + MPA (METs × min × days) + vigorous PA (METs × min × days)* [62]. Sedentary behavior was assessed through a single question about the weekly hours spent sitting. The Chinese questionnaire has been validated in previous studies (age range 15-55 years) and exhibits good test-retest reliability (intraclass correlation coefficient=0.79) [64].

Statistical Analysis

Data were analyzed using SPSS (version 28; IBM Corp) [65] and the PROCESS macro version 4.0 [66], a tool for mediation and moderation analysis. Two-tailed tests were used, with *P* values <.05 indicating statistical significance. Descriptive statistics described participant demographics, eHEALS, PA and sedentary behavior, and PL scores as means, (SDs) or frequencies (percentages). Two-tailed *t* tests evaluated gender differences in continuous and categorical variables. All continuous variables in the regression model were standardized.

ANOVAs assessed the significance of scores for participants in medicine and sports science. Normality, linearity, and homogeneity of variance were verified before analysis. Following the Pearson correlation, multiple linear regression analyzed the relationships among eHealth literacy, PA, and PL.

Mediation analysis followed established guidelines [67,68], requiring several conditions: (1) a significant path from eHealth literacy to MVPA (path c); (2) a significant association of eHealth literacy with PL (path a); (3) a significant path from PL to MVPA when controlling for eHealth literacy (path b); and (4) a significant indirect effect of PL (a×b) on the relationship between eHealth literacy and MVPA. Bootstrapping with 5000 samples assessed the 95% bias-corrected CI for the indirect effect (Preacher and Hayes [66]). A significant mediating effect was indicated if the CI for the indirect effect did not include zero. Separate regression analyses for gender differences in eHealth literacy PA and PL were also conducted due to observed variations.

Results

Participant Characteristics

The number of responses exceeded the anticipated sample size of 381 per university, reflecting high participant interest. In total, 410 responses were distributed to Beijing Normal University and 400 each to Beijing Jiaotong University and Beijing University of Traditional Chinese Medicine, with a total of 1210 questionnaires. Of these, 22 responses were not returned and 64 responses were incomplete, leaving 1124 usable responses (female, n=488, 43.4%; male, n=636, 56.6%). Overall, 86 (7.1%) of the 1210 responses were incomplete and subsequently excluded, leaving 389 responses from Beijing Normal University, 358 responses from Beijing Jiaotong University, and 377 valid responses from Beijing University of Traditional Chinese Medicine. Table 1 illustrates that participant ages ranged from 17 to 28 years, with a mean age of 19.46 (SD 1.58) years. Of the 1124 participants, a majority (n=758, 67.4%) were freshmen, and a smaller group (n=191, 16.9%) were sophomores. Regarding academic disciplines, the majority were science majors (n=726, 64.6%), followed by sports science (n=181, 16.1%), medical studies (n=137, 12.2%), and social sciences (n=80, 7.1%). In this study, “Science” refers specifically to traditional natural sciences, including disciplines such as physics, chemistry, and biology. Geographically, 70.7% (n=795) of participants originated from urban areas, while 29.3% (n=329) came from rural regions.

XSL•FO
RenderX

Table 1. Descriptive statistics of participants' characteristics.

Variables	Value (n=1124), n (%)
Gender, n (%)	
Female	488 (43.4)
Male	636 (56.6)
Age (years), mean (SD)	19.46 (1.58)
Education level, n (%)	
Freshman	758 (67.4)
Sophomore	191 (16.9)
Junior	94 (8.4)
Senior	43 (3.8)
Postgraduate	38 (3.4)
Major, n (%)	
Social Science	80 (7.1)
Medicine	137 (12.2)
Sports Science	181 (16.1)
Science	726 (64.6)
Residence, n (%)	
Urban area	795 (70.7)
Rural area	329 (29.3)
eHealth literacy, mean (SD)	32.63 (6.40)
Physical literacy, mean (SD)	31.68 (6.72)
Walking MET[a] minutes per week, mean (SD)	922.71 (1204.94)
Moderate MET minutes per week, mean (SD)	498.76 (865.55)
Vigorous MET minutes per week, mean (SD)	1181.2 (1961.24)
Moderate to vigorous physical activity MET minutes per week, mean (SD)	1679.96 (2619.45)
Total physical activity MET minutes per week, mean (SD)	2602.67 (3245.53)
Sitting total minutes per week, mean (SD)	2398.93 (1239.18)
Average sitting total minutes per day, mean (SD)	342.70 (177.03)

[a]MET: metabolic equivalent.

Descriptive Characteristics of eHealth Literacy and PL

Table 2 presents the average eHealth literacy score as 32.63 (SD 6.40), ranging from 8 to 40. It was noted that participants generally understood the availability of health resources on the internet (mean score, 4.12, SD 0.86), but scored lowest on using eHealth information for decision-making (mean score, 4.05, SD 0.87).

The mean score for total PL was 31.68 (SD 6.72), as detailed in Table 2. This encompassed 3 areas: confidence and physical competence (mean 11.89, SD 2.57), motivation (mean 12.63, SD 2.44), and interaction with the environment (mean 7.89, SD 1.83). Among these, the highest scoring item was awareness of the health benefits of sports (mean 4.32, SD 0.83), while the lowest was physical fitness relative to age (mean 3.73, SD 1.12).

XSL•FO
RenderX

Table 2. Descriptive statistics of eHealth literacy and physical literacy for university students.

Item	Value, mean (SD)
eHealth literacy (range 8-40)	32.63 (6.40)
Awareness	8.19 (1.67)
I know what health resources are available on the internet	4.12 (0.86)
I know where to find helpful health resources on the internet	4.07 (0.88)
Skills	12.23 (2.52)
I know how to find helpful health resources on the internet	4.08 (0.88)
I know how to use the internet to answer my questions about health	4.08 (0.88)
I know how to use the health information I find on the internet to help me	4.07 (0.88)
Evaluation	12.20 (2.47)
I have the skills I need to evaluate the health resources I find on the internet	4.07 (0.86)
I can tell high-quality health resources from low-quality health resources on the internet	4.08 (0.85)
I feel confident in using information from the internet to make health decisions	4.05 (0.87)
Physical literacy (range 8-40)	31.68 (6.72)
Confidence and physical competence	11.89 (2.57)
I possess adequate fundamental movement skills	4.21 (0.88)
I am physically fit, in accordance to my age	3.73 (1.12)
I am able to apply learned motor skills to other physical activities	3.95 (0.94)
Motivation	12.63 (2.44)
I appreciate myself or others doing sports	4.23 (0.89)
I am aware of the benefits of sports related to health	4.32 (0.83)
I aspire to know the current sports trend	4.07 (1.00)
Interaction with the environment	7.89 (1.83)
I have strong communication skills	3.96 (0.98)
I have strong social skills	3.94 (1.03)

Different Majors in PL, PA, and eHealth Literacy

Descriptive Statistics of the PA-Related Variables, PL, and eHealth Literacy

Significant differences in walking MET, MPA MET, vigorous PA MET, MVPA MET, and PL scores were observed across different majors, as shown in Table 3, although no significant variation in eHealth literacy scores was detected (P=.14).

XSL•FO
RenderX

Table 3. Descriptive statistics of physical activity, physical literacy, and eHealth literacy in different academic majors.

Item	Major, mean (SD)				*F* test (*df*)	*P* value
	Social Science (n=80)	Science (n=726)	Medicine (n=137)	Sports Science (n=181)		
Total MVPA[a]	1683.3 (2370.4)	1471.5 (2508.5)	823.1 (1161.0)	3163.3 (3321.6)	27.6 (3, 1120)	<.001
VPA[b]-MET[c] minutes per week	1136.8 (1626.5)	1016.7 (1886.2)	553.3 (868.9)	2336.1 (2487.1)	29.4 (3, 1120)	<.001
MPA[d]-MET minutes per week	546.5 (902.6)	454.8 (849.2)	269.8 (421.3)	827.2 (1058.5)	13.0 (3, 1120)	<.001
Walking-MET minutes per week	837.0 (1046.4)	924.4 (1236.0)	729.0 (1053.8)	1100.4 (1233.7)	2.6 (3, 1120)	.048
Physical literacy	31.7 (7.0)	31.4 (6.7)	30.2 (6.7)	33.8 (6.3)	8.7 (3, 1120)	<.001
eHealth literacy	32.3 (6.4)	32.6 (6.5)	31.9 (5.9)	33.5 (6.5)	1.9 (3, 1120)	.14

[a]MVPA: moderate to vigorous physical activity.

[b]VPA: vigorous physical activity.

[c]MET: metabolic equivalent.

[d]MPA: moderate physical activity.

Distribution of PA Levels Across Different Majors

Different majors in PA levels are shown in Table 3, which indicates that sports science majors reported the highest weekly MVPA (3163.3 MET min), in contrast to medicine majors who reported the lowest (823.1 MET min). Regarding low-intensity PA, like walking, medical students recorded the lowest levels (729 MET min), followed by those in social sciences (837 MET min), science (924.42 MET min), and sports science (1010.4 MET min).

Analysis of eHealth Literacy and PL Among Students Majoring in Sports Science and Medicine

Table S1 in Multimedia Appendix 1 shows significant differences in eHEALS (P=.02) and PL (P<.001) among sports science students and medicine major students. It also shows significant differences in confidence and physical competence (P<.001), interaction with the environment (P<.001), total PL (P<.001), and motivation (P=.002), in sports science and medicine students.

Gender Difference in PL, PA, and eHealth Literacy

The Distribution of PA Levels Across Different Genders

PA levels were categorized into 3 groups: low, moderate, and high. Low PA was defined as less than 600 MET minutes per week, high PA as more than 3000 MET minutes per week, and MPA as between 600 and 3000 MET minutes per week. Table S2 in Multimedia Appendix 1 indicates that nearly half (554/1124, 49.3%) of the participants fell into the MPA category. Among them, 18.1% (115/636) of male students were at the low level, while a higher proportion of female students (129/488, 26.4%) were classified as having low PA. The percentage of individuals at the high PA level was greater for male students (224/636, 35.2%) compared to female students.

Analysis of PA, PL, and eHealth Literacy in Different Genders

Table S3 in Multimedia Appendix 1 presents the differences between male and female students in BMI, PA, PL, and eHEALS. The vigorous physical activity, MPA, MVPA, and walking MET were significantly different between these 2 groups (P<.001). However, there were no significant differences for eHealth literacy (P=.23), and male participants had a higher BMI and PL than female participants.

Correlations Between PL, MVPA, and eHealth Literacy in Different Genders

Pearson correlations between PL, MVPA, and eHealth literacy are presented in Table S4 in Multimedia Appendix 1. eHealth literacy and MVPA were significantly correlated with PL (P<.001), and eHealth literacy was also significantly associated with MVPA (P<.001).

Mediation Model Test

Multiple linear regression was used to evaluate the relationships between eHealth literacy, PL, and MVPA, as outlined in Table 4. All identified correlations were positive and robust, confirming significant associations among these variables. The regression models formulated were as follows: $PL = 5.57 + 0.80 \times eHealth\ Literacy$, $MVPA = -390.64 + 63.46 \times eHealth\ Literacy$, and $PL = 30.711 + 0.001 \times MVPA$.

In model 1, significant pathways were observed from eHealth literacy to PL (β .75, SE 0.02; P<.001) and from eHealth literacy to MVPA (β=.22; P<.001), as depicted in Figure 1. Additionally, a significant link was found from MVPA to PL (β=.08; P<.001; Figure 1). Table 5 displays the mediating role of PA in the interaction between PL and eHealth literacy. The mediation analysis indicated a direct effect of eHealth literacy on PL at 0.78 and an indirect effect at 0.18. Furthermore, the total effect was calculated at 0.80. Bootstrap analysis with 5000 samples demonstrated an SE of 0.01 and a 95% bias-corrected CI ranging from 0.01 to 0.29, excluding zero. Thus, MVPA served as a

XSL•FO
RenderX

partial mediator, accounting for 2.22% of the total effect in the relationship between eHealth literacy and PL.

Model 2 explored the mediating effect of MVPA on the relationship between eHealth literacy and PL across genders, as visually represented in Figure 2 For male students, significant paths from eHealth literacy to PL (β=.81; *P*<.001) and from eHealth literacy to MVPA (β=.22; *P*<.001) were noted. For female students, eHealth literacy was significantly associated with both PL (β=.64; *P*<.001) and MVPA (β=.13; *P*<.001).

Table 5 further delineates the mediation effects within the relationship among eHealth literacy, MVPA, and PL-related variables. The mediation model indicated direct effects of 0.82 for male students and indirect effects of eHealth literacy on PL at 0.02, which were statistically significant. MVPA's mediating role accounted for 1.84% of the total effect, affirming its position as a partial mediator between eHealth literacy and PL. In female students, the mediation model of eHealth literacy on PL revealed direct effects of 0.717 and indirect effects of 0.016, with MVPA contributing to 2.16% of the total effect.

Table 4. Results of the standard linear regression analysis among eHealth literacy, PL[a], and MVPA[b].

Participants and item	β (SE)	*F* test (*df*)	95% CI	*R*	ΔR^2
All					
eHEALS[c]					
PL	.75 (0.02)[d]	1558.64 (1, 1122)	0.76-0.84	0.77	0.59
MVPA	.22 (12.06)[d]	27.68 (1, 1122)	39.79-87.13	0.76	0.58
MVPA					
PL	.08 (0.00)[d]	59.63 (1, 1122)	0.0004-0.0007	0.21	0.05
Male students					
eHEALS					
PL	.81 (0.02)[d]	1333.42 (1, 634)	0.79-0.88	0.82	0.68
MVPA	.22 (16.71)[d]	17.05 (1, 634)	36.18-101.80	0.16	0.03
MVPA					
PL	.07 (0.0001)[d]	33.64 (1, 634)	0.0004-0.0007	0.22	0.05
Female students					
eHEALS					
PL	.64 (0.04)[d]	370.08 (1, 486)	0.66-0.81	0.66	0.45
MVPA	.13 (15.89)[d]	8.13 (1, 486)	14.10-76.56	0.13	0.17
MVPA					
PL	.11 (0.0002)[d]	18.80 (1, 486)	0.0001-0.0006	0.19	0.04

[a]PL: physical literacy.

[b]MVPA: moderate to vigorous physical activity.

[c]eHEALS: eHealth literacy scale.

[d]Correlation is significant at the .01 level (2-tailed).

XSL•FO
RenderX

Figure 1. Model 1. Relationship between eHealth literacy and physical literacy with moderate to vigorous physical activity as a mediator (****P*<.001).

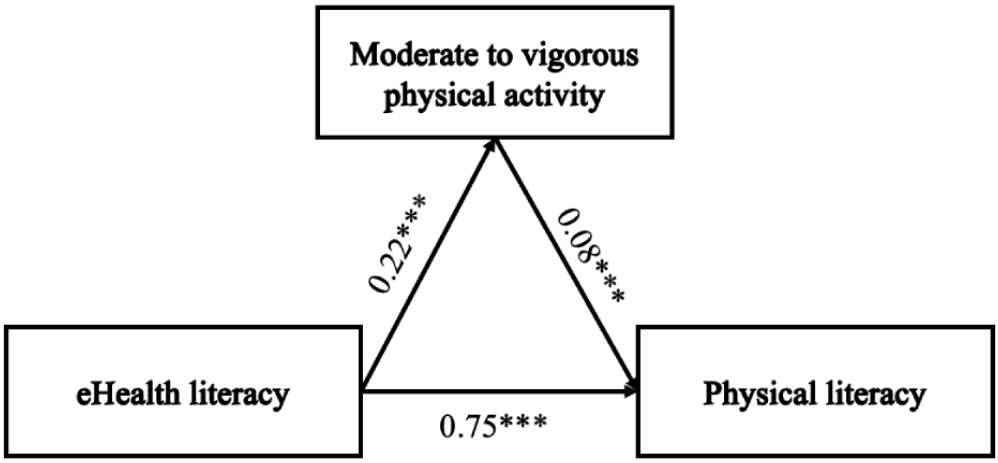

Table 5. Mediating effects of MVPA[a] on the relationship among eHealth literacy and PL[b] variables.

Mediating effect of MVPA on model	Direct effect			Indirect effect			Mediating effect (%)		
	All	Male	Female	All	Male	Female	All	Male	Female
eHEALS[c] and PL	0.78	0.82	0.72	0.18	0.02	0.016	2.22	1.84	2.16
eHEALS and CPC[d]	0.29	0.30	0.27	0.007	0.006	0.007	2.41	2.03	2.35
eHEALS and M[e]	0.26	0.28	0.23	0.004	0.003	0.005	1.36	0.95	2.01
eHEALS and IE[f]	0.20	0.21	0.17	0.003	0.003	0.003	1.65	1.3	1.55

[a]MVPA: moderate to vigorous physical activity.

[b]PL: physical literacy.

[c]eHEALS: eHealth literacy scale.

[d]CPC: confidence and physical competence.

[e]M: motivation.

[f]IE: interaction with the environment.

XSL•FO
RenderX

Figure 2. Model 2. The gender difference (male and female) in the relationship between eHealth literacy, physical literacy, and MVPA (***P<.001). MVPA: moderate to vigorous physical activity.

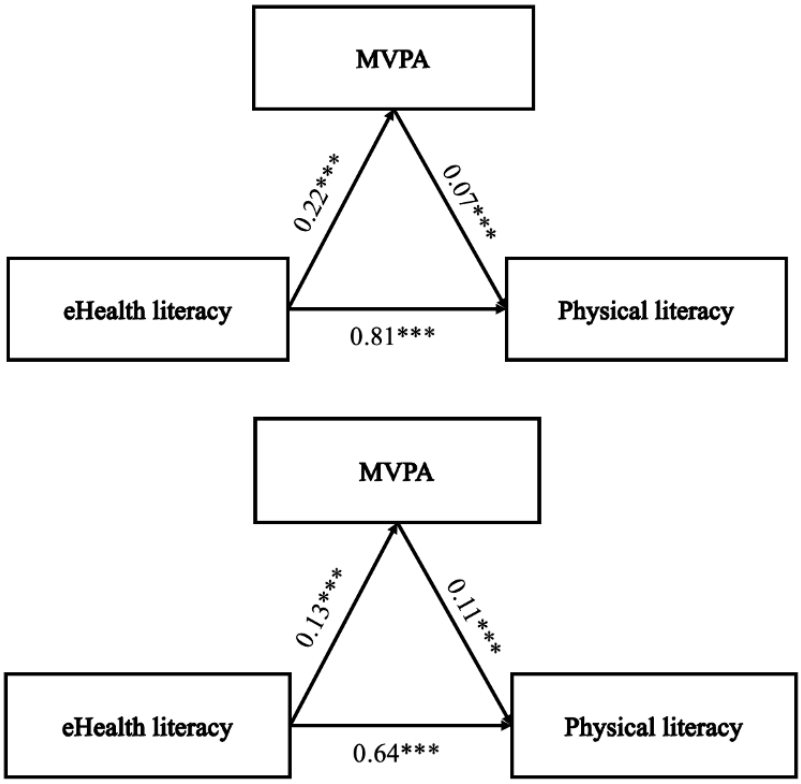

Discussion

Principal Findings

This investigation found robust positive associations between eHealth literacy and PL, supporting hypothesis 1. Additionally, the mediating role of MVPA was instrumental in delineating the connection between eHealth literacy and PL. Male students also showed higher PA and PL than female students, although no significant differences in eHealth literacy were observed. This study is pioneering in examining the linkage between eHealth literacy and PL among Chinese university students and is among the extensive investigations exploring MVPA's mediating effects. Overall, the findings indicate that both eHealth literacy and MVPA are [69] positively associated with [27,28,70] PL among university students.

Individuals With Higher eHealth Literacy Tend to Have Greater PL

This study identified eHealth literacy as a significant predictor of PL. Thus, university students with higher eHealth literacy are generally more physically literate. One mechanism linking eHealth literacy to improved PL is PA. Public health initiatives have increasingly focused on enhancing university students' eHealth literacy and PA, with numerous studies demonstrating that eHealth literacy is positively correlated with health behaviors including regular exercise and PA [27,28,70]. Additionally, both MVPA and health behaviors are strongly associated with PL [38]. The results of this study support the hypothesis that MVPA serves as a partial mediator in the relationship between eHealth literacy and PL. In essence, improving eHealth literacy to increase MVPA may enhance PL among university students.

eHealth Literacy Prevalence Among Chinese University Students

The average eHEALS score among the students was 32.63, surpassing scores reported in global studies [71-73]. This elevated level of eHealth literacy could be attributed to the enhanced availability of reputable health information sources in China. The Chinese government's initiatives to disseminate health information via social media and the accessibility of international health websites in Chinese may contribute to this trend [74]. Despite these high scores, many students demonstrated a lack of proficiency in distinguishing between high-quality and low-quality health information sources, often hesitating to make health decisions based on web-based information. These findings align with previous research [75,76].

Gender and Academic Major Differences in eHealth Literacy

Although male students achieved higher eHEALS scores than female students, the differences were not statistically significant, diverging from some prior studies in Taiwan [21,23,77]. Furthermore, the influence of academic majors on eHealth literacy levels was evident, with Chinese Medicine students displaying the lowest eHealth literacy. This finding contrasts with prior research indicating higher eHealth literacy among Western medical students [18,24]. The intensive curriculum and numerous internships in traditional Chinese medicine programs, which involve extensive memorization of classical texts, may reduce these students' engagement with digital health resources. However, sports science students recorded higher

XSL•FO
RenderX

eHEALS scores (33.5), a significant variance, supporting social cognitive theory [78] that environmental factors can influence health perceptions [79].

PA and PL Among Male Students

Consistent with previous research, this study found that PL influences PA levels, and participation in PA substantially affects PL [33]. An intriguing aspect of this study was the higher motivation for PA among university students, who exhibited lower confidence and competence in their physical activities. Moreover, gender differences in PL and PA were noted, with male students displaying higher levels than female students, corroborating findings from a study in southern China on the interplay between PL and PA [60]. Globally, men tend to be more physically active than women, with physical inactivity being more prevalent among women in nearly all countries [1,3]. Consequently, it is crucial to establish PA habits and motivate university students, particularly female students, to develop a lifelong commitment to PA.

A significant disparity was observed in PL and PA between medical students and sports science students. These findings align with a large quantitative study conducted in German universities assessing PA levels [69]. Sports science students, being more aware of the potential health consequences of physical inactivity, are more inclined to be physically active. Therefore, it is recommended that medical students increase their PA participation and engage in PL-related interventions and curricula. Further research should evaluate eHealth literacy in a larger cohort of Chinese university students majoring in medicine or health sciences.

Implications

Our findings have significant implications for educational practices in universities. Health educators are urged to enhance health education and develop robust health education programs to enhance overall health outcomes for university students. Given that physical education is mandatory in mainland China from primary school to university levels [80], integrating eHealth literacy into physical education classes could support holistic student development in PL. Additionally, incorporating content on MVPA into curricula could support students' PA and literacy levels. It is imperative for university-level eHealth interventions to use mobile health apps social media, and various digital resources to enhance students' skills in navigating and evaluating web-based health information [8].

Furthermore, scholars should investigate the factors contributing to gender disparities in PA and PL. Targeted health education programs at universities could significantly benefit female students by enhancing their PA and PL. Additionally, students in health-related fields, particularly medical students, require targeted educational interventions to ensure comprehensive mastery of eHealth resources. Previous studies indicate that women are more physically inactive [81,82] and have lower PL [38]. Therefore, enhancing their eHealth literacy and promoting PA could significantly improve their health outcomes.

Strengths and Limitations

This study is distinguished by its diverse participant base and the use of advanced analytical methods to elucidate MVPA's role in the dynamics between eHealth literacy and PL. However, it also has limitations. The participant pool was confined to 3 universities, suggesting that future research could broaden the scope to include more institutions. Moreover, the reliance on self-reported data for PA, PL, and eHealth literacy could introduce biases; subsequent studies might benefit from more objective data collection methods. Additionally, the cross-sectional design precludes causal inferences between eHealth literacy, MVPA, and PL. Finally, data collection coincided with the winter holiday, potentially affecting the typical PA levels of students.

Conclusions

This study elucidated the relationships among eHealth literacy, PL, and PA in a cohort of Chinese university students, highlighting the mediating role of MVPA. The results suggested that students with higher eHealth literacy and PA levels are likely to exhibit enhanced PL. Sports science students outperformed medical students in eHealth literacy, PL, and PA levels. Moreover, male students were more active and physically literate than female students, yet no significant gender differences in eHealth literacy were detected. Future research should aim to validate these findings through longitudinal studies and consider eHealth literacy and MVPA when designing PL intervention programs for university students.

Acknowledgments

The authors would like to thank all participants who were involved in this study, without whom the research would not have been possible.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

SJ and ASH designed the study. SJ conducted surveys and data analysis and wrote the first draft of the manuscript. SMC and JYYN revised all versions of the manuscript. ASH made some constructive suggestions and revised the final draft of the manuscript. All authors have read and agreed to the final draft of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 23 KB - jmir_v26i1e56386_app1.docx]

References

1. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
2. Kikuchi H, Inoue S, Lee IM, Odagiri Y, Sawada N, Inoue M, et al. Impact of moderate-intensity and vigorous-intensity physical activity on mortality. Med Sci Sports Exerc 2018;50(4):715-721. [doi: 10.1249/MSS.0000000000001463] [Medline: 29053480]
3. van Sluijs EMF, Ekelund U, Crochemore-Silva I, Guthold R, Ha A, Lubans D, et al. Physical activity behaviours in adolescence: current evidence and opportunities for intervention. Lancet 2021;398(10298):429-442 [FREE Full text] [doi: 10.1016/S0140-6736(21)01259-9] [Medline: 34302767]
4. Keating XD, Guan J, Piñero JC, Bridges DM. A meta-analysis of college students' physical activity behaviors. J Am Coll Health 2005;54(2):116-125. [doi: 10.3200/JACH.54.2.116-126] [Medline: 16255324]
5. Wang ZH, Dong YH, Song Y, Yang ZP, Ma J. Analysis of the prevalence and influencing factors of less than one hour of physical exercise among Chinese students aged 9 to 22 in 2014. Chin J Epidemiol 2017;38(3):341-345. [doi: 10.3760/cma.j.issn.0254-6450.2017.03.013] [Medline: 28329936]
6. Lepp A, Barkley JE, Sanders GJ, Rebold M, Gates P. The relationship between cell phone use, physical and sedentary activity, and cardiorespiratory fitness in a sample of U.S. college students. Int J Behav Nutr Phys Act 2013;10:79 [FREE Full text] [doi: 10.1186/1479-5868-10-79] [Medline: 23800133]
7. Gallè F, Sabella EA, Ferracuti S, de Giglio O, Caggiano G, Protano C, et al. Sedentary behaviors and physical activity of Italian undergraduate students during lockdown at the time of COVID-19 pandemic. Int J Environ Res Public Health 2020;17(17):6171 [FREE Full text] [doi: 10.3390/ijerph17176171] [Medline: 32854414]
8. Jiang S, Ng JYY, Chong KH, Peng B, Ha AS. Effects of eHealth interventions on 24-hour movement behaviors among preschoolers: systematic review and meta-analysis. J Med Internet Res 2024;26:e52905 [FREE Full text] [doi: 10.2196/52905] [Medline: 38381514]
9. Nagata JM, Vittinghoff E, Gabriel KP, Garber AK, Moran AE, Rana JS, et al. Moderate-to-vigorous intensity physical activity from young adulthood to middle age and metabolic disease: a 30-year population-based cohort study. Br J Sports Med 2022;56(15):847-853 [FREE Full text] [doi: 10.1136/bjsports-2021-104231] [Medline: 34521685]
10. Stamatakis E, Ahmadi MN, Elphick TL, Huang BH, Paudel S, Teixeira-Pinto A, et al. Occupational physical activity, all-cause, cardiovascular disease, and cancer mortality in 349,248 adults: prospective and longitudinal analyses of the MJ Cohort. J Sport Health Sci 2024;13(4):579-589 [FREE Full text] [doi: 10.1016/j.jshs.2024.03.002] [Medline: 38462173]
11. Nagata JM, Cortez CA, Dooley EE, Iyer P, Ganson KT, Gabriel KP. Moderate-to-vigorous intensity physical activity among adolescents in the USA during the COVID-19 pandemic. Prev Med Rep 2022;25:101685 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101685] [Medline: 35004134]
12. Wang J, Zhao X, Bi Y, Jiang S, Sun Y, Lang J, et al. Executive function elevated by long term high-intensity physical activity and the regulation role of beta-band activity in human frontal region. Cogn Neurodyn 2023;17(6):1463-1472. [doi: 10.1007/s11571-022-09905-z] [Medline: 37974584]
13. Yu H, Ma Q, Sun Y, Jiang S, Hu S, Wang X. Analyzing the effects of physical activity levels on aggressive behavior in college students using a chain-mediated model. Sci Rep 2024;14(1):5795 [FREE Full text] [doi: 10.1038/s41598-024-55534-3] [Medline: 38461174]
14. Vásquez PM, Durazo-Arvizu RA, Marquez DX, Argos M, Lamar M, Odoms-Young A, et al. Physical activity, stress, and cardiovascular disease risk: HCHS/SOL sociocultural ancillary study. Prev Med Rep 2020;20:101190 [FREE Full text] [doi: 10.1016/j.pmedr.2020.101190] [Medline: 32995142]
15. Chao DP. Associations between sociodemographic characteristics, eHealth literacy, and health-promoting lifestyle among university students in Taipei: cross-sectional validation study of the Chinese version of the eHealth literacy scale. J Med Internet Res 2024;26:e52314 [FREE Full text] [doi: 10.2196/52314] [Medline: 39024006]
16. Norman CD, Skinner HA. eHealth Literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
17. Stellefson M, Hanik B, Chaney B, Chaney D, Tennant B, Chavarria EA. eHealth literacy among college students: a systematic review with implications for eHealth education. J Med Internet Res 2011;13(4):e102 [FREE Full text] [doi: 10.2196/jmir.1703] [Medline: 22155629]

XSL•FO
RenderX

18. Tran HTT, Nguyen MH, Pham TTM, Kim GB, Nguyen HT, Nguyen NM, et al. Predictors of eHealth literacy and its associations with preventive behaviors, fear of COVID-19, anxiety, and depression among undergraduate nursing students: a cross-sectional survey. Int J Environ Res Public Health 2022;19(7):3766 [FREE Full text] [doi: 10.3390/ijerph19073766] [Medline: 35409448]
19. Cui GH, Li SJ, Yin YT, Chen LJ, Li JQ, Liang FY, et al. The relationship among social capital, eHealth literacy and health behaviours in chinese elderly people: a cross-sectional study. BMC Public Health 2021;21(1):45 [FREE Full text] [doi: 10.1186/s12889-020-10037-4] [Medline: 33407275]
20. Blebil AQ, Dujaili JA, Mohammed AH, Loh LL, Chung WX, Selvam T, et al. Exploring the eHealth literacy and mobile health application utilisation amongst Malaysian pharmacy students. J Telemed Telecare 2023;29(1):58-71. [doi: 10.1177/1357633X221077869] [Medline: 35188826]
21. Huang CL, Chiang CH, Yang SC. eHealth literacy and beliefs about medicines among Taiwanese college students: cross-sectional study. JMIR Med Inform 2021;9(11):e24144 [FREE Full text] [doi: 10.2196/24144] [Medline: 34851301]
22. Tsukahara S, Yamaguchi S, Igarashi F, Uruma R, Ikuina N, Iwakura K, et al. Association of eHealth literacy with lifestyle behaviors in university students: questionnaire-based cross-sectional study. J Med Internet Res 2020;22(6):e18155 [FREE Full text] [doi: 10.2196/18155] [Medline: 32579126]
23. Luo YF, Yang SC, Chen AS, Chiang CH. Associations of eHealth literacy with health services utilization among college students: cross-sectional study. J Med Internet Res 2018;20(10):e283 [FREE Full text] [doi: 10.2196/jmir.8897] [Medline: 30361201]
24. Hsu W, Chiang C, Yang S. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014;16(12):e287 [FREE Full text] [doi: 10.2196/jmir.3542] [Medline: 25499086]
25. Wang C, Wu X, Qi H. A comprehensive analysis of e-Health literacy research focuses and trends. Healthcare (Basel) 2021;10(1):66 [FREE Full text] [doi: 10.3390/healthcare10010066] [Medline: 35052230]
26. Liang W, Duan Y, Wang Y, Lippke S, Shang B, Lin Z, et al. Psychosocial mediators of web-based interventions for promoting a healthy lifestyle among chinese college students: secondary analysis of a randomized controlled trial. J Med Internet Res 2022;24(9):e37563 [FREE Full text] [doi: 10.2196/37563] [Medline: 36069840]
27. Aslan GK, Kartal A, Turan T, Yiğitoğlu GT, Kocakabak C. Association of electronic health literacy with health-promoting behaviours in adolescents. Int J Nurs Pract 2021;27(2):e12921. [doi: 10.1111/ijn.12921] [Medline: 33615625]
28. Li S, Cui G, Kaminga AC, Cheng S, Xu H. Associations between health literacy, eHealth literacy, and COVID-19-related health behaviors among Chinese college students: cross-sectional online study. J Med Internet Res 2021;23(5):e25600 [FREE Full text] [doi: 10.2196/25600] [Medline: 33822734]
29. Whitehead ME. Physical literacy. In: Grier J, editor. International Association of Physical Education and Sport for Girls and Women (IAPESWG) Congress, Melbourne, Australia, 31 July-06 August 1993. Melbourne, Australia: International Association of Physical Education and Sport for Girls and Women; 1993.
30. Ha AS, Chan W, Ng JYY. Relation between perceived barrier profiles, physical literacy, motivation and physical activity behaviors among parents with a young Child. Int J Environ Res Public Health 2020;17(12):4459 [FREE Full text] [doi: 10.3390/ijerph17124459] [Medline: 32575873]
31. Kwan MYW, Graham JD, Healey C, Paolucci N, Brown DM. Stopping the drop: examining the impact of a pilot physical literacy-based intervention program on physical activity behaviours and fitness during the transition into university. Int J Environ Res Public Health 2020;17(16):5832 [FREE Full text] [doi: 10.3390/ijerph17165832] [Medline: 32806584]
32. Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World. Geneva, Switzerland: World Health Organization; 2018.
33. Edwards LC, Bryant AS, Keegan RJ, Morgan K, Jones AM. Definitions, foundations and associations of physical literacy: a systematic review. Sports Med 2017;47(1):113-126 [FREE Full text] [doi: 10.1007/s40279-016-0560-7] [Medline: 27365029]
34. Brown DMY, Dudley DA, Cairney J. Physical literacy profiles are associated with differences in children's physical activity participation: a latent profile analysis approach. J Sci Med Sport 2020;23(11):1062-1067. [doi: 10.1016/j.jsams.2020.05.007] [Medline: 32475780]
35. Ma RS, Sum RKW, Li MH, Huang Y, Niu XL. Association between physical literacy and physical activity: a multilevel analysis study among Chinese undergraduates. Int J Environ Res Public Health 2020;17(21):7874 [FREE Full text] [doi: 10.3390/ijerph17217874] [Medline: 33121068]
36. Sun H. Operationalizing physical literacy: the potential of active video games. J Sport Health Sci 2015;4(2):145-149 [FREE Full text] [doi: 10.1016/j.jshs.2015.03.006]
37. Bopp T, Vadeboncoeur JD, Stellefson M, Weinsz M. Moving beyond the gym: a content analysis of YouTube as an information resource for physical literacy. Int J Environ Res Public Health 2019;16(18):3335 [FREE Full text] [doi: 10.3390/ijerph16183335] [Medline: 31510001]
38. Choi SM, Sum KWR, Leung FLE, Wallhead T, Morgan K, Milton D, et al. Effect of sport education on students' perceived physical literacy, motivation, and physical activity levels in university required physical education: a cluster-randomized trial. High Educ 2020;81(6):1137-1155. [doi: 10.1007/s10734-020-00603-5]

XSL•FO
RenderX

39. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023;25:e40778 [FREE Full text] [doi: 10.2196/40778] [Medline: 36716080]
40. Clark HJ, Dudley D, Barratt J, Cairney J. Physical literacy predicts the physical activity and sedentary behaviours of youth. J Sci Med Sport 2022;25(9):750-754. [doi: 10.1016/j.jsams.2022.04.008] [Medline: 35934659]
41. Wang X, Yue T, Mo PKH. The associations among cognitive social factors, eHealth literacy and health-promoting behaviors in Chinese adolescents. Health Promot Int 2022;37(6):daac143. [doi: 10.1093/heapro/daac143] [Medline: 36409148]
42. Moghaddam HR, Ranjbaran S, Babazadeh T. The role of e-health literacy and some cognitive factors in adopting protective behaviors of COVID-19 in Khalkhal residents. Front Public Health 2022;10:916362 [FREE Full text] [doi: 10.3389/fpubh.2022.916362] [Medline: 35942262]
43. Kim M, Kim B, Park S. Social support, eHealth literacy, and mHealth use in older adults with diabetes: moderated mediating effect of the perceived importance of app design. Comput Inform Nurs 2024;42(2):136-143. [doi: 10.1097/CIN.0000000000001081] [Medline: 38129323]
44. Zhang Y, Xu P, Sun Q, Baral S, Xi L, Wang D. Factors influencing the e-health literacy in cancer patients: a systematic review. J Cancer Surviv 2023;17(2):425-440 [FREE Full text] [doi: 10.1007/s11764-022-01260-6] [Medline: 36190672]
45. Jiang W, Luo J, Guan H. Gender difference in the relationship of physical activity and subjective happiness among Chinese university students. Front Psychol 2021;12:800515 [FREE Full text] [doi: 10.3389/fpsyg.2021.800515] [Medline: 34950093]
46. Amoah PA, Phillips DR. Socio-demographic and behavioral correlates of health literacy: a gender perspective in Ghana. Women Health 2020;60(2):123-139. [doi: 10.1080/03630242.2019.1613471] [Medline: 31092133]
47. Choi SM, Sum RKW, Leung EFL, Ng RSK. Relationship between perceived physical literacy and physical activity levels among Hong Kong adolescents. PLoS One 2018;13(8):e0203105 [FREE Full text] [doi: 10.1371/journal.pone.0203105] [Medline: 30148876]
48. Öztürk Ö, Aydoğdu O, Yıkılmaz SK, Feyzioğlu Ö, Pişirici P. Physical literacy as a determinant of physical activity level among late adolescents. PLoS One 2023;18(4):e0285032 [FREE Full text] [doi: 10.1371/journal.pone.0285032] [Medline: 37115768]
49. Li MH, Sum RKW, Sit CHP, Wong SHS, Ha ASC. Associations between perceived and actual physical literacy level in Chinese primary school children. BMC Public Health 2020;20(1):207 [FREE Full text] [doi: 10.1186/s12889-020-8318-4] [Medline: 32041576]
50. Boldovskaia A, Dias NMG, Silva MN, Carraça EV. Physical literacy assessment in adults: a systematic review. PLoS One 2023;18(7):e0288541 [FREE Full text] [doi: 10.1371/journal.pone.0288541] [Medline: 37450424]
51. Sample size calculator. Raosoft. 2004. URL: http://www.raosoft.com/samplesize.html [accessed 2024-10-14]
52. Wang J, Hao QH, Peng W, Tu Y, Zhang L, Zhu TM. Relationship between smartphone addiction and eating disorders and lifestyle among Chinese college students. Front Public Health 2023;11:1111477 [FREE Full text] [doi: 10.3389/fpubh.2023.1111477] [Medline: 37275494]
53. Luo W, Zhong BL, Chiu HFK. Prevalence of depressive symptoms among Chinese university students amid the COVID-19 pandemic: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2021;30:e31 [FREE Full text] [doi: 10.1017/S2045796021000202] [Medline: 33766163]
54. Zhai X, Ye M, Wang C, Gu Q, Huang T, Wang K, et al. Associations among physical activity and smartphone use with perceived stress and sleep quality of Chinese college students. Ment Health Phys Act 2020;18:100323 [FREE Full text] [doi: 10.1016/j.mhpa.2020.100323]
55. Ge Y, Xin S, Luan D, Zou Z, Liu M, Bai X, et al. Association of physical activity, sedentary time, and sleep duration on the health-related quality of life of college students in Northeast China. Health Qual Life Outcomes 2019;17(1):124 [FREE Full text] [doi: 10.1186/s12955-019-1194-x] [Medline: 31311564]
56. Ma Z, Wu M. The psychometric properties of the Chinese eHealth literacy scale (C-eHEALS) in a Chinese rural population: cross-sectional validation study. J Med Internet Res 2019;21(10):e15720 [FREE Full text] [doi: 10.2196/15720] [Medline: 31642811]
57. Britt RK, Collins WB, Wilson K, Linnemeier G, Englebert AM. eHealth literacy and health behaviors affecting modern college students: a pilot study of issues identified by the American college health association. J Med Internet Res 2017;19(12):e392 [FREE Full text] [doi: 10.2196/jmir.3100] [Medline: 29258979]
58. Zhang C, Liu Y, Xu S, Sum RKW, Ma R, Zhong P, et al. Exploring the level of physical fitness on physical activity and physical literacy among Chinese university students: a cross-sectional study. Front Psychol 2022;13:833461 [FREE Full text] [doi: 10.3389/fpsyg.2022.833461] [Medline: 35369138]
59. Sum RKW, Cheng CF, Wallhead T, Kuo CC, Wang FJ, Choi SM. Perceived physical literacy instrument for adolescents: a further validation of PPLI. J Exerc Sci Fit 2018;16(1):26-31 [FREE Full text] [doi: 10.1016/j.jesf.2018.03.002] [Medline: 30662489]
60. Ma RS, Sum RKW, Hu YN, Gao TY. Assessing factor structure of the simplified chinese version of perceived physical literacy instrument for undergraduates in mainland China. J Exerc Sci Fit 2020;18(2):68-73 [FREE Full text] [doi: 10.1016/j.jesf.2020.01.001] [Medline: 31998384]

XSL•FO
RenderX

61. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
62. Macfarlane DJ, Lee CCY, Ho EYK, Chan KL, Chan DTS. Reliability and validity of the Chinese version of IPAQ (short, last 7 days). J Sci Med Sport 2007;10(1):45-51. [doi: 10.1016/j.jsams.2006.05.003] [Medline: 16807105]
63. Sjostrom M, Ainsworth BE, Bauman A, Bull FC, Hamilton-Craig CR, Sallis JF, editors. Guidelines for Data Processing Analysis of the International Physical Activity Questionnaire (IPAQ)—Short and Long Forms. Geneva, Switzerland: IPAQ; 2005.
64. Macfarlane DJ, Lee CCY, Ho EYK, Chan KL, Chan DTS. Reliability and validity of the Chinese version of IPAQ (short, last 7 days). J Sci Med Sport 2007;10(1):45-51. [doi: 10.1016/j.jsams.2006.05.003] [Medline: 16807105]
65. IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp; 2021.
66. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput 2004;36(4):717-731. [doi: 10.3758/bf03206553] [Medline: 15641418]
67. Carrión GC, Nitzl C, Roldán JL. Mediation analyses in partial least squares structural equation modeling: guidelines empirical examples. In: Latan H, Noonan R, editors. Partial Least Squares Path Modeling: Basic Concepts, Methodological Issues and Applications. Cham, Switzerland: Springer International Publishing; 2017:173-195.
68. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annu Rev Psychol 2007;58:593-614 [FREE Full text] [doi: 10.1146/annurev.psych.58.110405.085542] [Medline: 16968208]
69. Diehl K, Brassat A, Hilger-Kolb J. Comparative physical activity as a global question to assess physical activity among university students. BMC Sports Sci Med Rehabil 2021;13(1):19 [FREE Full text] [doi: 10.1186/s13102-021-00247-7] [Medline: 33653387]
70. Cho H, Han K, Park BK. Associations of eHealth literacy with health-promoting behaviours among hospital nurses: a descriptive cross-sectional study. J Adv Nurs 2018;74(7):1618-1627. [doi: 10.1111/jan.13575] [Medline: 29575085]
71. Kim SH, Son YJ. Relationships between eHealth literacy and health behaviors in Korean adults. Comput Inform Nurs 2017;35(2):84-90. [doi: 10.1097/CIN.0000000000000255] [Medline: 27258808]
72. Li S, Cui G, Zhou F, Liu S, Guo Y, Yin Y, et al. The longitudinal relationship between eHealth literacy, health-promoting lifestyles, and health-related quality of life among college students: a cross-lagged analysis. Front Public Health 2022;10:868279 [FREE Full text] [doi: 10.3389/fpubh.2022.868279] [Medline: 35875048]
73. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet users. J Med Internet Res 2016;18(7):e192 [FREE Full text] [doi: 10.2196/jmir.5413] [Medline: 27432783]
74. Dashti S, Peyman N, Tajfard M, Esmaeeli H. E-health literacy of medical and health sciences university students in Mashhad, Iran in 2016: a pilot study. Electron Physician 2017;9(3):3966-3973 [FREE Full text] [doi: 10.19082/3966] [Medline: 28461871]
75. Dadaczynski K, Okan O, Messer M, Leung AYM, Rosário R, Darlington E, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021;23(1):e24097 [FREE Full text] [doi: 10.2196/24097] [Medline: 33395396]
76. Li S, Cui G, Yin Y, Wang S, Liu X, Chen L. Health-promoting behaviors mediate the relationship between eHealth literacy and health-related quality of life among Chinese older adults: a cross-sectional study. Qual Life Res 2021;30(8):2235-2243 [FREE Full text] [doi: 10.1007/s11136-021-02797-2] [Medline: 33661455]
77. Fung ICH, Hao Y, Cai J, Ying Y, Schaible BJ, Yu CM, et al. Chinese social media reaction to information about 42 notifiable infectious diseases. PLoS One 2015;10(5):e0126092 [FREE Full text] [doi: 10.1371/journal.pone.0126092] [Medline: 25946020]
78. Zhou J, Fan T. Understanding the factors influencing patient E-health literacy in online health communities (OHCs): a social cognitive theory perspective. Int J Environ Res Public Health 2019;16(14):2455 [FREE Full text] [doi: 10.3390/ijerph16142455] [Medline: 31295924]
79. Yang SC, Luo YF, Chiang CH. The associations among individual factors, eHealth literacy, and health-promoting lifestyles among college students. J Med Internet Res 2017;19(1):e15 [FREE Full text] [doi: 10.2196/jmir.5964] [Medline: 28073739]
80. Jin A. Physical education curriculum reform in China: a perspective from physical education teachers. Phys Educ Sport Pedagogy 2013;18(1):15-27. [doi: 10.1080/17408989.2011.623231]
81. Ha AS, Lonsdale C, Lubans DR, Ng FF, Ng JYY. Improving children's fundamental movement skills through a family-based physical activity program: results from the "Active 1 + FUN" randomized controlled trial. Int J Behav Nutr Phys Act 2021;18(1):99 [FREE Full text] [doi: 10.1186/s12966-021-01160-5] [Medline: 34273973]
82. The Lancet Public Health. Time to tackle the physical activity gender gap. Lancet Public Health 2019;4(8):e360 [FREE Full text] [doi: 10.1016/S2468-2667(19)30135-5] [Medline: 31345750]

Abbreviations

eHEALS: eHealth Literacy Scale
MET: metabolic equivalent

XSL•FO
RenderX

MPA: moderate physical activity
MVPA: moderate to vigorous physical activity
PA: physical activity
PL: physical literacy
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 15.01.24; peer-reviewed by C Zou, A Zalewska; comments to author 09.05.24; revised version received 09.08.24; accepted 23.09.24; published 04.11.24.

Please cite as:
Jiang S, Ng JYY, Choi SM, Ha AS
Relationships Among eHealth Literacy, Physical Literacy, and Physical Activity in Chinese University Students: Cross-Sectional Study
J Med Internet Res 2024;26:e56386
URL: https://www.jmir.org/2024/1/e56386
doi:10.2196/56386
PMID:

©Shan Jiang, Johan Y Y Ng, Siu Ming Choi, Amy S Ha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Associated With Digital Capacity for Health Promotion Among Primary Care Workers: Cross-Sectional Survey Study

Yining Wang[1*], BA; Hui Ren[2*], MSc; Shaotan Xiao[2*], MSc; Tian Meng[3]; Shuyue Sun[1], MSc; Siyu Yu[2], PhD; Qing Liu[2], MSc; Fan Wang[4], PhD

[1]School of Public Health, Fudan University, Shanghai, China

[2]Shanghai Pudong New Area Center for Disease Control and Prevention, Shanghai, China

[3]School of Accounting, Shanghai University of International Business and Economics, Shanghai, China

[4]Fudan Development Institute, Fudan University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Fan Wang, PhD
Fudan Development Institute
Fudan University
220 Handan Road
Shanghai, 200433
China
Phone: 86 21 55664081
Email: wangfan512@126.com

Abstract

Background: Health education and promotion are recognized as effective strategies for fostering healthy ageing, reducing the disease burden, and addressing health inequalities, particularly when delivered through digital media. Primary care workers are often regarded as the key providers of these interventions. Despite the strong practical significance and substantial individual demand, the use of digital media for delivering health promotion practices was not widespread in China. One of the main challenges identified is the providers' inadequate capacities. However, little is known about the digital capacity for health promotion among primary care workers.

Objective: This study aimed to investigate the levels of digital capacity for health promotion and its associated factors among community health workers.

Methods: A total of 1346 community health workers were recruited from across 47 communities in Shanghai, China, through cluster-stratified random sampling. The digital capacity for health promotion was measured using the revised version of the Digital Capabilities Framework. Web-based questionnaires were distributed to collect data from March 20 to March 29, 2024. Data were analyzed using descriptive statistics, independent t tests, one-way ANOVA, and linear hierarchical regression using Stata MP (version 17.0; StataCorp).

Results: We included 1199 participants. Among them, 47.5% (570/1199) had high digital media use for more than 19.6 hours per week, whereas 31.8% (381/1199) demonstrated high digital media trust. The average level of digital capacity for health promotion was 16.71 (SD 2.94) out of 25 points. Demographics, digital media usage–related characteristics, perceived usefulness and usability, attitudes, and behaviors were significant predictors of the capacities, explaining 44.4% of the total variance. Master's degree or above (β=.077; P=.013), perceived usability (β=.235; P<.001), attitudes toward digital media health promotion (β=.095; P=.002), and past digital media health promotion practices (β=.377; P<.001) had significantly positive associations with digital capacities for health promotion. However, senior (β=–.076; P=.008) or median (β=–.074; P=.01) titles had a significant negative association with capacity levels.

Conclusions: A digitally capable workforce is required for primary health care systems to take full advantage of digital media health promotion. Therefore, solutions are necessary to achieve enhanced capacities among health professionals, including public health policy making, community empowerment, and individual practices.

(J Med Internet Res 2024;26:e63054) doi:10.2196/63054

XSL•FO
RenderX

KEYWORDS

health promotion; digital capacity; primary care workers; Digital Capabilities Framework; online survey

Introduction

As various countries worldwide are committed to fostering healthy ageing, reducing the disease burden, and narrowing health inequalities, health education and promotion have been identified as effective ways to achieve these health targets [1,2]. In both developed and developing countries, health promotion equips the public with the knowledge about the determinants of health and skills to cope with risk factors [3]. It plays a vital role in controlling the outbreaks of infectious diseases, curbing the rising number of chronic diseases, fostering healthy lifestyles, and ultimately contributing to improved health outcomes [4]. Furthermore, with the development of health-related digitization, social media platforms such as Instagram and Facebook have become some of the most important sources of health information acquisition, particularly among the younger population [5,6]. This situation lays a solid foundation for carrying out multiple health promotion activities through digital media, as it offers the opportunity to reach specific target populations, lowers the costs of implementation, and customizes interventions to the needs of different individuals [7].

Medical professionals play a paramount role in conducting health promotion activities [8]. Specifically, primary care workers are urged to engage in health promotion as part of their daily routines [2]. Globally, primary care professionals, such as general practitioners and nurses, serve as the gatekeepers of the residents' health [9]. One of their imperative responsibilities entails the primary prevention of diseases before the first occurrence of undesirable symptoms, as opposed to secondary or tertiary prevention [7]. Even if a disease occurs, primary care workers can mobilize community resources to enhance community participation and reduce exposure to risk factors, thus enabling the efficient use of resources and fostering sustainability and resilience of the entire health system, simultaneously achieving cost savings and health gains [10]. Furthermore, primary care workers tend to have the most frequent contact with residents and obtain their deepest trust, which provides valuable insights that enable them to construct content that is relevant and accessible to the demands of the intended receivers [11,12].

China is one of the countries that have enacted to accelerate the implementation of health promotion through digital media. From the receivers' perspective, individuals in China have demonstrated a growing need to acquire health knowledge through digital media [13]. This heightened health awareness creates an extensive need for preventive care and promotion services to address potential risk factors [14]. Studies have demonstrated that 67.13% of older adults have health promotion demands beyond the general level [15]. Furthermore, high internet prevalence in China has precipitated a surge of online health information–seeking behaviors [16]. The COVID-19 pandemic has further exacerbated this situation [6,13]. Current content and traditional face-to-face education have to some extent failed to meet the increasing need for health information via digital media [17]. From the providers' perspective, relevant policies have been issued to support digitalized health promotion, particularly in primary health care (PHC) settings. In 2016, the State Council of the People's Republic of China proposed the Healthy China 2030 Initiative, a critical guideline steering China's health system development toward 2030, and incorporated health education into its strategies [18]. Since 2018, the government has provided funding to promote health promotion activities under the family physician contract, as family physicians are responsible for residents' health maintenance and the effective delivery of public health education [5]. Subsequently, in 2022, supporting guidelines were proposed to increase the supply of high-quality health promotion activities through various media platforms [19].

However, despite the strong political wills and substantial individual demands, PHC professionals in China allocate insufficient time to health promotion practices, particularly those disseminated through digital channels [20,21]. Consequently, the prevalence of digitalized health promotion remains limited. A notable impediment to the adoption is providers' lack of knowledge, awareness, and skills in using various digital media [22]. Several studies have investigated health care professionals' promotion capacities, revealing that capacity levels may be affected by sex, age, educational level, occupation, training frequency, and familiarity with the policy [5,8]. Another strand of research concentrated on the measurement of digital capacities among health care workers. These studies identified potential influencing factors including sociocultural context [22], organizational support [23,24], individual motivation and attitudes [25,26], and the perceived usefulness and usability of digital technology [27,28]. Despite the abundant literature, there is a scarcity of similar studies on these 2 topics conducted in primary care settings in China. Moreover, the digital capacity for health promotion among health care workers, the integration of these 2 strands of literature, has not been systematically explored.

This study aims to investigate the digital capacity for health promotion in a representative sample of community health workers in Shanghai, China, and to identify the influencing factors. This study contributes to the literature on digital capacity and health promotion capacity among health care professionals. Furthermore, gaining knowledge about the status quo and correlates can suggest feasible strategies for continuously improving the quality of health promotion services, addressing health inequities, and building resilient health systems.

Methods

Research Design

This cross-sectional study aims to examine the current status of digital capacity for health promotion among community health workers based on the Digital Capabilities Framework (DCF) created by Health Education England and identify associated factors. To achieve this study objective, a survey was conducted. The survey included an anonymous web-based

XSL•FO
RenderX

questionnaire to be completed by PHC providers in Shanghai, China.

Recruitment

The study involved health care workers in community health centers located in the Pudong New Area, one of the largest and most demographically representative districts of Shanghai, China. As of 2023, Pudong New Area spanned an area of 1210 km^2 and housed a population of 5.81 million residents, constituting approximately one-fourth of Shanghai's total population [29]. Participants were recruited through cluster-stratified random sampling based on the predetermined inclusion criteria. All 47 communities within the Pudong New Area were chosen in the sampling process. Health care providers were stratified into 4 distinct layers according to their occupations (clinicians, public health physicians, nurses, and medical technicians). Following this categorization, the participants were randomly selected from the 4 layers, ensuring that each facility contributed more than 5 individuals from each occupational category. The inclusion criteria were as follows: (1) health workers providing health services in community health centers in the Pudong New Area of Shanghai, (2) able to understand the research purpose and voluntarily participate in the research, and (3) able to read and understand in Chinese.

A web-based questionnaire was developed and disseminated through the working groups in WeChat. It was completed by community health workers voluntarily using the Questionnaire Star platform from March 20 to March 29, 2024. A total of 1346 community health workers completed the questionnaire. To ensure data quality, we excluded incomplete or noncompliant questionnaires for quality control and concluded with 1199 participants for the final analysis, resulting in a response rate of 89.08% (1199/1346) (Figure 1).

Figure 1. Flowchart for the sample inclusion.

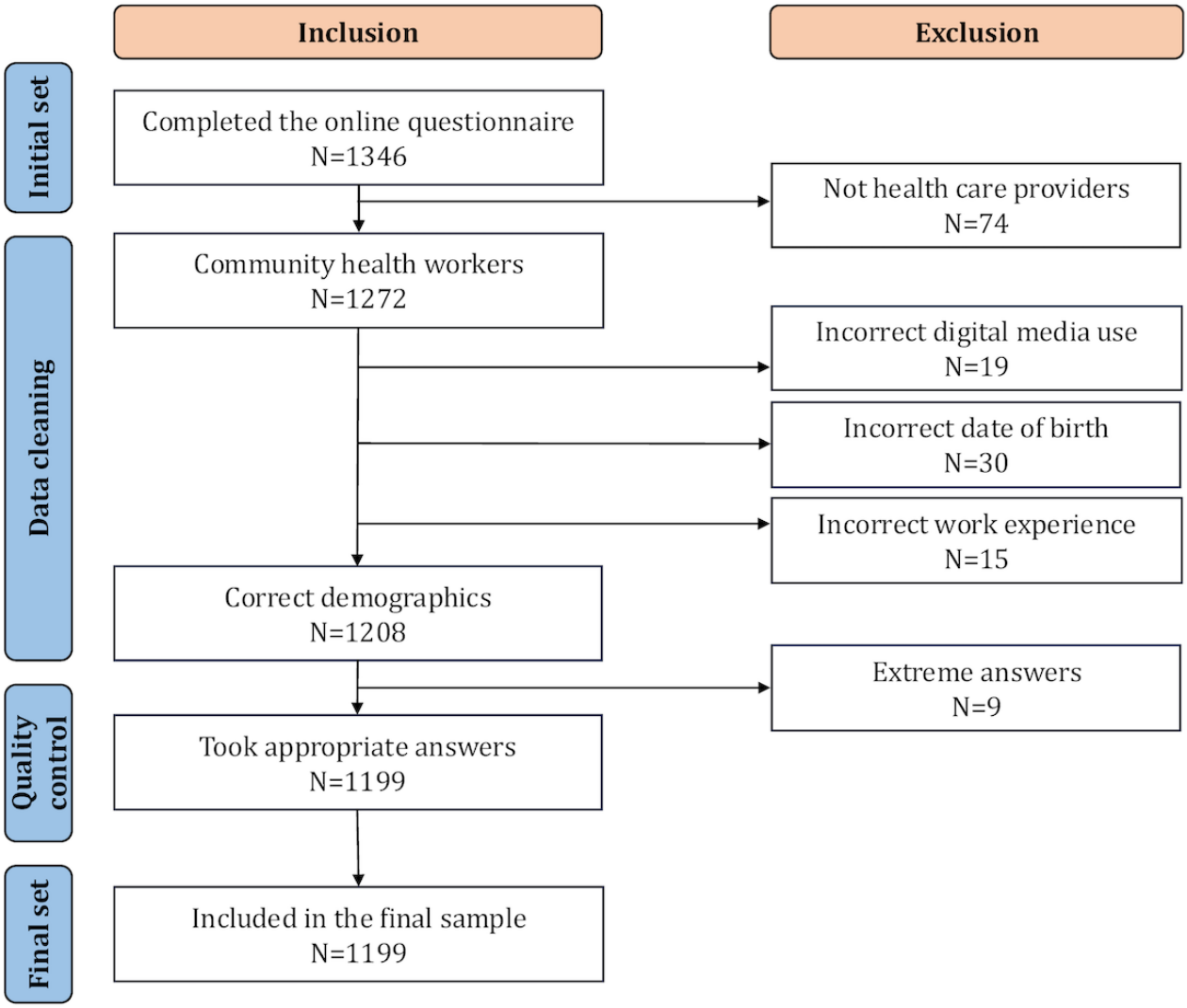

Ethical Considerations

This study received approval from the ethics review committee of the Shanghai Pudong New Area Center for Disease Control and Prevention (PDCDCLL-20240304-001). Participation was voluntary. Participants were informed of the purpose of the study and the possibility to opt out at any time. As a web-based survey, the completion and return of the self-administered questionnaire were regarded as providing informed consent. Measures were taken to ensure privacy and confidentiality by anonymizing the data, thus excluding any personal identifiers from the analysis. No compensation was offered.

XSL•FO
RenderX

Measures

A web-based questionnaire was designed to measure the capacity levels and potential associates. Pilot surveys were conducted among 145 community health workers and health promotion practitioners to confirm the effectiveness, clarity, and readability of the questionnaire. The results were used for revision and optimization of the questionnaire design but were not included in the final analysis. After slight modifications, the survey comprised 6 sections.

Sociodemographics

The following participant demographics were collected: sex, age, seniority, educational level, occupation, technical title, digital media usage, and digital media trust. We asked respondents to identify how many hours they spend on various digital media platforms per week as a measurement for usage. Regarding digital media trust, respondents rated how they trust digital media sources of information on a 5-point Likert scale ranging from 1 (totally distrust) to 5 (totally trust) [30,31]. Following previous studies [27,32], individuals who achieved a score of the mean value or greater were defined as having "high digital media usage or trust". In contrast, those who scored below average were regarded as having "low digital media usage or trust".

Digital Capacity for Health Promotion

The digital capacity for health promotion was measured according to the DCF created by Health Education England [33] and previous studies adopting this framework [34]. It incorporates the following 5 items: communication, collaboration, and participation; teaching, learning, and self-development; information, data, and content literacy; creation, innovation, and research; and technical proficiency. Each of the 5 items was self-rated by the participants using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The total score ranged from 5 to 25, with higher scores indicating higher levels of participants' competencies. For evaluating the validity and reliability, we used Cronbach α for internal consistency and Kaiser-Meyer-Olkin (KMO) value for construct validity for each section. Cronbach α value was 0.94 and the KMO value was 0.90 in the final analysis.

The Usefulness of Digital Media Health Promotion

Previous theories have defined perceived usefulness as whether technology has practical value in the completion of tasks or in the attainment of goals, which has been identified as an important aspect affecting digital capacity [27,35]. Thus, a 5-point Likert scale with 4 items (from 1: strongly disagree to 5: strongly agree) was designed for the measurement. The final Cronbach α value was 0.89 and the final KMO value was 0.79 for this domain.

The Usability of Digital Media Health Promotion

Perceived usability is defined as whether learning and adopting a new technology is convenient, simple, and devoid of great difficulties, and has also been demonstrated to significantly influence digital capacity [36-38]. Two items on a 5-point Likert scale (from 1: strongly disagree to 5: strongly agree) were used to measure participants' perceived usability. Cronbach α value was 0.82.

Attitudes Toward Digital Media Health Promotion

The participants' subjective positive or negative attitudes toward the use of digital media for health promotion were assessed using 4 items measured on a 5-point Likert scale (from 1: strongly disagree to 5: strongly agree) [28,39]. Cronbach α value was 0.93, and the final set had a KMO value of 0.85.

Digital Media Health Promotion Behaviors

Past behaviors of carrying out health promotion via digital media were also measured, including the participants' learning process, completed practice, and encouragement to colleagues. These 3 items are measured on a 5-point Likert scale (from 1: strongly disagree to 5: strongly agree). Cronbach α value was 0.90, and the KMO value was 0.75 in the final analysis.

Statistical Analysis

The collected data were analyzed using Microsoft Excel (Microsoft Inc) and Stata MP (version 17.0; StataCorp). Descriptive statistics such as means and SDs were used for continuous variables, and frequencies and proportions were used for categorical variables. We conducted independent *t* tests and one-way ANOVAs to compare the differences in digital capacities concerning general characteristics, as homoscedasticity was ensured by Bartlett equal variances tests. Where ANOVAs showed significant differences, we performed Bonferroni tests as post hoc procedures. Linear hierarchical regression analysis was performed to examine the factors influencing digital capacities. Each categorical variable was dummy-coded in the regression model. The potential multicollinearity of variables was assessed using variance inflation factors. All the statistical tests were 2-sided, with P value of <.05 considered statistically significant.

Results

Sample Characteristics

The sociodemographic characteristics of the participants are shown in Table 1. A total of 1199 participants were included, of whom 81.6% (978/1199) were female. The average age was 39.34 (SD 7.70) years, and the average working experience was 17.20 (SD 9.03) years. An overwhelming majority of participants (950/1199, 79.2%) were university graduates with a bachelor's degree, and 58.4% (700/1199) held median technical jobs. Among the participants, 27.5% (330/1199) were clinicians, 28.4% (341/1199) nurses, 24.7% (296/1199) public health physicians, and 19.4% (232/1199) medical technicians. Regarding digital media use and trust, 47.5% (570/1199) had high digital media use for more than 19.6 hours per week, whereas 31.8% (381/1199) demonstrated a high level of digital media trust.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of community health workers (N=1199).

Categories	Values
Sex, n (%)	
Male	221 (18.43)
Female	978 (81.57)
Age (years), n (%)	
≤35	414 (34.53)
36-45	492 (41.03)
≥46	293 (24.44)
Seniority (years), n (%)	
≤10	335 (27.94)
11-20	427 (35.61)
≥21	437 (36.45)
Educational level, n (%)	
Associate's degree or below	158 (13.18)
Bachelor's degree	950 (79.23)
Master's degree or above	91 (7.59)
Occupation, n (%)	
Clinician	330 (27.52)
Public health physician	296 (24.69)
Nurse	341 (28.44)
Medical technician	232 (19.35)
Technical title, n (%)	
Senior	106 (8.84)
Median	700 (58.38)
Primary	393 (32.78)
Digital media use, n (%)	
High	570 (47.54)
Low	629 (52.46)
Digital media trust, n (%)	
High	381 (31.78)
Low	818 (68.22)

Descriptive Statistics of the Study Variables

The results of descriptive statistics of the study variables are shown in Table 2. The participants' average score of digital capacity for health promotion was 16.71 (SD 2.94) out of 25 points. Information, data, and content literacies had the highest value of 3.39 (SD 0.63), followed by teaching, learning, and self-development with a slightly lower value of 3.38 (SD 0.63). Creation, innovation, and research had the lowest score of 3.30 (SD 0.69). The mean score for the usefulness of digital media use was 3.49 (SD 0.61). Participants exhibited the highest agreement with the statement that digital media health promotion fostered the development of personal capacity (mean 3.59, SD 0.68), while lowest to its role related to job promotion. Regarding the usability of digital media health promotion, the score for this domain was 3.21 (SD 0.65) on average. Participants thought that it was easier to adapt to the potential changes than to acquire the related knowledge and skills. In addition, the domains of attitudes and behaviors had average scores of 3.57 (SD 0.59) and 3.29 (SD 0.64), respectively.

XSL•FO
RenderX

Table 2. Study variables among participants (N=1199).

Item	Mean (SD)	Range
Digital capacity for health promotion		
Communication, collaboration, and participation	3.32 (0.66)	1-5
Teaching, learning, and self-development	3.38 (0.63)	1-5
Information, data, and content literacy	3.39 (0.63)	1-5
Creation, innovation, and research	3.30 (0.69)	1-5
Technical proficiency	3.32 (0.65)	1-5
Total scores	16.71 (2.94)	5-25
Usefulness of digital media health promotion		
Enhancement of professional value	3.57 (0.67)	1-5
Development of personal capacity	3.59 (0.68)	1-5
Promotion of technical title	3.35 (0.78)	1-5
Assistance in carrying out the work	3.47 (0.68)	1-5
Average score	3.49 (0.61)	1-5
Usability of digital media health promotion		
Easy to acquire related knowledge and skills	3.16 (0.72)	1-5
Easy to adapt to the changes in the work and life	3.26 (0.69)	1-5
Average score	3.21 (0.65)	1-5
Attitude toward digital media health promotion		
Meaningful to work on health promotion through digital media	3.63 (0.62)	1-5
A great way to work on health promotion through digital media	3.60 (0.64)	1-5
Not opposed to work on health promotion through digital media	3.62 (0.64)	1-5
Relevant to my work	3.44 (0.67)	1-5
Average score	3.57 (0.59)	1-5
Digital media health promotion behavior		
Learnt about it through various channels	3.33 (0.67)	1-5
Carried out in the scope of own work	3.25 (0.72)	1-5
Encouraged colleagues to carry on	3.30 (0.70)	1-5
Average score	3.29 (0.64)	1-5

Differences in Digital Capacity for Health Promotion According to General Characteristics

Differences in digital capacity for health promotion according to the general characteristics of the participants are reported in Table 3. The results revealed significant differences according to age ($P<.001$), seniority ($P<.001$), occupation ($P=.002$), technical title ($P<.001$), and digital media trust ($P<.001$).

XSL•FO
RenderX

Table 3. Digital capacity for health promotion according to general characteristics of the participants (N=1199).

Categories	Mean (SD)	*t*/*F* test (*df*)	*P* value
Sex		0.19 (1, 1197)	.66
Male	16.79 (2.99)		
Female	16.69 (2.92)		
Age (years)		22.41 (2, 1196)	<.001
≤35	17.42 (2.86)		
36-45	16.53 (2.88)		
≥46	16.00 (2.93)		
Seniority (years)		18.51 (2, 1196)	<.001
≤10	17.41 (2.96)		
11-20	16.75 (2.82)		
≥21	16.13 (2.91)		
Educational level		1.42 (2, 1196)	.24
Associate's degree or below	16.54 (2.83)		
Bachelor's degree	16.69 (2.92)		
Master's degree or above	17.18 (3.27)		
Occupation		5.04 (3, 1195)	.002
Clinician	16.20 (3.07)		
Public health physician	16.76 (3.07)		
Nurse	16.94 (2.79)		
Medical technician	17.02 (2.69)		
Technical title		13.88 (2, 1196)	<.001
Senior	16.18 (3.33)		
Median	16.44 (2.81)		
Primary	17.33 (2.94)		
Digital media use		0.31 (1, 1197)	.58
High	16.66 (3.00)		
Low	16.75 (2.88)		
Digital media trust		56.55 (1, 1197)	<.001
High	17.62 (2.81)		
Low	16.28 (2.90)		

Post hoc analyses using Bonferroni tests demonstrated that participant community health care workers aged 35 years or younger scored significantly higher than those aged between 36 and 45 years, which scored higher than those aged 46 years or older (all $P<.05$; Multimedia Appendix 1). Regarding seniority, participants who had worked for less than 10 years scored better (all $P<.05$). Nurses ($P=.006$) and medical technicians ($P=.006$) had higher digital capacities than clinicians. Furthermore, participants with a primary technical title scored better in their abilities than those with a median ($P<.001$) or senior title ($P=.001$).

Influencing Factors of Digital Capacity for Health Promotion

Table 4 shows the results of the linear hierarchical regression. Demographics and digital media usage-related characteristics alone explained approximately 9.2% of the variance in digital capacity for health promotion, reaching statistical significance ($F_{14,1184}=8.619$; $P<.001$). When including perceived usefulness and usability variables in model 3, the proportion of explained variance increased by 24.3% to an overall R^2 value of approximately 33.6% ($F_{16,1182}=37.299$; $P<.001$). Further including attitudes toward digital media health promotion into our model led to an additional increase in R^2 of 3.8%. Finally, the inclusion of health promotion behaviors led to an increase in R^2 by 7.1%, with an overall R^2 of 44.4% ($F_{18,1180}=52.347$;

XSL•FO
RenderX

P<.001). Thus, our model significantly improved at each stage of the hierarchical process.

Table 4. Model parameters of the linear hierarchical regression model for digital capacity for health promotion (N=1199)[a].

Model	Blocks	R^2	*F* test (*df*)	*P* value	$\triangle R^2$	$\triangle F$ test (*df*)	$\triangle P$ value
1	Demographics	0.054	5.594 (12, 1186)	<.001	0.054	5.594 (12, 1186)	<.001
2	Demographics + Digital media usage–related characteristics	0.092	8.619 (14, 1184)	<.001	0.039	25.387 (2, 1184)	<.001
3	Demographics + Digital media usage–related characteristics + Perceived usefulness and usability	0.336	37.299 (16, 1182)	<.001	0.243	216.133 (2, 1182)	<.001
4	Demographics + Digital media usage–related characteristics + Perceived usefulness and usability + Attitudes	0.373	41.334 (17, 1181)	<.001	0.038	70.702 (1, 1181)	<.001
5	Demographics + Digital media usage–related characteristics + Perceived usefulness and usability + Attitudes + Behaviors	0.444	52.337 (18, 1180)	<.001	0.071	150.580 (1, 1180)	<.001

[a]Variables included in the different hierarchical regression models are as follows: model 1 (sex, age group dummy coded, seniority group dummy coded, educational-level dummy coded, occupation dummy coded, and technical title dummy coded); model 2 (model 1 variables, digital media use group dummy coded, and digital media trust group dummy coded); model 3 (model 2 variables, perceived usefulness of digital media health promotion, and perceived usability of digital media health promotion); model 4 (model 3 variables, attitudes toward digital media health promotion); and model 5 (model 4 variables, digital media health promotion behaviors).

Table 5 reports the regression coefficients of the final hierarchical regression model. Of the control variables in blocks of demographics, those aged between 36 and 45 years had comparatively lower competencies than those aged 35 years or younger (*b*=–0.438, SE *B*=0.219, β=–.073; *P*=.046). Having a master's degree or above (*b*=0.850, SE *B*=0.342, β=.077; *P*=.01) increased the capacity score to 0.085 in comparison with associate's degree or below, whereas having a senior or median title decreased the score for –0.45 (*b*=–0.451, SE *B*=0.170, β=–.076; *P*=.008) and –0.76 (*b*=–0.761, SE *B*=0.301, β=–.074; *P*=.01) in comparison with a primary title, respectively. Furthermore, perceived usability (*b*=1.064, SE *B*=0.131, β=.235; *P*<.001), attitudes toward digital media health promotion (*b*=0.476, SE *B*=0.156, β=.095; *P*=.002), and digital media health promotion behaviors (*b*=1.727, SE *B*=0.141, β=.377; *P*<.001) were positively and significantly associated with higher scores in digital capacity for health promotion among community health workers.

XSL•FO
RenderX

Table 5. Regression coefficients for the final linear hierarchical regression model predicting the digital capacity for health promotion (N=1199)[a].

	b	SE	β	*P* value	95% CI
Sex (vs male)					
Female	-0.297	0.174	–.039	.088	–0.638 to 0.045
Age (years; vs ≤35 years)					
36-45	–0.438	0.219	–.073	.046	–0.867 to –0.009
≥46	–0.589	0.315	–.086	.061	–1.207 to 0.028
Seniority (years; vs ≤10 years)					
11-20	0.405	0.227	.066	.075	–0.041 to 0.851
≥21	0.471	0.306	.077	.12	–0.128 to 1.071
Educational level (vs associate's degree or below)					
Bachelor's degree	0.223	0.215	.031	.30	–0.198 to 0.645
Master's degree or above	0.850	0.342	.077	.013	0.178 to 1.521
Occupation (vs clinician)					
Public health physician	–0.201	0.195	–.030	.30	–0.584 to 0.182
Nurse	0.137	0.201	.021	.50	–0.258 to 0.532
Medical technician	0.141	0.206	.019	.49	–0.263 to 0.544
Technical title (vs primary)					
Median	–0.451	0.170	–.076	.008	–0.784 to –0.118
Senior	–0.761	0.301	–.074	.012	–1.352 to –0.171
Digital media use (vs low)					
High	–0.174	0.131	–.030	.19	–0.431 to 0.083
Digital media trust (vs low)					
High	0.205	0.146	.033	.16	–0.081 to 0.491
Perceived usefulness	0.173	0.143	.036	.23	–0.107 to 0.452
Perceived usability	1.064	0.131	.235	.23	–0.107 to 0.452
Attitudes toward digital media health promotion	0.476	0.156	.095	.002	0.169 to 0.783
Digital media health promotion behavior	1.727	0.141	.377	<.001	1.451 to 2.003
Constant	5.644	0.549	N/A[b]	<.001	4.567 to 6.721

[a]R^2=0.444, adjusted R^2=0.436, F=52.35, and P<.001.

[b]N/A: not applicable.

Discussion

Principal Findings in Relation to Previous Studies

This study examined the current levels of digital capacity for health promotion among community health workers in China and identified its influencing factors. Our findings suggested that the participants' average level of digital capacity for health promotion was 16.71 (SD 2.94) out of 25 points. Demographics, digital media usage–related characteristics, perceived usefulness and usability, attitudes, and behaviors were significant predictors of the capacity. Those with higher educational levels, perceived usability, more positive attitudes, and more frequent behaviors had significantly higher digital capacities for health promotion. However, the technical titles were negatively associated with capacity scores. In this digital era, strengthening the digital competencies of health care providers contributes to increasing access to health information, thereby improving the quality of practice, and raising public awareness of healthy lifestyles [26]. Furthermore, it has also demonstrated potential positive impacts on the health equality and resilience of the entire health system when facing public health emergencies and increasing chronic disease burdens.

Based on the success of the prior literature, our study specifically focused on the capacity levels for health promotion among health care providers. Previous studies have established different terms to refer to the concept of digital capacity, such as eHealth literacy and digital health competence, and have formulated operationalized frameworks for measurement and comparison [40,41]. However, few studies assessed digital capacity in the health promotion process, an important issue gaining increasingly wider popularity in the digital era [42-44]. Moreover, studies have been conducted in PHC settings to

XSL•FO
RenderX

investigate providers' skill development in health education and health promotion activities but were predominantly focused on specific programs, notably diabetes management, mental health care, and cancer prevention [17,45,46]. Evidence regarding the comprehensive assessment of PHC workers' capacity levels is scarce. In this context, we modified the DCF to incorporate the health promotion process into the measurement scale and observed that the average score among community health workers was 16.71 (SD 2.94) out of 25 points. Information, data, and content literacy was the highest item scoring 3.39 (SD 0.63), whereas creation, innovation, and research was the lowest item scoring 3.30 (SD 0.69). When compared with similar studies measuring health care providers' digital capacity or health promotion capacity, our study reported a relatively lower score when focusing specifically on the integration of these 2 capacities [25,27,47]. This was evidenced by the achieved scoring of only 66.9% (16.71/25), significantly lower than similar studies in China that achieved close to or above 80% of total scores in different scales [5,48]. Furthermore, studies in other countries also confirmed the relatively high level of wished scoring, with most studies conducted in resource-limited regions with limited internet penetration, unlike Shanghai [37,49,50]. This comparison indicates significant room for improvement in the digital capacity of community health workers when conducting health promotion activities in China.

Regarding the determinants, our results indicated that 5 blocks, namely, demographics, digital media usage–related characteristics, perceived usefulness and usability, attitudes, and behaviors, were all significant predictors. These findings are consistent with those of previous studies investigating factors associated with digital health literacy conducted elsewhere around the globe [50-52]. In our final regression model, attainment of a master's degree or above (β=.077; *P*=.01) was correlated with significantly greater capacity levels, possibly owning to their reception of past formal training curriculums and proficiency in simplifying intricate health concepts [53]. Conversely, having a senior (β=–.076; *P*=.008) or median title (β=–.074; *P*=.01) was significantly negatively associated with the score. This could possibly be elucidated by recent considerations of health education practices in job promotions among community health workers in China, motivating workers with primary titles to enhance their skills [54]. Furthermore, the propensity of the younger generation to engage with digital media in a proactive way might also contribute to this trend [55]. In addition to sociodemographics, perceived usability (β=.235; *P*<.001) was also a significant correlate, since health care providers tend to be more motivated and enthusiastic when having mastered the skills and knowledge of digital media health promotion [27]. They are more confident in performing promotional practices using digital tools, thus attaining a more desirable capacity level. Furthermore, digital capacity was influenced by current attitudes (β=.095; *P*=.002) and existing behaviors (β=.377; *P*<.001). This could be attributed to the fact that positive attitudes toward digital media fostered health workers' involvement in technology and their search for digital information [56]. By adopting digital media to source reliable data, synthesize understandable content, and disseminate it to various patients during the process of health popularization, their competencies can be progressively cultivated and strengthened.

Implications

These findings have valuable implications for public health policy making, community empowerment, and individual practices. These significant predictors can serve as important indicators for delivering targeted interventions to strengthen the PHC providers' digital capacity for health promotion in China. Specifically, a digitally capable workforce in PHC settings ensures equitable, accessible, and high-quality health promotion programs delivered through various digital media channels [10]. This improves the coverage of the most vulnerable or disadvantaged population groups, thereby contributing to a reduction in health inequalities [57]. Therefore, solutions aligned across policy-level interventions, community-level inputs, and individual-level efforts are urgently required. The results demonstrated that positive attitudes and previous engagement in digital health promotion were crucial facilitators fostered by the strategies at all 3 levels. Lower educational levels and perceived usability could be potential barriers, requiring measures at the community level.

At the policy level, relevant policies can be enacted to encourage health workers to use digital tools when carrying out health promotion activities. Strategies targeting economic incentives or other rewards can foster health workers' attitudes toward digital media health promotion, thus obtaining more desirable capacity levels [57]. Moreover, it is worth noting that recent years have witnessed the expansion of nurses' roles worldwide, particularly in health promotion responsibilities aimed at preventing chronic diseases and addressing unhealthy lifestyles [58]. The central government should thus commit to increasing the supply of nurses and enhancing their competencies by issuing relevant policies and guidelines. At the community level, resources should be invested to provide the necessary training programs to the workforce. Community health centers should develop purposeful training plans to familiarize health care workers with the use of digital tools and related theories, thereby obtaining higher usability levels [59]. Furthermore, a supportive organizational environment is crucial for translating policies into management and operational mechanisms. Possible approaches include advocating for benefits, addressing challenges encountered, and fostering implementation. By encouraging PHC workers to participate in digital health promotion activities, the institutional service capacity can also be strengthened and empowered [60]. At the individual level, community health workers should maintain positive attitudes and proactively engage in digital media health promotion activities to gain elevated capacities through multiple practices. They would in turn benefit from these policies and community-based measures by gaining a better understanding of the significance of health education activities and take advantage of various training programs to fill their capacity gaps.

Limitations and Directions for Future Research

Our study has several limitations. First, all of the study variables were self-reported by the participants through web-based questionnaires, with the possibility of reporting bias occurring

XSL•FO
RenderX

during the survey. Future studies can benefit from objectively evaluating the digital capacity of community health workers. Second, in this study, we primarily concentrated on the effects of individual factors fully considering broader contextual factors at the community and social levels. As previous studies have confirmed the connection between health workers' digital capacities and various cultural, economic, and political characteristics, more efforts should be invested in assessing factors at different levels simultaneously. Third, this study included only health workers in community health centers located in Shanghai, a tier-1 megacity in China. Thus, the results cannot be generalized to the entire population of health care providers nationwide. Future studies should be conducted with a group of participants representative of the nationwide health workforce nationwide to enhance the generalizability of the findings. Finally, this study examined the digital capacity of community health workers, without investigating further impacts on patients. Further research is required to evaluate the effects of providers' elevated capacities regarding improving patients' responsiveness and health outcomes.

Conclusions

Health promotion through digital media, particularly in primary care settings, has significant implications for improving the quality of health promotion services, addressing the increasing disease burden, and fostering the resilience of the health system. However, its implementation is still not extensive in primary care facilities in China, partly owing to the lack of relevant capacity among health workers. Our study found that the average level of digital capacity for health promotion among community health workers was 16.71 (SD 2.94) out of 25 points. Demographics, digital media usage–related characteristics, perceived usefulness and usability, attitudes, and behaviors were significant predictors of the capacity, explaining 44.4% of the total variance. Future interventions are urgently required at the policy, community, and individual levels to collectively cultivate primary care professionals with elevated capacities.

Acknowledgments

The authors are grateful to the Shanghai Center for Disease Control and Prevention for helping to carry out field research. They also thank Fudan-Sinarmas Thinktank Program for the support. This work was funded by the Science and Technology Development Fund of Pudong New Area (PKJ2023-Y74). The funder of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the paper.

Authors' Contributions

FW, HR, YW, SX, QL, and SS conceptualized the study; FW, HR, and SX collected data; YW, SY, and SS conducted data analysis; YW, TM, FW, HR, SX, and SY completed the first draft; FW, HR, YW, and SX revised the manuscript; and FW and HR reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Post hoc analyses using Bonferroni tests.
[DOCX File , 16 KB - jmir_v26i1e63054_app1.docx]

References

1. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023;25:e43542 [FREE Full text] [doi: 10.2196/43542] [Medline: 36951896]
2. Mahmud AJ, Olander E, Eriksén S, Haglund BJ. Health communication in primary health care -a case study of ICT development for health promotion. BMC Med Inform Decis Mak 2013;13:17 [FREE Full text] [doi: 10.1186/1472-6947-13-17] [Medline: 23363566]
3. Ghalavand H, Panahi S, Sedghi S. Opportunities and challenges of social media for health knowledge management: a narrative review. J Educ Health Promot 2020;9:144 [FREE Full text] [doi: 10.4103/jehp.jehp_754_19] [Medline: 32766329]
4. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]
5. Zhou C, Tan F, Lai S, Chen J, Cai Q, Yin X, et al. Health promotion capacity among Chinese healthcare professionals and its influence on preventive health service practices. J Multidiscip Healthc 2022;15:343-352 [FREE Full text] [doi: 10.2147/JMDH.S349449] [Medline: 35237040]
6. Bian D, Shi Y, Tang W, Li D, Han K, Shi C, et al. The influencing factors of nutrition and diet health knowledge dissemination using the WeChat official account in health promotion. Front Public Health 2021;9:775729 [FREE Full text] [doi: 10.3389/fpubh.2021.775729] [Medline: 34900917]

XSL•FO
RenderX

7. Stark AL, Geukes C, Dockweiler C. Digital health promotion and prevention in settings: scoping review. J Med Internet Res 2022;24(1):e21063 [FREE Full text] [doi: 10.2196/21063] [Medline: 35089140]
8. Luquis RR, Paz HL. Attitudes about and practices of health promotion and prevention among primary care providers. Health Promot Pract 2015;16(5):745-755. [doi: 10.1177/1524839914561516] [Medline: 25445979]
9. Jimenez G, Spinazze P, Matchar D, Koh Choon Huat G, van der Kleij RMJJ, Chavannes NH, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
10. Shilton T, Barry MM. The critical role of health promotion for effective universal health coverage. Glob Health Promot 2022;29(1):92-95. [doi: 10.1177/1757975920984217] [Medline: 33535893]
11. Eghtesadi M, Florea A. Facebook, instagram, reddit and TikTok: a proposal for health authorities to integrate popular social media platforms in contingency planning amid a global pandemic outbreak. Can J Public Health 2020;111(3):389-391 [FREE Full text] [doi: 10.17269/s41997-020-00343-0] [Medline: 32519085]
12. Li X, Lu J, Hu S, Cheng KK, De Maeseneer J, Meng Q, et al. The primary health-care system in China. Lancet 2017;390(10112):2584-2594. [doi: 10.1016/S0140-6736(17)33109-4] [Medline: 29231837]
13. Zhao X, Basnyat I. Online information and support seeking during COVID-19 lockdown in Wuhan: implications for health promotion. Health Promot Int 2022;37(3):daac057 [FREE Full text] [doi: 10.1093/heapro/daac057] [Medline: 35788308]
14. Zeng W, Wang C, Chen H, Tong B, Li D, Zou Z, et al. Health status and public health education for internal older migrants in China: evidence from a nationally representative survey. Front Public Health 2022;10:937361 [FREE Full text] [doi: 10.3389/fpubh.2022.937361] [Medline: 35937236]
15. Ren W, Ma X, Tarimo CS, Xing Y, Lv X, Liang Z. A study on the current state and equity level of the health promotion service demands among older adults in China. Int J Equity Health 2023;22(1):67 [FREE Full text] [doi: 10.1186/s12939-023-01882-x] [Medline: 37055791]
16. Sun M, Yang L, Chen W, Luo H, Zheng K, Zhang Y, et al. Current status of official WeChat accounts for public health education. J Public Health (Oxf) 2021;43(3):618-624. [doi: 10.1093/pubmed/fdz163] [Medline: 31974552]
17. McInnes S, Halcomb E, Ashley C, Kean A, Moxham L, Patterson C. An integrative review of primary health care nurses' mental health knowledge gaps and learning needs. Collegian 2022;29(4):540-548 [FREE Full text] [doi: 10.1016/j.colegn.2021.12.005] [Medline: 35043041]
18. State Council. Healthy China 2030 Initiative. 2016. URL: https://www.gov.cn/zhengce/2016-10/25/content_5124174.htm?eqid=ac00b02f000027f90000000464586447 [accessed 2024-05-20]
19. Health Commissions. Guidelines on the establishment of mechanisms for the publication and dissemination of health science knowledge through media platforms. 2022. URL: http://www.nhc.gov.cn/xcs/s7852/202107/1640094c16fa460abb1a804dc79a70ce/files/2bc3d0f02bd6460b90a57a7185e638ff.pdf [accessed 2024-05-20]
20. Zhong C, Luo Z, Liang C, Zhou M, Kuang L. An overview of general practitioner consultations in China: a direct observational study. Fam Pract 2020;37(5):682-688. [doi: 10.1093/fampra/cmaa039] [Medline: 32328659]
21. Wei Y, Wang F, Pan Z, Jin G, Wang D, Lu X, et al. Work content of general practitioners in Beijing, China: a multi-method Study. Front Public Health 2022;10:870224 [FREE Full text] [doi: 10.3389/fpubh.2022.870224] [Medline: 35570960]
22. Ma M, Li Y, Gao L, Xie Y, Zhang Y, Wang Y, et al. The need for digital health education among next-generation health workers in China: a cross-sectional survey on digital health education. BMC Med Educ 2023;23(1):541 [FREE Full text] [doi: 10.1186/s12909-023-04407-w] [Medline: 37525126]
23. Schierhout G, Praveen D, Patel B, Li Q, Mogulluru K, Ameer MA, et al. Why do strategies to strengthen primary health care succeed in some places and fail in others? Exploring local variation in the effectiveness of a community health worker managed digital health intervention in rural India. BMJ Glob Health 2021;6(5):e005003 [FREE Full text] [doi: 10.1136/bmjgh-2021-005003] [Medline: 34312146]
24. Birgel V, Decker L, Röding D, Walter U. Community capacity for prevention and health promotion: a scoping review on underlying domains and assessment methods. Syst Rev 2023;12(1):147 [FREE Full text] [doi: 10.1186/s13643-023-02314-1] [Medline: 37608344]
25. Cho H, Han K, Park BK. Associations of eHealth literacy with health-promoting behaviours among hospital nurses: a descriptive cross-sectional study. J Adv Nurs 2018;74(7):1618-1627. [doi: 10.1111/jan.13575] [Medline: 29575085]
26. Lee H, Nam HK, Zhao B, Jeong HR, Lim S, Chun A, et al. Analysis of digital capacity-related factors influencing health promotion participation and active aging of older adults residing in rural areas in South Korea: a structural equation model. Digit Health 2024;10:20552076241226958 [FREE Full text] [doi: 10.1177/20552076241226958] [Medline: 38269368]
27. Ahmed MH, Guadie HA, Ngusie HS, Teferi GH, Gullslett MK, Hailegebreal S, et al. Digital health literacy during the COVID-19 pandemic among health care providers in resource-limited settings: cross-sectional study. JMIR Nurs 2022;5(1):e39866 [FREE Full text] [doi: 10.2196/39866] [Medline: 36301671]
28. Kuek A, Hakkennes S. Healthcare staff digital literacy levels and their attitudes towards information systems. Health Informatics J 2020;26(1):592-612 [FREE Full text] [doi: 10.1177/1460458219839613] [Medline: 30983476]
29. People's Government of Pudong New Area. Statistical bulletin of national economic and social development of Pudong New Area of Shanghai in 2023. 2024. URL: https://www.pudong.gov.cn/zwgk/tjj_gkml_ywl_tjsj/2024/106/325948.html [accessed 2024-05-23]

XSL•FO
RenderX

30. Wei L, Huang Q. Retrospecting digital media use, negative emotions, and trust gaps during the COVID-19 pandemic in China: cross-sectional web-based survey. J Med Internet Res 2024;26:e49422 [FREE Full text] [doi: 10.2196/49422] [Medline: 38986127]
31. Tsai JY, Phua J, Pan S, Yang CC. Intergroup contact, COVID-19 news consumption, and the moderating role of digital media trust on prejudice toward asians in the United States: cross-sectional study. J Med Internet Res 2020;22(9):e22767 [FREE Full text] [doi: 10.2196/22767] [Medline: 32924948]
32. Assaye BT, Kassa M, Belachew M, Birhanu S, Worku A. Association of digital health literacy and information-seeking behaviors among physicians during COVID-19 in ethiopia: a cross-sectional study. Digit Health 2023;9:20552076231180436 [FREE Full text] [doi: 10.1177/20552076231180436] [Medline: 37312956]
33. Health Education England. A health and care digital capabilities framework. 2018. URL: https://hee.nhs.uk/sites/default/files/documents/Digital%20Literacy%20Capability%20Framework%202018.pdf [accessed 2024-05-16]
34. Crilly P, Fletcher J, Chandegra N, Khalefa A, Rouf SKM, Zein M, et al. Assessing the digital literacy levels of the community pharmacy workforce using a survey tool. Int J Pharm Pract 2023;31(1):55-61. [doi: 10.1093/ijpp/riac091] [Medline: 36472945]
35. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. JMIR Med Educ 2022;8(2):e38259 [FREE Full text] [doi: 10.2196/38259] [Medline: 35767323]
36. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002] [Medline: 28347450]
37. Tesfa GA, Yehualashet DE, Ewune HA, Zemeskel AG, Kalayou MH, Seboka BT. eHealth literacy and its associated factors among health professionals during the COVID-19 pandemic in resource-limited settings: cross-sectional study. JMIR Form Res 2022;6(7):e36206 [FREE Full text] [doi: 10.2196/36206] [Medline: 35737897]
38. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MDP, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]
39. Chereka AA, Demsash AW, Ngusie HS, Kassie SY. Digital health literacy to share COVID-19 related information and associated factors among healthcare providers worked at COVID-19 treatment centers in Amhara region, Ethiopia: a cross-sectional survey. Inform Med Unlocked 2022;30:100934 [FREE Full text] [doi: 10.1016/j.imu.2022.100934] [Medline: 35441087]
40. Battel-Kirk B, Barry MM. Has the development of health promotion competencies made a difference? A scoping review of the literature. Health Educ Behav 2019;46(5):824-842. [doi: 10.1177/1090198119846935] [Medline: 31064229]
41. Brørs G, Larsen MH, Hølvold LB, Wahl AK. eHealth literacy among hospital health care providers: a systematic review. BMC Health Serv Res 2023;23(1):1144 [FREE Full text] [doi: 10.1186/s12913-023-10103-8] [Medline: 37875882]
42. Liu S, Zhao H, Fu J, Kong D, Zhong Z, Hong Y, et al. Current status and influencing factors of digital health literacy among community-dwelling older adults in Southwest China: a cross-sectional study. BMC Public Health 2022;22(1):996 [FREE Full text] [doi: 10.1186/s12889-022-13378-4] [Medline: 35581565]
43. Li S, Cui G, Yin Y, Xu H. Associations between health literacy, digital skill, and eHealth literacy among older Chinese adults: a cross-sectional study. Digit Health 2023;9:20552076231178431 [FREE Full text] [doi: 10.1177/20552076231178431] [Medline: 37256010]
44. Zhao BY, Huang L, Cheng X, Chen TT, Li SJ, Wang XJ, et al. Digital health literacy and associated factors among internet users from China: a cross-sectional study. BMC Public Health 2024;24(1):908 [FREE Full text] [doi: 10.1186/s12889-024-18324-0] [Medline: 38539176]
45. Karasiewicz M, Chawłowska E, Lipiak A, Wi ckowska B. How to improve cancer prevention knowledge? A way to identify gaps and tackle the limited availability of health education services in primary health care using the European code against cancer. Front Public Health 2022;10:878703 [FREE Full text] [doi: 10.3389/fpubh.2022.878703] [Medline: 35586014]
46. Khoury ZH, Illesca P, Sultan AS. The role of primary care providers in oral health education for patients with diabetes. Patient Educ Couns 2021;104(6):1497-1499. [doi: 10.1016/j.pec.2020.11.020] [Medline: 33246872]
47. Yoğurtcu H, Ozturk Haney M. The relationship between e-health literacy and health-promoting behaviors of Turkish hospital nurses. Glob Health Promot 2022:17579759221093389. [doi: 10.1177/17579759221093389] [Medline: 35730122]
48. Sun H, Qian L, Xue M, Zhou T, Qu J, Zhou J, et al. The relationship between eHealth literacy, social media self-efficacy and health communication intention among Chinese nursing undergraduates: a cross-sectional study. Front Public Health 2022;10:1030887 [FREE Full text] [doi: 10.3389/fpubh.2022.1030887] [Medline: 36388375]
49. Do BN, Tran TV, Phan DT, Nguyen HC, Nguyen TTP, Nguyen HC, et al. Health literacy, eHealth literacy, adherence to infection prevention and control procedures, lifestyle changes, and suspected COVID-19 symptoms among health care workers during lockdown: online survey. J Med Internet Res 2020;22(11):e22894 [FREE Full text] [doi: 10.2196/22894] [Medline: 33122164]
50. Kritsotakis G, Andreadaki E, Linardakis M, Manomenidis G, Bellali T, Kostagiolas P. Nurses' eHealth literacy and associations with the nursing practice environment. Int Nurs Rev 2021;68(3):365-371. [doi: 10.1111/inr.12650] [Medline: 34004022]

XSL•FO
RenderX

51. Gartrell K, Han K, Trinkoff A, Cho H. Three-factor structure of the eHealth Literacy Scale and its relationship with nurses' health-promoting behaviours and performance quality. J Adv Nurs 2020;76(10):2522-2530. [doi: 10.1111/jan.14490] [Medline: 33463741]
52. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024;7(1):47 [FREE Full text] [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
53. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
54. Ministry of Human Resources and Social Security. Guiding opinions of the Ministry of Human Resources and Social Security, the National Health Commission, and the State Administration of Traditional Chinese Medicine on deepening the reform of the professional title system for health professionals and technicians. 2021 Jun 30. URL: https://www.gov.cn/zhengce/zhengceku/2021-08/05/content_5629566.htm?eqid=825085780004771600000000664672d5a [accessed 2024-05-28]
55. Yao Y, Zhang H, Liu X, Liu X, Chu T, Zeng Y. Bridging the digital divide between old and young people in China: challenges and opportunities. Lancet Healthy Longev 2021;2(3):e125-e126 [FREE Full text] [doi: 10.1016/S2666-7568(21)00032-5] [Medline: 36098110]
56. Yang E, Chang SJ, Ryu H, Kim HJ, Jang SJ. Comparing factors associated with eHealth literacy between young and older adults. J Gerontol Nurs 2020;46(8):46-56. [doi: 10.3928/00989134-20200707-02] [Medline: 32936926]
57. Maier CB, Budde H, Buchan J. Nurses in expanded roles to strengthen community-based health promotion and chronic care: policy implications from an international perspective; a commentary. Isr J Health Policy Res 2018;7(1):64 [FREE Full text] [doi: 10.1186/s13584-018-0257-5] [Medline: 30314532]
58. Bodenheimer T, Bauer L. Rethinking the primary care workforce—an expanded role for nurses. N Engl J Med 2016;375(11):1015-1017. [doi: 10.1056/NEJMp1606869] [Medline: 27626516]
59. Sangwan RK, Huda RK, Panigrahi A, Toteja GS, Sharma AK, Thakor M, et al. Strengthening breast cancer screening program through health education of women and capacity building of primary healthcare providers. Front Public Health 2023;11:1276853 [FREE Full text] [doi: 10.3389/fpubh.2023.1276853] [Medline: 38035296]
60. Aghaji A, Burchett HED, Hameed S, Gilbert C. Assessing the capacity of primary health care facilities in Nigeria to deliver eye health promotion: results of a mixed-methods feasibility study. PLOS Glob Public Health 2022;2(11):e0000645 [FREE Full text] [doi: 10.1371/journal.pgph.0000645] [Medline: 36962620]

Abbreviations

DCF: Digital Capabilities Framework
KMO: Kaiser-Meyer-Olkin
PHC: primary health care

Edited by A Mavragani; submitted 08.06.24; peer-reviewed by BA Odugbemi, M Colombo; comments to author 27.09.24; revised version received 10.10.24; accepted 07.11.24; published 20.12.24.

Please cite as:
Wang Y, Ren H, Xiao S, Meng T, Sun S, Yu S, Liu Q, Wang F
Factors Associated With Digital Capacity for Health Promotion Among Primary Care Workers: Cross-Sectional Survey Study
J Med Internet Res 2024;26:e63054
URL: https://www.jmir.org/2024/1/e63054
doi:10.2196/63054
PMID:

©Yining Wang, Hui Ren, Shaotan Xiao, Tian Meng, Shuyue Sun, Siyu Yu, Qing Liu, Fan Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integration of Patient-Reported Outcome Data Collected Via Web Applications and Mobile Apps Into a Nation-Wide COVID-19 Research Platform Using Fast Healthcare Interoperability Resources: Development Study

Johannes Benedict Oehm[1], MSc; Sarah Luise Riepenhausen[1]; Michael Storck[1], PhD; Martin Dugas[2], MD, MSc; Rüdiger Pryss[3], PhD; Julian Varghese[1], MD, MSc

[1]Institute of Medical Informatics, University of Münster, Münster, Germany
[2]Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany
[3]Institute of Clinical Epidemiology and Biometry, University of Würzburg, Würzburg, Germany

Corresponding Author:
Johannes Benedict Oehm, MSc
Institute of Medical Informatics
University of Münster
Albert-Schweizer-Campus 1, Gebäude 11
Münster, 48149
Germany
Phone: 49 251 83 58247
Fax: 49 251 83 52259
Email: johannes.oehm@uni-muenster.de

Abstract

Background: The Network University Medicine projects are an important part of the German COVID-19 research infrastructure. They comprise 2 subprojects: COVID-19 Data Exchange (CODEX) and Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing (COMPASS). CODEX provides a centralized and secure data storage platform for research data, whereas in COMPASS, expert panels were gathered to develop a reference app framework for capturing patient-reported outcomes (PROs) that can be used by any researcher.

Objective: Our study aims to integrate the data collected with the COMPASS reference app framework into the central CODEX platform, so that they can be used by secondary researchers. Although both projects used the Fast Healthcare Interoperability Resources (FHIR) standard, it was not used in a way that data could be shared directly. Given the short time frame and the parallel developments within the CODEX platform, a pragmatic and robust solution for an interface component was required.

Methods: We have developed a means to facilitate and promote the use of the German Corona Consensus (GECCO) data set, a core data set for COVID-19 research in Germany. In this way, we ensured semantic interoperability for the app-collected PRO data with the COMPASS app. We also developed an interface component to sustain syntactic interoperability.

Results: The use of different FHIR types by the COMPASS reference app framework (the general-purpose FHIR Questionnaire) and the CODEX platform (eg, Patient, Condition, and Observation) was found to be the most significant obstacle. Therefore, we developed an interface component that realigns the Questionnaire items with the corresponding items in the GECCO data set and provides the correct resources for the CODEX platform. We extended the existing COMPASS questionnaire editor with an import function for GECCO items, which also tags them for the interface component. This ensures syntactic interoperability and eases the reuse of the GECCO data set for researchers.

Conclusions: This paper shows how PRO data, which are collected across various studies conducted by different researchers, can be captured in a research-compatible way. This means that the data can be shared with a central research infrastructure and be reused by other researchers to gain more insights about COVID-19 and its sequelae.

(J Med Internet Res 2024;26:e47846) doi:10.2196/47846

XSL•FO
RenderX

KEYWORDS

Fast Healthcare Interoperability Resources; FHIR; FHIR Questionnaire; patient-reported outcome; mobile health; mHealth; research compatibility; interoperability; Germany; harmonized data collection; findable, accessible, interoperable, and reusable; FAIR data; mobile phone

Introduction

Background

Owing to the high number of individuals infected with SARS-CoV-2, lasting symptoms and sequelae are becoming increasingly common in the global population. Therefore, research on this topic is important. The clinical picture of these patients is usually referred to as post–COVID-19 condition. Regular follow-ups and continuous monitoring of patients with post–COVID-19 condition are crucial to collect data for research about this new disease.

As treatment for patients with post–COVID-19 condition is typically done in the home setting, the use of electronic patient-reported outcomes (ePROs) is the preferred means of collecting data about symptoms, well-being, and overall health, which is key for follow-up and gathering more information.

Objective

As aftercare and data gathering is time consuming and associated with high costs, sharing and reusing the data for multiple research projects is very beneficial. Therefore, it is necessary to provide a means to collect interoperable data and to create an infrastructure to exchange data.

During the peak of the pandemic, the Network University Medicine (NUM) was founded in Germany to create a central research infrastructure to collect and share medical data on COVID-19. As part of a small subproject, an app framework for ePRO collection was developed by a group of research experts and technical experts who had previous experience in conducting studies with patient-reported outcomes (PROs) and electronic data capture systems.

To reinforce the benefits of the research infrastructure provided by NUM, the objective of this study was to integrate the data collected by the ePRO apps into the central data repository. Therefore, an interface component (compass-interface-codex) was developed.

Methods

Prerequisites

When the first lockdown was imposed at the start of the COVID-19 pandemic, smartphones became an important tool for data collection. Researchers urgently needed data to better assess the typical natural history of the new disease owing to its threat to public health and the psychological and social impact of the lockdowns and social distancing measures.

At that time, many apps were developed at very short notice. They were often developed as isolated solutions for comparatively similar or even the same research problems, but the data were neither standardized nor interoperable, which limited their scientific use [1]. Hence, the Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing (COMPASS) project was initiated by the German Federal Ministry of Education and Research early during the pandemic. In this project, interdisciplinary teams of experts came together and collected the best practices for the development and deployment of pandemic apps in terms of technical, ethical, and regulatory requirements.

As part of the COMPASS project, a reference app framework for ePROs was developed, consisting of a native app and a web application. A screenshot of the noncustomized COMPASS native app with the default style is shown in Figure 1. A questionnaire editor was also developed to configure the apps. The app framework is available as open-source software on GitHub [2] and can be used and customized by anyone free of charge.

XSL•FO
RenderX

Figure 1. Main page of the Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing reference app with default styling.

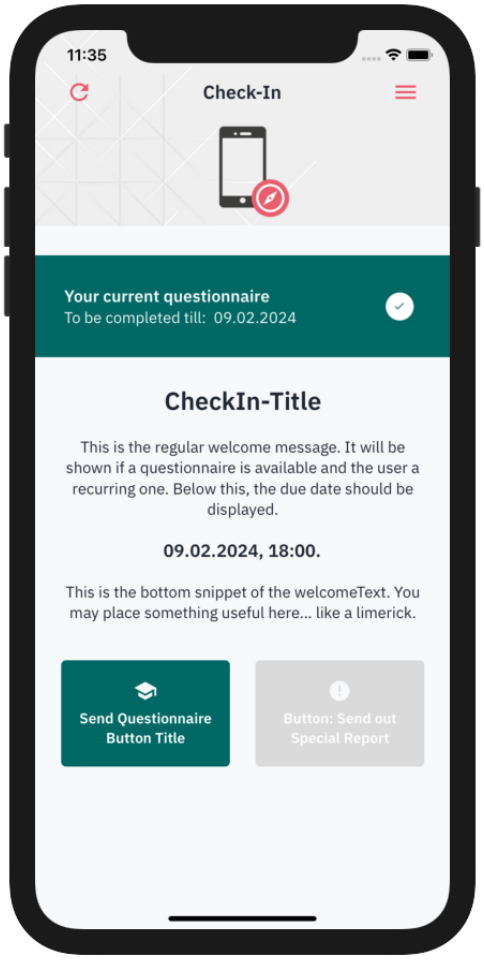

COMPASS is part of the NUM projects, which was established specifically for COVID-19 research through special funding from the German Federal Ministry of Education and Research as part of the crisis management for COVID-19 research. All 36 university hospitals in Germany are involved. A total of 13 projects were funded, including COMPASS and COVID-19 Data Exchange (CODEX) [3].

The German Corona Consensus (GECCO) data set [4] is a COVID-19–related core data set. It was defined at the start of the NUM projects to enable harmonized cross-site and cross-project data capture. The GECCO data set is defined in *Advanced Requirement Tooling using Data Elements, Codes, Object Identifiers and Rules*, and a set of Fast Healthcare Interoperability Resources (FHIR) profiles covering the Patient, Consent, Observation, Condition, Procedure, and MedicationStatement resource types. The elements are coded using *International Statistical Classification of Diseases and Related Health Problems-10th edition–German Modification* (*ICD-10–GM*), Logical Observation Identifiers Names and Codes (LOINC), Unified Code for Units of Measure, Anatomical Therapeutic Chemical (ATC), and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT).

Another important infrastructure project by NUM is the CODEX platform [5]. A secure and extensible platform was developed to collect GECCO data from the data integration centers of the German Medical Informatics Initiative [6] in a central location. This prevents the creation of data silos at local sites. There is a portal through which researchers can request access to the data. The CODEX platform is hosted in a secure environment at the society for scientific data processing (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen) [7].

Requirement Analysis

Before the technical requirement analysis, a group of experts within NUM-COMPASS gathered the requirements for research professionals.

An overview of the use cases, including the components and actors involved, is illustrated in Figure 2. Researchers as primary end users want to use the COMPASS app framework to conduct a study by designing questionnaires and deploying the app to capture data provided by the patients for their research. The captured data can then be published on the CODEX platform, so that others can perform secondary research. Therefore, data must be transformed between the data formats used by COMPASS and CODEX.

XSL•FO
RenderX

Figure 2. Use case diagram with the involved actors and components. CODEX: COVID-19 Data Exchange; COMPASS: Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing; GECCO: German Corona Consensus; PROM: patient-reported outcome measure.

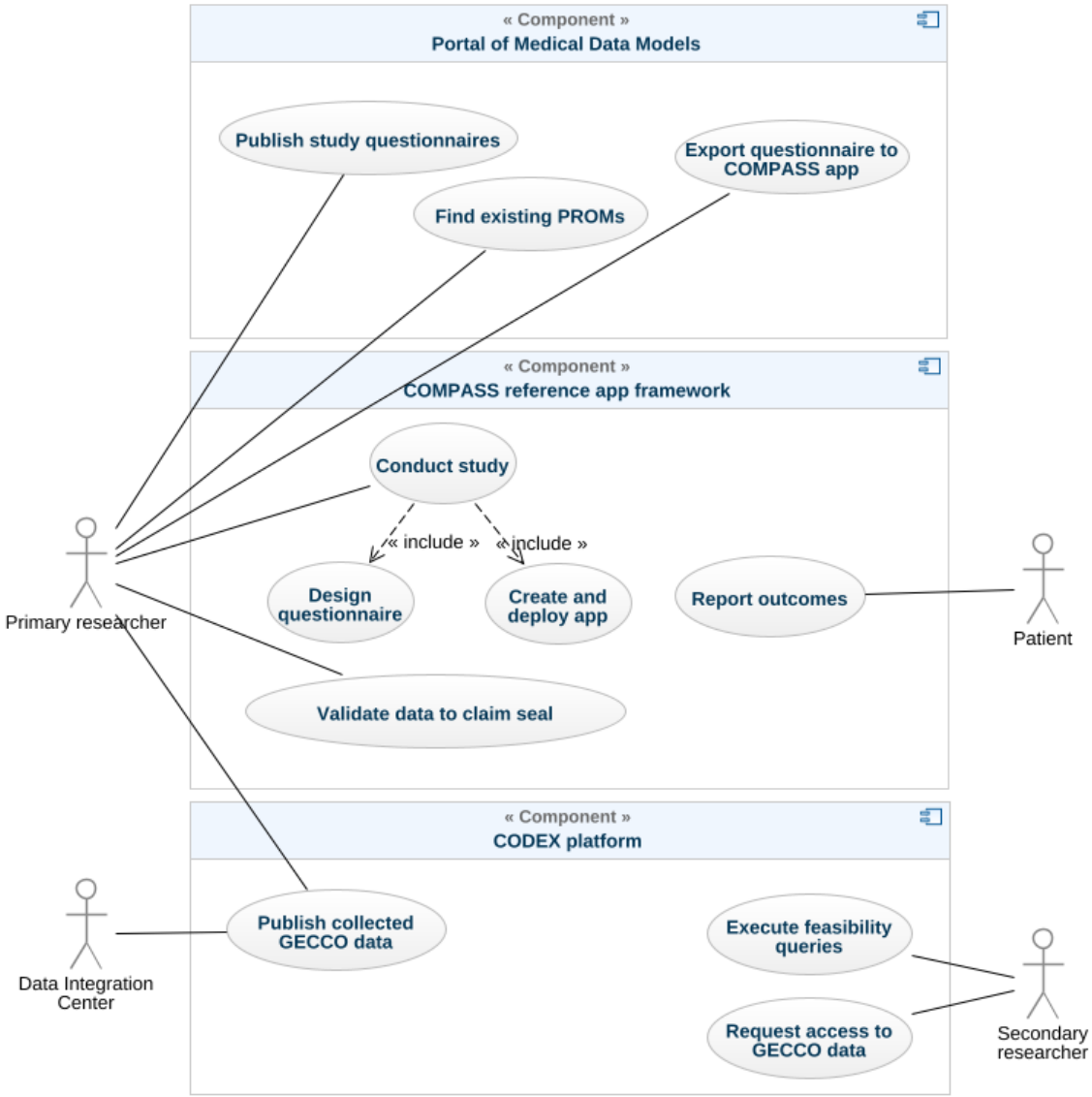

At this point, apart from the technical transformation, which is usually cumbersome but feasible after data collection, the incompatibility of the data elements collected across various studies was considered a major problem. Therefore, the use of core data sets was emphasized from the beginning. As the use of the GECCO data set was already defined for all NUM projects, it was a logical decision to promote its use. This means that if a new study commences, the data items to be collected should use the GECCO data set as a starting point.

Lack of awareness, lack of time, and lack of direct benefit were considered the main reasons why researchers ignore the need to capture data in a research-compatible way. To mitigate these effects, a First Contact Package [8] for COMPASS users was developed to explain good practices to researchers and recommend using the GECCO data set. To incentivize the use of the GECCO data set, a GECCO – Approved by NUM-COMPASS seal (Figure 3) was developed. To award this seal, Data4Life was tasked to develop a validation service called GECCO Conformance Checker [9].

XSL•FO
RenderX

Figure 3. German Corona Consensus (GECCO)–approved seal that can be issued to studies conducted using the GECCO data set.

To encourage the reuse of further study-specific data items beyond the GECCO data set, the Portal of Medical Data Models (MDM-Portal) [10] was recommended, and resources should be provided to facilitate integration. The MDM-Portal is a research infrastructure that provides public access to 25,157 data models (as of July 2023).

For cases in which the study has already started with a different data set, a "Late Mapping Scheme" to the GECCO data set has been created, but the emphasis is on using the GECCO data set from the beginning.

Technical Evaluation of Existing Components

The first step in developing the interface component was to examine the reference app framework.

The reference app framework consists of source code templates for building a native app and a web application. The native app template is based on the Gutenberg COVID-19 study [11] app and was developed by IBM Corp. The web application is based on CovApp developed by Data4Life. The questionnaire editor was developed from scratch by Healex Systems.

To start a new study with the app framework, a developer can fork the existing source code from GitHub and exchange icons and textual description templates. Extensive documentation and video tutorials [12] are available for reference. For customization, new features can be added to the implementation, which can then be published in the Apple App Store and Google Play Store.

To participate in a study, patients need to install the custom-built app from their smartphone's app store. To authorize the installed app for participation in the study, the patients have to scan a personalized QR code, which transmits the SubjectID to the device. Then, the app retrieves the questions from FHIR Questionnaire resources from the back end. These questions are then displayed to the user on a regular schedule.

After the user responses are collected, the app generates an FHIR QuestionnaireResponse resource that is sent as an asymmetrically encrypted binary large object (BLOB) and stored in the back end. Only researchers who have access to the private key can decrypt the BLOB with a Python downloader script, whereas the private key is not accessible in the back end.

Public cloud services can be used, as the responses remain encrypted in the app back end. This eliminates the need for the time-consuming configuration of firewall exceptions in hospital networks, increases the availability of the app back end, and saves hosting costs. Another major benefit is that it enables the use of cloud-based application programming interfaces for sending push notifications. This way, users can be automatically notified when new questionnaires need to be completed, which likely results in an increase in response rates [13].

The CODEX platform uses an EHRbase [14] instance as an openEHR-based central storage system. The transmission is done via an FHIR interface, but no Questionnaire resources are used. Instead, the platform requires FHIR resources, which are compliant to the GECCO profiles to receive the data. Using the FHIR Bridge [15], the GECCO profiles are mapped to the corresponding openEHR templates specifically created for the GECCO data set and stored in an EHRbase instance. This way, the data stored in CODEX are compatible not only with FHIR but also with openEHR.

Technical Requirement Analysis

Both the app and CODEX are using FHIR-based [16] interfaces for data exchange. Health Level–7 FHIR is an emerging standard in the field of health research [17]. FHIR is built upon XML, JSON, and Representational State Transfer technologies, which are comparatively lightweight and therefore well suited for mobile apps. The standard defines a collection of >140 resource types that define clinical concepts. These resource types can be adapted and used in a variety of use cases. The use of these resources to solve a specific interoperability problem is described in the implementation guides. In these implementation guides, FHIR resource types are usually constrained using the profiling mechanism.

Interoperability may require the conversion of data within the FHIR standard. In our case, there will be a variety of FHIR

XSL•FO
RenderX

Questionnaires used in the apps and a different set of resource types constrained by the GECCO profiles used by the CODEX platform to exchange data. The gap in between was identified as the biggest obstacle.

A FHIR Questionnaire resource stores only structural information. It consists of items that can be nested hierarchically. These items consist primarily of the text displayed to the user, the answer's data type, visibility conditions (enableWhen), the "required" attribute, and the linkId. The linkId is used for linkage between the answer in the QuestionnaireResponse and the corresponding question item in the Questionnaire.

The resource types constrained by the GECCO profiles (Patient, Observation, Condition, etc) are much stricter: They usually consist of a central code element linked to an external medical terminology such as LOINC, SNOMED CT, ATC, or *ICD-10–GM*. Therefore, they enforce a much higher degree of semantic interoperability of data. This is in contrast to Questionnaire and QuestionnaireResponse resources, which, on the basic level, only enforce syntactic interoperability.

The COMPASS apps use FHIR Questionnaires to keep the questions from various studies reusable. However, the CODEX platform cannot process FHIR Questionnaires of unknown origin. It is possible to store specific questionnaires in the FHIR Bridge and create a manual mapping to an openEHR template that has to be created specifically for the Questionnaire.

Zhu et al [18] used this approach to integrate the COVID-19 patient assessment questionnaire with the native app, which had already been integrated in the CODEX platform for data collection with CovApp [19]. For new studies with different questionnaires, this approach to integrate requires a lot of work on the CODEX platform for the connection of each new app, and it will not work if a study introduces new questionnaires or if existing questionnaires are changed at short notice.

In conclusion, to solve the interoperability problem, it is necessary to transform the items from the QuestionnaireResponses to their corresponding resources in the GECCO implementation guide, which is applicable for new studies conducted with the COMPASS app framework.

Point of Transformation

Before the implementation of an interface component could begin, an architectural decision had to be made about the location of the conversion from FHIR QuestionnaireResponse to the GECCO profiles.

The various options were identified and evaluated in an expert-led roundtable with the back end, native app, and web application developers.

Assessment of Existing Mapping Methods

To address the question of how items from QuestionnaireResponses can be mapped to the corresponding FHIR resource types, a literature search was performed by using the keywords "FHIR" and "Questionnaire" in the PubMed and MEDLINE databases. Furthermore, an in-depth examination of the FHIR standard was performed.

Evaluation

To verify the correctness of the mapped resources, the GECCO Conformance Checker developed by Data4Life was used. A test data set containing 10 patients with 20 QuestionnaireResponses was submitted to the central platform to verify the overall pipeline.

Ethical Considerations

All data used during the development and implementation process were synthetic. No data of real persons were used; therefore, ethics approval was not required.

Results

Point of Transformation

In the expert round, the following options were identified and their respective advantages and disadvantages were discussed:

1. On the smartphone: This would require 2 implementations for both the native app and the web application, thereby increasing costs. In addition, this solution would increase the download size of the apps but without adding any direct value for the patient. Owing to this fact and the risk of additional problems that could arise from outdated app versions transmitting outdated mappings, this option was discarded.
2. Inside the back end of the COMPASS project: Both the native app and the web application use the same back end for data submission. Thus, the back end would also be conceivable as the location of the transformation. However, this would require decryption of the QuestionnaireResponse BLOBs within untrusted cloud servers. This option would offset the privacy gains of the encrypted architecture and therefore had to be rejected.
3. What remained was the development of a separately running custom component that would run within the protected environment of the CODEX platform. Therefore, this component polls the common app back end periodically, downloads the encrypted BLOBs, decrypts them, and stores the transformed data in the internal data storage layer.

The final architecture and the flow of data are shown in Figure 4.

XSL•FO
RenderX

Figure 4. Overview of the overall architecture and the involved resource types and software components. The arrows indicate the data flow. Green arrows indicate that the data are only transport encrypted, whereas the brown arrows indicate the end-to-end encryption of the QuestionnaireResponse resource throughout the cloud-hosted back end. CODEX: COVID-19 Data Exchange; FHIR: Fast Healthcare Interoperability Resources; GECCO ODM: German Corona Consensus Operational Data Model; MDM: Medical Data Model.

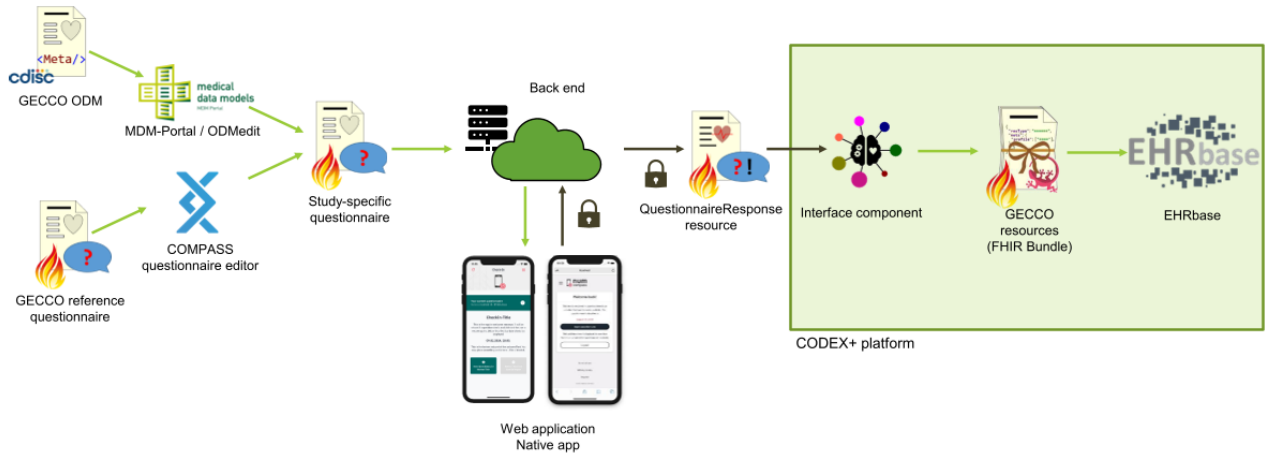

Assessment of Existing Mapping Methods

The literature search, as of September 1, 2021, returned the following results:

1. MEDLINE: 27 hits in the search for FHIR Questionnaire; none of them was relevant
2. Web of Science: 11 hits; none of them was relevant

In a publication, a mapping of a logical model that resulted from natural language processing parsing to other FHIR resource types was apparently performed [20].

The FHIR standard itself does not provide any suggestions about how to extract FHIR resources, but the Structured Data Capture Implementation Guide contains 3 different proposals—observation based, definition based, and FHIR Mapping Language [21]. All of them had different shortcomings for our use case, which are outlined in Multimedia Appendix 1, and an adaptation could not be done in a timely manner. Therefore, it was decided to perform the mapping with the Kotlin programming language.

Mapping Concept and Implementation

In the FHIR context, so-called logical models are commonly used to define medical data models by domain experts with minimal knowledge about FHIR resources. These logical models are later "translated" into FHIR profiles [22]. As GECCO's logical model is structurally more similar to a questionnaire than the GECCO FHIR profiles, we used it as the starting point for a project-specific logical model [23] and used the questions of the REDCap (Research Electronic Data Capture; Vanderbilt University) forms provided by Erbelding et al [24], which were developed for capturing GECCO data sets in hospitals, to develop the logical model further, so that it can be translated into a questionnaire.

Our logical model contains all the medical concepts and data elements included in the GECCO data set. Each data element is annotated with the corresponding data type and a textual question. The logical model was created with a well-defined mapping to the corresponding GECCO FHIR profiles in mind. An excerpt of the logical model is shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Excerpt of the structure of the logical model of the German Corona Consensus (GECCO) data set. Element identifier with associated questions are in parenthesis as well as the possible answer options.

- anamnesis: („Anamnesis/risk factors“)
  - hasChronicLungDisease („Does the patient suffer from a chronic lung disease?“): yes/no/unknown
  - chronicLungDisease: („Which chronic lung disease does the patient suffer from?“, enableWhen anamnesis.hasChronicLungDisease = yes)
    - asthma („Asthma“): yes/no/unknown
    - copd („COPD“): yes/no/unknown
    - (...)
  - (...)
- demographics („Demographics“)
  - biologicalSex („Biological sex“): male/female/unknown/x/d
  - (...)
- epidemiologicalFactors („Epidemiological factors“)
  - (...)
- complications („Complications“)
  - (...)
- (...)

The implementation was performed using the Kotlin programming language and the *HL7 Application Programming Interface* FHIR library [25]. The decryption of the encrypted QuestionnaireResponse BLOBs was implemented using the bouncycastle library. A Helm chart, which integrates into the CODEX platform infrastructure, was created for deployment.

The logical model is stored as a set of Kotlin classes. The elements are annotated with textual questions and LOINC or *ICD-10–GM codes*, if applicable. A class structure rather than JSON-based data formats to define the logical model, for example, was chosen because several common patterns could be found in the GECCO data set, which required >1 Questionnaire item per profile. For these common patterns, custom data types such as *YesNoUnknownWithDateTime* (vaccinations) and *YesNoUnknownWithSymptomSeverity* (symptoms) were defined, which have a specific mapping and a well-defined enableWhen behavior and can be used at multiple places. For example, it is pointless to ask the patient about the severity of their cough if they have previously answered that they do not have a cough.

Another benefit of using a class structure is the resulting type safety inside the transformation logic. If the logical model is updated, the affected profiles can easily be found in the process.

To improve the usability of the Questionnaire, in the anamnesis part, meta-elements were defined, which do not directly correspond to profiles of the GECCO data set. Those meta-elements ask the user for a specific category of preexisting diseases. If the user answers “No” or “Unknown” for this disease category, the apps will not show the list of all diseases in the data set using the enableWhen mechanism and the Condition resources will be set to “No” or “Unknown”, respectively.

By means of reflection, the GECCO reference questionnaire is generated from the logical model. This questionnaire can be used by researchers as a starting point for editing a custom questionnaire. It is stored as a reference questionnaire in the questionnaire editor to be able to import individual items from the GECCO data set.

Each item in the generated reference questionnaire has an extension with a reference to the path in the logical model. This path is transferred into the QuestionnaireResponse resource by the COMPASS apps. For other apps, the interface component is also able to transfer the extensions into the QuestionnaireResponse resource if the Questionnaire resource is known. This way, individual items of the reference questionnaire can be uniquely identified via the path in the logical model. The mapping is then executed based on this assignment. The generated GECCO profile instances are exported in the form of an FHIR bundle.

In total, the GECCO data set could be broken into 45 template functions for the different profiles and resources, and >430 data elements coded with SNOMED CT, LOINC, ICD-10-GM, or ATC were created. The template functions are available as part of the internal library “gecco-easy”, which can be used for manual transformation of other Questionnaire items.

The reference to the subject of the created resources is automatically set by the SubjectID used in the app, and the recording time of the resources is automatically set to the time the QuestionnaireResponse was submitted (refer to Table 1 for specifically set attributes).

XSL•FO
RenderX

Table 1. Mapping of the contextual information of the QuestionnaireResponse to the corresponding attributes of the specific resource types.

FHIR[a] resource type	Subject-specifying attribute	Recorded, time-specifying attribute
Patient	N/A[b]	N/A
Consent	Consent.patient	N/A
Observation	Observation.subject	Observation.effectiveDateTime
Condition	Condition.subject	Condition.recordedDate
Procedure	Procedure.subject	Procedure.performedDateTime[c]
MedicationStatement	MedicationStatement.subject	MedicationStatement.effectiveDateTime
Immunization	Immunization.patient	Immunization.occurenceDateTime[c]

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]N/A: not applicable.

[c]For domain-specific reasons, these attributes are not inferred by the context.

A problem we encountered with FHIR Questionnaires was the lack of support for multiple Codings per answerOption in the Questionnaire. The Condition.code, Observation.code, and Procedure.code elements are of type CodeableConcept, which means that multiple Codings can be used to describe a fact. In most cases, 1 Coding is sufficient for a distinct assignment to 1 answer option in the logical model. From that point, the interface component can emit all the required Codings for the target resource profile. However, 2 custom CodeSystems had to be created (Covid19-Vaccine and VentilationTypes).

The implemented solution was documented as part of the COMPASS implementation guide [26]. This way, other apps can use the extensions and thus the interface component for their needs.

Improvements to the Questionnaire Editor

To tackle the issue of lack of time that most researchers face during the Questionnaire development phase, an import function for GECCO items was added to the questionnaire editor. The import button is prominently placed among the other buttons to add new items. This way, it is hard to miss, and importing the existing GECCO items or item sets is more convenient than designing custom ones; however, it is still possible.

When a question or a set of questions is imported from the GECCO data set, it is highlighted accordingly in the user interface. Some functions such as adding or removing new answer options are disabled to avoid problems with different scales and the grade for symptoms.

A screenshot of the updated editor is shown in Figure 5.

XSL•FO
RenderX

Figure 5. Screenshot of the Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing questionnaire editor with German Corona Consensus (GECCO) and non-GECCO items inside a Fast Healthcare Interoperability Resources Questionnaire resource. On the left, the tree of questionnaire items is shown; on the right, the details of the selected item can be configured. The GECCO items are highlighted in the tree. The menu to add new items at the bottom left is expanded, revealing the option to add GECCO items along with the standard item data types.

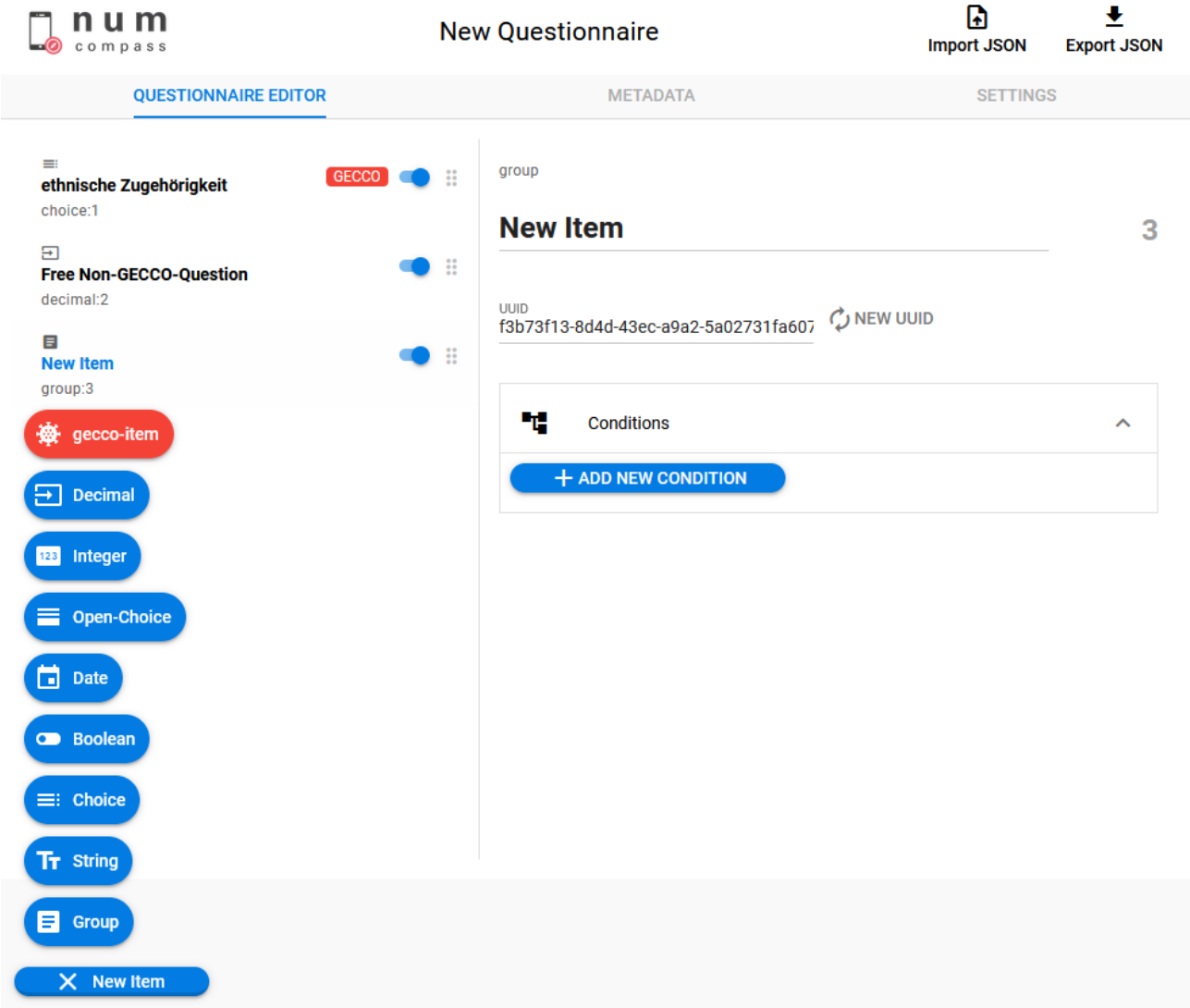

Integration With the MDM-Portal

To include additional PRO measures (PROMs), an integration with the MDM-Portal [27], which internally uses Clinical Data Interchange Standards Consortium Operational Data Model (ODM) files, was implemented.

To ensure interoperability with the existing apps, interface component, and GECCO FHIR profiles, a new variant of the existing ODM to FHIR Questionnaire converter [28] called ODM2COMPASS was created. For this purpose, the items within the GECCO form were annotated with their respective path in the logical model using ODM's Alias-Tag. The ODM2COMPASS converter uses this annotation to create the CompassGeccoItem extension in the resulting Questionnaire, which is needed by the interface component to map the response to the corresponding GECCO profiles.

Further adaptations to the existing converter include a conversion of MDM-Portal's adapted ODM-ConditionDef syntax to enableWhen element of FHIR Questionnaires. As FHIR Questionnaires only support 1 Coding per answerOption but ODM allows multiple aliases per CodeListItem, the reference Questionnaire had to be included. When there are multiple Codings or Alias-Tags available, the CodeSystem or Context used in the reference Questionnaire will be used in the export process by the converter. This way, the interface component can recognize not only the questions but also the answers. For provenance, the Questionnaire.url element is set to the MDM-Portal's URL of the form.

Researchers can create custom questionnaires by creating a new data model and using the built-in ODMedit tool [29], which also allows for carrying over parts of other forms inside the MDM-Portal. As the Alias-Tags are also adopted, when the resulting form is exported to the reference app framework, the interface component is still able to recognize elements from the original GECCO form and map them to the corresponding profiles.

This way, researchers can easily combine the GECCO data set with PROMs that contain items that are not available in GECCO such as overall well-being (eg, World Health Organization-5

XSL•FO
RenderX

Well-Being Index) or mental health (eg, Generalized Anxiety Disorder 7). Furthermore, researchers can publish their newly created questionnaires directly in the portal, increasing the level of adherence to the FAIR (findability, accessibility, interoperability, and reuse) criteria in the process.

Validation of the Created Resources

A generator script for QuestionnaireResponses was developed, which produces QuestionnaireResponses with random answers based on an input Questionnaire resource. These generated QuestionnaireResponses were then transformed into bundles of GECCO profiles by the interface component. This way, all possible answer options could be tested.

As part of the project, Data4Life was commissioned to develop the GECCO conformance checker. This component validates the entered resources for conformance with the FHIR profiles and creates a PDF file with the GECCO – Approved by NUM-COMPASS seal as an incentive for researchers [9]. A screenshot of such a validation result document can be seen in Figure 6. This conformance checker was used to verify whether, for each GECCO item, a resource was generated and all profiles were created correctly.

Figure 6. Screenshot of the result of the German Corona Consensus conformance check and comparison of the existing methods for Questionnaire extraction.

GECCO Konformitätsprüfung

Einreichende Organisation:	Klinik für Infektiologie - Universitätsklinikum Beispielstadt Spitalgasse 123 12345 Beispielstadt Germany
Geprüfte Anwendung:	Example Covid-19 App
Hersteller (Optional) :	
Version der geprüften Anwendung:	1.0.0
Prüfdatum:	2023-01-04

Die an den Konformitätsprüfer übermittelten Ressourcen konnten erfolgreich anhand der Spezifikationen des German Corona Consensus-Datensatzes (GECCO), Version 1.0.4, validiert werden.

Für die Konformitätsprüfung wurde(n) 67 Ressource(n) berücksichtigt:

Resourcentyp Condition
1 Beispiel(e) vom Profil
https://www.netzwerk-universitaetsmedizin.de/fhir/StructureDefinition/chronic-kidney-diseases
4 Beispiel(e) vom Profil
https://www.netzwerk-universitaetsmedizin.de/fhir/StructureDefinition/chronic-liver-diseases
7 Beispiel(e) vom Profil

XSL•FO
RenderX

In addition, further validation was conducted using EHRbase and FHIR Bridge. Both components represent the storage layer of the CODEX platform. The FHIR Bridge provides an FHIR Representational State Transfer application programming interface, which accepts and transforms the GECCO profiles to their corresponding openEHR templates. The resulting openEHR composition is then stored in the EHRbase instance. To this end, a variety of manual transformation functions have been defined inside FHIR Bridge, which strictly assume that the submitted resources conform to the GECCO profiles.

The final integration test was performed using a synthetic test data set containing 10 patients with 20 QuestionnaireResponses covering various corner cases. The test data were successfully transferred to the CODEX platform, and the correct mapping was manually validated.

Discussion

Principal Findings

We successfully created a means to integrate GECCO data that are collected by the COMPASS app framework into the CODEX platform in a way that can be efficiently applied to a variety of questionnaires. The generalizability of our approach makes it superior to the existing approach used by Zhu et al [18], which requires using a fixed questionnaire with the apps and a questionnaire-specific implementation at the central platform, which severely limits its practical use.

Normally, this flexibility in the design of the questionnaires would require the manual creation of questionnaires by the researchers to obtain interoperability with the central platform. After the creation of questionnaires, it would be necessary to create a manual mapping script for each questionnaire to assign the questionnaire items to the GECCO items and create the corresponding FHIR resources. This requires clinicians to have deep technical knowledge about programming languages and FHIR profiles. In contrast, our solution requires clinicians to have much less technical knowledge. It includes an interface component, which is compatible with different questionnaires, as long as the researchers provide the correct annotations. To create questionnaires with the said annotations, we provide 2 different questionnaire editors.

The COMPASS questionnaire editor was tailored to the feature set of the native app and the web application. The capability of importing questions is limited to the GECCO data set. This can be seen as an advantage, as it increases the hurdles to use data sets competing with the GECCO data set that cannot be integrated into the CODEX platform.

However, because the GECCO extension modules are in an early stage, it is better to facilitate the reuse of existing PROMs for areas that are not yet covered by the GECCO data set. This is where ODMedit, with its existing integration to the MDM-Portal, can shine. The MDM-Portal provides a vast number of PROMs, which can be easily used to extend the GECCO data set. However, owing to the generic nature of ODMedit and the involved conversion step between ODM and FHIR, it does not have the capability to configure features that are specific to FHIR Questionnaires and the COMPASS apps. A possible mitigation is using ODMedit to create the initial Questionnaire resource by importing items from the MDM-Portal and importing the Questionnaire into the COMPASS questionnaire editor to configure the exact behavior of the items in the app.

Limitations

Notably, the integration work conducted still has limitations that must be considered. The most important limitation is that few studies have been conducted not only with the COMPASS framework itself but also with the integrated version. Even if PRO-based studies are becoming increasingly important, the acceptance of the users in the long term must always be considered. Currently, it is unclear if this acceptance will be provided. As long-term use has not yet occurred, relatively little data have been generated. Therefore, it is technically not yet possible to assess whether all aspects of data collection have been sufficiently taken into account. In addition, the health care sector is changing at a very fast pace, which may make aspects of integration obsolete. Nevertheless, from a standard perspective, the project presented is currently unique in our view and shows that open and standard-compliant development is possible in the fast-moving app sector.

On the technical side, in the current interface component implementation, the data elements of the logical model may only appear exactly once in the Questionnaire. The reason for this limitation is that only 1 instance of the logical model is created during the transformation process. Theoretically, it is possible to split the Questionnaire into small parts that are mapped independently.

Considering that the apps do not support repeatable elements and the current workflow of completing and submitting a questionnaire on the same day, this limitation is not relevant for practical use.

The interface component is constrained to the GECCO data set. This was done on purpose, in the context of the project and with the intent to enforce research compatibility. This means that other data captured by the studies cannot be transferred to the CODEX platform, limiting the use for answering questions beyond the GECCO data set. However, it is possible to adopt the basic principle of logical models to other research databases and their data models.

The development of the interface component was done with the focus on providing syntactic and semantic interoperability. Depending on the study data protection design and the consent given by the patient, additional pseudonymization steps involving an additional, trusted third party might be necessary before the interface component can be used.

FHIR and FHIR Questionnaires

On the basis of our experience in the project, there was often the misconception, especially among clinicians, that if 2 systems “speak” FHIR, they are automatically compatible without considering the use of different implementation guides, resource types, or profiles to solve a specific problem.

We believe that through the use of 140 common resource types as building blocks, FHIR provides a solid foundation for

XSL•FO
RenderX

interoperability in health care. However, FHIR is not a "panacea", as there is the ongoing as there is the ongoing challenge of building bridges between different implementation guides, as seen in our case.

Given the increasing prevalence of the FHIR standard in both electronic data capture systems and other health information systems and the growing demand for interoperable solutions, closing the gap between questionnaires and other resource types is of great importance.

We believe that there are many use cases for mapping the data captured with questionnaires to other resource types for exchange with other existing systems.

It seems very promising to promote this by reusing the existing item blocks for new questionnaires that use some sort of annotation for mapping to other resource types. Templating mechanisms (importing item trees from other questionnaires) were nonexistent during the project's requirement analysis phase but seem to be of great interest, as was recently shown by the specification of the $assemble operation by the FHIR Structured Data Capture working group [30].

Outlook and Further Studies

The submission of GECCO FHIR resources is not limited to the CODEX platform. As GECCO profiles are also compliant with the German basic profiles, they can be integrated into a variety of FHIR-compatible software systems used in hospitals. For example, the responsible physician could monitor their patient's well-being with familiar software tools.

We will monitor the development of the GECCO data set extension modules [31] and add them to the interface component once new FHIR profiles are released.

Conclusions

We successfully created a concept to integrate PRO data into the German COVID-19 research infrastructure and implemented it.

Although educating and sensitizing researchers about syntactic and semantic interoperability during the design phase of a study already remains a top priority, we are convinced that the readiness for interoperable data collection can be increased as we provide researchers with a free, easy-to-use app framework that covers the semantic and technical aspects of integration. This will likely lead to the collection of more data in an interoperable way and contribute to secondary research.

Quickly gathering data about suddenly emerging infectious diseases such as monkey pox ("pandemic preparedness") and overseeing the aftereffects of the COVID-19 pandemic such as post–COVID-19 condition are becoming high priority for public health. Therefore, future contributions to knowledge gain through PRO apps are likely to increase. National data registries, which are fed by multiple sources such as PROs that collect data in a unified format, can play an important role by improving health research and therefore will benefit patients in the long run.

Acknowledgments

The authors would like to thank all those involved in the project, university members, and industry partners, who worked together with great effort to make all of this possible within the short time frame. In particular, the authors would like to thank Dagmar Krefting and Ulrich Sax, who helped them with their expansive expertise and background knowledge about metadata and findable, accessible, interoperable, and reusable data management. The Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing project is part of the German COVID-19 Research Network of University Medicine ("Netzwerk Universitätsmedizin"), funded by the German Federal Ministry of Education and Research (funding reference 01KX2021).

Data Availability

The source code of the interface component itself is available on GitHub [32]. The back end code of the Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing app can be found in the same repository [33]. The Helm charts and the configuration of the COVID-19 Data Exchange platform are not open source but EHRbase [34] and FHIR Bridge [35] as the main components of the central platform can be used independently and are available in their corresponding GitHub repositories.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of the existing methods for questionnaire extraction.
[PDF File (Adobe PDF File), 100 KB - jmir_v26i1e47846_app1.pdf]

References

1. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019 Aug 20;2:79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
2. Netzwerk Universitäts Medizin. GitHub. 2023. URL: https://github.com/NUMde/ [accessed 2023-07-17]

XSL•FO
RenderX

3. Netzwerk Universitätsmedizin. URL: https://www.netzwerk-universitaetsmedizin.de/ [accessed 2023-07-17]
4. Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 2020 Dec 21;20(1):341 [FREE Full text] [doi: 10.1186/s12911-020-01374-w] [Medline: 33349259]
5. Prokosch HU, Bahls T, Bialke M, Eils J, Fegeler C, Gruendner J, et al. The COVID-19 data exchange platform of the German University Medicine. Stud Health Technol Inform 2022 May 25;294:674-678. [doi: 10.3233/SHTI220554] [Medline: 35612174]
6. Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018 Jul;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
7. GWDG - IT in der Wissenschaft. Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. 2023. URL: https://www.gwdg.de/ [accessed 2023-07-17]
8. Muzoora MR, Riepenhausen S, Oehm J, Poyraz RA, Schaarschmidt M, Rieß EM, et al. Making COVID-19 data interoperable: GECCO recommendations by NUM-COMPASS. In: Proceedings of the 66th Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology. 2021 Presented at: GMDS '21; September 26-30; Virtual Event URL: https://www.egms.de/static/en/meetings/gmds2021/21gmds029.shtml [doi: 10.3205/21GMDS029]
9. NUMde/compass-num-conformance-checker: tool for checking FHIR data for conformance with the GECCO standard. GitHub. URL: https://github.com/NUMde/compass-num-conformance-checker [accessed 2023-07-17]
10. Dugas M, Hegselmann S, Riepenhausen S, Neuhaus P, Greulich L, Meidt A, et al. Compatible data models at design stage of medical information systems: leveraging related data elements from the MDM portal. Stud Health Technol Inform 2019 Aug 21;264:113-117. [doi: 10.3233/SHTI190194] [Medline: 31437896]
11. Gutenberg COVID-19 studie: overview. University Medical Center Mainz. URL: https://www.unimedizin-mainz.de/gcs/overview.html?L=1 [accessed 2023-07-17]
12. COMPASS. YouTube. URL: https://www.youtube.com/channel/UC_JVfREe2bDR87dPjV7Y8og [accessed 2023-07-17]
13. Bidargaddi N, Almirall D, Murphy S, Nahum-Shani I, Kovalcik M, Pituch T, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR mHealth and uHealth 2018 Nov 29;6(11):e10123 [FREE Full text] [doi: 10.2196/10123] [Medline: 30497999]
14. EHRbase: data is for life, not just for one system. EHRbase. URL: https://ehrbase.org/ [accessed 2023-07-17]
15. ehrbase/fhir-bridge: FHIR Bridge acts as a broker between an HL7 FHIR client and an openEHR server. GitHub. 2023. URL: https://github.com/ehrbase/fhir-bridge [accessed 2023-07-17]
16. Bender D, Sartipi K. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013 Presented at: CBMS '13; June 20-22; Porto, Portugal p. 326-331 URL: https://ieeexplore.ieee.org/document/6627810/similar [doi: 10.1109/cbms.2013.6627810]
17. Vorisek CN, Lehne M, Klopfenstein SA, Mayer PJ, Bartschke A, Haese T, et al. Fast healthcare interoperability resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
18. Zhu Y, Kächer F, Wiesenfeldt M, Krefting D. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF). In: Proceedings of the 67th Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology. 2022 Presented at: GMDS '22; August 21-25; Virtual Event URL: https://www.egms.de/dynamic/de/meetings/gmds2022/index.htm?main=1 [doi: 10.3205/22GMDS121]
19. CovApp. Charité. 2023. URL: https://covapp.charite.de/ [accessed 2023-07-17]
20. Zong N, Stone DJ, Sharma DK, Wen A, Wang C, Yu Y, et al. Modeling cancer clinical trials using HL7 FHIR to support downstream applications: a case study with colorectal cancer data. Int J Med Inform 2021 Jan;145:104308 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104308] [Medline: 33160272]
21. HL7.FHIR.UV.SDC\form data extraction - FHIR v4.0.1. HL7 International. 2023. URL: https://hl7.org/fhir/uv/sdc/extraction.html [accessed 2023-07-17]
22. Logical models in FHIR 2016. Hay on FHIR. URL: https://fhirblog.com/2016/10/17/logical-models-in-fhir/ [accessed 2023-07-17]
23. COVID-19 Research-Dataset. ART-DECOR. 2022. URL: https://art-decor.org/art-decor/decor-datasets--covid19f [accessed 2022-09-08]
24. Erbelding C, Sailer B, Stenzhorn H, Biergans S, Kohlmayer F, Kobak EM, et al. GECCO on FHIR- towards interoperable data on COVID-19. In: Proceedings of the 66th Annual Conference of the German Association for Medical Informatics. 2021 Presented at: GMDS '21; September 26-30; Virtual Event URL: https://www.egms.de/static/de/meetings/gmds2021/21gmds012.shtml [doi: 10.3205/21gmds012]
25. HAPI FHIR - the open source FHIR API for Java. HAPI FHIR. 2023. URL: https://hapifhir.io/index.html [accessed 2023-07-17]
26. Oehm J. Compass-implementation-guide/index. GitHub. URL: https://github.com/NUMde/compass-implementation-guide/blob/master/input/pagecontent/index.md [accessed 2023-07-17]

XSL•FO
RenderX

27. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford) 2016 Feb 11;2016:bav121 [FREE Full text] [doi: 10.1093/database/bav121] [Medline: 26868052]
28. Doods J, Neuhaus P, Dugas M. Converting ODM metadata to FHIR questionnaire resources. Stud Health Technol Inform 2016;228:456-460. [Medline: 27577424]
29. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016 Jun 01;16:65 [FREE Full text] [doi: 10.1186/s12874-016-0164-9] [Medline: 27245222]
30. HL7.FHIR.UV.SDC\modular forms - FHIR v4.0.1. HL7 International. URL: https://build.fhir.org/ig/HL7/sdc/modular.html [accessed 2023-07-17]
31. Lichtner G, Haese T, Brose S, Röhrig L, Lysyakova L, Rudolph S, et al. Extending the German corona consensus dataset (GECCO) to the immunization, pediatrics, and cardiology domains. medRxiv. Preprint posted online May 14, 2023 2023 [FREE Full text] [doi: 10.1101/2022.05.12.22274089]
32. NUMde / compass-interface-codex. GitHub. URL: https://github.com/NUMde/compass-interface-codex [accessed 2024-01-12]
33. NUMde / compass-numapp-backend. GitHub. URL: https://github.com/NUMde/compass-numapp-backend [accessed 2024-01-12]
34. ehrbase / ehrbase. GitHub. URL: https://github.com/ehrbase/ehrbase [accessed 2024-01-12]
35. ehrbase / fhir-bridge. GitHub. URL: https://github.com/ehrbase/fhir-bridge [accessed 2024-01-12]

Abbreviations

ATC: Anatomical Therapeutic Chemical
BLOB: binary large object
CODEX: COVID-19 Data Exchange
COMPASS: Coordination on Mobile Pandemic Apps Best Practice and Solution Sharing
ePRO: electronic patient-reported outcome
FAIR: findability, accessibility, interoperability, and reuse
FHIR: Fast Healthcare Interoperability Resources
GECCO: German Corona Consensus
***ICD-10–GM*:** *International Statistical Classification of Diseases and Related Health Problems-10th edition–German Modification*
LOINC: Logical Observation Identifiers Names and Codes
MDM-Portal: Portal of Medical Data Models
NUM: Network University Medicine
ODM: Operational Data Model
PRO: patient-reported outcome
PROM: patient-reported outcome measure
REDCap: Research Electronic Data Capture
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms

Edited by A Mavragani; submitted 03.04.23; peer-reviewed by H Ulrich, C Tziraki; comments to author 05.06.23; revised version received 30.07.23; accepted 12.12.23; published 27.02.24.

Please cite as:
Oehm JB, Riepenhausen SL, Storck M, Dugas M, Pryss R, Varghese J
Integration of Patient-Reported Outcome Data Collected Via Web Applications and Mobile Apps Into a Nation-Wide COVID-19 Research Platform Using Fast Healthcare Interoperability Resources: Development Study
J Med Internet Res 2024;26:e47846
URL: https://www.jmir.org/2024/1/e47846
doi:10.2196/47846
PMID:38411999

©Johannes Benedict Oehm, Sarah Luise Riepenhausen, Michael Storck, Martin Dugas, Rüdiger Pryss, Julian Varghese. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and External Validation of Machine Learning Models for Diabetic Microvascular Complications: Cross-Sectional Study With Metabolites

Feng He[1,2], MSc; Clarissa Ng Yin Ling[1], MD; Simon Nusinovici[1,3], PhD; Ching-Yu Cheng[1,3], PhD; Tien Yin Wong[1,3], PhD; Jialiang Li[2*], PhD; Charumathi Sabanayagam[1,3*], PhD

[1]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore

[2]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore

[3]Ophthalmology and Visual Sciences Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Charumathi Sabanayagam, PhD
Singapore Eye Research Institute
Singapore National Eye Centre
The Academia, 20 College Road, Discovery Tower Level 6
Singapore, 169856
Singapore
Phone: 65 6576 7286
Fax: 65 6225 2568
Email: charumathi.sabanayagam@seri.com.sg

Abstract

Background: Diabetic kidney disease (DKD) and diabetic retinopathy (DR) are major diabetic microvascular complications, contributing significantly to morbidity, disability, and mortality worldwide. The kidney and the eye, having similar microvascular structures and physiological and pathogenic features, may experience similar metabolic changes in diabetes.

Objective: This study aimed to use machine learning (ML) methods integrated with metabolic data to identify biomarkers associated with DKD and DR in a multiethnic Asian population with diabetes, as well as to improve the performance of DKD and DR detection models beyond traditional risk factors.

Methods: We used ML algorithms (logistic regression [LR] with Least Absolute Shrinkage and Selection Operator and gradient-boosting decision tree) to analyze 2772 adults with diabetes from the Singapore Epidemiology of Eye Diseases study, a population-based cross-sectional study conducted in Singapore (2004-2011). From 220 circulating metabolites and 19 risk factors, we selected the most important variables associated with DKD (defined as an estimated glomerular filtration rate <60 mL/min/1.73 m^2) and DR (defined as an Early Treatment Diabetic Retinopathy Study severity level ≥20). DKD and DR detection models were developed based on the variable selection results and externally validated on a sample of 5843 participants with diabetes from the UK biobank (2007-2010). Machine-learned model performance (area under the receiver operating characteristic curve [AUC] with 95% CI, sensitivity, and specificity) was compared to that of traditional LR adjusted for age, sex, diabetes duration, hemoglobin A_{1c}, systolic blood pressure, and BMI.

Results: Singapore Epidemiology of Eye Diseases participants had a median age of 61.7 (IQR 53.5-69.4) years, with 49.1% (1361/2772) being women, 20.2% (555/2753) having DKD, and 25.4% (685/2693) having DR. UK biobank participants had a median age of 61.0 (IQR 55.0-65.0) years, with 35.8% (2090/5843) being women, 6.7% (374/5570) having DKD, and 6.1% (355/5843) having DR. The ML algorithms identified diabetes duration, insulin usage, age, and tyrosine as the most important factors of both DKD and DR. DKD was additionally associated with cardiovascular disease history, antihypertensive medication use, and 3 metabolites (lactate, citrate, and cholesterol esters to total lipids ratio in intermediate-density lipoprotein), while DR was additionally associated with hemoglobin A_{1c}, blood glucose, pulse pressure, and alanine. Machine-learned models for DKD and DR detection outperformed traditional LR models in both internal (AUC 0.838 vs 0.743 for DKD and 0.790 vs 0.764 for DR) and external validation (AUC 0.791 vs 0.691 for DKD and 0.778 vs 0.760 for DR).

XSL•FO
RenderX

Conclusions: This study highlighted diabetes duration, insulin usage, age, and circulating tyrosine as important factors in detecting DKD and DR. The integration of ML with biomedical big data enables biomarker discovery and improves disease detection beyond traditional risk factors.

(J Med Internet Res 2024;26:e41065) doi:10.2196/41065

KEYWORDS

machine learning; diabetic microvascular complication; diabetic kidney disease; diabetic retinopathy; biomarkers; metabolomics; complication; adult; cardiovascular disease; metabolites; biomedical big data; kidney disease

Introduction

Diabetes is a complex metabolic disorder and a major global health problem of our time. In 2021, it affected 536.6 million adults worldwide, with a projected surge to 783.2 million by 2045 [1]. With the rapidly growing population with diabetes and the greater longevity over time, the burden of associated complications is expected to increase in parallel [2]. Among these, diabetic kidney disease (DKD) was estimated to develop in around 40% of the population with diabetes [3], while diabetic retinopathy (DR) would manifest in approximately 35% [4]. Left undetected and untreated, these microvascular complications could substantially elevate the risk of cardiovascular disease (CVD), end-stage renal disease, and permanent vision loss, resulting in compromised quality of life and shortened life expectancy [2,5,6]. Yet, timely and accurate diagnosis of DKD and DR remains a challenge because of their asymptomatic progression in the early stages [7]. Although factors such as age, sex, diabetes duration, hemoglobin A_{1c} (HbA_{1c}) %, systolic blood pressure (SBP), and BMI have been identified as the risk contributors for DKD and DR, they offer only partial insights into the variability of risk among individuals.

To enable risk prediction and exploration of the underlying metabolic pathways, metabolomics has been increasingly used for biomarker discovery in diabetes and its complications [7-10]. Abnormalities in amino acids and lipids metabolism have been linked to both DKD and DR [7]. Given the similar microvascular structure, physiology, and pathogenic features of the kidney and the eye, these common metabolic traits may indicate biochemical pathways shared by the 2 complications, or both being manifestations of a latent systematic condition [11]. However, to the best of our knowledge, investigations into the commonalities and differences between DKD and DR in terms of their metabolic basis have been somewhat scarce [7,12]. Moreover, prior research was often limited by inadequate sample sizes, lack of replication, and restricted data analysis methods [13].

In this study, we aimed to fill these gaps by analyzing 220 circulating metabolites and 19 established risk factors as predictors of prevalent DKD and DR in a retrospective Asian adult population with diabetes. We used 2 machine learning (ML) algorithms, logistic regression (LR) with Least Absolute Shrinkage and Selection Operator (LASSO [14]), and a gradient-boosting decision tree (GBDT [15]). Compared with traditional statistical methods, these ML algorithms excel in handling high-dimensional data with complex relationships and can quantify the relative contribution of individual variables through variable importance scores. Based on these scores, we selected the top variables to develop ML models for DKD and DR detection and validated them externally in UK biobank (UKBB [16]) data. Additionally, we developed reference models using LR adjusted for traditional risk factors. All models were evaluated using the area under the receiver operating characteristic curve (AUC) with 95% CI, sensitivity, and specificity. Finally, we discussed the potential role of the selected metabolites in DKD and DR with reference to previous studies.

Methods

Data Sets

We derived this study's data from the Singapore Epidemiology of Eye Diseases study (SEED [17]), a population-based cross-sectional study conducted in Singapore from 2004 to 2011. Detailed methodology has been reported elsewhere [17]. In brief, we recruited 10,033 adults aged 40-80 years using age-stratified random sampling. Participants completed interviewer-administered questionnaires, underwent ocular examinations, and provided samples for biochemical laboratory tests. The cohort included 3280 Malay (2004-2006, response rate 78.7%), 3400 Indian (2007-2009, response rate 75.6%), and 3353 Chinese (2009-2011, response rate 72.8%) individuals.

For external validation, we used data from UKBB, an open-access resource of prospective data set collected in the UK from 2007 to 2010, with over 500,000 participants recruited between the ages of 40 to 69 [16].

Ethical Considerations

Both SEED and UKBB studies were conducted per the Declaration of Helsinki. Ethical approval was obtained from the SingHealth Institutional Review Board for SEED (2018/2717, 2018/2921, 2018/2006, 2018/2594, 2018/2570, 2015/2279, and 2012/487/A) and from North West Multi-centre Research Ethics Committee for UKBB (21/NW/0157). Written informed consent was provided by all participants during the primary data collection. Due to the retrospective nature of our study and the use of deidentified health information, the SingHealth Institutional Review Board approved this study without requiring additional patient consent.

Definition of Outcome and Variables

In SEED, diabetes was defined as meeting any of the following criteria: HbA_{1c}% >6.5, random blood glucose >11.1 mmol/L, self-reported physician-diagnosed diabetes, or the use of antidiabetic medication including insulin. In UKBB, we applied

XSL•FO
RenderX

the same definition but additionally included individuals with DR if the aforementioned variables were missing.

DKD was defined as an estimated glomerular filtration rate (eGFR) <60 mL/min/1.73m^2 in people with diabetes. The eGFR values were calculated from blood creatinine concentrations using the chronic kidney disease epidemiology collaboration equation [18].

For SEED participants with diabetes, DR severity in each eye was graded from fundus photographs by certified ophthalmic graders according to the standard protocol of Early Treatment Diabetic Retinopathy Study (ETDRS) [2,19] with 6 stages: no DR (level 10-20), minimal (level 20), mild (level 35), moderate (levels 43 to 47), severe (level 53), and proliferative DR (levels >60). We defined the DR outcome as having an ETDRS level ≥20 in at least one eye (ie, any DR). In UKBB, DR severity was not graded. Therefore, we identified DR cases based on *the International Classification of Diseases, Tenth Revision (ICD-10)* code "H36.0" in their health-related outcomes (UKBB Data-Field: 41270) [16]. Supplementary analysis in SEED also considered moderate or worse DR (level >43 in at least one eye) as an additional outcome.

We considered a total of 239 variables for biomarker selection (Multimedia Appendix 1). Among these, 19 variables were identified through a literature review, comprising 6 traditional risk factors (age, sex, duration of diabetes, HbA_{1c}%, SBP, and BMI), and 13 extended risk factors related to lifestyle (alcohol consumption and smoking), medication use (insulin, anticholesterol, and antihypertensive medication), clinic or biochemistry (diastolic blood pressure [DBP], pulse pressure [PP], random blood glucose, total cholesterol, high-density lipoprotein [HDL] cholesterol, and low-density lipoprotein cholesterol), and comorbidity conditions (hypertension and history of CVD). Hypertension in both cohorts was defined as self-reported physician-diagnosed hypertension, SBP >140 mm Hg, DBP >80 mm Hg, or the use of antihypertensive medication.

Using nuclear magnetic resonance techniques (Nightingale Health), we quantified the concentration of 228 circulating metabolites from patients' blood samples. Of these, glycerol, pyruvate, and glutamine were not available for Malay individuals, creatinine was used in eGFR calculation and DKD outcome definition, while 4 metabolites (total, HDL, low-density lipoprotein cholesterols, and random blood glucose) were duplicated with those measured in biochemistry tests. Therefore, these variables were excluded from the analysis, leaving us with 220 metabolites from 15 categories (amino acids, apolipoproteins, cholesterol, cholesterol esters, fatty acids, fluid balance, free cholesterols, glycolysis-related metabolites, inflammation, ketone bodies, triglycerides, lipoprotein particle sizes, lipoprotein subclasses, lipoprotein lipid ratios, and other lipids).

Statistical Analysis

Inclusion and Exclusion Criteria

From the initial SEED data set of 10,033 participants, we excluded those free of diabetes (n=7069), missing metabolomics profiles (n=179), or missing more than 10% (n=13) of the data, to obtain a final study population of 2772 individuals. Similarly, we identified 5843 UKBB participants eligible for external validation after data cleaning (Figure 1).

XSL•FO
RenderX

Figure 1. Exclusion criteria and machine learning workflow. AUC: area under the receiver operating characteristic curve; DKD: diabetic kidney disease; DR: diabetic retinopathy; eGFR: estimated glomerular filtration rate; LASSO: Least Absolute Shrinkage and Selection Operator; ROC: receiver operating characteristic.

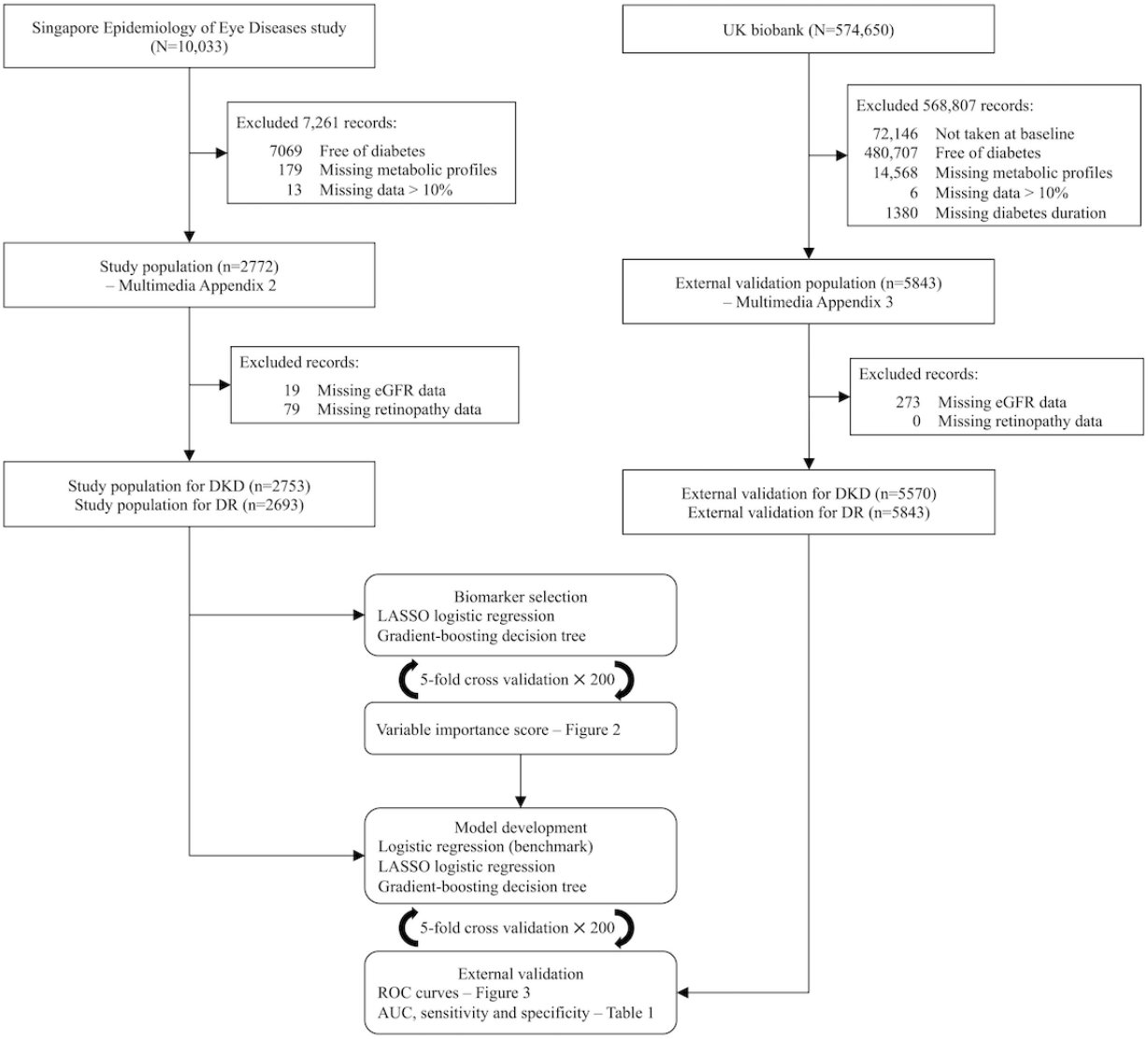

Descriptive Statistics

We categorized SEED participants into 4 groups based on their DKD and DR status, and summarized group characteristics as number (%), mean (SD), or median (IQR) as appropriate for each variable. Differences among groups were assessed using 2-sided *P* values derived from Pearson χ^2 tests or Fisher exact test for categorical variables, and Kruskal-Wallis rank sum tests for numeric variables. We also evaluated the interpopulation differences between SEED and UKBB using Pearson χ^2 tests and Mann-Whitney *U* tests as appropriate for each variable. Some subcategories may not add up due to the presence of missing data.

ML Algorithms

We used LASSO [14] and GBDT [15] to identify the key biomarkers and develop disease detection models for DR and DKD. LASSO is an extension of traditional LR that does not require the independence of variables. Therefore, LASSO is suitable for high-dimensional data sets where issues such as multicollinearity often arise. During parameter optimization, LASSO automatically shrinks the coefficients of the less-important variables to zero, while retaining nonzero coefficients for the important ones to achieve biomarker selection. Its strengths include relatively straightforward computation and parameter tuning compared to other ML algorithms. However, its scope is limited to examining only the linear associations between continuous variables and the log-odds. To account for possible nonlinear effects and variable interactions, we additionally implemented the GBDT algorithm, which constructs a sequence of interdependent decision trees to collectively make a decision. During the process, the algorithm assesses each variable's contribution to minimizing prediction errors, returning a score for its relative influence. We then applied a predefined threshold to select the most influential variables for disease detection based on their scores in each selection round. GBDT is known for its adaptability to various data distributions with generally strong performances, yet its

XSL•FO
RenderX

hyper-parameter tuning and computation can be rather time-consuming, with a higher chance of overfitting, and less transparency compared to LASSO.

Model Development and Validation

In SEED, we handled missing data by excluding variables and participants with high levels of missing information, yielding a final data set with each variable containing less than 6% (161/2772) missing data and a participant-level missing data percentage of 10% or less (23/239, Multimedia Appendix 2). Following this, missing data imputation was carried out by using mean values for numeric variables, and modes for categorical variables as appropriate. To mitigate selection bias arising from data division into training and validation sets, we averaged the results over 200 random repeats of 5-fold cross-validation. In each repeat, the imputed SEED data set was randomly divided into 5 subsets (ie, folds) of equal sample size by stratified sampling to ensure consistent case rates. Each fold (553/2772, 20% of data) took its turn as the validation set, while the remaining 4 folds (2219/2772, 80% of data) were used for model training and variable selection. From 200 replicates we generated 1000 sets of variable selection results, based on which we quantified each variable's contribution to the model's performance by calculating a variable importance score, as the variable's selection frequency during the repeated cross-validation process.

Next, we arranged the variables in descending order of their selection frequencies and included only those selected 990 times or more (ie, selection frequency 0.990) in the final model. To evaluate the performance of these new models, we performed another 200 random repeats of 5-fold cross-validation but used only the complete cases (ie, no missing data imputation). The final machine-learned models were compared with the multivariate LR models adjusted for the 6 established risk factors including age, sex, diabetes duration, HbA_{1c}%, SBP, and BMI. The performance of these models was evaluated in internal and external validation using AUC with 95% CI, sensitivity, and specificity. We reported sensitivity and specificity at the optimal threshold, where sensitivity equals specificity. Additionally, in the context of DR and DKD detection, where the cost of false negatives is generally higher than that of false positives, we prioritized higher sensitivity over higher specificity by setting the probability threshold at 0.8 sensitivity to compare the specificity. The AUC difference between machine-learned models and the traditional model was evaluated using the test by DeLong et al [20]. For the supplementary analysis, we examined the importance of the top machine-learned variables within nested models, to check for any marginal increase in AUC with additional variables. We also calculated the AUC for ML models with all the variables.

Metabolites Selection

Finally, we focused on the metabolites that were consistently selected by both ML algorithms. We quantified their associations with DKD and DR respectively, using odds ratios (OR) per SD increment with 95% CI, and *P* values from multivariate LR models adjusted for age, sex, diabetes duration, insulin use, HbA_{1c}%, PP, BMI, cholesterol, and HDL cholesterol.

We conducted all the analyses in R version 4.0.2. (R Foundation) and defined statistical significance as $P<.05$.

Results

Population Characteristics

Figure 1 illustrates the data cleaning and analysis workflow. In the SEED population with diabetes, 2674 people had information on both disease outcomes (Multimedia Appendix 3). Of these, 1657 (62%) had neither DKD nor DR, 338 (12.6%) had DKD but not DR, 496 (18.5%) had DR but not DKD, and 183 (6.8%) had both DKD and DR. People with DKD, regardless of their DR status, tended to be older, with higher PP, higher SBP, and lower DBP. They also had higher levels of HDL cholesterol, more history of hypertension and CVD, and were more likely to have used anticholesterol medication and antihypertensive medication. However, they had a lower smoking rate and reported less alcohol consumption than those without DKD. People with DR, regardless of their DKD status, tended to have a longer duration of diabetes, with the use of antidiabetic medication and insulin, higher HbA_{1c}%, and random blood glucose levels. They also had a lower BMI and lower total cholesterol levels. For people with both complications, the abovementioned characteristics further differed from those with neither complication. However, in terms of sex distribution, no significant difference was observed ($P=.09$).

We also found differences between SEED and UKBB in terms of demographics, lifestyle factors, biochemical laboratory results, and medical history (Multimedia Appendix 4). In particular, UKBB participants had a lower prevalence of both DKD (374/5570, 6.7%) and DR (355/5843, 6.1%) as compared to SEED (DKD prevalence: 555/2753, 20.2%, and DR prevalence: 685/2693, 25.4%), with only 0.7% (41/5570) having both complications. SEED had a median age of 61.7 (IQR 53.5-69.4) years and 49.1% (1361/2772) women, whereas UKBB had a median age of 61.0 (IQR 55.0-65.0) years and 35.8% (2090/5843) women. SEED was a multiethnic sample based in Singapore with 36.8% (1020/2772) Malay, 45.5% (1262/2772) Indian, and 17.7% (490/2772) Chinese individuals. In contrast, over 81.8% (4778/5843) of the UKBB participants were British, with the rest being Indian (217/5843, 3.7%), Irish (132/5843, 2.3%), Caribbean (93/5843, 1.6%), African (85/5843, 1.5%), or other ethnicities (538/5843, 9.2%).

Model Development and Validation

Figure 2 shows the top 50 variables arranged in descending orders of their variable importance. For DKD detection, LASSO identified 15 variables with a frequency exceeding the cutoff, while GBDT identified 13, so we used the top 15 for deriving the corresponding DKD models; then for DR, LASSO identified 10 variables, while GBDT identified 6, so we used the top 10 for DR.

XSL•FO
RenderX

Figure 2. Top 50 machine-learned variables for DKD and DR. %: metabolites to total lipids ratio; anti-HTN: antihypertensive medication use; ApoA1: apolipoprotein A1; C: cholesterol; CE: cholesterol esters; CVD: cardiovascular disease; DBP: diastolic blood pressure; DKD: diabetic kidney disease; DHA: docosahexaenoic acid; DHAFA: antihypertensive medication use; DR: diabetic retinopathy; FC: free cholesterol; GBDT: gradient-boosting decision tree; HDL: high-density lipoprotein; HDL3: high-density lipoprotein 3; IDL: intermediate-density lipoprotein; L: large; LA: linoleic acid; LASSO: Least Absolute Shrinkage and Selection Operator; LDL: low-density lipoprotein; M: medium; MUFA: monounsaturated fatty acids; P: particles; PL: phospholipids; PUFA: polyunsaturated fatty acids; S: small; SBP: systolic blood pressure; SFA: saturated fatty acids; TG: triglycerides; VLDL: very-low-density lipoprotein; XL: very large; XS: very small; XXL: extremely large.

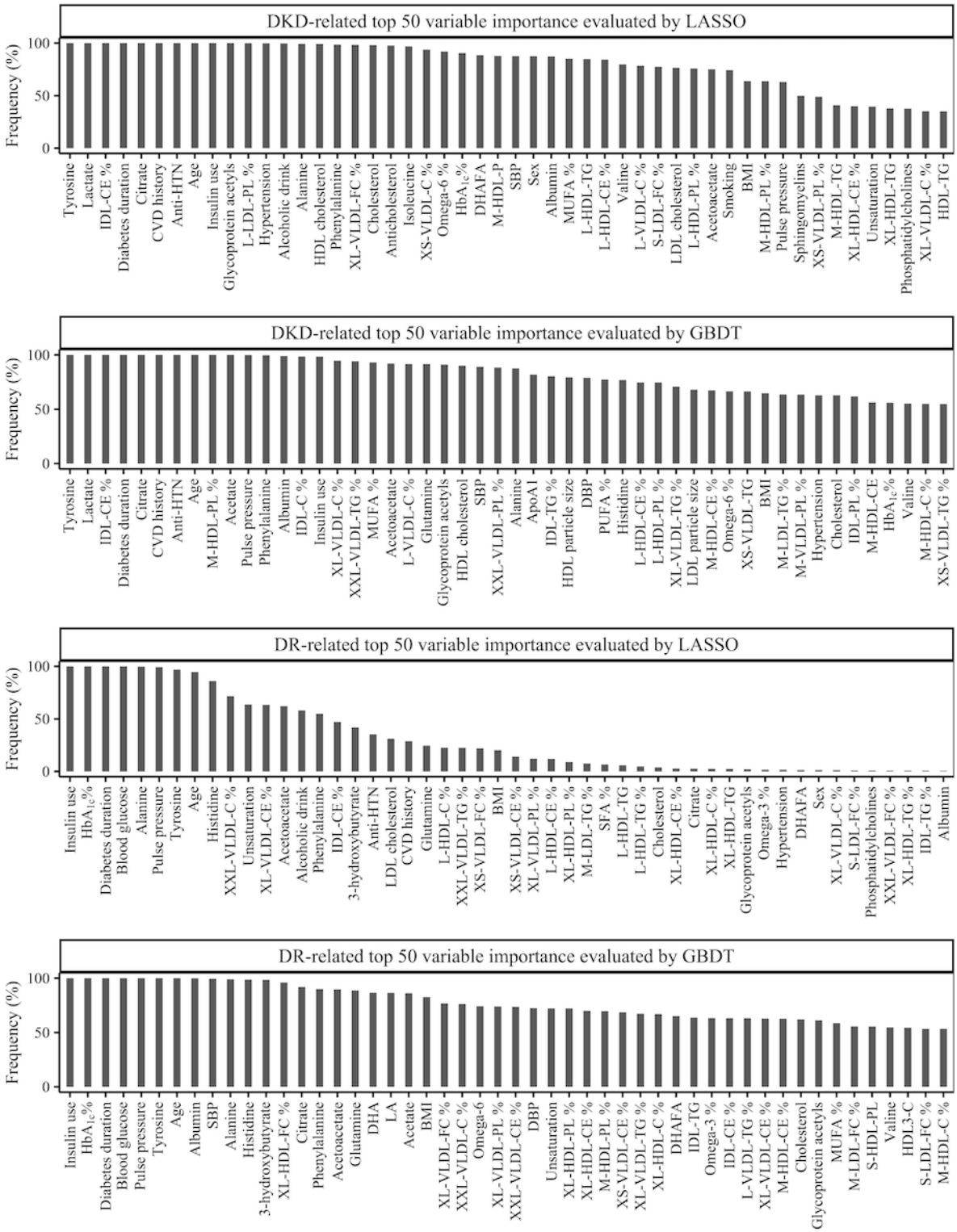

In the internal validation, GBDT models performed the best with an AUC of 0.838 for DKD and 0.790 for DR, followed by LASSO with AUC values of 0.832 and 0.779, respectively. In contrast, LR only achieved AUC scores of 0.743 for DKD, and 0.764 for DR. In the external validation using UKBB data, LASSO models exhibited the best performance with AUC values of 0.791 for DKD, and 0.778 for DR. GBDT models achieved

XSL•FO
RenderX

AUC scores of 0.738 and 0.778, respectively, while LR resulted in AUC values of 0.691 and 0.760 (Figure 3).

Figure 3. Receiver operating characteristic curves showing the model performances. AUC: area under the receiver operating characteristic curve; DKD: diabetic kidney disease; DR: diabetic retinopathy; GBDT: gradient-boosting decision tree; LASSO: Least Absolute Shrinkage and Selection Operator; LR: logistic regression; SEED: Singapore Epidemiology of Eye Diseases; UKBB: UK biobank.

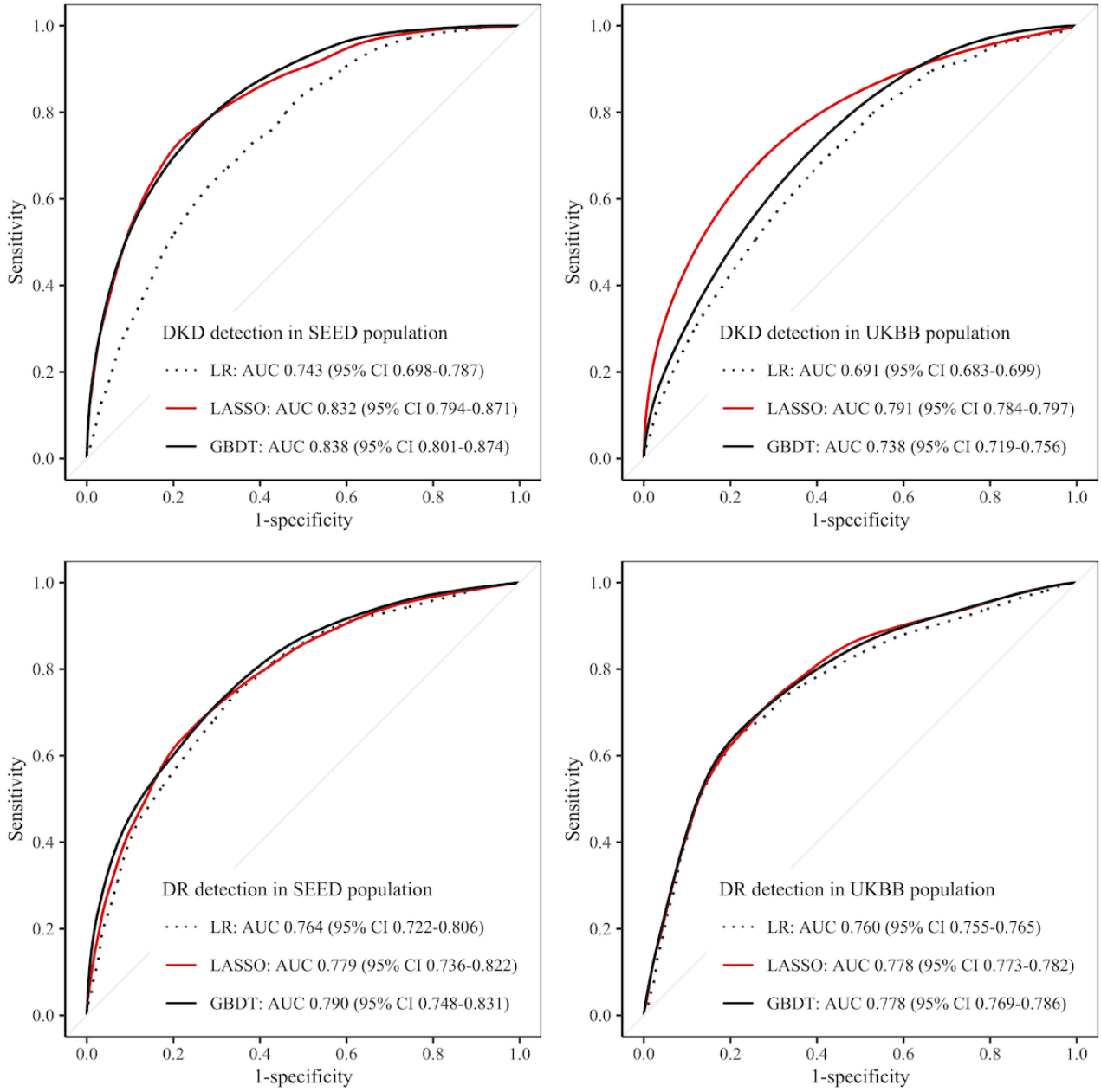

Further tests confirmed that the AUC scores of ML models were significantly higher than those obtained using traditional LR (2-sided $P<.001$). In terms of sensitivity and specificity at the optimal threshold (where sensitivity=specificity), LASSO and GBDT achieved comparable performance in internal validation (for DKD, 0.757 by LASSO vs 0.751 by GBDT; for DR, 0.708 by LASSO vs 0.709 by GBDT), and both were superior to LR (0.674 for DKD and 0.696 for DR). In external validation, LASSO performed the best, with 0.723 for DKD and 0.716 for DR. At 0.8 sensitivity, LASSO achieved 0.636 specificity for DKD detection in UKBB, and 0.617 for DR, outperforming the other 2 models (Table 1).

XSL•FO
RenderX

Table 1. Performance evaluation of machine-learned models.

Disease and data set	Algorithm	AUC[a] (95% CI)	Optimal SN[b]/SP[c]	SP at SN=0.8	Sample size, n[d]	Cases, n (%)
DKD[e]						
SEED[f]	LR[g]	0.743 (0.698-0.787)	0.674	0.54	2653	517 (19.5)
SEED	LASSO[h]	0.832 (0.794-0.871)	0.757	0.703	2666	532 (20)
SEED	GBDT[i]	0.838 (0.801-0.874)	0.751	0.709	2668	529 (19.8)
UKBB[j]	LR	0.691 (0.683-0.699)	0.635	0.472	5236	345 (6.6)
UKBB	LASSO	0.791 (0.784-0.797)	0.723	0.636	5090	333 (6.5)
UKBB	GBDT	0.738 (0.719-0.756)	0.666	0.535	5543	366 (6.6)
DR[k]						
SEED	LR	0.764 (0.722-0.806)	0.696	0.594	2597	653 (25.1)
SEED	LASSO	0.779 (0.736-0.822)	0.708	0.596	2514	628 (25)
SEED	GBDT	0.790 (0.748-0.831)	0.709	0.616	2598	655 (25.2)
UKBB	LR	0.760 (0.755-0.765)	0.707	0.571	5492	336 (6.1)
UKBB	LASSO	0.778 (0.773-0.782)	0.716	0.617	4678	280 (6)
UKBB	GBDT	0.778 (0.769-0.786)	0.715	0.604	4833	296 (6.1)

[a]AUC: area under the receiver operating characteristic curve.
[b]SN: sensitivity.
[c]SP: specificity.
[d]Final sample size after variable selection and missing data removal.
[e]DKD: diabetic kidney disease.
[f]SEED: Singapore Epidemiology of Eye Diseases.
[g]LR: logistic regression.
[h]LASSO: Least Absolute Shrinkage and Selection Operator.
[i]GBDT: gradient-boosting decision tree.
[j]UKBB: UK biobank.
[k]DR: diabetic retinopathy.

In the sensitivity analysis, we evaluated the potential improvement in AUC by introducing additional variables in nested models (Multimedia Appendix 5). The results supported the effectiveness of using the top 10 variables for detecting DR in SEED. In line with this, using all 239 variables did not improve beyond what was achieved with the top 10 in both internal validation (LASSO 0.776 and GBDT 0.783) and external validation (LASSO 0.754 and GBDT 0.774). Yet for DKD, incorporating additional variables beyond the top 15 resulted in improved AUC in SEED. Specifically, using all 239 variables yielded higher AUC values of 0.859 for LASSO and 0.842 for GBDT. However, this increase may be attributed to overfitting, as a similar pattern was not observed in external validation. Further, using all available variables resulted in lower AUC values for both LASSO (0.694) and GBDT (0.721) in the UKBB cohort. Consequently, we decided to maintain the DKD model with the top 15 variables for its simplicity and effectiveness.

Metabolites Selection

For DKD, both LASSO and GBDT selected 5 risk factors (duration of diabetes, history of CVD, antihypertensive medication use, age, and insulin use) and 4 metabolites (tyrosine, lactate, cholesterol esters to total lipid ratio in intermediate-density lipoprotein particles [IDL-CE%], and citrate). For any DR, both algorithms identified 6 risk factors (insulin use, HbA_{1c}%, duration of diabetes, random blood glucose, age, and PP) and 2 metabolites (tyrosine and alanine). In the supplementary analysis, tyrosine was again selected for moderate or worse DR detection (Multimedia Appendix 6).

Table 2 shows the association of the machine-learned metabolites with DKD and DR in multivariable LR models. We found tyrosine to be negatively associated with both DKD (OR 0.65, 95% CI 0.58-0.73; P<.001) and DR (OR 0.90, 95% CI 0.81-1.00; P=.047). High levels of alanine were associated with increased DR prevalence (OR 1.31, 95% CI 1.18-1.45; P<.001), but with decreased DKD prevalence (OR 0.72, 95% CI 0.64-0.80; P<.001). Similarly, a high level of lactate was linked to higher DR prevalence (OR 1.16, 95% CI 1.05-1.28; P=.004) but with lower DKD prevalence (OR 0.71, 95% CI 0.63-0.80; P<.001). Finally, a high level of citrate was associated with increased DKD prevalence (OR 1.90, 95% CI 1.70-2.12; P<.001), while high IDL-CE% was linked to decreased DKD prevalence (OR 0.46, 95% CI 0.40-0.53; P<.001).

XSL•FO
RenderX

Table 2. Association of selected metabolites with DKD[a] and DR[b].

Metabolite	DKD detection		DR detection	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Alanine	0.72 (0.64-0.80)	<.001	1.31 (1.18-1.45)	<.001
Tyrosine	0.65 (0.58-0.73)	<.001	0.90 (0.81-1.00)	.047
Citrate	1.90 (1.70-2.12)	<.001	0.93 (0.84-1.03)	.2
IDL-CE%[d]	0.46 (0.40-0.53)	<.001	0.89 (0.79-1.01)	.1
Lactate	0.71 (0.63-0.80)	<.001	1.16 (1.05-1.28)	.004

[a]DKD: diabetic kidney disease.

[b]DR: diabetic retinopathy.

[c]OR: odds ratio.

[d]IDL-CE%: cholesterol esters to total lipid ratio in intermediate-density lipoprotein particles.

Discussion

Principal Findings

ML selected age, use of insulin, duration of diabetes, and circulating tyrosine as the most important markers for DKD and DR detection in the SEED population with diabetes. Additionally, DKD was associated with the use of antihypertensive medications, CVD history, and 3 metabolites (lactate, citrate, and IDL-CE%), whereas DR was additionally linked to HbA_{1c}, random blood glucose, PP, and alanine.

The ML models developed in the SEED cohort with diabetes were externally validated using UKBB data. In both cohorts, ML models outperformed the traditional LR in terms of AUC, sensitivity, and specificity, demonstrating their potential to discover novel biomarkers and enable disease screening when integrated with health care and metabolite data.

Comparison With Prior Work

Our main data set included a comprehensive set of 19 risk factors and 220 circulating metabolites measured in 2772 individuals. The detailed patient profiling with a robust sample size allowed an opportunity to identify the markers most relevant to DKD and DR, offering insights into the systematic alteration of metabolism and underlying pathways.

In line with the previous literature [4,13], ML consistently identified 3 key factors—diabetes duration, age, and the use of insulin—as the top risk factors for both DKD and DR. However, ML also revealed novel aspects concerning some established risk factors for these conditions. For instance, our ML models exhibited a notable preference for PP over SBP and DBP in DR detection, supporting the study by Yamamoto et al [21] that PP, as a surrogate marker of arterial stiffness, reflected both the SBP elevation and DBP reduction, thereby carrying more predictive information for DR than other blood pressure metrics. Another example was assessing the relative importance of glycemia control indicators, where HbA_{1c}% consistently received a higher selection frequency in our ML models compared to random blood glucose levels. This was probably because HbA_{1c}% was averaged to reflect a mean shift with much less random noise, while random blood glucose data might carry more noise from life cycle changes and interindividual variability [22]. Interestingly, some well-established risk factors usually included such as sex did not appear in the top-ranking lists by ML, although this variable had been selected by traditional LR models on the same population in previous studies [23]. This could be because sex is an intrinsic component of other phenotypes. For instance, male sex was associated with CVD [24], a condition well-known to be linked to DKD and DR [2].

We noted high levels of tyrosine, an aromatic amino acid, to be negatively associated with the prevalence of DKD and DR, supporting the ADVANCE trial where increased tyrosine concentration was linked to a decreased risk of diabetic microvascular events (hazard ratio 0.78, 95% CI 0.67-0.91) [12]. Tyrosine is mainly synthesized from phenylalanine hydroxylation in the liver and kidney; impaired kidney function is therefore associated with reduced phenylalanine hydroxylase activity characterized by low blood tyrosine levels [12]. Additionally, tyrosine serves as a precursor to catecholamine neurotransmitters (dopamine, norepinephrine, and epinephrine), and plays a pivotal role in central nervous system functions and activities [25]. In metabolic disorders such as diabetes, reduced blood tyrosine can affect its uptake into the brain and also the synthesis and release of transmitters, thereby altering hormonal function, affective state, and blood pressure [25], potentially linked to higher microvascular risks.

Another amino acid selected by ML was alanine, and its higher concentration in blood was associated with higher DR prevalence, but lower DKD prevalence. In the ADVANCE trial, a negative association was reported between circulating alanine and an aggregated microvascular outcome, defined as new or worsening nephropathy or retinopathy (hazard ratio 0.86, 95% CI 0.76-0.98) [12]. In other research, high levels of blood amino acids such as alanine have been linked to inhibited insulin signaling to glucose transport, phosphorylation, and glycogen synthesis, causing insulin resistance, which impairs hepatic mitochondrial function in patients with diabetes and contributes to diabetes microvascular complications such as DR [26,27].

Among the ML-selected metabolites for DKD, high levels of lactate were found to be associated with lower DKD prevalence. However, another study conducted in the DKD population suggested an accumulation of acid due to abnormal lactate

XSL•FO
RenderX

metabolism causing fibrosis and mitochondrial abnormalities, leading to further kidney damage [28]. Nevertheless, further investigations are required to understand the complex association between circulating lactate and DKD. Another metabolite identified was citrate, which showed a positive association with DKD in SEED. In people with DKD, an elevated plasma citrate level has been linked to reduced organic anion transport and dysregulated mitochondrial functions of the kidney tissues [8]. Our study also highlighted the significance of IDL-CE% in diabetic microvascular complications, revealing its inverse association with prevalent DKD in the SEED population with diabetes. Aberrations in lipoprotein composition have been reported to be indicative of insulin resistance and impaired glucose tolerance in the general population, known to cause future diabetes [9,27]. In people with type 1 DKD, abnormality in lipoproteins clearance was further linked to impaired kidney function [29].

For biomarker discovery, traditional studies often relied on LR models to examine metabolites one by one separately [12,30], with stringent model assumptions and multiple testing corrections [9,31]. Herein, our ML approach was more efficient in the sense that it simultaneously examines all variables for potential associations. While LASSO was limited to detecting linear associations, we implemented GBDT as a complementary model to additionally assess nonlinear terms and high-order interactions. As was shown in Figure 2, DR-related variables in GBDT had higher selection frequencies than in LASSO, suggesting the existence of such complex associations. Still, LASSO models achieved superior performance in external validation, indicating a prominent contribution of linear associations to DKD and DR detection. Another highlight of our methodology was the repeated cross-validation, which reduced the influence of potential outliers and ensured the randomness of sampling, thereby generating results more robust than those relying on a fixed training set. This approach also allowed us to quantify individual contributions of variables based on their selection frequencies. However, alternative definitions of variable importance scores [32] may be considered in a future study to potentially refine the variable selection process.

Globally, three-quarters of those with diabetes live in low- and middle-income countries, particularly India and China [33]. While the risk profile of Asian patients with diabetes differs from that in high-income “Western” societies in terms of age, BMI, lifestyle, diet, and many other aspects [33], there has been limited discussion on Asian populations with diabetes. Our study, conducted in Asian populations (Malay, Indian, and Chinese), may contribute to the topic by allowing an opportunity to identify the commonalities and differences between DKD and DR in terms of circulating metabolic traits, offering insights into the systematic alteration of metabolism in diabetes.

Limitations

In total, 1 limitation of this study was that we did not separate study subjects by diabetes type. Since around 95% of the SEED population with diabetes had type 2 diabetes, our results would mainly reflect the variable associations with type 2 diabetes. While these associations hold potential for hypothesis generation and disease detection, the cross-sectional nature of our study highlights the need for caution in drawing causal inferences. Future longitudinal studies are warranted to establish temporal associations based on our current findings.

Another issue was data availability—albuminuria, an important indicator of kidney disease [13], and 3 metabolites (pyruvate, glycerol, and glutamine) were missing in Malay individuals of the SEED population. Hence, these variables were excluded from the analysis. Certain diabetic medications were reported to potentially interfere with circulating metabolite levels [12,34,35]. However, we could only account for the use of insulin due to limited data availability on UKBB medication profiles. Furthermore, we noted a difference in defining DR between our development data set (SEED), which used the standard ETDRS classification system, and the external test set (UKBB), where DR cases were identified using ICD-10 codes. This variance in outcome definition might have introduced some degree of misclassification in UKBB. Additionally, due to the absence of ETDRS-based severity scores in UKBB, validation of the supplementary model for moderate or worse DR in supplementary analysis was not feasible. Nevertheless, it was noteworthy that the features identified in SEED remained relevant for enhancing DR detection in the UKBB data set. However, we acknowledge that the discrepancy in DR definition underscores the need for caution when interpreting results, particularly in the context of clinical or practical applications. Future research may benefit from adopting a uniform approach to DR classification to minimize such limitations and enhance result validity.

Conclusions

In conclusion, current ML models developed using the SEED population with diabetes and subsequently validated in UKBB showed superior performance compared to traditional LR for DKD and DR detection. ML highlighted age, use of insulin, diabetes duration, and tyrosine as the most influential factors in DKD and DR. Additionally, DKD was associated with high levels of citrate, low levels of lactate, and low IDL-CE%; while DR was linked to higher levels of alanine. The integration of ML with health care data and metabolomics could facilitate biomarker selection and enable disease screening.

Acknowledgments

We thank all the participants of Singapore Epidemiology of Eye Diseases (SEED) and UK biobank (UKBB) for contributing their health profiles to support our research. ChatGPT (OpenAI) [36] was used solely for proofreading purposes, focusing on grammar and spelling corrections. All such revisions were reviewed and validated by human coauthors. This study was supported by the Singapore Ministry of Health’s National Medical Research Council (NMRC/StaR/016/2013, NMRC/CIRG/1371/2013,

XSL•FO
RenderX

NMRC/CIRG/1417/2015, and OFLCG/001/2017). The funders had no role in study design; data collection, management, analysis, and interpretation; preparation, and review of this paper; or decision to publish.

Data Availability

As the Singapore Epidemiology of Eye Diseases (SEED) study involves human participants, the data cannot be made freely available in this paper, the supplemental files, or a public repository due to ethical restrictions. Nevertheless, the data are available from the Singapore Eye Research Institutional Ethics Committee for researchers who meet the criteria for access to confidential data. Interested researchers can send data access requests to the Singapore Eye Research Institute. The UK biobank (UKBB) test data set was obtained from the UKBB (application 45925). Data cannot be shared publicly due to the violation of patient privacy and the absence of informed consent for data sharing.

Authors' Contributions

All authors contributed to the intellectual development of this paper. JL and CS designed this study and supervised data analysis. FH and CNYL wrote the initial draft. FH performed the statistical analyses. SN, CC, TYW, JL, and CS assisted in the interpretation of the analyzed data and provided critical corrections to this paper. CS is the guarantor of this work and as such had full access to all the data in this study and takes responsibility for the integrity of the data and accuracy of the data analysis. The final version of this paper has been seen and approved by all the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of 239 Singapore Epidemiology of Eye Diseases (SEED) variables included for machine learning feature selection.
[PDF File (Adobe PDF File), 77 KB - jmir_v26i1e41065_app1.pdf]

Multimedia Appendix 2
Distribution of missing data in Singapore Epidemiology of Eye Diseases (SEED), column-wise and row-wise histograms.
[PDF File (Adobe PDF File), 6 KB - jmir_v26i1e41065_app2.pdf]

Multimedia Appendix 3
Population characteristics of Singapore Epidemiology of Eye Diseases (SEED) stratified by diabetic kidney disease (DKD) and diabetic retinopathy (DR) status.
[PDF File (Adobe PDF File), 86 KB - jmir_v26i1e41065_app3.pdf]

Multimedia Appendix 4
Comparative analysis of population characteristics: Singapore Epidemiology of Eye Diseases (SEED) vs UK biobank.
[PDF File (Adobe PDF File), 77 KB - jmir_v26i1e41065_app4.pdf]

Multimedia Appendix 5
Area under the receiver operating characteristic curve performance of nested models in internal validation with varying numbers of machine-learned variables.
[PDF File (Adobe PDF File), 2656 KB - jmir_v26i1e41065_app5.pdf]

Multimedia Appendix 6
Top-50 machine-learned variables for moderate or worse diabetic retinopathy (DR) detection.
[PDF File (Adobe PDF File), 37 KB - jmir_v26i1e41065_app6.pdf]

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022;183:109119. [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
2. Sabanayagam C, Chee ML, Banu R, Cheng CY, Lim SC, Tai ES, et al. Association of diabetic retinopathy and diabetic kidney disease with all-cause and cardiovascular mortality in a multiethnic Asian population. JAMA Netw Open 2019;2(3):e191540 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.1540] [Medline: 30924904]

XSL•FO
RenderX

3. de Boer IH, Rue TC, Hall YN, Heagerty PJ, Weiss NS, Himmelfarb J. Temporal trends in the prevalence of diabetic kidney disease in the United States. JAMA 2011;305(24):2532-2539 [FREE Full text] [doi: 10.1001/jama.2011.861] [Medline: 21693741]
4. Yau JWY, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012;35(3):556-564 [FREE Full text] [doi: 10.2337/dc11-1909] [Medline: 22301125]
5. Valencia WM, Florez H. How to prevent the microvascular complications of type 2 diabetes beyond glucose control. BMJ 2017;356:i6505. [doi: 10.1136/bmj.i6505] [Medline: 28096078]
6. Eid S, Sas KM, Abcouwer SF, Feldman EL, Gardner TW, Pennathur S, et al. New insights into the mechanisms of diabetic complications: role of lipids and lipid metabolism. Diabetologia 2019;62(9):1539-1549 [FREE Full text] [doi: 10.1007/s00125-019-4959-1] [Medline: 31346658]
7. Wu T, Qiao S, Shi C, Wang S, Ji G. Metabolomics window into diabetic complications. J Diabetes Investig 2018;9(2):244-255 [FREE Full text] [doi: 10.1111/jdi.12723] [Medline: 28779528]
8. Sharma K, Karl B, Mathew AV, Gangoiti JA, Wassel CL, Saito R, et al. Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. J Am Soc Nephrol 2013;24(11):1901-1912 [FREE Full text] [doi: 10.1681/ASN.2013020126] [Medline: 23949796]
9. Ahola-Olli AV, Mustelin L, Kalimeri M, Kettunen J, Jokelainen J, Auvinen J, et al. Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four finnish cohorts. Diabetologia 2019;62(12):2298-2309 [FREE Full text] [doi: 10.1007/s00125-019-05001-w] [Medline: 31584131]
10. Quek DQY, He F, Sultana R, Banu R, Chee ML, Nusinovici S, et al. Novel serum and urinary metabolites associated with diabetic retinopathy in three Asian cohorts. Metabolites 2021;11(9):614 [FREE Full text] [doi: 10.3390/metabo11090614] [Medline: 34564429]
11. Wong CW, Wong TY, Cheng CY, Sabanayagam C. Kidney and eye diseases: common risk factors, etiological mechanisms, and pathways. Kidney Int 2014;85(6):1290-1302 [FREE Full text] [doi: 10.1038/ki.2013.491] [Medline: 24336029]
12. Welsh P, Rankin N, Li Q, Mark PB, Würtz P, Ala-Korpela M, et al. Circulating amino acids and the risk of macrovascular, microvascular and mortality outcomes in individuals with type 2 diabetes: results from the ADVANCE trial. Diabetologia 2018;61(7):1581-1591 [FREE Full text] [doi: 10.1007/s00125-018-4619-x] [Medline: 29728717]
13. Colhoun HM, Marcovecchio ML. Biomarkers of diabetic kidney disease. Diabetologia 2018;61(5):996-1011 [FREE Full text] [doi: 10.1007/s00125-018-4567-5] [Medline: 29520581]
14. Hastie T, Qian J, Tay K. An introduction to glmnet. In: CRAN R Repositary. New York, NY: Springer; 2021:1-14.
15. Greenwell B, Boehmke B, Cunningham J, GBMD. GBM: Generalized Boosted Regression Models. R Package Version 2.1.8; 2020. 2021. URL: https://www.rdocumentation.org/packages/gbm/versions/2.1.8 [accessed 2024-03-06]
16. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779 [FREE Full text] [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
17. Fang XL, Chong CCY, Thakur S, Da Soh Z, Teo ZL, Majithia S, et al. Ethnic differences in the incidence of pterygium in a multi-ethnic Asian population: the Singapore epidemiology of eye diseases study. Sci Rep 2021;11(1):501 [FREE Full text] [doi: 10.1038/s41598-020-79920-9] [Medline: 33436813]
18. Levey AS, Stevens LA, Schmid CH, Zhang YC, Castro III AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
19. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House Classification. ETDRS report number 10. Ophthalmology 1991;98(5 Suppl):786-806. [Medline: 2062513]
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837-845. [doi: 10.2307/2531595]
21. Yamamoto M, Fujihara K, Ishizawa M, Osawa T, Kaneko M, Ishiguro H, et al. Pulse pressure is a stronger predictor than systolic blood pressure for severe eye diseases in diabetes mellitus. J Am Heart Assoc 2019;8(8):e010627 [FREE Full text] [doi: 10.1161/JAHA.118.010627] [Medline: 30971163]
22. Cohen RM, Lindsell CJ. When the blood glucose and the HbA1c don't match: turning uncertainty into opportunity. Diabetes Care 2012;35(12):2421-2423 [FREE Full text] [doi: 10.2337/dc12-1479] [Medline: 23173128]
23. Wong TY, Cheung N, Tay WT, Wang JJ, Aung T, Saw SM, et al. Prevalence and risk factors for diabetic retinopathy: the Singapore Malay Eye Study. Ophthalmology 2008;115(11):1869-1875. [doi: 10.1016/j.ophtha.2008.05.014] [Medline: 18584872]
24. Bots SH, Peters SAE, Woodward M. Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. BMJ Glob Health 2017;2(2):e000298 [FREE Full text] [doi: 10.1136/bmjgh-2017-000298] [Medline: 28589033]
25. Fernstrom JD. Branched-chain amino acids and brain function. J Nutr 2005;135(6 Suppl):1539S-1546S [FREE Full text] [doi: 10.1093/jn/135.6.1539S] [Medline: 15930466]

XSL•FO
RenderX

26. Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF, et al. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. Cell Metab 2009;9(4):311-326 [FREE Full text] [doi: 10.1016/j.cmet.2009.02.002] [Medline: 19356713]
27. Gancheva S, Jelenik T, Álvarez-Hernández E, Roden M. Interorgan metabolic crosstalk in human insulin resistance. Physiol Rev 2018;98(3):1371-1415 [FREE Full text] [doi: 10.1152/physrev.00015.2017] [Medline: 29767564]
28. Lee DY, Kim JY, Ahn E, Hyeon JS, Kim GH, Park KJ, et al. Associations between local acidosis induced by renal LDHA and renal fibrosis and mitochondrial abnormalities in patients with diabetic kidney disease. Transl Res 2022;249:88-109 [FREE Full text] [doi: 10.1016/j.trsl.2022.06.015] [Medline: 35788054]
29. Sibley SD, Hokanson JE, Steffes MW, Purnell JQ, Marcovina SM, Cleary PA, et al. Increased small dense LDL and intermediate-density lipoprotein with albuminuria in type 1 diabetes. Diabetes Care 1999;22(7):1165-1170 [FREE Full text] [doi: 10.2337/diacare.22.7.1165] [Medline: 10388983]
30. Mutter S, Valo E, Aittomäki V, Nybo K, Raivonen L, Thorn LM, et al. Urinary metabolite profiling and risk of progression of diabetic nephropathy in 2670 individuals with type 1 diabetes. Diabetologia 2022;65(1):140-149 [FREE Full text] [doi: 10.1007/s00125-021-05584-3] [Medline: 34686904]
31. Shrier I, Pang M. Confounding, effect modification, and the odds ratio: common misinterpretations. J Clin Epidemiol 2015;68(4):470-474 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.12.012] [Medline: 25662008]
32. Williamson BD, Gilbert PB, Carone M, Simon N. Nonparametric variable importance assessment using machine learning techniques. Biometrics 2021;77(1):9-22 [FREE Full text] [doi: 10.1111/biom.13392]
33. Wong TY, Sabanayagam C. Strategies to tackle the global burden of diabetic retinopathy: from epidemiology to artificial intelligence. Ophthalmologica 2020;243(1):9-20 [FREE Full text] [doi: 10.1159/000502387] [Medline: 31408872]
34. Seheult J, Fitzpatrick G, Boran G. Lactic acidosis: an update. Clin Chem Lab Med 2017;55(3):322-333 [FREE Full text] [doi: 10.1515/cclm-2016-0438] [Medline: 27522622]
35. Bellomo R. Bench-to-bedside review: lactate and the kidney. Crit Care 2002;6(4):322-326 [FREE Full text] [doi: 10.1186/cc1518] [Medline: 12225607]
36. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-03-15]

Abbreviations

AUC: area under the receiver operating characteristic curve
CVD: cardiovascular disease
DBP: diastolic blood pressure
DKD: diabetic kidney disease
DR: diabetic retinopathy
eGFR: estimated glomerular filtration rate
ETDRS: Early Treatment Diabetic Retinopathy Study
GBDT: gradient-boosting decision tree
HbA1c: hemoglobin A1c
HDL: high-density lipoprotein
ICD-10: International Classification of Diseases, Tenth Revision
IDL-CE%: cholesterol esters to total lipid ratio in intermediate-density lipoprotein particles
LASSO: Least Absolute Shrinkage and Selection Operator
LR: logistic regression
ML: machine learning
OR: odds ratio
PP: pulse pressure
SBP: systolic blood pressure
SEED: Singapore Epidemiology of Eye Diseases
UKBB: UK biobank

XSL•FO
RenderX

Edited by A Mavragani; submitted 18.01.23; peer-reviewed by KH Hung, G Lim; comments to author 25.09.23; revised version received 12.10.23; accepted 19.12.23; published 28.03.24.

Please cite as:
He F, Ng Yin Ling C, Nusinovici S, Cheng CY, Wong TY, Li J, Sabanayagam C
Development and External Validation of Machine Learning Models for Diabetic Microvascular Complications: Cross-Sectional Study With Metabolites
J Med Internet Res 2024;26:e41065
URL: https://www.jmir.org/2024/1/e41065
doi:10.2196/41065
PMID:38546730

©Feng He, Clarissa Ng Yin Ling, Simon Nusinovici, Ching-Yu Cheng, Tien Yin Wong, Jialiang Li, Charumathi Sabanayagam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effects of Displaying the Time Targets of the Manchester Triage System to Emergency Department Personnel: Prospective Crossover Study

Jonas Bienzeisler[1], MSc; Guido Becker[2], MSc; Bernadett Erdmann[3], Dr med; Alexander Kombeiz[1], MSc; Raphael W Majeed[1,4], MSc, Dr biol hom; Rainer Röhrig[1], Dr med; Felix Greiner[5,6], MSc; Ronny Otto[6], BSc; Fabian Otto-Sobotka[7], Dipl Math, Dr rer nat; AKTIN Research Group[8]

[1]Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany

[2]Dedalus HealthCare, Bonn, Germany

[3]Klinikum Wolfsburg, Wolfsburg, Germany

[4]Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), German Center for Lung Research (DZL), Giessen, Germany

[5]Institute for Occupational and Maritime Medicine (ZfAM), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany

[6]Department of Trauma Surgery, Otto von Guericke University, Magdeburg, Germany

[7]Division of Epidemiology and Biometry, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

[8]see Acknowledgements

Corresponding Author:
Jonas Bienzeisler, MSc
Institute of Medical Informatics
Medical Faculty
RWTH Aachen University
Pauwelsstraße 30
Aachen, 52074
Germany
Phone: 49 24180 ext 88870
Email: jbienzeisler@ukaachen.de

Abstract

Background: The use of triage systems such as the Manchester Triage System (MTS) is a standard procedure to determine the sequence of treatment in emergency departments (EDs). When using the MTS, time targets for treatment are determined. These are commonly displayed in the ED information system (EDIS) to ED staff. Using measurements as targets has been associated with a decline in meeting those targets.

Objective: This study investigated the impact of displaying time targets for treatment to physicians on processing times in the ED.

Methods: We analyzed the effects of displaying time targets to ED staff on waiting times in a prospective crossover study, during the introduction of a new EDIS in a large regional hospital in Germany. The old information system version used a module that showed the time target determined by the MTS, while the new system version used a priority list instead. Evaluation was based on 35,167 routinely collected electronic health records from the preintervention period and 10,655 records from the postintervention period. Electronic health records were extracted from the EDIS, and data were analyzed using descriptive statistics and generalized additive models. We evaluated the effects of the intervention on waiting times and the odds of achieving timely treatment according to the time targets set by the MTS.

Results: The average ED length of stay and waiting times increased when the EDIS that did not display time targets was used (average time from admission to treatment: preintervention phase=median 15, IQR 6-39 min; postintervention phase=median 11, IQR 5-23 min). However, severe cases with high acuity (as indicated by the triage score) benefited from lower waiting times (0.15 times as high as in the preintervention period for MTS1, only 0.49 as high for MTS2). Furthermore, these patients were less likely to receive delayed treatment, and we observed reduced odds of late treatment when crowding occurred.

XSL•FO
RenderX

Conclusions: Our results suggest that it is beneficial to use a priority list instead of displaying time targets to ED personnel. These time targets may lead to false incentives. Our work highlights that working better is not the same as working faster.

(J Med Internet Res 2024;26:e45593) doi:10.2196/45593

KEYWORDS

EHR; emergency medicine; AKTIN, process management; crowding; triage system; electronic health record; health care; treatment; emergency department

Introduction

The use of triage systems, such as the Manchester Triage System (MTS) and the Emergency Severity Index (ESI), is a standard procedure to determine the sequence of treatment in emergency departments (EDs) [1-4]. The MTS and ESI are used for determining time targets for physician contact. These time targets are usually displayed to ED staff by the ED information system (EDIS). However, it is not known how this affects ED processing times. This study investigates the impact of displaying these time targets for treatment to physicians within the ED on processing times.

Such sociotechnical systems directly impact health care delivery. The data they produce can be used to improve efficiency and effectiveness in a learning health care system [5]. Efficient treatment is necessary because EDs around the world are experiencing an increase in the number of patients they have to treat. Therefore, EDs must determine who to treat to avoid negative outcomes due to inadequate inpatient capacity. Crowding is an everyday challenge for EDs that is commonly quantified by extreme patient occupancy (patients present in the ED), extended length of stay (LOS) in the ED, and waiting time between triage and treatment [6,7]. Crowding occurs if the demand for emergency care surpasses the available resources within the ED [7,8]. Triage systems are thus necessary to manage the treatment sequence and optimize patient flow [1-4].

These systems are used to quickly assign a triage score to patients arriving at the ED, which defines the priority of treatment. The scores are mostly based on the acuity of the patient's illness and aim at identifying the risk of negative patient outcomes and ensuring timely and adequate treatment [3]. It is mandatory for German EDs to put patients through a triage process or treat them directly. The majority of EDs in Germany use 5-tier triage systems, the most commonly used being the MTS and the ESI [1]. Using these, the acuity of all patients is categorized through a 5-tiered scale, resulting in 5 levels of urgency typically known to personnel by the associated color (Table 1).

Table 1. Definition of the German version of the Manchester Triage System (MTS)

Level	Triage score	Color	Acuity	Time target (minutes)
1	MTS1	Red	Immediate	0
2	MTS2	Orange	Very urgent	10
3	MTS3	Yellow	Urgent	30
4	MTS4	Green	Standard	90
5	MTS5	Blue	Nonurgent	120

With the MTS, a time target is determined by a certified nurse with flowchart diagrams. This target is the latest time that is acceptable until a physician consults the patient. In Germany, these times are between 0 (Level 1, MTS1) and 120 minutes (Level 5, MTS5), contrary to the international version, which sets a limit of 240 minutes (Level 5, MTS5) [1,9]. For example, a patient with an MTS score of MTS2 should be attended by a physician within 10 minutes in Germany. A retriage is possible at any time; however, it becomes mandatory if the proposed time target is not reached.

The triage process and subsequent treatment are usually implemented within the EDIS, supporting clinical documentation, patient tracking, and order management. The triage system and the implementation of the triage system in the EDIS thus represent potential target areas for reducing waiting times, improving patient flow, and, as a result, ED efficiency [10,11].

The triage system and its implementation in the EDIS directly impact the provision of care, waiting times, LOS, and overall patient flow (Figure 1) [1,9-12]. Formally structured and established triage systems have higher validity than informally structured systems, but overall performance varies considerably [13,14]. Using a Dutch version of the MTS in a before-and-after study, Storm-Versloot et al [15] could not see any effects of triaging patients on waiting times but in fact observed increased average waiting times (average time from admission to treatment=15 min) and increased average treatment times (average time from admission to treatment=14 min). However, urgent cases (Level 2) received treatment faster on average (average time from admission to treatment=4 min), and patient satisfaction with respect to waiting times was higher, especially among low-urgency patients who typically crowd the ED. Using computational experiments, van Bockstal and Maenhout [16] similarly associated triage with increased waiting times for patients with less severe injuries and decreased waiting times for acute patients. They also found that there was a beneficial effect on triage system resource consumption. Vegting et al [17], on the other hand, reported a general increase in the LOS of noncritically ill patients due to triage, attributing it to the

XSL•FO
RenderX

fragmentation of the provision of care. However, triage and waiting times are not only important quality management factors but also significantly impact patient satisfaction. Multiple studies suggest that patients care a lot about the time they spend in the ED and receiving information on the length of their projected waiting time, which demonstrates that patient satisfaction correlates strongly with waiting time [18,19]. Aside from these observational studies, ED routine records have been used to predict waiting times using, for example, machine learning [20] and statistical approaches [21,22].

The time it takes to receive treatment is thus understood to be a crucial metric for assessing quality within the ED. Policy makers have also recognized the importance of ED waiting times, but some of the approaches taken to improve them are controversial. In the United Kingdom, for example, the waiting time target that has been set for EDs is 4 hours [23]. It is well known that using measurements as targets is associated with a decline in meeting those targets [24]. This association, known as Goodhart's law [25] and originating from economics, has often been discussed in the context of health care systems, with some experts believing that setting time targets leads to false incentives within the health care system [20,24,26].

Aside from such legislative requirements, time targets have clinical relevance in managing incoming patient flow within the ED. To track waiting patients, EDISs that implement the MTS usually display the time target determined by the triage score. Such sensory cues add to the cognitive load of attending personnel [27,28]. Further, time pressure may have negative effects on physicians' performance [29,30].

Therefore, the objective of this study was to analyze the effect on waiting times of displaying treatment time targets provided by the MTS score in the EDIS.

We hypothesized that displaying these targets could alter treatment and waiting times, thus influencing patient flow and crowding. We expected, in accordance with Goodhart's law, that the practice of displaying time targets for patient treatment, inadvertently contributes to inefficient patient management within EDs. We proposed that the display of these time targets creates a false incentive structure for ED physician personnel. It could potentially lead them to prioritize meeting these targets over other critical aspects of patient care. This focus on time targets might result in suboptimal treatment decisions, potentially exacerbating patient wait times and lowering throughput, contrary to the intended purpose of these targets.

Figure 1. Data collection and variable definition. The evaluation of the intervention was based on routine documentation stemming from an emergency department (ED) information system. Timestamps describing internal patient flow were derived from routine medical records. The main target parameters were time to triage (first contact with a triaging nurse), waiting time from triage until treatment by the physician (first contact with the attending physician), and ED length of stay.

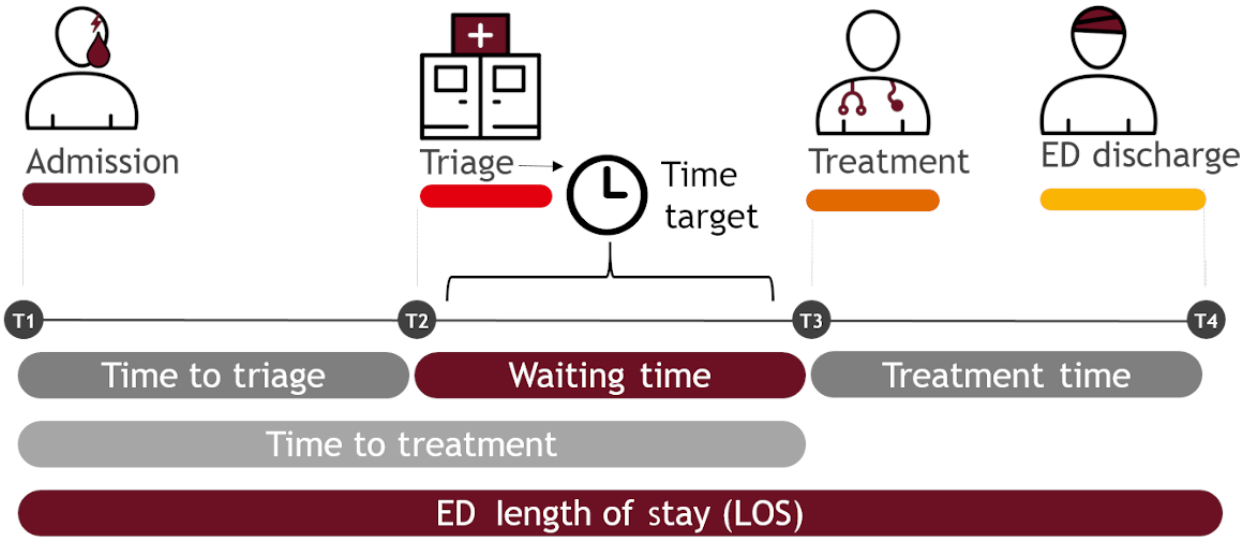

Methods

Overview

Based on routine documentation, we conducted a prospective crossover study of all adult patients treated before (preintervention) and after (postintervention) an update of the EDIS module used in the ED. The new version no longer displayed MTS time targets but a priority list (Figure 2). Patients thus received treatment with and without displaying the treatment time targets provided by the MTS score.

XSL•FO
RenderX

Figure 2. Study design and mock-ups of the emergency department information system (EDIS). In a prospective cross-over study, we evaluated the effect of displaying treatment time targets provided by the Manchester Triage System (MTS) score to emergency department personnel. Preintervention and postintervention took place before and after the update of an EDIS. The EDIS showed the time target determined by the MTS preintervention (triage time in the mock-up). Postintervention, a priority list (triage list in mock-up) was displayed instead of a time target. A red flag was displayed next to the color-coded MTS score when the time target was reached.

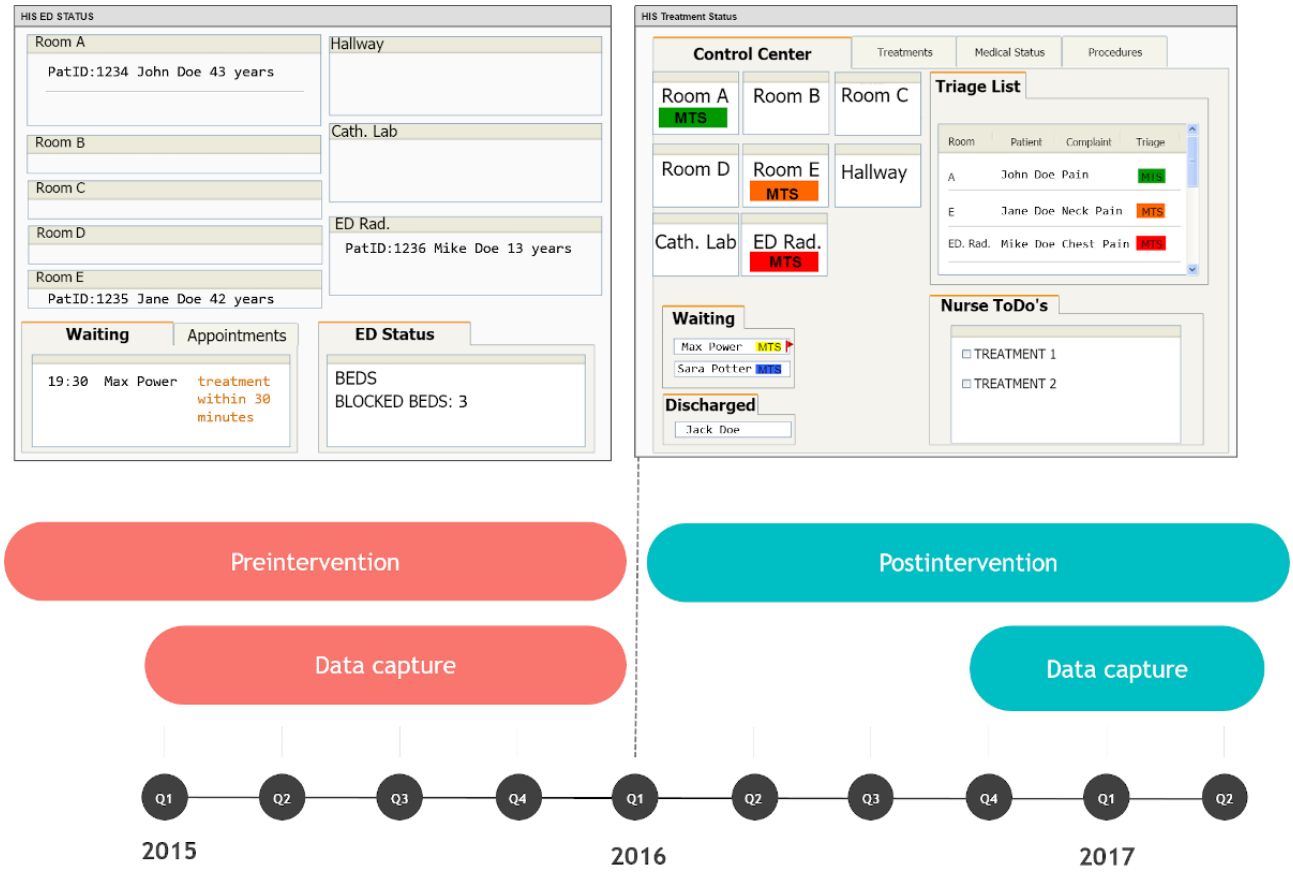

Study Context

We conducted the study within the ED of the Klinikum Wolfsburg in Germany, a central public hospital with around 70,000 patients per year. It is the only hospital to have an ED within a 17-km radius. It has a certified trauma center and processes 35,000 patients in the ED every year.

The standard procedure for patients arriving at the ED (in both phases) involved administrative registration, followed by a triage conducted by a nurse, retriage if necessary, treatment by a physician following a certain waiting time, and finally discharge. In urgent cases (eg, admission by ambulance), triage is bypassed if there are sufficient resources.

No training on triage practices was conducted during the study period. Staff did receive technical training from the software vendor before the implementation of the new module. The staffing levels, both for nursing and medical personnel, remained unchanged throughout the study. The medical staffing in the ED was fixed to a schedule. Following a fixed plan for day and night shifts, a team of 13.7 full-time equivalent positions for physicians directly associated with the ED and 29.2 full-time equivalent positions for nursing staff were maintained. The staffing model was applied on all weekdays, weekends, and public holidays. There were no notable changes in equipment during the study.

The EDIS used in the ED is a module of the monolithic hospital information system ORBIS (Dedalus Healthcare). The EDIS documentation method adhered to the ED medical record of the German Interdisciplinary Association for Intensive Care and Emergency Medicine. The captured routine documentation data are accessible through the alliance for information and communication technology in intensive care and emergency medicine (abbreviated in German as AKTIN) Emergency Department Data Registry operated by AKTIN in Germany [31,32].

Study Design

During the preintervention phase (January 1, 2014, to December 31, 2014), nurses carried out triage and documentation anonymously using paper-based MTS presentation flowcharts and documented data in a view only accessible to the nurses. Nursing care was documented externally. Time points of patient contact and triage score were displayed to physicians and nurses in a user interface (UI) developed in house. The UI used consisted of a list-based view of rooms and patients. Triage and time targets were visible to all personnel in a room overview. Patients were selected for treatment according to the displayed list. Physicians could open an ED medical record by clicking on triaged patients.

In the postintervention phase (October 13, 2015, to January 31, 2016), the Cockpit Notaufnahme EDIS module provided by

XSL•FO
RenderX

Dedalus Healthcare was used, which no longer displayed the MTS time target.

In the postintervention phase, data were collected after a transition period between the intervention phases to mitigate potential biases (Figure 2). During this interval, the adaptations to the new intervention were allowed to stabilize. Apart from using electronic rather than paper flowcharts for determining the MTS, the changes were purely cosmetic. Standard procedures for incoming patients remained the same. The new UI displayed patients, nurses, and physicians in a list according to the urgency of treatment without time targets. A room-based overview of patients was included in the new UI, which served as an overview for all personnel. Patients could be selected and assigned to rooms through drag and drop. A signal was displayed when time targets were exceeded.

Data Acquisition

Overview

We extracted process times at the end of the postintervention phase. Data were collected routinely by the EDIS, anonymized within the clinic, and provided through the infrastructure of the AKTIN Emergency Department Data Registry [31,32].

The main target parameters were the time until triage, waiting time from triage until treatment by the physician, and ED LOS (Figure 1). As a secondary outcome, we calculated a binary classification of late treatment as defined by the assigned triage score, which we interpreted as a proxy for ED efficiency (eg, “1” if waiting time was longer than 10 minutes for patients assigned the triage score MTS2).

Data Exclusion

Documentation errors may lead to implausible or missing timestamps. We considered a physician contact before the initial assessment nonevaluable and excluded nonpositive processing times as well as data with missing admission or triage timestamps. We classified LOS and waiting time that lie above 3 SDs (>3σ) from the mean (μ) of patients with the lowest acuity (MTS5), rounded to the next complete hour, as outliers. Any processing time between admission, triage, and treatment above 300 minutes or a LOS higher than 600 minutes resulted in exclusion. Further, we omitted data from the transition period.

Data Analysis

Overview

Timestamps were generated automatically by the EDIS when a nurse registered a patient, when a nurse opened the triage view and documented the triage score, when the physician opened the ED record, and when the patient was marked as discharged by the physician (Figure 1). We deduced direct physician contact from the occurrence of missing triage timestamps in combination with physician contact timestamps. In addition, we included data on the count of patients present at the time of physician contact (T3; Figure 1) using a counter that we set to “1” for the first patient in the data set. From the admission timestamp, we extracted the annual season, hour of the day, and whether the admission occurred on a working day (as opposed to a nonworking day). We corrected erroneous data with the clinic’s help and excluded implausible timestamps from further analysis (ie, timestamp treatment before timestamp administration). We performed all calculations using the statistical software R (version 3.6.1; R Core Team).

Primary Analysis

We extracted the entries from the EDIS into a standardized data entry form. The descriptive main analysis included the median and IQR for metric variables and the computation of absolute and relative frequencies for categorical variables. At the end of the intervention phase, we sought evaluative feedback regarding the intervention’s impact and effectiveness from the head of the ED and ED personnel.

Secondary Analysis

The secondary analysis consisted of generalized additive regression models. Response variables were waiting time until treatment in minutes and the binary occurrence of delayed treatment. The preselected set of covariates were the number of patients present within the ED at the time of physician contact, MTS score, study phase, working day, annual season, and hour of day. We assumed a gamma distribution with a log-link for positive waiting times (model 1) and used logistic regression for delayed treatment (model 2). The covariate hour of the day was included on a cyclic P-spline basis. The number of patients within the ED was modeled with a regular P-spline basis and 2-way interactions of MTS score and study phase, as well as patients present and study phase. The effects on waiting time are reported as multiplicative effects exp(\β) with 95% CIs. The logistic regression results are reported as odds ratios (ORs), with 95% CIs. The nonlinear effects estimated by splines are reported graphically.

Ethical Considerations

The study was approved by an ethics committee before the update of the EDIS module (Medical Ethical Committee Uni Oldenburg, Vote-No: 2016-05, Chair F Griesinger). Results are reported according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [33].

Results

Overview

In total, we extracted 48,822 data sets from the EDIS: 35,167 in the preintervention phase and 10,655 in the postintervention phase. Excluding 636 data sets with implausible timestamps from further analysis, we analyzed 45,186 data sets (Figure S1 in Multimedia Appendix 1). In general, the distribution of assigned triage scores differed between the preintervention and postintervention phases (Table 2). We noticed an upcoding postintervention an increase in the assignment of more urgent triage levels. However, most triaged patients were still assigned an MTS3 or MTS4 score in both phases. Very few patients in postintervention and none in preintervention were given the most urgent MTS score of MTS1. Thus, in the analyses, we focused on patients assigned a triage score of MTS 2-5.

XSL•FO
RenderX

Table 2. Comparative analysis of preintervention and postintervention study sample characteristics. Categorical variables are presented as frequencies (percentages) and were analyzed using chi-square tests. Continuous variables are presented as means with SDs and reported along with their medians and IQRs. We assumed the number of patients present per day and at physician contact to be normally distributed and compared them using independent sample 2-tailed *t* tests (two tailed). Nonnormally distributed waiting times were compared using Mann-Whitney *U* tests.

Characteristics	Preintervention phase (n=34,727)	Postintervention phase (n=10,459)	Total (N=45,186)	*P* value
Year, n (%)				<.001
2014	34,727 (100)	0 (0)	34,727 (76.9)	
2015	0 (0)	7556 (72.2)	7556 (16.7)	
2016	0 (0)	2903 (27.8)	2903 (6.4)	
Weekday, n (%)				.89
Working day	23,399 (67.4)	7040 (67.3)	30,439 (67.4)	
Nonworking day	11,328 (32.6)	3419 (32.7)	14,747 (32.6)	
Season of the year, n (%)				<.001
Fall	8698 (25)	4527 (43.3)	13,225 (29.3)	
Spring	8872 (25.5)	0 (0)	8872 (19.6)	
Summer	9003 (25.9)	0 (0)	9003 (19.9)	
Winter	8154 (23.5)	5932 (56.7)	14,086 (31.2)	
MTS[a] score, n (%)				<.001
LWBS[b]	0 (0)	38 (0.4)	38 (0.1)	
MTS1	0 (0)	15 (0.1)	15 (0)	
MTS2	220 (0.6)	186 (1.8)	406 (0.9)	
MTS3	4369 (12.6)	2004 (19.2)	6373 (14.1)	
MTS4	10,828 (31.2)	5049 (48.3)	15,877 (35.1)	
MTS5	3059 (8.8)	589 (5.6)	3648 (8.1)	
Direct contact	16,251 (46.8)	2578 (24.6)	18,829 (41.7)	
Adherence to MTS time target, n (%)				<.001
On-time	13,381 (72.4)	5190 (68.6)	18,571 (71.3)	
Late	5094 (27.6)	2376 (31.4)	7470 (28.7)	
Patients per day				.16
Missing data, n	0	0	0	
Mean (SD)	95.142 (10.618)	93.384 (13.732)	94.730 (11.435)	
Median (range)	95.000 (64.000-127.000)	93.000 (1.000-117.000)	95.000 (1.000-127.000)	
IQR	88.000-102.000	86.000-102.000	88.000-102.000	
Patients present at physician contact				<.001
Missing data, n	0	0	0	
Mean (SD)	12.479 (5.908)	15.108 (6.938)	13.073 (6.253)	
Median (range)	12.000 (0.000-33.000)	15.000 (1.000-41.000)	13.000 (0.000-41.000)	
IQR	8.000-16.000	10.000-20.000	8.000-17.000	
Waiting time in minutes				< .001
Missing data, n	16,252	2893	19,145	
Mean (SD)	55.831 (52.088)	58.898 (54.944)	56.722 (52.951)	
Median (range)	39.000 (0.000-294.000)	41.000 (0.000-291.000)	39.000 (0.000-294.000)	
IQR	17.000-80.000	17.000-85.000	17.000-81.000	

XSL•FO
RenderX

[a]MTS: Manchester Triage System.

[b]LWBS: left without being seen.

Primary Analysis

Depending on the severity of the injury, it took different times until a nurse assigned a patient a triage score. The less serious the complaint, the greater the MTS score, and the longer it took until the triage score was assessed. In the postintervention phase, patients with an MTS3 and MTS4 were triaged slightly faster than in the preintervention phase; those with an MTS5 were triaged slower (Tables S1-S5 in Multimedia Appendix 1).

The percentage of patients who received immediate treatment by a physician instead of first being triaged was higher in the preintervention phase (16,251/34,727, 46.8%) than in the postintervention phase (2578/10,459, 24.6%). Incoming patients that a physician attended to immediately received treatment faster in the postintervention phase (time from admission to treatment: preintervention phase=median 15, IQR 6-39 min; postintervention phase=median 11, IQR 5-23 min). In postintervention phase only, a small group of patients (n=38) left the ED after triage but before seeing a physician. Generally, we found that the time from patient registration until triage was slightly reduced postintervention, although still somewhat comparable.

The presence of patients had a greater impact, leading to increased waiting times at the triage stage, especially for low-urgency patients (Figure 3). In the preintervention phase, there was a rise in the time needed until triage, when more than 30 patients were present within the ED. However, we saw no such effect the postintervention phase. Daily patient numbers were comparable in the pre- and postintervention phases.

Figure 3. Waiting times from triage to treatment by assigned triage score and study phase. The old emergency department information system (EDIS) version (preintervention, green) displayed the time target determined by the Manchester Triage System (MTS), while the new EDIS version (postintervention, red) did not. The average waiting times from triage until physician contact varied for MTS5 (pre: 41 min, IQR 17-87; post: 67 min, IQR 31-121 min), MTS4 (pre: 46 min, IQR 20-88; post: 50 min, IQR: 23-95 min), MTS3 (pre: 26 min, IQR 12-53; post: 24 min, IQR 11-49 min), MTS2 (pre: 12 min, IQR 5-26 min; post: 10 min, IQR 5-16 min), and MTS1 (post: 4 min, IQR 2-14 min).

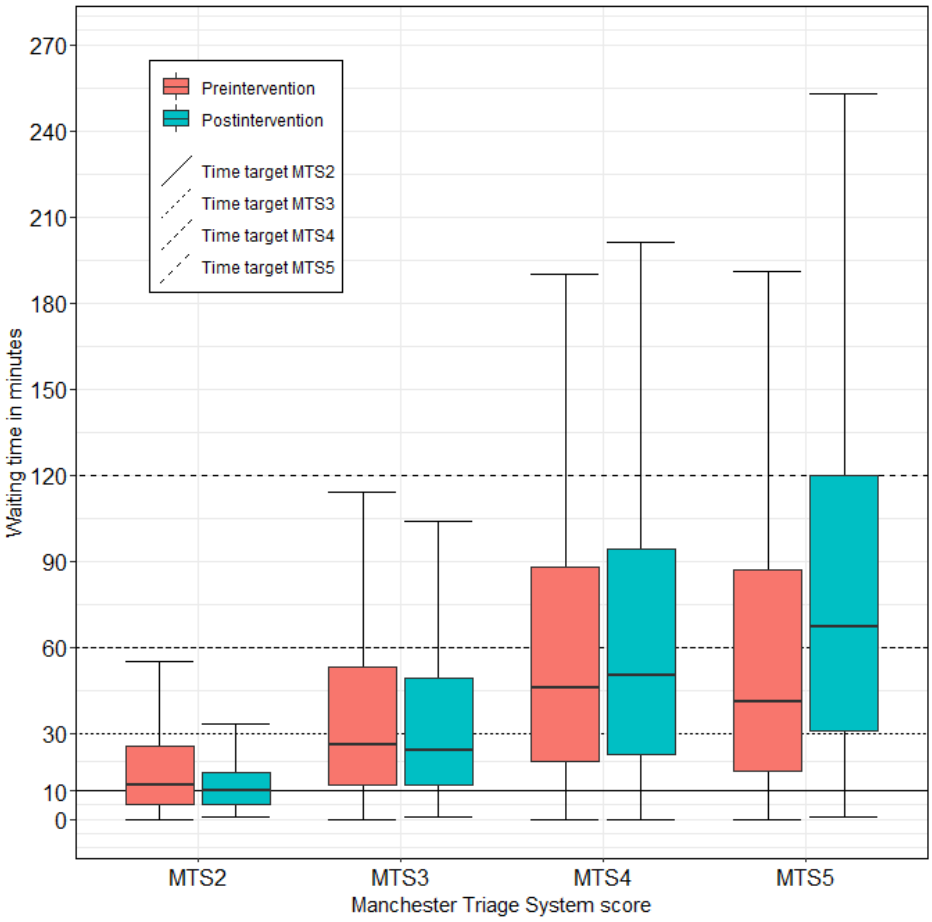

Crowding, as measured in patients present on average per visit, was higher in the postintervention phase (median 12, IQR 8-16) than in the preintervention phase (median 15, IQR 10-20). Accordingly, we observed an increased LOS in the postintervention phase. On average, patients stayed 29 minutes longer in the ED (preintervention phase=median 119, IQR 66-189 min; postintervention phase=median 148, IQR 88-226 min). Less severe cases with MTS4 and MTS5—the majority of cases—waited longer (Tables S1-S5 in Multimedia Appendix 1).

The waiting time from triage until physician contact increased on average as well (preintervention phase=median 39, IQR

XSL•FO
RenderX

17-80 min; postintervention phase=median 41, IQR 17-85 min), and the time target provided by the MTS was missed in a greater percentage of triaged cases postintervention (Δn =3,8%). The rate of timely treatment changed in the same fashion. Delayed treatment and retriage were more common in the postintervention phase than in the preintervention phase. This effect was mainly due to patients with triage scores of MTS4 and MTS5. For severely injured patients with MTS2 and MTS3, we observed a lower number of treatment instances occurring outside of the target scope. However, although these process times and statistics suggest impeded throughput, practicing physicians reported improved patient flow. Furthermore, the head of the ED reported perceiving an improvement in patient flow and treatment quality.

Secondary Analysis

In the secondary analysis of positive waiting times (model 1, Table S6 in Multimedia Appendix 1), we found that waiting times increased on average by a factor of 1.27 (CI) in the postintervention phase. ED crowding amplified this effect. However, the estimated interaction effects showed that waiting times postintervention were only 0.15 times as high as in the preintervention phase for MTS1, only 0.49 as high for MTS2, and only 0.68 as high for MTS3. These results can be multiplied by the main effects: that waiting times for MTS1 were, on average, only a third of the waiting times for MTS5, and waiting times for MTS2 were 0.68 of MTS5 waiting times. The effects of weekends and annual seasons in the model were negligible. On an average day, waiting times increased at around 6 AM and from 6 PM to midnight.

We observed some similarities when modeling delayed treatment (model 2; Table S7 in Multimedia Appendix 1). The odds of late treatment increased by 2.32 (CI) in the postintervention phase but were reduced by a factor of 0.32 for MTS2, by 0.44 for MTS3, and by 0.6 for MTS4. These results stand in positive contrast to the overall odds of delayed treatment of 8.79 for MTS2 compared to MTS5, 4.93 for MTS3, and 1.86 for MTS4. The effects of seasons and weekends are negligible. Similar to positive waiting times, certain hours of the day led to increased odds of delayed treatment—at around 6 AM and from 6 PM to 11 PM.

Discussion

Overview

The introduction of the new software module that did not show the time target for treatment but instead used a priority list resulted in prolonged waiting times for patients with lower acuity (MTS4 and MTS5) but reduced the odds of late treatment for patients with higher acuity (MTS2 and MTS3). Similar to what Storm-Versloot et al [15] observed, patient flow was thus found to be optimized in the second phase of the trial. Critical patients with higher urgency levels received more timely treatment and waited less time, an effect commonly associated with triage systems. When considering all patients, the LOS and waiting times increased. A longer LOS for all patients eventually leads to crowding and can be used to quantify crowding along with ED occupancy [7]. Indeed, we observed a slightly higher number of patients within the ED at the physician contact stage postintervention. Paradoxically, the head of the ED and the practicing physicians—professionals who are well aware of crowding and the daily implications of inadequate inpatient capacity—were delighted with the new software, refusing to change back to the old system for further investigation. While these remarks are anecdotal, multivariate regression models confirmed the perception that waiting times for critical patients had reduced. Furthermore, the probability of delayed treatment was also reduced.

The waiting time from triage to treatment was optimized, but the time until triage increased, while overcrowding did not affect the latter. The introduction of the new system led to a more sophisticated triage. Severe cases were treated more effectively, as perceived by the ED staff (Figure 4). Efficient treatment is also reflected in the number of patients sent straight to ED care instead of being triaged [34]; in the postintervention phase, more patients were triaged. We were able to quantify a positive effect of triage with regard to receiving timely treatment when the number of patients waiting increased (Figure 5). As one might expect, triage has little impact on patient processing when ED occupancy is low, and the ED is thus not crowded.

The findings support the idea that setting and displaying targets—in this case, displaying the time target for the respective MTS—may lead to false incentives. Time constraints may impact the thoroughness of care and the engagement in patient-centered decision-making [35]. We speculate that the knowledge of a time frame leads physicians to a less focused use of resources—particularly for medium urgent cases waiting. Previous work suggests that cues can inadvertently shape clinical priorities and decision-making [27,28]. This may lead to a phenomenon known as target behavior, where the focus of physicians shifts toward meeting set benchmarks rather than optimizing patient care based on clinical needs.

The removal of explicit time targets in this study likely altered the cognitive framework within which physicians operate, redirecting their focus from adhering to arbitrary time constraints to assessing and addressing patient needs more holistically, free from the distortions introduced by time pressures. We moved from strict time targets to a patient-centric model of postintervention.

Moreover, the change in the EDIS could have also disrupted habitual response patterns to visible time cues, necessitating a recalibration of treatment prioritization strategies that no longer relied on time targets as a primary directive [36]. This shift could have been further influenced by any briefings or guidance provided to physicians alongside the software change, which might have emphasized a more patient-centric approach to triage and treatment.

One could be tempted to generalize and conclude that waiting limits introduced by policymakers might lead to similar false incentives in much the same way. ED LOS may thus be insufficient as a quality indicator for ED care on its own and instead must be understood as a process metric, especially concerning ED crowding. These conclusions are speculative but similar to what we have previously observed in a larger sample [37].

XSL•FO
RenderX

Figure 4. Odds of late treatment by triage score and study phase for a fixed time (winter at 6 AM on a weekday) and patients present (n=13) with a standard normal deviation (α=5%). The time target for timely treatment is determined by the Manchester Triage System (MTS) score and was displayed to physicians at all times only in the old emergency department information system (EDIS) version (preintervention). Odds were calculated using a generalized additive regression model, assuming a logistic regression for delayed treatment. The odds of late treatment increased by 2.32 (CI) when using the new EDIS version (postintervention), which did not show the time target to physicians. However, the odds of late treatment were reduced by a factor of 0.32 for MTS2, 0.44 for MTS3, and 0.6 for MTS4.

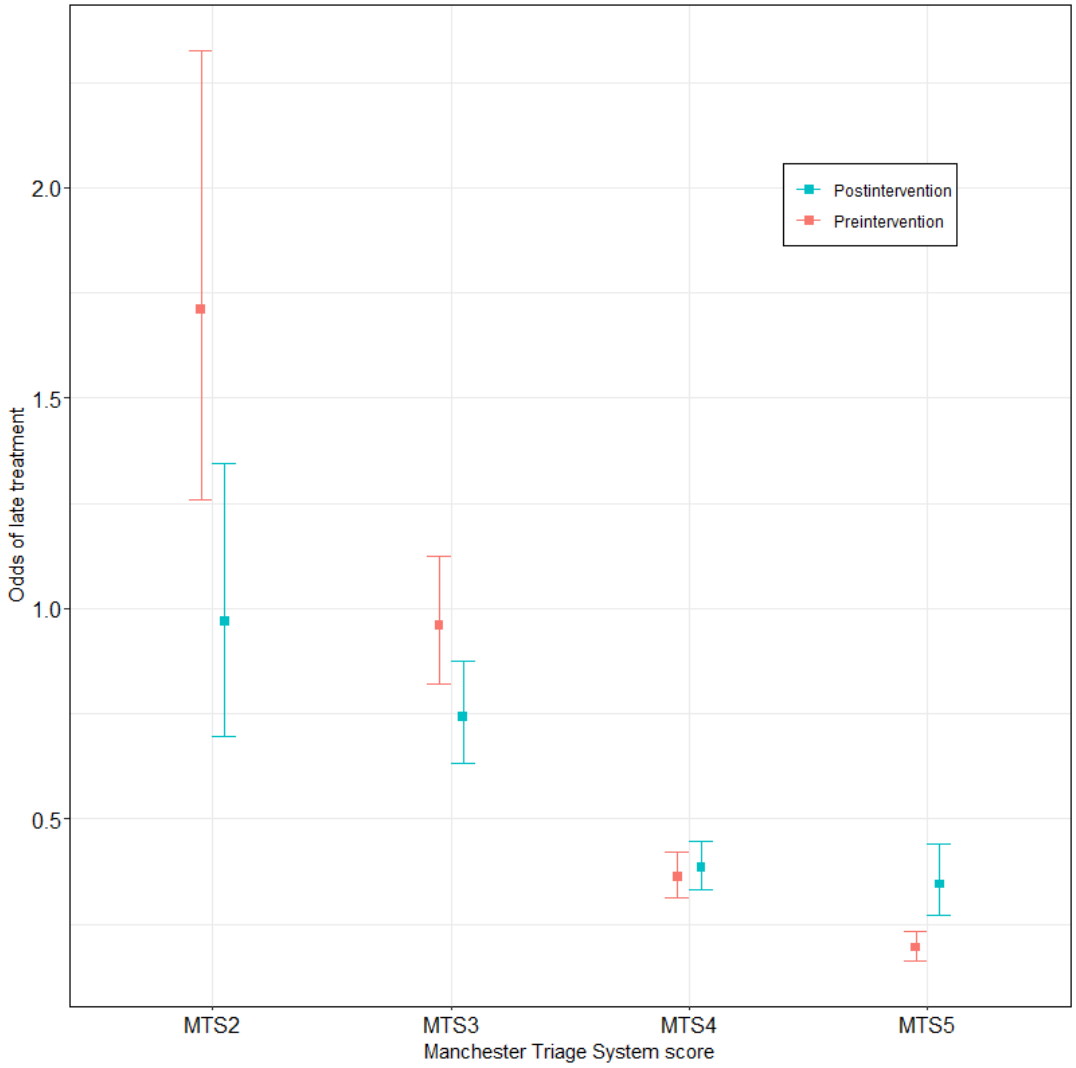

XSL•FO
RenderX

Figure 5. Odds of late treatment by the number of patients present within the emergency department (ED) and study phase for different triage scores A) MTS2, B) MTS3, C) MTS4, D) MTS5 at a fixed time (winter at 6 AM on a weekday) with a standard normal deviation (α=5%). The Manchester Triage System (MTS) score determines the time target for timely treatment. Preintervention, the time target was displayed to physicians at all times in the ED information system (EDIS), while it was not postintervention. The odds were calculated using a generalized additive regression model, assuming a logistic regression for delayed treatment. While the odds for late treatment in general increased by 2.32 (CI) when using the new EDIS version (postintervention) that did not show the time target to physicians, late treatment was considerably less likely when crowding (as indicated by the number of patients waiting) occurred. Severe cases (MTS2, MTS3, and MTS4) were less likely to receive late treatment postintervention. Patients with MTS5 were less likely to be treated on time postintervention.

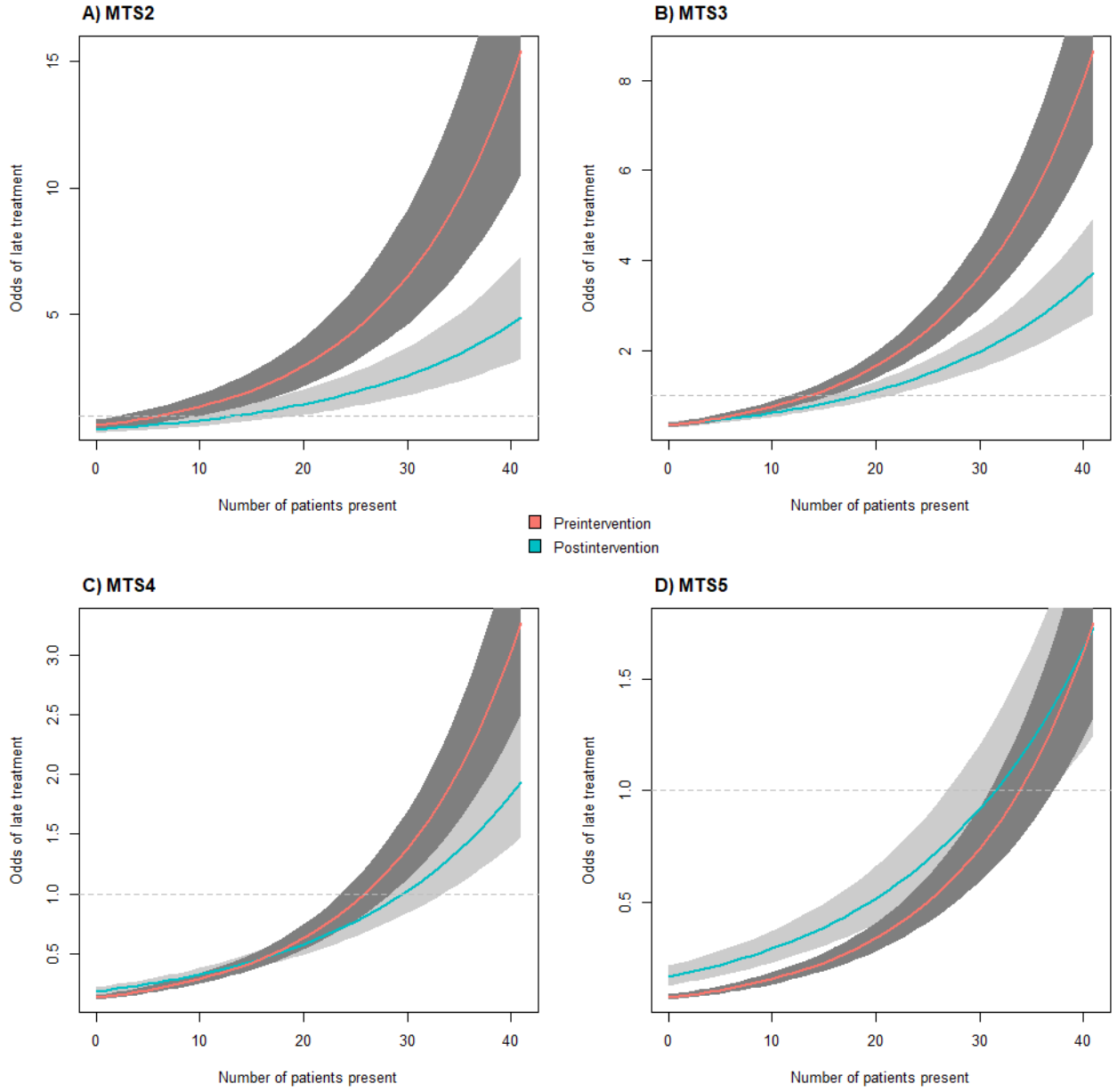

Limitations

This study had several limitations, making further research necessary. In the postintervention phase, data were collected over a shorter time period. However, the number of patients per hour of the day and the total number of patients were comparable. The selection bias should be low; outcomes were measured over time across the whole population of ED attendees. Further, our intervention targeted the delivery and organization of services within the ED and was hence on the service level [38]. A controlled evaluation was not possible. A previously proposed prospective study with an on-off study design [39] was not feasible. Because of this, we cannot assess the effects of minor changes accompanying the introduction of the new EDIS module. However, we assume that these can be neglected, as the changes were either cosmetic or impacted the triage process itself, which does not influence the process times after triage. The latter was the case with the changed triage process in the postintervention phase—the only significant noncosmetic change in this study. The workflow and the processing of patients did not change.

There is evidence that more formally structured triage leads to overtriage, especially when using the MTS [13,40-42]. The use of electronic presentation diagrams could thus explain the upcoding in treatment priority that we observed. This fails to

XSL•FO
RenderX

explain, especially for nonurgent cases, the fact that the time target for timely treatment was missed in greater numbers, yet patient flow improved. We were unable to assess the influence of raised awareness among ED personnel in general, given the setting of the study. Further work is necessary to systematically address the effects on ED personnel and the quality of treatment from a provider perspective. Further, future work will have to address patient satisfaction, as the majority of patients had a prolonged ED LOS. We intend to base further work on data from multiple hospitals, which is possible using the infrastructure of the AKTIN Emergency Department Data Registry [31].

Conclusions

The results suggest that it is beneficial not to display time targets when using triage systems, thus confirming the validity of Goodhart's law. Similar to what others have reported [15,17,19], we also showed that the update to the triage system had an unforeseen impact on ED waiting times. Rather than improving the quality of treatment through accelerating processes within the ED, we saw an improvement in patient flow for patients with more severe injuries. Although it was only anecdotal evidence, the improvement was much appreciated by the attending physicians and the head of the ED. Our work highlights that working better is not the same as working faster; working more quickly does not automatically imply better care or results. It is essential to discuss how time is spent instead of focusing on how to save it. Furthermore, our results suggest that using the number of patients present in the ED as an isolated metric for crowding can be misleading.

Acknowledgments

The authors acknowledge the following members of the AKTIN-Research-Group: Susanne Drynda, Department of Trauma Surgery, Otto von Guericke University; Wiebke Schirrmeister, Department of Trauma Surgery, Otto von Guericke University; Saskia Ehrentreich, Department of Trauma Surgery, Otto von Guericke University; Volker Thiemann, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg; Felix Walcher, Department of Trauma Surgery, Otto von Guericke University; Lucas Triefenbach, Institute of Medical Informatics, Medical Faculty, RWTH Aachen University; Insa Seeger Research Network Emergency and Intensive Care Medicine Oldenburg, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg; Tolga Philipp Naziyok, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg; Marc Wilken, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg; Myriam Lipprandt, Institute of Medical Informatics, Medical Faculty, RWTH Aachen University.

This publication was funded by the German Federal Ministry of Education and Research and Network of University Medicine 2.0: "NUM 2.0" (01KX2121), Project "AKTIN@NUM" (01KX1319A), and Project "AKTIN."

Data Availability

The data sets analyzed in this study are not publicly accessible and cannot be transmitted to a third party, in compliance with data protection regulations and the AKTIN Emergency Department Data Registry's policies on data use and access. However, they can be made available by the corresponding author upon reasonable request.

Authors' Contributions

The original draft was written by JB and critically revised by GB, BE, AK, RO, FG, RR, and FOS for important intellectual content. JB contributed to conceptualization, data curation, methodology, formal analysis, validation, and visualization of results together. RR, GB, and BE provided access to data and expertise in clinical processes in emergency departments and data management within clinical information systems. They contributed to conceptualization, investigation, data curation, and formal analysis. RO, FG, RR, and FOS provided methodological expertise and contributed to methodology and formal analysis. AK and JB developed software components of the AKTIN Emergency Department Data Registry. RR and BE secured funding and provided supervision. FOS directed the methodological design, visualization of results, ensured research validation, and supervised the work.

Conflicts of Interest

Guido Becker is employed by Dedalus HealthCare GmbH, which manufactures a product related to the subject matter. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 3981 KB - jmir_v26i1e45593_app1.docx]

References

XSL•FO
RenderX

1. Gräff I, Goldschmidt B, Glien P, Bogdanow M, Fimmers R, Hoeft A, et al. The German version of the Manchester Triage System and its quality criteria--first assessment of validity and reliability. PLoS One 2014;9(2):e88995 [FREE Full text] [doi: 10.1371/journal.pone.0088995] [Medline: 24586477]
2. Fernandes CMB, Tanabe P, Gilboy N, Johnson LA, McNair RS, Rosenau AM, et al. Five-level triage: a report from the ACEP/ENA Five-Level Triage Task Force. J Emerg Nurs 2005;31(1):39-50; quiz 118. [doi: 10.1016/j.jen.2004.11.002] [Medline: 15682128]
3. Iserson KV, Moskop JC. Triage in medicine, part I: concept, history, and types. Ann Emerg Med 2007;49(3):275-281. [doi: 10.1016/j.annemergmed.2006.05.019] [Medline: 17141139]
4. Moskop JC, Iserson KV. Triage in medicine, part II: underlying values and principles. Ann Emerg Med 2007;49(3):282-287. [doi: 10.1016/j.annemergmed.2006.07.012] [Medline: 17141137]
5. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010;2(57):57cm29 [FREE Full text] [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]
6. Jones P, Wells S, Ameratunga S. Towards a best measure of emergency department crowding: lessons from current Australasian practice. Emerg Med Australas 2018;30(2):214-221. [doi: 10.1111/1742-6723.12868] [Medline: 28941074]
7. Badr S, Nyce A, Awan T, Cortes D, Mowdawalla C, Rachoin JS. Measures of emergency department crowding, a systematic review. How to make sense of a long list. Open Access Emerg Med 2022;14:5-14 [FREE Full text] [doi: 10.2147/OAEM.S338079] [Medline: 35018125]
8. FitzGerald G, Jelinek GA, Scott D, Gerdtz MF. Emergency department triage revisited. Emerg Med J 2010;27(2):86-92 [FREE Full text] [doi: 10.1136/emj.2009.077081] [Medline: 20156855]
9. Slagman A, Greiner F, Searle J, Harriss L, Thompson F, Frick J, et al. Suitability of the German version of the Manchester Triage System to redirect emergency department patients to general practitioner care: a prospective cohort study. BMJ Open 2019;9(5):e024896 [FREE Full text] [doi: 10.1136/bmjopen-2018-024896] [Medline: 31064804]
10. Di Laura D, D'Angiolella L, Mantovani L, Squassabia G, Clemente F, Santalucia I, et al. Efficiency measures of emergency departments: an Italian systematic literature review. BMJ Open Qual 2021;10(3):e001058 [FREE Full text] [doi: 10.1136/bmjoq-2020-001058] [Medline: 34493488]
11. Baumlin KM, Shapiro JS, Weiner C, Gottlieb B, Chawla N, Richardson LD. Clinical information system and process redesign improves emergency department efficiency. Jt Comm J Qual Patient Saf 2010;36(4):179-185. [doi: 10.1016/s1553-7250(10)36030-2] [Medline: 20402375]
12. Mahmoodian F, Eqtesadi R, Ghareghani A. Waiting times in emergency department after using the emergency severity index triage tool. Arch Trauma Res 2014;3(4):e19507 [FREE Full text] [doi: 10.5812/atr.19507] [Medline: 25738132]
13. Storm-Versloot MN, Ubbink DT, Kappelhof J, Luitse JSK. Comparison of an informally structured triage system, the Emergency Severity Index, and the Manchester Triage System to distinguish patient priority in the emergency department. Acad Emerg Med 2011;18(8):822-829 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01122.x] [Medline: 21843217]
14. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
15. Storm-Versloot MN, Vermeulen H, van Lammeren N, Luitse JS, Goslings JC. Influence of the Manchester Triage System on waiting time, treatment time, length of stay and patient satisfaction; a before and after study. Emerg Med J 2014;31(1):13-18 [FREE Full text] [doi: 10.1136/emermed-2012-201099] [Medline: 23302504]
16. van Bockstal E, Maenhout B. A study on the impact of prioritising emergency department arrivals on the patient waiting time. Health Care Manag Sci 2019;22(4):589-614. [doi: 10.1007/s10729-018-9447-5] [Medline: 29725894]
17. Vegting IL, Alam N, Ghanes K, Jouini O, Mulder F, Vreeburg M, et al. What are we waiting for? factors influencing completion times in an academic and peripheral emergency department. Neth J Med 2015;73(7):331-340 [FREE Full text] [Medline: 26314716]
18. Taylor C, Benger JR. Patient satisfaction in emergency medicine. Emerg Med J 2004;21(5):528-532 [FREE Full text] [doi: 10.1136/emj.2002.003723] [Medline: 15333521]
19. Marohl R, Medghaltschi H, Herchenbach C. Patientenzufriedenheit in der notfallambulanz—einfluss der triage auf die patientenzufriedenheit. In: Abstracts zu vorträgen und postern der 10. Jahrestagung der deutschen gesellschaft interdisziplinäre. Notfall- und Akutmedizin 2015:15-16. [doi: 10.1007/s10049-015-0058-0]
20. Pak A, Gannon B, Staib A. Predicting waiting time to treatment for emergency department patients. Int J Med Inform 2021;145:104303. [doi: 10.1016/j.ijmedinf.2020.104303] [Medline: 33126060]
21. Ataman MG, Sarıyer G. Predicting waiting and treatment times in emergency departments using ordinal logistic regression models. Am J Emerg Med 2021;46:45-50. [doi: 10.1016/j.ajem.2021.02.061] [Medline: 33721589]
22. Sun Y, Teow KL, Heng BH, Ooi CK, Tay SY. Real-time prediction of waiting time in the emergency department, using quantile regression. Ann Emerg Med 2012;60(3):299-308. [doi: 10.1016/j.annemergmed.2012.03.011] [Medline: 22579492]
23. Trivedy M. If I were minister for health, I would … review the four-hour waiting time in the emergency department. J R Soc Med 2021;114(4):218-221 [FREE Full text] [doi: 10.1177/0141076820975363] [Medline: 33325759]
24. Crawford SM. Goodhart's law: when waiting times became a target, they stopped being a good measure. BMJ 2017;359:j5425. [doi: 10.1136/bmj.j5425] [Medline: 29180439]

XSL•FO
RenderX

25. Goodhart C, editor. Monetary Theory and Practice. London: Macmillan Education UK; 1984.
26. Campbell P, Boyle A, Higginson I. Should we scrap the target of a maximum four hour wait in emergency departments? BMJ 2017;359:j4857. [doi: 10.1136/bmj.j4857] [Medline: 29070598]
27. Wilken M, Hüske-Kraus D, Klausen A, Koch C, Schlauch W, Röhrig R. Alarm fatigue: causes and effects. Stud Health Technol Inform 2017;243:107-111. [Medline: 28883181]
28. Heng KW. Teaching and evaluating multitasking ability in emergency medicine residents—what is the best practice? Int J Emerg Med 2014;7:41 [FREE Full text] [doi: 10.1186/s12245-014-0041-4] [Medline: 25635201]
29. ALQahtani DA, Rotgans JI, Mamede S, ALAlwan I, Magzoub MEM, Altayeb FM, et al. Does time pressure have a negative effect on diagnostic accuracy? Acad Med 2016;91(5):710-716 [FREE Full text] [doi: 10.1097/ACM.0000000000001098] [Medline: 26826069]
30. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017;92(1):23-30 [FREE Full text] [doi: 10.1097/ACM.0000000000001421] [Medline: 27782919]
31. Ahlbrandt J, Brammen D, Majeed RW, Lefering R, Semler SC, Thun S, et al. Balancing the need for big data and patient data privacy--an IT infrastructure for a decentralized emergency care research database. Stud Health Technol Inform 2014;205:750-754. [Medline: 25160287]
32. Brammen D, Greiner F, Kulla M, Otto R, Schirrmeister W, Thun S, et al. Das AKTIN-notaufnahmeregister—kontinuierlich aktuelle daten aus der akutmedizin: ergebnisse des registeraufbaus und erste datenauswertungen aus 15 notaufnahmen unter besonderer berücksichtigung der vorgaben des Gemeinsamen bundesausschusses zur ersteinschätzung [AKTIN—the German emergency department data registry—real-time data from emergency medicine : implementation and first results from 15 emergency departments with focus on federal joint committee's guidelines on acuity assessment]. Med Klin Intensivmed Notfmed 2022;117(1):24-33 [FREE Full text] [doi: 10.1007/s00063-020-00764-2] [Medline: 33346852]
33. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13(Suppl 1):S31-S34 [FREE Full text] [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]
34. Cicolo EA, Nishi FA, Peres HHC, Monteiro da Cruz DDAL. Effectiveness of the Manchester triage system on time to treatment in the emergency department: a systematic review. JBI Evid Synth 2020;18(1):56-73. [doi: 10.11124/JBISRIR-2017-003825] [Medline: 31453842]
35. Tai-Seale M, McGuire T. Time is up: increasing shadow price of time in primary-care office visits. Health Econ 2012;21(4):457-476 [FREE Full text] [doi: 10.1002/hec.1726] [Medline: 21442688]
36. Wood W, Tam L, Witt MG. Changing circumstances, disrupting habits. J Pers Soc Psychol 2005;88(6):918-933. [doi: 10.1037/0022-3514.88.6.918] [Medline: 15982113]
37. Otto R, Blaschke S, Schirrmeister W, Drynda S, Walcher F, Greiner F. Length of stay as quality indicator in emergency departments: analysis of determinants in the German Emergency Department Data Registry (AKTIN registry). Intern Emerg Med 2022;17(4):1199-1209 [FREE Full text] [doi: 10.1007/s11739-021-02919-1] [Medline: 34989969]
38. Goodacre S. Uncontrolled before-after studies: discouraged by cochrane and the EMJ. Emerg Med J 2015;32(7):507-508 [FREE Full text] [doi: 10.1136/emermed-2015-204761] [Medline: 25820301]
39. Mathe N, Johnson ST, Wozniak LA, Majumdar SR, Johnson JA. Alternation as a form of allocation for quality improvement studies in primary healthcare settings: the on-off study design. Trials 2015;16:375 [FREE Full text] [doi: 10.1186/s13063-015-0904-x] [Medline: 26303892]
40. Hinson JS, Martinez DA, Cabral S, George K, Whalen M, Hansoti B, et al. Triage performance in emergency medicine: a systematic review. Ann Emerg Med 2019;74(1):140-152. [doi: 10.1016/j.annemergmed.2018.09.022] [Medline: 30470513]
41. Zachariasse JM, Seiger N, Rood PPM, Alves CF, Freitas P, Smit FJ, et al. Validity of the Manchester Triage System in emergency care: a prospective observational study. PLoS One 2017;12(2):e0170811 [FREE Full text] [doi: 10.1371/journal.pone.0170811] [Medline: 28151987]
42. Parenti N, Reggiani MLB, Iannone P, Percudani D, Dowding D. A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System. Int J Nurs Stud 2014;51(7):1062-1069. [doi: 10.1016/j.ijnurstu.2014.01.013] [Medline: 24613653]

Abbreviations

AKTIN: alliance for information and communication technology in intensive care and emergency medicine (abbreviated in German as AKTIN)
ED: emergency department
EDIS: emergency department information system
ESI: Emergency Severity Index
LOS: length of stay
MTS: Manchester Triage System
OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

UI: user interface

Edited by T Leung; submitted 09.01.23; peer-reviewed by A Tomar, J Walsh, T Aslanidis, R Matin, A Venkataraman; comments to author 23.12.23; revised version received 02.02.24; accepted 31.03.24; published 14.05.24.

Please cite as:
Bienzeisler J, Becker G, Erdmann B, Kombeiz A, Majeed RW, Röhrig R, Greiner F, Otto R, Otto-Sobotka F, AKTIN Research Group The Effects of Displaying the Time Targets of the Manchester Triage System to Emergency Department Personnel: Prospective Crossover Study
J Med Internet Res 2024;26:e45593
URL: https://www.jmir.org/2024/1/e45593
doi:10.2196/45593
PMID:38743464

©Jonas Bienzeisler, Guido Becker, Bernadett Erdmann, Alexander Kombeiz, Raphael W Majeed, Rainer Röhrig, Felix Greiner, Ronny Otto, Fabian Otto-Sobotka, AKTIN Research Group AKTIN Research Group. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Experiences of Electronic Health Records' and Client Information Systems' Use on a Mobile Device and Factors Associated With Work Time Savings Among Practical Nurses: Cross-Sectional Study

Satu Paatela[1], MHSc; Maiju Kyytsönen[1], MHSc; Kaija Saranto[2], PhD; Ulla-Mari Kinnunen[3], PhD; Tuulikki Vehko[1], PhD

[1]Health and Social Service System Research, Finnish Institute for Health and Welfare, Helsinki, Finland
[2]Faculty of Social Sciences and Business Studies, University of Eastern Finland, Kuopio, Finland
[3]Department of Health and Social Management, University of Eastern Finland, Kuopio, Finland

Corresponding Author:
Satu Paatela, MHSc
Health and Social Service System Research
Finnish Institute for Health and Welfare
Mannerheimintie 166
Helsinki, 00271
Finland
Phone: 358 29 524 77 22
Email: satu.paatela@thl.fi

Abstract

Background: The transmission of clinical information in nursing predominantly occurs through digital solutions, such as computers and mobile devices, in today's era. Various technological systems, including electronic health records (EHRs) and client information systems (CISs), can be seamlessly integrated with mobile devices. The use of mobile devices is anticipated to rise, particularly as long-term care is increasingly delivered in environments such as clients' homes, where computers are not readily accessible. However, there is a growing need for more user-centered data to ensure that mobile devices effectively support practical nurses in their daily activities.

Objective: This study aims to analyze practical nurses' experiences of using EHRs or CISs on a mobile device in their daily practice. In addition, it aims to examine the factors associated with work time savings when using EHRs/CISs on a mobile device.

Methods: A cross-sectional study using an electronic survey was conducted in spring 2022. A total of 3866 practical nurses participated in the survey based on self-assessment. The sample was limited to practical nurses who used EHRs or CISs on a mobile device and worked in home care or service housing within the social welfare or health care sector (n=1014). Logistic regression analysis was used to explore the factors associated with work time savings.

Results: The likelihood of perceiving work time savings was higher among more experienced EHR/CIS users compared with those with less experience (odds ratio [OR] 1.59, 95% CI 1.30-1.94). Participants with 0-5 years of work experience were more likely to experience work time savings compared with those who had worked 21 years or more (OR 2.41, 95% CI 1.43-4.07). Practical nurses in home care were also more likely to experience work time savings compared with those working in service housing (OR 1.95, 95% CI 1.23-3.07). A lower grade given for EHRs/CISs was associated with a reduced likelihood of experiencing work time savings (OR 0.76, 95% CI 0.66-0.89). Participants who documented client data in a public area were more likely to experience work time savings compared with those who did so in the nurses' office (OR 2.33, 95% CI 1.27-4.25). Practical nurses who found documentation of client data on a mobile device easy (OR 3.05, 95% CI 2.14-4.34) were more likely to experience work time savings compared with those who did not. Similarly, participants who believed that documentation of client data on a mobile device reduced the need to memorize things (OR 4.10, 95% CI 2.80-6.00) were more likely to experience work time savings compared with those who did not.

Conclusions: To enhance the proportion of practical nurses experiencing work time savings, we recommend that organizations offer comprehensive orientation and regular education sessions tailored for mobile device users who have less experience using EHRs or CISs and find mobile devices less intuitive to use.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e46954) doi:10.2196/46954

KEYWORDS

practical nurse; information and communication technology; electronic health record; client information system; documentation; mobile technology

Introduction

Information and communication technology (ICT) including electronic health records (EHRs) and client information systems (CISs) have become an increasingly important part of health care and social services in Finland [1-4]. In fact, EHR has been a common tool in Finnish health care for a long time, and from 2010 onward EHR availability has been 100% in public and private facilities [5]. EHRs include a comprehensive collection of patient health information (eg, narrative texts and laboratory data), with the collected data used in the care processes of the patient [6]. By contrast, CISs are more commonly used in the public social welfare sector for accessing, storing, and using client information and documents [3,4]. The Finnish Ministry of Social Affairs and Health has been at the helm of guiding the development of technological infrastructure and associated legislative efforts over the decades. Their eHealth and eWelfare strategy aims to improve the effective utilization of data in bolstering service renewal and citizen well-being at a national level [7].

With the development of the ICT infrastructure, the clinical information in nursing is nowadays mainly transmitted via digital solutions such as computers and mobile devices [8]. Particularly, mobile devices such as smartphones and tablets are commonly used in the health care sector [9,10], and nurses are known to use EHRs on mobile devices 3 times more often compared with doctors in a hospital environment [11]. Different technological systems such as EHRs can be integrated with mobile devices [9,11-13], and this allows social and health care professionals to document client data and exchange information related to the clients and service assignments in the system regardless of the time and location [11,13]. Thus, the ability to receive nursing information via a mobile device can promote the mobility and portability of care and enhance service flexibility [9,12].

In previous studies, nurses have been reported to perceive mobile devices as beneficial in their daily practice [14,15]. Mobile device use has been found to support nurses' workflow processes [11,16-18] as they need to spend less time on clinical documentation [17]. In fact, it is important to recognize that the use of different technological systems such as EHRs has been shown to provide nurses with more time for direct patient care and interprofessional communication [19]. Moreover, mobile devices have been reported to be useful for planning work, handling notes [14], and saving time for nurses [14,20,21]. Mobile devices may also reduce duplicate documentation [13] and potential documentation errors [17,20] because client data can be documented at the time of its occurrence. In addition, improved decision-making is one of the main advantages [9,20]. Mobile devices continuously provide the latest information on the situation of the clients, which can improve safety and the quality of care [14,18]. For instance, in home care settings, workers can review and plan real-time nursing interventions and tasks in the clients' home environment [13,16] because the daily assignments can be seen directly on the mobile device. The use of mobile devices can also contribute to client empowerment as nurses have easier access to clinical information and they can reply to clients' questions more readily [17].

By contrast, some challenges have emerged related to the use of EHRs/CISs on a mobile device. According to a recent study, the use of EHRs on a mobile device can negatively affect nurses' well-being because the use of mobile devices may increase time pressure and stress at work as a result of functional problems and changing information systems [10]. In home care settings, challenges have emerged, particularly concerning technical issues such as usability problems. This is because the information systems used on mobile devices are not always tailored to address the specific requirements of various working environments [22]. This in turn may lead to reduced workflows [22]. Additionally, there is a possibility of poor signal connectivity [18] and increased data security threats [17,18,23] when using mobile devices. Furthermore, some of the recent studies have observed that the use of EHRs itself may negatively impact the quality of communication between nurses and patients because nurses' attention is more focused on documentation tools such as computers rather than on the patient [24,25]. More generally, the lack of digital competence can affect how different digital tools are adopted in practice [26,27].

The European Union has promoted digitalization in society, including public services, with political consensus through the Digital Decade policy program [28]. The change in the demographic structure especially forces social and health care services to invest even more in technological solutions [29] such as mobile documentation [22]. An aging population increases the need to provide long-term care in home environments [30], and therefore, using EHRs and CISs on a mobile device is expected to grow as computers for professionals are rarely available at the patients' bedside in the home environment [14]. In the Finnish context, practical nurses often take care of needs related to the activities of daily living, for instance, in home care and service housing. Practical nurses in the social and health care profession are strictly regulated by law in Finland [31,32]. Practical nurses are required to have successfully obtained the Vocational Qualification in Social and Healthcare, which entails accruing 180 competence points [33]. Qualified practical nurses are registered with the National Supervisory Authority for Welfare and Health. They are employed across a diverse spectrum of careers within the social welfare and health care sectors, as well as in early childhood education and schools [34]. Practical nurses are the second largest occupational group in Finland and the largest group in the social welfare and health care sectors in Finland, with 79,800 people working as practical nurses at the end of 2020 [35].

XSL•FO
RenderX

As practical nurses form an important group of professionals, it is justified to study their ways of working and increase our knowledge about their experiences of EHR and CIS use on mobile devices. Some of the previous studies have investigated the use of mobile devices from the perspectives of registered nurses, nursing students, and doctors [10,11,14,15], but there is still limited understanding of the experiences of practical nurses. More user-centered data are needed to ensure that mobile devices fit into the changing clinical practice [18] and to improve health professionals' workflows in those work environments where mobile devices are commonly used. As patient care becomes increasingly complex [8] and health professionals are constantly required to work more efficiently [36], it is important to study whether mobile devices are as effective tools as they are expected to be in the daily activities of practical nurses [9,11,13,14,16,17].

Consequently, the aim of this study was to analyze practical nurses' experiences regarding their use of EHRs/CISs on a mobile device in their daily practice in home care and service housing settings in the social welfare and health care sectors. Furthermore, we examined the potential factors associated with work time savings when practical nurses were using EHRs/CISs on a mobile device.

Methods

Study Context

Finland is a Nordic welfare state where all citizens have universal access to health care and social welfare services. In the 2000s, long-term care for older people and persons with disabilities in Finland has changed from institutional care to more individualized services [37]. In Finland, long-term care is increasingly provided in service housing or in the home environment under social services. Service housing is available for those citizens who need support living independently. These facilities offer a range of services including meal provision, assistance with personal hygiene, and various health care services [38]. However, most older adults continue to reside in their own homes, where they can access home care services if needed. Home care encompasses health center–based home nursing and home help services [37,38]. Finland has a wide array of EHRs and CISs, which are used across both the health care and social welfare sectors [6].

Study Design and Sample

This was a cross-sectional study based on an electronic survey. Data were collected in the spring of 2022 over a 3-week period using a convenience sampling method. As of the end of 2020, there were 79,800 practical nurses employed in Finland [35], working across the social welfare and health care sectors, as well as in schools and early childhood education and care. Potential respondents were invited to participate in the survey through an email sent by 2 trade unions: The Finnish Union of Practical Nurses and The Union of Public and Welfare Sectors. The electronic survey was distributed to 54,030 members of the trade unions aged 18-65 years. The cover letter specified the study theme as the use of EHRs and CISs. However, previous studies indicate that not all members of the trade unions use EHRs/CISs in their daily practice. This is because practical nurses in social services may still rely on alternative solutions for documentation [6]. Despite this, 2 reminders were sent to potential participants. Ultimately, 3866 practical nurses responded to the survey, yielding a response rate of 7.16%.

In this study, the inclusion criteria for participation were 2-fold: (1) respondents must work as practical nurses and use an EHR or CIS, and (2) they must not be employed in school or early childhood education and care settings. These criteria were outlined in the first 2 questions of the survey, and the survey was closed for potential respondents who did not meet these criteria. The analysis was additionally narrowed down to practical nurses who indicated that they use EHRs or CISs on a mobile device and are employed in either home care or service housing settings (n=1014). Respondents working in other employment settings were excluded because of the limited number of mobile device users in those settings.

Instrument

The experiences of EHR systems among physicians were initially assessed in Finland through a national survey in 2010 [39,40]. Subsequently, the survey was refined and conducted again in 2014, 2017, and 2021 for physicians. Additionally, it was customized for registered nurses in 2017 [41-43] and for social care professionals (educated at a university or a university of applied sciences) in 2020 [44,45]. Since 2014, these national surveys have been carried out as part of the "Monitoring and Assessment of Social Welfare and Health Care Information System Services" (STePS) projects [6]. In a significant development, for the first time in 2022, the survey was customized and conducted for practical nurses as well. Before data collection, the survey underwent pretesting with 20 practical nurses. Questions regarding the use of EHRs and CISs on mobile devices were particularly emphasized, given their integral role in the workflow of many practical nurses. As a result, this study specifically centered on the utilization of a mobile device for the documentation of client data.

A total of 11 variables from the survey were covered in this study. The "Documentation of client data on a mobile device saves working time" variable was used as an outcome measure. To understand what kind of factors are connected to work time savings, the following variables were used: "Age," "Work experience," "Workplace," "Experience in using EHR/CIS," "Grade for EHR/CIS," "Most common place to document client data on a mobile device," "Received sufficient training to document client data on a mobile device," "Documentation of client data on a mobile device is easy," "Documentation of client data on a mobile device reduces the need to memorize," and "I can document everything I need on a mobile device." A total of 9 variables were recoded in the analysis and 2 variables were included as a continuous variable. The 5-point Likert scale was specified in 5 different variables as follows: 1=fully agree, 2=agree, 3=neither agree nor disagree, 4=disagree, and 5=fully disagree. To streamline the focus on the phenomena of interest and to ensure an adequate number of respondents in all categories, the response options were recoded as follows: 1 or 2=yes and 3-5=no. The included variables are presented in Table 1.

XSL•FO
RenderX

Table 1. The included variables in the analysis.

Variable	Item in the survey	Response options	Coded in the analysis
Age	Year of birth	1957-2003	18-34 years, 35-54 years, and 55-65 years
Work experience	How long have you worked as a practical nurse (or equivalent)?	1=under 1 year, 2=1-2 years, 3=3-5 years, 4=6-10 years, 5=11-15 years, 6=16-20 years, and 7=over 20 years	1-3=0-5 years, 4=6-10 years, 5=11-15 years, 6=16-20 years, and 7=21 years or more
Workplace	The place of the main employment	The 3 largest places of the main employment were included in the study: 1=hospital-based home care, 2=residential care home, and 3=home care	1 or 2=service housing and 3=home care
Experience in using EHR[a]/CIS[b]	How experienced do you consider yourself to be as an EHR/CIS user?	The answer options were rated from 1=beginner to 5=highly experienced	Included as a continuous variable
Grade for EHR/CIS	How would you rate the EHR/CIS you use on a mobile device?	On a scale from 4 to 10, with 4 being the lowest score and 10 being the highest score	Included as a continuous variable
Most common place to document client data on a mobile device	What is the most common place to document client data on a mobile device?	1=next to the patient, 2=in a public area (eg, corridor), 3=on the streets, 4=at the office/nurses' office, 5=in one's car, 6=in the public transport, and 7=other	1=next to the patient, 2=in a public area, 4=at the (nurses') office, 5=in one's car, and 3,6,7=other
Received sufficient training to document client data on a mobile device	I have received sufficient training to document client data on a mobile device	5-point Likert scale[c]	Binary variables: 1 or 2=yes and 3-5=no
Documentation of client data on a mobile device is easy	Documentation of client data on a mobile device is easy	5-point Likert scale	Binary variables: 1 or 2=yes and 3-5=no
Documentation of client data on a mobile device saves working time	Documentation of client data on a mobile device saves working time.	5-point Likert scale	Binary variables: 1 or 2=yes and 3-5=no
Documentation of client data on a mobile device reduces the need to memorize things	Documentation of client data on a mobile device reduces the need to memorize things.	5-point Likert scale	Binary variables: 1 or 2=yes and 3-5=no
I can document everything I need on a mobile device	I can document everything I need on a mobile device.	5-point Likert scale	Binary variables: 1 or 2=yes and 3-5=no

[a]EHR: electronic health record.

[b]CIS: client information system.

[c]The 5-point Likert scale was specified as follows: 1=fully agree, 2=agree, 3=neither agree nor disagree, 4=disagree, and 5=fully disagree.

Data Analysis

The data were analyzed using the statistical software SPSS Statistics version 29.0.0.0 (IBM, Inc.). The characteristics of the study participants were described using n (%). A binary logistic regression analysis was conducted to examine the association between independent and dependent variables. The "Documentation of client data on a mobile device saves working time" item was used as a dependent variable and 10 items were used as independent variables in the analysis. In establishing a model for the relationship between independent and dependent variables, we first tested the significance of each independent variable individually according to the Wald *F* test. Based on the *P* values ($P<.05$) of the Wald *F* test, the items "Age" and "I can document everything I need on a mobile device" were excluded from the regression analysis model. We included 8 other independent variables one by one in the model using a forward stepwise selection method. At each step, variables were chosen for the final model according to their effect on the model's goodness-of-fit measure, Nagelkerke R^2 (R^2_N), and *P* values of the Wald *F* test. The fully adjusted model included 7 independent variables, including "Experience in using EHR/CIS," "Work experience," "Workplace," "Grade for EHR/CIS," "Most common place to document client data on a mobile device," "Documentation of client data on a mobile device is easy," and "Documentation of client data on a mobile device reduces the need to memorize." The "Received sufficient training to document client data on a mobile device" item was omitted from the final model because it was no longer statistically significant ($P=.08$) after adjusting the final variable to the model. The fully adjusted model's R^2_N was 0.372. Variance inflation factor was used to secure a model without multicollinearity: the values indicated low correlation, which is acceptable in a regression model. The results of the fully adjusted regression analysis model are presented with *P* values, variance inflation factor, odds ratios, and their 95% CIs in Table 4.

XSL•FO
RenderX

Ethical Considerations

We considered ethical issues related to different phases of this study. Ethical approval for the study was provided by the Finnish Institute for Health and Welfare THL/1206/6.02.01/2022. Study participants were offered written information on the research and data processing in a cover letter and privacy notice [46]. Participants did not receive any compensation for their participation in the study. The research group has been committed to protecting the anonymity of the participants throughout the study process.

Results

Characteristics of the Mobile Device Users

Of the total of 1014 practical nurses who used EHRs or CISs on a mobile device, nearly one-half (471/1014, 46.45%) fell within the age range of 35-54 years. Additionally, there was a relatively high proportion of participants who were at least 55 years old. The work experience among participants was diverse and evenly distributed. For example, a portion of practical nurses (195/1014, 19.23%) had 0-5 years of experience as a practical nurse or equivalent, while others had worked for 21 years or more (238/1014, 23.47%). The majority of mobile device users (706/1014, 69.63%) were employed in home care, with the remainder working in service housing. Nearly half of the mobile device users (458/1014, 45.17%) rated their experience of using EHRs or CISs at level 4 (on a scale of 1 to 5, where 1 represents a beginner and 5 represents highly experienced). Only 4 practical nurses rated themselves as beginners in using EHRs/CISs. Additionally, the majority of practical nurses assessed the EHR/CIS system used via a mobile device as good (364/1014, 35.90%) or satisfactory (271/1014, 26.73%; Table 2).

XSL•FO
RenderX

Table 2. Characteristics of the study participants (n=1014).

Characteristics	Value, n (%)
Age (years)	
18-34	156 (15.38)
35-54	471 (46.45)
55-65	387 (38.17)
Missing data	0 (0)
Work experience (years)	
0-5	195 (19.23)
6-10	232 (22.88)
11-15	220 (21.70)
16-20	129 (12.72)
21 or more	238 (23.47)
Missing data	0 (0)
Workplace	
Service housing	308 (30.37)
Home care	706 (69.63)
Missing data	0 (0)
Experience using EHR[a]/CIS[b]	
1 (beginner)	4 (0.39)
2	47 (4.64)
3	287 (28.30)
4	458 (45.17)
5 (highly experienced)	218 (21.50)
Missing data	0 (0)
Grade for EHR/CIS	
10 (Excellent)	21 (2.07)
9 (Very good)	110 (10.85)
8 (Good)	364 (35.90)
7 (Satisfactory)	271 (26.73)
6 (Moderate)	140 (13.81)
5 (Adequate)	90 (8.88)
4 (Fail)	14 (1.38)
Missing data	4 (0.39)

[a]EHR: electronic health record.

[b]CIS: client information system.

Practical Nurses' Experiences of Documenting Client Data on a Mobile Device

The most prevalent location for documenting client data on a mobile device was next to the client (537/1014, 52.96%). Some practical nurses also documented client data in alternative settings such as in the car, at the office, or in public areas. Overall, mobile device users expressed relatively high satisfaction with the training they received for documenting client data on a mobile device (661/1014, 65.19%). The majority of mobile device users (648/1014, 63.91%) found it easy to document client data on a mobile device. Additionally, two-thirds of practical nurses (667/1014, 65.78%) reported that documenting client data on a mobile device saved them time. Furthermore, a vast majority of mobile device users (785/1014, 77.42%) agreed that documenting client data on a mobile device reduced the need to rely on memory. Less than one-half of the participants (418/1014, 41.22%) agreed that they could document everything they need on a mobile device (Table 3).

XSL•FO
RenderX

Table 3. Practical nurses' experiences of documenting client data on a mobile device (n=1014).

Variable	Value, n (%)
Most common place to document client data on a mobile device	
Next to the client	537 (52.96)
In a public area (eg, corridor)	135 (13.31)
At the (nurses') office	133 (13.12)
In one's car	175 (17.26)
Other	31 (3.06)
Missing data	3 (0.30)
Received sufficient training to document client data on a mobile device	
No	350 (34.52)
Yes	661 (65.19)
Missing data	3 (0.30)
Documentation of client data on a mobile device is easy	
No	361 (35.60)
Yes	648 (63.91)
Missing data	5 (0.49)
Documentation of client data on a mobile device saves working time	
No	343 (33.83)
Yes	667 (65.78)
Missing data	4 (0.39)
Documentation of client data on a mobile device reduces the need to memorize things	
No	222 (21.89)
Yes	785 (77.42)
Missing data	7 (0.69)
I can document everything I need on a mobile device	
No	594 (58.58)
Yes	418 (41.22)
Missing data	2 (0.20)

Factors Associated With Work Time Savings When Using EHRs/CISs on a Mobile Device

Several factors were associated with work time savings when using EHRs/CISs on a mobile device (Table 4). Experience of using EHRs/CISs (P<.001), work experience (P<.001), the workplace (P=.004), the grade given for the EHRs/CISs (P<.001), the statements "Documentation of patient data on a mobile device is easy" (P<.001) and "Documentation of patient data on a mobile device reduces the need to memorize things" (P<.001) had statistically significant associations with work time savings.

XSL•FO
RenderX

Table 4. The results of the fully adjusted logistic regression analysis model for the practical nurses' experience of work time savings when using EHRs[a]/CISs[b] on a mobile device.

Variable	Odds ratio (95% CI)	*P* value	Variance inflation factor
Experience in using EHR/CIS	1.59 (1.30-1.94)	*<.001* [c]	1.09
Work experience (years)		*<.001*	1.03
0-5	2.41 (1.43-4.07)	*.001*	
6-10	1.34 (0.84-2.13)	.23	
11-15	1.03 (0.64-1.67)	.89	
16-20	0.52 (0.31-0.87)	*.01*	
21 or more	1[d]	—[e]	
Workplace		*.004*	1.18
Home care	1.95 (1.23-3.07)		
Service housing	1		
Grade for EHR/CIS, on a scale of 4-10	0.76 (0.66-0.89)	*<.001*	1.40
The most common place to document client data on a mobile device		.08	1.17
Next to the client	1.66 (0.96-2.88)	.07	
In a public area (eg, corridor)	2.33 (1.27-4.25)	*.006*	
In one's car	1.84 (0.96-3.53)	.07	
Other	2.09 (0.73-5.94)	.17	
At the (nurses') office	1	—	
Documentation of client data on a mobile device is easy		*<.001*	1.44
Yes	3.05 (2.14-4.34)		
No	1		
Documentation of client data on a mobile device reduces the need to memorize things		*<.001*	1.20
Yes	4.10 (2.80-6.00)		
No	1		

[a]EHR: electronic health record.

[b]CIS: client information system.

[c]The level of statistical significance was set at *P*<.05 (italicized).

[d]Comparison group.

[e]Not applicable.

Practical nurses who considered themselves to be more experienced EHR/CIS users were more likely to perceive work time savings. Participants who had worked 0-5 years as a practical nurse were 2.41 times more likely to experience work time savings compared with those who had worked 21 years or more. Practical nurses who had worked 16-20 years had a lower likelihood of experiencing work time savings than those who had worked for 21 years or more. Furthermore, practical nurses who worked in home care settings were 1.95 times more likely to report work time savings compared with those participants who worked in service housing. Giving a lower grade for EHRs/CISs was associated with a lower likelihood of experiencing work time savings. Participants who documented client data in a public area were 2.33 times more likely to experience work time savings compared with those who documented client data at the (nurses') office. Moreover, those practical nurses who reported that the documentation of client data on a mobile device was easy were 3.05 times more likely to experience work time savings compared with others. Practical nurses who reported that the documentation of client data on a mobile device reduced their need to memorize things were 4.10 times more likely to experience work time savings compared with those who did not find mobile devices helpful in memorizing things.

Discussion

Principal Findings

The aim of the study was to analyze practical nurses' experiences of using EHRs/CISs on a mobile device in their daily practice. Our study findings indicate that practical nurses generally had positive experiences when documenting client data on a mobile device. Two-thirds of the participants perceived mobile devices as effective tools in their daily practice, as they

XSL•FO
RenderX

facilitated time savings in their work. The study revealed that a vast majority of the participants agreed that using EHRs/CISs on a mobile device reduced the need to memorize things. However, participants were less inclined to agree with the statement that they could document everything they needed on a mobile device.

Additionally, our study examined factors associated with work time savings when practical nurses used EHRs/CISs on a mobile device. Factors such as experience with the EHRs/CISs, work experience, workplace, the grade awarded for the EHRs/CISs, and responses to statements such as "Documentation of patient data on a mobile device is easy" and "Documentation of patient data on a mobile device reduces the need to memorize things" were all found to be associated with practical nurses' experiences of work time savings.

Limitations

This study has several limitations. First, the response rate of 7.16% (3866/54,030) was relatively low, which is common for web-based and lengthy surveys [47], especially those aimed at health care professionals [48]. Additionally, incorrect email addresses due to job changes or other reasons, as well as nonopened survey emails, may have further contributed to the low response rate. Therefore, the actual response rate might have been higher if calculated only for those who received and opened the survey invitation. Eventually, 3866/4533 (85.29%) survey clicks resulted in respondents completing the survey. However, it is worth noting that data collection occurred during a national industrial action organized by the trade unions, which could have complicated survey participation. Additionally, various work-related factors that practical nurses encounter in their daily routines, such as time constraints and interruptions, may have influenced survey response rates, especially considering that many union members use their work email as their contact information. Furthermore, the utilization of the convenience sampling method may restrict the generalizability of the results. However, the age distribution of the respondents mirrored that of individuals affiliated with national trade unions [49]. Additionally, the survey was available in both of Finland's official languages, Finnish and Swedish, potentially encouraging speakers of both languages to participate.

Second, while practical nurses are a common occupational group in Finland, their title may be less recognized in other regions worldwide. Indeed, long-term care may be provided by health care professionals with various occupational titles internationally. Nonetheless, we propose that the findings of our study can be applied to other nursing professions, such as registered nurses and health care assistants, who use mobile devices as documentation tools in their daily practice. Furthermore, it is important to acknowledge that Finland has a long-standing history of extensively using ICT tools in health care [5]. Moreover, Finland ranks among the global leaders in mobile data usage [50]. Consequently, the findings of this study may be particularly relevant and applicable to countries with similar levels of ICT development.

Third, the survey was customized for practical nurses in Finland for the first time, including the questions related to mobile device use. Given the low proportion of missing data, we can assume that respondents understood the various items of the instrument relatively well. Before distribution to participants, the instrument underwent pretesting with 20 practical nurses.

In future studies, it would be beneficial to investigate work time savings among users of specific EHR/CIS brands, as the grading of the system by respondents was strongly correlated with experiencing work time savings. Additionally, research should explore specific work environments, such as home care and service housing. Hence, conducting a subgroup analysis separately for practical nurses working in home care and service housing would be a valuable addition to future studies. Another important research avenue would be to explore the barriers that practical nurses may encounter when documenting next to the patient using a mobile device.

Comparison With Prior Work

To the best of our knowledge, this study marks the inaugural exploration of practical nurses' experiences regarding their use of EHRs/CISs on a mobile device. Our primary focus was to investigate whether the use of EHRs/CISs on a mobile device contributes to time savings for practical nurses, as well as to identify the factors associated with such savings. In the health care sector, saving work time is crucial because nursing professionals are tasked with a multitude of responsibilities in their daily practice. It is essential for them to have more time available for direct patient care and to minimize the time spent on indirect patient care activities, such as documentation [21].

This study revealed that two-thirds of practical nurses working in home care or service housing experienced work time savings when using EHRs/CISs on a mobile device. Comparable findings of work time savings have also been documented in previous studies involving health care professionals [14,20,21]. This study revealed that documenting client data in a public area, such as a corridor in a housing service, was over 2 times more likely to result in work time savings compared with documenting at the nurses' office, where computers are typically available. However, it is important to note that documenting sensitive client data on mobile devices in a public area may pose increased security risks, such as the potential loss or theft of the mobile device [23]. Therefore, mobile technology tools should incorporate essential security features, and organizations should establish clear policies regarding the management of mobile devices [51].

According to our study findings, work experience was linked to work time savings when using EHRs/CISs on a mobile device. Participants who had worked 0-5 years as practical nurses or in equivalent roles were over 2 times more likely to experience work time savings compared with those who had worked for over 21 years. We speculate that practical nurses with less work experience may perceive work time savings more frequently because they are accustomed to working with new technologies in their daily practice, and they may have received more recent orientation on using mobile devices. It is interesting to note that, in our analysis, age was not found to be significantly associated with work time savings when using EHRs/CISs on a mobile device. However, age may influence perceptions regarding the use of mobile devices. Findings from a previous study [52] have

XSL•FO
RenderX

suggested that older nurses are less inclined to use smartphones or acknowledge their benefits in acute care settings.

Additionally, our study revealed that practical nurses working in home care settings were nearly two times as likely to report work time savings compared with those working in service housing. This finding is unsurprising, considering that home care relies on mobility and necessitates the use of ICT tools directly at patients' homes [22]. This environment naturally fosters the integration of mobile technology into the daily practices of health care workers. An essential prerequisite for realizing the benefits of mobile technology is seamless integration with the existing information systems [18], such as EHRs/CISs. It could be hypothesized that mobile devices contribute to work time savings for practical nurses, especially in home care settings, by facilitating the documentation of client data immediately after completing daily tasks [13], such as next to the client. However, although practical nurses in this study often documented data next to the client, it was not identified as a statistically significant factor for work time savings. The immediacy afforded by mobile devices, allowing users to document client data promptly after interacting with the client, can alleviate the burden of memorization for health care professionals. According to our study findings, practical nurses who perceived that the documentation of client data reduced the need to rely on memory were 4 times more likely to report work time savings compared with those who did not find mobile devices helpful in reducing the need to memorize things.

Our study findings revealed that practical nurses who found the documentation of client data on a mobile device to be easy were over 3 times more likely to experience work time savings compared with those who did not find mobile devices easy to use. Furthermore, Zhang et al [53] discovered that nursing professionals in home care settings perceived mobile devices to be useful if the tools are easy to use. Overall, while usability issues related to health information systems, including EHRs, are widely recognized [2,54], much of the existing data are centered around the use of these systems on computers. However, it is important to note that using EHRs/CISs specifically on mobile devices may present additional challenges for social and health care professionals. For example, previous studies have indicated that mobile devices may be difficult to use, too small for daily practice [16], may not function properly at all times [13], and could be unstable due to potential internet connection problems [18,22].

In this study, the grade provided by respondents for the EHRs/CISs on a mobile device emerged as a factor associated with work time savings. Specifically, a lower grade for the EHRs/CISs was linked to a reduced likelihood of experiencing work time savings. As the grade for the EHRs/CISs may reflect user satisfaction to some extent, this finding underscores the significance of prioritizing user satisfaction regarding practical nurses' use of EHRs/CISs on mobile devices. User satisfaction has indeed garnered significant attention in previous studies [9,15], and its impact extends beyond work time savings. According to Hsiao and Chen [9], user satisfaction influences nurses' intention to continue using information systems on mobile devices, and perceived usefulness is often intertwined with user satisfaction. Furthermore, the quality of the information system and support from managers have been identified as significant predictors of user satisfaction [15], as well as technology adoption in general [18]. It is important to highlight that health care professionals who are more experienced users of information systems may offer valuable suggestions for improvements [9], underscoring the importance of involving these users in the development of EHRs/CISs to ensure user satisfaction with the system interfaces.

When assessing potential work time savings, it is crucial to take into account practical nurses' experiences with using EHRs/CISs. Our study results indicate that practical nurses with more experience in using EHRs/CISs were more likely to experience work time savings. Similarly, Villalba-Mora et al [26] discovered that health care professionals who frequently used health information technologies such as EHRs perceived these tools to be more useful. Additionally, previous experience with digital technologies is significant, as it aids health care professionals in integrating mobile devices into their daily practices [18].

Conclusions

This study contributes to the existing literature on the use of EHRs/CISs on a mobile device by practical nurses in their daily practice, as well as factors associated with work time savings. Our findings indicate that two-thirds of practical nurses perceived mobile devices as beneficial in home care and service housing settings, as they reported that documenting client data on a mobile device saved their working time. Experience in using EHRs/CISs, work experience, workplace, grade given for the EHRs/CISs, and perceptions regarding the ease of documentation and reduction in the need to memorize were all significantly associated with practical nurses' experiences of work time savings. Based on our findings, we recommend that special attention should be directed toward mobile device users who are less experienced in using EHRs/CISs or do not find mobile devices easy to use. Organizations should provide comprehensive orientation and regular education to health care professionals on the use of EHRs/CISs on mobile devices. Additionally, user satisfaction is a crucial aspect to consider in achieving work time savings among health care professionals who use EHRs/CISs on a mobile device, as demonstrated by our findings. Practical nurses who rated their EHRs/CISs more favorably were more likely to experience work time savings. Therefore, we suggest that end users, particularly those with more experience in using EHRs/CISs, should be involved in the development of EHRs/CISs to ensure better user satisfaction of system interfaces.

Acknowledgments

The research leading to these results received funding from the Ministry of Social Affairs and Health, Finland. The authors thank the Finnish Federation of Local and Basic Caregivers (SuPer ry) and the Union of Public and Welfare Sectors (JHL) for sending

XSL•FO
RenderX

the survey to their members and all the licensed practical nurses who responded to the survey. We also thank researcher Samuel Salovaara for commenting on the survey questionnaire. This study was supported by the Ministry of Social Affairs and Health, Finland (project 414919001). The funder' did not have any role in the study design, collection, analysis, interpretation of data, or writing of the report.

Data Availability

The data collected and analyzed during this study are not publicly available due to coregistratorship between the University of Eastern Finland and the Finnish Institute for Health and Welfare, the controllers of the data.

Conflicts of Interest

None declared.

References

1. Kinnunen U, Heponiemi T, Rajalahti E, Ahonen O, Korhonen T, Hyppönen H. Factors related to health informatics competencies for nurses—results of a national electronic health record survey. CIN: Computers, Informatics, Nursing 2019;37(8):420-429. [doi: 10.1097/cin.0000000000000511]
2. Kaihlanen A, Gluschkoff K, Hyppönen H, Kaipio J, Puttonen S, Vehko T, et al. The associations of electronic health record usability and user age with stress and cognitive failures among Finnish registered nurses: cross-sectional study. JMIR Med Inform 2020 Nov 18;8(11):e23623 [FREE Full text] [doi: 10.2196/23623] [Medline: 33206050]
3. Salovaara S, Ylönen K. Client information systems' support for case-based social work: experiences of Finnish social workers. Nordic Social Work Research 2021 Nov 10;12(3):364-378. [doi: 10.1080/2156857x.2021.1999847]
4. Ylönen K. The use of electronic information systems in social work. A scoping review of the empirical articles published between 2000 and 2019. European Journal of Social Work 2022 Apr 18:1-14. [doi: 10.1080/13691457.2022.2064433]
5. Reponen J, Keränen N, Ruotanen R, Tuovinen T, Haverinen J, Kangas M. Availability and use of e-health in Finland. In: Vehko T, editor. E-Health and E-Welfare of Finland: Checkpoint 2022. Helsinki, Finland: National Institute for Health and Welfare; 2022:61-95.
6. Vehko T, editor. E-Health and E-Welfare of Finland : Check Point 2022 [Report 2022_006]. Helsinki, Finland: National Institute for Health and Welfare; 2022.
7. Jormanainen V, Hämäläinen P, Reponen J. The Finnish healthcare and social care system and ICT-policies. In: Vehko T, editor. E-Health and E-Welfare of Finland: Checkpoint 2022. Helsinki, Finland: National Institute for Health and Welfare; 2022:61-95.
8. Silva AMDA, Mascarenhas VHA, Araújo SNM, Machado RDS, Santos AMRD, Andrade EMLR. Mobile technologies in the nursing area. Rev Bras Enferm 2018 Oct;71(5):2570-2578 [FREE Full text] [doi: 10.1590/0034-7167-2017-0513] [Medline: 30304191]
9. Hsiao J, Chen R. Understanding determinants of health care professionals' perspectives on mobile health continuance and performance. JMIR Med Inform 2019 Mar 18;7(1):e12350 [FREE Full text] [doi: 10.2196/12350] [Medline: 30882353]
10. Heponiemi T, Kaihlanen A, Gluschkoff K, Saranto K, Nissinen S, Laukka E, et al. The association between using a mobile version of an electronic health record and the well-being of nurses: cross-sectional survey study. JMIR Med Inform 2021 Jul 06;9(7):e28729 [FREE Full text] [doi: 10.2196/28729] [Medline: 34255704]
11. Lee Y, Park YR, Kim J, Kim JH, Kim WS, Lee J. Usage pattern differences and similarities of mobile electronic medical records among health care providers. JMIR Mhealth Uhealth 2017 Dec 13;5(12):e178 [FREE Full text] [doi: 10.2196/mhealth.8855] [Medline: 29237579]
12. Lin T. Mobile nursing information system utilization: the task-technology fit perspective. CIN:Computers, Informatics, Nursing 2014;32(3):129-137. [doi: 10.1097/cin.0000000000000039]
13. Han S, Juell-Skielse G, Smedberg, Aasi P, Nilsson AG. Benefits of mobile reporting systems in social home care: the case of seven Swedish municipalities. Int J Technol Assess Health Care 2014 Oct;30(4):409-415. [doi: 10.1017/S026646231400052X] [Medline: 25425318]
14. Johansson P, Petersson G, Saveman B, Nilsson G. Using advanced mobile devices in nursing practice--the views of nurses and nursing students. Health Informatics J 2014 Sep 02;20(3):220-231 [FREE Full text] [doi: 10.1177/1460458213491512] [Medline: 25183609]
15. Cheng C, Chan C, Chen L, Guo S. Evaluation of the implementation of a mobile nursing information system. Online Journal of Nursing Informatics 2019;23(3):1 [FREE Full text]
16. Schachner M, Sommer J, González Z, Luna D, Benítez S. Evaluating the feasibility of using mobile devices for nurse documentation. Nursing Informatics 2016;225:495-499. [doi: 10.3233/978-1-61499-658-3-495]
17. Ehrler F, Lovis C, Blondon K. A mobile phone app for bedside nursing care: design and development using an adapted software development life cycle model. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e12551 [FREE Full text] [doi: 10.2196/12551] [Medline: 30973339]

XSL•FO
RenderX

18. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
19. Moore E, Tolley C, Bates D, Slight S. A systematic review of the impact of health information technology on nurses' time. J Am Med Inform Assoc 2020 May 01;27(5):798-807 [FREE Full text] [doi: 10.1093/jamia/ocz231] [Medline: 32159770]
20. Mickan S, Tilson JK, Atherton H, Roberts NW, Heneghan C. Evidence of effectiveness of health care professionals using handheld computers: a scoping review of systematic reviews. J Med Internet Res 2013 Oct 28;15(10):e212 [FREE Full text] [doi: 10.2196/jmir.2530] [Medline: 24165786]
21. Pérez-Martí M, Casadó-Marín LC, Guillén-Villar A. Electronic records with tablets at the point of care in an internal medicine unit. Online Journal of Nursing Informatics 2020;24(3):e30512 [FREE Full text]
22. Rydenfält C, Persson J, Erlingsdottir G, Johansson G. eHealth services in the near and distant future in Swedish home care nursing. Comput Inform Nurs 2019 Jul;37(7):366-372. [doi: 10.1097/CIN.0000000000000536] [Medline: 31135467]
23. Vossebeld DM, Puik ECN, Jaspers JEN, Schuurmans MJ. Development process of a mobile electronic medical record for nurses: a single case study. BMC Med Inform Decis Mak 2019 Jan 14;19(1):11 [FREE Full text] [doi: 10.1186/s12911-018-0726-3] [Medline: 30642324]
24. Forde-Johnston C, Butcher D, Aveyard H. An integrative review exploring the impact of electronic health records (EHR) on the quality of nurse-patient interactions and communication. J Adv Nurs 2023 Jan 07;79(1):48-67 [FREE Full text] [doi: 10.1111/jan.15484] [Medline: 36345050]
25. Ali S, Kleib M, Paul P, Petrovskaya O, Kennedy M. Compassionate nursing care and the use of digital health technologies: a scoping review. Int J Nurs Stud 2022 Mar;127:104161. [doi: 10.1016/j.ijnurstu.2021.104161] [Medline: 35032743]
26. Villalba-Mora E, Casas I, Lupiañez-Villanueva F, Maghiros I. Adoption of health information technologies by physicians for clinical practice: the Andalusian case. Int J Med Inform 2015 Jul;84(7):477-485. [doi: 10.1016/j.ijmedinf.2015.03.002] [Medline: 25823578]
27. Zhu H, Andersen ST. Digital competence in social work practice and education: experiences from Norway. Nordic Social Work Research 2021 Mar 17;12(5):823-838. [doi: 10.1080/2156857x.2021.1899967]
28. European Commission. Shaping Europe's digital future. European Commission. 2021. URL: https://ec.europa.eu/digital-single-market/en/content/european-digital-strategy[accessed [accessed 2023-01-12]
29. Dequanter S, Steenhout I, Fobelets M, Gagnon M, Sasseville M, Bourbonnais A, et al. Technology implementation in care practices for community-dwelling older adults with mild cognitive decline: Perspectives of professional caregivers in Quebec and Brussels. Digit Health 2022 Nov 16;8:20552076221139693 [FREE Full text] [doi: 10.1177/20552076221139693] [Medline: 36420317]
30. Fritzell J, Agahi N, Jylhä M, Rostgaard T. Social inequalities in ageing in the Nordic countries. Eur J Ageing 2022 Jun 11;19(2):155-159 [FREE Full text] [doi: 10.1007/s10433-022-00702-5] [Medline: 35578718]
31. Act on health care professionals 559/1994. Finlex. URL: https://www.finlex.fi/en/laki/kaannokset/1994/en19940559 [accessed 2023-01-21]
32. Degree on health care professionals 564/1994. Finlex. URL: https://www.finlex.fi/en/laki/kaannokset/1994/en19940564 [accessed 2023-01-20]
33. Vocational qualification in social and health care. Finnish National Agency for Education. 2022. URL: https://eperusteet.opintopolku.fi/#/en/ammatillinen/7854765/tiedot [accessed 2022-12-20]
34. The Finnish Union of Practical Nurses. Practical nurse training. Practical nurse's education, training and registration. The Finnish Union of Practical Nurses. 2022. URL: https://www.superliitto.fi/in-english/practical-nurse-training/ [accessed 2022-12-20]
35. Statistics Finland. Shop sales assistants, health care assistants and nurses most common occupations in 2020. Statistics Finland. URL: https://www.stat.fi/en/publication/cktws35s04dru0b553lzi7aci [accessed 2022-12-15]
36. Lee T, Sun G, Kou L, Yeh M. The use of information technology to enhance patient safety and nursing efficiency. THC 2017 Oct 23;25(5):917-928. [doi: 10.3233/thc-170848]
37. Keskimaki I, Tynkkynen L, Reissell E, Koivusalo M, Syrja V, Vuorenkoski L, et al. Finland: health system review. Health Syst Transit 2019 Aug;21(2):1-166. [Medline: 31596240]
38. Social Welfare Act 1301/2014. Finlex. URL: https://finlex.fi/fi/laki/ajantasa/2014/20141301#L3P21 [accessed 2023-01-20]
39. Vänskä J, Viitanen J, Hyppönen H, Elovainio M, Winblad I, Reponen J, et al. Lääkärien arviot potilastietojärjestelmistä kriittisiä. Suomen Lääkärilehti 2010;52(65):4177-4183.
40. Viitanen J, Hyppönen H, Lääveri T, Vänskä J, Reponen J, Winblad I. National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int J Med Inform 2011 Oct;80(10):708-725. [doi: 10.1016/j.ijmedinf.2011.06.010] [Medline: 21784701]
41. Hyppönen H, Lääveri T, Hahtela N, Suutarla A, Sillanpää K, Kinnunen U, et al. Kyvykkäille käyttäjille fiksut järjestelmät? Sairaanhoitajien arviot potilastietojärjestelmistä 2017. FinJeHeW 2018 Mar 08;10(1):30-59. [doi: 10.23996/fjhw.65363]
42. Saranto K, Kinnunen U, Koponen S, Kyytsönen M, Hyyppönen H, Vehko T. Sairaanhoitajien valmiudet tiedonhallintaan sekä kokemukset potilas- ja asiakastietojärjestelmien tuesta työtehtäviin. FinJeHeW 2020 Oct 13;12(3):212-228. [doi: 10.23996/fjhw.95711]

XSL•FO
RenderX

43. Saranto K, Koponen S, Vehko T, Kivekäs E. Nurse managers' opinions of information system support for performance management: a correlational study. Methods Inf Med 2023 Jun 09;62(S 01):e63-e72 [FREE Full text] [doi: 10.1055/a-1978-9727] [Medline: 36379471]
44. Ylönen K, Salovaara S, Kaipio J, Tyllinen M, Tynkkynen E, Hautala S, et al. Sosiaalialan asiakastietojärjestelmissä paljon parannettavaa: käyttäjäkokemukset 2019. FinJeHeW 2020 Mar 05;12(1):30-43. [doi: 10.23996/fjhw.88583]
45. Salovaara S, Ylönen K, Silén M, Viitanen J, Lääveri T, Hautala S. Sosiaalialan korkeakoulutettujen ammattilaisten arviot asiakastietojärjestelmistä 2020. FinJeHeW 2022 May 09;14(2):191-207. [doi: 10.23996/fjhw.113710]
46. European Union. General Data Protection Regulation 2016/679. European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679 [accessed 2023-05-03]
47. Sammut R, Griscti O, Norman IJ. Strategies to improve response rates to web surveys: a literature review. Int J Nurs Stud 2021 Nov;123:104058. [doi: 10.1016/j.ijnurstu.2021.104058] [Medline: 34454334]
48. Timmins F, Ottonello G, Napolitano F, Musio ME, Calzolari M, Gammone M, et al. The state of the science-the impact of declining response rates by nurses in nursing research projects. J Clin Nurs 2023 Apr 02;32(7-8):e9-e11. [doi: 10.1111/jocn.16597] [Medline: 36460485]
49. Tilastot. The Finnish Union of Practical Nurses. URL: https://www.superliitto.fi/super-info/superin-toiminta/tilastot/ [accessed 2023-09-29]
50. Mobile data usage got a boost when we were immobile. Tefficient. URL: https://tefficient.com/mobile-data-usage-got-a-boost-when-we-were-immobile/#more-5938 [accessed 2022-02-21]
51. Dexheimer JW, Borycki EM. Use of mobile devices in the emergency department: a scoping review. Health Informatics J 2015 Dec 29;21(4):306-315 [FREE Full text] [doi: 10.1177/1460458214530137] [Medline: 24782479]
52. Flynn GAH, Polivka B, Behr JH. Smartphone use by nurses in acute care settings. CIN: Computers, Informatics, Nursing 2018;36(3):120-126. [doi: 10.1097/cin.0000000000000400]
53. Zhang H, Cocosila M, Archer N. Factors of adoption of mobile information technology by homecare nurses: a technology acceptance model 2 approach. Comput Inform Nurs 2010;28(1):49-56. [doi: 10.1097/NCN.0b013e3181c0474a] [Medline: 19940621]
54. Hyppönen H, Kaipio J, Heponiemi T, Lääveri T, Aalto A, Vänskä J, et al. Developing the national usability-focused health information system scale for physicians: validation study. J Med Internet Res 2019 May 16;21(5):e12875 [FREE Full text] [doi: 10.2196/12875] [Medline: 31099336]

Abbreviations

CIS: client information system
EHR: electronic health record
ICT: information and communication technology
STePS: Monitoring and Assessment of Social Welfare and Health Care Information System Services

Edited by T de Azevedo Cardoso; submitted 03.03.23; peer-reviewed by S Nissinen, S Ashraf, C Wang, R Zhang ; comments to author 19.09.23; revised version received 06.10.23; accepted 23.04.24; published 29.05.24.

Please cite as:
Paatela S, Kyytsönen M, Saranto K, Kinnunen UM, Vehko T
Experiences of Electronic Health Records' and Client Information Systems' Use on a Mobile Device and Factors Associated With Work Time Savings Among Practical Nurses: Cross-Sectional Study
J Med Internet Res 2024;26:e46954
URL: https://www.jmir.org/2024/1/e46954
doi:10.2196/46954
PMID:38809583

©Satu Paatela, Maiju Kyytsönen, Kaija Saranto, Ulla-Mari Kinnunen, Tuulikki Vehko. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Paradoxes of Digital Tools in Hospitals: Qualitative Interview Study

Marie Wosny[1,2], BSc, MSc; Livia Maria Strasser[1], BMed; Janna Hastings[1,2], BSc, MSc, MA, PhD

[1]School of Medicine, University of St Gallen, St.Gallen, Switzerland

[2]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

Corresponding Author:
Marie Wosny, BSc, MSc
School of Medicine
University of St Gallen
St.Jakob-Strasse 21
St.Gallen, 9000
Switzerland
Phone: 41 712243249
Email: mariejohanna.wosny@unisg.ch

Abstract

Background: Digital tools are progressively reshaping the daily work of health care professionals (HCPs) in hospitals. While this transformation holds substantial promise, it leads to frustrating experiences, raising concerns about negative impacts on clinicians' well-being.

Objective: The goal of this study was to comprehensively explore the lived experiences of HCPs navigating digital tools throughout their daily routines.

Methods: Qualitative in-depth interviews with 52 HCPs representing 24 medical specialties across 14 hospitals in Switzerland were performed.

Results: Inductive thematic analysis revealed 4 main themes: digital tool use, workflow and processes, HCPs' experience of care delivery, and digital transformation and management of change. Within these themes, 6 intriguing paradoxes emerged, and we hypothesized that these paradoxes might partly explain the persistence of the challenges facing hospital digitalization: the promise of efficiency and the reality of inefficiency, the shift from face to face to interface, juggling frustration and dedication, the illusion of information access and trust, the complexity and intersection of workflows and care paths, and the opportunities and challenges of shadow IT.

Conclusions: Our study highlights the central importance of acknowledging and considering the experiences of HCPs to support the transformation of health care technology and to avoid or mitigate any potential negative experiences that might arise from digitalization. The viewpoints of HCPs add relevant insights into long-standing informatics problems in health care and may suggest new strategies to follow when tackling future challenges.

(J Med Internet Res 2024;26:e56095) doi:10.2196/56095

KEYWORDS

health care; health care technology; health care information technology; hospital information technology; clinical information systems; health care professionals; experience; frustration; clinician burnout; technology implementation; paradoxes; digital tool; digital tools; hospital; hospitals; qualitative interview study; interview; interviews; Switzerland; thematic analysis

Introduction

Background

The burden placed upon health care professionals (HCPs) in hospitals by the complexities of digital environments is a recognized and well-reported challenge of our time [1,2]. While the increasing use of health information technologies (HITs), such as electronic health records (EHRs), clinical information systems, and clinical decision support systems, can improve patient care [3,4], suboptimal design and ineffectual workflow integration can lead to frustration and reduced job satisfaction and may be associated with burnout among HCPs [1,5,6]. In tertiary care facilities, nearly 50% of physicians' daily schedules involve computer-based responsibilities, with EHR management

XSL•FO
RenderX

consuming about 40% of their time [7]. Moreover, EHR design is known to correlate with clinician burnout due to information overload and high-volume messages from inbox notifications [8]. Burnout is experienced by 48% of hospital-based physicians, partially attributed to growing computerization, as well as time constraints, chaotic work settings, and other stressors [9,10], while in severe instances, it may lead to depression, substance abuse, and even suicide [11]. Furthermore, inadequate IT may contribute to systemic inefficiencies, diminishing overall economic efficiency within hospitals [12-14]. This consequently impacts national health care systems, which are already facing shortages in health care personnel [15].

The challenges related to technology usability in hospitals have persisted over decades [16-18], despite the regularly repeated promises of its potential to alleviate burdens [19-21], streamline clinical processes, such as data collection and synthesis; simplify documentation tasks [22]; enhance patient care; and improve the overall efficiency of health care systems [23]. One possible factor driving the persistence of challenges is the pace of change of the digital health care landscape. The result of each innovation that aims to address an existing challenge might create another challenge; for example, increasing digitalization may lead to the navigation of extensive data sets, management of a diverse array of digital tools, and the need to address ethical and data privacy concerns [2,24]. Furthermore, the growing use of unauthorized technologies, commonly referred to as *shadow IT*, is becoming more prevalent in hospitals, further complicating the digital health care ecosystem [25].

In the past, various theories have been developed to explain the factors influencing the adoption of technological innovations by individuals, such as the diffusion of innovation theory [26]; the Technology Acceptance Model that focuses on individuals' decision-making processes regarding the adoption and use of technology [27]; and the Work System Model, which focuses on understanding and optimizing the interaction between people, processes, and technology within organizational work systems [28]. The application of these theories and models in health care technologies has been extensively studied, emphasizing the importance of considering organizational and professional culture, while also highlighting that human factors outweigh technological factors in health informatics [29,30].

In previous research, we conducted a systematic review of the published qualitative literature on HCPs' experiences using digital tools in hospitals, with the aim to collect evidence on their benefits and challenges, particularly focusing on their personal experiences [31]. In this context, our findings revealed that clinicians' personal experiences were reported less frequently compared with the moderators or outcomes associated with the use of HIT and that these can often not be distinguished. To delineate this distinction, we developed a theoretical framework that aimed to understand the complex interplay between the use of digital tools, encompassing positive and negative experiences, positive and negative moderators that possibly affect their adoption and use, and the resulting corresponding positive and negative effects and outcomes [31]. While outcomes of digital tools use in health care can encompass improved patient care, enhanced workflow efficiency, and increased information availability, moderators (eg, training quality and interface design) do exist, which can exert either positive or negative influences on tool use. Furthermore, individual user experiences, comprising personal thoughts, emotions, and feelings, can be sharpened by both tool outcomes and use, consequently either fostering or impeding the adoption of digital tools. These direct emotional experiences when using technologies may account for the downstream impact on well-being [31].

Objective

Moreover, we discovered that the perspectives of HCPs themselves are often not sufficiently considered when designing technological solutions, despite evidence underscoring their importance in successful adoption [31-34]. To address this gap, this study aimed to gain new insights into the long-standing challenge of HIT from the perspective of HCPs through in-depth interviews by addressing the central research question "What is the lived experience of HCPs when using digital tools in hospital settings?" Switzerland offers a valuable research environment for a comprehensive exploration of HCPs' perspectives due to its diverse health care landscape [35]. Understanding and amplifying HCPs' experiences is vital to generate new insights into the mechanisms and tensions underscoring the ongoing challenges; explore current dilemmas and opportunities; and suggest alternative strategies for the future development, adoption, implementation, and evaluation of these tools.

Methods

Study Design and Setting

This qualitative descriptive interview study explored the firsthand experiences of HCPs using digital tools within hospital settings. Semistructured interviews were conducted to gain a deep understanding of HCPs' experiences. Unlike other research methods, qualitative interviews allow us to capture intricate aspects of individuals' experiences, including their thoughts, emotions, opinions, and perspectives [36-38]. Switzerland was chosen as the study location partly due to the limited exploration of its diverse health care digitalization landscape [31]. Moreover, Switzerland's health care system is decentralized; therefore, there are a variety of solutions implemented at different hospitals, offering a rich source of diverse experiences. The reporting of this study adheres to the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Multimedia Appendix 1) [39-41].

Ethical Considerations

This study obtained ethical research approval from the academic institution and was deemed out of scope for national ethics approval. In this study, all participants provided verbal consent directly before the interview was conducted and audio recorded. Participants were informed that their participation was voluntary and that they could withdraw from the study at any time.

Participant Recruitment and Selection Process

To participate, HCPs needed to be employed in a Swiss hospital, with at least 6 months of digital tool experience and willingness to participate in a 30- to 60-minute interview in the German or

XSL•FO
RenderX

English language. Students and non–hospital-based HCPs were excluded. For recruitment, convenience sampling was performed via web-based platforms, social media, clinic newsletters, hospital intranet, and printed flyers. Moreover, the snowballing technique was used to further enhance diversity as well as purposeful sampling to cover various medical disciplines. The study's purpose was communicated through the institute's website and sign-up form. Six participants withdrew due to time constraints, suitability reevaluation, or unavailability.

Data Collection Process

The authors conducted the interviews in person at hospitals, cafeterias, cafes, and the university or via video calls based on the participants' availability. The interviews took place from May to August 2023 and were conducted until data saturation was reached, that is, when no new concepts emerged [42]. A semistructured interview guide was developed with the objective of understanding the HCPs' experiences with digital tools. The questions were informed by the Technology Acceptance Model, which was used to understand and explain how HCPs perceive and adopt new technologies; the Work System Model, in order to explore how people, processes, and technology interact within a working system and to identify the different components of the health care work system; and a theoretical framework derived from our own previous systematic literature review, which explains the interplay between digital tool use, experience, moderators, and identified outcomes [31] (Multimedia Appendix 2) [27,28]. Interviews were conducted in the preferred languages of the HCPs—in Swiss German, German, or English. The interviews began with personal introductions and an outline of the study's structure. Sociodemographic questions were followed by queries about digital tool experiences, including type, integration, usefulness, trustworthiness, professional image impact, collaboration, patient interactions, and reflections. In the end, participants were invited to ask questions or add thoughts on their digital tool experience that were not covered. Interviews were audio recorded and transcribed verbatim using the artificial intelligence (AI)–based software *Spoke* (Spoke Software, Inc) [43]. Transcripts were reviewed for accuracy, deidentified, and assigned unique study IDs.. Participants reviewed the transcripts, and revisions were made accordingly based on their provided feedback. In cases of no feedback within a month, the transcript was considered accepted.

Data Analysis

Data collection and analysis proceeded concurrently by using ATLAS.ti software (ATLAS.ti Scientific Software Development GmbH) [40]. Deductive thematic analysis was used for the creation of a preliminary codebook based on the theoretical frameworks guiding the interview questions, followed by an inductive analysis to describe and synthesize HCPs' experience of digital tools in hospitals [27,28,31,44,45]. Two authors independently familiarized themselves with the data by rereading the interview transcripts and applied codes to the transcripts in their original language. Through iterative discussions, codes were expanded, harmonized, and merged into a single codebook. Notes were taken to capture data items, patterns, and connections, forming preliminary themes, patterned responses, or meanings within the data [41]. The third author reviewed the transcripts to ensure comprehensive data coverage. As the next step, all 3 authors collectively examined initial codes and data extracts, inductively grouping them into themes specific to HCPs' digital tool experiences in hospital settings. Themes were developed through analysis, combination, comparison, and visual mapping of code interrelations. Further examination led to the restructuring of themes and coded data into aggregated dimensions, ensuring a meaningful fit of codes and investigation of their alignment with the entire data set. Cross-connections among dimensions, themes, and subthemes were established. Narrative descriptions and definitions were created within the context of research inquiries into HCPs' digital tool experiences, with the identification of emergent subthemes. Finally, the discussion of findings and analysis of codes on the meta-level culminated in the identification and definition of paradoxical experiences and perceptions among HCPs regarding digital tools in the hospital, depicted through opposing codes. Drawing upon Parse's notion of paradox, a living paradox is delineated as a rhythmic oscillation of perspectives, with awareness arising from encountering the contradiction of opposites in daily navigation of value priorities while progressing toward the yet-to-be-realized [46]. To facilitate reporting, all quotes were translated into English. Finally, dimensions, themes, and subthemes were discussed with study participants to validate and ensure trustworthiness.

Results

Characteristics of Study Participants

A total of 52 interviews with an average duration of 27.38 (SD 9.1) minutes were conducted with HCPs representing 24 different medical specialties across 14 hospitals in Switzerland (Table 1). Sex distribution among the participants was balanced, with an average age of 40.31 (SD 10.18) years. Physicians comprised 71% (n=37) of the sample, with 42% (n=22) of the participants in senior roles and 29% (n=15) resident physicians. Registered nurses accounted for 29% (n=15) of the sample, with 19% (n=10) of the participants in senior positions and 10% (n=5) regular staff nurses. More than half of the HCPs (n=32, 62%) reported over a decade of experience with digital tools, typically acquired since the onset of their careers (Table 2).

XSL•FO
RenderX

Table 1. Distribution of study participants by medical disciplines (n=52).

Medical discipline	Participants, n (%)
Internal medicine	8 (15)
Neurosurgery	5 (10)
Emergency medicine	4 (8)
Intensive care	4 (8)
Oncology and hematology	4 (8)
Cardiology	3 (6)
Surgery	3 (6)
Transdisciplinary	3 (6)
Gastroenterology	2 (4)
Pneumology	2 (4)
Endocrinology and diabetology	1 (2)
Geriatrics	1 (2)
Gynecology	1 (2)
Hepatology	1 (2)
Infectious diseases	1 (2)
Nephrology	1 (2)
Ophthalmology	1 (2)
Orthopedics	1 (2)
Otorhinolaryngology	1 (2)
Pathology	1 (2)
Psychiatry	1 (2)
Radiation oncology	1 (2)
Radiology	1 (2)
Urology	1 (2)

XSL•FO
RenderX

Table 2. Sociodemographic and professional characteristics of study participants (n=52).

Participant characteristics	Participants
Hospital size, n (%)	
Large (>700 beds, >7000 staff)	11 (21)
Medium (500-700 beds, 3000-7000 staff)	32 (62)
Small (<500 beds, <3000 staff)	9 (17)
Role, n (%)	
Senior physician	22 (42)
Resident physician	15 (29)
Senior registered nurse	10 (19)
Staff registered nurse	5 (10)
Sex, n (%)	
Male	27 (52)
Female	25 (48)
Age (y), mean (SD)	40.31 (10.18)
20-29, n (%)	10 (19)
30-39, n (%)	17 (33)
40-49, n (%)	14 (27)
50-59, n (%)	9 (17)
>60, n (%)	2 (4)
Experience with digital tools, n (%)	
6 months to 1 year	3 (6)
>1 year to 5 years	10 (19)
>5 years to 10 years	7 (13)
>10 years	32 (62)
Employment status (%), n (%)	
100	39 (75)
90	3 (6)
80	7 (13)
70	2 (4)
60	1 (2)

HCPs' Experiences With Digital Tools in Hospitals

We identified 5962 quotations with 654 unique primary codes. Overall, negative sentiments about the use of digital tools have been mentioned more often by HCPs (n=3018, 50.62% quotations) compared with positive (n=1636, 27.44% quotations) or neutral (n=1308, 21.94% quotations) reports. Compared with our previous literature review and derived theoretical framework [31], which included 6 dimensions impacting digital tool use, including moderators, outcomes, and experiences, each of a positive and negative nature, this study reported a similar distribution of digital tool moderators, outcomes, and experiences (n=2976). Investigating these 6 dimensions, overall, 61.32% (1825/2976) of reported factors are moderators that influence the use of digital tools in positive (883/2976, 29.67%) or negative (942/2976, 31.65%) ways; 31.65% (942/2976) are outcomes, either positive (438/2976, 14.72%) or negative (504/2976, 16.93%); and only 7.02% (209/2976) true experiences and emotions, with only 1.88% (56/2976) positive and 5.14% (153/2976) negative reports that impact the overall well-being of clinicians.

Four aggregated themes emerged from the thematic analysis of the interview data, namely, (1) digital tool use, (2) workflow and processes, (3) HCP's experience of care delivery, and (4) digital transformation and management of change. Each theme in turn consisted of several subthemes (Figure 1).

XSL•FO
RenderX

Figure 1. Thematic map analysis of themes and subthemes from qualitative interviews with health care professionals.

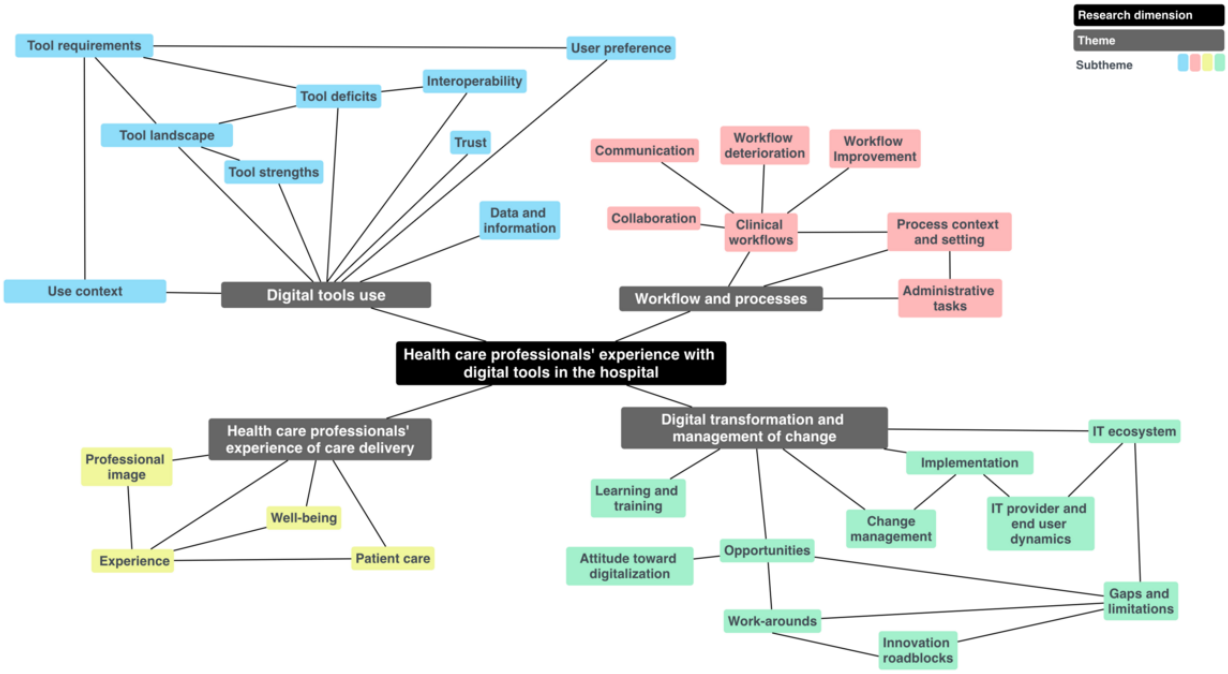

Theme 1: Digital Tool Use

HCPs (n=52) reported a comprehensive spectrum of >100 hardware and software tools, designed to serve both medical and administrative needs (Table 3). Reported tool strengths included remote access (n=15, 29%), system interoperability (n=14, 27%), and automation (n=11, 21%; Table 4 and Textbox 1). Reported tool challenges included interoperability issues (n=21, 40%) and technical malfunctions (n=21, 40%), compounded by counterintuitive (n=18, 35%) and user-unfriendly tool designs (n=17, 33%), as well as the existence of multiple individual tools (n=17, 33%) that sometimes ran in parallel and served identical purposes (Table S1 in Multimedia Appendix 3). These challenges lead to inefficiencies, tool avoidance, or the use of unauthorized shadow IT resources (n=12, 23%) such as private hardware, messaging apps, and social media platforms, which were often used for swift information sharing (Table S2 in Multimedia Appendix 3).

Participants (n=52) stated their preference for information systems that are both comprehensive and seamlessly integrated (n=18, 35%) and emphasized the importance of customization and modularity, aiming for a system that could be adapted to their specific needs, use contexts, and disciplines (n=6, 12%). For example, intensive care units and emergency medicine departments were reported to have distinct requirements, necessitating specialized tools that may not seamlessly integrate with the hospital's broader systems (n=11, 21%). HCPs reported that they often found themselves in a position where tools did not align with their specific needs. However, they reported that they often had to adapt without the option to reject tools (n=20, 38%) or choose alternative solutions (n=11, 21%) because decision-making is orchestrated by management (n=10, 19%) with limited consideration for end users' specific needs and preferences (n=12, 23%).

In the broader context, trust was central to digital tool adoption, and we observed an age-related difference (P=.02), with higher trust levels reported by HCPs aged 20 to 49 years compared to those aged ≥50 years. Nonetheless, most participants (n=36, 69%) expressed full trust in the tools, primarily due to oversight and safety confirmation offered by the local IT department (n=10, 19%).

XSL•FO
RenderX

Table 3. Most common reported digital tools in hospital settings (n=52).

Digital tools	Reports, n (%)
Medical tools	
CIS[a]	52 (100)
Dictation and speech recognition software	22 (42)
Laboratory information systems	20 (38)
Knowledge database	16 (31)
Picture archiving and communication system	16 (31)
Literature database	8 (15)
Remote patient monitoring tool	8 (15)
Surgery planning tool	8 (15)
Digital imaging and communication system	7 (13)
Surgical navigation system	6 (12)
Administrative tools	
Email program	33 (63)
MS Office 365 Suite	23 (44)
Videoconferencing tools	16 (31)
Shift scheduling software	12 (23)
Calendar tool	9 (17)
Performance tracking tool	9 (17)
HR[b] platform	6 (12)
Administration tool	4 (8)
Expense management tool	4 (8)
Hospital bed management system	3 (6)

[a]CIS: clinical information system.
[b]HR: human resource.

XSL•FO
RenderX

Table 4. Selection of commonly reported subthemes of the 4 identified main themes (n=52).

Themes and subthemes	Participants, n (%)
Digital tool strengths	
Remote access	15 (29)
System interoperability	14 (27)
Automation function	11 (21)
Provision of comprehensive overview	8 (15)
Intuitive tool design	7 (13)
Manual corrections possible	7 (13)
Automated data transfer	6 (12)
Digital tool deficits	
Interoperability issues	21 (40)
Technical malfunctions	21 (40)
Counterintuitive tool design	18 (35)
User-unfriendly tool design	17 (33)
Incompatible systems	17 (33)
Various specific and individual tools	17 (33)
Cumbersome tool visualization	13 (25)
Workflow improvement	
Gain in efficiency	37 (71)
Time-saving	34 (65)
Increased productivity	8 (15)
Better time resource management	6 (12)
Decreased workload	6 (12)
Mental support and assurance	4 (8)
Workflow deterioration	
Time consumption	34 (65)
Inefficiencies	22 (42)
More effort or cost than benefit	11 (21)
More distraction and less focus	9 (17)
No measurable outcome	9 (17)
Overflow of information	8 (15)
Work hindered	7 (13)
Positive experience	
Satisfied	14 (27)
Excited	8 (15)
Grateful	7 (13)
Relieved	7 (13)
Confident	5 (10)
Hope	5 (10)
Reassured	5 (10)
Negative experience	
Frustrated	46 (88)
Bothered and annoyed	21 (40)

XSL•FO
RenderX

Themes and subthemes	Participants, n (%)
Angry	12 (23)
Dissatisfied	12 (23)
Stressed	9 (17)
Helpless	8 (15)
Tedious	8 (15)

Textbox 1. Selection of supporting quotes on reported digital tool attributes, workflow impact, and user experience.

Digital tool strengths

- "I'll begin at home, where I have remote access. There, I can use the same interface as the one available at the hospital, and I find that very beneficial." [Senior physician]
- "One of the aspects I truly appreciate about the clinical information system is that everything functions within a single program. You do not need to open any external applications." [Resident physician]
- "What I found particularly advantageous was the seamless transfer of data from the surveillance monitor. It's reassuring that you don't have to painstakingly click through; it all happens automatically. That's a huge relief." [Staff nurse]

Digital tool deficits

- "I find it quite cumbersome because there are three different programs that don't communicate with each other." [Resident physician]
- "Moreover, the need to switch between these programs is inconvenient. They don't integrate smoothly, and we're certain that the key connection points between these systems are not 100% compatible. This often leads to issues." [Staff nurse]
- "When a problem arises the initial answer is often, 'We make a tool for that.' Subsequently, someone, usually at a significant cost, begins developing a tool. However, it's often rare to see truly effective, streamlined, and good results." [Senior physician]

Impact of digital tools on workflow improvement

- "So, the patient curve generated electronically is identical to the one produced manually. That's excellent." [Resident physician]
- "It's also a significant advantage that you can securely access information from different locations. For instance, in terms of planning, I can create the daily schedule from my office, while someone at the station office can simultaneously update that a patient is being discharged. I can instantly see these updates even when I'm not physically on-site." [Senior nurse]
- "We have transitioned into a nearly paperless clinic. Our entire workflow has been digitized with the assistance of Care Paths defined within our IT infrastructure. Every aspect of our operations has been successfully integrated." [Resident physician]
- "Working with this daily program can be quite frustrating due to its inherent inefficiency." [Resident physician]
- "And then the phone rings, pulling you out of your workflow once again. You're compelled to perform specific tasks on your phone, and only after completing those tasks can you return to your initial workflow. I believe this is a significant source of errors." [Staff nurse]
- "However, I strongly believe that one's efficiency increases when interruptions are minimized. [...] I've noticed that about half of the team, [...], tends to look at their phones or engage in other activities because they have received a message. This affects our interpersonal interactions, which are already rare. [...] I've observed that questions are often repeated, and important points get lost because some colleagues are distracted by notifications, messages, or emails [...]. Consequently, our morning meetings can easily become extended due to these digital distractions." [Senior physician]

Positive experiences of digital tool use

- "It's quite satisfying to swiftly access the data that's relevant to you, enabling you to perform assessments and make necessary adjustments to therapies or conduct consultations." [Resident physician]
- "I was able to schedule the session on my laptop, and it was incredibly satisfying. While it didn't reduce stress, it certainly provided a sense of relief." [Senior nurse]
- "The emotions involved were a mix of relief, in terms of time-saving, especially at midnight. Additionally, there was a sense of pride because this system is our creation, stemming from our team's effort." [Senior physician]

Negative experiences of digital tools use

- "You get completely blocked. I couldn't access the clinical information system and I couldn't work. It's extremely frustrating." [Staff nurse]
- "And then we have to cut back on patients. Or the resident physicians end up working overtime. When they have to deal with those interface management issues, they can't be with the patients or leave early. That's frustrating." [Senior physician]
- "Undoubtedly, this places a heavy psychological burden on us because we're aware that when the system isn't functioning, we can't take care of the patients who are in unstable conditions as they arrive. We must wait until the system is functioning again." [Staff nurse]

XSL•FO
RenderX

Theme 2: Workflow and Processes

While it was highlighted by the HCPs (n=52) that digital tools were promoting collaboration (n=24, 46%) and enhancing overall communication efficiency (n=7, 13%), the need for in-person interactions remained (n=18, 35%). However, it was reported that face-to-face interactions decreased (n=17, 33%) due to the use of digital tools and hospital logistics (n=5, 10%). Primary communication modes remained phone calls and emails, with some physicians having faced severe email overload (n=9, 17%) with >100 emails daily, while phone calls interrupted nurses' workflows (n=3, 6%).

Components associated with workflow improvement attributed to digital tools were mainly seamless integration (n=22, 42%), supportive tool features (n=17, 33%), and correctly mapped care paths (n=10, 19%). These factors were described as enhanced operational efficiency (n=37, 71%), time management (n=34, 65%), and overall productivity (n=8, 15%; Table 4; Textbox 1). By contrast, HCPs disclosed that workflow challenges often stemmed from the coexistence of paper and electronic documents (n=23, 44%), leading to documentation errors, time inefficiency, and increased work effort (Table S3 in Multimedia Appendix 3).

In addition, HCPs (n=29, 56%) stressed that the time constraints and inefficiencies that they encountered were linked to the intricacies of administrative processes, along with bureaucratic impediments (n=11, 21%), which often were enforced through digital tools, underlined by the conflict and inherent complexity between balancing standardization and considering the individualization of medical procedures (n=17, 33%; Table S4 in Multimedia Appendix 3).

Theme 3: HCPs' Experience of Care Delivery

The integration of digital tools within the hospital ecosystem and HCPs' (n=52) resulting experiences were multifaceted but primarily negative, characterized by frustration reported by most participants (n=46, 88%), annoyance (n=21, 40%), anger (n=12, 23%), dissatisfaction (n=12, 23%), and stress (n=9, 17%; Table 4 and Textbox 1). Notably, these negative experiences affect clinicians' well-being, leading to work-life balance challenges (n=10, 19%), heightened stress levels (n=9, 17%), cognitive overload (n=7, 13%), and mental burden (n=5, 10%; Table S5 in Multimedia Appendix 3). By contrast, positive experiences associated with digital tools, although less frequently reported, include satisfaction (n=14, 27%), enthusiasm (n=8, 15%), gratitude (n=7, 13%), and relief (n=7, 13%), mainly arising from digital tools' assistance with daily tasks (Table S6 in Multimedia Appendix 3).

The integration of digital tools significantly impacted HCPs' professional image and self-perception, with half of the participants (n=28, 54%) experiencing shifts in their professional identity. This included positive aspects, such as feeling better informed (n=10, 19%) and more proficient (n=6, 12%). They often viewed themselves as forward-thinking, driven by a desire for enhanced digitalization in health care (n=19, 37%). Conversely, negative shifts in HCPs' professional image were tied to increased time on computer tasks and documentation burdens (n=10, 19%), causing dissatisfaction due to role deviations and increased mental strain (n=5, 10%). Others (n=24, 46%) saw no significant changes, as digital tools were ingrained in their routines, particularly for those who grew up with digital tools (n=17, 33%).

Regarding decision-making, HCPs emphasized their irreplaceable human roles (n=8, 15%) based on their medical expertise and responsibilities. However, they also recognized significant reliance on digital tools (n=17, 33%) and their providers (n=3, 6%), especially in critical situations or system failures. Some HCPs noted changes in job expectations (n=5, 10%) and skillset requirements (n=5, 10%), including the need for IT proficiency, data interpretation, and rapid typing skills (n=5, 10%).

The core motivator for HCPs using innovative digital tools was to improve patient care and well-being (n=16, 31%), with an emphasis on patient safety (n=14, 27%), which was realized since HCPs were feeling better informed (n=10, 19%) and prepared (n=4, 8%) due to access to information. HCPs noted that digital tools had increased the quality of care (n=11, 21%), and mobile workstations facilitated improved patient-centered interactions (n=15, 29%). However, there were also reservations expressed about the potential transformation toward screen-centered care, which could lead to patients feeling disconnected and neglected (n=13, 25%).

HCPs anticipated the emergence of collaborative health care digital tools (n=6, 12%), emphasizing the importance of granting patients, their families, and caregivers access to health data for empowerment and shared decision-making (n=12, 23%), while also improving family and caregiver engagement through digital educational resources and visualization tools (n=2, 4%). Nevertheless, the abundant availability of medical information also raised concerns about misinformation and limited patient comprehension (n=2, 4%), particularly among patients with neurotic tendencies (n=3, 6%).

Theme 4: Digital Transformation and Management of Change

The HCPs (n=52) stated that their digital proficiency relied significantly on effective training (n=21, 40%). Although initial in-person training sessions at job commencement were standard, they were often perceived as insufficient and generic (n=12, 23%; Table S7 in Multimedia Appendix 3). Specialized training (n=7, 13%) and learning from experienced colleagues in professional settings or informal interactions (n=27, 52%) were stated as more effective in addition to experiential learning (n=33, 63%) and trial-and-error approaches (n=13, 25%). Overall, it was noted that proficiency improved over time and practice (n=21, 40%). However, despite recognizing training's importance, there was often a shortage of dedicated time (n=6, 12%) and a lack of prioritization (n=11, 21%), while controversially, an intriguing paradox emerged, with a subset of HCPs (n=4, 8%) reporting attendance in mandatory training that held no practical utility with their routines (Table S8 in Multimedia Appendix 3).

In the context of digital transformation, effective change management emerged as a pivotal factor for successful IT infrastructure implementation (n=9, 17%; Table S9 in

XSL•FO
RenderX

Multimedia Appendix 3). Furthermore, HCPs reported that they perceived smaller organizations to be more agile, demonstrating greater success in IT implementation than their larger counterparts (n=9, 17%). While HCPs appreciated the iterative refinement of digital systems (n=8, 15%) and emphasized the significance of pilot projects with feedback mechanisms (n=7, 13%), others in contrast reported unsuccessful tool introductions (n=8, 15%) leading to subsequent discontinuations due to technical issues, high effort intensity, or nontailored processes with no measurable benefits (Table S10 in Multimedia Appendix 3).

It was noted that the health care industry lagged far behind in digitalization efforts compared to other industries due to its complex nature (n=17, 33%). HCPs expressed underutilized technical possibilities (n=28, 54%) and gaps, with better hospital systems (n=19, 37%) and superior tools that existed but were inaccessible in their organization (n=6, 12%). In particular, challenges with eHealth applications and the efforts toward a national Electronic Patient Dossier in Switzerland were reported as problematic, often with a sense of pessimism (n=8, 15%).

Numerous innovation roadblocks that lead to these gaps were described, including monetary constraints (n=15, 29%), regulatory hurdles (n=14, 27%), and organizational structures (n=7, 13%). In response, HCPs found work-arounds (n=22, 42%) and shortcuts (n=17, 33%), and some HCPs reported developing customized IT solutions (n=8, 15%). Nearly half of the HCPs (n=22, 42%) highlighted ongoing changes and new clinical information system implementation at their institutions, with some HCPs (n=11, 21%) actively involved in implementation efforts.

Reported Paradoxes With Digital Tools in the Hospital

We identified 6 paradoxes associated with the experiences of HCPs with digital tools in the hospital (Table 5).

The first encompasses the "promise of efficiency and the reality of inefficiency," which describes the striking contrast between anticipated efficiency gains from health care digitalization and the real-world challenges and inefficiencies experienced during and after implementation. Numerous HCPs conveyed that the digital technologies they then used were time-consuming, inefficient, and burdened with cumbersome processes, while others underscored the advantages in efficiency and time-saving measures.

The second paradox is the "shift from face to face to interface," elaborating that health care team collaboration, spanning disciplines and extending externally, has transitioned from in person to digital interfaces, offering benefits and presenting challenges of maintaining a balance between both. HCPs noted that the transition toward screen-centric health care was reducing face-to-face interactions and had inadvertently hindered both internal and external collaboration along with regulatory constraints and systemic disparities. On the flip side, other HCPs have reported a prevailing reliance on in-person interactions as the default mode of operation. Moreover, it has been recognized that the integration of digital tools has enhanced communication efficiency.

The third paradox is defined as "juggling frustrations and dedication," as HCPs experience daily frustrations with digital tools, which prove to be inefficient with a lack of user-friendliness and cumbersome interfaces, ultimately impeding workflows and causing frustration and contributing to emotional exhaustion. Furthermore, there have been reports of varying job expectations among HCPs, with some noting substantial shifts while others' remained unchanged.

Another paradox is the "illusion of digital information access and trust," describing the disconnect between information availability and practical accessibility. A large number of HCPs (47/52, 90%) expressed concerns regarding the quality of data and information, highlighting inconsistencies stemming from the existence of multiple and differing information sources and human errors during data replication. Moreover, it has been reported that limited access to information, coupled with challenges in locating relevant information, exacerbates inefficiencies and increases the likelihood of errors in medical decision-making. In contrast, some HCPs recognized the advantages of enhanced access to information, facilitating informed decision-making. Furthermore, a subset of HCPs noted improvements in data quality, indicating a positive trend in this regard.

Moreover, a paradox at "the intersection of clinician workflows and patient care paths" emerged, which emphasizes the complexities that arise in health care workflows due to insufficient health care IT, disruptions, computer-centric tasks, and nonstandardized care paths. Most HCPs (45/52, 87%) voiced concerns that current health care IT systems disrupted rather than streamline processes, leading to a deterioration in workflow. This degradation is attributed to the proliferation of digital technologies, which introduce challenges such as email overload and distractions, as well as the need to manage multiple tools simultaneously, often necessitating frequent switching between them. Nevertheless, amid these challenges, workflow improvements have been observed, such as fast access to patient data and informed decision-making, efficient documentation and communication, and resource optimization.

Finally, the last paradox describes the "opportunities and challenges of shadow IT" and how HCPs use unauthorized software and hardware to tailor solutions and address workflow needs and gaps efficiently while being aware of the potential risks and data privacy concerns. Some HCPs (12/52, 23%) disclosed their use of unauthorized tools such as generative AI, chat applications, and social media platforms to expedite communication and access to information, despite the inherent risk of sharing unauthorized patient data. However, this practice was frequently conducted discreetly, and at the same time, HCPs highlighted their awareness of the potential risks involved and emphasized the importance of prioritizing data privacy concerns.

XSL•FO
RenderX

Table 5. Identified paradoxes with opposing reports and supporting quotes.

Paradox and opposing reports (quotes, n)	Exemplary quotes
The promise of efficiency and the reality of inefficiency	
• Time consumption (n=79) • Inefficiencies (n=52) • Cumbersome processes (n=48)	"You have to click on the same things day by day...And it consumes so much time and is inefficient. We could instead dedicate that time to the patients instead of sitting tired in front of the computer..." [Junior nurse]
• Gain in efficiency (n=122) • Time-saving (n=68)	"I am actually a big fan of technologies...because they primarily make our daily lives significantly easier, make everything more efficient, and also make it safer." [Assistant physician]
The shift from face-to face to interface	
• Shift to screen-centric care (n=21) • Reduced face-to-face interaction (n=32) • Internal collaboration hindered (n=15) • External collaboration hindered (n=15)	"I believe the direct interaction is gradually fading away. It feels like everyone is going their separate ways, immersed in their own tools, rather than functioning as a cohesive team." [Senior physician]
• In-person presence (n=22) • Internal collaboration promoted (n=33) • External collaboration promoted (n=18) • Improved communication (n=9)	"We definitely have less face-to-face involvement with the nursing team. However...this is not necessarily negative, as nurses value having slightly less direct interaction...Now, with nurses able to quickly access information in the clinical information system...there are fewer disruptions in the processes. This...is beneficial, although we may have slightly less interaction as a result." [Senior physician]
Juggling frustrations and dedication	
• Negative emotions (n=197) • Frustration (n=84)	"One feels incredibly hindered. I could not access my clinical information system, which prevents work. It is extremely frustrating." [Junior nurse]
• Positive emotions (n=71) • Satisfied (n=27)	"The ultimate aim is for the diagnosis to automatically appear digitally for physicians, ensuring it remains on their radar for proper treatment...The more insight I have into the required care, the more effectively I can practice medicine." [Senior physician]
The illusion of digital information access and trust	
• Poor data quality (n=25) • Different information sources (n=20) • Limited access to information (n=19) • Difficulties to find information (n=11)	"The quality of work has suffered due to this digitalization. Nowadays, diagnoses are simply copied from previous lists, and...are now six pages long because everything is copied and pasted." [Senior physician]
• More and better access to information (n=58) • Improved data quality (n=7)	"On one hand, digitalization has certainly brought us forward. We now have more concrete data, and the quality of that data has significantly improved." [Senior nurse]
The intersection of clinician workflows and patient care paths	
• Workflow deterioration (n=407)	"Among colleagues, the tool is seen as a disruptor, rather than something helpful." [Senior physician]
• Workflow improvements (n=277)	"We have transitioned into a nearly paperless clinic. Our entire workflow has been digitized with the assistance of Care Paths defined within our IT infrastructure." [Assistant physician]
The opportunities and challenges of shadow IT	
• Use of shadow IT (n=29)	"Recently, I tried to use ChatGPT to write reports. I was really very satisfied because I can gain a lot of time even though it is not perfect." [Assistant physician]
• Data privacy importance (n=17) • Data privacy concerns (n=8)	"Of course, security always entails some risk...I'm not really familiar with matters like data protection and such...But of course, it certainly also carries a risk." [Senior physician]

Discussion

Principal Findings

Despite the long-standing recognition of the potential of HIT, the realization of these promises seems perennially distant. Our study aimed to explore HCPs' perspectives on the ongoing digital transformation in hospitals and unveiled numerous negative experiences and frustrations. Paradoxes emerged from HCPs' perspectives, often highlighting opposing forces within reported experiences. Therefore, our findings underscore the substantial challenges posed by digital tools in hospitals and the impact of these challenges on HCPs' well-being. The development and implementation of effective hospital IT systems and digital tools face a narrow and challenging path to navigate these paradoxes, which are represented by tensions that often push in opposite directions. Given this context and

XSL•FO
RenderX

drawing from our findings, we discuss these emerging paradoxes as well as potential strategies that might help to navigate them.

The Promise of Efficiency and the Reality of Inefficiency

One paradox that was reported is the apparent promise of technological innovation compared to the actual experience of HCPs. The desire for efficiency gains through streamlined processes and improved patient care has long driven the adoption of digital tools in hospitals, with the optimistic outlook that technology may revolutionize health care delivery and administrative workflows [47,48]. However, this study revealed significant disparities between the ideation (the imagined promise) and the actual results of using digital tools in real-world hospital settings. Motivated by the aim to align discipline-specific needs with standardized patient care, standardizing medicine presents a complex paradox in itself and creates an inherent tension between the need for unification and customization of clinical information systems [49]. The aim for hospital information systems to establish seamless and fully integrated systems compatible with diverse tools, workflows, data models, and legacy system integrations represents a serious challenge [50,51]. This is further compounded by the interdepartmental and intrahospital variations, stemming from discipline-specific and regional needs. In addition, hospitals must maintain operations during IT system transitions, which necessitates thorough preparation, weighing the benefits and drawbacks of a disruptive system replacement versus a phased approach [2]. Another essential factor for effective use and integration of technologies into clinical decision-making is adequate training [52], which is currently often neglected. Moreover, while promising claims, for example, about the accuracy of medical AI as a digitalization initiative suggest that it can boost efficiency, this overlooks crucial factors such as the narrow definition of workload and the significant role of human factors in technology implementation, where human knowledge, competencies, and trust are pivotal in determining efficiency outcomes [53]. Given that many of the inefficiencies and frustrations reported in our study were due to system malfunctions or access challenges, one strategy to mitigate the possibility of inefficiencies introduced by digital tools might be for hospital IT systems to offer multiple access points and the possibility of diverse interfaces around common data stores [54]. Modularity and redundancy would allow backups for information capture during tool unavailability or system downtime, as well as tailored interfaces for different stakeholders and workflows [55]. This would also enable a staged rollout of novel innovations, minimize the risk of widespread system failure, and align with the dynamic nature and introduction of health care technology innovations [56].

Another key aspect that was reported is the emerging shifts in HCPs' identity, professional image, and self-perception due to digital technology integration. While some experience significant shifts, others, especially those accustomed to digital tools, see minimal to no changes. Nevertheless, providing tailored and effective training of these technologies and integrating these trainings into HCPs' schedules or freeing up time are essential for harnessing the full potential of health care technology [57]. In this regard, effective training sessions could encompass immersive digital simulations, real-time feedback, personalized learning, and adaptive modules [58-60]. Moreover, knowledge-sharing mechanisms and a culture of continuous learning are crucial for maximizing digital system use [61] as well as interprofessional training sessions between disciplines [62].

The Shift From Face to Face to Interface

Another paradox emerged in this study is the shift from traditional face-to-face interactions to digital interfaces, encompassing discipline-specific, interdisciplinary, multiprofessional, and external collaborations [63]. Multiprofessional collaborations in health care have been shown to improve patient outcomes while also benefiting HCPs by reducing redundant work [64]. However, this transition has not been without challenges, and there is a need for balancing in-person and digital collaboration, as HCPs noted distractions and a reduction in the ability to focus on human interactions during consultations since they increasingly rely on digital counterparts. Nevertheless, digital tools and increased data accessibility can promote all forms of collaboration through information sharing, remote access, and facilitation of swift consultations with colleagues. However, regulatory constraints based on regional mandates and the heterogeneous nature of the digital hospital systems currently limit external collaboration and set a major hurdle for a unified EHR approach. Achieving this requires harmonizing regulations and developing interoperable systems that enable the exchange of health information while upholding stringent data privacy standards and safeguards, especially in sensitive areas such as mental illness, substance use disorders, and sexual health, and ultimately ensuring adequate transparency while maintaining patient confidentiality [65,66]. To alleviate these challenges involved in the transition from face to face to interface and reduce digital overload, the introduction of communication guidelines and email etiquette is crucial for reducing stress, enhancing productivity, and fostering concentrated work phases. AI-supported systems can help to streamline these processes while enforcing a "no-devices" policy during time-boxed discussions, restricting email access, and optimizing inbox management, which can further increase engagement and minimize interruptions [67].

Juggling Frustrations and Dedication

The multifaceted frustrations experienced by HCPs in their daily roles negatively impact their well-being. Frustrations caused by technology and the resulting inefficiencies in work processes are impeding HCPs in clinical and administrative tasks. These frustrations paired with a constant sense of availability are disrupting work-life balance and are linked to emotional exhaustion, which can lead to burnout [14]. In accordance with previous research, it was found that increased time spent with computer tasks due to administrative duties and bureaucratic procedures adds complexity to health care work, affecting patient care and HCPs' well-being as well as career satisfaction [68]. One of the most striking findings of this study is the substantial frustration and stress that can emerge when digital tools do not function as expected, leading to delays in HCPs'

XSL•FO
RenderX

workflows and patient care. Concerns adding to this include not only tool reliability, security, and overall data privacy concerns but also missed opportunities, which all can lead to distrust or tool rejection. Moreover, the reported gap in access to necessary health care technologies and lack of decision rights further exacerbate HCPs' frustrations. Comparing these findings with those of other studies confirms that the cumulative effect of current IT challenges coupled with staffing shortages can result in extended working hours, overwork, and frustrations and highlights the urgent need for solutions [69]. Our study emphasized the paradox and the need to reduce HCPs' digital frustration while recognizing their job dedication. Involving HCPs in the choice, design, and implementation trade-offs in digital tool implementation is essential to empowering them as stakeholders and reducing associated negative emotions [70]. Moreover, harnessing the potential of innovations such as voice-based data entry and virtual smart assistants, as well modern language model technologies, can streamline administrative tasks, enhance data quality, and reduce costs [71].

The Illusion of Digital Information Access and Trust

Another paradox is the enormous and increasing amount of available data and knowledge, which carries positive and negative implications for HCPs' workflows related to information access, quality, and trust. Trust emerged as a central factor in the adoption of digital tools, with younger HCPs exhibiting higher levels of trust in digital tools compared to their counterparts aged ≥50 years. Research has consistently highlighted the significance of trust as a crucial component for the successful use and implementation of health care technologies. To establish trust, it is essential to align AI with existing values and consider social interactions and negotiations of values comprehensively [72].

One of the foremost obstacles reported is the illusion and disparity between information availability and actual accessibility, exacerbated by issues ranging from user account restrictions and unstable IT infrastructure to information overload and obscurity. The most critical issue reported is the duality of data quality. While robust data sets offer valuable insights and this is often a driver for adopting a digital strategy, HCPs frequently encountered data inconsistencies in patient reports due to various factors, including software limitations and human-generated copy-paste errors, which were also reported in other study settings [73,74]. This is especially critical for systems in which compromises in data integrity can lead to erroneous decisions affecting patient outcomes [75]. Hence, HCPs emphasized their role in making the ultimate medical decisions, underlining the need for expert knowledge to address data quality challenges. Moreover, limiting the copy-paste function in EHRs has been proposed to enhance report quality [76]; however, this would risk exacerbating the time spent in reporting. Participants in this study felt satisfaction when they had access to the right information at the right time, but in practice, what this meant differed from clinic to clinic and from HCP to HCP. To promote the desired flexible and individually customizable information systems to optimally support workflows within the realities of resource constraints, one possibility is to use the growing family of open-source infrastructures for dynamic, openly integrative solutions, which not only foster adaptability but also embrace collaboration [77].

The Intersection of Clinician Workflows and Patient Care Paths

Moreover, our study highlighted the complexities and intersection of HCPs' workflows and patient care paths, with most participants agreeing that they are increasingly exposed to a multitude of workflow interruptions and demands of administrative computer-centric work. Incessant email distractions have become a significant burden which, coupled with automated reminders and EHR alerts, create a constant challenge in prioritizing and managing correspondence. It has been indicated that information overload is raising physicians' cognitive load and is negatively affecting patient safety [78], while the increase in computer-centric work, with some HCPs spending up to 80% of their workday in front of screens, is negatively impacting patient care, in-person interactions, and their professional identities [79]. Therefore, clear communication guidelines and policies fostering minimal disruption during time-boxed discussions compounded by initiatives such as restricting email access, optimizing inbox management, and adhering to an appropriate email etiquette can significantly boost productivity while diminishing stress levels [67]. Moreover, the absence of standardized patient care paths complicates digital tool design, as health care technology often disrupts, rather than streamlines, workflows. In addition, numerous HCPs work in a hybrid manner, using both digital tools and paper-based processes in parallel, leading to significant inconsistencies [80]. In addition, technical and uncommunicated IT maintenance; system reboots, freezes, and crashes; and full blackouts disrupt clinical workflows, as these result in significant downtime and sometimes a return to paper-based processes.

The Opportunities and Challenges of Shadow IT

In addition, our study revealed that HCPs have many valuable insights in the digital domain, which they implement mainly through shadow IT. Due to limitations in existing digital tools and gaps, HCPs seek and develop their own solutions to fulfill their workflow requirements. Within this context, HCPs increasingly use unauthorized hardware and software to streamline processes, such as document imaging, scanning, and quick reference retrieval, as well as to promote more efficient collaboration, patient communication, and access to medical knowledge. This adaptive approach showcases the agility of HCPs in leveraging available resources; however, the associated security risks should not be disregarded. Other studies have revealed that health care institutions are grappling with the rise of shadow IT to complement shortcomings in hospital-provided IT resources, which can be beneficial, but also introduces vulnerabilities and potential access points for cyber threats [25]. Generative AI models and social media, along with chat apps, are now vital for information access, while also being used for exchanging sensitive patient data. The exponential increase in chat app use has improved communication but often occurs without HCPs' awareness of existing privacy and security regulations [81]. Although shadow IT is recognized, HCPs rather discreetly use it, showcasing their dedication to improving

XSL•FO
RenderX

efficiency. This innovative yet potentially risky approach underscores the imperative of addressing technological gaps and advancing hospital technology, signaling the need for change. Recognizing where workflow optimizations is reached through unauthorized tools can be of great interest for hospitals as well as HIT developers to inform workflow-aligned innovations for the central IT provision, as well as to identify vulnerabilities and access points for cyberthreats [25].

Limitations

This interview study has not only notable strengths but also limitations. It addresses gaps in the existing evidence by examining HCPs' experiences with digital tools in hospital settings. Extensive interviews with a diverse set of HCPs were conducted, offering a comprehensive overview of the digital tool landscape, rather than focusing solely on a specific type of tool or a specific workflow context. The rigorous approach ensured credibility through participant engagement, researcher reflexivity, and transparent data collection and analysis. Moreover, dependability was maintained by applying a literature-derived framework for data comprehension. However, convenience sampling might have introduced a self-selection bias, potentially impacting participants' attitudes and expectations. In addition, as the data are solely derived from interviews, these might lack structured observations of HCPs' interactions with digital tools. Besides this, the focus on hospital settings may limit generalizability to the primary care context, and perspectives of other stakeholders such as patients were only indirectly conveyed through HCPs. Finally, the study's location in Switzerland may constrain its applicability to other health care systems. Subsequent research could explore more discipline- or tool-specific insights as well as other contexts and health care environments.

Conclusions

Our study was designed to assess the experience of HCPs working with digital tools in hospitals. Despite the acknowledged benefits and possibilities offered by digital technologies in hospitals, these tools place a daily burden on HCPs that must be addressed. Within the ongoing digital transformation of almost every health care institution, significant paradoxes and conflicting perspectives regarding the optimal strategy for HIT development and implementation exist. Stakeholders must rethink the construction of digital tools and hospital IT systems and challenge the prevailing status quo. Health care tools and systems should be designed with flexibility and adaptability at their core to meet the diverse requirements of users in various contexts and should adopt a lightweight approach, enabling them to adapt to advancements and allowing local tailoring. To achieve this, a sufficient implementation strategy is crucial, and robust evaluation methods for these technologies are necessary, emphasizing a thorough examination of users' needs and the contexts in which they operate. Effective and tailored training is another critical aspect of ensuring that HCPs are equipped with the necessary skills to harness the potential of digital tools and data together with addressing the issue of information overload through improved communication guidelines.

In this study, paradoxes emerged from HCPs' perspectives, often highlighting opposing forces within reported experiences. Nevertheless, clinician experiences are frequently neglected in projects that aim to address HIT issues. Historically, emphasis has been placed on IT and commercial teams, relegating clinician insights more toward the periphery. However, the paradoxes identified from clinicians' viewpoints can provide invaluable insights that suggest new strategies to navigate paradoxes and address enduring challenges.

Acknowledgments

The authors would like to thank all clinicians who took the time to participate in this study, as well as those who kindly referred their colleagues, thereby contributing significantly to the recruitment process. In addition, the authors would like to thank all hospital and nursing managers who played a crucial role in effectively disseminating the call for participation, therefore contributing to the recruitment of participants. Furthermore, the authors would like to thank the dedicated secretaries who supported the scheduling of the interviews and hence greatly contributed to the success of this study. This research was funded by the School of Medicine at the University of St. Gallen in Switzerland.

Data Availability

Full interview transcripts and audio files are not publicly available due to data protection reasons; however, metadata, including the codebook and extracts that support the findings, are available on a Zenodo repository [82].

Authors' Contributions

MW contributed to the study design, conducted the interviews, analyzed the data, drafted the original manuscript, and reviewed as well as edited the draft manuscript. LMS analyzed the data and assisted in theme development and reviewed as well as edited the draft manuscript. JH is the supervising author and contributed to the study design, conducted the interviews, assisted in theme development, and reviewed as well as edited the draft manuscript. All authors read, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research): 32-item checklist.
[DOCX File , 39 KB - jmir_v26i1e56095_app1.docx]

Multimedia Appendix 2
Semistructured interview guide.
[DOCX File , 50 KB - jmir_v26i1e56095_app2.docx]

Multimedia Appendix 3
Additional tables with quotation counts and supportive quotes.
[DOCX File , 56 KB - jmir_v26i1e56095_app3.docx]

References

1. Thomas Craig KJ, Willis VC, Gruen D, Rhee K, Jackson GP. The burden of the digital environment: a systematic review on organization-directed workplace interventions to mitigate physician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):985-997 [FREE Full text] [doi: 10.1093/jamia/ocaa301] [Medline: PMC8068437]
2. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022 Jan 27;5(1):13 [FREE Full text] [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
3. Butcher CJ, Hussain W. Digital healthcare: the future. Future Healthc J 2022 Jul;9(2):113-117 [FREE Full text] [doi: 10.7861/fhj.2022-0046] [Medline: 35928188]
4. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407 [FREE Full text] [doi: 10.3390/ijerph20043407] [Medline: 36834105]
5. Patel V, Orchanian-Cheff A, Wu R. Evaluating the validity and utility of wearable technology for continuously monitoring patients in a hospital setting: systematic review. JMIR Mhealth Uhealth 2021 Aug 18;9(8):e17411 [FREE Full text] [doi: 10.2196/17411] [Medline: 34406121]
6. Wu DT, Xu C, Kim A, Bindhu S, Mah KE, Eckman MH. A scoping review of health information technology in clinician burnout. Appl Clin Inform 2021 May;12(3):597-620 [FREE Full text] [doi: 10.1055/s-0041-1731399] [Medline: 34233369]
7. Verma G, Ivanov A, Benn F, Rathi A, Tran N, Afzal A, et al. Analyses of electronic health records utilization in a large community hospital. PLoS One 2020;15(7):e0233004 [FREE Full text] [doi: 10.1371/journal.pone.0233004] [Medline: 32609757]
8. Yan Q, Jiang Z, Harbin Z, Tolbert PH, Davies MG. Exploring the relationship between electronic health records and provider burnout: a systematic review. J Am Med Inform Assoc 2021 Apr 23;28(5):1009-1021 [FREE Full text] [doi: 10.1093/jamia/ocab009] [Medline: 33659988]
9. Kane L. Physician burnout and depression report 2022: stress, anxiety, and anger. Medscape. URL: https://www.medscape.com/slideshow/2022-lifestyle-burnout-6014664 [accessed 2024-04-29]
10. Budd J. Burnout related to electronic health record use in primary care. J Prim Care Community Health 2023;14:21501319231166921 [FREE Full text] [doi: 10.1177/21501319231166921] [Medline: 37073905]
11. Stehman CR, Testo Z, Gershaw RS, Kellogg AR. Burnout, drop out, suicide: physician loss in emergency medicine, part I. West J Emerg Med 2019 May;20(3):485-494 [FREE Full text] [doi: 10.5811/westjem.2019.4.40970] [Medline: 31123550]
12. Ravaghi H, Afshari M, Isfahani P, Bélorgeot VD. A systematic review on hospital inefficiency in the Eastern Mediterranean region: sources and solutions. BMC Health Serv Res 2019 Nov 12;19(1):830 [FREE Full text] [doi: 10.1186/s12913-019-4701-1] [Medline: 31718648]
13. Asbu EZ, Masri MD, Naboulsi MA. Determinants of hospital efficiency: a literature review. Int J Healthc 2020 May 15;6(2):44. [doi: 10.5430/ijh.v6n2p44]
14. Tawfik DS, Sinha A, Bayati M, Adair KC, Shanafelt TD, Sexton JB, et al. Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study. J Med Internet Res 2021 Jul 06;23(7):e26817 [FREE Full text] [doi: 10.2196/26817] [Medline: 34255674]
15. Degen C, Li J, Angerer P. Physicians' intention to leave direct patient care: an integrative review. Hum Resour Health 2015 Sep 08;13(1):74 [FREE Full text] [doi: 10.1186/s12960-015-0068-5] [Medline: 26350545]
16. Berner ES, Detmer DE, Simborg D. Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. J Am Med Inform Assoc 2005;12(1):3-7 [FREE Full text] [doi: 10.1197/jamia.M1664] [Medline: 15492029]
17. Dunn Lopez K, Chin CL, Leitão Azevedo RF, Kaushik V, Roy B, Schuh W, et al. Electronic health record usability and workload changes over time for provider and nursing staff following transition to new EHR. Appl Ergon 2021 May;93:103359. [doi: 10.1016/j.apergo.2021.103359] [Medline: 33556884]
18. Carayon P, Hoonakker P. Human factors and usability for health information technology: old and new challenges. Yearb Med Inform 2019 Aug;28(1):71-77 [FREE Full text] [doi: 10.1055/s-0039-1677907] [Medline: 31419818]

XSL•FO
RenderX

19. Chen M, Decary M. Artificial intelligence in healthcare: an essential guide for health leaders. Healthc Manage Forum 2020 Jan;33(1):10-18. [doi: 10.1177/0840470419873123] [Medline: 31550922]
20. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
21. Ehrmann DE, Gallant SN, Nagaraj S, Goodfellow SD, Eytan D, Goldenberg A, et al. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nat Med 2022 Jul 31;28(7):1331-1333. [doi: 10.1038/s41591-022-01833-z] [Medline: 35641825]
22. Gandhi TK, Classen D, Sinsky CA, Rhew DC, Vande Garde N, Roberts A, et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? JAMIA Open 2023 Oct;6(3):ooad079 [FREE Full text] [doi: 10.1093/jamiaopen/ooad079] [Medline: 37655124]
23. Haleem A, Javaid M, Pratap Singh R, Suman R. Medical 4.0 technologies for healthcare: features, capabilities, and applications. Internet Thing Cyber Phys Syst 2022;2:12-30. [doi: 10.1016/j.iotcps.2022.04.001]
24. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
25. Baillette P, Barlette Y, Berthevas JF. Benefits and risks of shadow IT in health care: a narrative review of the literature. Inf Manag 2022;27(2):59-96. [doi: 10.3917/sim.222.0059]
26. Rogers EM. Diffusion of Innovations. New York, NY: Free Press of Glencoe; 1962.
27. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
28. Alter S. Work system theory: overview of core concepts, extensions, and challenges for the future. J Assoc Inf Syst 2013 Feb;14(2):72-121. [doi: 10.17705/1jais.00323]
29. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634 [FREE Full text] [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
30. Ward R. The application of technology acceptance and diffusion of innovation models in healthcare informatics. Health Policy Technol 2013 Dec;2(4):222-228. [doi: 10.1016/j.hlpt.2013.07.002]
31. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023 Oct 17;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
32. Laukka E, Huhtakangas M, Heponiemi T, Kanste O. Identifying the roles of healthcare leaders in HIT implementation: a scoping review of the quantitative and qualitative evidence. Int J Environ Res Public Health 2020 Apr 21;17(8):2865 [FREE Full text] [doi: 10.3390/ijerph17082865] [Medline: 32326300]
33. Desveaux L, Soobiah C, Bhatia RS, Shaw J. Identifying and overcoming policy-level barriers to the implementation of digital health innovation: qualitative study. J Med Internet Res 2019 Dec 20;21(12):e14994 [FREE Full text] [doi: 10.2196/14994] [Medline: 31859679]
34. Herrmann T, Pfeiffer S. Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. AI Soc 2022 Feb 18;38(4):1523-1542. [doi: 10.1007/s00146-022-01391-5]
35. Schusselé Filliettaz S, Berchtold P, Kohler D, Peytremann-Bridevaux I. Integrated care in Switzerland: results from the first nationwide survey. Health Policy 2018 Jun;122(6):568-576 [FREE Full text] [doi: 10.1016/j.healthpol.2018.03.006] [Medline: 29650245]
36. Tanwir F, Moideen S, Habib R. Interviews in healthcare: a phenomenological approach a qualitative research methodology. Int J Public Health 2021;4(2):10-15. [doi: 10.14302/issn.2641-4538.jphi-21-3881]
37. Knott E, Rao AH, Summers K, Teeger C. Interviews in the social sciences. Nat Rev Methods Primers 2022 Sep 15;2(1):73. [doi: 10.1038/S43586-022-00150-6]
38. Al-Busaidi ZQ. Qualitative research and its uses in health care. Sultan Qaboos Univ Med J 2008 Mar;8(1):11-19 [FREE Full text] [Medline: 21654952]
39. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
40. The #1 software for qualitative data analysis. ATLAS.ti. URL: https://atlasti.com/ [accessed 2024-04-29]
41. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
42. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
43. Spoke.App. SAS Spoke. URL: https://spoke.app/ [accessed 2024-04-29]
44. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
45. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 2016 Nov 29;5(1):80-92. [doi: 10.1177/160940690600500107]

XSL•FO
RenderX

46. Mitchell GJ. Living paradox in Parse's theory. Nurs Sci Q 1993 Aug 19;6(1):44-51. [doi: 10.1177/089431849300600111] [Medline: 8455873]
47. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022:202206e [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
48. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Bohr A, editor. Artificial Intelligence in Healthcare. Cambridge, MA: Academic Press; 2020:25-60.
49. Sinsky CA, Bavafa H, Roberts RG, Beasley JW. Standardization vs customization: finding the right balance. Ann Fam Med 2021;19(2):171-177 [FREE Full text] [doi: 10.1370/afm.2654] [Medline: 33685879]
50. Blum BI. Clinical information systems--a review. West J Med 1986 Dec;145(6):791-797 [FREE Full text] [Medline: 3811347]
51. Islam MM, Poly TN, Li YJ. Recent advancement of clinical information systems: opportunities and challenges. Yearb Med Inform 2018 Aug;27(1):83-90 [FREE Full text] [doi: 10.1055/s-0038-1667075] [Medline: 30157510]
52. Pang TY, Lee TK, Murshed M. Towards a new paradigm for digital health training and education in Australia: exploring the implication of the fifth industrial revolution. Applied Sciences 2023 Jun 05;13(11):6854. [doi: 10.3390/app13116854]
53. Jongsma KR, Sand M, Milota M. Why we should not mistake accuracy of medical AI for efficiency. NPJ Digit Med 2024 Mar 04;7(1):57 [FREE Full text] [doi: 10.1038/s41746-024-01047-2] [Medline: 38438477]
54. Calbimonte JP, Aidonopoulos O, Dubosson F, Pocklington B, Kebets I, Legris PM, et al. Decentralized semantic provision of personal health streams. J Web Semant 2023 Apr;76:100774. [doi: 10.1016/j.websem.2023.100774]
55. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1(Suppl 1):S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
56. Aguirre RR, Suarez O, Fuentes M, Sanchez-Gonzalez MA. Electronic health record implementation: a review of resources and tools. Cureus 2019 Sep 13;11(9):e5649 [FREE Full text] [doi: 10.7759/cureus.5649] [Medline: 31700751]
57. Jeyakumar T, McClure S, Lowe M, Hodges B, Fur K, Javier-Brozo M, et al. An education framework for effective implementation of a health information system: scoping review. J Med Internet Res 2021 Feb 24;23(2):e24691 [FREE Full text] [doi: 10.2196/24691] [Medline: 33625370]
58. Nabovati E, Jeddi FR, Ghaffari F, Mirhoseini F. The effects of simulation training on learning of health information systems: a scoping review. J Educ Health Promot 2022;11:4 [FREE Full text] [doi: 10.4103/jehp.jehp_17_21] [Medline: 35281403]
59. Song C, Shin SY, Shin KS. Implementing the dynamic feedback-driven learning optimization framework: a machine learning approach to personalize educational pathways. Appl Sci 2024 Jan 21;14(2):916. [doi: 10.3390/app14020916]
60. Fontaine G, Cossette S, Maheu-Cadotte MA, Mailhot T, Deschênes MF, Mathieu-Dupuis G, et al. Efficacy of adaptive e-learning for health professionals and students: a systematic review and meta-analysis. BMJ Open 2019 Aug 28;9(8):e025252 [FREE Full text] [doi: 10.1136/bmjopen-2018-025252] [Medline: 31467045]
61. Gawne S, Fish R, Machin L. Developing a workplace-based learning culture in the NHS: aspirations and challenges. J Med Educ Curric Dev 2020;7:2382120520947063 [FREE Full text] [doi: 10.1177/2382120520947063] [Medline: 32821850]
62. Zechariah S, Ansa BE, Johnson SW, Gates AM, Leo GD. Interprofessional education and collaboration in healthcare: an exploratory study of the perspectives of medical students in the United States. Healthcare (Basel) 2019 Oct 15;7(4):117 [FREE Full text] [doi: 10.3390/healthcare7040117] [Medline: 31618920]
63. Sell K, Hommes F, Fischer F, Arnold L. Multi-, inter-, and transdisciplinarity within the public health workforce: a scoping review to assess definitions and applications of concepts. Int J Environ Res Public Health 2022 Sep 01;19(17):10902 [FREE Full text] [doi: 10.3390/ijerph191710902] [Medline: 36078616]
64. Bosch B, Mansell H. Interprofessional collaboration in health care: lessons to be learned from competitive sports. Can Pharm J (Ott) 2015 Jul;148(4):176-179 [FREE Full text] [doi: 10.1177/1715163515588106] [Medline: 26448769]
65. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. P T 2017 Sep;42(9):572-575 [FREE Full text] [Medline: 28890644]
66. Shenoy A, Appel JM. Safeguarding confidentiality in electronic health records. Camb Q Healthc Ethics 2017 Apr 31;26(2):337-341. [doi: 10.1017/S0963180116000931] [Medline: 28361730]
67. Armstrong MJ. Improving email strategies to target stress and productivity in clinical practice. Neurol Clin Pract 2017 Dec;7(6):512-517 [FREE Full text] [doi: 10.1212/CPJ.0000000000000395] [Medline: 29431178]
68. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, et al. The impact of administrative burden on academic physicians: results of a hospital-wide physician survey. Acad Med 2017 Feb;92(2):237-243. [doi: 10.1097/ACM.0000000000001461] [Medline: 28121687]
69. Duong D, Vogel L. Overworked health workers are "past the point of exhaustion". CMAJ 2023 Feb 27;195(8):E309-E310 [FREE Full text] [doi: 10.1503/cmaj.1096042] [Medline: 36849179]
70. Martikainen S, Kaipio J, Lääveri T. End-user participation in health information systems (HIS) development: physicians' and nurses' experiences. Int J Med Inform 2020 May;137:104117. [doi: 10.1016/j.ijmedinf.2020.104117] [Medline: 32179254]
71. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 2024 Feb;177(2):210-220. [doi: 10.7326/M23-2772] [Medline: 38285984]

XSL•FO
RenderX

72. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150 [FREE Full text] [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
73. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag 2013;10(Fall):1c [FREE Full text] [Medline: 24159271]
74. Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. JAMA Netw Open 2020 Jun 01;3(6):e205867 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.5867] [Medline: 32515797]
75. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615 [FREE Full text] [doi: 10.2196/42615] [Medline: 37000497]
76. Cheng CG, Wu DC, Lu JC, Yu CP, Lin HL, Wang MC, et al. Restricted use of copy and paste in electronic health records potentially improves healthcare quality. Medicine (Baltimore) 2022 Jan 28;101(4):e28644 [FREE Full text] [doi: 10.1097/MD.0000000000028644] [Medline: 35089204]
77. Purkayastha S, Allam R, Maity P, Gichoya JW. Comparison of open-source electronic health record systems based on functional and user performance criteria. Healthc Inform Res 2019 Apr;25(2):89-98 [FREE Full text] [doi: 10.4258/hir.2019.25.2.89] [Medline: 31131143]
78. Nijor S, Rallis G, Lad N, Gokcen E. Patient safety issues from information overload in electronic medical records. J Patient Saf 2022 Sep 01;18(6):e999-1003 [FREE Full text] [doi: 10.1097/PTS.0000000000001002] [Medline: 35985047]
79. Toscano F, O'Donnell E, Broderick JE, May M, Tucker P, Unruh MA, et al. How physicians spend their work time: an ecological momentary assessment. J Gen Intern Med 2020 Nov;35(11):3166-3172 [FREE Full text] [doi: 10.1007/s11606-020-06087-4] [Medline: 32808212]
80. Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. J Am Med Inform Assoc 2017 Mar 01;24(2):246-250 [FREE Full text] [doi: 10.1093/jamia/ocw154] [Medline: 28011595]
81. Morris C, Scott RE, Mars M. WhatsApp in clinical practice-the challenges of record keeping and storage. A scoping review. Int J Environ Res Public Health 2021 Dec 20;18(24):13426 [FREE Full text] [doi: 10.3390/ijerph182413426] [Medline: 34949033]
82. Interview code analysis. Zenodo. URL: https://zenodo.org/records/10116088 [accessed 2024-04-29]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
HCP: health care professional
HIT: health information technology

Edited by K Williams; submitted 08.01.24; peer-reviewed by E Fioratou, S Rajput, T Makasi; comments to author 29.02.24; revised version received 21.03.24; accepted 16.04.24; published 15.07.24.

Please cite as:
Wosny M, Strasser LM, Hastings J
The Paradoxes of Digital Tools in Hospitals: Qualitative Interview Study
J Med Internet Res 2024;26:e56095
URL: https://www.jmir.org/2024/1/e56095
doi:10.2196/56095
PMID:39008341

©Marie Wosny, Livia Maria Strasser, Janna Hastings. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Five-Feature Models to Predict Preeclampsia Onset Time From Electronic Health Record Data: Development and Validation Study

Hailey K Ballard[1,2], BS; Xiaotong Yang[1], MS; Aditya D Mahadevan[3,4], BS; Dominick J Lemas[2,3,5], PhD; Lana X Garmire[1], PhD

[1]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, United States
[2]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States
[3]Center for Research in Perinatal Outcomes, University of Florida, Gainesville, FL, United States
[4]Department of Physiology and Aging, University of Florida, Gainesville, FL, United States
[5]Department of Obstetrics & Gynecology, University of Florida, Gainesville, FL, United States

Corresponding Author:
Lana X Garmire, PhD
Department of Computational Medicine and Bioinformatics
University of Michigan Medical School
Room 3366, Building 520, NCRC
1600 Huron Parkway
Ann Arbor, MI, 48105
United States
Phone: 1 734 615 0514
Email: lgarmire@gmail.com

Abstract

Background: Preeclampsia is a potentially fatal complication during pregnancy, characterized by high blood pressure and the presence of excessive proteins in the urine. Due to its complexity, the prediction of preeclampsia onset is often difficult and inaccurate.

Objective: This study aimed to create quantitative models to predict the onset gestational age of preeclampsia using electronic health records.

Methods: We retrospectively collected 1178 preeclamptic pregnancy records from the University of Michigan Health System as the discovery cohort, and 881 records from the University of Florida Health System as the validation cohort. We constructed 2 Cox-proportional hazards models: 1 baseline model using maternal and pregnancy characteristics, and the other full model with additional laboratory findings, vitals, and medications. We built the models using 80% of the discovery data, tested the remaining 20% of the discovery data, and validated with the University of Florida data. We further stratified the patients into high- and low-risk groups for preeclampsia onset risk assessment.

Results: The baseline model reached Concordance indices of 0.64 and 0.61 in the 20% testing data and the validation data, respectively, while the full model increased these Concordance indices to 0.69 and 0.61, respectively. For preeclampsia diagnosed at 34 weeks, the baseline and full models had area under the curve (AUC) values of 0.65 and 0.70, and AUC values of 0.69 and 0.70 for preeclampsia diagnosed at 37 weeks, respectively. Both models contain 5 selective features, among which the number of fetuses in the pregnancy, hypertension, and parity are shared between the 2 models with similar hazard ratios and significant *P* values. In the full model, maximum diastolic blood pressure in early pregnancy was the predominant feature.

Conclusions: Electronic health records data provide useful information to predict the gestational age of preeclampsia onset. Stratification of the cohorts using 5-predictor Cox-proportional hazards models provides clinicians with convenient tools to assess the onset time of preeclampsia in patients.

(J Med Internet Res 2024;26:e48997) doi:10.2196/48997

KEYWORDS

preeclampsia; survival analysis; risk prediction; pregnancy; prognosis; survival; risk; mortality; EHR; health records; maternal; machine learning; electronic health records

XSL•FO
RenderX

Introduction

Preeclampsia is a pregnancy-associated condition characterized by new-onset hypertension and proteinuria, typically diagnosed after 20 weeks of gestation in approximately 3%-5% of all pregnancies [1]. As one of the leading causes of maternal mortality and morbidity worldwide, it can lead to a more serious condition called eclampsia if left untreated [2]. Timely identification of preeclampsia is a key factor in pregnancy risk management and subsequent treatment. Current medical practice guideline recommends prevention therapy of low-dose aspirin on women at high risk for preeclampsia before the 13-week gestation period [3]. However, preeclampsia does not typically manifest itself clinically until after 20 weeks of gestation, through clinical markers such as blood pressure (BP), urinary protein excretion, mean arterial pressure, and placental growth factor levels. Moreover, the gestational age of preeclampsia onset can vary greatly across pregnancies [3]. Preeclampsia diagnosed before 34 weeks of gestation is called early-onset preeclampsia, and late-onset preeclampsia is diagnosed after 34 weeks [4]. To allow for maximal efficiency of prevention therapy, tools that accurately predict the onset time of preeclampsia and the patient risk will be extremely beneficial.

Previous studies have identified some qualitative risk factors of preeclampsia, including preeclampsia in a previous pregnancy, a multifetal pregnancy, chronic hypertension, kidney disease, diabetes before pregnancy, autoimmune disorders, as well as demographic factors including obesity, advanced maternal age, and race [5]. However, the quantitative importance of these risk factors relative to one another has not been adequately investigated. Haile et al [6] discuss how maternal age, weight, and history of preeclampsia significantly drive preeclampsia onset time, but many additional factors remain undefined. There is an unmet need to provide clinicians with tools to accurately identify which mothers are at risk for preeclampsia, and further identify when they will develop preeclampsia.

Prognosis modeling using population-level health data provides opportunities to systematically address both issues mentioned above [7]. These new models enable the investigation of risk factors (features) that may affect the gestational age at preeclampsia diagnosis, using the hazard ratio (HR), which indicates the importance of the risk factors. Each model outputs risk factors that influence preeclampsia development and predicts the gestational age at preeclampsia diagnosis for patients using the weighted impact of each feature. In addition, patients can be stratified into low-risk and high-risk preeclampsia groups, accompanied by differences in risk factors (features). These developed and validated prognosis models will allow clinicians to practically identify when an at-risk mother might develop preeclampsia and reveal any features associated with the onset time of preeclampsia that are not included in the current guidelines.

Methods

Data Source

The discovery cohort for this project was obtained from the University of Michigan (UM) Medicine Healthcare System. All deidentified pregnancy records between the years 2015 and 2021, with at least one preeclampsia diagnosis, based on the *ICD-10-CM* (*International Classification of Diseases, Tenth Revision, Clinical Modification*) codes, were extracted (Table S1 in Multimedia Appendix 1). Patients who were diagnosed with competing conditions (Table S1 in Multimedia Appendix 1) were removed from the cohort. Patients who did not have any electronic medical record (EMR) in the UM system within 20 weeks of the start of their pregnancy were also removed. Since preeclampsia is clinically defined after 20 weeks, all patients with a preeclampsia diagnosis before 20 weeks of gestation were dropped from the discovery cohort. A total of 1178 pregnancies remained in the UM discovery cohort after this data selection.

Following the same inclusion and exclusion criteria, the validation cohort was generated from the University of Florida Health System and contained 881 preeclamptic pregnancies from 2015 to 2021. The Integrated Data Repository managed the deidentification and transfer of patient data to the researchers.

Feature Extraction and Preprocessing

The electronic medical records include medical history, obstetric diagnostic codes entered during each unique pregnancy, demographics, medications, laboratory results, and vital signs (Table S2 in Multimedia Appendix 1). The baseline model initially used age at the start of pregnancy, race, pregnancy start date, date of the first preeclampsia diagnosis, gravidity, parity, and previous history of preeclampsia at the trimester it was diagnosed. In addition, medical histories based on *ICD-10-CM* diagnosis codes were extracted using the Elixhauser Comorbidities definitions [8]. Current diagnoses entered within 20 weeks of gestation were extracted using the same *ICD-10-CM* diagnosis codes and definitions.

The full model includes all features in the baseline model. In addition, laboratory results, vital signs, and medications ordered before 20 weeks of gestation were also added. Laboratory tests that included a complete blood count were considered (Table S2 in Multimedia Appendix 1). Vital signs included diastolic and systolic BP. Laboratory findings and vitals collected from the start of pregnancy (0 days gestation) to 20 weeks (140 days) gestation were included. The mean, maximum, minimum, and SD for each laboratory value were calculated. Medication records were retrieved based on previous reports that medications prescribed during pregnancy may be related to preeclampsia development [9]. Patients who did not have any laboratory finding or vital data collected and entered in the EMR system within the first 20 weeks (~15%) were assigned as “missing”. These missing values were imputed using the predictive mean matching algorithm from the R package “mice” [10], which has been shown to produce the least-biased results for data sets that use feature selection [11-13]. The standards for missing data used for multiple imputations were followed,

XSL•FO
RenderX

and imputation was performed on only the variables with no more than 20% missingness [14]. All numeric variables were log-transformed to adjust for skewness. Each feature in the medical history, clinical diagnosis, and medication categories was computed as a binary category: 1 for presence, and 0 for absence, to reduce feature dimensionality and improve interpretability. All analysis was conducted using R (version 4.2.2; The R Foundation) [15]. Data cleaning was carried out using the packages "dplyr" [16] and "gtsummary" [17].

Feature Selection, Model Construction, and Evaluations

The UM discovery data set was randomly divided into a training set (80%) and a hold-out testing set (20%) after multiple imputations on missing variables. A Cox-proportional hazards model with Least Absolute Shrinkage and Selection Operator (LASSO) regularization was conducted through 5-fold cross-validation, using the "glmnet" [18] package in R. We used cross-validation to select the optimal LASSO hyperparameter (lambda) that gave the smallest mean squared error and then performed bootstrapping with 1000 replicates to calculate a concordance index (C-index) and 95% CIs for each data set (training, testing, and validation). The baseline model had an optimal lambda of 0.0058 (Figure S1A in Multimedia Appendix 2) and the full model had an optimal lambda of 0.0066 (Figure S1B in Multimedia Appendix 2). The baseline model had 31 features and the full model had 92 features before selection. Following regularized feature selection using the LASSO method on the training data sets, both final models have 5 selected features. The output of the Cox-PH model is the log hazard ratio, also called the prognosis index (PI), which depicts the relative risk of a patient when compared with the baseline hazard of the population. The full model was constructed in the same way as the baseline model.

External validation on each finalized model (baseline and full models) was done through collaboration with the University of Florida (UF), where the electronic health record (EHR) data and patient characteristics are different. Each feature chosen by the model was able to be identified in the UF validation cohort except for the nonsteroidal anti-inflammatory drug (NSAID) medication prescription, which was not available at the time of collection.

The performance of each model was evaluated using the C-index with bootstrapping of 1000 replicates to calculate 95% CI and *P* values from log-rank tests. The C-index is a metric to compare the discriminative power of a risk prediction model that describes the frequency of concordant pairs among all pairs of patients included in the model construction [19]. We used the C-index calculated from the "cindex" [20] function. Low- and high-risk pregnancies were stratified based on the median PI score of the model, and Kaplan-Meier curves were plotted for each risk group. Their differences were tested with log-rank tests using the training data set, hold-out testing data set, and the validation data set separately to evaluate the discriminative power of the model. The log-rank test is a significance test in survival analysis, with the null hypothesis that 2 groups have identical distributions of survival time. Any log-rank *P* value below .05 is considered statistically significant in these analyses. Feature importance was evaluated in the Cox-PH model by their HR *P* values. HR describes the relative contribution of a feature to the patient's PI. In the context of our model, HRs above 1 shorten the gestational age of preeclampsia diagnosis, while HRs below 1 lengthen it.

We further measured model performance by calculating the sensitivity and specificity for each model, classified by predicting preeclampsia diagnosis by 34 and 37 weeks, respectively. We also plotted the area under the curve (AUC) from each testing data set for both models at both time points, using the "pROC" [21] package in R.

Ethical Considerations

The institutional review board (IRB) of the UM Medical School (HUM#00168171) and the UF (#201601899) approved the original data collection and the use of the discovery cohort. All authors have permission for the use of this data. IRB approval was not required for the secondary analysis presented here, as it was deemed exempt. [22].

Results

Study Design and Data Set Overviews

The overall study design is shown in Figure 1. The discovery cohort was extracted from patient records in the UM Health System from 2015 to 2022 with *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) code access. All patients with a preeclampsia diagnosis after 20 weeks of gestation were included in the cohort, and other exclusion criteria are detailed in the Methods section. The finalized UM discovery cohort consists of EMRs from 1178 pregnancies. Using the same inclusion and exclusion criteria, 881 pregnancies were identified in the validation data set from UF. The patient characteristics for each cohort are listed in Table 1. The average maternal age was 30.2 years (SD 5.67) in the discovery cohort and 29.1 years (SD 6.18) in the validation cohort. The mean gestational age of preeclampsia onset was 251 (SD 25.4) days for the discovery cohort and 257 (SD 25.9) days for the validation cohort. We constructed and validated 2 models using this data: (1) a baseline model using only patient medical history, demographics, and diagnoses of any new medical issues within the first 20 weeks of gestation; and (2) a full model including those features from the baseline model, as well as additional information on medication, laboratory findings, and vitals within the first 20 weeks of pregnancy.

XSL•FO
RenderX

Figure 1. Study design and workflow for the University of Michigan preeclampsia cohort (N=1178) and the University of Florida preeclampsia cohort (N=881), 2015-2021. The discovery cohort was obtained from the University of Michigan Health System and a validation cohort of similar size and time was obtained from the University of Florida Health System. We constructed 2 preeclampsia predictive models: baseline and full model. The input variables in baseline models include patients' demographics, lifestyle, comorbidities, and medical history (n=31) which were reduced to 5 features. The input for the full model includes additional lab tests and vital signs around preeclampsia diagnosis time, in addition to the variables in the baseline models (n=92), and was reduced to 5 features for the discovery cohort, and 4 features for the validation cohort. We trained the Cox-proportional hazards models with the Least Absolute Shrinkage and Selection Operator regularization, using 80% training from the University of Michigan discovery cohort. We tested it on 20% hold-out data from the same discovery cohort and validated it using the University of Florida validation cohort. Cox-PH: Cox proportional-hazard; LASSO: Least Absolute Shrinkage and Selection Operator; PE: preeclampsia; UF: University of Florida; UM: University of Michigan.

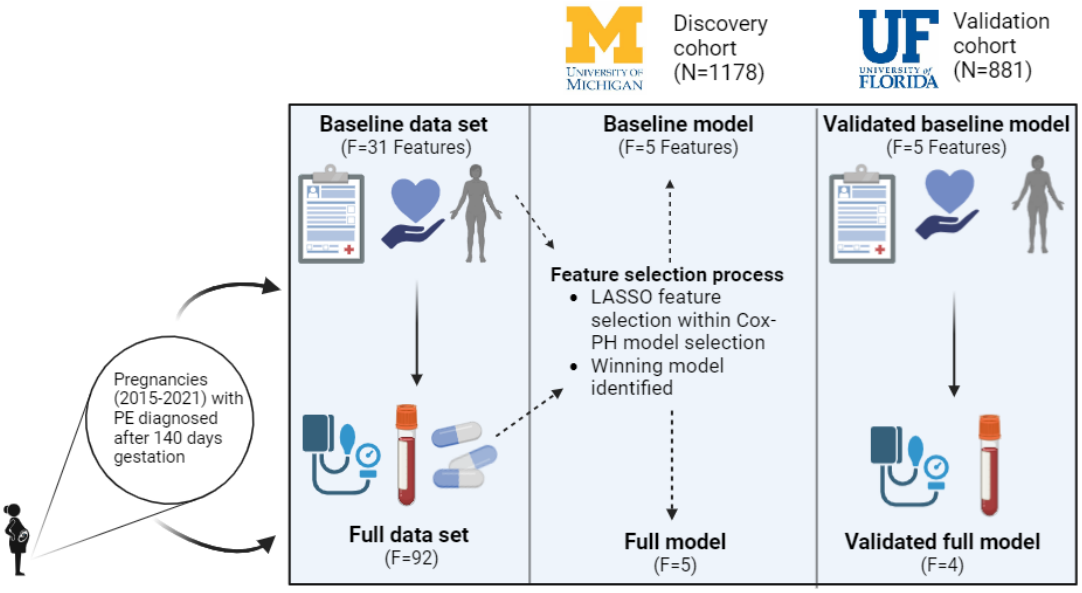

XSL•FO
RenderX

Table 1. Summaries of the characteristics of the University of Michigan preeclampsia discovery cohort (N=1178) and University of Florida preeclampsia validation cohort (N=881) for patients admitted from 2015 to 2021. Data are presented as the average (SD) or counts (% in the cohort).

Characteristics	Discovery cohort (N=1178)	Validation cohort (N=881)
Maternal age (years), mean (SD)	30.2 (5.67)	29.1 (6.18)
Gravidity, mean (SD)	2.31 (1.74)	2.82 (2.04)
Parity, mean (SD)	0.68 (1.12)	1.17 (1.5)
Number of fetuses, mean (SD)	1.07 (0.26)	1.04 (0.22)
Gestational age at PE onset (days), mean (SD)	251 (25.4)	257 (25.9)
Current smoker, n (%)	61 (5)	112 (13)
Current alcohol user, n (%)	311 (26)	184 (21)
Race or ethnicity, n (%)		
African American	195 (17)	335 (38)
Asian	74 (6)	19 (2)
Hispanic	58 (5)	4 (1)
History of PE[a]	184 (16)	117 (13)
History of PE diagnosed in the second trimester	66 (6)	3 (<1)
Medical history, n (%)		
Uncomplicated type I diabetes	34 (3)	19 (2)
Uncomplicated type II diabetes	62 (5)	22 (3)
Uncomplicated hypertension	201 (17)	81 (9)
Kidney disease	14 (1)	1 (<1)
Other clinical diagnoses within 20 weeks of gestation, n (%)		
Depression	265 (22)	19 (2)
Mood and anxiety disorder	318 (27)	0

[a]PE: preeclampsia.

Baseline Model

A baseline model was first built using medical history, demographics, and *ICD-10-CM* diagnosis codes of new medical conditions entered during the first 20 weeks of pregnancy. To build and test the model, we randomly split the data into an 80:20 ratio for training and testing data sets, and the Cox-PH model with LASSO (L1) regularization was built with the UM training data under 5-fold cross-validation. Alternatively, we explored ElasticNet (combined L1 and L2 regularization) as well as L2 penalization. However, the LASSO (L1) model overall performs better with higher C-indices and fewer features over these alternatives. We therefore chose LASSO as the regularization method (Table S3 in Multimedia Appendix 1).

We then applied this model to the 20% UM hold-out testing data and external UF validation cohort. The C-indices for the training, hold-out testing, and external validation data of the baseline model are 0.62 (95% CI 0.61-0.63), 0.64 (95% CI 0.60-0.65), and 0.61 (95% CI 0.59-0.63), confirming its validity. Table 2 shows the baseline model's C-index and corresponding 95% CI values for each data set. To further facilitate interpretation, we classified each preeclampsia diagnosis prediction by the timeline of its occurrence, specifically by gestational weeks 34 and 37, using the UM hold-out testing data set. Such simple binary classification shows a sensitivity of 0.74, specificity of 0.50, and AUC of 0.65 for preeclampsia diagnosed at 34 weeks (Table 2). It has improved performance for preeclampsia diagnosis by 37 weeks, with a sensitivity of 0.82, specificity of 0.50, and AUC of 0.69 (Table 2 and Multimedia Appendix 3).

Five features were selected for the baseline model. Their respective HRs and rankings in the multivariate Cox-PH are depicted in Figure 2A and Table 3. By the descending order of HR, these features are the number of fetuses in pregnancy of interest (HR 25.2; $P<.001$), parity (HR 2.08; $P<.001$), history of uncomplicated hypertension (HR 2.01; $P<.001$), history of uncomplicated type II diabetes (HR 1.87; $P<.001$), and a mood or anxiety disorder (HR 1.24; $P=.01$). All features increase preeclampsia risk and shorten the gestational age of preeclampsia diagnosis.

XSL•FO
RenderX

Table 2. Binarized performance for baseline and full models using 34- and 37-week preeclampsia diagnosis occurrences, measured using the hold-out testing data obtained from the randomly selected 20% patients from the University of Michigan discovery cohort.

Model version	34 weeks			37 weeks		
Metrics	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Baseline	0.74	0.50	0.65	0.82	0.50	0.70
Full	0.98	0.51	0.70	0.86	0.50	0.70

Figure 2. Gestational age of preeclampsia diagnosis baseline model features and performance. (A) Bar plot of hazard ratios of the selected features by Cox-proportional hazards method with Least Absolute Shrinkage and Selection Operator regularization. Ranging from smallest to largest hazard ratio: mood and anxiety disorder, diabetes, hypertension, parity, and number of fetuses. (B-D) Kaplan-Meier survival curves of high-risk (red) and low-risk (blue) pregnancies in the respective data sets, each with a log-rank test *P* value <.001. (B) University of Michigan training data set with a C-index of 0.62. (C) Hold-out testing set with a C-index of 0.64. (D) University of Florida validation data set with a C-index of 0.61.

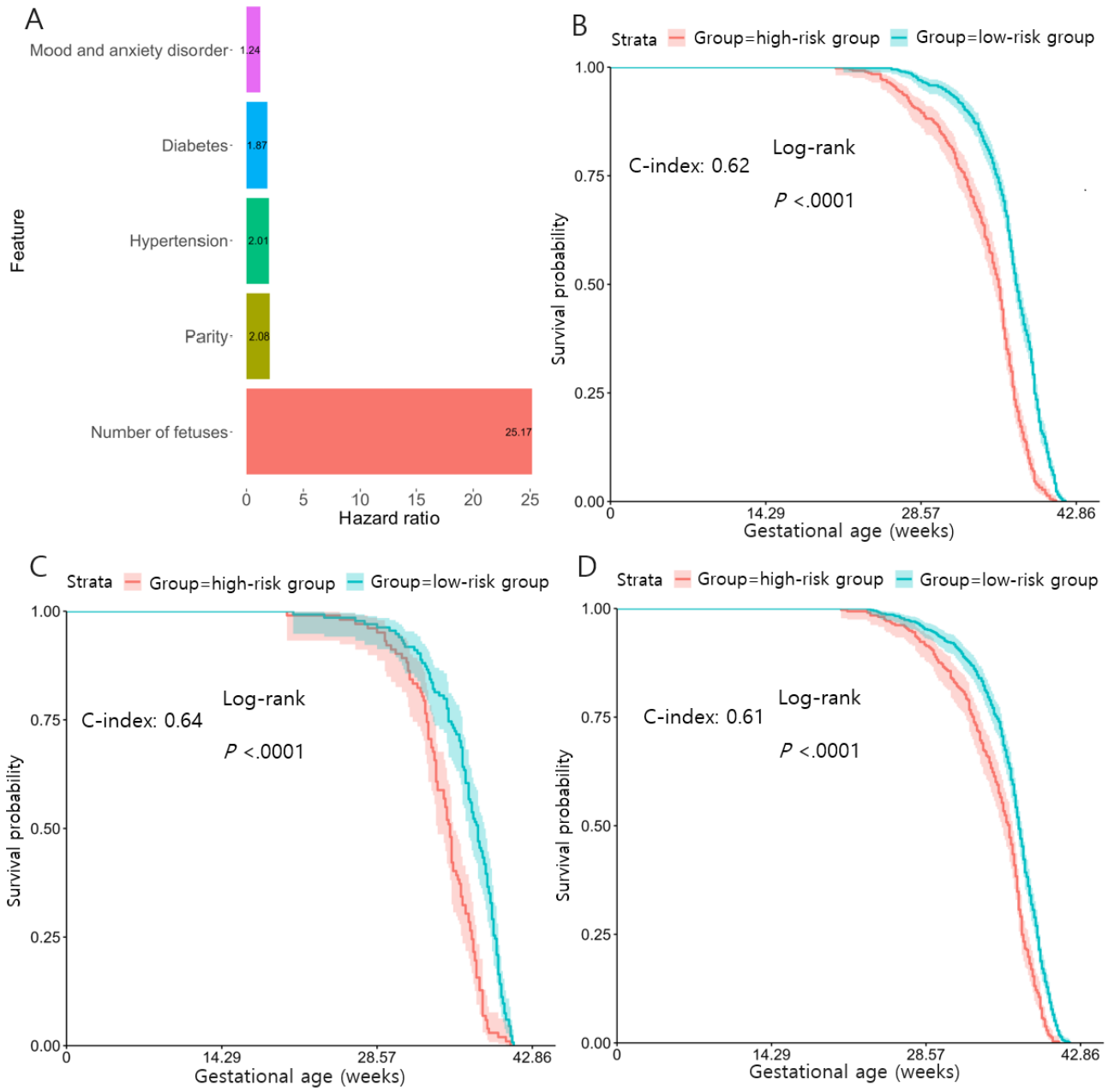

XSL•FO
RenderX

Table 3. The selected features in the baseline model to predict the gestational age of preeclampsia diagnosis.

Features	Hazard ratio (95% CI)	*P* value
Number of fetuses	25.2 (10.7-59.4)	<.001
Parity	2.08 (1.54-2.81)	<.001
History of uncomplicated hypertension	2.01 (1.68-2.40)	<.001
History of uncomplicated type II diabetes	1.87 (1.41-2.49)	<.001
Mood and anxiety disorder	1.24 (1.07-1.43)	.01

To evaluate the discriminative power of this model, patients from the training data set were dichotomized into high- and low-risk groups by stratifying the samples using the median of the predicted PI (PI=1.17) from the model. The 2 risk groups showed significant differences in prognosis (Figure 2B and Table S4 in Multimedia Appendix 1). The high-risk group was characterized by higher parity and number of fetuses, while the low-risk pregnancies had no prevalence of hypertension (*P*<.001) or diabetes (*P*<.001). The median PI value above was applied to categorize samples into high versus low-risk groups in the hold-out (PI=1.17) and validation data (PI=2.38), similar to others [23-25]. As shown in Figures 2C and 2D, the KM curves on these 2 risk groups are also significantly different (*P*<.001).

Full Model

We next evaluated a model with the addition of laboratory findings, vitals, and medications prescribed in the first 20 weeks of gestation to the clinical data used in the baseline model. We constructed the new Cox-PH model, or the "full model," in the same manner as the baseline model and obtained a 5-feature Cox-PH model (Figure 3A). Similar to the baseline model, LASSO regularization shows better overall performance than ElasticNet and L2 regularization and is chosen as the default (Table S3 in Multimedia Appendix 1). This new model reaches the C-indices of 0.66 (95% CI 0.64-0.67) and 0.69 (95% CI 0.64-0.70) for the training and hold-out testing data sets, respectively. It also yields a C-index of 0.61 (95% CI 0.60-0.63) on the UF validation cohort, despite missing 1 feature (NSAID medication) in the UF cohort. Table 2 lists the full-model C-indices and 95% CIs for each data set. Similar to the baseline model, to help interpretation, we classified each preeclampsia diagnosis prediction using the timeline of preeclampsia occurrence by gestational weeks 34 and 37, respectively, using the UM hold-out testing data set. It yields a sensitivity of 0.98, specificity of 0.51, and AUC of 0.70 for correctly predicting preeclampsia by week 34 (Table 2). The model has an improved correct diagnosis by week 37, with a sensitivity of 0.86, specificity of 0.50, and AUC of 0.70 (Table 2 and Multimedia Appendix 3).

The full model also yields 5 features, all with positive HRs (Figure 3A and Table 4). In descending order of HR, these features are maximum diastolic blood pressure (HR 21.7; *P*<.001), number of fetuses in current pregnancy (HR 21.1; *P*<.001), parity (HR 1.81; *P*<.001), history of uncomplicated hypertension (HR 1.79; *P*<.001), and NSAID medication prescription (HR 1.35; *P*<.001). Three of these features, namely the number of fetuses, history of uncomplicated hypertension, and parity features were also selected by the baseline model (Figure 3B). Table S5 in Multimedia Appendix 1 shows each of the features and their HRs in a univariate analysis. Their HRs across the baseline and full models remain very similar and had *P* values less than .05, suggesting that they are all significant in predicting preeclampsia onset time regardless of the other additional input information. Maximum diastolic BP and NSAID medication prescription are newly selected features unique to the full model (Figures 3A and 3B).

Like the baseline model, we stratified patients into high- versus low-risk groups using the median predicted PI value of 5.15 from the training data set (Figure 3C). The high-risk group was characterized by higher parity, a higher number of fetuses, and higher maximum diastolic BP (Table S4 in Multimedia Appendix 1). In contrast, the low-risk group had no history of hypertension and rare use of NSAID medication. BP had the most statistically significant difference (*P*<.001), as expected. The same median threshold was applied to the 20% hold-out testing data set (PI=5.08) and validation data (PI=5.18) for dichotomization (Figures 3D and 3E). KM curves on these 2 risk groups in the testing set have even more significant differences in their gestational age at diagnosis (*P*<.001). Both models are to be used by entering patient information in the predictors to predict when the patient may develop preeclampsia.

To determine the potential impact of missing data on modeling results, we explored building a baseline and full model with only cases that had complete BP data—the main selected feature in the full model. Table S6 in Multimedia Appendix 1 shows the selected features of both of these models. The complete cases baseline model had a training C-index of 0.63 and a testing C-index of 0.64. The complete cases full model had a training C-index of 0.67 and a testing C-index of 0.65. Due to similar performance and selected features, it can be safely assumed that imputation had little impact on the finalized models.

XSL•FO
RenderX

Figure 3. Gestational age of preeclampsia diagnosis full model features and performance. (A) Bar plot of hazard ratios of the selected features in the full model by Cox-proportional hazards method with Least Absolute Shrinkage and Selection Operator regularization. Ranging from smallest to largest hazard ratio: nonsteroidal anti-inflammatory drug use, hypertension, parity, number of fetuses, and maximum diastolic blood pressure. (B) The bubble plot of significant features from preeclampsia baseline and full models. The size of the bubbles represents the hazard ratio of each feature. The number of fetuses, parity, and hypertension were shared between both models with similar hazard ratios. (C-E) Kaplan-Meier survival curves of high-risk (red) and low-risk (blue) pregnancies in the respective data sets, each with a log-rank test *P* value <.001. (C) University of Michigan training data set with a concordance index of 0.62. (D) Hold-out testing set with a concordance index of 0.64. (E) University of Florida validation data set with a concordance index of 0.61. BP: blood pressure.

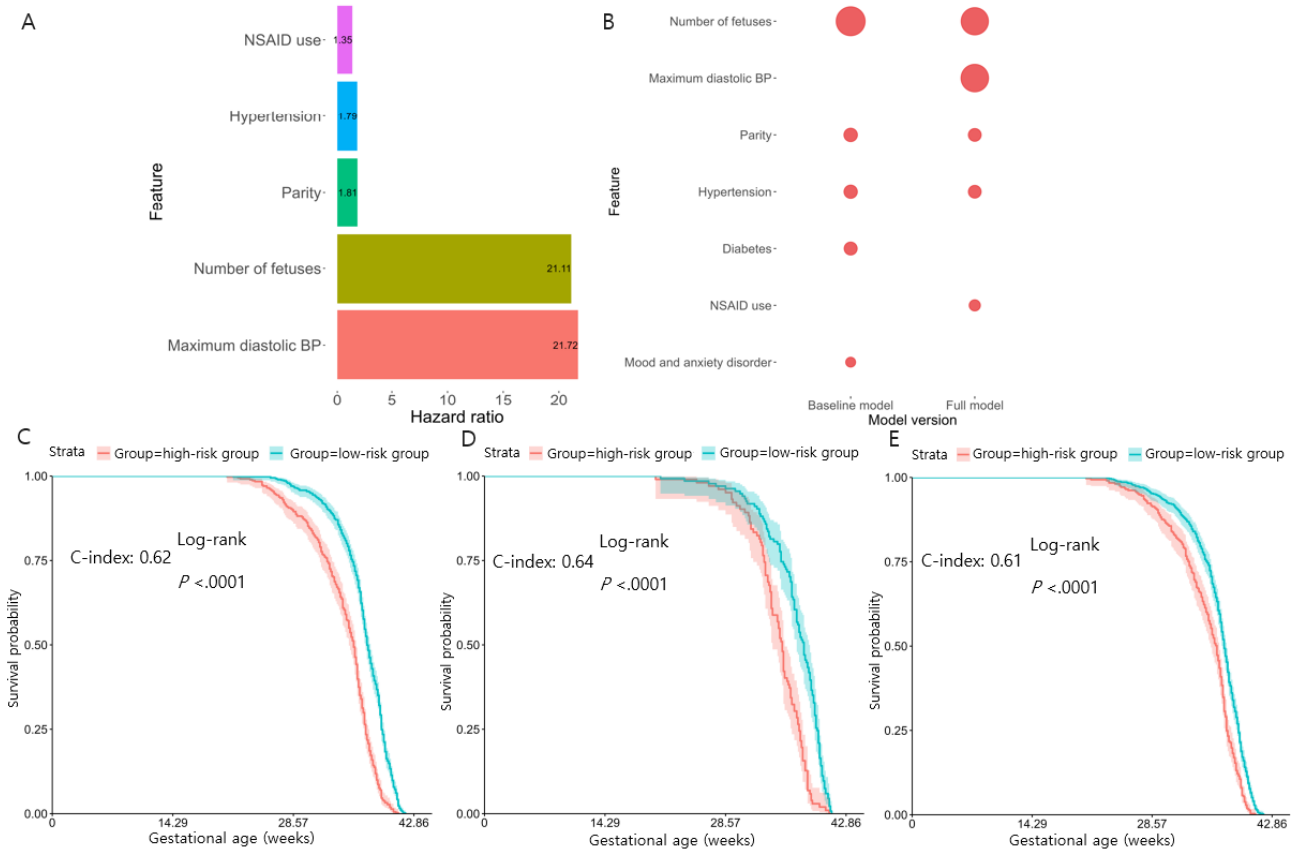

Table 4. Summary of the selected features in the full model to predict the gestational age upon preeclampsia diagnosis.

Features	Hazard ratio (95% CI)	*P* value
Maximum diastolic blood pressure	21.7 (7.93-59.8)	<.001
Number of fetuses	21.1 (9.88-45.1)	<.001
Parity	1.81 (1.37-2.39)	<.001
History of uncomplicated hypertension	1.79 (1.53-2.11)	<.001
NSAID[a] medication	1.35 (1.15-1.58)	<.001

[a]NSAID: nonsteroidal anti-inflammatory drug.

Discussion

Principal Results

This paper is the first of its kind to implement and externally validate a prognosis-predicting model for preeclampsia onset time using EHR data from the first 20 weeks of pregnancy [26]. These models confirmed that factors such as BP in the first 20 weeks of pregnancy, the number of fetuses, parity, and previous history of hypertension are associated with earlier preeclampsia onset time. Moreover, comorbidities such as gestational diabetes and anxiety, as well as NSAID medication, shorten preeclampsia onset time. The similar performance across validation and development data sets provides confidence in the accuracy of the predictive outputs.

Comparison With Previous Work

A recent study stratified patients with preeclampsia by gestational age to build classification models, resulting in many models that are difficult for clinicians to select from [27]. Moreover, these classification models cannot predict the gestational age of onset for an individual patient, thus failing to assist clinicians in making early decisions on delivery plans and proper antenatal care. Unlike most other accurate preeclampsia onset time prediction models, our models only use EMR data from the first 20 weeks of pregnancy and do not

XSL•FO
RenderX

require advanced testing inputs, such as biomarkers [27], enabling earlier use in clinics. In a systematic review of 68 preeclampsia prediction models [27], only 6% (4/68) of them were externally validated, and those not requiring complex biomarker features had much lower AUCs (0.58-0.61) than the models presented here (AUC 0.65-0.70), highlighting the accuracy of our models once validated against a different patient population.

Clinical and Research Implications

Due to the difficulty in predicting preeclampsia, accurate models that can identify women at high risk for preeclampsia can provide early targeted treatment as well as increased surveillance to reduce adverse outcomes [28]. The models here not only confirm the importance of some previously known risk factors, such as the number of fetuses, history of hypertension, and parity but also assign quantitative scores (weights) on the importance of these risk factors relative to each other. This is a significant advancement from most of the other studies focusing on a single risk factor. It also provides clinicians as well as pregnant women with quantitative tools to assess the onset time of preeclampsia more accurately, beyond the qualitative assessment of risks. Risk factors with higher weights can take a higher priority for clinicians to identify potential patients with preeclampsia. The fact that maximum diastolic BP had the highest HR in the full model confirms the importance of monitoring BP as early as possible, even before preeclampsia is diagnosed clinically [29]. More importantly, it identifies additional alarming factors to be considered in predicting preeclampsia diagnosis at gestational age, such as mood and anxiety disorder.

Further risk stratification of the survival models had slightly low specificity values in predicting the dichotomous diagnosis of preeclampsia at 34 and 37 weeks, suggesting that the continuous risk diagnosis has overall better performance compared with the simple binary prediction. However, the stratification may offer an easier way to identify women who may benefit more significantly from prevention therapy and need more medical attention from doctors for the possibility of preeclampsia. EHR-based models can serve as a screening test. For the patients that are potentially false positive for preeclampsia due to the lower specificity of the model, additional confirmative diagnostic tests using very specific biomarkers should be done, as practiced clinically.

Earlier studies using all pregnant women also revealed that mood and anxiety disorders increase the risk of preeclampsia [30]. We further show that within patients with preeclampsia, mood and anxiety disorders shorten the onset time of preeclampsia. This provides more context for clinicians to identify pregnant patients who present mood and anxiety disorders and provide preventative care to reduce preeclampsia onset risk. The molecular mechanism linking mood and anxiety disorders with preeclampsia is worth further research. We also show that NSAID use is positively associated with earlier onset of preeclampsia. However, aspirin is a common NSAID used by pregnant women at risk for preeclampsia early in pregnancy [31]. It was suggested that NSAID use may serve as a proxy for the interaction of many unmeasured risk factors [32]. Thus, the positive association of NSAID to the earlier onset of preeclampsia may indicate that it is a marker of high-risk preeclampsia in the population, rather than the cause of it.

Strengths and Limitations

A particular strength of the models here is their simplicity despite being quantitative. The models can also be generalized to different medical centers and hospitals, given the good accuracy when validated by vastly different institutions with different protocols, data collection, and data storage. There is a growing need for evidence-based and effective tools for clinicians to screen women at high risk of preeclampsia early in pregnancy, in the first and early second trimesters. This model supplies this need for early prediction models that previous models have not been able to fulfill [33]. Most clinical models recently published include many predictors from biomarkers and ultrasound markers that need special procedures [34], further suggesting that a simpler model on routinely collected clinical data is valuable to be implemented in a clinical setting. The main strength of this modeling for clinical use proposed here is providing more context in screening patients at risk for preeclampsia.

Our ultimate goal is to implement these models in the health care system, for example, starting from the University of Michigan. Potential challenges for implementing these models in a clinical setting include institutional buy-in, installation of the software in a HIPAA (Health Insurance Portability and Accountability Act)-compliant computing environment, and explaining the meaning of risk factors and model results to patients informatively without overly stressing them. In addition, these models may potentially require more active updating for improving accuracy, by considering additional multicenter data. Also, the current Cox-PH model is not designed to include longitudinal observations, limiting the kind of input variables to be incorporated into the model. Future work may benefit from more sophisticated modeling approaches [35]. Besides EHR, other omics information such as genetics, genomics, proteomics, and metabolomics using maternal blood samples [34] may be used, if they are available, to improve the model performance. However, implementing multimodal and complex models like this in the clinical setting is additionally challenging and would require more advanced modeling that can calculate individual risk scores for clinical application. It is also important to note the use of EHR data to extract medication prescriptions does not accurately capture the actual use or adherence of the medication by patients, and future research could be strengthened by combining data sources that provide such information.

Conclusions

In conclusion, this study reports prognosis models to predict the onset gestational age of preeclampsia with EMR data before the first 20 weeks of pregnancy. They identify clinical and physiological factors that clinicians should monitor as indicators of early preeclampsia development.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank UM Precision Health for providing technical support for data extraction in this study, the UF Integrated Data Repository, and the UF Health Office of the Chief Data Office for providing the analytic data set for this project. DJL was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (K01DK115632) and the UF Clinical and Translational Science Institute (UL1TR001427). LXG was supported by grants (K01ES025434) awarded through funds provided by the trans-National Institutes of Health Big Data to Knowledge initiative (R01 LM012373 and LM012907 awarded by the National Library of Medicine, and R01 HD084633 awarded by National Institute of Child Health and Human Development). ADM is supported by the National Center for Advancing Translational Science (5TL1TR001428). No funding sources listed were involved in the study design, collection, analysis, and interpretation of data, writing of the report, or decision to submit for publication.

Data Availability

The data sets generated during and/or analyzed during this study are not publicly available due to the presence of patient-protected health information. Data are available upon reasonable request and must be submitted on an individual basis to the home institution. Table S2 in Multimedia Appendix 1 lists all the EHR features extracted from the UM system that were considered in the starting model.

Authors' Contributions

LXG conceived this project and supervised the study. HKB conducted data analysis and wrote the manuscript. XY collaborated on data extraction of the University of Michigan cohort. ADM and DJL collaborated on validation using the University of Florida cohort. ADM provided clinical assessments and assistance. All authors have read, revised, and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 17 KB - jmir_v26i1e48997_app1.docx]

Multimedia Appendix 2
Lambdas from Least Absolute Shrinkage and Selection Operator (LASSO) regularization from the baseline and full preeclampsia (PE) prediction models. (A) Scatterplot of tested lambda values and associated errors from baseline model LASSO regularization. (B) Scatterplot of tested lambda values and associated errors from full model LASSO regularization.
[PNG File , 14 KB - jmir_v26i1e48997_app2.png]

Multimedia Appendix 3
AUC values of preeclampsia (PE) diagnosed at 34 and 37 weeks for the baseline and full PE prediction models. (A) Plot of the sensitivity, specificity, and area under the curve (AUC) values for the baseline model of predicting PE diagnosed at 34 weeks (red, AUC=0.654) and 37 weeks (green, AUC=0.694) for the testing dataset. (B) Plot of the sensitivity, specificity, and AUC values for the full model of predicting PE diagnosed at 34 weeks (red, AUC=0.697) and 37 weeks (green, AUC=0.700) for the testing data set.
[PNG File , 17 KB - jmir_v26i1e48997_app3.png]

References

1. Young BC, Levine RJ, Karumanchi SA. Pathogenesis of preeclampsia. Annu Rev Pathol Mech Dis 2010;5(1):173-192. [doi: 10.1146/annurev-pathol-121808-102149]
2. Al-Jameil N, Aziz Khan F, Fareed Khan M, Tabassum H. A brief overview of preeclampsia. J Clin Med Res 2014;6(1):1-7. [doi: 10.4021/jocmr1682w] [Medline: 24400024]
3. Chappell LC, Duckworth S, Seed PT, Griffin M, Myers J, Mackillop L, et al. Diagnostic accuracy of placental growth factor in women with suspected preeclampsia: a prospective multicenter study. Circulation 2013;128(19):2121-2131. [doi: 10.1161/CIRCULATIONAHA.113.003215] [Medline: 24190934]
4. E. G, Akurati L, Radhika K. Early onset and late onset preeclampsia-maternal and perinatal outcomes in a rural teritiary health center. Int J Reprod Contracept Obstet Gynecol 2018;7(6):2266-2269. [doi: 10.18203/2320-1770.ijrcog20182333]
5. Wainstock T, Sergienko R, Sheiner E. Who is at risk for preeclampsia? Risk factors for developing initial preeclampsia in a subsequent pregnancy. J Clin Med 2020;9(4):1103 [FREE Full text] [doi: 10.3390/jcm9041103] [Medline: 32294887]

XSL•FO
RenderX

6. Haile DB, Aguade AE, Fetene MZ. Joint modeling of hypertension measurements and time-to-onset of preeclampsia among pregnant women attending antenatal care service at arerti primary hospital, North Shoa, Ethiopia. Cogent Public Health 2022;9(1):202284. [doi: 10.1080/2331205x.2021.2022846]
7. Yang X, Ballard H, Mahadevan A. Deep learning-based prognosis prediction among preeclamptic pregnancies using electronic health record data. medRxiv 2022. [doi: 10.1101/2022.04.03.22273366]
8. Elixhauser comorbidity software refined for ICD-10-CM. URL: https://hcup-us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity_icd10.jsp [accessed 2023-02-07]
9. Bernard N, Forest J, Tarabulsy GM, Bujold E, Bouvier D, Giguère Y. Use of antidepressants and anxiolytics in early pregnancy and the risk of preeclampsia and gestational hypertension: a prospective study. BMC Pregnancy Childbirth 2019;19(1):146 [FREE Full text] [doi: 10.1186/s12884-019-2285-8] [Medline: 31039756]
10. Multivariate imputation by chained equations (mice) package. RDocumentation. URL: https://www.rdocumentation.org/packages/mice/versions/3.15.0 [accessed 2023-02-07]
11. Mera-Gaona M, Neumann U, Vargas-Canas R, López DM. Evaluating the impact of multivariate imputation by mice in feature selection. PLoS One 2021;16(12):e0261739 [FREE Full text] [doi: 10.1371/journal.pone.0261739] [Medline: 34914794]
12. Getz K, Hubbard RA, Linn KA. Performance of multiple imputation using modern machine learning methods in electronic health records data. Epidemiology 2023;34(2):206-215. [doi: 10.1097/EDE.0000000000001578] [Medline: 36722803]
13. Giorgi R, Belot A, Gaudart J, Launoy G, French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Stat Med 2008;27(30):6310-6331. [doi: 10.1002/sim.3476] [Medline: 19021241]
14. Dong Y, Peng CJ. Principled missing data methods for researchers. Springerplus 2013;2(1):222 [FREE Full text] [doi: 10.1186/2193-1801-2-222] [Medline: 23853744]
15. R: The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2023-02-07]
16. dplyr package. RDocumentation. URL: https://www.rdocumentation.org/packages/dplyr/versions/1.0.10 [accessed 2023-02-07]
17. gtsummary package. RDocumentation. URL: https://www.rdocumentation.org/packages/gtsummary/versions/1.6.3 [accessed 2023-02-07]
18. glmnet package. RDocumentation. URL: https://www.rdocumentation.org/packages/glmnet/versions/4.1-8/topics/glmnet [accessed 2023-02-07]
19. Friedman J, Hastie T, Tibshirani R. glmnet: LASSO and Elastic-Net regularized generalized linear models. 2022. URL: https://CRAN.R-project.org/package=glmnet [accessed 2023-02-07]
20. cindex function. RDocumentation. URL: https://www.rdocumentation.org/packages/pec/versions/2022.05.04/topics/cindex [accessed 2023-02-09]
21. pROC package. RDocumentation. URL: https://www.rdocumentation.org/packages/pROC/versions/1.18.5 [accessed 2023-12-17]
22. Kelly PA, Johnson ML. Just-in-Time IRB review: capitalizing on scientific merit review to improve human subjects research compliance. IRB: Ethics and Human Research 2005;27(2):6-10. [doi: 10.2307/3564480]
23. Ness RB, Roberts JM. Heterogeneous causes constituting the single syndrome of preeclampsia: a hypothesis and its implications. Am J Obstet Gynecol 1996;175(5):1365-1370. [doi: 10.1016/s0002-9378(96)70056-x]
24. English FA, Kenny LC, McCarthy F. Risk factors and effective management of preeclampsia. Integr Blood Press Control 2015;8:7-12 [FREE Full text] [doi: 10.2147/IBPC.S50641] [Medline: 25767405]
25. Paré E, Parry S, McElrath TF, Pucci D, Newton A, Lim KH. Clinical risk factors for preeclampsia in the 21st century. Obstet Gynecol 2014;124(4):763-770. [doi: 10.1097/AOG.0000000000000451] [Medline: 25198274]
26. Allahyari E, Foroushani A, Zeraati H, Mohammad K, Taghizadeh Z. A predictive model for the diagnosis of preeclampsia. J Reprod Infertil 2010;10(4):329 [FREE Full text]
27. De Kat AC, Hirst J, Woodward M, Kennedy S, Peters SA. Prediction models for preeclampsia: a systematic review. Pregnancy Hypertension 2019;16:48-66. [doi: 10.1016/j.preghy.2019.03.005]
28. von Dadelszen P, Magee LA, Roberts JM. Subclassification of preeclampsia. Hypertens Pregnancy 2003;22(2):143-148. [doi: 10.1081/PRG-120021060] [Medline: 12908998]
29. Hurrell A, Duhig K, Vandermolen B, Shennan AH. Recent advances in the diagnosis and management of pre-eclampsia. Fac Rev 2020;9:10 [FREE Full text] [doi: 10.12703/b/9-10] [Medline: 33659942]
30. Bullarbo M, Rylander R. Diastolic blood pressure increase is a risk indicator for pre-eclampsia. Arch Gynecol Obstet 2015;291(4):819-823. [doi: 10.1007/s00404-014-3476-1] [Medline: 25241271]
31. Qiu C, Williams MA, Calderon-Margalit R, Cripe SM, Sorensen TK. Preeclampsia risk in relation to maternal mood and anxiety disorders diagnosed before or during early pregnancy. Am J Hypertens 2009;22(4):397-402. [doi: 10.1038/ajh.2008.366] [Medline: 19197246]
32. LeFevre ML, U.S. Preventive Services Task Force. Low-dose aspirin use for the prevention of morbidity and mortality from preeclampsia: U.S. preventive services task force recommendation statement. Ann Intern Med 2014;161(11):819-826 [FREE Full text] [doi: 10.7326/M14-1884] [Medline: 25200125]

XSL•FO
RenderX

33. Low-dose aspirin use for the prevention of preeclampsia and related morbidity and mortality. URL: https://www.acog.org/en/clinical/clinical-guidance/practice-advisory/articles/2021/12/low-dose-aspirin-use-for-the-prevention-of-preeclampsia-and-related-morbidity-and-mortality [accessed 2023-03-30]
34. Benny PA, Alakwaa FM, Schlueter RJ, Lassiter CB, Garmire LX. A review of omics approaches to study preeclampsia. Placenta 2020;92:17-27 [FREE Full text] [doi: 10.1016/j.placenta.2020.01.008] [Medline: 32056783]
35. Tarca AL, Romero R, Benshalom-Tirosh N, Than NG, Gudicha DW, Done B, et al. The prediction of early preeclampsia: results from a longitudinal proteomics study. PLoS One 2019;14(6):e0217273 [FREE Full text] [doi: 10.1371/journal.pone.0217273] [Medline: 31163045]

Abbreviations

AUC: area under the curve
BP: blood pressure
C-index: concordance index
EHR: electronic health record
EMR: electronic medical record
HIPAA: Health Insurance Portability and Accountability Act
HR: hazard ratio
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
IRB: institutional review board
LASSO: Least Absolute Shrinkage and Selection Operator
NSAID: nonsteroidal anti-inflammatory drug
PI: prognosis index
UF: University of Florida
UM: University of Michigan

Edited by A Mavragani; submitted 15.05.23; peer-reviewed by S Nagavally, D Heider, B Puladi; comments to author 22.11.23; revised version received 17.01.24; accepted 30.05.24; published 14.08.24.

Please cite as:
Ballard HK, Yang X, Mahadevan AD, Lemas DJ, Garmire LX
Five-Feature Models to Predict Preeclampsia Onset Time From Electronic Health Record Data: Development and Validation Study
J Med Internet Res 2024;26:e48997
URL: https://www.jmir.org/2024/1/e48997
doi:10.2196/48997
PMID:

©Hailey K Ballard, Xiaotong Yang, Aditya D Mahadevan, Dominick J Lemas, Lana X Garmire. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study

Isabel Straw[1], BMedSci, BMBS, MPH, MRES; Geraint Rees[1], BMBCh, PhD; Parashkev Nachev[1], BMBCh, PhD

University College London, London, United Kingdom

Corresponding Author:
Isabel Straw, BMedSci, BMBS, MPH, MRES
University College London
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 020 3549 5969
Email: isabelstraw@doctors.org.uk

Abstract

Background: The presence of bias in artificial intelligence has garnered increased attention, with inequities in algorithmic performance being exposed across the fields of criminal justice, education, and welfare services. In health care, the inequitable performance of algorithms across demographic groups may widen health inequalities.

Objective: Here, we identify and characterize bias in cardiology algorithms, looking specifically at algorithms used in the management of heart failure.

Methods: Stage 1 involved a literature search of PubMed and Web of Science for key terms relating to cardiac machine learning (ML) algorithms. Papers that built ML models to predict cardiac disease were evaluated for their focus on demographic bias in model performance, and open-source data sets were retained for our investigation. Two open-source data sets were identified: (1) the University of California Irvine Heart Failure data set and (2) the University of California Irvine Coronary Artery Disease data set. We reproduced existing algorithms that have been reported for these data sets, tested them for sex biases in algorithm performance, and assessed a range of remediation techniques for their efficacy in reducing inequities. Particular attention was paid to the false negative rate (FNR), due to the clinical significance of underdiagnosis and missed opportunities for treatment.

Results: In stage 1, our literature search returned 127 papers, with 60 meeting the criteria for a full review and only 3 papers highlighting sex differences in algorithm performance. In the papers that reported sex, there was a consistent underrepresentation of female patients in the data sets. No papers investigated racial or ethnic differences. In stage 2, we reproduced algorithms reported in the literature, achieving mean accuracies of 84.24% (SD 3.51%) for data set 1 and 85.72% (SD 1.75%) for data set 2 (random forest models). For data set 1, the FNR was significantly higher for female patients in 13 out of 16 experiments, meeting the threshold of statistical significance (–17.81% to –3.37%; $P<.05$). A smaller disparity in the false positive rate was significant for male patients in 13 out of 16 experiments (–0.48% to +9.77%; $P<.05$). We observed an overprediction of disease for male patients (higher false positive rate) and an underprediction of disease for female patients (higher FNR). Sex differences in feature importance suggest that feature selection needs to be demographically tailored.

Conclusions: Our research exposes a significant gap in cardiac ML research, highlighting that the underperformance of algorithms for female patients has been overlooked in the published literature. Our study quantifies sex disparities in algorithmic performance and explores several sources of bias. We found an underrepresentation of female patients in the data sets used to train algorithms, identified sex biases in model error rates, and demonstrated that a series of remediation techniques were unable to address the inequities present.

(J Med Internet Res 2024;26:e46936) doi:10.2196/46936

KEYWORDS

artificial intelligence; machine learning; cardiology; health care; health equity; medicine; cardiac; quantitative evaluation; inequality; cardiac disease; performance; sex; management; heart failure

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) has been proposed as an effective solution to many health care challenges and depends on the construction of machine learning (ML) algorithms from health care data. Recent research has drawn attention to the possibility that algorithms may exhibit bias when applied to different demographic groups [1-6]. Such biases may widen health inequalities and negatively impact marginalized patients, such as female patients, minoritized racial and ethnic groups, and other neglected subpopulations [1-7].

Over the past 5 years, an increasing number of studies have quantified disparities in algorithmic performance for underserved populations [2-7]. Daneshjou and colleagues [2] demonstrated that state-of-the-art dermatology algorithms tend to perform worse on darker skin tones; Seyyed-Kalantari and colleagues [3] exposed biases in radiology algorithms; and Thompson and colleagues [4] reported increased false negative errors when classifying opioid misuse disorder for Black patients compared to White patients. Beyond specific diagnoses, researchers have demonstrated that infrastructural AI systems used in hospital settings can be subject to referral bias, demonstrated by Obermeyer and colleagues [5] who highlighted a hospital treatment allocation algorithm that overlooked the health needs of Black patients. Yet despite the increasing number of papers describing this issue, most of the current uses of biomedical AI technologies do not account for the problem of bias [5-8]. Here, we evaluate algorithmic inequity in ML algorithms used for predicting cardiac disease, focusing on heart failure (HF).

ML for HF

HF is a clinical syndrome in which the heart is unable to maintain a cardiac output adequate to meet the metabolic demands of the body [9]. Traditionally, algorithmic tools capable of identifying at-risk patients have played a key role in informing decisions on HF management and end-of-life care [10-12]. In recent years, ML algorithms that leverage biochemical data have been proposed as a superior alternative to traditional statistical models for identifying at-risk patients with HF [13]. A range of ML techniques outperforms traditional risk scores in forecasting HF-related events [13]. Yet given that existing medical research has described sex differences in both the presentation and management of HF, algorithms trained on existing data may perform differently for male versus female patients [14,15].

Sex Differences in HF

HF presents differently in female patients compared with male patients [14]. Female patients experience a wider range of symptoms, including higher fluid overload and lower health-related quality of life [14,15]. Moreover, female patients who present with HF are on average older, sustain a higher ejection fraction (EF) throughout later stages of the disease, and have a lower incidence of previous ischemic heart disease [15]. Furthermore, the biochemical tests used to detect cardiac disease have been demonstrated to perform less well for female patients [16]. Troponin is 1 key biomarker used to predict disease, which has been demonstrated to be less sensitive in female patients [16]. Standard troponin criteria fail to detect 1 out of 5 acute myocardial infarcts occurring in female patients [16]. Historically, the neglect of sex differences in cardiac pathophysiology has disadvantaged female patients, and if not considered during ML development, these inequities may manifest in the novel algorithms being integrated into cardiac care [14-19].

In our research, we scope the published literature reporting algorithms that predict HF and investigate whether existing papers give attention to bias in ML algorithms. Furthermore, we examine the data sets of existing models for demographic representation, evaluate demographic inequities in algorithmic performance, and assess the efficacy of a series of bias-mitigation techniques.

Methods

Study Design

Our analysis consists of two stages: (1) a literature review of papers describing ML models used to predict HF and (2) a quantitative analysis of identified models, evaluating inequities in algorithm performance. The flowchart in Figure 1 provides an overview of our approach.

XSL•FO
RenderX

Figure 1. A flowchart detailing the steps of our methodology, including (1) the initial literature search and qualitative evaluation of identified studies and (2) the identification of data sets and interrogation of algorithms for demographic bias. FAGTB: Fair Adversarial Gradient Tree Boosting; HF: heart failure; ML: machine learning; UCI: University of California Irvine.

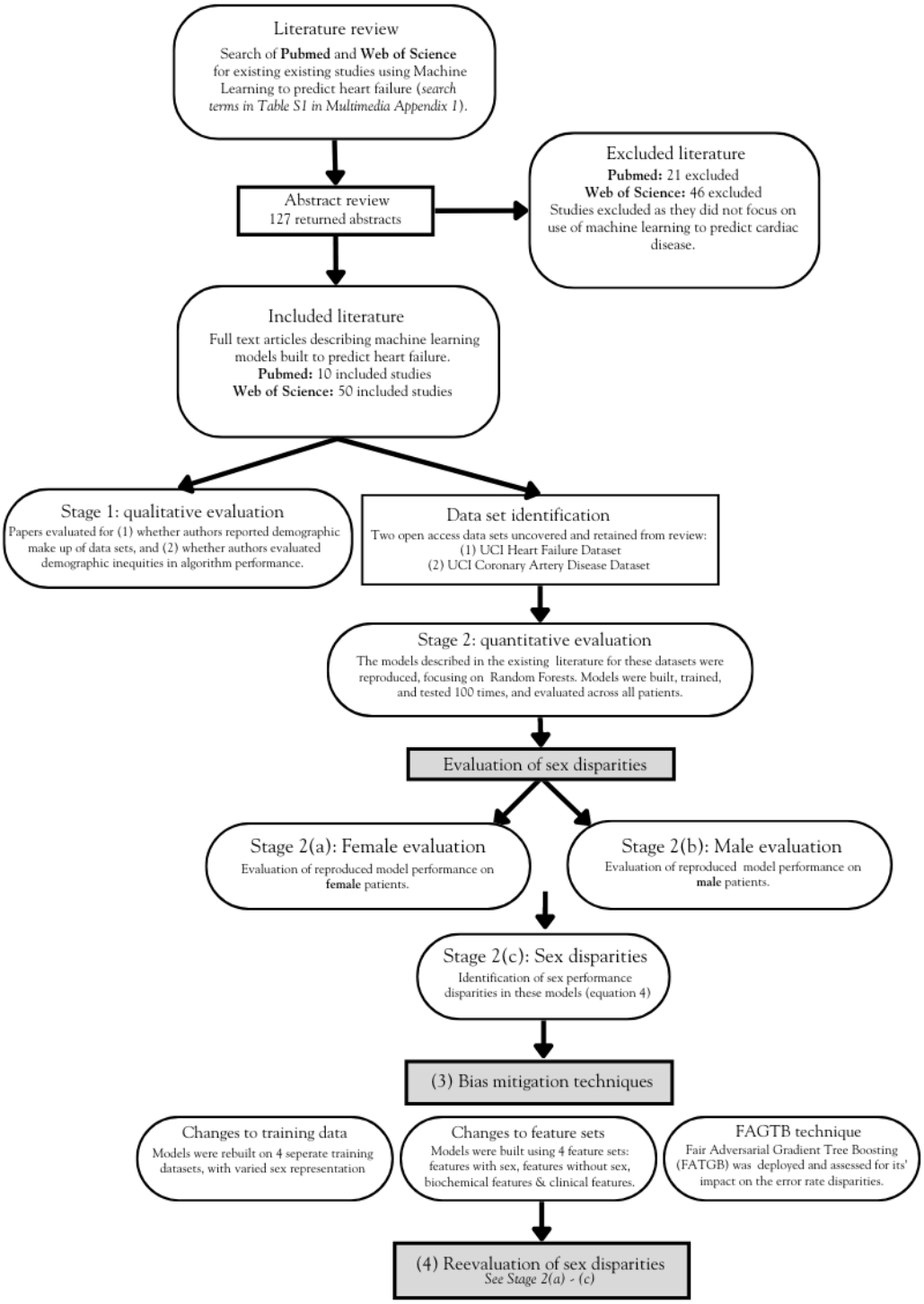

Stage 1 Literature Review: Qualitative Evaluation of Published Papers

We searched PubMed and Web of Science between April 1, 2022, and May 22, 2022, to identify ML algorithms used to predict cardiac disease adhering to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for systematic reviews (Figure 2 [20] and Tables S1 and S2 in Multimedia Appendix 1 [21,22]). All abstracts were reviewed, and papers were included for full-text review if they met the following criteria: (1) the target diagnosis was HF, (2) the model used biochemical markers to predict disease, and (3) the computational methods involved an ML approach (including supervised, unsupervised, and deep learning).

XSL•FO
RenderX

Figure 2. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only (PRISMA templated obtained from PRISMA at https://prisma-statement.org/prismastatement/flowdiagram.aspx) **Reasons for exclusion: Reason 1: The study did not focus on biochemical data or laboratory tests, instead utilizing different modalities (eg, visual data from radiological scans); Reason 2: The study did not use machine learning techniques (eg, it used traditional statistical methods); Reason 3: The study did not describe empirical research, involving the development of ML models for prediction of cardiac disease (eg, instead the paper was a review or commentary); Reason 4: The retrieved study was not a full paper, instead it was a conference or meeting abstract.

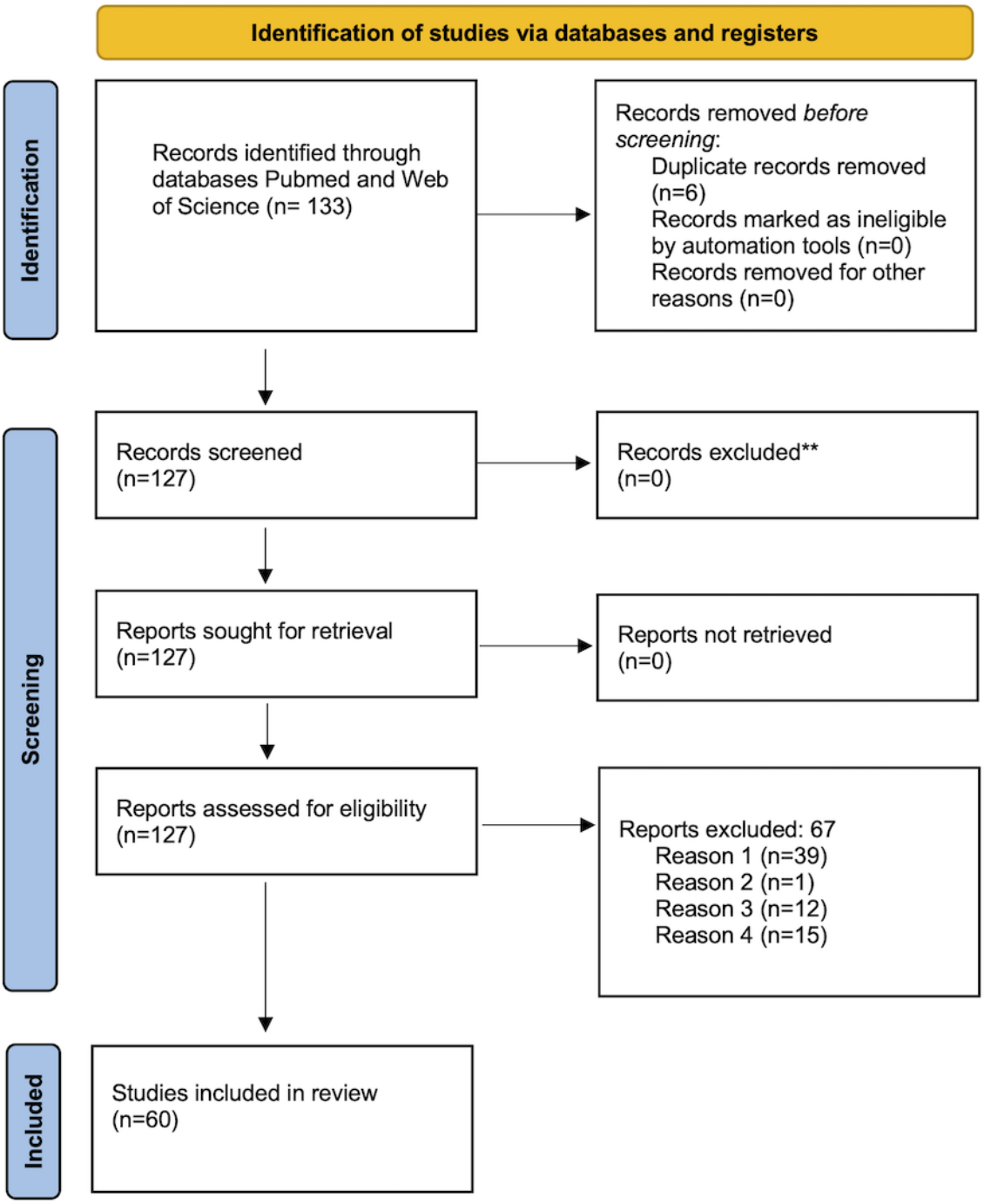

Of the retained papers, full texts were then reviewed to evaluate whether authors (1) reported the demographic make-up of data sets and (2) evaluated demographic inequities in algorithm performance, meaning that the authors specifically examined differences in algorithmic performance by demographic groups defined by protected characteristics [17].

Throughout the literature review, any identified open-source data sets were maintained for use in stage 2.

XSL•FO
RenderX

Stage 2: Quantitative Evaluation of Model Performance

Two open-source data sets were uncovered in our literature review: (1) data set 1: University of California Irvine for Heart Failure Prediction [21] and (2) data set 2: University of California Irvine Cleveland Heart Disease data set for identifying coronary artery disease (CAD) [22]. Descriptive statistics were performed on both data sets, evaluating the mean and variance of the data set variables for sexes separately, affected by disease or death (Table 1 and Tables S3-S5 in Multimedia Appendix 1).

Table 1. Descriptive statistics of the variables in data set 1 (heart failure; N=299), stratified by target (death) and sex[a].

Variables	Sex and death (target variable)[b]			
	Female (sex=0; n=105)		Male (sex=1; n=194)	
	Survived (HF[c] death=0)	Death (HF death=1)	Survived (HF death=0)	Death (HF death=1)
Total count, n (%)	71 (67.62)	34 (32.38)	132 (68.04)	62 (32.96)
Age (years), mean (SD)	58.6 (10.6)	62.2 (12.3)	58.8 (10.7)	66.9 (13.5)
Anemia (Boolean), mean (SD)	0.5 (0.5)	0.6 (0.5)	0.4 (0.5)	0.4 (0.5)
Creatinine phosphokinase (mcg/L), mean (SD)	462.0 (517.7)	507.7 (779.7)	582.8 (853.2)	759.3 (1532.3)
Diabetes mellitus (Boolean), mean (SD)	0.5 (0.5)	0.6 (0.5)	0.4 (0.5)	0.3 (0.5)
Ejection fraction (percentage), mean (SD)	41.9 (11.6)	37.5 (14.6)	39.4 (10.4)	31.2 (10.7)
High blood pressure (Boolean), mean (SD)	0.4 (0.5)	0.5 (0.5)	0.3 (0.5)	0.4 (0.5)
Platelets (kiloplatelets/mL), mean (SD)	289,757.6 (98,655.9)	259,512.7 (107,588.6)	254,232.4 (94,985.6)	254,663.7 (94,060.8)
Serum creatinine (mg/dL), mean (SD)	1.1 (0.6)	1.9 (1.6)	1.2 (0.7)	1.8 (1.4)
Serum sodium (mEq/L), mean (SD)	137.4 (3.6)	135.5 (6.7)	137.1 (4.2)	135.3 (3.8)
Smoking (Boolean), mean (SD)	0.0 (0.1)	0.1 (0.3)	0.5 (0.5)	0.4 (0.5)

[a]Full details of data set variables are available in Tanvir et al [21].

[b]For the death variable, a value of 1 indicates mortality.

[c]HF: heart failure.

Using these data sets, we rebuilt the ML algorithms described in the published literature and performed an additional analysis exploring inequities in algorithmic performance for demographic subgroups. As the only protected characteristic reported was sex, we focus on sex disparities in performance. Despite our initial aim to focus on HF, we retained an uncovered CAD data set to investigate whether trends identified for HF generalized to patients with CAD [22]. Tables S3 and S4 in Multimedia Appendix 1 provide details on data set 1 and data set 2, respectively.

Model Reproduction

We rebuilt the models described in the existing literature for these data sets, focusing on random forest (RF) algorithms, which have been widely reported to be the most effective models [23]. For both data sets, data was split into test or training subsets (0.7:0.3), RF models were built using SciKit Learn, and RF parameters were tuned using GridSearch CV (SciKit Learn). We adopted a bootstrapping approach to quantify uncertainty, such that models were built, trained, and tested 100 times, from which average results were derived with SD.

Statistical Analysis

Across the 100 runs, sex differences in each algorithm evaluation metric (equations 1-10) were calculated and averaged, with accompanying statistical tests performed to evaluate for statistical significance of any identified sex disparities. Our method for examining differences in algorithmic error rates builds on the foundational work from Buolamwini and Gebru [24], who demonstrated that a range of ML algorithms for facial recognition performed poorly on darker-skinned female patients. To evaluate for statistical significance, independent 2-tailed *t* tests were performed where the data was normally distributed, and Mann-Whitney *U* tests were performed where the data was not normally distributed. Kolmogorov-Smirnov tests were used to assess for normality [25].

Variations in Model Development

Overview

We then introduced a variety of changes to the model development, to evaluate the impact on the identified sex disparities in performance.

Changes to Model Training Data

In total, 1 widely proposed bias mitigation technique includes preprocessing the training data of a model to account for demographic representation, with previous research highlighting the benefit of training on demographically balanced or demographically stratified data sets [26]. We therefore created a range of data sets with varied sex representation and assessed for the impact on algorithm performance disparities. To form the sex-balanced data set, we used the oversampling function of *SMOTE()*, which has been proposed as an effective method for improving the representation of underserved populations in ML data sets [27]. The *SMOTE* package generates new minority

XSL•FO
RenderX

data points based on existing minority samples through linear interpolation [26,27]. Models were rebuilt as per the *Model Reproduction* section, using 4 different training data sets (sex-imbalanced, sex-balanced, and sex-specific; Tables S6 and S7 in Multimedia Appendix 1): (1) original sex-imbalanced training data, (2) sex-balanced training data, (3) female-only training data, and (4) male-only training data experiments.

Changes to Feature Selection

To understand why models make certain decisions, researchers in the domain of "explainable AI" have demonstrated how feature evaluation may provide important information regarding model performance for different subpopulations [26,28]. To do this, Shapley values have been widely accepted as a unified measure of feature importance since their proposal in 2017 [29].

In our experiments, we first perform an exploratory analysis, comparing feature importance for models trained on the male versus female data sets. Second, we create 4 feature subsets from the original data sets, to evaluate the impact of changing the feature selection on performance disparities. As described in the introduction, existing clinical research has described demographic differences in the biochemical and clinical markers of HF disease (eg, sex differences in EF and troponin levels) [16]. Thus, we delineate 4 different feature subsets that vary in this information, to examine whether certain feature subsets perform better for different demographic groups. These four feature subsets are described in detail in Tables S8 and S9 in Multimedia Appendix 1 and include (1) features with sex, (2) features without sex, (3) biochemical features, and (4) clinical features.

Our final series of experiments are therefore performed across the four training data sets (sex-imbalanced, sex-balanced, and sex-specific), and the four feature sets giving 16 total experiments: (1) original sex-imbalanced training data experiments (across four feature subsets), (2) sex-balanced training data experiments (across four feature subsets), (3) female training data experiments (across four feature subsets), and (4) male training data experiments (across four feature subsets)

Model Evaluation and Identification of Performance Disparities

Models are evaluated using global evaluation metrics (eg, accuracy) and specific error rates (eg, false negative rate [FNR]; equations 1-10). The difference between male and female scores is calculated to give a model's "sex performance disparity" (equation 10). To evaluate for statistical significance, Kolmogorov-Smirnov Tests were used to assess for the normality of the data, following which independent 2-tailed *t* tests were performed where the data were normally distributed, and Mann-Whitney *U* tests were performed where the data were not normally distributed.

Our choice of evaluation metrics is guided by the clinical consequence of each of these scores.

The existing research on algorithmic bias has highlighted the importance of examining error rates, particularly in medicine where a false negative clinically translates to missed diagnoses or opportunities for treatment [3-6,26]. As described by Afrose and colleagues [26], focusing on global metrics of performance such as area under the receiver operating characteristic curve scores can neglect subtler disparities arising from differences in error rates affecting subgroups. When selecting a bias assessment metric, previous studies have chosen to focus on FNR and false positive rate (FPR), due to the clinical implications of these errors [4,30,31]. Equations 5-8 places the error rates in their clinical context, demonstrating that the FNR represents missed diagnoses and potentially missed treatment. For the error rates, we use the threshold of 0.5, as we are investigating performance inequities in the existing reported models that used these default settings.

Error rate definitions are as follows:

Clinical implications of error rates are as follows:

True Positive Rate = Correct diagnosis that patient as disease **(5)**

False Positive Rate = Misdiagnosis of disease when patient is healthy **(6)**

True Negative Rate = Correct diagnosis that patient is healthy **(7)**

False Negative Rate = Misdiagnosis that patient is healthy when patient has disease **(8)**

The accuracy evaluation metric is calculated as follows:

Sex performance disparity is calculated as follows:

Sex performance disparity = Score for male patients (mean) – Score for female patients (mean) **(10)**

Fairness Techniques: Fair Adversarial Gradient Tree Boosting

We implemented a recent fairness technique to evaluate whether these approaches applied to bias in HF algorithms. The Fair Adversarial Gradient Tree Boosting (FAGTB) is a recent technique proposed by Grari et al [8] for mitigating bias in decision tree classifiers and the authors demonstrate the success of their technique on 4 data sets. The authors focus on 2 definitions of fairness: demographic parity and equalized odds [8]. The equalized odds metric focuses on model FPR and FNR, and hence we highlight this for our paper. A summary of these fairness metrics is provided in Section S1 in Multimedia Appendix 2 for further interest.

The definition of equalized odds is as follows:

XSL•FO
RenderX

To assess for the equalized odds the authors measure the disparate mistreatment, which computes the absolute difference between FPR and the FNR for both demographics.

The disparate FPR is calculated as follows:

The disparate FNR is calculated as follows:

We compare the performance of the FAGTB algorithm to a standard Gradient Tree Algorithm. As per the original FAGTB paper, we repeat 10 experiments randomly sampling 2 subsets (0.8:0.2) and report evaluation metrics for the test set.

Ethical Considerations

Ethical approval was not required for this study as all data used were sourced from publicly available open-source data sets [21,22] under a CC-BY 4.0 license. No direct patient contact or sensitive personal data was involved, ensuring compliance with research standards.

Results

Literature Review Search Results

Our search returned 127 papers, of which 60 met the criteria for full review and 3 highlighted sex differences in model performance. In the papers that reported sex, there was a consistent underrepresentation of female patients. No papers investigated racial or ethnic differences. Further, 1 paper focused specifically on female patients with HF, in which Tison et al [32] highlighted that HF was more common in people who were older, White, with a higher mean number of pregnancies, a higher BMI, and were less likely to have Medicare.

Descriptive Statistics and Feature Importance

Data Set 1 (HF)

The mean descriptive statistics for each feature present in the HF data set are provided in Table 1, which demonstrates subtle sex differences in the presentation of the disease. For HF deaths, male patients tend to be older than their female counterparts, with a higher creatinine phosphokinase, lower likelihood of diabetes, lower EF, and lower blood pressure.

Our exploratory analysis identified further sex differences on examining feature importance. Figure 3 compares the rankings of feature importance for ML models built to predict HF built from the female data set compared to the male data set. These differences are important as existing ML algorithms built on mixed-sex cohorts suggest that EF can be used alone for modeling, an approach that may disadvantage female patients [23].

Figure 3. Comparison of feature rankings for male and female patients, ordered by SHA*P* values. SHAP: Shapley additive explanations.

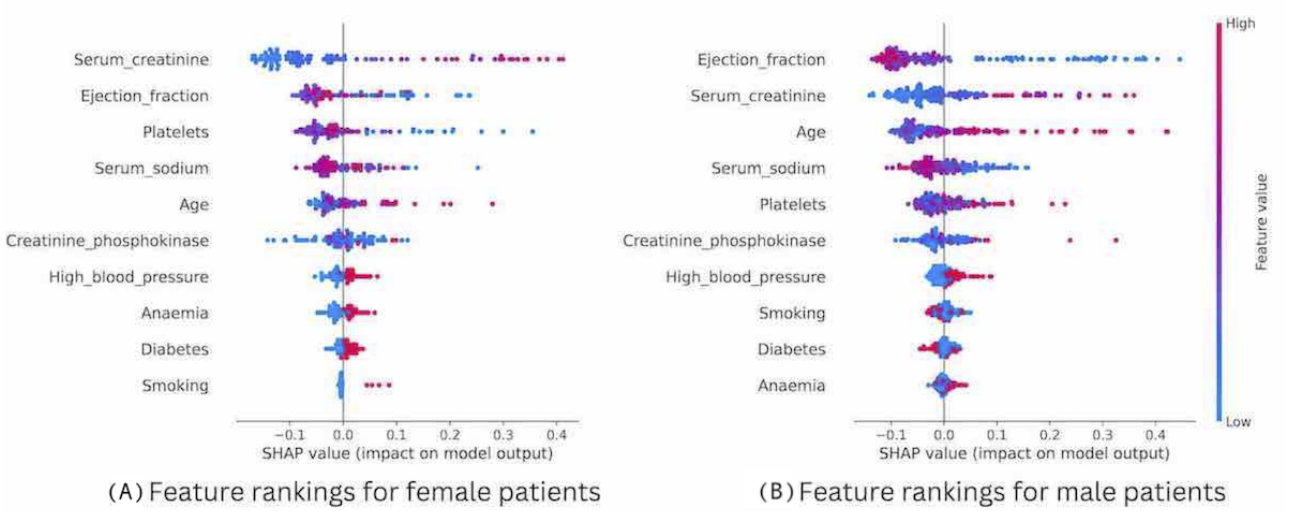

Data Set 2 (CAD)

Table S5 in Multimedia Appendix 1 provides details of the CAD data set and demonstrates that female patients with CAD have higher resting blood pressure and higher cholesterol compared to male patients. The categorical variable "resting electrocardiogram" is also higher for female patients, due to a higher incidence of left ventricular hypertrophy.

Model Results and Performance Disparities

We replicated the algorithms described in the existing literature, reproducing the same previously reported mean predictive accuracies of 84.24% (3.51 SD) for data set 1 and 85.72% (1.75 SD) for data set 2 [23]. In Tables 2 and 3, we present the disparity in performance for the sexes, where a positive value indicates a higher value for male patients (see equation 10).

For data set 1, Table 2 demonstrates that in 13 out of 16 experiments, the FNR is higher for female patients, meeting the threshold of statistical significance (mean difference of –17.81% to –3.37%; $P<.05$). Figure 4 represents this disparity in performance graphically, providing the point estimates of FNR for the sexes separately and highlighting that the disparity in FNR persisted across the variations in training data and selected features.

XSL•FO
RenderX

Table 2. Sex performance disparities for models built from data set 1 (heart failure disease)–sex performance disparities are calculated as the performance for male patients minus the performance for female patients (see equation 10). Thus, a positive value indicates a higher score for male patients and a negative value indicates a higher score for female patients. All disparities are presented alongside results of significance testing, where significant differences between the sexes are highlighted with a footnote (*P*<.05).

Disparity in model performance (score for male patients – score for female patients)	Feature subset used in model training							
	Features with sex	*P* value	Features without sex	*P* value	Biochemical features	*P* value	Clinical features	*P* value
Sex-imbalanced training data								
Accuracy disparity (%)	1.63	.03[a]	–0.72	.30	0.10	.88	–0.50	.49
ROC_AUC[b] disparity (%)	3.14	<.01[a]	0.43	.61	1.51	.09	0.47	.60
FNR[c] disparity (%)	–7.53	<.01[a]	–3.84	.02[a]	–5.15	.01[a]	–3.49	.049[a]
FPR[d] disparity (%)	1.26	.07	2.97	<.01[a]	2.11	<.01[a]	2.56	<.01[a]
Sex-balanced training data								
Accuracy disparity (%)	–4.78	<.01[a]	–7.25	<.01[a]	–9.42	<.01[a]	–3.63	<.01[a]
ROC_AUC disparity (%)	7.0	<.01[a]	4.27	<.01[a]	0.15	.83	8.32	<.01[a]
FNR disparity (%)	–17.81	<.01[a]	–13.91	<.01[a]	–3.37	.04[a]	–16.09	<.01[a]
FPR disparity (%)	3.90	<.01[a]	5.37	<.01[a]	3.07	<.001[a]	–0.54	.24
Female training data								
Accuracy disparity (%)	–10.95	<.01[a]	–9.75	<.01[a]	–12.32	<.01[a]	–9.64	<.01[a]
ROC_AUC disparity (%)	0.60	.57	0.57	.23	–2.92	<.01[a]	–0.53	.07
FNR disparity (%)	–7.42	<.01[a]	–10.91	<.01[a]	–2.24	.27	1.55	.01[a]
FPR disparity (%)	8.61	<.01[a]	9.77	<.01[a]	8.08	<.01[a]	–0.48	.04[a]
Male training data								
Accuracy disparity (%)	–5.46	<.01[a]	–5.73	<.01[a]	–8.73	<.01[a]	–2.46	<.01[a]
ROC_AUC disparity (%)	4.98	<.01[a]	4.54	<.01[a]	–1.59	.049[a]	8.32	<.01[a]
FNR disparity (%)	–13.96	<.01[a]	–13.32	<.01[a]	–1.68	.33	–16.58	<.01[a]
FPR disparity (%)	4.00	<.01[a]	4.24	<.01[a]	4.86	<.01[a]	–0.06	.35

[a]Indicates a statistically significant difference (*P*<.05) between the model's performance on male versus female patients.

[a]ROC_AUC: area under the receiver operating characteristic curve.

[b]FNR: false negative rate.

[c]FPR: false positive rate.

XSL•FO
RenderX

Table 3. Sex performance disparities for models built from data set 2 (coronary artery disease)—sex performance disparities are calculated as the performance for male patients minus the performance for female patients (see equation 10). Thus, a positive value indicates a higher score for male patients, and a negative value indicates a higher score for female patients. All disparities are presented alongside results of significance testing, where significant differences between the sexes are highlighted with a footnote (*P*<.05).

Disparity in model performance (score for male patients – score for female patients)	Feature subset used in model training							
	Features with sex	*P* value	Features without sex	*P* value	Biochemical features	*P* value	Clinical features	*P* value
Sex-imbalanced training data								
Accuracy disparity (%)	0.32	.50	0.64	.17	0.13	.80	0.25	.61
ROC_AUC[b] disparity (%)	3.86	<.01[a]	4.24	<.01[a]	3.05	<.01[a]	3.91	<.01[a]
FNR[c] disparity (%)	–11.66	<.01[a]	–12.52	<.01[a]	–10.81	<.01[a]	–12.38	<.01[a]
FPR[d] disparity (%)	3.94	<.01[a]	4.04	<.01[a]	4.71	<.01[a]	4.57	<.01[a]
Sex-balanced training data								
Accuracy disparity (%)	–4.01	<.01[a]	–5.12	<.01[a]	–7.32	<.01[a]	–2.86	<.01[a]
ROC_AUC disparity (%)	–3.89	.01[a]	–4.91	.01[a]	–7.18	<.001[a]	–2.75	<.01[a]
FNR disparity (%)	7.69	<.01[a]	10.54	<.01[a]	15.59	<.01[a]	6.61	<.01[a]
FPR disparity (%)	0.10	.87	–0.72	.19	–1.23	.29	–1.11	.06
Female training data								
Accuracy disparity (%)	–9.25	<.01[a]	–11.34	<.01[a]	–11.49	<.01[a]	–8.69	<.01[a]
ROC_AUC disparity (%)	–8.97	<.01[a]	–10.95	<.01[a]	–11.10	<.01[a]	–8.45	<.01[a]
FNR disparity (%)	18.98	<.01[a]	22.60	<.01[a]	27.23	<.01[a]	17.86	<.01[a]
FPR disparity (%)	–1.04	.07	–0.70	.20	–5.02	<.01[a]	–0.96	.09
Male training data								
Accuracy disparity (%)	6.38	<.01[a]	5.66	<.01[a]	–1.66	.02[a]	6.10	<.01[a]
ROC_AUC disparity (%)	6.30	<.01[a]	5.57	<.01[a]	1.52	.07	5.86	.01[a]
FNR disparity (%)	–10.12	<.01[a]	–10.10	<.001[a]	1.67	.17	–12.64	<.01[a]
FPR disparity (%)	–2.48	<.01[a]	–1.04	.07	1.38	.24	0.92	.15

[a]Indicates a statistically significant difference (*P*<.05) between the model's performance on male versus female patients. To determine statistical significance, the Kolmogorov-Smirnov tests were first run on the sex-stratified results to determine the distribution of data (normal or not). Independent 2-tailed *t* tests were used where data were normally distributed, and Mann-Whitney *U* tests were used when data were not normally distributed.

[b]ROC_AUC: area under the receiver operating characteristic curve.

[c]FNR: false negative rate.

[d]FPR: false positive rate.

A smaller disparity in the FPR was statistically significant for male patients in 13 out of 16 experiments (–0.48% to +9.77%; *P*<.05). The sex performance disparities in accuracy and area under the receiver operating characteristic curve varied depending on the underlying shifts in the error rates for each sex (Table 2 and Figure 5). On examining the individual error rates, we see consistencies in the sex disparities across feature sets, most notably an overprediction of disease for male patients (higher FPR) and an underprediction of disease for female patients (higher FNR: Table 2).

XSL•FO
RenderX

Figure 4. Data set 1 (heart failure): a series of violin plots showing the sex-stratified performance (false negative rate [0%-100%]) of the random forests trained across the 4 feature sets, on the different variations in training data. The plots show male (orange) and female (gray) FNR alongside each other, in groups of 4 (divided by a line) according to the training data used (sex-imbalanced, sex-balanced, female, and male). The feature set used is indicated within each training data group (features with sex, features without sex, biochemical features, and clinical features). See Multimedia Appendixes.

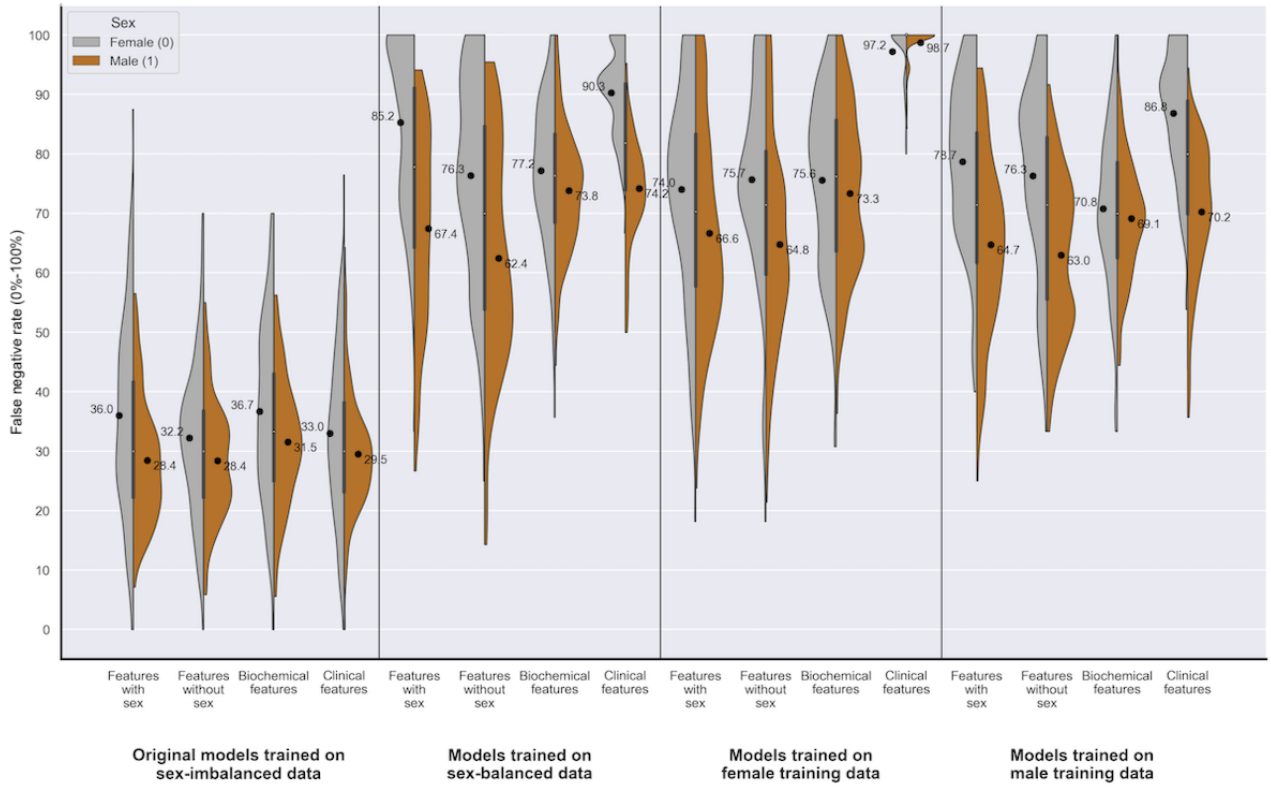

Figure 5. Data set 1 (heart failure): a series of violin plots showing the sex-stratified performance (accuracy [0%-100%]) of the random forests trained across the 4 feature sets, on the different variations in training data. The plots show male (orange) and female (gray) accuracy alongside each other, in groups of 4 (divided by a line) according to the training data used (sex-imbalanced, sex-balanced, female, and male). The feature set used is indicated within each training data group (features with sex, features without sex, biochemical features, and clinical features). See Multimedia Appendixes.

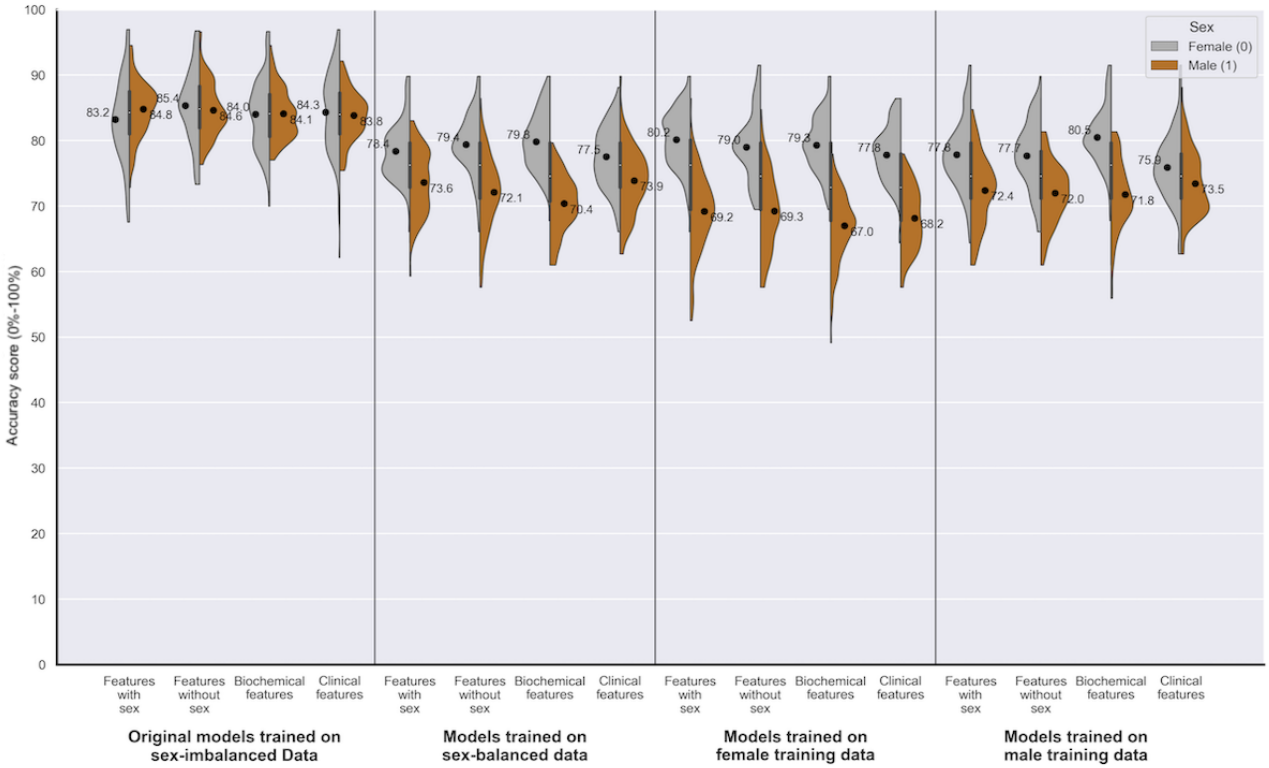

XSL•FO
RenderX

Our findings for data set 2 were similar to those for data set 1, such that models built on the original sex-imbalanced data set demonstrated a higher FNR for female patients (mean difference of –10.81% to –12.52%; $P<.05$; Table 3) and a higher FPR for male patients (3.94% to 4.71%; $P<.05$; Table 3). Figure 6 visualizes the disparity graphically, and demonstrates that, unlike data set 1, the disparity in error rates reversed when training on sex-balanced data and female-only data (Figure 6). Figure 7 illustrates the disparity in accuracy between the sexes, where we see that the direction of the disparity varies depending on the training data and feature set (Figure 7).

Figure 6. Data set 2 (coronary artery disease): a series of violin plots showing the sex-stratified performance (false negative rate [0%-100%]) of the random forests trained across the 4 feature sets, on the different variations in training data. The plots show male (orange) and female (gray) FNR alongside each other, in groups of 4 (divided by a line) according to the training data used (sex-imbalanced, sex-balanced, female, and male). The feature set used is indicated within each training data group (features with sex, features without sex, biochemical features, and clinical features). See Multimedia Appendixes. FNR: false negative rate.

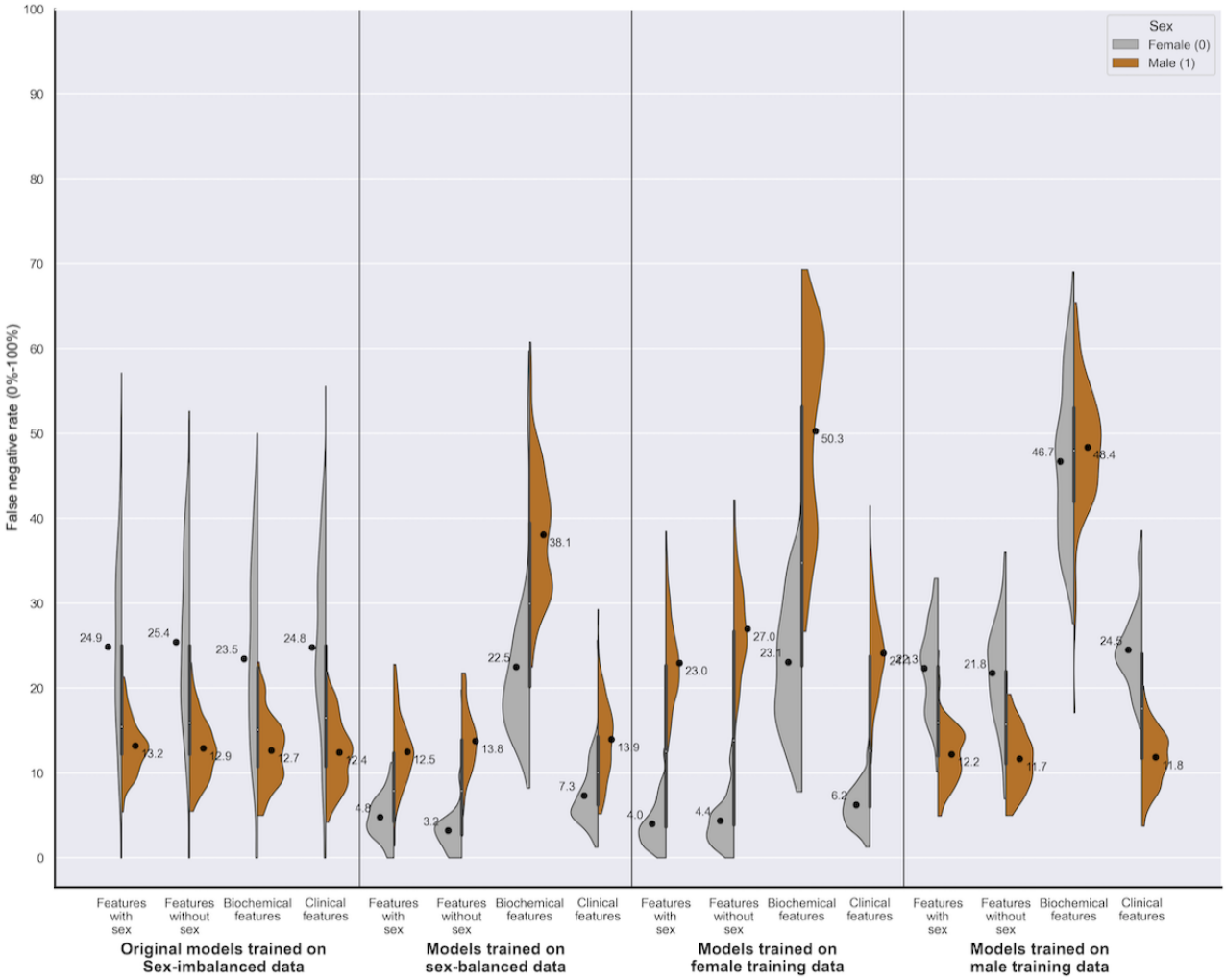

XSL•FO
RenderX

Figure 7. Data set 2 (coronary artery disease): a series of violin plots showing the sex-stratified performance (accuracy [0%-100%]) of the random forests trained across the 4 feature sets, on the variations in training data. The plots show male (orange) and female (gray) accuracy alongside each other, in groups of 4 (divided by a line) according to the training data used (sex-imbalanced, sex-balanced, female, and male). The feature set used is indicated within each training data group (features with sex, features without sex, biochemical features, and clinical features).

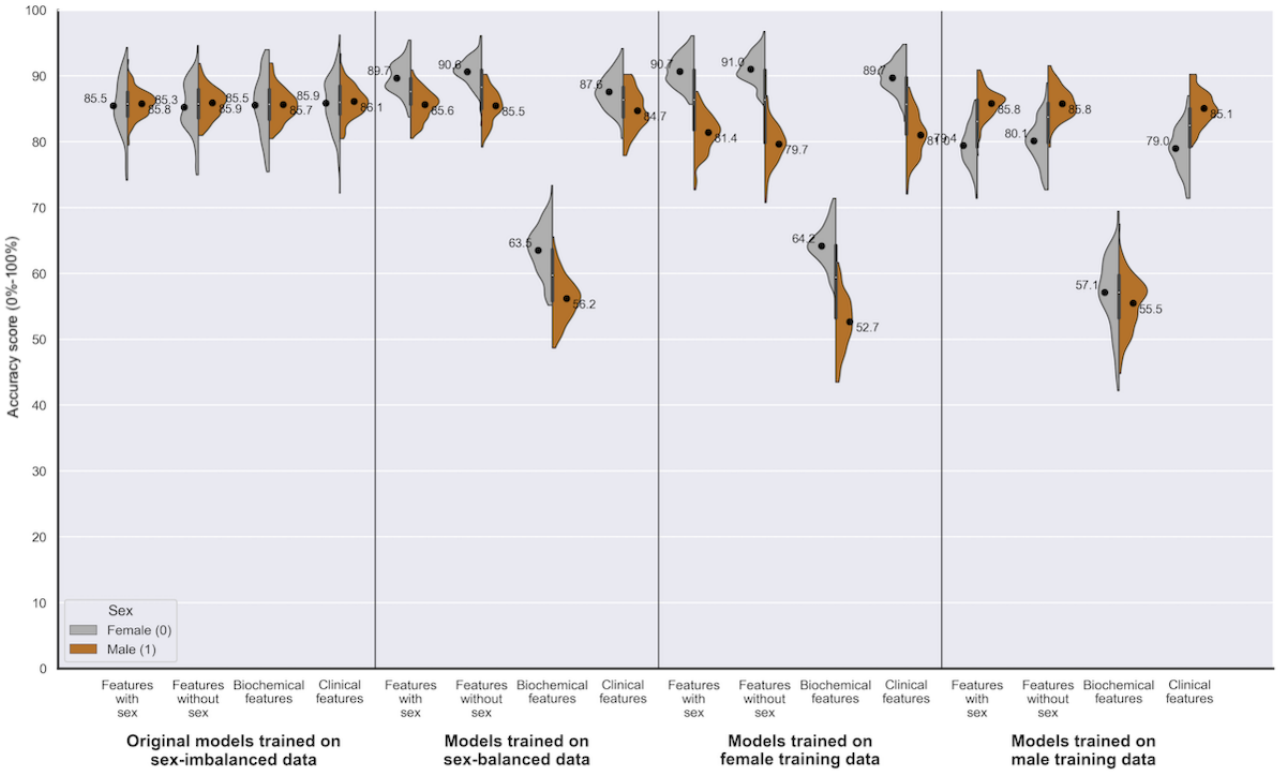

Variations in Training Data

Sex-Balanced Training Data

Training on sex-balanced data led to a fall in mean accuracy for all patients in data set 1 (76%, SD 3.46% vs 84.24%, SD 3.51%), with a more substantial drop in mean accuracy for male patients (73.61%, SD 4.84% vs 84.84%, SD 4.16%; Table 4 and Figure 5). The opposite trend was seen in data set 2, with models trained on sex-balanced data outperforming models trained on sex-imbalanced data for all patients (87.65%, SD 1.77% vs 85.72%, SD 1.75%) and for female patients (89.66%, SD 2.44% vs 85.48%, SD 4.12%; Table 4). The models trained on sex-balanced data in data set 2 reduced the FNR for both sexes when using the full feature set (female patients 4.79%, SD 2.58% vs 24.86%, SD 11.35%; male patients 12.48%, SD 4.11% vs 13.19%, SD 3.26%; Table 4 and Figure 6). The differences between the data sets may relate to underlying differences in the 2 cardiac conditions. Further, the failure to improve performance with sex-balanced training data may reflect the issues of mixing data that has conflicting indicators for disease.

XSL•FO
RenderX

Table 4. Model results when trained on sex-specific subsets for all patients and male or female patients separately, looking at the "features including sex" subset.

Results	Data set 1 (heart failure)				Data set 2 (coronary artery disease)			
	Sex-imbalanced training data (n=209)	Sex-balanced training data (n=272)	Female training data (n=136)	Male training data (n=136)	Sex-imbalanced training data (n=522)	Sex-balanced training data (n=715)	Female training data (n=358)	Male training data (n=358)
All patients, mean accuracy (SD)	84.24 (3.51)	76.0 (3.46)	74.68 (3.53)	75.12 (3.71)	85.72 (1.75)	87.65 (1.77)	86.06 (1.67)	82.63 (1.94)
Female patients, mean accuracy (SD)	83.21 (6.37)	78.39 (19.68)	80.15 (4.43)	77.85 (5.21)	85.48 (4.12)	89.66 (2.44)	90.69 (2.38)	79.44 (3.20)
Male patients, mean accuracy (SD)	84.84 (4.16)	73.61 (4.84)	69.20 (5.96)	72.39 (5.32)	85.80 (2.14)	85.65 (2.23)	81.44 (3.02)	85.82 (2.30).
Female patients, mean FNR[a] (SD)	35.98 (16.72)	85.25 (14.58)	74.04 (17.68)	78.66 (14.0)	24.86 (11.35)	4.79 (2.58)	4.00 (2.74)	22.32 (5.25)
Male patients, mean FNR (SD)	28.45 (10.41)	67.43 (16.6)	66.62 (17.32)	64.70 (14.9)	13.19 (3.26)	12.48 (4.11)	22.97 (5.20)	12.20 (3.41)

[a]FNR: false negative rate.

Sex-Specific Training Data

For data set 1, mean accuracy for all patients when trained on sex-imbalanced data (84.24%, SD 3.51%) falls when training both on female-specific data (74.68%, SD 3.53%) and male-specific training data (75.12%, SD 3.71%), likely related to the smaller training data. For data set 2, mean accuracy for all patients when trained on sex-imbalanced data (85.72%, SD 1.75%) improves when training on female-specific data (86.06%, SD 1.67%) and falls when training on male-specific training data (82.62%, SD 1.94%). The overall improvement seen in the data set 2 models when trained on female data, relates to the increase in accuracy for female patients (90.69%, SD 2.38% vs 85.48%, SD 4.12%) co-occurring with a smaller decrease in accuracy for male patients (81.44%, SD 3.02% vs 85.80%, SD 2.14%; Table 3 and Figure 7).

Unsurprisingly, performance for each sex is lowest when trained on the opposing sex (Table 4, Figures 4-7). In data set 1, same-sex training was preferable to opposite-sex training; however, this did not improve results compared to the models built from sex-imbalanced and sex-balanced training data, likely relating to the smaller sample size (Table 4). In contrast, data set 2 had greater training data available and demonstrated that sex-specific training is beneficial to both sexes above the sex-imbalanced models (Table 4).

Variations in Feature Sets

Models built on the biochemical features subset gave the worst performance in terms of accuracy and FNR (Figures 4-7). For data set 2, biochemical features included just cholesterol and fasting blood sugar, and so, the fall in performance may relate to information loss. Additionally, Table S5 in Multimedia Appendix 1 highlights the different biochemical profiles for male and female patients who were sick, with female patients who were sick demonstrating a far higher cholesterol level than their male counterparts (mean values: 279.2 female patients who were sick vs 247.5 male patients who were sick).

FAGTB Model

The disparity in false negative rate (DispFNR) was consistently higher than the disparity in false positive rate (Table 5). Compared to the Gradient Boosting Classifier, the FAGTB reduced the DispFNR for both data sets (data set 1: 0.20 vs 0.21; data set 2: 0.19 vs 0.28), however, the DispFNR that disadvantaged female patients persisted. The fall in DispFNR and disparity in false positive rate that occurred with FAGTB was associated with a fall in overall accuracy for both data sets.

XSL•FO
RenderX

Table 5. Results of bias mitigation with Fair Adversarial Gradient Tree Boosting (FAGTB).

Results on test set, averaged over 10 experiments	Gradient boosting classifier	FAGTB
Data set 1 (heart failure): experiments run on sex-imbalanced data with all features (averaged over 10 experiments)		
Accuracy	71.3	71.2
DispFPR[a]	0.08	0.08
DispFNR[b]	0.21	0.20
Data set 2 (coronary artery disease): experiments run on sex-imbalanced data with all features (averaged over 10 experiments)		
Accuracy	86.3	82.9
DispFPR	0.06	0.06
DispFNR	0.28	0.19

[a]DispFPR: disparity in false positive rate.

[b]DispFNR: disparity in false negative rate.

Discussion

Principal Findings

Our study sheds light on an important gap in existing cardiac ML research, with significant implications for digital health equity. We find that the majority of published ML studies predicting HF fail to acknowledge the underrepresentation of female patients in their data sets and do not perform stratified model evaluations, thus failing to assess sex disparities in algorithmic performance. Our secondary evaluation of 2 cardiac data sets exposed a neglected sex disparity in model performance, highlighting the importance of integrating these methods into future studies that use ML methods for cardiac modeling. In our approach, we identified several potential sources of algorithmic bias.

First, we detected the underrepresentation of female patients in training data sets that may produce inequalities in model fidelity. Despite introducing oversampling techniques to address this omission, the disparities in performance persisted suggesting that addressing data set representation alone is not a sufficient measure for mitigating bias. Further, our experiments demonstrated that oversampling could reduce overall performance, which may result from the mixing of conflicting data (ie, male vs female feature rankings). In addition, oversampling with synthetic instances solely from the data set at hand does not provide the machine with more information, it simply redirects attention and therefore cannot easily compensate for demographic underrepresentation [33]. When balancing the data set, our methods did not include undersampling due to our small data sets, however, this may be a potential avenue for future research.

Second, we considered featurization and highlighted sex differences in the biochemical manifestation of disease. In current clinical practice, the diagnostic parameters used for identifying pathology are drawn from research trials dominated by male physiology: it is perhaps unsurprising therefore that algorithms built from these data tend to underperform in female disease. There is a growing body of research that critiques the use of unisex thresholds in medicine for biochemical tests; our sex-stratified analysis of the cardiac data sets and the identified sex differences in feature rankings supports these proposals [16].

There are further sources of inequitable performance that our evaluation cannot distinguish between. It may be that the sex differences in the physiological expression of disease mean that the prediction is harder to extract from 1 population. As a result, 1 sex may require more complex models than another, with differing architecture and degrees of flexibility. It may also simply be that there are differences in the predictability of 1 group compared with another, such that if the physiology of 1 group is more opaque, it may ultimately not be possible to resolve the observed disparities. McCradden and colleagues [34] detail this challenge further in their review, highlighting that differences across groups may not always indicate inequity. There are complex causal relationships between biological, environmental, and social factors that underpin the differences in disease rates seen across population subgroups [34]. While models must not promote different standards of care according to protected characteristics, differences between groups may not necessarily reflect discriminatory practice [34].

Our research was limited by the available information in the data sets. The absence of race or ethnicity data precluded the evaluation of their effects. Furthermore, the absence of other demographic data in the studies we identified prevented the investigation of health inequities that might impact the LGBTQ+ (lesbian, gay, bisexual, transgender, queer) community, disadvantaged socioeconomic groups, or other subgroups. Previous research has described historic and institutional biases that contribute to worse health outcomes for these groups, and evolving AI systems require the same scrutiny to ensure these harms do not become embedded within digital systems [35-37].

Throughout this paper, we have used the terms male and female to reference biological sex, so as not to conflate sex and gender. With the ongoing problematic conflation of sex and gender in medicine, stratification of model performance by either sex or gender is often impossible, which was noted in our own work [35-37]. Beyond the features discussed above, there is a wide range of additional factors that we cannot account for. For example, creatinine phosphokinase was a key feature in HF modeling yet existing studies have demonstrated the variation

XSL•FO
RenderX

in these levels for manual laborers and athletes, illustrating how occupation may impact a patient's physiology [38].

To account for the complex interactions that potentiate disease, and the heterogeneous nature of patient cohorts, we require more complex modeling capable of capturing the full range of intersecting factors influencing patient health (eg, sex differences may be mediated by income). Unsupervised high-dimensional representation learning may be the path forward for this purpose [39]. In addition to improving representation, unsupervised techniques enable us to detect neglected subpopulations without predetermining a characteristic of interest, facilitating the identification of the previously overlooked disadvantaged. In this sense, AI may provide a route forward to uncovering and addressing bias, by deploying more complex modeling that can improve patient representation and by revealing previously neglected disparities in the provision of care.

Conclusions and Limitations

In our paper, we have identified inequities in the performance of cardiac ML algorithms. Our findings are limited by the small size of the uncovered data sets, reducing their potential generalizability, and hence we propose that larger studies focused on this issue are required. These data sets also came from the same source, as we found a limited number of open-access databases due to the confidential nature of patient data and issues of proprietary ownership. In addition, we focused on RF models to replicate the papers uncovered in our literature search; however, ML models may differ in their degrees of performance disparity, and an evaluation across the range of ML model options is an important next step.

In our paper we did not attempt to solve bias; instead, we highlighted a problem that exists throughout cardiology that requires further attention. The issue we have identified in these ML models is a foundational problem across medical modeling, in any instance where the use of an "average" is applied to a diverse population. It is possible that unsupervised ML and complex representational modeling may be a route forward for capturing heterogeneity in a previously unattainable manner and addressing issues of bias [39]. Our findings demonstrate that examining performance inequities across demographic subgroups is an essential approach for identifying biases in AI and preventing the perpetuation of inequalities in digital health systems.

Acknowledgments

The data sets analyzed during this study are publicly available. Data set 1 is available from the University of California Irvine Machine Learning Repository [21]. Data set 2 is available from the IEEE Dataport Repository [22]. This work was supported by UK Research and Innovation (UKRI; EP/S021612/1).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of literature search and data sets.
[PDF File (Adobe PDF File), 264 KB - jmir_v26i1e46936_app1.pdf]

Multimedia Appendix 2
Details of Fair Adversarial Gradient Tree Boosting.
[PDF File (Adobe PDF File), 92 KB - jmir_v26i1e46936_app2.pdf]

References

1. O'Neil C. Weapons of math destruction: how big data increases inequality and threatens democracy. New York City, U.S: Crown; 2017.
2. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv 2022;8(32):eabq6147 [FREE Full text] [doi: 10.1126/sciadv.abq6147] [Medline: 35960806]
3. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Biocomputing 2021:232-243 [FREE Full text] [doi: 10.1142/9789811232701_0022] [Medline: 33691020]
4. Thompson HM, Sharma B, Bhalla S, Boley R, McCluskey C, Dligach D, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. J Am Med Inform Assoc 2021;28(11):2393-2403 [FREE Full text] [doi: 10.1093/jamia/ocab148] [Medline: 34383925]
5. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
6. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020;3:81 [FREE Full text] [doi: 10.1038/s41746-020-0288-5] [Medline: 32529043]

XSL•FO
RenderX

7. Liu X, Hu P, Yeung W, Zhang Z, Ho V, Liu C, et al. Illness severity assessment of older adults in critical illness using machine learning (ELDER-ICU): an international multicentre study with subgroup bias evaluation. Lancet Digit Health 2023;5(10):e657-e667 [FREE Full text] [doi: 10.1016/S2589-7500(23)00128-0] [Medline: 37599147]
8. Grari V, Ruf B, Lamprier S, Detyniecki M. Fair adversarial gradient tree boosting. 2019 Presented at: 2019 IEEE International Conference on Data Mining (ICDM); November 8-11, 2019; Beijing, China p. 1060-1065 URL: https://ieeexplore.ieee.org/document/8970941 [doi: 10.1109/icdm.2019.00124]
9. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res 2022;118(17):3272-3287 [FREE Full text] [doi: 10.1093/cvr/cvac013] [Medline: 35150240]
10. Goldraich L, Beck-da-Silva L, Clausell N. Are scores useful in advanced heart failure? Expert Rev Cardiovasc Ther 2009;7(8):985-997. [doi: 10.1586/erc.09.68] [Medline: 19673676]
11. Treece J, Chemchirian H, Hamilton N, Jbara M, Gangadharan V, Paul T, et al. A review of prognostic tools in heart failure. Am J Hosp Palliat Med 2018;35(3):514-522. [doi: 10.1177/1049909117709468] [Medline: 28554221]
12. Thorvaldsen T, Benson L, Ståhlberg M, Dahlström U, Edner M, Lund LH. Triage of patients with moderate to severe heart failure: who should be referred to a heart failure center? J Am Coll Cardiol 2014;63(7):661-671 [FREE Full text] [doi: 10.1016/j.jacc.2013.10.017] [Medline: 24161453]
13. Escamilla AKG, Hassani AHE, Andres E. A comparison of machine learning techniques to predict the risk of heart failure. In: Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems. Switzerland AG: Springer; 2019:9-26.
14. Sullivan K, Doumouras BS, Santema BT, Walsh MN, Douglas PS, Voors AA, et al. Sex-specific differences in heart failure: pathophysiology, risk factors, management, and outcomes. Can J Cardiol 2021;37(4):560-571. [doi: 10.1016/j.cjca.2020.12.025] [Medline: 33383166]
15. Walsh MN, Jessup M, Lindenfeld J. Women with heart failure: unheard, untreated, and unstudied. J Am Coll Cardiol 2019;73(1):41-43 [FREE Full text] [doi: 10.1016/j.jacc.2018.10.041] [Medline: 30621949]
16. Sobhani K, Castro DKN, Fu Q, Gottlieb RA, Van Eyk JE, Merz CNB. Sex differences in ischemic heart disease and heart failure biomarkers. Biol Sex Differ 2018;9(1):43 [FREE Full text] [doi: 10.1186/s13293-018-0201-y] [Medline: 30223899]
17. Straw I. The automation of bias in medical Artificial Intelligence (AI): decoding the past to create a better future. Artif Intell Med 2020;110:101965. [doi: 10.1016/j.artmed.2020.101965] [Medline: 33250145]
18. Hamberg K. Gender bias in medicine. Womens Health (Lond) 2008;4(3):237-243 [FREE Full text] [doi: 10.2217/17455057.4.3.237] [Medline: 19072473]
19. Krieger N, Fee E. Man-made medicine and women's health: the biopolitics of sex/gender and race/ethnicity. Int J Health Serv 1994;24(2):265-283. [doi: 10.2190/LWLH-NMCJ-UACL-U80Y] [Medline: 8034393]
20. PRISMA flow diagram. PRISMA. URL: https://www.prisma-statement.org/prisma-2020-flow-diagram [accessed 2024-07-09]
21. Tanvir AAM, Bhatti SH, Aftab M, Raza MA. Heart failure clinical records data set F. University of California Irvine Machine Learning Repository. 2020. URL: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records [accessed 2024-05-17]
22. Siddhartha M. Heart disease dataset (comprehensive). IEEE Dataport. 2020. URL: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive [accessed 2024-05-17]
23. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 2020;20(1):16 [FREE Full text] [doi: 10.1186/s12911-020-1023-5] [Medline: 32013925]
24. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. 2018 Presented at: 1st Conference on Fairness, Accountability and Transparency, PMLR 81; February 23-24, 2018; New York, NY p. 77-91 URL: https://proceedings.mlr.press/v81/buolamwini18a.html
25. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019;22(1):67-72 [FREE Full text] [doi: 10.4103/aca.ACA_157_18] [Medline: 30648682]
26. Afrose S, Song W, Nemeroff CB, Lu C, Yao DD. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. Commun Med (Lond) 2022;2:111 [FREE Full text] [doi: 10.1038/s43856-022-00165-w] [Medline: 36059892]
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(3):321-357 [FREE Full text] [doi: 10.1016/s0887-6185(02)00103-2] [Medline: 12214817]
28. Islam SR, Eberle W, Ghafoor SK, Ahmed M. Explainable artificial intelligence approaches: a survey. arXiv Preprint posted online on January 23, 2021 [FREE Full text]
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: NIPS'17: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 4768-4777 URL: https://dl.acm.org/doi/proceedings/10.5555/3295222
30. Borgese M, Joyce C, Anderson EE, Churpek MM, Afshar M. Bias assessment and correction in machine learning algorithms: a use-case in a natural language processing algorithm to identify hospitalized patients with unhealthy alcohol use. AMIA Annu Symp Proc 2022;2021:247-254 [FREE Full text] [Medline: 35308909]

XSL•FO
RenderX

31. Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR Public Health Surveill 2020;6(4):e22400 [FREE Full text] [doi: 10.2196/22400] [Medline: 33090117]
32. Tison GH, Avram R, Nah G, Klein L, Howard BV, Allison MA, et al. Predicting incident heart failure in women with machine learning: the women's health initiative cohort. Can J Cardiol 2021;37(11):1708-1714 [FREE Full text] [doi: 10.1016/j.cjca.2021.08.006] [Medline: 34400272]
33. Pombo G, Gray R, Cardoso MJ, Ourselin S, Rees G, Ashburner J, et al. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models. Med Image Anal 2023;84:102723 [FREE Full text] [doi: 10.1016/j.media.2022.102723] [Medline: 36542907]
34. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020;2(5):e221-e223 [FREE Full text] [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
35. Safer JD, Coleman E, Feldman J, Garofalo R, Hembree W, Radix A, et al. Barriers to healthcare for transgender individuals. Curr Opin Endocrinol Diabetes Obes 2016;23(2):168-171 [FREE Full text] [doi: 10.1097/MED.0000000000000227] [Medline: 26910276]
36. Rutherford L, Stark A, Ablona A, Klassen BJ, Higgins R, Jacobsen H, et al. Health and well-being of trans and non-binary participants in a community-based survey of gay, bisexual, and queer men, and non-binary and two-spirit people across Canada. PLoS One 2021;16(2):e0246525 [FREE Full text] [doi: 10.1371/journal.pone.0246525] [Medline: 33571252]
37. Beckwith N, McDowell MJ, Reisner SL, Zaslow S, Weiss RD, Mayer KH, et al. Psychiatric epidemiology of transgender and nonbinary adult patients at an urban health center. LGBT Health 2019;6(2):51-61 [FREE Full text] [doi: 10.1089/lgbt.2018.0136] [Medline: 30707624]
38. Vejjajiva A, Teasdale GM. Serum creatine kinase and physical exercise. Br Med J 1965;1(5451):1653-1654 [FREE Full text] [doi: 10.1136/bmj.1.5451.1653] [Medline: 14295325]
39. Carruthers R, Straw I, Ruffle JK, Herron D, Nelson A, Bzdok D, et al. Representational ethical model calibration. NPJ Digit Med 2022;5(1):170 [FREE Full text] [doi: 10.1038/s41746-022-00716-4] [Medline: 36333390]

Abbreviations

AI: artificial intelligence
CAD: coronary artery disease
DispFNR: disparity in false negative rate
EF: ejection fraction
FAGTB: Fair Adversarial Gradient Tree Boosting
FNR: false negative rate
FPR: false positive rate
HF: heart failure
LGBTQ+: lesbian, gay, bisexual, transgender, queer
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest

Edited by A Mavragani; submitted 03.03.23; peer-reviewed by J Zeng, S Antani, E van der Velde, L Guo; comments to author 16.06.23; revised version received 13.10.23; accepted 04.05.24; published 26.08.24.

Please cite as:
Straw I, Rees G, Nachev P
Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study
J Med Internet Res 2024;26:e46936
URL: https://www.jmir.org/2024/1/e46936
doi:10.2196/46936
PMID:

©Isabel Straw, Geraint Rees, Parashkev Nachev. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of a Natural Language Processing Approach to Identify Diagnostic Errors and Analysis of Safety Learning System Case Review Data: Retrospective Cohort Study

Azade Tabaie[1,2], PhD; Alberta Tran[3], RN, CCRN, PhD; Tony Calabria[3], MA, CPHQ, CSSBB; Sonita S Bennett[1], MSc; Arianna Milicia[4], BSc; William Weintraub[5,6], MACC, MD; William James Gallagher[6,7], MD; John Yosaitis[6,8], MD; Laura C Schubel[4], MPH; Mary A Hill[9,10], MS; Kelly Michelle Smith[9,10], PhD; Kristen Miller[4,6], MSPH, MSL, CPPS, DrPH

[1]Center for Biostatistics, Informatics, and Data Science, MedStar Health Research Institute, Washington, DC, United States
[2]Department of Emergency Medicine, Georgetown University School of Medicine, Washington, DC, United States
[3]Department of Quality and Safety, MedStar Health Research Institute, Washington, DC, United States
[4]National Center for Human Factors in Healthcare, MedStar Health Research Institute, Washington, DC, United States
[5]Population Health, MedStar Health Research Institute, Washington, DC, United States
[6]Georgetown University School of Medicine, Washington, DC, United States
[7]Family Medicine Residency Program, MedStar Health Georgetown-Washington Hospital Center, Washington, DC, United States
[8]MedStar Simulation Training & Education Lab (SiTEL), MedStar Institute for Innovation, Washington, DC, United States
[9]Institute of Health Policy, Management & Evaluation, University of Toronto, Toronto, ON, Canada
[10]Michael Garron Hospital, Toronto, ON, Canada

Corresponding Author:
Azade Tabaie, PhD
Center for Biostatistics, Informatics, and Data Science
MedStar Health Research Institute
3007 Tilden Street NW
Washington, DC, 20008
United States
Phone: 1 202 244 9810
Email: azade.tabaie@medstar.net

Abstract

Background: Diagnostic errors are an underappreciated cause of preventable mortality in hospitals and pose a risk for severe patient harm and increase hospital length of stay.

Objective: This study aims to explore the potential of machine learning and natural language processing techniques in improving diagnostic safety surveillance. We conducted a rigorous evaluation of the feasibility and potential to use electronic health records clinical notes and existing case review data.

Methods: Safety Learning System case review data from 1 large health system composed of 10 hospitals in the mid-Atlantic region of the United States from February 2016 to September 2021 were analyzed. The case review outcome included opportunities for improvement including diagnostic opportunities for improvement. To supplement case review data, electronic health record clinical notes were extracted and analyzed. A simple logistic regression model along with 3 forms of logistic regression models (ie, Least Absolute Shrinkage and Selection Operator, Ridge, and Elastic Net) with regularization functions was trained on this data to compare classification performances in classifying patients who experienced diagnostic errors during hospitalization. Further, statistical tests were conducted to find significant differences between female and male patients who experienced diagnostic errors.

Results: In total, 126 (7.4%) patients (of 1704) had been identified by case reviewers as having experienced at least 1 diagnostic error. Patients who had experienced diagnostic error were grouped by sex: 59 (7.1%) of the 830 women and 67 (7.7%) of the 874 men. Among the patients who experienced a diagnostic error, female patients were older (median 72, IQR 66-80 vs median 67, IQR 57-76; P=.02), had higher rates of being admitted through general or internal medicine (69.5% vs 47.8%; P=.01), lower rates of cardiovascular-related admitted diagnosis (11.9% vs 28.4%; P=.02), and lower rates of being admitted through neurology

XSL•FO
RenderX

department (2.3% vs 13.4%; *P*=.04). The Ridge model achieved the highest area under the receiver operating characteristic curve (0.885), specificity (0.797), positive predictive value (PPV; 0.24), and F_1-score (0.369) in classifying patients who were at higher risk of diagnostic errors among hospitalized patients.

Conclusions: Our findings demonstrate that natural language processing can be a potential solution to more effectively identifying and selecting potential diagnostic error cases for review and therefore reducing the case review burden.

(J Med Internet Res 2024;26:e50935) doi:10.2196/50935

KEYWORDS

diagnostic error; electronic health records; machine learning; natural language processing; NLP; mortality; hospital; risk; length of stay; patient harm; diagnostic; EHR

Introduction

Diagnostic errors are an underappreciated cause of preventable mortality in hospitals, estimated to affect a quarter million hospital inpatients, and account for an estimated 40,000-80,000 deaths annually in the United States [1]. These errors pose a risk for severe patient harm [2,3], increase hospital length of stay [4], and made up 22% and accounted for US $5.7 billion of paid malpractice claims in hospitalized patients throughout a nearly 13-year period [5]. In their analysis of malpractice claims occurring in the US National Practitioner Database from 1999 to 2011, Gupta et al [5] found that diagnosis-related paid claims were most likely to be associated with death and cost (following surgery); among diagnosis-related paid claims, failure to diagnose was the most common subtype and was more likely than other types to be associated with mortality. Several factors have been proposed as contributors to inpatient diagnostic errors including time constraints related to the concurrent care of multiple patients, unpredictable workflows, distractions, and competing priorities for trainees. From their systematic review and meta-analysis, Gunderson et al [2] estimate that 250,000 diagnostic adverse events occur annually among hospitalized patients in the United States, and this is likely an underestimation of the problem due to several challenges in diagnostic error measurement [6].

Challenges in identifying and measuring diagnostic errors occur due to the evolving and iterative nature of the diagnostic process, making it difficult to determine when, if at all, a correct or more specific diagnosis could have been established by clinicians to start the appropriate treatment [6]. Since its landmark report, Improving Diagnosis in Health Care, the National Academies of Science, Engineering, and Medicine (NASEM) has produced a common understanding of diagnostic error that includes accuracy, timeliness, and communication of the explanation to the patient or patient's family member [3]. Diagnostic errors often involve missed opportunities related to various aspects of the diagnostic process [7-9] and diagnostic adverse events resulting in harm [10]. However, many hospitals currently do not capture or include surveillance for diagnostic errors, despite having robust systems in place to report and analyze patient safety issues [6,11,12].

A crucial first step to improving diagnosis in hospitals is the creation of programs to identify, analyze, and learn from diagnostic errors. Ongoing efforts by the Agency for Health Care Research and Quality have supported pragmatic measurement approaches for health organizations to build a diagnostic safety program and identify and learn from diagnostic errors such as those described in the Measure Dx resource [9]. One proposed and promising solution for hospitals to improve diagnostic surveillance is to build on existing efforts to collect patient safety data, root cause analyses, or other forms of case reviews for quality improvement purposes. Cases that have already been reviewed or investigated in the organization for general patient safety and quality purposes may be able to inform or be rereviewed for information and learning opportunities specific to diagnostic safety. Widely used case-based learning methodologies in particular, such as the "Learning From Every Death" initiative developed at Mayo Clinic [13] used both nationally and worldwide, offer an excellent opportunity for hospitals to augment their existing quality and safety efforts and support diagnostic safety.

Clinical notes in electronic health records (EHRs) written by health providers in free-text format are rich sources of a patient's diagnoses and care trajectory through hospitalization time. Approaches to processing free text, such as through natural language processing (NLP) and machine learning (ML), have demonstrated significant opportunities to improve quality and safety within health care organizations in diverse applications [14-16] such as cancer research [17,18] and infection prediction [19] to sleep issues [20] and neurological outcome prediction [21]. Besides its use in the diagnostic process, ML models proved to have added benefits when used in diagnostic error identification [22,23]. However, despite significant progress and evidence about the use of these ML and NLP approaches to improve patient safety, the use of ML and NLP approaches to diagnostic safety and surveillance has largely remained untapped. A 2022 study demonstrates how an academic medical center's implementation of an NLP-based algorithm to flag safety event reports for manual review enabled early detection of emerging diagnostic risks from large volumes of safety reports, and was among the first to apply an NLP approach to safety event reports to facilitate identification of COVID-19 related diagnostic errors [24]. Meanwhile, progress in the use of data mining approaches to develop electronic trigger tools offers promising methods to detect potential diagnostic events, promote organizational learning, and support the monitoring of data prospectively to identify patients at high risk for future adverse events [25]. To our knowledge, however, NLP has not yet been applied to case review data to facilitate the identification of diagnostic errors and understand its features and sources.

XSL•FO
RenderX

While free-text formatted clinical notes provide unique opportunities to incorporate ML models, the lack of reliable labels to represent diagnostic errors often limits the use of clinical notes for diagnostic safety surveillance efforts. The opportunity to train ML and NLP algorithms to identify diagnostic errors and opportunities depends on the collation of EHR data with existing efforts to identify diagnostic errors such as through case review findings from the Safety Learning System (SLS). To further explore the potential for this approach to be used to improve diagnostic safety surveillance, a rigorous evaluation of the feasibility and potential of using EHR and existing case review data is needed.

We hypothesized that ML and NLP methods can be applied to train models based on available case review data to examine content potentially related to diagnostic errors within EHR clinical notes. These approaches automatically identify features or information from free text using controlled vocabularies, rule sets, reference dictionaries, or lexicons.

Methods

Data Sets and Case Review Approach

We analyzed SLS data from 1 large health system comprised of 10 hospitals in the mid-Atlantic region of the United States. The SLS is one example of a holistic case review methodology delivered by health care organizations in the United States and globally. Established in 2015, the SLS builds upon the Mayo Clinic Mortality Review System of Huddleston et al [13] to review and analyze EHR data from patient mortality cases to find safety issues that could be found and mitigated. This approach was designed to enhance current quality improvement projects done within health organizations, providing a perspective and strategy based on the Safety II lens and rooted in the belief that every death provides an opportunity to improve care. With a Safety II lens, participating organizations use a holistic case review methodology designed to identify vulnerabilities in systems and processes of care delivery. Reviewers identify and translate these into different categories and labels to (1) define and quantify types of process of care and system failures contributing to adverse outcomes (errors) and (2) identify the components of the process of care and system failures that when fixed will improve performance (opportunities for improvement [OFIs]).

To ensure a sufficient cross-sampling of patients across different specialties and areas, patients are selected for case reviews at this health system based on their primary provider service line category (eg, medicine, surgery, etc) and hospital length of stay; patients in primary and ambulatory care settings are not included for case review selection. The case review process occurs according to the standardized SLS methodology and recommendations [13,26], and between at least 1 physician and 1 nurse within the health system who have both received training in the SLS approach. The case review outcome and identification of OFIs, including diagnostic OFIs, relies on the reviewer's consensus of any findings and through multiple multidisciplinary and multispecialty meetings that may involve a committee Chair member, clinical department leader, or escalation to other leadership.

Sample

We obtained SLS data from February 2016 to September 2021; data in later years were available but not included because of key changes to the case selection process made during and in response to the COVID-19 pandemic. All hospitalized adult patients older than 18 years were included for analysis, regardless of their hospitalization outcome (eg, mortality or discharge location). Pediatric and neonatal patients were excluded.

Ethical Considerations

The original data collection and study protocol was approved by the institutional review board (00001245) at MedStar Health Research Institute on August 26, 2019.

Variables

Data Extraction

Medical record number, encounter number, length of stay, age, date of birth, sex, diagnosis at the time of admission (ie, *ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*] diagnosis codes), mortality, OFI categories (eg, delayed or missed diagnosis and diagnostic opportunities), number of identified OFIs and diagnosis issues (eg, the accuracy of diagnosis and confirmation or fixation bias) were the features and patient identifiers which were extracted from SLS data [13,26].

Because chart reviews generally occur at a single point in time within the patient's care trajectory, they often do not contain information or details of the patient's full hospital course. However, clinical notes written by health care providers are rich sources of patient's health status throughout their hospitalization period [27-29]. Therefore, to supplement these chart review data, we additionally extracted and included all clinical notes from the EHR for patients who could be matched by patient identifiers (eg, encounter number and date of birth).

Coding Diagnostic Errors

Case reviewers can select any number of labels to describe a diagnosis issue or an OFI identified and agreed upon by consensus. For this study, diagnostic errors were defined by the available features from chart review pertaining to diagnosis and impacting the timeliness, accuracy, or communication of a diagnosis. Our definition of diagnostic errors was limited to the categories identified during chart reviews and recorded within the SLS data set; therefore, our diagnostic error definition does not include all aspects of the definition developed by the NASEM report [3]. Table 1 describes the SLS categories and values that were labeled as diagnostic errors and used to train our classification models. Patients were coded as having experienced a diagnostic error if one or more of the conditions listed in Table 1 were identified in their SLS case review.

XSL•FO
RenderX

Table 1. Indicators of diagnostic error in Safety Learning System data.

Feature from chart reviews	Value to indicate diagnostic error
OFI[a] category	Delayed or missed diagnosis
OFI category	Diagnostic opportunities
Diagnosis issues	accuracy of diagnosis
Diagnosis issues	Accuracy of interpretation of laboratory or test results
Diagnosis issues	Squirrel (red herring lab or test results)
Diagnosis issues	Confirmation or fixation bias
Diagnosis issues	Appropriateness of chosen tests or equipment given the patient's differential diagnosis

[a]OFI: opportunity for improvement.

NLP Approach

We used an NLP approach on critical incident reporting system data to explore the features and risk of diagnostic error among hospitalized patients.

Features From Free-Text Data

Descriptive statistical analyses were performed to identify any differences among age, length of stay, and mortality between the female and male patients who had experienced diagnostic errors.

All EHR clinical notes were transformed to lowercase. Extra white spaces, numbers, punctuations, and stop words were removed and words were stemmed. The term frequency-inverse document frequency (TF-IDF) matrix was calculated for each clinical note using the bag-of-words from the preprocessed EHR clinical notes [30]. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents and is a popular method to translate free text to numerical features in training ML models. The TF-IDF of a word in a document is calculated by multiplying 2 metrics: the number of times a word appeared in a document and the inverse document frequency of the word across a set of documents. TF-IDF is computationally efficient and easy to interpret. We excluded the most frequent words that had appeared in more than 95% of the EHR clinical notes, as these frequent words do not provide information to help with the classification. Moreover, we excluded the rare words that appeared in less than 5% of the EHR clinical notes [31].

In a TF-IDF matrix, the number of rows corresponds to the unique patients, and the number of columns represents the unique words found in EHR clinical notes. There are numerous unique words used in EHR clinical notes; therefore, the TF-IDF approach provides a high-dimensional input matrix for the classification task. The high-dimensional input matrix can lead to training inaccurate classifiers. To overcome that issue, we used the chi-square statistical test to select the most relevant words to identify diagnostic errors; therefore, if P values associated with a word (also called a feature) are less than .05, that word is selected and included in the feature matrix to train ML classification models.

Classification Models

In lieu of an existing model with the same objective in the literature, a simple logistic regression model was trained as the baseline classifier to identify patients within SLS data who were at higher risk of diagnostic error. Moreover, 3 forms of logistic regression models with regularization functions were trained on this data to compare classification performances and identify the best-performing model [32]: Least Absolute Shrinkage and Selection Operator (LASSO), Ridge, and Elastic Net.

- LASSO: for a more accurate prediction, LASSO regularization is used with a logistic regression model. The LASSO procedure encourages simple, sparse models which has fewer parameters in a way that the estimated coefficient of features with less effect will be set to zero. This characteristic makes LASSO well-suited for models showing high levels of multicollinearity or variable selection and parameter elimination is needed. LASSO is also called L1 regularization.
- Ridge: also called L2 regularization, Ridge is a regularization method used for models suffering from multicollinearity or high-dimensional feature space. Ridge regularization keeps all the features regardless of their effect on the model. However, it pushes the estimated coefficient of features with less effect toward zero to minimize their effect on the classification outcome. This characteristic of Ridge makes it well-suited when most features impact the outcome variable.
- Elastic Net: a logistic regression model with Elastic Net regularization is a weighted combination of LASSO (L1) and Ridge (L2) regularizations [33]. Elastic Net can remove the effect of the insignificant features by setting their estimated coefficient to zero and lower the effect of the less significant features by pushing their estimated coefficient toward zero while adding more weights to the more important features. From implementation and interpretation aspects, the Elastic Net model is simple to use. Such characteristics make this model an accepted baseline in ML-based studies [34].

The hyperparameters of the 3 classification models were optimized through cross-validation. All the analyses were conducted using Python 3 (Python Software Foundation).

XSL•FO
RenderX

Classification Performance Metrics

We calculated 7 common performance metrics reported for binary classifiers to compare the performance of the 4 classification models: area under receiver operating characteristics curve (AUROC), sensitivity or recall or true positive rate, specificity or true negative rate, positive predictive value (PPV) or precision, negative predictive value (NPV), F_1-score, and area under precision-recall curve (AUPRC). The 7 metrics take values between 0 and 1. Values closer to 1 indicate a well-performing classifier. Multimedia Appendix 1 presents the definition of the performance metrics used in this study. Figure 1 presents the summary of the methods used in this analysis.

Figure 1. A summary of the methods used in this study. EHR: electronic health record; LASSO: Least Absolute Shrinkage and Selection Operator; TF-IDF: term frequency-inverse document frequency.

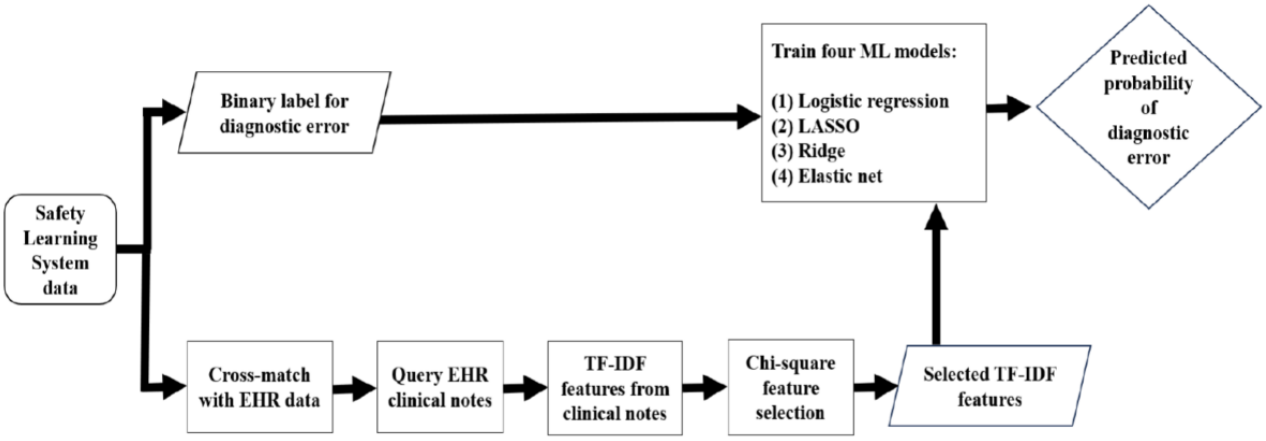

Results

Descriptive Summary

In total, there were 2184 unique patient records within SLS data from February 2016 to September 2021. EHR clinical notes were cross-matched, extracted, and included in analyses for 1704 (78%) of these SLS patient records. Of those patients with cross-matched EHR data, 126 (7.4%) patients had been identified by case reviewers as having experienced at least 1 diagnostic error. A total number of 20,848 EHR clinical notes associated with the 1704 unique patients were used in this study.

Patients who had experienced diagnostic errors were grouped by sex: 59 (7.1%) of the 830 women and 67 (7.7%) of the 874 men in the larger cross-matched sample had been found to have a diagnostic error. Table 2 presents the descriptive statistics between female and male patient groups. We applied the Wilcoxon rank sum test for numerical features (ie, age and length of stay), and the chi-square test for mortality rate, admission diagnosis, and admission department or specialty. Patients in the female group were older than the male group by a median of 72 (IQR 66-80) versus a median of 67 (IQR 57-76; P=.02). Compared to the male group, female patients who experienced diagnostic error had higher rates of being admitted through general or internal medicine (69.5% vs 47.8%; P=.01), lower rates of cardiovascular-related admitted diagnosis (11.9% vs 28.4%; P=.02), and lower rates of being admitted through neurology department (2.3% vs 13.4%; P=.04). We observed no differences between groups in mortality rates and length of stay.

XSL•FO
RenderX

Table 2. Descriptive statistics of patients.

	Patients who experienced diagnostic error		All patients	
	Female group (n=59)	Male group (n=67)	Female group (n=830)	Male group (n=874)
Age (in years), median (IQR)	72 (66-80)	67 (57-76)	72 (62-83)	69 (59-79)
Race, n (%)				
African American	38 (64)	42 (62)	429 (51.7)	429 (51.7)
Asian	0 (0)	0 (0)	12 (1.4)	12 (1.4)
Multiple	0 (0)	0 (0)	2 (0.2)	2 (0.2)
Not recorded	4 (6)	2 (2.9)	30 (3.6)	30 (3.6)
White	11 (18)	21 (31.3)	310 (37.3)	310 (37.3)
Other	6 (10)	2 (2.9)	47 (5.7)	47 (5.7)
Length of stay in days, median (IQR)	4 (6-10)	4 (8-14)	7 (4-12)	8 (4-12)
Mortality, n (%)				
Count	25 (42)	29 (43)	456 (54.9)	459 (52.5)
Admitting department or specialty, n (%)				
General or internal medicine or hospitalist	41 (69)	32 (47)	427 (51.4)	389 (44.5)
Cardiology	5 (8)	12 (17)	99 (11.9)	131 (14.9)
Critical care	6 (10)	6 (8)	117 (14.1)	142 (16.2)
Neurology	2 (3)	9 (13)	75 (9)	90 (10.3)
Pulmonary	1 (1)	1 (1)	22 (2.6)	31 (3.5)
Other	4 (6)	7 (10)	90 (10.8)	91 (10.4)
Admitting diagnosis, n (%)				
Cardiovascular	7 (11)	19 (28)	154 (18.6)	167 (19.1)
Respiratory	7 (11)	5 (7)	88 (10.6)	69 (7.9)
Sepsis	7 (11)	4 (5)	65 (7.8)	63 (7.2)
Altered mental status	1 (1)	2 (2)	36 (4.3)	28 (3.2)
Diabetes	1 (1)	1 (1)	6 (0.7)	3 (0.3)
Other	23 (38)	21 (31)	244 (29.4)	270 (30.9)
Admitting unit type, n (%)				
General care	54 (91)	60 (89)	144 (17.3)	179 (20.5)
Critical care	5 (8.5)	7 (10)	686 (82.7)	695 (79.5)
OFI[a] categories, n (%)				
Delayed or missed diagnosis	43 (72)	46 (68)	43 (5.2)	46 (5.3)
Diagnostic opportunities	15 (25)	16 (23)	15 (1.8)	16 (1.8)
Accuracy of diagnosis	1 (1)	4 (6)	1 (0.1)	4 (0.5)
Accuracy of interpretation of laboratory or test results	0 (0)	0 (0)	0 (0)	0 (0)
Squirrel (red herring lab or test results)	0 (0)	1 (1)	0 (0)	1 (0.1)
Confirmation or fixation bias	0 (0)	0 (0)	0 (0)	0 (0)
Appropriateness of chosen tests or equipment given patient's differential diagnosis	1 (1)	0 (0)	1 (0.1)	0 (0)
OFI unit type, n (%)				
Critical care	15 (25)	22 (32)	273 (32.9)	318 (36.4)
Emergency department	17 (28)	18 (26)	81 (9.8)	76 (8.7)

XSL•FO
RenderX

	Patients who experienced diagnostic error		All patients	
	Female group (n=59)	Male group (n=67)	Female group (n=830)	Male group (n=874)
General care	27 (45)	27 (40)	290 (34.9)	285 (32.6)

[a]OFI: opportunity for improvement.

Classification Models' Performance

Clinical notes were preprocessed for TF-IDF feature calculation. The bag-of-words included 2227 words, and each word was considered a feature (see Table S1 in Multimedia Appendix 2 for the top 100 words). We found that abscess, ascend, abnormality, scant, pair, and prefer were the top 5 features with the highest positive estimated coefficient (0.42 to 0.28); post, select, gave, muscl, hours, and unrespons were the top 5 features with the highest negative coefficients (–0.35 to –0.26). After applying the chi-square test, 250 features with a *P* value less than .05 were selected for the modeling process. All 4 ML classifiers were trained using the 250 selected features.

Table 3 presents the performances of the simple logistic regression and 3 regularized logistic regression models (LASSO, Ridge, and Elastic Net). The Ridge model achieved the highest AUROC (0.885), specificity (0.797), PPV (0.24), NPV (0.981), and F_1-score (0.369) in classifying patients who were at higher risk of diagnostic errors among hospitalized patients in SLS system. The simple logistic regression model obtained the highest AUPRC (0.537). The simple logistic regression model classified all patients as the ones with diagnostic errors; therefore, it achieved a sensitivity of 1, and specificity and NPV of 0.

Figures 2 and 3 present the receiver operating characteristics curves and precision-recall curves for the 4 classifiers in this study. Models that give ROC curves closer to the top-left corner indicate a better performance. The AUROC values represent the probability that a patient who experienced a diagnostic error, chosen at random, is ranked higher by the Ridge model than a randomly chosen patient who did not experience a diagnostic error. The higher value of AUPRC indicates that the Ridge model can identify patients who experienced diagnostic errors more precisely with fewer false positives compared to LASSO and Elastic Net models.

Table 3. Performance metrics. The performance metrics of 4 machine learning algorithms in classifying patients who are at higher risk of diagnostic errors among hospitalized patients. The numbers represent the average of the 5-fold cross-validation approach.

	Simple logistic regression	LASSO[a]	Ridge	Elastic Net
AUROC[b]	0.5	0.846	0.885	0.859
Sensitivity	1.0	0.802	0.802	0.802
Specificity	0	0.733	0.797	0.742
Positive predictive value	0.074	0.193	0.24	0.199
Negative predictive value	0	0.979	0.981	0.979
F_1-score	0.138	0.312	0.369	0.319
AUPRC[c]	0.537	0.361	0.491	0.411

[a]LASSO: Least Absolute Shrinkage and Selection Operator.

[b]AUROC: area under receiver operating characteristics curve.

[c]AUPRC: area under precision-recall curve.

XSL•FO
RenderX

Figure 2. Receiver operating characteristics curves. LASSO: Least Absolute Shrinkage and Selection Operator.

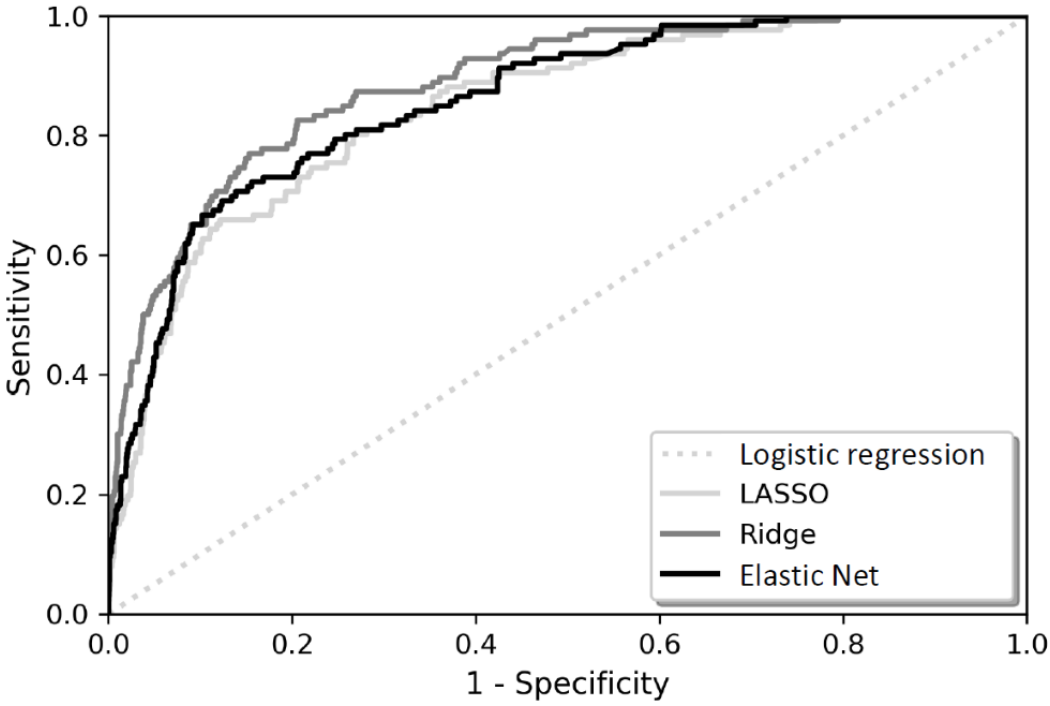

Figure 3. Precision recall curves. LASSO: Least Absolute Shrinkage and Selection Operator.

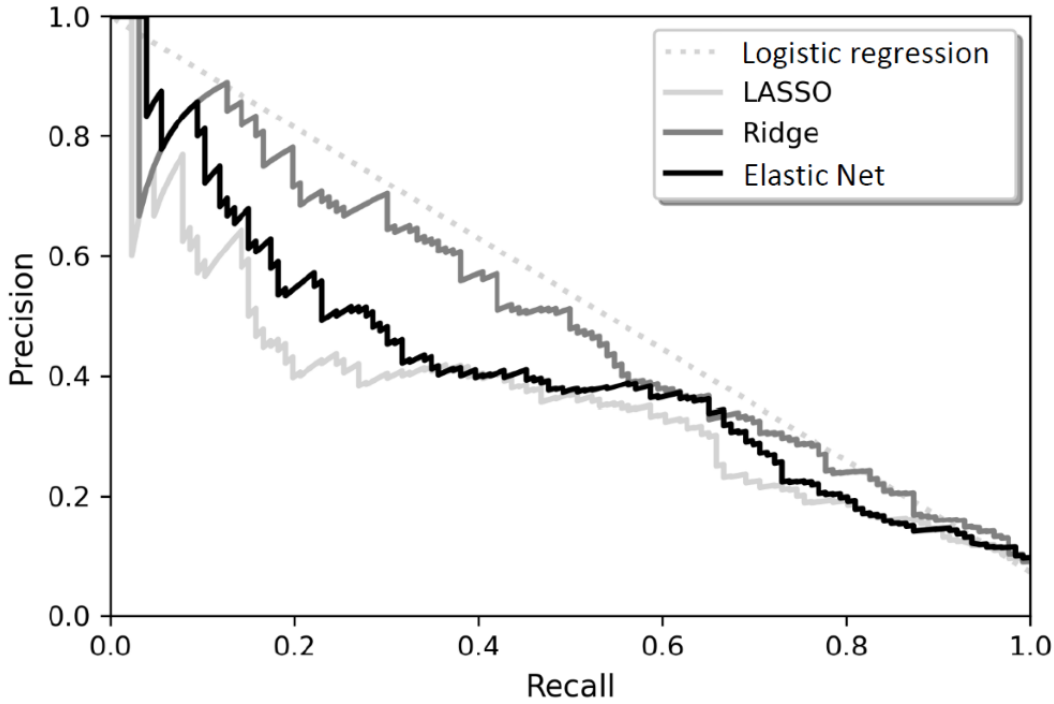

Discussion

Principal Findings

Our contribution is 2-fold; first, we integrated 2 data sources that are currently used by and available to many organizations across the United States, SLS and EHR data, to explore the use of ML and NLP algorithms to help identify diagnostic errors among hospitalized patients. Although case review methodologies offer rich insights into systems errors and OFIs, the predefined pull-down menus and structured data labels typically do not capture all the necessary clinical and contextual

XSL•FO
RenderX

details that are considered by reviewers. Therefore, a large portion of these case review data are stored in free-text narratives that typically record key information and judgments decided upon by the multidisciplinary reviewers. However, given persistent issues of staff shortage and lack of time in health care settings, it is becoming increasingly important to lower the burden of systematic EHR data reviews for health care providers while maintaining the review systems in place. Second, any developed ML and NLP approaches can potentially be incorporated to generate a diagnostic error risk score for each patient. The predicted risk score can be used in identifying and prioritizing patients for focused chart reviews, thus lowering the burden of systematic EHR data reviews for health care providers while maintaining the review systems in place.

To our knowledge, this study is the first attempt to apply and test several different ML classification models to identify diagnostic errors within routinely collected organizational case review data. Despite a substantial body of literature about the prevalence of diagnostic errors in hospital settings, current efforts to identify diagnostic errors generally rely on reviews of patient cases and data by clinical or quality teams that often are resource-intensive. ML classification models and NLP techniques offer an opportunity to generate diagnostic error risk scores to sort through large data sets and identify signals of potential diagnostic errors that can be flagged for further review. However, these classification models require a high number of observations (and identified diagnostic errors) to perform well, which might not be feasible for health organizations that are just beginning to identify diagnostic errors or may have limited personnel and efforts to perform high numbers of case reviews. In this study, we accessed nearly 2000 patient records (and of those, only 126 cases of diagnostic errors), which is considered to be a limited data sample size in the field of ML. However, techniques, such as feature selection and n-fold cross-validations, can potentially be approaches to address small sample size challenges [35].

Using the results of the simple logistic regression model as the baseline performance, we found that 3 regularization functions, namely LASSO, Ridge, and Elastic Net, boosted the performance of the baseline model. The Ridge model outperformed the rest of the models in terms of multiple performance metrics: AUROC of 0.885, specificity of 0.797, PPV of 0.24, NPV of 0.981, and F_1-score of 0.369. The Ridge algorithm tries to keep all features in the model even the features with a slight effect on the classification outcome. Since the patterns pointing at a diagnostic error were subtle in the clinical notes, even a small effect of a feature on the model's classification outcome could be important for the classification model to learn. On the other hand, the LASSO algorithm rigorously removes features that have a small effect on the classification outcome. The Elastic Net model is a weighted combination of LASSO and Ridge. The performance results presented in Table 3 show that the values achieved by the Elastic Net model lie between those of the LASSO and Ridge models.

Insights From Diagnostic Errors Within Free-Text Clinical Notes

We did not find the free text formatted clinical notes in the EHR to reflect any sort of direct language around diagnostic errors. Our analysis identified no use of the terms misdiagnosis, missed diagnosis, or diagnostic error within clinical notes, finding instead more subtle signals pointing at diagnostic errors such as "there may be a chance of misreading the test," or "insufficient data to make a diagnosis." Our findings demonstrate that NLP algorithms can be used to identify such patterns and find the associations between diagnostic errors and the subtle signals in the clinical notes. A natural extension of this work can focus on using other feature extraction methods, such as Bidirectional Encoder Representations from Transformers contextualized word embeddings, and explore the use of the pretrained language models for this objective.

We found that the presence of terms, such as abscess, abnormality, "cp" (chest pain), and dialysis in a patient's EHR clinical note were associated with reviewer-identified diagnostic errors (Multimedia Appendix 2). Misinterpretation of chest pain, specifically among female patients, has the potential to cause a cardiovascular-related diagnosis error [36]. Patients with chronic kidney disease are at higher risk of cardiovascular complications [37]. Missing such risk for a patient who is on dialysis, adds to the risk of diagnostic error.

Clinical and System Implications Around Diagnostic Inequity

Diagnostic inequity is defined as "the presence of preventable unwarranted variations in diagnostic process among population groups that are socially, economically, demographically, or geographically disadvantaged" [38]. Despite persistent and well-documented disparities in health care access and outcomes across different population groups, few studies have examined the association between diagnostic errors and health care disparities [39]. Recent evidence supports the notion that variation in diagnostic error rates across demographic groups may exist, particularly across sex. A systematic review of diagnostic errors in the emergency department, for example, found that female sex and non-White race were often associated with increased risk for diagnostic errors across several clinical conditions in emergency settings [40]. In cardiovascular medicine, a national cohort study of acute myocardial infarctions found that women were nearly twice as likely as men to receive the wrong initial diagnosis following signs of a heart attack [41]. Despite efforts to understand and reduce disparities in diagnosis and treatment, women not only continue to be understudied, underdiagnosed, and undertreated in cardiovascular medicine [42] but also may experience longer lengths of time to diagnosis than men in most patterns of disease diagnosis [43].

The analysis of case review data and other system-based data (eg, patient safety events or incident reporting) by subsets offer an opportunity to identify events in vulnerable patient populations and help sensitize clinicians to potential biases within the diagnostic process. To explore sex differences in diagnostic errors within our case review data, we statistically compared demographic and clinical differences between female

XSL•FO
RenderX

and male patients who had been identified in case reviews as having experienced diagnostic error or errors. We found that of those patients who had experienced diagnostic error or errors, the female group of patients were older, had higher rates of being admitted through general or internal medicine or hospitalist (vs specialty) departments, and had lower rates of having a cardiovascular diagnosis on admission. These preliminary results of this study revealed unexpected differences between male and female diagnostic error groups, offering novel insights that warrant further investigation to fully understand the mechanisms underlying these relationships and their implications for clinical decision-making and practice. Future uses of NLP can potentially support clinical and system-based approaches to capture and increase the evidence around structural biases or disparities in diagnoses. Individual cases from these types of data sources could be used as example narratives to engage clinicians and improve clinician learning, contributing to the development of tailored clinician and systemic interventions that can improve quality and equity throughout the diagnostic process.

Limitations

This study has several limitations. Our definition of diagnostic errors was limited to the categories and labels used within the SLS data set, reviewer interpretations of cases (subject to reviewer bias), and does not include all aspects of the definition developed by the NASEM report [3]. Despite several continued differences in definitions of diagnostic error in the peer-reviewed literature [8], we recommend that quality and safety teams within health systems use the NASEM definition for diagnostic error—including errors in communicating the diagnosis to the patient—to develop any definitions, categories, or labels used in their case review and surveillance initiatives. Although a time-consuming task, future studies could consider EHR data chart reviews to have the ground truth for the diagnostic error cases and add to the accuracy of the data set used for training the ML classifiers. Additionally, due to staffing challenges and shifting organizational priorities, case review selection varies by hospital and has changed over time, resulting in a relatively small sample size and also introducing the potential for bias. Our data came from a single health system and may reflect the specific language, culture, and practices occurring within the system and therefore may not be similar to that of other health systems. To enhance the external validity and generalizability of results, future efforts and research studies should consider the random selection of cases to evaluate both diagnostic and general quality issues within the organization; studies with larger sample sizes can build on our preliminary findings and test differences between clinical subgroups. Finally, our classification models were developed and evaluated based on a retrospective cohort from EHR; therefore, the performance may deteriorate when the method is applied to real-time data. Further work or future studies should be conducted to prospectively validate the models.

Conclusions

We performed an NLP approach and compared 4 techniques to classify patients who were at a higher risk of experiencing diagnostic error during hospitalization. Our findings demonstrate that NLP can be a potential solution to more effectively identifying and selecting potential diagnostic error cases for review, and therefore, reducing the case review burden.

Acknowledgments

This work was supported by the Agency for Health Care Research and Quality (grant 5R18HS027280-02).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Binary classification performance metrics.
[DOCX File , 24 KB - jmir_v26i1e50935_app1.docx]

Multimedia Appendix 2
The Estimated Coefficient from the Ridge Model.
[DOCX File , 28 KB - jmir_v26i1e50935_app2.docx]

References

1. Newman-Toker DE, Wang Z, Zhu Y, Nassery N, Saber Tehrani AS, Schaffer AC, et al. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "Big Three". Diagnosis (Berl) 2021 Feb 23;8(1):67-84 [FREE Full text] [doi: 10.1515/dx-2019-0104] [Medline: 32412440]
2. Gunderson CG, Bilan VP, Holleck JL, Nickerson P, Cherry BM, Chui P, et al. Prevalence of harmful diagnostic errors in hospitalised adults: a systematic review and meta-analysis. BMJ Qual Saf 2020 Dec;29(12):1008-1018. [doi: 10.1136/bmjqs-2019-010822] [Medline: 32269070]
3. Ball JR, Miller BT, Balogh EP. Improving Diagnosis in Health Care. Washington, DC: National Academies Press; 2015:472.

XSL•FO
RenderX

4. Hautz WE, Kämmer JE, Hautz SC, Sauter TC, Zwaan L, Exadaktylos AK, et al. Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. Scand J Trauma Resusc Emerg Med 2019 May 08;27(1):54 [FREE Full text] [doi: 10.1186/s13049-019-0629-z] [Medline: 31068188]
5. Gupta A, Snyder A, Kachalia A, Flanders S, Saint S, Chopra V. Malpractice claims related to diagnostic errors in the hospital. BMJ Qual Saf 2017 Aug 09;27(1):53-60. [doi: 10.1136/bmjqs-2017-006774] [Medline: 28794243]
6. Zwaan L, Singh H. Diagnostic error in hospitals: finding forests not just the big trees. BMJ Qual Saf 2020;29(12):961-964. [doi: 10.1136/bmjqs-2020-011099] [Medline: 32753410]
7. Singh H. Editorial: helping health care organizations to define diagnostic errors as missed opportunities in diagnosis. Jt Comm J Qual Patient Saf 2014;40(3):99-101. [doi: 10.1016/s1553-7250(14)40012-6] [Medline: 24730204]
8. Giardina TD, Hunte H, Hill MA, Heimlich SL, Singh H, Smith KM. Defining diagnostic error: a scoping review to assess the impact of the national academies' report improving diagnosis in health care. J Patient Saf 2022;18(8):770-778 [FREE Full text] [doi: 10.1097/PTS.0000000000000999] [Medline: 35405723]
9. Measure Dx: a resource to identify, analyze, and learn from diagnostic safety events. AHRQ: Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/patient-safety/settings/multiple/measure-dx.html [accessed 2023-06-22]
10. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Arch Intern Med 2010;170(12):1015-1021. [doi: 10.1001/archinternmed.2010.146] [Medline: 20585065]
11. Singh H. Diagnostic errors: moving beyond 'no respect' and getting ready for prime time. BMJ Qual Saf 2013;22(10):789-792 [FREE Full text] [doi: 10.1136/bmjqs-2013-002387] [Medline: 24048615]
12. Zwaan L, El-Kareh R, Meyer AND, Hooftman J, Singh H. Advancing diagnostic safety research: results of a systematic research priority setting exercise. J Gen Intern Med 2021;36(10):2943-2951 [FREE Full text] [doi: 10.1007/s11606-020-06428-3] [Medline: 33564945]
13. Huddleston JM, Diedrich DA, Kinsey GC, Enzler MJ, Manning DM. Learning from every death. J Patient Saf 2014;10(1):6-12. [doi: 10.1097/PTS.0000000000000053] [Medline: 24553440]
14. Arnaud E, Elbattah M, Gignon M, Dequen G. NLP-based prediction of medical specialties at hospital admission using triage notes. 2021 Presented at: IEEE 9th International Conference on Healthcare Informatics (ICHI); August 09-12, 2021; Victoria, British Columbia, Canada. [doi: 10.1109/ICHI52183.2021.00103]
15. Bansal A, Saraswat B, Sharma B, Nayan S, Kathuria K. Data extraction and integration from unstructured electronic health records. 2023 Presented at: 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS); November 01-03, 2023; Tashkent, Uzbekistan. [doi: 10.1109/ICTACS59847.2023.10390009]
16. Sezgin E, Hussain S, Rust S, Huang Y. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. JMIR Form Res 2023;7:e43014 [FREE Full text] [doi: 10.2196/43014] [Medline: 36881467]
17. Asri H, Mousannif H, Moatassime HA, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 2016;83:1064-1069. [doi: 10.1016/j.procs.2016.04.224]
18. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. Nat Rev Urol 2019;16(7):391-403. [doi: 10.1038/s41585-019-0193-3] [Medline: 31092914]
19. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
20. Nasiri S, Gari D. Attentive adversarial network for large-scale sleep staging. 2020 Presented at: Proceedings of the 5th Machine Learning for Healthcare Conference, PMLR; September, 2020; Atlanta, GA, USA p. 457-478 URL: https://proceedings.mlr.press/v126/nasiri20a.html
21. Tabaie A, Nemati S, Allen JW, Chung C, Queiroga F, Kuk W, et al. Assessing contribution of higher order clinical risk factors to prediction of outcome in aneurysmal subarachnoid hemorrhage patients. AMIA Annu Symp Proc 2019;2019:848-856 [FREE Full text] [Medline: 32308881]
22. Cheheltani R, King N, Lee S, North B, Kovarik D, Evans-Molina C, et al. Predicting misdiagnosed adult-onset type 1 diabetes using machine learning. Diabetes Res Clin Pract 2022;191:110029 [FREE Full text] [doi: 10.1016/j.diabres.2022.110029] [Medline: 35940302]
23. Hercus C, Hudaib A. Delirium misdiagnosis risk in psychiatry: a machine learning-logistic regression predictive algorithm. BMC Health Serv Res 2020;20(1):151 [FREE Full text] [doi: 10.1186/s12913-020-5005-1] [Medline: 32106845]
24. Shen L, Levie A, Singh H, Murray K, Desai S. Harnessing event report data to identify diagnostic error during the COVID-19 pandemic. Jt Comm J Qual Patient Saf 2022;48(2):71-80 [FREE Full text] [doi: 10.1016/j.jcjq.2021.10.002] [Medline: 34844874]
25. Murphy DR, Meyer AN, Sittig DF, Meeks DW, Thomas EJ, Singh H. Application of electronic trigger tools to identify targets for improving diagnostic safety. BMJ Qual Saf 2019;28(2):151-159 [FREE Full text] [doi: 10.1136/bmjqs-2018-008086] [Medline: 30291180]
26. The safety learning system®. Discover the Hidden Causes of Patient Harm. URL: http://hbhealthcaresafety.org/safety-learning-system/ [accessed 2023-06-22]

XSL•FO
RenderX

27. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; June, 2019; Minneapolis, Minnesota, USA. [doi: 10.18653/v1/W19-1909]
28. Huang K, Altosaar A, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv. Preprint posted online on April 10, 2019 2019:190405342. [doi: 10.48550/arXiv.1904.05342]
29. Zhang D, Thadajarassiri J, Sen C, Rundensteiner E. Time-aware transformer-based network for clinical notes series prediction. 2020 Presented at: Proceedings of the 5th Machine Learning for Healthcare Conference; September, 2020; Worcester, MA, USA URL: https://proceedings.mlr.press/v126/zhang20c/zhang20c.pdf
30. Jones KS. A statistical interpretation of term specificity and its application in retrieval. J Doc 1972;28(1):11-21. [doi: 10.1108/eb026526]
31. Tabaie A, Sengupta S, Pruitt ZM, Fong A. A natural language processing approach to categorise contributing factors from patient safety event reports. BMJ Health Care Inform 2023;30(1):e100731 [FREE Full text] [doi: 10.1136/bmjhci-2022-100731] [Medline: 37257922]
32. Aheto JMK, Duah HO, Agbadi P, Nakua EK. A predictive model, and predictors of under-five child malaria prevalence in Ghana: how do LASSO, Ridge and Elastic Net regression approaches compare? Prev Med Rep 2021;23:101475 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101475] [Medline: 34306999]
33. Zou H, Trevor H. Regularization and variable selection via the Elastic Net. J R Stat Soc Series B Stat Methodol 2005;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
34. Lee SJ, Weinberg BD, Gore A, Banerjee I. A scalable natural language processing for inferring BT-RADS categorization from unstructured brain magnetic resonance reports. J Digit Imaging 2020;33(6):1393-1400 [FREE Full text] [doi: 10.1007/s10278-020-00350-0] [Medline: 32495125]
35. Masino AJ, Harris MC, Forsyth D, Ostapenko S, Srinivasan L, Bonafide CP, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. PLoS One 2019;14(2):e0212665 [FREE Full text] [doi: 10.1371/journal.pone.0212665] [Medline: 30794638]
36. Mehta PK, Bess C, Elias-Smale S, Vaccarino V, Quyyumi A, Pepine CJ, et al. Gender in cardiovascular medicine: chest pain and coronary artery disease. Eur Heart J 2019;40(47):3819-3826 [FREE Full text] [doi: 10.1093/eurheartj/ehz784] [Medline: 31713592]
37. Jankowski J, Floege J, Fliser D, Böhm M, Marx N. Cardiovascular disease in chronic kidney disease: pathophysiological insights and therapeutic options. Circulation 2021;143(11):1157-1172 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.050686] [Medline: 33720773]
38. Giardina TD, Woodard LD, Singh H. Advancing diagnostic equity through clinician engagement, community partnerships, and connected care. J Gen Intern Med 2023;38(5):1293-1295 [FREE Full text] [doi: 10.1007/s11606-022-07966-8] [Medline: 36604388]
39. Ibrahim SA, Pronovost PJ. Diagnostic errors, health disparities, and artificial intelligence: a combination for health or harm? JAMA Health Forum 2021;2(9):e212430 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.2430] [Medline: 36218658]
40. Newman-Toker DE, Peterson SM, Badihian S, Hassoon A, Nassery N, Parizadeh D, et al. Diagnostic Errors in the Emergency Department: A Systematic Review [internet]. Rockville (MD): Agency for Healthcare Research and Quality; 2022.
41. Wu J, Gale CP, Hall M, Dondo TB, Metcalfe E, Oliver G, et al. Editor's choice—impact of initial hospital diagnosis on mortality for acute myocardial infarction: a national cohort study. Eur Heart J Acute Cardiovasc Care 2018;7(2):139-148 [FREE Full text] [doi: 10.1177/2048872616661693] [Medline: 27574333]
42. Burgess S. Understudied, under-recognized, underdiagnosed, and undertreated: sex-based disparities in cardiovascular medicine. Circ Cardiovasc Interv 2022;15(2):e011714. [doi: 10.1161/CIRCINTERVENTIONS.121.011714] [Medline: 35067073]
43. Sun TY, Hardin J, Nieva HR, Natarajan K, Cheng R, Ryan P, et al. Large-scale characterization of gender differences in diagnosis prevalence and time to diagnosis. MedRxiv. Preprint posted online on October 16, 2023 2023 [FREE Full text] [doi: 10.1101/2023.10.12.23296976] [Medline: 37873224]

Abbreviations

AUPRC: area under precision-recall curve
AUROC: area under receiver operating characteristic curve
EHR: electronic health record
ICD-10: International Statistical Classification of Diseases, Tenth Revision
LASSO: Least Absolute Shrinkage and Selection Operator
ML: machine learning
NASEM: National Academies of Science, Engineering, and Medicine
NLP: natural language processing
NPV: negative predictive value
OFI: opportunity for improvement

XSL•FO
RenderX

PPV: positive predictive value
SLS: Safety Learning System
TF-IDF: term frequency-inverse document frequency

Edited by S Ma, T Leung; submitted 17.07.23; peer-reviewed by D Chrimes, M Elbattah; comments to author 18.01.24; revised version received 21.03.24; accepted 20.06.24; published 26.08.24.

Please cite as:
Tabaie A, Tran A, Calabria T, Bennett SS, Milicia A, Weintraub W, Gallagher WJ, Yosaitis J, Schubel LC, Hill MA, Smith KM, Miller K
Evaluation of a Natural Language Processing Approach to Identify Diagnostic Errors and Analysis of Safety Learning System Case Review Data: Retrospective Cohort Study
J Med Internet Res 2024;26:e50935
URL: https://www.jmir.org/2024/1/e50935
doi:10.2196/50935
PMID:

©Azade Tabaie, Alberta Tran, Tony Calabria, Sonita S Bennett, Arianna Milicia, William Weintraub, William James Gallagher, John Yosaitis, Laura C Schubel, Mary A Hill, Kelly Michelle Smith, Kristen Miller. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Caregiver and Youth Characteristics That Influence Trust in Digital Health Platforms in Pediatric Care: Mixed Methods Study

Eric Chow[1], BSc; Alice Virani[2], PhD; Susan Pinkney[3], MA; Fatema S Abdulhussein[4], MD; Tibor van Rooij[3,5], PhD; Matthias Görges[3,6], PhD; Wyeth Wasserman[2,3], PhD; Jeffrey Bone[3], MSc; Holly Longstaff[7], PhD; Shazhan Amed[3,4], MD

[1]Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[2]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada
[3]BC Children's Hospital Research Institute, Vancouver, BC, Canada
[4]Department of Pediatrics, BC Children's Hospital, Vancouver, BC, Canada
[5]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada
[6]Department of Anesthesiology Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada
[7]Director Research Integration and Innovation, Provincial Health Services Authority, Vancouver, BC, Canada

Corresponding Author:
Shazhan Amed, MD
Department of Pediatrics
BC Children's Hospital
4480 Oak Street, Room K4-206
Vancouver, BC, V6H 3V4
Canada
Phone: 1 604 875 2117
Fax: 1 604 875 3231
Email: SAmed@cw.bc.ca

Abstract

Background: Combining patient-generated health data and digital health platforms may improve patient experience and population health, mitigate rising health care costs, reduce clinician burnout, and enable health equity. However, lack of trust may be a notable barrier to the data-sharing required by such platforms. Understanding sociodemographic, health, and personal characteristics will enable developers and implementers of such technologies to consider these in their technical design requirements.

Objective: This study aims to understand relationships between sociodemographic characteristics of caregivers of children or adolescents and trust in and willingness to use digital platforms to store and share personal health information for clinical care and research.

Methods: This study used a mixed methods approach, including surveys of caregivers of youth aged <18 years living in Canada or the United States and youth aged 16 to 17 years living in Canada, as well as web-based bulletin board discussions to further explore topics of trust in data sharing. Sociodemographic and survey data were tabulated and explored using proportional odds ordinal regression models. Comments from web-based group discussions were analyzed thematically using a coding approach to identify issues important to the participants.

Results: Survey data from 1128 caregivers (female participants: n=549, 48.7%; 36-50 years old: n=660, 58.5%; Canadian: n=603, 53.5%; urban population: n=494, 43.8%) were collected, of which 685 (60.7%) completed all questions. Data from 173 youth (female participants: n=73, 42.2%; urban population: n=94, 54.3%) were collected, of which 129 (74.6%) completed all questions, and data were available for analysis. Furthermore, among 40 participants, 23 (58%) caregivers contributed to the web-based discussion boards. Related to trust, living in a rural area (vs urban; odds ratio [OR] 0.66, 95% CI 0.46-0.95) resulted in lower concern for data privacy and security, while having an undergraduate (OR 1.82, 95% CI 1.30-2.55) or graduate degree (vs secondary or trade school; OR 2.50, 95% CI 1.68-3.73) resulted in higher levels of concern. Living with a chronic disease (OR 1.81, 95% CI 1.35-2.44) increased levels of concern regarding data privacy and security. Interestingly, those with chronic disease were more willing to use digital platforms for clinical care and share personal health information for not-for-profit research.

XSL•FO
RenderX

Caregivers were most concerned about data breaches involving data from their children but also highlighted that digital platforms would allow for better coordination of care for their children.

Conclusions: Our research confirms the willingness of caregivers and youth to use digital platforms for both clinical care delivery and research and suggests that the value of a digital platform may outweigh the risks of its use. Engagement of end users in co-designing such platforms has the potential to enhance digital trust. However, digital trust varies across sociodemographic groups; therefore, diverse end user engagement is necessary when designing digital applications.

(J Med Internet Res 2024;26:e53657) doi:10.2196/53657

KEYWORDS

pediatrics; patient trust; security; data privacy; data sharing; caregivers; patient engagement; co-design; personal health information; secondary use of data

Introduction

Background

Digital health can potentially advance the quintuple aim for health care improvement [1] by enhancing the patient experience, improving population health, mitigating rising health care costs, reducing clinician burnout, and enabling health equity [2,3]. Increased use of wearables that monitor health parameters, such as blood pressure, heart rate and rhythm, and interstitial glucose levels in real time produces vast amounts of patient-generated health data that, when combined with digital health platforms, can support remote patient monitoring, continuous (rather than episodic) care, and provide a more personalized care experience [4]. Moreover, patient-generated health and wellness data repositories provide research and quality improvement opportunities.

However, there are many obstacles to implementing digital health solutions [5]. Studies investigating the public's perspective on sharing digital health data for clinical care and research have reported concerns related to trust in data sharing, such as lack of anonymity, vulnerability to cyberattacks, and fear of data breaches leading to data misuse [6,7]. Many authors have sought to define the criteria for trustworthiness in digital platforms [8-12] and have revealed key themes, including ease of use and ease of platform use, personal recommendations from other known users, and safety and privacy protection measures [8-14]. The reputation of digital providers and the quality of information are also perceived as fostering trust [12].

Trustworthiness, however, is influenced by a range of sociocultural and political factors [8,12], yet few studies have measured their magnitude of influence. In addition, the use of artificial intelligence in medicine is increasing [15]. However, patients have expressed concerns related to the possibility of misdiagnosis and privacy breaches [16], further highlighting the importance of understanding factors that promote trust in the design of digital health platforms.

As a mechanism to enhance the patient experience and improve population health, we are developing a digital health platform (TrustSphere) [17] for the secure sharing of patient-generated health data between patients and clinicians that enables a collaborative clinical care experience. This digital platform also provides opportunities for patients to share their patient-generated health data with researchers via a digitized consent process. Our first test use case is children living with type 1 diabetes (T1D), one of the most common childhood chronic diseases [18]. T1D is characterized by absolute insulin deficiency resulting in impaired blood glucose level regulation and serious lifelong complications, such as cardiovascular disease, kidney failure, and blindness. To mitigate the risk of these complications, individuals living with T1D (and their caregivers) must carefully monitor glucose levels 24 hours a day. Modern diabetes technologies, such as continuous glucose monitoring systems have been a "game changer" where, instead of using a glucose meter that requires finger pricks 4 to 6 times per day, patients wear a sensor that sits just under their skin and measures glucose levels every 1 to 5 minutes, with the data "pushed" to a smart device in real time [19]. Studies show that the use of continuous glucose monitoring systems compared with standard "finger prick" blood glucose monitoring resulted in significantly improved control of glucose levels in children and youth living with T1D [20]. However, the integration of these patient-generated glucose levels with digital health platforms that support a collaborative clinical care experience and provide opportunities for patients to participate in research is lacking, partly due to a lack of digital trust.

This Study

We aimed to conduct a Canada and US-focused mixed methods study involving caregivers of children aged <18 years and youth aged 16 to 17 years to understand the relationship between sociodemographic characteristics (ie, sex, household income, level of education, rural vs urban locations, and experience with chronic disease) and "trust in" and "willingness to use" a digital platform to store and share personal health information (PHI) for clinical care and research. The United States and Canada are large North American countries with developed health care systems that are different yet share numerous similarities in their care models and associated challenges. Both nations grapple with escalating health care costs, inequitable access to care, and disparities in health outcomes. Moreover, there is a mounting level of concern in both countries regarding data security. We postulate that there will be differences in perspectives across different sociodemographic variables, and that understanding these differences will be important to consider in the design and prioritized features and functionalities of digital health platforms.

XSL•FO
RenderX

Methods

Study Population

Population groups that were approached for this survey study included caregivers of youth aged <18 years living in Canada or the United States (excluding Mexico) and youth aged 16 to 17 years living in Canada. Caregivers of children and youth living with T1D accessing care at the BC Children's Hospital Diabetes Clinic (Vancouver, BC) were also invited to participate.

All survey respondents were offered the opportunity to participate in web-based bulletin board discussion groups that explored the topics of trust in data sharing. To be eligible for the web-based bulletin board discussion groups, participants had to be aged >18 years, living in Canada, have at least one healthy child or a child with a chronic disease who is aged <18 years, and be able to read, write, and understand English.

Recruitment

Survey respondents were invited through the following methods: First, caregivers and youth living in Canada and the United States were invited by Insights West, a Canadian marketing research company that maintains a panel of volunteers to electively participate in web-based surveys and focus groups, along with their trusted panel partners (Dynata and Maru/Blue) from their list of adult volunteers. The youth included in this study were the children of the caregiver survey respondents living in Canada and were given parental consent to participate in this study. The target sample size was 1000; 1028 adult panel members and 173 youth responded. Caregivers of children living with T1D and receiving care at the BC Children's Hospital Diabetes Clinic were also recruited via a clinical registry. The survey invitation was sent to 232 caregivers and 100 responded, resulting in a response rate of 43%. No financial incentives or honorariums were offered for survey participation.

The web-based bulletin board discussions were facilitated through 2 separate group discussions over 3 days in February 2021. A unique ID code identified caregiver participants who expressed interest in participating in the qualitative study on the quantitative survey; caregivers were not individually identifiable. Individuals who expressed interest were recontacted and asked additional screening questions (ie, age >18 years, living in Canada, have at least one child aged <18 years, and able to read, write, and understand English) before they were invited to participate. Invited caregiver participants then provided informed consent and received a link to the web-based bulletin board discussions. Participants in the bulletin board discussions were offered an honorarium of CAD $75 (US $57) in appreciation of their time.

Data Collection

Overview

This mixed methods study was conducted from December 2020 to January 2021. The goal of the study was to gather caregiver and youth perspectives on elements of digital health delivery which included perspectives surrounding digital security, privacy and identity, ethics and informed consent, trust in digital health applications and platforms, sharing of digital health information, and perspectives on key features of an integrated digital platform for the delivery of clinical care and conduct of research.

Quantitative Methods

Survey Development

As no existing published validated surveys exploring these questions were available, the survey questions used in this study were collaboratively developed by the research team by drawing upon existing literature while also applying their expertise in qualitative and quantitative research methodologies, clinical care, patient engagement, privacy, procedural and substantive ethics and consent, digital health, and health informatics. The survey questions were also reviewed by 2 physicians in the Division of Endocrinology at BC Children's Hospital and one health informatics researcher with expertise in questionnaire development to ensure clarity and relevance to current clinical practice. The survey underwent rigorous refinement; however, it was not pilot-tested. The following description of a digital platform was provided in the survey:

> *A secure online platform that will be customized for child and youth patients and their caregivers, and will integrate a patient's health information such as diagnoses, medications and treatments, appointments, lab test results, wearable data (e.g. FitBit), etc. This platform would use secure and trusted digital identification, and follow the highest healthcare industry and public standards of privacy protection. The platform would help make it easier for children and families to access their health information and care plans, and to communicate directly with healthcare providers. It would also allow users to share their health information and care plans, if desired, with others involved in their child's care, as well as donate their data confidentially for research.*

The final survey for adults and youth comprised 32 to 36 questions (Multimedia Appendix 1) and 24 to 25 questions (Multimedia Appendix 2), respectively, depending on the responses and branching logic. The survey took 10 to 15 minutes to complete. The response options were predominantly Likert scales; however, some responses were binary, multiple selection, or rank order. There were no questions allowing for open-text responses. Respondents were able to skip questions, with no forced questions. All questions appeared on the screen except for branching questions that would only be displayed if relevant to the respondent's prior answer.

Survey Dissemination

The main survey was provided to all caregivers, and a modified version of this survey was provided to youth, in which 10 caregiver-specific questions were removed. Insights West or its partners sent out invitations to participate via email. Invitations included a brief outline of the survey topic, the approximate time required to complete the survey, and a unique link to the web-based survey hosted by Insights West where each participant could submit a singular survey response. To protect anonymity, participant identifiers were kept separate from survey responses. The same survey as above was sent out by

XSL•FO
RenderX

email by the clinic administrator to caregivers of children living with T1D and receiving care at BC Children's Hospital Diabetes Clinic. No reminder emails were sent after the initial invitation.

Qualitative Methods

Discussion Guide Development

The qualitative discussion guide was codeveloped by Insights West in collaboration with the study team. The web-based bulletin board discussion group included 26 questions (Multimedia Appendix 3) about trust, data privacy, research, and whether families would use a digital health platform like TrustSphere. Of note, 23 of the original 26 discussion board questions were analyzed for this paper. The 3 discarded questions were unrelated to trust.

Discussion Board

Participants were asked to spend approximately 15 to 20 minutes per day answering questions over the course of 3 days. The total time was 45 to 60 minutes, and individuals could stop participating anytime. A moderator at Insights West monitored the discussion group daily, and follow-up questions were asked publicly to all participants or privately to specific participants as appropriate to probe for additional details. The moderator of the discussion board periodically met with the research team to review the discussion board and to guide moderation. The study team members could freely view the web-based bulletin board discussion and communicate with the moderator to guide probes. Transcripts of the written discussion questions, answers, and follow-up questions were recorded for qualitative analysis.

Data Analysis

Statistical Analysis

Survey data were exported as an encrypted SPSS (IBM Corp) file and transferred to the research team through a secure file-sharing service. We used descriptive statistics to summarize respondent characteristics (both adult and youth) and to summarize responses to key questions around data storage, safety, trust, and use of a digital platform. The baseline characteristics (ie, age, gender, area of residence, level of education, and household income) of caregivers represented by respondents from Insights West and BC Children's Hospital were similar, and therefore, caregiver data from both survey cohorts were amalgamated into a comprehensive "adult" category. The youth cohort was analyzed separately. The adult category was further subcategorized into adults with and without chronic disease and adults with and without a child with chronic disease. To assess the relationship between survey responses to key questions and priori selected sociodemographic variables, we used multivariable proportional odds logistic regression models. Missing data and responses for survey questions were recorded but were not included in statistical analysis. Results were summarized as odds ratios (ORs) and corresponding 95% CIs. Analyses were conducted using R (version 4.0.4; R Foundation for Statistical Computing) and Microsoft Excel.

Qualitative Analysis

Qualitative transcripts were transferred to the research team through secure file sharing. Data gathered from the 23 trust-related questions in the bulletin board discussions were analyzed using an inductive coding approach [21] to identify common themes. Initial codes identified by 2 investigators (AV and HL) were discussed, consolidated, and used to independently analyze all transcripts. All coded data were then systematically reviewed by AV and HL to ensure agreement, after which inductive analysis was used to generate themes and subthemes [21].

Ethical Considerations

Ethics approval was obtained from the University of British Columbia/Children's & Women's Health Centre of British Columbia Research Ethics Board (approval number H20-03105, date of approval 2020-11-26, principal investigator: SA). Implied informed consent was used for surveys and discussion board participation. Findings were reported following the CROSS (Consensus-Based Checklist for Reporting of Survey Studies) checklist [20] for quantitative data and the SRQR (Standards for Reporting Qualitative Research) checklist [22] for reporting qualitative data for the bulletin board results as far as possible. The study data were deidentified with participant identifiers kept separate from survey responses. It is important to transparently state our team's positionality. We come from diverse academic, cultural, and personal backgrounds, including different sex, races, ethnicities, and socioeconomic statuses. We acknowledge that our varying experiences and perspectives shape our approaches to methodology (ie, survey development) and data interpretation (ie, measures of socioeconomic status and comfort or trust with technology), and that privilege and bias may impact our work. We engaged in dialog and critical reflection to navigate these complexities ethically and responsibly to enhance the rigor of our research [23].

Results

Quantitative Survey: Perspectives on Digital Security, Sharing of PHI, and the Value of a Digital Platform

Overview

Out of 1128 caregivers, 685 (60.7%) responded to all questions, with a slightly greater number of caregivers being located in Canada (n=603, 53.5%) versus the United States (n=522, 46.2%). Among 173 youth, 129 (74.5%) responded to all questions. All the youth were from Canada. Among the 1128 caregivers, 231 (20.4%) reported being diagnosed with a chronic health condition, and 198 (17.6%) reported having a child diagnosed with a chronic condition. Among the 231 caregivers with a chronic health condition, 69 (29.9%) also had a child with a chronic health condition (Table 1).

XSL•FO
RenderX

Table 1. Demographic characteristics of study participants (n=1301).

Characteristics	Adults (n=1128), n (%)	Youth (n=173), n (%)
Country of residence		
United States	522 (46.3)	__[a]
Canada	603 (53.5)	173 (100)
No response	3 (0.3)	—
Sex		
Male participants	562 (49.8)	95 (54.9)
Female participants	549 (48.7)	73 (42.2)
Other	4 (0.4)	2 (1.2)
Prefer not to answer	10 (0.9)	2 (1.2)
No response	3 (0.3)	1 (0.6)
Area[b]		
Urban	494 (43.8)	94 (54.3)
Suburban	475 (42.1)	53 (30.6)
Rural	137 (12.1)	22 (12.7)
Do not know or prefer not to answer	12 (1.1)	1 (0.6)
No response	10 (0.9)	3 (1.7)
Level of education[c]		
Secondary school or trade school	168 (14.9)	—
University or college (undergraduate degree)	669 (59.3)	—
University or college (graduate degree)	267 (23.7)	—
Prefer not to answer	16 (1.4)	—
No response	8 (0.7)	—
Household income[d]		
<CAD $75,000 (<US $57,000)	326 (28.9)	—
CAD $75,000-$150,000 (US $57,000-US $114,000)	529 (46.9)	—
>CAD $150,000 (>US $114,000)	216 (19.1)	—
Prefer not to answer	55 (4.9)	—
No response	2 (0.2)	—
Age (y)		
18-35	143 (12.7)	—
36-50	660 (58.5)	—
51-65	250 (22.2)	—
>65	52 (4.6)	—
Prefer not to answer	3 (0.3)	—
No response	20 (1.8)	—
Parent with chronic disease[e]		
Yes	231 (20.5)	—
No	766 (67.9)	—
Prefer not to answer	25 (2.2)	—
No response	106 (9.4)	—
Child with chronic disease		

XSL•FO
RenderX

Characteristics	Adults (n=1128), n (%)	Youth (n=173), n (%)
Yes	198 (17.6)	—
No	873 (77.4)	—
Prefer not to answer	21 (1.9)	—
No response	36 (3.2)	—
Families with children (y)		
<5	204 (19.8)	—
5-7	209 (20.3)	—
8-10	280 (27.2)	—
11-13	306 (29.8)	—
14-15	242 (23.5)	—
16-17	240 (23.3)	—

[a]Data not collected.

[b]Area was self-defined by respondents as living in either an urban, suburban, or rural community.

[c]Level of education was defined as either the highest level completed or having some completion or in the process of completing.

[d]Household income was reported in the currency of the survey respondents' country of residence; CAD $1=US $0.76.

[e]Chronic disease or illness was defined as medical conditions that lasted for a prolonged period and required ongoing medical care and lifestyle modifications to manage symptoms, prevent complications, and maintain one's quality of life.

Table 2 shows the perspectives of caregivers and youth, stratified by caregivers with or without a chronic disease and caregivers with or without a child with a chronic disease, on: (1) knowledge about how and where PHI is stored and who has access to it; (2) trust in health care providers (HCPs), governments, and organizations (ie, hospitals) in keeping PHI secure; (3) willingness to share PHI for not-for-profit health research; and (4) the value and use of a digital platform as described in the survey (Methods section). Caregivers and youth had a similar understanding of where PHI is stored and who can access it, as well as similar trust that HCPs, governments, and organizations implement regulations to keep PHI secure. Compared with caregivers, more youth were willing to share their PHI for research and on a digital platform for clinical care. A larger proportion of caregivers with a chronic disease or a child with a chronic disease reported they understood who could access their child's PHI compared with those without a child with a chronic disease. Further, among caregivers, having a child with a chronic disease resulted in greater trust that PHI is secure, more willingness to share data for research, and greater agreement that storing their child's PHI on a digital platform would be a positive change.

When asked about their level of concern regarding web data privacy and security, almost all caregivers (1039/1128, 92.11%), youth (160/173, 92.5%), caregivers with a chronic disease (218/231, 94.4%), and caregivers of a child with a chronic disease (181/198, 91.4%) reported that they were *very* or *somewhat* concerned. Caregivers living with a chronic disease represented the respondent group with the highest level of concern, with 76.2% (176/231) reporting being *very concerned*. Among respondents' selection of their top 3 choices of security processes that would make digital platforms more trustable, more than half of all caregivers indicated that features, such as multifactor authentication (539/1128, 47.78%) and notification of account changes and activity (504/1128, 44.68%) would improve their trust in a digital platform. These were closely followed by safety features such as using a trusted sign-in partner, for example, a banking or government services account (414/1128, 36.7%) and strong minimum password strength requirements (412/1128, 36.52%). The same 4 security features were most important for the youth (Multimedia Appendix 4).

XSL•FO
RenderX

Table 2. Perspectives of participants on aspects of data security and anticipated platform use and value (n=1301).

	All adults (agree; n=1128), n (%)	All youth (agree; n=173), n (%)	Adults with a chronic disease (agree; n=231), n (%)	Adults without a chronic disease (agree; n=766), n (%)	Caregivers of a child with a chronic disease (agree; n=198), n (%)	Caregivers of a child without a chronic disease (agree; n=873), n (%)
Data storage						
I have a clear understanding of how and where my/my child's PHI[a] is stored.	765 (67.8)	125 (72.2)	180 (77.9)	512 (66.8)	134 (67.7)	592 (67.8)
I have a clear understanding of who can access my/my child's PHI.	801 (71)	118 (68.2)	189 (81.8)	529 (69.1)	150 (75.8)	612 (70.1)
Trust						
I trust that my HCP[b] will keep my/my child's PHI secure.	976 (86.5)	145 (83.8)	203 (87.9)	654 (85.4)	184 (92.9)	750 (85.9)
I trust there are government regulations and practices in place to ensure that PHI is kept secure.	939 (83.2)	145 (83.8)	206 (89.2)	624 (81.5)	180 (90.9)	720 (82.5)
I trust there are organizational (eg, hospital) regulations and practices in place to ensure that PHI is kept secure.	969 (85.9)	144 (83.2)	208 (90.0)	642 (83.8)	188 (94.9)	736 (84.3)
Data sharing						
I am willing to share my/my child's PHI if it creates progress in nonprofit health research.	816 (72.3)	137 (79.2)	184 (79.7)	523 (68.3)	176 (88.9)	603 (69.1)
Digital platform						
A platform like this would be a positive change in how my/my child's PHI is stored.	790 (70)	138 (79.7)	171 (74.0)	515 (67.2)	169 (85.4)	585 (67.0)
I would share my/my child's PHI on this platform with multiple providers.	690 (61.2)	122 (70.5)	159 (68.8)	435 (56.8)	156 (78.8)	505 (57.8)
I would find it overwhelming to have to keep track of another digital account.	687 (60.9)	126 (72.8)	163 (70.6)	459 (59.9)	122 (61.6)	529 (60.6)

[a]PHI: personal health information.

[b]HCP: health care provider.

After being provided with a description of a digital platform (see above in Web-Based Survey Development), respondents were asked if they would be *likely*, *unlikely*, or *undecided* to use this platform. More youth (87/173, 50.3%) than caregivers (465/1128, 41.22%) stated they were *likely* to use the platform; however, a sizable portion of caregivers (407/1128, 36.08%) and youth (63/173, 36.4%) were *undecided*. A higher proportion of caregivers with a chronic condition (130/231, 56.3%) or who have a child with a chronic condition (130/198, 65.7%) responded that they were *likely* to use the described platform. Most caregivers were comfortable sharing their child's PHI on a digital platform, including demographics (694/1128, 61.52%), contact information (639/1128, 56.64%), laboratory test results (770/1128, 68.26%), diagnoses (758/1128, 67.19%), medications, procedures and treatments (780/1128, 69.15%), medical imaging (777/1128, 68.88%), a list of their child's HCPs (787/1128, 69.77%), health habits (such as physical activity and sleep habits; 788/1128, 69.85%), data from applications (751/1128, 66.58%), data from health devices (759/1128, 67.29%), infant feeding habits (760/1128, 67.38%), mental or emotional health (677/1128, 60.02%), immunization records (835/1128, 74.02%), family medical history (746/1128, 66.13%), dental health (850/1128, 75.35%), and allergies (838/1128, 74.29%). Overall, for all types of PHI, more caregivers with a chronic condition or caregivers of a child with a chronic condition were comfortable sharing their child's PHI on a digital platform (Multimedia Appendix 5). Table 3 shows respondents' perspectives on how helpful a digital platform would be for children and youth, caregivers, HCPs, and researchers.

XSL•FO
RenderX

Table 4 outlines the results of ordinal regression analysis for 3 key questions related to trust, willingness to share data for research, and the value of a digital platform.

Table 3. Perspective of participants on helpfulness of an integrated platform to different users ("How helpful do you think an integrated platform as described would be to/for"; n=1301).

Users	All adults (somewhat or very helpful; n=1128), n (%)	All youth (somewhat or very helpful; n=173), n (%)	Adults with chronic disease (somewhat or very helpful; n=231), n (%)	Adults without chronic disease (somewhat or very helpful; n=766), n (%)	Caregivers of a child with a chronic disease (somewhat or very helpful; n=198), n (%)	Caregivers of child without a chronic disease (somewhat or very helpful; n=873), n (%)
Children and youth	920 (81.5)	153 (88.4)	203 (87.9)	612 (79.9)	172 (86.9)	704 (80.6)
Parents and guardians	970 (86)	151 (87.3)	209 (90.5)	642 (83.8)	185 (93.4)	739 (84.7)
Doctors and other health care professionals	1023 (90.7)	162 (93.6)	215 (93.1)	689 (89.9)	193 (97.5)	789 (90.4)
Researchers	1013 (89.8)	154 (89.0)	210 (90.9)	681 (89.9)	192 (97.0)	774 (88.7)

XSL•FO
RenderX

Table 4. Ordinal regression analysisa: trust, willingness to share data, and likelihood of using the (described) digital platform.

	Q1: Level of concern around online data privacy[b] (OR[c]<1 indicates lower level of concern; OR>1 indicates higher level of concern), odds ratio (95% CI)	Q2: Willing to share data for research[d] (OR <1 indicates higher level of agreement; OR>1 indicates lower level of agreement), odds ratio (95% CI)	Q3: Likelihood of using a platform like this[e] (OR<1 indicates more likely; OR>1 indicates less likely), odds ratio (95% CI)
Age (y)			
18-35	1.18 (0.84-1.66)	0.85 (0.60-1.21)	0.63 (0.45-0.89)
36-50	1.00 (reference)	1.00 (reference)	1.00 (reference)
51-65	1.35 (1.03-1.77)	1.05 (0.79-1.39)	1.64 (1.25-2.17)
>65	1.91 (1.11-3.33)	0.72 (0.41-1.29)	2.39 (1.39-4.10)
Sex			
Male	1.00 (reference)	1.00 (reference)	1.00 (reference)
Female	0.81 (0.65-1.03)	0.86 (0.68-1.09)	1.31 (1.05-1.65)
Country			
Canada	1.00 (reference)	1.00 (reference)	1.00 (reference)
United States	1.27 (1.00-1.60)	1.06 (0.83-1.35)	0.67 (0.53-0.85)
Has a chronic disease			
Yes	1.81 (1.35-2.44)	0.71 (0.53-0.97)	0.63 (0.47-0.84)
No	1.00 (reference)	1.00 (reference)	1.00 (reference)
Child has a chronic disease			
Yes	1.42 (0.96-2.11)	0.51 (0.34-0.77)	0.34 (0.23-0.50)
No	1.00 (reference)	1.00 (reference)	1.00 (reference)
Area			
Urban	1.00 (reference)	1.00 (reference)	1.00 (reference)
Suburban	0.72 (0.56-0.92)	1.32 (1.03-1.69)	1.57 (1.23-2.01)
Rural	0.66 (0.46-0.95)	1.41 (0.97-2.04)	1.58 (1.11-2.26)
Education			
Secondary or trade school	1.00 (reference)	1.00 (reference)	1.00 (reference)
Undergraduate	1.82 (1.30-2.55)	0.88 (0.63-1.22)	1.11 (0.80-1.54)
Graduate	2.50 (1.68-3.73)	0.93 (0.62-1.38)	1.03 (0.70-1.53)
Income			
<CAD $75,000 (<US $57,000)	0.77 (0.58-1.02)	0.75 (0.57-1.00)	0.59 (0.45-0.78)
CAD $75,000-$150,000 (US $57,000-US $114,000)	1.00 (reference)	1.00 (reference)	1.00 (reference)
>CAD $150,000 (>US $114,000)	0.69 (0.48-0.98)	0.63 (0.44-0.92)	1.60 (0.42-0.85)

[a]Results of the ordinal regression are from a multivariate analysis. Each question included the same variables in their analysis.

[b]"In general, what is your level of concern regarding data privacy and security issues (eg, personal data being hacked, or companies such as Google or Facebook tracking your activities) when you are engaging in online activity?" (very unconcerned>very concerned).

[c]OR: odds ratio.

[d]"I'm willing to share some of my child's/children's health information confidentially if it helps create progress in nonprofit health research" (strongly agree>strongly disagree).

[e]"What is the likelihood that you would use a platform like this?" (very likely>very unlikely).

Trust

(Question: "In general, what is your level of concern regarding data privacy and security issues when you are engaging in online activity?"; OR<1=lower level of concern.) Caregivers living in suburban (OR 0.72, 95% CI 0.56-0.92) or rural areas (OR 0.66, 95% CI 0.46-0.95) were less likely to report a concern about web data privacy and security, compared with caregivers living

XSL•FO
RenderX

in urban areas. In addition, those who completed an undergraduate degree (OR 1.82, 95% CI 1.3-2.55) and graduate degree (OR 2.5, 95% CI 1.68-3.73) compared with those who only completed secondary or trade school had higher odds of reporting concern regarding data privacy and security. Caregivers living with a chronic disease had higher odds of reporting concern (OR 1.81, 95% CI 1.35-2.44) than caregivers without a chronic disease.

Sharing Data

(Question: "I'm willing to share some of my child's/children's health information confidentially if it helps create progress in nonprofit health research."; OR<1=higher level of willingness.) Compared with no chronic disease, caregivers living with a chronic disease (OR 0.71, 95% CI 0.53-0.97) or caring for a child with a chronic disease (OR 0.51, 95% CI 0.34-0.77) were more likely to be willing to share PHI for not-for-profit research.

Value of a Digital Platform

(Question: what is the likelihood that you would use a digital platform [as described in the survey]?; OR<1=more likely to use.) Compared with caregivers aged between 36 to 50 years, those aged between 18 to 35 years were more likely (OR 0.63, 95% CI 0.45-0.89), and those aged between 51 to 65 years (OR 1.64, 95% CI 1.25-2.17), and >65 years (OR 2.39, 95% CI 1.39-4.10) were less likely to use the described digital health platform. Compared with respondents located in Canada, those in the United States were more likely to use a digital platform (OR 0.67, 95% CI 0.53-0.85). Moreover, respondents living in suburban (OR 1.57, 95% CI 1.23-2.01) and rural (OR 1.58, 95% CI 1.11-2.26) areas were less likely to use a digital health platform when compared with those living in urban areas. Respondents living with a chronic health condition (OR 0.63, 95% CI 0.47-0.84) and those who have a child with a chronic condition (OR 0.34, 95% CI 0.23-0.5) were more likely to use such a platform.

Qualitative Bulletin Board: Perspectives on Digital Security, Sharing of PHI, and the Value of a Digital Platform

Of the 40 caregivers who expressed interest in participating, 23 (58%) caregivers completed the web-based bulletin board discussion group process. Among the 23 participants, 11 (48%) were caregivers of a child with a chronic disease and 12 (52%) were caregivers with a healthy child.

The most common theme raised in the web-based discussions related to digital security concerns and the fear of a data breach being somehow connected to their child. One subtheme was participants' worry that information they share about their child could be linked to their child in a way that might resurface in the future and impact job prospects or their ability to receive insurance. A second subtheme was parental concern that the information would be used for financial profit and be sold to third parties rather than be used altruistically. As stated during one web-based discussion:

> *[I am] not sure who would be really looking at this info...is it just the doctor that you're dealing with...or can it be looked at by the receptionist that's working at the office and gets ahold of your computer file and can get info on you?*

Another participant stated, "I would worry about data being secure or sold to third parties."

A second major theme was caregivers' recognition of the benefits of digital health platforms. Caregivers were generally open to sharing information and appeared amenable to using the digital platform. Caregivers thought the digital platform would offer several benefits; subthemes of identified potential benefits included: (1) being able to access and share information more easily (test results); (2) saving time and effort coordinating their child's health care; (3) being able to set or receive alerts and reminders for appointments, results, or action; (4) faster access to medical consultations or support (web-based or in person); (5) access to additional resources they might not be aware of; and (6) benefit to the child as they might be more comfortable interacting with the health care team through the digital platform rather than in-person.

As stated by one participant:

> *This is a great idea. The healthcare industry really needs to move into the 21st century. All people should have access to their own health records. As long as it is all secure, I am ok with it being online. I love the idea, especially in this COVID time, meeting my doctor virtually when it makes sense.*

A third major theme was the trustworthiness of individuals and institutions involved with digital health platforms. One key subtheme was the importance of a trusted source of information when learning about the digital health platform. Web-based discussion board participants noted that they would prefer to be told about the digital platform by a trusted source (such as their physician or clinician) with the benefits, ease of use, and security information clearly outlined in any materials provided to them. When asked if they would seek out digital services like applications if their child were diagnosed with a chronic health condition, most stated they would talk to a clinician first as they place much trust in their physician. When asked about the digital platform described in the survey specifically, most reported that they would use it if it were recommended to them by their physician. Yet, if their physician did not recommend it, they might consider using it if other trusted sources (community groups, friends and family, the media) gave it positive feedback. They also might search on the web; however, a recommendation from a trusted source would make them more receptive to trying a new digital platform. When asked if their physician recommended this digital platform, one participant stated:

> *I would be more likely if they recommended it, yes. I would assume they have been educated on the positive aspects of the app and can see how this would benefit the child.*

A second subtheme was trust in the recipients with whom they might share their health data. When asked with whom they were most comfortable sharing their child's PHI, most cited their family physician or other health professionals. Respondents were also willing to share PHI with hospitals, Canadian research

XSL•FO
RenderX

institutions, universities, specialists, their child's school, and government agencies (eg, Health Canada).

The fourth theme generated in the web-based discussions was the tension between the barriers and benefits of additional new digital accounts and technology. Over half of the web-based discussion group participants (13/23, 57%) reported that they would find it overwhelming to keep track of another digital account. Some were concerned that there might be a steep learning curve or that using the platform might be challenging. A few participants explained, "I don't need any more technology in my life."

However, other participants disagreed and stated that once they were over the initial learning curve, adopting this digital platform would allow their child to receive better care coordination, making it worthwhile. They explained that it was easy for the password to be saved on their phone, so they were not concerned about additional tasks required by the digital platform. As explained by one participant:

> *I have a lot of digital accounts, I feel it is the way of the world. I don't think one more would be an issue. Plus, there is a lot of information here, I feel like this is a one-stop shop for all our healthcare-related info. If I had to have multiple healthcare accounts, I would find it overwhelming, but since everything health-related seems to be in one spot, it is quite handy.*

Discussion

Principal Findings

We identified novel associations between sociodemographic factors and trust in digital health applications to share PHI for clinical care delivery and research among caregivers of children and youth aged 16 to 17 years. We found that living in an urban area (vs rural), having an undergraduate or graduate degree (vs secondary or trade school), and having chronic disease experience (vs no chronic disease experience) increased the level of concern regarding data privacy and security. Interestingly, those with chronic disease experience had the highest level of concern yet compared with those without chronic disease experience, were more willing to share PHI for not-for-profit research and were more likely to use a digital platform for clinical care and chronic disease management. Studies have mostly reported on the perspectives of adults' willingness to share personal data for research [6]. Our study adds to this growing literature by reporting on the perspectives of caregivers as delegates of their children on digital trust and digital platform use as it relates to sharing PHI for clinical care delivery and research.

Comparison to Previous Work

Lack of digital trust has notably hindered the widespread adoption of digital health platforms [6,24]. Health care is increasingly being delivered in digital and virtual environments, making strong digital trust and identity necessary to support 2-way interactions between clinicians and patients, and patient participation in research. To unlock the potential of digital health, it is critical to understand the differing perspectives on digital trust of citizens across different sociodemographic groups when designing digital health applications to optimize usability, feasibility, and adoption [25]. Like our study, other studies assessing perceptions of digital services, such as the internet or social media, have demonstrated that individuals with lower levels of education tended to be less concerned with web privacy [26,27], while those with a college or graduate degree were more likely to take additional security measures, such as encrypting emails to protect their privacy [26]. Age also influences the adoption of digital health technologies [28,29] and people's willingness to share personal health data for research, although our study did not show this. In a survey study, Nunes Vilaza et al [6] found that, compared with older individuals (>27 years), younger participants (<27 years) were more willing to share personal data for health research.

We observed that caregivers with chronic disease experience had the highest level of concern for data privacy and security yet were more likely to use a platform like TrustSphere for clinical care and share personal health data for not-for-profit research. Nunes Vilaza et al [6] found no difference between individuals who self-reported having good, very good, or excellent health compared with those who self-reported fair or poor health in their willingness to share PHI for research. Robbins et al [30] examined health application use stratified by self-reported health and chronic disease status. They found that individuals who self-reported very good or excellent health, compared with poor health, were likelier to download health applications. However, no significant difference in downloading health applications was found when examined based on chronic disease status. The results of our study show that trust in digital health may be connected to one's familiarity with the health care system and the challenges that patients face in accessing their PHI.

Studies show that individuals with chronic illness are more likely to make altruistic choices, including participating in clinical research [31,32]. In addition, individuals with chronic disease experiences may view the loss of privacy as worthwhile to progress medical research and to benefit others or future patients [6,33]. Involvement in clinical research and the use of digital health services is often an empowering experience for participants as many gain additional connections to health care professionals or other similar individuals through their participation or access relevant, practical information toward managing their illness [34-36]. Our survey also demonstrates that individuals are strongly motivated to share PHI if it positively impacts their children's or others' lives.

Similar to published literature [6,34], we found that attention to optimizing web privacy and security is critical when developing a new digital health platform. Establishing digital trust can be achieved via internal technological features as well as external validation of the technology by trusted sources. Internal security features include two-factor authentication and encrypted storage of PHI [37]. Our study also identified notification of account changes or activity, using a trusted sign-in partner, and strong minimum password requirements as important in gaining digital trust. External features that enhance digital trust include recommended use by trusted HCPs or health care organizations [38,39] or endorsements by friends, family,

XSL•FO
RenderX

or other patients or caregivers. In addition, research by Graham et al [25] highlights the significance of collaboration between researchers and patients, underscoring the importance of co-design approaches. This collaborative effort enables teams to gain deeper insights into the needs of end users (eg, security features), facilitating the development of interventions that are more aligned with user preferences and expectations [22]. Through iterative co-design processes, there is a potential for enhanced user engagement and overall user experience, ensuring that the mobile applications effectively address its users' needs.

Strengths and Limitations

Our study had both limitations and strengths. First, our survey instrument was not validated and was a voluntary, self-report survey introducing the potential for information, recall, social desirability, and sampling bias. We used Insight West's panel of volunteer participants to mitigate bias and maximize study validity by accessing a large sample size, yet we acknowledge our results should be interpreted with caution. For example, participants self-reported their chronic disease status and people who have inherent distrust in sharing information on the web might have been less likely to participate in our web-based survey. Despite using volunteer panels, our survey sample was overrepresented by participants living in urban or suburban areas and those with higher levels of education and household income. Furthermore, we could not stratify our findings by ethnicity or race or family structure (two-parent vs single-parent homes). Consequently, the validity of our survey results is challenged, potentially affecting the accuracy of estimations regarding relationships between variables and the generalizability of this study's findings to the entire North American population. Second, the regression analysis was exploratory in nature with the possibility of residual confounding, and therefore, results should be interpreted with caution. Third, our sample size for the qualitative bulletin boards was small, limiting our qualitative analyses and ability to triangulate quantitative and qualitative data. As such, these findings cannot be viewed as representative and are only presented in this paper to supplement the quantitative survey findings. Finally, our survey data were gathered at the peak of the COVID-19 pandemic, a period marked by a surge in enthusiasm and adoption of digital health tools out of necessity. Therefore, it is essential not to overlook the potential confounding effect of the pandemic on attitudes toward these technologies at the time of data collection. Study strengths included a mixed methods design and a robust sample size of >1000 respondents from across Canada and the United States, strengthening the generalizability of our results. Further, we included the perspectives of youth aged 16 to 17 years on digital trust, which is understudied.

Conclusions and Future Directions

Our research confirms that there is a willingness among caregivers and youth to use a digital platform like TrustSphere for clinical care delivery and to share their PHI for not-for-profit research. However, perceptions around digital trust vary across sociodemographic groups. Therefore, when designing digital applications, diverse engagement of end users is essential. The results of this study will inform the prioritization of the technological features of TrustSphere's "digital front door" and have validated the importance of engaging end users (patients or caregivers and health care professionals) as early as possible in the iterative co-design of TrustSphere to optimize the value of the digital tool and ultimately enhance digital trust. Broadly, this study provides much-needed guidance to researchers and technology developers on what it takes to overcome the barrier of digital trust that has, to date, impeded the comprehensive uptake of digital health platforms. Additional research is needed to characterize the digital needs of underrepresented or vulnerable groups to ensure that digital health solutions are accessible to all.

Acknowledgments

This work was supported, in part, by the University of British Columbia's Digital Technology Supercluster expansion research grant (AWD-015766; held by SA and WW). MG holds a Michael Smith Health Research BC scholar award (SCH-2020-0494).

The authors wish to thank all participating caregivers and adolescents for sharing their insights. The authors are thankful to the University of British Columbia's Office of the Vice-President, Research and Innovation, BC Children's Hospital Research Institute, BC Children's Hospital Foundation, and Canada's Digital Technology Supercluster for their support and contributions to the success of this project.

Authors' Contributions

EC completed data analysis and led the manuscript preparation. AV, HL, and SP designed data collection tools, performed data analysis, and assisted with manuscript preparation. TvR, FSA, MG, and WW provided input regarding data collection tools and supported manuscript preparation. JB led the data analysis. SA conceived the project, designed the study and data collection methods, supported data analysis and interpretation, and assisted with manuscript preparation.

Conflicts of Interest

MG holds a Michael Smith Health Research BC Scholar Award.

Multimedia Appendix 1
TrustSphere adult survey.

XSL•FO
RenderX

[DOCX File , 65 KB - jmir_v26i1e53657_app1.docx]

Multimedia Appendix 2
TrustSphere youth survey.
[DOCX File , 58 KB - jmir_v26i1e53657_app2.docx]

Multimedia Appendix 3
Web-based discussion board question list.
[DOCX File , 18 KB - jmir_v26i1e53657_app3.docx]

Multimedia Appendix 4
Security processes and mechanisms that would improve trust in a digital platform.
[DOCX File , 23 KB - jmir_v26i1e53657_app4.docx]

Multimedia Appendix 5
Comfort levels in sharing your child's personal health information.
[DOCX File , 32 KB - jmir_v26i1e53657_app5.docx]

References

1. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
2. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12(6):573-576 [FREE Full text] [doi: 10.1370/afm.1713] [Medline: 25384822]
3. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 08;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
4. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023 Jan;29(1):49-58. [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
5. Almoaber B, Amyot D. Barriers to successful health information exchange systems in Canada and the USA: a systematic review. Int J Healthc Inf Syst Inform 2017;12(1):44-63. [doi: 10.4018/IJHISI.2017010104]
6. Nunes Vilaza G, Coyle D, Bardram JE. Public attitudes to digital health research repositories: cross-sectional international survey. J Med Internet Res 2021 Oct 29;23(10):e31294 [FREE Full text] [doi: 10.2196/31294] [Medline: 34714253]
7. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014 Nov 05;14:1144 [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
8. Adjekum A, Ienca M, Vayena E. What is trust? Ethics and risk governance in precision medicine and predictive analytics. OMICS 2017 Dec;21(12):704-710 [FREE Full text] [doi: 10.1089/omi.2017.0156] [Medline: 29257733]
9. Seltzer E, Goldshear J, Guntuku SC, Grande D, Asch DA, Klinger EV, et al. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. BMC Med Inform Decis Mak 2019 Aug 08;19(1):157 [FREE Full text] [doi: 10.1186/s12911-019-0886-9] [Medline: 31395102]
10. Mikles SP, Snyder LE, Kientz JA, Turner AM. Why should I trust you? Supporting the sharing of health data in the interprofessional space of child development. AMIA Annu Symp Proc 2020;2020:840-849 [FREE Full text] [Medline: 33936459]
11. Hardiker NR, Grant MJ. Factors that influence public engagement with eHealth: a literature review. Int J Med Inform 2011 Jan;80(1):1-12 [FREE Full text] [doi: 10.1016/j.ijmedinf.2010.10.017] [Medline: 21112244]
12. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
13. Abelson J, Miller FA, Giacomini M. What does it mean to trust a health system? A qualitative study of Canadian health care values. Health Policy 2009 Jun;91(1):63-70. [doi: 10.1016/j.healthpol.2008.11.006] [Medline: 19117635]
14. Kelton K, Fleischmann KR, Wallace WA. Trust in digital information. J Am Soc Inf Sci 2007 Dec 04;59(3):363-374. [doi: 10.1002/asi.20722]
15. Gordon W. Growing use and confidence in artificial intelligence for care delivery. NEJM Catalyst 2022 Mar 16;3(4). [doi: 10.1056/cat.22.0095]
16. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of patients about artificial intelligence in health care. JAMA Netw Open 2022 May 02;5(5):e2210309 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.10309] [Medline: 35507346]
17. TrustSphere. URL: https://www.bcchr.ca/TrustSphere [accessed 2024-06-23]

XSL•FO
RenderX

18. Patterson CC, Karuranga S, Salpea P, Saeedi P, Dahlquist G, Soltesz G, et al. Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract 2019 Nov;157:107842. [doi: 10.1016/j.diabres.2019.107842] [Medline: 31518658]
19. Tauschmann M, Forlenza G, Hood K, Cardona-Hernandez R, Giani E, Hendrieckx C, et al. ISPAD clinical practice consensus guidelines 2022: diabetes technologies: glucose monitoring. Pediatr Diabetes 2022 Dec;23(8):1390-1405 [FREE Full text] [doi: 10.1111/pedi.13451] [Medline: 36537528]
20. Laffel LM, Kanapka LG, Beck RW, Bergamo K, Clements MA, Criego A, et al. Effect of continuous glucose monitoring on glycemic control in adolescents and young adults with type 1 diabetes: a randomized clinical trial. JAMA 2020 Jun 16;323(23):2388-2396 [FREE Full text] [doi: 10.1001/jama.2020.6940] [Medline: 32543683]
21. Thorne S. Beyond theming: making qualitative studies matter. Nurs Inq 2020 Jan;27(1):e12343. [doi: 10.1111/nin.12343] [Medline: 31990117]
22. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
23. Sharma A, Minh Duc NT, Luu Lam Thang T, Nam NH, Ng SJ, Abbas KS, et al. A consensus-based checklist for reporting of survey studies (CROSS). J Gen Intern Med 2021 Oct;36(10):3179-3187 [FREE Full text] [doi: 10.1007/s11606-021-06737-1] [Medline: 33886027]
24. van Kessel R, Hrzic R, O'Nuallain E, Weir E, Wong BL, Anderson M, et al. Digital health paradox: international policy perspectives to address increased health inequalities for people living with disabilities. J Med Internet Res 2022 Feb 22;24(2):e33819 [FREE Full text] [doi: 10.2196/33819] [Medline: 35191848]
25. Graham AK, Wildes JE, Reddy M, Munson SA, Barr Taylor C, Mohr DC. User-centered design for technology-enabled services for eating disorders. Int J Eat Disord 2019 Oct;52(10):1095-1107 [FREE Full text] [doi: 10.1002/eat.23130] [Medline: 31313370]
26. Blank G, Bolsover G, Dubois E. A new privacy paradox: young people and privacy on social network sites. SSRN J 2014 Apr [FREE Full text] [doi: 10.2139/ssrn.2479938]
27. O'Neil D. Analysis of internet users' level of online privacy concerns. Soc Sci Comput Rev 2001 Feb 01;19(1):17-31. [doi: 10.1177/089443930101900103]
28. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
29. Gordon NP, Hornbrook MC. Differences in access to and preferences for using patient portals and other eHealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. J Med Internet Res 2016 Mar 04;18(3):e50 [FREE Full text] [doi: 10.2196/jmir.5105] [Medline: 26944212]
30. Robbins R, Krebs P, Jagannathan R, Jean-Louis G, Duncan DT. Health app use among US mobile phone users: analysis of trends by chronic disease status. JMIR Mhealth Uhealth 2017 Dec 19;5(12):e197 [FREE Full text] [doi: 10.2196/mhealth.7832] [Medline: 29258981]
31. Seagle H, Jessee P, Nagy MC. Altruism: choices of healthy and chronically ill children. Early Child Dev Care 2002 Dec;172(6):573-583. [doi: 10.1080/03004430215097]
32. Baffoe-Bonnie MS. Lived experience with sickle cell disease: predictors of altruistic participation in clinical research. Soc Sci Med 2022 Nov;313:115353. [doi: 10.1016/j.socscimed.2022.115353] [Medline: 36191388]
33. Nebeker C, Lagare T, Takemoto M, Lewars B, Crist K, Bloss CS, et al. Engaging research participants to inform the ethical conduct of mobile imaging, pervasive sensing, and location tracking research. Transl Behav Med 2016 Dec;6(4):577-586 [FREE Full text] [doi: 10.1007/s13142-016-0426-4] [Medline: 27688250]
34. Taylor ML, Thomas EE, Vitangcol K, Marx W, Campbell KL, Caffery LJ, et al. Digital health experiences reported in chronic disease management: an umbrella review of qualitative studies. J Telemed Telecare 2022 Dec;28(10):705-717. [doi: 10.1177/1357633X221119620] [Medline: 36346938]
35. Estcourt S, Epton J, Epton T, Vaidya B, Daly M. Exploring the motivations of patients with type 2 diabetes to participate in clinical trials: a qualitative analysis. Res Involv Engagem 2016;2:34 [FREE Full text] [doi: 10.1186/s40900-016-0050-y] [Medline: 29507768]
36. Wannheden C, Åberg-Wennerholm M, Dahlberg M, Revenäs Å, Tolf S, Eftimovska E, et al. Digital health technologies enabling partnerships in chronic care management: scoping review. J Med Internet Res 2022 Aug 01;24(8):e38980 [FREE Full text] [doi: 10.2196/38980] [Medline: 35916720]
37. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res 2016;8(3):1560-1580 [FREE Full text] [Medline: 27186282]
38. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]

XSL•FO
RenderX

39. Guendelman S, Broderick A, Mlo H, Gemmill A, Lindeman D. Listening to communities: mixed-method study of the engagement of disadvantaged mothers and pregnant women with digital health technologies. J Med Internet Res 2017 Jul 05;19(7):e240 [FREE Full text] [doi: 10.2196/jmir.7736] [Medline: 28679489]

Abbreviations

CROSS: Consensus-Based Checklist for Reporting of Survey Studies
HCP: health care provider
OR: odds ratio
PHI: personal health information
SRQR: Standards for Reporting Qualitative Research
T1D: type 1 diabetes

Edited by N Cahill; submitted 20.10.23; peer-reviewed by AHY Lee, M Taylor, R Hrzic; comments to author 26.02.24; revised version received 14.04.24; accepted 21.08.24; published 28.10.24.

Please cite as:
Chow E, Virani A, Pinkney S, Abdulhussein FS, van Rooij T, Görges M, Wasserman W, Bone J, Longstaff H, Amed S
Caregiver and Youth Characteristics That Influence Trust in Digital Health Platforms in Pediatric Care: Mixed Methods Study
J Med Internet Res 2024;26:e53657
URL: https://www.jmir.org/2024/1/e53657
doi:10.2196/53657
PMID:

©Eric Chow, Alice Virani, Susan Pinkney, Fatema S Abdulhussein, Tibor van Rooij, Matthias Görges, Wyeth Wasserman, Jeffrey Bone, Holly Longstaff, Shazhan Amed. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Question Answering for Electronic Health Records: Scoping Review of Datasets and Models

Jayetri Bardhan[1], BE, MS; Kirk Roberts[2], BS, MS, PhD; Daisy Zhe Wang[1], BS, MS, PhD

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, United States

[2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

Corresponding Author:
Jayetri Bardhan, BE, MS
Department of Computer and Information Science and Engineering
University of Florida
1889 Museum Rd
Gainesville, FL, 32606
United States
Phone: 1 3528716584
Email: jayetri.bardhan@ufl.edu

Abstract

Background: Question answering (QA) systems for patient-related data can assist both clinicians and patients. They can, for example, assist clinicians in decision-making and enable patients to have a better understanding of their medical history. Substantial amounts of patient data are stored in electronic health records (EHRs), making EHR QA an important research area. Because of the differences in data format and modality, this differs greatly from other medical QA tasks that use medical websites or scientific papers to retrieve answers, making it critical to research EHR QA.

Objective: This study aims to provide a methodological review of existing works on QA for EHRs. The objectives of this study were to identify the existing EHR QA datasets and analyze them, study the state-of-the-art methodologies used in this task, compare the different evaluation metrics used by these state-of-the-art models, and finally elicit the various challenges and the ongoing issues in EHR QA.

Methods: We searched for articles from January 1, 2005, to September 30, 2023, in 4 digital sources, including Google Scholar, ACL Anthology, ACM Digital Library, and PubMed, to collect relevant publications on EHR QA. Our systematic screening process followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. A total of 4111 papers were identified for our study, and after screening based on our inclusion criteria, we obtained 47 papers for further study. The selected studies were then classified into 2 non–mutually exclusive categories depending on their scope: “EHR QA datasets” and “EHR QA models.”

Results: A systematic screening process obtained 47 papers on EHR QA for final review. Out of the 47 papers, 53% (n=25) were about EHR QA datasets, and 79% (n=37) papers were about EHR QA models. It was observed that QA on EHRs is relatively new and unexplored. Most of the works are fairly recent. In addition, it was observed that emrQA is by far the most popular EHR QA dataset, both in terms of citations and usage in other papers. We have classified the EHR QA datasets based on their modality, and we have inferred that Medical Information Mart for Intensive Care (MIMIC-III) and the National Natural Language Processing Clinical Challenges datasets (ie, n2c2 datasets) are the most popular EHR databases and corpuses used in EHR QA. Furthermore, we identified the different models used in EHR QA along with the evaluation metrics used for these models.

Conclusions: EHR QA research faces multiple challenges, such as the limited availability of clinical annotations, concept normalization in EHR QA, and challenges faced in generating realistic EHR QA datasets. There are still many gaps in research that motivate further work. This study will assist future researchers in focusing on areas of EHR QA that have possible future research directions.

(J Med Internet Res 2024;26:e53636) doi:10.2196/53636

KEYWORDS

medical question answering; electronic health record; EHR; electronic medical records; EMR; relational database; knowledge graph

XSL•FO
RenderX

Introduction

Motivation

Medical question answering (QA) may use biomedical journals, internet articles, and patient-specific data, such as that stored in the electronic health record (EHR), for QA. While there has been a great deal of work in medical QA [1-5], much of it does not help to answer patient-specific questions. In patient-specific QA, the answer is obtained from the patient's medical record (ie, the EHR). This differs from other medical QA tasks due to linguistic issues (eg, EHR notes are very different in terminology, grammar, style, and structure from biomedical articles) and privacy limitations (eg, most biomedical articles have a publicly available abstract while there are laws in most countries limiting the sharing of patient records). In addition, patient-specific QA also prevents the use of many common QA techniques (such as aggregating answers from different biomedical articles to give weight to a consensus opinion). All this merits the review of EHR QA separate from other medical QA approaches to properly scope its data and methods. In this review paper, our aim is to discuss all the recent approaches and methodologies used for QA on EHRs. There have been some reviews on medical QA [6,7], but none of the previous review papers have focused solely on EHR QA. To the best of our knowledge, this is the first work that does a scoping review of QA on EHRs and examines the various datasets and methodologies used in EHR QA. There are several aspects of EHR QA that merit analysis of scope.

One such aspect is data modality and the variety of methodological approaches available for EHR QA. The methodological approach used is determined by the format of the EHR data. EHRs contain structured and unstructured data. Structured EHR data are based on standardized terminologies and ontologies and are often available in the form of relational databases. By contrast, unstructured EHR data have minimal standardization and include data types such as textual notes and clinical imaging studies. Two kinds of approaches are used for QA on structured EHR data. In the first approach [5], the natural language questions are converted into structured queries (such as SQL). These queries are used to retrieve answers from the database. In the second approach [8], the structured EHR tables are converted into knowledge graphs, following which the natural language questions are converted into graph queries (such as SPARQL) to extract answers from the database. QA on unstructured clinical EHR notes is mostly performed as a reading comprehension task, where given a question and clinical notes as context, a span of text from the notes is returned as the answer. There can also be multimodal EHR QA, which can use both structured and unstructured EHR data for QA. The aim of this study is to identify the studies that use EHR QA. We have further narrowed our search to EHR QA studies that use natural language processing (NLP) techniques on the questions but may or may not use NLP on the answers. We have excluded studies in which questions are asked about images (eg, radiology scans). as these questions and datasets have an entirely different focus. While QA over medical images is also a critical area of research, focusing a systematic review specifically on QA over EHR text (ie, structured and unstructured EHR containing textual information) allows a more detailed, manageable, and methodologically consistent study. This focused approach can yield deeper insights and more practical recommendations for improving QA systems on structured and unstructured data in health care settings.

The second aspect of EHR QA is the access to raw medical data. Due to privacy restrictions on clinical data, the replication and sharing of methods have been reduced compared with QA in other domains. This has led to the emphasis on sharable EHR datasets on which QA benchmarks can be made. Medical Information Mart for Intensive Care (MIMIC; MIMIC-IV) [9] and the eICU Database [10] are large publicly available EHR databases for patients admitted to intensive care units. The MIMIC-III [11] database provides the foundation for many of the existing QA studies on EHRs. MIMIC-IV introduced in the year 2020 is a recent update to the MIMIC-III database. Finally, the National NLP clinical challenges (n2c2) datasets (previously known as Informatics for Integrating Biology and the Bedside ie, i2b2 datasets) are another repository of clinical notes that have been used by the clinical QA community to develop EHR QA datasets.

Another aspect that warrants a scoping review of EHR QA is to study its different applications, including information extraction, cohort selection, and risk score calculation. For instance, Datta and Roberts [12] used a 2-turn QA approach to extract spatial relations from radiology reports. Similarly, Xiong et al [13] used a QA approach with the help of a machine-reading comprehension (MRC) framework for cohort selection, where every selection criterion is converted into questions using simple rules. For example, the selection criteria "ALCOHOL-ABUSE" is converted to the question "Current alcohol use over weekly recommended limits?" Following this, state-of-the-art MRC models such as Bidirectional Encoder Representations from Transformers (BERT) [14], BiDAF [15], BioBERT [16], NCBI-BERT, RoBERTa [17], and BIMPM [18] are used to match question and passage pairs to select cohorts. Furthermore, Liang et al [19] demonstrate that QA over EHR data can improve risk score calculation.

Finally, EHR QA systems face a variety of challenges ranging from parsing natural language questions to retrieving answers. For structured data, the natural language question needs to be parsed and converted to a structured query which can be used to query, the database. Medical terms from the queries, such as *blood pressure* and *leukemia*, must be normalized into standard ontologies. Clinical text frequently uses acronyms for medical concepts. These abbreviations are often ambiguous (eg, *pt* can refer to the patient or physical therapy) [20] and so must be identified and standardized by the QA system before querying over the EHR database or clinical data. These problems are exacerbated by the fact that the standard NLP approaches to such issues require large amounts of labeled data from the domain of interest. Few such labeled EHR datasets exist. This is because annotating EHR QA datasets requires clinical expertise and is time-consuming. Existing general-domain QA systems provide erroneous results when they are not trained on clinical QA datasets. In addition, most of the data found in EHRs are complex and contain both missing and inconsistent information [21,22], which adds to the difficulty of performing

XSL•FO
RenderX

QA on EHRs. In the Discussion section, we have provided more detailed explanations of the various challenges of using QA on EHRs.

The wide variety of challenges and barriers discussed earlier motivates the need for a systematic scoping review of EHR QA literature. This paper identifies the articles that fall under the scope of EHR QA, identifies the difficult challenges faced in the task, and then enumerates both the data sources and QA methods that have been used to overcome such challenges. Finally, this paper also highlights the open issues in this field that demand future work in EHR QA.

Template-Based Dataset Generation

Before diving into the methodology and results of this review, it is helpful to introduce a common semiautomated approach for building EHR QA datasets, as all large EHR QA datasets use this approach. This also impacts the screening process described in the Methods section. While other methods, such as semantic parsing with grammar-based techniques, exist for generating EHR QA datasets [23,24], template-based dataset generation remains the most widely used approach. In general, large EHR QA datasets are often required to increase the performance of EHR QA models. However, the creation of these datasets necessitates subject expertise. The slot-filling approach to generate template-based datasets is a semiautomated process, and hence very popular. Most of the EHR QA datasets are template-based [5,8,25-27]. The steps to construct template-based QA datasets are illustrated using a flowchart in Figure 1.

Figure 1. Flowchart showing the process of template-based dataset generation. The dotted boundary shows the existing non-question answering (QA) natural language processing (NLP) dataset along with the electronic health record (EHR) data. Question templates (and logical form templates) are constructed based on the schema of the EHR data. Clinical expert annotations of non-QA tasks based on the same EHR data are used to slot-fill placeholders in question templates and generate QA pairs.

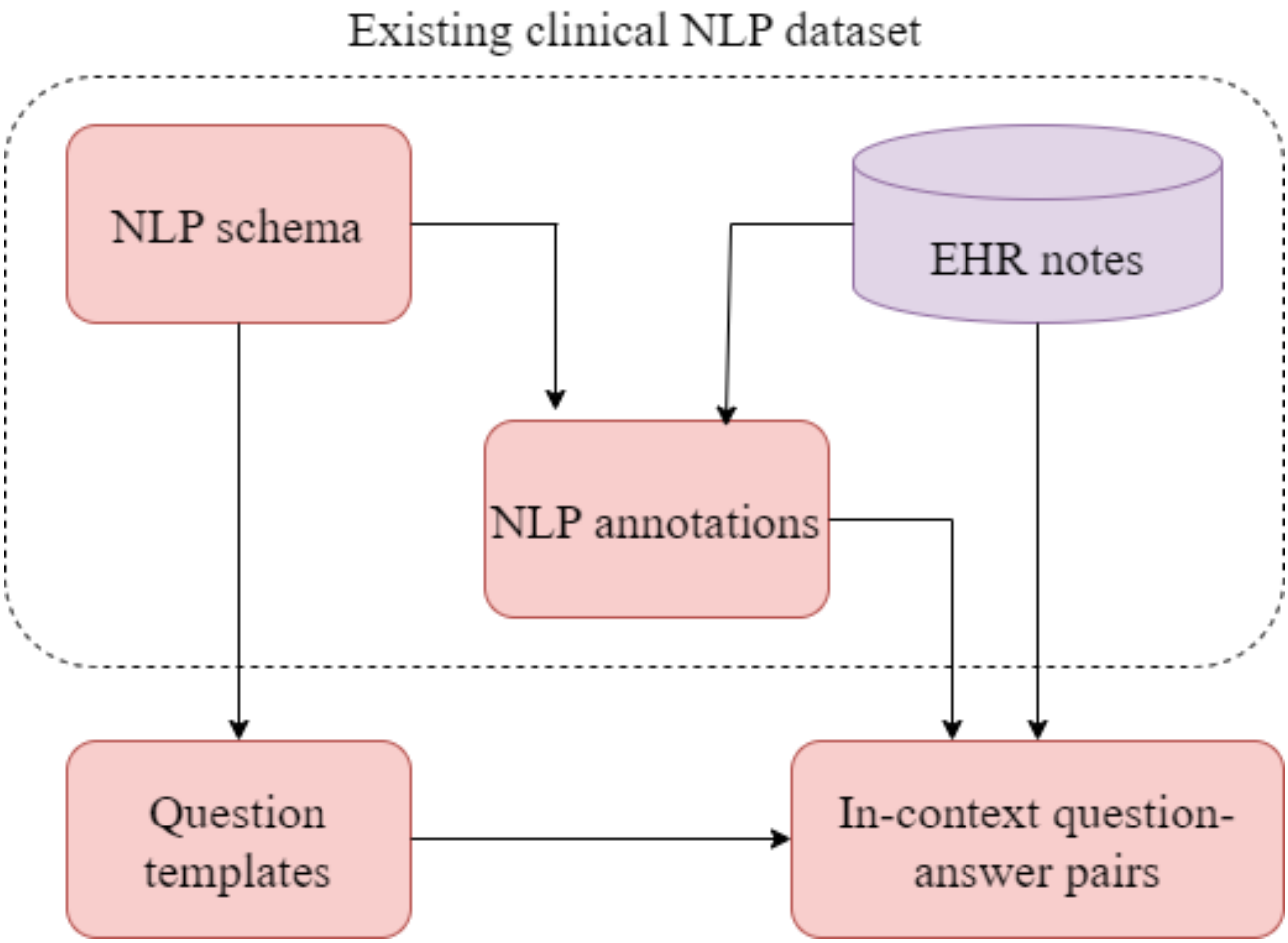

To minimize the need for clinical experts' involvement in the dataset generation process, existing annotations of other non-QA clinical tasks (such as entity recognition and relations learning) are used for generating EHR question-answer pairs. The existing clinical annotations are used as proxy-expert in the dataset generation process [26]. In the first step, template questions containing placeholders (in place of entities) are constructed. An example of a question template is "Has this patient ever been treated with |medication|?" Here, |medication|, |problem|, and |treatment| are some commonly used placeholders. These placeholders in the questions are then slot-filled to obtain QA pairs using the entities in the EHR data and database schema (for a structured EHR database) with the help of the existing annotations from the clinical NLP datasets. So, in a question template, such as "Has this patient ever been treated with |medication|?" entities such as "insulin" and "Tylenol" from the

XSL•FO
RenderX

EHR database and clinical notes (sharing the same entity type as |medication|) are slot-filled in the question template to obtain questions, such as "Has this patient ever been treated with insulin?" and "Has this patient ever been treated with Tylenol?" Following this approach, the RxWhyQA [27] and DrugEHRQA [25] datasets use the existing annotations from the 2018 n2c2 corpus, and the emrQA and emrKBQA datasets use annotations from 6 clinical tasks from the n2c2 repository [28-33].

Some EHR QA datasets, such as emrQA and emrKBQA, have used logical form templates in their template-based generation methods. Logical form templates are predefined structured representations of questions that provide a human-comprehensible symbolic representation, linking questions to answers. These are used to map EHR schema or ontology to represent relations in the questions. While generating these datasets, logical form templates are annotated by clinical experts for different question templates. For example, for the question template "what is the dosage of |medication|?" the annotated logical form template for emrQA is "MedicationEvent (|medication|) [dosage=x]." If more than 1 question template maps to the same logical form template, then they are considered paraphrases of each other. In the emrQA dataset, clinical expert annotations of non-QA tasks, such as entity recognition, relation learning, coreference, and medication challenge annotations (in the n2c2 repository), were used to slot-fill placeholders in question and logical form templates, which in turn were used to generate answers. This is shown in Figure 1. For example, the medication challenge in the n2c2 repository has annotations for medications and their corresponding dosage (eg, medication is nitroglycerin and the dosage is 40 mg). This was used to generate instances of the question "what is the dosage of |medication|?" along with instances of its corresponding logical form "MedicationEvent(|medication|) [dosage=x]." The dosage value, that is, 40 mg is the answer to the question. Similarly, the heart disease challenge dataset contains temporal information and was used to derive temporal-reasoning related question-answer pairs. The emrKBQA dataset used the same question templates and logical form templates of emrQA, which were then slot-filled using entities from the MIMIC-III knowledge base (KB) [11]. The answers of the emrKBQA dataset are present in the table cells of the MIMIC-III KB. The entity types used in the placeholders are test, problem, treatment, medication, and mode. So far, the slot-filling QA dataset generation process has proven to be the most common method of generating EHR QA datasets. This is because, while some manual annotation from domain experts is necessary, most of the process is automated.

Methods

Search Process

This study aims to review existing research on QA over EHRs. This includes papers on EHR QA datasets, QA models, and various approaches proposed over the years. We included papers related to QA in the clinical domain, specifically in EHRs. Papers in which EHRs are not used have been excluded. In this review, we define QA as the task of automatically providing precise, relevant answers to user queries from EHR data. This involves understanding and processing EHR data to extract and deliver specific information. We distinguish QA from broader interactive systems, such as conversational agents, chatbots, and general information retrieval systems, which may involve multiturn dialogue and do not focus solely on providing direct answers to questions. The scope of this review is specifically on structured and unstructured data within EHRs due to the unique challenges and methodologies involved in processing natural language and structured information. While medical images (eg, computed tomography, magnetic resonance imaging, and x-ray) and physical signals (eg, electrocardiograms and photoplethysmography) are critical components of EHRs, the techniques required to analyze these data types differ significantly from those used in structured and unstructured EHR data. Thus, studies focused on these modalities are excluded to maintain a clear and manageable focus on text-focused QA over structured and unstructured EHR data.

Each of the data sources has been queried to search for papers with the title having at least 1 of the following keywords: "clinical," "medical," "patient," "EHR," EMR," "Electronic Health Record(s)," or "Electronic Medical Record(s)." This should be used in combination with one or more of the keywords: "question answering," "questions," "text to SQL," "reading comprehension," "machine comprehension," "machine reading," or "queries." The search was limited to the period from January 1, 2005, to September 30, 2023, to review only recent works. We removed the duplicate studies after this.

Screening Process

We used a 2-step screening process. The first step involved reading the abstracts and titles of all the papers, including only papers that were about EHR QA. We also removed many irrelevant papers that focused on "clinical questions" and "patient questions" but did not use NLP. We also removed non research papers (such as PhD dissertations and books).

In the final stage of screening, a full-text review was used to screen the papers further. Papers that were about query engines and tools and which did not use natural language questions were removed. We excluded papers in languages other than English. We also removed papers that just had an abstract and did not contain full text. There were some papers that were about information retrieval systems not specifically QA. These were also excluded. Furthermore, we have excluded studies in which questions are asked about images or electrocardiograms, as these studies have an entirely different focus. After the 2-stage screening process, we performed forward snowballing that cited the previously included papers on Google Scholar.

For this study, all the authors (JB, KR, and DZW) jointly made the rules for inclusion and exclusion criteria that were used during the paper collection and screening process. On the basis of the rules decided, JB collected the papers and worked on the overall screening process. Papers that were borderline for inclusion were independently screened by KR and then resolved after discussion. The final list made during the full-text review process was again independently screened and reviewed by JB and KR, with conflicts being resolved after discussion.

XSL•FO
RenderX

Results

Search and Screening Results

We have fulfilled all PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) scoping review requirements and have attached a completed copy of the PRISMA checklist in Multimedia Appendix 1. The flowchart for conducting this study is shown in Figure 2.

In this record identification and collection step (ie, the search process), 4285 papers were collected (n=2790 from Google Scholar, n=114 from the ACM Digital Library, n=72 from the ACL Anthology, and n=1309 from PubMed). Following this, we removed the duplicate papers and obtained 4111 papers.

The first step of the screening process, the title and abstract screening step, yielded 126 papers. This was followed by the full-text review step, which yielded 37 papers. After the 2-stage screening process, we performed forward snowballing, adding 10 more papers to the list. We thus obtained 47 studies for EHR QA.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram for study on question answering over electronic health records (EHRs). NLP: natural language processing.

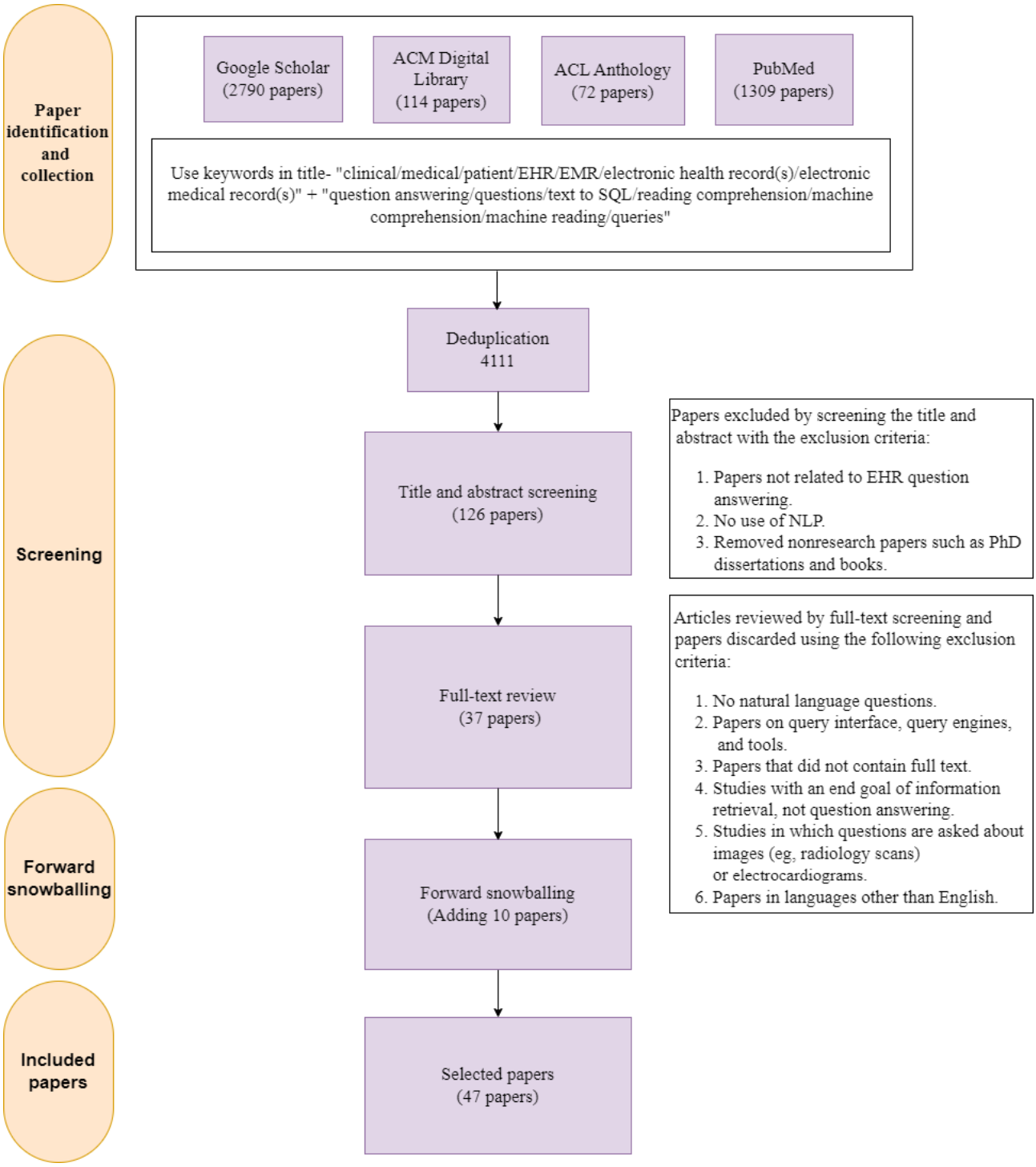

XSL•FO
RenderX

Classification of Selected Papers

This section presents the findings of our study about existing EHR QA papers.

Table 1 lists our final list of selected publications post screening and then classified the papers based on their scope: "EHR QA datasets" and "EHR QA models." We have further classified the studies on EHR QA models based on their function in the QA pipeline. "Full QA" denotes the papers on EHR QA models that are about end-to-end EHR QA systems. In the remaining part of the paper, we have provided our in-depth analysis of studies on QA using EHRs. In Multimedia Appendix 2, we have summarized our final list of selected papers.

Table 1. List of included papers in the systematic review and classification of selected papers based on their scope.

Type of study	References
EHR[a] QA[b] datasets	[5,8,23-27,34-51]
EHR QA models	
Question generation	[43]
Question paraphrasing	[52-54]
Question classification	[55,56]
Full QA	[5,8,25-27,38-42,48-51,57-73]

[a]EHR: electronic health record.

[b]QA: question answering.

Figure 3 illustrates the number of publications on EHR QA over the years. From Figure 3, it can be observed that this is a relatively new field, and most of the publications in this domain are fairly recent. In the following subsections, we discuss our findings on existing EHR QA datasets, the various models used for questioning over EHRs, and also the different evaluation metrics used.

Figure 3. Number of studies on electronic health record question answering over the years. Since this systematic review is conducted based on studies published before September 30, 2023, hence the number of studies shown for the year 2023 is recorded only for a period of 9 months.

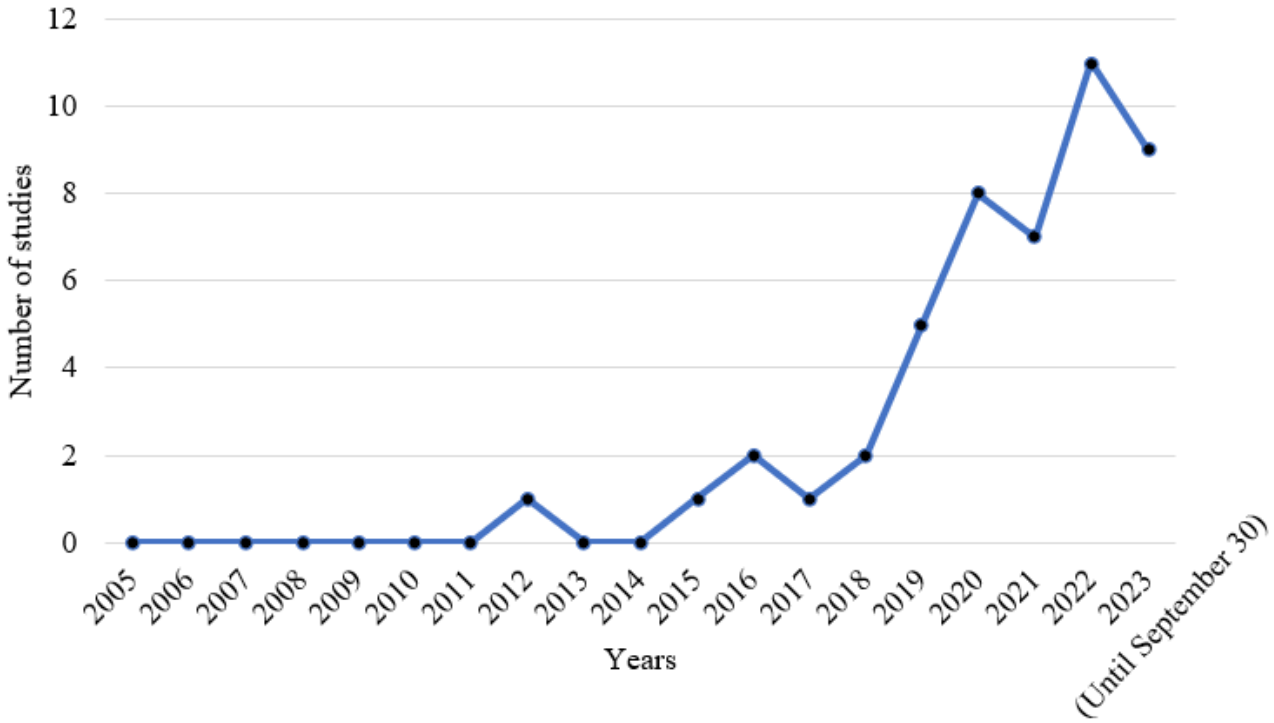

Datasets

Dataset Classification and Analysis

Table 2 displays the total number of citations for all the EHR QA. It also lists the number of studies included in our review that have used these datasets. Moreover, Table 2 classifies the EHR QA based on the accessibility of the datasets. We can observe from the figures that emrQA [26] is the most popular out of all the other EHR QA datasets. This is likely due to emrQA's size (1,295,814 question-logical forms and 455,837 question-answer pairs) and similarity to the Stanford QA dataset (SQuAD)-QA format.

XSL•FO
RenderX

Table 2. Popularity and accessibility of electronic health record (EHR) question answering (QA) datasets. We have listed the number of citations and the number of studies on EHR QA using the dataset. The information presented here is based on the data available on September 30, 2023.

Datasets	Number of citations	Number of studies on EHR QA using the datasets	Publicly available
emrQA [26]	151	11	Yes
MIMICSQL [5]	51	3	Yes
Yue et al [46]	40	0	No
MIMICSPARQL*[a] [41]	27	2	Yes
Yue et al [42]	18	0	Yes
Roberts and Demner-Fushman [23]	s18	3	No
emrKBQA [8]	15	0	No
Raghavan et al [34]	13	0	No
Roberts and Demner-Fushman [24]	10	1	No
Soni et al [44]	7	3	No
Fan [35]	7	0	Yes
DrugEHRQA [25]	5	0	Yes
DiSCQ[b] [43]	6	0	Yes
Oliveira et al [38]	3	0	No
RadQA[c] [37]	3	1	Yes
EHRSQL [36]	3	0	Yes
Kim et al [39]	2	0	Yes
ClinicalKBQA[d] [40]	2	0	No
Hamidi and Roberts [48]	1	0	No
MedAlign [49]	1	0	No
RxWhyQA [27]	0	0	Yes
Mishra et al [45]	0	0	No
CLIFT[e] [47]	0	0	No
Mahbub et al [50]	0	0	No
Dada et al [51]	0	0	No

[a]This dataset follows the original schema of Medical Information Mart for Intensive Care (MIMIC-III).

[b]DiSCQ: Discharge Summary Clinical Questions.

[c]RadQA: Radiology Question Answering Dataset.

[d]ClinicalKBQA: Clinical Knowledge Base Question Answering.

[e]CLIFT: Clinical Shift.

The classification of EHR QA datasets is shown in Figure 4. EHR QA datasets can be unimodal or multimodal. Unimodal EHR QA datasets are based on QA over 1 modality, which can be in the form of structured EHR data or unstructured EHR clinical notes. Multimodal EHR QA datasets use both modalities for QA over EHRs. The DrugEHRQA [25] and MedAlign [49] datasets are examples of multimodal EHR QA datasets that use structured and unstructured EHR data for QA. Figure 5 shows the size and modalities of the different EHR QA datasets.

It is to be noted that the dataset introduced in Mishra et al [45] uses 6 key questions (as can be observed from Figure 5), that is, the same 6 questions have been reused for all the articles. Multimedia Appendix 3 [5,8,23-27,34-51,74] summarizes the existing EHR QA datasets. The EHR databases or corpora contain answers to the questions. From the table in Multimedia Appendix 3, we can infer that most of the EHR QA datasets on structured EHR data use the MIMIC-III database [5,8,36,39,41], while most of the QA datasets on unstructured data use the n2c2 repository [26,27,35] or the clinical notes of MIMIC-III [37,42,43,45-48].

XSL•FO
RenderX

Figure 4. Classification of electronic health record (EHR) question answering (QA) datasets based on modality [5,8,23-27,34-51,74]. The datasets can be unimodal (based on structured or unstructured EHR data) or multimodal.

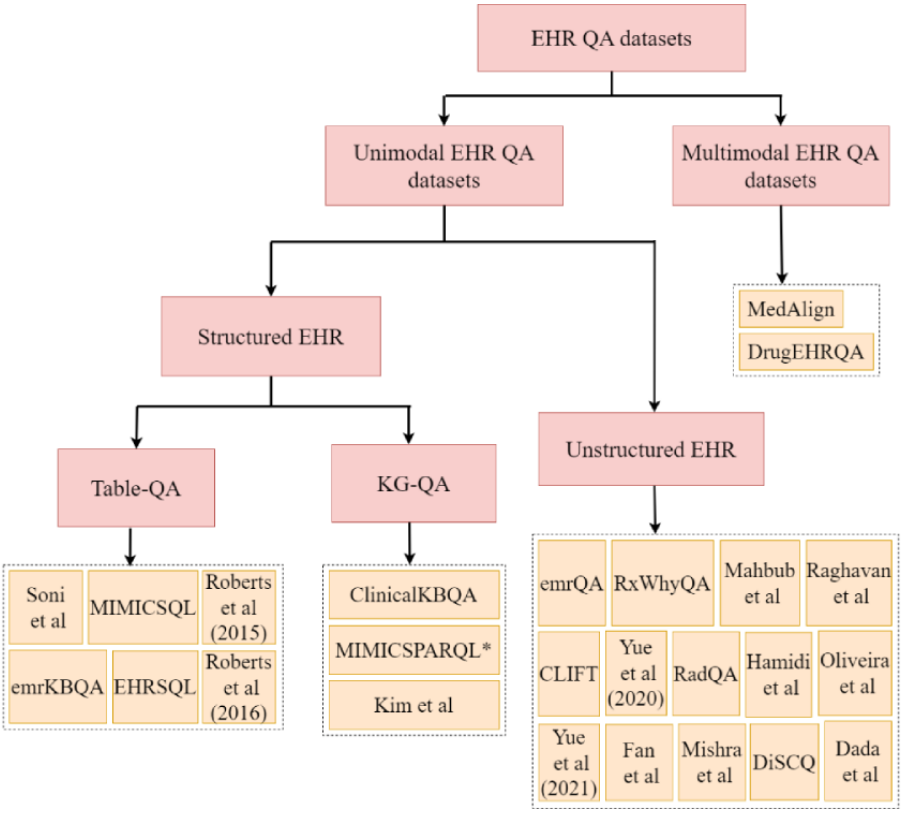

Figure 5. Plot of the total number of questions included in various electronic health record (EHR) question answering (QA) datasets and classification into unstructured, structured, and multimodal EHR QA datasets [5,8,23-27,34-51,74].

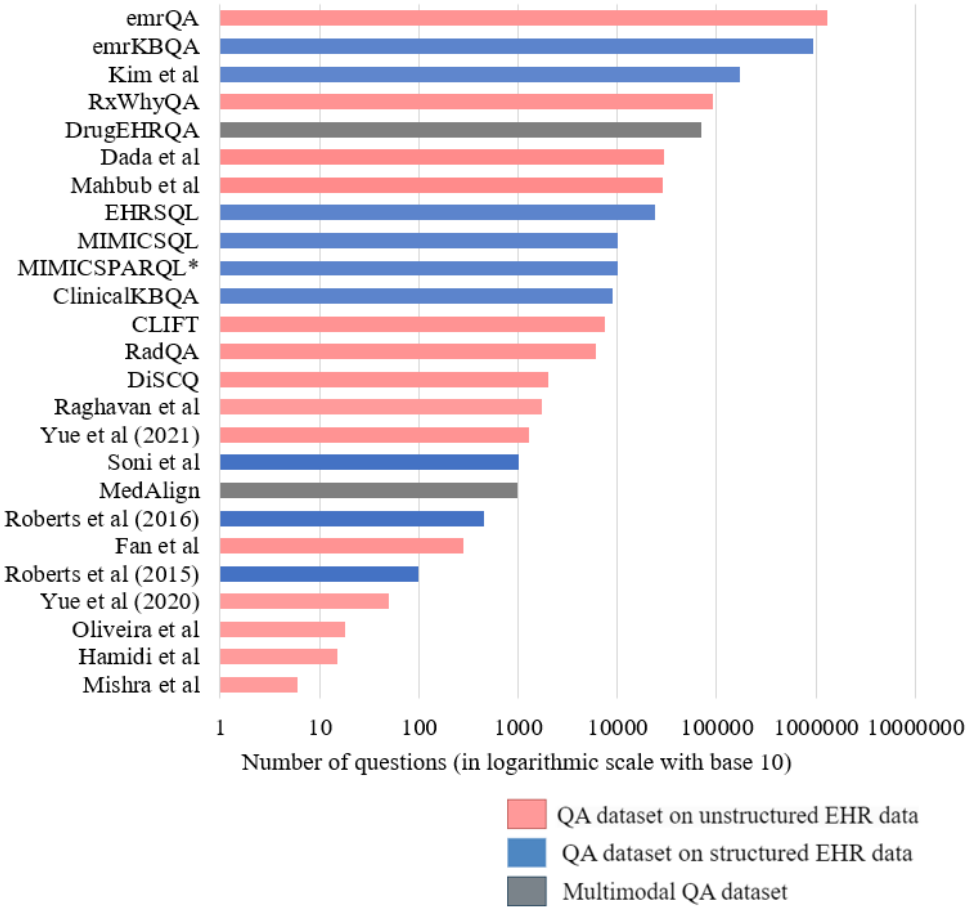

XSL•FO
RenderX

The following sections describe the QA datasets based on unimodal (structured or unstructured) and multimodal EHR data in detail.

QA Datasets Based on Unstructured EHR Data

Unstructured free text EHR data comprises discharge summaries, radiology reports, laboratory reports, medical images, progress notes, and many more note types. It accounts for roughly 80% of all EHR data [75]. One way to make use of this is to create a QA system that can extract answers from unstructured EHR data. Most of the QA datasets on unstructured clinical data are designed for the task of machine comprehension. Given clinical notes (containing patient information) and natural language questions, the objective of these tasks is to retrieve a span of text from the clinical notes as the answer.

The emrQA [26] is the most popular among the EHR QA datasets and contains 455,837 question-answer samples along with 1,295,814 question-logical form pairs. It relies on expertly annotated n2c2 datasets [28-33]. A semiautomatic, template-based process was used to generate the dataset. From Figure 5, we can observe that the emrQA is the largest EHR QA dataset overall.

Despite emrQA's popularity, it has some flaws. The emrQA dataset has attempted to simulate clinicians' questions using predefined templates and generating QA datasets by slot-filling with entities. Consequently, the questions in the emrQA dataset are not very realistic or relevant to the medical community. They are also highly repetitive. For example, it was shown in Yue et al [46] that the same model performance was obtained by sampling 5% to 20% of the dataset as with the entire dataset. This makes it necessary to create datasets that are more realistic and closer to real physicians' questions. Later, Yue et al [42] developed 975 human-verified questions along with 312 human-generated questions based on 36 discharge summaries from MIMIC-III's clinical notes. After randomly sampling 100 questions individually, the 975 human-verified questions and 312 human-generated questions, it was learned that 96% of the human-verified questions were obtained from the emrQA's templates, and 54% of the human-generated questions of Yue et al [42] used the same templates from emrQA.

The RxWhyQA dataset [27] and Fan [35] dataset have reasoning-based questions. The RxWhyQA dataset contains a combination of reasoning-based unanswerable and multi-answer questions. Similar to the emrQA dataset, RxWhyQA is also a template-based dataset and hence not very realistic. This made it necessary to create datasets that are more realistic and closer to real physicians' questions. The Discharge Summary Clinical Questions dataset [43] was created to address this issue and included questions about clinically relevant problems by gathering questions that clinicians could ask. It includes 2029 questions and >1000 triggers based on MIMIC-III discharge reports.

Most of the QA on unstructured EHR datasets is based on discharge summaries [26,27,35,43,45,74]. RadQA [37] and Dada et al [51] are the only 2 QA datasets that use radiology reports for QA. The types of questions used in the EHR QA datasets vary greatly from one another. emrQA covers different types of questions, including factual ("what" and "show me"), reasoning ("how" and "why"), and class prediction ("is" and "has"). However, the distribution of questions for the emrQA dataset is skewed; that is, most of the questions in the emrQA dataset start with "what." In comparison, the authors of RadQA claim that the questions in their dataset are more evenly distributed than emrQA. The RxWhyQA dataset [27] and Fan [35] are reasoning-based questions, and hence their questions have "why cues." Raghavan et al [34] predominantly have temporal questions along with questions on presence or absence (ie, "yes" or "no" questions) as well as questions on medications, tests, and procedures. Mishra et al [45], by contrast, restrict themselves to diagnosis-related questions. Table 3 compares some of the EHR QA datasets using unstructured EHR data for QA. Out of the 14 QA datasets on unstructured EHR notes, only 4 of them (RadQA [37], RxWhyQA [27], Hamidi and Roberts [48], and Dada et al [51]) contain unanswered questions.

XSL•FO
RenderX

Table 3. Comparison of different electronic health record (EHR) question answering (QA) datasets on unstructured data.

Dataset	Mode of dataset generation	Total questions, n	Unanswered questions, n	Average question length (tokens, n)	Total articles, n	Average article length (tokens, n)
emrQA [26]	Semiautomatically generated	1,295,814	0	8.6	2425	3825
RxWhyQA [27]	Automatically derived from the n2c2[b] 2018 ADEs[c] NLP[d] challenge	96,939	46,278	__[a]	505	—
Raghavan et al [34]	Human-generated (medical students)	1747	0	—	71	—
Fan [35]	Human-generated (author)	245	0	—	138	—
RadQA[e] [37]	Human-generated (physicians)	6148	1754	8.56	1009	274.49
Oliveira et al [38]	Human-generated (author)	18	0	—	9	—
Yue et al [42,74]	Trained question generation model paired with a human-in-the-loop	1287	0	8.7	36	2644
DiSCQ[f] [43]	Human-generated (medical experts)	2029	0	4.4	114	1481
Mishra et al [45]	Semiautomatically generated	6 questions or article	—	—	568	—
Yue et al [46]	Human-generated (medical experts)	50	0	—	—	—
CLIFT[g] [47]	Validated by human experts	7500	0	6.42, 8.31, 7.61, 7.19, and 8.40 for smoke, heart, medication, obesity, and cancer datasets	—	217.33, 234.18, 215.49, 212.88, and 210.16 for smoke, heart, medication, obesity, and cancer datasets, respectively
Hamidi and Roberts [48]	Human-generated	15	5	—	—	—
Mahbub et al [50]	Combination of manual exploration and rule-based NLP methods	28,855	—	6.22	2336	1003.98
Dada et al [51]	Human-generated (medical student assistants)	29,273	Unanswered questions available	—	1223	—

[a]Not applicable.

[b]n2c2: natural language processing clinical challenges.

[c]ADE: adverse drug events.

[d]NLP: natural language processing.

[e]RadQA: Radiology Question Answering Dataset.

[f]DiSCQ: Discharge Summary Clinical Questions.

[g]CLIFT: Clinical Shift.

QA Datasets Based on Structured EHR Data

EHR tables contain patient information, such as diagnoses, medications prescribed, treatments, procedures recommended, laboratory results details, and so on. It also includes a lot of temporal information, such as the date of admission, the date of discharge, and the duration of certain medications. The goal of QA tasks over structured databases is to translate the user's natural language question into a form that can be used to query the database.

The QA task on structured EHRs can be classified into 2 types based on the 2 most common forms of structured data: relational databases and knowledge graphs. The first type of QA task entails converting natural language questions into SQL (or logical form) queries that can be used to query the database. In the other type of approach, the EHR data exist in the form of knowledge graphs containing patient information, and the natural language questions are often converted into SPARQL queries to retrieve the answer. MIMICSQL, emrKBQA, and EHRSQL are examples of datasets that use table-based QA approaches whereas datasets such as Clinical Knowledge Base Question Answering (ClinicalKBQA) and MIMIC-SPARQL* use knowledge graph–based QA approaches.

MIMICSQL [5] is a large dataset used for question-to-SQL query generation tasks in the clinical domain. The MIMICSQL dataset is based on the tables of the MIMIC-III database.

XSL•FO
RenderX

emrKBQA [8] is the counterpart of the emrQA dataset for QA on structured EHRs. It is the largest QA dataset on structured EHR data (Figure 5) and contains 940,000 samples of questions, logical forms, and answers. EHRSQL [36] is a text-to-SQL dataset for 2 publicly available EHR databases—MIMIC-III [11] and eICU [10]. It is the only QA dataset on structured EHR data that contains unanswerable questions. Other QA datasets for structured EHR databases include MIMIC-SPARQL* [41] and ClinicalKBQA [40]. However, unlike previous table-based QA datasets, these are knowledge graph–based QA datasets.

The MIMICSQL dataset [5] was created by making changes to the MIMIC-III database's original schema. In total, 9 tables from the MIMIC-III database were merged into 5 tables to simplify the data structure. The derived tables and schemas were not the same as those found in actual hospitals and nursing homes. Therefore, a model trained on the MIMICSQL dataset will not be able to generalize to a real-world hospital setting. To address this issue, Park et al [41] introduced 2 new datasets—a graph-based EHR QA dataset (MIMIC-SPARQL*) and a table-based EHRQA dataset (MIMICSQL*). This was done to improve the analysis of EHR QA systems and to investigate the performance of each of these datasets. MIMICSQL [5] was modified to create MIMICSQL* to comply with the original MIMIC-III database schema [11]. The graph counterpart of the MIMICSQL* dataset is MIMIC-SPARQL*. Figure 6 compares the 2 datasets—MIMICSQL and MIMICSPARQL* based on the length of the questions and the length of SQL/SPARQL queries.

Figure 6. Average length of questions and SQL/SPARQL Protocol and RDF Query Language (SPARQL) queries (in tokens or words) for MIMICSQL and MIMICSPARQL datasets.

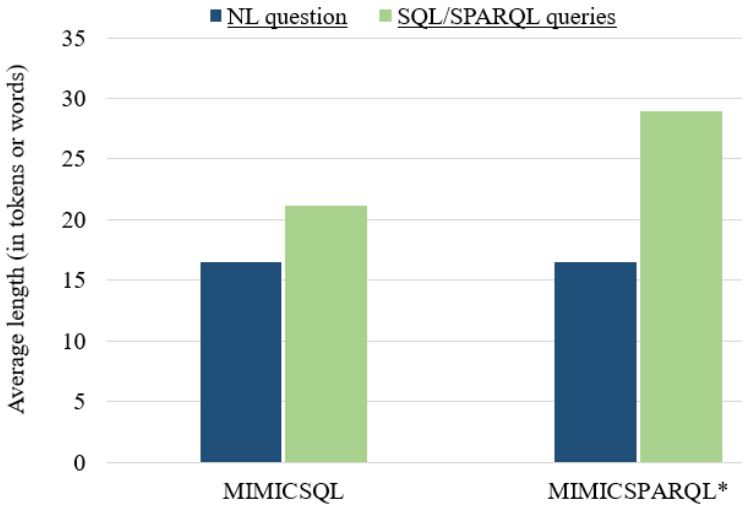

Wang et al [40] generated a clinical knowledge graph (ClinicalKB) with the help of clinical notes of n2c2 annotations and linked different patient information to perform KB QA. Furthermore, Wang et al [40] generated the ClinicalKBQA dataset that can answer statistics-related questions about different patients as well as questions specific to individual patient records.

Roberts and Demner-Fushman [23,24] and Soni et al [44] introduced datasets where logical forms (based on lambda calculus expressions) were created for questions to perform QA on EHR data (known as semantic parsing). Roberts and Demner-Fushman [23,24] generated a bottom-up grammar-based method that generates logical forms for question phrases. Soni et al [44] constructed the question-logical form dataset with the help of the Fast Healthcare Interoperability Resources server.

QA Datasets Based on Multimodal EHR Data

Multimodal QA is QA over >1 modality. QA over >1 modality can help in seeking more accurate answers while taking advantage of >1 source for QA. DrugEHRQA [25] is the first multimodal EHR QA dataset. It uses both structured tables of MIMIC-III and unstructured clinical notes for QA. The DrugEHRQA dataset is a template-based dataset containing medicine-related queries, its corresponding SQL queries for querying over multi-relational EHR tables, the retrieved answer from one or both modalities, as well as the final multimodal answer. The MedAlign dataset [49] also uses structured and unstructured EHR data for QA, but indirectly. The instructions and response pairs of the MedAlign dataset are based on XML markup documents that are derived from structured and unstructured EHR data.

Models and Approaches for QA on EHRs

This section describes the various QA models used in EHRs. QA tasks vary depending on the EHR modality because different information is found in different modalities. Most QA models on clinical notes use a MRC approach, that is, for a given question, the QA model is trained to predict the span of text containing the answer from the clinical note [26,27,38,42,48,50,51,59,65,66]. For QA over EHR tables, translating questions to SQL queries is one of the major approaches used to retrieve answers from the EHR tables [5,61,62,71]. The other approach is to transform the EHR relational database into a knowledge graph and perform a knowledge-graph QA task [39,41,61]. Table 4 summarizes all the QA models (full QA) used for EHRs. Multimedia Appendix 4 [5,8,23,25-27,38-42,44,48-51,57-74] contains further information about these models.

XSL•FO
RenderX

Table 4. Summary of models for question answering (QA) over electronic health records.

Papers	Model
Pampari et al [26]	• For QA task: DrQA's document reader and a multiclass logistic regression model for predicting class. • For question-to-logical form task: a sequence-to-sequence model is used with attention paradigm
Moon et al [27]	• Clinical BERTa model with incremental masking
Oliveira et al [38]	• BioBERTpt
Yue et al [42]	• For QA task: DrQA's DocReader and ClinicalBERT • For question generation task: QPPb module is used with base question generation models (NQGc, NQG++, and BERT-SQGd)
Hamidi and Roberts [48]	• ChatGPT (versions 3.5 and 4), Google Bard, and Claude
Fleming et al [49]	• 6 language models: GPT-4 (32 K tokens+multistep refinement), GPT-4 (32-K tokens), GPT-4 (2K tokens), Vicuña-13B (2K tokens), Vicuña-7B (2K tokens), and Vicuña-7B (2K tokens)
Mahbub et al [50]	• Baseline models: 4 state-of-the-art pretrained language models—BERT, BioBERT, BlueBERT, and ClinicalBERT for QA. • Modeling with transfer learning: sequential learning and adversarial learning
Dada et al [51]	• G-BERT and GM-BERT
Roberts and Patra [57]	• Hybrid semantic parsing method, uses rule-based methods along with a machine learning–based classifier.
Rawat and Li [59]	• Uses multilevel attention layers along with local and global context while answering questions
Rawat et al [60]	• Multitask learning with BERT and ERNIE [76] as the base model
Wen et al [64]	• BERT model trained on different data sources
Soni and Roberts [65]	• BERT, BioBERT, clinical BERT, and XLNet
Mairittha et al [66]	• BERT (large, uncased, whole word masking), BERT fine-tuned on SQuADe benchmark, BioBERT, and an extended BioBERT fine-tuned on unstructured EHR data
Moon et al [67]	• ClinicalBERT model fine-tuned on SQuAD-why dataset
Li et al [68]	• Clinical-Longformer and Clinical-BigBird language model
Yang et al [69]	• GatorTron language model
Lehman et al [73]	• 12 different language models (T5-Base, Clinical-T5-Base-Ckpt, Clinical-T5-Base, RoBERTa-Large, BioClinRoBERTa, GatorTron, T5-Large, Clinical-T5-Large, PubMedGPT, T5-XL, Flan-T5-XXL, and GPT-3)
Kang et al [70]	• KALAf
Wang et al [5]	• TREQSg
Raghavan et al [8]	• Min et al [77] for sequence-to-sequence task along with ParaGen and ParaDetect model
Pan et al [62]	• Medical text-to-SQL model

XSL•FO
RenderX

Papers	Model
Soni and Roberts [63]	• Tranx, Coarse2Fine, transformer, and lexicon-based
Tarbell et al [71]	• T5 language model for question-to-SQL task, along with data augmentation method for back-translation
quEHRy [72]	• End-to-end EHR QA pipeline with concept normalization (MetaMap), time frame classification, semantic parsing, visualization with question understanding, and query module for FHIRh mapping and processing
Kim et al [39]	• Program-based model
Wang et al [40]	• Attention-based aspect reasoning
Park et al [41]	• Seq2Seq model [78] and TREQS [5]
Schwertner et al [58]	• ENSEPROi framework
Bae et al [61]	• Unified encoder-decoder architecture that uses input masking
Bardhan et al [25]	• MultimodalEHRQA

[a]BERT: Bidirectional Encoder Representations from Transformers.
[b]QPP: question phrase prediction.
[c]NQG: Neural Question Generation.
[d]BERT-SQG: BERT-Sequential Question Generation.
[e]SQuAD: Stanford QA dataset.
[f]KALA: Knowledge-Augmented Language model Adaptation.
[g]TREQS: Translate-Edit Model for Question-to-SQL.
[h]FHIR: Fast Healthcare Interoperability Resources.
[i]ENSEPRO: Ensino de Serviços Proativos (in Portuguese), which translates to Teaching Proactive Services.

We can observe from Table 4 that over the years, DrQA's document reader, BERT, and ClinicalBERT are some of the most popular QA models used for unstructured clinical notes [26,27,42,50,60,64-67]. However, since the year 2022, there has been a sharp rise in the number of studies introducing new large language models (besides BERT and other variants of BERT) for MRC tasks [48,68,69,73]. For example, Clinical-Longformer and Clinical-BigBird [68] and GatorTron [69] language models were proposed for various tasks, including EHR QA. Hamidi and Roberts [48] also evaluated the performance of ChatGPT, Google Bard, and Claude for EHR QA. Lehman et al [73] is another study introduced in the year 2023 that evaluated different language models (T5-Base, Clinical-T5-Base-Ckpt, Clinical-T5-Base, RoBERTa-Large, BioClinRoBERTa, GatorTron, T5-Large, Clinical-T5-Large, PubMedGPT, T5-XL, Flan-T5-XXL, and GPT-3) for MRC task on EHR notes.

For QA over structured EHR tables, Translate-Edit Model for Question-to-SQL (TREQS) [5], Medical text-to-SQL (MedTS) [62], and T5 [71] models are used. TREQS [5] is a sequence-to-sequence model that uses a question encoder to convert the questions into vector representations, which are then decoded into SQL queries by the decoder. The generated SQL queries are further edited using an attentive-copying mechanism and recovery mechanism. The MedTS model [62] is another text-to-SQL model that uses a pretrained BERT model as an encoder and a grammar-based long short-term memory (LSTM) decoder to obtain an intermediate sequence. Experiments on the MIMICSQL dataset have shown that the MedTS model outperforms the TREQS model by 22.8% logical form accuracy and by 24.5% execution accuracy. Note that logical form accuracy and execution accuracy are some common evaluation metrics in text-to-SQL tasks. They are explained in detail in the Evaluation Metrics section. Some other examples of table-based QA methods include Tranx [79], Coarse2Fine [80], transformer-based model [63], lexicon-based models [63], quEHRy [72], and sequence-to-sequence tasks used with ParaGen and ParaDetect models [8].

Some models for QA over graph-based EHR are the sequence-to-sequence model [41], TREQS model [41], UniQA model [61], and attention-based aspect reasoning method for KBQA [40]. For most of these models [41,61], the EHR relational database (such as MIMIC-III) is converted into a knowledge graph, and a question-to-SPARQL task is performed to retrieve answers from the knowledge graph. The sequence-to-sequence model [78] uses a bidirectional LSTM as the encoder and uses LSTM decoder while having an attention paradigm. Unlike the TREQS model [5], the sequence-to-sequence model cannot handle out-of-vocabulary words. The UniQA model [61] uses a unified encoder-decoder architecture along with input-masking and value-recovering techniques; thus, it is robust to typos and mistakes in questions. The condition value of the query generated using the question-to-query model is compared with the values in the

XSL•FO
RenderX

database. This is called the condition value recovery technique. ROUGE-L score [81] is used to check the similarity between the values in the database to that of the condition values in the generated query. Then, the condition values are replaced with values most similar to those in the database. After applying the recovery technique, UniQA outperforms both the sequence-to-sequence model (by 74.6% logical form accuracy and 69.2% execution accuracy) and the TREQS model (by 14.2% logical form accuracy and 11.2% execution accuracy).

Most of the existing works discuss only QA on unimodal EHR data. Bardhan et al [25] have proposed a simple pipeline for multimodal QA on EHRs (called MultimodalEHRQA) that uses a modality selection network to choose the modality between structured and unstructured EHR as the preferred modality. If the selected modality obtained is "unstructured text," then QA is performed over the clinical notes using BERT or ClinicalBERT, and the span of text from the clinical notes is returned as the multimodal answer. Similarly, if the preferred modality selected is "structured tables," then a text-to-SQL task is performed using the TREQS model [5]. Further research is still needed to develop a multimodal QA model capable of handling the more challenging task of using answers from both structured and unstructured data to obtain a contextualized answer.

Evaluation Metrics

In this section, we discuss the different evaluation metrics used for EHR. Evaluation metrics are used to evaluate the efficacy of different models. Multimedia Appendix 5 lists the different evaluation metrics used in different EHR QA studies.

The type of QA task would determine the evaluation metrics used. For QA with MRC tasks (eg, in QA over clinical notes), exact match and F_1-score are the most popular metrics for evaluation [26,27,42,46,50,51,60,65,66,68]. Exact match refers to the percentage of predictions that exactly match the ground truth answers. In [26], an exact match is used to determine if the answer entity is included in the evidence. If not, it is determined whether the projected span of evidence is within a few characters of the actual evidence. Precision measures the number of tokens in a prediction that overlap with the correct answer compared with the total number of tokens in the prediction. Recall calculates the proportion of tokens in the correct answer that are included in the prediction compared with the total number of tokens in the correct answer. Precision and recall are represented using equations 1 and 2.

$$\text{Precision}=TP/(TP+FP) \quad \textbf{(1)}$$

$$\text{Recall}=TP/(TP+FN) \quad \textbf{(2)}$$

Where TP, FP, and FN represent true positives, false positives, and false negatives, respectively, at the token level. The F_1-score is a broader metric that calculates the average overlap between the prediction and the correct answer [6]. It is defined as the harmonic mean of precision and recall. This is represented using equation 3. Wen et al [64] and Moon et al [67] used exact match and partial match to assess the QA models for answering questions based on patient-specific clinical text. The F_1-score was used for weighing the partial match between the predicted token of words and the golden token of words. The F_1-score is calculated using the following equation:

$$F1=2\times(\text{Precision}\times\text{Recall}) / (\text{Precision}+\text{Recall}) \quad \textbf{(3)}$$

Evaluation metrics, such as logical form accuracy and execution accuracy, are commonly used for evaluating models responsible for table-based QA that use a question-to-SQL query-based approach [5,62,71]. They are also used for graph-based QA that uses a question-to-SPARQL query-based approach [41,61]. The logical form accuracy is calculated by making a string comparison between the predicted SQL/SPARQL queries and the ground truth queries, and execution accuracy is calculated by obtaining the ratio of the number of generated queries that produce correct answers to the total number of queries [5]. There are instances where execution accuracy might include questions where the generated SQL query is different from the ground truth query, but the returned answer is the same. Structural accuracy is another metric to evaluate models used for question-to-SQL/question-to-SPARQL query tasks [41,61]. Structural accuracy is similar to measuring logical form accuracy, except that it ignores the condition value tokens. Condition value refers to the string value or numeric value in the WHERE part of the SQL/SPARQL query. For example, in the SQL query "SELECT MAX(age) from patients WHERE Gender='F' and DoB>2020," "F" and 2020 are the condition values. The objective of using structural accuracy is to evaluate the accuracy of converting questions to SQL/SPARQL query structures, by not giving importance to the condition values (similar to the Spider dataset [82]). Raghavan et al [8] use exact match and denotation accuracy for evaluating clinical table-QA models. The framework involves the following 2 stages: (1) predicting logical forms for questions and (2) obtaining answers from the database with logical forms as input. Exact match is used for semantic parsing, while denotation accuracy is used to evaluate models for obtaining answers from logical forms. Denotation accuracy checks if the logical forms that are input to the model return the correct label answer, and the exact match is used to check if the logical forms generated are the same as the ground truth logical forms.

A variety of text-generating metrics have been used to evaluate question paraphrasing. Soni et al [52] used Bilingual Evaluation Understudy (BLEU) [83], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [84], and translation error rate (TER) [85] for evaluating paraphrasing models. The BLEU score evaluates how closely generated paraphrases (or candidate translations) resemble those in the reference. This is done with exact token matching. The BLEU score is calculated as follows:

$$\text{BLEU} = \text{brevity penalty}\times\exp\left(\sum w_n \log p_n\right) \quad \textbf{(4)}$$

where,

$$\text{Bravity penalty}=\min(1, \exp(1-\text{reference length}/\text{output length})) \quad \textbf{(5)}$$

$$p_n=\text{total number of candidate n-grams}/\text{total number of matched n-grams} \quad \textbf{(6)}$$

In equation 4, w_n represents the weight for each n-gram. The METEOR score, by contrast, uses synonyms and word stems. This is represented using the following equations:

XSL•FO
RenderX

$$METEOR = F \times (1 - Penalty) \quad \textbf{(7)}$$

where,

$$F = (Precision \times Recall)/(\alpha \times Precision + (1 - \alpha) \times Recall) \quad \textbf{(8)}$$

$$Penalty = \gamma \times (ch/m)\beta \quad \textbf{(9)}$$

In equation 9, "ch" is the number of chunks that match, and m is the number of uniforms that match between the prediction and the reference. The parameters α, β, and γ are adjusted to maximize the correlation with human judgments. The edit distance (the number of edits necessary to change one sentence into another) between generated and reference paraphrases is measured by the TER score. It is calculated by adding up all the edits, dividing that total by the number of words, and multiplying that result by 100, that is,

$$TER = (\text{number of edits} / \text{average number of reference words}) \times 100 \quad \textbf{(10)}$$

Discussion

Challenges and Existing Solutions

Limited Number of Clinical Annotations for Constructing EHR QA Datasets

There are very few clinical EHR annotations that are publicly available. The n2c2 repository is one of the very few public repositories that hosts EHR NLP datasets (that can be used to create template-based QA datasets). This is because creating these annotations requires a lot of manual work, which can be time-consuming, and at the same time requires domain knowledge [25,26]. For the same reasons, it was difficult to annotate EHR QA datasets. There are also some ethical issues and privacy concerns that need to be handled while constructing EHR QA datasets. This involves the deidentification of information related to patients.

Datasets such as emrQA [26] and ClinicalKBQA [40] are examples of template-based datasets that have used the available expert annotations of the n2c2 repository to generate large-scale patient-specific QA datasets using semiautomated methods, taking advantage of the limited clinical annotations. While questions in these datasets do not represent the true distribution of questions one would ask to EHR, their scale makes them valuable for transfer learning and method development.

Concept Normalization in Clinical QA

QA in any domain has its own challenges. However, clinical QA has added challenges. One major challenge is when different phrases are used for the same medical concept in the question and the database. Clinical normalization is used to deal with this issue. Clinical normalization involves recognizing the medical entities and terminologies and converting them into a singular clinical terminology or language. Many EHR QA datasets, such as emrQA, have used MetaMap [86] during the dataset generation process to map medical terminologies mentioned in the clinical text to the Unified Medical Language System (UMLS) Metathesaurus. However, it has been argued that concept normalization for EHR QA is fundamentally different than the task on clinical notes [72], so QA-specific datasets are clearly needed.

Generating Realistic EHR QA Datasets

It is necessary to ensure that questions in EHR QA datasets contain realistic questions that clinicians and patients would want answered from EHR data. To create realistic questions while constructing the EHRSQL dataset [36], a poll was created at a hospital to gather real-world questions that are frequently asked on the structured EHR data. The Discharge Summary Clinical Questions dataset [43] also included clinically relevant questions by collecting questions that physicians could ask. This ensured the use of medically relevant questions in the EHR QA datasets.

Adding more paraphrases to the QA dataset is another manner to ensure the questions are realistic. This is because, in a real-world scenario, the same question may be posed or stated in different manners. Generation of paraphrases may be machine-generated, human-generated [26], or it could be a combination of both [36]. Table 5 lists the number of paraphrases used per template in different EHR QA datasets.

Table 5. Summary of paraphrases used in various electronic health record (EHR) question answering (QA) datasets.

Dataset	Paraphrases per question type, mean	Method of generating paraphrases	Number of questions
MIMICSQL [5]	1	Human labor (crowdsourcing)	10,000
emrQA [26]	7	Human labor (templates generated by physicians were slot-filled)	1,000,000
emrKBQA [8]	7.5	Human labor (templates generated by physicians were slot-filled)	940,173
EHRSQL [36]	21	Human labor and machine learning	24,000

Open Issues and Future Work

Redundancy in the Types of Clinical Questions

Most of the existing EHR QA datasets are template-based datasets that are obtained by slot-filling. These datasets have several instances of the same type of templates that are slot-filled with various entities. Therefore, there is redundancy in the diversity of questions generated. This is still an ongoing issue that needs to be addressed.

Need for Multimodal EHR QA Systems

Clinical EHRs contain a vast amount of patient information. Structured EHR data contain highly complementary data that may or may not be present in the clinical notes. The information in structured and unstructured EHR data may contain

XSL•FO
RenderX

information that is similar, may contradict, or can provide additional context between these sources. There is a clear need for EHR QA systems that reason across both types of data.

DrugEHRQA [25] and MedAlign [49] datasets are the only multimodal EHR QA datasets (although MedAlign dataset is technically a pseudo-multimodal EHR QA dataset because the QA pairs of the MedAlign dataset are based on an XML markup that are derived from structured and unstructured EHR data). Bardhan et al [25] introduced a simple baseline QA model for multimodal EHR data, and further research is needed to develop a multimodal QA model that unifies the EHR data modalities to obtain a contextualized answer.

QA of EHRs on Unseen Paraphrased Questions

QA models trained on clinical question-answer pairs when tested on unseen paraphrased questions have historically produced poor results. There have been works that have tried to address this challenge. The model in Raghavan et al [8] uses paraphrasing detection and generation as a supplementary task to handle this issue. Another solution was discussed in Rawat et al [60]. Rawat et al [60] introduced a multitask learning approach where extractive QA and prediction of answer span were the primary tasks, with an auxiliary task of logical form prediction for the questions. However, this is still an ongoing issue that needs further work.

QA of EHRs on Unseen Data

QA models should be able to generalize to new clinical contexts and EHR questions. To study generalization, Yue et al [46] evaluated the performance of a model trained on the emrQA dataset on a new set of questions based on clinical notes of MIMIC-III. The experiment proved that the accuracy of the QA model dropped down by 40% when tested on unseen data. The same research group later proposed a solution [42]. They developed the CliniQG4QA framework, which uses question generation to obtain QA pairs for unseen clinical notes and strengthen QA models without the need for manual annotations. This was done using a sequence-to-sequence-based question phrase prediction model.

This issue was also addressed in question-to-SQL tasks for table-based EHR QA. Tarbell et al [71] introduced the MIMICSQL 2.0 data split (derived from the existing MIMICSQL dataset [5]) to test the generalizability of existing text-to-SQL models on EHRs. The performance of the TREQS [5] model on the MIMICSQL 2.0 data split was drastically poor (logical form accuracy of 0.068 and execution accuracy of 0.173 when trained on paraphrased questions and tested on paraphrased questions), thus showing the need for improvement. To improve generalizability of text-to-SQL tasks on EHR data, Tarbell et al [71] then introduced the use of the T5 model with the data augmentation method using back-translation and further adding out-of-domain training data to improve generalizability on text-to-SQL tasks. The proposed model, even though it outperformed the TREQS model (logical form accuracy of 0.233 and execution accuracy of 0.528 when trained on paraphrased questions and tested on paraphrased questions), still needs further improvement. More work is required in the future to overcome this challenge.

Progress of QA Models in Real Clinical Applications

Integrating QA systems into clinical workflows allows health care practitioners to access current medical information and recommendations, potentially lowering medical errors and improving patient care. Studies are now being conducted on QA models to determine their accuracy, safety, and reliability in clinical settings. These studies are critical for establishing their usefulness in real-world settings [87]. Efforts are underway to create user-friendly interfaces that allow health care providers to communicate more easily. Some QA models are being tested in cohort selection studies [13] and clinical trials to determine their efficacy and safety in real-world contexts. Deploying QA models in clinical contexts involves ethical problems about patient privacy, bias reduction, and transparency in decision-making. Addressing these concerns is critical for establishing acceptance among health care professionals and patients. To summarize, while QA models have considerable benefits for clinical practice and research, their implementation in real-world clinical applications necessitates resolving integration and ethical issues. To completely harness the power of QA models in health care, artificial intelligence researchers and physicians must keep working together.

Strengths

In this study, we presented the first scoping review for QA in EHRs. We methodologically collected and screened papers related to EHR QA from January 1, 2005, to September 30, 2023, and performed a thorough review of the existing studies on EHR QA. Then, we explored all the existing datasets, approaches, and evaluation metrics used in EHR QA. Furthermore, we identified the different modalities for QA over EHRs and described the approaches used for each. We have fulfilled all PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Review) requirements.

This review helps to identify the challenges faced in EHR QA. In addition, this study sheds light on the problems that have been solved along with the additional gaps that are still remaining. This will encourage researchers in this domain to pursue these open problems that have not yet been solved.

Limitations

Despite the strengths of this study, we note a few limitations. First, the search process was limited to a handful of EHR and QA-related keywords. There is a long tail in how these types of systems are described in the literature, but there is a possibility that we might have missed relevant studies that did not match this initial search criteria. We used forward snowballing to partially resolve this issue. This helped us to identify 10 additional papers that we had missed out on earlier. However, despite this, there is still a slim chance that we might have missed a few relevant studies in our final list. Furthermore, given the current expansion of research into EHR QA, we predict that new studies will be added to this list since our search.

Conclusions

In recent years, QA over EHRs has made significant progress. This is the first systematic or scoping review of QA over EHRs.

XSL•FO
RenderX

In this paper, we have provided a detailed review of the different approaches and techniques used for EHR QA. The study began by discussing the need for large domain-specific EHR QA datasets and then discussed the existing EHR QA datasets. We have reviewed the different unimodal EHR QA models used for both structured EHRs and unstructured EHRs, as well as QA models on multimodal EHRs. Then, we identified the major challenges in this field, such as the limited number of clinical annotations available for EHR QA dataset generation. We also talked about potential future directions in this field. It is a relatively new field with many unexplored challenges that require attention. This study should help future researchers explore various research directions within EHR QA and expand the horizons of research areas in this field.

Acknowledgments

This project has been funded by National Institutes of Health (grants R00LM012104, R21EB029575, and R01LM011934).

Authors' Contributions

For this study, JB, KR, and DZW proposed the idea of the study. All the authors jointly made the rules for inclusion and exclusion criteria. JB and KR contributed toward paper collection and the overall screening process and classified the papers based on their scope. JB conducted the initial analysis and drafted the manuscript. The manuscript of the paper was then critically reviewed by KR and DZW. All the authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e53636_app1.docx]

Multimedia Appendix 2
Summaries of selected papers.
[XLSX File (Microsoft Excel File), 24 KB - jmir_v26i1e53636_app2.xlsx]

Multimedia Appendix 3
Comparison of different electronic health record question answering datasets. The column “Database/Corpus” refers to the electronic health record database or clinical annotations on which the question answering datasets are based.
[DOCX File , 26 KB - jmir_v26i1e53636_app3.docx]

Multimedia Appendix 4
Summary of electronic health record question answering (QA) models. The model’s task can be of type machine-reading comprehension (MRC), KBQA, “question to SQL query,” and “question to SPARQL.” The “Answer type” column specifies the expected type of the electronic health record-QA model, indicating whether the answer is derived from tables, text notes, a knowledge graph, or a combination of sources. “Dataset”’ refers to the dataset used to evaluate the QA model.
[DOCX File , 29 KB - jmir_v26i1e53636_app4.docx]

Multimedia Appendix 5
Evaluation metrics used for evaluating different electronic health record-question answering models.
[DOCX File , 18 KB - jmir_v26i1e53636_app5.docx]

References

1. Demner-Fushman D, Mrabet Y, Ben Abacha A. Consumer health information and question answering: helping consumers find answers to their health-related information needs. J Am Med Inform Assoc 2020 Feb 01;27(2):194-201 [FREE Full text] [doi: 10.1093/jamia/ocz152] [Medline: 31592532]
2. Roberts K, Masterton K, Fiszman M, Kilicoglu H, Demner-Fushman D. Annotating question decomposition on complex medical questions. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC). 2014 Presented at: LREC 2014; May 26-31, 2014; Reykjavik, Iceland URL: https://experts.illinois.edu/en/publications/annotating-question-decomposition-on-complex-medical-questions
3. Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ clinical question answering system. AMIA Annu Symp Proc 2011;2011:171-180 [FREE Full text] [Medline: 22195068]

XSL•FO
RenderX

4. Lee M, Cimino J, Zhu HR, Sable C, Shanker V, Ely J, et al. Beyond information retrieval--medical question answering. AMIA Annu Symp Proc 2006;2006:469-473 [FREE Full text] [Medline: 17238385]
5. Wang P, Shi T, Reddy CK. Text-to-SQL generation for question answering on electronic medical records. In: Proceedings of The Web Conference 2020. 2020 Presented at: WWW '20; April 20-24, 2020; Taipei, Taiwan. [doi: 10.1145/3366423.3380120]
6. Mutabazi E, Ni J, Tang G, Cao W. A review on medical textual question answering systems based on deep learning approaches. Appl Sci 2021 Jun 11;11(12):5456. [doi: 10.3390/app11125456]
7. Athenikos SJ, Han H. Biomedical question answering: a survey. Comput Methods Programs Biomed 2010 Jul;99(1):1-24. [doi: 10.1016/j.cmpb.2009.10.003] [Medline: 19913938]
8. Raghavan P, Liang JJ, Mahajan D, Chandra R, Szolovits P. emrKBQA: a clinical knowledge-base question answering dataset. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP 2021; June 11, 2021; Online. [doi: 10.18653/v1/2021.bionlp-1.7]
9. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 03;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
10. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
11. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
12. Datta S, Roberts K. Fine-grained spatial information extraction in radiology as two-turn question answering. Int J Med Inform 2021 Nov 06;158:104628 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104628] [Medline: 34839119]
13. Xiong Y, Peng W, Chen Q, Huang Z, Tang B. A unified machine reading comprehension framework for cohort selection. IEEE J Biomed Health Inform 2022 Jan;26(1):379-387. [doi: 10.1109/JBHI.2021.3095478] [Medline: 34236972]
14. Rosenbloom L, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018 [FREE Full text] [doi: 10.5260/chara.21.2.8]
15. Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. arXiv Preprint posted online on November 5, 2016 [FREE Full text] [doi: 10.48550/arXiv.1611.01603]
16. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
17. Zhuang L, Wayne L, Ya S, Jun Z. A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics. 2021 Presented at: CCL 2021; August 13-15, 2021; Hohhot, China. [doi: 10.1007/978-3-030-84186-7_31]
18. Wang Z, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017 Presented at: IJCAI'17; August 19-25, 2017; Melbourne, Australia. [doi: 10.24963/ijcai.2017/579]
19. Liang JJ, Lehman E, Iyengar A, Mahajan D, Raghavan P, Chang CY, et al. Towards generalizable methods for automating risk score calculation. In: Proceedings of the 21st Workshop on Biomedical Language Processing. 2022 Presented at: BioNLP 2022; May 26, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.bionlp-1.42]
20. Newman-Griffis D, Divita G, Desmet B, Zirikly A, Rosé CP, Fosler-Lussier E. Ambiguity in medical concept normalization: an analysis of types and coverage in electronic health record datasets. J Am Med Inform Assoc 2021 Mar 01;28(3):516-532 [FREE Full text] [doi: 10.1093/jamia/ocaa269] [Medline: 33319905]
21. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013 Dec 17;1(3):1035 [FREE Full text] [doi: 10.13063/2327-9214.1035] [Medline: 25848578]
22. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Netw Open 2021 Feb 01;4(2):e210184 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0184] [Medline: 33635321]
23. Roberts K, Demner-Fushman D. Annotating logical forms for EHR questions. LREC Int Conf Lang Resour Eval 2016 May;2016:3772-3778 [FREE Full text] [Medline: 28503677]
24. Roberts K, Demner-Fushman D. Toward a natural language interface for EHR questions. AMIA Jt Summits Transl Sci Proc 2015 Mar 25;2015:157-161 [FREE Full text] [Medline: 26306260]
25. Bardhan J, Colas A, Roberts K, Wang DZ. DrugEHRQA: a question answering dataset on structured and unstructured electronic health records for medicine related queries. arXiv Preprint posted online on May 3, 2022 [FREE Full text] [doi: 10.48550/arXiv.2205.01290]
26. Pampari A, Raghavan P, Liang J, Peng J. emrQA: a large corpus for question answering on electronic medical records. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP 2018; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1258]

XSL•FO
RenderX

27. Moon S, He H, Jia H, Liu H, Fan JW. Extractive clinical question-answering with multianswer and multifocus questions: data set development and evaluation study. JMIR AI 2023 Jun 20;2:e41818 [FREE Full text] [doi: 10.2196/41818] [Medline: 38875580]
28. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc 2010;17(5):519-523 [FREE Full text] [doi: 10.1136/jamia.2010.004200] [Medline: 20819855]
29. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
30. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1):14-24 [FREE Full text] [doi: 10.1197/jamia.M2408] [Medline: 17947624]
31. Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc 2009;16(4):561-570 [FREE Full text] [doi: 10.1197/jamia.M3115] [Medline: 19390096]
32. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S20-S29 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.020] [Medline: 26319540]
33. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-791 [FREE Full text] [doi: 10.1136/amiajnl-2011-000784] [Medline: 22366294]
34. Raghavan P, Patwardhan S, Liang JJ, Devarakonda MV. Annotating electronic medical records for question answering. arXiv Preprint posted online on May 17, 2018 [FREE Full text] [doi: 10.48550/arXiv.1805.06816]
35. Fan J. Annotating and characterizing clinical sentences with explicit why-QA cues. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP 2019; June 7, 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1913]
36. Lee G, Hwang H, Bae S, Kwon Y, Shin W, Yang S, et al. EHRSQL: a practical text-to-SQL benchmark for electronic health records. In: Proceedings of the 36th Conference on Neural Information Processing Systems. 2022 Presented at: NeurIPS 2022; November 28-December 9, 2022; New Orleans, LA.
37. Soni S, Gudala M, Pajouhi A, Roberts K. RadQA: a question answering dataset to improve comprehension of radiology reports. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022 Presented at: LREC 2022; June 20-25, 2022; Marseille, France.
38. Oliveira LE, Schneider ET, Gumiel YB, Luz MA, Paraiso EC, Moro C. Experiments on Portuguese clinical question answering. In: Proceedings of the 10th Brazilian Conference on Intelligent Systems. 2021 Presented at: BRACIS 2021; November 29-December 3, 2021; Virtual Event. [doi: 10.1007/978-3-030-91699-2_10]
39. Kim D, Bae S, Kim S, Choi E. Uncertainty-aware text-to-program for question answering on structured electronic health records. arXiv Preprint posted online on March 14, 2022. [doi: 10.1007/978-3-030-96623-2_14]
40. Wang P, Shi T, Agarwal K, Choudhury S, Reddy CK. Attention-based aspect reasoning for knowledge base question answering on clinical notes. arXiv Preprint posted online on August 1, 2021. [doi: 10.1145/3535508.3545518]
41. Park J, Cho Y, Lee H, Choo J, Choi E. Knowledge graph-based question answering with electronic health records. arXiv Preprint posted online on October 19, 2020 [FREE Full text]
42. Yue X, Zhang XF, Yao Z, Lin S, Sun H. CliniQG4QA: generating diverse questions for domain adaptation of clinical question answering. arXiv Preprint posted online on October 30, 2020 [FREE Full text] [doi: 10.1109/bibm52615.2021.9669300]
43. Lehman E, Lialin V, Legaspi KY, Sy AJ, Pile PT, Alberto NR, et al. Learning to ask like a physician. arXiv Preprint posted online on June 6, 2022 [FREE Full text] [doi: 10.18653/v1/2022.clinicalnlp-1.8]
44. Soni S, Gudala M, Wang DZ, Roberts K. Using FHIR to construct a corpus of clinical questions annotated with logical forms and answers. AMIA Annu Symp Proc 2019;2019:1207-1215 [FREE Full text] [Medline: 32308918]
45. Mishra S, Awasthi R, Papay F, Maheshawari K, Cywinski JB, Khanna A, et al. DiagnosisQA: a semi-automated pipeline for developing clinician validated diagnosis specific QA datasets. medRxiv Preprint posted online on November 11, 2021 [FREE Full text] [doi: 10.1101/2021.11.10.21266184]
46. Yue X, Gutierrez BJ, Sun H. Clinical reading comprehension: a thorough analysis of the emrQA dataset. arXiv Preprint posted online on May 1, 2020 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.410]
47. Pal A. CLIFT: analysing natural distribution shift on question answering models in clinical domain. arXiv Preprint posted online on October 19, 2023 [FREE Full text]
48. Hamidi A, Roberts K. Evaluation of AI chatbots for patient-specific EHR questions. arXiv Preprint posted online on June 5, 2023 [FREE Full text]
49. Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, Thapa R, et al. MedAlign: a clinician-generated dataset for instruction following with electronic medical records. arXiv Preprint posted online on August 27, 2023 [FREE Full text] [doi: 10.48550/arXiv.2308.14089]
50. Mahbub M, Goethert I, Danciu I, Knight K, Srinivasan S, Tamang S, et al. Question-answering system extracts information on injection drug use from clinical notes. arXiv Preprint posted online on May 15, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.08777]

XSL•FO
RenderX

51. Dada A, Ufer TL, Kim M, Hasin M, Spieker N, Forsting M, et al. Information extraction from weakly structured radiological reports with natural language queries. Eur Radiol 2024 Jan;34(1):330-337 [FREE Full text] [doi: 10.1007/s00330-023-09977-3] [Medline: 37505252]
52. Soni S, Roberts K. A paraphrase generation system for EHR question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP 2019; August 1, 2019; Florence, Italy. [doi: 10.18653/v1/w19-5003]
53. Moon S, Fan J. How you ask matters: the effect of paraphrastic questions to BERT performance on a clinical SQuAD dataset. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. 2020 Presented at: EMNLP 2020; November 19, 2020; Online URL: https://aclanthology.org/2020.clinicalnlp-1.13 [doi: 10.18653/v1/2020.clinicalnlp-1.13]
54. Soni S, Roberts K. Paraphrasing to improve the performance of Electronic Health Records Question Answering. AMIA Jt Summits Transl Sci Proc 2020;2020:626-635 [FREE Full text] [Medline: 32477685]
55. Patrick J, Li M. An ontology for clinical questions about the contents of patient notes. J Biomed Inform 2012 Apr;45(2):292-306 [FREE Full text] [doi: 10.1016/j.jbi.2011.11.008] [Medline: 22142949]
56. Roberts K, Rodriguez L, Shooshan SE, Demner-Fushman D. Resource classification for medical questions. AMIA Annu Symp Proc 2016;2016:1040-1049 [FREE Full text] [Medline: 28269901]
57. Roberts K, Patra BG. A semantic parsing method for mapping clinical questions to logical forms. AMIA Annu Symp Proc 2017;2017:1478-1487 [FREE Full text] [Medline: 29854217]
58. Schwertner MA, Rigo SJ, Araujo DA, Silva AB, Eskofier B. Fostering natural language question answering over knowledge bases in oncology EHR. In: Proceedings of the IEEE 32nd International Symposium on Computer-Based Medical Systems. 2019 Presented at: CBMS 2019; June 5-7, 2019; Cordoba, Spain. [doi: 10.1109/cbms.2019.00102]
59. Singh Rawat BP, Li F, Yu H. Clinical judgement study using question answering from electronic health records. Proc Mach Learn Res 2019 Aug;106:216-229 [FREE Full text] [Medline: 31897452]
60. Rawat BP, Weng WH, Min SY, Raghavan P, Szolovits P. Entity-enriched neural models for clinical question answering. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. 2020 Presented at: BioNLP 2020; July 9, 2020; Online. [doi: 10.18653/v1/2020.bionlp-1.12]
61. Bae S, Kim D, Kim J, Choi E. Question answering for complex electronic health records database using unified encoder-decoder architecture. arXiv Preprint posted online on November 14, 2021 [FREE Full text]
62. Pan Y, Wang C, Hu B, Xiang Y, Wang X, Chen Q, et al. A BERT-based generation model to transform medical texts to SQL queries for electronic medical records: model development and validation. JMIR Med Inform 2021 Dec 08;9(12):e32698 [FREE Full text] [doi: 10.2196/32698] [Medline: 34889749]
63. Soni S, Roberts K. Toward a neural semantic parsing system for EHR question answering. AMIA Annu Symp Proc 2022;2022:1002-1011 [FREE Full text] [Medline: 37128406]
64. Wen A, Elwazir MY, Moon S, Fan J. Adapting and evaluating a deep learning language model for clinical why-question answering. JAMIA Open 2020 Apr;3(1):16-20 [FREE Full text] [doi: 10.1093/jamiaopen/ooz072] [Medline: 32607483]
65. Soni S, Roberts K. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020 Presented at: LREC 2020; May 11-16, 2020; Marseille, France.
66. Mairittha T, Mairittha N, Inoue S. Improving fine-tuned question answering models for electronic health records. In: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers. 2020 Presented at: UbiComp/ISWC '20 Adjunct; September 12-17, 2020; Virtual Event. [doi: 10.1145/3410530.3414436]
67. Moon S, He H, Fan JW. Effects of information masking in the task-specific finetuning of a transformers-based clinical question-answering framework. In: Proceedings of the IEEE 10th International Conference on Healthcare Informatics. 2022 Presented at: ICHI 2022; June 11-14, 2022; Rochester, MN. [doi: 10.1109/ichi54592.2022.00017]
68. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc 2023 Jan 18;30(2):340-347 [FREE Full text] [doi: 10.1093/jamia/ocac225] [Medline: 36451266]
69. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
70. Kang M, Baek J, Hwang SJ. KALA: knowledge-augmented language model adaptation. arXiv Preprint posted online on April 22, 2022 [FREE Full text] [doi: 10.48550/arXiv.2204.10555]
71. Tarbell R, Choo KK, Dietrich G, Rios A. Towards understanding the generalization of medical text-to-SQL models and datasets. arXiv Preprint posted online on March 22, 2023. [doi: 10.48550/arXiv.2303.12898]
72. Soni S, Datta S, Roberts K. quEHRy: a question answering system to query electronic health records. J Am Med Inform Assoc 2023 May 19;30(6):1091-1102 [FREE Full text] [doi: 10.1093/jamia/ocad050] [Medline: 37087111]
73. Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, et al. Do we still need clinical language models? arXiv Preprint posted online on February 16, 2023. [doi: 10.48550/arXiv.2302.08091]
74. Yue X, Zhang XF, Sun H. Annotated question-answer pairs for clinical notes in the MIMIC-III database (version 1.0.0). PhysioNet. 2021. URL: https://physionet.org/content/mimic-iii-question-answer/1.0.0/ [accessed 2024-09-06]

XSL•FO
RenderX

75. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Mar 17;4(1):ooab011 [FREE Full text] [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
76. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL 2019; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1139]
77. Min SY, Raghavan P, Szolovits P. Advancing Seq2seq with joint paraphrase learning. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. 2020 Presented at: ClinicalNLP 2020; November 19, 2020; Online. [doi: 10.18653/v1/2020.clinicalnlp-1.30]
78. Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: EMNLP 2015; September 17-21, 2015; Lisbon, Portugal URL: https://aclanthology.org/D15-1166 [doi: 10.18653/v1/d15-1166]
79. Yin P, Neubig G. TRANX: a transition-based neural abstract syntax parser for semantic parsing and code generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018 Presented at: EMNLP 2018; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-2002]
80. Dong L, Lapata M. Coarse-to-fine decoding for neural semantic parsing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Presented at: ACL 2018; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1068]
81. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. 2004 Presented at: ACL 2004; July 21-26, 2004; Barcelona, Spain URL: https://aclanthology.org/W04-1013
82. Yu T, Zhang R, Yang K, Yasunaga M, Wang D, Li Z, et al. Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP 2018; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1425]
83. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Presented at: ACL '02; July 7-12, 2002; Philadelphia, PA. [doi: 10.3115/1073083.1073135]
84. Agarwal A, Lavie A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. 2007 Presented at: StatMT '07; June 23, 2007; Prague, Czech Republic. [doi: 10.3115/1626355.1626389]
85. Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. 2006 Presented at: AMTA 2006; August 8-12, 2006; Cambridge, MA.
86. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
87. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Research Square Preprint posted online on February 28, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BLEU: Bilingual Evaluation Understudy
ClinicalKB: Clinical Knowledge Base
EHR: electronic health record
KB: knowledge base
LSTM: long short-term memory
MedTS: Medical text-to-SQL
METEOR: Metric for Evaluation of Translation with Explicit Ordering
MIMIC: Medical Information Mart for Intensive Care
MRC: machine-reading comprehension
n2c2: natural language processing clinical challenge
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Item for Systematic Reviews and Meta-Analyses extension for Scoping Review
QA: question answering
SPARQL: SPARQL Protocol and RDF Query Language

XSL•FO
RenderX

SQuAD: Stanford QA dataset
TER: translation error rate
TREQS: Translate-Edit Model for Question-to-SQL
UMLS: Unified Medical Language System

Edited by A Mavragani; submitted 16.10.23; peer-reviewed by D Hu, KM Kuo, S Bae; comments to author 06.05.24; revised version received 02.07.24; accepted 24.07.24; published 30.10.24.

Please cite as:
Bardhan J, Roberts K, Wang DZ
Question Answering for Electronic Health Records: Scoping Review of Datasets and Models
J Med Internet Res 2024;26:e53636
URL: https://www.jmir.org/2024/1/e53636
doi:10.2196/53636
PMID:

©Jayetri Bardhan, Kirk Roberts, Daisy Zhe Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Influence of Blood Sampling Service Process Reengineering on Medical Services Supply: Quasi-Experimental Study

Wenmin Liao[1*], MPH; Rong He[1*], ME; Zhonglian He[1*], ME; Nan Shi[1], MSc; Dan Li[1], ME; Aihua Zhuang[1], MBBS; Feng Gan[1], ME; Ying Sun[1], MD; Chaofeng Li[1], PhD

State Key Laboratory of Oncology, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Chaofeng Li, PhD
State Key Laboratory of Oncology
Guangdong Provincial Clinical Research Center for Cancer
Sun Yat-sen University Cancer Center
Dongfengdong road 651#
Guangzhou, 510060
China
Phone: 86 87343292
Email: lichaofeng@sysucc.org.cn

Abstract

Background: Tertiary hospitals in China are confronted with significant challenges due to limited spatial capacity and workforce constraints, leading to saturated allocation of medical resources and restricted growth in medical service provision. The incorporation of digital health into medical service process reengineering (MSPR) marks a pivotal transformation and restructuring of conventional health service delivery models. Specifically, the application of MSPR to blood sampling services processes reengineering (BSSPR) holds promise for substantially enhancing the efficiency and quality of medical services through streamlining and optimizing these procedures. However, the comprehensive impact of BSSPR has been infrequently quantified in existing research.

Objective: This study aims to investigate the influence of BSSPR on the efficiency and quality of medical services and to elucidate the key informative technological support points underpinning BSSPR.

Methods: Data were collected from both the new and old laboratory information systems from August 1, 2019, to December 31, 2021. A combination of statistical description, chi-square test, and *t* test was used to compare check-in time and waiting time of outpatients before and after the implementation of BSSPR. An interrupted time-series design was used to analyze the impact of BSSPR on medical service efficiency and quality, enabling the control of confounding variables, including changes in medical human resources and both long- and short-term temporal trends.

Results: BSSPR had an impact on the efficiency and quality of medical services. Notably, there was a significant increase in the number of patients receiving blood sampling services, with a daily service volume increase of ~150 individuals (P=.04). The average waiting time for patients decreased substantially from 29 (SD 36) to 11 (SD 11) minutes, indicating a marked improvement in patient experience. During the peak period, the number of patients receiving blood sampling services per working hour statistically increased from 9.56 to 16.77 (P<.001). The interrupted time-series model results demonstrated a reduction in patients' waiting time by an average of 26.1 (SD 3.8; 95% CI –33.64 to –18.57) minutes. Although there was an initial decline in the number of outpatients admitted following BSSPR implementation, an upward trend was observed over time (β=1.13, 95% CI 0.91-1.36).

Conclusions: BSSPR implementation for outpatients not only reduced waiting time and improved patients' experience but also augmented the hospital's capacity to provide medical services. This study's findings offer valuable insights into the potential advantages of BSSPR and underscore the significance of harnessing digital technologies to optimize medical service processes. This research serves as a foundational basis and provides scientific support for the promotion and application of BSSPR in other health care contexts. By continuing to explore and refine the integration of digital technologies in health care, we can further enhance patient outcomes and elevate the overall quality of medical services.

(J Med Internet Res 2024;26:e51412) doi:10.2196/51412

XSL•FO
RenderX

KEYWORDS

process reengineering; blood sampling; hospital administration; medical informatics; digital health; patient experience

Introduction

Optimizing medical services to provide faster, superior care that benefits a broader populace has consistently been the focus point of every health system, with this need becoming increasingly urgent in the post–COVID-19 era globally [1]. Three years into the COVID-19 pandemic, over 765 million confirmed cases [2] have persistently strained health systems, exacerbating the global medical supply-demand imbalance [3]. Drawing from years of pandemic response experience, medical institutions must not only deliver health services but also prevent cross-infections to ensure patients' safety [4,5], which has posed a huge challenge for the Chinese medical system given its vast population and limited medical resources. Since the 2009 medical reform, the overall efficiency of national medical resources has increased, yet most provinces have entered the stage of diminishing returns in the medical resource allocation scale [6]. Therefore, enhancing technical efficiency and promoting total factor productivity changes with technological advancement is of great practical significance in creating a secure, orderly, and efficient medical environment in the new context [7].

Due to spatial constraints and workforce limitations, tertiary hospitals face challenges related to the saturated allocation of medical resources and limited growth of medical services [8]. According to the China Health Statistical Yearbook 2020, there were 954,390 primary-level medical institutions in China, accounting for only 32.6% of patient visits, whereas 2,749 tertiary hospitals accounted for 35.5% of patient visits [9]. This conflict is particularly acute in tertiary cancer hospitals, where it is estimated that over 4.5 million new cancer cases and more than 3 million cancer-related deaths were diagnosed in China in 2020 [10]. Given the complexity of tumor treatment, it is not appropriate to use hierarchical diagnosis and treatment policy to divert patients.

In the face of these challenges, business process reengineering emerges as a potent strategy to achieve dramatic productivity improvements without compromising the quality and scope of core health care services [11]. Technological innovation through the "internet plus health care" paradigm [12], also known as digital health [13], can achieve the goal of providing high-quality medical services to more patients with limited resources. In this context, medical service process reengineering (MSPR) supported by digital health represents a fundamental thinking and reconstruction of the hospital's business process, aiming for overall improvement of cost, quality, service, and efficiency, and fostering a modern patient-centered medical environment.

In this study, a case of MSPR from a tertiary cancer hospital in Guangzhou, China, was presented. The Sun Yat-sen University Cancer Center (SYSUCC) is one of the oncology hospitals with the largest scale and the strongest academic strength in China, which integrates medical treatment, teaching, scientific research, and prevention. In 2020, SYSUCC had approximately 5000 outpatient visits daily, with an average of over 1000 daily blood tests. Before the reengineering of the blood sampling process, patients were required to queue up at the blood sampling center to print their individual barcodes after a doctor had prescribed a blood test. Once the barcode was printed, patients would need to queue up again for the actual blood sampling procedure. In cases where patients were prescribed multiple blood tests during a single clinic visit, they would take the centrally printed barcodes home for safekeeping and bring them back for their subsequent visit. The nurses would charge patients additional fees for consumables based on the different combinations of samples collected (Figure 1A). Consequently, the blood sampling center frequently experienced long queues, chaotic scenes, charge errors, and barcode allocation errors during peak hours.

Confronted with the challenges of high patient volume and limited service space, the blood sampling service process reengineering (BSSPR) was conducted in SUSYCC, supported by the Department of Information Technology. The BSSPR primarily encompassed the following aspects (Figure 1B): first, considering periodic blood sampling and testing needs of patients with cancer, doctors were required to batch blood tests after issuing prescriptions, with the system supporting batch appointment scheduling. Second, the system enabled appointment scheduling with a granularity of half an hour. Appointments can be made by doctors and patients through mobile apps or self-service machines, all of which could be modified at any time before the appointed day. Finally, a queuing and calling process was established within the blood sampling center: during the corresponding appointment period, patients could check in using mobile apps or self-service machines and await their call. Upon checking in, the blood sampling equipment would automatically standby with the barcode attached, allowing the nurse to sample blood immediately after calling.

The interrupted time-series (ITS) design is recognized as one of the most potent quasi-experimental designs, quantifying the effect of an intervention by comparing data before and after the intervention [14]. ITS necessitates a clearly defined intervention implementation timepoint and data collection at multiple equally spaced time points before and after the intervention (eg, weekly, monthly, or annually). By modeling preinterruption data to predict postinterruption outcomes and offering a counterfactual scenario for what would have transpired if the interruption had not occurred, akin to a comparator group in a randomized trial, ITS provides a range of effect estimates to describe the intervention's impact. For instance, a level change corresponds to the difference between the time point of interest and the predicted preintervention level, while a slope change corresponds to the difference between the postintervention trend and the preintervention trend [15,16]. In recent years, ITS has been extensively used to measure the effectiveness of health care policies and programs [17,18].

XSL•FO
RenderX

With BSSPR as the study event, this research aims to use the ITS design to explore the impact of MSPR on the efficiency or quality of medical services and discuss the key elements of successful MSPR implementation, thereby providing practical construction experience.

Figure 1. Blood sampling service process of outpatients (A) before and (B) after blood sampling service process reengineering.

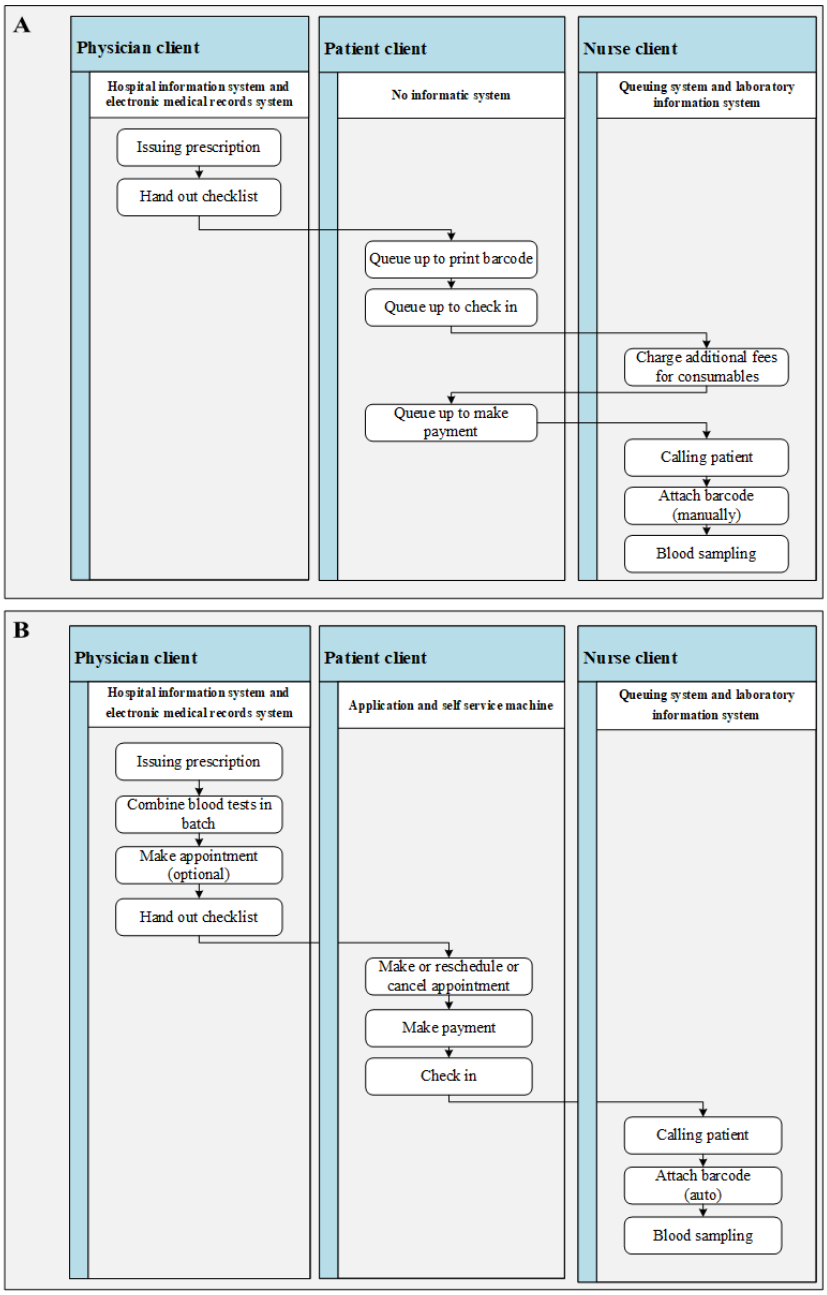

Methods

Research Data

Treating the BSSPR, which finished on August 1, 2020, in SYSYCC as an interest, for comparative analysis, data from the preceding and subsequent laboratory information system (LIS) were used. Specifically, data from the old LIS, representing the period before BSSPR, spanning from August 1, 2019, to October 30, 2020, were compared with those from the new LIS, reflecting the post-BSSPR era from September 28, 2020, to December 31, 2021. The old LIS dataset comprised anonymized patient names, medical record numbers, barcode numbers, barcode printing time, sampling time, sampling nurses, and blood test items. For the old LIS dataset, due to the lack of a reservation mechanism, the accurate check-in time of patients cannot be obtained; therefore, the barcode printing time is used as the check-in time of patients. Therefore, the waiting time was the difference between sampling time and barcode printing time. Conversely, the new LIS dataset included the same patient identifiers and barcode number but substituted barcode printing time with appointment and check-in times, thereby facilitating the calculation of waiting times after BSSPR. Specifically, the patient's waiting time is now the difference between sampling time and check-in time.

Initially, a comprehensive dataset encompassing 901,730 patient records was assembled. However, rigorous data cleaning was

XSL•FO
RenderX

conducted to ensure analytical rigor. Records of 125,397 individuals were excluded from the old LIS analysis due to time differences exceeding 1 day between sampling and barcode printing, which indicated potential barcode reuse from previous visits. In addition, 60,654 individuals undergoing the blood routine test project were omitted from both LIS datasets, as the hospital designated a dedicated window for this project, negating the need for appointments. Furthermore, 32,064 records pertaining to nonworking hours (before 6 AM or after 5:30 PM on weekdays, and all day on weekends) were excluded from both datasets to maintain temporal consistency. Consequently, the final analysis encompassed 683,615 blood collection records from patients visiting between 6 AM and 5:30 PM on weekdays during the study period.

Quantitative Analysis

The distribution of check-in time and waiting time of patients in the new and old LIS were statistically described. According to the check-in time distribution of outpatients, the peak period was set from 7 AM to 10 AM every day, and the rest was set as an off-peak period.

A series of two-tailed *t* test comparisons were used to compare the overall work efficiency per working hour before and after the implementation of the BSSPR, specifically assessing the average number of blood sampling tubes and patient encounters per hour per nurse. Comparisons were conducted separately for all day, peak hours, and off-peak hours, illuminating the effects of BSSPR on blood sampling efficiency across different time periods. Chi-square tests were used to compare the distribution of patient check-in time within 1 day and 1 week before and after BSSPR, respectively, to quantify the impact of BSSPR on the number of patients received.

ITS design [15] was used to analyze the impact of the BSSPR on the quantity and speed of medical services provision under the condition of controlling changes in medical human resources and long- and short-term time trends. The model framework is as follows:

$$y_{ct} = a + \beta_1 T_{ct} + \beta_2 \text{Trend}_t + \beta_3 T_c \times \text{Trend}_t + \beta_4 \text{wday}_t + \beta_5 \text{workforce}_t + \varepsilon$$

Wherein, the dependent variables y_{ct} are the daily total number of outpatients (Model 1) and the daily average waiting time of outpatients (Model 2) at day_t, respectively; T_{ct} is a dummy variable indicating before (coding as 0) and after (coding as 1) the BSSPR at day_t, which makes $_1$ the post-BSSPR level change of dependent variables; $Trend_t$ is the count of days from August 1, 2019, to day_t, which makes $_2$ the natural trend of dependent variables by day, used to control long-term time trends; and $_3$ indicates the post-BSSPR trend change of dependent variables. The temporal patterns (short-term time trend) were controlled using $wday_t$, indicating Monday to Friday, and medical human resource change was controlled by $workforce_t$, which denotes the total daily working hours of nurses at day_t. The authors used the unique identifiers of sampling nurses in the dataset to determine whether a nurse was engaged in blood sampling activities every half-hour and calculated the total amount of health care workforce (in working hours) dedicated to blood sampling each day.

R (version 4.0.2, R Development Core Team) was used to conduct all statistical analysis, with packages *plyr*, *stringr*, *dplyr*, *lubridate*, *doBy*, *ggplot2*, *ggtext*, and *Epi*.

Ethical Considerations

The study was granted ethical approval by the ethics committee of SYSUCC (B2023-486-01). Given the retrospective nature of the study, the requirement for informed consent was waived by the respective boards. Rigorous measures were put in place to anonymize all individual data, including the removal of identifiers like names, patient IDs, and contact details. Access to the data was strictly limited to authorized personnel, and it was stored securely to prevent any unauthorized access or disclosure, ensuring privacy protection while facilitating population-level analysis. No identification of individual participants or users was possible in any images of the paper. No financial or other forms of compensation were provided to participants in this study, as it was a retrospective analysis of existing data.

Results

Influence of Outpatient BSSPR on the Volume of Medical Service Provision

During the study period, a total of 683,615 patients underwent blood sampling, including 345,649 patients before BSSPR (1133 cases per day) and 337,966 patients after BSSPR (1285 cases per day). As shown in Figure 2, after BSSPR, the daily pattern of outpatients' visiting was the same as that before BSSPR, and the peak appeared from 7 AM to 10 AM. However, after BSSPR, the average number of outpatients who checked-in during peak hours increased significantly (P<.001), of whom, the average number of outpatients who checked-in from 8:30 AM to 9 AM and 9 AM to 9:30 AM were 147 (SD 35) and 157 (SD 45), respectively, while before BSSPR, the corresponding number of admitted outpatients were 128 and 121, respectively. The chi-square (χ^2) test showed a significant difference in the daily distribution of check-in time before and after BSSPR (χ^2=130.67; P<.001). Analysis of the distribution of outpatients on weekdays showed that by using the old blood sampling service process, the number of outpatients admitted showed a continuous downward trend from an average of 1384 (SD 504) outpatients on Monday to an average of 919 (SD 370) outpatients on Friday. After BSSPR, not only did the average number of outpatients admitted on each Monday reach 1437 (SD 377) but also the numbers on Thursday and Friday were significantly increased compared with those before BSSPR, up to more than 1100 (P=.04). In general, after BSSPR, the number of admitted outpatients of blood sampling centers increased significantly (χ^2=9.92; P=.04).

XSL•FO
RenderX

Figure 2. Distribution of outpatients' check-in time before and after blood sampling service process reengineering (BSSPR): (A) daily pattern and (B) weekly pattern.

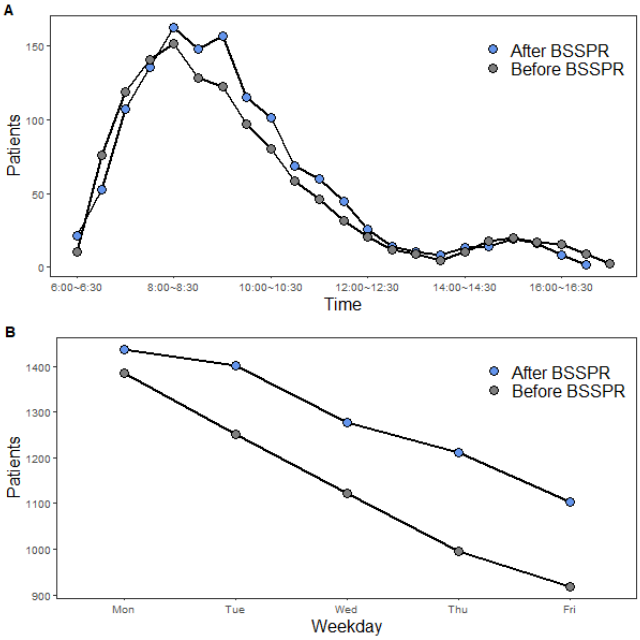

Comparative Analysis of Blood Sampling Efficiency Before and After BSSPR

The analysis of outpatients' waiting time before and after the BSSPR showed that before BSSPR, the average waiting time of patients in the whole day, peak period, and off-peak period was 29 (SD 36) minutes, 30 (SD 36) minutes, and 26 (SD 38) minutes, respectively. After BSSPR, the patients' average waiting time was 11 (SD 11) minutes, 11 (SD 10) minutes, and 10 (SD 13) minutes, respectively. As shown in Figure 3, after BSSPR, outpatients' waiting time was significantly reduced in all periods and remained stable across peak period and off-peak periods, yet the waiting time of outpatients in peak period increased sharply before BSSPR.

Figure 3. Distribution of outpatients' waiting time before and after blood sampling service process reengineering (BSSPR).

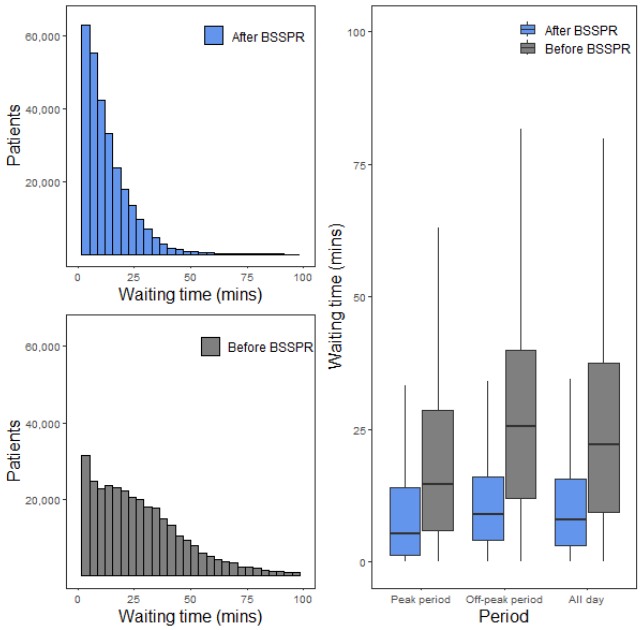

Table 1 shows the influence of the BSSPR on blood sampling efficiency; the number of sampling tubes per hour of the whole day and peak period before and after BSSPR were not significantly different (all $P>.05$), yet the average number of sampling tubes per hour of the off-peak period was 23.47 (SD 18.47) and 33.36 (SD 11.32), before and after BSSPR, respectively, and the difference is statistically significant ($P<.001$). By comparing the number of outpatients per hour, it can be found that, before and after BSSPR, the average number of patients admitted per hour was 9.56 (SD 7.67) and 16.77 (SD 4.11) in the whole day, respectively, and the difference was statistically significant ($P<.001$). Statistically significant differences were also shown in the peak period (mean 10.52, SD 8.93 vs 19.60, SD 4.62; $P<.001$) and the off-peak period

XSL•FO
RenderX

(mean 7.19, SD 6.56 vs 12.08, SD 5.64; *P*<.001). In the peak period and off-peak period, the number of admitted outpatients per work hour increased by 86% (9.08/10.52) and 68% (4.89/7.19) as a result of BSSPR.

Table 1. Comparative analysis on blood sampling efficiency.

Blood sampling efficiency	Before BSSPR[a], mean (SD)	After BSSPR, mean (SD)	*t* test (*df*)	*P* value
Sampling tubes (/working hour[b])				
All day	45.69 (26.63)	43.70 (12.89)	–0.59 (187)	.55
Peak period	51.52 (31.51)	53.12 (13.52)	0.40 (183)	.69
Off-peak period	23.47 (18.47)	33.36 (11.32)	4.17 (225)	<.001
Patients (/working hour)				
All day	9.56 (7.67)	16.77 (4.11)	7.35 (223)	<.001
Peak period	10.52 (8.93)	19.60 (4.62)	7.98 (221)	<.001
Off-peak period	7.19 (6.56)	12.08 (5.64)	5.73 (243)	<.001

[a]BSSPR: blood sampling service process reengineering.

[b]Working hour means 1 nurse working for 1 hour.

Attributed Influence of BSSPR on the Speed and Volume of Outpatient Medical Service Provision

The results of the ITS regression model are shown in Table 2. Although the BSSPR resulted in an average decrease of 106.62 patients per day in the peak period, a set of slight but significant increasing trends of patient counts after BSSPR have been detected in the whole-day, peak, and off-peak periods (β=1.13, .88, and .21, respectively; all *P*<.001). The effect of increasing the number of working hours on the number of patients received was statistically significant in every period (all *P*<.001), and an increase of 1 working hour could increase the number of admitted outpatients on average by 13.15 (SD 0.15).

The BSSPR reduced the waiting time of patients by 26.1 (95% CI –33.64 to –18.57) minutes in total, within the peak time by 27.52 (95% CI –35.47 to –19.56) minutes, and within the off-peak time by 22.33 (95% CI –30.46 to –14.20) minutes. Patients' waiting time was not significantly reduced by adding to the workforce (all *P*>.05). It has been detected that there is a very slight increasing trend in patients' waiting time after BSSPR, as well as during the peak period (β=0.02).

XSL•FO
RenderX

Table 2. Attributed influence of BSSPR[a] on the speed and volume of outpatient medical service provision.

Model	All day		Peak period		Off-peak period	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Model 1[b]						
Change after BSSPR	–78.87 (–208.12 to 50.38)	.23	–106.62 (–199.56 to –13.69)	.02	45.44 (–7.96 to 98.83)	.10
Trend after BSSPR	1.13 (0.91 to 1.36)	<.001	0.88 (0.71 to 1.04)	<.001	0.21 (0.12 to 0.31)	<.001
Natural trend	–0.50 (–0.58 to –0.42)	<.001	–0.32 (–0.39 to –0.26)	<.001	–0.16 (–0.2 to –0.12)	<.001
Working hour	13.15 (12.87 to 13.44)	<.001	14.87 (14.52 to 15.23)	<.001	10.16 (9.84 to 10.48)	<.001
Weekday (Monday as reference)						
Tuesday	–43.08 (–71.42 to –14.74)	.003	–31.51 (–52.28 to –10.73)	.003	–10.24 (–29.18 to 8.71)	.29
Wednesday	–72.62 (–89.26 to –55.97)	<.001	–41.85 (–60.73 to –22.97)	<.001	–42.76 (–58.8 to –26.72)	<.001
Thursday	–121.32 (–145.16 to –97.48)	<.001	–71.22 (–91.52 to –50.92)	<.001	–48.82 (–64.33 to –33.3)	<.001
Friday	–118.82 (–138.36 to –99.28)	<.001	–80.27 (–101.41 to –59.13)	<.001	–44.43 (–60.19 to –28.67)	<.001
Model 2[c]						
Change after BSSPR	–26.10 (–33.64 to –18.57)	<.001	–27.52 (–35.47 to –19.56)	<.001	–22.33 (–30.46 to –14.20)	<.001
Trend after BSSPR	0.02 (0.00 to 0.03)	.04	0.02 (0.00 to 0.04)	.03	0.005 (–0.01 to 0.02)	.65
Natural trend	–0.01 (–0.02 to 0.01)	.21	–0.01 (–0.03 to 0.00)	.11	0.01 (–0.01 to 0.02)	.52
Working hour	–0.02 (–0.07 to 0.02)	.32	–0.05 (–0.13 to 0.03)	.26	–0.09 (–0.23 to 0.05)	.19
Weekday (Monday as reference)						
Tuesday	–3.98 (–7.06 to –0.90)	.01	–2.75 (–5.82 to 0.32)	.08	–6.68 (–10.54 to –2.82)	.001
Wednesday	–3.97 (–7.08 to –0.87)	.01	–1.52 (–4.68 to 1.64)	.35	–9.00 (–12.76 to –5.25)	<.001
Thursday	–3.09 (–6.26 to 0.08)	.06	–0.30 (–3.55 to 2.94)	.85	–9.26 (–13.00 to –5.51)	<.001
Friday	–1.40 (–4.72 to 1.91)	.41	1.37 (–2.01 to 4.76)	.43	–8.40 (–12.30 to –4.50)	<.001

[a]BSSPR: blood sampling service process reengineering.

[b]The dependent variable is the daily total number of outpatients.

[c]The dependent variable is the daily average waiting time of outpatients.

Discussion

Principal Findings

This study took the BSSPR event of a tertiary oncology hospital in China as a focus and introduced the practice of MSPR through digital health. It was found that after BSSPR, the advanced management of patients' flow was effective, and medical resources were used more efficiently during off-peak hours such as Friday. BSSPR has not only improved the capacity of medical institutions to provide medical services but also significantly reduced the waiting time of outpatients. These results indicated the feasibility and applicability of using digital health to optimize the medical service process in the face of the bottleneck of tertiary hospitals in improving their ability to provide medical services.

Through our work, we have outlined several of the cores of MSPR to provide some practical experience. First, a common way of MSPR is to make medical service provision in a planned and orderly way as far as possible, such as proactive and rational management of patient flow in advance [19]. In this way, it not only avoids the adverse impact of the rush hour crowd on the medical environment but also reduces the waste of medical resources during the nonrush hour, to maximize the use of medical resources. The realization of this in BSSPR described in this study depends on the establishment of a unified appointment platform: making an appointment for blood sampling from multiple channels, including an electronic medical records system (by doctors), mobile apps, or self-service machines (by patients); sharing the same appointment pool and the same queue; and allowing patients to make, reschedule, and cancel appointments within certain rules [20].

This approach greatly reduces the invalid waiting time of patients, thus creating a win-win situation, which not only avoids the adverse impact of long waiting time on patients' satisfaction [21-23] and willingness to revisit [8,24] but also improves the use of health care services and guides the growth of hospital medical service [25]. Advanced management of patient flow may also make more effective use of medical resources in leisure time [26]. Before BSSPR, many patients would not choose Friday for blood tests, which may be due to the suspension of the clinic on weekends. In the absence of appointment rules, patients in other cities would consider it risky to see a doctor on Friday, because they would have to wait for a whole weekend

XSL•FO
RenderX

if they missed it. After BSSPR, the number of outpatient visits on Friday increased significantly, confirming the previously stated patient concerns and the powerful effect of the establishment of appointment rules in BSSPR on the allocation of medical resources over time.

Second, the MSPR related to patient services needs to pay much attention to user-friendliness and message feedback mechanisms. After BSSPR, doctors would hand out the outpatient examination checklist with the appointment rules and matters needing attention after issuing the examination order, so as to improve the friendliness of the process of appointment and check-in to patients. At the same time, an instant message will be sent after making appointments and changing appointments through mobile apps or self-service machines. The convenience of self-service operation and message transmission makes patients more active in operation. Although the results of the ITS model showed a decrease in the number of outpatient patients received by blood sampling center during the peak period after the implementation of BSSPR, this may be due to the diversion of patients from the peak period to another period at the early stage of the system's launch, which can also indicate that the medical resources of blood sampling center were overloaded before the implementation of MSPR. The upward trend in outpatient numbers after MSPR is implemented also shows that proper MSPR can improve hospital efficiency in the long run as patients adapt to the new service mode.

Third, the effectiveness of MSPR can be maximized only when MSPR is customized according to the characteristics of patients or diseases. In this case, according to the characteristics of patients with cancer undergoing periodic blood tests, the function of appointment by batch is realized. In other scenarios, for example, the return visit time of clinical research patients should be strictly calculated by the system, the follow-up visit time of postoperative patients should be personalized according to the follow-up results, and so on, and digital technology should be used to realize smart services.

Finally, leveraging detailed data recorded by the information system, we observed and analyzed service system operations to achieve lean management objectives, exemplifying this case's practical implications. In the actual scenario, the waiting time can be estimated by the length of the queue, which is beneficial for patients to choose the optimal route [27]. The calculated increase in the number of patients received and the decrease in waiting time due to the improvement of human resources can be used to flexibly allocate the number of nurses and equipment according to the immediate waiting time and the number of queuing people, so as to reduce the redundancy and shortage during peak hours [28]. In addition, calculating the number of visits and productivity of each nurse facilitates more accurate performance evaluations. Therefore, the optimal situation of patients, medical resource allocation, and service management can be achieved.

The novelty of this paper resides in its employment of an ITS design to appraise the impact of an MSPR initiative. By controlling both short- and long-term trends alongside health care human resource variations, this methodology enables the analysis of longitudinal administrative data, offering a robust retrospective assessment of digital technologies' efficacy within health care settings [29]. Moreover, our contribution extends to a comprehensive discourse on the pivotal steps involved in medical process reengineering.

However, the study had some limitations. First, the study's limitation primarily stems from its focused scope, encompassing a specific health care context and MSPR endeavor. As a single-center study, the findings' generalizability to diverse health care settings or interventions remains uncertain. Furthermore, the study's reliance on retrospective data analysis and ITS design may be constrained by inherent limitations, including data incompleteness or potential biases. An expanded exploration of confounding variables potentially influencing outcomes could have further strengthened the study. Nevertheless, this research presents valuable insights, underscoring the need for future endeavors to address the identified limitations and build upon our findings.

Conclusions

The implementation of BSSPR, an example of MSPR in a tertiary cancer hospital in China, not only improved the capacity of medical institutions to provide medical services but also significantly reduced the waiting time of outpatients. This provides a model for using digital technology in MSPR. Future research should focus on providing an evaluation model for digital MSPR to guide the efficient application of information technology in improving the efficiency of medical resources.

Acknowledgments

This study would not have been possible without the support of Sun Yat-sen University Cancer Center.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to their inclusion of sensitive hospital business operational data but are available from the corresponding author on reasonable request.

Authors' Contributions

ZH and CL contributed to conceptualization, WL and RH contributed to data curation and formal analysis, and all other authors reviewed and provided substantive comments. WL drafted the initial version of the manuscript with input from RH, ZH, and CL.

XSL•FO
RenderX

All other authors provided input and edits into the complete draft. The final draft for submission was prepared by WL and CL and approved by all authors.

Conflicts of Interest

None declared.

References

1. Hick JL, Biddinger PD. Novel Coronavirus and old lessons - preparing the health system for the pandemic. N Engl J Med 2020;382(20):e55. [doi: 10.1056/NEJMp2005118] [Medline: 32212515]
2. WHO COVID-19 dashboard. World Health Organization. URL: https://covid19.who.int [accessed 2024-10-10]
3. Khan S, Haleem A, Deshmukh SG, Javaid M. Exploring the impact of COVID-19 pandemic on medical supply chain disruption. J Ind Intg Mgmt 2021;6(2):235-255. [doi: 10.1142/s2424862221500147]
4. Lian W, Wen L, Zhou Q, Zhu W, Duan W, Xiao X, et al. Digital health technologies respond to the COVID-19 pandemic in a tertiary hospital in China: development and usability study. J Med Internet Res 2020;22(11):e24505 [FREE Full text] [doi: 10.2196/24505] [Medline: 33141679]
5. Shao S, Zhou Z, Li Y, Liu S, Lu L, Hou S, et al. Experiences and practices in the current prevention and control of the novel coronavirus pneumonia in China. Disaster Med Public Health Prep 2021;15(5):e7-e16 [FREE Full text] [doi: 10.1017/dmp.2020.173] [Medline: 32466822]
6. Leng Y, Liu W, Xiao N, Li Y, Deng J. The impact of policy on the intangible service efficiency of the primary health care institution- based on China's health care reform policy in 2009. Int J Equity Health 2019;18(1):14 [FREE Full text] [doi: 10.1186/s12939-018-0901-2] [Medline: 30665422]
7. Blobel B, Oemig F, Ruotsalainen P, Lopez DM. Transformation of health and social care systems-an interdisciplinary approach toward a foundational architecture. Front Med (Lausanne) 2022;9:802487 [FREE Full text] [doi: 10.3389/fmed.2022.802487] [Medline: 35402446]
8. Yan J, Yao J, Zhao D. Patient satisfaction with outpatient care in China: a comparison of public secondary and tertiary hospitals. Int J Qual Health Care 2021;33(1):mzab003. [doi: 10.1093/intqhc/mzab003] [Medline: 33404594]
9. Information center of the National Health Commission of China. China Health Statistical Yearbook 2020. 2021. URL: https://ghdx.healthdata.org/record/china-statistical-yearbook-2020 [accessed 2024-10-10]
10. Qiu H, Cao S, Xu R. Cancer incidence, mortality, and burden in China: a time-trend analysis and comparison with the United States and United Kingdom based on the global epidemiological data released in 2020. Cancer Commun (Lond) 2021;41(10):1037-1048 [FREE Full text] [doi: 10.1002/cac2.12197] [Medline: 34288593]
11. Carmichael BD. Business process reengineering: a remedy for health care. Healthc Manage Forum 1994;7(4):44-50. [doi: 10.1016/S0840-4704(10)61077-0] [Medline: 10140167]
12. Yang F, Shu H, Zhang X. Understanding "Internet Plus Healthcare" in China: policy text analysis. J Med Internet Res 2021;23(7):e23779 [FREE Full text] [doi: 10.2196/23779] [Medline: 34309581]
13. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. Eur Heart J 2016;37(18):1428-1438 [FREE Full text] [doi: 10.1093/eurheartj/ehv770] [Medline: 26873093]
14. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr 2013;13(6 Suppl):S38-S44. [doi: 10.1016/j.acap.2013.08.002] [Medline: 24268083]
15. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017;46(1):348-355 [FREE Full text] [doi: 10.1093/ije/dyw098] [Medline: 27283160]
16. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. J Clin Epidemiol 2020;122:1-11. [doi: 10.1016/j.jclinepi.2020.02.006] [Medline: 32109503]
17. Cruz M, Gillen DL, Bender M, Ombao H. Assessing health care interventions via an interrupted time series model: study power and design considerations. Stat Med 2019;38(10):1734-1752 [FREE Full text] [doi: 10.1002/sim.8067] [Medline: 30616298]
18. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Med Res Methodol 2019;19(1):137 [FREE Full text] [doi: 10.1186/s12874-019-0777-x] [Medline: 31272382]
19. Li H, Zhang Z, Li P, Nie H. Challenges and responses: a tertiary hospital in 2019-nCoV epidemic. Disaster Med Public Health Prep 2020;14(6):808-811 [FREE Full text] [doi: 10.1017/dmp.2020.93] [Medline: 32284090]
20. Pan C, Zhang D, Kon AWM, Wai CSL, Ang WB. Patient flow improvement for an ophthalmic specialist outpatient clinic with aid of discrete event simulation and design of experiment. Health Care Manag Sci 2015;18(2):137-155. [doi: 10.1007/s10729-014-9291-1] [Medline: 25012400]
21. Xie Z, Or C. Associations between waiting times, service times, and patient satisfaction in an endocrinology outpatient department: a time study and questionnaire survey. Inquiry 2017;54:46958017739527 [FREE Full text] [doi: 10.1177/0046958017739527] [Medline: 29161947]

XSL•FO
RenderX

22. Fazl Hashemi SME, Sarabi Asiabar A, Rezapour A, Azami-Aghdash S, Hosseini Amnab H, Mirabedini SA. Patient waiting time in hospital emergency departments of Iran: a systematic review and meta-analysis. Med J Islam Repub Iran 2017;31:79 [FREE Full text] [doi: 10.14196/mjiri.31.79] [Medline: 29445707]
23. Naiker U, FitzGerald G, Dulhunty JM, Rosemann M. Time to wait: a systematic review of strategies that affect out-patient waiting times. Aust Health Rev 2018;42(3):286-293. [doi: 10.1071/AH16275] [Medline: 28355525]
24. Lee S, Groß SE, Pfaff H, Dresen A. Waiting time, communication quality, and patient satisfaction: an analysis of moderating influences on the relationship between perceived waiting time and the satisfaction of breast cancer patients during their inpatient stay. Patient Educ Couns 2020;103(4):819-825. [doi: 10.1016/j.pec.2019.11.018] [Medline: 31767241]
25. Biya M, Gezahagn M, Birhanu B, Yitbarek K, Getachew N, Beyene W. Waiting time and its associated factors in patients presenting to outpatient departments at public hospitals of Jimma Zone, Southwest Ethiopia. BMC Health Serv Res 2022;22(1):107 [FREE Full text] [doi: 10.1186/s12913-022-07502-8] [Medline: 35078474]
26. Almomani I, AlSarheed A. Enhancing outpatient clinics management software by reducing patients' waiting time. J Infect Public Health 2016;9(6):734-743 [FREE Full text] [doi: 10.1016/j.jiph.2016.09.005] [Medline: 27663518]
27. Lin CKY, Ling TWC, Yeung WK. Resource allocation and outpatient appointment scheduling using simulation optimization. J Healthc Eng 2017;2017:9034737 [FREE Full text] [doi: 10.1155/2017/9034737] [Medline: 29104748]
28. Matta ME, Patterson SS. Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. Health Care Manag Sci 2007;10(2):173-194. [doi: 10.1007/s10729-007-9010-2] [Medline: 17608058]
29. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 2015;350:h2750 [FREE Full text] [doi: 10.1136/bmj.h2750] [Medline: 26058820]

Abbreviations

BSSPR: blood sampling service process reengineering
ITS: interrupted time-series
LIS: laboratory information system
MSPR: medical service process reengineering
SYSUCC: Sun Yat-sen University Cancer Center

Edited by A Schwartz; submitted 31.07.23; peer-reviewed by C Yanyan, C Peisong; comments to author 07.08.24; revised version received 22.08.24; accepted 24.09.24; published 12.11.24.

Please cite as:
Liao W, He R, He Z, Shi N, Li D, Zhuang A, Gan F, Sun Y, Li C
Influence of Blood Sampling Service Process Reengineering on Medical Services Supply: Quasi-Experimental Study
J Med Internet Res 2024;26:e51412
URL: https://www.jmir.org/2024/1/e51412
doi:10.2196/51412
PMID:

©Wenmin Liao, Rong He, Zhonglian He, Nan Shi, Dan Li, Aihua Zhuang, Feng Gan, Ying Sun, Chaofeng Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Hospital Length of Stay Prediction for Planned Admissions Using Observational Medical Outcomes Partnership Common Data Model: Retrospective Study

Haeun Lee[1,2], MS; Seok Kim[2], MPH; Hui-Woun Moon[2], MSN; Ho-Young Lee[2], MD, PhD; Kwangsoo Kim[3], PhD; Se Young Jung[4*], MD, MPH, PhD; Sooyoung Yoo[2*], PhD

[1]Department of Biomedical Informatics and Data Science, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, United States
[2]Office of eHealth Research and Businesses, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea
[3]Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[4]Department of Family Medicine, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Se Young Jung, MD, MPH, PhD
Department of Family Medicine, Seoul National University Bundang Hospital
172, Dolma-ro bundang-gu
Seongnam-si, 13605
Republic of Korea
Phone: 82 0317878845
Fax: 82 317874054
Email: syjung@snubh.org

Abstract

Background: Accurate hospital length of stay (LoS) prediction enables efficient resource management. Conventional LoS prediction models with limited covariates and nonstandardized data have limited reproducibility when applied to the general population.

Objective: In this study, we developed and validated a machine learning (ML)–based LoS prediction model for planned admissions using the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).

Methods: Retrospective patient-level prediction models used electronic health record (EHR) data converted to the OMOP CDM (version 5.3) from Seoul National University Bundang Hospital (SNUBH) in South Korea. The study included 137,437 hospital admission episodes between January 2016 and December 2020. Covariates from the patient, condition occurrence, medication, observation, measurement, procedure, and visit occurrence tables were included in the analysis. To perform feature selection, we applied Lasso regularization in the logistic regression. The primary outcome was an LoS of 7 days or longer, while the secondary outcome was an LoS of 3 days or longer. The prediction models were developed using 6 ML algorithms, with the training and test set split in a 7:3 ratio. The performance of each model was evaluated based on the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Shapley Additive Explanations (SHAP) analysis measured feature importance, while calibration plots assessed the reliability of the prediction models. External validation of the developed models occurred at an independent institution, the Seoul National University Hospital.

Results: The final sample included 129,938 patient entry events in the planned admissions. The Extreme Gradient Boosting (XGB) model achieved the best performance in binary classification for predicting an LoS of 7 days or longer, with an AUROC of 0.891 (95% CI 0.887-0.894) and an AUPRC of 0.819 (95% CI 0.813-0.826) on the internal test set. The Light Gradient Boosting (LGB) model performed the best in the multiclassification for predicting an LoS of 3 days or more, with an AUROC of 0.901 (95% CI 0.898-0.904) and an AUPRC of 0.770 (95% CI 0.762-0.779). The most important features contributing to the models were the operation performed, frequency of previous outpatient visits, patient admission department, age, and day of admission. The RF model showed robust performance in the external validation set, achieving an AUROC of 0.804 (95% CI 0.802-0.807).

Conclusions: The use of the OMOP CDM in predicting hospital LoS for planned admissions demonstrates promising predictive capabilities for stays of varying durations. It underscores the advantage of standardized data in achieving reproducible results.

XSL•FO
RenderX

This approach should serve as a model for enhancing operational efficiency and patient care coordination across health care settings.

(J Med Internet Res 2024;26:e59260) doi:10.2196/59260

KEYWORDS

length of stay; machine learning; Observational Medical Outcomes Partnership Common Data Model; allocation of resources; reproducibility of results; hospital; admission; retrospective study; prediction model; electronic health record; EHR; South Korea; logistic regression; algorithm; Shapley Additive Explanation; health care; clinical informatics

Introduction

Length of stay (LoS) in the health care system directly impacts optimal patient care provision, health care costs, and overall hospital efficiency [1]. An accurate prediction of LoS allows health care institutions to optimize utilization, maximize the availability of limited hospital resources, and ensure the effective management of hospital personnel [2,3]. Precise predictions can also facilitate the identification of patients at risk for long-term hospitalization, facilitating timely interventions to improve treatment [4]. Consequently, LoS optimization is widely recognized as an essential strategy for efficiently managing hospital resources and the overall quality of health care services [5,6].

Thus far, however, hospital LoS prediction models have focused primarily on patients with specific diseases and binary outcomes within certain specialties, limiting both their applicability across broader hospital settings and the precision of day-specific LoS analysis [7-10]. Levin et al [11] developed an automated discharge prediction model for use in the multidisciplinary rounding process. Sotoodeh et al [12] proposed a framework based on a hidden Markov model to predict LoS for patients in an intensive care unit using physiological measurements. Deep learning and artificial neural networks, incorporating the entire electronic health record (EHR) and clinical free-text notes, have been utilized to predict LoS outcomes, specifically targeting medical-surgical patients [9]. The Bayesian network model also predicted an expected discharge within 7 days for emergency department (ED) admissions [13,14].

Furthermore, previous LoS prediction models, constrained by single-center nonstandardized data with limited covariates, reduced generalizability. Historically, models often included limited clinical features and overlooked aspects such as hospitalist services and the characteristics of the hospital stay, limiting the scope and accuracy of information in the developed models. LoS is reflected by multiple factors, such as environmental conditions, health care system dynamics, and social drivers of health [15,16]. In contrast, previous models for predicting LoS mostly centered on demographics, lab results, and vital signs. Traditional LoS models ignored admission details, procedure codes, medications, and comorbidities, which are crucial for understanding patient recovery [1]. Additionally, prior studies often relied on data from a single medical center and nonstandardized sources like administrative data, the Medical Information Mart for Intensive Care (MIMIC), and EHRs for model development, leading to models with limited transferability between institutions due to varied data standards [17-19].

This study used standardized observational clinical data to address these challenges. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is a standardized framework designed to streamline and harmonize health care data from diverse sources [20]. The OMOP CDM strengthens the semantic interoperability of data and enables large-scale studies, including various health care research applications like predictive modeling [21,22]. Specifically, it enhances the reproducibility of research findings and ensures comparability of results, allowing researchers to seamlessly work with data from various health care systems [23]. It also reduces the effort required for data mapping and transformation, thereby streamlining the process and improving the prediction model's overall efficiency [22,23]. Adopting OMOP CDM in health care research contributes to more robust, scalable, and collaborative studies, ultimately improving the quality of health care research and patient care.

In this study, we aimed to develop and evaluate machine learning (ML)–based models for predicting hospital LoSs of 7 days or more for planned admissions using the OMOP CDM data. This study proposes multiclassification models for general hospitalization, targeting LoS durations of 3,4,5,6, and 7 days to reflect the complex clinical spectrum of health care conditions. These ML models are expected to facilitate an accurate prediction of LoS in various health care settings, aiding health care services in proactively implementing preventive measures to avoid unnecessary extensions of stay.

Methods

Data Source and Participants

We carried out an observational cohort study using EHR data in the OMOP CDM format (version 5.3) from 2 tertiary hospitals, the Seoul National University Bundang Hospital (SNUBH) for development and internal validation and Seoul National University Hospital (SNUH) for external validation. SNUBH and SNUH are tertiary general hospitals affiliated with universities in the Seoul metropolitan area and the capital city, respectively. They provide services for inpatients, outpatients, and ED patients. The EHR data from both hospitals encompass patient demographics, diagnoses, chief complaints, outpatient drug prescriptions, inpatient and ED drug administrations, operations, vital signs, laboratory test results, and in-hospital deaths [24]. Both hospitals use the same EHR system and have similar ETL processes for OMOP CDM. Out of a total of 961,672 admission episodes, 137,437 (14.3%) indexed hospitalizations with LoS ranging from 2 to 30 days between January 2016 and December 2020 were included in the study. The prediction timing was set within 30 days, aligning with

XSL•FO
RenderX

standard practices in most studies [25]. We included individuals who were over 18 years old, had at least 1 inpatient visit occurrence, were hospitalized in or after 2016, had at least 1 occurrence of any medical condition, survived hospitalization, and had no unplanned admissions during the study period. The study cohort of patients is provided in Figure 1.

Figure 1. Flowchart for model development and external validation study populations. A. Internal validation was conducted at Seoul National University Bundang Hospital (SNUBH) B. External validation was conducted at Seoul National University Hospital (SNUH). CDM: Common Data Model; LoS: length of stay.

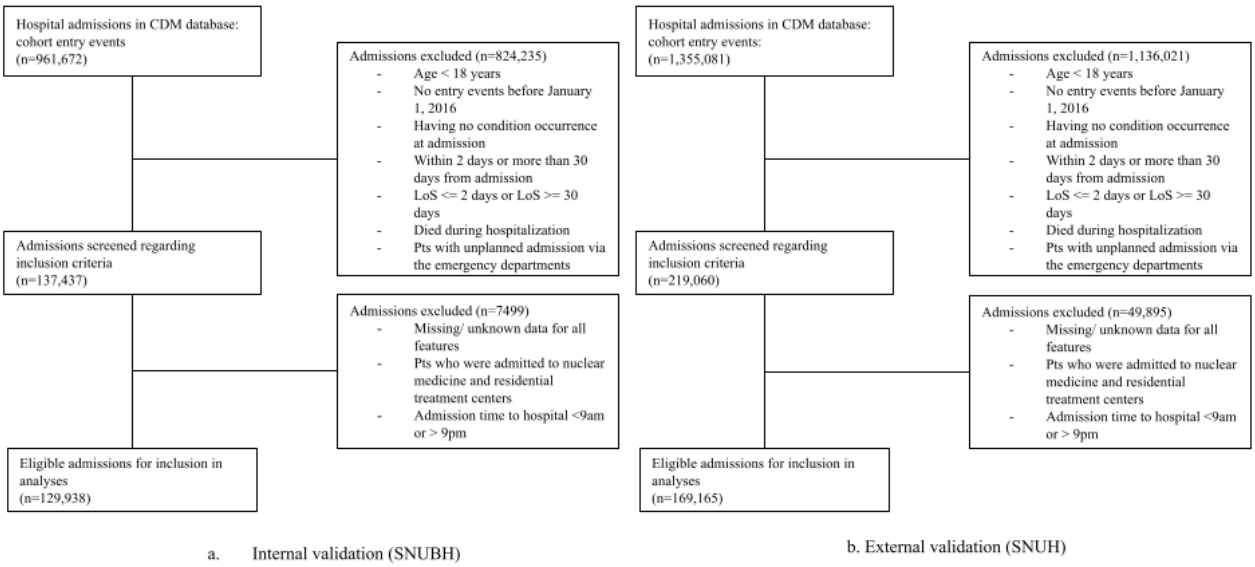

Ethical Considerations

Each participating site (SNUBH and SNUH) obtained an institutional review board (IRB) exemption (IRB no X-2103-672-907, IRB no E-2305-020-1429). Informed consent was waived due to the use of deidentified data within a secure internal network and the retrospective nature of the study.

Data Extraction and Predictors

We used an open-source Feature Extraction R package (version 3.0.1; R Foundation for Statistical Computing) to extract patients' clinical data before hospital admission from the OMOP CDM at SNUBH. Patient clinical characteristics, including condition occurrence, drug, observation, measurement, procedure, death, and visit occurrence tables were collected. The covariates included sex, age, the department at the visit, surgical operations, and lab results. After analyzing the LoS distribution for all patients, we excluded 2 departments, ophthalmology and nuclear medicine, from the model development due to their significantly shorter LoS (Figure S1 in Multimedia Appendix 1). Additional covariates that relied on the number of days (–365, –180, and –30) relative to the index date included diagnosis history, visit occurrences, and drug history. Severity scores such as the Charlson Comorbidity Index, Diabetes Complications Severity Index, CHADS2, and CHADS2-VAsc were also included as covariates. Additionally, custom covariates, such as the time, day of the week of admission, admission on holidays, day of the week of operation, whether operations were performed, diastolic and systolic blood pressure, body temperature, diagnosis at admission, and anthropometric measurements (eg, height and weight) were incorporated in the development of the prediction models.

Outcomes

The study classified hospitalizations lasting 7 days or longer using binary classification. In comparison, multiclassification was applied to categorize hospital stays lasting 3 days or longer, with specific subgroups for 3, 4, 5, 6, and 7 days. A hospitalization was defined as an inpatient visit during the study period, excluding unplanned admissions such as hospitalization through the emergency room [26].

Feature Selection and Model Development

After generating all the custom and noncustom covariates, logistic regression (LR) with Lasso regularization eliminated redundant collinear predictors from the models. We normalized continuous variables for uniform feature representation during data preprocessing and converted categorical variables into binary vectors using one-hot encoding. We removed variables from the analysis that had over 30% of the values missing [27]. Six ML models [5,28], including LR, random forest (RF), Extreme Gradient Boosting (XGB), Light Gradient Boosting (LGB), Gradient Boosting (GB), and multilayer perceptron (MLP) were evaluated to develop predictions for binary (LoS ≥7 days) and multiclass length of stay (LoS 3, 4, 5, 6 ≥7 days) outcomes based on the selected features. Model hyperparameters were optimized using 10-fold cross-validation via RandomizedSearchCV from Python (v.3.7.6) and scikit-learn library (v.0.22.1). We externally validated the models using SNUH data.

Model Evaluation

We evaluated model performance using 5 metrics: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity, precision, and specificity. A bootstrap resampling technique with 1000 replicates was employed to compute 95% CIs for all metrics.

XSL•FO
RenderX

The optimal threshold point of the Youden Index was used to assess the sensitivity and specificity of an LoS of 7 days or longer. Shapley Additive Explanations (SHAP) analysis interpreted the output of ML models and identified predictors that positively and negatively contribute to the model's prediction. Furthermore, calibration plots assessed the reliability of the prediction models and ascertained whether the predicted probabilities aligned with the actual outcomes across different probabilities. The Brier Score measured the accuracy of probabilistic predictions quantitatively. The performance of the models at SNUH was evaluated using AUROC and AUPRC, SHAP, and calibration plots alongside Brier Scores.

Statistical Analysis

Statistical analysis was conducted using R software (version 3.0.1) and Python (version 3.8.10).

Patient entry events consisted of 2 groups based on LoS <7 days and LoS ≥7 days. For the internal data set, comparisons between these groups were performed using standard 2-sample *t* tests for continuous variables and chi-squared tests of independence for categorical variables. To compare characteristics between internal and external validation cohorts, we used the *tableone* package in R, which generated descriptive statistics presenting continuous variables as mean (SD) and categorical variables as frequency counts and percentages. All the packages used are listed in Table S2 in Multimedia Appendix 1.

Results

Baseline Characteristics

During the study period, we identified 129,938 inpatient admission episodes at SNUBH and 169,165 at SNUH, each with an LoS ranging from 2 to 30 days. The mean age of the admitted patients were 58.4 (SD 15.4) years at SNUBH and 59.2 (SD 14.9) years at SNUH. The average LoS for planned admissions was 5.2 days at SNUBH and 5.4 days at SNUH. The baseline characteristics of both cohorts are presented in Table 1.

XSL•FO
RenderX

Table 1. Baseline characteristics and outcomes of planned admissions in the test and validation cohorts.

Characteristics	Internal validation set (SNUBH[a])			External validation set (SNUH[b])		
	LoS[c]< 7 (n=95,499)	LoS>=7 (n=34,439)	Overall (n=129,938)	LoS <7 (n=122,137)	LoS >=7 (n=47,028)	Overall (n=169,165)
Age (years), mean (SD)	57.5 (15.5)	60.9 (15)	58.4 (15.4)	59.4 (14.6)	58.8 (15.5)	59.2 (14.9)
Male sex, n (%)	42,117 (44.1)	16,762 (48.7)	58,879 (45.3)	63,956 (52.4)	21,986 (46.8)	85,942 (50.8)
Height (cm), mean (SD)	162 (8.75)	161 (9.08)	162 (8.84)	162 (8.70)	162 (8.86)	162 (8.75)
Weight (kg), mean (SD)	63.6 (12.7)	62.9 (12.4)	63.4 (12.6)	63.3 (12)	62.5 (12.3)	63 (12.1)
Body temperature, mean (SD)	36.7 (0.4)	36.8 (0.4)	36.7 (0.4)	36.4 (0.392)	36.5 (0.446)	36.5 (0.409)
Care site, n (%)						
Allergy	213 (0.2)	67 (0.2)	280 (0.2)	173 (0.1)	83 (0.2)	256 (0.2)
Anesthesiology	156 (0.2)	27 (0.1)	183 (0.1)	38 (0)	8 (0)	46 (0)
Cardiology	5408 (5.7)	433 (1.3)	5841 (4.5)	6580 (5.4)	565 (1.2)	7145 (4.2)
Dermatology	32 (0)	16 (0)	48 (0)	317 (0.3)	50 (0.1)	367 (0.2)
Endocrinology	564 (0.6)	185 (0.5)	749 (0.6)	812 (0.7)	337 (0.7)	1149 (0.7)
Gastroenterology	8326 (8.7)	1418 (4.1)	9744 (7.5)	27,452 (22.5)	3163 (6.7)	30,615 (18.1)
General surgery	15,299 (16)	8275 (24)	23,574 (18.1)	10,878 (8.9)	12,273 (26.1)	23,151 (13.7)
Hematology/oncology	11,961 (12.5)	3828 (11.1)	15,798 (12.2)	22,419 (18.4)	4761 (10.1)	27,180 (16.1)
Infectious diseases	135 (0.1)	222 (0.6)	357 (0.3)	132 (0.1)	178 (0.4)	310 (0.2)
Internal medicine	931 (1)	380 (1.1)	1,311 (1)	284 (0.2)	225 (0.5)	509 (0.3)
Nephrology	1477 (1.5)	468 (1.4)	1945 (1.5)	1606 (1.3)	1138 (2.4)	2744 (1.6)
Neurology	3147 (3.3)	621 (1.8)	3768 (2.9)	2102 (1.7)	972 (2.1)	3074 (1.8)
Psychiatry	258 (0.3)	1037 (3)	1295 (1)	832 (0.7)	1478 (3.1)	2310 (1.4)
Neurosurgery	6101 (6.4)	5200 (15.1)	11,301 (8.7)	4391 (3.6)	2300 (4.9)	6691 (4)
Obstetrics/gynecology	15,137 (15.9)	1484 (4.3)	16,621 (12.8)	10,564 (8.6)	3516 (7.5)	14,080 (8.3)
Ophthalmology	891 (0.9)	32 (0.1)	923 (0.7)	5007 (4.1)	169 (0.4)	5176 (3.1)
Orthopedic surgery	8235 (8.6)	3158 (9.2)	11,393 (8.8)	3926 (3.2)	5382 (11.4)	9308 (5.5)
Otolaryngology	5712 (6)	492 (1.4)	6204 (4.8)	5470 (4.5)	777 (1.7)	6247 (3.7)
Pediatrics	153 (0.2)	28 (0.1)	181 (0.1)	1084 (0.9)	715 (1.5)	1799 (1.1)
Physiatry	208 (0.2)	962 (2.8)	1170 (0.9)	61 (0.1)	233 (0.5)	294 (0.2)
Plastic surgery	749 (0.8)	1056 (3.1)	1805 (1.4)	1802 (1.5)	1572 (3.3)	3374 (2)
Pulmonary disease	3098 (3.2)	540 (1.6)	3638 (2.8)	2772 (2.3)	1002 (2.1)	3774 (2.2)
Rheumatology	246 (0.3)	172 (0.5)	418 (0.3)	1041 (0.9)	458 (1)	1499 (0.9)
Thoracic surgery	3430 (3.6)	2576 (7.5)	6006 (4.6)	2584 (2.1)	3233 (6.9)	5817 (3.4)
Urology	3112 (3.3)	1610 (4.7)	4722 (3.6)	7721 (6.3)	1446 (3.1)	9167 (5.4)
Admission hour, n (%)						
9-12	15,280 (16)	6215 (18)	21,495 (16.5)	12,492 (10.2)	6352 (13.5)	18,844 (11.1)
12-15	57,431 (60.1)	19,892 (57.8)	77,323 (59.5)	53,044 (43.4)	22,512 (47.9)	75,556 (44.7)
15-18	19,401 (20.3)	7253 (21.1)	26,554 (20.5)	43,135 (35.3)	14,631 (31.1)	57,766 (34.1)
18-21	3387 (3.5)	1079 (3.1)	4466 (3.4)	13,466 (11)	3533 (7.5)	16,999 (10)
Day of week of hospital admission, n (%)						
Monday	20,048 (21)	6759 (19.6)	26,807 (20.6)	24,266 (19.9)	9399 (20)	33,665 (19.9)
Tuesday	16,790 (17.6)	6758 (19.6)	23,548 (18.1)	20,142 (16.5)	8966 (19.1)	29,108 (17.2)
Wednesday	16,914 (17.7)	6546 (19)	23,460 (18.1)	20,641 (16.9)	8833 (18.8)	29,474 (17.4)

XSL•FO
RenderX

Characteristics	Internal validation set (SNUBH[a])			External validation set (SNUH[b])		
	LoS[c]< 7 (n=95,499)	LoS>=7 (n=34,439)	Overall (n=129,938)	LoS <7 (n=122,137)	LoS >=7 (n=47,028)	Overall (n=169,165)
Thursday	15,896 (16.6)	4733 (13.7)	20,629 (15.9)	20,027 (16.4)	5733 (12.2)	25,760 (15.2)
Friday	6987 (7.3)	2867 (8.3)	9854 (7.6)	8374 (6.9)	3109 (6.6)	11,483 (6.8)
Saturday	3116 (3.3)	1213 (3.5)	4329 (3.3)	3801 (3.1)	1905 (4.1)	5706 (3.4)
Sunday	15,748 (16.5)	5563 (16.2)	21,311 (16.4)	24,886 (20.4)	9083 (19.3)	33,969 (20.1)
Surgical operation, yes	45,776 (47.6)	23,850 (69.3)	69,626 (53.6)	30,128 (24.7)	22,164 (47.1)	52,292 (30.9)
Severity scores, mean (SD)						
$CHADS_2$[d]VASc[e]	1.3 (1.1)	1.4 (1.2)	1.3 (1.1)	1.3 (1.1)	1.4 (1.2)	1.3 (1.1)
$CHADS_2$	0.6 (1)	0.7 (1.1)	0.6 (1)	0.6 (1)	0.6 (1)	0.6 (1)
DCSI[f]	0.5 (1.2)	0.6 (1.4)	0.6 (1.3)	0.7 (1.3)	0.7 (1.4)	0.7 (1.3)
CCI[g]	2.1 (2.5)	2.2 (2.2)	2.11 (2.4)	2.6 (2.6)	2.5 (2.4)	2.6 (2.5)

[a]SNUBH: Seoul National University Bundang Hospital.

[b]SNUH: Seoul National University Hospital.

[c]LoS: length of stay.

[d]$CHADS_2$: clinical prediction tool used to test congestive heart failure, hypertension, age, diabetes, and stroke.

[e]VASc: The expanded version of $CHADS_2$ that also tests for transient ischemic attack, vascular disease, age, and sex category.

[f]DSCI: Diabetes Comorbidity Severity Index.

[g]CCI: Romano's adaptation of the Charlson Index.

Performance in Internal Validation

In the internal validation of the prediction model, the XGB model achieved the top performance in planned admissions for binary classification, with an AUROC of 0.891 (95% CI 0.887-0.894) and an AUPRC of 0.819 (95% CI 0.813-0.826) (Table 2). The LGB model showed the highest performance in the multiclassification, with an AUROC of 0.901 (95% CI 0.898-0.904), 0.836 (95% CI 0.831-0.841), 0.750 (95% CI 0.742-0.758), 0.766 (95% CI 0.757-0.774), and 0.890 (95% CI 0.887-0.893) for LoS 3, 4, 5, 6, and ≥ 7 days, respectively, and a macro AUPRC of 0.556 (Table 3). Overall, the predictive performance for LoS consistently showed good discriminatory ability, achieving an AUROC >0.800 in both the binary and multiclassification scenarios (Figure 2A, Table 2).

XSL•FO
RenderX

Table 2. Model performance for binary outcomes in the internal and external validation sets. All numbers are presented with 95% CI.

Data set and models	AUROC[a]	AUPRC[b]	Sensitivity	Specificity	PPV[c]	NPV[d]
Internal validation (n=38,847)						
LR[e]	0.853 (0.850-0.857)	0.744 (0.736-0.752)	0.771 (0.764-0.778)	0.786 (0.780-0.790)	0.642 (0.634-0.650)	0.873 (0.869-0.877)
RF[f]	0.881 (0.878-0.885)	0.802 (0.795-0.808)	0.604 (0.596-0.613)	0.922 (0.918-0.925)	0.794 (0.786-0.801)	0.824 (0.819-0.828)
XGB[g]	0.891 (0.887-0.894)	0.819 (0.813-0.826)	0.686 (0.679-0.695)	0.896 (0.893-0.900)	0.768 (0.760-0.775)	0.852 (0.848-0.856)
GB[h]	0.888 (0.884-0.891)	0.811 (0.804-0.818)	0.681 (0.673-0.690)	0.894 (0.891-0.898)	0.763 (0.755-0.770)	0.849 (0.845-0.853)
LGB[i]	0.889 (0.886-0.893)	0.816 (0.811-0.824)	0.661 (0.653-0.669)	0.906 (0.902-0.910)	0.778 (0.770-0.786)	0.843 (0.839-0.847)
MLP[j]	0.882 (0.878-0.885)	0.804 (0.798-0.811)	0.635 (0.627-0.643)	0.910 (0.906-0.913)	0.778 (0.770-0.786)	0.833 (0.829-0.838)
External validation (n = 169,165)						
LR	0.760 (0.757-0.763)	0.508 (0.503-0.514)	0.630 (0.622-0.632)	0.775 (0.773-0.777)	0.445 (0.441-0.449)	0.878 (0.876-0.880)
RF	0.804 (0.802-0.807)	0.545 (0.540-0.551)	0.412 (0.408-0.417)	0.923 (0.921-0.924)	0.605 (0.601-0.611)	0.845 (0.843-0.846)
XGB	0.774 (0.772-0.777)	0.523 (0.518-0.529)	0.473 (0.468-0.479)	0.879 (0.877-0.881)	0.530 (0.524-0.535)	0.853 (0.851-0.855)
GB	0.765 (0.763-0.768)	0.499 (0.493-0.504)	0.481 (0.476-0.486)	0.863 (0.861-0.865)	0.503 (0.498-0.508)	0.852 (0.850-0.854)
LGB	0.798 (0.796-0.801)	0.542 (0.537-0.548)	0.492 (0.487-0.497)	0.888 (0.887-0.890)	0.560 (0.555-0.565)	0.859 (0.857-0.860)
MLP	0.789 (0.786-0.791)	0.532 (0.527-0.538)	0.449 (0.444-0.454)	0.890 (0.898-0.901)	0.563 (0.558-0.569)	0.850 (0.848-0.852)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]LR: logistic regression.

[f]RF: random forest

[g]XGB: Extreme Gradient Boosting.

[h]GB: Gradient Boosting.

[i]LGB: Light Gradient Boosting.

[k]MLP: Multilayer Perceptron.

XSL•FO
RenderX

Table 3. Model performance for multiclass classification in the internal and external validation sets. All numbers are presented with a 95% CI.

Model and class	Internal validation set (n=38,847)			External validation set (n=169,165)		
	AUROC[a]	AUPRC[b]	Macroaveraged AUPRC	AUROC	AUPRC	Macroaveraged AUPRC
LR[c]			0.486			0.342
day 3 (class 0)	0.861 (0.857-0.866)	0.693 (0.683-0.702)		0.718 (0.715-0.720)	0.546 (0.542-0.550)	
day 4 (class 1)	0.805 (0.799-0.811)	0.543 (0.531-0.555)		0.697 (0.694-0.700)	0.340 (0.335-0.345)	
day 5 (class 2)	0.714 (0.705-0.722)	0.216 (0.206-0.229)		0.571 (0.567-0.576)	0.139 (0.136-0.143)	
day 6 (class 3)	0.710 (0.700-0.719)	0.226 (0.213-0.240)		0.596 (0.591-0.601)	0.115 (0.112-0.118)	
day 7 ≥ (class 4)	0.853 (0.849-0.857)	0.754 (0.746-0.762)		0.751 (0.748-0.754)	0.570 (0.566-0.575)	
RF[d]			0.534			0.379
day 3 (class 0)	0.883 (0.880-0.887)	0.751 (0.742-0.759)		0.754 (0.751-0.756)	0.620 (0.616-0.625)	
day 4 (class 1)	0.822 (0.817-0.828)	0.588 (0.576-0.600)		0.720 (0.717-0.723)	0.366 (0.361-0.371)	
day 5 (class 2)	0.737 (0.729-0.745)	0.250 (0.237-0.264)		0.589 (0.585-0.594)	0.158 (0.154-0.162)	
day 6 (class 3)	0.751 (0.742-0.760)	0.296 (0.280-0.312)		0.627 (0.622-0.632)	0.138 (0.134-0.142)	
day 7 ≥ (class 4)	0.872 (0.869-0.875)	0.784 (0.778-0.791)		0.797 (0.794-0.799)	0.613 (0.608-0.617)	
XGB[e]			0.546			0.363
day 3 (class 0)	0.899 (0.896-0.903)	0.770 (0.761-0.778)		0.753 (0.751-0.755)	0.609 (0.605-0.613)	
day 4 (class 1)	0.829 (0.825-0.835)	0.595 (0.583-0.606)		0.703 (0.700-0.706)	0.343 (0.338-0.348)	
day 5 (class 2)	0.735 (0.727-0.744)	0.251 (0.239-0.265)		0.570 (0.565-0.574)	0.135 (0.131-0.138)	
day 6 (class 3)	0.751 (0.742-0.760)	0.298 (0.282-0.315)		0.600 (0.594-0.604)	0.123 (0.119-0.126)	
day 7 ≥ (class 4)	0.887 (0.884-0.891)	0.814 (0.808-0.821)		0.785 (0.782-0.787)	0.607 (0.602-0.611)	
GB[f]			0.538			0.326
day 3 (class 0)	0.900 (0.896-0.903)	0.766 (0.757-0.774)		0.718 (0.716-0.721)	0.562 (0.558-0.567)	
day 4 (class 1)	0.829 (0.824-0.835)	0.593 (0.580-0.605)		0.621 (0.618-0.624)	0.253 (0.249-0.257)	
day 5 (class 2)	0.735 (0.727-0.743)	0.236 (0.225-0.248)		0.600 (0.595-0.604)	0.146 (0.143-0.150)	
day 6 (class 3)	0.754 (0.745-0.762)	0.285(0.270-0.301)		0.564 (0.559-0.569)	0.103 (0.101-0.106)	
day 7 ≥ (class 4)	0.888 (0.885-0.891)	0.812 (0.806-0.819)		0.754 (0.751-0.756)	0.565 (0.560-0.570)	
LGB[g]			0.556			0.383
day 3 (class 0)	0.901 (0.898-0.904)	0.770 (0.762-0.779)		0.766 (0.764-0.768)	0.626 (0.622-0.630)	

XSL•FO
RenderX

Model and class	Internal validation set (n=38,847)			External validation set (n=169,165)		
	AUROC[a]	AUPRC[b]	Macroaveraged AUPRC	AUROC	AUPRC	Macroaveraged AUPRC
day 4 (class 1)	0.836 (0.831-0.841)	0.606 (0.594-0.618)		0.724(0.721-0.727)	0.361 (0.375-0.386)	
day 5 (class 2)	0.750 (0.742-0.758)	0.270 (0.256-0.284)		0.595 (0.590-0.599)	0.146 (0.143-0.149)	
day 6 (class 3)	0.766 (0.757-0.774)	0.318 (0.302-0.335)		0.616 (0.611-0.621)	0.132 (0.128-0.135)	
day 7 ≥ (class 4)	0.890 (0.887-0.893)	0.818 (0.812-0.824)		0.808 (0.806-0.810)	0.632 (0.628-0.637)	
MLP[h]			0.523			0.346
day 3 (class 0)	0.887 (0.883-0.890)	0.736 (0.727-0.745)		0.735 (0.732-0.737)	0.582 (0.578-0.586)	
day 4 (class 1)	0.821 (0.816-0.827)	0.574 (0.561-0.586)		0.685 (0.682-0.689)	0.333(0.328-0.338)	
day 5 (class 2)	0.730 (0.722-0.739)	0.239 (0.228-0.253)		0.590 (0.586-0.594)	0.137 (0.134-0.140)	
day 6 (class 3)	0.736 (0.727-0.745)	0.276 (0.260-0.291)		0.603 (0.598-0.608)	0.120 (0.117-0.123)	
day 7 ≥ (class 4)	0.874 (0.871-0.878)	0.791 (0.784-0.798)		0.747(0.745-0.750)	0.559 (0.554-0.564)	

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]LR: logistic regression.

[d]RF: random forest

[e]XGB: Extreme Gradient Boosting.

[f]GB: Gradient Boosting.

[g]LGB: Light Gradient Boosting.

[h]MLP: Multilayer Perceptron.

Figure 2. Receiver operating characteristics curves (AUROCs) for length of stay binary classification in the internal validation set. A. Internal validation was conducted at Seoul National University Bundang Hospital (SNUBH) B. External validation was conducted at Seoul National University Hospital (SNUH). CDM: Common Data Model; LoS: length of stay. LR: logistic regression; RF: random forest; XGB: Extreme Gradient Boosting; GB: Gradient Boosting; LGB: Light Gradient Boosting; MLP: Multilayer Perceptron. ROC: receiver operating characteristic.

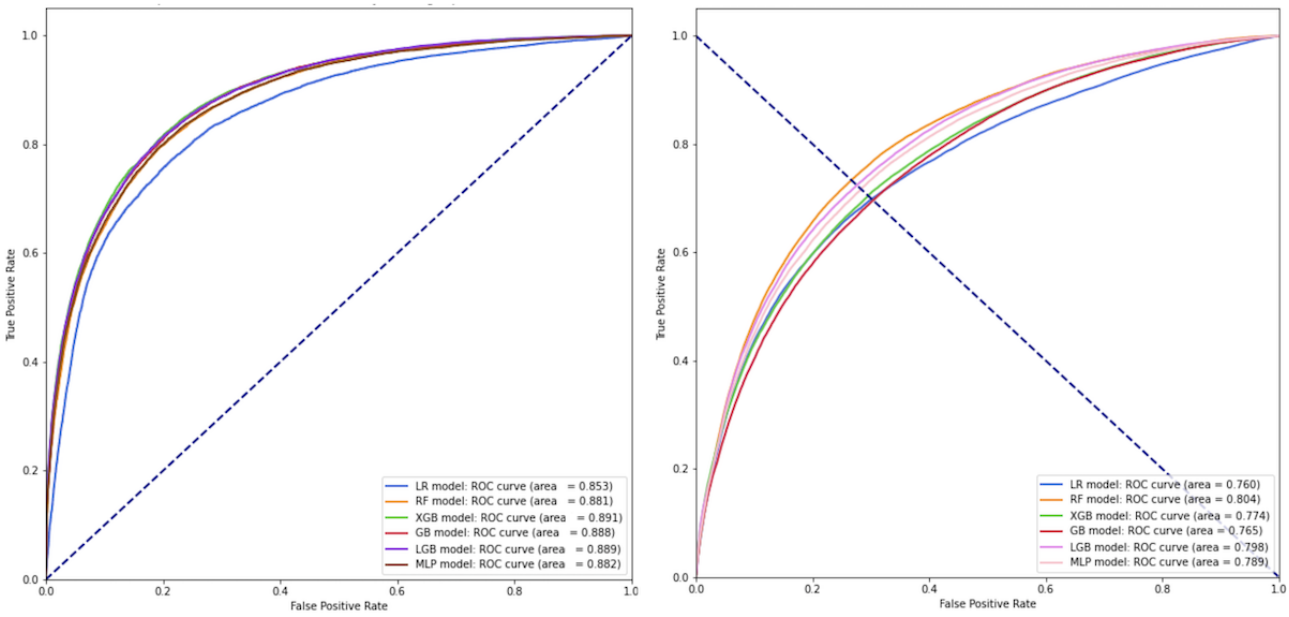

a. Internal validation (SNUBH)

b. External validation (SNUH)

XSL•FO
RenderX

SHAP Calibration Plots and Brier Scores in Internal Validation

The most important features contributing to the predictive performance of the models were the operation performed, the admitted patient's department, age, and severity scores such as the Charlson Index-Romano adaptation and the frequency of outpatient visits in the past 6 months (Figure 3A). In the multiclassification models, the SHAP analysis revealed distinct predictors, including surgical operations, blood pressure, and day of the week of hospital admission, which significantly influenced different hospital stay durations (Figure S2 in Multimedia Appendix 1). Specifically, surgical operations were significant predictors for LoS of 3 days and 7 days or longer, while admissions in obstetrics/gynecology and neurosurgery were relevant for LoS of 4 and 5 days. Additionally, hospital admissions on Wednesdays and Fridays were associated with a 6-day LoS (Figure S2 in Multimedia Appendix 1). The calibration plot and Brier scores indicated strong calibration performance, achieving approximately 0.14 scores in both the binary and multiclassification models (Figure 4A, Table S3 in Multimedia Appendix 1).

Figure 3. Shapley Additive Explanations (SHAP) feature importance analysis for the model predicting length of stay of 7 days or longer. A. Internal validation was conducted at Seoul National University Bundang Hospital (SNUBH) B. External validation was conducted at Seoul National University Hospital (SNUH).

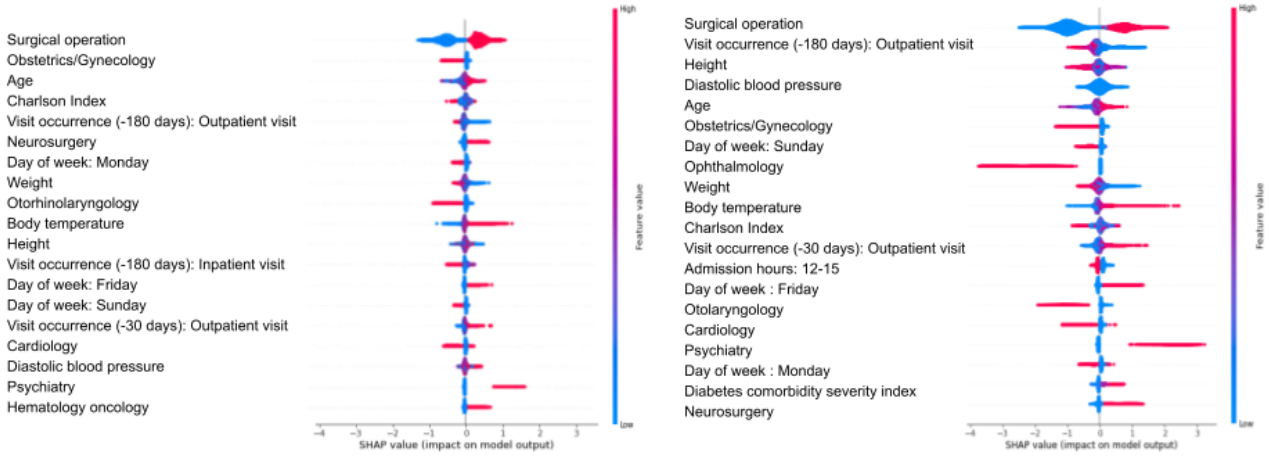

a. Internal validation (SNUBH)

b. External validation (SNUH)

Performance in the External Validation

In the external validation, the RF model performed the best in predicting an LoS of 7 days or longer with an AUROC of 0.804 (95% CI 0.802-0.807) and an AUPRC of 0.545 (95% CI 0.540-0.551) (Table 3). The AUROC decreased by 8% compared to the AUROC obtained from SNUBH (Figure 2B). The LGB model showed superior overall performance for multiclassification, with an average AUROC of 0.711 and a macro AUPRC of 0.383 across different LoS durations (Table 3). The AUROC showed a 0.93 reduction compared to the internal validation multiclassification model.

SHAP, Calibration Plots, and Brier Scores in the External Validation

The SHAP analysis identified significant predictors for hospital stays of 7 days or longer, including surgical procedures, elevated body temperatures, frequency of outpatient visits in the past month, hospital admissions on Fridays, and older age (Figure 3B). In the multiclassification analysis, akin to the prediction model's internal validation, cases with surgery performed showed an increased LoS in the hospital. Admission to the cardiology department generally led to discharge within 3 days. Specifically, admissions to the obstetrics/gynecology and hematology/oncology departments were associated with an average LoS of 4 and 5 days, respectively. Hospital admissions on Sundays and a higher number of outpatient visits in the past month resulted in hospital stays lasting more than 6 days (Figure S3 in Multimedia Appendix 1). The binary and multiclassification models accurately calibrated predicted probabilities and actual outcomes on the external validation data set (Figure 4B, Table S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 4. Calibration of length of stay prediction models for 7 days or longer in internal and external validation sets. A. Internal validation was conducted at Seoul National University Bundang Hospital (SNUBH) B. External validation was conducted at Seoul National University Hospital (SNUH). LR: logistic regression; RF: random forest; XGB: Extreme Gradient Boosting; GB: Gradient Boosting; LGB: Light Gradient Boosting; MLP: Multilayer Perceptron.

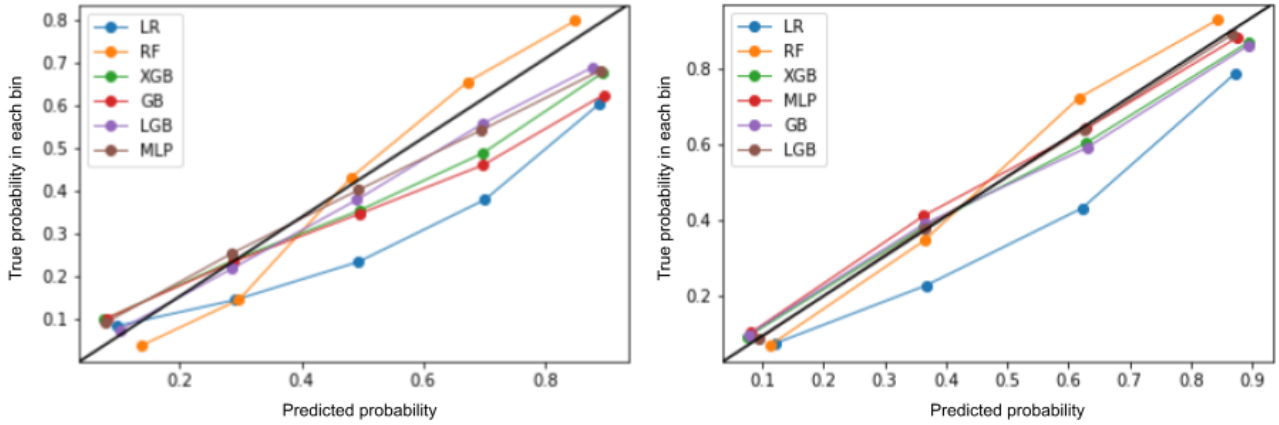

a. Internal validation (SNUBH)

b. External validation (SNUH)

Discussion

Principal Findings

In this paper, we proposed two ML models using the standardized vocabularies of the OMOP CDM: (1) a binary classification model predicting LoS of 7 days or more and (2) a multiclassification model categorizing LoS as 3, 4, 5, 6, and 7 days or longer. This study demonstrates that combining clinical and nonclinical characteristics from a standardized data source from different institutions yields moderate predictive accuracy and efficiently leverages the benefits of multicenter studies. To our knowledge, our model is the first to offer both binary and multiclass classification predictions for LoS using the OMOP CDM.

The ML algorithms using GB performed the highest binary classification performance, while the LGB model performed the best in the multiclass classification. The ML predictions identified surgical operation, diastolic blood pressure, and the day of week of hospital admissions as critical factors for predicting both an LoS of 7 days or more and for multiclass classification of LoS. The external validation revealed a model performance 8% lower than the internally calculated AUROC value. This decline in performance could be attributed to variations in patient populations and medical specialties across hospitals and disparities in converting EHR data into the standardized OMOP CDM format [29-31].

In this study, we conducted an extensive analysis across 46 departments, developing a model that predicted LoS from 3 to 7 days or longer. A multidepartment approach unraveled the complex dynamics of patient stays and illuminated the varied factors influencing each day's length of stay. Insights from this broad patient spectrum could equip health care professionals with essential tools for improving patient care strategies in various health care settings. These models may aid in optimizing health care resources and contribute to identifying patients most likely to be discharged on a particular day using classification algorithms.

Additionally, our LOS prediction model offers potential economic benefits by enabling better planning and management of bed occupancy, reducing costs associated with the inefficient use of hospital beds. By anticipating extended stays, hospitals can arrange necessary interventions in advance, potentially preventing prolonged hospitalizations and lowering treatment costs. Streamlining discharge processes based on accurate predictions reduces delays and frees up beds more quickly, leading to cost savings in terms of staffing and resource management. Although further studies are needed to empirically validate these benefits, our findings suggest that implementing such predictive models could support clinical and financial improvements in hospital management.

Earlier studies often used data sources such as administrative data, the MIMIC, and EHRs for model development, resulting in limited applicability across different institutions due to heterogeneous data standards (Table S1 in Multimedia Appendix 1). The variations in these data sources hindered scalability and the capacity to efficiently manage a wide range of routinely collected observational health data. Previously utilized data constrained the transferability of the model outputs to different patient populations across multiple medical specialties and institutions [5,17,32]. We designed prediction models that implemented the OMOP CDM standardized vocabularies and demonstrated their application across independent institutions, ensuring interoperability and applicability. Although one study predicted LoS from multicenter data using federated learning, it relied solely on administrative data and trained with 3 regression models, suggesting multicollinearity issues [33]. Unlike the aforementioned model, ours included hospital stay characteristics and patients' medications incorporating more records than the federated learning research. As a result, our model, including hospital admission information and comprehensive patient history, provides more reliable insights, covering a broader range of records.

Systematic reviews assessing efficacious health system interventions to avoid prolonged hospital LoS in high-risk populations indicate that reducing LoS requires a multifaceted approach, including clinical care and the logistics of care

XSL•FO
RenderX

coordination [4]. However, most previous studies primarily used demographics, lab results, and vital signs as input features [1]. Some authors integrated data from different modalities, combining free-text notes and demographic information with time-varying features, while only a few employed medications as predictive variables [1]. In this study, we incorporated clinical and nonclinical features, including medical specialties, severity scores, day and time of admission, and hospital admissions during the weekend. Time-dependent covariates classified as occurring in the past 1, 6, and 12 months include frequency of different types of visits, active ingredients in a drug group, and medical conditions to capture health care utilization patterns. This comprehensive data integration provides a longitudinal view of the patient, enabling more personalized care planning and effective discharge strategies tailored to each patient's recovery.

When comparing prediction models for a 7-day LoS using the OMOP CDM within the 2 tertiary hospitals, we found that surgical operation, admitting unit, patient age, severity scores, and admission day significantly influenced hospital stay duration. For example, our internal validation revealed that older patients admitted to the neurosurgery department and undergoing surgery often have longer LoS, aligning with the notion that older adult patients requiring neurosurgery tend to have poorer prognosis and generally require extended hospitalization [34]. In contrast, admissions to obstetrics/gynecology appeared to have a shorter LoS, likely because they are mostly elective surgeries, such as cesarean deliveries [35,36]. A lower frequency of outpatient visits in the past 6 months was associated with a longer LoS, indicating that patients who visit the hospital less frequently may not be managing their health effectively. Friday admissions, typically through the emergency room, were more likely to involve severe conditions, while the unavailability of elective surgeries during the weekend could result in a longer LoS [37]. Weekend admissions, particularly on Sundays and Mondays, were associated with shorter LoS, likely due to the scheduling of preplanned treatments and elective surgeries on weekdays [38]. Additionally, certain medical specialties were linked to extended LoS. Psychiatric admission often resulted in longer stays, attributed to the need for comprehensive assessments and stabilization in mental health care [39]. Similarly, admissions to hematology/oncology departments frequently reported prolonged stays for monitoring and managing side effects and complications arising from hematologic cancer treatments [40]. Higher body temperatures indicated more severe conditions, potentially leading to longer durations of illness [41].

Limitations

This study has several limitations. The models were only externally validated in a single hospital, which may limit their applicability to other health care settings with different patient populations and characteristics. However, health care systems with OMOP CDM can readily implement the methods proposed in this study using standardized vocabularies and analytic codes [17,23,42]. This work aids in developing a model that predicts the remaining LoS based on hospitalization progression during the admission period within a heterogeneous population through the Observational Health Data Sciences and Informatics (OHDSI) network study. Additionally, smaller patient numbers in the 5- and 6-day LoS categories in both the internal and external validation sets (Table S5 in Multimedia Appendix 1) may have affected the model performance for these subgroups. The models may not fully capture real-world performance due to the limited consideration of local context and resources such as team dynamics, hospital processes, staffing resources, social drivers of health, and administrative support—limitations stemming from nonstandard fields in the OMOP CDM. Using the standardized information we could capture from the CDM, we developed a reproducible model based on standardized variables that are less influenced by individual hospitals, such as clinical factors and hospital stay characteristics.

Conclusion

In this study, we utilized the OMOP CDM to predict the LoS in planned admissions, considering both binary (LoS ≥ 7 days) and multiclassification scenarios (LoS of 3, 4, 5, 6, and 7 days or more). The performance of the prediction models was consistent and confirmed through external validation. The proposed models showed ease of application in various clinical settings and reproducibility across institutions within the OHDSI community. The findings of this study may assist hospitals in effectively managing hospital resources, such as staffing, equipment, and supplies at different institutions.

Acknowledgments

This research was supported by a grant from the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number HI22C0471).

Data Availability

Because this study utilized a distributed research network, the data are kept within each institution and are not publicly available. However, the source code for cohort and covariates generation is available in the GitHub repository [43].

Conflicts of Interest

None declared.

XSL•FO
RenderX

Authors' Contributions

HL contributed to data cleansing, analysis, modeling, and manuscript writing. SK and H-WM were responsible for data collection, curation, interpretation, and verification. KK provided the data sets for external validation. H-YL critically reviewed the final manuscript. SY designed and supervised the research. SYJ provided clinical direction and assisted with drafting and finalizing the manuscript. SYJ and SY equally supervised the entire process as corresponding authors. All authors accessed the data, reviewed the study results, and approved the final manuscript.

Multimedia Appendix 1
Supplementary material for the study.
[DOCX File , 2386 KB - jmir_v26i1e59260_app1.docx]

References

1. Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N. A systematic review of the prediction of hospital length of stay: Towards a unified framework. PLOS Digit Health 2022 Apr 14;1(4):e0000017 [FREE Full text] [doi: 10.1371/journal.pdig.0000017] [Medline: 36812502]
2. Chang K, Tseng M, Weng H, Lin Y, Liou C, Tan T. Prediction of length of stay of first-ever ischemic stroke. Stroke 2002 Nov;33(11):2670-2674. [doi: 10.1161/01.str.0000034396.68980.39] [Medline: 12411659]
3. Bueno H, Ross JS, Wang Y, Chen J, Vidán MT, Normand SLT, et al. Trends in length of stay and short-term outcomes among Medicare patients hospitalized for heart failure, 1993-2006. JAMA 2010 Jun 02;303(21):2141-2147 [FREE Full text] [doi: 10.1001/jama.2010.748] [Medline: 20516414]
4. Siddique SM, Tipton K, Leas B, Greysen SR, Mull NK, Lane-Fall M, et al. Interventions to reduce hospital length of stay in high-risk populations: a systematic review. JAMA Netw Open 2021 Sep 01;4(9):e2125846 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.25846] [Medline: 34542615]
5. Jaotombo F, Pauly V, Fond G, Orleans V, Auquier P, Ghattas B, et al. Machine-learning prediction for hospital length of stay using a French medico-administrative database. J Mark Access Health Policy 2023 Nov 26;11(1):2149318 [FREE Full text] [doi: 10.1080/20016689.2022.2149318] [Medline: 36457821]
6. Gutierrez JMP, Sicilia M, Sanchez-Alonso S, Garcia-Barriocanal E. Predicting length of stay across hospital departments. IEEE Access 2021;9:44671-44680. [doi: 10.1109/access.2021.3066562]
7. Olle A. Predicting patient length of stay at time of admission using machine learning [thesis]. KTH Royal Institute of Technology School of Engineering Sciences in Chemistry, Biotechnology and Health. 2019. URL: https://www.diva-portal.org/smash/get/diva2:1338294/FULLTEXT01.pdf [accessed 2024-11-14]
8. Wu J, Lin Y, Li P, Hu Y, Zhang L, Kong G. Predicting prolonged length of ICU stay through machine learning. Diagnostics (Basel) 2021 Nov 30;11(12):2242 [FREE Full text] [doi: 10.3390/diagnostics11122242] [Medline: 34943479]
9. Ma X, Si Y, Wang Z, Wang Y. Length of stay prediction for ICU patients using individualized single classification algorithm. Comput Methods Programs Biomed 2020 Apr;186:105224. [doi: 10.1016/j.cmpb.2019.105224] [Medline: 31765937]
10. Lee J, Hsieh C, Lin C, Lin Y, Kao C. Prediction of hospitalization using artificial intelligence for urgent patients in the emergency department. Sci Rep 2021 Sep 30;11(1):19472 [FREE Full text] [doi: 10.1038/s41598-021-98961-2] [Medline: 34593930]
11. Levin S, Barnes S, Toerper M, Debraine A, DeAngelo A, Hamrock E, et al. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. BMJ Innov 2020 Dec 21;7(2):414-421. [doi: 10.1136/bmjinnov-2020-000420]
12. Sotoodeh M, Ho J. Improving length of stay prediction using a hidden Markov model. 2019 Presented at: AMIA Summits on Translational Science; May 6; Atlanta, GA p. 425.
13. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 08;1(1):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
14. Cai X, Perez-Concha O, Coiera E, Martin-Sanchez F, Day R, Roffe D, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. J Am Med Inform Assoc 2016 May;23(3):553-561 [FREE Full text] [doi: 10.1093/jamia/ocv110] [Medline: 26374704]
15. Eskandari M, Alizadeh Bahmani AH, Mardani-Fard HA, Karimzadeh I, Omidifar N, Peymani P. Evaluation of factors that influenced the length of hospital stay using data mining techniques. BMC Med Inform Decis Mak 2022 Oct 29;22(1):280 [FREE Full text] [doi: 10.1186/s12911-022-02027-w] [Medline: 36309751]
16. Tipton K, Leas B, Mull N. Interventions to decrease hospital length of stay. https://www.ncbi.nlm.nih.gov/books/NBK574435/. 2021 Sep. URL: https://www.ncbi.nlm.nih.gov/books/NBK574435/ [accessed 2022-12-10]
17. Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. BMC Med Inform Decis Mak 2021 Dec 18;21(1):351 [FREE Full text] [doi: 10.1186/s12911-021-01722-4] [Medline: 34922532]
18. Abujaber A, Fadlalla A, Nashwan A, El-Menyar A, Al-Thani H. Predicting prolonged length of stay in patients with traumatic brain injury: A machine learning approach. Intell-Based Med 2022;6:100052. [doi: 10.1016/j.ibmed.2022.100052]

XSL•FO
RenderX

19. Chiu W, Chan L, Masud JHB, Hong C, Chien Y, Hsu C, et al. Identifying risk factors for prolonged length of stay in hospital and developing prediction models for patients with cardiac arrest receiving targeted temperature management. Rev Cardiovasc Med 2023 Feb;24(2):55 [FREE Full text] [doi: 10.31083/j.rcm2402055] [Medline: 39077396]
20. Almeida JR, Silva LB, Bos I, Visser PJ, Oliveira JL. A methodology for cohort harmonisation in multicentre clinical research. Inform Med Unlocked 2021;27:100760. [doi: 10.1016/j.imu.2021.100760]
21. Carus J, Trübe L, Szczepanski P, Nürnberg S, Hees H, Bartels S, et al. Mapping the oncological basis dataset to the standardized vocabularies of a common data model: a feasibility study. Cancers (Basel) 2023 Aug 11;15(16):4059 [FREE Full text] [doi: 10.3390/cancers15164059] [Medline: 37627087]
22. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018 Aug 01;25(8):969-975 [FREE Full text] [doi: 10.1093/jamia/ocy032] [Medline: 29718407]
23. Kim J, Kim C, Kim K, Lee Y, Yu DH, Yun J, et al. Scalable infrastructure supporting reproducible nationwide healthcare data analysis toward FAIR stewardship. Sci Data 2023 Oct 04;10(1):674 [FREE Full text] [doi: 10.1038/s41597-023-02580-7] [Medline: 37794003]
24. Choi J, Yoo S, Song W, Kim S, Baek H, Lee JS, et al. Development and validation of a prognostic classification model predicting postoperative adverse outcomes in older surgical patients using a machine learning algorithm: retrospective observational network study. J Med Internet Res 2023 Nov 13;25:e42259 [FREE Full text] [doi: 10.2196/42259] [Medline: 37955965]
25. Gokhale S, Taylor D, Gill J, Hu Y, Zeps N, Lequertier V, et al. Hospital length of stay prediction tools for all hospital admissions and general medicine populations: systematic review and meta-analysis. Front Med (Lausanne) 2023 Aug 16;10:1192969 [FREE Full text] [doi: 10.3389/fmed.2023.1192969] [Medline: 37663657]
26. Huntley DA, Cho DW, Christman J, Csernansky JG. Predicting length of stay in an acute psychiatric hospital. Psychiatr Serv 1998 Aug;49(8):1049-1053. [doi: 10.1176/ps.49.8.1049] [Medline: 9712211]
27. McGee M, Bergasa N. Imputing missing data in clinical pilot studies. Southern Methodist University. 20. URL: https://www.smu.edu/dedman/academics/departments/statistics/research/~/media/b1cf5222f09d4708a6c89b5e5dd15113.ashx [accessed 2022-12-12]
28. Mekhaldi R, Caulier P, Chaabane S, Chraibi A, Piechowiak S. A comparative study of machine learning models for predicting length of stay in hospitals. J Inf Sci Eng 2021 Oct 22;37(5):1025-1038. [doi: 10.6688/JISE.202109_37(5).0003]
29. Glicksberg BS, Oskotsky B, Thangaraj PM, Giangreco N, Badgeley MA, Johnson KW, et al. PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. Bioinformatics 2019 Nov 01;35(21):4515-4518 [FREE Full text] [doi: 10.1093/bioinformatics/btz409] [Medline: 31214700]
30. Park K, Cho M, Song M, Yoo S, Baek H, Kim S, et al. Exploring the potential of OMOP common data model for process mining in healthcare. PLoS One 2023 Jan 3;18(1):e0279641 [FREE Full text] [doi: 10.1371/journal.pone.0279641] [Medline: 36595527]
31. Ryu B, Yoo S, Kim S, Choi J. Development of prediction models for unplanned hospital readmission within 30 days based on common data model: a feasibility study. Methods Inf Med 2021 Dec 28;60(S 02):e65-e75 [FREE Full text] [doi: 10.1055/s-0041-1735166] [Medline: 34583416]
32. Bahmani A, Alavi A, Buergel T, Upadhyayula S, Wang Q, Ananthakrishnan SK, et al. A scalable, secure, and interoperable platform for deep data-driven health management. Nat Commun 2021 Oct 01;12(1):5757 [FREE Full text] [doi: 10.1038/s41467-021-26040-1] [Medline: 34599181]
33. Rahman MM, Kundu D, Suha SA, Siddiqi UR, Dey SK. Hospital patients' length of stay prediction: A federated learning approach. J King Saud Univ-Com 2022 Nov;34(10):7874-7884. [doi: 10.1016/j.jksuci.2022.07.006]
34. Khan B, Haqqani U, Ullah S, Hamayun S, Bibi Z, Khanzada K. Duration of in-hospital stay for elective neurosurgical procedures in a tertiary care hospital. Cureus 2021 Jun;13(6):e15745 [FREE Full text] [doi: 10.7759/cureus.15745] [Medline: 34285851]
35. Jain RV, Jain V. Early recovery after surgery in elective obstetrics-gynaecology surgeries: prospective single-center pilot study. Int J Reprod Contracept Obstet Gynecol 2021 May 27;10(6):2302. [doi: 10.18203/2320-1770.ijrcog20212166]
36. Tamang T, Wangchuk T, Zangmo C, Wangmo T, Tshomo K. The successful implementation of the Enhanced Recovery After Surgery (ERAS) program among caesarean deliveries in Bhutan to reduce the postoperative length of hospital stay. BMC Pregnancy Childbirth 2021 Sep 18;21(1):637 [FREE Full text] [doi: 10.1186/s12884-021-04105-9] [Medline: 34537016]
37. Matoba M, Suzuki T, Ochiai H, Shirasawa T, Yoshimoto T, Minoura A, et al. Seven-day services in surgery and the "weekend effect" at a Japanese teaching hospital: a retrospective cohort study. Patient Saf Surg 2020 Jun 04;14(1):24 [FREE Full text] [doi: 10.1186/s13037-020-00250-w] [Medline: 32518591]
38. Khaliq AA, Broyles RW. Hospital admissions: who is admitted through the emergency department? Health Serv Manage Res 2006 Feb 01;19(1):13-22. [doi: 10.1258/095148406775322034] [Medline: 16438783]

XSL•FO
RenderX

39. Baeza FL, da Rocha NS, Fleck MP. Predictors of length of stay in an acute psychiatric inpatient facility in a general hospital: a prospective study. Braz J Psychiatry 2018 Jul 06;40(1):89-96 [FREE Full text] [doi: 10.1590/1516-4446-2016-2155] [Medline: 28700014]
40. Alkhaldy HY, Yahya AO, Algarni AM, Bakheet OSE, Assiri M, Saboor M. JAK2 mutation assessment in thrombotic events at unusual anatomical sites: insights from a high-altitude cohort. Int J Gen Med 2024 Oct 09;17:4551-4558 [FREE Full text] [doi: 10.2147/IJGM.S480705] [Medline: 39398483]
41. Erkens R, Wernly B, Masyuk M, Muessig J, Franz M, Schulze P, et al. Admission body temperature in critically ill patients as an independent risk predictor for overall outcome. Med Princ Pract 2019 Dec 2;29(4):389-395. [doi: 10.1159/000505126]
42. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed 2021 Nov;211:106394 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106394] [Medline: 34560604]
43. Length of stay prediction. URL: https://github.com/hlee110123/OHDSI---Length-of-Stay-Prediction/tree/main [accessed 2024-11-05]

Abbreviations

AUROC: area under the receiver operating characteristic curve
AUPRC: area under the precision-recall curve
ED: emergency department
EHR: electronic health record
ETL: extract, transform, and load
GB: Gradient Boosting
LGB: Light Gradient Boosting
LR: logistic regression
ML: machine learning
MIMIC: Medical Information Mart for Intensive Care
OHDSI: Observational Health Data Sciences and Informatics
OMOP CDM: Observational Medical Outcomes Partnership Common Data Model
RF: random forest
SHAP: Shapley Additive Explanations
SNUH: Seoul National University Hospital
SNUBH: Seoul National University Bundang Hospital
XGB: Extreme Gradient Boosting

Edited by A Schwartz; submitted 07.04.24; peer-reviewed by Y Chen, B Popoff; comments to author 11.06.24; revised version received 07.07.24; accepted 29.10.24; published 22.11.24.

Please cite as:
Lee H, Kim S, Moon HW, Lee HY, Kim K, Jung SY, Yoo S
Hospital Length of Stay Prediction for Planned Admissions Using Observational Medical Outcomes Partnership Common Data Model: Retrospective Study
J Med Internet Res 2024;26:e59260
URL: https://www.jmir.org/2024/1/e59260
doi:10.2196/59260
PMID:

©Haeun Lee, Seok Kim, Hui-Woun Moon, Ho-Young Lee, Kwangsoo Kim, Se Young Jung, Sooyoung Yoo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Antihypertensive Drug Recommendations for Reducing Arterial Stiffness in Patients With Hypertension: Machine Learning–Based Multicohort (RIGIPREV) Study

Iván Cavero-Redondo[1,2], PhD; Arturo Martinez-Rodrigo[3], PhD; Alicia Saz-Lara[1], PhD; Nerea Moreno-Herraiz[1], MSc; Veronica Casado-Vicente[4,5], PhD; Leticia Gomez-Sanchez[6], PhD; Luis Garcia-Ortiz[7,8,9], PhD; Manuel A Gomez-Marcos[7,8,10], PhD; RIGIPREV Investigators Group[1,7,11]

[1]CarVasCare Research Group, Facultad de Enfermería de Cuenca, Universidad de Castilla-La Mancha, Cuenca, Spain

[2]Facultad de Ciencias de la Salud, Universidad Auto´noma de Chile, Talca, Chile

[3]Department of Informatic Systems, University of Castilla - La Mancha, Cuenca, Spain

[4]Parquesol University Health Centre, West Valladolid Primary Healthcare Management, Castilla y León Regional Health Authority, Valladolid, Spain

[5]Department of Medicine, Dermatology and Toxicology, University of Valladolid, Valladolid, Spain

[6]Emergency Service, University Hospital of La Paz, Madrid, Spain

[7]Primary Care Research Unit of Salamanca, Salamanca Primary Healthcare Management, Institute of Biomedical Research of Salamanca, Salamanca, Spain

[8]Research Network on Chronicity, Primary Care and Health Promotion, Salamanca, Spain

[9]Department of Biomedical and Diagnostic Sciences, University of Salamanca, Salamanca, Spain

[10]Department of Medicine, University of Salamanca, Salamanca, Spain

[11]see Acknowledgements

Corresponding Author:
Alicia Saz-Lara, PhD
CarVasCare Research Group
Facultad de Enfermería de Cuenca
Universidad de Castilla-La Mancha
C. Santa Teresa Jornet s/n
Cuenca, 16001
Spain
Phone: 34 969179100
Email: alicia.delsaz@uclm.es

Abstract

Background: High systolic blood pressure is one of the leading global risk factors for mortality, contributing significantly to cardiovascular diseases. Despite advances in treatment, a large proportion of patients with hypertension do not achieve optimal blood pressure control. Arterial stiffness (AS), measured by pulse wave velocity (PWV), is an independent predictor of cardiovascular events and overall mortality. Various antihypertensive drugs exhibit differential effects on PWV, but the extent to which these effects vary depending on individual patient characteristics is not well understood. Given the complexity of selecting the most appropriate antihypertensive medication for reducing PWV, machine learning (ML) techniques offer an opportunity to improve personalized treatment recommendations.

Objective: This study aims to develop an ML model that provides personalized recommendations for antihypertensive medications aimed at reducing PWV. The model considers individual patient characteristics, such as demographic factors, clinical data, and cardiovascular measurements, to identify the most suitable antihypertensive agent for improving AS.

Methods: This study, known as the RIGIPREV study, used data from the EVA, LOD-DIABETES, and EVIDENT studies involving individuals with hypertension with baseline and follow-up measurements. Antihypertensive drugs were grouped into classes such as angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), β-blockers, diuretics, and combinations of diuretics with ACEIs or ARBs. The primary outcomes were carotid-femoral and brachial-ankle PWV, while the secondary outcomes included various cardiovascular, anthropometric, and biochemical parameters. A multioutput regressor

XSL•FO
RenderX

using 6 random forest models was used to predict the impact of each antihypertensive class on PWV reduction. Model performance was evaluated using the coefficient of determination (R^2) and mean squared error.

Results: The random forest models exhibited strong predictive capabilities, with internal validation yielding R^2 values between 0.61 and 0.74, while external validation showed a range of 0.26 to 0.46. The mean squared values ranged from 0.08 to 0.22 for internal validation and from 0.29 to 0.45 for external validation. Variable importance analysis revealed that glycated hemoglobin and weight were the most critical predictors for ACEIs, while carotid-femoral PWV and total cholesterol were key variables for ARBs. The decision tree model achieved an accuracy of 84.02% in identifying the most suitable antihypertensive drug based on individual patient characteristics. Furthermore, the system's recommendations for ARBs matched 55.3% of patients' original prescriptions.

Conclusions: This study demonstrates the utility of ML techniques in providing personalized treatment recommendations for antihypertensive therapy. By accounting for individual patient characteristics, the model improves the selection of drugs that control blood pressure and reduce AS. These findings could significantly aid clinicians in optimizing hypertension management and reducing cardiovascular risk. However, further studies with larger and more diverse populations are necessary to validate these results and extend the model's applicability.

(J Med Internet Res 2024;26:e54357) doi:10.2196/54357

KEYWORDS

antihypertensive; drugs; models; patients; pulse wave velocity; recommendations; hypertension; machine learning; drug recommendations; arterial stiffness; RIGIPREV

Introduction

Background

According to the Global Burden of Diseases, Injuries, and Risk Factors Study, the leading global risk factor for attributable deaths in 2019 was high systolic blood pressure, accounting for 10.8 million deaths (19.2% of the total) [1]. However, despite receiving treatment through both nonpharmacological and pharmacological measures, a significant percentage of patients continue to have poor blood pressure control [2]. Current clinical practice guidelines recommend individualized regimens for hypertension treatment [3,4]. In addition, arterial stiffness (AS), assessed through pulse wave velocity (PWV), is an independent cardiovascular risk factor with the capacity to predict morbidity and mortality from cardiovascular diseases [5,6]. Furthermore, there are numerous pharmacological strategies for hypertension treatment [7], and the effect on AS varies among these groups [8-12]. In a recently published network meta-analysis by our group [8] focusing on the outcomes of the 5 classes of drugs used in clinical practice for hypertension treatment, a significant decrease in PWV was observed with angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), and β-blockers (BBs) [8]. The meta-analysis conducted by Ong et al [12] compared the effects of antihypertensive drugs (except ARBs) with those of placebo and revealed that all antihypertensive drugs reduced the PWV. In short-term trials, ACEIs were more effective than calcium channel blockers (CCBs), and in long-term trials, ACEIs, CCBs, BBs, and diuretics were more effective than placebo. The meta-analysis conducted by Chen et al [11] analyzed the effects of ARBs versus other antihypertensive agents on PWV reduction and found no differences compared to other groups. The meta-analysis conducted by Shahin et al [9] examined the effect of ACEIs on AS compared to that of placebo or other antihypertensive agents and concluded that ACEIs reduce PWV but did not definitively establish whether the effect was superior to that of other antihypertensive agents. Finally, the meta-analysis conducted by Li et al [10] assessed the effect of ACEIs on PWV without finding any differences compared to other antihypertensive drugs but did find differences compared to placebo. Similarly, longitudinal studies that have analyzed the effects of ACEIs or ARBs alone or in combination with CCBs or diuretics have demonstrated their effectiveness in reducing PWV [13]. Therefore, all the results suggest that antihypertensive agents may have beneficial effects on AS, but the intensity of these effects may vary. However, no studies have analyzed whether the effects of antihypertensive drugs on AS differ based on the individual characteristics of patients with hypertension.

PWV is now recognized as a crucial indicator of AS, making it a valuable predictor marker for atherosclerosis [14]. This link between hypertension and cardiovascular risk is supported by a strong body of evidence [15]. This study highlights the importance of incorporating PWV into cardiovascular research and risk assessment. Several epidemiological studies, including the well-known Framingham Heart Study [16] and the Rotherdam study [17], have consistently shown a significant link between high PWV and a greater risk of cardiovascular illness and death [5,18,19]. This relationship is further supported by long-term studies that have linked PWV measurements to the occurrence of negative cardiovascular events, such as coronary artery disease and stroke [20]. Selecting PWV as the primary outcome measure provides a thorough understanding of the effects of antihypertensives on arterial health in line with our goal of reducing cardiovascular risk in patients with hypertension [8].

Machine learning (ML) techniques are ubiquitous in all fields of research, and their use in medicine is increasingly common, serving as tools to facilitate diagnosis, prediction, classification, or personalized recommendations based on individual patient characteristics. Unlike rule-based algorithms that require explicit programming, ML algorithms use self-learning techniques to discover sophisticated patterns in medical data, with the potential

XSL•FO
RenderX

to enhance personalized care [21]. Therefore, predictive analyses using ML methods are highly valued and practiced in real health care applications, with significant potential to develop recommendations that assist clinicians in making the most suitable decisions for each patient, as the complexity of medicine now exceeds the capacity of the human mind [21,22]. ML techniques have been widely used to predict cardiovascular diseases and their cardiovascular risk factors [23,24]. Numerous studies concerning hypertension have been conducted to predict which antihypertensive drug may be most beneficial for each patient with hypertension [25], analyze the profiles of different antihypertensive drugs [26], build a prediction model for antihypertensive drugs to be used in older patients with hypertension [27], analyze variables influencing blood pressure reduction [28], study variables predicting the risk of hypokalemia in patients with hypertension [29], or predict the onset of pregnancy-induced hypertension [30]. There are many ML studies investigating different personalized recommendations for patients, particularly reinforcement learning algorithms (ie, optimal dosing of medication [31,32] and optimal timing of intervention [33]). However, it is essential to consider the conclusion of the review conducted by the Journal of the American Heart Association in 2023 [22] regarding ML research in patients with hypertension, emphasizing that these ML techniques currently have significant deficiencies in terms of reporting quality, model validation, and bias in the algorithms used.

Objectives

The hypothesis of this study was that the effects of different antihypertensive drugs on PWV vary according to the individual characteristics of each person with hypertension. To the best of our knowledge, no study has analyzed which antihypertensive drug is most suitable for real clinical practice using ML techniques to reduce PWV. Therefore, we propose this study, with the main objective of developing a model to provide recommendations on the most suitable antihypertensive agent for reducing PWV, based on the individual characteristics of people with hypertension, using advanced ML techniques consisting of multiple random forest (RF) models within a multioutput regressor approach.

Methods

Design and Sample Characteristics

The Effectiveness of Antihypertensive Drugs on Arterial Stiffness in Hypertensive Adults Study (the RIGIPREV study) [34] represents a real-world longitudinal investigation involving participants with hypertension. This study drew upon data originating from the Influence of Different Risk Factors in Vascular Accelerated Aging study (the EVA study) [35] (NCT02623894), the Relationship of Central Blood Pressure and Pulse Wave Velocity With Target Organ Damage study (the LOD-DIABETES study) [36] (NCT01065155), and the Effectiveness of the Use of a Mobile Tool in Improving Lifestyle (the EVIDENT study) [37] (NCT02016014). Specifically, participants were selected from these 3 research endeavors if they were using antihypertensive drugs and possessed both baseline and follow-up measurements of variables essential for our primary analysis.

Ethical Considerations

The research protocols for the studies included in this pooled analysis were approved by the Drug Research Ethics Committee of Salamanca, with the following registration numbers: PI15/01039 and PI20/10569 (EVA study [35]), PI15/11/2015 (LOD-DIABETES study [36]), and PI83/06/2018 (EVIDENT study [37]). All the participants who participated in these studies provided written informed consent. During the development of the study, the principles of the Declaration of Helsinki [38] and the World Health Organization standards for observational studies were followed. The confidentiality of the participants included was always guaranteed in accordance with the provisions of Organic Law 3/2018, of 5 December, on Personal Data Protection and Guarantee of Digital Rights and Regulation (European Union) 2016/679 of the European Parliament and of the Council of 27 April 2016 , on Data Protection.

Intervention

The antihypertensive drugs used in the 3 studies were categorized into the following groups: BBs, diuretics, ACEIs, ARBs, and combinations of diuretics+ACEIs and diuretics+ARBs. While other antihypertensive agents from various groups (such as CCBs, α-adrenergic receptor antagonists, centrally acting agents, or direct-acting vasodilators) were used in these studies, they were excluded from our analysis due to their limited representation, each having samples of <10 participants (Textbox 1).

XSL•FO
RenderX

Textbox 1. Antihypertensive groups included in the study.

Antihypertensive groups

- β-blockers include acebutolol, atenolol, atenolol, betaxolol, bisoprolol, carteolol, esmolol, metoprolol, nadolol, oxprenolol, penbutolol, propranolol, timolol, celiprolol, carvedilol, labetalol, nebivolol, and pindolol.
- Diuretics include furosemide, bumetanide, torsemide, bendroflumethiazide, chlorothiazide, chlorthalidone, hydrochlorothiazide, indapamide, polythiazide, trichlormethiazide, amiloride, eplerenone, spironolactone, and triamterene.
- Angiotensin-converting enzyme inhibitors (ACEIs) include benazepril, captopril, cilazapril, enalapril, fosinopril, imidapril, lisinopril, moexipril, perindopril, quinapril, ramipril, trandolapril, and zofenopril.
- Angiotensin receptor blockers (ARBs) include candesartan, eprosartan, irbesartan, losartan, olmesartan, telmisartan, and valsartan.
- Combinations of diuretic+ACEI and diuretic+ARB: these categories involve combinations of medications from the diuretic, ACEI, and ARB classes.

Outcome

Our primary outcome measure centered on the change in carotid-femoral PWV (cfPWV) and brachial-ankle PWV (baPWV). These measurements were consistently collected across the 3 included studies, following their respective study protocols outlined elsewhere [35-37]. To facilitate comparison, we standardized these 2 types of PWV using normalized values.

Specifically, the following primary outcome measures were measured: the cfPWW was estimated using a SphygmoCor device (AtCor Medical Pty Ltd). By analyzing the carotid and femoral artery pulse waves in the supine position, the time delay was estimated and compared to that of the electrocardiogram R wave, and the cfPWV was calculated. Distances were measured with a tape measure from the sternal notch to the point where the sensor was placed in the carotid and femoral arteries [39]. The baPWV was estimated using a VaSera VS-1500 device (Fukuda Denshi Co Ltd) according to the manufacturer's instructions. The baPWV was estimated using the following equation: baPWV = ({0.5934×height (cm)+14.4724})/tba, where tba is the time interval between the arm and ankle waves [39].

Covariates

The variables collected and tests performed have been previously published in the protocol of the EVA study [35], the LOD-DIABETES study [36], and the EVIDENT II study [37]. The professionals who performed the tests and questionnaires followed a standardized protocol.

In this study, we meticulously considered an extensive array of covariates, each of which played a crucial role in shaping our analysis. These covariates included demographic factors such as age and sex; lifestyle characteristics such as smoking habits and alcohol consumption; adherence to the Mediterranean diet assessed with the Mediterranean Diet Adherence Screener questionnaire [40]; and medication use, which included antidiabetics, antiaggregants, anticlotting agents, and nonsteroidal anti-inflammatory drugs.

Furthermore, we incorporated a comprehensive set of anthropometric measurements, including height, weight, and waist circumference, along with vital cardiovascular parameters such as systolic and diastolic blood pressure and heart rate, and the measurements were performed according to the recommendations of the European Society of Hypertension [41]. Furthermore, we included parameters such as the ankle-brachial index, intima-media thickness, carotid-artery vascular index (CAVI), and augmentation index at 75 mm Hg (AIx75), ensuring that our statistical models were comprehensive and robust, thereby enhancing the validity of our findings. In addition, the Cornell index and left ventricular hypertrophy were measured.

We also included numerous biochemical variables. The participants in the 3 studies underwent a venous blood draw at the Salamanca Primary Care Research Unit between 8 AM and 9 AM on an empty stomach without smoking or consuming alcohol or caffeinated beverages during the previous 12 hours, and all samples were analyzed in the same laboratory. Specifically, we performed the following analytical determinations (Textbox 2).

XSL•FO
RenderX

Textbox 2. Analytical determinations performed in the study.

Analytical determinations

- Hematological variables, including red blood cells, hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular hemoglobin content, and mean corpuscular volume, were collected. Variables related to the distribution of erythrocyte sizes, such as the anisocytosis index, were also considered. Moreover, leukocyte subpopulations, including neutrophils, lymphocytes, monocytes, eosinophils, and basophils, red blood cell distribution width, along with platelet characteristics such as the mean platelet volume, were included in our covariate set.
- Regarding diabetes-related covariates, we incorporated fasting plasma glucose, glycated hemoglobin, and insulin levels as essential factors.
- In addition, lipid profiles, comprising total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, and triglycerides, were considered.
- Renal function markers such as uric acid, serum creatinine, and creatinine, as well as microalbuminuria, reflect kidney health. In addition, the albumin/creatinine ratio and estimated glomerular filtration rate according to the Chronic Kidney Disease Epidemiology Collaboration criteria were included.
- Liver function was assessed through markers, including aspartate aminotransferase and glutamic oxaloacetic transaminase, alanine aminotransferase and serum glutamic pyruvic transaminase, and gamma-glutamyl transferase.
- Inflammation markers, such as fibrinogen and C-reactive protein, were integrated into our covariate framework. Thyroid function was evaluated with the inclusion of thyroid-stimulating hormone and thyroxine levels, while vitamin D levels provided insight into nutritional status.

Data Standardization

The EVA [35], LOD-DIABETES [36], and EVIDENT [37] studies each followed unique protocols and meticulously selected participants based on predetermined criteria.

To achieve data standardization, our main priority was to ensure consistent measurement of key factors, particularly baseline and follow-up PWV assessments, which were integral to our primary analytical framework. It should be acknowledged that there was no universally agreed upon protocol for drug choice among the 3 studies, but for standardization, we categorized antihypertensive drugs into specific classes, including BB, diuretics, ACEI, ARB, and diuretic+ACEI and diuretic+ARB combinations. Due to their minimal representation within each group, certain medications were excluded from consideration. This was done to ensure substantial sample sizes for meaningful comparisons between groups, a crucial factor in our decision-making process. In addition, patients who consistently stayed on 1 drug throughout the follow-up period as well as those who may have switched between different drugs were included in the analysis.

This study investigated a wide range of antihypertensive drugs that were purposefully selected to enhance the scope and diversity of our research. We acknowledged the diversity and potential differences in drug options and how they are administered in various studies and thus used thorough statistical corrections. Our analysis considered factors such as age, lifestyle behaviors, and medication use, which were carefully incorporated to mitigate any potential biases stemming from these differences.

Statistical Analysis

Normal probability plots and the Kolmogorov-Smirnov test were used to verify the normality of the distribution of continuous variables. Descriptive data for the total sample and antihypertensive drugs are shown as the means and SDs or proportions (%), as appropriate. ANOVA for continuous variables and chi-square analysis for categorical variables were used to compare baseline variables according to antihypertensive drugs.

Data Preprocessing and Outcome Structuring

Once the patient data were organized, we addressed missing values within the feature set by imputing them using the median of the respective variables. For the target variables, we created 6 distinct output vectors, each corresponding to 1 of the 6 antihypertensive drugs under study: ACEIs, ARBs, BBs, diuretic, diuretic+ACEI, and diuretic+ARB. These output vectors capture the differences between the original PWV measurements and the PWV values observed during at least 1 year of follow-up for each patient.

Training and Assessment of Embedded RF ML Models for Multitarget Prediction

To predict medication effectiveness in decreasing PWV, we used a multioutput regressor framework consisting of 6 individual RF models, each corresponding to a different medication [42]. These RF models were independently trained on the same feature set but aimed to predict different target variables, specifically the effectiveness of the respective antihypertensive drugs in decreasing PWV.

To construct the predictive models, we characterized each patient using a feature input vector X, comprising medical attributes such as cholesterol levels, systolic blood pressure, and Mediterranean diet intake, among others. Nevertheless, given the high number of input variables in our study, various procedures for variable reduction were implemented. Initially, the average decrease in impurity for each model was calculated because of the inherent ability of RF algorithms to assess the importance of input features. This analysis was based on evaluating how each covariate contributes to reducing variance in the predictions within the trees. A feature is considered more important if its inclusion in tree splits leads to a greater reduction in variance, indicating that the predictions become more homogeneous and precise due to that feature. In addition, a permutation-based analysis was also conducted to address the potential high cardinality of the input variables, yielding

XSL•FO
RenderX

consistent results without unfairly favoring features with a high number of unique values.

Both variable reduction methodologies yielded similar results, leading to the identification of the top 10 most important features for each model and aiming to make the final recommendation system lighter and more interpretable. Thus, each RF model was finally trained using these 10 most important features exclusively in each case. Consequently, the global features used by the system were reduced from 64 to 38 variables, representing a reduction of >40%.

Initially, we combined the 3 studies and opted for an internal validation scheme. Thus, a holdout validation strategy was implemented, randomly distributing data into 2 folds for training and testing, with 80% of patients retained for training [43]. However, we also proceeded with an external validation phase to further test the model's robustness and generalizability across different datasets. However, given the limited sample size across the 3 databases, the synthetic minority oversampling technique (SMOTE) was used. The SMOTE algorithm works by synthetically generating new samples by selecting instances that are close to the feature space, drawing a line between the examples in the feature space, and creating a new instance at a point along the line [44].

Thus, considering the 3 studies that are part of this work, the samples from the EVIDENT II study were reserved for external validation because they contained the greatest proportion of samples in almost all the subgroups. Meanwhile, the remaining 2 databases were combined and synthetically augmented using the SMOTE algorithm to use the samples for training. These samples were augmented to approximately 80%-20% for training and testing, respectively.

The predictive power of the models was assessed on the validation set in both strategies using the coefficient of determination (R^2) and the mean squared error (MSE) [45]. R^2 quantifies the proportion of the variance in the dependent variable, PWV, predicted from the independent variables, with values closer to 1 indicating a better fit. MSE quantifies the alignment between predictions and actual outcomes.

Design of the Recommendation System

The core function of the recommendation system is to determine the most effective treatment for individual patients. For each patient's feature vector, the system runs all 6 RF models to predict the corresponding change in PWV. The system identifies the medication predicted to yield the most significant reduction in PWV, offering data-driven personalized treatment optimization.

This recommendation system provides a personalized approach to treatment optimization, allowing patients to benefit from the medication predicted to be most effective for them.

To provide additional insights into the recommendation system, a depth-level 10 decision tree was constructed. Tree growth was regulated using the Gini impurity index as the criterion for node splitting, stopping when the Gini index reached a value ≤20%. This decision tree serves as a valuable tool for understanding the most influential variables in the recommendation process and offers a transparent and interpretable framework.

Results

Characteristics of the Study Participants

The RIGIPREV study sample included a total of 194 participants with hypertension, of whom 96 (49.5%) were women. The mean age of the participants was 61.8 (SD 10.0) years. Overall, there were no differences in the baseline values of the participants included in each of the treatments except for mean corpuscular hemoglobin concentration (P=.02), mean corpuscular volume (MCV; P=.02), glycated hemoglobin (HbA_{1c}; P=.045), and thyroid-stimulating hormone (TSH; P=.02). Table 1 shows the baseline characteristics of the enrolled population.

XSL•FO
RenderX

Table 1. Baseline characteristics of patients with hypertension from the RIGIPREV study, a multicohort analysis conducted in Spain. This table summarizes the demographic, clinical, and biochemical characteristics of participants, including age, sex, BMI, comorbidities, and medication use. The study includes individuals with hypertension recruited from the EVA, LOD-DIABETES, and EVIDENT studies, focusing on their cardiovascular health and arterial stiffness as measured by pulse wave velocity.

Characteristic	Total (N=194)	ACEI[a] (n=44)	ARB[b] (n=47)	BB[c] (n=17)	Diuretic (n=14)	Diuretic+ACEI (n=28)	Diuretic+ARB (n=44)	*P* value
Study data, n (%)								
EVA study	56 (28.9)	8 (18.2)	15 (31.9)	6 (35.3)	5 (35.7)	9 (32.1)	13 (29.5)	—[d]
LOD-DIABETES study	57 (29.4)	14 (31.8)	14 (29.8)	3 (17.6)	2 (14.3)	12 (42.9)	12 (27.3)	—
EVIDENT II study	81 (41.8)	22 (50)	18 (38.3)	8 (47.1)	7 (50)	7 (25)	19 (43.2)	—
Demographic factors								
Age (y), mean (SD)	61.8 (10.0)	59.6 (10.0)	62.1 (8.9)	61.6 (9.5)	63.4 (10.6)	60.3 (11.9)	64.1 (9.9)	.37
Gender, n (%)								.19
Women	96 (49.5)	17 (38.6)	20 (42.6)	9 (52.9)	10 (71.4)	14 (50)	26 (59.1)	
Men	98 (50.5)	27 (61.4)	27 (57.4)	8 (47.1)	4 (28.6)	14 (50)	18 (40.9)	
Lifestyle characteristics								
Smoking, n (%)								.006
Yes	26 (13.4)	10 (22.7)	7 (14.9)	1 (5.9)	0 (0)	1 (3.6)	7 (15.9)	
No	97 (50.0)	12 (27.3)	25 (53.2)	8 (47.1)	12 (85.7)	20 (71.4)	20 (45.5)	
Former smoker	71 (36.6)	22 (50)	15 (31.9)	8 (47.1)	2 (14.3)	7 (25)	17 (38.6)	
Cigarette consumption (n), mean (SD)	5.75 (10.8)	8.3 (11.4)	6.7 (12.9)	5.2 (7.8)	0.7 (2.6)	2.8 (8.5)	5.7 (11.2)	.17
Alcohol consumption (g/wk), mean (SD)	55.5 (111.2)	47.6 (100.4)	58.3 (98.3)	38.2 (76.6)	19.2 (40.4)	51.2 (79.0)	81.4 (164.9)	.47
Mediterranean diet adherence (MEDAS[e] units), mean (SD)	7.4 (2.0)	7.2 (2.1)	8.0 (1.8)	7.8 (2.6)	6.6 (1.7)	6.5 (2.2)	7.5 (2.0)	.33
Medication use, n (%)								
Antidiabetics								.19
Yes	54 (27.8)	13 (29.5)	9 (19.1)	2 (11.8)	6 (42.9)	11 (39.3)	13 (29.5)	
No	140 (72.2)	31 (70.5)	38 (80.9)	15 (88.2)	8 (57.1)	17 (60.7)	31 (70.5)	
Lipid-lowering agents								.1
Yes	92 (47.4)	20 (45.5)	20 (42.6)	4 (23.5)	9 (64.3)	18 (64.3)	21 (47.7)	
No	102 (52.6)	24 (54.5)	27 (57.4)	13 (76.5)	5 (35.7)	10 (35.7)	23 (52.3)	
Antiaggregants								.51
Yes	43 (22.2)	10 (22.7)	15 (31.9)	3 (17.6)	2 (14.3)	6 (21.4)	7 (15.9)	
No	151 (77.8)	34 (77.3)	32 (68.1)	14 (82.4)	12 (85.7)	22 (78.6)	37 (84.1)	
Anticoagulants								.39
Yes	4 (2.1)	1 (2.3)	1 (2.1)	0 (0)	0 (0)	2 (7.1)	0 (0)	
No	190 (97.9)	44 (100)	47 (100)	17 (100)	14 (100)	26 (92.9)	44 (100)	
NSAIDs[f]								.64
Yes	1 (0.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.3)	
No	193 (99.5)	44 (100)	47 (100)	17 (100)	14 (100)	28 (100)	43 (97.8)	
Anthropometric measurements, mean (SD)								
Height (cm)	162.5 (10.3)	164.3 (8.7)	163.5 (10.5)	159.4 (8.3)	158.5 (13.2)	163.5 (11.2)	161.4 (10.7)	.31

XSL•FO
RenderX

Characteristic	Total (N=194)	ACEI[a] (n=44)	ARB[b] (n=47)	BB[c] (n=17)	Diuretic (n=14)	Diuretic+ACEI (n=28)	Diuretic+ARB (n=44)	*P* value
Weight (kg)	80.5 (15.3)	81.6 (13.4)	81.6 (17.0)	76.8 (12.9)	76.3 (16.1)	83.0 (17.2)	79.3 (14.6)	.63
Waist circumference (cm)	102.6 (10.4)	103.0 (8.5)	103.0 (12.1)	102.0 (10.8)	99.2 (10.1)	104.8 (11.9)	101.6 (9.2)	.64
Cardiovascular parameters, mean (SD)								
SBP[g] (mm Hg)	131.2 (16.4)	131.5 (13.9)	131.0 (16.4)	128.1 (18.7)	120.2 (11.9)	133.9 (18.1)	134.1 (17.1)	.11
DBP[h] (mm Hg)	81.0 (11.4)	79.3 (9.8)	81.9 (9.0)	78.0 (13.0)	76.6 (7.5)	83.5 (12.9)	82.9 (14.0)	.21
Heart rate (beats/min)	69.7 (11.6)	69.8 (12.1)	68.6 (12.2)	63.5 (9.9)	74.2 (11.4)	70.8 (10.9)	71.2 (11.2)	.14
ABI[i] (units)	1.1 (0.1)	1.0 (0.1)	1.0 (0.1)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)	.35
cfPWV[j] (m/s)	9.5 (2.6)	8.9 (2.3)	9.5 (2.5)	9.4 (2.5)	9.9 (1.9)	9.9 (3.0)	9.8 (3.0)	.67
baPWV[k] (m/s)	14.2 (2.3)	13.8 (2.3)	14.0 (2.4)	13.8 (2.0)	14.9 (2.4)	14.6 (2.3)	14.3 (2.3)	.53
CAVI[l] (m/s)	8.3 (1.3)	8.0 (1.3)	8.3 (1.4)	8.0 (1.0)	8.6 (1.3)	8.6 (1.3)	8.4 (1.3)	.37
AIx75[m] (%)	30.8 (10.8)	32.0 (12.5)	29.8 (10.2)	25.6 (7.9)	30.2 (13.2)	29.6 (10.4)	33.5 (9.8)	.21
IMT[n] (mm)	0.7 (0.1)	0.8 (0.1)	0.7 (0.1)	0.7 (0.1)	0.6 (0.1)	0.7 (0.1)	0.7 (0.1)	.07
IMT plaque, n (%)								.34
Yes	41 (21.1)	11 (25)	11 (23.4)	3 (17.6)	2 (14.3)	3 (10.7)	11 (25)	
No	107 (55.2)	19 (43.2)	22 (46.8)	8 (47.1)	9 (64.3)	23 (82.1)	26 (59.1)	
Cornell index (mm), mean (SD)	15.1 (5.6)	15.2 (4.8)	15.3 (5.6)	14.2 (5.6)	14.1 (3.6)	15.1 (7.5)	15.3 (5.3)	.98
LVH Cornell[o] (ms*mV), mean (SD)	1600.0 (610.6)	1499.8 (493.5)	1516.3 (610.2)	1643.9 (506.0)	1477.2 (378.0)	1662.8 (711.9)	1735.2 (700.7)	.55
Biochemical variables, mean (SD)								
Red blood cells (million/mm^3)	4.8 (0.4)	4.8 (0.4)	4.9 (0.4)	4.8 (0.3)	4.9 (0.6)	4.9 (0.3)	4.8 (0.4)	.87
Hemoglobin A (g/dL)	14.5 (1.3)	14.7 (1.3)	14.9 (1.2)	14.5 (1.0)	14.2 (1.1)	14.4 (1.5)	14.4 (1.3)	.39
Hematocrit (%)	43.3 (3.6)	43.9 (3.8)	44.1 (4.0)	43.4 (3.2)	42.6 (4.0)	42.9 (3.7)	42.4 (3.0)	.398
MCHC[p] (g/dL)	29.7 (2.2)	30.7 (1.5)	30.2 (1.5)	30.8 (1.3)	28.1 (2.9)	29.4 (2.5)	29.1 (2.4)	*.02*
CHbCM[q] (pg)	33.3 (0.9)	33.4 (0.9)	33.3 (0.8)	33.7 (0.5)	32.9 (1.3)	33.1 (1.1)	33.3 (0.8)	.61
MCV[r] (fl)	89.2 (5.7)	91.7 (4.1)	90.8 (3.3)	91.6 (3.7)	85.0 (8.0)	88.9 (6.5)	87.3 (6.2)	*.02*
Anisocytosis index (%)	12.8 (1.2)	12.8 (1.3)	12.2 (0.4)	12.6 (0.7)	12.8 (0.6)	12.7 (1.0)	13.4 (1.7)	.07
Leukocytes (million/mm^3)	6.8 (1.9)	7.1 (2.4)	7.0 (1.9)	7.0 (1.6)	7.1 (2.2)	7.1 (1.7)	6.1 (1.5)	.28
Neutrophils (million/mm^3)	3.8 (1.3)	4.0 (1.8)	3.9 (1.1)	3.9 (1.2)	3.5 (1.2)	4.1 (1.2)	3.3 (0.9)	.23
Lymphocytes (million/mm^3)	2.2 (0.7)	2.2 (0.6)	2.3 (1.1)	2.2 (0.6)	2.4 (0.7)	2.2 (0.6)	2.0 (0.5)	.63
Monocytes (million/mm^3)	0.5 (0.1)	0.5 (0.2)	0.5 (0.1)	0.5 (0.1)	0.4 (0.1)	0.5 (0.1)	0.5 (0.1)	.79
Eosinophils (million/mm^3)	0.2 (0.1)	0.1 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.3)	0.1 (0.1)	0.1 (0.1)	.27
Basophils (million/mm^3)	0.1 (0.1)	0.1 (0.1)	0.1 (0.0)	0.0 (0.0)	0.1 (0.2)	0.0 (0.0)	0.0 (0.0)	.59
Platelets (10^9/L)	231.4 (59.5)	229.0 (66.8)	230.5 (66.5)	239.2 (60.8)	247.3 (47.3)	231.0 (60.0)	225.4 (50.4)	.92
RDW[s] (%)	12.7 (0.9)	12.6 (0.8)	12.4 (0.3)	12.6 (0.5)	12.7 (0.6)	12.6 (0.7)	13.0 (1.4)	.89

XSL•FO
RenderX

Characteristic	Total (N=194)	ACEI[a] (n=44)	ARB[b] (n=47)	BB[c] (n=17)	Diuretic (n=14)	Diuretic+ACEI (n=28)	Diuretic+ARB (n=44)	*P* value
MPV[t] (fl)	9.4 (1.8)	10.2 (2.0)	9.0 (1.7)	9.3 (1.4)	9.5 (1.9)	8.9 (1.9)	9.6 (1.9)	.48
FPG[u] (mg/100 mL)	101.7 (27.2)	102.1 (26.4)	98.0 (20.9)	90.8 (9.8)	103.2 (24.6)	113.6 (38.8)	101.3 (29.0)	.11
HbA_{1c}[v] (%)	5.9 (0.8)	5.8 (0.6)	5.8 (0.7)	5.5 (0.4)	6.2 (1.0)	6.2 (1.2)	5.9 (0.8)	*.045*
Insulin (mIU/L)	12.4 (7.6)	11.8 (6.0)	14.2 (10.5)	11.5 (5.6)	12.5 (5.4)	12.7 (6.9)	11.1 (6.6)	.49
Total cholesterol (mg/dL)	191.3 (32.5)	190.9 (37.2)	194.7 (28.7)	191.5 (34.5)	184.2 (23.4)	189.7 (31.9)	191.2 (34.7)	.94
cLDL[w] (mg/dL)	95.9 (38.1)	89.6 (42.5)	95.5 (41.1)	97.1 (45.4)	83.3 (28.5)	104.1 (30.3)	101.2 (33.8)	.44
cHDL[x] (mg/dL)	69.5 (34.3)	75.8 (42.3)	70.2 (34.7)	69.6 (33.4)	76.7 (38.1)	56.8 (21.4)	68.1 (29.9)	.33
Triglycerides (mg/dL)	131.0 (61.5)	135.5 (67.1)	134.6 (54.4)	123.9 (49.0)	144.5 (69.8)	145.8 (72.6)	111.7 (54.6)	.2
Uric acid (mg/dL)	5.7 (1.4)	5.7 (1.4)	5.8 (1.3)	5.6 (1.6)	5.8 (0.8)	5.8 (1.5)	5.8 (1.4)	.99
Serum creatinine (mg/dL)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.2)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	.97
Creatinine (µg/min)	117.0 (66.6)	122.1 (66.8)	119.8 (65.7)	120.0 (67.1)	102.0 (63.8)	105.5 (68.7)	119.6 (68.9)	.86
Microalbuminuria (mg/day)	19.1 (48.7)	20.2 (52.5)	21.7 (37.7)	8.7 (10.2)	6.9 (8.2)	11.8 (18.8)	27.6 (76.6)	.596
Albumin/creatinine ratio	17.3 (39.0)	13.0 (28.9)	19.8 (35.4)	6.9 (7.9)	13.0 (23.0)	12.1 (25.1)	27.4 (62.2)	.55
eGFR[y] (mL/min/1.73 m^2)	85.2 (14.1)	89.2 (10.8)	85.0 (13.4)	85.8 (16.1)	81.5 (17.4)	86.0 (17.2)	82.1 (13.6)	.25
AST/GOT[z] (units/L)	23.6 (12.8)	20.4 (6.1)	21.2 (5.9)	20.4 (4.2)	24.5 (7.7)	26.7 (11.0)	28.0 (22.6)	.06
ALT/SGPT[aa] (units/L)	26.6 (18.5)	25.5 (14.3)	24.2 (9.7)	20.7 (8.0)	23.8 (11.3)	33.3 (25.1)	29.2 (26.6)	.18
GGT[ab] (units/L)	32.9 (37.4)	29.5 (24.2)	30.2 (25.0)	37.3 (40.8)	22.4 (7.9)	42.8 (46.7)	34.5 (53.7)	.56
Fibrinogen (mg/dL)	344.0 (78.3)	345.4 (65.7)	348.1 (94.9)	359.3 (108.4)	337.9 (97.8)	350.1 (62.7)	330.2 (59.5)	.82
CRP[ac] (mg/dL)	0.3 (0.4)	0.3 (0.3)	0.3 (0.5)	0.4 (0.8)	0.3 (0.3)	0.4 (0.3)	0.3 (0.4)	.95
TSH[ad] (mU/L)	2.4 (1.3)	2.0 (1.1)	2.2 (0.9)	1.8 (1.3)	3.1 (1.2)	2.6 (1.7)	2.9 (1.6)	*.02*
T4[ae] hormone (µg/dL)	1.1 (0.1)	1.1 (0.2)	1.1 (0.1)	1.1 (0.2)	1.1 (0.2)	1.2 (0.1)	1.1 (0.1)	.88

XSL•FO
RenderX

Characteristic	Total (N=194)	ACEI[a] (n=44)	ARB[b] (n=47)	BB[c] (n=17)	Diuretic (n=14)	Diuretic+ACEI (n=28)	Diuretic+ARB (n=44)	*P* value
Vitamin D (μg/dL)	23.4 (9.7)	21.5 (9.4)	25.1 (8.0)	27.1 (11.6)	26.1 (10.7)	24.9 (10.3)	18.6 (10.0)	.46

[a]ACEI: angiotensin-converting enzyme inhibitor.
[b]ARB: angiotensin receptor blocker.
[c]BB: β-blocker.
[d]Not available.
[e]MEDAS: Mediterranean Diet Adherence Screener.
[f]NSAID: nonsteroidal anti-inflammatory drug.
[g]SBP: systolic blood pressure.
[h]DBP: diastolic blood pressure.
[i]ABI: ankle brachial index.
[j]cfPWV: carotid-femoral pulse wave velocity.
[k]baPWV: brachial-ankle pulse wave velocity.
[l]CAVI: cardio-ankle vascular index.
[m]AIx75: augmentation index at 75 mm Hg.
[n]IMT: intima-media thickness.
[o]LVH Cornell: left ventricular hypertrophy diagnosed by Cornell.
[p]MCHC: mean corpuscular hemoglobin concentration.
[q]CHbCM: mean corpuscular hemoglobin content.
[r]MCV: mean corpuscular volume.
[s]RDW: red blood cell distribution width.
[t]MPV: mean platelet volume.
[u]FPG: fasting plasma glucose.
[v]HbA_{1c}: glycated hemoglobin.
[w]cLDL: low-density lipoprotein cholesterol.
[x]cHDL: high-density lipoprotein cholesterol.
[y]eGFR: estimated glomerular filtration rate.
[z]AST/SGOT: aspartate aminotransferase/glutamic oxaloacetic transaminase.
[aa]ALT/SGPT: alanine aminotransferase/serum glutamic pyruvic transaminase.
[ab]GGT: gamma-glutamyl transferase.
[ac]CRP: C-reactive protein.
[ad]TSH: thyroid-stimulating hormone.
[ae]T4: thyroxin.

Model Performance and Importance of Characteristics

Table 2 presents a summary of the final features used by each model in descending order of importance. The complexity of the factors influencing the effectiveness of treatments and the importance of considering a broad spectrum of clinical variables and biomarkers are shown in this table.

Specifically, the use of ACEIs was associated with the ability of weight and HbA_{1c} levels to predict the reduction in PWV. ARB, in contrast, emphasized cfPWV and low-density lipoprotein cholesterol levels as the most significant factors. BB indicated that TSH and diastolic blood pressure were the primary contributors. Diuretics exhibited high relevance for CAVI and waist circumference. In the case of diuretic+ACEI, AIx75 and uric acid were crucial factors, while diuretic+ARB were key factors for basophils and mean platelet volume. In summary, these findings underscore the intricate nature of the underlying mechanisms influencing the efficacy of various antihypertensive drugs in reducing PWV.

Furthermore, Table 3 shows the performance of the 6 predictive models used for each antihypertensive drug for both internal and external validation. The R^2 values ranged from 0.61 to 0.74 in the internal validation, with the highest R^2 value for the diuretic+ACEI group (R^2=0.74) and the lowest for the ACEI drug group (R^2=0.69). The MSE values for all drug models ranged between 0.08 for ACEIs and 0.22 for BBs. However, the model's performance in external validation significantly decreased compared to internal validation. Thus, the R^2 values ranged from 0.26 to 0.46, and the MSE values varied from 0.29 to 0.45.

XSL•FO
RenderX

Table 2. Most significant features contributing to the regression models evaluating the efficacy of each antihypertensive medication in the RIGIPREV study. This table highlights the key features that provide the most information for each regression model, which represents the efficacy of different antihypertensive drugs, including angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), β-blockers (BB), diuretics (DIU), and combinations.

Importance	ACEI	ARB	BB	DIU	DIU+ACEI	DIU+ARB
1	Weight	cfPWV[a]	TSH[b]	CAVI[c]	AIx75[d]	Basophils
2	HbA_{1c}[e]	cLDL[f]	DBP[g]	Waist circumference	Uric acid	MPV[h]
3	ALT/SGPT[i]	AST/GOT[j]	cfPWV	Vitamin D	ABI[k]	MCV[l]
4	Creatinine	CHbCM[m]	HbA_{1c}	cHDL[n]	LVH Cornell[o]	LVH Cornell
5	baPWV[p]	Alx75	Albumin and creatinine	cLDL	HbA_{1c}	DBP
6	Hemoglobin	Uric acid	cHDL	MPV	Triglycerides	Hematocrit
7	LVH Cornell	FPG[q]	Uric acid	Creatinine	cfPWV	TSH
8	cHDL	Total cholesterol	Triglycerides	RDW[r]	Height	cfPWV
9	Uric acid	GGT[s]	ABI	IMT[t] plaque	GGT	Monocytes
10	Red blood cell distribution width	eGFR[u]	eGFR	IMT	baPWV	IMT

[a]cfPWV: carotid-femoral pulse wave velocity.
[b]TSH: thyroid-stimulating hormone.
[c]CAVI: cardio-ankle vascular index.
[d]AIx75: augmentation index at 75 mm Hg.
[e]HbA_{1c}: glycated hemoglobin.
[f]cLDL: low-density lipoprotein cholesterol.
[g]DBP: diastolic blood pressure.
[h]MPV: mean platelet volume.
[i]ALT/SGPT: alanine aminotransferase/serum glutamic pyruvic transaminase.
[j]AST/GOT: aspartate aminotransferase/glutamic oxaloacetic transaminase.
[k]ABI: ankle-brachial index.
[l]MCV: mean corpuscular volume.
[m]CHbCM: mean corpuscular hemoglobin content.
[n]cHDL: high-density lipoprotein cholesterol.
[o]LVH Cornell: left ventricular hypertrophy diagnosed by Cornell.
[p]baPWV: brachial-ankle pulse wave velocity.
[q]FPG: fasting plasma glucose.
[r]RDW: red blood cell distribution width.
[s]GGT: gamma-glutamyl transferase.
[t]IMT: intima-media thickness.
[u]eGFR: estimated glomerular filtration rate.

XSL•FO
RenderX

Table 3. Performance evaluation of predictive models by medication using R2 and mean squared error (MSE) with internal and external validation. This table summarizes the performance of machine learning models predicting the impact of antihypertensive drugs on pulse wave velocity in patients with hypertension. Results are based on R^2 and mean square error for internal and external validation using data from the RIGIPREV study cohorts in Spain.

Drug	Internal validation		External validation	
	R^2	MSE	R^2	MSE
ACEI[a]	0.69	0.08	0.26	0.31
ARB[b]	0.72	0.09	0.42	0.38
BB[c]	0.61	0.22	0.31	0.45
Diuretic	0.71	0.12	0.40	0.29
Diuretic+ACEI	0.74	0.15	0.46	0.33
Diuretic+ARB	0.70	0.15	0.29	0.32

[a]ACEI: angiotensin-converting enzyme inhibitor.

[b]ARB: angiotensin receptor blocker.

[c]BB: β-blocker.

Analysis of Medication Recommendations

Figure 1 shows the distribution of patient recommendations by medication. There was substantial variability in medication recommendations compared to the original antihypertensive drugs taken by patients. ARB medication was the most frequently recommended, appearing in 26.8% (52/194) of the patients, followed by ACEI, which was recommended in 15.5% (30/194) of the patients. Notably, of the patients who were originally on ACEIs, the same recommendation (match rate) was made for only 36% (16/44) of the patients. For patients who were originally on ARBs, a substantially greater match rate of 55% (26/47) was observed. The combination antihypertensive drugs such as diuretics+ARB and diuretics+ACEI, 52% (23/44) and 43% (12/28), respectively, maintained the same combination medication regimens.

XSL•FO
RenderX

Figure 1. Comparison of original versus predicted pulse wave velocity (PWV) in patients with hypertension. This figure shows the comparison between the original PWV values and those predicted by the recommendation system in patients with hypertension from the RIGIPREV study. The model suggests optimized antihypertensive treatments based on individual patient characteristics. ACEI: angiotensin-converting enzyme inhibitor; ARB: angiotensin receptor blocker.

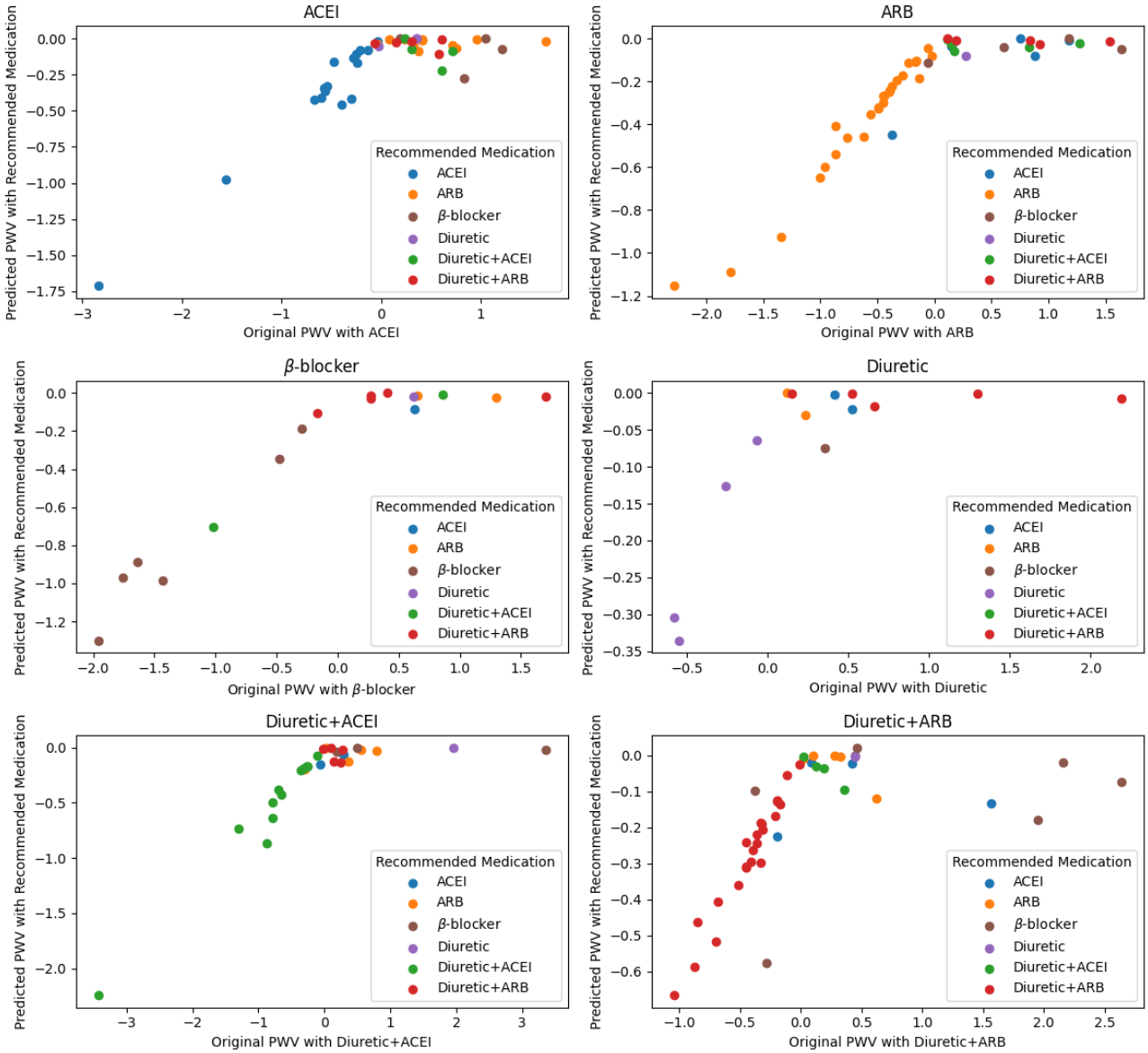

Modeling Medication Recommendations With a Decision Tree

Figure 2 shows the decision tree for selecting a suitable antihypertensive drug for PWV reduction based on patient characteristics. An overall accuracy of 84.02% was achieved when the decision tree was fitted to the medication recommendations generated by the original RF models, with 10 levels of depth. However, in terms of interpretability, only 5 levels are shown in Figure 2. After examination, the decision tree revealed that ARBs and ACEIs were the first antihypertensive drug options based on HbA_{1c}. At the second level, based on baseline cfPWV levels when a person is taking ACEI, the antihypertensive drug should be BBs or maintain ACEI, and based on total cholesterol when a person is taking ARA, diuretic+RA, or maintain ARA. Diuretics and diuretics+ACEI did not appear until the third level based on monocytes and MCV, respectively, and depended on the levels of the previous characteristics.

XSL•FO
RenderX

Figure 2. Visual decision tree representation of the underlying mechanisms and workflow of the recommendation system. This decision tree visualizes the workflow of the machine learning–based recommendation system, showing key predictors such as baseline pulse wave velocity (PWV) and patient characteristics. It was developed using data from the RIGIPREV study, conducted on patients with hypertension in Spain. ABI: ankle-brachial index; ACEI: angiotensin-converting enzyme inhibitor; AIx75: augmentation index at 75 mm Hg; ALT/SGPT: alanine aminotransferase/serum glutamic pyruvic transaminase; ARB: angiotensin receptor blocker; AST/SGOT: aspartate aminotransferase/glutamic oxaloacetic transaminase; baPWV: brachial-ankle PWV; BB: β-blocker; CAVI: cardio-ankle vascular index; cfPWV: carotid-femoral PWV; cHDL: high-density lipoprotein cholesterol; cLDL: low-density lipoprotein cholesterol; CRP: C-reactive protein; DBP: diastolic blood pressure; eGFR: estimated glomerular filtration rate; FPG: fasting plasma glucose; GGT: gamma-glutamyl transferase; HbA1c: glycated hemoglobin; IMT: intima-media thickness; LVH Cornell: left ventricular hypertrophy diagnosed by Cornell; MCV: mean corpuscular volume; MPV: mean platelet volume; RDW: red blood cell distribution width; TSH: thyroid-stimulating hormone.

Discussion

Effectiveness of the Predictive Models and Validation

In this study, we developed models using embedded RF ML models for multitarget prediction to provide recommendations on the most suitable antihypertensive drug for reducing PWV, based on the individual characteristics of participants with hypertension. The predictive models used demonstrated a high level of fit between the predictions and validation data, with R^2 values of approximately 0.65 for all the models in the internal validation. In addition, they exhibit low MSEs ranging from 0.08 to 0.22, indicating that the model predictions significantly align with the PWV values.

However, the performance of the model significantly decreased in external validation compared to its performance in internal validation. The R^2 value decreased by approximately 0.33 on average (SD 0.06), while the MSE increased by an average of 0.21 (SD 0.05). Although this decrease could indicate potential

XSL•FO
RenderX

overfitting of the regression models to the training data, it is important to consider that the use of SMOTE for generating training data could also play a role in this overfitting. The synthetic augmentation of data, while valuable for addressing sample size limitations, may introduce patterns or characteristics that are not present in the original dataset, potentially leading to models that are overly optimized for the synthetically enhanced training set rather than real-world data. This phenomenon underscores the complex balance between enhancing the dataset size for model robustness and ensuring that such augmentation accurately reflects the true data distributions.

Nevertheless, the results suggest that the models can explain a significant portion of the variance in PWV prediction and, consequently, affirm their reliability in providing recommendations in clinical practice.

Importance of Variables and Medication Recommendations

The intrinsic capability of RFs to evaluate and prioritize the importance of input features during training plays a pivotal role in understanding which variables crucially influence the predictions. This aligns seamlessly with our objective to guide clinicians in selecting the most appropriate medication by identifying key determinants of treatment effectiveness. Unlike other models, such as support vector machines, which might offer precision but lack the ability to inherently track feature importance, RF models provide unique insights into how specific variables impact outcomes, making them particularly valuable in scenarios where understanding these impacts is as crucial as achieving high prediction accuracy.

The variable weights for each of the 6 antihypertensive drug models for PWV reduction differed: ACEIs were associated with weight and HbA_{1c}, ARBs with cfPWV and low-density lipoprotein cholesterol, BBs with TSH and DBP, diuretics with CAVI and waist circumference, diuretic+ACEI with AIx75 and uric acid, and diuretic+ARB indicated basophils and mean platelet volume levels as the most important factors. Similarly, we found significant variability in the recommendations of the different models regarding the antihypertensive drug that most effectively reduced the PWV in each participant compared to the original antihypertensive drugs taken by the patients, with a matching rate of 55.3% for ARBs. Finally, the decision tree for selecting the most suitable antihypertensive drug to reduce PWV based on patient characteristics revealed that, at the first level, ARBs and ACEIs were the top choices based on HbA_{1c}, and at the second level, based on baseline cfPWV for ACEIs, the antihypertensive drug should be either BBs or continuation of ACEIs, while based on total cholesterol levels for ARBs, it should be diuretic+ARB or ARB alone.

Advantages of RF Models in the Clinical Context

RF models can handle a wide variety of variable types (both continuous and discrete) within their decision-making processes, effectively capturing the nonlinear relationships among variables. This is a significant advantage over linear models, which may struggle to interpret these complex relationships without prior data transformation. In addition, the ability of RF to manage high-dimensional data by identifying and selecting the most relevant features for model training (thereby reducing dimensionality without the need for manual intervention) distinguishes it from models such as k-nearest neighbors, which can suffer performance issues in high-dimensional spaces.

Currently, numerous clinical practice guidelines establish common guidelines for hypertension treatment and emphasize the importance of making a personalized choice regarding the most suitable antihypertensive drug [3,4]. However, the medical approach to choosing an antihypertensive drug is based on its ability to reduce blood pressure, individual patient needs, and potential side effects [3,4], without considering the effect of each antihypertensive drug on AS, although AS is an independent risk factor for morbidity and mortality from cardiovascular disease [5,6]. Therefore, we must choose an antihypertensive drug that considers both lowering blood pressure and reducing AS based on the individual characteristics of the participant. However, currently, physicians do not have sufficient information to consider all these aspects when making decisions. After demonstrating the beneficial effect [8-11] of antihypertensive drugs on AS, we analyzed individual characteristics to predict a greater reduction in AS for each antihypertensive drug using ML techniques to identify sophisticated patterns by analyzing each patient's variables and thus improving personalized care [21]. Furthermore, the embedded RF ML model used in this study is an integrated learning approach that offers excellent prediction capabilities and is tolerant of outliers [46].

Indeed, the essence of our prediction system lies precisely in its ability to infer potential outcomes for a patient based on information from other patients with similar characteristics who have been treated with various medications. In this regard, while a patient who has taken only one type of medication during the data collection period provides a direct outcome for that specific medication, the recommendation system uses 6 predictive models that have been trained with all patients who have taken different medications. Therefore, this learning process enables the system to predict how other medications could have impacted the patient by considering how patients with similar profiles responded to those other treatment options. For instance, by comparing these covariates and the observed outcomes in the dataset, the system can suggest that a patient who took ACEIs, for example, might benefit more from ARBs, based on patterns identified in patients with similar characteristics.

Comparison With Previous Studies

In recent years, numerous studies have examined ML techniques to assess which antihypertensive drug is most suitable for controlling blood pressure as well as the variables with the most significant weight in indicating its use, but we have not found studies that analyzed the effect on PWV. Among these studies, we aimed to construct a prediction model for antihypertensive drugs to be used in older patients with hypertension to help physicians quickly and rationally combine the most appropriate antihypertensive drugs based on their individual characteristics. The authors compared 5 models and found that the Light Gradient Boosting Machine algorithm predicted the best results, identifying different variables for each antihypertensive drug

XSL•FO
RenderX

analyzed [27]. Koren et al [47] applied ML techniques such as decision trees and neural networks to determine how to treat hypertension in a large group of patients from the Maccabi Healthcare Service in Israel. They found that the number of patients with hypertension treated with a single antihypertensive drug was greater than the number of patients with hypertension treated with combination therapy, and the monotherapy control rate was greater than the success rate of 2-, 3-, or 4-drug combinations. Liu et al [26] profiled 5 commonly used antihypertensive drugs (irbesartan, metoprolol, felodipine, amlodipine, and levamlodipine) and provided data for the customization of clinical antihypertensive drugs. All these studies could help tailor the administration of effective antihypertensive drugs to individual patients. Our study goes a step further and provides information on the most effective antihypertensive drugs for reducing AS on an individualized basis.

Conclusions and Limitations of This Study

In summary, this study included participants with hypertension on antihypertensive treatment from 3 research projects with at least 2 PWV measurements and a wide range of variables who were followed for at least 1 year. We developed individual recommendations to determine which antihypertensive drug achieves the greatest reductions in PWV based on individual characteristics. These results can assist physicians in making decisions regarding the prescription of antihypertensive drugs considering not only blood pressure reduction but also PWV reduction. The analysis also underscores the importance of a personalized approach to drug selection, considering a broad spectrum of patient characteristics. The differences in indicators found among antihypertensive drug groups will help guide our selection of antihypertensive drugs.

This study has several limitations, such as the omission of variables that have been shown to influence blood pressure levels, such as dietary salt intake [48], and the inclusion of variables that are not routinely measured in clinical practice, such as insulin levels. Consequently, the dataset considered for the investigated clinical question may not be the most appropriate, and the sample size is small. Nevertheless, the study also has strengths, as the data come from records of studies conducted in various research projects, and the measurement of variables was uniform and protocolized, a guarantee that cannot be provided when using electronic health record data. Given that these models learn from examples, data quality is fundamental; without meticulously selected and labeled data, models cannot be effectively built or evaluated. Furthermore, the participants who were treated with each antihypertensive drug were not randomized, but they were selected at the discretion of the clinician, which may bias our results because, although there were no statistically significant differences in most of the baseline variables, there were significant differences in mean corpuscular hemoglobin concentration, MCV, HbA_{1c}, and TSH. Finally, the main limitation of this study is related to the sample size. Although a train-test split of 80%-20% was used to evaluate the model's performance, both the training and test samples originated from the same databases. Consequently, the generalizability of the model to entirely new and external datasets remains untested. Ideally, in line with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis guidelines [49], an external validation using a completely independent dataset would have been performed to robustly ascertain the model's applicability across various patient populations. While the current approach provided valuable insights, a broader sample base in future studies could offer an even more robust understanding of the model's applicability across different patient groups.

Our research bridges a crucial gap in the current management of hypertension by not only prioritizing the reduction of blood pressure but also considering the effect on AS, as measured by PWV. AS is a widely accepted independent predictor of cardiovascular diseases. The importance of this lies in the customization of antihypertensive drug recommendations that not only effectively lower blood pressure but also aid in decreasing AS, thereby potentially reducing the overall risk of cardiovascular disease.

Our findings hold significant importance, and this can be better explained through the points mentioned subsequently. First, our study addresses a critical gap in current hypertension management by not only focusing on blood pressure reduction but also considering the impact on AS, as measured by PWV. AS is a well-established independent risk factor for cardiovascular diseases. The significance lies in tailoring antihypertensive drug recommendations that not only effectively lower blood pressure but also contribute to reducing AS, thus potentially lowering overall cardiovascular risk. Second, the personalized recommendations generated by our models offer a patient-centric approach to antihypertensive therapy. By considering individual characteristics and responses, clinicians can make more informed decisions, potentially enhancing treatment adherence and overall patient outcomes. Third, the use of advanced ML techniques, specifically the embedded RF model, enhances the reliability and sophistication of our recommendations. ML methods, when integrated appropriately, have the potential to reveal intricate patterns within complex datasets, providing more nuanced and accurate predictions than traditional approaches. Fourth, while the primary focus is on antihypertensive recommendations, it is essential to highlight that our study integrates evidence from 3 distinct studies (the EVA study, the LOD-DIABETES study, and the EVIDENT study), each with the following unique protocols. The careful consideration of key variables, consistent measurement of PWV, and adjustment for various covariates strengthen the validity of our results. In addition, the model performance metrics, including R^2 values and MSEs, indicate the robustness of our predictive models.

Furthermore, before the conclusions of this study, it is important to examine future work and the limitations of our study. Our analysis is primarily focused on data gathered at the initial and subsequent follow-up points. However, as we recognize the fluid nature of patient well-being, we are keen to examine the potential impact of incorporating multiple follow-up sessions within the data-gathering phase. This involves treating patients as dynamic variables during each follow-up, providing a more comprehensive understanding of their progressing health patterns. As we delve into our future work, addressing potential

XSL•FO
RenderX

transitions between antihypertensive medications for patients during the follow-up period holds great importance. To enhance our recommendation system, we delve into the implications of these changes and consider effective methodologies to account for them. It is imperative that we understand the interplay between patient characteristics, medication choices, and health outcomes. To enhance the strength of our recommendations, we will delve into the potential for causal inference within our model and explore the incorporation of supplementary outcomes.

In conclusion, in this study, we developed predictive models to provide recommendations regarding the most suitable antihypertensive drug for reducing PWV based on the individual characteristics of patients with hypertension. We found that at the first level, ARBs and ACEIs were the top choices based on HbA_{1c}, and at the second level, based on baseline cfPWV levels for ACEIs, antihypertensive drugs should be either BBs or ACEIs, while based on total cholesterol levels for ARBs, they should be diuretics+ARBs or ARBs alone. Therefore, our study provides a data-based reference for the customization of antihypertensive drugs to reduce PWV.

Acknowledgments

The authors would like to thank the patients and their families who participated in the studies from which the data presented here originated. This study was supported by all researchers of the Influence of Different Risk Factors in Vascular Accelerated Aging study (EVA study), the Relationship of Central Blood Pressure and Pulse Wave Velocity With Target Organ Damage study (LOD-DIABETES study), and the Effectiveness of the Use of a Mobile Tool in Improving Lifestyle (EVIDENT study). In addition, this study was supported by the Research Network on Chronicity, Primary Care and Health Promotion (RICAPPS). This study was funded by the Spanish Ministry of Science, Innovation and Universities, Instituto de Salud Carlos III through the projects PI21/00008 and PI21/00454, cofunded by the European Union (ERDF/ESF), through the projects RD21/0016/0010 and RD21/0016/0025 of the RICAPPS research network (Red de Investigación en Cronicidad, Atención Primaria y Promoción de la Salud), and cofunded by Next Generation European Union funds through the ERDF Recovery and Resilience Mechanism. The authors would also like to acknowledge the valuable contributions of the RIGIPREV Investigators Group. The members of the RIGIPREV Investigators Group are María-Carmen Patiño-Alonso, Carmela Rodríguez-Martín, Sara Mora-Simón, Leticia Gómez-Sánchez, Rosario Alonso-Domínguez, Cristina Lugones-Sánchez, Rosa María Bruno, Diana-Patricia Pozuelo-Carrascosa, Alicia Del-Saz-Lara, Irene Martínez-García, Iván Cavero-Redondo and Manuel Ángel Gómez-Marcos.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request. In addition, the code used to develop the machine learning models and recommendation system will be made publicly available through Mendeley library.

Authors' Contributions

IC-R and MAG-M prepared the original draft of the manuscript. All the authors had access to all the data in the study. The formal analysis was performed by AM-R. The methodology was developed by AM-R and IC-R. MAG-M and LG-O accessed and verified the data. Patients were recruited into any of the EVA studies, the LOD-DIABETES study, or the EVIDENT study by LG-S, VC-V, MAG-M, and LG-O. The study was conceptualized by IC-R, AM-R, AS-L, NM-H, LG-O, and MAG-M. AM-R was responsible for data curation and maintenance. Study supervision was performed by IC-R, LG-O, and MAG-M. Funding was acquired by IC-R, AS-L, LG-S, VC-V, and MAG-M. All authors wrote the manuscript and were responsible for the final decision to submit the manuscript for publication.

Conflicts of Interest

None declared.

References

1. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet 2021 Sep 11;398(10304):957-980. [doi: 10.1016/S0140-6736(21)01330-1] [Medline: 34450083]
3. Verdecchia P, Reboldi G, Angeli F. The 2020 International Society of Hypertension global hypertension practice guidelines - key messages and clinical considerations. Eur J Intern Med 2020 Dec;82:1-6. [doi: 10.1016/j.ejim.2020.09.001] [Medline: 32972800]

XSL•FO
RenderX

4. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J 2018 Sep 01;39(33):3021-3104. [doi: 10.1093/eurheartj/ehy339] [Medline: 30165516]
5. Ohkuma T, Ninomiya T, Tomiyama H, Kario K, Hoshide S, Kita Y, et al. Brachial-ankle pulse wave velocity and the risk prediction of cardiovascular disease: an individual participant data meta-analysis. Hypertension 2017 Jun;69(6):1045-1052. [doi: 10.1161/HYPERTENSIONAHA.117.09097] [Medline: 28438905]
6. Zhong Q, Hu MJ, Cui YJ, Liang L, Zhou MM, Yang YW, et al. Carotid-femoral pulse wave velocity in the prediction of cardiovascular events and mortality: an updated systematic review and meta-analysis. Angiology 2018 Aug;69(7):617-629. [doi: 10.1177/0003319717742544] [Medline: 29172654]
7. Thomopoulos C, Parati G, Zanchetti A. Effects of blood pressure lowering on outcome incidence in hypertension: 4. Effects of various classes of antihypertensive drugs--overview and meta-analyses. J Hypertens 2015 Mar;33(2):195-211. [doi: 10.1097/HJH.0000000000000447] [Medline: 25485720]
8. Cavero-Redondo I, Saz-Lara A, Lugones-Sánchez C, Pozuelo-Carrascosa DP, Gómez-Sánchez L, López-Gil JF, et al. Comparative effect of antihypertensive drugs in improving arterial stiffness in adults with hypertension (RIGIPREV study). A network meta-analysis. Front Pharmacol 2023 Sep 1;14:1225795 [FREE Full text] [doi: 10.3389/fphar.2023.1225795] [Medline: 37724181]
9. Shahin Y, Khan JA, Chetter I. Angiotensin converting enzyme inhibitors effect on arterial stiffness and wave reflections: a meta-analysis and meta-regression of randomised controlled trials. Atherosclerosis 2012 Mar;221(1):18-33. [doi: 10.1016/j.atherosclerosis.2011.12.005] [Medline: 22209214]
10. Li X, Chang P, Wang Q, Hu H, Bai F, Li N, et al. Effects of angiotensin-converting enzyme inhibitors on arterial stiffness: a systematic review and meta-analysis of randomized controlled trials. Cardiovasc Ther 2020;2020:7056184 [FREE Full text] [doi: 10.1155/2020/7056184] [Medline: 32190121]
11. Chen X, Huang B, Liu M, Li X. Effects of different types of antihypertensive agents on arterial stiffness: a systematic review and meta-analysis of randomized controlled trials. J Thorac Dis 2015 Dec;7(12):2339-2347. [doi: 10.3978/j.issn.2072-1439.2015.12.58] [Medline: 26793356]
12. Ong KT, Delerme S, Pannier B, Safar ME, Benetos A, Laurent S, et al. Aortic stiffness is reduced beyond blood pressure lowering by short-term and long-term antihypertensive treatment: a meta-analysis of individual data in 294 patients. J Hypertens 2011 Jun;29(6):1034-1042. [doi: 10.1097/HJH.0b013e328346a583] [Medline: 21519280]
13. Jatic Z, Skopljak A, Hebibovic S, Sukalo A, Rustempasic E, Valjevac A. Effects of different antihypertensive drug combinations on blood pressure and arterial stiffness. Med Arch 2019 Jun;73(3):157-162. [doi: 10.5455/medarh.2019.73.157-162] [Medline: 31391706]
14. Palombo C, Kozakova M. Arterial stiffness, atherosclerosis and cardiovascular risk: pathophysiologic mechanisms and emerging clinical indications. Vascul Pharmacol 2016 Mar;77:1-7. [doi: 10.1016/j.vph.2015.11.083] [Medline: 26643779]
15. Della Corte V, Tuttolomondo A, Pecoraro R, Di Raimondo D, Vassallo V, Pinto A. Inflammation, endothelial dysfunction and arterial stiffness as therapeutic targets in cardiovascular medicine. Curr Pharm Des 2016 Sep 07;22(30):4658-4668. [doi: 10.2174/1381612822666160510124801] [Medline: 27160758]
16. Mitchell GF, Hwang SJ, Vasan RS, Larson MG, Pencina MJ, Hamburg NM, et al. Arterial stiffness and cardiovascular events: the Framingham Heart Study. Circulation 2010 Mar 02;121(4):505-511. [doi: 10.1161/CIRCULATIONAHA.109.886655] [Medline: 20083680]
17. van Popele NM, Grobbee DE, Bots ML, Asmar R, Topouchian J, Reneman RS, et al. Association between arterial stiffness and atherosclerosis: the Rotterdam Study. Stroke 2001 Mar;32(2):454-460. [doi: 10.1161/01.str.32.2.454] [Medline: 11157182]
18. Sequí-Domínguez I, Cavero-Redondo I, Álvarez-Bueno C, Pozuelo-Carrascosa DP, Nuñez de Arenas-Arroyo S, Martínez-Vizcaíno V. Accuracy of pulse wave velocity predicting cardiovascular and all-cause mortality. A systematic review and meta-analysis. J Clin Med 2020 Jul 02;9(7):2080 [FREE Full text] [doi: 10.3390/jcm9072080] [Medline: 32630671]
19. Ben-Shlomo Y, Spears M, Boustred C, May M, Anderson SG, Benjamin EJ, et al. Aortic pulse wave velocity improves cardiovascular event prediction: an individual participant meta-analysis of prospective observational data from 17,635 subjects. J Am Coll Cardiol 2014 Feb 25;63(7):636-646 [FREE Full text] [doi: 10.1016/j.jacc.2013.09.063] [Medline: 24239664]
20. Sutton-Tyrrell K, Najjar SS, Boudreau RM, Venkitachalam L, Kupelian V, Simonsick EM, Health ABC Study. Elevated aortic pulse wave velocity, a marker of arterial stiffness, predicts cardiovascular events in well-functioning older adults. Circulation 2005 Jun 28;111(25):3384-3390. [doi: 10.1161/CIRCULATIONAHA.104.483628] [Medline: 15967850]
21. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. J Am Coll Cardiol 2021 Jan 26;77(3):300-313 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.030] [Medline: 33478654]
22. du Toit C, Tran TQ, Deo N, Aryal S, Lip S, Sykes R, et al. Survey and Evaluation of Hypertension Machine Learning Research. J Am Heart Assoc 2023 May 02;12(9):e027896 [FREE Full text] [doi: 10.1161/JAHA.122.027896] [Medline: 37119074]

XSL•FO
RenderX

23. Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular disease detection using ensemble learning. Comput Intell Neurosci 2022 Aug 16;2022:5267498 [FREE Full text] [doi: 10.1155/2022/5267498] [Medline: 36017452]
24. Chunyu L, Ran L, Junteng Z, Miye W, Jing X, Lan S, et al. Characterizing the critical features when personalizing antihypertensive drugs using spectrum analysis and machine learning methods. Artif Intell Med 2020 Apr;104:101841. [doi: 10.1016/j.artmed.2020.101841] [Medline: 32499008]
25. Silva GF, Fagundes TP, Teixeira BC, Chiavegatto Filho AD. Machine learning for hypertension prediction: a systematic review. Curr Hypertens Rep 2022 Nov 22;24(11):523-533. [doi: 10.1007/s11906-022-01212-6] [Medline: 35731335]
26. Liu C, Xu J, Liu R, Wang M, Zhuo Y, Su L, et al. Identifying drug-related attributes to personalise antihypertensive agents: the outcome report of patients receiving metoprolol therapy. BMC Med Inform Decis Mak 2021 Dec 30;21(1):370 [FREE Full text] [doi: 10.1186/s12911-021-01739-9] [Medline: 34969399]
27. Wang T, Yan Y, Xiang S, Tan J, Yang C, Zhao W. A comparative study of antihypertensive drugs prediction models for the elderly based on machine learning algorithms. Front Cardiovasc Med 2022 Dec 1;9:1056263 [FREE Full text] [doi: 10.3389/fcvm.2022.1056263] [Medline: 36531716]
28. Nuryunarsih D, Herawati L, Badi'ah A, Donsu JD, Okatiranti. Predicting changes in systolic and diastolic blood pressure of hypertensive patients in Indonesia using machine learning. Curr Hypertens Rep 2023 Nov 29;25(11):377-383 [FREE Full text] [doi: 10.1007/s11906-023-01261-5] [Medline: 37642805]
29. Lin Z, Cheng YT, Cheung BM. Machine learning algorithms identify hypokalaemia risk in people with hypertension in the United States National Health and Nutrition Examination Survey 1999-2018. Ann Med 2023 Dec 10;55(1):2209336 [FREE Full text] [doi: 10.1080/07853890.2023.2209336] [Medline: 37162442]
30. Chen Y, Huang X, Wu S, Guo P, Huang J, Zhou L, et al. Machine-learning predictive model of pregnancy-induced hypertension in the first trimester. Hypertens Res 2023 Sep 09;46(9):2135-2144. [doi: 10.1038/s41440-023-01298-8] [Medline: 37160966]
31. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov 22;24(11):1716-1720 [FREE Full text] [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
32. Ma S, Lee J, Serban N, Yang S. Deep attention Q-network for personalized treatment recommendation. In: Proceedings of the IEEE International Conference on Data Mining Workshops. 2023 Presented at: ICDMW 2023; December 1-4, 2023; Shanghai, China. [doi: 10.1109/icdmw60847.2023.00048]
33. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):57 [FREE Full text] [doi: 10.1186/s12911-019-0763-6] [Medline: 30961594]
34. Cavero-Redondo I, Saz-Lara A, García-Ortiz L, Lugones-Sánchez C, Notario-Pacheco B, Gómez-Sánchez L, et al. Comparative effect of antihypertensive drugs in improving arterial stiffness in hypertensive adults (RIGIPREV study). A protocol for network meta-analysis. Int J Environ Res Public Health 2021 Dec 18;18(24):13353 [FREE Full text] [doi: 10.3390/ijerph182413353] [Medline: 34948959]
35. Gomez-Marcos MA, Martinez-Salgado C, Gonzalez-Sarmiento R, Hernandez-Rivas JM, Sanchez-Fernandez PL, Recio-Rodriguez JI, et al. Association between different risk factors and vascular accelerated ageing (EVA study): study protocol for a cross-sectional, descriptive observational study. BMJ Open 2016 Jun 07;6(6):e011031 [FREE Full text] [doi: 10.1136/bmjopen-2016-011031] [Medline: 27267107]
36. Gómez-Marcos MA, Recio-Rodríguez JI, Rodríguez-Sánchez E, Castaño-Sánchez Y, de Cabo-Laso A, Sánchez-Salgado B, et al. Central blood pressure and pulse wave velocity: relationship to target organ damage and cardiovascular morbidity-mortality in diabetic patients or metabolic syndrome. An observational prospective study. LOD-DIABETES study protocol. BMC Public Health 2010 Mar 18;10(1):143 [FREE Full text] [doi: 10.1186/1471-2458-10-143] [Medline: 20298558]
37. Recio-Rodríguez JI, Martín-Cantera C, González-Viejo N, Gómez-Arranz A, Arietaleanizbeascoa MS, Schmolling-Guinovart Y, et al. Effectiveness of a smartphone application for improving healthy lifestyles, a randomized clinical trial (EVIDENT II): study protocol. BMC Public Health 2014 Mar 15;14(1):254 [FREE Full text] [doi: 10.1186/1471-2458-14-254] [Medline: 24628961]
38. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
39. Yamashina A, Tomiyama H, Takeda K, Tsuda H, Arai T, Hirose K, et al. Validity, reproducibility, and clinical significance of noninvasive brachial-ankle pulse wave velocity measurement. Hypertens Res 2002 May;25(3):359-364 [FREE Full text] [doi: 10.1291/hypres.25.359] [Medline: 12135313]
40. Schröder H, Fitó M, Estruch R, Martínez-González MA, Corella D, Salas-Salvadó J, et al. A short screener is valid for assessing Mediterranean diet adherence among older Spanish men and women. J Nutr 2011 Jun;141(6):1140-1145 [FREE Full text] [doi: 10.3945/jn.110.135566] [Medline: 21508208]
41. Mancia G, Fagard R, Narkiewicz K, Redón J, Zanchetti A, Böhm M, et al. 2013 ESH/ESC Guidelines for the management of arterial hypertension: the Task Force for the management of arterial hypertension of the European Society of Hypertension

XSL•FO
RenderX

(ESH) and of the European Society of Cardiology (ESC). J Hypertens 2013 Jul;31(7):1281-1357. [doi: 10.1097/01.hjh.0000431740.32696.cc] [Medline: 23817082]
42. Breiman L. Random forests. Mach Learn 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
43. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Cham, Switzerland: Springer; 2009.
44. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Inf Sci 2019 Dec;505:32-64. [doi: 10.1016/j.ins.2019.07.070]
45. Karran JC, Moodie EE, Wallace MP. Statistical method use in public health research. Scand J Public Health 2015 Nov 10;43(7):776-782. [doi: 10.1177/1403494815592735] [Medline: 26163023]
46. Wang Y, Xia ST, Tang Q, Wu J, Zhu X. A novel consistent random forest framework: Bernoulli random forests. IEEE Trans Neural Netw Learn Syst 2018 Aug;29(8):3510-3523. [doi: 10.1109/tnnls.2017.2729778]
47. Koren G, Nordon G, Radinsky K, Shalev V. Machine learning of big data in gaining insight into successful treatment of hypertension. Pharmacol Res Perspect 2018 Jun 24;6(3):e00396 [FREE Full text] [doi: 10.1002/prp2.396] [Medline: 29721321]
48. Huang L, Trieu K, Yoshimura S, Neal B, Woodward M, Campbell NR, et al. Effect of dose and duration of reduction in dietary sodium on blood pressure levels: systematic review and meta-analysis of randomised trials. BMJ 2020 Mar 24;368:m315 [FREE Full text] [doi: 10.1136/bmj.m315] [Medline: 32094151]
49. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015 Jan 06;13(1):1 [FREE Full text] [doi: 10.1186/s12916-014-0241-z] [Medline: 25563062]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
ARB: angiotensin receptor blocker
AS: arterial stiffness
baPWV: brachial-ankle pulse wave velocity
BB: β-blocker
CAVI: carotid-artery vascular index
CCB: calcium channel blocker
cfPWV: carotid-femoral pulse wave velocity
HbA1c: glycated hemoglobin
MCV: mean corpuscular volume
ML: machine learning
MSE: mean squared error
PWV: pulse wave velocity
RF: random forest
SMOTE: Synthetic Minority Oversampling Technique
TSH: thyroid-stimulating hormone

Edited by A Mavragani; submitted 07.11.23; peer-reviewed by H Ju, V Chauhan, S Ma; comments to author 11.02.24; revised version received 04.04.24; accepted 09.10.24; published 25.11.24.

Please cite as:
Cavero-Redondo I, Martinez-Rodrigo A, Saz-Lara A, Moreno-Herraiz N, Casado-Vicente V, Gomez-Sanchez L, Garcia-Ortiz L, Gomez-Marcos MA, RIGIPREV Investigators Group
Antihypertensive Drug Recommendations for Reducing Arterial Stiffness in Patients With Hypertension: Machine Learning–Based Multicohort (RIGIPREV) Study
J Med Internet Res 2024;26:e54357
URL: https://www.jmir.org/2024/1/e54357
doi:10.2196/54357
PMID:39585738

©Iván Cavero-Redondo, Arturo Martinez-Rodrigo, Alicia Saz-Lara, Nerea Moreno-Herraiz, Veronica Casado-Vicente, Leticia Gomez-Sanchez, Luis Garcia-Ortiz, Manuel A Gomez-Marcos, RIGIPREV Investigators Group. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Computable Phenotype Algorithm for Postvaccination Myocarditis/Pericarditis Detection Using Real-World Data: Validation Study

Matthew Deady[1], BA; Raymond Duncan[2], BA, MD; Matthew Sonesen[3], BA; Renier Estiandan[3], BA; Kelly Stimpert[4], BA; Sylvia Cho[5], PhD, MHS; Jeffrey Beers[1], MD; Brian Goodness[1], MA; Lance Daniel Jones[1], BA, MA; Richard Forshee[5], PhD; Steven A Anderson[5], PhD; Hussein Ezzeldin[5], PhD

[1]IBM Consulting, Bethesda, MD, United States

[2]Departments of Enterprise Information Services and Pediatrics, Cedars-Sinai Health System, Los Angeles, CA, United States

[3]Cedars-Sinai Health System, Enterprise Information Services, Los Angeles, CA, United States

[4]DRT Strategies, Inc., Arlington, VA, United States

[5]Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, United States

Corresponding Author:
Hussein Ezzeldin, PhD
Center for Biologics Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave
Silver Spring, MD, 20993
United States
Phone: 1 240 205 2215
Email: hussein.ezzeldin@fda.hhs.gov

Abstract

Background: Adverse events (AEs) associated with vaccination have traditionally been evaluated by epidemiological studies. More recently, they have gained attention due to the emergency use authorization of several COVID-19 vaccines. As part of its responsibility to conduct postmarket surveillance, the US Food and Drug Administration continues to monitor several AEs of interest to ensure the safety of vaccines, including those for COVID-19.

Objective: This study is part of the Biologics Effectiveness and Safety Initiative, which aims to improve the US Food and Drug Administration's postmarket surveillance capabilities while minimizing the burden of collecting clinical data on suspected postvaccination AEs. The objective of this study was to enhance active surveillance efforts through a pilot platform that can receive automatically reported AE cases through a health care data exchange.

Methods: We detected cases by sharing and applying computable phenotype algorithms to real-world data in health care providers' electronic health records databases. Using the fast healthcare interoperability resources standard for secure data transmission, we implemented a computable phenotype algorithm on a new health care system. The study focused on the algorithm's positive predictive value, validated through clinical records, assessing both the time required for implementation and the accuracy of AE detection.

Results: The algorithm required 200-250 hours to implement and optimize. Of the 6,574,420 clinical encounters across 694,151 patients, 30 cases were identified as potential myocarditis/pericarditis. Of these, 26 cases were retrievable, and 24 underwent clinical validation. In total, 14 cases were confirmed as definite or probable myocarditis/pericarditis, yielding a positive predictive value of 58.3% (95% CI 37.3%-76.9%). These findings underscore the algorithm's capability for real-time detection of AEs, though they also highlight variability in performance across different health care systems.

Conclusions: The study advocates for the ongoing refinement and application of distributed computable phenotype algorithms to enhance AE detection capabilities. These tools are crucial for comprehensive postmarket surveillance and improved vaccine safety monitoring. The outcomes suggest the need for further optimization to achieve more consistent results across diverse health care settings.

(J Med Internet Res 2024;26:e54597) doi:10.2196/54597

XSL•FO
RenderX

KEYWORDS

adverse event; vaccine safety; interoperability; computable phenotype; postmarket surveillance system; fast healthcare interoperability resources; FHIR; real-world data; validation study; Food and Drug Administration; electronic health records; COVID-19 vaccine

Introduction

Background

The US Food and Drug Administration (FDA) Center for Biologics Evaluation and Research (CBER) is responsible for ensuring the safety, purity, potency, and efficacy of biological products, including vaccines, allergenics, blood and blood products, and cells, tissues, and gene therapies [1]. In partnership with the Centers for Disease Control and Prevention, the FDA monitors vaccine safety through the vaccine adverse event reporting system (VAERS). Since the emergency use authorization of 4 COVID-19 vaccines—Pfizer-BioNTech, Moderna, Janssen, and Novavax—the FDA has been diligently analyzing adverse events (AEs) of special interest (AESIs) related to these vaccines and their boosters, including bivalent boosters.

VAERS receives spontaneous reports of suspected vaccine AEs from various sources, such as individuals, patients, clinical staff, and vaccine manufacturers. VAERS plays a vital role as an early warning system, relying on voluntary reports to monitor potential AEs. While this approach is invaluable, it can sometimes lead to underreporting, data quality issues, and challenges in establishing direct causality between vaccines and AEs. To enhance the robustness of monitoring, especially during public health emergencies, active surveillance efforts are essential in complementing and addressing the limitations of passive reporting systems like VAERS and FAERS. Established in 2017, CBER's Biologics Effectiveness and Safety (BEST) Initiative creates a robust postmarket surveillance system for biologic products by leveraging real-world data (RWD) from sources such as electronic health records (EHRs) and claims data [2]. Through collaborations with various organizations, BEST enhances the detection and reporting of biologics-related AEs by assessing extensive health care data and scientific expertise [3,4].

The key focus of BEST is developing interoperable, computable phenotype algorithms that use standardized language to identify potential AEs in EHR databases [2]. This approach allows for automatic or semiautomatic detection and reporting of AEs with minimal effort from health care providers. In a study with the first site (site #1) [5], BEST successfully validated these algorithms, demonstrating their effectiveness in identifying AEs such as myocarditis (inflammation of the myocardium) and pericarditis (inflammation of the pericardium) following vaccination.

The ultimate goal is to implement a standardized, automated case reporting system across different EHR databases, enabling precise and timely identification of postvaccination AEs and their subsequent reporting to the FDA for further evaluation. This system aims to improve overall vaccine safety by swiftly identifying potential increases in AE incidence rates [4].

Exchange Pilot Platform

To enhance CBER's capability for postmarket surveillance, accessing a diverse and representative set of data on a near-national scale is essential. In a traditional federated network, including additional partners to achieve this scale, although essential, is prohibitively time-intensive. Therefore, a more efficient approach is needed to access data on a national scale and to address the potential undercounting of AEs.

To obtain the necessary RWD for potential AE cases from health care providers nationwide, we developed a pilot cloud-based platform to exchange EHR data with the FDA for analysis and active surveillance efforts. CBER BEST Innovative Methods initiative collaborated with the eHealth Exchange (eHX), a health information exchange, to explore the semiautomated detection and reporting of potential COVID-19 vaccine–related AEs. This platform serves as a valuable tool for public health awareness, safety, and transparency while minimizing the burden on health care providers in collecting clinical data.

Figure 1 demonstrates how data are exchanged from the health care provider through eHX to the FDA BEST platform.

This platform uses recent regulations under the 21st Century Cures Act Final Rule, which mandates that all clinical data be available through an application programming interface in the Fast Healthcare Interoperability Resources (FHIR) standard developed by Health Level Seven, Inc. (HL7) to improve interoperability between health systems by simplifying data exchange. These regulations facilitate seamless and secure access, exchange, and use of electronic health information [6,7]. FHIR is a standard developed by HL7 to facilitate interoperability between health systems. It is designed to improve upon existing standards by reducing implementation complexity without losing information integrity [3].

The EHR data provided through the exchanges give the FDA access to richer clinical datasets compared to those currently submitted to VAERS, reducing the need for further translation to a common data model by one or both parties. These data include unstructured clinician notes, which are often critical for AE analysis. This pilot platform is a first step in establishing an automated reporting system for AEs. To scale nationally, computable phenotypes must be dependable across health care provider systems, allowing the transmission of probable AE reports to the BEST platform for further evaluation to support postmarket safety and effectiveness.

This study models a nationwide automated reporting system by evaluating the effective distribution and implementation of an interoperable, computable phenotype algorithm at a new health care partner site (site #2), using a different EHR platform than at the initial validation site (site #1) [8]. The algorithm detected postvaccination AE case candidates for automatic reporting through our exchange pilot platform. The study focused on the algorithm's implementation for identifying myocarditis and

XSL•FO
RenderX

pericarditis, rare but significant risk factors following COVID-19 vaccinations, at site #2. Clinical data for detected cases were transmitted through the health care exchange and validated to assess detection and reporting accuracy. The results were compared to those from the previous validation study using the same phenotype algorithm at site #1.

A common issue with any distributed algorithm, especially for health care data, is the potential drop in performance when an algorithm developed at one site (site #1) is applied to a different site (site #2). Therefore, it is often difficult to generalize results from algorithms developed at multiple sites, leading to significant differences in positive predictive value (PPV) between sites, which can hinder the use of algorithms at a national level. This study aimed to determine whether the results of the algorithm were generalizable by comparing the PPV results from site #1 to those from site #2. The long-term goal is to accurately detect postvaccination AE cases by applying standardized phenotypes across different EHR databases.

Distributed algorithms require significant overhead due to the volume of translation to different data models or code systems. This study assessed the efficacy of the phenotype algorithm by measuring the time required for implementation by the health care organization. It also aimed to determine whether the BEST pilot platform received all necessary patient case data to validate the postvaccination AEs such as myocarditis/pericarditis. The evaluation criteria included the percentage of cases with insufficient evidence (eg, missing lab tests or clinical outputs needed for the case definition assessment) to complete a validation determination. This study is part of the FDA's BEST initiative, and the findings from this validation study will help BEST staff to assess the long-term feasibility of scaling to a nationwide, automated detection using an active surveillance system.

Figure 1. CBER (Center for Biologics Evaluation and Research) BEST (Biologics Effectiveness and Safety) Exchange pilot platform. FDA: US Food and Drug Administration; FHIR: Fast Healthcare Interoperability Resources.

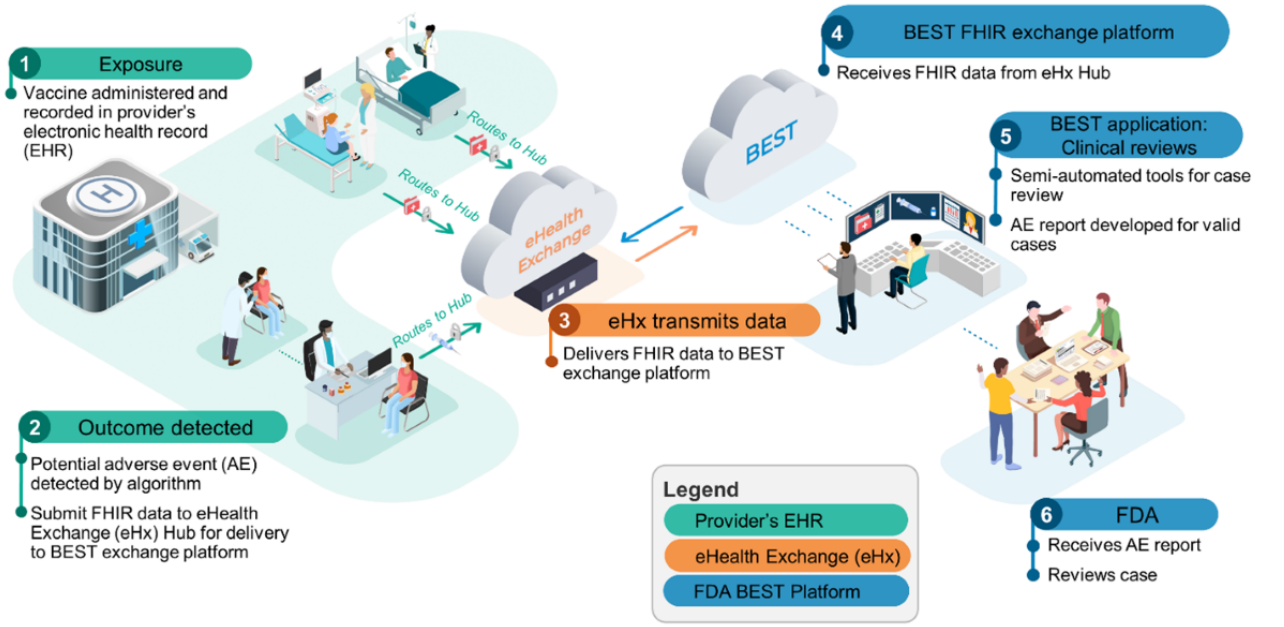

Methods

Overview

The methods section is crucial for ensuring the accuracy and reliability of research findings. This study's methodology encompasses the following aspects: computable phenotype development, phenotype distributed deployment, study period, data, and the process of reviewing medical records. Each methodological aspect plays a vital role in the comprehensive evaluation of our study's findings, ensuring the robustness and validity of the results.

Computable Phenotype Development

For this validation study, we applied the algorithm developed for the postvaccination myocarditis/pericarditis phenotype used in the validation study by Holdefer et al [5] in 2023. The algorithm was built using a myocarditis/pericarditis case definition, which was current at the time of the Holdefer study. While a new case definition [9] was published in 2022, no changes to the developed phenotype algorithm were recommended. Therefore, we continued using the original algorithm for detection and the new case definition for our clinician validation. However, we made one change to the phenotype algorithm used in the Holdefer (2023) validation study by adding a vaccine exposure requirement. This requirement is for the myocarditis/pericarditis diagnosis to occur within a risk window of 0-42 days post vaccination.

Table S1 in the Multimedia Appendix describes the definitions of the terms. The specific search terms used to develop the code lists for myocarditis/pericarditis case definition concepts are listed in Tables S2-S6 in the Multimedia Appendix. Myocarditis/pericarditis were chosen due to reports of potential safety concerns after the initial distribution of COVID-19 vaccinations [9]. Recent research has indicated a reduction in the incidence of myocarditis/pericarditis associated with boosters compared to the initial 2-dose regimen for mRNA vaccines,

XSL•FO
RenderX

likely due to interdose intervals [10,11]. However, myocarditis/pericarditis remain a monitoring concern given the recommendation for continued COVID-19 vaccinations.

Phenotype Distributed Deployment

The BEST team provided a newly developed phenotype algorithm and code lists to our exchange data partner at site #2 in 2 standardized formats: clinical quality language (CQL) and observational medical outcomes partnership standard query language. The partner then implemented the algorithm in their Epic Clarity reporting databases. Detection of an AE triggered a secure submission of patient demographics to the FDA BEST platform.

The BEST team subsequently initiated a data request to the eHX, using demographics to locate patients via the FHIR search functionality. Future updates will enable the automatic transmission of relevant patient information through eHX to the BEST platform upon detecting potential AEs at the partner site, eliminating the need for manual intervention. To evaluate the algorithm's sharing and application success, we interviewed the IT team members responsible for the implementation and documented the findings, including the estimated hours needed for implementation and testing. These are described in the results section.

Study Period

The study period was from December 14, 2020, when COVID-19 vaccinations received emergency use authorization, through April 28, 2023. The observation period began on December 14, 2019, to ensure all patients had at least 1 year of historical data, necessary to evaluate the clean window, as outlined in Figure 2.

Figure 2. Standard algorithm template logic. AE: adverse event.

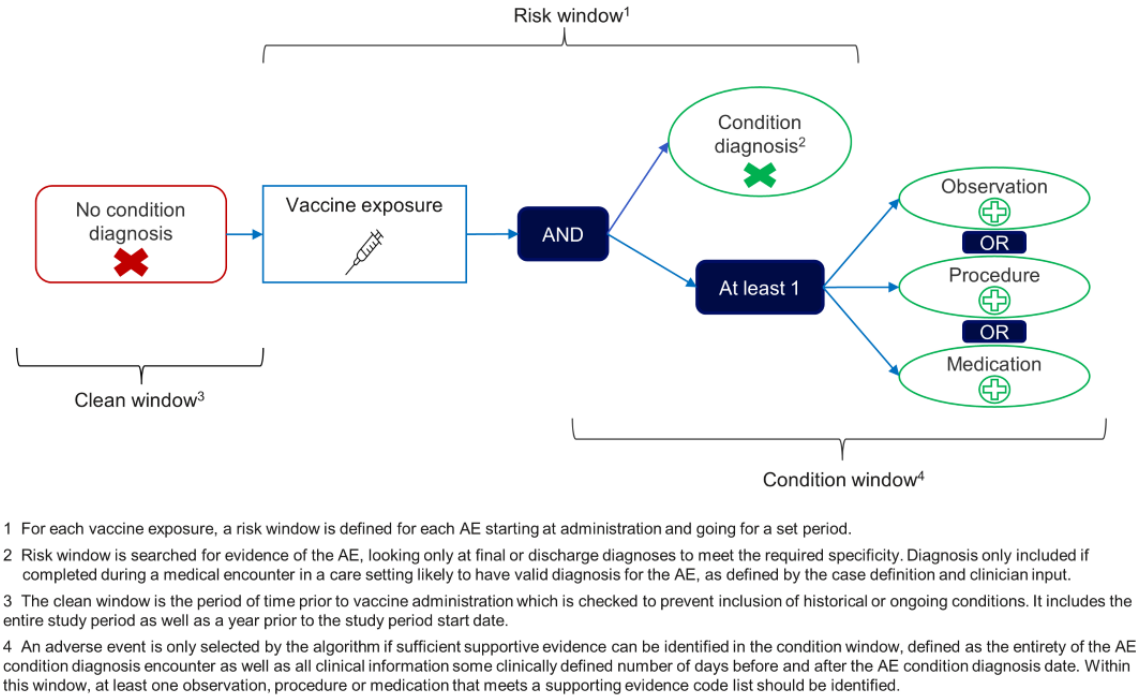

Data

The study population was derived from site #2, a health care provider partner recruited through their participation in the eHX system. eHX is a mature network of US exchange partners covering 77% of the nation's state and regional health information exchanges, including the federal government and private health care networks. It securely shares clinical information over the internet using a standardized approach [12] and prenegotiated common agreements, providing timely access to RWD without the need for individual data use negotiations with partners. During the study period, site #2 provided care to 694,151 patients across over 6.5 million medical encounters. Table 1 presents the patient and encounter demographic breakdown for total and inpatient encounters, the only care settings considered for the postvaccination myocarditis/pericarditis phenotype.

Cases were selected using a distributed computable phenotype algorithm, which the site #2 team translated into a standard query language query for their Epic Clarity reporting database. The clinical data for the selected cases were provided to the BEST study team through site #2's Epic FHIR application programming interface, which was queried by eHX and then sent to the BEST pilot platform. This included the data required for the algorithm, as well as additional data such as clinician notes, allergies, and procedures required for validation. A list of FHIR resources received for each patient identified by the algorithm is included in Table 2.

XSL•FO
RenderX

Table 1. Demographics of the academic health system for the study population (N=694,151).

Category	Inpatient		Total	
	Patients (n=78,946), n (%)	Encounters (114,254), n (%)	Patients (694,151), n (%)	Encounters (n=6,574,420), n (%)
Age (years)				
<5	2134 (2.70)	2328 (2.04)	30,256 (4.36)	137,068 (2.08)
5-17	888 (1.12)	1129 (0.99)	38,711 (5.58)	180,857 (2.75)
18-24	2017 (2.55)	2699 (2.36)	45,859 (6.61)	222,254 (3.38)
25-44	23,467 (29.73)	29,044 (25.42)	214,969 (30.97)	1,555,636 (23.66)
45-64	17,727 (22.45)	26,886 (23.53)	194,912 (28.08)	2,057,258 (31.29)
>65	32,677 (41.39)	52,131 (45.63)	167,988 (24.20)	2,418,968 (36.79)
Missing	36 (0.05)	37 (0.03)	1456 (0.21)	2379 (0.04)
Sex				
Male	45,327 (57.42)	63,534 (55.61)	317,393 (45.72)	2,808,578 (42.72)
Female	33,590 (42.55)	50,691 (44.37)	375,258 (54.06)	3,762,854 (57.23)
Missing or other	29 (0.04)	29 (0.03)	1500 (0.22)	2988 (0.05)
Race				
White	51,531 (65.27)	74,462 (65.17)	420,483 (60.58)	4,231,991 (64.37)
Black or African American	11,425 (14.47)	18,367 (16.08)	72,882 (10.50)	802,942 (12.21)
Asian or Pacific Islander	6797 (8.61)	9247 (8.09)	62,467 (9.00)	630,886 (9.60)
American Indian or Alaska Native	265 (0.34)	427 (0.37)	2104 (0.30)	25,755 (0.39)
Other	7440 (9.42)	10,085 (8.83)	57,815 (8.33)	537,872 (8.18)
Unknown	1171 (1.48)	1243 (1.09)	73,092 (10.53)	298,228 (4.54)
Declined to answer	317 (0.40)	423 (0.37)	5308 (0.76)	46,746 (0.71)
Ethnicity				
Hispanic	13,536 (17.15)	19,767 (17.30)	99,502 (14.33)	981,688 (14.93)
Non-Hispanic	63,771 (80.78)	92,673 (81.11)	507,626 (73.13)	5,206,326 (79.19)
Unknown	1639 (2.08)	1814 (1.59)	87,023 (12.54)	386,406 (5.88)

Table 2. Period of resources received for validation by resource type.

Resource	Date period
Allergy intolerance	Full clinical history
Condition	Full clinical history
Diagnostic report	Study period (12/14/2020-04/28/2023)
Document reference	Study period (12/14/2020-04/28/2023)
Encounter	Full clinical history
Immunization	Full clinical history
Location	All linked resources
Medication	All linked resources
Medication request	Full clinical history
Observation	Study period (12/14/2020-04/28/2023)
Patient	Single patient resource
Practitioner	All linked resources
Procedure	Full clinical history

XSL•FO
RenderX

Process for the Review of Medical Records

All cases selected by the algorithm and pulled into the BEST platform were reviewed using a chart tool for BEST clinician review. Each case was assigned to 2 clinicians. The clinical validation used a patient's comprehensive clinical history, including EHR data and all clinical notes. This dataset included information not used by the detection algorithm, such as different data types (eg, allergies, clinical notes, etc) and data that were filtered out (eg, admitting diagnosis, encounters with various care settings, etc). For each case, the reviewer evaluated whether the clinical data evidence met the specified case definition criteria. Relevant patient data for the case window were accessible and presented to the clinicians in a chart review application designed specifically for this purpose (see Figures S1-S7 in the Multimedia Appendix). Within the tool, clinicians could group items by type, search across all items and text, and request additional chart data to expand the window and access any available historical patient data.

All suspected AEs were validated using the conventional categories and definitions for rating diagnostic certainty of AEs: definite, probable, possible, or doubtful [13]. A case was classified as definite, probable, or possible if it met the published case definitions for that determination. Any case not meeting these criteria was labeled as doubtful. In cases of disagreement between the 2 clinicians, a third clinician made the final determination by reviewing the EHR data.

Clinicians could not confirm a case as positive (definite or probable) in case of insufficient information within the structured (data organized according to an explicit data model or data structure, often stored in a relational database) or unstructured (without a predefined data model, like free text) EHR data. In such instances, they marked the review as possible insufficient evidence, indicating that an AE might have occurred, but there was not enough documentation to meet the case definition requirements (eg, some case definitions may require specific test results that were not performed or not recorded in the EHR during patient care).

Statistical Analysis

Statistical analysis in this study involved the examination and interpretation of data to evaluate the accuracy and reliability of our algorithm for identifying AEs. This section covers PPV, CIs, and interrater reliability. Each of these measures provides a distinct perspective on the algorithm's performance and the consistency of clinical adjudication.

Positive Predictive Value

PPV was used to assess the accuracy of our algorithm. The PPV for each algorithm was determined as the proportion of positive AEs identified by the algorithm that were confirmed by clinical adjudication, out of all cases with sufficient evidence for clinical adjudication. This study's PPV was compared to the one in the Holdefer validation study [5], which reported a PPV of 83.5% (95% CI 74.9%-89.6%).

Cases flagged as having insufficient evidence for validation were excluded from the calculation. These cases were initially considered as possible AEs due to some evidence of an AE. However, they lacked one or more pieces of crucial evidence required to meet the case definition (ie, no record of results for a lab test expected for use in diagnosis).

As indicated in Table 3, 4 cases labeled definite or probable were marked as true positive (where the clinician agreed with the algorithm's prediction), while cases labeled possible or doubtful were marked as false positive (where the algorithm predicted a positive AE result, but clinicians could not confirm the occurrence of the AE).

Table 3. Case marking for PPV[a] calculation.

Case labeling	Case marking
Insufficient evidence	Not included in PPV calculation
Definite or probable	True positives
Possible or doubtful	False positives

[a]PPV: Positive predictive value.

The PPV for this study was calculated as follows:

PPV=true positives (definite+probable)/all cases selected by algorithm with sufficient evidence (definite+probable+possible+doubtful).

As a sensitivity analysis, we also calculated PPV by including the cases without sufficient evidence as false positives. This alternative PPV was calculated as follows:

PPV=true positives (definite+probable)/all cases selected by algorithm (definite+probable+possible+doubtful).

CIs

Since PPV is a binominal proportion, we calculated CIs for the performance metrics using the Agresti-Coull interval [14], the recommended method for estimating accurate CIs for binomial proportions such as PPV [15].

Interrater Reliability

Interrater reliability was used to measure the extent to which the 2 physicians agreed in their AE assessment. It was calculated using Cohen kappa between the first 2 reviewers. Cohen κ measures the agreement between 2 raters classifying instances into mutually exclusive groups [16].

Ethical Considerations

This study was conducted under the Food and Drug Administration's Sentinel initiative (Public Health Surveillance Activities), which are not deemed research per the guidance from the Office for Human Research Protections (OHRP), which

XSL•FO
RenderX

has determined that the regulations they administer (45 CFR part 46) do not apply to the activities that are included in the FDA's Sentinel initiative.

Results

Overview

The results section presents the key findings of the study organized under Population Sample, Phenotype Implementation Results, and Overall PPV and Interrater Reliability Results. These result findings detail the selection of cases, the implementation of the phenotype algorithm, and the evaluation of the algorithm's performance and reliability.

Population Sample

The computable phenotype algorithm selected cases from 114,254 inpatient encounters out of approximately 6.5 million total clinical encounters available during the study period. The algorithm identified 484 cases with a condition diagnosis and supporting evidence for myocarditis/pericarditis. Of those 484 cases, 30 met the algorithm's additional vaccine exposure criteria, and 26 could be pulled into the BEST platform for our clinical validation study.

The cases that could not be retrieved through the eHX had demographic information matching multiple patients in the EHR database. This prevented data transfer for privacy reasons since the FDA cannot currently confirm that both patients had a potential postvaccination AE. A CONSORT (Consolidated Standards of Reporting Trials) diagram is included in Figure 3 to illustrate how the steps of the algorithm in filtering encounters produce our validation sample.

As indicated in Figure 3 cases were not pulled in because of multiple matches for the provided patient demographics.

Figure 3. Study Population Consolidated Standards of Reporting Trials (CONSORT) diagram.

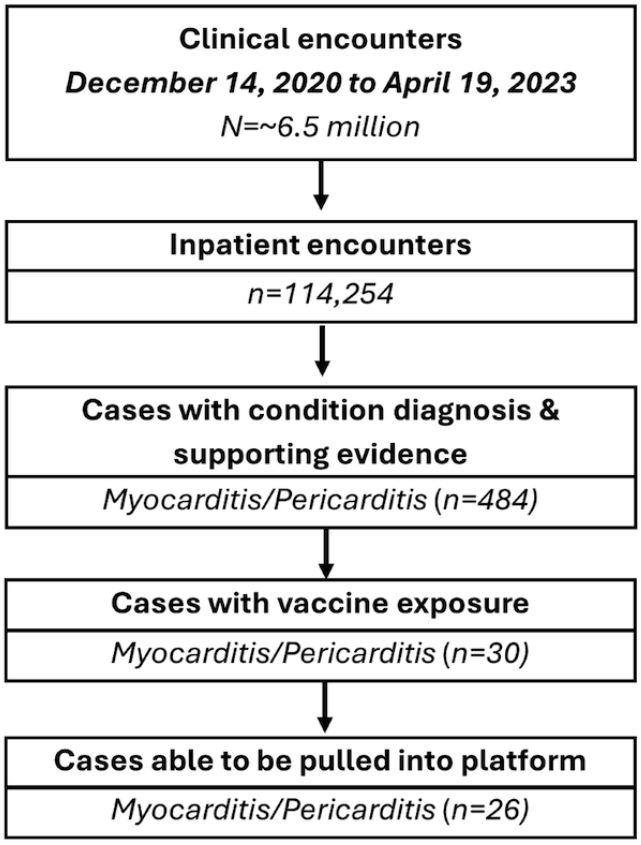

Phenotype Implementation Results

It took about 200-250 hours to successfully implement the phenotype on site #2's Epic Clarity database. Considering the size of the EHR database, most of this time was spent designing an efficient data extraction solution and testing the logic to ensure it accounted for all possible scenarios, particularly in patients with multiple immunizations or AE diagnosis codes. The implementation team at site #2 believes that the query implementation time could be substantially reduced to approximately 20 hours from the current 200-250 hours at the next partner site using the same Epic EHR software by reusing the queries from this implementation. This is presented in further detail in the discussion section.

There were no issues matching the interoperable code list to the Epic Clarity database.

Overall PPV and Interrater Reliability Results

The overall algorithm performance characteristics for the myocarditis/pericarditis computable phenotype as executed in the EHR system using cases with sufficient evidence are presented in Table 4.

For the 26 cases validated, PPV was 58.3% (95% CI 37.3%-76.9%), with approximately 7.7% of the data insufficient

XSL•FO
RenderX

to make a determination. The interrater reliability Cohen κ score for the 26 reviewed cases was 0.699, which shows substantial agreement between the clinicians. This measure suggested sufficient reliability since the value was greater than 0.61, although it did not meet the higher threshold of 0.80 suggested by other articles [16].

Table 4. Site #2 total validation sample results.

Partner and final case classification (*Cedars Sinai*)	Detected cases
Definite, n	3
Probable, n	11
Possible, n	0
Doubtful, n	10
Total true positive cases (definite+probable), n	14
Total false positive cases (possible+doubtful), n	10
Total cases with sufficient evidence, n	24
PPV (definite+probable/total cases with sufficient evidence; %), % (95% CI)	58.3 (37.3-76.9)
Possible (insufficient evidence), n	2
Total cases, n	26
% of Cases with insufficient evidence, %	7.7
PPV–alternate calculation (definite+probable/total cases; %), % (95% CI)	53.8 (34.2-72.5)

Discussion

Overview

The discussion section highlights the findings of our study on the computable phenotype for postvaccination myocarditis/pericarditis AE across 2 health care organizations, site #1 and site #2. The discussion covers Principal Findings, Comparison of Validation to Previous Site Validation, and Methodological Considerations and Limitations. Our study underscores the importance of continued research to refine these algorithms and ensure their success in revolutionizing public health surveillance.

Principal Findings

This study demonstrated that a simple, rules-based, postvaccination myocarditis/pericarditis AE-computable phenotype from one health care organization (site #1) [5] can be distributed and replicated at a separate health care organization (site #2) with comparable performance. However, the study also revealed a decline in PPV and a significant time cost for implementation at the new partner. The computable phenotype at site #2 had a PPV of 58.3% (95% CI 37.3%-76.9%), which was lower than that of 83.5% (95% CI 74.9%-89.6%) at site #1 [5]. This was driven largely by a higher percentage of cases with evidence of alternative diagnoses.

Additionally, the study showed that live, FHIR-formatted clinical data received through a health exchange-based platform are robust and interoperable enough for complex clinical case validation, in which we received direct data extracts from EHR tables from our health care partner from site #1 [5]. The proportion of cases identified by validating clinicians as containing insufficient data in this study was much lower than those in the previous study, where direct data extracts from EHR tables were used. Lastly, it demonstrated that our phenotypes could be implemented by health care analysts on EHR data unseen by algorithm developers, although with significant effort.

Feedback from our partner indicated that the phenotype took approximately 200-250 hours to implement. However, the implementation team believes that this time can be drastically reduced to approximately 20 hours by reusing the solution at subsequent data partners using the same Epic EHR software. They emphasized that the ease of deploying the algorithm is crucial for scaling the solution to other health organization data partners and provided several suggestions (as detailed in the section Methodological Considerations and Limitations) to make the process easier in the future.

Comparison of Validation to Previous Site Validation

The PPV of this study's validation was lower than the previously completed myocarditis/pericarditis validation (58.3%; 95% CI 37.3%-76.9%) versus (83.5%; 95% CI 74.9%-89.6%). Most false positives in this study occurred due to an alternative diagnosis that better explained the symptoms, such as myocardial infarction or pulmonary embolism [17]. This highlights the challenge of creating a reliable algorithm for complex diagnoses such as myocarditis or pericarditis, as they often present similarly to other conditions.

Possible explanations for the degradation in performance of the computable phenotype at site #2 include but are not limited to the following (Textbox 1).

If all cases with insufficient evidence were assumed negative, the PPVs of the 2 studies would be closer: 53.8% (95% CI 34.2%-72.5%) versus 63.7% (95% CI 55.2%-71.4%), and not statistically significant at $P<.10$.

XSL•FO
RenderX

Textbox 1. Possible explanations for the degradation in performance.

- Different population compositions: The populations from different hospital systems with varying geographies and demographics, potentially affecting medical treatment types.
- Algorithm changes: A vaccine exposure requirement was added [17].
- Case definition changes: An updated version of the case definition was used.
- Statistical noise from the small sample size: The 95% CI positive predictive value (PPV) for this study was 39.6%, between 37.3% and 76.9%, suggesting statistical noise. A 2-sample proportion test comparing PPVs for the 2 studies is significant at a $P<.01$. However, the number of cases with insufficient evidence complicates comparing algorithm PPVs. Due to the rarity of postvaccination myocarditis/pericarditis, we included all the cases identified by our algorithm at site #2.
- Different levels of evidence: Categorizing cases with insufficient evidence was challenging in both studies. We assumed that the cases from the PPV were as likely to be positive or negative as unreviewed cases, with missing details possibly not documented for various reasons. We excluded these cases from the PPV calculation. However, it is possible that the lack of evidence indicated a clinician's doubt about the diagnosis, leading to less detailed notes or fewer diagnostic tests. This could vary across sites due to differences in processes, patient populations, and available technology. Notably, the previous validation excluded a quarter of myocarditis cases due to insufficient evidence, compared to less than 8% in this study.

Methodological Considerations and Limitations

There were several limitations to this study. First, there were issues with measuring and comparing the PPV to our previous site (site #1) validation, as discussed above, including small sample sizes and challenges in handling insufficient evidence cases. Second, although the algorithms were designed to be simple to employ, they required customization from CQL or observational medical outcomes partnership queries to be implemented in the existing EHR system's available querying tools. Additionally, a substantial amount of time was needed to design and optimize the queries for large EHR databases. This customization must be made easier for health care data partners in the future.

We collected feedback from our data partner regarding potential concerns and improvements to make it easier to apply the algorithm to another site without sacrificing accuracy. Our data partner suggested that the query implementation time could be drastically reduced if we provided a design template for efficient phenotype queries and validations specific to an EHR vendor's system; added additional text description around some of the phenotype logic; and simplified the algorithms to make them easier to implement.

Future research involving a new partner site can be designed to assess the reduction in time with the template from site #2, along with simplified and improved algorithm documentation.

Even with the recommended improvements, the customization required makes scaling the algorithms nationwide challenging. In the future, the evolving landscape of health IT may facilitate the public health use cases of detecting and reporting postvaccination AESIs in a safe and secure manner that protects patient privacy over the process described in this paper.

A truly interoperable ability to query across health providers is a demonstrated need to address variations in underlying technology, standards, and connectivity. The stakeholders include the electronic clinical report program jointly managed by the Association of Public Health Laboratories, Centers for Disease Control and Prevention, and Council of State and Territorial Epidemiologists; the Electronic medical record Support for Public health network for case reporting [18]; and the Helios HL7 FHIR Accelerator, Aggregate Data initiative [19], which has recently focused on capacity reporting such as open beds, ICU beds, emergency department visits, and similar statistics supporting emerging public health responses.

These initiatives vary in the complexity of the queries or trigger codes that identify a data point. A common thread is the inclusion or centrality of FHIR standards or profiles. A challenge for the BEST use case is the need for more complex queries linking biologic exposures to potential outcomes. This could be achieved by EHRs supporting secure querying of patient cohorts with probable postvaccination AEs using CQL [20] or another interoperable query language. Reducing the burden of automatic detection of postvaccination AEs would help public health organizations improve AE surveillance with minimal additional burden to health care institutions and providers.

A final limitation of this study is that the algorithms were only applied to a limited number of sites. Moving forward, the algorithm's performance needs to be validated through exchanges at additional sites, preferably with a larger sample of potential postvaccination AE cases. This would ensure the following (Textbox 2).

XSL•FO
RenderX

Textbox 2. Algorithm generalizability and challenges in implementation across diverse healthcare settings and electronic health record systems.

- Performance generalizes to other health care sites: The algorithm has only been applied to site #1 and site #2. It may have different operating characteristics (positive predictive value, sensitivity, etc) at additional locations.
- Performance generalizes to other postvaccination adverse events (AEs): Other algorithms could not be validated due to sample size issues arising from the rarity of postvaccination AEs. These algorithms may be validated with a larger data partner or a series of validations at different data partners.
- Algorithms can be applied to other sites' electronic health record (EHR) implementations: Although the phenotype was applied to one Epic site, the results could differ for other Epic EHR implementations. We are currently researching variability in Fast Healthcare Interoperability Resources (FHIR) end points and how it might affect postmarket vaccine safety and effectiveness research (Deady et al, in preparation).
- Data received from other EHR vendor systems are similarly robust: Data captured or received through the exchange platform could vary among health care providers using other vendors' EHR systems, altering the validation results and reducing the usefulness of this method for postvaccination AE detection.

Conclusions

Active surveillance systems have the potential to enhance vaccine safety, contribute to the development and use of safer vaccines, and provide evidence to minimize the risk of postvaccination AEs [21]. These systems can also detect suspected or unknown AE cases. The algorithms described herein were developed and validated using a methodology that is recommended to be applied to and generalized for new EHR databases.

Before expanding the application of these algorithms, additional research is needed to improve their accuracy. This requires collaboration with additional partner EHR systems. Machine learning techniques could potentially be used to train the model to identify specific data patterns, creating a more effective algorithm compared to the rules-based methods that incorporate published case definition criteria and clinical subject matter experts. With sufficient data, machine learning approaches outperform rules-based approaches across domains, including the medical domain in which we received direct data extracts from EHR tables from our health care partner from site #1 [22]. However, machine learning methods may not be generalizable across different EHR systems because the data identified could be specific to individual health care organizations. Building a large dataset that combines multi-site data is challenging and costly due to concerns over infrastructure, regulations, privacy, and data standardization. Federated learning could be explored to address this problem. This method allows multiple sites to collaborate in training a global model without directly sharing data and has been used to train machine learning algorithms at EHR sites [23]. However, given the BEST initiatives' goal of reducing burden on providers, federated learning faces multiple barriers to becoming nationally feasible. In the meantime, current methods of conducting large-scale surveillance studies with health care provider networks remain critical for postmarket safety and surveillance [24].

BEST staff will use the findings from this validation study to inform the long-term feasibility of scaling to a nationwide, semi-automated detection approach for an active surveillance system. Further research and investigation are needed to enhance algorithm performance and integration across health care organizations for active surveillance in the interest of public health.

Acknowledgments

We thank the US Food and Drug Administration's (FDA's) Center for Biologics Evaluation and Research (CBER) team members and their partners for critical revisions to the manuscript. The description of eHealth Exchange and the ability to receive the data for the validation study would not have been possible without the assistance of Mike McCune, Kathryn Bingman, and Eric Heflin from eHealth Exchange. Special thanks are also due to our clinician reviewers Jeffrey Lenert, Tim Burrell, and Ananth Srikrishnan. Additional feedback on the manuscript development was provided by Ashley Holdefer, Arianna Sang, Nader Alawar, Mikhail Flom, Abhishek Pandey, and Erik Martinez from IBM Consulting. This research was funded through the FDA CBER BEST Initiative. Several co-authors hold commercial affiliations with IBM Consulting, DRT Consulting and Cedars-Sinai Health System. IBM Consulting (BG, JB, LJ, and MD), and Cedars-Sinai Health System (RD, MS, and RE) provided support in the form of salaries for authors but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views and opinions expressed here represent those of the authors and should not be considered as advice or guidance on behalf of the US Food and Drug Administration.

Conflicts of Interest

JB, LDJ, BG, and MD are and were employed by IBM while participating in the study. RD, MS, and RE are and were employed by Cedars-Sinai Health System. These authors have delivered clinical and epidemiology consulting engagement for public and private sector partners. These affiliations did not impact the study design, data collection and analysis, decision to publish, or preparation of the manuscript, and do not alter our adherence to JMIR policies on sharing data and materials. The opinions expressed are those of the authors and do not necessarily represent the opinions of their respective organizations. The remaining

XSL•FO
RenderX

authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Tables S1-S6 provide myocarditis and pericarditis case definition description, reference, computable phenotype search terms and code lists. Figure S1-S7 provide screenshots of the BEST Platform validation processes.
[DOCX File , 2401 KB - jmir_v26i1e54597_app1.docx]

References

1. FDA. CBER Vision & Mission. 2019. URL: https://www.fda.gov/about-fda/center-biologics-evaluation-and-research-cber/cber-vision-mission [accessed 2023-02-22]
2. FDA. Framework for FDA's Real-World Evidence Program. 2018. URL: https://www.fda.gov/drugs/ [accessed 2024-10-08]
3. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The fast health interoperability resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
4. Baker MA, Kaelber DC, Bar-Shain DS, Moro PL, Zambarano B, Mazza M, et al. Advanced clinical decision support for vaccine adverse event detection and reporting. Clin Infect Dis 2015;61(6):864-870 [FREE Full text] [doi: 10.1093/cid/civ430] [Medline: 26060294]
5. Holdefer AA, Pizarro J, Saunders-Hastings P, Beers J, Sang A, Hettinger AZ, et al. Development of interoperable computable phenotype algorithms for adverse events of special interest to be used for biologics safety surveillance: validation study. JMIR Public Health Surveill 2024;10:e49811 [FREE Full text] [doi: 10.2196/49811] [Medline: 39008361]
6. The ONC Cures Act, Final Rule. 2020. URL: https://fire.ly/regulations/onc-cures-act-final-rule/ [accessed 2024-10-08]
7. HHS. 21st century cures act: interoperability, information blocking, and the ONC health IT certification program. Federal Register 2020.
8. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2(9):e489-e492 [FREE Full text] [doi: 10.1016/S2589-7500(20)30186-2] [Medline: 32864600]
9. Oster ME, Shay DK, Su JR, Gee J, Creech CB, Broder KR, et al. Myocarditis cases reported after mRNA-based COVID-19 vaccination in the US from December 2020 to August 2021. J Am Med Assoc 2022;327(4):331-340 [FREE Full text] [doi: 10.1001/jama.2021.24110] [Medline: 35076665]
10. Buchan SA, Seo CY, Johnson C, Alley S, Kwong JC, Nasreen S, et al. Epidemiology of myocarditis and pericarditis following mRNA vaccination by vaccine product, schedule, and interdose interval among adolescents and adults in Ontario, Canada. JAMA Netw Open 2022;5(6):e2218505 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.18505] [Medline: 35749115]
11. CDC. Interim Clinical Considerations for Use of COVID-19 Vaccines in the United States. 2023. URL: https://www.cdc.gov/vaccines/covid-19/clinical-considerations/interim-considerations-us.html#extended-intervals-covid19-vaccine-does [accessed 2024-10-08]
12. ehealthExchange. Health Information Network. 2023. URL: https://ehealthexchange.org/ [accessed 2024-10-08]
13. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, et al. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 1981;30(2):239-245. [doi: 10.1038/clpt.1981.154] [Medline: 7249508]
14. Agresti A, Coull BA. Approximate is better than “Exact” for interval estimation of binomial proportions. Am Stat 1998;52(2):119-126. [doi: 10.1080/00031305.1998.10480550]
15. DasGupta A, Cai TT, Brown LD. Interval estimation for a binomial proportion. Statist Sci 2001;16(2):101-133.
16. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
17. Sexson Tejtel SK, Munoz FM, Al-Ammouri I, Savorgnan F, Guggilla RK, Khuri-Bulos N, et al. Myocarditis and pericarditis: case definition and guidelines for data collection, analysis, and presentation of immunization safety data. Vaccine 2022;40(10):1499-1511 [FREE Full text] [doi: 10.1016/j.vaccine.2021.11.074] [Medline: 35105494]
18. Rajamani S, Kayser A, Ruprecht A, Cassman J, Polzer M, Homan T, et al. Electronic case reporting (eCR) of COVID-19 to public health: implementation perspectives from the Minnesota Department of Health. J Am Med Inform Assoc 2022;29(11):1958-1966 [FREE Full text] [doi: 10.1093/jamia/ocac133] [Medline: 35904765]
19. ONC. Public Health Data Standardization - Deep Dive on USCDI+ and the Helios HL7® FHIR® Accelerator. 2022. URL: https://www.healthit.gov/sites/default/files/2022-10/ONC_Tech Forum PH Data Standardization_508.pdf [accessed 2024-10-08]
20. HL7. Clinical Quality Language (CQL). 2023. URL: https://cql.hl7.org/ [accessed 2024-10-08]
21. Cano ERMTSBHM. Chapter 21: Surveillance for Adverse Events Following Immunization Using the Vaccine Adverse Event Reporting System (VAERS). Atlanta, Georgia: CDC; 2021.

XSL•FO
RenderX

22. Ong MS, Klann JG, Lin KJ, Maron BA, Murphy SN, Natter MD, et al. Claims-based algorithms for identifying patients with pulmonary hypertension: a comparison of decision rules and machine-learning approaches. J Am Heart Assoc 2020;9(19):e016648 [FREE Full text] [doi: 10.1161/JAHA.120.016648] [Medline: 32990147]
23. Dang TK, Lan X, Weng J, Feng M. Federated learning for electronic health records. ACM Trans Intell Syst Technol 2022;13(5):1-17. [doi: 10.1145/3514500]
24. Schneider KL, Bell EJ, Zhou CK, Yang G, Lloyd P, Clarke TC, et al. Use of immunization information systems in ascertainment of COVID-19 vaccinations for claims-based vaccine safety and effectiveness studies. JAMA Netw Open 2023;6(5):e2313512 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.13512] [Medline: 37191962]

Abbreviations

AE: adverse event
AESI: adverse event of special interest
BEST: Biologics Effectiveness and Safety
CBER: Center for Biologics Evaluation and Research
CONSORT: Consolidated Standards of Reporting Trials
CQL: clinical quality language
EHR: electronic health record
eHX: eHealth Exchange
FDA: US Food and Drug Administration
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven, Inc
PPV: positive predictive value
RWD: real-world data
VAERS: vaccine adverse event reporting system

Edited by G Tsafnat; submitted 15.11.23; peer-reviewed by CK Zhou, L Zhang; comments to author 29.06.24; revised version received 23.08.24; accepted 21.09.24; published 25.11.24.

Please cite as:
Deady M, Duncan R, Sonesen M, Estiandan R, Stimpert K, Cho S, Beers J, Goodness B, Jones LD, Forshee R, Anderson SA, Ezzeldin H
A Computable Phenotype Algorithm for Postvaccination Myocarditis/Pericarditis Detection Using Real-World Data: Validation Study
J Med Internet Res 2024;26:e54597
URL: https://www.jmir.org/2024/1/e54597
doi:10.2196/54597
PMID:

©Matthew Deady, Raymond Duncan, Matthew Sonesen, Renier Estiandan, Kelly Stimpert, Sylvia Cho, Jeffrey Beers, Brian Goodness, Lance Daniel Jones, Richard Forshee, Steven A Anderson, Hussein Ezzeldin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Model Shifts of Machine Learning–Based Clinical Risk Prediction Models: Evaluation Study of Multiple Use Cases Across Different Hospitals

Patricia Cabanillas Silva[1*], MSc; Hong Sun[2,3*], PhD; Mohamed Rezk[1*], PhD; Diana M Roccaro-Waldmeyer[1], PhD; Janis Fliegenschmidt[4], BSc, MD; Nikolai Hulde[4], MD, PhD; Vera von Dossow[4], MD, PhD, Prof Dr; Laurent Meesseman[1], MD; Kristof Depraetere[1], MSc; Joerg Stieg[1], MSc; Ralph Szymanowsky[1], MSc; Fried-Michael Dahlweid[1], MD, PhD

[1]Dedalus HealthCare, Antwerp, Belgium

[2]Provincial Key Laboratory of Multimodal Perceiving and Intelligent Systems, Jiaxing University, Jiaxing, China

[3]Engineering Research Center of Intelligent Human Health Situation Awareness of Zhejiang Province, Jiaxing University, Jiaxing, China

[4]Institute of Anaesthesiology and Pain Therapy, Heart and Diabetes Centre North Rhine Westphalia, University Hospital of Ruhr-University Bochum, Bad Oeynhausen, Germany

[*]these authors contributed equally

Corresponding Author:
Mohamed Rezk, PhD
Dedalus HealthCare
Roderveldlaan 2
Antwerp, 2600
Belgium
Phone: 32 0784244010
Email: mohamed.rezk@dedalus.com

Abstract

Background: In recent years, machine learning (ML)–based models have been widely used in clinical domains to predict clinical risk events. However, in production, the performances of such models heavily rely on changes in the system and data. The dynamic nature of the system environment, characterized by continuous changes, has significant implications for prediction models, leading to performance degradation and reduced clinical efficacy. Thus, monitoring model shifts and evaluating their impact on prediction models are of utmost importance.

Objective: This study aimed to assess the impact of a model shift on ML-based prediction models by evaluating 3 different use cases—delirium, sepsis, and acute kidney injury (AKI)—from 2 hospitals (M and H) with different patient populations and investigate potential model deterioration during the COVID-19 pandemic period.

Methods: We trained prediction models using retrospective data from earlier years and examined the presence of a model shift using data from more recent years. We used the area under the receiver operating characteristic curve (AUROC) to evaluate model performance and analyzed the calibration curves over time. We also assessed the influence on clinical decisions by evaluating the alert rate, the rates of over- and underdiagnosis, and the decision curve.

Results: The 2 data sets used in this study contained 189,775 and 180,976 medical cases for hospitals M and H, respectively. Statistical analyses (*Z* test) revealed no significant difference ($P>.05$) between the AUROCs from the different years for all use cases and hospitals. For example, in hospital M, AKI did not show a significant difference between 2020 (AUROC=0.898) and 2021 (AUROC=0.907, Z=–1.171, $P=.242$). Similar results were observed in both hospitals and for all use cases (sepsis and delirium) when comparing all the different years. However, when evaluating the calibration curves at the 2 hospitals, model shifts were observed for the delirium and sepsis use cases but not for AKI. Additionally, to investigate the clinical utility of our models, we performed decision curve analysis (DCA) and compared the results across the different years. A pairwise nonparametric statistical comparison showed no differences in the net benefit at the probability thresholds of interest ($P>.05$). The comprehensive evaluations performed in this study ensured robust model performance of all the investigated models across the years. Moreover, neither performance deteriorations nor alert surges were observed during the COVID-19 pandemic period.

XSL•FO
RenderX

Conclusions: Clinical risk prediction models were affected by the dynamic and continuous evolution of clinical practices and workflows. The performance of the models evaluated in this study appeared stable when assessed using AUROCs, showing no significant variations over the years. Additional model shift investigations suggested that a calibration shift was present for certain use cases (delirium and sepsis). However, these changes did not have any impact on the clinical utility of the models based on DCA. Consequently, it is crucial to closely monitor data changes and detect possible model shifts, along with their potential influence on clinical decision-making.

(J Med Internet Res 2024;26:e51409) doi:10.2196/51409

KEYWORDS

model shift; model monitoring; prediction models; acute kidney injury; AKI; sepsis; delirium; decision curve analysis; DCA

Introduction

In recent years, machine learning (ML) algorithms for clinical risk predictions have emerged as a promising tool and have been widely used in health care applications [1-5]. However, most of such applications have been developed and evaluated on retrospective data [6,7], and their performance in live clinical settings remains understudied.

Wu et al [8] highlighted the importance of evaluating artificial intelligence (AI)–based medical devices with live clinical settings over different sites to address the shortcomings, such as overfitting to training data and bias against underrepresented subgroups. The evaluation of models' performances in live clinical settings is drawing more attention and is becoming more adopted in the field [9,10]. In our previous paper, we investigated clinical risk prediction models for heterogeneous patient populations from different hospitals and across different clinical use cases: delirium, sepsis, and acute kidney injury (AKI). By evaluating these models in both retrospective [7] and live clinical settings [11], we demonstrated that the diversity in patient populations translates into improved generalizability and a more comprehensive understanding [11]. A cohort study was also carried out to evaluate the clinical usefulness of our delirium prediction model [12]. These results led to a later review to conclude that our approach demonstrates the best clinical utility and highest performance when evaluated beyond the development setting [13].

However, even when evaluated in clinical settings, there remains a critical challenge that poses a significant threat to the reliability and performance of these models, which is the phenomenon of model shift [14-16]. Most ML-based prediction models use ML algorithms that leverage statistical methods to learn from clinical data. When clinical prediction models are deployed in nonstationary clinical environments [17], model deteriorations are observed over time in response to the dynamic nature of clinical environments [18]. The effectiveness and generalizability of ML-based clinical risk prediction models are greatly impacted when in such situations. As an example, during the COVID-19 pandemic, the alert rate of a sepsis prediction model was reported as more than double compared to that preceding the COVID-19 pandemic. As a consequence, the hospital had to shut down the prediction service [19,20].

This study intended to evaluate the model shift in ML-based clinical risk prediction models. Following the work of our previous study [7,11], we aimed to generalize our findings by applying the investigation to the delirium, sepsis, and AKI use cases at 1 community hospital and 1 specialized hospital. In this study, we used retrospective data to train different prediction models on clinical data from earlier years and then evaluated the model shift, as well as its impact on clinical decision-making, on data from later years, thus mimicking the situation of deploying prediction models in the real world. The evaluation covered periods preceding and during the COVID-19 pandemic, which allowed us to evaluate the impact of COVID-19. We investigated the impact of a model shift on model performance, model calibration, and clinical decisions. To the best of our knowledge, this is the first report that evaluated the impact of a model shift on such a scale.

Methods

Ethical Considerations

In accordance with Article 15.1 of the Professional Code for Healthcare Professionals of Germany, ethical approval is only required if human body materials are used or if the used data can be traced to a particular individual [21]. As no personal data were collected for this project and all used data were anonymized in compliance with the ethical and data protection obligations established by the European Commission [22], consultation or evaluation by an ethics committee was not required.

Study Design

We used a scalable method development approach to generate clinical risk prediction models for various conditions at 2 German hospitals, the Medius Klinik Nürtingen and the Herz- und Diabeteszentrum Nordrhein-Westfalen Bad Oeynhausen, referred to as hospitals M and H, respectively. Our previous investigations showed that our models' performance from live clinical workflows did not deteriorate compared to retrospective data. For instance, when comparing the area under the receiver operating characteristic curve (AUROC) on average across the 3 use cases and hospitals from live clinical workflows to that from retrospective data, the AUROC decreased slightly by 0.6 percentage points (from 94.8% to 94.2% at discharge) [7,11]. This study incorporated patient data from 2009 to 2021, which included demographic information, such as age and sex; clinical information, such as laboratory results; clinical notes, such as admission letters or nursing notes; vital signs; historical records; and medication. Tables S1 and S2 in Multimedia Appendix 1 provide an overview of data characteristics over the years for the 2 hospitals, respectively. We focused on 3 clinical use cases:

XSL•FO
RenderX

delirium, sepsis, and AKI. Table 1 displays the incidence of each condition at each hospital.

Table 1. Incidence for each use case in both hospitals.

Hospital	AKI[a] incidence (%)	Delirium incidence (%)	Sepsis incidence (%)
M	16.41	2.22	1.50
H	23.35	1.15	2.13

[a]AKI: acute kidney injury.

Analysis and Evaluation

We performed analyses and evaluations from different perspectives to systematically assess the model shift and its impact. Figure 1 provides an overview of these steps: First, we conducted a comprehensive analysis of the source data to detect any changes in them over time. Second, we evaluated the performance of the prediction model to determine whether there was any deterioration. Third, we examined any potential shift in the calibration curves. Lastly, we performed a decision analysis to determine whether there was a decline in clinical decision-making.

Figure 1. Overview of the investigation methodology.

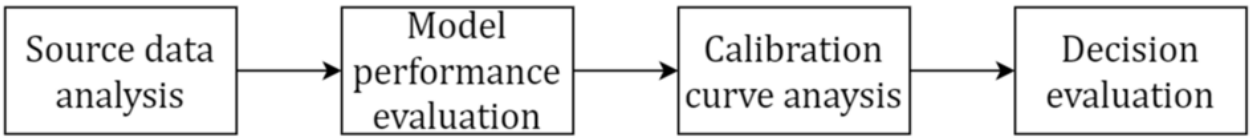

Data Source

As clinical prediction models may be negatively impacted by changes in clinical settings over time [18], one of the first steps in avoiding a model shift is to analyze the source data and look for changes in them. First, we looked at the number of samples per year, as this could drastically change. Second, it was important to track the incidence of the condition over the years as it could have been altered due to modifications in the coding system, the hospital protocol, or the appearance of new risk factors that cause an increase in incidence—for example, SARS-Cov2 as a risk factor for sepsis [23]. Third, it was necessary to closely examine data set characteristics. Among the key points were gender distribution over time, the availability of features throughout the years, and the length of observations and patient stay over the years.

Model Development

The prediction model design, data preparation, and training were implemented in an automated pipeline that provided a standardized method to install, configure, and execute the data preparation process, model training, and evaluation on a deployment-specific system. The automated pipeline started with source data analysis to detect abnormalities in the source data. It also automatically adapted the hyperparameters for the model training process. Risk prediction models were then generated based on the hospital data with such an automated pipeline and, therefore, calibrated with the electronic health record (EHR) data of the hospital. The generated models were automatically evaluated, and acceptance criteria were verified throughout the evaluation procedure.

Before data preparation, we applied the following inclusion criteria: (1) patients should be 18 years or older and must have a birth year available; (2) medical cases should contain information about the gender, the department, and the reason for admission; and (3) stay in the hospital should be for a maximum of 90 days. Next, input features were generated using a common data preparation pipeline, which used all the available data generated during the patient's stay. The features came from structured data, such as laboratory results, vital signs, and historical records, along with unstructured data, such as clinical notes. Details about the number of features used by the model for both hospitals are provided in Table S3 in Multimedia Appendix 1.

For preparing unstructured features, we developed a natural language processing (NLP) pipeline in 2 steps. The initial step involved using a trained TinyBERT model to perform named entity recognition (NER) on clinical text data, identifying disorders and clinical findings. The extracted entities were then passed as an input to a named entity normalization (NEN) model [24]. The NEN model maps the extracted entities to the *German Modification of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* (ICD-10-GM) codes. This 2-step process enhances the precision of information extraction and ensures a consistent representation of clinical concepts. (For more details, see the *NLP Models* section in Multimedia Appendix 2). The labels for sepsis and delirium were assigned according to the ICD-10-GM codes at discharge. The AKI labels were assigned based on the Kidney Disease Improving Global Outcomes (KDIGO) criteria [25].

To ensure the reliability of the study, the data sets were strictly split into training and evaluation sets at the patient level. Additionally, approximately 10% of the training set was exclusively reserved for validation, calibration, and threshold selection. The training set for hospital H included data from 2017 to 2018, while that for hospital M comprised data from 2009 to 2017. Consequently, the evaluation set contained the most recent years, spanning from 2019 for hospital H and from 2018 for hospital M. The choice of these specific years was due to differences in incidence across the years because of

XSL•FO
RenderX

mislabeling (for more details, see data source analysis in the *Results* section).

All the available information belonging to the same hospital stay was first aggregated to obtain a complete training record. Additionally, to mitigate the situation where the model was requested to generate predictions with limited information available, we applied data augmentation in the training set to generate a partial record. The partial data were constructed based on subsets of features of the original record and, combined with the complete record, enhanced the robustness of the clinical risk prediction model. Nevertheless, it is important to note that this approach was applied only to the training data, while the test set remained unmodified to reflect the real-life distribution of the population. The impact of this data augmentation was evaluated in our previous work [7], where the performance 24 hours/1 day/the next day after admission showed an improvement of around 4% in the AUROC. Furthermore, to avoid potential biases due to high class imbalance, the controls were downsampled to a 1:5 case:control ratio before training. More detailed information about the model development approach can be found in our previous paper [7].

Next, a clinical risk prediction model was trained separately for each condition (sepsis, delirium, and AKI) and each hospital (M and H), applying a standard model training methodology using/on transformer models [26]. As a result, we obtained 6 independent models. For the model training process, the previously mentioned features were used as inputs and the labels as targets of the model. However, it is important to note that certain features exhibited unreasonably strong correlations with the condition. Consequently, these features, identified as leaking features, were removed from the training data. Roughly, 2 categories of leaking features were considered: reverse causal correlations, which correspond to features that typically occur after onset (or clinical suspicion), and strong cofounders, such as the presence of specific but unrelated lab tests only taken in a high-risk population. The list of these features, defined by in-house clinical professionals, can be found in Table S4 in Multimedia Appendix 1.

Model Performance

A common pitfall of clinical prediction models is that their performance can deteriorate over time. Therefore, a key aspect to explore is the ability of the model to discriminate and sustain its performance over time. There are different metrics that can be used (eg, accuracy, sensitivity, specificity, precision) to evaluate a model's performance. In this study, we applied 2 methods to investigate the performance of a model: first, the AUROC, as it is an assessment metric that remains unaffected by incidence or threshold selection, and second, the calibration shift, which is the disagreement between the predicted probability and the observed number of events.

To detect a calibration shift, a well-calibrated model that makes probabilistic predictions that match real-world probabilities is required. There are several methodologies for calibrating a model, such as Platt scaling [27], isotonic regression [28], and temperature scaling (which is a single-parameter variant of Platt scaling) [29], among others. In this study, we used Platt scaling, which is more suitable for small data sets (like our sepsis and delirium data sets). Additionally, for plotting the calibration graph, we used a quantile binning approach, where each bin has the same number of samples, which reduces the noise in the curve by preventing different bins from having an unbalanced number of samples. There is no 1 unique way to measure the calibration of predictive models. Although some methods are frequently used and have specific strengths, all have limitations [30]. In this analysis, we focused on 2 metrics: expected calibration error (ECE) and maximum calibration error (MCE). The ECE measures the weighted mean absolute difference between confidence and accuracy, while the MCE measures the maximum difference between average confidence and accuracy across bins. In both metrics, larger values indicate greater miscalibration.

We generated calibration curves for each year within the test set , which allowed us to more easily detect a calibration shift by identifying deviations between the different curves. We observed calibration shifts for sepsis and delirium but not for AKI. Consequently, we expanded our experiments to assess whether the shift was related to the intrinsic nature of different conditions and also whether it was related to other factors, such as data size or incidence. Although other factors can contribute to the presence or absence of a shift, we investigated how the former 2 impacted the results in AKI, as no evidence of a shift was observed for this condition. Furthermore, the AKI use case exhibited a larger sample size and a higher incidence rate compared to the other 2 use cases. This provided us with an opportunity to conduct the following experiments: First, we randomly downsampled the AKI data set to assess the shift on different data scales. Second, we artificially reduced the incidence rate of AKI in the testing and calibration data sets to match the delirium incidence rate. These experiments aimed to assess the influence of both data size and incidence rate on the outcomes and whether the observed shift in sepsis and delirium is a consequence of a lower data size or incidence.

Decision Evaluation

The system calculated predictions throughout the hospitalization, and when a threshold was reached, an alert was shown. An alert may appear as soon as the patient is registered in the system if their past medical history is predictive enough. We conceptually created 5 categories (instead of the typical 4):

- True negative (TN; control): an alert is not shown correctly.
- False positive (FP; control): an alert is incorrectly shown at some point during the stay.
- False negative (FN; case): an alert is incorrectly not shown at all, although the patient has developed the condition at some point during the stay.
- True positive (TP; case): an alert is correctly shown before onset.
- Arguably FN/TP (case): an alert is correctly shown but not strictly before onset. The risk prediction is then diagnostic rather than strictly predictive, similar to other risk scores, such as the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) or Delirium Observation Scaling (DOS).

A severe limitation of ICD codes is that they convey no information about the time of onset of a condition. We were,

XSL•FO
RenderX

therefore, unable to distinguish between the last 2 groups, so we evaluated the occurrence of an alert at any time from admission until discharge (as opposed to no alert at all) and compared this alert rate to the finally assigned diagnosis.

Model performance evaluation normally focuses on discrimination metrics; however, it does not assess whether the risk model enhances clinical decision-making. Accordingly, we undertook an investigation to determine the potential impact on decision-making over time by analyzing alert rates and over- and underdiagnosis rates and performing decision curve analysis (DCA). First, we investigated whether there was any increase or decrease in alert rates across the years. Second, we examined the changes in over- and underdiagnosis rates. Overdiagnosis refers to overtreatment without any discernible benefit to the patient, which typically occurs in sensitive models, and is defined as the false-positive rate (FPR). Underdiagnosis, in contrast, refers to a failure to correctly diagnose a condition and is defined as the false-negative rate (FNR) [31,32]. The FPR and FNR are affected by threshold selection. Thus, in the pursuit of achieving an equilibrium between these 2 metrics, our threshold selection approach consisted of selecting a threshold that minimized the difference between under- and overdiagnosis rates. This threshold selection was performed in an independent set (10% of the training set) and after calibrating the model. Lastly, we performed DCA, where we assessed whether the benefit of using our model to make a clinical decision decreased over time. DCA allows us to evaluate and compare the clinical utility of predictive models using nonparametric statistical methods [33,34]. This method illustrates the net benefit related to a specific decision strategy across a spectrum of probabilities. Its application ensures that clinical decisions are not solely reliant on discriminative metrics but also consider the potential harms and benefits associated with such decisions. The net benefit is calculated according to the following formula:

where N is the total number of predictions and is the threshold probability.

Results

Data Source Analysis

The data sets used in this study contained 189,775 and 180,976 medical cases for hospitals M and H, respectively. The distribution and characteristics of the data sets for both hospitals are shown in Tables S1 and S2 in Multimedia Appendix 1. Hospital M is a community hospital, while hospital H is a specialized center for diabetes and heart disease, with a high volume of cardiac surgery and invasive, transcatheter procedures and intensive care. The 2 data sets exhibit common characteristics with respect to an increasing trend in the length of stay, resulting in a greater number of features and longer observations, and consistency in the proportion of women and men throughout the years (Tables S1 and S2 in Multimedia Appendix 1). However, it is important to note that there are some differences between the 2 data sets or different use cases. For instance, medication was not available as a feature until 2016 in hospital M and until 2018 in hospital H. Moreover, the incidence of sepsis and delirium in earlier years was lower than in later years, with a more notable disparity observed in delirium at hospital H. The increase in the delirium incidence rate is coincident with extensive efforts to improve the recognition and treatment of delirium. Based on these findings, we decided to use only data from the most recent years, specifically from 2017 for hospital H. This decision was made to mitigate the potential influence of mislabeling arising from the data source. Medical cases whose admission date was earlier than 2017 were removed. The real-life incidence for the validation and evaluation test was therefore not affected. As aforementioned, the case:control balance for the training data set was set to 1:5.

Model Performance

The performance of the clinical risk prediction model was evaluated at the end of the patient's hospital stay due to the lack of the time of onset and to avoid inconsistencies when the predictions were evaluated.

Performance Evaluation

Model evaluation results are presented in Tables 2 and 3 by showing AUROC and area under the precision recall curve (AUPRC) values. The AUPRC is a relevant metric when dealing with imbalanced data and detecting positive samples are of greater interest and concern. However, as the AUPRC is highly influenced by incidence, the model was evaluated based on the AUROC, which, in contrast to other metrics, such as sensitivity, specificity, precision, and the AUPRC, is independent of the threshold and incidence. Thus, it allowed us to compare model performance across different years. Table 2 shows that AUROC values remained stable for the 2 use cases at hospital M, with mean 90.37% (SD 0.32%), mean 96.95% (SD 0.50%), and mean 95.92% (SD 0.41%) for AKI, delirium, and sepsis, respectively. Similarly, Table 3 presents the values for hospital H, and the values for AKI, delirium, and sepsis were mean 91.25% (SD 0.29%), mean 94.24% (SD 0.44%), and mean 97.95% (SD 0.41%), respectively.

XSL•FO
RenderX

Table 2. Model performance and incidence in the test set of hospital M.

Use case and metrics	Years			
	2018	2019	2020	2021
AKI[a]				
Incidence (%)	14.57	14.81	14.95	17.64
Alert rate (%)	22.36	22.62	23.48	26.56
AUROC[b] (%)	90.54	90.40	89.84	90.69
AUPRC[c] (%)	65.48	65.87	64.99	71.01
Delirium				
Incidence (%)	2.18	1.92	3.17	2.63
Alert rate (%)	10.75	11.61	13.44	15.37
AUROC (%)	97.55	96.86	97.21	96.18
AUPRC (%)	42.93	34.94	51.26	40.89
Sepsis				
Incidence (%)	1.67	1.49	1.24	1.66
Alert rate (%)	11.86	13.12	14.65	17.17
AUROC (%)	96.29	96.37	95.45	95.55
AUPRC (%)	40.98	32.85	34.14	39.81

[a]AKI: acute kidney injury.
[b]AUROC: area under the receiver operating characteristic curve.
[c]AUPRC: area under the precision recall curve.

Table 3. Model performance and incidence in the test set of hospital H.

Use case and metrics	Years		
	2019	2020	2021
AKI[a]			
Incidence (%)	25.22	25.98	25.58
Alert rate (%)	32.70	34.36	32.76
AUROC[b] (%)	90.83	91.51	91.40
AUPRC[c] (%)	80.58	82.59	81.71
Delirium			
Incidence (%)	3.25	3.26	2.86
Alert rate (%)	20.56	21.25	20.46
AUROC (%)	94.70	94.39	93.64
AUPRC (%)	37.32	34.43	30.91
Sepsis			
Incidence (%)	3.70	3.86	3.89
Alert rate (%)	12.82	12.47	12.79
AUROC (%)	97.37	98.12	98.35
AUPRC (%)	73.63	77.04	77.78

[a]AKI: acute kidney injury.
[b]AUROC: area under the receiver operating characteristic curve.
[c]AUPRC: area under the precision recall curve.

XSL•FO
RenderX

These findings indicate low variations in the performance metrics; moreover, statistical tests were conducted to discern whether the variations in performance metrics over the years were statistically significant. First, we computed the SEs of the AUROCs using Hanley and McNeil's method [35]. Next, we performed a Z test to assess whether there were significant differences between different AUROCs. We did not observe any significant differences between the AUROCs across the different years for all use cases and all hospitals (P>.05). Detailed results are presented in Tables S5 and S6 in Multimedia Appendix 1. These results demonstrate that the model's performance did not exhibit persistent deterioration over time at either hospital.

Calibration Curve Analysis

In addition to the evaluation of the AUROC, which reflects model performance, we also investigated whether there was any potential shift in model calibration. Figure 2 shows the calibration curves, as well as their corresponding ECE and MCE values. Figure 2a displays the calibration curves of the 2 use cases at hospitals M and H, respectively. The curves before calibration indicated that the averaged predicted risks were higher than the observed event rate, which means our models overestimated the risk. Consequently, we calibrated our models by applying the Platt scaling method [27]. The curves after calibration showed that the predicted risks better matched the observed positive ratio.

Figure 2. Calibration curves and corresponding ECE and MCE values: (a) calibration curves for each use case and year before and after calibration, with 95% CIs; (b) ECE values across the years for each use case; and (c) MCE values over the years for the 3 use cases. AKI: acute kidney injury; ECE: expected calibration error; MCE: maximum calibration error.

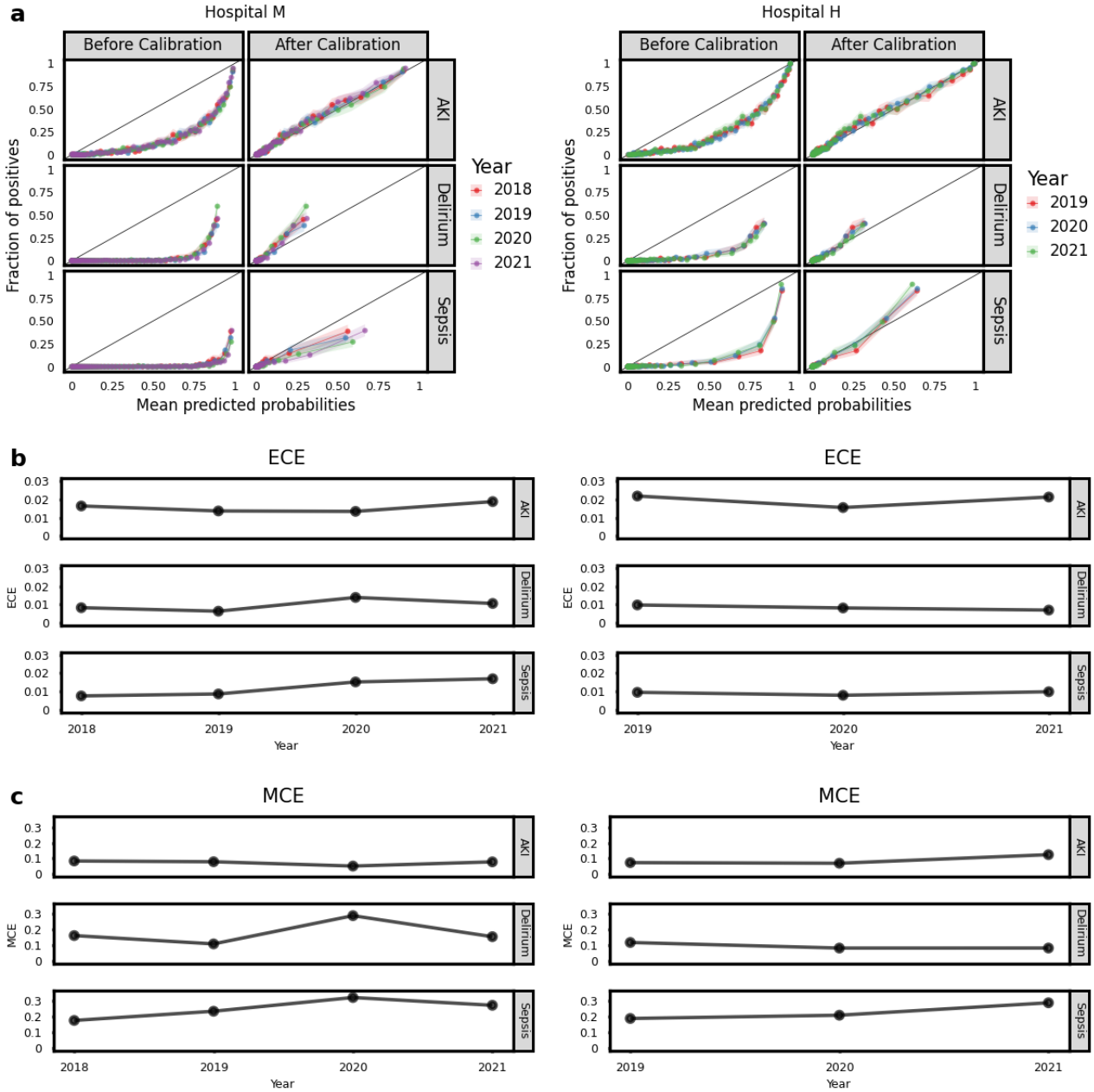

XSL•FO
RenderX

To better quantify the quality of the calibration curves, corresponding ECE (Figure 2b) and MCE (Figure 2c) values were generated for each calibrated curve. The error graphs did not show a clear increase in the shift across the evaluated years for any of the use cases at both hospitals, although small fluctuations were observed in the ECE and MCE curves in the delirium use case at hospital M. Furthermore, Figure 2a demonstrates that unlike the AKI use case, the sepsis and delirium use cases exhibited a calibration shift at hospital M, with the curves deviating from the diagonal line for higher probabilities. At hospital H, there was no evident calibration shift for both AKI and delirium use cases, although a slight shift was observed in its sepsis model.

Given the observed calibration shifts in sepsis and delirium, in contrast to the stability in AKI, we extended our investigations to examine whether the observed shift was intricately associated with inherent characteristics of the distinct conditions or whether it was influenced by other factors, such as variation in data size or incidence rates. Our results demonstrated that reducing the data size or incidence rates did not result in a calibration shift, and therefore, the observed calibration shift in sepsis and delirium could be use case specific or influenced by other factors (see Multimedia Appendix 2 for more details).

Decision Evaluation

The last evaluation focused on decision-making analysis. We first assessed the alert rates, that is, the number of alerts triggered when the predicted probability exceeds the predefined threshold. We observed a slight increase over the years for all 3 use cases, even though the AUROC and the incidence remained similar at hospital M (see Table 2). However, we did not observe a similar increase at hospital H (see Table 3), where both incidence and alert rates remained stable.

The over- and underdiagnosis rates were further evaluated based on the correctness of an alert. The thresholds selected from the threshold selection set (10% of the training set) at hospital M were 19.97, 2.40, and 2.67 for AKI, delirium, and sepsis, respectively, and at hospital H were 23.13, 2.79, and 2.93, respectively. Figure 3 shows the under- and overdiagnosis rates for the test sets from the 3 use cases on a yearly basis at both hospital M and hospital H. At hospital M, we also observed that although the underdiagnosis rate remained stable over the years, there were clear trends of an increase in the overdiagnosis rate in all 3 use cases. Such an increase in overdiagnosis is in line with the increase in the alert rate and is likely associated with the increased number of observations in patient records (Tables S1 and S2 in Multimedia Appendix 1), as observed in the data analysis. The maximum deviation of the under- and overdiagnosis curves was around 5% in overdiagnosis on sepsis observed at hospital M, which aligns with the fact that the alert rate also increased. Such an increase is partially in line with the report that the alert rate of sepsis prediction models increased during the COVID-19 pandemic [19,20]. However, in this investigation, the trend of such an increase started before the pandemic.

At hospital H, there were more fluctuations in the underdiagnosis rate. The 3 use cases behaved differently, hindering generalization. Although the number of observations in patient records also increased yearly at hospital H, in contrast to hospital M, its overdiagnosis rates remained stable for all 3 use cases. Furthermore, the underdiagnosis rate of sepsis decreased during the COVID-19 pandemic, while its overdiagnosis rate remained stable. This indicates that the performance of the sepsis prediction model at hospital H improved during the COVID-19 pandemic, which contradicts the findings of hospital M, as well as other reports [19,20].

In summary, the 2 hospitals behave differently in terms of over- and underdiagnosis rates. Therefore, site-specific monitoring schemes are required to detect such changes. Consequently, given the difficulty in determining whether a model shift occurred based solely on the under- and overdiagnosis rates, we additionally performed DCA to test for any changes in the clinical utility of the models.

To examine the potential impact of the model on decision-making over time, DCA was performed for each distinct year and use case. The net benefit was calculated across the full range of probability thresholds. However, the lower range of the probability thresholds is more relevant to clinicians as, arguably, most clinicians would agree that the medical and financial consequences of delirium, sepsis, and AKI outweigh the nuisance to the user and the cost and inconvenience of preventive measures and closer monitoring. It is necessary to find the optimal threshold that has a good balance between FPs and FNs to prevent the excessive generation of unnecessary alerts, which can lead to user desensitization and decreased responsiveness. Additionally, it is important to highlight that the initial position of the decision curves (threshold=0) is notably affected by the incidence rate of the particular use case. Therefore, to ensure that the differences between the curves did not stem from different incidence rates, we standardized the incidence rate across the years by augmenting the incidence of the years that had lower rates until a uniform incidence was achieved across all years.

Figure S1 in Multimedia Appendix 2 and Figure 4 show the decision curves for hospitals M and H, respectively, for each year across the 2 use cases. To test whether there was a significant difference in the decision-making between the years, we performed a pairwise nonparametric statistical comparison between the decision curves from 2 individual years for all possible year pairs [33,34]. For all the use cases and hospitals, we did not find any significant difference between any of the years (all $P>.05$) at the thresholds of interest. These results demonstrated that the model did not experience a degradation in clinical utility across the years.

XSL•FO
RenderX

Figure 3. Under- and overdiagnosis rates across years in hospitals M and H. AKI: acute kidney injury.

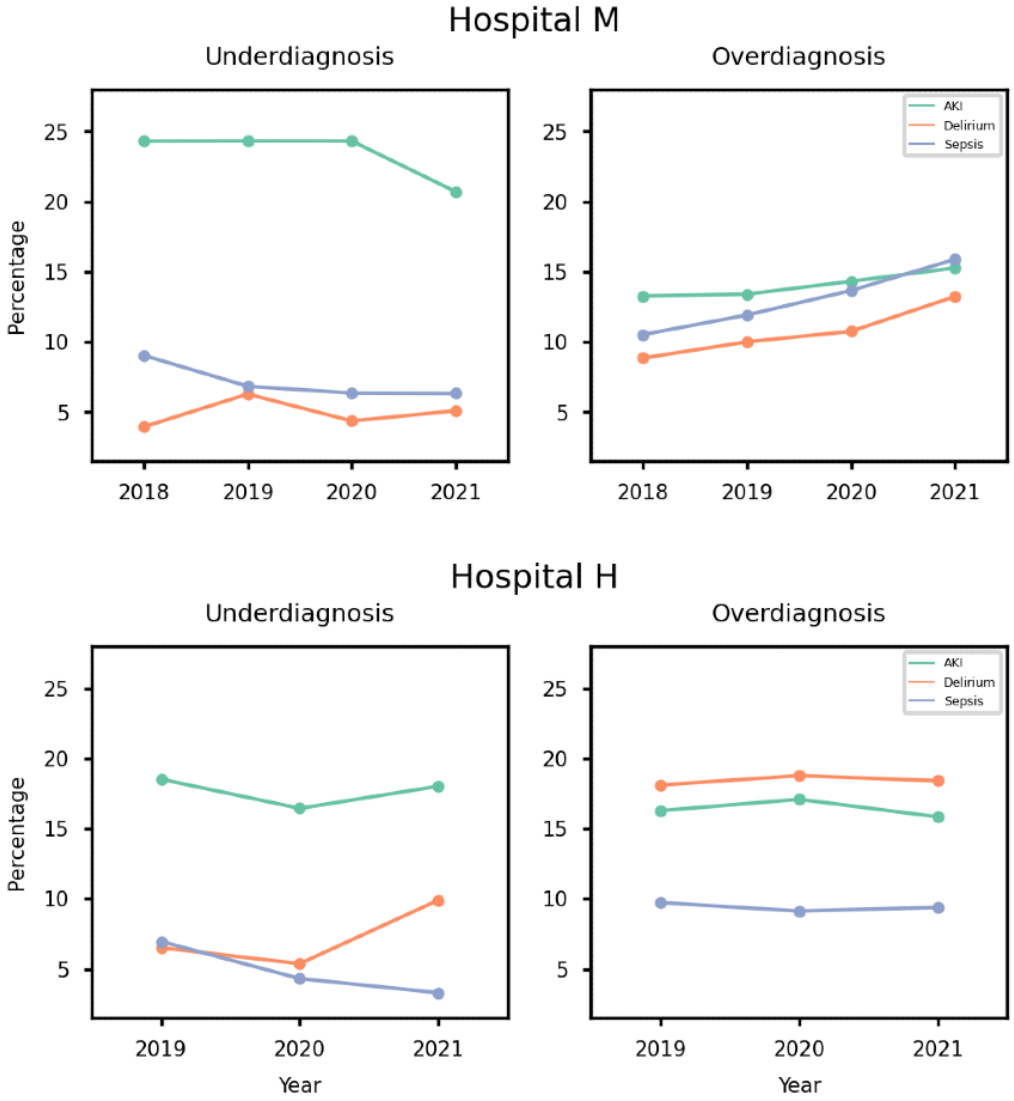

Figure 4. Decision curves for hospital M. Colored dotted lines show the net benefit of alerting all patients, gray lines represent the benefit of not alerting any patient, and colored solid lines show the net benefit of using the model on each of the years in the test. AKI: acute kidney injury.

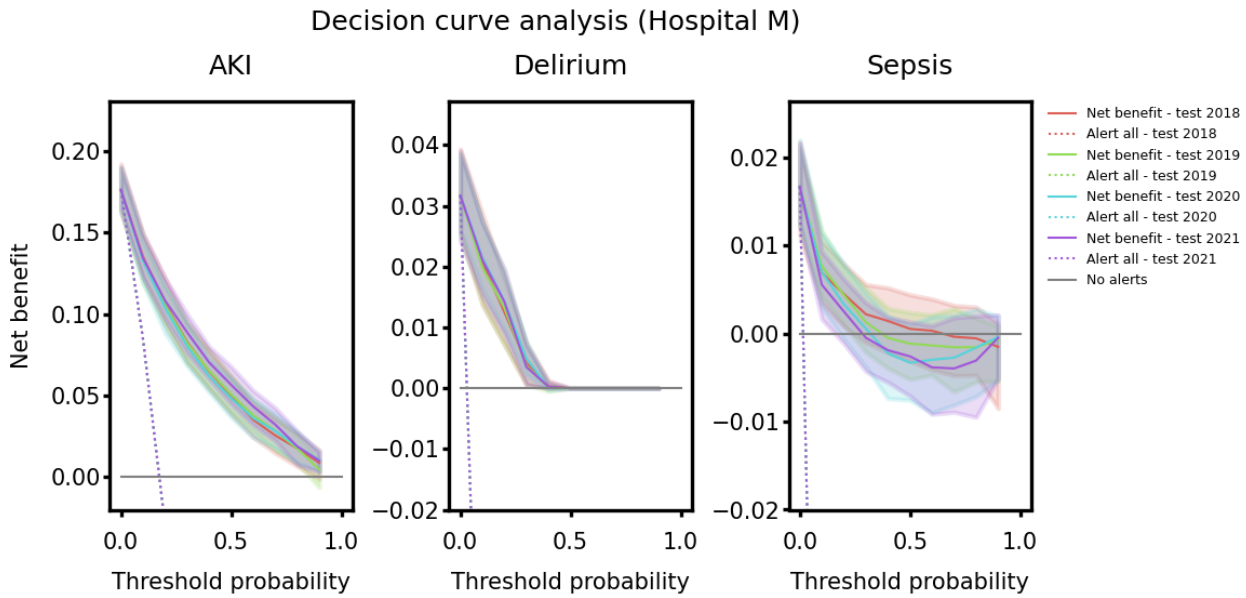

XSL•FO
RenderX

Discussion

Principal Findings

This paper evaluated the impact of a model shift on ML-based clinical risk prediction models. There was no significant model shift in the period under this investigation. The models' performances remained stable for all the 3 use cases, and there were no significant differences in the AUROCs between the years. Nevertheless, analysis of the calibration curves revealed a calibration shift in some use cases. The calibration curves of AKI were more stable compared to those for the delirium and sepsis prediction models at both hospitals in this study. No calibration shift was observed for AKI, even when the size of the sample and the rate of incidence were downscaled. These results suggest an intrinsic difference among the different use cases. Lastly, with respect to the interpretation of the predictions, the rate of alerts and overdiagnosis behaved differently across different hospitals: they consistently increased in the 3 use cases at hospital M and remained stable at hospital H. Furthermore, the examination of DCA results revealed a consistent pattern where the models exhibited no deterioration over time, affirming the stability of their clinical utility and decision-making.

These findings suggest that relying solely on AUROCs, the commonly used metric for assessing the performance of clinical risk prediction models, is inadequate in capturing the impact of a model shift. It is, therefore, essential to use comprehensive evaluations, such as calibration curve analysis, alert rate monitoring, and DCA, to promptly detect any kind of model shift. Additionally, recalibrating the model when the calibration curve significantly deviates from the diagonal may offer a viable solution without necessitating complete model retraining. This approach is effective because the chosen probability calibration approach does not affect the evaluation metrics as AUROCs or AUPRCs. However, in other situations where the performance or clinical usefulness of the model has deteriorated, retraining the model would be necessary. Furthermore, this study revealed that the potential impact of a model shift cannot be easily generalized, as the evaluations exhibited diverse behaviors across different use cases and hospitals. Therefore, implementing proper model calibration is crucial to avoid undesirable consequences, such as a significant increase in the alert rate [19,20].

Comparison With Prior Works

Previous research that investigated model shifts in the clinical domain often worked with a single use case at a single site. For instance, Davis et al [14] examined data shift scenarios in AKI risk prediction with 1 data set. Similarly, Minne et al [15] used a single data set to evaluate the effect of temporal changes on mortality prediction performance. Therefore, later studies have argued that the focus on developing new models in the clinical domain should introduce more heterogeneity in model development and model evaluation [16]. The study presented in this paper leveraged 2 large data sets from different hospitals and investigated 3 distinct use cases, thereby enhancing the generalizability of the findings. Furthermore, although previous studies [14,15] have focused on analyzing model drift in terms of performance, this study extended the evaluation to a broader set of metrics, assessing the evolution of clinical utility of the model over time. This approach significantly enhances the heterogeneity and robustness of the evaluation.

Motivations

ML-based prediction models are often closely tied to the system environment in which the models undergo training. After the models are deployed as real-world applications, their performance may deteriorate over time when the system experiences changes in its environment. The consequences of such changes largely impact the reliability of ML-based prediction models. Although the dynamic nature of the system is inevitable, timely detection of its impact is, therefore, of utmost importance to prevent model deterioration when it is used in clinical settings. This study aimed to investigate various approaches for monitoring the implications of model shifts by examining diverse use cases across different types of hospitals.

Strengths

This study performed comprehensive evaluations of model behaviors over a 4-year period. The evaluation was not limited to model performance, also assessing the consequence of the model shift on clinical decision-making. Such detailed evaluations were performed on 3 different use cases at both a specialized hospital and a generalized hospital so as to better generalize the findings of our study. To the best of our knowledge, this study is the first report to investigate the consequences of a model shift on such a comprehensive level and scale.

Limitations

This study has some limitations. First, the study was performed on retrospective data to simulate how prediction models would behave over time after they were trained. Evaluating model behavior in real-world clinical settings would provide more valuable insights. Second, the cause of the calibration shift observed on the delirium and sepsis use cases was not thoroughly analyzed in this study. We hypothesized that such a difference may be related to the different labeling strategies: the delirium and sepsis use cases were labeled with discharge ICD codes, while the AKI use case was labeled with the KDIGO algorithm. The use of ICD codes is imperfect owing to inevitable variations and inconsistencies in coding practice. However, lacking standard criteria for diagnosis, the use of such codes is an established method for retrospective identification [36]. Discharge ICD codes are likely more vulnerable to changes in coding policies, while diagnostic algorithms, such as KDIGO, that rely on lab values are less likely to change over time.

Future Directions

This study evaluated the impact of a model shift on ML-based clinical risk prediction models. As a future step, we plan to conduct the discussed evaluations on our prediction models after they are deployed in live EHR systems. In addition, we will further investigate the different behaviors of calibration curves between the AKI use case and the other 2 use cases. Specifically, we aim to explore the influence of various labeling policies on these behaviors.

XSL•FO
RenderX

Including retrospective data from earlier years may help build a more reliable model with more training samples; however, it may also increase the risk of data shift. We will further investigate how to make a good trade-off when selecting training data, thus significantly influencing the development of a good prediction model, while minimizing the model shift.

Conclusion

This study performed evaluations across different hospitals to investigate the impact of model shifts on various ML-based risk prediction models. Although the metrics of AUROCs remained stable for the evaluated use cases over the years, the analysis of calibration curves varied across certain use cases (ie, delirium and sepsis) at different hospitals. However, upon conducting more in-depth investigations using DCA, it was confirmed that these variations did not have any impact on the clinical utility of the models, as there was no evident degradation in their decision-making capabilities. Moreover, the COVID-19 pandemic has the potential to alter clinical diagnosis and treatment patterns over the long term. So, to obtain a more complete understanding of this impact in prediction models, a future study incorporating additional years of data would be highly valuable. As a summary, these findings suggest the need for appropriate monitoring schemes and timely model calibrations following the deployment of prediction models.

Acknowledgments

The authors would like to acknowledge the Herz- und Diabeteszentrum Nordrhein-Westfalen Bad Oeynhausen and the Medius Kliniken Nürtingen for assisting with model training, calibration, and evaluation with their retrospective data. The authors would also like to thank Martijn Vanbiervliet, Jos De Baerdemaeker, Patrick Deutschmann, and Harsh Grover for their support in preparing the software environment for this study.

JF, NH, and VvD received funding from the ARGUS project (grant 100126059) from the Ruhr-Universität Bochum. Authors from Dedalus received funding from the European Union's Horizon 2020 projects GenoMed4All (Genomics and Personalized Medicine for All Though Artificial Intelligence in Haematological Diseases; grant 101017549) and PERSIST (Patients-Centered Survivorship Care Plan After Cancer Treatments Based on Big Data and Artificial Intelligence Technologies; grant 875406).

Data Availability

Due to commercial restrictions, legal obligations, and data protection, data/code availability is not applicable to this paper.

Authors' Contributions

PCS, HS, and MR conceptualized the study and designed the evaluation methods. DMR provided input to consolidate the study. HS, MR, and PCS performed model preparation and performance evaluation. PCS performed source data analysis. PCS and MR performed statistical curve calibration and analysis. LM, JF, NH, and VvD provided evaluations from the clinical perspective. RS, JS, KD, and MD coordinated resources to perform the evaluation. HS and PCS wrote the original draft, MR prepared the results, HS made summarizations, and DMR and JF proofread the paper. JF, MD, and LM provided input as medical editors. All authors reviewed and edited the manuscript critically and have approved its final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data characteristics in hospitals H and M for each year, number of features used in hospitals H and M, leaking features, and *Z* test results for hospitals H and M.
[DOCX File , 27 KB - jmir_v26i1e51409_app1.docx]

Multimedia Appendix 2
Decision and calibration curve analysis results.
[DOCX File , 1309 KB - jmir_v26i1e51409_app2.docx]

References

1. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, MI in Healthcare Workshop Working Group. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med 2020 Mar 26;3(1):47 [FREE Full text] [doi: 10.1038/s41746-020-0254-2] [Medline: 32258429]
2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]

XSL•FO
RenderX

3. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
6. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1(1):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
7. Sun H, Depraetere K, Meesseman L, De Roo J, Vanbiervliet M, De Baerdemaeker J, et al. A scalable approach for developing clinical risk prediction applications in different hospitals. J Biomed Inform 2021 Jun;118:103783 [FREE Full text] [doi: 10.1016/j.jbi.2021.103783] [Medline: 33887456]
8. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021 Apr;27(4):582-584. [doi: 10.1038/s41591-021-01312-x] [Medline: 33820998]
9. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022 Jul 21;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
10. Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. J Am Med Inform Assoc 2020 Jul 01;27(9):1383-1392 [FREE Full text] [doi: 10.1093/jamia/ocaa113] [Medline: 32968811]
11. Sun H, Depraetere K, Meesseman L, Cabanillas Silva P, Szymanowsky R, Fliegenschmidt J, et al. Machine learning–based prediction models for different clinical risks in different hospitals: evaluation of live performance. J Med Internet Res 2022 Jun 07;24(6):e34295 [FREE Full text] [doi: 10.2196/34295] [Medline: 35502887]
12. Fliegenschmidt J, Hulde N, Preising M, Ruggeri S, Szymanowsky R, Meesseman L, et al. Leveraging artificial intelligence for the management of postoperative delirium following cardiac surgery. Eur J Anaesthesiol Intensive Care 2022 Dec 8;1:1-10. [doi: 10.1097/EA9.0000000000000010]
13. Strating T, Shafiee Hanjani L, Tornvall I, Hubbard R, Scott I. Navigating the machine learning pipeline: a scoping review of inpatient delirium prediction models. BMJ Health Care Inform 2023 Jul;30(1):1-8 [FREE Full text] [doi: 10.1136/bmjhci-2023-100767] [Medline: 37407226]
14. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017 Nov 01;24(6):1052-1061 [FREE Full text] [doi: 10.1093/jamia/ocx030] [Medline: 28379439]
15. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. Intensive Care Med 2012 Jan;38(1):40-46 [FREE Full text] [doi: 10.1007/s00134-011-2390-2] [Medline: 22042520]
16. van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023 Feb 24;21(1):70 [FREE Full text] [doi: 10.1186/s12916-023-02779-w] [Medline: 36829188]
17. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008 Jan;61(1):76-86. [doi: 10.1016/j.jclinepi.2007.04.018] [Medline: 18083464]
18. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. J Biomed Inform 2020 Dec;112:103611 [FREE Full text] [doi: 10.1016/j.jbi.2020.103611] [Medline: 33157313]
19. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
20. Wong A, Cao J, Lyons PG, Dutta S, Major VJ, Ötles E, et al. Quantification of sepsis model alerts in 24 US Hospitals before and during the COVID-19 pandemic. JAMA Netw Open 2021 Nov 01;4(11):e2135286 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.35286] [Medline: 34797372]
21. (Model) professional code* for physicians in Germany -MBO-Ä 1997 - **The resolutions of the 124st German Medical Assembly 2021 in Berlin. German Medical Association. 2021. URL: https://www.bundesaerztekammer.de/fileadmin/user_upload/BAEK/Themen/Internationales/Model-Professional_Code_for_Physicians_124_DAET_210505.pdf [accessed 2024-03-12]
22. Ethics and data protection. European Commission. 2021. URL: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf [accessed 2024-12-03]
23. da Silva Ramos FJ, de Freitas FGR, Machado FR. Sepsis in patients hospitalized with coronavirus disease 2019: how often and how severe? Curr Opin Crit Care 2021 Oct 01;27(5):474-479 [FREE Full text] [doi: 10.1097/MCC.0000000000000861] [Medline: 34292175]

XSL•FO
RenderX

24. Sung M, Jeon H, Lee J, Kang J. Biomedical entity representations with synonym marginalization. arXiv:200500239 [cs] [FREE Full text] Preprint posted online 2020 [doi: 10.18653/v1/2020.acl-main.335]
25. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract 2012 Aug 7;120(4):c179-c184 [FREE Full text] [doi: 10.1159/000339789] [Medline: 22890468]
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: NIPS 2017: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA.
27. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. Cambridge, MA: MIT Press; 1999:61-74.
28. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. USA: Association for Computing Machinery; 2002 Presented at: Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 23-26, 2002; Edmonton, Alberta. [doi: 10.1145/775047.775151]
29. Guo C, Pleiss G, Sun Y, Weinberger K. On calibration of modern neural networks. 2017 Presented at: ICML'17: 34th International Conference on Machine Learning; August 6-11, 2017; Sydney, NSW p. 1321-1330 URL: https://proceedings.mlr.press/v70/guo17a.html
30. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020 Apr 01;27(4):621-633 [FREE Full text] [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
31. Li Y, Wang H, Luo Y. Improving fairness in the prediction of heart failure length of stay and mortality by integrating social determinants of health. Circ: Heart Failure 2022 Nov;15(11):e009473. [doi: 10.1161/circheartfailure.122.009473]
32. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021 Dec 10;27(12):2176-2182. [doi: 10.1038/s41591-021-01595-0]
33. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019 Oct 04;3(1):18 [FREE Full text] [doi: 10.1186/s41512-019-0064-7] [Medline: 31592444]
34. Zhang Z, Rousson V, Lee W, Ferdynus C, Chen M, Qian X, written on behalf of AME Big-Data Clinical Trial Collaborative Group. Decision curve analysis: a technical note. Ann Transl Med 2018 Aug;6(15):308-308 [FREE Full text] [doi: 10.21037/atm.2018.07.02] [Medline: 30211196]
35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
36. Rothman M, Levy M, Dellinger R, Jones S, Fogerty R, Voelker K, et al. Sepsis as 2 problems: identifying sepsis at admission and predicting onset in the hospital using an electronic medical record-based acuity score. J Crit Care 2017 Apr;38:237-244 [FREE Full text] [doi: 10.1016/j.jcrc.2016.11.037] [Medline: 27992851]

Abbreviations

AKI: acute kidney injury
AUPRC: area under the precision recall curve
AUROC: area under the receiver operating characteristic curve
DCA: decision curve analysis
ECE: expected calibration error
EHR: electronic health record
FN: false negative
FNR: false-negative rate
FP: false positive
FPR: false-positive rate
ICD-10-GM: German Modification of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision
KDIGO: Kidney Disease Improving Global Outcomes
MCE: maximum calibration error
ML: machine learning
NEN: named entity normalization
NER: named entity recognition
NLP: natural language processing
TN: true negative
TP: true positive

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 03.08.23; peer-reviewed by SB Duan, P Ghasemi, A Mayampurath, M Rosenberg; comments to author 05.11.23; revised version received 30.01.24; accepted 16.10.24; published 13.12.24.

Please cite as:
Cabanillas Silva P, Sun H, Rezk M, Roccaro-Waldmeyer DM, Fliegenschmidt J, Hulde N, von Dossow V, Meesseman L, Depraetere K, Stieg J, Szymanowsky R, Dahlweid FM
Longitudinal Model Shifts of Machine Learning–Based Clinical Risk Prediction Models: Evaluation Study of Multiple Use Cases Across Different Hospitals
J Med Internet Res 2024;26:e51409
URL: https://www.jmir.org/2024/1/e51409
doi:10.2196/51409
PMID:

©Patricia Cabanillas Silva, Hong Sun, Mohamed Rezk, Diana M Roccaro-Waldmeyer, Janis Fliegenschmidt, Nikolai Hulde, Vera von Dossow, Laurent Meesseman, Kristof Depraetere, Joerg Stieg, Ralph Szymanowsky, Fried-Michael Dahlweid. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of Negative Online Reviews and Physician Responses on Health Consumers' Choice: Experimental Study

Xi Han[1], PhD; Yongxi Lin[1], MMSc; Wenting Han[2*], PhD; Ke Liao[1], MMSc; Kefu Mei[3*], MSc

[1]School of Business Administration, Guangdong University of Finance & Economics, Guangzhou, China

[2]School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, China

[3]Department of Internal Neurology, Xiangyang Central Hospital, Xiangyang, China

[*]these authors contributed equally

Corresponding Author:
Wenting Han, PhD
School of Management Science and Engineering
Shandong University of Finance and Economics
7366 Erhuan East Road, Yifu Building 5th FL.
Jinan, 250014
China
Phone: 86 15951933930
Email: hwt_2023@126.com

Abstract

Background: The COVID-19 pandemic has highlighted the importance of online medical services. Although some researchers have investigated how numerical ratings affect consumer choice, limited studies have focused on the effect of negative reviews that most concern physicians.

Objective: This study aimed to investigate how negative review features, including proportion (low/high), claim type (evaluative/factual), and physician response (absence/presence), influence consumers' physician evaluation process under conditions in which a physician's overall rating is high.

Methods: Using a 2×2×2 between-subject decision-controlled experiment, this study examined participants' judgment on physicians with different textual reviews. Collected data were analyzed using the t test and partial least squares–structural equation modeling.

Results: Negative reviews decreased consumers' physician selection intention. The negative review proportion (β=–0.371, *P*<.001) and claim type (β=–0.343, *P*<.001) had a greater effect on consumers' physician selection intention compared to the physician response (β=0.194, *P*<.001). A high negative review proportion, factual negative reviews, and the absence of a physician response significantly reduced consumers' physician selection intention compared to their counterparts. Consumers' locus attributions on the negative reviews affected their evaluation process. Physician attribution mediated the effects of review proportion (β=–0.150, *P*<.001), review claim type (β=–0.068, *P*=.01), and physician response (β=0.167, *P*<.001) on consumer choice. Reviewer attribution also mediated the effects of review proportion (β=–0.071, *P*<.001), review claim type (β=–0.025, *P*=.01), and physician response (β=0.096, *P*<.001) on consumer choice. The moderating effects of the physician response on the relationship between review proportion and physician attribution (β=–0.185, *P*<.001), review proportion and reviewer attribution (β=–0.110, *P*<.001), claim type and physician attribution (β=–0.123, *P*=.003), and claim type and reviewer attribution (β=–0.074, *P*=.04) were all significant.

Conclusions: Negative review features and the physician response significantly influence consumer choice through the causal attribution to physicians and reviewers. Physician attribution has a greater effect on consumers' physician selection intention than reviewer attribution does. The presence of a physician response decreases the influence of negative reviews through direct and moderating effects. We propose some practical implications for physicians, health care providers, and online medical service platforms.

(J Med Internet Res 2024;26:e46713) doi:10.2196/46713

KEYWORDS

negative review; proportion; claim type; attribution theory; physician-rating websites; consumer; physician response

XSL•FO
RenderX

Introduction

Background

The user-generated content (UGC) mechanism provides patients an opportunity to post online reviews about physicians' services [1]. These online physician reviews (OPRs) are attracting increasing consumer attention in recent years. In the beginning, health consumers mainly referred to these OPRs to book physicians offline [2]. With the development of online medical services, OPRs have become the most influential and important information source for selecting online physicians [3,4]. Particularly in the past few years, the COVID-19 pandemic has boosted the use of online medical services. Many consumers tend to seek online medical services to avoid physical touch, and they gradually get used to such services. Data from the Cyberspace Administration of China show that the number of online medical service users increased from 214 million in December 2020 to 300 million in June 2022, and 50% are new users with 1-2 years of experience [5]. In the consumerism era, the growing popularity of online medical services creates a need for physicians to embrace their online reputation and understand consumers' decision processes, even though they have an unfavorable opinion of OPRs [6,7]. Therefore, an important question needs to be answered: How do OPRs affect consumers' physician selection process in the online medical service context? The answer will be helpful for physicians to better manage their online reputation and promote their online performance.

Researchers have noted the OPR phenomenon and carried out studies on consumers' evaluation of OPRs for physician selection [8,9]. A few studies have mainly focused on the review valence and confirmed the effectiveness of OPRs. McBride [10] found that parents are more likely to choose a neighbor-recommended physician with positive online ratings than a recommended physician with no rating information. Online review valence is positively related to consumer choice, and negative reviews have a greater influence than positive reviews [11,12]. Some other studies have investigated the effect of different numerical ratings on consumer choice. Han et al [13] reported that the rating and review quantity affect consumer choice through perceived physician trustworthiness. Lu and Wu [14,15] found that the number of reviews is more effective in influencing patient decisions compared to the overall rating and that disease risk moderates the relationship between the service quality score and patient choice. Li et al [4,16] also found that a high online rating has a positive and significant effect on patient utility. Chen and Lee [17] noted that a physician's average rating increases the physician's patient flow and that rating credibility moderates the positive effect. Additionally, textual review features have been investigated as they can improve the predictive power of patient choice [18,19]. A randomized experiment found that the association of review style with review quantity affects users' attitude toward the rated physician [20]. The position, proportion, and length of negative physician reviews can also affect health consumers' selection intention [21,22].

Although studies have provided insights into the understanding of health consumers' decision-making behavior in the OPR context, there are still some important issues that need to be further explored. First, few studies have examined the effect of textual negative review features on helping narrow down consumer choice in the second stage of decision-making. As negative OPRs are inevitable for physicians, who worry that even 1 negative review can affect patients' choice and damage their reputation [23], more negative OPR features need researchers' attention. Second, the mechanism of how negative reviews work is overlooked. Although some studies have investigated the direct effect of textual review features [20,21], knowledge about how these reviews work is still limited. Investigating the mechanism will help us better understand the effect and provide more insights for stakeholders. Third, physician responses to OPRs are seldom discussed in consumers' decision-making process. To manage negative reviews, physicians are actively involved in online services and provide necessary responses to save their reputation [24]. Whether and how the presence of a physician response plays a positive role in promoting their performance needs further investigation.

To address gaps in the literature, drawing from attribution theory, our study aimed to explore the underlying mechanism of how negative review features and the physician response affect consumers' behavior in the online medical service environment. The main objectives of our study were threefold: Our first objective is to investigate the relationship between negative review features and consumers' physician evaluation. The second objective is to validate the causal locus attribution that mediates the relationship between negative review features and physician selection behavior. The last objective is to assess whether the physician's response moderates the effect of negative review features on a different locus attribution. Given that eHealth services are increasingly popular, we tested these arguments in the online medical service context to provide insights for stakeholders.

Literature Review and Theoretical Background

Negative Online Review Features

An online review is a negative or a positive statement made by former consumers about a product, which is available to numerous people via the internet [25]. Negative information is considered more informative and diagnostic than positive information, and negative cues attract more attention than positive ones [26]. Consumers always place more emphasis on negative reviews when making a judgment or purchase decision [27,28]. When the decision-making process focuses on the textual message, positive framing is less effective than negative framing [29]. In this sense, negative reviews always have a stronger influence on consumers' purchase intention than positive reviews do [11,30,31]. Thus, research always focuses on the influence of negative reviews on consumers' decisions.

Specifically, prior studies have focused on several features of negative textual reviews, such as quantity [32], emotional intensity [33], severity of failure [34,35], proportion [25,34,36,37], quality [25,36], type [38], claim [39], and presentation order [32]. All these features can significantly

XSL•FO
RenderX

affect consumer evaluation. In the online medical service context, physicians are no less worried about negative reviews than online merchants, even though most OPRs are positive [40]. Physicians worry that patients' subjective reviews will damage their reputation, and the proportion and claims of negative OPRs are the main concern [7]. To decrease physicians' concerns and provide insights into the understanding of consumers' information processing progress, this study aimed to examine whether the proportion and claim type of negative OPRs influence consumer choice through different mechanisms, particularly under the condition in which the physician's overall rating is high.

Physician Responses

As negative reviews have a notable impact on consumers' attitude and final decision [41], companies are increasingly using responses to reviews as a service recovery method to respond and apologize to complainants [41]. Responses to negative reviews can be understood as a service carried out to cocreate value with consumers and other stakeholders [42]. Regardless of the response strategy, managerial responses can create external effects on negative review posters and potential consumers [43]. Responses to negative reviews also mitigate the adverse effects of the negative reviews on the sellers' reputation and positively influence consumers' trust toward the sellers, as well as consumers' evaluation of the products and services [44].

Similar to online sellers, physicians are also under the threat of negative reviews. With increasingly more OPRs being generated, physicians have to respond to negative reviews to protect their reputation and decrease the negative effects. However, few studies have investigated the effect of physician responses. From a health consumer's perspective, since they try to make use of further information about the rated physician from the responses, the responses serve as an important cue for evaluation and decision-making. If there is no physician response, consumers can only rely on negative reviews, which may increase their concerns. As both negative reviews and managerial responses can affect consumers' evaluating process, there is a need to examine the effects of negative OPRs and physician responses, as well as the interactive effect on consumer behavior. Specifically, this study aimed to examine the direct and indirect effects of physician responses on consumers' final choice.

Attribution Theory

Attribution theory is a social cognition theory that addresses causal explanations and reasoning about the expressions and actions of another person [45]. This theory provides an understanding of how individuals interpret incidents based on their causal inferences, and it has been widely used in management and marketing research to explain individuals' evaluation processes [46,47]. In particular, attribution theory is an effective approach to understanding unfavorable incidents [48], such as negative employee reviews [49] and service failures [50]. Attribution theory has three dimensions, namely locus of control (internal or external), stability (stable or unstable), and controllability (controllable or uncontrollable) [48], and researchers often use different dimensions of the theory to fit their research questions and contexts [46]. In this study, we mainly focused on the "locus of control" dimension. The locus of causality establishes the cause as residing within the person (internal) or outside the person (external). The locus of control is related to who is to blame for undesirable or negative experiences [51].

In the OPR context, when health consumers encounter a negative OPR, they seek the underlying cause and show specific attitudes or emotions toward the physician, which affects their subsequent evaluation and selection behavior. In general, readers mainly attribute the origin of a negative review to the reviewer (external attribution) or the rated physician (internal attribution). However, how health consumers form different attribution loci is still unclear. Thus, for a deeper understanding of the psychological process that drives consumers' online physician selection behavior, this study further introduced physician attribution and reviewer attribution as the mediating variable in the link between negative review features and consumer choice.

Methods

Research Model and Hypothesis Development

Based on the literature and attribution theory, we constructed our research model (Figure 1). We examined the effects of negative review features on consumers' physician selection intention (hypothesis 1 [H1], H2). Furthermore, we investigated the mediating role of the attribution locus in the relationship between review features and consumer choice (H4, H5, H6) and how the presence of a physician response plays a role (H3, H7).

XSL•FO
RenderX

Figure 1. Proposed research model and hypotheses. H: hypothesis.

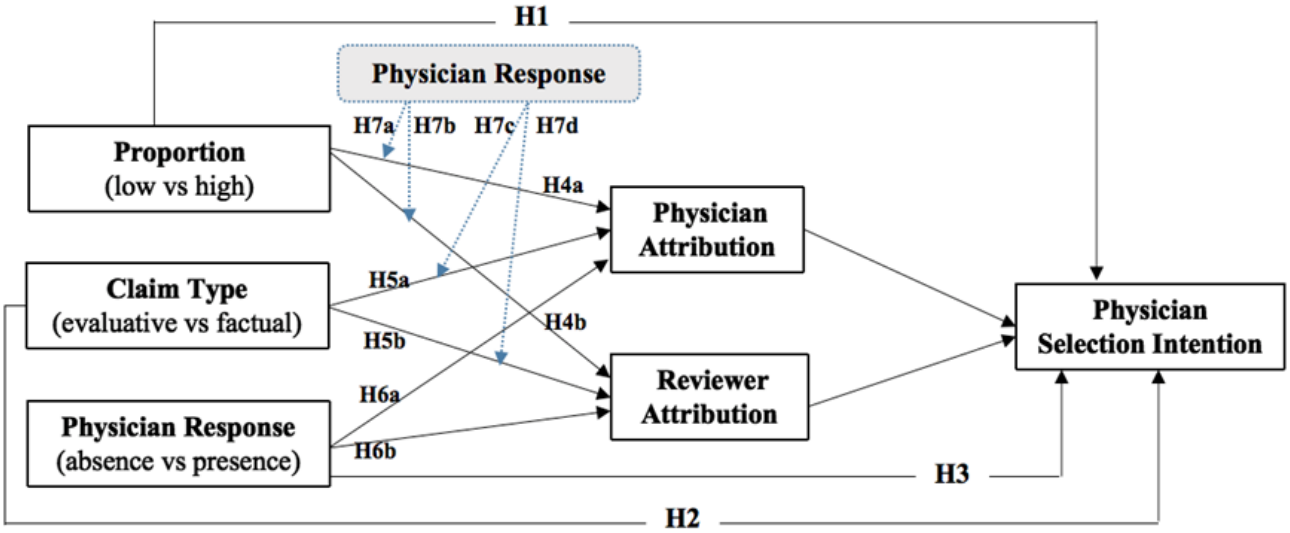

Negative OPRs and Physician Selection Intention

Although the majority of OPRs are positive, it is inevitable for physicians to receive negative reviews. In the era of information overload, the existence of large volumes of positive reviews makes consumers pay more attention to limited negative reviews [27]. The negative review features provoke consumers' deep evaluation in the second stage of decision-making after they have narrowed down their choices by rating and review quantity in the first stage.

The proportion of negative reviews is the number of negative reviews compared with overall online reviews. It is an important variable consumers consider when evaluating the online reviews of a desired object [52]. An increase in negative reviews escalates consumers' perception of product risk and in turn depresses their purchase intention [25,53]. A previous study confirmed that consumers' attitude lessens when the negative online review proportion rises [36]. In the OPR context, consumers' physician selection intention decreases as the negative review proportion increases, even if the physician's overall rating is relatively high. Based on these findings, we set out the first hypothesis (H1):

> *H1: A high proportion of negative OPRs decreases consumers' physician selection intention more than a low proportion of reviews.*

The other review feature that affects consumers' judgment is the claim type of a review: evaluative or factual. An evaluative claim includes emotional, subjective impressions or opinions about product attributes, whereas a factual claim includes numeric values and objective facts about product attributes [54]. The definitions indicate that the 2 types of claims differ in the levels of objectivity and verifiability [55]. Factual claims are easily verifiable and provide detailed indicators of the performance [56], while evaluative claims are susceptible to multiple interpretations and are often considered not objective. Therefore, consumers are likely to perceive factual reviews as more helpful for their decision-making than evaluative claims [57]. In the OPR context, some patients write detailed factual descriptions about their negative medical experiences, while other patients just express their anger and negative emotions. For a physician with a high overall rating, factual claims in negative reviews are more persuasive than evaluative claims. Hence, we reasonably hypothesized the following:

> *H2: Factual negative OPRs decrease consumers' physician selection intention more than evaluative reviews do.*

Responses to negative reviews are usually written positively to reduce the impact of the reviews. In e-commerce settings, a seller's apology and explanation of the cause of the service failure can be perceived as a repair effort by readers [58]. The responses provide additional and reasonable information for consumers' evaluation process. Prior studies have shown that responses have positive effects on complainants' repurchase intentions [59], and they can affect the extent to which consumers are satisfied with service recovery [60]. The responses of high-rated physicians who receive negative reviews demonstrate their responsibility and effectively decrease the influences of the negative reviews. However, an absence of responses increases information asymmetry and consumers' risk perception. Therefore, we proposed the following hypothesis:

> *H3: The presence of a physician response increases consumers' selection intention more than the absence of a physician response does.*

The Mediating Role of Physician Attribution and Reviewer Attribution

According to attribution theory, people always try to assess to what extent a certain entity is responsible for a negative or unexpected event [61]. Applied to the context of negative OPRs, health consumers always assess who is responsible for the negative reviews they encounter. The "locus of control" concept indicates that a negative review is caused mostly by either internal factors or external factors [61]. On the one hand, consumers may attribute a negative OPR to physician-related factors. In this case, consumers believe the negative information reflects what actually happens. On the other hand, consumers may attribute the cause or responsibility to reviewer-related factors, such as their motivations, moods, or bias [46,62]. In this case, the negative review is not (or is less) perceived as

XSL•FO
RenderX

reflecting reality but rather as the reviewer's personal subjective or prejudiced feeling. Due to information uncertainty and the complexity of medial issues, health consumers often attribute the causes of a negative review to both the physician and the reviewer, although to different degrees. When the level of physician attribution increases, consumers' willingness to select the physician decreases. On the contrary, consumers' physician selection intention increases when the level of reviewer attribution increases.

Regarding the formation of a different locus attribution, specific negative review features play an important role. In this study, the negative review proportion, claim type, and physician response may have an influence on consumers' attribution judgment.

Consumers' evaluation is based on their perceived conformity of reviews. A prior study on conformity research indicated that the greater the number of people with the same opinion, the greater the level of conformity to that opinion [36]. In other words, people are always influenced by the majority of the group, and the proportion of negative reviews from others could be a critical criterion for their attribution judgment. For a physician with a high overall rating, a higher proportion of negative reviews makes consumers believe the credibility of those reviews and confirms that the negative reviews are caused by the physician. Thus, consumers' physician attribution level increases and reviewer attribution level decreases. In the case of a low proportion of negative reviews, consumers trust positive reviews posted by others more and tend to believe that negative reviews are caused by reviewer-related factors. Thus, consumers' physician attribution level decreases and reviewer attribution level increases. As physician attribution and reviewer attribution are related to consumers' physician selection intention, we decided to also examine the mediating effects of physician attribution and reviewer attribution in our study. Therefore, we proposed the following hypotheses:

H4a: Physician attribution mediates the effect of the negative review proportion on consumer choice (proportion +→ physician attribution –→ physician selection intention).

H4b: Reviewer attribution mediates the effect of the negative review proportion on consumer choice (proportion –→ reviewer attribution +→ physician selection intention).

Review contents show detailed reasons from reviewers, and different review claim types have different degrees of persuasiveness [20]. Factual negative reviews describe the objective facts or results related to a physician's service, and these factual claims have high levels of credibility and verifiability. In this case, consumers tend to believe the negative claims to be true facts, and the physician needs to be responsible for the negative experience. Thus, consumers' physician attribution level increases and reviewer attribution level decreases. On the contrary, evaluative reviews reflect reviewers' subjective, emotional opinions and impressions about the physician [39]. The nonevidence-based reviews are considered an inaccurate and imprecise evaluation of the physician's performance. As the physician's overall rating is high, it is easy to attribute the negative reviews to the reviewers' personal bias. Therefore, consumers' physician attribution level decreases and reviewer attribution level increases. Considering the relationship between locus attribution and physician selection intention, we can reasonably believe that the attribution locus mediates the relationship between review claim type and consumer choice. Hence, the following hypotheses were proposed:

H5a: Physician attribution mediates the effect of the negative review claim type on consumer choice (claim type +→ physician attribution –→ physician selection intention).

H5b: Reviewer attribution mediates the effect of the negative review claim type on consumer choice (claim type –→ reviewer attribution +→ physician selection intention).

Most negative OPRs derive from misunderstandings between physicians and patients. The physician response can help develop a better relationship with complainants and create an external effect on potential consumers [43]. Both negative reviews and physician responses can provide insights into physician service. Physician responses provide additional information for potential consumers to make decisions, in addition to negative reviews. If there is no physician response, consumers can only use one-sided negative information to make causal attribution judgments, and the likelihood of physician attributing increases. A physician response not only shows the physician's responsibility but also reduces misunderstanding. In this case, consumers' physician attribution level is likely to decrease and their reviewer attribution level is likely to increase. Thus, a physician response makes consumers perceive a lower level of physician attribution than no response, while it has the opposite effect on reviewer attribution. As physician attribution and reviewer attribution are related to consumers' selection intention, the 2 locus attributions can be said to mediate the effect of the physician response on consumer choice. Based on this, the study hypothesized the following:

H6a: Physician attribution mediates the effect of the physician response on consumer choice (physician response –→ physician attribution –→ physician selection intention).

>H6b: Reviewer attribution mediates the effect of the physician response on consumer choice (physician response +→ reviewer attribution +→ physician selection intention).

The Moderating Role of Physician Responses

Responses to negative reviews can not only directly influence consumer attitude [49] but also indirectly decrease the effect of negative reviews on consumers' intention. As a communication strategy, managerial responses are targeted at problem solving and trust remedy. Responses are expected to reduce the unfavorable impacts of negative reviews on both potential and existing consumers [63]. Some studies have confirmed the moderating effect of responses on the relationship between negative reviews and sellers' performance. For example, responses have been found to moderate the effect of negative reviews on booking intention [64] and the number of bookings

XSL•FO
RenderX

[65]. Different from prior studies, we aimed to explore the moderating effects of managerial responses (absence or presence) on the relationship between negative review features and locus attribution. We proposed the following hypotheses:

> *H7a/H7b: The impact of the proportion of negative reviews on physician attribution/reviewer attribution decreases/increases in the presence of a physician response compared to the absence of a physician response.*
>
> *H7c/H7d: The impact of factual negative reviews on physician attribution/reviewer attribution decreases/increases in the presence of a physician response compared to the absence of a physician response.*

Study Design

Medical websites provide many physicians' rating information for consumers, including ratings, review quantity, and textual reviews. Health consumers tend to narrow down their choice set by numerical rating information in the first stage, while they deliberately read the textual reviews to make the final decisions in the second stage [13]. To examine the effect of negative textual reviews, this study used a 2 (negative review proportion: low vs high) × 2 (negative review type: evaluative vs factual) × 2 (physician response: absence vs presence) between-subject design in a decision-controlled setting, where participants were asked to select a physician from a fictitious website designed for the purpose of the experiment. A pretest and a main experiment were conducted. The pretest was used to assess the appropriateness of the stimulus manipulation, and the main experiment was used to test the effects of stimulus features.

Stimulus Design

To provide the participants with a realistic context, we designed a physician selection scenario and the fictitious stimulus resembled a famous medical website in China (91160 [66]). The website has more than 659,000 physicians and has served 814 million person-times up to September 2023. The characteristics of this website can be generalized to other medical websites. The selected physician works in a high-level hospital in a city, and we used a physician icon that did not show any additional features of the physician in order to control the influence of the physician's image. For textual reviews, we presented the newest 10 of 35 reviews on the first page. The detailed fictitious reviews are shown in Multimedia Appendix 1.

Proportion of Negative Reviews

Prior studies have considered 8 online reviews in their research [25,36]. Most experiments have used 12.5% or 25% as the low-proportion threshold and 50% or 62.5% as the high-proportion threshold [25,36,37]. One study in the OPR context used 1/6, 2/6, 3/6, 4/6, 5/6, and 6/6 to represent different levels of the negative review proportion [21]. In our study, participants were presented with 10 textual reviews. This design is prevalent on Chinese websites. When designating the proportion of negative reviews, we not only referred to prior research [25,36,37] but also considered the realistic situation of physician-rating websites. In a simulated page containing 10 textual reviews, 1 negative review is the minimal unit, which is common in realistic situations. Thus, we set 10% as a low proportion, which is also closer to prior research. Considering that physicians always get a few negative reviews on real physician-rating websites, we did not use a proportion exceeding 50% to represent a high proportion as in previous studies. Instead, after investigating a large number of physician reviews on the 91160 website [66], we think that designating a ratio of over 50% as a high proportion is not in line with the actual situation. Therefore, we chose 3 of 10 (30%) to represent a high proportion. It is not only different from the ratio of 10% but also conforms to reality. This designation is also more realistic than previous experiments. To control the influence of the negative review position [21], we set the negative reviews at 3rd, 6th, and 8th place in the high-proportion group and the 5th place in the low-proportion group.

Claim Type of Negative Reviews

All the 10 reviews in the manipulated setting were real reviews posted on the 91160 website [66], which increased the external validity. With the exception of the claim type and review valence, the reviews were identical, containing evaluations about physicians' skills, effect, and attitude. For a positive review, all attributes were described positively, whereas for a negative review, all attributes were described negatively. Additionally, factual claims included detailed descriptions of patients' experiences and the effects, whereas evaluative claims focused on patients' subjective feelings about attributes (unfriendliness, ineffective, incompetence).

Physician Responses

In this study, we only examined how the presence or absence of physician responses to negative reviews affects consumer evaluation. In the response absence scenario, the physician provides no response to any negative or positive review. In the response presence scenario, the physician provides a necessary explanation and encourages the reviewer to visit for further treatment again. This response strategy is widely used by physicians on the 91160 website [66]. Finally, 8 versions of web pages containing numerical rating information and textual reviews were created.

Pretest

We also conducted a pretest with 72 participants to assess the appropriateness of the stimulus manipulation. These participants were convenience samples recruited via researchers' WeChat, and they all had used OPRs. The participants were asked to rate a physician proposal for a neighbor, and each participant was randomly assigned to 1 alternative proposal. They used a 7-point Likert scale to rate the physician on 2 aspects: resemblance of the physician proposal to the 91160 website and understandability of the textual review information.

Main Experiment

First, we explained the purpose of the experiment. Next, during the experiment, we presented the following brief scenario:

> *Aunt Wang is your familiar neighbor, and she has a stomachache for a few days. She finds a physician on the 91160 website, which is a widely used medical*

XSL•FO
RenderX

website. You are requested to help her decide whether the physician is the right choice.

As Chinese people often help friends and family members select physicians, the scenario is prevalent in the real world. Next, we presented the physician's picture, which included only personal information and a numerical rating. The physician's rating was 9.1, and the review quantity is 35, which are average values on the 91160 website [66]. Participants were asked to report their intention to select the physician before they were exposed to negative textual reviews. Finally, 10 textual reviews of the physician were added. Participants were requested to read the reviews carefully and then complete questions about the manipulation effect, physician attribution, reviewer attribution, and physician selection intention. This procedure was consistent with the steps health consumers take to select a physician in the online environment.

Participants

We used the snowballing method to recruit participants. As students are not the main online medical service users, they were excluded in the first-round survey. First, 50 participants with different genders, ages, education levels, living cities, and career backgrounds were recruited via researchers' WeChat. Second, the 50 participants were requested to invite their WeChat friends with different socioeconomic statuses to finish our online experiment. We also used survey quotas based on gender, age, and education level, ensuring a higher diversity of the sample compared to standard convenience sampling techniques. All sessions of the experiment were conducted online, and participants were randomly assigned to 1 of 8 experimental scenarios.

Ethical Considerations

The study was approved by the Ethics Committee of the Guangdong University of Finance & Economics. All participants provided informed consent upon registration for the study. Participants were ensured that their privacy would be protected and their answers would only be used for academic research. Each participant was paid RMB 3 (US $0.42) after they conscientiously completed the questionnaire.

Measures

The dependent variable was physician selection intention, which was assessed using 3 items rated on a 7-point scale (ranging from 1 for completely disagree to 7 for completely agree). The scale was adapted from previous studies [13,21]. Physician attribution and reviewer attribution were also measured and included as mediating variables. The 2 constructs were adapted from a prior study [34]. For subsequent data analyses, we calculated the mean scores for all items of each construct.

After the questionnaire was initially developed, we invited 3 experts to assess content accuracy, validity, and ease of understanding before the experiment began. Additionally, a pilot test was conducted with 15 participants to verify the clarity of the questions and 2 items were slightly modified based on their feedback.

Manipulation Check

Based on previous studies [67,68], we conducted a manipulation check. Every participant was asked to rate the proportion of negative reviews, the negative review claim type, and the physician response.

Data Analysis

Using IBM SPSS Statistics and SmartPLS3, we applied 1-way ANOVA and partial least squares–structural equation modeling (PLS-SEM) to test the proposed hypotheses. PLS allows the simultaneous testing of mediation with minimum bias, resulting in greater appreciation of the complete effect. The method is better than simple linear regression, in which the mediation pathway is tested individually [69]. We conducted bootstrapping with 5000 resamples to obtain the *t* statistics and SEs. We also performed a full collinearity test to determine whether any constructs contained variance inflation factor (VIF) values of ≥3.3.

Results

Participant Details

In total, 550 participants provided complete data, and 91 (16.5%) surveys were discarded for too short response times or same values of all continuous measurement scales. Finally, 459 (83.5%) valid questionnaires were collected, with samples ranging from 51 to 67 (11.1%-14.6%) in the 8 experimental scenarios (Table 1). These participants were randomly grouped according to the last digit of their ID card number. The chi-square test verified that there were no significant differences (P=.24) in the characteristics of the participants among the 8 experimental scenarios.

Table 1. Details of the 8 experimental scenarios, including participants (N=459).

Scenario	Proportion of negative reviews (N=10), n (%)	Claim type	Response	Sample, n (%)
1	1 (10)	Evaluative	Absent	67 (14.6)
2	3 (30)	Evaluative	Absent	51 (11.1)
3	1 (10)	Factual	Absent	53 (11.5)
4	3 (30)	Factual	Absent	57 (12.4)
5	1 (10)	Evaluative	Present	55 (12.0)
6	3 (30)	Evaluative	Present	56 (12.2)
7	1 (10)	Factual	Present	59 (12.9)
8	3 (30)	Factual	Present	61 (13.3)

XSL•FO
RenderX

Demographic Characteristics

Participants' demographic characteristics are shown in Table 2. Of the 459 participants, 280 (61%) were 31-45 years old, 226 (49.2%) were female, and 262 (57.1%) were married. In addition, 314 (68.4%) participants had completed a college or higher level of education, and 323 (70.4%) had a monthly income of RMB 3001-12,000 (US $416.89-$1667.01).

Table 2. Demographic characteristics of the sample (N=459).

Characteristics	Participants, n (%)
Age (years)	
23-30	129 (28.1)
31-35	156 (33.9)
36-45	124 (27.1)
≥46	50 (10.9)
Gender	
Male	233 (50.8)
Female	226 (49.2)
Daily internet usage (hours)	
≤3	15 (3.3)
3 to <7	260 (56.6)
7 to <11	78 (17.0)
11 to ≤13	106 (23.1)
Monthly income (RMB)[a]	
≤3000 (US $416.75)	96 (20.9)
3001-8000 (US $416.89-$1111.34)	172 (37.5)
8001-12,000 (US $1111.48-$1667.01)	151 (32.9)
≥12,001 (US $1667.15)	40 (8.7)
Highest education	
Middle school	145 (31.6)
College/bachelor	154 (33.6)
Master/PhD	160 (34.8)
Marital status	
Married	262 (57.1)
Single	197 (42.9)

[a]An exchange rate of RMB 1.00=US $0.14 has been used.

Pretest

In the pretest to assess the appropriateness of stimulus manipulation, the mean score of resemblance was 6.47 (SD 0.94), and the mean score of review understandability was 5.78 (SD 1.07). No significant differences (P=.32) were observed between 3 manipulated factor designs in resemblance and understandability.

Manipulation Check

ANOVA results indicated that participants in the high-proportion group rated the proportion of negative reviews as higher compared to the low-proportion group ($mean_1$ 5.700, SD 0.947; $mean_2$ 2.290, SD 0.960; $F_{1,457}$=1464.039, P<.001). Participants exposed to factual negative reviews perceived the reviews as more objective than participants exposed to evaluative negative reviews ($mean_1$ 5.440, SD 1.030; $mean_2$ 2.640, SD 0.934; $F_{1,457}$=930.495, P<.001). Participants exposed to response presence scenarios perceived a higher level of response than those exposed to a response absence scenario ($mean_1$ 5.540, SD 1.052; $mean_2$ 2.530, SD 0.918; $F_{1,457}$=1065.037, P<.001). No other effects were significant. Thus, the manipulation was successful.

Model Assessment

We used internal consistency reliability, convergent validity, and discriminant validity to examine measures. As Table 3 shows, composite reliability (CR) values (above the lower limit of 0.7) indicated internal consistency reliability. As standardized factor loadings were >0.7 at a significance level of .001 and the average variance extracted (AVE) values were well above 0.5 [70], the convergent validity was supported. Furthermore, as

XSL•FO
RenderX

the heterotrait-monotrait (HTMT) ratio of the correlations was lower than the threshold of 0.85 and the HTMT CIs did not include 1.0, all constructs exhibited discriminant validity. Thus, the results provide supportive evidence for the constructs' reliability and validity.

Table 3. Assessment of reliability and convergent validity.

Construct and items	Mean (SD)	Loading	AVE[a]	CR[b]	Cronbach α
Physician attribution (PA)			0.772	0.910	0.852
PA1	3.08 (1.524)	0.882	__[c]	—	—
PA2	3.20 (1.507)	0.888	—	—	—
PA3	3.09 (1.522)	0.866	—	—	—
Reviewer attribution (RA)			0.753	0.901	0.836
RA1	4.78 (1.440)	0.865	—	—	—
RA2	4.62 (1.494)	0.885	—	—	—
RA3	4.38 (1.461)	0.853	—	—	—
Physician selection intention (PSI)			0.903	0.945	0.917
PSI1	4.42 (1.772)	0.920	—	—	—
PSI2	4.41 (1.808)	0.927	—	—	—
PSI3	4.43 (1.750)	0.930	—	—	—

[a]AVE: average variance extracted.

[b]CR: composite reliability.

[c]Not applicable.

Additionally, the standardized root mean square residual value for the structural model was <0.08 (0.076 for our model), which indicated good model fit. The R^2 values for physician attribution (R^2=0.221), reviewer attribution (R^2=0.416), and consumers' physician selection intention (R^2=0.517) were all larger than 0, supporting the predictive relevance of the model.

Collinearity Test

Data analysis results showed that pathological VIFs for all constructs ranged from 1.586 to 2.036, confirming that common method biases are not a threat for our study.

Hypotheses Test

Main Effects

Before the hypotheses test, we used paired-sample *t* tests to compare consumers' physician selection intention before and after they were exposed to negative textual reviews (Table 4). According to the *t* test results, consumers reported lower levels of physician selection intention after reading the detailed negative reviews ($mean_{before}$ 5.059, SD 1.493; $mean_{after}$ 4.419, SD 1.645; *P*<.001).

Table 4. Results of the t test of the main effects on the physician selection intention.

Review features	Mean (SD)	*P* value
Negative review exposure		<.001
Before	5.059 (1.493)	__[a]
After	4.419 (1.645)	—
Review proportion		<.001
High	3.760 (1.715)	—
Low	5.053 (1.292)	—
Claim type		<.001
Factual	3.805 (1.660)	—
Evaluative	5.020 (1.391)	—
Physician response		<.001
Present	4.774 (1.583)	—
Absent	4.072 (1.645)	—

[a]Not applicable.

XSL•FO
RenderX

We conducted ANOVA to test the relationship between review features and physician selection intention (Table 4). The proportion of negative online reviews significantly influenced consumer choice ($F_{1,457}$=83.618, P<.001). Participants in the high-proportion group (mean 3.760, SD 1.715) reported a significantly lower level of intention than those in the low-proportion group (mean 5.053, SD 1.292). Therefore, H1 was accepted.

The study yielded the same results for the effects of the claim type of negative reviews (Table 4). The claim type of negative reviews significantly affected consumer choice ($F_{1,457}$=72.404, P<.001). Factual negative reviews (mean 3.805, SD 1.660) led participants to report a significantly lower level of selection intention than evaluative negative reviews (mean 5.020, SD 1.391). Thus, H2 was supported.

In addition, physician response also had a significant influence on consumers' intention to select a physician with negative reviews ($F_{1,457}$=21.849, P<.001), as shown in Table 4. Participants exposed to negative reviews with a physician response present (mean 4.774, SD 1.583) reported a significantly higher level of intention compared to the response absence scenario (mean 4.072, SD 1.645). Therefore, H3 was supported.

Mediating Effects

Following the standard procedure applied in a previous study [69], we used the latent variable score generated in PLS-SEM as input for mediation and moderation effects. The results are shown in Figure 2.

Figure 2. Mediation and moderation effects of the structural model.

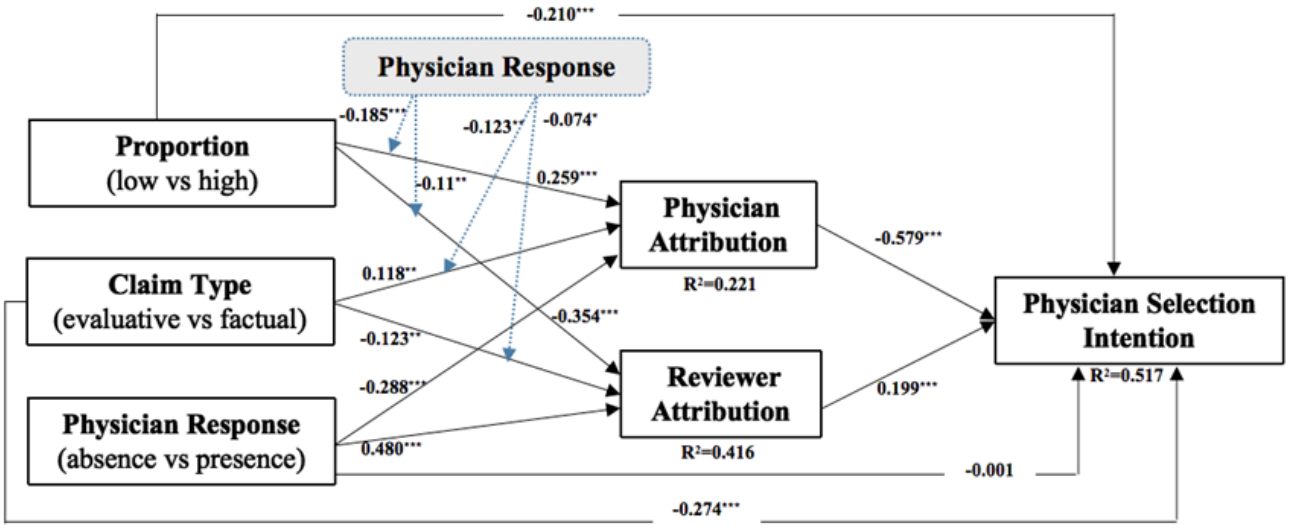

Table 5 shows the significant direct and indirect effects of review features. As shown in Figure 2 and Table 5, physician attribution (β=–0.579, 95% CI –0.640 to –0.511, P<.001) and reviewer attribution (β=0.199, 95% CI 0.137-0.263, P<.001) induced by negative review features were both significantly related to physician selection intention. Their effects were opposite, and physician attribution had a greater influence than reviewer attribution. In addition, 3 review features also had significant effects on consumers' locus attribution.

XSL•FO
RenderX

Table 5. Direct and indirect effects of review features on consumer choice.[a]

Hypothesis and parameters	Effect, β (95% CI)	SE	$t_{4.454}$ Test	*P* value
H[b]4a				
Proportion→PA[c]	0.259 (0.176 to 0.340)	0.041	6.253	<.001
PA→PSI[d]	–0.579 (–0.640 to –0.511)	0.033	17.505	<.001
Proportion→PA→PSI	–0.150 (–0.202 to –0.099)	0.026	5.701	<.001
H4b				
Proportion→RA[e]	–0.354 (–0.419 to –0.286)	0.034	10.301	<.001
RA→PSI	0.199 (0.137 to 0.263)	0.032	6.222	<.001
Proportion→RA→PSI	–0.071 (–0.100 to –0.045)	0.014	5.079	<.001
H5a				
Claim type→PA	0.118 (0.036 to 0.199)	0.042	2.778	.005
PA→PSI	–0.579 (–0.640 to –0.511)	0.033	17.505	<.001
Claim type→ PA→PSI	–0.068 (–0.118 to –0.021)	0.025	2.714	.01
H5b				
Claim type→RA	–0.123 (–0.198 to –0.051)	0.037	3.330	.001
RA→PSI	0.199 (0.137 to 0.263)	0.032	6.222	<.001
Claim type→RA→PSI	–0.025 (– 0.045 to –0.009)	0.009	2.587	.01
H6a				
Physician response→PA	–0.288 (–0.364 to –0.209)	0.040	7.280	<.001
PA→PSI	–0.579 (–0.640 to –0.511)	0.033	17.505	<.001
Physician response→PA→PSI	0.167 (0.118 to 0.214)	0.025	6.711	<.001
H6b				
Physician response→RA	0.480 (0.417 to 0.542)	0.032	14.894	<.001
RA→PSI	0.199 (0.137 to 0.263)	0.032	6.222	<.001
Physician response→RA→PSI	0.096 (0.065 to 0.130)	0.016	5.815	<.001

[a]Results are based on bootstrapping with 5000 subsamples (2-tailed).

[b]H: hypothesis.

[c]PA: physician attribution.

[d]PSI: physician selection intention.

[e]RA: reviewer attribution.

First, the proportion of negative reviews was significantly and positively associated with physician attribution (β=0.260, *P*<.001) and significantly and negatively associated with reviewer attribution (β=–0.355, *P*<.001). The mediation tests also indicated that mediations are in play, whereby the proportion of negative reviews influenced physician attribution and reviewer attribution, which in turn affected consumer selection intention. The total indirect effect size of the proportion of negative reviews on consumers' physician selection intention was –0.221 (95% CI –0.277 to –0.165, *P*=.02). Thus, H4a and H4b were supported.

Second, the negative review claim type was also significantly associated with physician attribution (β=0.118, *P*=.005) and reviewer attribution (β=–0.123, *P*=.001). Factual reviews induced a higher level of physician attribution and a lower level of reviewer attribution, while evaluative reviews induced a lower level of physician attribution and a higher level of reviewer attribution. Mediation analyses indicated that physician attribution and reviewer attribution significantly mediate the effects of the claim type on consumers' physician selection intention. The total indirect effect size of the negative review claim type on consumers' physician selection intention was –0.093 (95% CI –0.143 to –0.044, *P*=.02). Hence, H5a and H5b were supported.

Third, the physician response had a significantly opposite influence on physician attribution (β=–0.288, *P*<.001) and reviewer attribution (β=0.480, *P*<.001). Our mediation tests also indicated that physician attribution and reviewer attribution mediate the influence of the physician response on consumers' physician selection intention. The total indirect effect size of the physician response on consumer choice was 0.262 (95% CI 0.206-0.317, *P*<.001). Therefore, H6a and H6b were supported.

XSL•FO
RenderX

Considering the total effect size of 3 review features, we found that the negative review proportion (β=–0.371, 95% CI –0.447 to –0.295, *P*<.001) and review claim type (β=–0.343, 95% CI –0.420 to –0.267, *P*<.001) have greater effects than the physician response (β=0.194, 95% CI 0.110-0.259, *P*<.001).

Moderating Effects

The moderating effects of the physician response on the relationship between the negative review proportion and physician attribution (β=–0.185, *P*<.001), negative review proportion and reviewer attribution (β=–0.110, *P*<.001), claim type and physician attribution (β=0.123, *P*=.003), and claim type and reviewer attribution (β=0.074, *P*=.04) were all significant. The R^2 change in physician attribution (0.053) and reviewer attribution (0.035) significantly increased after the moderator was added. The results also indicated that the physician response decreases the effects of negative review features on physician attribution but increases the effects of negative review features on reviewer attribution. The significant moderating effects permitted subsequent investigation of indirect effects. Bootstrapping with 5000 resamples revealed that the physician response plays a significant moderating role in all the indirect relationships of negative review features with physician selection intention (Table 6), with 95% CIs never straddling 0. Thus, H7a-H7d were all supported.

Table 6. Moderating effects of the physician response.

Hypothesis and parameters	Effect, β (95% CI)	SE	$t_{4.454}$ Test	*P* value
H[a]7a				
Physician response×proportion→PA[b]	–0.185 (–0.266 to –0.105)	0.041	4.499	<.001
Physician response×proportion→PA→PSI[c]	0.107 (0.060 to 0.154)	0.024	4.468	<.001
H7b				
Physician response×proportion→RA[d]	–0.110 (–0.183 to –0.040)	0.036	3.022	.003
Physician response×proportion→RA→PSI	–0.022 (–0040 to –0.007)	0.008	2.610	.01
H7c				
Physician response×claim→PA	–0.123 (–0.205 to –0.041)	0.042	2.937	.003
Physician response×claim→PA→PSI	0.071 (0.022 to 0.118)	0.024	2.933	.003
H7d				
Physician response×claim→RA	–0.074 (–0.147 to –0.002)	0.036	2.053	.04
Physician response×claim→RA→PSI	–0.068 (–0.030 to –0.001)	0.026	2.463	.01

[a]H: hypothesis.
[b]PA: physician attribution.
[c]PSI: physician selection intention.
[d]RA: reviewer attribution.

Discussion

Principle Findings and Theoretical Contributions

The increasing proliferation of OPRs has enabled the wide usage of online medical services. Physicians should not only worry about negative reviews but also more actively obtain knowledge to better manage their online reputation [13,71]. To examine the effect of negative reviews, this study adopted attribution theory to analyze how review features influence consumers' physician selection intention. Past studies have confirmed that online reviews can influence consumer choice through attribution evaluation [72-74]. This study differs from these prior studies as we focused on negative physician reviews and different locus attributions. It explains how 2 simultaneously stimulated locus attributions can mediate the relationship between review features and consumer choice in the eHealth context. This is an important addition to the previous literature on attribution studies because it offers a deep understanding of health consumers' responses to negative reviews. The findings of this study provide insights into ways that physicians can best manage their negative online reviews. Several other key findings are worth noting.

First, the study shows that the proportion (high vs low) of negative OPRs influences consumer responses differently, and the influence is mediated by 2 coexisting locus attributions (physician attribution and reviewer attribution). A high proportion of negative OPRs significantly depresses consumers' physician selection intention compared to a low proportion of negative OPRs. This finding confirms the results of prior studies [21,34,53,75], which state that consumers' willingness tends to become more unfavorable as the proportion of negative reviews increases. However, the outcome is somewhat at odds with previous arguments that consumers' purchase intention is indifferent to exposure to high or low proportions of negative online reviews of popular products [25]. This study extends the understanding by introducing 2 mediators that explain the mechanism of how consumers evaluate the negative review proportion. As people are always influenced by the majority of the group, high exposure to negative OPRs makes consumers believe that the reviews are true. Hence, negative reviews are

XSL•FO
RenderX

more due to physician-related factors and less due to reviewer-related factors. These 2 different and coexisting causal locus attributions finally oppositely influence consumer choice.

Second, the claim type of negative OPRs also has an impact on consumers' physician selection intention via different casual locus attributions. A prior study has demonstrated that writing styles used in online reviews determine both consumers' perception of those messages and their choices [76], and researchers have obtained contradictory findings about the influence of review style on review acceptance [77]. In the OPR context, the review type also has an impact on how individuals evaluate the credibility of the review and the rated physician [78]. A previous study that suggested the interaction of review style and review quantity has an impact on consumers' decision to accept a physician, while the direct effect of a positive review style is not significant [20]. The finding differs somewhat from our study, which indicates that a negative review style influences consumer choice through different causal attributions. Focusing on negative OPRs, our study extends previous research by identifying 2 new mediators (physician attribution and reviewer attribution), in addition to review credibility and helpfulness. The 2 mediators can better explain consumer choice when they encounter a high-rated physician with negative reviews with different styles.

Third, the physician response not only has a direct influence on consumers' causal attribution evaluation but also moderates the effect of other negative review features.

As observers tend to make a default attribution to actors in a negative context, an appropriate response can provide additional information that leads consumers to acknowledge other possible causes and adjust default attribution [79]. In the OPR context, physicians with negative reviews often respond to help shift observers' attribution of negative reviews away from them. Physician responses remind consumers that the reviews are not objective and the reviewers may be responsible for the negative reviews. Therefore, physician responses increase consumers' attribution to the reviewers but decrease the attribution to physicians. This finding is similar to prior studies, which have confirmed that the effect is mediated by the reduced causal attribution of negative reviews to the firm [80]. However, our result is contrary to another study, which indicates that the presence of a response to negative reviews hurts consumers' purchase intention [81]. In addition, the presence of a physician response also increases the effect of negative review features on reviewer attribution and decreases the effect of negative review features on physician attribution. Our result is in accordance with a previous study that showed that responses moderate the effect of negative reviews on consumers' intention [64,65]. Our study also extends the findings of previous studies by simultaneously examining the direct and indirect effects of response presence and introducing a new path on which response presence plays a moderating role.

Finally, our study confirms that the effect of negative review features for a high-rated physician relies on how observers locate the cause of the negative reviews. The result can alleviate physicians' worries that even 1 negative review can affect patient choice and damage physicians' reputation [23]. For a physician with a relatively high rating, consumers further focus on negative reviews and make judgments about the cause after they make an initial decision based on the numerical rating information. This study examined the influence of numerical ratings and textual reviews in the same experiment and deepened our understanding of the consumer decision-making process.

Practical Implications

An increasing number of physicians are providing online medical services as a way to extend their offline services. Understanding the influencing mechanism of negative reviews can help them embrace online services more confidently. Our results offer managerial implications for physicians, health care providers, and online medical service platforms.

First, physicians and health care providers need to obtain more textual reviews as a way to reduce the proportion of negative reviews. Review quantity is positively related to review ratings [2], and more reviews can decrease the visibility of negative reviews. A low proportion of negative reviews increases consumers' causal attribution to reviewers and increases consumers' positive attitude. Second, physicians and health care providers should actively and appropriately respond to negative reviews. The presence of responses to negative reviews plays an important role in consumers' attribution allocation. Managerial responses not only greatly decrease physician attribution but also reduce the negative effects of reviews on physicians. There is no need for physicians to worry about negative reviews, if they provide effective responses that can activate consumers' causal attribution to the reviewers. Reviewer attribution helps improve the patient-physician relationship. Third, online medical service platforms should display negative OPRs appropriately and remind physicians to strengthen communication with patients and manage their negative reviews. Negative online reviews are important information sources for consumers' decision-making, and their position can greatly affect physicians' reputation. Online medical platforms should also enable consumers to notice negative OPRs among numerous reviews, while avoiding having unfair impacts on physicians' reputation.

Limitations

Even through the study was conducted with methodological rigor, there are still several limitations that deserve to be included in future research. First, the physician in our scenario is a high-rated physician. Although a high-rated physician is closer to reality [2], this may raise concerns in terms of the generalization of results to a low-rated physician. Future research may explore the hypotheses in the context of a low-rated physician. Second, we only tested the effects of the proportion and claim type of negative reviews for a physician. Other features, such as severity, length, position, and criticism target of negative reviews, were overlooked and deserve further investigation. Third, the response strategy in this study provides necessary explanations and encourages the reviewer to visit for further treatment again. Even though this strategy is prevalent on physician-rating websites, there are other different response styles that may affect the effectiveness of responses [82]; this also holds potential for future investigation. Fourth, sample bias may affect the generalizability of research results. The

XSL•FO
RenderX

convenient samples for this study mainly came from relatively young Chinese urban residents, who have a rich experience of using online reviews, and it is unclear whether the results are applicable to rural residents, who seldom use online reviews. Future research needs to analyze the evaluation behavior of this population.

Conclusion

The goal of this research was to understand how negative review features and physician responses affect consumer choice. The results show that the proportion of negative reviews (high vs low), the claim type of negative reviews (factual vs evaluative), and the physician response (presence vs absence) can affect consumers' physician selection intention through 2 different causal attributions (physician attribution and reviewer attribution). Physician attribution has greater influence on consumer choice than reviewer attribution does. The presence of a physician response also moderates the relationship between negative review features and causal attribution. We also provide some practical implications for physicians to manage their online reputation.

Acknowledgments

The research was supported by the National Natural Science Foundation of China (72274040), the Ministry of Education of China "Project of Humanities and Social Sciences" (22YJCZH045), and the Innovative Team Project of Guangdong Universities (2019WCXTD008). We wish to express our appreciation of the research participants.

Data Availability

Data will be made available on request.

Authors' Contributions

XH and KM contributed to data collection and writing of the paper. YL was responsible for data collection, data analysis, and conceptualization. KL and WH were responsible for data analysis and critically reviewed the paper. KM and WH contributed equally to the paper. Generative artificial intelligence was not used in any portion of the manuscript writing. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File , 1310 KB - jmir_v26i1e46713_app1.docx]

References

1. Liu JJ, Matelski J, Cram P, Urbach DR, Bell CM. Association between online physician ratings and cardiac surgery mortality. Circ Cardiovasc Qual Outcomes 2016 Nov 01;9(6):788-791. [doi: 10.1161/CIRCOUTCOMES.116.003016] [Medline: 27803091]
2. Han X, Li B, Zhang T, Qu J. Factors associated with the actual behavior and intention of rating physicians on physician rating websites: cross-sectional study. J Med Internet Res 2020 Jun 04;22(6):e14417 [FREE Full text] [doi: 10.2196/14417] [Medline: 32496198]
3. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. Public awareness, perception, and use of online physician rating sites. JAMA 2014 Feb 19;311(7):734-735. [doi: 10.1001/jama.2013.283194] [Medline: 24549555]
4. Li X, Chou S, Deily ME, Qian M. Comparing the impact of online ratings and report cards on patient choice of cardiac surgeon: large observational study. J Med Internet Res 2021 Oct 28;23(10):e28098 [FREE Full text] [doi: 10.2196/28098] [Medline: 34709192]
5. The 47th statistical report on internet development in China. Cyberspace Administration of China. 2022. URL: http://www.cac.gov.cn/2021-02/03/c_1613923423079314.htm [accessed 2024-02-23]
6. Menon AV. Do online reviews diminish physician authority? The case of cosmetic surgery in the U.S. Soc Sci Med 2017 May;181:1-8. [doi: 10.1016/j.socscimed.2017.03.046] [Medline: 28363107]
7. Samora JB, Lifchez SD, Blazar PE, American Society for Surgery of the Hand Ethics and Professionalism Committee. Physician-rating web sites: ethical implications. J Hand Surg Am 2016 Jan;41(1):104-10.e1. [doi: 10.1016/j.jhsa.2015.05.034] [Medline: 26304734]
8. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. Parental awareness and use of online physician rating sites. Pediatrics 2014 Oct;134(4):e966-e975 [FREE Full text] [doi: 10.1542/peds.2014-0681] [Medline: 25246629]
9. Ghosh S, Chakraborty S, Gupta N, Basu S. What ails physician review websites? A study of information needs of patients. Decis Support Syst 2023 Mar;166:113897. [doi: 10.1016/j.dss.2022.113897]

XSL•FO
RenderX

10. McBride DL. Parental use of online physician rating sites. J Pediatr Nurs 2015 Jan;30(1):268-269. [doi: 10.1016/j.pedn.2014.10.003] [Medline: 25450443]
11. Han X, Qu J, Zhang T. Exploring the impact of review valence, disease risk, and trust on patient choice based on online physician reviews. Telemat Informat 2019 Dec;45:101276. [doi: 10.1016/j.tele.2019.101276]
12. Carbonell G, Meshi D, Brand M. The use of recommendations on physician rating websites: the number of raters makes the difference when adjusting decisions. Health Commun 2019 Nov 17;34(13):1653-1662. [doi: 10.1080/10410236.2018.1517636] [Medline: 30222006]
13. Han X, Du JT, Zhang T, Han W, Zhu Q. How online ratings and trust influence health consumers' physician selection intentions: an experimental study. Telemat Informat 2021 Sep;62:101631. [doi: 10.1016/j.tele.2021.101631]
14. Lu W, Wu H. How online reviews and services affect physician outpatient visits: content analysis of evidence from two online health care communities. JMIR Med Inform 2019 Dec 02;7(4):e16185 [FREE Full text] [doi: 10.2196/16185] [Medline: 31789597]
15. Lu N, Wu H. Exploring the impact of word-of-mouth about physicians' service quality on patient choice based on online health communities. BMC Med Inform Decis Mak 2016 Nov 26;16(1):151 [FREE Full text] [doi: 10.1186/s12911-016-0386-0] [Medline: 27888834]
16. Li Y, Ma X, Song J, Yang Y, Ju X. Exploring the effects of online rating and the activeness of physicians on the number of patients in an online health community. Telemed J E Health 2019 Nov 23;25(11):1090-1098. [doi: 10.1089/tmj.2018.0192] [Medline: 30676279]
17. Chen Y, Lee S. User-generated physician ratings and their effects on patients' physician choices: evidence from Yelp. J Mark 2023 Mar 29;88(1):77-96. [doi: 10.1177/00222429221146511]
18. Xu Y, Armony M, Ghose A. The interplay between online reviews and physician demand: an empirical investigation. Manag Sci 2021 Dec;67(12):7344-7361. [doi: 10.1287/mnsc.2020.3879]
19. Wan Y, Peng Z, Wang Y, Zhang Y, Gao J, Ma B. Influencing factors and mechanism of doctor consultation volume on online medical consultation platforms based on physician review analysis. Internet Res 2021 Jun 24;31(6):2055-2075. [doi: 10.1108/intr-10-2020-0589]
20. Grabner-Kräuter S, Waiguny MKJ. Insights into the impact of online physician reviews on patients' decision making: randomized experiment. J Med Internet Res 2015 Apr 09;17(4):e93 [FREE Full text] [doi: 10.2196/jmir.3991] [Medline: 25862516]
21. Li S, Feng B, Chen M, Bell RA. Physician review websites: effects of the proportion and position of negative reviews on readers' willingness to choose the doctor. J Health Commun 2015 Apr;20(4):453-461 [FREE Full text] [doi: 10.1080/10810730.2014.977467] [Medline: 25749406]
22. Hu Y, Zhou H, Chen Y, Yao J, Su J. The influence of patient-generated reviews and doctor-patient relationship on online consultations in China. Electron Commer Res 2021 Aug 16;23(2):1115-1141. [doi: 10.1007/s10660-021-09506-8]
23. Schulz PJ, Rothenfluh F. Influence of health literacy on effects of patient rating websites: survey study using a hypothetical situation and fictitious doctors. J Med Internet Res 2020 Apr 06;22(4):e14134 [FREE Full text] [doi: 10.2196/14134] [Medline: 32250275]
24. Liu X, Hu M, Xiao BS, Shao J. Is my doctor around me? Investigating the impact of doctors' presence on patients' review behaviors on an online health platform. J Assoc Inf Sci Technol 2022 Feb 25;73(9):1279-1296. [doi: 10.1002/asi.24632]
25. Shihab MR, Putri AP. Negative online reviews of popular products: understanding the effects of review proportion and quality on consumers' attitude and intention to buy. Electron Commer Res 2018 Feb 28;19(1):159-187. [doi: 10.1007/s10660-018-9294-y]
26. Sperlich PW. Attribution: perceiving the causes of behavior. Edited by Edward E. Jones, David E. Kanouse, Harold H. Kelley, Richard E. Nisbett, Stuart Valins, and Bernard Weiner. (Morristown, N.J.: General Learning Press, 1972. Pp. 186. $8.95.). Am Polit Sci Rev 2014 Aug 01;70(2):617-618. [doi: 10.2307/1959677]
27. Skowronski JJ, Carlston DE. Social judgment and social memory: the role of cue diagnosticity in negativity, positivity, and extremity biases. J Pers Soc Psychol 1987 Apr;52(4):689-699. [doi: 10.1037/0022-3514.52.4.689]
28. Sen S, Lerman D. Why are you telling me this? An examination into negative consumer reviews on the Web. J Interact Mark 2007 Aug 01;21(4):76-94. [doi: 10.1002/dir.20090]
29. Maheswaran D, Meyers-Levy J. The influence of message framing and issue involvement. J Mark Res 1990 Aug;27(3):361. [doi: 10.2307/3172593]
30. Nejad MG, Amini M, Sherrell DL. The profit impact of revenue heterogeneity and assortativity in the presence of negative word-of-mouth. Int J Res Mark 2016 Sep;33(3):656-673. [doi: 10.1016/j.ijresmar.2015.11.005]
31. Charlett D, Garland R, Marr N. How damaging is negative word of mouth? Mark Bull 1995;6(1):42-50 [FREE Full text]
32. Liao H, Huang Z, Liu S. The effects of negative online reviews on consumer perception, attitude and purchase intention: experimental investigation of the amount, quality, and presentation order of eWOM. ACM Trans Asian Low-Resour Lang Inf Process 2021 May 05;20(3):1-21. [doi: 10.1145/3426883]
33. Wu Z, Guo T. The effect of negative online review's textual features on consumer's purchase intention: analysis on the perspective of self-construal. Sci Technol Dev 2019;15(6):582-587 [FREE Full text]

XSL•FO
RenderX

34. Du X, Wu Y, Gao H, Li M. Effect of negative online review and corporate response on customer purchase intention. J Syst Manag 2021;30(5):926-936.
35. S S, M.R A, Ponnam A. Can online service recovery interventions benignly alter customers' negative review evaluations? Evidence from the hotel industry. J Hosp Mark Manag 2018 Nov 26;28(6):711-742. [doi: 10.1080/19368623.2019.1544958]
36. Lee J, Park D, Han I. The effect of negative online consumer reviews on product attitude: an information processing view. Electron Commer Res Appl 2008 Sep;7(3):341-352. [doi: 10.1016/j.elerap.2007.05.004]
37. Weisstein FL, Song L, Andersen P, Zhu Y. Examining impacts of negative reviews and purchase goals on consumer purchase decision. J Retail Consum Serv 2017 Nov;39:201-207. [doi: 10.1016/j.jretconser.2017.08.015]
38. Joo B, Hwang S. Impact of negative review type, brand reputation, and opportunity scarcity perception on preferences of fashion products in social commerce. Fash Bus 2016 Sep 30;20(4):207-225. [doi: 10.12940/jfb.2016.20.4.207]
39. Kim M, Lee M. Effects of review characteristics and consumer regulatory focus on perceived review usefulness. Soc Behav Pers Int J 2015 Sep 18;43(8):1319-1333. [doi: 10.2224/sbp.2015.43.8.1319]
40. Okike K, Peter-Bibb TK, Xie KC, Okike ON. Association between physician online rating and quality of care. J Med Internet Res 2016 Dec 13;18(12):e324 [FREE Full text] [doi: 10.2196/jmir.6612] [Medline: 27965191]
41. Esmark Jones CL, Stevens JL, Breazeale M, Spaid BI. Tell it like it is: the effects of differing responses to negative online reviews. Psychol Mark 2018 Sep 27;35(12):891-901. [doi: 10.1002/mar.21142]
42. Park S, Allen JP. Responding to online reviews. Cornell Hosp Q 2012 Nov 30;54(1):64-73. [doi: 10.1177/1938965512463118]
43. Wang Y, Chaudhry A. When and how managers' responses to online reviews affect subsequent reviews. J Mark Res 2018 Apr 01;55(2):163-177. [doi: 10.1509/jmr.15.0511]
44. Ravichandran T, Deng C. Effects of managerial response to negative reviews on future review valence and complaints. Inf Syst Res 2023 Mar 28;34(1):319-341. [doi: 10.1287/isre.2022.1122]
45. Martinko MJ, Harvey P, Dasborough MT. Attribution theory in the organizational sciences: a case of unrealized potential. J Organ Behav 2010 Aug 25;32(1):144-149. [doi: 10.1002/job.690]
46. Kelley HH, Michela JL. Attribution theory and research. Annu Rev Psychol 1980 Jan;31(1):457-501. [doi: 10.1146/annurev.ps.31.020180.002325] [Medline: 20809783]
47. Abascal TE, Fluker M, Jiang M. Domestic demand for Indigenous tourism in Australia: understanding intention to participate. J Sustain Tour 2016 Jun 28;24(8-9):1350-1368. [doi: 10.1080/09669582.2016.1193187]
48. Hastie R. Causes and effects of causal attribution. J Pers Soc Psychol 1984 Jan;46(1):44-56. [doi: 10.1037/0022-3514.46.1.44]
49. Carpentier M, Van Hoye G. Managing organizational attractiveness after a negative employer review: company response strategies and review consensus. Eur J Work Org Psychol 2020 Jan 29;30(2):274-291. [doi: 10.1080/1359432x.2020.1718748]
50. Wu L, So KKF, Xiong L, King C. The impact of employee conspicuous consumption cue and physical attractiveness on consumers' behavioral responses to service failures. IJCHM 2019 Jan 14;31(1):21-40. [doi: 10.1108/ijchm-08-2017-0500]
51. Zhang Y, Prayag G, Song H. Attribution theory and negative emotions in tourism experiences. Tour Manag Perspect 2021 Oct;40:100904. [doi: 10.1016/j.tmp.2021.100904]
52. Schlosser AE. Can including pros and cons increase the helpfulness and persuasiveness of online reviews? The interactive effects of ratings and arguments. J Consum Psychol 2011 May 17;21(3):226-239. [doi: 10.1016/j.jcps.2011.04.002]
53. Xia L, Bechwati NN. Word of Mouse: the role of cognitive personalization in online consumer reviews. J Interact Advert 2008 Sep;9(1):3-13. [doi: 10.1080/15252019.2008.10722143]
54. Holbrook MB. Beyond attitude structure: toward the informational determinants of attitude. J Mark Res 2018 Dec 20;15(4):545-556. [doi: 10.1177/002224377801500404]
55. Darley WK, Smith RE. Advertising claim objectivity: antecedents and effects. J Mark 2018 Dec 13;57(4):100-113. [doi: 10.1177/002224299305700408]
56. Pratap Jain S, Buchanan B, Maheswaran D. Comparative versus noncomparative advertising: the moderating impact of prepurchase attribute verifiability. J Consum Psychol 2008 Jan 25;9(4):201-211. [doi: 10.1207/s15327663jcp0904_2]
57. Cowley E. Processing exaggerated advertising claims. J Bus Res 2006 Jun;59(6):728-734. [doi: 10.1016/j.jbusres.2005.12.004]
58. Wan Y, Zhang Y, Wang F, Yuan Y. Retailer response to negative online consumer reviews: how can damaged trust be effectively repaired? Inf Technol Manag 2022 Jun 01;24(1):37-53. [doi: 10.1007/s10799-022-00367-7]
59. Piehler R, Schade M, Hanisch I, Burmann C. Reacting to negative online customer reviews. J Serv Theory Pract 2019 Dec 02;29(4):401-414. [doi: 10.1108/jstp-10-2018-0227]
60. Bradley G, Sparks B. Explanations: if, when, and how they aid service recovery. J Serv Mark 2012 Feb 17;26(1):41-51. [doi: 10.1108/08876041211199715]
61. Weiner B. An attributional theory of achievement motivation and emotion. Psychol Rev 1985;92(4):548-573. [doi: 10.1037/0033-295x.92.4.548]
62. Chen Z, Lurie NH. Temporal contiguity and negativity bias in the impact of online word of mouth. J Mark Res 2013 Aug 01;50(4):463-476. [doi: 10.1509/jmr.12.0063]
63. Le LH, Ha Q. Effects of negative reviews and managerial responses on consumer attitude and subsequent purchase behavior: an experimental design. Comput Hum Behav 2021 Nov;124:106912. [doi: 10.1016/j.chb.2021.106912]

XSL•FO
RenderX

64. Casado-Díaz AB, Andreu L, Beckmann SC, Miller C. Negative online reviews and webcare strategies in social media: effects on hotel attitude and booking intentions. Curr Issues Tour 2018 Nov 18;23(4):418-422. [doi: 10.1080/13683500.2018.1546675]
65. Xu Y, Zhang Z, Law R, Zhang Z. Effects of online reviews and managerial responses from a review manipulation perspective. Curr Issues Tour 2019 Jun 08;23(17):2207-2222. [doi: 10.1080/13683500.2019.1626814]
66. Health160. 91160.com. URL: https://www.91160.com/ [accessed 2024-02-24]
67. Grewal D, Krishnan R, Baker J, Borin N. The effect of store name, brand name and price discounts on consumers' evaluations and purchase intentions. J Retail 1998;74(3):331-352. [doi: 10.1016/s0022-4359(99)80099-2]
68. Doney PM, Cannon JP. An examination of the nature of trust in buyer-seller relationships. J Mark 2018 Nov 19;61(2):35-51. [doi: 10.1177/002224299706100203]
69. Han X, Wang F, Lv S, Han W. Mechanism linking AR-based presentation mode and consumers' responses: a moderated serial mediation model. J Theor Appl Electron Commer Res 2021 Oct 10;16(7):2694-2707. [doi: 10.3390/jtaer16070148]
70. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 2018 Nov 28;18(1):39-50. [doi: 10.1177/002224378101800104]
71. Guetz B, Bidmon S. The impact of social influence on the intention to use physician rating websites: moderated mediation analysis using a mixed methods approach. J Med Internet Res 2022 Nov 14;24(11):e37505 [FREE Full text] [doi: 10.2196/37505] [Medline: 36374547]
72. Saleh MI. Attribution theory revisited: probing the link among locus of causality theory, destination social responsibility, tourism experience types, and tourist behavior. J Travel Res 2022 Sep 19;62(6):1309-1327. [doi: 10.1177/00472875221119968]
73. Weitzl W, Hutzinger C, Einwiller S. An empirical study on how webcare mitigates complainants' failure attributions and negative word-of-mouth. Comput Hum Behav 2018 Dec;89:316-327. [doi: 10.1016/j.chb.2018.07.012]
74. Swanson SR, Hsu MK. The effect of recovery locus attributions and service failure severity on word-of-mouth and repurchase behaviors in the hospitality industry. J Hosp Tour Res 2010 Oct 20;35(4):511-529. [doi: 10.1177/1096348010382237]
75. Bailey A. Thiscompanysucks.com: the use of the Internet in negative consumer - to - consumer articulations. J Mark Commun 2004 Sep;10(3):169-182. [doi: 10.1080/1352726042000186634]
76. Archak N, Ghose A, Ipeirotis PG. Deriving the pricing power of product features by mining consumer reviews. Manag Sci 2011 Aug;57(8):1485-1509. [doi: 10.1287/mnsc.1110.1370]
77. Lee K, Koo D. Effects of attribute and valence of e-WOM on message adoption: moderating roles of subjective knowledge and regulatory focus. Comput Hum Behav 2012 Sep;28(5):1974-1984. [doi: 10.1016/j.chb.2012.05.018]
78. Detz A, López A, Sarkar U. Long-term doctor-patient relationships: patient perspective from online reviews. J Med Internet Res 2013 Jul 02;15(7):e131 [FREE Full text] [doi: 10.2196/jmir.2552] [Medline: 23819959]
79. Ybarra O. Naive causal understanding of valenced behaviors and its implications for social information processing. Psychol Bull 2002 May;128(3):421-441. [doi: 10.1037/0033-2909.128.3.421] [Medline: 12002696]
80. Li C, Cui G, Peng L. Tailoring management response to negative reviews: the effectiveness of accommodative versus defensive responses. Comput Hum Behav 2018 Jul;84:272-284. [doi: 10.1016/j.chb.2018.03.009]
81. Mauri AG, Minazzi R. Web reviews influence on expectations and purchasing intentions of hotel potential customers. Int J Hosp Manag 2013 Sep;34:99-107. [doi: 10.1016/j.ijhm.2013.02.012]
82. Arendt F, Forrai M, Findl O. Dealing with negative reviews on physician-rating websites: an experimental test of how physicians can prevent reputational damage via effective response strategies. Soc Sci Med 2020 Dec;266:113422 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113422] [Medline: 33065499]

Abbreviations

AVE: average variance extracted
CR: composite reliability
HTMT: heterotrait-monotrait
OPR: online physician review
PLS-SEM: partial least squares–structural equation modeling
VIF: variance inflation factor

XSL•FO
RenderX

Edited by G Eysenbach; submitted 22.02.23; peer-reviewed by N Mungoli, Z Sijia, H Lu; comments to author 13.09.23; revised version received 03.10.23; accepted 17.02.24; published 12.03.24.

Please cite as:
Han X, Lin Y, Han W, Liao K, Mei K
Effect of Negative Online Reviews and Physician Responses on Health Consumers' Choice: Experimental Study
J Med Internet Res 2024;26:e46713
URL: https://www.jmir.org/2024/1/e46713
doi:10.2196/46713
PMID:38470465

©Xi Han, Yongxi Lin, Wenting Han, Ke Liao, Kefu Mei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Youth is Prized in Medicine, Old Age is Valued in Law: Analysis of Media Narratives Over 200 Years

Reuben Ng[1,2], PhD; Nicole Indran[1], BSocSci (Hons)

[1]Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore, Singapore

[2]Lloyd's Register Foundation Institute for the Public Understanding of Risk, National University of Singapore, Singapore, Singapore

Corresponding Author:
Reuben Ng, PhD
Lee Kuan Yew School of Public Policy
National University of Singapore
469C Bukit Timah Road
Singapore, 259772
Singapore
Phone: 65 66013967
Email: spprng@nus.edu.sg

Abstract

Background: This is the first study to explore how age has influenced depictions of doctors and lawyers in the media over the course of 210 years, from 1810 to 2019. The media represents a significant platform for examining age stereotypes and possesses tremendous power to shape public opinion. Insights could be used to improve depictions of older professionals in the media.

Objective: This study aims to understand how age shapes the portrayals of doctors and lawyers. Specifically, it compares the difference in sentiments toward younger and older doctors as well as younger and older lawyers in the media over 210 years.

Methods: Leveraging a 600-million-word corpus of American media publications spanning 210 years, we compiled top descriptors (N=478,452) of nouns related to youth × occupation (eg, younger doctor or physician) and old age × occupation (eg, older lawyer or attorney). These descriptors were selected using well-established criteria including co-occurrence frequency and context relevance, and were rated on a Likert scale from 1 (very negative) to 5 (very positive). Sentiment scores were generated for "doctor/physician," "young(er) doctor/physician," "old(er) doctor/physician," "lawyer/attorney," "young(er) lawyer/attorney," and "old(er) lawyer/attorney." The scores were calculated per decade for 21 decades from 1810 to 2019. Topic modeling was conducted on the descriptors of each occupation in both the 1800s and 1900s using latent Dirichlet allocation.

Results: As hypothesized, the media placed a premium on youth in the medical profession, with portrayals of younger doctors becoming 10% more positive over 210 years, and those of older doctors becoming 1.4% more negative. Meanwhile, a premium was placed on old age in law. Positive portrayals of older lawyers increased by 22.6% over time, while those of younger lawyers experienced a 4.3% decrease. In the 1800s, narratives on younger doctors revolved around their participation in rural health care. In the 1900s, the focus shifted to their mastery of new medical technologies. There was no marked change in narratives surrounding older doctors from the 1800s to the 1900s, though less attention was paid to their skills in the 1900s. Narratives on younger lawyers in the 1800s referenced their limited experience. In the 1900s, there was more focus on courtroom affairs. In both the 1800s and 1900s, narratives on older lawyers emphasized their prestige, especially in the 1900s.

Conclusions: Depending on the occupation, one's age may either be seen as an asset or a liability. Efforts must be expended to ensure that older professionals are recognized for their wealth of knowledge and skills. Failing to capitalize on the merits of an older workforce could ultimately be a grave disservice not only to older adults but to society in general.

(J Med Internet Res 2024;26:e45855) doi:10.2196/45855

KEYWORDS

older professionals; ageism; media; historical analysis; reframe aging; learned professions; psychomics

XSL•FO
RenderX

Introduction

Due to medical advances, older adults today are much healthier than before and make up a substantial portion of the labor market. The share of individuals aged 55 years and older either working or actively seeking employment has increased significantly since the 1990s, with approximately 37 million older Americans in the workforce as of March 2021 [1]. Never in history have older people been more integral to the workforce, and projections indicate a continuation of this trend [2]. Exploring how older workers have been portrayed in the media is essential for fostering a society that values its older population. In this study, we compare the portrayals of younger and older practitioners in the legal and medical spheres over the last 210 years.

The significance of our study lies in both conceptual and practical domains. From a conceptual standpoint, this study is one of the first to examine how portrayals of older workers in the media have changed over the span of 210 years. Existing research has looked only at how older adults in general are stereotyped in the media [3]. By analyzing the portrayals of older workers over such an extensive period, this study facilitates a more comprehensive understanding of how various social, cultural, or industry-related factors may have influenced these portrayals. This will in turn allow for better contextualization of current attitudes toward older workers. From a practical standpoint, insights from this study can inform interventions aimed at combating negative age stereotypes of older workers in the media.

The theory of social constructionism maintains that reality is neither objective nor fixed, but rather is constructed through social and cultural processes [4]. As a key agent for transmitting and disseminating information, the media plays a salient role in shaping the way reality is interpreted. According to the agenda-setting theory, the media molds public opinion by determining which issues are considered important and worthy of attention [5,6]. Similarly, cultivation theory posits that repeated exposure to certain messages in the media can alter people's attitudes and beliefs over time [7]. Given the power of the media to influence the collective conscience, it is critical to inspect the ways older workers have been depicted in the media.

Several studies have looked at how older people are portrayed in the media. Ng et al [3] discovered that age stereotypes in the United States have become more negative over the last 2 centuries. Meanwhile, some have observed that depictions of older persons in visual media have become more positive in Europe and America since the 1950s [8,9]. To date, there is a paucity of literature on how different subgroups of older people are stereotyped in the media. The existing studies have focused primarily on the portrayal of grandparents [10], with little attention dedicated to the stereotypes of older workers.

Over the decades, scholars have endeavored to dissect the nature of age stereotypes. Although there is copious evidence of a general negative bias in societal perceptions of older adults [3,11-15], age stereotypes are widely accepted as being multifaceted [16,17]. Negative age stereotypes include being frail and unfriendly, while positive ones include being warm and amiable. Psychologists have argued that the stereotypes applied to a given target tend to shift across contexts as only contextually meaningful information will be used to evaluate the individual [18,19]. This contextual malleability of stereotypes renders it a scholarly imperative to pinpoint the various circumstances in which positive and negative age stereotypes emerge.

Presently, most studies pertaining to older workers focus on age discrimination as well as stereotypes in the workplace [20,21]. Positive stereotypes of older workers class them as warm, reliable, and committed to the job [22,23], while negative ones include being less flexible and adaptable [23,24]. Meanwhile, even as younger workers are highly regarded in terms of physical ability, productivity, and creativity [23], they are often branded as being inexperienced, unreliable, and unmotivated [25].

The ways in which older workers are represented in the media can affect the public's attitudes toward old age and in turn the health of older persons. According to stereotype embodiment theory, the assimilation of age stereotypes into one's self-concept can affect one's health [26]. Negative age stereotypes are linked to poorer health outcomes such as a reduced sense of self-efficacy, a higher risk of depression as well as poorer immune or cardiovascular health [26-29]. Conversely, positive age stereotypes are associated with improved functional health, well-being, and longevity [26,28,29].

Research on how society evaluates older doctors is lacking. The few studies that have been carried out examined patients' preferences regarding the age of their physicians [30-33]. Although older physicians are commonly deemed more patient and reassuring [30,31], some believe they are more susceptible to dispensing a lower quality of care than their younger colleagues [32]. A recent study found that patients cared for by older practitioners had higher mortality rates than those treated by younger ones [34]. In the legal sphere, minimal attention has been devoted to examining the ways in which older lawyers are stereotyped. That said, prior scholarship has hinted at both the opportunities and challenges posed by older attorneys. On one hand, there has been discourse on the need to leverage the skills of older attorneys, be it by allowing them to register for emeritus pro bono status [35] or by offering them opportunities to train younger attorneys [36]. On the other hand, there have been fears that older attorneys' ability to continue their practice may be hampered by cognitive changes [35].

We test 2 hypotheses. First, to determine how age affects portrayals of doctors and lawyers in the media, we compare the difference in sentiments toward older and younger doctors, as well as older and younger lawyers. We hypothesize that the media will place a premium on youth in the medical profession where physical dexterity—a trait typically associated with younger people [37]—is prized across a range of specialties [38,39]. Specifically, we hypothesize that sentiments toward younger doctors will be more positive than toward older doctors over 21 decades in the media (hypothesis 1). Second, we hypothesize that the media will place a premium on old age in

XSL•FO
RenderX

the legal profession due to its inherent association with experience—a quality of paramount importance in the legal domain [35]. Specifically, we hypothesize that sentiments toward older lawyers will be more positive than toward younger lawyers over the same period (hypothesis 2).

Methods

Data Set

Following earlier work [10,40-42], we created the largest historical corpus—comprising 600 million words—of American media publications spanning 210 years (1810 to 2019) by merging the Corpus of Historical American English (COHA) from 1810 to 2009 with the Corpus of Contemporary American English (COCA) from 2010 to 2019 [43]. The combination of both corpora formed the largest structured historical English corpus with over 150,000 texts collected from newspapers, magazines, fiction, and nonfiction. Publications were extracted from major news outlets such as the *New York Times*, *Wall Street Journal*, and *USA Today*, as well as smaller ones like the *Atlantic Journal Constitution* and the *San Francisco Chronicle*. Material that was published at some point throughout the 210 years but that has ceased publication was included in the data set. The media represents a significant platform for examining age stereotypes as it possesses tremendous power to shape public opinion [4-6].

Target Nouns to Measure Occupation and Age × Occupation Stereotypes

Occupation

The Harris Poll [44] listed doctors (physicians) and lawyers (attorneys) as some of the most respected professions. Prior studies have also found that doctors (physicians) and lawyers (attorneys) are seen as some of the most functionally significant occupations in society [45]. To ensure the relevance and applicability of our findings, the terms we selected for analysis are those frequently used in American media to describe professionals in the legal and medical fields. These terms are “doctor,” “physician,” “lawyer,” and “attorney.”

Old Age × Occupation

The descriptors or adjectives for the following terms were compiled: “old doctors,” “older doctors,” “old physicians,” “older physicians,” “old lawyers,” “older lawyers,” “old attorneys,” and “older attorneys.” We took into consideration the fact that older workers may be referred to by other adjectives such as “aging,” “elder,” “elderly,” and “aged.” However, we used “old” and “older” as they evidenced the highest prevalence in the data set.

Youth × Occupation

The descriptors or adjectives for the following terms were compiled: “young doctors,” “younger doctors,” “young physicians,” “younger physicians,” “young lawyers,” “younger lawyers,” “young attorneys,” and “younger attorneys.” Although younger workers may be referred to by other adjectives including “junior,” we used “young” and “younger” to exclude words describing workers of a lower status.

Selection of Descriptors and Sentiment Scoring

The top descriptors that co-occurred most frequently—referred to as “collocates”—with each term were compiled per decade for 210 years based on the following inclusion criteria: (1) The collocate was present within 6 words before or after the target word (lexical proximity). Articles such as “the” and “a” were not included in the 6-word lexical span. If the target noun was the first word of a sentence, collocates from the preceding sentence were excluded. (2) The collocate referred to an older or younger person specifically (relevant context). (3) There was a mutual information score of 1.5 and above, which suggests semantic bonding, meaning that the collocate has a stronger association with the particular synonym than other words in the corpus [46]. We use the following formula to calculate the mutual information score:

“A” indicates the possibility of the target word A appearing, which is calculated by the frequency of the target word. “B” indicates the possibility of the collocate B appearing, which is calculated by the frequency of word B. “C” indicates the possibility of “A” and “B” appearing together, which is calculated by the frequency of collocate B appearing near the target word A. “SizeCorpus” refers to the size of the corpus or the number of words. Span is the span of words, that is, if there are 6 words to the left and 6 words to the right of the target word, span=12; log (2)=0.30103. This is an application of concordance analysis called “psychomics”, which has been used in past literature to analyze societal stereotypes [10,47-50]. This rigorous process culminated in 478,452 collocates (descriptors or adjectives).

To test both hypotheses, each collocate was rated using a sentiment engine on a scale from 1 (very negative) to 5 (very positive). This has proven to be a valid and reliable method of measuring words associated with age stereotypes [51] and follows previous corpus-based analyses [52]. Very negative collocates were rated 1 (eg, “frail” and “burden”), neutral collocates were rated 3 (eg, “transport” and “society”), and very positive collocates were rated 5 (eg, “venerable” and “cheerful”). For every noun per decade, we tabulated a mean score that was then weighted (by the number of times the respective noun appeared in that decade) to determine the respective sentiment score.

Analytic Strategy

Hypothesis 1 states that sentiments toward younger doctors in the media are more positive than toward older ones, while hypothesis 2 states that sentiments toward older lawyers in the media are more positive than toward younger lawyers. Both hypotheses were tested by analyzing the respective sentiment trends over 210 years (1810 to 2019) and thereafter determining whether the respective slopes were significantly different. Topic modeling was conducted on the descriptors of each occupation (eg, older doctor) in both the 1800s and 1900s using latent Dirichlet allocation (LDA) [53]. By probabilistically grouping words into topics, LDA identifies latent topics and clusters of words that co-occur frequently [53]. All data preprocessing, text analytics, and statistical analyses were done in Python 3.7

XSL•FO
RenderX

and OriginLab Corporation's OriginPro 2019b (OriginLab Corporation).

Ethical Considerations

No ethical approval was sought for this study due to the publicly available and nonidentifiable nature of the data. Moreover, this study was exempted from ethics review as it involved a secondary analysis of publicly available material.

Results

Youth Premium in Depictions of Doctors in the Media (Hypothesis 1)

We tested whether there was a youth premium in depictions of doctors in the media. As hypothesized, younger doctors or physicians evidenced more positive societal sentiments than older doctors or physicians over 21 decades from 1810 to 2019. Younger doctors enjoyed a 10% increase in positive portrayals over 210 years (β=0.01385, *P*=.04), while older doctors experienced a 1.4% decline in sentiments over the same period, though this trend was not statistically significant (Figure 1). The difference across both slopes achieved statistical significance, $F_{1,36}$=4.5602, *P*=.04, providing support for hypothesis 1.

Figure 1. The youth premium in the medical profession over 2 centuries from 1810 to 2019. Portrayals of younger doctors or physicians in the media became 10% more positive across 21 decades, while those of older doctors became 1.4% more negative, though this latter trend was not statistically significant. The shaded regions refer to the 95% confidence band of the respective lines. The sentiment score (y-axis) ranges from 1 (most negative) to 5 (most positive).

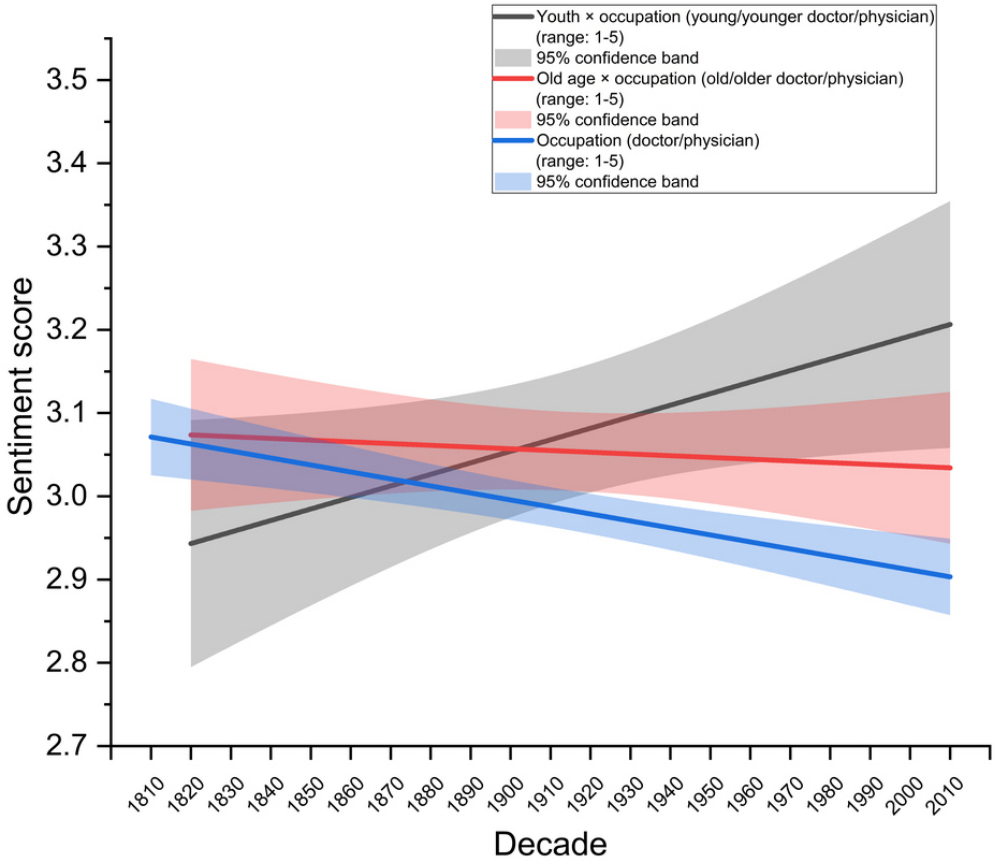

Old Age Premium in Depictions of Lawyers in the Media (Hypothesis 2)

We tested whether there was a premium of old age in depictions of lawyers in the media. As hypothesized, older lawyers or attorneys evidenced more positive sentiments compared with younger lawyers or attorneys across 21 decades from 1810 to 2019. Older lawyers experienced a 22.6% increase in positive portrayals over 21 decades (β=0.0302, *P*=.02). Conversely, younger lawyers experienced a 4.3% decline in sentiments over the same period (β=–0.00673, *P*≤.001). These results reflect an age premium for lawyers where older lawyers are portrayed more positively relative to younger ones (Figure 2). The difference across both slopes reached statistical significance, $F_{1,34}$=7.085, *P*=.01, supporting hypothesis 2.

XSL•FO
RenderX

Figure 2. The old-age premium in the legal fraternity over 2 centuries from 1810 to 2019. From 1810 to 2019, portrayals of older lawyers in the media became 22.6% more positive, compared with 4.3% more negative for younger lawyers. The shaded regions refer to the 95% confidence band of the respective lines. The sentiment score (y-axis) ranges from 1 (most negative) to 5 (most positive).

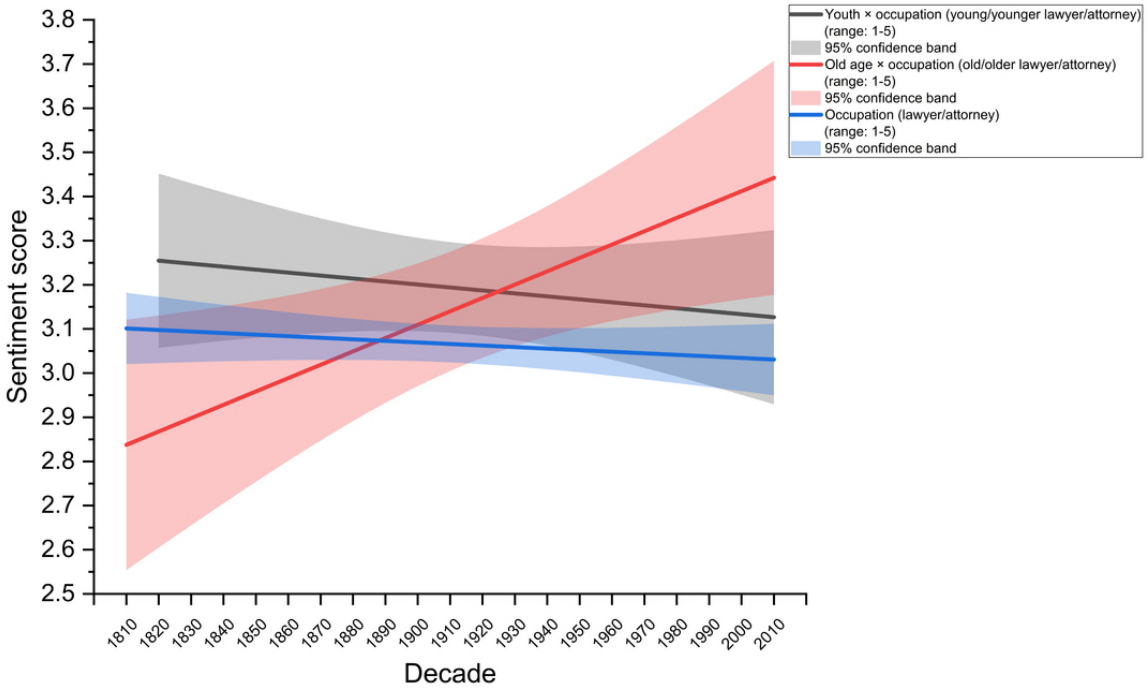

Summary of Insights from Topic Modeling

Topics were generated for old age × occupation and youth × occupation framing in the 1800s and 1900s. In the 1800s, narratives on younger doctors revolved around their participation in rural health care. In the 1900s, the focus shifted to their mastery of new medical technologies. Unlike younger doctors, there was no marked change in narratives surrounding older doctors from the 1800s to 1900s, though less attention was paid to their skills in the 1900s. Narratives on younger lawyers in the 1800s described them as having limited experience. In the 1900s, there was more focus on courtroom affairs. In both the 1800s and 1900s, narratives on older lawyers emphasized their prestige. This emphasis was more apparent in the 1900s as their seniority and experience became increasingly valuable in navigating the growing complexity of court cases. These results are summarized in Table 1.

XSL•FO
RenderX

Table 1. Summary of narratives in the media for old age × occupation (doctors or physicians and lawyers or attorneys) and youth × occupation (doctors or physicians and lawyers or attorneys) in the 1800s and 1900s.

Framing	1800s	1900s
Youth × occupation framing (doctors or physicians)	• Medical degree • Interactions with patients • Religion and medicine • Sickness and death in the hospital • Diagnosis and prescription • Rural health care	• Medical knowledge • High-quality care • Sickness and death in the hospital • Treatment and recovery • Financial woes • Consultations and medical appointments
Old age × occupation framing (doctors or physicians)	• Skill and expertise • Visits to the hospital • Religion and medicine • Sickness and death in the hospital	• Experience • Other occupations • Diagnosis and prescription • Sickness and death in the hospital • Surgery • Bedside manner • Royal visits
Youth × occupation framing (lawyers or attorneys)	• Prestige • Law and religion • Judicial district • Prosecution • Limited experience	• Courtroom affairs • Trial and prosecution • Politics • Lawsuits against politicians • Legal fees • Other occupations
Old age × occupation framing (lawyers or attorneys)	• Prestige • Legal guidance • Prosecution • Legal fees	• Seniority and experience • Courtroom proceedings • Legal guidance • Bar • Lawsuits against politicians • Family law

Youth × Occupation Framing (Doctors or Physicians) in the 1800s

Topic 1 consists of terms pertaining to obtaining a “medical degree” (“college,” “degree,” and “professor”). Topic 2 touches on “interactions with patients” (“laugh,” “smile,” and “boy”). Collocates alluding to “religion and medicine” are in topic 3 (“reverend,” “pray,” and “advise”). Topic 4 is about “sickness and death in the hospital” (“sick,” “dead,” and “fever”) while topic 5 is about “diagnosis and prescription” (“prescribe,” “medicine,” and “care”). Topic 6 focuses on “rural healthcare” (“village,” “horse,” “ride,” and “inquire”).

Youth × Occupation Framing (Doctors or Physicians) in the 1900s

Collocates in topic 1 paint younger physicians as having up-to-date “medical knowledge” and being proficient in the use of newer medical technologies (“treatment,” “medical,” “informatics,” and “technology”). Those in topic 2 hint at how younger doctors dispense “high-quality care” (“better” and “care”). Topic 3 is about “sickness and death in the hospital” (“sick,” “disease,” “and “die”). Topic 4 dwells on “treatment and recovery” (“treatment,” “cure,” and “best”). Topic 5 looks at the “financial woes” confronted by some patients (“shoulder,” “worry,” “bill,” and “enough”) and topic 6 at “consultations and medical appointments” (“consult,” “advise,” “and “visit”).

Old Age × Occupation Framing (Doctors or Physicians) in the 1800s

Topic 1 describes the “skill and expertise” of older doctors (“skill” and “experience”). Topic 2 covers “visits to the hospital” (“visit,” “consult,” and “opinion”). Topic 3 makes reference to “religion and medicine” (“divinity,” “village,” and “medicine”) and topic 4 to “sickness and death in the hospital” (“die” and “trouble”).

Old Age × Occupation Framing (Doctors or Physicians) in the 1900s

Topic 1 deals with the wealth of “experience” of older doctors (“experience” and “practitioner”). Comparisons of doctors with “other occupations” (“lawyer,” “surgeon,” and “dentist”) are in topic 2. Topic 3 is about “diagnosis and prescription” (“prescription,” “drug,” and “diagnosis”) and topic 4 is about “sickness and death in the hospital” (“disease” and “dead”). Topic 5 pertains to “surgery” (“operation,” “perform,” and “abortion”). Topic 6 looks at “bedside manner” (“laugh” and “smile”) and topic 7 involves “royal visits” (“royal,” “visit,” and “news”).

Youth × Occupation Framing (Lawyers or Attorneys) in the 1800s

Words in topic 1 portray the legal profession as one of “prestige” (“eminent,” “distinguished,” and “prominent”). Those in topic

XSL•FO
RenderX

2 concern "law and religion" ("clergyman," "judge," and "divine"). Topic 3 features terms related to the "judicial district" ("judicial," "district," "jury," and "statesman") and topic 4 features terms related to "prosecution" ("prosecute" and "defend"). Collocates in topic 5 imply that younger attorneys have "limited experience" ("less," "experience," and "knowledge").

Youth × Occupation Framing (Lawyers or Attorneys) in the 1900s

Topic 1 comprises terms related to "courtroom affairs" ("affairs," "drama," "judge," and "court"). Topic 2 circles around issues regarding "trial and prosecution" ("plead," "charge," and "criminal"). Matters regarding "politics" dominate topic 3 ("republican," "democratic," and "senate"). Specific "lawsuits against politicians" are in topic 4 ("settlement," "lawsuit, "allegation," and "complaint"). Topic 5 focuses on "legal fees" ("legal," "insurance," "interest," and "cost") and topic 6 compares the legal profession to "other occupations" ("accountant," "banker," "doctor," and "teacher").

Old Age × Occupation Framing (Lawyers or Attorneys) in the 1800s

Topic 1 depicts older lawyers as having a certain level of "prestige" ("eminent" and "respectable"). Topic 2 revolves around "legal guidance" ("opinion" and "advice"). Topic 3 covers "prosecution" ("prosecute," "evidence," and "judge") and topic 4 contains terms related to "legal fees" ("interest," "dollar," and "fee").

Old Age × Occupation Framing (Lawyers or Attorneys) in the 1900s

The idea that older attorneys have more "seniority and experience" is in topic 1 ("senior," "experience," "successful," and "complex"). Topic 2 discusses matters regarding "courtroom proceedings" ("judge," "defendant," and "trial"). Topic 3 is related to "legal guidance" ("hire" and "represent"). Topic 4 is about the "bar" ("association" and "conference") and topic 5 is about "lawsuits against politicians" ("file," "governor," and "republican"). Topic 6 involves "family law" ("divorce," "witness," and "son").

Discussion

In this study, we compared how sentiments toward older and younger doctors and lawyers have changed in the media over the last 2 centuries. Findings reveal that media outlets have placed a premium on youth in the medical enterprise but on old age in the legal fraternity.

Narratives on younger doctors in the media are more positive than on older doctors. Additionally, while narratives on older doctors have become more negative, narratives on younger doctors are experiencing the reverse trend. One possible reason for these trends is the rise of new medical technologies and techniques that younger doctors are usually assumed to be more conversant with. Moreover, medical practitioners are said to be the most proficient in the years immediately after completing residency training. There is a common belief that doctors further from training may rely on out-of-date clinical evidence and not adhere as rigidly to evidence-based guidelines as younger doctors [33]. As seen in our LDA results, from the 1900s, the media began emphasizing the merits of younger physicians by highlighting their up-to-date medical knowledge. Although there were portrayals of older doctors as being skilled in the 1800s, this changed in the 1900s, during which more attention was dedicated to their experience and bedside manner. Thus, there may have been a subconscious association of youth with medical competence in the media.

Narratives on older lawyers in the media are more positive than those on younger lawyers. While narratives on older lawyers have become more positive, the trend is the opposite for younger lawyers. Younger attorneys may be seen as more enthusiastic and willing to learn, but they may also be viewed as lacking the expertise of their older and more seasoned counterparts. The increasing specialization of the legal profession over time [54] may have resulted in age becoming more highly valued in the field. Older attorneys may be seen as possessing a deeper understanding of legal precedents and a greater ability to navigate the complexities and intricacies of legal matters [35]. Our LDA results uncovered that since the 1900s, older lawyers have been depicted in the media as having a wealth of expertise, which may render them better able to provide valuable legal guidance to clients and colleagues alike.

The issue of youth-directed ageism is beyond the ambit of this study. Nevertheless, it cannot go unacknowledged that narratives on younger lawyers have become increasingly negative. Over the years, clarion calls have been sounded for bullying in the legal profession to end. In 2018, the International Bar Association surveyed 7000 lawyers from 135 countries on the topic of bullying and sexual harassment in the profession [55]. Results from the survey lent empirical support to claims about the rampancy of bullying and sexual harassment in the legal enterprise, which may not be particularly shocking in view of the adversarial, hierarchical, and hypercompetitive nature of the job [55]. Younger legal professionals were found to be disproportionately affected by bullying, and respondents from the United States reported higher rates of both bullying as well as sexual harassment than the global average [55]. This may account for the increase in negativity associated with narratives on younger attorneys.

In line with social constructionism [4], agenda-setting [5,6], and cultivation theories [7], the media is instrumental in shaping public perceptions of older adults. Depictions of younger physicians as being well acquainted with the newest medical technologies, though particularly important as they begin their medical careers, may come at the expense of older doctors. The public may be conditioned to view younger doctors as superior to older ones, which could consequently foment ageist stereotypes. The fact that depictions of older lawyers have grown more positive over time is heartening as it shows that older lawyers are being celebrated for their seniority and expertise.

The finding that an age premium exists in depictions of lawyers but not in depictions of doctors is interesting. It may be that the age of a physician is perceived as more important as work in the medical setting may have life-threatening consequences. Indeed, scholars have contended that older physicians' waning

XSL•FO
RenderX

physical skills may pose major surgical risks [38,39]. Such risks may be thought to outweigh the potential lessons to be imparted by older doctors. Physical dexterity has comparatively little—if any—bearing on whether one is able to perform one's legal duties effectively, which may explain the differing results between portrayals of older doctors and lawyers.

This study yields some important insights. First, it is important that narratives on older adults in the media accurately mirror the contributions of older adults to society. More attention should be dedicated to the immense value they add to the workplace and economy in order that negative stereotypes are progressively removed from one's cognitive repertoire. Just as the media has highlighted the strengths of older lawyers, it may be worthwhile to increase media coverage of the unique strengths of older doctors and what they can bring to patient care. Media campaigns could be held to promote positive images of older doctors. Such campaigns could feature interviews with these doctors, details of their achievements, and testimonials from patients who have received exceptional care from them. Additionally, media outlets could offer training programs for journalists to sensitize them to the issue of ageism and to equip them with guidelines on writing about older doctors tactfully. To avoid perpetuating ageist stereotypes and to promote better psychological well-being among older persons [26], it is crucial to ensure accurate depictions of older professionals in the media. As only a small fraction of older professionals are covered in the media, the few depictions that do exist hold significant influence in shaping public attitudes toward aging.

Second, our results indicate that depending on the occupation, old age may be deemed more valuable than youth. There is therefore a need to explore how the merits of old age—skill, knowledge, and experience—in these occupational contexts could be extended to the way the older population is viewed as a whole. To this end, it may be useful to encourage more interaction between older and younger people in the workplace. Age diversity may translate into advantages in both human and social capital [56] as knowledge is passed from young to old and old to young [57]. Relatedly, it is imperative that older persons are eased into the retirement phase and not tossed into it unceremoniously. Opportunities to continue contributing to the workplace—such as through mentorship programs—should be provided so as to harness the treasury of skills and knowledge of this cohort. This will also allow older persons to transition more seamlessly into the next phase of life.

This study has various limitations. First, we acknowledge that an emphasis on occupational roles may wind up perpetuating the notion that older people are to be valued only if they remain economically productive [58]. However, our findings reveal that old age may evoke positive stereotypes in certain occupational contexts, which has implications on how negative stereotypes of older adults can be eliminated. Second, as only 2 occupational roles were used for analysis, the differences in portrayals of older and younger workers in other occupational settings remain an uncharted area of research. Furthermore, the occupations considered in this study are generally considered to be prestigious. Future scholarship could examine whether old age is valued in the context of blue-collar occupations. Third, our search only included the 2 professions as umbrella terms without specifying the various specialties within each profession, such as cardiologists, ophthalmologists, criminal defense prosecutors, and tax attorneys. Thus, our study only lays out in broad strokes how older lawyers and doctors are depicted in the media. Future studies could address this limitation by delving into the various subcategories.

Fourth, the data set used in this study only comprises American sources. The meanings ascribed to age and to occupations are likely to differ across cultural contexts. For instance, the idea of the American dream may have had an impact on the portrayal of the legal profession since it could be tied closely to the notion of wealth and success. A study that investigates how portrayals of older doctors and lawyers vary across cultures may hence be worth pursuing. Finally, it is important to acknowledge that LDA may not account for the context in which the words are used, making it difficult to discern whether ageist portrayals are propagated by journalists or other sources. In addition, LDA may not detect nuances in language such as tone, sarcasm, and irony. Further research is therefore required to determine the generalizability of our findings. Other directions for future study include an analysis of how narratives pertaining to select occupations have unfolded since the 2000s. It is likely that topics related to artificial intelligence and automation will emerge in these narratives. Surveys [59], interviews [60], and big data analytics [61,62] could also be used to explore the types of stereotypes linked to professionals of different age groups.

Conclusions

In seeking to eradicate ageism, perhaps a pressing question is not so much whether this phenomenon exists, but rather in what situations it manifests itself and to what degree. This study has demonstrated that depending on the occupation, one's age may either be seen as an asset or a liability. Moving forward, effort must be expended to ensure that older professionals are recognized for their wealth of knowledge and skills. Failing to capitalize on the merits of an aging workforce could ultimately be a grave disservice not only to older adults but to society in general.

Acknowledgments

We are grateful to W. Yang for preprocessing the data. We gratefully acknowledge the support of the Commonwealth Fund's Harkness Fellowship in Heathcare Policy and Practice and the Social Science Research Council SSHR Fellowship (MOE2018-SSHR-004). The funders had no role in study design, data collection, analysis, writing, and decision to publish.

XSL•FO
RenderX

Data Availability

Data are publicly available at English-Corpora.org [63].

Authors' Contributions

RN designed the study, developed the methodology, analyzed the data, wrote the paper, and acquired the funding. NI cowrote the paper.

Conflicts of Interest

None declared.

References

1. U.S. Bureau of Labor Statistics.: FRED URL: https://fred.stlouisfed.org/series/LNS11024230 [accessed 2021-12-21]
2. United States Census Bureau. 2020 census will help policymakers prepare for the incoming wave of aging boomers. Census.gov. 2019. URL: https://www.census.gov/library/stories/2019/12/by-2030-all-baby-boomers-will-be-age-65-or-older.html [accessed 2021-10-20]
3. Ng R, Allore HG, Trentalange M, Monin JK, Levy BR. Increasing negativity of age stereotypes across 200 years: evidence from a database of 400 million words. PLoS One 2015;10(2):e0117086 [FREE Full text] [doi: 10.1371/journal.pone.0117086] [Medline: 25675438]
4. Berger PL, Luckmann T. The Social Construction of Reality: A Treatise in the Sociology of Knowledge, 1st Edition. Garden City, NY: Penguin Books; 1966.
5. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Opin Q 1972;36(2):176-187. [doi: 10.1086/267990]
6. Lippmann W. Public Opinion, 1st edition. New York, NY, USA: Macmillan; 1922.
7. Gerbner G, Gross L, Morgan M, Signorielli N, Shanahan J. Growing up with television: cultivation processes. In: Bryant J, Zillmann D, editors. Media Effects: Advances in Theory and Research, Second Edition. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2002:43-67.
8. Loos E, Ivan L. Visual ageism in the media. In: Ayalon L, Tesch-Römer C, editors. Contemp Perspect Ageism. Cham: Springer International Publishing; 2018:163-176.
9. Ylänne V. Representations of ageing in the media. In: Twigg J, Martin W, editors. Routledge Handbook of Cultural Gerontology, Routledge International Handbooks. London: Routledge; 2015:369-376.
10. Ng R, Indran N. Role-based framing of older adults linked to decreased ageism over 210 years: evidence from a 600-million-word historical corpus. Gerontologist 2022;62(4):589-597 [FREE Full text] [doi: 10.1093/geront/gnab108] [Medline: 34323967]
11. Ng R, Indran N, Liu L. Ageism on Twitter during the COVID-19 pandemic. J Soc Issues 2022;78(4):842-859 [FREE Full text] [doi: 10.1111/josi.12535] [Medline: 35942488]
12. Ng R, Indran N. Hostility toward baby boomers on TikTok. Gerontologist 2022 Sep 07;62(8):1196-1206 [FREE Full text] [doi: 10.1093/geront/gnac020] [Medline: 35106587]
13. Ng R, Indran N. Innovations for an aging society through the lens of patent data. Gerontologist 2024 Feb 01;64(2):gnad015 [FREE Full text] [doi: 10.1093/geront/gnad015] [Medline: 37497634]
14. Ng R, Indran N. Does age matter? tweets about gerontocracy in the United States. J Gerontol B Psychol Sci Soc Sci 2023 Nov 14;78(11):1870-1878. [doi: 10.1093/geronb/gbad113] [Medline: 37584274]
15. Ng R, Indran N. Questions about aging and later life on Quora. Gerontologist 2024 (forthcoming).
16. Hummert ML. Multiple stereotypes of elderly and young adults: a comparison of structure and evaluations. Psychol Aging 1990;5(2):182-193. [doi: 10.1037//0882-7974.5.2.182] [Medline: 2378684]
17. Kite ME, Stockdale GD, Whitley BE, Johnson BT. Attitudes toward younger and older adults: an updated meta-analytic review. J Soc Issues 2005;61(2):241-266. [doi: 10.1111/j.1540-4560.2005.00404.x]
18. Bodenhausen GV, Todd AR, Richeson JA. Controlling prejudice and stereotyping: antecedents, mechanisms, and contexts. In: Nelson TD, editor. Handbook of Prejudice, Stereotyping, and Discrimination. New York: Psychology Press; 2009:111-135.
19. Casper C, Rothermund K, Wentura D. The activation of specific facets of age stereotypes depends on individuating information. Social Cognition 2011;29(4):393-414. [doi: 10.1521/soco.2011.29.4.393]
20. Kleissner V, Jahn G. Implicit and explicit measurement of work-related age attitudes and age stereotypes. Front Psychol 2020;11:579155 [FREE Full text] [doi: 10.3389/fpsyg.2020.579155] [Medline: 33123059]
21. Ng TWH, Feldman DC. Evaluating six common stereotypes about older workers with meta-analytical data. Personnel Psychology 2012;65(4):821-858. [doi: 10.1111/peps.12003]
22. Hassell BL, Perrewe PL. An examination of beliefs about older workers: do stereotypes still exist? J Organ Behavior 2006;16(5):457-468. [doi: 10.1002/job.4030160506]
23. Van Dalen HP, Henkens K, Schippers J. Dealing with older workers in Europe: a comparative survey of employers' attitudes and actions. J Eur Soc Policy 2009;19(1):47-60. [doi: 10.1177/0958928708098523]

XSL•FO
RenderX

24. Gringart E, Helmes E, Speelman CP. Exploring attitudes toward older workers among Australian employers: an empirical study. J Aging Soc Policy 2005;17(3):85-103. [doi: 10.1300/J031v17n03_05] [Medline: 16219595]
25. Finkelstein LM, Ryan KM, King EB. What do the young (old) people think of me? content and accuracy of age-based metastereotypes. Eur JWork Organ Psychol 2013;22(6):633-657. [doi: 10.1080/1359432x.2012.673279]
26. Levy B. Stereotype embodiment: a psychosocial approach to aging. Curr Dir Psychol Sci 2009;18(6):332-336 [FREE Full text] [doi: 10.1111/j.1467-8721.2009.01662.x] [Medline: 20802838]
27. Auman C, Bosworth HB, Hess TM. Effect of health-related stereotypes on physiological responses of hypertensive middle-aged and older men. J Gerontol B Psychol Sci Soc Sci 2005;60(1):P3-P10 [FREE Full text] [doi: 10.1093/geronb/60.1.p3] [Medline: 15643036]
28. Levy BR, Hausdorff JM, Hencke R, Wei JY. Reducing cardiovascular stress with positive self-stereotypes of aging. J Gerontol B Psychol Sci Soc Sci 2000;55(4):P205-P213 [FREE Full text] [doi: 10.1093/geronb/55.4.p205] [Medline: 11584877]
29. Levy BR, Slade MD, Kunkel SR, Kasl SV. Longevity increased by positive self-perceptions of aging. J Pers Soc Psychol 2002;83(2):261-270. [doi: 10.1037//0022-3514.83.2.261] [Medline: 12150226]
30. Mckinstry B, Yang SY. Do patients care about the age of their general practitioner? a questionnaire survey in five practices. Br J Gen Pract 1994;44(385):349-351 [FREE Full text]
31. Usta J, Antoun J, Ambuel B, Khawaja M. Involving the health care system in domestic violence: what women want. Ann Fam Med 2012;10(3):213-220 [FREE Full text] [doi: 10.1370/afm.1336] [Medline: 22585885]
32. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med 2005;142(4):260-273 [FREE Full text] [doi: 10.7326/0003-4819-142-4-200502150-00008] [Medline: 15710959]
33. MacRae H. Not too old, not too young: older women's perceptions of physicians. Can J Aging 2015;34(4):545-560. [doi: 10.1017/S0714980815000367] [Medline: 26649892]
34. Tsugawa Y, Newhouse JP, Zaslavsky AM, Blumenthal DM, Jena AB. Physician age and outcomes in elderly patients in hospital in the US: observational study. BMJ 2017;357:j1797 [FREE Full text] [doi: 10.1136/bmj.j1797] [Medline: 28512089]
35. Frye KA, Oten B. Senior moments: an examination of the ethical and practical considerations of our aging bar. J Am Acad Matrim Lawyers 2019;31:371-404 [FREE Full text]
36. Barnes C. Time to go. helping lawyers retire with dignity. Am Bar Assoc 2008;32(6) [FREE Full text]
37. Constansia RDN, Hentzen JEKR, Buis CI, Klaase JM, de Meijer VE, Meerdink M. Is surgical subspecialization associated with hand grip strength and manual dexterity? a cross-sectional study. Ann Med Surg (Lond) 2022;73:103159 [FREE Full text] [doi: 10.1016/j.amsu.2021.103159] [Medline: 34976387]
38. Blasier RB. The problem of the aging surgeon: when surgeon age becomes a surgical risk factor. Clin Orthop Relat Res 2009;467(2):402-411 [FREE Full text] [doi: 10.1007/s11999-008-0587-7] [Medline: 18975041]
39. Greenfield LJ, Proctor MC. When should a surgeon retire? Adv Surg 1999;32:385-393. [Medline: 9891754]
40. Ng R, Indran N. Reframing aging: foregrounding familial and occupational roles of older adults is linked to decreased ageism over two centuries. J Aging Soc Policy 2023:1-18 [FREE Full text] [doi: 10.1080/08959420.2023.2238538] [Medline: 37534573]
41. Ng R, Indran N. Impact of old age on an occupation's image over 210 years: an age premium for doctors, lawyers, and soldiers. J Appl Gerontol 2023 Jun;42(6):1345-1355. [doi: 10.1177/07334648231155025] [Medline: 37092180]
42. Ng R, Indran N. Role-based framing of older adults linked to decreased ageism over 210 years: evidence from a 600-million-word historical corpus. Gerontologist 2022 Apr 20;62(4):589-597 [FREE Full text] [doi: 10.1093/geront/gnab108] [Medline: 34323967]
43. Davies M. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. Corpora 2012;7(2):121-157. [doi: 10.3366/cor.2012.0024]
44. Doctors, military officers, firefighters, and scientists seen as among America's most prestigious occupations. The Harris Poll. 2014. URL: https://tinyurl.com/4e72ez3u [accessed 2021-09-10]
45. Thielbar G, Feldman SD. Occupational stereotypes and prestige. Soc Forces 1969;48(1):64-72. [doi: 10.2307/2575470]
46. Church KW, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguist 1990;16(1):22-29 [FREE Full text] [doi: 10.1093/oso/9780199292332.003.0019]
47. Ng R, Indran N, Suarez P. Communicating risk perceptions through batik art. JAMA 2023;330(9):790-791. [doi: 10.1001/jama.2023.14278] [Medline: 37486662]
48. Ng R, Indran N, Yang W. Portrayals of older adults in over 3000 films around the world. J Am Geriatr Soc 2023 Sep;71(9):2726-2735. [doi: 10.1111/jgs.18400] [Medline: 37314013]
49. Ng R, Tan YW. Diversity of COVID-19 news media coverage across 17 countries: the influence of cultural values, government stringency and pandemic severity. Int J Environ Res Public Health 2021 Nov 09;18(22):11768 [FREE Full text] [doi: 10.3390/ijerph182211768] [Medline: 34831524]
50. Ng R, Chow TYJ, Yang W. The impact of aging policy on societal age stereotypes and ageism. Gerontologist 2022 Apr 20;62(4):598-606 [FREE Full text] [doi: 10.1093/geront/gnab151] [Medline: 34636402]

XSL•FO
RenderX

51. Levy B, Langer E. Aging free from negative stereotypes: successful memory in China and among the American deaf. J Pers Soc Psychol 1994;66(6):989-997. [doi: 10.1037//0022-3514.66.6.989] [Medline: 8046582]
52. Ng R, Indran N. Reframing aging during COVID-19: familial role-based framing of older adults linked to decreased ageism. J Am Geriatr Soc 2022;70(1):60-66 [FREE Full text] [doi: 10.1111/jgs.17532] [Medline: 34674224]
53. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text] [doi: 10.7551/mitpress/1120.003.0082]
54. Banks CP. The State and Federal Courts: A Complete Guide to History, Powers, and Controversy. Santa Barbara, CA: ABC-CLIO; 2017.
55. Pender K. Us too? bullying and sexual harassment in the legal profession. International Bar Association. 2019. URL: https://apo.org.au/node/248266 [accessed 2021-12-25]
56. Li Y, Gong Y, Burmeister A, Wang M, Alterman V, Alonso A, et al. Leveraging age diversity for organizational performance: an intellectual capital perspective. J Appl Psychol 2021;106(1):71-91. [doi: 10.1037/apl0000497] [Medline: 32202816]
57. Froidevaux A, Alterman V, Wang M. Leveraging aging workforce and age diversity to achieve organizational goals: a human resource management perspective. In: Czaja S, Sharit J, James J, editors. Current and Emerging Trends in Aging and Work. Cham: Springer International Publishing; 2020:33-58.
58. van Dyk S. The appraisal of difference: critical gerontology and the active-ageing-paradigm. J Aging Stud 2014;31:93-103. [doi: 10.1016/j.jaging.2014.08.008] [Medline: 25456626]
59. Sima LC, Ng R, Elimelech M. Modeling risk categories to predict the longitudinal prevalence of childhood diarrhea in Indonesia. Am J Trop Med Hyg 2013 Nov;89(5):884-891 [FREE Full text] [doi: 10.4269/ajtmh.12-0540] [Medline: 24019442]
60. Yu CC, Tan L, Tang B, Liaw SY, Tierney T, Ho YY, et al. The development of empathy in the healthcare setting: a qualitative approach. BMC Med Educ 2022 Apr 04;22(1):245 [FREE Full text] [doi: 10.1186/s12909-022-03312-y] [Medline: 35379249]
61. Ng R, Levy B. Pettiness: conceptualization, measurement and cross-cultural differences. PLoS One 2018;13(1):e0191252 [FREE Full text] [doi: 10.1371/journal.pone.0191252] [Medline: 29385157]
62. Giest S, Ng R. Big data applications in governance and policy. PaG 2018 Nov 21;6(4):1-4. [doi: 10.17645/pag.v6i4.1810]
63. English-Corpora.org. URL: https://www.english-corpora.org/

Abbreviations

COCA: Corpus of Contemporary American English
COHA: Corpus of Historical American English
LDA: latent Dirichlet allocation

Edited by T Leung, G Eysenbach; submitted 19.01.23; peer-reviewed by L Allen, G Myreteg; comments to author 27.02.23; revised version received 13.03.23; accepted 31.08.23; published 26.03.24.

Please cite as:
Ng R, Indran N
Youth is Prized in Medicine, Old Age is Valued in Law: Analysis of Media Narratives Over 200 Years
J Med Internet Res 2024;26:e45855
URL: https://www.jmir.org/2024/1/e45855
doi:10.2196/45855
PMID:38530338

©Reuben Ng, Nicole Indran. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining the Role of Physician Characteristics in Web-Based Verified Primary Care Physician Reviews: Observational Study

Neil K R Sehgal[1,2], ME; Benjamin Rader[2,3*], PhD; John S Brownstein[2,4*], PhD

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, United States
[2]Computational Epidemiology Group, Boston Children's Hospital, Boston, MA, United States
[3]Department of Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital, Boston, MA, United States
[4]Department of Pediatrics, Harvard Medical School, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Neil K R Sehgal, ME
Department of Computer and Information Science
University of Pennsylvania
3330 Walnut St
Philadelphia, PA, 19104
United States
Phone: 1 215 898 9672
Email: neilsehgal99@gmail.com

Abstract

Background: Doctor review websites have become increasingly popular as a source of information for patients looking to select a primary care provider. Zocdoc is one such platform that allows patients to not only rate and review their experiences with doctors but also directly schedule appointments. This study examines how several physician characteristics including gender, age, race, languages spoken in a physician's office, education, and facial attractiveness impact the average numerical rating of primary care doctors on Zocdoc.

Objective: The aim of this study was to investigate the association between physician characteristics and patient satisfaction ratings on Zocdoc.

Methods: A data set of 1455 primary care doctor profiles across 30 cities was scraped from Zocdoc. The profiles contained information on the physician's gender, education, and languages spoken in their office. Age, facial attractiveness, and race were imputed from profile pictures using commercial facial analysis software. Each doctor profile listed an average overall satisfaction rating, bedside manner rating, and wait time rating from verified patients. Descriptive statistics, the Wilcoxon rank sum test, and multivariate logistic regression were used to analyze the data.

Results: The average overall rating on Zocdoc was highly positive, with older age, lower facial attractiveness, foreign degrees, allopathic degrees, and speaking more languages negatively associated with the average rating. However, the effect sizes of these factors were relatively small. For example, graduates of Latin American medical schools had a mean overall rating of 4.63 compared to a 4.77 rating for US graduates ($P<.001$), a difference roughly equivalent to a 2.8% decrease in appointments. On multivariate analysis, being Asian and having a doctor of osteopathic medicine degree were positively associated with higher overall ratings, while attending a South Asian medical school and speaking more European and Middle Eastern languages in the office were negatively associated with higher overall ratings.

Conclusions: Overall, the findings suggest that age, facial attractiveness, education, and multilingualism do have some impact on web-based doctor reviews, but the numerical effect is small. Notably, bias may play out in many forms. For example, a physician's appearance or accent may impact a patient's trust, confidence, or satisfaction with their physician, which could in turn influence their take-up of preventative services and lead to either better or worse health outcomes. The study highlights the need for further research in how physician characteristics influence patient ratings of care.

(J Med Internet Res 2024;26:e51672) doi:10.2196/51672

XSL•FO
RenderX

KEYWORDS

patient review websites; patient online review; telemedicine; internet; online review; online reviews; rating; physician review; physician reviews; doctor review; doctor reviews

Introduction

There has been a growing interest in understanding the factors that may influence a patient's perception of their physician and how these perceptions might in turn impact quality of care. Research has shown that patients often rely on nonverbal cues, such as a physician's attire, to form their opinions about their health care provider [1-6]. With the advent of various web-based review platforms, such as Healthgrades and Zocdoc, patients have easy access to a wealth of information about their health care providers, as well as a relatively straightforward means to express their satisfaction. Research suggests that these web-based ratings do influence patients' choices of medical providers [7-9]. On the web-based physician review platform Zocdoc, a half-star improvement in ratings on a scale of 1-5 stars leads to a 10% increase in appointments [10,11].

Past research is mixed on the factors influencing these web-based reviews. Some research suggests that patient experience and clinical quality directly affect web-based ratings [12,13]. Other research suggests that there exists minimal correlation between the quality and value of care or peer-assessed performance with web-based ratings, and that nonphysician characteristics such as staff friendliness and appointment wait times are instead the key determinants [14-16].

What remains clear, however, is that many different factors play a role [17,18]. Web-based reviews for female physicians, for example, have been found to be more emotional and informal compared to those for male physicians [19]. Moreover, female physicians receive lower web-based ratings than their male counterparts on some platforms, though this trend is not consistent across all platforms [19-23]. Other research finds that patient-physician racial concordance is associated with higher scores on internal patient satisfaction surveys [24]. One study of a direct-to-consumer telemedicine platform shows that, on average, patients report higher rates of dissatisfaction with Black and Asian physicians compared to White physicians [25].

However, the influence of physician factors on physician ratings is likely not limited to gender and race alone. Despite the growing influence of web-based review sites, much remains unknown about the various factors that may influence physician ratings. For instance, many studies have examined the association between patient attractiveness and physician care practices [26-29]. However, to our knowledge, no study has examined how physician attractiveness impacts patient satisfaction in real-world medical settings.

Other factors such as school ranking and multilingualism are similarly understudied. Existing research finds little correlation between medical school ranking and performance scores [30]. In addition, language-concordant care is associated with increased patient satisfaction [31,32]. However, the roles of ranking and multilingualism have not been examined in web-based review settings.

In this paper, we aim to explore the role of various physician characteristics, including gender, age, race, number of languages spoken in a physician's office, education, and facial attractiveness, in shaping patient perceptions of their physicians through ratings on a web-based doctor review site.

Methods

We collected primary care doctor profile data from Zocdoc, a platform that allows patients to rate and review their experiences with doctors. Physician variables included facial attractiveness, race, age, gender, language, and education. Outcome variables included overall satisfaction, bedside manner, and wait time scores from patient reviews. Descriptive statistics, the Wilcoxon rank sum test, and multivariate logistic regression were used to analyze the data.

Data Collection

We collected a data set of primary care doctor profiles from Zocdoc, a web-based platform that allows patients to search for and schedule telemedicine and in-person appointments with doctors, and also to rate and review their experiences. Notably, unlike many other sites that host physician reviews, Zocdoc only posts reviews from patients who have attended an appointment [33]. After appointments, patients receive emails from Zocdoc asking for feedback. Zocdoc is free for patients to use, and makes money by charging health care providers subscription or booking fees [34]. Health care providers may come from both independent practices and integrated networks. Patients can search for a physician using criteria such as condition, specialty, city, state, ZIP code, insurance carrier and plan, or a specific doctor's name. They can then filter by date, time of day, distance, gender, in-person or video consultations, hospital affiliation, and languages spoken by the physician.

Zocdoc does not maintain a public-centralized list of all providers on the platform. In order to systematically collect a large sample of providers, we used Zocdoc's search engine to identify primary care doctors within 50 miles of each of the 30 largest cities in the United States [35]. Specifying primary care doctor" in the search bar yields a variety of specialties, including internists, nurse practitioners, and physician assistants. We filtered these results specifically to profiles that listed their specialty as primary care doctor. We collected these profiles using a browser-based web scraping tool [36].

Physician Variables

From each profile, we collected the listed gender, languages spoken in the physician's office, type of medical school degree (allopathic or osteopathic), medical school institution, and downloaded the physician's profile picture. All physicians were either graduates of allopathic medical schools, which grant doctor of medicine degrees, or osteopathic medical schools, which grant doctor of osteopathic medicine degrees. We determined the geographical location of each medical school (Africa, the Caribbean, East Asia, Europe, Latin American, the

XSL•FO
RenderX

Middle East, South Asian, the mainland United States or Canada, other) by searching for each medical school on Google. The US medical schools were additionally coded for whether or not they were ranked in the top 30 of the US News & World Report's 2023-2024 Medical School Research rankings [37]. For analysis purposes, we divided the number of languages spoken into buckets (1, 2, 3, and ≥4). We additionally broke down the number of languages into 7 categories: European, East and Southeast Asian, South Asian, Middle Eastern, African, Caribbean, and Creole (Table S1 in Multimedia Appendix 1).

Through Face++, a commercial facial analysis software commonly used to infer demographic factors, we imputed a facial attractiveness rating for each profile [38-42]. Face++'s estimation of facial attractiveness has been found to correlate well (r=0.72) with human raters [42]. We use a second commonly used commercial facial analysis software, Kairos, to infer race and age, which has been found to outperform Face++'s race and age models [43-46]. One study from 2019 found the mean absolute error for Kairos' age model to be ±3.30 (SD 2.64) years when compared to human raters' estimates as ground truth [43]. Accuracy for Kairos' race model was 95.06% (95% CI 94.08%-95.93%) when compared to human raters [43]. Similar methods have been used to understand the diversity in hospital system faculty and medical editorial boards [47,48]. To examine potential nonlinearities and for interpretability purposes, we divided facial attractiveness and age into quartiles. For robustness, we compared Face++'s age prediction results with those of Kairos (Tables S2-S3 in Multimedia Appendices 2 and 3).

Outcome Variables

Each doctor profile was associated with 3 average scores from patient reviews: overall satisfaction, bedside manner, and wait time. Each score is rated on a scale of 1-5 stars. We additionally collected the total number of reviews. Profiles with 0 reviews were removed from the sample.

Statistical Analysis

The data are analyzed at the physician level. We used descriptive statistics to summarize the data and the Wilcoxon rank sum test to determine if there were significant differences in average review scores across gender, age quartile, facial attractiveness quartile, education, and number of languages spoken. Multivariable logistic regression was used to assess the association between these factors and average review scores. The outcome for the regression is a binary variable indicating whether or not the physician is above the 25th percentile of ratings to understand factors associated with very low ratings. All statistical analyses were conducted using R (version 4.1.2; The R Foundation).

Ethical Considerations

This study used publicly available data posted for public use by the providers, and therefore, did not require institutional review board approval.

Results

A total of 1521 primary care doctor profiles from Zocdoc were collected. Of these, 66 were missing an overall satisfaction rating and excluded, leaving a sample of 1455 primary care doctor profiles for analysis (Table 1). One profile lacked information on the medical school attended, one profile contained a profile picture from which facial attractiveness, age, and race could not be estimated, and one profile contained a profile picture for which only facial attractiveness could not be assessed. These 3 profiles were still included for analysis. Three cities—Houston, Chicago, and New York City—represented 49% of all primary care doctors in the sample (Table S4 in Multimedia Appendix 4). The majority of physicians were men (54%) with no nonbinary individuals represented in the study. The sample included speakers of 66 languages, with a mean of 1.73 languages spoken (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Characteristics of Zocdoc primary care profiles (N=1455[a]).

Characteristics	
Overall satisfaction rating, median (IQR)	4.82 (4.67-4.93)
Wait time rating, median (IQR)	4.67 (4.45-4.83)
Bedside manner rating, median (IQR)	4.88 (4.75-4.98)
Number of reviews, median (IQR)	57 (19-230)
Gender, n (%)	
Female	675 (46)
Male	780 (54)
Age (in years), mean (SD)	31 (8)
Race, n (%)	
Asian	598 (41)
Black	107 (7.4)
Hispanic	149 (10)
White	600 (41)
Facial attractiveness, mean (SD)	56 (12)
Top 30 medical school, n (%)	106 (7.3)
Degree, n (%)	
DO[b]	279 (19)
MD[c]	1176 (81)
Medical school location, n (%)	
Africa	28 (1.9)
Caribbean	141 (9.7)
East and Southeast Asia	40 (2.8)
Europe	68 (4.7)
Latin America	59 (4.1)
Middle East	38 (2.6)
Other	3 (0.2)
South Asia	180 (12)
United States or Canada	897 (62)
Number of languages spoken in office, n (%)	
1	858 (59)
2	326 (22)
3	165 (11)
≥4	106 (7.3)
Number of European languages, mean (SD)	1.29 (0.60)
Number of South Asian languages, mean (SD)	0.26 (0.73)
Number of East or Southeast Asian languages, mean (SD)	0.07 (0.33)
Number of Middle Eastern languages, mean (SD)	0.06 (0.26)
Number of African languages, mean (SD)	0.0158 (0.1543)
Number of Creole languages, mean (SD)	0.0034 (0.0585)

[a]Bedside rating and wait time rating were missing from 35 profiles; age, race, attractiveness, and medical school were missing from 1-2 profiles.
[b]DO: doctor of osteopathic medicine.

XSL•FO
RenderX

[c]MD: doctor of medicine.

Univariate Analysis

Average overall ratings on Zocdoc are typically highly positive with a median of 4.82 (IQR 4.67-4.93). We find significant but small differences in mean overall rating by physician characteristics including age quartile, facial attractiveness quartile, degree country, degree type, and number of languages spoken (Figure 1, Table S5 in Multimedia Appendix 5). For example, the median overall rating for primary care doctors in the first quartile of age is 4.83 (IQR 4.69-4.92) but 4.78 (IQR 4.61-4.93) for doctors in the fourth quartile (Wilcoxon rank sum test, *P*=.02). We find similar results when using Face++'s age estimates (Table S2 in Multimedia Appendix 2). The median overall rating for primary care doctors in the first quartile of facial attractiveness is 4.81 (IQR 4.66-4.93) but 4.84 (IQR 4.72-4.93) for doctors in the fourth quartile (*P*=.03). Osteopathic physicians have a median overall rating of 4.86 (IQR 4.74-4.96) compared to 4.81 (IQR 4.66-4.92) for allopathic physicians (*P*<.001). The median overall rating for primary care doctors with a US or Canadian degree is 4.85 (IQR 4.70-4.94) compared to 4.79 (IQR 4.61-4.90) for those with a foreign degree (*P*<.001). However, not all foreign-educated physicians have significantly lower scores. For example, Caribbean-educated physicians have a median rating of 4.84 (IQR 4.73-4.93). Physicians educated in Latin America, however, have a median overall rating of 4.71 (IQR 4.54-4.88), a gap of 0.14 in median overall ratings compared with the US- or Canadian-educated physicians. Doctors with 1 language spoken in their office have a median overall rating of 4.84 (IQR 4.69-4.95) while doctors with ≥4 languages have a median overall rating of 4.76 (IQR 4.52-4.85) (*P*<.001). We observe no significant differential effects in overall ratings by gender across the other variables.

Average bedside manner rating is similarly highly positive with a median of 4.88 (IQR 4.75-4.98). We find significant but small differences by physician characteristics including age quartile, degree type, medical school location, and number of languages spoken (Table S5 in Multimedia Appendix 5). For example, the median bedside manner rating for primary care doctors in the first quartile of age is 4.89 (IQR 4.77-4.97) but 4.84 (IQR 4.70-4.96) for doctors in the fourth quartile (*P*<.001). Osteopathic physicians have a median bedside manner rating of 4.91 (IQR 4.81-5.00) compared to 4.87 (IQR 4.73-4.97) for allopathic physicians (*P*<.001). The median bedside manner rating for primary care doctors with a US or Canadian degree is 4.90 (IQR 4.77-5.00) compared to 4.85 (IQR 4.70-4.95) for those with a foreign degree (*P*<.001). Doctors with 1 language spoken in their office have a median bedside manner rating of 4.90 (IQR 4.77-5.00) while doctors with ≥4 languages have a median bedside manner rating of 4.81 (IQR 4.58-4.91; *P*<.001).

The average wait time rating is similarly highly positive with a median of 4.67 (IQR 4.45-4.83). We find significant but small differences by physician characteristics including age quartile, facial attractiveness quartile, degree type, medical school location, and number of languages spoken (Table S5 in Multimedia Appendix 5). For example, the median wait time rating for primary care doctors in the first quartile of age is 4.72 (IQR 4.50-4.86) but 4.62 (IQR 4.38-4.79) for doctors in the fourth quartile (*P*<.001). The median wait time rating for primary care doctors in the first quartile of facial attractiveness is 4.64 (IQR 4.40-4.81) but 4.71 (IQR 4.53-4.85) for doctors in the fourth quartile (*P*=.003). Osteopathic physicians have a median wait time rating of 4.74 (IQR 4.54-4.88) compared to 4.65 (IQR 4.42-4.82) for allopathic physicians (*P*<.001). The median wait time rating for primary care doctors with a US degree is 4.69 (IQR 4.50-4.84) compared to 4.62 (IQR 4.40-4.81) for those with a foreign degree (*P*=002). Doctors with 1 language spoken in their office have a mean wait time rating of 4.70 (IQR 4.50-4.85) while doctors who speak ≥4 languages have a median wait time rating of 4.46 (IQR 4.24-4.67; *P*<.001).

Having excluded 66 profiles with zero reviews, the vast majority of profiles have multiple reviews with a median of 57 (IQR 19-230). We find significant differences in the number of reviews by facial attractiveness quartile, number of languages spoken, and degree type. For example, the median number of reviews for doctors below the 25th percentile of facial attractiveness is 50, but 73 for doctors above the 75th percentile (*P*=.009). The median number of reviews for doctors with 1 language is 42, while doctors with ≥4 languages have a median number of reviews of 244 (*P*<.001). The median number of reviews for osteopathic physicians is 44 compared to 61 for allopathic physicians (*P*=.01). Linear regression analysis finds no association between average overall rating and number of reviews.

XSL•FO
RenderX

Figure 1. Overall primary care physician Zocdoc ratings by physician demographic factor. A higher resolution version of this figure is available in Multimedia Appendix 6.

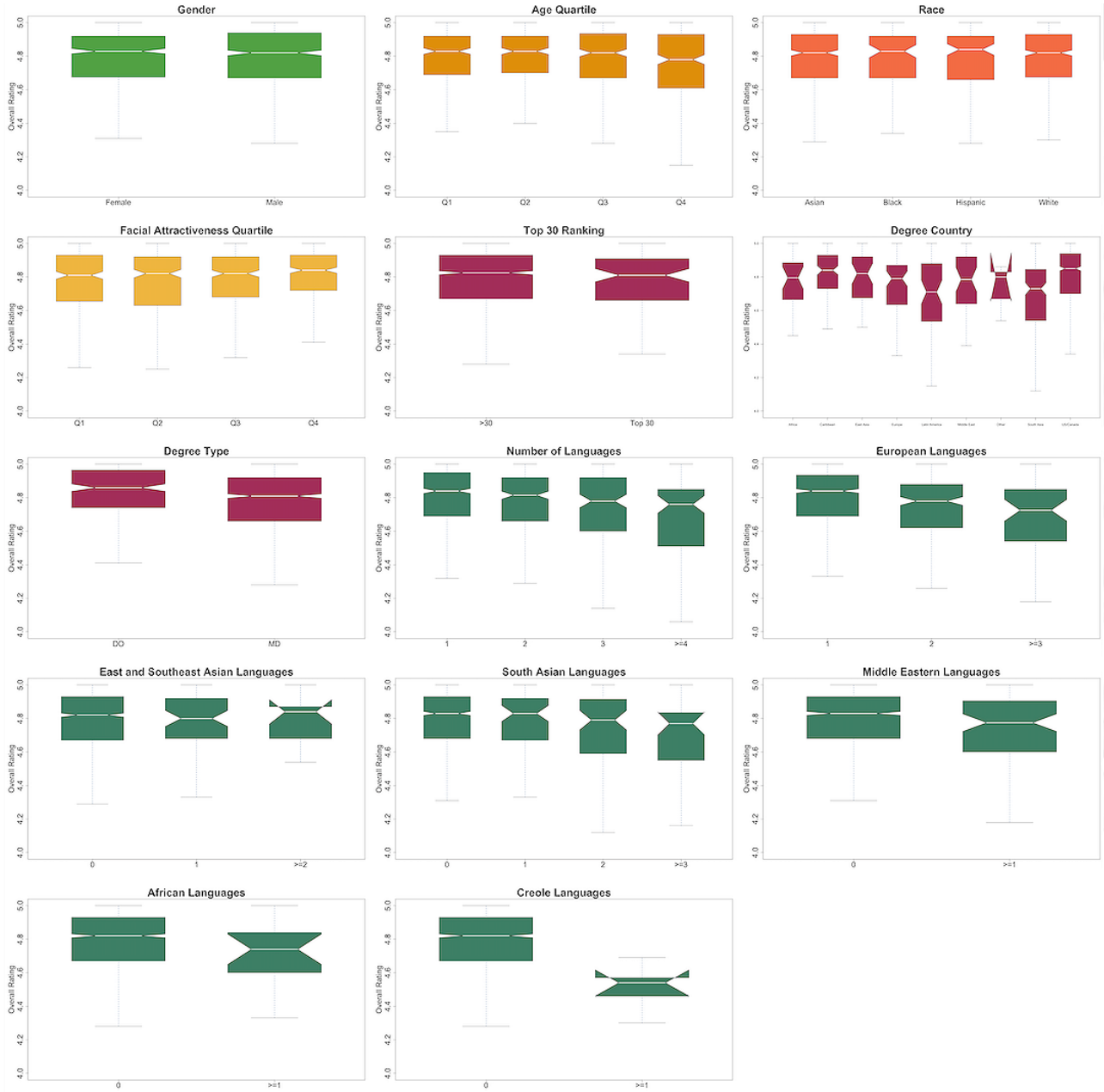

Multivariate Analysis

On multivariate analysis (Table 2), being older, attending a South Asian medical school, and speaking more European and Middle Eastern languages are associated with lower odds of higher overall ratings. For example, attending a South Asian medical school is associated with 0.32 (95% CI 0.20-0.51) greater odds of being in the bottom quartile of overall ratings relative to US or Canadian graduates. Variance inflation factor analysis suggests that there does not exist a problematic amount of collinearity, and all variables have a variance inflation factor below 5. Additionally, we do not find any strong influential outliers on binned residual analysis. We find similar results when using Face++'s Age predictions (Table S3 in Multimedia Appendix 3) and when treating age and facial attractiveness as discrete variables (Table S6 in Multimedia Appendix 7). We present results analyzing overall ratings using cut points at the 75th percentile and the median in Table S7 in Multimedia Appendix 8.

XSL•FO
RenderX

Table 2. Multivariate logistic regression analysis of factors associated with primary care physicians' overall, bedside manner, and wait time ratings on Zocdoc.

Characteristics	Overall[a], OR[b] (95% CI)	Bedside manner[c], OR (95% CI)	Wait time[d], OR (95% CI)
Gender			
Female	Reference level	Reference level	Reference level
Male	1.38 (0.97-1.97)	1.25 (0.88-1.79)	1.13 (0.79-1.61)
Age quartile			
Q1 (0-25)	Reference level	Reference level	Reference level
Q2 (25-30)	1.12 (0.76-1.64)	0.99 (0.67-1.45)	0.99 (0.67-1.47)
Q3 (30-37)	0.69 (0.46-1.04)	0.66 (0.44-0.99)[e]	0.71 (0.46-1.07)
Q4 (37-65)	0.49 (0.30-0.81)[f]	0.51 (0.31-0.84)[f]	0.68 (0.41-1.12)
Race			
White	Reference level	Reference level	Reference level
Asian	1.14 (0.82-1.59)	1.22 (0.88-1.70)	1.12 (0.80-1.57)
Black	0.98 (0.58-1.69)	1.08 (0.64-1.88)	0.67 (0.41-1.12)
Hispanic	0.88 (0.57-1.38)	0.94 (0.61-1.47)	1.22 (0.77-1.99)
Facial attractiveness quartile			
Q1 (0-47)	Reference level	Reference level	Reference level
Q2 (47-55)	0.81 (0.57-1.15)	0.87 (0.61-1.23)	0.87 (0.61-1.23)
Q3 (55-64)	0.98 (0.68-1.42)	0.98 (0.67-1.42)	1.07 (0.73-1.55)
Q4 (64-90)	1.09 (0.73-1.62)	0.87 (0.59-1.29)	1.18 (0.79-1.77)
Top 30 Ranking	0.68 (0.42-1.11)	0.56 (0.35-0.91)[e]	0.87 (0.53-1.43)
Region			
United States or Canada	Reference level	Reference level	Reference level
Africa	0.96 (0.39-2.55)	0.76 (0.31-1.94)	1.48 (0.60-4.00)
Caribbean	0.93 (0.59-1.51)	0.93 (0.59-1.51)	1.78 (1.06, 3.10)[e]
East or Southeast Asia	0.83 (0.37-1.95)	1.02 (0.45-2.50)	0.78 (0.36-1.76)
Europe	1.01 (0.54-1.97)	0.83 (0.45-1.58)	0.73 (0.40-1.36)
Latin America	0.56 (0.31-1.02)	0.63 (0.34-1.19)	0.73 (0.40-1.37)
Middle East	0.69 (0.32-1.58)	0.74 (0.34-1.68)	Reference level
Other	0.80 (0.07-17.7)	0.19 (0.01-2.03)	0.87 (0.08-19.3)
South Asia	0.32 (0.20-0.51)[g]	0.30 (0.19-0.48)[g]	Reference level
Degree			
DO[h]	Reference level	Reference level	Reference level
MD[i]	0.88 (0.60-1.29)	0.79 (0.53-1.16)	0.77 (0.52-1.13)
Number of European languages	0.78 (0.63-0.96)[e]	0.77 (0.62-0.95)[e]	0.62 (0.50, 0.77)[g]
Number of East or Southeast Asian languages	0.88 (0.59-1.36)	0.77 (0.52-1.17)	0.78 (0.53-1.18)
Number of South Asian languages	1.05 (0.87-1.29)	1.02 (0.84-1.24)	1.00 (0.83-1.22)
Number of Middle Eastern languages	0.58 (0.36-0.94)[e]	0.58 (0.36-0.94)[e]	0.50 (0.30, 0.84)[f]
Number of African languages	0.73 (0.34-1.64)	0.84 (0.40-2.02)	0.71 (0.33-1.53)
Number of Creole languages	0.14 (0.01-1.02)	Reference level	0.21 (0.01-1.57)

[a]Akaike information criterion (AIC)=1606; Bayesian information criterion (BIC)=1749; deviance=1552; and area under the receiver operating characteristic

XSL•FO
RenderX

curve (AUROC)=0.654.

[b]OR: odds ratio.

[c]AIC=1597; BIC=1739; deviance=1543; and AUROC=0.655.

[d]AIC=1542; BIC=1684; deviance=1488; and AUROC=0.666.

[e]$P<.05$.

[f]$P<.01$.

[g]$P<.001$.

[h]DO: doctor of osteopathic medicine.

[i]MD: doctor of medicine.

Discussion

This study aimed to examine the impact of physician characteristics, specifically gender, age, facial attractiveness, medical school ranking, foreign degree status, and number of languages spoken, on patients' ratings of primary care physicians on Zocdoc, a platform that only publishes reviews from verified patients. Our univariate findings show that older age, lower facial attractiveness, foreign degrees, allopathic degrees, and more languages spoken in office are negatively associated with the overall rating of primary care doctors. However, the effect sizes of these factors are small and may not be clinically significant. For example, we find a 0.14 gap in median overall rating between doctors with a US degree and doctors with a Latin American degree. Prior research identifies that a half-star improvement in ratings leads to a 10% increase in likelihood that a physician will fill an appointment [11]. A back-of-the-envelope calculation suggests that this 0.14 gap is equivalent to a 2.8% increase in appointments.

To our knowledge, this paper is the first to examine the association between the number of languages, age, facial attractiveness, degree type, foreign graduate status, school ranking, and race with web-based physician reviews, and the first to examine the association between the number of languages and school ranking on patient satisfaction in general.

Our findings on the role of degree type and school ranking are generally consistent with past research. For example, a national telephone survey found that patients of osteopathic physicians generally reported higher rates of satisfaction than patients of allopathic physicians [49]. Past research has found that osteopathic physicians are more likely than allopathic physicians to call patients by their first name, provide information on the underlying causes of their illnesses, and have conversations with them about the social, family, and emotional implications of their medical conditions, all of which may contribute to higher satisfaction rates [50]. In addition, past research finds little correlation between medical school ranking and patient mortality and readmission rates, suggesting an elite ranking has negligible impact on patient care [51].

While it may be easy to hypothesize reasons for differences based on age, facial attractiveness, or foreign degree status, it is not entirely clear what could drive the differences we observe with the number of languages. We find that more languages spoken is associated with a decrease in rating. The number of languages may serve as a proxy for foreign-born status, but we cannot be certain. It is possible that physicians who speak 2 languages are more likely to be bilingual US natives whereas those who speak more than 2 languages are more likely to be immigrants. Survey data have found that patients report lower satisfaction with international medical graduates [25,52]. Older Medicare patients notably have lower mortality rates when treated by international graduates compared to US graduates [53].

In addition, we did not find any numerical rating differences by race. This is inconsistent with past work on findings on a direct-to-consumer telemedicine platform in which patients report lower satisfaction with Asian and Black physicians [25]. Moreover, we did not find any numerical rating differences by gender, which is inconsistent with some previous papers on the topic [18,19,21,22,54]. Zocdoc, however, differs from many other physician review platforms in that reviews can only come from patients after they have received care from their respective physician [33]. On third-party independent platforms like RateMDs, Google Reviews, or Healthgrades, anyone may post a review regardless of whether they have actually seen the provider. Additionally, Zocdoc solicits reviews after each appointment which may counter the typical biases of only extremely satisfied or unsatisfied consumers leaving reviews, leading to larger, more representative samples. Lastly, Zocdoc only publishes patient reviews that do not violate their community standards barring those that include personal information, pricing specifics, profanity, claims about the accuracy of a provider's treatment or diagnosis, or promotional content.

Several limitations must be considered in interpreting the study results. First, our study only examined patient reviews of physicians on one web-based review platform and results may not generalize to other platforms. For instance, millennial women, New Yorkers, and residents of urban areas are disproportionately represented as patients on Zocdoc [55]. Moreover, while we find no differences by gender, female physicians on review platforms such as RateMDs and Google Reviews have been found to have lower numerical ratings than men [20-22]. This may be driven by a lack of verification in review postings or a lack of review moderation. Because Zocdoc moderates reviews, we cannot determine how results might change if reviews that violated community standards were included in the analysis. For example, it is possible that immigrants or women are more likely to receive lower but also more profane reviews. If Zocdoc removes these profane reviews, any disparities we are able to observe may be attenuated. In addition, because most physicians have a very high mean rating, there may be a ceiling effect that makes it difficult to discern relationships between ratings and physician factors. Second, we restrict the study to primary care physicians and the results

XSL•FO
RenderX

may differ for other specialties. For example, one small study of 271 sports medicine surgeons found gender differences in ratings on 1 of 3 platform studies [23]. Third, we limit our analysis to differences in numerical ratings. However, linguistic analyses may yield different types of bias; research on Zocdoc does find that text reviews of women physicians are more informal and emotional than reviews of men [19]. Fourth, we rely on automated face classification software to classify facial attractiveness, race, and age. However, while past social science work may have relied on human raters to code unstructured information, many social scientists have fully moved toward using automated algorithmic procedures [38-43,47,48,56-59]. Moreover, while we do not have access to the ground truth data on race and age, patients most likely do not have access to this information and instead are influenced by perceived race and age. Fifth, the number of languages spoken in a provider's office may correspond to the number of languages spoken by the physicians themselves or by their staff, and what this variable proxies is unclear. Sixth, the overall explanatory power of our multivariable models is fair and may be limited due to the many factors that play key roles in patient satisfaction that we are unable to observe. Seventh, our data do not permit any mechanistic or causal interpretations. Despite these limitations, our study represents an important step in understanding the potential biases in web-based doctor reviews and highlights the need for further research in this area.

Although our study suggests that physician factors have a real but limited impact on numerical ratings, it is important to note that bias may play out in many forms. For example, a physician's appearance or accent may impact a patient's trust, confidence, or satisfaction with their physician, which could in turn influence their take-up of preventative services and lead to either better or worse health outcomes [60]. Such a phenomenon would not be captured in our analysis of numerical ratings, but our results open the door to investigating such phenomenon across facial attractiveness, multilingualism, education, and age more deeply. In conclusion, this study provides insights into the association between physician characteristics and patients' web-based ratings of primary care physicians. Future research should consider textual analyses of reviews, investigate how factors like facial attractiveness interact with patient outcomes, and explore whether the findings of this study generalize to other medical specialties, review platforms, or patient populations. Ultimately, our findings underscore the need for greater awareness of potential biases in web-based doctor reviews and the importance of considering a range of factors in evaluating health care providers.

Data Availability

The data sets generated during or analyzed during this study are available in the Open Science Framework repository.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Languages.
[DOCX File , 17 KB - jmir_v26i1e51672_app1.docx]

Multimedia Appendix 2
Comparison of Kairos and Face++ on Age.
[DOCX File , 14 KB - jmir_v26i1e51672_app2.docx]

Multimedia Appendix 3
Multivariate Results using Face++ Age Predictions.
[DOCX File , 17 KB - jmir_v26i1e51672_app3.docx]

Multimedia Appendix 4
Locations of Primary Care Doctors.
[DOCX File , 15 KB - jmir_v26i1e51672_app4.docx]

Multimedia Appendix 5
Median Outcome by Physician Demographic Factor.
[DOCX File , 20 KB - jmir_v26i1e51672_app5.docx]

Multimedia Appendix 6
Higher resolution version of Figure 1. Overall primary care physician Zocdoc ratings by physician demographic factor.
[PDF File (Adobe PDF File), 1367 KB - jmir_v26i1e51672_app6.pdf]

XSL•FO
RenderX

Multimedia Appendix 7
Multivariate Logistic Regression Results with Discrete Age and Facial Attractiveness.
[DOCX File , 16 KB - jmir_v26i1e51672_app7.docx]

Multimedia Appendix 8
Multivariate Logistic Regression Results with Alternate Overall Rating Cut Point.
[DOCX File , 15 KB - jmir_v26i1e51672_app8.docx]

References

1. Hribar CA, Chandran A, Piazza M, Quinsey CS. Association between patient perception of surgeons and color of scrub attire. JAMA Surg 2023;158(4):421-423 [FREE Full text] [doi: 10.1001/jamasurg.2022.5837] [Medline: 36630142]
2. Rehman SU, Nietert PJ, Cope DW, Kilpatrick AO. What to wear today? Effect of doctor's attire on the trust and confidence of patients. Am J Med 2005;118(11):1279-1286. [doi: 10.1016/j.amjmed.2005.04.026] [Medline: 16271913]
3. Petrilli CM, Saint S, Jennings JJ, Caruso A, Kuhn L, Snyder A, et al. Understanding patient preference for physician attire: a cross-sectional observational study of 10 academic medical centres in the USA. BMJ Open 2018;8(5):e021239 [FREE Full text] [doi: 10.1136/bmjopen-2017-021239] [Medline: 29844101]
4. Jennings JD, Ciaravino SG, Ramsey FV, Haydel CH. Physicians' attire influences patients' perceptions in the urban outpatient orthopaedic surgery setting. Clin Orthop Relat Res 2016;474(9):1908-1918 [FREE Full text] [doi: 10.1007/s11999-016-4855-7] [Medline: 27116208]
5. Clark M, Shuja A, Thomas A, Steinberg S, Geffen J, Malespin M, et al. Patients' perceptions of gastroenterologists' attire in the clinic and endoscopy suite. Ann Gastroenterol 2018;31(2):237-240 [FREE Full text] [doi: 10.20524/aog.2017.0223] [Medline: 29507472]
6. Au S, Khandwala F, Stelfox HT. Physician attire in the intensive care unit and patient family perceptions of physician professional characteristics. JAMA Intern Med 2013;173(6):465-467. [doi: 10.1001/jamainternmed.2013.2732] [Medline: 23420343]
7. Burkle CM, Keegan MT. Popularity of internet physician rating sites and their apparent influence on patients' choices of physicians. BMC Health Serv Res 2015;15:416 [FREE Full text] [doi: 10.1186/s12913-015-1099-2] [Medline: 26410383]
8. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. Public awareness, perception, and use of online physician rating sites. JAMA 2014;311(7):734-735. [doi: 10.1001/jama.2013.283194] [Medline: 24549555]
9. Yaraghi N, Wang W, Gao GG, Agarwal R. How online quality ratings influence patients' choice of medical providers: controlled experimental survey study. J Med Internet Res 2018;20(3):e99 [FREE Full text] [doi: 10.2196/jmir.8986] [Medline: 29581091]
10. Xu Y, Armony M, Ghose A. The Interplay between Online Reviews and Physician Demand: An Empirical Investigation. Rochester, NY: SSRN Scholarly Paper; 2016.
11. Luca M, Vats S. Digitizing Doctor Demand: The Impact of Online Reviews on Doctor Choice. 2013. URL: https://www.aeaweb.org/conference/2014/retrieve.php?pdfid=55 [accessed 2024-06-25]
12. McGrath RJ, Priestley JL, Zhou Y, Culligan PJ. The validity of online patient ratings of physicians: analysis of physician peer reviews and patient ratings. Interact J Med Res 2018;7(1):e8 [FREE Full text] [doi: 10.2196/ijmr.9350] [Medline: 29631992]
13. Lu SF, Rui H. Can We Trust Online Physician Ratings? Evidence from Cardiac Surgeons in Florida. Rochester, NY: SSRN Scholarly Paper; 2014.
14. Daskivich TJ, Houman J, Fuller G, Black JT, Kim HL, Spiegel B. Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. J Am Med Inform Assoc 2018;25(4):401-407 [FREE Full text] [doi: 10.1093/jamia/ocx083] [Medline: 29025145]
15. Chen J, Presson A, Zhang C, Ray D, Finlayson S, Glasgow R. Online physician review websites poorly correlate to a validated metric of patient satisfaction. J Surg Res 2018;227:1-6. [doi: 10.1016/j.jss.2018.01.037] [Medline: 29804840]
16. Widmer RJ, Maurer MJ, Nayar VR, Aase LA, Wald JT, Kotsenas AL, et al. Online physician reviews do not reflect patient satisfaction survey responses. Mayo Clin Proc 2018;93(4):453-457 [FREE Full text] [doi: 10.1016/j.mayocp.2018.01.021] [Medline: 29622095]
17. Greaves F, Pape UJ, Lee H, Smith DM, Darzi A, Majeed A, et al. Patients' ratings of family physician practices on the internet: usage and associations with conventional measures of quality in the English National Health Service. J Med Internet Res 2012;14(5):e146 [FREE Full text] [doi: 10.2196/jmir.2280] [Medline: 23076301]
18. Gao GG, McCullough JS, Agarwal R, Jha AK. A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period. J Med Internet Res 2012;14(1):e38 [FREE Full text] [doi: 10.2196/jmir.2003] [Medline: 22366336]
19. Gupta S, Kayla J. Understanding gender bias toward physicians using online doctor reviews. Psychol Lang Commun 2022;26(1):18-41 [FREE Full text] [doi: 10.2478/plc-2022-0002]

XSL•FO
RenderX

20. Thawani A, Paul MJ, Sarkar U, Wallace BC. Are online reviews of physicians biased against female providers? 2019 Presented at: Proceedings of the 4th Machine Learning for Healthcare Conference; 2019; Durham, NC, USA p. 406-423 URL: https://proceedings.mlr.press/v106/thawani19a.html
21. Barnett J, Bjarnadóttir MV, Anderson D, Chen C. Understanding gender biases and differences in web-based reviews of sanctioned physicians through a machine learning approach: mixed methods study. JMIR Form Res 2022;6(9):e34902 [FREE Full text] [doi: 10.2196/34902] [Medline: 36074543]
22. Dunivin Z, Zadunayski L, Baskota U, Siek K, Mankoff J. Gender, soft skills, and patient experience in online physician reviews: a large-scale text analysis. J Med Internet Res 2020;22(7):e14455 [FREE Full text] [doi: 10.2196/14455] [Medline: 32729844]
23. Nwachukwu BU, Adjei J, Trehan SK, Chang B, Amoo-Achampong K, Nguyen JT, et al. Rating a sports medicine surgeon's "quality" in the modern era: an analysis of popular physician online rating websites. HSS J 2016;12(3):272-277 [FREE Full text] [doi: 10.1007/s11420-016-9520-x] [Medline: 27703422]
24. Takeshita J, Wang S, Loren AW, Mitra N, Shults J, Shin DB, et al. Association of racial/ethnic and gender concordance between patients and physicians with patient experience ratings. JAMA Netw Open 2020;3(11):e2024583 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.24583] [Medline: 33165609]
25. Martinez KA, Keenan K, Rastogi R, Roufael J, Fletcher A, Rood MN, et al. Association of racial/ethnic and gender concordance between patients and physicians with patient experience ratings. J Gen Intern Med 2020;35(9):2600-2606 [FREE Full text] [doi: 10.1007/s11606-020-06005-8] [Medline: 32632788]
26. Young JW. Symptom disclosure to male and female physicians: effects of sex, physical attractiveness, and symptom type. J Behav Med 1979;2(2):159-169. [doi: 10.1007/BF00846664] [Medline: 555486]
27. Hall JA, Ruben M, Swatantra. First impressions of physicians according to their physical and social group characteristics. J Nonverbal Behav 2020;44(2):279-299 [FREE Full text] [doi: 10.1007/s10919-019-00329-8]
28. Hadjistavropoulos HD, Ross MA, von Baeyer CL. Are physicians' ratings of pain affected by patients' physical attractiveness? Soc Sci Med 1990;31(1):69-72. [doi: 10.1016/0277-9536(90)90011-g] [Medline: 2367884]
29. Bordieri JE, Solodky ML, Mikos KA. Physical attractiveness and nurses' perceptions of pediatric patients. Nurs Res 1985;34(1):24-26. [Medline: 3844157]
30. Rao AR, Clarke D. Exploring relationships between medical college rankings and performance with big data. Big Data Anal 2019;4(1) [FREE Full text] [doi: 10.1186/s41044-019-0040-9]
31. Eskes C, Salisbury H, Johannsson M, Chene Y. Patient satisfaction with language--concordant care. J Physician Assist Educ 2013;24(3):14-22. [doi: 10.1097/01367895-201324030-00003] [Medline: 24261166]
32. Lopez Vera A, Thomas K, Trinh C, Nausheen F. A case study of the impact of language concordance on patient care, satisfaction, and comfort with sharing sensitive information during medical care. J Immigr Minor Health 2023;25(6):1261-1269 [FREE Full text] [doi: 10.1007/s10903-023-01463-8] [Medline: 36840903]
33. Luca. Verified Reviews. Zocdoc. 2013. URL: https://www.zocdoc.com/about/verifiedreviews/ [accessed 2023-04-25]
34. How Does Zocdoc Make Money. Zocdoc. URL: https://www.zocdoc.com/about/question/how-does-zocdoc-make-money/ [accessed 2023-04-25]
35. City and Town Population Totals: 2020-2021. United States Census Bureau. Washington, DC URL: https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html [accessed 2024-06-25]
36. Powerful Web Scraper for Regular and Professional use. 2013. URL: https://webscraper.io/ [accessed 2024-06-25]
37. The Best Medical Schools for Research, Ranked. U.S. News & World Report. URL: https://www.usnews.com/best-graduate-schools/top-medical-schools/research-rankings [accessed 2023-04-25]
38. Benjamin E, Luca M, Svirsky D. Racial discrimination in the sharing economy: evidence from a field experiment. Am Econ J: Appl Econ 2017:1-22 [FREE Full text] [doi: 10.1257/app.20160213]
39. Ouyang P, Wang J. Physician's online image and patient's choice in the online health community. Internet Res 2022;32(6):1952-1977 [FREE Full text] [doi: 10.1108/intr-04-2021-0251]
40. Kosinski M. Facial width-to-height ratio does not predict self-reported behavioral tendencies. Psychol Sci 2017 Nov;28(11):1675-1682. [doi: 10.1177/0956797617716929] [Medline: 28976810]
41. Jaeger B, Sleegers WW, Evans AM, Stel M, van Beest I. The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. J Econ Psychol 2019;75:102125 [FREE Full text] [doi: 10.1016/j.joep.2018.11.004]
42. Troncoso I, Luo L. Look the part? the role of profile pictures in online labor markets. Mark Sci 2023;42(6):1080-1100 [FREE Full text] [doi: 10.1287/mksc.2022.1425]
43. Jaeger B, Sleegers WWA, Evans AM. Automated classification of demographics from face images: a tutorial and validation. Social & Personality Psych 2020;14(3):e12520 [FREE Full text] [doi: 10.1111/spc3.12520]
44. Morgan A, Shah K, Tran K, Chino F. Racial, ethnic, and gender representation in leadership positions at national cancer institute-designated cancer centers. JAMA Netw Open 2021;4(6):e2112807 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.12807] [Medline: 34097046]
45. Alharbi M, Shihong H. A survey of incorporating affective computing for human-system co-adaptation. 2020 Presented at: Proceedings of the 2nd World Symposium on Software Engineering; 2020 September; New York, NY, USA p. 72-79 URL: https://doi.org/10.1145/3425329.3425343

XSL•FO
RenderX

46. Goel N, Rutagarama M, Faltings B. Tackling peer-to-peer discrimination in the sharing economy. 2020 Presented at: Proceedings of the 12th ACM Conference on Web Science; 2020 July; New York, NY, USA p. 365-361 URL: https://doi.org/10.1145/3394231.3397926
47. Mathis MS, Badewa TE, Obiarinze RN, Wilkinson LT, Martin CA. A novel use of artificial intelligence to examine diversity and hospital performance. J Surg Res 2021;260:377-382. [doi: 10.1016/j.jss.2020.07.081] [Medline: 33750544]
48. Toney C, Shroyer Mathis M, Martin C. The use of facial recognition software and published manuscripts to examine trends in surgical editorial board diversity. J Surg Res 2023;286:104-109. [doi: 10.1016/j.jss.2022.11.057] [Medline: 36803877]
49. Licciardone JC, Herron KM. Characteristics, satisfaction, and perceptions of patients receiving ambulatory healthcare from osteopathic physicians: a comparative national survey. J Am Osteopath Assoc 2001;101(7):374-385. [Medline: 11476027]
50. Carey TS, Motyka TM, Garrett JM, Keller RB. Do osteopathic physicians differ in patient interaction from allopathic physicians? An empirically derived approach. J Am Osteopath Assoc 2003;103(7):313-318. [Medline: 12884943]
51. Tsugawa Y, Blumenthal DM, Jha AK, Orav EJ, Jena AB. Association between physician medical school ranking and patient outcomes and costs of care: observational study. BMJ 2018;362:k3640 [FREE Full text] [doi: 10.1136/bmj.k3640] [Medline: 30257919]
52. Engelhardt KE, Matulewicz RS, DeLancey JO, Merkow RP, Quinn CM, Kreutzer L, et al. Physician characteristics associated with patient experience scores: implications for adjusting public reporting of individual physician scores. BMJ Qual Saf 2019;28(5):412-415. [doi: 10.1136/bmjqs-2018-008346] [Medline: 30487182]
53. Tsugawa Y, Jena AB, Orav EJ, Jha AK. Quality of care delivered by general internists in US hospitals who graduated from foreign versus US medical schools: observational study. BMJ 2017;356:j273 [FREE Full text] [doi: 10.1136/bmj.j273] [Medline: 28153977]
54. Kauff M, Anslinger J, Christ O, Niemann M, Geierhos M, Huster L. Ethnic and gender-based prejudice towards medical doctors? The relationship between physicians' ethnicity, gender, and ratings on a physician rating website. J Soc Psychol 2022;162(5):540-548. [doi: 10.1080/00224545.2021.1927944] [Medline: 34323667]
55. Fisher T. Do women prefer female doctors? it depends. In: The Paper Gown. New York: Zocdoc; 2018.
56. Sehgal NKR, Brownstein JS, Majumder MS, Tuli G. US COVID-19 clinical trial leadership gender disparities. Lancet Digit Health 2023;5(3):e109-e111 [FREE Full text] [doi: 10.1016/S2589-7500(23)00017-1] [Medline: 36828602]
57. Peng L, Cui G, Chung Y, Zheng W. The faces of success: beauty and ugliness premiums in e-commerce platforms. J. Mark 2020;84(4):67-85 [FREE Full text] [doi: 10.1177/0022242920914861]
58. Dietl H, Özdemir A, Rendall A. The role of facial attractiveness in tennis TV-viewership. Sport Manag Rev 2020;23(3):521-535 [FREE Full text] [doi: 10.1016/j.smr.2019.04.004]
59. Hyunkyu J. Judging an airbnb booking by its cover: how profile photos affect guest ratings. JCM 2022;39(4):371-382 [FREE Full text] [doi: 10.1108/jcm-01-2021-4341]
60. Frakes MD, Gruber J. Racial concordance and the quality of medical care: evidence from the military. In: Nber working paper series. Cambridge, MA: National Bureau of Economic Research; 2022.

Edited by A Mavragani; submitted 07.08.23; peer-reviewed by K Jordan, K Martinez; comments to author 11.11.23; revised version received 17.11.23; accepted 12.06.24; published 29.07.24.

Please cite as:
Sehgal NKR, Rader B, Brownstein JS
Examining the Role of Physician Characteristics in Web-Based Verified Primary Care Physician Reviews: Observational Study
J Med Internet Res 2024;26:e51672
URL: https://www.jmir.org/2024/1/e51672
doi:10.2196/51672
PMID:

©Neil K R Sehgal, Benjamin Rader, John S Brownstein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Classification of Patients' Judgments of Their Physicians in Web-Based Written Reviews Using Natural Language Processing: Algorithm Development and Validation

Farrah Madanay[1,2], PhD; Karissa Tu[3,4], BS; Ada Campagna[5,6], MA; J Kelly Davis[3,7], BA; Steven S Doerstling[8], MD; Felicia Chen[9], BS; Peter A Ubel[1,3], MD

[1]Sanford School of Public Policy, Duke University, Durham, NC, United States
[2]Center for Bioethics and Social Sciences in Medicine, University of Michigan Medical School, Ann Arbor, MI, United States
[3]Fuqua School of Business, Duke University, Durham, NC, United States
[4]School of Medicine, University of Washington, Seattle, WA, United States
[5]Center for Advanced Hindsight, Duke University, Durham, NC, United States
[6]Department of Sociology, University of California, Los Angeles, Los Angeles, CA, United States
[7]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, United States
[8]Department of Medicine, Stanford University, Stanford, CA, United States
[9]GrantScout, San Francisco, CA, United States

Corresponding Author:
Farrah Madanay, PhD
Center for Bioethics and Social Sciences in Medicine
University of Michigan Medical School
2800 Plymouth Rd
Bldg 14, G016
Ann Arbor, MI, 48109
United States
Phone: 1 8083524196
Email: madanafl@med.umich.edu

Abstract

Background: Patients increasingly rely on web-based physician reviews to choose a physician and share their experiences. However, the unstructured text of these written reviews presents a challenge for researchers seeking to make inferences about patients' judgments. Methods previously used to identify patient judgments within reviews, such as hand-coding and dictionary-based approaches, have posed limitations to sample size and classification accuracy. Advanced natural language processing methods can help overcome these limitations and promote further analysis of physician reviews on these popular platforms.

Objective: This study aims to train, test, and validate an advanced natural language processing algorithm for classifying the presence and valence of 2 dimensions of patient judgments in web-based physician reviews: interpersonal manner and technical competence.

Methods: We sampled 345,053 reviews for 167,150 physicians across the United States from Healthgrades.com, a commercial web-based physician rating and review website. We hand-coded 2000 written reviews and used those reviews to train and test a transformer classification algorithm called the Robustly Optimized BERT (Bidirectional Encoder Representations from Transformers) Pretraining Approach (RoBERTa). The 2 fine-tuned models coded the reviews for the presence and positive or negative valence of patients' interpersonal manner or technical competence judgments of their physicians. We evaluated the performance of the 2 models against 200 hand-coded reviews and validated the models using the full sample of 345,053 RoBERTa-coded reviews.

Results: The interpersonal manner model was 90% accurate with precision of 0.89, recall of 0.90, and weighted F_1-score of 0.89. The technical competence model was 90% accurate with precision of 0.91, recall of 0.90, and weighted F_1-score of 0.90. Positive-valence judgments were associated with higher review star ratings whereas negative-valence judgments were associated with lower star ratings. Analysis of the data by review rating and physician gender corresponded with findings in prior literature.

XSL•FO
RenderX

Conclusions: Our 2 classification models coded interpersonal manner and technical competence judgments with high precision, recall, and accuracy. These models were validated using review star ratings and results from previous research. RoBERTa can accurately classify unstructured, web-based review text at scale. Future work could explore the use of this algorithm with other textual data, such as social media posts and electronic health records.

(J Med Internet Res 2024;26:e50236) doi:10.2196/50236

KEYWORDS

web-based physician reviews; patient judgments; RoBERTa; natural language processing; text classification; machine learning; patient experience; patient-authored reviews; healthcare quality; patient care; psychology

Introduction

Patients increasingly turn to commercial physician rating and review websites to discuss their patient experiences and provide feedback to hospitals and providers [1,2]. Patient-authored reviews on these websites may capture factors of the patient experience not otherwise found in traditional patient experience surveys (eg, Press Ganey) or academic research (eg, interviews and questionnaires), such as insurance processing and appointment scheduling [3,4]. These websites have therefore gained increased attention among researchers seeking to better understand what patients care about and how commercial review data compare to other health care quality measures. For example, researchers analyzing commercial hospital reviews identified topics discussed by patients that were not covered in the current Hospital Consumer Assessment of Healthcare Providers and Systems survey, like nurse quality, staff compassion, and the technical aspects of care [3]. Other researchers found negative commercial reviews of surgeons focused on surgeon-independent factors, such as wait times and office staff, suggesting patients may consider factors beyond the patient-physician interaction when assessing quality [5].

Physician review websites potentially impact both patient choice and physician care quality. Some prospective patients rely on web-based physician ratings and reviews to help them choose physicians [4,6,7]. Research shows people prefer words to numbers and can easily comprehend review narratives over quantitative ratings [8]. Additionally, physicians use patient feedback conveyed in web-based written reviews to implement and improve quality measures, particularly related to patient communication [8].

The unstructured narrative text, however, presents a challenge for researchers seeking to make inferences from physician reviews. Methods previously used to identify patient judgments within written reviews include hand-coding [1] and dictionary-based approaches [9,10]. Hand-coding approaches, however, are time- and resource-intensive, which limits sample size [11]. Likewise, dictionary-based methods, such as Linguistic Inquiry and Word Count, use a context-independent bag-of-words approach, which may overlook misspellings, colloquialisms, and keywords and phrases not captured in prebuilt dictionaries [12,13].

In this paper, we present measures of precision, recall, and accuracy for an advanced natural language processing (NLP) algorithm, fine-tuned to identify the presence and valence of 2 dimensions of patient judgments in web-based physician reviews: interpersonal manner and technical competence. We use an algorithm called the Robustly Optimized BERT (Bidirectional Encoder Representations from Transformers) Pretraining Approach (RoBERTa), which we trained to classify our 2 judgment dimensions in written reviews and which has been successfully applied in other classification contexts (eg, Twitter) [13-15]. RoBERTa's novelty is in its transformer-based, bidirectional, context-aware approach, wherein it is pretrained on a large corpus of text but is fine-tunable for many NLP tasks, including text classification [15,16]. We validate this algorithm by correlating results with review star ratings and by comparing results with those found in prior literature.

Methods

Data Collection

We scraped physician profiles, ratings, and review data published on Healthgrades.com in April 2020. We collected primary care physician profiles associated with family medicine, internal medicine, and pediatrics, and surgeon profiles associated with general surgery; orthopedic surgery; and cosmetic, plastic, and reconstructive surgery. Healthgrades.com has a physician profile for every US physician with an active profile listed on the National Provider Identifier Registry [17]. In addition to physician profile characteristics, we scraped rating information and up to 20 of the most recent written reviews per physician. On Healthgrades.com, patients can elect to submit a star rating alone (ie, 1-5 stars, no fractions) or a star rating accompanied by a written review. The study was approved by the Duke University institutional review board and all data collected were publicly available and aggregated for research purposes. Our final sample included 345,053 reviews submitted for 167,150 physicians (primary care physicians and surgeons). Figure 1 shows a flow chart of our sample selection.

XSL•FO
RenderX

Figure 1. Sample selection flow chart.

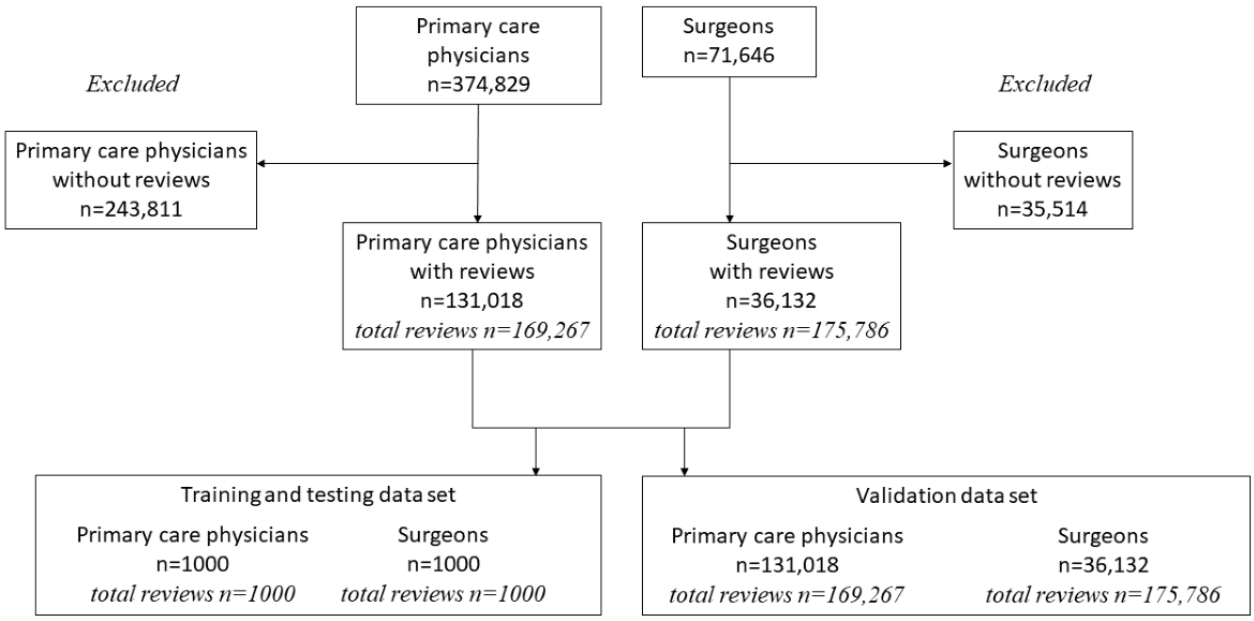

Coding Reviews for Interpersonal Manner and Technical Competence

Hand-Coding the Training Data

We first cleaned the text to convert non–ASCII-encoded characters to ASCII characters; examples included some apostrophes, dashes, and letters with accents (eg, blasé). We then trained our classification algorithm using a gold standard data set of rigorously hand-coded physician reviews [11]. We purposely sampled 2000 random reviews for equal representation of primary care physicians and surgeons, female and male physicians, and low-star (≤3 stars) and high-star (≥4 stars) review ratings. We achieved high interrater reliability with a subset of 300 double-coded reviews (Cohen κ range 0.74-0.85), before proceeding to independently code the remaining reviews.

We coded each review for the presence or absence of interpersonal manner and technical competence. Reviews could be coded for the presence of only 1 dimension, both dimensions, and neither dimension. Once we indicated the presence of a judgment dimension, we coded the valence of the judgment as positive or negative. If we did not code the presence of a judgment dimension, we would not have a valence indicated for that judgment. Figure 2 provides a diagram with illustrative examples showing how we hand-coded the presence and valence of the 2 judgment dimensions.

XSL•FO
RenderX

Figure 2. Diagram showing how real physician reviews were hand-coded for the presence and valence of interpersonal manner and technical competence.

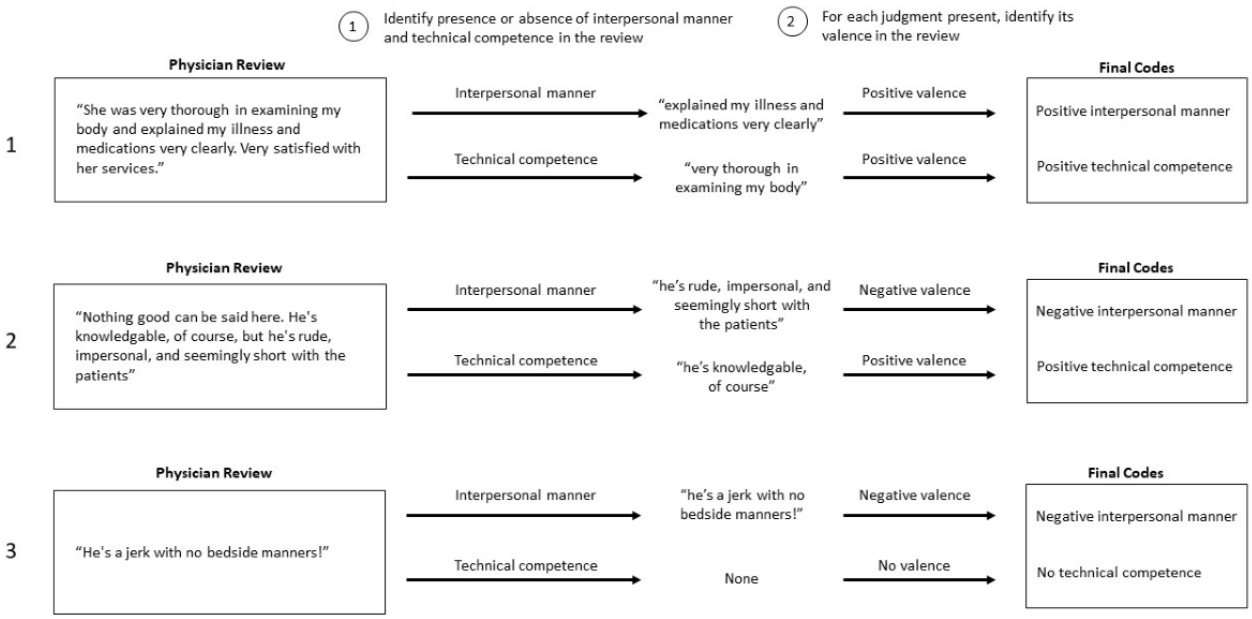

Training and Testing the Algorithm

To code patients' interpersonal manner and technical competence judgments in the sample of 345,053 reviews, we first used our hand-coded data to train RoBERTa, a transformer classification model. Transformers are neural network systems that use vectors to capture the meanings of words in context and are the main architecture underlying advanced NLP models [13,18]. These state-of-the-art NLP models improve upon prior NLP classification approaches, such as those based on dictionaries or fixed embeddings [13]. Specifically, RoBERTa builds bidirectional context-aware embeddings such that the vector representing the word changes depending on its context in the text [13,19]. RoBERTa is pretrained on Book Corpus (800 million words) and English Wikipedia (2500 million words), and can be fine-tuned with one additional level of training data for specific classification tasks [20]. We implemented RoBERTa with assistance from simple transformers [21], a wrapper library for the HuggingFace Transformers library [22]. In our study, each review used to train RoBERTa has its own sequence embedding and helps fine-tune the model to code the reviews for the presence and valence of interpersonal manner and technical competence.

We fine-tuned 2 multiclass classification models, 1 for classifying interpersonal manner and 1 for technical competence. We tuned each model using 1600 (80%) reviews randomly sampled from our hand-coded training data. We completed 6 iterations through our training data for both models. We used a test data set, or a set of 200 (10%) hand-coded reviews held out of the training data, to evaluate each model's fit on the training data set while further fine-tuning the model. Finally, we used a new set of 200 (10%) hand-coded reviews to provide an unbiased evaluation of the classification performance of each fully trained model for patients' interpersonal manner and technical competence judgments. Training and evaluation batch sizes for both models were 1024 sequences; both models used 6 training epochs, with final epoch running losses <0.01.

After training and testing the 2 models using the 2000 hand-coded reviews, we applied the fully trained models to all the reviews in our data set, including the 2000 reviews we hand-coded. This resulted in a data set with 345,053 reviews coded by RoBERTa for the presence and valence of interpersonal manner and technical competence judgments. We used Python and Google Colab to train RoBERTa on our judgment classification tasks and code our full sample of reviews.

Results

Evaluating the Accuracy of the 2 Classification Models

Our 2 classification models were highly accurate. The out-of-sample predictive accuracy for the interpersonal manner model was 90% with a weighted F_1-score of 0.89 (range 0.82-0.95), precision of 0.89 (range 0.85-0.94), and recall of 0.90 (range 0.80-0.96). The out-of-sample predictive accuracy for the technical competence model was also 90%, with a weighted F_1-score of 0.90 (range 0.90-0.92), precision of 0.91 (range 0.88-0.95), and recall of 0.90 (range 0.85-0.95). Table 1 details the classification performance metrics for the interpersonal manner and technical competence models.

XSL•FO
RenderX

Table 1. Fine-tuned transformer classification performance for interpersonal manner and technical competence judgments.

Classification model and valence[a]	Precision[b]	Recall[c]	F_1-score[d]
Interpersonal manner model			
No interpersonal manner	0.85	0.80	0.82
Negative interpersonal manner	0.88	0.89	0.88
Positive interpersonal manner	0.94	0.96	0.95
Accuracy	—[e]	—	0.90
Macro avg	0.89	0.88	0.88
Weighted average	0.89	0.90	0.89
Technical competence model			
No technical competence	0.88	0.91	0.90
Negative technical competence	0.95	0.85	0.90
Positive technical competence	0.89	0.95	0.92
Accuracy	—	—	0.90
Macro average	0.91	0.90	0.90
Weighted average	0.91	0.90	0.90

[a]Classification performance is based on a comparison to an evaluation data set of 200 reviews hand-coded by our team of researchers.

[b]Precision: number of true positives divided by the sum of true positives and false positives.

[c]Recall: number of true positives divided by the sum of true positives and false negatives.

[d]F_1-score: harmonic mean of precision and recall, given by [14].

[e]Not applicable.

Comparing Reviews Coded by RoBERTa and by Hand

As part of our final sample of 345,053 reviews, the 2000 hand-coded reviews from our training data set were recoded by RoBERTa. The interrater reliability between our hand-coding and RoBERTa was Cohen κ =0.96 for both interpersonal manner and technical competence. Comparing the RoBERTa codes with the original hand codes for these reviews, we found only 107 (5.4%) reviews had coding discrepancies. Of those, 49 (2.5%) reviews had the same interpersonal manner code but different technical competence code, 57 (2.9%) reviews had the same technical competence code but different interpersonal manner code, and 1 (0.05%) review had both different interpersonal manner and technical competence codes. Table 2 shows illustrative examples of coding discrepancies in our RoBERTa-coded reviews and our hand-coded reviews from the training data set.

XSL•FO
RenderX

Table 2. Illustrative examples of discrepancies in reviews coded by the RoBERTa[a] and by hand and reasoning underlying the discrepancies.

Review	RoBERTa coding		Hand-coding with reasoning	
	Interpersonal manner	Technical competence	Interpersonal manner	Technical competence
#metoo. He knows what he did.	0[b]	0	–[c] (Hints at sexual assault by the physician)	0
When someone is scared, I Think the Dr. should try to comfort them instead of telling them that when I finish this procedure, you will no longer be my patient . I'll refer you to someone else. So I wish him the best.	+[d]	0	– (Feels physician did not provide comfort)	0
very rude receptionists answering phones, unhelpful, and sarcastic, they should be replaced ! they cant be bothered especially trish! very rude!	–	0	0 (Discusses the interpersonal manner of the staff, not the physician)	0
Walking on the second day. My daughter also! I would send anyone I know his way... We're from out of town and after using a local doctor here and having to do it all over again the experience was amazing and recovery was shorter	0	0	0	+ (Perceives treatment as a success)
Awful experience! Thanks to another doctor that happened to see the urgency of my condition, I got the help that I needed. Had it been left to Dr. C[e], God only knows where I'd be today. Avoid him!!!	0	+	0	– (Perceives poor physician decision-making)
Excellent !!!!!! You will never ever find a better M.D. Caring, so professional an excellent surgeon with compassion??	+	+	+	0 (Does not discuss physician's expertise, treatment, or outcomes)
Very similar to all the one star ratings, if 0 stars were an option I'd choose that. The follow up on patients are non-existent, which makes it very obvious that the surgeon just wants $. The staff is always rude. I wish they would treat their patients and their family members how they would like their own to be treated. My mom has had two infections where they removed her lymph nodes after trying to call them about this several times, we took her to a different doctor to have the site drained	0	0	– (Feels physician prioritized money over care)	– (Critiques physician's lack of follow-up care)

[a]RoBERTa: Robustly Optimized BERT (Bidirectional Encoder Representations from Transformers) Pretraining Approach.

XSL•FO
RenderX

[b]"0" indicates no judgment coded.

[c]"–" indicates negative judgment coded.

[d]"+" indicates positive judgment coded.

[e]Physician surname reproduced here by only its first letter.

Testing the Validity of the 2 Classification Models

Overview

We tested the validity of our classification models using the full sample of RoBERTa-coded reviews. We validated our models in 2 ways. First, we related our valence coding with the review star ratings (ie, 1- to 5-star ratings submitted with each review), with the expectation that positive-valence judgments would be associated with higher star ratings and negative-valence judgments would be associated with lower star ratings. Second, we compared our findings with prior literature on patients' judgments of physicians in web-based physician reviews.

Testing for Associations Between Judgment Valence and Star Ratings

Using multilevel linear regressions, we analyzed associations between patients' interpersonal manner and technical competence judgment valences and review star ratings. We found evidence of construct validity: positive valences for both interpersonal manner and technical competence were significantly positively associated with review star ratings whereas negative valences for both judgment dimensions were significantly negatively associated with review star ratings. Compared with reviews with no or negative judgment, reviews with positive interpersonal manner were associated with 1.82 (95% CI 1.81-1.83; *P*<.001) more stars, and reviews with positive technical competence were associated with 1.50 (95% CI 1.49-1.51; *P*<.001) more stars. In contrast, compared with reviews with no or positive judgment, reviews with negative interpersonal manner were associated with 3.30 (95% CI –3.31 to –3.29; *P*<.001) fewer stars and reviews with negative technical competence were associated with 3.00 (95% CI –3.01 to –2.98; *P*<.001) fewer stars. Figure 3 displays mean review star ratings for each judgment dimension.

Figure 3. Mean review star ratings with SDs, for reviews with negative, no, and positive interpersonal manner or technical competence.

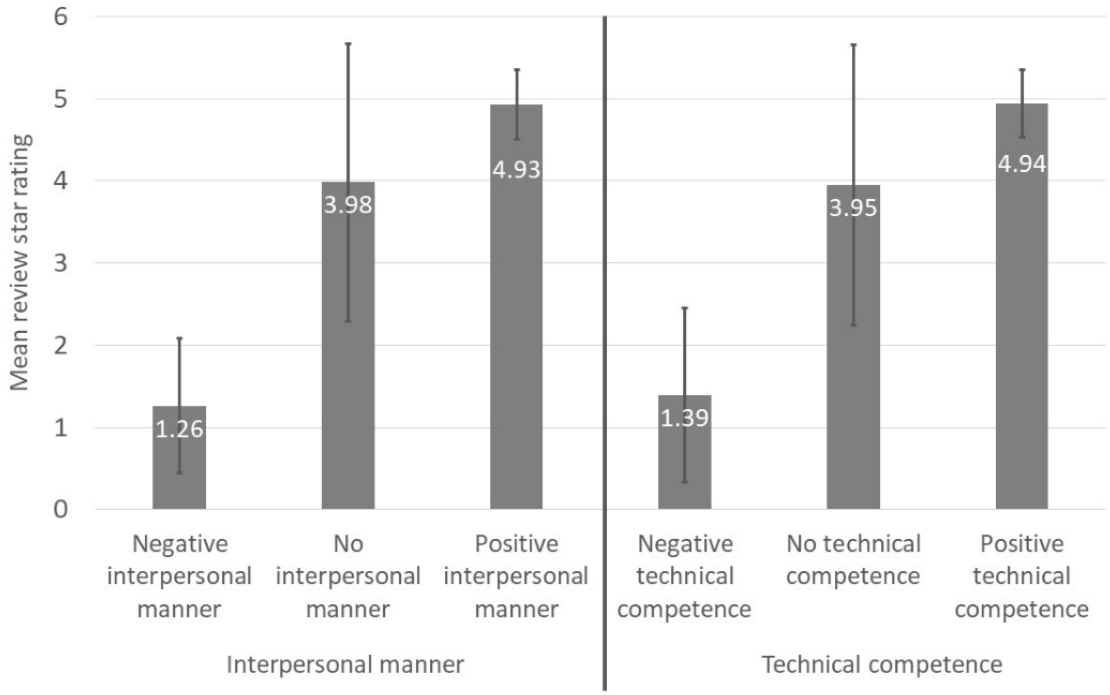

Testing Whether the Models Reproduce Prior Findings

One study analyzing 712 reviews determined that 69% of interpersonal manner reviews and 80% of technical competence reviews were positive [1]. We identified a similar pattern of majority positive reviews; 207,327 (81%) reviews mentioning interpersonal manner and 178,705 (82%) reviews mentioning technical competence were positive. Another study reported physicians who received reviews with interpersonal manner language were at least 2.39 times more likely to receive a 5-star review rating [9]. We similarly found physicians who received interpersonal manner reviews had 1.69 times the odds of receiving a 5-star review rating (95% CI 1.65-1.73; *P*<.001). When controlling for physician gender, specialty, age, and practicing state, as well as review word count, physicians with interpersonal manner reviews continued to have higher odds of receiving a 5-star review rating (odds ratio [OR] 2.22, 95% CI 2.17-2.28; *P*<.001).

Prior research also showed female physicians, compared with male physicians, had higher odds of receiving reviews mentioning interpersonal manner [9,10]. Our findings supported

XSL•FO
RenderX

these results: we determined female physicians had 1.56 times the odds of receiving a review mentioning interpersonal manner than male physicians (95% CI 1.53-1.59; $P<.001$). When controlling for physician specialty, age, practicing state, and review word count, female physicians still had significantly higher odds of receiving an interpersonal manner review (OR 1.19, 95% CI 1.17-1.22; $P<.001$).

One group of investigators demonstrated female physicians were more likely than male physicians to receive both reviews praising and reviews criticizing their interpersonal manner [10]. Consistent with these results, we found female physicians had 1.40 times the odds of receiving a negative review about their interpersonal manner than male physicians (95% CI 1.36-1.44; $P<.001$). This significant difference remained when including controls (OR 1.25, 95% CI 1.21-1.29; $P<.001$). Female physicians also had 1.18 times the odds of receiving a positive interpersonal manner review than male physicians (95% CI 1.15-1.20; $P<.001$); however, this gender difference was not significant when including controls (OR 1.02, 95% CI 1.00-1.04; $P=.05$).

Likewise, another study concluded that highly rated male physicians were 1.48 times more likely to receive reviews describing technical competence whereas highly rated female physicians were 2.11 times more likely to receive reviews describing interpersonal manner [23]. We found similar results: Highly rated female physicians had 1.76 higher odds of receiving an interpersonal manner review (95% CI 1.72-1.81; $P<.001$), which was still significant after including controls (OR 1.25, 95% CI 1.22-1.29; $P<.001$). Highly rated male physicians had 1.33 higher odds of receiving a technical competence review (95% CI 1.30-1.36; $P<.001$); however, with controls, this gender difference flipped, such that highly rated females were more likely to receive a technical competence review (OR 0.95, 95% CI 0.93-0.98; $P<.001$).

Discussion

Principal Results

Our 2 classification models identified the presence and valence of patients' interpersonal manner and technical competence judgments with high precision, recall, and accuracy. Our models identified these 2 judgment dimensions from a broad training data set inclusive of reviews for female and male physicians, for primary care physicians and surgeons, and for low- and high-star-rated reviews.

Our interpersonal manner and technical competence models underperformed in classifying reviews with no judgment and negative valence relative to reviews with positive valence. However, the overall predictive accuracy of both models (90%) was higher than the rate of hand-coding agreement among the 4 investigators (Cohen κ=0.84 and 0.77 for interpersonal manner and technical competence, respectively).

Our models produced classification metrics comparable to those found in other studies that use fine-tuned RoBERTa algorithms for coding tasks. For example, researchers who used RoBERTa to detect polarizing versus nonpolarizing rhetoric in tweets written by Congress members reported a model with 90% predictive accuracy and a weighted F_1-score of 0.90 [13]. They compared their results to the Valence Aware Dictionary and Sentiment Reasoner, a dictionary-based sentiment analysis model, which demonstrated a 68% accuracy and F_1-score of 0.75. Another study that used RoBERTa to classify 5 classes of mental illness in Reddit posts reported a model with an F_1-score of 0.86 [24]. These researchers showed RoBERTa outperformed both BERT and long short-term memory, a nontransformer neural network text classifier (86% accuracy vs 82% and 72%). Last, researchers who forecasted star ratings from physician reviews written on RateMDs.com demonstrated an 84.6% accuracy and a mean F_1-score of 0.83 with their RoBERTa model, which outperformed other NLP models [25]. In comparison, our interpersonal manner and technical competence models were each 90% accurate with weighted F_1-scores of 0.89 and 0.90, respectively.

Although we did not compare our own RoBERTa models to other NLP algorithms, our accuracy scores perform equal to or better than prior methods used to code patients' judgments in web-based physician reviews. For example, in one study, investigators who hand-coded reviews for 4 broad thematic categories, including interpersonal manner and technical competence, reported an interrater reliability range of κ=0.8-1.0 [1]. Another study, which used dictionary-based text analysis to code for positive and negative soft skills reported a mean accuracy of 0.76 (range 0.42-0.92) [10]. The rates of hand-coding agreement for interpersonal manner and technical competence among our own 4 investigators were Cohen κ=0.84 and 0.77, respectively. Last, research has shown RoBERTa outperforms both other pretrained models and traditional machine-learning models (eg, support vector machines and random forests) when used for text classification tasks in the health domain [16,26].

Our NLP classification models overcome several limitations of prior research using hand-coding and dictionary-based methods to identify the prevalence and valence of patient judgments in web-based physician reviews. Hand-coding, although considered the gold standard, is time-intensive, which limits scalability. Prior studies using multiple coders could only analyze data from sample sizes of fewer than 1000 physician reviews [1,27]. Dictionary-based approaches enable analysis of larger samples but are restricted to the keywords contained in their dictionaries. Dictionary-based or bag-of-words models may overlook misspellings and jargon (eg, "he butchered my surgery"), leading to false negatives. They may also misidentify judgments about nonphysician staff (eg, front desk worker and nurse) or pick up words that have different meanings in different contexts (eg, "he is thorough in his examinations" vs "she gives thorough explanations"), leading to false positives. Dictionary-based models also have difficulty distinguishing between words used positively or with negations (eg, "she was smart" vs "she was not smart"), which complicates valence estimates. Prior research on physician reviews using dictionary-based models could not determine valence and only coded reviews that contained at least 1 preselected dictionary keyword [9,10,28,29].

We validated our classification models by examining associations between our coded judgment valence and review

XSL•FO
RenderX

star ratings and by comparing our coded judgments to findings from prior studies. We found positive interpersonal manner and technical competence judgments were associated with higher review star ratings whereas negative judgments were associated with lower review star ratings. Additionally, we found similar patterns of results with other studies that have examined the presence and valence of patients' judgments in web-based physician reviews. Future research should examine how both interpersonal manner and technical competence judgments vary depending on both physician gender and specialty.

We demonstrate that fine-tuning RoBERTa classification models to code patients' interpersonal manner and technical competence judgments in web-based physician reviews offers a scalable, reliable, and accurate method for analyzing unstructured textual review data. To our knowledge, we are the first to use an advanced NLP algorithm to code a large data set of web-based physician reviews for patients' judgments. This algorithm successfully coded web-based physician reviews, suggesting that RoBERTa may also be used to code similar unstructured text, including reviews from other commercial physician review websites (eg, RateMDs and ZocDoc) and from traditional Press Ganey patient-experience surveys. Whereas research has begun to use BERT-based models to extract health care insights from triage notes and medical records [30-32], future research is needed to ascertain the effectiveness of RoBERTa models with more far-afield text, such as crowdfunding campaigns, social media posts, and recommendation letters.

We acknowledge patients' judgments of their physicians' technical competence should be taken with caution. Prior research has shown weak correlations between patients' assessments of technical care quality and evidence-based indicators from clinical records [33]. Certifying boards and professional societies are better equipped to assess physicians' technical skills, such as knowledge of diagnostic and therapeutic advances [34]; however, patients' written reviews may be useful in offering reports of what actually occurred during clinical encounters, such as whether the physician checked their blood pressure or offered a flu vaccine [35]. Thus, for patient reviews to improve clinical care quality, future classification models for technical competence may consider focusing more narrowly on patient reports of technical processes rather than general perceptions of physicians' technical skills.

Limitations

Our study has 3 broad limitations, which arise from both the training data and the algorithm output. First, despite representing a small sample of our data set, our 2000 hand-coded reviews took substantial time and resources. It is possible that future studies can use smaller samples to train RoBERTa for classification tasks while maintaining high accuracy. Although more research is needed, we explored this idea by re-fine-tuning our models with gradually smaller training data sets (ie, n=1000, 500, and 250 reviews). For both interpersonal manner and technical competence, accuracy decreased with smaller training samples (interpersonal manner: 90%, 87%, 82%, and 64%; technical competence: 90%, 86%, 72%, and 64%). This brief example shows that future researchers could use 1000, possibly 500, but likely not 250 hand-coded reviews to train RoBERTa for multiclass classification. However, researchers should consider the implications of not only accuracy but also precision and recall for each code.

Second, our RoBERTa models were only as good as our training data, and our training data was imperfect. Although hand-coding is considered the gold standard and team members rigorously followed a coding framework, biases in how individual coders identified patients' interpersonal manner and technical competence judgments may have influenced the RoBERTa models. The imperfect interrater reliability present within the hand-coded data set is evidence of differences between coders, which may have complicated the fine-tuning of the models. Additionally, the 2000 hand-coded reviews represented only 0.6% of all reviews in the final data set; thus, our models could have been overfitted to this relatively small training data set.

We also excluded the classification of certain reviews in our training data. Because of our own language barriers, we trained the models on reviews only written in English. Reviews written in other languages, such as Spanish, were not translated and thus received codes of no interpersonal manner and no technical competence, despite potentially describing either judgment dimension. Reviews written in languages other than English, however, represented a small proportion of the total reviews in our sample. In addition, we only trained the models to classify patient judgments of physicians' interpersonal manner and technical competence, ignoring other judgments. Other judgments categorized as neither interpersonal manner nor technical competence included global remarks (eg, "would definitely recommend to others!" or "the worst") and system-level comments about the office, staff, or other aspects of the health care experience (eg, "dingy building" or "his assistant was the most wonderful person I have ever met").

Third, our RoBERTa models had limitations. Although RoBERTa offers a more advanced NLP algorithm than dictionary-based methods, the algorithm may still not recognize cultural jargon. The first illustrative example of Table 2, in which RoBERTa did not recognize the connotation of the #metoo reference, demonstrates this limitation. Moreover, because the RoBERTa algorithm was not trained on a prebuilt dictionary but on a reference set of hand-coded reviews, it is difficult to determine specific words and phrases the models used when classifying interpersonal manner and technical competence judgments. This transparency limitation, often called "black box AI" is a common problem with deep learning algorithms that create their own neural networks for categorization [36,37].

Despite these limitations, practical benefits of applying advanced NLP algorithms, like RoBERTa, to physician reviews include the enhanced capability to review feedback on what patients like and dislike about their medical encounters at clinician, department, or hospital levels; assistance in discerning differences in physician reviews received in traditional versus web-based surveys; and support in identifying patient biases, if any, corresponding with physician demographics. The benefits of these large language models also extend beyond insights from physician reviews. For example, advanced NLP models can improve patient care through medical information retrieval

XSL•FO
RenderX

from medical literature, drug databases, and treatment guidelines; and through personalized clinical decision support by analyzing relevant patient data, such as medical histories, test results, and clinician notes. These models can also reduce physician workload via documentation assistance.

Conclusion

We coded a large data set of web-based physician reviews for the presence and valence of patients' interpersonal manner and technical competence judgments using RoBERTa, a pretrained NLP classification algorithm. We trained and tested our models using a gold standard data set of hand-coded reviews and demonstrated that our models accurately and reliably coded interpersonal manner and technical competence. We also validated the algorithm by comparing our RoBERTa-coded data set with review star ratings and results from prior literature. The RoBERTa algorithm overcomes text analysis limitations present in previous work by identifying patient judgments in a broad range of physician reviews accurately and at scale. Potential benefits of advanced NLP models pertain to web-based physician reviews and beyond, from helping physicians more efficiently assess patient feedback to improving physicians' workload and patient care.

Acknowledgments

FM was affiliated with the Sanford School of Public Policy at Duke University at the time of the study and is currently affiliated with the Center for Bioethics and Social Sciences in Medicine at the University of Michigan Medical School. KT was affiliated with the Fuqua School of Business at Duke University at the time of the study and is currently affiliated with the University of Washington School of Medicine. AC was affiliated with the Center for Advanced Hindsight at Duke University at the time of the study and is currently affiliated with the Department of Sociology at UCLA. JKD was affiliated with the Fuqua School of Business at Duke University at the time of the study and is currently affiliated with the Department of Population Health Sciences at Duke University School of Medicine.

Conflicts of Interest

None declared.

References

1. López A, Detz A, Ratanawongsa N, Sarkar U. What patients say about their doctors online: a qualitative content analysis. J Gen Intern Med 2012;27(6):685-692 [FREE Full text] [doi: 10.1007/s11606-011-1958-4] [Medline: 22215270]
2. Kilaru AS, Meisel ZF, Paciotti B, Ha YP, Smith RJ, Ranard BL, et al. What do patients say about emergency departments in online reviews? a qualitative study. BMJ Qual Saf 2016;25(1):14-24. [doi: 10.1136/bmjqs-2015-004035] [Medline: 26208538]
3. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, et al. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. Health Aff (Millwood) 2016;35(4):697-705 [FREE Full text] [doi: 10.1377/hlthaff.2015.1030] [Medline: 27044971]
4. Xu Y, Armony M, Ghose A. The interplay between online reviews and physician demand: an empirical investigation. Manage Sci 2021;67(12):7344-7361. [doi: 10.1287/mnsc.2020.3879]
5. Trehan SK, DeFrancesco CJ, Nguyen JT, Charalel RA, Daluiski A. Online patient ratings of hand surgeons. J Hand Surg Am 2016;41(1):98-103. [doi: 10.1016/j.jhsa.2015.10.006] [Medline: 26710742]
6. Murphy GP, Radadia KD, Breyer BN. Online physician reviews: is there a place for them? Risk Manag Healthc Policy 2019;12:85-89 [FREE Full text] [doi: 10.2147/RMHP.S170381] [Medline: 31191060]
7. Burkle CM, Keegan MT. Popularity of internet physician rating sites and their apparent influence on patients' choices of physicians. BMC Health Serv Res 2015;15:416 [FREE Full text] [doi: 10.1186/s12913-015-1099-2] [Medline: 26410383]
8. Emmert M, Meszmer N, Sander U. Do health care providers use online patient ratings to improve the quality of care? results from an online-based cross-sectional study. J Med Internet Res 2016;18(9):e254 [FREE Full text] [doi: 10.2196/jmir.5889] [Medline: 27644135]
9. Chen H, Pierson E, Schmer-Galunder S, Altamirano J, Jurafsky D, Leskovec J, et al. Gender differences in patient perceptions of physicians' communal traits and the impact on physician evaluations. J Womens Health (Larchmt) 2021;30(4):551-556. [doi: 10.1089/jwh.2019.8233] [Medline: 32857642]
10. Dunivin Z, Zadunayski L, Baskota U, Siek K, Mankoff J. Gender, soft skills, and patient experience in online physician reviews: a large-scale text analysis. J Med Internet Res 2020;22(7):e14455 [FREE Full text] [doi: 10.2196/14455] [Medline: 32729844]
11. Nelson LK, Burk D, Knudsen M, McCall L. The future of coding: a comparison of hand-coding and three types of computer-assisted text analysis methods. Sociol Methods Res 2018;50(1):202-237. [doi: 10.1177/0049124118769114]
12. Li W, Jin B, Quan Y. Review of research on text sentiment analysis based on deep learning. Open Access Library J 2020;07(03):1-8. [doi: 10.4236/oalib.1106174]
13. Ballard AO, DeTamble R, Dorsey S, Heseltine M, Johnson M. Dynamics of polarizing rhetoric in congressional tweets. Legislative Studies Qtrly 2022;48(1):105-144. [doi: 10.1111/lsq.12374]

XSL•FO
RenderX

14. Oliveira FB, Haque A, Mougouei D, Evans S, Sichman JS, Singh MP. Investigating the emotional response to covid-19 news on twitter: a topic modeling and emotion classification approach. IEEE Access 2022;10:16883-16897. [doi: 10.1109/access.2022.3150329]
15. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv 2019:190711692. [doi: 10.48550/arXiv.1907.11692]
16. Guo Y, Dong X, Al-Garadi M, Sarker A, Paris C, Aliod DM. Benchmarking of transformer-based pre-trained models on social media text classification datasets. 2020 Presented at: Proceedings of the the 18th annual workshop of the australasian language technology association; 2020; Australia.
17. Healthgrades frequently asked questions (FAQs). healthgrades.com. URL: https://www.healthgrades.com/content/faqs [accessed 2022-10-08]
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. Attention is all you need. 2017 Presented at: Advances in Neural Information Processing Systems; 2017 June 12; Long Beach, CA, USA.
19. Liao W, Zeng B, Yin X, Wei P. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Appl Intell 2020;51(6):3522-3533. [doi: 10.1007/s10489-020-01964-1]
20. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv 2018:181004805. [doi: 10.48550/arXiv.1810.04805]
21. Rajapkse T. Simple transformers. Python Software Foundation. URL: https://pypi.org/project/simpletransformers/ [accessed 2024-02-15]
22. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: state-of-the-art natural language processing. 2020 Presented at: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; Online p. 38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]
23. Haynes D, Pampari A, Topham C, Schwarzenberger K, Heath M, Zou J, et al. Patient experience surveys reveal gender-biased descriptions of their care providers. J Med Syst 2021;45(10):90. [doi: 10.1007/s10916-021-01766-z] [Medline: 34468879]
24. Murarka A, Radhakrishnan B, Ravichandran S. Classification of mental illnesses on social media using RoBERTa. 2021 Presented at: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis; 2021 Apr 19; IBM / Raleigh, NC p. 59-68.
25. Jhaveri Y, Gandhi T, Naik T, Nisar S, Sonawane P. Predicting doctor ratings from user reviews using deep learning. 2022 Presented at: 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC); 2022 May 12; Salem, India. [doi: 10.1109/icaaic53929.2022.9793206]
26. Al-Garadi MA, Yang Y, Cai H, Ruan Y, O'Connor K, Graciela G, et al. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med Inform Decis Mak 2021;21(1):27 [FREE Full text] [doi: 10.1186/s12911-021-01394-0] [Medline: 33499852]
27. Marrero K, King E, Fingeret AL. Impact of surgeon gender on online physician reviews. J Surg Res 2020;245:510-515. [doi: 10.1016/j.jss.2019.07.047] [Medline: 31446193]
28. Gupta S, Jordan K. Understanding gender bias toward physicians using online doctor reviews. Psychol Lang Commun 2022;26(1):18-41. [doi: 10.2478/plc-2022-0002]
29. Saifee DH, Hudnall M, Raja U. Physician gender, patient risk, and web-based reviews: longitudinal study of the relationship between physicians' gender and their web-based reviews. J Med Internet Res 2022;24(4):e31659 [FREE Full text] [doi: 10.2196/31659] [Medline: 35394435]
30. Arnaud E, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. HEALTHINF 2022;5:835-841. [doi: 10.5220/0011012800003123]
31. Zeng K, Pan Z, Xu Y, Qu Y. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. JMIR Med Inform 2020;8(7):e17832 [FREE Full text] [doi: 10.2196/17832] [Medline: 32609092]
32. Zhang H, Ni W, Li J, Zhang J. Artificial intelligence-based traditional Chinese medicine assistive diagnostic system: validation study. JMIR Med Inform 2020;8(6):e17608 [FREE Full text] [doi: 10.2196/17608] [Medline: 32538797]
33. Rao M, Clarke A, Sanderson C, Hammersley R. Patients' own assessments of quality of primary care compared with objective records based measures of technical quality of care: cross sectional study. BMJ 2006;333(7557):19 [FREE Full text] [doi: 10.1136/bmj.38874.499167.7C] [Medline: 16793783]
34. Beliveau ME, Nishimura RA, O'Gara P. Physician competence: a perspective from the practicing cardiologist. Methodist Debakey Cardiovasc J 2014;10(1):50-52 [FREE Full text] [doi: 10.14797/mdcj-10-1-50] [Medline: 24932364]
35. Coulter A. Can patients assess the quality of health care? BMJ 2006;333(7557):1-2 [FREE Full text] [doi: 10.1136/bmj.333.7557.1] [Medline: 16809674]
36. Ratti E, Graves M. Explainable machine learning practices: opening another black box for reliable medical AI. AI Ethics 2022;2(4):801-814. [doi: 10.1007/s43681-022-00141-z]
37. von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol 2021;34(4):1607-1622. [doi: 10.1007/s13347-021-00477-0]

XSL•FO
RenderX

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
NLP: natural language processing
OR: odds ratio
RoBERTa: Robustly Optimized BERT Pretraining Approach

Edited by Y Zhuang; submitted 28.06.23; peer-reviewed by B Senst, M Elbattah, A Tomar, T Huang; comments to author 26.01.24; revised version received 21.03.24; accepted 29.04.24; published 01.08.24.

Please cite as:
Madanay F, Tu K, Campagna A, Davis JK, Doerstling SS, Chen F, Ubel PA
Classification of Patients' Judgments of Their Physicians in Web-Based Written Reviews Using Natural Language Processing: Algorithm Development and Validation
J Med Internet Res 2024;26:e50236
URL: https://www.jmir.org/2024/1/e50236
doi:10.2196/50236
PMID:39088259

©Farrah Madanay, Karissa Tu, Ada Campagna, J Kelly Davis, Steven S Doerstling, Felicia Chen, Peter A Ubel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Electronic Health Record–Based Automated Self-Rescheduling Tool to Improve Patient Access: Retrospective Cohort Study

Smitha Ganeshan[1*], MD, MBA; Andrew W Liu[2*], BA; Anne Kroeger[3], BA; Prerna Anand[3], MA; Richard Seefeldt[3], MS; Alexis Regner[3], MA; Diana Vaughn[3], MBA; Anobel Y Odisho[2], MD, MPH; Michelle Mourad[1], MD

[1]Department of Medicine, University of California San Francisco, San Francisco, CA, United States

[2]Department of Urology, University of California San Francisco, San Francisco, CA, United States

[3]UCSF Health Faculty Practices, University of California San Francisco, San Francisco, CA, United States

[*]these authors contributed equally

Corresponding Author:
Smitha Ganeshan, MD, MBA
Department of Medicine
University of California San Francisco
505 Parnassus Avenue, #M1493
San Francisco, CA, 94117
United States
Phone: 1 415 514 1000
Email: smitha.ganeshan@ucsf.edu

Abstract

Background: In many large health centers, patients face long appointment wait times and difficulties accessing care. Last-minute cancellations and patient no-shows leave unfilled slots in a clinician's schedule, exacerbating delays in care from poor access. The mismatch between the supply of outpatient appointments and patient demand has led health systems to adopt many tools and strategies to minimize appointment no-show rates and fill open slots left by patient cancellations.

Objective: We evaluated an electronic health record (EHR)–based self-scheduling tool, Fast Pass, at a large academic medical center to understand the impacts of the tool on the ability to fill cancelled appointment slots, patient access to earlier appointments, and clinical revenue from visits that may otherwise have gone unscheduled.

Methods: In this retrospective cohort study, we extracted Fast Pass appointment offers and scheduling data, including patient demographics, from the EHR between June 18, 2022, and March 9, 2023. We analyzed the outcomes of Fast Pass offers (accepted, declined, expired, and unavailable) and the outcomes of scheduled appointments resulting from accepted Fast Pass offers (completed, canceled, and no-show). We stratified outcomes based on appointment specialty. For each specialty, the patient service revenue from appointments filled by Fast Pass was calculated using the visit slots filled, the payer mix of the appointments, and the contribution margin by payer.

Results: From June 18 to March 9, 2023, there were a total of 60,660 Fast Pass offers sent to patients for 21,978 available appointments. Of these offers, 6603 (11%) were accepted across all departments, and 5399 (8.9%) visits were completed. Patients were seen a median (IQR) of 14 (4-33) days sooner for their appointments. In a multivariate logistic regression model with primary outcome Fast Pass offer acceptance, patients who were aged 65 years or older (vs 20-40 years; P=.005 odds ratio [OR] 0.86, 95% CI 0.78-0.96), other ethnicity (vs White; P<.001, OR 0.84, 95% CI 0.77-0.91), primarily Chinese speakers (P<.001; OR 0.62, 95% CI 0.49-0.79), and other language speakers (vs English speakers; P=.001; OR 0.71, 95% CI 0.57-0.87) were less likely to accept an offer. Fast Pass added 2576 patient service hours to the clinical schedule, with a median (IQR) of 251 (216-322) hours per month. The estimated value of physician fees from these visits scheduled through 9 months of Fast Pass scheduling in professional fees at our institution was US $3 million.

Conclusions: Self-scheduling tools that provide patients with an opportunity to schedule into cancelled or unfilled appointment slots have the potential to improve patient access and efficiently capture additional revenue from filling unfilled slots. The demographics of the patients accepting these offers suggest that such digital tools may exacerbate inequities in access.

(J Med Internet Res 2024;26:e52071) doi:10.2196/52071

XSL•FO
RenderX

KEYWORDS

appointment; consultation; cost; digital health; digital tools; electronic health record; EHR; informatics; patient access; retrospective review; revenue; self-rescheduling tool; self-scheduling; waiting time

Introduction

Health care access continues to be a concern as patients endure long wait times to access care [1,2]. In an industry survey, more than half of hospital leaders said that it takes more than 2 weeks to schedule patients on average and that access to specialty care is worse than before the COVID-19 pandemic [3]. Many health systems lack an efficient process to manage unfilled slots from cancellations. Unfilled cancellations and no-shows may exacerbate patient access issues. The process of scheduling and rescheduling appointments often requires significant time and manual labor, as staff members spend time calling patients to schedule, confirm, and reschedule appointments [4].

At busy medical centers, the high volumes of patient visits make this manual process of maintaining waitlists and filling cancelled appointments impractical. More recently, health care systems have used a range of tools to improve appointment completion rates and the process of scheduling patients into unfilled slots. Electronic health record (EHR)–based scheduling tools, phone call reminders, and automated SMS text message appointment reminders can all help ensure patients select an appointment time that works with their schedule, have the option to reschedule, and are prompted to confirm their appointment. EHR-based scheduling tools that allow patients to view available appointments as well as schedule, cancel, and reschedule their appointments through an app or a web-based portal saw a dramatic rise in use during the COVID-19 pandemic, with studies showing improved patient satisfaction [5-10]. EHR self-scheduling tools have been deployed in a variety of use cases, including the COVID-19 vaccine, radiology imaging, and well-child primary care visit scheduling [10,11]. Tools that allow patients to view and schedule themselves into available appointments may allow earlier and easier access for patients, reduce staff burden, and increase clinic volume. More advanced self-scheduling tools can allow for the maintenance of an electronic waitlist that can provide patients with notifications when an earlier appointment is available. Despite these reported benefits, data quantifying the efficacy of these tools, their impact on patient access, and their financial value for health systems are needed to help spur adoption and investment [12].

The University of California, San Francisco (UCSF) Medical Center implemented an EHR module, Fast Pass (Epic Systems Corporation). The Fast Pass module, which is available through Epic's suite of self-scheduling tools, allows a patient to add themselves to a waitlist for an earlier appointment slot. When a slot becomes available, patients receive an automated notification through SMS text message or email prompting them to log into their patient portal (MyChart) to self-schedule into the new, earlier appointment slot or to keep their existing appointment slot. As an existing module within the Epic EHR, included as part of the MyChart patient portal, there was limited integration necessary, as would be required with a third-party vendor. While the Fast Pass tool had previously been available through Epic, there had been no enterprise-wide implementation effort at UCSF to facilitate its use.

The primary objective of this study was to describe the uptake of the enterprise-wide Fast Pass scheduling tool among patients, understand the impacts of Fast Pass on patient wait times for appointments, and determine the potential incremental dollar value of visits that may have otherwise remained unscheduled. Our secondary aim was to study the uptake of the tool by specialty, given the unique patient needs and workflows in each specialty.

Methods

Setting

UCSF Health is a large academic health system with 3 campuses, over 1000 inpatient beds, and 9 primary care practices serving approximately 90,000 patients. UCSF Health has approximately 45,000 hospital admissions and 1.7 million outpatient visits annually. As of January 2022, approximately 89% of adult ambulatory care patients were enrolled in UCSF's EHR-tethered patient portal. There are a total of 1538 different department entities across all service areas at UCSF. Before November 2022, 103 different departments were sending offers through Fast Pass, and by January 2023, the number of departments sending Fast Pass offers had increased to 220.

Fast Pass Implementation at UCSF

To use Fast Pass, a patient opts into the program through the patient portal and elects to receive notifications through email, SMS text messaging, or both regarding earlier appointment slots as they become available. Fast Pass offers are sent in batches in the evening to multiple patients for a single appointment slot, and patients have 12 hours to sign up for the earlier slot, after which the offer expires. When patients log into their patient portal, they can self-schedule into or decline the earlier appointment. When a patient declines, they have the opportunity to receive another offer at a different time. Beginning in November 2022, a central multidisciplinary team was formed to implement the Fast Pass tool across all departments at UCSF. Staff at the clinics were trained on using and implementing this tool in their respective clinics. A tip sheet was developed to facilitate patient education and awareness, and the tool was added to the UCSF digital and website promotions.

Data Acquisition

Fast Pass offer data from June 18, 2022, to March 9, 2023, for the 220 departments included in the Fast Pass implementation were extracted from the EHR. The demographic and scheduling data associated with the Fast Pass offers were included in the data extraction. Demographic data included patient age, gender, race and ethnicity, insurance financial class, marital status, and primary language. Fast Pass data included the offering department, provider name, visit type, whether the patient had an existing appointment, offer sent date and time, offered slot date and time, visit length, and the date and time the offer was

XSL•FO
RenderX

first viewed. Visit types were stratified by in-person office visits (including radiology and procedures), video visits, and nonbillable phone calls. Nonbillable phone calls were excluded from the analysis. Clinics offering the Fast Pass feature were grouped into their respective specialties by a physician-informaticist familiar with the UCSF system and EHR department entities.

The primary exposure of our retrospective cohort study was the dichotomous variable of whether the clinic had access to the enterprise Fast Pass tool. We only analyzed results from clinics that had access to Fast Pass in this study. Fast Pass appointment offers were also stratified according to their 4 possible categorical outcomes: accepted, declined, expired, or unavailable (offer accepted by another patient first). Additionally, Fast Pass–accepted appointments were stratified by 3 possible categorical appointment outcome end points: completed, no-show, and cancelled.

Data Analysis

Descriptive statistics were performed. We compared differences in the patient cohort stratified by Fast Pass offer acceptance status (accepted vs nonaccepted) using the chi-square test for categorical features. A multivariate logistic regression model was built with primary outcome of offer acceptance status to determine predictors of offer acceptance and the size of associations between primary outcome and the demographic factors extracted from the EHR as additional independent variables chosen by theoretical criteria. No stepwise regression was conducted. Offer acceptance status was used as the primary outcome due to Fast Pass functionality primarily being designed to increase the number of appointment slots filled rather than decrease the number of no-shows or cancellations. All analysis was conducted in R (version 3.5.1; R Foundation for Statistical Computing), and a value of $P<.05$ was considered significant.

Analysis of Fast Pass by Departments

Descriptive analyses were conducted on the outcomes of offers as well as the outcomes of accepted appointments, stratified by specialty.

Visit Dollar Value Calculation

We extracted the Fast Pass offers that were accepted and completed in a 9-month period and calculated the dollar value of the professional fees for these visits. We used the number of visits that Fast Pass added to the schedule within 7 days of the appointment in the 9 months of implementation, the payer mixes for these visits, and the average professional fee collections for the visits by specialty and by payer to estimate the revenue of the added visits. We hypothesize that visits scheduled within 7-days of an appointment were more likely to remain unfilled. These revenue numbers only include physician evaluation and management revenue, not facility or technical fees. We do not include downstream revenue associated with patient visits. While these Fast Pass visits do not represent true net new patient service revenue, they help fill a proportion of cancelled slots that may otherwise have gone unfilled and may create downstream capacity for new patients.

Ethical Considerations

This study was reviewed and approved by the UCSF institutional review board (22–35948) and was determined to not be human participants research. Epic Systems did not fund this study or participate in the analysis, and it is a paid vendor of UCSF.

Results

From June 18, 2022, to March 9, 2023, there were a total of 60,660 Fast Pass offers for 21,978 appointments sent to patients. As of March 17, 2023, there were 182 offers currently active or deleted on the clinic's end, for a total of 60,478 offers analyzed. The median (IQR) number of requests sent for each appointment was 2 (1-3). Out of 60,478 offers, there were 6703 (11%) offers accepted, for a monthly median of 139 visits across all departments. Of the 21,978 appointments sent, 6703 offers were accepted, for a 30.5% success rate. Of the accepted offers, 294 were beyond the time period of this analysis and therefore excluded, resulting in 6409 total analyzed accepted offers. Of the accepted offers within the time period of this study, 5399 appointments were completed, resulting in an overall completed appointment from the Fast Pass offer rate of 84.2%.

The cohort of patients who accepted their offer was significantly more likely to be older, male, White, have English as their primary language, and have an in-person visit, a physician visit, and a shorter appointment as part of their Fast Pass offer (Table 1). There was no difference in Fast Pass offer acceptance by insurance class. In addition, among patients that accepted a Fast Pass offer, patients who were older, male, of White race or ethnicity, had commercial insurance, saw a physician provider, and had a shorter appointment time were more likely to complete their appointment. There were no differences in appointment completion rates between visit types or primary languages.

In a multivariate regression model with the primary outcome of Fast Pass appointment offer acceptance, patients who were aged 65 years or older (odds ratio [OR] 0.86, 95% CI 0.78-0.96) versus patients who were aged younger than 40 years, other ethnicity (OR 0.84, 95% CI 0.77-0.91) versus White patients, primarily Chinese speakers (OR 0.62, 95% CI 0.49-0.79) and other language speakers (OR 0.71, 95% CI 0.57-0.87) versus English-speaking patients were less likely to accept an offer (Table 2). Male patients (OR 1.09, 95% CI 1.03-1.15) were more likely to accept their offer compared with female patients. In terms of the Fast Pass offer details, compared to office visits, phone calls were 41% less likely to be accepted (OR 0.59, 95% CI 0.38-0.86), while video visits were 29% more likely to be accepted (OR 1.29, 95% CI 1.20-1.38). A Fast Pass offer with a physician was 55% more likely to be accepted (OR 1.55, 95% CI 1.46-1.63) compared with a nonphysician offer, while appointments longer than 45 minutes (OR 1.20, 95% CI 1.09-1.32) and between 20 minutes and 45 minutes (OR 1.15, 95% CI 1.09-1.22) were both more likely to be accepted versus those less than 20 minutes. Patients were seen a median (IQR) of 14 (7-33) days before their originally scheduled appointment for those that eventually completed their visits from a Fast Pass offer (Figure 1).

XSL•FO
RenderX

Table 1. Demographics of Fast Pass offers stratified by offer acceptance.

	Accepted (n=6703), n (%)	Not accepted (n=53,775), n (%)	*P* value
Patient age (years)			.01
18-40	1637 (24.4)	12,307 (22.9)	
40-64	2839 (42.4)	23,033 (42.8)	
≥65	2227 (33.2)	18,435 (34.3)	
Gender			.004
Male	2585 (38.6)	19,649 (36.6)	
Female	4038 (60.2)	33,418 (62.1)	
Other or unknown	80 (1.2)	708 (1.3)	
Ethnicity			<.001
White	3807 (56.8)	28,825 (53.6)	
Black or African American	318 (4.7)	2717 (5.1)	
Hispanic or Latino	649 (9.7)	5102 (9.5)	
Asian, Native Hawaiian, or other Pacific Islander	1137 (17.0)	99,76 (18.6)	
Other or unknown	792 (11.8)	7155 (13.3)	
Primary language			<.001
English	6462 (96.4)	50,832 (94.5)	
Spanish	68 (1.0)	677 (1.3)	
Chinese	76 (1.1)	1110 (2.1)	
Other	97 (1.4)	1156 (2.1)	
Insurance			.43
Commercial	3805 (56.9)	30,217 (56.3)	
Medicare	703 (29.8)	5715 (30.0)	
Medicaid	1994 (10.5)	16,099 (10.6)	
Other	186 (2.8)	1676 (3.1)	
Provider type			<.001
Physician	3994 (59.6)	27,933 (48.1)	
Nonphysician	2709 (40.4)	25,842 (41.9)	
Offered visit type			<.001
Office visit, radiology, or procedure	5212 (77.9)	45,109 (84.0)	
Video visit	1465 (21.9)	291 (0.5)	
Nonbillable phone call	26 (0.4)	8375 (15.6)	
Offered appointment length (minutes)			<.001
<20	2813 (42.1)	24,580 (45.8)	
20-45	3183 (47.5)	24,215 (45.0)	
>45	700 (10.4)	4965 (9.2)	

XSL•FO
RenderX

Table 2. Multivariate predictive model for Fast Pass use.

Term	OR[a]	95% CI	*P* value
Patient age (vs <40 years)			
40-64	0.96	0.89-1.01	.11
≥65	0.86	0.78-0.96	.005
Sex (vs female)			
Male	1.09	1.03-1.15	.002
Other or unknown	0.91	0.71-1.15	.45
Ethnicity (vs White)			
Black or African American	0.92	0.81-1.04	.19
Hispanic or Latino	1.03	0.94-1.14	.49
Asian, Native Hawaiian, or other Pacific Islander	0.97	0.90-1.04	.39
Other or unknown	0.84	0.77-0.91	<.001
Primary Language (vs English)			
Spanish	0.79	0.60-1.02	.09
Chinese	0.62	0.49-0.79	<.001
Other	0.71	0.57-0.87	.001
Insurance (vs commercial)			
Medicare	1.06	0.97-1.16	.21
Medicaid	1.04	0.94-1.14	.46
Other	0.93	0.79-1.08	.36
Visit type (vs office visit)			
Nonbillable phone call	0.59	0.38-0.86	<.001
Video visit	1.29	1.20-1.38	<.001
Provider type (vs nonphysician)			
Physician	1.55	1.46-1.63	<.001
Appointment length (minutes)			
20-45	1.15	1.09-1.22	<.001
>45	1.20	1.09-1.32	<.001

[a]OR: odds ratio.

XSL•FO
RenderX

Figure 1. Distribution of number of days improvement from existing appointment and rescheduled appointment through Fast Pass.

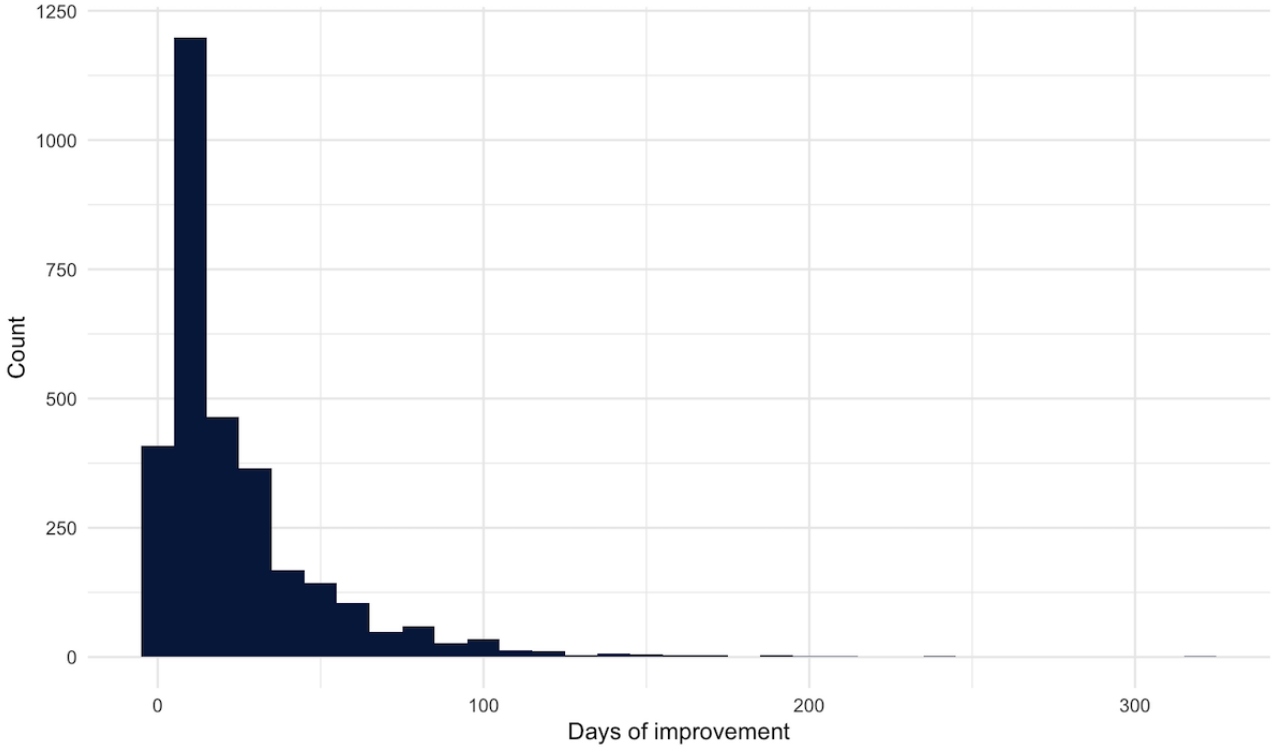

In a multivariate regression model with primary outcome Fast Pass appointment outcome post–offer acceptance, older patients (aged 40-64 years: OR 1.32, 95% CI 1.11-1.56; *P*=.001; and aged 65 years or older: OR 1.68, 95% CI 1.27-2.22; *P*<.001) versus patients aged 40 years or younger, and male patients (OR 1.35, 95% CI 1.17-1.57; *P*<.001) were more likely to complete their appointment. Patients with Medicaid (OR 0.51, 95% CI 0.42-0.63; *P*<.001), Medicare (OR 0.64, 95% CI 0.50-0.84; *P*<.001), and other insurance (OR 0.47, 95% CI 0.33-0.68; *P*<.001) versus commercial, and patients with appointment lengths longer than 45 minutes (OR 0.79, 95% CI 0.63-1.00; *P*=.05) compared with appointments less than 20 minutes in length were less likely to complete their appointment. There were no observed differences in the patient cohort in terms of race or ethnicity, primary language, visit type, or provider type for appointment completion.

There was differential uptake of Fast Pass among the specialties at our institution, both by offers accepted and by rates of completed appointments from Fast Pass offers (Figure 2). Oncology (49/211, 23.2%), nephrology (78/431, 18.1%), and pulmonology (70/394, 17.8%) were the specialties with the highest rates of fast pass scheduling offers accepted, while optometry (127/1831, 6.9%), radiology (1459/19134, 7.6%), and integrative medicine (114/283, 8.9%) had the lowest percentage of accepted fast pass offers. Rheumatology (139/155, 89.6%), endocrinology (69/77, 89.6%), and nephrology (66/75, 88.0%) had the highest rates of completed appointments using fast pass offers, while optometry (92/124, 74.1%), obstetrics and gynecology (314/429, 73.2%), and infectious diseases (31/42, 73.8%) had the lowest percentage (Figure 3).

In the Fast Pass revenue analysis, a total of 5387 of the 5399 completed visits added in the first 9 months of implementation across 25 specialties had payer information available and were included in the analysis (Table 3). These visits represent an estimated US $3 million in professional fees.

XSL•FO
RenderX

Figure 2. Outcomes of Fast Pass offers: unavailable, expired, declined, and accepted.

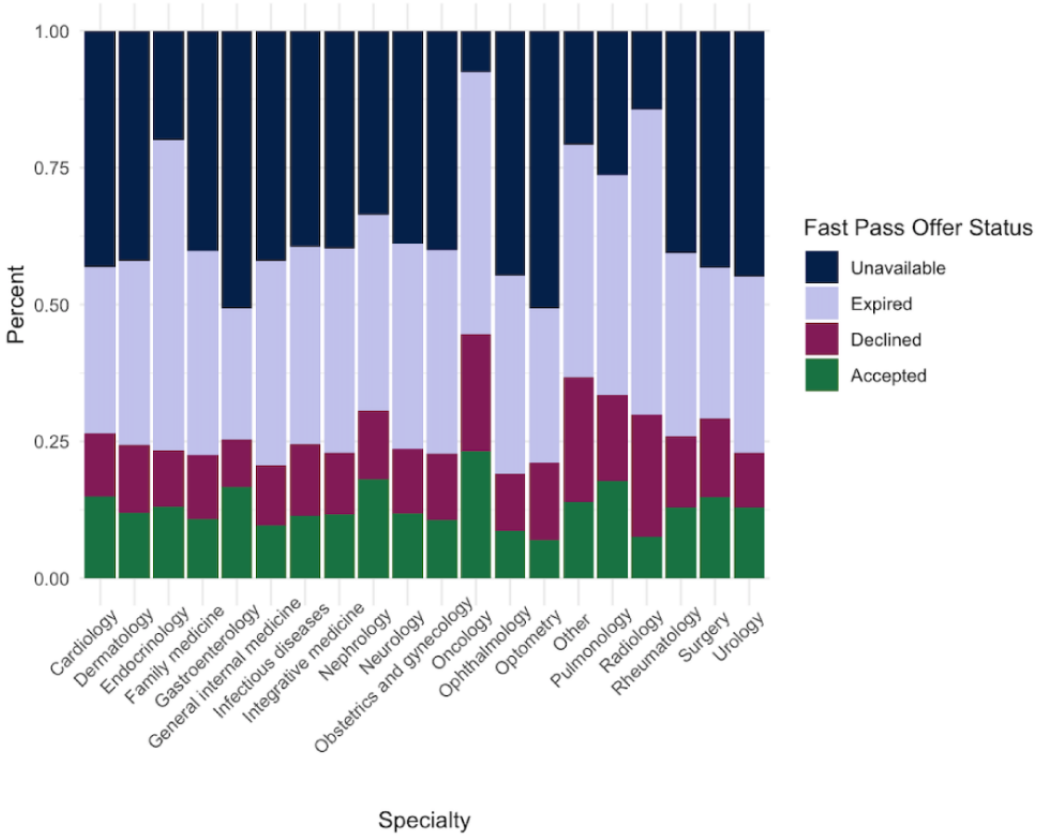

Figure 3. Outcomes of accepted Fast Pass offers: no show, canceled, and completed.

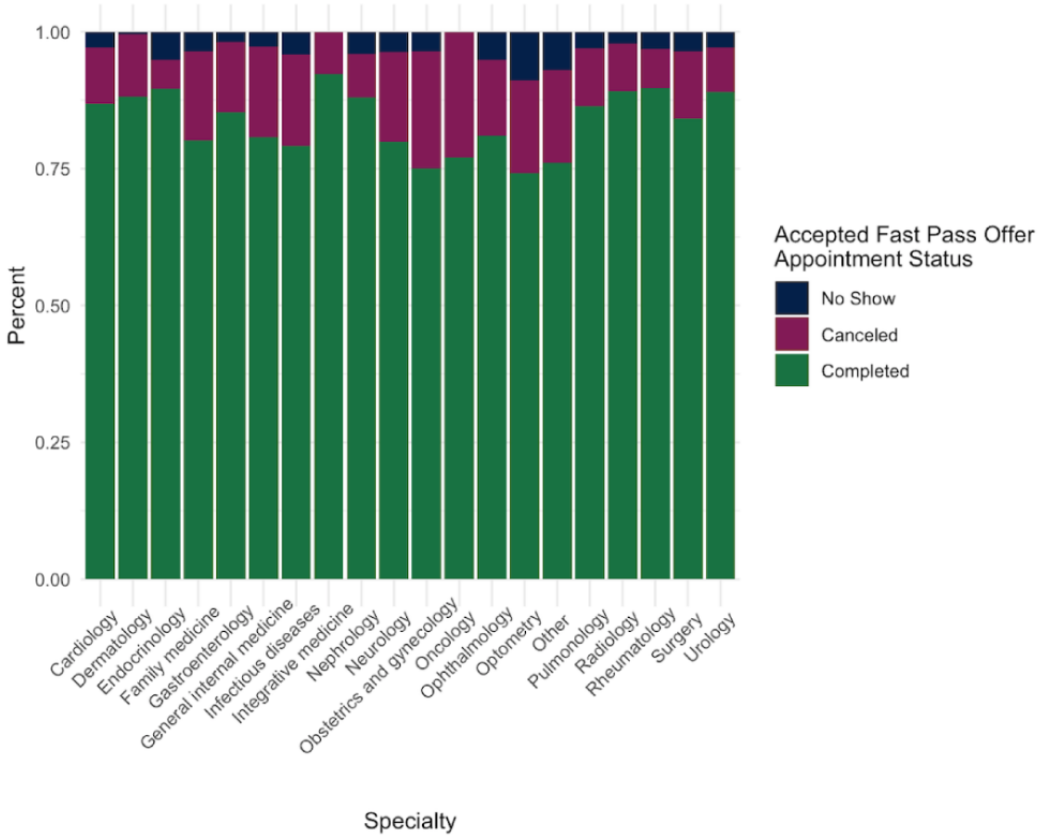

XSL•FO
RenderX

Table 3. Dollar value of physician fees from Fast Pass visits.

Specialty	Commercial, n	Medicare, n	Medicaid, n	Other, n	Total, N	Total revenue (US $)
Allergy and immunology	14	5	1	0	20	7819.00
Cardiology	126	187	15	1	329	103,446.66
Dermatology	183	21	4	1	209	72,043.80
Endocrinology	41	15	10	2	68	25,550.09
Family medicine	66	33	10	10	119	29,710.25
Gastroenterology	93	51	25	6	175	67,427.70
General internal medicine	326	120	28	11	485	130,770.54
Hematology	4	4	0	0	8	2760.16
Infectious diseases	11	17	3	0	31	9159.47
Integrative medicine	40	29	8	2	79	24,713.67
Nephrology	19	42	5	0	66	16,908.58
Neurology	76	81	13	4	174	53,757.85
Obstetrics and gynecology	244	29	29	11	313	113,294.91
Occupational health	1	0	0	0	1	275.89
Oncology	23	10	2	2	37	14,267.10
Ophthalmology	48	58	13	5	124	26,609.56
Optometry	37	32	16	6	91	20,016.01
Other	121	63	22	9	215	67,280.05
Otolaryngology	16	6	0	0	22	7907.58
Palliative care	2	4	1	0	7	1848.01
Pulmonology	22	27	7	1	57	17,875.52
Radiology	819	240	177	36	1272	1656,417.00
Rheumatology	82	40	17	0	139	49,700.57
Surgery	563	340	73	20	996	308,187.04
Transplant	1	2	1	0	4	1079.55
Urology	159	149	31	7	346	103,267.57
Total	3137	1605	511	134	5387	2,932,094

Discussion

Overview

FFast Pass is an EHR-based tool that allows patients to schedule into earlier appointment slots and allows clinics to fill unfilled appointments without the significant manual staff time needed to call and reschedule appointments in a timely manner. This study found that patients saw a median improvement in their appointment time slot between existing and rescheduled appointments of 14 days, which supports the ability of self-scheduling tools to facilitate quicker patient access to care. From June 18, 2022, to March 9, 2023, there were a total of 60,660 Fast Pass offers sent to patients. Of these offers, 6703 (11%) Fast Pass offers were accepted across all departments, for a monthly median of 139 visits across all departments. The median (IQR) number of requests for each appointment slot was 2 (1-2), which further highlights that this tool can free up staff time for direct patient care tasks or to assist patients who are less familiar with the EHR patient portal. This study builds on existing literature on the Fast Pass EHR tool by supporting its feasibility and implementation in a large tertiary health system and adding additional data on patient use of the tool, impacts across specialties, and potential revenue impacts [12].

When we analyzed Fast Pass offers by specialty, we found differential uptake of Fast Pass among the specialties at our institution. Oncology, nephrology, and pulmonology were the specialties with the highest rates of fast pass scheduling offers, while optometry, radiology, and integrative medicine had the lowest percentage of accepted fast pass offers. It is possible that our oncology, nephrology, and pulmonology clinics have longer wait times for ambulatory visits or higher cancellation rates, and therefore, the clinics and patients gain more value from automated Fast Pass offers and scheduling. Additional qualitative research may help us understand the drivers of differential uptake to better facilitate enterprise-wide implementation efforts of scheduling tools.

XSL•FO
RenderX

In our analysis, Fast Pass added 2576 patient service hours to the clinical schedule, with a median (IQR) of 251 (216-322) hours per month. The dollar value of physician fees from these visits through Fast Pass scheduling for 9 months of Fast Pass at our institution was approximately US $3 million. While this dollar value does not necessarily equate to the net additional revenue impact of Fast Pass, it does help quantify the financial impacts of self-scheduling tools since Fast Pass is an add-on tool in the EHR with minimal incremental cost [4]. Collecting additional longitudinal data will help us better understand the overall impacts of Fast Pass on ambulatory visit volumes, template use, patient access, and no-show rates.

This study builds on existing demographic data on digital tools, which show that patients who are White, in a younger age group, and have commercial insurance are more likely to use the tool [11,13-17]. These data raise concerns about equity in access to digital tools as health systems increasingly rely on EHR applications to facilitate front- and back-office functions, including triage and scheduling. Our data suggests that Fast Pass and similar tools may limit access to earlier appointments for older, non-White patients with a preferred language other than English. Given the personnel cost savings, these digital tools are likely to increase in use, and resources need to be channeled into supporting equitable adoption by all patients, or in the case of scheduling, ensuring that a system is put in place for offering patients earlier appointments based on clinical urgency. Possible interventions that may be undertaken to improve health equity in the area of self-scheduling include outreach of the tool in multiple languages or product design to increase accessibility for patients with low technological literacy.

Limitations

Our study has several limitations. This experience at an urban academic medical center with a wide clinical catchment area across California may not be generalizable to other regions of the country or other institutions, especially facilities with fewer subspecialty referrals or with different EHRs lacking a scheduling tool directly connected with their patient portal. This was also designed as a retrospective study. Future prospective natural experiments that test different modes of patient and clinician engagement could identify the key factors necessary for the successful implementation of this program enterprise-wide. In addition, the use of Fast Pass was limited to patients who had access to their MyChart portal. Previous data have demonstrated inequities in the ability to access and use MyChart, which may further compound inequities in using the Fast Pass tool. We did not have data to access patient digital literacy, which may play an important role in Fast Pass use. Further qualitative data are needed to better understand the specific patient and clinical workflow factors that lead to differential use of Fast Pass across specialties and departments.

Conclusions

Fast Pass, an EHR-based self-scheduling tool, afforded patients the opportunity for an earlier appointment and provided our medical center with the opportunity to efficiently capture revenue from cancelled appointments. We found that the patients accepting these offers were more likely to be older, male, and English-speaking, suggesting that these digital tools could exacerbate inequities in access.

Authors' Contributions

SG and AWL had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. SG, AWL, AYO, and MM conceived of the study concept and design and reviewed and provided critical feedback on manuscript drafts. AK, PA, RS, AR, and DV were instrumental in building and implementing the Fast Pass tool, provided feedback on the study concept and design, and also provided critical feedback on drafts. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Kyle MA, Tipirneni R, Thakore N, Dave S, Ganguli I. Primary care access during the COVID-19 pandemic: a simulated patient study. J Gen Intern Med 2021 Dec;36(12):3766-3771 [FREE Full text] [doi: 10.1007/s11606-021-06804-7] [Medline: 33904036]
2. Berchick ER, Hood E, Barnett JC. Health insurance coverage in the United States: 2017 current population reports. United States Census Bureau. 2018. URL: https://www.census.gov/library/publications/2018/demo/p60-264.html [accessed 2024-01-30]
3. Hospital patient volumes move towards 2019 levels. McKinsey & Company. URL: https://www.mckinsey.com/industries/healthcare/our-insights/survey-us-hospital-patient-volumes-move-back-towards-2019-levels [accessed 2023-07-27]
4. Judson TJ, Odisho AY, Neinstein AB, Chao J, Williams A, Miller C, et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. J Am Med Inform Assoc 2020 Jun 01;27(6):860-866 [FREE Full text] [doi: 10.1093/jamia/ocaa051] [Medline: 32267928]

XSL•FO
RenderX

5. Cao W, Wan Y, Tu H, Shang F, Liu D, Tan Z, et al. A web-based appointment system to reduce waiting for outpatients: a retrospective study. BMC Health Serv Res 2011;11:318 [FREE Full text] [doi: 10.1186/1472-6963-11-318] [Medline: 22108389]
6. Zhang X, Yu P, Yan J. Patients' adoption of the e-appointment scheduling service: a case study in primary healthcare. Stud Health Technol Inform 2014;204:176-181. [doi: 10.3233/978-1-61499-427-5-176]
7. Lyles CR, Nelson EC, Frampton S, Dykes PC, Cemballi AG, Sarkar U. Using electronic health record portals to improve patient engagement: research priorities and best practices. Ann Intern Med 2020 Jun 02;172(11 Suppl):S123-S129 [FREE Full text] [doi: 10.7326/M19-0876] [Medline: 32479176]
8. North F, Nelson EM, Majerus RJ, Buss RJ, Thompson MC, Crum BA. Impact of web-based self-scheduling on finalization of well-child appointments in a primary care setting: retrospective comparison study. JMIR Med Inform 2021 Mar 18;9(3):e23450 [FREE Full text] [doi: 10.2196/23450] [Medline: 33734095]
9. Han HR, Gleason KT, Sun CA, Miller HN, Kang SJ, Chow S, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019;6(4):e15038 [FREE Full text] [doi: 10.2196/15038] [Medline: 31855187]
10. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
11. Wallace LS, Angier H, Huguet N, Gaudino JA, Krist A, Dearing M, et al. Patterns of electronic portal use among vulnerable patients in a nationwide practice-based research network: from the OCHIN Practice-based Research Network (PBRN). J Am Board Fam Med 2016;29(5):592-603 [FREE Full text] [doi: 10.3122/jabfm.2016.05.160046] [Medline: 27613792]
12. Chung S, Martinez MC, Frosch DL, Jones VG, Chan AS. Patient-centric scheduling with the implementation of health information technology to improve the patient experience and access to care: retrospective case-control analysis. J Med Internet Res 2020 Jun 10;22(6):e16451 [FREE Full text] [doi: 10.2196/16451] [Medline: 32519970]
13. Denizard-Thompson NM, Feiereisel KB, Stevens SF, Miller DP, Wofford JL. The digital divide at an urban community health center: implications for quality improvement and health care access. J Community Health 2011;36(3):456-460. [doi: 10.1007/s10900-010-9327-5] [Medline: 21086028]
14. Anthony DL, Campos-Castillo C, Lim PS. Who isn't using patient portals and why? Evidence and implications from a national sample of US adults. Health Aff (Millwood) 2018;37(12):1948-1954 [FREE Full text] [doi: 10.1377/hlthaff.2018.05117] [Medline: 30633673]
15. Health information technology: HHS should assess the effectiveness of its efforts to enhance patient access to and use of electronic health information. U.S. GAO. URL: https://www.gao.gov/products/gao-17-305 [accessed 2021-10-24]
16. Grossman LV, Creber RMM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
17. Antonio MG, Petrovskaya O, Lau F. The state of evidence in patient portals: umbrella review. J Med Internet Res 2020;22(11):e23851 [FREE Full text] [doi: 10.2196/23851] [Medline: 33174851]

Abbreviations

EHR: electronic health record
UCSF: University of California, San Francisco

Edited by T de Azevedo Cardoso; submitted 21.08.23; peer-reviewed by H Veldandi, A Gangadhara Rao, C Asuzu; comments to author 21.11.23; revised version received 23.12.23; accepted 22.01.24; published 19.03.24.

Please cite as:
Ganeshan S, Liu AW, Kroeger A, Anand P, Seefeldt R, Regner A, Vaughn D, Odisho AY, Mourad M
An Electronic Health Record–Based Automated Self-Rescheduling Tool to Improve Patient Access: Retrospective Cohort Study
J Med Internet Res 2024;26:e52071
URL: https://www.jmir.org/2024/1/e52071
doi:10.2196/52071
PMID:38502159

©Smitha Ganeshan, Andrew W Liu, Anne Kroeger, Prerna Anand, Richard Seefeldt, Alexis Regner, Diana Vaughn, Anobel Y Odisho, Michelle Mourad. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Broadband Infrastructure Construction on Medical Resource Mismatch: Quasi-Natural Experiment From the Broadband China Policy

Yulin Chai[1], MS; Xiaoping Yuan[2], PhD; Lin Guo[1], PhD; Zhongming Chen[1], PhD

[1]School of Management, Shandong Second Medical University, Weifang, China

[2]School of Psychology, Shandong Second Medical University, Weifang, China

Corresponding Author:
Lin Guo, PhD
School of Management
Shandong Second Medical University
7166 Baotong West St
Weifang, 261053
China
Phone: 86 0536 8462573
Email: guolin19860818@163.com

Abstract

Background: Whether the construction of broadband infrastructure can alleviate the problem of mismatched medical resources is crucial to the national information strategy, residents' well-being, and social equity. However, the academic community lacks a comprehensive theoretical analysis and rigorous empirical research on this issue.

Objective: This study aims to construct a preliminary theoretical framework to scientifically assess the effects of broadband infrastructure development on the mitigation of health care resource mismatch from both theoretical and empirical perspectives, to explore the potential mechanisms of influence, and ultimately to develop several practical policy recommendations.

Methods: We first used a theoretical analysis to propose testable theoretical hypotheses and establish a preliminary theoretical framework. Then, based on balanced panel data from 300 cities from 2010 to 2021, a 2-way fixed effects difference-in-differences model was used for empirical testing. Mechanism tests, robustness analyses, and heterogeneity analyses were further conducted.

Results: The research findings demonstrate that the Broadband China Policy significantly reduces the degree of mismatch in medical resources by primarily using innovation effects and integration effects, resulting in a reduction of 13.2%. In addition, the heterogeneity analysis reveals that the central and eastern regions, cities with large populations, and areas with a high proportion of young people benefit more significantly.

Conclusions: This study fully confirms, both theoretically and empirically, that broadband infrastructure construction can effectively reduce the mismatch of medical resources not only by expanding the existing literature on the impact of broadband on public services but also by providing valuable opportunities for policy makers to optimize the allocation of medical resources.

(J Med Internet Res 2024;26:e53921) doi:10.2196/53921

KEYWORDS

broadband; infrastructure; medical resources; resource mismatch; Broadband China Policy

Introduction

Background

Medical resources encompass disease prevention, treatment, medication supply, and nursing care services. It includes hospitals, pharmaceuticals, and diagnostic resources [1]. The proper allocation of medical resources is crucial for assessing a country's comprehensive strength. China is experiencing a severe mismatch in medical resources, with uneven distribution among regions, inefficiency in medical services, and a lack of collaboration among medical institutions. The misalignment between medical resources and service providers exacerbates this situation [2,3]. Studies show a substantial geographic mismatch, with Grade 3A hospitals outnumbering primary medical institutions. The severe mismatch hampers social and economic development goals, exacerbates social inequality, and

XSL•FO
RenderX

diminishes residents' well-being [4-6]. Therefore, it is imperative to rectify this mismatch.

Globally, broadband is propelling a new wave of IT development. Many countries prioritize it as a strategic area to enhance national competitiveness. Since 2014, China has been implementing the Broadband China Policy (BCP) to promote broadband and network infrastructure construction. This study examines the impact of broadband infrastructure on medical resource mismatches (MMs), aiming to optimize government policies and improve the mismatch of medical resources.

In theory, the construction of broadband infrastructure can affect the efficiency of medical resource allocation. With the advent of the third industrial revolution, broadband and other network infrastructures have the potential to greatly enhance the mobility of factors, expedite resource processing and integration, and enhance the efficiency of resource allocation [7,8]. The construction of broadband infrastructure can enhance medical resource allocation efficiency and promote the integration of the internet and health care, providing real-time, convenient, and high-quality medical services [9]. In fact, the construction of broadband infrastructure has initially shown results. With the deepening of BCP promotion, the construction rate of provincial-, municipal-, and county-level regional health information platforms in China has reached 100%, 62.8%, and 46.4%, respectively. A total of >7000 public hospitals at or above the secondary level nationwide have access to the regional national health information platform, and >2200 tertiary hospitals have initially achieved interhospital medical service information sharing. Can the construction of broadband infrastructure truly reduce the mismatch in medical resources? What is the real effect, and what are the potential mechanisms? Unfortunately, the academic community has not paid enough attention to these questions. Existing research mainly qualitatively believes that BCP has potential in the medical and health field [10] and can alleviate the shortage of medical resources in remote areas [11] and promote the balanced allocation of medical resources [12]. This study will attempt to scientifically answer these questions from theoretical and empirical perspectives, which can effectively fill the gap in the academic community and have important theoretical value.

To identify the causal relationship between broadband infrastructure construction and mismatched medical resources, there are inherent endogeneity challenges. Broadband infrastructure construction may reduce the problem of mismatched medical resources, whereas regions with low levels of mismatched medical resources may indicate high levels of information technology, improving local broadband infrastructure. Fortunately, in 2013, the State Council issued the *Broadband China* strategy and implementation plan, gradually promoting infrastructure construction. This had a significant exogenous impact on broadband infrastructure levels, with pilot areas having higher levels and nonpilot areas having lower levels.

Policy Background

Broadband infrastructure is recognized as a strategic public asset that is essential for China's social development in the contemporary era. It plays a vital role in supporting various sectors and driving global information advancement. Broadband infrastructure, embodied by 5G technology and fiber optics, serves as the underpinning for worldwide digitalization, networking, and intelligent development, with significant potential to reshape industrial patterns and the societal division of labor. The global community, in pursuit of a sustained competitive advantage, has introduced various initiatives for the development of broadband infrastructure. Of note among these efforts are the European Union's *Connecting Europe Facility—Digital*, aiming to enhance broadband infrastructure coverage across Europe within 3 years; the United States's *Internet for All* program, designed to deploy broadband infrastructure in underserved areas by 2030 with an investment of US $45 billion; and Japan's *Digital Garden City Nation*, setting the target of achieving fiber-to-the-home coverage for 99.9% of households by the end of 2027. These endeavors underscore the worldwide recognition of broadband infrastructure as a crucial investment for national progress and development.

In 2013, the State Council of China issued the *Broadband China* strategy and implementation plan [13], which proposed the phased promotion of broadband and other network infrastructure. In 2015, *Guiding Opinions on Accelerating the Construction of High-Speed Broadband Networks and Promoting Network Speed Increase and Fee Reduction* was issued [14]. By the end of 2017, the goals were: 100 Mbps fiber optic access capability in urban areas of prefecture-level cities and most urban areas in nonprefecture-level cities; major cities achieved an average broadband access rate exceeding 30 Mbps and other urban areas reaching 20 Mbps; and comprehensive 4G network coverage in both urban and rural areas, with >80% of rural areas having fiber optic access.

The Ministry of Industry and Information Technology and the National Development and Reform Commission have chosen 120 *Broadband China* demo areas in 3 batches from 2014 to 2016. After selection, local governments will focus on expanding broadband use, increasing network speed and coverage, and advancing infrastructure. According to the construction standards [15], demo cities must meet specific standards. Therefore, the selected cities will reach a leading level in terms of broadband infrastructure construction in the country.

Amid widespread broadband coverage, the evolution of health care information has gained momentum. There has been a continual rollout of specifications for information construction in health care, hastening the integration of health care information systems and the collaborative development of multilevel hospitals. The COVID-19 pandemic has further accelerated the demand for integrated and standardized health care platform construction. Seizing the advantages of broadband infrastructure and driving digitization and information enhancement in health care represents a significant initiative for addressing the mismatch of medical resources. At present, the construction rates for regional medical and health care information platforms stand at 100% for the provincial level, 62.8% for the municipal level, and 46.4% for the county level. Moreover, >7000 public hospitals at the secondary level and above nationwide are linked to the regional population health

XSL•FO
RenderX

information platform. In addition, >2200 Grade 3A hospitals have made initial strides in achieving interconnectivity and sharing medical service information within their facilities [16]. The impact of broadband infrastructure construction on mitigating the mismatch of medical resources is yielding tangible outcomes.

Literature Review

Digital technology has completely transformed health care services, and the application of health information storage and retrieval has grown exponentially [10]. Most public and private medical institutions have announced digitization. The digitalization of medical information can improve profits, reduce waste, enhance the availability of commercial insurance for residents, and improve the efficiency of medical services to reduce residents' medical expenses [17,18]. Through electronic medical records, remote medical systems, and the Internet of Things, the limitations imposed by resources and location can be reduced, regional disparities can be narrowed, and better disease prevention and resource use can be achieved, thus relieving resource burdens. It also provides an effective platform for comprehensive connectivity and real-time interaction of medical resources, allowing for the rapid and automatic adjustment of rehabilitation strategies and the reconfiguration of medical resources based on specific patient needs [19,20].

Existing research on broadband infrastructure development mainly focuses on 3 aspects. The first is the impact of broadband infrastructure development on innovation. Broadband infrastructure will play an important role in the new round of technological and industrial revolutions by enhancing the level of information technology, promoting digital development, and alleviating resource mismatches to improve innovation capabilities [21,22]. The second is the impact of broadband infrastructure development on the environment. Broadband infrastructure development greatly improves greenhouse gas emission performance, showing significant emission reduction effects, and broadband connectivity promotes the efficiency of green innovation. It can also bring about network spillover effects, which can promote the diffusion and progress of green technologies and further influence green technology innovation [23-25]. The third is the impact of broadband infrastructure development on resource allocation. Broadband infrastructure development can improve resource mismatches, increase resource mobility, thereby reducing the cost of resource allocation, directly or indirectly improving the efficiency of resource allocation, addressing resource mismatches, and addressing spatial spillover effects [26-28]. On the basis of quasi-natural experiments using the BCP, it was found that the BCP has a significant positive impact on resource allocation efficiency. Accelerating the construction and application of broadband infrastructure can optimize decision-making processes and improve resource allocation efficiency. Simultaneously, BCP can enhance the overall level of information infrastructure, reduce resource mismatches, optimize resource allocation, and promote coordinated regional development [29,30].

Several factors contribute to MMs. First, economic development significantly affects resource allocation. Economically weak regions often lack sufficient medical resources, and improvements in income levels have a direct impact on resource availability. This leads to a loss of resources in underdeveloped areas, exacerbating the mismatch. In addition, corruption negatively influences resource allocation [31]. Second, population plays a key role in resource allocation. In remote and sparsely populated areas, the dispersed distribution of the population increases the likelihood of resource mismatches [32]. Moreover, allocating resources solely based on per capita or per area indicators can result in deviations due to higher medical demand among the elderly and children. Finally, factors such as the urbanization rate, financial investment, and per capita gross domestic product (GDP) also influence resource allocation, potentially leading to mismatches [33,34]. In addition, there are 4 unique reasons for MMs in China. These include low physician income and distorted salary systems, an imperfect medical service market hindering resource flow, varying degrees of scarcity of medical resources, and the impact of China's medical system on resource flow and pricing [35].

In addition, this study offers 3 key contributions. First, it establishes a theoretical framework for the correlation between broadband infrastructure construction and mismatched medical resources, exploring their basic relationship, heterogeneity, and potential mechanisms. This enriches the literature on broadband internet's impact on public services. Second, it presents empirical evidence from middle- and low-income nations, using authoritative data and rigorous causal strategies to assess the effects of broadband infrastructure on MMs. Finally, drawing from theoretical and empirical analyses, it furnishes theoretical guidance and decision-making support for government bodies to enhance broadband development and resource allocation optimization in the medical sector.

Theoretical Hypothesis

Basic Hypothesis

In the era of IT, broadband is vital to modern society, particularly for BCP's strategic positioning. By enhancing broadband access and mobile communication, BCP reduces information transmission time, lowers information asymmetry, and boosts public service efficiency [36,37]. Internet health care proved invaluable during the COVID-19 crisis, aiding epidemic control, reshaping physician-patient interactions, and facilitating resource use [38-40]. Information infrastructure development in medical systems alters service patterns, enhancing efficiency [41,42]. The internet empowers patients, tackles inequality, and controls medical costs. In this context, a hypothesis can be proposed:

- Hypothesis 1: BCP can reduce mismatches in the allocation of medical resources.

Heterogeneity Hypothesis

China's current resource allocation model favors the supply side, aggravating regional resource imbalances. Specifically, in health care, most top hospitals are concentrated in high-income regions, concentrating on high-quality medical services [43,44]. Tertiary hospitals, comprising 4% of all hospitals, employ >26% of medical staff and deliver more than 20% of medical services [45]. Eastern coastal regions have abundant medical resources,

XSL•FO
RenderX

whereas western areas lack resources, infrastructure, and education [46,47]. Although the BCP has improved nationwide broadband coverage, regional discrepancies persist. Broadband progress is notable in central and eastern regions, narrowing the gap with the West, which still lags behind in key areas [48,49]. From this, we propose the following hypothesis:

- Hypothesis 2: there are differences in the effects of the BCP between different regions, with the central and eastern regions showing greater effects than the western region.

Medical resources in China tend to concentrate in densely populated and economically developed top cities, especially in western regions [46]. The BCP has had a more significant impact on improving medical resource allocation and service quality in top cities than in ordinary prefecture-level cities, with top cities experiencing a 4.2% and 4.8% increase in human and material resources, respectively [50]. Nearly three-quarters of registered internet hospitals in China are concentrated in the top 11 large cities, highlighting their advantages in obtaining medical resources and providing services [51]. Implementing the BCP in top cities can improve local residents' health status and promote health care services' accessibility, generating spillover effects on surrounding areas, and promoting public health progress at a macrolevel. Thus, we propose the following hypothesis:

- Hypothesis 3: there are differences in the effects of the BCP between different population sizes, with the effect being greater in cities with larger populations.

In recent decades, global youth populations have faced challenges accessing medical resources and have been neglected in health and social policies. However, the rapid development of digital media and broadband technology presents new opportunities for this group, especially in urban areas with a higher concentration of adolescents [52]. Adolescents are highly active on the internet, with up to 95% using it frequently and having access to computers and tablets at home. This makes the internet an effective platform for adolescent self-care, education, and health-related services [53-55]. In addition, adolescents' susceptibility to social problems and their increased demand for medical resources and health information further highlight the significance of digital medical platforms and internet health resources in cities with a high proportion of youth populations [56]. Therefore, implementing network medical strategies and resource layouts (BCP) targeted at adolescents in these cities is crucial and more likely to be effective. On the basis of this, we propose the following hypothesis:

- Hypothesis 4: there are differences in the effectiveness of BCP between different population structures, with a greater effect in cities with a higher population of adolescents.

Methods

Data Sources

The urban-level data in this study mainly come from the National Bureau of Statistics of China, annual statistical yearbooks of various provinces over the years, annual statistical yearbooks of various cities over the years, annual health statistics yearbooks of China, and reports on the BCP released by the government. Considering data availability and excluding possible impacts from exogenous policy shocks, such as the 2008 global financial crisis and the largest-scale enterprise income tax reform in China, this study ultimately obtained a balanced panel data set covering 300 cities spanning the years 2010 to 2021, with 3600 observations.

Econometric Equation

In reference to the existing literature [57,58], we constructed the following 2-way fixed effects difference-in-differences (DID) model:

In this equation, MM_{jt} represents the degree of MM for city j in year t. BCP_{jt} represents the dummy variable for the BCP. $X`\theta$ represents a series of city-level control variables to control for other factors' influence; μ_j represents city fixed effects to control for unobservable factors that vary across cities; δ_t represents year fixed effects to control for unobservable factors that vary over time; ε_{jt} represents the error term. β_1 is the coefficient of interest in this study, and we expect it to be significantly negative. This indicates that the BCP significantly reduces the degree of MM in cities.

Variable Definition

Dependent Variable

The degree of MM in cities is the dependent variable in this study. A previous study defined resource mismatch at the enterprise level [59], stating that optimal resource allocation results in consistent marginal output across different enterprises, thus maximizing total output and efficiency. Conversely, resource mismatch distorts the input-output allocation ratio among different enterprises, leading to a loss in the total output and efficiency. Following their approach, we calculated the degree of MM as follows:

In this equation, τ_{kjt} represents the degree of capital mismatch in city j in year t, where a higher value indicates a greater distortion in the marginal output of capital. τ_{ljt} represents the degree of labor mismatch in city j in year t, where a higher value indicates a greater distortion in the marginal output of labor; α_j and β_j represent the output elasticities of medical capital and medical labor in cities, respectively, with $\alpha_j+\beta_j\neq1$ [60]; σ represents the elasticity of substitution between outputs of different cities, and this value generally falls between 3 and 10 [61]. To obtain a conservative estimate, in reference to the existing studies [62], we set $\sigma=3$; R represents the cost of capital use. In reference to the existing studies [59], we set R as 0.1 in this study; $P_{jt}Y_{jt}$ represents city medical expenditure, measured by health expenditure; ωL_{jt} represents the income of medical labor in city j, ω represents the average income in the medical industry within city j, and L represents the number of medical personnel in the city; and K_{jt} represents the amount of medical capital in the city, measured by the total number of hospitals

XSL•FO
RenderX

and beds in the city [8]. To facilitate interpretation of the results, we take the logarithm of the calculated results and obtain the dependent variable in this study, the degree of MM in cities.

Independent Variable

The core independent variable in this study is the BCP dummy variable. When a city is selected as a pilot city for the BCP in a particular year, the variable is assigned a value of 1 for that year and all subsequent years, whereas it is assigned a value of 0 for other cities. This variable (BCP) serves as the core independent variable in this study.

Control Variables

To minimize the potential bias caused by omitted variables, we selected 4 variables—economic level, government intervention, population structure, and urbanization rate—as control variables based on existing studies [26,63,64] and the actual situation.

The first is the economic level (Gdp). Because broadband infrastructure construction is positively correlated with regional economic levels, it is necessary to control for the economic level to identify the impact of broadband infrastructure construction on resource misallocation [65]. This study measures the economic level using *per capita GDP*.

The second is government intervention (Gov). Government intervention is an important factor leading to resource misallocation because the government may suppress the effectiveness of market resource allocation, resulting in market failure. The higher the proportion of fiscal expenditure to regional GDP, the stronger the government's intervention capacity and thus the higher the level of resource misallocation [66]. This study measures government intervention using the *proportion of government fiscal expenditure to regional GDP*.

The third is the population structure (Pop). The population structure has a significant impact on labor force allocation. The *population dividend* optimizes resource allocation efficiency through the flow of labor, and the unlimited supply of labor offsets the diminishing marginal returns of capital [67]. This study measures population structure using the *average number of employed persons per year divided by the average total population*.

The fourth is the urbanization rate (Urb). The agglomeration effect generated during the process of urbanization can promote the concentration of labor and capital within a region, expand the economic scope and labor market of urban centers and their peripheries, and improve the matching degree between capital, labor, and hospitals, thereby facilitating the flow of resources within the region, enhancing resource allocation efficiency, and reducing resource misallocation [66]. This study measures the urbanization rate using the *proportion of the nonagricultural population to the total population*. The major variables and their definitions are summarized in Table S1 in Multimedia Appendix 1 [59].

Ethical Considerations

According to China's Trial Guideline on the Review of Science and Technology Ethics, this study does not collect or use any personal identification information or sensitive data that may harm or disclose personal privacy. The study only focuses on nonhuman entities (cities) and does not involve any direct interaction with individuals or observation of individuals. In addition, this study does not involve any procedures or activities that may raise ethical concerns and complies with all applicable laws and regulations regarding research conduct. Therefore, no ethical approval is required.

Results

Descriptive Statistics

Table 1 presents the descriptive statistics for the main variables in this study. The mean of the MM is 1.008, with an SD of 0.360, indicating a significant variation in the misallocation of medical resources among different cities.

Table 1. Descriptive statistics (N=3600).

Variables	Values, mean (SD; range)
MM[a]	1.008 (0.360; 0.008-1.520)
BCP[b]	0.212 (0.409; 0.000-1.000)
Gdp[c]	2.481 (1.454; –0.157-5.142)
Gov[d]	0.556 (0.182; 0.218-0.892)
Pop[e]	0.264 (0.120; 0.044-0.472)
Urb[f]	0.452 (0.203; 0.082-0.831)

[a]MM: medical resource mismatch.
[b]BCP: Broadband China Policy.
[c]Gdp: economic level.
[d]Gov: government intervention.
[e]Pop: population structure.
[f]Urb: urbanization rate.

The mean of the BCP is 0.212 (SD 0.409), indicating that 21.22% (764/3600) of the observations in the sample period

XSL•FO
RenderX

were classified as the treatment group. The descriptive statistics of the other control variables are generally consistent with those of existing related studies and will not be reiterated here.

Basic Regression

Benchmark Regression Results

To test hypothesis 1, we conduct an empirical test based on equation 1. To demonstrate the effects of different factors, a stepwise regression is used. The specific regression results are presented in Table 2.

Table 2. Benchmark regression.

Variables	MM[a] (column 1)	MM (column 2)	MM (column 3)	MM (column 4)	MM (column 5)
BCP[b]					
Coefficient (SE)	–0.134 (0.0145)	–0.133 (0.0144)	–0.133 (0.0164)	–0.132 (0.0240)	–0.132 (0.0239)
P value	<.001	<.001	<.001	<.001	<.001
Gdp[c]					
Coefficient (SE)	N/A[d]	–0.00919 (0.00406)	–0.00931 (0.00407)	–0.00862 (0.00422)	–0.00862 (0.00402)
P value	N/A	.02	.02	.04	.03
Pop[e]					
Coefficient (SE)	N/A	–0.134 (0.0324)	–0.133 (0.0324)	–0.120 (0.0336)	–0.120 (0.0339)
P value	N/A	<.001	<.001	<.001	<.001
Gov[f]					
Coefficient (SE)	N/A	0.132 (0.0491)	0.130 (0.0491)	0.133 (0.0506)	0.133 (0.0498)
P value	N/A	.007	.008	.008	.008
Urb[g]					
Coefficient (SE)	N/A	–0.102 (0.0290)	–0.0976 (0.0290)	–0.0792 (0.0302)	–0.0792 (0.0287)
P value	N/A	<.001	.001	.009	.006
City fixed effect[h]	No	No	No	Yes	Yes
Year fixed effect[i]	No	No	Yes	Yes	Yes
Control[j]	No	Yes	Yes	Yes	Yes
Robust[k]	No	No	No	No	Yes
Observations, N	3600	3600	3600	3600	3600
Values, R^2	0.023	0.035	0.024	0.024	0.024

[a]MM: medical resource mismatch.
[b]BCP: Broadband China Policy.
[c]Gdp: economic level.
[d]N/A: not applicable.
[e]Pop: population structure.
[f]Gov: government intervention.
[g]Urb: urbanization rate.
[h]City fixed effect controls for unobservable factors that vary across cities.
[i]Year fixed effect controls for unobservable factors that vary over time.
[j]Control refers to a set of control variables as mentioned in the previous text.
[k]Robust indicates the use of robust SEs to account for heteroscedasticity.

Column 1 shows the ordinary least squares regression results without any control. The regression coefficient of the BCP on MM is –0.134, which is significant at the 1% level, indicating that the implementation of BCP has reduced the MM by 13.4%. However, it should be noted that this only reflects a correlation. Column 2 adds control variables to control for the impact of other observable factors. The implementation of the BCP has reduced the MM by 13.3%. Column 3 further adds year fixed effects to control for the influence of unobservable factors in different observation years. The result remains unchanged, with

XSL•FO
RenderX

only a slight increase in the SE. Column 4 further adds city fixed effects to control for the influence of unobservable factors in the different cities. The implementation of BCP has reduced the MM by 13.2%. Column 5 further uses robust SEs to control for the heteroscedasticity of the error term. The result is the same as before, with the implementation of BCP reducing the MM by 13.2%.

Due to the more complete control of the confounding effects of other factors, the baseline regression result of this study is based on column 5, and hypothesis 1 is clearly supported. The results were similar to those of existing studies [68], which studied the impact of the BCP on urban labor allocation and obtained similar results.

Looking at the control variables in column 5, Gdp reduces the MM by 0.86%, Pop reduces it by 12%, Gov increases it by 13.3%, and Urb reduces it by 7.92%, which is consistent with the previous expectations.

Dynamic Test

To conduct a DID regression, the parallel trends assumption must be satisfied between the control group and the experimental group. Otherwise, the regression results may be driven by preexisting differences between the groups. Drawing on the event study methodology commonly used in the field of finance, the following dynamic regression equation is constructed for testing:

In this equation, *event* represents the event period. In the period when the policy is implemented, event is set to 0. In the period before the policy implementation, event is set to –1. In the period after the policy implementation, event is set to +1, and so on for subsequent periods. Following the practice in mainstream literature, event=–1 is chosen as the base period. The interpretation of other variables remains the same and will not be repeated. The specific regression results are shown in Figure 1.

Figure 1. Dynamic test.

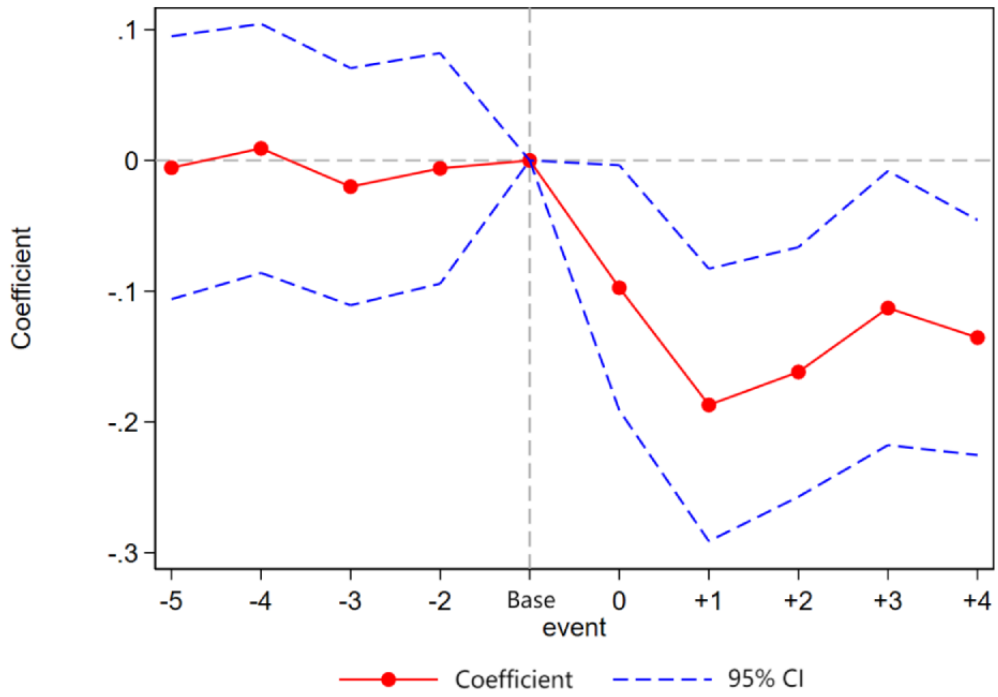

Before the policy implementation (ie, when event ≤–1), the coefficient of the BCP is not statistically significant, indicating that there are no significant preexisting differences between the pilot areas of the BCP and nonpilot areas, thus satisfying the parallel trends assumption. In the periods after the policy shock, the coefficient of BCP is significantly negative and decreases over time before stabilizing. Overall, the BCP has significantly reduced the degree of MM each year.

Placebo Test

The baseline regression results of this study may be due to chance grouping rather than a true reflection of the policy effect of BCP. To eliminate the interference of chance factors, we conducted a placebo test, arbitrarily designating a city as a false treatment group and then regressing to obtain a false regression coefficient. This operation was repeated 500 times, and the density of the 500 spurious regression coefficients obtained is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Placebo test.

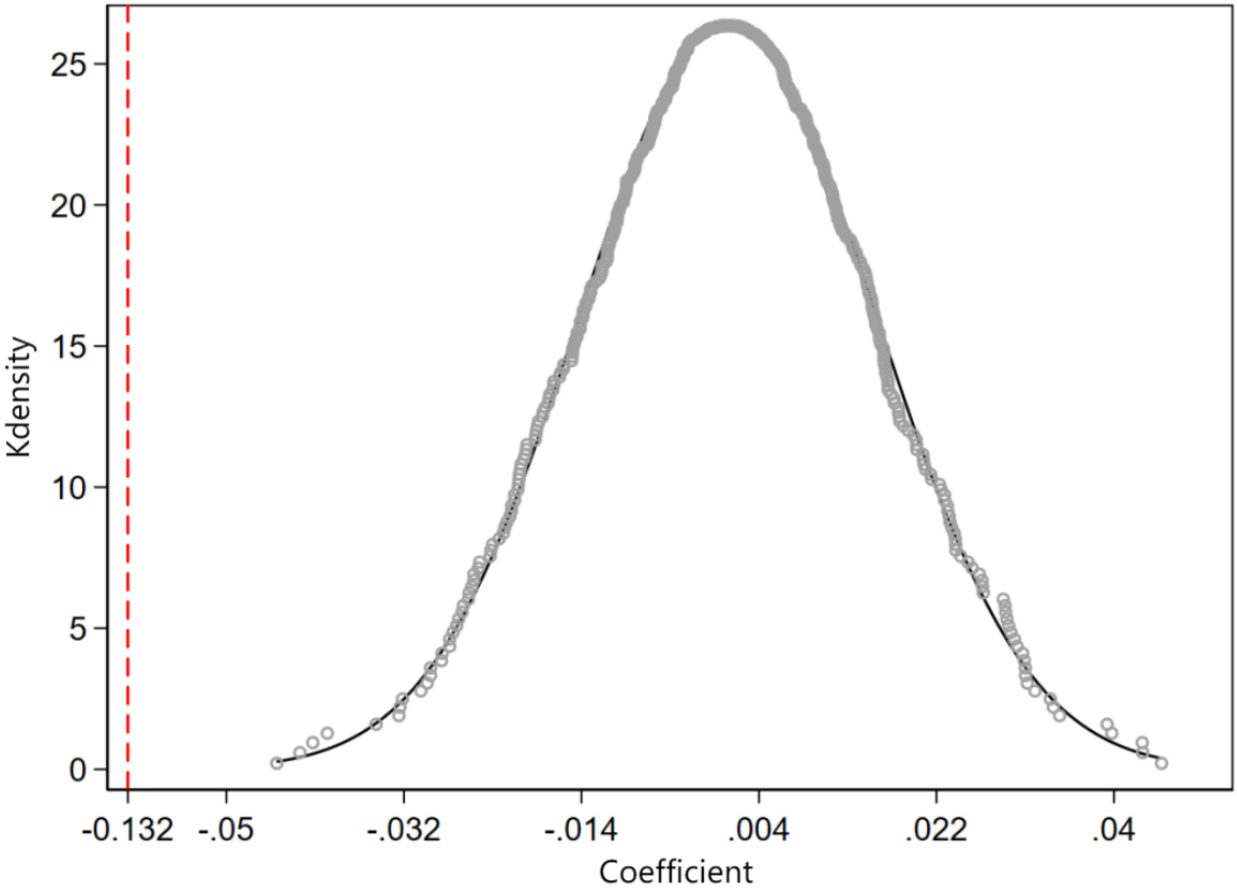

As shown in the figure, the 500 spurious regression coefficients are mostly concentrated around 0 and are significantly different from the real regression coefficient (red dashed line), indicating that the baseline results of this study are not due to chance factors.

Robustness Analysis

Overview

To ensure the robustness of the benchmark regression results, a series of robustness tests were conducted in this section, and the specific results are presented in Table 3.

XSL•FO
RenderX

Table 3. Robust regression results.

Variables	MM[a]_solow (column 1)	PSM-DID[b] (column 2)	IV[c]1 (column 3)	IV2 (column 4)
BCP[d]				
Coefficient (SE)	–0.218 (0.0737)	–0.123 (0.0208)	–0.183 (0.0395)	–0.119 (0.0273)
P value	.003	<.001	<.001	<.001
Kleibergen-Paap[e]	N/A[f]	N/A	32.748	23.946
Cragg-Donald[g]	N/A	N/A	33.661	23.809
City fixed effect[h]	Yes	Yes	Yes	Yes
Year fixed effect[i]	Yes	Yes	Yes	Yes
Control[j]	Yes	Yes	Yes	Yes
Robust[k]	Yes	Yes	Yes	Yes
Observations, N	3600	3600	3600	3600
Values, R^{2}[l]	0.010	0.023	N/A	N/A

[a]MM: medical resource mismatch.

[b]PSM-DID: propensity score matching difference-in-differences.

[c]IV: instrumental variable.

[d]BCP: Broadband China Policy.

[e]Kleibergen-Paap represents the Kleibergen-Paap rk LM (Lagrange Multiplier) statistic, which performs the underidentification test. The null hypothesis is that the core explanatory variable is not correlated with the IV. The results in the table clearly reject the null hypothesis.

[f]N/A: not applicable.

[g]Cragg-Donald represents the Cragg-Donald Wald F statistic, which performs the weak identification test. The null hypothesis is that the core explanatory variable is weakly correlated with the IV. The results in the table clearly exceed the eigenvalue at the 10% significance level suggested by Stock-Yogo.

[h]City fixed effect controls for unobservable factors that vary across cities.

[i]Year fixed effect controls for unobservable factors that vary over time.

[j]Control refers to a set of control variables as mentioned in the previous text.

[k]Robust indicates the use of robust SEs to account for heteroscedasticity.

[l]As the R^2 in the 2-stage IV method does not reflect the true goodness of fit, it is not reported here

Replacement of the Dependent Variable

In the previous calculation of the dependent variable MM, the elasticity values of medical labor and medical capital were measured using the Olley-Pakes method. In this section, following the existing studies [69], the Solow residual method was used for recalculation. Simultaneously, the substitution elasticity of urban output was replaced with an upper limit value of 10. The newly obtained dependent variable MM_solow was used, and the specific regression results are shown in column 1: the impact coefficient of the BCP on MM_solow is –0.218, which is significant at the 1% level, indicating that the implementation of BCP has reduced the degree of mismatch of medical resources by 21.8%, and the results remain robust.

Propensity Score Matching–DID Method

Considering that the pilot cities for the BCP may not have been arbitrarily selected and that there may be inherent differences between pilot and nonpilot cities, it is precisely these differences that drive the benchmark regression results. To address this concern, this study uses the propensity score matching–DID method to match pilot and nonpilot areas in a 1:1 ratio based on propensity score matching and then conducts a reregression. The results are shown in column 2: the impact coefficient of the BCP on MM is –0.123, which is significant at the 1% level, indicating that the implementation of the BCP has reduced the degree of MM by 12.3%, and the results remain robust. It should be noted that the premise of using the propensity score matching–DID method is that there are no significant differences in covariates between pilot and nonpilot areas. We conducted a covariate balance test; the specific regression results are presented in Table S2 in Multimedia Appendix 1.

Instrumental Variable Method

Considering the potential endogeneity issue between the BCP and the degree of MM, it is possible that cities with a lower degree of MM are also cities with higher levels of IT and economic development, making them more likely to be selected as pilot areas for the BCP. In addition, this study may also suffer from omitted variable bias and measurement error issues that could lead to endogeneity. Although the DID approach used in this study partially mitigated the endogeneity problem, to enhance result reliability, we further implemented an instrumental variable (IV) approach to address potential endogeneity issues, following the methodology of a previous study [68].

In reference to existing studies [68], we used the topographic fluctuation of cities as an instrument. On the one hand, cities

XSL•FO
RenderX

with larger topographic fluctuation not only increase the cost of broadband construction but also lead to a decrease in the signal quality of broadband networks. Therefore, cities with larger topographic fluctuations face greater difficulties in broadband construction, making them less likely to become pilot cities for the *Broadband China* initiative. On the other hand, topographic fluctuation is a geographical condition with strong exogeneity and does not affect labor allocation distortion. For ease of understanding, we took the reciprocal of topographic fluctuation as IV1. The specific regression results are presented in column 3: the impact coefficient of BCP on MM is –0.133, which is significant at the 1% level, indicating that the implementation of BCP has reduced the degree of MM by 18.3%, and the results remain robust. Furthermore, the results pass the overidentification test (Kleibergen-Paap) and weak identification test (Cragg-Donald).

In reference to existing studies [70], we used the distance from each city to Hangzhou as an instrument. On the one hand, in 2004, the Alibaba Group launched the Alipay service in Hangzhou, which promoted the development of digital finance. Therefore, cities closer to Hangzhou are more likely to have the foundation for digital infrastructure construction and application, making them more likely to be selected as pilot areas for the BCP. On the other hand, this IV is a historical distance variable, satisfying the exogeneity assumption [71,72]. Thus, we obtain IV2. The specific regression results are presented in column 4: the impact coefficient of BCP on MM is –0.119, which is significant at the 1% level, indicating that the implementation of BCP has reduced the degree of MM by 181.9%, and the results remain robust. Furthermore, the results pass the overidentification test (Kleibergen-Paap) and weak identification test (Cragg-Donald).

Mechanism Test

To determine which channels the BCP reduces the MM, this study explores and identifies 2 potential channels: the innovation effect and the integration effect. The mechanism test is conducted using a 3-step method, and the specific regression results are presented in Table 4.

Table 4. Mechanism test.

Variables	Patent (column 1)	MM[a] (column 2)	Integration (column 3)	MM (column 4)
BCP[b]				
Coefficient (SE)	0.225 (0.0807)	–0.127 (0.0238)	0.0786 (0.0365)	–0.126 (0.0236)
P value	.006	<.001	.03	<.001
Patent				
Coefficient (SE)	N/A[c]	–0.0249 (0.0051)	N/A	N/A
P value	N/A	<.001	N/A	N/A
Integration				
Coefficient (SE)	N/A	N/A	N/A	–0.0764 (0.0130)
P value	N/A	N/A	N/A	<.001
City fixed effect[d]	Yes	Yes	Yes	Yes
Year fixed effect[e]	Yes	Yes	Yes	Yes
Control[f]	Yes	Yes	Yes	Yes
Robust[g]	Yes	Yes	Yes	Yes
Observations, N	3600	3600	3600	3600
Values, R^2	0.006	0.030	0.007	0.034

[a]MM: medical resource mismatch.
[b]BCP: Broadband China Policy.
[c]N/A: not applicable.
[d]City fixed effect controls for unobservable factors that vary across cities.
[e]Year fixed effect controls for unobservable factors that vary over time.
[f]Control refers to a set of control variables as mentioned in the previous text.
[g]Robust indicates the use of robust SEs to account for heteroscedasticity.

Innovation Effect

Research has shown that enhancing broadband infrastructure can elevate urban innovation levels, with a reported increase of 1.94% in innovation for every unit of internet access [73]. It is evident that broadband infrastructure construction plays a crucial role in advancing China's regional innovation level by directly improving innovation levels and indirectly impacting innovation through the acceleration of human capital accumulation, financial development, and industrial upgrading [74]. The expansion of broadband networks not only broadens the spatial

XSL•FO
RenderX

range of network space functions and innovation resource allocation but also offers more opportunities and development space for innovation activities and the evolution of innovation models [75]. Furthermore, there exists a substantial correlation between innovation efficiency and resource allocation efficiency, with higher innovation efficiency leading to enhanced resource allocation efficiency [76]. Hence, it is anticipated that the BCP can alleviate resource mismatches by fostering regional innovation.

In reference to the existing studies [77], this study measures urban innovation levels based on the number of invention patents. Patent numbers are divided into 2 categories: patent applications and patent authorizations. Considering that policy shocks will have an impact on urban innovation levels in the current period and that there is a certain time lag from patent application to authorization, this study uses the number of invention patent applications in each city (Patent) to measure urban innovation levels. The number of city patent applications comes from the China Patent Public Announcement System on the website of the China National Intellectual Property Administration. The regression results are shown in columns 1 and 2: the coefficient of the BCP's impact on Patent is 0.225, and it is significant at the 1% level, indicating that the implementation of the BCP significantly improves urban innovation levels. The coefficient of Patent's impact on MM is –0.0249, and it is significant at the 1% level, indicating that the significant improvement in urban innovation levels reduces the mismatch of medical resources, confirming the theoretical expectations.

Integration Effect

Market segmentation restricts the free flow of resources. The development of broadband internet can promote market integration and improve resource mismatches [78]. Improving information technology infrastructure construction can achieve resource integration, enhance resource management levels, and improve production efficiency [79]. The internet accelerates production transformation and market integration. It optimizes resource flow between the upstream and downstream through information mechanisms, gradually reducing resource waste and promoting resource growth. In general, the BCP can improve resource mismatches in target areas through market integration [80,81].

Following the methodology of a previous study [82], this study uses *(number of regional hospitals/regional medical expenditure)/(number of national hospitals/national medical expenditure*) to measure regional integration (Integration). The regression results are shown in columns 3 and 4: the coefficient of BCP's impact on Integration is 0.0786, which is significant at the 5% level. This indicates that the implementation of BCP significantly improves urban integration levels. The coefficient of Integration's impact on MM is –0.0764, which is significant at the 1% level. This suggests that the significant improvement in integration levels significantly reduces the MM, confirming the theoretical expectations.

Heterogeneity Analysis

Overview

This section tests the heterogeneous hypotheses proposed earlier and conducts an empirical *P* value test for intergroup coefficient differences. The specific regression results are presented in Table 5.

XSL•FO
RenderX

Table 5. Heterogeneity results.

Variables	MM[a]		MM		MM	
	Column 1: central and eastern region	Column 2: western region	Column 3: high population size	Column 4: low population size	Column 5: high youth proportion	Column 6: low youth proportion
BCP[b]						
Coefficient (SE)	–0.159 (0.0292)	–0.0664 (0.0415)	–0.149 (0.0321)	–0.118 (0.0332)	–0.144 (0.0367)	–0.130 (0.0368)
P value	<.001	.11	<.001	<.001	<.001	<.001
Empirical *P* value[c]	.002	.002	.008	.008	.016	.016
City fixed effect[d]	No	No	No	Yes	Yes	Yes
Year fixed effect[e]	No	No	Yes	Yes	Yes	Yes
Control[f]	No	Yes	Yes	Yes	Yes	Yes
Robust[g]	No	No	No	No	Yes	Yes
Observations, N	2448	1152	1800	1800	1800	1800
Values, R^2	0.030	0.026	0.037	0.021	0.038	0.024

[a]MM: medical resource mismatch.

[b]BCP: Broadband China Policy.

[c]Empirical *P* value, testing for differences in coefficients between groups, was achieved using a 500-times self-help sampling method.

[d]City fixed effect controls for unobservable factors that vary across cities.

[e]Year fixed effect controls for unobservable factors that vary over time.

[f]Control refers to a set of control variables as mentioned in the previous text.

[g]Robust indicates the use of robust SEs to account for heteroscedasticity.

Regional Differences

Hypothesis 2 expects that there are differences in the effects of BCP between different regions, with a larger effect in the central and eastern regions compared with the western region. According to the regional classification standards of the National Bureau of Statistics of China, we classified Sichuan, Yunnan, Guizhou, Tibet, Chongqing, Shaanxi, Gansu, Qinghai, Xinjiang, Ningxia, Inner Mongolia, and Guangxi as the western region, while the rest are classified as the central and eastern regions. The group regression results are shown in columns 1 and 2: the BCP reduces MM by 15.9% in the central and eastern regions, but it does not have a significant impact in the western region. The empirical *P* value indicates a significant difference in the coefficients between the 2 groups. This confirms the expectation of hypothesis 2.

Population Size Differences

Hypothesis 3 expects that there are differences in the effects of BCP among different population sizes, with larger cities experiencing a greater effect. On the basis of the median total population at the end of each city, the sample is divided into 2 subsamples: a high population size and a low population size. The group regression results are shown in columns 3 and 4: the BCP reduces medical mismatch by 14.9% in the high population size sample, while it reduces it by 11.8% in the low population size sample. The empirical *P* value indicates a significant difference in the coefficients between the 2 groups. This confirms the expectation of hypothesis 3.

Population Structure Differences

Hypothesis 4 expects that there are differences in the effects of BCP among different population structures, with cities having a higher proportion of the young population experiencing a greater effect. Due to the lack of data on the proportion of the young population in the China Urban Statistical Yearbook, considering that the ratio of the young population receiving primary and secondary education is >90% and the ratio of those receiving higher education is close to 60% [83], we use the number of students in ordinary primary, secondary, and higher education institutions as a proxy for the young population. We divide the sample into 2 groups based on the median proportion of these students to the total population: high youth proportion and low youth proportion. The group regression results are shown in columns 5 and 6: the BCP reduces medical mismatch by 14.4% in the high youth proportion sample, while it reduces it by 13% in the low youth proportion sample. The empirical *P* value indicates a significant difference in the coefficients between the 2 groups. This confirms the expectation of hypothesis 4.

Discussion

Principal Findings

The theoretical analysis and empirical results of this study demonstrate that the construction of broadband infrastructure significantly reduces the MM by 13.2%. The potential realization channels are innovation effects and integration effects, and this impact is more pronounced in the central and

XSL•FO
RenderX

eastern regions, cities with large populations, and cities with a high proportion of young people.

Limitations

This study still has some limitations. For example, how to establish a mathematical model to accurately depict the quantitative relationship between the construction of broadband infrastructure and the MM from a mathematical perspective and incorporate mechanism analysis and heterogeneous analysis into this model, ultimately forming a systematic theoretical analysis framework, is worth exploring in the future.

Conclusions

This study proposes 4 policy recommendations. First, there is a need to promote the implementation of BCP more extensively. The benchmark regression results demonstrate that the BCP significantly reduces the mismatch of health care resources. This includes the expansion of BCP to develop information technology in more cities, enhancing the integrated use of broadband infrastructure in the health care service industry, and a comprehensive transformation of the health care service industry in various aspects. Furthermore, the digitization of health care construction and the development of information-based health services such as electronic health record systems, telemedicine services, and mobile health apps are imperative. Enhancing broadband infrastructure construction and using it as a precursor to improving the mechanism of health care resource allocation is crucial. This will strengthen broadband internet access in the health care industry, promote efficient resource exchange among health care institutions, optimize resource allocation using technologies such as big data and artificial intelligence, and reduce the mismatch of health care resources [84].

Second, there is a need to focus on the level of technological innovation and resource integration capacity. The mechanism test indicates that the BCP reduces the MM by enhancing the level of innovation and resource integration capacity. This entails accelerating the deployment of emerging broadband infrastructure, focusing on the coordination between broadband infrastructure and innovation policy and capacity, establishing interdisciplinary innovation centers, and promoting collaborative innovation among the health care, IT, and communication fields. Moreover, broadening and deepening the integration of broadband infrastructure and medical resources to guide better integration and allocation of interregional medical resources is essential. This can be achieved through the establishment of a unified resource management and scheduling platform [85].

Third, the prioritization of the BCP implementation in the central and eastern regions, cities with large populations, or regions with a high proportion of young people is recommended. The results of the heterogeneous regression indicate that the impact of BCP is more significant in these cities. This includes leveraging the leading role of key cities, reinforcing their role as leaders in health care information and broadband infrastructure, and capitalizing on their demonstration effect and radiation. Simultaneously, under the current conditions, it is crucial to introduce social capital, including demand, technology, and funds, to assist cities with relatively backward broadband infrastructure in overcoming their disadvantages and catching up with advanced cities [86].

Finally, macrocontrol should be strengthened to encourage factor mobility. Owing to differences in economic levels and support, there is a substantial gap in the allocation of medical resources in China, primarily due to the lack of medical factor flow. It is imperative to strengthen the macrocontrol and coordination capacity, formulate overall medical resource planning, clarify the objectives and development direction of medical resource allocation in various regions, improve the resource allocation system, and enhance the macrocontrol of the medical and health industries. Encouraging the flow of medical resource elements and increasing policy flexibility will help adapt to rapid changes in medical needs and technological development trends and better satisfy the medical needs of different levels [87].

Acknowledgments

This study was supported by the Doctoral Research Innovation Program of Weifang Medical University (grant 041174). The authors are thankful to the editors and reviewers for their comments and feedback.

Authors' Contributions

YC and LG designed and wrote the manuscript. XY and LG collected and analyzed the data. YC, XY, and LG designed and implemented the study. LG and ZC read and critically revised the manuscript and contributed to the study concept. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables.
[DOCX File , 14 KB - jmir_v26i1e53921_app1.docx]

References

XSL•FO
RenderX

1. Ran Z, Gao J. Spatial evolution and influencing factors of medical and health resources in big cities: a case study of central urban area of Zhengzhou City. Areal Res Dev 2023;42:53-59 In Chinese. [doi: 10.3969/j.issn.1003-2363.2023.02.009]
2. Ding J, Hu X, Zhang X, Shang L, Yu M, Chen H. Equity and efficiency of medical service systems at the provincial level of China's mainland: a comparative study from 2009 to 2014. BMC Public Health 2018;18(1):214 [FREE Full text] [doi: 10.1186/s12889-018-5084-7] [Medline: 29402260]
3. Pang R, Li S. Structural imbalance of medical resource allocation and "Kanbing Gui" in China: from the perspective of referral system. Mod Econ Sci 2022;44(3):97-110 In Chinese [FREE Full text]
4. Jin J, Wang J, Ma X, Wang Y, Li R. Equality of medical health resource allocation in China based on the Gini Coefficient Method. Iran J Public Health 2015;44(4):445-457 [FREE Full text] [Medline: 26056663]
5. Pu L. Fairness of the distribution of public medical and health resources. Front Public Health 2021;9:768728 [FREE Full text] [doi: 10.3389/fpubh.2021.768728] [Medline: 34858935]
6. Fu R, Zhang X, Yang D. Spatial equity of healthcare resource in Xinjiang. J Univ Chin Acad Sci 2023;40(1):69-78 In Chinese [FREE Full text] [doi: 10.7523/j.ucas.2021.0001]
7. Xue C, Meng Q, He X. Network infrastructure and the diffusion of technological knowledgevidence from a quasi-natural experiment. J Finance Econ 2020;46(4):48-62 In Chinese [FREE Full text] [doi: 10.16538/j.cnki.jfe.2020.04.004]
8. Liu C, Ma Q. Research on the influence of network infrastructure construction on total factor productivity growth: a quasi-natural experiment of "Broadband China" pilot policy. Chin J Popul Sci 2020(03):75-88 [FREE Full text]
9. Wu H, Ba N, Ren S, Xu L, Chai J, Irfan M, et al. The impact of internet development on the health of Chinese residents: transmission mechanisms and empirical tests. Socio-Econ Plan Sci 2022;81:101178. [doi: 10.1016/j.seps.2021.101178]
10. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health 2015;129(6):611-620. [doi: 10.1016/j.puhe.2015.04.001] [Medline: 26025176]
11. Hirko KA, Kerver JM, Ford S, Szafranski C, Beckett J, Kitchen C, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020;27(11):1816-1818 [FREE Full text] [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
12. Lv Q, Jiang Y, Qi J, Zhang Y, Zhang X, Fang L, et al. Using mobile apps for health management: a new health care mode in China. JMIR Mhealth Uhealth 2019;7(6):e10299 [FREE Full text] [doi: 10.2196/10299] [Medline: 31162131]
13. Circular of the state council on the issuance of the “broadband China” strategy and implementation programme. The Central People's Government of the People's Republic of China. URL: https://tinyurl.com/bdhwckud [accessed 2023-09-16]
14. Guiding opinions of the general office of the state council on accelerating the construction of high-speed broadband networks and promoting network speed-up and fee reduction. The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/zhengce/zhengceku/2015-05/20/content_9789.htm [accessed 2023-09-16]
15. Two departments to create “broadband China” demonstration cities. The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/gzdt/2014-01/16/content_2568722.htm [accessed 2023-09-16]
16. Multi-measures and measures to promote the construction of universal health information - China continues to enhance the public's sense of access to medical care. The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/xinwen/2022-09/02/content_5708099.htm [accessed 2023-09-16]
17. Lu WC, Tsai IC, Wang KC, Tang TA, Li KC, Ke YC, et al. Innovation resistance and resource allocation strategy of medical information digitalization. Sustainability 2021;13(14):7888. [doi: 10.3390/su13147888]
18. Han H, Hai C, Wu T, Zhou N. How does digital infrastructure affect residents' healthcare expenditures? Evidence from Chinese microdata. Front Public Health 2023;11:1122718 [FREE Full text] [doi: 10.3389/fpubh.2023.1122718] [Medline: 37213630]
19. Saeki M, Oyama S, Yoneda H, Shimoda S, Agata T, Handa Y, et al. Demonstration experiment of telemedicine using ultrasonography and telerehabilitation with 5G communication system in aging and depopulated mountainous area. Digit Health 2022;8:20552076221129074 [FREE Full text] [doi: 10.1177/20552076221129074] [Medline: 36262932]
20. Fan YJ, Yin YH, Xu LD, Zeng Y, Wu F. IoT-based smart rehabilitation system. IEEE Trans Ind Inf 2014;10(2):1568-1577. [doi: 10.1109/tii.2014.2302583]
21. Fan H, Fan L, Song Y. Network infrastructure construction and urban innovationvidence from a quasi-natural experiment. Ind Econ Rev 2022;13:52-67 In Chinese. [doi: 10.14007/j.cnki.cjpl.2022.04.004]
22. He X. Development trend of digital economy and China's strategic choice. Mod Econ Res 2013(3):39-43 [FREE Full text] [doi: 10.3969/j.issn.1009-2382.2013.03.008]
23. Dong F, Li Y, Qin C, Zhang X, Chen Y, Zhao X, et al. Information infrastructure and greenhouse gas emission performance in urban China: a difference-in-differences analysis. J Environ Manage 2022;316:115252. [doi: 10.1016/j.jenvman.2022.115252] [Medline: 35594820]
24. Fang Z, Razzaq A, Mohsin M, Irfan M. Spatial spillovers and threshold effects of internet development and entrepreneurship on green innovation efficiency in China. Technol Soc 2022;68:101844. [doi: 10.1016/j.techsoc.2021.101844]
25. Tang C, Xu Y, Hao Y, Wu H, Xue Y. What is the role of telecommunications infrastructure construction in green technology innovation? A firm-level analysis for China. Energy Econ 2021;103:105576. [doi: 10.1016/j.eneco.2021.105576]
26. Qian W, Liu H, Pan F. Digital economy, industry heterogeneity, and service industry resource allocation. Sustainability 2022;14(13):8020. [doi: 10.3390/su14138020]

XSL•FO
RenderX

27. Yu B, Feng M, Jiang N. Can internet development correct the misallocation of manufacturing resources?: a empirical test based on spatial Durbin model. R D Manag 2023;35(1):133-145 In Chinese. [doi: 10.13581/j.cnki.rdm.20211768]
28. Liu S, Wu Y, Wu Q. Can internet penetration improve human capital misallocation? Contemp Finance Econ 2022(6):12-25 In Chinese. [doi: 10.13676/j.cnki.cn36-1030/f.2022.06.005]
29. Tang C, Xue Y, Wu H, Irfan M, Hao Y. How does telecommunications infrastructure affect eco-efficiency? Evidence from a quasi-natural experiment in China. Technol Soc 2022;69:101963. [doi: 10.1016/j.techsoc.2022.101963]
30. Zhong X, Liu G, Chen P, Ke K, Xie R. The impact of internet development on urban eco-efficiency-a quasi-natural experiment of "Broadband China" pilot policy. Int J Environ Res Public Health 2022;19(3):1363 [FREE Full text] [doi: 10.3390/ijerph19031363] [Medline: 35162386]
31. Zhou L, Ampon-Wireko S, Dauda L, Xu X, Opuni Antwi M, Larnyo E. Empirical analysis of factors influencing healthcare efficiency among emerging countries. Healthcare (Basel) 2020;9(1):31 [FREE Full text] [doi: 10.3390/healthcare9010031] [Medline: 33396384]
32. Zhou Y, Zhao K, Han J, Zhao S, Cao J. Geographical pattern evolution of health resources in China: spatio-temporal dynamics and spatial mismatch. Trop Med Infect Dis 2022;7(10):292 [FREE Full text] [doi: 10.3390/tropicalmed7100292] [Medline: 36288033]
33. Wu X, Zhang Y, Guo X. Research on the equity and influencing factors of medical and health resources allocation in the context of COVID-19: a case of Taiyuan, China. Healthcare (Basel) 2022;10(7):1319 [FREE Full text] [doi: 10.3390/healthcare10071319] [Medline: 35885847]
34. Song X, Deng W, Zhou P. Spatial equity and influences of two-level public healthcare resources: a background to hierarchical diagnosis and treatment reform in China. Acta Geographica Sinica 2019;74(6):1178-1189 In Chinese. [doi: 10.11821/dlxb201906008]
35. Deng Z, Jiang N, Song S, Pang R. Misallocation and price distortions: a revenue decomposition of medical service providers in China. China Econ Rev 2021;65:101574. [doi: 10.1016/j.chieco.2020.101574]
36. Jiang Z, Xu S, Liu N. Information infrastructure, public service supply efficiency and urban innovation: an empirical study based on the "Broadband China" pilot policy. Urban Prob 2022(1):53-64 In Chinese. [doi: 10.13239/j.bjsshkxy.cswt.220106]
37. Anrijs S, Mariën I, De Marez L, Ponnet K. Excluded from essential internet services: examining associations between digital exclusion, socio-economic resources and internet resources. Technol Soc 2023;73:102211. [doi: 10.1016/j.techsoc.2023.102211]
38. He D, Gu Y, Shi Y, Wang M, Lou Z, Jin C. COVID-19 in China: the role and activities of Internet-based healthcare platforms. Glob Health Med 2020;2(2):89-95 [FREE Full text] [doi: 10.35772/ghm.2020.01017] [Medline: 33330783]
39. Chen X, Wu X, Zhang Q, Jing R, Cheng W, Tian J, et al. The construction and operational models of internet hospitals in China: a hospital-based survey study. BMC Health Serv Res 2023;23(1):669 [FREE Full text] [doi: 10.1186/s12913-023-09675-2] [Medline: 37344831]
40. Liu H, Liu Y. Construction of a medical resource sharing mechanism based on blockchain technology: evidence from the medical resource imbalance of China. Healthcare (Basel) 2021;9(1):52 [FREE Full text] [doi: 10.3390/healthcare9010052] [Medline: 33418859]
41. Hu Y, Wang Y, Lu S, Li Y. Impact of the development of information society on healthcare efficiency: empirical evidence from 31 Chinese provinces. Digit Health 2023;9 [FREE Full text] [doi: 10.1177/20552076231154375] [Medline: 36776406]
42. Yang K. Internet use, employment performance and the health of Chinese residents. Int Health 2022;14(3):222-235 [FREE Full text] [doi: 10.1093/inthealth/ihab034] [Medline: 34131727]
43. Liu W, Liu Y, Twum P, Li S. National equity of health resource allocation in China: data from 2009 to 2013. Int J Equity Health 2016;15:68 [FREE Full text] [doi: 10.1186/s12939-016-0357-1] [Medline: 27093960]
44. Yang B, Wen Q, Zhang Y, Wang X, Yin X, Li Q, et al. Economic value and characteristics of cloud pharmacy for children based on internet hospital in western China during the COVID-19 pandemic: cross-sectional survey study. Front Public Health 2022;10:1034450 [FREE Full text] [doi: 10.3389/fpubh.2022.1034450] [Medline: 36408041]
45. Zhang Y, Yang H, Pan J. Gaining from rational health planning: spatial reallocation of top-tier general hospital beds in China. Comput Ind Eng 2021;157:107344. [doi: 10.1016/j.cie.2021.107344]
46. Wan S, Chen Y, Xiao Y, Zhao Q, Li M, Wu S. Spatial analysis and evaluation of medical resource allocation in China based on geographic big data. BMC Health Serv Res 2021;21(1):1084 [FREE Full text] [doi: 10.1186/s12913-021-07119-3] [Medline: 34641850]
47. Li J, Chen X, Han X, Zhang G. Spatiotemporal matching between medical resources and population ageing in China from 2008 to 2017. BMC Public Health 2020;20(1):845 [FREE Full text] [doi: 10.1186/s12889-020-08976-z] [Medline: 32493251]
48. Liu C, Wang L. Does national broadband plan narrow regional digital divide? Evidence from China. Chin J Commun 2019;12(4):449-466. [doi: 10.1080/17544750.2019.1609539]
49. Li R, Shiu A. Internet diffusion in China: a dynamic panel data analysis. Telecommun Policy 2012;36(10-11):872-887. [doi: 10.1016/j.telpol.2012.06.004]
50. Tu C, Zang C, Tan Y, Zhou Y, Yu C. Can information infrastructure development improve the health care environment? Evidence from China. Front Public Health 2022;10:987391 [FREE Full text] [doi: 10.3389/fpubh.2022.987391] [Medline: 36091535]

XSL•FO
RenderX

51. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
52. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a Lancet commission on adolescent health and wellbeing. Lancet 2016;387(10036):2423-2478. [doi: 10.1016/s0140-6736(16)00579-1]
53. Park E, Kwon M. Health-related internet use by children and adolescents: systematic review. J Med Internet Res 2018;20(4):e120 [FREE Full text] [doi: 10.2196/jmir.7731] [Medline: 29615385]
54. Guldi M, Herbst CM. Offline effects of online connecting: the impact of broadband diffusion on teen fertility decisions. J Popul Econ 2016;30(1):69-91. [doi: 10.1007/s00148-016-0605-0]
55. Ralph LJ, Berglas NF, Schwartz SL, Brindis CD. Finding teens in TheirSpace: using social networking sites to connect youth to sexual health services. Sex Res Soc Policy 2011;8(1):38-49. [doi: 10.1007/s13178-011-0043-4]
56. Barros RP, Holanda PR, Sousa AD, Apostolico MR. Necessidades em saúde dos adolescentes na perspectiva dos profissionais da Atenção Primária à Saúde. Cien Saude Colet 2021;26(2):425-434 [FREE Full text] [doi: 10.1590/1413-81232021262.40812020] [Medline: 33605320]
57. Zhu H, Sang B, Zhang C, Guo L. Have industrial robots improved pollution reduction? A theoretical approach and empirical analysis. China World Econ 2023;31(4):153-172. [doi: 10.1111/cwe.12495]
58. Guo L, Chai Y, Yang C, Zhang L, Guo H, Yang H. Has smart city transition elevated the provision of healthcare services? Evidence from China's Smart City Pilot Policy. Digit Health 2023;9:20552076231197335 [FREE Full text] [doi: 10.1177/20552076231197335] [Medline: 37654714]
59. Hsieh CT, Klenow PJ. Misallocation and manufacturing TFP in China and India. Q J Econ 2009;124(4):1403-1448. [doi: 10.1162/qjec.2009.124.4.1403]
60. Olley GS, Pakes A. The dynamics of productivity in the telecommunications equipment industry. Econometrica 1996;64(6):1263. [doi: 10.2307/2171831]
61. Broda C, Weinstein DE. Globalization and the gains from variety. Q J Econ 2006;121(2):541-585. [doi: 10.1162/qjec.2006.121.2.541]
62. Wang W, Sun Z, Niu Z. Industrial policy, market competition and mismatch of resources. Economist 2014:22-32. [doi: 10.16158/j.cnki.51-1312/f.2014.09.015]
63. Li X. Digital rural construction, mismatch of resource elements and high-quality agricultural development. J Tech Econ Manag 2023(7):91-96 In Chinese. [doi: 10.3969/j.issn.1004-292X.2023.07.017]
64. Hu Z, Zhang M. Research on effect of new infrastructure construction on green technology innovation. Enterp Econ 2023;42(7):42-45 In Chinese [FREE Full text] [doi: 10.13529/j.cnki.enterprise.economy.2023.07.004]
65. Feng Y, Zhong S, Zhao J. Degree of marketization, resource mismatch and total factor productivity of enterprises. J Southwest Minzu University (Humanit Soc Sci Ed) 2020;41(5):100-112 [FREE Full text] [doi: 10.3969/j.issn.1004-3926.2020.05.016]
66. Pan Y, Gao H. Study on the resource mismatch effect of infrastructure investment. Reform 2019(7):62-72 [FREE Full text]
67. Liu R. Can regional economic integration improve the efficiency of resource allocation? Evidence from 26 cities in the Yangtze river delta. Nanjing J Soc Sci 2019(10):27-34 In Chinese. [doi: 10.15937/j.cnki.issn1001-8263.2019.10.004]
68. Niu Z, Cui B. Network infrastructure construction and labor allocation distortion: a quasi-natural experiment of the “broadband China” strategy. Stat Res 2022;39(10):133-148 In Chinese. [doi: 10.19343/j.cnki.11-1302/c.2022.10.009]
69. Bai J, Liu Y. Can outward foreign direct investment improve the resource misallocation of China. China Ind Econ 2018(1):60-78 In Chinese. [doi: 10.19581/j.cnki.ciejournal.20180115.002]
70. Zhang X, Wan G, Zhang J. Digital economy, financial inclusion, and inclusive growth. China Econ 2020;15(3):92-105. [doi: 10.19602/j.chinaeconomist.2020.05.07]
71. He Z, Song X. The mechanism and enlightenment of digital economy promoting employment: thinking after the coronavirus epidemic. Economist 2020(5):58-68 In Chinese. [doi: 10.16158/j.cnki.51-1312/f.2020.05.007]
72. Zhang X, Yang T, Wang C, Wan G. Digital finance and household consumption: theory and evidence from China. J Manage World 2020;36(11):48-63 In Chinese. [doi: 10.19744/j.cnki.11-1235/f.2020.0168]
73. Zhang H, Sun Z, Elahi E, Zhang Y. Internet development, level of industrial synergy, and urban innovation. Sustainability 2021;13(22):12410. [doi: 10.3390/su132212410]
74. Han X, Song W, Li B. Can the internet become a new momentum to improve the efficiency of regional innovation in China. China Ind Econ 2019:119-136 In Chinese. [doi: 10.19581/j.cnki.ciejournal.2019.07.007]
75. Zhang X. Research on evolution of innovation model under the condition of digital economy. Economist 2019:32-39 In Chinese. [doi: 10.16158/j.cnki.51-1312/f.2019.07.005]
76. Huang H, Zhang Z. Research on science and technology resource allocation efficiency in Chinese emerging strategic industries based on DEA model. China Soft Sci 2015(1):150-159 [FREE Full text] [doi: 10.3969/j.issn.1002-9753.2015.01.016]
77. Xu L, Hou L, Cheng G. Innovation Effect of National Big Data Comprehensive Pilot Zone. Sci Technol Prog Policy 2022;39(20):101-111 In Chinese [FREE Full text] [doi: 10.6049/kjjbydc.2021080751]
78. Bai J, Wang X, Bian Y. The impact of internet development on factors allocation distortion. J Quant Tech Econ 2022;39(11):71-90 In Chinese. [doi: 10.13653/j.cnki.jqte.2022.11.004]

XSL•FO
RenderX

79. Li X. Research on tourism industrial cluster and information platform based on internet of things technology. Int J Distrib Sens Netw 2019;15(7):155014771985884. [doi: 10.1177/1550147719858840]
80. Xie L. Distribution organization reconstruction in the internet age: backward supply chain integration perspective. China Ind Econ 2015(4):44-56 In Chinese. [doi: 10.19581/j.cnki.ciejournal.2015.04.005]
81. Han C, Zhang L. Can internet improve the resource misallocation of China: a empirical test based on the dynamic spatial Durbin model and threshold effect. Inq Econ Issues 2019;12:43-55 [FREE Full text]
82. Yu W, Wu S. Internet platform economy and the decreasing of market distortion. Finance Trade Econ 2020;41(5):146-160 In Chinese. [doi: 10.19795/j.cnki.cn11-1166/f.20200512.010]
83. Department of Development Planning. Basic national education development in 2022. Ministry of Education of the People's Republic of China. URL: http://www.moe.gov.cn/fbh/live/2023/55167/sfcl/202303/t20230323_1052203.html [accessed 2023-09-16]
84. Zhang H, Li J, Quan T. Has the construction of digital infrastructure promoted the integration of the digital economy and the real economy: a quasi - natural experiment based on the “broadband China” strategy. Inq Econ Issues 2023;10:1-15 [FREE Full text]
85. Zhong Z, Gao Z, Qin C. Network infrastructure construction and inter-city cooperative innovationvidence of the "Broadband China" pilot program and its spread. J Finance Econ 2022;48(3):79-93 In Chinese. [doi: 10.16538/j.cnki.jfe.20211113.402]
86. Bong Choi S, Williams C, Ho Ha S. Institutions and broadband internet diffusion in emerging economies: lessons from Korea and China. Innovation 2014;16(1):2-18. [doi: 10.5172/impp.2014.16.1.02]
87. Zhang K, Fu J, Shu Y. Researches on the impact of misallocation of medical resources on the performance of health care services based on GMM analysis of provincial dynamic panel data. Chin Health Serv Manag 2023;40(7):526-530 [FREE Full text]

Abbreviations

BCP: Broadband China Policy
DID: difference-in-differences
GDP: gross domestic product
IV: instrumental variable
MM: medical resource mismatch

Edited by T de Azevedo Cardoso, T McCall; submitted 24.10.23; peer-reviewed by Y Yang, S Wen, B Zhang; comments to author 12.02.24; revised version received 15.02.24; accepted 19.02.24; published 21.03.24.

Please cite as:
Chai Y, Yuan X, Guo L, Chen Z
The Impact of Broadband Infrastructure Construction on Medical Resource Mismatch: Quasi-Natural Experiment From the Broadband China Policy
J Med Internet Res 2024;26:e53921
URL: https://www.jmir.org/2024/1/e53921
doi:10.2196/53921
PMID:38512327

©Yulin Chai, Xiaoping Yuan, Lin Guo, Zhongming Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Economic Evaluations of Digital Health Interventions for Patients With Heart Failure: Systematic Review

Neily Zakiyah[1,2], PhD; Dita Marulin[1], MPharm; Mohammed Alfaqeeh[1], MPharm; Irma Melyani Puspitasari[1,2], Prof Dr; Keri Lestari[1,2], Prof Dr; Ka Keat Lim[3], PhD; Julia Fox-Rushby[3], Prof Dr

[1]Department of Pharmacology and Clinical Pharmacy, Faculty of Pharmacy, Universitas Padjadjaran, Bandung, Indonesia

[2]Center of Excellence for Pharmaceutical Care Innovation, Universitas Padjadjaran, Bandung, Indonesia

[3]Department of Population Health Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom

Corresponding Author:
Neily Zakiyah, PhD
Department of Pharmacology and Clinical Pharmacy
Faculty of Pharmacy
Universitas Padjadjaran
Jl. Raya Bandung Sumedang KM 21
Bandung, 45363
Indonesia
Phone: 62 22 7796200
Fax: 62 22 7796200
Email: neily.zakiyah@unpad.ac.id

Abstract

Background: Digital health interventions (DHIs) have shown promising results in enhancing the management of heart failure (HF). Although health care interventions are increasingly being delivered digitally, with growing evidence on the potential cost-effectiveness of adopting them, there has been little effort to collate and synthesize the findings.

Objective: This study's objective was to systematically review the economic evaluations that assess the adoption of DHIs in the management and treatment of HF.

Methods: A systematic review was conducted using 3 electronic databases: PubMed, EBSCOhost, and Scopus. Articles reporting full economic evaluations of DHIs for patients with HF published up to July 2023 were eligible for inclusion. Study characteristics, design (both trial based and model based), input parameters, and main results were extracted from full-text articles. Data synthesis was conducted based on the technologies used for delivering DHIs in the management of patients with HF, and the findings were analyzed narratively. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed for this systematic review. The reporting quality of the included studies was evaluated using the CHEERS (Consolidated Health Economic Evaluation Reporting Standards) guidelines.

Results: Overall, 27 economic evaluations were included in the review. The economic evaluations were based on models (13/27, 48%), trials (13/27, 48%), or a combination approach (1/27, 4%). The devices evaluated included noninvasive remote monitoring devices (eg, home telemonitoring using digital tablets or specific medical devices that enable transmission of physiological data), telephone support, mobile apps and wearables, remote monitoring follow-up in patients with implantable medical devices, and videoconferencing systems. Most of the studies (24/27, 89%) used cost-utility analysis. The majority of the studies (25/27, 93%) were conducted in high-income countries, particularly European countries (16/27, 59%) such as the United Kingdom and the Netherlands. Mobile apps and wearables, remote monitoring follow-up in patients with implantable medical devices, and videoconferencing systems yielded cost-effective results or even emerged as dominant strategies. However, conflicting results were observed, particularly in noninvasive remote monitoring devices and telephone support. In 15% (4/27) of the studies, these DHIs were found to be less costly and more effective than the comparators (ie, dominant), while 33% (9/27) reported them to be more costly but more effective with incremental cost-effectiveness ratios below the respective willingness-to-pay thresholds (ie, cost-effective). Furthermore, in 11% (3/27) of the studies, noninvasive remote monitoring devices and telephone support were either above the willingness-to-pay thresholds or more costly than, yet as effective as, the comparators (ie, not cost-effective). In terms of reporting quality, the studies were classified as *good* (20/27, 74%), *moderate* (6/27, 22%), or *excellent* (1/27, 4%).

XSL•FO
RenderX

Conclusions: Despite the conflicting results, the main findings indicated that, overall, DHIs were more cost-effective than non-DHI alternatives.

Trial Registration: PROSPERO CRD42023388241; https://tinyurl.com/2p9axpmc

(J Med Internet Res 2024;26:e53500) doi:10.2196/53500

KEYWORDS

digital health; telemonitoring; telehealth; heart failure; cost-effectiveness; systematic review; mobile phone

Introduction

Background

Heart failure (HF) is a complex and potentially fatal condition affecting approximately 26 million people worldwide and is associated with substantial morbidity and mortality [1,2]. The global impact of HF also imposes a significant economic burden, affecting patients and their families as well as communities [1,2]. Data from low- and middle-income countries (LMICs) indicate that mortality rates in patients with HF are higher in LMICs than in high-income countries (HICs) [3]. The overall estimated 1-year mortality rate for patients with HF in LMICs is 16.5% [3] compared to 8.3% in HICs. People living with HF also experience a significant decline in health-related quality of life (HRQoL) [4,5].

The growing availability of life-saving and evidence-based treatments, along with increasing life expectancy, suggests that there will be an increase in the prevalence of HF over time. This is attributed to the improved survival rates after an HF diagnosis and the aging population [1,6,7]. The rise in HF prevalence is leading to an increase in annual health care costs. In 2012, the estimated global annual cost of HF reached US $108 billion, with direct costs estimated at US $65 billion and indirect costs estimated at US $43 billion [8]. Considering a projected 22% increase in the cost of cardiovascular diseases (CVDs), HF-related expenses alone could potentially reach US $132 billion by 2030 [9]. Despite significant improvements in outcomes with medical therapy [7], readmission rates for patients hospitalized for HF are still high (ie, 50% within 6 months of discharge) [10,11]. Hospitalization rates have been shown to be correlated with disease severity, mortality, and lower HRQoL [5].

Considering the prevalence of HF and its substantial financial burden, there has been a global focus on cost-effective health care interventions aimed at providing effective and efficient support to patients, as well as a growing focus on the application of digital health interventions (DHIs), driven by the advanced integration of IT and mobile internet in health care practices [12]. The broad scope of digital health includes telehealth, teleconsultation, and telemonitoring using smartphone apps; telephone support; videoconferencing; noninvasive remote monitoring devices; wearables; implantable devices; and sensors [13,14]. In addition, emerging fields such as advanced computing sciences in big data, genomics for personalized medicine, and artificial intelligence have been recognized as DHIs [13,15-17]. DHIs are used by providers and other stakeholders to enhance access, reduce inefficiencies and costs, improve quality, and potentially incorporate personalized medicine to improve patients' clinical outcomes [14].

By using DHIs in the management of HF, it may be possible to prevent the progression of a patient's condition and potentially reduce health care costs [18]. HF is a chronic condition in which people often experience episodic deterioration. Improvement in disease monitoring can enable prompt identification of patient deterioration and facilitate timely interventions to restabilize the syndrome [19,20]. Implementing DHIs such as teleconsultation and remote monitoring can reduce unnecessary hospital visits, provide continuous disease monitoring, develop effective disease management, and improve clinical outcomes. However, because the landscape of DHIs is evolving rapidly, regulators, reimbursement authorities, and health care professionals often face challenges in evaluating the value of these technologies, as reflected in current recommendations in the international guidelines of the European Society of Cardiology and the American College of Cardiology [19,20]. Skepticism regarding the value of DHIs is partly driven by the limited large-scale studies that demonstrate a consistent impact and effectiveness [18].

Despite the growth in, and the integration of, DHIs in recent years, evidence from economic evaluations is limited. One systematic review found that telemedicine improved clinical outcomes and resulted in cost savings for patients with CVDs, concluding that it is more cost-effective than standard of care (SoC) [21]. This mirrors the broad findings of 2 other systematic reviews covering a diverse range of DHIs [22,23]. However, 2 reviews were more focused on CVDs than on HF [21,22], with 1 review not including search terms related to HF [21]; 2 reviews focused on either economic models or randomized controlled trials (RCTs) but not both [22,23]; and 2 reviews only considered a limited range of DHIs [21,23]. To date, no comprehensive systematic review has been conducted to evaluate the economic evaluations of DHIs specifically in patients with HF, considering evidence from both models and RCTs.

Objectives

The aim of this systematic review was to provide an overview and summarize published economic evaluations of DHIs in patients with HF that consider both models and analyses conducted alongside trial-based evaluations. Demonstrating the cost-effectiveness of DHIs will contribute to a better understanding of the potential economic implications of adopting these approaches.

XSL•FO
RenderX

Methods

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines [24], and the review was registered in PROSPERO [25]. The PRISMA checklist is provided in Multimedia Appendix 1.

Search Strategy

A systematic search was performed across 3 major electronic databases (PubMed, EBSCOhost, and Scopus) to identify economic evaluations of DHIs for patients with HF. Medical Subject Headings terms and text words related to "heart failure," "digital health," and "economic evaluation" were used to search from database inception to July 2023. Terms were combined using "OR" and "AND." Full details are provided in Multimedia Appendix 2.

Study Selection

The search results were exported to Mendeley Reference Manager (Elsevier Ltd) and checked for duplicates. Two reviewers (NZ and DM) independently performed a full-text review of the chosen articles after the preliminary title and abstract screening, using the inclusion and exclusion criteria detailed in Textbox 1.

Any disagreements were discussed, and a third reviewer was consulted for arbitration to arrive at a consensus if required. References were also searched for further relevant papers during the full-text reviews.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Type of study: a full economic evaluation of digital health interventions (DHIs) for the management of patients with heart failure (HF), categorized as cost-benefit analysis, cost-utility analysis, cost-effectiveness analysis, and cost-minimization analysis
- Intervention: any DHI for patients with HF comprising a digital intervention for transmitting medical information to improve patients' health status (DHIs are a broad concept encompassing eHealth, which refers to the application of information and communications technology in support of health and health-related fields; this includes the use of mobile devices such as smartphones and patient monitoring devices in medical and public health practices, commonly known as mobile health. DHIs also comprise emerging domains such as advanced computing sciences in big data, genomics, and artificial intelligence [13,26]. Standard of care was defined as the standard multidisciplinary management program [19,20], which includes regular planned follow-up for the purpose of safety and optimal drug dosing (standard of care with or without drug or exercise prescription), early detection of decompensation, and impact on disease progression that requires modification of the intervention or treatment regimen)
- Participants: adult patients with HF (aged ≥18 y)
- Time limits: searches were conducted for relevant articles published from the beginning of database entries to July 2023

Exclusion criteria

- Non-English studies, experimental and observational studies without economic evaluation, studies that did not report outcomes specific to HF, reviews, conference abstracts, and editorials

Data Extraction

Two reviewers conducted data extraction from the full-text articles independently using a predetermined form covering general study characteristics (author, country, and year of publication), study design (type of economic evaluation, perspective, model type, time horizon, discount rate, intervention vs comparator, outcome measures, and sensitivity analysis), primary outcomes, and quality of reporting. Only results related to DHIs for patients with HF were extracted when many interventions were evaluated. The primary outcomes collected were the cost-benefit ratio, cost savings, and cost-effectiveness of DHIs. Cost-effectiveness is represented by the incremental cost-effectiveness ratio (ICER) per quality-adjusted life year (QALY) gained or per intermediate outcome measure such as mortality or hospitalization.

Ethical Considerations

As we exclusively used published studies for this systematic review and did not involve patients or the public or conduct any patient interviews, a review by, or approval from, an institutional review board was not required.

Quality of Reporting

The CHEERS (Consolidated Health Economic Evaluation Reporting Standards) checklist was used to assess the reporting quality of each study [27]. The CHEERS checklist includes 28 items, with 1 point assigned to each item when the quality criterion is fulfilled (and 0 points for not entirely conforming to the relevant criterion) to generate a total score, with 28 (representing 100%) being the maximum score. On the basis of the scores, studies are classified into 4 quality categories: excellent (score: 100%), good (score: 75%-99%), moderate (score: 50%-74%), and low (score: ≤49%) [28]. This reflects reporting quality rather than a view of overall importance or methodological quality.

Analysis and Presentation of Results

The results are presented in a range of narrative tables by study. The included studies were categorized by the device or technology used for delivering DHIs in managing patients with HF (ie, noninvasive remote monitoring devices, telephone support, mobile apps and wearables, remote monitoring follow-up in patients with implantable medical devices, and videoconferencing systems) [13,17]. Secondary categorization separates results by model-based and RCT-based studies. Money

XSL•FO
RenderX

values were converted to 2023 US dollars using the Campbell and Cochrane Economics Methods Group–Evidence for Policy & Practice Information Centre Cost Converter [29]. If the study did not specify the costing year, publication year was assumed to be the year of costing. A 3×3 permutation matrix shows how each intervention's outcomes (improved, worsened, or unchanged) and costs (increased, decreased, or unchanged) compare with those of its comparator in the studies [30]. This permutation matrix also splits the findings by DHI type.

Results

Study Identification

The initial search retrieved 365 studies, of which 7 (1.9%) duplicates were excluded. Of the 358 studies left, the title and abstract screening process excluded 296 (82.7%). After a full-text screening of the remaining 62 studies, we excluded 37 (60%; 21/37, 57% were classified as partial economic evaluations, such as cost analysis, containing only descriptions of costs; 10/37, 27% did not report outcomes specifically for HF; and 6/37, 16% were conference abstracts), resulting in 25 (40%) out of 62 studies for inclusion in the analysis. Two extra studies were identified from reference reviews; thus, 27 studies [31-57] were included in this systematic review.

The selection process and flow diagram for the identification of studies are depicted in Figure 1.

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process.

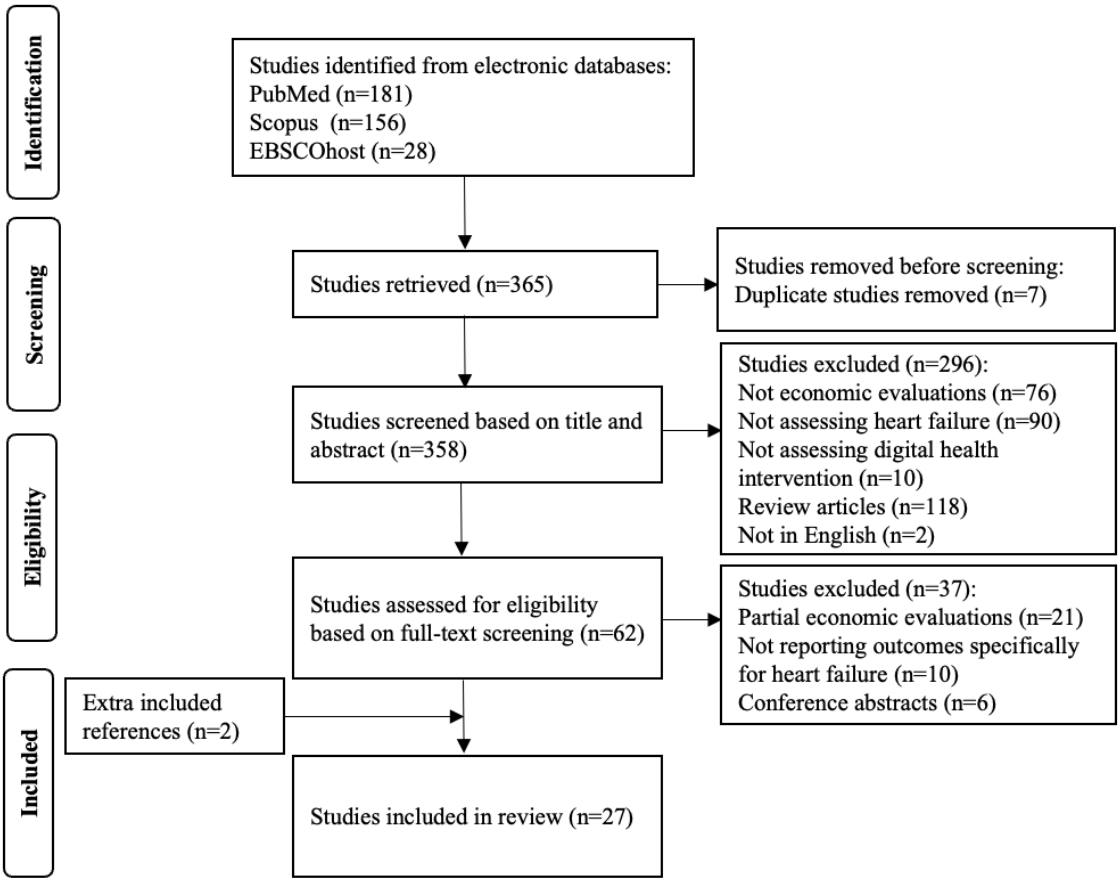

Study Characteristics and Design

Table 1 summarizes the general characteristics of the included studies. Of the 27 studies, 13 (48%) were conducted using a decision analytical model [31-35,40,47-51], 13 (48%) used trial-based data [36-39,41-46,54-57], and 1 (4%) used a combination of both [41]. The majority of the studies (25/27, 93%) were from HICs. Of the 27 studies, 6 (22%) were from the United States [34,43,44,51-53]; 3 (11%) each from the Netherlands [33,37,40] and the United Kingdom [32,35,50]; 2 (7%) each from Germany [36,38], Brazil [41,45], Canada [42,48], Italy [55,56], and Spain [46,49]; and 1 (4%) each from Australia [57], Poland [54], France [31], Hong Kong [47], and Denmark [39]. Of the 27 studies, 24 (89%) conducted a cost-utility analysis with cost and QALY as the outcome measures [31-35,37-40,42-51,54,56,57], 2 (7%) conducted a cost-effectiveness analysis (1/2, 50% with hospital readmission as the outcome measure [41] and 1/2, 50% with number of days alive and neither in hospital nor in inpatient care as the outcome measure [36]), and 1 (3%) conducted a cost-minimization analysis [55].

XSL•FO
RenderX

Table 1. General characteristics of included studies.

Study characteristics	Studies (n=27), n (%)
Type of economic evaluation	
Cost-utility analysis	24 (89)
Cost-effectiveness analysis	2 (7)
Cost-minimization analysis	1 (4)
Year of publication	
Before 2010	2 (7)
2011-2015	6 (22)
2016-2020	12 (45)
2021-2023	7 (26)
Region	
Europe	15 (56)
North and South America	9 (33)
Asia Pacific	3 (11)
Perspective	
Health care system	16 (59)
Health care provider	4 (15)
Health care system and societal	4 (15)
Societal	3 (11)
Study type	
Model based	13 (48)
Randomized controlled trial based	13 (48)
Combination	1 (4)
Time horizon	
Nonlifetime	15 (56)
Not stated	6 (22)
Lifetime	6 (22)
Outcome measures	
Quality-adjusted life years	24 (89)
Other effects	2 (7)
Not stated	1 (4)
Funding	
Nonprivate	15 (56)
Not stated	7 (26)
Private	3 (11)
No funding	2 (7)

Of the 27 studies, 15 (56%) used a time horizon of >1 year (3/15, 20% were RCTs) [31-35,40,44,45,47,48,50-54]; the time horizon in 6 (22%) studies was ≤1 year (all were RCTs) [36,39,41-43,46]; 6 (22%) did not state the time horizon (5/6, 83% were RCTs and 1/6, 17% was model based) [37,38,49,55-57]. Of the 27 studies, 15 (56%) received grants from public organizations [31,32,35,36,38-41,43,44,46,49,54,56,57], 3 (11%) received funding from industry [37,50,51], 2 (7%) received no funding [42,55], and 7 (26%) did not declare their funding source [33,34,45,47,48,52,53].

A little more than half of the studies (14/27, 52%) measured effectiveness with HRQoL (or utilities) using a patient-based EQ-5D instrument [31,32,34,35,37-40,48,50-53,56,57]. Other studies used the generic Short Form Health Survey-36 with norm-based scoring [42,44,54], the Minnesota Living with Heart

XSL•FO
RenderX

Failure Questionnaire [45], or a combination of both [36,43]. All studies included the direct costs of DHIs, such as the costs of the DHIs, inpatient and outpatient costs, monitoring and follow-up costs, and medication costs. Some of the studies (12/27, 44%) included nonmedical direct costs, such as travel and transportation costs [33,35,38,41,43,46,50,55]. Studies that used a societal perspective (6/27, 22%) also included indirect costs, such as productivity losses [33,38,43,46,55,56]. Details on the perspectives and included costs are provided in Multimedia Appendix 3 [19,20,31-57].

Cost-Effectiveness of Devices or Technologies Used for Delivering DHIs in the Management of Patients With HF

Overview

This subsection describes the nature of the DHIs assessed for cost-effectiveness and presents cost-effectiveness findings by type of DHI in order of the number of studies identified. Details on summaries and outcomes from the studies are provided in Table 2; and relative costs, effects, and main outcomes are presented in Table 3. Overall, of the 27 studies, 24 (89%) found the DHIs to be cost-effective [31-34,37-43,45-57], whereas 3 (11%) were not cost-effective, particularly home telemonitoring (HTM) and telephone support [35,36,44].

XSL•FO
RenderX

Table 2. Summary and quality assessment of the included studies.

Study	Country	Time horizon	Discount rate (%)	ICER[a] (in 2023 US $)	WTP[b] threshold (in 2023 US $/QALY[c])	Quality of reporting (CHEERS[d] checklist)
Noninvasive remote monitoring devices (n=9)						
Model based						
Caillon et al [31], 2022	France	10 y	2.5	• 8456/QALY • 5955/LY[e]	15,372	Excellent (score: 28/28, 100%)
Thokala et al [32], 2013	United Kingdom	30 y	3.5	• 20,715/QALY	34,895	Good (score: 25/28, 89%)
Albuquerque de Almeida et al [33], 2022	Netherlands	Lifetime	4	• SoC[f] vs HTM[g]: 45,277/QALY • SoC vs HTM+DA[h]: 36,422/QALY	105,146	Good (score: 24/28, 86%)
Jiang et al [34], 2020	United States	35 y	3	• 40,691/QALY • 37,641/QALY • 106,837/QALY	53,183	Moderate (score: 21/28, 75%)
Thokala et al [35], 2020	United Kingdom	Lifetime	3.5	• 72,028/QALY	29,904	Good (score: 22/28, 79%)
RCT[i,j] based						
Völler et al [36], 2022	Germany	1 y	__[k]	• −1474/d	—	Moderate (score: 20/28, 71%)
Boyne et al [37], 2013	Netherlands	—	N/A[l]	• 59,822/QALY	74,182	Moderate (score: 19/28, 68%)
Sydow et al [38], 2021	Germany	—	—	• Dominant (cost savings: 2358 per patient year)	—	Good (score: 21/28, 75%)
Vestergaard et al [39], 2020	Denmark	1 y	4	• 8020/QALY	31,064	Good (score: 25/28, 89%)
Telephone support (n=7)						
Model based						
Grustam et al [40], 2018	Netherlands	20 y	4	• UC[m] vs HTM: 17,597/QALY • UC vs NTS[n]: 11,661/QALY	112,811	Good (score: 25/28, 89%)
RCT based						
Ruschel et al [41], 2018	Brazil	6 mo	N/A	• PHS[o] framework: 332 per hospital readmission prevented; the private health care system, using a perspective of private health care system, the intervention was dominant (cost saving)	—	Good (score: 25/28, 89%)
Cui et al [42], 2013	Canada	1 y	N/A	• 3331/QALY	55,985	Good (score: 24/28, 86%)

XSL•FO
RenderX

Study	Country	Time horizon	Discount rate (%)	ICER[a] (in 2023 US $)	WTP[b] threshold (in 2023 US $/QALY[c])	Quality of reporting (CHEERS[d] checklist)
Hebert et al [43], 2008	United States	1 y	N/A	• Societal: 26,273/QALY • Payer: 5500/QALY	37,441	Good (score: 23/28, 82%)
Smith et al [44], 2008	United States	18 mo	—	• 212,586/QALY	144,744	Good (score: 21/28, 75%)
Bocchi et al [45], 2018	Brazil	Mean 2.47 (SD 1.75) y	—	• 4114/QALY	10,825	Moderate (score: 20/28, 71%)
Gonzalez-Guerrero et al [46], 2018	Spain	1 y	5	• Health care: 6611/QALY • Societal: 43,856/QALY	76,002	Good (score: 24/28, 86%)
Remote monitoring follow-up in patients with implantable medical devices (n=7)						
Model based						
Cowie et al [50], 2017	United Kingdom	10 y	3.5	• 31,177/QALY	32,351	Good (score: 25/28, 89%)
Schmier et al [51], 2016	United States	5 y	3	• 50,571/QALY	112,993	Good (score: 21/28, 75%)
Sandhu et al [52], 2016	United States	Lifetime	3	• 82,282/QALY	172,712	Good (score: 21/28, 75%)
Martinson et al [53], 2017	United States	5 y	3	• 13,855/QALY	56,496	Good (score: 23/28, 82%)
RCT based						
Niewada et al [54], 2021	Poland	Lifetime	3.5	• 56,333/QALY	91,300	Good (score: 21/28, 75%)
Calò et al [55], 2013	Italy	—	—	—	—	Moderate (score: 16/28, 57%)
Zanaboni et al [56], 2013	Italy	—	—	• Intervention dominant	62,166	Moderate (score: 20/28, 71%)
Mobile apps and wearables (n=3)						
Model based						
Jiang et al [47], 2021	Hong Kong, Special administrative region, China	10 y or until death, whichever occurred first	3	• 4380/QALY	49,949	Good (score: 21/28, 75%)
Boodoo et al [48], 2020	Canada	25 y	1.5	• 7127/QALY	40,119	Good (score: 25/28, 89%)
Cano Martin et al [49], 2014	Spain	—	3	• 16,064/QALY	—	Good (score: 21/28, 75%)
Videoconferencing system (n=1)						
RCT based						

XSL•FO
RenderX

Study	Country	Time horizon	Discount rate (%)	ICER[a] (in 2023 US $)	WTP[b] threshold (in 2023 US $/QALY[c])	Quality of reporting (CHEERS[d] checklist)
Hwang et al [57], 2018	Australia	—	—	• −3325/QALY (savings)	40,000	Good (score: 24/28, 86%)

[a]ICER: incremental cost-effectiveness ratio.
[b]WTP: willingness-to-pay.
[c]QALY: quality-adjusted life year.
[d]CHEERS: Consolidated Health Economic Evaluation Reporting Standards.
[e]LY: life-year.
[f]SoC: standard of care.
[g]HTM: home telemonitoring.
[h]DA: diagnostic algorithm.
[i]RCT: randomized controlled trial.
[j]RCT-based evaluation extended with a decision tree model (combination).
[k]Not stated.
[l]N/A: not applicable.
[m]UC: usual care.
[n]NTS: nurse telephone support.
[o]PHS: public health care system.

XSL•FO
RenderX

Table 3. Relative costs, effects, and main outcomes.

Relative cost	Relative effect		
	–[a]	0[b]	1[c]
–	No study	• DHI[d] is not cost-effective (1/27, 4%) • Noninvasive remote monitoring devices • SoC[e]+interactive bidirectional HTM[f] system (Motiva) vs SoC and patient diary to document health issues once a week [36][g]	• DHI is cost-effective (15/27, 56%) • Noninvasive remote monitoring devices • SCAD[h], home-based interactive telemonitoring service vs standard hospital-based care [31][i] • STSHM[j] interface+STSHH[k] contact+HTM vs SoC [32][i] • HTM+DA[l] vs SoC or HTM+DA vs HTM only [33][i] • Universal SoC+HTM for NYHA[m] class II to IV and class III to IV vs SoC [34][i] • Telephone support • HTM or nurse telephone support vs SoC+patient evaluation at the clinic every 4 months [40][i] • Nurse-led home visit vs regular visit to outpatient clinic [41][g,n] • HL[o] (nurses and health care providers providing telephone support)+SoC or HL+in-house monitoring+SoC vs SoC [42][g] • Nurse-led program vs SoC [43][g] • Mobile apps and wearables • Add-on HTM via app vs SoC [47][i] • HTM system (Medly) via app vs SoC, including specialized multidisciplinary HF[p] clinics [48]i • Remote monitoring follow-up in patients with implantable medical devices • Implantable hemodynamic sensor (CardioMEMS HF system) vs implantable usual care [50][i,q] • Implantable hemodynamic sensor (CardioMEMS HF system) vs implantable usual care [51][i,q] • Implantable hemodynamic sensor (CardioMEMS HF system) vs implantable usual care [52][i,q] • Implantable hemodynamic sensor (CardioMEMS HF system) vs implantable usual care [53][i,q] • HCTR[r], including telecare, telerehabilitation, and implantable+SoC vs SoC only [54][g] • DHI is not cost-effective (2/27, 7%) • Noninvasive remote monitoring devices • HTM vs SoC [35][i] • Telephone support • DM[s] (telephone support+augmented HTM) vs SoC [44][g]
0	No study	• DHI is cost-effective (1/27, 4%) • Noninvasive remote monitoring devices • HTM vs SoC [37][g]	N/A[t]

XSL•FO
RenderX

Relative cost	Relative effect		
	–[a]	0[b]	1[c]
1	No study	• DHI is cost saving (2/27, 7%) • Remote monitoring follow-up in patients with implantable medical devices • ICD[u] follow-up vs quarterly in-hospital follow-ups [55][g] • Videoconferencing system • Web-based telerehabilitation vs in-person center-based program [57][g]	• DHI is dominant (6/27, 22%) • Noninvasive remote monitoring devices • Additional noninvasive structured RPM[v] vs SoC [38][g] • HTM with a telekit (consisting of a tablet, a digital blood pressure monitor, and a scale) vs SoC [39][g] • Telephone support • HTM via telephone follow-up vs SoC [45][g] • DMP[w] vs postdischarge SoC [46][g] • Mobile apps and wearables • CardioManager app vs SoC [49][i] • Remote monitoring follow-up in patients with implantable medical devices • Wireless transmission–enabled ICD vs scheduled in-person evaluations [56][g]

[a]Digital health intervention has lower cost or lower effectiveness than the comparator.

[b]0: digital health intervention has the same cost and same effectiveness as the comparator.

[c]1: digital health intervention has a higher cost or higher effectiveness than the comparator.

[d]DHI: digital health intervention.

[e]SoC: standard of care (as defined by the European Society of Cardiology and the American Heart Association, American College of Cardiology, and Heart Failure Society of America, it is the standard multidisciplinary management program, which includes regular planned follow-up for the purpose of safety and optimal drug dosing [standard of care with or without drug or exercise prescription], early detection of decompensation, and impact on disease progression that requires modification of the intervention or treatment regimen).

[f]HTM: home telemonitoring.

[g]Randomized controlled trial based.

[h]SCAD: Suivi Clinique A Domicile (Clinical Follow-Up At Home).

[i]Model based.

[j]STSHM: structured telephone support via human-to-machine.

[k]STSHH: structured telephone support via human-to-human.

[l]DA: diagnostic algorithm.

[m]NYHA: New York Heart Association.

[n]Randomized controlled trial–based evaluation extended with a decision tree model (combination).

[o]HL: Health Lines.

[p]HF: heart failure.

[q]Implantable usual care described as patients with the device implanted but where the data were not used to guide management for remote monitoring.

[r]HCTR: hybrid comprehensive telerehabilitation.

[s]DM: disease management.

[t]N/A: not applicable.

[u]ICD: implantable cardioverter defibrillator.

[v]RPM: remote patient management.

[w]DMP: Disease management program.

Noninvasive Remote Monitoring Devices

Noninvasive remote monitoring devices (n=9) assessed for cost-effectiveness included HTM using medical devices and digital tablets. These devices enable the monitoring of a patient's vital parameters at home, including weight, blood pressure, heart rate, and heart rhythm. These devices enable the transmission of physiological data to the health care team, allowing for early detection of deterioration in patients with HF [31-39]. The prompt sending of these data to health care professionals for assessment facilitates the timely identification of significant changes and enables early interventions [37,38]. Early interventions help prevent complications and enable

XSL•FO
RenderX

patients to avoid emergency admissions, improving patient outcomes [31,32,39].

Most of the economic evaluations (7/9, 78%) of noninvasive remote monitoring devices were compared to SoC [19,20] as defined in the international guidelines [31,33-35,38,39]. Although the definitions of SoC are similar, some of the studies (2/9, 22%) provided additional details regarding the follow-up procedures, such as SoC with follow-up once a week [36] or 4 preplanned outpatient clinic visits [37]. Among the 9 studies, 4 (44%) used Markov models [31,32,34,35], 1 (11%) used a patient-level discrete-event simulation model [33], and 4 (44%) were trial based [36-39]. Some of the economic evaluations (4/9, 44%) showed that the implementation of noninvasive remote monitoring devices requires extra costs, mainly regarding the cost of HTM for HF management [31-34]. Nonetheless, the use of this technology was also accompanied by improved outcomes, such as improved HRQoL [31,32,34,35,37-39]. Although the majority of the results suggested that DHIs were cost-effective, the findings were conflicting. Although most of the studies (7/9, 78%) indicated that noninvasive remote monitoring devices for managing patients with HF were generally cost-effective [31-34,37-39], 22% (2/9) found dissimilar results: of these 2 studies, 1 (50%) conducted from the UK health care perspective reported that the incremental cost per QALY gained for HTM using noninvasive remote monitoring devices exceeded the acceptable willingness-to-pay (WTP) thresholds [35], while 1 (50%) conducted in Germany concluded that remote monitoring had higher costs and worse outcomes than SoC and was therefore not an efficient option [36].

Telephone Support

Structured telephone support (n=7), defined in the included studies, refers to the provision of HTM through self-care support or management by health care professionals, such as nurses, through regular telephone calls, typically on a monthly basis [42,43,45,46]. Of the 7 economic evaluations that used telephone support, 6 (86%) were based on RCTs [41-46], and 1 (14%) was model based [40]. The primary objective of telephone support includes assessing symptoms, reviewing current medications, and providing timely feedback to both physicians and patients [42]. The length of the intervention ranged from 4 to 30 months. The extra costs associated with the telephone support intervention compared to SoC included the costs of telephone calls and specialist follow-up visits. The outcomes measured included hospital readmission prevented over 24 weeks [41] and HRQoL [40,42,44-46]. The comparator SoC adhered to the definition provided in the guidelines [19,20], or it involved routine ambulatory evaluations in 3 to 4 months [45]. Overall, the results showed that telephone support was cost-effective compared to SoC (6/7, 86%) [40-43,45,46]. However, 1 (14%) of the 7 studies concluded that telephone support was not cost-effective in the United States because it surpassed the acceptable cost-effectiveness threshold as higher total costs in the intervention group were combined with a relatively small difference in health outcomes compared to the SoC group (ie, usual management by physicians) [44].

Remote Monitoring Follow-Up in Patients With Implantable Medical Devices

Of the 27 studies, 7 (26%) assessed the cost-effectiveness of remote monitoring follow-up in patients with implantable medical devices. Of these 7 studies, 3 (43%) were trial based [54-56], and 4 (57%) were model based [50-53]. The interventions included remote monitoring follow-up for patients using cardiac implantable electronic devices, which are used to manage conditions such as bradycardia and HF to prevent sudden cardiac death [58]. The 4 model-based studies [50-53] assessed the same device, that is, the CardioMEMS implantable hemodynamic sensor, which provides remote real-time pressure measurements from the pulmonary artery [59]. This wireless sensor transmits hemodynamic information to the patient database website, enabling health care professionals to promptly make decisions regarding treatment initiation and adjustments when changes in pulmonary artery pressure and signs of HF are detected. The comparator comprised usual care described as patients with the device implanted but where the data were not used to guide management for remote monitoring (implantable usual care) [50-53].

The 3 trial-based studies focused on remote monitoring follow-up with patients having cardiac implantable electronic devices compared to conventional follow-up [54-56]. The comparator included patients who typically attended regular follow-up visits at the clinic based on a predetermined calendar schedule [54]. Furthermore, 2 (67%) of these 3 studies provided specific details of their study settings: of these 2 studies, 1 (50%) described outpatient clinic visits every 3 to 6 months according to the standard schedule at the participating center [55], while 1 (50%) had scheduled in-office visits at 4, 8, 12, and 16 months [56]. Generally, costs associated with implantable devices followed by remote monitoring consist of the cost of visits to physicians and nurses and fees for the remote monitoring service, as well as the costs of the transmitter device, battery replacement, and cardiovascular treatment.

All studies indicated that implantable medical devices, especially for patients with severe HF (eg, New York Heart Association class III and class IV), were considered cost-effective [50-56]. In Italy, they were even deemed a dominant strategy, leading to improved health outcomes while incurring lower total costs [56].

Mobile Apps and Wearables

Of the 27 studies, 3 (11%) assessed the cost-effectiveness of providing HTM through expert counseling services via mobile apps [47-49]. The mobile apps provide a platform for patients to self-manage their heart condition [48]. The information section in the app contains a patient manual and medical information [49], while a separate section enables users to track their activity (physical activity and food consumption) and record health measurements such as vital signs [47-49]. In addition, the app includes a medication registry feature that allows patients to set reminders for medication administration times [49]. The features of mobile apps in the included studies were similar, that is, they included a feature that allowed the patient to transmit vital measurements (heart rate, blood pressure, and weight) daily to the HF management team,

XSL•FO
RenderX

followed by interpretation by experts and categorization of the patient's condition as well as feedback regarding the patient's condition such as medication dosage adjustments or recommendations for the patients to visit the emergency department. Reminders for patients to enter data were in the form of alarms [48,49]. All studies in this group conducted a model-based economic evaluation that compared add-on mobile apps to SoC [19,20]. The additional costs of mobile app technology rely mainly on monitoring and treatment, with outcomes captured as HRQoL. All included studies indicated that the mobile apps were cost-effective (ie, below the WTP thresholds in each setting) [47-49].

Videoconferencing System

Only 1 (4%) of the 27 studies assessed the cost-effectiveness of providing specialist consultation services to remote patients with HF via a videoconferencing system (known as telerehabilitation) [57]. A web-based commercial videoconferencing platform was used for synchronized audiovisual communication with groups of up to 4 participants [57]. The videoconferencing system equipment included a laptop computer and mobile broadband devices connected to 3G wireless broadband internet; in addition, the participants were provided a finger pulse oximeter, an automatic sphygmomanometer, free weights, and resistance bands [57]. In telerehabilitation, a physiotherapist supervised each training session, and a physiotherapist and a nurse led the information session [57]. The results suggested no significant differences in QALYs, but the health care costs per participant were significantly lower in the telerehabilitation group, with a savings of US $3325 per QALY [57].

Quality of Reporting

Of the 27 studies, 20 (74%) were rated *good* [32,33,35,38-44,46-54,57], 6 (22%) were rated *moderate* [34,36,37,45,55,56], and 1 (4%) was rated *excellent* [31]. Table 2 shows the percentage of items fulfilled by each study according to the CHEERS checklist.

The degree of adherence to the reporting criteria in the CHEERS checklist varied across the sections. Some items, such as background, intervention and comparator, study findings, generalizability, and funding, were adequately reported by all studies. The CHEERS checklist emphasizes the inclusion of essential and specific elements in the methods section, and nearly all studies included in this analysis comply with the checklist requirements, for example, the measurement and valuation of resources and costs (26/27, 96%) [31-33,35-57], perspective (26/27, 96%) [31-49,51-57], setting and location (25/27, 93%) [31-43,45-50,52-57], the measurement of outcomes (25/27, 93%) [31-36,38-48,50-57], and the selection of outcomes (24/27, 89%) [31-43,45-48,50-54,56,57]. However, there were certain items in the methods section that were reported less often; for instance, the approach to engagement with patients and other individuals affected by the study was only addressed in 2 (7%) of the 27 studies [31,39], and the impact of such engagement was discussed in only 1 (4%) of the 27 studies [31]. This limited reporting may be due to the fact that these items apply specifically to evaluation from trial-based data.

Discussion

Principal Findings

This systematic review comprehensively searched for, and summarized, the economic evaluations of various DHI devices used for managing patients with HF. In this review, we identified 27 studies, including both RCT- and model-based economic evaluations. The findings indicated that the types of DHI devices that were most frequently subjected to an economic evaluation were noninvasive remote monitoring devices (eg, HTM using digital tablets) and medical devices that enabled the transmission of physiological data, followed by telephone support, mobile apps and wearables, remote monitoring follow-up in patients with implantable medical devices (eg, implantable cardioverter defibrillators), and a videoconferencing system. The 27 studies, except for 2 (7%) from Brazil, were conducted in HICs, highlighting the lack of such assessment in LMICs. Despite the diverse range of devices and technologies used for delivering the interventions, the overall results demonstrated that DHIs are potentially more cost-effective than non-DHI alternatives or SoC.

Our findings suggest that HTM via mobile apps and wearables [47-49], home-based telerehabilitation using a videoconferencing system [57], and remote monitoring follow-up in patients with implantable medical devices [50,51,54-56] may be potentially dominant options in managing HF, with less total cost and higher effectiveness. The included studies demonstrate that remote monitoring follow-up in patients with implantable devices resulted in increased coverage of patient services, improved HRQoL [50-54,56], reduced years of life lost, and potentially reduced cost [55,56]. Of the 7 studies in this category, 4 (57%) focused on the economic evaluation of remote monitoring of intracardiac and pulmonary artery pressures in patients with HF via implantable hemodynamic monitoring devices in the United States and the United Kingdom [50-53]. The ICERs ranged from US $13,855 to US $82,782, and all were estimated to be below the respective WTP thresholds in each setting and deemed cost-effective. The subgroup analysis estimated that such a device might be more beneficial in terms of cost-effectiveness in patients with both types of HF: those with reduced ejection fraction and those with preserved ejection fraction [52]. This finding aligns with the updated guidelines recommending the consideration of monitoring pulmonary artery pressures using a wireless hemodynamic monitoring system, particularly for patients with symptoms of HF, to enhance clinical outcomes [20].

Nevertheless, the evidence is limited, especially in mobile apps and wearables as well as home-based telerehabilitation. While mobile apps and wearables are occasionally marketed directly to consumers for health and lifestyle maintenance, of the 27 studies, 3 (11%) focused on assessing HTM through HF-specific apps. The limited evidence and lack of clear app standards pose challenges for decision-makers to make recommendations, although the understanding of the importance of assessment and regulation regarding these DHIs is currently growing. All comparators (3/3, 100%) in these interventions consisted of

XSL•FO
RenderX

SoC, as defined in the guidelines [19,20], which improves the generalizability of the findings.

Furthermore, most base-case findings concerning HTM via noninvasive remote monitoring devices and telephone support indicate that HTM is more costly but more effective than conventional SoC comparators. The observed ICERs ranged from US $8020 [39] to US $106,837 [34] for noninvasive remote monitoring devices and from US $3331 [42] to US $212,586 [44] for telephone support, both per QALYs gained. While most of the studies (9/16, 56%) concluded that the ICERs remained below the WTP thresholds in their respective settings, thus making them cost-effective, certain countries with lower WTP thresholds yielded different findings (3/16, 19%), resulting in conflicting conclusions regarding the cost-effectiveness of these interventions. The WTP thresholds identified in the studies involving noninvasive remote monitoring devices ranged from US $15,372 [31] to US $105,146 [33], and all were derived from studies conducted in HICs. Hence, it is crucial to be cautious when applying these results in wider settings, especially in LMICs with relatively lower WTP-per-QALY threshold levels, because what may be considered a cost-effective intervention in settings with higher WTP thresholds could yield different outcomes in countries with lower WTP thresholds. Other included studies (4/16, 25%) indicated that HTM via noninvasive remote monitoring devices or telephone support can be a dominant strategy [38,39,45,46], with a lower total cost and higher effectiveness than SoC. The cost savings observed in these studies were primarily attributed to reduced hospitalization expenses, especially with regard to the noninvasive devices. This aligns with the primary objective of intervention in HF management because lowering cardiovascular-related hospitalizations and all-cause mortality represents an important clinical end point in most trials assessing HF treatments [60]. The effectiveness of HTM [32] and the cost of HF management [31] are identified as some of the most sensitive parameters that could influence the outcomes of the base-case analysis in model-based studies. Moreover, input parameters associated with hospitalization [33,34,41] and the cost of the intervention [41,44,45] are among the most sensitive variables in RCT-based studies.

In the analysis using noninvasive remote monitoring devices, the distribution of model-based and RCT-based studies is comparable and primarily reflects current updates because the majority (7/9, 78%) were published recently. In the case of telephone support, the economic evaluations are predominantly based on RCTs, with the most recent study dating back to 2018 [41,45,46]. The international guidelines from both the European Society of Cardiology and the American College of Cardiology written in the era before the COVID-19 pandemic do not recommend routine use of remote monitoring or HTM [61,62]. However, during and after the pandemic, the updated version of guidelines for the treatment and management of HF highlights the potential benefits of continuous monitoring of clinical parameters and optimizing care. Despite inconsistencies in its comparative effectiveness and cost-effectiveness, HTM is mentioned as a possible means of monitoring patients [19,20]. Previous evidence indicated that systems that focus on a health maintenance approach through continuous optimization by using DHIs such as noninvasive HTM and telephone support seem to reduce the risk for hospitalizations and all-cause mortality and subsequently improve HRQoL [63,64].

Our findings indicate inconsistencies in the cost-effectiveness of DHIs, which might be attributed to intervention variation. This variation makes it challenging to compare the different DHIs in terms of design, effectiveness, and cost-effectiveness. We stratified the findings by device, which allows a comparison of each technology and provides a better understanding of the cost and effectiveness of adopting DHIs. Economic evaluations of DHIs pose unique challenges compared to those of drugs and medical devices, primarily due to their interacting and evolving features. As observed in this review, most published economic evaluations of DHIs adhere to standard methodological recommendations for evaluating health care technologies, such as pharmaceutical drugs and medical devices. However, there is an argument that these methodological assumptions may not fully reflect the distinct nature of DHIs, which are typically complex interventions composed of multiple interacting components. Consequently, assessing their cost-effectiveness requires a broader evaluation of costs and effects. This evaluation should extend beyond using just 1 outcome measure, such as HRQoL, to include nonhealth benefits and costs beyond health care [65].

In this review, we observed that the incorporation of DHIs is generally associated with improved effectiveness, despite incurring higher total costs. Both short- and long-term time horizons were used in the included studies. The studies demonstrated improved cost-effectiveness of DHIs with a long-term time horizon (≥5 y), indicating the importance of considering a sufficient time horizon to assess the impact of the technology on outcomes. Economic evaluations conducted alongside RCTs tend to use a short time horizon, in line with the timeline of the trials. Determining the time horizon in economic evaluations is crucial because it determines the timing of costs and benefits and how long they should be spread out. When evaluating technology for patients with chronic conditions with long-term potential effects on both cost and health outcomes, assessments with a time horizon of ≤1 year may not consider benefits spread out over extended periods, potentially resulting in an underestimation of its cost-effectiveness. Combining data from RCTs with modeling that allows the projection of costs and effectiveness in the coming years could offer a viable solution to estimate the economic evaluations of DHIs more accurately.

A previous systematic review on HTM or structured telephone support programs for patients with HF suggested that these interventions were considered cost-effective compared to SoC [66]. Similar to these results, according to the included studies in this review, DHIs are generally more cost-effective than standard postdischarge care for managing HF. DHI systems, using infrastructure such as the telephone and the internet, allow patients to access cardiac rehabilitation programs from home and report signs of worsening conditions, regardless of location or the time of day. Such systems also enable remote patient monitoring, reducing the burden on hospitals and health care resources and potentially leading to overall cost savings [67]. The widespread availability of internet and telephone access in

XSL•FO
RenderX

patients' homes, combined with the ease and affordability of implementing remote monitoring systems in clinical practice, make DHIs a potentially cost-effective option [15].

Given the significant clinical and economic burden of HF in LMICs [8], the potential implementation of DHIs in these settings is promising. However, evidence on the cost-effectiveness of DHIs in LMICs is very limited, as observed in this review. To ensure successful implementation, there is a need to test the validity and reliability of DHIs, tailoring their function and design to the specific needs of programs in LMICs, thereby minimizing potential implementation challenges [68]. The transformative potential of digital health in improving health outcomes depends on substantial investment in governance, institutional capacity, and workforce training to navigate the evolving digital landscape of health systems [69]. Comprehensive evidence on the acceptability and cost-effectiveness of DHIs within specific settings in LMICs, including financial considerations, must be integrated into routine health budgets and budgeting processes to assess full-scale sustainability. Consequently, securing sufficient and sustainable financial resources, especially given the financial constraints in LMICs, is crucial. Mobilizing additional resources from development partners is essential in this regard. With strategic investments aligned with national digital health strategies, digital health has the potential to enhance care efficiency and cost-effectiveness, ultimately leading to improved health care service delivery [70].

Strengths and Limitations

The strength of our systematic review is that we assessed various DHIs—both decision-analytic model-based and trial-based economic evaluations of DHIs in managing HF globally—encompassing HTM, rehabilitation, and remote monitoring follow-up after cardiac device implantation. The results of this study may facilitate comparisons and assist policy makers in making informed decisions on how to improve the health outcomes of patients with HF.

Inevitably, our study has some limitations. Due to the variability of the methods, devices, and DHI technologies in the included studies, the comparability of studies is limited. We try to overcome this limitation by using a narrative approach; thus, the variations in methodology and study design can be observed thoroughly. It is important to note that nearly all included studies (25/27, 93%) are from HICs, and caution is warranted when generalizing their results, particularly to LMICs, due to differences in health care systems and resource availability. In addition, although we used a broad definition of DHIs that includes genomics for personalized medicine and artificial intelligence, we did not find any studies related to these concepts. This may be attributed to the existing gaps in clinical and cost-effectiveness evidence [71] when integrating these approaches in the context of HF. Nonetheless, the use of precision medicine, which holds the potential to improve clinical outcomes, represents a promising avenue for the future of precision medicine [72]. In addition, the search strategy used for this systematic review had some constraints. The search terms were constructed using the population, intervention, comparator, and outcomes method, emphasizing a predefined set of terms related to economic evaluations, HF, and DHIs. It is possible that this approach may have overlooked relevant studies that use different keywords. To mitigate this potential gap, we cross-checked the references of the included economic evaluations. Thus, even if we did overlook any, we anticipate that the number will be minimal. Furthermore, considering the variability in the DHIs, modeling approaches, ICER values, and WTP thresholds, it is crucial to perform economic evaluations customized to the specific setting and country. This is especially relevant for LMICs, where the choice of technology, analytical methods, and models should align with the local context.

Conclusions

This review includes 27 studies—model based, RCT based, and combination of both—that focus on economic evaluations of DHIs for patients with HF. The results indicated that noninvasive remote monitoring devices, followed by telephone support, mobile apps and wearables, remote monitoring follow-up in patients with implantable medical devices, and videoconferencing systems are the DHI devices most frequently subjected to economic evaluations in managing HF. Our main findings suggested that adopting DHIs as part of HF treatment and management, in general, requires extra costs but is accompanied by improved health outcomes as measured by HRQoL, compared to SoC, thus seeming to be cost-effective. However, this depends on each country's WTP thresholds for considering cost-effectiveness. The majority of the studies (25/27, 93%) are from HICs, and the findings may not be generalizable to LMICs. Improvement in the quality of reporting, especially in the methodology of further economic evaluations, would better inform the cost- and health-related outcomes of incorporating DHIs for patients with HF.

Acknowledgments

This study was supported by a grant from Universitas Padjadjaran, Bandung, Indonesia.

Authors' Contributions

This study was conceptualized by NZ, IMP, and KL. The protocol was developed by NZ, DM, and IMP. The search and data extraction were carried out by NZ and DM. Articles were screened for inclusion by NZ, DM, and IMP. The analysis was conducted by NZ, DM, and MA. NZ and JFR were responsible for supervision. NZ and DM wrote the original draft. All authors were responsible for reviewing and editing the final draft.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 34 KB - jmir_v26i1e53500_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 13 KB - jmir_v26i1e53500_app2.docx]

Multimedia Appendix 3
Perspective, costs, and sensitivity analysis.
[DOCX File , 274 KB - jmir_v26i1e53500_app3.docx]

References

1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GM, Coats AJ. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res 2023 Jan 18;118(17):3272-3287. [doi: 10.1093/cvr/cvac013] [Medline: 35150240]
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
3. Dokainish H, Teo K, Zhu J, Roy A, AlHabib KF, ElSayed A, et al. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. Lancet Glob Health 2017 Jul;5(7):e665-e672 [FREE Full text] [doi: 10.1016/S2214-109X(17)30196-1] [Medline: 28476564]
4. Lesman-Leegte I, Jaarsma T, Coyne JC, Hillege HL, van Veldhuisen DJ, Sanderman R. Quality of life and depressive symptoms in the elderly: a comparison between patients with heart failure and age- and gender-matched community controls. J Card Fail 2009 Feb;15(1):17-23. [doi: 10.1016/j.cardfail.2008.09.006] [Medline: 19181289]
5. Hessel FP. Overview of the socio-economic consequences of heart failure. Cardiovasc Diagn Ther 2021 Feb;11(1):254-262 [FREE Full text] [doi: 10.21037/cdt-20-291] [Medline: 33708497]
6. Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail 2013 May;6(3):606-619 [FREE Full text] [doi: 10.1161/HHF.0b013e318291329a] [Medline: 23616602]
7. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics-2021 update: a report from the American Heart Association. Circulation 2021 Feb 23;143(8):e254-e743 [FREE Full text] [doi: 10.1161/CIR.0000000000000950] [Medline: 33501848]
8. Cook C, Cole G, Asaria P, Jabbour R, Francis DP. The annual global economic burden of heart failure. Int J Cardiol 2014 Feb 15;171(3):368-376. [doi: 10.1016/j.ijcard.2013.12.028] [Medline: 24398230]
9. Bloom DE, Cafiero ET, Jané-Llopis E, Abrahams-Gessel S, Bloom LR, Fathima S, et al. The global economic burden of non-communicable diseases. World Economic Forum. 2011 Sep. URL: https://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf [accessed 2024-04-12]
10. Chun S, Tu JV, Wijeysundera HC, Austin PC, Wang X, Levy D, et al. Lifetime analysis of hospitalizations and survival of patients newly admitted with heart failure. Circ Heart Fail 2012 Jul 01;5(4):414-421 [FREE Full text] [doi: 10.1161/CIRCHEARTFAILURE.111.964791] [Medline: 22556322]
11. Desai AS, Stevenson LW. Rehospitalization for heart failure: predict or prevent? Circulation 2012 Jul 24;126(4):501-506. [doi: 10.1161/circulationaha.112.125435]
12. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. Eur Heart J 2016 May 07;37(18):1428-1438 [FREE Full text] [doi: 10.1093/eurheartj/ehv770] [Medline: 26873093]
13. Singhal A, Cowie MR. Digital health: implications for heart failure management. Card Fail Rev 2021 Mar;7:e08 [FREE Full text] [doi: 10.15420/cfr.2020.28] [Medline: 34035953]
14. What is digital health? U.S. Food & Drug Administration. 2020 Sep 22. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2023-02-07]
15. WHO Global Observatory for eHealth. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://apps.who.int/iris/handle/10665/44607 [accessed 2024-04-12]
16. Harnett B. Telemedicine systems and telecommunications. J Telemed Telecare 2006 Jun 23;12(1):4-15. [doi: 10.1258/135763306775321416] [Medline: 16438772]

XSL•FO
RenderX

17. WHO guideline: recommendations on digital interventions for health system strengthening: evidence and recommendations. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/311980 [accessed 2024-04-12]
18. Brahmbhatt DH, Cowie MR. Remote management of heart failure: an overview of telemonitoring technologies. Card Fail Rev 2019 May 24;5(2):86-92 [FREE Full text] [doi: 10.15420/cfr.2019.5.3] [Medline: 31179018]
19. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. Circulation 2022 May 03;145(18):e895-1032 [FREE Full text] [doi: 10.1161/CIR.0000000000001063] [Medline: 35363499]
20. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021 Sep 21;42(36):3599-3726 [FREE Full text] [doi: 10.1093/eurheartj/ehab368] [Medline: 34447992]
21. Farabi H, Rezapour A, Jahangiri R, Jafari A, Rashki Kemmak A, Nikjoo S. Economic evaluation of the utilization of telemedicine for patients with cardiovascular disease: a systematic review. Heart Fail Rev 2020 Nov 23;25(6):1063-1075. [doi: 10.1007/s10741-019-09864-4] [Medline: 31760593]
22. Jiang X, Ming WK, You JH. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019 Jun 17;21(6):e13166 [FREE Full text] [doi: 10.2196/13166] [Medline: 31210136]
23. Scherrenberg M, Falter M, Dendale P. Cost-effectiveness of cardiac telerehabilitation in coronary artery disease and heart failure patients: systematic review of randomized controlled trials. Eur Heart J Digit Health 2020 Nov;1(1):20-29 [FREE Full text] [doi: 10.1093/ehjdh/ztaa005] [Medline: 37056294]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Marulin D, Puspitasari IM, Zakiyah N. Economic evaluation of digital health interventions for patients with heart failure: a systematic review. PROSPERO. 2023. URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=388241 [accessed 2023-02-10]
26. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019 Jul 6. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2024-04-12]
27. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. Value Health 2022 Jan;25(1):3-9 [FREE Full text] [doi: 10.1016/j.jval.2021.11.1351] [Medline: 35031096]
28. Zakiyah N, van Asselt AD, Roijmans F, Postma MJ. Economic evaluation of family planning interventions in low and middle income countries; a systematic review. PLoS One 2016 Dec 19;11(12):e0168447 [FREE Full text] [doi: 10.1371/journal.pone.0168447] [Medline: 27992552]
29. CCEMG - EPPI-Centre cost converter. EPPI Centre. URL: https://eppi.ioe.ac.uk/costconversion [accessed 2022-04-10]
30. Nixon J, Khan KS, Kleijnen J. Summarising economic evaluations in systematic reviews: a new approach. BMJ 2001 Jun 30;322(7302):1596-1598 [FREE Full text] [doi: 10.1136/bmj.322.7302.1596] [Medline: 11431306]
31. Caillon M, Sabatier R, Legallois D, Courouve L, Donio V, Boudevin F, et al. A telemonitoring programme in patients with heart failure in France: a cost-utility analysis. BMC Cardiovasc Disord 2022 Oct 10;22(1):441 [FREE Full text] [doi: 10.1186/s12872-022-02878-1] [Medline: 36217130]
32. Thokala P, Baalbaki H, Brennan A, Pandor A, Stevens JW, Gomersall T, et al. Telemonitoring after discharge from hospital with heart failure: cost-effectiveness modelling of alternative service designs. BMJ Open 2013 Sep 18;3(9):e003250 [FREE Full text] [doi: 10.1136/bmjopen-2013-003250] [Medline: 24048626]
33. Albuquerque de Almeida F, Corro Ramos I, Al M, Rutten-van Mölken M. Home telemonitoring and a diagnostic algorithm in the management of heart failure in the Netherlands: cost-effectiveness analysis. JMIR Cardio 2022 Aug 04;6(2):e31302 [FREE Full text] [doi: 10.2196/31302] [Medline: 35925670]
34. Jiang X, Yao J, You JH. Telemonitoring versus usual care for elderly patients with heart failure discharged from the hospital in the United States: cost-effectiveness analysis. JMIR Mhealth Uhealth 2020 Jul 06;8(7):e17846 [FREE Full text] [doi: 10.2196/17846] [Medline: 32407288]
35. Thokala P, Dodd P, Baalbaki H, Brennan A, Dixon S, Lowrie K. Developing Markov models from real-world data: a case study of heart failure modeling using administrative data. Value Health 2020 Jun;23(6):743-750 [FREE Full text] [doi: 10.1016/j.jval.2020.02.012] [Medline: 32540232]
36. Völler H, Bindl D, Nagels K, Hofmann R, Vettorazzi E, Wegscheider K, et al. The first year of noninvasive remote telemonitoring in chronic heart failure is not cost saving but improves quality of life: the randomized controlled CardioBBEAT trial. Telemed J E Health 2022 Mar 21;28(11):1613-1622 [FREE Full text] [doi: 10.1089/tmj.2022.0021] [Medline: 35325562]
37. Boyne JJ, di van Asselt A, Gorgels AP, Steuten LM, de Weerd G, Kragten J, et al. Cost–effectiveness analysis of telemonitoring versus usual care in patients with heart failure: the TEHAF–study. J Telemed Telecare 2016 Jul 09;19(5):242-248. [doi: 10.1177/1357633x13495478]

XSL•FO
RenderX

38. Sydow H, Prescher S, Koehler F, Koehler K, Dorenkamp M, Spethmann S, et al. Cost-effectiveness of noninvasive telemedical interventional management in patients with heart failure: health economic analysis of the TIM-HF2 trial. Clin Res Cardiol 2022 Nov 11;111(11):1231-1244 [FREE Full text] [doi: 10.1007/s00392-021-01980-2] [Medline: 34894273]
39. Vestergaard AS, Hansen L, Sørensen SS, Jensen MB, Ehlers LH. Is telehealthcare for heart failure patients cost-effective? An economic evaluation alongside the Danish TeleCare North heart failure trial. BMJ Open 2020 Jan 27;10(1):e031670 [FREE Full text] [doi: 10.1136/bmjopen-2019-031670] [Medline: 31992604]
40. Grustam AS, Severens JL, de Massari D, Buyukkaramikli N, Koymans R, Vrijhoef HJ. Cost-effectiveness analysis in telehealth: a comparison between home telemonitoring, nurse telephone support, and usual care in chronic heart failure management. Value Health 2018 Jul;21(7):772-782 [FREE Full text] [doi: 10.1016/j.jval.2017.11.011] [Medline: 30005749]
41. Ruschel KB, Rabelo-Silva ER, Rohde LE, de Souza EN, Mussi CM, Polanczyk CA. Cost-effectiveness of a home visit program for patients with heart failure in Brazil: evidence from a randomized clinical trial. Value Health Reg Issues 2018 Dec;17:81-87. [doi: 10.1016/j.vhri.2018.03.006] [Medline: 29754015]
42. Cui Y, Doupe M, Katz A, Nyhof P, Forget EL. Economic evaluation of Manitoba Health Lines in the management of congestive heart failure. Healthc Policy 2013 Nov;9(2):36-50 [FREE Full text] [Medline: 24359716]
43. Hebert PL, Sisk JE, Wang JJ, Tuzzio L, Casabianca JM, Chassin MR, et al. Cost-effectiveness of nurse-led disease management for heart failure in an ethnically diverse urban community. Ann Intern Med 2008 Oct 21;149(8):540-548 [FREE Full text] [doi: 10.7326/0003-4819-149-8-200810210-00006] [Medline: 18936502]
44. Smith B, Hughes-Cromwick PF, Forkner E, Galbreath AD. Cost-effectiveness of telephonic disease management in heart failure. Am J Manag Care 2008 Feb;14(2):106-115 [FREE Full text] [Medline: 18269306]
45. Bocchi EA, da Cruz FD, BrandÃo SM, Issa V, Ayub-Ferreira SM, Brunner la Rocca HP, et al. Cost-effectiveness benefits of a disease management program: the REMADHE trial results. J Card Fail 2018 Oct;24(10):627-637. [doi: 10.1016/j.cardfail.2018.04.008] [Medline: 29723595]
46. González-Guerrero JL, Hernández-Mocholi MA, Ribera-Casado JM, García-Mayolín N, Alonso-Fernández T, Gusi N. Cost-effectiveness of a follow-up program for older patients with heart failure: a randomized controlled trial. Eur Geriatr Med 2018 Aug 06;9(4):523-532. [doi: 10.1007/s41999-018-0074-y] [Medline: 34674493]
47. Jiang X, Yao J, You JH. Cost-effectiveness of a telemonitoring program for patients with heart failure during the COVID-19 pandemic in Hong Kong: model development and data analysis. J Med Internet Res 2021 Mar 03;23(3):e26516 [FREE Full text] [doi: 10.2196/26516] [Medline: 33656440]
48. Boodoo C, Zhang Q, Ross HJ, Alba AC, Laporte A, Seto E. Evaluation of a heart failure telemonitoring program through a microsimulation model: cost-utility analysis. J Med Internet Res 2020 Oct 06;22(10):e18917 [FREE Full text] [doi: 10.2196/18917] [Medline: 33021485]
49. Cano Martín JA, Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Economic impact assessment from the use of a mobile app for the self-management of heart diseases by patients with heart failure in a Spanish region. J Med Syst 2014 Sep;38(9):96-97 [FREE Full text] [doi: 10.1007/s10916-014-0096-z] [Medline: 24994514]
50. Cowie MR, Simon M, Klein L, Thokala P. The cost-effectiveness of real-time pulmonary artery pressure monitoring in heart failure patients: a European perspective. Eur J Heart Fail 2017 May 07;19(5):661-669 [FREE Full text] [doi: 10.1002/ejhf.747] [Medline: 28176424]
51. Schmier JK, Ong KL, Fonarow GC. Cost-effectiveness of remote cardiac monitoring with the CardioMEMS heart failure system. Clin Cardiol 2017 Jul 08;40(7):430-436 [FREE Full text] [doi: 10.1002/clc.22696] [Medline: 28272808]
52. Sandhu AT, Goldhaber-Fiebert JD, Owens DK, Turakhia MP, Kaiser DW, Heidenreich PA. Cost-effectiveness of implantable pulmonary artery pressure monitoring in chronic heart failure. JACC Heart Fail 2016 May;4(5):368-375 [FREE Full text] [doi: 10.1016/j.jchf.2015.12.015] [Medline: 26874380]
53. Martinson M, Bharmi R, Dalal N, Abraham WT, Adamson PB. Pulmonary artery pressure-guided heart failure management: US cost-effectiveness analyses using the results of the CHAMPION clinical trial. Eur J Heart Fail 2017 May;19(5):652-660 [FREE Full text] [doi: 10.1002/ejhf.642] [Medline: 27647784]
54. Niewada M, Tabor B, Piotrowicz E, Piotrowicz R, Opolski G, Banach M, et al. Cost-effectiveness of telerehabilitation in patients with heart failure in Poland: an analysis based on the results of Telerehabilitation in the Heart Failure Patients (TELEREH-HF) randomized clinical trial. Kardiol Pol 2021;79(5):510-516 [FREE Full text] [doi: 10.33963/KP.15885] [Medline: 34125923]
55. Calò L, Gargaro A, de Ruvo E, Palozzi G, Sciarra L, Rebecchi M, et al. Economic impact of remote monitoring on ordinary follow-up of implantable cardioverter defibrillators as compared with conventional in-hospital visits. A single-center prospective and randomized study. J Interv Card Electrophysiol 2013 Jun 21;37(1):69-78. [doi: 10.1007/s10840-013-9783-9] [Medline: 23515883]
56. Zanaboni P, Landolina M, Marzegalli M, Lunati M, Perego GB, Guenzati G, et al. Cost-utility analysis of the EVOLVO study on remote monitoring for heart failure patients with implantable defibrillators: randomized controlled trial. J Med Internet Res 2013 May 30;15(5):e106 [FREE Full text] [doi: 10.2196/jmir.2587] [Medline: 23722666]
57. Hwang R, Morris NR, Mandrusiak A, Bruning J, Peters R, Korczyk D, et al. Cost-utility analysis of home-based telerehabilitation compared with centre-based rehabilitation in patients with heart failure. Heart Lung Circ 2019 Dec;28(12):1795-1803. [doi: 10.1016/j.hlc.2018.11.010] [Medline: 30528811]

XSL•FO
RenderX

58. Burri H, Sticherling C, Wright D, Makino K, Smala A, Tilden D. Cost-consequence analysis of daily continuous remote monitoring of implantable cardiac defibrillator and resynchronization devices in the UK. Europace 2013 Nov;15(11):1601-1608 [FREE Full text] [doi: 10.1093/europace/eut070] [Medline: 23599169]
59. Abraham WT, Adamson PB, Bourge RC, Aaron MF, Costanzo MR, Stevenson LW, et al. Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial. The Lancet 2011 Feb;377(9766):658-666. [doi: 10.1016/s0140-6736(11)60101-3]
60. Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Kirwan BA, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. The Lancet 2018 Sep;392(10152):1047-1057. [doi: 10.1016/s0140-6736(18)31880-4]
61. Ponikowski P, Voors A, Anker S, Bueno H, Cleland J, Coats A, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 2016 Jul 14;37(27):2129-2200 [FREE Full text] [doi: 10.1093/eurheartj/ehw128] [Medline: 27206819]
62. -. Correction to: 2016 ACC/AHA/HFSA focused update on new pharmacological therapy for heart failure: an update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on clinical practice guidelines and the Heart Failure Society of America. Circulation 2016 Sep 27;134(13):e298. [doi: 10.1161/CIR.0000000000000460]
63. Inglis SC, Clark RA, Dierckx R, Prieto-Merino D, Cleland JG. Structured telephone support or non-invasive telemonitoring for patients with heart failure. Cochrane Database Syst Rev 2015 Oct 31;2015(10):CD007228 [FREE Full text] [doi: 10.1002/14651858.CD007228.pub3] [Medline: 26517969]
64. Bashi N, Karunanithi M, Fatehi F, Ding H, Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews. J Med Internet Res 2017 Jan 20;19(1):e18 [FREE Full text] [doi: 10.2196/jmir.6571] [Medline: 28108430]
65. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: methodological issues and recommendations for practice. Pharmacoeconomics 2022 Apr 08;40(4):367-378 [FREE Full text] [doi: 10.1007/s40273-022-01130-0] [Medline: 35132606]
66. Pandor A, Thokala P, Gomersall T, Baalbaki H, Stevens JW, Wang J, et al. Home telemonitoring or structured telephone support programmes after recent discharge in patients with heart failure: systematic review and economic evaluation. Health Technol Assess 2013 Aug;17(32):1-207, v [FREE Full text] [doi: 10.3310/hta17320] [Medline: 23927840]
67. Seto E. Cost comparison between telemonitoring and usual care of heart failure: a systematic review. Telemed J E Health 2008 Sep;14(7):679-686. [doi: 10.1089/tmj.2007.0114] [Medline: 18817497]
68. Duggal M, El Ayadi A, Duggal B, Reynolds N, Bascaran C. Editorial: challenges in implementing digital health in public health settings in low and middle income countries. Front Public Health 2022 Jan 10;10:1090303 [FREE Full text] [doi: 10.3389/fpubh.2022.1090303] [Medline: 36703825]
69. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-04-12]
70. Kaboré SS, Ngangue P, Soubeiga D, Barro A, Pilabré AH, Bationo N, et al. Barriers and facilitators for the sustainability of digital health interventions in low and middle-income countries: a systematic review. Front Digit Health 2022 Nov 28;4:1014375 [FREE Full text] [doi: 10.3389/fdgth.2022.1014375] [Medline: 36518563]
71. Bombard Y, Ginsburg GS, Sturm AC, Zhou AY, Lemke AA. Digital health-enabled genomics: opportunities and challenges. Am J Hum Genet 2022 Jul 07;109(7):1190-1198 [FREE Full text] [doi: 10.1016/j.ajhg.2022.05.001] [Medline: 35803232]
72. Weldy CS, Ashley EA. Towards precision medicine in heart failure. Nat Rev Cardiol 2021 Nov 09;18(11):745-762. [doi: 10.1038/s41569-021-00566-9] [Medline: 34108678]

Abbreviations

CHEERS: Consolidated Health Economic Evaluation Reporting Standards
CVD: cardiovascular disease
DHI: digital health intervention
HF: heart failure
HIC: high-income country
HRQoL: health-related quality of life
HTM: home telemonitoring
ICER: incremental cost-effectiveness ratio
LMIC: low- and middle-income country
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QALY: quality-adjusted life year
RCT: randomized controlled trial
SoC: standard of care

XSL•FO
RenderX

WTP: willingness-to-pay

Edited by S Ma, T Leung; submitted 24.10.23; peer-reviewed by D Setiawan, T Gal; comments to author 18.03.24; revised version received 26.03.24; accepted 26.03.24; published 30.04.24.

Please cite as:
Zakiyah N, Marulin D, Alfaqeeh M, Puspitasari IM, Lestari K, Lim KK, Fox-Rushby J
Economic Evaluations of Digital Health Interventions for Patients With Heart Failure: Systematic Review
J Med Internet Res 2024;26:e53500
URL: https://www.jmir.org/2024/1/e53500
doi:10.2196/53500
PMID:38687991

©Neily Zakiyah, Dita Marulin, Mohammed Alfaqeeh, Irma Melyani Puspitasari, Keri Lestari, Ka Keat Lim, Julia Fox-Rushby. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Information and Communications Technology–Based Monitoring Service for Tailored Chronic Disease Management in Primary Care: Cost-Effectiveness Analysis Based on ICT-CM Trial Results

Sung-Hee Oh[1], MPH, PhD; Jae-Heon Kang[2*], MD, PhD; Jin-Won Kwon[1*], MPH, PhD

[1]Brain Korea 21 Four Community-Based Intelligent Novel Drug Discovery Education Unit, College of Pharmacy and Research Institute of Pharmaceutical Sciences, Kyungpook National University, Daegu, Republic of Korea

[2]Department of Family Medicine, Kangbuk Samsung Hospital, College of Medicine, Sungkyunkwan University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Jin-Won Kwon, MPH, PhD
Brain Korea 21 Four Community-Based Intelligent Novel Drug Discovery Education Unit, College of Pharmacy and Research Institute of Pharmaceutical Sciences
Kyungpook National University
80, Daehak-ro, Buk-gu
Daegu, 41566
Republic of Korea
Phone: 82 53 950 8580
Fax: 82 53 950 8557
Email: jwkwon@knu.ac.kr

Abstract

Background: Information and communications technology–based tailored management (TM) intervention is a novel automatic system in which a smartphone app for the management of patients with hypertension and diabetes, the provider web, and Bluetooth devices are linked. However, little evidence exists regarding the cost-effectiveness of the interventions using mobile apps.

Objective: This study aimed to assess the cost-effectiveness of TM intervention for adult patients with hypertension or diabetes in primary care compared with usual care (UC).

Methods: Cost-effectiveness analysis using a Markov model was conducted from the Korean health care system perspective. Based on 6-month outcome data from an information and communications technology–based tailored chronic disease management (ICT-CM) trial, effectiveness over a lifetime beyond the trial periods was extrapolated using a cardiovascular disease risk prediction model. Costs were estimated using ICT-CM trial data and national health insurance claims data. Health utility weights were obtained from the Korea National Health and Nutrition Examination Survey.

Results: In the base-case analysis, compared with UC, TM was more costly (US $23,157 for TM vs US $22,391 for UC) and more effective (12.006 quality-adjusted life-years [QALYs] for TM vs 11.868 QALYs for UC). The incremental cost-effectiveness ratio was US $5556 per QALY gained. Probabilistic sensitivity analysis showed that the probability of TM being cost-effective compared with UC was approximately 97% at an incremental cost-effectiveness ratio threshold of US $26,515 (KRW 35 million) per QALY gained.

Conclusions: Compared with UC, TM intervention is a cost-effective option for patients with hypertension or diabetes in primary care settings. The study results can assist policy makers in making evidence-based decisions when implementing accessible chronic disease management services.

(J Med Internet Res 2024;26:e51239) doi:10.2196/51239

KEYWORDS

mHealth; mobile health; smartphone application; cost-effectiveness analysis; hypertension; diabetes mellitus; primary care; artificial intelligence; applications digital health; mobile phone

XSL•FO
RenderX

Introduction

Hypertension and diabetes are major risk factors for cardiovascular diseases (CVD) associated with mortality and morbidity, imposing huge economic burdens [1,2]. Adequate blood pressure (BP) and glycemic control are important in health promotion and health care systems worldwide [3,4]. However, despite significant advances in evidence-based lifestyle modifications and pharmaceutical interventions, less than 40% of patients treated for hypertension or diabetes achieve the recommended target BP or blood glucose levels in Korea [5]. Therefore, novel interventions are needed to support patient efforts for behavior changes to promote healthy lifestyles and disease self-management [6,7]. Mobile health (mHealth) based on information and communication technology in primary care is an innovative approach to such interventions [8]. The widespread use of mobile phones combined with the ability to process and communicate data instantly enables real time individually tailored health care delivery and overcomes barriers of time and place [9].

A national pilot project for chronic disease management was recently conducted under the supervision of the Ministry of Health and Welfare to evaluate the effectiveness of information and communications technology (ICT)–based monitoring service for tailored chronic disease management in primary care in Korea [10,11]. The ICT-based tailored management (TM), using automated advanced systems for patients with hypertension and diabetes, provides a continuous and integrated customized health management service by linking the mHealth service platform (smartphone app) and the provider's (primary care physician and care coordinator) operation web combined with the patient's health information (examination and drug prescription information data from the National Health Insurance Service). Several mHealth interventions (eg, mobile phone SMS text messages, wearable monitoring devices, and telemedicine) for managing hypertension and diabetes have demonstrated efficacy [12], but evidence for interventions using smartphone apps is limited. Furthermore, their cost-effectiveness varied substantially based on the target disease and type of technology, and the results of economic evaluations have been inconsistent [13]. In particular, the cost-effectiveness of interventions using complex smartphone communications in older people cannot be judged due to a lack of information [14].

To our knowledge, no studies have examined the long-term cost-effectiveness of tailored management interventions for patients with hypertension or diabetes using automated advanced mobile technology in primary care. Although a large pilot project on TM intervention based on the highest smartphone penetration rate globally and advanced technology was conducted [15], it remains unclear whether the additional benefits from TM are sufficient to justify its use over and above usual care (UC) alone in terms of economic efficiency. This study aimed to estimate the cost-effectiveness of TM for adult patients older than 19 years of age with hypertension or diabetes in primary care compared with UC, using data collected in a pragmatic trial of ICT-based tailored chronic disease management (ICT-CM).

Methods

The ICT-Based Tailored Chronic Disease Management Trial

The ICT-CM was a pragmatic trial designed to test the real-world effectiveness of the ICT-based monitoring program among patients with chronic diseases in primary care [16]. Study participants were patients aged 19 years or older diagnosed with hypertension or diabetes, recruited from 8 clinics in Seoul and Gyeonggi-do, South Korea. Patients with myocardial infarction, stroke, end-stage renal disease, or liver failure within 1 year were excluded. Among the participants (n=1004) in the ICT-CM trial, those aged between 60 and <70 accounted for the highest proportion at 39% (392/1004), while those younger than 60 years accounted for 36.8% (369/1004), and those aged 70 and older accounted for 24.2% (243/1004). The average age was 62.4 years.

The study included a TM intervention group and a UC control group. Participants who received UC for hypertension and diabetes were treated at primary care institutions according to Korean clinical practice guidelines, and UC encompasses continuous monitoring and follow-up, appropriate medication management based on the patient's condition, and lifestyle modifications to promote healthy habits [17,18]. Participants who received TM intervention were treated using automated advanced mHealth systems by physicians and care coordinators. A care coordinator trained participants who received TM on how to use the technology, downloaded the mobile application Carecrew (Huraypositive Inc) available for both Android (Google) and iOS (Apple Inc) onto their own smartphones, and were given Bluetooth-enabled devices (sphygmomanometer and glucometer). Physicians established personalized management goals for each patient (referred to as a tailored care plan) based on the lifestyle survey results and clinical examination at baseline. The care plan, including each participant's target BP or blood glucose level, daily measurement frequency, and management priorities, was automatically sent to the Carecrew. The participants were asked to record their lifestyle such as diet and body weight and periodically check their blood glucose, BP, and medication use. Results measured by the devices were automatically uploaded to the provider's operation web (CareCrew Web by Huraypositive Inc). The physicians and care coordinators had access to this system to monitor the patient's condition constantly. Patients who missed self-recording were encouraged to self-measure and input lifestyle data through SMS text messages and phone calls. When the system reported out-of-range BP or glucose readings, a tailored mobile SMS text message was sent to the patient, and mobile-based feedback was performed for high-risk patients requiring intensive care. A customized examination voucher was issued for each patient and assessments were conducted at the clinic visit at 6 months. Source data from ICT-CM trial were verified through periodic on-site monitoring. Monitors identified any discrepancies between all case report form entries and the source data and issued queries. The queries were resolved by investigators, cooperating with monitors. The ICT-CM trial is described in detail in Multimedia Appendix 1.

XSL•FO
RenderX

Ethical Considerations

The study was approved by the Investigation Review Board of Kangbuk Samsung Hospital (IRB KBSMC 2020-07-026-012). Informed consent and the ability of participants to withdraw were provided in the ICT-CM trial used for source data in this study. Expenses, including intervention costs and examination fees incurred due to study participation, were covered by the study, but no financial compensation was paid. All data used in this study were anonymized and deidentified.

Model Overview

Long-term costs and health outcomes over the lifetime were compared for the TM intervention arm and UC arm. A target cohort population comprising Korean patients with hypertension or diabetes diagnosed in primary care, which was consistent with the populations examined in the ICT-CM trial. The Markov model was designed to mirror the natural progression of conditions in the population and the clinical pathway based on clinical guidelines. This model reflects the short-term outcomes of treatment (controlled-uncontrolled status of BP or blood glucose) in patients with hypertension or diabetes, their characteristics that were considered CVD risk factors, and the risk of long-term complications that patients may experience. The short-term outcomes were as observed in the ICT-CM trial at 6 months, and subsequent long-term outcomes were extrapolated from the trial data. The model was programmed in Microsoft Excel, supporting macro programming through Visual Basic for Applications. This study is reported according to the Consolidated Health Economic Evaluation Reporting Standards checklist [19].

Model Structure

The structure of the Markov model is presented in Figure 1. The model structure's validity was checked by reviewing the literature and expert opinion [20,21]. A Markov cohort could experience 1 of several health states, including "Well" with no major complications (ie, CVD), "CVD," "Death due to CVD," or "Death due to other cause." "CVD" was defined as a composite outcome of coronary heart disease (ie, myocardial infarction and angina), cerebrovascular diseases (ie, cerebral infarction, hemorrhagic stroke, and transient ischemic attack), heart failure, and peripheral artery disease. All patients in the TM or UC groups started the Markov process in the "Well" state. Patients could remain in the "Well" state or transfer to another state at every cycle length of 6 months, corresponding to the frequency of assessments in the ICT-CM study. From any health state, patients could transit to the "Death due to other cause" state, and from the "CVD" state, patients could transit to the "Death due to CVD" state. Because the age of the patient cohorts on entry into the model was 60 years, a treatment discontinuation rate of 5% was assumed, considering that TM intervention could be discontinued in the real-world setting. Patients who discontinued TM intervention continued with UC. Patients receiving the UC at the model start were assumed to remain on the treatment until the end of the time horizon or death, whichever occurred first.

XSL•FO
RenderX

Figure 1. Model structure. (A) Model framework. The short-term outcomes were as observed in the ICT-CM trial at 6 months, with subsequent long-term CVD events with extrapolation from the trial data. (B) State transition diagram. The CVD was defined as a composite outcome of coronary heart disease (ie, myocardial infarction and angina), cerebrovascular diseases (ie, cerebral infarction, hemorrhagic stroke, and transient ischemic attack), heart failure, and peripheral artery disease. BP: blood pressure; CVD: cardiovascular diseases; HbA1c: glycated hemoglobin A1c; HDL-C: high-density lipoprotein cholesterol; ICT: information and communications technology; ICT-CM: information and communications technology–based monitoring service for tailored chronic disease management.

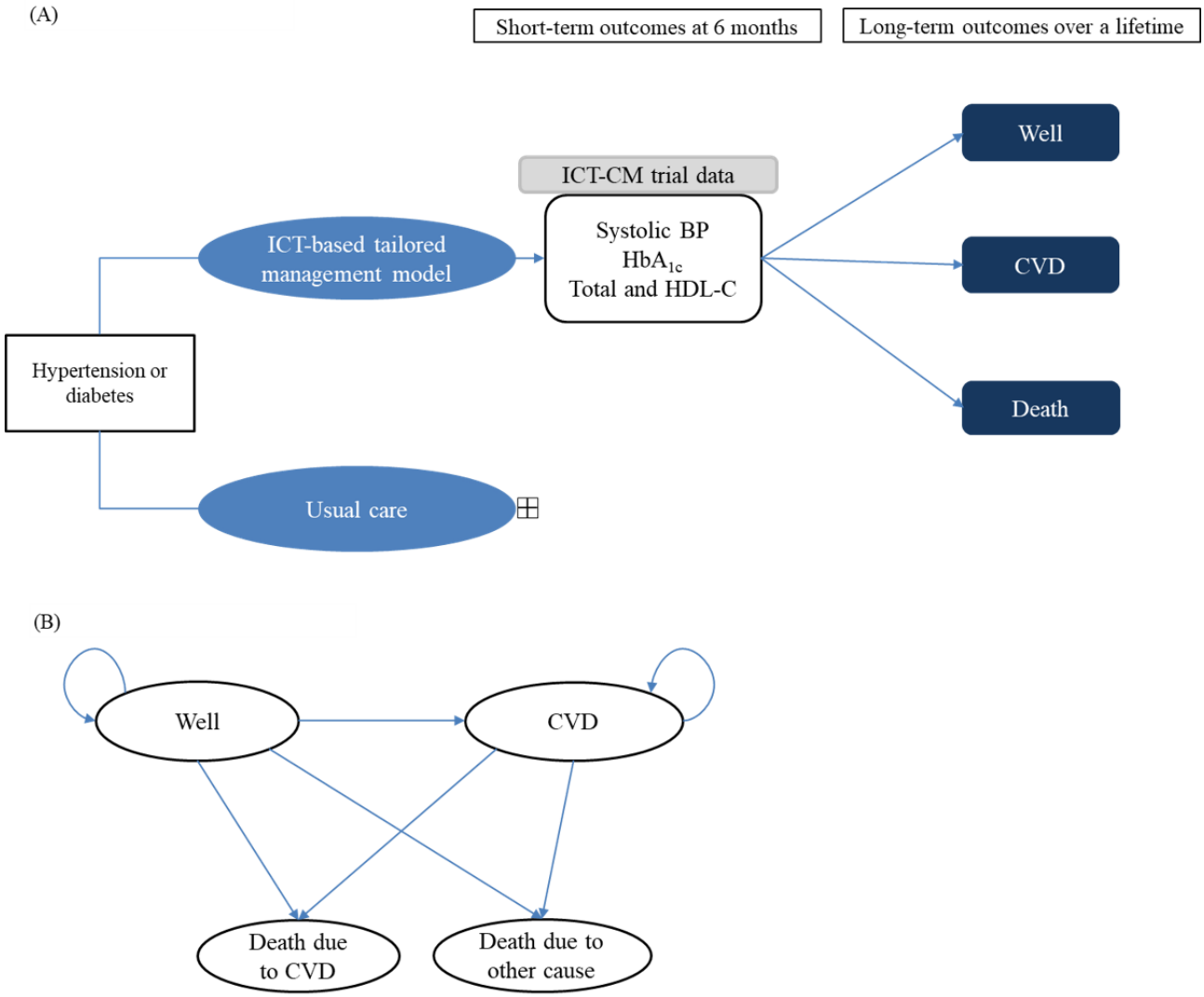

Clinical Outcomes

The outcomes derived from the ICT-CM trial used in the study are presented in Table 1. Through post hoc analysis using individual patient data from the ICT-CM trial, the CVD risk factor outcomes were measured for each treatment arm. Demographic information including age and medical history were examined in screening tests, and smoking history was assessed in health surveys using a complete self-health questionnaire at both baseline and 6 months. Total cholesterol, high-density lipoprotein-cholesterol, and glycated hemoglobin A_{1c} (HbA_{1c}) levels were assessed during clinical examinations at both baseline and 6 months, while BP was measured through body assessments at the same intervals. Details regarding the measurements and the outcomes assessed in the analysis are described in Multimedia Appendix 1. Using the input values associated with CVD risk factors obtained from the trial, the 10-year cardiovascular risk for each patient was calculated using the Framingham CVD risk prediction model [22]. The average risk estimate for each treatment arm was then converted into an annual probability of CVD events. The risk of CVD increased with the age of the patient cohort based on the age-related relative risk of CVD obtained from Kim et al [23]. The base-case regarding the long-term effect persistence assumed that the 6-month difference in BP or blood glucose between TM and UC decreased by 10% after 5 years, reflecting that the effectiveness may decline with age.

XSL•FO
RenderX

Table 1. Information and communications technology–based tailored chronic disease management trial participants' data used in the Framingham risk model.

Risk factor outcomes	All patients with hypertension or diabetes		Patients with hypertension		Patients with diabetes	
	TM[a]	UC[b]	TM	UC	TM	UC
Male, %	52.8	50.5	49.5	46.1	57.0	54.9
Change from baseline to 6 months in systolic BP[c], mean (mm Hg)	–1.6	4.3	4.9	3.3	2.7	5.4
Change from baseline to 6 months in diastolic BP, mean, mm Hg	–0.9	3.2	–2.9	3.9	1.6	2.3
Antihypertensive medication, %	55.6	55.5	100	99.4	66.4	63.2
Change from baseline to 6 months in total cholesterol, mean (mg/dL)	–2.4	–1.1	–5.7	–1.5	1.8	–0.7
Change from baseline to 6 months in HDL[d]-cholesterol, mean (mg/dL)	1.3	1.8	1.1	1.2	1.6	2.4
Change from baseline to 6 months in HbA_{1c}[e], mean (%)	–0.1	–0.1	–0.1	–0.0	0.0	–0.1
Current smoker, %	18.5	13.7	15.0	12.9	23.2	17.9

[a]TM: information and communications technology–based tailored management.

[b]UC: usual care.

[c]BP: blood pressure.

[d]HDL: high-density lipoprotein.

[e]HbA_{1c}: glycated hemoglobin A_{1c}.

All-cause and CVD-related mortality of the general population were based on cause-of-death statistics by age of the Korean Statistical Office [24]. The annual probability of death due to other-cause was calculated by excluding death due to CVD from all-cause death. The fatality rate for CVD by age was estimated as the number of deaths due to CVD among patients with CVD, using the cause-of-death statistics of the Korean Statistical Office [24] and the Korean National Health Insurance statistics by the Health Insurance Review and Assessment Service [25]. Table 2 presents a detailed description of the model inputs.

XSL•FO
RenderX

Table 2. Model inputs for patients with hypertension or diabetes.

Parameters	Estimate in base-case	Distribution[a]	Sources
Analysis setting			
Start age (year)	60	—[b]	ICT-CM[c] trial population
Time horizon (year)	Lifetime	—	Assumed
Mortality and risk of CVD[d]			
Probability of death due to CVD, %			Cause-of-death statistics of Korean Statistical Office; Korean national health insurance statistics by HIRA[e]
60-69 years	0.72	Beta	
70-79 years	1.36	Beta	
≥80 years	5.72	Beta	
Probability of death due to other cause, %			Cause-of-death statistics of Korean Statistical Office
60-69 years	0.61	Beta	
70-79 years	1.65	Beta	
≥80 years	7.01	Beta	
Probability of CVD events (1-year risk)[f], %			ICT-CM trial and Framingham CVD risk prediction model
TM[g]			
60-64 years	0.90	Beta	
65-69 years	0.90	Beta	
70-74 years	1.32	Beta	
75-79 years	1.32	Beta	
≥80 years	1.80	Beta	
Usual care			
60-64 years	1.37	Beta	
65-69 years	1.37	Beta	
70-74 years	1.91	Beta	
75-79 years	1.91	Beta	
≥80 years	2.59	Beta	
Age-related relative risk of CVD			Published data using national health insurance claims data by HIRA (Kim et al [23])
60-64 years	1.00	—	
65-69 years	1.39	—	
70-74 years	1.90	—	
75-79 years	2.17	—	
≥80 years	2.06	—	
Utility weights, mean (SE)			2007-2019 KNHANES[h] database
CVD			
60-69 years	0.859 (0.006)	Beta	
70-79 years	0.794 (0.008)	Beta	
≥80 years	0.747 (0.016)	Beta	
Well			
60-69 years	0.914 (0.002)	Beta	

XSL•FO
RenderX

Parameters	Estimate in base-case	Distribution[a]	Sources
70-79 years	0.858 (0.003)	Beta	
≥80 years	0.802 (0.006)	Beta	
Intervention costs (US $[i])			ICT-CM trial
Bluetooth-enabled devices costs per patient, one-off			
Sphygmomanometer	54.5	Gamma	
Glucometer	34.1	Gamma	
Equipment costs			
Initial development costs of application and operation web per 5 years	757,575.8	Gamma	
Annual network use costs	227,272.7	Gamma	
Annual system update costs	145,454.5	Gamma	
Total costs per patient, annuitized based on lifespan of 5 years	2.3	—	
Running costs per patient[j]	141.8	Gamma	
Treatment costs of health states (US $)			Analysis using national health insurance claims data by NHIS[k]
Acute phase CVD (first year costs after event)			
60-64 years	2676.7	Gamma	
≥65 years	3639.8	Gamma	
Chronic phase CVD (annual costs in subsequent years)			
60-64 years	1524.0	Gamma	
≥65 years	2580.1	Gamma	
Well (annual costs)			
60-64 years	960.6	Gamma	
≥65 years	1566.9	Gamma	

[a]Distributions used in probabilistic sensitivity analysis.

[b]Not applicable.

[c]ICT-CM: information and communications technology–based tailored chronic disease management.

[d]CVD: cardiovascular diseases.

[e]HIRA: Health Insurance Review and Assessment Service.

[f]In base-case for all patients, the 10-year cardiovascular risk was calculated with Framingham CVD risk prediction models based on the CVD risk factors obtained from the ICT-CM trial. 1-year probability was calculated from the 10-year risk values.

[g]TM: information and communications technology–based tailored management.

[h]KNHANES: Korea National Health and Nutrition Examination Survey.

[i]All costs are expressed in 2023 US $ using an exchange rate of US $1=1320 KRW.

[j]Through the TM trial, the annual running cost was calculated by dividing the care coordinator's labor cost by the number of managed patients (approximately 250 per year).

[k]NHIS: National Health Insurance Service.

Utility Weights

The utility weights for the health states of patients with hypertension or diabetes were derived from the Korea National Health and Nutrition Examination Survey (KNHANES, 2007-2019) data conducted by the Korea Centers for Disease Control and Prevention. KNHANES is a nationwide survey conducted annually for the representative general population in Korea [26]. The survey includes individual-level information on health-related quality of life measured by the EQ-5D developed by the EuroQol Group. The 5 dimensions of EQ-5D-3L comprise mobility, self-care, usual activities, pain or discomfort, and anxiety or depression, and each dimension was answered with 3 levels. The health state index scores calculated based on the tariff of Lee et al [27] range from 0 (where 0 is a health state equivalent to death) to 1 (perfect health), with higher scores indicating higher health utility. The utility weight of "Well" status was higher than that of "CVD" by 0.05 for patients in their sixties, and the utility weight decreased with age. However, the difference in utility weight

XSL•FO
RenderX

between the 2 health states remained similar. The utility weights by age are presented in Table 2.

Costs

All costs were expressed in 2023 US $ using an exchange rate of US $1=1320 KRW and adjusted for inflation where applicable using the Consumer Price Index for health care [28]. Intervention costs of patients who received TM were obtained from the TM trial. Bluetooth-enabled device costs per patient were applied as one-off costs to the model's initial health state ("Well"). Equipment costs included initial development, network use, and system update expenditure. The per capita total equipment costs were applied to the model, estimated based on the patient number of target cohorts reflecting smartphone penetration [23]. Running costs comprised costs associated with training and care coordinator. Through the TM trial, the number of manageable patients per care coordinator was investigated, and the annual running cost was calculated by dividing the care coordinator's labor cost by the number of managed patients (approximately 250 per year). Running and equipment, annuitized based on a lifespan of 5 years, were included in intervention costs over the lifetime time horizon.

The cost of "Well" and "CVD" health states were derived from data analysis of the national health insurance claims data [29], and were applied equally to the 2 arms. Each health state's costs comprised expenditure for diagnostic test, medication, surgical treatments, and resources use associated with inpatient or outpatient services. The costs of cardiovascular events including initial acute care and long-term care were estimated. The costs of acute phase CVD were applied to the model in the first year after event, and in subsequent years, annual costs of chronic phase CVD were applied. All cost inputs are presented in Table 2.

Cost-Effectiveness Analysis

In a base-case analysis of all patients with hypertension or diabetes, the cost-effectiveness analysis of TM compared with UC was conducted from the health care system perspective. A result was presented as an incremental cost-effectiveness ratio (ICER) calculated by dividing the incremental cost by the additional quality-adjusted life-year (QALY) gained. A 4.5% annual discount rate for both cost and QALY were performed with half-cycle corrections. Deterministic and probabilistic sensitivity analysis (PSA) was performed to address parameter uncertainty. The parameters, including probabilities, utility weights, costs, number of manageable patients per care coordinator, and discount rate, were varied through deterministic sensitivity analyses. PSA of 1000 Monte Carlo simulations was undertaken, in which each of the parameter estimates was sampled from its distribution listed in Table 2. PSA results were expressed as a cost-effectiveness plane and cost-effectiveness acceptability curve that shows graphically the probability of cost-effectiveness for all alternatives across a range of an ICER threshold of US $26,515 (KRW 35 million) per QALY [30]. In addition, several scenarios were explored to assess the structural uncertainty associated with model assumptions. The assumptions regarding the effect persistence between the 2 arms were tested by varying the period during which the 6-month effect difference lasts, from the 3 years beyond the trial period to the remainder of the lifetime. In consideration of treatment persistence, the scenario considering a treatment discontinuation rate of 0% or 10% in patients receiving TM was also evaluated. Finally, the model's time horizon varied from a lifetime to between 10 and 20 years.

A subgroup analysis was performed for patients with hypertension and diabetes. The Framingham risk model used in the base-case was reported to overestimate the risk of CVD in Koreans [31], and underestimate the risk of CVD in patients with diabetes [32]. Therefore, other CVD risk models were used in each subgroup. The Korean Hypertension cardiovascular (KH-CVD) risk model for patients with hypertension [33] and the United Kingdom prospective diabetes study (UKPDS) model that estimates the risk of coronary heart disease and stroke [32,34] for patients with diabetes were used to predict long-term effects. Table S1 in Multimedia Appendix 2 shows model inputs, such as probability of CVD events, utility, and costs, estimated for each subgroup.

Results

Base-Case Analysis

Compared with UC, the cost-effectiveness analysis indicated that TM was more effective (12.006 QALYs for TM vs 11.868 QALYs for UC) and more costly (US $23,157.4 for TM vs US $22,390.5 for UC). TM had incremental costs of US $766.9 and incremental QALYs of 0.138 compared with UC over the lifetime horizon for patients with hypertension or diabetes in South Korea. Through the base-case analysis, ICER was US $5556 per QALY gained (Table 3).

XSL•FO
RenderX

Table 3. Cost-effectiveness results in base-case and sensitivity analyses.

	Cost (US $[a])	Incremental cost (US $)	QALYs[b]	Incremental QALYs	ICER[c] (US $/QALY gained)
Base-case analysis					
UC[d]	22,390.5	—[e]	11.868	—	—
TM[f]	23,157.4	766.9	12.006	0.138	5556
Sensitivity analyses					
Probability of CVD[g] events					
Increase by 20% for TM					
UC	22,390.5	—	11.868	—	—
TM	23,467.8	1077.3	11.950	0.082	13,167
Decrease by 20% for TM					
UC	22,390.5	—	11.868	—	—
TM	22,824.5	433.9	12.067	0.198	2188
Increase by 20% for UC					
UC	22,808.9	—	11.794	—	—
TM	23,190.3	381.4	12.000	0.207	1846
Decrease by 20% for UC					
UC	21,923.1	—	11.952	—	—
TM	23,124.2	1201.1	12.012	0.061	19,787
Utility weights					
Increase by 20%					
UC	22,390.5	—	13.454	—	—
TM	23,157.4	766.9	13.584	0.130	5901
Decrease by 20%					
UC	22,390.5	—	9.495	—	—
TM	23,157.4	766.9	9.605	0.110	6945
Costs					
Increase by 20% for TM					
UC	22,390.5	—	11.868	—	—
TM	23,558.5	1167.9	12.006	0.138	8462
Decrease by 20% for TM					
UC	22,390.5	—	11.868	—	—
TM	22,756.4	365.8	12.006	0.138	2651
Number of manageable patients per care coordinator					
Increase by 20%					
UC	22,390.5	—	11.868	—	—
TM	22,833.0	442.5	12.006	0.138	3206
Increase by 40%					
UC	22,390.5	—	11.868	—	—
TM	22,611.0	220.4	12.006	0.138	1597
Discount rate					
0%					

XSL•FO
RenderX

	Cost (US $[a])	Incremental cost (US $)	QALYs[b]	Incremental QALYs	ICER[c] (US $/QALY gained)
UC	37,710.2	—	18.342	—	—
TM	38,777.1	1066.9	18.637	0.295	3612
3%					
UC	26,311.4	—	13.555	—	—
TM	27,155.7	844.3	13.731	0.176	4803
Effect persistence for TM					
Lasts 3 years beyond the trial period					
UC	22,390.5	—	11.868	—	—
TM	23,672.9	1282.3	11.914	0.046	28,170
Lasts 4 years beyond the trial period					
UC	22,390.5	—	11.868	—	—
TM	23,625.5	1235.0	11.922	0.053	23,192
Lasts 5 years beyond the trial period					
UC	22,390.5	—	11.868	—	—
TM	23,566.1	1175.6	11.931	0.063	18,631
Decrease by 10% per year beyond the trial period					
UC	22,390.5	—	11.868	—	—
TM	23,183.3	792.7	12.002	0.133	5940
Decrease by 10% per year after 3 years					
UC	22,390.5	—	11.868	—	—
TM	23,168.2	777.6	12.005	0.136	5708
Decrease by 10% per year after 7 years					
UC	22,390.5	—	11.868	—	—
TM	23,145.1	754.6	12.008	0.140	5386
Last over a lifetime					
UC	22,390.5	—	11.868	—	—
TM	23,100.5	710.0	12.017	0.148	4786
Treatment persistence					
10% discontinuation rate for TM					
UC	22,390.5	—	11.868	—	—
TM	23,180.2	789.6	12.006	0.138	5721
0% discontinuation rate for TM					
UC	22,390.5	—	11.868	—	—
TM	23,154.4	763.9	12.006	0.138	5535
Time horizon					
10 years					
UC	11,081.9	—	7.326	—	—
TM	11,696.7	614.7	7.350	0.024	25,201
20 years					
UC	19,825	—	10.960	—	—

XSL•FO
RenderX

	Cost (US $[a])	Incremental cost (US $)	QALYs[b]	Incremental QALYs	ICER[c] (US $/QALY gained)
TM	20,519.2	694.3	11.050	0.090	7731

[a]All costs are expressed in 2023 US $ using an exchange rate of US $1=1320 KRW.

[b]QALYs: quality-adjusted life-years.

[c]ICER: incremental cost-effectiveness ratio.

[d]UC: usual care.

[e]Not applicable.

[f]TM: information and communications technology–based tailored management.

[g]CVD: cardiovascular diseases.

Sensitivity Analysis

In the results of the sensitivity analyses, the ICER was mainly affected by the effect persistence of TM, time horizon, the probability of CVD events for UC, the probability of CVD events for TM, and TM intervention costs, in that order (Table 3 and Figure S1 in Multimedia Appendix 2). When a 10% decrease in the TM intervention's effectiveness was applied at 6 months and 7 years after the start of the intervention, all ICERs remained below a threshold of US $26,515. However, when the additional effect due to TM intervention disappeared at 3 to 5 years beyond the trial period, the ICER increased above the threshold after 3 years. The ICER was the highest at US $25,201/QALY in the 10-year time horizon scenario. When the probability of CVD events for UC was decreased by 20%, the ICER increased by 256.1% (US $19,787/QALY). As the time when horizon and the probability of CVD events for UC decreased, the additional QALY gains in TM decreased significantly. A 20% decrease in the TM intervention costs and an increase in manageable patients per care coordinator showed cost-saving results. However, TM was the cost-effective strategy in all variations assessed in the sensitivity analyses, except for the assumption that the TM effect lasts only 3 years.

The results of the PSA were presented in cost-effectiveness planes showing the distributions of incremental costs and effects between the 2 arms (Figure 2A). Most values were distributed below the cost-effective threshold in the northeast quadrant, indicating that the base-case results were robust. The cost-effectiveness acceptability curve (Figure 2B) showed that the probability of the ICER being below the threshold of US $26,515 per QALY gained was 97%.

XSL•FO
RenderX

Figure 2. Results of probabilistic sensitivity analysis. (A) Cost-effectiveness plane. (B) Cost-effectiveness acceptability curve. Dotted line on graph indicates an incremental cost-effectiveness ratio threshold. ICER: incremental cost-effectiveness ratio; ICT: information and communications technology; QALY: quality-adjusted life-year.

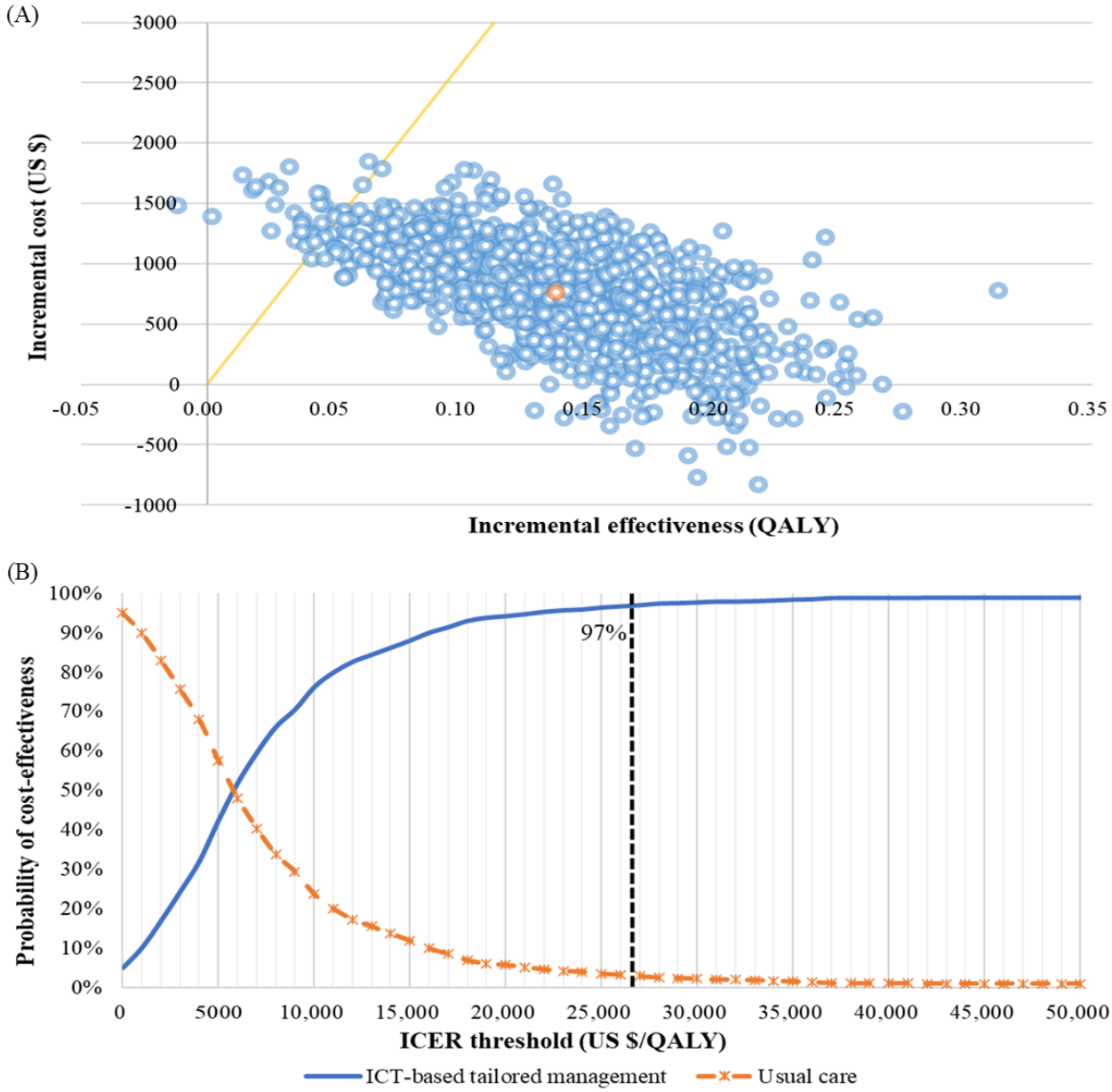

Subgroup Analysis

In the subgroup analysis of patients with hypertension, TM had US $678.7 higher costs and 0.154 higher QALYs than UC, presenting an ICER gain of US $4417 per QALY. In patients with diabetes, TM had US $885 higher costs and 0.102 higher QALYs than UC, presenting an ICER gain of $8687 per QALY (Table 4). The probability that TM would be cost-effective was 99% for treating patients with hypertension, and 86% for patients with diabetes (Figure S2 in Multimedia Appendix 2). When the long-term effect was estimated using the KH-CVD model for patients with hypertension and the UKPDS model for patients with diabetes, the ICER gain was US $8699 per QALY for patients with hypertension and US $10,409 per QALY for patients with diabetes. Consequently, TM was a cost-effective alternative in all subgroups (Table 4).

XSL•FO
RenderX

Table 4. Cost-effectiveness results for subgroups of patients with hypertension and diabetes.

	Cost (US $[a])	Incremental Cost (US $)	QALYs[b]	Incremental QALYs	ICER[c] (US $/QALY gained)
Patients with hypertension					
Long-term effect estimated using the Framingham model					
UC[d]	22,502	—[e]	11.848	—	—
TM[f]	23,180.6	678.7	12.002	0.154	4417
Long-term effect estimated using the KH-CVD[g] model					
UC	20,710.9	—	12.024	—	—
TM	21,637.1	926.2	12.131	0.106	8699
Patients with diabetes					
Long-term effect estimated using the Framingham model					
UC	26,537.5	—	11.789	—	—
TM	27,422.7	885.2	11.891	0.102	8687
Long-term effect estimated using the UKPDS[h] model					
UC	25,746.1	—	11.913	—	—
TM	26,703.2	957.1	12.005	0.092	10,409

[a]All costs are expressed in 2023 US $ using an exchange rate of US $1=1320 KRW.

[b]QALYs: quality-adjusted life-years.

[c]ICER: Incremental Cost Effectiveness Ratio.

[d]UC: usual care.

[e]Not applicable.

[f]TM: information and communications technology–based tailored management.

[g]KH-CVD: Korean Hypertension cardiovascular.

[h]UKPDS: United Kingdom prospective diabetes study.

Discussion

Principal Findings

In the economic evaluation model based on the CVD Risk Prediction Model with data from the ICT-CM pragmatic trial, we assessed the cost-effectiveness of TM for patients with hypertension or diabetes in primary care, compared with UC. The study found that TM, despite being more costly than UC, resulted in a better quality of life due to reduced cardiovascular events, making it a cost-effective option. PSA indicated that the probability of the ICER value lying below the threshold WTP was 97%.

Mobile technology is increasingly used to manage the chronic diseases of hypertension and diabetes because of its efficacy and accessibility [35]. Several devices using mobile technology improve patient care and health outcomes through lifestyle modification, promoting behavioral changes and increasing medication adherence [12]. Previous studies have shown that mHealth interventions using mobile phone SMS text messages, wearable monitoring devices, and telemedicine are cost-effective in managing chronic diseases [13]. With the continuous advance of technology, novel mobile systems using software programs such as smartphone apps and devices have recently emerged in Japan [36], and these have proven to be cost-effective for reducing BP [37]. Our study confirmed that automated algorithm-based advanced systems, in which mobile app for patients with hypertension and diabetes, patient management software for primary doctors, and Bluetooth-enabled devices were linked, had better treatment outcomes. This study is significant in that it proved the cost-effectiveness of the automated advanced mHealth system in patients with hypertension and diabetes in South Korea.

The model extrapolated effects over a lifetime based on the intervention effects at 6 months after starting the trial. The initial improvement in the HbA_{1c} levels at 6 months in the tailored mobile coaching group, including a mobile app, continued until 12 months in the randomized controlled trial despite partial cessation of interventions [38]. Another study showed that tailored phone- and home-monitoring interventions resulted in a continued greater improvement in systolic BP compared with usual care over a 2-year period [39]. Therefore, it was assumed in a base-case to be a 10% decrease in effect difference between the TM and UC groups after 5 years from the start of the intervention, which seems reasonable. In the sensitivity analyses that explored the assumption of effect persistence, the study demonstrated that the TM intervention in primary care was cost-effective compared with UC, even though the effect difference decreased by 10% immediately after the trial period. Conversely, if the added benefits of the TM intervention over UC were eliminated after 3 years, the intervention would not

XSL•FO
RenderX

be cost-effective. However, as the discontinuation rate of TM intervention was expected to be low (conservatively, 5% was applied in the model), the assumption that the effect disappears after 3 years despite ongoing treatment is considered unlikely in the real world.

In the subgroup analysis, patients with diabetes were associated with higher incremental costs and lower incremental QALYs in the TM group versus the UC group compared with patients with hypertension. Reduced equipment costs per capita due to a higher prevalence of hypertension than diabetes [23] may have contributed slightly to this result. Furthermore, 6-month outcomes, such as systolic BP and total cholesterol of the TM group compared with the UC group, improved more in patients with hypertension than in patients with diabetes, which would have resulted in better long-term effects and better QALY in patients with hypertension. While the rate of reduction in BP is relatively rapid after treatment [40], significant changes in HbA_{1c} may not occur in the short-term period after treatment change as A_{1c} reflects average glycemia over several months [41]. In a study evaluating long-term changes in HbA_{1c} for 7 years after visiting a tertiary hospital in patients with diabetes, HbA_{1c} decreased most strongly at 12 months and then was maintained thereafter [42]. Thus, in the subgroup analysis for patients with diabetes, the effect of the TM group might be underestimated compared with UC group. Nevertheless, our study found an 86% probability that the TM group would be cost-effective among patients with diabetes. Hypertension and diabetes are common comorbidities; each accelerates the other by being linked through vascular mechanisms and risk factors [43]. Therefore, implementing the TM program for patients with chronic illnesses such as hypertension or diabetes would efficiently prevent and manage CVD.

The Framingham model was used to estimate the long-term CVD outcome in a base-case, which is widely used to predict the risk of developing CVD worldwide. However, it has been reported to overestimate the CVD risk in the Asian population [31,44]. The accuracy of the Framingham model for patients with diabetes has also been debated [32]. As only 6% of the Framingham cohort was known to have diabetes, it was reported that the Framingham model tended to underestimate the risk in patients with diabetes [45]. Therefore, the study used other CVD risk models to evaluate the uncertainty due to the risk model. The KH-CVD model for patients with hypertension was developed based on the data from Korean Hypertension Cohort of 11,043 patients who were followed for 10 years at 6 national university hospitals, and the model validation demonstrated moderate predictive accuracy [33]. In tertiary hospitals, the risk of CVD may be underestimated due to strict patient management. The UKPDS model includes HbA_{1c} as a continuous variable, while the Framingham model includes dichotomous variables such as the presence or absence of diabetes as risk factors. The UKPDS model, which is sensitive to HbA_{1c} improvement, demonstrated enhanced precision in predicting CVD among patients with diabetes [32]. It exhibited acceptable discriminative performance for CVD [46]. However, the 6-month outcome from the ICT-CM trial in our study was insufficient to confirm significant improvement in HbA_{1c}; thus, it is speculated that the difference in effect between the 2 groups was underestimated. Hence, the ICER increased when estimating the CVD risk using the KH-CVD and UKPDS models compared with using the Framingham model. However, our results indicated that there was no significant change in the cost-effectiveness of TM.

The intervention cost of mHealth is a crucial factor for cost-effectiveness. Nomura et al [37] indicated that the cost of mobile app–based intervention for hypertension had the greatest impact on the ICER [37]. Although our deterministic sensitivity analysis results showed that the cost of TM intervention was not the most determinant, similar to previous studies [37,47], a reduction in intervention costs resulted in good cost-effectiveness. In particular, it might be possible that annual running costs, which account for the largest share of the TM intervention costs, could substantially decrease over time. The number of patients managed per care coordinator is expected to increase as the automation model using artificial intelligence is enhanced and care coordinators become more proficient, enabling efficient operation. This leads to reduced running costs and greater cost-effectiveness of intervention. The study found that the ICER decreased by more than 70% when the number of managed patients increased by 40% from 250 to 350 per year.

Limitations

Our study has some limitations. First, the TM intervention effect after the observed periods of the ICT-CM trial was unknown, but it was assumed to last for 5 years and then decline in the base-case. Our sensitivity analyses showed that the results remained robust to various potential reductions in effect, with the exception of no additional effect of TM after 3 years, which might be perceived as a very conservative approach. Further studies are needed to support the long-term use of various mHealth systems. Second, we used a Framingham CVD risk prediction model that was not specifically designed to predict outcomes among adults in South Korea. However, sensitivity analysis using the KH-CVD model developed in Korea demonstrated that TM intervention would still be cost-effective. Furthermore, the Framingham model has been retrospectively validated in a multiethnic Asian population in primary care settings in a 10-year cohort study [48]. Third, we could not consider whether cost-effectiveness differs in relation to lifestyle and metabolic risk changes by the amount of app usage. Nevertheless, various sensitivity analyses supported that the base-case result was robust.

Conclusions

From a Korean health care system perspective, TM intervention is a cost-effective strategy compared with UC in patients with hypertension or diabetes in primary care settings. The study can help policy makers seeking to implement accessible chronic disease management services reach at evidence-based decision-making.

XSL•FO
RenderX

Acknowledgments

This research was supported by a grant from the Korea Health Promotion Research and Development Project funded by the Ministry of Health and Welfare, South Korea (grant HS20C0041).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

J-WK and J-HK conceived the idea and achieved funding for the study. S-HO contributed to acquisition and analysis of data. J-WK, J-HK, and S-HO contributed to interpretation of data. S-HO wrote the first draft of the paper, and J-WK and J-HK helped revise the paper. J-WK and J-HK are the guarantor of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. The authors read and approved the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Information and Communications Technology–based tailored chronic disease management (ICT-CM) trial.
[DOCX File , 41 KB - jmir_v26i1e51239_app1.docx]

Multimedia Appendix 2
Model inputs for subgroups of patients with hypertension and diabetes. Tornado diagram for TM versus UC. Results of probabilistic sensitivity analysis for subgroups of patients with hypertension and diabetes. (A) Cost-effectiveness plane for patients with hypertension. (B) Cost-effectiveness acceptability curve for patients with hypertension. (C) Cost-effectiveness plane for patients with diabetes. (D) Cost-effectiveness acceptability curve for patients with diabetes.
[DOCX File , 895 KB - jmir_v26i1e51239_app2.docx]

References

1. Saeedi P, Salpea P, Karuranga S, Petersohn I, Malanda B, Gregg EW, et al. Mortality attributable to diabetes in 20-79 years old adults, 2019 estimates: results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract 2020;162:108086. [doi: 10.1016/j.diabres.2020.108086] [Medline: 32068099]
2. O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, INTERSTROKE investigators. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. Lancet 2016;388(10046):761-775. [doi: 10.1016/S0140-6736(16)30506-2] [Medline: 27431356]
3. Falaschetti E, Mindell J, Knott C, Poulter N. Hypertension management in England: a serial cross-sectional study from 1994 to 2011. Lancet 2014;383(9932):1912-1919. [doi: 10.1016/S0140-6736(14)60688-7] [Medline: 24881995]
4. Jeon JY, Kim DJ, Ko SH, Kwon HS, Lim S, Choi SH, Taskforce Team of Diabetes Fact Sheet of the Korean Diabetes Association. Current status of glycemic control of patients with diabetes in Korea: the fifth Korea national health and nutrition examination survey. Diabetes Metab J 2014;38(3):197-203 [FREE Full text] [doi: 10.4093/dmj.2014.38.3.197] [Medline: 25003073]
5. Oh SH, Kim D, Hwang J, Kang JH, Kwon Y, Kwon JW. Association of uncontrolled hypertension or diabetes mellitus with major adverse cardiovascular events and mortality in South Korea: population-based cohort study. JMIR Public Health Surveill 2023;9:e42190 [FREE Full text] [doi: 10.2196/42190] [Medline: 36735297]
6. American Diabetes Association. 1. Strategies for improving care. Diabetes Care 2016;39 Suppl 1:S6-S12. [doi: 10.2337/dc16-S004] [Medline: 26696683]
7. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Circulation 2019;140(11):e596-e646 [FREE Full text] [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
8. Savoy M, FAAFP, FABC, CPE, CMQ, FAAPL, et al. The role of primary care physicians in managing chronic disease. Dela J Public Health 2017;3(1):86-93 [FREE Full text] [doi: 10.32481/djph.2017.03.012] [Medline: 34466902]
9. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: a decade of evolution. J Med Internet Res 2013;15(5):e95 [FREE Full text] [doi: 10.2196/jmir.2430] [Medline: 23697600]
10. National Pilot Project for Chronic Disease Management in Primary Care: Korea Health Promotion Institute. URL: https://www.khealth.or.kr/ncd [accessed 2024-09-10]

XSL•FO
RenderX

11. Kang JH, Kwon JW, Oh SH. Establishmentupgrade of primary medical-based chronic disease patient monitoring service model using ICT. Korea Health Promotion Institute. 2023. URL: https://www.khepi.or.kr/kps/rsrhBusnRept/list?menuId=MENU00894 [accessed 2024-09-26]
12. Mao Y, Lin W, Wen J, Chen G. Impact and efficacy of mobile health intervention in the management of diabetes and hypertension: a systematic review and meta-analysis. BMJ Open Diabetes Res Care 2020;8(1):e001225 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001225] [Medline: 32988849]
13. Iribarren SJ, Cato K, Falzon L, Stone PW. What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. PLoS One 2017;12(2):e0170581 [FREE Full text] [doi: 10.1371/journal.pone.0170581] [Medline: 28152012]
14. Ghani Z, Jarl J, Sanmartin Berglund J, Andersson M, Anderberg P. The cost-effectiveness of mobile health (mHealth) interventions for older adults: systematic review. Int J Environ Res Public Health 2020;17(15):5295 [FREE Full text] [doi: 10.3390/ijerph17155290] [Medline: 32708016]
15. Kim HJ, Min JY, Kim HJ, Min KB. Association between psychological and self-assessed health status and smartphone overuse among Korean college students. J Ment Health 2019 Feb;28(1):11-16. [doi: 10.1080/09638237.2017.1370641] [Medline: 28868959]
16. Ford I, Norrie J. Pragmatic trials. N Engl J Med 2016;375(5):454-463 [FREE Full text] [doi: 10.1056/NEJMra1510059] [Medline: 27518663]
17. Lee HY, Shin J, Kim GH, Park S, Ihm SH, Kim HC, et al. 2018 Korean society of hypertension guidelines for the management of hypertension: part II-diagnosis and treatment of hypertension. Clin Hypertens 2019;25:20 [FREE Full text] [doi: 10.1186/s40885-019-0124-x] [Medline: 31388453]
18. Hur KY, Moon MK, Park JS, Kim SK, Lee SH, Yun JS, Committee of Clinical Practice Guidelines, Korean Diabetes Association. 2021 clinical practice guidelines for diabetes mellitus of the Korean diabetes association. Diabetes Metab J 2021;45(4):461-481 [FREE Full text] [doi: 10.4093/dmj.2021.0156] [Medline: 34352984]
19. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, CHEERS 2022 ISPOR Good Research Practices Task Force. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. Value Health 2022;25(1):3-9 [FREE Full text] [doi: 10.1016/j.jval.2021.11.1351] [Medline: 35031096]
20. Polisena J, Coyle D, Coyle K, McGill S. Home telehealth for chronic disease management: a systematic review and an analysis of economic evaluations. Int J Technol Assess Health Care 2009;25(3):339-349. [doi: 10.1017/S0266462309990201] [Medline: 19619353]
21. Teljeur C, Moran PS, Walshe S, Smith SM, Cianci F, Murphy L, et al. Economic evaluation of chronic disease self-management for people with diabetes: a systematic review. Diabet Med 2017;34(8):1040-1049. [doi: 10.1111/dme.13281] [Medline: 27770591]
22. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the framingham Heart Study. Circulation 2008;117(6):743-753. [doi: 10.1161/CIRCULATIONAHA.107.699579] [Medline: 18212285]
23. Kim H, Kim S, Han S, Rane PP, Fox KM, Qian Y, et al. Prevalence and incidence of atherosclerotic cardiovascular disease and its risk factors in Korea: a nationwide population-based study. BMC Public Health 2019;19(1):1112. [doi: 10.1186/s12889-019-7439-0]
24. Cause of Death Statistics: Korean Statistical Information Service. 2022. URL: https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B34E07&conn_path=I2 [accessed 2024-09-10]
25. Korean national health insurance statistics: healthcare bigdata hub. Health Insurance Review and Assessment Service. 2022. URL: http://opendata.hira.or.kr/ [accessed 2024-07-01]
26. Kweon S, Kim Y, Jang MJ, Kim Y, Kim K, Choi S, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). Int J Epidemiol 2014;43(1):69-77 [FREE Full text] [doi: 10.1093/ije/dyt228] [Medline: 24585853]
27. Lee YK, Nam HS, Chuang LH, Kim KY, Yang HK, Kwon IS, et al. South Korean time trade-off values for EQ-5D health states: modeling with observed values for 101 health states. Value Health 2009;12(8):1187-1193 [FREE Full text] [doi: 10.1111/j.1524-4733.2009.00579.x] [Medline: 19659703]
28. Consumer Price Index: Korean Statistical Information Service. 2022. URL: https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1J22001&conn_path=I2 [accessed 2024-09-22]
29. Seong SC, Kim YY, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening cohort (NHIS-HEALS) in Korea. BMJ Open 2017;7(9):e016640 [FREE Full text] [doi: 10.1136/bmjopen-2017-016640] [Medline: 28947447]
30. Song HJ, Lee EK. Evaluation of willingness to pay per quality-adjusted life year for a cure: a contingent valuation method using a scenario-based survey. Medicine (Baltimore) 2018;97(38):e12453 [FREE Full text] [doi: 10.1097/MD.0000000000012453] [Medline: 30235732]
31. Jee SH, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, et al. A coronary heart disease prediction model: the Korean Heart Study. BMJ Open 2014;4(5):e005025 [FREE Full text] [doi: 10.1136/bmjopen-2014-005025] [Medline: 24848088]

XSL•FO
RenderX

32. Stevens RJ, Kothari V, Adler AI, Stratton IM, United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56). Clin Sci (Lond) 2001;101(6):671-679. [Medline: 11724655]
33. Lee HY, Lee JY, Shin HG, Cho SH, Park KI, Oh GC, et al. The Korean hypertension cohort study: design and baseline characteristics. Korean J Intern Med 2021;36(5):1115-1125 [FREE Full text] [doi: 10.3904/kjim.2020.551] [Medline: 34289586]
34. Kothari V, Stevens RJ, Adler AI, Stratton IM, Manley SE, Neil HA, et al. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. Stroke 2002;33(7):1776-1781. [doi: 10.1161/01.str.0000020091.07144.c7] [Medline: 12105351]
35. Sim I. Mobile devices and health. N Engl J Med 2019;381(10):956-968. [doi: 10.1056/NEJMra1806949] [Medline: 31483966]
36. Kario K, Nomura A, Harada N, Tanigawa T, So R, Nakagawa K, et al. A multicenter clinical trial to assess the efficacy of the digital therapeutics for essential hypertension: rationale and design of the HERB-DH1 trial. J Clin Hypertens (Greenwich) 2020;22(9):1713-1722 [FREE Full text] [doi: 10.1111/jch.13993] [Medline: 32815648]
37. Nomura A, Tanigawa T, Kario K, Igarashi A. Cost-effectiveness of digital therapeutics for essential hypertension. Hypertens Res 2022;45(10):1538-1548 [FREE Full text] [doi: 10.1038/s41440-022-00952-x] [Medline: 35726085]
38. Lee DY, Park J, Choi D, Ahn HY, Park SW, Park CY. The effectiveness, reproducibility, and durability of tailored mobile coaching on diabetes management in policyholders: a randomized, controlled, open-label study. Sci Rep 2018;8(1):3642 [FREE Full text] [doi: 10.1038/s41598-018-22034-0] [Medline: 29483559]
39. Bosworth HB, Olsen MK, Grubber JM, Neary AM, Orr MM, Powers BJ, et al. Two self-management interventions to improve hypertension control: a randomized trial. Ann Intern Med 2009;151(10):687-695 [FREE Full text] [doi: 10.7326/0003-4819-151-10-200911170-00148] [Medline: 19920269]
40. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Circulation 2018;138(17):e426-e483. [doi: 10.1161/CIR.0000000000000597] [Medline: 30354655]
41. American Diabetes Association. Standards of medical care in diabetes--2013. Diabetes Care 2013;36 Suppl 1(Suppl 1):S11-S66 [FREE Full text] [doi: 10.2337/dc13-S011] [Medline: 23264422]
42. Kim H, Jung DY, Lee SH, Cho JH, Yim HW, Kim HS. Long-term changes in HbA1c according to blood glucose control status during the first 3 months after visiting a tertiary university hospital. J Korean Med Sci 2022;37(38):e281 [FREE Full text] [doi: 10.3346/jkms.2022.37.e281] [Medline: 36193638]
43. Li X, Wang J, Shen X, An Y, Gong Q, Li H, et al. Higher blood pressure predicts diabetes and enhances long-term risk of cardiovascular disease events in individuals with impaired glucose tolerance: twenty-three-year follow-up of the daqing diabetes prevention study. J Diabetes 2019;11(7):593-598 [FREE Full text] [doi: 10.1111/1753-0407.12887] [Medline: 30556339]
44. Liu J, Hong Y, D'Agostino RB, Wu Z, Wang W, Sun J, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese multi-provincial cohort study. JAMA 2004;291(21):2591-2599. [doi: 10.1001/jama.291.21.2591] [Medline: 15173150]
45. Kavaric N, Klisic A, Ninic A. Cardiovascular risk estimated by UKPDS risk engine algorithm in diabetes. Open Med (Wars) 2018;13:610-617 [FREE Full text] [doi: 10.1515/med-2018-0086] [Medline: 30847393]
46. Stephens JW, Ambler G, Vallance P, Betteridge DJ, Humphries SE, Hurel SJ. Cardiovascular risk and diabetes. Are the methods of risk prediction satisfactory? Eur J Cardiovasc Prev Rehabil 2004;11(6):521-528. [doi: 10.1097/01.hjr.0000136418.47640.bc] [Medline: 15580065]
47. Odnoletkova I, Ramaekers D, Nobels F, Goderis G, Aertgeerts B, Annemans L. Delivering diabetes education through nurse-led telecoaching. Cost-effectiveness analysis. PLoS One 2016;11(10):e0163997 [FREE Full text] [doi: 10.1371/journal.pone.0163997] [Medline: 27727281]
48. Chia YC, Gray SYW, Ching SM, Lim HM, Chinna K. Validation of the Framingham general cardiovascular risk score in a multiethnic Asian population: a retrospective cohort study. BMJ Open 2015;5(5):e007324 [FREE Full text] [doi: 10.1136/bmjopen-2014-007324] [Medline: 25991451]

Abbreviations

BP: blood pressure
CVD: cardiovascular diseases
HbA1c: glycated hemoglobin A1c
ICER: incremental cost-effectiveness ratio
ICT-CM: information and communications technology–based tailored chronic disease management
TM: information and communications technology–based tailored management
KH-CVD: Korean Hypertension cardiovascular

XSL•FO
RenderX

KNHANES: Korea National Health and Nutrition Examination Survey
mHealth: mobile health
PSA: probabilistic sensitivity analysis
QALY: quality-adjusted life-year
UC: usual care
UKPDS: United Kingdom prospective diabetes study

Edited by T de Azevedo Cardoso; submitted 26.07.23; peer-reviewed by W Lin, J Lee; comments to author 16.01.24; revised version received 07.03.24; accepted 14.08.24; published 11.10.24.

Please cite as:
Oh SH, Kang JH, Kwon JW
Information and Communications Technology–Based Monitoring Service for Tailored Chronic Disease Management in Primary Care: Cost-Effectiveness Analysis Based on ICT-CM Trial Results
J Med Internet Res 2024;26:e51239
URL: https://www.jmir.org/2024/1/e51239
doi:10.2196/51239
PMID:

©Sung-Hee Oh, Jae-Heon Kang, Jin-Won Kwon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Economic Impact of the PedAMINES App in Reducing Medication Errors in Pediatric Emergency Care: Cost-Effectiveness Analysis

Loïc Brunner[1], MSc; Johan N Siebert[2,3], MD, PD; Frédéric Ehrler[4], PhD; Sergio Manzano[2,3], MD; Joachim Marti[1], PhD

[1]Department of Epidemiology and Health Systems, Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland
[2]Department of Pediatric Emergency Medicine, Geneva Children's Hospital, Geneva University Hospitals, Geneva, Switzerland
[3]Faculty of Medicine, University of Geneva, Geneva, Switzerland
[4]Department of Radiology and Medical Informatics, Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Loïc Brunner, MSc
Department of Epidemiology and Health Systems
Center for Primary Care and Public Health (Unisanté)
University of Lausanne
Rue du Bugnon 44
Lausanne, 1011
Switzerland
Phone: 41 21 314 60 60
Email: brunner.loic2@gmail.com

Abstract

Background: The administration of drugs in pediatric emergency care is a time-consuming process and is associated with a higher occurrence of medication errors compared with adult care. This is attributed to the intricacies of administration, which involve calculating doses based on the child's weight or age. To mitigate the occurrence of adverse drug events (ADEs), the PedAMINES (Pediatric Accurate Medication in Emergency Situations; Geneva University Hospitals) mobile app has been developed. This app offers a step-by-step guide for preparing and administering pediatric drugs during emergency interventions by automating the dose calculation process. Although previous simulation-based randomized controlled trials conducted in emergency care have demonstrated the efficacy of the PedAMINES app in reducing drug administration errors, there is a lack of evidence regarding its economic implications.

Objective: This study aims to evaluate the cost-effectiveness of implementing the PedAMINES app for 4 emergency drugs: epinephrine, norepinephrine, dopamine, and midazolam.

Methods: The economic evaluation was conducted by combining hospital data from 2019, previous trial outcomes, information extracted from existing literature, and PedAMINES maintenance costs. The cost per avoided medication error was calculated, along with the number of administrations needed to achieve a positive return on investment. Subsequently, Monte Carlo simulations were used to identify the key parameters contributing to result uncertainty.

Results: The study revealed the number of preventable errors per administration for the 4 examined drugs: 0.513 for epinephrine, 0.484 for norepinephrine, 0.500 for dopamine, and 0.671 for midazolam. The cost-effectiveness ratios per ADE prevented were computed as follows: US $4808 for epinephrine, US $9705 for norepinephrine, US $6957 for dopamine, and US $2074 for midazolam. Accounting for the economic impact of ADEs, the analysis estimated that 16 administrations of epinephrine, 17 of norepinephrine and dopamine, and 13 of midazolam would be required to attain a positive return on investment. This corresponds to roughly one-third of the annual administrations at a major university hospital in Switzerland. The primary factors influencing the uncertainty in the estimated cost per ADE include the cost of maintenance of the app, the likelihood of an ADE resulting from an administration error, and the frequency of underdosing in the trial's control group.

Conclusions: A dedicated mobile app presents an economically viable solution to alleviate the health and economic burden of drug administration errors in in-hospital pediatric emergency care. The widespread adoption of this app is advocated to pool costs and extend the benefits on a national scale in Switzerland.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52077) doi:10.2196/52077

KEYWORDS

adverse drug event; health information technology; pediatric care; emergency care; ambulance care; economic evaluation; cost-effectiveness; epinephrine; norepinephrine; midazolam; dopamine; evidence-based; medical app; medication error; pediatric; child; pediatric emergency care; PedAMINES; Pediatric Accurate Medication in Emergency Situations; Switzerland; child care; mobile phone

Introduction

Medication errors are among the most common medical errors, with a significant health and economic impact on health care systems [1]. Incorrect drug doses increase the risk of adverse drug events (ADEs), which can adversely affect patient health outcomes, consume human and material resources, and increase health care costs [2-5]. Pediatric patients are at increased risk because most drugs administered intravenously to children are provided in vials originally prepared for the adult population, which must be dosed and prepared according to each child's individual weight, varying widely between age groups [6,7]. This risk may be further increased in pediatric emergency departments (PED) and pediatric intensive care units (PICUs) due to the complexity of stressful and high-risk situations encountered [8-10]. In addition, comorbidities and poor health status increase the likelihood of ADE [11,12]. Although health information technology (HIT) plays an increasingly important role in modern medicine, notably by improving the safety of the medication process [13-17], there is limited evidence of its effectiveness and efficiency [15], particularly in pediatrics [16-19]. Therefore, the need for clinical and cost-effectiveness evidence to support the implementation of digital health interventions to prevent medication error-related ADE in pediatric emergency care has become increasingly important.

Previous trials have demonstrated the ability of a mobile medical app, PedAMINES (Pediatric Accurate Medication in Emergency Situations; Geneva University Hospitals), to significantly reduce out-of-hospital and in-hospital medication errors for intravenous drug administration compared with conventional preparation methods during simulation-based pediatric resuscitation [20,21]. The app guides the user through the preparation and administration of drugs by automatically calculating the correct weight-based doses and providing a detailed preparation sequence. This study evaluated the cost-effectiveness and net economic benefit of using the app compared with conventional drug preparation methods to reduce the rates of medication errors and ADEs during intravenous drug administration in pediatric emergency care in a tertiary hospital. The hypothesis was that the use of the app could be a cost-effective strategy at the societal level to reduce medication errors and the associated economic burden in pediatric emergency medicine.

Methods

Intervention

PedAMINES provides automatic calculation support for the injection of drugs at pediatric doses. Registered as a medical device, it consists of 2 main parts. The first part features a welcome screen where the user is prompted to enter the patient's weight or age. Once this information is entered, the user is directed to a screen where each of the drugs listed for direct intravenous injection or continuous infusion can be selected from a menu on the left-hand side, with a detailed preparation procedure displayed on the right-hand panel, following a standardized and simplified path. The second part is an interface for editing, managing, and sharing drug lists with other users. Additional information about the app's interface has been published previously [17]. Multicenter randomized controlled trials have demonstrated the effectiveness of the app in reducing pediatric dosing errors during emergency drug injections in life-threatening critical situations [20,21]. Data from these trials are used below to examine the app's cost-effectiveness.

Economic Evaluation

Initially, the reduction in error rates for direct intravenous administration of pediatric doses of epinephrine using the app was evaluated in the prehospital setting [20]. Epinephrine was selected due to its frequent use as a first-line agent in emergency situations, especially in pediatric cardiopulmonary resuscitation (CPR) and its extensive coverage in scientific literature, making it a standard reference and comparator drug. Subsequently, the findings from prehospital settings were extrapolated to the PICU environment of a tertiary hospital in Switzerland.

In the economic evaluation, the costs (ie, maintenance, update, and training costs) as well as the health benefits (ie, avoided administration errors with associated consequences) and nonhealth benefits (avoided costs, time saved) of using the app in pediatric care at the Geneva University Hospitals were considered. First, a cost-effectiveness analysis was performed, expressed as the cost per error avoided. Second, a simple decision-analytic model was used to translate the reduction in error rates into prevented ADEs of varying severity and, ultimately, avoided costs. The cost per ADE avoided was then used to determine the minimum number of patients per year that would need to be treated with the app for the intervention to be cost-neutral.

The analysis was then extended to another commonly used direct intravenous drug in the prehospital setting, midazolam [20]. Finally, to assess the potential benefits of using the app in multiple contexts, the app's use for the continuous infusion of norepinephrine and dopamine in the in-hospital emergency care setting (ie, PED) was further investigated. The choice of these 2 vasopressors was based on previous experience with their use [21]. In addition, dosing recommendations and preparation methods for the continuous infusion of norepinephrine were similar to those for epinephrine [22]. The costs and benefits of using the application were evaluated separately for each drug to provide a conservative perspective on implementing the application for only one drug at a time.

XSL•FO
RenderX

Intervention Costs

We first estimated the direct costs of using the app from the provider's perspective. The upfront investment cost of developing the app was not included in the cost-effectiveness calculations (ie, US $186,026), but we did include the annual recurring costs of maintenance and upgrades. These costs were estimated to be 20% of the development costs (ie, US $37,205) per year.

The direct costs of using the app from the provider's perspective were initially estimated. The upfront investment cost of developing the app, totaling US $186,026, was not included in the cost-effectiveness calculations [23,24]. However, annual recurring costs for maintenance and upgrades, estimated at 20% of the development costs (US $37,205 per year), were included. It was assumed that hardware costs would be zero, as the app can be easily installed on standard hospital equipment such as smartphones or tablets commonly used in ambulances [25,26]. In addition, the cost of an initial 15-minute basic training session on the app, along with annual 15-minute refresher sessions for medical staff, was included. These training costs were valued based on the salaries of nurses and physicians in pediatric care at the institution's accounting services, resulting in an annual training cost of approximately US $5 for nurses and US $7.5 for physicians. For instance, in 2019, the PED employed 60 full-time equivalent nurses and 22.8 full-time equivalent physicians, resulting in a total annual training cost of about US $1400.

Intervention Outcomes and Benefits

The medication errors considered in the analysis are defined as deviations from the correct weight-based dose by more than or less than 10% [27]. The number of errors reported in patients with and without the assistance of the app for emergency drug preparation was previously identified [20]. Using this information, a ratio of preventable errors per epinephrine administration was calculated. Subsequently, the maintenance and training costs outlined earlier were used to quantify the resources invested per error avoided.

To estimate the avoided costs associated with the reduction in errors (N_{err}), the potentially prevented adverse drug events (N_{ADE}) attributable to error reduction were then estimated using scientific literature and hospital data as sources of information. ADEs were defined according to the criteria of the National Coordinating Council for Medication Error Reporting and Prevention, categorized as follows: (1) temporary patient harm requiring intervention (*Category E*) and (2) temporary patient harm requiring initial or prolonged hospitalization (*Category F*) [28]. The probability that a medication error would result in an ADE was calculated, along with the probability that an error would result in temporary harm requiring either intervention ($N_{Cat\ E}$) or prolonged hospitalization ($N_{Cat\ F}$) [29-32]. Due to the lack of evidence on mortality from medication errors in pediatric emergency care, a conservative assumption was made that an error would never result in patient death. Although some patients may experience multiple concurrent ADEs in 15% of cases [29], for simplicity, it was assumed that an error could result in at most 1 ADE.

$$NADE = N_{err} \cdot P_{(ADE)}$$

$$NCat\ E = NADE \cdot P_{(Cat\ E)}$$

$$NCat\ F = NADE \cdot P_{(Cat\ F)}$$

These ADEs were then converted into costs. For ADEs requiring intervention, the cost consequences were estimated as the average excess reimbursement received by the hospital for complication-related hospital stays (ie, +90%; Multimedia Appendix 1), which was applied to the average cost of a hospital stay. Studies reporting outcomes for ADEs requiring prolonged length of stay (LOS) in adult [33,34] and pediatric [35,36] populations were reviewed. However, the heterogeneity of these studies, which was highly dependent on the setting and the type of drug studied, did not allow an exact number of avoidable hospital days to be determined. Therefore, the increase in LOS due to opioid-related adverse events in surgical hospitalizations, equivalent to an additional 0.64 days [33], was used as a reference. The cost of these additional days resulting from an ADE was evaluated using the average daily hospital cost at the Geneva University Hospitals.

As an additional potential benefit, the economic value of time saved by using the app compared with standard drug dose calculations was included. The previously identified time savings per drug injection [20] were valued using corresponding wage information.

Economic Analysis

All costs and outcomes were identified for a single year, 2019. The cost and outcome data described above were used to calculate incremental cost-effectiveness ratios (ICERs; $ICER_{err}$ and $ICER_{ADE}$) [37]. The ICER is expressed as the ratio of the difference in cost between 2 strategies (ie, between the app [C_{app}] and conventional [C_{conv}] drug preparation methods) to the difference in effectiveness between the app (E_{app}) and the conventional method (E_{conv}).

Subsequently, an analysis was conducted that included the economic benefit of avoiding ADEs, and the results were expressed in terms of the minimum number of administrations required to achieve a positive return on investment (ROI). To perform this calculation, the costs of using the app (maintenance [*M*], update [*U*], and training costs [*T*]) were compared with its monetary benefits derived from the time saved during administration (*TA*).

A probabilistic sensitivity analysis of the parameters was then performed to reflect the uncertainty in the decision problem, using either a uniform or normal distribution where available (Table 1). The relative impact of uncertainty in key parameters on the number of uses required to achieve a positive ROI was also determined. In addition, Monte Carlo (MC) simulation [36] was used, involving random sampling based on probability distributions of the parameters, to provide further insights into the uncertainty surrounding the calculated values [38].

XSL•FO
RenderX

To maintain consistency with the timing of other data used and to mitigate the potential impacts of the COVID-19 pandemic on costs in 2020 and 2021, all costs in the following analyses were converted to US dollars based on the costs in CHF from the in-hospital trial conducted in 2019 [21] (exchange rate of CHF 1=US $1.03 as of December 31, 2019). All parameters used in the calculations are detailed for epinephrine (Table 1) and for norepinephrine, dopamine, and midazolam (Multimedia Appendix 2). Analyses were performed using Microsoft Excel 2016 (version 16.0.5254.1000).

Table 1. Variables and distributions used in the analysis.

Parameter	Distribution[a]	Mean	SD	CI 95%	Reference
Total number of epinephrine administrations investigated in the 2021 prehospital trial	Fixed value	76	__[b]	—	[20]
Preparation time to intravenous epinephrine administration in the experimental group (ie, with app support), in seconds	Normal	191.6	80.3	34.2-349	[20]
Time to intravenous epinephrine administration in the control group (ie, without app support), in seconds	Normal	201.1	73.4	56.2-344	[20]
Number of epinephrine overdoses (>10% of the prescribed dose) in the experimental group	Uniform	2.0	—	1.6-2.4	[20]
Number of epinephrine overdoses (>10% of prescribed dose) in the control group	Uniform	7.0	—	5.6-8.4	[20]
Number of epinephrine underdoses (>10% of prescribed dose) in the experimental group	Uniform	2.0	—	1.6-2.4	[20]
Number of epinephrine underdoses (>10% of prescribed dose) in the control group	Uniform	36.0	—	28.8-43.2	[20]
Probability of an ADE[c]	Normal	11.1%	1.1%	9.1%-13.5%	[28]
Probability of temporary patient harm and need for initial or prolonged hospitalization (category F)	Normal	3%	1.89%	0.6%-8%	[28]
Probability of temporary patient harm and complication (category E)	Normal	97%	1.89%	92%-99.4%	[28]
Daily cost of an inpatient stay in the PICU[d], in US $, 2019	Uniform	229.7	—	183.8-275.7	Hospital accounting
Increased cost of hospitalization due to complications: 91.84% of PICU cost, in US $, 2019	Uniform	22,080.0	—	17,664-26,496	(Multimedia Appendix 1)
Costs of using and maintaining the app, including the software and resources used to use the app each year.	Uniform	37,205.3	—	29,764.2-44,646.3	Expert opinion + [23,24]
Cost of training a nurse in the PED[e], in US $, 2019	Uniform	885.5	—	708.4-1052.6	Hospital accounting + expert opinion
Cost of training a physician in the PED, in US $, 2019	Uniform	522.7	—	418.1-627.3	Hospital accounting + expert opinion
Conditional length of stay, ie, the number of days of prolonged hospitalization in the event of an ADE.	Normal	0.6	0.1	0.4-0.9	[33]
US $: CHF exchange rate on 31 December 2019	Fixed value	1.03	—	—	[39]

[a]Where the information on the distribution of the variable was not available, a uniform distribution with a CI of ±20% was assumed.

[b]Not applicable.

[c]ADE: adverse drug event.

[d]PICU: pediatric intensive care unit.

[e]PED: pediatric emergency department.

Dopamine, Norepinephrine, and Midazolam Replication

The app supports the preparation of multiple drugs for direct intravenous injection or continuous infusion. Therefore, the same methodology described above was applied to 3 other injectable drugs: norepinephrine, dopamine, and midazolam. The use of the app with dopamine and norepinephrine has been studied in the in-hospital setting [21], while its use with midazolam has been investigated in the same prehospital study as epinephrine [20], demonstrating the app's value in different emergency contexts and preparation methods. All parameters for these additional analyses are provided in Multimedia Appendix 2. Replicating the analysis for different settings and

XSL•FO
RenderX

types of drugs is crucial for reinforcing the external validity of the results. The selected drugs include 2 for direct intravenous administration in a prehospital setting by paramedics and 2 for continuous infusion in an in-hospital setting by nurses and physicians.

Ethical Considerations

This study uses preexisting secondary data that was published before our analysis. The 2 principal trials referenced in this study underwent ethical review by the Geneva Cantonal Ethics Committee, as documented in the original publications [20,21]. For the first trial in 2016, the Geneva Cantonal Ethics Committee did not assign an approval code at the time, while for the second trial, the approval code was Req-2019-00773. As a result, additional ethical review was not required for this research. Furthermore, no individual participants from any of the studies used in this analysis can be identified.

Results

Cost Per Error Prevented Using the App for Epinephrine, Dopamine, Norepinephrine, and Midazolam

The number of errors and ADEs prevented using the app for direct intravenous epinephrine was calculated (Table 2). An error rate of 44 (57.9%) out of 76 injections was observed for epinephrine injection in the prehospital setting using conventional preparation methods, compared with 4 (5.4%) out of 74 injections using the app, resulting in a reduction in error rate of 53.8% (95% CI 38.4-66.4) [20]. The cost per prevented ADE was estimated at US $4808. Unfortunately, there was no available information on the cost of ADEs in Switzerland to contextualize this value. However, in Japan, the cost of an ADE was estimated at US $8258.23 [40], which is twice as high as the cost of preventing an ADE with the PedAMINES app. The literature provides limited information on the economic burden of ADEs in pediatric care. In the German emergency population, the mean cost of an ADE was €2743 (~US $2970) [41]. In the US population older than 64 years, the cost of ADE-related hospital admission has been estimated at US $17,796 [42]. These estimates highlight the challenges in determining ADE-related costs, which are highly dependent on context, the specific types of costs considered, and the population studied. However, they confirm the general range of costs associated with an ADE.

The number of errors ADEs prevented using the app for direct intravenous midazolam and continuous infusions of norepinephrine and dopamine was calculated (Table 2). Using the same calculation method as for epinephrine, the cost per error prevented for these 3 drugs was determined. It was found that preventing an ADE associated with norepinephrine would be the most expensive, while preventing an ADE associated with midazolam would be the least expensive. This analysis was repeated, accounting for an annual device cost of US $600 (acquisition of 2 tablets costing US $300 each annually), and showed similar results (more details in Multimedia Appendix 3).

Table 2. Preventable errors, preventable ADE[a], cost per prevented error, and cost per prevented ADE using the app for epinephrine, dopamine, norepinephrine, and midazolam.

Drugs	Preventable errors per administration, n	Preventable ADE per administration, n	Cost per prevented error, in US $	Cost per prevented ADE, in US $
Epinephrine (direct IV[b])	0.513	0.057	534	4808
Dopamine (continuous infusion)	0.500	0.056	772	6957
Norepinephrine (continuous infusion)	0.484	0.054	1077	9705
Midazolam (direct IV)	0.671	0.074	230	2074

[a]ADE: adverse drug event.

[b]IV: intravenous.

Number of Drug Administrations Required to Achieve a Positive ROI Using the App for Epinephrine, Dopamine, Norepinephrine, and Midazolam

Overall, the number of epinephrine administrations required to achieve a positive ROI was found to be 16 (Table 3). Extrapolating to 2019, when the PICU of Geneva University Hospitals reported 40 administrations, a positive ROI could have been achieved for the remaining 24 administrations.

Similarly, a comparable number of drug administrations required to achieve a positive ROI with the app was estimated (Table 3). In 2019, using the app in our PICU with any of these drugs would have achieved a positive ROI.

The number of administrations required to achieve a positive ROI and the results of 10,000 Monte Carlo simulations for each of these drugs are shown in Table 3. In 95% of the simulations, a positive ROI would be obtained with the app after a maximum of 23 administrations of epinephrine, a maximum of 24 administrations of dopamine, a maximum of 35 administrations of norepinephrine, and a maximum of 19 administrations of midazolam.

XSL•FO
RenderX

Table 3. Number of drug administrations required to achieve a positive ROI[a] and the result of Monte Carlo simulations.

Drugs	Drug administrations required to achieve a positive ROI (calculated), n	Drug administrations to achieve a positive ROI in 95% of cases (calculated), n	Total administrations at the PICU[b] in 2019, n
Epinephrine (direct IV[c])	16 (15.8)	23 (23.0)	40
Dopamine (continuous infusion)	17 (16.2)	24 (23.4)	100
Norepinephrine (continuous infusion)	17 (16.7)	35 (24.7)	141
Midazolam (direct IV)	13 (12.1)	19 (18.1)	250

[a]ROI: return on investment.

[b]PICU: pediatric intensive care unit.

[c]IV: intravenous.

Sensitivity Analysis

Sensitivity analysis of individual variables provided information on the relative importance of each parameter in the model used to predict the number of epinephrine administrations required to achieve a positive ROI. The independent effect of each parameter on the results is illustrated in a tornado plot (Figure 1). Maintenance and usage costs, the probability of occurrence of an ADE, and the frequency of underdosing in the control group were identified as the parameters with the strongest individual influence on the number of administrations required to achieve a positive ROI.

Figure 1. Deterministic sensitivity analysis (tornado plot): epinephrine. Relative importance of the uncertainty according to the variables used in the analysis. ADE: adverse drug event; cat: category; CLOS: conditional length of stay; ctrl grp: control group; epi: epinephrine; exp grp: experimental group; hosp: hospital; Nb: number; OD: overdoses; prob: probability; ROI: return on investment; UD: underdoses.

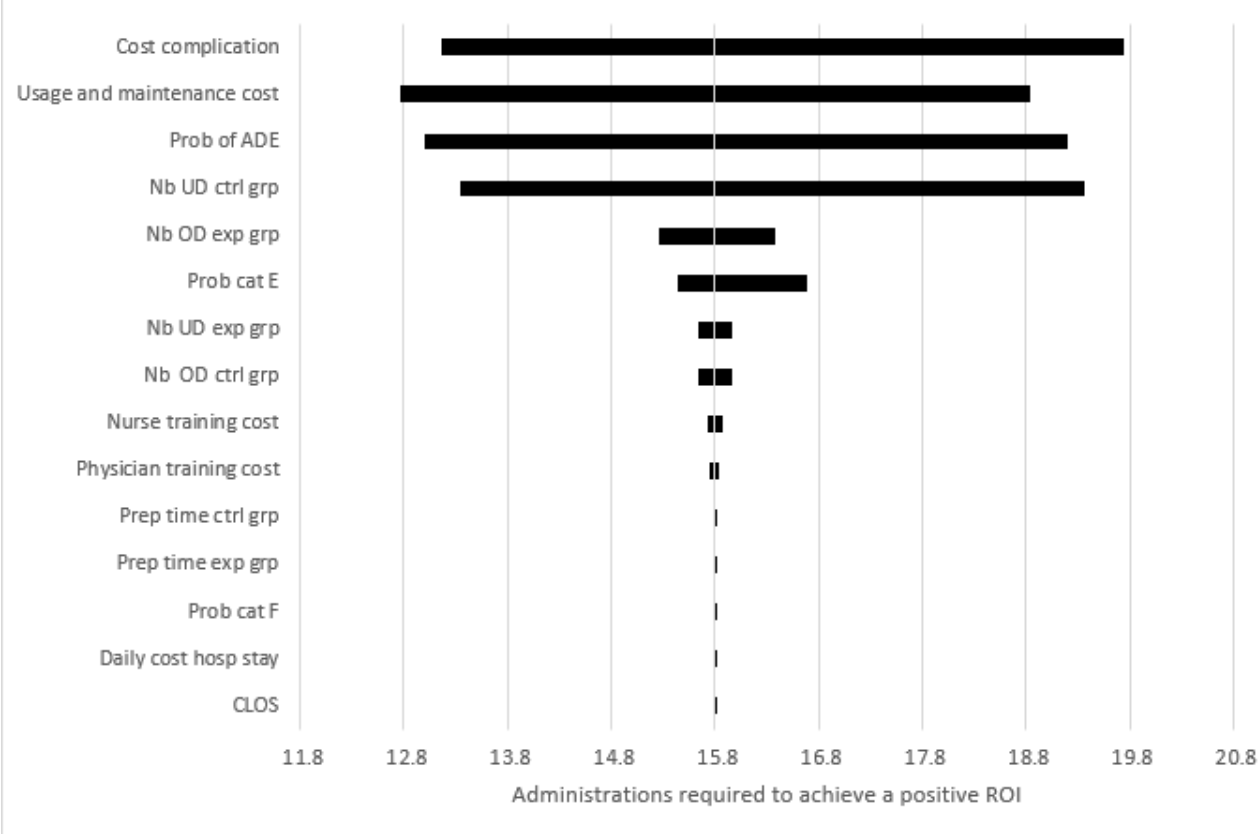

Probability distributions were used (Table 1) to assess the sensitivity of the results to parameter uncertainty. The distribution of results from the MC simulation is shown in a histogram (Figure 2). In this histogram, the gray bars represent the cumulative frequency, while the dark bars show the frequency of individual outcomes from the simulation. Subsequently, cumulative frequencies of these simulations based on probabilistic variations of the variables are displayed in Figure 3. In this figure, the grey line marks the threshold below which 95% of the simulations fall, while the black line represents the cumulative frequency of the Monte Carlo simulations. Out of 10,000 MC simulations, 95% achieved a positive ROI within 23 administrations. The sensitivity analysis reveals similar results for norepinephrine, midazolam, and dopamine. Therefore, cumulative simulation frequencies and tornado plots are provided in the supplementary material (Multimedia Appendices 4-5).

XSL•FO
RenderX

Figure 2. Proportion of simulations with a positive return on investment by the number of administrations: epinephrine. The number of administrations required to achieve a positive return on investment on 10,000 Monte Carlo simulations.

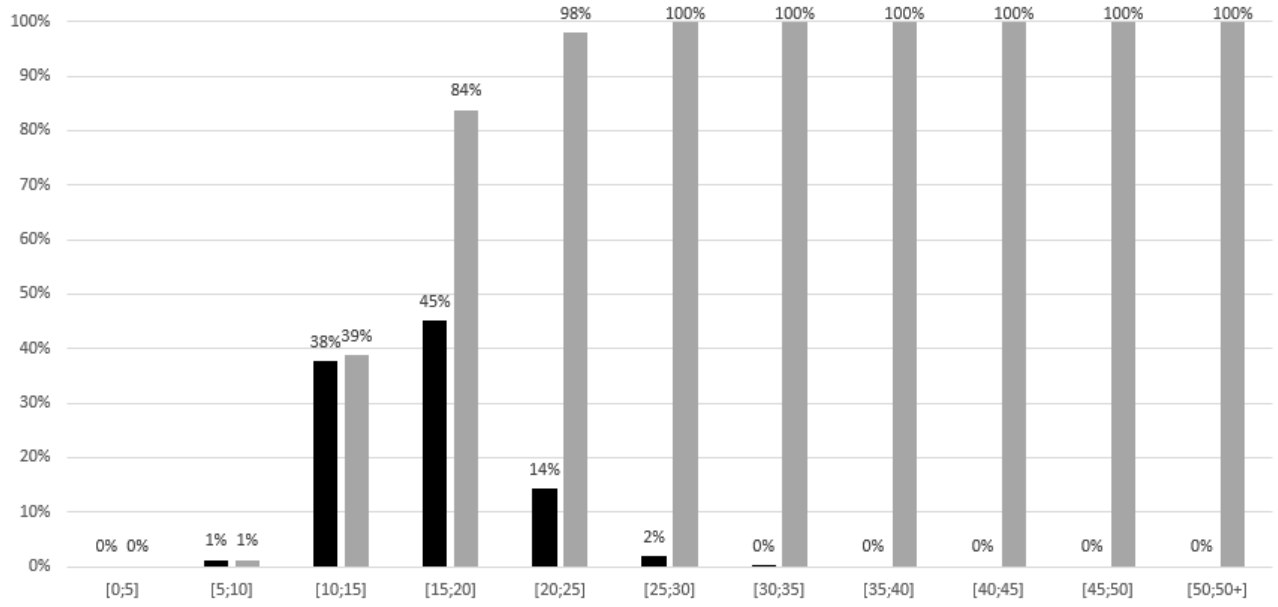

Figure 3. Sensitivity analysis, cumulative frequencies: epinephrine. Cumulative frequency of the number of administrations required to achieve a positive return on investment on 10,000 Monte Carlo simulations.

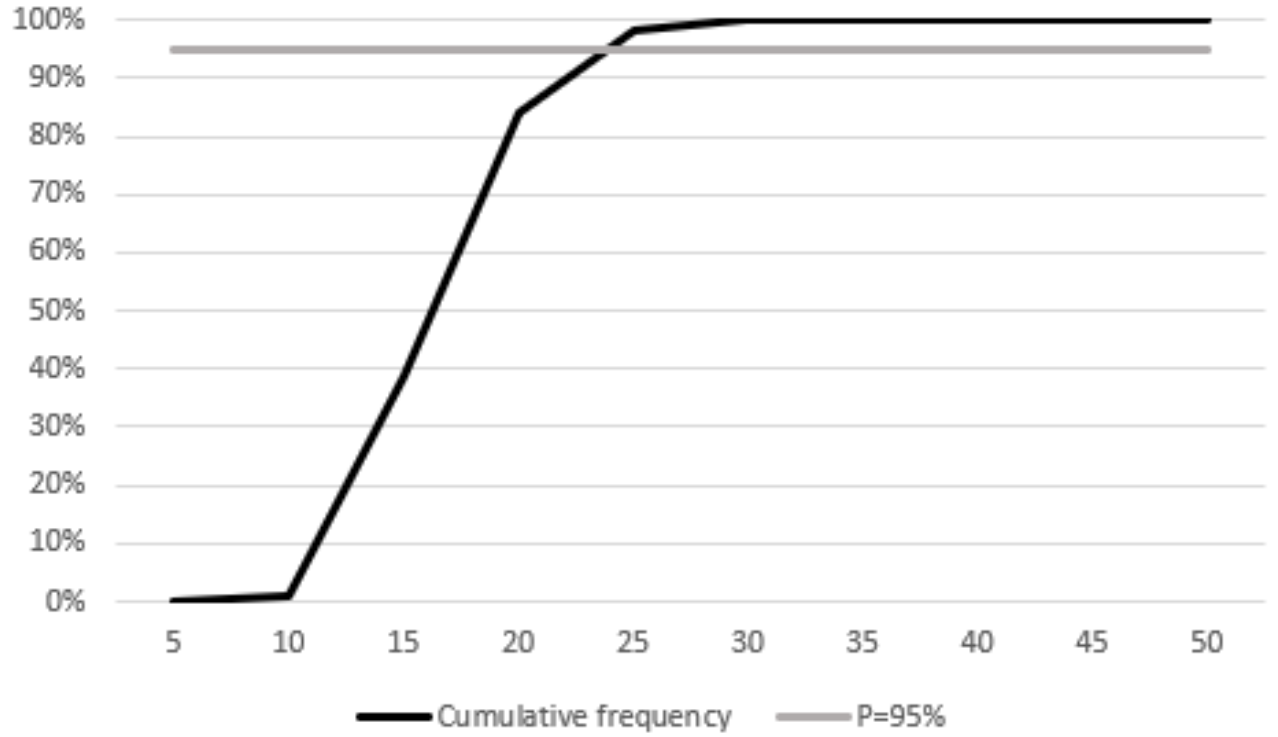

Discussion

Principal Findings

To the best of current knowledge, this study offers the most comprehensive economic analysis to date of using a medical mobile app to reduce medication errors and prevent ADEs in pediatric emergency care, compared with conventional drug preparation methods (Multimedia Appendix 6). The costs and benefits are presented from Geneva University Hospitals' perspective and could be extrapolated to the Swiss health care system. The cost per error prevented for epinephrine was US $ 534, and the cost per ADE prevented was US $ 4808. For epinephrine, the ROI becomes positive and profitable after the 16th injection, which is economically compelling given its frequent use in PICUs and PEDs globally. In 2019, Geneva University Hospitals administered epinephrine 40 times. This finding for epinephrine can be extended to other specific drugs. Similarly, considering the cost-effectiveness of midazolam, norepinephrine, and dopamine, it indicates that numerous other medications could derive advantages from the app for safe preparation and administration, especially high-risk drugs

XSL•FO
RenderX

susceptible to significant patient harm if improperly prepared (eg, chemotherapeutics, narcotics, anticoagulants, electrolytes). Furthermore, with no upfront development costs and only maintenance expenses, if the app is purchased, achieving a positive ROI would be accelerated, and the additional cost of preventing an ADE would be reduced.

Limitations and Future Work

This study has limitations, primarily related to the composite study framework used. First, ADE estimates were based on existing literature, which remains scarce in this field and limits generalization. Second, the partial reliance on data from simulated trials may be critiqued for its lack of external validity. However, high-fidelity simulations have been shown to effectively simulate real-life scenarios for evaluating research questions and technologies that are challenging to assess in real-world settings [43]. In addition, there is currently no data available on whether mobile apps reduce medication errors or ADEs in real-life settings. It is plausible that medication errors could potentially be higher in routine practice without app support due to external factors not accounted for in simulated trials. If so, this could underestimate the app's ability to further reduce medication errors and ADEs in real-world settings and, consequently, the marginal cost of error reduction.

Third, while evidence suggests comparable medication error rates between in-hospital and prehospital settings [20,21], prehospital medication error rates were applied to the in-hospital setting, acknowledging differences in contexts and health care providers involved. Fourth, data specific to drug volumes used in the PED of Geneva University Hospitals were not available as they were not separately recorded. Therefore, the number of drug administrations required to achieve a positive ROI was estimated based on drug volumes administered in the PICU of Geneva University Hospitals, assuming app use in both PED and PICU settings.

Fifth, the reported error rate from the literature may have been underestimated, as medication errors are often underreported due to a lack of recognition or reluctance to report [44-46]. Sixth, our study's cost list was not exhaustive. Extremely rare ADEs, such as death due to medication errors in pediatric care, were not included. In addition, indirect economic or health impacts, such as the effect of medication errors on patient quality of life, were not considered, presenting a conservative view of potential costs related to medication errors.

Finally, hospital systems may prefer integrated solutions within their workflows, and requiring staff to use personal devices without compensation could present challenges. However, PedAMINES, as an evidence-based app, provides a reliable and immediate solution to enhance medication safety, which may not always be achievable with existing hospital systems or in low-resource settings. Previous publications have demonstrated the app's usability and technology acceptance during simulated pediatric in- and out-of-hospital CPR, affirming its practical applicability and effectiveness [47].

While this study primarily focused on the use of PedAMINES for epinephrine, we have already assessed the app's effectiveness for other direct intravenous emergency drugs (midazolam, 10% dextrose, and sodium bicarbonate) [20], as well as for continuous infusions (dopamine and norepinephrine) [21]. The consistent decrease in risk to approximately 5% for all drugs, regardless of their varying degrees of preparation difficulty, reflects the app's ability to secure the preparation stage of the medication process, irrespective of the drugs. Future work should include real-world studies to validate the app's broader applicability and potential to reduce medication errors across different pediatric settings, such as onco-hematology or in low-resource countries. In addition, future studies should address the effectiveness of the app concerning age or weight categories separately. This approach will provide a clearer understanding of the factors influencing error rates in medication administration.

Comparison With Previous Work

The literature has evaluated various interventions to reduce the economic burden of medication errors, including electronic systems, process interventions, patient-centered approaches, and interprofessional education programs [1]. Specifically, smart pumps have demonstrated cost-effectiveness in pediatric care by reducing administration errors [48,49], achieving a positive ROI of 1.15. At Geneva University Hospital (HUG) in 2019, using PedAMINES for 40 annual epinephrine administrations would have yielded an ROI of 2.5. Unlike the previous trials of the authors, the impact of a mobile app on reducing drug administration errors in pediatric care has not yet been extensively studied. The findings of this study underscore the potential of consumer-grade technology to enhance medical care efficiency in preparing and administering weight-based emergency drug doses, thereby improving medication safety.

PedAMINES has effectively reduced medication errors and ADEs during the administration of pediatric drug doses, particularly with epinephrine, at a reasonable cost per ADE prevented. In the Netherlands, Jeserun et al [50] reported a cost of €17.69 (~US $19.16) per error prevented using barcode administration, which is lower than our findings with PedAMINES. However, this discrepancy may be influenced by the specific focus of this study, which evaluated the app's benefits with a limited range of drugs and settings.

These results are notably influenced by several key parameters: the cost of complications, the expenses for maintenance and usage, and the probability of ADEs. Given the constraints of our study and the limited evidence available for the Swiss context, we made necessary assumptions for these variables.

First, the annual maintenance and usage costs were estimated at 20% of the development cost, based on established models in the literature [23,24]. This figure reflects insights provided by PedAMINES developers regarding the ongoing resources required for app maintenance.

Second, the probability of ADE was informed by literature specific to pediatric emergency settings [28]. A conservative approach was adopted by assuming no fatalities associated with ADEs were prevented by the app, potentially underestimating its overall benefits. In England, drug error reduction software has been shown to prevent approximately 100 deaths across 745,170 admissions, with more harmful errors reported for

XSL•FO
RenderX

pediatric patients [51]. While it is challenging to determine the exact extent to which medication errors in pediatric care lead to fatalities, these results suggest that the use of PedAMINES could provide nonaccounted benefits, potentially preventing deaths due to medication errors and thereby yielding significant cost savings.

Our MC simulations, accounting for parameter uncertainty, demonstrated that a positive ROI was achieved in 95% of scenarios after just 23 epinephrine administrations. This contrasts with the 40 annual epinephrine administrations at HUG, highlighting the potential for PedAMINES to yield positive economic outcomes when scaled nationally or internationally and applied to multiple drugs.

Further research, ideally on a national or multinational scale, would be essential to confirm the broader impact and cost-effectiveness of integrating PedAMINES across various health care settings.

Conclusions

This study highlights PedAMINES as a consumer-grade, cost-effective mobile app designed to enhance medication safety, particularly in pediatric emergency care. This demonstrates its ability to swiftly achieve a positive ROI and underscore its potential economic value, especially if expanded to include other medications. However, the interpretation of the findings is constrained by the limitations inherent in literature review and simulation-based methodologies. Moving forward, it is crucial to validate in real-world settings whether the reduction in medication errors and ADEs achieved with PedAMINES translates into tangible economic benefits and improved patient outcomes. This empirical validation will provide deeper insights into the app's practical impact and its potential role in enhancing pediatric health care delivery.

Acknowledgments

This study was funded by the Swiss National Science Foundation (grant SNSF_32003B_182374). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

We would like to thank nurse Yann Levy-Jamet, Application Manager at Geneva University Hospitals, for generously supplying us with institutional data concerning drug administration within the pediatric intensive care unit.

Conflicts of Interest

The PedAMINES (Pediatric Accurate Medication in Emergency Situations) app is not available on the Google Play Store or the Apple App Store. All authors declare no competing interests. The PedAMINES mobile app is owned by Geneva University Hospitals and is not currently available for commercial use. As employees of Geneva University Hospitals, JNS, FE, and SM may receive institutional benefits if the app is commercialized in the future.

Multimedia Appendix 1
Average increase in cost weight in the cardiovascular category, 2019 (Swiss Diagnosis Related Groups).
[DOCX File , 22 KB - jmir_v26i1e52077_app1.docx]

Multimedia Appendix 2
Variables and distributions used in the analysis.
[DOCX File , 29 KB - jmir_v26i1e52077_app2.docx]

Multimedia Appendix 3
Sensitivity analysis with device cost added.
[DOCX File , 20 KB - jmir_v26i1e52077_app3.docx]

Multimedia Appendix 4
Sensitivity analysis, cumulative frequencies: norepinephrine, midazolam, and dopamine.
[DOCX File , 31 KB - jmir_v26i1e52077_app4.docx]

Multimedia Appendix 5
Deterministic sensitivity analysis (tornado plot): norepinephrine, midazolam, and dopamine.
[DOCX File , 33 KB - jmir_v26i1e52077_app5.docx]

Multimedia Appendix 6
Review of the economic evidence.
[DOCX File , 26 KB - jmir_v26i1e52077_app6.docx]

XSL•FO
RenderX

References

1. Ahsani-Estahbanati E, Sergeevich Gordeev V, Doshmangir L. Interventions to reduce the incidence of medical error and its financial burden in health care systems: a systematic review of systematic reviews. Front Med (Lausanne) 2022;9:875426 [FREE Full text] [doi: 10.3389/fmed.2022.875426] [Medline: 35966854]
2. Henry Basil J, Mohd Tahir NA, Menon Premakumar C, Mhd Ali A, Seman Z, Ishak S, et al. Clinical and economic impact of medication administration errors among neonates in neonatal intensive care units. PLoS One 2024;19(7):e0305538 [FREE Full text] [doi: 10.1371/journal.pone.0305538] [Medline: 38990851]
3. Leporini C, de Sarro C, Palleria C, Caccavo I, Piro B, Citraro R, et al. Pediatric drug safety surveillance: a 10-year analysis of adverse drug reaction reporting data in Calabria, Southern Italy. Drug Saf 2022;45(11):1381-1402 [FREE Full text] [doi: 10.1007/s40264-022-01232-w] [Medline: 36112324]
4. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patientsxcess length of stay, extra costs, and attributable mortality. Jama 1997;277(4):301-306. [doi: 10.1001/jama.277.4.301]
5. Formica D, Sultana J, Cutroneo PM, Lucchesi S, Angelica R, Crisafulli S, et al. The economic burden of preventable adverse drug reactions: a systematic review of observational studies. Expert Opin Drug Saf 2018;17(7):681-695. [doi: 10.1080/14740338.2018.1491547] [Medline: 29952667]
6. Tansuwannarat P, Vichiensanth P, Sivarak O, Tongpoo A, Promrungsri P, Sriapha C, et al. Characteristics and consequences of medication errors in pediatric patients reported to ramathibodi poison center: a 10-year retrospective study. Ther Clin Risk Manag 2022;18:669-681 [FREE Full text] [doi: 10.2147/TCRM.S363638] [Medline: 35795851]
7. Badgery-Parker T, Li L, Fitzpatrick E, Mumford V, Raban MZ, Westbrook JI. Child age and risk of medication error: a multisite children's hospital study. J Pediatr 2024;272:114087 [FREE Full text] [doi: 10.1016/j.jpeds.2024.114087] [Medline: 38705229]
8. Henry Basil J, Premakumar C, Mhd Ali A, Mohd Tahir NA, Mohamed Shah N. Prevalence, causes and severity of medication administration errors in the neonatal intensive care unit: a systematic review and meta-analysis. Drug Saf 2022;45(12):1457-1476. [doi: 10.1007/s40264-022-01236-6] [Medline: 36192535]
9. Alghamdi AA, Keers RN, Sutherland A, Ashcroft DM. Prevalence and nature of medication errors and preventable adverse drug events in paediatric and neonatal intensive care settings: a systematic review. Drug Saf 2019;42(12):1423-1436 [FREE Full text] [doi: 10.1007/s40264-019-00856-9] [Medline: 31410745]
10. Alsabri M, Eapen D, Sabesan V, Tarek Hassan Z, Amin M, Elshanbary AA, et al. Medication errors in pediatric emergency departments: a systematic review and recommendations for enhancing medication safety. Pediatr Emerg Care 2024;40(1):58-67. [doi: 10.1097/PEC.0000000000003108] [Medline: 38157396]
11. Agarwal S, Classen D, Larsen G, Tofil NM, Hayes LW, Sullivan JE, et al. Prevalence of adverse events in pediatric intensive care units in the United States. Pediatr Crit Care Med 2010;11(5):568-578. [doi: 10.1097/PCC.0b013e3181d8e405] [Medline: 20308932]
12. Larsen GY, Donaldson AE, Parker HB, Grant MJC. Preventable harm occurring to critically ill children. Pediatr Crit Care Med 2007;8(4):331-336. [doi: 10.1097/01.PCC.0000263042.73539.99] [Medline: 17417126]
13. Seidling HM, Bates DW. Evaluating the impact of health IT on medication safety. Stud Health Technol Inform 2016;222:195-205. [doi: 10.3233/978-1-61499-635-4-195]
14. Hutton K, Ding Q, Wellman G. The effects of bar-coding technology on medication errors: a systematic literature review. J Patient Saf 2021;17(3):e192-e206. [doi: 10.1097/PTS.0000000000000366] [Medline: 28234729]
15. Ahtiainen HK, Kallio MM, Airaksinen M, Holmström AR. Safety, time and cost evaluation of automated and semi-automated drug distribution systems in hospitals: a systematic review. Eur J Hosp Pharm 2020;27(5):253-262 [FREE Full text] [doi: 10.1136/ejhpharm-2018-001791] [Medline: 32839256]
16. Kahn S, Abramson EL. What is new in paediatric medication safety? Arch Dis Child 2019;104(6):596-599. [doi: 10.1136/archdischild-2018-315175] [Medline: 30154183]
17. Ehrler F, Siebert JN. PedAMINES: a disruptive mHealth app to tackle paediatric medication errors. Swiss Med Wkly 2020;150:w20335 [FREE Full text] [doi: 10.4414/smw.2020.20335] [Medline: 32920794]
18. Rinke ML, Bundy DG, Velasquez CA, Rao S, Zerhouni Y, Lobner K, et al. Interventions to reduce pediatric medication errors: a systematic review. Pediatrics 2014;134(2):338-360. [doi: 10.1542/peds.2013-3531] [Medline: 25022737]
19. Ruano M, Villamañán E, Pérez E, Herrero A, Álvarez-Sala R. New technologies as a strategy to decrease medication errors: how do they affect adults and children differently? World J Pediatr 2016;12(1):28-34. [doi: 10.1007/s12519-015-0067-6] [Medline: 26684316]
20. Siebert JN, Bloudeau L, Combescure C, Haddad K, Hugon F, Suppan L, et al. Effect of a mobile app on prehospital medication errors during simulated pediatric resuscitation: a randomized clinical trial. JAMA Netw Open 2021;4(8):e2123007 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.23007] [Medline: 34459905]
21. Siebert JN, Ehrler F, Combescure C, Lovis C, Haddad K, Hugon F, et al. A mobile device application to reduce medication errors and time to drug delivery during simulated paediatric cardiopulmonary resuscitation: a multicentre, randomised, controlled, crossover trial. Lancet Child Adolesc Health 2019;3(5):303-311. [doi: 10.1016/S2352-4642(19)30003-3] [Medline: 30797722]

XSL•FO
RenderX

22. Bougouin W, Slimani K, Renaudier M, Binois Y, Paul M, Dumas F, et al. Epinephrine versus norepinephrine in cardiac arrest patients with post-resuscitation shock. Intensive Care Med 2022;48(3):300-310. [doi: 10.1007/s00134-021-06608-7] [Medline: 35129643]
23. Ren Y, Xing T, Chen X, Chai X. Research on software maintenance cost of influence factor analysis and estimation method. 2011 Presented at: Proceedings of the 2011 3rd International Workshop on Intelligent Systems and Applications; May 28-29, 2011; Wuhan, China. [doi: 10.1109/isa.2011.5873461]
24. Singh C, Sharma N, Kumar N. Analysis of software maintenance cost affecting factors and estimation models. International Journal of Scientific & Technology Research 2019;8:276-281 [FREE Full text]
25. Lumley HA, Flynn D, Shaw L, McClelland G, Ford GA, White PM, et al. A scoping review of pre-hospital technology to assist ambulance personnel with patient diagnosis or stratification during the emergency assessment of suspected stroke. BMC Emerg Med 2020;20(1):1-21 [FREE Full text] [doi: 10.1186/s12873-020-00323-0] [Medline: 32336270]
26. Pulsiri N, Vatananan-Thesenvitz R, Sirisamutr T, Wachiradilokditors P. Save lives: a review of ambulance technologies in pre-hospital emergency medical services. 2019 Presented at: Proceedings of the 2019 Portland International Conference on Management of Engineering and Technology (PICMET); August 25-29, 2019; Portland, OR. [doi: 10.23919/picmet.2019.8893973]
27. Roumeliotis N, Pullenayegum E, Rochon P, Taddio A, Parshuram C. A modified delphi to define drug dosing errors in pediatric critical care. BMC Pediatr 2020;20(1):488 [FREE Full text] [doi: 10.1186/s12887-020-02384-3] [Medline: 33087096]
28. Takata GS, Mason W, Taketomo C, Logsdon T, Sharek PJ. Development, testing, and findings of a pediatric-focused trigger tool to identify medication-related harm in US children's hospitals. Pediatrics 2008;121(4):e927-e935. [doi: 10.1542/peds.2007-1779] [Medline: 18381521]
29. Ji HH, Song L, Xiao JW, Guo YX, Wei P, Tang TT, et al. Adverse drug events in Chinese pediatric inpatients and associated risk factors: a retrospective review using the Global Trigger Tool. Sci Rep 2018;8(1):2573 [FREE Full text] [doi: 10.1038/s41598-018-20868-2] [Medline: 29416072]
30. Cohen AL, Budnitz DS, Weidenbach KN, Jernigan DB, Schroeder TJ, Shehab N, et al. National surveillance of emergency department visits for outpatient adverse drug events in children and adolescents. J Pediatr 2008;152(3):416-421. [doi: 10.1016/j.jpeds.2007.07.041] [Medline: 18280852]
31. Sikdar KC, Alaghehbandan R, Macdonald D, Barrett B, Collins KD, Gadag V. Adverse drug events among children presenting to a hospital emergency department in Newfoundland and Labrador, Canada. Pharmacoepidemiol Drug Saf 2010;19(2):132-140. [doi: 10.1002/pds.1900] [Medline: 20033910]
32. Kaushal R, Bates DW, Landrigan C, McKenna KJ, Clapp MD, Federico F, et al. Medication errors and adverse drug events in pediatric inpatients. JAMA 2001;285(16):2114-2120. [doi: 10.1001/jama.285.16.2114] [Medline: 11311101]
33. Oderda GM, Said Q, Evans RS, Stoddard GJ, Lloyd J, Jackson K, et al. Opioid-related adverse drug events in surgical hospitalizations: impact on costs and length of stay. Ann Pharmacother 2007;41(3):400-406. [doi: 10.1345/aph.1H386] [Medline: 17341537]
34. Amelung S, Meid AD, Nafe M, Thalheimer M, Hoppe-Tichy T, Haefeli WE, et al. Association of preventable adverse drug events with inpatients' length of stay-a propensity-matched cohort study. Int J Clin Pract 2017;71(10):e12990. [doi: 10.1111/ijcp.12990] [Medline: 28873271]
35. Silva DCB, Araujo OR, Arduini RG, Alonso CFR, Shibata ARO, Troster EJ. Adverse drug events in a paediatric intensive care unit: a prospective cohort. BMJ Open 2013;3(2):e001868 [FREE Full text] [doi: 10.1136/bmjopen-2012-001868] [Medline: 23427200]
36. Silber JH, Rosenbaum PR, Even-Shoshan O, Shabbout M, Zhang X, Bradlow E, et al. Length of stay, conditional length of stay, and prolonged stay in pediatric asthma. Health Serv Res 2003;38(3):867-886 [FREE Full text] [doi: 10.1111/1475-6773.00150] [Medline: 12822916]
37. Espinosa O, Rodríguez-Lesmes P, Romano G, Orozco E, Basto S, Ávila D, et al. Use of cost-effectiveness thresholds in healthcare public policy: progress and challenges. Appl Health Econ Health Policy 2024. [doi: 10.1007/s40258-024-00900-5] [Medline: 38995492]
38. Zio E, Zio E. Monte Carlo Simulation: The Method. London, United Kingdom: Springer; 2013.
39. IMF. Representative exchange rates for selected currencies for December 2019. International Monetary Fund. URL: https://www.imf.org/external/np/fin/data/rms_mth.aspx?SelectDate=2019-12-31&reportType=REP [accessed 2024-01-17]
40. Iwasaki H, Sakuma M, Ida H, Morimoto T. The burden of preventable adverse drug events on hospital stay and healthcare costs in Japanese pediatric inpatients: the JADE study. Clin Med Insights Pediatr 2021;15:1179556521995833 [FREE Full text] [doi: 10.1177/1179556521995833] [Medline: 33746523]
41. Meier F, Maas R, Sonst A, Patapovas A, Müller F, Plank-Kiegele B, et al. Adverse drug events in patients admitted to an emergency department: an analysis of direct costs. Pharmacoepidemiol Drug Saf 2015;24(2):176-186. [doi: 10.1002/pds.3663] [Medline: 24934134]
42. Riaz M, Brown JD. Association of adverse drug events with hospitalization outcomes and costs in older adults in the USA using the nationwide readmissions database. Pharmaceut Med 2019;33(4):321-329. [doi: 10.1007/s40290-019-00286-z] [Medline: 31933184]

XSL•FO
RenderX

43. Cheng A, Auerbach M, Hunt EA, Chang TP, Pusic M, Nadkarni V, et al. Designing and conducting simulation-based research. Pediatrics 2014;133(6):1091-1101. [doi: 10.1542/peds.2013-3267] [Medline: 24819576]
44. Hobgood C, Bowen JB, Brice JH, Overby B, Tamayo-Sarver JH. Do EMS personnel identify, report, and disclose medical errors? Prehosp Emerg Care 2006;10(1):21-27. [doi: 10.1080/10903120500366011] [Medline: 16418087]
45. Stratton KM, Blegen MA, Pepper G, Vaughn T. Reporting of medication errors by pediatric nurses. J Pediatr Nurs 2004;19(6):385-392. [doi: 10.1016/j.pedn.2004.11.007] [Medline: 15637579]
46. Taylor JA, Brownstein D, Christakis DA, Blackburn S, Strandjord TP, Klein EJ, et al. Use of incident reports by physicians and nurses to document medical errors in pediatric patients. Pediatrics 2004;114(3):729-735. [doi: 10.1542/peds.2003-1124-L] [Medline: 15342846]
47. Siebert JN, Gosetto L, Sauvage M, Bloudeau L, Suppan L, Rodieux F, et al. Usability testing and technology acceptance of an mHealth app at the point of care during simulated pediatric in- and out-of-hospital cardiopulmonary resuscitations: study nested within 2 multicenter randomized controlled trials. JMIR Hum Factors 2022;9(1):e35399 [FREE Full text] [doi: 10.2196/35399] [Medline: 35230243]
48. Manrique-Rodríguez S, Sánchez-Galindo AC, López-Herce J, Calleja-Hernández M, Martínez-Martínez F, Iglesias-Peinado I, et al. Implementing smart pump technology in a pediatric intensive care unit: a cost-effective approach. Int J Med Inform 2014;83(2):99-105. [doi: 10.1016/j.ijmedinf.2013.10.011] [Medline: 24296271]
49. Silva RCLD, QuinellatoLouro T, Peregrino AADF, Silva CRLD, Marta CB, Itria A. Cost-effectiveness of infusion pumps to reduce errors in a Pediatric ICU. Rev Bras Enferm 2019;72(3):617-623 [FREE Full text] [doi: 10.1590/0034-7167-2018-0526] [Medline: 31269124]
50. Jessurun JG, Hunfeld NGM, van Dijk M, van den Bemt PMLA, Polinder S. Cost-effectiveness of central automated unit dose dispensing with barcode-assisted medication administration in a hospital setting. Res Social Adm Pharm 2022;18(11):3980-3987 [FREE Full text] [doi: 10.1016/j.sapharm.2022.07.006] [Medline: 35853809]
51. Sutherland A, Gerrard WS, Patel A, Randall M, Weston E. The impact of drug error reduction software on preventing harmful adverse drug events in England: a retrospective database study. BMJ Open Qual 2022;11(3):e001708 [FREE Full text] [doi: 10.1136/bmjoq-2021-001708] [Medline: 35820711]

Abbreviations

ADE: adverse drug event
CPR: cardiopulmonary resuscitation
HIT: health information technology
HUG: Geneva University Hospital
ICER: incremental cost-effectiveness ratio
LOS: length of stay
MC: Monte Carlo
PED: pediatric emergency department
PedAMINES: Pediatric Accurate Medication in Emergency Situations
PICU: pediatric intensive care unit
ROI: return on investment

Edited by A Coristine, A Mavragani; submitted 22.08.23; peer-reviewed by R Rivera, C Asuzu, J Walsh, CM Premakumar; comments to author 09.07.24; revised version received 19.07.24; accepted 20.08.24; published 25.10.24.

Please cite as:
Brunner L, Siebert JN, Ehrler F, Manzano S, Marti J
Evaluating the Economic Impact of the PedAMINES App in Reducing Medication Errors in Pediatric Emergency Care: Cost-Effectiveness Analysis
J Med Internet Res 2024;26:e52077
URL: https://www.jmir.org/2024/1/e52077
doi:10.2196/52077
PMID:

©Loïc Brunner, Johan N Siebert, Frédéric Ehrler, Sergio Manzano, Joachim Marti. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Typology of Social Media Use by Human Service Nonprofits: Mixed Methods Study

Jia Xue[1,2], PhD; Micheal L Shier[1], PhD; Junxiang Chen[3], PhD; Yirun Wang[4], MSc; Chengda Zheng[4], MI; Chen Chen[4], PhD

[1]Factor-Inwentash Faculty of Social Work, University of Toronto, Toronto, ON, Canada

[2]Faculty of Information, University of Toronto, Toronto, ON, Canada

[3]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States

[4]Artificial Intelligence for Justice Lab, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Jia Xue, PhD
Factor-Inwentash Faculty of Social Work
University of Toronto
246 Bloor Street West
Toronto, ON, M5S 1V4
Canada
Phone: 1 4169465429
Email: jia.xue@utoronto.ca

Abstract

Background: Nonprofit organizations are increasingly using social media to improve their communication strategies with the broader population. However, within the domain of human service nonprofits, there is hesitancy to fully use social media tools, and there is limited scope among organizational personnel in applying their potential beyond self-promotion and service advertisement. There is a pressing need for greater conceptual clarity to support education and training on the varied reasons for using social media to increase organizational outcomes.

Objective: This study leverages the potential of Twitter (subsequently rebranded as X [X Corp]) to examine the online communication content within a sample (n=133) of nonprofit sexual assault (SA) centers in Canada. To achieve this, we developed a typology using a qualitative and supervised machine learning model for the automatic classification of tweets posted by these centers.

Methods: Using a mixed methods approach that combines machine learning and qualitative analysis, we manually coded 10,809 tweets from 133 SA centers in Canada, spanning the period from March 2009 to March 2023. These manually labeled tweets were used as the training data set for the supervised machine learning process, which allowed us to classify 286,551 organizational tweets. The classification model based on supervised machine learning yielded satisfactory results, prompting the use of unsupervised machine learning to classify the topics within each thematic category and identify latent topics. The qualitative thematic analysis, in combination with topic modeling, provided a contextual understanding of each theme. Sentiment analysis was conducted to reveal the emotions conveyed in the tweets. We conducted validation of the model with 2 independent data sets.

Results: Manual annotation of 10,809 tweets identified seven thematic categories: (1) community engagement, (2) organization administration, (3) public awareness, (4) political advocacy, (5) support for others, (6) partnerships, and (7) appreciation. Organization administration was the most frequent segment, and political advocacy and partnerships were the smallest segments. The supervised machine learning model achieved an accuracy of 63.4% in classifying tweets. The sentiment analysis revealed a prevalence of neutral sentiment across all categories. The emotion analysis indicated that fear was predominant, whereas joy was associated with the partnership and appreciation tweets. Topic modeling identified distinct themes within each category, providing valuable insights into the prevalent discussions surrounding SA and related issues.

Conclusions: This research contributes an original theoretical model that sheds light on how human service nonprofits use social media to achieve their online organizational communication objectives across 7 thematic categories. The study advances our comprehension of social media use by nonprofits, presenting a comprehensive typology that captures the diverse communication objectives and contents of these organizations, which provide content to expand training and education for nonprofit leaders to connect and engage with the public, policy experts, other organizations, and potential service users.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51698) doi:10.2196/51698

KEYWORDS

human service nonprofits; sexual assault support centers; Canada; typology; theory; Twitter; machine learning; social media; tweet; tweets; nonprofit; nonprofits; crisis; sexual assault; sexual violence; sexual abuse; support center; support centers; communication; communications; organization; organizations; organizational; sentiment analysis; business; marketing

Introduction

Background

It has long been acknowledged that social media plays a significant role in facilitating stakeholder engagement between nonprofits and community members [1-3]. Human service nonprofits have recognized the potential of social media in securing donations; recruiting volunteers [4-7]; enhancing trust, accountability, and awareness [8]; and fostering partnerships [9]. However, research on the specific focus of social media engagement by human service nonprofits remains somewhat limited in the existing literature and practice [10]. Traditionally, social media in the nonprofit sector has been extensively explored in relation to its use for advocacy purposes [11-13]. Although investigation into the advocacy function of social media use is important within the human service nonprofit sector, as this is a key role played by human service nonprofits to promote and support social welfare development, the need of these organizations to engage with the wider community is much more expansive.

Human service organizations are complex entities involved in a wide range of activities to fulfill their missions, primarily focused on providing direct support to address negative social, economic, and political outcomes of social groups considered marginalized and to promote social welfare development through program development, public awareness, and advocacy efforts [14]. These human service nonprofits interact with diverse human resources (professional and volunteer), service users, and community groups; form partnerships across sectors (nonprofits, for-profit firms, and governments); and manage activities with for-profit (eg, social enterprises, social investors, and consumers), government (eg, contracting arrangements), and nonprofit (eg, through foundations) revenue sources and organizations. The use of social media is complicated further when considering its use for service user engagement, as there is an emerging body of literature on the use of social media for service delivery–related purposes [15,16].

To move beyond the advocacy-related function of social media by human service nonprofits, this research investigated in greater detail the various reasons why human service nonprofits are using social media within the sexual violence service delivery sector across Canada. The research merges social science, big data, and computer science to further enhance our knowledge and understanding of how human service nonprofit organizations use information communication technology. This study expands upon prior studies of nonprofit organizational communication research by using Twitter-based data (subsequently rebranded as X [X Corp]). Human-labeled tweets were used as training data, and a supervised machine learning approach was used to automatically predict content analytical themes in a Twitter corpus. The study builds a predictive classification model that uses a supervised machine learning algorithm to evaluate large social media data sets, resulting in a theoretical framework that categorizes the objectives of the sexual assault (SA) organization posts on social media. The overarching question that guides this research is as follows: “What are the different purposes of social media communication among SA centers in Canada?” This research is part of a greater effort to develop a strategic approach and educational information for training human service personnel and leaders on the use of social media to increase the capacity of human service nonprofits [17,18].

Literature Review

Current research indicates that nonprofit organizations are increasingly using social media to improve their communication strategies with the broader population. A primary focus of research in this area has been on the specific tangible ways of this type of engagement, including the volume of engagement and the focus of messaging, along with its directionality, and the emphasis of the posts being informative and practical [19-21]. For example, Guo and Saxton [22] have focused on the extent to which nonprofits are gaining attention and highlight that this is influenced by the size of an organization’s network, the frequency with which it communicates through social media, and the number of conversations an organization joins [22]. This research is important, as it highlights the mechanisms of social media use and the frequency; however, it does not provide sufficient insight into the various reasons for social media use and the outcomes of this communication strategy on different organizational functions or purposes, and particularly important within the realm of human service nonprofit organization, which may use social media to achieve a multitude of objectives.

In fact, research on social media use within human service nonprofits specifically has identified some hesitancy to use social media education or useful tools to focus on social media use [23,24], and there is limited scope among organizational personnel in applying its usefulness beyond promoting one’s organization and its services [25]. This lack of engagement has been determined to be influenced in part due to limited education and awareness of the utility of social media use in the human services sector and other key organizational dynamics such as organizational culture, funding, and size of the organization [6,24,26,27].

Furthermore, a strong focus within the literature has been on how social media has been impacted by market actors (such as donors), which has constrained the framing of social media messaging [20,28-30]. Likewise, challenges with social media use, such as breaches of confidentiality and its increased use for surveillance and accountability-related purposes [31], also act to constrain social media use. As a result, there is a need for greater conceptual clarity to support education and training on

XSL•FO
RenderX

the varied reasons for using social media to increase organizational outcomes [32-34].

This research seeks to address these gaps by investigating the wider range of social media use by human service nonprofits, establishing a typology of reasons for social media use beyond advocacy-related purposes. By doing so, it also addresses concerns regarding limited education and training within the sector on leveraging social media for diverse organizational objectives. Through the incorporation of machine learning and content analysis, this study contributes to a deeper understanding of nonprofit communication strategies and offers practical implications for improved social media engagement within the human service context.

Aim of the Study

This study investigates the objectives of social media engagement and the contents posted by human service nonprofit organizations on the social media platform Twitter, with a particular focus on SA service delivery centers in Canada. To achieve this aim, this study addresses the following research questions: (1) What is the typology and theoretical framework that effectively captures and categorizes the diverse online organizational communication objectives of SA centers as they use Twitter as a strategic tool to achieve their organizational outcomes? (2) How do the sentiments and emotions expressed in Twitter posts by SA centers vary in relation to different categories in the typology of online organizational communication, such as advocacy or public awareness? (3) How can machine learning and content analysis categorize and analyze the social media posts of these organizations, providing insights into their communication strategies?

Methods

Overview

This study used mixed research methods, including qualitative content analysis, supervised machine learning, unsupervised machine learning, thematic analysis, and sentiment analysis. To classify the full set of tweets, we first manually coded a subset of the full data set (n=10,809 tweets) into 7 emergent categories (Table 1). These human-labeled tweets were used as the training data set to train a supervised machine learning algorithm to classify the remaining tweets. Figure 1 illustrates the mixed methods approach.

Table 1. Coding protocol.

Label and classification	Definitions	Tweet samples
1. Community engagement	These tweets involve nonprofits engaging with their community beyond their primary mission. This may include sharing well-wishes, updates on the organization's activities, quotes or information about books, and other resources that the community might find valuable. Simply providing information about resources is a way to foster engagement.	RT @KidsHelpPhone: "Great article by @OmarMosleh about dealing w/teen grief. Includes quotes from @KidsHelpPhone http://bit.ly/9fqcJn"
2. Organization administration	These tweets pertain to the operation of the organization, such as its operating hours, recruitment of employees and volunteers, organization efforts, information about its activities, and soliciting donations.	"We'll be closing down the office early in preparation for New Year's Eve!"
3. Public awareness	These tweets contain information about social, environmental, or economic issues that require attention as well as causes or educational content concerning these issues or causes.	"In North America, 7% of women and 6% of men end up abused by their current or former partners."
4. Political advocacy	These tweets concern public policy or efforts or initiatives aimed at promoting more far-reaching social change. Based on charitable status, they may not be able to act politically. These tweets may involve the organization's efforts to draw attention to political issues or comment on government responses referring to what the government is doing to respond to the disaster.	RT @CKNW: "Changes coming to BC family law: Changes are being made to modernize BC's family law system. Attorney Genera..."
5. Support for others	These tweets pertain to the endorsement of local businesses. Support for others implies backing the initiatives of others that are not directly connected to one's own mission.	RT @RichmondYouth: "U-Connect Crew is giving back to the community w 3rd annual Winter Holiday Dinner! Volunteer or donate http://ow.ly/NbXP"
6. Partnerships	These tweets relate to the organization's endeavors to form partnerships and aid other organizations. Partnership involves collaborating with other organizations to further one's own mission or objective.	@gilmorepark "Happy New Year! Would love to chat with Rev Anna & send over some Outreach & Advocacy brochures for your community suppers."
7. Appreciation	These tweets concern expressing gratitude toward donors or volunteers for their contributions and dedication toward the organization.	"Best Buy in Richmond has donated a brand new TV to replace the one that broke at Nova Transition House! 3 cheers for @BBYCanada! Thank you!"

XSL•FO
RenderX

Figure 1. Mixed methods approach. API: application programming interface; LDA: latent Dirichlet allocation; SA: sexual assault.

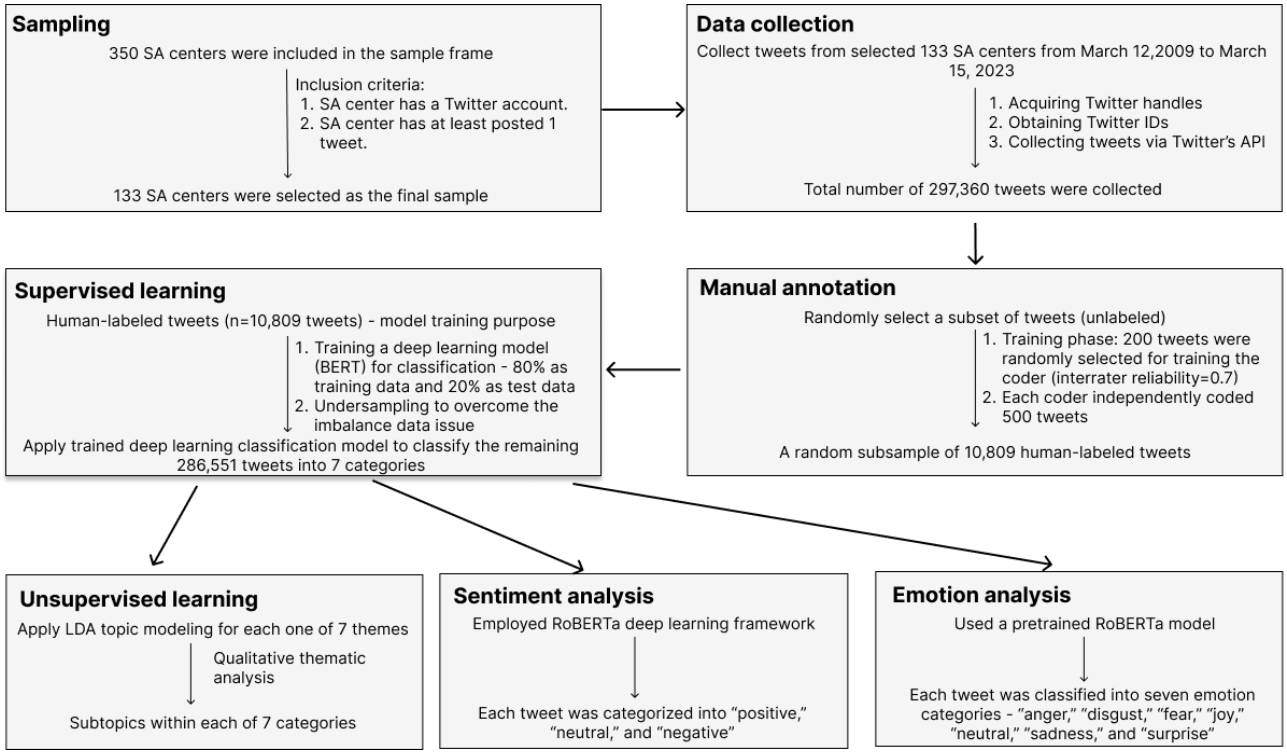

Sampling

To select SA centers in Canada, this study used a purposive sampling approach. Initially, a sampling frame was developed by combining the list of SA centers by province and territory from the Canadian Association of Sexual Assault Centres and the Sexual Assault Centres, Crisis Lines, and Support Services websites. After removing duplicates, the sample frame consisted of 350 SA centers across 10 provinces and 3 territories. The sample frame provided basic information about the centers, including their names, contact information (phone number and email), and website or URL. The inclusion criteria were twofold: (1) the SA center had an active Twitter account and (2) it had posted at least 1 tweet on its account. To verify the eligibility of these centers, the authors manually searched their home page and Twitter pages and conducted thorough Google searches. Ultimately, the Twitter accounts of 133 SA centers were included as the final sample for this study. These centers were from 9 provinces and the Northwest Territories (Prince Edward Island did not have any SA centers that used Twitter).

Data Collection

To collect tweets from SA centers, the authors followed the pipeline outlined in their papers, including acquiring Twitter handles, obtaining Twitter IDs, and collecting tweets via Twitter's application programing interface (API) [35-43]. The collected tweets encompassed the period from March 12, 2009, to March 15, 2023. The data set consisted of 297,360 tweets from 133 SA centers in Canada. The data sets are available for use by researchers upon request. First, a total of 91 unique Twitter handles (ie, @name) were obtained from the 133 SA centers in the sample, with 26 duplicate Twitter handles. Second, the 91 Twitter handles were converted into 91 Twitter IDs using 3 websites: TweeterID, CodeOfaNinja, and Comment Picker. Third, Twitter's premium search API and timeline end points (full-archive end point) were used to collect tweets posted by the sampled SA centers in Canada, starting from as early as 2006 (search tweets, 2019). Data collection concluded on March 15, 2023.

Manual Annotation

The purpose of manual annotation was to obtain human-labeled tweets categorized into different themes. These labeled tweets would serve as the training data set for classifying the entire corpus using a supervised machine learning approach. The coding protocol was developed based on prior literature on organizational communication research and adapted to suit the objectives of this study. Table 1 included the classification, labels, definitions, and sample tweets.

To ensure consistency, 2 authors (JX and MLS) provided training to the research assistants on the protocol and research goals. During the training phase, a random subset of 200 tweets was selected, and 2 research assistants were assigned to independently code them. This process was repeated 4 times (n=809 tweets) until an acceptable interrater reliability score of 0.7 was achieved for each of the 7 categories. Krippendorff α was used to determine the interrater reliability, which indicated substantial agreement.

Following the training phase, a subset of 10,000 tweets was randomly selected from the collected data. Research assistants were assigned to independently code a subset of 5000 tweets. The manual annotation data set consisted of a random subsample of 10,809 manually labeled tweets categorized into 7 themes from the full data set.

XSL•FO
RenderX

Construction of Predictive Classification Model

To create an accurate classification model for Twitter data, we used the BERT model [44]. BERT is a widely used natural language processing model that has been pretrained on various English language data sets, making it suitable for fine-tuning tasks such as sentence classification. To evaluate the performance of our model, we randomly selected 80% of the human-labeled tweets as training data, with the remaining 20% used as test data.

Due to the imbalanced distribution of classes in our data set, we used a 2-step strategy to train the machine learning model. First, we fine-tuned the BERT model with all the training data by minimizing the logistic loss [45]. Second, we applied a random undersampling process [46] to retrain the last layer of the BERT model (the classification layer) using this undersampled subset. The undersampling process randomly selected a subset of training data, ensuring an equal number of samples for each class. We chose the undersampling technique as it is less prone to overfitting the data compared to other methods such as oversampling [46].

In addition to using deep learning models in our study, we also used a range of traditional machine learning algorithms as benchmarks for performance comparison. Specifically, we trained models using linear regression, support vector machines with a radial basis function kernel, and support vector machine with a linear kernel. To represent features in these traditional algorithms, we chose the term frequency–inverse document frequency approach to convert our textual data into numerical vectors.

To evaluate the efficacy of these models, we computed the average sensitivity score based on the test data. The sensitivity score for a given class “k” denotes the probability that a sample will be classified by a model as belonging to class “k,” given that the sample truly belongs to that class. We calculated the mean of the sensitivity scores across all 7 classes as our final measurement. Following the training of the BERT model, we used it to classify the unlabeled 286,551 tweets into 7 categories.

Validation Data

To ensure the robust performance of our model across diverse contexts, we gathered 2 distinct independent data sets from Twitter and Facebook. Independent data set #1 was derived using the same sampling frame in this study. Our aim was to identify organizations active on Facebook but not on Twitter, thereby maintaining uniformity in organization type while varying the social media platform for further model validation. Using Apify software [47], we collected messages from 67 SA organizations and subsequently selected a random sample of 500 messages (n=2520). Independent data set #2 was obtained through a list of human service organizations (approximately 12,000) from the government of Canada’s list of charitable nonprofits (N=85,496). Of the approximately 86,000 charitable nonprofits in Canada, the list of human service organizations was developed through an assessment of the organizations’ website that shows an indication of providing some type of social service programing to a service user group. This frame enabled the identification of organizations with active Twitter accounts, thus ensuring consistency in the chosen social media platform while introducing variation in the type of organization for enhanced model validation. Following the collection of tweets via our API, a random sample of 500 tweets (n=15,696) was selected for data validation. We used the same manual annotation procedure for these 2 data sets to establish manual labels. This allowed us to directly compare the model’s predictions against these manual labels, serving as a method to assess the model’s effectiveness (Multimedia Appendix 1).

Sentiment Analysis

Sentiment analysis, sometimes referred to as opinion mining, involves the classification and analysis of people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions concerning various entities, including products, services, organizations, individuals, issues, events, topics, and their associated attributes [48]. Sentiment analysis applied to social media content has been extensively studied, and Twitter has the capability to promptly gauge public sentiments and emotions regarding a given topic [49]. For this analysis, we used RoBERTa, a deep learning framework [50]. We used a pretrained model [51] that was fine-tuned specifically for sentiment analysis of social media data. The model categorized each tweet into 1 of the 3 sentiments: “positive,” “neutral,” or “negative.” We converted a significant amount of textual data into quantitative sentiment scores and calculated the percentage of each sentiment within every category.

Emotion Analysis

Emotion analysis primarily focuses on capturing nuanced emotions, which contrasts with sentiment analysis, primarily concerned with detecting simple attitudes such as positivity and negativity. Mohammad [52] indicated that machines can infer people’s emotions in a limited way but are useful. We need to hold automatic emotion recognition systems to high standards by incorporating ethical considerations associated with each step of the detection process. In our study, we delve into emotion analysis within 7 categorized groups to investigate and compare potential emotional variations across different categories. For this analysis, we used a pretrained RoBERTa model [53], optimized for emotion analysis of social media data. This model classifies each tweet into 1 of the 7 emotion categories: “anger,” “disgust,” “fear,” “joy,” “neutral,” “sadness,” or “surprise.” We then calculated the percentage of tweets associated with each emotion category for each of our designated categories. Subsequently, we determined the percentage distribution of each emotion within each category. This measurement across various goals and intentions in social media communication provides valuable insights into the perspectives of both the public and SA issues, enhancing our overall understanding of these topics.

Topic Modeling for Tweets Categorized Into 7 Classes

The objective of this unsupervised machine learning work was to extract latent topics within each theme after categorizing tweets into 7 different categories. To achieve this, we used the latent Dirichlet allocation approach for topic modeling, which allowed us to group tweets into different topics. The initial step in this analysis stage was preprocessing, which enhanced model

XSL•FO
RenderX

performance by removing noisy data. We eliminated various elements, such as “mentions,” “emojis,” “hyperlinks,” “RT symbols,” and punctuations, and converted all tweets to lower case. Removing mention symbols eliminated irrelevant terms from the analysis, such as names of organizations and individuals. In addition, we removed stop words such as “the,” “is,” and “and” while retaining nouns, adjectives, and verbs related to events.

Once the data were preprocessed, we implemented latent Dirichlet allocation models using the Gensim library in Python. Our hyperparameter range was set from 1 to 30, and the similarity score served as our evaluation metric. By plotting the similarity score against the number of topics, we identified the turning point on the graph as our optimal hyperparameter. To further analyze the topics within each of the 7 categories, we extracted popular bigrams and reviewed a random sample of tweets. We used qualitative thematic analysis to assign underlying topic meanings to them.

Topic Evaluation

The team, consisting of domain experts and research assistants, summarized and evaluated the results of the topic modeling. Salient bigrams were used to summarize each topic, and similar topic themes were merged into higher-level categories, as per the machine learning approach described by Zhou et al [54].

Ethical Considerations

This study used publicly available Twitter data, eliminating the need for ethics approval or consent from organizations. The study data mentioned in this paper underwent processes of anonymization and deidentification. To guarantee full anonymity, all data that could potentially identify individuals or organizations, including users’ metadata and original tweets, have been carefully excluded from the data set.

Results

Overview

Our data set included 297,360 tweets and retweets from 133 SA support organizations in Canada. These tweets were posted from March 12, 2009, to March 14, 2023. Multimedia Appendix 2 illustrates a bar plot that summarizes the number of tweets collected for each year.

Manual Annotation and Class Distribution of Tweets

Among the data set consisting of 297,360 tweets and retweets, 10,809 tweets were manually annotated by humans following the coding protocol. As shown in Table 2, organization administration (5652/10,809, 52.29%) was the most frequent type of post, followed by community engagement (2322/10,809, 21.38%) and public awareness (1522/10,809, 14.08%). The smallest segment of tweets belonged to political advocacy (129/10,809, 1.19%) and partnerships (162/10,809, 1.5%). Multimedia Appendix 3 presents the histogram of class distributions of the human-labeled tweets. It is worth noting that the distribution of labels was imbalanced, with a smaller proportion of tweets (<5%) falling into the categories of political advocacy, support for others, and partnerships. Figure 2 presents a plot showing the percentage of each category plotted against the corresponding years.

Table 2. Comparison of human-annotated and machine learning classified tweets (N=297,360).

	Human annotation (n=10,809), n (%)	Supervised machine learning model (n=286,551), n (%)
Community engagement	2311 (21.38)	67,285 (23.5)
Organization administration	5652 (52.29)	50,620 (17.7)
Public awareness	1522 (14.08)	89,295 (31.2)
Political advocacy	129 (1.19)	14,504 (5.1)
Support for others	288 (2.66)	38,083 (13.3)
Partnerships	162 (1.5)	7,144 (2.5)
Appreciation	745 (6.89)	19,620 (6.8)

XSL•FO
RenderX

Figure 2. The relationship between the percentage of each category and the corresponding years.

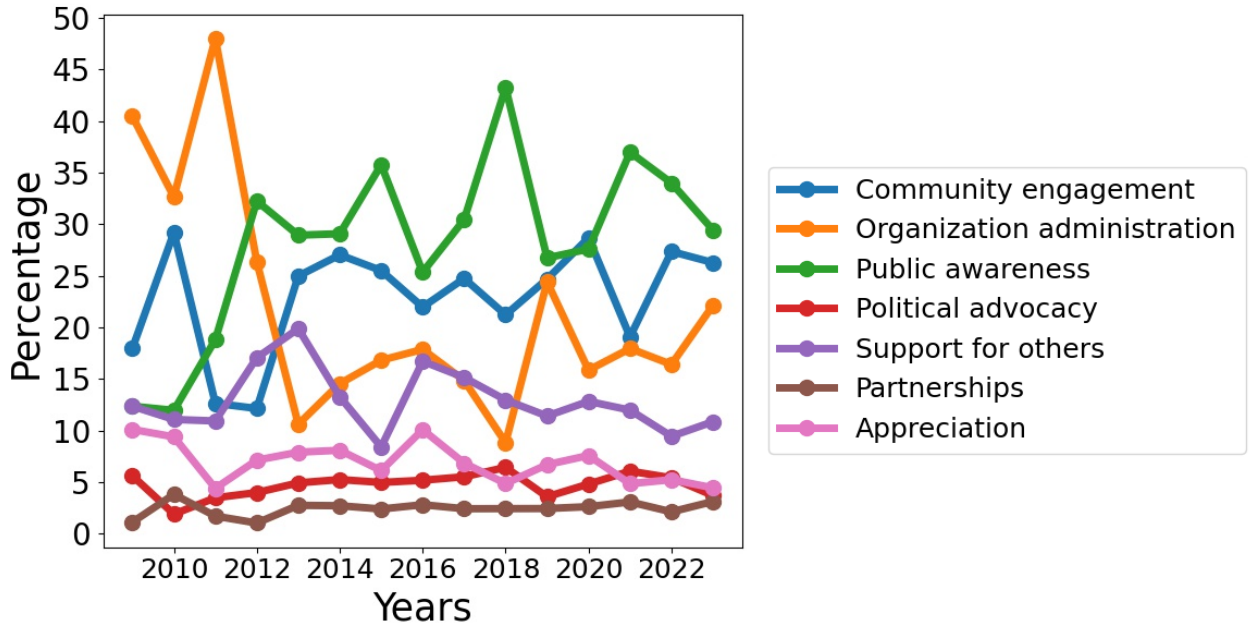

Performance of the Supervised Machine Learning Model

We evaluated the performance of the supervised machine learning model using the test set (20% of the human-labeled tweets). The accuracy of machine learning classification achieved by our trained model was 63.4%, which was higher than that of the human coders who labeled the training data set. This indicates an improvement in the BERT model's ability to accurately predict classifications compared to those made by human coders.

We analyzed and presented the confusion matrices in Tables 3 and 4. Table 3 displays the confusion matrix for the initial model without undersampling, whereas Table 4 illustrates the confusion matrix with undersampling applied. These matrices provide insights into the percentage of samples with actual labels that were correctly classified into the predicted class by the model. The sensitivity scores for each class are represented on the diagonal of the matrices, and the average sensitivity score across the 7 classes serves as an indicator of the overall performance of the model. As observed in the tables, the average sensitivity scores improved from 48.3% to 53.1%. As demonstrated in tables, BERT's performance surpassed that of traditional machine learning methods.

Table 3. Confusion matrix for the initial model without undersampling.

	Predicted classes (%)						
	Community engagement	Organization administration	Public awareness	Political advocacy	Support for others	Partnerships	Appreciation
Actual classes							
Community engagement	59.3	21.8	12.7	0.4	2.2	0.3	3.3
Organization administration	6.6	87.6	2.7	0.4	0.8	0.3	1.7
Public awareness	20.1	16.6	60.7	0.7	1.1	0.0	0.9
Political advocacy	22.6	24.2	38.7	12.9	1.6	0.0	0.0
Support for others	20.5	44.3	14.8	0.0	11.5	0.8	8.2
Partnerships	5.0	45.0	5.0	0.0	0.0	37.5	7.5
Appreciation	6.3	16.1	4.6	0.0	2.9	1.1	69.0

XSL•FO
RenderX

Table 4. Confusion matrix for initial model with undersampling.

	Predicted classes (%)						
	Community engagement	Organization administration	Public awareness	Political advocacy	Support for others	Partnerships	Appreciation
Actual classes							
Community engagement	57.0	11.4	12.7	3.6	11.4	1.6	2.3
Organization administration	8.2	75.8	3.3	1.3	7.2	2.7	1.4
Public awareness	17.4	9.1	58.9	5.1	6.8	1.5	1.1
Political advocacy	9.7	14.5	32.3	40.3	3.2	0.0	0.0
Support for others	20.5	25.4	13.1	1.6	30.3	5.7	3.3
Partnerships	2.5	20.0	0.0	0.0	10.0	55.0	12.5
Appreciation	5.7	9.8	4.0	0.6	13.2	5.2	61.5

Results of Predictive Classification of Unlabeled Data Using Machine Learning

The remaining 286,551 tweets were classified into 7 classes using the supervised machine learning algorithms. The classification results are presented in Table 2. The supervised machine learning model produced classifications that were similar to those of the human-annotated tweets in terms of the percentage of each category. The most frequent tweet class was public awareness (89,295/286,551, 31.2%), followed by community engagement (67,285/286,551, 23.5%) and organization administration (50,620/286,551, 23.5%). The 2 smallest categories of posts were partnerships (7144/286,551, 2.5%) and political advocacy (19,620/286,551, 6.8%).

Top Unigrams and Bigrams in the Tweets

We conducted an analysis to identify the most commonly used words and phrases in the tweets from the SA support organizations. To do this, we removed the stop words and generated a list of the top 30 most frequently occurring unigrams and bigrams, as presented in Multimedia Appendices 4 and 5. We observed that >220,000 tweets included a URL link, which directed users to news or events related to SA. In addition, approximately 100,000 tweets were retweets, with “rt” in the messages. The terms “women,” “support,” “sexual assault,” “sexual violence,” and “crisis line” were among the most commonly used terms in these tweets.

Sentiment and Emotion Analysis Results

We conducted sentiment and emotion analysis, and the summarized results can be found in Tables 5 and 6. The findings revealed that the neutral sentiment category surpassed both the negative and positive sentiment categories across all 7 classes.

Regarding the emotion analysis, most tweets from organizations were associated with the emotion of “fear.” In contrast, tweets discussing topics related to class 6 (partnerships) and class 7 (appreciation) exhibited the emotion of “joy.” Here are a few examples of tweets reflecting fear:

> *Every minute of every day, a Canadian woman or child is being sexually assaulted. #VAW*
>
> *Salau’s story is so symbolic of how universally disregarded, disrespected, and unprotected Black women are, even in our most vulnerable moments. #EndVAW #JusticeForToyin.*

Here is a tweet reflecting joy (appreciation):

> *We’ve seen that charity brings together amazing people to create great change and make meaningful impact in the lives of the people in their community. THANK YOU to the incredible supporters who make our work possible...*

These examples illustrate the emotional tone associated with different tweet categories, with fear being prevalent among organizational tweets and joy being linked to discussions on partnerships and appreciation.

Table 5. The sentiments for each category.

	Negative (%)	Neutral (%)	Positive (%)
Community engagement	6.7	72.2	21.1
Organization administration	0.5	91.8	7.6
Public awareness	18.4	80.3	1.4
Political advocacy	14.7	82.0	3.3
Support for others	1.3	69.4	29.4
Partnerships	0.3	73.9	25.8
Appreciation	0.6	25.8	73.6

XSL•FO
RenderX

Table 6. The emotions for each category.

	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Neutral (%)	Sadness (%)	Surprise (%)
Community engagement	4.5	1.1	52.3	12.6	24.5	3.1	1.9
Organization administration	1.5	0.1	59.3	10.9	24.8	2.1	1.3
Public awareness	8.7	3.2	72.2	1.3	10.1	3.9	0.6
Political advocacy	8.2	2.4	65.3	3.4	16.3	3.5	0.9
Support for others	3.8	0.1	59.3	21.4	11.1	1.7	2.5
Partnerships	2.1	0.1	53.4	27.6	13.6	0.7	2.4
Appreciation	3.8	0.0	34.6	54.8	4.3	0.9	1.5

Topic Modeling Results

The coded data set yielded distinct topics within each of the classes or categories. The identified topics, bigrams, and representative tweet examples are presented in Table 7. These themes provide insights into the prevalent topics and discussions within the data set, showcasing different aspects of the discourse surrounding SA and related issues.

XSL•FO
RenderX

Table 7. Sexual assault organizations tweets by topics.

Topics within the category	Bigrams (top 3-5)	Example tweet
1. Community engagement		
Experience and awareness of abuse	experience abuse, wear purple, everyday believe, know experience, red flag, woman know	RT @SHORECentreWR: “Don’t forget to wear your purple today!”
Support and information	post photo, need help, human right, safe way, information support, self-care, free violence	“Nova Scotians are grieving. Mandalas are a symbol of healing with roots in ancient cultural traditions. We invite you to create a mandala, photograph it, and send the photo to...will post the images to Instagram...”
Social media engagement	social media, friend family, find way, medium account, check pic, keep connected	“Great convo going on among Centre residents: social media and sharing too much information. Program like this supports personal growth”
2. Organization administration		
Sexual assault support services and helplines	support line, service available, online phone, crisis line, toll free	“...you want help figuring out a safe warm place to sleep. That phone line is staffed 24/7. We know this is a bandaid and not a permanent solution, but for now we’re focused on safety from this dangerous cold...”
Support groups and emotional support	support group, need help, join team, ticket sale, dropin support	“In response to requests from folks using the space, we decided to make Wednesdays a drop-in for women, trans*, and non-binary persons only. Monday mornings will continue to be open to all genders”
3. Public awareness		
Gender-based violence and advocacy	Genderbased, indigenous woman, woman kill, remember woman	RT @kevinkindred: “Given today’s news that there is no evidence of unmarked graves at the Shubenacadie residential school in NS, let’s take a moment to remember the names of the sixteen indigenous children known to have died at the school”
Sexual assault awareness and support	sexual assault, sexual abuse, experience sexual, rape culture	__[a]
Violence against women and rights advocacy	violence woman, support survivors, end violence, human right	@JulieSLalonde @ArielTroster “There are women right now in violent relationships who feel unsafe at home and are not leaving because there’s no place to go.”
4. Political advocacy		
Advocacy for survivors of sexual assault and justice	end violence, fund sexual, sign petition, criminal justice, assault law, federal government	RT @Pam_Palmater: “Join Dr. @cblackst &; I on YouTube LIVE Friday, Dec. 2nd at 2:00pm ET to get an update on federal gov/mt &; @AFN_Updates decision to file for judicial review of Can. Human Rights Tribunal decision to protect compensation order for First Nations kids.”
Legal aid for survivors of sexual assault	justice system, pour femme, femme service, national inquiry, coalition rape, legal aid	“We have continued to mask and offer testing all along. It’s not a political statement, it’s simply because we care about public health, about each other and everyone in our community. \np.s. it’s really not that hard to do.”
Indigenous rights and reconciliation	indigenous people, action reconciliation, canadian history, justice system, provincial government	RT @JarvisGoogoo: “At @SchulichLaw in March 2005 re Indigenous law, Candy Palmater said that if White women faced violence at the same rate as Aboriginal women did in this country, Canada would declare a national state of emergency”
Sexual violence and action for change	call action, reconciliation justice, truth stand, action reconciliation	RT @CSaulnierHfx: “I feel that there’s a growing collective spirit that is calling for political change, political action on this issue.” @FranklyLess #EndPoverty #nspoli @hfxexaminer
5. Support for others		
Support and advocacy for survivors of sexual assault	mental health, community health, make difference, public education, training forum	RT @SVPWR: “It’s #BellLetsTalk today, and we are so grateful for the local organizations doing so much to support mental health in #WatReg:...”
Campaigns and events to raise awareness about sexual assault	purple today, purple support, executive director, social medium, campaign launch, spread word	RT @welcomehousing: “Tomorrow is @worldpancreatic day &; we will be wearing purple for our late coordinator, Bryon Anderson #WPD15...”
6. Partnerships		
Support and fundraising for survivors of sexual violence	support survivors, raise funds, proceed donate, annual fundraiser, community partners, local business	“Special people deserve a special shout out. Tarek, Karen and their crew are working today to prepare and donate 120 hot meals. This is such a beautiful gift from Tarek’s. We’re taking the meals to individuals and families we support in hotels because #Halifax shelters are full.”

XSL•FO
RenderX

Topics within the category	Bigrams (top 3-5)	Example tweet
Campaigns and support for ending violence against women	support woman, support survivor, support service, raise money, raise awareness, supporting campaigns	"We are excited to be co-hosting an IWD Event, March 8th, 6:30 pm Atlantic with The Local Council of Women Halifax. Join us for 'Addressing Women\'s Housing Needs in Halifax: Exposing the intersections of invisibility.' Register for the event here:"
Programs and efforts to support women and children	download app, effort download, app gather	"Western researchers will be leading the way in developing the first free, Canada-wide app to help survivors of domestic violence find a path to safety and health. #innovation #stopdomesticviolence"
Community engagement and support	Join today, raise money, join tomorrow, silent auction	"We have many many community partners and our amazing show-stopping student Sam has made gifts to say thanks. Deliveries en route."
7. Appreciation		
Gratitude and appreciation for support	thank support, thank help, spread the word, silent auction, thank come, big thank	"@awarham thank you, a lot of people contributed their wisdom, knowledge and experience. and we've had fantastic professional, moral and financial support ❤ without all that, we wouldn't be here."
Thanking supporters and donors	generous donation, thank donate	"To the anonymous donor, thank you for this #RAK Comfort food for shelter residents on the heels of heavy news #NovaScotia #COVID19NS Also a great way to support local #shoplocal"
Gratitude for engagement and participation	Big thank, thank share, huge thank, walk mile, support survivors, mile shoe	"The walk has officially ended and we thank everyone who volunteered and walked the walk!!! #rrwalk2013"

[a]Not available.

Community Engagement

Approximately 20% of the tweets in the data set contained themes related to community engagement, generating 3 topics: "experience and awareness of abuse," "support and information," and "social media engagement." Topic 1 focuses on discussions related to experiences of abuse and raising awareness about it. Topic 2 revolves around providing support and information and promoting human rights in relation to SA. Topic 3 highlights engagement on social media platforms, connecting with friends and family, and finding ways to stay informed and connected.

Organization Administration

In the organization administration class, "sexual assault support services," "helplines," and "support groups and emotional support" were the salient topics. The first topic revolves around providing support for survivors of SA, and the tweets likely contain information about available services and crisis lines; promote helpline numbers; and emphasize the availability of support services. Topic 2 centers on support groups and emotional support for individuals impacted by sexual violence. The tweets may discuss the importance of support networks, encourage individuals to join support groups, and highlight the emotional support available.

Public Awareness

In the public awareness class, we identified 3 topics. Topic 1 focuses on discussions related to gender-based violence, and the tweets likely highlight the need to raise awareness, advocate for survivors, and address issues surrounding gender-based violence. Topic 2 centers on SA awareness, support for survivors, and efforts to combat sexual violence. The tweets may highlight initiatives such as awareness months, support services, survivor empowerment, and the importance of ending sexual violence. Topic 3 focuses on violence against women, advocating for women's rights, and addressing issues such as domestic violence and intimate partner violence. The tweets may discuss the importance of human rights, raise awareness about violence against women, and emphasize the need for support services.

Political Advocacy

This theme delves into the criminal justice system's response to SA cases and advocates for changes and reforms. It discusses specific initiatives, such as signing petitions and calling for justice for communities considered marginalized. It also mentions the importance of advocacy and policy work for Francophone women. Topic 1 revolves around discussions related to SA, advocating for survivors, and seeking justice. The tweets may address issues such as domestic violence, support for survivors, legal cases, and the need for systemic change in addressing SA and violence against women. Topic 2 focuses on support services and resources available for survivors of SA, including centers providing assistance and legal aid. The tweets may mention crisis centers, justice systems, confidential services, and the importance of providing support and resources to survivors. Topic 3 highlights discussions surrounding Indigenous rights, reconciliation, and addressing SA and violence within Indigenous communities. The tweets may emphasize actions for truth and reconciliation, support for Indigenous women, and the need for systemic change to combat violence and promote gender equality. Topic 4 addresses sexual violence in general, including domestic violence and violence against women. The tweets may focus on the need for action, standing against violence, justice systems, systemic change, and the role of government in addressing sexual violence.

Support for Others

This theme included 2 topics, highlighting the importance of community involvement, support, and collective efforts to combat violence and abuse. Topic 1 emphasizes "support and

XSL•FO
RenderX

advocacy for survivors of sexual assault," and the tweets may mention initiatives, organizations, and individuals working to support and address their mental health needs, highlighting the importance of community efforts in making a positive difference. Topic 2 focuses on the "campaigns and events to raise awareness about sexual assault," and the discussion within this topic often involves various campaigns and events, with tweets possibly mentioning actions such as wearing purple to demonstrate support as well as sharing information and promoting community campaigns.

Partnerships

This theme centers on the importance of partnerships, fundraising efforts, and collective efforts to address and prevent violence and abuse. We identified 4 topics within the theme. Topic 1 revolves around providing support and raising funds for survivors of sexual violence. The tweets may mention community partners, local businesses, and fundraising efforts aimed at supporting survivors. Topic 2 focuses on campaigns and initiatives to end violence against women. The tweets may mention supporting campaigns, raising funds, and providing support services for survivors of SA. The topic highlights the importance of community engagement and collective action in addressing violence against women. Topic 3 emphasizes programs and efforts aimed at supporting women and children who have experienced SA. The tweets may mention fundraising events, supporting local services, helplines, and providing assistance to survivors. Topic 4 revolves around community engagement and support related to sexual violence. The tweets may mention joining teams, spreading the word, and supporting survivors through initiatives such as silent auctions. The topic highlights the importance of community participation and collaboration in addressing sexual violence.

Appreciation

The least common theme revealed 3 topics, including "gratitude and appreciation for support," "thanking supporters and donors," and "gratitude for engagement and participation." This theme revolves around expressing gratitude and appreciation for support, donations, and contributions. Topic 1 revolves around expressing gratitude and appreciation for the support received. The tweets thank individuals, organizations, and community members, and the topic emphasizes the importance of acknowledging and recognizing the contributions of supporters. Topic 2 focuses on thanking supporters and donors for their contributions. The tweets express gratitude toward individuals and organizations for their generous donations and ongoing support. The topic highlights the significance of recognizing and thanking those who have contributed to the cause. Topic 3 focuses on expressing gratitude for engagement and participation. The tweets thank individuals for sharing information, participating in events such as walks or auctions, and making a difference. The topic emphasizes the importance of community involvement and active participation.

Discussion

Principal Findings

This study presents a comprehensive classification of Twitter messages that elucidate the reasons for social media use among SA support centers in Canada. Leveraging a data set of 297,360 tweets from 133 SA support organizations, the application of supervised machine learning enabled us to automatically predict content analytical themes within the Twitter corpus. First, we identified the emerging classifications of Twitter's use by human service nonprofits. The results indicated that Twitter is used by SA centers across Canada for various purposes. The identified classifications include (1) community engagement, (2) organization administration, (3) public awareness, (4) political advocacy, (5) support for others, (6) partnerships, and (7) appreciation. These categories reflect the multifaceted nature of human service nonprofits' communication strategies and their engagement with stakeholders. Second, the findings of this study contribute to the existing literature by expanding the understanding of social media use by human service nonprofits beyond the traditional focus on advocacy-related purposes. Although advocacy remains an important aspect, this research reveals that these organizations use social media to achieve a diverse range of objectives, such as raising public awareness, community engagement, and organization administration. Third, the sentiment and emotion analysis of tweets shed light on the emotional tone of different tweet categories. The prevalence of "fear" among organizational tweets underscores the gravity of the issues addressed by human service nonprofits. In contrast, the emotion of "joy" associated with the partnership and appreciation categories highlights the positive impact of community involvement and support. Fourth, the application of machine learning in this study has proven to be valuable in predicting content analytical themes in a large Twitter corpus. The predictive classification model outperformed human coders in terms of accuracy, indicating the potential of machine learning algorithms in analyzing social media data and gaining deeper insights into nonprofits communication strategies.

Typology and Theoretical Framework of Online Organizational Communication Objectives

A key discovery from our study pertains to the distribution of tweet categories. The analysis revealed that the most frequent type of posts falls under the "public awareness" category. Approximately one-third of the collected tweets were classified within this category, signifying that SA support organizations predominantly use social media platforms for advocating against issues related to intimate partner violence and sexual violence. These findings align with prior literature, highlighting how social media allows organizations to disseminate content aimed at increasing awareness while incurring minimal costs [8]. Our topic modeling results uncovered 3 salient themes within the public awareness category, which encompass tweets that emphasize the need to raise awareness, advocate for addressing gender-based violence issues, and support survivors of sexual violence and women's rights. This emphasis on awareness-raising activities reflects the pivotal role social media

XSL•FO
RenderX

plays in creating awareness for organizations and fostering interactions with donors and volunteers [55,56].

Community engagement emerges as the second most prominent reason for social media use by SA organizations in Canada, constituting approximately one-fourth of all collected tweets. This category generated 3 distinct topics through classifications, where community engagement entails nonprofits engaging with their community beyond their primary mission. These tweets include sharing well-wishes, updates on organization activities, quotes, and information about resources that the community may find valuable. Although some tweets do touch on providing information beyond their primary mission, the main focus remains on community engagement concerning sexual violence support. The nuanced examination of tweet contents provided by our topic modeling analysis sheds light on the multifaceted nature of community engagement efforts by SA support organizations. It is evident that these organizations recognize the significance of engaging with their communities regarding sexual violence and related matters through social media. Consistent with existing literature, our findings align with the view that social media offers an avenue for powerful participation and community engagement [57]. It also emphasizes the potential role of social media as a mechanism to raise awareness and inform the community of various initiatives and projects [58]. Although community engagement through social media remains an opportunity for SA support organizations to connect with their communities and market their services, further research should delve into the effectiveness of engagement across all aspects of design, delivery, and evaluation, particularly with regard to specific objectives such as supporting and combating sexual violence.

Sentiments and Emotions in Different Categories in the Typology

Sentiment analysis is used to examine the evaluative perspectives expressed within the text, with its importance stemming from its ability to comprehend the shifting dynamics, potential interventions, and predictive insights into public sentiment regarding trending events. It serves as a valuable tool for offering decision-making support to relevant authorities in the realms of public sentiment monitoring, intervention, and governance. Our sentiment and emotion analysis yielded valuable insights into the emotional tone of the tweets related to each identified online organizational communication objective. Most tweets exhibited a neutral sentiment, surpassing both the negative and positive categories across all 7 classes. This prevalence of neutrality could be attributed to the sensitive nature of the topic, as SA is a deeply distressing issue. However, it is worth noting that tweets related to partnerships and appreciation displayed a greater presence of joy as the dominant emotion, indicating positive sentiments associated with community engagement, support, and expressions of gratitude.

Machine Learning Classification

The evaluation of the machine learning model's performance on the test set showcased promising results. The model demonstrated improvement in average sensitivity scores, from 48.3% to 53.1%, indicating its ability to accurately classify tweets into their respective categories and provide reliable predictions for content analytical themes. Nonetheless, it is essential to acknowledge that there is still room for improvement in the model's performance, particularly in accurately classifying tweets related to partnerships and political advocacy, which constituted smaller segments of the data set.

Implications

Our research carries significant implications for both practitioners and policy makers in the field of SA support services. The typology we have developed represents a substantial advancement in research within this domain and provides a comprehensive framework for understanding how these organizations can effectively use Twitter to disseminate information and engage with the public at the message level. This typology empowers human service nonprofits to align their social media strategies with specific organizational goals. By understanding the different categories and topics of Twitter communication, these organizations can tailor their content and engagement strategies to maximize the impact and relevance of their online presence.

The study emphasizes the importance of social media as a potent communication tool for engaging with communities, raising public awareness, and providing essential information and support. Through an understanding of the diverse communication themes and strategies used by SA support organizations, practitioners can optimize their social media use to effectively reach and connect with their target audience.

On the basis of our research findings, we recommend that human service nonprofits invest in social media education and training for their personnel to enhance their understanding of how to use social media effectively. By building a strategic approach to social media use that aligns with organizational objectives, these nonprofits can maximize their impact and outreach, ultimately furthering their mission to support and advocate for survivors of SA.

Limitations

The study has certain limitations that need to be acknowledged. The predictive performance of the model is influenced by factors such as human annotation, interrater agreement, and the training data set. The trained model attempts to mimic the classification by the human coders, whose understanding of the tweet content and familiarity with the background and theoretical framework is critical to the study. In this regard, the study underwent 4 rounds of training to attain a satisfactory interrater reliability score. However, future studies could incorporate more human coders to enhance the accuracy of the results. Furthermore, our analysis focused solely on Twitter data from SA support organizations in Canada. This geographic and platform limitation may restrict the generalizability of our findings to other countries and social media platforms. Future research should consider expanding the scope to include a more diverse range of organizations and platforms. Finally, although our machine learning model demonstrated promising performance, there is still room for improvement. The accuracy of the model in classifying tweets related to partnerships and political advocacy was relatively lower compared to other categories.

XSL•FO
RenderX

Further refinement and fine-tuning of the model could enhance its accuracy and reliability.

Conclusions

In conclusion, our study offers valuable insights into the application of machine learning to understand the message-level communication purposes of SA support organizations on Twitter in Canada. By combining social science and computer science, we effectively analyzed a large data set and identified content analytical themes, sentiments, emotions, and topics within tweets. These findings enrich our understanding of how SA support organizations use social media for community engagement, public awareness, and organizational administration purposes. The implications extend to practitioners, policy makers, and organizational personnel, emphasizing the significance of education and training to maximize the benefits of social media in achieving organizational goals within the realm of SA support services.

Acknowledgments

The authors would like to extend gratitude to Qiaoru Zhang and Yiding Jin for their contributions to the data validation process of our study. Qiaoru Zhang diligently identified organizations with active Facebook or Twitter accounts, thereby ensuring the reliability and relevance of our data sources. Yiding Jin collected a sample of the Facebook data set.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Confusion matrix for the independent data sets.
[DOCX File , 16 KB - jmir_v26i1e51698_app1.docx]

Multimedia Appendix 2
The number of tweets collected each year.
[PNG File , 112 KB - jmir_v26i1e51698_app2.png]

Multimedia Appendix 3
The histograms of the categories in the human-labeled data: (1) community engagement, (2) organization administration, (3) public awareness, (4) political advocacy, (5) support for others, (6) partnerships, and (7) appreciation.
[PNG File , 75 KB - jmir_v26i1e51698_app3.png]

Multimedia Appendix 4
Top unigrams in the tweets.
[PNG File , 172 KB - jmir_v26i1e51698_app4.png]

Multimedia Appendix 5
Top bigrams in the tweets.
[PNG File , 183 KB - jmir_v26i1e51698_app5.png]

References

1. Enjolras B. Determinants of voluntary organizations' attention on Facebook: the case of Norwegian voluntary organizations. Nonprofit Volunt Sect Q 2022 Oct 11;52(5):1308-1330. [doi: 10.1177/08997640221129551]
2. Lai CH, Fu JS. Humanitarian relief and development organizations' stakeholder targeting communication on social media and beyond. Voluntas 2020 Mar 05;32(1):120-135. [doi: 10.1007/s11266-020-00209-6]
3. Xu W, Saxton GD. Does stakeholder engagement pay off on social media? A social capital perspective. Nonprofit Volunt Sect Q 2018 Aug 02;48(1):28-49. [doi: 10.1177/0899764018791267]
4. Bhati A, McDonnell D. Success in an online giving day: the role of social media in fundraising. Nonprofit Volunt Sect Q 2019 Aug 13;49(1):74-92. [doi: 10.1177/0899764019868849]
5. Priante A, Ehrenhard ML, van den Broek T, Need A, Hiemstra D. "Mo" together or alone? Investigating the role of fundraisers' networks in online peer-to-peer fundraising. Nonprofit Volunt Sect Q 2021 Dec 02;51(5):986-1009. [doi: 10.1177/08997640211057456]

XSL•FO
RenderX

6. Zhou H, Ye S. Legitimacy, worthiness, and social network: an empirical study of the key factors influencing crowdfunding outcomes for nonprofit projects. Voluntas 2018 Jun 4;30(4):849-864. [doi: 10.1007/s11266-018-0004-0]
7. Zhou H, Ye S. Fundraising in the digital era: legitimacy, social network, and political ties matter in China. Voluntas 2019 Apr 01;32(2):498-511. [doi: 10.1007/s11266-019-00112-9]
8. Bilgin Y, Kethüda Ö. Charity social media marketing and its influence on charity brand image, brand trust, and donation intention. Voluntas 2022 Jan 20;33(5):1091-1102. [doi: 10.1007/s11266-021-00426-7]
9. Livermore M, Verbovaya O. Doing collaboration: how organizations use Facebook to foster collaboration. Hum Serv Org Manage Leadership Gov 2016 Jun 07;40(5):553-571. [doi: 10.1080/23303131.2016.1197869]
10. Campbell DA, Lambright KT. Are you out there? Internet presence of nonprofit human service organizations. Nonprofit Volunt Sect Q 2019 Jun 06;48(6):1296-1311. [doi: 10.1177/0899764019852673]
11. Guo C, Saxton GD. Tweeting social change: how social media are changing nonprofit advocacy. Nonprofit Volunt Sect Q 2013 Jan 08;43(1):57-79. [doi: 10.1177/0899764012471585]
12. Halpin DR, Fraussen B, Ackland R. Which audiences engage with advocacy groups on Twitter? Explaining the online engagement of elite, peer, and mass audiences with advocacy groups. Nonprofit Volunt Sect Q 2020 Dec 21;50(4):842-865. [doi: 10.1177/0899764020979818]
13. Sitter KC, Curnew AH. The application of social media in social work community practice. Soc Work Educ 2016 Jan 14;35(3):271-283. [doi: 10.1080/02615479.2015.1131257]
14. Shier ML, Handy F. From advocacy to social innovation: a typology of social change efforts by nonprofits. Voluntas 2014 Nov 18;26(6):2581-2603. [doi: 10.1007/s11266-014-9535-1]
15. Cooner TS, Beddoe L, Ferguson H, Joy E. The use of Facebook in social work practice with children and families: exploring complexity in an emerging practice. J Technol Hum Serv 2019 Oct 22;38(2):137-158. [doi: 10.1080/15228835.2019.1680335]
16. O'Sullivan L, Hughes Z. Incorporating Facebook into nonprofit supports for family caregivers: reflections on its value and relevance. J Technol Hum Serv 2019 Jun 17;37(2-3):129-141. [doi: 10.1080/15228835.2019.1620669]
17. Tian F, Labban A, Shearer R, Gai Q. The impact of social media activity on nonprofit donations in China. Voluntas 2019 Oct 01;32(2):488-497 [FREE Full text] [doi: 10.1007/s11266-019-00168-7]
18. Zorn TE, Grant S, Henderson A. Strengthening resource mobilization chains: developing the social media competencies of community and voluntary organizations in New Zealand. Voluntas 2012 Mar 13;24(3):666-687. [doi: 10.1007/s11266-012-9265-1]
19. Comfort SE, Hester JB. Three dimensions of social media messaging success by environmental NGOs. Environ Commun 2019 Mar 25;13(3):281-286. [doi: 10.1080/17524032.2019.1579746]
20. Lam WF, Nie L. Online or offline? Nonprofits' choice and use of social media in Hong Kong. Voluntas 2019 Jun 24;31(1):111-128. [doi: 10.1007/s11266-019-00128-1]
21. Svensson PG, Mahoney TQ, Hambrick ME. Twitter as a communication tool for nonprofits: a study of sport-for-development organizations. Nonprofit Volunt Sect Q 2014 Oct 16;44(6):1086-1106. [doi: 10.1177/0899764014553639]
22. Guo C, Saxton GD. Speaking and being heard: how nonprofit advocacy organizations gain attention on social media. Nonprofit Volunt Sect Q 2017 Jun 12;47(1):5-26. [doi: 10.1177/0899764017713724]
23. Chan C. A scoping review of social media use in social work practice. J Evid Inf Soc Work 2016 Jul 15;13(3):263-276. [doi: 10.1080/23761407.2015.1052908] [Medline: 26176999]
24. Goldkind L. Social media and social service: are nonprofits plugged in to the digital age? Hum Serv Org Manage Leadership Gov 2015 Jun;39(4):380-396. [doi: 10.1080/23303131.2015.1053585]
25. Young JA. Facebook, Twitter, and blogs: the adoption and utilization of social media in nonprofit human service organizations. Hum Serv Org Manage Leadership Gov 2016 May 24;41(1):44-57. [doi: 10.1080/23303131.2016.1192574]
26. Ihm J, Kim E. When nonprofit organizations meet information and communication technologies: how organizational culture influences the use of traditional, digital, and sharing media. Voluntas 2021 Mar 02;32(3):678-694. [doi: 10.1007/s11266-021-00335-9]
27. Seo H, Vu HT. Transnational nonprofits' social media use: a survey of communications professionals and an analysis of organizational characteristics. Nonprofit Volunt Sect Q 2020 Feb 28;49(4):849-870. [doi: 10.1177/0899764020908340]
28. Huang S. NGO as sympathy vendor or public advocate? A case study of NGOs' participation in internet fundraising campaigns in China. Voluntas 2022 Feb 18;33(5):1064-1076. [doi: 10.1007/s11266-022-00463-w]
29. Meijer A. Co-production in an information age: individual and community engagement supported by new media. Voluntas 2012 Jul 18;23(4):1156-1172. [doi: 10.1007/s11266-012-9311-z]
30. Wu VC. Exploring donor influence and public engagement: computational and thematic analyses of social media messages. Voluntas 2022 Jul 14;34(4):813-829. [doi: 10.1007/s11266-022-00481-8]
31. Byrne J, Kirwan G, Mc Guckin C. Social media surveillance in social work: practice realities and ethical implications. J Technol Hum Serv 2019 Apr 04;37(2-3):142-158. [doi: 10.1080/15228835.2019.1584598]
32. Beaumont E, Chester P, Rideout H. Navigating ethical challenges in social media: social work student and practitioner perspectives. Aus Soc Work 2017 Mar 29;70(2):221-228 [FREE Full text] [doi: 10.1080/0312407x.2016.1274416]
33. Voshel EM, Wesala A. Social media and social work ethics: determining best practices in an ambiguous reality. J Soc Work Values Ethics 2015;12(1):67-76.

XSL•FO
RenderX

34. Stanfield D. Social media and social work education curriculum in Aotearoa New Zealand: an integrated framework. Adv Soc Work Welfare Educ 2020;21(2) [FREE Full text] [doi: 10.1080/02615479.2016.1154663]
35. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550 [FREE Full text] [doi: 10.2196/20550] [Medline: 33119535]
36. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS One 2020 Sep 25;15(9):e0239441 [FREE Full text] [doi: 10.1371/journal.pone.0239441] [Medline: 32976519]
37. Xue J, Chen J, Chen C, Hu R, Zhu T. The hidden pandemic of family violence during COVID-19: unsupervised learning of tweets. J Med Internet Res 2020 Nov 06;22(11):e24361 [FREE Full text] [doi: 10.2196/24361] [Medline: 33108315]
38. Su Y, Xue J, Liu X, Wu P, Chen J, Chen C, et al. Examining the impact of COVID-19 lockdown in Wuhan and Lombardy: a psycholinguistic analysis on Weibo and Twitter. Int J Environ Res Public Health 2020 Jun 24;17(12):4552 [FREE Full text] [doi: 10.3390/ijerph17124552] [Medline: 32599811]
39. Xiang X, Lu X, Halavanau A, Xue J, Sun Y, Lai PH, et al. Modern senicide in the face of a pandemic: an examination of public discourse and sentiment about older adults and COVID-19 using machine learning. J Gerontol B Psychol Sci Soc Sci 2021 Mar 14;76(4):e190-e200 [FREE Full text] [doi: 10.1093/geronb/gbaa128] [Medline: 32785620]
40. Xue J, Chen J, Gelles R. Using data mining techniques to examine domestic violence topics on Twitter. Violence Gender 2019 Jun;6(2):105-114. [doi: 10.1089/vio.2017.0066]
41. Xue J, Macropol K, Jia Y, Zhu T, Gelles RJ. Harnessing big data for social justice: an exploration of violence against women‐related conversations on Twitter. Hum Behav Emerg Tech 2019 Jul 26;1:269-279. [doi: 10.1002/hbe2.160]
42. Zheng C, Xue J, Sun Y, Zhu T. Public opinions and concerns regarding the Canadian Prime Minister's daily COVID-19 briefing: longitudinal study of YouTube comments using machine learning techniques. J Med Internet Res 2021 Feb 23;23(2):e23957 [FREE Full text] [doi: 10.2196/23957] [Medline: 33544690]
43. Xue J. Agenda-setting for intimate partner violence: exploring the role of social media United States-based Twitter. University of Pennsylvania. 2018. URL: https://assets-global.website-files.com/616733ce3e675341229a07bd/62aec0ca4580ec17599dca6e_Agenda-Setting_for_Intimate_Pa.pdf [accessed 2024-04-19]
44. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online October 11, 2018 [FREE Full text]
45. Brownlee J. Probability for Machine Learning: Discover How To Harness Uncertainty With Python. San Juan, Puerto Rico: Machine Learning Mastery; 2019.
46. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2022 Oct;6(5):429-449 [FREE Full text]
47. Build reliable web scrapers. Fast. Apify. URL: https://apify.com/ [accessed 2024-04-25]
48. Liu B. Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool; 2012.
49. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]
50. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv Preprint posted online July 26, 2019 [FREE Full text] [doi: 10.1090/mbk/121/79]
51. Hartmann J, Heitmann M, Schamp C, Netzer O. The power of brand selfies. J Mark Res 2021 Oct 16;58(6):1159-1177. [doi: 10.1177/00222437211037258]
52. Mohammad SM. Ethics sheet for automatic emotion recognition and sentiment analysis. Comput Linguist 2022;48(2):239-278. [doi: 10.1162/coli_a_00433]
53. Emotion English DistilRoBERTa-base. Hugging Face. URL: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/ [accessed 2024-04-18]
54. Zhou S, Zhao Y, Bian J, Haynos AF, Zhang R. Exploring eating disorder topics on Twitter: machine learning approach. JMIR Med Inform 2020 Oct 30;8(10):e18273 [FREE Full text] [doi: 10.2196/18273] [Medline: 33124997]
55. Cahalane C. Charities should be proactive and approach local authorities. The Guardian. 2013 Jul 3. URL: https://www.theguardian.com/voluntary-sector-network/2013/jul/03/charities-proactive-local-authorities [accessed 2024-04-10]
56. Quinton S, Fennemore P. Missing a strategic marketing trick? The use of online social networks by UK charities. J Philanthr Mark 2012 Oct 18;18(1):36-51. [doi: 10.1002/nvsm.1450]
57. Ross C. Social media for digital humanities and community engagement. In: Warwick C, Terras M, Nyhan J, editors. Digital Humanities in Practice. Cambridge, MA: Facet Publishing; 2012.
58. Steinmetz C, Rahmat H, Marshall N, Bishop K, Thompson S, Park M, et al. Liking, tweeting and posting: an analysis of community engagement through social media platforms. Urban Policy Res 2020 Jul 16;39(1):85-105. [doi: 10.1080/08111146.2020.1792283]

Abbreviations

API: application programing interface

XSL•FO
RenderX

SA: sexual assault

Edited by A Mavragani; submitted 08.08.23; peer-reviewed by MO Khursheed, N Hu; comments to author 31.08.23; revised version received 12.10.23; accepted 08.04.24; published 08.05.24.

Please cite as:
Xue J, Shier ML, Chen J, Wang Y, Zheng C, Chen C
A Typology of Social Media Use by Human Service Nonprofits: Mixed Methods Study
J Med Internet Res 2024;26:e51698
URL: https://www.jmir.org/2024/1/e51698
doi:10.2196/51698
PMID:38718390

©Jia Xue, Micheal L Shier, Junxiang Chen, Yirun Wang, Chengda Zheng, Chen Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User-Centered Framework for Implementation of Technology (UFIT): Development of an Integrated Framework for Designing Clinical Decision Support Tools Packaged With Tailored Implementation Strategies

Jessica Ray[1], MS, PhD; Emily Benjamin Finn[2], MPH; Hollyce Tyrrell[3], MSSW; Carlin F Aloe[2], MPH; Eliana M Perrin[4], MD, MPH; Charles T Wood[5], MPH, MD; Dean S Miner[6], MD; Randall Grout[7], MD, MS; Jeremy J Michel[8], MHS, MD; Laura J Damschroder[9], MS, MPH; Mona Sharifi[2], MPH, MD

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida College of Medicine, Gainesville, FL, United States
[2]Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States
[3]Academic Pediatric Association, McLean, VA, United States
[4]Department of Pediatrics, Johns Hopkins University School of Medicine and School of Nursing, Baltimore, MD, United States
[5]Department of Pediatrics, Duke University School of Medicine, Durham, NC, United States
[6]Departments of Pediatrics and Internal Medicine, East Carolina University, Greenville, NC, United States
[7]Department of Pediatrics, Indiana University School of Medicine and Regenstrief Institute, Indianapolis, IN, United States
[8]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[9]Implementation Pathways, LLC and Veterans Affairs Center for Clinical Management Research, Ann Arbor, MI, United States

Corresponding Author:
Jessica Ray, MS, PhD
Department of Health Outcomes and Biomedical Informatics
University of Florida College of Medicine
1889 Museum Road
Gainesville, FL, 32611
United States
Phone: 1 321 356 4346
Email: jessica.ray@ufl.edu

Abstract

Background: Electronic health record–based clinical decision support (CDS) tools can facilitate the adoption of evidence into practice. Yet, the impact of CDS beyond single-site implementation is often limited by dissemination and implementation barriers related to site- and user-specific variation in workflows and behaviors. The translation of evidence-based CDS from initial development to implementation in heterogeneous environments requires a framework that assures careful balancing of fidelity to core functional elements with adaptations to ensure compatibility with new contexts.

Objective: This study aims to develop and apply a framework to guide tailoring and implementing CDS across diverse clinical settings.

Methods: In preparation for a multisite trial implementing CDS for pediatric overweight or obesity in primary care, we developed the User-Centered Framework for Implementation of Technology (UFIT), a framework that integrates principles from user-centered design (UCD), human factors/ergonomics theories, and implementation science to guide both CDS adaptation and tailoring of related implementation strategies. Our transdisciplinary study team conducted semistructured interviews with pediatric primary care clinicians and a diverse group of stakeholders from 3 health systems in the northeastern, midwestern, and southeastern United States to inform and apply the framework for our formative evaluation.

Results: We conducted 41 qualitative interviews with primary care clinicians (n=21) and other stakeholders (n=20). Our workflow analysis found 3 primary ways in which clinicians interact with the electronic health record during primary care well-child visits identifying opportunities for decision support. Additionally, we identified differences in practice patterns across contexts necessitating a multiprong design approach to support a variety of workflows, user needs, preferences, and implementation strategies.

XSL•FO
RenderX

Conclusions: UFIT integrates theories and guidance from UCD, human factors/ergonomics, and implementation science to promote fit with local contexts for optimal outcomes. The components of UFIT were used to guide the development of Improving Pediatric Obesity Practice Using Prompts, an integrated package comprising CDS for obesity or overweight treatment with tailored implementation strategies.

Trial Registration: ClinicalTrials.gov NCT05627011; https://clinicaltrials.gov/study/NCT05627011

(J Med Internet Res 2024;26:e51952) doi:10.2196/51952

KEYWORDS

user-centered design; implementation science; clinical decision support; human factors; implementation; decision support; develop; development; framework; frameworks; design; user-centered; digital health; health technology; health technologies; need; needs; tailor; tailoring; guidance; guideline; guidelines; pediatric; pediatrics; child; children; obese; obesity; weight; overweight; primary care

Introduction

Electronic health record (EHR)–based clinical decision support (CDS) tools can accelerate clinicians' adoption of evidence-based practice [1,2]. However, the potential for a broad and sustained impact on care delivery is limited by institutional policies, EHR system constraints, and end user behaviors that can inhibit dissemination and implementation across heterogeneous settings [1-4]. Failing to situate CDS tools within existing workflows or overlooking human and system-level barriers or facilitators can produce unintended consequences. Poor EHR usability has been associated with clinician burnout, alert fatigue, and inappropriate care, suggesting potentially harmful effects of mismatches between technology, site-specific context, and user needs [5-7]. Successful dissemination and implementation across diverse clinical contexts require balancing the fidelity of core functions with adaptation to ensure compatibility with shared and divergent user needs and contexts [8].

User-centered design (UCD) offers a structured, iterative process to refine an intervention by identifying and meeting individual user needs and preferences at the outset and throughout design and development. This approach centers on users, clinical context, and user feedback throughout design and development. Well-executed UCD can elicit and address critical information about user cognition, affect, and work systems [9,10]. Although extant technology standards explicitly call for UCD, these processes are applied inconsistently during EHR tool development due to variability in technology developers' UCD knowledge and resources [10]. UCD embraces a variety of user engagement methods to distill and clarify the steps needed to optimally design systems for users. However, UCD models are designed to provide a process for design instead of theory-based approaches to guide the exploration of use context. Furthermore, UCD is often more focused on technology development than on strategies to support its adoption.

To facilitate adoption, the fields of implementation science and human factors/ergonomics (HFE) offer approaches for understanding important mediators and modifiers of the implementation of interventions like CDS. By providing theoretical and empirical insights into human-system interactions and cognitive processes, findings from these fields support design. For example, insights can be used to identify and tailor effective implementation strategies (eg, trainings, audits, and feedback) and have been bolstered recently through the application of the UCD process to design implementation strategies. In this study, we developed and applied an integrated framework, the User-Centered Framework for Implementation of Technology (UFIT), in preparation for a multisite clinical effectiveness-implementation randomized controlled trial, Improving Pediatric Obesity Practice Using Prompts (iPOP-UP; ClinicalTrials.gov NCT05627011) to improve guideline-concordant care for children with overweight or obesity in pediatric primary care. UFIT builds upon the UCD process model by layering HFE theories to ground our understanding of user-system interactions within the clinical setting and an implementation science framework to capture aspects of the broader context, informing implementation and sustained use of our CDS intervention.

Methods

Conceptual Foundations

We first developed a foundational conceptual framework to guide the development of the UFIT process model. As shown in Figure 1, we draw on multiple well-established frameworks and theoretical domains. We started by centering concepts from typical UCD approaches, placing the user at the center throughout design and development. Next, we link key HFE theoretical models (eg, Situation Awareness and Systems Engineering Initiative for Patient Safety) and 1 implementation science framework, the Consolidated Framework for Implementation Research (CFIR), to expand the understanding of factors that may help or hinder design or implementation of interventions [11-14].

XSL•FO
RenderX

Figure 1. Conceptual integration of models and theories. HFE: human factors/ergonomics; IS: implementation science; SEIPS: Systems Engineering for Patient Safety; UCD: user-centered design.

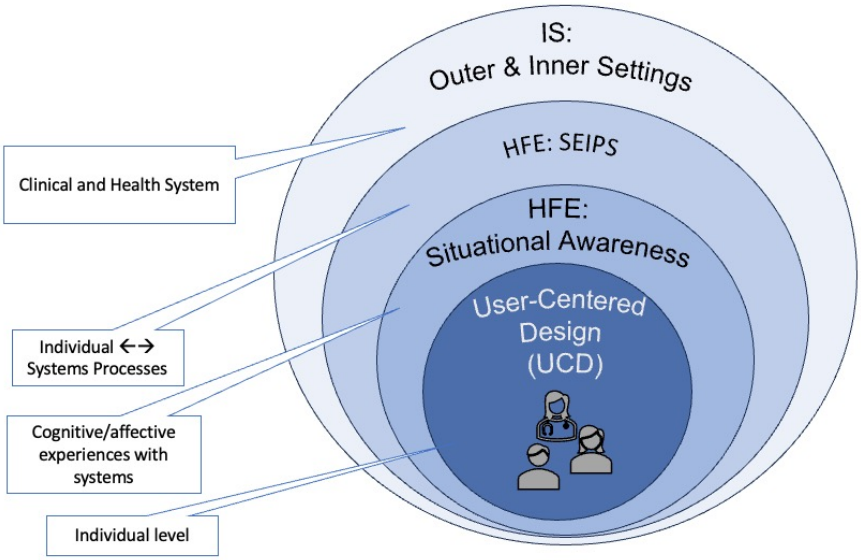

Incorporating HFE extends UCD's focus on the user by adding deeper consideration of the user's interactions across the broader clinical system and the local sociotechnical environment. Situation awareness theory expands our inquiry into human-system interactions by identifying the ways in which clinicians gather, process, and comprehend information in the clinical setting to predict future states and to inform their decision-making [11,15]. Health care environments are complex, where interactions occur among and between people engaged in the system (eg, clinician, patient, and chief medical information officer) as well as tools and technologies needed to complete tasks within the clinical environments [12,16]. Implementing a change in any one component of this type of complex sociotechnical environment can create unintended consequences. We used these insights to guide CDS design requirements to optimize human-system interactions that supported decision-making for best outcomes.

While prior work offers insight into features that can support the adoption of CDS, only recently have researchers sought to identify the implementation features that maximize the adoption (use) of CDS [17]. This recent work provides a foundation for understanding feature selection and paired implementation strategies that have been shown effective, yet it still leaves the important design considerations of how these strategies should be tailored to support local contexts. To tailor implementation strategies, we incorporated the CFIR, a "meta-theoretical" framework with well-defined constructs (ie, potential barriers and facilitators to effective implementation) grouped across five domains: (1) innovation characteristics, (2) outer setting, (3) inner setting, (4) the roles and characteristics of the individuals involved, and (5) implementation processes [13,18]. The CFIR provides a guide to systematically assess how each of these 5 domains may influence implementation. Incorporating components of the CFIR helped to inform understanding of multilayered complexities that influence end users' decisions to adopt CDS tools into their workflow. For example, we sought to understand not only clinicians' current and desired behaviors of interest (eg, "Individual Capability" and "Motivation") but also influences from multiple levels of the context within which clinicians act (eg, "Relative Priority" within the "Inner Setting" that may include the clinic) and (eg, "Local Conditions" within the "Outer Setting" that may include the broader health system). Incorporating the CFIR into UFIT helps to inform our understanding of the diverse needs and constraints of implementation sites more clearly than if we only used UCD, which does not explicitly cover these considerations.

UFIT Development and Application

Based on concepts from our foundational framework, we developed steps for designing an EHR-based CDS tool with companion implementation strategies that are tailored to the needs of multiple heterogeneous pediatric primary care clinics. The resulting 7-step framework, the User-Centered Framework for Implementation of Technology (UFIT), is presented in Figure 2. Concepts from our foundational framework (see Figure 1) are identified for each of the 7 steps. UFIT provides structured guidance for key actions and topics of inquiry for each step. UFIT helps to identify idiosyncratic workflows and site-specific processes to inform the design of a technology-based intervention with tailored implementation strategies.

We applied each step in UFIT to prepare for the iPOP-UP randomized controlled trial that aims to improve guideline-concordant care for children with overweight or obesity in pediatric primary care. iPOP-UP packages a clinical CDS with implementation strategies into a single bundle.

XSL•FO
RenderX

Figure 2. The User-Centered Framework for Implementation of Technology (UFIT): steps and methods for evaluation. CFIR: Consolidation Framework for Implementation Research; EHR: electronic health record; HFE: human factors/ergonomics; IS: implementation science; SEIPS: Systems Engineering for Patient Safety; SWOT: strengths, weaknesses, opportunities, and threats; UCD: user-centered design.

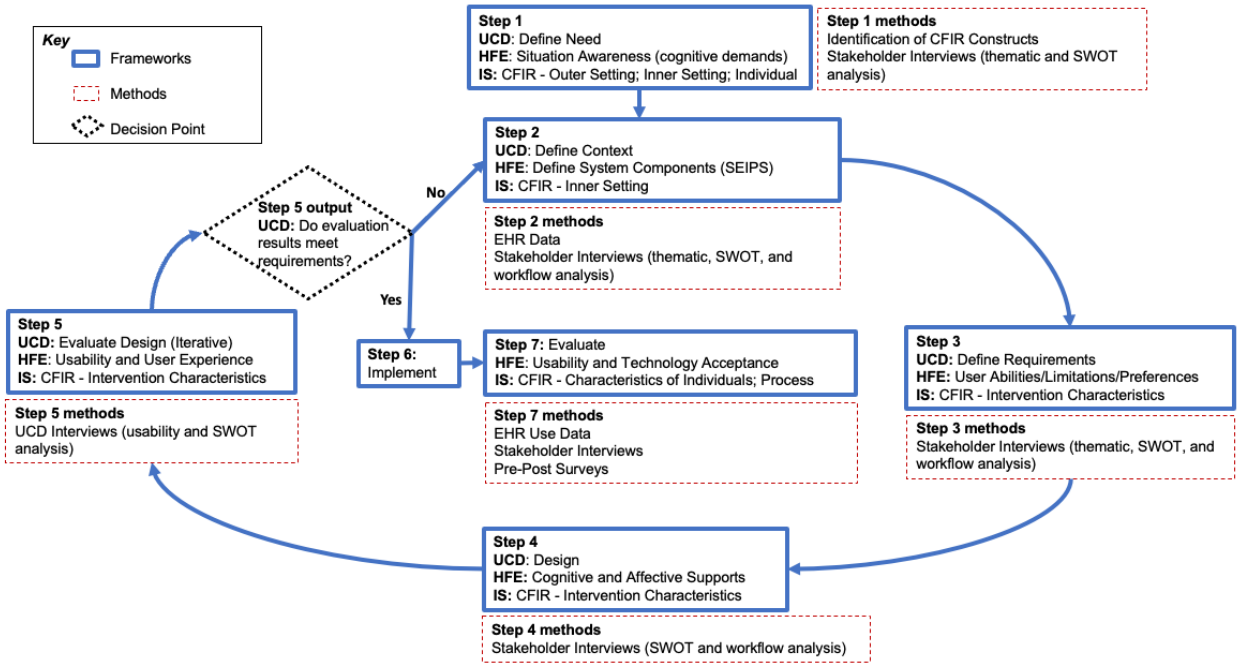

Preparation

Methods used to apply UFIT within our formative evaluation included qualitative and quantitative methods such as semistructured interviews with clinicians and staff, EHR data queries, and surveys (Table 1). Participants and data collection and analyses used throughout the process are detailed in Table 1. Institutional review boards at each site reviewed and determined the protocol for these data collections as exempt.

In preparation for engaging stakeholders in UFIT, 9 experts from our core transdisciplinary study team, with expertise in pediatric obesity, primary care practice patterns, health inequities, implementation science, and clinical informatics, weighted their perceived relevance of each of the original CFIR's 39 constructs to evaluating determinants of implementation of clinical practice guidelines and of adoption of CDS in pediatric primary care. This weighting was followed by a group discussion to achieve consensus; 26 constructs were identified as highly relevant priorities to inform the construction of the interview guides and survey used in our formative evaluation. While we used the original version of the CFIR to guide data collection, we mapped our findings and report based on the updated framework, published after our data collection was complete [9,10].

XSL•FO
RenderX

Table 1. Data collection participants, methods, objectives, and analyses within the User-Centered Framework for Implementation of Technology.

Objective, UFIT[a] steps, and participants	Methodology details		
	Method	Objectives	Analysis
Explore needs and context			
1			
Patients aged 2-18 years with BMI ≥85th percentile	EHR[b] data query	—[c]	Descriptive analysis to analyze patient demographics and EHR documentation of key clinician behaviors related to overweight or obesity management
1, 2, 3			
Level 1 stakeholders (anticipated end users)[d]	Semistructured interview	• Explore clinicians' goals, experiences, workflows, and expectations in addressing overweight and obesity in primary care • Explore perceptions of what changes are needed and possible solutions, including those beyond the constraints of current EHR systems	• Rapid qualitative analysis • Content analysis to code workflow data, graphically represented in Sankey diagrams • SWOT[e] analyses
Level 2 stakeholders (non-PCC[f] clinicians, staff who will be impacted by or influence implementation)[g]	Semistructured interview	• Explore clinicians' goals, experiences, workflows, and expectations in addressing overweight and obesity in primary care • Explore perceptions of what changes are needed and possible solutions, including those beyond the constraints of current EHR systems	SWOT analysis
Iterative design			
4, 5			
Level 1 stakeholders (anticipated end users)[d]	Focus group	• Confirm and refine workflow patterns • Gather feedback on initial design prototypes	• Field notes analyzed through summarization feedback for each CDS[h] tool presented. • Compiled data reviewed against prior findings, requirements, and workflow patterns to determine if additional modifications were needed.
Level 1 stakeholders (anticipated end users)[d]	Semistructured interview	• Confirm and refine workflow patterns • Gather feedback on initial design prototypes	—
Nursing and medical assistant staff	Semistructured interview	• Confirm and refine workflow patterns • Gather feedback on initial design prototypes	• Field notes were analyzed through summarization feedback for each CDS tool presented. • Compiled data were reviewed against prior findings, requirements, and workflow patterns to determine if additional design modifications were needed.
Evaluate trial impact			
7			
Level 1 stakeholders (anticipated end users)[d]	Semistructured interview	Gather feedback on intervention and implementation strategies	• Thematic analysis

XSL•FO
RenderX

Objective, UFIT[a] steps, and participants	Methodology details		
	Method	Objectives	Analysis
Level 1 stakeholders (anticipated end users)[d]	Postsurvey	Gather feedback on intervention and implementation strategies	• Descriptive analysis • Wilcoxon sum • $\chi 2$ tests
Patients aged 2-18 years with BMI ≥85th percentile	EHR data query	Gather feedback on intervention and implementation strategies	• Descriptive analysis to analyze patient demographics and EHR documentation of key clinician behaviors related to overweight or obesity management

[a]UFIT: User-Centered Framework for Implementation of Technology.

[b]EHR: electronic health record.

[c]Not available.

[d]Primary care clinicians included physicians with training in General Pediatrics, Internal Medicine-Pediatrics (Med-Peds), and Family Medicine; nurse practitioners; and physician assistants.

[e]SWOT: strengths, weaknesses, opportunities, and threats.

[f]PCC: primary care clinician.

[g]Level 2 stakeholders included Dietitians and healthy weight program clinicians and staff, informatics professionals, clinical site practice managers, and quality, safety, and human factors professionals.

[h]CDS: clinical decision support.

Defining Need, Context, and Requirements (UFIT Steps 1-3)

To understand local uptake of clinical practice guidelines and expert recommendations at primary care clinics, we first quantified clinicians' documented care for children with overweight or obesity through EHR data queried from 84 pediatric primary care practices affiliated with the 3 participating health systems all using Epic EHR systems. Together with primary care physicians with training in general pediatrics and medicine-pediatrics and expertise in clinical informatics from each site, we developed a data query protocol for all well-child visits for children aged 2-12 years with overweight or obesity, defined as a BMI at or above the 85th percentile for age and sex. From these data, we used descriptive analyses to summarize patient demographics and documentation of key clinician behaviors related to overweight or obesity management to define gaps and variability in guideline-concordant care within each health system.

Next, we sought to qualitatively understand primary care clinicians' (PCCs') behaviors and the cognitive and affective factors driving these behaviors. We developed semistructured interview guides to elicit cognitive features supporting clinicians' situation awareness [19]. To understand context, we applied an adapted HFE approach by having PCC stakeholders map their workflows during a well-child visit for children with overweight or obesity during the first round of interviews (Figures 3 and 4). To map the sociotechnical system, researchers often conduct observations capturing characteristics of system components and interactions among these components. For example, a clinician might interact with a patient in an exam room through a series of questions, then document the answers in the EHR, following local workflow structures and meeting external regulations and billing requirements. Given the broad range of systems and their geographic diversity, as well as the timing coinciding with the COVID-19 pandemic, we adapted the approach by conducting cognitive walkthroughs with a graphical sticky note exercise to map the system features and clinician workflow during our interviews. Based on our selected CFIR constructs, interview questions also explored prior experiences with CDS implementations in the participating organizations, clinician capability and motivation in managing overweight or obesity, and their agreement with associated guidelines. These questions encouraged "Reflecting and Evaluating," a key CFIR construct. The team also noted that CFIR's "Innovation Design" and "Tailoring Strategies" constructs are strengthened via UCD. A structured approach to CDS design and tailoring of implementation strategies within these contexts can optimize the overall performance and users' experiences while minimizing the potential for unintended consequences.

XSL•FO
RenderX

Figure 3. Example workflow mapping during data collection. AVS: after visit summary; EHR: electronic health record; MA: medical assistant.

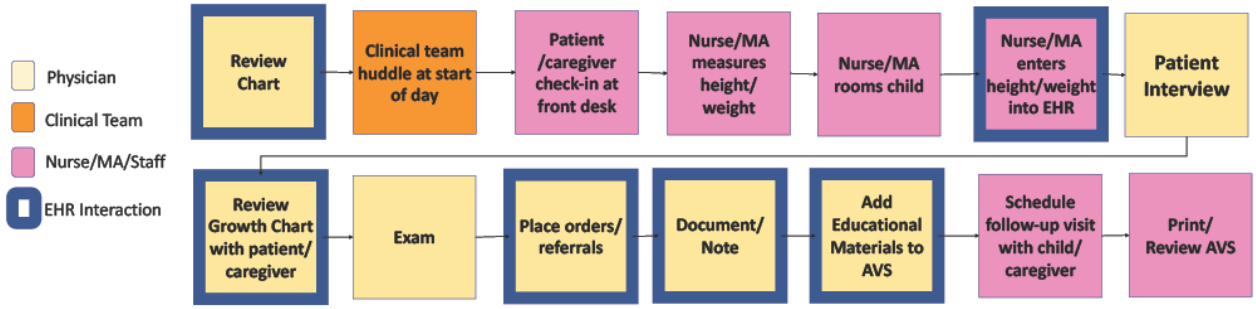

Figure 4. Sankey diagram of clinician workflow and EHR use patterns during pediatric primary care visit. EHR: electronic health record; PCC: primary care clinician.

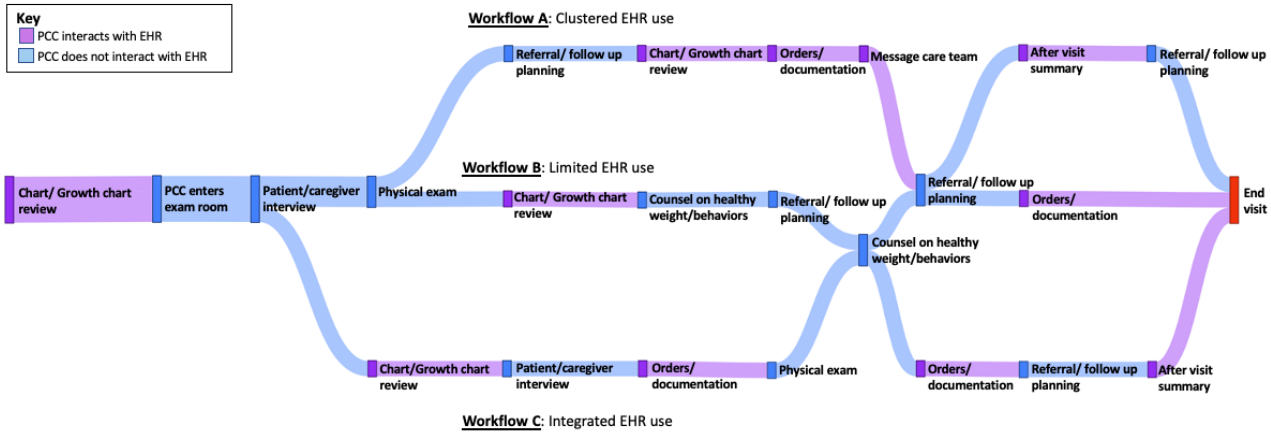

To achieve this, we purposively recruited 2 groups of key informants (clinician end users and non-PCC clinicians and staff who will be impacted by or influence implementation), defined by their proximity to using the CDS, (Table 1) for semistructured interviews to define the need, context, and requirements (UFIT steps 1, 2, and 3) at each site. Interviews lasted approximately 45 minutes. The PCC stakeholder interview guide was designed to produce four distinct categories of data: (1) barriers and facilitators of care for pediatric overweight or obesity, (2) workflow, (3) barriers and facilitators to EHR tool implementation and use, and (4) PCC needs and possible solutions to improve care for patients with overweight or obesity. Analysis for each of these sections included (1) rapid identification of themes for barriers and facilitators of care; (2) Sankey diagrams of clinician workflows; (3) strengths, weaknesses, opportunities, and threats analysis to understand barriers and facilitators to EHR tool implementation; and (4) rapid qualitative analysis, identifying new tools PCCs would find helpful and the potential users of those tools. Interviews were professionally transcribed and independently coded by at least 2 researchers using a rapid assessment procedure [11]. The 3-person interdisciplinary coding group with experience in human factors, public health, and pediatric health care research initially identified broad categories to define barriers and facilitators (eg, patient or family factors and clinician-level factors). We then defined subcategories and iteratively refined them as transcripts were coded and themes emerged within the barriers and facilitators. Coders summarized transcript segments and noted line numbers within each category or subcategory. Strengths, weaknesses, opportunities, and threats analyses were conducted using PCC interview transcripts or real-time field notes for additional stakeholder interviews. A content analysis of the steps clinicians took prior to and during a patient visit was completed to understand workflows and summarized using a Sankey diagram (Figure 3) to create a visualization of the findings from the content analysis.

To define design requirements (step 3), we mapped CDS components to initial interview findings including positive and negative experiences with EHR tools, perceptions of existing CDS tools, and desired CDS tools. We engaged our transdisciplinary study team to narrow the requirements generated from PCC interviews. Investigators from each site were asked to document existing tools that shared core functional components of the CDS to identify the baseline state and past experiences with EHR design in their local environments. The Sankey diagrams were then used to identify timing requirements for CDS tools within clinician workflows. In alignment with HFE, we used UFIT to outline user workload and competing priorities (eg, the need for CDS to facilitate the steps of a clinician's work in a 20-minute visit). To define initial design requirements for our implementation strategies, we linked potential barriers and threats to adoption using the Expert Recommendations for Implementing Change taxonomy of implementation strategies [12]. From CFIR, we consider the CDS characteristics and associated implementation strategies (eg, use "relative advantage" to ensure completing care tasks,

XSL•FO
RenderX

such as ordering labs based on new guidelines vs ordering labs saved as favorites that may not be in line with new guidelines).

Evaluating Design (UFIT Steps 4 and 5)

We translated data from step 3 into an initial set of design prototypes that we presented to our transdisciplinary study team to identify opportunities for further refinement and potential local EHR system constraints. Data from site-specific focus groups was used to confirm and refine workflow patterns. Next, we convened focus groups of PCCs in a first round of design and evaluation (steps 4 and 5) to elicit feedback on early design ideas and confirm workflow patterns. These sessions also allowed us to gather feedback on initial design prototypes and facilitate consensus building within each site to allow us to narrow design plans for subsequent rounds of user testing. The evaluation was informed by the HFE concepts of "usability" and "user experience" as well as key components of the "intervention characteristics" as detailed in CFIR [13,14]. Iterative rounds of design and testing (steps 4 and 5) consisted of individual interviews using images of refined CDS tools in a test environment applied to a hypothetical patient. Between our first and second rounds, we transitioned our prototypes from wireframe to the EHR test environment, from which we captured photos and videos. Beginning with our second round of evaluation, we conducted individual interviews with PCCs using a think-aloud interview guide based on EHR photos and videos to understand user preferences for the tools. Participants were prompted for summative feedback at the end of each session to elicit ideas on both technology design and potential implementation strategies to support adoption. Following each iteration, evaluation feedback was compared with data gathered in initial PCC interviews to confirm that our design was meeting user needs (step 1) and context (step 2) and to assess for any further refinement to our definitions of the context (step 2) and user requirements (step 3). While early rounds focused largely on CDS design testing, later rounds introduced potential implementation strategies to assess suitability. This iterative process was repeated based on meetings of the development team until all feasible user requirements were addressed. Drawing from the recent CFIR 2.0 Outcomes Addendum, we recognize that while formative evaluation may allow us to anticipate outcomes of an intervention and its implementation, only through summative evaluation of the implementation package will we be able to understand its actual impact [15].

Results

Defining Need, Context, and Requirements (UFIT Steps 1-3)

EHR data analysis demonstrated gaps in documented evidence of diagnosis and counseling. Specifically, among children aged 2-12 years with overweight or obesity, 57.3% (n=73,586) had any visit or problem list diagnosis reflecting elevated BMI; this ranged from a low of 18.7% (n=11,073) at the study site with the lowest such documentation to 71.2% (n=51,993) at the site with the highest. In addition, only 0.3% (n=73,586) of visits with such patients included documented nutrition or physical activity counseling.

We conducted qualitative interviews with 21 PCCs' end users and 20 Level 2 stakeholders who would potentially be impacted or involved with the CDS and implementation but were not direct end users. Qualitative analysis of PCC interviews elicited clinicians' goals, experiences, and expectations in addressing overweight or obesity in primary care. This included mapping both cognitive and affective dimensions driving behavior to better understand the need for decision support. PCC interviews demonstrated how clinicians encourage families to take the lead in conversations and ask permission to discuss weight, but fear of causing harm via weight stigma (clinician affect) is a barrier for some PCCs, leading to hesitancy about these conversations. Data from both PCC and additional stakeholder interviews further identified barriers and facilitators associated with CDS implementation. For example, participants spoke about prior implementations of EHR technologies and the types of implementation strategies (eg, live training sessions vs prerecorded webinars vs handouts) users in each setting preferred and why (eg, information being available when they needed it). Knowledge of these barriers provided us with targeted guidance on how, when, and where in the decision-making process to offer decision support.

To define context, workflow data captured in PCC interviews (1) provided an understanding of contextual facilitators and barriers to care, (2) identified key interactions with technology within the care context, and (3) mapped variation in workflows and EHR interactions. These data showed 3 general patterns of EHR use during a patient visit (Figure 4). The first pattern was characterized by users clustering their interactions with the EHR in the middle of the visit. Users from the second pattern described very limited use of the EHR while in the room with the patient and family, often only when showing the growth chart and placing orders at the end of a visit. The final pattern showed repeated interactions with the EHR integrated throughout the visit. These features provided the foundation for designing our CDS and implementation package to meet the needs of different sociotechnical systems and workflows. For example, clinics with colocated dietitians might offer nutritional counseling with a registered dietitian during a routine office visit, a very different workflow from clinics without this specialized support onsite.

To comprehensively characterize the context for the implementation of our CDS package, we examined how care for children with overweight or obesity varied across implementation sites. PCC interview data also defined different prioritized barriers across the CFIR domains of inner and outer settings specific to the patient and caregiver, the clinician and care team, and the health system. For example, different challenges were faced by patients in rural versus urban study sites, which might influence a PCC's management options or decisions. While there were commonalities expressed across settings, such as a safe place for exercise, the solutions to these challenges varied. For example, in urban settings, indoor activities might be recommended whereas in more rural settings there was often a focus on the types of outdoor activities. Barriers and facilitators rooted within the inner (eg, clinic) and outer setting (eg, health system) were considered important potential drivers influencing implementation, informing the

XSL•FO
RenderX

types of implementation strategies that would support these different contexts.

To define requirements, we examined each core functional component of the original trial components alongside data collected during our initial interviews pertaining to past experiences with EHR tools and implementation experiences. We considered how and when each design component would be used in the workflow as well as the anticipated impact on clinician situation awareness and workload. For example, while we anticipated labs would routinely be ordered near the middle to end of a visit, some clinic workflows incorporated early lab ordering to facilitate the information needs of onsite clinical staff who complete lab draws during the visit. Examining design requirements for implementation, we mapped interview findings to known strategies for implementation. For example, the implementation strategies of "identifying and preparing champion" and "conducting educational outreach" were suggested by several stakeholders as important for creating buy-in and supporting clinicians' adoption of the CDS [12,16].

Evaluating Design (UFIT Steps 4 and 5)

Early design sessions identified the need for multiple options for accessing support including a best practice advisory, health maintenance alerts, order sets, and support embedded in a standardized note template. Based on user feedback, health maintenance alerts were later removed from the design. Potential training methods were added in later rounds of iterative design and evaluation to elicit feedback on the types of implementation strategies that should be paired with the specific tools being designed and tested. Feedback on implementation strategies supported the need for multiple training strategies while largely shifting design focus away from audit and feedback strategies.

Discussion

Principal Findings

The design and development of iPOP-UP provided a concrete application of the UFIT as we progressed through our formative evaluation. The UFIT sought to capture a holistic understanding of users that includes their behaviors, cognitive and affective needs, and identifying relevant multilevel barriers and facilitators that are heterogeneous across clinical contexts. Insights from our transdisciplinary study team and direct feedback from users guided our design and iterative refinement as we tailored the iPOP-UP package for each study site. This included a multipronged approach to both CDS and implementation strategy design to meet the various needs and workflows from our qualitative findings. This approach draws on the strengths of UCD, HFE, and implementation science. UCD, which often focuses on initial technology development, provides the process while HFE brings into consideration of theories of situational awareness, and cognitive and affective experiences of users interacting with systems. Finally, the CFIR, which draws from implementation science, adds consideration of even broader contextual factors critical to implementation strategy development. Together these provided the necessary framework to both adapt a CDS and tailor implementation strategies that were combined into a single bundled product: iPOP-UP. Just as UCD methods have been applied more broadly to nontechnology settings [20], though iPOP-UP development was specific to a CDS and its implementation, UFIT is likely to be useful for other evidence-based interventions as well.

Comparison to Prior Work

While CDS tools are designed to support meaningful clinical change in local environments, maximizing adoption and scaling across heterogeneous settings presents many challenges that too often lead to failed implementation. Fitting CDS and integrated implementation strategies into local contexts are essential for high usability and sustained use, yet critical consideration of fit is often overlooked. Prior work has identified characteristics associated with successful CDS implementation [17]. Yet there still remains the need to understand how to tailor these characteristics across contexts. A user-centered, structured approach to tailoring and adapting the technology with context and tailoring implementation strategies that address user needs and navigate complex and diverse local contexts may facilitate dissemination, implementation, and sustained use of CDS. While UCD offers a proven approach to designing new technology, we incorporate HFE and implementation science frameworks into each step of this approach to more comprehensively guide the tailoring of CDS bundled implementation packages across diverse settings.

UCD is increasingly shaping intervention development and has recently been introduced within implementation science [18,21-23]. While existing frameworks, including the Accelerated Creation-to-Sustainment model, have begun to adopt iterative UCD approaches to implementation planning during the creation stage [24], our framework expands on this by providing detailed theoretical constructs to guide the simultaneous design of both technology and implementation strategies. The UFIT is a comprehensive action framework designed to simultaneously guide the design and development of CDS with implementation strategies.

Limitations

We developed the UFIT to guide our formative evaluation; as such, this presents some limitations. First, there are numerous frameworks within both HFE and implementation science that could inform UCD. We selected broad and well-established models, frameworks, and theories for additional structure throughout the process, and we recognize some readers may feel other theories warrant inclusion within our conceptual framework to further expand the steps of UCD for CDS design, development, and implementation. Second, the depth and complexity of the theories we included may require additional understanding for designers to apply to their own work. Our intention is to provide a shared understanding of the goals and topics of inquiry at each step. Finally, while we anticipate our framework will apply beyond CDS development and to chronic diseases beyond obesity, testing in other use cases will define its boundaries of application.

Conclusions

UFIT is a pragmatic framework expanding upon the steps of UCD by explicitly incorporating HFE and implementation science theories to holistically guide CDS design and implementation across diverse settings. It is a process framework

XSL•FO
RenderX

that incorporates the assessment of determinants to inform the simultaneous design of an optimized innovation with tailored implementation strategies [25]. We demonstrate the use of UFIT to optimize the tailoring of this package to local contexts across 3 large health systems. The conceptual framework presented here was developed through transdisciplinary collaboration, and it is our hope that it will provide the structured steps necessary to guide other teams to implement integrated UCD for the design of paired CDS and implementation strategies.

Acknowledgments

This work was conducted in collaboration with the Academic Pediatric Association's Continuity Research Network. Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under award R01MD014853. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. No generative AI was used in any portion of the manuscript writing.

Authors' Contributions

JR, MS, LJD, EBF, HT, and CFA provided the original conceptualization, methodology, and drafting of the writing of the manuscript. All authors provided critical review and editing.

Conflicts of Interest

JJM is the medical director for ECRI Guidelines' Trust, a content engine that supports the dissemination of guidelines for which he receives salary support but is not directly compensated or involved with the dissemination of individual guidelines.

References

1. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
2. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
3. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020;370:m3216 [FREE Full text] [doi: 10.1136/bmj.m3216] [Medline: 32943437]
4. Eichner J. Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations. Rockville, MD: Agency for Healthcare Research and Quality; 2010.
5. Melnick ER, Fong A, Nath B, Williams B, Ratwani RM, Goldstein R, et al. Analysis of electronic health record use and clinical productivity and their association with physician turnover. JAMA Netw Open 2021;4(10):e2128790 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.28790] [Medline: 34636911]
6. Murad DA, Tsugawa Y, Elashoff DA, Baldwin KM, Bell DS. Distinct components of alert fatigue in physicians' responses to a noninterruptive clinical decision support alert. J Am Med Inform Assoc 2022;30(1):64-72 [FREE Full text] [doi: 10.1093/jamia/ocac191] [Medline: 36264258]
7. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. J Biomed Inform 2018;78:134-143 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.005] [Medline: 29246790]
8. Jolles MP, Lengnick-Hall R, Mittman BS. Core functions and forms of complex health interventions: a patient-centered medical home illustration. J Gen Intern Med 2019;34(6):1032-1038 [FREE Full text] [doi: 10.1007/s11606-018-4818-7] [Medline: 30623387]
9. Norman DA, Draper SW. User Centered System Design: New Perspectives on Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates; 1986.
10. Ratwani RM, Fairbanks RJ, Hettinger AZ, Benda NC. Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. J Am Med Inform Assoc 2015;22(6):1179-1182 [FREE Full text] [doi: 10.1093/jamia/ocv050] [Medline: 26049532]
11. Gale RC, Wu J, Erhardt T, Bounthavong M, Reardon CM, Damschroder LJ, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019;14(1):11 [FREE Full text] [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
12. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]

XSL•FO
RenderX

13. Brooke J. SUS—a quick and dirty usability scale. Agency for Healthcare Research and Quality. URL: https://digital.ahrq.gov/sites/default/files/docs/survey/systemusabilityscale%2528sus%2529_comp%255B1%255D.pdf [accessed 2023-01-20]
14. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
15. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR outcomes addendum. Implement Sci 2022;17(1):7 [FREE Full text] [doi: 10.1186/s13012-021-01181-5] [Medline: 35065675]
16. Miech EJ, Rattray NA, Flanagan ME, Damschroder L, Schmid AA, Damush TM. Inside help: an integrative review of champions in healthcare-related implementation. SAGE Open Med 2018;6:2050312118773261 [FREE Full text] [doi: 10.1177/2050312118773261] [Medline: 29796266]
17. Liu S, Reese TJ, Kawamoto K, Del Fiol G, Weir C. A systematic review of theoretical constructs in CDS literature. BMC Med Inform Decis Mak 2021;21(1):102 [FREE Full text] [doi: 10.1186/s12911-021-01465-2] [Medline: 33731089]
18. Dopp AR, Parisi KE, Munson SA, Lyon AR. A glossary of user-centered design strategies for implementation experts. Transl Behav Med 2019;9(6):1057-1064 [FREE Full text] [doi: 10.1093/tbm/iby119] [Medline: 30535343]
19. Endsley MR. Situation awareness misconceptions and misunderstandings. J Cogn Eng Decis Mak 2015;9(1):4-32 [FREE Full text] [doi: 10.1177/1555343415572631]
20. Ray JM, Berg R, Sudikoff SN. PROcess for the Design of User-Centered Environments (PRODUCE): guiding change in the health care environment. Adv Health Care Manag 2019 Oct 24;18:19-33. [doi: 10.1108/S1474-823120190000018002] [Medline: 32077651]
21. Dopp AR, Parisi KE, Munson SA, Lyon AR. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. Implement Sci Commun 2020;1:17 [FREE Full text] [doi: 10.1186/s43058-020-00020-w] [Medline: 32885179]
22. Chen E, Neta G, Roberts MC. Complementary approaches to problem solving in healthcare and public health: implementation science and human-centered design. Transl Behav Med 2021;11(5):1115-1121 [FREE Full text] [doi: 10.1093/tbm/ibaa079] [Medline: 32986098]
23. Lyon AR, Koerner K. User-centered design for psychosocial intervention development and implementation. Clin Psychol (New York) 2016;23(2):180-200 [FREE Full text] [doi: 10.1111/cpsp.12154] [Medline: 29456295]
24. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]
25. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015;10:53 [FREE Full text] [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]

Abbreviations

CDS: clinical decision support
CFIR: Consolidated Framework for Implementation Research
EHR: electronic health record
HFE: human factors/ergonomics
iPOP-UP: Improving Pediatric Obesity Practice Using Prompts
PCC: primary care clinician
UCD: user-centered design
UFIT: User-Centered Framework for Implementation of Technology

Edited by G Eysenbach, T Leung; submitted 17.08.23; peer-reviewed by C Stipelman; comments to author 14.09.23; revised version received 30.11.23; accepted 17.02.24; published 21.05.24.

Please cite as:
Ray J, Finn EB, Tyrrell H, Aloe CF, Perrin EM, Wood CT, Miner DS, Grout R, Michel JJ, Damschroder LJ, Sharifi M
User-Centered Framework for Implementation of Technology (UFIT): Development of an Integrated Framework for Designing Clinical Decision Support Tools Packaged With Tailored Implementation Strategies
J Med Internet Res 2024;26:e51952
URL: https://www.jmir.org/2024/1/e51952
doi:10.2196/51952
PMID:38771622

XSL•FO
RenderX

©Jessica Ray, Emily Benjamin Finn, Hollyce Tyrrell, Carlin F Aloe, Eliana M Perrin, Charles T Wood, Dean S Miner, Randall Grout, Jeremy J Michel, Laura J Damschroder, Mona Sharifi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing and Evaluating Digital Public Health Interventions Using the Digital Public Health Framework DigiPHrame: A Framework Development Study

Tina Jahnel[1], BA, MA, PhD; Chen-Chia Pan[2,3,4], BA, MA; Núria Pedros Barnils[2,3], BA, MA; Saskia Muellmann[2,4], BA, MA, Dr. PH; Merle Freye[2,5], Dr. jur.; Hans-Henrik Dassow[2,6], Bsc, MA; Oliver Lange[2,7], Bsc, MA, Dr. rer. pol.; Anke V Reinschluessel[2,8,9], Bsc, Msc, PhD; Wolf Rogowski[2,7], Dr. oec. publ.; Ansgar Gerhardus[2,10], MA, MPH, Dr. med.

[1]Department of Health Services Research, Institute for Public Health and Nursing Research, University of Bremen, Bremen, Germany
[2]Leibniz ScienceCampus Digital Public Health, Bremen, Germany
[3]Department of Prevention and Health Promotion, Institute for Public Health and Nursing Research, University of Bremen, Bremen, Germany
[4]Leibniz Institute for Prevention Research and Epidemiology, Bremen, Germany
[5]Institute for Information, Health and Medical Law, University of Bremen, Bremen, Germany
[6]Institute for Philosophy, University of Bremen, Bremen, Germany
[7]Department of Health Care Management, Institute for Public Health and Nursing Research, University of Bremen, Bremen, Germany
[8]Digital Media Lab, University of Bremen, Bremen, Germany
[9]Human-Computer Interaction Group, University of Konstanz, Konstanz, Germany
[10]Department for Health Services Research, Institute for Public Health and Nursing Research, University of Bremen, Bremen, Germany

Corresponding Author:
Tina Jahnel, BA, MA, PhD
Department of Health Services Research
Institute for Public Health and Nursing Research
University of Bremen
Grazer Str 4
Bremen, 28359
Germany
Phone: 49 042121868808
Email: jahnel@uni-bremen.de

Abstract

Background: Digital public health (DiPH) interventions may help us tackle substantial public health challenges and reach historically underserved populations, in addition to presenting valuable opportunities to improve and complement existing services. However, DiPH interventions are often triggered through technological advancements and opportunities rather than public health needs. To develop and evaluate interventions designed to serve public health needs, a comprehensive framework is needed that systematically covers all aspects with relevance for public health. This includes considering the complexity of the technology, the context in which the technology is supposed to operate, its implementation, and its effects on public health, including ethical, legal, and social aspects.

Objective: We aimed to develop such a DiPH framework with a comprehensive list of core principles to be considered throughout the development and evaluation process of any DiPH intervention.

Methods: The resulting digital public health framework (DigiPHrame) was based on a scoping review of existing digital health and public health frameworks. After extracting all assessment criteria from these frameworks, we clustered the criteria. During a series of multidisciplinary meetings with experts from the Leibniz ScienceCampus Digital Public Health, we restructured each domain to represent the complexity of DiPH. In this paper, we used a COVID-19 contact–tracing app as a use case to illustrate how DigiPHrame may be applied to assess DiPH interventions.

Results: The current version of DigiPHrame consists of 182 questions nested under 12 domains. Domain 1 describes the current status of health needs and existing interventions; domains 2 and 3, the DiPH technology under assessment and aspects related to human-computer interaction, respectively; domains 4 and 5, structural and process aspects, respectively; and domains 6-12,

XSL•FO
RenderX

contextual conditions and the outcomes of the DiPH intervention from broad perspectives. In the CWA use case, a number of questions relevant during its development but also important for assessors once the CWA was available were highlighted.

Conclusions: DigiPHrame is a comprehensive framework for the development and assessment of digital technologies designed for public health purposes. It is a living framework and will, therefore, be updated regularly and as new public health needs and technological advancements emerge.

(J Med Internet Res 2024;26:e54269) doi:10.2196/54269

KEYWORDS

digital public health; DiPH; digital health; public health; framework; development; evaluation; guidance; program evaluation; telemedicine; intervention; technological advancement; opportunity; technology; implementation; digital public health framework; DigiPHrame; scoping review; COVID-19; contact-tracing app; contact tracing

Introduction

Background

The overarching goal of public health is to promote and improve the health and well-being of people and communities. In recent years, digital interventions specifically designed for public health purposes have emerged on a large scale. Digital public health (DiPH) interventions may help us tackle substantial public health challenges, including aging populations [1], the dual burden of noncommunicable and communicable diseases [2], and the health impacts of climate change [3]. Moreover, DiPH interventions present valuable opportunities to improve and complement existing health care services and reach historically underserved populations.

With the COVID-19 pandemic, we have seen how digital technologies may accelerate responses to public health emergencies. For example, digital contact-tracing apps have become a major component to monitor community transmission and curb the spread of the virus in a population [4]. Further, the development of information platforms for international real-time public health data has supported policy and decision makers in planning and executing containment strategies. Another relevant field that became more visible during the pandemic concerns public health education. Digital platforms of health authorities and national agencies played a critical role in rapidly engaging and educating the population through prompt dissemination of trusted and tailored public health information, while limiting the visibility of information from unreliable sources [5].

As with other health technologies, DiPH interventions need to be developed through an iterative process, considering a multitude of factors right from the beginning of the conceptualization process. However, these factors (eg, acceptability, usability, data security, and sustainability) are sometimes not well thought out during the development or not at all considered, often resulting in low-value interventions that are ineffective and burdensome and reduce both quality and efficiency. In turn, the development of DiPH interventions is often triggered through technological advancements (ie, what is possible) rather than current public health needs [6].

Although vast amounts of new health apps are launched in app stores regularly, the number of downloads for many of these apps generally stays notoriously low [7]. Individual decisions around the initial use, adoption, rejection, and continued use of an app might be influenced by concerns regarding data security and data protection issues, costs to purchase the app, or user-friendliness for different user groups [8,9]. Other societal aspects, such as sustainable financing and regulatory requirements, are described as challenges to fulfill public health functions. Thus, these aspects may influence the design of a DiPH intervention and need to be considered from the beginning of the development process [10].

During the development and evaluation process, a number of different stakeholders assess the potential impact of DiPH interventions (eg, tech companies, health insurances, governments, and health organizations). As such, for each DiPH intervention, a great variety of potential users and user environments must be considered. To systematically develop and evaluate DiPH interventions, a comprehensive framework is needed that systematically covers all aspects with relevance for public health. This includes considering the complexity of the technology, the context in which the technology is supposed to operate, its implementation, and its effects on public health (eg, ethical, legal, and social aspects). Such a comprehensive framework would cover all phases, from conceptualization to evaluation, of all types of DiPH interventions and all parties [11].

Existing Frameworks for Digital Health Interventions, Health Technologies, and Public Health Interventions

Interventions are often developed without a systematic method and without drawing on the evidence and theories. This point was made by Martin Eccles, Emeritus Professor of Clinical Effectiveness in the United Kingdom, in referring to a frequently used principle of intervention design, the ISLAGIATT (It Seemed Like A Good Idea At The Time!) principle. This means that we jump straight to intervention and crucially miss out understanding the behaviors we are trying to change or do not consider contextual facilitators and barriers for a successful implementation of the intervention. Frameworks that integrate a wide range of domains allow us to think ahead and help us avoid potential pitfalls before they occur so that we can design appropriate interventions based on this analysis [12].

Although frameworks for digital health interventions, health technologies, and public health interventions have been developed previously, to the best of our knowledge, no framework for the systematic development and assessment of digital interventions for public health purposes exists today. Assessment criteria for health-related technologies have been developed previously, although their focus generally lies on

XSL•FO
RenderX

either health technology [13,14] or digital health-relevant [15] aspects.

One prominent example of assessing various health technologies is health technology assessment (HTA). "HTA is a multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its lifecycle. The purpose is to inform decision-making in order to promote an equitable, efficient, and high-quality health system" [16]. Based on this methodology, various organizations have developed frameworks with different foci [13,14,17]. For instance, the European Network for Health Technology Assessment (EUnetHTA) developed the health technology core model for assessing the dimensions of value to facilitate the production and sharing of HTA information, such as evidence on efficacy, effectiveness, and patient aspects, to inform decisions. The model has a broad scope and offers a common ground to various stakeholders by offering a standard structure and a transparent set of proposed HTA questions [13]. HTA frameworks are generally applied to already developed technologies rather than providing standards for evaluation aspects that should be considered throughout development. However, this is important because existing interventions would likely be outdated by the time their assessment is finished.

Assessment frameworks specifically designed for the evaluation of digital health technology also exist. The National Institute for Health and Care Excellence (NICE) recently developed an Evidence Standards Framework (ESF) for digital health technologies [15], aiming at providing standards for clinical evidence of (novel) health technology's (cost-)effectiveness within the UK health and care system. Similar to other frameworks [18-21], the ESF lacks applicability to public health technologies, due to its focus on clinical outcomes. Other frameworks focus on evaluation and assessment criteria along the life cycle of digital health interventions yet still lacking a public health focus [22].

Digital interventions heavily rely on user interaction and engagement. However, public health frameworks generally do not include specific measures to assess usability, user experience, and the design aspects crucial for promoting sustained user engagement [23]. Furthermore, DiPH interventions often require integration into existing health care systems, which can be complex and fraught with interoperability, data security, and data protection challenges—issues that are often not properly addressed in public health frameworks [24]. Although these are just a few examples, they illustrate how unique aspects of DiPH interventions may fall short in existing public health frameworks.

Gaps and Objective

Together, we identified the following gaps:

- Absence of a framework for digital interventions in public health: Although frameworks for health technologies and public health interventions exist, there is no established framework specifically tailored for the systematic development and assessment of digital interventions in public health.
- Limited applicability of existing assessment frameworks.
- Inadequate consideration of usability and integration challenges.

Addressing these gaps requires the development of a comprehensive framework specifically tailored for digital interventions in public health, integrating diverse domains and considering usability, user experience, and integration challenges throughout the development and assessment process so that developers and assessors need not draw on multiple frameworks. The main focus of this paper was to present the current form of the digital public health framework (DigiPHrame) and describe its development process, followed by a use case to illustrate its application. More detailed information on the scoping review that served as a starting point to develop DigiPHrame can be found in the protocol that we preregistered on the Open Science Framework (OSF) [25]. The German contact-tracing app Corona-Warn-App (CWA), as a digital public warning system with a clear public health focus, was deemed as a suitable use case to illustrate the application of DigiPHrame.

Methods

Search Strategy

We developed DigiPHrame in several steps, as shown in Figure 1. First, we conducted a scoping review to identify existing frameworks for public health and digital health interventions (see the protocol and registration on the OSF [26]). See Table 1 for the eligibility criteria. For information sources, we searched journal papers in the electronic literature databases MEDLINE (via PubMed), Scopus, IEEE, CINAHL (via EBSCO), and PsycINFO (via Ovid). Our search strategy was first developed around our core concepts as our primary search keywords and Boolean operators: ("Public Health" [Title/Abstract] OR "Digital Health" [Title/Abstract]) AND Evaluation [Title] AND Framework [Title]). The search syntax was then expanded to include the synonyms, wildcards, and relevant subject terms of the primary keywords to increase the sensitivity of our searches. Next, we modified the subject terms and search field of the search syntax to adapt to each database (see Multimedia Appendix 1). We also manually searched relevant reviews' reference lists. The final search was completed on April 12, 2022, with no publication date limitations.

XSL•FO
RenderX

Figure 1. Methodological summary of the DigiPHrame development approach. DigiPHrame: digital public health framework.

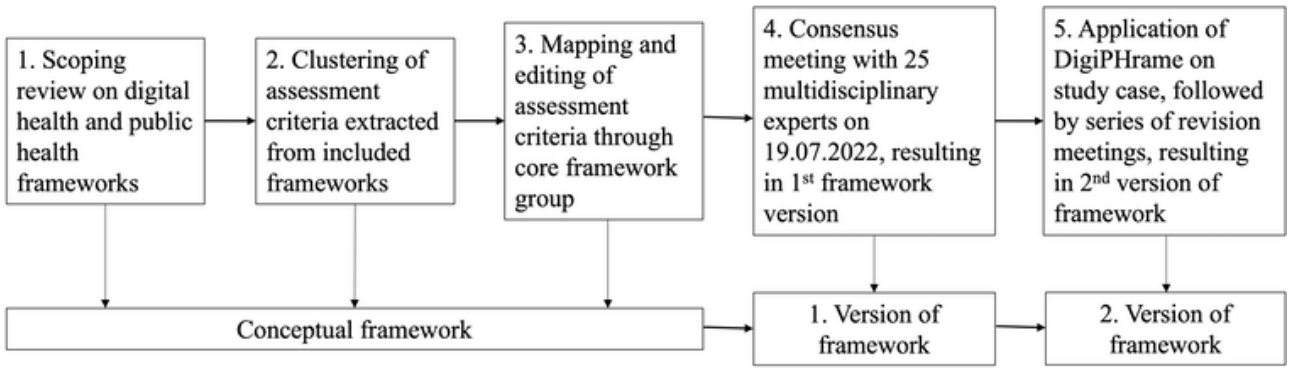

Table 1. Eligibility criteria for the scoping review to identify existing frameworks for public health and digital health interventions.

Criterion	Inclusion	Exclusion
Framework	Development or evaluation framework for health interventions related to public health or digital health	No framework or guidance in the report
Report	Framework or guidance outlining the standards, principles, criteria, or properties needed to support the systematic development or evaluation of health interventions aimed at health promotion or prevention with or without digital technologies	Framework or guidance not focusing on developing, monitoring, validating, or evaluating health interventions; not providing specific standards, principles, criteria, or properties; only designed for 1 specific tool; and not applicable to other health interventions and only applicable to pharmaceutical/surgical/clinical/rehabilitation interventions
Publication type	Journal papers, study/policy/program reported in gray literature	Comment, correction, letter, editorial, protocol, oral presentation, poster
Language	English	Other language than English
Access to full text	Access to full text of studies selected for data coding	No access to full text

Selection of Sources of Evidence

After deduplication, 4830 titles and abstracts were screened by 2 researchers independently, resulting in 433 (9%) full texts, which were then assessed by 2 independent researchers. Disagreements between researchers were resolved through dialogue, with the involvement of a third party, if necessary, although a definitive agreement score was not established. In total, 68 (15.7%, see Multimedia Appendix 4) papers were included for data extraction (see the Preferred Reporting Items for Systematic reviews and Meta-Analyses [PRISMA] flowchart in Multimedia Appendix 2 [27]).

Data Charting and Data Items

We extracted all pertinent assessment criteria from the frameworks identified through the scoping review. Initially, these criteria were assigned to HTA domains/subdomains [13], although several criteria could not be assigned due to thematic misfit (akin to deductive coding). Subsequently, new categories were formed (akin to inductive coding). One researcher performed the coding initially, followed by a collaborative examination of the coded sections by 2 researchers, leading to adjustments during the discussion process (eg, reassignment to other domains, reassignment to other subdomains within a domain, summarization of subdomains, and deletion of irrelevant domains or subdomains). Questions describing the subdomains were devised by us based on the criteria (here, too, a proposal was made by one person, followed by verification by a second person). We consulted additional literature for the categorization of ethical principles [28].

A group of multidisciplinary experts from the Leibniz ScienceCampus Digital Public Health (LSC DiPH) were assigned to the domains corresponding to their expertise for counseling. Each domain was restructured with proficient inputs to represent the complexity of DiPH. Where necessary, additional literature was consulted, especially when the included frameworks fell short of offering criteria specific to DiPH.

The Proposed Framework

A first draft of the proposed framework was sent to an interdisciplinary expert panel consisting of 105 members of the LSC DiPH. Feedback was gathered as unrestricted comments on the domains we developed. We reached out to experts from diverse fields, including medicine, public health, global health, psychology, sociology, human-computer interaction, (health) economics, informatics, sports science, medical biometry, architecture, urban planning, statistics, ethics, policy analytics, and law, assigning them domains based on their respective expertise. A deadline for feedback submission was set for July 18, 2022. Additionally, the same members of the LSC DiPH were invited to partake in a consensus meeting held on July 19, 2022. Participants were grouped into domain-specific discussions according to their areas of expertise, with these discussions being moderated by the DigiPHrame team. This resulted in the first version of the framework [29].

XSL•FO
RenderX

Use Case: Corona-Warn-App

Shortly after the COVID-19 pandemic in 2020, numerous digital contact-tracing apps were developed or proposed, with official government support in some territories and jurisdictions. The rationale was that contact tracing is an important tool in infectious disease control, but as the number of cases rises, time constraints make it more challenging to trace transmissions effectively. Digital contact tracing, especially if widely deployed, may be more effective than traditional methods of contact tracing [30].

COVID-19 apps include mobile apps for digital contact tracing (ie, identifying persons, or "contacts," who may have been in contact with an infected individual) deployed during the COVID-19 pandemic. Privacy concerns have been raised, especially about systems tracking users' geographical location. Alternatives include co-opting Bluetooth signals to log a user's proximity to other smartphones. For example, the open source CWA funded by the German government was based on proximity tracing using Bluetooth signals. The app provides a function for users to warn other users by uploading their positive test results anonymously on a voluntary basis to the CWA server. Users are then notified about any contacts with infected persons and can get tested on a voluntary basis.

The same experts were invited to a workshop on February 23, 2023, where the proposed framework was applied to a study case and tested for face validity. The case study was the CWA. Followed by several revision meetings by the framework team between February and May 2023, the second version of the proposed framework was finalized in May 2023 [31].

Ethical Considerations

Ethical approval is not applicable for this study since it did not involve human subjects.

Results

Characteristics of DigiPHrame

DigiPHrame comprises a set of criteria framed as open-ended questions clustered within domains that lead interested parties through a broad spectrum of crucial elements when developing and evaluating DiPH interventions. The evolution of domains and subdomains through the stepwise process, including the number of questions per subdomain in each version, can be found in Multimedia Appendix 3. The framework in its current form was uploaded to the LSC DiPH website and the OSF [32] in May 2023 and is a revised version of the original framework that was first published in July 2022. In total, DigiPHrame consists of 182 questions, structured by 12 domains (Figure 2).

Figure 2. Summary of DigiPHrame domains for developing and evaluating DiPH interventions. DiPH; digital public health; DigiPHrame: digital public health framework.

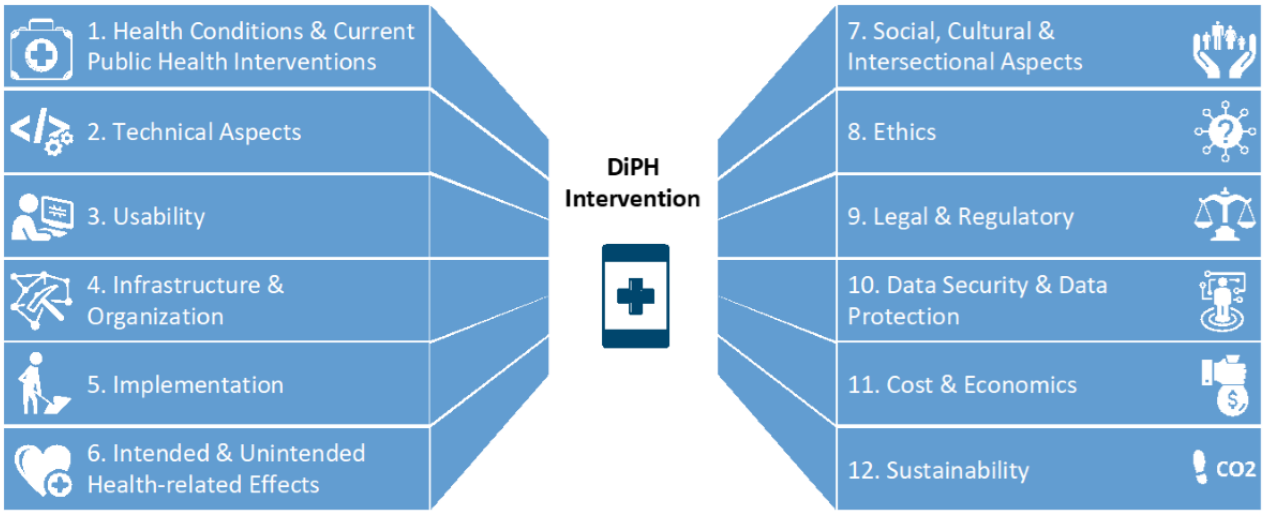

Domain 1 describes the current status of health needs and existing interventions; domains 2 and 3 are aimed at the DiPH technology under assessment and aspects related to human-computer interaction, respectively; domains 4 and 5 are aimed at structural and process aspects, respectively; and domains 6-12 assessment criteria address contextual conditions and the outcomes of the DiPH intervention from broad perspectives.

Next, we defined the domains and illustrated how DigiPHrame can be applied using the CWA as a use case. The CWA is a digital public warning system that was designed and developed during the COVID-19 pandemic and has a clear public health focus. We briefly outlined the purpose and characteristics of the CWA. From each domain, we applied 1 assessment question as an example.

Use Case: Corona-Warn-App

Domain 1: Health Conditions and Current Public Health Interventions

Domain 1 involves background information for DiPH interventions, describing the population, conditions, and observance of health inequities. Furthermore, this domain addresses current public health interventions and common alternatives.

Question 1.5 asks, "What is the expected level of digital literacy of the target population?" In the case of the CWA, the target

XSL•FO
RenderX

population comprises the entire population within a geographically delimited space. Therefore, the entire spectrum of digital literacy is to be expected. Thus, different forms of representation (eg, graphics, text, and sound) of risk exposure and other information related to the COVID-19 pandemic must be available, which was not the case and might have prevented people from using it.

Domain 2: Technology and Usability

Domain 2 guides one through assessing general technical aspects of the health technology of interest. The questions focus on what digital tools are applied and how aspects such as interoperability, data integration, internet connectivity, and others are integrated.

Question 2.17 is, "Does the software require an internet connection (eg, all the time, once in a while, or once)?" In the case of the CWA, an internet connection is necessary as the major functionality of warning people is distributed via the internet. Only a fraction of the available functions work completely without an internet connection, such as the contact diary. Generally, the app does not need a continuous internet connection. However, the device on which the app is installed needs to be connected to the internet, at best multiple times a day but at least once a day to sync the contacts and update on test results.

Domain 3: Usability

Domain 3 focuses on how usable the health technology system is in order to ensure that its users can perform the required tasks (ie, the intended function) safely, effectively, efficiently, and with satisfaction. Therefore, accessibility, user empowerment, credibility, and trustworthiness are also considered in this domain.

Question 3.3 asks, "Are the health technology and DiPH intervention available in relevant languages?" When the CWA was first launched, it was available only in German and English. Russian, 1 of the most spoken immigrant languages in Germany, was not provided in the CWA until much later versions. Since version 2.20.0 for iOS and version 2.20.4 for Android, the CWA is available in German, English, Turkish, Bulgarian, Polish, Romanian, and Ukrainian.

Domain 4: Infrastructure and Organization

Domain 4 on structural aspects considers the structure of the context in which a DiPH intervention is developed and implemented, as well as the involved stakeholders. Question 4.4 asks, "Is the DiPH intervention flexible to suit local, cultural, or social needs?" Initially, the German government promoted centralized storage of user data, which, according to the Federal Ministry of Health, would allow it to better track the spread of infections. However, this led to resistance from digital experts and data protectionists. As a consequence, the CWA was developed, with decentralized data collection across various servers. This approach ensured that the data could be decoupled, thereby hindering any potential tracing of app users.

Domain 5: Implementation

Domain 5 describes aspects to consider before and during integration of a DiPH intervention into the health care system to ensure that the intervention is delivered properly. The domain focuses on the theory used for implementing the DiPH intervention, infrastructure, process, and agents, as well as implementation outcomes and dissemination.

Question 5.9 asks, "Which implementation difficulties (eg, duration, scope, disruptivity, centrality, complexity, and the number of steps required) did the DiPH intervention encounter?" In case of the CWA, necessary features (eg, sharing test results and embedding vaccination certificates) were not initially available when the app was first launched in June 2020, but had to be continuously added to the app.

Contextual Conditions and Outcome-Related Domains

Domain 6: Intended and Unintended Health-Related Effects

Domain 6 considers the positive and negative effects of a DiPH intervention on physical, mental, and social health; the quality of life and well-being; and the knowledge, beliefs, and behavior of individuals and the population in the short, intermediate, and long terms.

Question 6.2 asks, "To what extent is the DiPH intervention expected to impact the physical, mental, and social health of the individual and the population?" With its goal to prevent infections, the CWA was expected to positively affect individuals' and, ultimately, population health. It is unclear how the large red warning sign displayed on users' smartphones when a high-risk contact with an infected person occurs would affect their mental health. Although generally accepted, the CWA was not used by the majority of the population and was widely discussed in terms of data privacy concerns prior to the launch of the app. With some individuals using the CWA and some not (including, sometimes, strong opinions in favor or against the benefit of the app within a social circle), this may have affected an individual's relationships and social health.

Domain 7: Social, Cultural and Intersectional Aspects

Domain 7 examines the societal, cultural, and intersectional dimensions pertinent to communities and groups of individuals, such as ethnic or demographic groups, people residing in the same neighborhood, those sharing common interests, or individuals with specific physical or mental conditions.

Question 7.5 asks, "Which factors in the society/community are relevant for DiPH intervention implementation?" In the case of the CWA, it is the availability of compatible smartphones (eg, older smartphones are not compatible), trust that data will be protected and not used for other purposes (eg, analog data from guests of restaurants, not the data from the app, were used to identify suspects of thefts), and willingness to enter one's data in the case of infection.

Domain 8: Ethics

Domain 8 addresses the moral considerations that arise from the implementation of DiPH interventions. The categorization of ethical principles is based on the influential *Principles of Biomedical Ethics* by Beauchamp and Childress [28].

Question 8.20 asks, "Does the DiPH intervention discriminate against particular segments of the target population?" Although efforts were successively visible to avoid discrimination, it took

XSL•FO
RenderX

too much time to offer the app in different languages frequently spoken in Germany. People using phones with older operating systems were also excluded from applying the app.

Domain 9: Legal and Regulatory Aspects

Domain 9 generates awareness about which areas of law must be considered when developing or evaluating DiPH interventions. It is not the purpose of the domain to pose every specific legal question that has to be answered in order to develop or evaluate DiPH interventions. Since laws differ from country to country, the domain helps detect fields of law and typical problems in those fields that could be relevant for developers and evaluators. The applicable law and its requirements depend on the country.

Question 9.6 asks, "Have you considered the potential reimbursement of DiPH interventions in a national health system (some countries may have specific requirements for reimbursement)?" This raises awareness about the requirements for reimbursement of the DiPH interventions in a national health system or for other payers. Regarding the CWA, the provider offered the intervention for free (without a reimbursement option) because the free-of-charge offer of the CWA promises a broader and quicker distribution of the app.

Domain 10: Data Security and Data Protection

Domain 10 focuses on the technological protection of data and, therefore, combines the aspects of data confidentiality, data integrity, data authenticity, data availability, and data controllability. Data protection relates to whether the system is allowed to process personal data.

If personal data are transferred to third parties, question 10.25 asks, "Is there is a legal basis for the transfer, and are the requirements of the legal basis fulfilled?" Regarding the case of the CWA, T-Systems International GmbH and SAP Deutschland SE & Co. KG are acting on the Robert Koch Institute's behalf. The legal basis is a contract that is binding on the processor with regard to the controller and that sets out the subject matter and duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects, and the obligations and rights of the controller (Art. 28(3) of the General Data Protection Regulation [GDPR]). Otherwise, the Robert Koch Institute only passes on data to third parties if it is legally obliged to do so or if this is necessary for legal action or criminal prosecution in the case of attacks on the app's technical infrastructure.

Domain 11: Cost and Economics

Domain 11 assesses DiPH interventions regarding whether they can be considered a rational use of scarce resources. Question 11.1 asks, "Which relevant costs and effects can be identified?" Considering the costs and effects of a DiPH intervention from the beginning could help compare it with other interventions and show that it is economically dominant (ie, it is at least as effective as but costs less than the alternative interventions). Further, this information might be the basis for health economic evaluation (see questions 11.4- 11.6) to see whether what it costs per health gain is considered acceptable by the payer. In the case of the CWA, there are various relevant costs of the intervention itself, such as development and operation (2020: €52.8, or US $57.5, million; 2021: €63.5, or US $69.1, million) and promotion (2020 and 2021: €13.7, or US $14.9, million) [16]. Taking a broader (societal) perspective, there might be further costs, such as costs of further testing when the CWA receives a warning and costs of unrelated survival gains or benefits, such as a reduction in the loss of earnings, reduction in hospitalizations and rehabilitation measures, and reduction in deaths [32]. However, to the best of our knowledge, the pandemic context and the decision process about the CWA lead to a situation where a decision was made without formally considering cost-effectiveness in comparison with alternative decision options.

Domain 12: Sustainability

Domain 12 asesses environmental, social, and economic sustainabilty. Given the goal to reduce carbon emissions in health care, question 12.1 asks, "Which resources are necessary to develop and maintain the DiPH intervention?" In the case of the CWA, servers need to run, which produces carbon emissions, and computers need to be obtained to ensure compatibility of health offices with the CWA. Measuring and evaluating these resource consumptions also allows decision makers to consider more climate-friendly design alternatives for DiPH interventions.

Application of DigiPHrame for the Development and Evaluation of DiPH Interventions

In the use case of the CWA, we highlighted a number of questions relevant during the development of the CWA but also important for assessors once the CWA was available. For example, developers needed to consider how the data would be collected and shared without interfering with data privacy and data protection laws. Similarly, assessors needed to find ways of evaluating the effectiveness of the CWA (eg, did the CWA prevent infections?) without relevant data (due to decentralized data storage, data from different individuals could not be connected, and thus, only estimates could be determined). In future scenarios, DigiPHrame can serve as a checklist for both developers and assessors to help them avoid overlooking key issues with relevance to the performance of the intervention. Although for some questions, it might be enough to use common sense (in the case of the CWA, it could be questions surrounding the usability of the app), for others, specialist expertise may be necessary (eg, questions regarding legal and regulatory issues).

DigiPHrame is agile and primarily user led (Textbox 1). We deliberately included the option of feedback loops in the framework to support the agile development process. Although it is advised to consider all domains and respective questions, developers may decide which domains are assessed at what stage of their development process and which questions are relevant for the respective DiPH intervention. For an intervention under development, a first orientation might be enough to understand whether it is worth continuing along the determined path or whether adjustments might be necessary. Developers may also decide to put specific questions on hold and revise them at a later stage in case any changes or additions need to be made to the DiPH intervention. Similarly, assessors may delay answering certain questions in case no robust evidence is available at the time to answer the questions.

XSL•FO
RenderX

Textbox 1. How to use the digital public health framework (DigiPHrame).

Users of DigiPHrame are encouraged to first answer a list of questions regarding a general description of their digital public health (DiPH) intervention. Providing general characteristics will help assessors better understand the DiPH intervention under assessment. DigiPHrame is further equipped with a standardized answer scheme to help developers in answering the questions and, if necessary, plan the next steps in the development process. For assessors, the answer scheme can serve as a checklist to tick off all relevant criteria.

DigiPHrame users can respond to each question using the provided answer scheme. The first 2 assessment indicators are "not applicable" when the question is irrelevant to the particular DiPH intervention and "assessment result" to provide the answer or additional information to the assessor. The last 3 columns of the answer scheme focus on the current status of the DiPH intervention during the assessment. These columns include "Assessment completed and sufficient" when the assessment is finished and satisfactory, "Assessment done but improvement needed" when the assessment is complete but indicates the need for improvements or changes to the DiPH intervention, and "Assessment only partially done or not possible yet" when the assessment is incomplete or not feasible at the moment.

Example answer scheme:

Criterion: population

Question: Who is the target population of the DiPH intervention?

Assessment indicator scheme:

- Not applicable (N/A): yes
- Assessment result: The entire population is at risk of getting infected with SARS-CoV-2.
- Assessment completed and sufficient: yes
- Assessment done but improvement needed: Briefly outline the necessary changes/expected date for revising the question.
- Assessment only partially done or not possible yet: Insert specific steps to be taken/expected date of completion.

Discussion

Principal Findings

A Unified Framework for Digital Interventions With a Public Health Focus

Although health-related digital technologies hold great potential for enhancing public health and addressing health-related inequalities at a relatively low cost, new developments are often driven by technological advancements and assessments and primarily revolve around clinical aspects of health. To the best of our knowledge, no existing frameworks consider digital interventions specifically designed for public health purposes. Additionally, previous frameworks primarily emphasize clinical aspects when addressing digital health technologies, neglecting the public health perspective. As an example, although the ESF [15] emphasizes clinical outcomes, crucial for any intervention's success, it omits essential aspects, such as sociocultural, ethical, legal, and sustainability factors, vital for effectively implementing DiPH interventions. DigiPHrame includes aspects regarding clinical outcomes (eg, domain 6: intended and unintended health-related effects), among others derived from the ESF, but also the above-mentioned factors. Moreover, although HTA frameworks [13] are often designed for evaluating existing technologies, our objective was to devise a comprehensive framework applicable across all stages of development and evaluation. DigiPHrame adopts a comprehensive public health perspective and can serve as a guide, specifically for developers and assessors throughout the entire development and assessment of DiPH interventions. DigiPHrame provides users with criteria concerning clinical effectiveness, technical functions, and usability, as well as organizational, legal, ethical, economic, and sociocultural aspects. Users have the flexibility to determine the relevant domains and assessment questions based on their specific needs and the stage of the process without relying on multiple development and assessment frameworks. Furthermore, the users of DigiPHrame are encouraged to take a broader view and may be inspired to include other perspectives that were not initially within their scope (eg, sociocultural aspects, ethics, and sustainability).

A Holistic Framework Covering Relevant Domains for DiPH Interventions

Additionally, a deeper understanding of contextual factors is necessary to assess what will work in one country versus another. These factors can either enhance or hinder the adoption and diffusion of DiPH technologies. Although many frameworks tend to overemphasize technical aspects, it is essential to acknowledge that various other factors influence success or failure. These factors include disparities in health expenditure, demographic conditions, health infrastructure, information and communication technology (ICT) skill levels, digital health literacy, clinical and patient engagement, and many more. Recognizing and understanding these key differences within and across countries is crucial for policy makers and other stakeholders in public health and DiPH. Although our framework considers these factors, future work needs to apply DigiPHrame in diverse contexts and countries to validate and continuously update the current version of the framework. In addition, although our framework aims to be universally applicable to various DiPH technologies, it will require revision as new public health needs and DiPH technologies emerge. Therefore, our framework can serve as the foundation for a development and assessment toolkit that developers, decision makers, and other users alike can use.

As we illustrated with the German contact-tracing app CWA, that was launched during the first wave of the COVID-19

XSL•FO
RenderX

pandemic, DigiPHrame can be applied for all stages, including design, implementation, and evaluation. This may have helped avoid potential pitfalls from the beginning that would have otherwise occurred further down the development process.

Strengths and Limitations

Our framework has several key strengths that set it apart. First, it is based on a comprehensive scoping review of digital health and public health frameworks (OSF [25]), ensuring a robust foundation. Additionally, we conducted scientific consensus meetings involving interdisciplinary experts, ensuring a breadth of perspectives in its development. Second, the assessment themes within our framework were derived from existing frameworks developed in various Western countries, including Germany, the United Kingdom, and the United States. This demonstrates the broad applicability of DigiPHrame across different geographical contexts, making it adaptable to diverse settings in high-income countries. Another strength of our framework is its universality. It is not limited to specific types of DiPH interventions and, therefore, can be applied to any digital intervention with the overarching aim of improving public health outcomes. This flexibility allows for its widespread application across a wide range of interventions. Furthermore, DigiPHrame is designed as a living framework that will evolve and adapt as technology advances. To do so, we will continue to revise the domains and questions and regularly test any changes for face validity using a variety of use cases. This will ensure that it remains relevant and up to date in the fast-paced DiPH landscape, accommodating emerging technologies and methodologies. Lastly, we incorporated input and expertise from various research fields throughout the entire development process of DigiPHrame. We fostered an interdisciplinary perspective by involving experts from different disciplines, including public health, epidemiology, psychology, philosophy, law, economics, human-computer interaction, and sociology, enriching the framework with diverse insights and knowledge.

Although our framework has several strengths, it is important to acknowledge certain limitations. First, going through the proposed framework might require significant time and expertise due to its complexity and depth. Nevertheless, it is flexible; it is up to the assessor to decide which domains and criteria are applicable to their specific case. This flexibility is advantageous, allowing the framework to be adapted to diverse contexts and DiPH interventions. However, it may also introduce subjectivity in the evaluation process, as different assessors may choose different domains and criteria, leading to varying outcomes. Ensuring transparency and consistency in domain selection could help mitigate this concern. Additionally, we intend to develop a condensed version of the framework focusing on the most critical domains and questions. Second, we engaged experts from diverse research fields to address potential inconsistencies during the development process. However, it is worth noting that the majority of our consultations did not extend to a broader geographical range, particularly in terms of incorporating specific aspects from low- and middle-income countries. It is crucial to recognize that contexts may differ significantly, including factors such as technology accessibility, digital health literacy, and legal requirements. Although DigiPHrame aims to be applicable across different geographical contexts, users of the framework are advised to consider and adhere to their local requirements and nuances. Furthermore, in our scoping review, we focused on primary prevention and health promotion but not on secondary and tertiary prevention (eg, rehabilitation). This could have limited the frameworks and criteria we found. Although as per our definition, DiPH focusses on primary prevention and health promotion, future research may also include frameworks focused on secondary and tertiary prevention. Lastly, we did not provide any evaluation methods along with the framework. As DigiPHrame evolves, however, our goal is to provide suitable existing methods and develop novel evaluation methods for DiPH interventions.

Conclusion

DigiPHrame is a comprehensive framework for the development and assessment of digital technologies designed for public health purposes. Our framework may assist in designing and evaluating DiPH interventions that serve public health needs rather than displaying technological advancements. Moreover, DigiPHrame may help avoid overlooking important aspects, such as acceptability, usability, data security, and sustainability, which would otherwise result in low-value interventions that are not user friendly, violate (data protection) laws, or are not sustainable. We aim to revise and improve DigiPHrame as new technologies emerge, and encourage developers and assessors to use and contribute to improving DigiPHrame.

Acknowledgments

The authors gratefully acknowledge the Leibniz ScienceCampus Bremen Digital Public Health support, jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen, and the Leibniz Institute for Prevention Research and Epidemiology (BIPS).

The authors would also like to thank Dorothee Jürgens, Sarah Janetzki, Sarah Forberger, and Jonathan Kolschen for their contributions to conducting the scoping review and data extraction for developing the first version of the framework.

Data Availability

The data collected and analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

TJ and AG conceived the concept of the manuscript. TJ drafted the first version of the manuscript. All authors contributed to the literature search and writing and editing of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search syntax.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v26i1e54269_app1.xlsx]

Multimedia Appendix 2
Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram.
[PDF File (Adobe PDF File), 69 KB - jmir_v26i1e54269_app2.pdf]

Multimedia Appendix 3
Evolution of domains and subdomains.
[DOCX File , 27 KB - jmir_v26i1e54269_app3.docx]

Multimedia Appendix 4
Included Reports.
[DOCX File , 19 KB - jmir_v26i1e54269_app4.docx]

References

1. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023 Jul 18;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
2. Rosen JM, Kun L, Mosher RE, Grigg E, Merrell RC, Macedonia C, et al. Cybercare 2.0: meeting the challenge of the global burden of disease in 2030. Health Technol (Berl) 2016;6(1):35-51 [FREE Full text] [doi: 10.1007/s12553-016-0132-8] [Medline: 27358760]
3. Rahimi-Ardabili H, Magrabi F, Coiera E. Digital health for climate change mitigation and response: a scoping review. J Am Med Inform Assoc 2022 Nov 14;29(12):2140-2152 [FREE Full text] [doi: 10.1093/jamia/ocac134] [Medline: 35960171]
4. Murray CJL, Alamro NMS, Hwang H, Lee U. Digital public health and COVID-19. Lancet Public Health 2020 Sep;5(9):e469-e470 [FREE Full text] [doi: 10.1016/S2468-2667(20)30187-0] [Medline: 32791051]
5. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
6. Zeeb H, Pigeot I, Schüz B, Leibniz-WissenschaftsCampus Digital Public Health Bremen. [Digital public health-an overview]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):137-144. [doi: 10.1007/s00103-019-03078-7] [Medline: 31919531]
7. Biviji R, Vest JR, Dixon BE, Cullen T, Harle CA. Factors related to user ratings and user downloads of mobile apps for maternal and infant health: cross-sectional study. JMIR Mhealth Uhealth 2020 Jan 24;8(1):e15663 [FREE Full text] [doi: 10.2196/15663] [Medline: 32012107]
8. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
9. Whitelaw S, Pellegrini D, Mamas M, Cowie M, Van Spall HGC. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021 Mar;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
10. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
11. Rodriguez-Villa E, Torous J. Regulating digital health technologies with transparency: the case for dynamic and multi-stakeholder evaluation. BMC Med 2019 Dec 03;17(1):226 [FREE Full text] [doi: 10.1186/s12916-019-1447-x] [Medline: 31801532]
12. Lorencatto F, Charani E, Sevdalis N, Tarrant C, Davey P. Driving sustainable change in antimicrobial prescribing practice: how can social and behavioural sciences help? J Antimicrob Chemother 2018 Oct 01;73(10):2613-2624. [doi: 10.1093/jac/dky222] [Medline: 30020464]

XSL•FO
RenderX

13. Lampe K, Mäkelä M, Garrido MV, Anttila H, Autti-Rämö I, Hicks NJ, European Network for Health Technology Assessment (EUnetHTA). The HTA core model: a novel method for producing and reporting health technology assessments. Int J Technol Assess Health Care 2009 Dec;25 Suppl 2:9-20. [doi: 10.1017/S0266462309990638] [Medline: 20030886]
14. Joore M, Grimm S, Boonen A, de Wit M, Guillemin F, Fautrel B. Health technology assessment: a framework. RMD Open 2020 Nov 03;6(3):e001289 [FREE Full text] [doi: 10.1136/rmdopen-2020-001289] [Medline: 33148786]
15. Unsworth H, Dillon B, Collinson L, Powell H, Salmon M, Oladapo T, et al. The NICE Evidence Standards Framework for digital health and care technologies - developing and maintaining an innovative evidence framework with global impact. Digit Health 2021;7:20552076211018617 [FREE Full text] [doi: 10.1177/20552076211018617] [Medline: 34249371]
16. What is health technology assessment (HTA)? International Network of Agencies for Health Technology Assessment. URL: https://www.inahta.org/ [accessed 2024-07-22]
17. Kirwin E, Round J, Bond K, McCabe C. A conceptual framework for life-cycle health technology assessment. Value Health 2022 Jul;25(7):1116-1123 [FREE Full text] [doi: 10.1016/j.jval.2021.11.1373] [Medline: 35779939]
18. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
19. Bradway M, Carrion C, Vallespin B, Saadatfard O, Puigdomènech E, Espallargues M, et al. mHealth assessment: conceptualization of a global framework. JMIR Mhealth Uhealth 2017 May 02;5(5):e60 [FREE Full text] [doi: 10.2196/mhealth.7291] [Medline: 28465282]
20. Khoja S, Durrani H, Scott RE, Sajwani A, Piryani U. Conceptual framework for development of comprehensive e-health evaluation tool. Telemed J E Health 2013 Jan;19(1):48-53. [doi: 10.1089/tmj.2012.0073] [Medline: 22957502]
21. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digit Health 2019 Jun;1(2):e52-e54 [FREE Full text] [doi: 10.1016/S2589-7500(19)30013-5] [Medline: 33323229]
22. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. it - Inf Technol 2019;61(5):253-263. [doi: 10.1515/itit-2019-0019]
23. Stratil JM, Baltussen R, Scheel I, Nacken A, Rehfuess EA. Development of the WHO-INTEGRATE evidence-to-decision framework: an overview of systematic reviews of decision criteria for health decision-making. Cost Eff Resour Alloc 2020 Feb 11;18(1):8 [FREE Full text] [doi: 10.1186/s12962-020-0203-6] [Medline: 32071560]
24. Vanderkruik R, McPherson ME. A contextual factors framework to inform implementation and evaluation of public health initiatives. Am J Eval 2016 Oct 03;38(3):348-359. [doi: 10.1177/1098214016670029]
25. Muellmann S, Pan C, Jahnel T, Forberger S, Jürgens D, Barnils N. Frameworks for the development and evaluation of digital technologies for public health: a scoping review protocol. OSF Registries. URL: https://osf.io/n8jge [accessed 2024-07-22]
26. Frameworks for the development and evaluation of digital technologies for public health. Open Science Framework. URL: https://osf.io/ku38m/ [accessed 2024-07-22]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Beauchamp T, Childress J. Principles of Biomedical Ethics. Oxford, UK: Oxford University Press; 2012.
29. Pan CN, Jürgens D, Muellmann S, Janetzki S, Kolschen J, Freye M. Developing and Assessing Digital Public Health Interventions: A Comprehensive Framework. 1st Version. Bremen: Leibniz ScienceCampus Digital Public Health; 2022.
30. Ellmann S, Maryschok M, Schöffski O, Emmert M. The German COVID-19 digital contact tracing app: a socioeconomic evaluation. Int J Environ Res Public Health 2022 Nov 02;19(21):14318 [FREE Full text] [doi: 10.3390/ijerph192114318] [Medline: 36361198]
31. Pozo-Martin F, Beltran Sanchez MA, Müller SA, Diaconu V, Weil K, El Bcheraoui C. Comparative effectiveness of contact tracing interventions in the context of the COVID-19 pandemic: a systematic review. Eur J Epidemiol 2023 Mar 16;38(3):243-266 [FREE Full text] [doi: 10.1007/s10654-023-00963-z] [Medline: 36795349]
32. Pan CN, Jürgens D, Muellmann S, Janetzki S, Kolschen J, Freye M. Developing and assessing digital public health interventions: a digital public health framework (DigiPHrame). OSF Registries. URL: https://osf.io/ub3w4 [accessed 2024-07-22]

Abbreviations

CWA: Corona-Warn-App
DiPH: digital public health
DigiPHrame: digital public health framework
ESF: Evidence Standards Framework
HTA: health technology assessment
LSC DiPH: Leibniz ScienceCampus Digital Public Health
OSF: Open Science Framework

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.11.23; peer-reviewed by L Maaß, G Humphreys, BC Silenou, V Zander; comments to author 29.01.24; revised version received 27.03.24; accepted 27.06.24; published 12.09.24.

Please cite as:
Jahnel T, Pan CC, Pedros Barnils N, Muellmann S, Freye M, Dassow HH, Lange O, Reinschluessel AV, Rogowski W, Gerhardus A Developing and Evaluating Digital Public Health Interventions Using the Digital Public Health Framework DigiPHrame: A Framework Development Study
J Med Internet Res 2024;26:e54269
URL: https://www.jmir.org/2024/1/e54269
doi:10.2196/54269
PMID:

©Tina Jahnel, Chen-Chia Pan, Núria Pedros Barnils, Saskia Muellmann, Merle Freye, Hans-Henrik Dassow, Oliver Lange, Anke V Reinschluessel, Wolf Rogowski, Ansgar Gerhardus. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Toward Guidelines for Designing Holistic Integrated Information Visualizations for Time-Critical Contexts: Systematic Review

Ahmed Mohammed Patel[1], PhD; Weston Baxter[1], PhD; Talya Porat[1], PhD

Dyson School of Design Engineering, Imperial College London, London, United Kingdom

Corresponding Author:
Ahmed Mohammed Patel, PhD
Dyson School of Design Engineering
Imperial College London
Imperial College Rd
South Kensington
London, SW7 2DB
United Kingdom
Phone: 44 07990035581
Email: ap19@ic.ac.uk

Abstract

Background: With the extensive volume of information from various and diverse data sources, it is essential to present information in a way that allows for quick understanding and interpretation. This is particularly crucial in health care, where timely insights into a patient's condition can be lifesaving. Holistic visualizations that integrate multiple data variables into a single visual representation can enhance rapid situational awareness and support informed decision-making. However, despite the existence of numerous guidelines for different types of visualizations, this study reveals that there are currently no specific guidelines or principles for designing holistic integrated information visualizations that enable quick processing and comprehensive understanding of multidimensional data in time-critical contexts. Addressing this gap is essential for enhancing decision-making in time-critical scenarios across various domains, particularly in health care.

Objective: This study aims to establish a theoretical foundation supporting the argument that holistic integrated visualizations are a distinct type of visualization for time-critical contexts and identify applicable design principles and guidelines that can be used to design for such cases.

Methods: We systematically searched the literature for peer-reviewed research on visualization strategies, guidelines, and taxonomies. The literature selection followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The search was conducted across 6 databases: ACM Digital Library, Google Scholar, IEEE Xplore, PubMed, Scopus, and Web of Science. The search was conducted up to August 2024 using the terms ("visualisations" OR "visualizations") AND ("guidelines" OR "taxonomy" OR "taxonomies"), with studies restricted to the English language.

Results: Of 936 papers, 46 (4.9%) were included in the final review. In total, 48% (22/46) related to providing a holistic understanding and overview of multidimensional data; 28% (13/46) focused on integrated presentation, that is, integrating or combining multidimensional data into a single visual representation; and 35% (16/46) pertained to time and designing for rapid information processing. In total, 65% (30/46) of the papers presented general information visualization or visual communication guidelines and principles. No specific guidelines or principles were found that addressed all the characteristics of holistic, integrated visualizations in time-critical contexts. A summary of the key guidelines and principles from the 46 papers was extracted, collated, and categorized into 60 guidelines that could aid in designing holistic integrated visualizations. These were grouped according to different characteristics identified in the systematic review (eg, gestalt principles, reduction, organization, abstraction, and task complexity) and further condensed into 5 main proposed guidelines.

Conclusions: Holistic integrated information visualizations in time-critical domains are a unique use case requiring a unique set of design guidelines. Our proposed 5 main guidelines, derived from existing design theories and guidelines, can serve as a starting point to enable both holistic and rapid processing of information, facilitating better-informed decisions in time-critical contexts.

(J Med Internet Res 2024;26:e58088) doi:10.2196/58088

XSL•FO
RenderX

KEYWORDS

visualization; design; holistic; integrated; time-critical; guidelines; pre-attentive processing; gestalt theory; situation awareness; decision-making; mobile phone

Introduction

Background

The need for quick and holistic processing and understanding of information or sensemaking is critical across multiple domains, particularly in health care. We define *quick* as ±30 seconds and *holistic* as representing *the bigger picture*, where all core components are referenced and individual components can only be understood in relation to the whole. A holistic visualization allows users to grasp the entire picture without needing to consult multiple sources. Examples include providing primary care practitioners with a holistic overview of the patient before starting a clinical consultation, with the requirement from clinicians to assimilate and understand the patient's key information within 30 seconds [1], or a holistic display for antibiotic use in intensive care that can be perceived in <40 seconds [2].

Enhanced processing speed, where an individual can understand and react to the information received in a holistic manner, can enable rapid or real-time response and contribute to heightened situational awareness and informed decision-making [3]. With the vast amount of information available from multiple and heterogeneous sources of data, considered organization is required more than ever to guide both the understanding and interpretation of the information [4-6].

While numerous guidelines exist for a range of visualizations, such as designing visualizations more broadly [5,7-10], step-by-step processes for visualizing data [9,11-13], designing dashboards [14-16], and designing glanceable visualizations [17,18], guidelines toward enhancing situational awareness [19] and visualization architecture for time-critical contexts, as well as merging support for decision-making with situational awareness [20], this study evidences that there are currently no guidelines for the design of holistic integrated visualizations enabling quick information processing and comprehension in time-critical contexts.

In this paper, we argue that, for time-critical contexts, holistic visualizations that integrate multidata variables into a single visual representation are a distinct type of information visualization with different needs and constraints. We define *integrated* as a single connected visual representation of multidata variables (ie, different data dimensions). Data dimensions pertain to the categories that define the structure of the information (ie, medications or allergies in a health care visualization or altitude or speed in a military aircraft display) and represent the characteristics of the described data. Data points are the individual variables and data instances within the data dimensions (ie, allergy to penicillin). An integrated presentation brings together diverse data in a single visual representation, allowing users to view and analyze the data in a way that is interconnected, showing the links and relationships within the data (ie, between medication and allergies). Thus, holistic integrated visualizations combine both holistic and integrated approaches. While *holistic* references the core information, allowing users to grasp the entire picture (ie, that the patient is an older adult female, who is obese, smokes, and has uncontrolled diabetes and hypertension), *integrated* provides a single connected visual representation of these different data dimensions, allowing the user to understand the relationships between the patient's characteristics (ie, between allergies and medications or health conditions and medications). As characteristics refer to the distinctive and discernible attributes, features, and qualities of something, the characteristics of holistic integrated information visualizations (HI-viz) are both holistic (offering the *bigger picture* and a comprehensive overview of key information) and integrated (a unified, connected visual representation), with higher information complexity (diverse data dimensions), and timely (enabling rapid processing of information in time-critical contexts). Toward defining design guidelines for this unique type of visualization, which we name HI-viz, we systematically searched the literature for established visualization design theories and guidelines from academic and industry sources, focusing on general design guidelines for visualizations and more specific guidelines for designing holistic, integrated, and timely visualizations.

In the following sections, we will position HI-viz displays against the current existing visualization types.

Visualization Types

Glanceable Visualizations

Glanceable visualizations are typically discussed in the context of smartwatches or smartphones [18,21-23] and characterized as "being able to quickly extract the essential information from a display by a quick glance" [18,24]. Example scenarios include a runner who has a few hundred milliseconds to glance at the smartwatch to check relevant indicators, such as heart rate, elevation, and distance covered, before refocusing on the trail ahead [21,23]. The quick information needs support the shifting of focus from the primary task, typically in use cases that are locomotive, and through displays that are portable or small in the case of a smartwatch. The information provided is immediate, featuring a simple appearance and encoding very few data dimensions [23], be it in the form of a graph [18,23] or abstracted representations such as an arrow indicating direction or the use of other familiar visual metaphors [21,22]. Previous human-computer interaction studies investigating smartwatch use and interactions [25,26] have categorized timescales of smartwatch glanceable interactions as *peeks* (≤5 seconds) and *interactions* (>5 seconds). Up to 5 seconds can be taken as the limits of a glanceable visualization without any physical interactions, where the user receives, understands, and acts on the information. However, the time a user actively attends to the content via a "glance" or "peek" is suggested to be shorter [23].

XSL•FO
RenderX

Individual Visualizations

Individual visualization types such as line, bar, radial, and donut graphs are established visual representations in glanceable visualizations [18,23,27]. The effectiveness of these varies. Comparative studies highlight that information presented in radial charts takes longer to gather than that presented in bar or donut charts [23,28]. A study on information representation types in smartwatches [29] highlighted "icon+text" as being the most common representation type and supports research suggesting that textual data representations are predominant in smartwatches [30]. The most common graphical representation types or information "visualization techniques" are categorized as trees, scatter plots, charts, tables, diagrams, and graphs [31,32]. Previous studies have evaluated the effectiveness of different visualization types [33-38] as well as effectiveness across a number of selected tasks [33,38-41]. While the various visualization types perform differently depending on the task at hand, the literature suggests a positive correlation between perceived accuracy and user preference in 2D visualization types (tables, line charts, bar charts, scatterplots, and pie charts) for facilitating accurate and quick completion of tasks (finding anomalies, finding clusters, correlation, derived value, distribution, finding extremum, order, retrieving value, filtering, and determining range). Average response times reported in the study ranged between 10 and 40 seconds [33].

Increased Complexity and Visualization Types

A traditional individual visualization type typically encodes up to 3 data dimensions, this being the number of unique attributes or discrete information points in the dataset [42]. When more data dimensions are included, the *complexity* is increased. Commonly, 3 to 4 data dimensions are encoded in a visualization [43]. For example, a line graph representing an individual's step count over the course of a week illustrates the days and number of steps—2 data dimensions. Adding a second individual to the graph provides a third data dimension. Adding data points within these dimensions (ie, additional weeks or more individuals) will not increase the complexity but may make it more crowded or complicated to visualize and understand the graph [43]. As the number of data dimensions that need to be visually encoded increases, the *complexity* of a visualization increases [43]. Although applications of visualizations encoding ≥4 data dimensions exist, it becomes more challenging to design and learn from them and illustrate their relationships [43,44]. Perceptually, an increase in dimensionality can also be overwhelming [42]. Beyond 3 data dimensions, visualization representations include *combined visualization types* (such as a line graph integrated within a bar graph); individual visualization types collectively tiled to present the information (*dashboards*); *diagrammatic representations* (ie, concept diagrams and scientific overviews) visualizing relationships between parameters, connecting and linking various themes to represent a multidimensional concept or topic [45,46]; or aesthetic presentations (*infographics* and *visual art*).

As the data dimensionality increases and visual representations change, the reader requires more time to understand the visualization [47,48]. There is also a shift from *explanatory visualizations*, where the data and narrative are provided, structured, and communicated, toward *exploratory visualizations*, where the viewer seeks the story that the data have to tell them through open interactions [43,49]. Iliinsky and Steele [43] further propose a hybrid explanation-exploration category where the provided data are curated and communicated, allowing for some exploration and interaction; thus, the reader can "...choose and constrain certain parameters, thereby discovering for herself whatever insights the dataset may have to offer." Adopting a hybrid approach and finding a balance between explanation and exploration contributes to some of the most successful visualizations of today [49]. *Dashboards* are an example of the hybrid exploration-explanation approach, where multiple individual visualizations provide a top-level overview, encouraging further exploration and analysis [50]. The literature characterizes dashboards as being "glanceable" [50,51] because of the tiled layout of singular simple charts or numbers, which facilitates the information to be monitored at a glance [52]. However, studies evaluating efficiency or task completion rate via dashboards highlight the engagement and time for task completion to be several minutes [53-55].

Infographics are common in data journalism, with the goal of making information presentable and digestible to readers [56,57]. Defined as "a larger graphic design that combines data visualisations, illustrations, text, and images together into a format that tells a complete story" [58], infographics are highly engaging [56] and exploratory [43,51], intended for "deeper exposition" [50]. Infographics provide a high-level overview or essence of complex, abstract, and dense information [59,60], typically static when published [61] and rendered through a manual illustration process, consequently being aesthetically rich but making it challenging to change and manipulate data points [43]. Beyond infographics, the *visual art* use case is highly exploratory, and data are translated into a visual form typically with a "unidirectional encoding of information" [43]. The viewer is presented with a highly visual but abstract representation and seeks to interpret and arrive at their own meaning of the data [43,62].

Existing visualizations support both explanatory and exploratory tasks [43]. *Explanatory visualizations* typically encode 1 to 3 data dimensions [42], communicating information and data, including glanceable visualizations and scientific individual types (eg, line, bar, radial, and donut graphs). Visualizations combining individual types (line graph included within a bar chart) encode ≥4 data dimensions [43]. *Exploratory visualizations* are higher in data dimensionality, typically where the viewer interacts with the visualization to seek the story communicated by the data (eg, infographics and visual art) [43,44]. Some visualization types, including dashboards, diagrammatic representations, and some infographics and holistic integrated visualizations, can be described as being both explanatory and exploratory through a combination of information communication and structured interaction [43,49]. HI-viz is positioned at the forefront of the explanatory-exploratory Venn diagram due to the heightened complexity and required rapid processing time akin to that for explanatory visualizations.

The proposed HI-viz is a unique visualization type in time-critical contexts (Figure 1). Characteristics include

XSL•FO
RenderX

heightened complexity with the presentation of multiple discrete data points, similar to a dashboard, but in a single integrated visual representation. It requires rapid processing, similar to glanceable visualizations, but with this being the primary task and requiring additional engagement. It can link and communicate relationships between information points, similar to diagrammatic representations, but in a curated, coherent, and concise manner. The requirements of this visualization type transition beyond the pre-attentive—a rapid subconscious processing of information in the visual environment where the viewer is drawn to elements of the design or data through pre-attentive attributes (color, form, movement, and spatial positioning) [10]—and "glanceable" timescales into a nuanced attentive area we define as "rapid attention processing." This stipulated segment of the timescale typically features glanceable visualizations and individual chart types that are usually lower in information density and complexity. Thus, the holistic integrated visualization has the potential to convey more information in a shorter time frame through a single visual representation (Figure 2).

Figure 1. Characteristics of different visualization types—categorization by information complexity (number of data dimensions represented), time (for interpretation), and being holistic (overview of core information) and integrated (representation of multi-data variables in a single connected visualization.

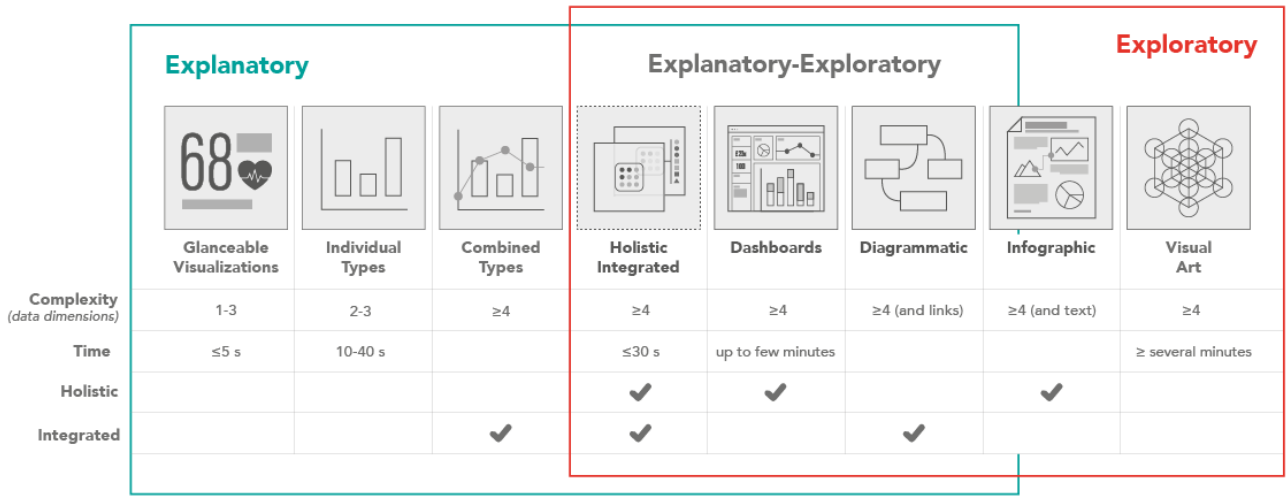

The synthesis of the theoretical underpinning (Figure 2) positions the information visualization types against *information* (from simple visualizations with few data dimensions to complex and rich visualizations with increased data dimensions) [18,23,42-44,47,48,50-52] and *time* (engagement duration) [18,21,23-26,33,47,48,50-55,62], across an explanatory-exploratory visualization task Venn diagram [43,49]. While reading the chart (Figure 2), one should keep in mind that it communicates the relative time and information complexity of each visualization type theoretically and provides a general comparison as synthesized through analysis of the literature. The authors have positioned visualization types considering information complexity (less or more information—data dimensions) and time (to understand the information). In practice, how the visualization types perform may vary depending on their physical implementation and information complexity.

XSL•FO
RenderX

Figure 2. Positioning of holistic integrated information visualizations (HI-viz)—information visualization types plotted against information complexity (from simple visualizations with few data points to complex/rich visualizations with increased data points) and time (engagement duration) across an explanatory-exploratory visualization task Venn diagram.

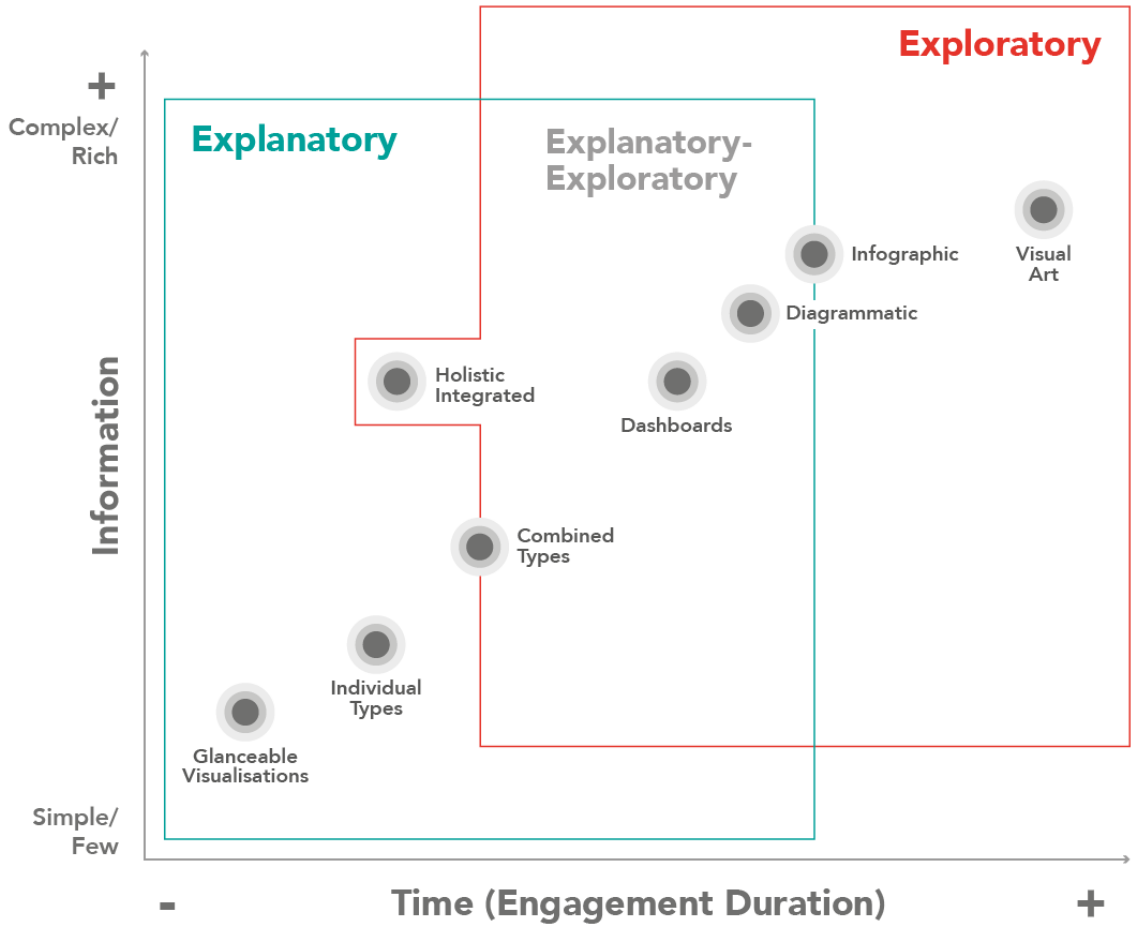

Examples of HI-viz in other domains such as sports, finance, and military applications have proven effective in giving a quick at-a-glance presentation of multivariate data providing both summary and context. Profiles of football players in fantasy leagues and computer games provide a holistic overview of key information regarding the professional athlete and their skills and abilities (eg, FIFA 22 video game player rankings [63]). The holistic integrated profile of Messi (Figure 3B [63]) quickly informs the viewer that the player is an excellent (93 overall aggregated score) male football player of Argentinean nationality; plays for Paris Saint-Germain Football Club (Paris) at club level; and excels at physical, mental, and technical attributes such as shooting, passing, and dribbling. This information captures a holistic overview of the player and their attributes. This FIFA 22 player card presents 12 core data dimensions pertaining to the player (name, photo, overall rating, position, nationality, club, and 6 individual performance attribute statistics) in a single integrated "player profile" visualization. The categories and numerical values are easily interpretable, allowing the viewer to understand player strengths and weaknesses as well as facilitating easy comparison between players holistically and across categories.

Similarly, Finviz [64], an interactive tree map of the stock market, provides users with an at-a-glance holistic summary of stocks categorized by industry and sectors against key performance metrics (price fluctuation and market cap). This single integrated view allows the viewer to see the overall market, how stocks compare within a sector, and relative sector performance. Color coding (red for losses and green for gains) and the size of the stock tiles immediately illustrate stock performance, performance across the market, and whether a sector is up or down. This visualization is high in information complexity, with hundreds of stocks and each tile communicating the stock ticker symbol, price change (percentage up or down), and market dominance (the size of the tile) within its respective sector and industry groupings.

XSL•FO
RenderX

Figure 3. (A) Finviz [64]; (B) FIFA 22 player rankings [63]; (C) Older Adults' Wellness by Le et al [4].

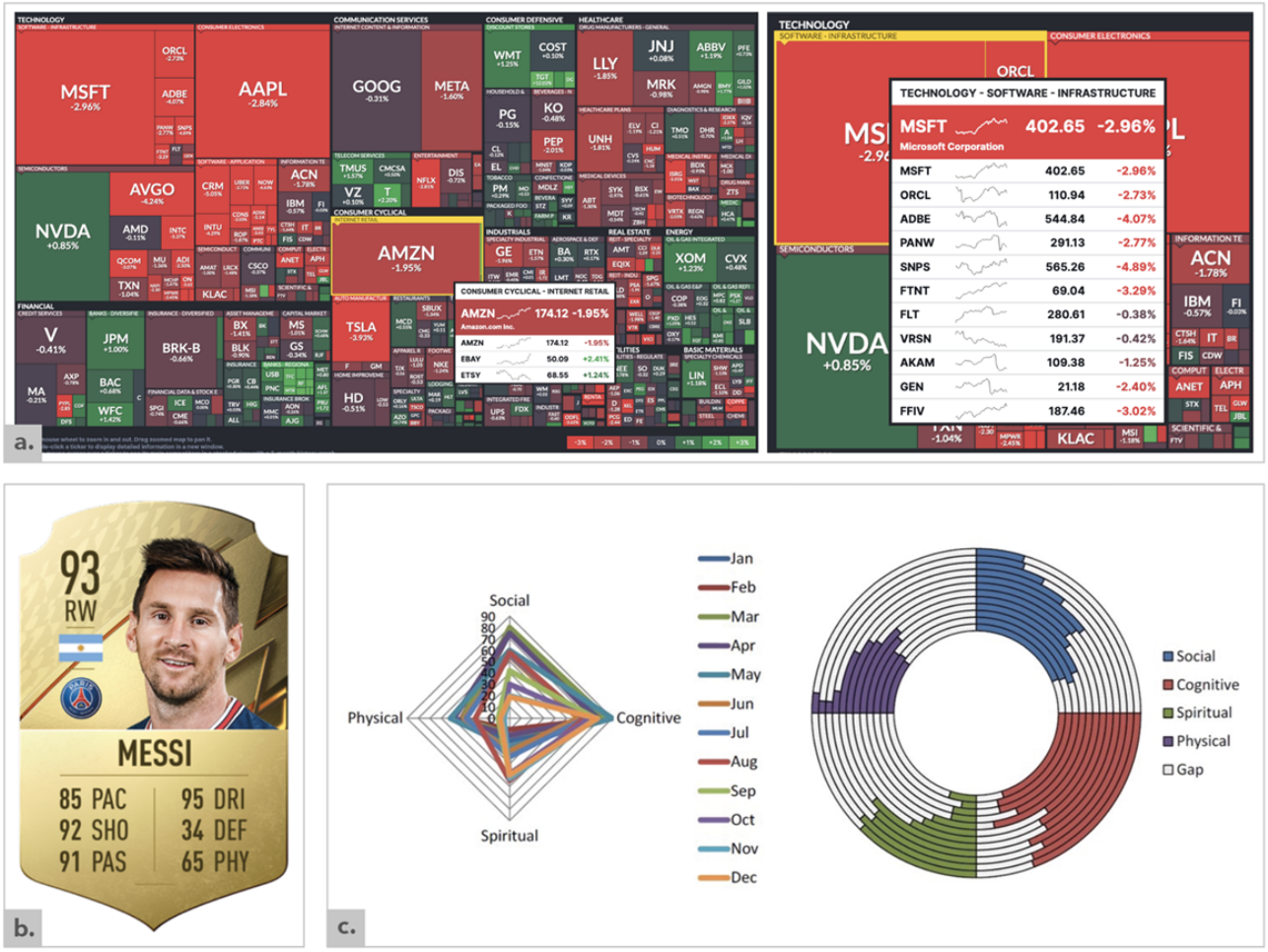

Military aircraft, too, amalgamate high information complexity and display a user interface for rapid information processing. Flight panels integrate disparate data dimensions such as altitude (above ground level), flight path, aircraft pitch and roll, current airspeed, vertical speed, aircraft heading, engine status (speed, fuel, temperature, and torque), outside air temperature, navigation data, alerts, and flight warnings. This provides an at-a-glance holistic overview for the pilot in a single display for rapid comprehension and decision-making. The integrated view enables the layering and combining of different types of information, allowing the pilot to track and process core information simultaneously.

Within health care, Le et al [4,65] present "an approach to capture older adults' wellness" through 3 alternate visualization techniques (stacked bar charts, polygons, and donut displays) in addition to an integrated measure of wellness—cognitive, physiological, social, and spiritual. This presents alternate approaches, nonexistent in electronic systems, combining the 4 factors into 1 graphical representation of overall wellness (Figure 3C [4]). The tool captures numerous data points pertaining to older adults' wellness over a 12-month period, with the holistic overview enabling quick interpretation of wellness indicators (physiological, social, spiritual, and cognitive) and correlations at a glance in a single integrated visual representation. The aim to provide additional granularity for health information, while maintaining the representation of overall wellness was an interesting paradigm, offering the viewer a quick overview in addition to further detail and layers of information. Similarly, a holistic integrated visualization in primary care can inform us, for example, that we are dealing with a complex older adult female patient who is obese, smokes, and has uncontrolled diabetes and hypertension. The information could be derived from different sources of data (ie, electronic health records, wellness sensors, and secondary care) and integrated into a single visual representation, and relationships between data dimensions could be highlighted through links (ie, medications and health problems or medications and allergies).

Motivation: Designing HI-viz

Guidelines and principles are useful for the systematic and effective creation of novel outputs following established tasks, techniques, and taxonomies [13,66]. Numerous studies have reviewed or presented guidelines for visualizations more broadly [5,7-10]. Their findings include overall design processes, principles, or frameworks for information visualization [9,11-13,66-69]; visualization taxonomies [70-78]; the grammar of graphics [10,79-81]; designing for storytelling [58,82-89]; designing dashboards [14-16]; multiple views [90,91]; designing infographics [58,84,86,92-95]; and glanceable interfaces [17,18]. Kohlhammer and Zeltzer [20] proposed a visualization architecture for time-critical contexts, aiming to merge support for decision-making with situational awareness. Their paper touches on the design of a user interface module and a

XSL•FO
RenderX

decision-centered visualization architecture, which shows the connections between modules (data access, decision focus, user interface, and submodules). However, no approaches, methods, principles, or guidelines were provided toward designing holistic integrated visualizations for time-critical contexts.

To design HI-viz, as the needs and constraints differ from those of other visualization use cases (Figures 1 and 2), a contextually appropriate set of guidelines is required combining established guidelines and principles for rapid processing, integrated presentation, and a holistic understanding.

Methods

We systematically searched the literature for peer-reviewed research on visualization strategies, guidelines, and taxonomies. The literature selection followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, an established method for systematic review of scientific publications (Multimedia Appendix 1) [96,97].

The following 6 databases were searched for published articles in peer-reviewed journals: ACM Digital Library, Google Scholar, IEEE Xplore, PubMed, Scopus, and Web of Science. The search was up to August 2024 using the terms (“visualisations” OR “visualizations”) AND (“guidelines” OR “taxonomy” OR “taxonomies”), with studies restricted to the English language. The search terms were intentionally broader to ensure that all relevant publications pertaining to visualizations, guidelines, and taxonomies were captured. The Boolean logic ensured that the search results included publications pertaining to visualizations (with both the UK and US spellings) as well as a minimum of 1 related concept (guidelines, taxonomy, or taxonomies). Search terms were screened across publication titles and abstracts.

Papers were included if they were published in English and related to strategies and methods (guidelines and taxonomies) for designing visualizations and if the guidelines related to one or more of the following: *holistic understanding for complexity* (designing visualizations for multidimensional data), *time* (ie, time-critical applications and designing for quick information processing), and *integrating* multidimensional data within a single visual representation. References from the included articles were examined to identify additional peer-reviewed and industry literature.

Papers were excluded if (1) they were duplicates; (2) they were not in English; (3) they focused on design guidelines for specific applications, such as augmented reality, virtual reality, or mixed reality, where the guidelines or taxonomies did not address integrated presentation; (4) they did not involve integrating or combining multidimensional data into a single visual representation; (5) they did not focus on guidelines for visualizing complex multidimensional data for holistic understanding; or (6) they did not address designing for rapid information processing.

Following title and abstract screening, the full texts of the remaining papers were checked for eligibility against the inclusion and exclusion criteria (Figure 4). The first author undertook the study search and screening for exclusion and inclusion of studies, which was confirmed to be accurate and complete by another reviewer (TP). Throughout the screening process, coauthors were consulted as needed in determining the eligibility of the studies.

Figure 4. Study Inclusion and Exclusion Criteria.

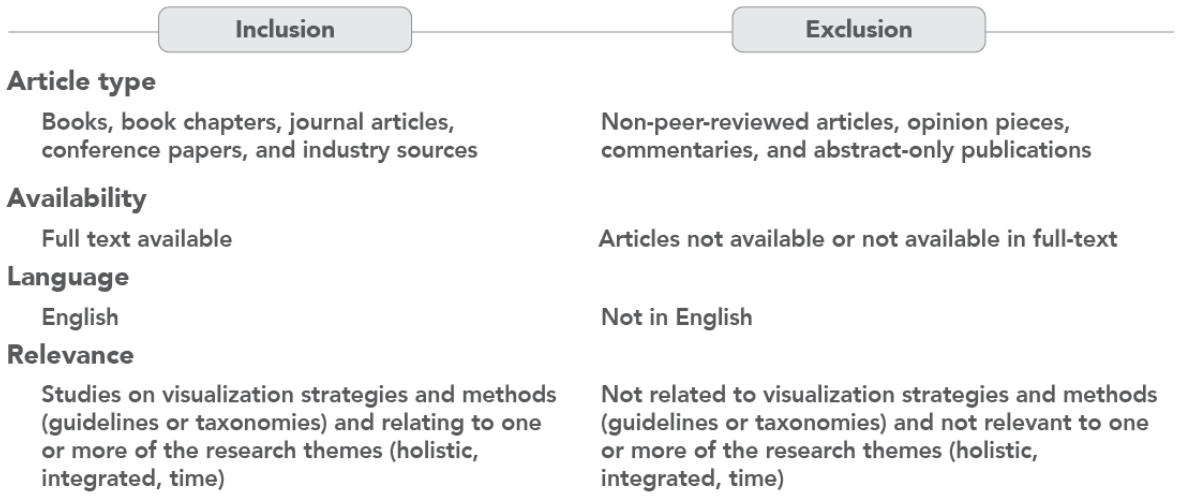

Through an inductive thematic analysis approach, themes emerging from the data were identified. Initial codes were elected by the first author. NVivo (version 12; QSR International) was used to code and analyze the data. Codes were discussed, grouped into broader themes, and iteratively reviewed among the 3 authors to reach a consensus.

Results

Overview

The systematic literature search resulted in a pool of 936 publications pertaining to visualization guidelines and taxonomies. A total of 90.9% (851/936) of the papers were excluded based on duplicity, non-English language, and title

XSL•FO
RenderX

and abstract screening. In total, 58% (49/85) of the remaining articles were excluded after reading the full text. A total of 36 papers met the inclusion criteria [9,17,67,69,71,90,98-127]. A further 10 publications were included upon examining references from included articles [10,19,128-133], including 2 from industry sources [5,134]. A total of 46 papers were included in the final review (Figure 5).

Figure 5. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram—flow of information through the different phases of the systematic review.

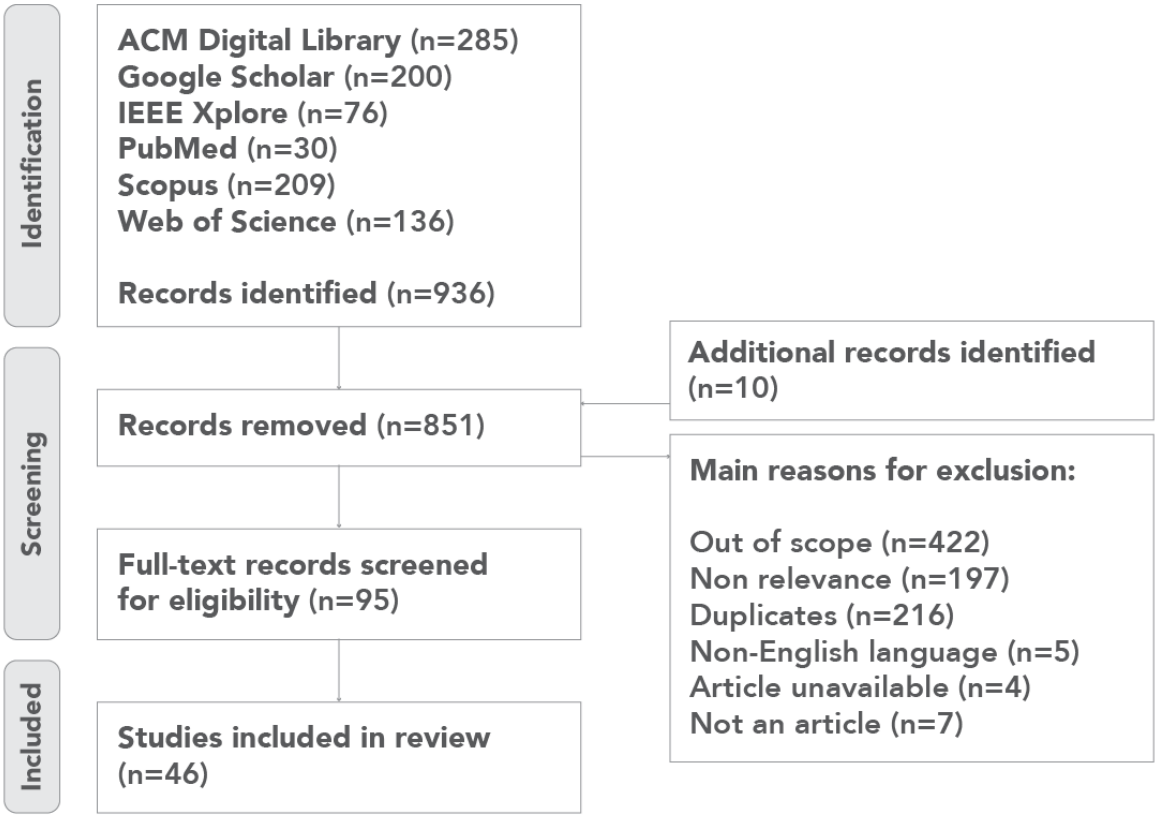

In total, 48% (22/46) of the papers were related to visualizing complexity to provide an overview or *holistic* understanding of multidimensional data (ie, prioritization of information, information reduction, organization of information, abstraction representations to communicate information, visual aggregates, and interacting with visualizations or to call and filter information); 28% (13/46) of the papers focused on *integrated presentation*, integrating or combining multidimensional data into a single visual representation (ie, through aggregation of data, integration of numerical and statistical values, and grouping approaches [ie, gestalt]); and 35% (16/46) of the papers pertained to *time* and designing for rapid information processing (ie, implementing pre-attentive processing attributes and features, gestalt principles, glanceable display guidelines, information abstraction, and interaction; minimizing task complexity; and enhancing situational awareness). A total of 65% (30/46) of the papers presented *general* information visualizations or visual communication guidelines and principles. In this category, we included more universal user-interface and user experience guidelines as well as considerations of information reduction, clarity of communication, information organization, interaction and navigation, relationships among and between information, minimization of task complexity, and alarm support and projection. Some of the papers (31/46, 67%) addressed more than one attribute (Table 1). A detailed mapping of the features of the holistic, integrated, time, and general guideline characteristics of the studies in this review is mapped in Multimedia Appendix 2 [5,10,17,19,67,69,71,90,98-134].

XSL•FO
RenderX

Table 1. Characteristics of the studies: holistic, integrated, time, and general design.

Study, year	Holistic (n=22)	Integrated (n=13)	Time (n=16)	General design (n=30)
Adnan et al [98], 2008	✓			✓
Ajani et al [99], 2022				✓
Andrews et al [100], 2011	✓	✓		✓
Bateman et al [101], 2010			✓	✓
Brath [102], 1997	✓			✓
Chang and Nesbitt [103], 2006		✓		✓
Cowan [104], 2015	✓			✓
De Carlo et al [105], 2022	✓			✓
De Weck et al [106], 2011				✓
Dix [107], 2013	✓			✓
Dybala et al [108], 2020		✓	✓	
Ellis and Dix [109], 2007				✓
Elmqvist and Fekete [9], 2010	✓	✓		
Endsley [19], 2012	✓			✓
Fekete and Plaisant [110], 2002	✓		✓	
Forsell and Johansson [69], 2010			✓	✓
Healey et al [111], 1995			✓	
Hick [128], 1952			✓	✓
Idrissov et al [112], 2020		✓		✓
Kelleher and Wagener [113], 2011		✓		✓
Kornhauser et al [114], 2009		✓	✓	✓
Krekhov et al [115], 2019			✓	
Ku et al [116], 2012	✓			
Lima [5], 2019			✓	✓
Matthews et al [17], 2006	✓		✓	✓
Midway [67], 2020	✓			✓
Miller [129], 1956	✓			✓
Moran [134], 2016	✓			✓
Moreira et al [117], 2020		✓	✓	
Murray et al [118], 2017				✓
Perer and Shneiderman [119], 2009	✓	✓		
Rechtin [120], 1991	✓			✓
Rodrigues et al [121], 2006	✓		✓	
Rosli et al [122], 2015		✓		
Seong and Nuamah [123], 2020		✓	✓	
Shaalan and Jusoh [124], 2020				✓
Shneiderman [71], 1996	✓			✓
Spence [130], 2001	✓		✓	✓
Triesman [131], 1985			✓	
Tufte [132], 1983				✓
van Ham and Perer [125], 2009	✓			

XSL•FO
RenderX

Study, year	Holistic (n=22)	Integrated (n=13)	Time (n=16)	General design (n=30)
Wagemans et al [126], 2012		✓		
Wang Baldonado et al [90], 2000	✓			
Ware [10], 2020		✓	✓	
Wenzel et al [127], 2003				✓
Yablonski [133], 2020	✓			✓

This systematic review highlights that, presently, no specific guidelines or principles were found to address all the characteristics of HI-viz, namely, holistic, integrated visualizations for time-critical contexts. Consideration of the established principles, guidelines, and taxonomies from the fields of visualization, design, and cognitive science and a combination thereof may aid in the design of holistic integrated visualizations for time-critical contexts. A summary of the key guidelines and principles identified in this review was extracted, collated, and categorized in Table 2. The categories in Table 2 were created through the following steps:

1. Repeated features were combined across the 4 characteristics identified in this systematic review (ie, gestalt principles, reduction, organization, abstraction, aggregates, and task complexity).
2. Interaction was deconstructed to provide prescriptive categories (ie, experience, overview, details on demand, and rearrangement).
3. Some guidelines were broad and closely aligned with existing categories. For example, the 4 guidelines pertaining to “glanceable” displays better aligned with organization, abstraction, visual clarity, and experience. Where the user experience or system principles (Hick’s law, Miller’s law, and Tesler’s law) aligned with more than one feature identified in this systematic review, they were better mapped across organization, reduction, and task complexity.
4. Individual unique features identified in this systematic review were maintained (ie, pre-attentive, prioritization, relationships, navigation, alarms, and visual clarity).

XSL•FO
RenderX

Table 2. Summary of key established design and visualization guidelines and principles for holistic integrated information visualizations (HI-viz).

Category, guideline, and principle	References
Pre-attentive	
G01. Use pre-attentive attributes (color, form, spatial positioning, and movement)	Ware [10], Fekete and Plaisant [110], Healey et al [111], Kornhauser et al [114], Krekhov and Kruger [115], Moreira et al [117], Rodrigues et al [121], Seong et al [123], Spence [130], Triesman [131]
Prioritization	
G02. Prioritize the core information and message to be communicated	Midway [67]
G03. Manage featurism through prioritization and flexibility	Endsley [19]
G04. Say no to feature creep—buck the load	Endsley [19]
Reduction	
G05. Remove the extraneous	Dix [107], Forsell and Johansson [69]
G06. Consider dataset reduction—efficiency and ease of use of the dataset	Forsell and Johansson [69]
G07. Use information filtering carefully	Endsley [19]
G08. Explicitly identify missing information	Endsley [19]
G09. Reduce display density but do not sacrifice coherence	Endsley [19]
Organization	
G10. Organize and group information around goals	Endsley [19], Cowan [104], Miller [129], Moran [134]
G11. Ensure spatial organization for user navigation—the distribution of elements	Forsell and Johansson [69]
G12. Ensure that consistency in design choices is maintained in similar or different contexts	Forsell and Johansson [69], Chang and Nesbitt [103]
G13. Maintain consistency in design elements that are similar	Lima [5], Matthews et al [17]
G14. Provide consistency and standardization on controls across different displays or systems	Endsley [19]
G15. Ensure logical consistency across modes and features	Endsley [19]
G16. Consider Miller's law of "seven, plus or minus two" for grouping information	Miller [129], Moran [134], Cowan [104]
G17. Remove clutter and "chart junk"	Ajani et al [99], Bateman et al [101], Ellis and Dix [109], Tufte [132]
G18. "Relationships among elements are what gives systems their added value"	Rechtin [120], Andrews et al [100], Dix [107], Idrissov et al [112], Murray et al [117]
Gestalt principles	
G19. Proximity—elements that are close together are perceptually grouped together	Ware [10], Chang and Nesbitt [103], Idrissov et al [112], Kornhauser et al [114], Moreira et al [117], Rosli and Cabrera [122], Seong et al [123], Wagemans et al [126]
G20. Similarity—similar elements tend to be grouped together	Ware [10], Chang and Nesbitt [103], Dybala et al [108], Idrissov et al [112], Kornhauser et al [114], Moreira et al [117], Rosli and Cabrera [122], Seong et al [123], Wagemans et al [126]
G21. Connectedness—linked elements express a relationship between them	Ware [10], Idrissov et al [112], Seong et al [123], Wagemans et al [126]
G22. Continuity—connections between visual elements are discernible if they are smooth and continuous	Ware [10], Chang and Nesbitt [103], Dybala et al [108], Idrissov et al [112], Moreira et al [117], Seong et al [123], Wagemans et al [126]
G23. Symmetry—symmetrical elements are perceived as members of 1 group	Ware [10], Idrissov et al [112], Seong et al [123], Wagemans et al [126]
G24. Closure—elements with a closed contour are perceived as shapes or object-like and contours with gaps are perceptually closed.	Ware [10], Chang and Nesbitt [103], Dybala et al [108], Idrissov et al [112], Moreira et al [117], Rosli and Cabrera [122], Seong et al [123], Wagemans et al [126]

XSL•FO
RenderX

Category, guideline, and principle	References
G25. Common region—elements within a contour or boundary are perceived as a group.	Ware [10], Idrissov et al [112], Moreira et al [117], Seong et al [123], Wagemans et al [126]
G26. Relative size—smaller elements inside larger elements are perceived as separate objects	Ware [10], Idrissov et al [112]
G27. Common fate—(in motion perception) elements moving together in the same direction tend to be perceived as a unified group	Ware [10], Chang and Nesbitt [103], Dybala et al [108], Idrissov et al [112], Kornhauser et al [114], Moreira et al [117], Seong et al [123], Wagemans et al [126]
Abstraction	
G28. Communicate the fundamental features of what is being designed clearly and accurately	Endsley [19], Forsell and Johansson [69], Cowan [104], De Weck et al [106], Dix [107], Miller [129], Moran [134]
G29. "Use abstraction representations" to distil and emphasize key aspects of the information in its most essential detail via iconic representations to communicate the information	Matthews et al [17], Bateman et al [101]
G30. Consider Tesler's law: for any system, there is a certain amount of complexity that cannot be reduced	Yablonski [133]
Aggregates	
G31. Entity budget—maintain an entity budget by controlling the quantity of data rendered on-screen to avoid visually overloading the viewer	Elmqvist and Fekete [9]
G32. Visual summary—aggregates should convey information about underlying data	Elmqvist and Fekete [9], Andrews et al [100]
G33. Visual simplicity—aggregates should be clean and simple	Elmqvist and Fekete [9], Kelleher and Wagener [113]
G34. Discriminability—aggregates should be distinguishable from data items	Elmqvist and Fekete [9]
G35. Fidelity—beware that abstractions may lie	Elmqvist and Fekete [9]
G36. Interpretability—aggregate items only so much so that the aggregation is still correctly interpretable within the visual mapping	Elmqvist and Fekete [9]
Visual clarity	
G37. Ensure that visuals are simple, well-understood representations that can be grasped when looking for the first time	Dix [107], Midway [67], Kelleher and Wagener [113], Kornhauser et al [114], Tufte [132]
G38. "Make visuals distinct," which helps improve the identification of elements and draws the viewer's eye to new or updated information	Matthews et al [17]
Alarms	
G39. Do not make people rely on alarms; provide projection support	Endsley [19]
G40. Support alarm confirmation activities	Endsley [19]
G41. Make alarms unambiguous	Endsley [19], Andrews et al [100]
G42. Ensure that alarms do not disrupt ongoing activities	Endsley [19]
G43. Support the assessment and diagnosis of multiple alarms	Endsley [19]
Experience	
G44. "Match user expectations," which will reduce learning effort	Matthews et al [17]
G45. Experience with the visualization should be intuitive and easy	Lima [5]
G46. Interactivity should be driven by user needs	Spence [130], Lima [5]
Overview	
G47. Gain an overview of information	Shneiderman [71], Andrews et al [100], De Carlo et al [105], Dix [107], Elmqvist and Fekete [9], Fekete and Plaisant [110], Ku et al [116], Van Ham and Perer [125]
G48. Zoom in on items of interest	Shneiderman [71], Wang Baldonado et al [90], Andrews et al [100], De Carlo et al [105], Dix [107], Fekete and Plaisant [110], Perer and Shneiderman [119]

XSL•FO
RenderX

Category, guideline, and principle	References
G49. Filter out uninteresting items	Shneiderman [71], Wang Baldonado et al [90], Adnan et al [98], Fekete and Plaisant [110], Perer and Shneiderman [119], Van Ham and Perer [125]
Details on demand	
G50. Details on demand; select an item or group and obtain information when needed	Shneiderman [71], Wang Baldonado et al [90], Adnan et al [98], Andrews et al [100], Brath [102], De Carlo et al [105], Fekete and Plaisant [110], Van Ham and Perer [125]
G51. Embrace dynamic, fast, and clever experiences; a hover can reveal information faster than clicking to open and clicking to close	Lima [5], Ware [10]
Relationships	
G52. View relationships between items	Shneiderman [71], Wang Baldonado et al [90], Andrews et al [100], Kelleher and Wagener [113], Ku et al [116], Perer and Shneiderman [119]
G53. Keep a history of actions to support undo, replay, and progressive refinement	Shneiderman [71]
G54. Strong contextual cues aid in navigating through the visualization and going back	Lima [5], Andrews et al [100]
G55. Allow for the extraction of subcollections and of the query parameters	Shneiderman [71]
Task complexity	
G56. Minimize task complexity and ensure minimal actions with respect to the number of actions necessary to accomplish a goal or task	Endsley [19], Forsell and Johansson [69]
G57. Flexibility—refers to the number of possible ways of achieving a goal	Forsell and Johansson [69], Chang and Nesbitt [103]
G58. Prompting—refers to all means that help know all alternatives when several actions are possible	Forsell and Johansson [69]
G59. Consider Hick's law, which states that "the time it takes to make a decision increases with the number and complexity of choices"	Hick [128]
Rearrangement	
G60. View the visualization from multiple angles or perspectives	Spence [130], Brath [102], De Carlo et al [105], Dix [107], Fekete and Plaisant [110], Perer and Shneiderman [119], Rodrigues et al [121]

Proposed HI-viz Guidelines

Overview

The 60 guidelines and principles extracted from this systematic review in Table 2 were grouped into 5 main interlinked guidelines for designing HI-viz. Through an inductive thematic analysis approach, 16 top-level categories emerging from the data were identified (Multimedia Appendix 2 and Table 2). Related guidelines were thematically grouped through the top-level categories and underlying principles that linked them, capturing the essence of each group of guidelines (ie, guidelines pertaining to *prioritization* and *reduction* of information were grouped with the common focus of prioritizing key information and reducing the extraneous). The categorization was iteratively reviewed and corroborated among all authors to reach a consensus.

The detailed descriptive and prescriptive guidance references the 60 theoretical guidelines (G01-G60) below.

Prioritize Core Information to Obtain a Holistic Overview of Task-Related Information

To ensure that a visualization provides a holistic overview for a specific task, prioritize the core information, establish hierarchy, reduce clutter, filter information carefully, and use visual cues to highlight important elements and improve user efficiency. Specifically, to attain a holistic overview, a hierarchy must be established in the dataset and design. The core information, message to be communicated, and features must be prioritized (G02 and G03), and the importance between them must be affirmed to aid the management of information in the visualization as well as reducing cognitive complexity on the user (G03). To convey hierarchy and rapidly call attention to prioritized information elements or areas, pre-attentive attributes such as color, form, spatial positioning, and movement should be used (G01). Extraneous information and features that are not core to the task and infrequently accessed must be identified and reduced (G05). This will prevent "feature creep," which is the gradual addition of nonessential features that can complicate the user experience (G04), and limit distractions from the core information needs and wants, contributing to the efficiency and ease of use (G06). The filtering of information must be used carefully (G07), reducing display density to support information

XSL•FO
RenderX

"at a glance" but not sacrificing coherence (G08 and G09) or augmenting the need to navigate and find information, consequently adversely increasing the time taken for information assimilation (G07). Thus, missing information must also be explicitly identified to highlight what is unavailable or not unimportant (G08) and preventing the user from having to search and navigate for data that are not provided.

Organize the Constituent Visualization Elements to Provide an Integrated Representation

To ensure an integrated visualization, organize information around the user's core goals and structure information effectively using grouping, layering, and visual principles into a single connected and combined representation. Specifically, to combine and integrate the constituent elements into a single integrated representation, focus and organize information around the users' core goals and provide the required information directly (G10). Ensure an efficient spatial distribution of information and features to aid user navigation (G11). Organize information by importance and commence with a holistic overview (macro) that captures and communicates the core message, subsequently implementing progressive disclosure to provide detail (micro) and context on demand.

Integrate the constituent elements into the data and design through "chunking" or grouping information while considering Miller's law of "seven, plus or minus two" as a limit for grouped information (G16). Use gestalt principles of organization (proximity, similarity, connectedness, continuity, closure and common region, and symmetry) to guide the arrangement and structuring of information into a connected, combined, and integrated whole (G19-G27). In layering information, use "visual aggregation" (sum or count, average, mode, extents, medians, percentiles, and distribution) to communicate information about underlying contents while not misrepresenting the data in the process of simplification (G31-G36), as well as abstraction to distil and emphasize key aspects of the information in its most essential detail via iconic representations to aid in the quick identification of information (G29). Consider Tesler's law: for any system, there is a certain amount of complexity that cannot be reduced (G30). Establish and communicate relationships between information and integrated elements through links and connections to guide insights and provide "added value" (G18).

Maintain Consistency and Clarity Across All Elements and Interactions to Ease Cognitive Demand

To reduce cognitive load when using the visualization, maintain consistency and clarity across design elements and interactions by standardizing attributes, modes, and features; simplifying visual elements; and using visual cues for effective notifications. Specifically, maintain consistency across similar design elements (G13), visual and spatial attributes (color and geometry [shape, size, scale, and position]; G11 and G12), and modes and features (G14, G15, and G28) to build familiarity and help manage complexity. Provide standardization across layout, format, and labels to reduce the time taken to search as well as helping develop and build a mental model to assimilate new or different information across the visualization (G14). Ensure clarity in the visual representation by eliminating extraneous elements and embellishments (G17); thus, every entity supports data insights. Develop simplicity in visual elements with well-understood representations that can be grasped when looking for the first time (G37) while being distinct to help aid the identification of elements and drawing the viewer's eye to new or updated information (G38). To rapidly call attention to new or updated information, use pre-attentive attributes (color, form, spatial positioning, and movement; G01). Communicate and provide updated or missing information and factors through discernible "alarms" or notifications (G40, G41, and G43), ensuring that these do not disrupt ongoing activities (G42) but provide projection support (G39).

Use Familiar Mental Models to Enable Intuitive Interactions

To facilitate intuitive interactions, design interfaces based on familiar mental models, use established affordances and contextual clues, and simplify tasks to reduce cognitive effort. Specifically, design the interface with existing mental models driven by user needs (G46). Use familiar, established interaction affordances to support intuitive use (G45) and reduce learning effort (G44). Minimize task complexity by ensuring a minimal number of actions to accomplish a goal or task (G56 and G59) and provide flexibility to communicate the number of possible ways to achieve a goal (G57), prompting with strong contextual clues to help know the alternatives when several actions are possible (G58). Support navigation by allowing users to undo actions, replay previous steps, and progress through the visualization (G53 and G54) and consider Hick's law, which states that "the time it takes to make a decision increases with the number and complexity of choices" (G59).

Afford Depth of Data Retrieval by Enabling Users to Zoom Into the Data

To ensure efficient data retrieval, design the visualization to allow users to navigate between overview and detailed information on demand, communicate relationships, and view data from multiple perspectives. Specifically, use established interactive approaches to minimize complexity. Commence with a holistic overview of the data (G47) and provide progressive disclosure through interactions, including zooming in on items of interest (G48), details on demand to obtain information when needed (G50), and a filter for uninteresting items (G49). A hover can reveal information faster than clicking to open and clicking to close (G51). Enable the user to rearrange or view the visualization from different perspectives (G60) and link elements to communicate the relationship between items (G52), guiding analytical reasoning and insights. Support the extraction and retention of information desired by the user once they have attained it (G55).

Discussion

Principal Findings

In this paper, we argue and establish that, in time-critical contexts, holistic visualizations that integrate multidata variables into a single visual representation are a distinct information visualization type with different needs and constraints. We position this visualization, called a HI-viz, in a 2D plot (time and information; Figure 2), highlighting that HI-viz has the

XSL•FO
RenderX

potential to convey more information in a shorter time frame. HI-viz in time-critical contexts requires approximately 30 seconds for information assimilation and processing, applying both pre-attentive and attentive processing (ie, requiring more engagement than glanceable displays but being quicker than dashboards) and integrating relatively large amounts of data (ie, similar in complexity to dashboards but as an integrated whole). This systematic review across 6 databases (ACM Digital Library, IEEE Xplore, PubMed, Scopus, Web of Science, and Google Scholar) confirms that, currently, no specific guidelines or principles were found to address all the characteristics of HI-viz, namely, holistic, integrated visualizations for time-critical contexts. As the needs and constraints differ from those of other visualization types, a contextually appropriate set of guidelines is required, combining established principles, guidelines, and taxonomies from the fields of visualization, design, and cognitive science to aid in the design of HI-viz for time-critical contexts.

HI-viz Guidelines

The 5 proposed HI-viz guidelines, derived from existing design theories and guidelines, are meant to provide design directions for practitioners and researchers working on HI-viz use cases based on theoretical grounding. Our work intentionally sought to offer practical prescriptive support. For instance, we focused on design guidelines that have observable properties (such as "make icons visually distinct") and excluded broad or conceptual principles (such as "less is more"). Such practicality is also shown through the adaptation from other known lists of guidelines (eg, the "Six Principles for Designing any Chart" by Lima [5] and "Four Principles for Glanceable Displays" by Matthews et al [17] that focus on conceptual and specific principles for visualizations and displays) together with contextual suggestions, resulting in grounded direction.

The first HI-viz guideline emphasizes the need to prioritize core information and reduce extraneous information for a holistic overview of the data. Excessive details and features irrelevant to the task can contribute to the cognitive overload and overwhelm the user [69,107], consequently reducing situational awareness and negatively impacting decision-making. Thus, reducing "visual clutter" [19] and establishing hierarchy in the data (ie, what is most important) [19,67] and design allows the user to quickly obtain the "bigger picture" and focus on the key information for rapid comprehension without sacrificing the depth of understanding. In time-critical contexts in which users rely on complex data, this facilitates a timely review of what is most important without having to switch between multiple displays or areas to assimilate information.

The second HI-viz guideline pertains to the integrated representation of information through structuring and organizing the diverse data dimensions. With larger datasets and higher information complexity, the grouping of information [104,129,134] and use of gestalt principles of organization [10,103,108,112,114,117,122,123,126] can bring together disparate data, facilitating patterns in groups over parts. Links and connections between information and elements provide the "added value" [100,107,112,118,120], facilitating relationships between data points in a concise and curated manner and, thus, distinguishing HI-viz from dashboards through the connected and integrated whole. Revisiting Finviz (Figure 3A [64]), the respective sector and industry groupings of stocks immediately illustrate stock performance across the market and whether a sector is up or down. This also resonates and aligns with Miller's law of "seven, plus or minus two" for grouping information [104], reducing cognitive overload and enabling the recall of information considering the capacity of working memory.

The third HI-viz guideline emphasizes consistency across all elements (design and interaction) and effective notifications to call attention to new or updated information. Implications of consistency across visual and spatial attributes allow the user to build familiarity and comprehend the information [19,69,103,104,106,107,129,134]. In time-critical domains such as clinical settings, inconsistent information communication can lead to misdiagnosis, incorrect treatments, and patient harm. Effective notifications can allow the user to quickly understand what information is new, updated, or missing [19,100] (eg, whether a patient's blood pressure results are recent or normal, high, or low; whether the data is old or missing data; and whether an up-to-date reading is required), improving task efficiency, accuracy, and user performance.

The fourth HI-viz guideline emphasizes the need for using familiar mental models for intuitive interactions. Implementing known affordances and contextual clues [5] ensures that the user requires minimal training, onboarding, and support. In a time-critical context, this contributes toward the efficient navigation and completion of the task [19,69,128]. The user knows how to complete the task; the number of possible ways and alternatives [69]; and how to undo, replay, and progress [5,71,100]. Thus, as per Hick's law [128], task complexity is reduced by limiting the choices and steps required, which can improve situational awareness, decision speed, and accuracy.

Here, a question arises on how user knowledge and familiarity with the data, visualization type, and task at hand influences information processing speed. For instance, the rapid comprehension of HI-viz might be strongly linked to a user's previous experience with the data and the specific visualization format used. Numerous research has shown that training or experience with the data or a specific format can impact the user's ability to understand and interact with the visualization and make decisions, consequently enhancing accuracy and efficiency [135-140]. Cognitive factors such as spatial and perceptual abilities, which differ between users, also contribute to task performance [141-145]. Future work can explore how different user groups (experts vs novices) interact with HI-viz and how previous knowledge, experience, or familiarity can impact task efficiency or information comprehension. What formats or modes of interaction cater best to varying levels of expertise? What level of detail is required for users to gather the core information?

The fifth HI-viz guideline references efficient data retrieval by allowing users to zoom in and interact with information on demand. This extends the information-seeking mantra by Shneiderman [71]—"overview first, zoom and filter, then details on demand"—a seminal principle in the design of interactive visualizations. For HI-viz, this approach affirms the

XSL•FO
RenderX

explanatory-exploratory positioning whereby users can navigate between the at-a-glance holistic overview and the detailed contextual information to attain insights across multiple layers of information [9,71,90,100,105,107,110,116,119,125]. Rearranging views or viewing data from multiple perspectives [102,105,107,110,119,121,130] can help communicate relationships within and between data dimensions and points [71,90,100,113,116,119]. Smart interactions, such as a hover [5,10], reduce unnecessary clicking between areas of the interface to gather and make sense of information (ie, reducing "click-fatigue" [1]) and, thus, facilitating quicker access to core information for task completion.

These guidelines provide a foundation for designing HI-viz that can contribute to improved situational awareness and informed decision-making [3] as well as enhancing user experience and improving information comprehension.

However, there is a trade-off between generality and specialization of guidelines [146], and therefore, these guidelines might not adequately support all time-critical domains. For example, additional specialized guidelines may be required in certain scenarios, such as in aviation, cybersecurity, finance, and military contexts. Thus, while we believe that we captured a broad range of key and relevant guidelines from both industry and academic literature as we searched the literature in a systematic way, this study acknowledges that the guidelines are derived from these databases and it is possible that other relevant guidelines to the design of HI-viz were missed. Such guidelines can be improved as future work continues to explore additional applications HI-viz, which may also highlight other theories. These guidelines serve as a starting point for designing holistic integrated visualizations in time-critical contexts and to initiate further research in this domain. Future work can further explore and contribute to the refinement of these HI-viz guidelines.

Future Work

Future work could examine the uses and value of these guidelines in different contexts and time-critical domains. For example, in the health care setting, HI-viz can be used to communicate large-scale scientific datasets for quick analysis and assist in collaborative or real-time informed decision-making by providing comprehensive and easily interpretable visual representation of complex patient data, consequently improving patient outcomes.

In a time-critical health care setting such as accident and emergency or intensive care units, HI-viz can be used to provide real-time visualizations of patient data, such as vital signs, symptoms, laboratory test results, medical history, medication or antibiotic administration [2], and triage scores. Where the fast-paced environment of an emergency room requires rapid assessment and prioritization of patients based on the severity of their condition, HI-viz can assist emergency staff in the quick identification and assessment of critical patients and prioritize their care based on the urgency and complexity of their needs as well as facilitating timely, informed decisions for appropriate interventions.

In primary care, general practitioners have highlighted a need to assimilate and understand the patient's key information within 30 seconds before a consultation [1]. There is a need to prioritize and present key information, including problem history, active and significant past, long-term conditions, medication, demographics (name, age, gender, and other socioeconomic details), and patient stratification, and present this efficiently to enhance situational awareness; enable informed decisions; and, consequently, improve patient health outcomes [1,147-150]. We aim to use the proposed guidelines in an iterative process to design and evaluate a holistic integrated visualization for supporting general practitioners in gaining a quick overview of the patient's health status before the clinical consultation as they have very limited time before and during clinical consultations [1,147,148,151,152]. The needs and priorities of stakeholders will be engaged through participatory methods (ie, focus groups, workshops, and semi-structured surveys) to ensure the acceptability of the visualization by its users. The effectiveness of the guidelines can be ascertained by how effectively the design intent is achieved (ie, enhancing efficiency, task time for retrieving information, and usability and user experience).

Conclusions

We argue that HI-viz in time-critical domains is a unique use case requiring a unique set of design guidelines. We propose 5 main guidelines derived from existing design theories and guidelines. We hope that these design guidelines will serve as a starting point to enable both holistic and rapid processing of information, facilitating better-informed decisions in time-critical contexts. As the amount of information available from multiple and heterogeneous sources of data is increasing rapidly, we see significant value in further developing and refining design guidelines for data visualization.

Acknowledgments

The authors thank the Engineering and Physical Sciences Research Council for their funding of this research.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 81 KB - jmir_v26i1e58088_app1.pdf]

Multimedia Appendix 2
Features of the holistic, integrated, time, and general guideline characteristics of the studies in the review.
[PDF File (Adobe PDF File), 88 KB - jmir_v26i1e58088_app2.pdf]

References

1. Patel AM, Porat T, Baxter WL. Enhancing situation awareness and decision making in primary care: clinicians' views. In: Proceedings of the 2020 IEEE International Conference on Healthcare Informatics. 2020 Presented at: ICHI '20; November 30-December 3, 2020; Oldenburg, Germany p. 1-9 URL: https://ieeexplore.ieee.org/document/9374398 [doi: 10.1109/ichi48887.2020.9374398]
2. Forsman J, Anani N, Eghdam A, Falkenhav M, Koch S. Integrated information visualization to support decision making for use of antibiotics in intensive care: design and usability evaluation. Inf Health Soc Care 2013 Aug 19;38(4):330-353. [doi: 10.3109/17538157.2013.812649]
3. Silva MA, Lee JM. Neurocognitive testing. In: Stein JP, editor. Reference Module in Neuroscience and Biobehavioral Psychology. Berkeley, CA: Elsevier; 2021:552-557.
4. Le T, Wilamowska K, Demiris G, Thompson H. Integrated data visualisation: an approach to capture older adults' wellness. Int J Electron Healthc 2012;7(2):89. [doi: 10.1504/ijeh.2012.049872]
5. Lima M. Six principles for designing any chart: an introduction new data visualization guideline to Google's. Medium. 2019. URL: https://medium.com/google-design/redefining-data-visualization-at-google-9bdcf2e447c6 [accessed 2022-02-22]
6. Bui AA, Hsu W. Medical data visualization: toward integrated clinical workstations. In: Bui AA, Taira RK, editors. Medical Imaging Informatics. Cham, Switzerland: Springer; 2010:139-193.
7. Featherstone C, van der Poel E. Human creativity in the data visualisation pipeline - a systematic literature review. In: Proceedings of the 2017 African Conference on Information Science and Technology. 2016 Presented at: ACIST '17; July 10-11, 2017; Cape Town, South Africa p. 13-27 URL: https://www.researchgate.net/publication/318393582_Human_creativity_in_the_data_visualisation_pipeline_-_A_systematic_literature_review
8. Engelbrecht L, Botha A, Alberts R. Designing the visualization of information. Int J Image Grap 2015 Apr 14;15(02):1540005. [doi: 10.1142/S0219467815400057]
9. Elmqvist N, Fekete JD. Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. IEEE Trans Vis Comput Graph 2010 May;16(3):439-454. [doi: 10.1109/tvcg.2009.84]
10. Ware C. Information Visualization: Perception for Design. 4th edition. Boca Raton, FL: Morgan Kaufmann Publishers; 2020.
11. Munzner T. A nested model for visualization design and validation. IEEE Trans Vis Comput Graph 2009 Nov;15(6):921-928. [doi: 10.1109/tvcg.2009.111]
12. Zhang Y, Sun Y, Padilla L, Barua S, Bertini E, Bertini AG. Mapping the landscape of COVID-19 crisis visualizations. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan p. 1-23 URL: https://dl.acm.org/doi/10.1145/3411764.3445381 [doi: 10.1145/3411764.3445381]
13. Munzner T. Visualization Design and Analysis: Abstractions, Principles, and Methods. Boca Raton, FL: CRC Press; 2012.
14. Ma Q, Millet B. Design guidelines for immersive dashboards. Proc Hum Factors Ergon Soc Annu Meet 2021 Nov 12;65(1):1524-1528. [doi: 10.1177/1071181321651177]
15. Bach B, Freeman E, Abdul-Rahman A, Turkay C, Khan S, Fan Y, et al. Dashboard design patterns. IEEE Trans Vis Comput Graph 2023;29(1):1-11. [doi: 10.1109/tvcg.2022.3209448]
16. Janes A, Sillitti A, Succi G. Effective dashboard design. Cutter IT J 2013:17-24 [FREE Full text]
17. Matthews TL, Blais D, Shick A, Mankoff J, Forlizzi J, Rohrbach S, et al. Evaluating glanceable visuals for multitasking (technical report no. UCB/EECS-2006-173). Electrical Engineering and Computer Sciences University of California. 2006. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-173.pdf [accessed 2024-04-29]
18. Neshati A, Sakamoto Y, Leboe-McGowan LC, Leboe-McGowan J, Serrano M, Irani P. G-sparks: glanceable sparklines on smartwatches. In: Proceedings of the 45th Graphics Interface Conference. 2019 Presented at: GI '19; May 28-31, 2019; Kingston, ON p. 1-9 URL: https://dl.acm.org/doi/10.20380/GI2019.23 [doi: 10.20380/GI2019.23]
19. Endsley MR. Designing for Situation Awareness: An Approach to User-Centered Design. 2nd edition. Boca Raton, FL: CRC Press; 2012.
20. Kohlhammer J, Zeltzer D. Towards a visualization architecture for time-critical applications. In: Proceedings of the 9th international conference on Intelligent user interfaces. 2004 Presented at: IUI '04; January 13-16, 2004; Funchal, Portugal p. 271-273 URL: https://dl.acm.org/doi/10.1145/964442.964502 [doi: 10.1145/964442.964502]

XSL•FO
RenderX

21. Blascheck T, Bentley F, Choe EK, Horak T, Isenberg P. Characterizing glanceable visualizations: from perception to behavior change. In: Lee B, Dachselt R, Isenberg P, Choe EK, editors. Mobile Data Visualization. Boca Raton, FL: CRC Press; 2021:1518-1576.
22. Wilkinson D, Bahirat P, Bahirat M, Lyu J, Alsubhi A, Qiu J, et al. Privacy at a glance: the user-centric design of glanceable data exposure visualizations. Proc Priv Enhancing Technol 2020(2):416-435 [FREE Full text] [doi: 10.2478/popets-2020-0034]
23. Blascheck T, Besancon L, Bezerianos A, Lee B, Isenberg P. Glanceable visualization: studies of data comparison performance on smartwatches. IEEE Trans Vis Comput Graph 2019 Jan;25(1):630-640. [doi: 10.1109/TVCG.2018.2865142]
24. Mullet K, Sano D. Designing Visual Interfaces: Communication Oriented Techniques. London, UK: Prentice Hall; 1995.
25. Pizza S, Brown B, McMillan D, Lampinen A. Smartwatch in vivo. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 5456-5469 URL: https://dl.acm.org/doi/10.1145/2858036.2858522 [doi: 10.1145/2858036.2858522]
26. Visuri A, Sarsenbayeva Z, van Berkel N, Goncalves J, Rawassizadeh R, Kostakos V, et al. Quantifying sources and types of smartwatch usage sessions. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO p. 3569-3581 URL: https://dl.acm.org/doi/10.1145/3025453.3025817 [doi: 10.1145/3025453.3025817]
27. Niess J, Knaving K, Kolb A, Woźniak PW. Exploring fitness tracker visualisations to avoid rumination. In: Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services. 2020 Presented at: MobileHCI '20; October 5-8, 2020; Oldenburg, Germany p. 1-11 URL: https://doi.org/10.1145/3379503.3405662 [doi: 10.1145/3379503.3405662]
28. Grioui F, Blascheck T. Study of heart rate visualizations on a virtual smartwatch. In: Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology. 2021 Presented at: VRST '21; December 8-10, 2021; Osaka, Japan p. 1-3 URL: https://dl.acm.org/doi/10.1145/3489849.3489913#:~:text=In%20this%20paper%2C%20we%20present,%2C%20moderate%2C%20high%20intensity) [doi: 10.1145/3489849.3489913]
29. Islam A, Bezerianos A, Blascheck T, Isenberg P. Visualizing information on watch faces: a survey with smartwatch users. In: Proceedings of the 2020 IEEE Visualization Conference. 2020 Presented at: VIS '20; October 25-30, 2021; Salt Lake City, UT p. 156-160 URL: https://ieeexplore.ieee.org/document/9331267 [doi: 10.1109/vis47514.2020.00038]
30. Schiewe A, Krekhov A, Kerber F, Daiber F, Krüger J. A study on real-time visualizations during sports activities on smartwatches. In: Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia. 2020 Presented at: MUM '20; November 22-25, 2020; Essen, Germany p. 18-31 URL: https://dl.acm.org/doi/10.1145/3428361.3428409 [doi: 10.1145/3428361.3428409]
31. Li Q. Overview of data visualization. In: Li Q, editor. Embodying Data: Chinese Aesthetics, Interactive Visualization and Gaming Technologies. Cham, Switzerland: Springer; 2020:17-47.
32. Cawthon N, Moere AV. Qualities of perceived aesthetic in data visualization. In: Proceedings of the 2007 conference on Designing for User eXperiences. 2007 Presented at: DUX '07; November 5-7, 2007; Chicago, IL p. 1 URL: https://dl.acm.org/doi/10.1145/1389908.1389920 [doi: 10.1145/1389908.1389920]
33. Saket B, Endert A, Demiralp C. Task-based effectiveness of basic visualizations. IEEE Trans Vis Comput Graph 2019 Jul 1;25(7):2505-2512. [doi: 10.1109/tvcg.2018.2829750]
34. Kay M, Kola T, Hullman JR, Munson SA. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 5092-5103 URL: https://dl.acm.org/doi/10.1145/2858036.2858558 [doi: 10.1145/2858036.2858558]
35. Kay M, Heer J. Beyond Weber's law: a second look at ranking visualizations of correlation. IEEE Trans Vis Comput Graph 2016 Jan;22(1):469-478. [doi: 10.1109/tvcg.2015.2467671]
36. Harrison L, Yang F, Franconeri S, Chang R. Ranking visualizations of correlation using Weber's law. IEEE Trans Vis Comput Graph 2014 Dec 31;20(12):1943-1952. [doi: 10.1109/tvcg.2014.2346979]
37. Eells WC. The relative merits of circles and bars for representing component parts. J Am Stat Assoc 1926 Jun;21(154):119-132. [doi: 10.1080/01621459.1926.10502165]
38. Croxton FE, Stryker RE. Bar charts versus circle diagrams. J Am Stat Assoc 1927 Dec;22(160):473-482. [doi: 10.1080/01621459.1927.10502976]
39. Dambacher M, Haffke P, Groß D, Hübner R. Graphs versus numbers: how information format affects risk aversion in gambling. Judgm Decis Mak 2023 Jan 01;11(3):223-242. [doi: 10.1017/S1930297500003077]
40. Correll M, Gleicher M. Error bars considered harmful: exploring alternate encodings for mean and error. IEEE Trans Vis Comput Graph 2014 Dec 31;20(12):2142-2151. [doi: 10.1109/tvcg.2014.2346298]
41. Porat T, Oron-Gilad T, Meyer J. Task-dependent processing of tables and graphs. Behav Inf Technol 2009 May;28(3):293-307. [doi: 10.1080/01449290701803516]
42. Pandre A. Dimensionality of visible data. WordPress. 2010. URL: https://apandre.wordpress.com/2010/02/02/dimensionality/ [accessed 2022-02-02]

XSL•FO
RenderX

43. Iliinsky N, Steele J. Designing Data Visualizations: Representing Informational Relationships. Sebastopol, CA: O'Reilly Media; 2011.
44. Sarkar D. The art of effective visualization of multi-dimensional data: strategies for effective data visualization. Medium. 2018. URL: https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57 [accessed 2024-04-29]
45. Rohwer A, Taylor M, Ryan R, Garner P, Oliver S. Enhancing public health systematic reviews with diagram visualization. Am J Public Health 2021 Jun;111(6):1029-1034. [doi: 10.2105/ajph.2021.306225]
46. Umoquit MJ, Tso P, Burchett HE, Dobrow MJ. A multidisciplinary systematic review of the use of diagrams as a means of collecting data from research subjects: application, benefits and recommendations. BMC Med Res Methodol 2011 Jan 27;11(1):16 [FREE Full text] [doi: 10.1186/1471-2288-11-11]
47. Nguyen QV, Miller N, Arness D, Huang W, Huang ML, Simoff S. Evaluation on interactive visualization data with scatterplots. Vis Inform 2020 Dec;4(4):1-10 [FREE Full text] [doi: 10.1016/j.visinf.2020.09.004]
48. Reitsma R, Marks A. The future of data: too much visualization, too little understanding? An empirical investigation into the comprehension of data visualizations. Dialectic 2019;2(2):12-34 [FREE Full text] [doi: 10.3998/dialectic.14932326.0002.207]
49. Wiederkehr B. Data visualization for exploration and explanation: inspiration and guidance for balancing interactive and narrative approaches when communicating with data. Medium. 2021. URL: https://blog.interactivethings.com/data-visualization-for-exploration-and-explanation-650afd901574 [accessed 2022-02-02]
50. Crisan A. The importance of data visualization in combating a pandemic. Am J Public Health 2022 Jun;112(6):893-895. [doi: 10.2105/ajph.2022.306857]
51. Sarikaya A, Correll M, Bartram LR, Tory MK, Fisher D. What do we talk about when we talk about dashboards? IEEE Trans Vis Comput Graph 2019 Jan;25(1):682-692. [doi: 10.1109/tvcg.2018.2864903]
52. Kruglov A, Strugar D, Succi G. Tailored performance dashboards-an evaluation of the state of the art. PeerJ Comput Sci 2021;7:e625 [FREE Full text] [doi: 10.7717/peerj-cs.625] [Medline: 34712785]
53. Brown B, Balatsoukas P, Williams R, Sperrin M, Buchan I. Multi-method laboratory user evaluation of an actionable clinical performance information system: implications for usability and patient safety. J Biomed Inform 2018 Jan;77:62-80 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.008] [Medline: 29146562]
54. Schall M, Cullen L, Pennathur P, Chen H, Burrell K, Matthews G. Usability evaluation and implementation of a health information technology dashboard of evidence-based quality indicators. Comput Inform Nurs 2017 Jun;35(6):281-288. [doi: 10.1097/CIN.0000000000000325] [Medline: 28005564]
55. Dolan JG, Veazie PJ, Russ AJ. Development and initial evaluation of a treatment decision dashboard. BMC Med Inform Decis Mak 2013 Apr 21;13:51 [FREE Full text] [doi: 10.1186/1472-6947-13-51] [Medline: 23601912]
56. Ferreira D, Goncalves J, Goncalves V, Barkhuus L, Dey AK. Contextual experience sampling of mobile application micro-usage. In: Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services. 2014 Presented at: MobileHCI '14; September 23-26, 2014; Toronto, ON p. 91-100 URL: https://dl.acm.org/doi/10.1145/2628363.2628367 [doi: 10.1145/2628363.2628367]
57. Albers M. Infographics: horrid chartjunk or quality communication. In: Proceedings of the 2014 IEEE International Professional Communication Conference. 2014 Presented at: IPCC '14; October 13-15, 2014; Pittsburgh, PA p. 1-4 URL: https://ieeexplore.ieee.org/abstract/document/7020344 [doi: 10.1109/ipcc.2014.7020344]
58. Krum R. Cool Infographics: Effective Communication with Data Visualization and Design. Hoboken, NJ: John Wiley & Sons; 2013.
59. Dunlap JC, Lowenthal PR. Getting graphic about infographics: design lessons learned from popular infographics. J Vis Lit 2016 Sep 08;35(1):42-59. [doi: 10.1080/1051144x.2016.1205832]
60. Lamb GR, Polman JL, Newman A, Smith CG. Science news infographics: teaching students to gather, interpret, and present information graphically. Sci Teacher 2014;081(03):25-30. [doi: 10.2505/4/tst14_081_03_25]
61. Toth C. Revisiting a genre: teaching infographics in business and professional communication courses. Bus Commun Q 2013 Nov 12;76(4):446-457. [doi: 10.1177/1080569913506253]
62. Ennis BB, Hinton S, Whitelaw M. Playing with complexity: an approach to exploratory data visualisation. Australian Council of University Art and Design Schools. 2011. URL: https://acuads.com.au/conference/article/playing-with-complexity-an-approach-to-exploratory-data-visualisation/ [accessed 2024-04-29]
63. FIFA 22 player rankings: Ronaldo bumped out of top two, Messi still No. 1. ESPN. URL: https://www.espn.com/soccer/blog-the-toe-poke/story/4474417/fifa-22-player-ratings-ronaldo-bumped-out-of-top-two-messi-still-no-1 [accessed 2023-03-23]
64. Finviz Financial Visualizations. URL: https://finviz.com/map.ashx?t=sec [accessed 2023-03-23]
65. Le T, Reeder B, Yoo D, Aziz R, Thompson HJ, Demiris G. An evaluation of wellness assessment visualizations for older adults. Telemed J E Health 2015 Jan;21(1):9-15 [FREE Full text] [doi: 10.1089/tmj.2014.0012] [Medline: 25401414]
66. Akanmu S, Jamaludin Z. Information visualization techniques usage model. In: Proceedings of the 4th International Conference on Computing and Informatics. 2013 Presented at: ICOCI '13; August 28-30, 2013; Sarawak, Malaysia p. 28-30 URL: https://soc.uum.edu.my/icoci/2023/icoci2013/PDF/PID45.pdf

XSL•FO
RenderX

67. Midway SR. Principles of effective data visualization. Patterns 2020 Dec;1(9):100141. [doi: 10.1016/j.patter.2020.100141]
68. Eppler MJ, Burkhard RA. Visual representations in knowledge management: framework and cases. J Knowl Manag 2007;11(4):112-122 [FREE Full text] [doi: 10.1108/13673270710762756]
69. Forsell C, Johansson J. An heuristic set for evaluation in information visualization. In: Proceedings of the 2010 International Conference on Advanced Visual Interfaces. 2010 Presented at: AVI '10; May 26-28, 2010; Rome, Italy p. 199-206 URL: https://dl.acm.org/doi/10.1145/1842993.1843029 [doi: 10.1145/1842993.1843029]
70. Ribecca S. Home page. The Data Visualisation Catalogue. 2015. URL: http://datavizcatalogue.com/ [accessed 2024-04-29]
71. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages. 1996 Presented at: VL '96; September 3-6, 1996; Boulder, CO p. 336-343 URL: https://doi.org/10.1109/VL.1996.545307 [doi: 10.1016/b978-155860915-0/50046-9]
72. Heer J, Kong N, Agrawala M. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009 Presented at: CHI '09; April 4-9, 2009; Boston, MA p. 1303-1312 URL: https://dl.acm.org/doi/10.1145/1518701.1518897 [doi: 10.1145/1518701.1518897]
73. Keim DA. Information visualization and visual data mining. IEEE Trans Vis Comput Graph 2002;8(1):1-8 [FREE Full text] [doi: 10.1109/2945.981847]
74. Lengler R, Eppler MJ. Towards a periodic table of visualization methods of management. In: Proceedings of the 2007 IASTED International Conference on Graphics and Visualization in Engineerin. 2007 Presented at: GVE '07; January 3-5, 2007; Clearwater, FL p. 83-88 URL: https://dl.acm.org/doi/10.5555/1712936.1712954
75. Amar R, Stasko J. BEST PAPER: a knowledge task-based framework for design and evaluation of information visualizations. In: Proceedings of the 2004 Conference on IEEE Symposium on Information Visualization. 2004 Presented at: INFVIS' 04; October 10-12, 2004; Austin, TX p. 143-150 URL: https://ieeexplore.ieee.org/document/1382902 [doi: 10.1109/infvis.2004.10]
76. Amar R, Eagan J, Stasko J. Low-level components of analytic activity in information visualization. In: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization. 2005 Presented at: INFOVIS '05; October 23-25, 2005; Washington, DC p. 111-117 URL: https://dl.acm.org/doi/10.1109/INFOVIS.2005.24 [doi: 10.1109/infvis.2005.1532136]
77. Valiati ER, Pimenta MS, Freitas CM. A taxonomy of tasks for guiding the evaluation of multidimensional visualizations. In: Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization. 2006 Presented at: BELIV '06; May 23, 2006; Venice, Italy p. 1-6 URL: https://dl.acm.org/doi/10.1145/1168149.1168169 [doi: 10.1145/1168149.1168169]
78. Lee B, Plaisant C, Parr CS, Fekete JD, Henry N. Task taxonomy for graph visualization. In: Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization. 2006 Presented at: BELIV '06; May 23, 2006; Venice, Italy p. 1-5 URL: https://dl.acm.org/doi/10.1145/1168149.1168168 [doi: 10.1145/1168149.1168168]
79. Wickham H. A layered grammar of graphics. J Comput Graph Stat 2010 Jan;19(1):3-28. [doi: 10.1198/jcgs.2009.07098]
80. Bostock M, Heer J. Protovis: a graphical toolkit for visualization. IEEE Trans Vis Comput Graph 2009 Nov;15(6):1121-1128. [doi: 10.1109/tvcg.2009.174]
81. Wilkinson L. The Grammar of Graphics. Cham, Switzerland: Springer; 2006.
82. Figueiras A. Narrative visualization: a case study of how to incorporate narrative elements in existing visualizations. In: Proceedings of the 18th International Conference on Information Visualisation. 2014 Presented at: IV '14; July 16-18, 2014; Paris, France p. 46-52 URL: https://ieeexplore.ieee.org/document/6902879 [doi: 10.1109/iv.2014.79]
83. Khataei A, Lau D. Recommender narrative visualization. In: Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research. 2013 Presented at: CASCON '13; November 18-20, 2013; Toronto, ON p. 415-421 URL: https://dl.acm.org/doi/10.5555/2555523.2555588
84. Kosara R, Mackinlay J. Storytelling: the next step for visualization. Computer 2013 May;46(5):44-50. [doi: 10.1109/mc.2013.36]
85. Katz J. Designing Information: Human Factors and Common Sense in Information Design. Hoboken, NJ: John Wiley & Sons; 2012.
86. Kirk A. Data Visualization: A Successful Design Process. Birmingham, UK: Packt Publishers; 2012.
87. Hullman J, Diakopoulos N. Visualization rhetoric: framing effects in narrative visualization. IEEE Trans Vis Comput Graph 2011 Dec;17(12):2231-2240. [doi: 10.1109/tvcg.2011.255]
88. Segel E, Heer J. Narrative visualization: telling stories with data. IEEE Trans Vis Comput Graph 2010 Nov;16(6):1139-1148. [doi: 10.1109/tvcg.2010.179]
89. Pereira FC. Creativity and Artificial Intelligence: A Conceptual Blending Approach. Munich, Germany: Walter de Gruyter; 2007.
90. Wang Baldonado MQ, Woodruff A, Kuchinsky A. Guidelines for using multiple views in information visualization. In: Proceedings of the 2000 Working Conference on Advanced Visual Interfaces. 2000 Presented at: AVI '00; May 1, 2000; Palermo, Italy p. 110-119 URL: https://dl.acm.org/doi/10.1145/345513.345271 [doi: 10.1145/345513.345271]

XSL•FO
RenderX

91. North C, Shneiderman B. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In: Proceedings of the 2000 Working Conference on Advanced Visual Interfaces. 2000 Presented at: AVI '00; May 1, 2000; Palermo, Italy p. 128-135 URL: https://dl.acm.org/doi/10.1145/345513.345282 [doi: 10.1145/345513.345282]
92. Byrne L, Angus D, Wiles J. Acquired codes of meaning in data visualization and infographics: beyond perceptual primitives. IEEE Trans Vis Comput Graph 2016 Jan 31;22(1):509-518. [doi: 10.1109/tvcg.2015.2467321]
93. Beegel JM. Infographics for Dummies. Hoboken, NJ: John Wiley & Sons; 2014.
94. Lankow J, Ritchie J, Crooks R. Infographics: The Power of Visual Storytelling. Stoughton, MA: Adams Media; 2012.
95. Wickham H, Lawrence M, Cook D, Buja A, Hofmann H, Swayne DF. The plumbing of interactive graphics. Comput Stat 2008 May 10;24(2):207-215. [doi: 10.1007/s00180-008-0116-x]
96. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
97. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
98. Adnan WA, Daud NG, Noor NL. Expressive information visualization taxonomy for decision support environment. In: Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology. 2008 Presented at: ICCIT '08; November 11-13, 2008; Busan, South Korea p. 88-93 URL: https://ieeexplore.ieee.org/document/4682007 [doi: 10.1109/iccit.2008.330]
99. Ajani K, Lee E, Xiong C, Knaflic CN, Kemper W, Franconeri S. Declutter and focus: empirically evaluating design guidelines for effective data communication. IEEE Trans Vis Comput Graph 2022 Oct 1;28(10):3351-3364. [doi: 10.1109/tvcg.2021.3068337]
100. Andrews C, Endert A, Yost B, North C. Information visualization on large, high-resolution displays: issues, challenges, and opportunities. Inf Vis 2011 Aug 22;10(4):341-355. [doi: 10.1177/1473871611415997]
101. Bateman S, Mandryk RL, Gutwin C, Genest A, McDine D. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In: Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems. 2010 Presented at: CHI '10; April 10-15, 2010; Atlanta, Georgia p. 2573-2582 URL: https://dl.acm.org/doi/10.1145/1753326.1753716 [doi: 10.1145/1753326.1753716]
102. Brath R. 3D interactive information visualization: guidelines from experience and analysis of applications. In: Proceedings of the 7th International Conference on Human-Computer Interaction. 1997 Presented at: HCI '17; August 24-29, 1997; San Francisco, CA p. 865-868 URL: https://www.researchgate.net/publication/221096538_3D_Interactive_Information_Visualization_Guidelines_from_Experience_and_Analysis_of_Applications
103. Chang D, Nesbitt KV. Developing Gestalt-based design guidelines for multi-sensory displays. In: Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop. 2006 Presented at: MMUI '05; September 13, 2005; Sydney, Australia p. 9-16 URL: https://dl.acm.org/doi/10.5555/1151804.1151807
104. Cowan N. George Miller's magical number of immediate memory in retrospect: observations on the faltering progression of science. Psychol Rev 2015 Jul;122(3):536-541 [FREE Full text] [doi: 10.1037/a0039035] [Medline: 25751370]
105. De Carlo G, Langer P, Bork D. Rethinking model representation - a taxonomy of advanced information visualization in conceptual modeling. In: Proceedings of the 41st International Conference on Conceptual Modeling. 2022 Presented at: ER '22; October 17-20, 2022; Hyderabad, India p. 35-51 URL: https://doi.org/10.1007/978-3-031-17995-2_3 [doi: 10.1007/978-3-031-17995-2]
106. De Weck OL, Roos D, Magee CL. Engineering Systems: Meeting Human Needs in a Complex Technological World. Cambridge, MA: MIT Press; 2011.
107. Dix A. Introduction to information visualisation. In: Proceedings of the 2012 Conference on Information Retrieval Meets Information Visualization. 2013 Presented at: PROMISE '12; January 23-27, 2012; Zinal, Switzerland p. 1-27 URL: https://link.springer.com/chapter/10.1007/978-3-642-36415-0_1 [doi: 10.1007/978-3-642-36415-0_1]
108. Dybala MP, Butterfield JK, Hendren-Santiago BK, Hara M. Pancreatic islets and gestalt principles. Diabetes 2020 Sep;69(9):1864-1874 [FREE Full text] [doi: 10.2337/db20-0304] [Medline: 32669392]
109. Ellis G, Dix A. A taxonomy of clutter reduction for information visualisation. IEEE Trans Vis Comput Graph 2007 Nov;13(6):1216-1223. [doi: 10.1109/tvcg.2007.70535]
110. Fekete JD, Plaisant C. Interactive information visualization of a million items. In: Proceedings of the 2002 IEEE Symposium on Information Visualization. 2002 Presented at: INFOVIS '02; October 28-29, 2002; Boston, MA p. 117-124 URL: https://ieeexplore.ieee.org/document/1173156 [doi: 10.1109/infvis.2002.1173156]
111. Healey CG, Booth KS, Enns JT. Visualizing real-time multivariate data using preattentive processing. ACM Trans Model Comput Simul 1995 Jul;5(3):190-221. [doi: 10.1145/217853.217855]
112. Idrissov A, Škec S, Maier AM. Visualising systems: mapping system features and interactive information visualisations in design. Proc Des Soc Des Conf 2020 Jun 11;1:2295-2304. [doi: 10.1017/dsd.2020.138]
113. Kelleher C, Wagener T. Ten guidelines for effective data visualization in scientific publications. Environ Model Softw 2011 Jun;26(6):822-827 [FREE Full text] [doi: 10.1016/j.envsoft.2010.12.006]

XSL•FO
RenderX

114. Kornhauser D, Wilensky U, Rand W. Design guidelines for agent based model visualization. J Artif Soc Soc Simul 2009;12(2):1 [FREE Full text]
115. Krekhov A, Kruger J. Deadeye: a novel preattentive visualization technique based on dichoptic presentation. IEEE Trans Vis Comput Graph 2019 Jan;25(1):936-945. [doi: 10.1109/tvcg.2018.2864498]
116. Ku CH, Nguyen JH, Leroy G. TASC - crime report visualization for investigative analysis: a case study. In: Proceedings of the 13th International Conference on Information Reuse & Integration. 2012 Presented at: IRI '12; August 8-10, 2012; Las Vegas, NV p. 466-473 URL: https://ieeexplore.ieee.org/document/6303045 [doi: 10.1109/iri.2012.6303045]
117. Moreira J, Mendes D, Gonçalves D. Incidental visualizations: pre-attentive primitive visual tasks. In: Proceedings of the 2020 International Conference on Advanced Visual Interfaces. 2020 Presented at: AVI '20; September 28-October 2, 2020; Salerno, Italy p. 1-9 URL: https://dl.acm.org/doi/10.1145/3399715.3399841 [doi: 10.1145/3399715.3399841]
118. Murray P, McGee F, Forbes AG. A taxonomy of visualization tasks for the analysis of biological pathway data. BMC Bioinformatics 2017 Feb 15;18(S2):21. [doi: 10.1186/S12859-016-1443-5]
119. Perer A, Shneiderman B. Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. IEEE Comput Grap Appl 2009 May;29(3):39-51. [doi: 10.1109/mcg.2009.44]
120. Rechtin E. Systems Architecting: Creating & Building Complex Systems. New York, NY: Prentice Hall; Jun 23, 1991.
121. Rodrigues JF, Traina AJ, de Oliveira CM, Traina C. Reviewing data visualization: an analytical taxonomical study. In: Proceedings of the 10th International Conference on Information Visualisation. 2006 Presented at: IV '06; July 5-7, 2006; London, UK p. 713-720 URL: https://ieeexplore.ieee.org/abstract/document/1648338 [doi: 10.1109/iv.2006.94]
122. Rosli MH, Cabrera A. Gestalt principles in multimodal data representation. IEEE Comput Grap Appl 2015 Mar;35(2):80-87. [doi: 10.1109/mcg.2015.29]
123. Seong Y, Nuamah J, Yi S. Guidelines for cybersecurity visualization design. In: Proceedings of the 24th Symposium on International Database Engineering & Applications. 2020 Presented at: IDEAS '20; August 12-14, 2020; Seoul, Republic of Korea p. 1-6 URL: https://dl.acm.org/doi/10.1145/3410566.3410606 [doi: 10.1145/3410566.3410606]
124. Shaalan D, Jusoh S. Visualization in medical system interfaces: UX guidelines. In: Proceedings of the 12th International Conference on Electronics, Computers and Artificial Intelligence. 2020 Presented at: ECAI '20; June 25-27, 2020; Bucharest, Romania p. 1-8 URL: https://ieeexplore.ieee.org/document/9223236 [doi: 10.1109/ecai50035.2020.9223236]
125. van Ham F, Perer A. “Search, show context, expand on demand”: supporting large graph exploration with degree-of-interest. IEEE Trans Vis Comput Graph 2009 Nov;15(6):953-960 [FREE Full text] [doi: 10.1109/tvcg.2009.108]
126. Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, et al. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. Psychol Bull 2012;138(6):1172-1217. [doi: 10.1037/a0029333]
127. Wenzel S, Bernhard J, Jessen U. A taxonomy of visualization techniques for simulation in production and logistics. In: Proceedings of the 2003 Winter Simulation Conference. 2003 Presented at: WSC '03; December 7-10, 2003; New Orleans, LA p. 729-736 URL: https://ieeexplore.ieee.org/document/1261489 [doi: 10.1109/wsc.2003.1261489]
128. Hick WE. On the rate of gain of information. Q J Exp Psychol 1952 Mar 01;4(1):11-26. [doi: 10.1080/17470215208416600]
129. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956 Mar;63(2):81-97. [doi: 10.1037/h0043158]
130. Spence R. Information Visualization. New York, NY: AMC Press; 2001.
131. Treisman A. Preattentive processing in vision. Compute Visio Image Process 1985 Aug;31(2):156-177. [doi: 10.1016/s0734-189x(85)80004-9]
132. Tufte ER. The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press; 1983.
133. Yablonski J. Tesler's law? In: Yablonski J, editor. Laws of UX: Using Psychology to Design Better Products & Services. New York, NY: O′Reilly Publishers; 2020:87.
134. Moran K. How chunking helps content processing. Nielsen Norman Group. 2016. URL: https://www.nngroup.com/articles/chunking/ [accessed 2022-02-22]
135. Eberhard K. The effects of visualization on judgment and decision-making: a systematic literature review. Manag Rev Q 2023 Aug 25;73(1):167-214. [doi: 10.1007/S11301-021-00235-8]
136. Hilton NZ, Ham E, Nunes KL, Rodrigues NC, Frank C, Seto MC. Using graphs to improve violence risk communication. Crim Justice Behav 2017 Sep 22;44(5):678-694. [doi: 10.1177/0093854816668916]
137. Conati C, Carenini G, Hoque E, Steichen B, Toker D. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. Comput Graph Forum 2014 Jul 12;33(3):371-380. [doi: 10.1111/cgf.12393]
138. Raschke RL, Steinbart PJ. Mitigating the effects of misleading graphs on decisions by educating users about the principles of graph design. J Inf Syst 2008;22(2):23-52. [doi: 10.2308/jis.2008.22.2.23]
139. Meyer J. Performance with tables and graphs: effects of training and a visual search model. Ergonomics 2000 Nov 10;43(11):1840-1865. [doi: 10.1080/00140130050174509]
140. Boy J, Rensink RA, Bertini E, Fekete JD. A principled way of assessing visualization literacy. IEEE Trans Vis Comput Graph 2014 Dec 31;20(12):1963-1972. [doi: 10.1109/tvcg.2014.2346984]

XSL•FO
RenderX

141. Ziemkiewicz C, Ottley A, Crouser RJ, Chauncey K, Su SL, Chang R. Understanding visualization by understanding individual users. IEEE Comput Grap Appl 2012 Nov;32(6):88-94. [doi: 10.1109/mcg.2012.120]
142. Conati C, Maclaren H. Exploring the role of individual differences in information visualization. In: Proceedings of the 2008 Working Conference on Advanced Visual Interfaces. 2008 Presented at: AVI '08; May 28-30, 2008; Napoli, Italy p. 199-206 URL: https://dl.acm.org/doi/10.1145/1385569.1385602 [doi: 10.1145/1385569.1385602]
143. Cohen CA, Hegarty M. Individual differences in use of external visualisations to perform an internal visualisation task. Appl Cogn Psychol 2007 Aug 22;21(6):701-711. [doi: 10.1002/acp.1344]
144. Velez MC, Silver D, Tremaine M. Understanding visualization through spatial ability differences. In: Proceedings of the 2005 Conference on IEEE Visualization. 2005 Presented at: VIS '05; October 23-28, 2005; Minneapolis, MN p. 511-518 URL: https://ieeexplore.ieee.org/document/1532836 [doi: 10.1109/visual.2005.1532836]
145. Chen C, Czerwinski M. Spatial ability and visual navigation: an empirical study. New Rev Hypermedia Multimed 1997 Jan;3(1):67-89. [doi: 10.1080/13614569708914684]
146. Amershi S, Weld DS, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, Scotland p. 1-13 URL: https://dl.acm.org/doi/10.1145/3290605.3300233 [doi: 10.1145/3290605.3300233]
147. Patel A, Porat T, Baxter W. What Errors Impact Situation Awareness and Decision Making for Primary Care Clinicians? 2023 Presented at: Society for Medical Decision-Making 18th Biennial European Conference; May 21-23, 2023; Berlin, Germany URL: https://openaccess.cms-conferences.org/publications/book/978-1-958651-55-1/article/978-1-958651-55-1_7 [doi: 10.54941/ahfe1003473]
148. Patel AM. The design and evaluation of an information visualization to improve the efficiency and understanding of patient health state. In: Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics. 2023 Presented at: ICHI '23; June 26-29, 2023; Houston, TA p. 26-29 URL: https://www.computer.org/csdl/proceedings-article/ichi/2023/026300a477/1SN7uYfHGKs [doi: 10.1109/ichi57859.2023.00071]
149. Patel A, Porat T, Baxter W. A taxonomy of situation awareness failure factors in primary care. In: Proceedings of the 2023 International Conference on Applied Human Factors and Ergonomics. 2023 Presented at: AHFE '23; July 20-24, 2023; San Francisco, CA p. 1-8 URL: https://openaccess.cms-conferences.org/publications/book/978-1-958651-55-1/article/978-1-958651-55-1_7 [doi: 10.54941/ahfe1003473]
150. Patel AM, Porat T, Baxter W. A taxonomy of situation awareness failure factors in primary care poster presentation. In: Proceedings of the 2023 International Conference on Applied Human Factors and Ergonomics. 2023 Presented at: CHI '23; June 26-29, 2023; Houston, TX p. 26 URL: https://openaccess.cms-conferences.org/publications/book/978-1-958651-55-1/article/978-1-958651-55-1_7 [doi: 10.1109/ichi57859.2023.00086]
151. Porat T, Delaney B, Kostopoulou O. The impact of a diagnostic decision support system on the consultation: perceptions of GPs and patients. BMC Med Inform Decis Mak 2017 Jun 2;17(1):79. [doi: 10.1186/s12911-017-0477-6]
152. Porat T, Kostopoulou O, Woolley A, Delaney BC. Eliciting user decision requirements for designing computerized diagnostic support for family physicians. J Cogn Eng Decis Mak 2015 Oct 07;10(1):57-73. [doi: 10.1177/1555343415608973]

Abbreviations

HI-viz: holistic integrated information visualization
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 05.03.24; peer-reviewed by J Sleigh, M Endsley; comments to author 08.08.24; revised version received 29.08.24; accepted 24.09.24; published 20.11.24.

Please cite as:
Patel AM, Baxter W, Porat T
Toward Guidelines for Designing Holistic Integrated Information Visualizations for Time-Critical Contexts: Systematic Review
J Med Internet Res 2024;26:e58088
URL: https://www.jmir.org/2024/1/e58088
doi:10.2196/58088
PMID:

©Ahmed Mohammed Patel, Weston Baxter, Talya Porat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Reverse Empowerment and Proactive Motivations on Physicians' Online Knowledge Sharing in Digital Platforms: Survey Study

Jingyuan Su[1], PhD; Kathy Ning Shen[2], PhD; Xitong Guo[3], PhD

[1]School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou, China

[2]College of Business and Economics, Human Capital Research Center, United Arab Emirates University, Al Ain, United Arab Emirates

[3]eHealth Research Institute, School of Management, Harbin Institute of Technology, Harbin, China

Corresponding Author:
Kathy Ning Shen, PhD
College of Business and Economics
Human Capital Research Center
United Arab Emirates University
PO Box 15551
Al Ain
United Arab Emirates
Phone: 971 556543825
Fax: 971 03 7135270
Email: ningshen@uaeu.ac.ae

Abstract

Background: Digital platforms offer a venue for patients and physicians to exchange health information and provide health care services outside traditional organizational contexts. Previous studies have seldom focused on the factors that drive the proactivity of physicians' online behavior. Additionally, there is limited research exploring the influence of patients on physicians' online behavior, particularly from the perspective of patients possessing power.

Objective: This study aims to investigate the effect of patient-empowering behaviors on physicians' online knowledge sharing and uncover the potential mechanisms of proactivity. Based on the proactive motivational model and empowerment theory, we propose the existence of a reverse empowerment process, where empowerment flows from patients to physicians. We suggest that patient-empowering behaviors may drive physicians' online knowledge sharing as a form of proactivity. Specifically, 3 proactive motivational states—knowledge-sharing self-efficacy, sharing meaning, and positive professional affect—mediate this relationship. Additionally, platform extrinsic rewards, as a contextual factor, have a moderating effect.

Methods: To validate our proposed research model, we conducted a survey in China using the WJX platform, targeting physicians engaged in online knowledge sharing. The measurement instrument utilizes validated items adapted from prior research, using a 5-point Likert scale. We collected 257 valid responses, ensuring that participation was both anonymous and voluntary. Data analysis was performed in 2 stages. The first stage assessed the measurement model for reliability and validity, using the Harman 1-factor test and confirmatory factor analysis. The second stage used partial least squares-structural equation modeling to examine the direct, moderation, and mediation effects among the constructs, with bootstrapping used for significance testing. This comprehensive approach ensures a robust analysis of the proposed hypotheses and contributes to the overall validity of our research model.

Results: Perceived patient-empowering behaviors significantly and positively influence physicians' online knowledge sharing (β=0.27, P<.001). Knowledge-sharing self-efficacy (effect=0.06, P=.04), sharing meaning (effect=0.12, P<.001), and positive professional affect (effect=0.10, P=.003) each partially mediate the effect of patient-empowering behaviors on physicians' online knowledge sharing. The overall proactive motivational states play a complete mediation role, meaning the entire indirect effect of the model is significant (effect=0.29, P<.001), while the direct effect in the model is nonsignificant (effect=0.07, P=.26). Additionally, platform extrinsic rewards significantly and negatively moderate the effect of sharing meaning on physicians' online knowledge sharing (β=–0.13, P=.001).

Conclusions: This study is the first to recognize and examine proactivity as an alternative mediating mechanism for physicians' online knowledge sharing, highlighting the active role of patients in empowering physicians. It makes a significant contribution

XSL•FO
RenderX

to the existing literature on empowerment, eHealth, and proactive behavior. Additionally, the findings offer valuable guidance for designing and managing digital platforms to ensure service sustainability.

(J Med Internet Res 2024;26:e59904) doi:10.2196/59904

KEYWORDS

physician behavior; online knowledge sharing; proactivity; patient empowerment; digital platforms; health communication

Introduction

Background

In recent decades, there has been a significant shift, with more physicians embracing digital platforms for professional engagement. Platforms such as TikTok (ByteDance Ltd.) and HaoDF (Beijing Interactive Peak Technology Co., Ltd.) have implemented incentive mechanisms encouraging certified physicians to create and maintain personal channels for distributing health-related information. Research suggests that online information exchanges between physicians and patients offer numerous benefits, including enhanced patient outcomes, stronger physician-patient relationships, and reduced health disparities [1,2]. Online knowledge sharing generally refers to the exchange of health-related information between key stakeholders, such as physicians and patients, facilitated by digital technology platforms [3,4]. Digital platforms, particularly third-party online health platforms, provide opportunities for alternative interactions between patients and physicians beyond organizational boundaries, thereby challenging traditional health professional practices and expectations [5,6]. As a result, physicians' participation in online knowledge sharing, such as responding to patient inquiries or posting health-related content, is considered an extra-role behavior. This activity is often carried out without organizational enforcement or immediate rewards and, in many ways, deviates from the professional practices traditionally defined by organizations and the health care industry [7].

Most prior research on physicians' online knowledge sharing uses social exchange theory, framing it as a trade-off between the benefits and costs of participation [3,8-10]. This approach presupposes predefined expectations and exchange counterparts, which may not be present on open digital platforms. Other studies adopt an individualistic perspective, using motivation theory to investigate the intrinsic and extrinsic drivers of online knowledge sharing [5,6,11,12]. Although insightful, this perspective may overlook or underemphasize the role of altruism in physicians' online knowledge sharing.

The adoption of digital platforms represents a paradigm shift from traditional health care, where physicians were regarded as the sole authority and patients as passive recipients of information [7]. On these platforms, patients actively influence physician behavior through actions such as information-seeking, commentary, advocacy, and even boycotts. While previous research has acknowledged the role of patients by examining platform features such as ratings, reviews, and gifts as motivators for knowledge sharing [8,11,13], there is a notable lack of studies directly addressing the significant influence patients exert on physicians' online knowledge sharing.

This research aims to deepen the understanding of physicians' online knowledge sharing by exploring the roles of proactivity and patient behavior. Proactivity involves actively seeking to make a difference in others' lives or the surrounding environment to achieve better future outcomes [14]. It is characterized by self-initiated efforts (self-starting) to improve processes (situation-changing) and enhance effectiveness over the long term (future-focused) [15]. Previous studies have highlighted the role of proactivity in knowledge sharing. For example, Mittal et al [16] argued that prosocial behavior is inherently proactive, while Kang et al [17] identified users' knowledge-sharing behaviors within a firm's knowledge management system as examples of proactivity [17]. To encourage proactivity, both digital platforms and patients play pivotal roles. Rather than viewing patients as passive participants in physician-patient interactions, this study examines the active, empowering role that patients can assume.

Accordingly, the central research question we seek to address is: What influence does patient behavior have on the proactivity of physicians' online knowledge sharing on digital platforms? The primary objective of this study is to investigate the impact of patient behavior on physicians' proactive engagement in online knowledge sharing and to uncover the underlying mechanisms. Drawing on Parker et al's [14] proactive motivation model and empowerment theory [18], we propose a comprehensive framework and research hypotheses, outlined as follows.

Proactive Motivation Model

Most prior research relies on 2 primary perspectives to explain extra-role or prosocial behaviors, such as knowledge sharing. The first is social exchange theory, which posits that these behaviors stem from individuals weighing perceived benefits against costs [3,8-10]. However, the assumption that individuals act rationally and can accurately calculate benefits and costs is problematic, particularly given the uncertainties involved in engaging with digital platforms. Although insightful, this approach also dismisses the possibility of irrational considerations. Alternatively, basic psychological needs theory [19], rooted in self-determination theory, identifies 3 fundamental psychological needs: autonomy, competence, and relatedness. The satisfaction of these needs is thought to significantly influence overall motivation and behavior. Studies following this line of reasoning adopt a self-centric approach to explore the intrinsic factors (eg, sense of self-worth and enjoyment of helping others [5,6]) and extrinsic factors (eg, social and economic returns [11,12]) that drive extra-role or prosocial behaviors. Although useful, this approach may overlook or underemphasize other perspectives, such as altruism. Additionally, while it effectively identifies a broad range of

XSL•FO
RenderX

motivational sources, it does not fully capture the motivational states that drive action.

Digital platforms offer physicians greater autonomy and incentives to explore new ways of engaging with and supporting patients [7,13]. However, participation on these platforms is voluntary, effortful, and uncertain in terms of the promised returns. What motivates physicians to overcome the resistance and challenges posed by strict professional and organizational expectations?

Consistent with the argument made by Mittal et al [16] that prosocial behavior is inherently proactive, physicians' online knowledge sharing reflects their proactivity. This involves self-initiated efforts aimed at improving patients' health outcomes and ensuring long-term effectiveness. According to Parker et al's [14] framework, individual proactive behavior can be predicted by 3 proximal proactive motivational states, namely, "can do," "reason to," and "energized to" motivation. Can-do motivation pertains to individuals' self-assessment of their ability to exhibit proactive behavior, or self-efficacy [6]. Reason-to motivation refers to individuals' intrinsic drive to perceive a compelling reason to act proactively, often linked to a sense of purpose or meaning [20]. Energized-to motivation involves the activation of positive affect that enhances proactive tendencies [16].

Additionally, the impact of proactive motivational states on proactive behavior depends on individual differences and contextual factors. Notably, proactive personality, defined as an individual's disposition to engage in active role orientations by initiating change and influencing the environment, is a key individual difference [14]. Furthermore, contextual factors, such as the initiative climate [21] and leadership [22], are thought to shape or amplify proactivity over time.

This study proposes these 3 proactive motivational pathways to better understand physicians' proactive behavior in the context of online knowledge sharing. To explore the underlying mechanisms driving physicians' proactivity on digital platforms, the study focuses on knowledge-sharing self-efficacy, sharing meaning, and positive professional affect as the 3 proactive motivational states for physicians. Compared with existing literature, Parker et al's [14] framework provides a relevant and novel perspective for understanding physicians' online knowledge sharing.

The digital platforms under examination include 2 distinct sets of contextual factors. One involves the incentive structures within the platforms, such as ranking systems or monetization features, which encourage physicians to contribute high-quality content—a dynamic driven by platform extrinsic rewards. The other pertains to the influence of patients, who act as significant social agents, representing a key contextual factor in this context [13,23]. Subsequent sections will explore the contextual impact of patients from an empowerment perspective.

Reverse Empowerment From Patients to Physicians

Prior research has applied empowerment theory to explain the contextual influences on proactive behavior [17,24]. Empowerment refers to a process that enables individuals to gain power [25]. Traditionally, it has been viewed as the delegation of authority and decision-making power from higher- to lower-level members within an organization [26]. In a traditional health care setting, the exchange of health solutions between physicians and patients has predominantly involved unidirectional information flows from health care providers to recipients [27,28], following a paternalistic approach [29]. As a result, most research assumes that patients are empowered [30], and among the 2 key user groups of digital platforms—patients and physicians—patients have received the most attention [31-33]. Patients have become increasingly empowered through self-diagnosis [34], peer support [35], and the use of artificial intelligence to enhance health care quality [36,37]. By contrast, the perspectives of health care service providers, particularly physicians, have received limited attention [38,39]. There is a growing need to incorporate health care professionals' views on the role of patient empowerment in digital platforms [40].

In the digital context, the traditional patient-physician relationship and paternalistic approach have been both challenged and reshaped [29]. As patients gain greater access to health information and become more actively involved in health decision-making, physicians face new challenges arising from emerging technologies and must explore their evolving roles as medical experts [41]. Therefore, it is essential to recognize and investigate the active role of patients to foster a more cooperative and effective patient-physician relationship [42].

In the marketing discipline, reverse empowerment—where service receivers empower providers—has received empirical support [18]. As customers have first-hand experience with products and services, their interactions with frontline staff can influence employees' creativity in decision-making. A previous study identified 4 activities that can empower service providers: emphasizing the value of the information provided by service receivers, involving them in decision-making processes, expressing confidence in their abilities, and granting them increased autonomy during service interactions [18]. Such customer-initiated support has been shown to stimulate proactive behaviors in service providers. Reverse empowerment has been found to foster creativity [18] and career advancement [43] among service providers, playing a pivotal role in influencing service performance [44].

In the context of digital platforms, patients' online engagement similarly influences providers' actions [7]. For example, patients' demand for medical advice and information can serve as a primary motivator for physician engagement, while patient behaviors—such as one-on-one chats, information seeking, advocacy, and feedback—significantly impact physicians' actions. Therefore, we consider patient behaviors a critical contextual factor in examining their influence on physicians' online knowledge sharing. Similar to customer-empowering behaviors, patient-empowering behaviors are defined as actions taken by patients that inspire physicians and enable them to take steps toward achieving desired outcomes on digital platforms.

XSL•FO
RenderX

Research Model and Hypotheses

The Impact of Perceived Patient-Empowering Behaviors on Physicians' Online Knowledge Sharing

Figure 1 illustrates the proposed research model based on the above literature discussion. Empowering behaviors, in general, serve as positive responses that motivate the replication of specific actions [22]. Therefore, when physicians' online knowledge-sharing behaviors are reinforced by positive feedback from patients, they are more likely to sustain these behaviors, leading to a cycle of continuous and beneficial online knowledge-sharing practices.

Figure 1. Research model and hypotheses.

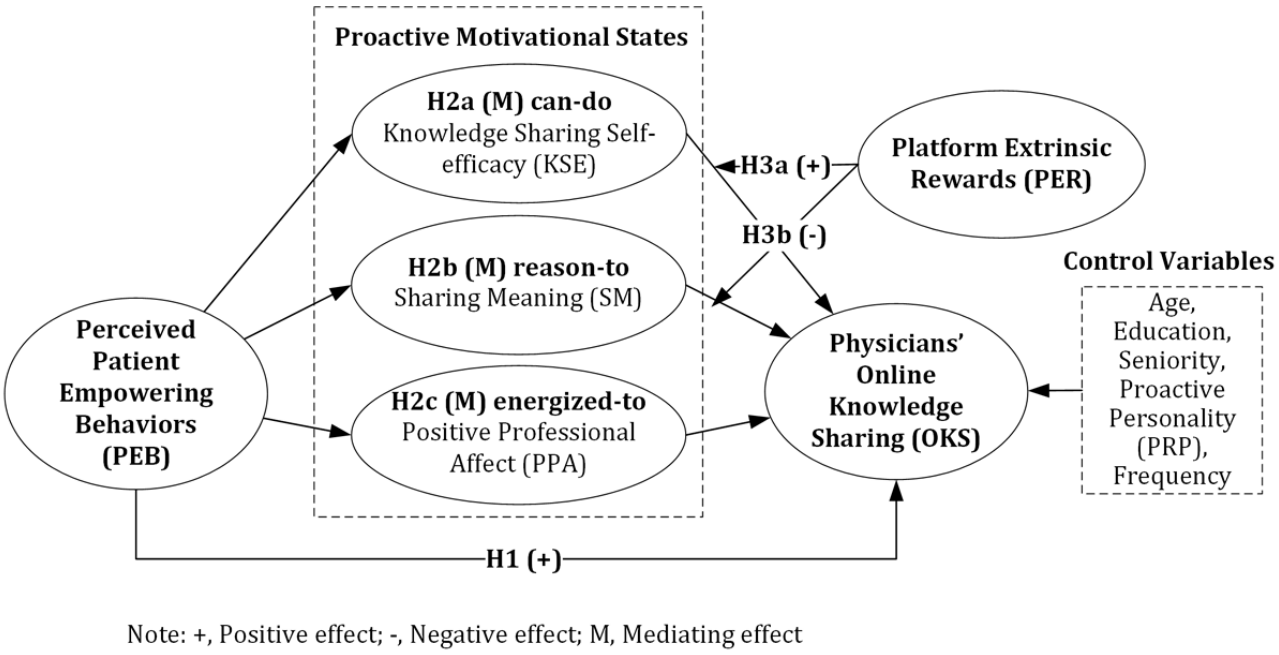

Drawing from reverse empowerment research in marketing [18], we identify 4 distinct online actions enacted by patients as empowering behaviors: highlighting the usefulness of physicians' information, involving physicians in the decision-making process, expressing confidence in physicians' abilities, and granting physicians autonomy in knowledge sharing. First, usefulness refers to the perceived value that patients attribute to online health information in addressing their health-related issues and queries. Prior research has highlighted the positive impact of perceived usefulness on knowledge-sharing behavior within virtual communities [45]. Second, patients increasingly view digital platforms as valuable tools for decision-making, largely due to the provision of unbiased information [46]. Effective assistance from physicians in decision-making requires patients to share their concerns and health status. Third, patients are more likely to share their health information with physicians they trust and perceive as competent [47]. The perception of being trusted enhances a physician's confidence and interest in online knowledge sharing, thereby fostering sustained engagement from physicians [48,49]. Lastly, online knowledge sharing is not limited to physicians' domain expertise as solution providers. As Mesko et al [41] suggested, "the characteristics of a physician-as-idol could shift from self-confident to curious, from rule follower to creative, and from lone hero to team worker." Encouragement from patients can stimulate physicians to explore and become more creative. Therefore, given the pivotal role of patients as key social entities on digital platforms, their empowering behaviors are likely to stimulate physicians' participation in online knowledge sharing. We propose the following:

- Hypothesis 1: Perceived patient-empowering behaviors are positively associated with p
  hysicians' online knowledge sharing.

Mediating Role of Proactive Motivational States

Digital platforms can flatten organizational hierarchies and transcend geographic boundaries, enabling physicians of varying ranks and locations to engage more actively. While patient-empowering behaviors enhance the likelihood of physicians' participation in online knowledge sharing, the unique nature of interactions on these platforms requires the identification of contingencies that may influence this relationship. Unlike traditional face-to-face interactions, digital platform exchanges are not constrained by treatment relationships, and there is often an abundance of information available about both parties. The perception of physicians as valuable or helpful is shaped not only by their qualifications and professional experience but also by their online content and communication with patients [8,13]. Therefore, for patient-empowering behaviors to drive proactive knowledge sharing, they must instigate internal transformations in physicians.

Anchored in the proactive motivation model [14], this study defines 3 proactive motivational states that drive physicians' online knowledge sharing: can-do motivation (knowledge-sharing self-efficacy), reason-to motivation (sharing

XSL•FO
RenderX

meaning), and energized-to motivation (positive professional affect). These motivational states are nurtured by patient-empowering behaviors, which in turn influence physicians' actions on digital platforms. The framework highlights not only how physicians feel empowered but also how they transform this empowerment into proactive behaviors.

First, knowledge-sharing self-efficacy, as a form of can-do motivation, refers to physicians' confidence in their professional abilities and their effectiveness in communicating online to share valuable knowledge with patients [14]. Patient-empowering behaviors, such as demonstrating patients' confidence in physicians' capabilities during service delivery, enable physicians to perceive themselves as possessing high levels of expertise, skill, and competence. This positive feedback from patients contributes to an increase in physicians' self-efficacy in knowledge sharing. Research shows that individuals with higher levels of knowledge-sharing self-efficacy are more likely to offer valuable advice and support to others [50]. This is because a strong sense of efficacy fosters a heightened sense of responsibility to use their expertise for the benefit of others, as demonstrated in studies on self-efficacy and prosocial behavior [51]. Therefore, we hypothesize:

- Hypothesis 2a: Knowledge-sharing self-efficacy, as can-do motivation, mediates the relationship between perceived patient-empowering behaviors and physicians' online knowledge sharing.

Second, sharing meaning, as a form of reason-to motivation, reflects the significance and value that physicians attribute to their online behavior [14]. Physicians who perceive greater meaning in sharing knowledge are likely to experience increased self-worth and fulfillment, which can enhance the quality of their service provision [5,9]. Individuals who find meaning in their actions are more inclined to engage in behaviors that reinforce this positive self-perception [52], such as sharing knowledge with others. Moreover, when patient-empowering behaviors highlight the value and effectiveness of physicians' contributions, it reinforces the importance of their role, deepening their commitment to knowledge sharing. This heightened sense of significance not only makes physicians feel that their contributions are meaningful but also fosters a sense of responsibility and enthusiasm to continue their online knowledge sharing activities. Hence, we hypothesize the following:

- Hypothesis 2b: Sharing meaning, as reason-to motivation, mediates the relationship between perceived patient-empowering behaviors and physicians' online knowledge sharing.

Lastly, as energized-to motivation, positive professional affect refers to the positive emotions and attitudes physicians associate with their professional experiences [53]. As participation in digital platforms is voluntary and can sometimes be perceived as a distraction from core responsibilities, physicians need to experience positive emotions related to their professional role to remain engaged. Positive professional affect also represents the internalization of online knowledge sharing into physicians' professional identities, helping reconcile the tension between online presence and traditional professional expectations [54]. This factor can influence physicians' proactivity by shaping their professional aspirations, aligning their actions with their core professional values and goals [21]. When physicians experience a strong sense of positive affect toward their profession, they are more likely to engage in proactive behaviors. These activities reinforce their positive self-concept and contribute to their professional fulfillment, thereby energizing their motivation to engage in online knowledge sharing. Hence, we hypothesize the following:

- Hypothesis 2c: Positive professional affect, as energized-to motivation, mediates the relationship between perceived patient-empowering behaviors and physicians' online knowledge sharing.

Moderating Role of Extrinsic Rewards

Platform extrinsic rewards serve to validate individuals' capabilities and reinforce their expectations regarding online knowledge sharing [5,9,11,23]. Physicians with high self-efficacy are confident in their ability to share knowledge online. Consequently, in the presence of favorable platform extrinsic rewards, they are more likely to be motivated to engage in online knowledge sharing and confident in the outcomes of their efforts, thus establishing a cycle of positive reinforcement through a self-fulfilling prophecy [55]. Previous research has indicated that rewards effectively moderate the impact of self-efficacy on both effort [56] and performance [57]. Hence, we hypothesize the following:

- Hypothesis 3a: Platform extrinsic rewards positively moderate the relationship between knowledge-sharing self-efficacy and physicians' online knowledge sharing.

According to the motivation crowding theory, platform-based extrinsic incentives have the potential to displace an individual's intrinsic motivation to act based on their own internal needs [54]. Similarly, cognitive evaluation theory posits that when external rewards are perceived as controlling, they can undermine intrinsic motivation by shifting the perceived locus of causality from internal to external. In the context of physicians' online knowledge sharing, platform extrinsic rewards can diminish contributors' intrinsic drive, thereby impeding their participation. This aligns with studies suggesting that external interventions can negatively moderate the connection between intrinsic benefits and the intention to contribute to online feedback systems [58]. As sharing meaning reflects a sense of self-worth, it represents an intrinsic factor influencing physicians' behaviors [14]. Hence, we hypothesize the following:

- Hypothesis 3b: Platform extrinsic rewards negatively moderate the relationship between sharing meaning and physicians' online knowledge sharing.

Lastly, the impact of platform extrinsic rewards on moderating the relationship between positive professional affect and online knowledge sharing may yield ambivalent outcomes. On the one hand, platform extrinsic rewards tied to the quality and quantity of physicians' online knowledge sharing could empower physicians with high levels of positive affect in their occupations, motivating them to increase their contributions and achieve their desired occupational expectations and goals

XSL•FO
RenderX

[21]. On the other hand, similar to the substitution effect of extrinsic motivation on intrinsic motivation [59], the pursuit of extrinsic rewards could potentially diminish the intrinsic drive derived from the positive affect inherent to physicians' occupation. The moderating effect of platform extrinsic rewards is likely contingent on the magnitude of the rewards. However, due to limited knowledge of the actual extrinsic rewards experienced by physicians on digital platforms and the ambiguous evidence regarding their moderating effect, we refrain from formulating a hypothesis.

Study Objectives

The main objective of this research is to explore the influence of patient-empowering behaviors on physicians' online knowledge sharing, examine the mediating role of proactive motivational states, namely, can-do, reason-to, and energized-to motivation in this relationship, and assess the moderating effect of platform extrinsic rewards on the connection between proactive motivational states and physicians' online knowledge sharing. To empirically test our theoretical framework, we collected data via a survey targeting physicians engaged in online knowledge sharing and analyzed the data using structural equation modeling. The results provide strong support for the proposed research model.

Methods

Ethical Considerations

The study was conducted in accordance with the Institutional Review Board of the School of Management at Harbin Institute of Technology. All participants were thoroughly informed about the nature of the study and the voluntary nature of their participation in the web-based survey. They were assured that all data collected would be treated with strict confidentiality and managed in accordance with established ethical standards. The survey was anonymous, with no personal or identifying information being collected or accessible to the researchers.

Data Collection

This survey was conducted in China. Three companies, with physicians as their clients, were approached and agreed to facilitate data collection. In these companies, salespersons maintain direct service and business relationships with physicians. Approximately 50 medical salespersons were randomly selected from these companies to support the data collection.

The questionnaire was designed in an electronic format using the WJX platform. Salespersons distributed the questionnaire to physicians via WeChat (Tencent Holdings Limited) or provided it during offline visits. Physicians could access the questionnaire by scanning a QR code or clicking on a forwarded link through WeChat on their mobile phones. Additionally, a snowball sampling method was used, where participating salespersons and physicians were asked to forward the questionnaire to other physicians they knew.

We included a screening question at the beginning of the questionnaire to identify physicians with no experience in online knowledge sharing. Physicians without this experience were excluded from the sample. To ensure anonymity, no personal or identifying information was collected from participants. Additionally, no incentive mechanism was offered to the salespersons or respondents. To detect potential errors associated with the online survey, we included reverse-coded questions and ensured that participation was entirely voluntary and anonymous. To prevent duplication, we collected a WeChat ID or phone number as an identifier, ensuring that each physician could complete the questionnaire only once.

A total of 571 responses were collected over 1 month, including 327 physicians with experience in online knowledge sharing and 244 without such experience, as indicated by the screening question. Invalid responses—such as those with no response variance, incomplete answers, or contradictory information—were examined and filtered out, leaving 257 valid responses for data analysis and hypothesis testing. More detailed information on data collection issues can be found in Multimedia Appendix 1 [60].

Measurement Instrument

To validate the research model, a survey study was conducted using items adapted from previous research to ensure both validity and contextual relevance. A 5-point Likert scale was used, with responses ranging from 1=strongly disagree to 5=strongly agree. Physicians' online knowledge sharing was measured with 4 items adapted from Zhang et al [5]. The scale for patient-empowering behaviors was derived from Dong et al's [18] customer-empowering behaviors scale, retaining key aspects of their original framework. Four dimensions, each with 2 items, were used to measure perceived patient-empowering behaviors, except for the dimension of autonomy. One autonomy item ("Customer allowed me to make important decisions to satisfy his/her needs") was excluded from the research context, as physicians do not make decisions for patients but instead provide advice. In total, 7 items were used to assess perceived patient-empowering behaviors. The 3 dimensions of proactive motivational states were measured using 3 items for knowledge-sharing self-efficacy, based on Lin et al [50]; 3 items for sharing meaning, adapted from Spreitzer [61]; and a measurement for positive professional affect, adapted from Watson et al [62]. Platform extrinsic rewards were assessed with 3 items from Yang and Lai [63].

To examine how physicians' characteristics may influence their online behaviors, control variables were considered. Proactivity personality was measured using an adaptation of the scale from Parker et al [14]. Nominal scales were used to assess seniority, age, frequency, and education level.

To ensure the validity of the measurement during the translation of the instrument from English to Chinese, we used a back-translation method. Three experts in the field of information systems reviewed the instrument for clarity, question wording, ease of understanding, logical consistency, item sequence, and contextual relevance, suggesting minor modifications to the wording and item order. A pilot study involving 40 physicians was then conducted using the revised questionnaire, which enhanced clarity and reduced ambiguities. We recorded the time taken by physicians to complete the survey as a criterion for evaluating response quality. Items with

XSL•FO
RenderX

unsatisfactory loadings were removed as a result. Detailed information about the survey and its items is provided in Multimedia Appendix 2.

Statistical Analysis Methods

This study utilizes SPSS (version 26; IBM Corp.) and SmartPLS (version 3.0; SmartPLS GmbH) software to conduct multiple stages of data analysis.

In the first stage, we conducted a measurement model examination to assess the validity and reliability of the constructs in our study. First, we checked for potential selection bias by comparing early and late respondents across all variables. Second, to examine common method bias, we performed the Harman 1-factor test using SPSS [64]. Third, confirmatory factor analysis was conducted to assess the reliability and validity of the research model. All factor loadings exceeded the 0.70 threshold, indicating good convergent validity. Composite reliability and Cronbach α values were above 0.70, and the average variance extracted values surpassed 0.50, demonstrating strong internal consistency and reliability. Additionally, the square root of each average variance extracted exceeded the correlations with other variables, confirming discriminant validity. Finally, collinearity among predictors was assessed, with variance inflation factor values below the 3.3 threshold, indicating no multicollinearity issues [65].

In the second stage, partial least squares-structural equation modeling was used to examine the structural model, including direct, moderation, and mediation effects, in order to test the proposed hypotheses [66]. Partial least squares-structural equation modeling, a robust technique based on bootstrapping, was selected for its flexibility in handling less stringent distributional assumptions and smaller sample sizes compared with other methods. The structural model was assessed through path estimations and the model's explanatory power. A bootstrapping resampling approach with 5000 resamples was used to evaluate the significance of the paths, providing the explained variance of endogenous variables (R^2), standardized path coefficients (β), and the significance of these estimates (P value). Additionally, to rigorously test the mediation effects of proactive motivational states, we used bootstrapping, which does not require the assumption of normality for the sampling distribution. Finally, interaction terms were used to statistically test the moderating effects. The moderating role was visually depicted through a multigroup analysis. The data set was divided into 2 groups based on the mean value of platform extrinsic rewards, creating high- and low-reward subsets. The difference in slopes between these 2 groups demonstrated how platform extrinsic rewards moderated the relationship.

Results

Descriptive Analysis

Table 1 presents the demographic characteristics of the respondents. The sample (N=257) showed a gender balance, with 103 male physicians (40.1%) and 154 female physicians (59.9%). The majority of respondents were aged between 31 and 40 years (132/257, 51.4%). Most platforms provided payment consultation services, ranking systems, or gift incentives to encourage physicians' online engagement. The frequency with which respondents used these platforms over the past 3 months varied among physicians.

XSL•FO
RenderX

Table 1. Demographic characteristics of sample participants (N=257).

Demographic variables	Values, n (%)
Gender	
Male	103 (40.1)
Female	154 (59.9)
Age (years)	
<30	37 (14.4)
31-40	132 (51.4)
41-50	65 (25.3)
51-60	20 (7.8)
>60	3 (1.2)
Education level	
Bachelor's degree	109 (42.4)
Master's degree	98 (38.1)
PhD degree	50 (19.5)
Seniority	
Resident physician	45 (17.5)
Attending physician	105 (40.9)
Deputy chief physician	76 (29.6)
Chief physician	31 (12.1)
Frequency of online knowledge sharing in the last 3 months	
Once a day	25 (9.7)
Once every 3 days	30 (11.7)
Once a week	50 (19.5)
Once every 2 weeks	23 (8.9)
Once a month	129 (50.2)
Digital platforms (multichoice)	
HaoDF	80 (31.1)
Dingxiang Doctor	117 (45.5)
Weiyi	34 (13.2)
Chunyu Doctor	37 (14.4)
PingAn Doctor	43 (16.7)
Others	90 (35.0)

Measurement Model

First, the results of the selection bias test indicate no significant differences between early and late respondents (online knowledge sharing, dependent variable, P=.92). Second, the results of the Harman 1-factor test revealed that the first and largest factor of the unrotated solution explained 41.3% of the total variance, which is below the 50% threshold, suggesting no evidence of common method bias in this study. Third, as shown in Tables 2 and 3, the results demonstrated good convergent validity, strong internal consistency, reliability, and discriminant validity. The variance inflation factor values ranged from 1.064 to 2.167, all below the threshold of 3.3, indicating no multicollinearity issues in this study.

XSL•FO
RenderX

Table 2. Construct reliability and convergent validity.

Constructs	Factor loadings	Composite reliability	Average variance extracted	Cronbach α
OKS[a]				
OKS1	0.900	0.936	0.786	0.909
OKS2	0.864	N/A[b]	N/A	N/A
OKS3	0.892	N/A	N/A	N/A
OKS4	0.889	N/A	N/A	N/A
PEB[c]				
PEB1	0.786	0.920	0.620	0.898
PEB2	0.810	N/A	N/A	N/A
PEB3	0.773	N/A	N/A	N/A
PEB4	0.766	N/A	N/A	N/A
PEB5	0.793	N/A	N/A	N/A
PEB6	0.795	N/A	N/A	N/A
PEB7	0.790	N/A	N/A	N/A
KSE[d]				
KSE1	0.812	0871	0.693	0.780
KSE2	0.867	N/A	N/A	N/A
KSE3	0.817	N/A	N/A	N/A
SM[e]				
SM1	0.890	0.921	0.795	0.871
SM2	0.886	N/A	N/A	N/A
SM3	0.899	N/A	N/A	N/A
PPA[f]				
PPA1	0.898	0.973	0.798	0.968
PPA2	0.871	N/A	N/A	N/A
PPA3	0.916	N/A	N/A	N/A
PPA4	0.884	N/A	N/A	N/A
PPA5	0.892	N/A	N/A	N/A
PPA6	0.918	N/A	N/A	N/A
PPA7	0.924	N/A	N/A	N/A
PPA8	0.898	N/A	N/A	N/A
PPA10	0.838	N/A	N/A	N/A
PER[g]				
PER1	0.879	0.911	0.773	0.853
PER2	0.880	N/A	N/A	N/A
PER3	0.879	N/A	N/A	N/A
PRP[h]				
PRP1	0.916	0.890	0.731	0.839
PRP2	0.906	N/A	N/A	N/A
PRP4	0.733	N/A	N/A	N/A

[a]OKS: online knowledge sharing.

XSL•FO
RenderX

[b]N/A: not applicable.
[c]PEB: patient-empowering behaviors.
[d]KSE: knowledge sharing self-efficacy.
[e]SM: sharing meaning.
[f]PPA: positive professional affect.
[g]PER: platform extrinsic rewards.
[h]PRP: proactive personality.

Table 3. Discriminant validity[a].

Constructs	OKS[b]	PEB[c]	SM[d]	KSE[e]	PPA[f]	PER[g]	PRP[h]
OKS	0.887	__[i]	—	—	—	—	—
PEB	0.471[j]	0.788	—	—	—	—	—
SM	0.635[j]	0.427[j]	0.892	—	—	—	—
KSE	0.473[j]	0.533[j]	0.392[j]	0.833	—	—	—
PPA	0.554[j]	0.526[j]	0.430[j]	0.474[j]	0.893	—	—
PER	0.635[j]	0.415[j]	0.505[j]	0.379[j]	0.540[j]	0.879	—
PRP	0.086	0.110	0.07	0.074	–0.045	–0.036	0.855

[a]Diagonal numbers are average variance extracted square root, others are correlation coefficients.
[b]OKS: online knowledge sharing.
[c]PEB: patient-empowering behaviors.
[d]SM: sharing meaning.
[e]KSE: knowledge sharing self-efficacy.
[f]PPA: positive professional affect.
[g]PER: platform extrinsic rewards.
[h]PRP: proactive personality.
[i]Not applicable.
[j]$P<.01$.

Structural Model

As illustrated in model A of Figure 2, the direct effect demonstrates that perceived patient-empowering behaviors are positively associated with online knowledge sharing (β=0.27, $P<.001$). Model A explains 51% of the variance in online knowledge sharing, with the direct effect contributing an additional 6.2% variance compared with the baseline model, which includes only the control variables. Therefore, hypothesis 1 is supported.

Regarding the control variables, physicians' age and education level are negatively correlated with online knowledge sharing, whereas the frequency of participation and platform extrinsic rewards show positive correlations with online knowledge sharing.

Mediation Effect

Model B, as delineated in Figure 2, demonstrates a positive correlation between perceived patient-empowering behaviors and the 3 proactive motivational states: knowledge-sharing self-efficacy (β=0.53, $P<.001$), sharing meaning (β=0.43, $P<.001$), and positive professional affect (β=0.53, $P<.001$). Furthermore, these 3 proactive motivational states—knowledge-sharing self-efficacy, sharing meaning, and positive professional affect—are positively associated with online knowledge sharing among physicians, with standardized path coefficients of 0.11 ($P=.03$), 0.29 ($P<.001$), and 0.20 ($P=.001$), respectively. The explanatory power of this model accounts for 61% of the variance in online knowledge sharing, representing a 16% increase in variance explained over the baseline model that included only control variables.

XSL•FO
RenderX

Figure 2. Analysis results of the structural model. (A) Model with direct effects; (B) model with mediating effects.

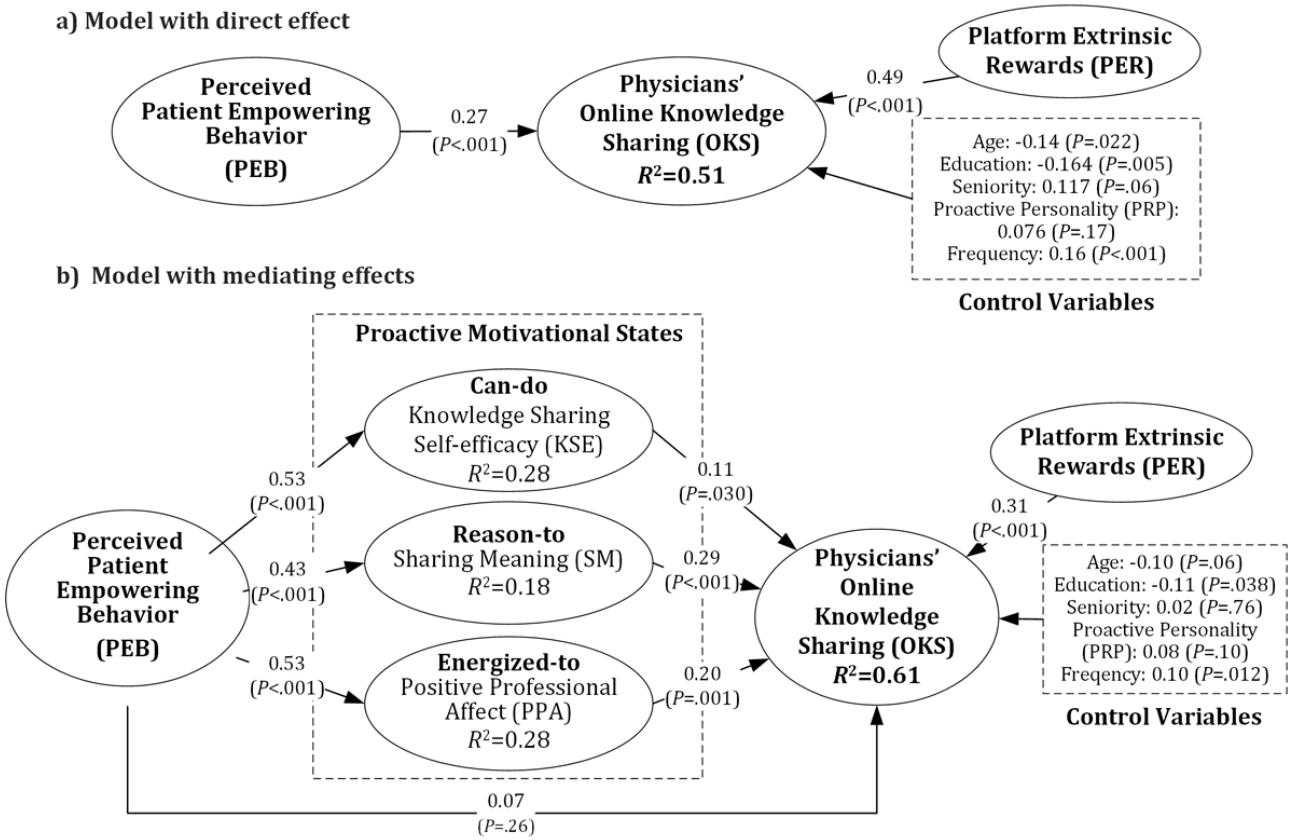

The application of this nonparametric resampling method reveals significant findings for the entire indirect effect (effect=0.29, *P*<.001), while the direct effect is statistically insignificant (effect=0.07, *P*=.26). These results underscore the complete mediation role of proactive motivational states in the relationship between physicians' perceived patient-empowering behaviors and their engagement in online knowledge sharing. Additionally, the partial mediation effects are significant for knowledge-sharing self-efficacy (effect=0.06, *P*=.04), sharing meaning (effect=0.12, *P*<.001), and positive professional affect (effect=0.10, *P*=.003). These findings indicate that each of these motivational states individually mediates the relationship between physicians' perceived patient-empowering behaviors and their online knowledge sharing to some extent, thereby supporting hypothesis 2a-c (Table 4).

Table 4. Results of testing mediation effects for proactive motivational states.

Mediation effect	Effect	*P* value	Bias-corrected 95% CI
Total effect	0.36	<.001	0.23 to 0.48
Total indirect effect	0.29	<.001	0.19 to 0.39
Specific indirect effect			
Knowledge sharing self-efficacy	0.06	.04	0.01 to 0.11
Sharing meaning	0.12	<.001	0.07 to 0.19
Positive professional affect	0.10	.003	0.04 to 0.16
Direct effect	0.07	.26	–0.05 to 0.20

Moderating Effect

The analysis presented in Table 5 evaluates the moderating effect of platform extrinsic rewards. A single significant finding emerges, specifically regarding the interaction between sharing meaning and online knowledge sharing (β=–0.13, *P*=.001). This result supports hypothesis 3b, while no evidence was found to support hypothesis 3a.

XSL•FO
RenderX

Table 5. The moderating effect of platform extrinsic rewards.

Path	Coefficient (β)	*P* value
PER[a] × KSE[b]→OKS[c]	–0.03	.57
PER × SM[d]→OKS	–0.13	.001

[a]PER: platform extrinsic rewards.
[b]KSE: knowledge sharing self-efficacy.
[c]OKS: online knowledge sharing.
[d]SM: sharing meaning.

Figure 3 graphically illustrates the moderating role of platform extrinsic rewards in the relationship. The figure highlights a nuanced dynamic in how the level of platform extrinsic rewards affects physicians' online knowledge sharing. Specifically, with a higher level of platform extrinsic rewards (depicted by the dotted line in Figure 3), increases in sharing meaning correspond to a more gradual change in physicians' online knowledge sharing. Conversely, in scenarios with lower platform extrinsic rewards (depicted by a solid line in Figure 3), increases in sharing meaning result in a more pronounced rise in physicians' online knowledge sharing.

Figure 3. Interaction between platform extrinsic rewards and sharing meaning in influencing physicians' online knowledge sharing.

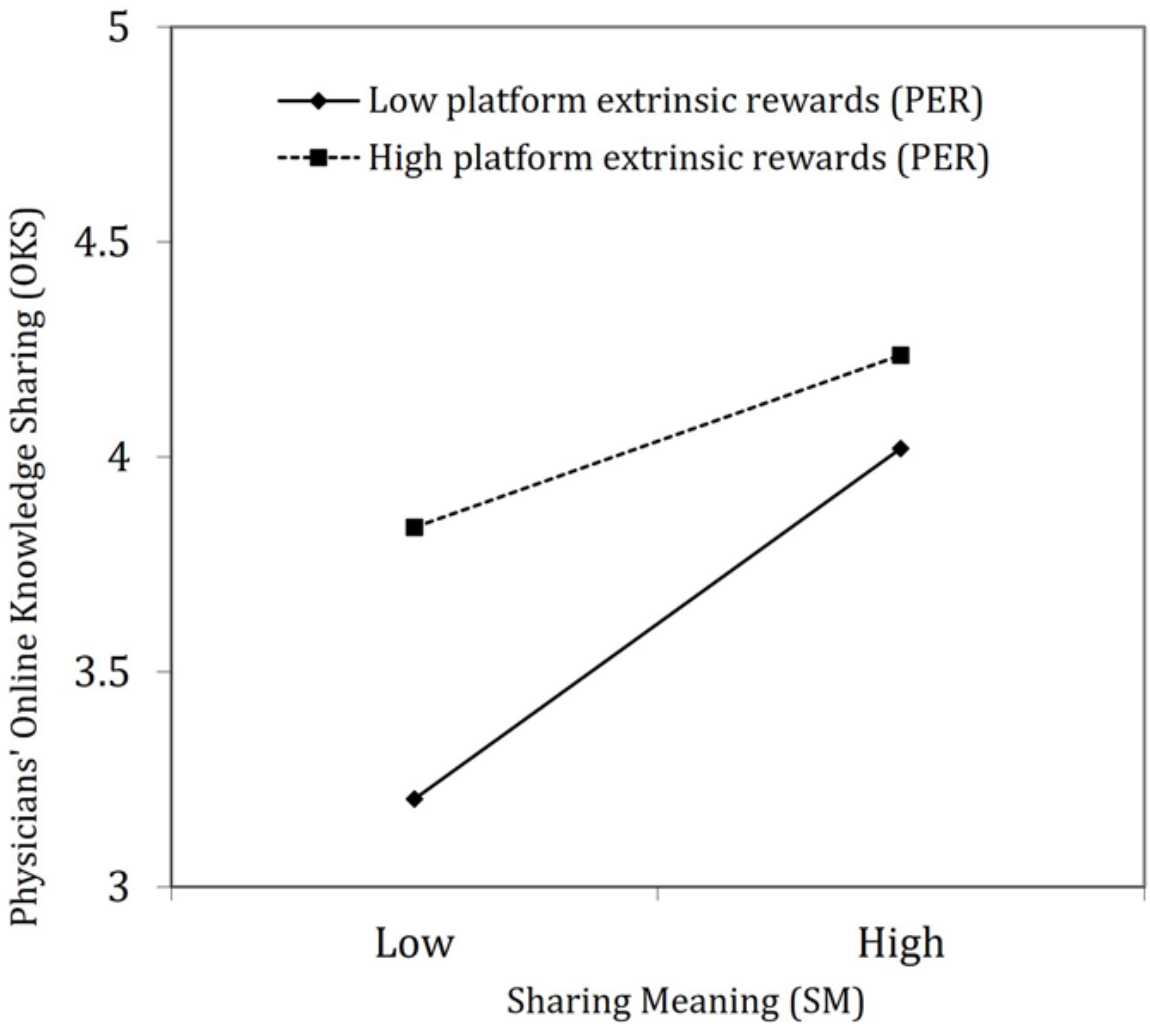

Discussion

Principal Findings

This study highlights how patient behaviors can enhance and empower physicians' online knowledge sharing, mediated by proactive motivational states, including knowledge-sharing self-efficacy (can-do motivation), sharing meaning (reason-to motivation), and positive professional affect (energized-to motivation). Furthermore, platform extrinsic rewards, as a contextual factor, negatively moderate the relationship between sharing meaning and online knowledge sharing.

XSL•FO
RenderX

Specifically, first, our empirical evidence supports the phenomenon of reverse empowerment from patients to physicians within digital platforms. Encounters with patient-empowering behaviors across 4 dimensions make physicians more inclined to engage in online knowledge sharing. This relationship aligns with findings by Dong et al [18] and Zhang et al [13], although, in this context, the behavior of the "service provider" (knowledge contributors) is oriented toward a future exchange.

Second, our results further validate the mediating role of proactive motivational states, triggered by perceived patient-empowering behaviors. Notably, these proactive motivational states fully mediate the influence of perceived patient-empowering behaviors on online knowledge sharing, especially when considering the potential effects of platform extrinsic rewards. Each motivational state—knowledge-sharing self-efficacy, sharing meaning, and positive professional affect—partially contributes to the overall mediating effect. These results confirm the viability of conceptualizing physicians' online knowledge sharing as an aspect of proactive behavior, presenting proactivity as an alternative framework to explain physicians' engagement in online knowledge sharing [17]. The considerable mediation effect exerted by proactive motivational states underscores the desired nature of social dynamics between patients and physicians on digital platforms. It highlights the impactful role of patient-driven empowering actions in enhancing physicians' perceived sharing meaning, knowledge-sharing self-efficacy, and positive professional affect, thereby amplifying knowledge dissemination.

Third, our findings align with the negative moderating effect of platform extrinsic rewards on the link between sharing meaning and online knowledge sharing, consistent with existing literature suggesting that extrinsic motivators may undermine intrinsic motivation [59]. The absence of a moderating effect of platform extrinsic rewards on the impact of knowledge-sharing self-efficacy could be attributed to the predominantly high self-efficacy levels among the physicians surveyed. This suggests a possible limit to the effectiveness of platform extrinsic rewards, as individuals with already high self-efficacy do not seem to experience further enhancement in their online knowledge-sharing behaviors.

Regarding control variables, the significant positive correlation between platform extrinsic rewards and online knowledge sharing corroborates previous studies within the knowledge-sharing motivation theory domain [5,9]. The demographic characteristics of participating physicians indicate that digital platforms tend to attract younger and less experienced physicians. The pursuit of professional advancement, the acquisition of broader experience, and the availability of such platforms might serve as key drivers for platform engagement.

Theoretical Implications

Our findings offer 3 important contributions to the existing literature. First, our study enriches the existing empowerment literature by investigating the reverse empowerment process from patients to physicians. Traditional scholarship on empowerment predominantly conceptualizes the empowerment process as emanating from the more authoritative party toward those in subordinate positions, such as empowerment initiatives from organizations [25,61], leaders [67], or service providers [68] directed at employees, subordinates, or service recipients. The digital environment, however, disrupts traditional information hierarchies, endowing previously perceived "lesser-powered" entities with empowerment through the proactive seeking and dissemination of information. While marketing research has started to probe into consumers acting as a prominent source of empowerment to motivate service employees toward taking control of the service process [18], the notion of patients empowering physicians has not garnered sufficient attention. Our investigation stands as a new effort to scrutinize and validate the substantial influence patients wield in elevating physicians on digital platforms, thereby completing the loop of empowerment and augmenting comprehension of the genesis of an empowerment cycle and productive patient-physician dynamics.

Second, this study enhances the eHealth literature by characterizing physicians' online knowledge sharing as proactive conduct. The advent of digital technology presents challenges and opportunities to traditional professional roles and workplace settings. Proactive behavior, epitomized as self-initiated, future-oriented actions aimed at altering one's work environment, roles, or self, is vital for individuals and organizations to navigate the demands of a rapidly evolving milieu [69]. Despite extensive coverage by management researchers aiming to broaden the scope of job performance within the confines of organizational settings [70], the proactive aspect of physicians' online knowledge sharing has largely been overlooked. By applying the proactive behavior paradigm to eHealth research within the ambit of open digital platforms, our study not only deepens the existing understanding but also crucially opens new avenues for future exploration.

Moreover, our examination advances the understanding of proactive behavior within a novel milieu and its contextual determinants. Previous inquiries have scrutinized the contextual variables influencing proactive behaviors within formal organizational structures, such as leadership, work design, and social dynamics [14], which may manifest differently within the context of digital health platforms. Given the reciprocal nature of the physician-patient relationship, social stimuli (perceived patient-empowering behaviors) alongside platform design attributes (platform extrinsic rewards) emerge as catalysts for physicians to adopt proactive stances within digital platforms. By identifying and empirically demonstrating the significance of situational conditions that foster proactive behaviors among physicians in this emergent context, our analysis makes a valuable contribution to the discourse on proactive conduct.

Practical Implications

Considering the increasing popularity and adoption of digital channels by traditional health care service providers, the findings of this study offer valuable insights into constructing and sustaining effective digital platforms. First, it is crucial for platforms to view patients or general participants not merely as information seekers but as key catalysts in the empowerment

XSL•FO
RenderX

cycle. It is essential, therefore, to acknowledge and incentivize their empowering actions, which play a crucial role in sustaining the ongoing engagement of physicians and other experts. For instance, to increase patient-empowering behaviors, managers might enhance patients' ability to affirm the competence of physicians or provide simpler ways for patients to acknowledge and commend other physicians' contributions. Furthermore, platforms should focus on improving the visibility of relevant patient-empowering behaviors to physicians, such as by highlighting the positive outcomes of online consultations.

Second, while the deployment of platform extrinsic rewards is critical, it alone may prove inadequate in catalyzing physicians' proactive engagement. It is advisable for platforms to incorporate proactive emotional states into their metrics for designing, assessing, and refining the efficacy of incentives. Additionally, given that patient-empowering behaviors attenuate the mediating influence of sharing meaning, it becomes necessary for platforms to strike a harmonious balance between patient-empowering behaviors and strategies that promote them.

Lastly, the phenomenon of reverse empowerment may have implications for clinical practices, especially in contexts where physicians engage in regular interactions with patients. There is an opportunity for patients to collaborate with physicians during the provision of clinical services. Consequently, clinical institutions should strive to raise patients' awareness of how their empowering behavior could lead to more proactive services from physicians, thereby fostering a more dynamic and reciprocal patient-physician interaction.

Limitations and Future Directions

Although this research provides valuable insights into both theoretical and practical domains, several limitations underscore the need for further scholarly exploration. First, the scope of this study is confined to third-party open platforms, excluding a comprehensive analysis of the various organizational affiliations of physicians. In light of the growing embrace and valuation of digital channels by health care providers, the findings of this study may have broader implications for physicians' professional performance. Future research should investigate organizational platforms and explore potential interactions between online participation and organizational factors in driving physicians' proactive efforts.

Second, to attract domain expertise and encourage continuous engagement, platforms often compete by designing various incentives. While this study considers extrinsic rewards as perceived by physicians to account for varying exposure levels to different platform incentives, it would be beneficial to expand research to directly examine these platform incentives and their impact on patient-physician interactions.

Third, given the cross-sectional design of our study, the ability to empirically substantiate causality is limited. Nonetheless, our findings support the value of considering proactivity as an alternative mechanism and underscore the significant effect of perceived patient-empowering behaviors. This design allowed us to capture physicians' personal perceptions and focus on the internal mechanisms driving their behavior. While our study sets the stage for understanding these perceptual dynamics, it also paves the way for future research to track dynamic behavioral changes using a longitudinal design, further clarifying the time-based causal effects between physicians' proactive behavior changes and patient behavior.

Conclusions

Drawing upon proactive motivation theory and the concept of the reverse empowerment process, this study conceptualizes physicians' online knowledge sharing as a form of proactive behavior and investigates the underlying mechanisms by identifying contextual factors and empirically validating their impact on physicians' online knowledge sharing. The findings suggest that the reverse empowerment process, specifically perceived patient-empowering behaviors, can facilitate physicians' online knowledge sharing by fostering proactive motivational states such as knowledge-sharing self-efficacy, sharing meaning, and positive professional affect. These motivational states emerge as pivotal mediators, while platform extrinsic rewards may attenuate the positive relationship between sharing meaning and online knowledge sharing. This study makes significant contributions to the existing literature on empowerment, eHealth, and proactive behavior. The findings also offer valuable insights for developing strategies to ensure the sustainability of services in the design and management of digital health platforms.

Acknowledgments

The authors express their gratitude to the editors for their guidance and to the anonymous reviewers for their constructive suggestions. This study was funded by the National Natural Science Foundation of China (grants 72301249, 72125001, 72431004, 72071054, 72293584), the UAE University Start-up Grant (G00003520), the Research Start-up Funding from Zhejiang Sci-Tech University (grant 22092166-Y), and the Chinese Academy of Engineering's Strategic Research and Consulting Project (2024-XBZD-18-05).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 60 KB - jmir_v26i1e59904_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Measurement scales.
[DOCX File , 20 KB - jmir_v26i1e59904_app2.docx]

References

1. Goh JM, Gao G, Agarwal R. The creation of social value: can an online health community reduce rural-urban health disparities? MISQ 2016 Jan 1;40(1):247-263. [doi: 10.25300/MISQ/2016/40.1.11]
2. Cao X, Wang D. The role of online communities in reducing urban–rural health disparities in China. Asso for Info Science & Tech 2018 Feb 27;69(7):890-899. [doi: 10.1002/asi.24013]
3. Yan Z, Wang T, Chen Y, Zhang H. Knowledge sharing in online health communities: a social exchange theory perspective. Information & Management 2016 Jul;53(5):643-653. [doi: 10.1016/j.im.2016.02.001]
4. Wu H, Deng Z. Knowledge collaboration among physicians in online health communities: a transactive memory perspective. International Journal of Information Management 2019 Dec;49:13-33. [doi: 10.1016/j.ijinfomgt.2019.01.003]
5. Zhang X, Liu S, Deng Z, Chen X. Knowledge sharing motivations in online health communities: a comparative study of health professionals and normal users. Computers in Human Behavior 2017 Oct;75:797-810. [doi: 10.1016/j.chb.2017.06.028]
6. Singh JB, Chandwani R, Kumar M. Factors affecting Web 2.0 adoption: exploring the knowledge sharing and knowledge seeking aspects in health care professionals. JKM 2018 Jan 08;22(1):21-43. [doi: 10.1108/jkm-08-2016-0320]
7. Wang L, Yan L, Zhou T, Guo X, Heim GR. Understanding physicians' online-offline behavior dynamics: an empirical study. Information Systems Research 2020 Jun;31(2):537-555. [doi: 10.1287/isre.2019.0901]
8. Guo S, Guo X, Fang Y, Vogel D. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. Journal of Management Information Systems 2017 Aug 17;34(2):487-519. [doi: 10.1080/07421222.2017.1334480]
9. Chen J, Lan Y, Chang Y, Chang P. Exploring doctors' willingness to provide online counseling services: the roles of motivations and costs. Int J Environ Res Public Health 2019 Dec 22;17(1):110 [FREE Full text] [doi: 10.3390/ijerph17010110] [Medline: 31877861]
10. Li J, Tang J, Jiang L, Yen DC, Liu X. Economic success of physicians in the online consultation market: a signaling theory perspective. International Journal of Electronic Commerce 2019 Mar 26;23(2):244-271. [doi: 10.1080/10864415.2018.1564552]
11. Yang H, Ju X. Investigating the influences of motivators on physician contribution behaviors in online health community: offline status as a moderator. 2016 Presented at: 20th Pacific Asia Conference on Information Systems PACIS; June 27 to July 1, 2016; Chiayi, Taiwan p. 11 URL: https://aisel.aisnet.org/pacis2016/11/
12. Zhang X, Guo X, Lai K, Yi W. How does online interactional unfairness matter for patient–doctor relationship quality in online health consultation? The contingencies of professional seniority and disease severity. European Journal of Information Systems 2018 Nov 27;28(3):336-354. [doi: 10.1080/0960085X.2018.1547354]
13. Zhang X, Guo F, Xu T, Li Y. What motivates physicians to share free health information on online health platforms? Information Processing & Management 2020 Mar;57(2):102166. [doi: 10.1016/j.ipm.2019.102166]
14. Parker SK, Bindl UK, Strauss K. Making things happen: a model of proactive motivation. Journal of Management 2010 May 14;36(4):827-856. [doi: 10.1177/0149206310363732]
15. Crant J. Proactive behavior in organizations. Journal of Management 2000;26(3):435-462 [FREE Full text] [doi: 10.1016/s0149-2063(00)00044-1]
16. Mittal S, Sengupta A, Agrawal NM, Gupta S. How prosocial is proactive: developing and validating a scale and process model of knowledge-based proactive helping. Journal of Management & Organization 2018 Mar 21;26(4):625-650. [doi: 10.1017/jmo.2017.80]
17. Kang YJ, Lee JY, Kim H. A psychological empowerment approach to online knowledge sharing. Computers in Human Behavior 2017 Sep;74:175-187. [doi: 10.1016/j.chb.2017.04.039]
18. Dong Y, Liao H, Chuang A, Zhou J, Campbell EM. Fostering employee service creativity: joint effects of customer empowering behaviors and supervisory empowering leadership. J Appl Psychol 2015 Sep;100(5):1364-1380. [doi: 10.1037/a0038969] [Medline: 25774571]
19. Van den Broeck A, Ferris DL, Chang C, Rosen CC. A review of self-determination theory's basic psychological needs at work. Journal of Management 2016 Mar 09;42(5):1195-1229. [doi: 10.1177/0149206316632058]
20. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
21. Hong Y, Liao H, Raub S, Han JH. What it takes to get proactive: an integrative multilevel model of the antecedents of personal initiative. J Appl Psychol 2016 May;101(5):687-701. [doi: 10.1037/apl0000064] [Medline: 26653528]
22. Binyamin G, Brender-Ilan Y. Leaders's language and employee proactivity: enhancing psychological meaningfulness and vitality. European Management Journal 2018 Aug;36(4):463-473. [doi: 10.1016/j.emj.2017.09.004]
23. Wang J, Chiu Y, Yu H, Hsu Y. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: longitudinal study. J Med Internet Res 2017 Dec 21;19(12):e427 [FREE Full text] [doi: 10.2196/jmir.9082] [Medline: 29269344]

XSL•FO
RenderX

24. Kirkman BL, Rosen B. Beyond self-management: antecedents and consequences of team empowerment. AMJ 1999 Feb;42(1):58-74. [doi: 10.5465/256874]
25. Zimmerman MA. Psychological empowerment: issues and illustrations. Am J Community Psychol 1995 Oct;23(5):581-599 [FREE Full text] [doi: 10.1007/BF02506983] [Medline: 8851341]
26. Ertürk A, Vurgun L. Retention of IT professionals: examining the influence of empowerment, social exchange, and trust. Journal of Business Research 2015 Jan;68(1):34-46. [doi: 10.1016/j.jbusres.2014.05.010]
27. Polese F, Tartaglione A, Cavacece Y. Patient empowerment for health care service quality improvements: a value co-creation view. 2016 Presented at: 19th Toulon-Verona International Conference Excellence in Services; September 5-6, 2016; Huelva, Spain URL: https://d1wqtxts1xzle7.cloudfront.net/89356649/Polese-Moretta-Tartaglione-Cavacece-libre.pdf?1659894051=&response-content-disposition=inline%3B+filename%3DPatient_empowerment_for_health
28. McAllister M, Dunn G, Payne K, Davies L, Todd C. Patient empowerment: the need to consider it as a measurable patient-reported outcome for chronic conditions. BMC Health Serv Res 2012 Jun 13;12:157 [FREE Full text] [doi: 10.1186/1472-6963-12-157] [Medline: 22694747]
29. Grünloh C, Myreteg G, Cajander ?, Rexhepi H. "Why do they need to check me?" Patient participation through eHealth and the doctor-patient relationship: qualitative study. J Med Internet Res 2018 Jan 15;20(1):e11 [FREE Full text] [doi: 10.2196/jmir.8444] [Medline: 29335237]
30. François J, Audrain-Pontevia A, Boudhraâ S, Vial S. Assessing the influence of patient empowerment gained through mental health apps on patient trust in the health care provider and patient compliance with the recommended treatment: cross-sectional study. J Med Internet Res 2024 Feb 12;26:e48182 [FREE Full text] [doi: 10.2196/48182] [Medline: 38345851]
31. Grimm M, Link E, Albrecht M, Czerwinski F, Baumann E, Suhr R. Exploring functions and predictors of digital health engagement among German internet users: survey study. J Med Internet Res 2023 Jun 28;25:e44024 [FREE Full text] [doi: 10.2196/44024] [Medline: 37379058]
32. Ploderer B, Rezaei Aghdam A, Burns K. Patient-generated health photos and videos across health and well-being contexts: scoping review. J Med Internet Res 2022 Apr 12;24(4):e28867 [FREE Full text] [doi: 10.2196/28867] [Medline: 35412458]
33. Jiang F, Liu Y, Hu J, Chen X. Understanding health empowerment from the perspective of information processing: questionnaire study. J Med Internet Res 2022 Jan 11;24(1):e27178 [FREE Full text] [doi: 10.2196/27178] [Medline: 35014957]
34. Radionova N, Ög E, Wetzel A, Rieger MA, Preiser C. Impacts of symptom checkers for laypersons' self-diagnosis on physicians in primary care: scoping review. J Med Internet Res 2023 May 29;25:e39219 [FREE Full text] [doi: 10.2196/39219] [Medline: 37247214]
35. Kjærulff EM, Andersen TH, Kingod N, Nexø MA. When people with chronic conditions turn to peers on social media to obtain and share information: systematic review of the implications for relationships with health care professionals. J Med Internet Res 2023 Apr 17;25:e41156 [FREE Full text] [doi: 10.2196/41156] [Medline: 37067874]
36. Lammons W, Silkens M, Hunter J, Shah S, Stavropoulou C. Centering public perceptions on translating AI into clinical practice: patient and public involvement and engagement consultation focus group study. J Med Internet Res 2023 Sep 26;25:e49303 [FREE Full text] [doi: 10.2196/49303] [Medline: 37751234]
37. Gould DJ, Dowsey MM, Glanville-Hearst M, Spelman T, Bailey JA, Choong PFM, et al. Patients' views on AI for risk prediction in shared decision-making for knee replacement surgery: qualitative interview study. J Med Internet Res 2023 Sep 18;25:e43632 [FREE Full text] [doi: 10.2196/43632] [Medline: 37721797]
38. Yousef CC, Salgado TM, Farooq A, Burnett K, McClelland LE, Abu Esba LC, et al. Health care providers' acceptance of a personal health record: cross-sectional study. J Med Internet Res 2021 Oct 21;23(10):e31582 [FREE Full text] [doi: 10.2196/31582] [Medline: 34569943]
39. Monaco A, Palmer K, Holm Ravn Faber N, Kohler I, Silva M, Vatland A, et al. Digital health tools for managing noncommunicable diseases during and after the COVID-19 pandemic: perspectives of patients and caregivers. J Med Internet Res 2021 Jan 29;23(1):e25652 [FREE Full text] [doi: 10.2196/25652] [Medline: 33464206]
40. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 09;23(2):e19910 [FREE Full text] [doi: 10.2196/19910] [Medline: 33560233]
41. Mesko B, Győrffy Z. The rise of the empowered physician in the digital health era: viewpoint. J Med Internet Res 2019 Mar 26;21(3):e12490 [FREE Full text] [doi: 10.2196/12490] [Medline: 30912758]
42. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]
43. Guo G, Cheng B. Effects of customer empowering behaviors on employees' career growth: Perspective of self-determination theory [顾客授权行为对员工职业成长的影响：自我决定理论视角]. Acta Psychol Sin 2022 Jul 13;53(2):215-228. [doi: 10.3724/SP.J.1041.2021.00215]
44. Öksüz M. The effect of customer empowering behaviors on service performance in hospitality industry. International Journal of Contemporary Economics Administrative Sciences 2021 Nov;11(1):193-215. [doi: 10.5281/zenodo.5136861]
45. Yu T, Lu L, Liu T. Exploring factors that influence knowledge sharing behavior via weblogs. Computers in Human Behavior 2010 Jan;26(1):32-41. [doi: 10.1016/j.chb.2009.08.002]

XSL•FO
RenderX

46. Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2011 Oct 05(10):CD001431. [doi: 10.1002/14651858.CD001431.pub3] [Medline: 21975733]
47. Erdem SA, Harrison-Walker LJ. The role of the internet in physician–patient relationships: the issue of trust. Business Horizons 2006 Sep;49(5):387-393. [doi: 10.1016/j.bushor.2006.01.003]
48. Almadana AV, Suharnomo S, Perdhana MS. Can generational differences and feeling trusted improve knowledge-sharing behavior? Consequences of high-performance work systems. JWL 2021 Sep 22;34(2):200-214. [doi: 10.1108/jwl-05-2021-0058]
49. Hashim KF, Tan FB. The mediating role of trust and commitment on members' continuous knowledge sharing intention: a commitment-trust theory perspective. International Journal of Information Management 2015 Apr;35(2):145-151. [doi: 10.1016/j.ijinfomgt.2014.11.001]
50. Lin MJ, Hung S, Chen C. Fostering the determinants of knowledge sharing in professional virtual communities. Computers in Human Behavior 2009 Jul;25(4):929-939. [doi: 10.1016/j.chb.2009.03.008]
51. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
52. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 2000 Oct;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]
53. Dahl AJ, Milne GR, Peltier JW. Digital health information seeking in an omni-channel environment: a shared decision-making and service-dominant logic perspective. Journal of Business Research 2021 Mar;125:840-850. [doi: 10.1016/j.jbusres.2019.02.025]
54. Deci EL, Eghrari H, Patrick BC, Leone DR. Facilitating internalization: the self-determination theory perspective. J Pers 1994 Mar;62(1):119-142. [doi: 10.1111/j.1467-6494.1994.tb00797.x] [Medline: 8169757]
55. Malik MAR, Butt AN, Choi JN. Rewards and employee creative performance: moderating effects of creative self‐efficacy, reward importance, and locus of control. J Organ Behavior 2014 Jul 02;36(1):59-74. [doi: 10.1002/job.1943]
56. Dissanayake I, Mehta N, Palvia P, Taras V, Amoako-Gyampah K. Competition matters! Self-efficacy, effort, and performance in crowdsourcing teams. Information & Management 2019 Dec;56(8):103158. [doi: 10.1016/j.im.2019.04.001]
57. Stirin Tzur K, Ganzach Y, Pazy A. On the positive and negative effects of self-efficacy on performance: reward as a moderator. Human Performance 2016 Jul 13;29(5):362-377. [doi: 10.1080/08959285.2016.1192631]
58. Tong Y, Wang X. Understanding the intention of information contribution to online feedback systems from social exchange and motivation crowding perspectives. 2007 Presented at: 40th Annual Hawaii International Conference on System Sciences 2007, HICSS; January 3-6, 2007; Waikoloa, HI p. 28 URL: https://ieeexplore.ieee.org/document/4076427/ [doi: 10.1109/hicss.2007.585]
59. Frey BS, Jegen R. Motivation crowding theory. Journal of Economic Surveys 2002 Dec 16;15(5):589-611. [doi: 10.1111/1467-6419.00150]
60. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
61. Spreitzer GM. Psychological, empowerment in the workplace: dimensions, measurement and validation. Academy of Management Journal 1995 Oct 01;38(5):1442-1465. [doi: 10.2307/256865]
62. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology 1988;54(6):1063-1070. [doi: 10.1037/0022-3514.54.6.1063]
63. Yang H, Lai C. Motivations of Wikipedia content contributors. Computers in Human Behavior 2010 Nov;26(6):1377-1383. [doi: 10.1016/j.chb.2010.04.011]
64. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
65. Kock N, Lynn G. Lateral collinearity and misleading results in variance-based SEM: an illustration and recommendations. JAIS 2012 Jul;13(7):546-580. [doi: 10.17705/1jais.00302]
66. Gefen, Rigdon, Straub. Editor's Comments: an update and extension to SEM guidelines for administrative and social science research. MIS Quarterly 2011;35(2):iii. [doi: 10.2307/23044042]
67. Martin SL, Liao H, Campbell EM. Directive versus empowering leadership: a field experiment comparing impacts on task proficiency and proactivity. AMJ 2013 Oct;56(5):1372-1395. [doi: 10.5465/amj.2011.0113]
68. Khenfer J, Shepherd S, Trendel O. Customer empowerment in the face of perceived Incompetence: effect on preference for anthropomorphized brands. Journal of Business Research 2020 Sep;118:1-11. [doi: 10.1016/j.jbusres.2020.06.010]
69. Griffin MA, Neal A, Parker SK. A new model of work role performance: positive behavior in uncertain and interdependent contexts. AMJ 2007 Apr;50(2):327-347. [doi: 10.5465/amj.2007.24634438]
70. Strauss K, Griffin MA, Rafferty AE. Proactivity directed toward the team and organization: the role of leadership, commitment and role‐breadth self‐efficacy. British J Management 2009 Aug 03;20(3):279-291. [doi: 10.1111/j.1467-8551.2008.00590.x]

XSL•FO
RenderX

Edited by A Schwartz; submitted 25.04.24; peer-reviewed by H Ju, SB Alavi; comments to author 08.08.24; revised version received 22.09.24; accepted 11.11.24; published 29.11.24.

Please cite as:
Su J, Shen KN, Guo X
Impact of Reverse Empowerment and Proactive Motivations on Physicians' Online Knowledge Sharing in Digital Platforms: Survey Study
J Med Internet Res 2024;26:e59904
URL: https://www.jmir.org/2024/1/e59904
doi:10.2196/59904
PMID:

©Jingyuan Su, Kathy Ning Shen, Xitong Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Application of Patient-Reported Outcome Measurements in Adult Tumor Clinical Trials in China: Cross-Sectional Study

Yan Jia[1,2*]; Qi Li[1,2*]; Xiaowen Zhang[1], MS; Yi Yan[3]; Shiyan Yan[4], PhD; Shunping Li[5], PhD; Wei Li[6], PhD; Xiaowen Wu[7], PhD; Hongguo Rong[1,8*], PhD; Jianping Liu[1,8], PhD

[1]Center for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China
[2]Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing, China
[3]School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China
[4]College of Acupuncture and Massage, Beijing University of Chinese Medicine, Beijing, China
[5]Centre for Health Management and Policy Research, Shandong University, Shandong, China
[6]International Research Center for Medicinal Administration, Peking University, Beijing, China
[7]Peking University Cancer Hospital & Institute, Peking University, Beijng, China
[8]Institute for Excellence in Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Hongguo Rong, PhD
Center for Evidence-Based Chinese Medicine
Beijing University of Chinese Medicine
No. 11 Beisanhuan East Road, Chaoyang District
Beijing, 100029
China
Phone: 86 (10)64286757
Email: hgrong@hsc.pku.edu.cn

Abstract

Background: International health policies and researchers have emphasized the value of evaluating patient-reported outcomes (PROs) in clinical studies. However, the characteristics of PROs in adult tumor clinical trials in China remain insufficiently elucidated.

Objective: This study aims to assess the application and characteristics of PRO instruments as primary or secondary outcomes in adult randomized clinical trials related to tumors in China.

Methods: This cross-sectional study identified tumor-focused randomized clinical trials conducted in China between January 1, 2010, and June 30, 2022. The ClinicalTrials.gov database and the Chinese Clinical Trial Registry were selected as the databases. Trials were classified into four groups based on the use of PRO instruments: (1) trials listing PRO instruments as primary outcomes, (2) trials listing PRO instruments as secondary outcomes, (3) trials listing PRO instruments as coprimary outcomes, and (4) trials without any mention of PRO instruments. Pertinent data, including study phase, settings, geographic regions, centers, participant demographics (age and sex), funding sources, intervention types, target diseases, and the names of PRO instruments, were extracted from these trials. The target diseases involved in the trials were grouped according to the *American Joint Committee on Cancer Staging Manual, 8th Edition*.

Results: Among the 6445 trials examined, 2390 (37.08%) incorporated PRO instruments as part of their outcomes. Within this subset, 26.82% (641/2390) listed PRO instruments as primary outcomes, 52.72% (1260/2390) as secondary outcomes, and 20.46% (489/2390) as coprimary outcomes. Among the 2,155,306 participants included in these trials, PRO instruments were used to collect data from 613,648 (28.47%) patients as primary or secondary outcomes and from 74,287 (3.45%) patients as coprimary outcomes. The most common conditions explicitly using specified PRO instruments included thorax tumors (217/1280, 16.95%), breast tumors (176/1280, 13.75%), and lower gastrointestinal tract tumors (173/1280, 13.52%). Frequently used PRO instruments included the European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire–30, the visual analog scale, the numeric rating scale, the Traditional Chinese Medicine Symptom Scale, and the Pittsburgh Sleep Quality Index.

Conclusions: Over recent years, the incorporation of PROs has demonstrated an upward trajectory in adult randomized clinical trials on tumors in China. Nonetheless, the infrequent measurement of the patient's voice remains noteworthy. Disease-specific

XSL•FO
RenderX

PRO instruments should be more effectively incorporated into various tumor disease categories in clinical trials, and there is room for improvement in the inclusion of PRO instruments as clinical trial end points.

(J Med Internet Res 2024;26:e45719) doi:10.2196/45719

KEYWORDS

patient-reported outcomes; tumor; cross-sectional study; quality of life; outcome study

Introduction

Background

Patient-reported outcome (PRO) instruments are defined as any report regarding a patient's health status obtained directly from the patient, excluding interpretation of the patient's responses by clinicians or other individuals [1]. PRO data consist of information obtained directly from patients concerning their health status, symptoms, treatment adherence, physical and social functioning, health-related quality of life, and satisfaction with health care [2-4]. Serving as noninvasive, comprehensive, and patient-centered metrics, PROs play a pivotal role in enhancing patient engagement, facilitating informed clinical decisions, and improving patient-clinician communication [5-9]. High-quality PRO measures examined in rigorous trials can evaluate treatment effectiveness, assess patient adherence to treatment, guide drug research, and inform health care policies [2,5]. In addition, some PRO instruments could supplement safety data and contribute to the assessment of tolerability (eg, Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events [PRO-CTCAE]) [2,5].

In particular, PROs are valuable end points in trials of disabling, chronic, and incurable conditions because they systematically capture the patients' perspectives in a scientifically rigorous way [3,10,11]. Recognizing their importance, clinical trials focused on tumors are increasingly incorporating PRO instruments as primary or secondary outcomes [12-15]. The European Commission has indicated the priority of preventing cancer and ensuring a high quality of life for patients with cancer within the framework of Europe's Beating Cancer Plan [16]. The incorporation of PROs in clinical trials offers distinct advantages, including improvements in health-related quality of life, patient-clinician communication, and economic benefits from reduced health care use [17-20]. To uphold best practices in tumor clinical trials that use PROs, several methodological recommendations have emerged in recent years, such as SPIRIT-PRO (Standard Protocol Items: Recommendations for Interventional Trials–Patient-Reported Outcome), CONSORT-PRO (Consolidated Standards of Reporting Trials–Patient-Reported Outcome), SISAQOL (Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints), and other relevant guidelines [2-4,21]. However, PRO measures often receive lower priority in the design of oncology-related clinical trials when compared to survival, imaging, and biomarker-related outcomes [22].

Objectives

In China, PROs are increasingly being used in clinical trials, but there are challenges as well. A cross-sectional survey of interventional clinical trials conducted in China revealed that only 29.7% of the included trials listed PRO instruments as primary or secondary outcomes [23]. Moreover, there is a notable absence of comprehensive assessments evaluating the application of PRO instruments in tumor clinical trials in China. Unlike previous cross-sectional studies that encompassed all types of clinical trials, our study primarily examined adult tumor clinical trials in China that have listed PRO instruments as primary or secondary outcomes, referencing the methodologies and reporting patterns of a previous study [23]. We extracted the registration information of adult randomized clinical trials conducted in China to systematically analyze the application of PRO instruments in tumor clinical trials, aiming to evaluate the application of PRO instruments in adult tumor clinical trials in China and provide potential directions for further investigation.

Methods

Study Design

This cross-sectional study was designed to describe the characteristics of adult tumor clinical trials conducted in China between January 1, 2010, and June 30, 2022, that listed PRO instruments as primary or secondary outcomes. All clinical trials should be registered, and data of clinical trials were collected from 2 clinical trial registries, namely ClinicalTrials.gov and the Chinese Clinical Trial Registry, with public disclosure. We conducted data retrieval and export in July 2022. The clinical trials covered 34 provincial-level administrative regions in accordance with the 2019 version of China's administrative divisions. We further sought to describe the PRO instruments frequently used in trials encompassing diverse target tumor conditions.

Data Collection Strategy

This study focused on interventional randomized clinical trials conducted in China involving participants aged ≥18 years (Figure 1). Duplicate trials with 2 registration identification numbers were treated as a single trial (ClinicalTrials.gov records were retained). The evaluation of tumor clinical trials included three types of information: (1) basic information (registration number, registration date, scientific name, recruiting country, and other information), (2) key information (outcome, target disease, and age and sex of participants), and (3) characteristic information (main sponsor's location, study settings, number of setting centers, study stage, funding source, and intervention type).

XSL•FO
RenderX

Figure 1. Trial exclusion and classification criteria. PRO: patient-reported outcome.

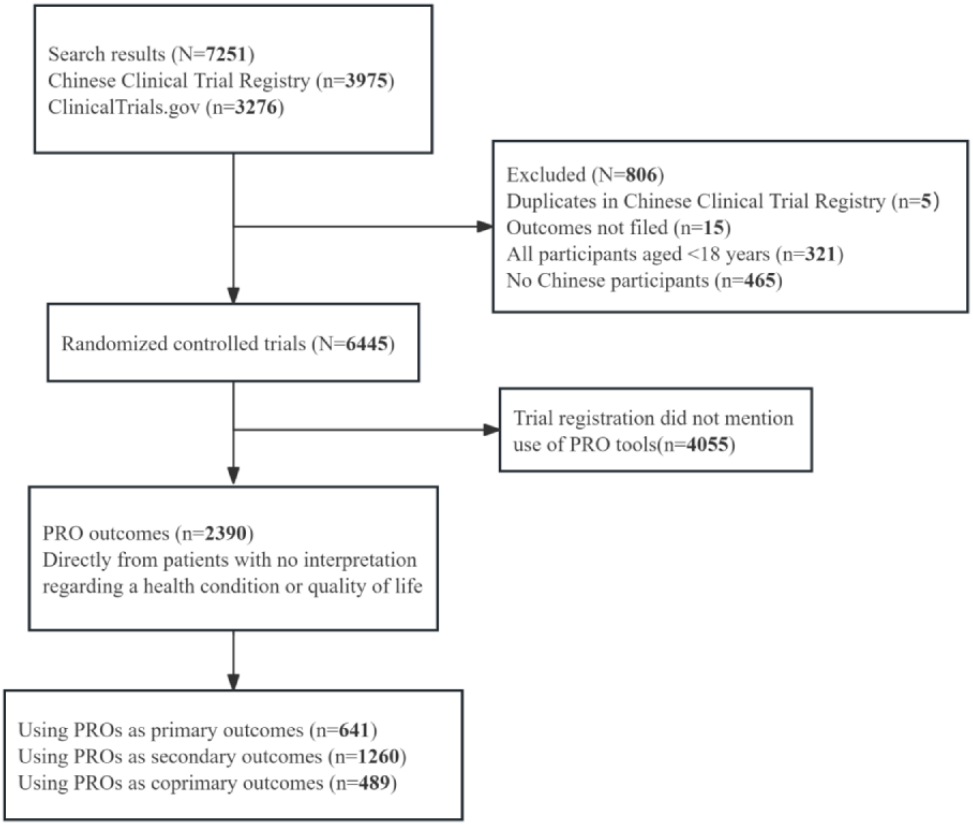

Data Classification

PRO instruments were defined by the US Food and Drug Administration in 2009 [1] as any report about a patient's health status obtained directly from the patient, excluding interpretation of the patient's response by clinicians or other individuals. Trials using PRO instruments as primary or secondary outcomes were considered PRO trials. On the basis of a previous study of PRO labeling of new US Food and Drug Administration–approved drugs (2016-2020) [24], eligible trials were classified into four groups: (1) trials that listed PRO instruments as primary outcomes, (2) trials that listed PRO instruments as secondary outcomes, (3) trials that listed PRO instruments as coprimary outcomes, and (4) trials without any mention of PRO instruments.

Statistical Analysis

Data related to the characteristics of the included trials (clinical phase, study setting, participant age and sex, region of the primary sponsor, setting center, number of PROs, funding source, and type of intervention) were extracted independently by 2 authors with a predesigned data extraction table. Owing to the varied categories and wide variation of target diseases, we classified similar target diseases based on classifications from the *American Joint Committee on Cancer Staging Manual, 8th Edition* (Multimedia Appendix 1). On the basis of this categorization of diseases, we consolidated the PRO instruments used in each trial to identify those used most frequently. We conducted quantitative analysis only on items that listed the names of PRO instruments for a more detailed understanding of the commonly used evaluation tools. All data analyses were performed using Stata (version 14.0; StataCorp LLC).

Ethical Considerations

According to the Common Rule (45 CFR part 46) of the US Department of Health and Human Services (Office for Human Research Protections), this study is exempt from institutional review board approval and the requirement for informed patient consent because it did not involve clinical data or human participants. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines designed for observational studies in epidemiology.

Results

Trial Characteristics

Table 1 presents a comprehensive overview of the included trials. The study included 7251 tumor-focused randomized controlled trials conducted in China between January 1, 2010, and June 30, 2022. Of these 7251 trials, 3276 (45.18%) were sourced from ClinicalTrials.gov, and 3975 (54.82%) were identified from the Chinese Clinical Trial Registry. Of these 7251 trials, after excluding 806 (11.12%) trials (n=5, 0.6% duplicates; n=465, 57.7% non-Chinese trials; n=321, 39.8% trials involving children; and n=15, 1.9% trials with incomplete reports), 6445 (88.88%) eligible trials were identified for analysis.

XSL•FO
RenderX

Table 1. Characteristics of all identified trials and patient-reported outcome (PRO)–related trials.

Characteristics	Trials (n=6445), n (%)	PRO trials (n=2390), n (%)
Phase		
Early phase[a]	1317 (20.43)	575 (24.06)
2	873 (13.55)	218 (9.12)
3	1004 (15.58)	284 (11.88)
4	779 (12.09)	269 (11.26)
Other[b]	1514 (23.49)	537 (22.47)
Unclear	958 (14.86)	507 (21.21)
Setting		
Hospital	6034 (93.62)	2256 (94.39)
Community	3 (0.05)	3 (0.13)
Other[c]	300 (4.65)	96 (4.02)
Unclear	108 (1.68)	35 (1.46)
Age (y)		
≥18	6098 (94.62)	2252 (94.23)
>65	100 (1.55)	42 (1.76)
Unclear	247 (3.83)	96 (4.02)
Sex		
Male only	267 (4.14)	107 (4.48)
Female only	1000 (15.52)	410 (17.15)
Male and female	5170 (80.22)	1869 (78.2)
Unclear	8 (0.12)	4 (0.17)
Regions of China		
Southwest	517 (8.02)	207 (8.66)
Northeast	185 (2.87)	83 (3.47)
North	797 (12.37)	321 (13.43)
Northwest	173 (2.68)	63 (2.64)
East	3745 (58.11)	1309 (54.77)
South	682 (10.58)	286 (11.97)
Central	340 (5.28)	120 (5.02)
Other[d]	6 (0.09)	1 (0.04)
Centers involved		
Single	5626 (87.29)	2046 (85.61)
Multiple	716 (11.11)	312 (13.05)
Unclear	103 (1.6)	32 (1.34)
PRO instruments used		
1-3	N/A[e]	2144 (89.71)
4-6	N/A	218 (9.12)
7-9	N/A	25 (1.05)
≥10	N/A	3 (0.13)
Funding source		
Industry	752 (11.67)	186 (7.78)

XSL•FO
RenderX

Characteristics	Trials (n=6445), n (%)	PRO trials (n=2390), n (%)
Nonindustry institutions	5443 (84.45)	2120 (88.7)
Combination[f]	219 (3.4)	71 (2.97)
Unclear	31 (0.48)	13 (0.54)

[a]The early phase trials included a clinical pretest as well as phase 0 and phase 1 trials.

[b]Diagnostic new technique clinical study, inspection technology, and trials involving multiple phases.

[c]Rehabilitation center, nursing home, campus, centers for disease control, home, and research institute.

[d]The trials were conducted in China, but their sponsor was based overseas.

[e]N/A: not applicable.

[f]Combination trials were funded partly by industry and partly by nonindustry institutions, such as universities, hospitals, and so on.

Of the 2,155,306 participants recruited in all included trials, 139,297 (6.46%) were involved in trials with PRO instruments as primary outcomes, 400,064 (18.56%) in trials with PRO instruments as secondary outcomes, and 74,287 (3.45%) in trials with PRO instruments as coprimary outcomes. Among the 6445 trials included, 2390 (37.08%) used PRO instruments as either primary or secondary outcomes, while 4055 (62.92%) did not use any PRO instrument.

The majority of the studies (6098/6445, 94.62%) did not impose any age restrictions on participants (children were excluded). In trials involving PROs, the proportion of older participants (aged >65 y; 42/2390, 1.78%) was slightly higher than in those without PROs (100/6445, 1.55%). Among all trials that incorporated PRO measurements, 17.15% (410/2390) included only female participants, while 4.48% (107/2390) included only male participants. Furthermore, in trials involving only female participants, the vast majority (974/1000, 97.4%) studied breast and female reproductive organ tumors. In trials exclusively involving male participants, more than half (135/267, 50.5%) centered around male genital organ tumors.

Regarding trial phases, of the 6445 clinical trials, early phase trials were the most prevalent (n=1317, 20.43%), followed by phase 3 trials (n=1004, 15.58%), phase 2 trials (n=873, 13.56%), and phase 4 trials (n=779, 12.09%). Of the 2390 PRO-related trials, early phase trials were again the most common (n=575, 24.06%), followed by phase 3 trials (n=284, 11.88%), phase 4 trials (n=269, 11.26%), and phase 2 trials (n=218, 9.12%).

Most of the trials (6034/6445, 93.62%) were conducted in hospitals, with hardly any (3/6445, 0.05%) conducted in community settings. More than half of the primary sponsors were located in eastern China (3745/6445, 58.11%), followed by northern (797/6445, 12.37%) and southern (682/6445, 10.58%) China, while 18.85% (1215/6445) of the primary sponsors were situated in other regions of China, such as the southwestern, central, northwestern, and northeastern regions. Similar patterns were observed for studies involving PROs. The majority of the major sponsors (1916/2390, 80.17%) originated from the eastern, northern, and southern regions of China, while 19.79% (473/2390) hailed from the southwestern, central, northeastern, and northwestern regions. There were differences in the proportions of PRO trials were noted among different provinces; the distribution of PRO instruments across Chinese provinces can be found in Multimedia Appendix 2.

Moreover, 87.29% (5626/6445) of the trials were single-center trials, and only 11.11% (716/6445) were multicenter trials. Similar phenomena were observed for PRO-related studies, but multicenter trials accounted for a slightly higher percentage (312/2390, 13.05%). Of the 2390 PRO trials, 2144 (89.71%) used 1 to 3 PRO instruments, followed by 4 to 6 (n=218, 9.12%) and 7 to 9 (n=25, 1.05%) PRO instruments. The majority of the trials were nonindustry-funded trials (5443/6445, 84.45%), while 11.67% (752/6445) were industry-funded trials.

Table 2 shows the frequency of intervention types used across different trial classifications. The data indicated that more than a third of the included trials used drugs as the intervention (2496/6445, 38.73%), followed by combination therapies (1350/6445, 20.95%) and surgery (1044/6445, 16.2%). Among clinical trials involving drug interventions, nearly four-tenths (989/2496, 39.62%) used PRO instruments as their outcomes. Trials using drugs as the intervention exhibited a higher incidence of using PRO instruments as their primary or coprimary outcomes (468/989, 47.32%) compared to trials using other intervention types.

XSL•FO
RenderX

Table 2. Frequency of intervention types used across different trial classifications.

Intervention	Trials (n=6445), n (%)	PRO[a] trials			
		Trials, n/N (%)	Primary outcome, n/N (%)	Secondary outcome, n/N (%)	Coprimary outcome, n/N (%)
Drug	2496 (38.73)	989/2496 (39.62)	250/989 (25.28)	521/989 (52.68)	218/989 (22.04)
Biological	400 (6.21)	73/400 (18.25)	8/73 (10.96)	61/73 (83.56)	4/73 (5.48)
Surgery	1044 (16.2)	329/1044 (31.51)	124/329 (37.69)	184/329 (55.93)	21/329 (6.38)
Radiation	187 (2.9)	57/187 (30.48)	11/57 (19.3)	44/57 (77.19)	2/57 (3.51)
Combination	1350 (20.95)	375/1350 (27.78)	52/375 (13.87)	297/375 (79.2)	26/375 (6.93)
Other[b]	968 (15.02)	567/968 (58.57)	196/567 (34.57)	153/567 (26.98)	218/567 (38.45)

[a]PRO: patient-reported outcome.

[b]Other interventions included acupuncture, physical exercise, and psychosocial treatment.

Conditions and Participants

The annual counts of tumor clinical trials are listed in Figure 2. During the study period—from January 1, 2010, to June 30, 2022—the number of tumor clinical trial registrations exhibited a consistent upward trajectory, paralleled by a commensurate increase in the number of clinical trials related to PROs.

Figure 2. Number of tumor clinical trials analyzed. PRO: patient-reported outcome.

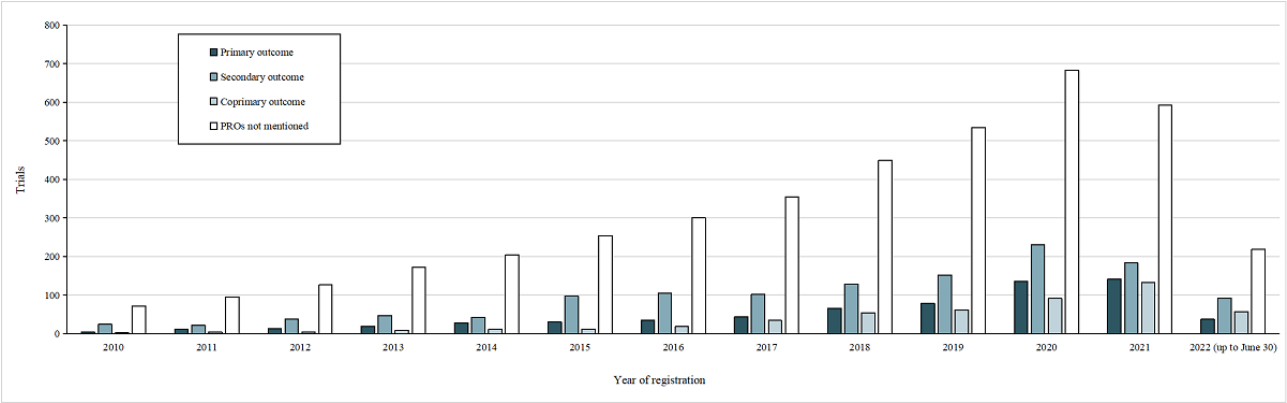

Figures 3 and 4 depict the distribution of trial counts and corresponding participant numbers across different tumor types, respectively, wherein PROs served as outcomes. Among the 2390 tumor-related trials that used PRO instruments as primary or secondary outcomes, the top 5 tumors were thorax (448/2390, 18.74%), upper gastrointestinal tract (306/2390, 12.8%), lower gastrointestinal tract (300/2390, 12.55%), breast (289/2390, 12.09%), and head and neck (177/2390, 7.41%) tumors. Trials regarding female reproductive organ (168/2390, 7.03%) and hepatobiliary system (146/2390, 6.11%) tumors were also frequently observed. Male genital organ tumors (56/2390, 2.34%), central nervous system tumors (51/2390, 2.13%), endocrine system tumors (47/2390, 1.97%), and urinary tract tumors (33/2390, 1.38%) all accounted for proportions ranging from 1% to 5%, and hematologic malignant tumors (22/2390, 0.92%), neuroendocrine tumors (14/2390, 0.59%), bone tumors (8/2390, 0.33%), skin tumors (4/2390, 0.17%), ophthalmic tumors (2/2390, 0.08%), and soft tissue sarcoma (1/2390, 0.04%) constituted <1% of the trials.

XSL•FO
RenderX

Figure 3. Number of trials with patient-reported outcomes (PROs).

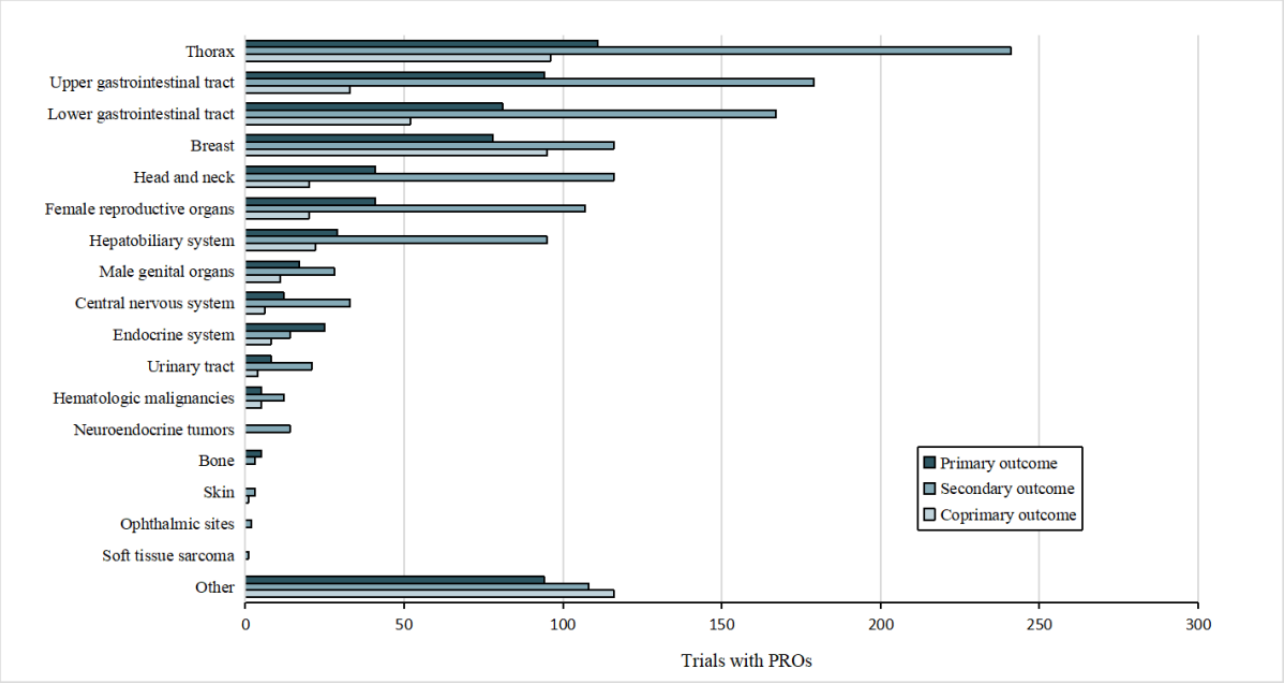

Figure 4. Number of participants with patient-reported outcomes (PROs).

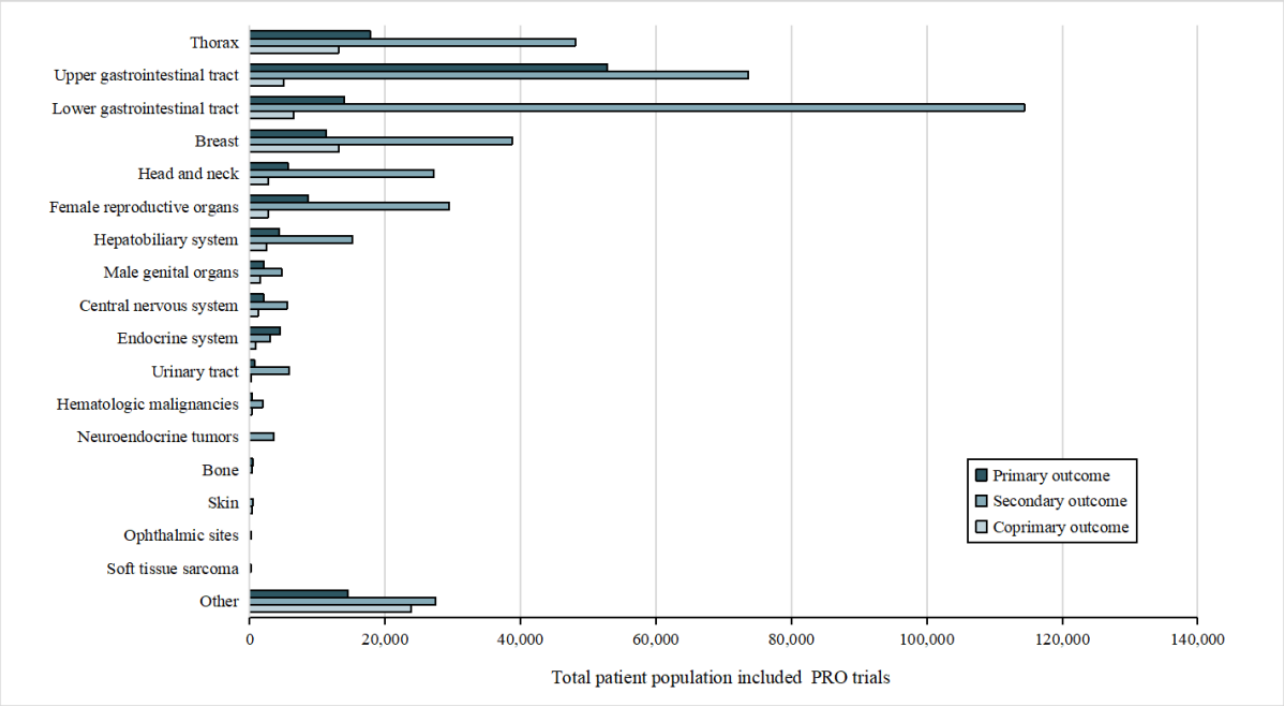

Among the 613,648 participants enrolled in these PRO trials, 134,940 (22%) were diagnosed with lower gastrointestinal tract tumors, 131,470 (21.42%) with upper gastrointestinal tract tumors, and 79,068 (12.88%) with thorax tumors. Furthermore, there were a number of patients with breast tumors (63,238/613,648, 10.31%), female reproductive organ tumors (440,975/613,648, 6.68%), head and neck tumors (35,642/613,648, 5.81%), or hepatobiliary system tumors (22,044/613,648, 3.59%), each involving >10,000 patients. By contrast, conditions with <10,000 participants encompassed central nervous system tumors (8897/613,648, 1.45%), endocrine system tumors (8472/613,648, 1.38%), male genital organ tumors (8357/613,648, 1.36%), urinary tract tumors (6784/613,648, 1.11%), neuroendocrine tumors (3539/613,648, 0.58%), hematologic malignant tumors (2629/613,648, 0.43%), bone tumors (825/613,648, 0.13%), skin tumors (311/613,648, 0.05%), ophthalmic tumors (274/613,648, 0.04%), and soft tissue sarcoma (266/613,648, 0.04%).

PRO Instruments Used in Clinical Trials

Table 3 presents the number of explicitly specified PROs where trials precisely listed the names of the PRO instruments and the number of implicitly specified PROs where trials referenced patients' subjective feelings without specifying the instruments

XSL•FO
RenderX

used, separately for the 3 trial types. Specifically, the trial that specified the PRO instruments used was classified into "explicitly specified PROs," and the trial that did not specify the instruments used was classified into "implicitly specified PROs." It was evident that in primary and coprimary outcome trial sets, a greater number of trials explicitly listed the PRO instruments compared to those that did not specify the instruments used. Among the 3 trial types, the coprimary outcome category exhibited the highest proportion of explicitly specified PROs (339/489, 69.3%).

Table 3. Number of explicitly specified patient-reported outcomes (PROs) and implicitly specified PROs used in different trial classifications (n=2390).

Classification	Explicitly specified PROs, n (%)	Implicitly specified PROs, n (%)
Primary outcome (n=641)	335 (52.3)	306 (47.7)
Secondary outcome (n=1260)	606 (48.1)	654 (51.9)
Coprimary outcome (n=489)	339 (69.3)	150 (30.7)

Tables 4-6 display the frequency of use of PRO scales for different diseases under the 3 categories. In trials using PRO instruments as coprimary outcomes, the visual analog scale (VAS) and the numeric rating scale (NRS) were the most commonly used scales for various tumors. For trials using PRO instruments as primary outcomes, the VAS was the most commonly used scale for various diseases. For trials using PRO instruments as secondary outcomes, the most commonly used scale for each disease was the European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30 (EORTC QLQ-C30).

XSL•FO
RenderX

Table 4. Frequency of patient-reported outcome (PRO) instruments used as primary outcomes in different classifications of tumor trials by condition.

Condition	Trials (n=1280), n (%)	Instrument frequency in trials with PRO instruments as primary outcomes			
		Trials, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)
Thorax tumors	217 (16.95)	56/217 (25.81)	VAS[a], 25/56 (44.64)	NRS[b], 9/56 (16.07)	EORTC QLQ-LC43[c], 4/56 (7.14)
Breast tumors	176 (13.75)	41/176 (23.3)	VAS, 14/41 (34.15)	SF-36[d], 4/41 (9.76)	PSQI[e], 3/41 (7.32)
Lower gastrointestinal tract tumors	173 (13.52)	43/173 (24.86)	VAS, 14/43 (32.56)	IPSS[f], 4/43 (9.3)	LARS[g], 4/43 (9.3)
Upper gastrointestinal tract tumors	140 (10.94)	46/140 (32.86)	VAS, 13/46 (28.26)	EORTC QLQ-C30[h], 12/46 (26.09)	EORTC QLQ-STO22[i], 8/46 (17.39)
Head and neck tumors	101 (7.89)	21/101 (20.79)	VAS, 3/21 (14.29)	UW-QOL[j], 3/21 (14.29)	NRS, 2/21 (9.52)
Female reproductive organ tumors	85 (6.64)	20/85 (23.53)	VAS, 6/20 (30)	NRS, 3/20 (15)	EORTC QLQ-C30, 3/20 (15)
Hepatobiliary system tumors	67 (5.23)	14/67 (20.9)	VAS, 6/14 (42.86)	EORTC QLQ-C30, 3/14 (21.43)	QoR-40[k], 1/14 (7.14)
Male genital organ tumors	31 (2.42)	11/31 (35.48)	VAS, 7/11 (63.64)	NRS, 3/11 (27.27)	EORTC QLQ-C30, 2/11 (18.18)
Endocrine system tumors	30 (2.34)	16/30 (53.33)	VAS, 9/16 (56.25)	IDS[l], 2/16 (12.5)	EORTC QLQ-C30, 1/16 (6.25)
Central nervous system tumors	25 (1.95)	5/25 (20)	VAS, 2/5 (40)	QoR-40, 1/5 (20)	NRS, 1/5 (20)
Urinary tract tumors	17 (1.33)	4/17 (23.53)	VAS, 1/4 (25)	IIEF-15[m], 1/4 (25)	QoR-15[n], 1/4 (25)
Bone tumors	8 (0.63)	5/8 (62.5)	NRS, 2/5 (40)	VAS, 2/5 (40)	SF-36, 1/5 (20)
Hematologic malignant tumors	9 (0.7)	2/9 (22.22)	NRS, 1/2 (50)	TCMSS[o], 1/2 (50)	N/A[p]
Neuroendocrine tumors	10 (0.78)	N/A	N/A	N/A	N/A
Skin tumors	3 (0.23)	N/A	N/A	N/A	N/A
Ophthalmic tumors	1 (0.08)	N/A	N/A	N/A	N/A
Soft tissue sarcoma	0 (0)	N/A	N/A	N/A	N/A
Other tumors	187 (14.61)	51/187 (27.27)	VAS, 18/51 (35.29)	NRS, 11/51 (21.57)	EORTC QLQ-C30, 4/51 (7.84)

[a]VAS: visual analog scale.

[b]NRS: numeric rating scale.

[c]EORTC QLQ-LC43: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Lung Cancer 43.

[d]SF-36: 36-item Short Form Health Survey.

[e]PSQI: Pittsburgh Sleep Quality Index.

[f]IPSS: International Prostate Symptom Score.

[g]LARS: Low Anterior Resection Syndrome.

[h]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30.

[i]EORTC QLQ-STO22: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Stomach 22.

[j]UW-QOL: University of Washington Quality of Life Questionnaire.

[k]QoR-40: Quality of Recovery-40.

[l]IDS: Involvement-Detachment Scale.

[m]IIEF-15: International Index of Erectile Function-15.

[n]QoR-15: Quality of Recovery-15.

[o]TCMSS: Traditional Chinese Medicine Symptom Scale.

[p]N/A: not applicable.

XSL•FO
RenderX

Table 5. Frequency of patient-reported outcome (PRO) instruments used as secondary outcomes in different classifications of tumor trials by condition.

Conditions	Trials (n=1280), n (%)	Instrument frequency in trials with PRO instruments as secondary outcomes			
		Trials, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)
Thorax tumors	217 (16.95)	98/217 (45.16)	EORTC QLQ-C30[a], 33/98 (33.67)	FACT-L[b], 18/98 (18.37)	EORTC QLQ-LC13[c], 15/98 (15.31)
Breast tumors	176 (13.75)	67/176 (38.07)	EORTC QLQ-C30, 23/67 (34.33)	FACT-B[d], 16/67 (23.88)	EORTC QLQ-BR23[e], 13/67 (19.4)
Lower gastrointestinal tract tumors	173 (13.52)	94/173 (54.34)	EORTC QLQ-C30, 38/94 (40.43)	VAS[f], 21/94 (22.34)	Wexner Scale, 11/94 (11.7)
Upper gastrointestinal tract tumors	140 (10.94)	72/140 (51.43)	EORTC QLQ-C30, 41/72 (56.94)	EORTC QLQ-OES18[g], 15/72 (20.83)	VAS, 13/72 (18.06)
Head and neck tumors	101 (7.89)	68/101 (67.33)	EORTC QLQ-C30, 38/68 (55.88)	EORTC QLQ-H&N35[h], 34/68 (50)	NRS[i], 6/68 (8.82)
Female reproductive organ tumors	85 (6.64)	49/85 (57.65)	EORTC QLQ-C30, 18/49 (36.73)	VAS, 5/49 (10.2)	EORTC QLQ-CX24[j], 5/49 (10.2)
Hepatobiliary system tumors	67 (5.23)	37/67 (55.22)	EORTC QLQ-C30, 18/37 (48.65)	EORTC QLQ-HCC18[k], 7/37 (18.92)	VAS, 5/37 (13.51)
Male genital organ tumors	31 (2.42)	10/31 (32.26)	FACT-P[l], 4/10 (40)	BPI-SF[m], 3/10 (30)	FACT-G[n], 3/10 (30)
Endocrine system tumors	30 (2.34)	7/30 (23.33)	VAS, 3/7 (42.86)	QoR-40[o], 1/7 (14.29)	SF-36[p], 1/7 (14.29)
Central nervous system tumors	25 (1.95)	14/25 (56)	VAS, 8/14 (57.14)	NRS, 4/14 (28.57)	QoR-15[q], 3/14 (21.43)
Urinary tract tumors	17 (1.33)	11/17 (64.71)	EORTC QLQ-C30, 3/11 (27.27)	VAS, 3/11 (27.27)	WHOQOL-BREF[r], 2/11 (18.18)
Neuroendocrine tumors	10 (0.78)	10/10 (100)	EORTC QLQ-C30, 6/10 (60)	EORTC QLQ-PAN26[s], 2/10 (20)	VAS, 2/10 (20)
Hematologic malignant tumors	9 (0.7)	4/9 (44.44)	EORTC QLQ-C30, 2/4 (50)	FACIT[t], 2/4 (50)	EQ-5D-5L, 1/4 (25)
Skin tumors	3 (0.23)	3/3 (100)	EORTC QLQ-C30, 1/3 (33.33)	VAS, 1/3 (33.33)	HF-QoL[u], 1/3 (33.33)
Bone tumors	8 (0.63)	3/8 (37.5)	EORTC QLQ-C30, 2/3 (66.67)	BPI-SF, 1/3 (33.33)	N/A[v]
Ophthalmic tumors	1 (0.08)	1/1 (100)	EORTC QLQ-C30, 1/1 (100)	EORTC QLQ-OPT30[w], 1/1 (100)	N/A
Soft tissue sarcoma	0 (0)	N/A	N/A	N/A	N/A
Other tumors	187 (14.61)	58/187 (31.02)	VAS, 12/58 (20.69)	EORTC QLQ-C30, 12/58 (20.69)	NRS, 9/58 (15.52)

[a]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30.

[b]FACT-L: Functional Assessment of Cancer Therapy–Lung.

[c]EORTC QLQ-LC13: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Lung Cancer 13.

[d]FACT-B: Functional Assessment of Cancer Therapy–Breast.

[e]EORTC QLQ-BR23: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Breast Cancer 23.

[f]VAS: visual analog scale.

[g]EORTC QLQ-OES18: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Oesophageal Cancer 18.

[h]EORTC QLQ-H&N35: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Head and Neck Cancer 35.

[i]NRS: numeric rating scale.

[j]EORTC QLQ-CX24: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Cervical Cancer 24.

[k]EORTC QLQ-HCC18: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Hepatocellular Carcinoma 18.

[l]FACT-P: Functional Assessment of Cancer Therapy–Prostate.

[m]BPI-SF: Brief Pain Inventory–Short Form.

XSL•FO
RenderX

[n]FACT-G: Functional Assessment of Cancer Therapy–General.

[o]QoR-40: Quality of Recovery-40.

[p]SF-36: 36-item Short Form Health Survey.

[q]QoR-15: Quality of Recovery-15.

[r]WHOQOL-BREF: World Health Organization Quality of Life Brief Version.

[s]EORTC QLQ-PAN26: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Pancreatic Cancer 26.

[t]FACIT: Functional Assessment of Chronic Illness Therapy.

[u]HF-QOL: Hand-Foot Skin Reaction and Quality of Life.

[v]N/A: not applicable.

[w]EORTC QLQ-OPT30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Ophthalmic Cancer 30.

XSL•FO
RenderX

Table 6. Frequency of patient-reported outcome (PRO) instruments used as coprimary outcomes in different classifications of tumor trials by condition.

Conditions	Trials (n=1280), n (%)	Instrument frequency in trials with PRO instruments as coprimary outcomes			
		Trials, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)
Thorax tumors	217 (16.95)	63/217 (29.03)	VAS[a], 22/63 (34.92)	NRS[b], 17/63 (26.98)	EORTC QLQ-C30[c], 16/63 (25.4)
Breast tumors	176 (13.75)	68/176 (38.64)	VAS, 22/68 (32.35)	NRS, 13/68 (19.12)	QoR-15[d], 9/68 (13.24)
Lower gastrointestinal tract tumors	173 (13.52)	36/173 (20.81)	VAS, 13/36 (36.11)	EORTC QLQ-C30, 7/36 (19.44)	QoR-15, 6/36 (16.67)
Upper gastrointestinal tract tumors	140 (10.94)	22/140 (15.71)	VAS, 7/22 (31.82)	NRS, 7/22 (31.82)	EORTC QLQ-C30, 3/22 (13.64)
Head and neck tumors	101 (7.89)	12/101 (11.88)	NRS, 4/12 (33.33)	EORTC QLQ-C30, 3/12 (25)	TNSS[e], 2/12 (16.67)
Female reproductive organ tumors	85 (6.64)	16/85 (18.82)	VAS, 6/16 (37.5)	BCS[f], 2/16 (12.5)	NRS, 2/16 (12.5)
Hepatobiliary system tumors	67 (5.23)	16/67 (23.88)	NRS, 7/16 (43.75)	EORTC QLQ-C30, 2/16 (12.5)	PSQI[g], 2/16 (12.5)
Male genital organ tumors	31 (2.42)	10/31 (32.26)	VAS, 3/10 (30)	ICIQ-SF[h], 3/10 (30)	FACT-P[i], 2/10 (20)
Endocrine system tumors	30 (2.34)	7/30 (23.33)	QoR-15, 3/7 (42.86)	VAS, 3/7 (42.86)	HADS[j], 1/7 (14.29)
Central nervous system tumors	25 (1.95)	6/25 (24)	VAS, 5/6 (83.33)	EORTC IADL-BN32[k], 1/6 (16.67)	N/A[l]
Urinary tract tumors	17 (1.33)	2/17 (11.76)	NRS, 1/2 (50)	QoR-15, 1/2 (50)	VAS, 1/2 (50)
Hematologic malignant tumors	9 (0.7)	3/9 (33.33)	SAS[m], 1/3 (33.33)	SDS[n], 1/3 (33.33)	NRS, 1/3 (33.33)
Bone tumors	8 (0.63)	N/A	N/A	N/A	N/A
Neuroendocrine tumors	10 (0.78)	N/A	N/A	N/A	N/A
Skin tumors	3 (0.23)	N/A	N/A	N/A	N/A
Ophthalmic tumors	1 (0.08)	N/A	N/A	N/A	N/A
Soft tissue sarcoma	0 (0)	N/A	N/A	N/A	N/A
Other tumors	187 (14.61)	78/187 (41.71)	NRS, 24/78 (30.77)	VAS, 14/78 (17.95)	PSQI, 12/78 (15.38)

[a]VAS: visual analog scale.

[b]NRS: numeric rating scale.

[c]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30.

[d]QoR-15: Quality of Recovery-15.

[e]TNSS: Total Nasal Symptom Score.

[f]BCS: Bruggemann Comfort Scale.

[g]PSQI: Pittsburgh Sleep Quality Index.

[h]ICIQ-SF: International Consultation on Incontinence Questionnaire–Short Form.

[i]FACT-P: Functional Assessment of Cancer Therapy–Prostate.

[j]HADS: Hospital Anxiety and Depression Scale.

[k]EORTC IADL-BN32: European Organisation for Research and Treatment of Cancer Instrumental Activities of Daily Living in Patients With Brain Tumors-32.

[l]N/A: not applicable.

[m]SAS: Self-Rating Anxiety Scale.

[n]SDS: Self-Rating Depression Scale.

To analyze the overall application of scales in explicitly specified PROs by condition, we examined the specific PRO instruments used in trials that explicitly mentioned the PRO instruments as primary or secondary outcomes (Table 7). Of the 1280 trials, 321 (25.08%) used the EORTC QLQ-C30 (Multimedia Appendix 3), which was the most commonly used PRO scale. Of note, the EORTC QLQ-C30 was the most commonly used scale in trials concerning lower gastrointestinal

XSL•FO
RenderX

tract, upper gastrointestinal tract, head and neck, female reproductive organ, hepatobiliary system, bone, neuroendocrine, skin, and ophthalmic tumors as well as hematologic malignancies. In addition, the VAS was used in 24.77% (317/1280) of the trials (Multimedia Appendix 3), predominating in trials involving thorax, breast, male genital organ, endocrine system, central nervous system, and urinary tract tumors. The NRS was also frequently used (169/1280, 13.2%) in cancer trials. More targeted scales have been used for different tumor diseases; for example, the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ)–Head and Neck Cancer 35 (36/101, 35.6%) was more common in head and neck tumor trials, the EORTC QLQ–Oesophageal Cancer 18 (15/140, 10.7%) and the EORTC QLQ–Stomach 22 (14/140, 10%) were frequently observed in upper gastrointestinal cancer trials, the EORTC QLQ–Colorectal Cancer 29 (14/173, 8.1%) scale was prevalent in lower gastrointestinal cancer trials, the EORTC QLQ–Hepatocellular Carcinoma 18 (8/67, 12%) was frequently found in hepatobiliary system tumor trials, the Functional Assessment of Cancer Therapy (FACT)–Lung (21/217, 9.7%) and the EORTC QLQ–Lung Cancer 13 (19/217, 8.8%) commonly featured in thorax tumor trials, the FACT–Breast (29/176, 16.5%) and the EORTC QLQ–Breast Cancer 23 (16/176, 9.1%) were frequently seen in breast cancer trials, the EORTC QLQ–Ovarian Cancer 28 (6/85, 7%) was a typical scale used in female reproductive organ tumor trials, the FACT–Prostate (7/31, 23%) was often used in male genital organ tumor trials, and the FACT–Anemia (1/9, 11%) and the FACT–Lymphoma (1/9, 11%) were common choices in hematologic malignant tumor trials.

XSL•FO
RenderX

Table 7. Frequency of use of patient-reported outcome (PRO) instruments by condition.

Conditions	Trials (n=1280), n (%)	Instrument, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)	Instrument, n/N (%)
Thorax tumors	217 (16.95)	VAS[a], 57/217 (26.27)	EORTC QLQ-C30[b], 53/217 (24.42)	NRS[c], 32/217 (14.75)	FACT-L[d], 21/217 (9.68)	EORTC QLQ-LC13[e], 19/217 (8.76)
Breast tumors	176 (13.75)	VAS, 43/176 (24.43)	FACT-B[f], 29/176 (16.48)	EORTC QLQ-C30, 26/176 (14.77)	EORTC QLQ-BR23[g], 16/176 (9.09)	NRS, 15/176 (8.52)
Lower gastrointestinal tract tumors	173 (13.52)	EORTC QLQ-C30, 49/173 (28.32)	VAS, 48/173 (27.75)	NRS, 15/173 (8.67)	EORTC QLQ-CR29[h], 14/173 (8.09)	Wexner Scale, 14/173 (8.09)
Upper gastrointestinal tract tumors	140 (10.94)	EORTC QLQ-C30, 56/140 (40)	VAS, 33/140 (23.57)	NRS, 17/140 (12.14)	EORTC QLQ-OES18[i], 15/140 (10.71)	EORTC QLQ-STO22[j], 14/140 (10)
Head and neck tumors	101 (7.89)	EORTC QLQ-C30, 41/101 (40.59)	EORTC QLQ-H&N35[k], 36/101 (35.64)	NRS, 12/101 (11.88)	VAS, 11/101 (10.89)	PG-SGA[l], 8/101 (7.92)
Female reproductive organ tumors	85 (6.64)	EORTC QLQ-C30, 21/85 (24.71)	VAS, 17/85 (20)	NRS, 6/85 (7.06)	SDS[m], 6/85 (7.06)	EORTC QLQ-OV28[n], 6/85 (7.06)
Hepatobiliary system tumors	67 (5.23)	EORTC QLQ-C30, 23/67 (34.33)	VAS, 13/67 (19.4)	NRS, 9/67 (13.43)	EORTC QLQ-HCC18[o], 8/67 (11.94)	TCMSS[p], 5/67 (7.46)
Male genital organ tumors	31 (2.42)	VAS, 10/31 (32.26)	FACT-P[q], 7/31 (22.58)	BPI[r], 4/31 (12.9)	IPSS[s], 4/31 (12.9)	NRS, 4/31 (12.9)
Endocrine system tumors	30 (2.34)	VAS, 15/30 (50)	QoR-15[t], 4/30 (13.33)	EORTC QLQ-C30, 2/30 (6.67)	NRS, 2/30 (6.67)	QoR-40[u], 2/30 (6.67)
Central nervous system tumors	25 (1.95)	VAS, 15/25 (60)	NRS, 4/25 (16)	QoR-15, 3/25 (12)	QoR-40, 2/25 (8)	PCSQ[v], 2/25 (8)
Urinary tract tumors	17 (1.33)	VAS, 5/17 (29.41)	EORTC QLQ-C30, 3/17 (17.65)	NRS, 2/17 (11.76)	QoR-15, 2/17 (11.76)	WHOQOL-BREF[w], 2/17 (11.76)
Hematologic malignant tumors	9 (0.7)	EORTC QLQ-C30, 2/9 (22.22)	NRS, 2/9 (22.22)	EQ-5D, 1/9 (11.11)	FACT-An[x], 1/9 (11.11)	FACT-Lym[y], 1/9 (11.11)
Bone tumors	8 (0.63)	EORTC QLQ-C30, 2/8 (25)	VAS, 2/8 (25)	NRS, 2/8 (25)	BPI, 1/8 (12.5)	SF-36[z], 1/8 (12.5)
Neuroendocrine tumors	10 (0.78)	EORTC QLQ-C30, 6/10 (60)	EORTC QLQ-PAN26[aa], 2/10 (20)	VAS, 2/10 (20)	N/A[ab]	N/A
Skin tumors	3 (0.23)	EORTC QLQ-C30, 1/3 (33.33)	VAS, 1/3 (33.33)	HF-QoL[ac], 1/3 (33.33)	N/A	N/A
Ophthalmic tumors	1 (0.08)	EORTC QLQ-C30, 1/1 (100)	EORTC QLQ-OPT30[ad], 1/1 (100)	N/A	N/A	N/A
Soft tissue sarcoma	0 (0)	N/A	N/A	N/A	N/A	N/A
Other tumors	187 (14.61)	VAS, 44/187 (23.53)	NRS, 44/187 (23.53)	EORTC QLQ-C30, 33/187 (17.65)	PSQI[ae], 15/187 (8.02)	BFI[af], 11/187 (5.88)

[a]VAS: visual analog scale.

[b]EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30.

[c]NRS: numeric rating scale.

[d]FACT-L: Functional Assessment of Cancer Therapy–Lung.

[e]EORTC QLQ-LC13: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Lung Cancer 13.

XSL•FO
RenderX

[f]FACT-B: Functional Assessment of Cancer Therapy–Breast.
[g]EORTC QLQ-BR23: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Breast Cancer 23.
[h]EORTC QLQ-CR29: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Colorectal Cancer 29.
[i]EORTC QLQ-OES18: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Oesophageal Cancer 18.
[j]EORTC QLQ-STO22: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Stomach 22.
[k]EORTC QLQ-H&N35: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Head and Neck Cancer 35.
[l]PG-SGA: Patient-Generated Subjective Global Assessment.
[m]SDS: Self-Rating Depression Scale.
[n]EORTC QLQ-OV28: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Ovarian Cancer 28.
[o]EORTC QLQ-HCC18: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Hepatocellular Carcinoma 18.
[p]TCMSS: Traditional Chinese Medicine Symptom Scale.
[q]FACT-P: Functional Assessment of Cancer Therapy–Prostate.
[r]BPI: Brief Pain Inventory.
[s]IPSS: International Prostate Symptom Score.
[t]QoR-15: Quality of Recovery-15.
[u]QoR-40: Quality of Recovery-40.
[v]PCSQ: Preparedness for Colorectal Cancer Surgery Questionnaire.
[w]WHOQOL-BREF: World Health Organization Quality of Life Brief Version.
[x]FACT-An: Functional Assessment of Cancer Therapy–Anemia.
[y]FACT-Lym: Functional Assessment of Cancer Therapy–Lymphoma.
[z]SF-36: 36-item Short Form Health Survey.
[aa]EORTC QLQ-PAN26: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Pancreatic Cancer 26.
[ab]N/A: not applicable.
[ac]HF-QoL: Hand-Foot Skin Reaction and Quality of Life.
[ad]EORTC QLQ-OPT30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Ophthalmic Cancer 30.
[ae]PSQI: Pittsburgh Sleep Quality Index.
[af]BFI: Brief Fatigue Inventory.

Discussion

Principal Findings

This cross-sectional study depicted the general characteristics of adult tumor clinical trials incorporating PROs in China and analyzed the application of PRO instruments in randomized clinical trials of tumors to provide potential directions for future research and serve as a reference for tumor clinical practice. The findings revealed that a significant proportion, specifically 62.92% (4055/6445) of the included trials, missed the opportunity to capture patients' subjective evaluations. Of the trials with PRO instruments as end points, 26.82% (641/2390) used PRO instruments as primary outcomes, 52.72% (1260/2390) as secondary outcomes, and 20.46% (489/2390) as coprimary outcomes. The majority of PRO trials (2144/2390, 89.71%) used 1 to 3 PRO instruments. Given that PROs can authentically represent patients' subjective experiences and evaluations, they should receive heightened emphasis in the context of tumor clinical trials. However, in light of the small proportion of tumor-related randomized clinical trials assessing PROs, policy makers and standard-setting bodies are recommended to further promote the collection of PROs in such trials in China.

This study delved into the yearly distribution of tumor clinical trials, indicating a notable surge in the use of PRO instruments as end points between January 1, 2010, and June 30, 2022. Among the trials incorporating PROs, early phase trials constituted the largest proportion (575/2390, 24.06%), followed by phase 3 (284/2390, 11.88%) and phase 4 (269/2390, 11.26%) trials. A retrospective cross-sectional study suggested a potential correlation between the use of PROs in late-stage trials and improved drug outcomes, such as overall survival [25]. However, the omission of PROs in late-stage trial results may reduce the value of patient participation in these trials. Previous work has shown that the concern regarding funding for PRO research seems significant, and additional funding was needed—and considered important—to pay for the use of PRO instruments to collect relevant data [26]. This may also be the reason why, among the included studies, there were few PRO tumor trials funded by industry. Relevant policies could provide more financial support for PRO tumor trials. In addition, our study indicated that the application of PRO instruments was more prevalent in trials involving drug interventions. PRO instruments can serve as valuable tools for assessing patient experiences during treatment, which is an essential aspect of drug discovery [27], and their absence can result in the exclusion of critical information, such as opportunities for patient-centered support programs and insights into benefit-risk profiles [27].

In accordance with prior research [23], our study also identified regional differences in the use of PROs. Tumor trials were more prevalent in the eastern, northern, and southern regions of China—especially in Shanghai, Beijing, Guangdong, and Jiangsu—and the adoption of PRO measurements followed a similar pattern. Conversely, in other regions of China, especially in the northwestern and northeastern regions—such as Qinghai, Tibet and Heilongjiang—both the overall number of tumor clinical trials and those incorporating PRO instruments as end

XSL•FO
RenderX

points were conspicuously lower. These results indicated the relationship between the volume of tumor clinical trials and the adoption of PRO tools. In addition, other factors such as economic conditions and medical resources also played an important role in this phenomenon [28]. Relevant policies can continue to encourage medical resources to be tilted toward rural and less developed areas. Remarkably, the study suggested that in resource-constrained remote regions, simplified applications of PRO instruments may be considered in tumor clinical trials. Moreover, our investigation revealed a lower prevalence of industry-funded trials in tumor clinical trials in China. This discrepancy may be attributed to previous findings that tumor trials were characterized by increased risk and costliness [29].

This study further found that thorax tumors, breast tumors, and lower gastrointestinal tract tumors were the most common conditions in trials with explicit PRO instruments. This might be related to variances in tumor incidence and different clinical concerns [30]. In the primary and coprimary outcome trial sets, a higher proportion of trials explicitly listed the PRO instruments as end points compared to those not specifying PROs, underscoring the normative inclination to formalize the acquisition and application of PRO instruments. Adherence to guidelines and standardization of PRO application is essential to maximize the application of PRO trial data, enhance their impact, and minimize research waste [31]. In particular, studies have shown that the standardized PROs were conducive to making trials or clinical treatments more scientifically rigorous and ethically sound [32-35]. Therefore, the need to standardize the application of PRO instruments remains important, with an increased emphasis on explicitly specifying PRO instruments in clinical trials.

This study analyzed the frequency of the use of PRO instruments in different classifications of trials by medical condition and found that the VAS and the NRS were the most commonly used in trials where PROs were designated as coprimary outcomes. Meanwhile, in all trials that used PRO instruments as outcomes, the VAS and the NRS were consistently prevalent. This prevalence can be attributed to the precision, simplicity, and sensitivity of VAS scores, as well as the ease of use and standardized format of the NRS for assessing subjective indicators [36-38]. In addition, almost 90% of patients with cancer would experience pain during the course of their illness [39]. The pain is both prevalent and burdensome for patients, but there is a lack of objective evaluation indices available for this purpose [40,41]. Consequently, the VAS emerged as the preferred choice for pain assessment in clinical research. Similarly, the NRS, with its user-friendly nature and standardized format, has been the preferred tool for pain assessment [36-38]. PROs continue to represent the gold standard for evaluating patients' core pain outcomes [42-44]. In this study, among the trials that used PRO instruments as secondary outcomes, the EORTC QLQ-C30 was the most commonly used (223/606, 36.8%), which might be attributed to the significance of addressing quality-of-life concerns for patients with tumors. This study also scrutinized the prevalent PRO instruments used in various medical conditions and found that the quality-of-life scale was frequently used in clinical trials involving tumors. The high frequency of the EORTC QLQ-C30 and FACT scale groups underscored the widespread application of these instruments in assessing patients' quality of life in cancer clinical trials in China. Specific modules in the EORTC QLQ scale system, such as the EORTC QLQ–Breast Cancer 23, the EORTC QLQ–Lung Cancer 13, and the EORTC QLQ–Colorectal Cancer 29, have been widely used in various cancer diseases [45,46]. Similarly, specific modules in the FACT scales, such as FACT–Lung (lung cancer), FACT–Breast (breast cancer), and FACT–Prostate (prostate cancer), have exhibited a high rate of use in cancer clinical trials in China. The extensive use of various PRO scales indicates a growing awareness and acceptance of PRO instruments, which, in turn, encourages the development of more effective and reliable PRO instruments. PRO instruments can be divided into universal and disease-specific PRO instruments. Considering the heterogeneity of symptom types in patients with tumors, symptom assessment should be performed for specific diseases [47]. However, in different tumor trials, the explicitly specified PRO instruments were primarily quality-of-life scales, the VAS, and the NRS, suggesting a need for the application of disease-specific PRO scales for different tumor types in clinical trials. It is suggested that according to the heterogeneity of diseases, experts from different fields should be brought together to develop or improve the disease-specific scale through participatory and consensus approaches under the guidance of relevant guidelines [33,47,48]. Acceptance of the scale by a wide range of stakeholders would be beneficial to improve the quality and specificity of the scale [48]. Training of clinicians and researchers on disease-specific scales is recommended. In addition, regarding the implementation of PRO measurement, it can be attempted as part of routine clinical care delivery for corresponding diseases, as well as continuous quality improvement as a clinical care priority [48].

This study undertook an in-depth analysis of the fundamental aspects of tumor clinical trials encompassing PROs in China, involving categorizing tumors and assessing the application of specific PRO tools for each tumor type. The findings underscore the critical importance of integrating PRO measures into tumor clinical trials in China and the need to standardize the use of PRO instruments within these trials. In recent years, the Chinese government has attached great importance to the application of PRO instruments in clinical trials. To encourage the patient-centered concept of new drug development and make reasonable use of PRO instruments, the National Medical Products Administration formulated the *Guiding Principles for the Application of Patient Reported Outcomes in Drug Clinical Research and Development* in 2022. To further promote these guiding principles, the relevant departments can educate researchers about the importance of regulating the application of PRO instruments, provide an interpretation of these principles to researchers, and advise them to follow the guidelines. We encourage researchers to communicate relevant information to regulators in a timely manner to conduct higher-quality clinical trials, such as the background of the study, the type of study, and the scale used. Policy makers should further formulate and implement pertinent policies, and PRO application platforms need to be developed and promoted to accelerate rational use of PROs in tumor clinical trials. It is recommended to define

XSL•FO
RenderX

or form an institution or department to coordinate and standardize the use of PROs in clinical trials [49]. The institution or department can provide researchers with some support, such as methodological guidance for PRO applications, interpretation of relevant guidelines, and guidance on internet technologies. Efforts should also be made to encourage communication and collaboration among policy makers, researchers, and medical institutions to promote the high-quality application of PROs in clinical trials. Furthermore, it is crucial to train clinicians in how to use PRO instruments in clinical practice. Ideally, this training can be part of standard medical education programs in the future. The most successful and effective way of training involved real patient cases and problem-based learning using audio and video clips, which could enable clinicians to know how to use PRO instruments and refer to the PRO data [50]. Researchers are encouraged to follow relevant guidelines and principles and actively engage in conducting high-quality tumor clinical trials to improve well-established PRO protocols and enrich the array of available PRO instruments, thereby advancing personalized population health. In addition, it is suggested to encourage and provide relevant support to patients who have difficulties in completing the PRO reports [51].

Limitations

It is important to acknowledge several limitations to this study. First, we excluded trials lacking detailed end point information, which may have introduced bias into the results. Second, the inclusion of trials that have not yet commenced participant recruitment, although necessary for our investigation, may have inflated the reported outcomes. Finally, the exclusion of trials involving children due to their limited expressive ability and the potential influence of parental reporting on outcomes may have introduced bias in the findings.

Conclusions

In China, the incorporation of PROs has demonstrated an upward trajectory in adult randomized clinical trials of tumors in recent years. Nonetheless, the infrequent measurement of the patient's voice remains noteworthy. This study highlights the need for a more comprehensive evaluation of patients' experiences in adult tumor clinical trials in China. The incorporation of patients' subjective feelings in the context of tumor diseases is necessary. Disease-specific PRO instruments should be widely used in different categories of tumor disease. Pertinent policies should be formulated and implemented, and PRO application platforms need to be developed and promoted as well. In addition, researchers should actively engage in conducting high-quality tumor clinical trials. There is room for improvement in the standardization of PROs in China.

Acknowledgments

This work was supported by the high-level traditional Chinese Medicine Key Subjects Construction Project of the National Administration of Traditional Chinese Medicine—Evidence-Based Traditional Chinese Medicine (zyyzdxk-2023249).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HR and JL conceived of the presented idea. YJ and QL coordinated the data collection and analysis. XW, YY, and YJ performed the data extraction. YJ and QL wrote the first draft of the paper; and SY, SL, WL, and XW provided inputs for subsequent drafts. JL and HR provided comments related to the presentation of the findings and critically reviewed the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Classification of specific diseases.
[PDF File (Adobe PDF File), 114 KB - jmir_v26i1e45719_app1.pdf]

Multimedia Appendix 2
The number of trials with patient-reported outcomes in each province of China.
[PNG File , 232 KB - jmir_v26i1e45719_app2.png]

Multimedia Appendix 3
Patient-reported outcome tests used most frequently.
[PDF File (Adobe PDF File), 107 KB - jmir_v26i1e45719_app3.pdf]

XSL•FO
RenderX

References

1. Patrick DL, Burke LB, Powers JH, Scott JA, Rock EP, Dawisha S, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 2007;10 Suppl 2:S125-S137 [FREE Full text] [doi: 10.1111/j.1524-4733.2007.00275.x] [Medline: 17995471]
2. Calvert M, Kyte D, Mercieca-Bebber R, Slade A, Chan AW, King MT, the SPIRIT-PRO Group, et al. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO extension. JAMA 2018 Feb 06;319(5):483-494 [FREE Full text] [doi: 10.1001/jama.2017.21903] [Medline: 29411037]
3. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013 Feb 27;309(8):814-822. [doi: 10.1001/jama.2013.879] [Medline: 23443445]
4. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006 Oct 11;4:79 [FREE Full text] [doi: 10.1186/1477-7525-4-79] [Medline: 17034633]
5. The Lancet Neurology. Patient-reported outcomes in the spotlight. Lancet Neurol 2019 Nov;18(11):981. [doi: 10.1016/s1474-4422(19)30357-6]
6. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. J Eval Clin Pract 2006 Oct;12(5):559-568. [doi: 10.1111/j.1365-2753.2006.00650.x] [Medline: 16987118]
7. Greenhalgh J, Gooding K, Gibbons E, Dalkin S, Wright J, Valderas J, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. J Patient Rep Outcomes 2018 Dec;2:42 [FREE Full text] [doi: 10.1186/s41687-018-0061-6] [Medline: 30294712]
8. Chan G, Bezuidenhout L, Walker L, Rowan R. The Impact on Life questionnaire: validation for elective surgery prioritisation in New Zealand prioritisation criteria in orthopaedic surgery. N Z Med J 2016 Apr 01;129(1432):26-32. [Medline: 27356249]
9. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013 Jan 28;346:f167 [FREE Full text] [doi: 10.1136/bmj.f167] [Medline: 23358487]
10. Tong A, Oberbauer R, Bellini MI, Budde K, Caskey FJ, Dobbels F, et al. Patient-reported outcomes as endpoints in clinical trials of kidney transplantation interventions. Transpl Int 2022 May 20;35:10134 [FREE Full text] [doi: 10.3389/ti.2022.10134] [Medline: 35669971]
11. Qian Y, Walters SJ, Jacques R, Flight L. Comprehensive review of statistical methods for analysing patient-reported outcomes (PROs) used as primary outcomes in randomised controlled trials (RCTs) published by the UK's Health Technology Assessment (HTA) journal (1997-2020). BMJ Open 2021 Sep 06;11(9):e051673 [FREE Full text] [doi: 10.1136/bmjopen-2021-051673] [Medline: 34489292]
12. Dai W, Feng W, Zhang Y, Wang XS, Liu Y, Pompili C, et al. Patient-reported outcome-based symptom management versus usual care after lung cancer surgery: a multicenter randomized controlled trial. J Clin Oncol 2022 Mar 20;40(9):988-996 [FREE Full text] [doi: 10.1200/JCO.21.01344] [Medline: 34995100]
13. Kouzy R, Abi Jaoude J, Lin D, Nguyen ND, El Alam MB, Ludmir EB, et al. Patient-reported outcome measures in pancreatic cancer receiving radiotherapy. Cancers (Basel) 2020 Sep 02;12(9):2487 [FREE Full text] [doi: 10.3390/cancers12092487] [Medline: 32887363]
14. Singhal U, Skolarus TA, Gore JL, Parry MG, Chen RC, Nossiter J, et al. Implementation of patient-reported outcome measures into health care for men with localized prostate cancer. Nat Rev Urol 2022 May;19(5):263-279. [doi: 10.1038/s41585-022-00575-4] [Medline: 35260844]
15. Stover AM, Basch EM. Using patient-reported outcome measures as quality indicators in routine cancer care. Cancer 2016 Feb 01;122(3):355-357 [FREE Full text] [doi: 10.1002/cncr.29768] [Medline: 26619153]
16. Albreht T. Europe's beating cancer plan-a new step towards more comprehensive and equitable cancer control in Europe. Eur J Public Health 2021 Jul 13;31(3):456-457. [doi: 10.1093/eurpub/ckab014] [Medline: 34255853]
17. Lizán L, Pérez-Carbonell L, Comellas M. Additional value of patient-reported symptom monitoring in cancer care: a systematic review of the literature. Cancers (Basel) 2021 Sep 15;13(18):4615 [FREE Full text] [doi: 10.3390/cancers13184615] [Medline: 34572842]
18. Graupner C, Kimman ML, Mul S, Slok AH, Claessens D, Kleijnen J, et al. Patient outcomes, patient experiences and process indicators associated with the routine use of patient-reported outcome measures (PROMs) in cancer care: a systematic review. Support Care Cancer 2021 Feb;29(2):573-593 [FREE Full text] [doi: 10.1007/s00520-020-05695-4] [Medline: 32875373]
19. Engstrom T, Tanner S, Lee WR, Forbes C, Walker R, Bradford N, et al. Patient reported outcome measure domains and tools used among adolescents and young adults with cancer: a scoping review. Crit Rev Oncol Hematol 2023 Jan;181:103867. [doi: 10.1016/j.critrevonc.2022.103867] [Medline: 36427770]
20. Yang LY, Manhas DS, Howard AF, Olson RA. Patient-reported outcome use in oncology: a systematic review of the impact on patient-clinician communication. Support Care Cancer 2018 Jan;26(1):41-60. [doi: 10.1007/s00520-017-3865-7] [Medline: 28849277]

XSL•FO
RenderX

21. Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, et al. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. Lancet Oncol 2020 Feb;21(2):e83-e96. [doi: 10.1016/s1470-2045(19)30790-9]
22. Basch E, Geoghegan C, Coons SJ, Gnanasakthy A, Slagle AF, Papadopoulos EJ, et al. Patient-reported outcomes in cancer drug development and US regulatory review: perspectives from industry, the food and drug administration, and the patient. JAMA Oncol 2015 Jun 01;1(3):375-379. [doi: 10.1001/jamaoncol.2015.0530] [Medline: 26181187]
23. Zhou H, Yao M, Gu X, Liu M, Zeng R, Li Q, et al. Application of patient-reported outcome measurements in clinical trials in China. JAMA Netw Open 2022 May 02;5(5):e2211644 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.11644] [Medline: 35544134]
24. Gnanasakthy A, Norcross L, DeMuro Romano C, Carson RT. A review of patient-reported outcome labeling of FDA-approved new drugs (2016-2020): counts, categories, and comprehensibility. Value Health 2022 Apr;25(4):647-655 [FREE Full text] [doi: 10.1016/j.jval.2021.10.006] [Medline: 35365309]
25. Haslam A, Herrera-Perez D, Gill J, Prasad V. Patient experience captured by quality-of-life measurement in oncology clinical trials. JAMA Netw Open 2020 Mar 02;3(3):e200363 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0363] [Medline: 32129865]
26. Watkins Bruner D, Bryan CJ, Aaronson N, Blackmore CC, Brundage M, Cella D, et al. Issues and challenges with integrating patient-reported outcomes in clinical trials supported by the National Cancer Institute–sponsored clinical trials networks. J Clin Oncol 2007 Nov 10;25(32):5051-5057. [doi: 10.1200/jco.2007.11.3324]
27. Basch E, Dueck AC. Patient-reported outcome measurement in drug discovery: a tool to improve accuracy and completeness of efficacy and safety data. Expert Opin Drug Discov 2016 Aug;11(8):753-758. [doi: 10.1080/17460441.2016.1193148] [Medline: 27310432]
28. Liniker E, Harrison M, Weaver JM, Agrawal N, Chhabra A, Kingshott V, et al. Treatment costs associated with interventional cancer clinical trials conducted at a single UK institution over 2 years (2009-2010). Br J Cancer 2013 Oct 15;109(8):2051-2057 [FREE Full text] [doi: 10.1038/bjc.2013.495] [Medline: 24064969]
29. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics 2019 Apr 01;20(2):273-286 [FREE Full text] [doi: 10.1093/biostatistics/kxx069] [Medline: 29394327]
30. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl) 2021 Mar 17;134(7):783-791 [FREE Full text] [doi: 10.1097/CM9.0000000000001474] [Medline: 33734139]
31. Rivera SC, Kyte DG, Aiyegbusi OL, Slade AL, McMullan C, Calvert MJ. The impact of patient-reported outcome (PRO) data from clinical trials: a systematic review and critical analysis. Health Qual Life Outcomes 2019 Oct 16;17(1):156 [FREE Full text] [doi: 10.1186/s12955-019-1220-z] [Medline: 31619266]
32. Di Maio M, Gallo C, Leighl NB, Piccirillo MC, Daniele G, Nuzzo F, et al. Symptomatic toxicities experienced during anticancer treatment: agreement between patient and physician reporting in three randomized trials. J Clin Oncol 2015 Mar 10;33(8):910-915 [FREE Full text] [doi: 10.1200/JCO.2014.57.9334] [Medline: 25624439]
33. Howell D, Molloy S, Wilkinson K, Green E, Orchard K, Wang K, et al. Patient-reported outcomes in routine cancer clinical practice: a scoping review of use, impact on health outcomes, and implementation factors. Ann Oncol 2015 Sep;26(9):1846-1858 [FREE Full text] [doi: 10.1093/annonc/mdv181] [Medline: 25888610]
34. Crites JS, Chuang C, Dimmock A, Hwang W, Johannes B, Paranjape A, et al. PROs in the balance: ethical implications of collecting patient reported outcome measures in the electronic health record. Am J Bioeth 2016 Mar 16;16(4):67-68. [doi: 10.1080/15265161.2016.1146543] [Medline: 26982933]
35. Kyte D, Draper H, Calvert M. Patient-reported outcome alerts: ethical and logistical considerations in clinical trials. JAMA 2013 Sep 25;310(12):1229-1230. [doi: 10.1001/jama.2013.277222] [Medline: 24065005]
36. Jensen MP. The validity and reliability of pain measures in adults with cancer. J Pain 2003 Feb;4(1):2-21 [FREE Full text] [doi: 10.1054/jpai.2003.1] [Medline: 14622723]
37. Caraceni A, Brunelli C, Martini C, Zecca E, De Conno F. Cancer pain assessment in clinical trials. A review of the literature (1999-2002). J Pain Symptom Manage 2005 May;29(5):507-519 [FREE Full text] [doi: 10.1016/j.jpainsymman.2004.08.014] [Medline: 15904753]
38. Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, et al. Studies comparing Numerical Rating Scales, Verbal Rating Scales, and Visual Analogue Scales for assessment of pain intensity in adults: a systematic literature review. J Pain Symptom Manage 2011 Jun;41(6):1073-1093 [FREE Full text] [doi: 10.1016/j.jpainsymman.2010.08.016] [Medline: 21621130]
39. Hjermstad MJ, Fainsinger R, Kaasa S, European Palliative Care Research Collaborative (EPCRC). Assessment and classification of cancer pain. Curr Opin Support Palliat Care 2009 Mar;3(1):24-30. [doi: 10.1097/SPC.0b013e3283260644] [Medline: 19365158]
40. Santoni A, Santoni M, Arcuri E. Chronic cancer pain: opioids within tumor microenvironment affect neuroinflammation, tumor and pain evolution. Cancers (Basel) 2022 Apr 30;14(9):2253 [FREE Full text] [doi: 10.3390/cancers14092253] [Medline: 35565382]

XSL•FO
RenderX

41. Morishita S, Hirabayashi R, Tsubaki A, Aoki O, Fu JB, Onishi H, et al. Relationship between balance function and QOL in cancer survivors and healthy subjects. Medicine (Baltimore) 2021 Nov 19;100(46):e27822 [FREE Full text] [doi: 10.1097/MD.0000000000027822] [Medline: 34797310]
42. Comer SD, Zacny JP, Dworkin RH, Turk DC, Bigelow GE, Foltin RW, et al. Core outcome measures for opioid abuse liability laboratory assessment studies in humans: IMMPACT recommendations. Pain 2012 Dec;153(12):2315-2324 [FREE Full text] [doi: 10.1016/j.pain.2012.07.035] [Medline: 22998781]
43. Dworkin RH, Turk DC, McDermott MP, Peirce-Sandner S, Burke LB, Cowan P, et al. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. Pain 2009 Dec;146(3):238-244. [doi: 10.1016/j.pain.2009.08.019] [Medline: 19836888]
44. Younger J, McCue R, Mackey S. Pain outcomes: a brief review of instruments and techniques. Curr Pain Headache Rep 2009 Feb 23;13(1):39-43 [FREE Full text] [doi: 10.1007/s11916-009-0009-x] [Medline: 19126370]
45. Wan C, Meng Q, Yang Z, Tu X, Feng C, Tang X, et al. Validation of the simplified Chinese version of EORTC QLQ-C30 from the measurements of five types of inpatients with cancer. Ann Oncol 2008 Dec;19(12):2053-2060 [FREE Full text] [doi: 10.1093/annonc/mdn417] [Medline: 18653703]
46. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 03;85(5):365-376. [doi: 10.1093/jnci/85.5.365] [Medline: 8433390]
47. Taylor F, Reasner DS, Carson RT, Deal LS, Foley C, Iovin R, et al. Development of a symptom-based patient-reported outcome instrument for functional dyspepsia: a preliminary conceptual model and an evaluation of the adequacy of existing instruments. Patient 2016 Oct 28;9(5):409-418 [FREE Full text] [doi: 10.1007/s40271-016-0164-1] [Medline: 27020447]
48. Howell D, Fitch M, Bakker D, Green E, Sussman J, Mayo S, et al. Core domains for a person-focused outcome measurement system in cancer (PROMS-Cancer Core) for routine care: a scoping review and Canadian Delphi Consensus. Value Health 2013;16(1):76-87 [FREE Full text] [doi: 10.1016/j.jval.2012.10.017] [Medline: 23337218]
49. Eton DT, Beebe TJ, Hagen PT, Halyard MY, Montori VM, Naessens JM, et al. Harmonizing and consolidating the measurement of patient-reported information at health care institutions: a position statement of the Mayo Clinic. Patient Relat Outcome Meas 2014 Feb 10;5:7-15 [FREE Full text] [doi: 10.2147/PROM.S55069] [Medline: 24550683]
50. Santana MJ, Haverman L, Absolom K, Takeuchi E, Feeny D, Grootenhuis M, et al. Training clinicians in how to use patient-reported outcome measures in routine clinical practice. Qual Life Res 2015 Jul;24(7):1707-1718 [FREE Full text] [doi: 10.1007/s11136-014-0903-5] [Medline: 25589231]
51. Riedl D, Lehmann J, Rothmund M, Dejaco D, Grote V, Fischer MJ, et al. Usability of electronic patient-reported outcome measures for older patients with cancer: secondary analysis of data from an observational single center study. J Med Internet Res 2023 Sep 21;25:e49476 [FREE Full text] [doi: 10.2196/49476] [Medline: 37733409]

Abbreviations

CONSORT-PRO: Consolidated Standards of Reporting Trials–Patient-Reported Outcome
EORTC QLQ: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire
EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire-30
FACT: Functional Assessment of Cancer Therapy
PRO: patient-reported outcome
PRO-CTCAE: Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events
SISAQOL: Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints
SPIRIT-PRO: Standard Protocol Items: Recommendations for Interventional Trials–Patient-Reported Outcome
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 14.01.23; peer-reviewed by Y Chu, L Guo; comments to author 24.10.23; revised version received 29.10.23; accepted 09.02.24; published 08.05.24.

Please cite as:
Jia Y, Li Q, Zhang X, Yan Y, Yan S, Li S, Li W, Wu X, Rong H, Liu J
Application of Patient-Reported Outcome Measurements in Adult Tumor Clinical Trials in China: Cross-Sectional Study
J Med Internet Res 2024;26:e45719
URL: https://www.jmir.org/2024/1/e45719
doi:10.2196/45719
PMID:38718388

XSL•FO
RenderX

©Yan Jia, Qi Li, Xiaowen Zhang, Yi Yan, Shiyan Yan, Shunping Li, Wei Li, Xiaowen Wu, Hongguo Rong, Jianping Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of Patient-Reported Outcomes in a Medical Oncology Setting (the iPROMOS Study): Type II Hybrid Implementation Study

Natasha Anne Roberts[1,2], BN, PhD; Anita Pelecanos[3], BSc, BHSc; Kimberly Alexander[4], BHlthSc, BNurs, PhD; David Wyld[5,6], MBBS; Monika Janda[1], PhD

[1]The University of Queensland Centre for Clinical Research, Herston, Australia
[2]Surgical Treatment and Rehabilitation Service Metro North Health and University of Queensland, Herston, Australia
[3]QIMR Berghofer Medical Research Institute, Herston, Australia
[4]School of Nursing, Queensland University of Technology, Kelvin Grove, Australia
[5]Cancer Care Services, Royal Brisbane and Women's Hospital, Herston, Australia
[6]The University of Queensland Clinical School, St Lucia, Australia

Corresponding Author:
Natasha Anne Roberts, BN, PhD
The University of Queensland Centre for Clinical Research
Herston Road
Herston, 4029
Australia
Phone: 61 7 36468100
Email: natasha.roberts@uq.edu.au

Abstract

Background: Clinical trials have demonstrated that patient-reported outcome measures (PROMs) can improve mortality and morbidity outcomes when used in clinical practice.

Objective: This study aimed to prospectively investigate the implementation of PROMs in routine oncology. Outcomes measured included improved symptom detection, clinical response to symptom information, and health service outcomes.

Methods: Two of 12 eligible clinics were randomized to implement symptom PROMs in a medical oncology outpatient department in Australia. Randomization was carried out at the clinic level. Patients in control clinics continued with usual care; those in intervention clinics completed a symptom PROM at presentation. This was a pilot study investigating symptom detection, using binary logistic models, and clinical response to PROMs investigated using multiple regression models.

Results: A total of 461 patient encounters were included, consisting of 242 encounters in the control and 222 in the intervention condition. Patients in these clinics most commonly had head and neck, lung, prostate, breast, or colorectal cancer and were seen in the clinic for surveillance and oral or systemic treatments for curative, metastatic, or palliative cancer care pathways. Compared with control encounters, the proportion of symptoms detected increased in intervention encounters (odds ratio 1.05, 95% CI 0.99-1.11; *P*=.08). The odds of receiving supportive care, demonstrated by nonroutine allied health review, increased in the intervention compared with control encounters (odds ratio 3.54, 95% CI 1.26-9.90; *P*=.02).

Conclusions: Implementation of PROMs in routine care did not significantly improve symptom detection but increased the likelihood of nonroutine allied health reviews for supportive care. Larger studies are needed to investigate health service outcomes.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12618000398202; https://tinyurl.com/3cxbemy4

(J Med Internet Res 2024;26:e55841) doi:10.2196/55841

KEYWORDS

implementation science; iPARIHS; clinical practice; intervention; implementation; facilitator; facilitation; patient-reported outcomes, patient-reported outcome measures; oncology; symptom; symptoms; detection; investigate; service; services; clinic; clinics; Australia; binary logistic models; regression model; regression models; patient; patients; supportive care

XSL•FO
RenderX

Introduction

The use of symptom patient-reported outcome measures (PROMs) in stringent randomized clinical trials has been shown to improve oncology outpatient care outcomes. Randomized clinical trials reported on individual patient outcomes such as improved physical functioning, better treatment adherence, self-efficacy, quality of life, and overall survival [1-4]. For health services, the use of PROMs for symptom measurement has led to a reduction in emergency presentations and hospital admissions, without increasing clinical workloads [5-7]. These benefits are thought to be due to improved symptom detection by clinicians enabling better supportive care [8-10].

Despite this evidence and an increased uptake of PROMs in routine care, successful implementation in routine care has been inconsistent due to the complex nature of PROMs interventions [4]. More than 2 decades ago, the review by Greenhalgh and Meadows [11] identified that PROMs are challenging to implement, and this was a persisting concern. Recently, guidelines from the European Society for Medical Oncology identified that evidence informing implementation of PROMs in clinical practice is still scarce and further studies that shed light on the barriers to implementation are needed [4].

In 2017, a prospective pilot study (iPROMOS) was designed to investigate the feasibility of investigating implementation of PROMs in medical oncology outpatient care using a type II hybrid implementation clinical trial method [12]. This study structured its implementation approach around the recommendations of the integrated-Promoting Action on Research Implementation in Health Services (iPARIHS) implementation science framework and found that despite some challenges, implementing PROMs in routine oncology clinics was feasible and acceptable [13]. It also showed that it is important to measure underlying processes to ensure that the intervention was delivered in the way it is intended [14]. It was proposed that since PROMs were successfully implemented, clinical benefits such as improved symptom detection and corresponding supportive care would occur. This paper reports on these clinical outcome results of the iPROMOS study.

Methods

Study Design

A hybrid feasibility implementation clinical trial was designed to measure both intervention and implementation outcomes [15]. The trial randomized at the level of the outpatient clinic with preimplementation data compared with that of postimplementation data across 2 clinics, with weeks and clinic as fixed effects [12]. The introduction of the intervention was phased for practical reasons, mainly to limit the demand on local resources. The purpose of this study design was to investigate implementation of PROMs taking into consideration the learnings of previous research identifying randomization at the level of the patient was not optimal [5,11,16]. Implementation design features were structured following the iPARIHS framework [17].

The intervention (defined as the "innovation construct" according to the iPARIHS framework) contained 3 components:

1. Completion of the PROM (the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events [PRO-CTCAE] [18] core set [19]) by patients via a touch screen computer when visiting the outpatient department. The PRO-CTCAE was provided for patient completion using REDCap (Research Electronic Data Capture) electronic survey and data capture web application hosted at the Queensland University of Technology [20,21].

The PRO-CTCAE core set evaluates symptoms of fatigue, insomnia, pain, anorexia, dyspnea, anxiety, depression, peripheral sensory neuropathy, constipation, diarrhea, nausea, and vomiting, generating a summary symptom report of all responses by the patient from that encounter.

1. A generated summary of the patient PROM information was made available to medical, nursing, and other clinical teams who reviewed the patient as a part of the patient encounter. Making the report available was integrated into existing workflows.
2. The response by clinical teams to the generated summary of PROM information was documented on a case report form and the patient's medical record.

The facilitator supported the implementation of the intervention, including being available in-person to support patients to complete the PROM, and staff to obtain the PROM report, if needed. Facilitation actions have been reported as a part of the study findings [13].

This study took place in the medical oncology outpatient department in a large tertiary teaching and training hospital in Southeast Queensland, within the largest public health service in Australia. The hospital serves as a quaternary referral center with specialist medical oncology, hematology, radiation oncology, surgical oncology specialties, medical imaging, nuclear medicine, pathology services, and a high-acuity clinical research unit.

All patients attending the included clinics were eligible to participate if they provided informed consent. At this hospital, 12 medical oncology clinics were potentially eligible, of which 2 clinics were randomly chosen for this study, using a random draw of clinics names in blank envelopes, first by day of the week (so that the study took place on the same day) and then the clinic type. During the preintervention data collection phase both clinics continued to provide usual care (staff noted symptoms or adverse events for their patients that were discovered during anamnesis in the medical chart) [22]. For each patient encounter, staff members documented patient symptoms and symptom details, and whether a patient had a consultation with allied health in the patient chart, a case report was offered initially but this did not get used. This study was conducted over 20 calendar weeks. Clinic 1 then started with the intervention at week 5, while clinic 2 continued with usual care for another 3 months until week 17 and then started the same intervention [14].

The primary outcome was the proportion of doctors and nursing staff documenting symptoms during each patient encounter. A

XSL•FO
RenderX

sample size calculation used estimated effect sizes as published by Berry et al [23]. Given a baseline symptom detection level of 75%, it was estimated that 500 participant encounters would be needed to show improvement by 10% or more with 80% power. This was a feasibility study, and it was not to focus on a difference between a particular outcome and having it powered based on this end point.

Secondary outcomes were the proportion of patient encounters with a response to PROMs information, including whether they received additional supportive care by allied health staff (recorded as seen by allied health); proportion of patient encounters that proceeded to an emergency department presentation; and subsequent hospital admissions.

Exploratory outcomes were added to the analysis. These included (1) an analysis of whether the clinician type (doctor or nurse) had an influence on the clinicians' response to symptom information, and (2) an analysis of any emergent care, other than an emergency presentation. A sensitivity analysis was performed to assess whether public holidays that fell at calendar weeks 8, 13, and 16 had any effect on findings.

Clinic characteristics and outcomes were summarized using descriptive statistics, including frequency and percentages. To assess whether the intervention resulted in a change to the proportion of symptoms assessed by clinicians and additional supportive care consultations, binary logistic regression models were used. Three separate univariable models were fitted for intervention, week (as continuous measure) and clinic, and all 3 clinician types were included in a multivariable model. *P* values less than .05 were indicated as statistically significant. A fixed effect for clinic was chosen for better statistical efficiency.

Emergent care (patient-initiated ad hoc phone calls to clinics, cancer care coordinators, or additional treatments through outpatients), emergency presentations, and hospital admissions were presented as counts and proportions. The study protocol has been published [12].

Ethical Considerations

The iPROMOS study protocol underwent full review by the Royal Brisbane & Women's Hospital Human Research Ethics Committee (HREC/17/QRBW/416). Participants completed PROMs after providing informed consent via the touch screen computer. Participation was voluntary. All participant data were made nonidentifiable and stored on password-protected hospital servers.

Results

The consort diagram is presented in Figure 1. There were 464 patient encounters from 421 patients recruited between March and June 2018. The majority of patients (421/464, 91%) had only 1 encounter, while 15 patients had more than 1 (2-7) encounter.

Patient characteristics by intervention period and clinic are shown in Table 1. Reasons for clinic attendance included new patient appointments, intravenous treatments (including standard of care and clinical trials), oral treatments (including standard of care and clinical trials), follow-up, or surveillance for either localized or metastatic disease.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram for the iPROMOS study.

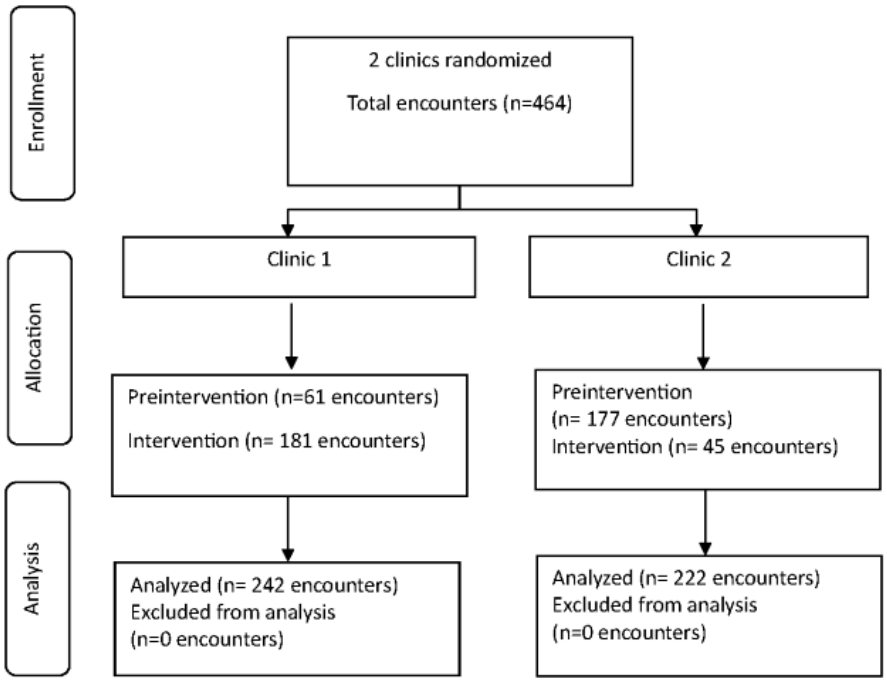

XSL•FO
RenderX

Table 1. Patient encounter characteristics by intervention phase and clinic.

	Preintervention (n=242)		Intervention (n=222)	
	Clinic 1 (n=61)	Clinic 2 (n=181)	Clinic 1 (n=177)	Clinic 2 (n=45)
Treatment, n (%)				
None	2 (3)	1 (1)	1 (1)	0 (0)
New	5 (8)	27 (15)	26 (15)	2 (4)
Intravenous	43 (70)	91 (50)	99 (56)	23 (51)
Intravenous clinical trial	2 (3)	2 (1)	23 (13)	4 (9)
Oral	1 (2)	18 (10)	6 (3)	6 (13)
Oral clinical trial	0 (0)	0 (0)	1 (1)	1 (2)
Follow-up	7 (11)	34 (19)	16 (9)	8 (18)
Surveillance	1 (2)	8 (4)	5 (3)	1 (2)
Stage, n (%)				
Localized	44 (75)	82 (46)	120 (69)	19 (42)
Metastatic	14 (24)	95 (53)	50 (29)	26 (58)
Recurrence	1 (2)	1 (1)	5 (3)	0 (0)
Palliative	0 (0)	1 (1)	0 (0)	0 (0)
Diagnosis, n (%)				
Thyroid	4 (7)	2 (1)	18 (10)	0 (0)
Lung	8 (13)	1 (1)	41 (23)	2 (4)
Colorectal	0 (0)	4 (2)	0 (0)	0 (0)
Breast	0 (0)	50 (28)	0 (0)	10 (22)
Prostate	0 (0)	68 (38)	0 (0)	22 (49)
Gynecologic	0 (0)	14 (8)	0 (0)	5 (11)
Head and neck	49 (80)	0 (0)	117 (66)	0 (0)
Genitourinary	0 (0)	42 (23)	1 (1)	6 (13)

For the primary outcome of symptom detection, 125 of 242 (52%) recorded patient encounters with a doctor had a symptom detected during the preintervention period. During the intervention period, 137 of 222 (62%) patient encounters with doctors had a symptom detected (*P*=.08). For encounters with a nurse, 43 of 97 (44%) identified a symptom before the intervention. During the intervention period, 99 of 130 (76%) documented encounters recorded a symptom detected (*P*=.004).

There was an increase of 3.54 (95% CI 1.26-9.90; *P*=.02) fold in the odds of being seen by allied health in the intervention period compared with the preintervention period after accounting for study week and clinic. A sensitivity analysis of considering week as calendar week instead of study week number (Table S1 in Multimedia Appendix 1) yielded similar results.

Exploratory analyses of symptom detection by clinician type (Table 2) identified that 125 of 242 (52%) recorded patient encounters with a doctor had a symptom detected during the preintervention period, compared with 137 of 222 (62%) patient encounters with doctors during the intervention period. For encounters with a nurse, 43 of 97 (44%) identified a symptom during the preintervention period, and 99 of 130 (76%) identified a symptom during the intervention period. Exploration of models including clinician type indicated that when a doctor identified a symptom (but not nurse) the odds that the patients would be seen by allied health increased during the intervention period (odds ratio 2.32, 95% CI 1.19-4.52; *P*=.013) (Table S2 in Multimedia Appendix 1).

Compared to the preintervention period, emergent care through the service line increased, while emergency presentations and hospital admissions decreased during the intervention period. Before the intervention, there were 57/242 (24%) encounters resulting in emergent care, 27/242 (12%) presenting to an emergency department, and 14/242 (1%) unplanned admissions to hospital. During the intervention period, 60/222 (27%) encounters resulted in required emergent care, 8/222 (0.5%) resulted in presentations to an emergency department, and 8/222 (0.5%) unplanned admissions to hospital.

XSL•FO
RenderX

Table 2. Summaries and logistic regression models of symptom identification or referral where week is defined as the study week number.

Total, N		Outcome, n (%)		Unadjusted OR[a,b] (95% CI)	*P* values	Adjusted OR[c] (95% CI)	*P* values
		No	Yes				
Outcome: doctor identified symptom							
Phase[d]							
Preintervention	242	117 (48)	125 (52)	Reference	.029	Reference	.43
Intervention	222	85 (38)	137 (62)	1.51 (1.04-2.18)		1.29 (0.68-2.44)	
Study week number	464			1.07 (1.03-1.11)	.001	1.05 (0.99-1.11)	.079
Clinic[d]							
1	238	106 (45)	132 (55)	Reference	.66	Reference	.41
2	226	96 (42)	130 (58)	1.09 (0.75-1.57)		1.24 (0.74-2.09)	
Outcome: nurse identified symptom							
Phase [e]							
Preintervention	97	54 (56)	43 (44)	Reference	<.001	Reference	.91
Intervention	130	31 (24)	99 (76)	4.01 (2.27-7.08)		0.95 (0.37-2.41)	
Study week number	227			1.13 (1.06-1.20)	<.001	1.14 (1.04-1.25)	.004
Clinic[e]							
1	153	41 (27)	112 (73)	Reference	<.001	Reference	<.001
2	74	44 (59)	30 (41)	0.25 (0.14-0.45)		0.22 (0.10-0.50)	
Outcome: seen by allied health							
Phase[f]							
Preintervention	242	231 (95)	11 (5)	Reference	<.001	Reference	.016
Intervention	222	179 (81)	43 (19)	5.04 (2.53-10.06)		3.54 (1.26-9.90)	
Study week number	464			0.99 (0.94-1.05)	.77	0.92 (0.84-1.01)	.069
Clinic[f]							
1	238	187 (79)	51 (21)	Reference	<.001	Reference	<.001
2	226	223 (99)	3 (1)	0.05 (0.02-0.16)		0.10 (0.03-0.36)	
Outcome: allied health identified symptom							
Phase[g]							
Preintervention	11	0 (0)	11 (100)	__[h]	—	—	—
Intervention	43	8 (19)	35 (81)	—	—	—	—
Study week number	54				—	—	—
Clinic[g]							
1	51	8 (16)	43 (84)	—	—	—	—
2	3	0 (0)	3 (100)	—	—	—	—

[a]Unadjusted ORs: each predictor runs as a separate model for each outcome.
[b]OR: odd ratio.
[c]Adjusted OR: all 3 predictors included in 1 model for each outcome.
[d]No: n=202; Yes: n=262.
[e]No: n=85; Yes: n=142.
[f]No: n=410; Yes: n=54.
[g]No: n=8; Yes: n=46.
[h]Not applicable.

XSL•FO
RenderX

Discussion

Our findings identified that the use of symptom PROMs in oncology outpatient care may potentially benefit clinical and health service outcomes, in line with other clinical studies. As a pilot study, we established that it would be feasible to conduct larger studies. Implementation outcomes were previously reported. We did not identify a statistically significant association between the intervention and symptom identification but there was an association between intervention and the clinical response to symptoms, that is, whether patients received a formal supportive care consultation by allied health. Participants in the intervention period had an increased odds of being seen by allied health compared to those in the preintervention period. This was more likely if a doctor identified a symptom but not if a nurse identified the symptom. While the sample was not powered to detect a significant difference [4], it appears that emergency presentations and unplanned hospital admissions reduced during the intervention period compared with the preintervention period. Emergent care by cancer care coordinators increased during the intervention period.

This study was designed to investigate the structured implementation of a PROMs intervention on clinical outcomes. Larger data sets from clinical practice cohort studies have proven effective in identifying health service benefits of PROMs intervention [6,24-26]. However, the mechanisms underlying these outcomes have not been clear. The use of an implementation science framework, iPARIHS [17], guided implementation and potentially brought to light the clinical mechanisms underlying outcomes. In addition, the use of randomization at the level of clinic aimed to mitigate selection or performance bias, which has been an identified concern in other study designs [5,11]. We identified that more allied health teams and cancer care coordinators were engaged through increased supportive care provision, offering a potential explanation for why the use of PROMs has impacts on health service outcomes.

In this study, symptom detection was not significantly different between the preintervention and intervention groups after the sensitivity analysis, even though the literature has broadly understood that PROMs improve symptom identification [8-10]. This result may be because medical assessment using CTCAE criteria [22] is routine practice for assessing adverse events in the unit and a requirement across the large number of clinical trials recruiting on an ongoing basis. What is also a unique aspect of the setting is cancer care coordinator provision of emergent care in line with an emergency avoidance model of care, where patients will be brought into the outpatients' department to avoid emergency presentations [27,28]. This service may have also contributed to allied health encounters and health service outcomes and brings attention to the multidisciplinary nature of cancer care.

The electronic system used for patient reporting and clinician review was basic in nature, but it also gave flexibility to tailor it to workflows, communication, and referral processes. The introduction of any technology into a clinical setting can be challenging. For PROMs, this literature supports the use of a simple design with ongoing quality improvement methods to ensure the intervention fidelity [29], which supports the approach taken in this study.

This was a small clinical trial, but it offers learnings for the future. The iPROMOS study highlights the need to understand how to build from a small intense facilitated implementation for scale-up and sustainability. Economic analyses of implementation in such a context of a constantly changing complex clinical environment are needed [29-31]. The potential clinical benefits amplified by the COVID-19 pandemic have forced us to reconsider how clinical care is delivered but ongoing research is needed. The heterogeneity of the populations included in this study indicates that the results may be generalizable [5,32].

The study design was explanatory and documented a series of public holidays impacting the preplanned carrying out of the study across sequential weeks which is a potential limitation in this work. However, a sensitivity analysis did not identify this to affect the results. It cannot be assumed that the presence of symptoms across the phases of this study was equal. Future research could include a reference assessment by a researcher blinded to the intervention. The use of PROMs was not clearly consistently associated with symptom detection across clinician groups. Understanding this variation across clinician groups warrants further research. In addition, patients with head and neck cancer were a larger proportion of one of the randomized clinics. At the facility, these patients are engaged in a multidisciplinary clinic prior to starting treatment and additional funded access to supportive care during treatment. It is unclear whether this contextual factor influenced outcomes. With the large number of patients seen at the facility, and the small number of patients reported in this study, it is possible that the health service outcomes described are random in nature.

Acknowledgments

The authors thank Gary Power and Anita McGrath, consumer representatives, who contributed to the study design and methods.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Summaries and logistic regression models of symptom identification or referral where week is defined as the calendar week: sensitivity analysis Summaries and logistic regression models of symptom identification or referral using study week.
[DOCX File , 26 KB - jmir_v26i1e55841_app1.docx]

References

1. Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol 2016;34(6):557-565 [FREE Full text] [doi: 10.1200/JCO.2015.63.0830] [Medline: 26644527]
2. Denis F, Lethrosne C, Pourel N, Molinier O, Pointreau Y, Domont J, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients. J Natl Cancer Inst 2017;109(9):1-8. [doi: 10.1093/jnci/djx029] [Medline: 28423407]
3. Warrington L, Absolom K, Conner M, Kellar I, Clayton B, Ayres M, et al. Electronic systems for patients to report and manage side effects of cancer treatment: systematic review. J Med Internet Res 2019;21(1):e10875 [FREE Full text] [doi: 10.2196/10875] [Medline: 30679145]
4. Di Maio M, Basch E, Denis F, Fallowfield L, Ganz P, Howell D, ESMO Guidelines Committee. The role of patient-reported outcome measures in the continuum of cancer clinical care: ESMO clinical practice guideline. Ann Oncol 2022;33(9):878-892. [doi: 10.1016/j.annonc.2022.04.007] [Medline: 35462007]
5. Absolom K, Warrington L, Hudson E, Hewison J, Morris C, Holch P, et al. Phase III randomized controlled trial of eRAPID: eHealth intervention during chemotherapy. J Clin Oncol 2021;39(7):734-747 [FREE Full text] [doi: 10.1200/JCO.20.02015] [Medline: 33417506]
6. Fromme EK, Eilers KM, Mori M, Hsieh Y, Beer TM. How accurate is clinician reporting of chemotherapy adverse effects? A comparison with patient-reported symptoms from the quality-of-life questionnaire C30. J Clin Oncol 2004;22(17):3485-3490. [doi: 10.1200/JCO.2004.03.025] [Medline: 15337796]
7. Billingy N, Tromp V, Aaronson N, Hoek RJA, Bogaard HJ, Onwuteaka-Philipsen BD, SYMPRO-Lung Consortium. Quality of life after patient-initiated vs physician-initiated response to symptom monitoring: the SYMPRO-Lung trial. J Natl Cancer Inst 2023;115(12):1515-1525 [FREE Full text] [doi: 10.1093/jnci/djad159] [Medline: 37603720]
8. Di Maio M, Gallo C, Leighl NB, Piccirillo MC, Daniele G, Nuzzo F, et al. Symptomatic toxicities experienced during anticancer treatment: agreement between patient and physician reporting in three randomized trials. J Clin Oncol 2015;33(8):910-915 [FREE Full text] [doi: 10.1200/JCO.2014.57.9334] [Medline: 25624439]
9. Laugsand EA, Sprangers MA, Bjordal K, Skorpen F, Kaasa S, Klepstad P. Health care providers underestimate symptom intensities of cancer patients: a multicenter European study. Health Qual Life Outcomes 2010;8:104 [FREE Full text] [doi: 10.1186/1477-7525-8-104] [Medline: 20858248]
10. Atkinson TM, Li Y, Coffey CW, Sit L, Shaw M, Lavene D, et al. Reliability of adverse symptom event reporting by clinicians. Qual Life Res 2012;21(7):1159-1164 [FREE Full text] [doi: 10.1007/s11136-011-0031-4] [Medline: 21984468]
11. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. J Eval Clin Pract 1999;5(4):401-416. [doi: 10.1046/j.1365-2753.1999.00209.x] [Medline: 10579704]
12. Roberts NA, Mudge A, Alexander K, Wyld D, Janda M. The iPROMOS protocol: a stepped-wedge study to implement routine patient-reported outcomes in a medical oncology outpatient setting. BMJ Open 2019;9(2):e027046 [FREE Full text] [doi: 10.1136/bmjopen-2018-027046] [Medline: 30804038]
13. Roberts NA, Janda M, Stover AM, Alexander KE, Wyld D, Mudge A, ISOQOL PROMs/PREMs in Clinical Practice Implementation Science Work Group. The utility of the implementation science framework "Integrated Promoting Action on Research Implementation in Health Services" (i-PARIHS) and the facilitator role for introducing patient-reported outcome measures (PROMs) in a medical oncology outpatient department. Qual Life Res 2021;30(11):3063-3071 [FREE Full text] [doi: 10.1007/s11136-020-02669-1] [Medline: 33089474]
14. Roberts NA, Alexander K, Wyld D, Janda M. Statistical process control assessed implementation fidelity of patient-reported outcome measures (PROMs) in routine care. J Clin Epidemiol 2020;127:76-86. [doi: 10.1016/j.jclinepi.2020.06.022] [Medline: 32569618]
15. Pinnock H, Barwick M, Carpenter CR, Eldridge S, Grandes G, Griffiths CJ, StaRI Group. Standards for Reporting Implementation Studies (StaRI) Statement. BMJ 2017;356:i6795 [FREE Full text] [doi: 10.1136/bmj.i6795] [Medline: 28264797]
16. Basch E. Patient-reported outcomes—harnessing patients' voices to improve clinical care. N Engl J Med 2017;376(2):105-108. [doi: 10.1056/NEJMp1611252] [Medline: 28076708]
17. Harvey G, Kitson A. PARIHS revisited: from heuristic to integrated framework for the successful implementation of knowledge into practice. Implement Sci 2016;11:33 [FREE Full text] [doi: 10.1186/s13012-016-0398-2] [Medline: 27013464]
18. Reeve BB, Mitchell SA, Dueck AC, Basch E, Cella D, Reilly CM, et al. Recommended patient-reported core set of symptoms to measure in adult cancer treatment trials. J Natl Cancer Inst 2014;106(7):1-8 [FREE Full text] [doi: 10.1093/jnci/dju129] [Medline: 25006191]

XSL•FO
RenderX

19. Basch E, Reeve BB, Mitchell SA, Clauser SB, Minasian LM, Dueck AC, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). J Natl Cancer Inst 2014;106(9):1-11 [FREE Full text] [doi: 10.1093/jnci/dju244] [Medline: 25265940]
20. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
21. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
22. Trotti A, Colevas AD, Setser A, Basch E. Patient-reported outcomes and the evolution of adverse event reporting in oncology. J Clin Oncol 2007;25(32):5121-5127. [doi: 10.1200/JCO.2007.12.4784] [Medline: 17991931]
23. Berry DL, Hong F, Halpenny B, Partridge AH, Fann JR, Wolpin S, et al. Electronic self-report assessment for cancer and self-care support: results of a multicenter randomized trial. J Clin Oncol 2014;32(3):199-205 [FREE Full text] [doi: 10.1200/JCO.2013.48.6662] [Medline: 24344222]
24. Girgis A, Durcinoska I, Arnold A, Descallar J, Kaadan N, Koh E, et al. Web-based Patient-Reported Outcome Measures for Personalized Treatment and Care (PROMPT-Care): multicenter pragmatic nonrandomized trial. J Med Internet Res 2020;22(10):e19685 [FREE Full text] [doi: 10.2196/19685] [Medline: 33118954]
25. Flores LT, Bennett AV, Law EB, Hajj C, Griffith MP, Goodman KA. Patient-reported outcomes vs. clinician symptom reporting during chemoradiation for rectal cancer. Gastrointest Cancer Res 2012;5(4):119-124 [FREE Full text] [Medline: 23077685]
26. Greenhalgh J, Gooding K, Gibbons E, Dalkin S, Wright J, Valderas J, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. J Patient Rep Outcomes 2018;2:42 [FREE Full text] [doi: 10.1186/s41687-018-0061-6] [Medline: 30294712]
27. Button E, Northfield S, Smith M, Wyld D, Nasato G, Yates P. How do cancer care services in Australia take care of their patients when they require emergency care? We need more evidence. Aust Health Rev 2020;44(3):497-498. [doi: 10.1071/AH19118] [Medline: 31619319]
28. Northfield S, Button E, Wyld D, Gavin NC, Nasato G, Yates P. Taking care of our own: a narrative review of cancer care services-led models of care providing emergent care to patients with cancer. Eur J Oncol Nurs 2019;40:85-97. [doi: 10.1016/j.ejon.2019.02.005] [Medline: 31229211]
29. Basch E, Rocque G, Mody G, Mullangi S, Patt D. Tenets for implementing electronic patient-reported outcomes for remote symptom monitoring during cancer treatment. JCO Clin Cancer Inform 2023;7:e2200187 [FREE Full text] [doi: 10.1200/CCI.22.00187] [Medline: 36857630]
30. Boulanger M, Petrillo L, Temel J. Listen to the patient: integrating patient-reported outcomes into clinical care. J Natl Cancer Inst 2023;115(12):1451-1453. [doi: 10.1093/jnci/djad174] [Medline: 37740689]
31. Roberts NA, Young AM, Duff J. Using implementation science in nursing research. Semin Oncol Nurs 2023;39(2):151399 [FREE Full text] [doi: 10.1016/j.soncn.2023.151399] [Medline: 36894448]
32. Schliep ME, Alonzo CN, Morris MA. Beyond RCTs: innovations in research design and methods to advance implementation science. EBP Adv Corner 2018;11(3-4):82-98. [doi: 10.1080/17489539.2017.1394807]

Abbreviations

iPARIHS: integrated-Promoting Action on Research Implementation in Health Services
PRO-CTCAE: Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events
PROM: patient-reported outcome measure
REDCap: Research Electronic Data Capture

Edited by Y Zhuang; submitted 27.12.23; peer-reviewed by A von Huben, G Carot-Sans; comments to author 05.04.24; revised version received 25.04.24; accepted 13.05.24; published 27.08.24.

Please cite as:
Roberts NA, Pelecanos A, Alexander K, Wyld D, Janda M
Implementation of Patient-Reported Outcomes in a Medical Oncology Setting (the iPROMOS Study): Type II Hybrid Implementation Study
J Med Internet Res 2024;26:e55841
URL: https://www.jmir.org/2024/1/e55841
doi:10.2196/55841
PMID:39190468

XSL•FO
RenderX

©Natasha Anne Roberts, Anita Pelecanos, Kimberly Alexander, David Wyld, Monika Janda. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Collection of Components to Design Clinical Dashboards Incorporating Patient-Reported Outcome Measures: Qualitative Study

Anja Yvonne Bischof[1], PhD; David Kuklinski[1], PhD; Irene Salvi[1], MSc; Carla Walker[1], MA; Justus Vogel[1], PhD; Alexander Geissler[1], Prof Dr

Chair of Health Economics, Policy and Management, School of Medicine, University of St. Gallen, St. Gallen, Switzerland

Corresponding Author:
Anja Yvonne Bischof, PhD
Chair of Health Economics, Policy and Management
School of Medicine
University of St. Gallen
St. Jakob-Strasse 21
St. Gallen, 9000
Switzerland
Phone: 41 712243220
Email: anja.bischof@unisg.ch

Abstract

Background: A clinical dashboard is a data-driven clinical decision support tool visualizing multiple key performance indicators in a single report while minimizing time and effort for data gathering. Studies have shown that including patient-reported outcome measures (PROMs) in clinical dashboards supports the clinician's understanding of how treatments impact patients' health status, helps identify changes in health-related quality of life at an early stage, and strengthens patient-physician communication.

Objective: This study aims to determine design components for clinical dashboards incorporating PROMs to inform software producers and users (ie, physicians).

Methods: We conducted interviews with software producers and users to test preselected design components. Furthermore, the interviews allowed us to derive additional components that are not outlined in existing literature. Finally, we used inductive and deductive coding to derive a guide on which design components need to be considered when building a clinical dashboard incorporating PROMs.

Results: A total of 25 design components were identified, of which 16 were already surfaced during the literature search. Furthermore, 9 additional components were derived inductively during our interviews. The design components are clustered in a generic dashboard, PROM-related, adjacent information, and requirements for adoption components. Both software producers and users agreed on the primary purpose of a clinical dashboard incorporating PROMs to enhance patient communication in outpatient settings. Dashboard benefits include enhanced data visualization and improved workflow efficiency, while interoperability and data collection were named as adoption challenges. Consistency in dashboard design components is preferred across different episodes of care, with adaptations only for disease-specific PROMs.

Conclusions: Clinical dashboards have the potential to facilitate informed treatment decisions if certain design components are followed. This study establishes a comprehensive framework of design components to guide the development of effective clinical dashboards incorporating PROMs in health care practice.

(J Med Internet Res 2024;26:e55267) doi:10.2196/55267

KEYWORDS

clinical dashboards; design components; patient-reported outcome measures (PROMs); patient-reported outcomes (PROs)

XSL•FO
RenderX

Introduction

Technological advances have created a multitude of innovative prevention and treatment options in medicine. Despite providing benefits to patients, this evolution created settings of increasingly complex medical decision-making [1] and an increased distancing of physicians from their patients [2]. The use of clinical dashboards can support physicians in dealing with various relevant sources of information, such as clinical parameters or patient-reported outcomes (PROs). A clinical dashboard can serve as a data-driven clinical decision support tool if capable of querying multiple data sources and visually representing key performance indicators in a single frame. The added value of a dashboard comes from its ability to provide a concise overview of key information along the treatment process [3]. Theoretically, visualizing data on clinical dashboards should decrease time spent on data gathering, cognitive overload, and time to task completion and should improve situation awareness and compliance with evidence-based safety guidelines [4]. However, the frustration of physicians increases as systems are not running smoothly and sometimes even hinder physicians in their daily work [5].

Incorporating the patient perspective into clinical dashboards by patient-reported outcome measures (PROMs) offers considerable potential to enhance clinical outcomes and patient's health-related quality of life (HRQoL) [6-9]. Furthermore, it can result in a more patient-centered treatment approach and improved care processes [10,11]. PROMs assess, for instance, patients' HRQoL, functional impairments, and disabilities. In addition, PROMs can report the nature and severity of symptoms in a standardized way [11,12], independent of physicians' or caregivers' interpretation [13,14]. The use of PROMs in clinical dashboards has, thus, been shown to support physicians' understanding of how treatment impacts symptom scores [11] and to help identify HRQoL deterioration at an early stage [12].

Research on the development of dashboards, including PROMs, has been conducted for outpatient treatment of prostate cancer, overactive bladder, multiple sclerosis, and various surgical procedures [12,15-18]. Furthermore, most of the existing literature has focused on disease-specific PROMs [9,11,16,19,20]. Accordingly, the potential of incorporating a generic PROM, such as the EQ-5D, into a dashboard has not been investigated yet. Incorporating a generic PROM enables comparability across diseases and an overall assessment of the patient's HRQoL. However, research shows that physicians will only use clinical dashboards if they perceive an additional value [11].

Current literature focuses considerably more often on dashboards for chronic diseases [11,12,16,17,20-22] compared with dashboards on 1-time acute care interventions [18,23]. Furthermore, the literature reveals the relevance of complementary data as important design components of clinical dashboards incorporating PROMs. Components such as patient information (such as demographic information or most recent health updates) [20,24,25] or clinical data (such as lab results and medication data) [20,24,26,27] are highlighted. Furthermore, the inclusion of additional design components such as past assessment scores (ie, the visualization of PRO scores over time) [9,11,27,28], peer-group comparisons (ie, analyzing a patient's PRO scores with a peer group) [17,27,29], PRO-related and overall health-related goals (ie, goals concerning PROs or overall health developed during patient-physician interaction) [16,27], alerts (ie, in case of critical values) [9,12], free-write in features (ie, additional space for physician's notetaking) [9,28], and dashboard customizability (ie, degree of how far individual wishes of users may be respected) [16,17,24] are promoted (for more information, please refer to Multimedia Appendix 1).

However, due to the literature's inconsistency on design components for clinical dashboards, we aim to develop a guide on which design components to consider when building clinical dashboards incorporating PROMs. Therefore, we investigate the following 2 research questions: (1) What are the design components of clinical dashboards incorporating generic and disease-specific PROMs? (2) Do the design components for clinical dashboards incorporating PROMs differ for acute conditions with a 1-time intervention versus chronic conditions?

Methods

Overview

We conducted interviews with software producers and users, that is, physicians, to identify design components for user-friendly clinical dashboards to enhance the quality of care [18]. By interviewing 2 stakeholder groups, we collected insights on 2 key perspectives (software producers vs users).

The recruitment of the interviewees followed a consecutive process: first, we recruited in our own network, and then, through a snowballing technique, we were able to contact software producers and users outside of our network. The conduction of interviews ended once thematic saturation was reached, that is, when no additional information was gained from further interviews [30]. The interviews were semistructured, and findings from a previous literature search (see Multimedia Appendix 1) guided the development of the interview questionnaire. Initially, AYB developed the interview guide, while IS, DK, and AG collaborated to refine it. To finalize the questionnaire, a pilot interview was conducted with a key opinion leader in this subject area. After the first 5 interviews, we iteratively tailored the interview structure to align with the evolving interview dynamics, facilitating a focused exploration of our research priorities. This adaptive approach involved selecting pertinent questions toward the gravitation points that seemed to emerge from our preestablished question catalog, thus augmenting the inductive explanatory insights.

All interviews were conducted by 2 authors (AYB and IS) to ensure that the same content was covered. Both interviewers were female, had received training in qualitative methods as part of their PhD studies, and had no previous relationship with any of the interviewees. All interviews took place between November 02, 2022, and May 15, 2023. Interviewees decided on the interview date and time.

Before each interview, 3 introductory slides were presented to all interviewees to ensure a common understanding of clinical

XSL•FO
RenderX

dashboards, outline the underlying research questions, and clarify content-related questions. All interviews covered the following main areas: (1) general questions on dashboards and PROM usage, (2) questions about market penetration (software producers only), (3) usage of the dashboard, (4) dashboard development and data collection, (5) feature assessment, and (6) role of the patient (Multimedia Appendix 2).

We conducted the one-to-one interviews on the web (by Zoom [Zoom Video Communications] or Microsoft Teams) in English or German. During the interviews, both interviewers took notes. All interviews were audio-recorded and transcribed verbatim to perform detailed data analysis. We sent the transcripts back to the interviewees to receive approval and then anonymized all participant details. For the analysis of the interviews, we applied thematic coding to uncover potential patterns and themes within the data [31]. To identify relevant information for the design components, first, deductive coding was used to cross-validate the findings from the literature search. Second, inductive coding allowed us to identify and classify information not covered by the literature search. AYB and CW coded all interviews and conducted the analysis with Atlas.ti Windows (version 22) [32]. The study is reported in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) 32-item checklist to ensure transparency and reliability (Multimedia Appendix 3) [33].

Ethical Considerations

This study was performed in line with the principles of the Declaration of Helsinki. The Ethics Committee of the University of St. Gallen was informed about this research and issued a letter of exemption.

Results

Overview

In total, we conducted 16 interviews with 6 software producers (I1 to I6) and 10 users (I7 to I16) until thematic saturation was reached (male=10, female=6). All software producer interviewees were from different companies and had different positions. A total of 8 out of 10 user interviewees were department heads with different specialties, such as pneumology, oncology, pediatrics, or orthopedics. This allowed us to analyze distinctive needs for design components in clinical dashboards. One of the remaining user interviewees was a chief hospital innovation officer, and the other interviewee was a member of hospital quality management. These specific interviewees lacked a formal medical background. However, based on their employment within a hospital setting, they routinely interacted closely with diverse medical practitioners and, thus, can be assumed to have a profound understanding of physicians' needs for clinical dashboards. The software producer interviews lasted between 29 and 45 minutes, whereas the user interviews had a duration of 20 to 36 minutes (Multimedia Appendix 4).

The interviews with software producers as well as users revealed that they perceived differentiation in design depending on disease (1-time acute vs chronic condition) as not essential. Therefore, we did not distinguish between the design requirements of different disease types but instead focused on the design principles for clinical dashboards incorporating PROMs in general.

General Information

Overview

To evaluate the importance of the key design components, we asked both interview groups to outline the type of disease, the setting, the type of PROM used, and the targeted key user for their clinical dashboard. Software producers mostly indicated they prefer dashboards that are applicable to chronic and acute conditions similarly (4/6) (Multimedia Appendix 5 contains the full set of codes). For users, we observed that they preferred a dashboard for either a chronic or acute disease, which often correlated with their professional focus. Both interview groups rated the outpatient setting as the preferable area of application (8/16). Both interview groups mentioned using a disease-specific PROM only (6/16) or combining generic and disease-specific PROMs (9/16). However, none of the interviewees preferred to use a generic PROM only, whereas I11 acknowledged that his team only uses generic PROMs for research purposes. Furthermore, both interview groups favored physicians (13/16), followed by other health care professionals (such as nurses or physiotherapists) (7/16) as key users. The patient was also recognized as a potential user of clinical dashboards (8/16); however, there were some discrepancies on how much and which content should be displayed to patients. I13 also highlighted that clinical dashboards may be valuable tools for relatives supporting patients in their daily activities to grasp the patient's health status better.

Data Collection

The most frequently mentioned use case for data collection was for micro perspectives (ie, patient-physician communication and intrapatient comparison; 12/16), especially by users (8/10), followed by meso perspectives (ie, comparison of patient groups within departments or institutions) for both interview groups (5/16; Table 1). Furthermore, the most often mentioned purposes of reporting were "communication" (14/16) and "better basis for decision-making (for physician)" (9/16; Table 1).

For the point-in-time of data collection, there was a clear tendency (11/16) of interviewees for patients to answer the questionnaires independently at home. However, I2 mentioned that it is perceivable that the first-time data collection is conducted in the waiting room because "[patients] get an explanation by a physician or a nurse, but after that, they are on their own." Also, interviewee I14 had a similar perception of dealing with the data collection process. For the type of data collection, a clear tendency toward web-based tools such as platforms or phone apps was indicated (10/16).

The table shows the ranking of design components' features according to the number of mentions by software producers and users. Behind the features, the share of interviewees mentioning this design component's feature is indicated. As the interviewees could provide more than 1 answer or no answer for each design component, shares can add up to more or less than 100%. The quantification of the responses can only be interpreted in the context of this study. However, it does not allow for generalizable statements.

XSL•FO
RenderX

Table 1. Overview of software producer and user responses on how data collection should be implemented in the clinical dashboard.

Software producer responses (n/N)	User responses (n/N)
Level of reporting	
Micro (4/6)	Micro (8/10)
Macro (2/6)	Meso (4/10)
Meso (1/6)	Macro (1/10)
Purpose of reporting	
Improved communication (6/6)	Improved communication (8/10)
Better basis for physician's decision (4/6)	Better basis for physician's decision (5/10)
Real-time tracking (2/6) and shared decision-making (2/6)	Shared decision-making (2/10)
Data collection	
Digital (5/6)	Digital (5/10)
__[a]	Paper-based (4/10)
Time of data collection	
Independent at home (4/6)	Independent at home (7/10)
Before the appointment (in the waiting room) (2/6)	Before the appointment (in the waiting room) (6/10)

[a]Not applicable.

Dashboard Components

Concerning the dashboard components, representatives of both interview groups mentioned that some patient information (such as patient photographs, demographic information, or contact details of other care team members) needs to be presented on the dashboard (9/16). An interviewee additionally emphasized that she expects the patient information to include,

> *[…] key events. So, surgeries need to be shown. For example, in cancer, the start of chemotherapy, completion of chemotherapy, started radiation, that you can understand what is going on in the background of those patients.* [I16]

Except for I1 and I5, all other software producers rated clinical data (eg, lab results or medication data) as meaningful information that must be included in the clinical dashboard. In contrast, users tended rather not to overfill the clinical dashboard with clinical data as this information might be available from another source, as a participant mentioned:

> *[…] and then if they [the physicians] want to have the clinical information on the patient, they just open up the EMR [electronic medical record] and check it, which is another tab in the Chrome app.* [I8]

The free write-in design component did not resonate well in either of the 2 groups. A total of 6 interviewees replied that a free write-in space is not useful, whereas only 2 interviewees perceived this design component as very beneficial. Nevertheless, when having such a free write-in box included, users (3/10) wanted it as an additional source of information for some specific variables.

Software producers and users rated the past assessment PRO score as one of the most crucial design components in a clinical dashboard incorporating PROMs. From the software producer perspective, one mentioned:

> *This [past assessment PRO score] is very well received, simply the score progression up and down visually, so to speak.* [I6]

Another participant had a similar perception:

> *This [past assessment PRO score] is absolutely relevant because it is above all changes in these questionnaires that are significant. In any case, it is very important to look at the progression, not just the individual value.* [I9]

Another well-perceived design component was the peer-group comparison. All software producer interviewees (6/6) agreed on the inclusion of this feature. The users (7/10) appreciated the inclusion, too. However, 3 users did not perceive an added value in the peer-group comparison due to uncertainties in interpreting the applied PRO (I9), the missing guidance on the relevant factors to compare different patient groups (I12), or a clear reasoning on why peer group comparison is useful (I15).

Although only 1 user emphasized applying PRO-related goals, overall health-related goals were discussed more controversially within both groups. Software producer I4 reported that their dashboard includes a component where the physician can develop overall health goals together with the patient and check goal achievement, whereas I6 stated that this is not part of their dashboard. From the user perspective, this design component could add value to the patient-physician communication, as the goal statement makes the aim or expectations of the patients explicit (I9 and I16).

Concerning alerts, software producers and users have different perceptions of whether alerts should be included in the dashboard. Software producers (5/6) mentioned that an alert

XSL•FO
RenderX

function needs to be included. Generally, producers were indifferent regarding whether the alert is in real time (4/6) or only appears during the appointment (4/6). A participant mentioned, however, that liability for medical risks and errors is unclear when real-time alerts are used:

> *[…] if [a score is] three days red or is really critical and no push notification is sent, the error is on our side. However, if a push notification is sent and the doctor does not react, the error lies with the doctor. And that is a bit of a grey area, where we still have to figure out how it's actually done.* [I4]

In contrast, 4 users preferred not to include alerts. I8 and I11 mentioned that this feature previously existed in their dashboard, but the acceptance was not high enough in their teams, which made them stop using it. Furthermore, I12 raised the issue that additional interpretations for physicians are required to ensure that they completely understand what the deterioration or improvement in a score means. I10 favored the alerts during the appointment to highlight critical factors and to facilitate comparison over time. I13 and I16 preferred real-time alerts allowing the treating physician to react directly to the patient's issues.

Concerning the dashboard's customizability, all interviewees agreed that some degree of customizability is required to meet different user needs. Although most users preferred customizability from a standard set (7/10), software producers still seemed undecided whether a clinical dashboard should be adapted to individual needs (2/6) or "off the shelf" (3/6). Nevertheless, software producers (5/6) agreed that scalability is only achievable in case clinical dashboards equipped with features defined in a standard set are provided.

Findings From Inductive Coding

Whereas the above-presented results were derived from the deductive procedure, the subsequent inductive procedure allowed us to identify additional aspects that were not explicitly discussed in the current literature. These aspects included, on the one hand, the added value of clinical dashboards incorporating PROMs and visualization possibilities of PROMs and, on the other hand, various barriers, reducing the uptake of the clinical dashboard. Furthermore, the interviewees also discussed the patient's access to their data controversially.

Although the visualization of the PRO scores was not extensively discussed in the literature, we asked our interviewees about their preferences. Only 2 out of 6 software producers indicated that the index and dimensional scores are always provided in their dashboards. The remaining software producers elaborated on the visualization of their PRO scores. Contrarily, almost all users (9/10) mentioned their preferences for the visualization: 4/10 users preferred to have index and dimensional scores available, 3/10 users prioritized dimensional scores, whereas 2/10 users voted for the visualization of index scores. For instance, I7 voted for the visualization of dimensional scores by stating, "*we must of course know the dimensionality and different aspects.*" In contrast, an interviewee argued:

> *I like [index] scores better, as I said, but because we have these individual questions like there are ten questions, and you can have a summary score of it and the system plots every question on a trend. I think that is rather messy because then you have like ten different color graphs just projected over each other, and you can click them on, or off. So it's easy, but for me it's less informative.* [I11]

Concerning added value provided by the clinical dashboard, software producers emphasized the visualization of results (5/6), enhanced workflow efficiency (4/6), improved data comparability (3/6), and higher patient satisfaction (3/6). Users perceived the biggest advantages of a clinical dashboard in the visualization of results (8/10), enhanced workflow efficiency (6/10), and improved data comparability (5/10). Higher patient satisfaction (2/10) was not among the most often named advantages.

Potential barriers to implementing the full potential of clinical dashboards are interoperability between various systems for both interview groups (12/16) and the burdensome collection of data for users (6/10). The reasons for the burdensome collection were various. For instance, a participant highlighted organizational issues by stating:

> *There is one burden that falls more on the admin staff and the secretaries, which is to make sure that the right patient is put on the right pathway with the right PROMs. So, you need to pre-program all this. Then you must either have an automatic integration saying this patient is going to see this because it needs to be personalized to the patient pathway. So how do you make sure that the right patients get the right thing without having an army of extra people?* [19]

Furthermore, a participant raised technical challenges in collecting PROs by expressing:

> *We had a problem for a year now that we had these computers in the waiting room and for some reason they were broken or disconnected. It took us a very long time to get them because there were all these people taking ownership of these silly computers.* [I11]

Furthermore, software producers mentioned legal consequences in case of displaying inadequate information (4/6) and the licensing of the PROM questionnaires (3/6) as potential barriers. Users rather perceived the lack of good data (4/10), legal consequences (3/10), and nonintuitive use (3/10) as additional barriers.

Except for 2 users (I7 and I14), all interviewees agreed that patients benefit from access to the information stored on the clinical dashboard. A total of 9 of the interviewees expect the patients to develop a better understanding and interpretation of their own health status. One software producer summarized their efforts as

> *[…] what we are doing now is that we have patient apps, and the patients can also see their own results in their own app. They can also say to the caregiver: 'hey, I see this or that', and so we involve the patients into it. I think it is important to make it patient-friendly also.* [I2]

XSL•FO
RenderX

Users highlighted similar scenarios, for instance:

> *When reading this data with PROMs, it's very clear. It's data that they have entered. If they see that they're functioning or the ability to walk a block or go on a bus is going down, they can relate to it and understand 'Oh, actually I thought I was doing better than that' or 'I'm happy, actually I forgot that I couldn't and now I can.* [I8]

This quote also represents the statement of I14, whereas I15 additionally emphasizes the positive correlation between an enhanced understanding of one's own health status and increased treatment adherence. Especially the score visualization supports the patient's understanding of their own health status (6/16). Still, users (7/10) acknowledged that more explanation of the meaning of the data is required to avoid patient's misinterpretation of the data. Furthermore, 2 users claimed that physicians and patients need different levels of aggregation of available data concerning depth of analysis and interpretation complexity to avoid overwhelming patients. Finally, I6 and I8 stated they would refuse to share clinical data with patients as clinical outcomes might frighten patients or lead to misinterpretations, such as outlined by a user:

> *If you look at in oncology or radio oncology to PSA levels, which are an indicator but not an actual sign of cancer, you wouldn't want the patients to learn by themselves. 'Okay, my PSA level are up again, and I don't know what it means and what is going on.' With this type of outcomes, you need to be very careful on how you get them across […] and this needs professional support.* [I18]

Findings From Deductive and Inductive Coding

The derivation of deductive and inductive codes led to insights into how different design components are related to each other (Figure 1). To build clinical dashboards incorporating PROMs, not only the selected PROMs on the dashboard are crucial, but rather the complete construct of the clinical dashboard. In addition, generic dashboard components need to be respected as well. Furthermore, aspects such as awareness of barriers, patients' perspectives, and the potential of added value need to be considered when reflecting on the adoption of clinical dashboards (Multimedia Appendix 6 contains a full description of all topics).

Figure 1. Overview on design components for developing clinical dashboards incorporating patient-reported outcome measures. PRO: patient-reported outcome; PROM: patient-reported outcome measure.

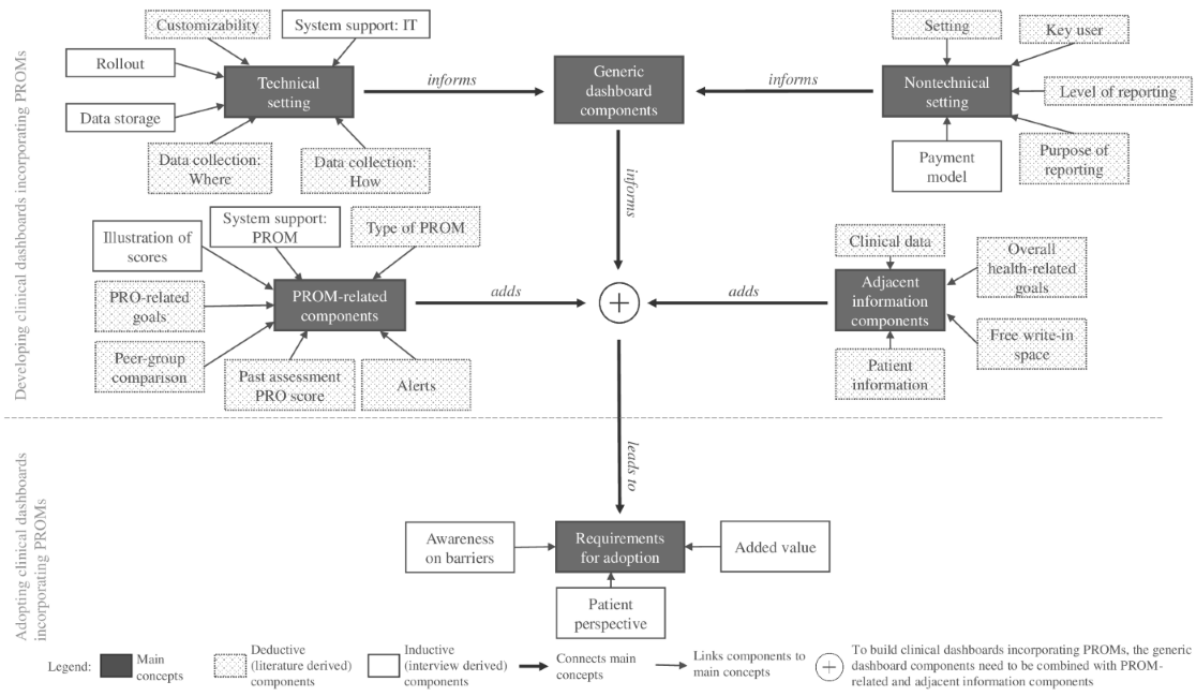

Discussion

Principal Findings

The qualitative research approach allowed for a more profound understanding of the design components for clinical dashboards incorporating generic and disease-specific PROMs. Our developed framework (Figure 1) displays that we found 16 design components based on a literature search and identified 9 additional design components through interviews with software producers and users. The interviews revealed that both interview groups had a similar perception of the use case of a clinical dashboard incorporating PROMs.

The use of a dashboard primarily strengthens patient-physician communication in outpatient settings. It should include a combination of disease-specific and generic PROMs. Physicians and other health care professionals were the most frequently mentioned key users. In addition, preferences for visualizing PROs, whether as index or dimensional scores, are subjective and dependent on individual inclinations. According to software producers and users, dashboards provide benefits, especially

XSL•FO
RenderX

through the visualization of data and the enhancement of workflow efficiency. The main barriers to dashboard implementation were seen in the interoperability between various systems and the collection of data.

In addition, the interviews revealed that dashboard content should not vary across diseases, except by an adaptation of the disease-specific PROM. This project showed that not only design components directly displayed by the clinical dashboard incorporating PROMs but also other features, such as providing system support or the definition of key users, are relevant topics for a successful implementation.

Comparison With Previous Work

Concerning the use of PROMs, studies show that it is common to use generic PROMs, particularly the EQ-5D, to track the health status of patients. Notably, researchers see the standardized format and content as a major benefit that facilitates its usability across different diseases and patient groups [34]. Furthermore, the multilingual questionnaire allows large-scale analyses [35,36]. Not only were disease-specific PROMs complementing generic PROMs, but their combined value was seen as more valuable than their sum, as in the case of the EQ-5D and the COPD (chronic obstructive pulmonary disease) Assessment Test (CAT) combination [37]. Interestingly, similar findings emerged from studies developing PROM dashboards for various disease areas. Baeksted et al [28] and Hassett et al [9] found it relevant to include the cancer care–specific PRO of Common Terminology Criteria for Adverse Events (CTCAE) in their dashboards, and Nicolas-Boluda et al [22] incorporated endometriosis-specific indicators into their dashboard. We found that disease-specific PROMs are preferred over generic PROMs by the interviewees. Nevertheless, a combined use of both PROMs was seen as beneficial, too.

Enhanced visualization of the patient's PROs over time can facilitate the moderation role of the physician [12,19,38] in line with the growing demand for shared decision-making. We found a similar outcome as the interviewees perceived the purpose of a clinical dashboard incorporating PROMs in the improved communication between patient and physician. Furthermore, Desantis et al [12] found that using a clinical dashboard improves the workflow and communication of changes in the HRQoL between the patient and physician. Following this argument, interviewees emphasized that a clinical dashboard incorporating PROMs may also serve as a basis for better-informed decisions by physicians as various key information points are displayed in 1 report. Therefore, additional information besides the PRO data must be displayed to display a holistic picture of the patient's general health conditions. Patient information (such as sociodemographic information or most recent health updates) and clinical data (such as lab results and medication lists) are required, which was previously highlighted in the literature [20,24,25] and also rated as essential by the interviewees in our study.

Furthermore, the interviewees promoted the design component "past assessment PRO score" and ranked it as one of the most important ones. Changes in the patient's health status become apparent and traceable, and thus, enable the initiation of countermeasures immediately. Similar reasons are also provided in the literature [9,28]. Another well-perceived design component for analyzing PROs was the peer-group comparison. Interviewees claimed that this design component facilitates understanding whether a patient's health status is above or below the norm. Hartzler et al [17] highlighted that peer-group comparison should match patients in age and treatment. In addition, Ragouzeos et al [27] advise indicating a "normal" range.

Design components where interviewees did not indicate a clear tendency were PRO-related and overall health goals and the function of alerts. Cronin et al [16] found that patients want to set and evaluate goals over time, which could be supported by a clinical dashboard. However, Liu et al [26] warned that setting goals might further pressure the patient. Furthermore, the challenge of including alerts is to decide on the appropriate alert level [39] and avoid "alert fatigue" of physicians [20].

Limitations

Our study has 3 limitations. First, the interviewees were mainly from Western countries, especially Switzerland and Germany. A generalization of our results to other countries should, therefore, be carefully assessed. Second, the user interviewees were most often specialists, and we did not control for their years of experience. This research can be seen as a starting point in building guidance on what needs to be considered when developing clinical dashboards incorporating PROMs. Therefore, we wanted to include a broad range of opinions and needs on critical design components. Future research should consider various dashboard stakeholders from different countries with varying years of professional experience to further evaluate our proposed design components. Third, the design components were not built and evaluated in a real-world setting but only discussed during the interviews. We aimed to develop a theoretical framework of design components based on software producers' capabilities and users' needs. Therefore, future research approaches can build up on our findings to refine clinical dashboards and test them in clinical practice.

Conclusions

In our study, we identified 25 design components and consolidated them into a framework of 4 main concepts. This framework strives to work as a guiding approach where researchers and practitioners can orient themselves on components to be included when building clinical dashboards incorporating PROMs. Design components such as peer-group comparison or alert thresholds require further investigation in future research, as knowledge on determining appropriate peer groups or setting alerts based on PRO thresholds needs to be improved. In addition, future research efforts may test the real-world applicability of the proposed design components by either discussing them in focus groups or implementing them in a dashboard in clinical practice.

XSL•FO
RenderX

Acknowledgments

We would like to thank Dr Gouke Bonsel and Dr Bas Janssen for their support during the research process and for providing valuable insights. Finally, we thank all software producers and users who took the time for an interview and transparently talked to us about their wishes and requirements for the clinical dashboards incorporating patient-reported outcome measures. This study is financially supported by the EuroQol Research Foundation. However, the EuroQol Research Foundation played no role in the study design, data collection, analysis and interpretation of data, or writing of the manuscript.

Authors' Contributions

All authors contributed to the conception and design of the study. AYB and IS acquired the data. AYB and IS prepared the data for analyses. AYB, DK, and CW analyzed and interpreted the data. AYB drafted the manuscript. DK and AG supervised the study. All authors substantively revised the work. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of design components and categories of components which laid the foundation for user and software producer interviews.
[PDF File (Adobe PDF File), 201 KB - jmir_v26i1e55267_app1.pdf]

Multimedia Appendix 2
Semi-structured interview guide for software producer and user interviews.
[PDF File (Adobe PDF File), 184 KB - jmir_v26i1e55267_app2.pdf]

Multimedia Appendix 3
Consolidated Criteria for Reporting Qualitative Research (COREQ).
[PDF File (Adobe PDF File), 192 KB - jmir_v26i1e55267_app3.pdf]

Multimedia Appendix 4
Overview of software producer and user interviewee's characteristics.
[PDF File (Adobe PDF File), 105 KB - jmir_v26i1e55267_app4.pdf]

Multimedia Appendix 5
Output of thematic coding.
[PDF File (Adobe PDF File), 275 KB - jmir_v26i1e55267_app5.pdf]

Multimedia Appendix 6
Explanation of topics from deductive and inductive coding.
[PDF File (Adobe PDF File), 285 KB - jmir_v26i1e55267_app6.pdf]

References

1. Mulley AG, Trimble C, Elwyn G. Stop the silent misdiagnosis: patients' preferences matter. BMJ 2012;345:e6572. [doi: 10.1136/bmj.e6572] [Medline: 23137819]
2. Barry MJ, Edgman-Levitan S. Shared decision making--pinnacle of patient-centered care. N Engl J Med 2012;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
3. Wilbanks BA, Langford PA. A review of dashboards for data analytics in nursing. Comput Inform Nurs 2014;32(11):545-549. [doi: 10.1097/CIN.0000000000000106] [Medline: 25310224]
4. Khairat SS, Dukkipati A, Lauria HA, Bice T, Travers D, Carson SS. The impact of visualization dashboards on quality of care and clinician satisfaction: integrative literature review. JMIR Hum Factors 2018;5(2):e22 [FREE Full text] [doi: 10.2196/humanfactors.9328] [Medline: 29853440]
5. Tawfik DS, Sinha A, Bayati M, Adair KC, Shanafelt TD, Sexton JB, et al. Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study. J Med Internet Res 2021;23(7):e26817 [FREE Full text] [doi: 10.2196/26817] [Medline: 34255674]
6. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. BMC Health Serv Res 2013;13:211 [FREE Full text] [doi: 10.1186/1472-6963-13-211] [Medline: 23758898]

XSL•FO
RenderX

7. Devlin NJ, Appleby J. Getting the Most Out of PROMs. In: Putting Health Outcomes at the Heart of NHS Decision-Making. London, England: King's Fund; 2010:83.
8. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. J Eval Clin Pract 1999;5(4):401-416. [doi: 10.1046/j.1365-2753.1999.00209.x] [Medline: 10579704]
9. Hassett MJ, Cronin C, Tsou TC, Wedge J, Bian J, Dizon DS, et al. eSyM: an electronic health record-integrated patient-reported outcomes-based cancer symptom management program used by six diverse health systems. JCO Clin Cancer Inform 2022;6:e2100137 [FREE Full text] [doi: 10.1200/CCI.21.00137] [Medline: 34985914]
10. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform 2015;84(2):87-100 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.10.001] [Medline: 25453274]
11. Fautrel B, Alten R, Kirkham B, de la Torre I, Durand F, Barry J, et al. Call for action: how to improve use of patient-reported outcomes to guide clinical decision making in rheumatoid arthritis. Rheumatol Int 2018;38(6):935-947 [FREE Full text] [doi: 10.1007/s00296-018-4005-5] [Medline: 29564549]
12. Desantis D, Baverstock RJ, Civitarese A, Crump RT, Carlson KV. A clinical perspective on electronically collecting patient-reported outcomes at the point-of-care for overactive bladder. Can Urol Assoc J 2016;10(11-12):E359-E366 [FREE Full text] [doi: 10.5489/cuaj.3757] [Medline: 27942273]
13. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, PROMIS Cooperative Group. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):S3-S11 [FREE Full text] [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
14. Lipscomb J, Reeve BB, Clauser SB, Abrams JS, Bruner DW, Burke LB, et al. Patient-reported outcomes assessment in cancer trials: taking stock, moving forward. J Clin Oncol 2007;25(32):5133-5140. [doi: 10.1200/JCO.2007.12.4644] [Medline: 17991933]
15. Jongen P, Sinnige L, van Geel B, Verheul F, Verhagen W, van der Kruijk R, et al. The interactive web-based program MSmonitor for self-management and multidisciplinary care in multiple sclerosis: concept, content, and pilot results. Patient Prefer Adherence 2015;9:1741-1750 [FREE Full text] [doi: 10.2147/PPA.S93783] [Medline: 26715841]
16. Cronin RM, Conway D, Condon D, Jerome RN, Byrne DW, Harris PA. Patient and healthcare provider views on a patient-reported outcomes portal. J Am Med Inform Assoc 2018;25(11):1470-1480 [FREE Full text] [doi: 10.1093/jamia/ocy111] [Medline: 30239733]
17. Hartzler AL, Izard JP, Dalkin BL, Mikles SP, Gore JL. Design and feasibility of integrating personalized PRO dashboards into prostate cancer care. J Am Med Inform Assoc 2016;23(1):38-47 [FREE Full text] [doi: 10.1093/jamia/ocv101] [Medline: 26260247]
18. Hartzler AL, Chaudhuri S, Fey BC, Flum DR, Lavallee D. Integrating patient-reported outcomes into spine surgical care through visual dashboards: lessons learned from human-centered design. EGEMS (Wash DC) 2015;3(2):1133 [FREE Full text] [doi: 10.13063/2327-9214.1133] [Medline: 25988187]
19. LeRouge C, Hasselquist MB, Kellogg L, Austin E, Fey BC, Hartzler AL, et al. Using heuristic evaluation to enhance the visual display of a provider dashboard for patient-reported outcomes. EGEMS (Wash DC) 2017;5(2):6 [FREE Full text] [doi: 10.13063/2327-9214.1283] [Medline: 29930968]
20. Rudin RS, Perez S, Rodriguez JA, Sousa J, Plombon S, Arcia A, et al. User-centered design of a scalable, electronic health record-integrated remote symptom monitoring intervention for patients with asthma and providers in primary care. J Am Med Inform Assoc 2021;28(11):2433-2444 [FREE Full text] [doi: 10.1093/jamia/ocab157] [Medline: 34406413]
21. Elm JJ, Daeschler M, Bataille L, Schneider R, Amara A, Espay AJ, et al. Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson's disease data. NPJ Digit Med 2019;2:95 [FREE Full text] [doi: 10.1038/s41746-019-0169-y] [Medline: 31583283]
22. Nicolas-Boluda A, Oppenheimer A, Bouaziz J, Fauconnier A. Patient-reported outcome measures in endometriosis. J Clin Med 2021;10(21):5106 [FREE Full text] [doi: 10.3390/jcm10215106] [Medline: 34768627]
23. Heath EL, Ackerman I, Lorimer M, Rainbird S, O'Donohue G, Brock A, et al. National implementation of an electronic patient-reported outcome measures program for joint replacement surgery: pilot study. JMIR Form Res 2022;6(4):e30245 [FREE Full text] [doi: 10.2196/30245] [Medline: 35394439]
24. Tsangaris E, Edelen M, Means J, Gregorowitsch M, O'Gorman J, Pattanaik R, et al. User-centered design and agile development of a novel mobile health application and clinician dashboard to support the collection and reporting of patient-reported outcomes for breast cancer care. BMJ Surg Interv Health Technol 2022;4(1):e000119. [doi: 10.1136/bmjsit-2021-000119] [Medline: 35464815]
25. Taxter A, Johnson L, Tabussi D, Kimura Y, Donaldson B, Lawson E, et al. Co-design of an electronic dashboard to support the coproduction of care in pediatric rheumatic disease: human-centered design and usability testing. J Particip Med 2022;14(1):e34735 [FREE Full text] [doi: 10.2196/34735] [Medline: 35133283]
26. Liu LH, Garrett SB, Li J, Ragouzeos D, Berrean B, Dohan D, et al. Patient and clinician perspectives on a patient-facing dashboard that visualizes patient reported outcomes in rheumatoid arthritis. Health Expect 2020;23(4):846-859 [FREE Full text] [doi: 10.1111/hex.13057] [Medline: 32270591]

XSL•FO
RenderX

27. Ragouzeos D, Gandrup J, Berrean B, Li J, Murphy M, Trupin L, et al. "Am I OK?" using human centered design to empower rheumatoid arthritis patients through patient reported outcomes. Patient Educ Couns 2019;102(3):503-510 [FREE Full text] [doi: 10.1016/j.pec.2018.10.016] [Medline: 30446358]
28. Baeksted C, Pappot H, Nissen A, Hjollund NH, Mitchell SA, Basch E, et al. Feasibility and acceptability of electronic symptom surveillance with clinician feedback using the patient-reported outcomes version of common terminology criteria for adverse events (PRO-CTCAE) in danish prostate cancer patients. J Patient Rep Outcomes 2017;1(1):1 [FREE Full text] [doi: 10.1186/s41687-017-0005-6] [Medline: 29757324]
29. Strachna O, Cohen MA, Allison MM, Pfister DG, Lee NY, Wong RJ, et al. Case study of the integration of electronic patient-reported outcomes as standard of care in a head and neck oncology practice: obstacles and opportunities. Cancer 2021;127(3):359-371 [FREE Full text] [doi: 10.1002/cncr.33272] [Medline: 33107986]
30. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3:77-101 [FREE Full text]
32. Atlas.ti 22 Windows. ATLAS.ti scientific software development GmbH. URL: https://atlasti.com/ [accessed 2024-09-06]
33. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. Zhou T, Guan H, Wang L, Zhang Y, Rui M, Ma A. Health-related quality of life in patients with different diseases measured with the EQ-5D-5L: a systematic review. Front Public Health 2021;9:675523 [FREE Full text] [doi: 10.3389/fpubh.2021.675523] [Medline: 34268287]
35. Devlin NJ, Brooks R. EQ-5D and the EuroQol group: past, present and future. Appl Health Econ Health Policy 2017;15(2):127-137 [FREE Full text] [doi: 10.1007/s40258-017-0310-5] [Medline: 28194657]
36. Sen RK, Shetti V, Mukhopadhyay R, Tripathy SK, Saini G, Dip SK, et al. Satisfaction and health-related quality of life following hip and knee arthroplasty surgeries in Indian patients: a cross-sectional study. Indian J Orthop 2022;56(5):918-926 [FREE Full text] [doi: 10.1007/s43465-021-00589-x] [Medline: 35547347]
37. Szentes BL, Schwarzkopf L, Kirsch F, Schramm A, Leidl R. Measuring quality of life in COPD patients: comparing disease-specific supplements to the EQ-5D-5L. Expert Rev Pharmacoecon Outcomes Res 2020;20(5):523-529. [doi: 10.1080/14737167.2019.1662302] [Medline: 31512524]
38. Watson L, Delure A, Qi S, Link C, Chmielewski L, Photitai ?, et al. Utilizing patient reported outcome measures (PROMs) in ambulatory oncology in alberta: digital reporting at the micro, meso and macro level. J Patient Rep Outcomes 2021;5(Suppl 2):97 [FREE Full text] [doi: 10.1186/s41687-021-00373-3] [Medline: 34637003]
39. Mooney K, Whisenant M, Beck S. Symptom care at home: a comprehensive and pragmatic PRO system approach to improve cancer symptom care. Med Care 2019;57(Suppl 5 Suppl 1(Suppl 5 1)):S66-S72 [FREE Full text] [doi: 10.1097/MLR.0000000000001037] [Medline: 30531525]

Abbreviations

CAT: COPD Assessment Test
COPD: chronic obstructive pulmonary disease
COREQ: Consolidated Criteria for Reporting Qualitative Research
CTCAE: Common Terminology Criteria for Adverse Events
HRQoL: health-related quality of life
PRO: patient-reported outcome
PROM: patient-reported outcome measure

Edited by S He; submitted 07.12.23; peer-reviewed by G Schmajuk, M Sedlmayr; comments to author 13.02.24; revised version received 09.05.24; accepted 07.07.24; published 02.10.24.

Please cite as:
Bischof AY, Kuklinski D, Salvi I, Walker C, Vogel J, Geissler A
A Collection of Components to Design Clinical Dashboards Incorporating Patient-Reported Outcome Measures: Qualitative Study
J Med Internet Res 2024;26:e55267
URL: https://www.jmir.org/2024/1/e55267
doi:10.2196/55267
PMID:

©Anja Yvonne Bischof, David Kuklinski, Irene Salvi, Carla Walker, Justus Vogel, Alexander Geissler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.10.2024. This is an open-access article distributed under the

XSL•FO
RenderX

terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancing Clinicians' Use of Electronic Patient-Reported Outcome Measures in Outpatient Care: Mixed Methods Study

Veerle van Engen[1], MSc; Igna Bonfrer[1], PhD; Kees Ahaus[1], Prof Dr; Monique Den Hollander-Ardon[2], PhD; Ingrid Peters[2*], PhD; Martina Buljac-Samardzic[1*], PhD

[1]Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, Netherlands

[2]Department of Quality and Patient Care, Erasmus Medical Center, Rotterdam, Netherlands

[*]these authors contributed equally

Corresponding Author:
Veerle van Engen, MSc
Erasmus School of Health Policy & Management
Erasmus University Rotterdam
Burgemeester Oudlaan 50
Rotterdam, 3062 PA
Netherlands
Phone: 31 010 408 8555
Email: vanengen@eshpm.eur.nl

Abstract

Background: Despite the increasing use of patient-reported outcome measures (PROMs) for collecting self-reported data among hospital outpatients, clinicians' use of these data remains suboptimal. Insight into this issue and strategies to enhance the use of PROMs are critical but limited.

Objective: This study aimed to examine clinicians' use of PROM data for value-based outpatient consultations and identify efforts to enhance their use of PROMs in a Dutch university hospital. First, we aimed to investigate clinicians' use of outpatients' PROM data in 2023, focusing on adoption, implementation, and maintenance. Second, we aimed to develop insights into the organizational-level strategies implemented to enhance clinicians' use of PROM data from 2020 to 2023. This included understanding the underlying rationales for these strategies and identifying strategies that appeared to be missing to address barriers or leverage facilitators. Third, we aimed to explore the key factors driving and constraining clinicians' use of PROMs in 2023.

Methods: We integrated data from 4 sources: 1-year performance data on clinicians' use of PROMs (n=70 subdepartments), internal hospital documents from a central support team (n=56), a survey among clinicians (n=47), and interviews with individuals contributing to the organizational-level implementation of PROMs (n=20). The Reach, Effectiveness, Adoption, Implementation, and Maintenance framework was used to analyze clinicians' adoption, implementation, and maintenance of PROMs. Strategies were analyzed using the Expert Recommendations for Implementing Change taxonomy, and results were structured around the constructs of capability, opportunity, and motivation.

Results: On average, around 2023, clinicians accessed PROM data for approximately 3 of 20 (14%) patients during their outpatient consultation, despite numerous strategies to improve this practice. We identified issues in adoption, implementation, and maintenance. The hospital's strategies, shaped organically and pragmatically, were related to 27 (37%) out of 73 Expert Recommendations for Implementing Change strategies. These strategies focused on enhancing clinicians' capability, opportunity, and motivation. We found shortcomings in the quality of execution and completeness of strategies in relation to addressing all barriers and leveraging facilitators. We identified variations in the factors influencing the use of PROMs among frequent PROM users, occasional users, and nonusers. Challenges to effective facilitation were apparent, with certain desired strategies being unfeasible or impeded.

Conclusions: Enhancing clinicians' use of PROMs has remained challenging despite various strategies aimed at improving their capability, opportunity, and motivation. The use of PROMs may require more substantial changes than initially expected, necessitating a shift in clinicians' professional attitudes and practices. Hospitals can facilitate rather than manage clinicians' genuine use of PROMs. They must prioritize efforts to engage clinicians with PROMs for value-based outpatient care. Specific

XSL•FO
RenderX

attention to their professionalization may be warranted. Tailored strategies, designed to address within-group differences in clinicians' needs and motivation, hold promise for future efforts.

(J Med Internet Res 2024;26:e60306) doi:10.2196/60306

KEYWORDS

patient-reported outcome measure; value-based health care; implementation; clinician; behavior; barrier; facilitator; strategies; professionalization; mixed methods

Introduction

Background

Despite the increasing volume of patient-reported outcome measures (PROMs) data collected in hospitals [1], clinicians' use of these data remains suboptimal [2,3]. PROMs are tools designed to capture patients' own perceptions of their health, functioning, and quality of life [4,5], often through electronic surveys [1,3]. In outpatient care, clinicians can use a patient's PROM data to screen and monitor important outcomes and involve patients in medical decisions [6]. In this way, PROMs facilitate value-based health care (VBHC) [7,8], aiming to optimally allocate resources to achieve outcomes that matter to patients.

Achieving and maintaining consistent use of PROM data by clinicians presents difficulties, requiring them to adapt their data use behavior and acquire new skills [3,9,10]. Therefore, to realize the full benefits of PROMs, it is imperative to understand the factors influencing clinicians' use of PROMs and develop strategies that facilitate and motivate their use in outpatient care [8,10,11].

Previous Work

Clinicians' use of PROM data has been proposed as an important metric to evaluate implementation success of PROMs [12]; however, it has received limited attention thus far. The underuse of PROM data is concerning because the effectiveness of PROMs relies on clinicians acting upon the received data [13-16]. Moreover, clinicians' explicit use is essential for encouraging continued patient participation in future PROMs [17,18].

While comprehensive overviews of barriers to and facilitators of PROMs implementation exist [12,19-21], these often lack a specific focus on clinicians' perspectives and behaviors [9] and leave gaps in understanding how to respond to the identified factors. While some studies propose hypothetical strategies on the basis of identified barriers [22,23], insights from real-world experiences are limited. A few exceptions provide insights into clinicians' experiences with PROMs [10,24-27], their self-reported use [24], and implemented strategies [12,20,28]. Thus, there remains a critical need to learn from both successful and less successful implementation experiences aimed at enhancing clinicians' use of PROMs [10,25,29,30].

Case

This study draws attention to clinicians' use of PROMs in a Dutch university hospital, Erasmus Medical Center (Erasmus MC), and the organizational-level strategies initiated to enhance this practice. Adopting a stepwise approach, the first specialty departments began collecting electronic PROMs from their outpatients around 2017, aimed at fostering VBHC [31-33]. New subdepartments continued to join this initiative, continuing into 2024. A central support team (CST) facilitated and coordinated this change, comprising around 6 full-time equivalent and an integrated IT team.

At Erasmus MC, patients complete PROMs before some of their outpatient consultations. To prevent survey overload, a maximum frequency for PROMs completion is set in agreement with subdepartments. This means that during outpatient consultations, clinicians will have a subset of patients for whom new PROM data are available. Clinicians are expected to review these data and discuss them with the outpatient to inform care decisions and actions. This process is facilitated by a consultation room dashboard that visualizes outcomes over time and compares them to reference values. This dashboard shows summary scores as trend lines and allows clinicians to access patient responses to specific items. Clinicians can access this dashboard, which is fully integrated into the patient's electronic health record (EHR) since 2020, both during preparation for consultations and during the consultations themselves. For the first time, patients will receive feedback on their completed PROMs during their consultation. Individually, they can only access item-level responses without visualizations that aid in interpretation. In future, a separate patient dashboard is planned to offer these additional features.

Figure 1 [34-36] outlines the process of using PROMs, the specific tools used, and it provides a visual representation of the dashboard. In addition to the PROMs items, the patient survey includes an open-ended question designed to capture topics that the patients wish to discuss during their consultation. Multimedia Appendix 1 provides further information on dashboarding.

XSL•FO
RenderX

Figure 1. The process of patient-reported outcome measures (PROMs) completion and discussion (top-left), the 3 tiers of PROMs tools used (top-right), and an impression of the dashboard designs (bottom). For the generic tier, the hospital used the Patient-Reported Outcome Measure Information System (PROMIS; version 1.2; Global Health). Within the domain-specific tier, PROMIS short forms (SFs) were used among the nononcological population, including version 2.0, Physical function 4a; version 1.0, Anxiety 4a; version 1.0, Depression 4a; version 1.0, Fatigue 4a; version 1.1, Pain interference 4a; version 1.0, Sleep disturbance 4a; and version 1.0, Satisfaction with participation in social roles 4a. The European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30 CAT) was used for the oncological population. Subdepartments could select additional disease-specific PROMs.

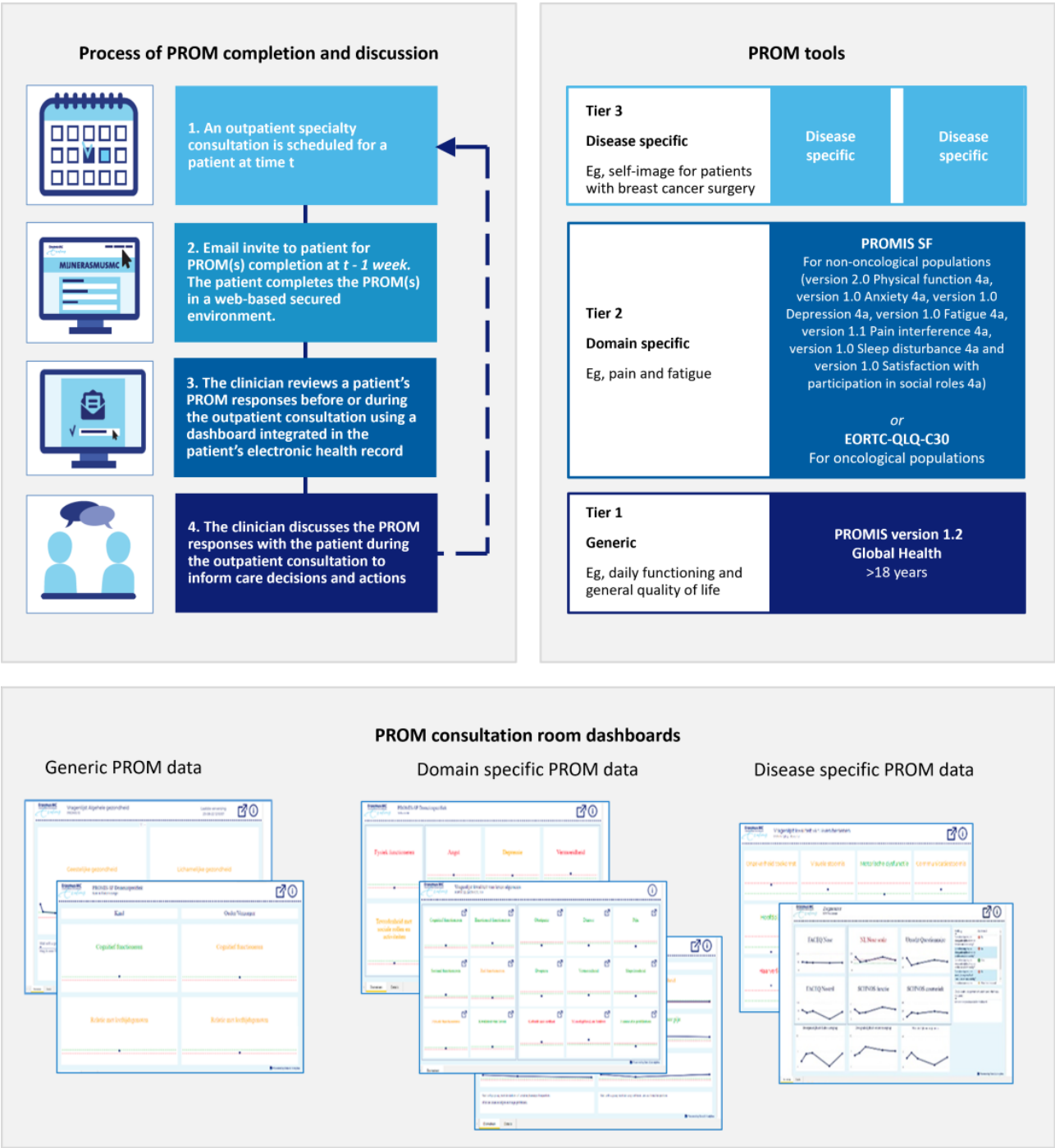

By the end of 2023, PROMs were delivered to nearly one-quarter of the outpatient population from 70 subdepartments. However, clinicians' use of PROM data remained limited. This is despite numerous strategies organically and pragmatically applied by the CST to facilitate and encourage use, including those previously reported as supportive [12,20,28]. This discrepancy highlights the need for further investigation and makes Erasmus MC an intriguing case for investigation.

Objectives

The objectives of this study are 3-fold. First, we aimed to outline the issues with clinicians' underuse of outpatients' PROM data in Erasmus MC around 2023, focusing on adoption, implementation, and maintenance. Second, we aimed to develop knowledge on the organizational-level strategies implemented to enhance clinicians' use of PROM data from 2020 to 2023. This includes understanding the underlying rationales for these

XSL•FO
RenderX

strategies and identifying any strategies that appear missing to address barriers or leverage facilitators. Third, we aimed to investigate the key factors driving and constraining clinicians' use of PROMs in 2023.

This approach helps understand how implementation unfolds in a real-life context using standardized frameworks to ensure that findings can be compared with other cases and replicated in different settings [37]. In addition, the findings may inform future strategies aimed at facilitating and promoting clinicians' use of PROMs.

Methods

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and exempted from formal approval under the Dutch Medical Research Involving Human Subjects Act by the Medical Research Ethics Committee of the Erasmus School of Health Policy & Management (ETH2223-0225). All participants consented to participation and the publication of quotes. Survey data were collected in a way that ensured participants remained anonymous, and interview data were pseudonymized. All data were stored in a secured space. Participants did not receive compensation.

Data Sources

Overview

This study used 4 sources of data: implementation outcome indicators, a survey, hospital documents, and interviews. Data collection was part of a broader study on the implementation of VBHC [33]. Implementation outcome indicators were used specifically for objective 1, whereas the other data sources were used to address objectives 2 and 3.

Implementation Outcome Indicators

The hospital monitored the degree to which clinicians accessed the PROM consultation room dashboard. We extracted the percentage of completed PROMs accessed by clinicians from the hospital's monitoring system, covering the 12-month period from February 2023 to January 2024. We used the aggregated data that eventually included 70 subdepartments. Investigating subdepartmental differences in the use of PROMs was outside the scope of this study. We analyzed both the average percentage of completed PROMs accessed by clinicians during the week of the patient's consultation and on the day of the consultation. While the former indicates the use of PROMs to prepare for consultations, the latter serves as a proxy for discussion with the patient.

Hospital Documents

Internal hospital documents were accessed through the digital workspace of the CST, to which author VvE was granted access. This workspace was established around 2020. A systematic search was conducted using the following keywords: compliance, evaluation, barrier, facilitator, challenge, plan, intervention, and strategy. This yielded 56 files. These were then screened for data pertaining to factors facilitating or impeding clinicians' use of PROMs or strategies aimed at enhancing their use, followed by data extraction. If files referred to other files, these were also considered for inclusion. Subsequently, the list of extracted strategies was cross-validated with author MDH-A, a member of the CST, to distinguish between planned and executed strategies.

Survey

In January 2023, a collective survey, named EMC23, was digitally distributed to all 194 clinicians across the 35 subdepartments that collected PROM data from outpatients in 2022. One clinician involved in the survey design was excluded from participation. The participants remained anonymous and untraceable. The survey comprised 26 items (Multimedia Appendix 2), developed based on preliminary insight from document analysis. The items delved into clinicians' self-reported use of PROMs, prominent factors influencing this, and their overall satisfaction with the implementation process and outcomes achieved. In addition, the respondents were asked to evaluate a subset of implemented strategies and had the opportunity to provide comments. Two reminders were sent. As shown in Table 1, 47 responses were included in the analysis.

XSL•FO
RenderX

Table 1. Data sources, sample sizes, and participant characteristics.

Data source, theme, and description	Values
Survey in January 2023 (N=194)	
Respondents, n (%)	
Total	57 (29.4)
Excluded	10 (5.2)
• Demographic questions answered (n=5) • Not providing patient care (n=2) • PROMs[a] not yet available (n=2) • No familiarity with PROMs (n=1)	
Complete	42 (21.6)
Sex of included respondents (n=47), n (%)	
Female	35 (74.5)
Male	12 (25.5)
Age (y), mean (SD; range)	46 (8.2; 31-64)
Function of included respondents (n=47), n (%)	
Medical specialist	30 (63.8)
Nurse	12 (25.5)
Other (eg, psychologist or resident in training)	5 (10.6)
Interviews conducted between February 2023 and April 2023 (n=20), n (%)	
Participants	
Member of executive board	1 (5)
Director of quality and patient safety	1 (5)
Head of VBHC[b] implementation	2 (10)
Member of steering committee	2 (10)
Lead of VBHC program team	3 (15)
Member of VBHC program team	9 (45)
External consultant	1 (5)
Clinician in VBHC	1 (5)
Sex	
Female	14 (70)
Male	6 (30)

[a]PROM: patient-reported outcome measure.

[b]VBHC: value-based health care.

Interviews

For a process evaluation on the implementation of VBHC [33], author VvE conducted 20 semistructured interviews with individuals involved in the central implementation effort during the past decade (Table 1). Two individuals refused participation for personal reasons. Participants were identified through documents and snowball sampling. Interviews were recorded and transcribed verbatim. Given the hospital's focus on implementing PROMs in their move to VBHC, interviewees provided valuable insights into the hospital's efforts and experiences in enhancing clinicians' use of PROMs. The portion of the interview guide that elicited relevant comments for this study is presented in Multimedia Appendix 3. We reanalyzed these transcripts, extracting information on barriers and facilitators for clinicians to use PROMs, strategies and their rationales, and areas for improvement. These data complemented the other sources mentioned in the Data Sources section, enriching the findings of this study.

Theoretical Lenses and Data Analysis

Overview

Given the intricate nature of the implementation of PROMs, adopting a multifaceted research approach is imperative [12], typically including contextual factors, the innovation (PROMs), strategies to support effective implementation of the innovation, and implementation outcomes interact, as illustrated in Figure

XSL•FO
RenderX

2. These elements align with the objectives of this study. Therefore, our methods were plural, incorporating the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework [38]; the Expert Recommendations for Implementing Change (ERIC) taxonomy [39] and associated 9 clusters [40]; and the COM-B constructs [41]. We reviewed multiple strategy taxonomies and evaluation frameworks as summarized by Nilsen [42], selecting the aforementioned frameworks due to their alignment with this study's objectives and data, widespread use, compatibility with the health care setting, and clarity in presenting results. These frameworks were used for post hoc, deductive data analysis; these frameworks are explained in subsequent sections.

Figure 2. Study objectives. PROM: patient-reported outcome measure.

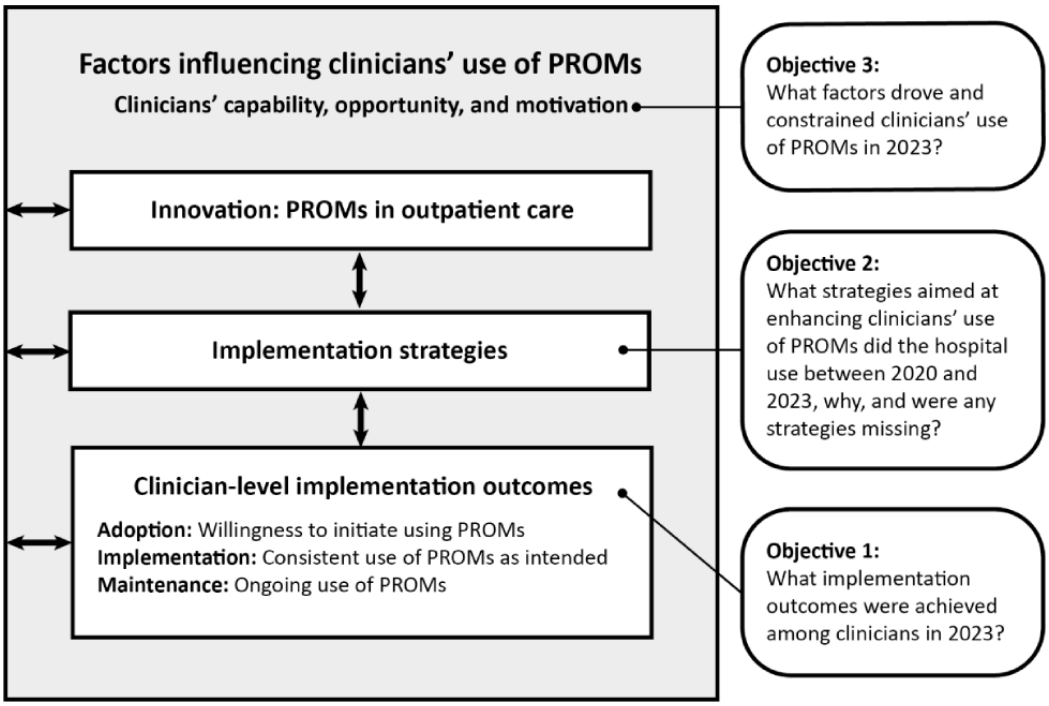

Objective 1: Implementation Outcomes Among Clinicians in 2023

The RE-AIM framework [38] guides the planning and evaluation of programs by examining 5 outcomes: *reach*, *effectiveness*, *adoption*, *implementation*, and *maintenance*. This study focused on the latter 3 outcomes at the individual level rather than at the department or hospital level. While the collection of PROMs among outpatients was a departmental choice, individual clinicians subsequently needed to adopt, implement, and maintain using PROMs. In this study, *adoption* denoted individual clinicians' willingness to initiate the use of PROM data from their patients. *Implementation* denoted the consistency of clinicians' use of PROMs as intended, that is, acknowledging outcomes and discussing results with the patient during their consultation to inform decisions. *Maintenance* referred to the clinicians' ongoing use of PROMs. *Reach* and *effectiveness* have not been directly studied in this research but were indirectly addressed, as they are contingent upon the extent to which clinicians adopt, implement, and maintain PROM use [8,11].

We examined the outcomes *adoption*, *implementation*, and *maintenance* through *survey* data, reporting item-level frequencies for two key questions: (1) How frequently do you examine patients' responses to PROMs? with the following response options: *I never used PROMs*, *I stopped looking at PROMs*, *occasionally*, *as often as possible*, and *always*; and (2) To what extent do you discuss these outcomes with the patient during the consultation? (for those who examine PROMs) with the following response options: *never*, *occasionally*, *as often as possible*, and *always*. We supplemented these data with insights from the hospital's implementation outcomes indicators, for which no further analyses were required.

Objective 2: Implemented and Missing Strategies From 2020 to 2023

We retrospectively analyzed Erasmus MC's implementation efforts. First, we documented all strategies applied by the hospital from 2020 to 2023 based on *document* and *interview* data, focusing on understanding why these strategies were chosen, particularly in response to barriers or facilitators. For each strategy, we aimed to deductively specify the actor, action, target of the action, temporality, dose, and implementation outcome affected and provide a justification (Multimedia Appendix 4) [43].

Besides describing the practical application of strategies at Erasmus MC, this study mapped discrete ERIC strategies [39] to the identified strategies using a coding manual developed by Fridberg et al [44]. ERIC provides a compilation of 73 discrete strategies that one can use to effectively implement an innovation in a health care setting [39], which have been mapped into 9 strategy clusters [40]. These clusters are mentioned in the Results section. The manual excluded the discrete strategy of facilitation due to overlap with other strategies. This study also included 10 strategies proposed as additions to the ERIC taxonomy: provide stakeholders with the possibility to attend educational meetings [44], recruit clinicians with competence

XSL•FO
RenderX

in the innovation [44], provide stakeholders with resources [44], act as a role model [44], assess and redesign workflow [45], create web-based learning communities [45], engage community resources [45], plan for outcome evaluation [46], obtain worker feedback about the implementation plan [46], and communication as a strategy category [47].

To deepen our understanding, we included stakeholder reflections on strategies from *documents*, *interviews*, and *survey* data, including comments on their feasibility and perceived utility. We coded these data by mapping them to the respective strategy they pertained to, allowing us to cluster relevant information accordingly. Subsequently, we integrated these qualitative data into the results per strategy. The dataset was too limited to conduct thematic analyses on a per-strategy basis. Furthermore, we assessed whether identified barriers or facilitators in our data were addressed through these strategies. If not, these cases were labeled as *missing strategies*.

We observed that our data aligned effectively with the constructs of capability, opportunity, and motivation, which are components of the COM-B model [41]. Therefore, we structured our results around these constructs, finding them intuitive and straightforward for conveying our findings.

Objective 3: Key Factors Influencing Clinicians' Use of PROMs in 2023

On the basis of document analysis, we identified key barriers and facilitators to clinicians' use of PROMs. In the survey, clinicians rated the perceived influence of these factors on their PROM use (Multimedia Appendix 2). They also had the opportunity to identify additional factors.

We analyzed item-level scores for the entire survey population by grouping *agree* and *partly agree* into a single category. Likewise, *disagree* and *partly disagree* were grouped into a single category. The categories *neutral* and *no opinion* remained unchanged. Moreover, we conducted subsample analyses to explore differences among different user profiles. Clinicians were allocated to 3 groups depending on their self-reported use of PROMs: *nonusers*, *frequent users*, and *occasional users*. Nonusers included clinicians who reported never attempting to use PROMs or ceasing to use PROMs. Frequent users comprised those who reported using PROMs as often as possible or always. Clinicians who reported occasional use remained unchanged.

Results

Objective 1: Implementation Outcomes Among Clinicians in 2023

Implementation Outcome Indicators

The hospital monitored that, on average, from February 2023 to January 2024, clinicians accessed PROM data for nearly 3 out of 20 patients (14%; SD 1.6%) on the day of each patient's consultation. This data served as a proxy for discussing PROMs with the patient. During the week of the patient's consultation, which included the use of PROM data for triage and consult preparations, this percentage was slightly higher, averaging at 19.5% (SD 2.4%). While variations in the use of PROM data were apparent across subdepartments and individual clinicians, these specifics were outside the scope of this study.

Yearly data suggested a slight, albeit minimal, increase in the percentage of completed PROMs of unique patients accessed by clinicians using the dashboard (Figure 3). It is important to note that during 2024, the number of participating subdepartments grew from 38 to 70, implying that around half of the clinicians were in the adoption and implementation stages, while the other half were potentially moving toward the maintenance stage.

Figure 3. Patient-reported outcome measure (PROM) dashboard use rates among clinicians from February 2023 to January 2024.

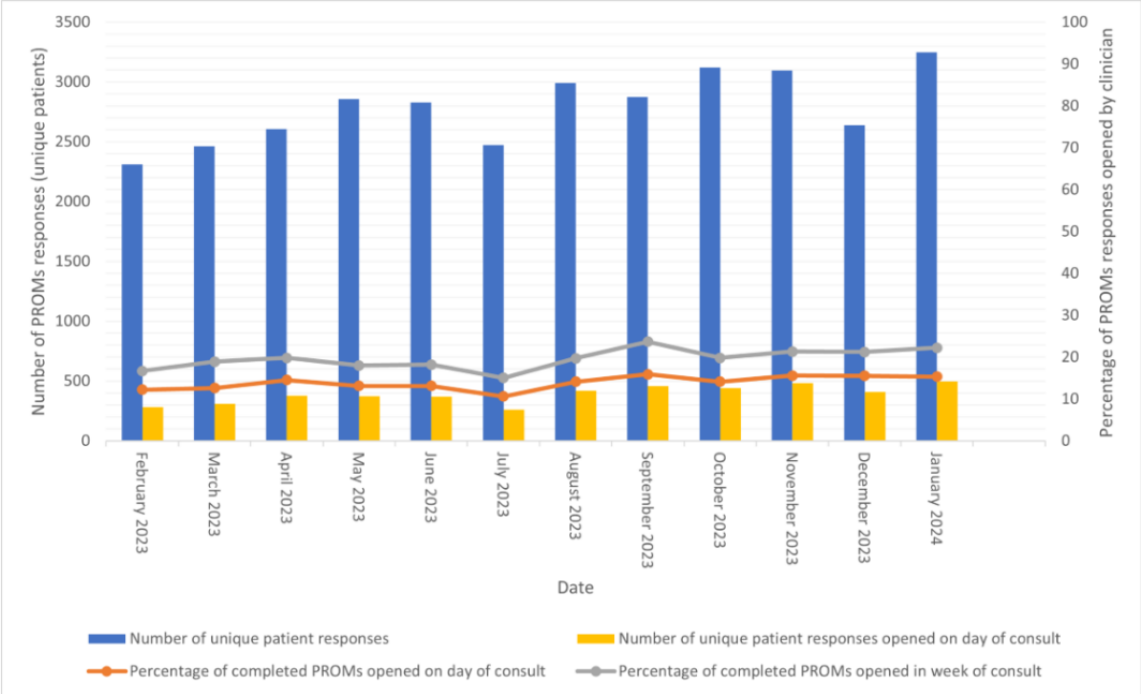

XSL•FO
RenderX

Survey Data

Clinicians' self-reported use of PROMs, as derived from our survey in early 2023, revealed issues across all 3 RE-AIM outcomes. Adoption issues were apparent, with around one-tenth (5/47, 11%) of the respondents never attempting to review their patients' PROMs responses (Table 2). Implementation issues, characterized by inconsistent use of PROMs as intended, became evident, with more than half (27/47, 58%) of the clinicians examining completed PROMs infrequently. Moreover, among those who examined PROMs, less than one-third (10/35, 29%) reported always discussing PROM data with patients. Less than one-fifth of the respondents (6/35, 17%) did not engage in this activity at all. Furthermore, 15% (7/47) ceased examining PROMs, indicating a maintenance issue.

Table 2. Clinicians' self-reported use of patient-reported outcome measures (PROMs).

Items and answers	Respondents, n (%)
Frequency of examining patients' responses to PROMs (n=47)	
Never attempted	5 (10.6)
Ceased	7 (14.9)
Occasionally	17 (36.2)
As often as possible	10 (21.3)
Always	8 (17)
Frequency of discussing PROMs with patients (among those examining PROMs at least occasionally; n=35)	
Never	6 (17.1)
Occasionally	10 (28.6)
As often as possible	9 (25.7)
Always	10 (28.6)

Objective 2: Implemented and Missing Strategies From 2020 to 2023

Overview

Table 3 summarizes the strategies that the hospital implemented alongside their corresponding ERIC strategies, structured according to the COM-B constructs of capability, opportunity, and motivation. The hospital used 27 (37%) out of 73 discrete ERIC strategies to enhance clinicians' use of PROMs alongside 5 proposed additions to the ERIC taxonomy. These are elaborated upon in the subsequent section and specified in Multimedia Appendix 4.

Strategies were drawn from 8 (89%) out of 9 strategy clusters, including using evaluative and iterative strategies, providing interactive assistance, adapting and tailoring to the context, developing stakeholder interrelationships, training and educating stakeholders, supporting clinicians, engaging consumers, and changing infrastructure. No strategies from the cluster financial strategies were used. Qualitatively, it appears that clinicians who used PROMs more extensively rated the supportiveness of the initiated strategies higher than those who used PROMs less. A few strategies were found to be missing, leaving barriers unaddressed or facilitators not leveraged.

Tracking and generalizing strategy specifications, such as temporality and dose, for all strategies was challenging due to the organic and pragmatic application of strategies tailored to specific subdepartment needs and the lack of a comprehensive documented program theory. Having been developed during preceding pilot phases and, most of the strategies were available from the onset of the hospital-wide implementation program in 2020 and refined in the course of time.

XSL•FO
RenderX

Table 3. Strategies applied by the hospital, structured around the COM-B constructs of capability, opportunity, and motivation and corresponding strategies from the Expert Recommendations for Implementing Change (ERIC) taxonomy.

COM-B construct targeted by the strategy and strategies by the CST[a], except otherwise noted	ERIC strategies (except when referenced and noted with N[b])
Capability	
Conducted a kick-off session and shared a self-developed written and video manual	• 15[b]. Conduct educational meetings • 29. Develop educational materials • 31. Distribute educational materials
Developed and distributed a pocket guide on discussing PROMs[c] with patients	• 29. Develop educational materials • 31. Distribute educational materials
Developed and conducted training and highlighted external training opportunities	• 29. Develop educational materials • 31. Distribute educational materials • 52. Promote network weaving
Made training flexible (e-trainings) and sometimes accredited	• N1. Provide stakeholders with the possibility to attend educational meetings [44] • 29. Develop educational materials • 69. Create or change credentialing and/or licensure standards
Was accessible for questions and tailored support	• 43. Make training dynamic • 63. Tailor strategies • 8. Centralize technical assistance
Provided coaching on the job	• 16. Conduct educational outreach visits • 71. Use train-the-trainer strategies
Organized peer-to-peer discussion events	• 20. Create a learning collaborative
Opportunity	
Created hospital-wide awareness about PROMs and VBHC[d], aiming to extend reach to extramural parties	• N2. Communication [47]
Visualized PROMs in a dashboard and integrated this dashboard into the EHR[e]	• 12. Change record systems
Facilitated patients to take the initiative to discuss PROMs	• 50. Prepare patients to be active participants
Facilitated quick actions on PROMs outcomes and streamlined other care processes	• 12. Change record systems • N3. Assess and redesign workflow [45]
Implemented a reminder for PROMs in the EHR	• 12. Change record systems • 58. Remind clinicians
Motivation	
The executive board verbally expressed commitment to achieving VBHC and included it as a hospital aim.	• 40. Involve executive boards • 44. Mandate change • N2. Communication [47]
The executive board monitored departments' implementation status and complimented the departments excelling in VBHC.	• 56. Purposely re-examine the implementation • N2. Communication [47]
Clinicians participated in the steering committee.	• 64. Use advisory boards and workgroups
The CST communicated about VBHC to achieve and sustain clinicians' interest.	• N2. Communication [47] • 41. Involve patients and consumers and family members • 7. Capture and share local knowledge • 35. Identify and prepare champions • 36 Identify early adopters
Provided tailored education and held discussions with each subdepartment on PROMs and VBHC	• 15. Conduct educational meetings

XSL•FO
RenderX

COM-B construct targeted by the strategy and strategies by the CST[a], except otherwise noted	ERIC strategies (except when referenced and noted with N[b])
Enabled subdepartments to try PROMs with a subset of patients	• 14. Conduct cyclical small tests of change
Aimed to demonstrate the impact of PROMs	• N4. Plan for outcome evaluation [46] • 24. Develop academic partnerships
Gave tailored feedback on subdepartments and clinicians' use of PROMs	• 56. Purposely reexamine the implementation • 27. Develop and organize quality monitoring systems • 48. Organize clinician implementation team meetings
Enabled clinicians to adapt PROMs and their uses	• 51. Promote adaptability
Extended the use cases of PROMs	• 51. Promote adaptability
Adapted patient information to alleviate clinicians' concerns about accountability	• N2. Communication [47]
Generic	
Conducted formal and informal evaluations and used the obtained insights to adjust the implementation plan	• N5. Obtain worker feedback about the implementation plan [46]

[a]CST: central support team.

[b]Numbered strategies as per Waltz et al [40].

[c]PROM: patient-reported outcome measure.

[d]VBHC: value-based health care.

[e]EHR: electronic health record.

Capability-Associated Strategies

To enhance clinicians' knowledge and skills in using PROMs, the CST organized kick-off presentations at each department when PROMs were installed. Clinicians received written instructions on using the PROMs dashboard and interpreting outcomes via email, which were later supplemented with video instructions. In 2022, a pocket guide, with exemplary sentences to discuss PROMs was cocreated with clinicians (internal document dated July 14, 2022). Training possibilities, both internal and external, were highlighted. Internal training was often conducted in collaboration with the person-centered care group, focusing on using PROMs in shared decision-making and among patients with limited literacy. The latter aimed to address limitations from clinicians believing that "PROMs cannot be discussed with the patient in question" (internal document, July 14, 2022). Flexible learning resources, such as e-trainings, were developed in response to time constraints:

> *Attending a conference for half a day can already be challenging, let alone undertaking a lengthy training.* [Interviewee 19]

In addition, efforts were made to align trainings with clinicians' accreditation requirements.

However, interviewee 19 noted that training had not been an essential component of the program since 2020; interviewee 10 emphasized a tailored approach that addressed the specific needs of individual subdepartments. The CST increasingly reached out to subdepartments with low use of PROMs. They addressed barriers in perceived capability. One such barrier is clinicians feeling unable to influence certain PROMs outcomes that are affected by external factors, such as divorces impacting the quality of life. This perception led some clinicians to disregard these data (internal document, July 14, 2022). In response, the CST emphasized the importance of empathic listening and referring patients with critical PROMs outcomes to other physicians. Furthermore, since 2022, they provided on-the-job coaching, which proved valuable:

> *You must take them [a substantial portion of clinicians] by the hand for a moment. Where do you click?* [Interviewee 19]

They also organized a few peer-to-peer discussion events each year, which interviewee 20 recognized as more impactful than traditional theoretical instruction. However, attendance issues arose with clinicians not always able to participate in these sessions. While survey results indicated that around 60% (26/43) of respondents were dissatisfied with the training or coaching received, skill issues were not a prominent factor impeding their use of PROMs (Table 4).

XSL•FO
RenderX

Table 4. Prominent reasons for clinicians to use patient-reported outcome measures (PROMs), both in total and as per 3 user profiles.

Item, post hoc categories, and answer options	All respondents who used PROMs at least once (n=42)	User profiles		
		Nonusers (n=7)[a]	Occasional users (n=17)	Frequent users (n=18)
Most prominent factors or reasons for why I used PROMs, n (%)				
Motivation				
Benefits for the patient	16 (38)	0 (0)	5 (29)	11 (61)
Benefits for myself	15 (36)	0 (0)	5 (29)	10 (56)
Opportunity				
It is expected from me.	19 (45)	3 (43)	9 (53)	7 (39)
Patient requests	5 (12)	2 (29)	3 (18)	0 (0)
Calculation				
Number of factors selected, mean (SD)	1.3 (0.9)	0.7 (0.5)	1.3 (1.0)	1.6 (0.9)

[a]There are lower sample population values compared to the total population value because this question was posed to ceased users only (n=7), excluding those who never attempted PROMs (n=5).

Opportunity-Associated Strategies

Another set of strategies targeted clinicians' social and physical environments to create opportunities for using PROMs. Regarding clinicians' social context, the CST aimed to create legitimacy and enable clinicians to discuss PROM data during multidisciplinary consultations, even with colleagues who had not yet initiated PROMs. They focused on raising hospital-wide awareness about PROMs and VBHC through mass communication via email, intranet, and marketing in public spaces. One clinician described a challenge in this endeavor:

> *Emails get buried under more important ones, and newsletters go unread. However, too little communication also doesn't seem to work (you never do it right).* [Internal document, August 31, 2020; clinician]

Interviewee 14 also noted difficulties arising from the organization's size, with hard-to-reach *islands*.

Anticipated future strategies include providing information about PROMs to general practitioners and external providers involved in patient care. Clinicians requested this strategy to facilitate communication with external health care professionals, such as making them aware of possible referrals on the basis of PROMs scores. However, this strategy was deemed feasible once PROMs were used among all patients, as partial implementation would require excessive manual effort (internal document, November 23, 2020; interviewee 19). Furthermore, a dashboard enabling patients to review their own outcomes will be developed to increase patient initiative in discussing PROMs:

> *If patients can see their outcomes, they'll probably ask more questions about it. So, it will get used more by clinicals, both intentionally and unintentionally.* [Interviewee 2]

Currently, to encourage patient initiative, patients are provided with resources such as an animated video explaining how to discuss PROMs during consultations.

Regarding the physical environment, several clinicians experienced or assumed PROM use in outpatient care as incompatible with their workload:

> *It takes far too much time, and we don't get any extra.* [Survey respondent 8]

To mitigate this issue, the hospital sought to streamline the care process and simplify PROM use, as extending available consultation time was not feasible. Interviewee 10 explained as follows:

> *Small adjustments can make a real difference for a doctor who may not necessarily be receptive to changing their practice.*

As per interviewee 15, integration of the PROMs dashboard into the EHR was considered a major improvement. Two other features reduced the registration burden: quick copying of PROMs scores into patient notes and a referral template for easily referring patients needing further attention on the basis of PROM outcomes (internal document, May 11, 2023). However, 79% (33/42) of the survey respondents had no opinion on the usefulness of the referral letter, possibly indicating unawareness of this feature. In addition, the CST implemented a reminder by prominently displaying the PROMs completion status on the patients' EHR front page. This also eased the workflow by eliminating the need to open the PROMs dashboard to verify completion. More than one-third (15/42, 36%) of the survey respondents perceived this feature as helpful.

Moreover, to save time, PROM data were used in triage and the developed PROMs-IT infrastructure was used to have patients report their medication and lifestyle. This extension reduced the necessity for discussions on these topics during consultations (internal document, May 11, 2023). Despite these efforts, perceived time constraints remained the most frequently cited barrier among clinicians as of early 2023 (Table 4). Loading time delays of the PROMs dashboard exacerbated these constraints, resulting in decreased commitment and temporary cessation of PROM use. These delays were resolved, and a

XSL•FO
RenderX

dedicated campaign was organized to encourage the reuptake of PROMs.

Motivation-Associated Strategies

The use of PROMs remained nonobligatory, focusing on genuine engagement rather than mere compliance. The inclusion of clinicians in the steering committee aimed to create "visible standard-bearers from within the healthcare sector" (internal document, May 20, 2019) to enhance buy-in. The CST frequently communicated about PROMs and VBHC to achieve and sustain interest. They shared stories highlighting the benefits of PROMs from both patients and clinicians via a monthly newsletter and intranet. One of the ideas was that "if it's beneficial for the patient, it usually has added value for the clinician as well" (internal document, July 14, 2022). Stories of championing clinicians were noted as a key facilitator in documents and interviews (internal document, August 31, 2020; interviewee 3). However, approximately half (21/43, 49%) of the survey respondents did not feel that the stories of other health care professionals positively influenced their attitude toward working with PROMs, with only 16% (7/43)—primarily consistent PROM users—reporting that they did. In communication, the CST also emphasized the importance of clinicians' discussion of PROMs on patient adherence to complete subsequent PROMs, supported by local evidence (internal document, June 7, 2023). Sometimes, the CST chose not to overtly reference VBHC, focusing instead on visible outcomes for clinicians, such as PROMs enhancing the patient-clinician conversation. Interviewee 3 explained as follows:

> *Sometimes your sales pitches need to be much flatter than what's underneath it all.*

This adjusted framing aimed to address resistance among clinicians who viewed VBHC sceptically, perceiving it, for example, as a "dull, container concept" (interviewee 6) and "a hype" (survey respondent 11). As another reason, terminology shifts in the national policy created confusion and was overwhelming (interviewee 3). Nonetheless, some clinicians perceived the hospital's communication efforts negatively. Survey respondent 60 mentioned, "extremely selective advertising," and survey respondent 11 noted that "a hallelujah story is being told that overlooks the differences in patient populations."

Concurrently, formal pressures increased: the executive board verbally committed to achieving VBHC, supported by video messages and integration into hospital aims. Externally, government programs supported PROMs and VBHC [48,49], and external audits increasingly evaluated PROM use (internal document, July 14, 2022). These factors directly motivated clinicians and empowered department heads to prioritize PROM use among their employees, according to interviewee 14. Since 2022, the executive board requested subdepartments to report on their VBHC implementation status thrice yearly and complimented high-performing departments (internal document, July 14, 2022). The CST proposed that the executive board explicitly request information on clinicians' use of PROMs rather than merely their inquiries among outpatients (interviewee 19). However, this proposal has not been implemented thus far, and the reasons for this remain unknown to the CST. Survey respondent 11 expressed concerns about such an approach:

> *My biggest fear is that it [PROMs] becomes an outcome, that the departments that use it are deemed "better" than those that do not.*

This fear may have stemmed from clinicians' belief that PROMs do not suit all patient groups or consultations equally (internal document, July 14, 2022).

To facilitate adoption, subdepartments were enabled to try PROMs with a subset of patients. Interviewee 19 highlights the importance of clinicians developing direct experience with PROMs:

> *[T]he true impact of PROM data becomes evident when you experience it.*

However, implementing PROMs among a subset of patients had an unintended consequence:

> *If we decide, "We'll do it for X and Y but not for Z"... well, then you forget about it.* [Interviewee 7]

Forgetting to use PROMs was the second-most cited barrier by clinicians (Table 4). To stimulate sufficient PROM survey volume, several strategies aimed to enhance patient response rates to PROMs. Furthermore, the CST aimed to evidence the benefits of PROMs and their impact on professionals' work experience, including time investment. Interviewee 6 stated the following:

> *The key question is, do you have proof? That is the ultimate killer question.*

However, these efforts were hindered by limitations in defining measurable outcome measures, data availability, and difficulties in consolidating local evidence and making it convincing to other disciplines. Moreover, the hospital conducted both formal and informal evaluations of PROMs among clinicians to gain insights into their experiences, motivations, and needs, subsequently adapting implementation plans on the basis of these findings. This included prioritizing disease-specific PROMs, extending training opportunities, and expanding PROM data use to areas beyond outpatient care, such as triage, research, and care pathway improvements, tapping into different clinician motivations (interviewee 17). In addition, the CST alleviated clinicians' concerns about accountability by adapting patient communication to clarify patients' responsibility for timely contact in critical situations via traditional means.

To achieve and maintain effective implementation, the CST provided feedback on PROM use to subdepartments. The frequency and format of feedback was tailored to departmental preferences, with some favoring nonanonymous feedback, such as "[clinician's name] has accessed this many PROMs" (interviewee 19), for a competitive aspect, while others preferred anonymity. However, nearly half of the survey respondents (20/43, 47%) experienced infrequent feedback. Furthermore, local departments were increasingly empowered to adapt PROMs and their application to their context, addressing limitations of uniform solutions on buy-in, perceived ownership, and local fit (interviewee 18). Subdepartments could incorporate disease-specific items that were deemed critical:

XSL•FO
RenderX

> *[a generic PROMs] is excellent for (big) data collection but lacks clinical value; it's too general.* [Internal document, August 31, 2020]

Moreover, they could determine patient eligibility criteria, adjust inquiry frequency, and choose from multiple dashboard layouts (internal document, May 11, 2023). Individual clinicians were also enabled to manually exclude patients from PROMs if needed, aiming to mitigate challenges from upset patients:

> *The feeling of incompetence [to complete a PROM] among the less literate population takes a huge toll on their outpatient visits.* [Survey respondent 60]

> *Even the son of one of my patients complains about it.* [Survey respondent 7]

Moreover, it aimed to overcome a specific drawback noted by survey respondent 4:

> *[Patient] inclusion in studies is less due to PROMs, so that irritates me.*

Overall, interviewee 20 perceived the following:

> *Once you give a team the feeling of autonomy or autonomy itself, then you already win 80% of the battle to get them to use it.*

Missing Strategies

Unaddressed barriers also appeared in relation to clinicians' capability, opportunity, and motivation. Various interviewees highlighted shortcomings in contemporary education, noting a lack of emphasis on developing the professional attitudes necessary for effectively using PROMs in VBHC. This includes insufficient focus on understanding the impact of disease on the individual, their social environment, and work (interviewee 20), as well as limitations in shared decision-making, inclusive care, and data-driven improvement work (interviewee 19). Interviewee 14 pointed out that VBHC requires clinicians to take shared responsibility for patient outcomes rather than dismissing certain issues as outside their scope, observing that this attitude is "only really embraced by a few people." Interviewee 18 emphasized another limitation in education, stating the following:

> *It has become checkbox medicine.... [Residents are] afraid to color outside the lines of those protocols, guidelines, and contractual agreements.*

In addition, there seem to be limitations in clinicians' capacity for change and medical leadership (internal document, July 13, 2020). Interviewee 11 noted that clinicians are typically "quite conservative" and find changing behavior "very difficult." Interviewee 20 observed that PROMs resonate more with nurses, who focus more on caring and benefit from longer consultation times, while physicians typically focus on curing. The CST found that departments with a dedicated nurse specialist or physician assistant achieved higher use of PROMs compared to those without such staff. Some subdepartments tasked these nurses with discussing patients' PROM data and flagging any issues for the medical specialist who would see the patient next (internal document, May 11, 2023). However, the CST struggled to leverage this facilitator across the hospital due to budget constraints in local subdepartments to hire nurse specialists or physician assistants and the nonbillability of PROMs discussions by nurses. Overall, the CST believes that patients' experiences of symptoms, functioning, and quality of life should be a central focus in all care activities rather than being narrowed to nurse consultations alone.

There also appeared limitations in the structural integration of PROMs for VBHC into the onboarding of new staff. Survey respondent 39 noted the following:

> *I started working here a year ago, there hasn't been a structured introduction to this topic, I have no experience with it.*

Interviewee 19 explained that despite 2 attempts by the program team to include PROMs in onboarding, these efforts faced rejection due to PROMs not yet being uniformly adopted across all departments.

Behavior change was further hindered by limitations in opportunity and motivation. Clinicians' contemporary feelings of work pressure and well-being were considered to have a negative impact:

> *People feel exhausted, like stretched rubber bands nearing their breaking point.... They think: "What now again? Will this add further strain to me?"* [Interviewee 6]

Interviewee 19 noted the challenge of change fatigue alongside "a pervasive aversion to any form of registration and data handling." A fine line existed between meeting data needs and experiencing an overwhelming availability of data. Interviewee 11 noted a lack of tension and urgency to use PROMs:

> *There is no fire. There are no patients dying if you don't use PROMs.*

Interviewee 18 noted the challenge from the ongoing volume-based rather than value-based payment, impeding clinicians' opportunity to use PROMs to make value-based decisions:

> *We ask people to play a different game of soccer without changing the rules. Resultantly, no one ends up playing differently.*

Objective 3: Key Factors Influencing Clinicians' Use of PROMs From Early 2023

Overview

Despite implementing numerous strategies (objective 2), the hospital continued to face limitations in clinicians' use of PROMs (objective 1). Therefore, as the third objective, we aimed to broaden our understanding of the key factors influencing clinicians' use of PROMs in early 2023. We first highlight the primary motivations driving clinicians to use PROMs, followed by the key factors contributing to suboptimal use.

Reasons to Use PROMs

Survey respondents, excluding those who never attempted using PROMs, were asked to select the most prominent reasons for using (or having used) PROMs. Response options were identified from document analysis. Respondents had the

XSL•FO
RenderX

opportunity to add additional factors, but this option was not used.

Among all survey respondents, 38% (16/42) reported using PROMs primarily because of patient benefits and 36% (15/42) reported using it because of personal benefits (Table 4). Nearly one-third (12/42, 29%) experienced both types of benefits. In addition, 45% (19/42) of respondents used PROMs because they felt it was expected of them, with 31% (13/42) citing this expectation as the sole reason for using PROMs. Only 12% (5/42) used PROMs because of patient requests.

Table 4 also shows responses among 3 user profiles: *nonusers*, *occasional users*, and *frequent users* (refer to the Methods section). Perceived benefits were the primary driver for frequent users, while feeling expected to use PROMs was the most frequently selected reason among occasional users and those who ceased use. Frequent PROM users selected more factors compared to less frequent users, suggesting that they perceived more reasons to use PROMs.

Barriers and Satisfaction

Survey respondents were also asked to select the most prominent factors that impeded their use of PROMs, if experienced, in a multiple-choice question. Among all respondents, key issues were time constraints (24/44, 55%) alongside dashboard functioning limitations (24/44, 55%), followed by forgetting (22/44, 50%; Table 5). Limitations in reinforcement and believing to have a limited impact on improving PROMs domains were rarely selected as reasons impeding PROM use.

Disaggregating into the 3 user profiles, we found that dashboard functioning limitations impeded all groups. While all nonusers perceived time constraints, this factor was less frequently selected among the other profiles. Nonusers often cited a misalignment between PROMs and how they preferred to work as a substantial factor. Among occasional users, forgetting to use PROMs was a critical issue, a challenge also faced by current nonusers. The frequent PROM users felt impeded by the low volume of completed PROMs. Overall, those who used PROMs less frequently or not at all selected more factors compared to more frequent users.

Respondents were also asked to grade the outcomes of implementing PROMs in outpatient care and the implementation process; respectively, average scores were 4.9 and 5.4, both out of 10. Clinicians who used PROMs more frequently were more satisfied (Table 5).

XSL•FO
RenderX

Table 5. Prominent factors that contribute to suboptimal use of patient-reported outcome measures (PROMs), if experienced, both in total and as per 3 user profiles.

Item, post hoc categories, and answer options	All respondents (n=44[a])	User profiles		
		Nonusers (n=11[a])	Occasional users (n=15[a])	Frequent users (n=18)
Most prominent factors contributing to suboptimal use of PROMs, if experienced; n (%)				
Capability				
Finding it difficult or perceiving not to possess necessary skills	8 (18)	3 (27)	4 (27)	1 (6)
Opportunity				
Dashboard functioning issues	24 (55)	8 (73)	8 (53)	8 (44)
Not enough time	24 (55)	11 (100)	8 (53)	5 (28)
Low volume of completed PROMs	17 (39)	4 (36)	5 (33)	8 (44)
A colleague discusses PROMs[b]	2 (5)	0 (0)	2 (13)	0 (0)
Motivation				
Not in my routine, I forget it	22 (50)	8 (73)	10 (67)	4 (22)
Misalignment with how I prefer to work	12 (27)	6 (55)	5 (33)	1 (6)
No added benefits for patients or myself	11 (25)	3 (27)	5 (33)	3 (17)
PROMs do not fit my patient population	7 (16)	3 (27)	2 (13)	2 (11)
I have limited influence on improving PROM domains	4 (9)	0 (0)	3 (20)	1 (6)
Too little reinforcement	2 (5)	(18)	0 (0)	0 (0)
Calculation				
Number of factors selected, mean (SD)	3.0 (1.7)	4.4 (1.3)	3.5 (1.6)	1.8 (1.1)
Average grade from 1 to 10				
PROMs implementation process, mean (SD, range)	5.4 (2.4, 1-9)	2.7 (1.7, 1-6)	5.3 (1.8, 3-8)	7.1 (1.6, 4-9)
PROMs implementation, mean outcomes (SD, range)	4.9 (2.3, 1-9)	2.7 (1.9, 1-6)	4.8 (1.6, 2-8)	6.2 (1.9, 3-9)

[a]There are lower sample population values compared to the total population value due to missing data from incomplete responses.

[b]Some local departments adapted professional roles and workflows (strategy addition [45]) so that physician assistants or nurses were responsible for discussing patients' PROM data and signaling any issues that require attention to the medical specialist who sees the patient next (internal document, May 11, 2023).

Discussion

Overview

Clinicians' effective use of PROM data is crucial for maximizing the benefits of PROMs and ensuring their ongoing use [13-17]; however, it remains challenging to achieve [3,9,10]. This study aimed to (1) explore how clinicians use individual patients' PROM data for value-based outpatient specialty care in a Dutch university hospital, (2) identify organizational strategies used to enhance PROM data use, and (3) uncover key motivations and barriers that persist.

Our findings indicate that clinicians' use of PROM data remains suboptimal despite numerous strategies aimed at enhancing their capability, opportunity, and motivation. We identified shortcomings in the execution quality and comprehensiveness of applied strategies. Factors that influenced PROM use varied among clinicians, and some strategies were hindered or controversial, adding to the complexity of effective change facilitation.

Clinicians' Use of PROMs

On average, between early 2023 and 2024, clinicians accessed PROM data for approximately 3 of 20 (14%) patient responders on the day of the patient's consultation, serving as a proxy for discussing outcomes during the consultation. We found clinicians not trying PROMs (an adoption issue), using PROMs inconsistently (an implementation issue), and stopping the use of PROMs (a maintenance issue), underscoring the inherent challenges in altering clinicians' practices and the need for multifaceted strategies. The persistent challenges in PROM use rates, both in our study and others [3], contrast with a study reporting higher self-reported use rates [24] and clinicians prioritizing discussions about PROMs in outpatient consultations [17].

XSL•FO
RenderX

Strategies to Enhance PROM use

The hospital used 27 (37%) out of 73 ERIC strategies [39] to enhance clinicians' use of PROMs, with the only category not used being financial strategies [40]. Our study also confirmed 5 strategies that scholars proposed as additions to ERIC: communication [47], providing stakeholders with the possibility to attend educational meetings [44], assessing and redesigning workflow [45], obtaining employee's feedback about the implementation plan [46], and planning for outcome evaluation [46].

Strategies appeared to address clinicians' capability, opportunity, and motivation to use PROMs, aligning with the crucial constructs in the COM-B model.

Regarding *capability*, the results of this study confirmed that training clinicians in using PROMs for VBHC is crucial [27]. This underscores the need for integration into (interprofessional) education and ongoing training efforts for current staff as well as effective onboarding for new employees. Effective use of PROMs appears to require more substantial and impactful changes than initially anticipated, requiring a shift in clinicians' roles, professional attitudes, and ingrained practices. These changes include extending accountability and collaboration as well as focusing on care in addition to cure. There appears need for attention to professionalization [50,51], including the development of change capability [52] and nurse or medical leadership [51,53-55]. For training current staff, individualized approaches such as coaching on the job appeared helpful. Moreover, attention should be given to providing clinicians with opportunities for training, considering their busy schedules [44].

Regarding *opportunity*, strategies were directed at creating both social and physical opportunities to use PROMs. The hospital made substantial efforts to simplify and streamline the use of PROMs by adopting strategies identified by other scholars as supportive, such as integrating PROMs into the EHR with graphical interpretation support [10,15,25,56,57], quick copy of PROMs outcomes into patient notes [57], and using PROMs to collect previsit information on medication and lifestyle [56]. Nevertheless, more than half (24/44, 55%) of the clinicians surveyed in early 2023 cited time constraints as a barrier to using PROMs, alongside a temporary issue of long loading times of the PROMs dashboard. Time has been reported as a barrier to PROM use previously [7,24,27]. While PROMs may take more time due to inexperience [31,58], research has also shown that PROMs do not lengthen consultation time [59,60] and some clinicians experience time savings [7,58]. These discrepancies warrant attention. Possibly, emphasis may be needed to manage clinicians' assumptions and perceptions of time. Moreover, it is critical that PROMs and associated dashboarding function as needed, representing a cornerstone. In addition, the current payment system hinders the use of PROMs, requiring clinicians to focus on outcomes while reimbursing based on volume of health services. Moreover, it restricts the roles nurses can assume regarding PROMs, as their PROMs discussions are not always reimbursable.

Regarding motivation, most clinicians were driven by the perceived benefits of PROMs for both themselves and their patients. Smaller groups were motivated to use PROMs due to role expectations and patient–initiated discussions, making these potential levers that hospitals could tap on. This may be particularly important to address, as prior research suggests that professionals may expect patients to initiate discussions about PROMs, while patients believe this responsibility lies with the professionals [26]. We recommend that hospitals clarify the motivations for implementing PROMs in outpatient care and encourage professionals to explore and be driven by their own motivations. Conversely, we found that a quarter (12/44, 27%) of clinicians, especially those who have now quit using PROMs, did not perceive PROMs to align with how they prefer to work. This highlights a potential area for further research.

Challenges also seem to stem from local circumstances, with staff feeling pressured and lacking the tension and evidence needed to change. By measuring patient and professional outcomes and experiences, such as the quality of conversations, pre-PROMs implementation and post-PROMs implementation [61], hospitals could instill a sense of urgency for change and provide the requested evidence.

The positioning of PROMs as either a voluntary tool or a core element of care impacts the range of strategies that can be used. In the observed case, PROM use remained voluntary to prevent their ceremonial use [62]. However, there was an increase in formal pressures, both internally and externally, including government focus on outcomes and the value of care [48,49]. To address limitations related to clinicians viewing PROMs and VBHC as supplementary rather than integral to their work, hospitals may consider embedding these practices into their core operations. To exemplify, integrating PROMs and VBHC into hospital strategy, communications, education, career development discussions, control cycles, and mainstream IT could facilitate institutionalization [50,63].

Challenges in Strategy Selection and Implementation

The hospital selected and implemented strategies to enhance the use of PROM use in a pragmatic and organic manner without a comprehensive and detailed documented program theory. They did not evaluate the effectiveness of the strategies used. However, finding linear and singular cause-and-effect linkages may be unfeasible given the complexities of change [64]. Complex change typically includes multiple and interacting factors influencing behavior as well as characteristics of nonlinearity and path dependency [64-66].

Challenges in effective change facilitation included unfeasible strategies due to limitations in manpower and budget as well as constraints inherent in the staged implementation approach. For example, the latter necessitated a simultaneous focus on adoption, implementation, and maintenance. Moreover, certain strategies had unintended consequences. For example, initially implementing PROMs for a subset of patients facilitated adoption but ultimately led to infrequent use. Half (22/44, 50%) of the clinicians cited forgetting to use PROMs because it is not a part of their routine, which echoes prior findings [27] and presents a substantial risk to maintaining PROM use [67]. Scheduling consultations for which PROMs are sent together may be a potential solution worth investigating further.

XSL•FO
RenderX

Furthermore, we identified challenges stemming from the diffusing nature of factors influencing clinicians' use of PROMs. We identified variations in critical factors among frequent users, occasional users, and those who have never attempted or have quit using PROMs. For instance, nonusers more often cited perceived time constraints as a prominent barrier compared to frequent users and felt that PROMs did not align with their preferred way of working. Consequently, increased understanding and tailoring of strategies on the basis of behavioral determinants appears promising. In addition, we anticipate that the hospital's approach could benefit from increased focus on achieving higher levels of workforce activation [68]. Above all, a CST can only facilitate and encourage but not manage clinicians' use of PROMs; each individual clinician should ultimately be engaged and take leadership in using PROMs.

Future Research

Future research could explore the factors contributing to differences in the adoption, implementation, and maintenance of PROMs across subdepartments. Conducting interviews could help delve deeper into specific findings, particularly in areas such as routines, time perceptions, and professionalization. Investigating the association between clinicians' characteristics and their use of PROMs would be valuable. Deeper exploration of specific strategies, including their adaptations over time, for example, using Justin Smith's Longitudinal Implementation Strategy Tracking System [69], could also provide valuable insights.

Limitations and Strengths

Regarding the scope of the study, we focused on clinicians' use of PROM data, but their role is broader and likely includes motivating nonresponding patients to complete subsequent PROMs. We focused on reporting whether strategies were used rather than evaluating their execution and effectiveness. Our emphasis on organizational-level strategies led us to overlook decentralized strategies and their potential interactions. Regarding our data sources, the low survey response rate necessitates cautious interpretation. However, survey respondents reported varying frequencies of PROM use and satisfaction, which suggests the inclusion of a diverse range of perspectives. The timing of the survey followed a period of limited dashboard functioning, which likely influenced the results. Interview data were collected as part of a broader research question, so not all respondents systematically provided input on strategies and challenges associated with PROM use. This limitation hindered our ability to quantify the percentage of interviewees supporting specific arguments and potentially limited completeness. In addition, there may be missing documents that were not captured by the screening keywords. The specification of strategies and the application of ERIC coding may be open to interpretation and hence introduce bias. However, we aimed to enhance rigor by using a coding manual.

Despite these limitations, our study is a pioneering effort to investigate the underuse of PROMs and document practice-developed strategies to enhance their use among clinicians. Furthermore, our study uniquely highlights how implementation factors impact clinicians differently in a hospital setting, suggesting the potential benefit of tailored strategies.

Conclusions

Achieving clinicians' use of PROMs in value-based outpatient care has remained challenging despite various strategies aimed at enhancing their capability, opportunity, and motivation. Their experienced barriers and motivations were diverse and sometimes difficult to address, adding to the complexity of effective facilitation. PROM use may require more substantial changes than initially expected, necessitating a shift in clinicians' professional attitudes and practices. Hospitals can facilitate and encourage rather than manage clinicians' genuine use of PROMs. They must prioritize efforts to engage clinicians with PROMs for value-based outpatient care. Specific attention to their professionalization is warranted. Tailored strategies designed to address differences in clinicians' needs and motivation hold promise for future efforts and could also facilitate the successful implementation of other eHealth initiatives.

Acknowledgments

Erasmus School of Health Policy & Management, Erasmus University Rotterdam, the Netherlands, funded this study. The authors are grateful to all individuals who participated in this research. They would like to thank all those who contributed to the implementation of patient-related outcome measures. The authors would also like to thank the reviewers for their valuable suggestions.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

The study was conceptualized by VvE, MB-S, KA, and IB. VvE was involved in the study's methodology, validation, formal analysis, investigation, data curation, writing of the original draft, visualization, and project administration. MB-S was involved in the study's methodology and supervision. KA and IB supervised the study, and IP and MDH-A were involved in the investigation and resource collection. All the authors were involved in reviewing and editing the manuscript.

XSL•FO
RenderX

Conflicts of Interest

Authors MDH-A and IP disclose employment at the studied hospital and membership of the value-based health care implementation team. They did not participate in data collection and analysis to guarantee minimal bias. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Patient-reported outcome measure dashboarding visualizations.
[DOCX File , 414 KB - jmir_v26i1e60306_app1.docx]

Multimedia Appendix 2
Survey.
[DOCX File , 52 KB - jmir_v26i1e60306_app2.docx]

Multimedia Appendix 3
Semistructured interview questions.
[DOCX File , 21 KB - jmir_v26i1e60306_app3.docx]

Multimedia Appendix 4
Specification of strategies.
[DOCX File , 53 KB - jmir_v26i1e60306_app4.docx]

References

1. Bates DW, Deerberg-Wittram J, Katz G, Braeger K, Hirsch LS, Kaplan GS, et al. Using publicly reported global hospital rankings to improve dissemination of patient-reported outcome measures (PROMs). NEJM Catal 2023 Oct 18;4(11):51. [doi: 10.1056/cat.23.0097]
2. Calvert M, Kyte D, Price G, Valderas JM, Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. BMJ 2019 Jan 24;364:k5267. [doi: 10.1136/bmj.k5267] [Medline: 30679170]
3. Hjollund NH. Fifteen years' use of patient-reported outcome measures at the group and patient levels: trend analysis. J Med Internet Res 2019 Sep 30;21(9):e15856 [FREE Full text] [doi: 10.2196/15856] [Medline: 31573943]
4. Snyder CF, Jensen RE, Segal JB, Wu AW. Patient-reported outcomes (PROs): putting the patient perspective in patient-centered outcomes research. Med Care 2013 Aug;51(8 Suppl 3):S73-S79 [FREE Full text] [doi: 10.1097/MLR.0b013e31829b1d84] [Medline: 23774513]
5. Acquadro C, Berzon R, Dubois D, Leidy NK, Marquis P, Revicki D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. Value Health 2003;6(5):522-531 [FREE Full text] [doi: 10.1046/j.1524-4733.2003.65309.x] [Medline: 14627058]
6. Damman OC, Jani A, de Jong BA, Becker A, Metz MJ, de Bruijne MC, et al. The use of PROMs and shared decision-making in medical encounters with patients: an opportunity to deliver value-based health care to patients. J Eval Clin Pract 2020 Apr;26(2):524-540 [FREE Full text] [doi: 10.1111/jep.13321] [Medline: 31840346]
7. van Engen V, Bonfrer I, Ahaus K, Buljac-Samardzic M. Identifying consensus on activities that underpin value-based healthcare in outpatient specialty consultations, among clinicians. Patient Educ Couns 2023 Apr;109:107642 [FREE Full text] [doi: 10.1016/j.pec.2023.107642] [Medline: 36696878]
8. Jani A, Jungmann S, Gray M. Shifting to triple value healthcare: reflections from England. Z Evid Fortbild Qual Gesundhwes 2018 Feb;130:2. [doi: 10.1016/j.zefq.2018.01.002] [Medline: 29402597]
9. Wolff AC, Dresselhuis A, Hejazi S, Dixon D, Gibson D, Howard AF, et al. Healthcare provider characteristics that influence the implementation of individual-level patient-centered outcome measure (PROM) and patient-reported experience measure (PREM) data across practice settings: a protocol for a mixed methods systematic review with a narrative synthesis. Syst Rev 2021 Jun 09;10(1):169 [FREE Full text] [doi: 10.1186/s13643-021-01725-2] [Medline: 34108024]
10. van Egdom LS, Oemrawsingh A, Verweij LM, Lingsma HF, Koppert LB, Verhoef C, et al. Implementing patient-reported outcome measures in clinical breast cancer care: a systematic review. Value Health 2019 Oct;22(10):1197-1226 [FREE Full text] [doi: 10.1016/j.jval.2019.04.1927] [Medline: 31563263]
11. Harvey G, Kitson A, Harvey G. Implementing Evidence-Based Practice in Healthcare: A Facilitation Guide. New York, NY: Routledge; 2015.
12. Stover AM, Haverman L, van Oers HA, Greenhalgh J, Potter CM, ISOQOL PROMs/PREMs in Clinical Practice Implementation Science Work Group. Using an implementation science approach to implement and evaluate patient-reported outcome measures (PROM) initiatives in routine care settings. Qual Life Res 2020 Jul 10;30(11):3015-3033. [doi: 10.1007/s11136-020-02564-9] [Medline: 32651805]

XSL•FO
RenderX

13. Aiyegbusi OL, Roydhouse J, Rivera SC, Kamudoni P, Schache P, Wilson R, et al. Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection. Nat Commun 2022 Oct 12;13(1):6026 [FREE Full text] [doi: 10.1038/s41467-022-33826-4] [Medline: 36224187]
14. Unni E, Coles T, Lavallee DC, Freel J, Roberts N, Absolom K. Patient adherence to patient-reported outcome measure (PROM) completion in clinical care: current understanding and future recommendations. Qual Life Res 2023 Sep 11;33(1):28111-28190. [doi: 10.1007/s11136-023-03505-y] [Medline: 37695476]
15. Øvretveit J, Zubkoff L, Nelson EC, Frampton S, Knudsen JL, Zimlichman E. Using patient-reported outcome measurement to improve patient care. Int J Qual Health Care 2017 Oct 01;29(6):874-879. [doi: 10.1093/intqhc/mzx108] [Medline: 29025001]
16. Hjollund NH, Valderas JM, Kyte D, Calvert MJ. Health data processes: a framework for analyzing and discussing efficient use and reuse of health data with a focus on patient-reported outcome measures. J Med Internet Res 2019 May 21;21(5):e12412 [FREE Full text] [doi: 10.2196/12412] [Medline: 31115347]
17. de Jel DV, Young-Afat DA, Ooms-Renckens MM, Smeele LE, Rakhorst HA, DHNA study group. Patients' and healthcare professionals' perspectives on better use of patient-reported outcome measures in head and neck cancer. Value Health 2023 Aug;26(8):1210-1216 [FREE Full text] [doi: 10.1016/j.jval.2023.04.009] [Medline: 37164042]
18. Van Der Wees PJ, Nijhuis-Van Der Sanden MW, Ayanian JZ, Black N, Westert GP, Schneider EC. Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries. Milbank Q 2014 Dec;92(4):754-775 [FREE Full text] [doi: 10.1111/1468-0009.12091] [Medline: 25492603]
19. Foster A, Croot L, Brazier J, Harris J, O'Cathain A. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. J Patient Rep Outcomes 2018 Dec;2:46 [FREE Full text] [doi: 10.1186/s41687-018-0072-3] [Medline: 30363333]
20. Gelkopf M, Mazor Y, Roe D. A systematic review of patient-reported outcome measurement (PROM) and provider assessment in mental health: goals, implementation, setting, measurement characteristics and barriers. Int J Qual Health Care 2021 Mar 05;34(Suppl 1):ii13–ii27. [doi: 10.1093/intqhc/mzz133] [Medline: 32159763]
21. Silveira Bianchim M, Crane E, Jones A, Neukirchinger B, Roberts G, Mclaughlin L, et al. The implementation, use and impact of patient reported outcome measures in value-based healthcare programmes: a scoping review. PLoS One 2023;18(12):e0290976 [FREE Full text] [doi: 10.1371/journal.pone.0290976] [Medline: 38055759]
22. Adeoye-Olatunde OA, Curran GM, Jaynes HA, Hillman LA, Sangasubana N, Chewning BA, et al. Preparing for the spread of patient-reported outcome (PRO) data collection from primary care to community pharmacy: a mixed-methods study. Implement Sci Commun 2022 Mar 14;3(1):29 [FREE Full text] [doi: 10.1186/s43058-022-00277-3] [Medline: 35287764]
23. van Oers HA, Teela L, Schepers SA, Grootenhuis MA, Haverman L, ISOQOL PROMs and PREMs in Clinical Practice Implementation Science Group. A retrospective assessment of the KLIK PROM portal implementation using the Consolidated Framework for Implementation Research (CFIR). Qual Life Res 2021 Nov;30(11):3049-3061 [FREE Full text] [doi: 10.1007/s11136-020-02586-3] [Medline: 32803626]
24. Teela L, van Muilekom MM, Kooij LH, Gathier AW, van Goudoever JB, Grootenhuis MA, et al. Clinicians' perspective on the implemented KLIK PROM portal in clinical practice. Qual Life Res 2020 May 28;30(11):3267-3277. [doi: 10.1007/s11136-020-02522-5] [Medline: 32468402]
25. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. BMJ Qual Saf 2014 Jun;23(6):508-518. [doi: 10.1136/bmjqs-2013-002524] [Medline: 24505110]
26. Damman OC, Verbiest MEA, Vonk SI, Berendse HW, Bloem BR, de Bruijne MC, et al. Using PROMs during routine medical consultations: The perspectives of people with Parkinson's disease and their health professionals. Health Expect 2019 Oct;22(5):939-951 [FREE Full text] [doi: 10.1111/hex.12899] [Medline: 31199574]
27. Westerink HJ, Bresser CC, Garvelink MM, van Uden-Kraan CF, Zouitni O, Bart HA, Santeon Patient Participation Study Group. The use of outcome data in patient consultations from the healthcare professionals' and patients' perspectives: a mixed methods study. Patient Educ Couns 2024 Jan;118:108043 [FREE Full text] [doi: 10.1016/j.pec.2023.108043] [Medline: 37925975]
28. Fernández-Salido M, Alhambra-Borrás T, Casanova G, Garcés-Ferrer J. Value-based healthcare delivery: a scoping review. Int J Environ Res Public Health 2024 Jan 25;21(2):134 [FREE Full text] [doi: 10.3390/ijerph21020134] [Medline: 38397625]
29. Henderson NL, Basch EM, Rocque GB. Learning to overcome barriers to electronic patient-reported outcome implementation: lessons from the field. JCO Oncol Pract 2024 Jun;20(6):751-753. [doi: 10.1200/OP.24.00104] [Medline: 38502891]
30. Roberts NA, Janda M, Stover AM, Alexander KE, Wyld D, Mudge A, ISOQOL PROMs/PREMs in Clinical Practice Implementation Science Work Group. The utility of the implementation science framework "integrated promoting action on research implementation in health services" (i-PARIHS) and the facilitator role for introducing patient-reported outcome measures (PROMs) in a medical oncology outpatient department. Qual Life Res 2021 Nov;30(11):3063-3071 [FREE Full text] [doi: 10.1007/s11136-020-02669-1] [Medline: 33089474]
31. Offerman M, Dronkers E, Baatenburg de Jong R. Healthcare monitor supports value-based health care for head and neck cancer patients. In: van Weert N, Hazelzet J, editors. Personalized Speciality Care. Cham, Switzerland: Springer; 2021:55-62.

XSL•FO
RenderX

32. van Egdom LS, Lagendijk M, van der Kemp MH, van Dam JH, Mureau MA, Hazelzet JA, et al. Implementation of value based breast cancer care. Eur J Surg Oncol 2019 Jul;45(7):1163-1170. [doi: 10.1016/j.ejso.2019.01.007] [Medline: 30638807]
33. van Engen V, Buljac-Samardzic M, Baatenburg de Jong R, Braithwaite J, Ahaus K, Den Hollander-Ardon M, et al. A decade of change towards value-based health care at a Dutch University Hospital: a complexity-informed process study. Health Res Policy Syst 2024 Aug 05;22(1):94 [FREE Full text] [doi: 10.1186/s12961-024-01181-z] [Medline: 39103922]
34. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880 [FREE Full text] [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
35. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, et al. PROMIS adult health profiles: efficient short-form measures of seven health domains. Value Health 2019 May;22(5):537-544. [doi: 10.1016/j.jval.2019.02.004] [Medline: 31104731]
36. Petersen MA, Aaronson NK, Arraras JI, Chie W, Conroy T, Costantini A, European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group. The EORTC CAT Core-the computer adaptive version of the EORTC QLQ-C30 questionnaire. Eur J Cancer 2018 Sep;100:8-16 [FREE Full text] [doi: 10.1016/j.ejca.2018.04.016] [Medline: 29936066]
37. McHugh S, Sinnott C, Racine E, Timmons S, Byrne M, Kearney PM. 'Around the edges': using behaviour change techniques to characterise a multilevel implementation strategy for a fall prevention programme. Implement Sci 2018 Aug 20;13(1):113 [FREE Full text] [doi: 10.1186/s13012-018-0798-6] [Medline: 30126418]
38. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [Medline: 10474547]
39. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
40. Waltz TJ, Powell BJ, Matthieu MM, Damschroder LJ, Chinman MJ, Smith JL, et al. Use of concept mapping to characterize relationships among implementation strategies and assess their feasibility and importance: results from the Expert Recommendations for Implementing Change (ERIC) study. Implement Sci 2015 Aug 07;10:109 [FREE Full text] [doi: 10.1186/s13012-015-0295-0] [Medline: 26249843]
41. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
42. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015;10:53 [FREE Full text] [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
43. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 01;8:139 [FREE Full text] [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
44. Fridberg H, Wallin L, Tistad M. Tracking, naming, specifying, and comparing implementation strategies for person-centred care in a real-world setting: a case study with seven embedded units. BMC Health Serv Res 2022 Nov 24;22(1):1409 [FREE Full text] [doi: 10.1186/s12913-022-08846-x] [Medline: 36424611]
45. Perry CK, Damschroder LJ, Hemler JR, Woodson TT, Ono SS, Cohen DJ. Specifying and comparing implementation strategies across seven large implementation interventions: a practical application of theory. Implement Sci 2019 Dec 21;14(1):32 [FREE Full text] [doi: 10.1186/s13012-019-0876-4] [Medline: 30898133]
46. Bunger AC, Powell BJ, Robertson HA, MacDowell H, Birken SA, Shea C. Tracking implementation strategies: a description of a practical approach and early findings. Health Res Policy Syst 2017 Feb 23;15(1):15 [FREE Full text] [doi: 10.1186/s12961-017-0175-y] [Medline: 28231801]
47. Boyd MR, Powell BJ, Endicott D, Lewis CC. A method for tracking implementation strategies: an exemplar implementing measurement-based care in community behavioral health clinics. Behav Ther 2018 Jul;49(4):525-537 [FREE Full text] [doi: 10.1016/j.beth.2017.11.012] [Medline: 29937255]
48. Outcome based healthcare 2018-2022. Ministry of Healthcare Welfare and Sports. 2018. URL: https://www.government.nl/topics/quality-of-healthcare/documents/reports/2018/07/02/outcome-based-healthcare-2018-2022 [accessed 2024-04-29]
49. Integraal zorg akkoord [integrale care agreement]. Ministry of Healthcare Welfare and Sports. 2022. URL: https://www.rijksoverheid.nl/documenten/rapporten/2022/09/16/integraal-zorgakkoord-samen-werken-aan-gezonde-zorg [accessed 2024-04-29]
50. Reay T, Goodrick EH. Institutionalization and professionalization. In: Ferlie E, Montgomery K, editors. The Oxford Handbook of Health Care Management. Oxford, UK: Oxford University Press; 2016.
51. Keijser W, Huq JL, Reay T. Enacting medical leadership to address wicked problems. BMJ Lead 2020 Mar 27;4(1):12-17. [doi: 10.1136/leader-2019-000137]
52. Long JC, Roberts N, Francis-Auton E, Sarkies MN, Nguyen HM, Westbrook JI, et al. Implementation of large, multi-site hospital interventions: a realist evaluation of strategies for developing capability. BMC Health Serv Res 2024 Mar 06;24(1):303 [FREE Full text] [doi: 10.1186/s12913-024-10721-w] [Medline: 38448960]
53. Berghout MA, Oldenhof L, van der Scheer WK, Hilders CG. From context to contexting: professional identity un/doing in a medical leadership development programme. Sociol Health Illn 2020 Feb;42(2):359-378 [FREE Full text] [doi: 10.1111/1467-9566.13007] [Medline: 31646655]

XSL•FO
RenderX

54. Keijser WA, Handgraaf HJ, Isfordink LM, Janmaat VT, Vergroesen PA, Verkade JM, et al. Development of a national medical leadership competency framework: the Dutch approach. BMC Med Educ 2019 Nov 28;19(1):441 [FREE Full text] [doi: 10.1186/s12909-019-1800-y] [Medline: 31779632]
55. Jeffers BR, Astroth KS. The clinical nurse leader: prepared for an era of healthcare reform. Nurs Forum 2013;48(3):223-229. [doi: 10.1111/nuf.12032] [Medline: 23889201]
56. Wagle NW. Implementing patient-reported outcome measures. NEJM Catal 2017:5 [FREE Full text] [doi: 10.1056/CAT.17.0373]
57. Rocque GB, Dent DN, Ingram SA, Caston NE, Thigpen HB, Lalor FR, et al. Adaptation of remote symptom monitoring using electronic patient-reported outcomes for implementation in real-world settings. JCO Oncol Pract 2022 Oct 28:OP2200360. [doi: 10.1200/OP.22.00360] [Medline: 36306496]
58. van der Willik EM, Milders J, Bart JA, Bos WJ, van Ittersum FJ, Ten Dam MA, et al. Discussing results of patient-reported outcome measures (PROMs) between patients and healthcare professionals in routine dialysis care: a qualitative study. BMJ Open 2022 Nov 17;12(11):e067044 [FREE Full text] [doi: 10.1136/bmjopen-2022-067044] [Medline: 36396312]
59. Engelen V, Detmar S, Koopman H, Maurice-Stam H, Caron H, Hoogerbrugge P, et al. Reporting health-related quality of life scores to physicians during routine follow-up visits of pediatric oncology patients: is it effective? Pediatr Blood Cancer 2012 May;58(5):766-774. [doi: 10.1002/pbc.23158] [Medline: 21584933]
60. Stover A, Irwin DE, Chen RC, Chera BS, Mayer DK, Muss HB, et al. Integrating patient-reported outcome measures into routine cancer care: cancer patients' and clinicians' perceptions of acceptability and value. EGEMS (Wash DC) 2015;3(1):1169 [FREE Full text] [doi: 10.13063/2327-9214.1169] [Medline: 26557724]
61. Grol R, Wensing M. Effective implementation of change in healthcare: a systematic approach. In: Wensing M, Grol R, Grimshaw J, editors. Improving Patient Care: The Implementation of Change in Health Care. Hoboken, NJ: Wiley-Blackwell; 2020:45-71.
62. Greenwood R, Raynard M, Kodeih F, Micelotta ER, Lounsbury M. Institutional complexity and organizational responses. Acad Manag Ann 2011 Jun;5(1):317-371. [doi: 10.5465/19416520.2011.590299]
63. Reay T, Chreim S, Golden‐Biddle K, Goodrick E, Williams BE, Casebeer A, et al. Transforming new ideas into practice: an activity based perspective on the institutionalization of practices. J Manag Stud 2013 Aug 16;50(6):963-990. [doi: 10.1111/joms.12039]
64. Pettigrew AM. Longitudinal field research on change: theory and practice. Organ Sci 1990 Aug;1(3):267-292. [doi: 10.1287/orsc.1.3.267]
65. Braithwaite J, Churruca K, Long JC, Ellis LA, Herkes J. When complexity science meets implementation science: a theoretical and empirical analysis of systems change. BMC Med 2018 Apr 30;16(1):63 [FREE Full text] [doi: 10.1186/s12916-018-1057-z] [Medline: 29706132]
66. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Dec 20;16(1):95 [FREE Full text] [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
67. Rogers E. Diffusions Of Innovations. 2nd edition. Cambridge, MA: Simon & Schuster; 1971.
68. Gray CF, Parvataneni HK, Bozic KJ. Value-based healthcare: "physician activation": healthcare transformation requires physician engagement and leadership. Clin Orthop Relat Res 2020 May;478(5):954-957 [FREE Full text] [doi: 10.1097/CORR.0000000000001234] [Medline: 32217862]
69. Smith JD, Norton WE, Mitchell SA, Cronin C, Hassett MJ, Ridgeway JL, et al. The longitudinal implementation strategy tracking system (LISTS): feasibility, usability, and pilot testing of a novel method. Implement Sci Commun 2023 Nov 28;4(1):153. [doi: 10.1186/S43058-023-00529-W]

Abbreviations

CST: central support team
EHR: electronic health record
Erasmus MC: Erasmus Medical Center
ERIC: Expert Recommendations for Implementing Change
PROM: patient-reported outcome measure
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
VBHC: value-based health care

XSL•FO
RenderX

Edited by G Tsafnat; submitted 07.05.24; peer-reviewed by T Broughton, N Henderson, O Eilayyan; comments to author 30.06.24; revised version received 05.08.24; accepted 31.08.24; published 18.10.24.

Please cite as:
van Engen V, Bonfrer I, Ahaus K, Den Hollander-Ardon M, Peters I, Buljac-Samardzic M
Enhancing Clinicians' Use of Electronic Patient-Reported Outcome Measures in Outpatient Care: Mixed Methods Study
J Med Internet Res 2024;26:e60306
URL: https://www.jmir.org/2024/1/e60306
doi:10.2196/60306
PMID:39422999

©Veerle van Engen, Igna Bonfrer, Kees Ahaus, Monique Den Hollander-Ardon, Ingrid Peters, Martina Buljac-Samardzic. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Association Between Linguistic Characteristics of Physicians' Communication and Their Economic Returns: Mixed Method Study

Shuang Geng[1], PhD; Yuqin He[1], BSc; Liezhen Duan[1], BSc; Chen Yang[1], PhD; Xusheng Wu[2], MSc; Gemin Liang[1], BSc; Ben Niu[1], DPhil

[1]College of Management, Shenzhen University, Shenzhen, China

[2]Shenzhen Health Development Research and Data Management Center, Shenzhen, China

Corresponding Author:
Chen Yang, PhD
College of Management
Shenzhen University
Xueyuan Road 1066
Shenzhen, 518055
China
Phone: 86 18565859968
Email: yangc@szu.edu.cn

Abstract

Background: Web-based health care has the potential to improve health care access and convenience for patients with limited mobility, but its success depends on active physician participation. The economic returns of internet-based health care initiatives are an important factor that can motivate physicians to continue their participation. Although several studies have examined the communication patterns and influences of web-based health consultations, the correlation between physicians' communication characteristics and their economic returns remains unexplored.

Objective: This study aims to investigate how the linguistic features of 2 modes of physician-patient communication, instrumental and affective, determine the physician's economic returns, measured by the honorarium their patients agree to pay per consultation. We also examined the moderating effects of communication media (web-based text messages and voice messages) and the compounding effects of different communication features on economic returns.

Methods: We collected 40,563 web-based consultations from 528 physicians across 4 disease specialties on a large, web-based health care platform in China. Communication features were extracted using linguistic inquiry and word count, and we used multivariable linear regression and K-means clustering to analyze the data.

Results: We found that the use of cognitive processing language (ie, words related to insight, causation, tentativeness, and certainty) in instrumental communication and positive emotion–related words in affective communication were positively associated with the economic returns of physicians. However, the extensive use of discrepancy-related words could generate adverse effects. We also found that the use of voice messages for service delivery magnified the effects of cognitive processing language but did not moderate the effects of affective processing language. The highest economic returns were associated with consultations in which the physicians used few expressions related to negative emotion; used more terms associated with positive emotions; and later, used instrumental communication language.

Conclusions: Our study provides empirical evidence about the relationship between physicians' communication characteristics and their economic returns. It contributes to a better understanding of patient-physician interactions from a professional-client perspective and has practical implications for physicians and web-based health care platform executives.

(J Med Internet Res 2024;26:e42850) doi:10.2196/42850

KEYWORDS

web-based health care; instrumental communication; affective communication; linguistic features; economic returns; linguistic inquiry and word count

XSL•FO
RenderX

Introduction

Background

Web-based health care platforms offer environments where patients can consult physicians and pay for their services remotely. These platforms are particularly helpful for patients residing in rural areas with limited access to medical resources and patients with limited mobility [1]. Physicians also benefit from providing web-based consultations, both in terms of economic returns and social returns, such as improving their reputation [2]. In recent years, web-based health care has been upheld as a national health priority in China. The number of web-based hospitals surged to 2700, and the number of users surged to 360 million in 2022 [3]. Physicians affiliated with offline hospitals can also provide services on third-party health care platforms, such as the Dingxiang Doctor and Haodf websites [4,5]. Unlike offline service prices, web-based consultation prices are not governed by certain pricing standards on these third-party health care platforms [6]. Thus, physicians can charge higher or lower fees per web-based consultation than their offline consultations. Meanwhile, patients can register as platform users, consult physicians, and make payment. Most of these health care platforms offer both synchronous telemedicine and asynchronous, message-based consultation services. We focus specifically on asynchronous, message-based consultations on the third-party health care platforms.

The quality of patient-physician interactions is vital for consultation outcome and patient satisfaction [7,8]. Although many studies have investigated the communication styles and features during web-based health care consultation [7-10], scant attention has been devoted to its associations with the economic returns of physicians. Characterizing physician-patient communications at a more granular level and exploring their associations with physicians' economic returns can provide important guidance for physicians to improve their communication skills and economic benefits. The findings also have implications for the amelioration of platform incentive mechanisms.

Previous studies have identified 2 types of physician-patient interactions during web-based consultations: affective, which focuses on expressing care for the patients [8], and instrumental, which focuses on addressing the patients' health problems [11]. Instrumental interactions typically involve discussions about the severity of the disease or syndrome, its causes, and potential treatment plans [12,13]. They are problem-solving oriented and emphasize the provision of medical information and expertise to patients. These 2 types of interactions are the 2 components of the professional-client interaction theory proposed by Ben-Sira [11] and are crucial for effective web-based health care consultations [14].

This study investigated the associations between these 2 types of interactions and the economic returns of physicians in the context of the Chinese web-based health care system. Specifically, we focused on asynchronous, text message–based consultation services, in which physicians are paid on a per-consultation basis. Physicians usually receive a larger portion (ie, 90%) of the consultation fee, with the remaining portion paid to the platform. We measured the economic returns of the physician as the consultation payment on a per-consultation basis. On the one hand, the consultation prices are initially established by physicians and range from RMB 30 (US $4.19) to RMB 699 (US $97.65) [4]. However, patients can make a choice among physicians, and they are only willing to pay a higher price for better service. Thus, physicians need to adjust the consultation prices to maintain the patient base, and the economic returns are regarded as being mutually determined by the patient and physician. During the consultation process, physicians can choose to use text or voice message for service delivery. Therefore, besides the association between communication characteristics and economic returns, we were also interested in how the associations change vis-à-vis different communication media (text vs voice).

Instrumental and Affective Linguistic Features

Communication quality is critical for patient satisfaction and effective use of medical resources [15]. Previous studies have investigated the quality of patient-physician interactions using self-reported survey [7,16-18], statistical method [9,19,20], and manual labeling combined with text classification [14]. We derived insights from sociopsychological studies and used Linguistic Inquiry and Word Count (LIWC) to extract the linguistic indicators. LIWC provides a psychometrically validated dictionary comprising approximately 6400 words, word stems, and selected emoticons [21,22]. So far, LIWC has been widely used to identify emotional and cognitive dimensions in social, health care, and personality psychology research [23-26]. For instance, Liu et al [27] used LIWC to ascertain positive emotion words in replies provided by physicians to indicate their emotional support to patients.

In instrumental interactions, physicians demonstrate their expertise and cognitive thinking process. Hence, physicians tend to use words related to the cognitive process, presenting linguistic features regarding insight, causation, discrepancy, tentativeness, and certainty to deliver disease knowledge to patients [28]. Affective interactions articulate the emotions of physicians including the emotional support they provide to their patients to alleviate patient anxiety [29]. Notably, the emotional expressions of physicians could comprise diverse feelings such as happiness, anxiety, anger, and sadness. Thus, we included both positive and negative emotional linguistic features for affective interactions. Specifically, we use a simplified Chinese LIWC dictionary established according to the LIWC lexicon and its Chinese version [30-32]. In total, there are 9 linguistic features for the 2 types of interactions, as shown in Figure 1. Example words for each feature are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. The research model. H1: hypothesis 1; H2: hypothesis 2; H3: hypothesis 3; H4: hypothesis 4.

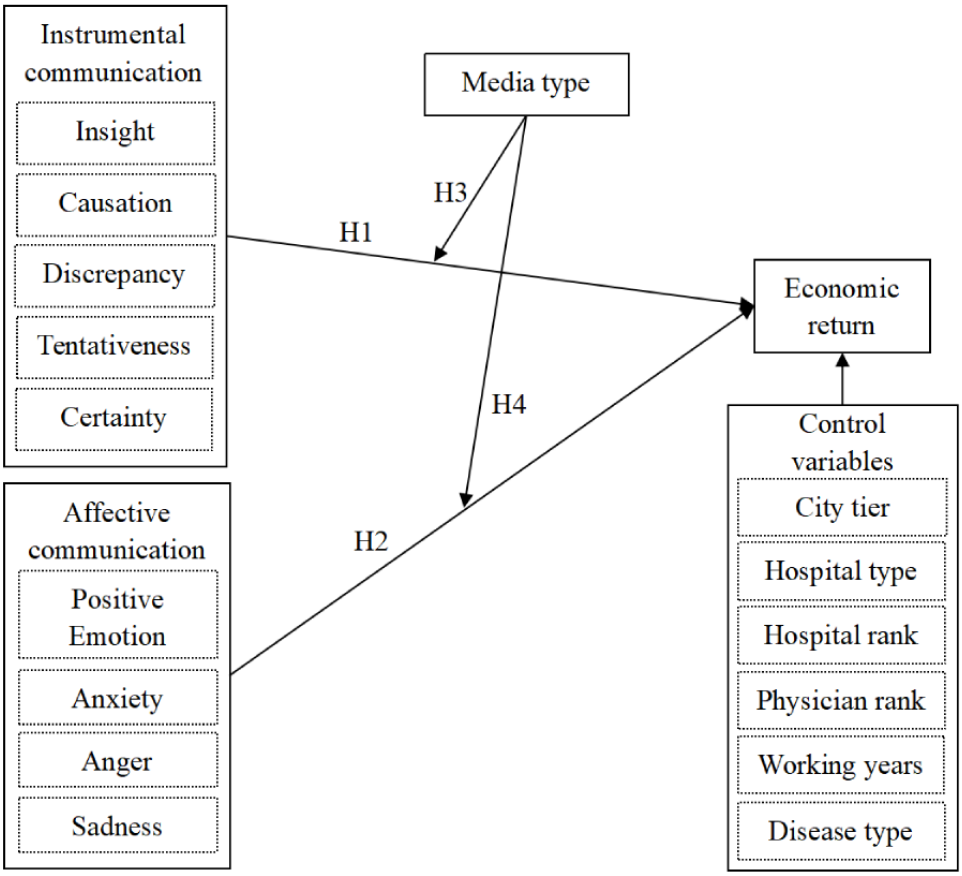

Methods

Research Model

Figure 1 displays our conceptual model. This study explored the associations between the communication behaviors of physicians and their economic returns in 3 stages. First, it probed the impact of instrumental and affective communication of physicians on their economic returns. Second, it investigated the potential moderating effects of communication media (voice vs text). Third, it investigated the interaction patterns among different communication features and identified the pattern yielding the highest economic returns.

Instrumental Interactions by Physicians and Their Economic Returns

Patients seek knowledge related to their health conditions through consultations with physicians because of their professional capital [33]. Thus, physicians are expected to provide disease-specific information, suggestions, or guidance to patients to improve communication quality and patient satisfaction. Physicians' instrumental behaviors manifest mainly through the use of words related to the cognitive process that facilitate the inculcation of due medical knowledge in patients [34]. Thus, instrumental interactions are vital for physicians to achieve a consensus with their patients regarding relevant disease knowledge, treatment plans, and problem solutions.

Economic returns incentivize physicians by satisfying their financial needs. In our research context, health care platforms allow physicians to levy consultation charges as they deem fit [6]; patients can also offer monetary gifts to their physicians. According to the capital exchange theory [2], physicians use their decision capital to perform exchange actions. To a certain degree, the consultation service charges reflect the self-evaluation of physicians vis-à-vis their decision capital and service quality. In contrast, patients may attribute high expectations for the consultation process and outcomes when they agree to pay high service charges. Therefore, we postulated that high economic returns could correlate to more intense instrumental interactions. Specifically, we hypothesized the following: The instrumental communication features of physicians positively influence their economic returns (hypothesis 1).

Affective Interactions by Physicians and Their Economic Returns

Physicians' affective communications involve the expression of care about the feelings of their patients as human beings rather than medical cases [11]. During the consultation process, patients are often emotionally occupied and could unintentionally express their emotional concerns. Many studies have emphasized the psychological needs of patients, citing anxiety and distress as critical issues for health care services [7,8,14,35]. The expression of feelings such as empathy, sympathy, caring, and concern to patients could help mitigate the stress or anxiety induced in patients by their health issues [36]. Hence, patient satisfaction is also associated with the affective behaviors of physicians. In addition, many patients are ignorant of the complexities and technical challenges of their treatment solutions [6,11]. Hence, affective behaviors by

XSL•FO
RenderX

physicians can influence patients' perceptions about their service quality. Patients may place higher expectations for web-based consultation with higher charges in terms of the sensed emotional support. Therefore, we hypothesized the following: The affective communication features of physicians positively influence their economic returns (hypothesis 2).

The Moderating Effects of Communication Media

Physicians and patients communicate primarily via textual and vocal means in asynchronous web-based health consultations. The media synchronicity theory asserts that media characteristics function significantly in information transmission and processing [37]. Vocal communication can deliver the speech signals of physicians, such as pronunciation and intonation, which textual communication using character data cannot [38]. Therefore, vocal communication allows patients to discern the subtleties of the thoughts or concerns of their physicians about their diseases. Moreover, vocal communication also adds to the multiplicity of cues to support complex tasks [14]. In addition, physicians can use vocal communication to deliver immediate and convenient feedback because typing messages is more time consuming. However, it cannot be presumed that vocal communication is inherently better than textual communication because the latter allows physicians time to think about the inquiry in great depth.

Communication comprises 2 primary processes according to the media synchronicity theory: the conveyance of information and the convergence of meaning [37]. Thus, we relate instrumental communication to the process of conveying information and associate affective communication with the process of achieving the convergence of meaning. The match between media capabilities and communication purposes could cause differences in communication performance. We further investigated the potential moderating effects of communication media (text vs voice) on the correlations between the communication features and economic returns of physicians. Therefore, we hypothesized the following: the associations between the instrumental communication features of physicians and their economic returns are moderated by the type of communication media (hypothesis 3), and the associations between the affective communication features of physicians and their economic returns are moderated by the type of communication media (hypothesis 4).

Data Collection

Objective secondhand data were collected from the Dingxiang Doctor website [4], a predominant web-based health consultation platform in China for >20 years. So far, the website serves >5.5 million active users and 2.1 million participating physicians. The Dingxiang Doctor website categorizes physicians into 39 disease specialties, which deliver different treatment solutions according to patients' disease severity. We selected physicians who treat severe diseases (malignant tumors and heart disease) and who treat less severe diseases (digestive and endocrine diseases) to control interference from the disease type and severity [39]. We used a data crawling application to collect historical physician-patient consultation records. Each consultation instance contained several rounds of physician-patient interactions. Figure 2 presents a sample historical consultation record encompassing physician-related information (title, professional domain, disease department, hospital, and overall patient rating), patient consultation inquiries, physician's replies, and consultation service charges. We collected data from 2 periods: June 2021 and December 2022. We excluded some duplicated or invalid samples containing errors and missing values and ultimately analyzed 40,563 consultation instances generated from 528 physicians between September 2016 and December 2022. Of the 528 physicians, 54 (10.2%) were directors, 154 (29.2%) were associate directors, 231 (43.8%) were attending physicians, and 89 (16.9%) were residents. The studied physicians had worked for an average of 14.34 (SD 7.01) years, and their average consultation fees were RMB 36.53 (US $5.10; SD 21.37).

XSL•FO
RenderX

Figure 2. A sample consultation record page (accessed on August 29, 2022).

Ethical Considerations

Crawling techniques were used to collect data about web-based consultation from a public domain data set visible to all platform users. The studied consultations were not conducted for research purposes. The crawling program followed the Robots Exclusion Protocol. The home pages of the concerned physicians provided only the publicly available information of the physician (such as name, designations, and hospitals) and only the patient consultation information (content, time, and price) that patients had agreed to make public. The patient names were automatically anonymized on the home pages. In the data set handling process, we took precautionary measures to guarantee data security. We also applied to the institutional review board of Shenzhen University for the ethical review of the research project and obtained due approval (202300004) for the study protocol. The institutional review board had waived the requirement to obtain informed consent for this study.

Dependent Variable

The payment received by individual physicians on a per-consultation basis was used to measure their economic returns. We used the log value of the returns in the empirical model. Notably, physicians could vary their consultation charges for different patients.

Independent Variables

The instrumental and affective communication features were estimated from the messages transmitted by physicians for each consultation instance. We used a software named Textmind [31] to extract and quantify the instrumental and affective linguistic features as ratio values between 0 and 1. Textmind is a simplified Chinese-language analysis program developed based on LIWC and Chinese LIWC by the Computational Cyber-Psychology Lab at the Institute of Psychology at the Chinese Academy of Sciences [31]. It delivers a 1-stop solution from word segmentation to language feature analysis and has been effectively used in multiple Chinese-language studies [21,22].

Moderating Variables

We classified the physician messages as textual and vocal communication by introducing an additional feature labeled as message media. Voice messages were assigned a value of 1, and textual messages were assigned a value of 0 [14].

Control Variables

We controlled for the designations and working years of physicians, levels and rankings of their hospitals, development levels of cities in which the hospitals are located, and disease types. Physicians are accorded 4 designations on the Dingxiang

XSL•FO
RenderX

Doctor website: director, associate director, attending physician, and resident physician. Physicians who are assigned high designations are assumed to have more experience in the treatment of particular diseases. Hospital types include private, public, and 3A. The term *3A* denotes tertiary hospitals equipped with more staff members and more patient beds (>501 beds) [40]. These hospitals are usually located in urban cities and provide high-level, specialized medical services. For example, a 3A hospital should set up at least 13 designated clinical departments and have >60 m^2 of space on each floor [40]. Hospital rankings may also influence patient expectations. According to Fudan's 2021 China hospital rankings [41], we classified institutions as ranked in the top 100 hospitals or not ranked in the top 100 hospitals. Disease types included malignant tumors, heart disease, digestive disease, and endocrine disease. The abovementioned control variables, apart from working years, were treated as categorical variables in the model estimating process. They are transformed into dummy variables.

Table 1 presents the descriptive statistics of all the variables, and the correlations among variables are reported in Multimedia Appendix 2. We used Spearman correlation instead of Pearson correlation because of the nonnormality of many LIWC indicators [25]. The correlations among variables (Multimedia Appendix 2) provide the basis for further analysis of the proposed model.

Table 1. Descriptive statistics.

Variable	Values, mean (SD; range)
Economic returns	36.527 (42.154; 1-699)
Insight	0.107 (0.088; 0-1.162)
Causation	0.058 (0.050; 0-0.643)
Discrepancy	0.144 (0.100; 0-1.551)
Tentativeness	0.121 (0.089; 0-1.083)
Certainty	0.033 (0.037; 0-0.754)
Positive emotion	0.078 (0.088; 0-1.048)
Anxiety	0.011 (0.021; 0-0.500)
Anger	0.002 (0.007; 0-0.200)
Sadness	0.004 (0.011; 0-0.387)
Voice service	0.090 (0.285; 0-1)
Working years	14.340 (7.908; 2-46)

Results

Model Estimation

We first conducted multiple linear regression analyses using SPSS (version 22; IBM Corp) to test our hypotheses about the impact of instrumental and affective communications on economic returns of physicians (log value). Table 2 reports the analysis results, which disclose that our independent variables and control variables explained 31.9% of the variance in the economic returns of physicians. The effect size (f^2) is 0.468, and the regression model is significant at the 0.001 level. The variance inflation factor values ranged between 1.052 and 2.566, indicating that our study did not evince the problem of multicollinearity [42].

XSL•FO
RenderX

Table 2. Results of regression analysis[a,b].

	B (SE)	β	*t* test (*df*)[c]	*P* value	VIF[d]
Constant	3.015 (0.006)	N/A[e]	516.813 (40,538)	<.001	N/A
Insight	0.248 (0.025)	.053	10.013 (40,538)	<.001	1.686
Causation	0.155 (0.039)	.019	4.012 (40,538)	<.001	1.326
Discrepancy	−0.294 (0.026)	−.072	−11.169 (40,538)	<.001	2.482
Tentativeness	0.189 (0.030)	.041	6.394 (40,538)	<.001	2.451
Certainty	0.528 (0.051)	.048	10.271 (40,538)	<.001	1.321
Positive emotion	0.283 (0.023)	.061	12.402 (40,538)	<.001	1.444
Anxiety	0.048 (0.083)	.002	0.574 (40,538)	.57	1.093
Anger	0.132 (0.253)	.002	0.520 (40,538)	.60	1.052
Sadness	−0.228 (0.153)	−.006	−1.489 (40,538)	.14	1.063
Working years	0.006 (0.000)	.118	17.930 (40,538)	<.001	2.566
Hospital rank	0.216 (0.004)	.239	52.578 (40,538)	<.001	1.235
Disease type	N/A	N/A	N/A	<.001	1.266
Physician rank	N/A	N/A	N/A	<.001	1.697
Hospital type	N/A	N/A	N/A	<.001	1.053
City tier	N/A	N/A	N/A	<.001	1.253

[a]R^2=0.319.

[b]F_{24}=792.686 (*P*<.001).

[c]2-tailed test.

[d]VIF: variance inflation factor.

[e]N/A: not applicable.

Direct Effect

The results presented in Table 2 indicate that instrumental communication using words indicating insight (B=0.248; *P*<.001), causation (B=0.155; *P*<.001), tentativeness (B=0.189; *P*<.001), and certainty (B=0.528; *P*<.001) was positively associated with the economic returns of physicians. Affective communication using language expressing positive emotion (B=0.283; *P*<.001) was also linked with high economic returns. In contrast, the use of terms conveying discrepancy (B=−0.294; *P*<.001) by physicians was negatively related to their economic returns. Surprisingly, the use of words related to anxiety (B=0.048; *P*=.57), anger (B=0.132; *P*=.60), and sadness (B=−0.228; *P*=.14) did not affect the economic returns of physicians.

We also conducted 2 robustness tests to evaluate the results' stability. First, because the feature values are small, we scaled the independent variables up by 100-fold and performed the multiple regression analysis again, which would not change the model's fitness [43]. Second, we conducted quantile regression analysis to provide a more detailed analysis of the relationships between the communication features and economic returns of physicians. The results are reported in Multimedia Appendices 3 and 4, and they demonstrated the robustness of our multiple linear regression results.

Moderating Effect

On the basis of the regression results, we further tested the moderating effects of communication media (text vs voice) using PROCESS Model 1 in SPSS (version 22) [44]. Table 3 displays the results of the moderating effects of communication media on significant influencing paths presented in Table 2. The findings reveal that communication media function significantly in moderating the influencing paths from the use of insight (B=0.169; *P*=.09), causation (B=0.403; *P*=.005), tentativeness (B=0.268; *P*<.001), and certainty (B=0.839; *P*<.001). However, communication media do not influence the use of words related to discrepancy (B=0.097; *P*=.21) and positive emotion (B=−0.159; *P*=.12). Figure 3 summarizes the results of the testing of our hypotheses.

XSL•FO
RenderX

Table 3. Summary of the moderating effects.

Test and path	B (SE)	*t* test (*df*)[a]	*P* value
1			
Insight → economic returns	0.366 (0.030)	12.283 (40,551)	<.001
Voice service → economic returns	0.044 (0.011)	3.925 (40,551)	<.001
Insight×voice service → economic returns	0.169 (0.100)	1.680 (40,551)	.09
2			
Causation → economic returns	0.310 (0.047)	6.575 (40,551)	<.001
Voice service → economic returns	0.034 (0.011)	3.097 (40,551)	<.001
Causation×voice service → economic returns	0.403 (0.145)	2.782 (40,551)	.005
3			
Discrepancy → economic returns	−0.285 (0.032)	−9.026 (40,551)	<.001
Voice service → economic returns	0.044 (0.013)	3.395 (40,551)	<.001
Discrepancy×voice service → economic returns	0.097 (0.077)	1.262 (40,551)	.21
4			
Tentativeness → economic returns	0.141 (0.036)	3.955 (40,551)	<.001
Voice service → economic returns	0.024 (0.013)	1.858 (40,551)	.06
Tentativeness×voice service → economic returns	0.268 (0.082)	3.285 (40,551)	<.001
5			
Certainty → economic returns	0.586 (0.063)	9.311 (40,551)	<.001
Voice service → economic returns	0.030 (0.010)	3.125 (40,551)	<.001
Certainty×voice service → economic returns	0.839 (0.197)	4.260 (40,551)	<.001
6			
Positive emotion → economic returns	0.242 (0.028)	8.777 (40,551)	<.001
Voice service → economic returns	0.067 (0.009)	7.242 (40,551)	<.001
Positive emotion×voice service → economic returns	−0.159 (0.102)	−1.558 (40,551)	.12

[a]2-tailed test.

XSL•FO
RenderX

Figure 3. The results of main and moderating effects. The numbers indicate B values. NS: not significant; *P=.09; ***P<.001.

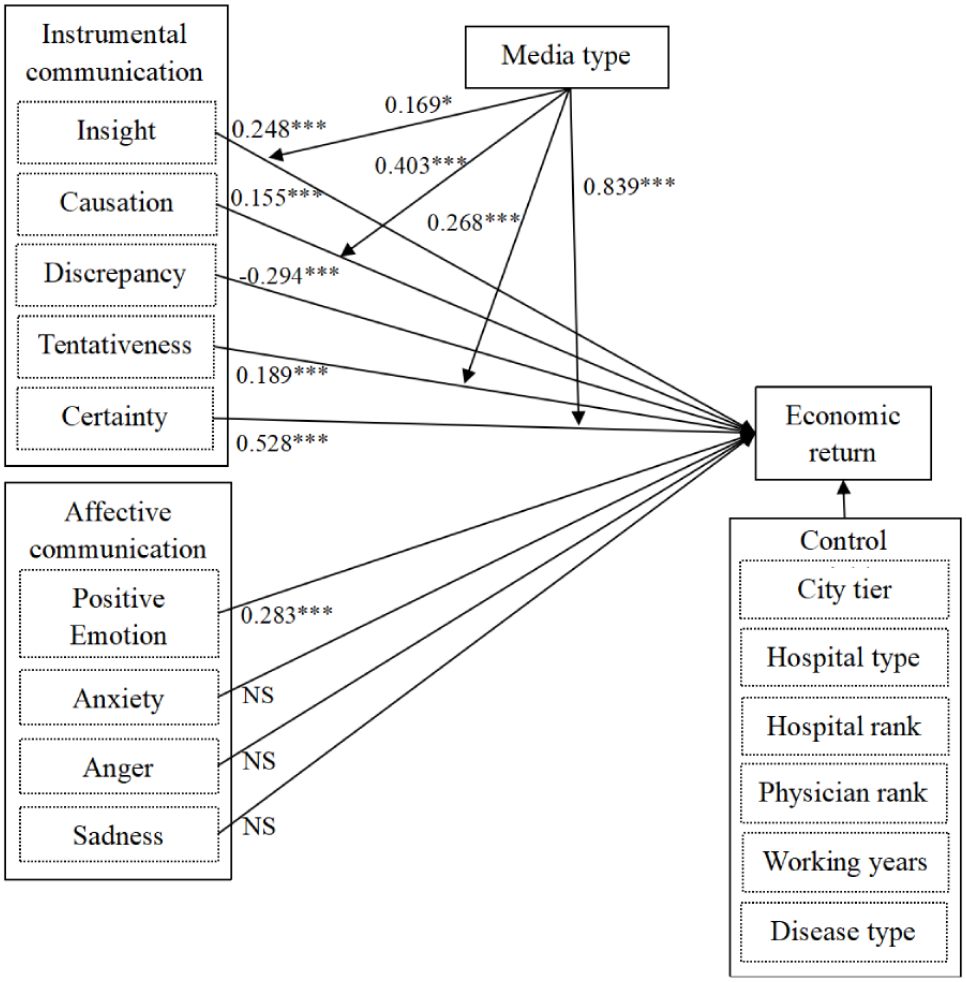

Figure 4 further illustrates the moderating effects. We found that the use of voice messages could magnify the impact of insight-related words (simple slope=0.535; $t_{40,551}$=5.324; P<.001), causation-related terms (simple slope=0.713; $t_{40,551}$=5.055; P<.001), tentativeness-related terms (simple slope=0.409; $t_{40,551}$=4.967; P<.001), and certainty-related terms (simple slope=1.425; $t_{40,551}$=7.434; P<.001) on the economic returns of physicians. However, the use of textual and vocal messages did not vary the effects of terms communicating discrepancy and positive emotion.

XSL•FO
RenderX

Figure 4. Results of the moderating effects.

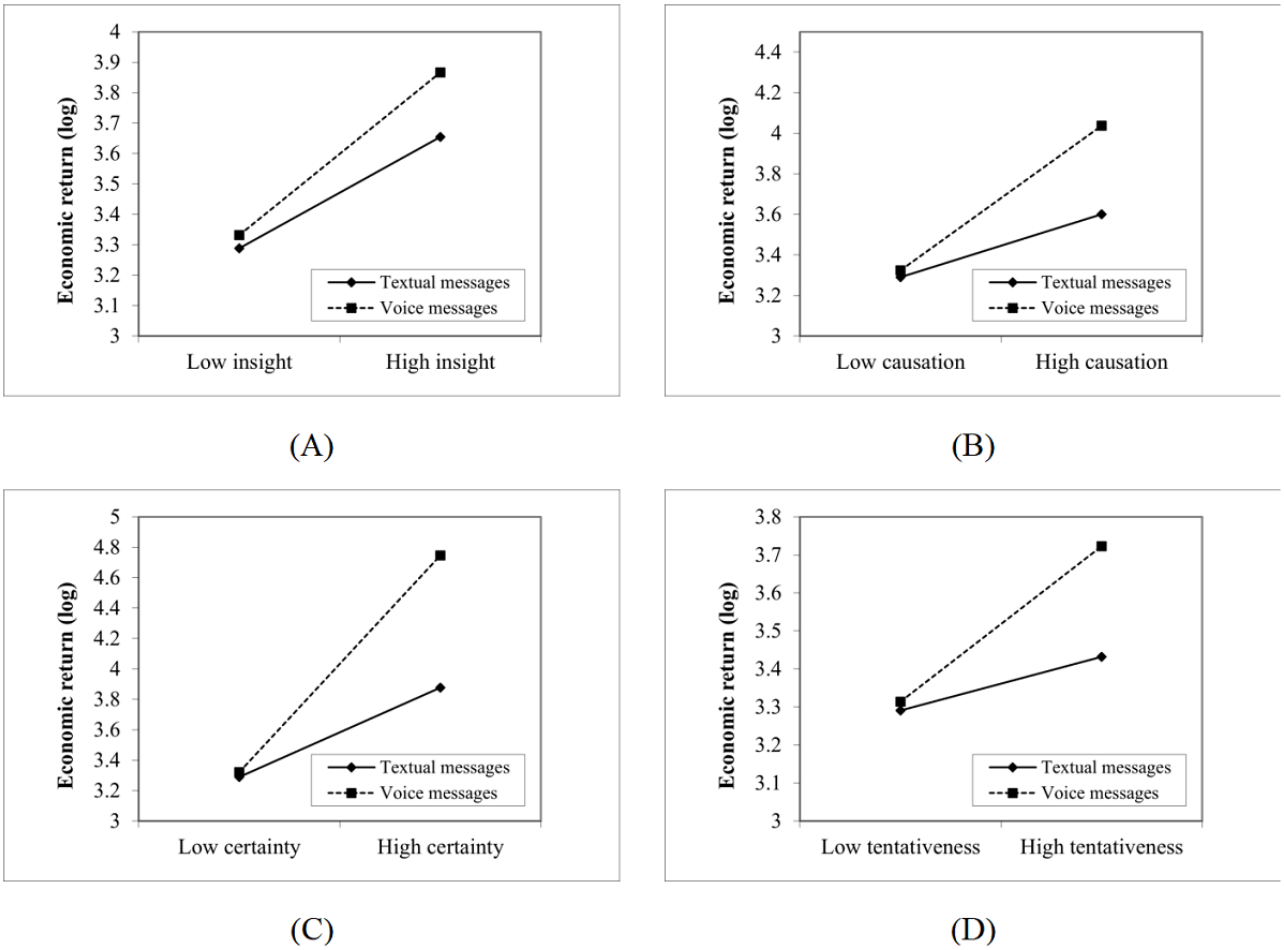

Communication Patterns

The regression results indicated that communication features exerted heterogeneous effects on the economic returns of physicians. In many cases, physicians offered both instrumental and affective responses to patients at different stages of the consultations. Thus, we investigated the compounding effects of discrete communicating features to attain deep insight into communication patterns. First, we transformed the physician responses into a sequence of messages according to the number of replies. The longest sequence comprised 16 messages. Second, we calculated the linguistic feature values for 9 psychological linguistic features for each message in the sequence and constructed a data set composed of 9×16 dimensions for every consultation instance. We entered 0 values for the remaining dimensions of consultations with sequence length <16, because no information was offered after the last message. Thus, the data set encompassed the 9 communication features at different stages of the consultations. Third, we used the K-means clustering method to cluster the consultation instances into multiple groups corresponding to discrete communication patterns. We then selected the optimal number of clusters by using ANOVA to compare the economic returns of physicians across the different groups. The T2 test by Tamhane [45] was selected for the analysis of the variance because the sample variance did not satisfy homogeneity. The results disclosed 4 clusters as the optimal number. Therefore, we obtained 4 groups of consultation instances associated with different levels of economic returns, as depicted in Multimedia Appendix 5.

Table 4 presents the detailed results of the T2 test by Tamhane [45]. The results elucidated that the third group commanded the highest economic returns, whereas the second group received the lowest gain. The mean revenue from consultations in the third group was 12.2% higher than the average earnings of the first group, 21.2% higher than the standard remunerations of the second group, and 14.6% higher than the average payments received by the fourth group.

XSL•FO
RenderX

Table 4. ANOVA results for economic returns (log)[a,b].

Group	1		2		3		4	
	D (SE)	*P* value	D (SE)	*P* value	D (SE)	*P* value	D (SE)	*P* value
1	N/A[c]	N/A[b]	0.090 (0.026)	.004	–0.122 (0.010)	<.001	0.024 (0.009)	.03
2	–0.090 (0.026)	.004	N/A[b]	N/A[b]	–0.212 (0.026)	<.001	–0.066 (0.026)	.06
3	0.122 (0.010)	<.001	0.212 (0.026)	<.001	N/A[b]	N/A[b]	0.146 (0.009)	<.001
4	–0.024 (0.009)	.03	0.066 (0.026)	.06	–0.146 (0.009)	<.001	N/A[b]	N/A[b]

[a]The values in the table represent the difference (%) between the vertical and horizontal category labels.

[b]F_3=87.971 (P<.001).

[c]N/A: not applicable.

To further elucidate the communication patterns of each group, we divided the 9 linguistic features into 3 dimensions according to the regression results: instrumental interactions, affective communication using positive emotions, and affective communication using negative emotions. As Multimedia Appendix 5 displays, we then visualized the communication features for each group using the mean values. The sequence length was set to 16 according to the longest sequence of physician messages. Multimedia Appendix 6 describes the visualization results for all 9 linguistic features.

Figure 5 shows that consultations in the third group tended to include longer sequences than those categorized into other groups, signifying that physicians designated to the third group replied more frequently. These physicians also used more terms related to positive emotions at the beginning of their consultations and later offered more instrumental suggestions. In contrast, physicians in the second group used more terms related to positive and negative emotions, offered instrumental suggestions in their second reply, and then tended to end the consultation with a short response. Surprisingly, physicians in the fourth group used few emotion-related terms and provided moderate amounts of instrumental suggestions at the beginning of their consultations, and their economic returns were slightly higher than that of those in the second group.

XSL•FO
RenderX

Figure 5. Distributions of communication features across the 4 groups: (A) instrumental communication, (B) affective communication (positive emotion), and (C) affective communication (negative emotion).

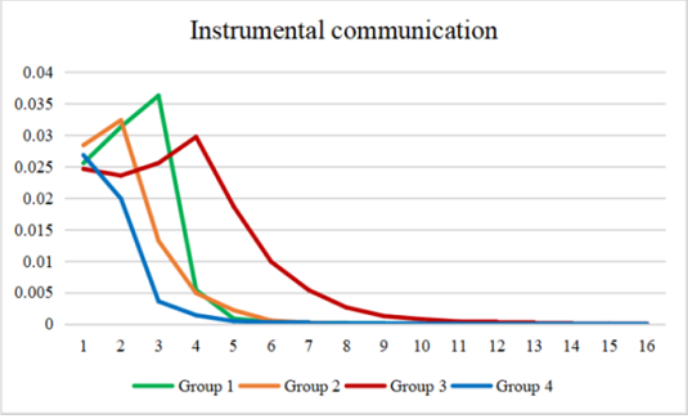

(A)

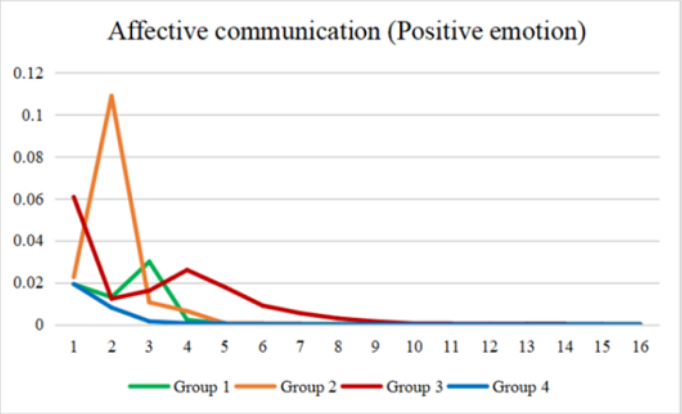

(B)

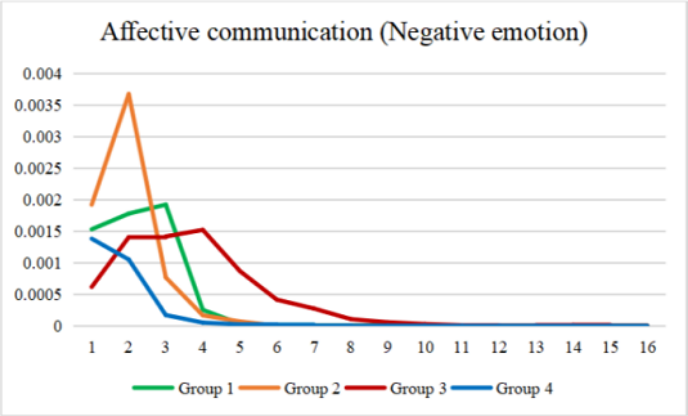

(C)

Discussion

Principal Findings

This study aimed to investigate, at a more granular level, the associations between communication behaviors (instrumental and affective) and economic returns of physicians. We also explored the moderating effects of communication media (text vs voice). Our results indicate that the use of words indicating insight, causation, tentativeness, and certainty and the use of words indicating positive emotion were positively associated with the economic returns of physicians. In contrast, the use of terms conveying discrepancy by physicians was negatively related to their economic returns. The use of voice media by physicians intensified the impact of terms related to insight, causation, tentativeness, and certainty. The pattern analysis results indicate that physicians who responded to patients more

XSL•FO
RenderX

frequently, communicated positive emotions at the beginning of the consultations, and provided more instrumental suggestions afterward achieved the highest economic returns.

Our findings align with the results reported in previous studies [2,6,7,14,46] and offer additional feature-level details. Moreover, the provision of logical and assured explanations helps reduce uncertainty in patients, who gain knowledge about the causes and consequences of their diseases. Specificity and clarity in expression and information delivery also improve communication quality [47]. Therefore, it is reasonable that physicians with superior instrumental communication skills achieve high economic returns. Moreover, patients may harbor great expectations from physicians with high consultation fees [6], which could drive physicians to devote more effort to cognitive thinking before providing answers. However, the increased use of discrepancy-related words (ie, should and would) could suggest that physicians notice divergences between their suggestions and patient behaviors and highlight them during conversations. Thus, they may unintentionally deliver dissatisfaction signals to patients. Hence, instrumental communication using discrepancy-related words was negatively associated with the economic returns of physicians.

Our results also indicate that affective communication encompassing more terms related to positive emotions (ie, happy, love, and nice) was positively linked with high economic returns of physicians. This result is congruent with previous findings that the enhancement of service quality mandates the delivery of emotional support for patients [8,11,33], which is also a vital patient-centered communication skill [48]. The patients' psychological needs for empathy and care are satisfied when physicians demonstrate compassion and encouragement [7]. Consequently, physicians with better affective communication skills tend to receive high economic returns. However, the articulation of negative emotions such as anger, anxiety, and sadness was not related to the economic returns attained by physicians. Perhaps, the use of words conveying anger, anxiety, and sadness by physicians could deliver negative signals to patients and undermine their confidence in their physicians, even though such terms could also transmit empathy. Physicians are suggested to use language related to understanding, respecting, and supporting to express empathy to patients [47].

The moderating effects analysis revealed that the choice of text or voice media for communication can moderate the influence exerted by certain linguistic features. The media synchronicity theory posits that the use of media supporting high synchronicity and multiplicity of cues is more suited to complex communication [37]. Explaining the causes of diseases or syndromes during web-based consultations may be deemed as information transmission processes with high requirements for media synchronicity. Therefore, such messages become more significantly influential through voice media, which is characterized by high synchronicity [38]. These findings suggest that physicians who make smart use of voice media to communicate with patients tend to obtain high economic returns.

Our pattern analysis results unveiled the compounding effects of multiple communication features. Providing positive emotional support to patients at the beginning of the consultation can fulfill the psychological needs of patients before satisfying their knowledge requirements [47]. This process enhances the outcomes of instrumental communication because patients may find it difficult to comprehend disease knowledge and treatment suggestions if they are emotionally occupied. Therefore, consultations that first delivered positive emotional support and later supplied the instrumental comments were found to yield high levels of economic returns. This partially aligns with the previous finding that interaction frequency positively affects patient satisfaction [19]. In contrast, consultations that simultaneously offered emotional support and instrumental comments in a limited number of replies yielded low levels of economic returns. These findings imply that physicians who address the emotional needs of patients before offering professional advice are more likely to obtain high economic returns.

Limitations and Future Directions

Despite the contributions of our study, we must indicate a few limitations. First, our data were collected from the Dingxiang Doctor website in China. The generalization of our study's findings in other countries would require us to obtain data from multiple websites in many other countries. Second, although our data contain a relatively large set of consultation cases, we should collect data sets encompassing long time durations and generate a panel data set that would guarantee the robustness of our findings. Third, we did not include patients' personal preferences in our conceptual model such as patients' education level because of the limited access to patients' personal information. We believe future study that investigates the impact of patients' preferences will contribute novel insight into this research issue. Finally, we can expand our study by relating the communication features of physicians to their medical domain knowledge to provide deep insight into the service quality of web-based health care consultations.

Conclusions

This study demonstrates that the economic returns of physicians are associated with their communication features and the media used for web-based health care consultations. This study adopted a psychological and linguistic perspective to offer methodological referential value for relevant prospective studies of web-based physician-patient interactions. Moreover, it supplements the limited literature relating to the economic returns received by physicians through web-based health care platforms. The findings deliver important practical directions for improving the quality of web-based consultation services provided by physicians.

XSL•FO
RenderX

Acknowledgments

This study was supported by the National Natural Science Foundation of China (71901150 and 72334004), Guangdong Basic and Applied Basic Research Foundation (2022A1515012077 and 2023A1515012515), Guangdong Philosophy and Social Science Planning Project (GD23XGL113), and Guangdong Province Innovation Team (2021WCXTD002).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Linguistic features related to cognitive and affective processes.
[DOCX File , 12 KB - jmir_v26i1e42850_app1.docx]

Multimedia Appendix 2
Pair-wise correlations (Spearman correlation).
[DOCX File , 17 KB - jmir_v26i1e42850_app2.docx]

Multimedia Appendix 3
Regression results (independent variables×100).
[DOCX File , 17 KB - jmir_v26i1e42850_app3.docx]

Multimedia Appendix 4
Results of quantile regression analysis.
[DOCX File , 16 KB - jmir_v26i1e42850_app4.docx]

Multimedia Appendix 5
Levels of economic returns across the 4 groups.
[PNG File , 54 KB - jmir_v26i1e42850_app5.png]

Multimedia Appendix 6
Distribution of communication features.
[DOCX File , 61 KB - jmir_v26i1e42850_app6.docx]

References

1. Lu X, Zhang R. Association between eHealth literacy in online health communities and patient adherence: cross-sectional questionnaire study. J Med Internet Res 2021 Sep 13;23(9):e14908 [FREE Full text] [doi: 10.2196/14908] [Medline: 34515638]
2. Guo S, Guo X, Fang Y, Vogel D. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. J Manag Inf Syst 2017;34(2):487-519. [doi: 10.1080/07421222.2017.1334480]
3. The cyberspace administration of China released the "digital China development report (2022)". Cyberspace Administration of China. 2023 May 23. URL: http://www.cac.gov.cn/2023-05/22/c_1686402318492248.htm [accessed 2023-06-18]
4. Dingxiang Doctor. URL: https://dxy.com/ [accessed 2023-06-18]
5. Haodf. URL: https://www.haodf.com/ [accessed 2023-06-18]
6. Chiu YL, Wang JN, Yu H, Hsu YT. Consultation pricing of the online health care service in China: hierarchical linear regression approach. J Med Internet Res 2021 Jul 14;23(7):e29170. [doi: 10.2196/29170] [Medline: 34259643]
7. Roongruangsee R, Patterson P, Ngo LV. Professionals' interpersonal communications style: does it matter in building client psychological comfort? J Serv Mark 2022;36(3):379-397. [doi: 10.1108/JSM-09-2020-0382]
8. Atanasova S, Kamin T, Petrič G. The benefits and challenges of online professional-patient interaction: comparing views between users and health professional moderators in an online health community. Comput Hum Behav 2018 Jun;83:106-118. [doi: 10.1016/j.chb.2018.01.031]
9. Cao B, Huang W, Chao N, Yang G, Luo N. Patient activeness during online medical consultation in China: multilevel analysis. J Med Internet Res 2022 May 27;24(5):e35557 [FREE Full text] [doi: 10.2196/35557] [Medline: 35622403]
10. Wu D, Lowry PB, Zhang D, Tao Y. Patient trust in physicians matters-understanding the role of a mobile patient education system and patient-physician communication in improving patient adherence behavior: field study. J Med Internet Res 2022 Dec 20;24(12):e42941 [FREE Full text] [doi: 10.2196/42941] [Medline: 36538351]
11. Ben-Sira Z. Affective and instrumental components in the physician-patient relationship: an additional dimension of interaction theory. J Health Soc Behav 1980 Jun;21(2):170-180. [Medline: 7391531]

XSL•FO
RenderX

12. Ben-Sira Z. The function of the professional's affective behavior in client satisfaction: a revised approach to social interaction theory. J Health Soc Behav 1976 Mar;17(1):3-11. [Medline: 1270783]
13. Webster C, Sundaram DS. Effect of service provider's communication style on customer satisfaction in professional services setting: the moderating role of criticality and service nature. J Serv Mark 2009;23(2):103-113. [doi: 10.1108/08876040910946369]
14. Tan H, Yan M. Physician-user interaction and users' perceived service quality: evidence from Chinese mobile healthcare consultation. Inf Technol People 2020;33(5):1403-1426. [doi: 10.1108/itp-01-2019-0039]
15. Vermeir P, Vandijck D, Degroote S, Peleman R, Verhaeghe R, Mortier E, et al. Communication in healthcare: a narrative review of the literature and practical recommendations. Int J Clin Pract 2015 Nov;69(11):1257-1267 [FREE Full text] [doi: 10.1111/ijcp.12686] [Medline: 26147310]
16. Lee D. HEALTHQUAL: a multi-item scale for assessing healthcare service quality. Serv Bus 2016 Jun 13;11:491-516. [doi: 10.1007/s11628-016-0317-2]
17. Izumi N, Matsuo T, Matsukawa Y. Associations among physician-patient communication, patient satisfaction, and clinical effectiveness of overactive bladder medication: a survey of patients with overactive bladder. J Clin Med 2022 Jul 14;11(14):4087 [FREE Full text] [doi: 10.3390/jcm11144087] [Medline: 35887847]
18. Chen L, Tang H, Guo Y. Effect of patient-centered communication on physician-patient conflicts from the physicians' perspective: a moderated mediation model. J Health Commun 2022 Mar 04;27(3):164-172. [doi: 10.1080/10810730.2022.2071505] [Medline: 35579286]
19. Yang H, Guo X, Wu T. Exploring the influence of the online physician service delivery process on patient satisfaction. Decis Support Syst 2015 Oct;78:113-121. [doi: 10.1016/j.dss.2015.05.006]
20. Yang Y, Zhang X, Lee PK. Improving the effectiveness of online healthcare platforms: an empirical study with multi-period patient-doctor consultation data. Int J Prod Econ 2019 Jan;207:70-80. [doi: 10.1016/j.ijpe.2018.11.009]
21. Yuan C, Hong Y, Wu J. Personality expression and recognition in Chinese language usage. User Model User Adap Inter 2020 Aug 25;31:121-147. [doi: 10.1007/s11257-020-09276-2]
22. Guan L, Hao B, Cheng Q, Yip PS, Zhu T. Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model. JMIR Ment Health 2015 May 12;2(2):e17 [FREE Full text] [doi: 10.2196/mental.4227] [Medline: 26543921]
23. Andy A, Andy U. Understanding communication in an online cancer forum: content analysis study. JMIR Cancer 2021 Sep 07;7(3):e29555 [FREE Full text] [doi: 10.2196/29555] [Medline: 34491209]
24. Donnellan WJ, Warren JG. Emotional word use in informal carers of people living with dementia: linguistic analysis of online discussion forums. JMIR Aging 2022 Jun 17;5(2):e32603 [FREE Full text] [doi: 10.2196/32603] [Medline: 35713942]
25. Marengo D, Azucar D, Longobardi C, Settanni M. Mining Facebook data for quality of life assessment. Behav Inf Technol 2021;40(6):597-607. [doi: 10.1080/0144929x.2019.1711454]
26. Jha D, Singh R. Analysis of associations between emotions and activities of drug users and their addiction recovery tendencies from social media posts using structural equation modeling. BMC Bioinformatics 2020 Dec 30;21(Suppl 18):554 [FREE Full text] [doi: 10.1186/s12859-020-03893-9] [Medline: 33375934]
27. Liu Y, Ren C, Shi D, Li K, Zhang X. Evaluating the social value of online health information for third-party patients: is uncertainty always bad? Inf Process Manag 2020 Sep;57(5):102259. [doi: 10.1016/j.ipm.2020.102259]
28. Seraj S, Blackburn KG, Pennebaker JW. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. Proc Natl Acad Sci U S A 2021 Mar 16;118(7):e2017154118 [FREE Full text] [doi: 10.1073/pnas.2017154118] [Medline: 33526594]
29. Meggiolaro E, Berardi MA, Andritsch E, Nanni MG, Sirgo A, Samorì E, et al. Cancer patients' emotional distress, coping styles and perception of doctor-patient interaction in European cancer settings. Palliat Support Care 2016 Jun;14(3):204-211. [doi: 10.1017/S1478951515000760] [Medline: 26155817]
30. Geng S, Niu B, Feng Y, Huang M. Understanding the focal points and sentiment of learners in MOOC reviews: a machine learning and SC - LIWC - based approach. Br J Educ Technol 2020 Sep;51(5):1785-1803. [doi: 10.1111/bjet.12999]
31. Gao R, Hao B, Li H, gao Y, Zhu T. Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In: Proceedings of the International Conference, BHI 2013. 2013 Presented at: International Conference, BHI 2013; October 29-31, 2013; Maebashi, Japan. [doi: 10.1007/978-3-319-02753-1_36]
32. Huang CL, Chung CK, Hui N, Lin YC, Seih YT, Lam BC, et al. The development of the Chinese linguistic inquiry and word count dictionary. Chin J Psychol 2012;54(2):185-201.
33. Chen S, Guo X, Wu T, Ju X. Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. Inf Process Manag 2020 Sep;57(5):102253. [doi: 10.1016/j.ipm.2020.102253]
34. Kelly KM, Ellington L, Schoenberg N, Agarwal P, Jackson T, Dickinson S, et al. Linking genetic counseling content to short-term outcomes in individuals at elevated breast cancer risk. J Genet Couns 2014 Oct;23(5):838-848 [FREE Full text] [doi: 10.1007/s10897-014-9705-8] [Medline: 24671341]
35. Soós MJ, Coulson NS, Davies EB. Exploring social support in an online support community for tourette syndrome and tic disorders: analysis of postings. J Med Internet Res 2022 Oct 04;24(10):e34403 [FREE Full text] [doi: 10.2196/34403] [Medline: 36194454]

XSL•FO
RenderX

36. Keating DM. Spirituality and support: a descriptive analysis of online social support for depression. J Relig Health 2013 Sep;52(3):1014-1028. [doi: 10.1007/s10943-012-9577-x] [Medline: 22322336]
37. Dennis AR, Fuller RM, Valacich JS. Media, tasks, and communication processes: a theory of media synchronicity. MIS Q 2008 Sep;32(3):575-600. [doi: 10.2307/25148857]
38. Liu S, Zhang M, Gao B, Jiang G. Physician voice characteristics and patient satisfaction in online health consultation. Inf Manag 2020 Jul;57(5):103233. [doi: 10.1016/j.im.2019.103233]
39. China statistical yearbook 2021. Institute for Health Metrics and Evaluation. URL: https://ghdx.healthdata.org/record/china-statistical-yearbook-2021 [accessed 2022-12-25]
40. Yifa GW. Notice of the national health commission on issuing the accreditation standards for tertiary hospitals (2020 edition). China Government Website. 2022 Dec 21. URL: http://www.gov.cn/zhengce/zhengceku/2020-12/28/content_5574274.htm [accessed 2023-02-01]
41. China's hospital and specialty reputation rankings. CN-Healthcare. URL: http://rank.cn-healthcare.com/fudan/national-general [accessed 2023-02-01]
42. Petter S, Straub D, Rai A. Specifying formative constructs in information systems research. MIS Q 2007 Dec;31(4):623-656. [doi: 10.2307/25148814]
43. Wooldridge JM. Introductory Econometrics: A Modern Approach. Mason, OH: South-Western Cengage Learning; 2012.
44. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: The Guilford Press; 2013.
45. Tamhane AC. Multiple comparisons in model I one-way ANOVA with unequal variances. Commun Stat Theory Methods 1977;6(1):15-32. [doi: 10.1080/03610927708827466]
46. Li J, Zheng H, Duan X. Factors influencing the popularity of a health-related answer on a Chinese question-and-answer website: case study. J Med Internet Res 2021 Sep 28;23(9):e29885 [FREE Full text] [doi: 10.2196/29885] [Medline: 34581675]
47. Hashim MJ. Patient-centered communication: basic skills. Am Fam Physician 2017 Jan 01;95(1):29-34 [FREE Full text] [Medline: 28075109]
48. Jiang S. How does patient-centered communication improve emotional health? An exploratory study in China. Asian J Commun 2018;28(3):298-314. [doi: 10.1080/01292986.2017.1413402]

Abbreviations

LIWC: Linguistic Inquiry and Word Count

Edited by G Eysenbach; submitted 21.09.22; peer-reviewed by W Zhang, E Nichele, Y Yang; comments to author 23.11.22; revised version received 24.02.23; accepted 17.11.23; published 11.01.24.

Please cite as:
Geng S, He Y, Duan L, Yang C, Wu X, Liang G, Niu B
The Association Between Linguistic Characteristics of Physicians' Communication and Their Economic Returns: Mixed Method Study
J Med Internet Res 2024;26:e42850
URL: https://www.jmir.org/2024/1/e42850
doi:10.2196/42850
PMID:38206657

©Shuang Geng, Yuqin He, Liezhen Duan, Chen Yang, Xusheng Wu, Gemin Liang, Ben Niu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Individual Characteristics in the Comprehension of Pandemic Video Communication: Randomized Controlled Between-Subjects Design

Daniel Adrian Lungu[1], PhD; Jo Røislien[1], PhD; Ionica Smeets[2], PhD; Siri Wiig[1], PhD; Kolbjørn Kallesten Brønnick[1], PhD

[1]SHARE – Centre for Resilience in Healthcare, Department of Quality and Health Technology, Faculty of Health Sciences, University of Stavanger, Stavanger, Norway

[2]Science Communication and Society, Institute of Biology, Leiden University, Leiden, Netherlands

Corresponding Author:
Daniel Adrian Lungu, PhD
SHARE – Centre for Resilience in Healthcare
Department of Quality and Health Technology, Faculty of Health Sciences
University of Stavanger
PO Box 8600
Stavanger, 4021
Norway
Phone: 47 91532327
Email: daniel.a.lungu@uis.no

Abstract

Background: Video played an important role in health communication throughout the COVID-19 pandemic. It was used to communicate pandemic information to the public, with a variety of formats, presenters, and topics. Evidence regarding the effectiveness of video features is available, while how individual characteristics of recipients influence communication comprehension is still limited.

Objective: This study aimed to test 6 individual characteristics and assess their effect on the comprehension of pandemic video communication.

Methods: Short health communication videos were presented to a large sample of subjects, receiving questionnaire responses from 1194 participants. Individual characteristics consisted of age, sex, living area, education level, income level, and belief in science. Communication comprehension consisted of both perceived and objective comprehension. The data were analyzed by multiple linear regression.

Results: Age had a negative effect on both perceived and objective comprehension—age was negatively associated with comprehension. There were sex differences, with higher perceived comprehension and lower objective comprehension among female than male individuals. Living in an urban or a rural area had no significant effect (all $P>.05$). The level of education and income had a positive effect on both subjective and objective comprehension. Finally, the belief in science had a positive effect on perceived comprehension ($P<.001$) but did not have a statistically significant effect on objective comprehension ($P=.87$).

Conclusions: The main differences between those who think they understand pandemic communication and those who comprehend it better are sex (female individuals have a higher perception of having comprehended, while male individuals have higher levels of objective comprehension) and belief in science (higher belief in science leads to higher perceived comprehension, while it does not have any impact on objectively understanding the message conveyed).

(J Med Internet Res 2024;26:e48882) doi:10.2196/48882

KEYWORDS

video communication; COVID-19 pandemic; comprehension; health literacy; video; videos; health communication; psychology; perception; comprehend; understanding; coronavirus disease 2019

XSL•FO
RenderX

Introduction

Pandemic Video Communication

A pandemic is a comprehensive public health concern, and effective communication is key to handling the challenges. Unsurprisingly, the COVID-19 pandemic spawned massive efforts from authorities worldwide to communicate a wide range of important information to the public. Effective communication is essential for reducing anxiety and fear in the population [1,2]; building trust in authorities [3,4]; disseminating information and addressing misinformation [5-7]; and encouraging behavioral changes such as social distancing [8,9], hygiene practices, and vaccination [8-10].

Although COVID-19 communication occurred through all media channels, video was among the most used [11-13]. Some videos were developed by governmental agencies, whereas others were developed by public and private media, enterprises, and individual users. While some videos were distributed through traditional media, a massive amount was distributed on social media (eg, TikTok, YouTube, and Facebook) and instant messaging platforms [14-17]. A previous study identified major differences between COVID-19 videos created by public health authorities and professional creators, identifying a creative gap and a limited reach of the videos created by public health authorities [18].

Besides the reach, comprehension is one of the main aims of pandemic video communication, to ensure that the message is conveyed to the population. In a large randomized controlled trial, Vandormael et al [19] reported an increase in knowledge of a COVID-19 animated video distributed through social media. A study from Tan et al [20] found that the quality of COVID-19 vaccine videos distributed on video-sharing platforms was generally low and led to limited comprehension, with slight variations across the platforms.

Recipients' Characteristics

On the other hand of the communication process, there is the public, also called recipients of the message. A recent scoping review identified available evidence regarding the effect of recipients' individual factors on health communication outcomes [21], highlighting the importance of the following 6 factors: age, sex, living area, education, income, and belief in science. Evidence is mixed for the sociodemographic variables, ranging from positive to negative to no effects of age [14,17,22-26], sex [14,19,23,26-29], living area [30-33], education [22,23,26,34-39], and income [34,36,40] on health communication outcomes. Instead, belief in science was found to have a positive effect on health communication outcomes. According to Farias et al [41], belief in science can help individuals to cope with stress and anxiety, similar to religious belief. Research has found that belief in science influences physical distancing behavior: there was less compliance to physical distancing in countries with less belief in science [42] or mask wearing, and higher belief in science predicted higher mask wearing in public spaces [43] during the COVID-19 period.

Studies have identified some features of effective pandemic video communication [7,18,44], but the gap persists regarding the role that individual characteristics of recipients play in the comprehension of pandemic video communication. This study aims to fill this gap by assessing the effect of 6 individual factors (age, sex, living area, education level, income, and belief in science) on the comprehension of pandemic video communication.

The relevance of this study derives from the fact that while video communication factors (eg length, visuals, tone, etc) can be manipulated to increase effectiveness, the individual characteristics of pandemic video communication recipients cannot, and therefore public health authorities should be aware of potential differences when developing pandemic communication videos.

Methods

Design and Video Creation

This study analyzed individual characteristics from a trial using a full-factorial, between-subjects randomized controlled experiment [45]. The design of the videos was done by manipulating 3 variables: the source, the topic, and the call to action.

The source had two levels: (1) in one case, the presenter was introduced as an infectious diseases specialist, and (2) in the other, he was presented as a salesman. We chose 3 topics to communicate: exponential growth, handwashing, and the burden of pandemics on health care systems. The call to action had two levels: one with a motivational call to action at the end of the video, and one without. The design led to 12 video versions to be created. A professional scriptwriter was hired to create 6 scripts (3 topics with 2 tones), while the third variable was manipulated through the on-screen text introducing the speaker. After 2 rounds of revision, the scripts were approved by the researchers and the writer.

Based on the scripts, 12 videos were produced in collaboration with the Department for Development of Digital Learning Resources of the University of Stavanger. To avoid confounding factors, all other variables that the experimental ones were controlled for. We used the same professional actor to shoot all the videos, he wore the same clothes (white shirt and grey suit), and we had a neutral gray background for all videos (Figure 1). All videos were approximately the same length (range 54 s to 1 min 25 s). Videos are stored in the OpenScience repository of the University of Stavanger [46].

XSL•FO
RenderX

Figure 1. The thumbnail of the videos.

Participants

The Norwegian Air Ambulance Foundation (NAAF), one of the Covcom project [45] partners, helped recruit participants among their member database. The member base of NAAF is made of donating members (who support the foundation through regular financial donations) and prospect members (who are interested and support in nonfinancial ways, eg, following and distributing content on their social media accounts). At the time of the data collection, there were approximately 230,000 donating members and 110,000 prospect members. As the former has been previously involved in research activities, we decided to recruit the study participants among the donating members. An a priori power analysis assuming multiple linear regression with 6 predictor variables, a small effect size of F=0.02, and 95% statistical power revealed that the minimum sample size was 1050 participants. From previous experiences, the response rates of NAAF members in previous research studies were approximately 10%. Therefore, we adopted a conservative approach and invited 12,000 people to participate, so that a response rate of 8.75% (1050/12,000) would be sufficient to reach the minimum sample size. Participants were randomly assigned to 1 of the 12 video versions and were invited by email. The email contained a brief explanation of the study and a link to participate.

Among the invited, 54% (6481/12,000) were male; ages ranged from 18 to 90 years, with a median value of 64.8 (IQR 58-73) years.

Data Collection and Management

NAAF's IT infrastructure was used to collect the data. This choice came with 2 advantages: first, participants would receive the email from NAAF, an organization they are familiar with, therefore lowering the risk of emails being perceived as spam. Second, as per General Data Protection Regulation (GDPR) requirements, no personal information had to be shared. The data were collected from June 1 to June 9, 2021.

Participants were sent an email containing 1 of the 12 videos, and a link to the survey created on the SurveyMonkey platform. Participants were randomly assigned to 1 of the 12 video versions. The data collection was anonymous: the IP addresses of respondents were not collected, and the email addresses were used only for inviting participants, without being stored or linked to responses. Respondents were not asked for any personal or sensitive data.

A unique identifier was generated for each response, and responses were stored in a database. This database is not linked to the NAAF's members database.

Ethical Considerations

Since the project did not intend to collect or process any personal or sensitive data, and because the data collection was anonymous (the IP address was not collected), the study was exempt from obtaining approval from the Norwegian Centre for Research Data or the Ethical Committee. In addition, the legal department of NAAF ensured that the study was compliant with the GDPR and national data processing regulations.

Participants gave their informed consent digitally at the beginning of the survey. Participation was voluntary and no incentives were offered to participants to complete the survey. Consent to use the image was granted from the actor depicted in Figure 1.

Measures and Data Analysis

The questionnaire collected 6 individual characteristics: age, sex, education, income level, place of living, and belief in science. The Belief in Science scale [41] was used to define a

XSL•FO
RenderX

baseline of the level of scientific belief of participants. The Belief in Science scale is a measurement tool of attitudes toward science, where science shares similarities with religion in terms of the comforting role it plays in individuals' lives.

Comprehension is one of the main aims of health communication and was assessed as both subjective (or perceived) and objective comprehension. Here, subjective comprehension indicates the extent to which someone believes to have understood that information, while objective comprehension refers to the ability to understand the information and to incorporate it into one's knowledge [47]. Subjective comprehension was measured by a 4-item, 5-point Likert scale, with internal consistency of the scale being assessed by the Cronbach α. Objective comprehension was measured by 1 single-choice question. There were 3 different questions that matched the message enclosed in the video. Participants were presented with 4 answer alternatives, of which 1 was correct. The English translation of the questionnaire is available in the Multimedia Appendix 1.

Multiple linear regression analysis was performed in R Studio (R Core Team) [48], with comprehension as the dependent variable and age, sex, education, income, place of living, and belief in science as independent variables. The model was also tested against generalized additive model analysis to account for nonlinear data such as the ceiling effect.

Results

We collected 1194 complete replies from 12,000 invites, corresponding to a response rate of 9.95%. After removing 2 responses that failed the attention checks, a total of 1192 valid responses were included for analysis. This is greater than the threshold of 1050 subjects resulting from the power analysis.

The mean age of participants was 64.77 (SD 11.31) years, female individuals represented 46.14% (550/1192) of the total, and 53.1% (644/1192) lived in an urban context. Those with higher education (at least a college or university degree) comprised 53.69% (640/11921) and those earning more than NOK 500,000 per year comprised 41.86% (499/1192). The mean belief in science score, measured on a 1 to 6 scale, was 4.42 (SD 0.84). All values, including the details by video version, are available in Table 1 below.

Table 1. Characteristics of study participants (n=1192).

Group or version	Participants, n	Age (years), mean (SD)	Female individuals, n (%)	Living in city, n (%)	High education, n (%)	Income >NOR 500,000[a], n (%)	Belief in science, mean (SD)
1	98	63.48 (11.64)	44 (44.9)	52 (53.06)	59 (60.2)	45 (45.92)	4.38 (0.88)
2	108	65.53 (12.04)	50 (46.3)	62 (57.41)	59 (54.63)	43 (39.81)	4.18 (0.85)
3	108	64.60 (10.45)	48 (44.44)	58 (53.7)	51 (47.22)	50 (46.3)	4.44 (0.77)
4	98	63.28 (11.61)	47 (47.96)	50 (51.02)	54 (55.1)	44 (44.9)	4.42 (0.80)
5	93	68.00 (10.77)	44 (47.31)	48 (51.61)	53 (56.99)	36 (38.71)	4.25 (0.88)
6	94	66.48 (11.08)	34 (36.17)	51 (54.26)	50 (53.19)	41 (43.62)	4.48 (0.66)
7	107	62.70 (11.84)	53 (49.53)	58 (54.21)	71 (66.36)	55 (51.4)	4.50 (0.77)
8	103	65.78 (10.2)	48 (46.6)	63 (61.17)	49 (47.57)	33 (32.04)	4.56 (0.84)
9	85	65.59 (10.05)	47 (55.29)	43 (50.59)	38 (44.71)	32 (37.65)	4.39 (0.87)
10	93	64.29 (14.79)	42 (45.16)	44 (47.31)	52 (55.91)	35 (37.63)	4.53 (0.92)
11	96	63.96 (10.35)	41 (42.71)	47 (48.96)	50 (52.08)	42 (43.75)	4.48 (0.82)
12	109	64.06 (9.85)	52 (47.71)	57 (52.29)	54 (49.54)	43 (39.45)	4.45 (0.90)
Total	1192	64.77 (11.31)	550 (46.14)	633 (53.1)	640 (53.69)	499 (41.86)	4.42 (0.84)

[a]Equivalent to USD $45,000, using a conversion rate of 1 NOK=0.09 USD.

We found that participants reported high levels of comprehension, as the mean for subjective comprehension was 5.48 (SD 0.63) on a 1 to 6 scale. Moreover, there was a limited degree of variation between video versions, with *z* scores ranging from –0.57 to 0.33. Regarding objective comprehension, we observed a mean value of 0.78 (SD 0.42) on a 0 to 1 scale, meaning that 78% (930/1192) of participants correctly understood the information in the video. In this case, between-groups variation was greater compared to subjective comprehension, with *z* scores ranging from –0.76 to 0.51. The complete overview is available in Table 2 below.

The data confirmed the internal consistency of the objective comprehension scale (α=.87).

Figure 2 displays the levels of subjective and objective comprehension by topic, source, and call to action.

XSL•FO
RenderX

Table 2. Subjective and objective comprehension by video version.

Version	Subjective comprehension *z* scores	Objective comprehension *z* scores
1	–0.02	0.00
2	0.00	–0.02
3	0.13	–0.15
4	0.19	0.00
5	–0.57	0.07
6	–0.46	0.07
7	–0.13	0.05
8	0.02	0.10
9	0.00	–0.76
10	0.16	–0.54
11	0.22	0.49
12	0.33	0.51

Figure 2. Objective versus subjective comprehension of the 12 video versions. The color displays the topic, the shape represents the source, and the filling shows the call to action. Exp.: exponential; HC: health care.

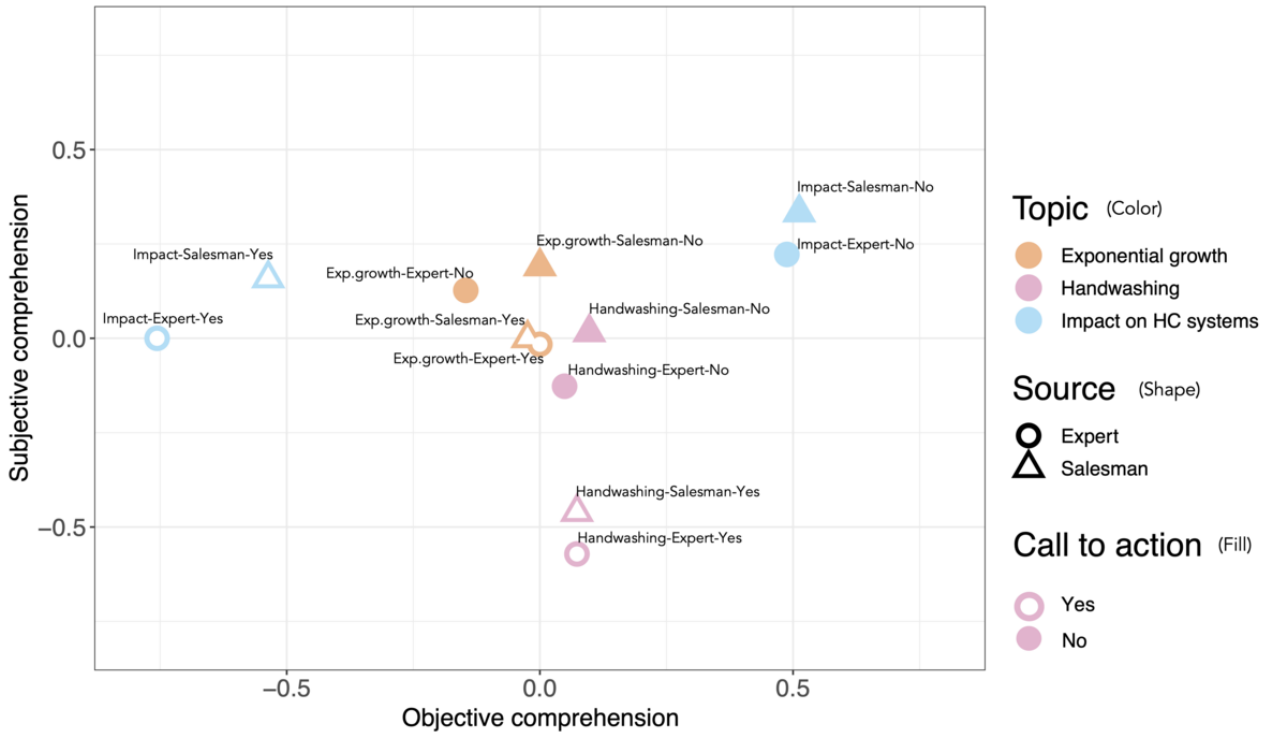

An overview of the multiple linear regression analysis is presented in Table 3. We found that age had an impact on comprehension, with increased age being associated with lower comprehension, both subjective and objective. Moreover, the effect was larger on subjective comprehension than objective comprehension (β=–.13 vs β=–.07), and both effects were statistically significant (*P*<.001 and *P*=.02).

Our data revealed that sex had a statistically significant effect on both perceived and objective comprehension (*P*<.001 and *P*=.02) and that although effects were comparable in magnitude, they had different directions (Table 3). In fact, female individuals had a higher degree of subjective comprehension than males (β=–.10), while male individuals had a higher level of objective comprehension than females (β=.07).

The data showed that whether participants resided in an urban or rural area did not have any statistically significant effect on either subjective or objective comprehension (all *P*>.05).

XSL•FO
RenderX

Table 3. Overview of the multiple linear regression analysis.

Predictor	Subjective comprehension		Objective comprehension	
	Standardized β	*P* value	Standardized β	*P* value
Age	–.13	<.001	–.07	.02
Sex	–.10	<.001	.07	.02
Residence	.03	.37	–.03	.32
Education	.11	<.001	.14	<.001
Income	.07	.03	.07	.02
Belief in science	.23	<.001	<.01	.87

Instead, we found that education had a statistically significant effect on comprehension, where higher levels of education were associated with both higher subjective (β=.11; *P*<.001) and objective (β=.14; *P*<.001) comprehension. Similarly, we observed that income had a statistically significant effect on both subjective (β=.07; *P*=.03) and objective (β=.07; *P*=.02) comprehension.

Finally, the data showed that the belief in science had no significant effect on objective comprehension (*P*=.87), while it had a relatively high effect on subjective comprehension (β=.23; *P*<.001), meaning that the higher the belief in science, the higher the level of perceived comprehension.

These effect sizes and significance were also confirmed by generalized additive model analysis.

Discussion

Principal Findings

The results of this randomized controlled trial displayed mixed effects of individual characteristics on the comprehension of pandemic video communication, both confirming existing evidence and providing novel insights.

Age

Our findings add evidence to the body of literature investigating how younger and older adults comprehend pandemic videos, as we found that increasing age is associated with lower comprehension, both perceived and objective. The older the people, the less they correctly understood the pandemic-related information presented in the video, and to an even greater extent, the less they perceived that they had comprehended that information.

This result is in line with previous research that suggests that attention should be given to age differences when developing pandemic communication videos [49,50].

Sex

Sex was found to have a different effect on subjective and objective comprehension. While female individuals scored higher than male individuals on perceived comprehension, they scored lower on objective comprehension. In other words, males were less overconfident than females in their capacity to understand the pandemic-related information in the video. This result is surprising in light of previous evidence showing that men tend to be more overconfident than women [51,52]. A possible explanation is that the confidence gap between men and women varies with age. In a research study involving 8665 people, Zenger and Folkman [53] found that the confidence gap had an inverted trend in the mid-40s and widened after. After this age, women became more confident and overconfident than men. Given that the participants in our study mainly belong to the age group of 60 years or older, it is plausible that we observed the same trend as reported by Zenger and Folkman [53].

Living Area

Rural or urban differences have been extensively reported in the literature in multiple disciplines, including several health domains [54,55]. We did not find any significant rural or urban difference in the comprehension of pandemic-related video information. This result is likely due to the structure of Norwegian society, as very little differences exist in terms of education and socioeconomic status (SES) between people living in rural and urban areas [56,57].

Socioeconomic Status

The effect of education and income on the pandemic video communication outcomes has been investigated previously [19,37,58]. We found a significant and positive effect of education and income on comprehension, both perceived and objective. The higher the education and income level, the higher the likelihood of having higher levels of confidence in one's ability to understand pandemic-related information (perceived comprehension) and to actually understand it (objective comprehension). Our findings add supporting evidence of the role that education and income play in social differences, encouraging policy makers to keep these differences into account when designing health communication strategies [59-61].

Belief in Science

Results concerning the effect of belief in science on comprehension are contrasting: the higher the belief in science, the higher the perceived comprehension, while not leading to any statistically significant increase in objective comprehension. In other words, those with higher belief in science will understand as much as others, while they feel they understand more. A possible explanation for this finding could be rooted in the Dunning-Kruger effect, where individuals with higher confidence in their understanding may overestimate their actual knowledge [62]. This overconfidence might stem from their trust in scientific sources, which increases their perceived

XSL•FO
RenderX

comprehension but does not necessarily reflect their actual understanding of the content. Prior research has shown that trust in science and scientific authorities can influence how individuals perceive and evaluate scientific information. People with higher beliefs in science may experience a confirmation bias, wherein they interpret information in a way that aligns with their preexisting beliefs and trust, thereby increasing their subjective sense of understanding [63]. Moreover, those with a strong belief in science might possess a higher level of scientific self-efficacy, which enhances their confidence in processing and understanding scientific information, leading to a higher perceived comprehension [64]. However, self-efficacy does not always correlate with actual performance [65], which might explain why their objective comprehension did not show a statistically significant increase.

Implications

Our findings contribute to the growing, but still limited, extent of evidence on pandemic communication to support authorities in effectively delivering complex information to the public. Our main result is that one-size-fits-all pandemic video communication does not exist: the same message will be comprehended differently by the younger and the older adults, male and female individuals, those with lower and higher levels of education, those who earn less and those who earn more, and those who have a low and those who have a high belief in science.

The implication of this is that health authorities should make different versions of pandemic videos. As the older adults had lower comprehension than the younger adults, videos aimed at the former could be more explicative, simpler in language, and distributed more intensively and through different channels, to favor reach, viewing, and consequently comprehension [66-68].

In addition, in the context of limited resources and to maximize outcomes, more attention could be given to those with a lower SES. In addition to the previous recommendations, as in the Norwegian context, minority individuals often belong to the lower SES groups, do not speak Norwegian, and live in more closed communities, pandemic communication videos could be created (or subtitled) in their native languages [60].

These 5 characteristics (age, sex, education, income, and belief in science), which are a limited subset of all potentially significant individual attributes, indicate that authorities should to a much higher degree understand the different target groups and how these groups access and interpret information. This will then allow authorities to create tailor-made pandemic videos that can reach and effectively communicate information to the public.

At the same time, we should pay attention to the potential drawbacks of popularizing complex scientific information. As demonstrated by Scharrer et al [69], making science easier to understand for laypeople inclines them to underrate their dependence on experts. This is a concrete risk with increased saliency during pandemics as decisions are generally made under time pressure and based on partial and uncertain information.

Directions for Future Research

Future research should explore targeted communication strategies that enhance comprehension among older adults since they will be an increasingly larger share of the population and have a relatively low level of understanding of the pandemic information presented in our videos. Future studies could test if videos using simpler language or interactive elements aid their comprehension of pandemic-related information. In addition, future research should investigate how confidence and comprehension in pandemic communication vary across different age and sex groups, and if the lack of urban-rural differences in comprehension holds true in other countries with more pronounced rural-urban divides. Moreover, developing and testing interventions aimed at reducing educational and income-related disparities in health communication comprehension is crucial, and we encourage researchers to address this matter. Furthermore, we need a deeper understanding of the mechanisms of belief in science and its impact on both perceived and objective comprehension; creating and testing interventions to align perceived and actual comprehension would enhance the understanding of these mechanisms. Finally, expanding the research to include a broader range of individual attributes beyond age, sex, place of living, education, income, and belief in science will help better tailor pandemic communication strategies to effectively reach and inform all segments of the population.

Limitations

Our limitations include the choice of the sample, meaning that findings apply to a subset of the Norwegian population, those who are members of the NAAF, while no conclusion can be drawn about their generalizability to a national level. The mean age of our sample is higher than the mean age of the national population. Moreover, as donating members of the NAAF, we have reason to believe that their SES is higher than the general population. Additional country-specific factors such pose a challenge to the generalizability of our findings internationally. Moreover, the ceiling effect, and therefore little variation, in subjective comprehension levels might hide effects that we were not able to detect. Finally, the use of one single objective comprehension question is another limitation. Our choice was due to the absence of a validated scale that could be applied to all three topics communicated in our study.

Conclusions

In this study, we investigated how the comprehension of pandemic-related video information is influenced by recipients' characteristics. Our findings indicate that younger individuals understand and perceive their comprehension of pandemic information better than older individuals. Female individuals report higher levels of perceived comprehension, while male individuals exhibit higher levels of objective understanding. In addition, higher education and income levels are associated with greater comprehension, both objective and subjective. Belief in science plays an important role for perceived comprehension, while it does not have any effect on people's ability to objectively comprehend pandemic-related information delivered through video.

XSL•FO
RenderX

These findings are relevant because effective communication during pandemics is vital for public health, ensuring that citizens are accurately informed and can take appropriate actions to protect themselves and others [70,71]. By highlighting the varying levels of comprehension across different demographic groups, our study underscores the need for tailored communication strategies that account for these differences. This approach is essential to maximize the reach and effectiveness of public health messages, particularly in diverse populations.

Our study contributes to the growing body of evidence on health communication [72-74] by providing insights into how different segments of the population understand pandemic-related information. It reinforces the notion that a one-size-fits-all approach is ineffective; instead, public health authorities should create multiple versions of pandemic videos to address the unique needs of various groups. These findings can guide future communication efforts, ensuring that critical health information is accessible and comprehensible to all citizens, ultimately enhancing public health outcomes during pandemics and other health crises.

Acknowledgments

The COVID Communication: Fighting a Pandemic through Translating Science project has received funding from the Trond Mohn Foundation (under grant agreement TMS2020TMT10) and the University of Stavanger. The authors acknowledge the Norwegian Air Ambulance Foundation and their members for their valuable contribution to this study.

Data Availability

The videos generated during and analyzed during this study are available in the OpenScience repository of the University of Stavanger [46].

Authors' Contributions

This study was conducted by an interdisciplinary team of researchers. DAL and KKB conceived the study, discussed the methodological approach, and analyzed the data. DAL drafted the initial manuscript. All the authors (JR, IS, SW, and KKB) provided critical feedback on the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File , 97 KB - jmir_v26i1e48882_app1.docx]

References

1. Wright RN, Faul L, Graner JL, Stewart GW, LaBar KS. Psychosocial determinants of anxiety about the COVID-19 pandemic. J Health Psychol 2022;27(10):2344-2360. [doi: 10.1177/13591053211030981] [Medline: 34348495]
2. Straker N. Vicissitudes of death anxiety during the COVID-19 pandemic. Psychodyn Psychiatry 2021;49(3):384-387. [doi: 10.1521/pdps.2021.49.3.384] [Medline: 34478325]
3. Raffetti E, Mondino E, Di Baldassarre G. COVID-19 vaccine hesitancy in Sweden and Italy: the role of trust in authorities. Scand J Public Health 2022;50(6):803-809 [FREE Full text] [doi: 10.1177/14034948221099410] [Medline: 35656576]
4. Skirbekk H, Magelssen M, Conradsen S. Trust in healthcare before and during the COVID-19 pandemic. BMC Public Health 2023;23(1):863 [FREE Full text] [doi: 10.1186/s12889-023-15716-6] [Medline: 37170208]
5. Risk communication and community engagement (RCCE) action plan guidance COVID-19 preparedness and response. World Health Organization. 2020 Mar 16. URL: https://www.who.int/publications-detail-redirect/risk-communication-and-community-engagement-(rcce)-action-plan-guidance [accessed 2024-02-19]
6. Weisæth L, Tønnessen A. Fear, information and control during a pandemic. Tidsskrift for Den norske legeforening. 2022. URL: https://tidsskriftet.no/en/2020/06/kronikk/fear-information-and-control-during-pandemic [accessed 2024-02-19]
7. Berg SH, Shortt MT, Thune H, Røislien J, O'Hara JK, Lungu DA, et al. Differences in comprehending and acting on pandemic health risk information: a qualitative study using mental models. BMC Public Health 2022;22(1):1440 [FREE Full text] [doi: 10.1186/s12889-022-13853-y] [Medline: 35902839]
8. Winograd DM, Fresquez CL, Egli M, Peterson EK, Lombardi AR, Megale A, et al. Rapid review of virus risk communication interventions: directions for COVID-19. Patient Educ Couns 2021;104(8):1834-1859 [FREE Full text] [doi: 10.1016/j.pec.2021.01.024] [Medline: 33583651]
9. Valenti A, Mirabile M, Cannone E, Boccuni F, Dionisi P, Fortuna G, et al. The impact of COVID-19 pandemics on the development of health risk communication: challenges and opportunities. Int J Environ Res Public Health 2022;20(1):645 [FREE Full text] [doi: 10.3390/ijerph20010645] [Medline: 36612966]

XSL•FO
RenderX

10. Nan X, Iles IA, Yang B, Ma Z. Public health messaging during the COVID-19 pandemic and beyond: lessons from communication science. Health Commun 2022;37(1):1-19. [doi: 10.1080/10410236.2021.1994910] [Medline: 34724838]
11. Lazris A, Rifkin E. Utilizing Effective Risk Communication in COVID-19: Highlighting the BRCT. Cham, Switzerland: Springer International Publishing; 2021.
12. Abrams EM, Greenhawt M. Risk communication during COVID-19. J Allergy Clin Immunol Pract 2020;8(6):1791-1794 [FREE Full text] [doi: 10.1016/j.jaip.2020.04.012] [Medline: 32304834]
13. Elleström L. The modalities of media: a model for understanding intermedial relations. In: Media Borders, Multimodality and Intermediality. London, United Kingdom: Palgrave Macmillan UK; 2010:11-48.
14. Sit SMM, Gong WJ, Ho SY, Lai AYK, Wong BYM, Wang MP, et al. A population study on COVID-19 information sharing: sociodemographic differences and associations with family communication quality and well-being in Hong Kong. Int J Environ Res Public Health 2022;19(6):3577 [FREE Full text] [doi: 10.3390/ijerph19063577] [Medline: 35329264]
15. Sun S, Liu Z, Zhai Y, Wang F. COVID-19 vaccines on TikTok: a big-data analysis of entangled discourses. Int J Environ Res Public Health 2022;19(20):13287 [FREE Full text] [doi: 10.3390/ijerph192013287] [Medline: 36293868]
16. Hillyer GC, Basch CH, Basch CE. Coverage of transmission of COVID-19 information on successive samples of youTube videos. J Community Health 2021;46(4):817-821 [FREE Full text] [doi: 10.1007/s10900-020-00956-z] [Medline: 33400077]
17. Shi CF, So MC, Stelmach S, Earn A, Earn DJD, Dushoff J. From science to politics: COVID-19 information fatigue on YouTube. BMC Public Health 2022;22(1):816 [FREE Full text] [doi: 10.1186/s12889-022-13151-7] [Medline: 35461254]
18. Shortt MT, Smeets I, Wiig S, Berg SH, Lungu DA, Thune H, et al. Shortcomings in public health authorities' videos on COVID-19: limited reach and a creative Gap. Front. Commun 2021;6:764220. [doi: 10.3389/fcomm.2021.764220]
19. Vandormael A, Adam M, Greuel M, Gates J, Favaretti C, Hachaturyan V, et al. The effect of a wordless, animated, social media video intervention on COVID-19 prevention: online randomized controlled trial. JMIR Public Health Surveill 2021;7(7):e29060 [FREE Full text] [doi: 10.2196/29060] [Medline: 34174778]
20. Tan RY, Pua AE, Wong LL, Yap KYL. Assessing the quality of COVID-19 vaccine videos on video-sharing platforms. Explor Res Clin Soc Pharm 2021;2:100035 [FREE Full text] [doi: 10.1016/j.rcsop.2021.100035] [Medline: 34568867]
21. Lungu DA, Røislien J, Wiig S, Shortt MT, Ferrè F, Berg SH, et al. The role of recipient characteristics in health video communication outcomes: scoping review. J Med Internet Res 2021;23(12):e30962 [FREE Full text] [doi: 10.2196/30962] [Medline: 34967758]
22. Danila MI, Outman RC, Rahn EJ, Mudano AS, Thomas TF, Redden DT, et al. A multi-modal intervention for activating patients at risk for osteoporosis (APROPOS): rationale, design, and uptake of online study intervention material. Contemp Clin Trials Commun 2016;4:14-24 [FREE Full text] [doi: 10.1016/j.conctc.2016.06.010] [Medline: 27453960]
23. Hillen MA, de Haes HCJM, Verdam MGE, Smets EMA. Does source of patient recruitment affect the impact of communication on trust? Patient Educ Couns 2014;95(2):226-230. [doi: 10.1016/j.pec.2014.01.008] [Medline: 24534265]
24. Nurmi J, Knittle K, Ginchev T, Khattak F, Helf C, Zwickl P, et al. Engaging users in the behavior change process with digitalized motivational interviewing and gamification: development and feasibility testing of the precious app. JMIR Mhealth Uhealth 2020;8(1):e12884 [FREE Full text] [doi: 10.2196/12884] [Medline: 32003750]
25. Berry DL, Halpenny B, Hong F, Wolpin S, Lober WB, Russell KJ, et al. The personal patient profile-prostate decision support for men with localized prostate cancer: a multi-center randomized trial. Urol Oncol 2013;31(7):1012-1021 [FREE Full text] [doi: 10.1016/j.urolonc.2011.10.004] [Medline: 22153756]
26. Bekalu MA, Bigman CA, McCloud RF, Lin LK, Viswanath K. The relative persuasiveness of narrative versus non-narrative health messages in public health emergency communication: evidence from a field experiment. Prev Med 2018;111:284-290 [FREE Full text] [doi: 10.1016/j.ypmed.2017.11.014] [Medline: 29154794]
27. Albert NM, Dinesen B, Spindler H, Southard J, Bena JF, Catz S, et al. Factors associated with telemonitoring use among patients with chronic heart failure. J Telemed Telecare 2017;23(2):283-291. [doi: 10.1177/1357633X16630444] [Medline: 26869144]
28. Hannawa AF. "Explicitly implicit": examining the importance of physician nonverbal involvement during error disclosures. Swiss Med Wkly 2012;142:w13576 [FREE Full text] [doi: 10.4414/smw.2012.13576] [Medline: 22573515]
29. Prieto-Pinto L, Lara-Díaz MF, Garzón-Orjuela N, Herrera D, Páez-Canro C, Reyes JH, et al. Effectiveness assessment of maternal and neonatal health video clips in knowledge transfer using neuromarketing tools: a randomized crossover trial. PLoS One 2019;14(5):e0215561 [FREE Full text] [doi: 10.1371/journal.pone.0215561] [Medline: 31067282]
30. Laaksonen M, Prättälä R, Lahelma E. Sociodemographic determinants of multiple unhealthy behaviours. Scand J Public Health 2003;31(1):37-43. [doi: 10.1080/14034940210133915] [Medline: 12623523]
31. Dreger S, Buck C, Bolte G. Material, psychosocial and sociodemographic determinants are associated with positive mental health in Europe: a cross-sectional study. BMJ Open 2014;4(5):e005095 [FREE Full text] [doi: 10.1136/bmjopen-2014-005095] [Medline: 24871540]
32. Hearst MO, Sirard JR, Forsyth A, Parker ED, Klein EG, Green CG, et al. The relationship of area-level sociodemographic characteristics, household composition and individual-level socioeconomic status on walking behavior among adults. Transp Res Part A Policy Pract 2013;50:149-157 [FREE Full text] [doi: 10.1016/j.tra.2013.01.006] [Medline: 23729994]

XSL•FO
RenderX

33. Wang S, Li B, Wu Y, Ungvari GS, Ng CH, Fu Y, et al. Relationship of sleep duration with sociodemographic characteristics, lifestyle, mental health, and chronic diseases in a large chinese adult population. J Clin Sleep Med 2017;13(3):377-384 [FREE Full text] [doi: 10.5664/jcsm.6484] [Medline: 27998377]
34. Grindel CG, Brown L, Caplan L, Blumenthal D. The effect of breast cancer screening messages on knowledge, attitudes, perceived risk, and mammography screening of African American women in the rural South. Oncol Nurs Forum 2004;31(4):801-808. [doi: 10.1188/04.ONF.801-808] [Medline: 15252434]
35. Hickey S, Merz EL, Malcarne VL, Gunsauls DC, Huang J, Sadler GR. Breast cancer education for the deaf community in American sign language. Oncol Nurs Forum 2013;40(3):E86-E91. [doi: 10.1188/13.ONF.E86-E91] [Medline: 23615152]
36. Perez SL, Kravitz RL, Bell RA, Chan MS, Paterniti DA. Characterizing internet health information seeking strategies by socioeconomic status: a mixed methods approach. BMC Med Inform Decis Mak 2016;16:107 [FREE Full text] [doi: 10.1186/s12911-016-0344-x] [Medline: 27506607]
37. Higashi RT, Sweetenham JW, Israel AD, Tiro JA. COVID-19 communication from seven health care institutions in north Texas for english- and Spanish-speaking cancer patients: mixed method website study. JMIR Cancer 2021;7(3):e30492 [FREE Full text] [doi: 10.2196/30492] [Medline: 34346886]
38. Dudley MZ, Limaye RJ, Omer SB, O'Leary ST, Ellingson MK, Spina CI, et al. Factors associated with referring close contacts to an app with individually-tailored vaccine information. Vaccine 2020;38(13):2827-2832 [FREE Full text] [doi: 10.1016/j.vaccine.2020.02.019] [Medline: 32098739]
39. McKenzie LB, Roberts KJ, Collins CL, Clark RM, Smith KC, Manganello J. Maternal knowledge, attitudes, and behavioral intention after exposure to injury prevention recommendations in the news media. J Health Commun 2019;24(7-8):625-632. [doi: 10.1080/10810730.2019.1646357] [Medline: 31378155]
40. Palacio AM, Vidot DC, Tamariz LJ, Uribe C, Hazel-Fernandez L, Li H, et al. Can we identify minority patients at risk of nonadherence to antiplatelet medication at the time of coronary stent placement? J Cardiovasc Nurs 2017;32(6):522-529 [FREE Full text] [doi: 10.1097/JCN.0000000000000387] [Medline: 28060084]
41. Farias M, Newheiser AK, Kahane G, de Toledo Z. Scientific faith: belief in science increases in the face of stress and existential anxiety. J Exp Soc Psychol 2013;49(6):1210-1213 [FREE Full text] [doi: 10.1016/j.jesp.2013.05.008] [Medline: 24187384]
42. Brzezinski A, Kecht V, van Dijcke D, Wright AL. Science skepticism reduced compliance with COVID-19 shelter-in-place policies in the United States. SSRN Journal. Preprint posted online on May 4, 2020 URL: https://www.ssrn.com/abstract=3587990 [accessed 2022-01-31]
43. Stosic MD, Helwig S, Ruben MA. Greater belief in science predicts mask-wearing behavior during COVID-19. Pers Individ Dif 2021;176:110769 [FREE Full text] [doi: 10.1016/j.paid.2021.110769] [Medline: 33612904]
44. Lungu DA, Røislien J, Berg SH, Smeets I, Shortt MT, Thune H, et al. Assessing the effect of nonvisual information factors in pandemic-related video communication: randomized controlled between-subjects experiment. J Med Internet Res 2023;25:e42528 [FREE Full text] [doi: 10.2196/42528] [Medline: 37610820]
45. Røislien J, O'Hara JK, Smeets I, Brønnick K, Berg SH, Shortt MT, et al. Creating effective, evidence-based video communication of public health science (COVCOM Study): protocol for a sequential mixed methods effect study. JMIR Res Protoc 2022;11(3):e34275 [FREE Full text] [doi: 10.2196/34275] [Medline: 35147500]
46. Lungu DA. Covcom test videos. DataverseNO. 2022. URL: https://doi.org/10.18710/EZQR78 [accessed 2024-10-26]
47. Lima Sanches C, Augereau O, Kise K. Estimation of reading subjective understanding based on eye gaze analysis. PLoS One 2018 Oct 25;13(10):e0206213 [FREE Full text] [doi: 10.1371/journal.pone.0206213] [Medline: 30359411]
48. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria URL: https://www.R-project.org/ [accessed 2024-10-26]
49. Paakkari L, Okan O. COVID-19: health literacy is an underestimated problem. Lancet Public Health 2020;5(5):e249-e250 [FREE Full text] [doi: 10.1016/S2468-2667(20)30086-4] [Medline: 32302535]
50. Wolfe K, Sirota M, Clarke ADF. Age differences in COVID-19 risk-taking, and the relationship with risk attitude and numerical ability. R Soc Open Sci 2021;8(9):201445 [FREE Full text] [doi: 10.1098/rsos.201445] [Medline: 34603740]
51. Lundeberg MA, Fox PW, Punccohar J. Highly confident but wrong: gender differences and similarities in confidence judgments. Journal of Educational Psychology 1994;86(1):114-121. [doi: 10.1037//0022-0663.86.1.114]
52. Bengtsson U, Kjellgren K, Hallberg I, Lundin M, Mäkitalo Å. Patient contributions during primary care consultations for hypertension after self-reporting via a mobile phone self-management support system. Scand J Prim Health Care 2018;36(1):70-79 [FREE Full text] [doi: 10.1080/02813432.2018.1426144] [Medline: 29343156]
53. Zenger J, Folkman J. Research: women score higher than men in most leadership skills. Harvard Business Review. 2019. URL: https://hbr.org/2019/06/research-women-score-higher-than-men-in-most-leadership-skills [accessed 2024-02-14]
54. Flood D, Geldsetzer P, Agoudavi K, Aryal KK, Brant LCC, Brian G, et al. Rural-urban differences in diabetes care and control in 42 low- and middle-income countries: a cross-sectional study of nationally representative individual-level data. Diabetes Care 2022;45(9):1961-1970 [FREE Full text] [doi: 10.2337/dc21-2342] [Medline: 35771765]
55. Saenz JL, Downer B, Garcia MA, Wong R. Cognition and context: rural-urban differences in cognitive aging among older Mexican adults. J Aging Health 2018;30(6):965-986 [FREE Full text] [doi: 10.1177/0898264317703560] [Medline: 28553815]

XSL•FO
RenderX

56. Ulset VS, Venter Z, Kozák M, Nordbø ECA, von Soest T. Increased nationwide use of green spaces in Norway during the COVID-19 pandemic. Environ Int 2023;180:108190 [FREE Full text] [doi: 10.1016/j.envint.2023.108190] [Medline: 37717522]
57. Reas DL, Nygård JF, Svensson E, Sørensen T, Sandanger I. Changes in body mass index by age, gender, and socio-economic status among a cohort of Norwegian men and women (1990-2001). BMC Public Health 2007;7:269 [FREE Full text] [doi: 10.1186/1471-2458-7-269] [Medline: 17903273]
58. He C, Liu H, He L, Lu T, Li B. More collaboration, less seriousness: investigating new strategies for promoting youth engagement in government-generated videos during the COVID-19 pandemic in China. Comput Human Behav 2022;126:107019 [FREE Full text] [doi: 10.1016/j.chb.2021.107019] [Medline: 34602727]
59. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promotion International 2000;15(3):67. [doi: 10.1093/heapro/15.3.259]
60. Kreps GL, Sparks L. Meeting the health literacy needs of immigrant populations. Patient Educ Couns 2008;71(3):328-332. [doi: 10.1016/j.pec.2008.03.001] [Medline: 18387773]
61. Marmot M. The influence of income on health: views of an epidemiologist. Health Aff (Millwood) 2002;21(2):31-46. [doi: 10.1377/hlthaff.21.2.31] [Medline: 11900185]
62. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999;77(6):1121-1134. [doi: 10.1037//0022-3514.77.6.1121] [Medline: 10626367]
63. Lewandowsky S, Gignac GE, Vaughan S. The pivotal role of perceived scientific consensus in acceptance of science. Nature Clim Change 2012;3(4):399-404. [doi: 10.1038/nclimate1720]
64. Bandura A. Self-Efficacy: The Exercise of Control. New York, NY: W H Freeman/Times Books/ Henry Holt & Co; 1997.
65. Kleitman S, Gibson J. Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. Learning and Individual Differences 2011;21(6):728-735. [doi: 10.1016/j.lindif.2011.08.003]
66. Kreuter M, Skinner C. Tailoring: what's in a name? Health Educ Res 2000;15(1):1-4. [doi: 10.1093/her/15.1.1] [Medline: 10788196]
67. Kreuter MW, McClure SM. The role of culture in health communication. Annu Rev Public Health 2004;25:439-455. [doi: 10.1146/annurev.publhealth.25.101802.123000] [Medline: 15015929]
68. Morrow D, Leirer V, Altieri P, Fitzsimmons C. When expertise reduces age differences in performance. Psychol Aging 1994;9(1):134-148. [doi: 10.1037//0882-7974.9.1.134] [Medline: 8185861]
69. Scharrer L, Rupieper Y, Stadtler M, Bromme R. When science becomes too easy: science popularization inclines laypeople to underrate their dependence on experts. Public Underst Sci 2017;26(8):1003-1018. [doi: 10.1177/0963662516680311] [Medline: 27899471]
70. Communicating risk in public health emergencies: a WHO guideline for emergency risk communication (ERC) policy and practice. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/259807/9789241550208-eng.pdf?sequence=2 [accessed 2024-10-26]
71. Seale H, Heywood AE, Leask J, Sheel M, Thomas S, Durrheim DN, et al. COVID-19 is rapidly changing: examining public perceptions and behaviors in response to this evolving pandemic. PLoS One 2020;15(6):e0235112 [FREE Full text] [doi: 10.1371/journal.pone.0235112] [Medline: 32574184]
72. Reynolds B, Quinn Crouse S. Effective communication during an influenza pandemic: the value of using a crisis and emergency risk communication framework. Health Promot Pract 2008;9(4 Suppl):13S-17S. [doi: 10.1177/1524839908325267] [Medline: 18936255]
73. Glik DC. Risk communication for public health emergencies. Annu Rev Public Health 2007;28:33-54. [doi: 10.1146/annurev.publhealth.28.021406.144123] [Medline: 17222081]
74. Vaughan E, Tinker T. Effective health risk communication about pandemic influenza for vulnerable populations. Am J Public Health 2009;99 Suppl 2(Suppl 2):S324-S332. [doi: 10.2105/AJPH.2009.162537] [Medline: 19797744]

Abbreviations

GDPR: General Data Protection Regulation
NAAF: Norwegian Air Ambulance Foundation
SES: socioeconomic status

XSL•FO
RenderX

Edited by A Mavragani; submitted 10.05.23; peer-reviewed by C Baur, F van de Baan; comments to author 07.12.23; revised version received 08.03.24; accepted 27.09.24; published 04.12.24.

Please cite as:
Lungu DA, Røislien J, Smeets I, Wiig S, Brønnick KK
Individual Characteristics in the Comprehension of Pandemic Video Communication: Randomized Controlled Between-Subjects Design
J Med Internet Res 2024;26:e48882
URL: https://www.jmir.org/2024/1/e48882
doi:10.2196/48882
PMID:

©Daniel Adrian Lungu, Jo Røislien, Ionica Smeets, Siri Wiig, Kolbjørn Kallesten Brønnick. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Discussions of Cannabis Over Patient Portal Secure Messaging: Content Analysis

Vishal A Shetty[1,2], MS; Christina M Gregor[1], BS; Lorraine D Tusing[1], BA; Apoorva M Pradhan[1], MPH; Katrina M Romagnoli[1,3], MLiS, MS, PhD; Brian J Piper[1,4], MS, PhD; Eric A Wright[1,5], MPH, PharmD

[1]Center for Pharmacy Innovation and Outcomes, Geisinger, Danville, PA, United States

[2]Department of Health Promotion and Policy, University of Massachusetts, Amherst, MA, United States

[3]Department of Population Health Sciences, Geisinger, Danville, PA, United States

[4]Department of Medical Education, Geisinger Commonwealth School of Medicine, Scranton, PA, United States

[5]Department of Bioethics and Decision Sciences, Geisinger College of Health Sciences, Gesinger, Danville, PA, United States

Corresponding Author:
Vishal A Shetty, MS
Department of Health Promotion and Policy
University of Massachusetts
715 North Pleasant St.
Amherst, MA, 01003
United States
Phone: 1 413 230 4015
Email: vashetty@geisinger.edu

Abstract

Background: Patient portal secure messaging allows patients to describe health-related behaviors in ways that may not be sufficiently captured in standard electronic health record (EHR) documentation, but little is known about how cannabis is discussed on this platform.

Objective: This study aimed to identify patient and provider secure messages that discussed cannabis and contextualize these discussions over periods before and after its legalization for medical purposes in Pennsylvania.

Methods: We examined 382,982 secure messages sent by 15,340 patients and 6101 providers from an integrated health delivery system in Pennsylvania, United States, from January 2012 to June 2022. We used an unsupervised natural language processing approach to construct a lexicon that identified messages explicitly discussing cannabis. We then conducted a qualitative content analysis on a random sample of identified messages to understand the medical reasons behind patients' use, the primary purposes of the cannabis-related discussions, and changes in these purposes over time.

Results: We identified 1782 messages sent by 1098 patients (7.2% of total patients in the study) and 800 messages sent by 430 providers (7% of total providers in the study) as explicitly discussing cannabis. The most common medical reasons for use stated by patients in 190 sampled messages included pain or a pain-related condition (50.5% of messages), anxiety (13.7% of messages), and sleep (11.1% of messages). We coded 56 different purposes behind the mentions of cannabis in patient messages and 33 purposes in 100 sampled provider messages. In years before the legalization (2012-2016), patient and provider messages (n=20 for both) were primarily driven by discussions about cannabis screening results (38.9% and 76.5% of messages, respectively). In the years following legalization (2017-2022), patient messages (n=170) primarily involved seeking assistance to facilitate medical use (35.2% of messages) and reporting current use (25.3% of messages). Provider messages (n=80) were driven by giving assistance with medical marijuana access (27.5% of messages) and stating that they were unable to refer, prescribe or recommend medical marijuana (26.3% of messages).

Conclusions: Patients showed a willingness to discuss cannabis use over patient portal secure messages and expressed interest in use after the legalization of medical marijuana. Some providers responded to patient inquiries with assistance in obtaining access to medical marijuana, while others cautioned patients on the risks of use. Insight into cannabis-related discussions through secure messages can help health systems determine opportunities to improve care processes around patients' cannabis use, and providers should be supported to communicate accurate and consistent information.

(J Med Internet Res 2024;26:e63311) doi:10.2196/63311

XSL•FO
RenderX

KEYWORDS

patient portal; secure message; marijuana; patient-provider communication; message content; content analysis; United States; pain; anxiety; depression; insomnia; electronic messaging; electronic health record; EHR; cannabis

Introduction

Medical marijuana (defined as derivatives of the *Cannabis sativa* plant used to ease symptoms for medical conditions certified by a clinician and acquired through state-sanctioned dispensaries) [1] has been legalized in 38 US states, of which 24 have legalized recreational use as of 2024 [2]. In Pennsylvania, the Medical Marijuana Act was signed into law in 2016, legalizing the use of medical marijuana for 17 medical conditions [3]. There are now 24 "qualifying conditions" for which patients can claim eligibility to participate in the state's medical marijuana program [4]. Across the United States, many patients are using marijuana to manage medical conditions, including intractable pain, anxiety and depression, and insomnia [5-10]. This patient use comes despite known risks associated with use, such as cannabis use disorder, impaired cognitive functioning and motor coordination, mood disorders, and exacerbation of psychotic disorders [11-15]. Patient use is also seen despite limited evidence for the efficacy of medical marijuana in treating many health conditions for which patients commonly use [16-18]. Given the potential consequences of cannabis (defined as any of several psychoactive preparations derived from the plant genus *Cannabis* used for recreational or medical purposes, which may or may not be legal depending on the state, preparation, and source) [19], misuse among patient populations, several recent studies have focused on documentation of cannabis use in electronic health records (EHRs) to identify patients who use and understand the nature of their use [20-22]. However, such effort has not been extended to other sources of unstructured textual clinical information, such as secure electronic messaging within patient portal systems.

Patient portal systems have been widely implemented across health care organizations in the United States [23]. Among organizations that have a patient portal, patient adoption is estimated to increase between 5% and 10% per year [24]. In addition, large multispecialty health systems have seen substantial increases in patient portal message volume within the last decade [25]. From 2020 to 2023, the number of patients with an active patient portal account at Geisinger, an integrated health delivery system in Pennsylvania, increased 59.3% from 533,660 to 850,208, and the number of patients who sent secure messages increased 146.3% from 223,630 to 550,740. These patients sent 1,178,288 and 1,832,288 total secure messages over this period, respectively. Secure messaging within patient portals enables asynchronous communication between patients and providers, which allows patients to express their thoughts and explain health-related behaviors in ways that cannot be sufficiently captured in standard EHR documentation [26-28]. Investigation of cannabis-related discussions by secure messaging can give health systems insight into the dynamics of patient use and reveal how providers navigate this topic with patients.

In this study, we aimed to identify patient and provider secure messages that discussed cannabis and contextualize these discussions around the legalization of medical marijuana in Pennsylvania. We used an unsupervised natural language processing approach to construct a lexicon to flag messages explicitly discussing cannabis. We then conducted a qualitative content analysis to understand the medical reasons behind patients' use (if any) and the overall purposes of the cannabis-related discussions.

Methods

Study Population and Dataset

This study was conducted at Geisinger, one of the larger integrated health systems in the United States, serving approximately 1 million patients annually across 45 counties in central and northeast Pennsylvania. Geisinger comprises 9 acute care hospitals, an alcohol and chemical dependency treatment center, 133 specialty and primary care centers, 2 research centers, and an insurance company, Geisinger Health Plan (GHP) [29]. The study population included patients who had been flagged as cannabis users (medical or recreational) with at least one inpatient or outpatient encounter from January 2013 to June 2022, and at least one secure message sent a year before or a year after the date of their cannabis use indication. Messages were eligible for the analysis if they were designated in the "Patient Medical Advice Request" category of platform messages. These are patient-initiated messages which represent the primary dialogue between patients and providers. Flagged cannabis users were identified by discrete indications in the "Smart Data Element," "Problem List," "Encounter Diagnosis," "Labs," "Medications," and "Social History" areas of the EHR (more details in Multimedia Appendix 1 for details on discrete indications of cannabis use). A total of 15,340 adult (age≥18 years) patients were included in the study population.

Message Selection

Among the study population, all patient messages were sent a year before and a year after their discrete cannabis use identification, and all corresponding provider response messages were extracted (up to and including June 2022). To identify secure messages that discussed cannabis, we filtered messages that contained at least one mention of a cannabis-related keyword based on a constructed lexicon. We started with an initial lexicon derived from previous literature [22]. We excluded terms related to cannabinoid-based prescription drugs such as "Epidiolex," "Marinol," and "Dronabinol," as the focus of our analysis was on plant-based cannabis products. To refine our lexicon and mitigate against the misclassification of messages, we constructed word similarity lists using Word2Vec. Word2vec is an unsupervised natural language processing method that generates vector representations for each word in a set of documents (eg, secure messages) [30]. Words that appear in similar contexts will have similar representations. For each word in our initial lexicon, we generated the top 10 words

XSL•FO
RenderX

that most frequently appear in similar contexts (more details in Multimedia Appendix 2 for the initial lexicon and generated similarity lists).

Members of the lexicon were removed if no similarity list was generated (ie, the word appeared fewer than 5 times in all messages) or if its similarity list did not contain at least two words related to cannabis. A word in a similarity list was added to the lexicon if it had a meaning related to cannabis or was a misspelling of a member of the lexicon. Note the similarity lists for the term "pot" revealed that it appeared frequently as cannabis-related words but also in unrelated contexts. To minimize misclassification in these cases, we removed all messages that contained "neti pot," "netti pot," "netty pot," and "netting pot." Our final lexicon included the following words: "cannabis," "cannabinoid," "cbd," "edible," "edibles," "hemp," "ilera," "indica," "mariajuana," "marihuana," "marijuana," "marijuanna," "marj," "marjuana" "mj," "mmj," "pot," "sativa," "thc," "tincture," "tinctures," "weed." Text processing was conducted using Python (version 3.8.10; Python Software Foundation), and Word2vec representations were generated using the library Gensim [31].

Codebook Development

Messages discussing cannabis were organized into four groups: (1) patient messages sent from 2012-2016, (2) patient messages sent from 2017 to 2022, (3) provider messages sent from 2012 to 2016, and (4) provider messages sent from 2017 to 2022. Messages were organized by the years in which they were sent in order to separate those that occurred before and after the state medical marijuana program was signed into law in Pennsylvania (2017 being the first full year after the passing of the law) [3]. One member of the research team (VAS) reviewed a random sample of 10 messages from each group to identify broad concepts in the text related to patient cannabis use and the purpose behind the mention of cannabis. This step was repeated until a point of saturation in concepts was reached [32]. 50 patient messages and 50 provider messages were reviewed (20 messages from 2012-2016 groups, 30 from 2017-2022 groups). This process produced 2 initial codebooks for patient and provider messages separately. Each codebook was reviewed by members of the research team.

Our analysis of message content was conducted using a random sample of at least 10% of messages from each group (minimum of 20 messages) to reach sufficiency for analysis [33]. Using the initial codebooks, two coders (VAS and CMG) independently coded 5 messages from each group. Discrepancies in codes were discussed, and modifications were made to the codebook. This process was repeated until an interrater reliability (Cohen κ) [34] of at least 0.75 was achieved; two rounds of double coding were needed to achieve sufficient interrater reliability [35]. A κ of 0.66 was achieved in the first round, and 0.81 was achieved in the second round. The remaining messages in each of the four message groups were then split evenly between the two coders for single coding. After each coder had reviewed their respective half of messages for a group, suggestions for new codes and modifications to existing codes were discussed between the coders and a third member (LDT). Changes were made to the appropriate codebook, and messages in that group were recoded over iterative rounds until no changes were suggested. This process was repeated for each message group to finalize codebooks.

For patient messages, five primary content categories were coded: (1) correctly classified message (ie, the keyword which flagged the message as pertaining to a plant-based product), (2) message author (patient or nonpatient), (3) time of cannabis use, (4) reason for cannabis use, and (5) purpose of cannabis mention in message. "Time of cannabis" use codes were mutually exclusive. An exception was made if a message described the use of multiple different cannabis products. Multiple "reasons for cannabis use" and "purposes of the cannabis mention" could be coded in each message. "Reason for cannabis use" codes included explicitly named health issues patients used, were using, or wanted to use cannabis for. "Purpose of the cannabis mention" codes were determined based on the coders' judgment of the primary drivers behind the patient bringing up cannabis in the message.

For provider messages, four primary content categories were coded: (1) correctly classified message, (2) message recipient (patient or nonpatient), (3) reason for cannabis use, and (4) purpose of cannabis mention. "Reason for cannabis use" codes were determined based on the provider's description of what the patient had used or wanted to use for (more details in Multimedia Appendix 3 for the finalized codebooks with code definitions).

Analysis and Synthesis

Frequencies were calculated for all finalized codes. We used two proportion *z* tests to calculate differences in code frequencies between messages sent from 2012-2016 and messages sent from 2017-2022. We conducted a thematic analysis based on frequency patterns in codes to synthesize the information collected from messages [33]. Each coder independently rereviewed example messages of coded categories and analyzed code frequencies to construct qualitative memos in which themes were chosen to capture how patients and providers have been using secure messages to discuss cannabis, with a particular focus on how message content reflected the legal status of medical marijuana. Each coder justified chosen themes by citing the data that supported a theme and by describing the broader implications of the theme in terms of patient-provider interactions around cannabis use. Coders comprehensively discussed themes to come to a consensus, and themes were reviewed and further refined with the full research team.

Ethical Considerations

This study was approved by the Geisinger Institutional Review Board (IRB00008345). All procedures were performed in accordance with the relevant guidelines and regulations.

Results

Message, Patient, Provider Characteristics

Among 206,699 messages sent by 15,340 patients from 2012-2022, we identified 1782 messages (0.9%) from 1098 patients (7.2%; 1.6 average messages per patient) as explicitly

XSL•FO
RenderX

mentioning cannabis (either medical marijuana or recreational). Among 176,283 messages sent by 6101 providers from 2012-2022, we identified 800 messages (0.5%) from 430 providers (7.0%; 1.9 average messages per provider) as explicitly mentioning cannabis. The proportion of patient messages mentioning cannabis increased from 0.3% from 2012-2016 to 0.97% from 2017-2022 (Figure 1). The proportion of provider messages mentioning cannabis increased from 0.25% from 2012-2016 to 0.51% from 2017-2022. The largest increase in the volume of patient and provider messages mentioning cannabis occurred from 2017 to 2018 (57 to 251, 340.4% and 34 to 108, 217.6% increase, respectively). This coincided with the opening of the first medical marijuana dispensary in Pennsylvania in February 2018 [36]. For all patient and provider messages from 2017 to 2018, volume increased by 47.5% (11,835 to 17,454) and 48.3% (10,638 to 15,773), respectively (more details in Multimedia Appendices 4 and 5 for trends in patient and provider message volume over time).

Figure 1. Proportion of all secure messages that include a mention of cannabis from 2012-2022.

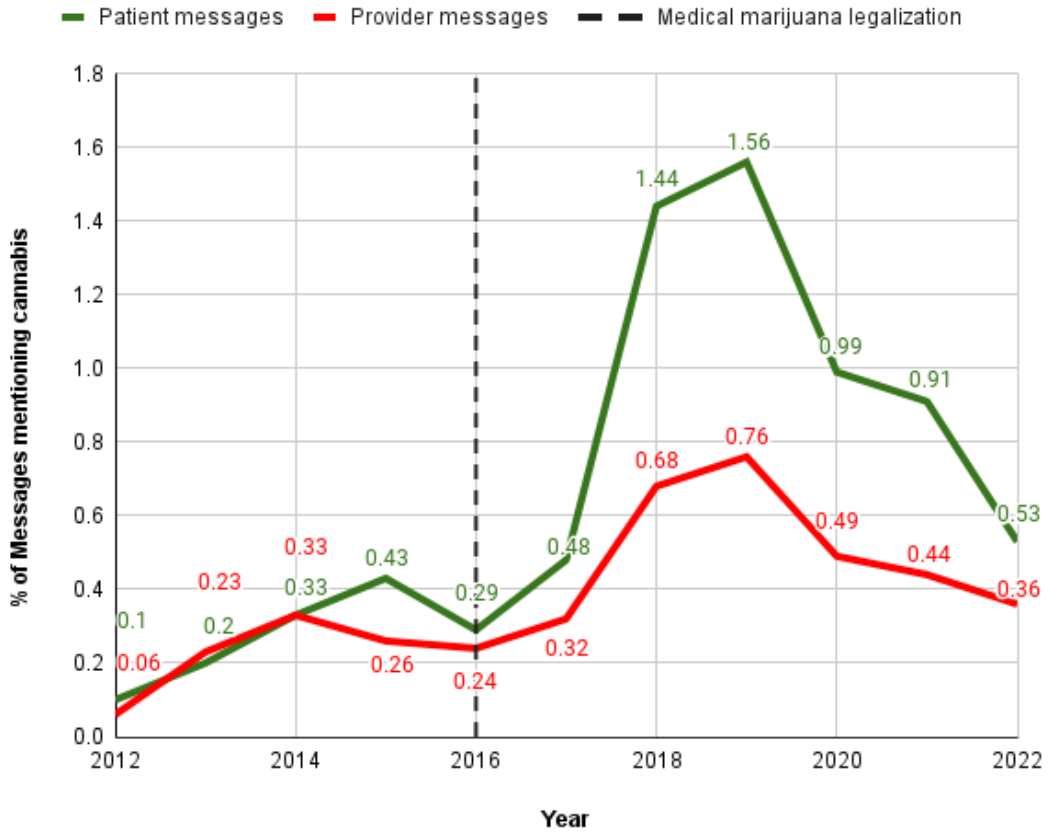

Patients in the total study population were primarily White non-Hispanic (84.9%, 13,028/15,340), female (60.6%, 9301/15,340), had a mean age of 39.7 (SD 16), and were insured under GHP (37.9%, 5812/15,340). Patient messages were primarily sent by those insured under GHP (40.4%, 83,644/206,699), a commercial plan (24.2%, 50,081/206,699), Medicare (16.9%, 35,029/206,699), and Medicaid (12.7%, 26,323/206,699). Among patients who either sent or only received a message mentioning cannabis (n=1098 and n=150, respectively), most were White non-Hispanic (90%, 1118/1248), female (63%, 786/1248), had a mean age of 41.8 (SD 15.6), and were insured under GHP (38.9%, 486/1248). Patient messages discussing cannabis were primarily sent by those insured under GHP (38%, 678/1782), a commercial plan (26.5%, 472/1782), Medicare (19.3%, 344/1782), and Medicaid (10.4%, 185/1782). Provider messages to patients in the study population were primarily sent from family medicine (38.9%, 68,511/176,283), obstetrics and gynecology (6.8%, 12,056/176,283), gastroenterology (5.1%, 9078/176,283), neurology (5.1%, 8963/176,283), and internal medicine (4.7%, 8350/176,283). Messages were primarily sent by physicians (33.7%, 59,488/176,283), registered nurses (15.3%, 26,943/176,283), and physician assistants (13.5%, 23,865/176,283). Provider messages discussing cannabis were primarily sent from family medicine (50.9%, 407/800), neurology (13.1%, 105/800), internal medicine (5.1%, 41/800), gastroenterology (4.1%, 33/800), and psychiatry (3.4%, 27/800). Messages were primarily sent by physicians (61.1%, 489/800), physician assistants (15.5%, 124/800), and registered nurses (11%, 88/800).

Message Content Frequencies

In the coded patient messages sent from both 2012-2016 and 2017-2022, at least half included a health-related reason for having used, using, or wanting to use cannabis (Table 1). Patients named 34 different reasons for use, though there may be overlap between some, such as anxiety (1/18, 5.6% of messages in 2012-2016; 25/162, 15.4% in 2017-2022) and racing thoughts (1/162, 0.6% of messages in 2017-2022). Note that Table 1 only includes the three most prevalent categories.

XSL•FO
RenderX

Table 1. Frequency of patient message content categories related to cannabis by year of message.

Content category	2012-2016 (N=20), n (%)	2017-2022 (N=170), n (%)	*P* value[a]
Correctly classified message			
Yes	18 (90)	162 (95.3)	—[b]
No, not related to cannabis	1 (5)	8 (4.7)	—
No, reference to Rx	1 (5)	—	—
Message author[c]			
Patient	17 (94.4)	153 (94.4)	—
Patient proxy	1 (5.6)	9 (5.6)	—
Time of use[c,d]			
Past, but not current	5 (27.8)	18 (11.1)	.04
Current	5 (27.8)	53 (32.7)	.67
Interest or intent to use	5 (27.8)	80 (49.4)	.08
None	3 (16.7)	18 (11.1)	.48
Reason for use[c,e]			
Not specified	9 (50)	70 (43.2)	.58
Pain or pain condition	12 (66.7)	84 (51.9)	.23
Anxiety	1 (5.6)	25 (15.4)	.26
Sleep	—	21 (13)	.10
Purpose of cannabis mention[c,e]			
Seeking assistance to facilitate use	4 (22.2)	57 (35.2)	.27
Seeking guidance on recommendation to use	1 (5.6)	23 (14.2)	.31
Statement of plan to use	—	17 (10.5)	.15
Reporting current use	3 (16.7)	41 (25.3)	.42
Reporting no current use	5 (27.8)	13 (8)	.01
Discussion of cannabis screening results	7 (38.9)	9 (5.6)	<.001

[a]*P* values were calculated using two proportion *z* tests.

[b]Not available.

[c]Percentages are based on values of correctly classified messages (ie, n=18 for 2012-2016 messages, n=162 for 2017-2022 messages).

[d]Some patients indicated using different cannabis products at varying times, in which case they were assigned to multiple "Time of use" categories.

[e]Only includes a subset of coded categories.

Reason for use of coded inpatient messages (more details in Multimedia Appendix 6 for frequencies of all coded categories). The most frequently mentioned reason for use in both 2012-2016 and 2017-2022 was pain or a pain-related condition, which included general pain (7/18, 38.9% of messages in 2012-2016; 49/162, 30.2% in 2017-2022), pain in a specific body part (2/18, 11.2% of messages in 2012-2016; 21/162, 13.1% in 2017-2022), migraine (3/18, 16.7% of messages in 2012-2016; 6/162, 3.7% in 2017-2022), and fibromyalgia (8/162, 4.9% of messages in 2017-2022). We coded 56 different purposes behind the mentions of cannabis in patient messages. From 2017-2022, messages about cannabis were largely driven by patients seeking assistance to facilitate medical marijuana use (57/162, 35.2%), which included patients requesting formal approval to use (ie, physical documentation or a diagnosis of a medical marijuana qualifying condition to be reflected in their EHR; 32/162, 19.8%), patients asking for help with access, such as a request to help with accessing a medical marijuana card (9/162, 5.6%), patients asking for a referral to a certifying provider (13/162, 8%), and patients asking for a medical marijuana prescription (6/162, 3.7%). Note that "prescription" is the language used by patients, but providers cannot prescribe medical marijuana (for a description of Pennsylvania's medical marijuana access process for patients; more details in Multimedia Appendix 7). From 2012-2016, a significantly higher proportion of messages were about cannabis screening results (7/18, 38.9%) and involved patients reporting no current cannabis use (5/18, 27.8%) compared with messages from 2017-2022 (9/162, 5.6% and 13/162, 8%; $P<.01$ and $P=.01$).

Fewer than 20% of provider messages from both 2012-2016 and 2017-2022 specified a reason for a patient's use of cannabis (Table 2). Providers indicated 10 different reasons for patient

XSL•FO
RenderX

use, 7 of which were named in patient messages. Like patient messages, pain or a pain-related condition was the most frequently mentioned reason for use (2/17, 11.8% of messages in 2012-2016; 8/80, 10% in 2017-2022), with general pain being the most common of these mentions (1/17, 5.9% of messages in 2012-2016; 7/80, 8.8% in 2017-2022). We coded 33 different purposes behind the mentions of cannabis in provider messages. The most common purpose from 2017-2022 was to assist patients in accessing medical marijuana (22/80, 27.5%). Providers also frequently stated they were unable to refer, prescribe, or recommend medical marijuana (21/80, 26.3%), citing health system policy and the limitations in the scope of their practice as primary justifications (8/80, 10%, and 7/80, 8.8%, respectively). From 2012-2016, a significantly higher proportion of messages from providers involved information about a cannabis screening result (13/17, 76.5%) compared with messages from 2017-2022 (7/80, 8.8%; $P<.01$). These messages from 2012-2016 included the provider informing the patient of a positive result or that they had violated a medication use agreement to not use cannabis (11/17, 64.7% and 7/17, 41.2%, respectively).

Table 2. Frequency of provider message content categories related to cannabis by year of message.

Content category	2012-2016 (N=20), n (%)	2017-2022 (N=80), n (%)	*P* value[a]
Correctly classified message			
Yes	17 (85)	80 (100)	__[b]
No, not related to cannabis	3 (15)	—	—
No, reference to Rx	—	—	—
Message recipient[c]			
Patient	17 (100)	75 (93.8)	—
Patient proxy	—	5 (6.3)	—
Reason for use[c,d]			
Not specified	14 (82.4)	65 (81.3)	.91
Pain or pain condition	2 (11.8)	8 (10)	.83
Anxiety	—	4 (5)	.35
Sleep	—	1 (1.3)	.65
Purpose of cannabis mention[c,d]			
Providing assistance with medical marijuana access	—	22 (27.5)	.01
Explanation of potential drug-to-drug interaction	2 (11.8)	12 (15)	.73
Explanation of other associated adverse events	—	8 (10)	.17
Unable to refer, prescribe, or recommend use	2 (11.8)	21 (26.3)	.20
Providing information on cannabis screening	13 (76.5)	7 (8.8)	<.001

[a]*P* values were calculated using two proportion *z* tests.

[b]Not available.

[c]Percentages are based on values of correctly classified messages (ie, n=18 for 2012-2016 messages, n=162 for 2017-2022 messages).

[d]Only includes a subset of coded categories.

Thematic Analysis

Our analysis of the frequency patterns in message content produced 6 themes. The themes reflect the primary takeaways of the coded data, with differences between message groups related to the legalization of medical marijuana being a key focus. Not all coded categories were used as a part of this analysis. Representative patient and provider messages are given for each theme. Note that select quotes are given for each theme, and those that contain inaccurate information were labeled with [sic].

1. Many patient reasons for using cannabis were expressed in messages, with mixed connections to Pennsylvania qualifying conditions. Between patient and provider messages, 37 distinct reasons for patients having used or wanting to use medical marijuana were named. This includes multiple variations of the same type of health issue. For example, 11 pain-related issues were named. Out of the 37 reasons, 19 can be linked to a Pennsylvania qualifying condition for medical marijuana, and only 9 of 27 reasons if pain-related issues are considered together (more details in Multimedia Appendix 8 for a list of medical marijuana qualifying conditions in Pennsylvania). While patients primarily mentioned reasons for medical marijuana use that could be considered qualifying conditions, such as pain and anxiety, there were a number of cases where patients discussed use for purposes that had no obvious connection to a qualifying condition. For example, sleep was a commonly cited reason for a patient's use or interest in use:

XSL•FO
RenderX

> *I wanted to talk about the possibility of getting a marijuana card. I've been having a hard time falling asleep and staying asleep. I've tried trazodone, but I don't like the "hung over" feeling the next day. How do I go about getting this? Thanks.*

Providers did not often indicate a patient's reason for use, and most of these few instances referenced pain or anxiety. When these reasons were mentioned, providers tended to endorse the use or relay the positive experiences of other patients. One provider wrote, "Medical marijuana can certainly be helpful to reduce chronic pain…" Another wrote, "Unfortunately I do not have expertise in marijuana. We have heard a lot of success with it for pain, anxiety control, etc"

2. After the legalization of medical cannabis, patients expressing interest in use was the primary driver of cannabis-related discussions. Nearly 50% of patient messages after 2016 expressed interest in using or explicit intent to use medical marijuana. Patients primarily expressed their interest in using medical marijuana through direct requests to a provider for approval to use:

> *The doctor I contacted regarding medical marijuana needs a diagnosis code for my anxiety and recommendation from a doctor before I can schedule an appointment with him. Is that something you can provide me?*

Patients also expressed their interest by seeking recommendations from a provider on whether to use:

> *... I have heard a lot lately about medical marijuana for various conditions and was wondering what your opinion on it is… I guess I'm wondering if it could be something that would be able to help my mental health issues, as well as my physical ones…*

Other patients were more forthcoming, explicitly stating their plans to use medical marijuana:

> *Just to inform you, I got my medical marijuana card. They are starting with a CBD tincture. I will be going to pick it up today or tomorrow.*

3. In response to patient interest in using medical marijuana, some providers tried to support this demand while others took a more cautious approach. In the face of receiving patient inquiries about accessing medical marijuana and requests for approval to use, many providers attempted to give some level of assistance. This assistance ranged from simply directing a patient to the state medical marijuana program website to giving a list of certifying providers to outright certifying patients in limited cases. Providers who assisted in this way were primarily from the family medicine department. A physician assistant from family medicine sent the following message:

> *Referrals are not needed to obtain a medical marijuana card - insurance will not pay for this type of treatment. You just need to schedule an intake with a doctor who has gone through the credentialing to evaluate you for this. If you are deemed an appropriate candidate, you then go to a pharmacy [sic] where the medical marijuana is dispensed. [Name of physician] is a local doctor who can do this. I copied and pasted his information below.*

At the same time, other providers took a more cautious approach in guiding patients around medical marijuana use. This included informing about potential drug interactions or other adverse outcomes and recommendations to not use medical marijuana. The distribution of departments that these messages were sent from was heterogeneous, including family medicine, internal medicine, psychiatry, psychology, rheumatology, neurology, cardiology, pharmacy, gastroenterology, and obstetrics and gynecology. For instance, a physician from the neurology department gave a warning to a patient:

> *In general, I do not recommend the use of medical marijuana for people with cognitive concerns given that it can have some brain impairing effects. CBD alone may or may not have brain impairing side effects and until that is clarified, I also don't in general recommend its use in those with cognitive concerns.*

4. Patients expressed a willingness to admit and discuss use after 2016, a contrast from previous years characterized by patients distancing themselves from use. From 2012-2016, more messages included statements about current cannabis nonuse than current use. Patients tended to be defensive and denied any cannabis use in response to a positive screening result:

> *I think you misunderstood! I was not lying, since I signed the agreement I stopped the pot because this pain is real and pot is not worth the pain, which is why I said it's an old habit and that it will not be in my urine anymore…*

The tenor and content of provider messages from 2012 to 2016 may help to explain these patient responses. Cannabis was primarily brought up by providers to inform patients of a positive screening result and that they had violated their medication use agreement. In 2013, a family medicine physician sent the following message:

> *Marijuana ingestion is not accidental and is in violation of your medication usage agreement. Simply put, this clinic will never be writing for any controlled substance ever again.*

However, more messages included statements about current use than current nonuse from 2017 to 2022. Patients were transparent about cannabis use, bringing it up on their own to provide more context around their situation. For instance, one patient was upfront about their use and why they use it: "I smoke marijuana due to my anxiety, and it helps me sleep…"

5. In the early years of legalized medical marijuana, some providers signaled that they were not in a position or were unwilling to engage in discussions about medical marijuana. Over 25% of provider messages from 2017-2022 included a statement that the provider was unable to refer, prescribe, or recommend medical marijuana use. Despite this lack of engagement from providers on the topic of medical marijuana, 52.5% (42/80) of these messages still attempted to support patients' requests. In the following message, a provider cites their own limited knowledge and the absence of a health system

XSL•FO
RenderX

medical marijuana policy to abstain from making a clear recommendation. It is important to note that Geisinger has had a medical marijuana policy since 2017 acknowledging the status of marijuana as a Schedule I substance from the federal Controlled Substance Act and, in accordance, does not advocate for medical marijuana use. However, it does allow for self-administration of legal products under medical supervision within facilities and allows providers to share information about medical marijuana and register with the Commonwealth to certify patients with medical conditions recognized by Pennsylvania. While this provider appears to be unaware of the policy, they still try to direct the patient toward the relevant resources:

> *I'm not sure if marijuana would help you with your symptoms, and I personally do not have any experience with medical marijuana. There is no medical marijuana policy in Geisinger [sic], but there is a government program (Pennsylvania's medical marijuana program). You can go to their website and find out about the locations and providers who prescribe [sic] medical marijuana… [URL]*

Discussion

Principal Findings

We identified 1782 messages from 1098 patients and 800 messages from 430 providers using a Word2Vec-constructed lexicon of cannabis key terms as explicitly discussing cannabis (both recreational and medical marijuana). A further content and thematic analysis revealed changes in the purposes behind discussions of cannabis between years before the legalization of medical marijuana in Pennsylvania and years after legalization. Before legalization, patients and providers primarily discussed cannabis in the context of a patient's positive toxicology screening result. In the years after legalization, patients sought access to medical marijuana and guidance around its use, while providers were split between supporting requests and cautioning patients on the risks of cannabis (ie, informing about potential drug interactions or other adverse outcomes and recommending not to use medical marijuana). Patients were also willing to disclose that they were currently using cannabis, often specifying the health-related issue they were using it for.

We found in our content analysis that just under half of patient messages from 2017 to 2022 expressed interest in using cannabis for a health-related reason, indicating that some patients are using the patient portal as a vehicle to engage providers about cannabis as a clinical option. While cannabis-related discussions over secure messaging have not been explored, previous studies have examined the general message content of other patient populations and topics, such as COVID-19 [27], veterans [28], and veterans with diabetes [26]. These studies found patients most frequently sent messages that were transactional (eg, appointment scheduling, administrative requests), followed by informational (eg, updates on health status or care) or interactional (eg, request for input on health-related issues) [26,28]. This is consistent with our own findings on the content of patient messages after 2016 if seeking assistance to facilitate medical marijuana use is considered transactional (eg, request for an appointment with or a referral to a certifying provider, request for a diagnosis code that reflects a qualifying condition), reporting current cannabis use or nonuse is considered informational, and seeking guidance on cannabis use is considered interactional. Our observations of patients' use of secure messaging suggest after legalization, some patients viewed medical marijuana similarly to more clinically established medications and treatment options. As a result, these patients turned to providers as a primary point of access as they would for other therapies. Several studies have reported sharp increases in cannabis use across various populations in the immediate aftermath of state legalization (medical or recreational) [37-39]. In Pennsylvania and other states that have recently legalized medical marijuana, it should be expected patients will express their interest in the use, and secure messaging may be used to express this interest.

In several states with medical marijuana programs, providers have expressed concern over the risks of patient use, particularly with respect to drug interactions [40-43]. This parallels our own finding of providers cautioning patients around use, including by informing them about potential drug interactions. Much of the concern from providers may be the result of uncertainty and an overall lack of knowledge on medical marijuana [40,41,43,44]. Despite this uncertainty, some providers believe marijuana has medical efficacy for certain conditions, such as intractable pain and terminal illness [41,43]. This may explain our finding that in over half of the messages where a provider claimed they could not refer, prescribe [sic], or recommend medical marijuana, they followed up with assistance for the patient to access medical marijuana. Patients in our study primarily named pain, anxiety, and sleep as reasons for using or wanting to use medical marijuana, and these are the most commonly reported reasons for use in patients across the United States [10,45]. Providers may be more comfortable supporting use for reasons about which they are frequently approached [43,46]. However, patients in our study brought up numerous other reasons for use that would not directly be considered a qualifying condition in Pennsylvania. Some of these reasons included cramps, periods, inflammation, skin irritation, mood swings, vertigo, tremors, and attention deficit disorder. Provider caution is warranted in the absence of an informed position, particularly for conditions that do not qualify for use under a state's medical marijuana program. However, some patients have reported frustration and distrust in providers who come across as overly negative or uneducated about medical marijuana [47,48]. Health systems should support providers with updated evidence-based guidelines on medical marijuana use for medical conditions [17], and providers must be prepared to communicate this information over mediums such as secure messaging to respond to patient inquiries with appropriate and accurate information.

Stigma exists as a part of the experience of medical marijuana users, which can lead to stress, anxiety, isolation, and avoidance of interacting with health care professionals [47,49-51]. Despite California legalizing medical marijuana in 1996, patient users have reported fears of being stigmatized more than a decade after legalization, including by their own physicians [50]. We found more patients openly admitted to using cannabis in

XSL•FO
RenderX

postlegalization years versus prelegalization years. However, many patients may choose not to discuss their cannabis use over secure messaging due to a variety of factors, including stigma. It is unknown whether secure messaging as a communication platform facilitates or hinders patient disclosure of sensitive information such as cannabis use. A recent investigation found among patients with hypertension, diabetes, or both, younger patients were less likely to share clinical updates using secure messaging than older patients, while men were less likely to seek medical guidance than women [52]. At the same time, a study from California found that the population of medical marijuana users in the state is composed of a higher proportion of males and skewed younger than the general population [45]. While there may be many patients who are not disclosing their cannabis use [51], the patient portal platform may uniquely select patients who are more willing to disclose sensitive information. Patients with access to a patient portal account are more likely to have a primary care clinician, health insurance, and higher educational attainment than patients without access [53]. Lower socioeconomic status has been associated with higher levels of patient distrust of physicians [54], so patients who are less willing to share information may not be as represented in a population of patient portal users. Still, in messages sent before legalization, we found cannabis was primarily discussed in the context of positive screening results when a patient was "caught" rather than when the patient volunteered information on their use themselves. A previous investigation found evidence of patients in Pennsylvania withholding cannabis use information before legalization [55]. This shift in tenor prelegalization versus postlegalization may signal a change in patient attitudes toward sharing information related to cannabis use. Future studies should examine changes in cannabis use disclosure over time in states that recently legalized cannabis (for medical or recreational use), as well as the potential role secure messaging plays in facilitating disclosure.

Limitations

Our study is the first to analyze the content of secure messages to understand cannabis-related discussions between patients and providers, but it is limited in several ways. First, of the messages we identified as explicitly discussing cannabis, relatively few occurred during the prelegalization period, likely because of the illegality of cannabis for any reason in the state during this period, as well as a lower overall message volume compared to recent years. The low message volume limited the concepts we were able to derive from our content analysis. However, among the messages we did analyze, a considerable proportion (7/18, 38.9% for patients and 13/17, 76.5% for providers) were related to discussions of cannabis screening results and the consequences of those results. This may be a common pattern in the message content for discussions of any illicit substance. Future work is needed to compare the volume and content of secure messages discussing cannabis versus other illicit substances, particularly as states continue to enact medical and recreational marijuana policies as well as after cannabis becomes a Schedule III substance. Second, our data includes messages sent a year before and a year after a patient in the study population had been flagged as a cannabis user in their EHR. It is possible that the patient message content we observed was influenced by the fact these patients had been flagged at some point as cannabis users. For example, patients who send cannabis-related messages after being flagged may be more open to discussing their use, and patients who express interest in cannabis over secure messages may be easier to flag when they eventually start using. Third, we focused our analysis at the message level, as our objective was to identify and contextualize explicit mentions of cannabis in individual messages. However, by excluding replies in our analysis, there may be messages in which we left out the full context behind the purpose of the cannabis mention. Excluding replies without an explicit mention of cannabis also means we are likely underestimating the number of "true" messages that discuss cannabis. However, we found that messages often included multiple topics and that replies may not address cannabis at all, even if the initial message explicitly mentioned cannabis. Future studies can expand on this work by analyzing entire message threads related to cannabis to understand the dynamics of patient-provider conversations on this topic, including how specific patient inquiries and disclosures correlate with specific provider responses. Finally, medical marijuana is currently legal in Pennsylvania, but the recreational use of cannabis is not as of this writing (December 2024). Further research is necessary to determine the generalizability of these findings to states with legal recreational marijuana or to different patient populations (ie, English as a second language patients or minors).

Conclusions

We identified patient and provider secure messages at an integrated health delivery system discussing cannabis use and performed a content analysis to further contextualize these messages around the legalization of medical marijuana. After legalization in Pennsylvania, patients expressed interest in medical marijuana use and a willingness to disclose information related to their use. Providers were split in their responses to patient inquiries, with some aiding with access to medical marijuana while others emphasized its risks to patients. Secure messaging through the patient portal offers a channel through which providers can engage patients about cannabis use, but health systems can support providers in communicating consistent and appropriate guidance to prevent harmful use, particularly as patient portal adoption grows and providers experience a greater burden to respond to patient messages. Health systems should also continue to explore ways in which secure messaging can be used to facilitate patient sharing of cannabis use information to optimize their health care.

Acknowledgments

Geisinger is the recipient of research support from Ascend Wellness Holdings, an approved medical marijuana clinical registrant in Pennsylvania. The funder was not involved in the study or the decision to submit for publication.

XSL•FO
RenderX

Authors' Contributions

VAS made substantial contributions to the study's conception and design, analysis and interpretation of the results, and manuscript drafting. CMG and LDT provided substantial contributions to the analysis and interpretation of the results. AMP made substantial contributions to the acquisition of the data and the interpretation of the results. KMR gave substantial guidance on the study's design and interpretation of the results. BJP and EAW made substantial contributions to the study conception and interpretation of the results. All authors reviewed the manuscript critically for important intellectual content and exercised final approval of the version to be published.

Conflicts of Interest

BJP was (2019-21) part of an osteoarthritis research team supported by Pfizer and Eli Lilly. The other authors have no disclosures.

Multimedia Appendix 1
Algorithms for identifying discrete indications of cannabis use in electronic health records.
[DOCX File , 39 KB - jmir_v26i1e63311_app1.docx]

Multimedia Appendix 2
Word2vec word similarity lists (with cosine similarity scores) for cannabis lexicon.
[DOCX File , 25 KB - jmir_v26i1e63311_app2.docx]

Multimedia Appendix 3
Final patient and provider cannabis codebooks.
[DOCX File , 34 KB - jmir_v26i1e63311_app3.docx]

Multimedia Appendix 4
Volume of secure messages that include a mention of cannabis from 2012-2021.
[PNG File , 26 KB - jmir_v26i1e63311_app4.png]

Multimedia Appendix 5
Volume of messages among all patients in study population and responding providers from 2012-2021.
[PNG File , 26 KB - jmir_v26i1e63311_app5.png]

Multimedia Appendix 6
Frequencies for all patient and provider code categories.
[DOCX File , 32 KB - jmir_v26i1e63311_app6.docx]

Multimedia Appendix 7
Medical marijuana access process in Pennsylvania.
[DOCX File , 17 KB - jmir_v26i1e63311_app7.docx]

Multimedia Appendix 8
Pennsylvania medical marijuana qualifying conditions from 2016-2024.
[DOCX File , 19 KB - jmir_v26i1e63311_app8.docx]

References

1. Mayo Clinic. Medical marijuana. 2021. URL: https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/medical-marijuana/art-20137855 [accessed 2024-12-25]
2. National Conference of State Legislatures. State medical cannabis laws. URL: https://www.ncsl.org/health/state-medical-cannabis-laws [accessed 2024-09-30]
3. Medical Marijuana Act - Enactment. Pennsylvania General Assembly. 2016. URL: https://www.pa.gov/agencies/ethics/laws-acts-and-regulations/medical-marijuana-act.html [accessed 2024-03-25]
4. Pennsylvania Medical Marijuana Program. Pennsylvania Department of Health. URL: https://www.health.pa.gov/topics/programs/Medical%20Marijuana/Pages/Medical%20Marijuana.aspx [accessed 2024-03-25]
5. Kosiba JD, Maisto SA, Ditre JW. Patient-reported use of medical cannabis for pain, anxiety, and depression symptoms: Systematic review and meta-analysis. Soc Sci Med 2019 Jul;233:181-192. [doi: 10.1016/j.socscimed.2019.06.005] [Medline: 31207470]

XSL•FO
RenderX

6. Jennings JM, Johnson RM, Brady AC, Dennis DA. Patient perception regarding potential effectiveness of cannabis for pain management. J Arthroplasty 2020 Dec;35(12):3524-3527. [doi: 10.1016/j.arth.2020.06.051] [Medline: 32684396]
7. Salazar CA, Tomko RL, Akbar SA, Squeglia LM, McClure EA. Medical cannabis use among adults in the Southeastern United States. Cannabis 2019 Feb 09;2(1):53-65 [FREE Full text] [Medline: 30815638]
8. Ryan-Ibarra S, Induni M, Ewing D. Prevalence of medical marijuana use in California, 2012. Drug Alcohol Rev 2015 Mar;34(2):141-146. [doi: 10.1111/dar.12207] [Medline: 25255903]
9. Boehnke KF, Scott JR, Litinas E, Sisley S, Williams DA, Clauw DJ. Pills to pot: observational analyses of cannabis substitution among medical cannabis users with chronic pain. J Pain 2019 Jul;20(7):830-841 [FREE Full text] [doi: 10.1016/j.jpain.2019.01.010] [Medline: 30690169]
10. Azcarate PM, Zhang AJ, Keyhani S, Steigerwald S, Ishida JH, Cohen BE. Medical reasons for marijuana use, forms of use, and patient perception of physician attitudes among the US population. J Gen Intern Med 2020 Jul;35(7):1979-1986 [FREE Full text] [doi: 10.1007/s11606-020-05800-7] [Medline: 32291715]
11. Leung J, Chan GCK, Hides L, Hall WD. What is the prevalence and risk of cannabis use disorders among people who use cannabis? a systematic review and meta-analysis. Addict Behav 2020 Oct;109:106479. [doi: 10.1016/j.addbeh.2020.106479] [Medline: 32485547]
12. Karila L, Roux P, Rolland B, Benyamina A, Reynaud M, Aubin H, et al. Acute and long-term effects of cannabis use: a review. Curr Pharm Des 2014;20(25):4112-4118. [doi: 10.2174/13816128113199990620] [Medline: 24001294]
13. Volkow ND, Baler RD, Compton WM, Weiss SR. Adverse health effects of marijuana use. N Engl J Med 2014 Jun 05;370(23):2219-2227. [doi: 10.1056/nejmra1402309]
14. Hall W. What has research over the past two decades revealed about the adverse health effects of recreational cannabis use? Addiction 2015 Jan;110(1):19-35. [doi: 10.1111/add.12703] [Medline: 25287883]
15. Hill KP. Cannabis use and risk for substance use disorders and mood or anxiety disorders. JAMA 2017 Mar 14;317(10):1070-1071. [doi: 10.1001/jama.2016.19706] [Medline: 28291877]
16. Levinsohn EA, Hill KP. Clinical uses of cannabis and cannabinoids in the United States. J Neurol Sci 2020 Apr 15;411:116717. [doi: 10.1016/j.jns.2020.116717] [Medline: 32044684]
17. Abrams DI. The therapeutic effects of cannabis and cannabinoids: An update from the National Academies of Sciences, Engineering and Medicine report. Eur J Intern Med 2018 Mar;49:7-11. [doi: 10.1016/j.ejim.2018.01.003] [Medline: 29325791]
18. Whiting PF, Wolff RF, Deshpande S, Di Nisio M, Duffy S, Hernandez AV, et al. Cannabinoids for medical use: A systematic review and meta-analysis. JAMA 2015;313(24):2456-2473. [doi: 10.1001/jama.2015.6358] [Medline: 26103030]
19. World Health Organization. Cannabis. Alcohol, Drugs and Addictive Behaviours. URL: https://www.who.int/teams/mental-health-and-substance-use/alcohol-drugs-and-addictive-behaviours/drugs-psychoactive/cannabis [accessed 2024-03-25]
20. Lapham GT, Matson TE, Carrell DS, Bobb JF, Luce C, Oliver MM, et al. Comparison of medical cannabis use reported on a confidential survey vs documented in the electronic health record among primary care patients. JAMA Netw Open 2022 May 02;5(5):e2211677 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.11677] [Medline: 35604691]
21. Matson TE, Carrell DS, Bobb JF, Cronkite DJ, Oliver MM, Luce C, et al. Prevalence of medical cannabis use and associated health conditions documented in electronic health records among primary care patients in Washington state. JAMA Netw Open 2021 May 03;4(5):e219375 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.9375] [Medline: 33956129]
22. Sajdeya R, Mardini MT, Tighe PJ, Ison RL, Bai C, Jugl S, et al. Developing and validating a natural language processing algorithm to extract preoperative cannabis use status documentation from unstructured narrative clinical notes. J Am Med Inform Assoc 2023 Jul 19;30(8):1418-1428 [FREE Full text] [doi: 10.1093/jamia/ocad080] [Medline: 37178155]
23. Heath S. Patient Engagement HIT. Patient Portal Adoption Tops 90%, But Strong Patient Use is Needed. URL: https://patientengagementhit.com/news/patient-portal-adoption-tops-90-but-strong-patient-use-is-needed [accessed 2024-01-26]
24. Schillinger D, McNamara D, Crossley S, Lyles C, Moffet HH, Sarkar U, et al. The next frontier in communication and the ECLIPPSE study: Bridging the linguistic divide in secure messaging. J Diabetes Res 2017;2017:1348242 [FREE Full text] [doi: 10.1155/2017/1348242] [Medline: 28265579]
25. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
26. Robinson SA, Zocchi M, Purington C, Am L, DeLaughter K, Vimalananda VG, et al. Secure messaging for diabetes management: content analysis. JMIR Diabetes 2023 Mar 23;8:e40272 [FREE Full text] [doi: 10.2196/40272] [Medline: 36951903]
27. Alpert JM, Campbell-Salome G, Gao C, Markham MJ, Murphy M, Harle CA, et al. Secure messaging and COVID-19: a content analysis of patient-clinician communication during the pandemic. Telemed J E Health 2022 Jul;28(7):1028-1034 [FREE Full text] [doi: 10.1089/tmj.2021.0316] [Medline: 34767741]
28. Shimada SL, Petrakis BA, Rothendler JA, Zirkle M, Zhao S, Feng H, et al. An analysis of patient-provider secure messaging at two Veterans Health Administration medical centers: message content and resolution through secure messaging. J Am Med Inform Assoc 2017 Sep 01;24(5):942-949 [FREE Full text] [doi: 10.1093/jamia/ocx021] [Medline: 28371896]
29. Geisinger. About Geisinger. URL: https://www.geisinger.org/about-geisinger [accessed 2024-03-25]

XSL•FO
RenderX

30. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. URL: http://arxiv.org/abs/1301.3781 [accessed 2013-01-16]
31. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. 2010 Presented at: Proceedings of LREC 2010 Workshop; 2010; Valetta, MT: University of Malta; May 22, 2010; Valetta URL: https://www.researchgate.net/publication/255820377_Software_Framework_for_Topic_Modelling_with_Large_Corpora
32. Taylor SJ, Bogdan R, DeVault ML. Introduction to Qualitative Research Methods: A Guidebook and Resource. Hoboken: John Wiley & Sons, Inc; 2015.
33. Pope C, Ziebland S, Mays N. In: Pope C, Mays N, editors. Analysing Qualitative Data. United States: Blackwell Publishing Ltd; 2006.
34. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960 Apr 01;20(1):37-46. [doi: 10.1177/001316446002000104]
35. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363. [Medline: 15883903]
36. Scolforo M. AP News. Pennsylvania gives approval to first medical pot dispensary. URL: https://apnews.com/general-news-ae795803544d411ea954b6127fbba054 [accessed 2024-01-26]
37. Fedorova EV, Ataiants J, Wong CF, Iverson E, Lankenau SE. Changes in medical cannabis patient status before and after cannabis legalization in California: associations with cannabis and other drug use. J Psychoactive Drugs 2022;54(2):129-139 [FREE Full text] [doi: 10.1080/02791072.2021.1926604] [Medline: 34044753]
38. Levine M, Jontz A, Dabrowski P, Claudius IA, Kreisler R, Yee N, et al. Prevalence of marijuana use among trauma patients before and after legalization of medical marijuana: The Arizona experience. Subst Abus 2021;42(3):366-371. [doi: 10.1080/08897077.2020.1784359] [Medline: 32693706]
39. Paschall MJ, Grube JW, Biglan A. Medical marijuana legalization and marijuana use among youth in Oregon. J Prim Prev 2017 Jun;38(3):329-341 [FREE Full text] [doi: 10.1007/s10935-017-0476-5] [Medline: 28484894]
40. Kondrad E, Reid A. Colorado family physicians' attitudes toward medical marijuana. J Am Board Fam Med 2013;26(1):52-60 [FREE Full text] [doi: 10.3122/jabfm.2013.01.120089] [Medline: 23288281]
41. Szyliowicz D, Hilsenrath P. Medical marijuana knowledge and attitudes: a survey of the California Pharmacists Association. J Prim Care Community Health 2019;10:2150132719831871 [FREE Full text] [doi: 10.1177/2150132719831871] [Medline: 30795717]
42. Lombardi E, Gunter J, Tanner E. Ohio physician attitudes toward medical Cannabis and Ohio's medical marijuana program. J Cannabis Res 2020 Apr 21;2(1):16 [FREE Full text] [doi: 10.1186/s42238-020-00025-1] [Medline: 33526115]
43. Philpot LM, Ebbert JO, Hurt RT. A survey of the attitudes, beliefs and knowledge about medical cannabis among primary care providers. BMC Fam Pract 2019 Jan 22;20(1):17 [FREE Full text] [doi: 10.1186/s12875-019-0906-y] [Medline: 30669979]
44. Worster B, Ashare RL, Hajjar E, Garber G, Smith K, Kelly EL. Clinician attitudes, training, and beliefs about cannabis: an interprofessional assessment. Cannabis Cannabinoid Res 2023 Jun;8(3):547-556. [doi: 10.1089/can.2021.0022] [Medline: 34978882]
45. Reinarman C, Nunberg H, Lanthier F, Heddleston T. Who are medical marijuana patients? Population characteristics from nine California assessment clinics. J Psychoactive Drugs 2011;43(2):128-135. [doi: 10.1080/02791072.2011.587700] [Medline: 21858958]
46. Sideris A, Khan F, Boltunova A, Cuff G, Gharibo C, Doan LV. New York physicians' perspectives and knowledge of the State Medical Marijuana Program. Cannabis Cannabinoid Res 2018;3(1):74-84 [FREE Full text] [doi: 10.1089/can.2017.0046] [Medline: 29662957]
47. Ryan J, Sharts-Hopko N. The experiences of medical marijuana patients: a scoping review of the qualitative literature. J Neurosci Nurs 2017 Jun;49(3):185-190. [doi: 10.1097/JNN.0000000000000283] [Medline: 28471927]
48. Pedersen W, Sandberg S. The medicalisation of revolt: a sociological analysis of medical cannabis users. Sociol Health Illn 2013 Jan;35(1):17-32 [FREE Full text] [doi: 10.1111/j.1467-9566.2012.01476.x] [Medline: 22827932]
49. Bottorff JL, Bissell LJL, Balneaves LG, Oliffe JL, Capler NR, Buxton J. Perceptions of cannabis as a stigmatized medicine: a qualitative descriptive study. Harm Reduct J 2013 Feb 16;10:2 [FREE Full text] [doi: 10.1186/1477-7517-10-2] [Medline: 23414118]
50. Satterlund TD, Lee JP, Moore RS. Stigma among California's medical marijuana patients. J Psychoactive Drugs 2015;47(1):10-17 [FREE Full text] [doi: 10.1080/02791072.2014.991858] [Medline: 25715067]
51. Piper BJ, DeKeuster RM, Beals ML, Cobb CM, Burchman CA, Perkinson L, et al. Substitution of medical cannabis for pharmaceutical agents for pain, anxiety, and sleep. J Psychopharmacol 2017 May;31(5):569-575. [doi: 10.1177/0269881117699616] [Medline: 28372506]
52. Heisey-Grove D, Rathert C, McClelland LE, Jackson K, DeShazo JP. Patient and clinician characteristics associated with secure message content: retrospective cohort study. J Med Internet Res 2021 Aug 19;23(8):e26650 [FREE Full text] [doi: 10.2196/26650] [Medline: 34420923]

XSL•FO
RenderX

53. El-Toukhy S, Méndez A, Collins S, Pérez-Stable EJ. Barriers to patient portal access and use: evidence from the Health Information National Trends Survey. J Am Board Fam Med 2020;33(6):953-968 [FREE Full text] [doi: 10.3122/jabfm.2020.06.190402] [Medline: 33219074]
54. Armstrong K, Ravenell KL, McMurphy S, Putt M. Racial/ethnic differences in physician distrust in the United States. Am J Public Health 2007 Jul;97(7):1283-1289. [doi: 10.2105/AJPH.2005.080762] [Medline: 17538069]
55. Chang JC, Holland CL, Tarr JA, Rubio D, Rodriguez KL, Kraemer KL, et al. Perinatal illicit drug and marijuana use. Am J Health Promot 2017 Jan;31(1):35-42 [FREE Full text] [doi: 10.4278/ajhp.141215-QUAL-625] [Medline: 26559718]

Abbreviations

EHR: electronic health record
GHP: Geisinger Health Plan

Edited by A Mavragani; submitted 16.06.24; peer-reviewed by J Alpert, L Yang; comments to author 20.09.24; revised version received 07.10.24; accepted 29.10.24; published 12.12.24.

Please cite as:
Shetty VA, Gregor CM, Tusing LD, Pradhan AM, Romagnoli KM, Piper BJ, Wright EA
Discussions of Cannabis Over Patient Portal Secure Messaging: Content Analysis
J Med Internet Res 2024;26:e63311
URL: https://www.jmir.org/2024/1/e63311
doi:10.2196/63311
PMID:39666375

©Vishal A Shetty, Christina M Gregor, Lorraine D Tusing, Apoorva M Pradhan, Katrina M Romagnoli, Brian J Piper, Eric A Wright. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study

Anne Herrmann-Werner[1,2], MME, Prof Dr Med; Teresa Festl-Wietek[1], MSc, Dr Rer Nat; Friederike Holderried[1,3], MME, Dr Med; Lea Herschbach[1], MSc; Jan Griewatz[1], MA; Ken Masters[4], Prof Dr; Stephan Zipfel[2], Prof Dr Med; Moritz Mahling[1,5], Dr Med, MHBA

[1]Tübingen Institute for Medical Education, Faculty of Medicine, University of Tübingen, Tübingen, Germany
[2]Department of Psychosomatic Medicine and Psychotherapy, University Hospital Tübingen, Tübingen, Germany
[3]University Department of Anesthesiology and Intensive Care Medicine, University Hospital Tübingen, Tübingen, Germany
[4]Medical Education and Informatics Department, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Oman
[5]Department of Diabetology, Endocrinology, Nephrology, Section of Nephrology and Hypertension, University Hospital Tübingen, Tübingen, Germany

Corresponding Author:
Teresa Festl-Wietek, MSc, Dr Rer Nat
Tübingen Institute for Medical Education
Faculty of Medicine
University of Tübingen
Elfriede-Aulhorn-Strasse 10
Tübingen, 72076 Tübingen
Germany
Phone: 49 7071 29 73715
Email: teresa.festl-wietek@med.uni-tuebingen.de

Abstract

Background: Large language models such as GPT-4 (Generative Pre-trained Transformer 4) are being increasingly used in medicine and medical education. However, these models are prone to "hallucinations" (ie, outputs that seem convincing while being factually incorrect). It is currently unknown how these errors by large language models relate to the different cognitive levels defined in Bloom's taxonomy.

Objective: This study aims to explore how GPT-4 performs in terms of Bloom's taxonomy using psychosomatic medicine exam questions.

Methods: We used a large data set of psychosomatic medicine multiple-choice questions (N=307) with real-world results derived from medical school exams. GPT-4 answered the multiple-choice questions using 2 distinct prompt versions: detailed and short. The answers were analyzed using a quantitative approach and a qualitative approach. Focusing on incorrectly answered questions, we categorized reasoning errors according to the hierarchical framework of Bloom's taxonomy.

Results: GPT-4's performance in answering exam questions yielded a high success rate: 93% (284/307) for the detailed prompt and 91% (278/307) for the short prompt. Questions answered correctly by GPT-4 had a statistically significant higher difficulty than questions answered incorrectly (P=.002 for the detailed prompt and P<.001 for the short prompt). Independent of the prompt, GPT-4's lowest exam performance was 78.9% (15/19), thereby always surpassing the "pass" threshold. Our qualitative analysis of incorrect answers, based on Bloom's taxonomy, showed that errors were primarily in the "remember" (29/68) and "understand" (23/68) cognitive levels; specific issues arose in recalling details, understanding conceptual relationships, and adhering to standardized guidelines.

Conclusions: GPT-4 demonstrated a remarkable success rate when confronted with psychosomatic medicine multiple-choice exam questions, aligning with previous findings. When evaluated through Bloom's taxonomy, our data revealed that GPT-4 occasionally ignored specific facts (remember), provided illogical reasoning (understand), or failed to apply concepts to a new situation (apply). These errors, which were confidently presented, could be attributed to inherent model biases and the tendency to generate outputs that maximize likelihood.

(J Med Internet Res 2024;26:e52113) doi:10.2196/52113

XSL•FO
RenderX

KEYWORDS

answer; artificial intelligence; assessment; Bloom's taxonomy; ChatGPT; classification; error; exam; examination; generative; GPT-4; Generative Pre-trained Transformer 4; language model; learning outcome; LLM; MCQ; medical education; medical exam; multiple-choice question; natural language processing; NLP; psychosomatic; question; response; taxonomy

Introduction

The recent developments in artificial intelligence (AI) have transformative potential for various fields, including medicine [1] and medical education [2]. In November 2022, OpenAI launched GPT-3 (Generative Pre-trained Transformer 3), a large language model (LLM) [3]. Its high-quality performance surprised even experts and generated huge public interest (particularly in school and higher education settings, where GPT-3 prompted manifold discussions on its potential benefits and harms) [4,5].

In medical education, LLMs have the potential to revolutionize current teaching approaches and thus ultimately improve physician performance and health care outcomes. However, before LLMs are thoroughly integrated into medical education, their performance in this context must be comprehensively assessed. It is especially important to evaluate the capabilities of AI and LLMs within educational theoretical frameworks.

One of the most-used frameworks in medical education is Bloom's taxonomy [6,7] of learning outcomes, first introduced in 1956. Briefly, Bloom and subsequent colleagues developed a hierarchical classification of cognitive processes, ordered from lower-order cognitive skills—such as knowledge recall (remember) and comprehension (understand)—to higher-order thinking—such as application (apply), analysis (analyze), evaluation (evaluate), and creation (create) [8].

Since its first publication in 1956, this taxonomy has been used as a common language for educational instructors and still influences the field of medical education [8]. With his work, Bloom provided a significant contribution to what is now known as outcome-based education [9] and laid the foundation for other educational theories, such as Miller's pyramid of clinical competencies [10,11]. While Bloom's taxonomy is widely used and offers a structured approach to learning outcomes, some educators believe that its hierarchical nature might not always represent the complexity of learning [12].

Although derived from human learning processes, Bloom's taxonomy provides an ideal framework to describe the cognitive processes that underlie success and failure. Recently, LLMs have been assessed for their (surprisingly mostly good) performance in various fields of medicine, ranging from specific subjects to board exams [13-15]. However, the errors made by LLMs have not been evaluated in detail. For example, while LLMs might successfully recall facts (remember), they might struggle to apply those facts to a different context, or vice versa. We acknowledge that applying terms such as "remember" and "struggle" are anthropomorphisms used for ease of reading, as an LLM currently does neither and merely generates responses based on language-usage statistical probabilities using a "next-word prediction paradigm" [16].

Therefore, we aimed to use Bloom's taxonomy to gain a better understanding of the failures of LLMs. For human medicine education and the aforementioned use cases for LLMs, multiple-choice questions (MCQs) remain a primary written exam form and are used for summative and formative assessments [17]. In Bloom's taxonomy, MCQs are often used to assess lower-order cognitive skills, such as knowledge recall (remember) and comprehension (understand), but they may also probe higher-order thinking, such as application (apply), analysis (analyze), and evaluation (evaluate) [18]. Thus, MCQs offer a suitable lens for evaluating different cognitive processes.

A medical field that relies heavily on language and factual understanding is instrumental to elucidating cognitive processes and correct or incorrect reasoning. Given its interplay of psychological, social, and biological factors, psychosomatic medicine offers such a case. The field's heavy reliance on verbal and written communication for diagnosis and treatment makes it particularly challenging. Additionally, the combination of strict diagnostic criteria with a nuanced understanding of the patient's language makes it an ideal testing ground for the capabilities of language models.

We present a mixed methods study designed to explore how GPT-4 performs in terms of Bloom's taxonomy. First, we assessed the performance of GPT-4 with a large set of psychosomatic medicine exam questions and compared the results to responses from a cohort of medical students, thereby providing human comparison and quality indicators. For a deeper understanding of the results, we used qualitative methods to comprehend the model's performance and to assess the strengths and weaknesses of LLMs in relation to Bloom's taxonomy. The findings of this study provide critical insights into the practical applications and limitations of LLMs (such as GPT-4) in medical education.

Methods

Exams

A total of 16 examinations from winter term 2014-2015 to summer term 2022 were retrieved from the integrated management system of the Department of Psychosomatic Medicine and Psychotherapy faculty's web-based exam program (Figure 1 provides a graphical illustration of our methodological approach). In addition to question stems, answers, and distractors, the system also offers quality criteria for each individual question.

XSL•FO
RenderX

Figure 1. Illustration of our methodological approach. API: application programming interface; GPT-4: Generative Pre-trained Transformer 4; MCQ: multiple-choice question.

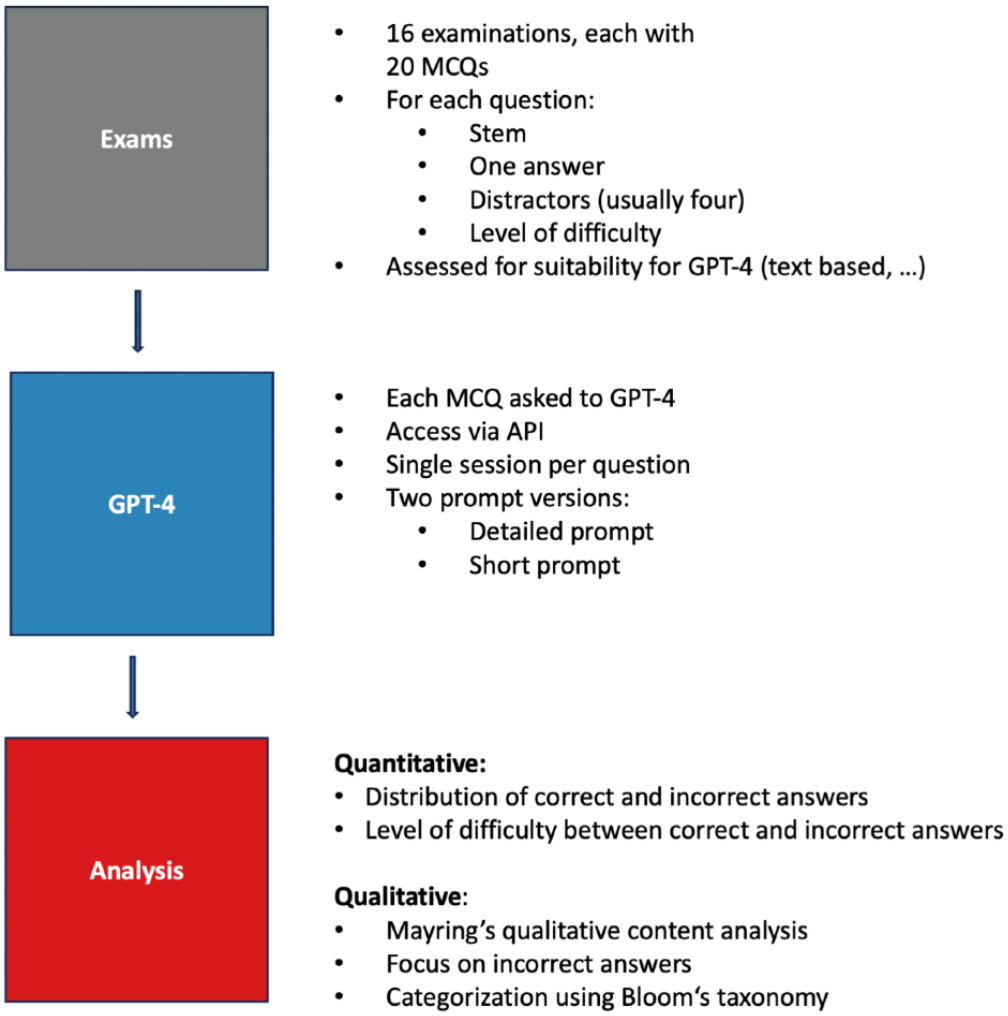

Each examination consists of 20 MCQs, with each MCQ having 1 answer and 4 distractors. Diagnostic and therapeutic questions cover topics concerning anxiety disorders, depression, eating, and somatoform and trauma disorders. The examinations also include questions concerning motivational interviewing techniques.

To compare the performance of GPT-4 to the student's performance, we retrieved the level of difficulty from the system, calculated as the average score attained in the respective question. Undergraduate medical students take the examination containing these questions at the end of their third year, after having attended 7.5 hours of lectures and 18 hours of practical classes on psychosomatic medicine and psychotherapy. To pass, a student is usually required to answer 12 of 20 questions correctly; an adjustment of this passing score is possible if, for example, a question has a too-poor performance.

We assessed the questions for their suitability to be analyzed by GPT-4. From a total of 320 questions, 13 (4%) were excluded, including questions that were not single-choice answers (n=6), ambiguous questions (n=3), questions featuring a graphic that had to be analyzed (n=2), and questions that covered a case represented in multiple questions (n=2).

We used GPT-4 to answer every question (model "gpt-4," OpenAI LP). For each question, we generated a detailed prompt version and a short prompt version. The prompts were created by the authors in an iterative process using the web interface ChatGPT Plus to achieve ideal performance. The most relevant difference between the versions was that the detailed prompt included a command to critically reflect on the answer and justify the choice made. Table 1 provides an example of a detailed prompt. We used the application programming interface (API) provided by OpenAI to post the questions to GPT-4 and retrieve the answers (access dates were March 21 and March 22, 2023). Each question and answer was posted in its own session. All interactions with GPT-4 were conducted in German, the original language of the examination questions; for the purposes of this paper, the questions were manually translated into English.

XSL•FO
RenderX

Table 1. Examples of detailed and short prompt versions.

Prompt version	Example
Detailed prompt	You have all the knowledge of psychosomatic medicine and have to answer an exam question. Please elaborate on the following multiple-choice question. Only one of the five answer choices is correct. Please consider carefully and choose one answer. Give a detailed reason for your answer. At the end of the reasoning, please add the letter of the answer you chose with the following notation: !A! (that is, A between two exclamation marks if you chose answer A). The question reads: A 49-year-old teacher is undergoing rehabilitation treatment after anterior wall infarction. It has been observed that he skips therapy appointments. The wife is concerned because he shows no interest in her visits. When asked, he states that he sees no point in further treatment because he has little hope that his condition will improve. However, none of that really matters anymore. He is also not afraid of another heart attack; on the contrary, then his misery would come to an end. Which suspected diagnosis in the psychosomatic field are you most likely to make? Answer A: Unspecific somatisation disorder; Answer B: Post-traumatic stress disorder; Answer C: Pseudodementia; Answer D: Generalised anxiety disorder; Answer E: Depression.
Short prompt	Please act like a specialist in psychosomatic medicine. Answer the following multiple-choice question and briefly explain your answer: A 49-year-old teacher is undergoing rehabilitation treatment after anterior wall infarction. It has been observed that he skips therapy appointments. The wife is concerned because he shows no interest in her visits. When asked, he states that he sees no point in further treatment because he has little hope that his condition will improve. However, none of that really matters anymore. He is also not afraid of another heart attack; on the contrary, then his misery would come to an end. Which suspected diagnosis in the psychosomatic field are you most likely to make? Answer A: Unspecific somatisation disorder; Answer B: Post-traumatic stress disorder; Answer C: Pseudodementia; Answer D: Generalised anxiety disorder; Answer E: Depression.

Data Analysis

The responses given by GPT-4 were compared to the answers indicated by the answer index (eg, “A” or “C”) and stored in Excel (version 16.0.10394.20022; Microsoft Corporation).

Quantitative Data Analysis

Quantitative analyses and figure generation were performed using R (R version 4.3.1; R Core Team) statistical software [19]. Briefly, we combined all tables with relevant data—that is, answers from GPT-4 and the aggregated data of the students’ exams (such as item difficulty). For each prompt version, we analyzed the ratio of correctly answered questions versus incorrectly answered questions. We further compared the question difficulty (taken from the aggregated student data) across questions answered correctly and incorrectly by GPT-4. The difficulty of a question is operationalized as the proportion of students answering a question correctly, with 0 representing a very difficult question and 1 a very easy question [20]. A Wilcoxon rank sum test was used to test for statistical significance. A level of $P<.05$ was considered statistically significant. If not stated otherwise, the results are given in medians and IQRs.

Qualitative Data Analysis

A total of 2 authors (TFW and FH) separately coded each text response. The answers from GPT-4 were analyzed inductively and iteratively according to Mayring’s [21] qualitative content analysis, as described previously by our group [22]. The goal of the analysis was defined in line with the answers to the examination questions. For the main category, we used the correct or incorrect answer to the question, then further focused primarily on incorrect answers.

In the answer texts, individual reasoning was categorized according to Bloom’s taxonomy as revised by Krathwohl [8]. Briefly, we used the following definitions of the cognitive domains for our rating procedure:

1. Remember: retrieving relevant knowledge from long-term memory.
2. Understand: determining the meaning of instructional messages, including oral, written, and graphic communication.
3. Apply: carrying out or using a procedure in a given situation.
4. Analyze: breaking down material into its constituent parts and detecting how the parts relate to 1 another and to an overall structure or purpose.
5. Evaluate: making judgments based on criteria and standards.

In the second step, each of the raters coded the answers using MAXQDA (version 12.3.2; VERBI software). To obtain the same level of abstraction when building the categories, the raters revised the codes together and agreed on the final categories, paraphrasing representative examples and building a hierarchy of categories based on the found codes. Subsequently, both initial raters independently worked through the material again. Each rater individually analyzed the answers given by GPT-4 and built codes using MAXQDA, including the main classification (correct or wrong answer), followed by the category of Bloom’s taxonomy and an example. When they could not agree on a category, 2 other experts were consulted in order to reach a consensus. When GPT-4’s responses were wrong, the explanation was analyzed using the levels of Bloom’s taxonomy (remember, understand, apply, analyze, evaluate, and create) [8].

Ethical Considerations

The Ethics Committee of the Faculty of Medicine at University Hospital Tübingen approved the study (number 076/2023A). All data were kept anonymous and were not associated with individual students.

XSL•FO
RenderX

Results

Quantitative Results

Distribution of Correctly and Incorrectly Answered Questions

For the detailed prompt, GPT-4 answered 92.5% (284/307) of the questions correctly; for the short prompt, the success rate was 90.6% (278/307). The distribution is shown in Figure 2.

Figure 2. Distribution of correctly and incorrectly answered questions by prompt version.

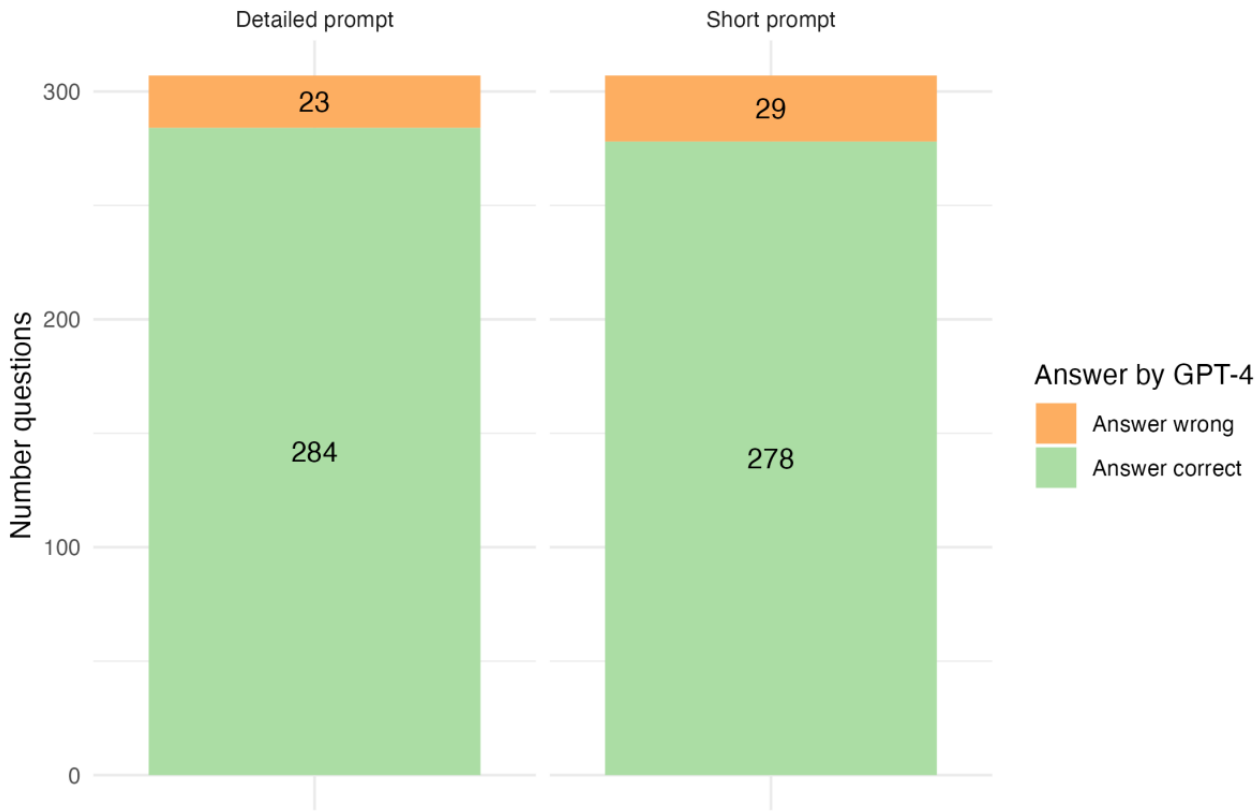

Question Difficulty

Across all questions, the median difficulty was 0.892 (IQR 0.705-0.949). The distribution of the question difficulty for correctly and incorrectly answered questions is displayed in Figure 3.

For the detailed prompt, questions answered correctly had a higher difficulty (median 0.900, IQR 0.737-0.952) compared with questions answered incorrectly (median 0.705, IQR 0.380-0.885). This difference was statistically significant (P=.002).

In the analysis of the short prompt, we also found a lower difficulty for incorrectly answered questions (median 0.708, IQR 0.500-0.864) compared with correctly answered questions (median 0.904, IQR 0.741-0.953). Here as well, a significant difference was detected between the correctly and incorrectly answered questions (P<.001).

XSL•FO
RenderX

Figure 3. Question difficulty by prompt version and answer correctness.

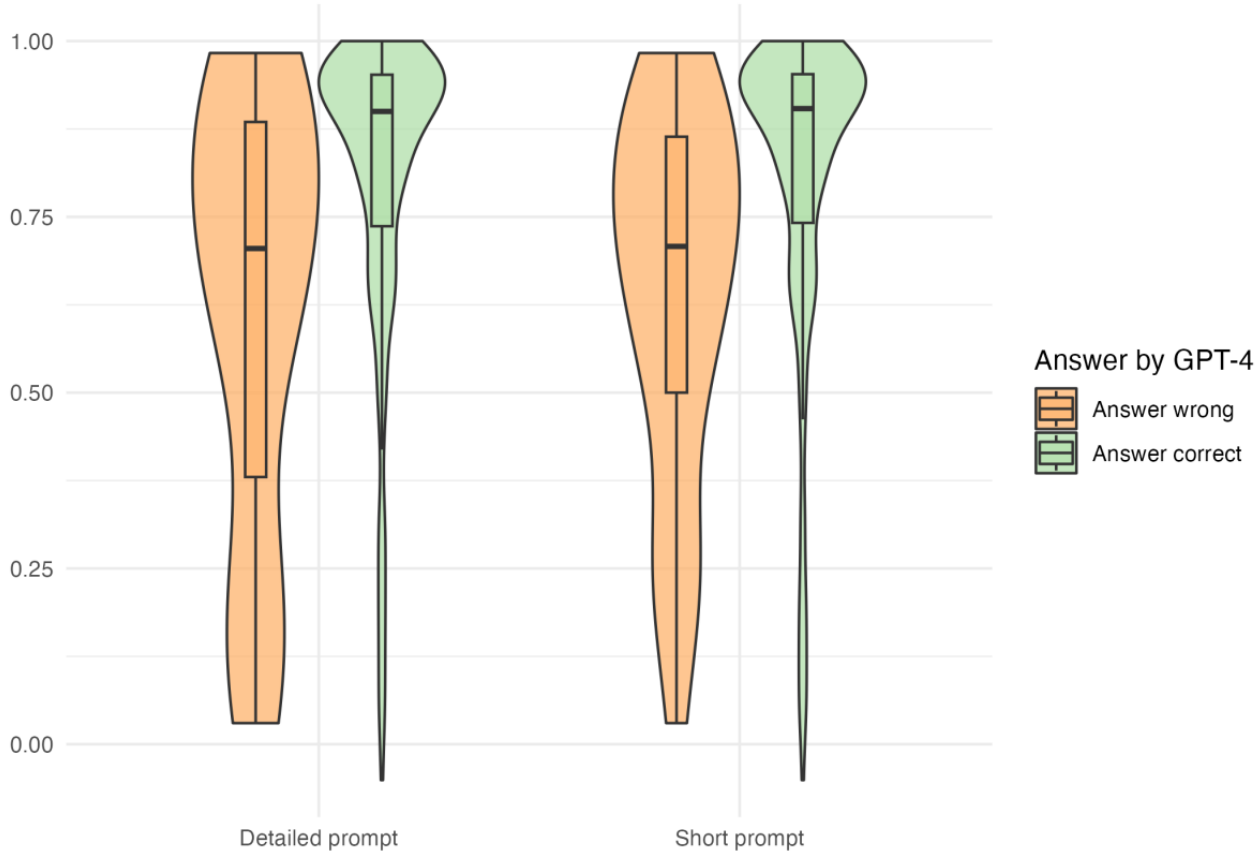

Exam Scores

We further analyzed the performance of GPT-4 for all 16 individual exams (Figure 4). Regardless of the prompt version, GPT-4 never scored below 78.9% and thus always passed the exams. Furthermore, 3 exams (exam "WS17/18" for both prompt versions and exams "SS16" and "SS19" for the detailed prompt only) were passed with a score of 100%.

XSL•FO
RenderX

Figure 4. Performance of GPT-4 for all relevant terms for the detailed prompt (red) and short prompt (blue). GPT-4: Generative Pre-trained Transformer 4.

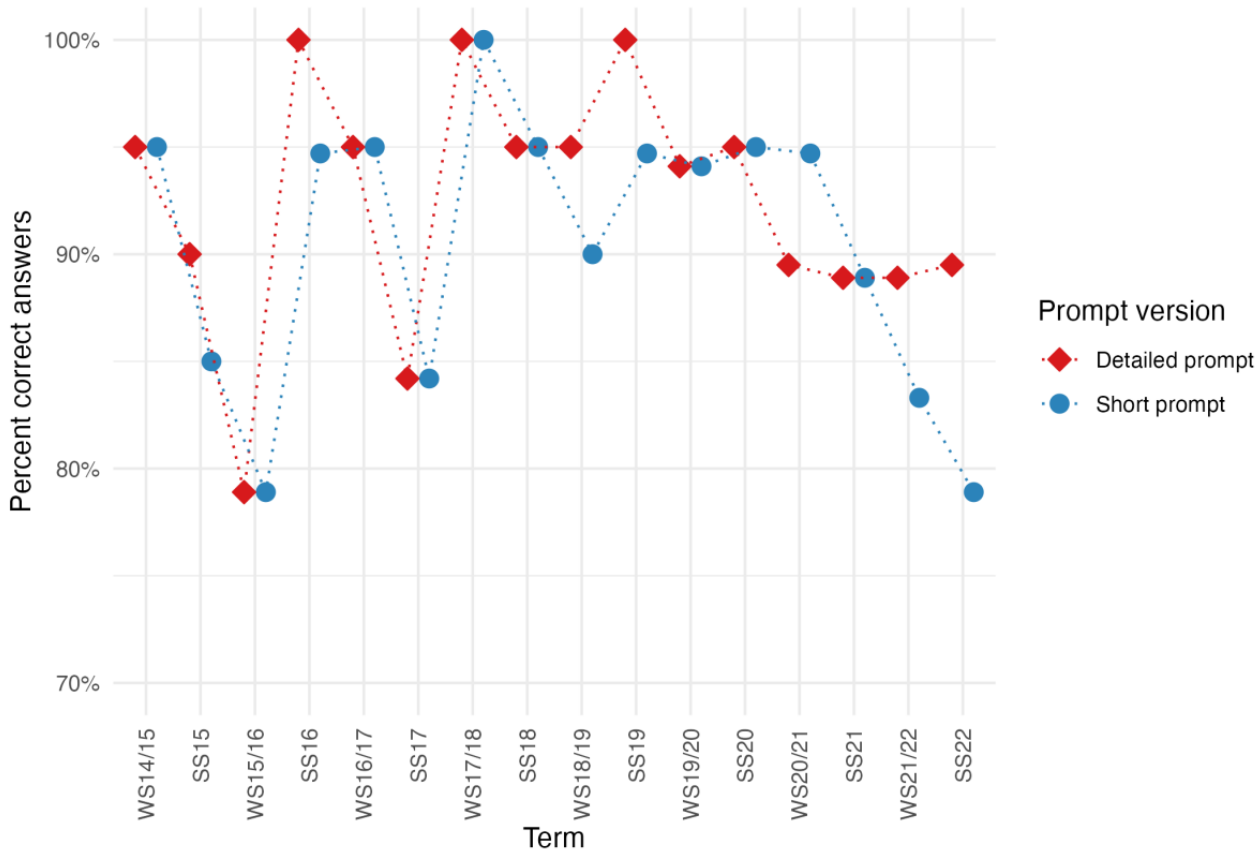

Qualitative Results

After the quantitative results were obtained, we analyzed the answers given by GPT-4 in detail using content analysis.

Correctly Answered Questions

When GPT-4 answered a question correctly, it provided either a detailed or short explanation (Table 2). In the detailed version, GPT-4 also commented on the distractors. For correctly answered questions, the content analysis revealed logical reasoning and the correct use of facts. With regard to the facts cited, no answers were found to be either implausible or incorrect.

XSL•FO
RenderX

Table 2. Correct answers by GPT-4 (Generative Pre-trained Transformer 4) to the question described in Table 1 for the detailed and short prompts.

Prompt version	Correct answer
Detailed prompt	Explanation: After careful consideration of the various answer options and taking into account all the knowledge of psychosomatic medicine, answer E: Depression is the most plausible.
	In this case, the 49-year-old teacher shows clear signs of emotional impairment after an anterior wall infarction. These include disinterest in therapy appointments, lack of interest in his wife's visits, hopelessness about his recovery, and no sense of fear of another heart attack – on the contrary, he sees the prospect of an end to his "misery" as a positive thing. All these signs point to a possible depressive episode.
	Answer A: Nonspecific somatization disorder can be ruled out because the patient does not have a multitude of physical symptoms without a demonstrable organic cause.
	Answer B: Post-traumatic stress disorder could be considered due to the event of the infarction, but the typical symptoms such as memories or flashbacks of the traumatic event, avoidance behaviour, and increased irritability are missing in the description of the case.
	Answer C: Pseudodementia refers to cognitive impairment caused by depressive symptoms rather than actual dementia. Although this would be possible in the case of depression, the question does not provide enough information about cognitive impairment to support this diagnosis.
	Answer D: Generalized Anxiety Disorder does not fit the case as the patient states that he is not afraid of having another heart attack and overall does not show excessive fear or anxiety.
	Therefore, answer E: Depression is the most plausible tentative diagnosis in the psychosomatic field in this case.
Short prompt	Answer E: Depression
	The symptoms described, such as lack of interest in further treatment, lack of hope for improvement, dejection and even death wishes, most likely indicate depression. The patient shows signs of anhedonia (inability to feel pleasure or interest in activities) and negative thought patterns. In such cases, psychosomatic assessment and appropriate treatment is important to aid the patient's recovery.

Incorrectly Answered Questions

When a question was answered incorrectly by GPT-4, the explanation's reasoning was analyzed using Bloom's taxonomy, with its domains "remember," "understand," "apply," "analyze," "evaluate," and "create" (Table 3). The most mistakes were found to be related to "remember," followed by "understand" and "apply," and only 1 erroneous reasoning example was found for "evaluate." In the following sections, examples for each category are given in detail.

Table 3. Count of incorrect reasonings by GPT-4 (Generative Pre-trained Transformer 4) for each category of Bloom's taxonomy. For this analysis, all answers (detailed prompts and short prompts) were analyzed together.

Category of Bloom's taxonomy	Count (n=68), n
Remember	29
Understand	23
Apply	15
Analyze	0
Evaluate	1
Create	0

Bloom's Taxonomy: Remember

This section provides an example of a question for which GPT-4 failed to remember facts. The question presented the case of a man with recurrent physical complaints without organic cause; the complaints persisted for 10 months, making it appropriate for a diagnosis of "undifferentiated somatoform disorder" (F45.1 according to the *International Classification of Diseases, 10th Revision* [ICD-10]). A "somatization disorder" (F45.0) cannot be chosen, as this requires "symptoms of at least 2 years duration" [23].

However, GPT-4 incorrectly chose "somatization disorder." In its reasoning, GPT-4 stated:

> *…I would choose answer C: Somatization disorder (F45.0) as the most appropriate diagnosis for this patient. Somatization disorder (F45.0) is characterized by persistent, multiple, and recurrent physical complaints for which no sufficient organic cause can be found. The symptoms can occur in different organ or functional systems and are usually severe enough to interfere with the patient's daily life.*

In its recall, GPT-4 did not mention the required duration of the symptoms stated in the ICD-10. Thus, from what can be observed in the answer, the concept used by GPT-4 missed at least 1 important fact, and this failure could be attributed to GPT-4 not "remembering."

Bloom's Taxonomy: Understand

An example of GPT-4 not understanding some basic concepts can be found in a question concerning anorexia nervosa. The

XSL•FO
RenderX

answer relates to the fact that anorexia nervosa can be accompanied by transient hypercholesterolemia, which does not necessitate treatment [24,25]. However, GPT-4 picked one of the distractors, which states that the basic metabolic rate increases during phases of low body weight and decreases after refeeding [26]—but the opposite is true. In its explanation, GPT-4 thought about these facts (and even remembered them correctly in sentence 3!) but did not seem to understand how they relate and interact.

> *In patients with anorexia nervosa, the basal metabolic rate is initially elevated at the onset of weight gain and decreases dramatically during refeeding. Food intake should be adjusted accordingly. Basal metabolic rate here refers to the amount of energy the body needs at rest to maintain its basic functions. In people with anorexia nervosa, the body is initially very efficient at obtaining energy due to persistent nutrient deficiency. However, as refeeding progresses, the basal metabolic rate drops dramatically as the body re-learns to store energy.*

Bloom's Taxonomy: Apply

The third level of Bloom's taxonomy involves the application of previously learned content to other areas. An error in this domain was evident when we asked GPT-4 a question about the diagnosis and treatment of a depressive episode. The correct answer was that symptoms or criteria must be present for at least 14 days, which is clearly required by the ICD-10 [23]. However, GPT-4 rates this answer as incorrect and implies that this time frame is more of a rough guideline than a strict minimum requirement.

> *In the diagnosis of depressive episodes, the time criterion is an important factor. However, many experts believe that a period of at least two weeks is rather a rough guideline. Some patients can also go through a depressive episode in a shorter time or fulfil the criteria for a longer time.*

This reveals an issue at the "application" level of Bloom's taxonomy. GPT-4 seemed to "remember" the fact about the temporal criterion and "understood" its concept. However, it struggled with the "application" of this rule, proposing flexibility where there is conventional rigidity. This can be interpreted as a misapplication of the learned rule and, thus, a mistake made at the application level of Bloom's taxonomy.

Bloom's Taxonomy: Evaluate

GPT-4 made a mistake at the "evaluate" level of Bloom's taxonomy in a more complex question testing one's judgment regarding the ideal therapy for a patient. Due to the complexity, the question is presented here in full:

> *A 57-year-old ready-to-wear clothing-maker is referred to the psychosomatic outpatient clinic by her family doctor because of persistent back pain. She reports "terrible" pain in her lower back. She has been unable to work for seven months because of it and can hardly sleep. No one was able to help her; the family doctor just sent her from doctor to doctor. (The findings from orthopaedics, neurology, neurosurgery, and radiology that she brought with her show minor disc changes on several levels, but no neurological deficits). She does not want to undergo surgery because "they probably could not earn enough money from her as a patient with public insurance". Now the family doctor has prescribed her an antidepressant, which she will not take. She still has hope for improvement, but if it continues like this, she does not know whether she will put an end to her life. What is the first course of action indicated?*
>
> *A: Immediate admission to a psychiatric ward, against the patient's will if necessary.*
>
> *B: Initiate inpatient psychosomatic treatment.*
>
> *C: Observation of the symptoms under medication with NSAIDs.*
>
> *D: Obtain a neurosurgical second opinion after current imaging.*
>
> *E: Clarification of secondary disease gain and immediate termination of incapacity for work.*

When we used the short prompt, GPT-4 chose the answer B: Plan a psychosomatic inpatient treatment. In its brief explanation, GPT-4 correctly judged why this was the most likely answer.

However, when we used the detailed prompt, GPT-4 chose the distractor E and sent the patient back to work effectively immediately, arguing:

> *After careful consideration, answer E is most appropriate, but with modification. A detailed explanation of psychosomatic correlations and the possible role of secondary illness gain might help the patient better understand her situation and is thus indicated. However, immediate cessation of incapacity should not be forced, as this could potentially exacerbate the patient's pain and frustration. It would make sense to first seek outpatient psychosomatic treatment to help the patient cope with her pain and focus on long-term recovery.*

In this example, it is evident that GPT-4 knows the facts (remember), understands them (understand), applies them to this situation (apply), and correctly analyzes the situation (analyze). GPT-4 can even be observed to evaluate, but it fails at evaluating the consequences of its decision. While GPT-4 seems to be aware of this and wants to modify the given answer, it incorrectly evaluates the answer options and does not recommend inpatient psychosomatic treatment.

Discussion

Principal Findings

Tools based on LLMs are rapidly evolving and influencing medicine [1] and medical education [2]. In this study, we show that GPT-4 yields high accuracy in answering psychosomatic medicine questions, with success rates well above 90%. This confirms results from other studies, which also show high exam scores in, for example, the USMLE (United States Medical Licensing Examination) [14] and other board exams [27].

XSL•FO
RenderX

Although we detected high correct-answer rates for both our detailed prompt and short prompt versions, GPT-4 answered more questions correctly when we used the detailed prompt (284/307, 92.5% vs 278/307, 90.6%). While this study was not designed to assess reasons for this difference, 2 likely explanations for it deserve attention. First, the difference rates for the 2 prompt versions could represent a true difference in the LLM's performance with respect to the prompt version [28]. Second, the difference could be due to a random variation that is known to occur even when the same prompt is used more than once [29,30].

We were also interested in GPT-4's performance compared with that of medical students. Here, our analysis revealed that the questions answered correctly by GPT-4 were significantly easier than the questions that were incorrectly answered. This difference could be observed for both the detailed prompt and short prompt versions. For further comparison, it should be noted that question difficulty is not a fixed or static variable but rather is dependent and calculated on the basis of the responses of human students [20].

However, in order to understand why GPT-4 fails at some questions, we further analyzed incorrectly answered questions using a qualitative approach. It is well known that incorrect or inaccurate information is an important issue with LLMs [31]. Bloom's taxonomy has emerged as a frequently used standard to describe the cognitive process underlying learning [8]. To the best of our knowledge, the levels in Bloom's taxonomy at which GPT-4 commits cognitive errors have not yet been elucidated. Thus, we performed a detailed assessment of the answers and reasoning provided by GPT-4.

In our analysis, we found that most errors were made at the lowest level of Bloom's taxonomy, labeled "remember." In these answers, GPT-4 failed at naming or using a specific fact, as evident in the text response. In the example presented in the previous section, GPT-4 named most of the diagnostic criteria for a somatization disorder but did not mention the time criterion. In this context, it is important to note that GPT-4 has been trained with publicly available and licensed data (although these are not specified in detail by OpenAI) [32]. The information needed in this example is publicly available in the ICD-10 [23] and thus is expected to be included in the GPT-4 training data. Indeed, when asked by the authors, ChatGPT (using GPT-4) states that its training data include details on the ICD-10.

In a recent study, Johnson et al [33] evaluated ChatGPT for its accuracy in providing medical information. Using a quantitative approach, they found that GPT-3.5 provided medical answers ranging from "mostly correct" to "almost correct." It is important to note that these results were generated using GPT-3.5, an older model than the model we used: GPT-4 is claimed to be "40% more likely to produce factual responses than GPT-3.5" [32] and exhibits better performance in medical exams [27]. However, the results reported by Johnson et al [33] are in line with our findings: while GPT-4 uses most facts correctly and completely, it sometimes fails with respect to specific details. In psychosomatic medicine, we observed this to be a diagnostic criterion; while this can be an issue, missing some specific facts in other areas can make all the difference for patient outcomes. Thus, it is important for those using GPT-4 in medicine to keep in mind that specific facts can be wrong or missing.

Some errors were found at the second level of Bloom's taxonomy, labeled "understand." While GPT-4 generally showed good reasoning capability [32] and errors were sparse, we were surprised that some answers yielded obvious logical flaws, as seen in the example from the previous section. In its response, GPT-4 confidently presents a set of sentences that do not correlate logically. Mechanistically, language models such as GPT use likelihood maximization, generating text based on what most likely follows [16,34]. However, this approach can result in what is called hallucinations, or "content that is nonsensical or unfaithful to the provided source content" [34]. The resulting medical information might sound very confident but be incorrect [35], thereby posing a significant threat for medical applications [5]. This raises ethical concerns around the use of AI systems for patient-related work, particularly as GPT-4's algorithms and ethical models are unknown and variable [4]. Because GPT-4 is not considered to be sentient, it neither knows nor cares about the accuracy of its responses.

We also detected some mistakes that represented Bloom's taxonomy level labeled "application." In our representative example presented in the previous section, GPT-4 was quite flexible in applying a very strict time criterion. This can be interpreted in the context of the process of training LLMs. Although little has been published about this process, classification systems probably represent only a small amount of the data available on a certain subject. It can be further assumed that information designed for the public might not be as specific as strict diagnostic criteria because it serves another audience. Thus, following GPT-4's likelihood maximization approach, LLMs might neglect a specific but likely underrepresented piece of information. Furthermore, GPT-4 has been observed to perform poorly in pure calculation tasks [35], probably also challenging strictly numerical criteria. This is not entirely surprising, as GPT-4 is a LLM (with an emphasis on language) and is not intended to be used as a calculator.

We found only 1 mistake that could be classified at the Bloom's taxonomy level labeled "evaluate," in which GPT-4 incorrectly judged a medically complex situation.

Limitations

This study has some limitations that merit discussion. First, we used Bloom's taxonomy. While it provides good operationalization for cognitive processes, the taxonomy represents a continuum wherein more than 1 level can be activated in a single question [8]. Nevertheless, we observed that most errors could be attributed to only 1 cognitive level. Second, we chose questions from psychosomatic medicine because many aspects of this field can be covered by written language and do not require images or many numbers. However, psychosomatic medicine is a specialty in which treatment can be individualized and especially complex, making categorical judgment harder and possibly reducing generalizability beyond this field. Third, since we only used GPT-4 as our LLM, we cannot judge if our implications hold true for other models.

XSL•FO
RenderX

Fourth, our data were acquired 1 time, at a specific date. As the performance of GPT-4 varies over time, this could reduce generalizability [36]. Finally, in this study, we compared GPT-4's performance with that of medical students by using the difficulty index of each exam question. It is important to note that this difficulty index is specific to the particular cohort of medical students who participated in the representative exam.

Conclusion

In summary, we found that GPT-4 performs extremely well on psychosomatic medicine questions. Questions answered correctly by GPT-4 were also easier for human students than questions answered incorrectly, as shown by the level of question difficulty. When analyzing the mistakes of GPT-4, we found that most errors corresponded to lower-order cognitive levels, particularly "remember" and "understand." While we found some mistakes for "apply," very few or no errors were found for "analyze" and "evaluate" ("create" could not be assessed in this study). To the best of our knowledge, this study is the first to describe the cognitive levels at which GPT-4 makes mistakes in the context of psychosomatic medicine.

This study has important implications. First, GPT-4 is already capable of answering many questions in (psychosomatic) medicine, and thus, should the technology be made available, it could reduce the effectiveness of summative assessment. Second, GPT-4 sometimes fails at exact facts, correct understanding, and application of knowledge; however, without exact knowledge, these failures are hard to recognize. Thus, the output generated by GPT-4 must be checked for accuracy, especially in those domains. Our research can also help in model training, and future studies can use our results to correlate model training and LLM outcomes.

Acknowledgments

The authors would like to acknowledge the support of the Open Access Publishing Fund of the University of Tübingen.

Data Availability

The data sets used and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AHW and MM were responsible for designing and conducting the study, as well as the acquisition, analysis, and interpretation of data. AHW and MM also drafted the first version of the manuscript. TFW, FH, and LH were involved in data analyses and interpretation. JG, KM, and SZ made substantial contributions to the study design. All authors critically revised the manuscript, and all authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Reddy S. Explainability and artificial intelligence in medicine. Lancet Digit Health 2022;4(4):e214-e215 [FREE Full text] [doi: 10.1016/S2589-7500(22)00029-2] [Medline: 35337639]
2. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
3. ChatGPT: optimizing language models for dialogue. OpenAI. 2022. URL: https://web.archive.org/web/20221130180912/https://openai.com/blog/chatgpt/ [accessed 2023-08-07]
4. Masters K. Ethical use of artificial intelligence in health professions education: AMEE guide no. 158. Med Teach 2023;45(6):574-584 [FREE Full text] [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
6. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I: Cognitive Domain. London: Longman Green & Co; 1956.
7. Anderson LW, Krathwohl DR. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition. New York: Longman; 2001.
8. Krathwohl DR. A revision of Bloom's Taxonomy: an overview. Theory Into Practice 2010;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
9. Harden RM. Outcome-based education: the future is today. Med Teach 2007;29(7):625-629. [doi: 10.1080/01421590701729930] [Medline: 18236247]
10. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63-S67 [FREE Full text] [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]

XSL•FO
RenderX

11. Taylor DCM, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE guide no. 83. Med Teach 2013;35(11):e1561-e1572 [FREE Full text] [doi: 10.3109/0142159X.2013.828153] [Medline: 24004029]
12. Horne S. Learning hierarchies: a critique. Educ Psychol 2006;3(1):63-77. [doi: 10.1080/0144341830030106]
13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
14. Kung TH, Cheatham M, Medenilla A, Sillos C, de Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
15. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
16. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv. Preprint posted online on March 22, 2023 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.12712]
17. Epstein RM. Assessment in medical education. N Engl J Med 2007;356(4):387-396. [doi: 10.1056/NEJMra054784] [Medline: 17251535]
18. Case SM, Swanson DB. Constructing Written Test Questions For the Basic and Clinical Sciences, Third Edition. Philadelphia, PA: National Board of Medical Examiners; 2002.
19. R: a language and environment for statistical computing. The R Project for Statistical Computing. URL: https://www.R-project.org/ [accessed 2024-01-02]
20. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. Article in German. GMS Z Med Ausbild 2006;23(3):23 [FREE Full text]
21. Mayring P. Qualitative inhaltsanalyse. In: Mey G, Mruck K, editors. Handbuch qualitative Forschung in der Psychologie. Wiesbaden: VS Verlag für Sozialwissenschaften; 2010:601-613.
22. Moldt JA, Festl-Wietek T, Madany-Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023;28(1):2182659 [FREE Full text] [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
23. ICD-10: international statistical classification of diseases and related health problems: tenth revision, 2nd ed. World Health Organization. 2004. URL: https://apps.who.int/iris/handle/10665/42980 [accessed 2023-08-05]
24. Ohwada R, Hotta M, Oikawa S, Takano K. Etiology of hypercholesterolemia in patients with anorexia nervosa. Int J Eat Disord 2006;39(7):598-601. [doi: 10.1002/eat.20298] [Medline: 16791856]
25. Weinbrenner T, Züger M, Jacoby GE, Herpertz S, Liedtke R, Sudhop T, et al. Lipoprotein metabolism in patients with anorexia nervosa: a case-control study investigating the mechanisms leading to hypercholesterolaemia. Br J Nutr 2004;91(6):959-969 [FREE Full text] [doi: 10.1079/BJN20041151] [Medline: 15182399]
26. Kosmiski L, Schmiege SJ, Mascolo M, Gaudiani J, Mehler PS. Chronic starvation secondary to anorexia nervosa is associated with an adaptive suppression of resting energy expenditure. J Clin Endocrinol Metab 2014;99(3):908-914 [FREE Full text] [doi: 10.1210/jc.2013-1694] [Medline: 24302748]
27. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
28. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. : Association for Computational Linguistics; 2021 Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; November 7–11, 2021; Online and Punta Cana, Dominican Republic p. 3045-3059. [doi: 10.18653/v1/2021.emnlp-main.243]
29. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6(1):9 [FREE Full text] [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
30. Dale R. GPT-3: what's it good for? Nat Lang Eng 2020;27(1):113-118 [FREE Full text] [doi: 10.1017/S1351324920000601]
31. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
32. GPT-4. OpenAI. URL: https://openai.com/gpt-4 [accessed 2023-08-09]
33. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Research Square. Preprint posted online on February 28 2023 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
34. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38. [doi: 10.1145/3571730]
35. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023;307(5):e230987. [doi: 10.1148/radiol.230987] [Medline: 37191491]
36. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv. Preprint posted online on July 18, 2023 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.09009]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
API: application programming interface
GPT: Generative Pre-trained Transformer
ICD-10: International Classification of Diseases, 10th Revision
LLM: large language model
MCQ: multiple-choice question
USMLE: United States Medical Licensing Examination

Edited by T de Azevedo Cardoso; submitted 23.08.23; peer-reviewed by X Li, A Gencer; comments to author 26.08.23; revised version received 15.09.23; accepted 07.12.23; published 23.01.24.

Please cite as:
Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, Zipfel S, Mahling M
Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study
J Med Internet Res 2024;26:e52113
URL: https://www.jmir.org/2024/1/e52113
doi:10.2196/52113
PMID:38261378

©Anne Herrmann-Werner, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, Moritz Mahling. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identifying Functional Status Impairment in People Living With Dementia Through Natural Language Processing of Clinical Documents: Cross-Sectional Study

John Laurentiev[1*], MS; Dae Hyun Kim[1,2,3*], MD, ScD; Mufaddal Mahesri[1], MPH, MD; Kuan-Yuan Wang[4], MD, PhD; Lily G Bessette[1], MS; Cassandra York[1], BS; Heidi Zakoul[1], BA; Su Been Lee[1], BA; Li Zhou[1,2*], MD, PhD; Kueiyu Joshua Lin[1,2,5*], MD, ScD

[1]Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States

[2]Harvard Medical School, Boston, MA, United States

[3]Marcus Institute for Aging Research, Hebrew SeniorLife, Boston, MA, United States

[4]National Cheng Kung University Hospital, Tainan, Taiwan

[5]Massachusetts General Hospital, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Kueiyu Joshua Lin, MD, ScD
Department of Medicine
Brigham and Women's Hospital
1620 Tremont St. Suite 3030
Boston, MA, 02120
United States
Phone: 1 617 278 0930
Fax: 1 617 232 8602
Email: jklin@bwh.harvard.edu

Abstract

Background: Assessment of activities of daily living (ADLs) and instrumental ADLs (iADLs) is key to determining the severity of dementia and care needs among older adults. However, such information is often only documented in free-text clinical notes within the electronic health record and can be challenging to find.

Objective: This study aims to develop and validate machine learning models to determine the status of ADL and iADL impairments based on clinical notes.

Methods: This cross-sectional study leveraged electronic health record clinical notes from Mass General Brigham's Research Patient Data Repository linked with Medicare fee-for-service claims data from 2007 to 2017 to identify individuals aged 65 years or older with at least 1 diagnosis of dementia. Notes for encounters both 180 days before and after the first date of dementia diagnosis were randomly sampled. Models were trained and validated using note sentences filtered by expert-curated keywords (filtered cohort) and further evaluated using unfiltered sentences (unfiltered cohort). The model's performance was compared using area under the receiver operating characteristic curve and area under the precision-recall curve (AUPRC).

Results: The study included 10,000 key-term–filtered sentences representing 441 people (n=283, 64.2% women; mean age 82.7, SD 7.9 years) and 1000 unfiltered sentences representing 80 people (n=56, 70% women; mean age 82.8, SD 7.5 years). Area under the receiver operating characteristic curve was high for the best-performing ADL and iADL models on both cohorts (>0.97). For ADL impairment identification, the random forest model achieved the best AUPRC (0.89, 95% CI 0.86-0.91) on the filtered cohort; the support vector machine model achieved the highest AUPRC (0.82, 95% CI 0.75-0.89) for the unfiltered cohort. For iADL impairment, the Bio+Clinical bidirectional encoder representations from transformers (BERT) model had the highest AUPRC (filtered: 0.76, 95% CI 0.68-0.82; unfiltered: 0.58, 95% CI 0.001-1.0). Compared with a keyword-search approach on the unfiltered cohort, machine learning reduced false-positive rates from 4.5% to 0.2% for ADL and 1.8% to 0.1% for iADL.

Conclusions: In this study, we demonstrated the ability of machine learning models to accurately identify ADL and iADL impairment based on free-text clinical notes, which could be useful in determining the severity of dementia.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e47739) doi:10.2196/47739

KEYWORDS

activities of daily living; ADLs; clinical note; dementia; electronic health record; EHR; functional impairment; instrumental activities of daily living; iADLs; machine learning; natural language processing; NLP

Introduction

In the United States, over 6 million people are living with Alzheimer disease or related dementia, and this number is projected to increase to 13 million by 2050 [1]. As dementia progresses, the affected individuals lose the ability to carry out everyday activities, including basic activities of daily living (ADLs) and instrumental ADLs (iADLs), which are fundamental skills required to independently care for oneself and serve as an indicator of a person's functional status [2]. Assessment of ADLs and iADLs is also essential in determining the severity of dementia and the care needs of older adults [3]. This information is important in predicting a patient's risk of mortality, long-term nursing home admission, and health care use.

Despite the significance of assessing disability in performing ADLs and iADLs in people living with dementia, the assessment is not routinely done in clinical practice; if performed, it is often documented within the unstructured clinical notes of a patient's electronic health record (EHR), making it difficult to readily locate. This process could be improved with natural language processing (NLP) and machine learning. NLP has been applied to health care research in a variety of ways, including quantifying changes in social media posts, better understanding the mental health impacts of COVID-19 [4], and extracting cancer phenotypes from clinical note text [5]. An NLP approach can convert free-text information on ADLs and iADLs from an EHR into structured data. The structured data of ADL and iADL can then be readily used in clinical care and research to perform statistical modeling for outcome prediction, patient phenotyping, and confounding or risk adjustment. In this study, we seek to develop and validate machine learning models that can identify clinical note text containing information on ADL and iADL impairment in people living with dementia.

Methods

Setting and Data Sources

We used data from the Research Patient Data Repository (RPDR) [6] linked to Medicare fee-for-service parts A (inpatient coverage), B (outpatient coverage), and D (prescription benefits) claims data for over 700,000 individuals from 2007 to 2017. The RPDR includes longitudinal EHR data from 2 tertiary hospitals, 3 community hospitals, and more than 35 primary care centers in Greater Boston, Massachusetts. This data set includes demographic information, inpatient and outpatient diagnoses and procedures, medical orders and drug prescriptions, vital signs, laboratory and radiology test results, and free-text notes and reports from inpatient and ambulatory encounters. We linked the EHR with Medicare claims data to reduce information leakage of the EHR due to care provided outside of our EHR [7,8]. The linkage was done deterministically by the unique Medicare beneficiary number, date of birth, and sex, with a success linkage rate of 98.7% [9]. Medicare is a US federal health insurance program that currently covers approximately 50 million Americans by providing medical and prescription drug coverage to individuals aged 65 years or older and to younger individuals with disabilities. The Medicare claims database contains longitudinal, individual-level data on health care use, diagnoses, diagnostic tests, procedures, and pharmacy-filled prescriptions.

Study Population

From the linked RPDR-claims data from 2007 to 2017, we identified older adults aged 65 years or older with at least 1 diagnosis of dementia using validated algorithms with positive predictive values of 78%-92% to define dementia [10]. The first date of dementia diagnosis during our study period was the cohort entry (index) date. We further required the study cohort to have at least 365 days of continuous enrollment in Medicare parts A and B and at least 1 admission, progress, or discharge note from an inpatient encounter or ambulatory visit within 365 days before the index date.

Labeled Data Set Development

Patient clinical notes were split into sentences using the Medical Text Extraction, Reasoning, and Mapping System NLP system [11]. In our preliminary exploration, the information relevant for ADL and iADL was noted to be sparse, resulting in a highly imbalanced data set with little input data for model development. To increase the prevalence of data containing information relevant to ADL and iADL impairment in our model training data set, we created a lexicon of ADL- and iADL-related key terms curated with physician expert guidance. A total of 3 medical doctors (DHK, MM, and KJL) came up with the initial list of terms with automatically generated synonyms, followed by list refinement based on clinical knowledge (see Table S1 in Multimedia Appendix 1 for a final list of the terms). We filtered and kept sentences that included at least 1 key term from the final lexicon using a filtering tool coded in Python. Notes were split into sentences but retained a context window of 250 characters before and after key terms. Each medical doctor was paired with a research assistant to form a review team. The 3 review teams first conducted preliminary reviews on the same set of notes, filtered by our key term lexicon. They discussed sentences with different classifications regarding ADL and iADL impairment and assessed interrater agreement (κ was 85.2%, 89.8%, and 85.4% between the 3 teams after 5 rounds of training sessions). Then the review teams manually labeled 10,000 randomly selected filtered sentences in 2743 clinical notes from 441 patients for evidence of ADL or iADL impairment. The data were randomly split into a 70% subset (filtered training cohort) and a 30% subset (filtered validation cohort). We used the training cohort to train and tune model parameters by 5-fold cross-validation and used the validation

XSL•FO
RenderX

cohort to test the performance. Models were further evaluated using 1000 randomly selected, unfiltered sentences from an independent set of 80 patients for assessing generalizability.

Classifier Development

We implemented 5 commonly used statistical models in machine learning literature: logistic regression, support vector machine (SVM), least absolute shrinkage and selection operator (LASSO) regression [12], random forest [13], implemented using the Python Scikit-learn module [14], and gradient boosting, implemented using the Python XGBoost module [15]. Training data for these 5 models were represented as unigrams transformed using term frequency—inverse document frequency [16]. We also implemented a hierarchical attention-based deep learning model consisting of a convolutional neural network and long short-term memory network, developed in a previous study [17]. Additionally, we implemented a model derived from Bio+Clinical bidirectional encoder representations from transformers (BERT), a contextualized word representation model based on BioBERT, and trained further on Medical Information Mart for Intensive Care (MIMIC) data [18-21]. We performed 5-fold cross-validation on the training cohort to tune the parameters for each model based on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) metrics. Tuned model performance was then evaluated on the filtered validation cohort, and generalizability was tested on the unfiltered validation cohort (ie, the 1000 unfiltered sentence set). See Tables S2 and S3 in Multimedia Appendix 1 for the final parameters of the ADL and iADL impairment classifiers, respectively.

Ethical Considerations

The study was approved by the institutional review board of Brigham and Women's Hospital, Boston, Massachusetts (2018P002462). Personal health information was used as minimally as possible within the needs of the study. Data were not shared with any individuals not directly involved in the study.

Results

Study Sample

The filtered cohort was used to extract our data set of 10,000 filtered sentences; it comprises 441 people with a mean age of 82.7 (SD 7.9) years. A total of 64% (n=283) of the people were female, 88% (n=389) were White, and 4% (n=19) were Black. During the 365-day baseline period, the mean frailty score in our cohort was 0.26 (SD 0.08) with the most commonly observed comorbidities being urinary tract infections (n=192, 43.5%), history of falls (n=172, 39%), failure to thrive (n=89, 20.2%), incontinence (n=73, 16.5%), pressure ulcers (n=57, 12.9%), and dysphagia (n=50, 11.3%). As far as health care use is concerned, people on average had 13 (SD 7.8) unique medications along with a mean of 11.5 (SD 8.8) outpatient visits in the baseline period. Furthermore, the mean number of baseline hospitalizations and emergency room visits were 1.3 (SD 1.7) and 2.4 (SD 2.7), respectively. Our key-term filtered data set contained 1628 (16.3%) sentences annotated as positive (ie, containing relevant information) for ADL impairment (1128 in the training subset and 500 in the internal evaluation subset) and 323 (3.2%) sentences annotated as positive for iADL impairment (234 in the training subset and 89 in the internal evaluation subset). In contrast, the unfiltered data set was used to extract the external validation cohort of 1000 unfiltered sentences from 80 patients that contained 7 (0.7%) sentences labeled positive for ADL impairment and 4 (0.4%) sentences labeled positive for iADL impairment. Compared to the filtered data set, the unfiltered data set has a comparable mean age and race composition but a slightly higher female percentage (56/80, 70% vs 283/441, 64.2%). The baseline comorbidity profile of the 2 study cohorts was largely comparable, except that the prevalence of aspiration pneumonia was higher in the filtered than unfiltered set. The health care use was also noted to be slightly higher in the filtered than in the unfiltered set (Table 1).

XSL•FO
RenderX

Table 1. Selected characteristics of the filtered and unfiltered study data set.

Covariate	Filtered 10,000-sentence sample (n=441)	Unfiltered 1000-sentence sample (n=80)	Absolute standardized difference
Age (years), mean (SD)	82.7 (7.9)	82.8 (7.5)	0.02
Age categories (years), n (%)			
65-70	36 (8.2)	5 (6)	0.07
71-75	51 (11.6)	9 (11)	0.01
76-80	68 (15.4)	12 (15)	0.01
81-85	96 (21.8)	17 (21)	0.01
>85	190 (43.1)	37 (46)	0.06
Sex, n (%)			
Female	283 (64.2)	56 (70)	0.12
Male	158 (35.9)	24 (30)	0.12
Race, n (%)			
White	389 (88.2)	70 (88)	0.02
Black	19 (4.3)	4 (5)	0.03
Other	33 (7.5)	6 (8)	0.00
Frailty score, mean (SD)	0.26 (0.1)	0.25 (0.1)	0.06
Comorbidities, n (%)			
Falls	172 (39.0)	30 (38)	0.03
Pressure ulcer	57 (12.9)	11 (14)	0.02
Failure to thrive or body weight loss	89 (20.2)	15 (19)	0.04
Use of feeding tube	7 (1.6)	2 (3)	0.06
Aspiration pneumonia	33 (7.5)	1 (1)	0.31
UTI[a]	192 (43.6)	34 (43)	0.02
Incontinence	73 (16.6)	14 (18)	0.03
Dysphagia (eating problems)	50 (11.3)	6 (8)	0.13
Medications, n (%)			
1st line dementia medication use	49 (11.1)	7 (9)	0.08
Memantine use	21 (4.8)	4 (5)	0.01
Antipsychotic medication use	38 (8.7)	8 (10)	0.05
Health care use, mean (SD)			
Hospitalizations, n	1.32 (1.7)	1.25 (2)	0.04
ER[b] visits, n	2.40 (2.7)	2.00 (2)	0.16
Outpatient visits, n	11.54 (8.8)	10.15 (9)	0.16
Medications, n	13.06 (7.8)	11.94 (8)	0.14
Nursing home stay of >100 days in baseline, n (%)	17 (3.9)	6 (8)	0.16
Hospice care, n (%)	12 (2.7)	4 (5)	0.12

[a]UTI: urinary tract infection.

[b]ER: emergency room.

Performance of the ADL Model

AUROC and AUPRC performance across all models for ADL impairment detection in the training set and both evaluation sets are shown in Table 2 with receiver operating characteristic (ROC) and precision-recall curves for the filtered validation cohort evaluation in Figures 1 and 2, respectively. While most models scored high AUROC across data sets, there was more notable variation in AUPRC scores, particularly in the unfiltered validation set. LASSO performed best in training set

XSL•FO
RenderX

cross-validation, with an AUROC of 0.958 and AUPRC of 0.865. Top predictors of the LASSO model include "incontinence," "tube," "incontinent," "PEG (percutaneous endoscopic gastrostomy)," "bathing," "dressing," "feeding," "total parenteral nutrition (TPN)," "assistance," and "toileting." These features tended to have high importance for the remaining models, as well as "gastrostomy" and "body." As shown in Table 2, most ADL models performed similarly on the filtered validation cohort, with the random forest model achieving slightly better AUROC and AUPRC measures (0.971 and 0.887, respectively) than the others. All models' AUROC scores improved for the unfiltered validation cohort, explained by the notable imbalance of the data set—only 0.7% (7/1000) cases were positive for ADL impairment in the unfiltered validation cohort versus 16.7% (500/3000) in the filtered validation cohort. Unfiltered validation AUROC was highest for the deep learning model (0.991). The AUPRC scores of all models decreased for the unfiltered validation cohort prediction, with the SVM model's score the highest (0.822) and dropping the least. Though trained on more data than the deep learning model, the data that the Bio+Clinical BERT model is pretrained on is not specific to the Mass General Brigham (MGB) EHR. This is likely why Bio+Clinical BERT exhibited lower performance than the deep learning model, which was trained entirely on our annotated MGB data set.

Table 2. Activities of daily living classifier performance. Italic values represent the optimal performance in each data set.

Model	AUROC[a] (95% CI)	AUPRC[b] (95% CI)
Filtered cohort		
Training set		
Deep learning	0.952 (0.945-0.960)	0.864 (0.844-0.881)
Bio+Clinical BERT[c]	0.870 (0.842-0.897)	0.826 (0.789-0.862)
Logistic regression	0.955 (0.949-0.962)	0.855 (0.837-0.872)
LASSO[d]	*0.958 (0.952-0.965)*	*0.865 (0.846-0.883)*
Random forest	0.953 (0.946-0.960)	0.857 (0.838-0.875)
SVM[e]	0.954 (0.946-0.960)	0.854 (0.835-0.872)
XGBoost	0.955 (0.948-0.962)	0.848 (0.826-0.869)
Validation set		
Deep learning	0.961 (0.951-0.971)	0.880 (0.852-0.906)
Bio+Clinical BERT	0.873 (0.852-0.891)	0.847 (0.823-0.869)
Logistic regression	0.963 (0.954-0.971)	0.871 (0.841-0.896)
LASSO	0.962 (0.954-0.970)	0.870 (0.841-0.896)
Random forest	*0.971 (0.964-0.977)*	*0.887 (0.859-0.913)*
SVM	0.963 (0.954-0.971)	0.877 (0.849-0.902)
XGBoost	0.961 (0.951-0.969)	0.873 (0.846-0.898)
Unfiltered validation cohort		
Deep learning	*0.991 (0.986-0.994)*	0.817 (0.746-0.882)
Bio+Clinical BERT	0.785 (0.582-0.999)	0.621 (0.227-0.901)
Logistic regression	0.981 (0.971-0.990)	0.737 (0.644-0.817)
LASSO	0.969 (0.954-0.983)	0.675 (0.573-0.769)
Random forest	0.990 (0.984-0.995)	0.806 (0.723-0.880)
SVM	0.986 (0.975-0.994)	*0.822 (0.748-0.887)*
XGBoost	0.978 (0.959-0.992)	0.771 (0.680-0.846)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]BERT: bidirectional encoder representations from transformers.

[d]LASSO: least absolute shrinkage and selection operator.

[e]SVM: support vector machine.

XSL•FO
RenderX

Figure 1. Receiver operating characteristic curves for activity of daily living impairment prediction performance on the filtered validation subset. AUC: area under the curve; BERT: bidirectional encoder representations from transformers; LASSO: least absolute shrinkage and selection operator; SVM: support vector machine.

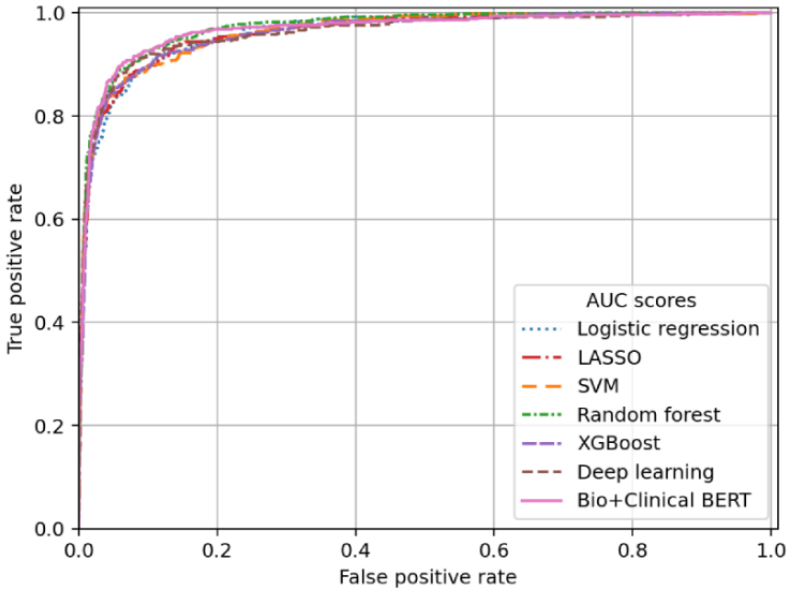

Figure 2. Precision-recall curves for activity of daily living impairment prediction performance on the filtered validation subset. BERT: bidirectional encoder representations from transformers; LASSO: least absolute shrinkage and selection operator; SVM: support vector machine.

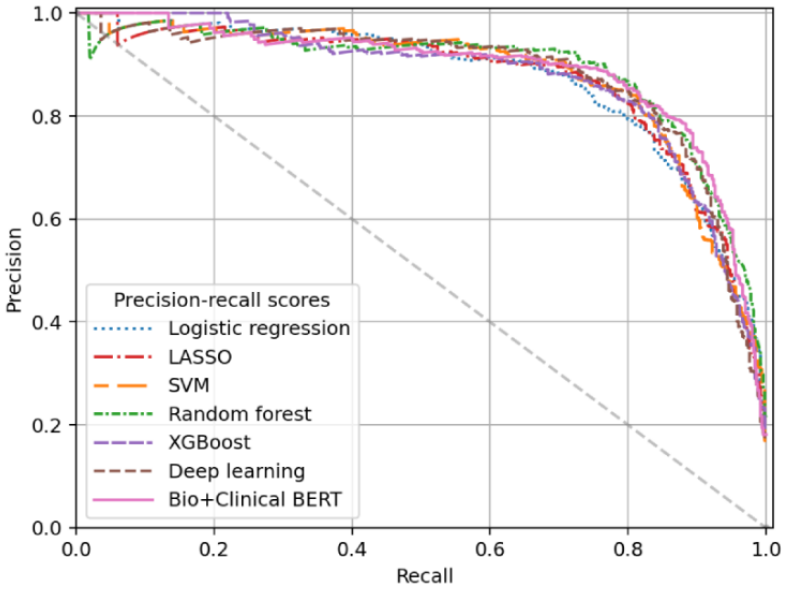

Performance of the iADL Model

Evaluation data sets were more imbalanced for the iADL impairment classification task—the filtered validation cohort had a 3.0% positive rate for iADL impairment and just 0.4% in the unfiltered validation cohort. This resulted in wide CIs for reported performance. Across data sets, AUROC scores remained high for all models except the deep learning and Bio+Clinical BERT models, which may have been hindered due to the low prevalence of positive training instances (n=234, 3.3% of the training set). Table 3 presents model performance results for iADL impairment detection in the training and 2 evaluation sets. Logistic regression and XGBoost performed best in training set cross-validation AUROC (0.97), while SVM produced the highest AUPRC (0.735). Top predictors of the LASSO model include iADL-related terms such as “cooking,” “shopping,” “management,” “laundry,” “finances,” “meals,” “cleaning,” “food,” and “medication.” These features also tend to have high importance for the remaining models, along with “husband,” “drives,” and “bills.” The XGBoost model’s AUROC scores were best for both data sets (0.995 for filtered validation and 0.991 for unfiltered validation), while the Bio+Clinical BERT model had the highest AUPRC scores for each validation data set (0.551 filtered and 0.568 unfiltered). ROC curves for the filtered validation cohort iADL classification appear in Figure 3, and precision-recall curves are provided in Figure 4.

XSL•FO
RenderX

Table 3. Instrumental activities of daily living classifier performance. Italic values represent the optimal performance in each data set.

Model	AUROC[a] (95% CI)	AUPRC[b] (95% CI)
Filtered cohort		
Training set		
Deep learning	0.948 (0.931-0.964)	0.677 (0.617-0.736)
Bio+Clinical BERT[c]	0.860 (0.797-0.918)	0.730 (0.625-0.826)
Logistic regression	*0.970 (0.958-0.980)*	0.714 (0.656-0.766)
LASSO[d]	0.961 (0.945-0.975)	0.704 (0.644-0.758)
Random forest	0.966 (0.951-0.979)	0.722 (0.668-0.774)
SVM[e]	0.968 (0.955-0.980)	*0.735 (0.679-0.786)*
XGBoost	*0.970 (0.956-0.981)*	0.703 (0.644-0.765)
Validation set		
Deep learning	0.806 (0.243-1.00)	0.551 (0.003-1.00)
Bio+Clinical BERT	0.830 (0.777-0.876)	*0.758 (0.679-0.818)*
Logistic regression	0.952 (0.901-0.998)	0.396 (0.067-0.803)
LASSO	0.978 (0.935-0.999)	0.414 (0.155-0.869)
Random forest	0.941 (0.863-0.998)	0.309 (0.062-0.744)
SVM	0.934 (0.792-0.998)	0.430 (0.125-0.831)
XGBoost	*0.995 (0.988-0.999)*	0.528 (0.255-0.925)
Unfiltered validation cohort		
Deep learning	0.794 (0.191-1.00)	0.568 (0.002-1.00)
Bio+Clinical BERT	0.750 (0.499-1.00)	*0.584 (0.001-1.00)*
Logistic regression	0.960 (0.869-1.00)	0.538 (0.014-1.00)
LASSO	0.986 (0.972-0.999)	0.271 (0.042-0.833)
Random forest	0.945 (0.828-1.00)	0.521 (0.011-1.00)
SVM	0.959 (0.867-1.00)	0.456 (0.022-1.00)
XGBoost	*0.991 (0.972-1.00)*	0.552 (0.050-1.00)

[a]AUROC: area under the receiver operating characteristic curve.
[b]AUPRC: area under the precision-recall curve.
[c]BERT: bidirectional encoder representations from transformers.
[d]LASSO: least absolute shrinkage and selection operator.
[e]SVM: support vector machine.

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curves for instrumental activity of daily living impairment prediction performance on the filtered validation subset. AUC: area under the curve; BERT: bidirectional encoder representations from transformers; LASSO: least absolute shrinkage and selection operator; SVM: support vector machine.

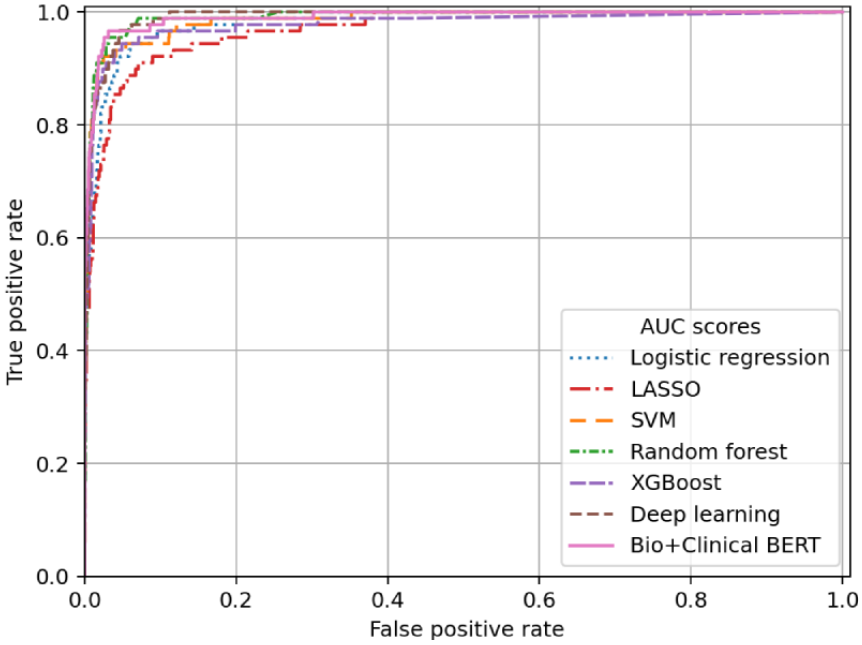

Figure 4. Precision-recall curves for instrumental activity of daily living impairment prediction performance on the filtered validation subset. BERT: bidirectional encoder representations from transformers; LASSO: least absolute shrinkage and selection operator; SVM: support vector machine.

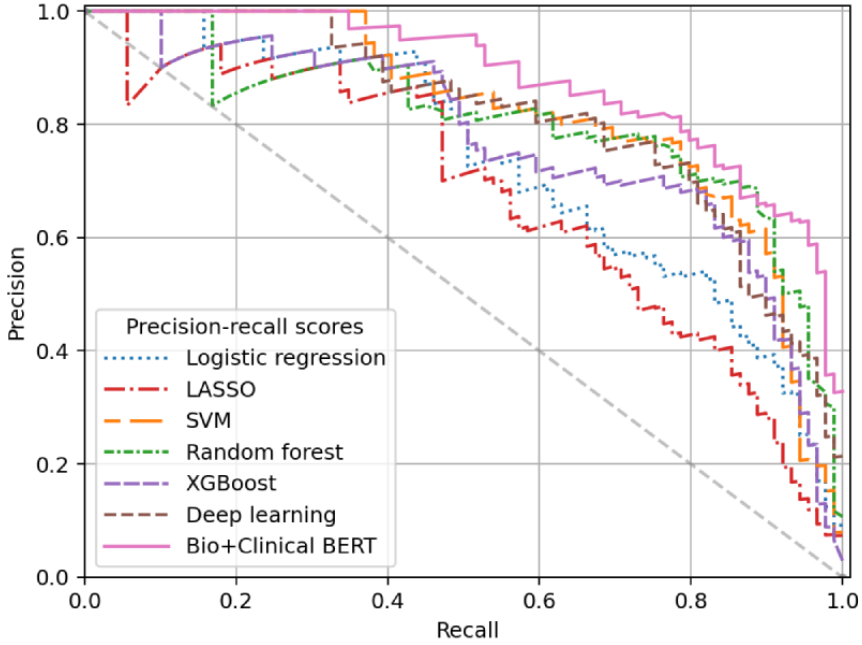

Clinical Implications

Table S4 in Multimedia Appendix 1 provides examples identified by our NLP classifier of various ways that ADL and iADL impairment can present in patient note text. The most prevalent key term categories (Table S1 in Multimedia Appendix 1) across the unfiltered cohort data set appear in Tables S5 and S6 in Multimedia Appendix 1 for ADL and iADL impairment, respectively. As shown by these tables, common ADL- and iADL-related terms appear in just a small fraction of patient note sentences. Using NLP can demonstrably improve the ability to locate evidence of diverse ADL and iADL impairments within patient notes. We performed a key term search on the unfiltered cohort data set using our ADL and iADL lexicons to see how well such an approach identified patient note sentences with evidence of ADL and iADL impairment. Key term search proved to be a sensitive approach—no false negatives occurred for ADL impairment identification and 1 (0.1%) occurred for iADL impairment identification. The false positive rate, however, was greater (45/1000, 4.5% for ADL vs 18/1000, 1.8% for iADL) when compared with the highest performing machine learning models (2/1000, 0.2% for ADL vs 1/1000, 0.1% for iADL).

XSL•FO
RenderX

Discussion

Principal Findings

Among people living with dementia in a cohort using US multicenter EHRs linked with Medicare claims data, we developed and validated NLP models to determine evidence of ADL and iADL impairment. Although the proportion of sentences in clinical notes that contained ADL and iADL information was low, our best-performing models effectively identified relevant sentences, with an AUROC of 0.990 (95% CI 0.984-0.995) for ADL (random forest) and 0.991 (95% CI 0.972-1.00) for iADL (XGBoost).

Identifying people living with dementia who have difficulty with basic ADLs and iADLs is important for clinical care and population health management. The degree of ADL and iADL impairment is associated with dementia severity and progression. The iADLs begin to decline at the mild cognitive impairment stage [22]. Widely used dementia severity scales, such as the Clinical Dementia Rating and Functional Assessment Staging Tool, require assessment of iADLs and ADLs. As the ability to perform iADLs and ADLs declines with the progression of dementia [23], early detection of iADL and ADL impairment can lead to early rehabilitation to preserve their daily function. In acute hospital care settings, assessment of iADL and ADL function could help identify those at risk of loss of independence and arrange care transition interventions [24]. Moreover, ADL dependence is a risk factor for falls in community-dwelling adults with dementia [25]. Similarly, iADL impairment is predictive of 30-day readmission and can be helpful in identifying high-risk patients for early interventions [26].

Despite the importance of ADL and iADL assessment, documentation of this information is neither standardized nor available in most EHR and claim data. As a result, measures of ADL and iADL impairment are not included in prediction models of readmission. Recently, an effort to use machine learning methods to extract ADLs and iADLs information from EHR free-text notes or reports showed a potential to improve risk prediction or clinical decision support systems. The iADL impairment identified using machine learning was predictive of 30-day readmission [26]. Similarly, geriatric syndromes that are not documented in structured EHR data can be further identified in unstructured clinical notes in the EHR using NLP algorithms [27]. It has also been shown that frailty described in clinical notes was associated with greater health care use [28].

Comparison to Previous Work

Our work adds to previous research by showing the utility of NLP and machine learning algorithms to identify ADL and iADL information in unstructured EHR data with high accuracy for older adults with dementia. ADL and iADL impairment information from clinical notes of people living with dementia can help researchers identify medically stable and ambulatory older adults with dementia and specific functional levels who can be enrolled in clinical trials. In addition, information on ADL and iADL function is an important confounder in administrative claims-based studies of medical interventions in the fields of geriatrics, neurology, rehabilitation medicine, and family medicine. Combining our NLP approach with other data from the EHR could further improve the validity of EHR-based analysis. The extent of confounding and further adjustment has become possible through EHR-claims linkage in clinical research networks [8,29].

Study Limitations

Our study has several limitations. Our model is based on a US metropolitan academic care delivery network. Because of the subjectivity of self-reported information, variations in documentation conventions, and different demographic or cultural backgrounds of the study population, it is unclear if our findings can be generalizable to other health care systems. Additionally, our findings were based on a relatively small cohort with 441 patients in the filtered set and 80 in the unfiltered set, so validation in a larger cohort, preferably with a different demographic profile, is needed to confirm generalizability. Additionally, the measure and conceptualization of iADLs can be complex due to the differences between cultural norms and gender roles. For example, women have greater health-related iADL limitations than men [30]. Cross-national variations in ADL and iADL impairment may reflect item-response bias due to culture-based gender norms rather than actual differences in disability levels [31].

Conclusion

In conclusion, we have developed models to determine ADL and iADL impairment among US Medicare beneficiaries using NLP and EHR unstructured data. Because ADL and iADL are typically not available as structured EHR data, our models can enhance researchers' ability to identify subgroups among people living with dementia according to their ADL and iADL dependency. Our models can improve patient phenotyping and confounding adjustment in EHR data that are used in comparative effectiveness and safety research.

Acknowledgments

This study was funded by the National Institute on Aging (1RF1AG063381-01) and the National Library of Medicine (1R01LM013204-01A1).

Data Availability

The data sets generated and analyzed during this study are not publicly available as they contain protected health information from patient health records.

XSL•FO
RenderX

Authors' Contributions

JL was responsible for methodology, software, validation, formal analysis, investigation, data curation, writing the original draft, reviewing, editing, and visualization. DHK was involved in conceptualization, methodology, investigation, data curation, writing the original draft, reviewing, editing, and supervision. MM contributed to formal analysis, data curation, writing, reviewing, and editing. KYW was responsible for writing the original draft, reviewing, and editing. LGB, SBL, and CY were involved in data curation and project administration. HZ conducted data curation, writing, reviewing, editing, and project administration. LZ was responsible for conceptualization, methodology, investigation, resources, writing, reviewing, editing, and supervision. KJL was involved in conceptualization, methodology, investigation, resources, data curation, writing the original draft, reviewing, editing, supervision, and funding acquisition.

Conflicts of Interest

DHK has been supported by grants from the National Institute on Aging of the National Institutes of Health for unrelated work. He received a personal fee from Alosa Health and VillageMD for unrelated work. The funder had no role in the design, collection, analysis, interpretation of the data, or the decision to submit the manuscript for publication.

Multimedia Appendix 1
ADL-and iADL-related key term categories, ADL model parameters, iADL model parameters, excerpts of NLP-identified patient notes with evidence of ADL/iADL impairment, prevalence of ADL term groups across positive-labeled cases of ADL impairment in the unfiltered data set, and prevalence of iADL term groups across positive-labeled cases of iADL impairment in the unfiltered data set.
[DOCX File , 26 KB - jmir_v26i1e47739_app1.docx]

References

1. 2023 Alzheimer's disease facts and figures. The Alzheimer's Association. URL: https://www.alz.org/media/Documents/alzheimers-facts-and-figures.pdf [accessed 2024-01-11]
2. Edemekong PF, Bomgaars DL, Sukumaran S, Caroline C. Activities of Daily Living. Treasure Island, FL: StatPearls Publishing; 2023.
3. Hung CH, Hung GU, Wei CY, Tzeng RC, Chiu PY. Function-based dementia severity assessment for vascular cognitive impairment. J Formos Med Assoc 2021;120(1 Pt 2):533-541 [FREE Full text] [doi: 10.1016/j.jfma.2020.07.001] [Medline: 32653387]
4. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: observational study. J Med Internet Res 2020;22(10):e22635 [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
5. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 2019;79(21):5463-5470 [FREE Full text] [doi: 10.1158/0008-5472.CAN-19-0579] [Medline: 31395609]
6. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a research patient data repository. AMIA Annu Symp Proc 2006;2006:1044 [FREE Full text] [Medline: 17238663]
7. Lin KJ, Glynn RJ, Singer DE, Murphy SN, Lii J, Schneeweiss S. Out-of-system care and recording of patient characteristics critical for comparative effectiveness research. Epidemiology 2018;29(3):356-363 [FREE Full text] [doi: 10.1097/EDE.0000000000000794] [Medline: 29283893]
8. Lin KJ, Rosenthal GE, Murphy SN, Mandl KD, Jin Y, Glynn RJ, et al. External validation of an algorithm to identify patients with high data-completeness in electronic health records for comparative effectiveness research. Clin Epidemiol 2020;12:133-141 [FREE Full text] [doi: 10.2147/CLEP.S232540] [Medline: 32099479]
9. Lin KJ, Singer DE, Glynn RJ, Blackley S, Zhou L, Liu J, et al. Prediction score for anticoagulation control quality among older adults. J Am Heart Assoc 2017;6(10):e006814 [FREE Full text] [doi: 10.1161/JAHA.117.006814] [Medline: 28982676]
10. Wilkinson T, Ly A, Schnier C, Rannikmäe K, Bush K, Brayne C, et al. Identifying dementia cases with routinely collected health data: a systematic review. Alzheimers Dement 2018;14(8):1038-1051 [FREE Full text] [doi: 10.1016/j.jalz.2018.02.016] [Medline: 29621480]
11. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc 2011;2011:1639-1648 [FREE Full text] [Medline: 22195230]
12. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 2018;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
13. Breiman L. Random forests. Mach Learn 2001;45(1):5-32 [FREE Full text]
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]

XSL•FO
RenderX

15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
16. Tf-idf weighting. The Stanford Natural Language Processing Group. URL: https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html [accessed 2022-09-21]
17. Yang J, Wang L, Phadke NA, Wickner PG, Mancini CM, Blumenthal KG, et al. Development and validation of a deep learning model for detection of allergic reactions using safety event reports across hospitals. JAMA Netw Open 2020;3(11):e2022836 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22836] [Medline: 33196805]
18. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. ArXive Preprint posted online on April 6, 2019 [FREE Full text] [doi: 10.18653/v1/w19-1909]
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
21. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. ArXive Preprint posted online on October 9, 2019 [FREE Full text] [doi: 10.18653/v1/2020.emnlp-demos.6]
22. Jekel K, Damian M, Wattmo C, Hausner L, Bullock R, Connelly PJ, et al. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. Alzheimers Res Ther 2015;7(1):17 [FREE Full text] [doi: 10.1186/s13195-015-0099-0] [Medline: 25815063]
23. Liu KPY, Chan CCH, Chu MML, Ng TYL, Chu LW, Hui FSL, et al. Activities of daily living performance in dementia. Acta Neurol Scand 2007;116(2):91-95. [doi: 10.1111/j.1600-0404.2007.00800.x] [Medline: 17661793]
24. Zurlo A, Zuliani G. Management of care transition and hospital discharge. Aging Clin Exp Res 2018;30(3):263-270. [doi: 10.1007/s40520-017-0885-6] [Medline: 29313293]
25. Salvà A, Roqué M, Rojano X, Inzitari M, Andrieu S, Schiffrin EJ, et al. Falls and risk factors for falls in community-dwelling adults with dementia (NutriAlz trial). Alzheimer Dis Assoc Disord 2012;26(1):74-80. [doi: 10.1097/WAD.0b013e318215ca90] [Medline: 22354139]
26. Schiltz NK, Dolansky MA, Warner DF, Stange KC, Gravenstein S, Koroukian SM. Impact of instrumental activities of daily living limitations on hospital readmission: an observational study using machine learning. J Gen Intern Med 2020;35(10):2865-2872 [FREE Full text] [doi: 10.1007/s11606-020-05982-0] [Medline: 32728960]
27. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc 2018;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
28. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. BMC Geriatr 2017;17(1):248 [FREE Full text] [doi: 10.1186/s12877-017-0645-7] [Medline: 29070036]
29. Lin KJ, Schneeweiss S. Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. Clin Pharmacol Ther 2016;100(2):147-159. [doi: 10.1002/cpt.359] [Medline: 26916672]
30. Sheehan CM, Tucker-Drob EM. Gendered expectations distort male-female differences in instrumental activities of daily living in later adulthood. J Gerontol B Psychol Sci Soc Sci 2019;74(4):715-723 [FREE Full text] [doi: 10.1093/geronb/gbw209] [Medline: 28158847]
31. Jang SN, Kawachi I. Why do older Korean adults respond differently to activities of daily living and instrumental activities of daily living? A differential item functioning analysis. Ann Geriatr Med Res 2019;23(4):197-203 [FREE Full text] [doi: 10.4235/agmr.19.0047] [Medline: 32743312]

Abbreviations

ADL: activity of daily living
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
EHR: electronic health record
iADL: instrumental activity of daily living
LASSO: least absolute shrinkage and selection operator
MGB: Mass General Brigham
MIMIC: Medical Information Mart for Intensive Care

XSL•FO
RenderX

NLP: natural language processing
ROC: receiver operating characteristic
RPDR: Research Patient Data Repository
SVM: support vector machine

Edited by T Leung; submitted 30.03.23; peer-reviewed by M Elbattah, H Shah, Y Zheng; comments to author 17.06.23; revised version received 30.06.23; accepted 31.10.23; published 13.02.24.

Please cite as:
Laurentiev J, Kim DH, Mahesri M, Wang KY, Bessette LG, York C, Zakoul H, Lee SB, Zhou L, Lin KJ
Identifying Functional Status Impairment in People Living With Dementia Through Natural Language Processing of Clinical Documents: Cross-Sectional Study
J Med Internet Res 2024;26:e47739
URL: https://www.jmir.org/2024/1/e47739
doi:10.2196/47739
PMID:38349732

©John Laurentiev, Dae Hyun Kim, Mufaddal Mahesri, Kuan-Yuan Wang, Lily G Bessette, Cassandra York, Heidi Zakoul, Su Been Lee, Li Zhou, Kueiyu Joshua Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adverse Event Signal Detection Using Patients' Concerns in Pharmaceutical Care Records: Evaluation of Deep Learning Models

Satoshi Nishioka[1], PhD; Satoshi Watabe[1], BSc; Yuki Yanagisawa[1], PhD; Kyoko Sayama[1], MSc; Hayato Kizaki[1], MSc; Shungo Imai[1], PhD; Mitsuhiro Someya[2], BSc; Ryoo Taniguchi[2], PhD; Shuntaro Yada[3], PhD; Eiji Aramaki[3], PhD; Satoko Hori[1], PhD

[1]Division of Drug Informatics, Keio University Faculty of Pharmacy, Tokyo, Japan

[2]Nakajima Pharmacy, Hokkaido, Japan

[3]Nara Institute of Science and Technology, Nara, Japan

Corresponding Author:
Satoko Hori, PhD
Division of Drug Informatics
Keio University Faculty of Pharmacy
1-5-30 Shibakoen
Minato-ku
Tokyo, 105-8512
Japan
Phone: 81 3 5400 2650
Email: satokoh@keio.jp

Abstract

Background: Early detection of adverse events and their management are crucial to improving anticancer treatment outcomes, and listening to patients' subjective opinions (patients' voices) can make a major contribution to improving safety management. Recent progress in deep learning technologies has enabled various new approaches for the evaluation of safety-related events based on patient-generated text data, but few studies have focused on the improvement of real-time safety monitoring for individual patients. In addition, no study has yet been performed to validate deep learning models for screening patients' narratives for clinically important adverse event signals that require medical intervention. In our previous work, novel deep learning models have been developed to detect adverse event signals for hand-foot syndrome or adverse events limiting patients' daily lives from the authored narratives of patients with cancer, aiming ultimately to use them as safety monitoring support tools for individual patients.

Objective: This study was designed to evaluate whether our deep learning models can screen clinically important adverse event signals that require intervention by health care professionals. The applicability of our deep learning models to data on patients' concerns at pharmacies was also assessed.

Methods: Pharmaceutical care records at community pharmacies were used for the evaluation of our deep learning models. The records followed the SOAP format, consisting of subjective (S), objective (O), assessment (A), and plan (P) columns. Because of the unique combination of patients' concerns in the S column and the professional records of the pharmacists, this was considered a suitable data for the present purpose. Our deep learning models were applied to the S records of patients with cancer, and the extracted adverse event signals were assessed in relation to medical actions and prescribed drugs.

Results: From 30,784 S records of 2479 patients with at least 1 prescription of anticancer drugs, our deep learning models extracted true adverse event signals with more than 80% accuracy for both hand-foot syndrome (n=152, 91%) and adverse events limiting patients' daily lives (n=157, 80.1%). The deep learning models were also able to screen adverse event signals that require medical intervention by health care providers. The extracted adverse event signals could reflect the side effects of anticancer drugs used by the patients based on analysis of prescribed anticancer drugs. "Pain or numbness" (n=57, 36.3%), "fever" (n=46, 29.3%), and "nausea" (n=40, 25.5%) were common symptoms out of the true adverse event signals identified by the model for adverse events limiting patients' daily lives.

XSL•FO
RenderX

Conclusions: Our deep learning models were able to screen clinically important adverse event signals that require intervention for symptoms. It was also confirmed that these deep learning models could be applied to patients' subjective information recorded in pharmaceutical care records accumulated during pharmacists' daily work.

(J Med Internet Res 2024;26:e55794) doi:10.2196/55794

KEYWORDS

cancer; anticancer drug; adverse event; side effect; patient-reported outcome; patients' voice; patient-oriented; patient narrative; natural language processing; deep learning; pharmaceutical care record; SOAP

Introduction

Increasing numbers of people are expected to develop cancers in our aging society [1-3]. Thus, there is increasing interest in how to detect and manage the side effects of anticancer therapies in order to improve treatment regimens and patients' quality of life [4-8]. The primary approaches for side effect management are "early signal detection and early intervention" [9-11]. Thus, more efficient approaches for this purpose are needed.

It has been recognized that patients' voices concerning adverse events represent an important source of information. Several studies have indicated that the number, severity, and time of occurrence of adverse events might be underevaluated by physicians [12-15]. Thus, patient-reported outcomes (PROs) have recently received more attention in the drug evaluation process, reflecting patients' real voices. Various kinds of PRO measures have been developed and investigated in different disease populations [16,17]. Health care authorities have also encouraged the pharmaceutical industry to use PROs for drug evaluation [18,19], and it is becoming more common to take PRO assessment results into consideration for drug marketing approval [20,21]. Similar trends can be seen in the clinical management of individual patients. Thus, health care professionals have an interest in understanding how to appropriately gather patients' concerns in order to improve safety management and clinical decisions [22-24].

The applications of deep learning for natural language processing have expanded dramatically in recent years [25]. Since the development of a high-performance deep learning model in 2018 [26], attempts to apply cutting-edge deep learning models to various kinds of patient-generated text data for the evaluation of safety events or the analysis of unscalable subjective information from patients have been accelerating [27-31]. Most studies have been conducted to use patients' narrative data for pharmacovigilance [27,32-35], while few have been aimed at improvement of real-time safety monitoring for individual patients. In addition, there have been some studies on adverse event severity grading based on health care records [36-39], but none has yet aimed to extract clinically important adverse event signals that require medical intervention from patients' narratives. It is important to know whether deep learning models could contribute to the detection of such important adverse event signals from concern texts generated by individual patients.

To address this question, we have developed deep learning models to detect adverse event signals from individual patients with cancer based on patients' blog articles in online communities, following other types of natural language processing–related previous work [40,41]. One deep learning model focused on the specific symptom of hand-foot syndrome (HFS), which is one of the typical side effects of anticancer treatments [42], and another focused on a broad range of adverse events that impact patients' activities of daily living [43]. We showed that our models can provide good performance scores in targeting adverse event signals. However, the evaluation relied on patients' narratives from the patients' blog data used for deep learning model training, so further evaluation is needed to ensure the validity and applicability of the models to other texts regarding patients' concerns. In addition, the blog data source did not contain medical information, so it was not feasible to assess whether the models could contribute to the extraction of clinically important adverse event signals.

To address these challenges, we focused on pharmaceutical care records written by pharmacists at community pharmacies. The gold standard format for pharmaceutical care records in Japan is the SOAP (subjective, objective, assessment, plan)-based document that follows the "problem-oriented system" concept proposed by Weed [44] in 1968. Pharmacists track patients' subjective concerns in the S column, provide objective information or observations in the O column, give their assessment from the pharmacist perspective in the A column, and suggest a plan for moving forward in the P column [45,46]. We considered that SOAP-based pharmaceutical care records could be a unique data source suitable for further evaluation of our deep learning models because they contain both patients' concerns and professional health care records by pharmacists, including the medication prescription history with time stamps. Therefore, this study was designed to assess whether our deep learning models could extract clinically important adverse event signals that require intervention by medical professionals from these records. We also aimed to evaluate the characteristics of the models when applied to patients' subjective information noted in the pharmaceutical care records, as there have been only a few studies on the application of deep learning models to patients' concerns recorded during pharmacists' daily work [47-49].

Here, we report the results of applying our deep learning models to patients' concern text data in pharmaceutical care records, focusing on patients receiving anticancer treatment.

Methods

Data Source

The original data source was 2,276,494 pharmaceutical care records for 303,179 patients, created from April 2020 to

XSL•FO
RenderX

December 2021 at community pharmacies belonging to the Nakajima Pharmacy Group in Japan [50]. To focus on patients with cancer, records of patients with at least 1 prescription for an anticancer drug were retrieved by sorting individual drug codes (YJ codes) used in Japan (YJ codes starting with 42 refer to anticancer drugs). Records in the S column (ie, S records) were collected from the patients with cancer as the text data of patients' concerns for deep learning model analysis.

Deep Learning Models

The deep learning models used for this research were those that we constructed based on patients' narratives in blog articles posted in an online community and that showed the best performance score in each task in our previous work (ie, a Bidirectional Encoder Representations From Transformers [BERT]–based model for HFS signal extraction [42] and a T5-based model for adverse event signal extraction [43]). BERT [26] and T5 [51] both belong to a type of deep learning model that has recently shown high performance in several studies [29,52]. Hereafter, we refer to the deep learning model for HFS signals as the HFS model, the model for any adverse event signals as All AE (ie, all or any adverse events) model, and the model for adverse event signals limited to patients' activities of daily living as the AE-L (adverse events limiting patients' daily lives) model. It was also confirmed that these deep learning models showed similar or higher performance scores for the HFS, All AE, or AE-L identification tasks using 1000 S records randomly extracted from the data source of this study compared to the values obtained in our previous work [42,43] (the performance scores of sentence-level tasks from our previous work are comparable, as the mean number of words in the sentences in the data source in our previous work was 32.7 [SD 33.9], which is close to that of the S records used in this study, 38.8 [SD 29.4]). The method and results of the performance-level check are described in detail in Multimedia Appendix 1 [42,43]. We applied the deep learning models to all text data in this study without any adjustment in setting parameters from those used in constructing them based on patient-authored texts in our previous work [42,43].

Evaluation of Extracted S Records by the Deep Learning Models

In this study, we focused on the evaluation of S records that our deep learning models extracted as HFS or AE-L positive. Each positive S record was assessed as if it was a true adverse event signal, a sort of adverse event symptom, whether or not an intervention was made by health care professionals. We also investigated the kind of anticancer treatment prescription in connection with each adverse event signal identified in S records.

To assess whether an extracted positive S record was a true adverse event signal, we used the same annotation guidelines as in our previous work [43]. In brief, each S record was treated as an "adverse event signal" if any untoward medical occurrence happened to the patient, regardless of the cause. For the AE-L model only, if a positive S record was confirmed as an adverse event signal, it was further categorized into 1 or more of the following adverse event symptoms: "fatigue," "nausea," "vomiting," "diarrhea," "constipation," "appetite loss," "pain or numbness," "rash or itchy," "hair loss," "menstrual irregularity," "fever," "taste disorder," "dizziness," "sleep disorder," "edema," or "others."

For the assessment of interventions by health care professionals and anticancer treatment prescriptions, information from the O, A, and P columns and drug prescription history in the data source were investigated for the extracted positive S records. The interventions by health care professionals were categorized in any of the following: "adding symptomatic treatment for the adverse event signal," "dose reduction or discontinuation of causative anticancer treatment," "consultation with physician," "others," or "no intervention (ie, just following up the adverse event signal)." The actions categorized in "others" were further evaluated individually. For this assessment, we also randomly extracted 200 S records and evaluated them in the same way for comparison with the results from the deep learning model. Prescription history of anticancer treatment was analyzed by primary category of mechanism of action (MoA) with subcategories if applicable (eg, target molecule for kinase inhibitors).

Applicability Check to Other Text Data Including Patients' Concerns

To check the applicability of our deep learning models to data from a different source, interview transcripts from patients with cancer were also evaluated. The interview transcripts were created by the Database of Individual Patient Experiences-Japan (DIPEx-Japan) [53]. DIPEx-Japan divides the interview transcripts into sections for each topic, such as "onset of disease" and "treatment," and posts the processed texts on its website. Processing is conducted by accredited researchers based on qualitative research methods established by the University of Oxford [54]. In this study, interview text data created from interviews with 52 patients with breast cancer conducted from January 2008 to October 2018 were used to assess whether our deep learning models can extract adverse event signals from this source. In total, 508 interview transcripts were included with the approval of DIPEx-Japan.

Ethical Considerations

This study was conducted with anonymized data following approval by the ethics committee of the Keio University Faculty of Pharmacy (210914-1 and 230217-1) and in accordance with relevant guidelines and regulations and the Declaration of Helsinki. Informed consent specific to this study was waived due to the retrospective observational design of the study with the approval of the ethics committee of the Keio University Faculty of Pharmacy. To respect the will of each individual stakeholder, however, we provided patients and pharmacists of the pharmacy group with an opportunity to refuse the sharing of their pharmaceutical care records by posting an overview of this study at each pharmacy store or on their web page regarding the analysis using pharmaceutical care records. Interview transcripts from DIPEx-Japan were provided through a data sharing arrangement for using narrative data for research and education. Consent for interview transcription and its sharing from DIPEx-Japan was obtained from the participants when the interviews were recorded.

XSL•FO
RenderX

Results

Data Set

From the original data source of 2,180,902 pharmaceutical care records for 291,150 patients, S records written by pharmacists for patients with a history of at least 1 prescription of an anticancer drug were extracted. This yielded 30,784 S records for 2479 patients with cancer (Table 1). The mean and median number of words in the S records were 38.8 (SD 29.4) and 32 (IQR 20-50), respectively. We applied our deep learning models, HFS, All AE, and AE-L, to these 30,784 S records for the evaluation of the deep learning models for adverse event signal detection.

For interview transcripts created by DIPEx-Japan, the mean and median number of words were 428.9 (SD 160.9) and 416 (IQR 308-526), respectively, in the 508 transcripts for 52 patients with breast cancer.

Table 1. Characteristics of the pharmaceutical care records.

Item	Values
Duration	April 2020 to December 2021 (1 year and 9 months)
Patients	
Total, n	291,150
At least 1 anticancer drug prescription, n (%)	2479 (0.9)
Pharmaceutical care records	
SOAP[a] records for all patients, n	2,180,902
S records for patients with at least 1 anticancer drug prescription, n (%)	30,784 (1.4)
Words in S[b] records for patients with at least 1 anticancer drug prescription	
Mean (SD)	38.8 (29.4)
Median (IQR)	32 (20-50)

[a]SOAP: subjective, objective, assessment, plan.

[b]S: subjective.

Application of the HFS Model

Overview

First, we applied the HFS model to the S records for patients with cancer. The BERT-based model was used for this research as it showed the best performance score in our previous work [42].

S Records Extracted as HFS Positive

The S records extracted as HFS positive by the HFS model (Table 2) amounted to 167 (0.5%) records for 119 (4.8%) patients. A majority of the patients had 1 HFS-positive record in their S records (n=91, 76.5%), while 2 patients had as many as 6 (1.7%) HFS-positive records. When we examined whether the extracted S records were true adverse event signals or not, 152 records were confirmed to be adverse event signals, while the other 15 records were false-positives. All the false-positive S records were descriptions about the absence of symptoms or confirmation of improving condition (eg, “no diarrhea, mouth ulcers, or limb pain so far” or “the skin on the soles of my feet has calmed down a lot with this ointment”). Some examples of S records that were predicted as HFS positive by the model are shown in Table S1 in Multimedia Appendix 2.

The same examination was conducted with interview transcripts from DIPEx-Japan. Only 1 (0.2%) transcript was extracted as HFS positive by the HFS model, and it was a true adverse event signal (100%). The actual transcript extracted as HFS positive is shown in Table S2 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. S[a] records extracted as HFS[b] positive by the HFS model.

Item	Values
Positive patients (n=2479), n (%)	119 (4.8)
Positive S records (n=30,784), n (%)	167 (0.5)
Positive S records per patient, mean (SD)	1.40 (0.92)
Breakdown of patient numbers by number of positive S records, n (%)	
1 record	91 (76.5)
2 records	18 (15.1)
3 records	4 (3.4)
4 records	4 (3.4)
6 records	2 (1.7)
Adverse event signal, n (%)	
Yes	152 (91)
No (ie, false-positive[c])	15 (9)

[a]S: subjective.

[b]HFS: hand-foot syndrome.

[c]All false-positive S records were denial of symptoms or confirmation of improving condition.

Interventions by Health Care Professionals

The 167 S records extracted as HFS positive as well as 200 randomly selected records were checked for interventions by health care professionals (Figure 1). The proportion showing any action by health care professionals was 64.1% for 167 HFS-positive S records compared to 13% for the 200 random S records. Among the actions taken for HFS positives, "adding symptomatic treatment" was the most common, accounting for around half (n=79, 47.3%), followed by "other" (n=18, 10.8%). Most "other" actions were educational guidance from pharmacists, such as instructions on moisturizing, nail care, or application of ointment and advice on daily living (eg, "avoid tight socks").

Figure 1. Interventions in cases with HFS-positive S records and in random cases. HFS: hand-foot syndrome; HP: health care professional.

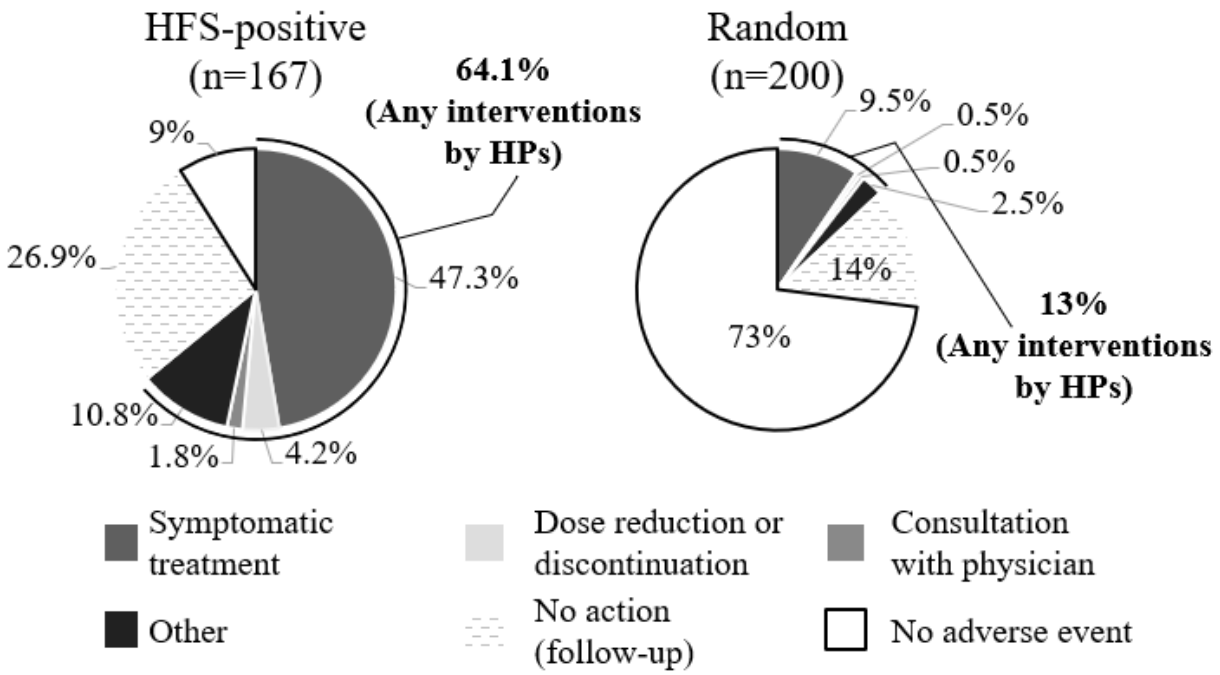

Anticancer Drugs Prescribed

The types of anticancer drugs prescribed for HFS-positive patients are summarized based on the prescription histories in Table 3. For the 152 adverse event signals identified by the HFS model in the previous section, the most common MoA class of anticancer drugs used for the patients was antimetabolite (n=62, 40.8%), specifically fluoropyrimidines (n=59, 38.8%). Kinase

XSL•FO
RenderX

inhibitors were next (n=49, 32.2%), with epidermal growth factor receptor (EGFR) inhibitors and multikinase inhibitors as major subgroups (n=28, 18.4% and n=14, 9.2%, respectively). The third and fourth most common MoAs were aromatase inhibitors (n=24, 15.8%) and antiandrogen or estrogen drugs (n=7, 4.6% each) for hormone therapy.

Table 3. MoA classes of anticancer drugs prescribed for patients with adverse event signals identified by the HFS model (n=152).

Anticancer drugs	Adverse event signals identified by HFS model, n (%)
Antimetabolites	
Overall	62 (40.8)
Fluoropyrimidine	59 (38.8)
Folate analog	3 (2)
Kinase inhibitors	
Overall	49 (32.2)
EGFR[a]	28 (18.4)
Multi	14 (9.2)
VEGF[b]	6 (3.9)
HER2[c]	1 (0.7)
Aromatase inhibitors	24 (15.8)
Antiandrogens	7 (4.6)
Antiestrogens	7 (4.6)
CDK4/6[d] inhibitors	3 (2)
Alkylating agents	1 (0.7)

[a]EGFR: epidermal growth factor receptor.

[b]VEGF: vascular endothelial growth factor.

[c]HER2: human epidermal growth factor receptor-2.

[d]CDK4/6: cyclin-dependent kinase 4/6.

Application of the All AE or AE-L model

Overview

The All AE and AE-L models were also applied to the same S records for patients with cancer. The T5-based model was used for this research as it gave the best performance score in our previous work [43].

S Records Extracted as All AE or AE-L positive

The numbers of S records extracted as positive were 7604 (24.7%) for 1797 patients and 196 (0.6%) for 142 patients for All AE and AE-L, respectively. In the case of All AE, patients tended to have multiple adverse event positives in their S records (n=1315, 73.2% of patients had at least 2 positives). In the case of AE-L, most patients had only 1 AE-L positive (n=104, 73.2%), and the largest number of AE-L positives for 1 patient was 4 (2.8%; Table 4).

We focused on AE-L evaluation due to its greater importance from a medical viewpoint and lower workload for manual assessment, considering the number of positive S records. Of the 197 AE-L–positive S records, it was confirmed that 157 (80.1%) records accurately extracted adverse event signals, while 39 (19.9%) records were false-positives that did not include any adverse event signals (Table 4). The contents of the 39 false-positives were all descriptions about the absence of symptoms or confirmation of improving condition, showing a similar tendency to the HFS false-positives (eg, “The diarrhea has calmed down so far. Symptoms in hands and feet are currently fine” and “No symptoms for the following: upset in stomach, diarrhea, nausea, abdominal pain, abdominal pain or stomach cramps, constipation”). Examples of S records that were predicted as AE-L positive are shown in Table S3 in Multimedia Appendix 2.

The deep learning models were also applied to interview transcripts from DIPEx-Japan in the same manner. The deep learning models identified 84 (16.5%) and 18 (3.5%) transcripts as All AE or AE-L positive, respectively. Of the 84 All AE–positive transcripts, 73 (86.9%) were true adverse event signals. The false-positives of All AE (n=11, 13.1%) were categorized into any of the following 3 types: explanations about the disease or its prognosis, stories when their cancer was discovered, or emotional changes that did not include clear adverse event mentions. With regard to AE-L, all the 18 (100%) positives were true adverse event signals (Table S4 in Multimedia Appendix 2). Examples of actual transcripts extracted as All AE or AE-L positive are shown in Table S5 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 4. S[a] records extracted as positive by the All AE[b] or AE-L[c] model.

Item	Values
All AE positive	
Positive patients (n=2479), n (%)	1797 (72.5)
Positive S records (n=30,784), n (%)	7604 (24.7)
Positive S records per patient, mean (SD)	4.23 (4.17)
Breakdown of patient numbers by number of positive S records, n (%)	
1 record	482 (26.8)
2-5 records	849 (47.3)
6-10 records	337 (18.8)
11-20 records	114 (6.3)
>20 records	15 (0.8)
AE-L positive	
Positive patients (n=2479), n (%)	142 (5.7)
Positive S records (n=30,784), n (%)	196 (0.6)
Positive S records per patient, mean (SD)	1.38 (0.72)
Breakdown of patient numbers by number of positive S records, n (%)	
1 record	104 (73.2)
2 records	26 (18.3)
3 records	8 (5.6)
4 records	4 (2.8)
Adverse event signal, n (%)	
Yes	157 (80.1)
No (ie, false-positive[d])	39 (19.9)

[a]S: subjective.

[b]All AE: all (or any of) adverse event.

[c]AE-L: adverse events limiting patients' daily lives.

[d]All false-positive S records were denial of symptoms or confirmation of improving condition.

Interventions by Health Care Professionals

Whether or not interventions were made by health care professionals was investigated for the 196 AE-L–positive S records. As in the HFS model evaluation, data from 200 randomly selected S records were used for comparison (Figure 2). In total, 91 (46.4%) records in the 196 AE-L–positive records were accompanied by an intervention, while the corresponding figure in the 200 random records was 26 (13%) records. The most common action in response to adverse event signals identified by the AE-L model was "adding symptomatic treatment" (n=71, 36.2%), followed by "other" (n=11, 5.6%). "Other" included educational guidance from pharmacists, inquiries from pharmacists to physicians, or recommendations for patients to visit a doctor.

XSL•FO
RenderX

Figure 2. Interventions in cases with AE-L-positive S records and in random cases. AE-L: adverse events limiting patients' daily lives; HP: health care professional.

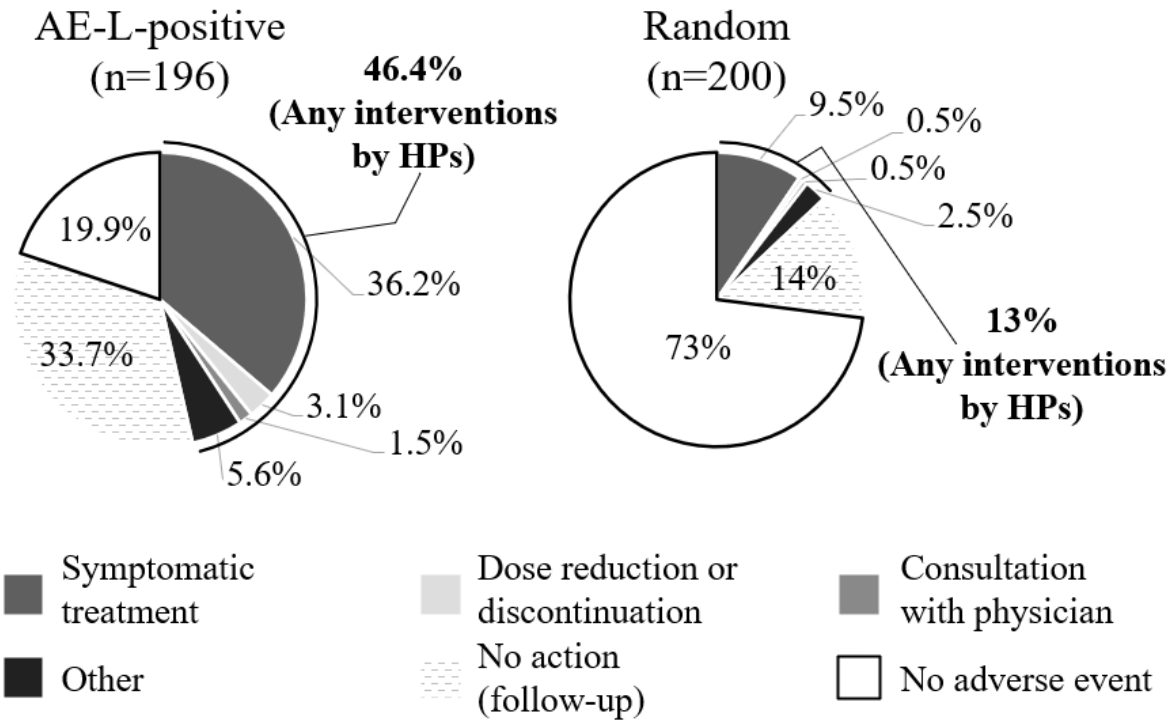

Anticancer Drugs Prescribed

The types of anticancer drugs prescribed for patients with adverse event signals identified by the AE-L model were summarized based on the prescription histories (Table 5). In connection with the 157 adverse event signals, the most common MoA of the prescribed anticancer drug was antimetabolite (n=62, 39.5%) and fluoropyrimidine (n=53, 33.8%), which accounted for the majority. Kinase inhibitor (n=31, 19.7%) was the next largest category with multikinase inhibitor (n=14, 8.9%) as the major subgroup. These were followed by antiandrogen (n=27, 17.2%), antiestrogen (n=10, 6.4%), and aromatase inhibitor (n=10, 6.4%) for hormone therapy.

XSL•FO
RenderX

Table 5. MoA classes of anticancer drugs prescribed for patients with adverse event signals identified by the AE-L model (n=157).

Anticancer drugs	Adverse event signals identified by AE-L model, n (%)
Antimetabolites	
Overall	62 (39.5)
Fluoropyrimidine	53 (33.8)
Trifluridine	5 (3.2)
Purine analog	3 (1.9)
Folate analog	1 (0.6)
Ribonucelotide reductase inhibitor	1 (0.6)
Kinase inhibitors	
Overall	31 (19.7)
Multi	14 (8.9)
EGFR[a]	9 (5.7)
JAK[b]	3 (1.9)
VEGF[c]	2 (1.3)
BTK[d]	2 (1.3)
FLT3[e]	1 (0.6)
Antiandrogens	27 (17.2)
Antiestrogens	10 (6.4)
Aromatase inhibitors	10 (6.4)
Topoisomerase inhibitors	6 (3.8)
PARP[f] inhibitors	6 (3.8)
CDK4/6[g] inhibitors	4 (2.5)
Alkylating agents	1 (0.6)
Anti-CD20[h] antibodies	1 (0.6)

[a]EGFR: epidermal growth factor receptor.
[b]JAK: janus kinase.
[c]VEGF: vascular endothelial growth factor.
[c]HER2: human epidermal growth factor receptor-2.
[d]BTK: bruton tyrosine kinase.
[e]FLT3: FMS-like tyrosine kinase-3.
[f]PARP: poly-ADP ribose polymerase.
[g]CDK4/6: cyclin-dependent kinase 4/6.
[h]CD20: cluster of differentiation 20.

Adverse Event Symptoms

For the 157 adverse event signals identified by the AE-L model, the symptoms were categorized according to the predefined guideline in our previous work [43]. “Pain or numbness” (n=57, 36.3%) accounted for the largest proportion followed by “fever” (n=46, 29.3%) and “nausea” (n=40, 25.5%; Table 6). Symptoms classified as “others” included chills, tinnitus, running tears, dry or peeling skin, and frequent urination. When comparing the proportion of the symptoms associated with or without interventions by health care professionals, a trend toward a greater proportion of interventions was observed in “fever,” “nausea,” “diarrhea,” “constipation,” “vomiting,” and “edema” (Figure 3, black boxes). On the other hand, a smaller proportion was observed in “pain or numbness,” “fatigue,” “appetite loss,” “rash or itchy,” “taste disorder,” and “dizziness” (Figure 3, gray boxes).

XSL•FO
RenderX

Table 6. Symptoms of adverse event signals identified by the AE-L model from subjective records (n=157).

Symptoms	Adverse event signals identified by the AE-L model, n (%)
Pain or numbness	57 (36.3)
Fever	46 (29.3)
Nausea	40 (25.5)
Fatigue	23 (14.6)
Rash or itchy	17 (10.8)
Appetite loss	16 (10.2)
Diarrhea	16 (10.2)
Constipation	11 (7)
Vomiting	8 (5.1)
Taste disorder	7 (4.5)
Dizziness	6 (3.8)
Edema	1 (0.6)
Menstrual irregularity	0 (0)
Hair loss	0 (0)
Sleep disorder	0 (0)
Others	21 (13.4)

Figure 3. Symptoms of adverse event signals identified by the AE-L model from subjective records by intervention (n=157). AE-L: adverse events limiting patients' daily lives.

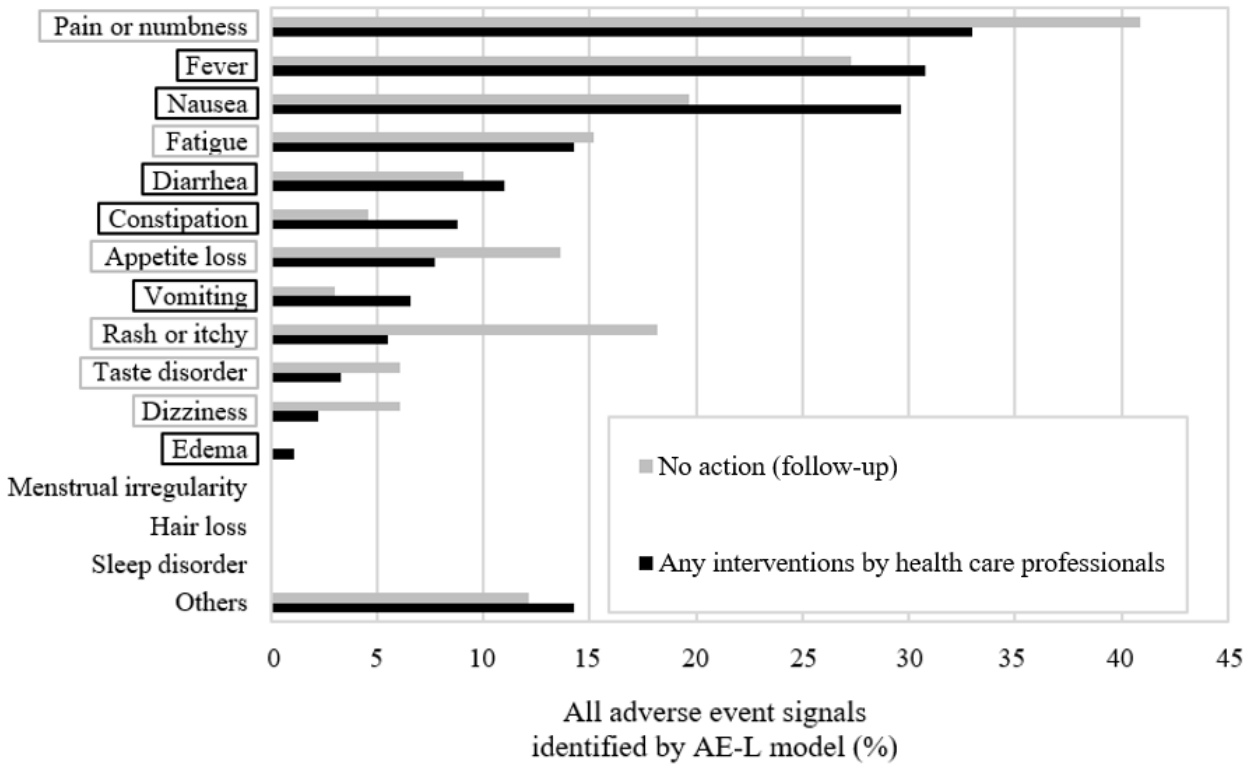

Discussion

Overview

This study was designed to evaluate our deep learning models, previously constructed based on patient-authored texts posted in an online community, by applying them to pharmaceutical care records that contain both patients' subjective concerns and medical information created by pharmacists. Based on the results, we discuss whether these deep learning models can extract clinically important adverse event signals that require medical intervention, and what characteristics they show when applied to data on patients' concerns in pharmaceutical care records.

XSL•FO
RenderX

Performance for Adverse Event Signal Extraction

The first requirement for the deep learning models is to extract adverse event signals from patients' narratives precisely. In this study, we evaluated the proportion of true adverse event signals in positive S records extracted by the HFS or AE-L model. True adverse event signals amounted to 152 (91%) and 157 (80.1%) for the HFS and AE-L models, respectively (Tables 2 and 4). Given that the proportion of true adverse event signals in 200 randomly extracted S records without deep learning models was 54 (27%; categories other than "no adverse event" in Figures 1 and 2), the HFS and AE-L models were able to concentrate S records with adverse event mentions. Although 15 (9%) for the HFS model and 39 (19.9%) for the AE-L model were false-positives, it was confirmed all of the false-positive records described a lack of symptoms or confirmation of improving condition. We considered that such false-positives are due to the unique feature of pharmaceutical care records, where pharmacists might proactively interview patients about potential side effects of their medications. As the data set of blog articles we used to construct the deep learning models included few such cases (especially comments on lack of symptoms), our models seemed unable to exclude them correctly. Even though we confirmed that the proportion of true "adverse event" signals extracted from the S records by the HFS or AE-L model was more than 80%, the performance scores to extract true "HFS" or "AE-L" signals were not so high based on the performance check using 1000 randomly extracted S records (F_1-scores were 0.50 and 0.22 for true HFS and AE-L signals, respectively; Table S1 in Multimedia Appendix 1). It is considered that the performance to extract true HFS and AE-L signals was relatively low due to the short length of texts in the S records, providing less context to judge the impact on patients' daily lives, especially for the AE-L model (the mean word number of the S records was 38.8 [SD 29.4; Table 1], similar to the sentence-level tasks in our previous work [42,43]). However, we consider a true adverse event signal proportion of more than 80% in this study represents a promising outcome, as this is the first attempt to apply our deep learning models to a different source of patients' concern data, and the extracted positive cases would be worthy of evaluation by a medical professional, as the potential adverse events could be caused by drugs taken by the patients.

When the deep learning models were applied to DIPEx-Japan interview transcripts, including patients' concerns, the proportion of true adverse event signals was also more than 80% (for All AE: n=73, 86.9% and for HFS and AE-L: n=18, 100%). The difference in the results between pharmaceutical care S records and DIPEx-Japan interview transcripts was the features of false-positives, descriptions about lack of symptoms or confirmation of improving condition in S records versus explanations about disease or its prognosis, stories about when their cancer was discovered, or emotional changes in interview transcripts. This is considered due to the difference in the nature of the data source; the pharmaceutical care records were generated in a real-time manner by pharmacists through their daily work, where adverse event signals are proactively monitored, while the interview transcripts were purely based on patients' retrospective memories. Our deep learning models were able to extract true adverse event signals with an accuracy of more than 80% from both text data sources in spite of the difference in their nature. When looking at future implementation of the deep learning models in society (discussed in the Potential for Deep Learning Model Implementation in Society section), it may be desirable to further adjust deep learning models to reduce false-positives depending upon the features of the data source.

Identification of Important Adverse Events Requiring Medical Intervention

To assess whether the models could extract clinically important adverse event signals, we investigated interventions by health care professionals connected with the adverse event signals that are identified by our deep learning models. In the 200 randomly extracted S records, only 26 (13%) consisted of adverse event signals, leading to any intervention by health care professionals. On the other hand, the proportion of signals associated with interventions was increased to 107 (64.1%) and 91 (46.4%) in the S records extracted as positive by the HFS and AE-L models, respectively (Figures 1 and 2). These results suggest that both deep learning models can screen clinically important adverse event signals that require intervention from health care professionals. The performance level in screening adverse event signals requiring medical intervention was higher in the HFS model than in the AE-L model (n=107, 64.1% vs n=91, 46.4%; Figures 1 and 2). Since the target events were specific and adverse event signals of HFS were narrowly defined, which is one of the typical side effects of some anticancer drugs, we consider that health care providers paid special attention to HFS-related signals and took action proactively. In both deep learning models, similar trends were observed in actions taken by health care professionals in response to extracted adverse event signals; common actions were attempts to manage adverse event symptoms by symptomatic treatment or other mild interventions, including educational guidance from pharmacists or recommendations for patients to visit a doctor. More direct interventions focused on the causative drugs (ie, "dose reduction or discontinuation of anticancer treatment") amounted to less than 5%; 7 (4.2%) for the HFS model and 6 (3.1%) for the AE-L model (Figures 1 and 2). Thus, it appears that our deep learning models can contribute to screening mild to moderate adverse event signals that require preventive actions such as symptomatic treatments or professional advice from health care providers, especially for patients with less sensitivity to adverse event signals or who have few opportunities to visit clinics and pharmacies.

Ability to Catch Real Side Effect Signals of Anticancer Drugs

Based on the drug prescription history associated with S records extracted as HFS or AE-L positive, the type and duration of anticancer drugs taken by patients experiencing the adverse event signals were investigated. For the HFS model, the most common MoA of anticancer drug was antimetabolite (fluoropyrimidine: n=59, 38.8%), followed by kinase inhibitors (n=49, 32.2%, of which EGFR inhibitors and multikinase inhibitors accounted for n=28, 18.4% and n=14, 9.2%, respectively) and aromatase inhibitors (n=24, 15.8%; Table 3).

XSL•FO
RenderX

It is known that fluoropyrimidine and multikinase inhibitors are typical HFS-inducing drugs [55-58], suggesting that the HFS model accurately extracted HFS side effect signals derived from these drugs. Note that symptoms such as acneiform rash, xerosis, eczema, paronychia, changes in the nails, arthralgia, or stiffness of limb joints, which are common side effects of EGFR inhibitors or aromatase inhibitors [59,60], might be extracted as closely related expressions to those of HFS signals. When looking at the MoA of anticancer drugs for patients with adverse event signals identified by the AE-L model, antimetabolite (fluoropyrimidine) was the most common one (n=53, 33.8%), as in the case of those identified by the HFS model, followed by kinase inhibitors (n=31, 19.7%) and antiandrogens (n=27, 17.2%; Table 5). Since the AE-L model targets a broad range of adverse event symptoms, it is difficult to rationalize the relationship between the adverse event signals and types of anticancer drugs. However, the type of anticancer drugs would presumably closely correspond to the standard treatments of the cancer types of the patients. Based on the prescribed anticancer drugs, we can infer that a large percentage of the patients had breast or lung cancer, indicating that our study results were based on data from such a population. Thus, a possible direction for the expansion of this research would be adjusting the deep learning models by additional training with expressions for typical side effects associated with standard treatments of other cancer types. To interpret these results correctly, it should be noted that we could not investigate anticancer treatments conducted outside of the pharmacies (eg, the time-course relationship with intravenously administered drugs would be missed, as the administration will be done at hospitals). To further evaluate how useful this model is in side effect signal monitoring for patients with cancer, comprehensive medical information for the eligible patients would be required.

Suitability of the Deep Learning Models for Specific Adverse Event Symptoms

Among the adverse event signals identified by the AE-L model, the type of symptom was categorized according to a predefined annotation guideline that we previously developed [43]. The most frequently recorded adverse event signals identified by the AE-L model were “pain or numbness” (n=57, 36.3%), “fever” (n=46, 29.3%), and “nausea” (n=40, 25.5%; Table 6). Since the pharmaceutical care records had information about interventions by health care professionals, the frequency of the presence or absence of the interventions for each symptom was examined. A trend toward a greater proportion of interventions was observed in “fever,” “nausea,” “diarrhea,” “constipation,” “vomiting,” and “edema” (Figure 3, black boxes). There seem to be 2 possible explanations for this: these symptoms are of high importance and require early medical intervention or effective symptomatic treatments are available for these symptoms in clinical practice so that medical intervention is an easy option. On the other hand, a trend for a smaller proportion of adverse event signals to result in interventions was observed for “pain or numbness,” “fatigue,” “appetite loss,” “rash or itchy,” “taste disorder,” and “dizziness” (Figure 3, gray boxes). The reason for this may be the lack of effective symptomatic treatments or the difficulty of judging whether the severity of these symptoms justifies medical intervention by health care providers. In either case, there may be room for improvement in the quality of medical care for these symptoms. We expect that our research will contribute to a quality improvement in safety monitoring in clinical practice by supporting adverse event signal detection in a cost-effective manner.

Potential for Deep Learning Model Implementation in Society

Although we evaluated our deep learning models using pharmaceutical care records in this study, the main target of future implementation of our deep learning models in society would be narrative texts that patients directly write to record their daily experiences. For example, the application of these deep learning models to electronic media where patients record their daily experiences in their lives with disease (eg, health care–related e-communities and disease diary applications) could enable information about adverse event signal onset that patients experience to be provided to health care providers in a timely manner. Adverse event signals can automatically be identified and shared with health care providers based on the concern texts that patients post to any platform. This system will have the advantage that health care providers can efficiently grasp safety-related events that patients experience outside of clinic visits so that they can conduct more focused or personalized interactions with patients at their clinic visits. However, consideration should be given to avoid an excessive burden on health care providers. For instance, limiting the sharing of adverse event signals to those of high severity or summarizing adverse event signals over a week rather than sharing each one in a real-time manner may be reasonable approaches for medical staff. We also need to think about how to encourage patients to record their daily experiences using electronic tools. Not only technical progress and support but also the establishment of an ecosystem where both patients and medical staff can feel benefit will be required. Prospective studies with deep learning models to follow up patients in the long term and evaluate outcomes will be needed. We primarily looked at patient-authored texts as targets of implementation, but our deep learning models may also be worth using medical data including patients’ subjective concerns, such as pharmaceutical care S records. As this study confirmed that our deep learning models are applicable to patients’ concern texts tracked by pharmacists, it should be possible to use them to analyze other “patient voice-like” medical text data that have not been actively investigated so far.

Limitations

First, the major limitation of this study was that we were not able to collect complete medical information of the patients. Although we designed this study to analyze patients’ concerns extracted by the deep learning models and their relationship with medical information contained in the pharmaceutical care records, some information could not be tracked (eg, missing history of medical interventions or anticancer treatment at hospitals as well as diagnosis of patients’ primary cancers). Second, there might be a data creation bias in S records for patients’ concerns by pharmacists. For example, symptoms that have little impact on intervention decisions might less likely be recorded by them. It should be also noted that the characteristics

XSL•FO
RenderX

of S records may not be consistent at different community pharmacies.

Conclusions

Our deep learning models were able to screen clinically important adverse event signals that require intervention by health care professionals from patients' concerns in pharmaceutical care records. Thus, these models have the potential to support real-time adverse event monitoring of individual patients taking anticancer treatments in an efficient manner. We also confirmed that these deep learning models constructed based on patient-authored texts could be applied to patients' subjective information recorded by pharmacists through their daily work. Further research may help to expand the applicability of the deep learning models for implementation in society or for analysis of data on patients' concerns accumulated in professional records at pharmacies or hospitals.

Acknowledgments

This work was supported by Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (KAKENHI; grant 21H03170) and Japan Science and Technology Agency, Core Research for Evolutional Science and Technology (CREST; grant JPMJCR22N1), Japan. Mr Yuki Yokokawa and Ms Sakura Yokoyama at our laboratory advised SN about the structure of pharmaceutical care records. This study would not have been feasible without the high quality of pharmaceutical care records created by many individual pharmacists at Nakajima Pharmacy Group through their daily work.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SN and SH designed the study. SN retrieved the subjective records of patients with cancer from the data source for the application of deep learning models and organized other data for subsequent evaluations. SN ran the deep learning models with the support of SW. SN, YY, and KS checked the adverse event signals for each subjective record that was extracted as positive by the models for hand-foot syndrome or adverse events limiting patients' daily lives and evaluated the adverse event signal symptoms, details of interventions taken by health care professionals, and types of anticancer drugs prescribed for patients based on available data from the data source. HK and SI advised on the study concept and process. MS and RT provided pharmaceutical records at their community pharmacies along with advice on how to use and interpret them. SY and EA supervised the natural language processing research as specialists. SH supervised the study overall. SN drafted and finalized the paper. All authors reviewed and approved the paper.

Conflicts of Interest

SN is an employee of Daiichi Sankyo Co, Ltd. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Performance evaluation of deep learning models.
[DOCX File , 31 KB - jmir_v26i1e55794_app1.docx]

Multimedia Appendix 2
Examples of S records and sample interview transcripts.
[DOCX File , 41 KB - jmir_v26i1e55794_app2.docx]

References

1. Global cancer observatory: cancer over time. World Health Organization. URL: https://gco.iarc.fr/overtime/en [accessed 2023-07-02]
2. Mattiuzzi C, Lippi G. Current cancer epidemiology. J Epidemiol Glob Health 2019;9(4):217-222 [FREE Full text] [doi: 10.2991/jegh.k.191008.001] [Medline: 31854162]
3. Montazeri F, Komaki H, Mohebi F, Mohajer B, Mansournia MA, Shahraz S, et al. Editorial: disparities in cancer prevention and epidemiology. Front Oncol 2022;12:872051 [FREE Full text] [doi: 10.3389/fonc.2022.872051] [Medline: 35719956]
4. Lasala R, Santoleri F. Association between adherence to oral therapies in cancer patients and clinical outcome: a systematic review of the literature. Br J Clin Pharmacol 2022;88(5):1999-2018 [FREE Full text] [doi: 10.1111/bcp.15147] [Medline: 34784654]
5. Pudkasam S, Polman R, Pitcher M, Fisher M, Chinlumprasert N, Stojanovska L, et al. Physical activity and breast cancer survivors: importance of adherence, motivational interviewing and psychological health. Maturitas 2018;116:66-72 [FREE Full text] [doi: 10.1016/j.maturitas.2018.07.010] [Medline: 30244781]

XSL•FO
RenderX

6. Markman M. Chemotherapy-associated neurotoxicity: an important side effect-impacting on quality, rather than quantity, of life. J Cancer Res Clin Oncol 1996;122(9):511-512. [doi: 10.1007/BF01213547] [Medline: 8781565]
7. Jitender S, Mahajan R, Rathore V, Choudhary R. Quality of life of cancer patients. J Exp Ther Oncol 2018;12(3):217-221. [Medline: 29790313]
8. Di Nardo P, Lisanti C, Garutti M, Buriolla S, Alberti M, Mazzeo R, et al. Chemotherapy in patients with early breast cancer: clinical overview and management of long-term side effects. Expert Opin Drug Saf 2022;21(11):1341-1355 [FREE Full text] [doi: 10.1080/14740338.2022.2151584] [Medline: 36469577]
9. Cuomo RE. Improving cancer patient outcomes and cost-effectiveness: a Markov simulation of improved early detection, side effect management, and palliative care. Cancer Invest 2023;41(10):858-862. [doi: 10.1080/07357907.2023.2287485] [Medline: 37991389]
10. Pulito C, Cristaudo A, La Porta C, Zapperi S, Blandino G, Morrone A, et al. Oral mucositis: the hidden side of cancer therapy. J Exp Clin Cancer Res 2020;39(1):210 [FREE Full text] [doi: 10.1186/s13046-020-01715-7] [Medline: 33028357]
11. Bartal A, Mátrai Z, Szûcs A, Liszkay G. Main treatment and preventive measures for hand-foot syndrome, a dermatologic side effect of cancer therapy. Magy Onkol 2011;55(2):91-98 [FREE Full text] [Medline: 21655474]
12. Basch E, Jia X, Heller G, Barz A, Sit L, Fruscione M, et al. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. J Natl Cancer Inst 2009;101(23):1624-1632 [FREE Full text] [doi: 10.1093/jnci/djp386] [Medline: 19920223]
13. Basch E. The missing voice of patients in drug-safety reporting. N Engl J Med 2010;362(10):865-869 [FREE Full text] [doi: 10.1056/NEJMp0911494] [Medline: 20220181]
14. Fromme EK, Eilers KM, Mori M, Hsieh YC, Beer TM. How accurate is clinician reporting of chemotherapy adverse effects? A comparison with patient-reported symptoms from the Quality-of-Life Questionnaire C30. J Clin Oncol 2004;22(17):3485-3490 [FREE Full text] [doi: 10.1200/JCO.2004.03.025] [Medline: 15337796]
15. Liu L, Suo T, Shen Y, Geng C, Song Z, Liu F, et al. Clinicians versus patients subjective adverse events assessment: based on patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). Qual Life Res 2020;29(11):3009-3015. [doi: 10.1007/s11136-020-02558-7] [Medline: 32564293]
16. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expect 2021;24(4):1015-1024 [FREE Full text] [doi: 10.1111/hex.13254] [Medline: 33949755]
17. Pérez-Alfonso KE, Sánchez-Martínez V. Electronic patient-reported outcome measures evaluating cancer symptoms: a systematic review. Semin Oncol Nurs 2021;37(2):151145 [FREE Full text] [doi: 10.1016/j.soncn.2021.151145] [Medline: 33773879]
18. Patient-reported outcome measures: use in medical product development to support labeling claims. U.S. Food & Drug Administration. 2009. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims [accessed 2023-11-26]
19. Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man—the use of patient-reported outcome (PRO) measures in oncology studies—scientific guideline. European Medicines Agency. 2016. URL: https://www.ema.europa.eu/en/appendix-2-guideline-evaluation-anticancer-medicinal-products-man-use-patient-reported-outcome-pro [accessed 2023-11-26]
20. Weber SC. The evolution and use of patient-reported outcomes in regulatory decision making. RF Q 2023;3(1):4-9 [FREE Full text]
21. Teixeira MM, Borges FC, Ferreira PS, Rocha J, Sepodes B, Torre C. A review of patient-reported outcomes used for regulatory approval of oncology medicinal products in the European Union between 2017 and 2020. Front Med (Lausanne) 2022;9:968272 [FREE Full text] [doi: 10.3389/fmed.2022.968272] [Medline: 36035431]
22. Newell S, Jordan Z. The patient experience of patient-centered communication with nurses in the hospital setting: a qualitative systematic review protocol. JBI Database System Rev Implement Rep 2015;13(1):76-87 [FREE Full text] [doi: 10.11124/jbisrir-2015-1072] [Medline: 26447009]
23. Yagasaki K, Takahashi H, Ouchi T, Yamagami J, Hamamoto Y, Amagai M, et al. Patient voice on management of facial dermatological adverse events with targeted therapies: a qualitative study. J Patient Rep Outcomes 2019;3(1):27 [FREE Full text] [doi: 10.1186/s41687-019-0116-3] [Medline: 31049724]
24. Giardina TD, Korukonda S, Shahid U, Vaghani V, Upadhyay DK, Burke GF, et al. Use of patient complaints to identify diagnosis-related safety concerns: a mixed-method evaluation. BMJ Qual Saf 2021;30(12):996-1001 [FREE Full text] [doi: 10.1136/bmjqs-2020-011593] [Medline: 33597282]
25. Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomputing 2022;470:443-456. [doi: 10.1016/j.neucom.2021.05.103]
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, MN, USA p. 4171-4186.

XSL•FO
RenderX

27. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
28. Sim JA, Huang X, Horan MR, Stewart CM, Robison LL, Hudson MM, et al. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: a systematic review. Artif Intell Med 2023;146:102701. [doi: 10.1016/j.artmed.2023.102701] [Medline: 38042599]
29. Weissenbacher D, Banda JM, Davydova V, Estrada-Zavala D, Gascó Sánchez L, Ge Y, et al. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. 2022 Presented at: Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task; October 12-17, 2022; Gyeongju, Republic of Korea p. 221-241 URL: https://aclanthology.org/2022.smm4h-1.54/
30. Matsuda S, Ohtomo T, Okuyama M, Miyake H, Aoki K. Estimating patient satisfaction through a language processing model: model development and evaluation. JMIR Form Res 2023;7(1):e48534 [FREE Full text] [doi: 10.2196/48534] [Medline: 37707946]
31. Yu D, Vydiswaran VGV. An assessment of mentions of adverse drug events on social media with natural language processing: model development and analysis. JMIR Med Inform 2022;10(9):e38140 [FREE Full text] [doi: 10.2196/38140] [Medline: 36170004]
32. Liu X, Chen H. A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports. J Biomed Inform 2015;58:268-279 [FREE Full text] [doi: 10.1016/j.jbi.2015.10.011] [Medline: 26518315]
33. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015;22(3):671-681 [FREE Full text] [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
34. Kakalou C, Dimitsaki S, Dimitriadis VK, Natsiavas P. Exploiting social media for active pharmacovigilance: the PVClinical social media workspace. Stud Health Technol Inform 2022;290:739-743. [doi: 10.3233/SHTI220176] [Medline: 35673115]
35. Bousquet C, Dahamna B, Guillemin-Lanne S, Darmoni SJ, Faviez C, Huot C, et al. The adverse drug reactions from patient reports in social media project: five major challenges to overcome to operationalize analysis and efficiently support pharmacovigilance process. JMIR Res Protoc 2017;6(9):e179 [FREE Full text] [doi: 10.2196/resprot.6463] [Medline: 28935617]
36. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 2019;132:103971. [doi: 10.1016/j.ijmedinf.2019.103971] [Medline: 31630063]
37. Jacobsson R, Bergvall T, Sandberg L, Ellenius J. Extraction of adverse event severity information from clinical narratives using natural language processing. Pharmacoepidemiol Drug Saf 2017;26(S2):37 [FREE Full text]
38. Liang C, Gong Y. Predicting harm scores from patient safety event reports. Stud Health Technol Inform 2017;245:1075-1079. [doi: 10.3233/978-1-61499-830-3-1075] [Medline: 29295267]
39. Jiang G, Wang L, Liu H, Solbrig HR, Chute CG. Building a knowledge base of severe adverse drug events based on AERS reporting data using semantic web technologies. Stud Health Technol Inform 2013;192(1-2):496-500. [doi: 10.3233/978-1-61499-289-9-496] [Medline: 23920604]
40. Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in Japanese. JMIR Med Inform 2018;6(3):e11021 [FREE Full text] [doi: 10.2196/11021] [Medline: 30262450]
41. Watanabe T, Yada S, Aramaki E, Yajima H, Kizaki H, Hori S. Extracting multiple worries from breast cancer patient blogs using multilabel classification with the natural language processing model bidirectional encoder representations from transformers: infodemiology study of blogs. JMIR Cancer 2022;8(2):e37840 [FREE Full text] [doi: 10.2196/37840] [Medline: 35657664]
42. Nishioka S, Watanabe T, Asano M, Yamamoto T, Kawakami K, Yada S, et al. Identification of hand-foot syndrome from cancer patients' blog posts: BERT-based deep-learning approach to detect potential adverse drug reaction symptoms. PLoS One 2022;17(5):e0267901 [FREE Full text] [doi: 10.1371/journal.pone.0267901] [Medline: 35507636]
43. Nishioka S, Asano M, Yada S, Aramaki E, Yajima H, Yanagisawa Y, et al. Adverse event signal extraction from cancer patients' narratives focusing on impact on their daily-life activities. Sci Rep 2023;13(1):15516 [FREE Full text] [doi: 10.1038/s41598-023-42496-1] [Medline: 37726371]
44. Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(11):593-600. [doi: 10.1056/NEJM196803142781105] [Medline: 5637758]
45. Podder V, Lew V, Ghassemzadeh S. SOAP Notes. Treasure Island, FL: StatPearls Publishing; 2023.
46. Shenavar Masooleh I, Ramezanzadeh E, Yaseri M, Sahere Mortazavi Khatibani S, Sadat Fayazi H, Ali Balou H, et al. The effectiveness of training on daily progress note writing by medical interns. J Adv Med Educ Prof 2021;9(3):168-175 [FREE Full text] [doi: 10.30476/jamp.2021.88747.1357] [Medline: 34277848]

XSL•FO
RenderX

47. Grothen AE, Tennant B, Wang C, Torres A, Sheppard BB, Abastillas G, et al. Application of artificial intelligence methods to pharmacy data for cancer surveillance and epidemiology research: a systematic review. JCO Clin Cancer Inform 2020;4:1051-1058 [FREE Full text] [doi: 10.1200/CCI.20.00101] [Medline: 33197205]
48. Ohno Y, Kato R, Ishikawa H, Nishiyama T, Isawa M, Mochizuki M, et al. Using the natural language processing system MedNER-J to analyze pharmaceutical care records. medRxiv Preprint posted online on October 2, 2023. [doi: 10.1101/2023.09.28.23295887]
49. Ranchon F, Chanoine S, Lambert-Lacroix S, Bosson JL, Moreau-Gaudry A, Bedouch P. Development of artificial intelligence powered apps and tools for clinical pharmacy services: a systematic review. Int J Med Inform 2023;172:104983. [doi: 10.1016/j.ijmedinf.2022.104983] [Medline: 36724730]
50. Nakajima Pharmacy. URL: https://www.nakajima-phar.co.jp/ [accessed 2023-12-07]
51. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1-67 [FREE Full text]
52. Pathak A. Comparative analysis of transformer based language models. Comput Sci Inf Technol 2021:165-176 [FREE Full text] [doi: 10.5121/csit.2021.110111]
53. DIPEx Japan. URL: https://www.dipex-j.org/ [accessed 2024-02-04]
54. Herxheimer A, McPherson A, Miller R, Shepperd S, Yaphe J, Ziebland S. Database of patients' experiences (DIPEx): a multi-media approach to sharing experiences and information. Lancet 2000;355(9214):1540-1543. [doi: 10.1016/S0140-6736(00)02174-7] [Medline: 10801187]
55. Lara PE, Muiño CB, de Spéville BD, Reyes JJ. Hand-foot skin reaction to regorafenib. Actas Dermosifiliogr 2016;107(1):71-73. [doi: 10.1016/j.adengl.2015.11.005]
56. Zaiem A, Hammamia SB, Aouinti I, Charfi O, Ladhari W, Kastalli S, et al. Hand-foot syndrome induced by chemotherapy drug: case series study and literature review. Indian J Pharmacol 2022;54(3):208-215 [FREE Full text] [doi: 10.4103/ijp.ijp_175_21] [Medline: 35848692]
57. McLellan B, Ciardiello F, Lacouture ME, Segaert S, Van Cutsem E. Regorafenib-associated hand-foot skin reaction: practical advice on diagnosis, prevention, and management. Ann Oncol 2015;26(10):2017-2026 [FREE Full text] [doi: 10.1093/annonc/mdv244] [Medline: 26034039]
58. Ai L, Xu Z, Yang B, He Q, Luo P. Sorafenib-associated hand-foot skin reaction: practical advice on diagnosis, mechanism, prevention, and management. Expert Rev Clin Pharmacol 2019;12(12):1121-1127. [doi: 10.1080/17512433.2019.1689122] [Medline: 31679411]
59. Tenti S, Correale P, Cheleschi S, Fioravanti A, Pirtoli L. Aromatase inhibitors-induced musculoskeletal disorders: current knowledge on clinical and molecular aspects. Int J Mol Sci 2020;21(16):5625 [FREE Full text] [doi: 10.3390/ijms21165625] [Medline: 32781535]
60. Lacouture ME, Melosky BL. Cutaneous reactions to anticancer agents targeting the epidermal growth factor receptor: a dermatology-oncology perspective. Skin Therapy Lett 2007;12(6):1-5 [FREE Full text] [Medline: 17762902]

Abbreviations

AE-L: adverse events limiting patients' daily lives
All AE: all or any adverse event
BERT: Bidirectional Encoder Representations From Transformers
DIPEx-Japan: Database of Individual Patient Experiences-Japan
EGFR: epidermal growth factor receptor
HFS: hand-foot syndrome
MoA: mechanism of action
PRO: patient-reported outcome
SOAP: subjective, objective, assessment, plan

Edited by G Eysenbach; submitted 25.12.23; peer-reviewed by CY Wang, L Guo; comments to author 24.01.24; revised version received 14.02.24; accepted 09.03.24; published 16.04.24.

Please cite as:
Nishioka S, Watabe S, Yanagisawa Y, Sayama K, Kizaki H, Imai S, Someya M, Taniguchi R, Yada S, Aramaki E, Hori S
Adverse Event Signal Detection Using Patients' Concerns in Pharmaceutical Care Records: Evaluation of Deep Learning Models
J Med Internet Res 2024;26:e55794
URL: https://www.jmir.org/2024/1/e55794
doi:10.2196/55794
PMID:38625718

XSL•FO
RenderX

©Satoshi Nishioka, Satoshi Watabe, Yuki Yanagisawa, Kyoko Sayama, Hayato Kizaki, Shungo Imai, Mitsuhiro Someya, Ryoo Taniguchi, Shuntaro Yada, Eiji Aramaki, Satoko Hori. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning for Predicting Risk and Prognosis of Acute Kidney Disease in Critically Ill Elderly Patients During Hospitalization: Internet-Based and Interpretable Model Study

Mingxia Li[1,2*], PhD; Shuzhe Han[3*], MM; Fang Liang[4], MD; Chenghuan Hu[1], MD; Buyao Zhang[1], MM; Qinlan Hou[1], MM; Shuangping Zhao[1,5,6], MD

[1]Department of Critical Care Medicine, Xiangya Hospital of Central South University, Changsha, China
[2]Department of Critical Care Medicine, ZhuJiang Hospital of Southern Medical University, Guangzhou, China
[3]Department of Obstetrics and Gynecology, 967th Hospital of the Joint Logistics Support Force of the Chinese People's Liberation Army, Dalian, China
[4]Department of Hematology and Critical Care Medicine, The Third Xiangya Hospital, Central South University, Changsha, China
[5]National Clinical Research Center for Geriatric Disorders, Changsha, China
[6]Hunan Provincial Clinical Research Center of Intensive Care Medicine, Changsha, China
[*]these authors contributed equally

Corresponding Author:
Shuangping Zhao, MD
Department of Critical Care Medicine
Xiangya Hospital of Central South University
No 87, Xiangya Road, Kaifu District
Changsha, 410008
China
Phone: 86 1397495302
Email: zhshping@csu.edu.cn

Abstract

Background: Acute kidney disease (AKD) affects more than half of critically ill elderly patients with acute kidney injury (AKI), which leads to worse short-term outcomes.

Objective: We aimed to establish 2 machine learning models to predict the risk and prognosis of AKD in the elderly and to deploy the models as online apps.

Methods: Data on elderly patients with AKI (n=3542) and AKD (n=2661) from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database were used to develop 2 models for predicting the AKD risk and in-hospital mortality, respectively. Data collected from Xiangya Hospital of Central South University were for external validation. A bootstrap method was used for internal validation to obtain relatively stable results. We extracted the indicators within 24 hours of the first diagnosis of AKI and the fluctuation range of some indicators, namely delta (day 3 after AKI minus day 1), as features. Six machine learning algorithms were used for modeling; the area under the receiver operating characteristic curve (AUROC), decision curve analysis, and calibration curve for evaluating; Shapley additive explanation (SHAP) analysis for visually interpreting; and the Heroku platform for deploying the best-performing models as web-based apps.

Results: For the model of predicting the risk of AKD in elderly patients with AKI during hospitalization, the Light Gradient Boosting Machine (LightGBM) showed the best overall performance in the training (AUROC=0.844, 95% CI 0.831-0.857), internal validation (AUROC=0.853, 95% CI 0.841-0.865), and external (AUROC=0.755, 95% CI 0.699–0.811) cohorts. In addition, LightGBM performed well for the AKD prognostic prediction in the training (AUROC=0.861, 95% CI 0.843-0.878), internal validation (AUROC=0.868, 95% CI 0.851-0.885), and external (AUROC=0.746, 95% CI 0.673-0.820) cohorts. The models deployed as online prediction apps allowed users to predict and provide feedback to submit new data for model iteration. In the importance ranking and correlation visualization of the model's top 10 influencing factors conducted based on the SHAP value, partial dependence plots revealed the optimal cutoff of some interventionable indicators. The top 5 factors predicting the risk of AKD were creatinine on day 3, sepsis, delta blood urea nitrogen (BUN), diastolic blood pressure (DBP), and heart rate, while the top 5 factors determining in-hospital mortality were age, BUN on day 1, vasopressor use, BUN on day 3, and partial pressure of carbon dioxide ($PaCO_2$).

XSL•FO
RenderX

Conclusions: We developed and validated 2 online apps for predicting the risk of AKD and its prognostic mortality in elderly patients, respectively. The top 10 factors that influenced the AKD risk and mortality during hospitalization were identified and explained visually, which might provide useful applications for intelligent management and suggestions for future prospective research.

(J Med Internet Res 2024;26:e51354) doi:10.2196/51354

KEYWORDS

acute kidney disease; AKD; machine learning; critically ill patients; elderly patients; Shapley additive explanation; SHAP

Introduction

Acute kidney injury (AKI), as a common complex heterogeneous syndrome in critically ill patients, is associated with an increased risk of death and adverse renal events [1-3]. AKI is more common in elderly patients in the intensive care unit (ICU), with sustained impaired renal function associated with a poor prognosis for survival [4,5]. Therefore, in this study, we focused on AKI in the elderly with renal function impairment for more than 7 days during hospitalization, that is, acute kidney disease (AKD). In 2012, Kidney Disease: Improving Global Outcomes (KDIGO) first proposed the term “AKD,” but at this time, it was viewed as a period of kidney pathology following AKI, not as an independent definition [6]. In a comparative study on the epidemiology of AKD, patients with AKI who developed AKD had a higher risk of chronic kidney disease (CKD) and end-stage renal disease (ESRD), suggesting the potential clinical research value for AKD as a novel term [7]. In 2017, the Acute Disease Quality Initiative (ADQI) 16 Workgroup presented an expert consensus on AKD, defining AKD as AKI with KDIGO stage 1 or higher within 7-90 days of the first diagnosis of AKI [8]. There is a distinct disease course between AKI, AKD, and CKD. AKD highlights the key intervenable period in the transformation process from AKI to CKD and lays the foundation for the construction of management strategies for renal function recovery.

Recently, a multicenter study indicated that more than half of the hospitalized patients with AKI developed AKD, which increased the risk of long-term mortality [9]. Nevertheless, an epidemiological study by Andonovic et al [10] found that patients with AKD in the ICU have a higher short-term risk of death but no statistically significant difference in long-term survival. Moreover, Chen et al [11] reported that patients with AKD are more likely to require long-term dialysis. Considering the high incidence and mortality rate of AKD, researchers have conducted exploratory early warning studies on AKD. Current predictive research on AKD has focused primarily on patients with AKI during hospitalization, sepsis-related AKI, coronary heart disease, and renal cell carcinoma postoperatively [9,12-14]. However, it should be noted that elderly patients with diminished renal function have received insufficient attention, and little is known about the AKD risk and prognostic mortality in the elderly. Further, some studies have explored the use of artificial intelligence (AI) algorithms to predict the onset and progression of diseases, but only a few have developed user-friendly online prediction apps for clinical practice. Zhou et al [15] established an online calculator using Extreme Gradient Boosting (XGBoost) to predict AKI in patients with sepsis-associated acute respiratory distress syndrome. Regrettably, this risk calculator has not been validated externally to determine generalization capacity. Thus, it is imperative to use big data and AI technology to conduct research on the diagnosis and prognosis prediction of AKD in the elderly, and transform the AI models into internet-based apps to assist clinicians in timely intervention to maximize the improvement in renal function and survival outcomes.

Therefore, this study intended to develop 2 machine learning prediction models: one was a model for predicting the risk of AKD in critically ill patients during hospitalization to address problems regarding renal function recovery and early detection of AKD; the other was for predicting the in-hospital mortality in AKD to deal with the adverse outcomes of AKD. In addition, Shapley additive explanation (SHAP) analysis was used to rank the importance and visualize the correlation of the features affecting the occurrence and outcome of AKD [16]. Importantly, we deployed the models with the most comprehensive performance as web-based prediction apps to facilitate doctors' decision-making.

Methods

Ethical Considerations

This study was conducted strictly in accordance with the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [17]. The MIMIC-IV database was approved by the Ethics Review Boards of the MIT and the Beth Israel Deaconess Medical Center. This study obtained access and download permission from the MIMIC-IV database (no. 41817305) and passed the retrospective ethics review of the Medical Ethics Committee of Xiangya Hospital Central South University (no. 202105200). Due to the deprivacy of the data for this retrospective study, it was exempted from patients’ informed consent.

Study Design

A retrospective cohort study was conducted using electronic health records (EHRs) from the Medical Information Mart for Intensive Care IV (MIMIC-IV) data and patients’ data from Xiangya Hospital Central South University (Hunan, China) [18]. In June 2022, the Massachusetts Institute of Technology (MIT) released the revised version of MIMIC-IV 2.0, which contained in-hospital diagnosis and treatment records, as well as in- and out-of-hospital death information for about 40,000 patients in ICUs, and achieved data privacy by deleting patient identification numbers and drifting data through time. The MIMIC-IV cohorts were used to develop machine learning predictive models. Further, the Department of Critical Care

XSL•FO
RenderX

Medicine at Xiangya Hospital Central South University is the National Key Clinical Specialty. Approximately 2500 critically ill patients are admitted to the department each year for the treatment of various diseases. Patients were enrolled in the Department of Critical Care Medicine from 2017 to 2021 for external validation of models.

Data on AKI in the Elderly

Inclusion and Exclusion Criteria

Data on AKI in elderly patients in the ICU were collected for the construction and external verification of an early warning model for the risk of AKD. According to the Chinese Healthy Elderly Standard issued by the National Health Commission of China, those 60 years old and above were defined as the elderly. The inclusion criteria were (1) age≥60 years, (2) ICU stay of at least 48 hours, (3) EHRs of patients admitted to the ICU for the first time, and (4) patients with AKI who met KDIGO criteria. The exclusion criteria were (1) patients with ESRD and (2) missing data on the diagnosis of AKI. Figure 1 shows the data extraction process in detail.

Figure 1. Flowchart for the selection of elderly patients with AKI and AKD. AKD: acute kidney disease; AKI: acute kidney injury; ESRD: end-stage renal disease; ICU: intensive care unit; LightGBM: Light Gradient Boosting Machine; MIMIC-IV: Medical Information Mart for Intensive Care IV.

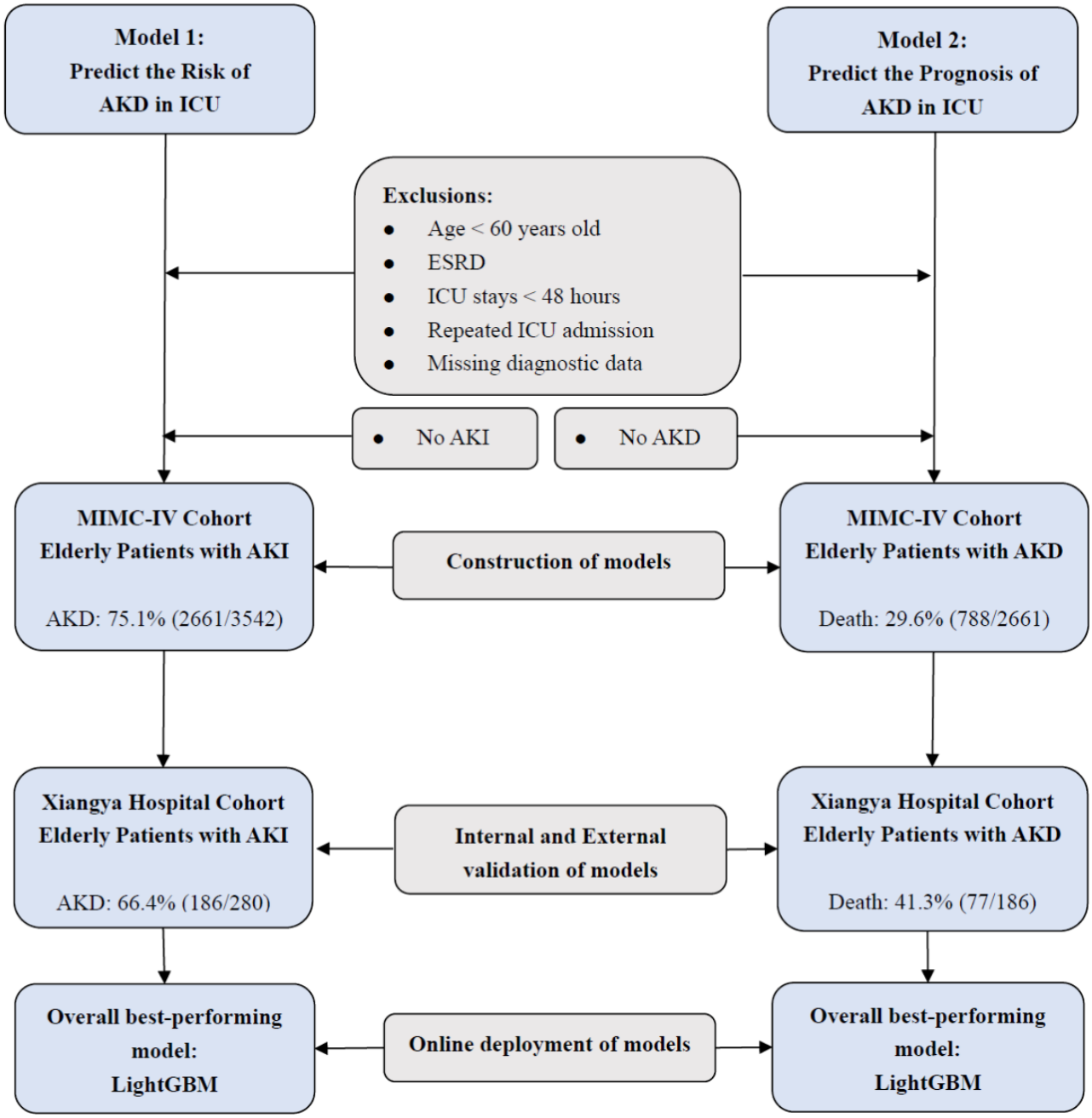

XSL•FO
RenderX

Outcome Definition

The occurrence of AKD during hospitalization was considered the outcome of the risk prediction study. AKI was diagnosed and staged in accordance with the AKI guidelines issued by the KDIGO in 2012 [6]. According to the expert consensus of the ADQI-16 Workgroup in 2017, AKD was defined as the presence of at least stage 1 AKI within 7-90 days after the initial diagnosis of AKI [8]. In this study, patients with AKI who met this definition during hospitalization were regarded as having AKD (Multimedia Appendix 1).

Data Extraction

Navicat Premium (version 15.0.13) was used for MIMIC-IV database management and PostgreSQL (version 9.6; PostgreSQL Global Development Group) for variable extraction. Patients with AKI were identified based on their serum creatinine and urine output levels. Patients with AKI stage higher than or equal to 1 between 7 days following AKI and discharge were considered to have AKD during their hospitalization. Finally, 33 variables were determined and extracted from the Xiangya Hospital data set and the MIMIC-IV database, including age, gender, and the AKI stage as basic characteristics; sepsis, hypertension, diabetes, chronic kidney disease (CKD), chronic pulmonary disease (CPD), and chronic liver disease (CLD) as comorbidities; mechanical ventilation (MV), renal replacement therapy (RRT), and vasopressors as interventions; heart rate, respiratory rate, systolic blood pressure (SBP), diastolic blood pressure (DBP) as vital signs; and white blood cell (WBC) count, red blood cell (RBC) count, hemoglobin, hematocrit, potassium, calcium, anion gap, partial pressure of oxygen (PaO_2), partial pressure of carbon dioxide ($PaCO_2$), pH, glucose, blood urea nitrogen (BUN), and serum creatinine as laboratory tests. These examination indicators were measured on day 1 of AKI diagnosis. We also obtained BUN and serum creatinine levels on day 3 following AKI diagnosis, as well as corresponding delta BUN and delta creatinine values on day 3 minus day 1.

Data on AKD in the Elderly

Inclusion and Exclusion Criteria

Following the aforementioned study on AKI in the elderly, further extracted the data concerning critically ill patients with AKD to construct and verify a model that predicted poor prognostic mortality during hospitalization. The inclusion criteria of elderly patients with AKD were as follows: (1) age≥60 years, (2) length of stay in the ICU for more than 48 hours, (3) patients admitted to the ICU for the first time, and (4) patients with AKD who met the ADQI consensus of 2017. The exclusion criteria were (1) patients who had ESRD and (2) missing data related to AKD diagnosis. A detailed description of the data extraction can be seen in Figure 1.

Outcome Definition and Data Extraction

In-hospital death was the outcome of the prognostic prediction study of AKD in the elderly.

This study on prognostic mortality prediction and the aforementioned study on risk prediction of AKD in the elderly were similar in terms of the content and timing for extracting 33 variables.

Construction and Validation of Models

Several supervised learning algorithms were selected to solve classification prediction problems in this study: logistic regression model (LRM), XGBoost, Light Gradient Boosting Machine (LightGBM), multilayer perceptron (MLP), random forest (RF), and the K-nearest neighbor (KNN) algorithm. Two models were developed using the MIMIC-IV cohort: one for predicting AKD occurrence among the elderly and the other for predicting the prognostic mortality in AKD. To prevent overfitting and improve generalization, a 10-fold cross-validation method was applied to assess the models, and the final models were constructed based on repeated iterations. Multimedia Appendix 2 shows the optimal hyperparameters of the AKD risk and AKD mortality models. Through the GridSearchCV module, we conducted a grid search that traversed all parameter values and returned the parameter combination that provided the best overall performance. The models constructed from the MIMIC-IV database were internally validated by bootstrap resampling with replacement to evaluate performance. The established training model was externally validated with the Xiangya Hospital cohort.

Evaluation and Deployment of Models

The classification prediction effect of the models was evaluated using the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) under the optimal cutoff value. Moreover, a calibration curve was developed to determine the models' predictive accuracy, and clinical decision curve analysis (DCA) was performed to assess their clinical utility. To enhance the interpretability of machine learning black-box models, we performed SHAP analysis by visualizing each feature's marginal contribution to the models' prediction in importance-ranking plots and showing how each feature impacts the outcome in partial dependence plots. Lastly, we selected machine learning algorithm models with the best comprehensive performance from the training and validation cohorts and deployed them to the online server for the convenience of clinical workers or patients. The web-based apps were managed by Heroku.

Statistical Analysis

Data were analyzed using Python (version 3.9.7) and R (version 4.2.0; R Foundation for Statistical Computing). Variables with a missing ratio higher than 35% were deleted, and the *mice* package (version 3.14.0) in R was used to fill in the missing values using the multiple imputation method. In data preprocessing, the Z-score method was used to scale the continuous variables with the *StandardScaler* function. Categorical variables were represented as numbers (percentages) and compared between groups using the chi-square test. Depending on whether the continuous variables were normally distributed, the mean (SD) or median (IQR) was expressed and compared using the 2-tailed *t* test or the Mann-Whitney *U* test. By analyzing the Youden index, the optimal cutoff value of the receiver operating characteristic (ROC) curve was calculated,

XSL•FO
RenderX

as well as the sensitivity, specificity, PPV, and NPV of the models. Statistical significance was set at $P<.05$.

Results

Model 1: Predicting the risk of AKD in the Elderly

Baseline Characteristics

In the study on the risk prediction of AKD in critically ill elderly patients during hospitalization, a total of 3542 elderly patients with AKI from the MIMIC-IV database and 280 from Xiangya Hospital were retrospectively included after screening by the inclusion and exclusion criteria as the training and external validation cohorts, respectively. AKD incidence was 75.1% (2661/3542) in the training cohort and 66.4% (186/280) in the external validation cohort. A comparison of baseline characteristics and stratification of the 2 cohorts according to the presence or absence of AKD is shown in Multimedia Appendix 3. In the MIMIC-IV cohort, patients with AKD had a higher proportion of comorbidities (sepsis, diabetes, CKD, and CPD) and a lower proportion of hypertension ($P<.05$); a higher proportion of interventions (MV, RRT, and vasopressor use; $P<.05$); a higher heart rate and lower SBP and DBP in terms of vital signs ($P<.05$); and higher potassium, anion gap, BUN on day 1 (following AKI diagnosis), serum creatinine on day 1, BUN on day 3 (following AKI diagnosis), and serum creatinine on day 3 and lower PaO_2 in terms of laboratory tests ($P<.05$). Furthermore, more patients with AKD in the Xiangya Hospital cohort were males and had stage 3 AKI ($P<.05$). The features of the Xiangya Hospital cohort with a similar trend to the MIMIC-IV cohort were as follows: diabetes, CKD, RRT, vasopressor use, potassium, anion gap, PaO_2, BUN on day 1, serum creatinine on day 1, BUN on day 3, and serum creatinine on day 3, while hypertension and SBP had opposite trends and statistical results ($P<.05$).

Model Comparison

We included all the variables shown in Multimedia Appendix 3 in the model construction since these indicators are common and easily obtainable in clinical practice. Table 1 shows the performance comparison of the 6 machine learning models for predicting AKD risk in the training, internal validation, and external cohorts. In the training cohort, the algorithm with the greatest performance was LightGBM, with AUROC=0.844 (95% CI 0.831-0.857), sensitivity=0.788 (95% CI 0.759-0.814), specificity=0.761 (95% CI 0.745-0.777), PPV=0.522 (95% CI 0.495-0.549), and NPV=0.915 (95% CI 0.903-0.927). In the external validation cohort, the best-predicting model was the LRM, which had AUROC=0.763 (95% CI 0.707-0.818), sensitivity=0.830 (95% CI 0.738-0.899), specificity=0.586 (95% CI 0.512-0.658), PPV=0.503 (95% CI 0.422-0.584), and NPV=0.872 (95% CI 0.800-0.925). LightGBM also demonstrated the ability to distinguish patients at a higher risk of AKD during hospitalization in the validation cohort: AUROC=0.853 (95% CI 0.841-0.865), sensitivity=0.817 (95% CI 0.791-0.842), specificity=0.759 (95% CI 0.742-0.775), PPV=0.534 (95% CI 0.507-0.560), and NPV=0.925 (95% CI 0.913-0.936) in the internal validation cohort and AUROC=0.755 (95% CI 0.699-0.811), sensitivity=0.851 (95% CI 0.763-0.916), specificity=0.597 (95% CI 0.523-0.668), PPV=0.516 (95% CI 0.435-0.597), and NPV=0.888 (95% CI 0.819-0.937) in the external cohort. Figures 2A and 2B provide the ROC curves of the prediction models in the training and external cohorts, among which LightGBM showed the best overall performance. Multimedia Appendix 4 exhibits the ROC curves for the internal validation cohort. We selected the 3 algorithms (LightGBM, RF, XGBoost) with better performance in the validation cohort to conduct DCA; Figure 2C shows that when the threshold probability of AKD reached 60%, the net benefit ratio of taking intervention measures was 0.5, showing good clinical applicability of LightGBM. Further, the calibration curves for the 3 algorithms are presented in Figure 2D displaying the relative consistency between predictions and actual values. However, according to Figure 2D, when the threshold probability was low, the prediction probability of the model was high with overfitting, which was consistent with Figure 2C: when the threshold probability was low, the net benefit of the model hardly increased.

XSL•FO
RenderX

Figure 2. ROC curves, DCA , and calibration curves of AKD risk prediction models. AKD: acute kidney disease; AUROC: area under the receiver operating characteristic curve; DCA: decision curve analysis; KNN: K-nearest neighbor; LightGBM: Light Gradient Boosting Machine; LRM: logistic regression model; MLP: multilayer perceptron; RF: random forest; ROC: receiver operating characteristic; XGBoost: Extreme Gradient Boosting.

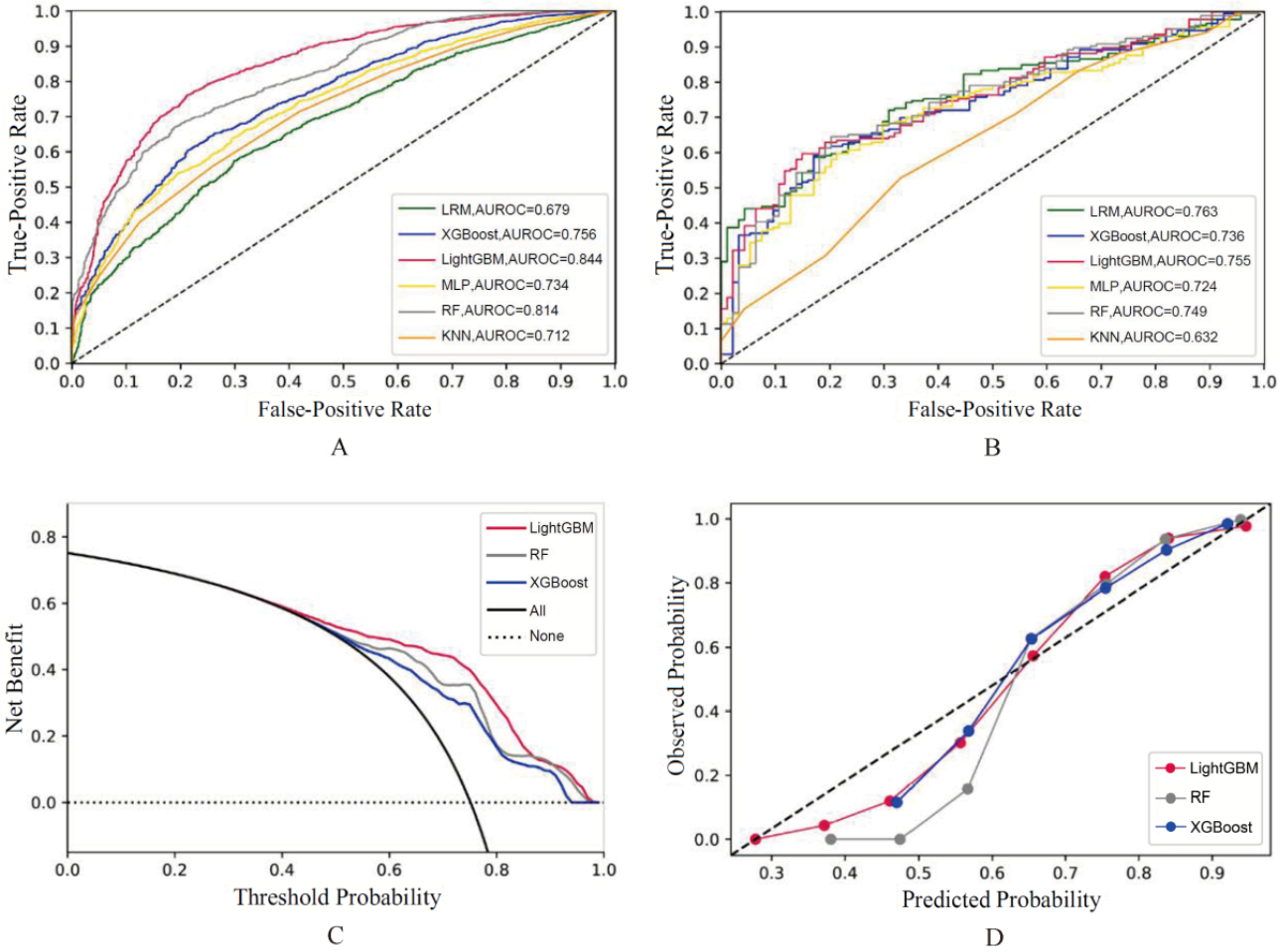

XSL•FO
RenderX

Table 1. Performance of the AKD[a] risk prediction models for elderly patients.

Cohort and models	AUROC[b] (95% CI)	Cutoff	Sensitivity (95% CI)	Specificity (95% CI)	PPV[c] (95% CI)	NPV[d] (95% CI)
Training cohort						
LRM[e]	0.679 (0.660-0.698)	0.749	0.701 (0.670-0.732)	0.575 (0.556-0.594)	0.353 (0.331-0.376)	0.853 (0.836-0.869)
XGBoost[f]	0.756 (0.740-0.773)	0.751	0.787 (0.758-0.813)	0.605 (0.586-0.624)	0.397 (0.374-0.421)	0.895 (0.880-0.909)
LightGBM[g]	0.844 (0.831-0.857)	0.724	0.788 (0.759-0.814)	0.761 (0.745-0.777)	0.522 (0.495-0.549)	0.915 (0.903-0.927)
MLP[h]	0.734 (0.717-0.751)	0.774	0.806 (0.778-0.832)	0.539 (0.520-0.558)	0.367 (0.345-0.389)	0.894 (0.877-0.908)
RF[i]	0.814 (0.800-0.828)	0.748	0.810 (0.783-0.836)	0.671 (0.653-0.689)	0.449 (0.425-0.474)	0.914 (0.901-0.927)
KNN[j]	0.712 (0.694-0.730)	0.789	0.742 (0.712-0.771)	0.558 (0.539-0.577)	0.357 (0.335-0.380)	0.867 (0.850-0.883)
Internal validation cohort						
LRM	0.669 (0.650-0.688)	0.677	0.710 (0.679-0.740)	0.566 (0.547-0.585)	0.356 (0.333-0.378)	0.853 (0.835-0.869)
XGBoost	0.684 (0.665-0.703)	0.657	0.614 (0.582-0.647)	0.663 (0.645-0.681)	0.381 (0.356-0.407)	0.836 (0.820-0.852)
LightGBM	0.853 (0.841-0.865)	0.722	0.817 (0.791-0.842)	0.759 (0.742-0.775)	0.534 (0.507-0.560)	0.925 (0.913-0.936)
MLP	0.719 (0.701-0.737)	0.739	0.751 (0.722-0.779)	0.587 (0.568-0.606)	0.380 (0.357-0.403)	0.875 (0.859-0.890)
RF	0.823 (0.809-0.837)	0.745	0.844 (0.819-0.868)	0.653 (0.634-0.671)	0.450 (0.426-0.475)	0.926 (0.913-0.937)
KNN	0.692 (0.674-0.711)	0.789	0.731 (0.701-0.760)	0.552 (0.532-0.571)	0.355 (0.333-0.377)	0.859 (0.841-0.875)
External validation cohort						
LRM	0.763 (0.707-0.818)	0.787	0.830 (0.738-0.899)	0.586 (0.512-0.658)	0.503 (0.422-0.584)	0.872 (0.800-0.925)
XGBoost	0.736 (0.678-0.794)	0.825	0.809 (0.714-0.882)	0.613 (0.539-0.683)	0.514 (0.430-0.596)	0.864 (0.793-0.917)
LightGBM	0.755 (0.699-0.811)	0.899	0.851 (0.763-0.916)	0.597 (0.523-0.668)	0.516 (0.435-0.597)	0.888 (0.819-0.937)
MLP	0.724 (0.665-0.784)	0.764	0.702 (0.599-0.792)	0.683 (0.611-0.749)	0.528 (0.437-0.618)	0.819 (0.750-0.876)
RF	0.749 (0.692-0.806)	0.778	0.798 (0.702-0.874)	0.645 (0.572-0.714)	0.532 (0.446-0.616)	0.863 (0.795-0.916)
KNN	0.632 (0.566-0.699)	0.789	0.670 (0.566-0.764)	0.527 (0.453-0.600)	0.417 (0.338-0.500)	0.760 (0.677-0.831)

[a]AKD: acute kidney disease.

[b]AUROC: area under the receiver operating characteristic curve.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]LRM: logistic regression model.

[f]XGBoost: Extreme Gradient Boosting.

[g]LightGBM: Light Gradient Boosting Machine.

[h]MLP: multilayer perceptron.

[i]RF: random forest.

[j]KNN: K-nearest neighbor.

XSL•FO
RenderX

Model Interpretability

To better explain the clinical significance of certain features, this study quantified the features' importance as SHAP values. As shown in Figure 3A, variables were given a ranking based on their contribution to the risk prediction of AKD, with creatinine on day 3, sepsis, delta BUN, DBP, heart rate, delta creatinine, creatinine on day 1, respiratory rate, pH, and diabetes as the top 10 predictors of developing AKD during hospitalization in the elderly. Figure 3B shows a detailed relationship between each feature and AKD risk, indicating that the positively related features were as follows: creatinine on day 3, sepsis, delta BUN, heart rate, delta creatinine, creatinine on day 1, pH, and diabetes (the higher the value of these features or the presence of complications, the higher the probability of developing AKD in elderly patients with AKI). Further, the protective effect was associated with a higher DBP. However, the relationship between respiratory rate and AKD during hospitalization was not clearly demonstrated. Furthermore, partial dependence plots were drawn in Figure 4 for the first 4 continuous variables in Figure 3A. The partial dependence plots visually displayed the global relationship between feature and risk distribution. According to Figure 4A, the change curve between creatinine on day 3 (abscissa) and AKD risk (ordinate) indicated a cutoff value of 110 for this feature, meaning that when the creatinine level on day 3 exceeded 110 umol/L, the risk of AKD during hospitalization also increased. Similarly, Figures 4B, 4C, and 4D demonstrate that 0 was the cutoff for delta BUN (positive correlation), 80 mmHg for the DBP (negative correlation), and 110 beats/minute for the heart rate (positive correlation). Thus, targeted feature management may assist in reducing the risk of AKD in elderly patients with AKI during hospitalization based on the cutoff values in partial dependency plots.

Figure 3. Feature-ranking plots (A) and summary plots (B) of LightGBM for predicting AKD risk. AKD: acute kidney disease; BUN: blood urea nitrogen; DBP: diastolic blood pressure; LightGBM: Light Gradient Boosting Machine; SHAP: Shapley additive explanation.

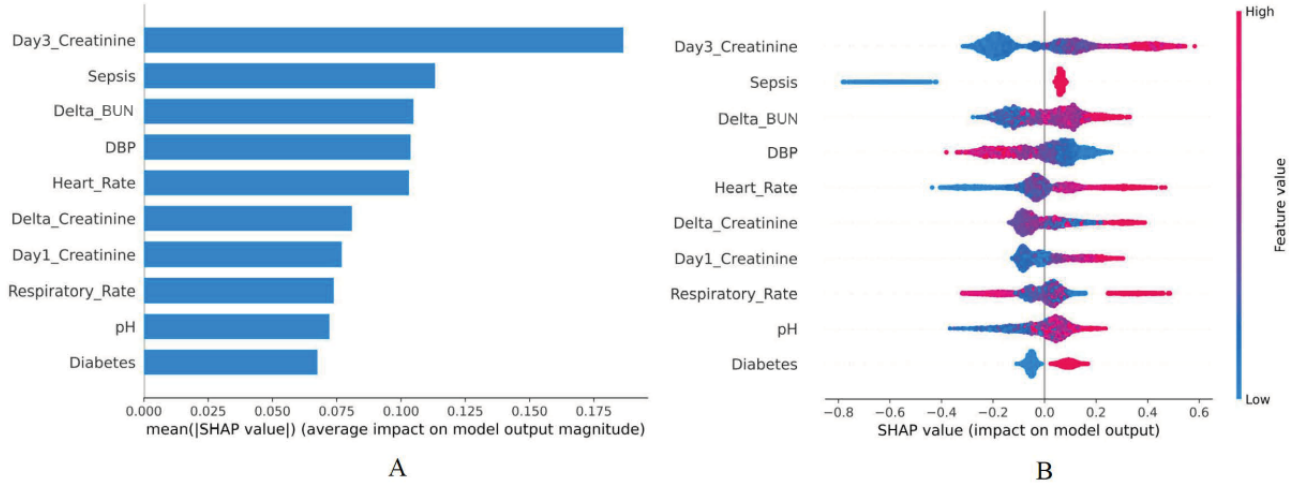

Model Application

We deployed the LightGBM algorithm as an online app because the LightGBM AKD risk model had a relatively high AUROC in the training, internal validation, and external cohorts. After the 10-fold cross-validation grid search, the LightGBM hyperparameters were finally tuned as follows: "num_leaves": 10, "max_depth": 5, "max_bin": 135, "min_data_in_leaf": 11, "feature_fraction": 1.0, "bagging_fraction": 1.0, "bagging_freq": 45, "lambda_l1": 0.0, "lambda_l2": 0.001, "min_split_gain": 0.4. Further, a web-based app for predicting AKD risk in the elderly was designed, which could be accessed online at any time by medical staff or patients (Multimedia Appendix 5) [19]. For an elderly patient with AKI being diagnosed for the first time in the ICU, physicians collected and input all variables' values correctly in Multimedia Appendix 5 and then clicked the Predict button to obtain the predicted result (AKD or non-AKD) during hospitalization. Moreover, users could enter variables' values and the author's email address and click the Feedback button, enabling new data to be sent to the author to facilitate model iteration. When the result showed that the patient was at high risk of AKD, early intervention could be implemented based on the partial dependence plots in Figure 4 and interventionable indicators might be controlled as close to the cutoff value as possible to prevent the progression of AKI and reduce the risk of AKD.

XSL•FO
RenderX

Figure 4. Partial dependence plots of LightGBM model for predicting AKD risk. AKD: acute kidney disease; BUN: blood urea nitrogen; DBP: diastolic blood pressure; LightGBM: Light Gradient Boosting Machine.

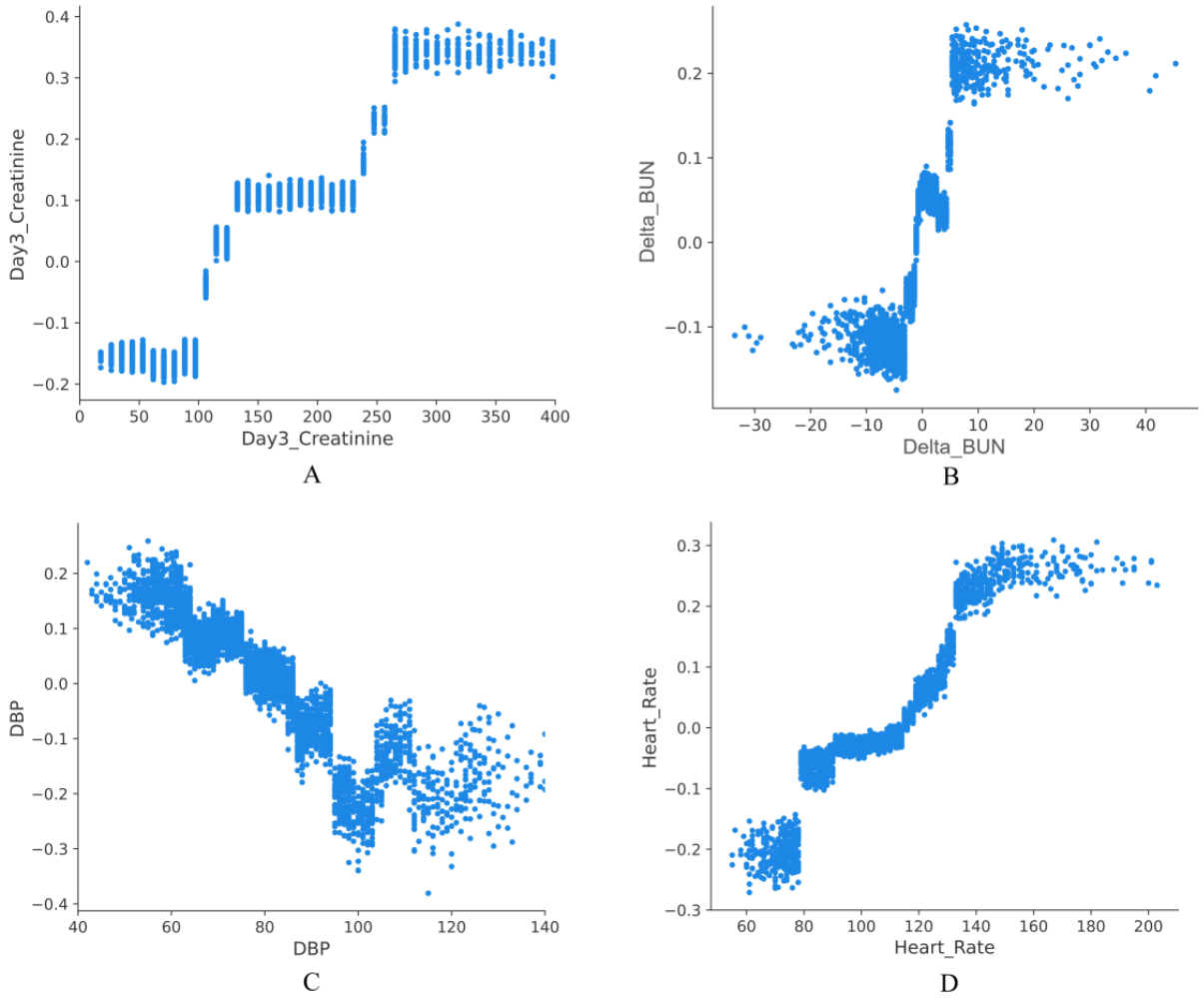

Model 2 : Predicting Prognostic Mortality in Elderly Patients With AKD

Baseline Characteristics

In this study on predicting the prognostic mortality in elderly patients with AKD, a total of 2661 elderly patients with AKD from the MIMIC-IV database (training cohort) and 186 from Xiangya Hospital (external validation cohort) were screened out and enrolled. The in-hospital mortality of elderly patients with AKD was 29.6% (788/2661) in the training cohort and 41.3% (77/186) in the external validation cohort. Multimedia Appendix 6 provides the differences in baseline characteristics between the 2 cohorts stratified by in-hospital death. In the MIMIC-IV cohort, compared with survivors, patients who died in the hospital were more likely to be older (P<.05), with a higher proportion of comorbidities (sepsis, CKD, and CLDl P<.05); a higher proportion of interventions (RRT and vasopressor use; P<.05); a higher heart rate and lower SBP in terms of vital signs (P<.05); and higher WBC count, potassium, anion gap, glucose, BUN on day 1, creatinine on day 1, BUN on day 3, creatinine on day 3, delta BUN, and delta creatinine and lower RBC count, hemoglobin, hematocrit, and $PaCO_2$ in terms of laboratory tests (P<.05). Additionally, in the Xiangya Hospital cohort, sepsis, RRT, vasopressor use, heart rate, anion gap, and BUN on day 1 had similar statistical trends to those in the MIMIC-IV cohort (P<.05).

Model Comparison

The performance of the in-hospital death prediction model for AKD in elderly patients in the training and external cohorts is presented in Table 2. In the training cohort, the best-performing algorithm was XGBoost, with AUROC=0.870 (95% CI 0.853-0.886), sensitivity=0.772 (95% CI 0.752-0.791), specificity=0.793 (95% CI 0.763-0.821), PPV=0.594 (95% CI 0.564-0.624), and NPV=0.899 (95% CI 0.883-0.913). In the external validation cohort, the LRM provided the best prediction, with AUROC=0.772 (95% CI 0.701-0.843), sensitivity=0.706 (95% CI 0.612-0.790), specificity=0.740 (95% CI 0.628-0.834), PPV=0.640 (95% CI 0.532-0.739), and NPV=0.794 (95% CI 0.700-0.869). However, comprehensively comparing the prediction performance of the training cohort and the generalization of the validation cohort, the LightGBM algorithm showed good overall performance, as demonstrated by an AUROC of 0.861 (95% CI 0.843-0.878) in the training cohort, 0.868 (95% CI 0.851-0.885) in the internal validation cohort, and 0.746 (95% CI 0.673-0.820) in the external cohort, in accordance with the ROC curves in Figure 5A, Multimedia Appendix 7, and Figure 5B. Figure 5C indicates that in DCA, when the probability of death during hospitalization reached 10%, the net benefit of intervention measures was 0.2,

XSL•FO
RenderX

suggesting good efficacy of LightGBM. Moreover, the calibration curve in Figure 5D shows that the predicted curve of the model surrounded the actual probability line, indicating relative accuracy.

Table 2. Performance of the AKD[a] prognostic mortality models for elderly patients.

Cohort and models	AUROC[b] (95% CI)	Cutoff	Sensitivity (95% CI)	Specificity (95% CI)	PPV[c] (95% CI)	NPV[d] (95% CI)
Training cohort						
LRM[e]	0.698 (0.675-0.721)	0.308	0.683 (0.661-0.704)	0.619 (0.584-0.653)	0.451 (0.421-0.481)	0.810 (0.790-0.829)
XGBoost[f]	0.870 (0.853-0.886)	0.312	0.772 (0.752-0.791)	0.793 (0.763-0.821)	0.594 (0.564-0.624)	0.899 (0.883-0.913)
LightGBM[g]	0.861 (0.843-0.878)	0.334	0.798 (0.779-0.816)	0.754 (0.722-0.784)	0.610 (0.579-0.641)	0.885 (0.869-0.899)
MLP[h]	0.731 (0.709-0.753)	0.332	0.737 (0.716-0.757)	0.603 (0.568-0.637)	0.491 (0.459-0.523)	0.815 (0.796-0.833)
RF[i]	0.844 (0.826-0.862)	0.351	0.887 (0.872-0.901)	0.632 (0.597-0.666)	0.702 (0.667-0.736)	0.851 (0.835-0.867)
KNN[j]	0.717 (0.695-0.740)	0.313	0.662 (0.640-0.683)	0.648 (0.614-0.682)	0.447 (0.418-0.476)	0.817 (0.797-0.837)
Internal validation cohort						
LRM	0.720 (0.697-0.742)	0.301	0.676 (0.642-0.709)	0.669 (0.647-0.690)	0.449 (0.420-0.479)	0.838 (0.818-0.856)
XGBoost	0.810 (0.790-0.830)	0.290	0.793 (0.763-0.822)	0.686 (0.665-0.707)	0.503 (0.474-0.532)	0.893 (0.876-0.908)
LightGBM	0.868 (0.851-0.885)	0.303	0.828 (0.799-0.854)	0.753 (0.733-0.773)	0.573 (0.543-0.602)	0.916 (0.901-0.929)
MLP	0.750 (0.728-0.771)	0.339	0.633 (0.598-0.667)	0.744 (0.724-0.764)	0.497 (0.465-0.529)	0.835 (0.817-0.853)
RF	0.706 (0.605-0.759)	0.330	0.625 (0.561-0.637)	0.761 (0.745-0.776)	0.567 (0.533-0.599)	0.776 (0.760-0.790)
KNN	0.725 (0.703-0.748)	0.278	0.762 (0.730-0.792)	0.558 (0.535-0.580)	0.408 (0.382-0.434)	0.854 (0.833-0.873)
External validation cohort						
LRM	0.772 (0.701-0.843)	0.285	0.706 (0.612-0.790)	0.740 (0.628-0.834)	0.640 (0.532-0.739)	0.794 (0.700-0.869)
XGBoost	0.698 (0.620-0.776)	0.385	0.706 (0.612-0.790)	0.636 (0.519-0.743)	0.605 (0.490-0.712)	0.733 (0.638-0.815)
LightGBM	0.746 (0.673-0.820)	0.312	0.716 (0.621-0.798)	0.740 (0.628-0.834)	0.648 (0.539-0.747)	0.796 (0.703-0.871)
MLP	0.770 (0.699-0.841)	0.339	0.789 (0.700-0.861)	0.701 (0.586-0.800)	0.701 (0.586-0.800)	0.789 (0.700-0.861)
RF	0.716 (0.639-0.792)	0.337	0.606 (0.507-0.698)	0.740 (0.628-0.834)	0.570 (0.467-0.669)	0.767 (0.664-0.852)
KNN	0.602 (0.519-0.685)	0.188	0.312 (0.227-0.408)	0.909 (0.822-0.963)	0.483 (0.399-0.567)	0.829 (0.679-0.928)

[a]AKD: acute kidney disease.
[b]AUROC: area under the receiver operating characteristic curve.
[c]PPV: positive predictive value.
[d]NPV: negative predictive value.
[e]LRM: logistic regression model.
[f]XGBoost: Extreme Gradient Boosting.
[g]LightGBM: Light Gradient Boosting Machine.
[h]MLP: multilayer perceptron.
[i]RF: random forest.
[j]KNN: K-nearest neighbor.

XSL•FO
RenderX

Figure 5. ROC curves, DCA , and calibration curves of AKD prognostic mortality prediction models. AKD: acute kidney disease; AUROC: area under the receiver operating characteristic curve; DCA: decision curve analysis; KNN: K-nearest neighbor; LightGBM: Light Gradient Boosting Machine; LRM: logistic regression model; MLP: multilayer perceptron; RF: random forest; ROC: receiver operating characteristic; XGBoost: Extreme Gradient Boosting.

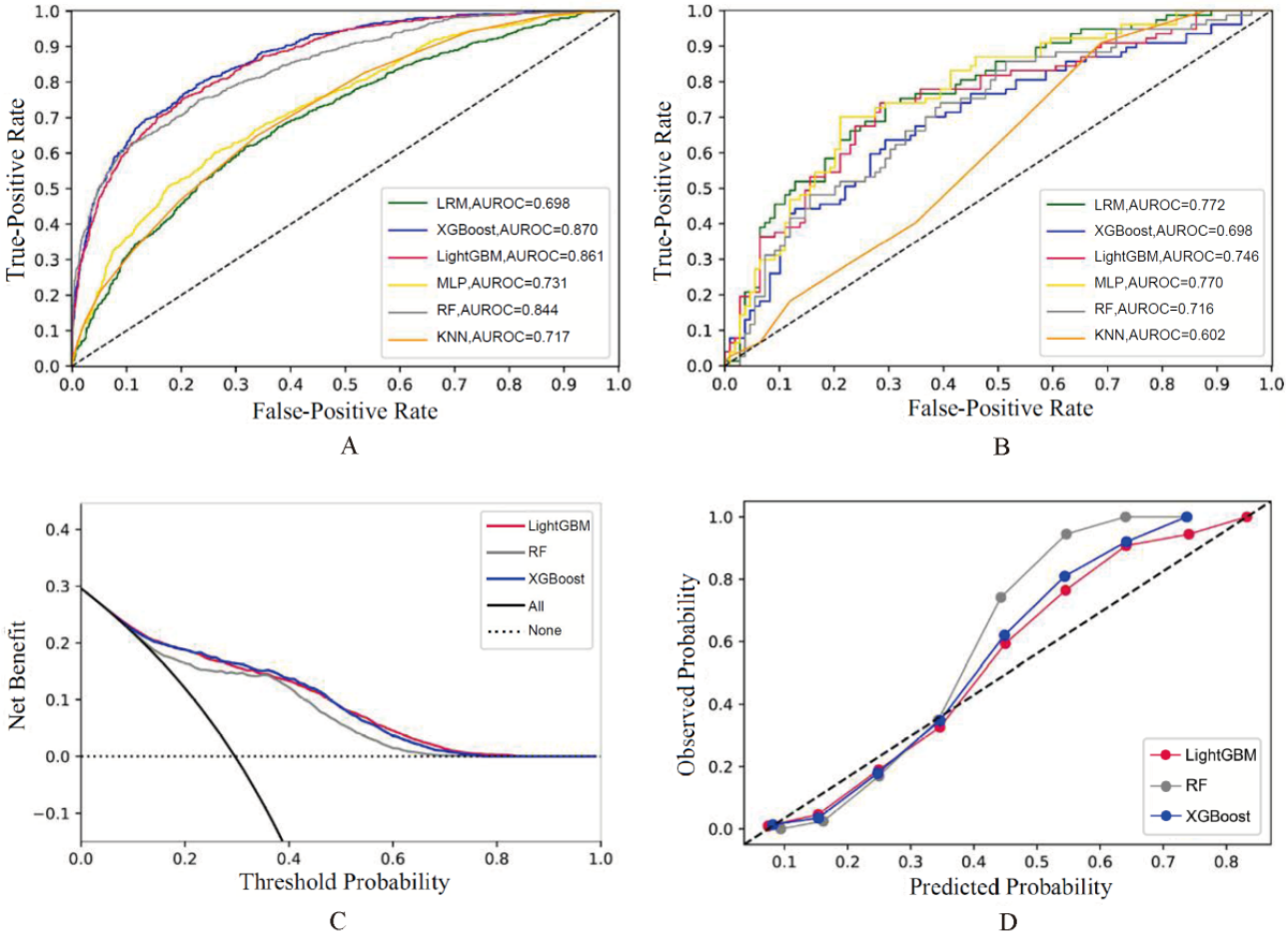

Model Interpretability

Using SHAP values, we performed a visual analysis of a model for predicting AKD prognostic mortality of the elderly. Figure 6A shows the top 10 predictors of in-hospital death in patients with AKD, as follows: age, BUN on day 1, vasopressor use, BUN on day 3, $PaCO_2$, RRT, delta creatinine, RBC count, respiratory rate, and creatinine on day 1. Figure 6B indicates a more detailed representation of the positive and negative relationships between features and outcomes. The risk of death due to AKD during hospitalization was positively associated with the following features: older age, higher BUN on day 1, use of vasopressors, higher BUN on day 3, higher $PaCO_2$, use of RRT, higher delta creatinine, and creatinine on day 1. The RBC count and respiratory rate were higher in hospitalized survivors among elderly patients with AKD. For the first 4 continuous variables in Figure 6A, partial dependence plots were drawn (Figure 7). According to Figure 7A, the probability of in-hospital death increased from 0 when the patient reached 75 years of age. Similarly, Figures 7B, 7C, and 7D show that the cutoff values of BUN on day 1, BUN on day 3 and $PaCO_2$ affecting the risk of death were 15 mmol/L, 10 mmol/L, and 45 mmHg, respectively, which might contribute to guiding patients' management and reducing the in-hospital risk of death for those with AKD.

XSL•FO
RenderX

Figure 6. Feature-ranking plots (A) and summary plots (B) of LightGBM for predicting prognostic mortality in AKD. AKD: acute kidney disease; BUN: blood urea nitrogen; $PaCO_2$: partial pressure of carbon dioxide; LightGBM: Light Gradient Boosting Machine; RBC: red blood cell; RRT: renal replacement therapy; SHAP: Shapley additive explanation.

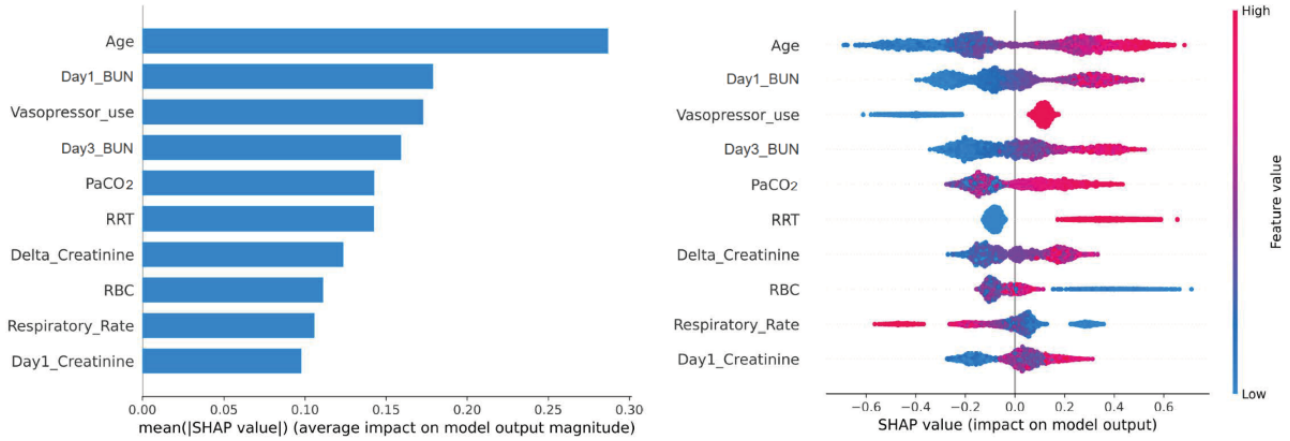

Model Application

We chose the LightGBM algorithm, which exhibited good AUROC values in the training and validation cohorts, to deploy the prognostic mortality function in the online version of the AKD model. The optimal combination of hyperparameters for the LightGBM prognostic model was as follows: “num_leaves”: 10, “max_depth”: 4, “max_bin”: 35, “min_data_in_leaf”: 100, “feature_fraction”: 1.0, “bagging_fraction”: 0.7, “bagging_freq”: 5, “lambda_l1”: 0.0, “lambda_l2”: 0.1, “min_split_gain”: 0.0. The web-based app to predict in-hospital death in elderly patients with AKD could be accessed online (Multimedia Appendix 8) [20]. When elderly patients were diagnosed with AKD for the first time, we entered all the indicators correctly on the web page and clicked the Predict button to predict the prognosis (death or survival) of elderly patients with AKD during hospitalization. Additionally, if users found a prediction error, they could enter the variable’s value and their own email address and then click the Feedback button, enabling the corresponding data to be automatically sent to the author’s email address. Using this feedback function could facilitate the collection of new data for model iterations. At the same time, according to the cutoff value shown in the partial dependence plots in Figure 7, targeted interventions were performed on patients at risk of death due to AKD, with the potential to improve the survival of patients with AKD.

XSL•FO
RenderX

Figure 7. Partial dependence plots of LightGBM for predicting prognostic mortality in patients with AKD. AKD: acute kidney disease; BUN: blood urea nitrogen; LightGBM: Light Gradient Boosting Machine; $PaCO_2$: partial pressure of carbon dioxide.

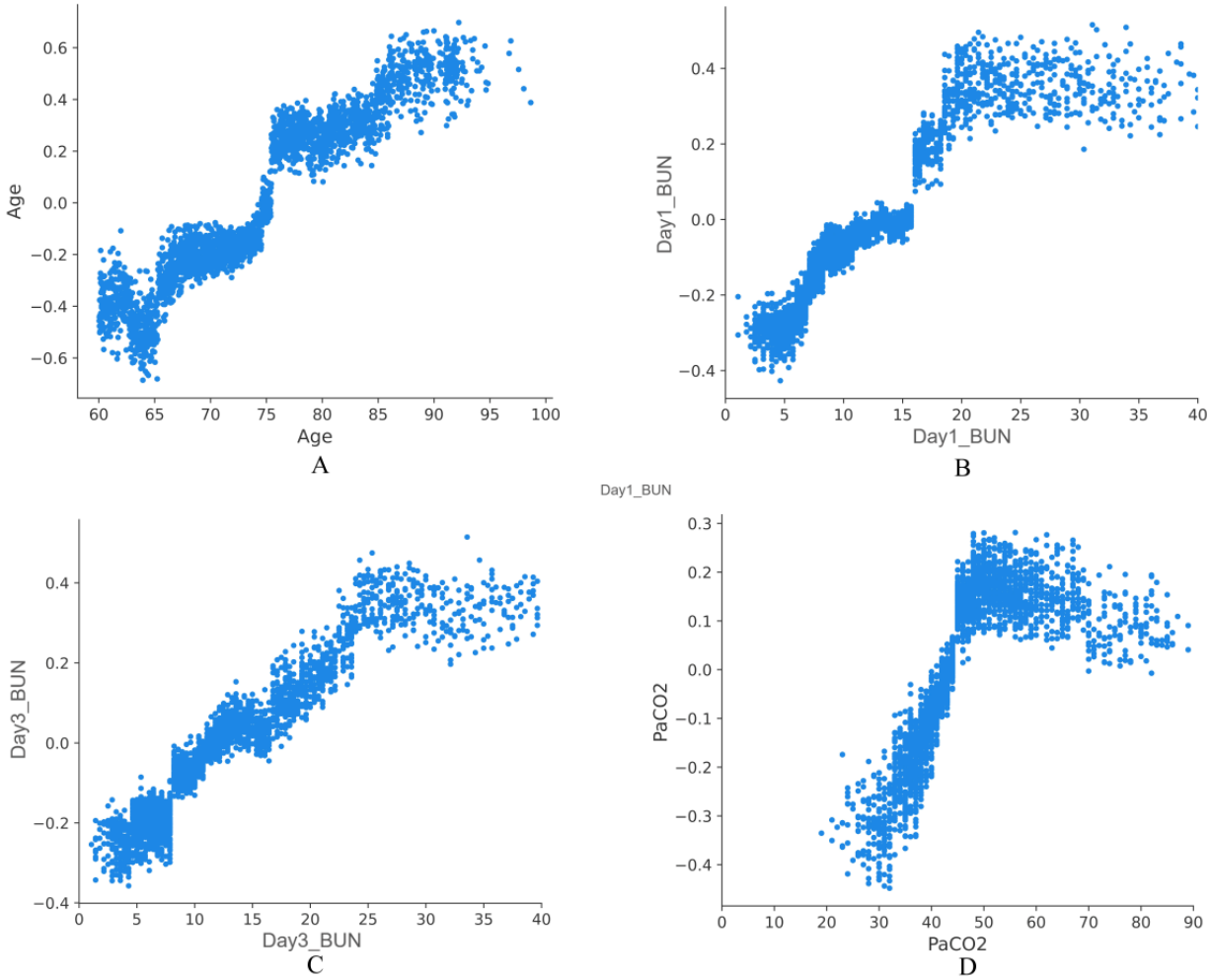

Discussion

Principal Findings

Predicting the Risk of AKD in the Elderly

As part of this study, we focused on model construction and feature analysis for AKD risk during hospitalization, and LightGBM was selected as the best algorithm for online deployment (training cohort AUROC=0.844, 95% CI 0.831-0.857; validation cohort AUROC=0.755, 95% CI 0.699-0.811). To the best of our knowledge, our study was the first to analyze the risk characteristics of AKD in critically ill elderly patients during hospitalization and to develop an easy-to-use online AKD risk identification app.

In addition to basic information, comorbidities, vital signs, and laboratory indicators on day 1 of AKI diagnosis, some indicators on day 3 and their fluctuations were also selected as features, including creatinine on day 3, BUN on day 3, delta BUN, and delta creatinine. A previous study found that the maximum creatinine level is reached on day 3 within 1-5 days after cardiac surgery in elderly patients [21]. Treiber et al [22] demonstrated that in neonatal patients with perinatal hypoxia, the AUROC of serum creatinine on day 3 after birth as a single predictor for AKI is 0.660, indicating a certain predictive value. Similar to these studies, our study revealed that serum creatinine on day 3 is higher than that on day 1 of AKD diagnosis in the elderly, ranking first in the feature of the AKD risk prediction model. Thus, serum creatinine on day 3 might be considered a focused experimental indicator for clinical research on patients with AKD in the ICU.

Delta BUN is commonly used to evaluate changes in renal function; however, the definition of the specific delta BUN varies. According to a study on patients with acute heart failure, delta BUN refers to the difference between the day before and after the administration of loop diuretics, but there was no statistical difference between the treatment and control groups (P>.05) [23]. Moreover, delta BUN was defined as the difference between 1 year after transplantation and at transplantation to evaluate renal function in a retrospective study conducted by Ewald et al [24]. In our study, we found that delta BUN (day 3 – day 1 after AKI diagnosis) is significantly positively correlated with AKD in elderly patients, with higher BUN on day 3 than on day 1 (delta BUN>0). Wu et al [25] also observed a gradual increase in BUN after AKI, in which BUN peaked at day 3 following cisplatin-induced AKI. Additionally, we determined that delta creatinine and creatinine on day 1 are associated with an increased risk of AKD during hospitalization. In a prospective study on adult patients after cardiac surgery, researchers defined delta creatinine as baseline – first postoperative creatinine and concluded that delta creatinine

XSL•FO
RenderX

combined with biomarkers has a good predictive effect on mortality[24]. Furthermore, Garner et al [26] defined delta creatinine to be higher than 26 μmol/L within 30 days of admission, enabling 98% of hospitalized patients with AKI to be identified.

At present, many studies have examined the factors associated with sepsis-related AKI, such as age, CKD, diabetes, infective endocarditis, and intra-abdominal infections [27-30]. However, there are relatively few studies conducted on sepsis and AKD. According to a single-center retrospective study, 46.9% of patients with sepsis developed AKD; in other words, sepsis is a critical factor contributing to the development of AKD in patients with AKI [31]. As a result of this study, it was also found that sepsis has a significant influence on renal function recovery of elderly patients with AKI.

Renal dysfunction is primarily caused by insufficient renal perfusion, indicating that improving the patient's hemodynamics to increase perfusion pressure might be an effective strategy for reversing kidney damage [32]. Previously, it was demonstrated that the DBP might be a valuable target for hemodynamic therapy in AKI by affecting renal perfusion [33]. Additionally, we found that the DBP is a major factor in the occurrence of AKD in elderly patients with AKI and that the risk of AKD gradually increases as the DBP decreases from 80 mmHg. However, a study of patients with severe coronary artery disease found that the risk of AKD is higher when the DBP is less than 50 mmHg [34]. This difference in the cutoff of the DBP for predicting AKD might reflect varying patient populations. As an indicator of overall health, the heart rate is affected by many factors, including pain stimulation, temperature, blood volume, and inflammatory responses. In a randomized controlled trial of β-receptor blockers in heart failure, maintaining a heart rate of 60 beats/minute was found to be beneficial to patient outcomes [35]. Additionally, a heart rate higher than 100 beats/minute might be a predictor of sepsis in patients not on advanced life support [36]. Our study also revealed that AKD is more likely to develop when the heart rate exceeds 110 beats/minute.

There is evidence that an abnormal respiratory rate could interfere with the baroreceptor reflex and cardiovascular variability [37]. We also found that an excessively high or low respiratory rate might adversely affect renal function recovery and lead to AKD in elderly patients with AKI. Metabolic acidosis is a common and life-threatening homeostatic disorder in the ICU, especially in patients with sepsis [38]. Furthermore, acidosis-related hemodynamic changes and decreased pH also contribute to the risk of AKI [39]. However, there has been relatively little attention given to metabolic alkalosis resulting from mass gastric fluid loss, a compensatory response to respiratory acidosis, or excess diuresis in critically ill patients. In a retrospective study of patients with septic shock, metabolic alkalosis was a significant predictor of the length of stay [40]. Likewise, we found that elevated pH is also a predictor of patients with AKD in the ICU, suggesting persistent renal impairment. Diabetes as a chronic disease is preventable and controllable. Currently, some studies have indicated that AKI is more common among patients with diabetes and that diabetes might increase AKI risk [41]. According to a national study of hospitalization trends in AKI in the United States between 2000 and 2014, the incidence of AKI among patients with diabetes was significantly higher than among patients without diabetes [42]. Our study also demonstrated that diabetes contributes to the development of AKD in elderly patients with AKI in the ICU.

Predicting Prognostic Mortality in Elderly Patients With AKD

After analyzing and predicting the risk of AKD during hospitalization for elderly patients with AKI in the ICU, a further machine learning prediction study was conducted on the hospital prognostic mortality of patients with AKD. Finally, the LightGBM algorithm was selected and deployed as a user-friendly web app, which performed well in both the training (AUROC=0.861, 95% CI 0.843-0.878) and external validation (AUROC=0.746, 95% CI 0.673-0.820) cohorts. In our opinion, this study was the first to construct and validate online machine learning models for continuously predicting the AKD risk and prognostic mortality in elderly patients.

Notably, we found that among the top 10 significant variables for predicting the occurrence of AKD in patients with AKI and predicting the prognostic mortality in AKD, delta creatinine, creatinine on day 1, and respiratory frequency all had good predictive values. In the prediction of hospital death in the elderly with AKD, creatinine on day 1 following renal injury was proportional to the likelihood of death. Some studies have shown that serum creatinine and mortality risk are significantly correlated. Thongprayoon et al [43] concluded that the serum creatinine level is a reliable predictor of mortality in critically ill patients. According to a retrospective study by Pooja et al [44], hepatorenal syndrome–related death is independently affected by high serum creatinine levels. Further, our study observed a positive correlation between delta creatinine and hospital mortality in elderly patients with AKD, which also appeared relevant to the risk of developing AKD in elderly patients with AKI. In patients with liver cirrhosis, researchers have found that creatinine variability (ie, delta creatinine) can serve as an effective indicator for predicting mortality [45]. Bradypnea is often seen in patients with central respiratory failure, sleep apnea syndrome, and high intracranial pressure. There is evidence to suggest that slow breathing can lead to death in patients with many diseases, including traumatic brain injury and stroke [46,47]. Additionally, Gooneratne et al [48] revealed that slow breathing with drowsiness results in increased mortality in the elderly. In this study, we also found that slow breathing is highly predictive of mortality in elderly patients with AKD.

In previous studies, age has been found to be a critical factor in the development of AKI [49]. Further, a retrospective study of Chinese multicenter patients with AKI revealed that age is an independent predictor of AKD progression in the logistic regression model but is not considered a risk factor for death [9]. As found in our study, age ranked first among the factors influencing in-hospital death in elderly patients with AKD, with patients aged over 75 years showing higher mortality and those 60-75 years old not appearing to present a significant risk of death. Patients in the ICU with septic shock or cardiogenic

XSL•FO
RenderX

shock usually require vasopressors, such as norepinephrine, epinephrine, and dopamine. Plurad et al [50] found that early administration of vasopressors in the ICU is independently associated with the risk of death regardless of the volume status at admission. In our study, vasopressor use was also a key predictor of mortality in elderly patients with AKD. Research has indicating that BUN is closely related to mortality risk in those with critical illness, acute pancreatitis, and heart failure [51-53]. In our study, we found that BUN on day 1 and day 3 after AKI diagnosis contributes to the risk of death due to AKD. Importantly, the cutoff values for BUN on day 1 and day 3 to predict in-hospital death were determined to be 15 and 20 mmol/L, respectively. However, Wernly et al [54] calculated 9.7 mmol/L as the optimal cutoff for BUN using Youden's index to predict patients' mortality in the ICU. Among patients with acute myocardial infarction complicated by cardiogenic shock, Zhu et al [55] observed that patients with BUN levels higher than 8.95 mmol/L on admission have an adverse short-term outcome [55].

As the main indicator of respiratory health and acid-base homeostasis, $PaCO_2$ levels higher than 45 mmHg often indicate the presence of hypercapnia. In a prospective observational study, patients with hypercapnia experienced higher in-hospital and long-term mortality [56]. Additionally, we observed that elderly patients with AKD who had $PaCO_2$ levels higher than 45 mmHg are more likely to die during hospitalization. RRT has been widely used as an effective intervention in patients with AKI, acute severe pancreatitis, and poisoning. Our study suggested that elderly patients with AKD requiring RRT might be at higher risk of in-hospital death, similar to the fact that patients with AKI in need for RRT usually have poorer survival and less renal function recovery, although RRT could delay or even stop this adverse process [57]. Based on our findings, AKD is associated with decreased RBC counts, which are observed in aplastic anemia, iron deficiency anemia, and massive bleeding. Recently, the RBC distribution width (RDW) has been widely regarded as a predictor of prognosis, especially in patients with coronary heart disease, AKI, and CKD [58-60]. Nevertheless, because of the high missing proportion of the RDW, we did not include it in our constructed model.

Strengths

We focused on elderly individuals with AKD for the first time, identified features affecting AKD risk and prognostic mortality, and developed 2 web-based prediction apps. After the users input the apps' URLs on the mobile phone or computer and manually entered the variables' values, they could click the Predict button to obtain the predictions or the Feedback button to send us new data. Although our online apps are easy to use, the calibration tool deployed by Sun et al [61] is more convenient and can be automated for use at different hospitals without manual data preparation, which could serve as a reference for further iterative development. Of note, data sets from Xiangya Hospital (China) were used for external validation, with good performance. However, the performance of the AKI prediction model in different sites has shown significant degradation [61], which might be due to the following measures we took to minimize performance degradation caused by a data shift. First, we adopted relatively strict inclusion and exclusion criteria to reduce the heterogeneity of enrolled patients. In addition, AKI and AKD were defined based on laboratory measurements, which prevented errors in medical record text recognition. Third, the units of the variables in the MIMIC-IV and Xiangya Hospital cohorts were unified. Finally, we added the Feedback button to 2 online forecasting apps to gather new training data through user feedback, that is, coping with data shifts by adhering to the fundamental principle of increasing training data.

Limitations

However, there are still some limitations. First, since detailed information about patients after discharge was lacking, the emphasis was placed on AKD diagnosis and prognosis during hospitalization. Second, the prediction models were based on machine learning classification algorithms, which could only identify a high or a low risk of AKD and patients' survival but could not display detailed risk values. Finally, although the 2 prediction models constructed in this study were externally validated and demonstrated good generalization abilities, additional variables, such as biomarkers, were needed to ensure better performance, as well as prospective experiments to further evaluate the online apps.

Conclusion

In conclusion, 2 online apps with machine learning algorithms were successfully constructed and deployed for predicting the AKD risk and prognostic mortality in elderly patients. SHAP can intuitively explain the rankings of importance, threshold values for partial features, and positive or negative correlations between features and outcomes, thereby aiding medical staff in early identification and targeted management to promote renal function recovery and patient survival to a certain extent.

Acknowledgments

The author thanks all the contributors of the MIMIC-IV database and all the enrolled patients, in appreciation of their contribution to the research of medical big data. This study was supported by the National Key R&D Program of China (no. 2020YFC2005000), the Natural Science Foundation of the Hunan Province of China (no. 2020JJ4929), and the Research Project of the Hunan Provincial Health Commission of China (no. 202217015418).

Data Availability

The data sets used and analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

ML and SH contributed equally to this work. SZ and CH were responsible for study design; BZ and QH for data extraction; ML and FL for data analysis; ML and SH for model construction and application; and ML and SZ for manuscript preparation. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The timeline plot of AKI and AKD during hospitalization. AKD: acute kidney disease; AKI: acute kidney injury.
[PDF File (Adobe PDF File), 172 KB - jmir_v26i1e51354_app1.pdf]

Multimedia Appendix 2
The optimal hyperparameters.
[PDF File (Adobe PDF File), 102 KB - jmir_v26i1e51354_app2.pdf]

Multimedia Appendix 3
The characteristics of elderly patients with AKI. AKI: acute kidney injury.
[PDF File (Adobe PDF File), 151 KB - jmir_v26i1e51354_app3.pdf]

Multimedia Appendix 4
The ROC curves in the internal validation of AKD risk prediction models.
[PDF File (Adobe PDF File), 586 KB - jmir_v26i1e51354_app4.pdf]

Multimedia Appendix 5
The internet-based app of the LightGBM model for predicting the AKD risk. AKD: acute kidney disease; LightGBM: Light Gradient Boosting Machine.
[PDF File (Adobe PDF File), 240 KB - jmir_v26i1e51354_app5.pdf]

Multimedia Appendix 6
The characteristics of elderly patients with AKD. AKD: acute kidney disease.
[PDF File (Adobe PDF File), 151 KB - jmir_v26i1e51354_app6.pdf]

Multimedia Appendix 7
The ROC curves in the internal validation of prognostic mortality prediction models.
[PDF File (Adobe PDF File), 593 KB - jmir_v26i1e51354_app7.pdf]

Multimedia Appendix 8
The internet-based app of the LightGBM model for predicting AKD mortality. AKD: acute kidney disease; LightGBM: Light Gradient Boosting Machine.
[PDF File (Adobe PDF File), 243 KB - jmir_v26i1e51354_app8.pdf]

References

1. Ronco C, Bellomo R, Kellum JA. Acute kidney injury. Lancet 2019 Nov 23;394(10212):1949-1964. [doi: 10.1016/S0140-6736(19)32563-2] [Medline: 31777389]
2. Hoste EAJ, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. Intensive Care Med 2015 Aug 11;41(8):1411-1423. [doi: 10.1007/s00134-015-3934-7] [Medline: 26162677]
3. Meyer D, Mohan A, Subev E, Sarav M, Sturgill D. Acute kidney injury incidence in hospitalized patients and implications for nutrition support. Nutr Clin Pract 2020 Dec 03;35(6):987-1000. [doi: 10.1002/ncp.10595] [Medline: 33140897]
4. Kane-Gill SL, Sileanu FE, Murugan R, Trietley GS, Handler SM, Kellum JA. Risk factors for acute kidney injury in older adults with critical illness: a retrospective cohort study. Am J Kidney Dis 2015 Jun;65(6):860-869. [doi: 10.1053/j.ajkd.2014.10.018] [Medline: 25488106]
5. Li Q, Zhao M, Wang X. The impact of transient and persistent acute kidney injury on short-term outcomes in very elderly patients. Clin Interv Aging 2017;12:1013-1020 [FREE Full text] [doi: 10.2147/CIA.S135241] [Medline: 28721029]

XSL•FO
RenderX

6. Lameire N, Kellum JA, KDIGO AKI Guideline Work Group. Contrast-induced acute kidney injury and renal support for acute kidney injury: a KDIGO summary (part 2). Crit Care 2013 Feb 04;17(1):205. [doi: 10.1186/cc11455] [Medline: 23394215]
7. See EJ, Polkinghorne KR, Toussaint ND, Bailey M, Johnson DW, Bellomo R. Epidemiology and outcomes of acute kidney diseases: a comparative analysis. Am J Nephrol 2021 Apr 27;52(4):342-350. [doi: 10.1159/000515231] [Medline: 33906191]
8. Chawla LS, Bellomo R, Bihorac A, Goldstein SL, Siew ED, Bagshaw SM, Acute Disease Quality Initiative Workgroup 16. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative (ADQI) 16 Workgroup. Nat Rev Nephrol 2017 Apr 27;13(4):241-257. [doi: 10.1038/nrneph.2017.2] [Medline: 28239173]
9. Xiao Y, Cheng W, Wu X, Yan P, Feng L, Zhang N, et al. Novel risk models to predict acute kidney disease and its outcomes in a Chinese hospitalized population with acute kidney injury. Sci Rep 2020 Sep 24;10(1):15636. [doi: 10.1038/s41598-020-72651-x] [Medline: 32973230]
10. Andonovic M, Traynor JP, Shaw M, Sim MA, Mark PB, Puxty KA. Short- and long-term outcomes of intensive care patients with acute kidney disease. EClinicalMedicine 2022 Feb;44:101291. [doi: 10.1016/j.eclinm.2022.101291] [Medline: 35198917]
11. Chen Y, Wu M, Mao C, Yeh Y, Chen T, Liao C, et al. Severe acute kidney disease is associated with worse kidney outcome among acute kidney injury patients. Sci Rep 2022 Apr 20;12(1):6492. [doi: 10.1038/s41598-022-09599-7] [Medline: 35444219]
12. He J, Lin J, Duan M. Application of machine learning to predict acute kidney disease in patients with sepsis associated acute kidney injury. Front Med (Lausanne) 2021 Dec 10;8:792974. [doi: 10.3389/fmed.2021.792974] [Medline: 34957162]
13. Chen Y, Jenq C, Hsu C, Yu Y, Chang C, Fan P, et al. Acute kidney disease and acute kidney injury biomarkers in coronary care unit patients. BMC Nephrol 2020 Jun 01;21(1):207. [doi: 10.1186/s12882-020-01872-z] [Medline: 32487168]
14. Hu X, Liu D, Qiao Y, Zheng X, Duan J, Pan S, et al. Development and Validation of a Nomogram Model to Predict Acute Kidney Disease After Nephrectomy in Patients with Renal Cell Carcinoma. Cancer Manag Res 2020;12:11783-11791 [FREE Full text] [doi: 10.2147/CMAR.S273244] [Medline: 33235506]
15. Zhou Y, Feng J, Mei S, Zhong H, Tang R, Xing S, et al. MACHINE LEARNING MODELS FOR PREDICTING ACUTE KIDNEY INJURY IN PATIENTS WITH SEPSIS-ASSOCIATED ACUTE RESPIRATORY DISTRESS SYNDROME. Shock 2023 Mar 01;59(3):352-359. [doi: 10.1097/SHK.0000000000002065] [Medline: 36625493]
16. Van den Broeck G, Lykov A, Schleich M, Suciu D. On the tractability of SHAP explanations. J Artif Intell Res 2022 Jun 23;74:851-886. [doi: 10.1613/jair.1.13283]
17. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323. [doi: 10.2196/jmir.5870] [Medline: 27986644]
18. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L, Mark R. MIMIC-IV (version 2). PhysioNet. 2022 Jun 12. URL: https://physionet.org/content/mimiciv/2.0/ [accessed 2024-04-23]
19. AKD prediction in elderly AKD patients. Heroku. URL: https://predict---akd-f59a631788c2.herokuapp.com/ [accessed 2024-04-23]
20. Mortality prediction in elderly AKD patients. Heroku. URL: https://predict---death-260031eeda1c.herokuapp.com/ [accessed 2024-04-23]
21. Ristikankare A, Pöyhiä R, Kuitunen A, Skrifvars M, Hämmäinen P, Salmenperä M, et al. Serum cystatin C in elderly cardiac surgery patients. Ann Thorac Surg 2010 Mar;89(3):689-694. [doi: 10.1016/j.athoracsur.2009.11.018] [Medline: 20172110]
22. Treiber M, Gorenjak M, Pecovnik Balon B. Serum cystatin-C as a marker of acute kidney injury in the newborn after perinatal hypoxia/asphyxia. Ther Apher Dial 2014 Feb 03;18(1):57-67. [doi: 10.1111/1744-9987.12054] [Medline: 24499085]
23. Yamamoto T, Miura S, Shirai K, Urata H. Renoprotective benefit of tolvaptan in acute decompensated heart failure patients with loop diuretic-resistant status. J Clin Med Res 2019 Jan;11(1):49-55. [doi: 10.14740/jocmr3671] [Medline: 30627278]
24. Ewald C, Swanson BJ, Vargas L, Grant WJ, Mercer DF, Langnas AN, et al. Including colon in intestinal transplantation: a focus on post-transplant renal function - a retrospective study. Transpl Int 2020 Feb 10;33(2):142-148. [doi: 10.1111/tri.13523] [Medline: 31523865]
25. Wu C, Liu L, Xu W, Jian J, Zhang N, Wang X, et al. [Correlation analysis of inflammatory response and Klotho expression in renal tissue of mice with acute renal injury induced by cisplatin]. Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 2019 Aug;35(8):702-706. [Medline: 31638567]
26. Garner AE, Lewington AJP, Barth JH. Detection of patients with acute kidney injury by the clinical laboratory using rises in serum creatinine: comparison of proposed definitions and a laboratory delta check. Ann Clin Biochem 2012 Jan 30;49(Pt 1):59-62. [doi: 10.1258/acb.2011.011125] [Medline: 22130632]
27. Bagshaw SM, Lapinsky S, Dial S, Arabi Y, Dodek P, Wood G, Cooperative Antimicrobial Therapy of Septic Shock (CATSS) Database Research Group. Acute kidney injury in septic shock: clinical outcomes and impact of duration of hypotension prior to initiation of antimicrobial therapy. Intensive Care Med 2009 May 9;35(5):871-881. [doi: 10.1007/s00134-008-1367-2] [Medline: 19066848]

XSL•FO
RenderX

28. Bagshaw S, Uchino S, Bellomo R, Morimatsu H, Morgera S, Schetz M, BeginningEnding Supportive Therapy for the Kidney (BEST Kidney) Investigators. Septic acute kidney injury in critically ill patients: clinical characteristics and outcomes. Clin J Am Soc Nephrol 2007 May;2(3):431-439. [doi: 10.2215/CJN.03681106] [Medline: 17699448]
29. Hoste E, Lameire N, Vanholder R, Benoit D, Decruyenaere J, Colardyn F. Acute renal failure in patients with sepsis in a surgical ICU: predictive factors, incidence, comorbidity, and outcome. J Am Soc Nephrol 2003 Apr;14(4):1022-1030. [doi: 10.1097/01.asn.0000059863.48590.e9] [Medline: 12660337]
30. Fan Z, Jiang J, Xiao C, Chen Y, Xia Q, Wang J, et al. Construction and validation of prognostic models in critically Ill patients with sepsis-associated acute kidney injury: interpretable machine learning approach. J Transl Med 2023 Jun 22;21(1):406. [doi: 10.1186/s12967-023-04205-4] [Medline: 37349774]
31. Flannery AH, Li X, Delozier NL, Toto RD, Moe OW, Yee J, et al. Sepsis-associated acute kidney disease and long-term kidney outcomes. Kidney Med 2021 Jul;3(4):507-514.e1. [doi: 10.1016/j.xkme.2021.02.007] [Medline: 34401718]
32. Gattinoni L, Brazzi L, Pelosi P, Latini R, Tognoni G, Pesenti A, et al. A trial of goal-oriented hemodynamic therapy in critically ill patients. SvO2 Collaborative Group. N Engl J Med 1995 Oct 19;333(16):1025-1032. [doi: 10.1056/NEJM199510193331601] [Medline: 7675044]
33. Legrand M, Dupuis C, Simon C, Gayat E, Mateo J, Lukaszewicz A, et al. Association between systemic hemodynamics and septic acute kidney injury in critically ill patients: a retrospective observational study. Crit Care 2013 Nov 29;17(6):R278. [doi: 10.1186/cc13133] [Medline: 24289206]
34. Gong K, Xie X. An interpretable ensemble model of acute kidney disease risk prediction for patients in coronary care units. In: AI and Analytics for Smart Cities and Service Systems: Proceedings of the 2021 INFORMS International Conference on Service Science. Switzerland: Springer; 2021:76-90.
35. Kotecha D, Flather MD, Altman DG, Holmes J, Rosano G, Wikstrand J, Hjalmarson, Beta-Blockers in Heart Failure Collaborative Group. Heart rate and rhythm and the benefit of beta-blockers in patients with heart failure. J Am Coll Cardiol 2017 Jun 20;69(24):2885-2896. [doi: 10.1016/j.jacc.2017.04.001] [Medline: 28467883]
36. Wahab A, Smith RJ, Lal A, Flurin L, Malinchoc M, Dong Y, et al. CHARACTERISTICS AND PREDICTORS OF PATIENTS WITH SEPSIS WHO ARE CANDIDATES FOR MINIMALLY INVASIVE APPROACH OUTSIDE OF INTENSIVE CARE UNIT. Shock 2023 May 01;59(5):702-707 [FREE Full text] [doi: 10.1097/SHK.0000000000002112] [Medline: 36870069]
37. Radaelli A, Mancia G, Balestri G, Bonfanti D, Castiglioni P. Respiratory patterns and baroreflex function in heart failure. Sci Rep 2023 Feb 08;13(1):2220. [doi: 10.1038/s41598-023-29271-y] [Medline: 36755066]
38. Kellum JA, Bellomo R, Kramer DJ, Pinsky MR. Etiology of metabolic acidosis during saline resuscitation in endotoxemia. Shock 1998 May;9(5):364-368. [doi: 10.1097/00024382-199805000-00009] [Medline: 9617887]
39. Husain-Syed F, Slutsky AS, Ronco C. Lung-Kidney Cross-Talk in the Critically Ill Patient. Am J Respir Crit Care Med 2016 Aug 15;194(4):402-414. [doi: 10.1164/rccm.201602-0420CP] [Medline: 27337068]
40. Kreü S, Jazrawi A, Miller J, Baigi A, Chew M. Alkalosis in critically ill patients with severe sepsis and septic shock. PLoS One 2017 Jan 3;12(1):e0168563. [doi: 10.1371/journal.pone.0168563] [Medline: 28045915]
41. Advani A. Acute kidney injury: a bona fide complication of diabetes. Diabetes 2020 Nov;69(11):2229-2237. [doi: 10.2337/db20-0604] [Medline: 33082271]
42. Pavkov ME, Harding JL, Burrows NR. Trends in hospitalizations for acute kidney injury—United States, 2000–2014. MMWR Morb Mortal Wkly Rep 2018 Mar 16;67(10):289-293. [doi: 10.15585/mmwr.mm6710a2] [Medline: 29543788]
43. Thongprayoon C, Cheungpasitporn W, Kashani K. Serum creatinine level, a surrogate of muscle mass, predicts mortality in critically ill patients. J Thorac Dis 2016 May;8(5):E305-E311. [doi: 10.21037/jtd.2016.03.62] [Medline: 27162688]
44. Pooja B, Rajpurohit S, Nagaraju SP, Musunuri B, Bhat G, Shetty S. Clinical profile and outcome of hepatorenal syndrome in tertiary care hospital based on new ICA criteria: a retrospective study. J Clin Exp Hepatol 2022;12:S41. [doi: 10.1016/j.jceh.2022.07.111]
45. Cullaro G, Hsu C, Lai JC. Variability in serum creatinine is associated with waitlist and post-liver transplant mortality in patients with cirrhosis. Hepatology 2022 Oct 15;76(4):1069-1078. [doi: 10.1002/hep.32497] [Medline: 35357707]
46. Saadat S, Akbari H, Khorramirouz R, Mofid R, Rahimi-Movaghar V. Determinants of mortality in patients with traumatic brain injury. Ulus Travma Acil Cerrahi Derg 2012 May;18(3):219-224 [FREE Full text] [Medline: 22864713]
47. Bayir A, Ak A, Kara H, Sahin TK. Serum and cerebrospinal fluid magnesium levels, Glasgow Coma Scores, and in-hospital mortality in patients with acute stroke. Biol Trace Elem Res 2009 Jul 23;130(1):7-12. [doi: 10.1007/s12011-009-8318-9] [Medline: 19165425]
48. Gooneratne N, Richards K, Joffe M, Lam R, Pack F, Staley B, et al. Sleep disordered breathing with excessive daytime sleepiness is a risk factor for mortality in older adults. Sleep 2011 Apr 01;34(4):435-442. [doi: 10.1093/sleep/34.4.435] [Medline: 21461321]
49. Shen J, Chu Y, Wang C, Yan S. Risk factors for acute kidney injury after major abdominal surgery in the elderly aged 75 years and above. BMC Nephrol 2022 Jun 23;23(1):224. [doi: 10.1186/s12882-022-02822-7] [Medline: 35739472]
50. Plurad D, Talving P, Lam L, Inaba K, Green D, Demetriades D. Early vasopressor use in critical injury is associated with mortality independent from volume status. J Trauma 2011 Sep;71(3):565-70; discussion 570. [doi: 10.1097/TA.0b013e3182213d52] [Medline: 21908995]

XSL•FO
RenderX

51. Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig J, et al. Blood urea nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU. PLoS One 2018 Jan 25;13(1):e0191697. [doi: 10.1371/journal.pone.0191697] [Medline: 29370259]
52. Wu BU, Johannes RS, Sun X, Conwell DL, Banks PA. Early changes in blood urea nitrogen predict mortality in acute pancreatitis. Gastroenterology 2009 Jul;137(1):129-135. [doi: 10.1053/j.gastro.2009.03.056] [Medline: 19344722]
53. Cauthen CA, Lipinski MJ, Abbate A, Appleton D, Nusca A, Varma A, et al. Relation of blood urea nitrogen to long-term mortality in patients with heart failure. Am J Cardiol 2008 Jun 01;101(11):1643-1647. [doi: 10.1016/j.amjcard.2008.01.047] [Medline: 18489944]
54. Wernly B, Lichtenauer M, Vellinga NA, Boerma EC, Ince C, Kelm M, et al. Blood urea nitrogen (BUN) independently predicts mortality in critically ill patients admitted to ICU: A multicenter study. Clin Hemorheol Microcirc 2018;69(1-2):123-131. [doi: 10.3233/CH-189111] [Medline: 29758935]
55. Zhu Y, Sasmita BR, Hu X, Xue Y, Gan H, Xiang Z, et al. Blood urea nitrogen for short-term prognosis in patients with cardiogenic shock complicating acute myocardial infarction. Int J Clin Pract 2022 Mar 15;2022:9396088. [doi: 10.1155/2022/9396088] [Medline: 35685591]
56. Vonderbank S, Gibis N, Schulz A, Boyko M, Erbuth A, Gürleyen H, et al. Hypercapnia at Hospital Admission as a Predictor of Mortality. Open Access Emerg Med 2020;12:173-180 [FREE Full text] [doi: 10.2147/OAEM.S242075] [Medline: 32617025]
57. Cerdá J, Liu K, Cruz D, Jaber B, Koyner J, Heung M, AKI Advisory Group of the American Society of Nephrology. Promoting Kidney Function Recovery in Patients with AKI Requiring RRT. Clin J Am Soc Nephrol 2015 Oct 07;10(10):1859-1867 [FREE Full text] [doi: 10.2215/CJN.01170215] [Medline: 26138260]
58. Oh HJ, Park JT, Kim J, Yoo DE, Kim SJ, Han SH, et al. Red blood cell distribution width is an independent predictor of mortality in acute kidney injury patients treated with continuous renal replacement therapy. Nephrol Dial Transplant 2012 Feb 28;27(2):589-594. [doi: 10.1093/ndt/gfr307] [Medline: 21712489]
59. Hu Y, Liu H, Fu S, Wan J, Li X. Red blood cell distribution width is an independent predictor of AKI and mortality in patients in the coronary care unit. Kidney Blood Press Res 2017 Dec 8;42(6):1193-1204. [doi: 10.1159/000485866] [Medline: 29227977]
60. Zhang T, Li J, Lin Y, Yang H, Cao S. Association between red blood cell distribution width and all-cause mortality in chronic kidney disease patients: a systematic review and meta-analysis. Arch Med Res 2017 May;48(4):378-385. [doi: 10.1016/j.arcmed.2017.06.009] [Medline: 28916240]
61. Sun H, Depraetere K, Meesseman L, Cabanillas Silva P, Szymanowsky R, Fliegenschmidt J, et al. Machine learning-based prediction models for different clinical risks in different hospitals: evaluation of live performance. J Med Internet Res 2022 Jun 07;24(6):e34295. [doi: 10.2196/34295] [Medline: 35502887]

Abbreviations

ADQI: Acute Disease Quality Initiative
AI: artificial intelligence
AKD: acute kidney disease
AKI: acute kidney injury
AUROC: area under the receiver operating characteristic curve
BUN: blood urea nitrogen
CKD: chronic kidney disease
CLD: chronic liver disease
CPD: chronic pulmonary disease
DBP: diastolic blood pressure
DCA: decision curve analysis
EHR: electronic health record
ESRD: end-stage renal disease
ICU: intensive care unit
KDIGO: Kidney Disease: Improving Global Outcomes
KNN: K-nearest neighbor
LightGBM: Light Gradient Boosting Machine
LRM: logistic regression model
MIMIC-IV: Medical Information Mart for Intensive Care IV
MLP: multilayer perceptron
MV: mechanical ventilation
NPV: negative predictive value
PPV: positive predictive value
$PaCO_2$: partial pressure of carbon dioxide

XSL•FO
RenderX

PaO2: partial pressure of oxygen
RBC: red blood cell
RDW: RBC distribution width
RF: random forest
ROC: receiver operating characteristic
RRT: renal replacement therapy
SBP: systolic blood pressure
SHAP: Shapley additive explanation
WBC: white blood cell
XGBoost: Extreme Gradient Boosting

Edited by T de Azevedo Cardoso; submitted 28.07.23; peer-reviewed by H Sun, F Zhang; comments to author 05.12.23; revised version received 23.01.24; accepted 17.04.24; published 01.05.24.

Please cite as:
Li M, Han S, Liang F, Hu C, Zhang B, Hou Q, Zhao S
Machine Learning for Predicting Risk and Prognosis of Acute Kidney Disease in Critically Ill Elderly Patients During Hospitalization: Internet-Based and Interpretable Model Study
J Med Internet Res 2024;26:e51354
URL: https://www.jmir.org/2024/1/e51354
doi:10.2196/51354
PMID:38691403

©Mingxia Li, Shuzhe Han, Fang Liang, Chenghuan Hu, Buyao Zhang, Qinlan Hou, Shuangping Zhao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Improving the Prognostic Evaluation Precision of Hospital Outcomes for Heart Failure Using Admission Notes and Clinical Tabular Data: Multimodal Deep Learning Model

Zhenyue Gao[1*], MS; Xiaoli Liu[2*], PhD; Yu Kang[3], MD; Pan Hu[2], MD; Xiu Zhang[3], MD; Wei Yan[2], MD; Muyang Yan[2], MD; Pengming Yu[3], MD; Qing Zhang[3], MD; Wendong Xiao[1], PhD; Zhengbo Zhang[2], PhD

[1]Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

[2]Center for Artificial Intelligence in Medicine, The General Hospital of People's Liberation Army, Beijing, China

[3]Department of Cardiology, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Zhengbo Zhang, PhD
Center for Artificial Intelligence in Medicine
The General Hospital of People's Liberation Army
28 Fuxing Road
Beijing, 100853
China
Phone: 86 010 68295454
Email: zhangzhengbo@301hospital.com.cn

Abstract

Background: Clinical notes contain contextualized information beyond structured data related to patients' past and current health status.

Objective: This study aimed to design a multimodal deep learning approach to improve the evaluation precision of hospital outcomes for heart failure (HF) using admission clinical notes and easily collected tabular data.

Methods: Data for the development and validation of the multimodal model were retrospectively derived from 3 open-access US databases, including the Medical Information Mart for Intensive Care III v1.4 (MIMIC-III) and MIMIC-IV v1.0, collected from a teaching hospital from 2001 to 2019, and the eICU Collaborative Research Database v1.2, collected from 208 hospitals from 2014 to 2015. The study cohorts consisted of all patients with critical HF. The clinical notes, including chief complaint, history of present illness, physical examination, medical history, and admission medication, as well as clinical variables recorded in electronic health records, were analyzed. We developed a deep learning mortality prediction model for in-hospital patients, which underwent complete internal, prospective, and external evaluation. The Integrated Gradients and SHapley Additive exPlanations (SHAP) methods were used to analyze the importance of risk factors.

Results: The study included 9989 (16.4%) patients in the development set, 2497 (14.1%) patients in the internal validation set, 1896 (18.3%) in the prospective validation set, and 7432 (15%) patients in the external validation set. The area under the receiver operating characteristic curve of the models was 0.838 (95% CI 0.827-0.851), 0.849 (95% CI 0.841-0.856), and 0.767 (95% CI 0.762-0.772), for the internal, prospective, and external validation sets, respectively. The area under the receiver operating characteristic curve of the multimodal model outperformed that of the unimodal models in all test sets, and tabular data contributed to higher discrimination. The medical history and physical examination were more useful than other factors in early assessments.

Conclusions: The multimodal deep learning model for combining admission notes and clinical tabular data showed promising efficacy as a potentially novel method in evaluating the risk of mortality in patients with HF, providing more accurate and timely decision support.

(J Med Internet Res 2024;26:e54363) doi:10.2196/54363

XSL•FO
RenderX

KEYWORDS

heart failure; multimodal deep learning; mortality prediction; admission notes; clinical tabular data; tabular; notes; deep learning; machine learning; cardiology; heart; cardiac; documentation; prognostic; prognosis; prognoses; predict; prediction; predictions; predictive

Introduction

Heart failure (HF), a syndrome of impaired heart function, represents the advanced stage of various cardiac conditions [1-3]. With its substantial influence on both morbidity and mortality, HF poses a formidable challenge to human health and societal progress [4-6].

As a potentially life-threatening condition, particularly when accompanied by advanced organ dysfunction or severe complications, a considerable portion of patients with HF may require immediate access to advanced, high-technology, life-saving care, which is typically available only in intensive care units (ICUs) [7]. Studies have indicated that approximately 10% to 51% of patients with HF admitted to hospitals in the United States are subsequently admitted to ICUs [8,9]. It has also been found that ICU-admitted patients with HF experience significantly higher adjusted in-hospital mortality rates compared to those admitted solely to hospitals [10]. The in-hospital mortality rate for patients with HF receiving treatment in an ICU has been reported as 10.6%, in contrast to the overall in-hospital mortality rate of 4% for all patients with HF [11]. Given this substantially higher mortality rate, accurate prediction of in-hospital mortality could empower physicians to implement early interventions and tailor individualized treatments [12,13]. Consequently, there is an increasing need for the development of predictive models that can effectively identify individuals at a heightened risk of mortality in the ICU.

Most previous research works have applied statistical analysis or machine learning techniques using structured administrative data from electronic health records to identify significant risk predictors that trigger adverse outcomes [14-19]. However, HF disease often develops rapidly, and while some sensitive biomarkers, such as N-terminal pro–b-type natriuretic peptide (NT-proBNP), tend to increase in reactivity after the disease progresses, their efficiency is limited due to their high cost and inability to be measured in real time [20,21]. Recently, there has been a growing acknowledgment of the importance of clinical narratives in clinical decision-making [22,23]. The narrative notes at admission, such as chief complaint, history of present illness, physical examination, medical history, and admission medication, play a central role in health care communication. They represent a more comprehensive and personalized account of patient history and assessments [24]. Harnessing the potential of clinical narratives can largely enhance patient care and contribute to the improvement of predictive models for prognosis [25,26]. Exploiting the potential of clinical narratives and modeling them by multimodal deep learning (DL) approaches can enhance the precision of patient care and contribute to the improvement of predictive models for in-hospital mortality.

We aim to design a multimodal DL model and explore the infusion approaches to improve evaluation performance using tabular data and admission notes. The cross-modal model, which characterizes textual, categorical, and continuous variables separately, significantly outperforms the unimodal models on the multicenter and prospective validation sets. We believe our findings will motivate data-centric studies to more precisely characterize the illness severity of patients with HF.

Methods

Study Design

An overview of the study flow is shown in Figure 1A. First, we acquired patients' admission notes and tabular data, and these two single modalities were separately embedded to obtain the feature and status representation. The categorical and continuous variables in tabular data were characterized separately. Next, a feature-fusing DL network was applied to integrate the two modalities and achieved the model development. Then, a fully connected DL network was used to predict the in-hospital outcome, our primary outcome of interest. Two postexplanation approaches were adopted to increase the credibility of the model. Finally, the internal, prospective, and external validation with multiple evaluation metrics were accomplished. Our study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines for prognosis studies.

XSL•FO
RenderX

Figure 1. (A) Workflow of the multimodal deep learning model with the fusion of textual and tabular data. (B) An overview of the inclusion criteria with all study cohorts. BERT: Bidirectional Encoder Representations from Transformers; eICU-CRD: eICU Collaborative Research Database; HF: heart failure; *ICD-10: International Classification of Diseases, Tenth Revision*; ICU: intensive care unit; MIMIC: Medical Information Mart for Intensive Care.

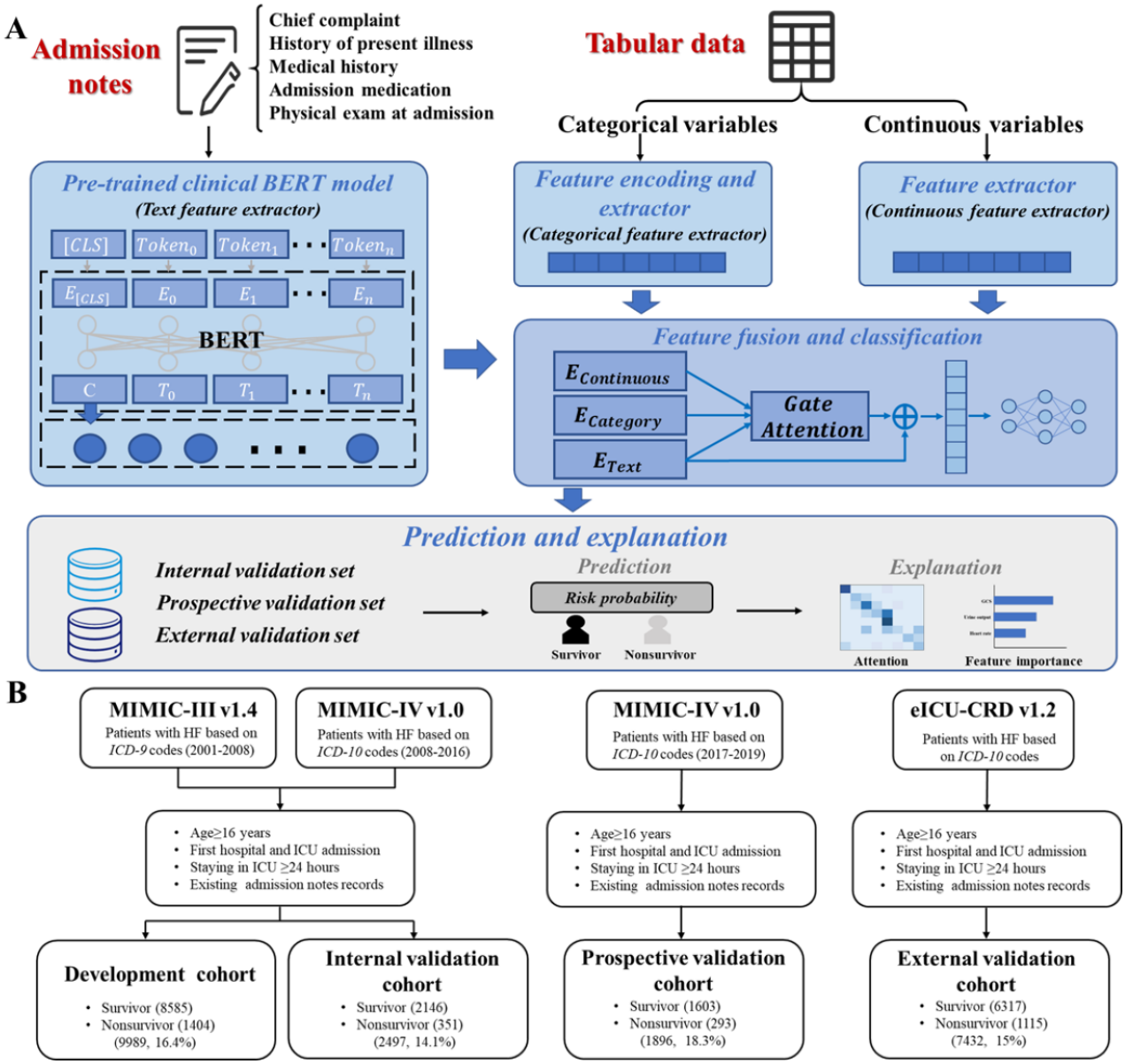

Data Sets and Cohorts

The cohorts for this multicenter retrospective cohort study were derived from 3 open-access clinical databases, including the Medical Information Mart for Intensive Care v1.4 (MIMIC-III; CareVue) and MIMIC-IV v1.0, collected from the Beth Israel Deaconess Medical Center in Boston from 2001 to 2008 and 2008 to 2019, respectively [27,28], and the eICU Collaborative Research Database v1.2 (eICU-CRD), collected from 208 hospitals in United States from 2014 to 2015 [29]. We included all first-time ICU admissions for patients with HF aged ≥16 years according to the International Classification of Diseases diagnostic codes. We excluded patients who stayed in the ICU for less than 24 hours and did not have admission notes. Patients were divided into 4 cohorts to support adequate model evaluation, including the development, internal validation, prospective validation, and external validation cohorts. These cohorts correspond to different stages of model development and evaluation. The development cohort consisted of a subset of data used to create or develop the predictive model. Internal validation, prospective validation, and external validation cohorts helped to check if the model had learned patterns that generalized well to new, unseen data. Each of these cohorts played a crucial role in different stages of model development and validation, ensuring that the predictive model was accurate, reliable, and applicable to new and diverse data sets or situations. The inclusion criteria of them are displayed in Figure 1B.

Data Extraction

Our target is to provide early clinical decision support during ICU admissions. The 5 types of commonly recorded notes were extracted, including chief complaint, history of present illness, medical history, admission medications, and physical exam. In

XSL•FO
RenderX

Figure S1 in Multimedia Appendix 1, an example of an admission note is shown with highlights. Meanwhile, 6 types of clinical variables were collected for model development, as follows: (1) basic information of age, gender, weight, BMI, and Charlson Comorbidity Index; (2) vital signs, such as Glasgow Coma Scale, heart rate, respiratory rate, and systolic blood pressure; (3) laboratory tests, including glucose, creatinine, white blood cell, and total bilirubin; (4) urine output; (5) treatments received, including mechanical ventilation; and (6) physical frailty assessments, including activity and fall risks. Representative statistical features were calculated based on the type of variable, such as the maximum, minimum, and mean values. The median value of each feature was used to impute missing values for continuous variables except for FiO_2 ([fraction of inspired oxygen] with the imputation of 21%), with a missing ratio limitation of less than 30%. Details about all types of candidate variables are provided in Table S1 in Multimedia Appendix 1. Their missing ratio is shown in Table S2 in Multimedia Appendix 1.

Model Development and Output

The model was constructed based on a supervised multimodal DL framework, which mainly included feature extractors and a feature fusion module. A pretrained Bidirectional Encoder Representations from Transformers (BERT) module was used for learning the presentation of clinical notes [30]. In the preliminary experiments, we used all the text chunks (the same subset from the training set for the model) to compare the performances of different pretrained clinical BERT models. We found that clinical BERT [31] demonstrated the best comprehensive performance (Table S3 in Multimedia Appendix 1). In the fusion module, a gate attention mechanism [32] was introduced to aggregate the embedded features of clinical notes and tabular data using attention scores; this module finally output the predicted risk probability of in-hospital deaths through a fully connected layer (Figure S2 in Multimedia Appendix 1). The maximum predicted value of all text chunks for a patient was adopted as the optimal risk prediction score. Further detailed information on model building and training is present in the Multimedia Appendix 1.

Model Explanation

In the pursuit of explicating the underlying mechanisms of the DL model and facilitating a comprehensive visualization of pivotal insights, we embarked on an intricate analysis of the pivotal terminologies instrumental in shaping predictions within the developed model. To achieve this, we used the Integrated Gradients (IG) technique [33] to enhance our comprehension of the BERT model's inner workings and the rationale behind its predictions. This technique hinges on computing gradients with respect to input features, gauging each feature's contribution to the model's prediction. IG offers an intuitive understanding of model predictions by quantifying different features' contributions, aiding clinicians and researchers in comprehending the model's decisions [34,35]. At the same time, IG demonstrates stability across diverse samples and model architectures, yielding consistent explanatory outcomes, crucial in the face of clinical data diversity and complexity [36]. Consequently, it is considered a reliable analytical tool, helping to assess how each word in the input sequence influences the model's predictions for our research. Simultaneously, we harnessed the SHapley Additive exPlanations (SHAP) technique to unravel the importance of clinical variables in structured tabular data. We computed Shapley values to rank the important clinical variables. Shapley values involve a game theory–based approach to explain the prediction of DL models. They measure the contribution of a given feature value to the difference between the actual prediction and the mean prediction. It is important to note that higher SHAP values signify a heightened pertinence of specific terms in influencing the model's predictions, whereas relatively diminished SHAP values connote a less pronounced impact. The IG technique exhibits a similar pattern.

Leveraging the IG and SHAP techniques offers valuable insights into the intricate relationship between input features and prediction outcomes, contributing to a more comprehensive understanding of the model's decision-making process.

Model Validation

The discrimination performance of our prediction model was assessed on the internal (MIMIC, 2001-2016), prospective (MIMIC, 2017-2019) and external (eICU-CRD, 2014-2015) validation cohorts. This assessment compared the model against different single modalities covering notes, tabular data, and a combination of both. The importance of the 5 types of notes for outcome assessment was also examined separately. We trained 5 predictive models based on the tabular data and individual clinical notes. It should be mentioned that the chief complaint was absent in the external validation cohort. Three evaluation metrics were calculated along with their corresponding 95% CIs, the area under the receiver operating characteristic curve (AUROC), F_1-score, and the area under the precision-recall curve.

Statistics Analysis

The median (IQR) values for continuous variables are presented. The *t* test (2-tailed) or the Wilcoxon Rank Sum Test was used when appropriate to compare survivors and nonsurvivors of HF. Categorical variables were reported by total numbers and percentages. Two-sided *P* values of less than .05 were considered statistically significant.

Ethical Considerations

This study was exempt from institutional review board approval due to the retrospective design and lack of direct patient intervention. All data from patients were retrospectively collected from the electronic health care records systems (in the form of third-party public databases or hospital health care systems), which originated from daily clinical work.

All data were de-identified before the analysis. Third-party public databases (MIMIC-IV, MIMIC-III, and eICU-CRD) were used in this study. The institutional review boards of the Massachusetts Institute of Technology (number 0403000206) and Beth Israel Deaconess Medical Center (number 2001-P-001699/14) approved the use of the database for research.

XSL•FO
RenderX

The requirement for individual patient consent was waived because the study did not impact clinical care, all protected health information was deidentified, and all available data in the databases were anonymous.

Results

Patient Characteristics

A total of samples from 12,486 (14.1%) patients with HF were collected from MIMIC-III and MIMIC-IV joint data sets between 2001 and 2016; they were randomly divided into a development set and an independent internal validation set. Additionally, 1896 (18.3%) patients with HF were extracted from MIMIC-IV from 2017 to 2019 for a prospective validation set. For the external validation set, 7432 (15%) patients with HF were extracted from the eICU-CRD data set. Baseline characteristics are summarized in Table 1. The proportion of patients with in-hospital mortality in the 4 cohorts ranges from 14% to 19%. Detailed comparisons of survivors and nonsurvivors in all study cohorts are shown in Tables S4-S7 in Multimedia Appendix 1.

Table 1. The comparison of the total study cohorts for model development and validation.

Characteristics	Development set (n=9989)	Internal validation set (n=2497)	Prospective validation set (n=1896)	External validation set (n=7432)
Basic information				
Age (years), median (IQR)	75 (65-84)	75 (64-83)	74 (64-82)	73 (62-82)
Female sex, n (%)	4637 (46.4)	1151 (46.1)	786 (41.5)	3470 (46.7)
BMI (kg/m^2), median (IQR)	27.7 (24.1-32.6)	27.5 (23.8-32.6)	28.8 (24.4-33.8)	28.9 (24.4-35.0)
CCI[a] score, median (IQR)	7.0 (5.0-8.0)	7.0 (5.0-9.0)	7.0 (6.0-9.0)	5.0 (4.0-6.0)
Physical frailty (fall risk), n (%)	3334 (33.4)	836 (33.5)	1896 (100)	1374 (18.5)
Activity, n (%)				
Bed	7347 (73.8)	1801 (72.6)	1064 (56.2)	3955 (83.5)
Sit	1759 (17.7)	444 (17.9)	441 (23.3)	279 (5.9)
Stand	843 (8.5)	235 (9.5)	388 (20.5)	504 (10.6)
Notes recorded proportion, n (%)				
Chief complaint	8592 (86.0)	2146 (85.9)	1659 (87.5)	0 (0)
History of present illness	9866 (98.8)	2455 (98.3)	1680 (88.6)	7186 (96.7)
Medical history	9730 (97.4)	2427 (97.2)	1675 (88.3)	7188 (96.7)
Admission medication	9018 (90.3)	2257 (90.4)	1650 (87.0)	4387 (59.0)
Physical exam	9186 (92.0)	2303 (92.2)	1644 (86.7)	5104 (68.7)
Outcome				
Days before ICU[b] admission, median (IQR)	0.1 (0.0-1.1)	0.1 (0.0-1.3)	0.1 (0.0-0.9)	0.2 (0.1-0.8)
Days of ICU admission, median (IQR)	3.0 (1.8-5.5)	3.0 (1.8-5.5)	2.9 (1.7-5.3)	2.8 (1.8-4.9)
Days of hospital admission, median (IQR)	8.9 (5.7-14.6)	9.2 (5.8-14.7)	9.8 (6.0-15.7)	7.2 (4.3-11.9)
Death in hospital, n (%)	1404 (16.4)	351 (14.1)	293 (18.3)	1115 (15.0)

[a]CCI: Charlson Comorbidity Index.

[b]ICU: intensive care unit.

Model Performance Evaluation

We present the discrimination performance on internal, prospective, and external validation sets by receiver operating characteristic curves of the optimal models after tuning the hyperparameters (Figure 2). The AUROCs of the multimodal model were significantly higher than the two unimodal models in all 3 types of validation evaluations. They were 0.838 (95% CI 0.827-0.851), 0.849 (95% CI 0.841-0.856), and 0.767 [0.762-0.772] for the internal, prospective, and external validation sets, respectively. Specifically, the design details of the unimodal models were as follows: the text-based unimodal model used clinical BERT, leveraging its capabilities in contextualizing clinical text data; on the other hand, the tabular unimodal model used a fully connected model structure, tailored to effectively process structured tabular data from the tables. More comparisons on baseline models, such as random forest and logistic regression, and all evaluation metrics for these models in the 3 validation types are presented in Table S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. The receiver operating characteristic curve comparison of the optimal multimodal deep learning model against different single modalities in the (A) internal validation set, (B) prospective validation set, and (C) external validation set. Values inside brackets are 95% CIs.

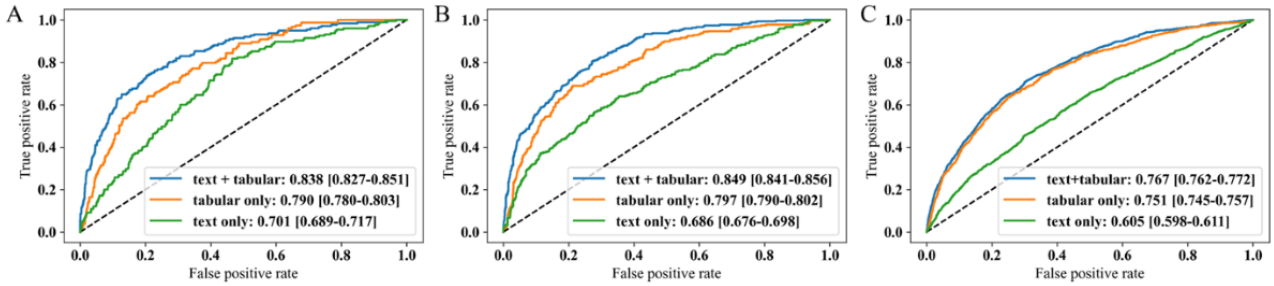

Contribution of Individual Part in Clinical Notes

The performance contributions of the 5 types of clinical notes (including chief complaint, history of present illness, medical history, admission medication, and physical exam) were separately evaluated by combining them with clinical variables to retrain all prediction models. Figure 3 and Table S9 (Multimedia Appendix 1) display the AUROC comparisons with the full model. We found the individual contributions were much lower than the overall contribution in all validation cohorts. Specifically, medical history and physical exam contained more information that was useful in assessing the prognosis of patients with HF compared to other note types.

Figure 3. The reduction of discrimination when separately combining the individual clinical note with tabular data in the (A) internal validation set, (B) prospective validation set, and (C) external validation set. AUC: area under the receiver operating characteristic curve.

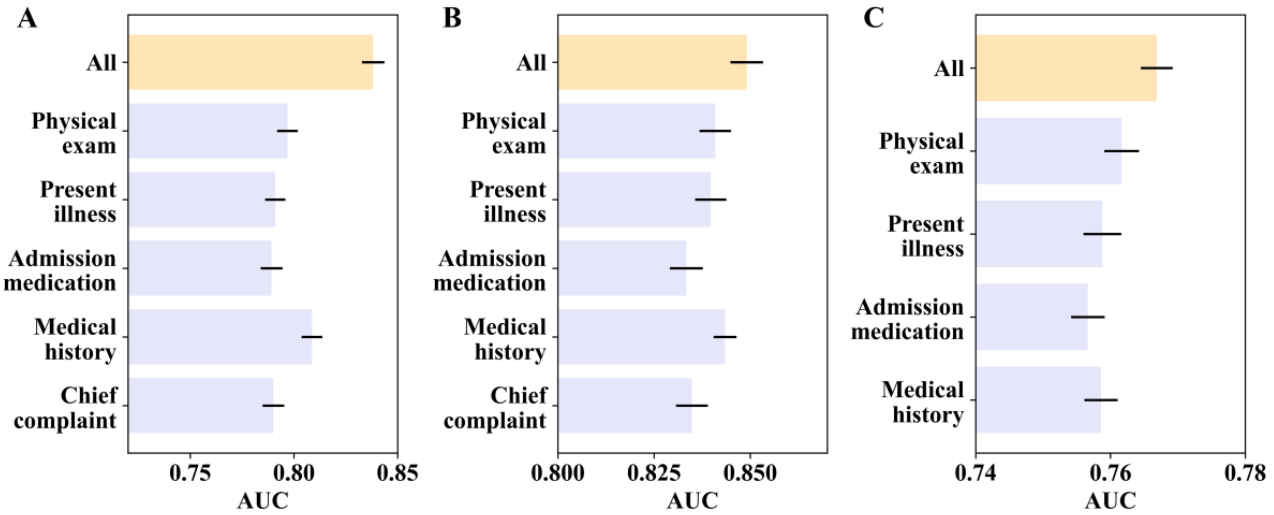

Clinical Notes Visualization and Interpretation

We applied the IG method to study the attribution of the prediction of a deep network to its input features, aiming to provide explanation for individual predictions. IG is computed based on the gradient of the prediction outputs with respect to the input words. Higher IG values denote the greater significance of a word to the model's prediction, whereas lower values indicate lesser importance. We derived IG values for all tokens present in the clinical notes of each patient within the test data set, extracting those tokens with higher IG values. It is important to note that, due to BERT's tokenization process, inputs are represented as tokens rather than individual words. For instance, the phrase "the patient has been extubated" is tokenized into "the patient has been ex ##tub ##ated" as the input sequence [35]. To enhance readability, we conducted postprocessing by excluding numbers, tokens with only 1 or 2 characters, and separators. A clinical expert assessed the clinical significance of tokens and their associated IG values in the context of mortality prediction. The sorted tokens are illustrated in Figure 4A.

The analysis identified commonly ubiquitous clinical terms like "in," "to," and "with," which were segregated due to their limited potential in distinguishing prognostic variations. Among the top 20 clinically meaningful indicators vital for mortality prediction, intriguing insights emerged upon clinical interpretation. For instance, "Failure" and "Pain," the leading predictors, denote prevalent symptoms within ICU care and can mirror disease severity and disability. Indicators 4 and 9 align with pulmonary pathology, their elevated importance reflecting the gravity of respiratory conditions and the necessity for ICU interventions, such as mechanical ventilation. Additional indicators such as "pneumonia" and "fall" manifest acute illness, carrying prognostic weight in mortality prediction. Clinical cues, such as "status," "reflex," and "shock," correspond to mental well-being, with their significance in prognosis attributed to the association of delirium with adverse outcomes.

XSL•FO
RenderX

Figure 4. Explanation of the developed deep learning model. (A) The top 20 tokens of clinical notes evaluated by the Integrated Gradient values in 3 types of language scenarios. (B) The top 20 clinical variables evaluated by the Shapley values. BUN: blood urea nitrogen; CCI: Charlson comorbidity index; GCS: Glasgow Coma Scale; ICU: intensive care unit; PTT: partial thromboplastin time; SBP: systolic blood pressure; SHAP: SHapley Additive exPlanations; SpO_2: oxygen saturation.

A

Rank	Meaningful indicators	Unclear token	Common words
1	failure	##ent	to
2	pain	##tub	in
3	unable	##ema	cell
4	lung	ed	pupils
5	vent	##phe	line
6	transferred	##ated	course
7	reactive	##al	status
8	negative	##yt	commands
9	pulmonary	##yt	multiple
10	poor	##pox	but
11	loss	##f	scan
12	arrest	##ymph	large
13	liver	##hage	flag
14	shock	##ate	lobe
15	emerge	##usion	respiratory
16	fall	##pit	organ
17	response	ne	ray
18	reflex	dem	elderly
19	disease	##hos	cool
20	pneumonia	##dated	on

B

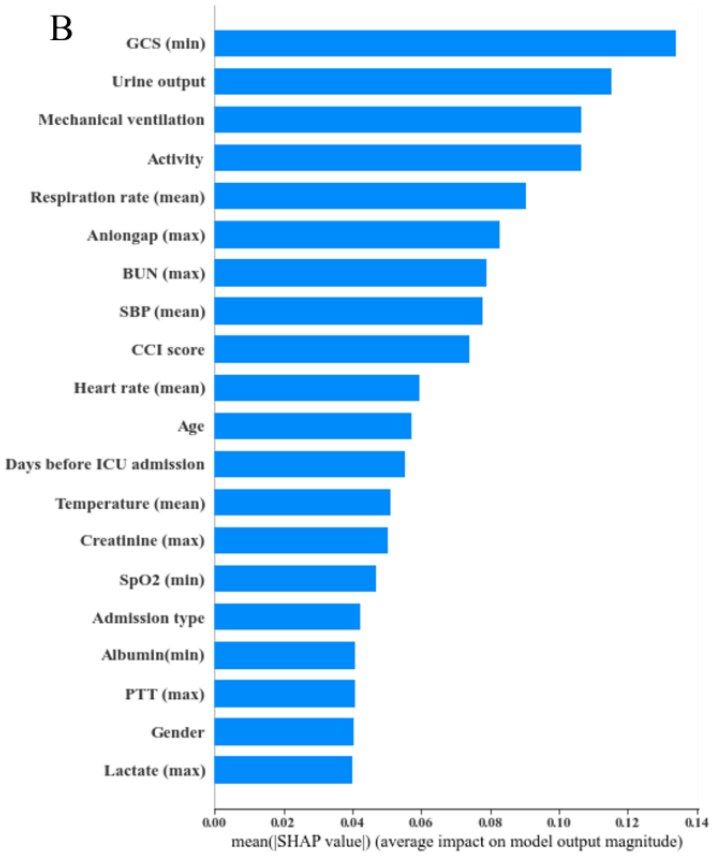

Clinical Variables Feature Analysis

We ranked the important clinical variables using the SHAP technique. The top 20 out of 52 clinical variables (Figure 4) show that for structured tabular data, the highest ranked variables also correlate with disease severity and poorer prognosis. These variables represent clinically important information, such as mental status, using the Glasgow Coma Scale, urine output, mechanical ventilation, activity, and respiratory rate measurements.

Discussion

Principal Findings

This retrospective prognostic study aimed to develop, validate, and explain a multimodal DL prediction model for in-hospital outcomes in critically ill patients with HF. The model was constructed based on the admission notes and records from the first ICU admission day. Simultaneously, we compared the difference between multimodal and unimodal models and explored the individual importance of admission notes in the clinical practice of HF. We found that multimodality could further enhance the model's ability and credibility to evaluate outcomes compared to unimodality.

Emerging clinical data sets provide an opportunity for the DL techniques to study the problem of in-hospital mortality prediction. Compared to previous related work, which mostly considers single modality or simply concatenates embeddings from different modalities, our work demonstrates a novel approach. We separately embed texts as well as categorical and continuous variables to integrate multimodal knowledge and leverage clinical notes information for better predictions. Our comprehensive experiments demonstrate that our proposed model outperforms the models using single modality (text-only AUROC: 0.701; tabular-only AUROC: 0.790) by achieving high performance (AUROC: 0.838).

The fusion method we used integrated two different modalities—unstructured clinical notes and structured clinical variables—into a universal shareable space using a transformer block. It was efficient to leverage clinical notes and integrate tabular data. Meanwhile, the novel application of an attention mechanism on clinical data enhanced the model's ability to focus on evaluating the target in models when fusing multimodal information. Our ablation study, as shown in Figures S3-S5 (Multimedia Appendix 1), on domain adaptive pretraining and task adaptive fine-tuning with multiple BERTs verified the significance of pretraining and fine-tuning, when implementing BERT models on natural language text, especially on domain-specific clinical notes.

In the multimodal model, the proportion of clinical variables, especially continuous variables, was much higher compared to other parts (Figure S6 in Multimedia Appendix 1). The analysis and visualization of important words in clinical notes also yielded interesting findings. The ranking of words by IG values

XSL•FO
RenderX

provided face validity, indicating that some of the important words used for prediction were clinically related to diseases trajectory, such as the severity of respiratory disease or mental status. Some of the unspecified words, such as "disease," used in diverse scenarios, were more difficult to interpret as isolated words. Lastly, some of the clinically meaningful words can change significantly with negation, such as "fall." In the future, we will use more techniques, such as the NegEx algorithm, to consider negation of keywords to better explain the clinical words' meanings.

There are some limitations in our study. First, our model leveraged the electronic health records data based on patients' ICU admission and the first day of admission to predict in-hospital death risk. It did not include recorded data during the treatment, which might reduce the evaluation performance of the model. Second, we simplify the feature extraction, using the maximum, minimum, or mean statistical values to characterize all data throughout the day. Such simplification ignored the changes in time series, and it might have caused the loss of useful information. Time series data will be considered in our future study. Finally, we recommend that the model need to be calibrated using local data to avoid assessment bias.

Conclusions

In this multicenter prognostic study, we developed and validated an attention-multimodal DL model for in-hospital outcome prediction of patients with HF and explored the approaches that can improve the evaluation precision by simultaneously characterizing both admission notes and tabular data. The AUROCs of our model were significantly higher than those of unimodal models in all validation sets. The clinical variables included in the study made a particularly significant contribution to the overall results, with the data from the clinical notes exhibiting a much lower contribution. The model shows good predictive and explainable performance to potentially support the precise decision-making and disease management of critically ill patients with HF.

Acknowledgments

We gratefully acknowledge the guidance and assistance of Dr Max Shen from Beth Israel Deaconess Medical Center (BIDMC).

This work was supported in part by the National Natural Science Foundations of China (NSFC) under grants 62173032 and 62171471, the Foshan Science and Technology Innovation Special Projects (grant BK22BF005), the Regional Joint Fund of the Guangdong Basic and Applied Basic Research Fund (grant 2022A1515140109).

Qing Zhang (qzhang2000cn@163.com), Wendong Xiao (wdxiao@ustb.edu.cn), and Zhengbo Zhang (zhangzhengbo@301hospital.com.cn) are co-corresponding authors of this manuscript.

Authors' Contributions

ZG and XL had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. They also contributed to the conceptualization and design of the study. YK, PH, XZ, and WY contributed to the acquisition, analysis, or interpretation of data. MY and PY conducted statistical analysis. ZG and XL drafted the manuscript. PY, QZ, ZZ, and WX obtained funding for the study and supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional statistics.
[DOCX File , 466 KB - jmir_v26i1e54363_app1.docx]

References

1. Katz AM, Rolett EL. Heart failure: when form fails to follow function. Eur Heart J 2016 Feb 01;37(5):449-454. [doi: 10.1093/eurheartj/ehv548] [Medline: 26497163]
2. Ormerod JO, Ashrafian H, Frenneaux MP. Impaired energetics in heart failure - a new therapeutic target. Pharmacol Ther 2008 Sep;119(3):264-274. [doi: 10.1016/j.pharmthera.2008.05.007] [Medline: 18602947]
3. Pagliaro BR, Cannata F, Stefanini GG, Bolognese L. Myocardial ischemia and coronary disease in heart failure. Heart Fail Rev 2020 Jan 22;25(1):53-65. [doi: 10.1007/s10741-019-09831-z] [Medline: 31332663]
4. Murphy SP, Ibrahim NE, Januzzi JL. Heart failure with reduced ejection fraction: a review. JAMA 2020 Aug 04;324(5):488-504. [doi: 10.1001/jama.2020.10262] [Medline: 32749493]
5. Savarese G, Becher P, Lund L, Seferovic P, Rosano G, Coats A. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res 2023 Jan 18;118(17):3272-3287. [doi: 10.1093/cvr/cvac013] [Medline: 35150240]
6. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. Nat Rev Cardiol 2016 Dec;13(6):368-378 [FREE Full text] [doi: 10.1038/nrcardio.2016.25] [Medline: 26935038]

XSL•FO
RenderX

7. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. BMJ Open 2021 Jul 23;11(7):e044779 [FREE Full text] [doi: 10.1136/bmjopen-2020-044779] [Medline: 34301649]
8. Safavi KC, Dharmarajan K, Kim N, Strait KM, Li S, Chen SI, et al. Variation exists in rates of admission to intensive care units for heart failure patients across hospitals in the United States. Circulation 2013 Feb 26;127(8):923-929. [doi: 10.1161/circulationaha.112.001088]
9. van Diepen S, Bakal JA, Lin M, Kaul P, McAlister FA, Ezekowitz JA. Variation in critical care unit admission rates and outcomes for patients with acute coronary syndromes or heart failure among high - and low - volume cardiac hospitals. JAHA 2015 Mar 10;4(3):e001708. [doi: 10.1161/jaha.114.001708]
10. Wunsch H, Angus DC, Harrison DA, Collange O, Fowler R, Hoste EAJ, et al. Variation in critical care services across North America and Western Europe*. Critical Care Medicine 2008;36(10):2787-27e8. [doi: 10.1097/ccm.0b013e318186aec8]
11. Adams KF, Fonarow GC, Emerman CL, LeJemtel TH, Costanzo MR, Abraham WT, ADHERE Scientific Advisory CommitteeInvestigators. Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE). Am Heart J 2005 Feb;149(2):209-216. [doi: 10.1016/j.ahj.2004.08.005] [Medline: 15846257]
12. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. Eur J Heart Fail 2020 Jan;22(1):139-147 [FREE Full text] [doi: 10.1002/ejhf.1628] [Medline: 31721391]
13. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC Heart Fail 2020 Jan;8(1):12-21 [FREE Full text] [doi: 10.1016/j.jchf.2019.06.013] [Medline: 31606361]
14. Soliman A, Agvall B, Etminani K, Hamed O, Lingman M. The price of explainability in machine learning models for 100-day readmission prediction in heart failure: retrospective, comparative, machine learning study. JMIR 2023 Oct 27;25:e46934 [FREE Full text] [doi: 10.2196/46934] [Medline: 37889530]
15. White-Williams C, Rossi LP, Bittner VA, Driscoll A, Durant RW, Granger BB, American Heart Association Council on CardiovascularStroke Nursing; Council on Clinical Cardiology;Council on EpidemiologyPrevention. Addressing social determinants of health in the care of patients with heart failure: a scientific statement from the American heart association. Circulation 2020 Jun 02;141(22):e841-e863 [FREE Full text] [doi: 10.1161/CIR.0000000000000767] [Medline: 32349541]
16. Abraham WT, Fonarow GC, Albert NM, Stough WG, Gheorghiade M, Greenberg BH, OPTIMIZE-HF InvestigatorsCoordinators. Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). J Am Coll Cardiol 2008 Jul 29;52(5):347-356 [FREE Full text] [doi: 10.1016/j.jacc.2008.04.028] [Medline: 18652942]
17. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC Heart Fail 2020 Jan;8(1):12-21 [FREE Full text] [doi: 10.1016/j.jchf.2019.06.013] [Medline: 31606361]
18. McGilvray MM, Heaton J, Guo A, Masood MF, Cupps BP, Damiano M, et al. Electronic health record-based deep learning prediction of death or severe decompensation in heart failure patients. JACC Heart Fail 2022 Sep;10(9):637-647 [FREE Full text] [doi: 10.1016/j.jchf.2022.05.010] [Medline: 36049815]
19. Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. JMIR 2022 Aug 09;24(8):e38082 [FREE Full text] [doi: 10.2196/38082] [Medline: 35943767]
20. Kaya SI, Cetinkaya A, Ozcelikay G, Samanci SN, Ozkan SA. Approaches and challenges for biosensors for acute and chronic heart failure. Biosensors (Basel) 2023 Feb 16;13(2):282 [FREE Full text] [doi: 10.3390/bios13020282] [Medline: 36832048]
21. Hudson SR, Chan D, Ng LL. Change in plasma volume and prognosis in acute decompensated heart failure: an observational cohort study. J R Soc Med 2016 Sep 08;109(9):337-346 [FREE Full text] [doi: 10.1177/0141076816661316] [Medline: 27609799]
22. Goh KH, Wang L, Yeow AYK, Ding YY, Au LSY, Poh HMN, et al. Prediction of readmission in geriatric patients from clinical notes: retrospective text mining study. J Med Internet Res 2021 Oct 19;23(10):e26486 [FREE Full text] [doi: 10.2196/26486] [Medline: 34665149]
23. Clapp MA, Kim E, James KE, Perlis RH, Kaimal AJ, McCoy TH, et al. Comparison of natural language processing of clinical notes with a validated risk-stratification tool to predict severe maternal morbidity. JAMA Netw Open 2022 Oct 03;5(10):e2234924 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.34924] [Medline: 36197662]
24. Özyılmaz E, Özkan Kuşçu Ö, Karakoç E, Boz A, Orhan Tıraşçı G, Güzel R, et al. Worse pre-admission quality of life is a strong predictor of mortality in critically ill patients. Turk J Phys Med Rehabil 2022 Mar 01;68(1):19-29 [FREE Full text] [doi: 10.5606/tftrd.2022.5287] [Medline: 35949964]
25. Thapa NB, Nischay B, Sattar S, Sona T. Hospital readmission prediction using clinical admission notes. 2022 Presented at: ACSW '22: Proceedings of the 2022 Australasian Computer Science Week; February 14 - 18; Brisbane, Australia. [doi: 10.1145/3511616.3513115]

XSL•FO
RenderX

26. van Aken B, Papaioannou JM, Mayrdorfer M, Budde K, Gers F, Loeser A. Clinical outcome prediction from admission notes using self-supervised knowledge integration. arXiv Preprint posted online on Feb 8, 2021. [doi: 10.48550/arXiv.2102.04110]
27. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
28. Johnson A, Stone D, Celi L, Pollard T. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 2018 Jan 01;25(1):32-39. [doi: 10.1093/jamia/ocx084] [Medline: 29036464]
29. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5(1):180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
31. Huang K, Altossar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
32. Rahman W, Hasan MK, Lee S, Bagher Zadeh AA, Mao C, Morency LP, et al. Integrating multimodal information in large pretrained transformers. Proc Conf Assoc Comput Linguist Meet 2020 Jul;2020:2359-2369 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.214] [Medline: 33782629]
33. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017 Presented at: Proceedings of the 34th International Conference on Machine Learning; August 6 - 11; Sydney, Australia p. 3319-3328. [doi: 10.5555/3305890.3306024]
34. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet 2023 Feb 03;24(2):125-137. [doi: 10.1038/s41576-022-00532-2] [Medline: 36192604]
35. Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, et al. A multimodal transformer: fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. AMIA Annu Symp Proc 2022;2022:719-728 [FREE Full text] [Medline: 37128451]
36. Zhang Y, Tino P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell 2021 Oct;5(5):726-742. [doi: 10.1109/tetci.2021.3100641]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
DL: deep learning
eICU-CRD: eICU Collaborative Research Database
HF: heart failure
ICU: intensive care unit
IG: Integrated Gradients
MIMIC: Medical Information Mart for Intensive Care
NT-proBNP: N-terminal pro–b-type natriuretic peptide
SHAP: SHapley Additive exPlanations

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 07.11.23; peer-reviewed by E Kawamoto, MO Khursheed; comments to author 05.12.23; revised version received 01.01.24; accepted 19.03.24; published 02.05.24.

Please cite as:
Gao Z, Liu X, Kang Y, Hu P, Zhang X, Yan W, Yan M, Yu P, Zhang Q, Xiao W, Zhang Z
Improving the Prognostic Evaluation Precision of Hospital Outcomes for Heart Failure Using Admission Notes and Clinical Tabular Data: Multimodal Deep Learning Model
J Med Internet Res 2024;26:e54363
URL: https://www.jmir.org/2024/1/e54363
doi:10.2196/54363
PMID:38696251

©Zhenyue Gao, Xiaoli Liu, Yu Kang, Pan Hu, Xiu Zhang, Wei Yan, Muyang Yan, Pengming Yu, Qing Zhang, Wendong Xiao, Zhengbo Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of an Explainable Deep Learning Model to Predict In-Hospital Mortality for Patients With Acute Myocardial Infarction: Algorithm Development and Validation Study

Puguang Xie[1], PhD; Hao Wang[1], MD; Jun Xiao[1], MD; Fan Xu[1], MS; Jingyang Liu[1], MS; Zihang Chen[2], MS; Weijie Zhao[2], MS; Siyu Hou[3], MS; Dongdong Wu[4], PhD; Yu Ma[1*], MD; Jingjing Xiao[3*], PhD

[1]Chongqing Emergency Medical Centre, Chongqing University Central Hospital, School of Medicine, Chongqing University, Chongqing, China
[2]Bioengineering College, Chongqing University, Chongqing, China
[3]Bio-Med Informatics Research Centre & Clinical Research Centre, Xinqiao Hospital, Army Medical University, Chongqing, China
[4]Medical Big Data Research Centre, Chinese People's Liberation Army General Hospital, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Jingjing Xiao, PhD
Bio-Med Informatics Research Centre & Clinical Research Centre, Xinqiao Hospital, Army Medical University
No. 183 Xinqiao Street, Shapingba District
Chongqing, 400037
China
Phone: 86 18502299862
Email: shine636363@sina.com

Abstract

Background: Acute myocardial infarction (AMI) is one of the most severe cardiovascular diseases and is associated with a high risk of in-hospital mortality. However, the current deep learning models for in-hospital mortality prediction lack interpretability.

Objective: This study aims to establish an explainable deep learning model to provide individualized in-hospital mortality prediction and risk factor assessment for patients with AMI.

Methods: In this retrospective multicenter study, we used data for consecutive patients hospitalized with AMI from the Chongqing University Central Hospital between July 2016 and December 2022 and the Electronic Intensive Care Unit Collaborative Research Database. These patients were randomly divided into training (7668/10,955, 70%) and internal test (3287/10,955, 30%) data sets. In addition, data of patients with AMI from the Medical Information Mart for Intensive Care database were used for external validation. Deep learning models were used to predict in-hospital mortality in patients with AMI, and they were compared with linear and tree-based models. The Shapley Additive Explanations method was used to explain the model with the highest area under the receiver operating characteristic curve in both the internal test and external validation data sets to quantify and visualize the features that drive predictions.

Results: A total of 10,955 patients with AMI who were admitted to Chongqing University Central Hospital or included in the Electronic Intensive Care Unit Collaborative Research Database were randomly divided into a training data set of 7668 (70%) patients and an internal test data set of 3287 (30%) patients. A total of 9355 patients from the Medical Information Mart for Intensive Care database were included for independent external validation. In-hospital mortality occurred in 8.74% (670/7668), 8.73% (287/3287), and 9.12% (853/9355) of the patients in the training, internal test, and external validation cohorts, respectively. The Self-Attention and Intersample Attention Transformer model performed best in both the internal test data set and the external validation data set among the 9 prediction models, with the highest area under the receiver operating characteristic curve of 0.86 (95% CI 0.84-0.88) and 0.85 (95% CI 0.84-0.87), respectively. Older age, high heart rate, and low body temperature were the 3 most important predictors of increased mortality, according to the explanations of the Self-Attention and Intersample Attention Transformer model.

Conclusions: The explainable deep learning model that we developed could provide estimates of mortality and visual contribution of the features to the prediction for a patient with AMI. The explanations suggested that older age, unstable vital signs, and metabolic disorders may increase the risk of mortality in patients with AMI.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e49848) doi:10.2196/49848

KEYWORDS

acute myocardial infarction; mortality; deep learning; explainable model; prediction

Introduction

Background

Acute myocardial infarction (AMI) refers to myocardial ischemia caused by acute blockage of coronary arteries, which is one of the most common cardiovascular emergencies [1]. AMI remains a leading cause of mortality worldwide, and patients with AMI have an estimated in-hospital mortality of 7% to 14% [2,3]. Therefore, the prediction of in-hospital mortality and assessment of risk factors for patients with AMI is of great interest to clinicians. Predicting the risk of mortality and conducting an individualized analysis of the risk factors in patients with AMI at initial diagnosis is critical for developing targeted interventions and improving the prognosis [4].

A substantial effort has been devoted to developing models for predicting mortality in patients with AMI [5]. In the past, generalized linear models (eg, the logistic regression [LR] model and Cox proportional hazard model) have been used to predict mortality in patients with AMI [6-9]. However, the generalized linear models fail to capture the nonlinear relations of the massive, high-dimensional, and incomplete medical data, which hinder the practical use and further clinical application of the models [10]. With the rapid development of artificial intelligence, tree-based machine learning models, such as random forest (RF) [11] and extreme gradient boosting (XGBoost) [12], were used to address problems related to the prediction of mortality and achieved promising performance [13-15]. Nevertheless, tree-based models tend to be ineffective in the face of web-based learning and sparse classification features. More importantly, tree-based models failed to develop the model with powerful deep learning modules through the backpropagation method, which limits the further improvement of the model performance [16]. Deep learning is the most advanced branch of artificial intelligence, which can automatically learn feature representations from training data and provide new prospects for solving the existing limitations in tree-based models [17]. However, state-of-the-art deep learning models specific to tabular data have not been applied to the prediction of mortality in patients with AMI. In addition, while a complex model may perform better than a simpler one, it could lead to black-box problems, meaning that only the inputs and outputs of the model can be seen and it is difficult to understand how the inputs affect the predictions, which limit their clinical acceptance and raise ethical and legal questions [18]. The proposal of methods that could provide explanations for black-box models might facilitate clinicians in understanding predictions and making faster and more accurate treatment decisions [9,19,20]. At present, there is a lack of tools to predict in-hospital mortality of patients with AMI based on tabular data–specific deep learning models, and the potential of state-of-the-art deep learning models for predicting mortality in AMI is unclear. In addition, most mortality prediction models for patients with AMI proposed in previous studies are black-box models, which makes it difficult for clinicians to understand the predictions and use the models in clinical practice.

Contribution

In this study, we chose 3 state-of-the-art deep learning models to predict in-hospital mortality of patients with AMI and compared their performance with linear and tree-based models. In addition, we used an algorithm to interpret the best-performing black-box prediction model to obtain a tool for individualized mortality prediction and risk factor assessment in patients with AMI.

Methods

Data Source

This retrospective study included consecutive patients hospitalized with AMI from Chongqing University Central Hospital between July 2016 and December 2022, the Medical Information Mart for Intensive Care-IV (MIMIC-IV; version 2.2) database [21], and the Electronic Intensive Care Unit Collaborative Research Database (eICU-CRD; version 2.0) [22]. Chongqing University Central Hospital is a city-designated emergency tertiary medical center in China with a permanent population of >32 million. The MIMIC-IV database and eICU-CRD are 2 large free public databases. The MIMIC-IV database contains case information from 299,712 patients admitted to the intensive care unit or the emergency department at the Beth Israel Deaconess Medical Center from 2008 to 2019. The eICU-CRD included information from 139,367 patients from 208 hospitals in the United States between 2014 and 2015.

Ethical Considerations

Access to the MIMIC-IV database and eICU-CRD was approved by the Massachusetts Institute of Technology (Cambridge, Massachusetts) and Beth Israel Deaconess Medical Center (Boston, Massachusetts), and consent was obtained for the data collection. These databases are public, all patient data are anonymized, and data extracted from the databases do not require individual informed consent. PX attended a series of courses offered by the National Institutes of Health and was granted access to these databases after passing the required assessment (record ID 51524821). The study was conducted following the Declaration of Helsinki. Patient data obtained from the Chongqing University Central Hospital were also anonymized. Ethics approval was obtained from the Chongqing University Central Hospital Ethics Committee, and informed consent was waived.

Population

All patients diagnosed with AMI according to the *ICD* (*International Classification of Diseases*) in the Chongqing University Central Hospital between July 2016 and December 2022, MIMIC-IV database and eICU-CRD were included. *ICD-9* (*International Classification of Diseases, Ninth Revision*) code

XSL•FO
RenderX

410 and *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) code I21 were used to identify the patients with AMI. For patients with AMI admitted multiple times during the study period, only the first admission was included. Exclusion criteria were (1) patients aged <18 years, (2) patients with unclear clinical outcome, and (3) patients with >20% of missing data (Figure 1).

Figure 1. Study profile: (A) Inclusion and exclusion criteria for the training and internal test data set; and (B) Inclusion and exclusion criteria for the external validation data set. AMI: acute myocardial infarction; eICU-CRD: Electronic Intensive Care Unit Collaborative Research Database; MIMIC-IV: Medical Information Mart for Intensive Care-IV.

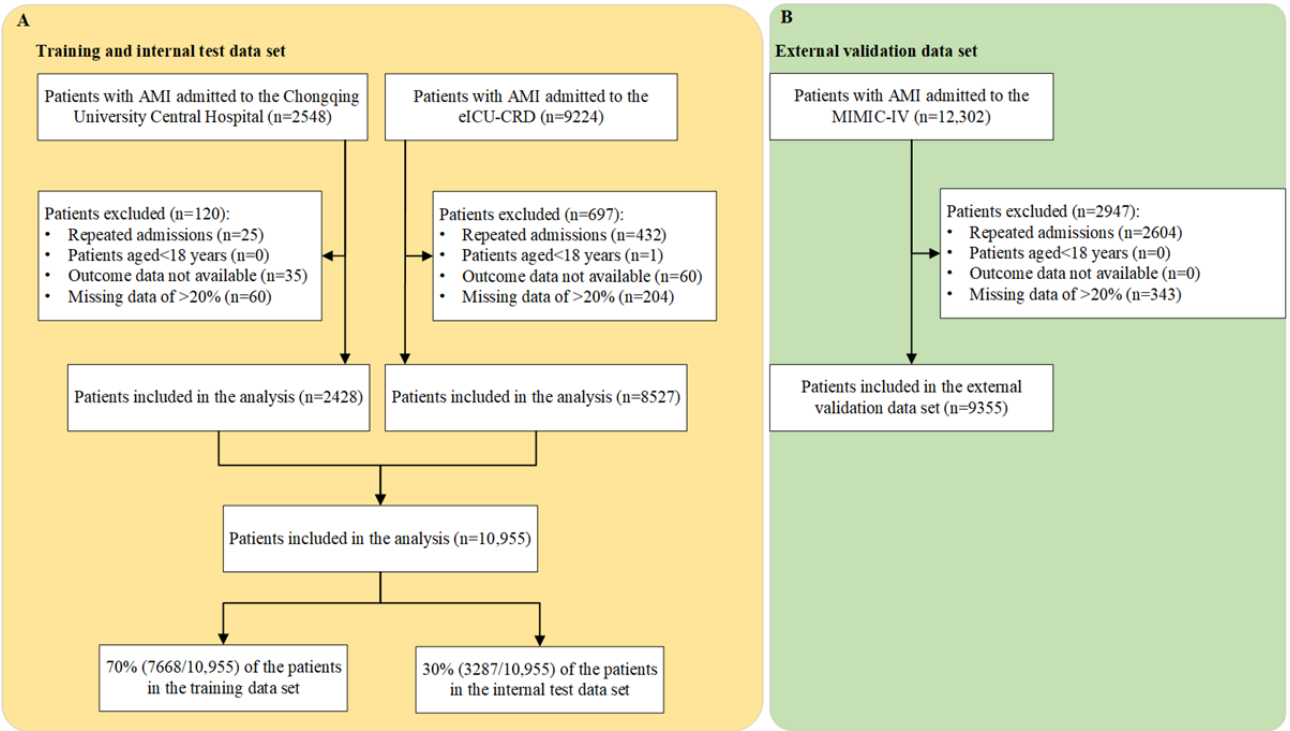

Data Extraction

Patient data at Chongqing University Central Hospital were collected by reviewing electronic medical records. The data in the MIMIC-IV database and eICU-CRD were obtained using SQL. A total of 39 variables were incorporated into this study to determine the prediction models, including demographics, vital signs, laboratory parameters, comorbidities and drug use data based on previous studies, and clinical relevance and experience. All vital signs and laboratory data were the first measurements after admission. Only variables with <30% of missing values across the entire data set were considered for inclusion. This inclusion criterion, unlike the exclusion criterion for patients, was applicable to variables in the entire database. The predetermined threshold was selected according to the evidence from previous studies and experience [23,24]. The missing values were filled by the K-nearest neighbors (KNN) algorithm, and the hyperparameter K for KNN was determined using a 5-fold cross-validation in the training data set. The KNN model trained in the training data set was then used to fill in the missing values in the internal test and external validation data sets. In total, 24,004 missing data points were successfully filled by the KNN model. The clinical outcome assessed in this study was all-cause in-hospital mortality.

Statistical Analysis

Patients with AMI in the training data set were divided into a survival group and a nonsurvival group according to the clinical outcome. Shapiro-Wilk test was applied for the normality test. Continuous variables were presented as medians (IQRs), and categorical variables were presented as numbers and percentages. To test the differences within the groups, the 2-tailed *t* test and Kruskal-Wallis test were used for normal and nonnormal continuous variables, respectively. For categorical variables, between-group differences were evaluated using the chi-square test. *P* value <.05 was regarded as statistically significant.

Model Development

The data of the Chongqing University Central Hospital and eICU-CRD were merged and shuffled. The shuffled data set was then randomly divided into the training data set and internal test data set in a 7:3 ratio. The reason for this combination was that the data of the eICU-CRD were collected from 208 hospitals in the United States, and the generalization ability of the model may be improved by combining the data of eICU-CRD with the data from Chongqing University Central Hospital into a larger multicenter training data set. Patient data in the MIMIC-IV database were used for an independent external validation, which is independent of the Chongqing University Central Hospital and eICU-CRD data.

Data preprocessing tasks were performed before training the prediction models. The KNN algorithm and *z* score were used to fill in missing values and normalize continuous variables, respectively.

XSL•FO
RenderX

A total of 9 models were used to predict in-hospital mortality in patients with AMI, including state-of-the-art deep learning models (eg, Self-Attention and Intersample Attention Transformer [SAINT] [25], TabTransformer [26], and TabNet [27]), a basic deep learning model (eg, multilayer perceptron [MLP] [28]), tree-based machine learning models (eg, RF [11], XGBoost [12], light gradient boosting machine [29], and categorical boosting [30]), and a linear model (LR [31]). An optimization procedure based on nested 5-fold cross-validation Bayesian optimization was used for each model with 100 iterations to determine the hyperparameters. Additional details on hyperparameters setting are provided in Tables S1-S8 in Multimedia Appendix 1. The complete training data set was then used to train the final models with the optimal hyperparameters. Code for model development is publicly available [32].

Model Evaluation

The area under the receiver operating characteristic curve (AUC) was used to assess the performance of the prediction models in the internal test data set and external validation data set. The cutoff thresholds were determined according to the Youden index, and 5 metrics were further calculated, including balanced accuracy, sensitivity, specificity, precision, and F_1-score. CIs were calculated using 1000 bootstrap samples. The calibration curve and Brier score were used to evaluate the reliability of probability estimates of the prediction models; decision curve analysis was used to evaluate the clinical utility of the models.

Model Explanation

The Shapley Additive Explanations (SHAP) algorithm was used to interpret the predictions to obtain global and local influences of the relevant features on the predictions. With a set of feature values of a patient, a SHAP value represents how much a single feature's value influences the difference between the actual prediction and the average prediction of the model in interaction with other features. Briefly, the mean prediction of the model plus the sum of the SHAP values of a sample is equal to the prediction of that sample [33,34].

We first used the SHAP algorithm to visualize the global relative importance of features to the predictions. The rationality of the abovementioned results was then verified by statistical methods. Patients in the training data set were further divided into a survival group and a nonsurvival group. Statistical test was used to evaluate the differences in features between the groups, and the test results were compared with those obtained by the SHAP algorithm. Finally, we linearly mapped the SHAP value to the probability of driving an increase or decrease in predicted mortality, quantifying and visualizing the contribution of each feature to the prediction in individuals.

Results

Statistical Analysis Results

A total of 2428 consecutive patients with AMI who were admitted to Chongqing University Central Hospital between July 2016 and December 2022 and 8527 patients with AMI from the eICU-CRD were included in the analysis. A total of 10,955 patients were randomly divided into a training data set of 7668 (70%) patients and an internal test data set of 3287 (30%) patients. The median age of patients in the training and internal test data sets was 66 (IQR 56-76) years, with 65.6% (5030/7668) and 68.09% (2238/3287) of male participants, respectively. In total, 8.74% (670/7668) of the patients each in the training data set and 8.73% (287/3287) of the patients in the internal test data set died during their hospital stay (Table 1).

A total of 9355 patients with AMI from the MIMIC-IV database were included for independent external validation. The median age of patients in the external validation group was 72 (IQR 62-82) years, higher than that in the training or internal test data set. In total, 61.17% (5722/9355) of the patients were male participants, and 9.12% (853/9355) of the patients died during the hospitalization (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics of patients with acute myocardial infarction included in the models (N=10,955).

Variables	Training data set (n=7668)	Internal test data set (n=3287)	External validation data set (n=9355)
Demographic data			
Age (years), median (IQR)	66 (56-76)	66 (56-76)	72 (62-82)
Female participants, n (%)	2638 (34.4)	1049 (31.9)	3633 (38.8)
Vital signs, median (IQR)			
Heart rate (beats/minute)	81 (70-96)	81 (70-95)	83 (72-96)
Respiratory rate (minute^{-1})	20 (17-21)	20 (17-21)	18 (16-20)
Systolic blood pressure (mm Hg)	129 (111-149)	129 (112-148)	126 (109-145)
Diastolic blood pressure (mm Hg)	76 (64-88)	76 (64-88)	69 (58-82)
Temperature (°C)	36.6 (36.4-36.8)	36.6 (36.4-36.8)	36.7 (36.4-37)
Laboratory data, median (IQR)			
Troponin T or I (ng/mL)	0.74 (0.10-4.85)	0.69 (0.10-4.93)	0.32 (0.11-1.01)
White blood cell (k/μ)	10.6 (8.2-13.6)	10.7 (8.3-13.7)	9.30 (7.1-12.4)
Red blood cell (m/μ)	4.42 (3.89-4.87)	4.43 (3.91-4.87)	3.88 (3.34-4.41)
Hemoglobin (g/dL)	13.5 (11.7-14.8)	13.4 (11.8-14.9)	11.6 (9.9-13.2)
Hematocrit (%)	40.1 (35.4-43.9)	40.2 (35.7-44)	35.2 (30.5-39.6)
Platelet (k/μ)	215 (173-265)	217 (174-268)	208 (164-262)
Prothrombin time (seconds)	13.3 (12.3-14.5)	13.3 (12.3-14.5)	12.9 (11.8-14.7)
International normalized ratio	1.06 (1-1.2)	1.05 (1-1.2)	1.2 (1.1-1.3)
Sodium (mEq/L)	138 (136-140)	138 (136-140)	139 (137-141)
Potassium (mEq/L)	4.00 (3.70-4.37)	4 (3.7-4.4)	4.2 (3.8-4.5)
Chloride (mEq/L)	103.3 (100-106)	103.4 (100-106)	103 (100-106)
Calcium (mg/dL)	8.90 (8.5-9.3)	8.9 (8.41-9.3)	8.7 (8.3-9.1)
Magnesium (mg/dL)	1.99 (1.8-2.2)	1.99 (1.77-2.19)	2 (1.8-2.2)
Creatinine (mg/dL)	1.04 (0.84-1.38)	1.03 (0.82-1.40)	1.1 (0.8-1.6)
Blood urea nitrogen (mg/dL)	19 (14- 28)	19 (14-28)	21 (15-35)
Comorbidities, n (%)			
Hypertension	4616 (60.2)	1965 (59.8)	3962 (42.4)
Hyperlipidemia	820 (10.7)	397 (12.1)	4940 (52.8)
Diabetes	2516 (32.8)	1087 (33.1)	3813 (40.8)
Heart failure	1987 (25.9)	854 (26)	4298 (45.9)
Renal failure	1648 (21.5)	760 (23.1)	4446 (47.5)
Cerebral vascular disease	788 (10.3)	348 (10.6)	1106 (11.8)
Peripheral vascular disease	1062 (13.8)	439 (13.4)	1324 (14.2)
Chronic obstructive pulmonary disease	864 (11.3)	386 (11.7)	688 (7.4)
Operations, n (%)			
Coronary arteriography	3855 (50.3)	1663 (50.6)	2839 (30)
Percutaneous coronary intervention	3083 (40.2)	1359 (41.3)	2954 (31.6)
Drugs, n (%)			
Aspirin	4941 (64.4)	2126 (64.7)	8534 (91.2)
Clopidogrel	2070 (27)	872 (26.5)	4385 (46.9)
Ticagrelor	1491 (19.4)	629 (19.1)	622 (6.6)
Angiotensin-converting enzyme inhibitor	1827 (23.8)	765 (23.3)	4001 (42.8)

XSL•FO
RenderX

Variables	Training data set (n=7668)	Internal test data set (n=3287)	External validation data set (n=9355)
Angiotensin receptor blockers	240 (3.1)	118 (3.6)	1158 (12.4)
Statins	3728 (48.6)	1596 (48.6)	8276 (88.5)
β-Blockers	3565 (46.5)	1555 (47.3)	7625 (81.5)
Clinical outcome			
All-cause in-hospital mortality, n (%)	670 (8.7)	287 (8.7)	853 (9.1)
Length of hospital stay (days), median (IQR)	4.98 (2.50-9.25)	4.95 (2.42-9.51)	5.33 (2.80-9.85)

Model Performance

The AUCs of 9 models in the internal test data set and external validation data set are shown in Figures 2A and 2B (AUC=1 indicates perfect prediction and AUC=0.5 indicates random prediction), and the other 5 evaluation metrics, including balanced accuracy, sensitivity, specificity, precision, and F_1-score are listed in Tables 2 and 3.

Figure 2. Performance of the models in predicting in-hospital mortality of patients with acute myocardial infarction in the internal test set and external validation set: (A) and (B), AUCs of the models; (C) and (D), calibration plots of the models; and (E) and (F), decision curves of the models. AUC: area under the receiver operating characteristic curve; CatBoost: categorical boosting; LightGBM: light gradient boosting machine; LR: logistic regression; MLP: multilayer perceptron; RF: random forest; SAINT: Self-Attention and Intersample Attention Transformer; XGBoost: extreme gradient boosting.

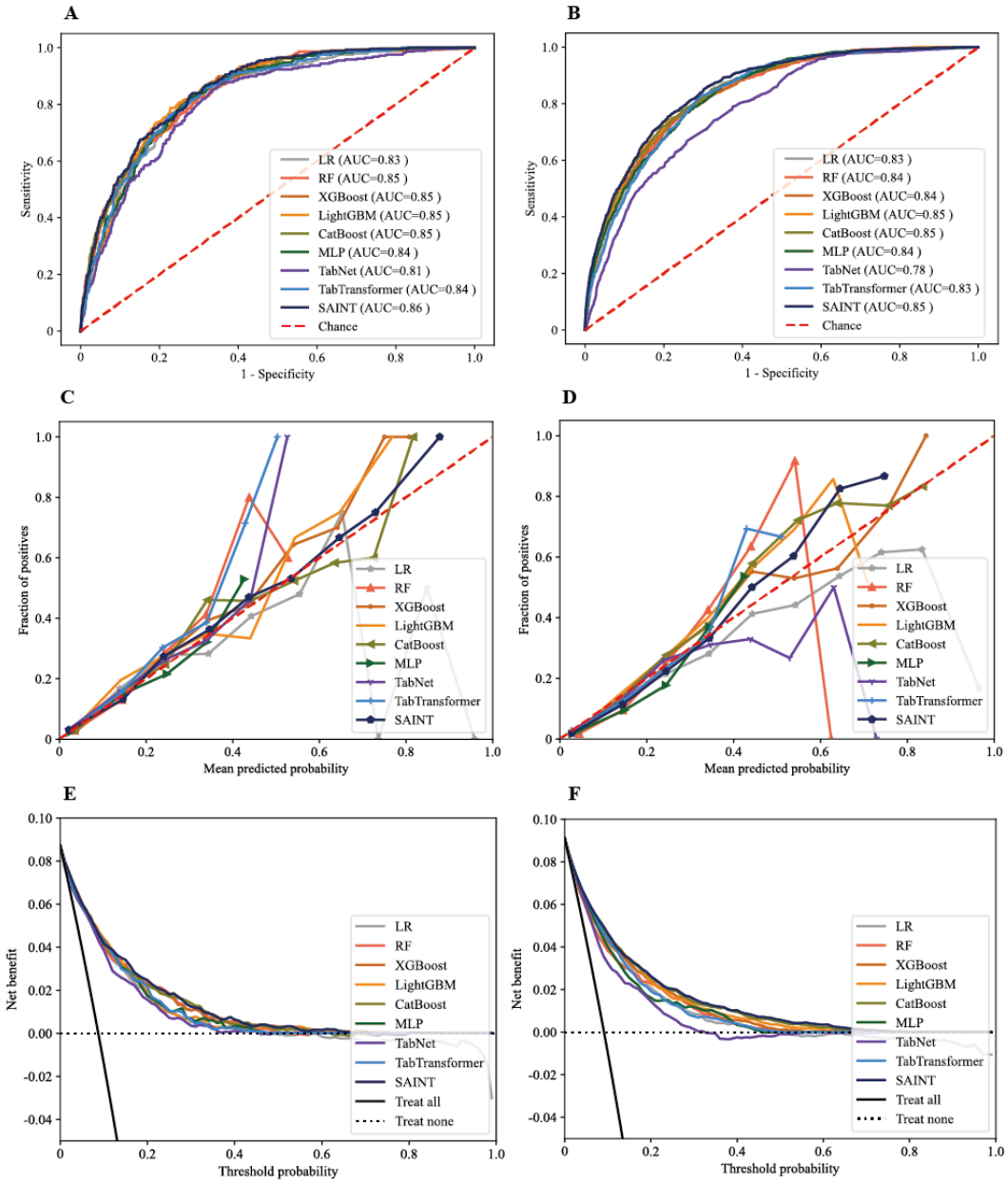

XSL•FO
RenderX

Table 2. Values of the evaluation metrics of the models in the internal test data set.

Models	Balanced accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F_1-score (95% CI)	AUC[a] (95% CI)
Logistic regression	0.76 (0.74-0.78)	0.85 (0.81-0.89)	0.67 (0.65-0.68)	0.20 (0.17-0.22)	0.32 (0.29-0.35)	0.83 (0.81-0.85)
Random forest	0.72 (0.69-0.75)	0.58 (0.52-0.64)	0.87 (0.86-0.88)	0.30 (0.26-0.34)	0.39 (0.35-0.43)	0.85 (0.82-0.87)
Extreme gradient boosting	0.75 (0.73-0.78)	0.74 (0.69-0.78)	0.77 (0.75-0.78)	0.23 (0.21-0.26)	0.35 (0.32-0.39)	0.85 (0.83-0.87)
Light gradient boosting machine	0.71 (0.68-0.73)	0.52 (0.47-0.58)	0.89 (0.88-0.91)	0.32 (0.28-0.36)	0.40 (0.35-0.44)	0.85 (0.83-0.87)
Categorical boosting	0.76 (0.73-0.78)	0.74 (0.69-0.79)	0.78 (0.76-0.80)	0.24 (0.21-0.27)	0.37 (0.33-0.40)	0.85 (0.83-0.87)
Multilayer perceptron	0.77 (0.74-0.79)	0.79 (0.75-0.84)	0.74 (0.73-0.76)	0.23 (0.20-0.25)	0.32 (0.35-0.39)	0.84 (0.82-0.86)
TabNet	0.76 (0.73-0.78)	0.86 (0.82-0.90)	0.65 (0.64-0.67)	0.19 (0.17-0.21)	0.31 (0.28-0.34)	0.81 (0.79-0.84)
TabTransformer	0.77 (0.75-0.79)	0.85 (0.81-0.89)	0.69 (0.67-0.71)	0.21 (0.18-0.23)	0.33 (0.30-0.37)	0.84 (0.82-0.86)
Self-Attention and Intersample Attention Transformer	0.76 (0.73-0.78)	0.74 (0.69-0.79)	0.77 (0.75-0.78)	0.23 (0.21-0.26)	0.36 (0.32-0.39)	0.86 (0.84-0.88)

[a]AUC: area under the receiver operating characteristic curve.

Table 3. Values of the evaluation metrics of the models in the external validation data set.

Models	Balanced accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F_1-score (95% CI)	AUC[a] (95% CI)
Logistic regression	0.73 (0.74-0.76)	0.91 (0.89-0.93)	0.58 (0.57-0.59)	0.18 (0.17-0.19)	0.30 (0.28-0.31)	0.83 (0.82-0.85)
Random forest	0.75 (0.73-0.77)	0.70 (0.67-0.73)	0.80 (0.79-0.81)	0.26 (0.24-0.28)	0.38 (0.36-0.40)	0.84 (0.82-0.85)
Extreme gradient boosting	0.75 (0.74-0.77)	0.82 (0.79-0.84)	0.69 (0.68-0.70)	0.21 (0.19-0.22)	0.33 (0.31-0.35)	0.84 (0.83-0.85)
Light gradient boosting machine	0.73 (0.71-0.75)	0.57 (0.54-0.61)	0.89 (0.88-0.89)	0.34 (0.31-0.36)	0.42 (0.40-0.45)	0.85 (0.85-0.86)
Categorical boosting	0.76 (0.75-0.77)	0.81 (0.78-0.84)	0.71 (0.70-0.72)	0.22 (0.21-0.23)	0.35 (0.33-0.36)	0.85 (0.83-0.86)
Multilayer perceptron	0.75 (0.74-0.76)	0.89 (0.86-0.91)	0.62 (0.61-0.63)	0.19 (0.18-0.20)	0.31 (0.29-0.33)	0.84 (0.82-0.85)
TabNet	0.69 (0.68-0.70)	0.93 (0.91-0.94)	0.45 (0.44-0.46)	0.14 (0.14-0.15)	0.25 (0.24-0.26)	0.78 (0.77-0.80)
TabTransformer	0.75 (0.74-0.76)	0.92 (0.90-0.94)	0.57 (0.57-0.59)	0.18 (0.17-0.19)	0.30 (0.28-0.31)	0.83 (0.82-0.85)
Self-Attention and Intersample Attention Transformer	0.77 (0.76-0.78)	0.87 (0.85-0.90)	0.67 (0.66-0.68)	0.21 (0.20-0.22)	0.34 (0.32-0.36)	0.85 (0.84-0.87)

[a]AUC: area under the receiver operating characteristic curve.

In general, the SAINT model performed best in both the internal test data set and the external validation data set among the 9 prediction models, with an AUC of 0.86 (95% CI 0.84-0.88) and 0.85 (95% CI 0.84-0.87), respectively. The corresponding balanced accuracy of the internal test data set and external validation data set were 0.76 (95% CI 0.73-0.78) and 0.77 (95% CI 0.76-0.78), sensitivities were 0.74 (95% CI 0.69-0.79) and 0.87 (95% CI 0.85-0.90), specificities were 0.77 (95% CI 0.75-0.78) and 0.67 (95% CI 0.66-0.68), precision values were 0.23 (95% CI 0.21-0.26) and 0.21 (95% CI 0.20-0.22), and F_1-score were 0.36 (95% CI 0.32-0.39) and 0.34 (95% CI 0.32-0.36), respectively. The calibration curves demonstrated good consistency between the predicted and observed probabilities for the SAINT model with curves close to the 45° line, and the Brier scores were 0.07 (95% CI 0.01-0.21) and 0.07 (95% CI 0.02-0.17) in the internal test data set and external validation data set, respectively (Figures 2C and 2D). In addition, the decision curve analysis (Figures 2E and 2F) illustrated that the SAINT model was clinically beneficial in both the internal test data set and external validation data set.

Moreover, we tried to train the model using different data sets and evaluated the performance of the models in different external validation data sets. The results showed that the SAINT model performed better than other models (Figures S2 and S3 and Tables S9-S12 in Multimedia Appendix 1). Furthermore, we evaluated the model's predictive power using the measurements obtained within the first 2 hours of patient admission. This evaluation aimed to assess the predictive power of the models for the early diagnostic stage of patients with AMI. The results demonstrated that although the SAINT model still outperformed the other models, there was a clear decrease in the predictive power of all models, with an average decrease of 0.06 (SD 0.02) in the AUC. (Figure S4 and Table S13 in Multimedia Appendix 1).

Model Explanation Results

The relative importance and contribution of the top 20 features to the SAINT model are shown in Figure 3A, which was obtained by the SHAP algorithm interpreting the predictions of the SAINT model. The importance and contribution of the full features to the SAINT model are presented in Figure S1 in Multimedia Appendix 1. Age was ranked as the most important

XSL•FO
RenderX

feature for the prediction model, followed by heart rate, temperature, white blood cell level, systolic blood pressure, magnesium level, prothrombin time, calcium level, creatinine level, and sodium level. Considering the impact of age on the predictions as an example, advanced age was commonly associated with a higher risk of mortality, and younger age generally drove the predictions toward survival (red dots represent a high value of the features, which on the right side of the x-axis represents an increased risk of mortality). Similar explanations can be applied to other features, and most interpretations were consistent with clinical experience. It is notable that features can drive predictions in different directions (increasing or decreasing mortality predictions) for patients with different features, unlike previous scoring systems (eg, Global Registry of Acute Coronary Events risk score [35] and Thrombolysis in Myocardial Infarction score [36]), where the influence of a given feature value on the prediction is invariant. The 9 most important features for the prediction models showed a statistically significant difference between the survival group and the nonsurvival group in the training data set ($P<.05$), which further supports the rationality of the explanations of the features (Table 4).

We further mapped the SHAP values to the probability of driving an increase or decrease in mortality and visualized the application of the method to 1 deceased patient and 1 surviving patient in the external validation data set (Figures 3B and 3C). In the case of the deceased patient, the patient was female and aged 89 years with a history of heart failure and renal failure. The model predicted that the risk of mortality of the patients was 0.551. High levels of serum sodium (185 mEq/L) and serum magnesium (3.2 mg/dL) drove a 0.29 and 0.21 increase in the risk of mortality, respectively, whereas a high serum chloride level was associated with a 0.17 reduction in the risk of mortality. A similar explanation can be applied to other features. The prediction was driven by 39 features used for model training, and the sum of the SHAP values for all features plus the baseline risk equaled the predicted risk of mortality. The baseline risk score was obtained by calculating the average predicted risk of mortality among all patients in the training data set (0.079). Therefore, the SHAP algorithm made our model explainable at both the global and individual levels.

Figure 3. The impact of the input features on predictions: (A) the effect of the input features on the model at the global level. Each dot represents the influence of a feature on the prediction for a patient. Red dots indicate a higher value of the features, and blue dots indicate a lower value of the features. A dot on the left side of the x-axis represents that the feature value decreased mortality prediction, and a dot on the right side of the x-axis represents that the feature value increased mortality prediction, (B) an example of risk factor analysis for a patient in the external validation set (in-hospital death), and (C) an example of risk factor analysis for a patient in the external validation set (survival at hospital discharge). SHAP: Shapley Additive Explanations.

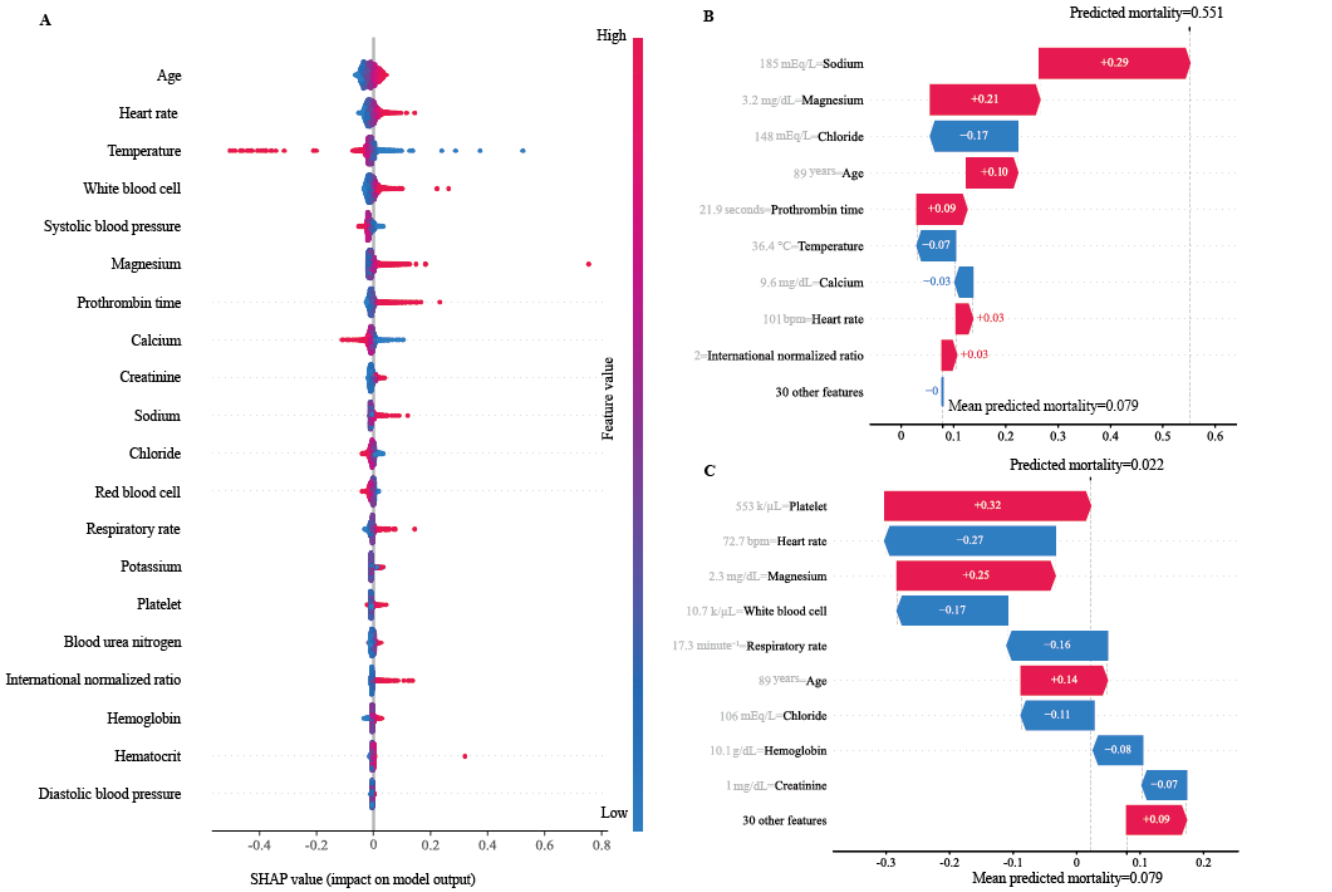

XSL•FO
RenderX

Table 4. Baseline characteristics of patients with acute myocardial infarction in the training data set by clinical outcomes (n=7668).

Variables	Survival group (n=6998)	Nonsurvival group (n=670)	*P* value[a]
Demographic data			
Age (years), median (IQR)	66 (56-75)	75 (66-83)	*<.001*
Sex, n (%)			<.001
Male	4661 (66.6)	369 (55.1)	
Female	2337 (33.4)	301 (44.9)	
Vital signs, median (IQR)			
Heart rate (beats/minute)	80 (69-94)	92 (78-108)	*<.001*
Respiratory rate (minute^{-1})	20 (17-21)	20 (18-25)	*<.001*
Systolic blood pressure (mm Hg)	130 (113-149)	116 (98-142)	*<.001*
Diastolic blood pressure (mm Hg)	77 (65-88)	68 (57-81)	*<.001*
Temperature (°C)	36.6 (36.4-36.8)	36.5 (36.1-36.8)	*<.001*
Laboratory data, median (IQR)			
Troponin T or I (ng/mL)	0.66 (0.10-4.47)	1.77 (0.26-8.78)	*<.001*
White blood cell (k/μ)	10.4 (8.2-13.4)	12.4 (9.1-16.8)	*<.001*
Red blood cell (m/μ)	4.45 (3.94-4.88)	4.05 (3.48-4.59)	*<.001*
Hemoglobin (g/dL)	13.6 (11.8-14.9)	12.1 (10.3-14)	*<.001*
Hematocrit (%)	40.4 (35.9-44)	37 (32.1-42.4)	*<.001*
Platelet (k/μ)	215 (174-264)	213 (153-270)	.06
Prothrombin time (seconds)	13.2 (12.2-14.3)	14.5 (13.2-17.2)	*<.001*
International normalized ratio	1.04 (1-1.16)	1.2 (1.08-1.49)	*<.001*
Sodium (mEq/L)	138 (136-140)	138 (135-141)	.06
Potassium (mEq/L)	4 (3.7-4.3)	4.2 (3.73-4.70)	*<.001*
Chloride (mEq/L)	103.6 (100.4-106)	103 (98.5-106)	*<.001*
Calcium (mg/dL)	8.9 (8.5-9.3)	8.6 (8.18-9.10)	*<.001*
Magnesium (mg/dL)	1.99 (1.80-2.19)	2 (1.71-2.30)	*<.001*
Creatinine (mg/dL)	1.01 (0.82-1.30)	1.42 (1.06-2.14)	*<.001*
Blood urea nitrogen (mg/dL)	19 (14-27)	27 (18-41)	*<.001*
Comorbidities, n (%)			
Hypertension	4209 (60.1)	407 (60.7)	.76
Hyperlipidemia	767 (11)	53 (7.9)	*.02*
Diabetes	2249 (32.1)	267 (39.9)	*<.001*
Heart failure	1688 (24.1)	299 (44.6)	*<.001*
Renal failure	1333 (19)	315 (47)	*<.001*
Cerebral vascular disease	670 (9.6)	118 (17.6)	*<.001*
Peripheral vascular disease	971 (13.9)	91 (13.6)	.83
Chronic obstructive pulmonary disease	749 (10.7)	115 (17.2)	*<.001*
Operations, n (%)			
Coronary arteriography	3670 (52.4)	185 (27.6)	*<.001*
Percutaneous coronary intervention	2920 (41.7)	163 (24.3)	*<.001*
Drugs, n (%)			
Aspirin	4583 (65.5)	358 (53.4)	*<.001*

XSL•FO
RenderX

Variables	Survival group (n=6998)	Nonsurvival group (n=670)	*P* value[a]
Clopidogrel	1943 (27.8)	127 (19)	*<.001*
Ticagrelor	1413 (20.2)	78 (11.6)	*<.001*
Angiotensin-converting enzyme inhibitor	1777 (25.4)	50 (7.5)	*<.001*
Angiotensin receptor blocker	233 (3.3)	7 (1)	*.001*
Statins	3512 (50.2)	216 (32.2)	*<.001*
β-Blockers	3369 (48.1)	196 (29.3)	*<.001*
Clinical outcome, median (IQR)			
Length of hospital stay (days)	5.15 (2.59-9.34)	3.71 (1.23-8.27)	*<.001*

[a]Italicized *P* values (<.05) were considered statistically significant.

Discussion

Principal Findings

In this multicenter, retrospective study, we trained 3 state-of-the-art deep learning models on patients with AMI from Chongqing University Central Hospital and patient data from the eICU-CRD and compared their performance with 4 advanced tree-based models, an LR model, and an MLP. We performed both internal and external validation (MIMIC-IV database) to confirm the predictive ability of the models and ensure the reliability of the results. In addition, the SHAP algorithm was used to interpret the predictions of the best-performing model (SAINT) to quantify and visualize the impact of the features on the predicted mortality at global and individual levels.

The models developed in this study had similar AUC in the internal test data set and the external validation data set, which demonstrated a good generalization ability of the models. The training data set we used was collected from 209 centers in Asia and the United States, and the multicenter data contributed to improving the performance of the model, especially in terms of generalization ability. However, although the specificities of the models in the internal test data set were generally higher and the sensitivities were lower than that in the external validation, the models tended to overestimate patient mortality in the external validation data set, leading to false positive results. The SAINT model is a hybrid deep learning–based approach specifically designed to solve tabular data problems and outperforms tree-based models over a variety of benchmark tasks [25]. In this study, the SAINT model performed best in both the internal test data set and the external validation data set among the 9 prediction models, with the highest AUC and lowest Brier score. In addition, the decision curve analysis indicated that the SAINT model could provide net benefits across an extensive range of threshold values. The SAINT model attained the highest balanced accuracy during evaluation on the external validation data set, indicating that the model had strong classification ability. Nevertheless, it is worth noting that both precision and specificity metrics associated with the SAINT model did not exhibit high values, implying a proclivity of the model to overestimate the mortality risk among patients with AMI. According to these results, state-of-the-art deep learning methods have greater potential in accurately predicting mortality in patients with AMI than linear, MLP, and tree-based models.

The SHAP algorithm was used to obtain an interpretation of the predictions of the SAINT model, and the impact of most features on the predictions is consistent with clinical practice and previous evidence. For instance, advanced age, unstable vital signs, presence of infection, metabolic disorders, and coagulation disorders lead to increased predicted mortality. The results of the model interpretation indicated that advanced age was ranked as the most important risk factor for mortality. In addition, as patients age, their physical function and resistance decline, which may increase their risk of mortality [37,38]. Unstable vital signs (eg, elevated heart rate and respiratory rate and decreased levels of systolic blood pressure and temperature) usually indicate that the patient is in a critical condition, and such patients tend to have a worse prognosis [39,40]. Similarly, the presence of infection (eg, elevated levels of white blood cells) may impair physical function and reduce the effectiveness of treatment [41]. In addition, metabolic and coagulation disorders (eg, elevated levels of magnesium and sodium and increased prothrombin time) are associated with an increased risk of mortality [42].

LR models and Cox proportional hazard models are commonly used tools for predicting the mortality of patients with AMI [6]. However, the rapid development of medical information technology brings massive, nonlinear, and high-dimensional data, which challenge the generalized linear model [43]. Tree-based machine learning models (eg, RF, XGBoost, and light gradient boosting machine) achieved substantial gains in the predictive accuracy of tabular data [4,13,14]. Nevertheless, tree-based models were developed directly using the information entropy of data without calculating the loss function and performing backpropagation, which limits the model from taking advantage of powerful deep learning modules and conducting web-based learning [29,30]. Recently, deep learning models have received extensive attention in the screening, diagnosis, and prognosis prediction of diseases, and their performance even outperformed that of clinicians in some aspects [44]. A recent, large population-based study by Mohammad et al [45] revealed the tremendous potential of deep learning models to identify the risk of adverse outcomes in patients with AMI. However, the prediction model used in this study is the most basic deep learning model (MLP), and the ability of state-of-the-art deep learning models to predict the mortality of patients with AMI remains unclear. In addition, the model used in the study by Mohammad et al [45] is regarded as a black-box

XSL•FO
RenderX

model due to the complex internal calculation process, that is, we can only see the input and final predictions of the model but cannot understand how the model makes decisions. We used the SHAP algorithm to quantify and visualize the features that drive the predictions, which combines the advantages of complex deep learning models and simple linear models to solve the problem of insufficient predictive ability of the linear model and black-box model of the complex deep learning model. The effect of each feature on the prediction is not isolated but derived from the interaction with the other features, which makes it superior to the fixed feature weight in the linear models [33]. Furthermore, the SHAP algorithm not only provides the impact of the features on the model but also provides the interpretation at an individual level. Such a patient-level interpretable model allows clinicians to combine the predictions and explanations with their empirical knowledge to facilitate decision-making [19].

Limitations

This study has several limitations. First, to enable the models to learn more comprehensive information and improve performance, up to 39 features were used to develop the model. Nevertheless, the variables included in the study were limited by the retrospective design and the use of publicly available databases. Second, there were missing values in the raw data; although the KNN algorithm was used to complete them, the real data could provide more valuable information. This limitation was particularly significant during the early stages of patient hospital admission when the available predictors were limited. Even with the application of the KNN algorithm to complete the missing values, the predictive power of the model could be compromised. Third, although the SHAP algorithm was used to obtain the explanations of the models, the impact of features on the predictions is not equal to the association in the causal chain. In addition, features such as age and sex, while influencing predictions, were not subject to human intervention. However, the explanations of the relationship between features and predictions may guide us in examining causation.

Conclusions

In conclusion, in this multicenter study, we developed and externally validated 9 prediction models for in-hospital mortality in patients with AMI. Our results indicate that state-of-the-art deep learning models may provide a promising alternative solution to further improve the predictive ability of mortality in patients with AMI. The explanations of the model suggest that older age, unstable vital signs, and metabolic disorders may increase the risk of mortality in patients with AMI. Transparent and explainable models are expected to support clinical practice, trial design, and future research. However, future prospective studies are needed to confirm the findings.

Acknowledgments

This study was supported by Independent Research Project of Medical Engineering Laboratory of Chinese PLA General Hospital (2022SYSZZKY07), National Natural Science Foundation of China (62076247, 72374206), National Key R&D Program of China (2021ZD0140408), and Technology Innovation and Application Development Special Project of Chongqing, China (CSTB2023TIAD-KPX0047).

Data Availability

The data sets from the Medical Information Mart for Intensive Care-IV database and Electronic Intensive Care Unit Collaborative Research Database presented in this study can be found in web-based repositories [46,47]. The data set from Chongqing University Central Hospital is not publicly available but is available from the corresponding author on reasonable request.

Authors' Contributions

PX was responsible for the study design, data acquisition, analysis, and interpretation and for preparing the first draft of the manuscript. HW, JX, FX, JL, ZC, WZ, SH, and DW were responsible for recruiting study participants and providing clinical input to the study design. YM and JX designed the study and helped write the manuscript. All authors had an opportunity to review the manuscript and approved its final submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 1018 KB - jmir_v26i1e49848_app1.docx]

References

1. Reed GW, Rossi JE, Cannon CP. Acute myocardial infarction. Lancet 2017 Jan;389(10065):197-210. [doi: 10.1016/s0140-6736(16)30677-8]

XSL•FO
RenderX

2. Cui Y, Hao K, Takahashi J, Miyata S, Shindo T, Nishimiya K, et al. Age-specific trends in the incidence and in-hospital mortality of acute myocardial infarction over 30 years in Japan— report from the Miyagi AMI registry study. Circ J 2017;81(4):520-528. [doi: 10.1253/circj.cj-16-0799]
3. Szummer K, Wallentin L, Lindhagen L, Alfredsson J, Erlinge D, Held C, et al. Improved outcomes in patients with ST-elevation myocardial infarction during the last 20 years are related to implementation of evidence-based treatments: experiences from the SWEDEHEART registry 1995-2014. Eur Heart J 2017 Nov 01;38(41):3056-3065. [doi: 10.1093/eurheartj/ehx515] [Medline: 29020314]
4. Figtree GA, Vernon ST, Hadziosmanovic N, Sundström J, Alfredsson J, Arnott C, et al. Mortality in STEMI patients without standard modifiable risk factors: a sex-disaggregated analysis of SWEDEHEART registry data. Lancet 2021 Mar 20;397(10279):1085-1094. [doi: 10.1016/S0140-6736(21)00272-5] [Medline: 33711294]
5. Castro-Dominguez Y, Dharmarajan K, McNamara RL. Predicting death after acute myocardial infarction. Trends Cardiovasc Med 2018 Feb;28(2):102-109. [doi: 10.1016/j.tcm.2017.07.011] [Medline: 28826668]
6. Funk M, Pooley-Richards RL. Predicting hospital mortality in patients with acute myocardial infarction. Am J Crit Care 1994 May;3(3):168-176. [Medline: 8038844]
7. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol 2004 Nov;57(11):1138-1146. [doi: 10.1016/j.jclinepi.2004.04.003] [Medline: 15567629]
8. Amin AP, Spertus JA, Reid KJ, Lan X, Buchanan DM, Decker C, et al. The prognostic importance of worsening renal function during an acute myocardial infarction on long-term mortality. Am Heart J 2010 Dec;160(6):1065-1071. [doi: 10.1016/j.ahj.2010.08.007] [Medline: 21146659]
9. Huang Y, Zheng Z, Ma M, Xin X, Liu H, Fei X, et al. Improving the performance of outcome prediction for inpatients with acute myocardial infarction based on embedding representation learned from electronic medical records: development and validation study. J Med Internet Res 2022 Aug 03;24(8):e37486. [doi: 10.2196/37486] [Medline: 35921141]
10. Mood C. Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur Sociol Rev 2010 Mar 09;26(1):67-82. [doi: 10.1093/esr/jcp006]
11. Breiman L. Random forests. Mach Learn 2001;45:5-32 [FREE Full text]
12. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA p. 785-794 URL: https://dl.acm.org/doi/10.1145/2939672.2939785 [doi: 10.1145/2939672.2939785]
13. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol 2021 Jun 01;6(6):633-641. [doi: 10.1001/jamacardio.2021.0122] [Medline: 33688915]
14. Oliveira M, Seringa J, Pinto FJ, Henriques R, Magalhães T. Machine learning prediction of mortality in acute myocardial infarction. BMC Med Inform Decis Mak 2023 Apr 18;23(1):70. [doi: 10.1186/s12911-023-02168-6] [Medline: 37072766]
15. Xie P, Yang C, Yang G, Jiang Y, He M, Jiang X, et al. Mortality prediction in patients with hyperglycaemic crisis using explainable machine learning: a prospective, multicentre study based on tertiary hospitals. Diabetol Metab Syndr 2023 Mar 11;15(1):44. [doi: 10.1186/s13098-023-01020-1] [Medline: 36899433]
16. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. IEEE Trans Neural Netw Learn Syst 2022:1-21 [FREE Full text] [doi: 10.1109/tnnls.2022.3229161]
17. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access 2019;7:53040-53065. [doi: 10.1109/access.2019.2912200]
18. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion 2022 Jan;77:29-52. [doi: 10.1016/j.inffus.2021.07.016] [Medline: 34980946]
19. Liang Y, Li S, Yan C, Li M, Jiang C. Explaining the black-box model: a survey of local interpretation methods for deep neural networks. Neurocomputing 2021 Jan;419:168-182. [doi: 10.1016/j.neucom.2020.08.011]
20. Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. J Med Internet Res 2022 Aug 09;24(8):e38082. [doi: 10.2196/38082] [Medline: 35943767]
21. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. Author correction: MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 16;10(1):31. [doi: 10.1038/s41597-023-01945-2] [Medline: 36646711]
22. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5(1):180178. [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
23. Payrovnaziri SN, Xing A, Salman S, Liu X, Bian J, He Z. Assessing the impact of imputation on the interpretations of prediction models: a case study on mortality prediction for patients with acute myocardial infarction. AMIA Jt Summits Transl Sci Proc 2021;2021:465-474. [Medline: 34457162]
24. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021 Oct 27;8(1):140. [doi: 10.1186/s40537-021-00516-9] [Medline: 34722113]

XSL•FO
RenderX

25. Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. Preprint posted online June 2, 2021.
26. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: tabular data modeling using contextual embeddings. Preprint posted online December 11, 2020 [FREE Full text]
27. Arik S, Pfister T. TabNet: attentive interpretable tabular learning. AAAI Conf Artif Intell 2021;35(8):6679-6687 [FREE Full text] [doi: 10.1609/aaai.v35i8.16826]
28. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943 Dec;5(4):115-133. [doi: 10.1007/bf02478259]
29. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 3149-3157 URL: https://dl.acm.org/doi/10.5555/3294996.3295074
30. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018 Presented at: NIPS '18; December 3-8, 2018; Montréal, QC p. 6639-6649 URL: https://dl.acm.org/doi/abs/10.5555/3327757.3327770
31. DeMaris A. A tutorial in logistic regression. J Marriage Fam 1995 Nov;57(4):956. [doi: 10.2307/353415]
32. PuguangXie/AMI. GitHub. URL: https://github.com/PuguangXie/AMI [accessed 2024-04-29]
33. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan 17;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
34. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 4768-4777 URL: https://dl.acm.org/doi/10.5555/3295222.3295230
35. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, Global Registry of Acute Coronary Events Investigators. Predictors of hospital mortality in the global registry of acute coronary events. Arch Intern Med 2003 Oct 27;163(19):2345-2353. [doi: 10.1001/archinte.163.19.2345] [Medline: 14581255]
36. Morrow DA, Antman EM, Charlesworth A, Cairns R, Murphy SA, de Lemos JA, et al. TIMI risk score for ST-elevation myocardial infarction: a convenient, bedside, clinical score for risk assessment at presentation: An intravenous nPA for treatment of infarcting myocardium early II trial substudy. Circulation 2000 Oct 24;102(17):2031-2037. [doi: 10.1161/01.cir.102.17.2031] [Medline: 11044416]
37. Cosentino N, Resta ML, Somaschini A, Campodonico J, Lucci C, Moltrasio M, et al. Acute kidney injury and in-hospital mortality in patients with ST-elevation myocardial infarction of different age groups. Int J Cardiol 2021 Dec 01;344:8-12. [doi: 10.1016/j.ijcard.2021.09.023] [Medline: 34537309]
38. Căruntu F, Bordejevic DA, Buz B, Gheorghiu A, Tomescu MC. Independent predictors of in-hospital and 1-year mortality rates in octogenarians with acute myocardial infarction. Rev Cardiovasc Med 2021 Jun 30;22(2):489-497. [doi: 10.31083/j.rcm2202056] [Medline: 34258917]
39. Lv J, Wang C, Gao X, Yang J, Zhang X, Ye Y, China Acute Myocardial Infarction Registry study group. Development and validation of dynamic models to predict postdischarge mortality risk in patients with acute myocardial infarction: results from China Acute Myocardial Infarction Registry. BMJ Open 2023 Mar 29;13(3):e069505. [doi: 10.1136/bmjopen-2022-069505] [Medline: 36990493]
40. Li L, Zhang X, Wang Y, Yu X, Jia H, Hou J, et al. A novel risk score to predict in-hospital mortality in patients with acute myocardial infarction: results from a prospective observational cohort. Front Cardiovasc Med 2022;9:840485. [doi: 10.3389/fcvm.2022.840485] [Medline: 35463775]
41. Tan L, Xu Q, Shi R. A nomogram for predicting hospital mortality in intensive care unit patients with acute myocardial infarction. Int J Gen Med 2021;14:5863-5877. [doi: 10.2147/IJGM.S326898] [Medline: 34566426]
42. Mizuguchi Y, Konishi T, Nagai T, Sato T, Takenaka S, Tada A, et al. Prognostic value of admission serum magnesium in acute myocardial infarction complicated by malignant ventricular arrhythmias. Am J Emerg Med 2021 Jun;44:100-105. [doi: 10.1016/j.ajem.2021.02.005] [Medline: 33582610]
43. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/JBHI.2016.2636665] [Medline: 28055930]
44. Egger J, Gsaxner C, Pepe A, Pomykala KL, Jonske F, Kurz M, et al. Medical deep learning-a systematic meta-review. Comput Methods Programs Biomed 2022 Jun;221:106874. [doi: 10.1016/j.cmpb.2022.106874] [Medline: 35588660]
45. Mohammad MA, Olesen KK, Koul S, Gale CP, Rylance R, Jernberg T, et al. Development and validation of an artificial neural network algorithm to predict mortality and admission to hospital for heart failure after myocardial infarction: a nationwide population-based study. Lancet Digit Health 2022 Jan;4(1):e37-e45. [doi: 10.1016/s2589-7500(21)00228-4]
46. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.2). PhysioNet. 2023 Jan 6. URL: https://physionet.org/content/mimiciv/2.2/ [accessed 2024-04-29]
47. eICU-CRD v2.0 documentation. eICU Collaborative Research Database. URL: https://eicu-crd.mit.edu/about/eicu/ [accessed 2024-04-29]

XSL•FO
RenderX

Abbreviations

AMI: acute myocardial infarction
AUC: area under the receiver operating characteristic curve
eICU-CRD: Electronic Intensive Care Unit Collaborative Research Database
ICD: International Classification of Diseases
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Statistical Classification of Diseases, Tenth Revision
KNN: K-nearest neighbors
LR: logistic regression
MIMIC-IV: Medical Information Mart for Intensive Care-IV
MLP: multilayer perceptron
RF: random forest
SAINT: Self-Attention and Intersample Attention Transformer
SHAP: Shapley Additive Explanations
XGBoost: extreme gradient boosting

Edited by T de Azevedo Cardoso; submitted 12.06.23; peer-reviewed by S Schallmoser, JW Li, Y Wei; comments to author 14.09.23; revised version received 02.10.23; accepted 02.04.24; published 10.05.24.

Please cite as:
Xie P, Wang H, Xiao J, Xu F, Liu J, Chen Z, Zhao W, Hou S, Wu D, Ma Y, Xiao J
Development and Validation of an Explainable Deep Learning Model to Predict In-Hospital Mortality for Patients With Acute Myocardial Infarction: Algorithm Development and Validation Study
J Med Internet Res 2024;26:e49848
URL: https://www.jmir.org/2024/1/e49848
doi:10.2196/49848
PMID:38728685

©Puguang Xie, Hao Wang, Jun Xiao, Fan Xu, Jingyang Liu, Zihang Chen, Weijie Zhao, Siyu Hou, Dongdong Wu, Yu Ma, Jingjing Xiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Potential of Large Language Models in Health Care: Delphi Study

Kerstin Denecke[1], Dr rer nat; Richard May[2], MEng; LLMHealthGroup[3]; Octavio Rivera Romero[4,5], PhD

[1]Bern University of Applied Sciences, Biel, Switzerland
[2]Harz University of Applied Sciences, Wernigerode, Germany
[3]see Acknowledgments
[4]Instituto de Ingeniería Informática (I3US), Universidad de Sevilla, Sevilla, Spain
[5]Department of Electronic Technology, Universidad de Sevilla, Sevilla, Spain

Corresponding Author:
Kerstin Denecke, Dr rer nat
Bern University of Applied Sciences
Quallgasse 21
Biel, 2502
Switzerland
Phone: 41 323216794
Email: kerstin.denecke@bfh.ch

Abstract

Background: A large language model (LLM) is a machine learning model inferred from text data that captures subtle patterns of language use in context. Modern LLMs are based on neural network architectures that incorporate transformer methods. They allow the model to relate words together through attention to multiple words in a text sequence. LLMs have been shown to be highly effective for a range of tasks in natural language processing (NLP), including classification and information extraction tasks and generative applications.

Objective: The aim of this adapted Delphi study was to collect researchers' opinions on how LLMs might influence health care and on the strengths, weaknesses, opportunities, and threats of LLM use in health care.

Methods: We invited researchers in the fields of health informatics, nursing informatics, and medical NLP to share their opinions on LLM use in health care. We started the first round with open questions based on our strengths, weaknesses, opportunities, and threats framework. In the second and third round, the participants scored these items.

Results: The first, second, and third rounds had 28, 23, and 21 participants, respectively. Almost all participants (26/28, 93% in round 1 and 20/21, 95% in round 3) were affiliated with academic institutions. Agreement was reached on 103 items related to use cases, benefits, risks, reliability, adoption aspects, and the future of LLMs in health care. Participants offered several use cases, including supporting clinical tasks, documentation tasks, and medical research and education, and agreed that LLM-based systems will act as health assistants for patient education. The agreed-upon benefits included increased efficiency in data handling and extraction, improved automation of processes, improved quality of health care services and overall health outcomes, provision of personalized care, accelerated diagnosis and treatment processes, and improved interaction between patients and health care professionals. In total, 5 risks to health care in general were identified: cybersecurity breaches, the potential for patient misinformation, ethical concerns, the likelihood of biased decision-making, and the risk associated with inaccurate communication. Overconfidence in LLM-based systems was recognized as a risk to the medical profession. The 6 agreed-upon privacy risks included the use of unregulated cloud services that compromise data security, exposure of sensitive patient data, breaches of confidentiality, fraudulent use of information, vulnerabilities in data storage and communication, and inappropriate access or use of patient data.

Conclusions: Future research related to LLMs should not only focus on testing their possibilities for NLP-related tasks but also consider the workflows the models could contribute to and the requirements regarding quality, integration, and regulations needed for successful implementation in practice.

(J Med Internet Res 2024;26:e52399) doi:10.2196/52399

XSL•FO
RenderX

KEYWORDS

large language models; LLMs; health care; Delphi study; natural language processing; NLP; artificial intelligence; language model; Delphi; future; innovation; interview; interviews; informatics; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; implementation

Introduction

Background

A large language model (LLM) is a machine learning model that encodes complex patterns of language use derived from vast quantities of input texts [1,2]. Modern LLMs use neural network architectures, typically enhanced with a transformer attention mechanism that captures associative relationships between words based on shared context [3]. They use attention or self-attention to identify how distant data elements influence and depend on one another. Specifically, transformers learn context by tracking relationships in sequential data such as words in a sentence. Typically, transformer-based models are trained in 2 phases: the *pretraining phase* focuses on generic representation learning, and the *transfer learning phase* focuses on adjusting the model to an application-specific prediction task [4]. The pretrained models, which are often trained on large data sets (eg, Wikipedia, Reddit, biomedical literature, or public medical data sets), are tuned to be used for a wider set of tasks and can be fine-tuned for specific tasks [5]. In this study, we consider LLMs that use transformers for tasks related to medical natural language processing (NLP) tasks.

First described in 2017 by researchers from Google [3], LLMs are very well suited to NLP [6,7] for tasks such as machine translation [8], document summarization [9], natural language generation [10], and emotion recognition [11]. For example, Yang et al [12] explored LLMs for clinical concept extraction. Specifically, they tested 4 architectures—Bidirectional Encoder Representations From Transformers (BERT) [13], Robustly optimized BERT approach, A lite BERT, and Efficiently Learning an Encoder that Classifies Token Replacements Accurately—and achieved F_1-scores between 93% and 95%. The public has gained widespread awareness of LLMs starting in 2022 with the release of ChatGPT, which uses a generative pretrained transformer model. The studies on ChatGPT demonstrate a huge potential, but some have identified limitations [14]. For instance, Cocci et al [15] assessed ChatGPT as a tool for providing medical information to patients in the context of urology. They compared output generated by ChatGPT to that provided by a board-certified urologist. Although they concluded that this approach has "the potential to enhance health outcomes and patient satisfaction" [15], they also identified an inadequate appropriateness and quality of responses. Many use case–specific appraisals of LLM technology beyond the ones outlined here are becoming available. For this reason, it is useful to identify and summarize the overall potential benefits and risks of these technologies.

Several individual opinion papers on the strengths, weaknesses, opportunities, and threats (SWOT) of LLMs in general, and ChatGPT in particular, have been published recently. Farrokhnia et al [16] conducted a SWOT analysis and outlined ChatGPT's strengths and weaknesses with a specific focus on education. In a systematic review, Garg et al [17] described articles on the use of ChatGPT in clinical practice and medical research to summarize the potential of ChatGPT in diagnosing and treating patients as well as its possible contributions to medical research. Lee et al [18] have also studied the benefits, limits, and risks of ChatGPT in the medical domain. Although multiple studies have assessed the quality of LLMs for various tasks in the health care domain [19-21], a comprehensive description of the potential benefits and risks of LLMs, along with their relative importance according to multiple participants from a range of settings, has yet to be provided.

We believe that aggregated researchers' opinions are needed because individual studies can never reflect the entire potential and risks might have been overlooked. It is essential to ensure a responsible and safe use to protect patients' interests and foster trust in artificial intelligence (AI)–driven health care technologies. Such work can guide future research and development efforts to effectively address specific health care challenges.

Objectives

Overall, the objectives of our study were to seek researchers' opinions on (1) the likelihood that LLMs will be adopted in health care and for what purposes, (2) the likely benefits of LLMs in health care, (3) the shortcomings and risks of LLM adoption in health care, (4) the requirements for the adoption of LLMs in health care, and (5) the reliability of LLMs in future health care.

Where there was substantial agreement among respondents, the responses were analyzed to identify the SWOT of LLMs and derive practical and research implications.

Methods

Overview

We followed a 2-step process. First, a modified Delphi method [22] was used to aggregate opinions on the potential and limitations of LLMs in health care. Second, to increase the practical relevance of the analysis, the results were aggregated into the SWOT of LLMs. The Delphi method is widely used to evaluate consensus, or lack thereof, among participants [23]. Similarly, the Delphi method has been considered suitable for exploratory idea generation on complex and multidisciplinary topics, especially if the objective of the research is the analysis of new trends [24-28]. In addition, this method has been widely applied to the health care domain [19-21,29-32].

Our iterative process consisted of four stages (Figure 1): (1) preparatory phase, (2) Delphi rounds, (3) data processing and analysis, and (4) concluding and reporting.

XSL•FO
RenderX

Figure 1. Stages of the study process.

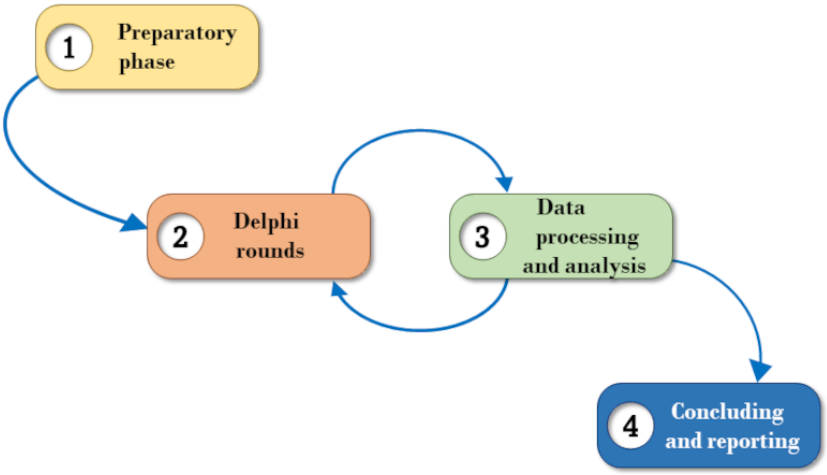

Study Development

Preparatory Phase

The first stage in the defined Delphi process sought to develop the initial questionnaire for the first round. An iterative development process was followed. First, potential items for inclusion were identified by KD and RM by formulating questions and statements referring to demographic data as well as data related to the SWOT analysis (Textbox 1). Then, a third researcher (ORR) reviewed all the questions and provided feedback. Discrepancies were resolved through consensus among the 3 researchers.

Textbox 1. Questions driving the strengths, weaknesses, opportunities, and threats analysis.

Internal features

- Strengths
  - What are the advantages of large language models (LLMs) in health care?
  - What are the achievements of LLMs in health care?
  - Do LLMs already significantly improve digitized processes in health care?
- Weaknesses
  - Which disadvantages of applying LLMs in health care exist?
  - Are LLMs sufficiently developed to be actually reliable?
  - Do LLMs support saving costs as well as ecological sustainability?

External features

- Opportunities
  - What are current trends that support LLMs in health care?
  - What are shortcomings in health care that can be exploited by LLMs?
  - Do LLMs benefit from specific developments in both artificial intelligence and health care?
- Threats
  - Which risks could emerge for health professionals, patient care, data protection, or general health IT due to the adoption of LLMs?
  - Do LLMs contribute to discrimination in health care?
  - Are health professionals sufficiently prepared for successfully adopting and using LLMs without losing competencies?
  - Are LLMs useful and accepted by health professionals and patients?

We used a SWOT analysis [33] as it is a strategic planning method that takes into account both the internal and external features of a technology. Internal features refer to strengths (eg, scalability and innovative features) and weaknesses (eg, domain suitability and technical limitations) of the technology. In contrast, external features include opportunities (eg, regulatory support and market demands) as well as technology threats (eg, legal risks and competitor technologies). Overall, a SWOT

XSL•FO
RenderX

analysis provides a basis for strategic decision-making by determining how to leverage strengths, address weaknesses, exploit opportunities, and mitigate threats. First, we collected questions driving our SWOT analysis (Textbox 1). These questions formed the basis to formulate questions for the Delphi questionnaire. After analyzing the answers to the Delphi questions, the answers in which agreement was achieved were used in turn to answer the SWOT analysis questions.

The final version of the initial (round 1) questionnaire consisted of 23 items organized into 5 sections (Multimedia Appendix 1). The first section (7 questions) collected participants' demographic data. The second section included 8 open-ended questions on perceived benefits and shortcomings of using LLMs in health contexts. The third section consisted of 1 statement assessing the potential extent to which future health care could rely on LLMs. Responses on agreement were collected via a 5-point Likert scale with the following options: "very high extent, high extent, neither high nor small extent, small extent, and very small extent." The fourth section consisted of 15 statements to be judged on a 5-point Likert scale with the following options—"strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree"—assessing the future of LLMs. The last section included 2 statements with answers given on a 5-point Likert scale ("very high impact, impact beyond average, average impact, impact below average, and very low to no impact") assessing the potential impact of using LLMs on health care.

A web version of the final round-1 questionnaire was created using the Microsoft Forms software (Microsoft Corp). A researcher (ORR) tested its usability and accessibility using several browsers in a laptop (Google Chrome version 116.0.5845.140, Microsoft Edge version 116.0.1938.62, and Mozilla Firefox version 46.0) and 2 updated browsers (Google Chrome and Samsung Internet Browser apps) in a Samsung Galaxy A52s and resolved these issues before deployment.

Delphi Rounds

After review of round-1 responses, some questions were reworded, and some items were added to the 5-point Likert questions for the round-2 questionnaire. Our approach was to carry all questions and items from a given round to the subsequent round even when agreement was reached in the previous round. We defined agreement as follows: when >75% of the researchers assigned a score equal to 4 or 5, agreement was achieved on this opinion. Participants were given the opportunity to change their minds during the subsequent round, and we were able to calculate the stability of their responses. Participants received the link to access the corresponding e-Delphi questionnaire via email. In addition, each participant received a feedback report including their responses to the questionnaire of the immediately previous round and figures showing the distribution of participants' responses in percentages for each question and item.

Data Processing and Analysis

Quantitative data were analyzed using descriptive statistics in the Microsoft Excel software (Microsoft Corp). We calculated the median and IQR for the responses to each 5-point Likert question in each round. In addition, scores assigned to each item in 2 consecutive rounds were passed as parameters of the Wilcoxon matched-pairs signed rank test to estimate the stability of each item. This test was conducted using the R software (version 3.6.3; R Foundation for Statistical Computing).

The qualitative data were analyzed following a thematic analysis [34]. After reading several times for familiarization, KD and ORR independently coded all data. The codification was aimed at identifying some suggestions to modify the current questionnaire and any additional factors to be included as new items in the subsequent round. Coding was compared, and discrepancies were resolved through consensus among the coders. All factors were then grouped into themes or categories through consensus and added to the next round's questionnaire.

We followed the recommendations of von der Gracht [35] for finding consensus. In this regard, we considered that agreement with an item was reached when the IQR of the participants' responses to this item in the round was ≤1. The IQR is usually found to be a suitable consensus criterion for 4- or 5-unit scales. Following this criterion, we defined "agreement" with an item in a given round as the IQR of the participants' responses being ≤1 and defined "disagreement" otherwise.

As it is recommended by von der Gracht [35], we also defined the stability between rounds as follows. Participants' responses to an item in 2 consecutive rounds were considered stable when the median of these responses failed to show a statistically significant difference between the rounds. We used the Wilcoxon matched-pairs signed rank test to assess the stability in these responses. This test is commonly used to assess the stability of responses in 2 consecutive rounds in Delphi studies. Following these criteria, we considered that participants' responses to an item in 2 consecutive rounds were stable when the results of the Wilcoxon matched-pairs signed rank test did not show a statistically significant difference and considered them unstable otherwise.

Finally, we defined the following stopping criteria for our Delphi process: (1) agreement was reached on all items and no new items were identified, (2) items for which agreement was not reached showed stability in 2 consecutive rounds, and (3) the panel size was reduced by >30% from the initial sample size.

Concluding and Reporting

We reported our study according to the Conducting and Reporting Delphi Studies guidelines (Multimedia Appendix 2) [36]. Although these were developed for consensus reporting in palliative care, they are relevant for other health care Delphi studies. The results tables include all statements and information on judgments, consensus, and stability.

Delphi Participants

The Delphi method uses a purposely selected panel of participants to provide their feedback on a subject. The sample need not be statistically representative, and therefore, there is no standard method to calculate the panel size. Some guidelines have suggested the involvement of 15 participants [37]. However, Delphi studies in the health care domain have often involved 20 to 30 participants [36], and panels for this research

XSL•FO
RenderX

typically have <50 participants. To achieve a gender-balanced sample size, we aimed to enroll 20 to 30 participants with representation from each of 6 continents (North America, South America, Europe, Africa, Asia, and Australia and Oceania). Specifically, we actively recruited health informaticians and researchers in the field of NLP in health care. We gathered members of the health informatics community from the International Medical Informatics Association Participatory Health and Social Media Working Group and the authors' peer networks. In addition, researchers from the health NLP community were identified from research papers on health NLP, specifically on generative pretrained transformers and LLMs in health care, and invited via email. To acknowledge participation in this study, we offered coauthorship as a group to participants who completed all 3 Delphi rounds.

This study was conducted between April 2023 and June 2023. The recruitment of participants for round 1 was conducted from April 10, 2023, to May 1, 2023. For rounds 2 and 3, participants had 2 weeks to respond, with a reminder sent after 1 week. Round 2 was available for completion from May 15, 2023, to May 28, 2023; round 3 was to be completed from June 12, 2023, to June 25, 2023. In round 3, an additional reminder was sent 1 day before the deadline.

Ethical Considerations

The study design was submitted to the ethics committee of the canton of Bern, which confirmed that no ethics approval was necessary (Req-2023-00427). Participants were invited to take part in the various rounds of the web-based Delphi study via email. Participation was voluntary, and they were informed that by submitting their responses to the form, they consented to their participation. Contact information for the corresponding author for any questions related to the study was provided both in the email and in the form header for each round.

Results

Characteristics of the Panel

An invitation was posted to the mailing list of the International Medical Informatics Association Participatory Health and Social Media Working Group, and 45 invitations were sent to individual researchers. The round-1 questionnaire had 28 respondents (Table 1). The round-2 questionnaire had 23 participants (return rate of 23/28, 82%), and the round-3 questionnaire had 21 participants (return rate of 21/23, 91%). Most of the participants in round 3 (18/21, 86%) had >10 years of experience in their field. Academia was the most represented sector (20/21, 95%). The panel engaged a diverse set of disciplines, with computer science or engineering being the most frequent work discipline followed by health informatics and medicine. The panel had participants from 3 continents: North America, Europe, and Australia and Oceania.

The panel reported a range of expertise in LLMs: 10% (2/21) were experts in LLMs, 29% (6/21) used their basic functions regularly, 29% (8/28) knew how they work, and 33% (7/21) had tested ChatGPT but had only basic knowledge of the underlying technology.

XSL•FO
RenderX

Table 1. Summary of participant characteristics.

Characteristics	Round 1 (n=28), n (%)	Round 2 (n=22), n (%)	Round 3 (n=21), n (%)
Gender			
Female	7 (25)	5 (23)	5 (24)
Male	21 (75)	17 (77)	16 (76)
Nonbinary	0 (0)	0 (0)	0 (0)
Education or background			
Computer science or engineering	13 (46)	11 (50)	10 (48)
Health informatics	11 (39)	9 (41)	9 (43)
Medicine	8 (29)	7 (32)	7 (33)
Nursing	1 (4)	1 (5)	1 (5)
Other health sciences	3 (11)	2 (9)	2 (10)
Other	2 (7)	1 (5)	1 (5)
Years of work experience			
<5	3 (11)	2 (9)	2 (10)
5-10	1 (4)	1 (5)	1 (5)
>10	24 (86)	19 (86)	18 (86)
Sector[a]			
Academia	26 (93)	21 (95)	20 (95)
Public health sector	5 (18)	4 (18)	4 (19)
Private health sector	2 (7)	0 (0)	0 (0)
Continent			
Europe	21 (75)	16 (73)	15 (71)
Australia and Oceania	3 (11)	3 (14)	3 (14)
North America	4 (14)	4 (18)	4 (19)
Level of experience with LLMs[b]			
I am an expert. I know and apply nearly all of their functions.	3 (11)	2 (9)	2 (10)
I use their basic functions regularly.	7 (25)	6 (27)	6 (29)
I know how they work.	8 (29)	7 (32)	6 (29)
I tested ChatGPT but have only basic knowledge of the underlying technology.	9 (32)	7 (32)	7 (33)
I have no knowledge.	1 (4)	0 (0)	0 (0)

[a]Some participants have several affiliations to different sectors.

[b]LLM: large language model.

Incorporation of New Items Based on Responses to the Open-Ended Questions

The responses to the open-ended questions in round 1 led to 65 new items for inclusion in rounds 2 and 3. A total of 4 new items were added in the round-3 questionnaire based on the responses to the open-ended questions. One statement was excluded in round 3 as it was already covered by another statement. In total, 7 items of round 2 were adapted according to the participants' input. Figure 2 presents the evolution of the questionnaire in each round. The complete list of times in rounds 2 and 3 are available in Multimedia Appendix 3. The evolution of agreement in rounds 2 and 3 are available in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 2. Data collection in the Delphi study. RR: response rate.

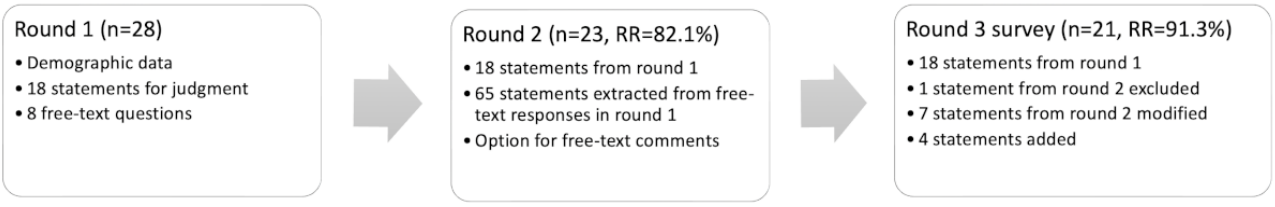

Perceived Likelihood That LLMs Will Support Health Care Tasks

The items referring to the perceived likelihood that LLMs will support health care tasks were grouped into 3 dimensions: support in clinical tasks, documentation tasks, and medical research and education. Table 2 presents the results obtained in the final round. Consensus was not reached for only 1 item, "Design of chemical compositions of new drugs," which belongs to the medical research and education dimension. However, only 57% (12/21) of the participants agreed that the likelihood of these systems supporting this task would be high (score of 4 or 5), with this being the item with the lowest percentage. A total of 95% (19/20) of the evaluated items were scored with a high likelihood of being supported by LLMs. Among all these tasks, all the participants agreed that LLM-based systems will act as virtual health assistants for patient education.

Table 2. Perceived likelihood that large language models will support health care tasks.

Item	Agreement and disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
Clinical tasks				
Virtual health assistant for patients (education)	Agreement	21 (100)	9 (43)	Yes
Automatic follow-up in chronic diseases	Agreement	20 (95)	11 (52)	Yes
Virtual health assistant for patients (medical assistance and information)	Agreement	20 (95)	8 (38)	Yes
Prediction of risk of disease development	Agreement	20 (95)	8 (38)	Yes
Virtual health assistant for patients (answering queries)	Agreement	19 (91)	14 (67)	Yes
Diagnostic process	Agreement	19 (91)	9 (43)	Yes
Patient triage	Agreement	18 (86)	7 (33)	Yes
Automatic treatment plan generation	Agreement	17 (81)	5 (24)	Yes
Verbalizing interactions	Agreement	16 (76)	5 (24)	Yes
Documentation tasks				
Automatic clinical encoding	Agreement	20 (95)	12 (57)	Yes
Virtual health assistants for administrative tasks	Agreement	20 (95)	9 (43)	Yes
Summarization	Agreement	19 (91)	7 (33)	Yes
Automatic structuring of clinical narratives	Agreement	19 (91)	7 (33)	Yes
Medical charting assistance	Agreement	18 (86)	8 (38)	Yes
Generation of layperson summaries	Agreement	17 (81)	5 (24)	Yes
Medical research and education				
Literature review and research	Agreement	19 (91)	10 (48)	Yes
Clinical trial matching	Agreement	19 (91)	8 (38)	Yes
Development of educational resources	Agreement	18 (86)	6 (29)	Yes
Automatic generation of guidelines	Agreement	16 (76)	5 (24)	Yes
Design of the chemical compositions of new drugs	Disagreement	12 (57)	3 (14)	N/A[a]

[a]N/A: not applicable.

Benefits of Using LLMs in Health Care

Consensus was not reached in only 1 of the items evaluated regarding the benefits of LLMs, and it showed stability between rounds 2 and 3. The participants agreed on 7 potential benefits of using LLM-based systems in health care (Table 3): "More efficient data handling and extraction," "Improved process automation," "Improved quality of health services,"

XSL•FO
RenderX

"Personalized care," "Improved health outcomes," "Faster diagnosis and treatment," and "Facilitated patient-professional interaction." Most (13/21, 62%) of the participants felt that these systems will not reduce health care costs.

Table 3. Participants' perceptions on the general benefits of using large language models in health care.

Item	Agreement or disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
More efficient data handling and extraction	Agreement	20 (95)	9 (43)	Yes
Improved process automation	Agreement	20 (95)	7 (33)	Yes
Improved quality of health services	Agreement	19 (91)	8 (38)	No
Personalized care	Agreement	19 (91)	4 (19)	Yes
Improved health outcomes	Agreement	17 (81)	4 (19)	Yes
Faster diagnosis and treatment	Agreement	17 (81)	3 (14)	Yes
Facilitated patient-professional interaction	Agreement	18 (76)	6 (29)	Yes
Improved clinical communication	Agreement	15 (71)	2 (10)	Yes
Increased caregiver empowerment	Agreement	15 (71)	1 (5)	Yes
Reduced workload for health care professionals	Disagreement	15 (71)	6 (29)	Yes
Resource optimization	Agreement	15 (71)	4 (19)	No
Reduction of human errors	Agreement	15 (71)	1 (5)	Yes
Improved interoperability	Agreement	14 (67)	2 (10)	N/A[a]
Reduced health care costs	Agreement	8 (38)	1 (5)	Yes

[a]N/A: not applicable.

Shortcomings and Risks of LLM-Based Systems in Health Care

The shortcomings of using LLM-based systems in health care were grouped into 5 dimensions (Table 4): "Risks to health care," "Risks to the medical profession," "Risks to patients," "Risks related to data protection," and "Risks to the health IT field." Agreement was reached in almost two-thirds (29/50, 58%) of the evaluated items. All items except the "accessibility issues" item showed stability in round 3.

XSL•FO
RenderX

Table 4. Shortcomings and risks of large language model (LLM)–based systems in health care.

Item	Agreement or disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
Risks to health care				
Cybersecurity risks	Agreement	18 (86)	10 (48)	Yes
Risk of misinformation of patients	Agreement	18 (86)	9 (43)	Yes
Ethical risks	Agreement	17 (81)	14 (67)	Yes
Risk of biased decisions	Agreement	17 (81)	6 (29)	Yes
Risk of inaccurate communication	Agreement	16 (76)	7 (33)	Yes
Lack of explainability of system decision-making processes	Disagreement	15 (71)	9 (43)	Yes
Risk of increasing health inequities	Disagreement	14 (67)	6 (29)	Yes
Limited interoperability of generated outputs	Agreement	13 (62)	4 (19)	Yes
Risk of dehumanization of care	Disagreement	13 (62)	2 (10)	Yes
Risk of errors (HCPs[a])	Agreement	12 (57)	3 (14)	Yes
Negative clinical outcomes	Disagreement	12 (57)	4 (19)	Yes
Risk of information overload of patients	Disagreement	10 (48)	3 (14)	Yes
Risk of information overload of HCPs	Disagreement	9 (43)	3 (14)	Yes
Risks to the medical profession				
Overconfidence in LLM-based models	Agreement	16 (76)	5 (24)	Yes
Impact on jobs in the health care sector	Disagreement	15 (71)	4 (19)	Yes
Misdiagnosis due to wrong generated results	Agreement	14 (67)	3 (14)	Yes
Liability for errors made by LLM-based systems	Agreement	13 (62)	4 (19)	Yes
Lack of understanding of the underlying technology	Disagreement	13 (62)	2 (10)	Yes
Risk of losing knowledge and competencies	Disagreement	9 (43)	2 (10)	Yes
Risk of attempts to replace health care practitioners with tools	Disagreement	8 (38)	2 (10)	Yes
Loss of communication skills	Disagreement	6 (29)	2 (10)	Yes
Loss of trust of patients in HCPs	Disagreement	6 (29)	1 (5)	Yes
Reduced need for medical professionals	Disagreement	2 (10)	0 (0)	Yes
Risks to patients				
Risk of inaccurate communication	Agreement	15 (71)	2 (10)	Yes
Lack of transparency of system use	Disagreement	13 (62)	3 (14)	Yes
Wrong personal health decisions due to the use of unverified information	Agreement	13 (62)	1 (5)	Yes
Accessibility issues	Disagreement	11 (52)	4 (19)	No
Loss of patient-professional contact	Disagreement	11 (52)	3 (14)	Yes
Incorrect treatment plans	Disagreement	11 (52)	1 (5)	Yes
Incorrect diagnoses	Disagreement	10 (48)	2 (10)	Yes
Loss of trust in HCPs	Disagreement	6 (29)	1 (5)	Yes
Risks related to data protection				
Use of unregulated cloud services may risk data security and privacy	Agreement	18 (86)	7 (33)	Yes
Disclosure of sensitive patient data during training and inference	Agreement	18 (86)	5 (24)	Yes
Breach of patient confidentiality	Agreement	17 (81)	4 (19)	Yes

XSL•FO
RenderX

Item	Agreement or disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
Fraudulent use of information	Agreement	16 (76)	6 (29)	Yes
Vulnerabilities in data storage systems or communication channels	Agreement	16 (76)	5 (24)	Yes
Risk that individual patient data may be accessed or used inappropriately	Agreement	16 (76)	5 (24)	Yes
Breach of GDPR[b]	Agreement	15 (71)	4 (19)	Yes
Risk of patient reidentification	Agreement	15 (71)	4 (19)	Yes
Uncontrolled access by third parties	Disagreement	14 (67)	8 (38)	Yes
Risks to the health IT field				
Unresolved responsibilities for system errors or wrong outputs hamper adoption of LLM-based systems	Agreement	18 (86)	6 (29)	Yes
Developing and delivering solutions compliant with regulations is complex for health IT companies	Agreement	17 (81)	5 (24)	Yes
Competitive pressure leads to market release of LLM-based systems of low quality	Agreement	17 (81)	5 (24)	Yes
Lack of understanding of clinical risks leads to systems that can harm patients	Agreement	16 (76)	4 (19)	Yes
Financial constraints at health care institutions for maintenance of LLM-based systems will hamper the adoption of high-quality systems	Agreement	15 (71)	5 (24)	Yes
Lack of skilled workers for developing LLM-based systems will hamper the development of high-quality systems	Agreement	15 (71)	3 (14)	Yes
A missing standard quality assessment framework for LLM-based systems will lead to low-quality systems released to market	Agreement	14 (67)	5 (24)	Yes
LLM-based systems will lack integration into clinical systems	Agreement	14 (66.7)	1 (5)	Yes
Missing reimbursement models for LLM-based systems hamper the adoption of technology	Disagreement	10 (48)	6 (29)	Yes
Companies' lack of competence to ensure the development of systems compliant with regulations	Agreement	10 (48)	2 (10)	Yes

[a]HCP: health care professional.

[b]GDPR: General Data Protection Regulation.

Regarding the risks to health care, agreement was reached on 54% (7/13) of the evaluated items. The experts agreed that the use of LLM-based systems can lead to 5 of the evaluated risks: "Cybersecurity risks," "Risk of misinformation of patients," "Ethical risks," "Risks of biased decisions," and "Risks of inaccurate communication." On the other hand, the experts did not believe that the use of these systems could provoke information overload in health care professionals (HCPs) or patients.

Agreement was reached on 30% (3/10) of the evaluated risks to the medical profession. Only 1 of the evaluated risks to the medical profession, "Overconfidence in LLM-based systems," was recognized by most of the experts. The experts did not believe that the use of these systems will lead to a reduced need for medical professionals.

Agreement was reached on 25% (2/8) of the risks to patients ("Risk of inaccurate communication" and "Wrong personal health decisions due to the use of unverified information"). None of the risks were scored with a 4 or 5 by >75% of the experts.

Regarding the risks related to data protection, only in 1 item, "Uncontrolled access by third parties," agreement was not reached. The experts believed that LLM-based systems could lead to 6 risks related to data protection: "Use of unregulated cloud services may risk data security and privacy," "Disclosure of sensitive patient data during training and inference," "Breach of patient confidentiality," "Fraudulent use of information," "Vulnerabilities in data storage systems or communication channels," and "Risk of individual patient data may be accessed or used inappropriately."

XSL•FO
RenderX

Agreement was not reached in only 10% (1/10) of the risks to the health IT field. The experts agreed that LLM-based systems can lead to 4 risks in this dimension: "Unresolved responsibilities for system error or wrong outputs hamper adoption of LLM-based systems," "Developing and delivering solutions compliant with regulations is complex for health IT companies," "Competitive pressure leads to market release of LLM-based systems of low quality," and "Lack of understanding of clinical risks leads to systems that can harm patients."

Needs for Future Adoption and Implementation of High-Quality LLM-Based Systems

In all items related to future adoption and implementation of high-quality LLM-based systems included in this section, agreement and stability were reached in round 3 (Table 5). All items were considered relevant by most of the experts. All experts agreed that successful adoption of LLM-based systems in practice requires training of HCPs and quality assessment standards.

Table 5. Experts' agreement on requirements for the successful adoption of large language model (LLM)–based systems in the health care domain.

Item	Agreement or disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
Successful adoption in practice requires training of HCPs[a]	Agreement	21 (100)	15 (71)	Yes
Successful adoption in practice requires quality assessment standards	Agreement	21 (100)	15 (71)	Yes
Successful adoption in practice requires regulations on data privacy for such systems	Agreement	20 (95)	15 (71)	Yes
Successful adoption in practice requires proper standards for data security and data privacy	Agreement	20 (95)	16 (76)	Yes
Successful adoption in practice requires algorithm or vigilance	Agreement	20 (95)	14 (67)	Yes
Successful adoption in practice requires training of health IT personnel	Agreement	20 (95.2)	12 (57)	Yes
Successful adoption of LLM-based systems supporting the decision-making process in practice requires co-design of new workflows with HCPs	Agreement	20 (95)	11 (52)	Yes
Successful adoption in practice requires regulations on data ownership	Agreement	18 (86)	14 (67)	Yes
Successful adoption in practice requires guidelines for interpretation of the results of LLM-based systems and their use in clinical practice	Agreement	18 (86)	14 (67)	Yes
Successful adoption in practice requires a cultural change in health care	Agreement	17 (81)	12 (57)	No
Successful adoption in practice requires integration with existing EHRs[b] if the LLM-based system is to support the decision-making process	Agreement	17 (81)	11 (52)	Yes
Successful adoption in practice requires the adaptation of jobs in the health care domain	Agreement	17 (81)	6 (29)	Yes
Successful adoption in practice requires reimbursement models for LLM-based systems and their use in health care	Agreement	16 (76)	5 (24)	Yes

[a]HCP: health care professional.

[b]EHR: electronic health record.

Reliability of Systems Based on LLMs

Agreement was reached on all the characteristics evaluated regarding reliability, and they were considered relevant by the experts (Table 6). All participants agreed on half of the characteristics: "The system is tested in real settings," "The system outputs are reproducible," "The system outputs are reliable," "The system is robust against a wide range of inputs," "Quality of the data underlying the system is ensured," "The system is tested in simulated settings with real users," and "The system is validated for accuracy."

XSL•FO
RenderX

Table 6. Requirements for reliable systems based on large language models (LLMs).

Item	Agreement or disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
The system is tested in real settings	Agreement	21 (100)	17 (81)	Yes
The system outputs are reproducible	Agreement	21 (100)	15 (71)	Yes
The system outputs are reliable	Agreement	21 (100)	14 (67)	Yes
The system is robust against a wide range of inputs	Agreement	21 (100)	13 (62)	Yes
Quality of the data underlying the system is ensured	Agreement	21 (100)	12 (57)	Yes
The system is tested in simulated settings with real users	Agreement	21 (100)	11 (52)	Yes
The system is validated for accuracy	Agreement	21 (100)	10 (48)	Yes
The system meets federal regulations	Agreement	20 (95)	15 (71)	Yes
The system is interoperable with existing health care systems	Agreement	20 (95)	11 (52)	Yes
Control mechanisms or human-in-the-loop processes are established to ensure reliability of LLM-based systems	Agreement	19 (91)	11 (52)	Yes
A standardized quality assessment is available for the system	Agreement	19 (91)	11 (52)	Yes
The system has been proven to be noninferior in a variety of clinical settings	Agreement	19 (91)	7 (33)	Yes
Explanations of the reasoning behind model predictions and recommendations are available	Agreement	19 (91)	6 (29)	Yes
The system can solve easy routine tasks with nearly 100% accuracy	Agreement	17 (81)	7 (33)	Yes

Future of LLMs

All statements on the future of LLMs reached stability in round 3 (Table 7). Agreement was not reached on 3 of these statements. Most of the experts agreed on 5 of the statements: “LLMs will be combined with other technologies in future health applications,” “Applications based on LLMs will be used by HCPs,” “LLMs will have an impact on future technologies in healthcare,” “Applications based on LLMs will be used by patients,” and “LLMs will replace other technologies.” Results from round 3 are shown in Multimedia Appendix 5 in more detail.

XSL•FO
RenderX

Table 7. Experts' opinions on statements related to the future of large language model (LLM)–based systems in health care.

Item	Agreement or disagreement	Scores of 4 or 5 (n=21), n (%)	Scores of 5 (n=21), n (%)	Stability
LLMs will be combined with other technologies in future health applications	Agreement	21 (100)	10 (48)	Yes
Applications based on LLMs will be used by health care professionals	Agreement	21 (100)	2 (10)	Yes
LLMs will have an impact on future technologies in health care	Agreement	20 (95)	5 (24)	Yes
Applications based on LLMs will be used by patients	Agreement	18 (86)	7 (33)	Yes
LLMs will replace other technologies	Agreement	18 (86)	2 (10)	Yes
The medical device regulation hampers the introduction of solutions based on LLMs	Disagreement	14 (67)	1 (5)	Yes
Solutions based on LLMs will help address the shortage of skilled health professionals	Agreement	14 (67)	0 (0)	Yes
To what extent will future health care rely on LLM-based solutions?	Agreement	12 (57)	0 (0)	Yes
LLM-based solutions will contribute to discrimination in health care because they rely on biased data	Agreement	11 (52)	1 (5)	Yes
I consider LLMs, specifically their resource consumption, ecologically sustainable	Disagreement	7 (33)	1 (5)	Yes
Students of medicine will lose competencies through the increased use of LLMs	Disagreement	7 (33)	1 (5)	Yes
The introduction of LLMs in digital health solutions will result in cost savings in the health sector	Agreement	6 (29)	1 (5)	Yes
LLMs will be replaced by other technologies in the coming 5 years	Agreement	6 (29)	0 (0)	Yes
Health care professionals (physicians and nurses) will lose competencies through the increased use of LLMs	Agreement	5 (24)	1 (5)	Yes
Solutions based on LLMs will offend the sensibilities of health care professionals	Agreement	5 (24)	0 (0)	Yes
Patients will lose competencies through the increased use of LLMs	Agreement	4 (19)	0 (0)	Yes
Solutions based on LLMs will offend the sensibilities of patients	Agreement	2 (10)	0 (0)	Yes
Solutions based on LLMs will offend the sensibilities of other people involved in the care process	Agreement	2 (10)	0 (0)	Yes

Discussion

Principal Findings

Our study describes the views of sector experts on the future potential benefits and risks of LLMs in health care as well as requirements for adoption. The results show that LLMs are expected to be used within virtual health assistants, for helping with various tasks, and for supporting patient education. Key benefits identified included improved data handling, process automation, service quality, personalized care, and faster diagnosis. However, experts also warned of potential risks, such as cybersecurity threats, misinformation, ethical concerns, and decision bias. Key privacy risks included potential breaches of confidentiality and data storage vulnerabilities. This study also highlights the complexity of regulatory compliance and the risk of low-quality system releases due to market pressures. To effectively integrate LLM-based systems into health care, there is a consensus on the need to train HCPs and set quality standards to ensure a balanced approach between reaping the benefits and managing the risks.

In the following sections, we consider the points on which there was agreement in answering the questions that drove our SWOT analysis of LLM-based systems and contextualize the findings with previous work. Implications for practitioners as well as researchers are presented.

Comparison to Prior Work

Our study describes many use cases and tasks in which systems based on LLMs can be used to support HCPs and patients. These fall into 3 groups: support in clinical tasks, support in documentation tasks, and support in medical research and education.

XSL•FO
RenderX

Support in Clinical Tasks

Clinical tasks concern patient-professional interaction. In this context, virtual assistants based on LLMs could be useful in several cases, such as providing medical assistance and information to patients, answering their questions, or conducting patient education. This is in line with the research published on intelligent agents and systems with conversational user interfaces. Although many of the existing agents described are rule based [38], reports on LLM-based systems are emerging [39]. Other clinical tasks with which LLMs could help include triage and diagnostic tasks as well as risk assessment and treatment plan generation. As exemplified, LLMs can be used for predicting the occurrence of chronic diseases in a patient based on information from clinical notes [40].

Support in Documentation Tasks

Documentation tasks that experts expect to be supported in the future by LLM-based systems include automatic clinical coding, summarization of clinical documents, automatic structuring of clinical narratives, and medical charting assistance. Again, there is a nascent literature describing these uses. Yang et al [12] presented an LLM-based approach to clinical concept extraction. López-García et al [41] analyzed model performance of LLMs for automatic clinical coding in Spanish. Lentzen et al [5] studied LLM accuracy for automatic structuring of clinical notes in German. LLMs have demonstrated significant performance gains for medical problem summarization tasks [42]. We note preliminary reports of the use of LLMs to generate radiological reports from images [43]. These examples demonstrate the potential for use cases for medical charting assistance.

Support in Medical Research and Education

LLMs could support literature review and research, clinical trial matching, guideline generation, and educational resource development. Tian et al [44] successfully tested LLM-based system named entity recognition for parsing clinical trial eligibility criteria.

Other Strengths

Experts agreed that LLMs can lead to improved quality of care, better health outcomes, optimized clinical communication, reduced human error, personalized care, and increased caregiver empowerment. These use cases have preliminary evidence in the literature. However, the benefits are still hypothetical, as LLM-based systems have not yet been implemented in daily practice solutions. It remains to be proven whether LLM-based systems can significantly improve digitized processes in health care or reduce costs.

There are potential risks with the use of LLMs to health care in general, to the medical profession, to patients, and to privacy rights.

Risks to the Medical Profession

A potential risk to the medical profession is overreliance on automated systems. HCPs may place too much trust in the results generated by LLMs, leading to potential complacency and reliance on technology over clinical judgment [45]. Of concern is the potential for misdiagnosis due to incorrect results produced by LLMs if results are not carefully validated and cross-checked by HCPs.

Human factors may lead to errors such as incorrect use or interpretation of LLM outputs by HCPs. There is a risk that HCPs may make mistakes when relying on information generated by an LLM-based system, potentially affecting patient care and safety. This raises the issue of liability in the event of system failure, which can become a complex legal issue [46]. Possible implications and unintended consequences of LLM-based systems must be considered now before systems are used in practice.

In addition, increased reliance on LLM-based systems could lead to a loss of skills among HCPs. As they become more accustomed to using automated outputs, they may rely less on their own knowledge and skills, which could affect their clinical decision-making. Reliance on LLM-based systems during medical education could result in trainees never acquiring knowledge gained by previous trainee cohorts. Finally, the introduction of LLM-based systems in health care could potentially offend the sensibilities of some HCPs. Some professionals may feel uncomfortable or threatened by the move toward greater automation as it may seem to devalue the human aspect of health care.

Risks to Patients

One significant risk to patients is the potential for inaccurate information sharing with them. LLMs may produce outputs that are difficult for patients to understand, leading to misunderstandings or incomplete information exchange during critical health care interactions. Another important concern is the risk of people making poor personal health decisions based on unverified or inaccurate information from LLM outputs. People may misinterpret the information or act on it without proper validation, which may result in making decisions that do not align with their goals for care, preferences, or values. In addition, the introduction of LLMs in health care could offend the sensibilities of some patients. They may feel uncomfortable or uneasy with the idea of their diagnosis or treatment decisions relying on automated technology, potentially leading to a sense of detachment, a perceived lack of personalized care, or disengagement with the health care system. Patients in health care systems that require significant patient cost sharing may be at risk of higher copayments if the costs of implementing transforming technologies are passed on.

Risks Related to Data Protection

Cloud services are often used for the training of LLMs. The use of unregulated cloud services introduces potential vulnerabilities, with data security and privacy at risk. During the training and inference processes, sensitive patient data may be inadvertently exposed, possibly compromising patient confidentiality. One of the most significant threats is the potential for fraudulent or illegal use of patient information, which could lead to patient reidentification. If patient data are not adequately protected, identity theft, denial of care, or other malicious activities may result.

Vulnerabilities in data storage systems or communication channels increase the risk of data breaches [47]. Cyberattacks

XSL•FO
RenderX

or data leaks through insecure channels could compromise patient information, particularly in light of increasing system complexity and increasing number of features. If not adequately protected following international standards such as the International Organization for Standardization and International Electrotechnical Commission 27000 series, patient data may be accessed or used for purposes other than the intended medical care and clinic operations. In addition, the implementation of LLMs in health care must comply with data protection regulations such as the General Data Protection Regulation. Exploiting vulnerabilities could also threaten patient safety. In this context, some LLM-based applications can be considered safety-critical systems because they store and process information that is needed for patient care (eg, medication data) [47].

General Risks

Ethical considerations are another area of concern as LLMs could inadvertently lead to biased decisions based on inadequately representative training data. This could lead to unfair or discriminatory outcomes for some patient populations. The reliance of LLMs on biased data could perpetuate and amplify existing inequalities in health care [48,49]. Furthermore, the implementation of LLMs in health care settings may raise ethical dilemmas and offend certain individuals involved in the care process. Greco et al [50] claimed that there is still only "little discussion...provided to avoid or mitigate the bias of transformed-based models, in terms of different aspects of the users/patients, such as culture, age and gender."

Reliability of Systems

We identified a number of issues relevant to LLM-based systems' reliability. First, careful, standardized quality testing is essential. LLM-based systems should be shown to be noninferior in a variety of clinical settings. First attempts regarding reporting guidelines for AI-based clinical decision support systems have been made [51]. Second, LLM-based systems should produce reproducible and reliable outputs and be robust to a wide range of inputs. Third, control and reasoning mechanisms need to be established for LLM-based systems' implementation in practice to ensure reliability and explainability. This point was also confirmed by Greco et al [50] in their survey on LLMs for mental health issues. Kelly et al [52] found that for some scenarios, AI cannot replace or replicate human contact. Fourth, the quality of the underlying data must be ensured. Finally, LLM-based systems must comply with federal regulations.

In summary, there are several disadvantages or challenges of applying LLMs in health care. Reliability of the systems in the real world still has to be proven.

Opportunities Contributing LLM Use in Practice

Our panel of experts identified several opportunities that could contribute to the successful implementation of LLMs in health care. LLM-based solutions could free up capacity, mitigating the effects of the current shortage of skilled HCPs. Our experts agreed that LLMs have the potential to increase the efficiency of clinical processes and improve quality. Clinical processes can be optimized through more efficient data handling and extraction using LLMs, process automation, resource optimization, and improved interoperability. LLMs can help make faster diagnoses and reduce time needed for patient education.

Work on standardized data exchange (eg, using Health Level 7 and Fast Healthcare Interoperability Resources) is progressing in health care, which will enable the integration of LLM outputs into existing health IT systems. The experts agreed that LLMs will be combined with other technologies in future health care applications and may therefore influence future health care technologies. Furthermore, some other technologies might be replaced by LLM-based solutions.

Threats to Adoption of LLM in Practice

One set of threats is related to IT companies that develop products integrating LLMs. A lack of understanding of clinical risks potentially caused by LLMs may result in systems that have the potential to harm patients. A shortage of system development skills could hinder the development of high-quality systems. Furthermore, companies may lack the expertise and competence required to ensure the development of LLM-based systems that are fully compliant with health care regulations and that function as intended. These regulatory challenges could be barriers to the successful implementation of LLMs in health care settings.

Kokol et al [53] have previously raised concerns about the quality of digital health solutions because "neither the volume, distribution nor scope of the quality research content related to digital health solutions is satisfactory." They assert that there is a risk of reducing the quality of care due to subpar software and software-based tools.

Another set of threats is related to health care institutions, workflow integration, and maintenance of LLM-based systems. Financial constraints may hinder the adoption of high-quality LLM-based systems as maintenance costs may become prohibitive. A similar observation was made by Sezgin et al [54] in their work on using GPT-3 in the US health care system. Accordingly, there is a risk that LLM-based systems may not be properly integrated with existing clinical systems, leading to inefficiencies, suboptimal performance, and perhaps harm to patients.

We identified several factors that should be considered for a successful adoption of LLM-based systems in health care. At a minimum, HCPs and health IT personnel must be trained appropriately to use LLM-based systems. Without appropriate, adequate, and sufficient training, LLM implementation teams and the HCPs they serve cannot hope to use LLM-based systems.

A second aspect is quality and data security. Quality assessment standards are needed for LLM solutions, and data privacy and ownership regulations need to be considered or developed when none exist. Also required are appropriate standards for data security and privacy [55]. Monitoring the output of algorithms through algorithm or vigilance [56] is essential to ensure patient safety. Unresolved accountability for system failures or incorrect outcomes may hinder the widespread adoption of these technologies. Langlais et al [45] considered AI applications in

XSL•FO
RenderX

cardiology and called for a framework for accountability in cases of system error.

A third factor is integration into the clinical workflow. Integration with existing health IT systems, including electronic health records, is a complex and unavoidable challenge for a successful LLM implementation. Otherwise, seamless communication and coordination may be hindered. A culture change in health care and reimbursement models are needed to ensure success. The Digital Health Validitron offered by researchers of the University of Melbourne is one of the first attempts to ensure that digital health solutions are successfully designed, developed, validated, and evaluated [57]. It enables secure co-design, testing, and refinement of ideas for digital health solutions in a laboratory environment in preparation for implementation. Such a platform could also be useful in the current stage of evolving LLM-based solutions in health care.

Strengths and Limitations of This Study

Although this Delphi study involved experienced participants, the panel was not widely representative. We acknowledged participation in the Delphi panel by offering coauthorship as a group. This could have attracted participation of persons who are not very experienced with LLMs. Furthermore, participants were mainly from computer science, health informatics, and medicine, and almost all (20/21, 95%) worked in academia. Most participants (15/21, 71%) were from Europe, with only one-third from other regions (ie, Australia or Oceania 3/21, 14% in round 3 and North America 4/21, 19% in round 3). European countries have national health care systems and strong privacy regulations for the management of citizens' personal data. Responses from participants from regions under other circumstances could be rendered insignificant due to the demographic mixture of the respondents. Regulatory, economic, and infrastructural concerns affecting LLM adoption and use may not be reflected. The participants may have had expertise in health informatics or related fields, but only one-third of participants (7/21, 33%) had experience on LLMs limited to testing ChatGPT without having a comprehensive knowledge of LLMs. This could of course affect the reliability of the judgments. The selection of the participants was biased in that we contacted persons on an individual basis based on their publication record and from an industry working group.

The number of participants decreased over 3 rounds but remained within the 75% requirement for participation—stability was not achieved for all items even though their number was small. Future research should investigate whether the items that did not show stability could do so in research involving more rounds or a different group of experts. The design of our study (answer collection only through web-based surveys) meant that participants had no interaction or discussions. Such interactions might have been useful only to clarify the formulations and harmonize the viewpoints. Although the final panel comprised 21 participants, it is not possible to affirm the completeness of the SWOT analysis. We might have missed aspects or aggregated them with other items when they should have been considered separately. In addition, a significant presence of patients or citizens who regularly interact with the health care system might yield different perspectives.

In round 1, we collected a large number of use cases for LLMs. We removed some of them from subsequent rounds as they were not specific to health care. For example, one expert suggested "Building of socio-sanitary services: processing healthcare data with social data (from e.g., the city hall, the public water/gas companies or the treasury department, among others) could be helpful to address specific needs of population at-risk." Integrating health care data and citizen science is clearly interesting. However, we focused rather on clinical applications. In this way, we might have excluded interesting and relevant cases.

Some participants also commented in their free textual comments in rounds 2 and 3 that their judgments depended on whether LLMs are used to replace HCPs or assist them. Our statements were not formulated clearly enough to capture this distinction and might have led to misunderstanding or bias.

The items on which the participants disagreed over 3 rounds are related to five topics: (1) impact on jobs ("Reduced workload for healthcare professionals," "Impact on jobs in the healthcare sector," "Risk of attempts to replace healthcare practitioners," and "Reduced need for medical professionals"), (2) patient-physician relationship ("Risk of dehumanization of care," "Risk of information overload of patients," "Risk of information overload of HCPs," "Loss of trust of patients in HCPs," "Loss of patient-professional contact," and "Loss of trust in HCPs"), (3) quality and transparency of results ("Lack of explainability of system," "decision making processes," "Negative clinical outcomes," "Lack of transparency of use," "Incorrect treatment plans," and "Incorrect diagnoses"), (4) accessibility and equity ("Risk of increasing health inequities" and "Accessibility issues"), and (5) skills ("Lack of understanding of the underlying technology," "Risk of losing knowledge and competencies," and "Loss of communication skills").

The disagreements could be related to the different viewpoints and experiences of the participants or a different understanding of how transformers will be integrated into health care processes. For example, the impact on jobs in the health care sector might be judged differently when we envision scenarios in which transformers are used with or without a human in the loop. In addition, personal fears of potentially losing their job could be relevant when judging the items related to future jobs. Variations in how transformer-based technologies are regulated in participants' countries may have affected their experiences with transformers at the time of the study, thereby influencing their opinions and expectations of the technology.

Practical Implications

On the basis of our findings, we describe key implications for practitioners. These are intended to create awareness regarding the successful and efficient adoption, development, and deployment of LLMs in the health care ecosystem and to optimally benefit from their existing advantages for practice.

Clinical Deployment and Training

Overall, the effectiveness of LLMs in clinical settings relies on the skills and familiarity of the HCPs using LLMs. In this context, it is necessary to improve HCPs' skills to ensure proper

XSL•FO
RenderX

and successful integration into (existing) clinical workflows (eg, for supporting clinical documentation). Moreover, we recommend designing and implementing LLM-based applications in human-in-the-loop processes (ie, involving HCPs in the development process in a collaborative feedback loop), especially for validation tasks. In this way, HCPs will become familiar with the technology more quickly and build the necessary trust in it and its decision-making process.

Quality Assessment

The development of standardized quality assessment frameworks for LLM-based systems is essential to ensure the release and adoption only of systems that have achieved a minimum standard of quality. Thus, regulators in collaboration with HCPs should establish guidelines, standards, and benchmarks ensuring patient safety, secure data processing and storing, and privacy while supporting the technology's innovation potential, and developers should build products that meet these standards. Developers should not only validate the accuracy of the results with HCP oversight but also perform simulations under real conditions, including configurations of the underlying systems. Testing of such applications in diverse scenarios is required for successful implementation and deployment, especially to address regulatory concerns and build HCP and patient trust. LLM-based systems are expected to be able to complete routine tasks in some cases with nearly 100% accuracy.

Data Security and Privacy

It is crucial to achieve compliance with well-established data protection regulations such as the General Data Protection Regulation, the Health Insurance Portability and Accountability Act, Genetic Information Nondiscrimination Act, and data security standards such as the International Organization for Standardization and International Electrotechnical Commission 27000 series to ensure secure LLM-based applications. This is critical in health care to avoid increased risks to patients, including but not limited to physical, psychological, legal, economic, and reputational harm as well as discrimination and loss of access to health care services.

In addition, there are multiple potential threats and risks associated with the use of third-party services (eg, cloud services) in LLM training. Consequently, reliable and robust data security measures as well as comprehensive security policies are needed to address the security goals of confidentiality, integrity, and secure availability. Depending on the application, the 3 goals of information security—authorization, accountability, and nonrepudiation—should also be met.

Ethical Considerations

Biases and potentially discriminatory outcomes, which are typically not only based on biased training data but also influenced by human perceptions and activities, pose a threat to health care if not anticipated and avoided. Ethical guidelines need to be developed to mitigate these risks, particularly in the context of reliable and ethical outcomes. To address these and other risks related to LLMs, Porsdam Mann et al [58] promoted transparency and engagement in open discussions that will allow LLM developers to demonstrate their commitment to and practice of responsible and ethical practices.

Research Implications

Our Delphi process articulated a number of future research considerations.

Clinical Deployment and Workflow Integration

Clinical workflows can be highly complex, making the successful integration of LLMs quite challenging. Interdisciplinary research approaches are needed to address this concern. Existing workflows can be extended, or new workflows can be created that encompass a user-centered, efficiency-oriented design of LLM-based applications. In each case, real-world evaluations are needed to investigate their actual efficiency and reliability and any potential hurdles such as their functional acceptance by HCPs and patients. The long-term impact of integrating LLMs into clinical workflows should be measured in the context of HCP skills, patient engagement and satisfaction, and overall health care and process quality. It will be important to identify the appropriate outcomes to measure and determine how best to assess them both immediately after implementation and throughout the product life span.

Quality Improvement

One of the major issues of LLMs is biased data and outcomes. Techniques that can ensure debiased and equal training are needed. Moreover, we highlight the need for controlled environments for validation and refinement purposes, such as the Digital Health Validitron. Similar platforms and frameworks could effectively facilitate the secure development and evaluation of LLMs for health care. Note that the real-world operational quality of LLMs should be evaluated in different clinical settings and patient populations to successfully test possible application configurations as well as user perceptions. Furthermore, cost-benefit analyses based on comprehensive evaluations including economic impacts are also necessary in this context. In summary, the impact of LLMs on cost savings and environmental sustainability needs further evaluation.

Ensuring Explainability

Due to the increasing spread and use of AI applications, it is important to ensure transparent, accurate, and interpretable explanation outcomes. This requirement also applies to LLMs even more in health care given the need to ensure patient safety. By making the entire decision-making process interpretable, concerns related to so-called black box AI can be addressed. Thus, we emphasize the need for techniques or even frameworks to achieve explainable LLMs to improve trust but also HCPs' and patients' understanding.

Conclusions

In this paper, we reported expert agreed opinions regarding the SWOT of LLM-based systems in health care. Many use cases we collected have yet to come to fruition. However, it is a work in progress in terms of research and development of LLMs for many tasks in the health care domain. There are substantial threats to the successful implementation of LLMs, which include the quality and quality assessment, regulatory aspects, and integration with workflows. Research in these areas could

XSL•FO
RenderX

contribute to the acceptance in real-world settings and reliability of LLM-based products in health care. We conclude by recommending that research should not only focus on testing the possibilities of LLMs for natural language–related tasks but also consider the workflows that these models could contribute to and the requirements regarding quality, integration, and regulations that are necessary for a successful implementation in practice. With this approach, it will be possible to generate a real impact in health care.

Acknowledgments

The authors acknowledge the participation of the LLMHealthGroup in this Delphi panel. LLMHealthGroup members: Antonio de Arriba-Muñoz, Wendy Chapman, James CL Chow, Shauna Davies, Rebecca Grainger, Boris V Janssen, Shaoxiong Ji, Markus Kreuzthaler, Augustin Lecler, Chris Paton, Carolyn Petersen, Juan Ramón Lacalle, Denis Remedios, Jorge Ropero, Jose L Sevillano, Emre Sezgin, Vicente Traver, Jesús Daniel Trigo, and Karin Verspoor. This study did not receive third-party funding. Generative artificial intelligence (Open AI ChatGPT and DeepL) was used to a limited extent to reformulate single sentences.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

RM, KD, and ORR contributed to conceptualization. ORR was responsible for data curation and formal analysis of the data. KD and RM developed the methodology. KD, ORR, and RM were writing the original draft. The LLMHealthGroup was reviewing the paper draft. All authors have read and approved the current version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Items of the first-round questionnaire.
[DOCX File , 17 KB - jmir_v26i1e52399_app1.docx]

Multimedia Appendix 2
Summary of Conducting and Reporting Delphi Studies recommendations.
[DOCX File , 16 KB - jmir_v26i1e52399_app2.docx]

Multimedia Appendix 3
New items and rewording of the round-2 and round-3 questionnaires.
[DOCX File , 21 KB - jmir_v26i1e52399_app3.docx]

Multimedia Appendix 4
Evolution of agreement across rounds.
[DOCX File , 40 KB - jmir_v26i1e52399_app4.docx]

Multimedia Appendix 5
Percentage of participants assigning each score in round 3.
[DOCX File , 2057 KB - jmir_v26i1e52399_app5.docx]

References

1. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135 [FREE Full text] [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
2. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DS, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023 Jul 24;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61]
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA URL: https://dl.acm.org/doi/10.5555/3295222.3295349

XSL•FO
RenderX

4. Lentzen M, Linden T, Veeranki S, Madan S, Kramer D, Leodolter W, et al. A transformer-based model trained on large scale claims data for prediction of severe COVID-19 disease progression. IEEE J Biomed Health Inform 2023 Sep;27(9):4548-4558. [doi: 10.1109/JBHI.2023.3288768] [Medline: 37347632]
5. Lentzen M, Madan S, Lage-Rupprecht V, Kühnel L, Fluck J, Jacobs M, et al. Critical assessment of transformer-based AI models for German clinical notes. JAMIA Open 2022 Nov 15;5(4):ooac087 [FREE Full text] [doi: 10.1093/jamiaopen/ooac087] [Medline: 36380848]
6. Gillioz A, Casas J, Mugellini E, Khaled OA. Overview of the transformer-based models for NLP tasks. In: Proceedings of the 2020 Federated Conference on Computer Science and Information Systems. 2020 Presented at: FedCSIS 2020; September 6-9, 2020; Sofia, Bulgaria. [doi: 10.15439/2020F20]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Wang W, Yang Z, Gao Y, Ney H. Transformer-based direct hidden markov model for machine translation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop. 2021 Presented at: ACL-IJCNLP 2021; August 5-6, 2021; Virtual Event. [doi: 10.18653/v1/2021.acl-srw.3]
9. Liu Y, Lapata M. Hierarchical transformers for multi-document summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL 2019; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1500]
10. Li H, Wang AY, Liu Y, Tang D, Lei Z, Li W. An augmented transformer architecture for natural language generation tasks. In: Proceedings of the International Conference on Data Mining Workshops. 2019 Presented at: ICDMW 2019; November 8-11, 2019; Beijing, China. [doi: 10.1109/icdmw48858.2019.9024754]
11. Wang Y, Gu Y, Yin Y, Han Y, Zhang H, Wang S, et al. Multimodal transformer augmented fusion for speech emotion recognition. Front Neurorobot 2023 May 22;17:1181598 [FREE Full text] [doi: 10.3389/fnbot.2023.1181598] [Medline: 37283784]
12. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc 2020 Dec 09;27(12):1935-1942 [FREE Full text] [doi: 10.1093/jamia/ocaa189] [Medline: 33120431]
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL-HLT 2019; June 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/n18-2]
14. Chow JC, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023 Apr 5;6:1166014 [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
15. Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis 2024 Mar;27(1):103-108. [doi: 10.1038/s41391-023-00705-y] [Medline: 37516804]
16. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. Innov Educ Teach Int 2023 Mar 27:1-15. [doi: 10.1080/14703297.2023.2195846]
17. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect 2023 Sep 11;13(3):183-191 [FREE Full text] [doi: 10.34172/hpp.2023.22] [Medline: 37808939]
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
19. Vedula SS, Ghazi A, Collins JW, Pugh C, Stefanidis D, Meireles O, et al. Artificial intelligence methods and artificial intelligence-enabled metrics for surgical education: a multidisciplinary consensus. J Am Coll Surg 2022 Jun 01;234(6):1181-1192 [FREE Full text] [doi: 10.1097/XCS.0000000000000190] [Medline: 35703817]
20. Collins JW, Marcus HJ, Ghazi A, Sridhar A, Hashimoto D, Hager G, et al. Ethical implications of AI in robotic surgical training: a Delphi consensus statement. Eur Urol Focus 2022 Mar;8(2):613-622 [FREE Full text] [doi: 10.1016/j.euf.2021.04.006] [Medline: 33941503]
21. Shinners L, Aggar C, Grace S, Smith S. Exploring healthcare professionals' perceptions of artificial intelligence: validating a questionnaire using the e-Delphi method. Digit Health 2021 Mar 23;7:20552076211003433 [FREE Full text] [doi: 10.1177/20552076211003433] [Medline: 33815816]
22. McKenna HP. The Delphi technique: a worthwhile research approach for nursing? J Adv Nurs 1994 Jun;19(6):1221-1225. [doi: 10.1111/j.1365-2648.1994.tb01207.x] [Medline: 7930104]
23. Hsu C, Sandford B. The Delphi technique: making sense of consensus. Pract Assess Res Eval 2007;12:1-8. [doi: 10.7275/PDZ9-TH90]
24. Neely A. Production/operations management: research process and content during the 1980s. Int J Oper Prod Manag 1993;13(1):5-18. [doi: 10.1108/01443579310023963]
25. Meredith JR, Raturi A, Amoako‐Gyampah K, Kaplan B. Alternative research paradigms in operations. J Oper Manag 1989 Oct;8(4):297-326. [doi: 10.1016/0272-6963(89)90033-8]

XSL•FO
RenderX

26. Akkermans H, Bogerd P, Vos B. Virtuous and vicious cycles on the road towards international supply chain management. J Econ Stud 1999 May;19(5/6):565-582. [doi: 10.1108/01443579910260883]
27. Akkermans HA, Bogerd P, Yücesan E, van Wassenhove LN. The impact of ERP on supply chain management: exploratory findings from a European Delphi study. Eur J Oper Res 2003 Apr 16;146(2):284-301. [doi: 10.1016/s0377-2217(02)00550-7]
28. Daniel EM, White A. The future of inter-organisational system linkages: findings of an international Delphi study. Eur J Inf Syst 2017 Dec 19;14(2):188-203. [doi: 10.1057/palgrave.ejis.3000529]
29. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform 2019 Aug;28(1):41-46 [FREE Full text] [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
30. Lam K, Iqbal FM, Purkayastha S, Kinross JM. Investigating the ethical and data governance issues of artificial intelligence in surgery: protocol for a Delphi study. JMIR Res Protoc 2021 Feb 22;10(2):e26552 [FREE Full text] [doi: 10.2196/26552] [Medline: 33616543]
31. Boulkedid R, Abdoul H, Loustau M, Sibony O, Alberti C. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. PLoS One 2011;6(6):e20476 [FREE Full text] [doi: 10.1371/journal.pone.0020476] [Medline: 21694759]
32. Humphrey-Murto S, Varpio L, Wood TJ, Gonsalves C, Ufholz LA, Mascioli K, et al. The use of the Delphi and other consensus group methods in medical education research: a review. Acad Med 2017 Oct;92(10):1491-1498. [doi: 10.1097/ACM.0000000000001812] [Medline: 28678098]
33. Houben G, Lenie K, Vanhoof K. A knowledge-based SWOT-analysis system as an instrument for strategic planning in small and medium sized enterprises. Decis Support Syst 1999 Aug;26(2):125-135. [doi: 10.1016/s0167-9236(99)00024-x]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. von der Gracht HA. Consensus measurement in Delphi studies: review and implications for future quality assurance. Technol Forecast Soc Change 2012 Oct;79(8):1525-1536. [doi: 10.1016/j.techfore.2012.04.013]
36. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017 Sep;31(8):684-706. [doi: 10.1177/0269216317690685] [Medline: 28190381]
37. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. Int J Clin Pharm 2016 Jun;38(3):655-662 [FREE Full text] [doi: 10.1007/s11096-016-0257-x] [Medline: 26846316]
38. Denecke K, May R. Investigating conversational agents in healthcare: application of a technical-oriented taxonomy. Procedia Comput Sci 2023;219:1289-1296. [doi: 10.1016/j.procs.2023.01.413]
39. Xygi E, Andriopoulos AD, Koutsomitropoulos DA. Question answering chatbots for biomedical research using transformers. In: Proceedings of the International Conference on Artificial Intelligence in Information and Communication. 2023 Presented at: ICAIIC 2023; February 20-23, 2023; Bali, Indonesia. [doi: 10.1109/icaiic57133.2023.10066979]
40. Saigaonkar S, Narawade V. Predicting chronic diseases using clinical notes and fine-tuned transformers. In: Proceedings of the IEEE Bombay Section Signature Conference. 2022 Presented at: IBSSC 2022; December 8-10, 2022; Mumbai, India. [doi: 10.1109/ibssc56953.2022.10037512]
41. López-García G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Transformers for clinical coding in Spanish. IEEE Access 2021 May 13;9:72387-72397. [doi: 10.1109/access.2021.3080085]
42. Gao Y, Miller T, Xu D, Dligach D, Churpek MM, Afshar M. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. Proc Int Conf Comput Ling 2022 Oct;2022:2979-2991 [FREE Full text] [Medline: 36268128]
43. Nooralahzadeh F, Gonzalez NP, Frauenfelder T, Fujimoto K, Krauthammer M. Progressive transformer-based generation of radiology reports. In: Proceedings of the Empirical Methods in Natural Language Processing. 2021 Presented at: EMNLP 2021; November 7-11, 2021; Punta Cana, Dominican Republic. [doi: 10.18653/v1/2021.findings-emnlp.241]
44. Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, et al. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. 2021 Presented at: BCB '21; August 1-4, 2021; Gainesville, FL. [doi: 10.1145/3459930.3469560]
45. Langlais É, Thériault-Lauzier P, Marquis-Gravel G, Kulbay M, So DY, Tanguay JF, et al. Novel artificial intelligence applications in cardiology: current landscape, limitations, and the road to real-world applications. J Cardiovasc Transl Res 2023 Jun;16(3):513-525. [doi: 10.1007/s12265-022-10260-x] [Medline: 35460017]
46. Banja JD, Hollstein RD, Bruno MA. When artificial intelligence models surpass physician performance: medical malpractice liability in an era of advanced artificial intelligence. J Am Coll Radiol 2022 Jul;19(7):816-820. [doi: 10.1016/j.jacr.2021.11.014] [Medline: 35120881]
47. May R, Biermann C, Krüger J, Saake G, Leich T. A systematic mapping study of security concepts for configurable data storages. In: Proceedings of the 26th ACM International Systems and Software Product Line Conference - Volume A. 2022 Presented at: SPLC '22; September 12-16, 2022; Graz, Austria. [doi: 10.1145/3546932.3546994]
48. Raza S, Osivand Pour P, Bashir SR. Fairness in machine learning meets with equity in healthcare. arXiv Preprint posted online on August 14, 2023. [FREE Full text] [doi: 10.1609/aaaiss.v1i1.27493]

XSL•FO
RenderX

49. Denecke K, Baudoin CR. A review of artificial intelligence and robotics in transformed health ecosystems. Front Med (Lausanne) 2022;9:795957 [FREE Full text] [doi: 10.3389/fmed.2022.795957] [Medline: 35872767]
50. Greco CM, Simeri A, Tagarelli A, Zumpano E. Transformer-based language models for mental health issues: a survey. Pattern Recognit Lett 2023 Mar;167:204-211. [doi: 10.1016/j.patrec.2023.02.016]
51. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904 [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
52. Kelly S, Kaye SA, Oviedo-Trespalacios O. What factors contribute to the acceptance of artificial intelligence? a systematic review. Telemat Inform 2023 Feb;77:101925. [doi: 10.1016/j.tele.2022.101925]
53. Kokol P, Vošner HB, Kokol M, Završnik J. The quality of digital health software: should we be concerned? Digit Health 2022 Jun 16;8:20552076221109055 [FREE Full text] [doi: 10.1177/20552076221109055] [Medline: 35746952]
54. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022 Feb 10;10(2):e32875 [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
55. Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
56. Embi PJ. Algorithmovigilance-advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Netw Open 2021 Apr 01;4(4):e214622 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.4622] [Medline: 33856479]
57. Capurro D, Huckdale K, Chapman W, Pantano V. The digital health validitron. Australian Healthcare and Hospitals Association. 2022. URL: https://issuu.com/aushealthcare/docs/the_health_advocate_august_2022/s/16504451 [accessed 2024-05-01]
58. Porsdam Mann S, Earp BD, Nyholm S, Danaher J, Møller N, Bowman-Smart H, et al. Generative AI entails a credit–blame asymmetry. Nat Mach Intell 2023 May 04;5(5):472-475. [doi: 10.1038/s42256-023-00653-1]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
HCP: health care professional
LLM: large language model
NLP: natural language processing
SWOT: strengths, weaknesses, opportunities, and threats

Edited by T de Azevedo Cardoso; submitted 02.09.23; peer-reviewed by R Yang, M Abdalla; comments to author 02.10.23; revised version received 10.10.23; accepted 19.04.24; published 13.05.24.

Please cite as:
Denecke K, May R, LLMHealthGroup, Rivera Romero O
Potential of Large Language Models in Health Care: Delphi Study
J Med Internet Res 2024;26:e52399
URL: https://www.jmir.org/2024/1/e52399
doi:10.2196/52399
PMID:38739445

©Kerstin Denecke, Richard May, LLMHealthGroup, Octavio Rivera Romero. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Based Prediction of Suicidal Thinking in Adolescents by Derivation and Validation in 3 Independent Worldwide Cohorts: Algorithm Development and Validation Study

Hyejun Kim[1,2*], BS; Yejun Son[1,3*], MS; Hojae Lee[1,4*], MS; Jiseung Kang[5,6], PhD; Ahmed Hammoodi[7], BS; Yujin Choi[1,8], BS; Hyeon Jin Kim[1,4], MS; Hayeon Lee[1], MS; Guillaume Fond[9], PhD; Laurent Boyer[9], PhD; Rosie Kwon[1], MS; Selin Woo[1], PhD; Dong Keon Yon[1,3,4,10], MD, PhD

[1]Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[2]Department of Applied Information Engineering, Yonsei University, Seoul, Republic of Korea

[3]Department of Precision Medicine, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[4]Department of Regulatory Science, Kyung Hee University, Seoul, Republic of Korea

[5]Division of Sleep Medicine, Harvard Medical School, Boston, MA, United States

[6]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, United States

[7]Department of Business Administration, Kyung Hee University School of Management, Seoul, Republic of Korea

[8]Department of Korean Medicine, Kyung Hee University College of Korean Medicine, Seoul, Republic of Korea

[9]Assistance Publique–Hôpitaux de Marseille (APHM), CEReSS–Health Service Research and Quality of Life Center, Aix–Marseille University, Marseille, France

[10]Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Dong Keon Yon, MD, PhD
Center for Digital Health
Medical Science Research Institute
Kyung Hee University College of Medicine
23 Kyungheedae–ro
Dongdaemun–gu
Seoul, 02447
Republic of Korea
Phone: 82 2 6935 2476
Fax: 82 504 478 0201
Email: yonkkang@gmail.com

Abstract

Background: Suicide is the second-leading cause of death among adolescents and is associated with clusters of suicides. Despite numerous studies on this preventable cause of death, the focus has primarily been on single nations and traditional statistical methods.

Objective: This study aims to develop a predictive model for adolescent suicidal thinking using multinational data sets and machine learning (ML).

Methods: We used data from the Korea Youth Risk Behavior Web-based Survey with 566,875 adolescents aged between 13 and 18 years and conducted external validation using the Youth Risk Behavior Survey with 103,874 adolescents and Norway's University National General Survey with 19,574 adolescents. Several tree-based ML models were developed, and feature importance and Shapley additive explanations values were analyzed to identify risk factors for adolescent suicidal thinking.

Results: When trained on the Korea Youth Risk Behavior Web-based Survey data from South Korea with a 95% CI, the XGBoost model reported an area under the receiver operating characteristic (AUROC) curve of 90.06% (95% CI 89.97-90.16), displaying superior performance compared to other models. For external validation using the Youth Risk Behavior Survey data from the United States and the University National General Survey from Norway, the XGBoost model achieved AUROCs of 83.09% and 81.27%, respectively. Across all data sets, XGBoost consistently outperformed the other models with the highest AUROC score,

XSL•FO
RenderX

and was selected as the optimal model. In terms of predictors of suicidal thinking, feelings of sadness and despair were the most influential, accounting for 57.4% of the impact, followed by stress status at 19.8%. This was followed by age (5.7%), household income (4%), academic achievement (3.4%), sex (2.1%), and others, which contributed less than 2% each.

Conclusions: This study used ML by integrating diverse data sets from 3 countries to address adolescent suicide. The findings highlight the important role of emotional health indicators in predicting suicidal thinking among adolescents. Specifically, sadness and despair were identified as the most significant predictors, followed by stressful conditions and age. These findings emphasize the critical need for early diagnosis and prevention of mental health issues during adolescence.

(J Med Internet Res 2024;26:e55913) doi:10.2196/55913

KEYWORDS

adolescent; machine learning; Shapley additive explanations; SHAP value; suicidal thinking; XGBoost; mental health; predictive model; risk behavior

Introduction

Adolescent suicide stands out as a prominent global public health concern, with its rank as the second leading cause of death among young populations underscoring its severity [1,2] Notably, adolescence is a phase characterized by an amplified suicide risk [3]. Concerningly, some geographic regions are experiencing a surge in suicide clusters, where the instances of suicide exceed the typical levels [4]. Research into these clusters indicates that individuals younger than 25 years are up to 4 times more likely to be affected by suicide [5]. Since suicide is preventable in the early stages, there is a pressing need for action through rigorous mental health strategies and proactive educational interventions [6].

While various methodologies have been proposed to prevent suicidal thinking in adolescents, many lack empirical outcomes and often fail to identify key determinants [7-9]. A significant gap remains in accurately assessing the risk of suicidal thinking for individual adolescents [2,10,11]. Recent advances in machine learning (ML) methodologies have shown promise in addressing the challenges of adolescent suicidal tendencies. Studies leveraging boosted ML [12], daily data analysis through classification and regression trees [13], and risk and protective factor frameworks [14] have begun to unpack the complex interplay of factors contributing to suicidal thinking among adolescents. However, these studies have also highlighted limitations, including a focus on specific socioeconomic or short-term predictors and a lack of comprehensive risk profiles integrating emotional, social, and psychological variables [12-14].

Therefore, in this study, we developed a predictive model for suicidal thinking among adolescents, using advanced ML algorithms. Addressing the gaps identified in earlier research, our model incorporates a broader array of factors, including family dynamics, emotional well-being, academic performance, and general health indicators. Across distinct adolescent cohorts from South Korea, Norway, and the United States, we aimed for a comprehensive multinational approach. By refining our approach based on previous studies' insights, this research aims to highlight the preventability of suicide and influence mental health clinicians and policy makers to develop more effective preventive measures and supportive programs.

Methods

Study Design and Participants

This study aimed to develop an ML model to predict suicidal thinking among Korean adolescents. Our approach used multiple variables extracted from 3 distinct, large-scale international data sources: the Korea Youth Risk Behavior Web-based Survey (KYRBS) [15,16], the Youth Risk Behavior Survey (YRBS), and Norway's nationwide University National General Survey (Ungdata).

Data Preparation and Harmonization

Initial data preprocessing involved adjusting the sample sizes after the removal of missing values: KYRBS from 1,145,178 to 566,875, YRBS from 438,566 to 103,874, and Ungdata from 89,077 to 19,574. We analyzed data from adolescents aged between 13 and 18 years who participated in the KYRBS from 2009 to 2021, the YRBS in 2021, and Ungdata from 2017 to 2019. The primary outcome, termed "current suicidal thinking," was derived from participants' affirmative responses to the question, "During the past 12 months, did you ever seriously consider attempting suicide?" This outcome indicated that participants had contemplated serious suicidal thinking at least once in the preceding year. The analysis considered several covariates: region, age, sex, BMI (kg/m^2), academic achievement, household income, smoking status, alcoholic consumption, stress status, feelings of sadness and despair, exercise habits, and screen time (Figure S1 in Multimedia Appendix 1) [17].

We harmonized the data sets for XGBoost model compatibility, addressing the challenge posed by different variable configurations within the same questions. Our preprocessing aligned each variable across the KYRBS, YRBS, and Ungdata data sets, ensuring they matched in terms of content and format. Recognizing the potential disparities in variable configurations across these data sets, we standardized the variable names, formats, and scales, focusing on key features such as demographic information, behavioral factors, psychosocial aspects, and environmental influences that could serve as predictors for suicidal thinking. To ensure consistency, we adopted the following strategic approach to cases where YRBS and Ungdata were missing certain features: by calculating the median of the missing variables in the KYRBS data set, we were able to effectively impute the missing values to maintain

XSL•FO
RenderX

the integrity and comparability of the data set compilations. Through variable alignment and addressing missing data, we successfully harnessed the diverse strengths of each data set, facilitating the development of a comprehensive model designed to address adolescent suicidal thinking effectively.

ML Model Development

Our ML model underwent training and validation to ensure its predictive accuracy in identifying suicidal thinking. We used the KYRBS data set to build a model tailored to predict suicidal thinking among Korean adolescents aged between 13 and 18 years. Recognizing the intricate characteristics of the data, we used a variety of tree-based ML techniques, including XGBoost, adaptive boosting (AdaBoost), light gradient-boosting machine (LightGBM), and random forest to train the data set for our modeling process [18]. Before this, data preprocessing measures, such as addressing missing values and encoding categorical variables, were executed to maintain data integrity and optimize the data for the modeling phase.

We adopted the 10-fold cross-validation method, dividing the initial data set into 10 equal-sized subsets, to rigorously assess the performance efficacy of the ML model. Of these, 9 are designated for model training, while the remaining subset serves as validation [19]. The process iterates 10 times, ensuring each subset undergoes validation at least once. During each cycle, we computed various performance metrics such as area under the receiver operating characteristic (AUROC) curve, sensitivity, specificity, accuracy, and balanced accuracy, along with their respective 95% CIs [20-25]. The 95% CIs provide a range of possible values for the model's performance metrics, allowing us to assess the stability and generalizability of the model. For a visual representation of the model efficacy, we used visualization methods, primarily the receiver operating characteristic curve. After 10 iterations, the metrics from each were averaged to determine the final performance evaluation.

We trained our model on the KYRBS data set using 10-fold cross-validation. The model trained on the KYRBS data set was then externally validated with YRBS and Ungdata data sets preprocessed with the same column structure as KYRBS. This rigorous process reinforced the reliability of our model's performance trained on the KYRBS data set [26]. Among the 4 models tested, XGBoost consistently yielded the highest AUROC scores across all data sets, leading to its selection as the primary model.

We performed hyperparameter tuning using GridSearchCV to optimize the performance of the XGBoost model, prioritizing the maximization of the AUROC score to determine the optimal hyperparameter combination. Hyperparameters were carefully selected for improved performance: the gbtree booster was used for its effectiveness in classification tasks, and the logloss evaluation metric was chosen to ensure accurate probability estimations. We set the learning rate at 0.08 to balance training speed with model accuracy, and the max depth was capped at 5 to prevent overfitting while allowing the model to capture complex patterns. Additionally, 350 trees (n_estimators) were used to construct a robust model, with further adjustments made to parameters like "scale_pos_weight" and subsample to address class imbalance and enhance model stability. These adjustments were important in refining our model's predictive capabilities and are detailed in Table S1 in Multimedia Appendix 1. In order to interpret and gain insights into the model predictions, we used Shapley additive explanations (SHAP) values, a unified measure derived from cooperative game theory. Data set variables were analyzed with SAS software (version 9.3; SAS Institute Inc), and ML analysis was performed using Python (version 3.11.4; Python Software Foundation). The main Python libraries used are as follows: NumPy (version 1.26.0; Python Software Foundation) for data arrays and operations, and Pandas (version 2.1.0; Python Software Foundation) for data manipulation and analysis. All 3—scikit-learn (version 1.2.2; scikit-learn development team), TensorFlow-gpu (version 2.6.0; Tensor development team), and Keras (version 2.6.0; Keras development team)—were used for constructing and training ML models [27]. Additionally, the SHAP package (version 0.42.1) was used to interpret the ML models and for its explanation capabilities [28].

Software and Libraries

All computations, model training, and evaluations were executed using Python (version 3.11.4). Key libraries from our toolbox included scikit-learn (version 1.2.2), NumPy (version 1.24.0), and Pandas (version 2.1.0) for ML tasks and data wrangling. Visualization was facilitated using Matplotlib (version 3.7.2) and Seaborn (version 0.12.2).

Ethical Considerations

The study protocol was approved by the institutional review board of the Korean Disease Control and Prevention Agency (2014-06EXP-02-P-A), the US Centers for Disease Control and Prevention (#1969.0), the Norwegian Centre for Research Data and Data Protection Office of Inland Hospital Trust (18778329) and by the local law of the Population Health Promotion Act 19 (117058) form of the Korean government. All participants provided written informed consent. This research followed the guidelines outlined in the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement (Table S2 in Multimedia Appendix 1).

Results

Demographic Characteristics

This study was conducted to develop a ML-based predictive model for suicidal thinking among adolescents aged between 13 and 18 years. After collecting independent data from 3 countries, covariates were standardized for the ML prediction modeling process (Figures 1 and 2).

XSL•FO
RenderX

Figure 1. Study architecture. AdaBoost: adaptive boosting; KYRBS: Korea Youth Risk Behavior Web-based Survey; LightGBM: light gradient-boosting machine; Ungdata: Norwegian nationwide Ungdata surveys; YRBS: Youth Risk Behavior Survey.

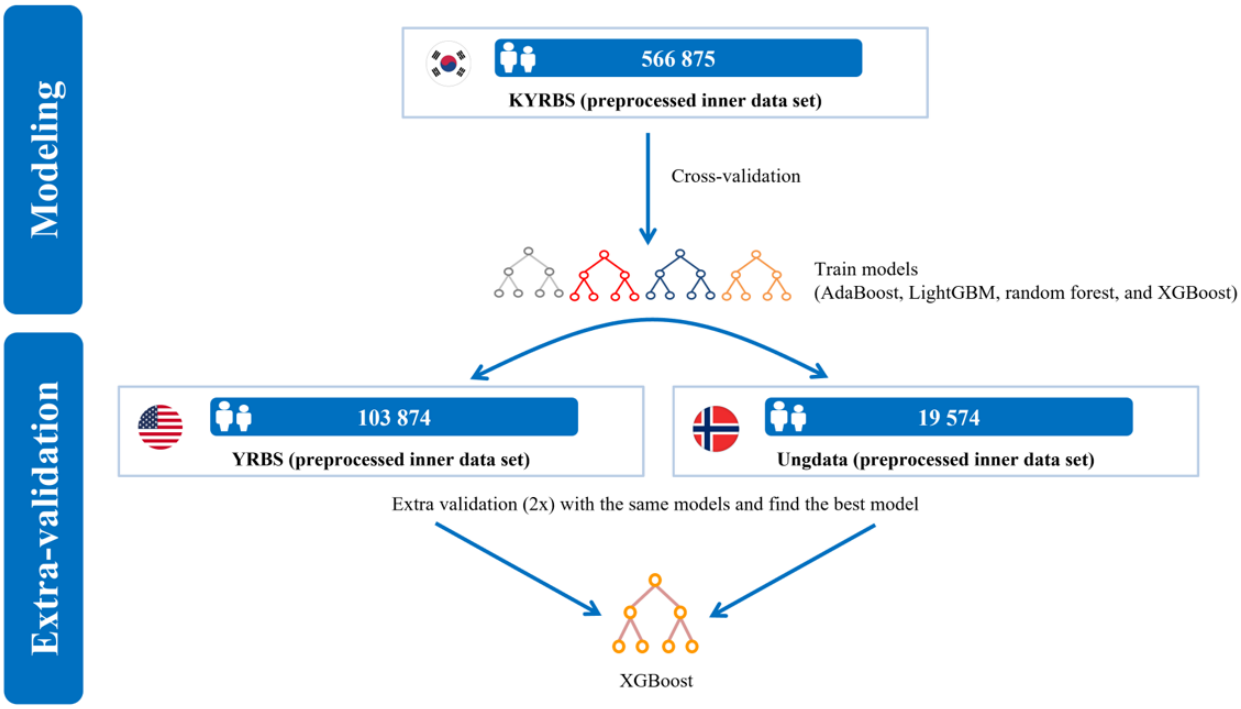

Figure 2. Model architecture. The original Korea Youth Risk Behavior Web-based Survey (KYRBS) data set was partitioned into the original data set for model development, with performance assessed using area under the receiver operating characteristic (AUROC) curve scores. Selected high-performing models were further validated using an external Youth Risk Behavior Survey (YRBS) data set. The validation results were derived from the original data set, external results from the additional YRBS data set, and the Norwegian nationwide Ungdata survey (Ungdata) data set. CV: cross validation.

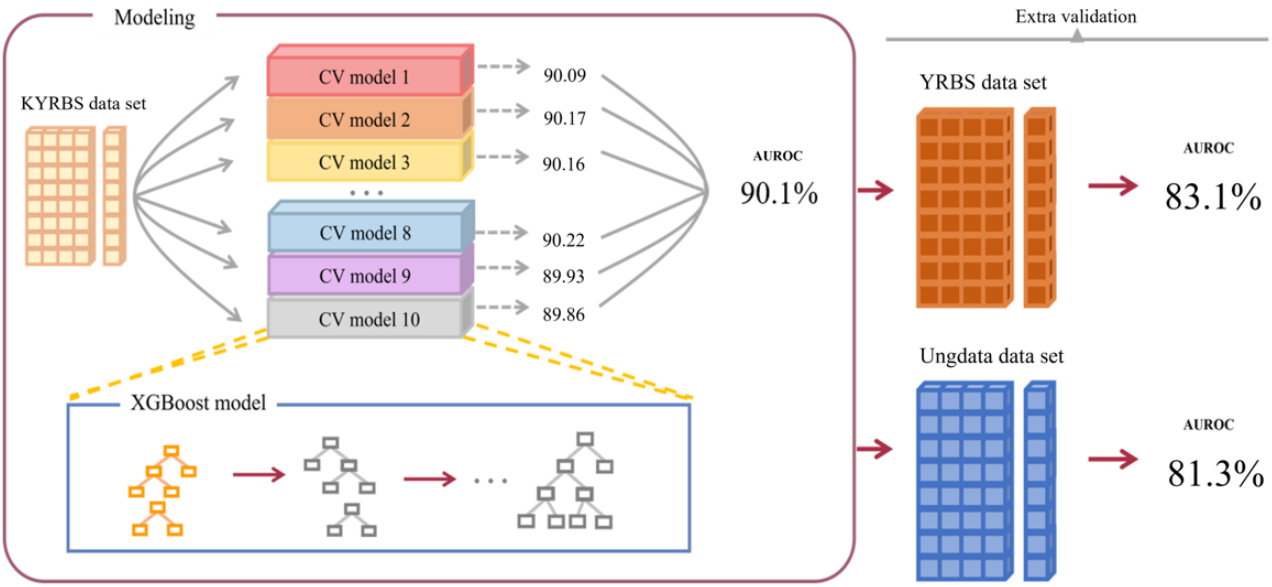

The distribution of age in the initial training cohort for KYRBS from South Korea, which was used to build the prediction model, was as follows: aged 13 years (94,923/566,875, 16.74%), aged 14 years (98,624/566,875, 17.4%), aged 15 years (100,490/566,875, 17.73%), aged 16 years (90,057/566,875, 15.89%), aged 17 years (92,071/566,875, 16.24%), and aged 18 years (90,710/566,875, 16%). For the external validation cohort using YRBS from the United States, the age distribution was as follows: aged 13 years (351/103,874, 0.34%), aged 14 years (21,095/103,874, 20.31%), aged 15 years (28,016/103,874, 26.97%), aged 16 years (25,929/103,874, 24.96%), aged 17 years (22,405/103,874, 21.57%), and aged 18 years (6078/103,874, 5.85%). Another external validation stage using Ungdata from Norway had the following age distribution: aged 13 years (5039/19,574, 25.74%), aged 14 years (4874/19,574, 24.9%), aged 15 years (5034/19,574, 25.72%), aged 16 years (3181/19,574, 16.25%), aged 17 years (845/19,574, 4.32%), and aged 18 years (601/19,574, 3.07%) (Table 1).

XSL•FO
RenderX

Table 1. Demographic characteristics of Korea Youth Risk Behavior Web-based Survey (KYRBS) data from South Korea (2009-2021), Youth Risk Behavior Survey (YRBS) data from the United States (2021), and University National General Survey (Ungdata) from Norway (2017-2019).

Characteristics	KYRBS (n=566,875)	YRBS (n=103,874)	Ungdata (n=19,574)
Region, n (%)			
Urban	259,453 (45.77)	N/A[a]	N/A
Rural	307,422 (54.23)	N/A	N/A
Age (years), n (%)			
13	94,923 (16.74)	351 (0.34)	5039 (25.74)
14	98,624 (17.4)	21,095 (20.31)	4874 (24.9)
15	100,490 (17.73)	28,016 (26.97)	5034 (25.72)
16	90,057 (15.89)	25,929 (24.96)	3181 (16.25)
17	92,071 (16.24)	22,405 (21.57)	845 (4.32)
18	90,710 (16)	6078 (5.85)	601 (3.07)
Sex, n (%)			
Male	289,311 (51.04)	51,842 (49.91)	9812 (50.13)
Female	277,564 (48.96)	52,032 (50.09)	9762 (49.87)
BMI[b], n (%)			
Underweight	44,539 (7.86)	12,541 (12.07)	N/A
Normal	423,384 (74.69)	47,592 (45.82)	N/A
Overweight	48,951 (8.64)	14,439 (13.9)	N/A
Obese	50,001 (8.82)	29,302 (28.21)	N/A
Academic achievement, n (%)			
Low (0-19th percentiles)	56,799 (10.02)	N/A	N/A
Lower-middle (20th -39th percentiles)	132,774 (23.42)	N/A	N/A
Middle (40th-59th percentiles)	162,447 (28.66)	N/A	N/A
Upper-middle (60th-79th percentiles)	146,401 (25.83)	N/A	N/A
High (80th-100th percentiles)	68,454 (12.08)	N/A	N/A
Household income, n (%)			
Low (0-19th percentiles)	19,200 (3.39)	N/A	8466 (43.25)
Lower-middle (20th -39th percentiles)	82,314 (14.52)	N/A	6557 (33.5)
Middle (40th-59th percentiles)	274,093 (48.35)	N/A	3523 (18)
Upper-middle (60th-79th percentiles)	148,848 (26.26)	N/A	818 (4.18)
High (80th-100th percentiles)	42,420 (7.48)	N/A	210 (1.07)
Smoking status, n (%)			
Nonsmoker	467,707 (82.51)	84,984 (81.81)	16,594 (84.78)
Smoker	99,168 (17.49)	18,890 (18.19)	2980 (15.22)
Alcohol consumption, n (%)			
Nondrinker	478,305 (84.38)	76,571 (73.72)	9673 (49.42)
Drinker	88,570 (15.62)	27,303 (26.28)	9901 (50.58)
Stress status[c], n (%)			
Low to moderate	337,938 (59.61)	70,817 (68.18)	N/A
High to severe	228,937 (40.39)	33,057 (31.82)	N/A
Sadness and despair, n (%)			
Low to moderate	401,253 (70.78)	64,190 (61.8)	14,522 (74.19)

XSL•FO
RenderX

Characteristics	KYRBS (n=566,875)	YRBS (n=103,874)	Ungdata (n=19,574)
High to severe	165,622 (29.22)	39,684 (38.2)	5052 (25.81)
Exercise status, n (%)			
Not enough	496,475 (87.58)	87,510 (84.25)	N/A
Enough	70,400 (12.42)	16,364 (15.75)	N/A
Suicide thinking in the past year, n (%)			
No	482,613 (85.14)	80,561 (77.56)	17,652 (90.18)
Yes	84,262 (14.86)	23,313 (22.44)	1922 (9.82)
Screentime status, n (%)			
Low to moderate	165,382 (29.17)	27,569 (26.54)	994 (5.08)
High to severe	401,493 (70.83)	76,305 (73.46)	18,580 (94.92)

[a]N/A: not applicable.

[b]BMI was divided into 4 groups according to the National Growth Charts: underweight (0-4th percentiles), normal (5th-84th percentiles), overweight (85th-94th percentiles), and obese (95th-100th percentiles).

[c]Stress was defined by receipt of mental health counseling owing to stress.

Both the initial training cohort and the external validation cohorts took into account socioeconomic backgrounds, such as household income and academic achievement, as well as risk behaviors such as alcohol consumption, smoking, and screen time. Additionally, factors that could potentially influence mental health, such as feelings of sadness and despair, were also considered. Inconsistencies or missing values in validation sets were addressed by implementing median imputation from the primary training data. Such thorough demographic incorporation bolsters our model performance, offering a nuanced understanding of suicidal thinking in adolescents.

ML Model Results

Table S1 in Multimedia Appendix 1 and Table 2 present the process of hyperparameter tuning the XGBoost model and the evaluation of our models conducted on data sets from 3 distinct countries using 5 performance metrics. Notably, XGBoost emerged as the frontrunner among the 4 tested models by getting hyperparameters of booster: gbtree, eval_metric: logloss, learning_rate: 0.08, max_depth: 5, n_estimators: 350, scale_pos_weight: 2, subsample: 0.09 (Table S1 in Multimedia Appendix 1). When put to the training on the KYRBS data set from South Korea, with a 95% CI, the XGBoost model reported an AUROC of 90.06 (95% CI 89.97-90.16), sensitivity was 82.11 (95% CI 81.67-82.55), specificity was 82.16 (95% CI 81.68-82.63), accuracy was 82.13 (95% CI 82.01-82.26), and balanced accuracy was 82.13 (95% CI 82.01-82.26), consistently displaying superior results compared to the other 3 models. During the external validation assessment, the model evaluation was conducted without considering the 95% CI. For the external model validation, using the YRBS from the United States, the XGBoost model achieved an AUROC of 83.09%, sensitivity of 80.26%, specificity of 75.52%, accuracy of 76.58%, and balanced accuracy of 77.89%. For the external validation using the Ungdata from Norway, the XGBoost model achieved an AUROC of 81.27%, sensitivity of 79.19%, specificity of 80%, accuracy of 79.92%, and balanced accuracy of 79.60%. Across all data sets, XGBoost consistently outperformed all other models with the highest AUROC score, which was selected as the most optimal model (Figure 3).

XSL•FO
RenderX

Table 2. Various performance metrics for 4 different machine learning algorithms in the Korea Youth Risk Behavior Web-based Survey (KYRBS) data set for validation, the Youth Risk Behavior Survey (YRBS) data set for extra validation, and another external validation at the Norwegian nationwide University National General Survey (Ungdata) surveys data set.

Data set and model	AUROC[a]	Sensitivity	Specificity	Accuracy	Balanced accuracy
South Korea, % (95% CI)					
XGBoost	90.06 (89.97-90.16)	82.11 (81.67-82.55)	82.16 (81.68-82.63)	82.13 (82.01-82.26)	82.13 (82.01-82.26)
AdaBoost[b]	85.26 (85.14-85.38)	78.08 (77.94-78.21)	78.09 (77.96-78.22)	78.08 (77.95-78.21)	78.08 (77.95-78.21)
LightGBM[c]	84.64 (84.53-84.75)	78.96 (78.88-79.04)	78.96 (78.88-79.04)	78.96 (78.88-79.04)	78.96 (78.88-79.04)
Random forest	84.10 (83.99-84.21)	76.91 (76.73-77.10)	77.06 (76.98-77.15)	76.99 (76.86-77.11)	76.99 (76.86-77.11)
United States, %					
XGBoost	83.09	80.26	75.52	76.58	77.89
AdaBoost	82.13	76.75	77.06	76.99	76.91
LightGBM	83.93	79.85	75.87	76.76	77.86
Random forest	83.20	78.28	77.43	77.62	77.86
Norway, %					
XGBoost	81.27	79.19	80.00	79.92	79.60
AdaBoost	79.14	79.19	80.00	79.92	79.60
LightGBM	84.14	77.98	81.95	81.87	79.97
Random forest	79.23	79.50	79.50	79.55	79.97

[a]AUROC: area under the receiver operating characteristic.

[b]AdaBoost: adaptive boosting.

[c]LightGBM: light gradient-boosting machine.

Figure 3. Receiver operating characteristic (ROC) curves with CIs for the Korea Youth Risk Behavior Web-based Survey (KYRBS), Youth Risk Behavior Survey (YRBS), and University National General Survey (Ungdata) of the XGBoost model.

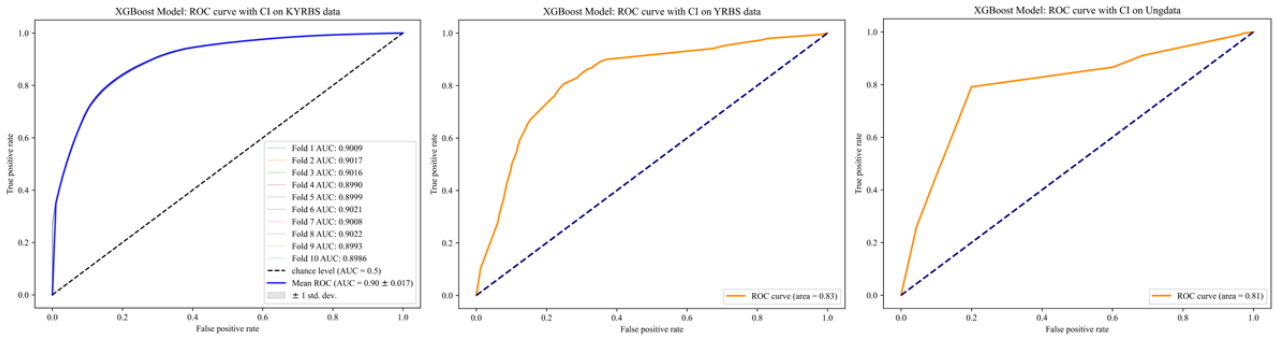

Feature Importance

Table 3 shows the feature importance derived from the XGBoost model, illustrating the relative contributions of each feature to predicting suicidal thinking. Notably, feelings of sadness and despair emerge as the most dominant predictor, accounting for 57.4% of the influence, followed by stress status at 19.8%. Subsequent factors include age (5.7%), household income (4%), academic achievement (3.4%), sex (2.1%), and others contributing less than 2% each.

XSL•FO
RenderX

Table 3. Feature importance of XGBoost.

Feature	Importance, %
Sadness and despair	57.4
Stress status	19.8
Age	5.7
Household income	4
Academic achievement	3.4
Sex	2.1
Smoking status	1.7
BMI	1.6
Alcohol consumption	1.3
Exercise status	1.3
Screentime status	1
Region	0.8

SHAP Value

We addressed a deeper visual interpretation of the SHAP values within our ML model (Figure S2 in Multimedia Appendix 1) [29]. Figure S3 in Multimedia Appendix 1 provides a waterfall plot, distinctively showcasing the cumulative contribution of each feature to a single prediction. We interpreted individual predictions by starting from the initial estimate and sequentially incorporating the influence of each feature to reach the final prediction. E[f(x)] refers to the average predicted output of the model across the entire data set, providing insights into the model's overall prediction tendency. The starting point of the illustration, denoted as E[f(x)]=0.83, represents the model's average prediction for the given data set. Among the variables, sadness and despair stood out, boosting the prediction by 1.17 and ranking as the most influential factor. Conversely, stress status and sex reduced the prediction by 0.86 and 0.22, respectively. This visualization offers a clear insight into the profound influence each feature wields in predicting adolescent suicidal thinking. Our ML model notably underscores a substantial reliance on the sadness, despair, and stress status features (Figure S3 in Multimedia Appendix 1).

Code Availability

Based on the results of the ML model, we established a web-based app for policy implementation or health system management to support in their decision-making process for cases involving suicidal thinking prediction use in adolescents [30]. An example of a web interface and the results are shown in Figure S4 in Multimedia Appendix 1. Custom code for the website is available on the web [31].

Discussion

Key Findings

Our research represents a pioneering machine-learning initiative for predicting suicidal thinking among adolescents. We sourced distinct data sets from South Korea (KYRBS), the United States (YRBS), and Norway (Ungdata). This provided comprehensive analysis related to socioeconomic indicators and key mental health influencers such as alcohol consumption, smoking status, and feelings of sadness and despair. Importantly, our findings highlight XGBoost as the optimal predictive model, achieving an AUROC of 88.6% with the KYRBS data set. External validation with the data from the United States and Norway yielded AUROCs of 82.9% and 83.6%, respectively. The most significant predictor of suicidal thinking was sadness and despair, with a feature influence of 61%, followed by stress status at 19.6%. Using SHAP values, we further emphasized the pivotal roles of sadness, despair, and stress in predicting suicidal thinking in adolescents. To enhance the practical application of our research, we have developed a web-based platform to visualize the prediction model, accompanied by a mobile interface, enhancing its accessibility and user experience. This dual-platform system provides a more methodical and analytical approach for the public to comprehend and manage potential suicidal concerns.

Plausible Mechanism

The close relationship between feelings of sadness and despair and suicidal thinking in adolescents can be understood from various perspectives, encompassing biological and environmental factors [32]. During adolescence, the brain undergoes significant development, especially in the prefrontal cortex, which controls impulses and emotions [33]. Persistent sadness can interfere with adolescent brain development, resulting in a perpetual state of negative emotions. This increases their risk of suicidal thinking due to feelings of despair and impulsive actions [34].

The influence of external portrayals, be it from peers or the media, cannot be underestimated. When adolescents confront additional adversities, such as bullying, social isolation, or academic failures, these inherent stressors are amplified. Adolescents exposed to narratives associating despair with suicidal behaviors might inadvertently absorb these sentiments. This phenomenon, known as "suicide contagion," postulates that exposure to others' suicidal actions can reshape an individual's perspective [35]. The confluence of these environmental stressors and limited emotional regulation

XSL•FO
RenderX

capacities heightens their vulnerability, potentially leading them to view suicide as a viable solution to their emotional turmoil [36].

Furthermore, due to their developmental stage, many adolescents have not yet acquired the necessary emotional coping strategies [37]. When faced with intense stressors without these tools, some may come to view suicide as their only way to escape from increasingly desperate circumstances.

This emotional vulnerability is further compounded by physiological changes. The stress response, regulated by the hypothalamic-pituitary-adrenal axis, is intensified during adolescence [38]. Heightened sensitivity of the hypothalamic-pituitary-adrenal axis results in increased cortisol production in response to stress [39]. Chronic exposure to these elevated cortisol levels not only exacerbates feelings of sadness and despair but also directly contributes to an increased vulnerability to suicidal thoughts [40]. Understanding these relationships makes it evident that both the emotional responses induced by stress and the physiological effects of stress significantly influence the increased propensity for suicidal thinking during this critical phase of life.

Strengths and Limitations

The limitations of this study should be stated. One primary concern pertains to the use of self-reported data, which exposes our results to potential biases, such as recall and social desirability. While this approach offers insights directly from the participants, such susceptibilities might skew the data and ultimately affect the model's performance. It is also worth noting that the foundational training data were sourced predominantly from adolescents in South Korea [41,42]. This could amplify specific cultural or racial attributes distinctive to Korean adolescents. Equally important to note is that establishing a direct cause-and-effect relationship between significant risk factors and adolescent suicidal thinking remains elusive. This study, while expansive, does not determine if suicidal thinking is a cause or an effect of other risk factors. Further research is needed to unravel these complex interconnections. The potential for overfitting is another critical limitation to consider. Our comprehensive model, regardless of its use of 10-fold cross-validation, might inadvertently capture anomalies rather than genuine patterns [43]. Furthermore, our method for managing missing data, especially through median imputation, poses the risk of introducing unintended biases, which could impact the model's performance [44]. Lastly, while predicting suicide attempts rather than suicide ideation may be crucial in suicide-related research, the low prevalence of suicide attempts presents limitations in constructing ML models: thus, we have developed predictive models for suicide ideation. With our predictive model, policy researchers, physicians, and community neighbors can develop individualized prevention strategies for these adolescents.

Despite these limitations, the strengths of this study are manifold. Notably, the SHAP value analysis highlights the importance of diverse demographic and behavioral indicators in understanding suicidal thinking among adolescents. This study provides nuanced insight into each feature's unique influence and the model's decision-making process [29]. The robustness of the model is also noteworthy, as evidenced by its uniform effectiveness across data sets spanning South Korea, the United States, and Norway. Such wide-reaching efficacy suggests the adaptability of the model to different cultural and demographic landscapes. Another strength lies in the real-world applicability of our research. The development of our advanced web-based platform and mobile interface marks a notable advancement in practical application. By offering a user-friendly interface tailored for both desktop and mobile users, we enhance accessibility and promote greater self-awareness about suicidal risk. This can encourage individuals to seek timely professional assistance or supportive resources, serving as a preventive measure against severe mental health crises [45].

Clinical and Policy Implications

In light of the findings, several crucial policy implications emerge. Foremost, the significant role of sadness and despair as predictors underscores the necessity to prioritize mental health support for adolescents [46]. This prominence not only necessitates immediate interventions but also stresses the vital role of education and awareness initiatives, targeting both risk behaviors and associated mental health ramifications. Early identification of suicidal thinking is paramount [47]. Recognizing these indications enables health care professionals to initiate early intervention strategies. This preemptive approach should include tailored counseling, support group engagements, or intensive therapeutic interventions. This can prevent the progression toward actual suicide attempts, which might be driven by mixed emotions or even an intent just to signal distress [48]. Moreover, there is a pressing need to bolster educational and awareness campaigns concerning suicide. Such campaigns serve to equip adolescents with the tools and knowledge necessary, encouraging them to navigate challenges related to risky behaviors and maintain positive mental health perspectives [49].

Conclusion

In addressing the pressing global concern of adolescent suicide, this study used ML to offer novel insights into preemptive detection. By integrating diverse data sets across 3 nations, the study highlighted the superiority of the XGBoost model in predicting suicidal thinking, achieving remarkable AUROCs of 90.06% (95% CI 89.97-90.16; KYRBS from South Korea; discovery), 83.09% (YRBS from the United States; extra validation), and 81.27% (Ungdata from Norway; extra validation). Our findings emphasize the significant role of emotional health indicators in predicting suicidal thinking among adolescents. Specifically, sadness and despair proved to be the most influential predictors, followed by stress status and age. Through our robust, cross-culturally validated model and its accessibility through web-based platforms, we underscore the potential for timely interventions and offer a promising blueprint for future mental health strategies and preventive measures for at-risk adolescents.

XSL•FO
RenderX

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (HI22C1976). The funders had no role in study design, data collection, data analysis, data interpretation, or the writing of the report.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request. The study protocol and statistical code are available from DKY. The data set can be accessed from the Korean Disease Control and Prevention Agency, US Centers for Disease Control and Prevention, and Norwegian Social Research Institute (NOVA) through a data use agreement.

Authors' Contributions

DKY had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. HK, YS, Hojae Lee, RK, SW, and DKY were responsible for study concept and design; acquisition, analysis, or interpretation of data; drafting of the manuscript; and statistical analysis. All authors were responsible for critical revision of the manuscript for important intellectual content and approved the final version before submission. Study supervision was done by RK, SW, and DKY. DKY supervised the study and is the guarantor for this study. HK, YS, and Hojae Lee contributed equally as first authors. RK, SW, and DKY contributed equally as corresponding authors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study population; SHAP value of XGBoost model; SHAP waterfall plot; deployed web-based application to provide suicidal thinking prediction among adolescents; process of hyperparameter tuning for the XGBoost model on KYRBS data set; and transparent reporting of a multivariable prediction model for individual.
[DOCX File , 410 KB - jmir_v26i1e55913_app1.docx]

References

1. Hughes JL, Horowitz LM, Ackerman JP, Adrian MC, Campo JV, Bridge JA. Suicide in young people: screening, risk assessment, and intervention. BMJ 2023;381:e070630 [FREE Full text] [doi: 10.1136/bmj-2022-070630] [Medline: 37094838]
2. Lee H, Cho JK, Park J, Lee H, Fond G, Boyer L, et al. Machine learning-based prediction of suicidality in adolescents with allergic rhinitis: derivation and validation in 2 independent nationwide cohorts. J Med Internet Res 2024;26:e51473 [FREE Full text] [doi: 10.2196/51473] [Medline: 38354043]
3. Hawton K, Hill NTM, Gould M, John A, Lascelles K, Robinson J. Clustering of suicides in children and adolescents. Lancet Child Adolesc Health 2020;4(1):58-67. [doi: 10.1016/S2352-4642(19)30335-9] [Medline: 31606323]
4. India State-Level Disease Burden Initiative Suicide Collaborators. Gender differentials and state variations in suicide deaths in India: the Global Burden of Disease Study 1990-2016. Lancet Public Health 2018;3(10):e478-e489 [FREE Full text] [doi: 10.1016/S2468-2667(18)30138-5] [Medline: 30219340]
5. Méndez-Bustos P, Fuster-Villaseca J, Lopez-Castroman J, Jiménez-Solomon O, Olivari C, Baca-Garcia E. Longitudinal trajectories of suicidal ideation and attempts in adolescents with psychiatric disorders in Chile: study protocol. BMJ Open 2022;12(2):e051749 [FREE Full text] [doi: 10.1136/bmjopen-2021-051749] [Medline: 35193905]
6. Mann JJ, Michel CA, Auerbach RP. Improving suicide prevention through evidence-based strategies: a systematic review. Am J Psychiatry 2021;178(7):611-624 [FREE Full text] [doi: 10.1176/appi.ajp.2020.20060864] [Medline: 33596680]
7. Mars B, Heron J, Klonsky ED, Moran P, O'Connor RC, Tilling K, et al. Predictors of future suicide attempt among adolescents with suicidal thoughts or non-suicidal self-harm: a population-based birth cohort study. Lancet Psychiatry 2019;6(4):327-337 [FREE Full text] [doi: 10.1016/S2215-0366(19)30030-6] [Medline: 30879972]
8. Kwak CW, Ickovics JR. Adolescent suicide in South Korea: risk factors and proposed multi-dimensional solution. Asian J Psychiatr 2019;43:150-153. [doi: 10.1016/j.ajp.2019.05.027] [Medline: 31151083]
9. Jeong SC, Kim JY, Choi MH, Lee JS, Lee JH, Kim CW, et al. Identification of influencing factors for suicidal ideation and suicide attempts among adolescents: 11-year national data analysis for 788,411 participants. Psychiatry Res 2020;291:113228. [doi: 10.1016/j.psychres.2020.113228] [Medline: 32562930]
10. Mayes TL, Carmody T, Rush AJ, Nandy K, Emslie GJ, Kennard BD, et al. Predicting suicidal events: a comparison of the Concise Health Risk Tracking Self-Report (CHRT-SR) and the Columbia Suicide Severity Rating Scale (C-SSRS). Psychiatry Res 2023;326:115306 [FREE Full text] [doi: 10.1016/j.psychres.2023.115306] [Medline: 37364504]

XSL•FO
RenderX

11. Kim N, Song JY, Yang H, Kim MJ, Lee K, Shin YH, et al. National trends in suicide-related behaviors among youths between 2005-2020, including COVID-19: a Korean representative survey of one million adolescents. Eur Rev Med Pharmacol Sci 2023;27(3):1192-1202 [FREE Full text] [doi: 10.26355/eurrev_202302_31226] [Medline: 36808368]
12. Park H, Lee K. Using boosted machine learning to predict suicidal ideation by socioeconomic status among adolescents. J Pers Med 2022;12(9):1357 [FREE Full text] [doi: 10.3390/jpm12091357] [Medline: 36143142]
13. Czyz EK, Koo HJ, Al-Dajani N, King CA, Nahum-Shani I. Predicting short-term suicidal thoughts in adolescents using machine learning: developing decision tools to identify daily level risk after hospitalization. Psychol Med 2023;53(7):2982-2991 [FREE Full text] [doi: 10.1017/S0033291721005006] [Medline: 34879890]
14. Weller O, Sagers L, Hanson C, Barnes M, Snell Q, Tass ES. Predicting suicidal thoughts and behavior among adolescents using the risk and protective factor framework: a large-scale machine learning approach. PLoS One 2021;16(11):e0258535 [FREE Full text] [doi: 10.1371/journal.pone.0258535] [Medline: 34731169]
15. Oh J, Lee M, Lee H, Yang H, Park J, Rahmati M, et al. Hand and oral hygiene practices of South Korean adolescents before and during the COVID-19 pandemic. JAMA Netw Open 2023;6(12):e2349249 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.49249] [Medline: 38147331]
16. Oh J, Kim M, Rhee SY, Rahmati M, Koyanagi A, Smith L, et al. National trends in the prevalence of screen time and its association with biopsychosocial risk factors among korean adolescents, 2008-2021. J Adolesc Health 2024;74(3):504-513. [doi: 10.1016/j.jadohealth.2023.10.021] [Medline: 38085201]
17. Wilimitis D, Turer RW, Ripperger M, McCoy AB, Sperry SH, Fielstein EM, et al. Integration of face-to-face screening with real-time machine learning to predict risk of suicide among adults. JAMA Netw Open 2022;5(5):e2212095 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.12095] [Medline: 35560048]
18. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
19. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods 2016;13(9):703-704 [FREE Full text] [doi: 10.1038/nmeth.3968]
20. Kwon R, Lee H, Kim MS, Lee J, Yon DK. Machine learning-based prediction of suicidality in adolescents during the COVID-19 pandemic (2020-2021): derivation and validation in two independent nationwide cohorts. Asian J Psychiatr 2023;88:103704. [doi: 10.1016/j.ajp.2023.103704] [Medline: 37541104]
21. Kim J, Kim SC, Kang D, Yon DK, Kim JG. Classification of Alzheimer's disease stage using machine learning for left and right oxygenation difference signals in the prefrontal cortex: a patient-level, single-group, diagnostic interventional trial. Eur Rev Med Pharmacol Sci 2022;26(21):7734-7741 [FREE Full text] [doi: 10.26355/eurrev_202211_30122] [Medline: 36394721]
22. Lee SW. Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee. Life Cycle 2022;2:e1 [FREE Full text] [doi: 10.54724/lc.2022.e1]
23. Kim MS, Lee H, Lee SW, Kwon R, Rhee SY, Lee JA, et al. Long-term autoimmune inflammatory rheumatic outcomes of COVID-19 : a binational cohort study. Ann Intern Med 2024;177(3):291-302 [FREE Full text] [doi: 10.7326/M23-1831] [Medline: 38437702]
24. Yoo HW, Jin HY, Yon DK, Effenberger M, Shin YH, Kim SY, et al. Non-alcoholic fatty liver disease and COVID-19 susceptibility and outcomes: a Korean nationwide cohort. J Korean Med Sci 2021;36(41):e291 [FREE Full text] [doi: 10.3346/jkms.2021.36.e291] [Medline: 34697932]
25. Lee SW, Kim SY, Moon SY, Yang JM, Ha EK, Jee HM, et al. Estimating COVID-19 infection and severity risks in patients with chronic rhinosinusitis: a Korean nationwide cohort Study. J Allergy Clin Immunol Pract 2021;9(6):2262-2271.e2 [FREE Full text] [doi: 10.1016/j.jaip.2021.03.044] [Medline: 33931377]
26. Barish M, Bolourani S, Lau LF, Shah S, Zanos TP. External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. Nat Mach Intell 2020;3(1):25-27 [FREE Full text] [doi: 10.1038/s42256-020-00254-2]
27. Chung H, Ko H, Lee H, Yon DK, Lee WH, Kim T, et al. Development and validation of a deep learning model to diagnose COVID-19 using time-series heart rate values before the onset of symptoms. J Med Virol 2023;95(2):e28462. [doi: 10.1002/jmv.28462] [Medline: 36602055]
28. Kim J, Lee H, Lee J, Rhee SY, Shin JI, Lee SW, et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning: a post hoc analysis of a diagnostic trial and validation of an external additional trial. Alzheimers Res Ther 2023;15(1):127 [FREE Full text] [doi: 10.1186/s13195-023-01268-9] [Medline: 37481573]
29. Chen H, Lundberg SM, Lee SI. Explaining a series of models by propagating Shapley values. Nat Commun 2022;13(1):4512 [FREE Full text] [doi: 10.1038/s41467-022-31384-3] [Medline: 35922410]
30. Probability for suicidal thinking. Streamlit. URL: https://suicidalthinkingpredict3validation.streamlit.app/ [accessed 2024-04-28]
31. CenterForDH/suicidality. GitHub. URL: https:// github.com/CenterForDH/suicidalthinking_predict_3validation [accessed 2024-04-28]

XSL•FO
RenderX

32. Malhi GS, Mann JJ. Depression. Lancet 2018;392(10161):2299-2312. [doi: 10.1016/S0140-6736(18)31948-2] [Medline: 30396512]
33. Kolk SM, Rakic P. Development of prefrontal cortex. Neuropsychopharmacology 2022;47(1):41-57 [FREE Full text] [doi: 10.1038/s41386-021-01137-9] [Medline: 34645980]
34. Straub J, Brown R, Malejko K, Bonenberger M, Grön G, Plener PL, et al. Adolescent depression and brain development: evidence from voxel-based morphometry. J Psychiatry Neurosci 2019;44(4):237-245 [FREE Full text] [doi: 10.1503/jpn.170233] [Medline: 30720261]
35. Khasawneh A, Madathil KC, Dixon E, Wiśniewski P, Zinzow H, Roth R. Examining the self-harm and suicide contagion effects of the blue whale challenge on YouTube and Twitter: qualitative study. JMIR Ment Health 2020;7(6):e15973 [FREE Full text] [doi: 10.2196/15973] [Medline: 32515741]
36. Shapero BG, Farabaugh A, Terechina O, DeCross S, Cheung JC, Fava M, et al. Understanding the effects of emotional reactivity on depression and suicidal thoughts and behaviors: moderating effects of childhood adversity and resilience. J Affect Disord 2019;245:419-427. [doi: 10.1016/j.jad.2018.11.033] [Medline: 30423470]
37. Gruhn MA, Compas BE. Effects of maltreatment on coping and emotion regulation in childhood and adolescence: a meta-analytic review. Child Abuse Negl 2020;103:104446. [doi: 10.1016/j.chiabu.2020.104446] [Medline: 32200195]
38. Häusl AS, Brix LM, Hartmann J, Pöhlmann ML, Lopez J, Menegaz D, et al. The co-chaperone Fkbp5 shapes the acute stress response in the paraventricular nucleus of the hypothalamus of male mice. Mol Psychiatry 2021;26(7):3060-3076 [FREE Full text] [doi: 10.1038/s41380-021-01044-x] [Medline: 33649453]
39. Russell G, Lightman S. The human stress response. Nat Rev Endocrinol 2019;15(9):525-534. [doi: 10.1038/s41574-019-0228-0] [Medline: 31249398]
40. Thoma MV, Rohleder N, Rohner SL. Clinical ecopsychology: the mental health impacts and underlying pathways of the climate and environmental crisis. Front Psychiatry 2021;12:675936 [FREE Full text] [doi: 10.3389/fpsyt.2021.675936] [Medline: 34093283]
41. Kim Y, Choi S, Chun C, Park S, Khang YH, Oh K. Data resource profile: the Korea Youth Risk Behavior Web-based Survey (KYRBS). Int J Epidemiol 2016;45(4):1076-1076e [FREE Full text] [doi: 10.1093/ije/dyw070] [Medline: 27380796]
42. Woo HG, Park S, Yon H, Lee SW, Koyanagi A, Jacob L, et al. National trends in sadness, suicidality, and COVID-19 pandemic-related risk factors among South Korean adolescents from 2005 to 2021. JAMA Netw Open 2023;6(5):e2314838 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.14838] [Medline: 37223902]
43. Valavi R, Elith J, Lahoz‐Monfort JJ, Guillera‐Arroita G. BlockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol 2018;10(2):225-232 [FREE Full text] [doi: 10.1111/2041-210x.13107]
44. Yang J, Soltan AAS, Eyre DW, Clifton DA. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nat Mach Intell 2023;5(8):884-894 [FREE Full text] [doi: 10.1038/s42256-023-00697-3] [Medline: 37615031]
45. Dinh-Le C, Chuang R, Chokshi S, Mann D. Wearable health technology and electronic health record integration: scoping review and future directions. JMIR Mhealth Uhealth 2019;7(9):e12861 [FREE Full text] [doi: 10.2196/12861] [Medline: 31512582]
46. Rogers A, Obst S, Teague SJ, Rossen L, Spry EA, Macdonald JA, et al. Association between maternal perinatal depression and anxiety and child and adolescent development: a meta-analysis. JAMA Pediatr 2020;174(11):1082-1092 [FREE Full text] [doi: 10.1001/jamapediatrics.2020.2910] [Medline: 32926075]
47. Stanley B, Mann JJ. The need for innovation in health care systems to improve suicide prevention. JAMA Psychiatry 2020;77(1):96-98. [doi: 10.1001/jamapsychiatry.2019.2769] [Medline: 31577340]
48. Robinson J, Bailey E, Witt K, Stefanac N, Milner A, Currier D, et al. What works in youth suicide prevention? A systematic review and meta-analysis. EClinicalMedicine 2018;4-5:52-91 [FREE Full text] [doi: 10.1016/j.eclinm.2018.10.004] [Medline: 31193651]
49. Thorn P, Hill NT, Lamblin M, Teh Z, Battersby-Coulter R, Rice S, et al. Developing a suicide prevention social media campaign with young people (The #Chatsafe Project): co-design approach. JMIR Ment Health 2020;7(5):e17520 [FREE Full text] [doi: 10.2196/17520] [Medline: 32391800]

Abbreviations

AdaBoost: adaptive boosting
AUROC: area under the receiver operating characteristic
KYRBS: Korea Youth Risk Behavior Web-based Survey
LightGBM: light gradient-boosting machine
ML: machine learning
SHAP: Shapley additive explanations
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
Ungdata: University National General Survey

XSL•FO
RenderX

YRBS: Youth Risk Behavior Survey

Edited by Y Hong; submitted 29.12.23; peer-reviewed by F Alam, D Brent; comments to author 19.03.24; revised version received 24.03.24; accepted 25.03.24; published 17.05.24.

Please cite as:
Kim H, Son Y, Lee H, Kang J, Hammoodi A, Choi Y, Kim HJ, Lee H, Fond G, Boyer L, Kwon R, Woo S, Yon DK
Machine Learning–Based Prediction of Suicidal Thinking in Adolescents by Derivation and Validation in 3 Independent Worldwide Cohorts: Algorithm Development and Validation Study
J Med Internet Res 2024;26:e55913
URL: https://www.jmir.org/2024/1/e55913
doi:10.2196/55913
PMID:38758578

©Hyejun Kim, Yejun Son, Hojae Lee, Jiseung Kang, Ahmed Hammoodi, Yujin Choi, Hyeon Jin Kim, Hayeon Lee, Guillaume Fond, Laurent Boyer, Rosie Kwon, Selin Woo, Dong Keon Yon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis

Mikaël Chelli[1], MSc, MD; Jules Descamps[2], MSc, MD; Vincent Lavoué[1], MD; Christophe Trojani[1], MD, PhD; Michel Azar[1], MD; Marcel Deckert[3], PhD; Jean-Luc Raynier[1], MSc, MD; Gilles Clowez[1], MD; Pascal Boileau[1], MD, PhD; Caroline Ruetsch-Chelli[3], MSc, MD

[1]Institute for Sports and Reconstructive Bone and Joint Surgery, Groupe Kantys, Nice, France

[2]Orthopedic and Traumatology Unit, Hospital Lariboisière, Assistance Publique-Hôpitaux de Paris, Paris, France

[3]Université Côte d'Azur, INSERM, C3M, Team Microenvironment, Signalling and Cancer, Nice, France

Corresponding Author:
Mikaël Chelli, MSc, MD
Institute for Sports and Reconstructive Bone and Joint Surgery
Groupe Kantys
7 Avenue, Durante
Nice, 06000
France
Phone: 33 4 93 16 76 40
Email: mikael.chelli@gmail.com

Abstract

Background: Large language models (LLMs) have raised both interest and concern in the academic community. They offer the potential for automating literature search and synthesis for systematic reviews but raise concerns regarding their reliability, as the tendency to generate unsupported (hallucinated) content persist.

Objective: The aim of the study is to assess the performance of LLMs such as ChatGPT and Bard (subsequently rebranded Gemini) to produce references in the context of scientific writing.

Methods: The performance of ChatGPT and Bard in replicating the results of human-conducted systematic reviews was assessed. Using systematic reviews pertaining to shoulder rotator cuff pathology, these LLMs were tested by providing the same inclusion criteria and comparing the results with original systematic review references, serving as gold standards. The study used 3 key performance metrics: recall, precision, and F_1-score, alongside the hallucination rate. Papers were considered "hallucinated" if any 2 of the following information were wrong: title, first author, or year of publication.

Results: In total, 11 systematic reviews across 4 fields yielded 33 prompts to LLMs (3 LLMs×11 reviews), with 471 references analyzed. Precision rates for GPT-3.5, GPT-4, and Bard were 9.4% (13/139), 13.4% (16/119), and 0% (0/104) respectively (P<.001). Recall rates were 11.9% (13/109) for GPT-3.5 and 13.7% (15/109) for GPT-4, with Bard failing to retrieve any relevant papers (P<.001). Hallucination rates stood at 39.6% (55/139) for GPT-3.5, 28.6% (34/119) for GPT-4, and 91.4% (95/104) for Bard (P<.001). Further analysis of nonhallucinated papers retrieved by GPT models revealed significant differences in identifying various criteria, such as randomized studies, participant criteria, and intervention criteria. The study also noted the geographical and open-access biases in the papers retrieved by the LLMs.

Conclusions: Given their current performance, it is not recommended for LLMs to be deployed as the primary or exclusive tool for conducting systematic reviews. Any references generated by such models warrant thorough validation by researchers. The high occurrence of hallucinations in LLMs highlights the necessity for refining their training and functionality before confidently using them for rigorous academic purposes.

(J Med Internet Res 2024;26:e53164) doi:10.2196/53164

KEYWORDS

artificial intelligence; large language models; ChatGPT; Bard; rotator cuff; systematic reviews; literature search; hallucinated; human conducted

XSL•FO
RenderX

Introduction

The advent of artificial intelligence (AI) has led to significant advancements in various fields, including medical research. Large language models (LLMs), such as ChatGPT (OpenAI), could assist academic researchers in a variety of tasks, including writing scientific papers. These models have the potential to streamline the way researchers conduct literature searches, synthesize findings, and draft systematic reviews [1]. However, there is ongoing debate surrounding their reliability, ethical considerations, and appropriate use in academic publishing.

Recently, editorials and opinion papers have been published addressing the use of LLMs in the scientific community. One such example is an editorial in *The Lancet Digital Health*, which discusses the potential benefits and challenges of implementing AI in medical research [2]. As the application of LLMs such as ChatGPT in research settings grows, concerns have arisen regarding their accuracy, the potential for generating misleading or false information, and the ethical implications of using AI-generated content without proper disclosure.

While it is known that ChatGPT can help researchers write papers [3-5], controversy exists about whether it should be used at all, whether its use should be disclosed, and whether it should be listed as an author or not [6]. These debates raise important questions about the role of AI in scientific research and the potential consequences of using LLMs in generating systematic reviews and other research outputs [7].

In this study, we aim to address these concerns by systematically evaluating the reliability of ChatGPT and Bard (subsequently rebranded Gemini; Google AI) [8] in the context of searching for and synthesizing peer-reviewed literature for systematic reviews. We will compare their performance to that of traditional methods used by researchers, investigate the extent of the "hallucination" phenomenon, and discuss potential ethical and practical considerations for using ChatGPT and Bard in academic publishing. By providing evidence-based insights into the capabilities and limitations of LLMs in medical research, we hope to contribute to the ongoing debate about the role of AI in the research ecosystem and guide researchers in making informed decisions about using LLMs in their work.

Methods

Ethical Considerations

Ethics approval is not required, as human participants were not involved in this research. Consent for publication has been provided from all identifiable persons in the figures.

Study Design

This study follows a sequential design, chosen for its ability to progressively build on each preceding phase, thus ensuring a comprehensive evaluation of the LLMs in the context of a systematic review. The process initiated with a systematic review search on PubMed, followed by the retrieval of selected papers. Subsequently, the methodology of these papers served as inputs to the LLM, which is tasked to search for papers using the same inclusion criteria as the systematic reviews. The final phase involves a comparison of the LLM results with the systematic review references, which act as the ground truth, thus providing a robust evaluation of the LLMs' ability to replicate the results of human-conducted systematic reviews. The ethical considerations of using AI, specifically LLMs, in research were carefully evaluated.

Systematic Review Search on PubMed

On July 27, 2023, a literature search was performed on PubMed to find literature published in the English language during 2020. The selected year aligns with ChatGPT's training cut-off point in September 2021, ensuring that the AI model has access to the comprehensive scope of literature for the given year. The focus was directed toward systematic reviews of randomized clinical trials pertaining to shoulder rotator cuff pathology. This prevalent condition spans multiple disciplines inclusive of surgery, anesthesiology, sports medicine, and physical therapy, thereby positioning it as an optimal candidate for this multidisciplinary appraisal. In addition, the collective clinical and scientific experience of the research team on the topic furnished a critical review of the references obtained from the PubMed search and the LLMs [9-12].

An electronic search of PubMed was conducted using a combination of keywords, including "shoulder," "rotator cuff," and "randomized" (Multimedia Appendix 1). The search was restricted to papers published in 2020 and filtered to retrieve only systematic reviews and meta-analyses. Titles and abstracts were scrutinized, and papers indicating a systematic review of randomized studies on rotator cuff pathology were selected for further analysis.

Exclusion criteria were applied to eliminate papers that did not meet our study focus. Papers were excluded if they were not systematic reviews, if their primary concern did not pertain to rotator cuff pathology, if written in a language other than English, or if they included nonrandomized clinical studies.

Two independent reviewers (MC and PB) screened titles, abstracts, and full texts retrieved by this query. Differences between reviewers were reconciled with a third reviewer (JD). To ensure the selection of relevant systematic reviews, the reviewers applied exclusion criteria that consisted of systematic reviews including nonrandomized studies and papers that were not systematic reviews. The eligibility of the selected systematic reviews was further validated by assessing their adherence to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [13]. Additionally, the registration status of these reviews was verified in the PROSPERO database [14].

For each paper referenced in the systematic reviews, information on the paper title, author list, country (based on the first author's affiliation with PubMed), journal name, journal date and issue, DOI, and open access status was collected. We assessed the hypothesis that LLMs may favor publicly available papers in their results by using a broad definition of "open access." This definition included open access through the journal or any full-text PDF available on another server and accessible through a Google search (eg, ResearchGate or university website).

XSL•FO
RenderX

Systematic Review on LLMs: Paper Retrieval

For each new request, a fresh chatbot session was initiated to prevent any carryover effect from previous queries, ensuring the validity of the results. We prompted ChatGPT and Bard with a precise query to identify papers that could be included in the systematic review. The structure of the prompt consisted of a statement about the physician's and researcher's current work, followed by the inclusion criteria for the studies in the review (Figure 1). The criteria specified randomized controlled trials with specific participant criteria and interventions comparing 2 different treatments. LLMs were asked to provide references to randomized studies on the topic, excluding papers published after 2020 and systematic reviews or meta-analyses. To assess the impact of the prompt's specificity on the search results of LLMs, we tested 2 versions of the prompt for each request. One specifying the minimum number of papers to be found and the other without specifying this minimum number, thus providing us with an opportunity to assess if the presence or absence of a target number influences the LLMs' search results. The query that led to the largest number of results was retained for this study.

Figure 1. Captured screenshots demonstrating a prompt to large language models.

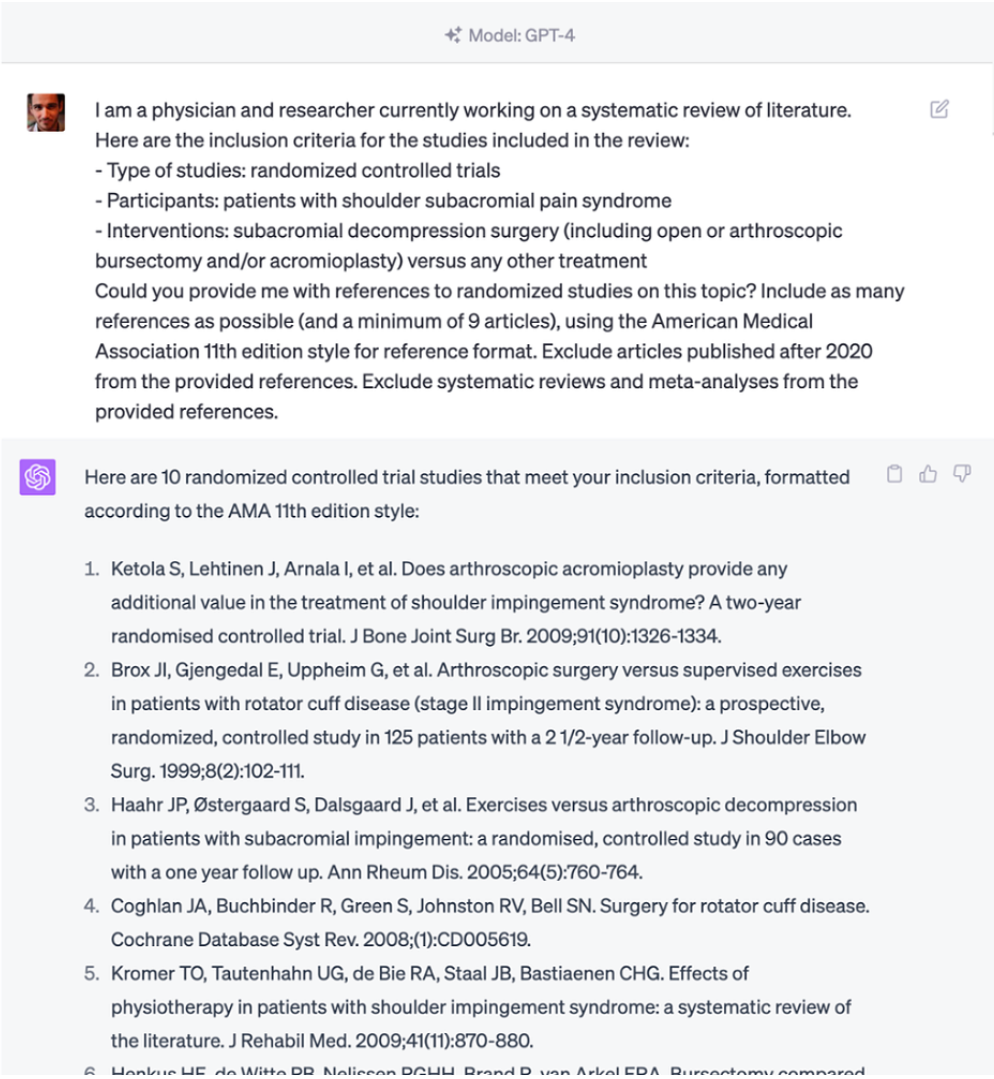

For each paper provided by LLMs, information on the existence or hallucination status of the paper, authors' list, country (based on the first author's affiliation on PubMed), open-access status, inclusion in the original systematic review, randomization status, participant criteria adherence, intervention criteria adherence, exclusion of systematic reviews (as requested in the prompt), and accuracy of the provided information (authors' list, journal, year and issue, title, and DOI) was collected. We also verified if the paper was published before 2021, as requested in the prompt.

Papers were considered hallucinated if any 2 of the following information were wrong: title, first author, or year of publication. The hallucination rate was calculated to quantify the proportion of LLM-generated references that were irrelevant, incorrect, or unsupported by the available literature, offering insights into the extent of spurious or inaccurate information production by the LLMs.

For noncomparative studies, the intervention criteria were considered adequate if at least 1 of the 2 interventions was studied in the proposed reference. For comparative studies, the intervention criteria were considered adequate if both interventions were studied in the proposed reference.

XSL•FO
RenderX

Comparison of LLMs Results

The sample size was determined based on an anticipated 10% rate of systematic review references overlooked by LLMs, with an assumed power of 90% and an α of .05. This calculation yielded a requisite of 80 references for the comparison. The PubMed search yielded 11 systematic reviews (Figure 2), each with an average of 9.9 (SD 6.6; range 3-23) references. The evaluation of the LLMs was predicated on three widely used metrics: (1) recall, representing the proportion of genuinely pertinent papers from the original systematic reviews accurately identified and retrieved by the LLMs; (2) precision, quantifying the proportion of papers retrieved by the LLMs that are verifiably present in the original systematic reviews; and (3) F_1-score, which serves as an aggregate metric encapsulating both the recall and precision values (Table 1).

Figure 2. Flow diagram of included systematic reviews.

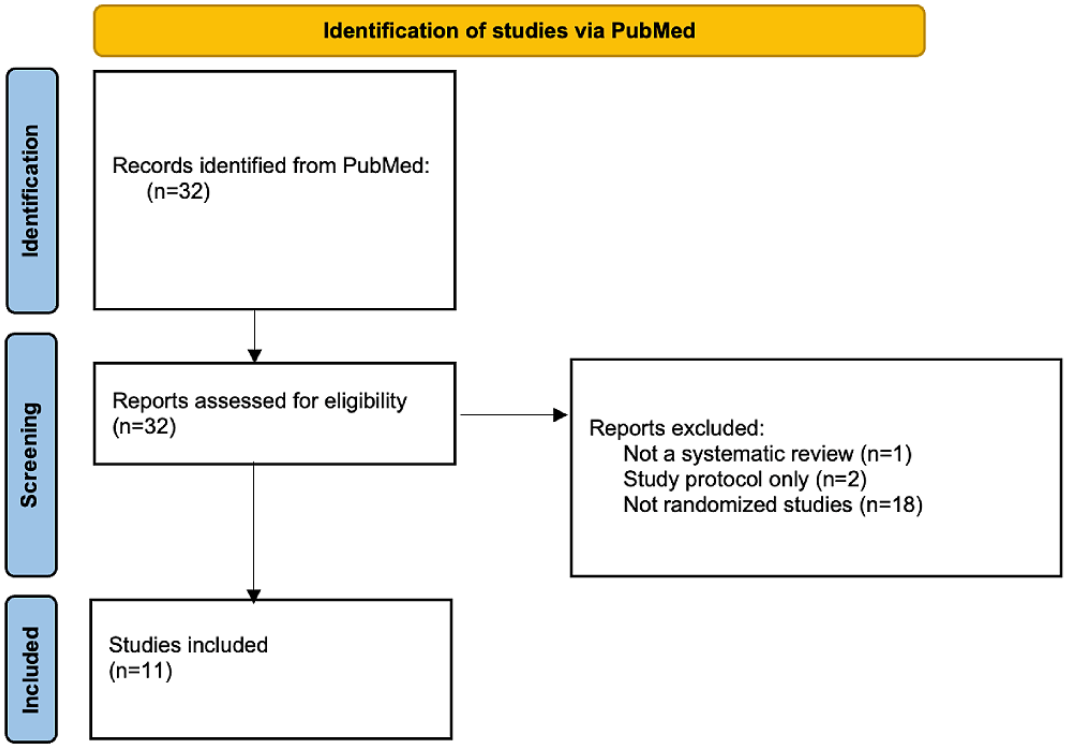

Table 1. Recall, precision, and F1-score.

	Papers provided as an output by LLMs[a]	Papers not provided as an output by LLMs
Papers cited by systematic reviews	True positive	False negative
Papers not cited by systematic reviews	False positive	True negative

[a]LLM: large language model.

where TP represents true positive, FN represents false negative, and FP represents false positive.

The LLMs incorporated in this study included GPT-3.5 (text-davinci-002-render-sha, July 19 version; OpenAI), GPT-4 (gpt-4-32k-0314, July 19 version; OpenAI) [15], and Bard (PaLM version 2.0, released on July 13, 2023; Google AI). We conducted chi-square tests to compare each piece of information extracted from LLMs' responses, including authors' nationalities and the open-access status of the retrieved papers. The significance threshold used was $P<.05$. Statistical analysis was performed with EasyMedStat (version 3.24).

Results

In total, 11 systematic reviews were identified in 4 fields (Table 2): physiotherapy (3 papers), sports medicine (3 papers), orthopedic surgery (3 papers), and anesthesiology (2 papers), leading to 33 prompts to LLMs (3 tested LLMs×11 systematic reviews). LLM prompts returned references in 32 of 33 cases: Bard did not return any result for the systematic review about "subacromial analgesia via continuous infusion catheter." In most cases, the number of references returned by LLMs was greater or equal to that of the original papers (Table 2). Overall, 471 references were included in this study and analyzed.

Papers identified by LLMs were present in the original systematic reviews (precision) in 9.4% (13/139), 13.4%

XSL•FO
RenderX

(16/119), and 0% (0/104) of cases for GPT-3.5, GPT-4, and Bard (*P*<.001), respectively. Conversely, 11.9% (13/109) of papers from the systematic reviews (recall) were retrieved by GPT-3.5, and 13.7% (15/109) by GPT-4. No paper from the systematic reviews was retrieved by Bard (*P*<.001; Table 3).

The hallucination rates were, respectively, 39.6% (55/139), 28.6% (34/119), and 91.4% (95/104) for GPT-3.5, GPT-4, and Bard (*P*<.001). When analyzing the papers retrieved by GPT that were not hallucinated (n=84 for GPT-3.5 and n=85 for GPT-4), the following criteria were successfully identified (Figure 3): randomized studies (33/84, 39% vs 42/85, 49%; *P*=.24), participant criteria (49/84, 57% vs 57/85, 67%; *P*=.24), intervention criteria (58/84, 69% vs 72/85, 85%; *P*=.03), not a systematic review (69/84, 81% vs 66/85, 78%; *P*=.73), and published before 2021 (84/84, 100% vs 85/85, 100%; *P*>.99). In total, 9 papers retrieved by Bard were not hallucinated. This limited sample was not appropriate for further inferential statistics.

Regarding the same nonhallucinated papers retrieved by GPT, the following bibliographic information were considered accurate (Figure 4): authors list (73/84, 87% vs 74/85, 87%; *P*>.99), journal title (81/84, 96% vs 85/85, 100%; *P*=.12), date and issue (71/84, 84% vs 81/85, 95%; *P*=.02), paper title (83/84, 99% vs 84/85, 99%; *P*>.99), and DOI (13/82, 16% vs 17/84, 20%; *P*=.59).

Open-access papers were selected in 27.5% (30/109) of original systematic reviews, 38% (32/84) of GPT-3.5 papers, and 36% (31/85) of GPT-4 papers (*P*=.24). Papers from American authors were selected in 16.5% (18/109) of original systematic reviews, 44% (37/84) of GPT-3.5 papers, and 33% (28/85) of GPT-4 papers (*P*<.001).

Table 2. Systematic reviews included in the study and the count of papers retrieved by original authors and large language models.

Systematic review	Field	PRISMA[a] guidelines	PROSPERO registration	Papers in the original paper, n	Papers returned by GPT-3.5, n	Papers returned by GPT-4, n	Papers returned by Bard, n
Lähdeoja et al [16]	Surgery	Yes	Yes	9	10	10	9
Catapano et al [17]	Sports medicine	Yes	No	5	7	5	5
Gutiérrez-Espinoza et al [18]	Physiotherapy	Yes	Yes	7	15	7	7
Chen et al [19]	Sports medicine	Yes	No	18	18	18	16
An et al [20]	Anesthesiology	Yes	Yes	9	9	9	0
Craig et al [21]	Surgery	Yes	No	23	22	22	23
Naunton et al [22]	Physiotherapy	Yes	Yes	7	10	8	7
Malliaras et al [23]	Physiotherapy	Yes	Yes	3	5	7	5
Simpson et al [24]	Sports medicine	Yes	Yes	18	18	18	23
Belk et al [25]	Anesthesiology	Yes	No	5	10	8	5
Belk et al [8]	Surgery	Yes	No	5	15	7	5
Total		11/11	6/11	109	139	119	104

[a]PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Table 3. Evaluative metrics of the assessed large language models.

Metric	GPT-3.5	GPT-4	Bard
True positive	13	16	0
False positive	126	103	104
False negative	96	93	109
Recall (%)	11.9	13.7	0
Precision (%)	9.4	13.4	0
F_1-score (%)	10.5	14	0

XSL•FO
RenderX

Figure 3. Efficiency of the tested large language models in complying to inclusion and exclusion criteria. With the exception of the "paper exists" criteria, hallucinated papers were excluded from this analysis.

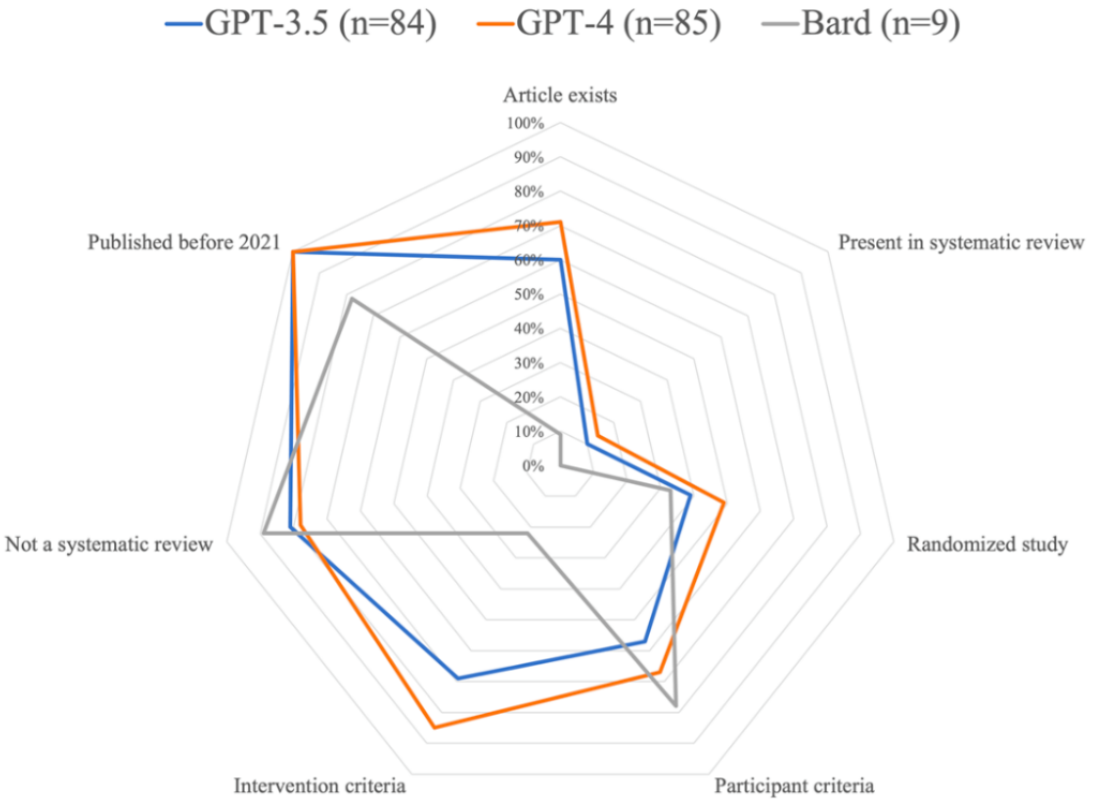

Figure 4. Efficiency of the tested large language models in generating accurate bibliographic information of the retrieved papers. Hallucinated papers were excluded from this analysis.

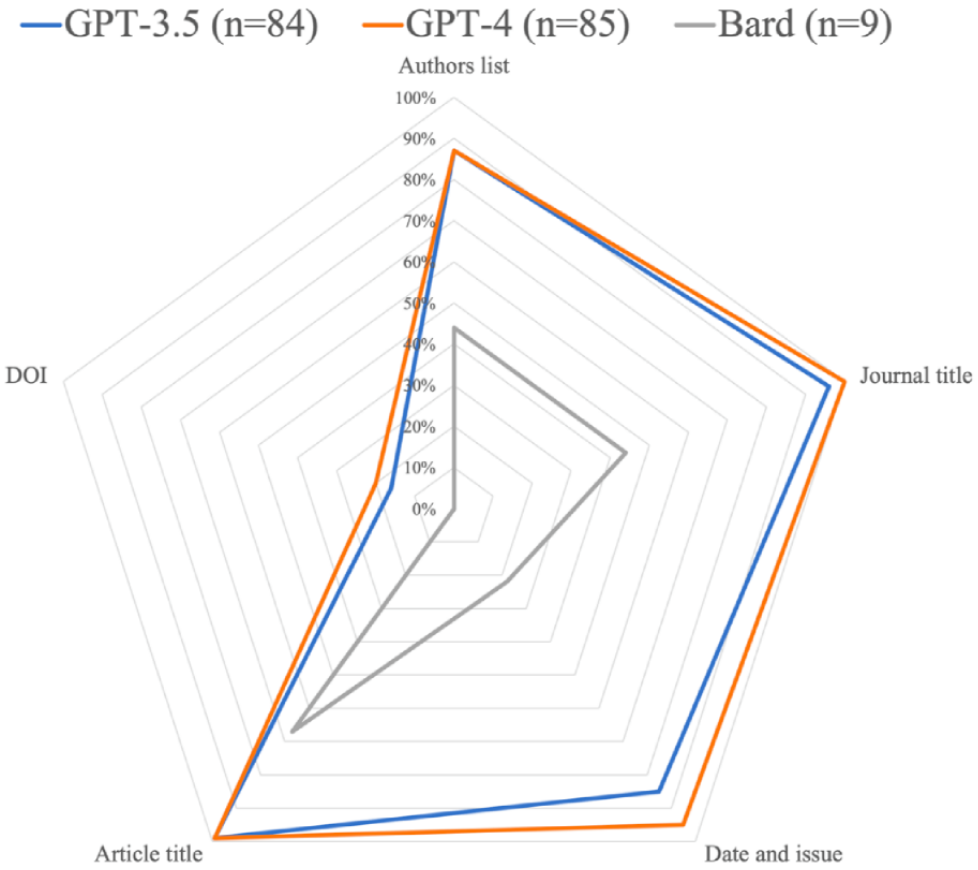

XSL•FO
RenderX

Discussion

Principal Findings

The most important finding of this study is that using LLMs such as ChatGPT and Bard to conduct systematic reviews for a common condition such as rotator cuff disease can generate misleading or "hallucinated" references, exceeding a 25% rate.

This concern has been broached in previous literature [26-29], but our study provides an experimental design to probe the matter more deeply. OpenAI, the developer of ChatGPT, acknowledges this issue, stating that their model "occasionally generates plausible but incorrect or nonsensical responses" [30]. As LLMs increasingly assist academic researchers in producing scientific literature, this phenomenon warrants careful scrutiny.

When comparing the 3 models tested, GPT-4 was the most efficient to retrieve nonhallucinated references, while GPT-3.5 produced 39.6% (55/139) of nonexisting references. Bard, however, appears ill-suited for conducting systematic reviews in the selected areas, with 91.3% (95/104) of the references failing to correlate with legitimate papers. Bard seemed to have a try-and-repeat approach, providing multiple versions of hallucinated papers with close titles and journal names (Figure 5).

Despite this, LLMs typically encouraged users to conduct their own systematic reviews, recognizing the necessity of human involvement. However, in none of our queries did the LLMs ask to verify the authenticity of the produced citations. Nonetheless, the convincing verisimilitude of the references generated by LLMs presents a risk for incautious researchers, potentially undermining the quality of scientific bibliographies if improperly used (Figures 5 and 6). Moreover, the efficiency of LLMs in retrieving original papers from systematic reviews ranged from negligible to modest (0/109, 0% to 15/109, 13.8%), emphasizing that researchers should not overly rely on these tools for systematic reviews. Nevertheless, in numerous instances, both ChatGPT and Bard "encouraged [users] to conduct their own research" (Figure 5), a suggestion that appears crucial considering the findings of this study.

It could be expected that LLMs were not able to retrieve the same references as authors of systematic reviews. However, this study also reveals that LLMs, despite being provided with the same eligibility criteria as those in the original systematic reviews, were not able to consistently apply them. For instance, the criterion of "randomized study" was adhered to in only 39% (33/84) to 49% (42/85) of nonhallucinated papers generated by ChatGPT, even when the term "randomized" appeared in the title or abstract of the papers from the original systematic reviews. The same finding was observed for the "not a systematic review" criterion, which was not respected in 20.1% (36/179) of cases, while the publicly available information of the produced papers clearly states the nature of these studies.

These discrepancies could potentially stem from the underlying statistical nature of these LLMs, which predict subsequent text (tokens) based on a model reinforced by human feedback [31]. However, as human supervision does not extend to validating the accuracy of LLM outputs, especially in specialized fields like medicine, inaccuracies can prevail.

In the case of nonhallucinated papers, however, ChatGPT demonstrated significant efficiency in retrieving accurate bibliographic information like the exact paper title, the authors' list, and the journal title.

Potential biases in LLMs due to training on biased data sets and the risk of perpetuating stereotypes have been highlighted [2]. Our findings suggest that American authors were more frequently represented in ChatGPT references. However, further investigation across diverse medical fields is warranted to ascertain whether these LLMs may introduce such biases definitively.

Figure 5. Instances of hallucinated papers. In total, 4 of 5 paper titles commence with "Tranexamic acid for the prevention of bleeding in arthroscopic" and were allegedly published in the journal Arthroscopy.

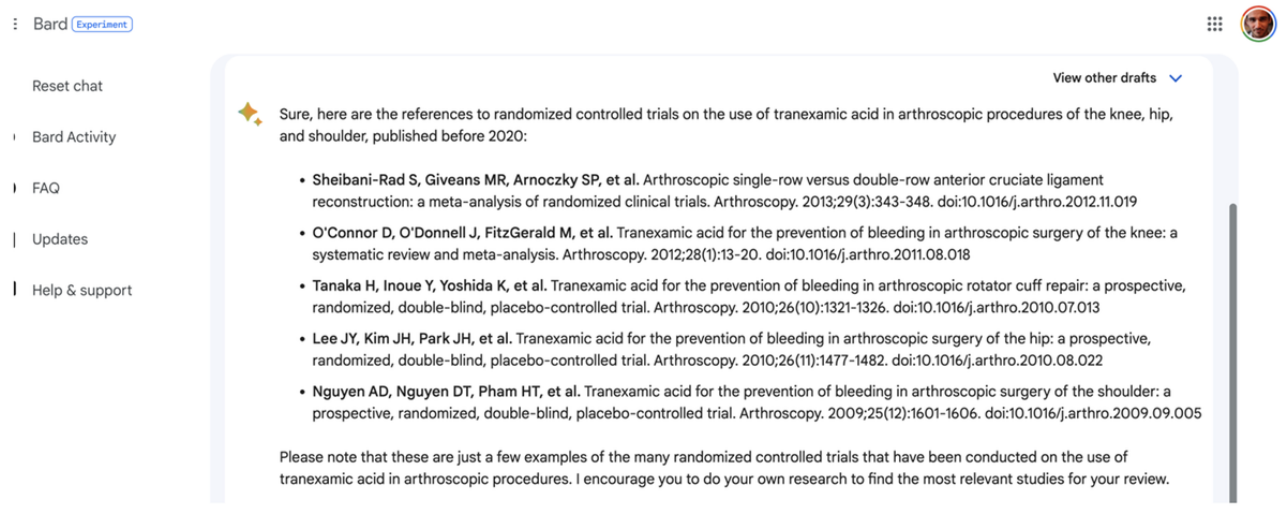

XSL•FO
RenderX

Figure 6. Instance of a hallucinated reference. (A) The output of a large language model. (B and C) Authentic papers with similarities in title and author list, potentially serving as original data for large language model reference generation.

Strengths and Limitations

This investigation, by virtue of its specific and circumscribed parameters, comes with several inherent limitations. The scope of the study was exclusively focused on systematic reviews related to shoulder rotator cuff pathology. Consequently, it must be recognized that the findings might not be universally applicable across diverse medical specialties or disciplines. The examination was also restricted to 3 LLMs, specifically GPT-3.5, GPT-4, and Bard. The landscape of available language models is vast and continually evolving, and it is conceivable that different models might yield divergent results. In addition, the field lacks established guidelines for leveraging LLMs to optimize accuracy. Notwithstanding rigorous attempts to devise specific, comprehensive prompts, it remains plausible that alternative queries could generate more precise outcomes. This fact underscores the multifaceted nature of the challenge and the need for further research in this domain.

The choice of prompt plays a crucial role in determining the output generated by LLMs. During the exploratory phase of our study, various prompt versions were tested. While our study did not focus on identifying the optimal prompts, several techniques used in our prompts appeared to enhance output quality: specifying a minimum number of papers (a minimum of 9 papers); using bullet points to delineate criteria such as “type of studies,” “participants,” and “interventions”; and explicitly instructing to “exclude systematic reviews and meta-analyses.” Introducing prompts by specifying the researcher’s profession provides additional context, aligning with recommendations from LLM providers. Finally, enforcing a specific reference style format facilitated the retrieval of vital information, including authors’ names, journal titles, publication dates, and DOIs when available.

Our decision not to provide the initial PubMed results list to LLMs for assessing paper eligibility was deliberate, aimed at preserving study integrity and interpretability. While providing the list might enhance LLM accuracy, it introduces bias by guiding models toward replicating the provided set rather than autonomously identifying relevant studies. Our study design, though sacrificing some precision, ensures that LLM results reflect genuine capabilities in navigating scientific literature independently.

Future Directions

LLMs present a highly efficient instrument that may aid academics in the drafting of research papers. However, upon analyzing the findings of this study, it becomes imperative to emphasize that the bibliographic references proposed by the AI are not intrinsically trustworthy. These citations necessitate human validation, focusing on the authors, the title, and the subject matter.

We thereby deduce that, in the context of GPT iterations, user verification is indispensable for preserving the scientific integrity and relevance of the output. A statement or a scholarly usage guideline should be prominently featured before the tool is used or should be integrated into the software itself to outline its lack of liability for any inaccuracies in the citation of papers. This is paramount as such errors could potentially mislead a considerable number of users. We also propose that the application of GPT-based chatbots for tasks such as spelling correction, proofreading, or text restructuring ought to be explicitly mentioned within the materials and methods section of academic writings.

Conclusions

ChatGPT and Bard exhibit the capacity to generate convincingly authentic references for systematic reviews but also yield

XSL•FO
RenderX

hallucinated papers in 28.6% (34/119) to 91.3% (95/104) of cases. Among the models tested, GPT-4 displayed superior performance in generating legitimate and relevant references but, like the other models, largely failed to respect the established eligibility criteria. Given their current state, LLMs such as ChatGPT and Bard should not be used as the sole or primary means for conducting systematic reviews of literature, and it is crucial that references generated by these tools undergo rigorous validation by the authors of scientific papers.

Acknowledgments

The entirety of this paper was composed by the authors of this research. As nonnative English speakers, the authors used ChatGPT to refine the English language used in the paper [32]. Importantly, all modifications suggested by ChatGPT underwent meticulous evaluation and approval by the authors to ensure accuracy and clarity. ChatGPT was not used for bibliographic reference retrieval.

Data Availability

All papers accessed by the large language models (LLMs) were publicly available, and no proprietary or subscription-based sources were used without appropriate access rights. LLMs did not have access to any sensitive or private patient data. The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MC and PB screened titles, abstracts, and full texts retrieved by PubMed search. Differences between reviewers were reconciled with a third reviewer (JD). CR-C designed the study and drafted and critically reviewed the paper. MC, JD, VL, CT, MA, MD, JLR, GC, PB, and CR-C confirmed that they had full access to all the data in the study and accept responsibility to submit for publication and commented on, revised, and approved the final version of the paper.

Conflicts of Interest

MC is a co-founder and shareholder of EasyMedStat.

Multimedia Appendix 1
Search strategy.
[DOCX File , 13 KB - jmir_v26i1e53164_app1.docx]

References

1. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Lib Hi Tech News 2023;40(3):26-29. [doi: 10.1108/lhtn-01-2023-0009]
2. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023;5(3):e102 [FREE Full text] [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
3. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
4. Biswas S. ChatGPT and the future of medical writing. Radiology 2023;307(2):e223312 [FREE Full text] [doi: 10.1148/radiol.223312] [Medline: 36728748]
5. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care 2023;27(1):75 [FREE Full text] [doi: 10.1186/s13054-023-04380-2] [Medline: 36841840]
6. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023;613(7945):620-621. [doi: 10.1038/d41586-023-00107-z] [Medline: 36653617]
7. Zheng H, Zhan H. ChatGPT in scientific writing: a cautionary tale. Am J Med 2023;136(8):725-726.e6. [doi: 10.1016/j.amjmed.2023.02.011] [Medline: 36906169]
8. Belk JW, Kraeutler MJ, Houck DA, Chrisman AN, Scillia AJ, McCarty EC. Biceps tenodesis versus tenotomy: a systematic review and meta-analysis of level I randomized controlled trials. J Shoulder Elbow Surg 2021;30(5):951-960. [doi: 10.1016/j.jse.2020.11.012] [Medline: 33373685]
9. Azar M, Van der Meijden O, Pireau N, Chelli M, Gonzalez JF, Boileau P. Arthroscopic revision cuff repair: do tendons have a second chance to heal? J Shoulder Elbow Surg 2022;31(12):2521-2531. [doi: 10.1016/j.jse.2022.04.024] [Medline: 35671929]
10. Boileau P, Andreani O, Schramm M, Baba M, Barret H, Chelli M. The effect of tendon delamination on rotator cuff healing. Am J Sports Med 2019;47(5):1074-1081. [doi: 10.1177/0363546519835491] [Medline: 30943087]
11. Muccioli C, Chelli M, Caudal A, Andreani O, Elhor H, Gauci MO, et al. Rotator cuff integrity and shoulder function after intra-medullary humerus nailing. Orthop Traumatol Surg Res 2020;106(1):17-23 [FREE Full text] [doi: 10.1016/j.otsr.2019.11.004] [Medline: 31882328]

XSL•FO
RenderX

12. Boileau P, Baqué F, Valerio L, Ahrens P, Chuinard C, Trojani C. Isolated arthroscopic biceps tenotomy or tenodesis improves symptoms in patients with massive irreparable rotator cuff tears. J Bone Joint Surg Am 2007;89(4):747-757. [doi: 10.2106/JBJS.E.01097] [Medline: 17403796]
13. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
14. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev 2012;1:2 [FREE Full text] [doi: 10.1186/2046-4053-1-2] [Medline: 22587842]
15. OpenAI. GPT-4 technical report. ArXiv Preprint posted online on March 15, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
16. Lähdeoja T, Karjalainen T, Jokihaara J, Salamh P, Kavaja L, Agarwal A, et al. Subacromial decompression surgery for adults with shoulder pain: a systematic review with meta-analysis. Br J Sports Med 2020;54(11):665-673 [FREE Full text] [doi: 10.1136/bjsports-2018-100486] [Medline: 30647053]
17. Catapano M, Zhang K, Mittal N, Sangha H, Onishi K, de Sa D. Effectiveness of dextrose prolotherapy for rotator cuff tendinopathy: a systematic review. PM R 2020;12(3):288-300. [doi: 10.1002/pmrj.12268] [Medline: 31642203]
18. Gutiérrez-Espinoza H, Araya-Quintanilla F, Cereceda-Muriel C, Álvarez-Bueno C, Martínez-Vizcaíno V, Cavero-Redondo I. Effect of supervised physiotherapy versus home exercise program in patients with subacromial impingement syndrome: a systematic review and meta-analysis. Phys Ther Sport 2020;41:34-42. [doi: 10.1016/j.ptsp.2019.11.003] [Medline: 31726386]
19. Chen X, Jones IA, Togashi R, Park C, Vangsness CT. Use of platelet-rich plasma for the improvement of pain and function in rotator cuff tears: a systematic review and meta-analysis with bias assessment. Am J Sports Med 2020;48(8):2028-2041 [FREE Full text] [doi: 10.1177/0363546519881423] [Medline: 31743037]
20. An VVG, Farey JE, Karunaratne S, Smithers CJ, Petchell JF. Subacromial analgesia via continuous infusion catheter vs. placebo following arthroscopic shoulder surgery: a systematic review and meta-analysis of randomized trials. J Shoulder Elbow Surg 2020;29(3):471-482. [doi: 10.1016/j.jse.2019.11.010] [Medline: 32067710]
21. Craig RS, Goodier H, Singh JA, Hopewell S, Rees JL. Shoulder replacement surgery for osteoarthritis and rotator cuff tear arthropathy. Cochrane Database Syst Rev 2020;4(4):CD012879 [FREE Full text] [doi: 10.1002/14651858.CD012879.pub2] [Medline: 32315453]
22. Naunton J, Street G, Littlewood C, Haines T, Malliaras P. Effectiveness of progressive and resisted and non-progressive or non-resisted exercise in rotator cuff related shoulder pain: a systematic review and meta-analysis of randomized controlled trials. Clin Rehabil 2020;34(9):1198-1216. [doi: 10.1177/0269215520934147] [Medline: 32571081]
23. Malliaras P, Johnston R, Street G, Littlewood C, Bennell K, Haines T, et al. The efficacy of higher versus lower dose exercise in rotator cuff tendinopathy: a systematic review of randomized controlled trials. Arch Phys Med Rehabil 2020;101(10):1822-1834. [doi: 10.1016/j.apmr.2020.06.013] [Medline: 32682937]
24. Simpson M, Pizzari T, Cook T, Wildman S, Lewis J. Effectiveness of non-surgical interventions for rotator cuff calcific tendinopathy: a systematic review. J Rehabil Med 2020;52(10):jrm00119 [FREE Full text] [doi: 10.2340/16501977-2725] [Medline: 32830280]
25. Belk JW, McCarty EC, Houck DA, Dragoo JL, Savoie FH, Thon SG. Tranexamic acid use in knee and shoulder arthroscopy leads to improved outcomes and fewer hemarthrosis-related complications: a systematic review of level I and II studies. Arthroscopy 2021;37(4):1323-1333. [doi: 10.1016/j.arthro.2020.11.051] [Medline: 33278534]
26. Hillier M. Why does ChatGPT generate fake references? TECHE. 2023. URL: https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/ [accessed 2023-05-17]
27. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clin Proc Digit Health 2023;1(3):226-234 [FREE Full text] [doi: 10.1016/j.mcpdig.2023.05.004]
28. Day T. A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. Prof Geogr 2023;75(6):1024-1027 [FREE Full text] [doi: 10.1080/00330124.2023.2190373]
29. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
30. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt [accessed 2023-05-18]
31. Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C, et al. Learning to summarize from human feedback. ArXiv Preprint posted online on September 2, 2020 [FREE Full text] [doi: 10.48550/arXiv.2009.01325]
32. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023;5(6):e333-e335 [FREE Full text] [doi: 10.1016/S2589-7500(23)00083-3] [Medline: 37120418]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 27.09.23; peer-reviewed by U Kgmu, L Carrasco-Ribelles; comments to author 02.01.24; revised version received 22.01.24; accepted 21.02.24; published 22.05.24.

Please cite as:
Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, Raynier JL, Clowez G, Boileau P, Ruetsch-Chelli C
Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis
J Med Internet Res 2024;26:e53164
URL: https://www.jmir.org/2024/1/e53164
doi:10.2196/53164
PMID:38776130

©Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, Caroline Ruetsch-Chelli. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI Quality Standards in Health Care: Rapid Umbrella Review

Craig E Kuziemsky[1], BSc, BCom, PhD; Dillon Chrimes[2], BSc, MSc, PhD; Simon Minshall[2], BSc, MSc; Michael Mannerow[1], BSc; Francis Lau[2], BSc, MSc, MBA, PhD

[1]MacEwan University, Edmonton, AB, Canada
[2]School of Health Information Science, University of Victoria, Victoria, BC, Canada

Corresponding Author:
Craig E Kuziemsky, BSc, BCom, PhD
MacEwan University
10700 104 Avenue
7-257
Edmonton, AB, T5J4S2
Canada
Phone: 1 7806333290
Email: kuziemskyc@macewan.ca

Abstract

Background: In recent years, there has been an upwelling of artificial intelligence (AI) studies in the health care literature. During this period, there has been an increasing number of proposed standards to evaluate the quality of health care AI studies.

Objective: This rapid umbrella review examines the use of AI quality standards in a sample of health care AI systematic review articles published over a 36-month period.

Methods: We used a modified version of the Joanna Briggs Institute umbrella review method. Our rapid approach was informed by the practical guide by Tricco and colleagues for conducting rapid reviews. Our search was focused on the MEDLINE database supplemented with Google Scholar. The inclusion criteria were English-language systematic reviews regardless of review type, with mention of AI and health in the abstract, published during a 36-month period. For the synthesis, we summarized the AI quality standards used and issues noted in these reviews drawing on a set of published health care AI standards, harmonized the terms used, and offered guidance to improve the quality of future health care AI studies.

Results: We selected 33 review articles published between 2020 and 2022 in our synthesis. The reviews covered a wide range of objectives, topics, settings, designs, and results. Over 60 AI approaches across different domains were identified with varying levels of detail spanning different AI life cycle stages, making comparisons difficult. Health care AI quality standards were applied in only 39% (13/33) of the reviews and in 14% (25/178) of the original studies from the reviews examined, mostly to appraise their methodological or reporting quality. Only a handful mentioned the transparency, explainability, trustworthiness, ethics, and privacy aspects. A total of 23 AI quality standard–related issues were identified in the reviews. There was a recognized need to standardize the planning, conduct, and reporting of health care AI studies and address their broader societal, ethical, and regulatory implications.

Conclusions: Despite the growing number of AI standards to assess the quality of health care AI studies, they are seldom applied in practice. With increasing desire to adopt AI in different health topics, domains, and settings, practitioners and researchers must stay abreast of and adapt to the evolving landscape of health care AI quality standards and apply these standards to improve the quality of their AI studies.

(J Med Internet Res 2024;26:e54705) doi:10.2196/54705

KEYWORDS

artificial intelligence; health care artificial intelligence; health care AI; rapid review; umbrella review; quality standard

XSL•FO
RenderX

Introduction

Growth of Health Care Artificial Intelligence

In recent years, there has been an upwelling of artificial intelligence (AI)–based studies in the health care literature. While there have been reported benefits, such as improved prediction accuracy and monitoring of diseases [1], health care organizations face potential patient safety, ethical, legal, social, and other risks from the adoption of AI approaches [2,3]. A search of the MEDLINE database for the terms "artificial intelligence" and "health" in the abstracts of articles published in 2022 alone returned >1000 results. Even by narrowing it down to systematic review articles, the same search returned dozens of results. These articles cover a wide range of AI approaches applied in different health care contexts, including such topics as the application of machine learning (ML) in skin cancer [4], use of natural language processing (NLP) to identify atrial fibrillation in electronic health records [5], image-based AI in inflammatory bowel disease [6], and predictive modeling of pressure injury in hospitalized patients [7]. The AI studies reported are also at different AI life cycle stages, from model development, validation, and deployment to evaluation [8]. Each of these AI life cycle stages can involve different contexts, questions, designs, measures, and outcomes [9]. With the number of health care AI studies rapidly on the rise, there is a need to evaluate the quality of these studies in different contexts. However, the means to examine the quality of health care AI studies have grown more complex, especially when considering their broader societal and ethical implications [10-13].

Coiera et al [14] described a "replication crisis" in health and biomedical informatics where issues regarding experimental design and reporting of results impede our ability to replicate existing research. Poor replication raises concerns about the quality of published studies as well as the ability to understand how context could impact replication across settings. The replication issue is prevalent in health care AI studies as many are single-setting approaches and we do not know the extent to which they can be translated to other settings or contexts. One solution to address the replication issue in AI studies has been the development of a growing number of AI quality standards. Most prominent are the reporting guidelines from the Enhancing the Quality and Transparency of Health Research (EQUATOR) network [15]. Examples include the CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) extension for reporting AI clinical trials [16] and the SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) extension for reporting AI clinical trial protocols [17]. Beyond the EQUATOR guidelines, there are also the Minimum Information for Medical AI Reporting standard [18] and the Minimum Information About Clinical Artificial Intelligence Modeling checklist [19] on the minimum information needed in published AI studies. These standards mainly focus on the methodological and reporting quality aspects of AI studies to ensure that the published information is rigorous, complete, and transparent.

Need for Health Care AI Standards

However, there is a shortcoming of standard-driven guidance that spans the entire AI life cycle spectrum of design, validation, implementation, and governance. The World Health Organization has published six ethical principles to guide the use of AI [20] that cover (1) protecting human autonomy; (2) promoting human well-being and safety and the public interest; (3) ensuring transparency, explainability, and intelligibility; (4) fostering responsibility and accountability; (5) ensuring inclusiveness and equity; and (6) promoting AI that is responsive and sustainable. In a scoping review, Solanki et al [21] operationalized health care AI ethics through a framework of 6 guidelines that spans the entire AI life cycle of data management, model development, deployment, and monitoring. The National Health Service England has published a best practice guide on health care AI on *how to get it right* that encompasses a governance framework, addressing data access and protection issues, spreading the good innovation, and monitoring uses over time [22]. To further promote the quality of health care AI, van de Sande et al [23] have proposed a step-by-step approach with specific AI quality criteria that span the entire AI life cycle from development and implementation to governance.

Despite the aforementioned principles, frameworks, and guidance, there is still widespread variation in the quality of published AI studies in the health care literature. For example, 2 systematic reviews of 152 prediction and 28 diagnosis studies have shown poor methodological and reporting quality that have made it difficult to replicate, assess, and interpret the study findings [24,25]. The recent shifts beyond study quality to broader ethical, equity, and regulatory issues have also raised additional challenges for AI practitioners and researchers on the impact, transparency, trustworthiness, and accountability of the AI studies involved [13,26-28]. Increasingly, we are also seeing reports of various types of AI implementation issues [2]. There is a growing gap between the expected quality and performance of health care AI that needs to be addressed. We suggest that the overall issue is a lack of awareness and of the use of principles, frameworks, and guidance in health care AI studies.

This rapid umbrella review addressed the aforementioned issues by focusing on the principles and frameworks for health care AI design, implementation, and governance. We analyzed and synthesized the use of AI quality standards as reported in a sample of published health care AI systematic review articles. In this paper, AI quality standards are defined as guidelines, criteria, checklists, statements, guiding principles, or framework components used to evaluate the quality of health care AI studies in different domains and life cycle stages. In this context, quality covers the trustworthiness, methodological, reporting, and technical aspects of health care AI studies. Domains refer to the disciplines, branches, or areas in which AI can be found or applied, such as computer science, medicine, and robotics. The findings from this review can help address the growing need for AI practitioners and researchers to navigate the increasingly complex landscape of AI quality standards to plan, conduct, evaluate, and report health care AI studies.

XSL•FO
RenderX

Methods

Overview

With the increasing volume of systematic review articles that appear in the health care literature each year, an umbrella review has become a popular and timely approach to synthesize knowledge from published systematic reviews on a given topic. For this paper, we drew on the umbrella review method in the typology of systematic reviews for synthesizing evidence in health care by MacEntee [29]. In this typology, umbrella reviews are used to synthesize multiple systematic reviews from different sources into a summarized form to address a specific topic. We used a modified version of the Joanna Briggs Institute (JBI) umbrella review method to tailor the process, including developing of an umbrella review protocol, applying a rapid approach, and eliminating duplicate original studies [30]. Our rapid approach was informed by the practical guide to conducting rapid reviews in the areas of database selection, topic refinement, searching, study selection, data extraction, and synthesis by Tricco et al [31]. A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of our review process is shown in Figure 1 [32]. A PRISMA checklist is provided in Multimedia Appendix 1 [32].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram based on the work by Page et al [32]. JBI: Joanna Briggs Institute.

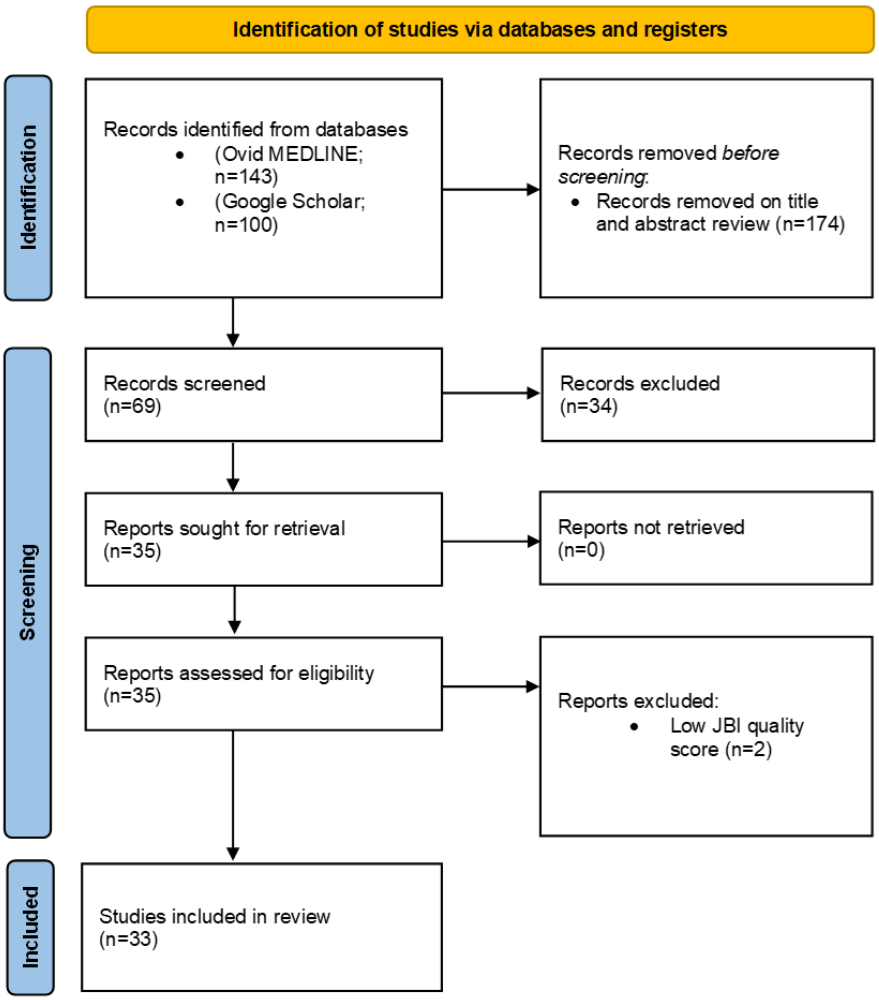

Objective and Questions

The objective of this rapid umbrella review was to examine the use of AI quality standards based on a sample of published health care AI systematic reviews. Specifically, our questions were as follows:

1. What AI quality standards have been applied to evaluate the quality of health care AI studies?
2. What key quality standard–related issues are noted in these reviews?
3. What guidance can be offered to improve the quality of health care AI studies through the incorporation of AI quality standards?

Search Strategy

Our search strategy focused on the MEDLINE database supplemented with Google Scholar. Our search terms consisted of “artificial intelligence” or “AI,” “health,” and “systematic review” mentioned in the abstract (refer to Multimedia Appendix 2 for the search strings used). We used the .TW search field tag as it searches on title and abstract as well as fields such as abstract, Medical Subject Heading terms, and Medical Subject

XSL•FO
RenderX

Heading subheadings. Our rationale to limit the search to MEDLINE with simple terms was to keep the process manageable, recognizing the huge volume of health care AI–related literature reviews that have appeared in the last few years, especially on COVID-19. One author conducted the MEDLINE and Google Scholar searches with assistance from an academic librarian. For Google Scholar, we restricted the search to the first 100 citations returned.

Inclusion Criteria

We considered all English-language systematic review articles published over a 36-month period from January 1, 2020, to December 31, 2022. The review could be any type of systematic review, meta-analysis, narrative review, qualitative review, scoping review, meta-synthesis, realist review, or umbrella review as defined in the review typology by MacEntee [29]. The overarching inclusion criteria were AI and health as the focus. To be considered for inclusion, the review articles must meet the following criteria:

1. Each original study in the review is described, where an AI approach in the form of a model, method, algorithm, technique, or intervention is proposed, designed, implemented, or evaluated within a health care context to address a particular health care problem or topic area.
2. We define AI as a simulation of the approximation of human intelligence in machines that comprises learning, reasoning, and logic [33]. In that approximation, AI has different levels of adaptivity and autonomy. Weak AI requires supervision or reinforced learning with human intervention to adapt to the environment, with low autonomous interaction. Strong AI is highly adaptive and highly autonomous via unsupervised learning, with no human intervention.
3. We looked through all the articles, and our health care context categorization was informed by the stated settings (eg, hospital) and purpose (eg, diagnosis) mentioned in the included reviews.
4. The review can include all types of AI approaches, such as ML, NLP, speech recognition, prediction models, neural networks, intelligent robotics, and AI-assisted and automated medical devices.
5. The review must contain sufficient detail on the original AI studies, covering their objectives, contexts, study designs, AI approaches, measures, outcomes, and reference sources.

Exclusion Criteria

We excluded articles if any one of the following applied:

1. Review articles published before January 1, 2020; not accessible in web-based format; or containing only an abstract
2. Review articles in languages other than English
3. Earlier versions of the review article with the same title or topic by the same authors
4. Context not health care–related, such as electronic commerce or smart manufacturing
5. The AI studies not containing sufficient detail on their purpose, features, or reference sources
6. Studies including multiple forms of digital health technologies besides AI, such as telehealth, personal health records, or communication tools

Review Article Selection

One author conducted the literature searches and retrieved the citations after eliminating duplicates. The author then screened the citation titles and abstracts against the inclusion and exclusion criteria. Those that met the inclusion criteria were retrieved for full-text review independently by 2 other authors. Any disagreements in final article selection were resolved through consensus between the 2 authors or with a third author. The excluded articles and the reasons for their exclusion were logged.

Quality Appraisal

In total, 2 authors applied the JBI critical appraisal checklist independently to appraise the quality of the selected reviews [30]. The checklist has 11 questions that allow for *yes*, *no*, *unclear*, or *not applicable* as the response. The questions cover the areas of review question, inclusion criteria, search strategy and sources, appraisal criteria used, use of multiple reviewers, methods of minimizing data extraction errors and combining studies, publication bias, and recommendations supported by data. The reviews were ranked as high, medium, and low quality based on their JBI critical appraisal score (≥0.75 was high quality, ≥0.5 and <0.75 was medium quality, and <0.5 was low quality). All low-quality reviews were excluded from the final synthesis.

Data Extraction

One author extracted data from selected review articles using a predefined template. A second author validated all the articles for correctness and completeness. As this review was focused on AI quality standards, we extracted data that were relevant to this topic. We created a spreadsheet template with the following data fields to guide data extraction:

1. Author, year, and reference: first author last name, publication year, and reference number
2. URL: the URL where the review article can be found
3. Objective or topic: objective or topic being addressed by the review article
4. Type: type of review reported (eg, systematic review, meta-analysis, or scoping review)
5. Sources: bibliographic databases used to find the primary studies reported in the review article
6. Years: period of the primary studies covered by the review article
7. Studies: total number of primary studies included in the review article
8. Countries: countries where the studies were conducted
9. Settings: study settings reported in the primary studies of the review article
10. Participants: number and types of individuals being studied as reported in the review article
11. AI approaches: the type of AI model, method, algorithm, technique, tool, or intervention described in the review article

XSL•FO
RenderX

12. Life cycle and design: the stage or design of the AI study in the AI life cycle in the primary studies being reported, such as requirements, design, implementation, monitoring, experimental, observational, training-test-validation, or controlled trial
13. Appraisal: quality assessment of the primary studies using predefined criteria (eg, risk of bias)
14. Rating: quality assessment results of the primary studies reported in the review article
15. Measures: performance criteria reported in the review article (eg, mortality, accuracy, and resource use)
16. Analysis: methods used to summarize the primary study results (eg, narrative or quantitative)
17. Results: aggregate findings from the primary studies in the review article
18. Standards: name of the quality standards mentioned in the review article
19. Comments: issues mentioned in the review article relevant to our synthesis

Removing Duplicate AI Studies

We identified all unique AI studies across the selected reviews after eliminating duplicates that appeared in them. We retrieved full-text articles for every tenth of these unique studies and searched for mention of AI quality standard–related terms in them. This was to ensure that all relevant AI quality standards were accounted for even if the reviews did not mention them.

Analysis and Synthesis

Our analysis was based on a set of recent publications on health care AI standards. These include (1) the AI life cycle step-by-step approach by van de Sande et al [23] with a list of AI quality standards as benchmarks, (2) the reporting guidelines by Shelmerdine et al [15] with specific standards for different AI-based clinical studies, (3) the international standards for evaluating health care AI by Wenzel and Wiegand [26], and (4) the broader requirements for trustworthy health care AI across the entire life cycle stages by the National Academy of Medicine (NAM) [8] and the European Union Commission (EUC) [34]. As part of the synthesis, we created a conceptual organizing scheme drawing on published literature on AI domains and approaches to visualize their relationships (via a Euler diagram) [35]. All analyses and syntheses were conducted by one author and then validated by another to resolve differences.

For the analysis, we (1) extracted key characteristics of the selected reviews based on our predefined template; (2) summarized the AI approaches, life cycle stages, and quality standards mentioned in the reviews; (3) extracted any additional AI quality standards mentioned in the 10% sample of unique AI studies from the selected reviews; and (4) identified AI quality standard–related issues reported.

For the synthesis, we (1) mapped the AI approaches to our conceptual organizing scheme, visualized their relationships with the AI domains and health topics found, and described the challenges in harmonizing these terms; (2) established key themes from the AI quality standard issues identified and mapped them to the NAM and EUC frameworks [8,34]; and (3) created a summary list of the AI quality standards found and mapped them to the life cycle phases by van de Sande et al [23].

Drawing on these findings, we proposed a set of guidelines that can enhance the quality of future health care AI studies and described its practice, policy, and research implications. Finally, we identified the limitations of this rapid umbrella review as caveats for the readers to consider. As health care, AI, and standards are replete with industry terminologies, we used the acronyms where they are mentioned in the paper and compiled an alphabetical acronym list with their spelled-out form at the end of the paper.

Results

Summary of Included Reviews

We found 69 health care AI systematic review articles published between 2020 and 2022, of which 35 (51%) met the inclusion criteria. The included articles covered different review types, topics, settings, numbers of studies, designs, participants, AI approaches, and performance measures (refer to Multimedia Appendix 3 [36-68] for the review characteristics). We excluded the remaining 49% (34/69) of the articles because they (1) covered multiple technologies (eg, telehealth), (2) had insufficient detail, (3) were not specific to health care, or (4) were not in English (refer to Multimedia Appendix 4 for the excluded reviews and reasons). The quality of these reviews ranged from JBI critical appraisal scores of 1.0 to 0.36, with 49% (17/35) rated as high quality, 40% (14/35) rated as moderate quality, and 6% (2/35) rated as poor quality (Multimedia Appendix 5 [36-68]). A total of 6% (2/35) of the reviews were excluded for their low JBI scores [69,70], leaving a sample of 33 reviews for the final synthesis.

Regarding review types, most (23/33, 70%) were systematic reviews [37-40,45-51,53-57,59-64,66,67], with the remaining being scoping reviews [36,41-44,52,58,65,68]. Only 3% (1/33) of the reviews were meta-analyses [38], and another was a rapid review [61]. Regarding health topics, the reviews spanned a wide range of specific health conditions, disciplines, areas, and practices. Examples of conditions were COVID-19 [36,37,49,51,56,62,66], mental health [48,65,68], infection [50,59,66], melanoma [57], and hypoglycemia [67]. Examples of disciplines were public health [36,37,56,66], nursing [42,43,61], rehabilitation [52,64], and dentistry [55,63]. Areas included mobile health and wearables [41,52,54,65], surveillance and remote monitoring [51,61,66], robotic surgeries [47], and biobanks [39]. Practices included diagnosis [37,47,49,58,59,62], prevention [47], prediction [36,38,49,50,57], disease management [41,46,47,58], and administration [42]. Regarding settings, less than half (12/33, 36%) were explicit in their health care settings, which included multiple sources [36,42,43,50,54,61], hospitals [45,49], communities [44,51,58], and social media groups [48]. The number of included studies ranged from 794 on COVID-19 [49] to 8 on hypoglycemia [67]. Regarding designs, most were performance assessment studies using secondary data sources such as intensive care unit [38], imaging [37,62,63], and biobank [39] databases. Regarding participants, they included patients, health care providers, educators, students, simulated cases, and those who use social

XSL•FO
RenderX

media. Less than one-third of the reviews (8/33, 24%) mentioned sample sizes, which ranged from 11 adults [44] to 1,547,677 electronic medical records [40] (refer to Multimedia Appendix 3 for details).

Regarding AI approaches, there were >60 types of AI models, methods, algorithms, tools, and techniques mentioned in varying levels of detail across the broad AI domains of computer science, data science with and without NLP, and robotics. The main AI approaches were ML and deep learning (DL), with support vector machine, convolutional neural network, neural network, logistic regression, and random forest being mentioned the most (refer to the next section for details). The performance measures covered a wide range of metrics, such as diagnostic and prognostic accuracies (eg, sensitivity, specificity, accuracy, and area under the curve) [37-40,46-48,53,57,59,63,67], resource use (eg, whether an intensive care unit stay was necessary, length of stay, and cost) [37,58,62], and clinical outcomes (eg, COVID-19 severity, mortality, and behavior change) [36,37,49,56,62,65]. A few reviews (6/33, 18%) focused on the extent of the socioethical guidelines addressed [44,51,55,58,66,68]. Regarding life cycle stages, different schemes were applied, including preprocessing and classification [48,57], data preparation-preprocessing [37,38], different stages of adoption (eg, knowledge, persuasion, decision making, implementation) [44], conceptual research [42], model development [36,37,40,42,45,46,50-56,58-64,66,67], design [43], training and testing [38,42,45,50-53,58,61-64], validation [36-38,40,45,46,50,51,53,55,56,58-64,67], pilot trials [65], public engagement [68], implementation [42,44,60-62,66,68], confirmation [44], and evaluation [42,43,53,60-62,65] (refer to Multimedia Appendix 3 for details). It is worth noting that the period covered for our review did not include any studies on large language models (LLMs). LLM studies became more prevalent in the literature in the period just after our review.

Use of Quality Standards in Health Care AI Studies

To make sense of the different AI approaches mentioned, we used a Euler diagram [71] as a conceptual organizing scheme to visualize their relationships with AI domains and health topics (Figure 2 [36,41-43,47,48,51-54,56-58,60,62,65,67]). The Euler diagram shows that AI broadly comprised approaches in the domains of computer science, data science with and without NLP, and robotics that could be overlapping. The main AI approaches were ML and DL, with DL being a more advanced form of ML through the use of artificial neural networks [33]. The diagram also shows that AI can exist without ML and DL (eg, decision trees and expert systems). There are also outliers in these domains with borderline AI-like approaches mostly intended to enhance human-computer interactions, such as social robotics [42,43], robotic-assisted surgery [47], and exoskeletons [54]. The health topics in our reviews spanned the AI domains, with most falling within data science with or without NLP. This was followed by computer science mostly for communication or database and other functional support and robotics for enhanced social interactions that may or may not be AI driven. There were borderline AI approaches such as programmed social robotics [42,43] or AI-enhanced social robots [54]. These approaches focus on AI enabled social robotic programming and did not use ML or DL. Borderline AI approaches also included virtual reality [60] and wearable sensors [65,66,68].

Regarding AI life cycle stages, we harmonized the different terms used in the original studies by mapping them to the 5 life cycle phases by van de Sande et al [23]: 0 (preparation), I (model development), II (performance assessment), III (clinical testing), and IV (implementation). Most AI studies in the reviews mapped to the first 3 life cycle phases by van de Sande et al [23]. These studies would typically describe the development and performance of the AI approach on a given health topic in a specific domain and setting, including their validation, sometimes done using external data sets [36,38]. A small number of reviews reported AI studies that were at the clinical testing phase [60,61,66,68]. A total of 7 studies were described as being in the implementation phase [66,68]. On the basis of the descriptions provided, few of the AI approaches in the studies in the AI reviews had been adopted for routine use in clinical settings [66,68] with quantifiable improvements in health outcomes (refer to Multimedia Appendix 6 [36-68] for details).

Regarding AI quality standards, only 39% (13/33) of the reviews applied specific AI quality standards in their results [37-40,45,46,50,54,58,59,61,63,66], and 12% (4/33) mentioned the need for standards [55,63,68]. These included the Prediction Model Risk of Bias Assessment Tool [37,38,58,59], Newcastle-Ottawa Scale [39,50], Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies [38,59], Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Machine Learning Extension [50], levels of evidence [61], Critical Appraisal Skills Program Clinical Prediction Rule Checklist [40], Mixed Methods Appraisal Tool [66], and CONSORT-AI [54]. Another review applied 7 design justice principles as the criteria to appraise the quality of their AI studies [68]. There were also broader-level standards mentioned. These included the European Union ethical guidelines for trustworthy AI [44]; international AI standards from the International Organization for Standardization (ISO); and AI policy guidelines from the United States, Russia, and China [46] (refer to Multimedia Appendix 6 for details). We updated the Euler diagram (Figure 2 [36,41-43,47,48,51-54,56-58,60,62,65,67]) to show in red the health topics in reviews with no mention of specific AI standards.

XSL•FO
RenderX

Figure 2. Euler diagram showing the overlapping artificial intelligence (AI) domains and health topics. Health topics in red are from reviews with no mention of specific AI quality standards. Health-related subjects in blue are from reviews with mention of AI quality standards. DL: deep learning; mHealth: mobile health.

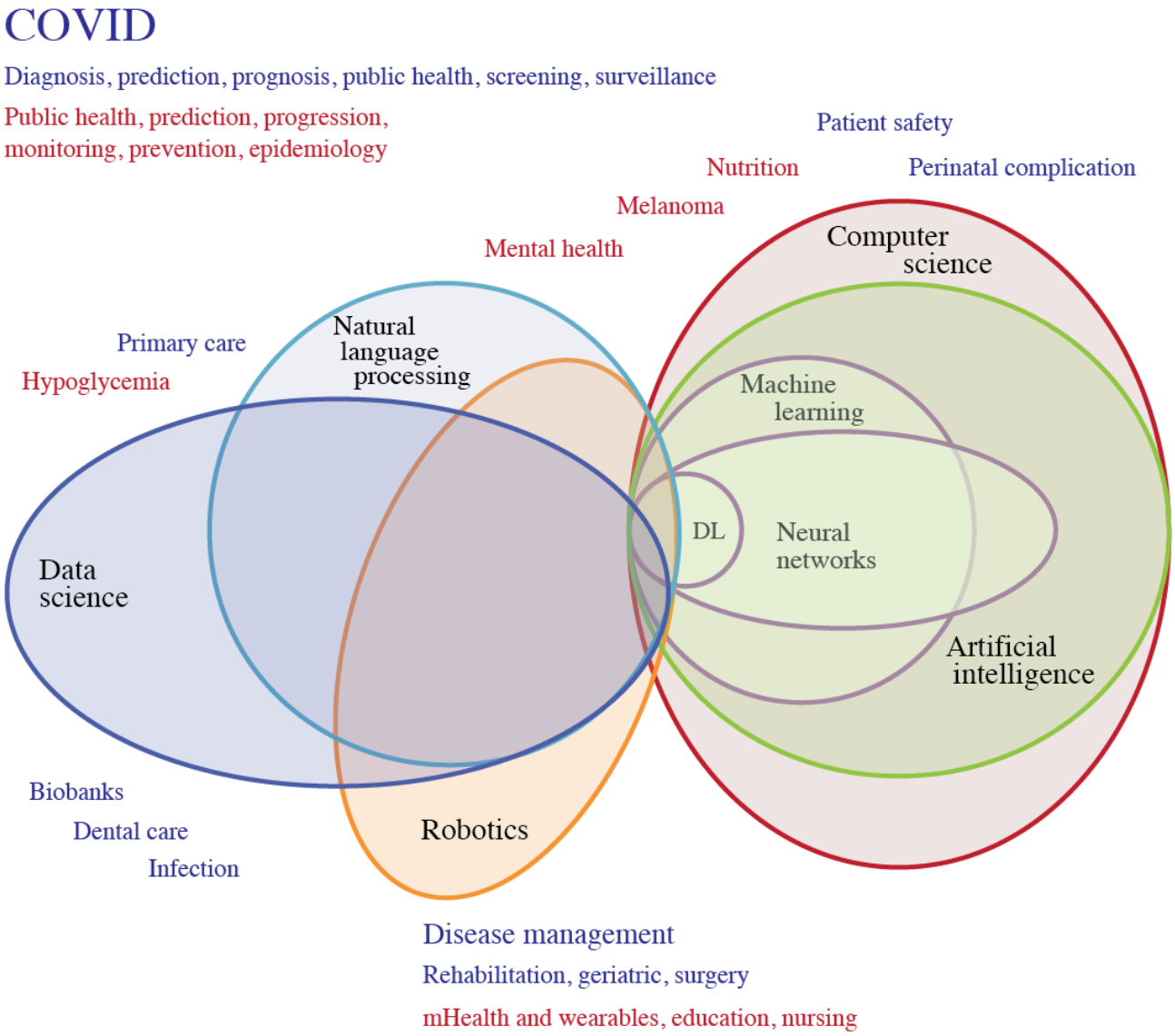

Of the 178 unique original AI studies from the selected reviews that were examined, only 25 (14%) mentioned the use of or need for specific AI quality standards (refer to Multimedia Appendix 7 [36-68] for details). They were of six types: (1) reporting—COREQ (Consolidated Criteria for Reporting Qualitative Research), Strengthening the Reporting of Observational Studies in Epidemiology, Standards for Reporting Diagnostic Accuracy Studies, PRISMA, and EQUATOR; (2) data—Unified Medical Language System, Food and Drug Administration (FDA) Adverse Event Reporting System, MedEx, RxNorm, Medical Dictionary for Regulatory Activities, and PCORnet; (3) technical—ISO-12207, FDA Software as a Medical Device, EU-Scholarly Publishing and Academic Resources Coalition, Sensor Web Enablement, Open Geospatial Consortium, Sensor Observation Service, and the American Medical Association AI recommendations; (4) robotics—ISO-13482 and ISO and TC-299; (5) ethics—Helsinki Declaration and European Union AI Watch; and (6) regulations—Health Insurance Portability and Accountability Act (HIPAA) and World Health Organization World Economic Forum. These standards were added to the list of AI quality standards mentioned by review in Multimedia Appendix 6.

A summary of the harmonized AI topics, approaches, domains, the life cycle phases by van de Sande et al [23], and quality standards derived from our 33 reviews and 10% of unique studies within them is shown in Table 1.

XSL•FO
RenderX

Table 1. Summary of artificial intelligence (AI) approaches, domains, life cycle phases, and quality standards in the reviews.

Review, year	Topics	Approaches; examples only (source from original review)	Domains[a]	Life cycle phases	Quality standards[b]
Abd-Alrazaq et al [36], 2020	Public health, risk prediction, and COVID-19	CNN[c], SVM[d], RF[e], DT[f], and LoR[g]	Data science with NLP[h]	Phase 0, I, and II[i]	Not mentioned; *Helsinki declaration*[b]
Adamidi et al [37], 2021	Public health, COVID-19, screening, diagnosis, and prognosis	AB[j], ARMED[k], BE[l], BNB[m], and CNN	Data science	Phase 0, I, and II	PROBAST[n], *TRIPOD*[b,o]*FDA-SaMD*[b,p], and *STROBE*[b,q]
Barboi et al [38], 2022	Prediction, mortality, and ICU[r]	ANN-ELM[s], DT, ELM[t], ensemble LSTM[u], and ESICU-LA[v]	Data science with NLP	Phase 0, I, and II	CHARMS[w] and PROBAST
Battineni et al [39], 2022	Biobanks	CNN and SFCN[x]	Data science and computer science	Phase 0 and I, maybe II	NOS[y]
Bertini et al [40], 2022	Perinatal and complications	AB, ANN[z], DT, EN[aa], and GAM[ab]	Computer science	Phase 0, I, and II with need for phase-III clinical testing	CASP[ac]
Bhatt et al [41], 2022	mHealth[ad] and disease management	DL[ae] and FL[af]	Data science	Phase 0 and I	Not mentioned
Buchanan et al [42], 2020	Administration, clinical practice, policy, and research	ML[ag], SAR[ah], CDSS[ai], and chatbots	Data science and robotics (x)	Phase 0, implied need for nurses to be involved in all phases	Not mentioned; *COREQ*[aj,b], *ISO-13482*[ak,b], *EU-SPARC*[al,b], and *ISO and TC299*, *AM*[am,b]
Buchanan et al [43], 2021	Education	ML, virtual avatar applications, chatbots, wearable armband with ML, and predictive analysis	Data science and robotics (x)	Phase 0, implied need for nurses to be part of co-design at all stages	Not mentioned
Chew and Achananuparp [44], 2022	General	Chatbots, weak AI, image recognition, AI diagnosis, and NLP	Data science with NLP and robotics	Implied phase 0	Not mentioned
Choudhury et al [45], 2020	Patient safety outcomes	ANN, BICMM[an], BNC[ao], C4.5[ap], and CPH[aq]	Data science	Phase 0 and I	ISO and IEC[ar] 23053, ISO 22100-5, Laskai, NIST[as], and OECD-AI[at]
Choudhury and Asan [46], 2020	Geriatrics and disease management	AUC[au], AI, BCP-NN[av], BCPNN[aw], and BNM[ax]	Data science with NLP	Phase 0 and I	TRIPOD, TRIPOD-ML[ay] (cited but not used in AI studies; cited NIST standards), *FAERS*[az,b], *MedEx*[b], *RxNorm*[b], *MedDRA*[ba,b], *PCORnet*[b], and *MADE1.0*[bb,b]
Eldaly et al [47], 2022	Lymphedema, prevention, diagnosis, and disease management	ANN, ANFIS[bc], chatbots, DT, and EML[bd]	Data science and robotics	Phase 0 and I	Not mentioned
Le Glaz et al [48], 2021	Mental health	C4.5, CRA[be], CUI[bf], DT, and KM[bg]	Data science with NLP and computer science	Phase 0 and I	Not mentioned; *UMLS*[bh,b]
Guo et al [49], 2021	Prediction, diagnosis, prognosis, and COVID-19	3DQI[bi], ACNN[bj], AB, AI, and ANN	Data science	Phase 0 and I	Not mentioned
Hassan et al [50], 2021	Prediction and sepsis	InSight, LASSO[bk], LR, MCRM[bl], and MLR[bm]	Data science and computer science	Phase 0, I, and II	TRIPOD and NOS

XSL•FO
RenderX

Review, year	Topics	Approaches; examples only (source from original review)	Domains[a]	Life cycle phases	Quality standards[b]
Huang et al [51], 2022	Telemedicine, monitoring, and COVID-19	CNN-TF[bn], IRRC-NN[bo], IoT[bp]-based wearable monitoring device, NVHODL[bq], and SVM	Data science	Phase 0 and I	Not mentioned
Kaelin et al [52], 2021	Pediatric and rehabilitation	NLP, ML, computer vision, and robotics	Data science, computer science, and robotics	Phase 0 and I	Not mentioned; *HIPAA*[br,b] and *COREQ*[b]
Kirk et al [53], 2021	Nutrition	BC[bs], CNN, DL, DT, and EM[bt]	Data science and computer science	Phase 0, I, and II; the mapping to these phases is questionable	Not mentioned (not even PRISMA[bu])
Loveys et al [54], 2022	Geriatrics and interventions	AI-enhanced robots, social robots, environmental sensors, and wearable sensors	Data science, computer science, and robotics (x)	Phase 0 and I	Revised Cochrane risk-of-bias tool for RCTs[bv], cluster RCTs, and non-RCTs (ROBINS-I[bw])
Mörch et al [55], 2021	Dentistry and ethics	DL, DSP[bx], ML, and NN[by]	Data science	Phase 0 and I	Mentioned need for SPIRIT[bz] and TRIPOD, used 2018 Montreal Declaration as AI ethical framework
Payedimarri et al [56], 2021	Public health, interventions, and COVID-19	ABS[ca], LiR[cb], NN, and TOPSIS[cc]	Data science	Phase 0 and I	Not mentioned
Popescu et al [57], 2022	Cancer and prediction melanoma	ABC[cd], autoencoder, CNN, combined networks, and DCNN[ce]	Data science and computer science	Phase 0 and I	Not mentioned
Abbasgholizadeh Rahimi et al [58], 2021	Primary care, diagnosis, and disease management	AL[cf], AR[cg], BN[ch], COBWEB[ci], and CH[cj]	Data science with NLP	Phase 0 and I	PROBAST
Sahu et al [59], 2022	Detection, neonatal, and sepsis	AR-HMM[ck], LiR, LoR, MLoR[cl], and NN	Data science	Phase 0 and I	CHARMS and PROBAST
Sapci and Sapci [60], 2020	Education	AI, DL, ITS[cm], ML, and NLP	Data Science with NLP and computer science (x)	Phase 0 and I for AI education and training	Not mentioned; *AMA–augmented intelligence*[cn,b]
Seibert et al [61], 2021	Ethics	APS[co], ML, ES[cp], hybrid, and NLP	Data science, computer science, and robotics	Phase 0, I, II, and III	Risk of bias; levels of evidence from I to VII and not applicable; *SWE*[cq,b], *OGC*[cr,b], *SOS*[cs,b], *COREQ*[b], and *STROBE*[b]
Syeda et al [62], 2021	Epidemiology, diagnosis, disease progression, and COVID-19	ANN, BiGAN[ct], CNN, DT, and DL	Data science	Phase 0, I, and II	Not mentioned
Talpur et al [63], 2022	Dentistry and caries	ADA-NN[cu], ANN, CNN, F-CNN[cv], and FFBP-ANN[cw]	Data science	Phase 0, I, and II	Risk-of-bias assessment, no other standards mentioned
Vélez-Guerrero et al [64], 2021	Rehabilitation	AL, ANN, AFM[cx], ATC[cy], and AFC[cz]	Data science, computer science, and robotics	Phase 0, I, and II	Need standardized protocol for clinical evaluation, FDA[da] regulatory standards
Welch et al [65], 2022	Diagnosis, pediatrics, and psychiatry	ML and wearable biosensors	Data science with NLP and computer science (x)	Phase 0, I, and II	Not mentioned

XSL•FO
RenderX

Review, year	Topics	Approaches; examples only (source from original review)	Domains[a]	Life cycle phases	Quality standards[b]
Zhao et al [66], 2021	Public health, surveillance, and COVID-19	AI, AR, ML, physiological monitoring, and sensory technologies	Data science (x)	Phase 0, I, II, III, and IV; the mapping to these phases is questionable	MMAT[db] for study quality, Asadi framework for ethics, and *STARD*[dc,b]
Zheng et al [67], 2022	Detection and hypoglycemia	ML or rule-based NLP	Data science with NLP (rule-based)	Phase 0, I, and II	Not mentioned

XSL•FO
RenderX

Review, year	Topics	Approaches; examples only (source from original review)	Domains[a]	Life cycle phases	Quality standards[b]
Zidaru et al [68], 2020	Mental health	ML, NLP, sentiment analysis, VR[dd], and wearable biosensors	Data science (x)	Phase 0, I, II, III, and IV; the mapping to these phases is questionable	Need standards for evaluating safety, outcomes, acceptability, explainability, and inclusive design; *EU AI Watch*[de,b], *FDA-SaMD*[b], *EQUATOR*[df,b], *WHO-WEF governance*[dg,b], *CCC AI road map*[dh,b], and *ISO, IEC, or IEEE-12207*[di,b]

[a]Borderline AI approaches in the AI domains are identified with *(x)*.

[b]Italicized entries are AI quality standards mentioned only in the original studies in the reviews.

[c]CNN: convolutional neural network.

[d]SVM: support vector machine.

[e]RF: random forest.

[f]DT: decision tree.

[g]LoR: logistic regression.

[h]NLP: natural language processing.

[i]Phase 0: preparation before model development; phase I: AI model development; phase II: assessment of AI performance and reliability; phase III: clinical testing of AI; and phase IV: implementing and governing AI.

[j]AB: adaptive boosting or adaboost.

[k]ARMED: attribute reduction with multi-objective decomposition ensemble optimizer.

[l]BE: boost ensembling.

[m]BNB: Bernoulli naïve Bayes.

[n]PROBAST: Prediction Model Risk of Bias Assessment Tool.

[o]TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

[p]FDA-SaMD: Food and Drug Administration–Software as a Medical Device.

[q]STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.

[r]ICU: intensive care unit.

[s]ANN-ELM: artificial neural network extreme learning machine.

[t]ELM: ensemble machine learning.

[u]LSTM: long short-term memory.

[v]ESICULA: super intensive care unit learner algorithm.

[w]CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies.

[x]SFCN: sparse fully convolutional network.

[y]NOS: Newcastle-Ottawa scale.

[z]ANN: artificial neural network.

[aa]EN: elastic net.

[ab]GAM: generalized additive model.

[ac]CASP: Critical Appraisal Skills Programme.

[ad]mHealth: mobile health.

[ae]DL: deep learning.

[af]FL: federated learning.

[ag]ML: machine learning.

[ah]SAR: socially assistive robot.

[ai]CDSS: clinical decision support system.

[aj]COREQ: Consolidated Criteria for Reporting Qualitative Research.

[ak]ISO: International Organization for Standardization.

[al]EU-SPARC: Scholarly Publishing and Academic Resources Coalition Europe.

[am]AMS: Associated Medical Services.

[an]BICMM: Bayesian independent component mixture model.

[ao]BNC: Bayesian network classifier.

[ap]C4.5: a named algorithm for creating decision trees.

XSL•FO
RenderX

[aq]CPH: Cox proportional hazard regression.
[ar]IEC: international electrotechnical commission.
[as]NIST: National Institute of Standards and Technology.
[at]OECD-AI: Organisation for Economic Co-operation and Development–artificial intelligence.
[au]AUC: area under the curve.
[av]BCP-NN: Bayesian classifier based on propagation neural network.
[aw]BCPNN: Bayesian confidence propagation neural network.
[ax]BNM: Bayesian network model.
[ay]TRIPOD-ML: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Machine Learning.
[az]FAERS: Food and Drug Administration Adverse Event Reporting System.
[ba]MedDRA: Medical Dictionary for Regulatory Activities.
[bb]MADE1.0: Medical Artificial Intelligence Data Set for Electronic Health Records 1.0.
[bc]ANFIS: adaptive neuro fuzzy inference system.
[bd]EML: ensemble machine learning.
[be]cTAKES: clinical Text Analysis and Knowledge Extraction System.
[bf]CUI: concept unique identifier.
[bg]KM: k-means clustering.
[bh]UMLS: Unified Medical Language System.
[bi]3DQI: 3D quantitative imaging.
[bj]ACNN: attention-based convolutional neural network.
[bk]LASSO: least absolute shrinkage and selection operator.
[bl]MCRM: multivariable Cox regression model.
[bm]MLR: multivariate linear regression.
[bn]CNN-TF: convolutional neural network using Tensorflow.
[bo]IRRCN: inception residual recurrent convolutional neural network.
[bp]IoT: internet of things.
[bq]NVHDOL: notal vision home optical-based deep learning.
[br]HIPAA: Health Insurance Portability and Accountability Act.
[bs]BC: Bayesian classifier.
[bt]EM: ensemble method.
[bu]PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[bv]RCT: randomized controlled trial.
[bw]ROBINS-I: Risk of Bias in Non-Randomised Studies of Interventions.
[bx]DSP: deep supervised learning.
[by]NN: neural network.
[bz]SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials.
[ca]ABS: agent based simulation.
[cb]LiR: linear regression.
[cc]TOPSIS: technique for order of preference by similarity to ideal solution.
[cd]ABC: artificial bee colony.
[ce]DCNN: deep convolutional neural network.
[cf]AL: abductive learning.
[cg]AR: automated reasoning.
[ch]BN: Bayesian network.
[ci]COBWEB: a conceptual clustering algorithm.
[cj]CH: computer heuristic.
[ck]AR-HMM: auto-regressive hidden Markov model.
[cl]MLoR: multivariate logistic regression.
[cm]ITS: intelligent tutoring system.
[cn]AMA: American Medical Association.
[co]APS: automated planning and scheduling.
[cp]ES: expert system.
[cq]SWE: software engineering.
[cr]OGC: open geospatial consortium standard.

XSL•FO
RenderX

[cs]SOS: start of sequence.
[ct]BiGAN: bidirectional generative adversarial network.
[cu]ADA-NN: adaptive dragonfly algorithms with neural network.
[cv]F-CNN: fully convolutional neural network.
[cw]FFBP-ANN: feed-forward backpropagation artificial neural network.
[cx]AFM: adaptive finite state machine.
[cy]ATC: anatomical therapeutic chemical.
[cz]AFC: active force control.
[da]FDA: Food and Drug Administration.
[db]MMAT: Mixed Methods Appraisal Tool.
[dc]STARD: Standards for Reporting of Diagnostic Accuracy Study.
[dd]VR: virtual reality.
[de]EU: European Union.
[df]EQUATOR: Enhancing the Quality and Transparency of Health Research.
[dg]WHO-WEF: World Health Organization World Economic Forum.
[dh]CCC: concordance correlation coefficient.
[di]IEEE: Institute of Electrical and Electronics Engineers.

There were also other AI quality standards not mentioned in the reviews or their unique studies. They included guidelines such as the do no harm road map, Factor Analysis of Information Risk, HIPAA, and the FDA regulatory framework mentioned by van de Sande et al [23]; AI clinical study reporting guidelines such as Clinical Artificial Intelligence Modeling and Minimum Information About Clinical Artificial Intelligence Modeling mentioned by Shelmerdine et al [15]; and the international technical AI standards such as ISO and International Electrotechnical Commission 22989, 23053, 23894, 24027, 24028, 24029, and 24030 mentioned by Wenzel and Wiegand [26].

With these additional findings, we updated the original table of AI standards in the study by van de Sande et al [23] showing crucial steps and key documents by life cycle phase (Table 2).

XSL•FO
RenderX

Table 2. Use of health care standards in the reviews mapped to the life cycle phases by van de Sande et al [23].

	Standards and corresponding reviews[a]
Life cycle phase 0: preparation before AI[b] model development	
Define the problem and engage stakeholders	• Do no harm road map (Wiens)
Search for and evaluate available models	• FDA[c] devices (Benjamens) • ÉCLAIR[d] (Omoumi)
Identify, collect data, and account for bias	• FHIR[e] (Mandel) • FAIR[f] (Wilkinson) • Validation (Riley) • PROBAST[g] (Moons and Wolff): *Adamidi et al* [37][a], Barboi et al [38], Abbasgholizadeh Rahimi et al [58], and Sahu et al [59]
Handle privacy	• HIPAA[h] (OOTA[i]): Battineni et al [39] • GDPR[j] (EU[k])
Ethical principles, frameworks, and guidelines	• WMA[l] (Helsinki declaration): *Abd-Alrazaq et al* [36][a] • Ethics for mental health technology (WEF[m]): Zidaru et al [68] • Digital disease technology detection (SORMAS[n]): *Zhao et al* [13][a] • Asadi framework of ethics (Asadi): *Zhao et al* [15][a] • Ethical guidelines (EU): Chew and Achananuparp [44] • Ethical principles and framework (Montreal): Mörch et al [55] • Ethics and governance of AI health (WHO[o]): Zidaru et al [68]
Life cycle phase I: AI model development	
Check applicable regulations	• Proposed regulatory framework (FDA) • Harmonized rules on AI (EU)
Prepare data	• Preprocessing data (Ferrao)
Train and validate	• ML[p] cardiac imaging (Juarez-Orozco)
Evaluate performance and report results	• AI guide (Park and Han): Mörch et al [55] • TRIPOD[q] (Collins): Adamidi et al [37] • TRIPOD-ML[r] (Collins): Hassan et al [50] and Mörch et al [55] • CLAIM[s] (Mongan): Shelmerdine et al [15][t] • CHARMS[u] (Moons): Barboi et al [38] and Sahu et al [59] • PRISMA-DTA[v] (McInnes): Shelmerdine et al [15][t] • MI-CLAIM[w] (Norgeot): Shelmerdine et al [15][t] • MINIMAR[x] (Hernandez-Boussard): Shelmerdine et al [15][t] • NOS[y] (Lo): Hassan et al [50] and Battineni et al [39] • LOE[z] (Concato): *Seibert et al* [61][a] • MMAT[aa] (Hong): *Zhao et al* [66][a] • Clinical Prediction Rule Checklist (CASP[ab]): Bertini et al [40] • STARD[ac] (checklist): *Zhao et al* [66][a] • COREQ[ad] (checklist): *Buchanan et al* [42][a], *Kaelin et al* [52][a], and *Seibert et al* [61][a]
Life cycle phase II: assessment of AI performance and reliability	
Externally validate model or concept	• Validation (Ramspek and Riley) • Generalizability (Futoma) • Risk of bias (?): Talpur et al [63] • MADE1.0[ae] (Dandala): *Choudhury and Asan* [46][a]
Simulate results and prepare for clinical study	• DECIDE-AI[af] (steering group): Shelmerdine et al [15][t]

XSL•FO
RenderX

	Standards and corresponding reviews[a]
Life cycle phase III: clinically testing AI	
Design and conduct clinical study	• SPIRIT-AI[ag] (Cruz): Mörch et al [55] • Explanations (Barda) • CONSORT-AI[ah] (Liu): Loveys et al [54] • Revised Cochrane RoB 2[ai] (Sterne): Loveys et al [54] • ROBINS-I[aj] (RoB 2 for non-RCTs[ak]; Sterne): Loveys et al [54] • STROBE[al] (checklists): *Adamidi et al* [37][a]
Life cycle phase IV: implementing and governing AI	
Legal approval	• AI-ML[am] medical devices (Muehlematter)
Safely implement model	• TAM[an] (Jauk) • ML model facts (Sendak)
Model and data governance	• FAIR (Wilkinson): *Adamidi et al* [37][a] • SaMD[ao] clinical evaluation (FDA): *Adamidi et al* [37][a] • Quality management system (IMDRF[ap]): *Adamidi et al* [37][a]
Responsible model use	• Ethics ambient intelligence (Martinz-Martin)
Standards in the reviews mapped to multiple phases	

XSL•FO
RenderX

	Standards and corresponding reviews[a]
Design justice principles	• 10 principles—International Design Justice Network: Zidaru et al [68]
Study quality	• Reporting guidelines (EQUATOR[aq]): Zidaru et al [68]
Policy	• China—AI governance (Laskai): Choudhury et al [45] • Federal engagement plan (NIST[ar]): Choudhury et al [45] • Russian AI policy (OECD[as]): Choudhury et al [45] • AMA[at] AI recommendations (link): Sapci and Sapci [60] • CCC[au] AI road map (link): Zidaru et al [68][a] • EU (AI Watch): Zidaru et al [68][a]
Technical and interoperability	• Software life cycle ISO[av] and IEEE[aw]-12207 (link): Zidaru et al [68][a] • OGC[ax]: *Seibert et al* [61][a] • SWE[ay]: *Seibert et al* [61][a] • SOS[az]: *Seibert et al* [61][a] • AI concepts and terminology (ISO and IEC[ba] 22989): *Seibert et al* [61][a] • Framework for AI systems (ISO and IEC 23053): Wenzel and Wiegand [26][t] • AI risk management (ISO and IEC 23894): Wenzel and Wiegand [26][t] • AI bias (ISO and IEC 24027): Wenzel and Wiegand [26][t] • AI trustworthiness (ISO and IEC 24028): Wenzel and Wiegand [26][t] • AI robustness (ISO and IEC 24029-1): Wenzel and Wiegand [26][t] • AI use cases (ISO and IEC 24030): Wenzel and Wiegand [26][t] • Safety of machinery (ISO 22100-5): *Choudhury et al* [45][a]
Terminology standards	• FAERS[bb]: *Choudhury et al* [45][a] • Medication information extract clinical notes (MedEx): *Choudhury et al* [45][a] • Medical prescription normalized (RxNorm): *Choudhury et al* [45][a] • MedDRA[bc]: *Choudhury et al* [45][a] • Patient-centered clinical research network (PCORnet): *Choudhury et al* [45][a] • UMLS[bd]: *Choudhury et al* [45][a]
Robotics	

XSL•FO
RenderX

	Standards and corresponding reviews[a]
Partnership for R&D[be] and innovation	• European robotics partnership (SPARC[bf]): *Buchanan et al* [42][a]
Robotic standardization and safety	• Robotics standardization (ISO and TC[bg] 299): *Buchanan et al* [42][a]
Robotic devices for personal care	• Personal care robots (13482: 2014): *Buchanan et al* [42][a] • Vocabulary (ISO 8373: 2021): *Buchanan et al* [42][a]

[a]Italicized references are original studies cited in the reviews, and references denoted with the footnote *t* are those cited in our paper but not present in any of the reviews.
[b]AI: artificial intelligence.
[c]FDA: Food and Drug Administration.
[d]ECLAIR: Evaluate Commercial AI Solutions in Radiology.
[e]FHIR: Fast Healthcare Interoperability Resources.
[f]FAIR: Findability, Accessibility, Interoperability, and Reusability.
[g]PROBAST: Prediction Model Risk of Bias Assessment Tool.
[h]HIPAA: Health Insurance Portability and Accountability Act.
[i]OOTA: Office of The Assistant Secretary.
[j]GDPR: General Data Protection Regulation.
[k]EU: European Union.
[l]WMA: World Medical Association.
[m]WEF: World Economic Forum.
[n]SORMAS: Surveillance, Outbreak Response Management and Analysis System.
[o]WHO: World Health Organization.
[p]ML: machine learning.
[q]TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.
[r]TRIPOD-ML: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis—Machine Learning.
[s]CLAIM: Checklist for Artificial Intelligence in Medical Imaging.
[t]References denoted with the footnote *t* are those cited in our paper but not present in any of the reviews.
[u]CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies.
[v]PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy.
[w]MI-CLAIM: Minimum Information About Clinical Artificial Intelligence Modeling.
[x]MINIMAR: Minimum Information for Medical AI Reporting.
[y]NOS: Newcastle-Ottawa Scale.
[z]LOE: level of evidence.
[aa]MMAT: Mixed Methods Appraisal Tool.
[ab]CASP: Critical Appraisal Skills Programme.
[ac]STARD: Standards for Reporting of Diagnostic Accuracy Studies.
[ad]COREQ: Consolidated Criteria for Reporting Qualitative Research.
[ae]MADE1.0: Model Agnostic Diagnostic Engine 1.0.
[af]DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision-Support Systems Driven by Artificial Intelligence.
[ag]SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence.
[ah]CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence.
[ai]RoB 2: Risk of Bias 2.
[aj]ROBINS-I: Risk of Bias in Non-Randomised Studies of Interventions.
[ak]RCT: randomized controlled trial.
[al]STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.
[am]AI-ML: artificial intelligence–machine learning.
[an]TAM: Technology Acceptance Model.
[ao]SaMD: Software as a Medical Device.
[ap]IMDRF: International Medical Device Regulators Forum.
[aq]EQUATOR: Enhancing the Quality and Transparency of Health Research.
[ar]NIST: National Institute of Standards and Technology.
[as]OECD: Organisation for Economic Co-operation and Development.

XSL•FO
RenderX

[at]AMA: American Medical Association.
[au]CCC: Computing Community Consortium.
[av]ISO: International Organization for Standardization.
[aw]IEEE: Institute of Electrical and Electronics Engineers.
[ax]OGC: Open Geospatial Consortium.
[ay]SWE: Sensor Web Enablement.
[az]SOS: Sensor Observation Service.
[ba]IEC: International Electrotechnical Commission.
[bb]FAERS: Food and Drug Administration Adverse Event Reporting System.
[bc]MedDRA: Medical Dictionary for Regulatory Activities.
[bd]UMLS: Unified Medical Language System.
[be]R&D: research and development.
[bf]SPARC: Scholarly Publishing and Academic Resources Coalition.
[bg]TC: technical committee.

Quality Standard–Related Issues

We extracted a set of AI quality standard–related issues from the 33 reviews and assigned themes based on keywords used in the reviews (Multimedia Appendix 8 [36-68]). In total, we identified 23 issues, with the most frequently mentioned ones being clinical utility and economic benefits (n=10); ethics (n=10); benchmarks for data, model, and performance (n=9); privacy, security, data protection, and access (n=8); and federated learning and integration (n=8). Table 3 shows the quality standard issues by theme from the 33 reviews. To provide a framing and means of conceptualizing the quality-related issues, we did a high-level mapping of the issues to the AI requirements proposed by the NAM [8] and EUC [20]. The mapping was done by 2 of the authors, with the remaining authors validating the results. Final mapping was the result of consensus across the authors (Table 4).

XSL•FO
RenderX

Table 3. Summary of quality standard–related issues in the reviews.

Key themes	Quality issues	Reviews
Ethics	Guidelines needed, 10 issues—prudence, equity, privacy and intimacy, democratic participation, solidarity, responsibility, diversity inclusion, well-being, respect for autonomy, and sustainable development (Mörch et al [55]); the individual, organizational and society levels of the ethical framework by Asadi et al (Zhao et al [66])	• Abd-Alrazaq et al [36] • Bertini et al [40] • Buchanan et al [43] • Le Glaz et al [48] • Loveys et al [54] • Mörch et al [55] • Abbasgholizadeh Rahimi et al [58] • Seibert et al [61] • Zhao et al [66] • Zidaru et al 69]
Benefits, cost-effective, economic, external and clinical validation, and clinical utility	Need clinical validation and demonstration of economic benefits and clinical utility in real-world settings	• Bertini et al [40] • Eldaly et al [47] • Kirk et al [53] • Abbasgholizadeh Rahimi et al [58] • Sahu et al [59] • Seibert et al [61] • Syeda et al [62] • Vélez-Guerrero et al [64] • Welch et al [65] • Zidaru et al [68]
Benchmarks—models, data, and performance	Need standardized and comparable AI[a] models, parameters, evaluation measures, and gold standards	• Barboi et al [38] • Bertini et al [40] • Choudhury et al [45] • Choudhury and Asan [46] • Hassan et al [50] • Kirk et al [53] • Syeda et al [62]
Integration, federated learning, decision fusion, and ability to aggregate results	Need to integrate heterogeneous data from multiple sources and combine multiple classifier outputs into a common decision	• Abd-Alrazaq et al [36] • Adamidi et al [37] • Bhatt et al [41] • Guo et al [49] • Kirk et al [53] • Popescu et al [57] • Vélez-Guerrero et al [64] • Zheng et al [67]
Privacy, security, open data, access, and protection	Need agreements and processes on privacy, security, access, open data, and data protection	• Buchanan et al [43] • Choudhury and Asan [46] • Le Glaz et al [48] • Guo et al [49] • Loveys et al [54] • Seibert et al [61]
Education, web-based learning, learning experience, and competencies	Need education for the public, patients, students, and providers, including web-based learning and building competencies and as part of formal curricula	• Abd-Alrazaq et al [36] • Chew and Achananuparp [44] • Choudhury and Asan [46] • Kirk et al [53] • Sapci and Sapci [60] • Seibert et al [61]
Explainability	Enhance acceptability, understandability, and interpretability of solutions and ability to convey them to patients	• Abd-Alrazaq et al [36] • Adamidi et al [37] • Bhatt et al [41] • Kirk et al [53] • Syeda et al [62] • Vélez-Guerrero et al [64]

XSL•FO
RenderX

Key themes	Quality issues	Reviews
Reporting standards	Standardized reporting of study details to allow for comparison and replication	• Abd-Alrazaq et al [36] • Barboi et al [38] • Bertini et al [40] • Choudhury and Asan [46] • Abbasgholizadeh Rahimi et al [58] • Zheng et al [67]
Transparency	Need openness and being accountable through the entire AI life cycle	• Adamidi et al [37] • Barboi et al [38] • Bhatt et al [41] • Chew and Achananuparp [44] • Le Glaz et al [48] • Huang et al [51]
Trust and trustworthiness	Need ethical guidelines to ensure confidence, truthfulness, and honesty with the design, use, and impact of AI systems	• Adamidi et al [37] • Bhatt et al [41] • Chew and Achananuparp [44] • Huang et al [51] • Abbasgholizadeh Rahimi et al [58] • Zidaru et al [68]
Safety	Need to ensure patient safety from harm	• Chew and Achananuparp [44] • Choudhury et al [45] • Choudhury and Asan [46] • Seibert et al [61] • Zidaru et al [68]
Bias—SDOH[b] and assessment	Need to consider sociodemographic variables and adequate sample sizes	• Adamidi et al [37] • Le Glaz et al [48] • Abbasgholizadeh Rahimi et al [58] • Syeda et al [62] • Vélez-Guerrero et al [64]
Co-design and engagement—user, provider, and public	Meaningful participation at all life cycle stages	• Buchanan et al [42] • Buchanan et al [43] • Huang et al [51] • Abbasgholizadeh Rahimi et al [58] • Zidaru et al [68]
Technology maturity or feasibility, acceptance, and usability	Need user-friendly and mature AI systems with proven benefits to increase adoption	• Chew and Achananuparp [44] • Eldaly et al [47] • Seibert et al [61]
Regulation and legal	Need legal framework and laws to ensure appropriate safe use and liability protection	• Bhatt et al [41] • Choudhury and Asan [46] • Abbasgholizadeh Rahimi et al [58] • Seibert et al [61]
Context and time dependency	Purpose of AI models, health care context, and time lags have mediating effect	• Choudhury et al [45] • Kaelin et al [52] • Payedimarri et al [56]
Data integration and preprocessing	Need to integrate different variables and include multilevel data preprocessing to reduce dimensionality	• Guo et al [49] • Kirk et al [53]
Design justice, equity, and fairness	Need design justice principles to engage the public and ensure a fair and equitable AI system	• Buchanan et al [43] • Zidaru et al [68]
Personalized care and targeted interventions	Select the best AI algorithms and outputs to customize care for specific individuals	• Battineni et al [39] • Kaelin et al [52]
Quality—data and study	Need well-designed studies and quality data to conduct AI studies	• Talpur et al [63] • Welch et al [65] • Buchanan et al [43]

XSL•FO
RenderX

Key themes	Quality issues	Reviews
Social justice and social implications	Need to balance human caring needs with AI advances, understanding the societal impact of AI interventions	
Governance	Need governance on the collection, storage, use, and transfer of data; being accountable and transparent with the process	• Choudhury et al [45]
Self-adaptability	Need adaptable and flexible AI systems that can improve over time	• Vélez-Guerrero et al [64]

[a]AI: artificial intelligence.

[b]SDOH: social determinants of health.

XSL•FO
RenderX

Table 4. Quality standard–related issues by theme mapped to the National Academy of Medicine (NAM) and European Union Commission (EUC) requirements.

Key themes	NAM[a]	EUC[b]	Reviews
Ethics	T6-2: ethics and fairness	1—rights, agency, and oversight; 7—minimizing and reporting negative impact	• Abd-Alrazaq et al [36] • Bertini et al [40] • Buchanan et al [43] • Le Glaz et al [48] • Loveys et al [54] • Mörch et al [55] • Abbasgholizadeh Rahimi et al [58] • Seibert et al [61] • Zhao et al [66] • Zidaru et al [68]
Benefits, cost-effective, economic, external and clinical validation, and clinical utility	B5-1: accuracy and outcome change; T6-2: cost, revenue and value, and safety and efficacy; and T6-3: improvement and assessments	6—environmentally friendly and sustainable; 7—documenting trade-off and ability to redress	• Bertini et al [40] • Eldaly et al [47] • Kirk et al [53] • Abbasgholizadeh Rahimi et al [58] • Sahu et al [59] • Seibert et al [61] • Syeda et al [62] • Vélez-Guerrero et al [64] • Welch et al [65] • Zidaru et al [68]
Benchmarks—models, data, and performance	B5-1: accuracy and outcome change; T6-2: cost, revenue and value, and safety and efficacy; and T6-3: defining success and after-action assessments	2—accuracy, reliability, and reproducibility; 3—data quality and integrity	• Barboi et al [38] • Bertini et al [40] • Choudhury et al [45] • Choudhury and Asan [46] • Hassan et al [50] • Kirk et al [53] • Syeda et al [62]
Integration, federated learning, decision fusion, and ability to aggregate results	T6-2: data environment and interoperability	3—data governance; 4—traceability and explainability	• Abd-Alrazaq et al [36] • Adamidi et al [37] • Bhatt et al [41] • Guo et al [49] • Kirk et al [53] • Popescu et al [57] • Vélez-Guerrero et al [64] • Zheng et al [67]
Privacy, security, open data, access, and protection	T6-2: cybersecurity and privacy	2—resilience; 3—data privacy, protection, and access	• Buchanan et al [43] • Choudhury and Asan [46] • Le Glaz et al [48] • Guo et al [49] • Loveys et al [54] • Seibert et al [61]
Education, web-based learning, learning experience, and competencies	T6-3: education and support	7—minimizing and reporting negative impact	• Abd-Alrazaq et al [36] • Chew and Achananuparp [44] • Choudhury and Asan [46] • Kirk et al [53] • Sapci and Sapci [60] • Seibert et al [61]
Explainability	N/A[c]	4—explainability and communication	• Abd-Alrazaq et al [36] • Adamidi et al [37] • Bhatt et al [41] • Kirk et al [53] • Syeda et al [62] • Vélez-Guerrero et al [64]

XSL•FO
RenderX

Key themes	NAM[a]	EUC[b]	Reviews
Reporting standards	__[d]	3—privacy and data protection; 7—auditability, documenting trade-offs, and ab	• Abd-Alrazaq et al [36] • Barboi et al [38] • Bertini et al [40] • Choudhury and Asan [46] • Abbasgholizadeh Rahimi et al [58] • Zheng et al [67]
Transparency	—	4—traceability and communication	• Adamidi et al [37] • Barboi et al [38] • Bhatt et al [41] • Chew and Achananuparp [44] • Le Glaz et al [48] • Huang et al [51]
Trust and trustworthiness	—	All 7 assessment list items	• Adamidi et al [37] • Bhatt et al [41] • Chew and Achananuparp [44] • Huang et al [51] • Abbasgholizadeh Rahimi et al [58] • Zidaru et al [68]
Safety	T6-2: safety and efficacy	2—resilience and safety	• Chew and Achananuparp [44] • Choudhury et al [45] • Choudhury and Asan [46] • Seibert et al [61] • Zidaru et al [68]
Bias—SDOH[e] and assessment	B5-1: accuracy	5—bias avoidance	• Adamidi et al [37] • Le Glaz et al [48] • Abbasgholizadeh Rahimi et al [58] • Syeda et al [62] • Vélez-Guerrero et al [64]
Co-design and engagement—user, provider, and public	B5-1: target users; T6-2: patient, family, and consumer engagement; and T6-3: stakeholder consensus	1—rights, agency, and oversight; 5—participation	• Buchanan et al [42] • Buchanan et al [43] • Huang et al [51] • Abbasgholizadeh Rahimi et al [58] • Zidaru et al [68]
Technology maturity or feasibility, acceptance, and usability	—	5—accessibility and universal design	• Chew and Achananuparp [44] • Eldaly et al [47] • Seibert et al [61]
Regulation and legal	T6-2: regulatory issues	7—audibility and minimizing and reporting negative impact	• Bhatt et al [41] • Choudhury and Asan [46] • Abbasgholizadeh Rahimi et al [58] • Seibert et al [61]
Context and time dependency	T6-3: problem identification	2—accuracy, reliability, and reproducibility	• Choudhury et al [45] • Kaelin et al [52] • Payedimarri et al [56]
Data integration and preprocessing	—	3—data quality and integrity	• Guo et al [49] • Kirk et al [53]
Design justice, equity, and fairness	T6-2: ethics and fairness	1—rights, agency, and oversight; 6—environmentally friendly and sustainable	• Buchanan et al [43] • Zidaru et al [68]
Personalized care and targeted interventions	B5-1: downstream interventions, target users, and capacity to intervene	1—rights, agency, and oversight; 5—bias avoidance, universal design, and accessibility	• Battineni et al [39] • Kaelin et al [52]
Quality—data and study	—	3—data quality and integrity	• Talpur et al [63] • Welch et al [65]

XSL•FO
RenderX

Key themes	NAM[a]	EUC[b]	Reviews
Social justice and social implications	B5-1: downstream interventions and desired outcome change	1—rights, agency, and oversight; 6—social impact and society and democracy	• Buchanan et al [43]
Governance	T6-2: organizational capabilities, data environment, and personal capacity	3—data quality and integrity; data access	• Choudhury et al [45]
Self-adaptability	—	1—oversight	• Vélez-Guerrero et al [64]

[a]B5-1: key considerations in model development; T6-2: key considerations for institutional infrastructure and governance; and T6-3: key artificial intelligence tool implementation concepts, considerations, and tasks.

[b]1—human agency and oversight; 2—technical robustness and safety; 3—privacy and data governance; 4—transparency; 5—diversity, nondiscrimination, and fairness; 6—societal and environmental well-being; and 7—accountability.

[c]N/A: not applicable.

[d]Themes not addressed.

[e]SDOH: social determinants of health.

We found that all 23 quality standard issues were covered in the AI frameworks by the NAM and EUC. Both frameworks have a detailed set of guidelines and questions to be considered at different life cycle stages of the health care AI studies. While there was consistency in the mapping of the AI issues to the NAM and EUC frameworks, there were some differences across them. Regarding the NAM, the focus was on key aspects of AI model development, infrastructure and governance, and implementation tasks. Regarding the EUC, the emphasis was on achieving trustworthiness by addressing all 7 interconnected requirements of accountability; human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, nondiscrimination, and fairness; and societal and environmental well-being. The quality standard issues were based on our analysis of the review articles, and our mapping was at times more granular than the issues from the NAM and EUC frameworks. However, our results showed that the 2 frameworks do provide sufficient terminology for quality standard–related issues. By embracing these guidelines, one can enhance the buy-in and adoption of the AI interventions in the health care system.

Discussion

Principal Findings

Overall, we found that, despite the growing number of health care AI quality standards in the literature, they are seldom applied in practice, as is shown in a sample of recently published systematic reviews of health care AI studies. Of the reviews that mentioned AI quality standards, most were used to ensure the methodological and reporting quality of the AI studies involved. At the same time, the reviews identified many AI quality standard–related issues, including those broader in nature, such as ethics, regulations, transparency, interoperability, safety, and governance. Examples of broader standards mentioned in a handful of reviews or original studies are the ISO-12207, Unified Medical Language System, HIPAA, FDA Software as a Medical Device, World Health Organization AI governance, and American Medical Association augmented intelligence recommendations. These findings reflect the evolving nature of health care AI, which has not yet reached maturity or been widely adopted. There is a need to apply appropriate AI quality standards to demonstrate the transparency, robustness, and benefits of these AI approaches in different AI domains and health topics while protecting the privacy, safety, and rights of individuals and society from the potential unintended consequences of such innovations.

Another contribution of our study was a conceptual reframing for a systems-based perspective to harmonize health care AI. We did not look at AI studies solely as individual entities but rather as part of a bigger system that includes clinical, organizational, and societal aspects. Our findings complement those of recent publications, such as an FDA paper that advocates for a need to help people understand the broader system of AI in health care, including across different clinical settings [72]. Moving forward, we advocate for AI research that looks at how AI approaches will mature over time. AI approaches evolve through different phases of maturity as they move from development to validation to implementation. Each phase of maturity has different requirements [23] that must be assessed as part of evaluating AI approaches across domains as the number of health care applications rapidly increases [73]. However, comparing AI life cycle maturity across studies was challenging as there were a variety of life cycle terms used across the reviews, making it hard to compare life cycle maturity in and across studies. To address this issue, we provided a mapping of life cycle terms from the original studies but also used the system life cycle phases by van de Sande et al [23] as a common terminology for AI life cycle stages. A significant finding from the mapping was that most AI studies in our selected reviews were still at early stages of maturity (ie, model preparation, development, or validation), with very few studies progressing to later phases of maturity such as clinical testing and implementation. If AI research in health systems is to evolve, we need to move past single-case studies with external data validation to studies that achieve higher levels of life cycle maturity, such as clinical testing and implementation over a variety of routine health care settings (eg, hospitals, clinics, and patient homes and other community settings).

Our findings also highlighted that there are many AI approaches and quality standards used across domains in health care AI studies. To better understand their relationships and the overall construct of the approach, our applied conceptual organizing

XSL•FO
RenderX

scheme for harmonized health care characterizes AI studies according to AI domains, approaches, health topics, life cycle phases, and quality standards. The health care AI landscape is complex. The Euler diagram shows multiple AI approaches in one or more AI domains for a given health topic. These domains can overlap, and the AI approaches can be driven by ML, DL, or other types (eg, decision trees, robotics). This complexity is expected to increase as the number of AI approaches and range of applications across all health topics and settings grows over time. For meaningful comparison, we need a harmonized scheme such as the one described in this paper to make sense of the multitude of AI terminology for the types of approaches reported in the health care AI literature. The systems-based perspective in this review provides the means for harmonizing AI life cycles and incorporating quality standards through different maturity stages, which could help advance health care AI research by scaling up to clinical validation and implementation in routine practice. Furthermore, we need to move toward explainable AI approaches where applications are based on clinical models if we are to move toward later stages of AI maturity in health care (eg, clinical validation, and implementation) [74].

Proposed Guidance

To improve the quality of future health care AI studies, we urge AI practitioners and researchers to draw on published health care AI quality standard literature, such as those identified in this review. The type of quality standards to be considered should cover the trustworthiness, methodological, reporting, and technical aspects. Examples include the NAM and EUC AI frameworks that address trustworthiness and the EQUATOR network with its catalog of methodological and reporting guidelines identified in this review. Also included are the Minimum Information for Medical AI Reporting guidelines and technical ISO standards (eg, robotics) that are not in the EQUATOR. Components that should be standardized are the AI ethics, approaches, life cycle stages, and performance measures used in AI studies to facilitate their meaningful comparison and aggregation. The technical standards should address such key design features as data, interoperability, and robotics. Given the complexities of the different AI approaches involved, rather than focusing on the underlying model or algorithm design, one should compare their actual performance based on life cycle stages (eg, degree of accuracy in model development or assessment vs outcome improvement in implementation). The summary list of the AI quality standards described in this paper is provided in Multimedia Appendix 9 for those wishing to apply them in future studies.

Implications

Our review has practice, policy, and research implications. For practice, better application of health care AI quality standards could help AI practitioners and researchers become more confident regarding the rigor and transparency of their health care AI studies. Developers adhering to standards may help make AI approaches in domains less of a black box and reduce unintended consequences such as systemic bias or threats to patient safety. AI standards may help health care providers better understand, trust, and apply the study findings in relevant clinical settings. For policy, these standards can provide the necessary guidance to address the broader impacts of health care AI, such as the issues of data governance, privacy, patient safety, and ethics. For research, AI quality standards can help advance the field by improving the rigor, reproducibility, and transparency in the planning, design, conduct, reporting, and appraisal of health care AI studies. Standardization would also allow for the meaningful comparison and aggregation of different health care AI studies to expand the evidence base in terms of their performance impacts, such as cost-effectiveness, and clinical outcomes.

Limitations

Despite our best effort, this umbrella review has limitations. First, we only searched for peer-reviewed English articles with “health” and “AI” as the keywords in MEDLINE and Google Scholar covering a 36-month period. It is possible to have missed relevant or important reviews that did not meet our inclusion criteria. Second, some of the AI quality standards were only published in the last few years, at approximately the same time when the AI reviews were conducted. As such, it is possible for AI review and study authors to have been unaware of these standards or the need to apply them. Third, the AI standard landscape is still evolving; thus, there are likely standards that we missed in this review (eg, Digital Imaging and Communications in Medicine in pattern recognition with convolutional neural networks [75]). Fourth, the broader socioethical guidelines are still in the early stages of being refined, operationalized, and adopted. They may not yet be in a form that can be easily applied when compared with the more established methodological and reporting standards with explicit checklists and criteria. Fifth, our literature review did not include any literature reviews on LLMs [76], and we know there are reviews of LLMs published in 2023 and beyond. Nevertheless, our categorization of NLP could coincide with NLP and DL in our Euler diagram, and furthermore, LLMs could be used in health care via approved chatbot applications at an early life cycle phase, for example, using decision trees first to prototype the chatbot as clinical decision support [77] before advancing it in the mature phase toward a more robust AI solution in health care with LLMs. Finally, only one author was involved in screening citation titles and abstracts (although 2 were later involved in full-text review of all articles that were screened in), and there is the possibility that we erroneously excluded an article on the basis of title and abstract. Despite these limitations, this umbrella review provided a snapshot of the current state of knowledge and gaps that exist with respect to the use of and need for AI quality standards in health care AI studies.

Conclusions

Despite the growing number of AI standards to assess the quality of health care AI studies, they are seldom applied in practice. With the recent unveiling of broader ethical guidelines such as those of the NAM and EUC, more transparency and guidance in health care AI use are needed. The key contribution of this review was the harmonization of different AI quality standards that could help practitioners, developers, and users understand the relationships among AI domains, approaches, life cycles, and standards. Specifically, we advocate for common terminology on AI life cycles to enable comparison of AI

XSL•FO
RenderX

maturity across stages and settings and ensure that AI research scales up to clinical validation and implementation.

Acknowledgments

CK acknowledges funding support from a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN/04884-2019). The authors affirm that no generative artificial intelligence tools were used in the writing of this manuscript.

Authors' Contributions

CK contributed to conceptualization (equal), methodology (equal), data curation (equal), formal analysis (equal), investigation (equal), and writing—original draft (lead). DC contributed to conceptualization (equal), methodology (equal), data curation (equal), formal analysis (equal), investigation (equal), and visualization (equal). SM contributed to conceptualization (equal), methodology (equal), data curation (equal), formal analysis (equal), investigation (equal), and visualization (equal). MM contributed to conceptualization (equal), methodology (equal), data curation (equal), formal analysis (equal), and investigation (equal). FL contributed to conceptualization (equal), methodology (lead), data curation (lead), formal analysis (lead), investigation (equal), writing—original draft (equal), visualization (equal), project administration (lead), and supervision (lead).

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 31 KB - jmir_v26i1e54705_app1.docx]

Multimedia Appendix 2
PubMed search strings.
[DOCX File , 32 KB - jmir_v26i1e54705_app2.docx]

Multimedia Appendix 3
Characteristics of the included reviews.
[XLSX File (Microsoft Excel File), 26 KB - jmir_v26i1e54705_app3.xlsx]

Multimedia Appendix 4
List of excluded reviews and reasons.
[XLSX File (Microsoft Excel File), 16 KB - jmir_v26i1e54705_app4.xlsx]

Multimedia Appendix 5
Quality of the included reviews using Joanna Briggs Institute scores.
[XLSX File (Microsoft Excel File), 763 KB - jmir_v26i1e54705_app5.xlsx]

Multimedia Appendix 6
Health care artificial intelligence reviews by life cycle stage.
[DOCX File , 51 KB - jmir_v26i1e54705_app6.docx]

Multimedia Appendix 7
Quality standards found in 10% of unique studies in the selected reviews.
[XLSX File (Microsoft Excel File), 305 KB - jmir_v26i1e54705_app7.xlsx]

Multimedia Appendix 8
Quality standard–related issues mentioned in the artificial intelligence reviews.
[DOCX File , 54 KB - jmir_v26i1e54705_app8.docx]

Multimedia Appendix 9
Summary list of artificial intelligence quality standards.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v26i1e54705_app9.xlsx]

XSL•FO
RenderX

References

1. Saleh L, Mcheick H, Ajami H, Mili H, Dargham J. Comparison of machine learning algorithms to increase prediction accuracy of COPD domain. In: Proceedings of the 15th International Conference on Enhanced Quality of Life and Smart Living. 2017 Presented at: ICOST '17; August 29-31, 2017; Paris, France p. 247-254 URL: https://doi.org/10.1007/978-3-319-66188-9_22 [doi: 10.1007/978-3-319-66188-9_22]
2. Gerke S, Minssen T, Cohen IG. Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc 2020:295-336 [FREE Full text] [doi: 10.2139/ssrn.3570129]
3. Čartolovni A, Tomičić A, Lazić Mosler E. Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review. Int J Med Inform 2022 May;161:104738. [doi: 10.1016/j.ijmedinf.2022.104738] [Medline: 35299098]
4. Das K, Cockerell CJ, Patil A, Pietkiewicz P, Giulini M, Grabbe S, et al. Machine learning and its application in skin cancer. Int J Environ Res Public Health 2021 Dec 20;18(24):13409 [FREE Full text] [doi: 10.3390/ijerph182413409] [Medline: 34949015]
5. Elkin P, Mullin S, Mardekian J, Crowner C, Sakilay S, Sinha S, et al. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: evaluation and case-control study. J Med Internet Res 2021 Nov 09;23(11):e28946 [FREE Full text] [doi: 10.2196/28946] [Medline: 34751659]
6. Kawamoto A, Takenaka K, Okamoto R, Watanabe M, Ohtsuka K. Systematic review of artificial intelligence-based image diagnosis for inflammatory bowel disease. Dig Endosc 2022 Nov;34(7):1311-1319 [FREE Full text] [doi: 10.1111/den.14334] [Medline: 35441381]
7. Anderson C, Bekele Z, Qiu Y, Tschannen D, Dinov ID. Modeling and prediction of pressure injury in hospitalized patients using artificial intelligence. BMC Med Inform Decis Mak 2021 Aug 30;21(1):253 [FREE Full text] [doi: 10.1186/s12911-021-01608-5] [Medline: 34461876]
8. Matheny M, Israni ST, Ahmed M, Whicher D. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC: National Academy of Medicine; 2019.
9. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open 2020 Oct;3(3):326-331 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa033] [Medline: 33215066]
10. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc 2022 Apr 13;29(5):983-989 [FREE Full text] [doi: 10.1093/jamia/ocac002] [Medline: 35045179]
11. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019 Dec 01;26(12):1651-1654 [FREE Full text] [doi: 10.1093/jamia/ocz130] [Medline: 31373357]
12. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
13. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]
14. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc 2018 Aug 01;25(8):963-968 [FREE Full text] [doi: 10.1093/jamia/ocy028] [Medline: 29669066]
15. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform 2021 Aug 23;28(1):e100385 [FREE Full text] [doi: 10.1136/bmjhci-2021-100385] [Medline: 34426417]
16. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ 2020 Sep 09;370:m3164-m3148 [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]
17. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. BMJ 2020 Sep 09;370:m3210 [FREE Full text] [doi: 10.1136/bmj.m3210] [Medline: 32907797]
18. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, Shah NH. MINIMAR (MINimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020 Dec 09;27(12):2011-2015 [FREE Full text] [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]
19. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020 Sep;26(9):1320-1324 [FREE Full text] [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]
20. Ethics and governance of artificial intelligence for health: WHO guidance. Licence: CC BY-NC-SA 3.0 IGO. World Health Organization. URL: https://apps.who.int/iris/bitstream/handle/10665/341996/9789240029200-eng.pdf [accessed 2024-04-05]
21. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI Ethics 2022 Jul 19;3(1):223-240 [FREE Full text] [doi: 10.1007/s43681-022-00195-z]

XSL•FO
RenderX

22. Joshi I, Morley J. Artificial intelligence: how to get it right: putting policy into practice for safe data-driven innovation in health and care. National Health Service. 2019. URL: https://transform.england.nhs.uk/media/documents/NHSX_AI_report.pdf [accessed 2024-04-05]
23. van de Sande D, Van Genderen ME, Smit JM, Huiskens J, Visser JJ, Veen RE, et al. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ Health Care Inform 2022 Feb 19;29(1):e100495 [FREE Full text] [doi: 10.1136/bmjhci-2021-100495] [Medline: 35185012]
24. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022 Jan 13;22(1):12 [FREE Full text] [doi: 10.1186/s12874-021-01469-6] [Medline: 35026997]
25. Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open 2020 Mar 23;10(3):e034568 [FREE Full text] [doi: 10.1136/bmjopen-2019-034568] [Medline: 32205374]
26. Wenzel MA, Wiegand T. Towards international standards for the evaluation of artificial intelligence for health. In: Proceedings of the 2019 ITU Kaleidoscope: ICT for Health: Networks, Standards and Innovation. 2019 Presented at: ITU K '19; December 4-6, 2019; Atlanta, GA p. 1-10 URL: https://ieeexplore.ieee.org/abstract/document/8996131 [doi: 10.23919/ituk48006.2019.8996131]
27. Varghese J. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. Visc Med 2020 Dec;36(6):443-449 [FREE Full text] [doi: 10.1159/000511930] [Medline: 33442551]
28. Nyarairo M, Emami E, Abbasgholizadeh S. Integrating equity, diversity and inclusion throughout the lifecycle of artificial intelligence in health. In: Proceedings of the 13th Augmented Human International Conference. 2022 Presented at: AH '22; May 26-27, 2022; Winnipeg, MB p. 1-4 URL: https://dl.acm.org/doi/abs/10.1145/3532530.3539565 [doi: 10.1145/3532530.3539565]
29. MacEntee MI. A typology of systematic reviews for synthesising evidence on health care. Gerodontology 2019 Dec 06;36(4):303-312. [doi: 10.1111/ger.12439] [Medline: 31691365]
30. Aromataris E, Fernandez R, Godfrey C, Holly C, Khalil H, Tungpunkom P. Umbrella reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, South Australia: Joanna Briggs Institute; 2020.
31. Tricco AC, Langlois EV, Straus SE. Rapid reviews to strengthen health policy and systems: a practical guide. Licence CC BY-NC-SA 3.0 IGO. World Health Organization. 2017. URL: https://apps.who.int/iris/handle/10665/258698 [accessed 2024-04-05]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4th edition. London, UK: Pearson Education; 2021.
34. Ethics guidelines for trustworthy AI. European Commission, Directorate-General for Communications Networks, Content and Technology. URL: https://data.europa.eu/doi/10.2759/346720 [accessed 2024-04-05]
35. Lloyd N, Khuman AS. AI in healthcare: malignant or benign? In: Chen T, Carter J, Mahmud M, Khuman AS, editors. Artificial Intelligence in Healthcare: Recent Applications and Developments. Singapore, Singapore: Springer; 2022:1-46.
36. Abd-Alrazaq A, Alajlani M, Alhuwail D, Schneider J, Al-Kuwari S, Shah Z, et al. Artificial intelligence in the fight against COVID-19: scoping review. J Med Internet Res 2020 Dec 15;22(12):e20756 [FREE Full text] [doi: 10.2196/20756] [Medline: 33284779]
37. Adamidi ES, Mitsis K, Nikita KS. Artificial intelligence in clinical care amidst COVID-19 pandemic: a systematic review. Comput Struct Biotechnol J 2021;19:2833-2850 [FREE Full text] [doi: 10.1016/j.csbj.2021.05.010] [Medline: 34025952]
38. Barboi C, Tzavelis A, Muhammad LN. Comparison of severity of illness scores and artificial intelligence models that are predictive of intensive care unit mortality: meta-analysis and review of the literature. JMIR Med Inform 2022 May 31;10(5):e35293 [FREE Full text] [doi: 10.2196/35293] [Medline: 35639445]
39. Battineni G, Hossain MA, Chintalapudi N, Amenta F. A survey on the role of artificial intelligence in biobanking studies: a systematic review. Diagnostics (Basel) 2022 May 09;12(5):1179 [FREE Full text] [doi: 10.3390/diagnostics12051179] [Medline: 35626333]
40. Bertini A, Salas R, Chabert S, Sobrevia L, Pardo F. Using machine learning to predict complications in pregnancy: a systematic review. Front Bioeng Biotechnol 2021 Jan 19;9:780389 [FREE Full text] [doi: 10.3389/fbioe.2021.780389] [Medline: 35127665]
41. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mHealth: scoping review. JMIR Mhealth Uhealth 2022 Jun 09;10(6):e35053 [FREE Full text] [doi: 10.2196/35053] [Medline: 35679107]
42. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on the domains of nursing: scoping review. JMIR Nurs 2020 Dec 17;3(1):e23939 [FREE Full text] [doi: 10.2196/23939] [Medline: 34406963]

XSL•FO
RenderX

43. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on nursing education: scoping review. JMIR Nurs 2021 Jan 28;4(1):e23933 [FREE Full text] [doi: 10.2196/23933] [Medline: 34345794]
44. Chew HS, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. J Med Internet Res 2022 Jan 14;24(1):e32939 [FREE Full text] [doi: 10.2196/32939] [Medline: 35029538]
45. Choudhury A, Renjilian E, Asan O. Use of machine learning in geriatric clinical care for chronic diseases: a systematic literature review. JAMIA Open 2020 Oct;3(3):459-471 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa034] [Medline: 33215079]
46. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599 [FREE Full text] [doi: 10.2196/18599] [Medline: 32706688]
47. Eldaly AS, Avila FR, Torres-Guzman RA, Maita K, Garcia JP, Serrano LP, et al. Artificial intelligence and lymphedema: state of the art. J Clin Transl Res 2022 Jun 29;8(3):234-242 [FREE Full text] [Medline: 35813896]
48. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 04;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
49. Guo Y, Zhang Y, Lyu T, Prosperi M, Wang F, Xu H, et al. The application of artificial intelligence and data integration in COVID-19 studies: a scoping review. J Am Med Inform Assoc 2021 Aug 13;28(9):2050-2067 [FREE Full text] [doi: 10.1093/jamia/ocab098] [Medline: 34151987]
50. Hassan N, Slight R, Weiand D, Vellinga A, Morgan G, Aboushareb F, et al. Preventing sepsis; how can artificial intelligence inform the clinical decision-making process? A systematic review. Int J Med Inform 2021 Jun;150:104457. [doi: 10.1016/j.ijmedinf.2021.104457] [Medline: 33878596]
51. Huang JA, Hartanti IR, Colin MN, Pitaloka DA. Telemedicine and artificial intelligence to support self-isolation of COVID-19 patients: recent updates and challenges. Digit Health 2022 May 15;8:20552076221100634 [FREE Full text] [doi: 10.1177/20552076221100634] [Medline: 35603328]
52. Kaelin VC, Valizadeh M, Salgado Z, Parde N, Khetani MA. Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: scoping review. J Med Internet Res 2021 Nov 04;23(11):e25745 [FREE Full text] [doi: 10.2196/25745] [Medline: 34734833]
53. Kirk D, Catal C, Tekinerdogan B. Precision nutrition: a systematic literature review. Comput Biol Med 2021 Jun;133:104365 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104365] [Medline: 33866251]
54. Loveys K, Prina M, Axford C, Domènec Ò, Weng W, Broadbent E, et al. Artificial intelligence for older people receiving long-term care: a systematic review of acceptability and effectiveness studies. Lancet Healthy Longev 2022 Apr;3(4):e286-e297 [FREE Full text] [doi: 10.1016/S2666-7568(22)00034-4] [Medline: 35515814]
55. Mörch CM, Atsu S, Cai W, Li X, Madathil SA, Liu X, et al. Artificial intelligence and ethics in dentistry: a scoping review. J Dent Res 2021 Dec 01;100(13):1452-1460. [doi: 10.1177/00220345211013808] [Medline: 34060359]
56. Payedimarri AB, Concina D, Portinale L, Canonico M, Seys D, Vanhaecht K, et al. Prediction models for public health containment measures on COVID-19 using artificial intelligence and machine learning: a systematic review. Int J Environ Res Public Health 2021 Apr 23;18(9):4499 [FREE Full text] [doi: 10.3390/ijerph18094499] [Medline: 33922693]
57. Popescu D, El-Khatib M, El-Khatib H, Ichim L. New trends in melanoma detection using neural networks: a systematic review. Sensors (Basel) 2022 Jan 10;22(2):496 [FREE Full text] [doi: 10.3390/s22020496] [Medline: 35062458]
58. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HT, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 03;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]
59. Sahu P, Raj Stanly EA, Simon Lewis LE, Prabhu K, Rao M, Kunhikatta V. Prediction modelling in the early detection of neonatal sepsis. World J Pediatr 2022 Mar 05;18(3):160-175 [FREE Full text] [doi: 10.1007/s12519-021-00505-1] [Medline: 34984642]
60. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
61. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, et al. Application scenarios for artificial intelligence in nursing care: rapid review. J Med Internet Res 2021 Nov 29;23(11):e26522 [FREE Full text] [doi: 10.2196/26522] [Medline: 34847057]
62. Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, et al. Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR Med Inform 2021 Jan 11;9(1):e23811 [FREE Full text] [doi: 10.2196/23811] [Medline: 33326405]
63. Talpur S, Azim F, Rashid M, Syed SA, Talpur BA, Khan SJ. Uses of different machine learning algorithms for diagnosis of dental caries. J Healthc Eng 2022 Mar 31;2022:5032435-5032413 [FREE Full text] [doi: 10.1155/2022/5032435] [Medline: 35399834]
64. Vélez-Guerrero MA, Callejas-Cuervo M, Mazzoleni S. Artificial intelligence-based wearable robotic exoskeletons for upper limb rehabilitation: a review. Sensors (Basel) 2021 Mar 18;21(6):2146 [FREE Full text] [doi: 10.3390/s21062146] [Medline: 33803911]

XSL•FO
RenderX

65. Welch V, Wy TJ, Ligezka A, Hassett LC, Croarkin PE, Athreya AP, et al. Use of mobile and wearable artificial intelligence in child and adolescent psychiatry: scoping review. J Med Internet Res 2022 Mar 14;24(3):e33560 [FREE Full text] [doi: 10.2196/33560] [Medline: 35285812]
66. Zhao IY, Ma YX, Yu MW, Liu J, Dong WN, Pang Q, et al. Ethics, integrity, and retributions of digital detection surveillance systems for infectious diseases: systematic literature review. J Med Internet Res 2021 Oct 20;23(10):e32328 [FREE Full text] [doi: 10.2196/32328] [Medline: 34543228]
67. Zheng Y, Dickson VV, Blecker S, Ng JM, Rice BC, Melkus GD, et al. Identifying patients with hypoglycemia using natural language processing: systematic literature review. JMIR Diabetes 2022 May 16;7(2):e34681 [FREE Full text] [doi: 10.2196/34681] [Medline: 35576579]
68. Zidaru T, Morrow EM, Stockley R. Ensuring patient and public involvement in the transition to AI-assisted mental health care: a systematic scoping review and agenda for design justice. Health Expect 2021 Aug 12;24(4):1072-1124 [FREE Full text] [doi: 10.1111/hex.13299] [Medline: 34118185]
69. Khan ZF, Alotaibi SR. Applications of artificial intelligence and big data analytics in m-Health: a healthcare system perspective. J Healthc Eng 2020;2020:8894694 [FREE Full text] [doi: 10.1155/2020/8894694] [Medline: 32952992]
70. Sarker S, Jamal L, Ahmed SF, Irtisam N. Robotics and artificial intelligence in healthcare during COVID-19 pandemic: a systematic review. Rob Auton Syst 2021 Dec;146:103902 [FREE Full text] [doi: 10.1016/j.robot.2021.103902] [Medline: 34629751]
71. Herrmann H. The arcanum of artificial intelligence in enterprise applications: toward a unified framework. J Eng Technol Manag 2022 Oct;66:101716. [doi: 10.1016/j.jengtecman.2022.101716]
72. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021 Apr 05;27(4):582-584. [doi: 10.1038/s41591-021-01312-x] [Medline: 33820998]
73. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. New York, NY: Academic Press; 2020:25-57.
74. Jacobson FL, Krupinski EA. Clinical validation is the key to adopting AI in clinical practice. Radiol Artif Intell 2021 Jul 01;3(4):e210104 [FREE Full text] [doi: 10.1148/ryai.2021210104] [Medline: 34350416]
75. Tran K, Bøtker JP, Aframian A, Memarzadeh K. Artificial intelligence for medical imaging. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. New York, NY: Academic Press; 2020:143-161.
76. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 2023 Aug 24;16(1):45-74 [FREE Full text] [doi: 10.1007/s12559-023-10179-8]
77. Chrimes D. Using decision trees as an expert system for clinical decision support for COVID-19. Interact J Med Res 2023 Jan 30;12:e42540 [FREE Full text] [doi: 10.2196/42540] [Medline: 36645840]

Abbreviations

AI: artificial intelligence
CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
DL: deep learning
EQUATOR: Enhancing the Quality and Transparency of Health Research
EUC: European Union Commission
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
ISO: International Organization for Standardization
JBI: Joanna Briggs Institute
LLM: large language model
ML: machine learning
NAM: National Academy of Medicine
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence

XSL•FO
RenderX

Edited by S Ma, T Leung; submitted 19.11.23; peer-reviewed by K Seibert, K Washington, X Yan; comments to author 21.03.24; revised version received 03.04.24; accepted 04.04.24; published 22.05.24.

Please cite as:
Kuziemsky CE, Chrimes D, Minshall S, Mannerow M, Lau F
AI Quality Standards in Health Care: Rapid Umbrella Review
J Med Internet Res 2024;26:e54705
URL: https://www.jmir.org/2024/1/e54705
doi:10.2196/54705
PMID:38776538

©Craig E Kuziemsky, Dillon Chrimes, Simon Minshall, Michael Mannerow, Francis Lau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effect of Artificial Intelligence on Patient-Physician Trust: Cross-Sectional Vignette Study

Anna G M Zondag[1], MSc; Raoul Rozestraten[2], MSc; Stephan G Grimmelikhuijsen[2], PhD; Karin R Jongsma[3], PhD; Wouter W van Solinge[1], PhD; Michiel L Bots[3], MD, PhD; Robin W M Vernooij[3,4], PhD; Saskia Haitjema[1], MD, PhD

[1]Central Diagnostic Laboratory, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[2]Utrecht University School of Governance, Utrecht University, Utrecht, Netherlands

[3]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[4]Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht, Netherlands

Corresponding Author:
Anna G M Zondag, MSc
Central Diagnostic Laboratory
University Medical Center Utrecht
Utrecht University
Heidelberglaan 100
Utrecht, 3584 CX
Netherlands
Phone: 31 631117922
Email: a.g.m.zondag@umcutrecht.nl

Abstract

Background: Clinical decision support systems (CDSSs) based on routine care data, using artificial intelligence (AI), are increasingly being developed. Previous studies focused largely on the technical aspects of using AI, but the acceptability of these technologies by patients remains unclear.

Objective: We aimed to investigate whether patient-physician trust is affected when medical decision-making is supported by a CDSS.

Methods: We conducted a vignette study among the patient panel (N=860) of the University Medical Center Utrecht, the Netherlands. Patients were randomly assigned into 4 groups—either the intervention or control groups of the high-risk or low-risk cases. In both the high-risk and low-risk case groups, a physician made a treatment decision with (intervention groups) or without (control groups) the support of a CDSS. Using a questionnaire with a 7-point Likert scale, with 1 indicating "strongly disagree" and 7 indicating "strongly agree," we collected data on patient-physician trust in 3 dimensions: competence, integrity, and benevolence. We assessed differences in patient-physician trust between the control and intervention groups per case using Mann-Whitney *U* tests and potential effect modification by the participant's sex, age, education level, general trust in health care, and general trust in technology using multivariate analyses of (co)variance.

Results: In total, 398 patients participated. In the high-risk case, median perceived competence and integrity were lower in the intervention group compared to the control group but not statistically significant (5.8 vs 5.6; *P*=.16 and 6.3 vs 6.0; *P*=.06, respectively). However, the effect of a CDSS application on the perceived competence of the physician depended on the participant's sex (*P*=.03). Although no between-group differences were found in men, in women, the perception of the physician's competence and integrity was significantly lower in the intervention compared to the control group (*P*=.009 and *P*=.01, respectively). In the low-risk case, no differences in trust between the groups were found. However, increased trust in technology positively influenced the perceived benevolence and integrity in the low-risk case (*P*=.009 and *P*=.04, respectively).

Conclusions: We found that, in general, patient-physician trust was high. However, our findings indicate a potentially negative effect of AI applications on the patient-physician relationship, especially among women and in high-risk situations. Trust in technology, in general, might increase the likelihood of embracing the use of CDSSs by treating professionals.

(J Med Internet Res 2024;26:e50853) doi:10.2196/50853

KEYWORDS

patient-physician relationship; trust; clinical decision support; artificial intelligence; digital health; decision support system

XSL•FO
RenderX

Introduction

Background

It was John McCarthy who coined the term "artificial intelligence" (AI) at the Dartmouth conference in 1956 and defined it as "the science and engineering of making intelligent machines, especially intelligent computer programs" [1]. However, it was only in the 1990s, after the first so-called "AI winter," that interest in AI began to increase again [2]. Since then, AI applications have been on the rise, ranging from self-driving cars to AI-powered web search [3]. Development and implementation of AI in health care is similarly increasing. It is believed that AI has the potential to improve every facet of health care—screening, diagnosis, prognosis, and treatment [4]. Although the use of AI in routine clinical care is still in the early stages, it has already shown promise in specific medical fields, such as radiology, for the recognition of complex patterns in imaging data [3,5]. Hospital-wide strategic programs have been initiated to develop predictive AI algorithms based on routine clinical care data in various hospitals [6]. The goal of these projects is, among others, to integrate AI algorithms in clinical decision support systems (CDSSs). CDSSs are often classified as either knowledge-based or non–knowledge-based systems. Knowledge-based CDSSs provide an output by evaluating a certain rule, which is programmed based on evidence or practice. Non–knowledge-based CDSSs use AI techniques, such as machine learning, for decision support and prediction [7]. AI-based CDSSs are often developed when dealing with complex, high-dimensional, and large amounts of data (ie, big data), such as routine care data. By linking patient information to evidence-based knowledge, a CDSS can provide case-specific information, which may support physicians in developing more personalized judgments and recommendations [8]. Despite these efforts, however, to date, only a fraction of all developed AI-based algorithms have been implemented in clinical care [9].

The debate on AI-based CDSSs in health care has mainly focused on the technical aspects of the technology [10,11], including questions like "How well does the predictive algorithm perform in terms of, for example, recall and precision?" or "What is the importance of its features?" To date, less attention has been given to the acceptability of using AI-based CDSSs in terms of patients' or physicians' trust in CDSSs even though these are crucial for the implementation and acceptance of these systems [12,13]. Some concerns revolve around how the patient-physician relationship is directly affected by the integration of AI applications into clinical practice. This concern arises due to all the new possibilities that AI offers, such as decision support, patient dashboards, and eHealth [14-16]. Figure 1 illustrates how the relationship between a patient and physician could be influenced by AI [17,18]. Studies have shown that algorithms developed to diagnose or predict a disease often perform as well as, and sometimes even better than, a physician [19]. As a result, physicians may be able to make more informed decisions and subsequently improve patient care, which was the incentive to start the Applied Data Analytics in Medicine (ADAM) project at the University Medical Center (UMC) Utrecht [6]. On the other hand, to build and maintain interpersonal trust, patient involvement in the decision-making process is important [20-23]. However, these algorithms, especially those that include AI, are sometimes perceived as "black boxes," which could potentially lead to a reduction in trust in the physician, even though these algorithms can be beneficial to the patient's care process. In medicine, trust is considered a central aspect of the patient-physician relationship. Without trust, treatments have proven to be less effective, and patients are more prone to ask for a second opinion [24].

XSL•FO
RenderX

Figure 1. Relationships between the patient, health care, and data (adapted from Groenhof [17], which is published under Creative Commons Attribution 4.0 International License [18]).

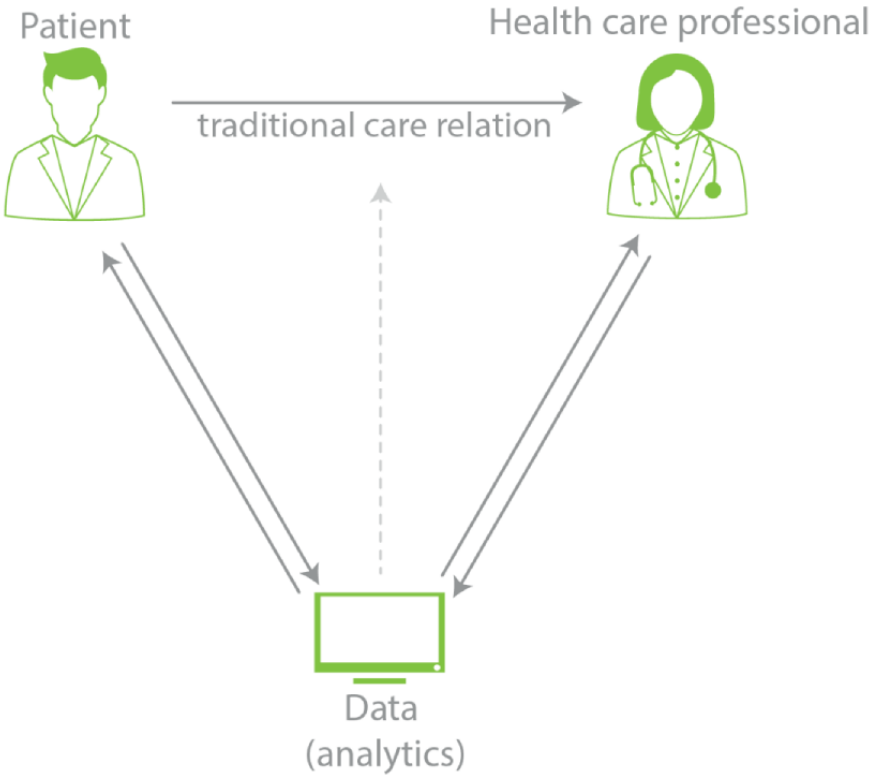

Trust

To the best of our knowledge, we are among the first to study the impact of AI-based CDSSs on patient-physician trust in both a clinical high-risk and low-risk setting. Trust is a complex and multidimensional concept [25]. Over the past decades, a lot of research has been conducted on this notion and various dimensions of trust have been described in the academic literature [26]. The most commonly named dimensions of trust in a physician are privacy and confidentiality [27,28], compassion [27,29], reliability and dependence [28], competence [27-31], communication [32], and honesty [33]. However, these dimensions of trust are often studied separately. With so many described trust dimensions, it is difficult to develop a framework specifically for patient-physician trust. Therefore, a framework that integrates these components of trust in a physician has not been established yet. In social sciences, trust is commonly studied in 3 dimensions: benevolence, competence, and integrity [34-37]. In a clinical context, benevolence is the extent to which a patient perceives a physician as caring about the patient's personal and health interests. Competence is the extent to which a patient perceives a physician as competent, capable, effective, and professional. Integrity concerns the extent to which a patient sees a physician as honest and truthful, handling the patient's sensitive information with care and confidentially.

We aim to study the extent to which these aspects of patient-physician trust are affected by AI-based CDSSs.

Methods

We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for observational (cross-sectional) studies.

Study Setting

We conducted a cross-sectional vignette study at the UMC Utrecht, the Netherlands. We reached out to all members of the UMC Utrecht patient panel (N=860) as part of the ADAM program (2017-2020; currently known as the Digital Health Department within the UMC Utrecht) [6]. In brief, the ADAM program consisted of 11 projects all aiming to answer clinical questions to personalize health care using routine care data and innovative methodology, such as AI. The overall goal of the ADAM program was to investigate the need to apply this health care innovation on a large scale at the UMC Utrecht to establish a learning health care system [6].

For our study, we simulated possible future clinical decision support applications of 2 of the ADAM projects. The Neonatal Intensive Care Unit (NICU) application concerned data of babies between 24 and 32 weeks old admitted in the NICU using an algorithm that was trained on a data set from the Wilhelmina Children's Hospital of the UMC Utrecht. This CDSS tool aimed to predict late-onset sepsis using the algorithm to prevent unnecessary antibiotic use and enable timely treatment [6]. The rheumatoid arthritis (RA) CDSS worked on the premise that disease flares can be predicted based on data about the course of the disease, patient characteristics, and information about current treatment. The aim of the RA application was to predict flares to support data-driven reduction of high preventative

XSL•FO
RenderX

doses of RA medication [6,38]. This algorithm was trained on data of patients with RA of the UMC Utrecht [38].

Study Population

Members of the patient panel (N=860) were all current and former patients of the UMC Utrecht willing to participate in research from the UMC Utrecht. The majority (693/860, 80.6%) of the panel was between 45 and 80 years of age, 57.3% (493/860) had higher education levels, and 52.8% (454/860) of the patient panel were women.

Questionnaire

In our study, we focused on interpersonal trust. We used the definition from Mayer et al [34] to define interpersonal trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party."

We studied 3 dimensions of trust: competence, integrity, and benevolence. We used an adapted version of the "Trust in Physician Scale" to assess patient-physician trust. The "Trust in Physician Scale" is a questionnaire consisting of 11 items and is commonly used to measure patient-physician trust [28,39]. Table 1 illustrates how the 11 questions of the Trust in Physician Scale were adapted and divided into the 3 trust dimensions.

In addition to the questions of the Trust in Physician Scale, we added several self-constructed questions to study these aspects of trust more comprehensively. We added the following 3 statements in the integrity dimension:

- I have the feeling that this physician is not holding anything back from me (scale 1-7).
- I have the feeling that this physician is being honest with me (scale 1-7).
- I trust this physician is handling my medical data with care (scale 1-7).

One additional statement was added in the benevolence dimension, as follows: "This physician's recommendation is in my personal best interest (scale 1-7)."

Besides trust in the physician, we added 2 additional questions about trust in health care and technology in general on a 7-point Likert scale (1: "no trust at all"; 7: "fully trust"), as follows:

- Please indicate your level of trust in health care in general (scale 1-7).
- Please indicate your level of trust in technology in general (scale 1-7).

Additionally, we asked the panel members their age, sex, and education level.

Subsequently, we performed a factor analysis on the questionnaire, including the Trust in Physician Scale and the self-constructed questions, to check whether the predefined dimensions were present in the questionnaire. The factor analysis revealed the correlations between the items of the questionnaire and subdivided them into factors. Thereafter, we calculated Cronbach α to measure the degree of coherence between the 3 trust dimensions [40]. A Cronbach α of more than 0.75 was considered high, and thus, acceptable, meaning that all items of the dimension measured the same concept [41].

Table 1. Trust in Physician Scale questions divided by dimensions of trust, rated on a 7-point Likert scale, with 1 indicating "strongly disagree" and 7 indicating "strongly agree," as answer possibilities.

Dimension	Statement
Benevolence	• This physician really cares about me as a person. • This physician's recommendation takes into account my wishes and puts them first. • I trust this physician's recommendation will put my medical needs above other interests. • I feel that with this recommendation the physician does everything to help me.
Competence	• This physician is a real expert in taking care of medical problems like mine. • I trust this physician's judgment about this medical problem. • I trust this physician's recommendation and therefore have no need for a second opinion. • I trust my physician's recommendation so much that I will follow it.
Integrity	• I trust this physician will tell me if he/she makes a mistake in this recommendation. • If this physician tells me this recommendation, then I also believe that this recommendation is correct. • I feel that this physician keeps the medical information used for this recommendation private.

Design of Data Collection

Data were collected in April 2019. The questionnaire included 1 of the 2 hypothetical cases in the form of a so-called "vignette." These vignettes were developed together with physicians from the neonatology and rheumatology department as well as the ADAM program staff to make them as realistic as possible (eg, to assess whether the physician's communication style in the vignette reflected clinical practice) and to ensure that the AI-based CDSS applications of both projects were accurately described. The vignettes were tested on master's degree students of Public Management from Utrecht University and employees of the ADAM program (N=36) first to study whether they could empathize well with the situation presented to them. This test resulted in a mean score of 4.6 for the NICU and 5.7 for the RA vignette on a 7-point Likert scale (the higher the score, the better they could empathize with the situation presented). We processed the feedback we received from the test participants to further improve the vignettes. After successful testing, we randomly divided the panel members into 4 groups (Multimedia Appendix 1). The groups were presented with either a life-threatening (high-risk) vignette in the NICU

XSL•FO
RenderX

(from now on referred to as the "high-risk case"), or a non–life-threatening (low-risk) vignette regarding RA (referred to as the "low-risk case"). Details of the vignettes are described in Multimedia Appendix 2. The high-risk case described a baby in the NICU that possibly had sepsis. In this case, we asked the panel members to empathize with the role of the baby's parents. The low-risk case described a patient with RA, and panel members were asked to empathize with the role of the patient. In both situations, the physician made a treatment recommendation. This recommendation was made by a physician only in the control group and by a physician supported by a CDSS application in the intervention group. After reading the vignette, the panel members were asked to fill in the questionnaire. Possible answers were given using a 7-point Likert scale, with 1 indicating "strongly disagree" and 7 indicating "strongly agree."

We tracked the time it took panel members to read the vignettes. To minimize the potential loss of statistical power due to panel members not being diligent and motivated to complete the survey, we excluded the 2.5% slowest and fastest readers from all analyses. Figure 2 illustrates the patient inclusion and exclusion procedure in a flow diagram.

Figure 2. Inclusion procedure of the University Medical Center Utrecht panel members.

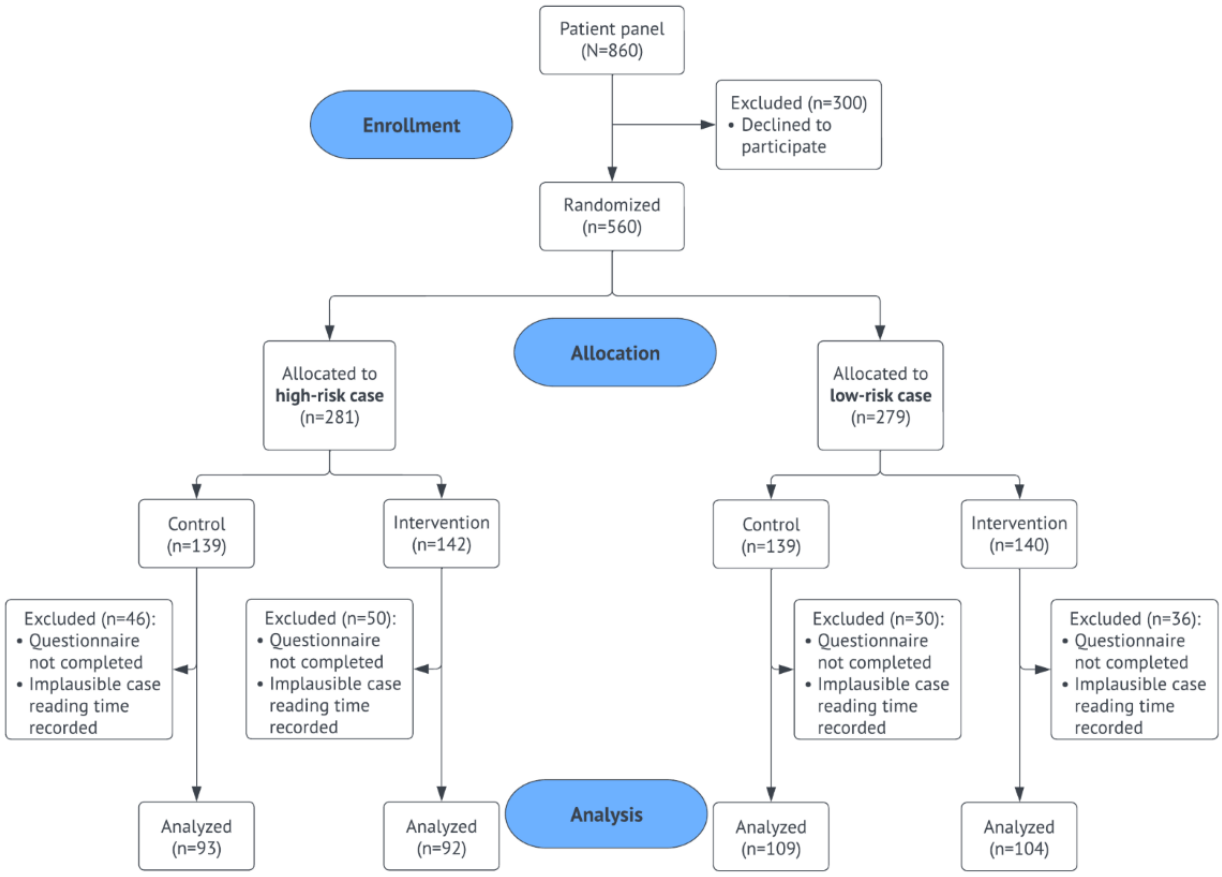

Data Analyses

We presented the results of all groups in medians (IQRs), stratified by case. We compared perceived trust between the intervention and control group using the Mann-Whitney *U* test (Wilcoxon rank sum test).

We assessed the presence or absence of effect modification using a two-way multivariate analysis of variance (MANOVA) or multivariate analysis of covariance (MANCOVA), as appropriate. The participant's sex, age, education level, general trust in health care, and general trust in technology were considered potential effect modifiers in the relation between the use of a CDSS and patient-physician trust. We further stratified the results by sex in case of effect modification to examine this potential effect.

All analyses were performed using SPSS edition 26 (IBM SPSS Inc) [42]. Differences with a 2-sided *P* value of <.05 were considered statistically significant.

Ethical Considerations

All participants provided informed consent digitally for this study. The panel members received an introductory text with information about this study. At the end of the introductory text, the privacy statement informed the panel members that by clicking the "Continue" button, they agreed to the use of their data in this study. Data were anonymized and IP addresses were not stored.

The questionnaires were administered in Qualtrics in accordance with Utrecht University guidelines and stored on the Yoda server, which is a research data management service from the Utrecht University, among others, enabling secure storage of the research data [43,44]. The methods have been performed

XSL•FO
RenderX

based on relevant guidelines and regulations. The institutional review board of UMC Utrecht waived ethical approval because it ruled that this study did not concern a medical research question.

Results

Demographics of the Study Population

In total, 398/860 (46.3%) panel members of the UMC Utrecht patient panel participated in this study. Characteristics of the study population can be seen in Table 2. Participants were older in the intervention group of both cases. In the high-risk case, the intervention group consisted of more men (52/92, 56.5%), whereas the control group consisted of more women (54/93, 58.1%). In the low-risk case, both the control and intervention groups consisted of more women (55/109, 50.5% and 56/104, 53.8%, respectively). The education level was similar across the groups. In all groups, the majority of participants (62/93, 66.7%; 60/92, 65.9%; 69/109, 63.3%; and 68/104, 66.0%) had a high education level, that is, a degree from a university of applied sciences or higher.

Table 2. Demographics of the study population by case.

Dimension	High-risk case		Low-risk case	
	Control (n=93)	Intervention (n=92)	Control (n=109)	Intervention (n=104)
Age, mean (SD)	57.9 (11.7)	58.9 (12.1)	60.1 (12.0)	61.6 (10.0)
Sex, n (%)				
Men	39 (41.9)	52 (56.5)	54 (49.5)	48 (46.2)
Women	54 (58.1)	40 (43.5)	55 (50.5)	56 (53.8)
Education level, n (%)				
Lower	31 (33.3)	31 (34.1)	40 (36.7)	35 (34)
Higher	62 (66.7)	60 (65.9)	69 (63.3)	68 (66)

Measured Trust in the Physician

Based on the results of the factor analysis and Cronbach α calculation, we assessed trust on both a multidimensional and a unidimensional scale. The reason to use a unidimensional scale was that the factor analysis initially showed 2 factors (Table S1 in Multimedia Appendix 3); however, after removing the 2 items that loaded strongly on both factors and 1 item that primarily loaded on the second factor, only 1 factor remained. The reason for using a multidimensional scale was that the Cronbach α of the 3 dimensions was considered high enough (>0.75) to measure trust on a multidimensional scale (Table S2 in Multimedia Appendix 3).

Overall, trust in the physician was high, with a median of 5.8 (control group IQR 5.0-6.5; intervention group IQR 4.7-6.2) in the high-risk case and 6.0 (control group IQR 5.3-6.5; intervention group IQR 5.1-6.5) in the low-risk case on a 7-point Likert scale (Multimedia Appendix 4). The high-risk case showed a lower median for the integrity of the physician using a CDSS (6.3, IQR 5.3-6.8 vs 6.0, IQR 5.0-6.7; U=3590.0; P=.06; Table 3) compared to the physician who did not use a CDSS, but these results were not statistically significant. Similarly, perceived competence was lower in the intervention group compared to the control group (median 5.8, IQR 4.8-6.5 vs 5.6, IQR 4.3-6.3; U=3771.5; P=.16). We observed no between-group differences in perceived competence, integrity, and benevolence of the low-risk group.

In the analyses exploring whether the results were different among subgroups, in the high-risk case, the effect of a CDSS application on the perceived competence of the physician depended on the participant's sex ($F_{1,181}$=4.694; P=.03; Multimedia Appendix 5). In women, perceived competence and integrity were significantly lower in physicians who used a CDSS compared to physicians who did not (U=740.5; P=.009 and U=756.0; P=.01, respectively; Table 4), whereas no such statistically significant differences were found in men.

In the low-risk case, results showed that the effect of the CDSS application on the perceived benevolence and integrity depended on the participant's trust in technology in general ($F_{1,209}$=6.943; P=.009 and $F_{1,209}$=4.119; P=.04, respectively; Multimedia Appendix 5). An increase in the participant's trust in technology, in general, led to an increased perceived integrity and benevolence of physicians using a CDSS. This increase was more significant in the intervention group compared to the control group.

XSL•FO
RenderX

Table 3. Measured levels of trust of the 3 dimensions between groups, by case.

Dimension	High-risk case				Low-risk case			
	Control (n=93)	Intervention (n=92)	*U* test[a]	*P* value[b]	Control (n=109)	Intervention (n=104)	*U* test	*P* value
Competence, median (IQR)	5.8 (4.8-6.5)	5.6 (4.3-6.3)	3771.5	.16	6.0 (5.3-6.5)	6.0 (5.0-6.5)	5319.0	.43
Integrity, median (IQR)	6.3 (5.3-6.8)	6.0 (5.0-6.7)	3590.0	.06	6.0 (5.5-6.8)	6.0 (5.3-6.8)	5698.0	.95
Benevolence, median (IQR)	5.8 (5.0-6.6)	5.8 (4.9-6.2)	4055.5	.54	6.0 (5.4-6.6)	6.0 (5.3-6.6)	5829.0	.72

[a]*U* test: Wilcoxon test statistic.

[b]The *P* values in the table indicate the difference between the control and intervention groups.

Table 4. Measured levels of trust of the 3 dimensions in the high-risk case, stratified by sex.

Dimension	Men				Women			
	Control (n=39)	Intervention (n=52)	*U* test[a]	*P* value[b]	Control (n=54)	Intervention (n=40)	*U* test	*P* value
Competence, median (IQR)	5.0 (4.5-6.3)	5.9 (4.3-6.4)	1111.0	.43	6.0 (5.2-6.8)	5.4 (4.0-6.0)	740.5	.009
Integrity, median (IQR)	6.3 (5.3-6.8)	6.1 (5.0-6.8)	1001.0	.92	6.5 (5.3-7.0)	5.8 (5.0-6.3)	756.0	.01
Benevolence, median (IQR)	5.8 (5.0-6.4)	6.0 (5.1-6.4)	1064.5	.68	5.7 (5.0-6.8)	5.5 (4.8-6.0)	918.5	.21

[a]*U*: Wilcoxon test statistic.

[b]*P* values indicate the difference between the control and intervention groups.

Discussion

Principal Findings

We aimed to assess the extent to which patient-physician trust was affected by using a CDSS application. We found that, in general, trust in physicians was high. Nonetheless, in the high-risk case, we observed that trust, in terms of competence and integrity, was lower in physicians using a CDSS compared to physicians who did not, and that the differences were larger in women than in men. No differences were found in the low-risk case. However, we found that perceived benevolence and integrity in the low-risk case depended on the participant's trust in technology in general.

Comparison With Prior Literature

To our knowledge, this study is among the first to examine the impact of AI-based CDSSs on patient-physician trust in both a clinical high-risk and low-risk setting. There are several possible explanations for our results. First, prior to our study, we asked several physicians of the UMC Utrecht about their expectations regarding CDSS applications. Some anticipated that the use of CDSS applications in clinical practice could evoke a critical response from patients, believing that CDSS applications "first must prove their value before being accepted by patients." Regarding competence, physicians mentioned that the use of a CDSS application could raise doubts about their professionalism, as it could come across as "being dependent on such an application." This dependence and the associated loss of competence, known as "deskilling," is a known concern of the introduction of clinical decision support applications in health care [45]. Some of these concerns seem reasonable, as our study indicates that patients' trust in their physicians can decrease when a CDSS is used. Moreover, the decrease found in trust in terms of the integrity of a physician who uses a CDSS could be explained by concerns about data protection. Patients may feel they have no control over what happens with their personal data. However, some studies indicated that privacy was not an issue for some patients, as long as they trusted their physician [46,47]. Yakar et al [48] studied the general population's view on the use of AI in health care through a web-based survey in the Netherlands and found less trust in AI than initially hypothesized. They, and some others, also found that women were less trusting of AI than men [48,49]. A reason for this could be that the women participating in this study were aware of the potential presence of gender bias in AI, as this issue has often been raised in previous research and media [50-52]. Additionally, women generally are considered to have more risk aversion than men [53]. Both these reasons combined could explain why the effect of using a CDSS on the perceived trust in the physician in the high-risk case depended on the participant's sex and why this was not the case in the low-risk case.

In the low-risk case, our results showed that a patient's general trust in technology played a significant role in the effect that CDSS applications could have on the perceived integrity and benevolence of the physician. An increase in the perceived integrity and benevolence of the physician seemed to be more likely in patients who had more trust in technology in general. This is in line with previous research investigating the association between trust in technology in general and trust in the implementation of AI in medicine [48,54].

Implications

Results from this study underline that mainly communicating about the technological aspects of AI algorithms in health care may ultimately hamper the successful implementation and acceptance of these algorithms in clinical practice. Personalized health care starts with putting the patient's values first.

XSL•FO
RenderX

Therefore, we believe that the development of AI-based CDSSs in health care should be transparent in terms of methods, alternatives, and understandability. In addition, patients should be involved early in the developmental phase of the CDSS, and the use of AI should be explained well during the visit with the health care professional. Furthermore, AI applications should not only be developed to predict the prognosis of a certain disease (precision medicine) but also to improve and facilitate discussions between patients and physicians about the patient's illness and personal care needs (shared decision-making). The discussion between the patient and the physician may be improved by being transparent about the CDSS application. This potentially leads to more trust in the application, which, as shown in this study, could lead to an increase in the perceived integrity and benevolence of the physician. Moreover, it seems sensible to distinguish between the types of situations or patient populations in the development of AI algorithms for CDSS in health care; the results from our study indicate that less favorable reactions occur when AI is used in more life-threatening situations, which could also be due to the different patient populations, for example, babies in an acute situation versus adults in a chronic situation. By implementing AI-based CDSS in less risky situations first, patients could get acquainted with such CDSSs before implementing them in high-risk situations. Further study is warranted to establish the best implementation strategy regarding AI-based CDSSs in clinical care.

We recommend replicating this study while providing a clear explanation of the AI application. For example, by explaining how it works, how patient data are handled, and the benefits of such AI-based CDSSs for the patients. In addition, further research with a larger sample size and in different preventive care and cure settings, and across several age groups, should be conducted to confirm the results found in this study. External validation of our results will inform us whether our results are similar in, for example, patients in other health care settings or areas. Additionally, qualitative research methods should be considered to gain more insights into the reasons patient-physician trust might decline when using a CDSS from a patient perspective.

Strengths and Limitations

Vignette studies have been criticized in the past for their limitations. First, a frequently heard criticism of vignette studies is the lack of reality due to their hypothetical nature [55]. However, we minimized this by creating both hypothetical cases in collaboration with physicians working in the involved departments—the neonatology and rheumatology departments—and tested whether the participants could empathize well with the situation during a pilot study. Second, the participants in this study were (or still are) patients of the UMC Utrecht. This made it easier for the participants to empathize with the vignettes because they may have been in similar situations themselves, and they indicated they could (median 6.0, IQR 5.0-7.0 in all 4 groups). Third, it has been shown that participants of vignette studies are not always diligent and motivated to complete surveys or take these kinds of experiments seriously, which may reduce statistical power [56]. Participants could, for example, have put the questionnaire aside for a while, or might not have read the described situations properly [57]. We reduced this effect by tracking the time the participants took to read the vignettes and by excluding the fastest and the slowest 2.5% of readers from the analyses. Therefore, participants spending a remarkably long or short time reading the vignette were not included.

Conclusions

Trust in physicians is generally high among patients. However, our findings point toward a negative effect that AI applications can have on the patient-physician relationship, especially among women. We, therefore, believe that, for a successful adoption of AI applications in clinical practice, patients should be involved in both the development and implementation of such applications. Moreover, a broader societal discussion needs to take place about humane values and AI to gain insights into how we want AI to influence our lives when we encounter health care. In addition, clear communication to patients and society about the functions of the AI applications and what personal data they use seems equally important.

Acknowledgments

AGMZ was supported by the European Union's Horizon 2020 research and innovation program under grant agreement number 101017331 (ODIN).

The funding source was not involved in the study design, analysis, and interpretation of the data nor the writing and in the decision to submit the manuscript for publication.

Authors' Contributions

AGMZ, RR, SGG, WWvS, MLB, RWMV, and SH contributed to the conceptualization and methodology of the project. AGMZ and RR analyzed the data. AGMZ wrote the draft manuscript. All authors contributed to the interpretation of the results and critically reviewed the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Study design.
[DOC File , 28 KB - jmir_v26i1e50853_app1.doc]

Multimedia Appendix 2
Details of the high-risk and low-risk vignettes.
[DOC File , 33 KB - jmir_v26i1e50853_app2.doc]

Multimedia Appendix 3
Questionnaire factor analysis and Cronbach α.
[DOC File , 47 KB - jmir_v26i1e50853_app3.doc]

Multimedia Appendix 4
Unidimensional trust scores, stratified by case.
[DOC File , 28 KB - jmir_v26i1e50853_app4.doc]

Multimedia Appendix 5
Multivariate regression analyses to test potential effect modification.
[DOC File , 56 KB - jmir_v26i1e50853_app5.doc]

References

1. What is AI? Basic questions. Stanford University. URL: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html [accessed 2022-01-06]
2. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Dec;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. Tran AQ, Nguyen LH, Nguyen HSA, Nguyen CT, Vu LG, Zhang M, et al. Determinants of intention to use artificial intelligence-based diagnosis support system among prospective physicians. Front Public Health 2021;9:755644 [FREE Full text] [doi: 10.3389/fpubh.2021.755644] [Medline: 34900904]
5. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 2018 Mar;15(3 Pt B):504-508. [doi: 10.1016/j.jacr.2017.12.026] [Medline: 29402533]
6. Haitjema S, Prescott TR, van Solinge WW. The applied data analytics in medicine program: lessons learned from four years' experience with personalizing health care in an academic teaching hospital. JMIR Form Res 2022 Jan 28;6(1):e29333. [doi: 10.2196/29333] [Medline: 35089145]
7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
8. Melton GB, McDonald CJ, Tang PC, Hripcsak G. Electronic health records. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics. 5th Edition. Cham, Switzerland: Springer Nature; 2021:467-509.
9. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, Ghadiri P, Grad R, Poenaru D, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022 Aug 09;10(8):e36199 [FREE Full text] [doi: 10.2196/36199] [Medline: 35943793]
10. Stai B, Heller N, McSweeney S, Rickman J, Blake P, Vasdev R, et al. Public perceptions of artificial intelligence and robotics in medicine. J Endourol 2020 Oct 01;34(10):1041-1048 [FREE Full text] [doi: 10.1089/end.2020.0137] [Medline: 32611217]
11. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
12. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
13. Gerke S, Minssen T, Cohen IG. Ethical and legal challenges of artificial intelligence-driven health care. SSRN Journal 2020:295-336. [doi: 10.2139/ssrn.3570129]
14. Lucivero F, Jongsma KR. A mobile revolution for healthcare? Setting the agenda for bioethics. J Med Ethics 2018 Oct;44(10):685-689 [FREE Full text] [doi: 10.1136/medethics-2017-104741] [Medline: 29907579]
15. Schermer M. Telecare and self-management: opportunity to change the paradigm? J Med Ethics 2009 Nov;35(11):688-691. [doi: 10.1136/jme.2009.030973] [Medline: 19880706]
16. Fiske A, Buyx A, Prainsack B. The double-edged sword of digital self-care: physician perspectives from Northern Germany. Soc Sci Med 2020 Sep;260:113174 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113174] [Medline: 32659512]

XSL•FO
RenderX

17. Groenhof TKJ. General discussion: how to make a learning healthcare system work? In: Towards a Learning Healthcare System. Enschede, The Netherlands: Gildeprint B.V; 2020:215-232.
18. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-05-16]
19. Tucker I. AI cancer detectors. The Guardian. 2018 Jun 10. URL: https://www.theguardian.com/technology/2018/jun/10/artificial-intelligence-cancer-detectors-the-five [accessed 2024-03-06]
20. Bardes CL. Defining "patient-centered medicine". N Engl J Med 2012 Mar 01;366(9):782-783. [doi: 10.1056/NEJMp1200070] [Medline: 22375968]
21. Weston WW. Informed and shared decision-making: the crux of patient-centered care. CMAJ 2001 Aug 21;165(4):438-439 [FREE Full text] [Medline: 11531054]
22. Reuben DB, Tinetti ME. Goal-oriented patient care--an alternative health outcomes paradigm. N Engl J Med 2012 Mar 01;366(9):777-779. [doi: 10.1056/NEJMp1113631] [Medline: 22375966]
23. Kitson A, Marshall A, Bassett K, Zeitz K. What are the core elements of patient-centred care? A narrative review and synthesis of the literature from health policy, medicine and nursing. J Adv Nurs 2013 Jan;69(1):4-15. [doi: 10.1111/j.1365-2648.2012.06064.x] [Medline: 22709336]
24. LaVeist TA, Isaac LA, Williams KP. Mistrust of health care organizations is associated with underutilization of health services. Health Serv Res 2009 Dec 13;44(6):2093-2105 [FREE Full text] [doi: 10.1111/j.1475-6773.2009.01017.x] [Medline: 19732170]
25. Jones GR, George JM. The experience and evolution of trust: implications for cooperation and teamwork. AMR 1998 Jul;23(3):531-546. [doi: 10.5465/amr.1998.926625]
26. Gill H, Boies K, Finegan JE, McNally J. Antecedents of trust: establishing a boundary condition for the relation between propensity to trust and intention to trust. J Bus Psychol 2005 Apr;19(3):287-302. [doi: 10.1007/s10869-004-2229-8]
27. Mechanic D, Schlesinger M. The impact of managed care on patients' trust in medical care and their physicians. JAMA 1996 Jun 05;275(21):1693-1697. [Medline: 8637148]
28. Anderson LA, Dedrick RF. Development of the Trust in Physician scale: a measure to assess interpersonal trust in patient-physician relationships. Psychol Rep 1990 Dec;67(3 Pt 2):1091-1100. [doi: 10.2466/pr0.1990.67.3f.1091] [Medline: 2084735]
29. Emanuel EJ, Dubler NN. Preserving the physician-patient relationship in the era of managed care. JAMA 1995 Jan 25;273(4):323-329. [Medline: 7815662]
30. Gray BH. Trust and trustworthy care in the managed care era. Health Aff (Millwood) 1997 Jan;16(1):34-49. [doi: 10.1377/hlthaff.16.1.34] [Medline: 9018941]
31. Caterinicchio RP. Testing plausible path models of interpersonal trust in patient-physician treatment relationships. Soc Sci Med. Part A: Med Psychol Med Sociol 1979 Jan;13:81-99. [doi: 10.1016/0271-7123(79)90011-7]
32. Mechanic D. The functions and limitations of trust in the provision of medical care. J Health Polit Policy Law 1998 Aug;23(4):661-686. [doi: 10.1215/03616878-23-4-661] [Medline: 9718518]
33. Thom DH, Campbell B. Patient-physician trust: an exploratory study. J Fam Pract 1997 Feb;44(2):169-176. [Medline: 9040520]
34. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. AMR 1995 Jul;20(3):709. [doi: 10.2307/258792]
35. Grimmelikhuijsen S, Knies E. Validating a scale for citizen trust in government organizations. Int Rev Adm Sci 2015 Sep 03;83(3):583-601. [doi: 10.1177/0020852315585950]
36. McEvily B, Tortoriello M. Measuring trust in organisational research: review and recommendations. J Trust Res 2011 Apr;1(1):23-63. [doi: 10.1080/21515581.2011.552424]
37. McKnight DH, Choudhury V, Kacmar C. Developing and validating trust measures for e-commerce: an integrative typology. ISR 2002 Sep;13(3):334-359. [doi: 10.1287/isre.13.3.334.81]
38. van der Leeuw MS, Messelink MA, Tekstra J, Medina O, van Laar JM, Haitjema S, et al. Using real-world data to dynamically predict flares during tapering of biological DMARDs in rheumatoid arthritis: development, validation, and potential impact of prediction-aided decisions. Arthritis Res Ther 2022 Mar 23;24(1):74 [FREE Full text] [doi: 10.1186/s13075-022-02751-8] [Medline: 35321739]
39. Pearson SD, Raeke LH. Patients' trust in physicians: many theories, few measures, and little data. J Gen Intern Med 2000 Jul;15(7):509-513 [FREE Full text] [doi: 10.1046/j.1525-1497.2000.11002.x] [Medline: 10940139]
40. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm 2008 Dec 01;65(23):2276-2284. [doi: 10.2146/ajhp070364] [Medline: 19020196]
41. Gottems LBD, Carvalho EMPD, Guilhem D, Pires MRGM. Good practices in normal childbirth: reliability analysis of an instrument by Cronbach's Alpha. Rev Lat Am Enfermagem 2018 May 17;26:e3000 [FREE Full text] [doi: 10.1590/1518-8345.2234.3000] [Medline: 29791667]
42. IBM SPSS Statistics for Windows. IBM. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-26 [accessed 2024-03-21]
43. Qualtrics - een enquêtetool. Utrecht University. URL: https://students.uu.nl/qualtrics-een-enquetetool [accessed 2022-05-02]

XSL•FO
RenderX

44. Yoda - a research data management service. Utrecht University. URL: https://www.uu.nl/en/research/yoda/about-yoda/why-yoda [accessed 2022-05-02]
45. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017 Aug 08;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
46. Belfrage S, Helgesson G, Lynøe N. Trust and digital privacy in healthcare: a cross-sectional descriptive study of trust and attitudes towards uses of electronic health data among the general public in Sweden. BMC Med Ethics 2022 Mar 04;23(1):19 [FREE Full text] [doi: 10.1186/s12910-022-00758-z] [Medline: 35246118]
47. Haselager DR, Boers SN, Jongsma KR, Vinkers CH, Broekman ML, Bredenoord AL. Breeding brains? Patients' and laymen's perspectives on cerebral organoids. Regen Med 2020 Dec;15(12):2351-2360. [doi: 10.2217/rme-2020-0108] [Medline: 33471559]
48. Yakar D, Ongena YP, Kwee TC, Haan M. Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. Value Health 2022 Mar;25(3):374-381 [FREE Full text] [doi: 10.1016/j.jval.2021.09.004] [Medline: 35227448]
49. Kaplan AD, Kessler TT, Brill JC, Hancock PA. Trust in artificial intelligence: meta-analytic findings. Hum Factors 2023 Mar 28;65(2):337-359 [FREE Full text] [doi: 10.1177/00187208211013988] [Medline: 34048287]
50. Paul K. Disastrous lack of diversity in AI industry perpetuates bias, study finds. The Guardian. URL: https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study [accessed 2022-09-08]
51. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020;3:81 [FREE Full text] [doi: 10.1038/s41746-020-0288-5] [Medline: 32529043]
52. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378. [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
53. van Diemen J, Verdonk P, Chieffo A, Regar E, Mauri F, Kunadian V, et al. The importance of achieving sex- and gender-based equity in clinical trials: a call to action. Eur Heart J 2021 Aug 17;42(31):2990-2994 [FREE Full text] [doi: 10.1093/eurheartj/ehab457] [Medline: 34352884]
54. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. Intell -Based Med 2020 Nov;1-2:100001. [doi: 10.1016/j.ibmed.2020.100001]
55. Krupnikov Y, Levine A. Expressing versus revealing preferences in experimental research. In: Sourcebook for Political Communication Research. 1st Edition. London, UK: Routledge; 2010:9780203938669.
56. Blom-Hansen J, Morton R, Serritzlew S. Experiments in public management research. IPMJ 2015 Mar 11;18(2):151-170. [doi: 10.1080/10967494.2015.1024904]
57. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: detecting satisficing to increase statistical power. J Exp Soc Psychol 2009 Jul;45(4):867-872. [doi: 10.1016/j.jesp.2009.03.009]

Abbreviations

ADAM: Applied Data Analytics in Medicine
AI: artificial intelligence
CDSS: clinical decision support system
MANCOVA: multivariate analysis of covariance
MANOVA: multivariate analysis of variance
NICU: Neonatal Intensive Care Unit
RA: rheumatoid arthritis
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UMC: University Medical Center

Edited by Y Zhuang; submitted 17.07.23; peer-reviewed by Z Su, E Vashishtha, PH Liao, P Velmovitsky; comments to author 26.01.24; revised version received 21.03.24; accepted 16.04.24; published 28.05.24.

Please cite as:
Zondag AGM, Rozestraten R, Grimmelikhuijsen SG, Jongsma KR, van Solinge WW, Bots ML, Vernooij RWM, Haitjema S
The Effect of Artificial Intelligence on Patient-Physician Trust: Cross-Sectional Vignette Study
J Med Internet Res 2024;26:e50853
URL: https://www.jmir.org/2024/1/e50853
doi:10.2196/50853
PMID:38805702

XSL•FO
RenderX

©Anna G M Zondag, Raoul Rozestraten, Stephan G Grimmelikhuijsen, Karin R Jongsma, Wouter W van Solinge, Michiel L Bots, Robin W M Vernooij, Saskia Haitjema. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of the Clinical Efficacy and Trust in AI-Assisted Embryo Ranking: Survey-Based Prospective Study

Hyung Min Kim[1], PhD; Hyoeun Kang[1], MS; Chaeyoon Lee[1], BS; Jong Hyuk Park[2], PhD; Mi Kyung Chung[3], PhD; Miran Kim[4], MD, PhD; Na Young Kim[5], MD; Hye Jun Lee[6], MD

[1]AI Lab, Kai Health, Seoul, Republic of Korea
[2]IVF Clinic, Miraewaheemang Hospital, Seoul, Republic of Korea
[3]IVF Clinic, Seoul Rachel Fertility Center, Seoul, Republic of Korea
[4]Department of Obstetrics & Gynecology, Ajou University School of Medicine, Suwon, Republic of Korea
[5]IVF Clinic, HI Fertility Center, Seoul, Republic of Korea
[6]Kai Health, Seoul, Republic of Korea

Corresponding Author:
Hye Jun Lee, MD
Kai Health
217 Teheran-ro #306, Yeoksam-dong, Gangnam-gu
Seoul, 06142
Republic of Korea
Phone: 82 1052299697
Email: hyejunlee@gmail.com

Abstract

Background: Current embryo assessment methods for in vitro fertilization depend on subjective morphological assessments. Recently, artificial intelligence (AI) has emerged as a promising tool for embryo assessment; however, its clinical efficacy and trustworthiness remain unproven. Simulation studies may provide additional evidence, provided that they are meticulously designed to mitigate bias and variance.

Objective: The primary objective of this study was to evaluate the benefits of an AI model for predicting clinical pregnancy through well-designed simulations. The secondary objective was to identify the characteristics of and potential bias in the subgroups of embryologists with varying degrees of experience.

Methods: This simulation study involved a questionnaire-based survey conducted on 61 embryologists with varying levels of experience from 12 in vitro fertilization clinics. The survey was conducted via Google Forms (Google Inc) in three phases: (1) phase 1, an initial assessment (December 23, 2022, to January 22, 2023); (2) phase 2, a validation assessment (March 6, 2023, to April 5, 2023); and (3) phase 3 an AI-guided assessment (March 6, 2023, to April 5, 2023). Inter- and intraobserver assessments and the accuracy of embryo selection from 360 day-5 embryos before and after AI guidance were analyzed for all embryologists and subgroups of senior and junior embryologists.

Results: With AI guidance, the interobserver agreement increased from 0.355 to 0.527 and from 0.440 to 0.524 for junior and senior embryologists, respectively, thus reaching similar levels of agreement. In a test of accurate embryo selection with 90 questions, the numbers of correct responses by the embryologists only, embryologists with AI guidance, and AI only were 34 (38%), 45 (50%), and 59 (66%), respectively. Without AI, the average score (accuracy) of the junior group was 33.516 (37%), while that of the senior group was 35.967 (40%), with $P<.001$ in the t test. With AI guidance, the average score (accuracy) of the junior group increased to 46.581 (52%), reaching a level similar to that of the senior embryologists of 44.833 (50%), with $P=.34$. Junior embryologists had a higher level of trust in the AI score.

Conclusions: This study demonstrates the potential benefits of AI in selecting embryos with high chances of pregnancy, particularly for embryologists with 5 years or less of experience, possibly due to their trust in AI. Thus, using AI as an auxiliary tool in clinical practice has the potential to improve embryo assessment and increase the probability of a successful pregnancy.

(J Med Internet Res 2024;26:e52637) doi:10.2196/52637

XSL•FO
RenderX

KEYWORDS

assisted reproductive technology; in vitro fertilization; artificial intelligence; intraobserver and interobserver agreements; embryos; embryologists

Introduction

Infertility affects 1 in 6 couples worldwide, making in vitro fertilization (IVF) a widely sought-after solution. However, the success rate of IVF remains relatively low, typically ranging from 20% to 30% [1,2]. Amidst the ongoing efforts to improve IVF outcomes, the paramount challenge lies in selecting the most viable embryo for transfer, as embryo quality is critical for a successful outcome.

Traditionally, embryologists rely on morphological assessment for selecting embryos [3-5], which involves the observation of embryos under a microscope and assigning grades based on criteria, such as blastocyst expansion stage, inner cell mass, and trophectoderm development. Some laboratories select euploid embryos through preimplantation genetic testing [6]. Even after the preimplantation genetic testing, morphological assessment remains crucial to selecting the most viable one from multiple euploid embryos.

However, this reliance on morphological assessment raises concerns because of its inherent subjectivity and the substantial variability observed in both intra- and interobserver assessments [7-11]. This variability underscores the pressing need for standardized methods of embryo evaluation across laboratories and the IVF industry.

Recent advancements have introduced artificial intelligence (AI) as a complementary tool for morphological evaluation of embryos. By leveraging deep learning techniques, AI-based systems can predict IVF outcomes by learning from extensive sets of embryo images, thus reducing human bias and potentially providing more objective and accurate results [12-17]. Several studies have reported significant advantages of AI-selected embryos in terms of pregnancy rates, compared to embryos chosen through traditional morphological assessment by embryologists [12,15,18].

Despite the notable progress in research on embryo selection using AI, its widespread adoption hinges on proving its clinical efficacy and securing the trust of clinicians. However, demonstrating the clinical efficacy of AI in a real-world setting poses challenges, as pregnancy outcomes can only be observed for selected embryos and not for those left unselected. Furthermore, the ultimate decision regarding which embryos to transfer rests squarely with clinicians, who may either embrace or question AI recommendations. In cases where clinicians opt not to trust AI, the resulting IVF outcomes may not accurately reflect the precision of the AI model's predictions.

Previous attempts to address these critical questions have involved simulation studies that compared the accuracy of and pregnancy rates facilitated by AI-driven embryo selection and assessments by embryologists [19,20]. However, these studies predominantly relied on historical data from embryo grading records provided by embryologists from various laboratories. This retrospective approach inadvertently introduced substantial sources of variability. Intra- and interobserver agreements among embryologists, which arise from the differing standards and criteria for embryo evaluation, played a significant role in shaping the results. Moreover, the lack of standardization of embryo evaluation across laboratories further complicates the interpretation of the findings.

In addition to these challenges, previous simulation studies have not examined the nuanced demographic profiles and attitudes of embryologists themselves. Understanding these factors can shed light on the broader dynamics at play and offer insights into how AI may be influenced by the unique characteristics and perspectives of the medical professionals involved.

In this study, we aimed to evaluate the clinical efficacy of AI in embryo selection by simulating a clinical setting in which embryologists ranked the embryos for transfer. We assessed the intra- and interobserver agreements of the embryologists' evaluations to emphasize the need for alternative embryo selection methods. Moreover, we compared the accuracy of embryologists with and without AI assistance, as well as AI-only selection to substantiate the clinical efficacy of AI. Additionally, we analyzed and compared the results for subgroups of embryologists with varying levels of experience to identify distinct practice patterns and levels of trust in AI. We believe that the results of this simulation study can serve as a foundational step for large-scale clinical investigations.

Methods

Study Design

This study was a prospective cohort study in which a web-based questionnaire-based survey was conducted among embryologists with varying degrees of experience. The intra- and interobserver agreements of the evaluation of 360 day-5 embryos by the embryologists and the use of a self-developed AI tool were assessed as references for embryologists.

Ethical Considerations

The embryo images used in our study were collected from 7 IVF clinics in accordance with the Declaration of Helsinki and the institutional review boards (IRBs) of Miraewaheemang Hospital (2022-RESEARCH-01), Good Moonhwa Hospital (GMH-2022-01), the HI Fertility Center (HIRB 2022-01), Seoul Rachel Fertility Center (RTR-2022-01), Ajou University Hospital (AJIRB-MED-MDB-21-716), Pusan National University Hospital (2204-003-113), and Seoul National University Bundang Hospital (B-2208-772-104). Because this study used retrospective data collected through the IRBs of the aforementioned institutions, informed consent was waived, and the personal information contained in the data were deidentified.

Survey

The survey was conducted in three phases via Google Forms: (1) phase 1, an initial assessment (December 23, 2022, to January 22, 2023); (2) phase 2, a validation assessment (March

XSL•FO
RenderX

6, 2023, to April 5, 2023); and (3) phase 3, an AI-guided assessment (March 6, 2023, to April 5, 2023). The participants received an email with the link to the Google Forms and an attachment of original embryo images included in the questions in case they wanted an enlarged view. The submitted survey results were collected and used for analysis.

To measure intra- and interobserver agreements, the initial assessment and the validation assessment were performed 1 month apart and the questions in the initial assessment were identical to those in the validation assessment. Intraobserver agreement was analyzed based on the differences in responses between the initial and validation assessments (phases 1 and 2), while interobserver agreement was assessed based on the average of the responses. To assess the efficacy of AI guidance, the AI-guided assessment (phase 3) was performed right after the validation assessment.

The questionnaire was divided into 2 major sections—items designed to analyze the accuracy of embryo selection and demographic information (age, sex, highest educational level, and tenure). Each assessment consisted of 90 questions and each question consisted of images of 3 embryos that did not result in clinical pregnancy and 1 embryo that did. In the initial and validation assessments, the embryologists were asked to arrange the images in the order of the embryos with the highest likelihood of pregnancy. In the AI-guided assessment, AI scores were provided alongside the embryo images, and the embryologists were asked to reorder the embryos based on their perceived likelihood of pregnancy while considering the AI scores (Figure 1).

Figure 1. Questions on embryo selection from the web survey. (A) Phase 2 of the survey—question on embryo selection without AI scores. (B) Phase 3 of the survey—question on embryo selection with AI scores. AI: artificial intelligence.

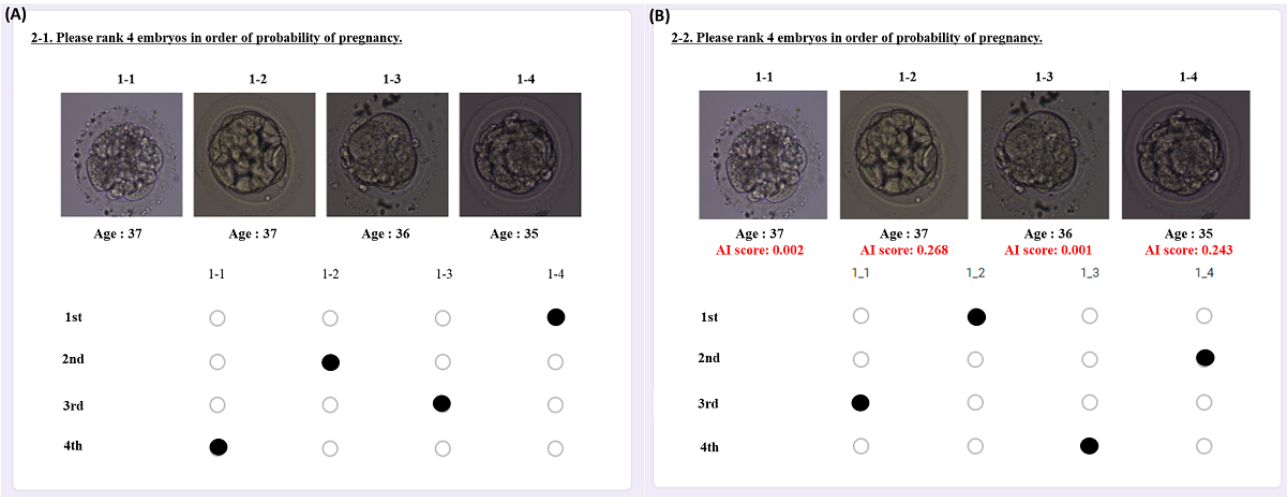

Out of 90 questions, 70 questions featured day-5 embryos at the same developmental stage and the remaining 20 questions used randomly selected day-5 embryos regardless of developmental stage to compare the effect of the developmental stage of embryos in embryo selection. To minimize the effect of women's age, each question contained images of embryos from women of the same age group; younger than 37 years of age and 37 years of age and older.

After the survey, the embryologists were asked to express their opinions on the aspects of (1) the difficulty of the test and the reason for their opinion about the difficulty; (2) the reason for the criteria that they considered when ranking the selected embryos (eg, age, AI score, and image); (3) the reason for changing the answer based on AI score; and (4) consideration when selecting embryos conventionally, as opposed to using a test approach.

Participants

We conducted an initial assessment with 34 embryologists to build a baseline and included 27 more embryologists for validation and AI-guided assessment. A total of 61 embryologists were recruited from 12 different IVF clinics and had clinical experience of 1 to over 30 years. All participants gave informed consent web-based before participating, and those who refused to give consent were excluded. Among them, 50 Korean embryologists were certified by the Korean Association for Clinical Embryologists and conducted an average of 40-150 cases per month. Of the remaining embryologists, 9 were from Malaysia and 2 were from the United States. These participants were divided into two groups, (1) a junior group consisting of embryologists with embryo grading experience of more than or equal to 5 years and (2) a senior group consisting of embryologists with embryo grading experience of more than 5 years.

All 61 participants successfully completed the validation and AI-guided assessments without any loss to follow up. To analyze the results of the validation and AI-guided assessments conducted before and after referring to the AI scores, 29 additional participants were recruited. Statistical analyses were conducted on all participants together and between groups (Figure 2), respectively. From the junior group, 18, 31, and 31 embryologists participated in the initial, validation, and AI-guided assessments, respectively, and from the senior group 16, 30, and 30 embryologists participated in the initial, validation, and AI-guided assessments, respectively.

XSL•FO
RenderX

Figure 2. Study design and participant flow diagram according to the survey phase. AI: artificial intelligence.

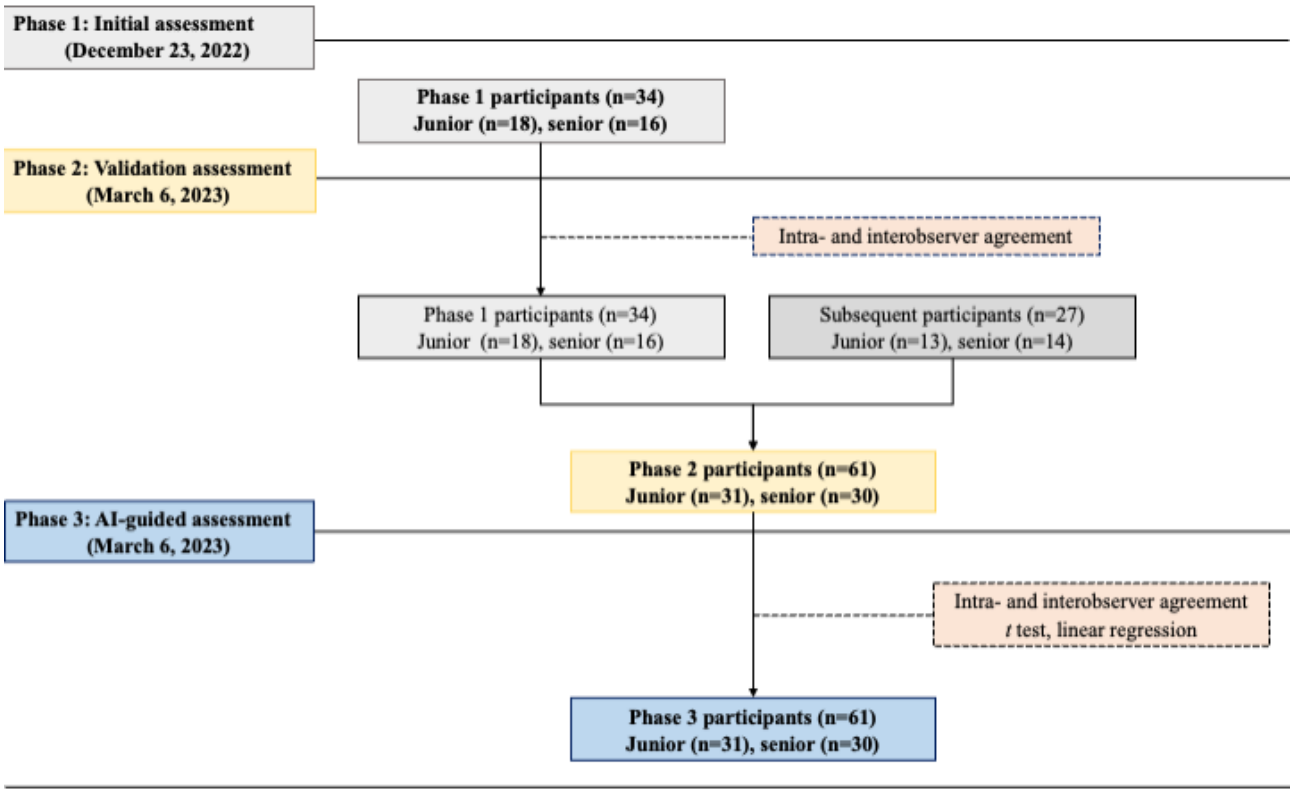

Evaluation of the Impact of AI Guidance on Embryo Selection

We measured the accuracy of embryo selection over 4 cycles to determine how much it increased when the embryologists were guided by AI. The first cycle refers to the case in which the embryo selected as rank 1 resulted in clinical pregnancy, and the second cycle refers to the case in which the embryo selected as rank 1 or 2 resulted in clinical pregnancy. Likewise, the third cycle means that the embryo selected as 1, 2, or 3 resulted in clinical pregnancy, and the fourth cycle means 100% because there are 4 examples per question.

Evaluation of Trust in AI

The model used to infer the AI scores per image was trained on 2555 day-5 embryo images collected from 7 Korean IVF clinics. A total of 2555 images were divided into 2 sets—a training data set consisting of 2043 (80%) images and a model performance test data set containing 512 (20%) images. We then used a 3-fold cross-validation approach to further divide the 2043 training images into 3 separate folds. Each fold was used for both training and validation of the model, and performance was assessed using a fixed model performance test data set. When trained using the ResNet50 architecture, the performance resulted in an area under the receiver operating characteristic of 0.716 and an accuracy of 0.663 [21]. A total of 360 embryo images used in this survey were extracted from 512 images in the data set that were not used to train the AI model. For our questions in the clinical study, the accuracy was 59 (66%), which is similar to the accuracy of the model test set. In assessing the accuracy of the AI model's predictions, we used a method centered on the alignment of AI scores with the actual clinical outcomes. For each set of embryo images, encompassing 4 images per case, our criterion for determining accuracy was the extent to which the AI's highest score matched with the embryo led to a successful pregnancy.

To determine the extent to which embryologists relied on AI scores to modify their responses, we defined the AI trust level as follows:

The formula resulted in a trust level of 0-1 for each embryologist. A level of 0 indicates that the embryologist performed all modifications to their responses on a subjective basis with no reference to the AI, whereas a level of 1 indicates that the embryologist performed all modifications so that their ranking was as consistent with the AI score as possible.

Statistical Analysis

Statistical analyses were performed to assess the consistency of embryo scoring between the embryologists. The Cohen κ coefficient was used to evaluate the intraobserver agreement between the scores given by the same embryologist at 2 different time points [22], whereas Fleiss κ coefficient was used for interobserver agreement between the scores given by different embryologists [23]. The κ coefficient was subsequently construed as excellent (≥0.80), good (0.60-0.79), moderate (0.40-0.59), poor (0.20-0.39), or very poor (<0.20) levels of intra- and interobserver agreements [24]. In this study, we used a 2-tailed *t* test and linear regression to compare the selection accuracy and AI trust level between junior and senior groups. IBM SPSS Statistics (version 29.0; IBM Corporation) was used to quantify intra- and interobserver agreements, while the Python

XSL•FO
RenderX

(version 3.8.0; Python Software Foundation) was used to conduct *t* tests and linear regression analyses. When performing the *t* test, as our data consisted of more than 30 samples per group, we assumed a normal distribution based on the central limit theorem.

Results

Demographics

The demographics of the 2 groups (Table 1) show that the junior group comprised 25 (81%) female participants and 6 (19%) male participants. The most common age subgroup of the junior group was the 30-39 (n=16, 52%), followed by 20-29 (n=14, 45%) and 40-49 (n=1, 3%). The senior group, on the other hand, had a proportion of 21 (70%) female participants and 9 (30%) male participants, with the largest proportion of embryologists (n=10, 33%) being in their 40s and 50s, followed by those in their 30s (n=9, 30%) and 20s (n=1, 4%). Over half of the junior and senior groups (n=18, 58% and n=15, 50%, respectively) had a master's degree. The junior group comprised 13 (42%) university graduates and no doctoral degree holders. The senior group had a high percentage of doctoral degree holders (n=10, 33%), followed by bachelor's degrees (n=5, 17%). In relation to embryo assessment expertise, the junior group had an average of 1.6 years of experience, while the senior group had an average of 13.2 years of experience.

Table 1. Demographic characteristics of the junior and senior groups.

Characteristics	Junior	Senior
Sex, n (%)		
Male	6 (19)	9 (30)
Female	25 (81)	21 (70)
Age (years), n (%)		
20-29	14 (45)	1 (4)
30-39	16 (52)	9 (30)
40-49	1 (3)	10 (33)
≥50	0 (0)	10 (33)
Highest level of education, n (%)		
Bachelor's degree	13 (42)	5 (17)
Master's degree	18 (58)	15 (50)
Doctoral degree	0 (0)	10 (33)
Experience in embryo selection (years), mean (SD)	1.6 (1.9)	13.2 (7.4)

Intra- and Interobserver Agreements

The evaluation of intraobserver agreement revealed a Cohen κ score of 0.662 between the initial and validation assessments, indicating good (0.60-0.79) concordance between the responses of 1 embryologist at 2 separate time points. The coefficients for the junior and senior groups were 0.659 and 0.664, respectively, indicating that the less experienced group was able to provide consistent responses at a similar level as the experienced group (Table 2). Additionally, the correlation coefficients between the validation and AI-guided assessments were 0.735 for the overall population and 0.698 and 0.773 for the junior and senior groups, respectively. This indicates that all participants showed improved consistency after the AI guidance.

Table 2. Results of the evaluation of the intra- and interobserver agreements.

Characteristics	Intraobserver agreement, Cohen κ coefficient (95% CI)		Interobserver agreement, Fleiss κ coefficient (95% CI)	
	Without AI[a] (phase 1[b] vs phase 2[c])	With AI vs without AI (phase 2 vs phase 3[d])	Without AI (phase 2)	With AI (phase 3)
Overall	0.662 (0.631-0.692)	0.735 (0.670-0.770)	0.392 (0.389-0.395)	0.521 (0.518-0.524)
Junior	0.659 (0.603-0.714)	0.698 (0.653-0.744)	0.355 (0.349-0.360)	0.527 (0.521-0.532)
Senior	0.664 (0.604-0.710)	0.773 (0.722-0.825)	0.440 (0.434-0.445)	0.524 (0.518-0.529)

[a]AI: artificial intelligence.
[b]Initial assessment.
[c]Validation assessment.
[d]AI-guided assessment

Next, we measured the Fleiss κ coefficient for interobserver agreement between multiple embryologists and found that validation assessment showed a poor coefficient (0.20-0.39) of 0.392 for the total population. Interobserver agreement within

XSL•FO
RenderX

the junior group was indicated by a poor coefficient of 0.355, whereas within the senior group, the agreement was indicated by a moderate coefficient of 0.440 (Table 2). After referring to the AI, the coefficient was moderate at 0.521 for the entire population, and the junior and senior groups showed improved concordance of 0.527 and 0.523, respectively. This result suggests that the junior group could make judgments with consistency that were similar to those of the senior group after the AI's guidance.

Impact of AI Guidance on Embryo Selection

In the first cycle, the accuracy of the embryologists was 34 (38%) and that of the AI model was 59 (66%; Table 3). When the embryologist was guided by the AI score, the accuracy rate increased to 45 (50%). The AI model outperformed embryologists in selecting an embryo that led to pregnancy by 25 (28%), and embryologists with AI guidance outperformed embryologists without AI guidance by 11 (12%). The difference in accuracy between the embryologists and AI model was 20 (23%) in the second cycle and 13 (15%) in the third cycle, and the performance gap between the embryologists with AI guidance and AI model was reduced to 11 (13%) in the second cycle and 7 (8%) in the third cycle.

Table 3. Comparison of cumulative accuracies correct responses of embryologists, embryologists with AI guidance, and AI for the prediction of clinical pregnancy on a test of accurate embryo selection with 90 questions.

Characteristics	Cumulative accuracycorrect responses, n (%)			
	First cycle	Second cycle	Third cycle	Fourth cycle
Embryologists' selection	34 (38)	57 (63)	73 (81)	90 (100)
Embryologists' selection with AI[a] guidance	45 (50)	66 (73)	79 (88)	90 (100)
AI selection	59 (66)	77 (86)	86 (96)	90 (100)

[a]AI: artificial intelligence.

Relationship Between AI Trust Level and Embryo Selection Accuracy

The relationship between AI trust level and embryo selection accuracy was determined through statistical analysis of the accuracy of embryo selection by the junior and senior groups. The analysis revealed that in the validation assessment, the responses of the junior group differed significantly from those of the senior groups, with a *P* value of <.001 in the *t* test (Figure 3). The average accuracy of embryo selection by the junior group was 33.516 (SD 3.688), while that of the senior group was 35.967 (SD 2.580), indicating that embryologists with over 5 years of experience had significantly higher embryo selection ability.

Figure 3. Within and between-group *t* test results. AI: artificial intelligence.

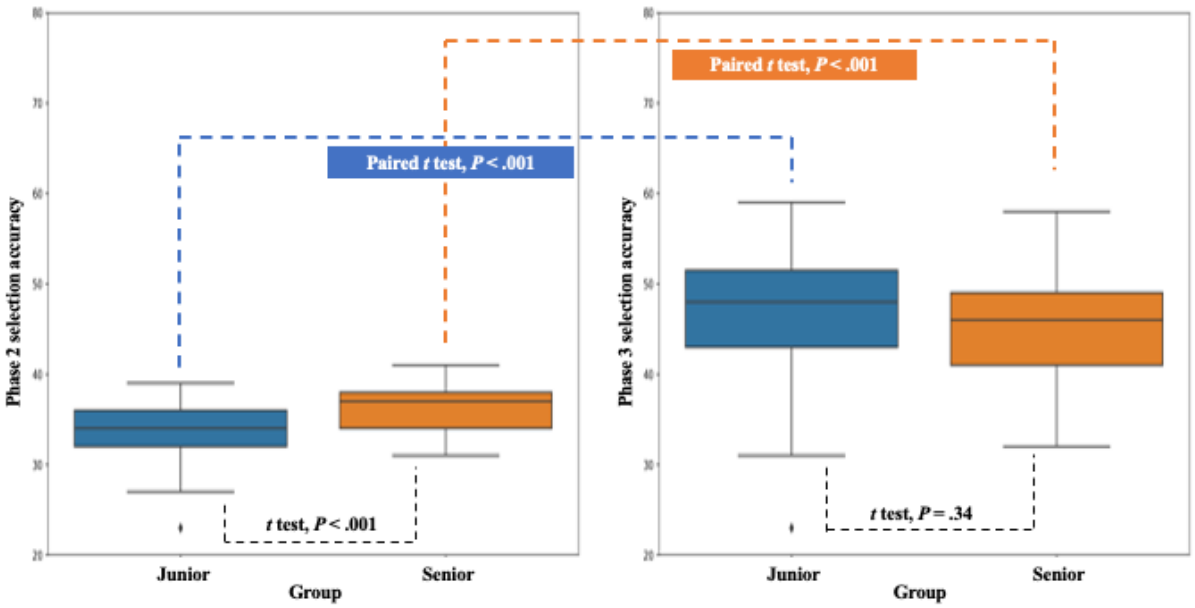

When comparing the 2 groups in the AI-guided assessment, the mean score of the junior group was 46.581 (SD 7.967), and the mean score of the senior group was 44.833 (SD 6.772), with *P*=.34, showing no significant difference between the 2 groups.

XSL•FO
RenderX

In addition, a paired *t* test was performed between the validation and AI-guided assessments to determine whether there was a significant difference in selection accuracy before and after the groups referred to the AI score. The *P* value was less than .001 for both groups, confirming that the score increased after referring to AI.

Before checking the relationship between the AI trust level and embryo selection accuracy, we tested whether there was a difference in AI trust levels between the 2 groups. The AI trust level of the junior group was 0.581 (SD 0.244), whereas that of the senior group was 0.443 (SD 0.278). The *P* value of the *t* test was .047, which confirmed that the confidence of the junior group was significantly higher than that of the senior group (Figure 4). Subsequently, we performed a regression analysis with AI trust level as the independent variable and embryo selection score as the dependent variable. We found that the scores of both groups increased with the increase in the trust level. In addition, the regression coefficient of the junior group was 29.209, compared to 22.870 for the senior group; therefore, the slope of the embryo selection score with trust level increased sharply (Table 4). We further examined the distribution of the denominator in the AI trust level calculation by group. The distribution spanned from the smallest (n=29) to the largest number of different questions (n=61). We found that the junior group had 4 (13%) cases in which the top-ranked embryo was in the category in which 40 or fewer questions differed from the AI, while the senior group had 7 (23%) cases. This suggests that the senior embryologists gave a higher number of answers that were similar to those of the AI (Multimedia Appendix 1).

Figure 4. Results of the groupwise analyses of AI trust level through *t* test and regression analysis. AI: artificial intelligence.

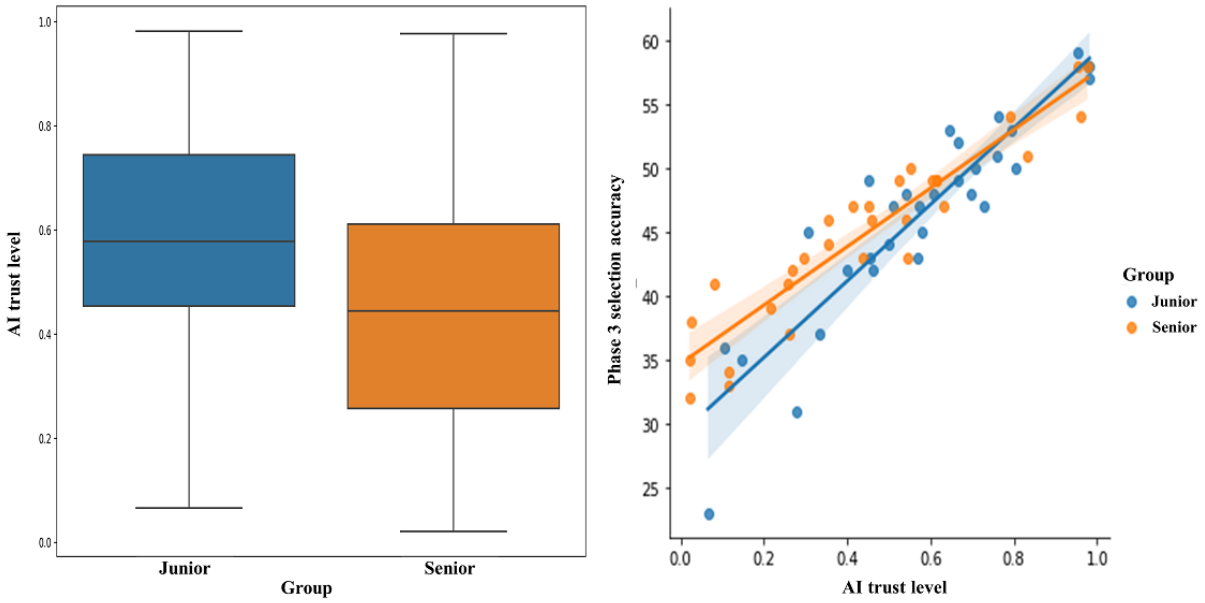

Table 4. Results of the groupwise regression analysis.

Characteristics	Coefficient (95% CI)	SE	*P* value
Junior embryologist	29.209 (24.996-34.821)	1.514	<.001
Senior embryologist	22.870 (19.582-26.158)	1.605	<.001

Discussion

Principal Results

The primary objective of this study was to demonstrate the potential clinical benefits of integrating AI into the embryo-ranking process, ultimately increasing the likelihood of achieving a clinical pregnancy. The area under the receiver operating characteristic curve and the accuracy of the model used in this study were 0.716 and 0.663, respectively. This was comparable to previous AI models developed upon 2D static images, with an accuracy of 0.64 [12] and the area under the receiver operating characteristic ranging from 0.6 to 0.7 [25]. In this study, we demonstrated that the AI model's number of correct test responses of 59 of 90 (66%) outperformed that of the embryologists, who had 34 (38%). Previous studies used the historical data of embryologists' morphological grading to simulate embryo selection and found that the embryologists' accuracy was between 0.47 and 0.65. Our study recruited a large number of embryologists rather than using the historical data, which might have resulted from the difference in the accuracy. Our investigation reveals several critical findings that shed light on the role of AI in this context.

Our study revealed that intra- and interobserver agreements among embryologists in ranking embryos improved with the

XSL•FO
RenderX

assistance of AI. It is noted that the embryologists, particularly junior embryologists, exhibited relatively low interobserver agreement, but this was mitigated by the AI's guidance, effectively leveling the performance between junior and senior embryologists. Regarding intraobserver agreement, while the Cohen κ score of 0.662 is statistically considered good (0.60-0.79), its clinical implications may differ, as evidenced by embryologists changing their responses to identical questions in 23 (25%) of questions over 1 month. This variability underscored the need for a comparative analysis to assess the accuracy of embryo selection before and after the introduction of AI guidance. We used an AI model that demonstrated industrial standard performance and achieved an area under the receiver operating characteristic curve of 0.716 [21]. This performance metric aligns with previous studies that used 2D images of day-5 embryos [12,25]. Furthermore, the positive correlation between AI scores and traditional manual grading by embryologists reinforced trust in the AI model (Multimedia Appendix 2).

To further substantiate the clinical benefits of AI, we conducted a blinded test in which embryologists ranked embryos without knowing their future outcomes. The findings were noteworthy as they indicated that the highest accuracy of selecting the most viable embryos was achieved by AI models followed by embryologists with AI guidance and embryologists without AI guidance. This observation suggests that our AI model has the potential to assist embryologists in the selection of the most viable embryos, thereby increasing the probability of successful pregnancies per cycle, while potentially reducing the time to conception.

This study was designed to closely simulate a clinical setting. Unlike previous simulation studies [19,25], we leveraged the embryologists' actual rankings rather than the rankings derived from their historical grading records. Although the morphological evaluation methodology is well established, the criteria for grading vary, resulting in limited intra- and interobserver agreements. Previous studies have used manual grades as numeric scores mapped from alphanumeric historical grades [25] or used random and Gardner-based scores as proxies for embryological accuracy [19]. These approaches face challenges in translating grades into ranks owing to the nonlinear nature of the embryo grading system.

Furthermore, we compared 3 distinct scenarios—embryologist only, embryologists with AI guidance, and AI-only rankings. Previous studies have predominantly focused on comparing the embryologists' independent assessments with AI-only evaluations [12,13]. However, it is imperative to include a scenario in which embryologists are guided by AI, as this closely mirrors the most likely clinical scenario, in which AI aids, rather than replaces, human judgment due to liability concerns.

In addition, we controlled for blastocyst developmental stages in 70 of 90 questions and maintained consistent age groups across all questions, allowing us to compare the outcomes between stage-controlled and random-stage questions. This approach closely mirrored the clinical context of embryo ranking. We observed the following rates of correct responses: 22 (32%) and 12 (61%) without AI, 30 (44%) and 14 (71%) with AI guidance, and 42 (60%) and 17 (85%) for AI-only rankings in 70 blastocyst stage-adjusted questions and 20 random-stage questions, respectively. Notably, the accuracy of the responses to questions involving randomly selected embryos with different blastocyst stages exceeded that of the stage-adjusted questions. This observation suggests that AI may offer the most substantial benefits in scenarios where embryologists frequently encounter assessment challenges. Our research design, which focused on embryos at similar stages, proved to be the most suitable for evaluating the clinical efficacy of the AI model. A comprehensive questionnaire was also administered to gain deeper insight. The survey revealed significantly higher levels of trust in AI among junior embryologists than among their senior counterparts. Although junior embryologists initially exhibited lower accuracy rates than their senior peers before AI guidance, their performance improved and converged with those of their seniors after the AI intervention. An intriguing trend emerged from the regression analysis of confidence—for every 1 unit increase in confidence, the junior group demonstrated a more substantial increase of 29.2 points, compared to the senior group's increase of 22.87 points (Table 4). Interestingly, the current level of trust in AI appeared relatively modest, with 38 (62%) of the surveyed embryologists indicating that their ranking considerations included embryo morphology, age, and AI score. In contrast, 17 (28%) of the surveyed embryologists prioritized embryo morphology, AI scores, and age in their ranking considerations. This underscores the need for further research to establish clinical efficacy and foster trust among embryologists, particularly their senior counterparts.

Limitations

The AI model that we developed is highly effective in analyzing 2D static images. In the practical context of embryo selection, embryologists can assess embryos from multiple perspectives under a microscope. However, our experiment necessitated judgments based on single images captured from a single viewpoint. Additionally, our data set comprised images captured by embryologists before embryo transfer; consequently, we lacked comprehensive kinetic information throughout the entire developmental process. Given the above limitations, we considered studies that covered complete embryonic development, such as time-lapse video analyses [26-30]. However, this method presents a set of challenges. Time-lapse equipment is expensive and requires embryologists to visually monitor the entire process, which requires considerable time and effort. Interestingly, prior research has suggested that using the final image taken on day 5 yields a predictive performance for pregnancy outcomes similar to that achieved with time-lapse images capturing the entire developmental process [31]. This insight led us to make a strategic decision to leverage our expertise in 2D image analysis to design a cost-effective and time-efficient experimental setup.

All the embryologists surveyed in this study reported varying levels of difficulty in the embryo ranking task, with the majority describing it as slightly difficult 32 (52%) or moderately difficult 22 (36%). Their perceived reasons for this difficulty include factors such as image quality and fixed focus. The embryo images used in this study were collected from various IVF

XSL•FO
RenderX

clinics by introducing variations in magnification and color. This variability may have contributed to the less precise responses, as embryologists selected embryos under conditions that differed from standard practices. Therefore, for more accurate comparisons between AI and embryologists, future experiments should be conducted by collecting images of uniform size, magnification, and color within a single institution.

Conclusions

To date, there is a lack of practical research on the extent to which AI can assist researchers in embryo selection. In this study, we demonstrated that AI is crucial for successfully selecting embryos that provide high chances of pregnancy. This effect was particularly pronounced among embryologists with less than 5 years of experience who had more trust in AI scores. Thus, this study suggests that using AI as an auxiliary tool in clinical practice has the potential to enhance embryo assessment and increase the probability of a successful pregnancy.

Acknowledgments

The authors thank the 61 embryologists who participated in the survey. This study was supported by the Starting Growth Technological R&D Program (TIPS Program; S3318926) funded by the Ministry of SMEs and Startups in 2022, as well as by the Korea Medical Device Development Fund grant funded by the Korean government (Ministry of Science and ICT, Ministry of Trade, Industry and Energy, Ministry of Health & Welfare, and Ministry of Food and Drug Safety; RS-2023-00254175).

Data Availability

The data sets analyzed during this study are available in the AI Hub repository [32].

Authors' Contributions

HMK and HJL contributed to the conceptualization. HMK contributed to the methodology. HMK and CL helped with the software. JHP, MKC, NYK, and MK performed the validation. HMK performed the formal analysis. JHP, MKC, NYK, and MK contributed to the acquisition of data. HMK and HK performed writing—original draft preparation. All authors contributed to the writing—review and editing. HMK performed the visualization. HJL did the supervision. All authors have read and agreed to the published version of the paper.

Conflicts of Interest

No potential conflict of interest relevant to this paper is reported. HMK reports stock options from Kai Health. HK, CL, MK, NYK, JHP and MKC have no conflicts of interest to disclose. HJL is a shareholder of Kai Health.

Multimedia Appendix 1
Distribution of questions in which top-ranked embryos differed between embryologists and artificial intelligence by group.
[DOCX File , 32 KB - jmir_v26i1e52637_app1.docx]

Multimedia Appendix 2
Correlation between artificial intelligence scores and traditional manual grading.
[DOCX File , 73 KB - jmir_v26i1e52637_app2.docx]

References

1. Wang J, Sauer MV. In vitro fertilization (IVF): a review of 3 decades of clinical innovation and technological advancement. Ther Clin Risk Manag 2006;2(4):355-364 [FREE Full text] [doi: 10.2147/tcrm.2006.2.4.355] [Medline: 18360648]
2. Chow DJX, Wijesinghe P, Dholakia K, Dunning KR. Does artificial intelligence have a role in the IVF clinic? Reprod Fertil 2021;2(3):C29-C34 [FREE Full text] [doi: 10.1530/RAF-21-0043] [Medline: 35118395]
3. Racowsky C, Vernon M, Mayer J, Ball GD, Behr B, Pomeroy KO, et al. Standardization of grading embryo morphology. J Assist Reprod Genet 2010;27(8):437-439 [FREE Full text] [doi: 10.1007/s10815-010-9443-2] [Medline: 20532975]
4. Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. Hum Reprod 2011;26(6):1270-1283 [FREE Full text] [doi: 10.1093/humrep/der037] [Medline: 21502182]
5. Hill MJ, Richter KS, Heitmann RJ, Graham JR, Tucker MJ, DeCherney AH, et al. Trophectoderm grade predicts outcomes of single-blastocyst transfers. Fertil Steril 2013;99(5):1283-1289.e1 [FREE Full text] [doi: 10.1016/j.fertnstert.2012.12.003] [Medline: 23312233]
6. Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. Cochrane Database Syst Rev 2020;9(9):CD005291 [FREE Full text] [doi: 10.1002/14651858.CD005291.pub3] [Medline: 32898291]

XSL•FO
RenderX

7. Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single day 5 embryo for transfer: a multicenter study. Hum Reprod 2017;32(2):307-314 [FREE Full text] [doi: 10.1093/humrep/dew330] [Medline: 28031323]
8. Bormann CL, Thirumalaraju P, Kanakasabapathy MK, Kandula H, Souter I, Dimitriadis I, et al. Consistency and objectivity of automated embryo assessments using deep neural networks. Fertil Steril 2020;113(4):781-787.e1 [FREE Full text] [doi: 10.1016/j.fertnstert.2019.12.004] [Medline: 32228880]
9. Adolfsson E, Andershed AN. Morphology vs morphokinetics: a retrospective comparison of inter-observer and intra-observer agreement between embryologists on blastocysts with known implantation outcome. JBRA Assist Reprod 2018;22(3):228-237 [FREE Full text] [doi: 10.5935/1518-0557.20180042] [Medline: 29912521]
10. Paternot G, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. Reprod Biol Endocrinol 2009;7:105 [FREE Full text] [doi: 10.1186/1477-7827-7-105] [Medline: 19788739]
11. Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. Hum Reprod 2013;28(12):3215-3221 [FREE Full text] [doi: 10.1093/humrep/det366] [Medline: 24070998]
12. VerMilyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum Reprod 2020;35(4):770-784 [FREE Full text] [doi: 10.1093/humrep/deaa013] [Medline: 32240301]
13. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digit Med 2019;2:21 [FREE Full text] [doi: 10.1038/s41746-019-0096-y] [Medline: 31304368]
14. Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online 2020;41(4):585-593. [doi: 10.1016/j.rbmo.2020.07.003] [Medline: 32843306]
15. Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, et al. Performance of a deep learning based neural network in the selection of human blastocysts for implantation. Elife 2020;9:e55301 [FREE Full text] [doi: 10.7554/eLife.55301] [Medline: 32930094]
16. Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Hum Reprod 2019;34(6):1011-1018 [FREE Full text] [doi: 10.1093/humrep/dez064] [Medline: 31111884]
17. Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. Fertil Steril 2020;114(5):914-920 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.09.157] [Medline: 33160513]
18. Friedenthal J, Hernandez-Nieto C, Roth RM, Slifkin R, Gounko D, Lee JA, et al. Clinical implementation of algorithm-based embryo selection is associated with improved pregnancy outcomes in single vitrified warmed euploid embryo transfers. J Assist Reprod Genet 2021;38(7):1647-1653 [FREE Full text] [doi: 10.1007/s10815-021-02203-3] [Medline: 33932196]
19. Diakiw SM, Hall JMM, VerMilyea M, Lim AYX, Quangkananurug W, Chanchamroen S, et al. An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos. Reprod Biomed Online 2022;45(6):1105-1117 [FREE Full text] [doi: 10.1016/j.rbmo.2022.07.018] [Medline: 36117079]
20. Bori L, Meseguer F, Valera MA, Galan A, Remohi J, Meseguer M. The higher the score, the better the clinical outcome: retrospective evaluation of automatic embryo grading as a support tool for embryo selection in IVF laboratories. Hum Reprod 2022;37(6):1148-1160 [FREE Full text] [doi: 10.1093/humrep/deac066] [Medline: 35435210]
21. Kim HM, Ko T, Kang H, Choi S, Park JH, Chung MK, et al. Improved prediction of clinical pregnancy using artificial intelligence with enhanced inner cell mass and trophectoderm images. Sci Rep 2024;14(1):3240 [FREE Full text] [doi: 10.1038/s41598-024-52241-x] [Medline: 38331914]
22. Rau G, Shih YS. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. J Engl Acad Purp 2021;53:101026. [doi: 10.1016/j.jeap.2021.101026]
23. Rücker G, Schimek-Jasch T, Nestle U. Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using Fleiss' kappa. Methods Inf Med 2012;51(6):489-494. [doi: 10.3414/ME12-01-0005] [Medline: 23160666]
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]
25. Loewke K, Cho JH, Brumar CD, Maeder-York P, Barash O, Malmsten JE, et al. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. Fertil Steril 2022;117(3):528-535 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.11.022] [Medline: 34998577]
26. Zaninovic N, Irani M, Meseguer M. Assessment of embryo morphology and developmental dynamics by time-lapse microscopy: is there a relation to implantation and ploidy? Fertil Steril 2017;108(5):722-729 [FREE Full text] [doi: 10.1016/j.fertnstert.2017.10.002] [Medline: 29101997]
27. Athayde Wirka K, Chen AA, Conaghan J, Ivani K, Gvakharia M, Behr B, et al. Atypical embryo phenotypes identified by time-lapse microscopy: high prevalence and association with embryo development. Fertil Steril 2014;101(6):1637-1648.e1 [FREE Full text] [doi: 10.1016/j.fertnstert.2014.02.050] [Medline: 24726214]

XSL•FO
RenderX

28. Kirkegaard K, Agerholm IE, Ingerslev HJ. Time-lapse monitoring as a tool for clinical embryo assessment. Hum Reprod 2012;27(5):1277-1285 [FREE Full text] [doi: 10.1093/humrep/des079] [Medline: 22419744]
29. Aparicio-Ruiz B, Romany L, Meseguer M. Selection of preimplantation embryos using time-lapse microscopy in in vitro fertilization: state of the technology and future directions. Birth Defects Res 2018;110(8):648-653. [doi: 10.1002/bdr2.1226] [Medline: 29714056]
30. Khan A, Gould S, Salzmann M. Segmentation of developing human embryo in time-lapse microscopy. 2016 Presented at: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); April 13-16, 2016; Prague, Czech Republic p. 930-934. [doi: 10.1109/isbi.2016.7493417]
31. Chen M, Wei S, Hu J, Yuan J, Liu F. Does time-lapse imaging have favorable results for embryo incubation and selection compared with conventional methods in clinical in vitro fertilization? A meta-analysis and systematic review of randomized controlled trials. PLoS One 2017;12(6):e0178720 [FREE Full text] [doi: 10.1371/journal.pone.0178720] [Medline: 28570713]
32. AI Hub repository. AI Hub. URL: https://www.aihub.or.kr/ [accessed 2021-05-29]

Abbreviations

AI: artificial intelligence
IRB: institutional review board
IVF: in vitro fertilization

Edited by T de Azevedo Cardoso; submitted 12.09.23; peer-reviewed by M Zhou; comments to author 12.10.23; revised version received 23.10.23; accepted 06.05.24; published 03.06.24.

Please cite as:
Kim HM, Kang H, Lee C, Park JH, Chung MK, Kim M, Kim NY, Lee HJ
Evaluation of the Clinical Efficacy and Trust in AI-Assisted Embryo Ranking: Survey-Based Prospective Study
J Med Internet Res 2024;26:e52637
URL: https://www.jmir.org/2024/1/e52637
doi:10.2196/52637
PMID:38830209

©Hyung Min Kim, Hyoeun Kang, Chaeyoon Lee, Jong Hyuk Park, Mi Kyung Chung, Miran Kim, Na Young Kim, Hye Jun Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4

Adi Lahat[1,2*], MD; Kassem Sharif[1,3*], MD; Narmin Zoabi[1], MD; Yonatan Shneor Patt[3], MD; Yousra Sharif[4], MD; Lior Fisher[3], MD; Uria Shani[3], MD; Mohamad Arow[3], MD; Roni Levin[3], MD; Eyal Klang[5], MD

[1]Department of Gastroenterology, Chaim Sheba Medical Center, Affiliated with Tel Aviv University, Ramat Gan, Israel

[2]Department of Gastroenterology, Samson Assuta Ashdod Medical Center, Affiliated with Ben Gurion University of the Negev, Be'er Sheva, Israel

[3]Department of Internal Medicine B, Sheba Medical Centre, Tel Aviv, Israel

[4]Department of Internal Medicine C, Hadassah Medical Center, Jerusalem, Israel

[5]Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
Adi Lahat, MD
Department of Gastroenterology
Chaim Sheba Medical Center
Affiliated with Tel Aviv University
Tel Hashomer
Ramat Gan, 5262100
Israel
Phone: 972 5302060
Email: zokadi@gmail.com

Abstract

Background: Artificial intelligence, particularly chatbot systems, is becoming an instrumental tool in health care, aiding clinical decision-making and patient engagement.

Objective: This study aims to analyze the performance of ChatGPT-3.5 and ChatGPT-4 in addressing complex clinical and ethical dilemmas, and to illustrate their potential role in health care decision-making while comparing seniors' and residents' ratings, and specific question types.

Methods: A total of 4 specialized physicians formulated 176 real-world clinical questions. A total of 8 senior physicians and residents assessed responses from GPT-3.5 and GPT-4 on a 1-5 scale across 5 categories: accuracy, relevance, clarity, utility, and comprehensiveness. Evaluations were conducted within internal medicine, emergency medicine, and ethics. Comparisons were made globally, between seniors and residents, and across classifications.

Results: Both GPT models received high mean scores (4.4, SD 0.8 for GPT-4 and 4.1, SD 1.0 for GPT-3.5). GPT-4 outperformed GPT-3.5 across all rating dimensions, with seniors consistently rating responses higher than residents for both models. Specifically, seniors rated GPT-4 as more beneficial and complete (mean 4.6 vs 4.0 and 4.6 vs 4.1, respectively; $P<.001$), and GPT-3.5 similarly (mean 4.1 vs 3.7 and 3.9 vs 3.5, respectively; $P<.001$). Ethical queries received the highest ratings for both models, with mean scores reflecting consistency across accuracy and completeness criteria. Distinctions among question types were significant, particularly for the GPT-4 mean scores in completeness across emergency, internal, and ethical questions (4.2, SD 1.0; 4.3, SD 0.8; and 4.5, SD 0.7, respectively; $P<.001$), and for GPT-3.5's accuracy, beneficial, and completeness dimensions.

Conclusions: ChatGPT's potential to assist physicians with medical issues is promising, with prospects to enhance diagnostics, treatments, and ethics. While integration into clinical workflows may be valuable, it must complement, not replace, human expertise. Continued research is essential to ensure safe and effective implementation in clinical environments.

(J Med Internet Res 2024;26:e54571) doi:10.2196/54571

KEYWORDS

ChatGPT; chat-GPT; chatbot; chatbots; chat-bot; chat-bots; natural language processing; NLP; artificial intelligence; AI; machine learning; ML; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model;

XSL•FO
RenderX

practical models; internal medicine; ethics; ethical; ethical dilemma; ethical dilemmas; bioethics; emergency medicine; EM medicine; ED physician; emergency physician; emergency doctor

Introduction

Artificial intelligence (AI), particularly chatbot systems, is becoming an instrumental tool in health care, aiding clinical decision-making and patient engagement [1,2]. These systems, using natural language processing (NLP), offer support to physicians in various clinical scenarios [3]. OpenAI's GPT technology represents a leading example of this innovation [4].

The integration potential of AI in health care is not only a technological advancement; it is a paradigm shift in how medical professionals approach patient care. Leveraging AI, clinicians can access real-time data analysis, personalized treatment recommendations, and even predictive insights into patient health trends. The convergence of AI with traditional medical practices is fostering a new era of precision medicine, where treatments are tailored to individual patient needs and preferences. This personalized approach enhances patient satisfaction and outcomes, while also optimizing resource allocation within the health care system.

Earlier GPT models, like GPT-3.5, were recognized for generating human-like text and have potential across various fields, including medicine, where they can assist in diagnostics, therapeutic planning, telehealth, and patient education [5-15]. Recent AI algorithm's ability to synthesize vast information and recognize complex patterns can enhance clinical accuracy, and reduce errors [16].

With the introduction of GPT-4, an advancement in AI's capabilities, it becomes imperative to evaluate its comparative effectiveness in clinical scenarios [17]. This model, with improved language comprehension and generation, presents a promising opportunity for refined health care applications. Past studies have highlighted the clinical potential of ChatGPT (GPT-3.5), which managed to score approximately 60% on the USMLE (United States Medical Licensing Examination) [18]. Moreover, the more advanced GPT-4 achieved an impressive 87% [19].

Recent advancements in AI technology have paved the way for innovative applications in medical education and patient support, particularly through AI-powered chatbots. Chow et al [20] developed a chatbot using the IBM Watson Assistant platform to make radiotherapy knowledge more accessible to the general public. This chatbot, designed with a user-friendly interface, engages users in a digital question-and-answer format, enhancing their understanding of radiotherapy. Despite its positive feedback, the study highlights the need for improvements in conversational approaches and language inclusivity. In another study, Rebelo et al [21] created a novel web-based assistant to explain the radiation treatment process comprehensively. Using IBM Watson's NLP features, this chatbot guides users through the complex radiotherapy process, from diagnosis to treatment delivery, and has been tested for effectiveness in knowledge transfer. Both studies underscore the potential of AI chatbots in transforming medical communication and education, offering personalized, digital, and reassuring platforms for patients and the general public to access critical health information. These initiatives reflect the growing importance of AI in addressing the need for accessible and accurate medical information, particularly in areas, like radiotherapy, where patient and family education is crucial.

However, the reliability of ChatGPT in delivering health information to patients and health care providers remains questionable. ChatGPT differs from specialized medical chatbots as it is not trained on data sets curated by medical professionals, raising concerns about the accuracy of its medical advice and ethical issues related to patient safety and data privacy. As ChatGPT evolves as a disruptive technology in health care, it faces several challenges. These include its database being possibly outdated, risks of misinformation, and the need for its integration into the medical field to be carefully monitored and guided by ethical frameworks and professional standards [22].

In light of all of the above, the goal of this study is to analyze the performance of ChatGPT-3.5 and ChatGPT-4 in addressing complex clinical and ethical dilemmas and to illustrate their potential role as allies in health care decision-making.

Our study aims to compare the free version of GPT-3.5 with the paid version of GPT-4 in a medical context, focusing on accessibility and performance for a diverse audience. This comparison is vital for understanding the practicality and effectiveness of AI in health care, especially for users with limited resources. By evaluating both versions, we can assess the performance gap, guide resource allocation, and address the democratization of AI technology in medicine. This approach not only helps in setting realistic expectations for the free version's capabilities but also underscores the ethical and social implications of AI accessibility. Such a comparison is crucial for informing both users and developers about the practical use of AI in medical contexts, ensuring the study remains relevant and beneficial to a wide range of potential users including those without access to the paid version.

Furthermore, we aimed to assess and compare the performance of ChatGPT in responding to medical questions across different categories, namely emergency medicine, internal medicine, and ethical dilemmas. We specifically focus on comparing the ratings provided by 2 distinct groups: medical residents and senior physicians. The rationale behind this comparison lies in the differing levels of clinical experience and expertise between these groups. Medical residents, being in the earlier stages of their training, may approach medical questions with a fresh perspective and rely more on foundational knowledge and guidelines. On the other hand, senior physicians, with their extensive clinical experience, may offer nuanced insights and consider broader contextual factors in their assessments. By evaluating responses from both groups, we aimed to gain a comprehensive understanding of ChatGPT's performance across various medical domains and discern potential differences in the depth and accuracy of responses based on clinical experience levels. We believe that this comparison can provide valuable insights into the utility and limitations of using AI language

XSL•FO
RenderX

models, like ChatGPT, in supporting medical decision-making across different stages of clinical training and practice, and to contribute to a broader understanding of AI's expanding role in the medical field.

Methods

Study Design

In this study, 3 forms of comparisons were conducted to evaluate the performance of ChatGPT in responding to medical questions. First, we compared the responses generated by 2 versions of the model, GPT-3.5 and GPT-4, to assess any differences in the quality and accuracy of responses between the 2 iterations. This comparison aimed to elucidate potential improvements or variations in performance introduced by advancements in the underlying AI architecture. Second, we compared the evaluations provided by senior physicians and medical residents on the responses generated by ChatGPT. This analysis aimed to explore potential disparities in the perceived quality and usefulness of responses based on the level of clinical experience and expertise of the evaluator. Finally, we analyzed responses across different types of medical questions, including emergency medicine, internal medicine, and ethical dilemmas, to assess the model's performance across diverse clinical scenarios. This comparison aimed to identify any variations in the model's ability to provide relevant and accurate information across various medical domains. By conducting these 3 forms of comparisons, we aimed to comprehensively evaluate the capabilities of ChatGPT in addressing medical queries across different contexts and user perspectives.

Question Development and Categorization

A total of 4 experienced physicians specialized in internal medicine and emergency medicine collaborated to formulate 176 questions. Questions were generated based on clinical experience and designed to follow clinical guidelines.

The questions were written by 2 senior physicians in internal medicine and emergency medicine with reference to board questions used for the assessment of doctors in training, questions were chosen to reflect different aspects of medicine as seen in everyday practice. Verification was done independently by 2 different physicians and answers were checked to comply with updated guidelines.

Questions were mainly classified into one of three categories, (1) emergency room (ER): questions addressing emergency scenarios; (2) internal medicine: questions concerning diagnosis, treatment, and management of internal medicine conditions; and (3) ethical: questions focusing on ethical considerations in medical practice.

These questions were created to reflect real-world clinical scenarios, aligning with prevailing clinical guidelines and ethical norms.

Selection and Configuration of Models

The study engaged OpenAI's GPT-3.5 and GPT-4 models, recognized for their advanced capabilities in NLP. They were queried through OpenAI's Application Programming Interface to generate answers to the formulated questions.

Prompting the Models

The models were prompted using the specific instruction: "Please answer this {question_type} question in clear, concise, concrete, full, bullets: {question}."

Grading Participants and Process

A total of 8 clinicians from 2 tertiary medical centers, including 4 senior physicians and 4 residents, independently graded the responses. Senior physicians were all active in daily clinical practice, including emergency care within the ER department, between 2 and 4 years subsequent to the completion of residency training. All residents were in their last year of residency (3 out of 4 years) and experienced with internal and emergency medicine. The grading criteria are provided in Textbox 1.

A comprehensive grading guide ensured uniformity in the application of the criteria.

XSL•FO
RenderX

Textbox 1. Grading criteria.

Accuracy

Does it reflect medical understanding (scale 1-5). The answers were graded as follows:

- Very inaccurate: the response shows a fundamental misunderstanding of medical concepts.
- Somewhat inaccurate: the answer contains more incorrect than correct medical information.
- Moderately accurate: the response is generally correct but includes some inaccuracies.
- Mostly accurate: the answer is largely accurate with minor errors or omissions.
- Completely accurate: the response reflects a high level of medical understanding with accurate and precise information.

Relevance

Does the response directly address the asked question, or does it deviate to unrelated subtopics (scale 1-5). The answers were graded as follows:

- Not relevant: the response is completely off-topic or unrelated to the question.
- Slightly relevant: the answer addresses the question but includes significant unrelated information.
- Moderately relevant: the response is relevant but includes some tangential content.
- Highly relevant: the answer is directly related to the question with minimal unrelated details.
- Completely relevant: the response precisely addresses the question without any deviation.

Clarity

How clear is the provided information (scale 1-5). The answers were graded as follows:

- Very unclear: the response is confusing, poorly articulated, or difficult to comprehend.
- Somewhat clear: the answer has some clarity but may require additional explanation.
- Clear: the response is understandable with a reasonable level of clarity.
- Very clear: the answer is well-explained and easy to follow.
- Exceptionally clear: the response is articulated in an exceptionally straightforward and comprehensible manner.

Beneficial

Does the response significantly aid the decision-making process (scale 1-5). The answers were graded as follows:

- Not beneficial: the response provides no useful aid for decision-making.
- Slightly beneficial: the answer offers limited assistance in the decision-making process.
- Moderately beneficial: the response is somewhat helpful for decision-making.
- Highly beneficial: the answer significantly aids the decision-making process.
- Extremely beneficial: the response is exceptionally valuable and decisively aids in making informed decisions.

Completeness

Does the response cover all necessary information required to fully answer the question (scale 1-5). The answers were graded as follows:

- Very incomplete: the response leaves out crucial information necessary to answer the question.
- Somewhat complete: the answer includes some necessary information but is missing significant aspects.
- Moderately complete: the response covers a fair amount of the necessary information.
- Mostly complete: the answer is almost complete with only minor omissions.
- Fully complete: the response comprehensively covers all information required to fully answer the question.

Statistical Analysis

To evaluate the performance differences between GPT-3.5 and GPT-4, the nonparametric Mann-Whitney *U* tests were used. The Mann-Whitney *U* test was chosen to compare GPT-3.5 and GPT-4 scores, as our data are ordinal. This nonparametric test is ideal for data not meeting normal distribution assumptions. This nonparametric test was also used to discern grading disparities between senior physicians and residents, representing 2 heterogeneous grader groups with potentially varying evaluation patterns. Furthermore, to probe into the potential differences in the responses based on the 3 distinct question categories (emergency medicine, internal medicine, and ethical questions), Mann-Whitney *U* tests were once again applied.

XSL•FO
RenderX

For all statistical tests, a *P* value of less than .05 was considered indicative of a statistically significant difference.

Results

Overview of ChatGPT Performance

An evaluation was conducted to assess the performance of 2 generative models, GPT-4 and GPT-3.5, in terms of accuracy, relevance, clarity, benefit to the decision-making process, and completeness of the response. A total of 176 questions were analyzed by both residents and seniors, yielding a consistent and objective assessment. Examples of emergency medicine, internal medicine, and ethical questions and the corresponding GPT-3.5 and GPT-4 detailed answers are presented in Table 1.

GPT-4 demonstrated statistically significant superiority in all assessed criteria compared with GPT-3.5. The analysis shows that the models differed notably in their ability to provide clear, relevant, and beneficial responses, with GPT-4 consistently outperforming GPT-3.5. The results of GPT-4 and GPT-3.5 are summarized in Table 2.

XSL•FO
RenderX

Table 1. A comparison between the responses of ChatGPT-4 and ChatGPT-3.5 to questions from 3 different disciplines: emergency medicine, internal medicine, and ethical dilemmas.

Discipline	Question	ChatGPT-4	ChatGPT-3.5
Emergency medicine	How would you manage a patient presenting with signs of necrotizing fasciitis?	• Immediately initiate broad-spectrum intravenous antibiotics including coverage for gram-positive, gram-negative, and anaerobic bacteria • Administer intravenous fluid resuscitation to maintain hemodynamic stability • Provide the patient with adequate analgesia for pain management • Consult a surgeon urgently for possible surgical debridement. Necrotizing fasciitis is a surgical emergency and tissue debridement is critical in the management • Order laboratory tests including complete blood count, electrolytes, renal function, liver function, coagulation profile, and blood cultures • Obtain imaging, such as x-ray, CT[a] scan, or magnetic resonance imaging, if the diagnosis is uncertain. However, do not delay treatment for imaging • Consider administration of IVIG[b] as adjunctive therapy. • Monitor the patient closely in an intensive care setting • Provide supportive care including nutritional support and prevention of secondary infections • Consider tetanus prophylaxis if the patient's immunization status is unknown or not up-to-date • Arrange for psychosocial support as necrotizing fasciitis can have a significant impact on mental health • Implement contact precautions to prevent the spread of infection to other patients and health care workers	• Call for emergency medical assistance immediately • Begin supportive care including oxygenation and intravenous fluids • Administer broad-spectrum antibiotics that cover both gram-positive and gram-negative bacteria • Obtain blood cultures and wound cultures • Surgical consultation should be obtained immediately • Consider early surgical intervention including debridement of affected tissue • Monitor the patient closely for signs of sepsis and organ dysfunction • Administer pain medication as needed • Educate the patient and family about the disease and potential complications

XSL•FO
RenderX

Discipline	Question	ChatGPT-4	ChatGPT-3.5
Internal medicine	How should you handle a patient with known cirrhosis presenting with new-onset ascites?	• Confirm the diagnosis: ascertain the patient's symptoms such as abdominal bloating, weight gain, and discomfort. Perform a physical examination to assess for fluid waves or shifting dullness • Order diagnostic tests: arrange for an ultrasound or CT scan to confirm ascites. A paracentesis may be performed to analyze the ascitic fluid and rule out infection or malignancy • Manage underlying cirrhosis: ensure the patient is on optimal therapy for cirrhosis, which may include medications like diuretics, beta-blockers, or lactulose • Sodium restriction: advise the patient to limit their sodium intake to less than 2 grams per day • Diuretics: start the patient on a diuretic regimen, often spironolactone with or without furosemide, to help mobilize the fluid • Monitor for complications: regularly monitor the patient for complications of cirrhosis and ascites such as hepatic encephalopathy, spontaneous bacterial peritonitis, and hepatorenal syndrome • Consider referral for liver transplantation: if the patient's condition continues to worsen despite optimal medical therapy, consider referral to a transplant center for evaluation for liver transplantation • Patient education: educate the patient about the importance of medication adherence, dietary sodium restriction, and regular follow-up, and also, inform about the signs of complications	• Obtain a detailed medical history, including the duration and severity of ascites, any recent changes in medication or diet, and any history of alcohol or drug abuse • Perform a thorough physical examination including a careful assessment of the abdomen for fluid accumulation, signs of liver disease, and any evidence of hepatic encephalopathy • Order laboratory tests including liver function tests, complete blood count, electrolytes, and renal function tests • Monitor the patient closely for signs of complications, such as infection, renal failure, or hepatic encephalopathy, and adjust treatment as needed • Educate the patient about the importance of adhering to a low-sodium diet, avoiding alcohol and other hepatotoxic substances, and following up regularly with their health care provider to monitor their condition.
Ethical questions	Should a health care provider give placebos to patients who insist on receiving treatment when no treatment is necessary?	• Ethically, health care providers should not give placebos to patients who insist on receiving treatment when no treatment is necessary • The use of placebos without patient consent is deceptive and violates the principles of patient autonomy and informed consent, which are fundamental to medical ethics • It can undermine the trust between the patient and the health care provider. If discovered, it could damage the patient's confidence in their health care provider or the health care system as a whole • It may also delay necessary treatment if the patient's condition changes or worsens, as the placebo effect is temporary and does not treat underlying conditions • However, some argue that if a placebo can alleviate patient anxiety or perceived symptoms, it may be considered in the patient's best interest. This approach should only be taken with full transparency and patient consent • In conclusion, it is generally considered unethical to administer placebos without patient knowledge or consent, and it is crucial to respect patient autonomy, promote open communication, and provide appropriate care based on medical necessity	• No, health care providers should not give placebos to patients who insist on receiving treatment when no treatment is necessary • Giving placebos is unethical and goes against the principle of informed consent. • It can also harm the patient's trust in health care providers and the medical profession as a whole • Health care providers should educate patients on the lack of necessity for treatment and offer alternative forms of support such as counseling or therapy • If a patient continues to insist on treatment, it is important to have an open and honest conversation about the potential risks and benefits of treatment and the lack of evidence supporting the use of placebos

[a]CT: computed tomography.

[b]IVIG: intravenous immunoglobulin.

XSL•FO
RenderX

Table 2. Comparative analysis of GPT-4 and GPT-3.5 performance.

Criteria	GPT-4, mean (SD)	GPT-3.5, mean (SD)	*P* value
Accuracy	4.5 (0.8)	4.2 (0.9)	<.001
Relevance	4.5 (0.7)	4.4 (0.8)	<.001
Clarity	4.6 (0.7)	4.4 (0.9)	<.001
Beneficial	4.3 (0.9)	3.9 (1.0)	<.001
Completeness	4.3 (0.9)	3.7 (1.1)	<.001
Total	4.4 (0.8)	4.1 (1.0)	<.001

Distribution of Ratings for GPT-4 and GPT-3.5

GPT-4 consistently scored between 4.29 and 4.55, excelling in clarity with an average of 4.55, indicating clear and direct information delivery. In contrast, GPT-3.5 scores ranged from 3.92 to 4.37, with its lowest in beneficiality at 3.92, reflecting variability in aiding decision-making.

For all the rating dimensions, GPT-4 generally received higher ratings than its counterpart, GPT-3.5. The mode—the most frequently given rating—for GPT-4 consistently achieved the maximum score of 5 across all categories, except for the benefit category, where the mode was approximately 4.5.

In contrast, the distribution of ratings for GPT-3.5 presented a wider spread, with modes fluctuating between scores of 3 and 5, contingent on the category. In particular, the categories of benefit and completeness exhibited a broad spread of ratings for GPT-3.5, indicative of greater variability in the responses.

Comparison Between Residents and Seniors

In evaluating the models' responses based on the reviewer type (Table 3), significant differences were discerned between residents and senior physicians.

Table 3. Comparison between residents' and seniors' assessments of GPT-4 and GPT-3.5 answers.

Criteria	Residents GPT-4, mean (SD)	Seniors GPT-4, mean (SD)	*P* value GPT-4	Residents GPT-3.5, mean (SD)	Seniors GPT-3.5, mean (SD)	*P* value GPT-3.5
Accuracy	4.3 (0.9)	4.7 (0.5)	<.001	4.1 (1.0)	4.3 (0.7)	<.001
Relevance	4.3 (0.8)	4.7 (0.6)	<.001	4.2 (0.9)	4.6 (0.8)	<.001
Clarity	4.4 (0.7)	4.7 (0.6)	<.001	4.2 (0.9)	4.5 (0.9)	<.001
Beneficial	4.0 (1.0)	4.6 (0.7)	<.001	3.7 (1.0)	4.1 (1.0)	<.001
Completeness	4.1 (1.0)	4.6 (0.6)	<.001	3.6 (1.2)	3.9 (0.9)	<.001

For both GPT-4 and GPT-3.5, seniors consistently rated the responses higher across all criteria. The variation was most notable in the assessment of the model's benefits and completeness.

In the case of GPT-4, senior physicians rated it as more beneficial (4.6 vs 4.0, $P<.001$) and more complete (4.6 vs 4.1, $P<.001$) compared with the residents. A similar trend was observed for GPT-3.5, with seniors appreciating its benefits (4.1 vs 3.7, $P<.001$) and completeness (3.9 vs 3.5, $P<.001$) more than the residents. This finding suggests a potential influence of clinical experience on the perception of AI-generated responses.

Performance Across Question Types

In the evaluation of both GPT-4 and GPT-3.5, distinctions were evident across question types (emergency medicine, internal medicine, and ethical questions). Ethical queries consistently received the highest ratings (Figures 1 and 2 present analysis for GPT-4 and GPT-3.5 across all reviewers, while Tables 4 and 5 present analysis separately for residents and seniors). Higher grades were given for ethical answers by residents and seniors alike and in both models of GPT (Tables 4 and 5).

XSL•FO
RenderX

Figure 1. Performance of GPT-3.5 according to question subject.

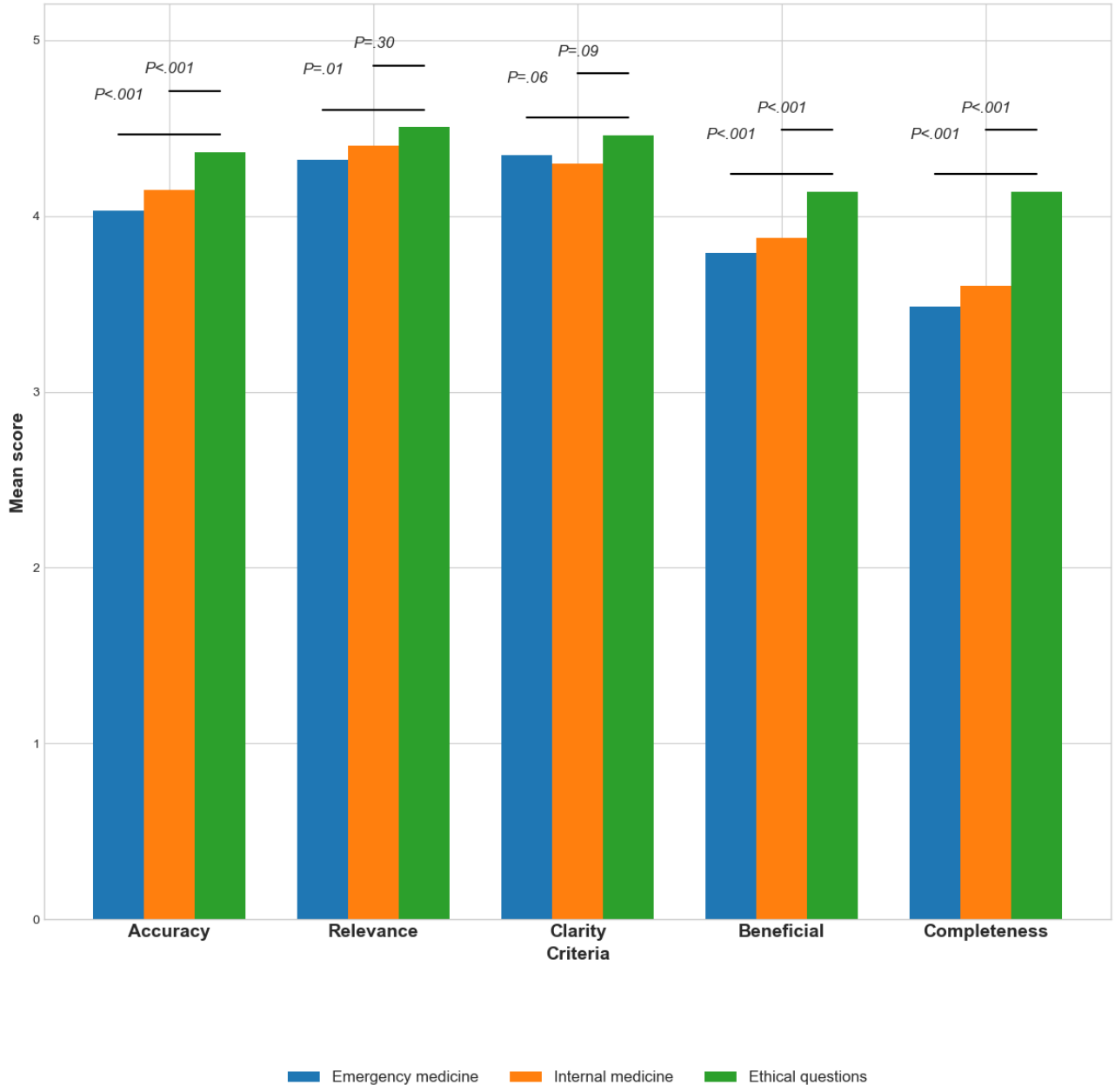

XSL•FO
RenderX

Figure 2. Performance of GPT-4 according to question subject.

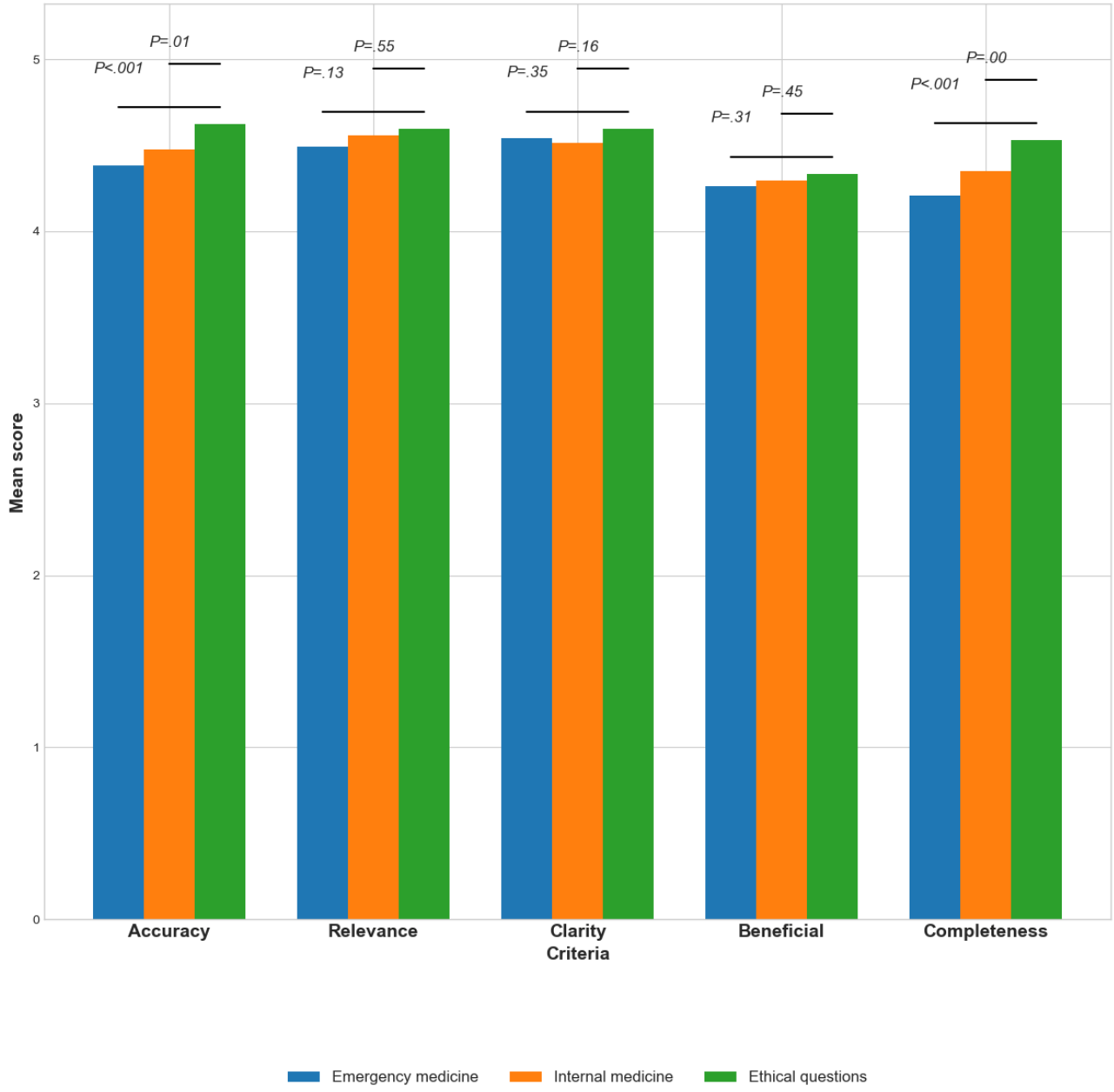

XSL•FO
RenderX

Table 4. Comparison between residents' and seniors' assessment of medical and ethical GPT-4 answers.

Criteria	Medical GPT-4, mean (SD)	Ethical GPT-4, mean (SD)	*P* value
Residents' assessment			
Accuracy	4.2 (0.9)	4.4 (0.8)	.17
Relevance	4.3 (0.8)	4.4 (0.7)	.52
Clarity	4.4 (0.7)	4.5 (0.7)	.17
Beneficial	3.9 (1.0)	4.1 (0.9)	.03
Completeness	4.0 (1.0)	4.3 (0.8)	<.001
Total	4.2 (0.9)	4.3 (0.8)	<.001
Seniors' assessment			
Accuracy	4.6 (0.6)	4.9 (0.3)	<.001
Relevance	4.7 (0.6)	4.8 (0.5)	.14
Clarity	4.7 (0.7)	4.7 (0.6)	.85
Beneficial	4.6 (0.7)	4.5 (0.8)	.63
Completeness	4.5 (0.7)	4.8 (0.5)	<.001
Total	4.6 (0.7)	4.7 (0.6)	<.001

Table 5. Comparison between residents' and seniors' assessment of medical and ethical GPT-3.5 answers.

Criteria	Medical GPT-3.5, mean (SD)	Ethical GPT-3.5, mean (SD)	*P* value
Residents' assessment			
Accuracy	4.0 (1.0)	4.2 (0.9)	.08
Relevance	4.1 (0.9)	4.3 (0.8)	.004
Clarity	4.2 (0.9)	4.3 (0.8)	.11
Beneficial	3.6 (1.0)	3.9 (1.0)	<.001
Completeness	3.4 (1.2)	3.9 (1.0)	<.001
Total	3.9 (1.0)	4.1 (0.9)	<.001
Seniors' assessment			
Accuracy	4.1 (0.8)	4.6 (0.5)	<.001
Relevance	4.6 (0.9)	4.7 (0.6)	.70
Clarity	4.5 (1.0)	4.6 (0.8)	.16
Beneficial	4.0 (1.0)	4.4 (1.0)	<.001
Completeness	3.7 (0.9)	4.3 (0.7)	<.001
Total	4.2 (1.0)	4.5 (0.8)	<.001

For GPT-4, mean scores were fairly consistent across the criteria of accuracy, and completeness, with significant differences among the question types (Figure 1). Specifically, completeness revealed statistically significant differences among emergency, internal, and ethical questions (4.2, SD 1.0; 4.3, SD 0.8; and 4.5, SD 0.7; *P*<.001).

For GPT-3.5, differences among question types were often statistically significant, as demonstrated in Figure 2. Accuracy ranged from 4.0 (SD 0.9) for ED questions to 4.4 (SD 0.8) for ethical questions (*P*<.001). Similar trends were observed in beneficial (*P*<.001) and completeness (*P*<.001).

Discussion

Principal Findings

In our study, we evaluated ChatGPT-3.5 and ChatGPT-4's ability to address complex clinical and ethical dilemmas. Both models showed promise as tools to aid physicians in decision-making. Notably, ChatGPT-4 outperformed ChatGPT-3.5 across all parameters. This was validated by both resident and senior medical practitioners.

In an effort to provide a thorough, wide-ranging, and accurate evaluation, our study included an extensive compilation of queries, totaling 176 questions. These questions spanned various topics found within the domains of emergency medicine, internal

XSL•FO
RenderX

medicine, and ethical considerations, mirroring the real-life challenges that doctors frequently face in their daily clinical work. Both groups of evaluators—senior physicians and interns—assessed each question, aiming to capture diverse perspectives on the perceived benefits at different levels of expertise. To the best of our knowledge, this is the most exhaustive study conducted on this particular subject.

Overall, both ChatGPT models received high grades in terms of accuracy, relevance, clarity, benefit, and completeness. However, GPT-4 scored higher in all criteria assessed including total mean grades (Table 2).

These findings are promising when considering the potential benefits of implementing an NLP model, like ChatGPT, into the field of medicine, and in agreement with this literature [10-15]. The visible advantages are manifold, with one of the standout features being the rapid retrieval of information and examination of the literature. ChatGPT's strength lies in its capacity to quickly access a wide array of medical data from various sources. By offering doctors immediate entry to the newest findings, clinical standards, and specific cases, ChatGPT acts as a catalyst for keeping them aligned with the ever-changing medical landscape. This ability enhances physicians' capacity to make educated judgments when dealing with intricate or uncommon medical scenarios.

Additionally, ChatGPT's aptitude in understanding natural language equips it to thoroughly examine patient symptoms and medical backgrounds. This can lead to the suggestion of possible diagnoses, as well as offering alternative diagnoses for reflection. While it does not substitute for hands-on clinical expertise, ChatGPT proves to be an invaluable asset in assisting physicians to pinpoint diagnostic avenues and contemplate less evident conditions [16].

Furthermore, ChatGPT's skill in scrutinizing medical publications and clinical studies empowers it to put forth suitable treatment recommendations rooted in the most up-to-date scientific proof. Doctors can leverage ChatGPT to weigh different treatment approaches, potential adverse effects, and counterindications, thus paving the way for customized and well-informed therapeutic choices [23].

The superior performance grades attributed to ChatGPT-4 align with expectations and are consistent with prior research comparing these 2 models within the health care domain [19,24-26]. This highlights the enhanced language model that ChatGPT-4 possesses, having benefitted from a broader and more varied data set during its development phase. Such improvements allow ChatGPT-4 to detect more complex linguistic patterns, thereby improving its ability to comprehend and generate contextually relevant responses. Furthermore, the inclusion of a more substantial corpus of medical literature, scientific articles, and ethical guidelines equips ChatGPT-4 with a wider base of knowledge. This rich repository of information empowers the model to provide more comprehensive and nuanced answers when faced with medical inquiries and ethical dilemmas.

Our study also explored an intriguing research question related to ChatGPT's role in supporting decision-making in the nuanced area of ethical dilemmas. Since ethical considerations are a fundamental part of medical practice, they often create intricate scenarios for medical professionals. ChatGPT has the potential to act as a significant tool in this aspect, providing insights into ethical principles and the resolutions of past cases. While the final ethical determinations are the responsibility of the physician, ChatGPT's assistance can guide them through multifaceted ethical conundrums, thereby enhancing the focus on patient-centered care. Hence, our data reveals that questions related to ethics consistently garnered the highest evaluations (as seen in Table 4 and Figures 1 and 2). These observations emphasize ChatGPT's impressive ability to grapple with ethical dilemmas, regardless of their innate complexity. The model's performance in this vital area is praiseworthy, even outperforming its evaluations in responding to information-based questions.

Interestingly, for both GPT-4 and GPT-3.5, senior physicians consistently rated the responses higher across all criteria. The variation was most notable in the assessment of the model's benefits and completeness. These differences might be explained by the combination of higher experience with complexity, higher familiarity with research and guidelines, higher critical analysis skills, more experience navigating ethical dilemmas, and a broader interdisciplinary perspective possessed by senior physicians that likely contributes to their higher ratings of ChatGPT's responses. However, further research is warranted to confirm our results.

Limitations

Our study had several limitations. First, the expert panel that formulated the questions was composed of only 4 experts, while the group that assessed the questions included 8 physicians. Although the evaluating group was quite diverse, the findings might not accurately reflect the views of the broader physician community within these specialties. This small, potentially nonrepresentative sample could lead to biases in focus areas, subjective interpretations, and variability in expertise. Subjectivity in assessing responses, especially in relevance and clarity, combined with individual preconceptions about AI's capabilities, might skew results. Expanding the expert pool and incorporating diverse perspectives from different subspecialties and health care settings, along with methodologies to adjust for individual biases, could mitigate these limitations in future studies.

Second, the assessment of ChatGPT's efficacy relied on subjective judgments from 2 groups of physicians, which could introduce bias and inconsistency. Nonetheless, medical inquiries often encompass intricate matters that defy simple quantification, and the study intended to gauge its applicability to everyday medical work. Hence, personal assessment plays a vital role in this examination, ensuring that ChatGPT's responses are pertinent, lucid, evidence-supported, legitimate, and worthwhile.

Third, this study concentrated solely on ChatGPT's capability to respond to questions within specific subdisciplines of internal medicine, emergency medicine (ER), and ethics, leaving its potential in other medical fields unexplored. Additional studies are required to scrutinize ChatGPT's performance across a

XSL•FO
RenderX

broader spectrum of medical areas. Fourth, our examination was restricted to both versions of ChatGPT, raising the possibility that the findings might have differed with an alternate language model. Further research is needed to ascertain how applicable our conclusions might be to other large language models and varying contexts.

Conclusion

In conclusion, the potential of ChatGPT in aiding physicians in addressing common medical problems is promising. As technology continues to advance, integrating ChatGPT into clinical workflows may become a valuable asset, enhancing diagnostic accuracy, treatment decisions, and ethical considerations. Nevertheless, it is essential to acknowledge that ChatGPT's role should complement rather than replace clinical expertise and human judgment. As the technology evolves, further research and validation studies are warranted to optimize its abilities, ensuring safe and effective use in clinical settings.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

AL, KS, and EK designed and performed the research and wrote the paper; NZ, YSP, YS, LF, US, MA, and RL answered the questions, contributed to the analysis, and revised the paper critically. All authors approved the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original Questions and Responses Generated by GPT-3.5 and GPT-4.
[XLSX File (Microsoft Excel File), 137 KB - jmir_v26i1e54571_app1.xlsx]

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
2. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
3. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online on May 28, 2020.
4. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog 2019;1(8):9.
5. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. USA: Advanced Analytics, LLC; 2019.
6. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci Rep 2023 Mar 13;13(1):4164. [doi: 10.1038/s41598-023-31412-2] [Medline: 36914821]
7. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. Hepatol Commun 2023;7(4):e0097 [FREE Full text] [doi: 10.1097/HC9.0000000000000097] [Medline: 36972383]
8. Lahat A, Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? J Telemed Telecare 2023:1357633X231155520. [doi: 10.1177/1357633X231155520] [Medline: 36760131]
9. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
10. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
11. Rasmussen MLR, Larsen A, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefes Arch Clin Exp Ophthalmol 2023;261(10):3041-3043. [doi: 10.1007/s00417-023-06078-1] [Medline: 37129631]
12. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33(6):1790-1796 [FREE Full text] [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]

XSL•FO
RenderX

13. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg 2023;47(5):1985-1993 [FREE Full text] [doi: 10.1007/s00266-023-03338-7] [Medline: 37095384]
14. Yeo YH, Samaan JS, Ng WH, Ting P, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
15. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023;7(2):pkad015 [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
17. Voss C, Surendran A, LaValle S. OpenAI GPT-4: supercharging AI's language understanding. OpenAI blog 2023.
18. Liévin V, Hother C, Winther O. Can large language models reason about medical questions? arXiv Preprint posted online on July 17, 2022.
19. Nori H, King N, McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv Preprint posted online on March 20, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.13375]
20. Chow JCL, Wong V, Sanders L, Li K. Developing an AI-assisted educational chatbot for radiotherapy using the IBM Watson assistant platform. Healthcare (Basel) 2023;11(17):2417 [FREE Full text] [doi: 10.3390/healthcare11172417] [Medline: 37685452]
21. Rebelo N, Sanders L, Li K, Chow JCL. Learning the treatment process in radiotherapy using an artificial intelligence-assisted chatbot: development study. JMIR Form Res 2022;6(12):e39443 [FREE Full text] [doi: 10.2196/39443] [Medline: 36327383]
22. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6:1166014 [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
23. Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol 2024;281(4):2081-2086. [doi: 10.1007/s00405-023-08104-8] [Medline: 37405455]
24. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
25. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023;15(6):e40822 [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
26. He N, Yan Y, Wu Z, Cheng Y, Liu F, Li X, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. J Telemed Telecare 2023:1357633X231181922. [doi: 10.1177/1357633X231181922] [Medline: 37350055]

Abbreviations

AI: artificial intelligence
ER: emergency room
NLP: natural language processing

Edited by T de Azevedo Cardoso; submitted 15.11.23; peer-reviewed by F Liu, B Oronti; comments to author 26.01.24; revised version received 02.02.24; accepted 29.04.24; published 27.06.24.

Please cite as:
Lahat A, Sharif K, Zoabi N, Shneor Patt Y, Sharif Y, Fisher L, Shani U, Arow M, Levin R, Klang E
Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4
J Med Internet Res 2024;26:e54571
URL: https://www.jmir.org/2024/1/e54571
doi:10.2196/54571
PMID:

©Adi Lahat, Kassem Sharif, Narmin Zoabi, Yonatan Shneor Patt, Yousra Sharif, Lior Fisher, Uria Shani, Mohamad Arow, Roni Levin, Eyal Klang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence–Based Electrocardiographic Biomarker for Outcome Prediction in Patients With Acute Heart Failure: Prospective Cohort Study

Youngjin Cho[1,2*], MD, PhD; Minjae Yoon[1*], MD; Joonghee Kim[2,3], MD, PhD; Ji Hyun Lee[1], MD, PhD; Il-Young Oh[1], MD, PhD; Chan Joo Lee[4], MD, PhD; Seok-Min Kang[4], MD, PhD; Dong-Ju Choi[1], MD, PhD

[1]Division of Cardiology, Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Gyeonggi-do, Republic of Korea

[2]ARPI Inc, Seongnam, Gyeonggi-do, Republic of Korea

[3]Department of Emergency Medicine, Seoul National University Bundang Hospital, Seongnam, Gyeonggi-do, Republic of Korea

[4]Division of Cardiology, Department of Internal Medicine, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Dong-Ju Choi, MD, PhD
Division of Cardiology, Department of Internal Medicine
Seoul National University Bundang Hospital
Seoul National University College of Medicine
82 Gumi-ro 173 Beon-gil
Bundang-gu
Seongnam, Gyeonggi-do, 13620
Republic of Korea
Phone: 82 317877007
Fax: 82 317877041
Email: djchoi@snubh.org

Abstract

Background: Although several biomarkers exist for patients with heart failure (HF), their use in routine clinical practice is often constrained by high costs and limited availability.

Objective: We examined the utility of an artificial intelligence (AI) algorithm that analyzes printed electrocardiograms (ECGs) for outcome prediction in patients with acute HF.

Methods: We retrospectively analyzed prospectively collected data of patients with acute HF at two tertiary centers in Korea. Baseline ECGs were analyzed using a deep-learning system called Quantitative ECG (QCG), which was trained to detect several urgent clinical conditions, including shock, cardiac arrest, and reduced left ventricular ejection fraction (LVEF).

Results: Among the 1254 patients enrolled, in-hospital cardiac death occurred in 53 (4.2%) patients, and the QCG score for critical events (QCG-Critical) was significantly higher in these patients than in survivors (mean 0.57, SD 0.23 vs mean 0.29, SD 0.20; $P<.001$). The QCG-Critical score was an independent predictor of in-hospital cardiac death after adjustment for age, sex, comorbidities, HF etiology/type, atrial fibrillation, and QRS widening (adjusted odds ratio [OR] 1.68, 95% CI 1.47-1.92 per 0.1 increase; $P<.001$), and remained a significant predictor after additional adjustments for echocardiographic LVEF and N-terminal prohormone of brain natriuretic peptide level (adjusted OR 1.59, 95% CI 1.36-1.87 per 0.1 increase; $P<.001$). During long-term follow-up, patients with higher QCG-Critical scores (>0.5) had higher mortality rates than those with low QCG-Critical scores (<0.25) (adjusted hazard ratio 2.69, 95% CI 2.14-3.38; $P<.001$).

Conclusions: Predicting outcomes in patients with acute HF using the QCG-Critical score is feasible, indicating that this AI-based ECG score may be a novel biomarker for these patients.

Trial Registration: ClinicalTrials.gov NCT01389843; https://clinicaltrials.gov/study/NCT01389843

(J Med Internet Res 2024;26:e52139) doi:10.2196/52139

XSL•FO
RenderX

KEYWORDS

acute heart failure; electrocardiography; artificial intelligence; deep learning

Introduction

Heart failure (HF) is a major global health problem affecting millions of people worldwide, leading to significant morbidity, mortality, and health care expenditure [1-3]. Although several valuable biomarkers such as N-terminal prohormone of brain natriuretic peptide (NT-proBNP) [4,5] and cardiac troponins [6] have been introduced for patients with HF, their use in routine clinical practice is often constrained by their cost and limited availability.

Electrocardiogram (ECG) is an essential and cost-effective tool for evaluating cardiovascular diseases. ECG is widely available, noninvasive, and provides real-time information about cardiac electrical activity, which is crucial for detecting arrhythmias, ischemia, and other cardiac abnormalities. With advances in artificial intelligence (AI) and deep learning, there has been growing interest in employing AI algorithms to analyze ECG data and predict outcomes in patients with various cardiovascular conditions [7,8].

In this study, we investigated the utility of an AI algorithm that analyzes printed ECG images for outcome prediction in patients with acute HF. These findings will demonstrate the potential of AI-assisted ECG analysis for predicting outcomes in these patients, potentially overcoming the cost and availability constraints of current biomarkers.

Methods

Study Population

This was a substudy of the prospective multicenter Korean Acute Heart Failure (KorAHF) registry, which enrolled 5625 consecutive patients upon initial hospital admission for acute HF at 10 tertiary university hospitals in Korea. Details on the KorAHF registry objectives, design, and population are available on the clinical trial registration site (ClinicalTrials.gov NCT01389843) and have been published previously [9,10]. Briefly, patients who had signs or symptoms of HF and met one of the following criteria were eligible for enrollment in the KorAHF registry: (1) lung congestion or (2) objective left ventricular systolic dysfunction or structural heart disease findings. There were no exclusion criteria.

In this study, we retrospectively analyzed the prospectively collected data from 1254 patients who were hospitalized for acute HF from March 2011 to February 2014 at 2 out of 10 participating tertiary centers (Seoul National University Bundang Hospital and Severance Hospital) using the KorAHF registry (Figure 1). Additional ECG image data were collected for this study.

Figure 1. Flowchart of the training and validation study populations. The AI-ECG analyzer, named Quantitative ECG (QCG), was developed using 47,194 annotated ECG images of over 32,968 patients who visited the emergency department of SNUBH between 2017 and 2019. The QCG analyzer was applied to ECGs from a subpopulation of the KorAHF registry including patients with AHF enrolled between 2011 and 2014. *Internal validation results for two QCG scores. AHF: acute heart failure; AI: artificial intelligence; AUC: area under the curve; ECG: electrocardiogram; ED: emergency department; KorAHF: Korean Acute Heart Failure; LV: left ventricular; SNUBH: Seoul National University Bundang Hospital.

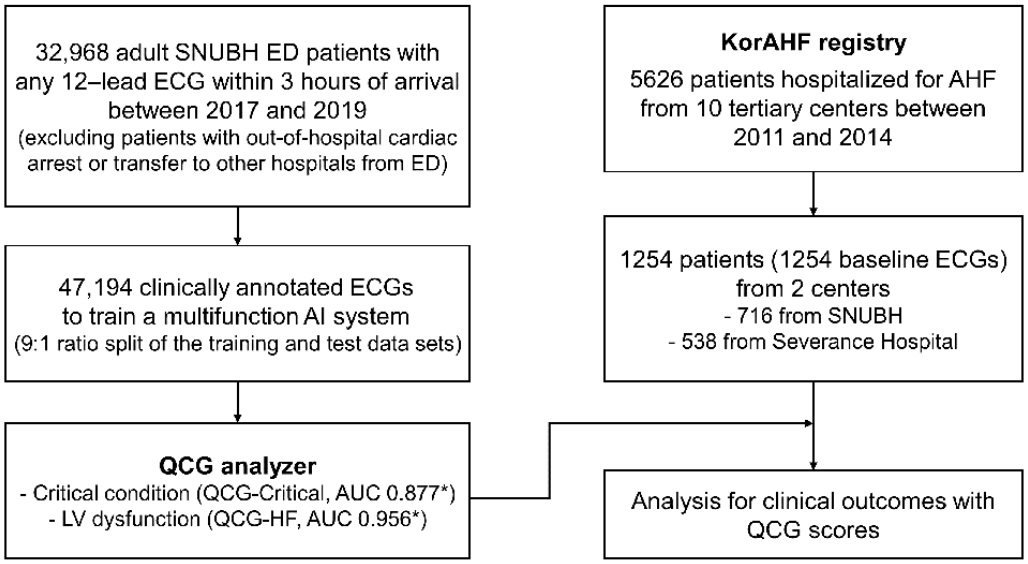

XSL•FO
RenderX

Ethical Considerations

This study conformed with the principles outlined in the Declaration of Helsinki. The study protocol was approved by the institutional review board at Seoul National University Bundang Hospital (No. B-1104-125-014) and Severance Hospital (No. 2022-2166-001). The need for written informed consent was waived by the institutional review board. Our research strictly adheres to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [11].

Clinical Follow-Up and Endpoints

Data collection methods have been previously described [9]. Briefly, data on patients' clinical manifestations, biochemical parameters, medication, and outcome were collected using a web-based case report form for up to 60 months by research nurses. Outcome data on patients lost to follow-up were additionally collected from national death records.

The primary endpoint of this study was all-cause mortality. Secondary outcomes included in-hospital outcomes, especially in-hospital mortality. All deaths were considered to be cardiac-related unless a definite noncardiac cause could be established. All outcome data reported from the participating centers were reviewed by an independent clinical event adjudicating committee.

AI Algorithm

Quantitative ECG (QCG) is an AI analyzer composed of an encoder part and multiple task-specific networks. The encoder part is a modified convolutional neural network with residual connections, squeeze excitation modules, and a nonlocal block. The task-specific networks are multilayer percetron models. The encoder part accepts 2D ECG images as input to produce a common numerical feature vector for downstream tasks. The encoder part was pretrained on 49,731 open ECGs using self-supervised learning schemes and then fine-tuned on 47,194 annotated ECG images of over 32,968 patients who visited the Emergency Department of Seoul National University Bundang Hospital between 2017 and 2019 using multitask learning schemes. The tasks include the classification of 12 rhythms (with 35 subtypes) and production of 10 digital biomarkers correlated with the risk of (1) being critically ill (shock, respiratory failure, or cardiac arrest), (2) cardiac ischemia (acute coronary syndrome, ST-elevation myocardial infarction, or myocardial injury as defined by an elevated troponin level), (3) cardiac dysfunction (pulmonary edema, left and right heart dysfunction, pulmonary hypertension, and clinically significant pericardial effusion), and (4) hyperkalemia. Several validation studies of the system have been published previously [12-14]. The collection of these AI algorithms has been developed into a mobile app (ECG Buddy, ARPI), which has been approved by the Korean Ministry of Food and Drug Safety.

In this study, two QCG features were evaluated: QCG-Critical for critical conditions such as shock or mortality and QCG-HF for a reduced echocardiographic left ventricular ejection fraction (LVEF) of <40%. The QCG scores, representing probability, ranged from 0 to 1.0, with 0 indicating low and 1.0 indicating high probability. With a 9:1 ratio split of the training and test data sets, the internal validation results for these two QCG features showed an area under the curve (AUC) of 0.877 for QCG-Critical and 0.956 for QCG-HF. The composition of the training and validation data sets is presented as a flowchart in Figure 1.

Statistical Analysis

Categorical variables are reported as frequencies (percentages) and continuous variables are expressed as means (SD) or medians (IQR). The two key AI-driven scores (QCG-Critical and QCG-HF) were analyzed as continuous variables. The Student *t* test and χ^2 (or Fisher exact) test were used to compare the baseline clinical characteristics between the two groups. The discrimination performance of QCG scores for in-hospital outcomes was evaluated using receiver operating characteristic (ROC) curve analysis. The AUC values were compared using the DeLong test. The logistic regression model was used to estimate the odds ratios (ORs) and 95% CIs. Survival analysis was performed using the Kaplan-Meier method, and the Cox proportional hazard model was used to estimate the hazard ratios (HRs) and 95% CIs for the clinical outcomes. Multivariable analysis was performed with the inclusion of clinically relevant variables.

All tests were two-tailed and a *P* value <.05 was considered statistically significant. Statistical analyses were performed using R programming version 4.3.0 (The R Foundation for Statistical Computing).

Results

Baseline Characteristics

Data of 1254 patients (716 from Seoul National University Bundang Hospital and 538 from Severance Hospital) were analyzed. Among the 1254 patients, 53 (4.2%) experienced in-hospital cardiac death. The baseline characteristics of the study population according to the in-hospital outcomes are shown in Table 1. Compared with survivors, patients who died in the hospital were older, had a higher prevalence of ischemic heart disease, lower LVEF, and higher NT-proBNP levels. By contrast, atrial fibrillation (AF) was more frequent in survivors. The QCG-Critical and QCG-HF scores were significantly higher in patients who experienced in-hospital cardiac death than in survivors (*P*<.001) (Table 1 and Figure S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Baseline characteristics.

Characteristics	Total (n=1254)	In-hospital cardiac death (n=53)	Survivors (n=1201)	*P* value
Age (years), mean (SD)	69.8 (14.7)	74.0 (14.5)	69.6 (14.1)	.03
Male, n (%)	673 (53.7)	29 (54.7)	644 (53.6)	.99
Hypertension, n (%)	843 (67.2)	31 (58.5)	812 (67.6)	.22
Diabetes mellitus, n (%)	499 (39.8)	23 (43.4)	476 (39.6)	.69
Cerebrovascular disease, n (%)	224 (17.9)	7 (13.2)	217 (18.1)	.47
Chronic kidney disease, n (%)	212 (28.3)	11 (20.8)	212 (28.3)	.69
Ischemic heart disease, n (%)	365 (29.1)	25 (47.2)	340 (28.3)	.005
Valvular heart disease, n (%)	217 (17.3)	9 (17.0)	208 (17.3)	>.99
De novo HF[a], n (%)	612 (48.8)	29 (54.7)	583 (48.5)	.46
Atrial fibrillation, n (%)	417 (34.7)	10 (10.9)	417 (34.7)	.03
QRS duration≥120 ms, n (%)	318 (25.4)	17 (32.1)	301 (25.1)	.32
LVEF[b] (%), mean (SD)	35.3 (14.7)	28.5 (11.9)	35.6 (14.7)	.002
NT-proBNP[c] (pg/mL), mean (SD)	10,373 (11,915)	17,035 (1900)	10,092 (11,879)	<.001
QCG[d] scores, mean (SD)				
QCG-Critical	0.30 (0.21)	0.57 (0.23)	0.29 (0.20)	<.001
QCG-HF	0.65 (0.31)	0.78 (0.18)	0.64 (0.31)	<.001

[a]HF: heart failure.

[b]LVEF: left ventricular ejection fraction.

[c]NT-proBNP: N-terminal prohormone of brain natriuretic peptide.

[d]QCG: Quantitative electrocardiogram artificial intelligence system.

Predictors of In-Hospital Cardiac Death

In the univariable logistic regression analysis, the QCG-Critical and QCG-HF scores were significant predictors of in-hospital cardiac death (Table 2). Other than QCG scores, echocardiographic LVEF, NT-proBNP level, age, ischemic heart disease, and AF were significantly correlated with in-hospital cardiac death (Table 2).

XSL•FO
RenderX

Table 2. Predictors of in-hospital cardiac death.

Variables	Univariate analyses		Model 1[a,b]		Model 2[b,c]	
	OR[d] (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
QCG[e] parameters (per 0.1 increase)						
QCG-Critical	1.66 (1.47-1.87)	<.001	1.68 (1.47-1.92)	<.001	1.59 (1.36-1.87)	<.001
QCG-HF[f]	1.21 (1.08-1.37)	.001	1.22 (1.08-1.39)	.002	1.02 (0.84-1.24)	.82
LVEF[g] (per 5% decrease)	1.21 (1.07-1.37)	.002	1.26 (1.09-1.45)	.001	*1.29 (1.10-1.51)*	*.02*
NT-proBNP[h] (per 1000 pg/ml increase)	1.03 (1.01-1.05)	<.001	1.04 (1.01-1.06)	<.001	*1.03 (1.00-1.05)*	*.002*
Clinical and demographic factors as covariates						
Age	1.03 (1.00-1.05)	.03	*1.03 (1.00-1.06)*	*.03*	*1.04 (1.01-1.08)*	*.009*
Male	1.05 (0.60-1.82)	.88	*1.21 (0.66-2.23)*	*.53*	*1.00 (0.48-2.07)*	*.99*
Hypertension	0.68 (0.39-1.18)	.17	*0.51 (0.27-0.99)*	*.05*	*0.42 (0.20-0.90)*	*.03*
Diabetes mellitus	1.17 (0.67-2.03)	.58	*0.72 (0.38-1.36)*	*.31*	*0.56 (0.26-1.19)*	*.13*
Chronic kidney disease	1.22 (0.62-2.41)	.56	*1.08 (0.49-2.34)*	*.85*	*1.06 (0.41-2.74)*	*.90*
Cerebrovascular disease	0.69 (0.31-1.55)	.37	*0.64 (0.26-1.53)*	*.31*	*0.54 (0.18-1.68)*	*.29*
Ischemic heart disease	2.26 (1.30-3.93)	.004	*3.00 (1.51-5.97)*	*.002*	*2.54 (1.12-5.76)*	*.03*
Valvular heart disease	0.98 (0.47-2.03)	.95	*2.00 (0.84-4.75)*	*.12*	*1.94 (0.68-5.50)*	*.21*
ADHF[i] (vs de novo)	0.78 (0.45-1.36)	.38	*0.54 (0.26-1.11)*	*.10*	*0.51 (0.22-1.19)*	*.12*
Atrial fibrillation	0.44 (0.22-0.88)	.02	*0.73 (0.34-1.58)*	*.42*	*0.80 (0.32-2.00)*	*.63*
QRS duration>120 ms	1.41 (0.78-2.55)	.25	*0.94 (0.48-1.83)*	*.85*	*1.23 (0.57-2.64)*	*.60*

[a]Adjusted for age, sex, hypertension, diabetes, chronic kidney disease, cerebrovascular disease, ischemic heart disease, valvular heart disease, heart failure type, atrial fibrillation, and QRS duration.

[b]When a variable was included as a covariate for adjustment, it was not adjusted for itself and QCG-Critical was added to the adjustment model (presented in italics).

[c]Adjusted for the same covariates as model 1 and further adjusted for left ventricular ejection fraction and N-terminal prohormone of brain natriuretic peptide.

[d]OR: odds ratio.

[e]QCG: Quantitative electrocardiogram.

[f]HF: heart failure.

[g]LVEF: left ventricular ejection fraction.

[h]NT-proBNP: N-terminal prohormone of brain natriuretic peptide.

[i]ADHF: acute decompensated heart failure.

After adjustment for age, sex, hypertension, diabetes, chronic kidney disease, cerebrovascular disease, ischemic heart disease, valvular heart disease, HF type, AF, and QRS duration, the two QCG scores remained significant predictors of in-hospital cardiac death. Moreover, the QCG-Critical score was an independent predictor of in-hospital cardiac death after further adjustment for echocardiographic LVEF and NT-proBNP level (OR 1.59, 95% CI 1.36-1.87; *P*<.001).

In a subgroup analysis, the QCG-Critical score was a significant predictor of in-hospital cardiac death regardless of the initial rhythm (AF or sinus rhythm), QRS width (wide or narrow), hypertension, diabetes, HF etiology (ischemic or nonischemic), HF type (de novo or acute decompensated HF), and LVEF (HF with reduced ejection fraction vs HF with preserved or mildly reduced ejection fraction), after adjustment for other clinical parameters (Figure 2).

XSL•FO
RenderX

Figure 2. Subgroup analysis results for predicting in-hospital cardiac death. Adjusted ORs are presented for a 0.1 increase in the QCG-Critical score *Adjusted for age, sex, hypertension, diabetes, chronic kidney disease, cerebrovascular disease, ischemic heart disease, valvular heart disease, HF type, atrial fibrillation, and QRS duration (QRSd). ADHF: acute decompensated heart failure; CMP: cardiomyopathy; HF: heart failure; LVEF: left ventricular ejection fraction; OR, odds ratio.

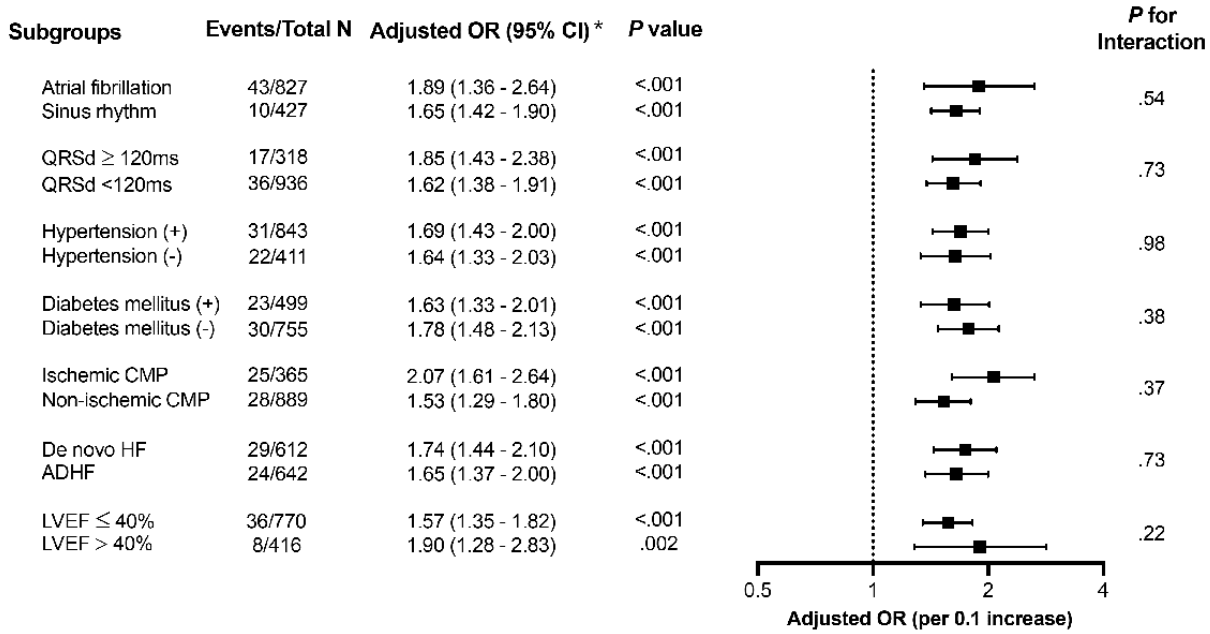

QCG-Critical Score and In-Hospital Cardiac Death

The QCG-Critical score was significantly higher in patients who experienced cardiac death within 1 day, 2 days, or during hospitalization than in survivors (Figure 3A). When the performance of the QCG-Critical score for predicting these events was analyzed using ROC curves, the AUC values for 1- and 2-day mortality and in-hospital cardiac death were 0.936, 0.917, and 0.821, respectively (Figure 3B).

Figure 3. Performance of the QCG-Critical score for predicting in-hospital cardiac death (IHCD). (A) The QCG-Critical score was significantly higher in patients who experienced cardiac death within 1 day (D1), 2 days (D2), and during hospitalization than in survivors. The box-and-whisker plot is presented with 5th to 95th percentiles. (B) Performance of the QCG-Critical score presented as receiver operating characteristic curves. AUC: area under the curve; QCG: Quantitative electrocardiogram.

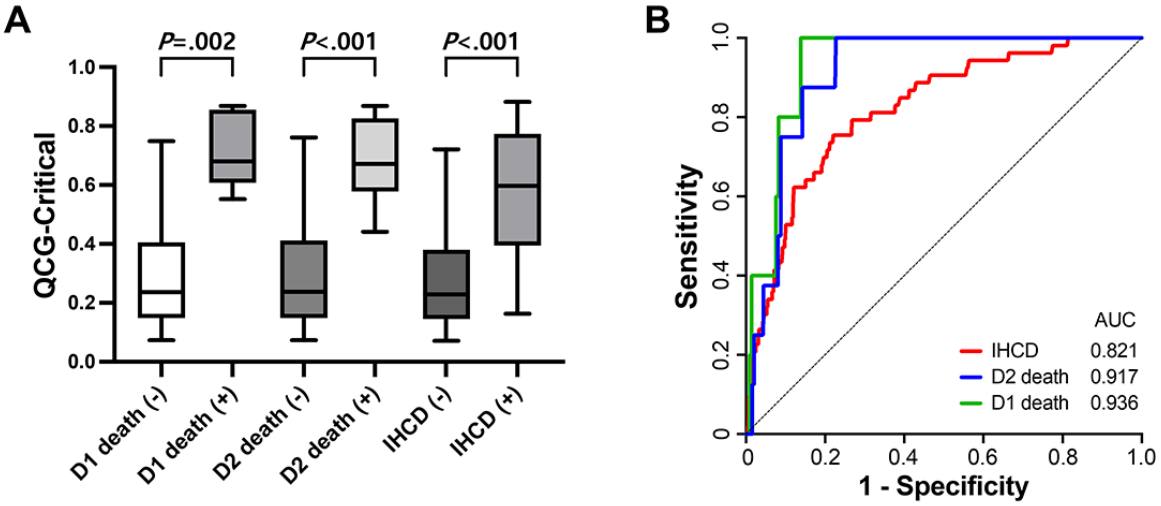

Comparatively, the AUC values of echocardiographic LVEF and NT-proBNP level for predicting in-hospital cardiac death were 0.642 (P<.001 vs QCG-Critical) and 0.720 (P=.07 vs QCG-Critical) (Figure 4A). The AUC value of the QCG-Critical score (0.821) was significantly (P=.02). higher than that of model 1 (0.705) established using traditional clinical variables, including age, sex, hypertension, diabetes, chronic kidney disease, cerebrovascular disease, ischemic heart disease, valvular heart disease, HF type, AF, and QRS duration. In addition, when the QCG-Critical score was added to model 1, it significantly enhanced the prediction for in-hospital cardiac death (AUC of model 1=0.705 vs AUC of model 1 with QCG-Critical=0.843; P<.001) (Figure 4B). When NT-proBNP and LVEF were further included in model 1 (model 2), the QCG-Critical score again demonstrated additional predictive value for in-hospital cardiac death compared to model 2 alone (AUC of model 2=0.787 vs AUC of model 2 with QCG-Critical=0.863; P=.01) (Figure 4C).

XSL•FO
RenderX

Figure 4. Receiver operating characteristic curves for predicting in-hospital cardiac mortality. (A) The AUC value of the QCG-Critical score was 0.821 and tended to be higher than that of echocardiographic LVEF, NT-proBNP, or a model utilizing traditional clinical variables (model 1). (B) Performance of prediction model 1 and with the addition of the QCG-Critical score. (C) Further incorporation of NT-proBNP and LVEF into the model 1 (model 2). The QCG-Critical score demonstrated additional predictive value for in-hospital cardiac mortality than model 2 alone. Model 1 includes age, sex, hypertension, diabetes, chronic kidney disease, cerebrovascular disease, ischemic heart disease, valvular heart disease, heart failure type, atrial fibrillation, and QRS duration. NT-proBNP and LVEF were further incorporated into model 2. **P* value for comparison with the AUC of QCG-Critical score. AUC: area under the curve; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; QCG: Quantitative electrocardiogram.

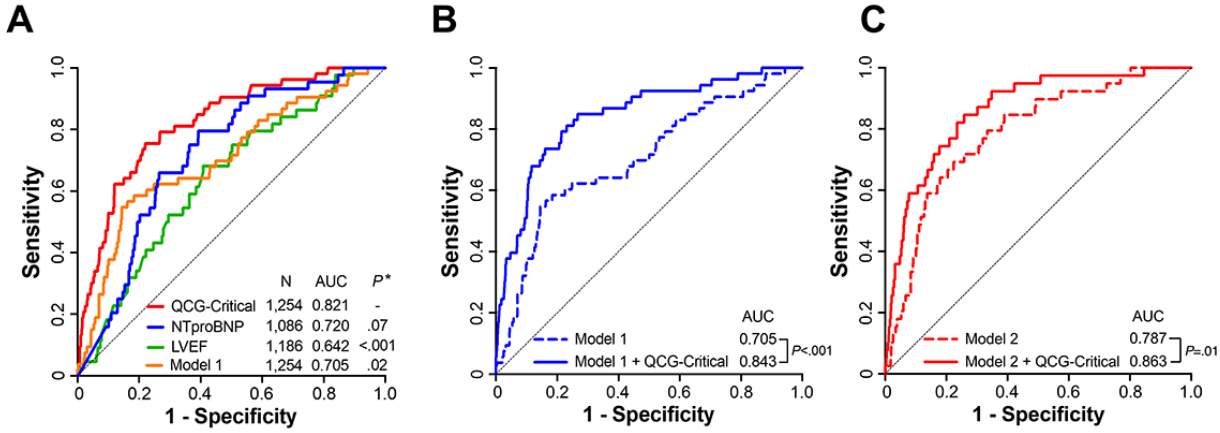

QCG-Critical Score and Long-Term Outcomes

During a median follow-up of 2.7 years, 508 deaths occurred in the study population. To further analyze the performance of the QCG-Critical score for outcome prediction, we divided patients into three QCG-Critical score groups based on arbitrary cut-off values of 0.25 and 0.50 and then conducted survival analysis (Figure 5).

Figure 5. Kaplan-Meier curves for long-term mortality according to QCG-Critical scores. HR: hazard ratio; QCG: Quantitative electrocardiogram.

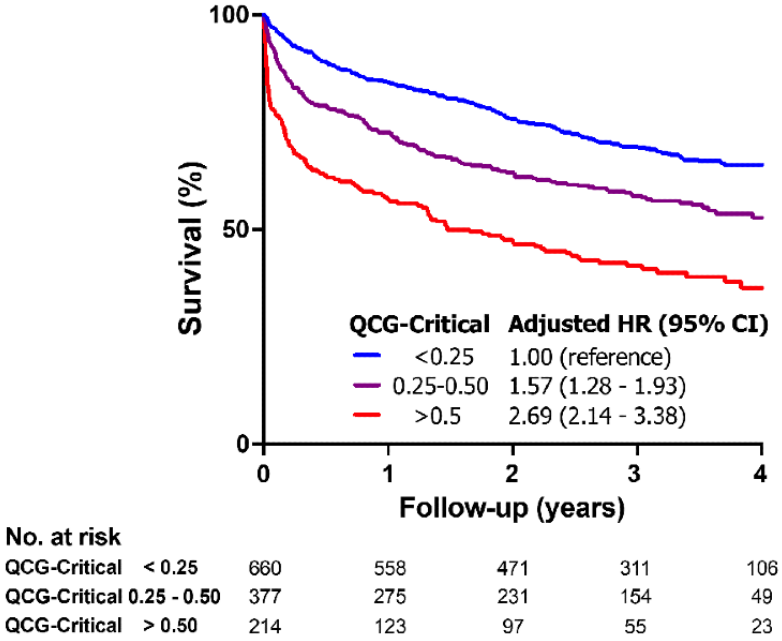

After adjustment for age, sex, comorbidities, HF etiology and type, AF, and QRS widening, patients with higher QCG-Critical scores had significantly higher all-cause mortality rates during follow-up than those with lower QCG-Critical scores (<0.25). The adjusted HRs for patients with QCG-Critical scores between 0.25 and 0.50 and for patients with QCG-Critical scores higher than 0.50 were 1.57 (95% CI 1.28-1.93) and 2.69 (95% CI 2.14-3.38), respectively (all *P*<.001). With additional adjustment for LVEF and NT-proBNP to the previous model, the adjusted HRs were 1.61 and 2.27, respectively, consistent with the main analysis (Figure S2 in the Multimedia Appendix 1).

In a subgroup analysis, a higher QCG-Critical score (>0.50 vs ≤0.50) was significantly correlated with all-cause mortality during follow-up, regardless of the initial rhythm (AF or sinus rhythm), QRS width (wide or narrow), hypertension, diabetes, HF etiology, HF type (de novo or acute decompensated HF), and LVEF (HF with reduced ejection fraction vs HF with preserved or mildly reduced ejection fraction), after adjustment for other clinical parameters (Figure S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Discussion

Predicting outcomes in patients with HF is important for guiding management and improving prognosis [15] but is often hindered by the complexity of HF pathophysiology and the presence of other comorbidities. Recently, AI algorithms based on big data from medical records have been found to be helpful in predicting the outcomes of patients with HF [16,17]; however, these algorithms are difficult to apply in daily practice and their performance requires further improvements. In this study, the QCG-Critical score, a newly developed AI-based ECG score, was well correlated with early mortality and in-hospital cardiac death during the index after adjusting for traditional clinical risk factors. Moreover, the QCG-Critical score was an independent predictor of long-term all-cause mortality in this population, suggesting that this AI-based ECG score may serve as a novel biomarker for these patients.

ECG is a cost-effective, widely available, and easy-to-perform test, and is therefore often used as a first-line evaluation for patients with cardiovascular diseases. ST-elevation myocardial infarction is a quintessential disease where ECG evaluation is critical for a timely diagnosis. Although ECG is not deterministic for an HF diagnosis, several studies have demonstrated that some ECG features are correlated with the characteristics of HF [18]. In addition, the presence of AF or QRS widening may represent ECG features reflecting unfavorable underlying hemodynamics, thus correlating with a poor prognosis [19,20]. More subtle ECG changes have also been suggested as predictors of a poor prognosis in patients with HF; however, these require high levels of experience and skill for interpretation, which may limit their applicability [21].

Theoretically, the ECG signal may contain information regarding the electric and mechanical activities of the diseased heart beyond a physician's perception. With the assistance of AI, ECG may provide valuable information beyond its current usage. For example, Attia et al [22] reported that LVEF reduction may be detected by ECG using AI. This new application of AI-ECG was reproduced by other researchers [23,24]. In this study, the QCG-HF score also showed good performance in predicting reduced echocardiographic LVEF of less than 40%, with an AUC value of 0.884 (Figure S4 in Multimedia Appendix 1). Notably, in the above-mentioned studies, the AI-ECG–predicted LVEF was correlated with the prognosis of patients with chronic HF, whereas the AI-based ECG score had a predictive value in patients with acute HF in this study. Thus, to the best of our knowledge, this study represents an initial effort in terms of predicting the outcomes of acute HF using AI-based ECG interpretation.

The QCG-Critical score was originally trained to detect critical medical conditions that may result in shock or mortality within 1 day [12]. In this study, the QCG-Critical score predicted early cardiac mortality in patients with acute HF with high accuracy. The AUC value of the QCG-Critical score was higher than that of echocardiographic LVEF for the prediction of in-hospital cardiac death and was also higher than the AUC value of the serum NT-proBNP level, but without statistical significance. Notably, the QCG-Critical score was available for all 1254 patients enrolled in the KorAHF study, whereas LVEF and NT-proBNP results were not available in 68 (5.4%) and 168 (13.4%) patients, respectively. Considering that the KorAHF study enrolled patients from tertiary centers in Korea, a high proportion of patients with acute HF might not have the opportunity to benefit from these echocardiographic or serum biomarker tests in real-world practice. Because ECG is a widely available evaluation tool and QCG scores are derived from ECG images, the QCG-Critical score may serve as an adequate alternative biomarker for risk stratification of patients with acute HF in real-world settings with limited resources. This score may also be useful even in well-equipped centers because it would be available immediately after the ECG exam, without requiring additional waiting for echocardiography or laboratory tests. This may be beneficial for timely risk stratification in the emergency department. The QCG-Critical score was not only correlated with in-hospital cardiac death but also showed a strong association with long-term mortality. In addition, the subgroup analysis demonstrated a consistent correlation between the QCG-Critical score and clinical outcomes. These results emphasize the potential of AI-based ECG interpretation as a novel biomarker in this field.

This study has several limitations. First, the study population predominantly consisted of Asian patients; hence, further studies are needed to validate our results across different ethnicities. Second, the AI algorithm tested in this study was derived from one of the participating centers (Seoul National University Bundang Hospital). However, there was a temporal difference between patient enrollment for algorithm training (2017 to 2019) and the test population (KorAHF enrollment, 2011 to 2014), and another external center (Severance Hospital) was involved in this study. Nevertheless, this may limit the generalizability of our findings. Third, the ECG format may affect the algorithm's performance. Although the manufacturers of the ECG devices used in the two participating hospitals differed (Philips PageWriter TC 30 and TC 70 at Seoul National University Bundang Hospital and GE Healthcare MAC 5500 and MAC VU360 at Severance Hospital), there was no significant difference in the AI algorithm performance between the hospitals. However, because the system uses printed ECG images as input, there may be problematic scenarios where the qualities of the images influence the predictive power of the biomarkers. Although some recent AI algorithm–based studies suggest further interpretation analysis, the QCG system does not support gradient-weighted class-activation mapping or similar visualization for model explainability due to the custom network architecture used. Therefore, we could not evaluate which part of the ECG images the system uses for each prediction.

In conclusion, predicting outcomes in patients with acute HF using the newly developed AI-based ECG score appears to be feasible. Thus, this score may serve as a novel biomarker for patients with HF, potentially overcoming the cost and availability constraints of current biomarkers.

XSL•FO
RenderX

Acknowledgments

YC and MY are co-first authors, and SMK and DJC are co-corresponding authors. Author SMK can be reached at smkang@yuhs.ac for correspondence.

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant RS-2023-00265933). This work was also supported by the Seoul National University Bundang Hospital Research Fund (grant 13-2024-0009).

Conflicts of Interest

JK and YC are employed by ARPI Inc and hold dual roles as CEO and Head of Research Collaboration Center, respectively. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Distribution of QCG-Critical scores (Figure S1); Kaplan-Meier curves for long-term mortality according to the QCG-Critical scores and adjusted HRs with additional adjustment for LVEF and NT-proBNP (Figure S2); subgroup analysis results for predicting long-term mortality (Figure S3); performance of the QCG-HF score for diagnosing left ventricular dysfunction (Figure S4).
[DOCX File , 621 KB - jmir_v26i1e52139_app1.docx]

References

1. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA Guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2022 May 03;79(17):e263-e421. [doi: 10.1016/j.jacc.2021.12.012] [Medline: 35379503]
2. McDonagh T, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, ESC Scientific Document Group. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021 Sep 21;42(36):3599-3726 [FREE Full text] [doi: 10.1093/eurheartj/ehab368] [Medline: 34447992]
3. Park JJ, Lee CJ, Park S, Choi J, Choi S, Park S, et al. Heart failure statistics in Korea, 2020: a report from the Korean Society of Heart Failure. Int J Heart Fail 2021 Oct;3(4):224-236 [FREE Full text] [doi: 10.36628/ijhf.2021.0023] [Medline: 36262554]
4. Januzzi JL, Rehman SU, Mohammed AA, Bhardwaj A, Barajas L, Barajas J, et al. Use of amino-terminal pro-B-type natriuretic peptide to guide outpatient therapy of patients with chronic left ventricular systolic dysfunction. J Am Coll Cardiol 2011 Oct 25;58(18):1881-1889 [FREE Full text] [doi: 10.1016/j.jacc.2011.03.072] [Medline: 22018299]
5. Masson S, Latini R, Anand IS, Barlera S, Angelici L, Vago T, Val-HeFT Investigators. Prognostic value of changes in N-terminal pro-brain natriuretic peptide in Val-HeFT (Valsartan Heart Failure Trial). J Am Coll Cardiol 2008 Sep 16;52(12):997-1003 [FREE Full text] [doi: 10.1016/j.jacc.2008.04.069] [Medline: 18786480]
6. Peacock WF, De Marco T, Fonarow GC, Diercks D, Wynne J, Apple FS, ADHERE Investigators. Cardiac troponin and outcome in acute heart failure. N Engl J Med 2008 May 15;358(20):2117-2126. [doi: 10.1056/NEJMoa0706824] [Medline: 18480204]
7. Verbrugge FH, Reddy YNV, Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, et al. Detection of left atrial myopathy using artificial intelligence-enabled electrocardiography. Circ Heart Fail 2022 Jan;15(1):e008176 [FREE Full text] [doi: 10.1161/CIRCHEARTFAILURE.120.008176] [Medline: 34911362]
8. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019 Sep 07;394(10201):861-867. [doi: 10.1016/S0140-6736(19)31721-0] [Medline: 31378392]
9. Lee SE, Cho H, Lee H, Yang H, Choi J, Jeon E, et al. A multicentre cohort study of acute heart failure syndromes in Korea: rationale, design, and interim observations of the Korean Acute Heart Failure (KorAHF) registry. Eur J Heart Fail 2014 Jun 02;16(6):700-708 [FREE Full text] [doi: 10.1002/ejhf.91] [Medline: 24797348]
10. Cho Y, Cho S, Oh I, Lee JH, Park JJ, Lee H, et al. Implantable cardioverter-defibrillator utilization and its outcomes in Korea: data from Korean Acute Heart Failure Registry. J Korean Med Sci 2020 Nov 30;35(46):e397 [FREE Full text] [doi: 10.3346/jkms.2020.35.e397] [Medline: 33258331]
11. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
12. Kim D, Hwang JE, Cho Y, Cho H, Lee W, Lee JH, et al. A retrospective clinical evaluation of an artificial intelligence screening method for early detection of STEMI in the emergency department. J Korean Med Sci 2022 Mar 14;37(10):e81 [FREE Full text] [doi: 10.3346/jkms.2022.37.e81] [Medline: 35289140]

XSL•FO
RenderX

13. Choi YJ, Park MJ, Ko Y, Soh M, Kim HM, Kim CH, et al. Artificial intelligence versus physicians on interpretation of printed ECG images: diagnostic performance of ST-elevation myocardial infarction on electrocardiography. Int J Cardiol 2022 Sep 15;363:6-10. [doi: 10.1016/j.ijcard.2022.06.012] [Medline: 35691440]
14. Park J, Yoon Y, Cho Y, Kim J. Feasibility of artificial intelligence-based electrocardiography analysis for the prediction of obstructive coronary artery disease in patients with stable angina: validation study. JMIR Cardio 2023 May 02;7:e44791 [FREE Full text] [doi: 10.2196/44791] [Medline: 37129937]
15. Fonarow GC, Albert NM, Curtis AB, Stough WG, Gheorghiade M, Heywood JT, et al. Improving evidence-based care for heart failure in outpatient cardiology practices: primary results of the Registry to Improve the Use of Evidence-Based Heart Failure Therapies in the Outpatient Setting (IMPROVE HF). Circulation 2010 Aug 10;122(6):585-596. [doi: 10.1161/CIRCULATIONAHA.109.934471] [Medline: 20660805]
16. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
17. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J Am Heart Assoc 2018 Apr 12;7(8):e008081 [FREE Full text] [doi: 10.1161/JAHA.117.008081] [Medline: 29650709]
18. Hendry PB, Krisdinarti L, Erika M. Scoring system based on electrocardiogram features to predict the type of heart failure in patients with chronic heart failure. Cardiol Res 2016 Jun;7(3):110-116 [FREE Full text] [doi: 10.14740/cr473w] [Medline: 28197277]
19. Lee JH, Park JJ, Cho Y, Oh IY, Yoo BS, Kim JJ, et al. Prognostic implication of ventricular conduction disturbance pattern in hospitalized patients with acute heart failure syndrome. Korean Circ J 2019 Jul;49(7):602-611 [FREE Full text] [doi: 10.4070/kcj.2018.0290] [Medline: 30891964]
20. Wang NC, Maggioni AP, Konstam MA, Zannad F, Krasa HB, Burnett JC, Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study With Tolvaptan (EVEREST) Investigators. Clinical implications of QRS duration in patients hospitalized with worsening heart failure and reduced left ventricular ejection fraction. JAMA 2008 Jun 11;299(22):2656-2666. [doi: 10.1001/jama.299.22.2656] [Medline: 18544725]
21. Das MK, Maskoun W, Shen C, Michael MA, Suradi H, Desai M, et al. Fragmented QRS on twelve-lead electrocardiogram predicts arrhythmic events in patients with ischemic and nonischemic cardiomyopathy. Heart Rhythm 2010 Jan;7(1):74-80. [doi: 10.1016/j.hrthm.2009.09.065] [Medline: 20129288]
22. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019 Jan 7;25(1):70-74. [doi: 10.1038/s41591-018-0240-2] [Medline: 30617318]
23. Choi J, Lee S, Chang M, Lee Y, Oh GC, Lee H. Deep learning of ECG waveforms for diagnosis of heart failure with a reduced left ventricular ejection fraction. Sci Rep 2022 Aug 20;12(1):14235. [doi: 10.1038/s41598-022-18640-8] [Medline: 35987961]
24. Kwon J, Kim K, Jeon K, Kim HM, Kim MJ, Lim S, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. Korean Circ J 2019 Jul;49(7):629-639 [FREE Full text] [doi: 10.4070/kcj.2018.0446] [Medline: 31074221]

Abbreviations

AF: atrial fibrillation
AI: artificial intelligence
AUC: area under the curve
ECG: electrocardiogram
HF: heart failure
HR: hazard ratio
KorAHF: Korean Acute Heart Failure
LVEF: left ventricular ejection fraction
NT-proBNP: N-terminal prohormone of brain natriuretic peptide
OR: odds ratio
QCG: Quantitative electrocardiogram
ROC: receiver operating characteristic

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.08.23; peer-reviewed by S Khan, J Zeng; comments to author 08.02.24; revised version received 22.02.24; accepted 29.05.24; published 03.07.24.

Please cite as:
Cho Y, Yoon M, Kim J, Lee JH, Oh IY, Lee CJ, Kang SM, Choi DJ
Artificial Intelligence–Based Electrocardiographic Biomarker for Outcome Prediction in Patients With Acute Heart Failure: Prospective Cohort Study
J Med Internet Res 2024;26:e52139
URL: https://www.jmir.org/2024/1/e52139
doi:10.2196/52139
PMID:

©Youngjin Cho, Minjae Yoon, Joonghee Kim, Ji Hyun Lee, Il-Young Oh, Chan Joo Lee, Seok-Min Kang, Dong-Ju Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Harnessing Artificial Intelligence to Predict Ovarian Stimulation Outcomes in In Vitro Fertilization: Scoping Review

Rawan AlSaad[1], PhD; Alaa Abd-alrazaq[1], PhD; Fadi Choucair[2], PhD; Arfan Ahmed[1], PhD; Sarah Aziz[1], MSc; Javaid Sheikh[1], MD

[1]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha, Qatar

[2]Reproductive Medicine Unit, Sidra Medicine, Doha, Qatar

Corresponding Author:
Rawan AlSaad, PhD
AI Center for Precision Health
Weill Cornell Medicine-Qatar
Education City, Street 2700
Doha
Qatar
Phone: 974 44928830
Email: rta4003@qatar-med.cornell.edu

Abstract

Background: In the realm of in vitro fertilization (IVF), artificial intelligence (AI) models serve as invaluable tools for clinicians, offering predictive insights into ovarian stimulation outcomes. Predicting and understanding a patient's response to ovarian stimulation can help in personalizing doses of drugs, preventing adverse outcomes (eg, hyperstimulation), and improving the likelihood of successful fertilization and pregnancy. Given the pivotal role of accurate predictions in IVF procedures, it becomes important to investigate the landscape of AI models that are being used to predict the outcomes of ovarian stimulation.

Objective: The objective of this review is to comprehensively examine the literature to explore the characteristics of AI models used for predicting ovarian stimulation outcomes in the context of IVF.

Methods: A total of 6 electronic databases were searched for peer-reviewed literature published before August 2023, using the concepts of IVF and AI, along with their related terms. Records were independently screened by 2 reviewers against the eligibility criteria. The extracted data were then consolidated and presented through narrative synthesis.

Results: Upon reviewing 1348 articles, 30 met the predetermined inclusion criteria. The literature primarily focused on the number of oocytes retrieved as the main predicted outcome. Microscopy images stood out as the primary ground truth reference. The reviewed studies also highlighted that the most frequently adopted stimulation protocol was the gonadotropin-releasing hormone (GnRH) antagonist. In terms of using trigger medication, human chorionic gonadotropin (hCG) was the most commonly selected option. Among the machine learning techniques, the favored choice was the support vector machine. As for the validation of AI algorithms, the hold-out cross-validation method was the most prevalent. The area under the curve was highlighted as the primary evaluation metric. The literature exhibited a wide variation in the number of features used for AI algorithm development, ranging from 2 to 28,054 features. Data were mostly sourced from patient demographics, followed by laboratory data, specifically hormonal levels. Notably, the vast majority of studies were restricted to a single infertility clinic and exclusively relied on nonpublic data sets.

Conclusions: These insights highlight an urgent need to diversify data sources and explore varied AI techniques for improved prediction accuracy and generalizability of AI models for the prediction of ovarian stimulation outcomes. Future research should prioritize multiclinic collaborations and consider leveraging public data sets, aiming for more precise AI-driven predictions that ultimately boost patient care and IVF success rates.

(J Med Internet Res 2024;26:e53396) doi:10.2196/53396

KEYWORDS

artificial intelligence; AI; AI models; AI model; in vitro fertilization; IVF; ovarian stimulation; infertility; fertility; ovary; ovaries; reproductive; reproduction; gynecology; prediction; predictions; predictive; prediction model; ovarian; adverse outcome; fertilization; pregnancy

XSL•FO
RenderX

Introduction

Background

Infertility is a global health issue affecting millions of people of reproductive age [1]. A recent report from the World Health Organization (WHO) indicates that 1 in 6 (17.5%) adults worldwide experience infertility during their lifetime [2]. In the United States, among married women aged 15 to 49 years with no prior births, about 1 in 5 (19%) are unable to conceive after 1 year of trying [3]. One of the most common forms of assisted reproductive technologies is in vitro fertilization (IVF). During an IVF procedure, the ovaries are first stimulated to produce multiple eggs. Once mature, these eggs are surgically retrieved, fertilized with sperm in the laboratory, and finally transferred back into a woman's uterus.

The IVF treatment process is lengthy, financially burdensome, emotionally taxing, and physically demanding for the couple, all without any guaranteed success of pregnancy [4,5]. The success rate is influenced by a range of factors, with some stemming from the patient's quality of gametes and others associated with the expertise and the service quality of the clinic. With increasing age ovarian reserve decreases, leading to a decline in both natural fertility rates and the success rates of IVF programs. For instance, the live birth rate after IVF for women aged younger than 35 years is 41.6%. However, this rate drops to 29.6% for those aged 35-37 years and diminishes to just 9.2% for the 41-42 years age group [3]. This declining probability of success with age adds another layer of uncertainty and complexity to the expected outcomes of the IVF treatment.

To enhance pregnancy rates in IVF, ovarian stimulation protocols are used to stimulate the growth of multiple follicles, allowing the retrieval of several oocytes. This approach permits the selection of 1 or multiple embryos for implantation. In addition to their effectiveness, stimulated cycles can lead to complications such as ovarian hyperstimulation syndrome (OHSS) [6], a potentially life-threatening condition. On the other hand, understimulation might result in an insufficient number of metaphase II (MII) oocytes, thereby reducing the IVF success rate [6].

There is a growing body of evidence suggesting that ovarian stimulation could potentially have adverse impacts on the quality of oocytes and embryos, ultimately affecting clinical results [7,8]. It has been proposed that careful tailoring of controlled ovarian hyperstimulation protocols is crucial for both effective treatment and subsequently achieving optimized success in IVF [9,10]. Personalizing controlled ovarian hyperstimulation protocols involves making a series of critical clinical decisions that are essential for optimizing the success rate. These decisions include selecting the most appropriate stimulation protocol, determining the optimal starting dose of gonadotropins, assessing the potential need for adjunctive agents, establishing the frequency of ultrasonographies and blood tests to monitor follicular growth, and deciding on the optimal time for initiating final oocyte maturation, as well as selecting the triggering agent [11,12]. Currently, the common practice for making these decisions predominantly depends on individual clinician expertise. While this expertise is valuable, the approach can be subjective, potentially leading to clinic-to-clinic variability in outcomes. Consequently, there is an urgent need for more consistent and data-driven approaches to enhance the precision of these important decisions during ovarian stimulation, thereby optimizing IVF success rates.

Artificial intelligence (AI) is being increasingly used in various fields of medicine, including reproductive medicine [13-15]. In the realm of assisted reproductive technologies, AI has the potential to revolutionize by personalizing treatment plans, enhancing embryo selection accuracy, and developing new technologies [16,17]. The recent integration of AI into IVF treatment has significantly expanded the literature on AI models tailored for IVF procedures. Several machine learning (ML) approaches, including artificial neural networks, support vector machines (SVMs), decision trees, and random forests have been used for clinical decision-making in IVF. These models cover a wide range of IVF-related tasks, encompassing pretreatment counseling [18-20], hormone dosage optimization [21,22], monitoring stimulation responses [23-25], sperm analysis [26], and evaluating embryo culture and development [27-29].

Research Problem and Aim

With the rapid advancement of AI, many studies used ML as a promising methodology for tailoring controlled ovarian stimulation strategies and improving IVF outcomes. Many reviews have attempted to synthesize the evidence related to the application of AI models in IVF settings [30-36]. Although these studies provide valuable insights, they also exhibit several significant limitations such as, first, the methodological approach—a significant proportion of these reviews adopted a narrative literature review approach rather than a systematic methodology. This choice complicates the extraction of objective and replicable conclusions. Second, the scope of the study—the breadth of these reviews often encompassed outcomes at different stages of IVF treatment. Such a generalized focus may overlook the significant implications of ovarian stimulation parameters. Third, AI models' characteristics—there was a lack of detailed descriptions regarding the features, types, and specifications of the AI models used. Furthermore, the characteristics of the clinical data used for the training and validation of these models were often neglected. Fourth, search sources—comprehensive searches across relevant databases, including MEDLINE, IEEE Xplore, and ACM Digital Library, were not consistently performed. Finally, clinical application—there was a limited discussion about the direct clinical applicability and implications of the AI models in the ovarian stimulation phase, which is essential for practitioners seeking to implement these findings.

In light of the growing intersection between AI and reproductive medicine, and to address the limitations of previous review studies, this scoping review aims to examine the current landscape of AI models used for predicting ovarian stimulation outcomes within the context of IVF. Specifically, our study aims to address the research questions such as (1) what are the characteristics of AI models currently used to predict ovarian stimulation outcomes in IVF treatments? (2) What are the characteristics of IVF treatment cycles incorporated into these AI models? (3) What are the characteristics of the data used to

XSL•FO
RenderX

train these AI models for the prediction of ovarian stimulation outcomes?

Methods

Overview

To achieve the aim of this study, we conducted a scoping review in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [37]. The PRISMA-ScR checklist pertinent to this review can be found in Multimedia Appendix 1 [37]. The following subsections present a detailed description of the methods used in this review.

Search Strategy

To retrieve relevant studies, we searched 6 electronic databases on June 13, 2023—Scopus, MEDLINE (through Ovid), Embase (through Ovid), ACM Digital Library, IEEE Xplore, and Google Scholar. We then scheduled a biweekly automatic search over a span of 10 weeks, concluding on August 22, 2023. Due to the extensive number of results from Google Scholar, which ranks them by relevance, we only assessed the top 100 entries, equivalent to 10 pages. To ensure a thorough review, we screened the reference lists of our primary selected studies (ie, backward reference list checking) and considered studies that referenced our primary selections (ie, forward reference list checking). Our search criteria were composed of 2 main categories of terms—terms associated with IVF (eg, in vitro fertilization, assisted reproductive technologies, and intracytoplasmic sperm injection) and terms related to AI (eg, artificial intelligence, machine learning, and deep learning). The detailed search queries for each of the databases can be found in Multimedia Appendix 2.

Study Eligibility Criteria

This review focused on studies that investigated the use of AI methods to predict and monitor the outcomes of the ovarian stimulation phase during IVF cycles. We did not set any limitations based on age or ethnicity. However, we excluded studies related to non–IVF-assisted reproductive technologies, natural conception, individuals without fertility issues, or those using IVF solely for fertility preservation, egg freezing, or egg donation. We included studies that used AI models exclusively fed with data from the stimulation phase. Studies that incorporated variables into their prediction models beyond the ovarian stimulation phase—such as oocyte fertilization, embryo culture, embryo grading, cryopreservation, and luteal phase support, which take place post-oocyte retrieval—were not considered.

In terms of predicted outcomes, our focus was on two categories, (1) intermediate IVF outcomes, including the number of oocytes retrieved, the number of mature oocytes, follicle count and size, the number of fertilized oocytes, and the number of top-quality embryos and (2) IVF cycle clinical outcomes namely implantation, chemical and clinical pregnancy, and live birth rates. We excluded papers that focused on post-IVF cycle outcomes, such as neonatal health and complications, maternal health post delivery, menstrual cycle regularity post-IVF treatment, the impact on subsequent IVF cycles, and long-term ovarian health.

Our review focused on studies using any AI form, ranging from ML, deep learning, and supervised and unsupervised learning, particularly within the IVF ovarian stimulation phase. We excluded studies that did not deploy AI or ML techniques, as well as those relying solely on statistical methods. Additionally, we excluded studies that did not provide sufficient details about the AI technique used or its specific role in the ovarian stimulation process.

In terms of study design, this review encompassed both retrospective and prospective study designs. Regarding the type of publication, we included peer-reviewed articles, theses, dissertations, and conference articles. We excluded non–peer-reviewed articles, preprints, reviews, opinion papers, research letters, commentaries, editorials, case studies, conference abstracts, posters, and protocols. Studies were limited to those published in English, with no constraints on the year of publication.

Study Selection

The study selection was conducted in 3 phases. First, we used EndNote X9 (Clarivate) to eliminate duplicates from the retrieved studies. Then, we screened the titles and abstracts of the remaining articles. In the final phase, we evaluated the full texts of the studies shortlisted in the preceding step. Two reviewers independently undertook the selection process, resolving disagreements through discussions. To measure the level of agreement between the 2 reviewers, we used Cohen κ [38], which yielded a value of 0.86.

Data Extraction

Two reviewers used Microsoft Excel to independently gather data on study metadata, IVF treatment cycles, AI algorithms, and data used in AI algorithm development. Disagreements were settled through discussion. The data extraction form for this review was piloted with 5 studies and can be found in Multimedia Appendix 3.

Data Synthesis

We synthesized the data extracted from the included studies using a narrative approach, summarizing and detailing the information through text, tables, and figures. First, we outlined the metadata of the included studies, such as the publication year and country. We then presented the characteristics of the IVF treatment cycles featured in the studies, including aspects like fertilization procedures, stimulation protocols, trigger medications, and outcome measures. Next, we summarized the AI algorithms used, identifying their aims, types, and validation methods. Finally, we detailed the specifics of the data used for AI algorithm development, covering aspects like sample size, number of features, data sources, and data types.

Results

Search Results

Figure 1 illustrates the search results from the preselected databases, totaling 1348 records. From this number, 444 were

XSL•FO
RenderX

identified and eliminated as duplicates using reference management software (EndNote X9). A review of the titles and abstracts of the subsequent 904 articles led to the exclusion of 829. Among the 75 left, full texts for 3 articles were unavailable. Upon full-text screening of the 72 accessible full texts 46 were discarded for various reasons, as shown in Figure 1. An additional 4 articles were identified through backward and forward referencing. In total, 30 articles were selected for inclusion in this review [39-68].

Figure 1. Flowchart of the study selection process. AI: artificial intelligence; IVF: in vitro fertilization.

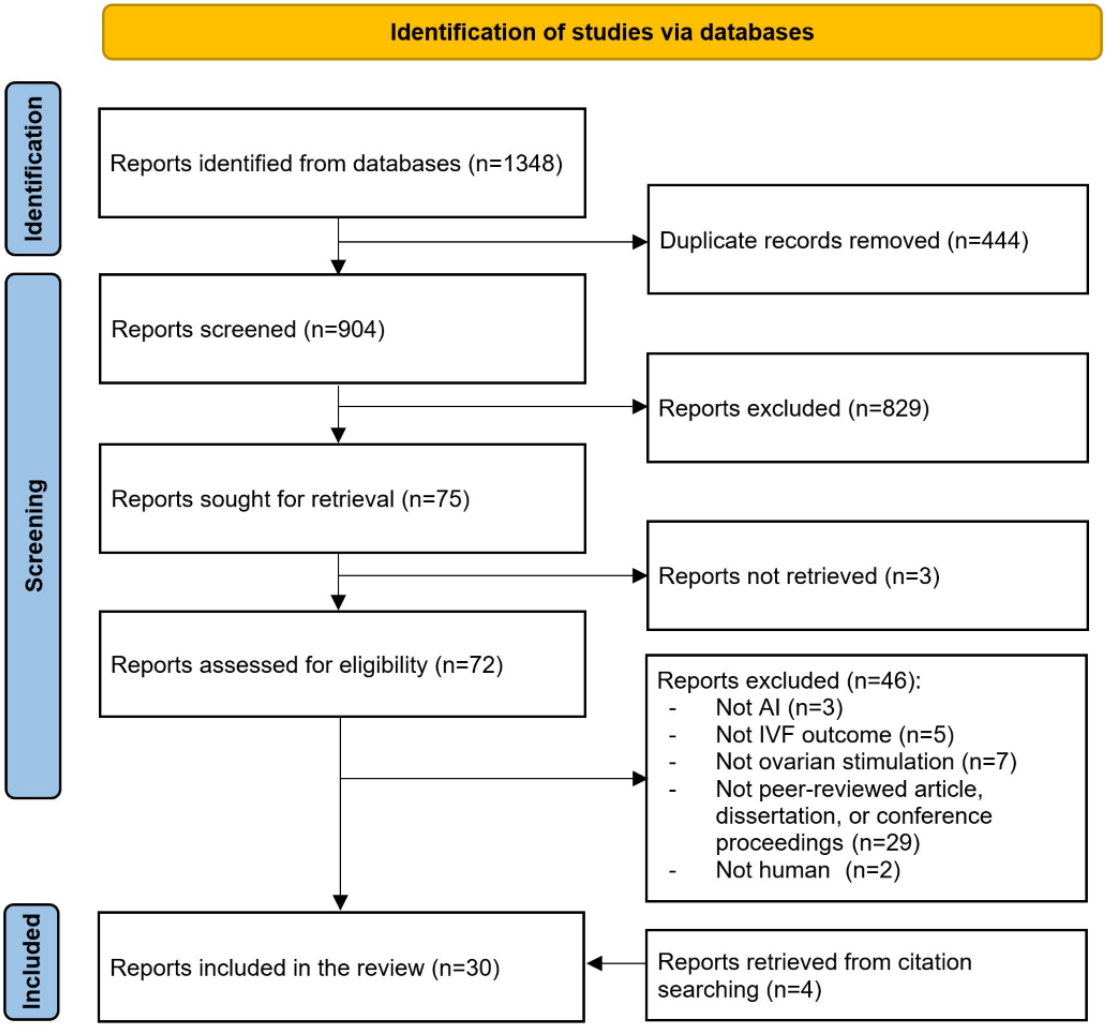

Characteristics of Included Studies

The studies included were published between 2000 and 2023 (Table 1). The largest number of studies were published in 2022 (n=11, 37%) and 2021 (n=6, 20%). The included studies originated from 13 different countries (Table 1). China was the leading country of publication with over a quarter (n=11, 37%) of the studies, followed by the United States (n=6, 20%). Of these, 87% (n=26) were peer-reviewed journal articles, while the remainder 13% (n=4) were conference papers. Regarding the research design, 77% (n=23) of the studies were retrospective and 23% (n=7) were prospective. As for data collection sites, single-site studies constituted the majority at 80% (n=24) whereas, multisite studies accounted for 17% (n=5). The number of participants in the studies varied from 4 to 30,278, with an average of 4798.5 (SD 8004.4; Table 1). The mean age of participants, reported in 20 studies, ranged from 18 to 50 years, with an average of 34.6 (SD 2.7). In the 16 studies that reported the mean BMI, the range was between 14.4 and 50.9, with an average BMI of 23.1 (SD 1.0). Multimedia Appendix 4 [39-68] provides detailed characteristics of each included study.

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Feature	Values	References
Study characteristics		
Year of publication, n (%)		
2023	4 (13)	[55,59,64,68]
2022	11 (37)	[42-47,49,52-54,60]
2021	6 (20)	[39,48,61,63,66,67]
2020	5 (13)	[41,51,57,58,65]
2018	1 (3)	[50]
2016	1 (3)	[40]
2013	1 (3)	[56]
2000	1 (3)	[62]
Type of publication, n (%)		
Journal article	26 (87)	[40-49,51-60,62,64-68]
Conference paper	4 (13)	[39,50,61,63]
Country of publication, n (%)		
China	11 (37)	[42,47,49,53-55,59,64-67]
United States	6 (20)	[44,45,48,51,52,58]
Greece	2 (7)	[46,60]
United Kingdom	2 (7)	[57,62]
Others	9 (30)	[39-41,43,50,56,61,63,68]
Study design		
Research design, n (%)		
Retrospective	23 (77)	[41-45,47-54,57-61,63-68]
Prospective	7 (23)	[39,40,46,55,56,62]
Single or multisite, n (%)		
Single-site	24 (80)	[39-42,47-49,51-67]
Multisite	5 (17)	[43-46,68]
NR[a]	1 (3)	[50]
Number of participants		
Mean (SD)	4798.5 (8004.4)	[40,42,44-46,48-50,52-54,56-68]
Range	4-30278	[40,42,44-46,48-50,52-54,56-68]
Patient characteristics		
Women's age (years)		
Mean (SD)	34.6 (2.7)	[41-43,46-49,52-58,60,62,64,65,67,68]
Range	18-50	[39,40,44,45,50,52,59,61,63,66]
Women's BMI		
Mean (SD)	23.1 (1.0)	[41-43,46-49,51,53-56,60,64,65,68]
Range	14.4-50.9	[39,40,44,45,50,52,57-59,61-63,66,67]

[a]NR: not reported.

Features of IVF Treatment Cycles

In the studies we examined, fertilization was accomplished through either IVF, intracytoplasmic sperm injection (ICSI), or a combination of both. Specifically, IVF was used in 14 out of 30 studies (47%), ICSI in 3 out of 30 (10%), and a combination of IVF and ICSI in 11 out of 30 studies (37%; Table 2). Regarding the stimulation protocols used, the

XSL•FO
RenderX

gonadotropin-releasing hormone (GnRH) antagonist protocol was the most prevalent which featured in 57% (n=17) of the studies. Following closely behind was the GnRH agonist protocol which was used in 47% (n=14) of the studies. In terms of trigger regimen, human chorionic gonadotropin (hCG) was the most common, used in 67% (n=20) of the studies. GnRH agonist was noted in 10% (n=3) of studies, while one-third (n=10, 33%) of the studies did not specify the trigger medication used.

XSL•FO
RenderX

Table 2. Features of IVF[a] treatment cycles.

Feature	Studies, n (%)	References
Fertilization method		
IVF	14 (47)	[40,43,45,51-56,58,61,63,65,68]
Mix of IVF and ICSI[b]	11 (37)	[41,42,46,47,49,57,59,62,64,66,67]
ICSI	3 (10)	[44,48,60]
NR[c]	2 (7)	[39,50]
Stimulation protocol		
GnRH[d] antagonist protocol	17 (57)	[41,42,44,47-49,51-58,64-66,68]
GnRH agonist protocol	14 (47)	[40,42,47-49,51-54,56,58,59,62,66]
Natural cycle IVF	2 (7)	[48,60]
Mild stimulation protocol	2 (7)	[48,53]
NR	9 (30)	[39,43,45,46,50,61,63,67,68]
Others	5 (17)	[47,49,52,54,66]
Trigger medication		
hCG[e]	20 (67)	[40-42,47,48,51-60,62,64-67]
GnRH agonist	3 (10)	[41,42,52]
NR	10 (33)	[39,43-46,49,50,61,63,68]
Outcome measures		
Number of oocytes retrieved	14 (47)	[43-45,48,52-55,60,63-66,68]
Blastocyst development	5 (17)	[48,56,58,64,68]
Live birth delivery	3 (10)	[40,41,46]
Number and size of follicles	3 (10)	[50,61,62]
Treatment management and optimization	3 (10)	[49,51,57]
Clinical pregnancy	2 (7)	[47,67]
Moderate or severe OHSS[f] incidence	2 (7)	[42,55]
Hormone levels poststimulation	2 (7)	[45,59]
Oocyte viability	1 (3)	[39]
Ground truth reference		
Microscopy images	20 (67)	[39,42-45,48,50,52-58,60,63-66,68]
Ultrasonography scans	5 (17)	[40,47,51,61,62]
Laboratory tests	4 (13)	[45,47,51,59]
Live birth delivery	4 (13)	[40,41,46,67]
Medications	1 (3)	[49]

[a]IVF: in vitro fertilization.
[b]ICSI: intracytoplasmic sperm injection.
[c]NR: not reported.
[d]GnRH: gonadotropin-releasing hormone.
[e]hCG: human chorionic gonadotropin.
[f]OHSS: ovarian hyperstimulation syndrome.

The outcome measures targeted by these studies spanned a broad spectrum, categorized into 9 primary groups (Table 3). The most frequently recorded outcome was the number of oocytes retrieved, reported in 47% (n=14) of the studies. Other significant outcomes included blastocyst development in 17% (n=5); live birth delivery, number and size of follicles, and treatment management and optimization each in 10% (n=3) of the studies.

XSL•FO
RenderX

Table 3. Features of AI[a] algorithms.

Feature	Studies, n (%)	References
Aim of AI algorithm		
Prediction of ovarian response	14 (47)	[42,44,52-55,57,60,63-68]
IVF[b] treatment management and optimization	8 (27)	[43,45,48,49,51,52,57,59]
Follicular monitoring or assessment	7 (23)	[39,40,50,53,56,61,62]
Prediction of live birth	4 (13)	[40,41,46,67]
Prediction of fertilization and embryo development	1 (3)	[58]
Prediction of pregnancy	1 (3)	[47]
AI algorithm used		
Support vector machine	9 (30)	[39,40,49-51,53,54,59,66]
Random forest	8 (27)	[51,52,56-58,63,64,66]
Gradient boosting	6 (20)	[47,48,52,63,66,68]
Linear regression	6 (20)	[42,43,45,52,55,63]
Decision tree	5 (17)	[41,53,58,63,66]
K-nearest neighbors	5 (17)	[40,44,52,53,63]
Logistic regression	5 (17)	[46,51,59,65,66]
Artificial neural networks	4 (13)	[49,51,54,66]
Multilayer perceptron	4 (13)	[41,46,53,62]
Convolutional neural network	3 (10)	[39,53,61]
Others	5 (17)	[40,51,52,61,67]
NR[c]	1 (3)	[60]
Type of validation		
Hold-out cross-validation	27 (90)	[39-49,51-56,59-67]
K-fold cross-validation	7 (23)	[44,48,57,64,65,67,68]
Leave-one-out cross-validation	2 (7)	[40,50]
NR	1 (3)	[58]
Evaluation metrics		
Area under the curve (AUC-ROC)	16 (53)	[39-49,51-56,59-68]
Sensitivity	10 (33)	[39,40,50-53,55,59,60,65]
Accuracy	8 (27)	[39,40,50-53,58,60]
Specificity	8 (27)	[39,40,50,53,55,59,60,65]
Positive predictive value (PPV)	7 (23)	[40,51-53,59,60,65]
Root mean squared error or mean squared error	6 (20)	[41,49,54,57,64,68]
Mean absolute error	5 (17)	[44,45,52,63,68]
Negative predictive value (NPV)	5 (17)	[40,53,59,60,65]
C-index	2 (7)	[42,48,66]
R-squared	2 (7)	[44-46]
Mean absolute percentage error	2 (7)	[63,68]
Precision	2 (7)	[39,59]
Correlation coefficient (*r*)	2 (7)	[49,60]
Others	4 (13)	[43,53,54,61]
NR	1 (3)	[62]

XSL•FO
RenderX

[a]AI: artificial intelligence.

[b]IVF: in vitro fertilization.

[c]NR: not reported.

To determine these outcomes, diverse tools were used as ground truth references. These references were categorized into 5 primary groups. The predominant reference was microscopy images, used in 67% (n=20) of the studies, followed by ultrasonography scans and laboratory tests, used in 17% (n=5) and 13% (n=4) of the studies respectively. Multimedia Appendix 5 [39-68] shows features of IVF treatment cycles in each included study.

Features of AI Algorithms

The AI algorithms in the analyzed studies primarily focused on 6 main objectives (Table 3). The most common aim was the prediction of ovarian response (n=14, 47%). This was followed by IVF treatment management and optimization in 27% (n=8) and follicular monitoring or assessment in 23% (n=7).

Different AI algorithms were used across the studies (Table 3). The SVM was used in 30% (n=9), while random forest algorithms were applied in 27% (n=8). Gradient boosting and linear regression each appeared in 20% (n=6) of the studies.

When it comes to the validation methods of the AI algorithms, the hold-out cross-validation approach dominated, being used in 90% (n=27) of the studies. Other methods like k-fold cross-validation were reported in 23% (n=7) and leave-one-out cross-validation in 7% (n=2) of the studies.

Different evaluation metrics were adopted across the studies (Table 3). The area under the curve (AUC) was the most prominent, used in 53% (n=16) of the studies. Following that, sensitivity was featured in 33% (n=10) of studies, while both accuracy and specificity were used in 27% (n=8) of the studies. For a detailed view, Multimedia Appendix 6 [39-68] provides insights into the features of AI algorithms across each study.

Features of Data Used in AI Algorithms Development

The vast majority of studies (n=29, 97%) specified the sample size used as an input for AI algorithms (see Table 4). These sizes varied widely, ranging between 24 to 37,062 records, encompassing diverse categories such as women participants, follicles, cycles, or images. On average, the sample size was 5666.6 (SD 9525.0). Notably, no studies used public data sets; all opted for closed data sets, which are not available to the public.

XSL•FO
RenderX

Table 4. Features of data used in artificial intelligence algorithms development.

Feature	Values	References
Sample size		
Mean (SD)	5666.6 (9525.0)	[40-68]
Range	24-37062	[40-68]
Data sources, n (%)		
Closed	30 (100)	[39-68]
Open	0 (0)	N/A[a]
Data types, n (%)		
Demographics data	22 (73)	[41-49,51-55,57,59,64-68]
Laboratory data	18 (60)	[41-47,49,51-56,59,62,64,65]
Radiology data	15 (50)	[44,45,48,50-54,57,60,61,63-65,68]
Medical history	14 (47)	[41-43,45-47,51,53,63-68]
Anthropometry data	13 (43)	[41-45,47,48,51,53,59,63-65]
Medications	6 (20)	[52-54,58,64,66]
IVF[b] cycle data	6 (20)	[45,48,53,64,66-68]
Microscopy images	3 (10)	[39,44,53]
Genetic data	3 (10)	[40,59,68]
Lifestyle data	3 (10)	[41,47,52]
Others	1 (3)	[51]
Number of features		
Mean (SD)	979.6 (5207.2)	[39-50,52-68]
Range	2-28054	[39-50,52-68]

[a]N/A: not applicable.

[b]IVF: in vitro fertilization.

Of the studies reviewed, 97% (n=29) of the studies specified the number of features used in both model development and validation. The number of features across these studies ranged from 2 to 28,054, with an average of 979.6 (SD 5207.2).

The data used as the underlying input for these algorithms were diverse. We categorized these data into 10 distinct types—patient demographics, laboratory data, radiology data, medical history, anthropometry data, medications, current IVF cycle data, microscopy images, genetic data, and lifestyle data. In the studies reviewed, patient demographics were the most used data type, reported in 73% (n=22) of the studies. Laboratory data, including baseline and dynamic hormonal levels such as follicle stimulating hormone, luteinizing hormone, estradiol, progesterone, and anti-Mullerian hormone were used in 60% (n=18) of the studies. Radiology data sources, such as pelvic ultrasonographies, transvaginal ultrasonographies, and magnetic resonance imaging (MRI) were observed in 50% (n=15) of the studies. Medical history, encompassing surgical history, reproductive history, and chronic conditions was used in 47% (n=14) of the studies. Anthropometry data were used in 43% (n=13). Current patient medications were considered in 20% (n=6), as was the current IVF cycle data, including details of stimulation protocols and oocytes. Microscopy images, including those of oocytes, embryos, and blastocysts, were used in 10% (n=3) of the studies. Multimedia Appendix 7 [39-68] provides insights into the features of data used in AI algorithm development across each study.

Discussion

Principal Findings

In the development of AI-driven predictive models, forecasting the outcome plays a pivotal role. This scoping review explores the features of AI models that use ovarian stimulation phase characteristics to predict outcomes in IVF cycles. The literature examined primarily featured models that incorporated patient demographic information, notably women's age at the time of treatment, in 73% of the studies. Additionally, laboratory data such as basal and dynamic hormonal levels (follicle stimulating hormone [FSH], antral follicle count [AFC], progesterone, and anti-Mullerian hormone [AMH]) were commonly included in 60% of the studies.

Both patient age and hormonal tracking are crucial inputs during controlled ovarian stimulation, impacting dose and protocol selection, as well as the clinical outcomes of IVF cycles [69-72]. The reviewed studies indicate that the GnRH antagonist protocol is frequently used as a stimulation protocol. Although the optimal protocol should be customized to accommodate each

XSL•FO
RenderX

patient's distinct requirements and characteristics, this trend reflects an increasing preference for the GnRH antagonist in recent times [73]. Its widespread adoption in many clinics is attributed to the shorter treatment duration, diminished side effects, reduced overall treatment burden, and outcomes that are on par with other protocols [74-76].

A noticeable trend in the reviewed studies was the diversity of AI algorithms used across studies. The SVM stood out, being used in 30% (n=9) of the studies, as indicated in Table 3. The preference for SVM likely stems from its capability to handle high-dimensional IVF data, its robustness against overfitting ensuring effective generalization to new data points, and its ability to navigate nonlinear relationships [77]. Ovarian stimulation prediction models, which involve parameters such as hormonal levels, age, genetic markers, and past medical history, can benefit from SVM's comprehensive high-dimensional data analysis. Furthermore, as IVF data sets may not always be comprehensive, the risk of overfitting, where the model becomes overly tailored to the training data and performs poorly on new data, increases. Therefore, SVM, with its inherent regularization properties, offers robustness against overfitting, making it a reliable choice for such data sets.

When assessing prediction models, various evaluation metrics come into consideration, with AUC being the most frequently used. The advantage of AUC lies in its consistency; while metrics such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) can vary based on different cutoff values, AUC remains unaffected by such variations.

We also observed that hold-out cross-validation was the most frequently used validation method. This observation can be attributed to 2 factors. First, the hold-out method is favored for its simplicity and efficiency, as it is computationally less demanding than alternatives such as k-fold cross-validation. Given the potential complexity of some AI models and the extensive training durations they may require, the hold-out approach can accelerate the evaluation process. Second, when predicting ovarian stimulation outcomes, data might exhibit inherent variability due to individual patient differences. In such cases, the hold-out method offers a stable and consistent validation method, reducing the variability that might arise from recurrent data partitioning and testing.

Regarding the aim of the AI algorithms, predicting ovarian response was the primary objective in almost half (47%) of the studies. This included the prediction of the number of oocytes and metaphase II oocytes retrieved, the number of pronuclear stage (2PNs), and categorizing ovarian response (from hyporesponders to hyperresponders). These parameters are pivotal as they directly influence the success rate of the IVF cycle [69,72,78]. The second prevalent outcome was IVF treatment management and optimization, accounting for nearly a third (27%) of the studies. This encompassed predicting oocyte maturation trigger timing, determining the optimal retrieval day, strategizing follow-up options (adjusting, personalizing, and reducing in-person visits), and optimizing the starting dose of gonadotropins. The current emphasis on managing IVF treatment highlights the importance of precision and personalization in optimizing both procedural outcomes and the patient experience throughout the fertility treatment [79-81].

As mentioned earlier, the primary objective of AI algorithms in the included studies was to predict ovarian response, which was quantified by the number and quality of oocytes. Hence, microscopy images naturally emerged as the principal ground truth reference, used in two-thirds of these studies to establish labels for training and validate the AI models. Microscopy images are valuable due to their high-resolution techniques, which provide detailed visualization, enabling accurate oocyte counting after preparation and staining. In addition to oocyte quantification, microscopy can critically assess oocyte quality by examining features such as the thickness of the zona pellucida, the size of the perivitelline space, the clarity of the ooplasm, and the alignment of polar bodies [82,83]. These detailed morphological insights underscore the value of microscopy images as an essential reference for training AI models to predict ovarian responses.

Based on the findings of our study, the GnRH antagonist protocol is frequently used as a stimulation protocol. Although the optimal protocol should be customized to accommodate each patient's distinct requirements and traits, this trend reflects an increasing preference for the GnRH antagonist in recent times [73]. Its widespread adoption in many clinics is attributed to the shorter treatment duration, diminished side effects, reduced overall treatment burden, and outcomes that are on par with other protocols [74-76].

Notably, 80% of the studies included in the analysis were limited to a single fertility clinic, prompting concerns about their generalizability to a broader context. Such models might overlook diverse patient demographics and protocols across clinics. To ensure the broader applicability of AI models in predicting ovarian stimulation outcomes, multicentric studies encompassing diverse clinical settings are essential.

More so, all the included studies relied on closed data sets that are not publicly available, with none using public data sets. This reflects the limited availability of comprehensive public IVF data sets, with only a few examples of public data sets such as the Human Fertilisation and Embryology Authority (HFEA) research data [84]. While closed data sets are valuable, they come with inherent limitations, often capturing only specific patient demographics or treatment methods. Thus, AI models developed exclusively on these data sets might introduce bias, potentially limiting their generalizability across diverse populations and treatments.

Research and Practical Implications

The performance of AI models in predicting ovarian stimulation outcomes was not assessed in this review. Systematic reviews and meta-analyses are required to evaluate their efficacy. Moreover, future studies should compare the performance of various AI models based on different data types (eg, ultrasonography data versus hormonal level data) and stimulation protocols. Conducting systematic reviews of these studies will assist clinicians, data scientists, and researchers in identifying the most significant features and accurate AI algorithms in predicting ovarian stimulation outcomes.

XSL•FO
RenderX

While the majority of the studies in this review are retrospective (77%), it is crucial to consider the implications this design might have on the applicability of AI models in predicting ovarian stimulation outcomes. Retrospective studies rely on previously recorded data, which can sometimes be incomplete, inaccurate, or biased. Consequently, such studies often encounter challenges in establishing causality and may be influenced by issues related to data integrity or selection bias. These limitations can, in turn, impact the accuracy and reliability of AI models trained on such data. On the other hand, prospective studies, which actively monitor outcomes against predictions, offer direct insights into AI model performance in real-world scenarios. They ensure better control over variables and reduce potential biases. A balanced approach that incorporates both retrospective and prospective designs is essential to fully harness the potential of AI in predicting and monitoring IVF outcomes.

Another important aspect to consider in AI models is bias in data selection and reporting. This bias can occur when training data are limited to IVF cycles with outcomes that exclude cycles of no response or cycles in which no embryos or eggs were collected. This type of bias can skew the model's understanding of real-world scenarios and potentially lead to inaccurate predictions or recommendations. Therefore, it is crucial to ensure that training data are representative of the full spectrum of IVF outcomes to mitigate this bias and enhance the model's reliability.

Another consideration that demands careful vigilance is the potential for AI-related failures that could pose safety concerns for both clinicians and patients. These errors may be linked to noise and artifacts in the input data, data shifts occurring after AI training, and unexpected variations in real-world scenarios. Further research is necessary to address these challenges. In some AI applications, it may be necessary to design fail-safe systems that can safeguard against potentially harmful recommendations in clinical practice. These systems could include mechanisms for continuous learning from new scenarios and errors as they are detected in practice [85].

Based on our observations in the reviewed studies, there was a prevalent use of closed, nonpublic data sets. This finding strongly emphasizes the imperative need to foster the development and use of accessible public data sets in the realm of IVF research. Such data sets, even when anonymized, present a more comprehensive and diverse data pool, encompassing a wide spectrum of patient profiles, stimulation protocols, and IVF outcomes from different regions and populations [86]. Thus, prioritizing initiatives that promote the establishment of public IVF data sets is essential in driving forward more reliable, accurate, and robust AI-powered prediction models in the field of IVF. In parallel, to ensure both appropriate transparency and intellectual property protection, it is essential to effectively address challenges posed by data confidentiality, along with the competition between clinics [87].

Another key consideration is the heterogeneity of data sources generated by different types of equipment and decentralized data collection systems. Given that each fertility clinic operates with distinct workflows and procedures, these differences must be accounted for during data analysis and modeling. Embracing standardization in data generation processes is a crucial endeavor that all involved organizations should prioritize. Establishing and implementing consistent quality standards would streamline data modeling from diverse sources and reduce data variability [88]. This consistency is not only important for comprehensive algorithm validation but also vital before advancing to potential production and commercialization.

Leveraging accurate and efficient AI models for predicting ovarian stimulation outcomes in IVF clinical practice offers transformative advantages. First, these AI models can accurately analyze complex patterns within large data sets, capturing significant predictors of ovarian response that might be overlooked by conventional approaches. This allows for tailored ovarian stimulation strategies, potentially optimizing follicular response and, subsequently enhancing the likelihood of successful IVF outcomes. Second, accurate predictions also mitigate risks associated with over- or understimulation, thus enhancing patient safety. Furthermore, with precise predictions, future research, particularly a systematic review and meta-analysis focusing on the performance of AI models for the prediction of ovarian stimulation outcomes, could offer more definitive insights into their efficacy and potential to enhance clinical practice. Such studies are essential not only for evaluating the effectiveness of AI tools but also for assessing their safety, cost-effectiveness, and overall contribution to improving patient care.

Limitations

In this review, we excluded studies that incorporated additional parameters from stages following oocyte retrieval in IVF (eg, oocyte fertilization, embryo culture, embryo grading, cryopreservation, and luteal phase support) along with ovarian stimulation parameters into their prediction models. While our primary objective was to identify characteristics of AI models specific to ovarian stimulation, which pertains solely to stimulation parameters, our findings may not be generalizable to contexts where parameters from the omitted stages are relevant to the research question. We included only studies published in the English language. Consequently, we may have overlooked relevant studies published in other languages. A notable limitation of this review is our inability to comment on the performance of AI models in predicting ovarian stimulation outcomes. This particular aspect is beyond the scope of this review and a comprehensive assessment would require systematic reviews that thoroughly examine the quality of evidence and potential biases. Another potential limitation of this study is that due to the rapidly evolving nature of the AI field, it is possible that new records in both academic and gray literature may have been published after our search concluded and before our publication.

Conclusions

In this scoping review, we explored the landscape of AI models applied to predict ovarian stimulation outcomes in IVF cycles. Notable findings from the literature include the frequent use of patient demographic information, especially age and hormonal levels for predictions, with SVM emerging as the preferred algorithm. A recurring theme was the reliance on single-center studies and closed data sets, spotlighting the immediate need

XSL•FO
RenderX

for more expansive, public data sets and multicentric collaborations to enhance generalizability and robustness. To further refine AI's impact on IVF care, emphasis should be on comprehensive data collection, enhanced transparency, and standardized methodologies. Implementing these approaches has the potential to facilitate the development of AI-driven predictions that are more precise and applicable, ultimately leading to improved patient care and higher IVF success rates.

Acknowledgments

No relevant funding, custom code, or mathematical algorithm was used in this study.

Data Availability

The data generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RA and AAa conceptualized and designed the study under the guidance and supervision of JS. RA and AAa conducted database searches, screened the studies, extracted and synthesized the data, and drafted the paper. FC, AA, SA, and JS reviewed and revised the paper. All authors have read and approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 116 KB - jmir_v26i1e53396_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 31 KB - jmir_v26i1e53396_app2.docx]

Multimedia Appendix 3
Data extraction form.
[DOCX File , 37 KB - jmir_v26i1e53396_app3.docx]

Multimedia Appendix 4
Characteristics of each included study.
[DOCX File , 26 KB - jmir_v26i1e53396_app4.docx]

Multimedia Appendix 5
Features of in vitro fertilization treatment cycles.
[DOCX File , 24 KB - jmir_v26i1e53396_app5.docx]

Multimedia Appendix 6
Features of artificial intelligence algorithms.
[DOCX File , 24 KB - jmir_v26i1e53396_app6.docx]

Multimedia Appendix 7
Features of data used in developing artificial intelligence algorithms.
[DOCX File , 24 KB - jmir_v26i1e53396_app7.docx]

References

1. Nik Hazlina NH, Norhayati MN, Shaiful Bahari I, Nik Muhammad Arif NA. Worldwide prevalence, risk factors and psychological impact of infertility among women: a systematic review and meta-analysis. BMJ Open 2022;12(3):e057132 [FREE Full text] [doi: 10.1136/bmjopen-2021-057132] [Medline: 35354629]
2. 1 in 6 people globally affected by infertility. World Health Organization (WHO). 2023. URL: https://www.who.int/news/item/04-04-2023-1-in-6-people-globally-affected-by-infertility [accessed 2023-09-13]

XSL•FO
RenderX

3. Society for Assisted Reproductive Treatments (SART). Annual Summary Report. 2020. URL: https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx?reportingYear=2020#patient-cumulative [accessed 2024-06-11]
4. Chen H, Li J, Cai S, Zeng S, Yin C, Kuang W, et al. Impact of body mass index (BMI) on the success rate of fresh embryo transfer in women undergoing first in vitro fertilization/intracytoplasmic sperm injection (IVF/ICSI) treatment. Int J Obes (Lond) 2022;46(1):202-210. [doi: 10.1038/s41366-021-00978-0] [Medline: 34628467]
5. Dabbagh Rezaeiyeh R, Mehrara A, Mohammad Ali Pour A, Fallahi J, Forouhari S. Impact of various parameters as predictors of the success rate of fertilization. Int J Fertil Steril 2022 Apr;16(2):76-84 [FREE Full text] [doi: 10.22074/IJFS.2021.531672.1134] [Medline: 35639653]
6. Datta AK, Maheshwari A, Felix N, Campbell S, Nargund G. Mild versus conventional ovarian stimulation for IVF in poor, normal and hyper-responders: a systematic review and meta-analysis. Hum Reprod Update 2021;27(2):229-253 [FREE Full text] [doi: 10.1093/humupd/dmaa035] [Medline: 33146690]
7. Bosch E, Labarta E, Kolibianakis E, Rosen M, Meldrum D. Regimen of ovarian stimulation affects oocyte and therefore embryo quality. Fertil Steril 2016;105(3):560-570 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.01.022] [Medline: 26826273]
8. Greb RR, Behre HM, Simoni M. Pharmacogenetics in ovarian stimulation - current concepts and future options. Reprod Biomed Online 2005;11(5):589-600. [doi: 10.1016/s1472-6483(10)61167-4] [Medline: 16409708]
9. La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. Hum Reprod Update 2014;20(1):124-140. [doi: 10.1093/humupd/dmt037] [Medline: 24077980]
10. Grisendi V, La Marca A. Individualization of controlled ovarian stimulation in vitro fertilization using ovarian reserve markers. Minerva Ginecol 2017;69(3):250-258. [doi: 10.23736/S0026-4784.17.04044-8] [Medline: 28271699]
11. Youssef MAF, van Wely M, Mochtar M, Fouda UM, Eldaly A, El Abidin EZ, et al. Low dosing of gonadotropins in in vitro fertilization cycles for women with poor ovarian reserve: systematic review and meta-analysis. Fertil Steril 2018;109(2):289-301 [FREE Full text] [doi: 10.1016/j.fertnstert.2017.10.033] [Medline: 29317127]
12. Kuc P. The most effective starting dose of gonadotropins in the first IVF cycle in patients under 35 years of age - GnRH agonist versus antagonist protocol. Hum Reprod 2016;31:448-448. [doi: 10.26226/morressier.573c1514d462b80296c98c5e]
13. Tur-Kaspa I. Artificial intelligence (AI) to optimize ovarian stimulation for IVF: from stimulation protocols to total and mature oocyte yield. Hum Reprod 2022;37:I478-I479. [doi: 10.1093/humrep/deac107.568]
14. Papatheodorou A. Successful implementation of an end-to-end artificial intelligence (AI) platform in a busy IVF clinic: a prospective observational study. Hum Reprod 2022;37:284-284. [doi: 10.1093/humrep/deac107.190]
15. Meseguer M, Maor RU, Alegre L, Del Gallego R, Pellicer A, Seidman DS, et al. Automated halo identification: a novel predictive feature for IVF success identified through an artificial intelligence (AI) algorithm. Fertil Steril 2019;112(3):e157. [doi: 10.1016/j.fertnstert.2019.07.528]
16. Medenica S, Zivanovic D, Batkoska L, Marinelli S, Basile G, Perino A, et al. The future is coming: artificial intelligence in the treatment of infertility could improve assisted reproduction outcomes-the value of regulatory frameworks. Diagnostics (Basel) 2022;12(12) [FREE Full text] [doi: 10.3390/diagnostics12122979] [Medline: 36552986]
17. Gullo G, Basile G, Cucinella G, Greco ME, Perino A, Chiantera V, et al. Fresh vs. frozen embryo transfer in assisted reproductive techniques: a single center retrospective cohort study and ethical-legal implications. Eur Rev Med Pharmacol Sci 2023;27(14):6809-6823 [FREE Full text] [doi: 10.26355/eurrev_202307_33152] [Medline: 37522692]
18. Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. J Transl Med 2019;17(1):317 [FREE Full text] [doi: 10.1186/s12967-019-2062-5] [Medline: 31547822]
19. Liao S, Pan W, Dai W, Jin L, Huang G, Wang R, et al. Development of a dynamic diagnosis grading system for infertility using machine learning. JAMA Netw Open 2020;3(11):e2023654 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.23654] [Medline: 33165608]
20. Bachelot G. The InfertilScore, a couple-based machine learning model in patients with idiopathic infertility. Hum Reprod 2020;35:1468-1469.
21. Manas N. Development and validation of an artificial intelligence algorithm that matches a clinician ability to select the best follitropin dose for ovarian stimulation. Hum Reprod 2021;36:427-427. [doi: 10.1093/humrep/deab130.636]
22. Choo CW, Kim JH. A prospective cohort study to develop a treatment algorithm for controlled ovarian stimulation to select a starting dose of recombinant follicle stimulating hormone based on patient characteristics and ovarian response (fame study). Fertil Steril 2021;116(3):e196-e197. [doi: 10.1016/j.fertnstert.2021.07.538]
23. Tan SL, Child TJ, Gulekli B. In vitro maturation and fertilization of oocytes from unstimulated ovaries: predicting the number of immature oocytes retrieved by early follicular phase ultrasonography. Am J Obstet Gynecol 2002;186(4):684-689. [doi: 10.1067/mob.2002.122146] [Medline: 11967491]
24. Eldar-Geva T, Margalioth EJ, Ben-Chetrit A, Gal M, Robertson DM, Healy DL, et al. Serum inhibin B levels measured early during FSH administration for IVF may be of value in predicting the number of oocytes to be retrieved in normal and low responders. Hum Reprod 2002;17(9):2331-2337. [doi: 10.1093/humrep/17.9.2331] [Medline: 12202421]

XSL•FO
RenderX

25. Ferrand T, Boulant J, He C, Chambost J, Jacques C, Pena CA, et al. Predicting the number of oocytes retrieved from controlled ovarian hyperstimulation with machine learning. Hum Reprod 2023;38(10):1918-1926 [FREE Full text] [doi: 10.1093/humrep/dead163] [Medline: 37581894]
26. You JB, McCallum C, Wang Y, Riordon J, Nosrati R, Sinton D. Machine learning for sperm selection. Nat Rev Urol 2021;18(7):387-403. [doi: 10.1038/s41585-021-00465-1] [Medline: 34002070]
27. Diakiw S, Hall JMM, VerMilyea MD, Amin J, Aizpurua J, Giardini L, et al. Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. Hum Reprod 2022;37(8):1746-1759 [FREE Full text] [doi: 10.1093/humrep/deac131] [Medline: 35674312]
28. Diakiw S. An artificial intelligence model that was trained on pregnancy outcomes for embryo viability assessment is highly correlated with gardner score. Hum Reprod 2021;36:128-129. [doi: 10.1093/humrep/deab128.046]
29. La Marca A, Capuzzo M, Longo M, Imbrogno MG, Spedicato GA, Fiorentino F, et al. The number and rate of euploid blastocysts in women undergoing IVF/ICSI cycles are strongly dependent on ovarian reserve and female age. Hum Reprod 2022;37(10):2392-2401. [doi: 10.1093/humrep/deac191] [Medline: 36006017]
30. Zaninovic N, Elemento O, Rosenwaks Z. Artificial intelligence: its applications in reproductive medicine and the assisted reproductive technologies. Fertil Steril 2019;112(1):28-30 [FREE Full text] [doi: 10.1016/j.fertnstert.2019.05.019] [Medline: 31277764]
31. Khan HL, Khan S, Bhatti S, Abbas S. Role of artificial intelligence in quality assurance in ART: a review. FandR 2022;05(01):1-7. [doi: 10.1142/s2661318223300015]
32. Raimundo JM, Cabrita P. Artificial intelligence at assisted reproductive technology. Procedia Comput Sci 2021;181:442-447. [doi: 10.1016/j.procs.2021.01.189]
33. Letterie G. Artificial intelligence and assisted reproductive technologies: 2023. Ready for prime time? Or not. Fertil Steril 2023;120(1):32-37. [doi: 10.1016/j.fertnstert.2023.05.146] [Medline: 37210064]
34. Letterie G. Three ways of knowing: the integration of clinical expertise, evidence-based medicine, and artificial intelligence in assisted reproductive technologies. J Assist Reprod Genet 2021;38(7):1617-1625. [doi: 10.1007/s10815-021-02159-4] [Medline: 33870475]
35. Rutvij D, Gupta S, Mishra A. Artificial intelligence in assisted reproductive technology: present and future. Int J Infertil Fetal Med 2020;11(3). [doi: 10.5005/jp-journals-10016-1208]
36. Riegler M, Stensen MH, Witczak O, Andersen JM, Hicks SA, Hammer HL, et al. Artificial intelligence in the fertility clinic: status, pitfalls and possibilities. Hum Reprod 2021;36(9):2429-2442. [doi: 10.1093/humrep/deab168] [Medline: 34324672]
37. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
38. Higgins J, Deeks J. Selecting studies and collecting data. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. New York: John Wiley & Sons; 2008.
39. Barucic D. Automatic evaluation of human oocyte developmental potential from microscopy images. 2021 Presented at: 17th International Symposium on Medical Information Processing and Analysis; 2021; Campinas, Brazil p. 12088. [doi: 10.1117/12.2604010]
40. Borup R, Thuesen LL, Andersen CY, Nyboe-Andersen A, Ziebe S, Winther O, et al. Competence classification of cumulus and granulosa cell transcriptome in embryos matched by morphology and female age. PLoS One 2016;11(4):e0153562. [doi: 10.1371/journal.pone.0153562] [Medline: 27128483]
41. Brás de Guimarães B, Martins L, Metello JL, Ferreira FL, Ferreira P, Fonseca JM. Application of artificial intelligence algorithms to estimate the success rate in medically assisted procreation. Reprod Med 2020;1(3):181-194. [doi: 10.3390/reprodmed1030014]
42. Cao M, Liu Z, Lin Y, Luo Y, Li S, Huang Q, et al. A personalized management approach of OHSS: development of a multiphase prediction model and smartphone-based app. Front Endocrinol (Lausanne) 2022;13:911225 [FREE Full text] [doi: 10.3389/fendo.2022.911225] [Medline: 35872996]
43. Correa N, Cerquides J, Arcos JL, Vassena R. Supporting first FSH dosage for ovarian stimulation with machine learning. Reprod Biomed Online 2022;45(5):1039-1045. [doi: 10.1016/j.rbmo.2022.06.010] [Medline: 35915001]
44. Fanton M, Nutting V, Rothman A, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation. Reprod Biomed Online 2022;45(6):1152-1159 [FREE Full text] [doi: 10.1016/j.rbmo.2022.07.010] [Medline: 36096871]
45. Fanton M, Nutting V, Solano F, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation. Fertil Steril 2022;118(1):101-108 [FREE Full text] [doi: 10.1016/j.fertnstert.2022.04.003] [Medline: 35589417]
46. Fragoulakis V, Mantis A, Christoforidis N, Dovas D, Deftereos S, Lainas T, et al. Follitropin alpha for assisted reproduction: an analysis based on a non-interventional study in Greece. Curr Med Res Opin 2022;38(12):2227-2235. [doi: 10.1080/03007995.2022.2131303] [Medline: 36193626]

XSL•FO
RenderX

47. Fu K, Li Y, Lv H, Wu W, Song J, Xu J. Development of a model predicting the outcome of fertilization cycles by a robust decision tree method. Front Endocrinol (Lausanne) 2022;13:877518 [FREE Full text] [doi: 10.3389/fendo.2022.877518] [Medline: 36093079]
48. Hariton E, Chi EA, Chi G, Morris JR, Braatz J, Rajpurkar P, et al. A machine learning algorithm can optimize the day of trigger to improve in vitro fertilization outcomes. Fertil Steril 2021;116(5):1227-1235 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.06.018] [Medline: 34256948]
49. Hua L, Zhe Y, Jing Y, Fujin S, Jiao C, Liu L. Prediction model of gonadotropin starting dose and its clinical application in controlled ovarian stimulation. BMC Pregnancy Childbirth 2022;22(1):810 [FREE Full text] [doi: 10.1186/s12884-022-05152-6] [Medline: 36333671]
50. Kashiwaki R, Ishikawa T, Matubayashi H. Classification of follicles by ultrasonic images. 2018 Presented at: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC); October 10, 2018; Miyazaki, Japan p. 368-371. [doi: 10.1109/smc.2018.00072]
51. Letterie G, Mac Donald A. Artificial intelligence in in vitro fertilization: a computer decision support system for day-to-day management of ovarian stimulation during in vitro fertilization. Fertil Steril 2020;114(5):1026-1031 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.06.006] [Medline: 33012555]
52. Letterie G, MacDonald A, Shi Z. An artificial intelligence platform to optimize workflow during ovarian stimulation and IVF: process improvement and outcome-based predictions. Reprod Biomed Online 2022;44(2):254-260. [doi: 10.1016/j.rbmo.2021.10.006] [Medline: 34865998]
53. Liang X, Liang J, Zeng F, Lin Y, Li Y, Cai K, et al. Evaluation of oocyte maturity using artificial intelligence quantification of follicle volume biomarker by three-dimensional ultrasound. Reprod Biomed Online 2022;45(6):1197-1206 [FREE Full text] [doi: 10.1016/j.rbmo.2022.07.012] [Medline: 36075848]
54. Liu L, Shen F, Liang H, Yang Z, Yang J, Chen J. Machine learning-based modeling of ovarian response and the quantitative evaluation of comprehensive impact features. Diagnostics (Basel) 2022;12(2) [FREE Full text] [doi: 10.3390/diagnostics12020492] [Medline: 35204580]
55. Ma C, Xu H, Wang H, Feng G, Han Y, Alpadi K, et al. An online tool for predicting ovarian responses in unselected patients using dynamic inhibin B and basal antimüllerian hormone levels. Front Endocrinol (Lausanne) 2023;14:1074347 [FREE Full text] [doi: 10.3389/fendo.2023.1074347] [Medline: 36742391]
56. O'Gorman A, Wallace M, Cottell E, Gibney MJ, McAuliffe FM, Wingfield M, et al. Metabolic profiling of human follicular fluid identifies potential biomarkers of oocyte developmental competence. Reproduction 2013;146(4):389-395. [doi: 10.1530/REP-13-0184] [Medline: 23886995]
57. Robertson I, Chmiel FP, Cheong Y. Streamlining follicular monitoring during controlled ovarian stimulation: a data-driven approach to efficient IVF care in the new era of social distancing. Hum Reprod 2021;36(1):99-106 [FREE Full text] [doi: 10.1093/humrep/deaa251] [Medline: 33147345]
58. Sadruddin S, Barnett B, Ku L, Havemann D, Mucowski S, Herrington R, et al. Maternal serum concentration of anti-Müllerian hormone is a better predictor than basal follicle stimulating hormone of successful blastocysts development during IVF treatment. PLoS One 2020;15(10):e0239779 [FREE Full text] [doi: 10.1371/journal.pone.0239779] [Medline: 33044971]
59. Shi Y, Miao B, Ai X, Cao P, Gao J, Xu Y, et al. Identification of common genetic polymorphisms associated with down-regulated gonadotropin levels in an exome-wide association study. Fertil Steril 2023;120(3 Pt 2):671-681. [doi: 10.1016/j.fertnstert.2023.03.031] [Medline: 37001689]
60. Simopoulou M, Galatis D, Maziotis E, Pantou A, Giannelou P, Grigoriadis S, et al. Development of a predictive model for luteal phase oocyte retrieval in poor responders undergoing natural cycle IVF. Sci Rep 2022;12(1):7695 [FREE Full text] [doi: 10.1038/s41598-022-11602-0] [Medline: 35545640]
61. Srivastava D, Gupta S, Kudavelly S, Ga R. Unsupervised deep learning based longitudinal follicular growth tracking during IVF cycle using 3D transvaginal ultrasound in assisted reproduction. 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); November 1-5, 2021; Mexico p. 3209-3212. [doi: 10.1109/EMBC46164.2021.9630495]
62. Thomas N, Goodacre R, Timmins EM, Gaudoin M, Fleming R. Fourier transform infrared spectroscopy of follicular fluids from large and small antral follicles. Hum Reprod 2000;15(8):1667-1671. [doi: 10.1093/humrep/15.8.1667] [Medline: 10920083]
63. Tikhaeva K. Gradient boosting predictive model of ovarian response for hormonal therapy in infertility treatment. In: Russian Federation. 2021 Presented at: Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference; 2021; Russia.
64. Wei J, Xiong D, Zhang Y, Zeng J, Liu W, Ye F. Predicting ovarian responses to the controlled ovarian hyperstimulation in elderly infertile women using clinical measurements and random forest regression. Eur J Obstet Gynecol Reprod Biol 2023;288:153-159 [FREE Full text] [doi: 10.1016/j.ejogrb.2023.07.012] [Medline: 37544248]
65. Xu H, Feng G, Wang H, Han Y, Yang R, Song Y, et al. A novel mathematical model of true ovarian reserve assessment based on predicted probability of poor ovarian response: a retrospective cohort study. J Assist Reprod Genet 2020;37(4):963-972 [FREE Full text] [doi: 10.1007/s10815-020-01700-1] [Medline: 32318905]

XSL•FO
RenderX

66. Yan S, Jin W, Ding J, Yin T, Zhang Y, Yang J. Machine-intelligence for developing a potent signature to predict ovarian response to tailor assisted reproduction technology. Aging (Albany NY) 2021;13(13):17137-17154 [FREE Full text] [doi: 10.18632/aging.203032] [Medline: 33999860]
67. Zhu H, Zhao C, Xiao P, Zhang S. Predicting the likelihood of live birth in assisted reproductive technology according to the number of oocytes retrieved and female age using a generalized additive model: a retrospective cohort analysis of 17,948 cycles. Front Endocrinol (Lausanne) 2021;12:606231 [FREE Full text] [doi: 10.3389/fendo.2021.606231] [Medline: 33995268]
68. Zieliński K, Pukszta S, Mickiewicz M, Kotlarz M, Wygocki P, Zieleń M, et al. Personalized prediction of the secondary oocytes number after ovarian stimulation: a machine learning model based on clinical and genetic data. PLoS Comput Biol 2023;19(4):e1011020 [FREE Full text] [doi: 10.1371/journal.pcbi.1011020] [Medline: 37104276]
69. Lebovitz O, Haas J, Mor N, Zilberberg E, Aizer A, Kirshenbaum M, et al. Predicting IVF outcome in poor ovarian responders. BMC Womens Health 2022;22(1):395 [FREE Full text] [doi: 10.1186/s12905-022-01964-y] [Medline: 36180882]
70. Adebayo FO, Ameh N, Adesiyun AG, Ekele BA, Wada I. Correlation of female age with outcome of IVF in a low-resource setting. Int J Gynaecol Obstet 2023;161(1):283-288. [doi: 10.1002/ijgo.14545] [Medline: 36330818]
71. Hsieh H, Su J, Wang S, Huang Y. Age effect on in vitro fertilization pregnancy mediated by anti-Mullerian hormone (AMH) and modified by follicle stimulating hormone (FSH). BMC Pregnancy Childbirth 2020;20(1):209 [FREE Full text] [doi: 10.1186/s12884-020-02875-2] [Medline: 32272911]
72. Amini P, Ramezanali F, Parchehbaf-Kashani M, Maroufizadeh S, Omani-Samani R, Ghaheri A. Factors associated with in vitro fertilization live birth outcome: a comparison of different classification methods. Int J Fertil Steril 2021;15(2):128-134 [FREE Full text] [doi: 10.22074/IJFS.2020.134582] [Medline: 33687166]
73. Singh N, Kashyap A, Patel G, Vatsa R. Using GnRH agonists in combination with hCG in antagonist ART cycles: a scoping review of recent evidences. JBRA Assist Reprod 2023;27(4):717-726 [FREE Full text] [doi: 10.5935/1518-0557.20230029] [Medline: 37579268]
74. Kadoura S, Alhalabi M, Nattouf AH. Conventional GnRH antagonist protocols versus long GnRH agonist protocol in IVF/ICSI cycles of polycystic ovary syndrome women: a systematic review and meta-analysis. Sci Rep 2022;12(1):4456 [FREE Full text] [doi: 10.1038/s41598-022-08400-z] [Medline: 35292717]
75. He F, Hu W, Yong L, Li Y. Triggering of ovulation for GnRH-antagonist cycles in normal and low ovarian responders undergoing IVF/ICSI: a systematic review and meta-analysis of randomized trials. Eur J Obstet Gynecol Reprod Biol 2023;289:65-73 [FREE Full text] [doi: 10.1016/j.ejogrb.2023.08.014] [Medline: 37639817]
76. Hebisha SA, Aboelazm BA, Sallam HN. GnRH antagonist cetrorelix administration before hCG for protection of ovarian hyperstimulation syndrome. J Obstet Gynaecol India 2017;67(4):270-274 [FREE Full text] [doi: 10.1007/s13224-016-0952-5] [Medline: 28706366]
77. Boser E, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. 1992 Presented at: Proceedings of the 5th Annual Workshop on Computational Learning Theory; July 1, 1992; New York. [doi: 10.1145/130385.130401]
78. Jamil M, Debbarh H, Kabit A, Ennaji M, Zarqaoui M, Senhaji WR, et al. Impact of the number of retrieved oocytes on IVF outcomes: oocyte maturation, fertilization, embryo quality and implantation rate. Zygote 2023;31(1):91-96. [doi: 10.1017/S096719942200065X] [Medline: 36533391]
79. Doroftei B, Ilie O, Anton N, Marcu O, Scripcariu I, Ilea C. A narrative review discussing the efficiency of personalized dosing algorithm of follitropin delta for ovarian stimulation and the reproductive and clinical outcomes. Diagnostics (Basel) 2023;13(2) [FREE Full text] [doi: 10.3390/diagnostics13020177] [Medline: 36672987]
80. Mol BW, Bossuyt PM, Sunkara SK, Garcia Velasco JA, Venetis C, Sakkas D, et al. Personalized ovarian stimulation for assisted reproductive technology: study design considerations to move from hype to added value for patients. Fertil Steril 2018;109(6):968-979 [FREE Full text] [doi: 10.1016/j.fertnstert.2018.04.037] [Medline: 29935655]
81. Bila J, Dotlic J, Radjenovic SS, Vidakovic S, Tulic L, Micic J, et al. Predictive value of basal serum progesterone for successful IVF in endometriosis patients: the need for a personalized approach. J Pers Med 2022;12(10) [FREE Full text] [doi: 10.3390/jpm12101639] [Medline: 36294778]
82. Heindryckx B, De Gheselle S, Lierman S, Gerris J, De Sutter P. Efficiency of polarized microscopy as a predictive tool for human oocyte quality. Hum Reprod 2011;26(3):535-544. [doi: 10.1093/humrep/deq376] [Medline: 21233107]
83. Edwards R. Oocyte and follicle quality and success with IVF. Reprod Biomed Online 2002;3(3):263-264. [doi: 10.1016/s1472-6483(10)62046-9] [Medline: 12513865]
84. Data research. Human Fertilisation and Embryology Authority. 2023. URL: https://www.hfea.gov.uk/about-us/data-research/ [accessed 2023-09-05]
85. Ellahham S, Ellahham N, Simsekler MCE. Application of artificial intelligence in the health care safety context: opportunities and challenges. Am J Med Qual 2020;35(4):341-348. [doi: 10.1177/1062860619878515] [Medline: 31581790]
86. Fernandez EI, Ferreira AS, Cecílio MHM, Chéles DS, de Souza RCM, Nogueira MFG, et al. Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. J Assist Reprod Genet 2020;37(10):2359-2376 [FREE Full text] [doi: 10.1007/s10815-020-01881-9] [Medline: 32654105]

XSL•FO
RenderX

87. Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. J Assist Reprod Genet 2019;36(4):591-600 [FREE Full text] [doi: 10.1007/s10815-019-01408-x] [Medline: 30690654]
88. Swain J, VerMilyea MT, Meseguer M, Ezcurra D, Fertility AI Forum Group. AI in the treatment of fertility: key considerations. J Assist Reprod Genet 2020;37(11):2817-2824 [FREE Full text] [doi: 10.1007/s10815-020-01950-z] [Medline: 32989510]

Abbreviations

2PN: pronuclear stage
AI: artificial intelligence
AUC: area under the curve
GnRH: gonadotropin-releasing hormone
hCG: human chorionic gonadotropin
HFEA: Human Fertilisation and Embryology Authority
ICSI: intracytoplasmic sperm injection
IVF: in vitro fertilization
MII: metaphase II
ML: machine learning
MRI: magnetic resonance imaging
NPV: negative predictive value
OHSS: ovarian hyperstimulation syndrome
PPV: positive predictive value
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SVM: support vector machine
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 05.10.23; peer-reviewed by G Letterie, G Gullo, S Ma, T Shahzad, A ur Rehman; comments to author 19.03.24; revised version received 08.04.24; accepted 22.05.24; published 05.07.24.

Please cite as:
AlSaad R, Abd-alrazaq A, Choucair F, Ahmed A, Aziz S, Sheikh J
Harnessing Artificial Intelligence to Predict Ovarian Stimulation Outcomes in In Vitro Fertilization: Scoping Review
J Med Internet Res 2024;26:e53396
URL: https://www.jmir.org/2024/1/e53396
doi:10.2196/53396
PMID:38967964

©Rawan AlSaad, Alaa Abd-alrazaq, Fadi Choucair, Arfan Ahmed, Sarah Aziz, Javaid Sheikh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating and Enhancing Large Language Models' Performance in Domain-Specific Medicine: Development and Usability Study With DocOA

Xi Chen[1,2*], MD, PhD; Li Wang[1,2*], BMed; MingKe You[1,2], MMSc; WeiZhi Liu[1,2]; Yu Fu[3]; Jie Xu[4], DHM; Shaoting Zhang[4], PhD; Gang Chen[1,2], MD; Kang Li[4,5,6*], PhD; Jian Li[1,2*], MD

[1]Sports Medicine Center, West China Hospital, Sichuan University, Chengdu, China
[2]Department of Orthopedics and Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China
[3]West China Hospital, West China School of Medicine, Sichuan University, Chengdu, China
[4]Shanghai Artificial Intelligence Laboratory, OpenMedLab, Shanghai, China
[5]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
[6]Med-X Center for Informatics, Sichuan University, Chengdu, China
[*]these authors contributed equally

Corresponding Author:
Jian Li, MD
Sports Medicine Center, West China Hospital, Sichuan University
No. 37, Guoxue Alley, Wuhou District
Chengdu, 610041
China
Phone: 86 18980601388
Email: lijian_sportsmed@163.com

Abstract

Background: The efficacy of large language models (LLMs) in domain-specific medicine, particularly for managing complex diseases such as osteoarthritis (OA), remains largely unexplored.

Objective: This study focused on evaluating and enhancing the clinical capabilities and explainability of LLMs in specific domains, using OA management as a case study.

Methods: A domain-specific benchmark framework was developed to evaluate LLMs across a spectrum from domain-specific knowledge to clinical applications in real-world clinical scenarios. DocOA, a specialized LLM designed for OA management integrating retrieval-augmented generation and instructional prompts, was developed. It can identify the clinical evidence upon which its answers are based through retrieval-augmented generation, thereby demonstrating the explainability of those answers. The study compared the performance of GPT-3.5, GPT-4, and a specialized assistant, DocOA, using objective and human evaluations.

Results: Results showed that general LLMs such as GPT-3.5 and GPT-4 were less effective in the specialized domain of OA management, particularly in providing personalized treatment recommendations. However, DocOA showed significant improvements.

Conclusions: This study introduces a novel benchmark framework that assesses the domain-specific abilities of LLMs in multiple aspects, highlights the limitations of generalized LLMs in clinical contexts, and demonstrates the potential of tailored approaches for developing domain-specific medical LLMs.

(J Med Internet Res 2024;26:e58158) doi:10.2196/58158

KEYWORDS

large language model; retrieval-augmented generation; domain-specific benchmark framework; osteoarthritis management

XSL•FO
RenderX

Introduction

The rapid development of large language models (LLMs) has shown promising potential in the medical field, as demonstrated by their ability to pass the United States Medical Licensing Examination and diagnose clinical conditions [1-3]. The promising performance of LLMs in the general medical field warrants further research and exploration of their clinical performance in domain-specific medical scenarios [4,5].

Osteoarthritis (OA) is one of the most prevalent and debilitating diseases that causes pain, disability, and loss of function [6]. The global prevalence of OA is approximately 7.6% (595 million people) as of 2020 [7]. The management of OA requires complex strategies that encompass a variety of pharmacological treatments, lifestyle alterations, rehabilitation, and surgical interventions across multiple disciplines. Effective management of this condition necessitates the integration of extensive evidence-based medical data and the consideration of individual circumstances [6].

Although some LLMs have achieved commendable results in general medical question-answer (QA) tasks, substantial limitations persist in their clinical capability, particularly in complex and multifaceted diseases such as OA [8]. However, the data sets used to train LLMs are predominantly composed of general medical knowledge and lack in-depth, domain-specific content. Existing research indicates that current training data and benchmarking methodologies may be inadequate for LLMs to acquire the necessary domain-specific knowledge and clinical capabilities [4].

Additionally, LLMs may lack the ability to translate their knowledge into clinical proficiency. Despite possessing sound knowledge about certain diseases, effectively applying this knowledge to disease diagnosis remains challenging for LLMs [9]. This observation highlights the need to train and evaluate LLMs using data sets that are more closely aligned with clinical applications, thereby bridging the gap between theoretical knowledge and practical clinical usage.

To address these challenges, we proposed to build a data set that focuses on specific medical diseases, which should encompass updated evidence-based medical knowledge capable of providing both physicians and patients with expert disease-related information. In addition, real-world cases featuring patient information and treatment decisions encountered in clinical practice should be included. This repository can serve as a benchmark for testing the performance of LLMs in specific medical domains, such as OA management.

Moreover, despite the fact that LLMs have demonstrated impressive capabilities, their internal mechanisms remain unclear. This lack of transparency poses unnecessary risks to downstream applications [10], which is particularly crucial in the medical field and constitutes a significant ethical consideration [11]. Retrieval-augmented generation (RAG) offers a solution for explainability, as the RAG technique enables large models to identify the source of their answers when responding to questions. RAG is an AI framework that improves LLMs by integrating relevant information from external knowledge bases, thus enhancing the accuracy and reliability of the model's responses while also providing efficient and cost-effective access to updated external data [12]. Therefore, the integration of RAG and prompt engineering could enable the model to assimilate external knowledge bases and adhere to instructions to respond in a predetermined manner.

In general, we propose a data set framework that encompasses updated evidence-based medical knowledge, and real-world cases may effectively examine the capabilities of LLMs in clinical practice. The integration of RAG and prompt engineering may allow trained LLMs such as GPT-4 to acquire domain-specific abilities. Moreover, the management of OA serves as an ideal example in terms of its clinical significance and data volume on this research topic. Therefore, this study aimed to curate a data set for OA management, evaluate knowledge of updated evidence-based medicine for LLMs and their capabilities in clinical scenarios, and adopt RAG and instruction prompts to enhance these capabilities.

Methods

Overview

This study curated an OA management data set based on clinical guidelines and real-world cases. A benchmark was developed to evaluate the clinical knowledge and capabilities of LLMs for OA management. DocOA was built with instruction prompts and RAG and was tested along with other LLMs. Figure 1 illustrates the flow diagram of the study.

XSL•FO
RenderX

Figure 1. Study flow diagram. GIQA: guideline-item question-answer; MOQA: management options question-answer; QA: question-answer; real-case QA; TSQA: treatment strategy QA.

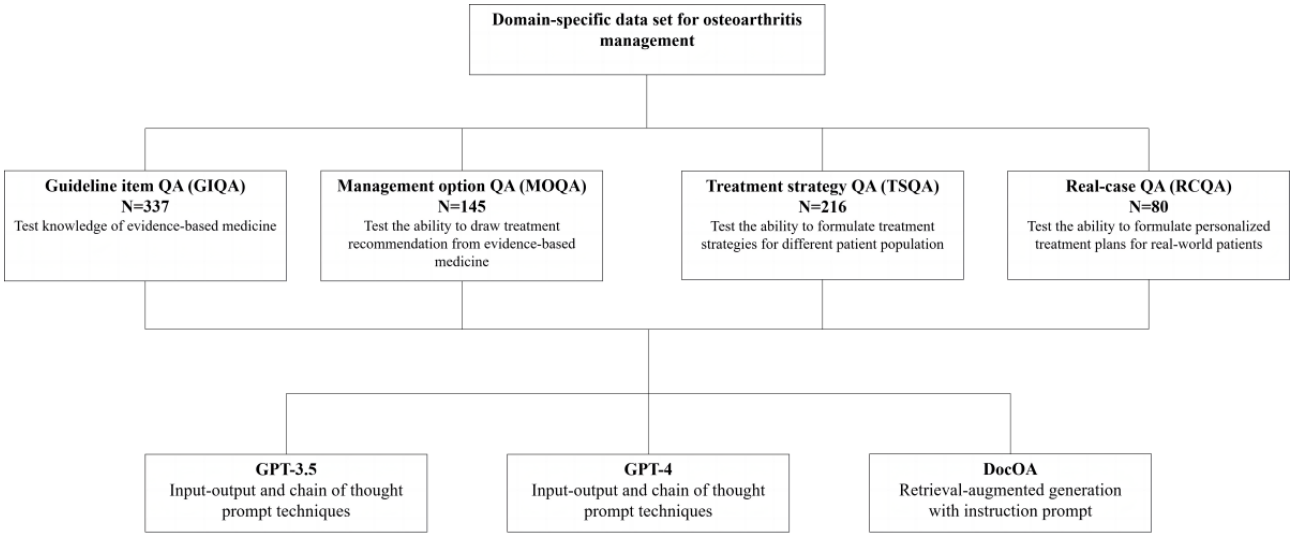

Data Set

This data set was developed based on key clinical guidelines and real-world patients. After the panel discussion, 6 well-acknowledged guidelines and data from 80 real-world patients were selected that included various aspects of OA management. The following guidelines were included: American Academy of Orthopedic Surgeons management of OA of the knee (Nonarthroplasty) [13]; NICE (National Institute for Health and Care Excellence) guideline for OA in over 16s [14]; Osteoarthritis Research Society International (OARSI) guidelines for the nonsurgical management of knee, hip, and polyarticular OA [15]; Royal Australian College of General Practitioners Guideline for the management of knee and hip OA [16]; American College of Rheumatology/Arthritis Foundation (ACR) Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee [17]; European League Against Rheumatism (EULAR) recommendations for the nonpharmacological core management of hip and knee OA [18]. Between April 1, 2023, and October 1, 2023, a total of 80 patients diagnosed with OA and who had received OA management at our hospital were randomly selected. The patient information, including age, sex, height, weight, BMI, laterality of knee involvement, medical history, level of pain, mechanical symptoms, physical examination results, and radiographic findings, was retrieved. All identifiable information was concealed to maintain confidentiality.

The OA benchmark aims to test the clinical capabilities of LLMs at 4 levels within the context of evidence-based medicine, ranging from domain-specific knowledge to clinical capabilities. The benchmark assesses the performance of LLMs pertaining to OA knowledge, summarizing the knowledge to formulate recommendations for specific management options, providing tailored management options for different patient populations, and formulating personalized management plans for real-world cases.

Assistant With RAG and Instruction Prompting

DocOA, a specialized assistant, was developed based on the GPT-4-1106-preview model, which integrates instruction prompts and RAG to enhance performance. The instruction prompt emphasized its role in providing evidence-based medical insights and personalized management programs guided by evidence-based medicine. The DocOA strictly adheres to facts, avoids speculation, and clearly states its limitations. Moreover, it maintains a professional and informative tone suitable for medical discussions.

RAG has been used to respond to various OA-related queries. The RAG integrates a model's language generation capabilities with a retrieval system, enabling access to specific information from external sources [19]. Of the several RAG techniques and data structures tested, the retrieval function from OpenAI was adopted, and the most optimal data structure was selected and converted into the JSON format for optimal retrieval accuracy. In response to OA-related queries, the RAG enables the assistant to dynamically pull relevant data from the external data set as it generates responses. The workflow of the assistant is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Workflow.

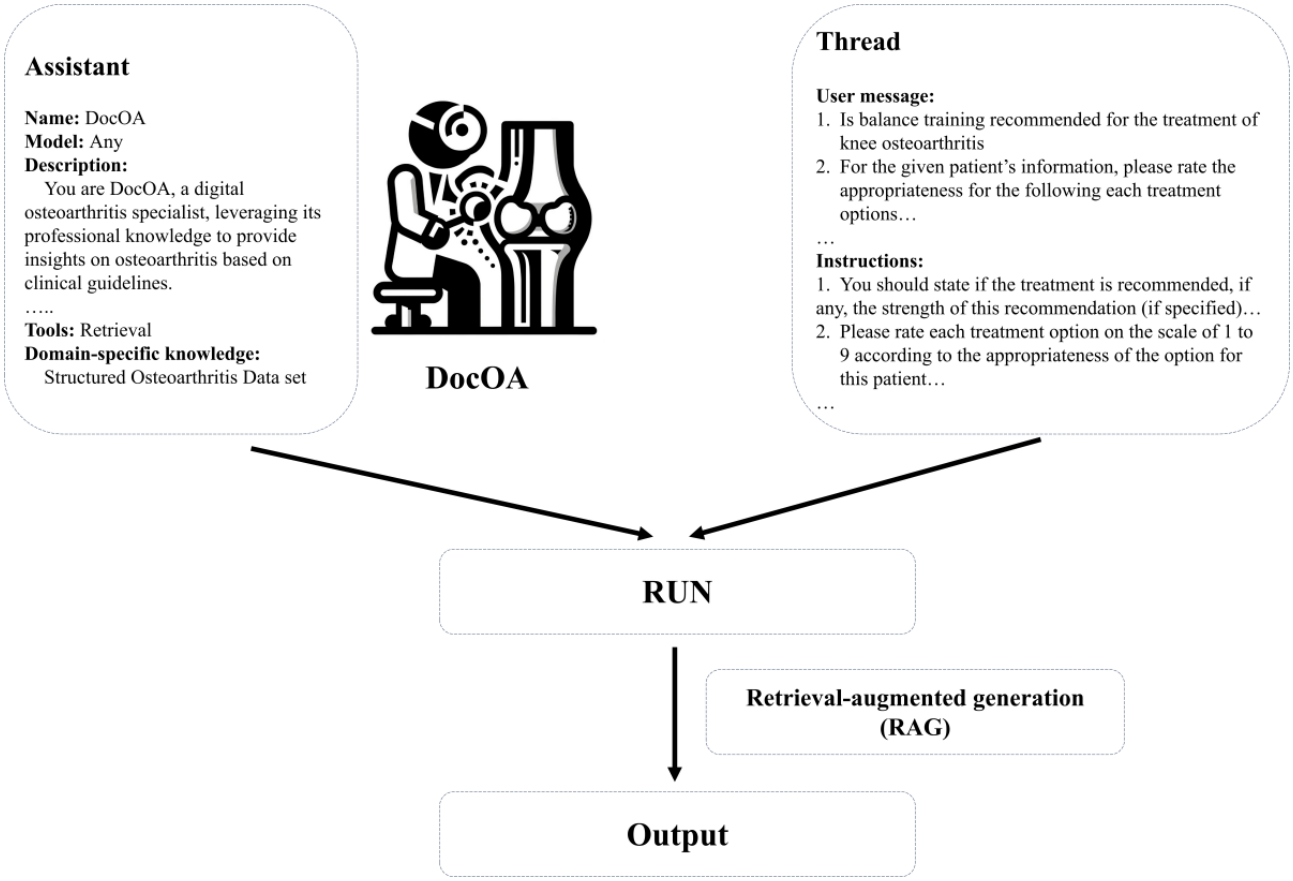

The assistant block details the core of the DocOA system, using a base model of GPT-4-1106-preview. The description within the block serves as an instruction prompt that outlines the system's role. The assistant's functionality includes a retrieval tool that accesses external data set.

The system receives input from thread blocks in the form of questions about OA management and requests evaluation. This includes user messages with specific queries regarding OA treatment and detailed instructions for the system to follow.

The execution command then triggers the processing of input data through the DocOA system. Through RAG, DocOA incorporates external information from the knowledge database and follows the instructions to generate the final output. Simultaneously, based on instructional prompts, DocOA is capable of providing final outputs tailored to the user, whether they are patients or professional doctors, in the corresponding style.

Models Testing

DocOA and the 2 base models, GPT-3.5 and GPT-4, were tested against the OA benchmark. Each question was presented 5 times to each model to assess the robustness of its performance. Additionally, the zero-shot chain of thoughts prompt technique was tested for GPT-3.5 and GPT-4 to determine whether it outperformed the input-output technique.

Evaluation of LLMs' Performance

Objective Evaluation

The model-generated responses were compared with predefined correct answers for each subset of the benchmark. An answer was considered accurate if LLM provided correct knowledge (recommendation status and recommendation strength) about the treatment option and predicted the correct treatment recommendation (treatment appropriateness) for a specific patient profile or individual patient.

The human evaluation framework is an effective approach for identifying the gap between LLMs and clinical experts [3]. In this study, human evaluation was performed by both physicians and patients. A total of 80 items from the OA benchmark were randomly selected for a detailed human evaluation framework.

A panel of 5 experienced physicians evaluated the outputs from LLMs. The sequence of answers was randomized and the generating models were anonymized to ensure that the evaluation was conducted without any knowledge of the model that generated them. Specifically, each evaluator independently assessed the sampled responses. All sampled responses were compiled into an Excel (Microsoft Corp) spreadsheet. The model labels corresponding to the responses were concealed. The evaluation metrics were established based on a previous study with modifications [3]. The physician assessed the quality of the responses in the following domains: inaccurate content, relevance, hallucinations, missing content, likelihood of possible harm, extent of possible harm, and possibility of bias. The ability of LLM to achieve correct comprehension, retrieval, and reasoning was assessed using the method described in a previous study [20]. Patient evaluation was conducted by assessing the

XSL•FO
RenderX

user intent fulfillment and helpfulness of the content. The detailed descriptions of each human evaluation metric are provided in Multimedia Appendix 1.

Statistical Analysis

All statistical analyses were performed using the SPSS 25.0 software (IBM Corp) and GraphPad Prism (version 8; GraphPad Software). Discontinuous data are expressed as incidence and rate and analyzed using the chi-square test for differences. A *P* value less than .05 indicated statistical significance.

Ethical Considerations

Real-case QA (RCQA) includes treatment recommendations for 80 real cases. The study have obtained approval from the Medical Ethics Committee of Sichuan University (approval number: 2023; review number 2277). Participants were informed that the study would need to use relevant data including clinical symptoms, imaging reports and physical examination data, but they would not receive any intervention measures and their personal identifiable information was anonymized. This study does not involve compensation.

Results

OA Benchmark

The benchmark comprised 4 subsets of QA evaluations designed to test the performance of LLMs across a spectrum ranging from domain-specific knowledge to practical capability. Guideline-item QA (GIQA), which was developed based on specific items extracted from the clinical guidelines, evaluates the LLMs' knowledge of these well-established standards. The GIQA comprised 337 items. Management options QA (MOQA) included summarized recommendations for specific treatments from the included clinical guidelines. The MOQA, which comprised 145 items, evaluated LLMs' knowledge of specific treatment options as well as their ability to summarize medical evidence. Treatment strategy QA (TSQA), which included treatment recommendations for different patient populations, provided treatment recommendations based on the patient's age, clinical presentation, and other factors. The TSQA, which comprised 216 items, evaluated the capability of LLMs to derive treatment recommendations for specific patient types. RCQA included treatment recommendations for 80 real-world patients. The RCQA, which comprised 80 items, evaluated LLMs' capability in formulating treatment recommendations in a more complicated scenario in which individual information is provided, mirroring real-world clinical decision-making. This data set is available on GitHub [21]. Examples of each QA type are shown in Multimedia Appendix 2.

Objective Evaluation

The accuracy of GPT-3.5 in GIQA, MOQA, TSQA, and RCQA was 0.26, 0.22, 0.01, and 0.03, respectively. The accuracy of GPT-4 in GIQA, MOQA, TSQA, and RCQA was 0.38, 0.30, 0.07, and 0.01, respectively. The accuracy of DocOA in GIQA, MOQA, TSQA, and RCQA was 0.92, 0.87, 0.88, and 0.72, respectively. The accuracy of each model against the benchmark is presented in Table 1 and Figure 3A. As shown in Figure 3A, the degree of accuracy significantly decreased: GIQA MOQA TSQA RCQA. As shown in Figure 3B, DocOA reported 111 failures in accessing the external data set, which accounted for 12.4% of the inaccurate answers generated.

Figure 3. Results of objective evaluation. (A) Accuracy of each model against each subset of benchmark. (B) Inaccuracy analysis for DocOA due to wrong answer and failure to access external data. GIQA: guideline item question-answer (QA); MOQA: management option QA; TSQA: treatment strategy QA; RCQA: real-case QA.

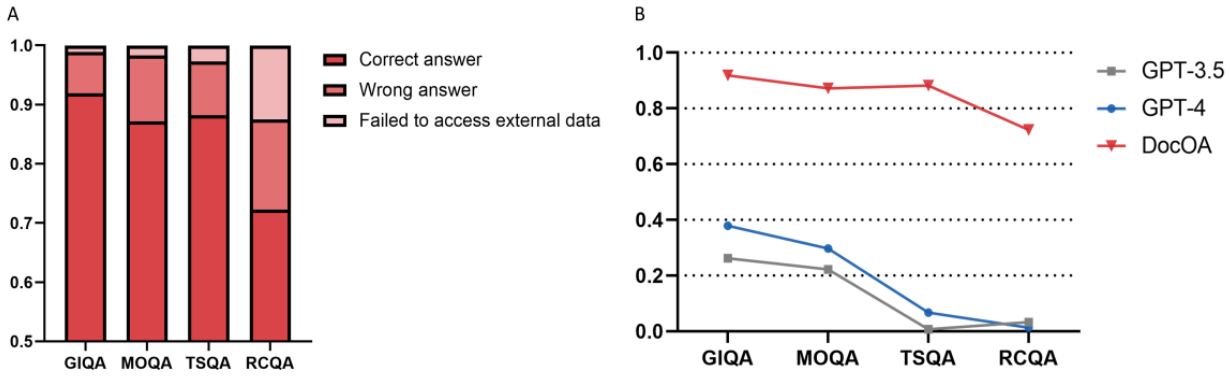

XSL•FO
RenderX

Table 1. Accuracy of each model against the osteoarthritis benchmark.

	GPT-3.5	GPT-4	DocOA	*P* value
Osteoarthritis benchmark	0.16	0.24	0.88	<.001[b]
Guideline item QA[a]	0.26	0.38	0.92	<.001[b]
Management option QA	0.22	0.30	0.87	<.001[b]
Treatment strategy QA	0.01	0.07	0.88	<.001[b]
Real-case QA	0.03	0.01	0.72	<.001[c]

[a]QA: question-answer.

[b]Further analysis showed all pairwise comparisons had a *P* value less than .05.

[c]Further analysis showed a *P* value is .056 for GPT-3.5 vs GPT-4, <.001 for GPT-3.5 vs DocOA, and <.001 for GPT-4 vs DocOA.

Zero-shot chain of thoughts prompt techniques were adopted for GPT-3.5 and GPT-4. Compared with the input-output prompt technique, no significant improvements in model performance were observed. The results are summarized in Table 2.

Table 2. Accuracy of different prompt techniques against the osteoarthritis benchmark.

	GPT-3.5			GPT-4		
	IO[a]	COT[b]	*P* value	IO	COT	*P* value
Osteoarthritis benchmark	0.16	0.17	.41	0.24	0.23	.52
Guideline item QA[c]	0.26	0.28	.03	0.38	0.38	.80
Management option QA	0.22	0.20	.004	0.30	0.27	.002
Treatment strategy QA	0.02	0.03	<.001	0.07	0.07	.79
Real-case QA	0.03	0.01	<.001	0.01	0.01	.20

[a]Input-output prompt technique.

[b]Zero-shot chain of thought prompt technique.

[c]QA: question-answer.

Human Evaluation Results

From each of the GIQA, MOQA, TSQA, and RCQA, 20 items were randomly selected along with the corresponding responses generated by each model. A total of 1200 outputs were evaluated by physicians and patients. The results of the human evaluations of GPT-3.5, GPT-4, and DocOA revealed distinct outcomes across several aspects. A few examples have been selected and summarized in Multimedia Appendix 3, demonstrating how to apply the human evaluation framework to assess sampled responses.

Multimedia Appendix 4 shows the human evaluation results for the models' output. The rate of inaccuracy was the highest for GPT-3.5 (57%, n=1200), followed by GPT-4 and DocOA at 50% (n=1200) and 19.3% (n=1200), respectively. All the models achieved high relevance and infrequently produced hallucinatory content in their responses. GPT-3.5 had a higher proportion of responses with missing content (22%, n=1200) than GPT-4 (16.4%, n=1200) or DocOA (16.5%, n=1200). GPT-3.5 presented a higher likelihood of generating harmful content (20%, n=1200) than GPT-4 (11.3%, n=1200) and DocOA (8.3%, n=1200). Moreover, GPT-3.5 was associated with a higher risk of causing severe harm (10.5%, n=1200) than GPT-4 (5.5%, n=1200) and DocOA (3.5%, n=1200). The likelihoods of potentially biased content were 13.3% (n=1200), 9.5% (n=1200), and 2.8% (n=1200) for GPT-3.5, GPT-4, and DocOA, respectively. The results of the human evaluation for each subset benchmark are listed in Multimedia Appendices 5-7. The results showed a substantial decrease in performance in terms of inaccurate content and missing content (GIQA MOQA TSQA RCQA).

Figure 4 shows the results of the LLMs' ability to assess correct comprehension, correct retrieval, and correct reasoning. Regarding the correct comprehension of the question, the response rate of DocOA was 91% (n=1200), followed by GPT-4 (86%, n=1200) and GPT-3.5 (82.5%, n=1200). DocOA was able to correctly recall and present complete, relevant information in 65.8% (n=1200) of the responses, followed by GPT-4 (14.3%, n=1200) and GPT-3.5 (12.0%, n=1200). In terms of subset evaluation, the results showed comparable performance in comprehension and reasoning among the different models, whereas a substantial performance decrease was found in correct retrieval across GIQA, MOQA, TSQA, and RCQA.

XSL•FO
RenderX

Figure 4. Results of human evaluation for LLMs' comprehension, retrieval, and reasoning ability. (A) Evidence of correct comprehension. (B) Evidence of correct retrieval. (C) Evidence of correct reasoning.

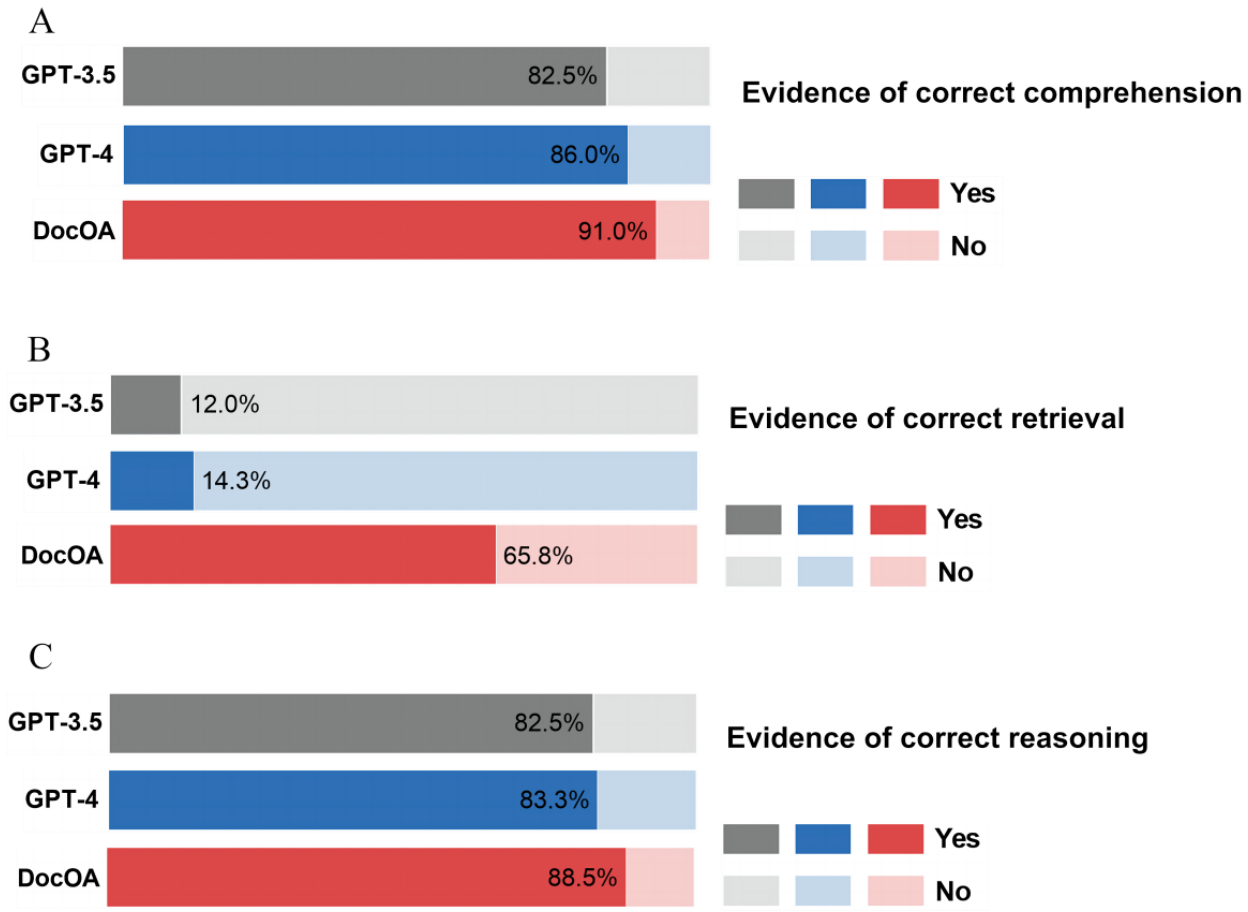

The results of patient evaluations are shown in Figure 5. DocOA achieved a success rate of 71.3% in fulfilling patient intention, with GPT-4 at 39.8% (n=1200) and GPT-3.5 at 36.5% (n=1200). Of the responses generated by DocOA, 75.8% (n=1200) were considered to be at least somewhat helpful, compared to 47% (n=1200) for GPT-3.5 and 47.75% (n=1200) for GPT-4. For GPT-3.5 and GPT-4, the subset evaluation showed a substantial decrease in intent fulfillment and helpfulness as the tasks shifted from domain-specific knowledge to personalized treatment recommendations.

Figure 5. Results of human evaluation by patients. (A) Patient intent fulfillment. (B) Patient helpfulness.

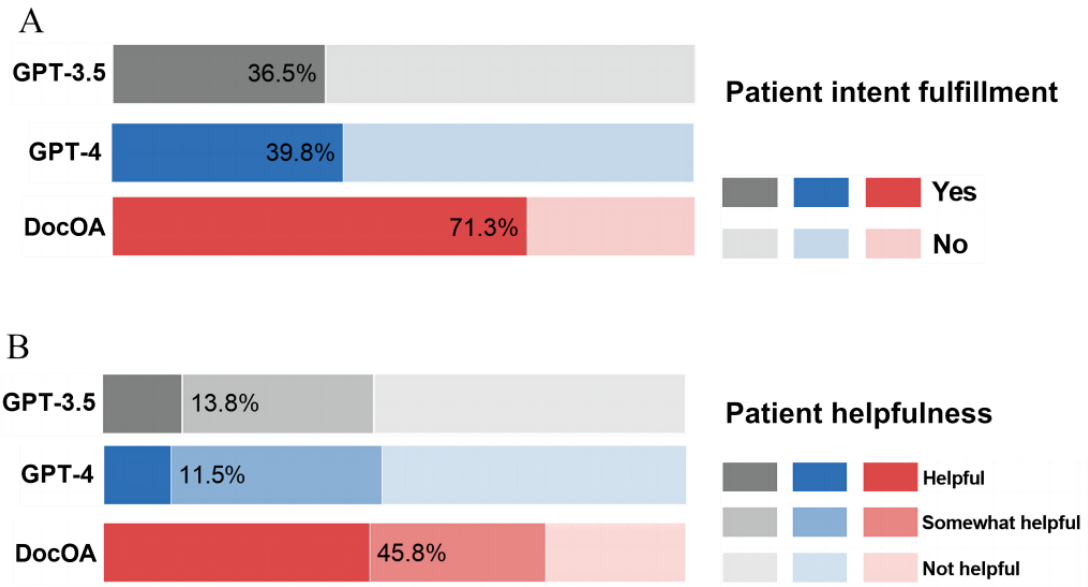

XSL•FO
RenderX

Discussion

This study introduced a benchmark framework to assess the performance of LLMs in specific medical domains. Using OA as a case study, this framework is the first to evaluate LLMs across a spectrum, from domain-specific knowledge to clinical applications in specific disease management. The incorporation of human evaluation provides multiple dimensions of assessment that are of considerable interest to clinical practitioners and patients, making it an essential tool for evaluating the clinical capabilities of LLMs.

The study found that the integration of RAG and instructional prompts substantially improved the domain-specific capabilities and explainability of general-purpose LLMs without additional training. By contrast, general-purpose models such as GPT-3.5 and GPT-4 exhibit unsatisfactory performance when benchmarked against OA management. Additionally, both models demonstrated a marked decline in performance as tasks shifted from domain-specific knowledge to personalized treatment recommendations. Overall, the findings of this study demonstrate a cost-effective method for evaluating and enhancing the capabilities of LLMs in specialized medical fields.

Domain-Specific Medical Benchmark

Although benchmarks targeting general medical knowledge have been previously developed [22-24], recent research has suggested that these benchmarks are only preliminary indicators of medical knowledge. The absence of tailored benchmarks in specific domains remains a potential challenge for evaluating the clinical effectiveness of LLMs [4,25]. Therefore, we developed a domain-specific benchmark focused on disease management for OA, which was selected for its prevalence, substantial disease burden, and complexity of its management strategies [6,7]. This benchmark was designed to test the domain-specific knowledge and clinical capabilities of LLMs. The benchmark comprised 4 parts, each testing the ability of LLMs at different levels, including the ability to provide evidence-based knowledge, summarizing knowledge to formulate recommendations, providing management recommendations for different patient populations, and formulating personalized management plans for real-world patients. The benchmark was constructed based on established clinical guidelines and real-world patient information. Clinical guidelines offer comprehensive reviews of updated evidence and expert opinions, making them reliable sources of domain-specific medical knowledge. Through panel discussions involving physicians and data scientists, the questions were designed in a hybrid format, integrating both definitive and interpretative elements. Using this benchmark, we confirmed that general-purpose LLMs exhibit suboptimal performance in specialized domains. A significant performance gap was observed between domain-specific knowledge and clinical proficiency. This highlights the challenges faced by general-purpose LLMs in effectively applying specialized knowledge to clinical scenarios.

Human Evaluation Framework

Human evaluation is a crucial component in assessing the medical capabilities of LLMs and offers a multidimensional assessment of their clinical capabilities. In this study, the human evaluation framework was modified based on a previous study, and hallucinations and relevance were added as additional criteria [3,4]. The evaluation criteria included accuracy, relevance, hallucinations, omissions, potential harm, and biased content. Moreover, the performance of LLMs in question comprehension, information retrieval, and medical reasoning was evaluated, as these are crucial abilities in tailoring patient-specific treatment. Patient evaluations primarily determine how responses address the user's intent and helpfulness. Although previous studies indicate a notable gap between objective benchmarking and human evaluation, our findings reveal a smaller discrepancy [3]. This could be attributed to the different knowledge domains and designs of the QA structure in this benchmark. The results of our study suggest that the GPT-3.5, GPT-4, and DocOA performed well in terms of hallucinations, comprehension, reasoning, and relevance. DocOA outperformed the other models in terms of accurate information, correct retrieval, and helpfulness as perceived by patients. This indicates that although generalized models are proficient in some areas, they remain inadequate in delivering the qualified responses required in a clinical context.

Augmenting LLM With Domain-Specific Ability

Several techniques are available for developing medical LLMs, which primarily include integrating domain-specific knowledge during the training phase through techniques such as reinforcement learning with human feedback [26-28]. However, in this study, we focused on enhancing already-trained LLMs, such as GPT-4, by using a suite of techniques, including RAG and instruction-based prompts. Similar methodologies have been applied to the development of specialized LLMs for chemical domains [29].

This approach was adopted for the following reasons: first, augmenting an existing model such as GPT-4 is more cost-effective than training a new model from scratch; second, advanced general-purpose models have been trained on diverse data sets, providing a broad base of general knowledge that can be beneficial for understanding and contextualizing domain-specific information; third, techniques such as RAG and specialized prompting offer the convenience of being adjustable and refined over time, enabling easy adaptability to new evidence in the fast-evolving field of medicine.

Among these reasons, the role of RAG in this study needs to be emphasized. Explainability has always been a problem to be addressed in the application of LLMs, and it is also one of the ethical considerations in their application in the medical field. RAG enables the model to identify the source on which the generated answer was based, which significantly improves the explainability of the model. In this study, the knowledge base was structured in a way that allows RAG to identify the clinical evidence that the answer was based on. When DocOA generates an OA-relevant answer, it identifies the evidence on which the answer is based and makes it clear for professionals to evaluate the rationale and accuracy of its response. In general, provided with reliable and professional sources, RAG significantly improved the explainability and accuracy of LLMs in medical care.

XSL•FO
RenderX

Our results demonstrated that GPT-4 can effectively acquire domain-specific knowledge and clinical capabilities in the management of OA through a combination of approaches, including RAG and instruction prompts. This strategy can also be applied cost-effectively to other medical domains.

Nonetheless, the efficacy of the RAG is contingent upon factors such as the size and quality of the data, retrieval techniques used, and the underlying architecture of the LLM in use [30]. The evaluation results showed that DocOA was able to correctly understand the clinical question and avoid hallucinations. The fundamental reason for the errors stems from the inherent limitations of the RAG technology, where the retrieved information may sometimes be incomplete or may not adequately address the question.

In RAG, the query process is the key determinant of the information retrieved from the external knowledge base. In the query process, text is converted into an array of floating points. The entire array corresponds to a point in an n-dimensional space, which is known as the text vector, also referred to as an embedding [31,32]. The distance between vectors, which corresponds to semantic similarity, can be calculated. Therefore, by calculating semantic similarity, the query process retrieves the information from the external knowledge base that is considered most relevant to the query posted by the LLM. However, this process is sometimes uncertain because the most appropriate answer does not necessarily yield the highest similarity score. For example, in GIQA, the question is, “Is Walking aids, assistive technology and adaptations at home and at work recommended for Physical Treatment of Knee osteoarthritis according to 2013 EULAR recommendations for the nonpharmacological core management of hip and knee osteoarthritis?” For this question, out of the 5 responses generated by LLM, 4 incorrectly state “recommend” when the correct answer should be “may be used.” The error occurred because only a portion of the information was retrieved. However, neither the complete information nor the key information, which contains the answer, was retrieved. More specifically, if the retrieved information was “The frequent use of assistive technology and the high satisfaction rates with its use indicate that walking aids, assistive technology, and adaptations are important and useful for people with hip or knee OA.” The model might be misled to generate the incorrect answer.

Generally, enhancing the effectiveness of RAG is an ongoing effort in the field of computer science, and the effectiveness of RAG is likely to keep increasing with the rapid development of LLMs and RAG technologies.

The Potential Value of DocOA in Future Clinical Practice

DocOA has shown promising potential in clinical practice due to its proficient OA management knowledge and the ability to explain its answer. Previous studies have shown general-purpose LLMs to be limited in medical specialties such as nephrology, ophthalmology, and OA. The versatile and unverified training data could contain misleading medical information, which led to inaccurate answers. On the other hand, DocOA has shown good results in providing evidence-based OA management knowledge and formulating personalized treatment plans. To adopt DocOA in clinical practice, it is designed to answer patient queries and doctor queries differently, as shown in Figure 6. When generating answers for medical professionals, it is able to generate answers based on selected high-quality clinical evidence and accurately identify the corresponding references, thus providing explainable and evidence-based answers.

XSL•FO
RenderX

Figure 6. Examples of clinical practice. (A) How DocOA provides exercise guidance for patients with OA. (B) How DocOA provides doctors with clinical evidence for exercise guidance in OA, as well as the clinical evidence the answer is based on.

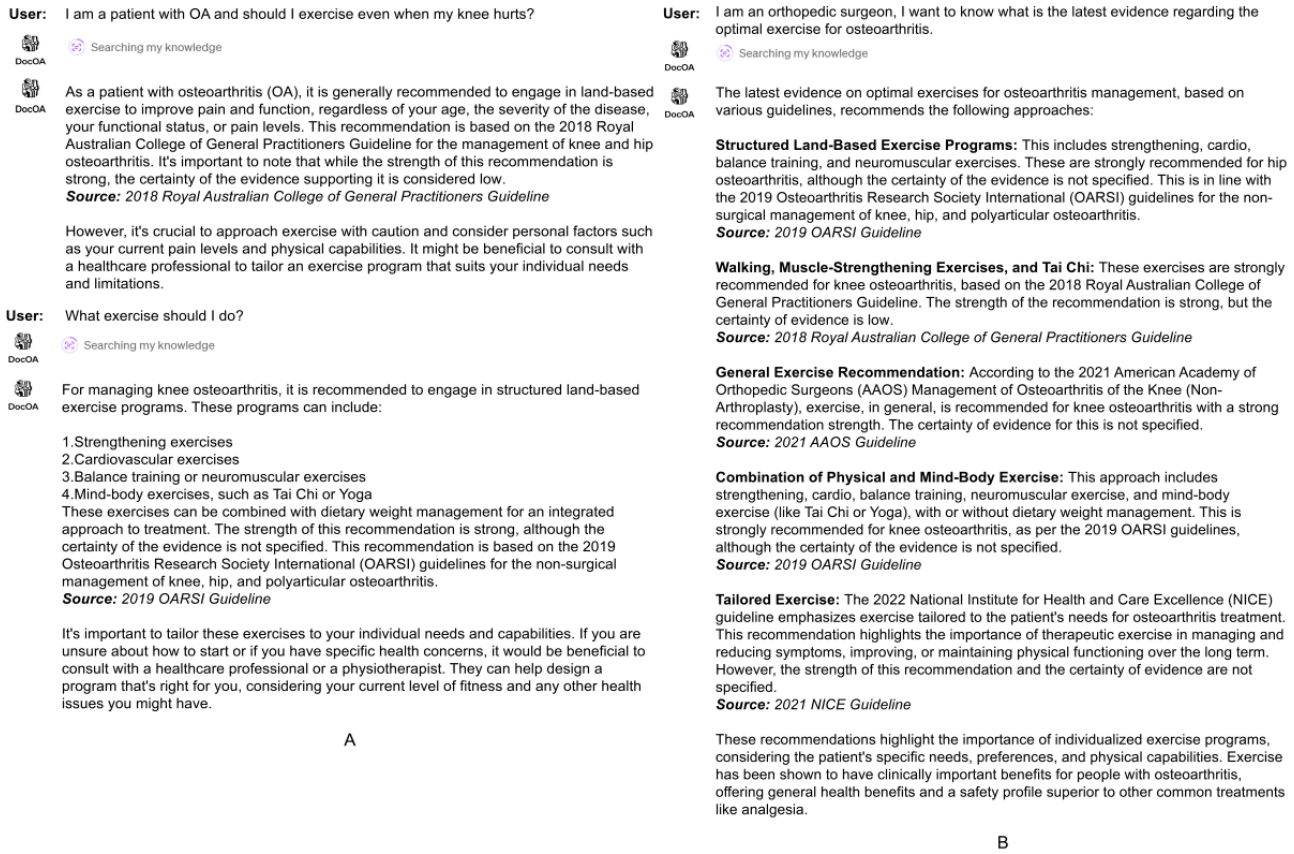

Limitations

This study has several potential limitations that need to be addressed. First, OA management is highly complex, and our current data set remains limited and requires continuous supplementation and updating. Therefore, establishing specialized groups dedicated to building and updating these LLM databases is imperative for diverse medical applications. Second, our reliance on English sources could restrict the applicability and inclusivity of our findings across different linguistic and cultural contexts. Third, although there have been human evaluations of clinical case data, the augmented model has not yet been tested in a real-world clinical setting. These limitations highlight the necessity for ongoing development and a comprehensive, multidimensional approach for evaluating LLMs in the medical field.

Conclusions

In this study, we introduce a novel benchmark framework designed to evaluate the capabilities of LLMs in specific medical domains, with OA serving as a case study. This framework assesses LLMs in terms of medical knowledge, evidence summarization, and clinical capabilities. Through a combination of objective measures and human evaluations, we identified the limitations of generalized LLMs in clinical contexts. Furthermore, our study demonstrated that DocOA, which integrates RAG and instructional prompts, significantly improves both the domain-specific performance and the explainability of LLMs. This approach is a potentially cost-effective strategy for developing domain-specific medical LLMs.

Acknowledgments

This study was funded by the Key Research and Development Fund of Sichuan Science and Technology Planning Department (grant 2022YFS0372; received by JL) and the Youth Research Fund of Sichuan Science and Technology Planning Department (grant 23NSFSC4894; received by XC).

Data Availability

The data sets generated during or analyzed during this study are available in the GitHub repository Osteoarthritis-Benchmark [21]. The data citation should include a persistent identifier (eg, web URL or DOI) cited per journal style in the reference list.

XSL•FO
RenderX

Authors' Contributions

XC and LW coordinated the research. JL and KL served as co-corresponding authors and co-guided the conduct of this study. JL was responsible for the clinical medicine related part, while KL was in charge of the computer science-related part. MKY, WZL, YF, and GC participated in the research design and improvement of the clinical medicine related part, while JX and SZ engaged in the research design and improvement of the computer science–related part.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of human evaluation framework.
[PDF File (Adobe PDF File), 104 KB - jmir_v26i1e58158_app1.pdf]

Multimedia Appendix 2
Benchmark framework for osteoarthritis management.
[PDF File (Adobe PDF File), 470 KB - jmir_v26i1e58158_app2.pdf]

Multimedia Appendix 3
Examples for human evaluation.
[PDF File (Adobe PDF File), 107 KB - jmir_v26i1e58158_app3.pdf]

Multimedia Appendix 4
Results of human evaluation for the assessment of responses.
[PDF File (Adobe PDF File), 136 KB - jmir_v26i1e58158_app4.pdf]

Multimedia Appendix 5
Human evaluation results for GPT-3.5 across GIQA, MOQA, TSQA and RCQA.
[PDF File (Adobe PDF File), 38 KB - jmir_v26i1e58158_app5.pdf]

Multimedia Appendix 6
Human evaluation results for GPT-4 across GIQA, MOQA, TSQA and RCQA.
[PDF File (Adobe PDF File), 81 KB - jmir_v26i1e58158_app6.pdf]

Multimedia Appendix 7
Human evaluation results for DocOA across GIQA, MOQA, TSQA and RCQA.
[PDF File (Adobe PDF File), 81 KB - jmir_v26i1e58158_app7.pdf]

References

1. Nori H, King N, McKinney SM, Carignan S, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375v1 2023:1-33.
2. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. J Am Med Assoc 2023;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin A, Park H, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. J Am Med Assoc 2023;330(9):818-820. [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
6. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019;393(10182):1745-1759. [doi: 10.1016/S0140-6736(19)30417-9] [Medline: 31034380]
7. GBD 2021 Osteoarthritis Collaborators. Global, regional, and national burden of osteoarthritis, 1990-2020 and projections to 2050: a systematic analysis for the global burden of disease study 2021. Lancet Rheumatol 2023;5(9):e508-e522 [FREE Full text] [doi: 10.1016/S2665-9913(23)00163-7] [Medline: 37675071]

XSL•FO
RenderX

8. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024;7(1):41 [FREE Full text] [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
9. Li J, Chen X, Liu W, Wang L, Guo Y, You M, et al. One is not enough: multi-agent conversation framework enhances rare disease diagnostic capabilities of large language models. Res Square 2023:1-16. [doi: 10.21203/rs.3.rs-3757148/v1]
10. Haiyan Z, Hanjie C, Fan Y, Ninghao L, Huiqi D, Hengyi C, et al. Explainability for Large Language Models: A Survey. arXiv:2309.01029 2023:1-38.
11. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
12. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS 2020:1-16.
13. Brophy RH, Fillingham YA. AAOS clinical practice guideline summary: management of osteoarthritis of the knee (nonarthroplasty), third edition. J Am Acad Orthop Surg 2022;30(9):e721-e729. [doi: 10.5435/jaaos-d-21-01233]
14. Health N, Excellence C. Osteoarthritis in Over 16s: Diagnosis and Management. France: NICE; 2022.
15. Bannuru RR, Osani MC, Vaysbrot EE, Arden NK, Bennell K, Bierma-Zeinstra SMA, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. Osteoarthritis Cartilage 2019;27(11):1578-1589 [FREE Full text] [doi: 10.1016/j.joca.2019.06.011] [Medline: 31278997]
16. Guideline for the Management of Knee and hip Osteoarthritis, 2nd edition. Australia: The Royal Australian College of General Practitioners; 2018.
17. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American College of Rheumatology/Arthritis Foundation guideline for the management of osteoarthritis of the hand, hip, and knee. Arthritis Rheumatol 2020;72(2):220-233 [FREE Full text] [doi: 10.1002/art.41142] [Medline: 31908163]
18. Fernandes L, Hagen KB, Bijlsma JW, Andreassen O, Christensen P, Conaghan PG, European League Against Rheumatism (EULAR). EULAR recommendations for the non-pharmacological core management of hip and knee osteoarthritis. Ann Rheum Dis 2013;72(7):1125-1135 [FREE Full text] [doi: 10.1136/annrheumdis-2012-202745] [Medline: 23595142]
19. Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. AAAI 2023:17754-17762.
20. Feng SY, Khetan V, Sacaleanu B, Gershman A, Hovy E. CHARD: clinical health-aware reasoning across dimensions for text generation models. arXiv:2210.04191 2023:1-16.
21. geteff1 / Osteoarthritis-Benchmark. GitHub. URL: https://github.com/geteff1/Osteoarthritis-Benchmark [accessed 2024-07-09]
22. Xiao Z, Ji W, Zhiyang H, Xien L, Ying S. Medical Exam Question Answering with Large-scale Reading Comprehension. 2018 Presented at: The Thirty-Second AAAI Conference on Artificial Intelligence; February; New Orleans, Louisiana, USA.
23. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023;6(1):210 [FREE Full text] [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
24. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
25. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
26. Kraljevic Z, Shek A, Bean D, Bendayan R, Teo J, Dobson R. MedGPT: medical concept prediction from clinical narratives. arXiv:2107.03134 2021:1-6.
27. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4(1):86 [FREE Full text] [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
28. Tian Y, Gan R, Song Y, Zhang J, Zhang Y. ChiMed-GPT: a chinese medical large language model with full training regime and better alignment to human preferences. Comput Sci: Comput Language 2023. [doi: 10.48550/arXiv.2311.06025]
29. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature 2023;624(7992):570-578 [FREE Full text] [doi: 10.1038/s41586-023-06792-0] [Medline: 38123806]
30. Zhang P, Xiao S, Liu Z, Dou Z, Nie J. Retrieve anything to augment large language models. Comput Sci: Inform Retrieval 2023.
31. Lee J, Griepp DW, Burgess CJ, Petrone B, Bitterman AD, Cohn RM. The AAOS 2019 clinical practice guidelines for the management of rotator cuff injuries are unbiased and incorporate a diverse body of literature. Arthrosc Sports Med Rehabil 2022;4(2):e559-e565 [FREE Full text] [doi: 10.1016/j.asmr.2021.11.017] [Medline: 35494298]
32. Jang B, Kim I, Kim JW. Effective training data extraction method to improve influenza outbreak prediction from online news articles: deep learning model study. JMIR Med Inform 2021;9(5):e23305 [FREE Full text] [doi: 10.2196/23305] [Medline: 34032577]

XSL•FO
RenderX

Abbreviations

ACR: American College of Rheumatology/Arthritis Foundation
EULAR: European League Against Rheumatism
GIQA: guideline-item question-answer
LLM: large language model
MOQA: management options question-answer
NICE: National Institute for Health and Care Excellence
OA: osteoarthritis
OARSI: Osteoarthritis Research Society International
QA: question-answer
RAG: retrieval-augmented generation
RCQA: real-case QA
TSQA: treatment strategy QA

Edited by Q Jin; submitted 07.03.24; peer-reviewed by S Matsuda, L Wang; comments to author 08.04.24; revised version received 28.04.24; accepted 03.06.24; published 22.07.24.

Please cite as:
Chen X, Wang L, You M, Liu W, Fu Y, Xu J, Zhang S, Chen G, Li K, Li J
Evaluating and Enhancing Large Language Models' Performance in Domain-Specific Medicine: Development and Usability Study With DocOA
J Med Internet Res 2024;26:e58158
URL: https://www.jmir.org/2024/1/e58158
doi:10.2196/58158
PMID:38833165

©Xi Chen, Li Wang, MingKe You, WeiZhi Liu, Yu Fu, Jie Xu, Shaoting Zhang, Gang Chen, Kang Li, Jian Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Human-AI Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation

Nadine Bienefeld[1], PD, PhD; Emanuela Keller[2], MD, Prof Dr; Gudela Grote[1], Prof Dr

[1]Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

[2]Institute of Intensive Care Medicine, Department of Neurosurgery, University Hospital and University of Zurich, Zurich, Switzerland

Corresponding Author:
Nadine Bienefeld, PD, PhD
Department of Management, Technology, and Economics
ETH Zurich
Zurich
Switzerland
Phone: 41 44 633 45 95
Email: nbienefeld@ethz.ch

Abstract

Background: Artificial intelligence (AI) holds immense potential for enhancing clinical and administrative health care tasks. However, slow adoption and implementation challenges highlight the need to consider how humans can effectively collaborate with AI within broader socio-technical systems in health care.

Objective: In the example of intensive care units (ICUs), we compare data scientists' and clinicians' assessments of the optimal utilization of human and AI capabilities by determining suitable levels of human-AI teaming for safely and meaningfully augmenting or automating 6 core tasks. The goal is to provide actionable recommendations for policy makers and health care practitioners regarding AI design and implementation.

Methods: In this multimethod study, we combine a systematic task analysis across 6 ICUs with an international Delphi survey involving 19 health data scientists from the industry and academia and 61 ICU clinicians (25 physicians and 36 nurses) to define and assess optimal levels of human-AI teaming (level 1=no performance benefits; level 2=AI augments human performance; level 3=humans augment AI performance; level 4=AI performs without human input). Stakeholder groups also considered ethical and social implications.

Results: Both stakeholder groups chose level 2 and 3 human-AI teaming for 4 out of 6 core tasks in the ICU. For one task (monitoring), level 4 was the preferred design choice. For the task of patient interactions, both data scientists and clinicians agreed that AI should not be used regardless of technological feasibility due to the importance of the physician-patient and nurse-patient relationship and ethical concerns. Human-AI design choices rely on interpretability, predictability, and control over AI systems. If these conditions are not met and AI performs below human-level reliability, a reduction to level 1 or shifting accountability away from human end users is advised. If AI performs at or beyond human-level reliability and these conditions are not met, shifting to level 4 automation should be considered to ensure safe and efficient human-AI teaming.

Conclusions: By considering the sociotechnical system and determining appropriate levels of human-AI teaming, our study showcases the potential for improving the safety and effectiveness of AI usage in ICUs and broader health care settings. Regulatory measures should prioritize interpretability, predictability, and control if clinicians hold full accountability. Ethical and social implications must be carefully evaluated to ensure effective collaboration between humans and AI, particularly considering the most recent advancements in generative AI.

(J Med Internet Res 2024;26:e50130) doi:10.2196/50130

KEYWORDS

AI in health care; human-AI teaming; sociotechnical systems; intensive care; ICU; AI adoption; AI implementation; augmentation; automation, health care policy and regulatory foresight; explainable AI; explainable; human-AI; human-computer; human-machine;

XSL•FO
RenderX

ethical implications of AI in health care; ethical; ethic; ethics; artificial intelligence; policy; foresight; policies; recommendation; recommendations; policy maker; policy makers; Delphi; sociotechnical

Introduction

The rapid development of artificial intelligence (AI) and machine learning offers unprecedented opportunities for supporting physicians and nurses (hereafter clinicians) in a wide range of clinical and administrative tasks [1,2]. Despite these promises, however, integrating AI into clinical practice remains slow, with significant implementation hurdles [3-5].

One key obstacle is failing to account for the broader sociotechnical system (STS) in which humans and AI collaborate. Disregarding the seamless integration of AI with human work practices and the overall systems in which they operate can not only lead to a lack of acceptance [6,7]; but also introduce unwanted errors, patient safety risks, and, in the long run, increase rather than decrease costs [8].

To fully harness the complementarity between humans and machines [9], advancements in machine capabilities must be accompanied by simultaneously building human competencies for successfully collaborating with machines [8]. Contrary to popular belief, more capable systems do not necessarily reduce human effort or errors. Paradoxically, complex systems like AI can translate into higher rather than lower demand for skilled personnel and the need for enhanced expertise, ultimately leading to decreased efficiency and potential performance losses [10]. Furthermore, numerous accidents in aviation, for example, the crash of Air France 447 [11], serve as tragic demonstrations that increased system complexity and a shift from manual to supervisory control are challenging for humans in terms of maintaining situational awareness, the loss of control, or deskilling [12-15].

The lack of transparency in today's black-box AI systems raises the additional question of who should be responsible for the system's outcomes. If humans are to maintain overall authority over system goals and their attainment, both accountability and control over system functioning must be afforded to humans, requiring transparency, predictability, and means to influence the systems [16]. In most of today's AI applications, these conditions are not met, exposing clinicians to significant legal and professional repercussions and raising questions about reassigning accountability to other entities (eg, AI development firms or health insurers) [17].

This study integrates different viewpoints on these issues by data scientists developing AI systems for intensive care unit (ICU) and the clinicians ultimately using these AI systems. We thus tackle the question of how humans can effectively collaborate with AI from a holistic STS perspective [8] and in a context where AI could support overworked clinicians based on AI solutions using high volumes of patient data [18,19].

Contrary to technology-driven initiatives, our human-first approach considers the complementarity between humans and machines to create joint cognitive systems (humans and AI) that achieve better outcomes than humans or AI could achieve on their own [20]. Just because a particular AI application has the potential to augment or automate a given task does not mean that it should do so at all costs. We thus explore the question of social desirability and ethical acceptability of task automation or augmentation, both from the perspectives of clinicians currently performing these tasks and data science experts developing AI solutions.

By defining optimal human-AI teaming in clinical practice and considering the broader sociotechnical context in which AI operates, this study contributes to a more effective, safe, socially acceptable, and ethically sound use of AI in health care. The resulting decision framework can assist hospital managers, policy makers, and legislators in making well-balanced decisions about AI implementation and use to reduce the workload for clinicians and advance the overall quality of care.

Methods

Overview

This multimethod study combines (1) a systematic sociotechnical analysis of ICU work tasks with (2) an international Delphi survey among n=19 data science experts to assess optimal levels of augmentation and automation of these tasks; and (3) n=61 semistructured interviews with clinicians exploring their views on AI augmentation and automation.

Ethical Considerations

Ethics approval was obtained from the local ethics committee at ETH Zurich (number EK 2019-N-51 and EK 2019-N-190), and informed consent was obtained from all participants before data collection.

Sociotechnical Work Task Analysis

The sociotechnical work task analysis was conducted using COMPASS (Complementary Analysis of Socio-technical Systems), a well-established framework for assessing work tasks and systems based on STS theory [20-22]. Detailed observational notes were analyzed using COMPASS [22] to identify 6 core tasks performed by ICU clinicians (Textbox 1). Observations were carried out as part of a related study [23], during which the first author systematically observed all work-related activities performed by ICU clinicians for 30 morning shifts (8.5 hours each). Additional data were gathered from hospital documents, such as descriptions of job profiles, professional competencies, and organizational charts. Two ICU department heads checked and validated the accuracy and adequacy of the analysis concerning the correctness of medical knowledge and adequate use of terminology [24].

XSL•FO
RenderX

Textbox 1. Core tasks performed by intensive care unit clinicians (based on observation and COMPASS work system analysis).

- Monitoring patient data (derived from biosensors such as vital signs, parameters from artificial ventilation, or laboratory values).
- Documenting clinical information.
- Analyzing medical data (eg, from reports, articles, test results, or images).
- Prescribing medication or treatment.
- Diagnostic decision-making.
- Interacting with patients.

Delphi Survey

To assess data science experts' agreement on the levels of human-AI teaming and technological feasibility to augment or automate each of the 6 core tasks, we conducted an international Delphi survey [25,26]. The Delphi survey is an iterative process method aiming to forecast future (technological) developments and attain consensus among a group of experts regarding questions where there are no clear right or wrong answers and where there is limited or contradictory information. The Delphi method has been shown to outperform other group consensus-finding methods regarding accuracy and efficacy [27], making it a commonly used technique in health information management and health care (eg, [25]).

Selection of International Data Science Experts

The quality of Delphi surveys heavily depends on the adequate choice of experts [28]. Therefore, we purposefully selected each expert based on their internationally renowned expertise in the fields of bioinformatics, bioengineering, and health data science. Within these fields, we included academic researchers (professors) from global top-tier universities and data scientists employed by global health care technology manufacturers. Suitable participants were identified through academic publications, participation at conferences, topic-based newsletters [29], and a search of companies and job descriptions on the professional social network LinkedIn, aiming to create a heterogeneous sample regarding geographical regions. Based on the recommended number of participants for Delphi surveys between 10 and 35 [30], we invited 20 data science experts (DS 1-20), out of whom 19 agreed to participate (response rate=95%). Data scientists were 21% female (the relatively low proportion of female experts represents the gender distribution in the field of data science), between ages 29 and 47, and had between 8 and 21 years of experience as professionals in their domain. A total of 37% of experts were employed by global health care technology manufacturers and 67% were employed as faculty of top-tier universities worldwide; 42% of experts were from the AMER (North, Central, and South America) region; 42% from the EMEA (Europe, the Middle-East, and Africa) region; and 16% from the APAC (Asia Pacific, excluding China) region.

Survey Development and Data Collection

As an entry point to the Delphi survey and to provide common ground among experts, we used Russel and Norvig's definition of AI as "machines that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving" [31] and described each of the 6 core tasks performed by ICU clinicians in short vignettes to illustrate the ICU work system and workflow (Multimedia Appendix 1). Experts were then asked to anonymously select which level of automation and augmentation they deemed best suited to enable safe and effective human-AI teaming for each task from a technological feasibility perspective. The levels of human-AI teaming were developed based on Johnson and colleagues' taxonomy that goes beyond the traditional levels of automation [32], in that it "more effectively models the technology, the human, and the work [system] together" [33] (Table 1). In addition, data scientists were asked to indicate which of the core tasks performed in the ICU *should or should not* be augmented or automated by AI based on their subject matter expertise, including social and ethical considerations beyond merely considering technological feasibility. For each answer, participants were asked to provide an open-text explanation to justify their choices. The first 3 experts to respond to our initial invitation were chosen to pilot-test the Delphi survey, which resulted in only minor changes. Experts submitted their responses anonymously via email in multiple survey rounds. At each data collection point, 2 participation reminders were sent via email. There were no dropout cases.

Statistical data about the group's collective choices and the qualitative answers given by data scientists explaining their choices were anonymously fed back to all participants in each consecutive round. As recommended by Sumsion [34], agreements above 70% were considered as consensus. This goal was achieved in round 3 and the Delphi survey was terminated then.

XSL•FO
RenderX

Table 1. Results from the Delphi survey—data science experts' agreement in 3 Delphi survey rounds.

Task	Level 1: AI is unable to significantly augment human performance %[a]			Level 2: AI augments human performance (humans benefit from AI augmentation but are always needed), %			Level 3: humans augment AI performance (AI could perform alone, but human input increases reliability), %			Level 4: AI automates tasks without human input (AI performs reliably; humans serve as troubleshooters), %			Should AI be used?, %	
	1	2	3	1	2	3	1	2	3	1	2	3	Yes	No
Monitoring patient data	—[b]	—	—	—	—	—	—	—	—	63.6	63.6	72.7	92	8
Documenting clinical information	—	—	—	—	—	—	90.1	90.1	90.1	—	—	—	100	—
Analyzing medical data	—	—	—	—	—	—	81.8	90.1	100	—	—	—	92	8
Prescribing medication or treatment	—	—	—	81.8	90.1	90.1	—	—	—	—	—	—	92	8
Diagnostic Decision-Making	—	—	—	81.8	90.1	90.1	—	—	—	—	—	—	53	47
Interacting with patients	90.1	90.1	90.1	—	—	—	—	—	—	—	—	—	—	100

[a]Percentage values are used as a measure of agreement among experts.

[b]Not applicable.

Survey Data Analysis

The Qualtrics survey web-based platform [35] was used to develop the survey and collect data. SPSS (version 24; IBM) was used for statistical data analysis.

Interviews With ICU Clinicians

To examine ICU clinicians' perspectives on the social and ethical implications of AI-enabled automation and augmentation technologies, we conducted 61 semi-structured interviews as part of a larger research project [36]. Interview questions related to clinicians' vision of future human-AI teaming solutions and the distribution of control and accountability (Multimedia Appendix 2). Interviews were conducted in private offices, lasted between 60 and 90 minutes, and were audio-recorded and manually transcribed ad verbatim.

Selection of Clinicians

In line with grounded theory [37], we used a theoretical sampling approach to include both professional groups of ICU physicians and nurses (25 physicians and 36 nurses). Clinicians were 56% female and had between 2 and 30 years of experience after completing their initial education as a registered nurse (RN) or board-certified ICU physician. All informants were directly involved in care delivery. We gave each informant a code showing their professional position (attending physician [AP], resident physician [RP], and RN) and personal identifier (1–61).

Interview Data Analysis

The analysis of the interview content was conducted following the grounded theory methodology [37]. The principal investigator engaged in the process of open coding, methodically labeling each discrete conceptual unit within the interview transcripts. This procedure entailed the aggregation of related codes into broader categories and themes. Subsequently, a review of the related textual segments was performed to verify the consistent alignment of the data with the identified themes, a step that is pivotal for maintaining the integrity of qualitative inquiry [38]. The software MAXQDA (version 2024; VERBI GmbH) was used to facilitate qualitative data analysis [39]. The presentation of the qualitative findings adheres to the protocols endorsed by the Academy of Medicine [40].

Results

Data Science Expert Perspective

Table 1 summarizes the results from the Delphi survey. Data science experts' assessments about the technological potential to augment or automate each ICU task were quite similar from the start, and increasing consensus between 72.7% and 100% was reached across the 3 Delphi survey rounds. Data science experts also showed high levels of agreement (92%-100%) about which of the 6 tasks *should or should not* be augmented or automated by AI based on social and ethical considerations, except for one task: For the task "diagnostic decision-making" 53% of data science experts recommended the use of AI, whereas 47% did not.

In what follows, we provide a detailed account of how the consensus-building process among data science experts unfolded as part of the Delphi survey process. For each task and level of human-AI teaming, we provide illustrative statements by data science experts justifying their choices.

Monitoring Patient Data

In the first round, 63.6% of data science experts agreed that AI could eventually fully automate this task and that human assistance would bring no further benefits (level 4). This result remained the same also in the second round but increased to a 72.7% consensus in the third round. One data science expert justified this choice by stating:

XSL•FO
RenderX

> *Large longitudinal data like heart rate, BP, etc., are much harder to interpret for humans. It's like weather forecasting where you have so many data points and supercomputers can more accurately model weather than humans. Even the most experienced clinicians are overwhelmed with today's data overload and since human monitoring is not feasible at all times, full automation without human interference is really the best option (DS 17).*

A total of 27.3% of experts selected the design choice "AI could do it, but humans increase reliability" in the first and second rounds:

> *Monitoring of the signs is possible to learn by ML; however, having human help when an obvious mistake happens increases efficiency and reliability and seems unavoidable. AI could not do it as reliably and reproducibly as doing it together with the human (DS 05).*

Regarding the question as to whether "monitoring patient data" *should* be automated by AI (level 4), all but one expert (92%) agreed that there were no social or ethical concerns:

> *Automated monitoring is bound to be better than what there is now. Even if the system fails at times, it's not that patients will die from this (DS 11).*

Documenting Clinical Information

In the first round, 90.1% of experts agreed that AI could eventually automate this task, but human assistance would increase reliability (level 3). Expert agreements remained stable throughout all 3 rounds. The following argument was given by a data science expert, further explicating their choice:

> *There are a lot of approaches that involve humans in the loop such as active learning/weak supervision that can greatly augment the reliability of AI techniques used for these types of tasks. With greater data standardization and curated datasets, the goal of greater automation becomes more and more feasible but for now, physicians will have to be in the loop (DS 07).*

Also, 100% of experts agreed that this task *should* be augmented by AI, for instance, as one data science expert argued:

> *That's [the task of documenting clinical information] what [clinicians] don't like doing and where AI can really help so they can focus on more interesting things (DS 15).*

Analyzing Medical Data

Already in round one, most data science experts (81.8%) agreed that AI could eventually do this task autonomously but that human assistance would increase reliability (level 3), for instance, stating that:

> *So many fields are already doing automated reporting and with humans in the loop, this becomes easier and easier for AI systems to automate and for medical professionals to interpret. Again, not fully autonomous but moving closer on the continuum towards it (DS 09).*

Two data science experts adopted the consensus opinion in the second and third rounds of the Delphi process, resulting in a 100% agreement. One of them provided the following explanation for changing from level 4 to level 3:

> *I agree. The process of automatically producing "draft analysis reports" can be fully automated but if we take human sense-making as a kind of "assistance" then I would also say that this [level 3] is the right choice (DS 18).*

Furthermore, 92% of experts agreed that there were no social or ethical concerns and that this task *should* be augmented by AI (level 3).

Prescribing Medication or Treatment

In the first round, 81.8% of data science experts agreed that this task could be augmented by AI but would always require humans in the lead (level 2). This opinion increased to 90.1% in rounds 2 and 3. The following arguments were given by data science experts, explaining their choices:

> *Although automated diagnostics and algorithmic medication prescriptions have become extremely accurate because mistakes are very costly, a medical professional should always review and approve the suggestions produced by the software (DS 10).*

One expert (9.1%) argued that AI could perform this task with human assistance (level 3) and did not change his or her opinion throughout the Delphi survey process based on the following explanation:

> *I don't think human oversight is always required. It will be like how self-driving cars are gradually phasing into our society as well. There are certain use cases where automated diagnostics can yield greater results in a well-designed system that triages events earlier in the healthcare system and prevents issues later downstream that have bigger healthcare impacts in terms of patient health and costs (DS 19).*

In total, 92% of data science experts agreed that this task *should* be augmented by AI as long as humans are accountable and in the lead (level 2):

> *AI tools will never be ready to prescribe meds without a human in the loop simply because they don't take legal liability for the decisions which ultimately will have to be made by the physician (DS 06).*

Diagnostic Decision-Making

In the first round, 81.8% of data science experts indicated that AI could augment this task, but humans would always be required (level 2), and in the second and third rounds, 90.1% selected this choice. As one data science expert stated:

> *I think our [AI] solutions are far too narrow for this type of decision-making. They [AI solutions] can assist but need human input always (DS 05).*

Two experts (18.2%) chose level 3 for this task in the first round stating that:

XSL•FO
RenderX

> *Diagnostics and prognosis can be automated really well with AI/ML but depends on how it is accepted. But in my opinion, having 100% human supervision defeats the purpose of AI/ML (DS 16).*

Concerning social and ethical considerations 47% of data science experts believed that diagnostic decision-making *should not* be augmented by AI, even if it was technologically feasible to do so:

> *I strongly believe that for as long as we [data scientists] are still struggling with issues of bias, transparency, and equity etc., we should shy away from such consequential tasks [diagnostic decision-making]. ML can be utilized in the form of a "recommendation engine", but doctors should never rely on it fully (DS 19).*

Besides the ethical issues of bias, transparency, and equity, some data science experts also felt that using AI in diagnostic decision-making would undermine physicians' role identities, which could be problematic.

> *This is what being a doctor is all about. Applying one's knowledge and experience to diagnose patients is the very core of medicine. It would be like messing with the Hippocratic oath and taking away the reason why they chose to become a doctor in the first place (DS 08).*

Interacting With Patients

With a strong 90.1% agreement in all 3 rounds and no changes throughout, experts believed that the task of patient interaction cannot be augmented or automated by AI.

> *AI/ML technologies simply cannot properly mimic human soft skills which are essential in these sorts of interactions (DS 13).*

Only one expert argued that AI could augment humans in their patient interactions (level 2) by stating:

> *In most cases, it is challenging for AI/ML to interact empathically with patients, but there are some new use cases with promising results that could potentially augment this task (DS 07).*

However, regardless of the technological possibilities of AI mimicking empathic patient interactions, all data science experts (100%) agreed that AI *should not* be used to interact with patients, mainly based on social concerns and what it means to be human:

> *Medicine is all about empathy, one human caring for another [...]. Therefore, [interactions with patients] should never be replaced by an AI [ro]bot or the like (DS 11).*
>
> *The Role of AI is to provide more time for physicians to do what this empathy-driven field requires most: human interaction and connection. No AI should interfere with that (DS 17).*

Clinician Perspective

In the following, we report clinicians' views on how to best team up with AI when augmenting or automating ICU tasks and conclude with a brief comparison between the 2 stakeholder groups.

Monitoring Patient Data

Like data scientists, clinicians acknowledged the benefits of automating the monitoring task. Many clinicians, however, pointed out the need for highly reliable systems so they would no longer have to constantly supervise AI, which defeats the purpose of gaining time and increasing efficiency:

> *Taking over the monitoring [task] would really save us time and reduce our workload. But of course only if the AI is so good that we don't have to constantly monitor it. Because until now, with AI [systems], we still have to do the monitoring so it is more like a double effort and I actually have less time for the patient (RN 43).*

Documenting Clinical Information

Clinicians perceived the documentation task as "the most time-consuming [task] of all," imagining a future where AI could "take away that burden" (RP 12) and welcoming high levels of automation for this task. As one AP explained, their future roles would ideally consist of merely checking for potential errors in AI-produced documents:

> *This [automation of clinical documentation] would be a huge relief, especially for residents because they are the ones who spend the most time on administrative tasks. I see them sit here [in the ICU] for hours on end way past they clock out [after the end of their shift] and I think to myself, "All those years of medical school to do what - office work!" (AP 39).*

Analyzing Medical Data

Clinicians saw immense potential in using AI to aid with the analysis of medical data, but only if they could "stay in the driver's seat." Most clinicians said they would always want "to decide whether to follow the advice [given by AI] or not" (AP 11). One nurse gave the example of using AI to analyze medical data so they could foresee the risk of delirium in ICU patients and initiate prophylactic actions before problems occurred:

> *All the vast amounts of data that we can use to predict the onset of conditions such as delirium. That would be brilliant because delirious patients have a considerably higher mortality rate and are very resource-intensive. In some ICUs, [the use of AI to analyze medical data] is already available, for instance, to predict sepsis and so on. A lot is going on at the moment and I see a huge opportunity [for AI] to recognize all these risks in patients so that we can simply say, "Oh yes, that makes sense" and take prophylactic action (RN 36).*

Prescribing Medication or Treatment

Similar to data scientists, clinicians realized AI's potential to augment the task of prescribing medication and treatment, but they stressed the importance of keeping the ultimate decision

XSL•FO
RenderX

power in their own hands. As one RP explained this important precondition of using AI for this task:

> *Yes, of course, an AI system that can provide us with information on adverse drug interactions or patient-specific intolerance, for example, would be great because we humans are much more prone to error for these kinds of tasks than machines. This also goes in the direction of personalized medicine and that is the future for sure. But one thing that must not happen is that the system then directly initiates a therapy, that would be too dangerous (RP 58).*

One use case that was discussed intensively by clinicians was the use of AI to assist decision-making around which treatment patients should receive based on ethical considerations and long-term prognosis:

> *AI could help us a lot in deciding whether and if so which therapy really makes sense for a patient in terms of prognosis and quality of life long-term. Now, we have interdisciplinary ethics boards for this [decision], but the outcome for each patient is mostly unclear. If AI could show all the facts and the prognosis and perhaps also visualize them, our decisions would ultimately be more ethical, objective, comprehensible, and better in terms of health economics for sure (RN 30).*

Diagnostic Decision-Making

Unlike data scientists believing that clinicians would be worried about their loss of status or power, not a single clinician worried about themselves. Rather than focusing on the changes AI would incur in their professional lives, clinicians focused solely on the benefits AI would bring to their patients. As one AP explained, putting patient benefits front and center is what matters most in any decision they make:

> *What's important here is not whether we like [the changes] or not but how the patient can benefit from [AI solutions]. The goal should be that AI can help as many patients as possible. I mean it's well known that it is more difficult for us [humans] to diagnose rare diseases and that we have all kinds of cognitive biases. That's where I see the big potential of AI, that it can help [improve diagnostic accuracy] (AP 01).*

Unlike data science experts, issues relating to biased or unfair AI algorithms did not feature in clinicians' answers. Instead, clinicians highlighted how important it was to "see every patient as a unique human being" and to assess each patient "holistically including individual, social, and environmental factors" (RN 45):

> *Diagnostic support from AI is great. But where I would have problems is if every patient was treated in the same standardized way and we could no longer have a say. Because every patient is unique; one patient feels fine with a blood pressure of 60 and another patient is lying flat on the floor already. Medicine cannot be generalized, it cannot be standardized following some statistical norms, that would be a setback (RN 43).*

Interacting With Patients

Clinicians agreed with data scientists on the importance of social interaction, empathy, and human connection in patient interactions. They too were unable to imagine a future where AI would interact socially and demonstrate human-level empathy with patients:

> *It has been proven that people need human connection and closeness to get well. Our social skills, our empathy, the care received from human to human are enormously important for overcoming even the most difficult crisis situations. I can't imagine that an AI system, a robot, will ever be able to do this. Especially not here in the ICU where patients are always on the border between life and death (RN 36).*

Furthermore, interactions with patients were seen as one of the core tasks why clinicians had chosen their profession and one that they hoped they could invest more time in, thanks to AI augmentation and automation:

> *The human aspect in nursing is actually the reason why I chose this profession. I think it would be incredibly hard for me if it was just a case of operating the AI, maybe turning the patient to his side once a shift and that would be it. So that's what I'm hoping for with AI, that we'll find time again to care for our patients more closely, to talk to them or their relatives in peace, to do all the actual caring again. Unfortunately, there's really not enough time for that anymore these days (RN 17).*

To summarize, clinicians' and data scientists' perspectives regarding the optimal levels of human-AI teaming, and ethical and social concerns were very similar across all tasks. However, clinicians provided a more nuanced picture regarding the effects different levels of augmentation and automation would have on their ability to validate AI's performance. Specifically, they argued for higher rather than lower levels of automation for the tasks of monitoring and documentation, provided that AI performance was highly reliable. Clinicians agreed with data scientists that they would need to retain the ultimate control and decision-making power for the tasks of analyzing medical data, prescribing medication or treatment, and clinical decision-making, but not because they feared losing professional status or purpose but because they were worried about patient safety. Clinicians worried less about bias or inequity in AI but more about standardization and failing to see each patient as a unique human being, that is, that the recommendation is plausible given a specific patient situation in the clinical context. Finally, for the task of interacting with patients, both stakeholder groups were clear about not wanting AI to interfere because they both valued "real" human empathy and human-to-human connection.

Discussion

Principal Findings

In this study, we identified 6 core tasks characterizing the work in hospital ICUs and assessed optimal levels of human-AI teaming for each task, considering data science experts' and

XSL•FO
RenderX

ICU clinicians' perspectives. Such a human-centered assessment based on STS theory contributes to the ongoing debates about AI use, as well as the social and ethical implications of AI in health care in the following important ways.

First, contrary to the predominant focus on what AI can and will be able to achieve, our approach considers the strengths and weaknesses of both humans and AI in a complementary fashion. Such an approach reduces the risks of misalignment and aims at achieving better overall system outcomes than humans or AI could achieve alone [8]. The ICU context served as an ideal context due to the availability of large amounts of patient data and the need for technological support. The methodology itself can be used in any work system in health care and beyond.

Second, whereas the current consensus on the future of AI in health care suggests that augmentation is always better than automation, the results from our Delphi study and the responses from clinicians themselves paint a more nuanced picture. Augmentation (levels 2 and 3) was considered the ideal form of human-AI teaming for 4 out of 6 tasks, but only if certain requirements were met. From data scientists' perspectives, these requirements include transparency, predictability, and sufficient means to influence the system [16,21]. Currently, the goals of higher transparency and predictability are not always attained due to black-boxed AI systems. Clinicians said they would not need to understand the AI algorithms as long as they provided interpretable results (eg, people can drive a car without knowing how the engine works). Involving clinicians in the co-design of interpretable rather than fully transparent systems could thus be a solution to solving the explainable AI conundrum [41]. Given the high safety risks for patients, these considerations are particularly important for "diagnostic decision-making," "prescribing medication or treatment," and "analyzing medical data." Also, both stakeholder groups agreed that at levels 2 and 3, both the control over and responsibility for system outcomes must reside with clinicians.

If full human control is not possible (due to a lack of interpretability or because the necessary skills or knowledge are absent or lost over time), legal responsibility would need to shift to another entity (eg, AI providers). In cases where AI applications advance to a level of robustness and near-perfect reliability yet human control is not guaranteed, higher levels of automation can increase overall system safety and efficiency [14]. For "monitoring patient data," data science experts and clinicians chose level 4 automation because it is not humanly possible to continuously monitor automated systems.

Looking into the future and considering the fast-moving developments in generative AI (GenAI) applied to health care [42], tasks such as "documenting clinical information" or "analyzing medical data" could benefit from higher levels of automation but also introduce new ethical and security risks [43]. Because these models are capable of combining multiple modalities (text, image, code, or video), AI applications may soon no longer be restricted to performing single tasks. Contemplating the state-of-the-art performance of such future AI applications on a wide variety of problems [44,45], they have the potential to automate tasks such as "documenting clinical information" by integrating and conjointly analyzing laboratory, visual, and patient history data. Also, while such systems currently do make mistakes, they are also able to self-correct when prompted to do so [2].

Third, for the task of interacting with patients, both data science experts and the clinicians in this study believed AI should not assist humans even if it were technologically feasible. They unanimously agreed that compassion and empathy lie at the core of the physician-patient or nurse-patient relationship and that these innately human qualities should remain in the sole hands of humans. Intriguingly, ChatGPT (version 4) from OpenAI [46] was recently rated as more "empathetic" than human physicians by a panel of expert physicians assessing physician-patient interactions posted on social media [47]. GenAI could be used to assist clinicians in brainstorming ideas on how to best prepare for difficult patient conversations or to simulate which questions patients are likely to have. This is not to say that clinicians lack these skills altogether, but that they often lack the time to prepare for challenging patient interactions [2]. After all, if AI can help free up time for tasks spent away from patients, clinicians' number one desire—(re)gaining time spent with their patients—could finally be realized [48].

Finally, the question of whether "diagnostic decision-making" should be augmented and automated by AI or not revealed the largest disagreement both within the groups of data science experts (53% yes vs 47% no) and across data science experts and clinicians. Contrary to the 47% of data science experts who believed AI should *not* be used for this task, all clinicians in our study thought that it should, as long as they could retain the final decision-making power. Concerns about the bias or fairness of AI algorithms did not feature in clinicians' responses. Instead, clinicians stressed the importance of making holistic patient assessments and that every patient is a unique human being with specific circumstances that must be considered. Moreover, some data science experts were concerned that AI augmentation or automation of the diagnostic decision-making task would threaten clinicians' role identity and professional sense of meaning and that AI should not be used for these reasons. Clinicians, however, put the benefits AI could bring to their patients above any such concerns. Finally, from a sociotechnical perspective, one important question remains: if AI applications would one day be capable of augmenting diagnostic accuracy beyond human levels, how will clinicians retain the ability to build the necessary medical expertise to control the results? Also, with increasingly accurate and capable systems, the risk of overreliance is likely to increase, which could lead to long-term consequences of deskilling and loss of expertise [49].

Implications for Policy and Practice

Based on these results, Figure 1 shows a decision diagram aimed at assisting hospital managers, IT managers, and policy makers in assessing optimal levels of human-AI teaming that enable safe and effective AI use. The decision process is based on the assumption that reliability, interpretability, and accountability are key factors in defining optimal levels of human-AI teaming by augmenting and automating a task. Furthermore, regardless of technological capabilities, the question of whether there are any ethical or social concerns to be considered should always

XSL•FO
RenderX

be asked. As depicted in Figure 1, if AI reliability is near perfect and there are no ethical and social concerns, level 4 automation is advised. If AI does not (yet) perform highly reliably but system interpretability is granted—or if accountability is shifted away from human users—level 2 or 3 is recommended. Finally, if AI systems are neither reliable nor interpretable and accountability remains with human users, level 1 is advised, or AI should not be used. Textbox 2 summarizes this study's key takeaways and implications for policy making and practice.

Figure 1. Decision flowchart summarizing recommended levels for human-AI teaming based on study results and STS literature. AI: artificial intelligence; STS: sociotechnical system.

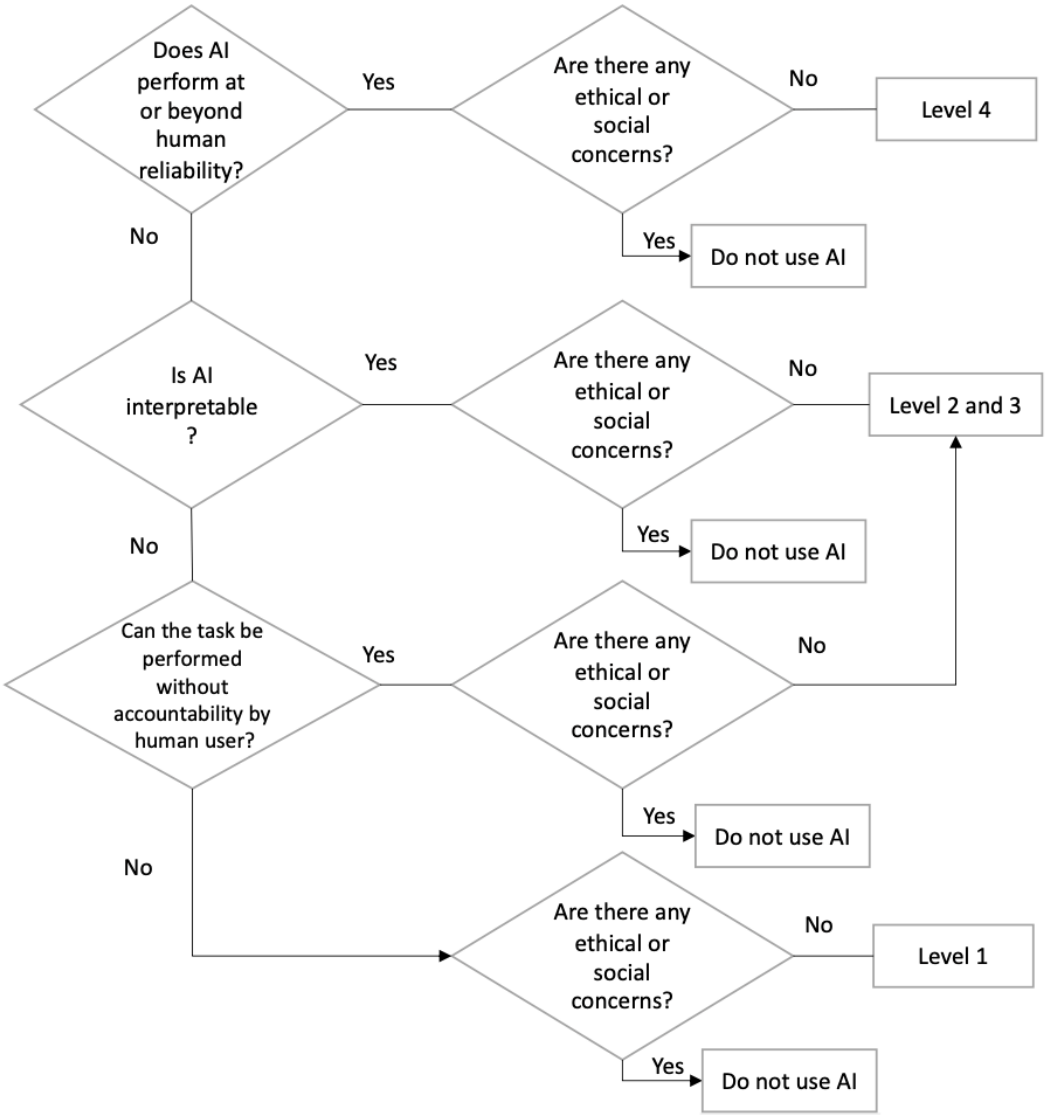

XSL•FO
RenderX

Textbox 2. Implications for policy and practice.

1. Adopting an STS perspective can prevent costly misalignment between AI, humans, and their work and facilitate implementation into clinical practice.
2. Optimal human-AI teaming is based on considering the strengths and weaknesses of humans and AI in a complementary fashion, thus creating better outcomes than either one could achieve alone.
3. Clinicians should be involved in co-designing future AI applications (eg, [41]) to leverage the complementarity of humans and AI and to continuously monitor human-AI teaming effectiveness post-implementation.
4. Clinicians need specific knowledge, skills, and attitudes to effectively work in human-AI teams; human-centered interaction design is not enough.
5. At the level of augmentation (levels 2 and 3), clinicians must be enabled to understand which data are used to make predictions, interpret results within the clinical context, and have adequate means of controlling the system.
6. If AI achieves near perfect (human-level or above) reliability yet the conditions 1-5 are not met, level 4 automation can create higher safety and efficiency.
7. All relevant stakeholders (eg, clinicians, patients, regulators, management, and patient safety experts) must decide whether AI should or should not be used based on social and ethical considerations beyond technological feasibility.
8. Regulatory foresight must address the question of control and accountability and reconsider liability assignments and insurance arrangements.
9. The rapid evolution of AI solutions needs frequent reevaluations of the STS.
10. Policy around AI-enabled in-context learning is needed, especially regarding new generations of generative AI and LLMs [1].

Limitations and Future Research

Several limitations must be considered when interpreting the results of this study. First, as we used the example of ICUs to forecast the future use of AI to augment or automate various tasks, the peculiarities of other medical domains in terms of tasks, level of risk, and ethical concerns should be specified in future research.

Second, although the Delphi technique has proven to be an ideal method for developing probable future scenarios based on experts' present knowledge and outperforms comparable interactive group forecasting techniques [27], the actual realization of the forecast depends on multiple other factors, such as the availability of resources, legal frameworks, or the acceptance by clinicians. Although the clinicians in our study agreed with data science experts about the optimal levels of human-AI teaming for most tasks, there were some differences, such as the importance of considering a patient situation holistically (eg, social context or environmental factors), beyond measurable data. Also, in the rapidly evolving landscape of AI, expert assessments may change and will need to be adapted over time. Future research, and ideally also policy making, should adopt our proactive approach and regularly ask a panel of data science experts, clinicians, and additional stakeholders, such as health insurers or hospital management, about their views. In that way, all stakeholder groups can be one step ahead of the upcoming technological change. This presents an alternative route to the currently predominant technology-driven approaches to AI implementation in health care.

Third, cultural aspects are known to influence people's perceptions regarding the social and ethical considerations of AI use [50]. For this reason, we paid special attention to including a broad sample of data science experts from the AMER, EMEA, and APAC regions. In this study, we found no notable differences between these geographical areas, but we acknowledge that we failed to recruit data science experts from China and Africa. Future research should involve China and African countries, especially. Also, the clinicians in our study, although representing multiple nationalities, are not a fully representative sample of the global ICU clinician population. Even though most ICUs adhere to similar standards and work tasks, there may be other factors, such as the availability of resources, to consider.

Lastly, despite the widespread use of the Delphi technique, particularly in medical informatics research [25,26], several concerns have been raised. Anonymity, a central characteristic of the Delphi technique, is intended to promote unbiased assessments from experts by eliminating social desirability bias. Nevertheless, it has been argued that anonymity may lead to hasty judgments, as experts feel relieved from the responsibility of defending their responses [51]. Additionally, the Delphi technique necessitates the disclosure of interim results in each round to facilitate the generation of a group consensus, which some scholars contend compromises independent judgment because this disclosure may exert social pressure on outliers to revise their assessments, thereby promoting group conformity instead of genuine changes in opinion [52]. Hence, as noted by McKenna [53], group consensus should not be regarded as the only “correct answer,” but seen as a way to structure the discussion among experts in domains where no right or wrong answers exist. In this study, we believe we achieved this goal by additionally asking experts to provide open-text answers explaining their motivation to change previous ratings.

Conclusions

With the overall aim of enabling safe, socially accepted, and ethically sound use of AI to reduce workload and improve quality of care, this study offers valuable insights into the potential of human-AI teaming in ICUs and wider health care settings. By adopting an STS perspective, we emphasize the importance of considering human and AI strengths and weaknesses in a complementary fashion to optimize outcomes and minimize misalignment. The findings challenge the prevailing notion that augmentation is always preferable to

XSL•FO
RenderX

automation by demonstrating a more nuanced picture of the ideal interaction levels for each task.

We propose several key principles for implementing AI in health care for policy and practice, for example, incorporating user-centered co-design, promoting transparency and predictability, and ensuring control over AI systems in cases where accountability resides with users. Moreover, assessing ethical and social implications when deciding on AI applications is important, especially considering the promising new developments in GenAI. Regulatory foresight and knowledge about optimal fit between work tasks and levels of human-AI teaming will be crucial in shaping the future of AI in health care and beyond.

Acknowledgments

Our immense gratitude goes to all the clinicians and data science experts who devoted their time to participate in this study and to Sophie Kerstan for her invaluable support with clinician interviews. This work was supported by funding from the Swiss National Science Foundation (grant 407740_187331). This information or content and conclusions are those of the authors and should neither be construed as the official position or policy of nor should any endorsements be inferred by the Swiss National Science Foundation.

Authors' Contributions

NB and GG contributed to the initial study design and preliminary analysis plan. EK provided her medical expertise to set up the study, verify findings from the work system analysis, and formulate questions. NB conducted all parts of the study and implemented the study protocol. NB and GG contributed to the review and analysis of the results. NB and GG contributed to the preparation of the manuscript. All authors approved the final version of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Delphi Survey Vignettes to describe ICU tasks and workflow. ICU: intensive care unit.
[PDF File (Adobe PDF File), 124 KB - jmir_v26i1e50130_app1.pdf]

Multimedia Appendix 2
Interview guidelines.
[PDF File (Adobe PDF File), 158 KB - jmir_v26i1e50130_app2.pdf]

References

1. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
3. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
4. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
5. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019;21(7):e13659 [FREE Full text] [doi: 10.2196/13659] [Medline: 31293245]
6. Abdullah R, Fakieh B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. J Med Internet Res 2020;22(5):e17620 [FREE Full text] [doi: 10.2196/17620] [Medline: 32406857]
7. Olaye IM, Seixas AA. The gap between AI and bedside: participatory workshop on the barriers to the integration, translation, and adoption of digital health care and AI startup technology into clinical practice. J Med Internet Res 2023;25:e32962 [FREE Full text] [doi: 10.2196/32962] [Medline: 37129947]
8. Johnson M, Vera AH. No AI is an island: the case for teaming intelligence. AI Mag 2019;40(1):16-28. [doi: 10.1609/aimag.v40i1.2842]
9. Hollnagel E, Woods DD. Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. Boca Raton: CRC Press; 2005.
10. Blackhurst J, Gresham J, Stone M. The autonomy paradox. Armed Forces J 2011:40.

XSL•FO
RenderX

11. Final report on the accident on 1st June 2009 to the airbus A330-203 registered F-GZCP operated by air France flight AF 447 Rio de Janeiro - Paris. BEA. 2012. URL: https://bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf [accessed 2024-06-21]
12. Billings CE. Aviation Automation?: The Search for a Human-Centered Approach. Boca Raton: Lawrence Erlbaum Associates Publishers; 1997.
13. Cook RI, Woods DD. Operating at the sharp end: the complexity of human error. In: Bogner MS, editor. Hum Err Med. USA: CRC Press; 2018:255-310.
14. Endsley MR. From here to autonomy: lessons learned from human-automation research. Hum Factors 2017;59(1):5-27. [doi: 10.1177/0018720816681350] [Medline: 28146676]
15. Lee JD. Review of a pivotal human factors article: “humans and automation: use, misuse, disuse, abuse”. Hum Factors 2008;50(3):404-410. [doi: 10.1518/001872008x288547]
16. Boy G, Grote G. The authority issue in organizational automation. In: Boy GA, editor. The Handbook of Human-Machine Interaction. USA: CRC Press; 2017:131-150.
17. Helberger N, Diakopoulos N. ChatGPT and the AI act. Internet Policy Rev 2023;12(1) [FREE Full text] [doi: 10.14763/2023.1.1682]
18. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021;47(7):750-760 [FREE Full text] [doi: 10.1007/s00134-021-06446-7] [Medline: 34089064]
19. van de Sande D, Van Genderen ME, Smit JM, Huiskens J, Visser JJ, Veen RER, et al. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ Health Care Inform 2022;29(1):e100495 [FREE Full text] [doi: 10.1136/bmjhci-2021-100495] [Medline: 35185012]
20. Grote G, Ryser C, Waefler T, Windischer A, Weik S. KOMPASS: a method for complementary function allocation in automated work systems. Int J Hum-Comput Stud 2000 Feb;52(2):267-287. [doi: 10.1006/ijhc.1999.0289]
21. Boos D, Grote G, Guenter H. A toolbox for managing organisational issues in the early stage of the development of a ubiquitous computing application. Pers Ubiquit Comput 2013;17(6):1261-1279. [doi: 10.1007/s00779-012-0634-y]
22. Waefler T, Grote G, Windischer A, Ryser C. KOMPASS: a method for complementary system design. In: Handbook of Cognitive Task Design. Human Factors and Ergonomics. USA: Lawrence Erlbaum Associates Publishers; 2003:477-502.
23. Bienefeld N, Keller E, Grote G. Can AI help with healthcare shortages? A work system analysis for enhancing job satisfaction and wellbeing. JMIR Preprints Preprint posted online Jul 14, 2023. [doi: 10.2196/preprints.50852]
24. Seale C. Quality in qualitative research. Qual Inq 1999;5(4):465-478. [doi: 10.1177/107780049900500402]
25. Ermolina A, Tiberius V. Voice-controlled intelligent personal assistants in health care: international delphi study. J Med Internet Res 2021;23(4):e25312 [FREE Full text] [doi: 10.2196/25312] [Medline: 33835032]
26. Hasson F, Keeney S, McKenna H. Research guidelines for the delphi survey technique. J Adv Nurs 2000;32(4):1008-1015. [Medline: 11095242]
27. Graefe A, Armstrong JS. Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. Int J Forecast 2011;27(1):183-195. [doi: 10.1016/j.ijforecast.2010.05.004]
28. Rowe G, Wright G, McColl A. Judgment change during Delphi-like procedures: the role of majority influence, expertise, and confidence. Technol Forecast Soc Change 2005;72(4):377-399. [doi: 10.1016/j.techfore.2004.03.004]
29. Chen E, Johri S, Rajpurkar P, Topol E. Doctor Penguin. 2020. URL: https://doctorpenguin.com/newsletters [accessed 2024-06-21]
30. Wright G, Lawrence MJ, Collopy F. The role and validity of judgment in forecasting. Int J Forecast 1996;12(1):1-8. [doi: 10.1016/0169-2070(96)00674-7]
31. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. US: Pearson Education Limited; 2010.
32. Onnasch L, Wickens CD, Li H, Manzey D. Human performance consequences of stages and levels of automation: an integrated meta-analysis. Hum Factors 2014;56(3):476-488. [doi: 10.1177/0018720813501549] [Medline: 24930170]
33. Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M. Coactive design: designing support for interdependence in joint activity. J Hum-Robot Interact 2014;3(1):43-69. [doi: 10.5898/jhri.3.1.johnson]
34. Sumsion T. The delphi technique: an adaptive research tool. Br J Occup Ther 2016;61(4):153-156. [doi: 10.1177/030802269806100403]
35. Qualtrics. 2020. URL: https://www.qualtrics.com [accessed 2024-06-21]
36. Bienefeld N. Human-AI Teaming in High-Risk Work: The Role of Cognitive-Affective Frames, Sociomaterial Practices, and Socio-Technical System Integration. Habilitation. Zurich: ETH; 2024.
37. Glaser BG, Strauss AL. Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Routledge; 2017.
38. Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research. Organ Res Methods 2012;16(1):15-31. [doi: 10.1177/1094428112452151]
39. Kuckartz U, Rädiker S. Analyzing Qualitative Data with MAXQDA: Text, Audio, and Video. Germany: Springer International Publishing; 2020.

XSL•FO
RenderX

40. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
41. Bienefeld N, Boss J, Luethy R, Brodbeck D, Azzati J. Solving the explainable AI conundrum: how to bridge the gap between clinicians needs and developers goal. npj Digit. Med 2023 Jan 07;6(94):1-7 [FREE Full text]
42. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic Narrative review. JMIR Med Inform 2024;12(1):e52073 [FREE Full text] [doi: 10.2196/52073] [Medline: 38506918]
43. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. J Am Med Assoc 2023;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
44. Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A. A Generalist Agent. 2019. URL: https://openreview.net/forum?id=1ikK0kHjvj [accessed 2022-11-11]
45. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
46. Large language model. OpenAI ChatGPT. 2024. URL: https://chat.openai.com/chat [accessed 2024-06-21]
47. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
48. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. J Am Med Assoc 2018;319(1):19-20. [doi: 10.1001/jama.2017.19198] [Medline: 29261830]
49. Beane M. Shadow learning: building robotic surgical skill when approved means fail. Adm Sci Q 2018;64(1):87-123. [doi: 10.1177/0001839217751692]
50. Wong PH. Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. Philos Technol 2020;33(4):705-715. [doi: 10.1007/s13347-020-00413-8]
51. Sackman H. Summary evaluation of delphi. Policy Anal 1975;1(4):693-718. [doi: 10.1093/jaoac/58.4.692]
52. Goodman CM. The delphi technique: a critique. J Adv Nurs 1987;12(6):729-734. [doi: 10.1111/j.1365-2648.1987.tb01376.x] [Medline: 3320139]
53. McKenna HP. The delphi technique: a worthwhile research approach for nursing? J Adv Nurs 1994;19(6):1221-1225. [doi: 10.1111/j.1365-2648.1994.tb01207.x] [Medline: 7930104]

Abbreviations

AI: artificial intelligence
AMER: North, Central, and South America
AP: attending physician
APAC: Asia Pacific, excluding China
EMEA: Europe, the Middle-East, and Africa
COMPASS: Complementary Analysis of Socio-technical Systems
GenAI: generative AI
ICU: intensive care unit
RN: registered nurse
RP: resident physician
STS: sociotechnical system

Edited by A Mavragani; submitted 20.06.23; peer-reviewed by B Chaudhry, D Juzwishin, E Vashishtha, WB Lee; comments to author 12.03.24; revised version received 03.05.24; accepted 23.05.24; published 22.07.24.

Please cite as:
Bienefeld N, Keller E, Grote G
Human-AI Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation
J Med Internet Res 2024;26:e50130
URL: https://www.jmir.org/2024/1/e50130
doi:10.2196/50130
PMID:

XSL•FO
RenderX

©Nadine Bienefeld, Emanuela Keller, Gudela Grote. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Early Detection of Pulmonary Embolism in a General Patient Population Immediately Upon Hospital Admission Using Machine Learning to Identify New, Unidentified Risk Factors: Model Development Study

Ori Ben Yehuda[1*], BSc, MSc; Edward Itelman[2,3*], MD; Adva Vaisman[2], MD; Gad Segal[2], MD; Boaz Lerner[1], BA, MSc, PhD

[1]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[2]Education Authority, Chaim Sheba Medical Center, Faculty of Health Science and Medicine, Tel-Aviv University, Tel-Aviv, Israel

[3]Cardiology Division, Rabin Medical Center, Petach-Tikva, Israel

[*]these authors contributed equally

Corresponding Author:
Boaz Lerner, BA, MSc, PhD
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
POB 653
Beer-Sheva, 84105
Israel
Phone: 972 +972544399763
Email: boaz@bgu.ac.il

Abstract

Background: Under- or late identification of pulmonary embolism (PE)—a thrombosis of 1 or more pulmonary arteries that seriously threatens patients' lives—is a major challenge confronting modern medicine.

Objective: We aimed to establish accurate and informative machine learning (ML) models to identify patients at high risk for PE as they are admitted to the hospital, before their initial clinical checkup, by using only the information in their medical records.

Methods: We collected demographics, comorbidities, and medications data for 2568 patients with PE and 52,598 control patients. We focused on data available prior to emergency department admission, as these are the most universally accessible data. We trained an ML random forest algorithm to detect PE at the earliest possible time during a patient's hospitalization—at the time of his or her admission. We developed and applied 2 ML-based methods specifically to address the data imbalance between PE and non-PE patients, which causes misdiagnosis of PE.

Results: The resulting models predicted PE based on age, sex, BMI, past clinical PE events, chronic lung disease, past thrombotic events, and usage of anticoagulants, obtaining an 80% geometric mean value for the PE and non-PE classification accuracies. Although on hospital admission only 4% (1942/46,639) of the patients had a diagnosis of PE, we identified 2 clustering schemes comprising subgroups with more than 61% (705/1120 in clustering scheme 1; 427/701 and 340/549 in clustering scheme 2) positive patients for PE. One subgroup in the first clustering scheme included 36% (705/1942) of all patients with PE who were characterized by a definite past PE diagnosis, a 6-fold higher prevalence of deep vein thrombosis, and a 3-fold higher prevalence of pneumonia, compared with patients of the other subgroups in this scheme. In the second clustering scheme, 2 subgroups (1 of only men and 1 of only women) included patients who all had a past PE diagnosis and a relatively high prevalence of pneumonia, and a third subgroup included only those patients with a past diagnosis of pneumonia.

Conclusions: This study established an ML tool for early diagnosis of PE almost immediately upon hospital admission. Despite the highly imbalanced scenario undermining accurate PE prediction and using information available only from the patient's medical history, our models were both accurate and informative, enabling the identification of patients already at high risk for PE upon hospital admission, even before the initial clinical checkup was performed. The fact that we did not restrict our patients to those at high risk for PE according to previously published scales (eg, Wells or revised Genova scores) enabled us to accurately assess the application of ML on raw medical data and identify new, previously unidentified risk factors for PE, such as previous pulmonary disease, in general populations.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48595) doi:10.2196/48595

KEYWORDS

pulmonary embolism; deep vein thrombosis; venous thromboembolism; imbalanced data; clustering; risk factors; Wells score; revised Genova score; hospital admission; machine learning

Introduction

Pulmonary Embolism Diagnosis—a Challenge Faced by Modern Medicine

Pulmonary embolism (PE) occurs when a blood clot enters and blocks (either fully or partially) the pulmonary veins, usually because of a dislodged thrombosis in the deeper veins of the lower limbs (also termed "deep vein thrombosis" [DVT]), which is an aspect of the venous thromboembolism (VTE) phenomenon [1]. After myocardial infarction and cerebral stroke, PE is the third most common cause of death from cardiovascular diseases. The worldwide financial burden of PE is immense. Reports have found the estimated cost of a thromboembolic event in the United States to be US $3000 to US $10,000, the total annual cost related to VTE to be US $33,000, and that of VTE complications to be more than US $40,000, with the US health care system's total annual expenses due to VTE reaching US $12 billion [2,3]. This cost is not significantly different in Europe. By considering the general population admitted to the hospital, we can help reduce mortality rate of later-diagnosed patients, employ early prevention means, and better use hospital resources [4], thereby alleviating this financial burden and improve health care.

Early Identification of Patients at Risk of PE

Along the VTE continuum, there is a clear correlation between the rapidity of diagnosis and the risk of mortality [5]. Various risk factors and symptoms of PE have been suggested over the years. In addition to previous cases, which increase the risk of recurrence, other patient characteristics associated with this condition include being older than 70 years, active malignancy, congestive heart failure, chronic obstructive pulmonary disease, systolic arterial hypertension, presenting with tachypnea, and right ventricular hypokinesis on echocardiography [1]. New or worsening shortness of breath, chest pain, or sustained hypotension without an obvious alternative cause [5], as well as cigarette smoking, diabetes, obesity, and any type of immobility, is also a contributing factor [6]. For suspected PE, hospital wards commonly use scoring methods designed to predict plausibility, such as the Wells score and the revised Geneva score, to decide whether to send a patient for confirmatory tests [7]. However, it is difficult to compare the Wells score for PE with scoring methods that rely entirely on medical facts, and the revised Geneva score, which is based entirely on clinical variables rather than the experience of the physician, also requires a reasonable pretest probability or initial suspicion of PE. Moreover, while the former score is considered subjective, the latter is ineffective in safely reducing the number of unnecessary computed tomographic scans [8].

In current practice, after a clinical probability assessment, the diagnostic workup should be tailored to the severity of the clinical presentation based on whether the patient's condition is hemodynamically stable or unstable. In patients with hemodynamic stability, the diagnosis of PE should follow a sequential diagnostic workup potentially consisting of D-dimer blood testing and (if necessary) multidetector computed tomographic angiography (CTA; considered to be the gold standard diagnostic procedure for PE) or ventilation-perfusion scanning (mainly in cases where CTA is contraindicated). Patients with a high pretest probability for PE should not undergo D-dimer blood testing but should immediately be sent for a confirmatory test [5].

However, many cases of PE are diagnosed late or misdiagnosed, and there are 2 main reasons for this. The first is that risk factors, which are presumably known, have yet to be proven connected to the disease, and the second is that the diagnostic methods currently used in hospitals are limited; they are not accurate, safe, or sufficiently available for routine use. Over the years, studies on PE have focused mainly on improving its diagnosis by analyzing CTA images, and when considering early diagnosis, the focus was on risk factors and symptoms that may indicate an elevated or high risk for PE.

Prediction Tools for Patients at Risk of PE

In the recent years, research on prediction tools for PE has studied several aspects of the disease, for example, by suggesting clinical prediction rules (eg, the Wells and revised Genova scores) [9] and comparing them [10], predicting PE in clinically suspected patients [11,12], predicting adverse outcomes in PE [13], and the application of statistical means for prediction [11-13].

In addition, with the advent of artificial intelligence (AI) as a predictive tool, several studies applied AI to the diagnosis of PE. For example, 1 study [14] used AI and non-AI tools to predict PE-imaging outcomes based on the patient's electronic medical record to provide a patient-specific risk score for those referred for computed tomography. Other studies used feature engineering and an artificial neural network to predict PE [15] or clustering to differentiate typical clusters of patients [16]. Note, however, that all of these studies, like many others, were performed using the data of patients during their hospital stay.

In a recent study, Ryan et al [17] devised an AI algorithm that predicts the chances for PE during hospitalization. The algorithm used data from laboratory tests conducted during the patient's hospital stay and achieved performance values of 81% sensitivity, 35%-70% specificity, and an area under the curve of 0.67-0.85. However, their approach does not meet a challenge faced by clinicians, which is the need to make a diagnosis immediately upon a patient's emergency department presentation when clinical data are only partially available and laboratory test results are still unavailable. Moreover, if we use the performance values above to calculate the positive predictive value (PPV), that is, precision, measuring the chances of a true-positive result of all results predicted as positive (which

XSL•FO
RenderX

was not reported), we obtain a value of 1.8%, which reflects too many false positives (ie, on average, only 1.8 of 100 subjects predicted as patients with PE are real patients). A system based on such an algorithm will fail to become a practical diagnostic tool, since it did not address the very high inherent imbalance in the problem in favor of non-PE (ie, the very low prevalence of 309 patients with PE among the 60,297 patients tested [17]).

Another recent study by Shen et al [18] concentrated on data collected during the first 3 hours of hospitalization, mainly vital signs and laboratory tests results, from high-risk patients (patients who recently had heart failure, have a history of specific diseases, or currently have cancer, risk factors that are also included in the Wells and Geneva screening scores). Although we acknowledge the potential value of this study, we question its applicability to the general population of hospital admissions. We note that making PE predictions for patients who have already met screening score requirements (and failing to detect those who have not), are at high risk for PE, and collecting data after patients' hospitalization are not relevant when the aim is to screen the general population immediately upon admission. Second, although predictions are based on patients who have already met screening score requirements and are at high risk for PE and based on data after patients' hospitalization, the PPV performance reported is only of 20%, again due to a failure to consider the imbalanced inhospital data (3% PE prevalence), which also makes the models proposed by the authors less practical for PE screening on hospital admission.

In contrast, this study focuses on the identification of patients at high risk for PE in the general patient population presenting at emergency departments, using only their demographic and medical history data, prior to the generation of any inhospital data, and using 2 methodologies to tackle the imbalance in PE prevalence in hospital admission data.

The Aim of This Study

Our study had three main goals: (1) to identify new PE risk factors, among those available to the medical staff upon hospital admission, that may have been overlooked; (2) to accurately predict PE as early as possible upon patient admission, as assessed by all performance measures (including the PPV), in order to create a practical diagnostic tool for PE; and (3) to group patients using the newly identified and already known risk factors in order to find subgroups of patients who are at higher risk of PE upon hospital admission. That is, the main goals of this study were not only to establish an AI tool for early diagnosis of PE but also to identify new, previously unidentified risk factors for PE that clinicians should have in mind when caring for their general population of patients—and not only those already defined as having high risk of PE—and to use these risk factors to identify informative subgroups of patients at risk of PE upon hospital admission. The fact that we did not restrict our patients to those with high risk for PE according to previously published scales (eg, Wells and revised Geneva scores) enabled us to accurately assess the application of our AI algorithms on raw medical data.

There are 3 novelties in our study. First, unlike previous studies, our study concentrates on the data of a general population of patients admitted to tertiary care and not just those with a high probability or suspicion of PE. Since straightforward modeling of PE is prone to imbalance due to the low disease prevalence among the admitted patient population, where only a minority of patients admitted to the hospital will later be diagnosed with PE, our second novelty is in proposing 2 methodologies aimed at overcoming the highly imbalanced scenario that provide reasonable PPV values. The third novelty is in a new performance measure we apply to a conventional clustering algorithm to identify clusters that maximize the minority-to-majority (PE to non-PE) ratio and thereby help us focus on PE risk factors and groups of patients at risk on hospital admission from the additional angle of a clustering measure.

In the sections that follow, we describe the data used in our study and how we addressed the imbalance challenge ("Methods" section), our modeling methods ("A Methodology for Meeting Our 3 Main Goals on Hospital Admission" section), and derived results ("Results" section) before summarizing and proposing future research directions ("Discussion" section).

Methods

In this section, we present the patient cohort and database, along with our data understanding and preparation methods. Our methods for tackling imbalance, which are described in Multimedia Appendix 1 [19-34], use the features and past diagnoses that we identify in this section as being most effective in differentiating PE from control patients.

Patients

The data were obtained by Sheba Medical Center (SMC) using the MDClone ADAMS Platform, maximizing collaboration with synthetic data while maintaining patient privacy and maximizing data use [35]. Balancing patient privacy, legal, compliance, and security issues often hinders individuals, teams, and organizations from working together to share data. However, the MDClone ADAMS Platform overcomes these common obstacles by allowing users to access and share data and information across both internal and external entities with synthetic data safely and securely. It provides data not only for research purposes but also for real-time clinical applications, which ensures the high reliability of the extracted data. Several publications have validated the usage of MDClone-generated synthetic data for epidemiological studies. In 1 study, Foraker et al [36] thoroughly compared the synthetic and real data of septic patients and did not find differences in the statistical conclusions derived. Benaim et al [37] also validated the application of MDClone-generated synthetic data in 5 observational studies and concluded that predictability was highly preserved. In this study, 2600 patients diagnosed with PE and 53,250 patients with non-PE diagnoses (controls) were represented by their demographic, clinical, laboratory test results, and medical history data.

Ethical Considerations

The synthetic data were sampled from the real-world, raw data distribution such that the statistical characteristics of their distribution replicate those of the true distribution and do not contain identifiable information included in the real data. Clinical events in the synthetic samples are slightly and

XSL•FO
RenderX

randomly shifted in time to avoid the possibility of de-anonymizing the data. To protect patient privacy, the SMC database consists of synthetic data generated from the data of all patients who were hospitalized in the center's internal medicine departments between 2008 and 2020. Patients' data were retrieved, and the MDClone system was used after the Sheba Medical Center institutional review board approval of this study (7864-20-SMC).

Data Understanding and Preparation

We suggest a model capable of predicting PE in a patient upon his or her admission to the hospital, based only on the data available in the patient's electronic medical record before admission. We focus exclusively on data available prior to emergency department admission, as these are the most universally accessible data worldwide. This approach allowed us to overcome the challenge stemming from variability in hospital-specific clinical decisions and routines, which could limit the broader applicability of our findings. We collected demographics, comorbidities, and chronic medications data for 2568 patients with PE and 52,598 control patients. While patient records for the age and sex variables were complete, the BMI variable was frequently missing; thus, we removed records with missing BMI values. This left us with the data of 1942 patients with PE (PE=1) and 44,697 control patients (PE=0). Table 1 provides some descriptive statistics for age, sex, and BMI for the patients with PE and controls. Since the 2 populations were of unequal sample sizes and variances, to check significance, we applied a Welch's *t* test (unequal variance 2-tailed *t* test). According to this test, patients with PE were significantly older (mean 69.4, SD 16.8 vs mean 65.3, SD 17.5; $P<.001$) and significantly more obese (mean 27.2, SD 5.9 vs mean 26.8, SD 5.3; $P<.001$). Moreover, female patients were diagnosed with PE significantly more often than male patients (1097/1942, 56.5% vs 845/1942, 43.5%; $P<.001$). Note, however, that our data were obtained for all patients admitted to the internal department of SMC, and, thus, case subjects and control subjects were not matched, which may explain some of the differences between the 2 groups.

Since the time of past diagnoses is not always recorded in hospital records, and sometimes it is even based on patient memory, we ignored this parameter and treated each of these clinical events as a binary variable indicating whether a patient had received this diagnosis in the past. Then we merged 19 past diagnoses into 9 categories of diagnostic "families," as is frequently done. In addition to the clinical motivation, this merging also helped us unify sparse diagnoses into denser categories. We treated past PE as a stand-alone category and did not associate it with the thrombosis category, as would be customary, because it is a known risk factor for PE, and we wanted to examine it as a "stand-alone" diagnosis in our model. Also, we converted the only chronic medication variable existing in our data, the registration date of a patient receiving treatment with anticoagulants, into a binary variable indicating whether the patient had received such treatment in the past. Table 2 shows the prevalence of the 9 categories of past diagnoses and that of anticoagulant use for the 2 patient groups. Except for the gastrointestinal past diagnosis, the differences in prevalence between the 2 groups for all past diagnoses are statistically significant according to the Welch's *t* test.

Using the PrefixSpan algorithm, which is designed to discover sequential patterns in sequence databases [38], we looked for common combinations of 2 or more past diagnoses (plus the single variable indicating the use of anticoagulants) in each of the PE and control patients. In general, we found that the patients with PE had a richer background of past diagnoses than the control patients, with only 15.7% (305/1942) of them not having any past diagnosis, compared with 46.8% (20,918/44,697) of the control patients ($P<.001$). In addition, we found that the 2 patient groups showed a high incidence of cardiovascular disease in the past. As for the combinations, pulmonary and cardiovascular background diagnoses were the most common, with 13.1% (254/1942) of the patients with PE having both (compared with only 3.1% [1386/44,697] of the control patients), while the combination of cardiovascular and past PE affected 9.1% (177/1942) (447/44,697; 1% control), pulmonary and past PE affected 9% (175/1942; 447/44,697, 1% of the control group), and cardiovascular and anticoagulant usage affected 7.3% (142/1942; 1609/44,697, 3.6% of the control group). We also identified 16 combinations of 3 variables and 1 combination of 4 variables (anticoagulants, pulmonary, cardiovascular, and past PE) affecting more than 1% (20/1942) of patients with PE. It is notable that the control patients did not have a single combination of 3 or more variables in more than 1% of the population. Figure 1 presents the prevalence (%) of past diagnoses and chronic medications for PE and control patients for single diagnoses, pairs of past diagnoses, and combinations of 3 past diagnoses. The past diagnoses identified here served as the basis for patient representation when tackling imbalance in training and testing PE classifiers (Multimedia Appendix 1).

Table 1. Descriptive statistics for patients with pulmonary embolism (PE) and control patients.

Variable	Sex distribution		*P* value
	PE[a]	Control	
Sex (%)			<.001
Male	43.5	55.4	
Female	56.5	44.6	
Age (years), mean (SD)	69.4 (16.8)	65.3 (17.5)	<.001
BMI, mean (SD)	27.2 (5.9)	26.8 (5.3)	.003

[a]PE: pulmonary embolism.

XSL•FO
RenderX

Table 2. Composition and prevalence of past diagnosis categories and anticoagulant use in our database.

Category	Past diagnosis	Prevalence (%)		*P* value
		PE[a]	Control	
Surgery	• Past surgery documentation • Past surgery procedure	6.9	5.3	<.01
Pulmonary	• Pneumonia • Pleural effusion • Chronic obstructive pulmonary disease • Restrictive lung disease	30.2	7.6	<.001
Thrombosis	• Deep vein thrombosis • Coagulopathies	15.1	3.3	<.001
Cardiovascular	• Ischemic heart disease • Atrial fibrillation or atrial flutter • Congestive heart failure	33.6	36	.03
Kidney	• Chronic kidney disease	8.8	6.2	<.001
Gastrointestinal	• Inflammatory bowel disease	0.9	1.3	.07
Joints	• Rheumatoid arthritis	1.6	0.7	<.001
Neurologic or psychiatric	• Dementia • Mental disorders • Mood disorders • Psychosis	10.0	4.7	<.001
Past PE	• Past PE	39.5	1.1	<.001
Anticoagulants	• Anticoagulants	13.8	6.1	<.001

[a]PE: pulmonary embolism.

XSL•FO
RenderX

Figure 1. Prevalence for a single past diagnosis or medication in each of the study populations (top left), a pair of past diagnoses or medications in each population (top right), and a triple past diagnosis or medication in each population (bottom). PE: pulmonary embolism.

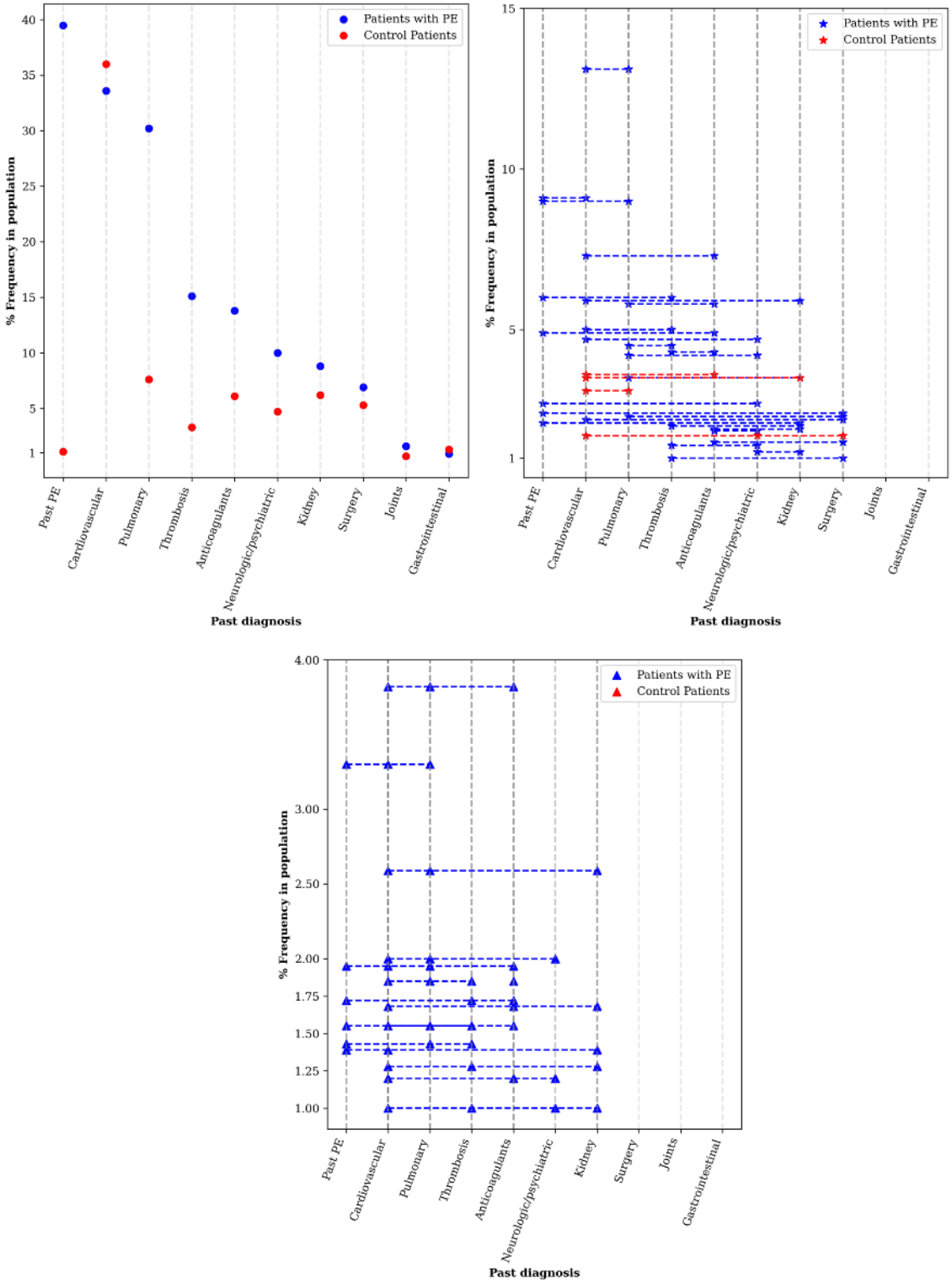

A Methodology for Meeting Our 3 Main Goals on Hospital Admission

The 3 main goals of this study were to identify new PE risk factors, among those available to the medical staff upon hospital admission that may have been overlooked, to accurately predict PE as early as possible upon patient admission, and to group subgroups of patients who have the newly identified together with known risk factors and are at higher risk of PE. To achieve the first goal, we identified important features (previous diagnoses in the patient's medical records) according to the classifier, validated this identification using an ablation study, and focused the analysis on important individual diagnoses that may be considered potential risk factors. To achieve the second goal, we used performance measures dedicated to imbalanced scenarios to estimate the ability of the 2 methodologies we

XSL•FO
RenderX

suggested (Multimedia Appendix 1) and establish a common benchmark that would enable the comparison of the 2 methods. To achieve the third goal, we clustered the population on hospital admission to identify some homogenous clusters that maximize the minority-to-majority (PE to non-PE) ratio and then focused on risk factors that established these risk groups.

Feature Importance—Initial Identification of Potential Risk Factors

We used a random forest (RF) consisting of a collection of tree-structured classifiers, each trained on bootstrapped samples of the training data, that randomly searches across subsets of the 13 input variables (the 10 in Table 2 plus the 3 in Table 1) to determine the forest's trees. The output of the classifier is a majority vote of the trees. To maximize its performance, we adjusted 3 hyperparameters of the RF: maximum depth—the longest path possible from a tree root to a tree leaf; minimum sample split—the minimal number of samples required to split an internal node; and the split criterion—either the Gini impurity or the information gain [39,40]. Our prior research demonstrates the accuracy and efficiency of the RF classifier in several clinical domains [41-45]. To increase confidence in the classifiers obtained by methods 1 and 2 (a description of the methods is provided in Multimedia Appendix 1), we repeated the evaluation of the methods over 10 data permutations and reported the averaged performance.

The RF model measures the contribution (ie, importance) of its input variables (features) to classification using the Gini index, measuring the chance of misclassifying a random record based on the a priori probabilities of each class in a particular split in a tree [39]. However, evaluation over 10 data permutations yielded 10 lists of feature importance for each method. Therefore, to determine feature importance, we proposed to evaluate the statistical difference between 10 feature lists of both methods using a Friedman-Nemenyi test [46], as described in Multimedia Appendix 2.

Accurate PE Prediction Under the Imbalanced Scenario

To perform an in-depth comparison of the performance of methods 1 and 2, we needed to evaluate them using the same data both in the training and tests sets, because method 1 uses more control subjects than method 2. Therefore, we evaluated the methods using a positive (PE) class measure, the true-positive rate (TPR), and divided the data into training and test sets in 2 steps (Figure 2):

1. We created the data sets for method 2 by creating a balanced test set: 20% of the PE subjects and the same number of control subjects and left the remaining subjects for training.
2. We created the data sets for method 1 by keeping the same PE test and training sets used in method 2. Then, we balanced the remaining control patients (beyond those used in method 2) between the control training and test sets to include 80% and 20% of the original control data set, respectively.

To focus the evaluation of methods 1 and 2 on their performance on patients with PE, we compared the TPR values of the 2 methods using a univariate 2-tailed *t* test [47].

Figure 2. Division into training and test sets to compare methods 1 and 2. X and Y are the numbers of patients equal to 20% of the PE and control sets, respectively. PE: pulmonary embolism.

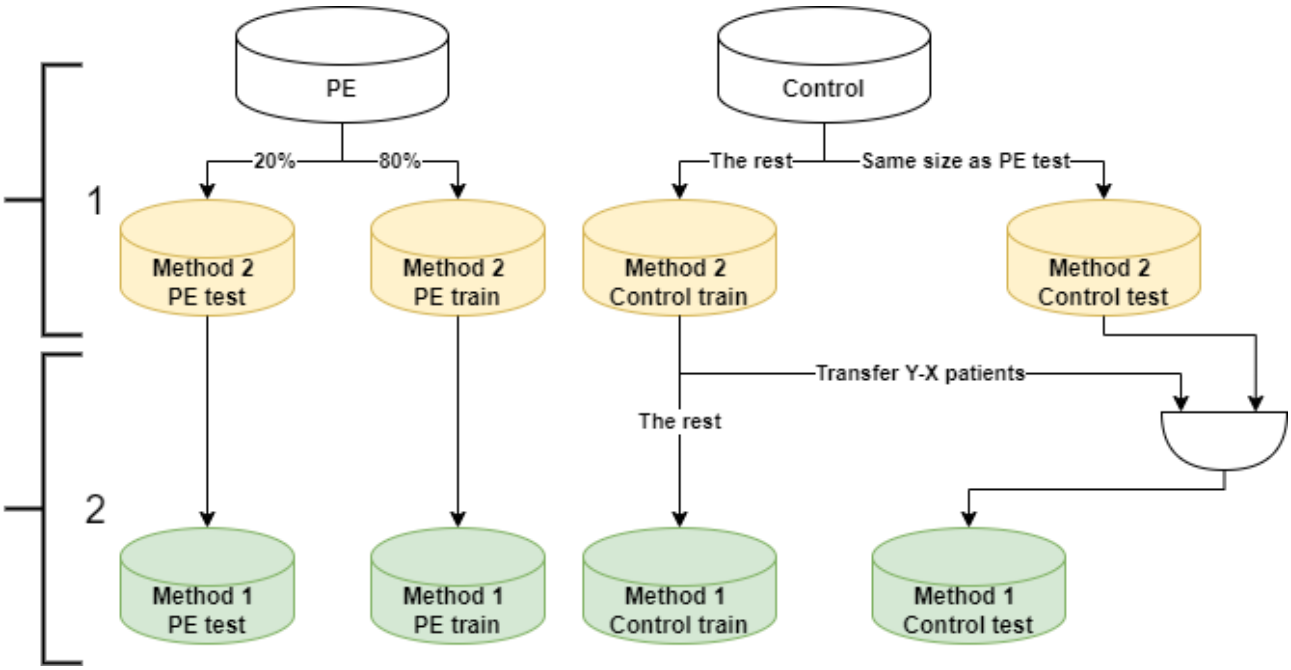

Clustering to Informative Homogeneous Subgroups for PE

To add to our study additional angle on the characteristics of patients at risk for PE on hospital admission, we wanted to find as homogeneous subgroups as possible for PE using a different performance measure than those already used. To achieve this goal, first we relaxed the categorization of the 19 indications into the 9 categories we created earlier whether there were specific past diagnoses that had not been detected in the process

XSL•FO
RenderX

performed so far that may be considered as potential risk factors for PE. Second, we established different possible feature representations for patients based on all combinations of features (past diagnoses and demographics data). Third, we suggested to maximize a performance measure that accounts for the percentages of patients with PE in a clustering scheme's clusters to simultaneously solve 2 patient representation problems that are usually solved separately, and their optimization is based on results of each other. These are the optimal feature representation and clustering scheme (number of clusters). Using this measure and the *K*-means algorithm, we clustered patients while considering different patient representations (combinations of possible feature representations and possible clustering schemes with $K=\{2, 5\}$) to derive that which yields clusters that maximize the percentage of patients with PE compared with the PE prevalence in the data (1942/46,639, 4.16%).

Results

We evaluated our methodology under the imbalanced scenario according to five criteria: (1) Gini-based feature significance, (2) classification performance measures, (3) ablation-based feature significance, (4) TPR-driven maximization, and (5) identification of potential risk factors.

Gini-Based Feature Significance

Using 10 data permutations, we applied methods 1 and 2 (Figures S1 and S2 in Multimedia Appendix 1) to predict PE and obtained 10 feature importance lists for the 13 features for each method. Testing each method separately using a Friedman-Nemenyi test, past PE was ranked the most significant by both methods, and age, BMI, pulmonary, and past thrombosis (Tables 1 and 2) were ranked not significantly differently than past PE by both methods, as well as anticoagulant usage for method 1 and sex for method 2. The other 6 features (ie, past diagnoses: surgery, cardiovascular, kidney, gastrointestinal, joints, and neurologic or psychiatric) were ranked significantly differently than past PE.

Classification Performance Measures

Tables 3 and 4 present confusion matrices for methods 1 and 2, respectively, and Table 5 presents performance measures of the best model based on the 7 most important features for each method (derived as described in the "Gini-Based Feature Significance" section) averaged over 10 data permutations. The tables demonstrate that method 1 achieved higher performance values on most measures than method 2 (and similar true-negative values) in most cases except for precision and the F_1-score. The very low precision value and very high negative predictive values of method 1 suggest that the decision threshold that was selected by this method is low, allowing more subjects to pass the threshold, among them (because of the imbalance) more controls that are thus incorrectly predicted as PE, reducing precision. However, on the other hand, this "low" threshold guarantees that almost all subjects that are identified as controls are indeed controls, increasing the negative predictive values to almost a perfect value. This "low" threshold also ensures predicting more patients with PE as PE and fewer as controls, both of which increase the TPR. These mechanisms represent a trade-off in the calculation of the main performance measure, as is expected in imbalanced scenarios.

Table 3. Method 1—confusion matrix.

	True negative	True positive	Total predicted
Predicted negative	0.83	0.01	0.84
Predicted positive	0.13	0.03	0.16
Total true	0.96	0.04	1

Table 4. Method 2—confusion matrix.

	True negative	True positive	Total predicted
Predicted negative	0.44	0.16	0.6
Predicted positive	0.06	0.34	0.4
Total true	0.5	0.5	1

Table 5. Performance measures (the values of measures for the best method are italicized).

Measure	Accuracy	Precision	TPR[a]	TNR[b]	NPV[c]	F_1-score	GM[d]
Method 1	*0.86*	0.19	*0.75*	0.86	*0.99*	0.30	*0.80*
Method 2	0.78	*0.85*	0.68	*0.88*	0.73	*0.76*	0.77

[a]TPR: true-positive rate.

[b]TNR: true-negative rate.

[c]NPV: negative predictive values.

[d]GM: geometric mean.

XSL•FO
RenderX

Ablation-Based Feature Significance

While the selection of features was Gini-based, we wanted to back this selection based on classification performance and an ablation study [48]. For each method, we compared a classifier based on the 7 features that are the union of the important features identified for both methods (past PE, age, BMI, pulmonary, thrombosis, anticoagulants, and sex), a classifier that is based on each method's 6 selected features, and 6 classifiers in which each is missing 1 of the 6 important features of a method. We trained each classifier on 10 data permutations and examined statistical difference between the geometric mean values of the classifiers using the validation set and the Friedman-Nemenyi test for each method separately. We observed that only those classifiers missing past PE, pulmonary, or thrombosis were significantly worse in their geometric mean value than the full classifier (the one that included all 7 important features). We also observed that there was no significant difference between classifiers of the 2 methods. This analysis reinforces the results presented in the "Gini-Based Feature Significance" section about the significance of past PE, pulmonary, and thrombosis to PE identification.

TPR-Driven Maximization

To examine whether one method was superior to the other on the positive (PE) group when using the same patients with PE in training and testing, we compared the TPR values of the best models produced by the 2 methods using 10 data permutations and the univariate 2-tailed *t* test. We found that there was no significant difference between the 2 methods regarding the TPR results at 95% CI.

Identification of Potential Risk Factors

To deepen our understanding of PE on hospital admission, we relaxed the categorization of the 19 indications into the 9 categories we created earlier (in the "Data Understanding and Preparation" section) to examine whether there were specific past diagnoses that had not been detected in the process performed so far, which may be considered as potential risk factors for PE. Simultaneously, we also attempted to find as homogeneous a group as possible of patients with PE. We hoped that such a group would be characterized differently than other groups, hinting at potential PE risk factors.

First, we examined the prevalence of each of the past diagnoses existing in our data for each of the patients with PE and control patients. We found that the prevalence of several of these past diagnoses, such as past PE, pneumonia, and DVT, was higher in patients with PE than in control patients (Figure 3). On the other hand, the only past diagnosis having an incidence higher than 1% in control patients compared with patients with PE was ischemic heart disease.

XSL•FO
RenderX

Figure 3. Prevalence of an original past diagnosis or medication in each of the study populations. CHF: congestive heart failure; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; DVT: deep vein thrombosis; IBD: inflammatory bowel disease; IHD: ischemic heart disease; PE: pulmonary embolism.

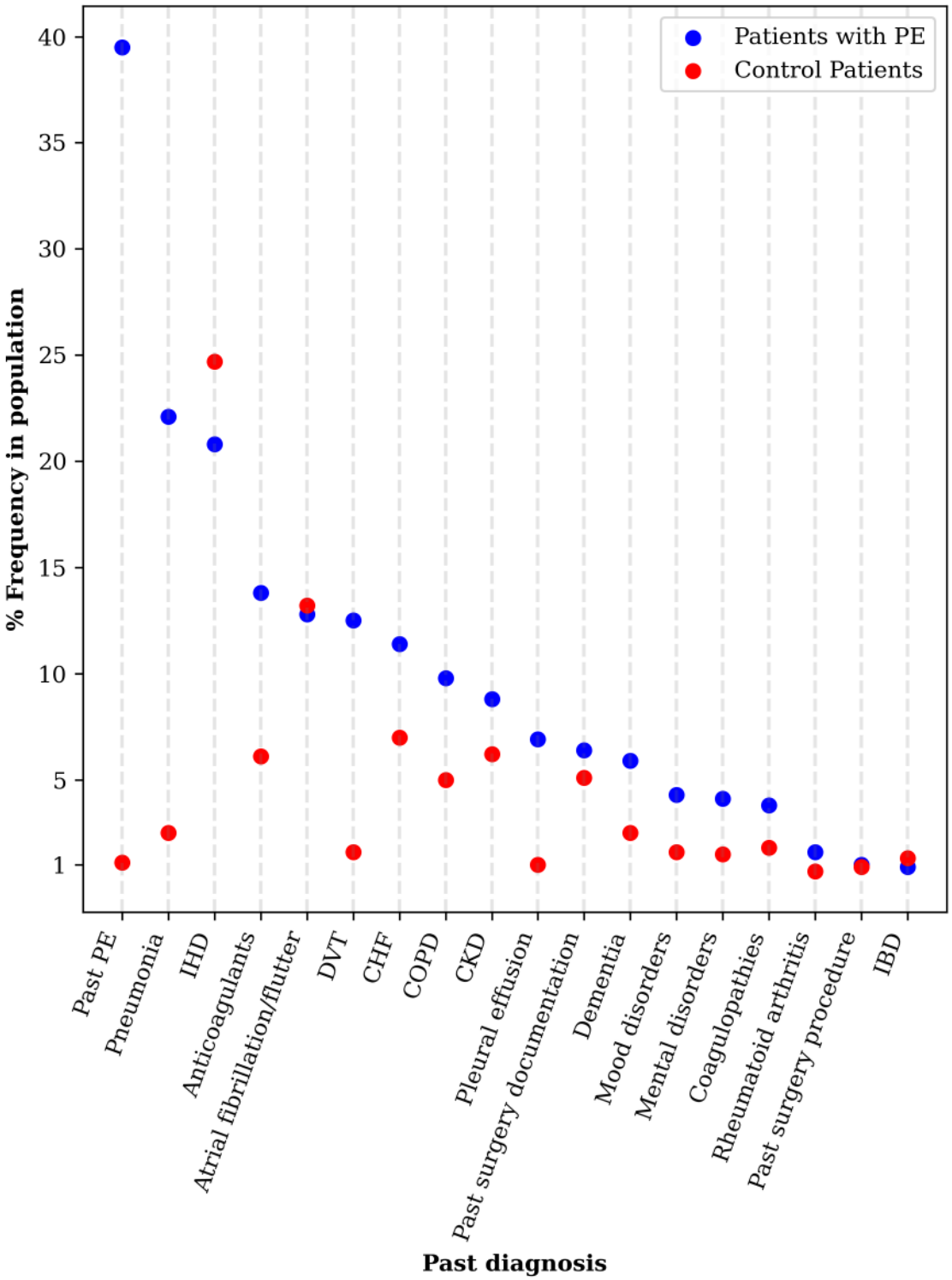

Next, to ensure our ability to find the most homogeneous group of patients with PE, we clustered, using the *K*-means algorithm, patient representations using different combinations of the features [49]. The only features we chose to omit from the clustering process were those with a frequency below a 2% threshold in both populations: past surgery procedure, rheumatoid arthritis, inflammatory bowel disease, psychosis, and restrictive lung disease. Therefore, we were left with 18 possible features for patient representation: 3 demographic (age, sex, and BMI), 14 past diagnoses, and 1 chronic medication (anticoagulants). Because the prevalence of patients with PE in our data was very low (1942/46,639, 4.16%), we looked for a cluster or more where the percentage of patients with PE was significantly higher than this threshold.

XSL•FO
RenderX

After simultaneously considering possible clustering schemes (K={2, 5}) and patient representations based on a combination of features, we found 2 cluster schemes that yielded interesting results. The first scheme (Figure 4), which yielded the cluster with the highest frequency of patients with PE (705/1120, 63%; cluster 4 in Table 6), is a 5-cluster scheme in which each patient is represented by sex, age, BMI, and past PE, past pneumonia, history of atrial fibrillation or atrial flutter, past DVT, history of mental disorders, and history of coagulopathies. We see that cluster 4, which is the most PE-populated cluster of this scheme, represents a profile that is based on both demographic and medical history features, suggesting complex relations between such characteristics that are nontrivial and need further exploration. Cluster 4 contained 1120 patients (1120/46,639, 2.4% of all patients): 705 were patients with PE (705/1942, 36.3% of all patients with PE in our data) and only 415 were controls. Prominent characteristics of this cluster compared with the others are that all its patients had a past diagnosis of PE (see Table 7), a relatively high incidence of past DVT (12% compared with 2% in any of the other clusters) and past pneumonia (11% compared with 4% or less in any of the other clusters), and a mediocre incidence of mental disorders, and they more often took anticoagulants. Patients in this cluster were equally mixed by sex. The other clusters had only a small incidence of such past diagnoses and were composed of single-sex patients (Table 7). That is, this clustering scheme made a clear distinction between a single small, but very PE dominant, cluster 4, and 4 other non-PE clusters that are mainly dominated by demographics (sex and age) and BMI characteristics. Besides showing past PE and DVT as risk factors, which are known to clinicians (see, eg, the study by Scarvelis and Wells [50]), patients in cluster 4 also reveal past pneumonia history as a factor, which is evidence of a new risk factor of PE.

The second clustering scheme (Figure 5), which resulted in 2 clusters (clusters 2 and 4) with a prevalence of patients with PE of 61% (427/701 and 340/549) or more (Table 8), was a 5-cluster schema in which patients were represented by sex, age, BMI, and past PE, past pneumonia, history of mental disorders, and use of anticoagulants. All patients in these 2 clusters had a past diagnosis of PE and a higher incidence of past diagnosis of pneumonia and usage of anticoagulants (Table 9), higher BMI values, and, interestingly, they were divided into only women (cluster 2) or only men (cluster 4). Also, cluster 3, which has a relatively high frequency of patients with PE—20% (301/1470) compared with 4.16% in the full database (Table 8), is characterized by all patients having a past diagnosis of pneumonia. That is, the chances of developing PE according to this scheme are mainly related to either past PE and past pneumonia or, to a lesser degree, only past pneumonia. Although the incidence values of a past diagnosis of mental disorders and usage of anticoagulants were low, these 2 variables were chosen for patient representation by both clustering schemes, because their contribution to patient segregation was significant in maximizing the cluster PE frequency we optimized in the clustering. The atrial fibrillation or flutter diagnosis was selected only by the first clustering scheme, producing clusters with either all subjects or none of them with this diagnosis, as this partition helped maximize the target measure. Finally, in addition to past PE, which is arguably the most important variable for marking patients with suspected PE, past diagnoses of pneumonia and DVT may also lead to PE. Thus, we recommend overattention to admitted patients with these past diagnoses.

XSL•FO
RenderX

Figure 4. Radar plot of characteristics of the first clustering scheme. A point of a particular cluster on the axis of any feature constitutes the average normalized value (between 0 and 1) for the patients in the cluster. DVT: deep vein thrombosis; PE: pulmonary embolism.

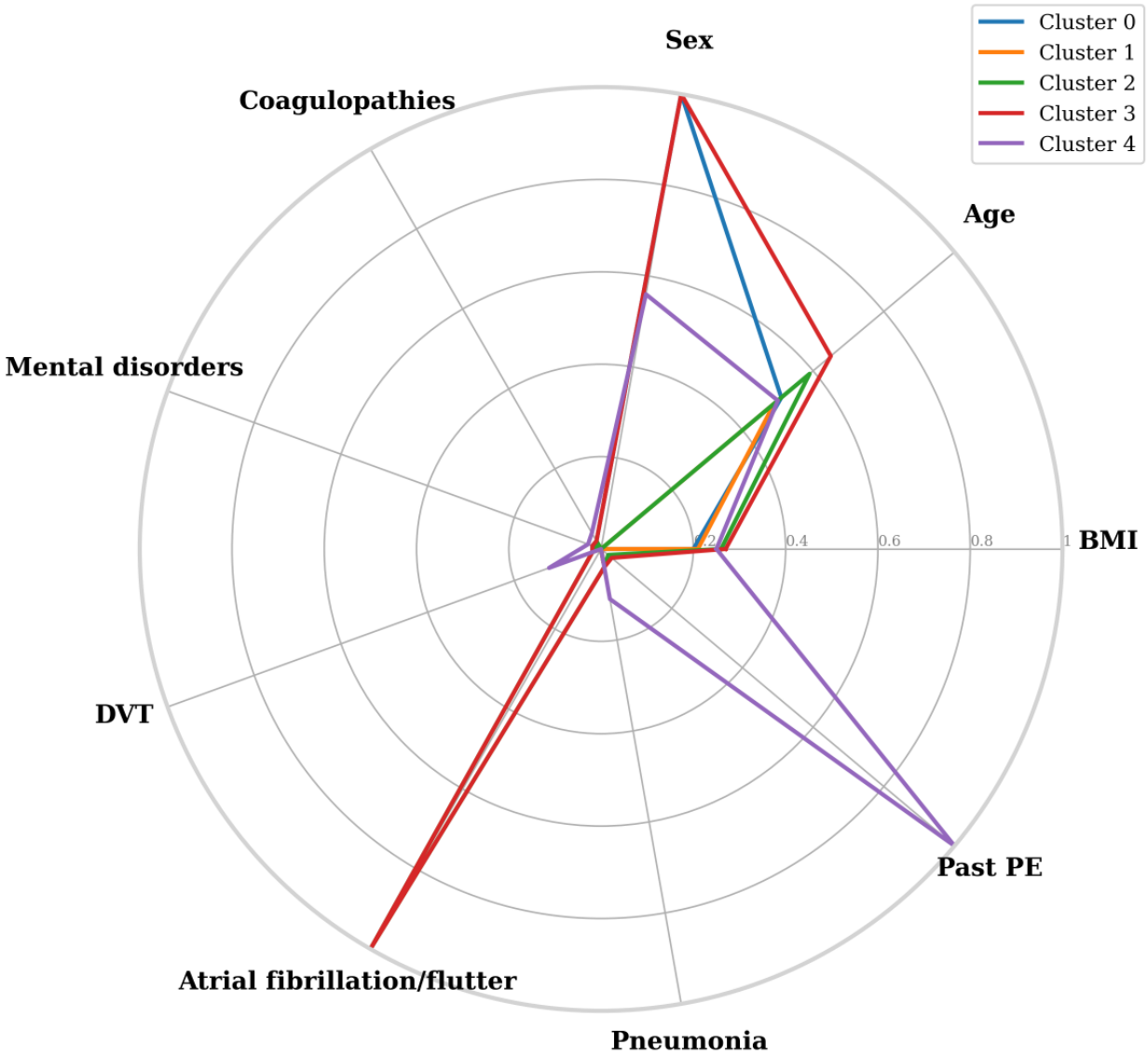

Table 6. PE[a]-control prevalence in the clusters of the first clustering scheme (the highest PE prevalence for a cluster is italicized).

Cluster, n	Relative size	Patients, n		PE/total
		PE	Control	
0	0.38	564	17,046	0.03
1	0.47	426	21,395	0.02
2	0.07	106	3179	0.03
3	0.06	141	2662	0.05
4	0.02	705	415	*0.63*

[a]PE: pulmonary embolism.

XSL•FO
RenderX

Table 7. Average feature values in the clusters of the first clustering scheme.

Cluster, n	Sex	Age (years)	BMI	Past PE[a]	Pneumonia	Atrial fibrillation/flutter	DVT[b]	Mental disorders	Coagulopathies
0	1	65.5	26.5	0	0.03	0	0.02	0.02	0.02
1	0	63.2	26.8	0	0.03	0	0.02	0.01	0.01
2	0	72.0	27.7	0.02	0.04	1	0.02	0.01	0.02
3	1	75.6	27.9	0.03	0.04	1	0.02	0.02	0.02
4	0.56	64.3	27.6	1	0.11	0	0.12	0.03	0.04

[a]PE: pulmonary embolism.

[b]DVT: deep vein thrombosis.

Figure 5. Radar plot of characteristics of the second clustering scheme. A point of a particular cluster on the axis of any feature constitutes the average normalized value (between 0 and 1) for the patients in the cluster. PE: pulmonary embolism.

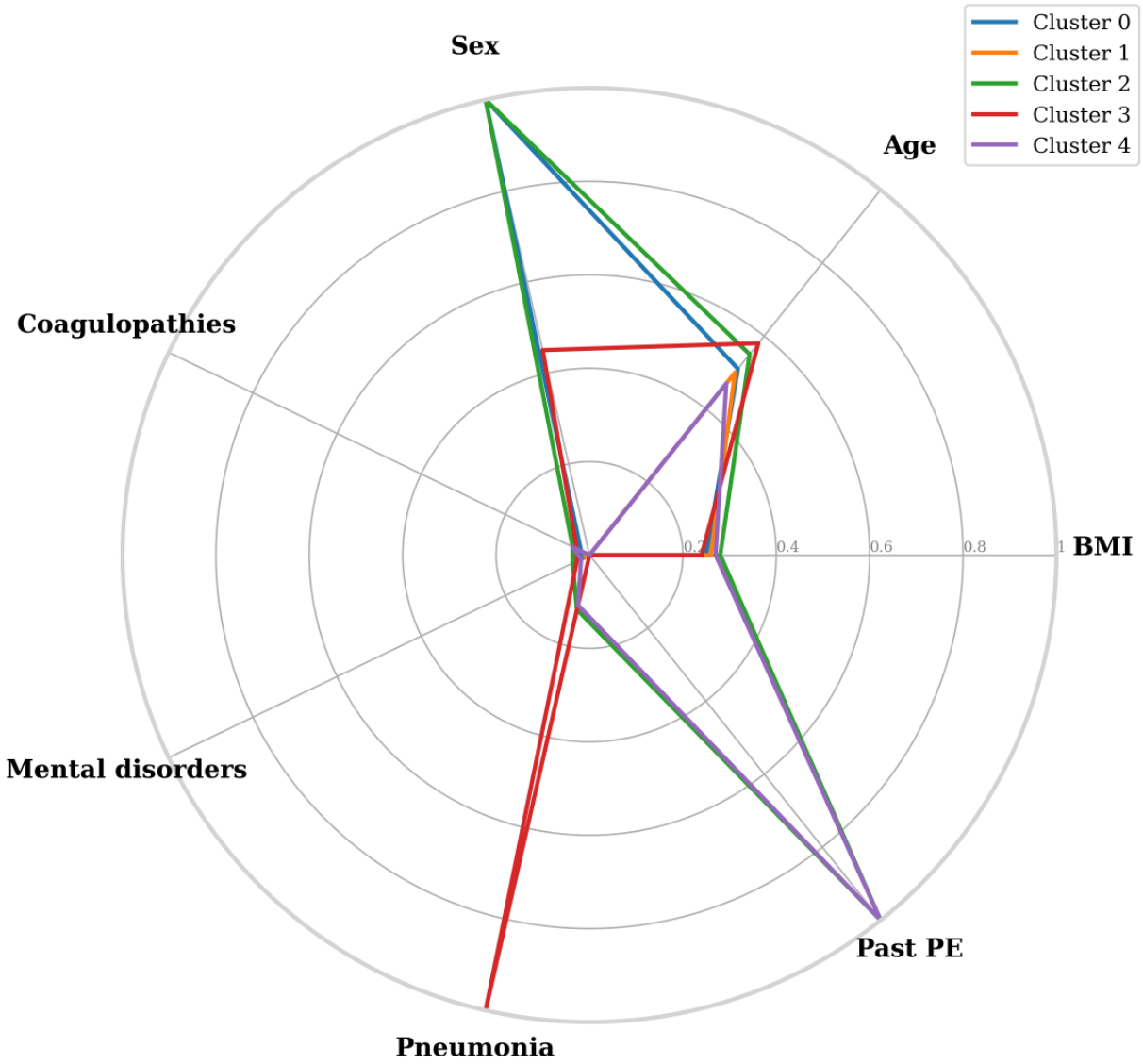

XSL•FO
RenderX

Table 8. Descriptive statistics of the clusters of the second clustering scheme[a].

Cluster, n	Relative size	Patients, n		PE/total
		PE[b]	Control	
0	0.42	510	19,161	0.03
1	0.52	364	23,884	0.02
2	0.02	427	274	*0.61*
3	0.03	301	1169	0.20
4	0.01	340	209	*0.62*

[a]Italicized values indicate the high prevalence of patients with pulmonary embolism in clusters 2 and 4 compared with the prevalence of 4.16% of these patients in the cohort.

[b]PE: pulmonary embolism.

Table 9. Average value of each feature in each cluster of the second clustering scheme[a].

Cluster, n	Sex	Age (years)	BMI	Past PE[b]	Pneumonia	Mental disorders	Coagulopathies
0	1	66.7	26.7	0	0	0.02	0.02
1	0	64.2	27.0	0	0	0.01	0.01
2	*1*	67.3	27.8	*1*	0.12	0.04	0.04
3	0.45	69.3	26.5	0	*1*	0.03	0.03
4	*0*	63.0	27.5	*1*	0.11	0.02	0.04

[a]All patients of clusters 2 and 4 (see Table 8), clusters that are mostly populated by patients with PE, have a past diagnosis of pulmonary embolism and higher incidence of past diagnosis of pneumonia and usage of anticoagulants. While cluster 2 composes of only women relatively old, cluster 4 includes only men relatively young (italicized values).

[b]PE: pulmonary embolism.

Discussion

We are not the first to address the potential benefits in applying AI to PE, as AI does not require additional clinician input or cause workflow disruption by automatically screening a broad inpatient population [14-18]. These studies reported the application of AI in interpreting chest images [14] or their accompanying radiology reports [51] and in using laboratory test results and vital signs [17,18] of already diagnosed patients with PE. In addition, these studies often targeted PE at patients who are already at elevated risk of PE [18].

To better understand and predict the occurrence and diagnosis of PE, we have suggested tools that may be clinically available on patient hospital admission—the earliest point in time to model PE in the hospital. Also, we addressed the entire population of patients presenting to a tertiary hospital, without narrowing them to only those with high probability of PE, as required by the Wells or revised Genova scores, to screen the full admitted population and not only those with higher risk of PE. Using 13 years of data from around 50,000 patients with PE and those with no PE, we considered the medical history of patients presenting in a hospital, where their prior chance of being diagnosed with PE was 4%. To create reliable models, we suggested 2 methods to address this highly imbalanced clinical scenario. The first empirically sets a classifier decision threshold to account for the minority-to-majority ratio in the imbalanced data, and the second uses an ensemble of balanced classifiers evaluating the PE class versus equal-sized disjoint portions of the imbalanced non-PE class. To identify significant predictive past diagnoses, we considered diagnoses in 2 resolutions, separately and categorized, and proposed a nonparametric statistical test to consolidate feature importance lists obtained over data permutations. An ablation study validated results based on the test. Lists produced by the 2 methods were almost identical, and classifiers trained using them provided similar performance, validating the correctness of both methods. Past diagnoses of PE, pulmonary diseases, and thrombosis, together with age, sex, BMI, and usage of anticoagulants contributed to 80% accuracy in early identification of both patients with PE and those with no PE, with no statistical difference between the methods.

Although only 4% of the patients in our data set had a definitive diagnosis of PE, we identified, by joint optimization of feature representation and a clustering scheme, subgroups comprising more than 60% PE-positive patients. These subgroups facilitated a more in-depth analysis to identify demographic and past diagnosis characteristics, demographic-clinical relations, and potential PE risk factors. Some of our findings are not conventionally considered. For example, in medical practice, a past diagnosis of pulmonary disease on admission is related to a chronic respiratory disease displaying respiratory symptoms, but we show here that this may also be an indication of PE. This is an important clinical pearl that should prompt clinicians worldwide, with and without access to AI predictive algorithms.

In summary, despite the high imbalance in the data and the very early stage of modeling, our prediction and clustering models

XSL•FO
RenderX

were both accurate and informative in identifying patients at high risk for PE, at the time of hospital admission, before a patient is even seen by a hospital doctor. The main advantage in our findings, which is relevant to every health care professional, is the fact that applying AI enables whole-patient-population analysis without the need to adhere to previously published criteria (eg, Wells or revised Genova scores), which are not sensitive enough [52], or to limit its use to high-risk patients. In the case of PE, this approach defines previous pulmonary disease, whether with past PE or alone, as a new, significant risk factor to consider.

A limitation of our research is working with data, which, due to privacy restrictions, were synthesized from real data. The difference between real data and synthetic data derived from real data is that the synthetic data include times of diagnosis and measurements that are randomly shifted to up to a year before or after the real time and thus could not be verified. Therefore, in our study, the existence of past diagnoses was expressed using binary variables (whether a diagnosis exists or not) and not by the times of diagnosis, which is vital information that could have contributed greatly to the models.

Our future research will use real data and also assimilate data from the emergency department (validating and amending, as necessary, the admission model) and then the internal ward, whenever data are collected, which may necessitate the use of dynamic modeling. In addition, while our second method was applied to a balanced test set, it should be further examined on a nonbalanced database that reflects the true PE prevalence. Finally, beyond being the basis for predictive and informative models of PE, we see our workflow as suitable for other diseases, especially since in almost all of them, an imbalanced scenario is involved.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Methods developed for imbalanced data.
[DOCX File , 166 KB - jmir_v26i1e48595_app1.docx]

Multimedia Appendix 2
A statistical test to consolidate different recommendations on feature importance.
[DOCX File , 14 KB - jmir_v26i1e48595_app2.docx]

References

1. Goldhaber SZ, Bounameaux H. Pulmonary embolism and deep vein thrombosis. Lancet 2012;379(9828):1835-1846. [doi: 10.1016/S0140-6736(11)61904-1] [Medline: 22494827]
2. Ruppert A, Steinle T, Lees M. Economic burden of venous thromboembolism: a systematic review. J Med Econ 2011;14(1):65-74 [FREE Full text] [doi: 10.3111/13696998.2010.546465] [Medline: 21222564]
3. Grosse SD, Nelson RE, Nyarko KA, Richardson LC, Raskob GE. The economic burden of incident venous thromboembolism in the United States: a review of estimated attributable healthcare costs. Thromb Res 2016;137:3-10 [FREE Full text] [doi: 10.1016/j.thromres.2015.11.033] [Medline: 26654719]
4. Gartner D, Kolisch R, Neill DB, Padman R. Machine learning approaches for early DRG classification and resource allocation. INFORMS J Comput 2015;27(4):597-808. [doi: 10.1287/ijoc.2015.0655]
5. Vacca VM, Jehle J. Acute pulmonary embolism. Nursing 2013;43(3):25-26 [FREE Full text] [doi: 10.1097/01.NURSE.0000427353.73708.9d] [Medline: 23411548]
6. Morrone D, Morrone V. Erratum: acute pulmonary embolism: focus on the clinical picture. Korean Circ J 2018;48(5):365-381 [FREE Full text] [doi: 10.4070/kcj.2017.0998] [Medline: 29968441]
7. Ma Y, Huang J, Wang Y, Wu T, Cai D, Liu Y, et al. Comparison of the Wells score with the revised Geneva score for assessing pretest probability of pulmonary embolism in hospitalized elderly patients. Eur J Intern Med 2016;36:e18-e19. [doi: 10.1016/j.ejim.2016.09.003] [Medline: 27650507]
8. Wong DD, Ramaseshan G, Mendelson RM. Comparison of the Wells and revised Geneva scores for the diagnosis of pulmonary embolism: an Australian experience. Intern Med J 2011;41(3):258-263. [doi: 10.1111/j.1445-5994.2010.02204.x] [Medline: 20214691]
9. Le Gal G, Righini M, Roy PM, Sanchez O, Aujesky D, Bounameaux H, et al. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. Ann Intern Med 2006;144(3):165. [doi: 10.7326/0003-4819-144-3-200602070-00004] [Medline: 16461960]
10. Ceriani E, Combescure C, Le Gal G, Nendaz M, Perneger T, Bounameaux H, et al. Clinical prediction rules for pulmonary embolism: a systematic review and meta-analysis. J Thromb Haemost 2010;8(5):957-970 [FREE Full text] [doi: 10.1111/j.1538-7836.2010.03801.x] [Medline: 20149072]

XSL•FO
RenderX

11. van Es N, Takada T, Kraaijpoel N, Klok FA, Stals MAM, Büller HR, et al. Diagnostic management of acute pulmonary embolism: a prediction model based on a patient data meta-analysis. Eur Heart J 2023;44(32):3073-3081 [FREE Full text] [doi: 10.1093/eurheartj/ehad417] [Medline: 37452732]
12. Miniati M, Bottai M, Monti S, Salvadori M, Serasini L, Passera M. Simple and accurate prediction of the clinical probability of pulmonary embolism. Am J Respir Crit Care Med 2008;178(3):290-294. [doi: 10.1164/rccm.200802-207OC] [Medline: 18436792]
13. Wicki J, Perrier A, Perneger TV, Bounameaux H, Junod AF. Predicting adverse outcome in patients with acute pulmonary embolism: a risk score. Thromb Haemost 2000;84(4):548-552. [Medline: 11057848]
14. Banerjee I, Sofela M, Yang J, Chen JH, Shah NH, Ball R, et al. Development and performance of the pulmonary embolism Result Forecast Model (PERFORM) for computed tomography clinical decision support. JAMA Netw Open 2019;2(8):e198719 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.8719] [Medline: 31390040]
15. Rucco M, Sousa-Rodrigues D, Merelli E, Johnson JH, Falsetti L, Nitti C, et al. Neural hypernetwork approach for pulmonary embolism diagnosis. BMC Res Notes 2015;8:617 [FREE Full text] [doi: 10.1186/s13104-015-1554-5] [Medline: 26515513]
16. Ji QY, Wang MF, Su CM, Yang QF, Feng LF, Zhao LY, et al. Clinical symptoms and related risk factors in pulmonary embolism patients and cluster analysis based on these symptoms. Sci Rep 2017;7(1):14887 [FREE Full text] [doi: 10.1038/s41598-017-14888-7] [Medline: 29097743]
17. Ryan L, Maharjan J, Mataraso S, Barnes G, Hoffman J, Mao Q, et al. Predicting pulmonary embolism among hospitalized patients with machine learning algorithms. Pulm Circ 2022;12(1):e12013 [FREE Full text] [doi: 10.1002/pul2.12013] [Medline: 35506114]
18. Shen J, Casie Chetty S, Shokouhi S, Maharjan J, Chuba Y, Calvert J, et al. Massive external validation of a machine learning algorithm to predict pulmonary embolism in hospitalized patients. Thromb Res 2022;216:14-21 [FREE Full text] [doi: 10.1016/j.thromres.2022.05.016] [Medline: 35679633]
19. Zhang H, Zhang H, Pirbhulal S, Wu W, De Albuquerque VHC. Active balancing mechanism for imbalanced medical data in deep learning–based classification models. ACM Trans Multimedia Comput Commun Appl 2020;16(1s):1-15. [doi: 10.1145/3357253]
20. Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognit Lett 2020;136:190-197. [doi: 10.1016/j.patrec.2020.05.035]
21. Kvamme H, Borgan O, Scheel I. Time-to-event prediction with neural networks and Cox regression. J Mach Learn Res 2019;20:1-30.
22. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 2018;61:863-905. [doi: 10.1613/jair.1.11192]
23. Lerner B, Yeshaya J, Koushnir L. On the classification of a small imbalanced cytogenetic image database. IEEE/ACM Trans Comput Biol Bioinform 2007;4(2):204-215 [FREE Full text] [doi: 10.1109/TCBB.2007.070207] [Medline: 17473314]
24. Halbersberg D, Wienreb M, Lerner B. Joint maximization of accuracy and information for learning the structure of a Bayesian network classifier. Mach Learn 2020;109:1039-1099. [doi: 10.1007/s10994-020-05869-5]
25. Halbersberg D, Lerner B. Young driver fatal motorcycle accident analysis by jointly maximizing accuracy and information. Accid Anal Prev 2019;129:350-361. [doi: 10.1016/j.aap.2019.04.016] [Medline: 31201968]
26. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning. ACM Comput Surv 2019;52(4):1-36. [doi: 10.1145/3343440]
27. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
28. Lobo JM, Jiménez‐Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. GEB 2007;17:145-151. [doi: 10.1111/j.1466-8238.2007.00358.x]
29. Kuncheva LI, Arnaiz-González Á, Díez-Pastor JF, Gunn IAD. Instance selection improves geometric mean accuracy: a study on imbalanced data classification. Prog Artif Intell 2019;8(2):215-228. [doi: 10.1007/s13748-019-00172-4]
30. Jadhav AS. A novel weighted TPR-TNR measure to assess performance of the classifiers. Expert Syst Appl 2020;152:113391. [doi: 10.1016/j.eswa.2020.113391]
31. Zhu Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognit Lett 2020;136:71-80. [doi: 10.1016/j.patrec.2020.03.030]
32. Provost F. Machine learning from imbalanced data sets 101. Proc. AAAI-2000 Work 2000;Online:3 [FREE Full text]
33. Lee W, Seo K. Active downsampling for binary classification with an imbalanced dataset. 2021 Presented at: Proceedings of the IISE Annual Conference; 2021; Columbia, MO p. 187-192.
34. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. Lect Notes Comput Sci 2005;3408:345-359. [doi: 10.1007/978-3-540-31865-1_25]
35. MDClone. The world's most powerful healthcare data platforme. URL: https://www.mdclone.com/ [accessed 2024-03-08]
36. Foraker RE, Yu SC, Gupta A, Michelson AP, Pineda Soto JA, Colvin R, et al. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. JAMIA Open 2020;3(4):557-566 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa060] [Medline: 33623891]

XSL•FO
RenderX

37. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. JMIR Med Inform 2020;8(2):e16492 [FREE Full text] [doi: 10.2196/16492] [Medline: 32130148]
38. Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. 2001 Presented at: International Conference on Data Engineering; June 06 2024; Canada. [doi: 10.1109/icde.2001.914830]
39. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 2009;10:213 [FREE Full text] [doi: 10.1186/1471-2105-10-213] [Medline: 19591666]
40. Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowledge Data Eng 2003;15(6):1437-1447. [doi: 10.1109/tkde.2003.1245283]
41. Gordon J, Lerner B. Insights into amyotrophic lateral sclerosis from a machine learning perspective. J Clin Med 2019;8(10):1578 [FREE Full text] [doi: 10.3390/jcm8101578] [Medline: 31581566]
42. Hever G, Cohen L, O'Connor MF, Matot I, Lerner B, Bitan Y. Machine learning applied to multi-sensor information to reduce false alarm rate in the ICU. J Clin Monit Comput 2020;34(2):339-352. [doi: 10.1007/s10877-019-00307-x] [Medline: 30955160]
43. Kashi S, Polak RF, Lerner B, Rokach L, Levy-Tzedek S. A Machine-Learning Model for automatic detection of movement compensations in stroke patients. IEEE Trans Emerg Topics Comput 2021;9(3):1234-1247. [doi: 10.1109/tetc.2020.2988945]
44. Avisar H, Guardia-Laguarta C, Area-Gomez E, Surface M, Chan AK, Alcalay RN, et al. Lipidomics prediction of Parkinson's disease severity: a machine-learning analysis. J Parkinsons Dis 2021;11(3):1141-1155 [FREE Full text] [doi: 10.3233/JPD-202476] [Medline: 33814463]
45. Avisar H, Guardia-Laguarta C, Surface M, Papagiannakis N, Maniati M, Antonellou R, et al. Lipid level alteration in human and cellular models of alpha synuclein mutations. NPJ Parkinsons Dis 2022;8(1):52 [FREE Full text] [doi: 10.1038/s41531-022-00313-y] [Medline: 35468903]
46. Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1-30.
47. Yildiz OT, Aslan O, Alpaydin E. Multivariate statistical tests for comparing classification algorithms. : Springer; 2011 Presented at: International Conference on Learning and Intelligent Optimization; 2011; Berlin, Heidelberg p. 1-15 URL: https://doi.org/10.1007/978-3-642-25566-3_1
48. Alam MZ, Rahman MS, Rahman MS. A random forest based predictor for medical data classification using feature ranking. Informatics Med Unlocked 2019;15:100180. [doi: 10.1016/j.imu.2019.100180]
49. Sinaga KP, Yang MS. Unsupervised K-Means clustering algorithm. IEEE Access 2020;8:80716-80727. [doi: 10.1109/access.2020.2988796]
50. Scarvelis D, Wells PS. Diagnosis and treatment of deep-vein thrombosis. CMAJ 2006;175(9):1087-1092. [doi: 10.1503/cmaj.060366]
51. Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med 2019;97:79-88 [FREE Full text] [doi: 10.1016/j.artmed.2018.11.004] [Medline: 30477892]
52. Shen JH, Chen HL, Chen JR, Xing JL, Gu P, Zhu BF. Comparison of the Wells score with the revised Geneva score for assessing suspected pulmonary embolism: a systematic review and meta-analysis. J Thromb Thrombolysis 2016;41(3):482-492. [doi: 10.1007/s11239-015-1250-2] [Medline: 26178041]

Abbreviations

AI: artificial intelligence
CTA: computed tomographic angiography
DVT: deep vein thrombosis
ML: machine learning
PE: pulmonary embolism
PPV: positive predictive value
RF: random forest
SMC: Sheba Medical Center
TPR: true-positive rate
VTE: venous thromboembolism

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, T Leung, G Eysenbach; submitted 29.04.23; peer-reviewed by D Gartner, JH Rajendran, J Zeng; comments to author 10.10.23; revised version received 02.12.23; accepted 30.04.24; published 30.07.24.

Please cite as:
Ben Yehuda O, Itelman E, Vaisman A, Segal G, Lerner B
Early Detection of Pulmonary Embolism in a General Patient Population Immediately Upon Hospital Admission Using Machine Learning to Identify New, Unidentified Risk Factors: Model Development Study
J Med Internet Res 2024;26:e48595
URL: https://www.jmir.org/2024/1/e48595
doi:10.2196/48595
PMID:

©Ori Ben Yehuda, Edward Itelman, Adva Vaisman, Gad Segal, Boaz Lerner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Machine Learning Model for Predicting In-Hospital Mortality in Chinese Patients With ST-Segment Elevation Myocardial Infarction: Findings From the China Myocardial Infarction Registry

Jingang Yang[1], MD, PhD[‡]; Yingxue Li[2], PhD; Xiang Li[2], PhD; Shuiying Tao[1], MD; Yuan Zhang[2], PhD; Tiange Chen[2], PhD; Guotong Xie[2], PhD; Haiyan Xu[1], MD, PhD; Xiaojin Gao[1], MD, PhD; Yuejin Yang[1], MD, PhD

[1]State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[2]Ping An Healthcare and Technology, Beijing, China

[‡]China Myocardial Infarction Registry Research Group

Corresponding Author:
Yuejin Yang, MD, PhD
State Key Laboratory of Cardiovascular Disease
Fuwai Hospital, National Center for Cardiovascular Diseases
Chinese Academy of Medical Sciences & Peking Union Medical College
No 167, Beilishi Road
Xicheng District
Beijing, 100037
China
Phone: 86 13701151408
Email: Yangyjfw@126.com

Abstract

Background: Machine learning (ML) risk prediction models, although much more accurate than traditional statistical methods, are inconvenient to use in clinical practice due to their nontransparency and requirement of a large number of input variables.

Objective: We aimed to develop a precise, explainable, and flexible ML model to predict the risk of in-hospital mortality in patients with ST-segment elevation myocardial infarction (STEMI).

Methods: This study recruited 18,744 patients enrolled in the 2013 China Acute Myocardial Infarction (CAMI) registry and 12,018 patients from the China Patient-Centered Evaluative Assessment of Cardiac Events (PEACE)-Retrospective Acute Myocardial Infarction Study. The Extreme Gradient Boosting (XGBoost) model was derived from 9616 patients in the CAMI registry (2014, 89 variables) with 5-fold cross-validation and validated on both the 9125 patients in the CAMI registry (89 variables) and the independent China PEACE cohort (10 variables). The Shapley Additive Explanations (SHAP) approach was employed to interpret the complex relationships embedded in the proposed model.

Results: In the XGBoost model for predicting all-cause in-hospital mortality, the variables with the top 8 most important scores were age, left ventricular ejection fraction, Killip class, heart rate, creatinine, blood glucose, white blood cell count, and use of angiotensin-converting enzyme inhibitors (ACEIs) and angiotensin II receptor blockers (ARBs). The area under the curve (AUC) on the CAMI validation set was 0.896 (95% CI 0.884-0.909), significantly higher than the previous models. The AUC for the Global Registry of Acute Coronary Events (GRACE) model was 0.809 (95% CI 0.790-0.828), and for the TIMI model, it was 0.782 (95% CI 0.763-0.800). Despite the China PEACE validation set only having 10 available variables, the AUC reached 0.840 (0.829-0.852), showing a substantial improvement to the GRACE (0.762, 95% CI 0.748-0.776) and TIMI (0.789, 95% CI 0.776-0.803) scores. Several novel and nonlinear relationships were discovered between patients' characteristics and in-hospital mortality, including a U-shape pattern of high-density lipoprotein cholesterol (HDL-C).

Conclusions: The proposed ML risk prediction model was highly accurate in predicting in-hospital mortality. Its flexible and explainable characteristics make the model convenient to use in clinical practice and could help guide patient management.

Trial Registration: ClinicalTrials.gov NCT01874691; https://clinicaltrials.gov/study/NCT01874691

(J Med Internet Res 2024;26:e50067) doi:10.2196/50067

XSL•FO
RenderX

KEYWORDS

ST-elevation myocardial infarction; in-hospital mortality; risk prediction; explainable machine learning; machine learning; acute myocardial infarction; myocardial infarction; mortality; risk; predication model; china; clinical practice; validation; patient management; management

Introduction

Acute myocardial infarction (AMI) is a major cause of hospitalization and mortality in China, while ST-segment elevation myocardial infarction (STEMI) accounts for over 80% of myocardial infarctions [1-3]. It is critical to accurately predict the risks of in-hospital mortality for patients with STEMI to improve prognosis. Traditionally, most risk prediction models have been based on generalized linear regression methods [4,5]. Although straightforward to understand and apply, these models require parametric assumptions [6,7]. For example, using the logistic regression (LR) method, the Global Registry in Acute Coronary Events (GRACE) [4] and Thrombolysis in Myocardial Infarction (TIMI) risk scores [5] oversimplified the complexity of the real association among variables and outcome, resulting in poor predictive accuracy [8,9]. Recently, machine learning (ML) techniques have been increasingly used for predicting different clinical events in cardiovascular disease [10-12] and have achieved higher accuracy than traditional models. However, ML models, often built on a large number of variables, are difficult to use in clinical practice due to the need for extensive input data and the challenge of identifying specific therapeutic targets. The complexity and ambiguity of ML models require a shift toward explainable artificial intelligence (XAI) methods to guarantee that the model outputs are comprehensible for end users [13]. Moreover, ML models tend to use a large number of variables [14-16]. However, in clinical practice, where many scenarios are unknown, a significant challenge is how to apply the model more flexibly when some variables are missing. Therefore, we aimed to develop an ML risk prediction model for in-hospital mortality in patients with STEMI that is not only highly accurate but also explainable and flexible with the number of input variables (tolerant to the missing variables), making it easy to use in clinical practice.

Methods

Data Description

The patients included in this study were from the China Acute Myocardial Infarction (CAMI) registry [3], organized and conducted by the Fuwai Hospital, National Center for Cardiovascular Diseases, China, from January 2013 to September 2014. The methodology of the CAMI registry (NCT01874691) has been previously described [3]. In short, the CAMI registry was a prospective, nationwide, multicenter observational study for patients with AMI. The registry included 3 levels of hospitals (provincial, prefecture, and county), reflecting the typical Chinese governmental and administrative model and providing broad geographic representation across all provinces and municipalities across mainland China. Patients with AMI were consecutively enrolled, and data were collected upon their arrival and throughout their hospital stay until discharge. Data were collected, validated, and submitted by trained clinical cardiologists or cardiovascular fellows to ensure accuracy and reliability at each participating site. Patients diagnosed as non-STEMI (NSTEMI) or lack of in-hospital mortality status were excluded from the study.

The CAMI registry data were used for model derivation and internal validation. Patients with STEMI hospitalized in 2014 (n=9616, 51.3%) were used to derive the model, while those hospitalized in 2013 (n=9125, 48.7%) were used for internal validation. An independent cohort of patients from the China Patient-Centered Evaluative Assessment of Cardiac Events (PEACE)-Retrospective Acute Myocardial Infarction Study [2], another nationally representative sample of patients with STEMI spanning from 2001 to 2011 (N=12,108), was also used to externally validate the proposed risk prediction model. We only selected 10 important variables to carry out the validation, with the aim of assessing the proposed risk prediction model's flexibility when applied in daily clinical practice. The internal validation set sampled at a different time point, along with the independent external validation set, were both used to assess the model's reproducibility and generalizability to new and different patients.

Ethical Considerations

Both study protocols conformed to the ethical guidelines of the 1975 Declaration of Helsinki and were approved by the ethics review board committee of Fuwai Hospital (431) [2,3]. Written informed consent was obtained from eligible patients before registration. All data were anonymized.

Main Outcome

The main outcome was all-cause in-hospital mortality, defined as death for any reason during hospitalization.

Predictor Variables

The patients with STEMI included in the CAMI cohort were characterized by a total of 89 variables (Table S1 in Multimedia Appendix 1), including social demographics, presentation characteristics, laboratory tests, treatment, medical history, and more [3]. The patients with STEMI included in the China PEACE-Retrospective Acute Myocardial Infarction Study [2] were characterized by 10 variables, including age, weight, Killip class, heart rate, systolic blood pressure (SBP), glucose, creatinine, white blood cell (WBC) count, high-density lipoprotein cholesterol (HDL-C), and use of angiotensin-converting enzyme inhibitors (ACEIs) or angiotensin II receptor blockers (ARBs).

Explainable ML Analysis

Model Construction

The predictive model was developed using the Extreme Gradient Boosting (XGBoost) [17] approach based on the CAMI derivation set. XGBoost ensembles [18] a series of relatively weak base classifiers (typically decision trees) into a stronger

XSL•FO
RenderX

one sequentially and has achieved state-of-the-art results in many clinical challenges [10,19]. Apart from its highly competitive and accurate predictive performance, we chose the XGBoost method for its ability to handle missing data automatically [17]. Users do not need to impute the missing values when deriving, validating, and applying the XGBoost model. XGBoost provides the importance score of each variable, representing the frequency that a variable is used across all trees. The hyperparameters in the XGBoost model were tuned by 5-fold cross-validation on the derivation set.

Model Interpretation

The Shapley Additive Explanations (SHAP) method [20] was used to interpret the derived XGBoost model. It offers explanations on how the XGBoost model makes predictions and interprets the complex nonlinear relationship among the predictors and outcomes [19]. This method has been applied recently in other clinical studies [10,19]. SHAP represents the predicted risk as a cumulative effect of contributing variables for each prediction. The variable impact values that SHAP computes essentially represent the change in the predicted risk of the XGBoost model when we observe a feature (such as the weight of a patient) versus when we do not observe the feature (such as not knowing a patient's weight).

Model's Flexibility in Application

XGBoost's ability to handle missing values automatically makes it a robust and flexible choice for dealing with input variables. Users are free to input any number of available variables and leave other unrecorded ones as "N/A" (not available) values. Several experiments were conducted to assess the XGBoost model's flexibility. First, we retained the top 20, 15, and 10 most important variables and replaced the others with "N/A" values on the CAMI derivation set. Second, we randomly reduced the number of available variables from 89 to 10 in the CAMI validation set (Multimedia Appendix 1). Third, we included 10 variables from the independent China PEACE data set for our analysis.

Statistical Analysis

Descriptive statistics were estimated as mean (SD) for the continuous variables and frequency (percentage) for the categorical ones. The missing rates for each variable were also calculated. Missing values were imputed using the chained equation method proposed in the Multiple Imputation by Chained Equations (MICE) algorithm [21], as the models being compared—namely, lasso LR, random forest, TIMI scores, and GRACE scores—cannot handle missing data automatically. The discrimination ability was estimated by the area under the curve (AUC). Isotonic regression [22] was used downstream of the XGBoost model to adjust the predictions [23,24]. The calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test [25] on the CAMI derivation set. Additionally, a decile plot of observed versus predicted risk was used to visualize the calibration.

Results

Overview

The in-hospital mortality rate was 6.9% (n=663), 6.8% (n=621), and 9.3% (n=1132) in the CAMI derivation, validation, and China PEACE sets, respectively. The descriptive statistics of the CAMI and China PEACE data set are illustrated in Table S2 in Multimedia Appendix 1, while the missing rates are listed in Table S3 in Multimedia Appendix 1.

Prediction of In-Hospital Mortality

Figure 1 illustrates the receiver operating characteristic (ROC) curves of all the compared models. XGBoost produced the highest discrimination performance for in-hospital mortality with an AUC of 0.896 (95% CI 0.884-0.909; $P<.05$) on CAMI validation set, better than the 2 compared ML methods: random forest (AUC 0.861, 95% CI 0.845-0.876) and LR with lasso penalty (0.850, 95% CI 0.834-0.866). The XGBoost model also exhibited a significant improvement over the 2 well-established models: GRACE score (AUC 0.809, 95% CI 0.790-0.828) and TIMI score (AUC 0.782, 95% CI 0.763-0.800). All comparisons were statistically significant when $P<.05$.

XSL•FO
RenderX

Figure 1. Receiver operating characteristic (ROC) curve of different compared models on the China Acute Myocardial Infarction (CAMI) validation set. GRACE: Global Registry in Acute Coronary Events; LR: logistic regression; TIMI: thrombolysis in myocardial infarction; XGBoost: Extreme Gradient Boosting.

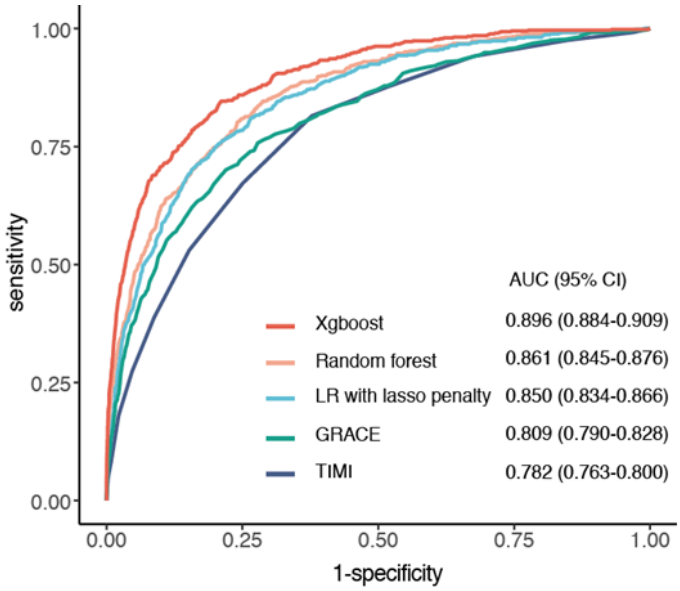

The Hosmer-Lemeshow statistic for the XGBoost model was 2.378 (*P*=.97), indicating a very good calibration result. The decile plot further confirmed strong agreement between XGBoost predicted probability and the observed in-hospital mortality risk (Figure 2).

Figure 2. Calibration plot for the Extreme Gradient Boosting (XGBoost) model. The calibration plot shows the relationship between the observed and predicted in-hospital mortality, grouped by deciles of predicted risk. The XGBoost model showed excellent calibration with the observed in-hospital mortality.

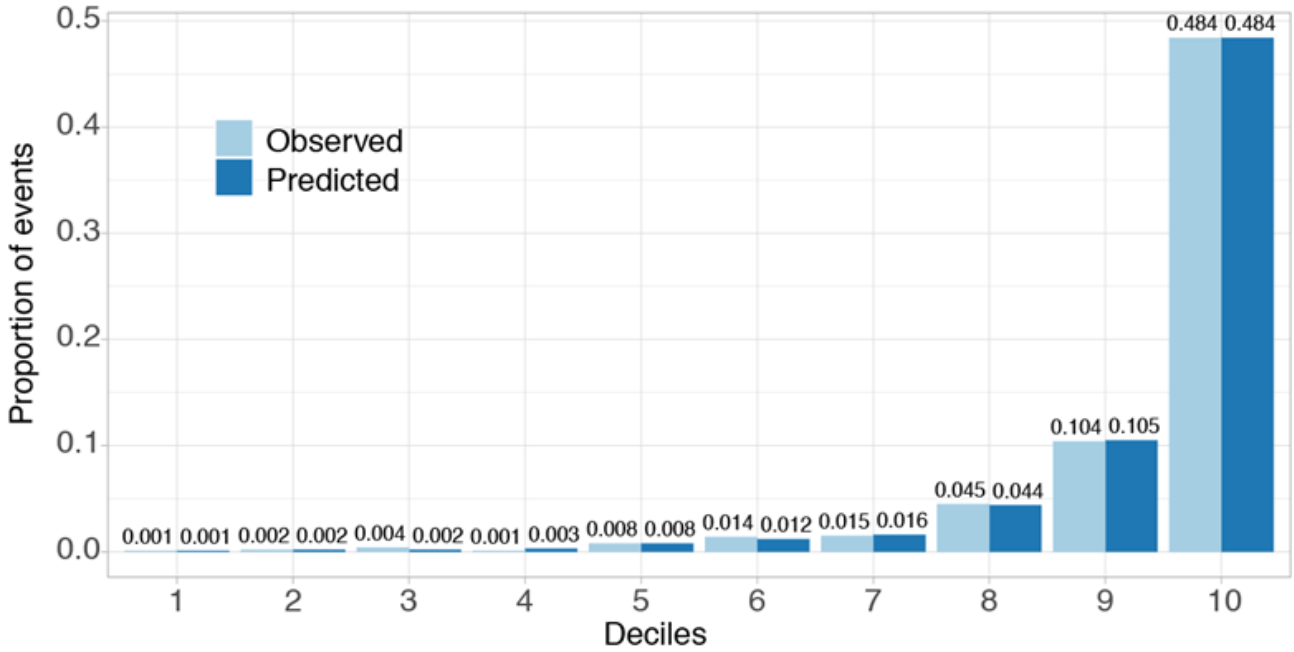

The hyperparameters for XGBoost and random forest, tuned by 5-fold cross-validation, are listed in Tables S4 and S5 in Multimedia Appendix 1.

Model Interpretation

Figure 3 illustrates the variable importance score in the XGBoost model, reflecting the frequency with which a variable was used across all trees. Age was the most important predictor of in-hospital mortality, followed by left ventricular ejection fraction (LVEF), Killip class, heart rate, creatinine, and blood glucose.

XSL•FO
RenderX

Figure 3. Variable importance score in the Extreme Gradient Boosting (XGBoost) model. A total of 48 variables with importance scores over 0 are illustrated. The color of the bar represents the variable category. ACEI: angiotensin-converting enzyme inhibitor; ARB: angiotensin II receptor blocker; AWI: anterior wall infarction; CCB: calcium channel blocker; CKMB: creatine kinase Mb isoenzyme; GRAN: neutrophilic granulocyte; HCT: Hematocrit; HDL-C: high-density lipoprotein-cholesterol; LDL-C: low-density lipoprotein-cholesterol; LVEDd: left ventricular end diastolic diameter; LVEF: left ventricular ejection fraction; PCI: percutaneous coronary intervention; SBP: systolic blood pressure; TC: total cholesterol; TCM: traditional Chinese medicine; TG: triglyceride; WBC: white blood cell.

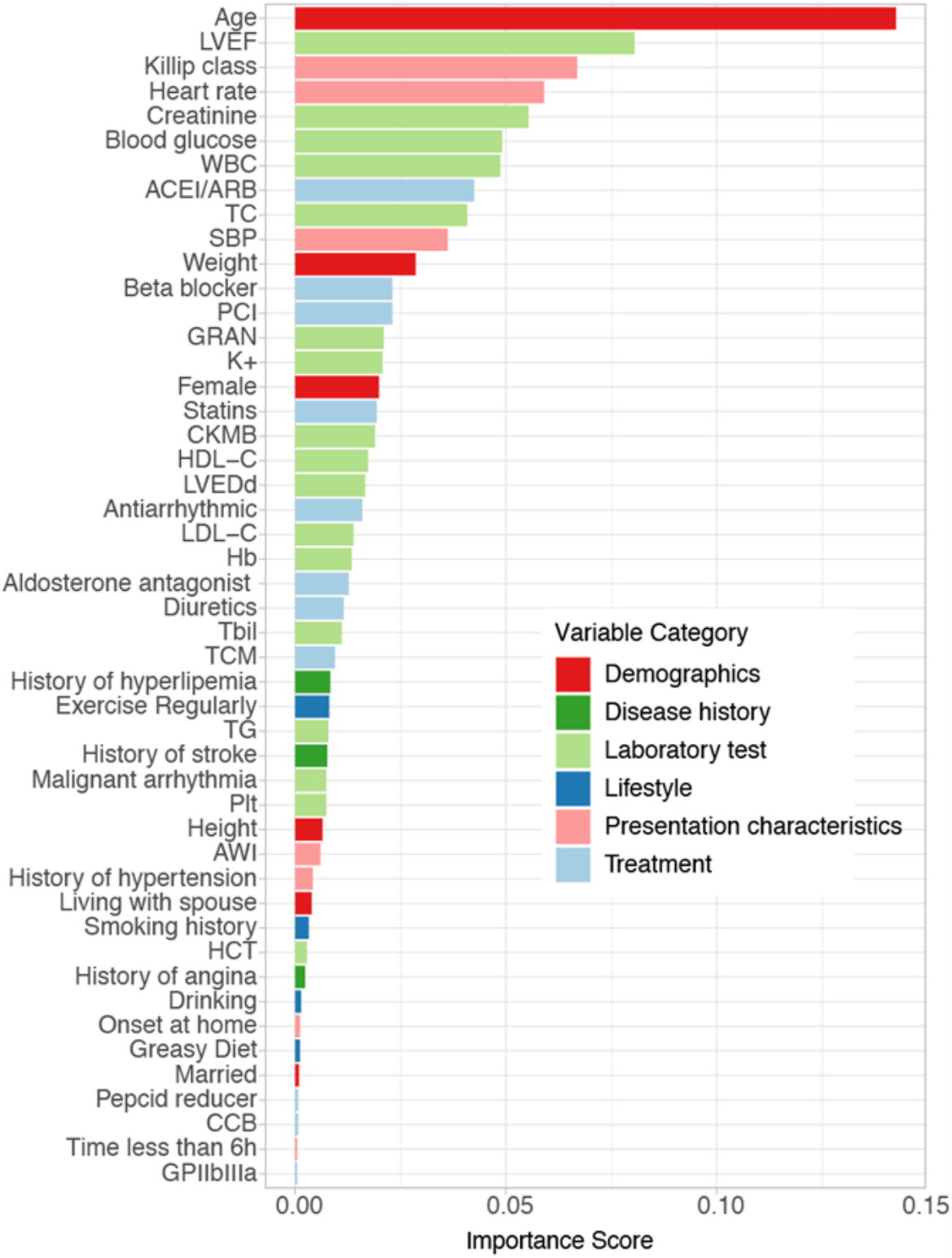

Figure 4 explains the rationale behind the model's prediction of an individual's risk. It displays the relative contributions of all features toward the predicted risk of in-hospital mortality. For instance, a predicted risk value of 0.01 for illustrated patient A was influenced by variables such as Killip class, LVEF, age, weight, and use of ACEI/ARB, among others. The red bars in Figure 3 indicate variables that increase the risk (pushing to the right), while the blue bars indicate variables that decrease the risk (pushing to the left). The length of each bar corresponds to the magnitude of its effect.

XSL•FO
RenderX

Figure 4. Sample of predicted risk of in-hospital mortality for a selected patient. ACEI: angiotensin-converting enzyme inhibitor; ARB: angiotensin II receptor blocker; HR: heart rate; Glu: glucose, LVEF: left ventricular ejection fraction; PCI: percutaneous coronary intervention; SBP: systolic blood pressure; WBC: white blood cell.

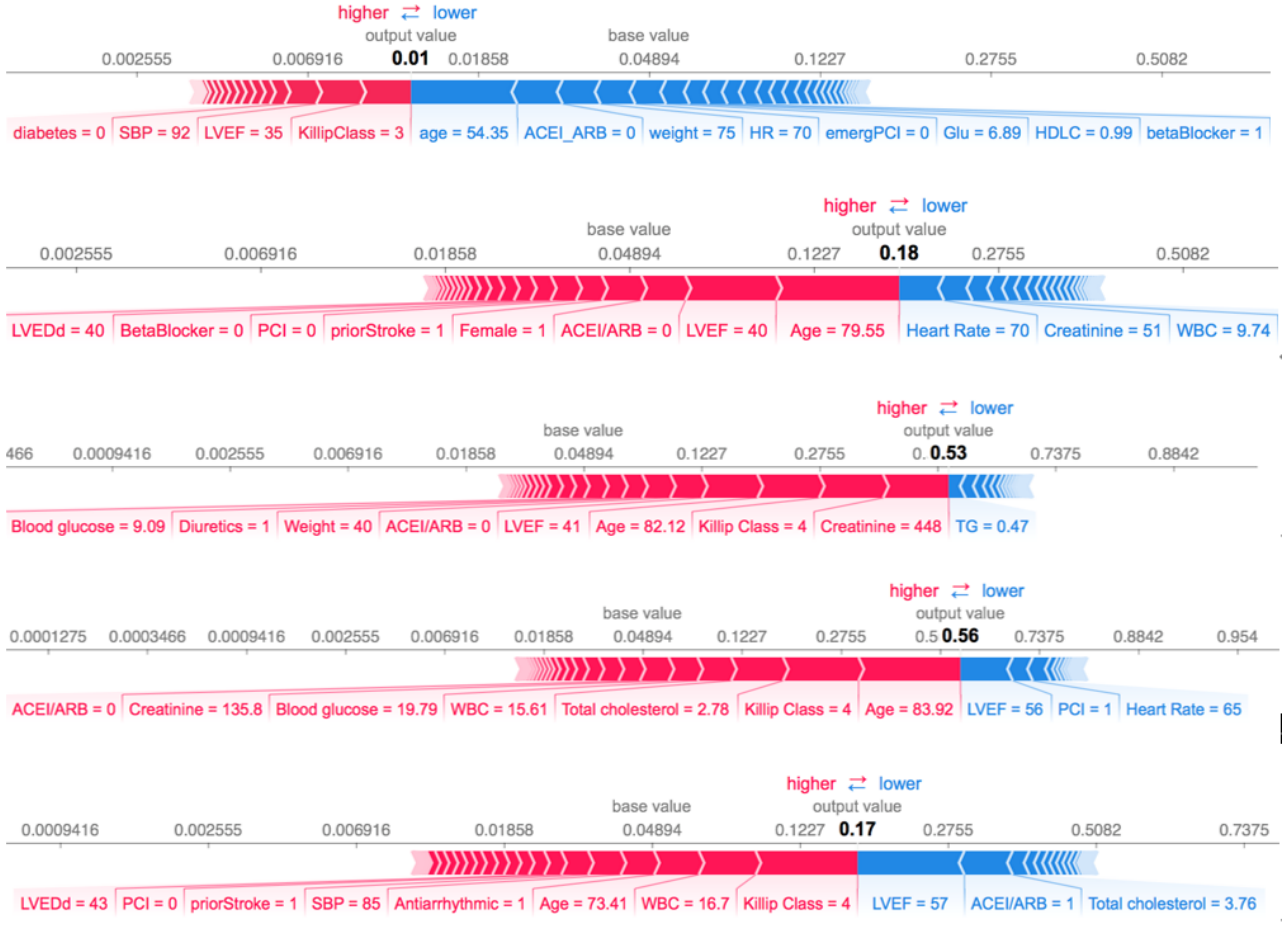

Figure 5 shows important novel and nonlinear relationships between individual variables and in-hospital mortality risk captured by the XGBoost model. For example, when age was less than 56 years, its attribution to in-hospital mortality was consistent and increased linearly after age was higher than 56 (J-shaped relationship). The heart rate variable displayed an S-shaped relationship with in-hospital mortality risk. The risk increased linearly after the heart rate was higher than 73 bpm and almost doubled until it reached 125 bpm. LVEF followed an inverted S-shaped pattern. Creatinine increased linearly until 26 and became constant after that (inverted J-shaped relationship), similar to WBC. Higher blood glucose reflected an increased in-hospital mortality risk. Variables like total cholesterol, SBP, and weight showed an L-shaped pattern. An N-shaped relationship was shown for neutrophilic granulocytes. Patients with neutrophilic granulocytes between 77% and 90% were predicted to have a higher in-hospital mortality risk. HDL-C displayed a U-shaped pattern. For potassium, a value between 4.13 and 4.49 mmol/L predicted the lowest in-hospital mortality risk.

XSL•FO
RenderX

Figure 5. Effect of varying individual variable values on the in-hospital mortality. These partial dependence plots show the change in in-hospital mortality risk for all values of a given feature. The grey histograms on each plot show the distribution of values for that variable in the derivation data set. Each point in red was obtained from one prediction. The green lines were fitted from the points by smoothed conditional means. HDL-C: high-density lipoprotein cholesterol; LVEF: left ventricular ejection fraction; SBP: systolic blood pressure; WBC: white blood cell.

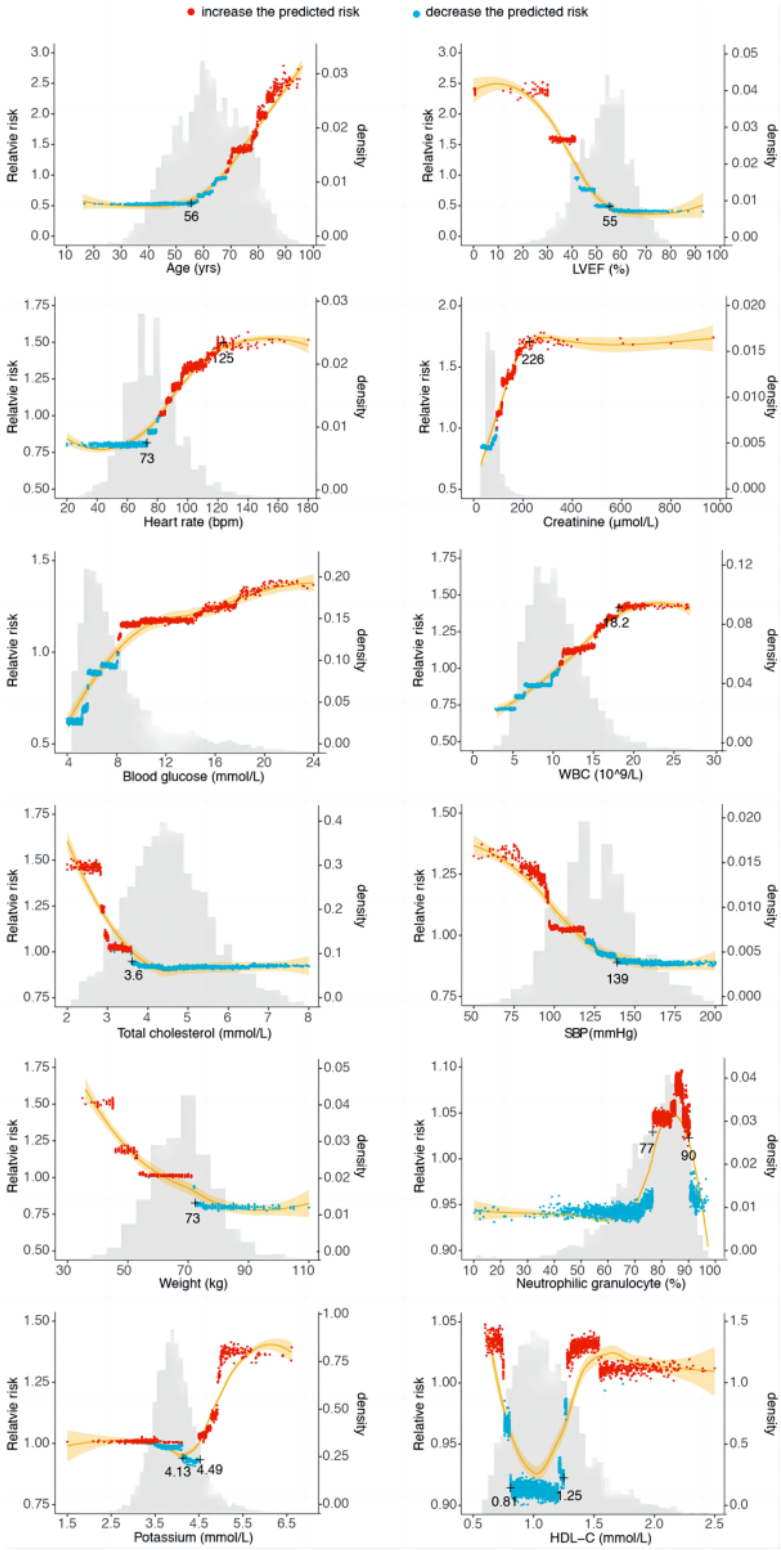

Flexibility of the Predictive Model

When we retained the top 20, 15, and 10 most important variables (Figure 2) and replaced others as "N/A" values in the CAMI validation set, the XGBoost model still achieved an AUC of 0.892 (95% CI 0.879-0.905), 0.885(95% CI 0.872-0.899), and 0.877(95% CI: 0.862 0.891), respectively. When the number of retained variables was reduced randomly from 89 to 10, the AUC decreased from 0.896 to 0.825 (SD 0.020) (20 available variables) to 0.810 (SD 0.011) (10 available variables) (Figure S1 in Multimedia Appendix 1). When the XGBoost model was validated on the China PEACE data set with the top 10 available variables (Figure 2), it achieved an AUC of 0.840 (95% CI 0.829-0.852). For comparison, the TIMI score and GRACE

XSL•FO
RenderX

score applied to the China PEACE data set gained AUCs of 0.762 (95% CI 0.748-0.776) and 0.789 (95% CI 0.776-0.803). The XGBoost model still significantly outperformed the conventional TIMI and GRACE risk score models.

For practical convenience, we embedded the XGBoost prediction model in a web-based calculator that required only the top 10 most important variables as inputs [19].

Discussion

In this study, we proposed a risk model that predicted in-hospital mortality for patients with STEMI by incorporating the ML method XGBoost and the model interpretation approach SHAP. The model we constructed had excellent performance in terms of high predictive accuracy, high tolerance to missing values (flexibility), and good clinical interpretability. Importantly, we identified the top 7 clinical factors affecting in-hospital mortality as age, LVEF, Killip class, heart rate, creatinine, glucose, and WBC. Among these, LVEF, glucose, and WBC were not included in the current traditional predictive models. Although creatinine is also included in the GRACE score, its relationship with mortality is not a simple linear one. The predictive value of glucose and WBC exceeds that of other variables in traditional predictive models, such as blood pressure, weight, and medical history (hypertension, diabetes, and angina). We believe that these findings can help doctors understand the value of ML models and uncover the pathophysiological significance of certain clinical variables in myocardial infarction.

While traditional statistical models such as TIMI and GRACE, as recommended by current guidelines [26], are useful and user-friendly, their overly simplified nature may result in inadequate predictive accuracy for risk classification and decision-making [8]. First, these models are developed based on a limited number of variables and may not encompass comprehensive information. Second, the LR method used by these models requires strong assumptions, including a linear relationship under the logit function, independence of observations, and no multicollinearity among variables [7,8,25,27]. This results in underestimating the complexity of the real association among variables and outcomes.

In contrast, ML methods can handle a larger number of variables, require no parametric assumptions, and can learn the complex relationships hidden in the data automatically [9]. The XGBoost method overcomes these limitations by generating a series of classification and regression trees (CARTs) with each one learning the residuals of its predecessors. The boosting mechanism gives the model a strong predictive power. As observed, the XGBoost model achieved an impressive AUC of 0.896 (95% CI 0.884-0.909) on the CAMI validation set, outperforming the other methods and proving to be a more powerful and effective tool for clinical risk prediction.

The XGBoost model's ability to tolerate missing values makes it well-suited for clinical applications, where incomplete variables are frequent [28-30]. While most ML methods achieve accuracy and precision by learning from a large number of variables, they often lose practicality because it is usually difficult to collect all the predictors used in the model in clinical practice. In such cases, missing values must be imputed if clinicians still want to apply the model. The proposed XGBoost model overcomes this weakness thanks to its ability to deal with missing values. We demonstrated that the XGBoost model's performance is relatively robust when faced with incomplete data compared to the traditional LR model. Even with only the top 10 important variables, the XGBoost model achieved an AUC of 0.877(95% CI 0.862-0.891) on the CAMI validation set. On the independent China PEACE set with only the top 10 important variables available, XGBoost gained an AUC of 0.840 (95% CI 0.829-0.852) compared to TIMI 0.762 (95% CI 0.748-0.776) and GRACE 0.789 (95% CI 0.776-0.803). These results demonstrated the XGBoost model's flexibility and generalization ability, which could alleviate concerns about the feasibility of applying complex ML models in clinical practice.

Another concern about the complex ML approaches applied in clinical practice is their lack of transparency. Unlike the widely employed LR method, whose coefficients clearly indicate the effect of predictive factors on the outcome, the black-box nature of complex ML algorithms applied in medical tasks has been seriously criticized and doubted in recent years [8,9]. To address this issue, our study used SHAP to interpret how the predicted risk was determined for individual patients and uncover the complex relationship between predictors and outcomes embedded in the XGBoost model.

Our results showed that HDL-C displayed a U-shaped relationship with in-hospital mortality among patients with STEMI. In the previous studies, Madsen et al [31] reported a U-shaped association between HDL-C and mortality, using data from 52,268 men and 64,240 women enrolled in 2 prospective population-based studies. Similarly, Bowe et al [32] found a U-shaped relationship between HDL-C and the risk of all-cause mortality in patients with kidney disease. For the variable potassium, our result showed that the patients with STEMI with potassium levels ranging from 4.13 to 4.49 mmol/L had the lowest in-hospital mortality risk, while levels greater than 4.5 mmol/L increased the mortality risk. Clinical practice guidelines recommend maintaining serum potassium levels between 4.0 and 5.0 mmol/L in patients with acute myocardial infarction (AMI) [33,34]. However, recent studies have challenged these guidelines, reporting that potassium levels greater than 4.5 mmol/L are associated with increased mortality [35-37]. Our study found that creatinine >1.1mg/dl (94.5/L) contributed to a higher in-hospital mortality risk. A previous study [38] reported that an elevated serum creatinine level (defined as creatinine ≥1.2 mg/dl) predicted a higher long-term mortality risk in patients with AMI.

For the variable blood glucose, our results showed that levels less than 8.15 mmol/L were safer for patients with STEMI. Another study reported that the best cutoff values for 30-day mortality among patients with STEMI were 149 mg/dL (8.27 mmol/L) for those without diabetes, 231 mg/dL (12.82 mmol/L) for those with diabetes, and 169 mg/dL (9.38 mmol/L) for all patients [39]. For the variable WBC, our result showed that a higher WBC count was associated with higher in-hospital mortality risk, with a safer threshold being less than 10.77/L. Cannon et al [40] reported that mortality at 30 days showed a curvilinear increase with increasing WBC count, with mortality

XSL•FO
RenderX

rising in patients with WBC count >10,000 /dL (P<.0001). Previous studies often investigated this relationship by categorizing or binning continuous variables and regressing the outcome on the categorical variables. However, this approach is heavily influenced by predefined cutoffs and cannot provide a continuous picture of the relationship. In contrast, our model offered more thorough and quantitative insights into the exact change in risk induced by specific patient characteristics. By interpreting how each variable contributed to in-hospital mortality, our study could help clinicians identify specific therapeutic targets and further guide patient management.

Our research has a certain guiding significance for clinical implementation. First, the new model is significantly superior to traditional GRACE and TIMI models, helping doctors predict patient prognosis. Second, ML has identified several variables not included in past models, which may serve as potential targets for clinical intervention or provide further understanding of the pathophysiology of disease development, such as WBC and blood glucose. Third, while clinicians often find it difficult to understand the variables selected by ML, adopting the XGBoost model and model interpretation approach SHAP further increases accuracy by capturing nonlinear relationships among the predictors and outcomes. This offers a clear explanation for why ML can improve predictive efficiency, thus enhancing clinicians' understanding of the performance improvement of ML. Methodologically, we used internal validation and a large sample size of independent external validation, all leading to consistent conclusions.

However, despite the superior performance of the proposed XGBoost model, several limitations still exist. First, the proposed XGBoost model was derived and validated on the Chinese STEMI patient cohort. Further validation is needed to confirm its efficiency on more general cohorts. Second, the study was designed prospectively, but this research is a retrospective analysis, so the variables recruited in our study may be limited. The model may be more powerful if more informative variables were added.

In conclusion, the proposed ML model in our paper demonstrated strong advantages in predictive ability, flexibility, and interpretability. Although some results need further study and verification, we have shown the benefits of complex models in the field of disease predictions. We offered a web calculator for convenient application, and we hope our study can help augment and extend the effectiveness of cardiologists to improve patient care and promote incorporating ML into daily practice.

Acknowledgments

This work was supported by the Twelfth Five-Year Planning Project of the Scientific and Technological Department of China (2011BAI11B02) and the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS; 2016-I2M-1-009 and 2020-I2M-C&T-B-050).

Authors' Contributions

YY conceived the study. JY contributed to the literature search and the development of the manuscript under the supervision of YY. YL and XL contributed to the data analysis. ST contributed to literature screening. YZ, TC, and GX contributed to data extraction and assessment. HX and XG contributed to the revision. All authors contributed to the critical review of the manuscript and approved the final draft. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. YY is the guarantor of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information and tables.
[DOCX File , 79 KB - jmir_v26i1e50067_app1.docx]

References

1. Gao R, Patel A, Gao W, Hu D, Huang D, Kong L, CPACS Investigators. Prospective observational study of acute coronary syndromes in China: practice patterns and outcomes. Heart 2008 May 01;94(5):554-560. [doi: 10.1136/hrt.2007.119750] [Medline: 17932092]
2. Li J, Li X, Wang Q, Hu S, Wang Y, Masoudi FA, China PEACE Collaborative Group. ST-segment elevation myocardial infarction in China from 2001 to 2011 (the China PEACE-Retrospective Acute Myocardial Infarction Study): a retrospective analysis of hospital data. Lancet 2015 Jan 31;385(9966):441-451 [FREE Full text] [doi: 10.1016/S0140-6736(14)60921-1] [Medline: 24969506]
3. Xu H, Li W, Yang J, Wiviott SD, Sabatine MS, Peterson ED, CAMI Registry study group. The China Acute Myocardial Infarction (CAMI) Registry: A national long-term registry-research-education integrated platform for exploring acute myocardial infarction in China. Am Heart J 2016 May;175:193-201.e3 [FREE Full text] [doi: 10.1016/j.ahj.2015.04.014] [Medline: 27179740]

XSL•FO
RenderX

4. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, Global Registry of Acute Coronary Events Investigators. Predictors of hospital mortality in the global registry of acute coronary events. Arch Intern Med 2003 Oct 27;163(19):2345-2353. [doi: 10.1001/archinte.163.19.2345] [Medline: 14581255]
5. Morrow DA, Antman EM, Charlesworth A, Cairns R, Murphy SA, de Lemos JA, et al. TIMI risk score for ST-elevation myocardial infarction: A convenient, bedside, clinical score for risk assessment at presentation: An intravenous nPA for treatment of infarcting myocardium early II trial substudy. Circulation 2000 Oct 24;102(17):2031-2037. [doi: 10.1161/01.cir.102.17.2031] [Medline: 11044416]
6. Long JS, Freese J. Regression Models for Categorical Dependent Variables using Stata, 2nd Edition. USA: Stata Press; Jun 01, 2006:273-278.
7. Berry W, Feldman S, Stanley FD. Multiple Regression in Practice. Los Angeles, CA: Sage Publications Inc; 1985:0803920547.
8. Ayer T, Chhatwal J, Alagoz O, Kahn CE, Woods RW, Burnside ES. Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics 2010 Jan;30(1):13-22 [FREE Full text] [doi: 10.1148/rg.301095057] [Medline: 19901087]
9. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996 Nov;49(11):1225-1231. [doi: 10.1016/s0895-4356(96)00002-9] [Medline: 8892489]
10. Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. Eur Heart J 2020 Jan 14;41(3):359-367 [FREE Full text] [doi: 10.1093/eurheartj/ehz565] [Medline: 31513271]
11. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 2017 Feb 14;38(7):500-507 [FREE Full text] [doi: 10.1093/eurheartj/ehw188] [Medline: 27252451]
12. Khosla A, Cao Y, Lin C, Chiu H, Hu J, Lee H, editors. An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010 Presented at: 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 25-28; Washington, DC. [doi: 10.1145/1835804.1835830]
13. Salih A, Boscolo Galazzo I, Gkontra P, Lee AM, Lekadir K, Raisi-Estabragh Z, et al. Explainable artificial intelligence and cardiac imaging: toward more interpretable models. Circ: Cardiovascular Imaging 2023 Apr;16(4). [doi: 10.1161/circimaging.122.014519] [Medline: 37042240]
14. Lee W, Lee J, Woo S, Choi SH, Bae J, Jung S, et al. Machine learning enhances the performance of short and long-term mortality prediction model in non-ST-segment elevation myocardial infarction. Sci Rep 2021 Jun 18;11(1):12886 [FREE Full text] [doi: 10.1038/s41598-021-92362-1] [Medline: 34145358]
15. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol 2021 Jun 01;6(6):633-641 [FREE Full text] [doi: 10.1001/jamacardio.2021.0122] [Medline: 33688915]
16. Pieszko K, Hiczkiewicz J, Budzianowski P, Rzeźniczak J, Budzianowski J, Błaszczyński J, et al. Machine-learned models using hematological inflammation markers in the prediction of short-term acute coronary syndrome outcomes. J Transl Med 2018 Dec 03;16(1):334 [FREE Full text] [doi: 10.1186/s12967-018-1702-5] [Medline: 30509300]
17. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17; San Francisco, CA p. 13-17. [doi: 10.1145/2939672.2939785]
18. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001 Oct 1;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
19. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760. [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]
20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. arXiv ;21(2) Preprint posted online on April 10 2019. [doi: 10.5260/chara.21.2.8]
21. Buuren SV, Groothuis-Oudshoorn K. Multivariate imputation by chained equations in R. J Stat Soft 2011;45(3). [doi: 10.18637/jss.v045.i03]
22. de Leeuw J. Correctness of Kruskal's algorithms for monotone regression with ties. Psychometrika 1977 Mar;42(1):141-144. [doi: 10.1007/bf02293750]
23. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. 2001 Presented at: 18th International Conference on Machine Learning; June 28-July 1; San Francisco, CA URL: https://dl.acm.org/doi/10.5555/645530.655658
24. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. 2002 Presented at: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 23-26; Edmonton, AB. [doi: 10.1145/775047.775151]

XSL•FO
RenderX

25. Hosmer JD, Lemeshow S, Sturdivant R. Applied Logistic Regression. Hoboken, NJ: John Wiley & Sons; 2013:0470582472.
26. O'Gara P, Kushner F, Ascheim D, Casey DJ, Chung M, de LJ. 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 2013:e362-e425. [doi: 10.1161/CIR.0b013e3182742cf6] [Medline: 23247304]
27. Pregibon D. Logistic regression diagnostics. Ann Statist 1981 Jul 1;9(4). [doi: 10.1214/aos/1176345513]
28. Molenberghs G, Kenward M. Missing Data in Clinical Studies. Bogner Regis, UK: John Wiley & Sons Ltd; 2007:0470510439.
29. DeSouza CM, Legedza ATR, Sankoh AJ. An overview of practical approaches for handling missing data in clinical trials. J Biopharm Stat 2009 Nov 21;19(6):1055-1073. [doi: 10.1080/10543400903242795] [Medline: 20183464]
30. Lin J, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform 2008 Feb;41(1):1-14 [FREE Full text] [doi: 10.1016/j.jbi.2007.06.001] [Medline: 17625974]
31. Madsen C, Varbo A, Nordestgaard B. Extreme high high-density lipoprotein cholesterol is paradoxically associated with high mortality in men and women: two prospective cohort studies. Eur Heart J 2017 Aug 21;38(32):2478-2486. [doi: 10.1093/eurheartj/ehx163] [Medline: 28419274]
32. Bowe B, Xie Y, Xian H, Balasubramanian S, Zayed MA, Al-Aly Z. High density lipoprotein cholesterol and the risk of all-cause mortality among US veterans. Clin J Am Soc Nephrol 2016 Aug 11;11(10):1784-1793. [doi: 10.2215/cjn.00730116]
33. Amsterdam E, Wenger N, Brindis R, Casey DJ, Ganiats T, Holmes D, et al. 2014 AHA/ACC guideline for the management of patients with non–st-elevation acute coronary syndromes: executive summary. Circulation 2014 Dec;130(25):2354-2394. [doi: 10.1161/cir.0000000000000133]
34. Cohn JN, Kowey PR, Whelton PK, Prisant LM. New guidelines for potassium replacement in clinical practice: a contemporary review by the National Council on Potassium in Clinical Practice. Arch Intern Med 2000 Sep 11;160(16):2429-2436. [doi: 10.1001/archinte.160.16.2429] [Medline: 10979053]
35. Choi JS, Kim YA, Kim HY, Oak CY, Kang YU, Kim CS, et al. Relation of serum potassium level to long-term outcomes in patients with acute myocardial infarction. Am J Cardiol 2014 Apr 15;113(8):1285-1290. [doi: 10.1016/j.amjcard.2014.01.402] [Medline: 24560065]
36. Goyal A, Spertus JA, Gosch K, Venkitachalam L, Jones PG, Van DBG, et al. Serum potassium levels and mortality in acute myocardial infarction. JAMA 2012 Jan 11;307(2):157-164. [doi: 10.1001/jama.2011.1967] [Medline: 22235086]
37. Krogager ML, Eggers-Kaas L, Aasbjerg K, Mortensen RN, Køber L, Gislason G, et al. Short-term mortality risk of serum potassium levels in acute heart failure following myocardial infarction. Eur Heart J Cardiovasc Pharmacother 2015 Oct 27;1(4):245-251 [FREE Full text] [doi: 10.1093/ehjcvp/pvv026] [Medline: 27418967]
38. Yamaguchi J, Kasanuki H, Ishii Y, Yagi M, Nagashima M, Fujii S, HIJC Investigators. Serum creatinine on admission predicts long-term mortality in acute myocardial infarction patients undergoing successful primary angioplasty: data from the Heart Institute of Japan Acute Myocardial Infarction (HIJAMI) Registry. Circ J 2007 Sep;71(9):1354-1359 [FREE Full text] [doi: 10.1253/circj.71.1354] [Medline: 17721010]
39. Planer D, Witzenbichler B, Guagliumi G, Peruga JZ, Brodie BR, Xu K, et al. Impact of hyperglycemia in patients with ST-segment elevation myocardial infarction undergoing percutaneous coronary intervention: the HORIZONS-AMI trial. Int J Cardiol 2013 Sep 10;167(6):2572-2579. [doi: 10.1016/j.ijcard.2012.06.054] [Medline: 22795245]
40. Cannon CP, McCabe CH, Wilcox RG, Bentley JH, Braunwald E. Association of white blood cell count with increased mortality in acute myocardial infarction and unstable angina pectoris. OPUS-TIMI 16 Investigators. Am J Cardiol 2001 Mar 01;87(5):636-9, A10. [doi: 10.1016/s0002-9149(00)01444-2] [Medline: 11230853]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
AUC: area under the curve
ARB: angiotensin II receptor blocker
AMI: acute myocardial infarction
CAMI: China Acute Myocardial Infarction
CART: classification and regression tree
GRACE: Global Registry in Acute Coronary Events
HDL-C: high-density lipoprotein cholesterol
LVEF: left ventricular ejection fraction
LR: logistic regression
MICE: Multiple Imputation by Chained Equations
ML: machine learning
NSTEMI: non–ST-segment elevation myocardial infarction
PEACE: Patient-Centered Evaluative Assessment of Cardiac Events
ROC: receiver operating characteristic
SBP: systolic blood pressure
SHAP: Shapley Additive Explanations

XSL•FO
RenderX

STEMI: ST-segment elevation myocardial infarction
TIMI: Thrombolysis In Myocardial Infarction
WBC: white blood cell
XGBoost: Extreme Gradient Boosting
XAI: explainable artificial intelligence

Edited by T de Azevedo Cardoso; submitted 19.06.23; peer-reviewed by H Sun, L Borges; comments to author 11.01.24; revised version received 25.03.24; accepted 18.06.24; published 30.07.24.

Please cite as:
Yang J, Li Y, Li X, Tao S, Zhang Y, Chen T, Xie G, Xu H, Gao X, Yang Y
A Machine Learning Model for Predicting In-Hospital Mortality in Chinese Patients With ST-Segment Elevation Myocardial Infarction: Findings From the China Myocardial Infarction Registry
J Med Internet Res 2024;26:e50067
URL: https://www.jmir.org/2024/1/e50067
doi:10.2196/50067
PMID:

©Jingang Yang, Yingxue Li, Xiang Li, Shuiying Tao, Yuan Zhang, Tiange Chen, Guotong Xie, Haiyan Xu, Xiaojin Gao, Yuejin Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A 3D and Explainable Artificial Intelligence Model for Evaluation of Chronic Otitis Media Based on Temporal Bone Computed Tomography: Model Development, Validation, and Clinical Application

Binjun Chen[1,2*], MD; Yike Li[3*], MD, PhD; Yu Sun[4*], MD, PhD; Haojie Sun[1,2], MD; Yanmei Wang[1,2], MD, PhD; Jihan Lyu[1,2], MD; Jiajie Guo[5], PhD; Shunxing Bao[6], PhD; Yushu Cheng[7], MD; Xun Niu[4], MD; Lian Yang[8], MD; Jianghong Xu[1,2], MD, PhD; Juanmei Yang[1,2], MD, PhD; Yibo Huang[1,2], MD, PhD; Fanglu Chi[1,2], MD, PhD; Bo Liang[8*], MD; Dongdong Ren[1,2*], MD, PhD

[1]ENT Institute and Department of Otorhinolaryngology, Eye & ENT Hospital, Fudan University, Shanghai, China

[2]NHC Key Laboratory of Hearing Medicine Research, Eye & ENT Hospital, Fudan University, Shanghai, China

[3]Department of Otolaryngology—Head and Neck Surgery, Vanderbilt University Medical Center, Nashville, TN, United States

[4]Department of Otorhinolargnology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[5]State Key Laboratory of Digital Manufacturing Equipment and Technology, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China

[6]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, United States

[7]Department of Radiology, Eye & ENT Hospital, Fudan University, Shanghai, China

[8]Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Yike Li, MD, PhD
Department of Otolaryngology—Head and Neck Surgery
Vanderbilt University Medical Center
1215 Medical Center Drive
Nashville, TN, 37232
United States
Phone: 1 6153438146
Email: yike.li.1@vumc.org

Abstract

Background: Temporal bone computed tomography (CT) helps diagnose chronic otitis media (COM). However, its interpretation requires training and expertise. Artificial intelligence (AI) can help clinicians evaluate COM through CT scans, but existing models lack transparency and may not fully leverage multidimensional diagnostic information.

Objective: We aimed to develop an explainable AI system based on 3D convolutional neural networks (CNNs) for automatic CT-based evaluation of COM.

Methods: Temporal bone CT scans were retrospectively obtained from patients operated for COM between December 2015 and July 2021 at 2 independent institutes. A region of interest encompassing the middle ear was automatically segmented, and 3D CNNs were subsequently trained to identify pathological ears and cholesteatoma. An ablation study was performed to refine model architecture. Benchmark tests were conducted against a baseline 2D model and 7 clinical experts. Model performance was measured through cross-validation and external validation. Heat maps, generated using Gradient-Weighted Class Activation Mapping, were used to highlight critical decision-making regions. Finally, the AI system was assessed with a prospective cohort to aid clinicians in preoperative COM assessment.

Results: Internal and external data sets contained 1661 and 108 patients (3153 and 211 eligible ears), respectively. The 3D model exhibited decent performance with mean areas under the receiver operating characteristic curves of 0.96 (SD 0.01) and 0.93 (SD 0.01), and mean accuracies of 0.878 (SD 0.017) and 0.843 (SD 0.015), respectively, for detecting pathological ears on the 2 data sets. Similar outcomes were observed for cholesteatoma identification (mean area under the receiver operating

XSL•FO
RenderX

characteristic curve 0.85, SD 0.03 and 0.83, SD 0.05; mean accuracies 0.783, SD 0.04 and 0.813, SD 0.033, respectively). The proposed 3D model achieved a commendable balance between performance and network size relative to alternative models. It significantly outperformed the 2D approach in detecting COM ($P \leq .05$) and exhibited a substantial gain in identifying cholesteatoma ($P < .001$). The model also demonstrated superior diagnostic capabilities over resident fellows and the attending otologist ($P < .05$), rivaling all senior clinicians in both tasks. The generated heat maps properly highlighted the middle ear and mastoid regions, aligning with human knowledge in interpreting temporal bone CT. The resulting AI system achieved an accuracy of 81.8% in generating preoperative diagnoses for 121 patients and contributed to clinical decision-making in 90.1% cases.

Conclusions: We present a 3D CNN model trained to detect pathological changes and identify cholesteatoma via temporal bone CT scans. In both tasks, this model significantly outperforms the baseline 2D approach, achieving levels comparable with or surpassing those of human experts. The model also exhibits decent generalizability and enhanced comprehensibility. This AI system facilitates automatic COM assessment and shows promising viability in real-world clinical settings. These findings underscore AI's potential as a valuable aid for clinicians in COM evaluation.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000036300; https://www.chictr.org.cn/showprojEN.html?proj=58685

(J Med Internet Res 2024;26:e51706) doi:10.2196/51706

KEYWORDS

artificial intelligence; cholesteatoma; deep learning; otitis media; tomography, x-ray computed; machine learning; mastoidectomy; convolutional neural networks; temporal bone

Introduction

Chronic otitis media (COM) represents a recurrent inflammatory condition inside the tympanic cavity [1]. COM encompasses various forms, including chronic suppurative otitis media (CSOM) and cholesteatoma, each with unique histological characteristics. CSOM involves the accumulation and discharge of purulent fluid, affecting an estimated 330 million people worldwide, with approximately half experiencing hearing loss [2]. Cholesteatoma is characterized by the buildup of keratinized squamous epithelium, which has the potential to erode auditory structures and exhibits a notable tendency for relapse. Accurate identification and differentiation of COM types are crucial for effective disease management and surgical planning [3]. Mastoidectomy, which involves the removal of part of the temporal bone, is the conventional surgical approach for COM. However, less invasive techniques such as endoscopic tympanoplasty are gaining favor for treating CSOM and other noncholesteatoma conditions due to their potential for reduced structural damage and faster recovery [4-9].

Temporal bone computed tomography (CT) is vital for assessing COM and aiding in surgical planning, especially when initial otoscopic examinations have restricted views and yield inconclusive findings [10]. Offering a cost-effective alternative to magnetic resonance imaging (MRI), CT is instrumental in distinguishing cholesteatoma from CSOM by detecting osseous erosion in the tympanum. Although studies have shown that clinicians are capable of diagnosing COM based on CT alone [11-17], distinguishing between COM subtypes poses greater challenges to the human eye. Moreover, interpreting temporal bone CT scans requires specialized training and experience, which may not be universally available across otolaryngologists.

Artificial intelligence (AI) is making remarkable advancements in health care. Deep learning (DL) models, particularly convolutional neural networks (CNNs), have demonstrated enhanced efficiency and reduced errors in disease diagnoses and prediction of clinical outcomes [18-21]. While a few recent papers have reported CNN models in evaluating COM with accuracy scores ranging from 0.77 to 0.85, these studies primarily relied on otoscopic or single-layer CT scans [22,23]. These 2D representations may not be optimal for revealing pathological changes in concealed or peripheral anatomical structures, such as the attic space and the mastoid air cells. In addition, the inherent "black box" nature of DL models, where decision-making strategies are challenging to understand, has been a common criticism [24,25]. This lack of comprehensibility hinders the widespread adoption of AI models in clinical practice.

In light of these challenges, this study aimed to create an explainable, 3D CNN model for the automatic interpretation of temporal bone CT scans. The model was designed to pinpoint the region of interest (ROI) and identify pathological and cholesteatomatous conditions in a 3D fashion. Comprehensive benchmarks against baseline methods and human experts on distinct data sets were conducted to demonstrate the robustness and generalizability of this model. In addition, heat map generation was used to highlight potential pathological changes in CT scans and elucidate the model's rationale for making predictions. These features were integrated into an AI system for the automatic, end-to-end evaluation of COM, which was subsequently assessed in clinical settings. The overarching goal of this system is to support clinicians in making informed decisions for common otologic conditions, thereby enhancing efficiency, reliability, and transparency.

Methods

Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was granted by the institutional review boards at Vanderbilt University Medical Center (191804) and the Eye, Ear, Nose and Throat (EENT) Hospital of Fudan University (2019076). Informed consent was waived as all data were de-identified. The observational study, which aimed to assess the model's viability in aiding

XSL•FO
RenderX

preoperative assessment, was registered with the Chinese Clinical Trial Register (ChiCTR: 2000036300). No compensation was provided to any study participants.

Participants

Data were retrospectively obtained from patients admitted for middle ear surgeries from December 2015 to July 2021 at EENT Hospital. Patients diagnosed with acute otitis media, any inner or external ear diseases, or those with missing temporal bone CT scan were excluded, resulting in 1661 patients eligible for model development. An extra data set containing 108 patients with COM was collected from Wuhan Union (WU) Hospital for external validation (Figure 1).

Figure 1. Flowchart of data retrieval. CT: computed tomography; EAC: external auditory canal; EENT: Eye, Ear, Nose, and Throat Hospital of Fudan University; TM: tympanic membrane; WU: Wuhan Union Hospital.

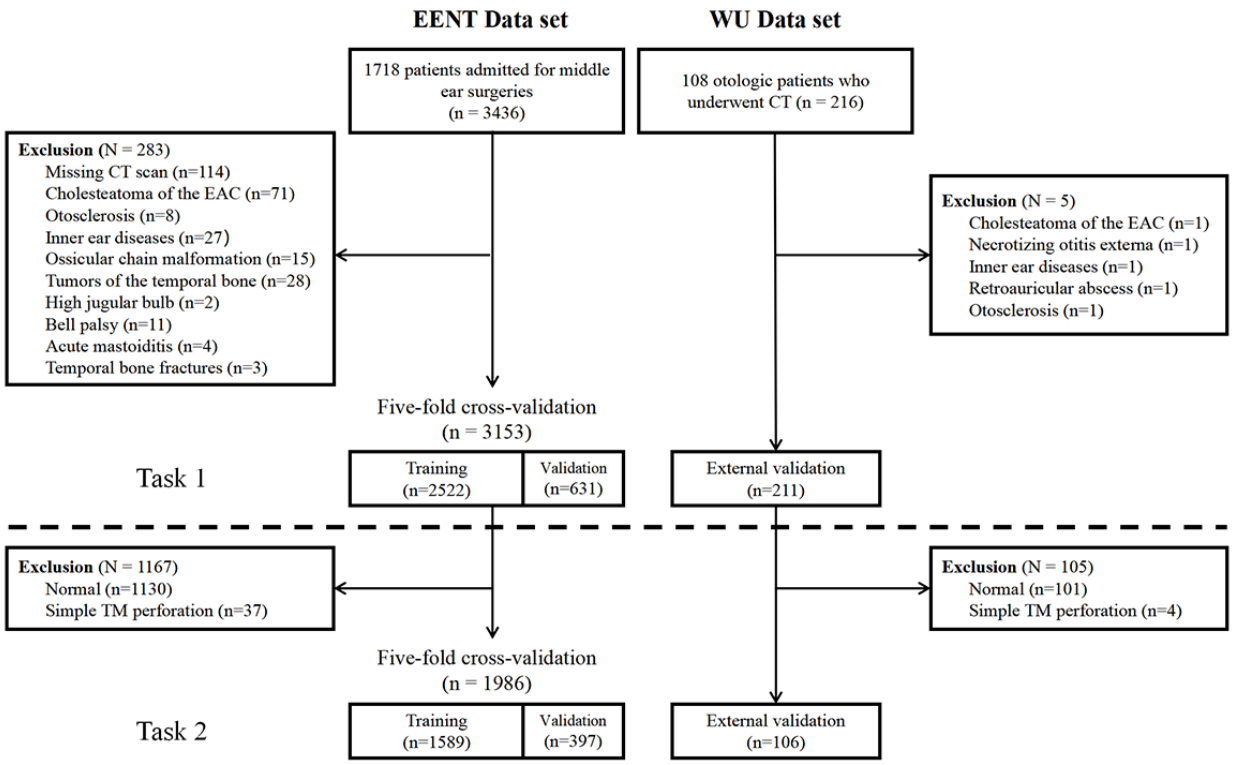

Temporal Bone CT Scans

As part of the routine preoperative assessment, each patient underwent at least 1 temporal bone CT, conducted from the lower margin of the external auditory meatus to the top margin of the petrous bone using a SOMATOM Sensation 10 CT scanner (Siemens Inc) at the EENT Hospital. The scanning parameters were as follows: matrix (512 × 512), field of view (220 mm × 220 mm), tube voltage (140 kV), tube current (100 mAs), section thickness (0.6-0.75 mm), window width (4000 HU), and window level (700 HU). CT scans from the WU Hospital were obtained using a SOMATOM Plus 4 model (Siemens Inc) with different settings for field of view (100 mm), voltage (120 kV), and thickness (0.75 mm). All images were saved in the DICOM format.

Label Assignment

All eligible ears were treated as independent cases and assigned ground truth labels based on their diagnoses (Table 1). Each label was verified according to intraoperative findings and pathology reports for operated ears and using a combination of history, ear examination, audiogram results, and imaging findings for unoperated ears. In cases of unoperated ears, a “normal” label was assigned when there was an absence of ear symptoms, hearing loss, or signs of inflammation. A diagnosis of CSOM was assigned when chronic purulent discharge, conductive hearing loss, and the presence of a perforated tympanic membrane or soft tissue shadow in the tympanic cavity were observed. Cholesteatoma was considered if keratin debris was identified, or if there were signs of osseous damage along with retraction or perforation of the pars flaccida [22]. Two otolaryngology residents with full access to patients’ medical records independently reviewed these labels as unblinded annotators. Any discrepancies were addressed with senior specialists until a consensus was reached. All data were deidentified and stored on password-protected computers.

XSL•FO
RenderX

Table 1. Summary of patient characteristics and label assignment.

Characteristics	EENT[a] data set (N=1661; number of ears=3153)	WU[b] data set (N=108; N=211)
Patient age (years), mean (SD)	41.1 (16.6)	39.8 (14.0)
Patient sex, n (%)		
Male	832 (50.1)	49 (45.4)
Female	829 (49.9)	59 (54.6)
Diagnosis per ear, n (%)		
Normal	1130 (35.8)	101 (47.9)
Cholesteatoma	728 (23.1)	30 (14.2)
CSOM[c]	1011 (32.1)	69 (32.7)
Tympanosclerosis	142 (4.5)	2 (0.1)
Cholesterol granuloma	72 (2.3)	1 (0.05)
OME[d]	41 (1.3)	7 (3.3)
Adhesive otitis media	29 (0.1)	1 (0.05)
Task 1 labels, n (%)		
Normal	1130 (35.8)	101 (47.9)
Pathological	2023 (64.2)	110 (52.1)
Task 2 labels, n (%)		
Cholesteatoma	728 (36.7)	28 (26.4)
Noncholesteatoma	1258 (63.3)	78 (73.6)

[a]EENT: Eye, Ear, Nose, and Throat Hospital of Fudan University.

[b]WU: Wuhan Union Hospital.

[c]CSOM: chronic suppurative otitis media.

[d]OME: otitis media with effusion.

Model Architecture

The framework consists of 2 functionally distinct units: a region proposal network for 3D segmentation of ROI, and a classification network for generating predictions. Both networks are established based on CNN models.

Region Proposal Network

This network is designed to extract the middle ear on each side from a full set of temporal bone CT scan (Figure 2A). It contains a YOLO (You Only Look Once; v5) model that is trained to detect and locate 2 auditory structures, including the internal auditory canal and the horizontal semicircular canal, in a series of 2D axial CT scans [26]. These landmarks, positioned at or around the central level of the middle ear, possess unique graphical appearances recognizable by the object detection model. In our recent study, this model demonstrated a 100% success rate in identifying the middle ear region from temporal bone CT scans [22]. Subsequently, a 3D data matrix (150 × 150 × 32) of the ROI is extracted based on the center coordinates of these 2 structures on each side.

XSL•FO
RenderX

Figure 2. An overview of the AI framework. (A) The region proposal network used to locate landmark structures and segment the 3D ROI from the original CT scans. (B) The classification network based on a 3D convolutional neural network architecture and trained to perform 2 classification tasks. (C) The gradient heatmaps generated to highlight the critical regions for decision-making. Conv: convolution; CT: computed tomographic; FC: fully connected; MP: max pooling; ReLu: rectified linear unit; ROI: region of interest.

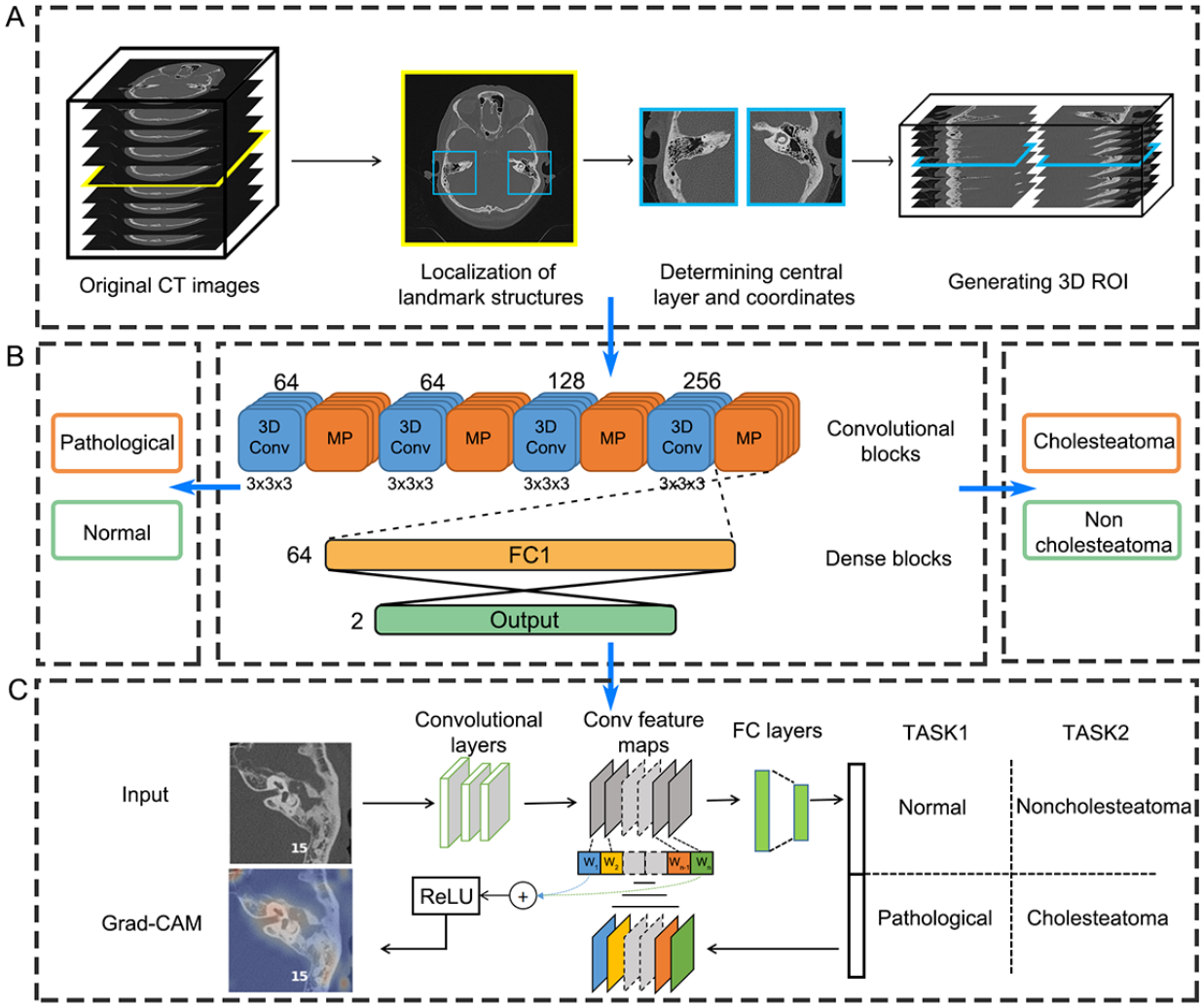

Classification Network

A 3D CNN model is built to interpret the extracted ROI and classify different types of conditions (Figure 2B). This model features 4 convolution blocks and 2 dense blocks (Table 2). Each convolution block consists of a 3D convolutional layer to summarize graphical features along all axes of the input image, followed by a max-pooling layer for downsampling these features and another layer for batch normalization. These high-level features are then pooled and passed to the fully connected layers of the dense blocks, where the diagnosis is predicted based on the calculated probability of each class by a softmax function. A dropout layer is applied to prevent overfitting [27].

XSL•FO
RenderX

Table 2. Architecture of the 3D convolutional neural network model.

Block and kernel inputs	Settings
Convolution 1	
Conv3D[a]	(3,3,3,64)
MaxPooling3D[b]	(2,2,2)
BatchNormalization[c]	
Convolution 2	
Conv3D	(3,3,3,64)
MaxPooling3D	(2,2,2)
BatchNormalization	
Convolution 3	
Conv3D	(3,3,3,128)
MaxPooling3D	(2,2,2)
BatchNormalization	
Convolution 4	
Conv3D	(3,3,3,256)
MaxPooling3D	(2,2,2)
BatchNormalization	
GlobalAveragePooling3D[d]	
Dense 1	
Fully connected	64
Dropout	0.3
Output	
Fully connected	2

[a]Conv3D: 3D convolutional layer.

[b]MaxPooling3D: 3D max pooling layer.

[c]BatchNormalization: batch normalization layer.

[d]GlobalAveragePooling3D: layer performing global average pooling for 3D data.

Model Training and Testing

Task 1—Detection of COM

The first classification model was trained in a binary task distinguishing between normal and pathological ears in all cases (n=3153). The training and testing procedures involved 5-fold cross-validation on the internal (EENT) data set. Specifically, the data set was evenly partitioned into 5 nonoverlapping subsets in a random, stratified fashion. In each iteration, 1 subset was reserved for testing (n=631), while the remaining 4 were used for training (n=2522). Model performance metrics were averaged over 5 iterations of this process. During each training session, a random 20% of training images (n=504) were allocated for validation. Training was set for 1000 epochs with an initial learning rate of 0.0001, and the Adam optimizer was used to dynamically adjust the algorithm's learning capability and minimize errors [28]. Early termination was implemented if no further decrease in validation loss was observed for a consecutive 10 epochs. These hyperparameters were determined based on the resultant model performance and training efficiency shown in a preliminary study. The trained model was also evaluated on the external data set (n=211) in each round.

Task 2—Identification of Cholesteatoma

The second classification model was trained to specifically identify cholesteatoma on selected CT scans that displayed signs of inflammation in the middle ears. This task was designed to simulate a common clinical scenario where clinicians need to differentiate cholesteatoma from other types of COM in patients with positive imaging findings. The aim was to provide a preoperative assessment of the risk of cholesteatoma, assisting clinicians in surgical planning [3,29]. For this task, a subset of CT scans with visible soft tissue density or increased opacification in the middle ear or mastoid was selected from both the internal (n=1986) and external sets (n=106). The remaining methods, including extraction of ROI, network architecture, and the training and testing procedures, were consistent with those used in the first task.

XSL•FO
RenderX

Ablation Study

To refine model selection and gain a better understanding of the network's behavior, an ablation study was performed to compare the proposed classification network with 3 alternative models, each incorporating modifications to certain features. Specifically, the number of convolutional blocks was decreased and increased by 1 in alternative model 1 and model 2, respectively, and a different size of filter was applied in model 3 (Tables S1-S3 in Multimedia Appendix 1). To ensure adequate statistical power for detecting differences across models, experiments were conducted on the main data set using the same methodology as outlined in the preceding sections.

Benchmarking Against the 2D Approach

To investigate whether the use of 3D CT scans may enhance diagnostic performance, a benchmark study was designed to compare the proposed system with a baseline model using 2D images. This baseline model, previously established by our team, uses transfer learning on a pretrained Inception-V3 (Google LLC) model [22]. In this study, the base model of Inception-V3 was retained, and the final classification layer was customized with a binary output. Training and validation were conducted in the same manner as the 3D model, except that only a single CT scan at the central layer of the ROI was used as the input for the 2D model. All image-preprocessing techniques and hyperparameter settings remained consistent with those outlined in the previous study [22].

Benchmarking Against Human Experts

Another benchmark test was performed against human experts to provide an additional unbiased evaluation of the proposed system. Seven human specialists with a broad range of qualifications were recruited to perform both tasks based on the same image data. The participants included 2 senior otologists, each with 12 years of clinical experience, 1 senior head and neck radiologist with 21 years of experience, 1 attending otologist with 7 years of experience, and 3 otolaryngology residents with 3, 3, and 2 years of experience, respectively. Each expert was provided only with the CT scans and instructed to make a task-specific diagnosis to each ear (task 1: normal or pathological; task 2: cholesteatoma or noncholesteatoma). The test data for clinicians comprised a random selection of 244 ears from the EENT set and all eligible ears from the WU set. To assess intrarater reliability, a random replication of 10% of test cases (n=48) was mixed with these data. All test cases (N=502) had not been previously seen by any experts. They were anonymized, shuffled, and stored on a password-protected computer along with spreadsheets to record each expert's diagnoses for these cases.

Generation of Heat Maps

Gradient-Weighted Class Activation Mapping was used to visualize model's rationale for decision-making (Figure 2C). In essence, this approach leverages the gradients of the target class flowing into the final convolutional layer to produce a coarse localization heat map, highlighting the critical regions in the image [30]. In this study, heat maps were generated in a 3D fashion and rescaled to match the original images using TensorFlow 2.11 in Python 3.91 (Python Core Team) [31].

Clinical Applications

The validated model was integrated into a Python program, enabling the automated assessment of COM from raw CT inputs to the generation of explainable diagnoses in an end-to-end fashion (see the section "Data Availability Statements" and Multimedia Appendix 2). To evaluate its viability in assisting otologists in clinical settings, this system was used with a prospective cohort of patients undergoing middle ear surgeries at EENT hospital from November 2023 to January 2024 in a single-arm observational study. Preoperative model predictions, along with routine assessments, were provided to 2 senior otologists, who were given autonomy to determine surgical strategies based on their discretion. Surgeons were surveyed regarding the use of model-generated information in their decision-making processes for these cases. Model predictions were used to analyze the selection of surgical approaches and to measure model performance against pathological findings. Hearing gain was assessed by comparing the air conduction threshold at 2 weeks postoperatively with the baseline.

Statistical Analysis

Descriptive statistics were applied as appropriate. The overall predictability of a model was evaluated by the area under the receiver operating characteristic (AUROC) curve. The optimal cutoff threshold on the curve was determined at the point with minimal distance to the upper left corner on the validation set and subsequently applied to the test set. The numbers of correctly and incorrectly classified cases were displayed in a confusion matrix, and these were used to calculate the performance metrics, including accuracy, recall, specificity, precision, and F_1-score. These metrics offer comprehensive insights into the model's performance, covering overall correctness in identifying both positives and negatives (accuracy), sensitivity in detecting positive cases (recall), capability in ruling in patients (specificity), propensity for preventing false alarms (precision), and effectiveness in identifying positive cases while minimizing false positives and false negatives (F_1-score). They were derived as shown in Textbox 1. Results are averaged over 5 iterations of cross-validation or external validation and presented as mean (SD). Intrarater consistency was evaluated using Cohen kappa. Significance was determined through pairwise 2-tailed *t* test for difference in performance between models and via 1-way analysis of variance between the proposed model and human experts. The alpha level was set at .05. Statistical analyses were conducted using Python 3.91 [31].

XSL•FO
RenderX

Textbox 1. The calculation of performance metrics.

Accuracy = (True positive + True negative)/Total sample size

Recall = True positive/(True positive + False negative)

Specificity = True negative/(True negative + False positive)

Precision = True negative/(True negative + False negative)

F_1-score=2 × True positive/(2 × True positive + False positive + False negative)

Results

ROI Extraction

The region proposal network successfully extracted the 3D ROI containing the critical anatomies on each side, including the tympanic cavity and sinus tympani (Figure 3). This has been confirmed by manual inspection of the generated images in all cases from both data sets.

Figure 3. Generation of the 3D ROI. The region proposal network identifies landmark structures in each of the full-sized sequential CT slices and determines the center of the middle ear on each side. A 3D image comprising 32 stacks of axial slices in 150 × 150 pixels is subsequently segmented. This ROI encompasses an extensive range of critical anatomies within the temporal bone for the evaluation of COM. CT: computed tomographic; HSC: horizontal semicircular canal; IAC: internal auditory canal; ROI: region of interest.

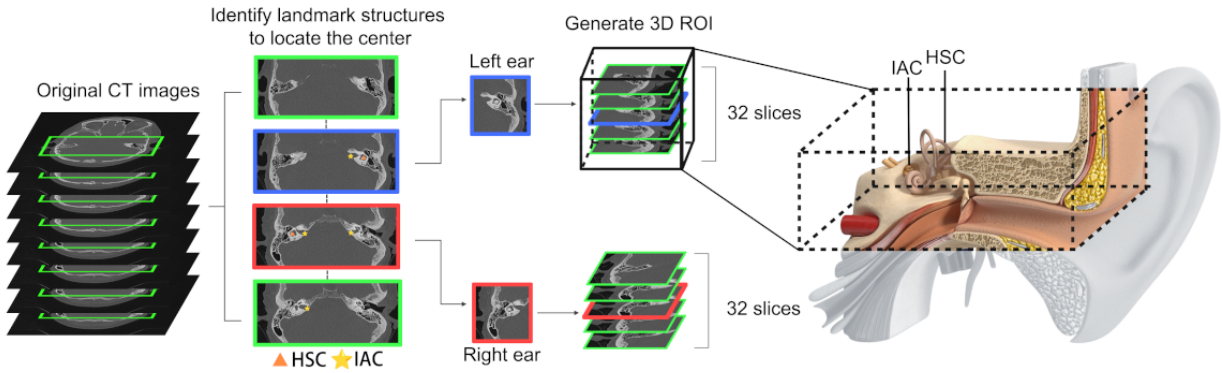

Task 1

Our model exhibited decent performance in identifying pathological changes in the middle ear, achieving a mean accuracy of 87.8%, recall of 85.3%, specificity of 91.3%, and precision of 93.3% on the internal data set (Table 3). It also demonstrated a near-perfect AUROC score of 0.96. These performance metrics remained generally consistent on the external data set, with a comparable AUROC score of 0.93, indicating reasonable generalizability (Figure 4).

XSL•FO
RenderX

Table 3. Performance of the baseline 2D and the proposed 3D models.

Task and model	Size (MB)	Data set	Accuracy, mean (SD)	Recall, mean (SD)	Specificity, mean (SD)	Precision, mean (SD)	F_1-score, mean (SD)	AUROC[a], mean (SD)	*P* value
1									
3D	14.2	EENT[b]	0.878 (0.017)	0.853 (0.032)	0.913 (0.067)	0.933 (0.045)	0.89 (0.012)	0.00959 (0.00011)	.003
2D	274	EENT	0.861 (0.019)	0.845 (0.028)	0.883 (0.052)	0.909 (0.036)	0.875 (0.016)	0.00939 (0.00013)	N/A[c]
3D	14.2	WU[d]	0.843 (0.015)	0.756 (0.047)	0.934 (0.021)	0.924 (0.018)	0.83 (0.022)	0.00933 (0.0001)	.05
2D	274	WU	0.821 (0.023)	0.744 (0.078)	0.901 (0.046)	0.891 (0.039)	0.808 (0.036)	0.00918 (0.00012)	N/A
2									
3D	14.2	EENT	0.783 (0.04)	0.808 (0.025)	0.77 (0.054)	0.652 (0.06)	0.721 (0.042)	0.00853 (0.0003)	<.001
2D	274	EENT	0.67 (0.037)	0.716 (0.144)	0.646 (0.119)	0.523 (0.044)	0.596 (0.036)	0.00744 (0.00025)	N/A
3D	14.2	WU	0.812 (0.033)	0.614 (0.085)	0.878 (0.031)	0.626 (0.078)	0.618 (0.069)	0.00826 (0.00055)	<.001
2D	274	WU	0.676 (0.103)	0.479 (0.224)	0.741 (0.185)	0.41 (0.086)	0.411 (0.096)	0.00714 (0.00049)	N/A

[a]AUROC: area under the receiver operating characteristic curve.

[b]EENT: Eye, Ear, Nose, and Throat Hospital of Fudan University.

[c]N/A: not applicable.

[d]WU: Wuhan Union Hospital.

XSL•FO
RenderX

Figure 4. Receiver operating characteristic plots for the benchmark tests. The curve and the shaded area indicate the mean (1 SD) of a model, respectively. Clinical experts are marked by colored asterisks for individual performance and by an open circle for averaged performance. The dotted diagonal line represents a random classifier. AUC: area under the curve; EENT: Eye, Ear, Nose, and Throat Hospital of Fudan University; WU: Wuhan Union Hospital.

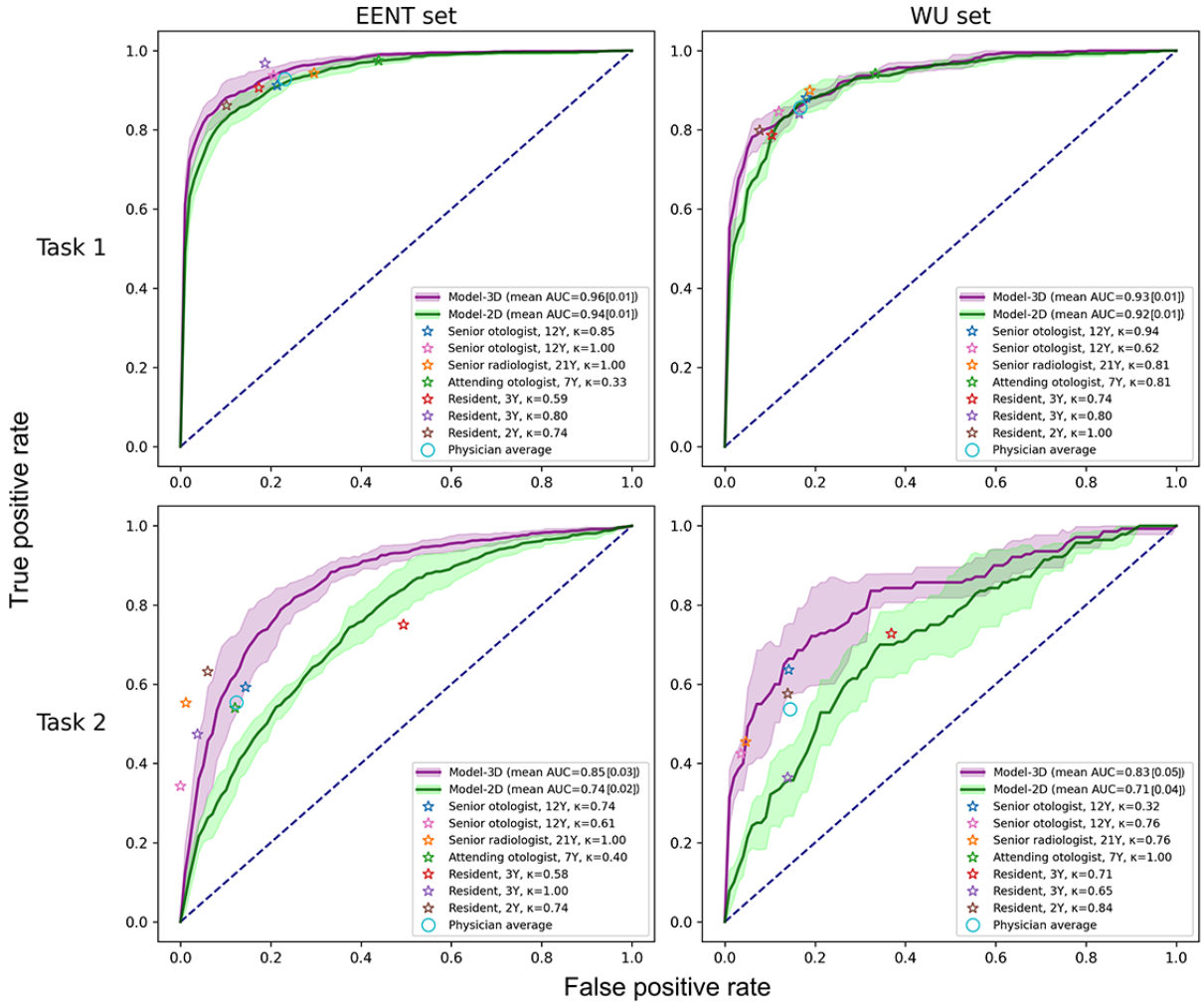

Task 2

This model also demonstrated satisfactory predictive capabilities in differentiating between cholesteatoma and noncholesteatomatous cases. On both data sets, the model managed to correctly identify whether a case involved cholesteatoma in approximately 4 out of 5 instances (with accuracies of 78.3% and 81.3%). Generalizability was further supported by the comparable AUROC scores of 0.85 and 0.83 on the internal and the external data set, respectively (Table 3).

Ablation Study

This model exhibited a reasonable balance between predictability and efficiency (Table 4). Compared with models 1 and 3, it achieved significantly better performance in both tasks ($P<.01$). In addition, despite having approximately 60% fewer parameters, the proposed model demonstrated equivalent performance to model 2 in both tasks ($P=.26$ and .91, respectively), indicating its enhanced computational efficiency.

XSL•FO
RenderX

Table 4. Ablation study on the 3D classification network.

Task and model	Size (MB)	Accuracy, mean (SD)	Recall, mean (SD)	Specificity, mean (SD)	Precision, mean (SD)	F_1-score, mean (SD)	AUROC[a], mean (SD)	*P* value
1								
Proposed	14.2	0.878 (0.017)	0.853 (0.032)	0.913 (0.067)	0.933 (0.045)	0.89 (0.012)	0.00959 (0.00011)	N/A[b]
Model 1	4.0	0.858 (0.03)	0.827 (0.046)	0.901 (0.058)	0.921 (0.043)	0.87 (0.028)	0.00947 (0.00019)	<.001
Model 2	34.5	0.884 (0.014)	0.862 (0.021)	0.914 (0.041)	0.933 (0.03)	0.895 (0.012)	0.00961 (0.00009)	.26
Model 3	64.8	0.864 (0.022)	0.851 (0.062)	0.887 (0.074)	0.914 (0.053)	0.878 (0.019)	0.0095 (0.00019)	.003
2								
Proposed	14.2	0.783 (4.0)	0.808 (0.025)	0.77 (0.054)	0.652 (0.06)	0.721 (0.042)	0.00853 (0.0003)	N/A
Model 1	4.0	0.758 (0.048)	0.712 (0.118)	0.783 (0.065)	0.636 (0.064)	0.668 (0.075)	0.00817 (0.0006)	.006
Model 2	34.5	0.782 (0.036)	0.795 (0.071)	0.775 (0.074)	0.659 (0.071)	0.716 (0.032)	0.00862 (0.00031)	.91
Model 3	64.8	0.756 (0.056)	0.76 (0.059)	0.754 (0.109)	0.634 (0.088)	0.685 (0.037)	0.00826 (0.000047)	.003

[a]AUROC: area under the receiver operating characteristic curve.

[b]N/A: not applicable.

Benchmarks

Compared with the 2D approach, the 3D network demonstrated significantly superior performance in both tasks across data sets ($P \leq .05$). In particular, the proposed model exhibited a substantial performance gain in differentiating between cholesteatoma and noncholesteatomata, with an increase of more than 10% in all outcome metrics on both data sets (Table 3).

This model also matched or even surpassed the diagnostic capabilities of human experts in both tasks (Figure 4). It exhibited marginally superior performance compared with human eyes in the first task ($P=.05$) and significantly outperformed them in the visually challenging task 2 ($P<.001$). Post hoc pairwise comparisons revealed that the model excelled over the attending otologist in task 1 and 2 resident fellows in task 2, rivaling all senior clinicians (Table 5). Similar results were shown across the breakdown of data sources, with a notable finding that the model outperformed a senior otologist in task 2 on the EENT subset (Table S4 in Multimedia Appendix 1). Moreover, the proposed model demonstrated perfect consistency, surpassing all human experts who exhibited higher SDs in all outcome metrics and lower scores of intrarater reliability.

XSL•FO
RenderX

Table 5. Benchmark performance against human experts.

Task and rater	Accuracy	Recall	Specificity	Precision	F_1-score	Kappa values	*P* value
1							
The 3D model, mean (SD)	0.878 (0.017)	0.853 (0.032)	0.913 (0.067)	0.933 (0.045)	0.89 (0.012)	0.01 (0.00)	N/A[a]
Expert average, mean (SD)	0.857 (0.022)	0.898 (0.042)	0.804 (0.094)	0.86 (0.05)	0.876 (0.013)	0.0082 (0.0009)	.05
Senior otologist A: 12 Y[b]	87.3%	89.8%	84.1%	87.9%	88.8%	0.75	.79
Senior otologist B: 12 Y	85.7%	89.9%	80.4%	85.6%	87.7%	0.92	.49
Senior radiologist: 21 Y	85.4%	92.4%	76.3%	83.4%	87.7%	0.87	.37
Attending otologist: 7 Y	81.1%	96.0%	61.9%	76.5%	85.2%	0.73	.002
Resident A: 3 Y	85.9%	85.5%	86.4%	89.1%	87.2%	0.71	.56
Resident B: 3 Y	87.4%	91.2%	82.5%	87.1%	89.1%	0.81	.74
Resident C: 2 Y	86.8%	83.5%	91.2%	92.4%	87.7%	0.92	.96
2							
The 3D model, mean (SD)	0.843 (0.015)	0.756 (0.047)	0.934 (0.021)	0.924 (0.018)	0.83 (0.022)	0.01 (0.00)	N/A
Expert average, mean (SD)	0.741 (0.052)	0.549 (0.123)	0.865 (0.139)	0.772 (0.135)	0.622 (0.061)	0.0072 (0.0012)	<.001
Senior otologist A: 12 Y	73.8%	36.7%	98.2%	93.0%	52.6%	0.70	.07
Senior otologist B: 12 Y	75.8%	60.6%	85.7%	73.3%	66.3%	0.47	.25
Senior radiologist: 21 Y	79.5%	52.3%	97.0%	91.9%	66.7%	0.86	.82
Attending otologist: 7 Y	74.5%	55.0%	87.0%	73.2%	62.8%	0.74	.11
Resident A: 3 Y	63.8%	74.3%	56.9%	52.9%	61.8%	0.67	<.001
Resident B: 3 Y	72.3%	44.0%	90.9%	76.2%	55.8%	0.77	.02
Resident C: 2 Y	78.8%	61.5%	89.9%	79.8%	69.4%	0.80	.96

[a]N/A: not applicable.

[b]Y: years of experience in clinical practice.

Visual Assessment of Heat Maps

Heat Maps from both models consistently highlighted the tympanic cavity and mastoid that manifested pathological findings characteristic of the target condition (Figure 5). Specifically, the first model generated a hot signal indicative of soft tissue density in an affected middle ear (Figure 5A), while the signal remained subdued in a normal ear (Figure 5B). Similarly, the second model revealed a distinct hot spot in a cholesteatomatous ear exhibiting the classic patterns of tympanum widening and ossicular destruction [17,32,33] (Figure 5C). In contrast, a case of CSOM showing intact ossicles surrounded by soft tissue shadows in a normal-sized tympanic cavity did not exhibit a corresponding hot spot (Figure 5D). These observations reflect that the AI's decision-making strategy aligns reasonably well with established human knowledge for both tasks.

XSL•FO
RenderX

Figure 5. Examples of heat maps. The heat maps, generated in 3D fashion, are superimposed on the original computed tomographic scans and flattened to a series of 2D images for demonstration purpose. (A-B) A pathological and a normal ear, respectively. (C-D) A cholesteatoma and a noncholesteatoma case, respectively. Area marked by hot signals indicate the presence of graphic patterns contributing to a "positive" prediction (ie, a pathological ear in task 1 and a cholesteatoma in task 2).

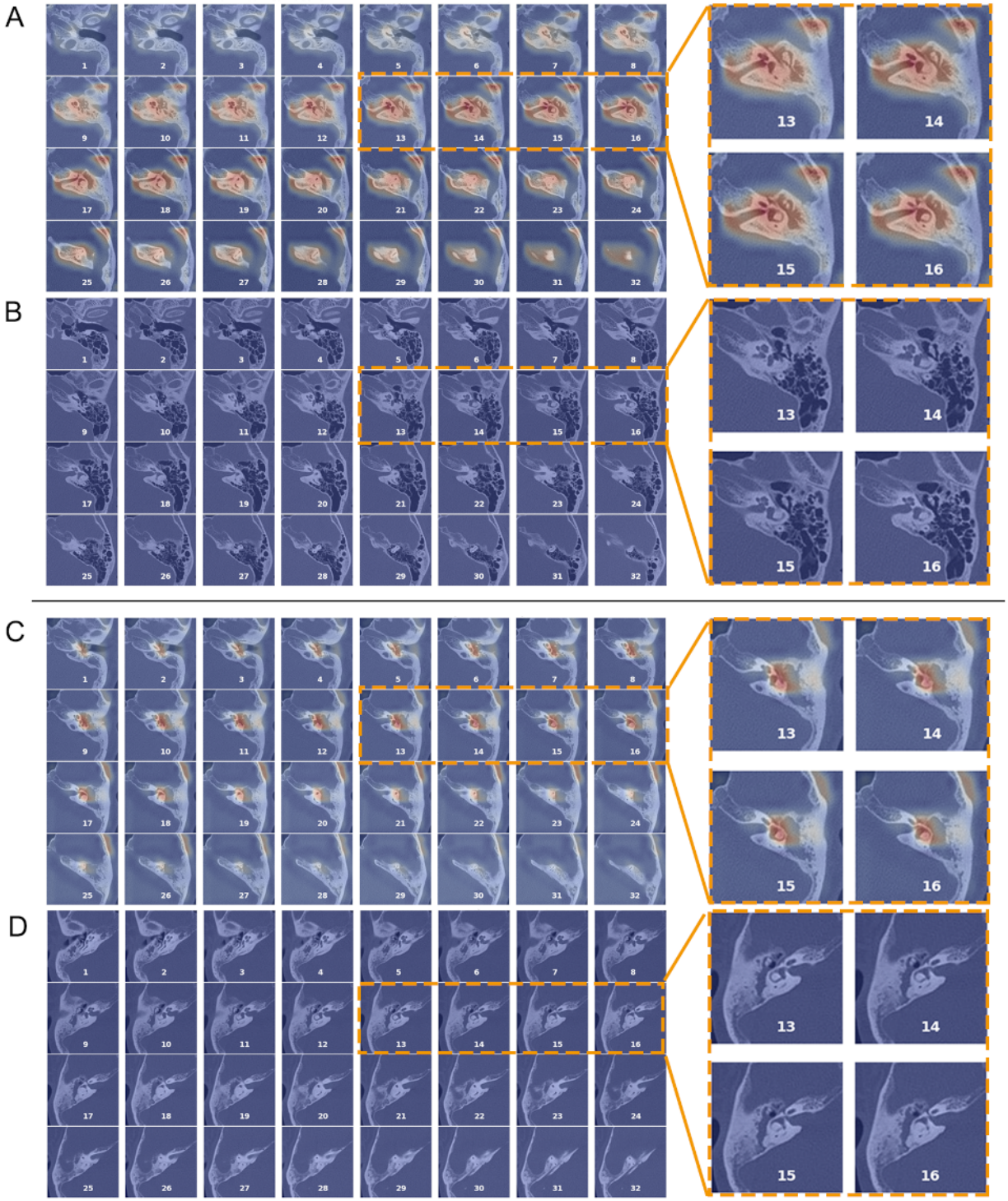

Clinical Use

The automatic evaluation system, incorporating the validated 3D model and the heatmap visualization technique, was evaluated for its viability in aiding preoperative assessment in 121 patients with COM (mean age 46.8, SD 16.1 years, 40.5% male). This system achieved an overall accuracy of 81.8% in distinguishing between cholesteatoma and noncholesteatoma cases. Sixty-nine ears were identified as free of cholesteatoma by the model, all of which received minimally invasive tympanoplasty under endoscopy. During the procedure, 9 ears (13.0%) revealed signs of cholesteatoma, and 5 of them required additional bone-grinding technique for complete removal of the mass. Cholesteatoma was initially predicted in 52 ears, with 37

XSL•FO
RenderX

(71.2%) of them undergoing canal-wall-down mastoidectomy. In the remaining 15 ears, the treating surgeons opted for endoscopic tympanoplasty, overriding the conventional technique for the model's predicted diagnosis. Clinicians reported that the model predictions aligned with their initial judgment or helped with their decision-making in 90.1% (109/121) cases. Postoperative hearing results were obtained in 87.6% (106/121) patients who maintained follow-up. Both groups of ears showed normal recovery, with a mean hearing gain of 8.5 (SD 15.6) and 5.5 (SD 18.1) dB, respectively (Figure 6).

Figure 6. Postoperative hearing gain for the operated ears with available audiometry outcomes (n=106). Data are categorized according to model predictions. Predictions that agree with the pathological results are denoted by close symbols, while open symbols indicate disagreements. Circles and triangles represent the treatment of endoscopic tympanoplasty and mastoidectomy, respectively. The error bars indicate ±1 SD from the mean.

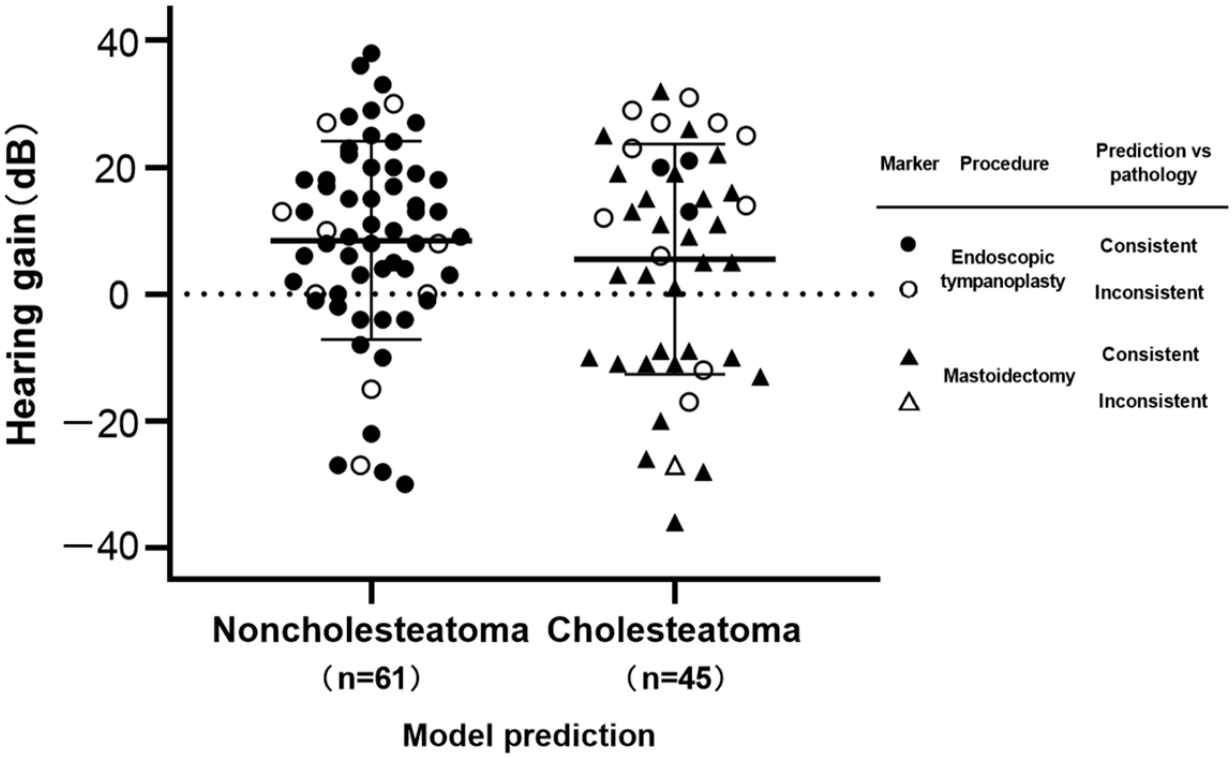

Discussion

Principal Results

This study demonstrates the robustness and generalizability of an AI model based on 3D CNN for the detection and differential diagnosis of COM using temporal bone CT scans. This model leverages multidimensional diagnostic information from the middle ear, resulting in a significant performance improvement compared with the traditional 2D approach. The framework exhibits comparable or even superior performances to human experts in otologic tasks with clinical significance and visual challenges, especially for classifying between cholesteatoma and noncholesteatomatous cases. In addition, the novel heatmap technique allows inspection of the AI's logic for decision-making, thereby enhancing the transparency of this model. The resulting AI system serves to automate summarization of critical radiologic findings and enables efficient evaluation of COM with minimum manual input. It provides tangible benefit in assisting otologists during preoperative assessment and results in favorable clinical outcomes that are comparable with historical results [34-37]. These findings further support the clinical viability and advantages of AI technology, which is expected to improve efficiency, reduce errors, and facilitate precision medicine in health care in the new era of big data.

Comparison With Prior Work

A few AI models have recently been developed to classify common middle ear conditions, such as CSOM, otitis media with effusion, and cholesteatoma [38-41]. However, these models were primarily based on traditional otoscopic images, which are potentially limited by a narrow field of view and insufficient diagnostic information. Temporal bone CT scans, which are increasingly used in otologic workup by virtue of its accessibility, rich amount of anatomical information, and adequate sensitivity in revealing pathological changes, have also been explored in a limited number of studies [22,42-45]. Although these AI models demonstrated decent AUROC scores (eg, >0.9) in common classification tasks, they were all trained to generate predictions based on 2D single-layer CT scans. A potential drawback is the increased likelihood of missing small or peripheral pathological changes (eg, an attic cholesteatoma) and the resultant false negatives.

Efforts were made in this study to establish a 3D approach to take full advantage of all available anatomical information and achieve a better coverage of the tympanum and the mastoid.

XSL•FO
RenderX

Inspection of the extracted ROI suggests that all critical anatomies are visible. Results from the benchmark test indicate that the proposed 3D model outperforms the state-of-the-art 2D approach by a modest performance gain in the detection of COM and by a much larger extent in differentiating between cholesteatoma and noncholesteatoma. This finding has several implications. First, both models are generally adequate in identifying common abnormal patterns from the CT, which are graphically characterized by increased opacification or soft tissue shadows in the middle ear cavity and indicative of pathological conditions in general. This is a relatively simple visual task, during which diagnostic information obtained from a single 2D CT slice is likely sufficient for the purpose and extra findings from other layers provide only minimal contribution to the decision-making. Second, the 3D model has huge advantage over the 2D approach in differentiating cholesteatoma from other types of COM. This task is known to be more visually challenging for humans, often requiring detection of subtle osseous erosions from multiple CT slices, as quite a few pathological changes caused by cholesteatoma are peripheral or noncharacteristic [32,33]. A substantial increase in each outcome measure justifies the advantage of the current 3D model for this task. Moreover, this 3D model has only a simple network structure with a small size (14.5 MB) as opposed to a complex and large-sized 2D network (274 MB), suggesting both higher computational efficiency and performance of the 3D approach. Finally, the AUROC of 0.92-0.94 and accuracy scores of 82.1%-86.1% achieved by the 2D network in this study in detecting COM were equivalent to historical results (0.92%-86%, respectively) in our previous study [22], further indicating the reliability of these findings and potentially the intrinsic limit of using single-layer CT scan for this task. To the best of our knowledge, this is the first study showing quantitative evidence to support the advantage of a 3D CNN model in 2 common otologic tasks based on temporal bone CT scans. It also advances beyond prior retrospective research by showcasing the practicality and benefits of the model in a clinical environment.

Clinical Implications

Cholesteatoma exhibits distinct histology marked by local invasiveness and a propensity for recurrence. The imperative for successful outcomes necessitates complete removal of the mass, particularly because recurrent cholesteatoma complicates revision surgery [46]. Suspected cases often require a canal-wall-down mastoidectomy to expose the tympanum, resulting in an open cavity and a permanently altered sound conduction pathway [46]. Accumulating evidence suggests that noncholesteatoma may spare from mastoidectomy and benefit from minimally invasive procedures such as endoscopic tympanoplasty [47,48]. Therefore, the current AI system holds potential value for otologists in surgical planning. Ears with a low risk of cholesteatoma, as identified by the model, could potentially be treated by less invasive procedures that retain the integrity of canal wall, leading to reduced procedural time and enhanced recovery [6,7,9,49,50]. This clinical merit is supported by the superior benchmark performance in identifying cholesteatoma and the favorable outcomes observed in the prospective study.

While detecting COM in task 1 involves spotting any pathological patterns on CT, which may not fully capture the differences between models in diagnostic capabilities, the increased visual challenges in identifying cholesteatoma substantiate the advantages of the proposed 3D approach for this task. In this study, the 3D model outperformed junior clinicians and demonstrated equivalent or superior performance to senior experts in identifying cholesteatoma based on CT. Notably, the 3D model achieved outcomes that were on par with or better than those based on human interpretation of MRI, which, despite its higher sensitivity, is a more expensive diagnostic method [22,43,45,51-53]. These findings underscore the 3D model's potential as a reliable and cost-effective alternative, offering sufficient COM evaluation with CT alone, thereby reducing the need for the pricier MRI.

The findings from the prospective study indicate that the model is efficacious in clinical environments, especially in distinguishing cholesteatoma from noncholesteatoma. Feedback from our clinical team highlights that the system serves as a reliable and streamlined source for a second opinion. Before surgery, the treating physician can rapidly identify essential details such as the lesion's location and properties, using the model's diagnostic output, and heatmaps. Concordance between the model's predictions and the physician's initial assessment bolsters confidence in surgical planning, thereby streamlining the diagnostic and therapeutic process. In contrast, discrepancies between the model's results and the physician's judgment prompt a detailed case reassessment or team consultation, aiding in the validation of a suitable treatment plan or preparing for intraoperative modifications. This process provides timely advisory support for complex cases, encouraging meticulous evaluation by the physician, minimizing errors, and keeping the clinician's cognitive load in check without compromising their autonomy in decision-making.

It should be noted that even for seasoned otologists and radiologists, who are adept at quickly and accurately reading temporal bone CT scans, a second opinion can add an extra layer of confidence to their assessments. For novice clinicians, who may find the diagnostic process more challenging and time-intensive (Table 5), the model may offer substantial improvements in both the accuracy and the speed of diagnosing and managing COM. This is particularly beneficial for physicians in smaller medical facilities or those early in their careers. Looking ahead, the integration of this model into electronic medical systems or cloud-based servers stands to streamline the provision of immediate second opinions or enable physicians from diverse locations to upload imaging data for dependable diagnostic insights. Such technological progress is poised to advance individualized COM treatments in the big data era, boosting efficiency, reducing costs, and enhancing the quality of patient care.

Research Insights

Efforts were undertaken in this study to demystify the criticized nontransparency of DL models, characterized by intricate decision-making strategies within multilayer architectures [30,54]. The nonlinear interactions among these components can yield incomprehensible logic and untraceable predictions

XSL•FO
RenderX

vulnerable to bias or errors, posing a significant challenge to the widespread application of AI in health care. To address this issue, heatmaps, and specifically, the Gradient-Weighted Class Activation Mapping technique, have been used as a method to inspect AI's strategy and enhance human interpretation in a parsimonious manner [55-57]. In this study, the strategy learned by our models to focus on the middle ear and mastoid regions appeared reasonable and aligned with human knowledge in interpreting CT for COM, reinforcing the reliability of this framework. These informative heatmaps can aid clinicians in understanding and validating AI predictions for specific cases, or serve as educational tools for training medical students or junior residents in reading temporal bone CT scans. Ultimately, this approach presents a viable solution for developing explainable AI models for clinical tasks.

Overfitting is a common concern with DL models, especially when data are limited or sourced from a single institute. It can lead to poor performance on new data despite promising results on the original data set. Previous DL models were trained on monocentric CT scans with participant counts ranging from 61 to 562. Lack of external validation and small sample sizes may raise concern about potential overfitting of these models [22,42,43]. Several approaches were used in this study to enhance the generalizability of our framework. First, our models underwent cross-validation on a major data set comprising more than 3000 ears, the largest sample size reported to date. Second, these models were evaluated on external data with different patient origins and image properties. Third, several machine learning methods were applied to minimize the risk of models being tuned to the random features, including early termination of training and the use of a dropout function to decrease the interdependency among network nodes [27]. Consistent performance metrics across data sets in both tasks substantiated the generalizability of this framework. Moreover, the region proposal method proved applicable to CT scans from both sources, demonstrating adaptability despite differences in CT scanner, scan settings, and image quality.

Limitations

This study has several limitations. First, although an external data set was obtained from a hospital in a different city, patients in both data sets shared a common racial background. Further validation on data collected from patients with diverse origins may be necessary to ensure the generalizability of these models. Second, the research was constrained to 2 binary classification tasks relevant to COM. Incorporating additional diagnostic tasks, such as assessing the ossicular chain's integrity and forecasting auditory outcomes, may enrich the diagnostic toolkit. Third, the models were exclusively trained to analyze CT scans, potentially not leveraging AI's full potential in COM evaluation. Comprehensive diagnostics often involve synthesizing information from patient history, clinical symptoms, ear examinations, audiological testing, otoscopy, and various imaging techniques. Overreliance on CT scans alone may introduce limitations in performance and may not always lead to conclusive diagnoses (Figure 7). Fourth, the ablation study examined a limited array of model alternatives. Despite achieving notable performance through initial model structure refinement, future endeavors should include ongoing optimization of the model architecture and detailed analysis of network component functions to optimize the trade-off between model efficacy and computational demands. In addition, this study did not place extensive emphasis on exploring common ethical issues, such as patient privacy, data security, and human autonomy, which are critical considerations in the clinical application of AI and warrant ongoing attention. Finally, this study reported initial findings from the clinical application of the AI system in a small, prospective cohort without a control group. Although the main objective was to show that the current model is ready for clinical implementation, a thorough assessment of the model's clinical benefits will be conducted in an upcoming clinical trial with a more rigorous research design.

XSL•FO
RenderX

Figure 7. Examples of misclassified cases. (A) A pathological ear showing a small-sized soft tissue density near the ossicles (arrows) with no evident sign of osseous erosion or mastoid opacification. (B) A case of cholesteatoma showing soft tissue density (asterisks) but with a visually intact ossicular chain (arrows) and a normal-sized tympanic cavity.

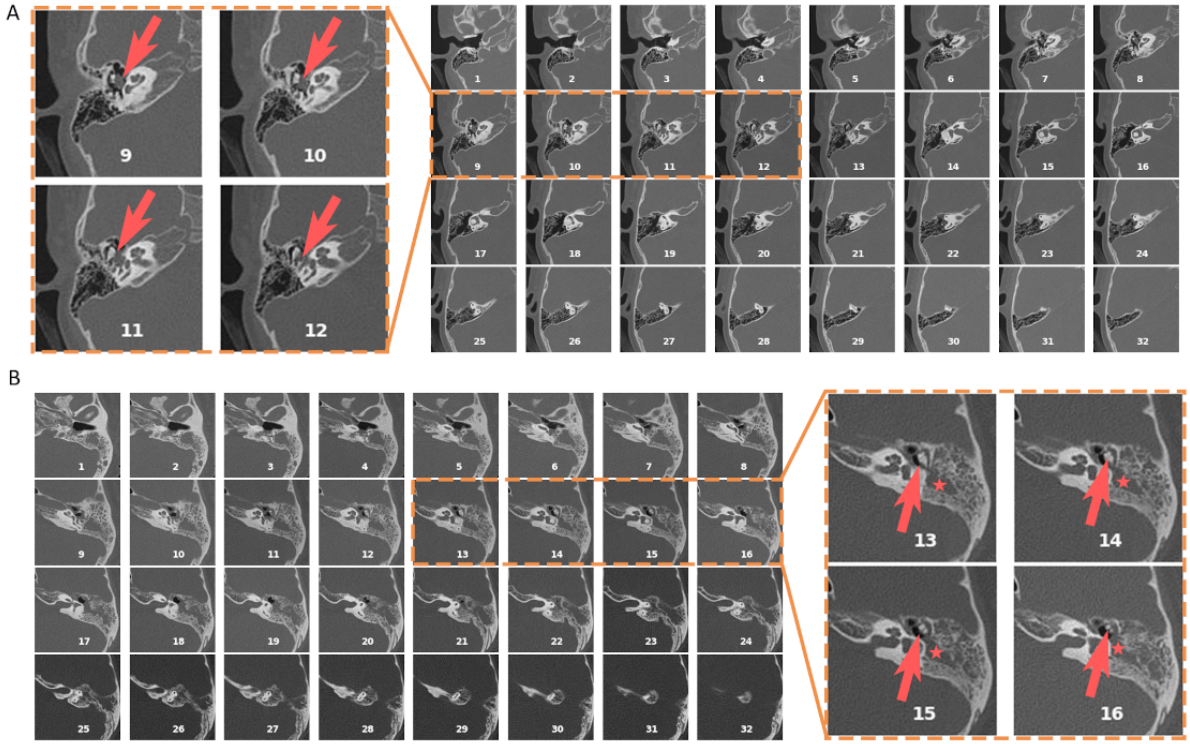

Future Research

Future studies will focus on leveraging novel techniques to enhance model performance and evaluate the effectiveness in larger-scale controlled trials. For example, new models will be trained to perform additional tasks, including evaluation of ossicular chain and forecasting postoperative hearing, which may enhance features of the current AI framework. A broader data set will be compiled from hospitals worldwide to assess and refine the generalizability of these models. Moreover, future models will potentially incorporate multiple sources of clinical information with a fusion layer for generating predictions, mimicking human decision-making strategies, and potentially enhancing model robustness. Ongoing efforts will also be made to refine model architectures and to address ethical issues associated with the use of AI in health care. An active learning framework may be established to integrate feedback loops, allowing clinicians to provide input to the model. This approach is expected to support ongoing model enhancement and reinforcement learning based on human feedback. In the next stage, multicenter, prospective human trials will be conducted to assess the practical benefits of implementing these AI models in clinical contexts. The ultimate goal of this research line is to establish a robust AI system that can assist clinicians with reliability, efficiency, and transparency in the evaluation and management of ear diseases.

Conclusions

This study presents a 3D CNN model trained to detect pathological changes and identify cholesteatoma based on temporal bone CT scans. The model's performance significantly surpasses the baseline 2D approach, reaching a level comparable with or even exceeding that of human experts in both tasks. The model also exhibits decent generalizability and enhanced comprehensibility through the gradient heatmaps. The resulting AI system allows automatic assessment of COM and shows promising viability in real-word clinical settings. These findings imply the potential of AI as a valuable tool for aiding clinicians in the evaluation of COM. Future research will involve enhancing models with additional source of diagnostic information to perform various clinical tasks and evaluating the benefits of AI models in large-scale controlled trials.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (grants 81970889 to Fanglu Chi and 82271166, 81970880, and 81771017 to Dongdong Ren, U22A20249, 52188102, 52027806 to Jiajie Guo); Natural Science Foundation of Shanghai (grant 22ZR1410100 to Dongdong Ren); the "Zhuo-Xue Plan" of Fudan University (Dongdong Ren); Heng-Jie special technical support plan (Dongdong Ren); and the Shanghai Outstanding Young Medical Talent Program (Dongdong Ren). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

XSL•FO
RenderX

Data Availability

The complete source code and trained models are available in Multimedia Appendices 1 and 2 as well as in a public repository, which can be accessed in GitHub [58]. The automatic evaluation system is available for individual use with a detailed instruction manual and a walk-through tutorial. The data sets generated during this study may be obtained from the corresponding authors upon reasonable request.

Authors' Contributions

YL conceptualized and designed the study, reviewed and analyzed the data, performed computer programming, developed and evaluated the AI models, wrote and edited the manuscript, prepared the figures, and supervised the entire project; BC retrieved, validated and analyzed the data, drafted the manuscript, and prepared the figures; YS provided data resources and validated the data. HS, YW, and JL retrieved data and evaluated the models. XN and LY retrieved and validated the data. JG acquired funding support, validated the model, and edited the manuscript. SB performed model validation and deployment. YC, JX, and JY evaluated the models. YH and BL provided data resources; FC provided data resources and funding support; and DR conceptualized the study, provided funding support and data resources, and edited the manuscript. All authors have reviewed, discussed, and approved the manuscript. No generative artificial intelligence tool was used during the preparation of this manuscript. BC, YL, and YS have contributed equally to this study and are credited as co-first authors. YL, BL, and DR are designated as the corresponding authors. Contact details for correspondence are as follows: Yike Li, Department of Otolaryngology-Head and Neck Surgery, Vanderbilt University Medical Center, 1215 21st Avenue South, Rm. 10410, Medical Center East, Nashville, TN 37232, USA. E-mail: yike.li.1@vumc.org. ORCID: 0000-0001-8465-130X; Bo Liang, Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Jiefang Avenue #1277, Wuhan, Hubei, 430022, China. E-mail: xiehelb@163.com. ORCID: 0000-0002-3494-4187; DongDong Ren, Department of Otorhinolaryngology, Eye, Ear, Nose and Throat Hospital, 83 Fenyang Road, Shanghai, 200031, China. E-mail: dongdongren@fudan.edu.cn. ORCID: 0000-0002-2889-9375.

Conflicts of Interest

YL served as an associate editor for the Journal of Medical Internet Research at the time of manuscript submission. YL has abstained from participating in any peer-reviewed or editorial decision-making processes related to this article.

Multimedia Appendix 1
Additional tables.
[DOCX File , 24 KB - jmir_v26i1e51706_app1.docx]

Multimedia Appendix 2
Source code for model development and validation.
[ZIP File (Zip Archive), 69 KB - jmir_v26i1e51706_app2.zip]

References

1. Schilder AGM, Chonmaitree T, Cripps AW, Rosenfeld RM, Casselbrant ML, Haggard MP, et al. Otitis media. Nat Rev Dis Primers 2016;2(1):16063. [doi: 10.1038/nrdp.2016.63] [Medline: 27604644]
2. World Health Organization. Chronic Suppurative Otitis Media: Burden of Illness and Management Options. Geneva, Switzerland: World Health Organization; 2004.
3. Lustig L, Limb C, Baden R, LaSalvia M. Chronic Otitis Media, Cholesteatoma, and Mastoiditis in Adults. Waltham, MA (citirano 145 2019): UpToDate; 2018.
4. Takahashi M, Motegi M, Yamamoto K, Yamamoto Y, Kojima H. Endoscopic tympanoplasty type I using interlay technique. J Otolaryngol Head Neck Surg 2022;51(1):45. [doi: 10.1186/s40463-022-00597-3] [Medline: 36397175]
5. Ohki M, Kikuchi S, Tanaka S. Endoscopic type 1 tympanoplasty in chronic otitis media: comparative study with a postauricular microscopic approach. Otolaryngol Head Neck Surg 2019;161(2):315-323. [doi: 10.1177/0194599819838778]
6. Hsu Y, Kuo C, Huang T. A retrospective comparative study of endoscopic and microscopic tympanoplasty. J Otolaryngol Head Neck Surg 2018;47(1):44. [doi: 10.1186/s40463-018-0289-4]
7. Yang Q, Wang B, Zhang J, Liu H, Xu M, Zhang W. Comparison of endoscopic and microscopic tympanoplasty in patients with chronic otitis media. Eur Arch Otorhinolaryngol 2022;279(10):4801-4807. [doi: 10.1007/s00405-022-07273-2]
8. Tsetsos N, Vlachtsis K, Stavrakas M, Fyrmpas G. Endoscopic versus microscopic ossiculoplasty in chronic otitis media: a systematic review of the literature. Eur Arch Otorhinolaryngol 2020;278(4):917-923. [doi: 10.1007/s00405-020-06182-6]
9. Tarabichi M, Ayache S, Nogueira JF, Al Qahtani M, Pothier DD. Endoscopic management of chronic otitis media and tympanoplasty. Otolaryngol Clin North Am 2013;46(2):155-163. [doi: 10.1016/j.otc.2012.12.002]
10. Watts S, Flood LM, Clifford K. A systematic approach to interpretation of computed tomography scans prior to surgery of middle ear cholesteatoma. J Laryngol Otol 2000;114(4):248-253. [doi: 10.1258/0022215001905454]

XSL•FO
RenderX

11. Selwyn D, Howard J, Cuddihy P. Pre-operative prediction of cholesteatomas from radiology: retrospective cohort study of 106 cases. J Laryngol Otol 2019;133(06):477-481. [doi: 10.1017/s0022215119001154]
12. Songu M, Altay C, Onal K, Arslanoglu S, Balci MK, Ucar M, et al. Correlation of computed tomography, echo-planar diffusion-weighted magnetic resonance imaging and surgical outcomes in middle ear cholesteatoma. Acta Otolaryngol 2015;135(8):776-780. [doi: 10.3109/00016489.2015.1021931]
13. Mahmutoğlu AS, Celebi I, Sahinoğlu S, Cakmakçi E, Sözen E. Reliability of preoperative multidetector computed tomography scan in patients with chronic otitis media. J Craniofac Surg 2013;24(4):1472-1476. [doi: 10.1097/scs.0b013e31829031b1]
14. Pandey AK, Bapuraj JR, Gupta AK, Khandelwal N. Is there a role for virtual otoscopy in the preoperative assessment of the ossicular chain in chronic suppurative otitis media? Comparison of HRCT and virtual otoscopy with surgical findings. Eur Radiol 2009;19(6):1408-1416. [doi: 10.1007/s00330-008-1282-5] [Medline: 19153741]
15. Chee NW, Tan TY. The value of pre-operative high resolution CT scans in cholesteatoma surgery. Singapore Med J 2001;42(4):155-159.
16. Zaman SU, Rangankar V, Muralinath K, Shah V, Pawar R. Temporal bone cholesteatoma: typical findings and evaluation of diagnostic utility on high resolution computed tomography. Cureus 2022;14(3):e22730. [doi: 10.7759/cureus.22730] [Medline: 35386487]
17. Gaurano JL, Joharjy IA. Middle ear cholesteatoma: characteristic CT findings in 64 patients. Ann Saudi Med 2004;24(6):442-447. [doi: 10.5144/0256-4947.2004.442] [Medline: 15646162]
18. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954-961. [doi: 10.1038/s41591-019-0447-x] [Medline: 31110349]
19. Mikhael PG, Wohlwend J, Yala A, Karstens L, Xiang J, Takigami AK, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. J Clin Oncol 2023;41(12):2191-2200. [doi: 10.1200/jco.22.01345]
20. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol 2021;22(1):132-141. [doi: 10.1016/s1470-2045(20)30535-0] [Medline: 33387492]
21. Li Y, Guo J, Yang P. Developing an image-based deep learning framework for automatic scoring of the pentagon drawing test. J Alzheimers Dis 2022;85(1):129-139. [doi: 10.3233/jad-210714]
22. Wang Y, Li Y, Cheng Y, He Z, Yang J, Xu J, et al. Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. Ear Hear 2020;41(3):669-677. [doi: 10.1097/aud.0000000000000794]
23. Sundgaard JV, Harte J, Bray P, Laugesen S, Kamide Y, Tanaka C, et al. Deep metric learning for otitis media classification. Med Image Anal 2021;71:102034. [doi: 10.1016/j.media.2021.102034] [Medline: 33848961]
24. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886. [doi: 10.1136/bmj.l886]
25. Castelvecchi D. Can we open the black box of AI? Nature 2016;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
26. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. : IEEE; 2016 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; Las Vegas, NV, USA p. 779-788. [doi: 10.1109/CVPR.2016.91]
27. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J machine Learn Res 2014;15(1):1929-1958.
28. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2014;13:54-29 Preprint published online Dec 2014 [FREE Full text]
29. Tseng C, Lai M, Wu C, Yuan S, Ding Y. Comparison of the efficacy of endoscopic tympanoplasty and microscopic tympanoplasty: a systematic review and meta‐analysis. Laryngoscope 2016;127(8):1890-1896. [doi: 10.1002/lary.26379] [Medline: 27861950]
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. 2017 Presented at: Proceedings of the IEEE International Conference on Computer Vision; 2017; Venice, Italy p. 618-626.
31. Python Core Team. Python: a dynamic, open source programming language. Python Software Foundation. URL: https://www.python.org/ [accessed 2024-06-01]
32. Baráth K, Huber AM, Stämpfli P, Varga Z, Kollias S. Neuroradiology of cholesteatomas. AJNR Am J Neuroradiol 2011;32(2):221-229. [doi: 10.3174/ajnr.a2052]
33. Gulati M, Gupta S, Prakash A, Garg A, Dixit R. HRCT imaging of acquired cholesteatoma: a pictorial review. Insights Imaging 2019;10(1):92. [doi: 10.1186/s13244-019-0782-y] [Medline: 31578644]
34. Daneshi A, Daneshvar A, Asghari A, Farhadi M, Mohebbi S, Mohseni M. Endoscopic versus microscopic cartilage myringoplasty in chronic otitis media. Iran J Otorhinolaryngol 2020;32(112):263-269. [doi: 10.22038/IJORL.2020.44015.2453]

XSL•FO
RenderX

35. Prasad SC, Melia CL, Medina M, Vincenti V, Bacciu A, Bacciu S, et al. Long-term surgical and functional outcomes of the intact canal wall technique for middle ear cholesteatoma in the paediatric population. Acta Otorhinolaryngol Ital 2014;34(5):354-361.
36. Wood CB, O'Connell BP, Lowery AC, Bennett ML, Wanna GB. Hearing outcomes following type 3 tympanoplasty with stapes columella grafting in canal wall down mastoidectomy. Ann Otol Rhinol Laryngol 2019;128(8):736-741. [doi: 10.1177/0003489419841400]
37. Chamoli P, Singh CV, Radia S, Shah AK. Functional and anatomical outcome of inside out technique for cholesteatoma surgery. Am J Otolaryngol 2018;39(4):423-430. [doi: 10.1016/j.amjoto.2018.04.008] [Medline: 29748054]
38. Wu Z, Lin Z, Li L, Pan H, Chen G, Fu Y, et al. Deep learning for classification of pediatric otitis media. Laryngoscope 2020;131(7):E2344-E2351. [doi: 10.1002/lary.29302]
39. Pichichero ME. Can machine learning and AI replace otoscopy for diagnosis of otitis media? Pediatrics 2021;147(4):e2020049584. [doi: 10.1542/peds.2020-049584]
40. Tseng CC, Lim V, Jyung RW. Use of artificial intelligence for the diagnosis of cholesteatoma. Laryngoscope Invest Otolaryngol 2023;8(1):201-211. [doi: 10.1002/lio2.1008]
41. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: an automated machine learning approach. Laryngoscope 2020;130(6):1408-1413. [doi: 10.1002/lary.28292] [Medline: 31532858]
42. Duan B, Guo Z, Pan L, Xu Z, Chen W. Temporal bone CT-based deep learning models for differential diagnosis of primary ciliary dyskinesia related otitis media and simple otitis media with effusion. Am J Transl Res 2020;14(7):4728-4735.
43. Eroğlu O, Eroğlu Y, Yıldırım M, Karlıdag T, Çınar A, Akyiğit A, et al. Is it useful to use computerized tomography image-based artificial intelligence modelling in the differential diagnosis of chronic otitis media with and without cholesteatoma? Am J Otolaryngol 2022;43(3):103395. [doi: 10.1016/j.amjoto.2022.103395]
44. Khosravi M, Jabbari Moghaddam Y, Esmaeili M, Keshtkar A, Jalili J, Tayefi Nasrabadi H. Classification of mastoid air cells by CT scan images using deep learning method. J Big Data 2022;9(1):1-14. [doi: 10.1186/s40537-022-00596-1]
45. Wang Z, Song J, Su R, Hou M, Qi M, Zhang J, et al. Structure-aware deep learning for chronic middle ear disease. Expert Syst Appl 2022;194:116519. [doi: 10.1016/j.eswa.2022.116519]
46. Tomlin J, Chang D, McCutcheon B, Harris J. Surgical technique and recurrence in cholesteatoma: a meta-analysis. Audiol Neurotol 2013;18(3):135-142. [doi: 10.1159/000346140]
47. Lee S, Lee DY, Seo Y, Kim YH. Can endoscopic tympanoplasty be a good alternative to microscopic tympanoplasty? a systematic review and meta-analysis. Clin Exp Otorhinolaryngol 2019;12(2):145-155. [doi: 10.21053/ceo.2018.01277] [Medline: 30674106]
48. Trinidade A, Page JC, Dornhoffer JL. Therapeutic mastoidectomy in the management of noncholesteatomatous chronic otitis media. Otolaryngol Head Neck Surg 2016;155(6):914-922. [doi: 10.1177/0194599816662438] [Medline: 27484233]
49. Wu L, Liu Q, Gao B, Huang S, Yang N. Comparison of endoscopic and microscopic management of attic cholesteatoma: a randomized controlled trial. Am J Otolaryngol 2022;43(3):103378. [doi: 10.1016/j.amjoto.2022.103378] [Medline: 35177254]
50. Toulouie S, Block‐Wheeler NR, Rivero A. Postoperative pain after endoscopic vs microscopic otologic surgery: a systematic review and meta-analysis. Otolaryngol Head Neck Surg 2022;167(1):25-34. [doi: 10.1177/01945998211041946] [Medline: 34491858]
51. Profant M, Sláviková K, Kabátová Z, Slezák P, Waczulíková I. Predictive validity of MRI in detecting and following cholesteatoma. Eur Arch Otorhinolaryngol 2012;269(3):757-765. [doi: 10.1007/s00405-011-1706-8]
52. Lin M, Sha Y, Sheng Y, Chen W. Accuracy of 2D blade turbo gradient- and spin-echo diffusion weighted imaging for the diagnosis of primary middle ear cholesteatoma. Otol Neurotol 2022;43(6):e651-e657. [doi: 10.1097/MAO.0000000000003521] [Medline: 35261384]
53. Sharifian H, Taheri E, Borghei P, Shakiba M, Jalali AH, Roshanfekr M, et al. Diagnostic accuracy of non‐echo‐planar diffusion‐weighted MRI versus other MRI sequences in cholesteatoma. J Med Imaging Radiat Oncol 2012;56(4):398-408. [doi: 10.1111/j.1754-9485.2012.02377.x] [Medline: 22883647]
54. Montavon G, Lapuschkin S, Binder A, Samek W, Müller K. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 2017;65:211-222. [doi: 10.1016/j.patcog.2016.11.008]
55. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V. A deep learning and grad-CAM based color visualization approach for fast detection of Covid-19 cases using chest X-ray and CT-scan images. Chaos Solitons Fractals 2020;140:110190. [doi: 10.1016/j.chaos.2020.110190] [Medline: 32836918]
56. Cheng C, Ho T, Lee T, Chang C, Chou C, Chen C, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29(10):5469-5477. [doi: 10.1007/s00330-019-06167-y] [Medline: 30937588]
57. He T, Guo J, Chen N, Xu X, Wang Z, Fu K, et al. MediMLP: using grad-CAM to extract crucial variables for lung cancer postoperative complication prediction. IEEE J Biomed Health Inform 2020;24(6):1762-1771. [doi: 10.1109/jbhi.2019.2949601] [Medline: 31670685]
58. huntlylee / 3D-Otitis-Media. GitHub. URL: https://github.com/huntlylee/3D-Otitis-Media [accessed 2024-08-01]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
COM: chronic otitis media
CSOM: chronic suppurative otitis media
CT: computed tomography
DL: deep learning
EENT: Eye, Ear, Nose, and Throat Hospital of Fudan University
MRI: magnetic resonance imaging
ROI: region of interest
WU: Wuhan Union Hospital
YOLO: You Only Look Once

Edited by S Ma; submitted 09.08.23; peer-reviewed by Q Chen, SA Javed, CN Hang; comments to author 02.11.23; revised version received 30.11.23; accepted 29.05.24; published 08.08.24.

Please cite as:
Chen B, Li Y, Sun Y, Sun H, Wang Y, Lyu J, Guo J, Bao S, Cheng Y, Niu X, Yang L, Xu J, Yang J, Huang Y, Chi F, Liang B, Ren D
A 3D and Explainable Artificial Intelligence Model for Evaluation of Chronic Otitis Media Based on Temporal Bone Computed Tomography: Model Development, Validation, and Clinical Application
J Med Internet Res 2024;26:e51706
URL: https://www.jmir.org/2024/1/e51706
doi:10.2196/51706
PMID:

©Binjun Chen, Yike Li, Yu Sun, Haojie Sun, Yanmei Wang, Jihan Lyu, Jiajie Guo, Shunxing Bao, Yushu Cheng, Xun Niu, Lian Yang, Jianghong Xu, Juanmei Yang, Yibo Huang, Fanglu Chi, Bo Liang, Dongdong Ren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Differences in Fear and Negativity Levels Between Formal and Informal Health-Related Websites: Analysis of Sentiments and Emotions

Abigail Paradise Vit[1*], PhD; Avi Magid[1,2*], PhD

[1]Department of Information Systems, The Max Stern Yezreel Valley College, Emek Yezreel, Israel
[2]Management, Rambam Health Care Campus, Haifa, Israel
[*]all authors contributed equally

Corresponding Author:
Abigail Paradise Vit, PhD
Department of Information Systems, The Max Stern Yezreel Valley College
Emek Yezreel 1
Emek Yezreel, 1930600
Israel
Phone: 972 509903930
Email: abigailparadise@gmail.com

Abstract

Background: Searching for web-based health-related information is frequently performed by the public and may affect public behavior regarding health decision-making. Particularly, it may result in anxiety, erroneous, and harmful self-diagnosis. Most searched health-related topics are cancer, cardiovascular diseases, and infectious diseases. A health-related web-based search may result in either formal or informal medical website, both of which may evoke feelings of fear and negativity.

Objective: Our study aimed to assess whether there is a difference in fear and negativity levels between information appearing on formal and informal health-related websites.

Methods: A web search was performed to retrieve the contents of websites containing symptoms of selected diseases, using selected common symptoms. Retrieved websites were classified into formal and informal websites. Fear and negativity of each content were evaluated using 3 transformer models. A fourth transformer model was fine-tuned using an existing emotion data set obtained from a web-based health community. For formal and informal websites, fear and negativity levels were aggregated. *t* tests were conducted to evaluate the differences in fear and negativity levels between formal and informal websites.

Results: In this study, unique websites (N=1448) were collected, of which 534 were considered formal and 914 were considered informal. There were 1820 result pages from formal websites and 1494 result pages from informal websites. According to our findings, fear levels were statistically higher (t_{2753}=3.331; P<.001) on formal websites (mean 0.388, SD 0.177) than on informal websites (mean 0.366, SD 0.168). The results also show that the level of negativity was statistically higher (t_{2753}=2.726; P=.006) on formal websites (mean 0.657, SD 0.211) than on informal websites (mean 0.636, SD 0.201).

Conclusions: Positive texts may increase the credibility of formal health websites and increase their usage by the general public and the public's compliance to the recommendations. Increasing the usage of natural language processing tools before publishing health-related information to achieve a more positive and less stressful text to be disseminated to the public is recommended.

(J Med Internet Res 2024;26:e55151) doi:10.2196/55151

KEYWORDS

emotions; sentiment; health websites; fear

Introduction

Background

A wide range of medical information is available on the web and can be used by individuals who are not health care professionals to gain a better understanding of health and illness, as well as to provide an explanation for their symptoms [1,2]. However, it should be noted that health-related search information may have a significant impact on people's decisions and concerns, including the ability to diagnose and assess their

XSL•FO
RenderX

own health care needs, deciding whether or not to consult with a physician for assistance with diagnosis and treatment, and their approach to maintaining their own and their families' health [3].

People can experience anxiety when reading information on the internet [4-6]. Exposing people without medical knowledge to complex medical terminology may put them at risk of self-diagnosis and self-treatment that could be harmful to them [7]. According to White and Horvitz [1], web search engines can escalate health concerns and various factors contribute to this escalation, including the amount and distribution of health-related content viewed by users, the presence of escalatory terminology on the pages visited, and a user's propensity to escalate concerns in order to find more plausible explanations for their ailments. A condition known as "cyberchondria" is the creation of unfounded concerns about common symptoms based upon web-based searches and other sources of information [1].

People search health-related information on various topics. Previous studies showed that the most searched health-related topics were the cancer diseases, cardiovascular diseases, and infectious diseases [8]. Web search of health-related topics may lead to formal websites, originating from an official health organization such as the World Health Organization, and to informal websites, originating from unofficial health-related websites, including blogs of people without appropriate medical training [9]. Both formal and informal health-related websites may contain information that may increase the reader's fear while reading the information [10]. However, whether there is a difference in fear and negativity levels between health-related information that appears in formal health-related websites and health-related information that appears in informal websites remains questionable.

Sentiment and emotion analysis are natural language processing (NLP) techniques for the identification of sentiment and emotion from speech or voice data, images, or text data [11-13]. In sentiment analysis, the overall polarity of a text is identified—whether it is positive, negative, or neutral. Emotion analysis involves the identification of specific emotions expressed in a text. In accordance with Ekman's [14] model, several basic emotions can be identified, such as happiness, sadness, anger, disgust, surprise, and fear.

Different methods can be used to detect sentiment and emotions in a given text, and in recent years transformer models have become increasingly popular, especially in NLP. Transformer model is a deep neural network that learns context and meaning using self-attention [15]. Text classification using the transformer model achieved similar accuracy to human annotations. In recent years, several pretrained language models have been introduced, including BERT [15], XLNet [16], RoBERTa [17], DistilBERT [18], and ALBERT [19]. Our study aimed to assess the association between the type of health-related website (formal or informal) and 2 outcome variables: the level of fear and the level of negativity by using transformer for sentiment and emotion detection.

Related Work

The analysis of emotions and sentiments in health information has become increasingly popular in recent years, particularly with the rise of social media as platforms for sharing health-related information [20]. This section provides an overview of studies that investigated the analysis of emotions and sentiments in health-related information.

Sentiment Analysis in Web-Based Health Information

Studies have examined sentiments of patients with cancer in web-based forums [21,22]. Cabling et al [22] analyzed the feelings and opinions of web-based breast cancer support group users regarding tamoxifen, hormone-based therapy used to treat breast cancer. Results showed that the most active users were significantly more positive than the least active users, who were more negative. Users with higher cancer stages were less likely to post, focusing on side effects and associated anxiety and sadness. In contrast, users with lower cancer stages were more likely to post, remain active, and encourage social support. Carrillo-de-Albornoz et al [23] analyzed more than 3500 posts on web-based health forums concerning breast cancer, Crohn disease, and various allergies. They trained different machine learning algorithms to automatically classify patient-authored content into positive, negative, and neutral categories. Using word embeddings, they predicted the polarity of patient-authored content with an accuracy of 0.7, which indicates the times that the algorithms correctly classified the sentiment of the posts, outperforming more traditional methods.

In addition, other studies have examined sentiment analysis of patient responses online [24-27]. Greaves et al [24], for instance, used machine learning to understand patients' comments about their care in web-based comments posted about hospitals on the English National Health Service website. A sentiment analysis technique was used to categorize web-based free-text comments made by patients as positive or negative. Ali et al [25] analyzed 26 different threads on 3 medical forums and determined whether the posts were positive, negative, or neutral regarding hearing loss.

Considering the analysis of emotions in web-based health information, Khanpour and Caragea [28] proposed ConvLexLSTM to identify fine-grained emotion types in health-related web-based posts in web-based health communities. This model combines Convolutional Neural Network features with lexicon-based features, which are fed into a Long Short Term Memory network. They demonstrate that ConvLexLSTM outperforms strong baselines and prior works. Later in 2020, Sosea and Caragea [29] presented CANCEREMO, an emotion data set created from a web-based health community and annotated with 8 fine-grained emotions. Their most effective BERT model achieves an average F_1-score of 71%.

Sentiment Analysis of Web-Based Health-Related Information on Social Media

An extensive amount of health data is disseminated via social media platforms such as Twitter and Facebook [30]. Studies in this category analyze users' sentiments regarding disease and mental health, for example, sentiments in responses to YouTube

XSL•FO
RenderX

videos related to proanorexia and antianorexia [31]. Similarly, Roccetti et al [32] explore attitudes toward Crohn disease on Facebook and Twitter. Furthermore, Ricard et al [33] examine the potential of community-generated social media content for detecting depression on Instagram.

Health discourse on social media has been significantly affected by the COVID-19 pandemic [34]. Public attitudes and opinions regarding vaccination have been the subject of several studies [35-41]. Using active learning to select a large population of health care professionals' Twitter accounts, Elyashar et al [42] analyzed health care professionals' web-based discussions as expressed in web-based discussions published on Twitter during the COVID-19 pandemic. According to their analysis, the intensity of emotion in their discourses decreased with the pandemic waves in 2020. They revealed a decline in joy and an increase in sadness, fear, and disgust in that year's discourses. Aduragba et al [43] propose EmoBERT, an emotion-based variant of the BERT transformer model, able to learn emotion representations during major disease outbreaks using the Twitter platform and outperform the state-of-the-art.

Christensen et al [44] analyzed 172 major news outlets across 11 countries. As a result of keyword-based frequency analysis, the percentage of vaccination-related papers in all collated English-language papers was calculated. The Vader Python module, developed by CJ Hutto and Eric Gilbert, quantified sentiment polarization. The number of negatively polarized papers increased from 6698 in 2015-2019 to 28,552 in 2020-2021 (the COVID-19 pandemic period).

In contrast to previous studies, which analyzed textual data posted by users and their posts, reviews, and comments for sentiment analysis, our objective is to gain a better understanding of the sentiments and emotions expressed on formal and informal health-related websites. By focusing primarily on the differences in negative and fear levels expressed on formal and informal health-related websites, we hope to provide a more comprehensive understanding of fear and negative sentiment that is provided to the public through formal and informal health-related websites.

Methods

Overview

Figure 1 summarizes the seven main phases we followed to conduct the proposed study: (1) based on literature, identifying the symptoms associated with each family disease, (2) conducting Google searches for each symptom, (3) retrieving the top 200 pages returned by Google for each symptom, (4) classifying each website as formal or informal by manual labeling, (5) creating a data set of websites symptom results, (6) applying content, sentiment, and emotion analysis to the text of each website, and (7) aggregating the results to fear and negativity levels.

Figure 1. An overview of the main phases performed.

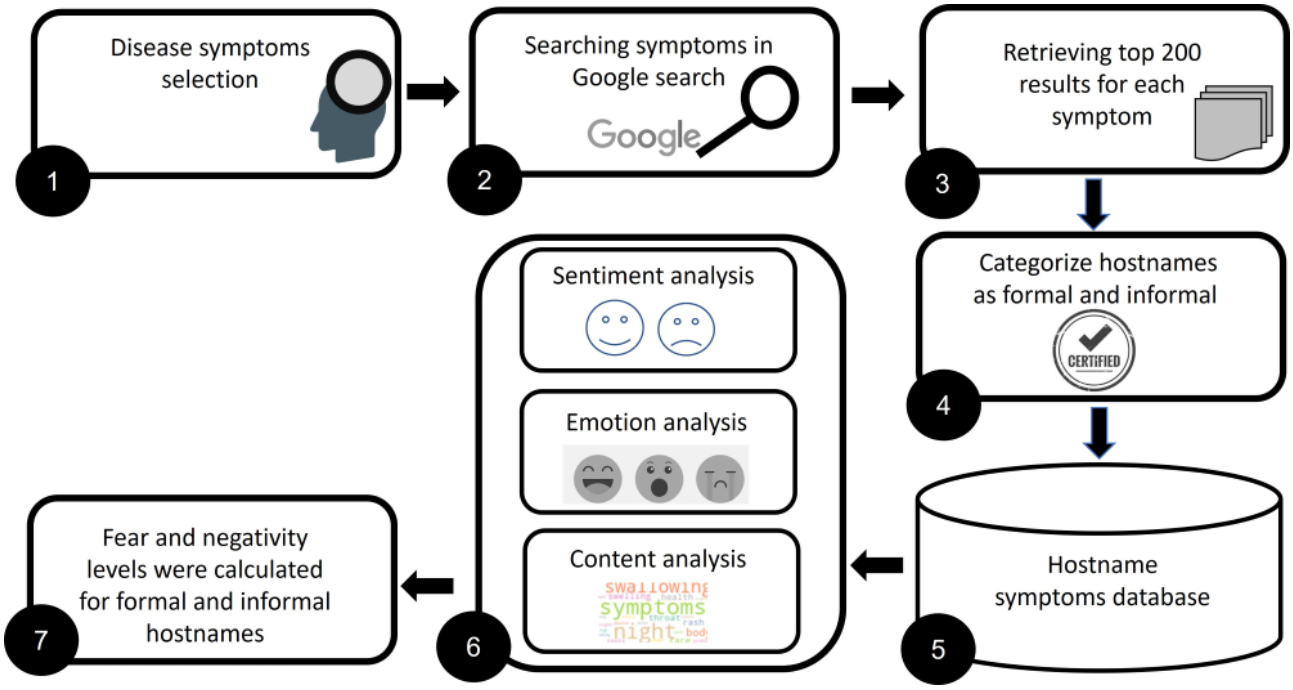

Identification of Family Disease Symptoms

Common diseases were first considered, including infectious diseases, cancer, and cardiovascular diseases. Symptoms for selected diseases were identified using existing literature regarding common symptoms associated with these diseases [8-10,45,46]. The list of symptoms is shown in Table 1.

XSL•FO
RenderX

Table 1. Symptoms related to each disease family.

Diseases family	Related symptoms
Infectious	• Fever • Headache • Muscle strain • Cough • Shortness of breath • Weakness
Cancer	• Bleeding • Cough • Chronic cough • Itch • Pain • Back pain • Rash • Swelling • Recurrent pneumonia • Diarrhea • Abdominal pain • Rectal bleeding • Blood in urine • Difficulty or pain while swallowing solid food • Vomiting • Choking on food • Enlarged lymph nodes • Night sweats • Breast tenderness • Pain or burning sensation • Breast warmth or redness • Breast lump • Breast pain or burning • Weakness
Cardiovascular	• Chest pain • Chest pressure • Chest tightness • Shortness of breath

Data Collection

We conducted a web search to retrieve websites containing symptoms of selected diseases, including infectious diseases, cancer diseases, and cardiovascular diseases. In order to search the websites, we used Google Web Search using an application programming interface provided by the RapidAPI website.

We conducted a search for each symptom and added the top 200 search results to our website's database. Our database is a comma-separated values file containing the entire website content. We obtained the hostname for each website using the application programming interface. In this study, the term website refers to the hostname of a specific website.

We used self-written Python code (version 3.9; Python Software Foundation) to decompose each website's content into sentences. Since long text contains multiple emotions [47], it was convenient to divide the content into sentences. We applied standard text preprocessing techniques to each content, such as removing line breaks, nonalphabetic words, stop words, and hyperlinks using natural language toolkit (NLTK, version 3.8.1).

Ethical Considerations

The original data collection was approved by our institutional review board. This analysis received approval from The Emek Yezreel College Ethical Review Board (approval number 2024-74 YVC EMEK). The research did not involve human subjects (the research data included only websites). Therefore, informed consent was not applicable, privacy and confidentiality of the data were not applicable, and compensation details were not applicable.

The Classification of Website as Formal or Informal

The received websites were manually categorized into 2 categories: formal websites and informal websites. A clear and consistent set of criteria based on expert domain knowledge guided the categorization process.

The criteria for formal health websites refer to those operated by an official government ministry or organization or authority, such as the ministry of health, the Centers for Disease Control and Prevention, the World Health Organization, and official websites of established health care facilities such as hospitals and academic medical centers (not including private clinics). Websites that did not meet these criteria were categorized as

XSL•FO
RenderX

informal websites. Among the informal websites were blogs, news websites, entertainment websites, and so forth.

Content Analysis

To analyze the content of each family disease, we calculated the average term frequency—inverse document frequency (TF-IDF). The TF-IDF measures the importance of words within each family disease. A term frequency (TF) measures how frequently a particular term appears in the content of a website. It is calculated by dividing the number of times a term appears in a website's content by the number of terms found across all family disease websites. Inverse document frequency (IDF) measures the importance of a term in the entire data set of family disease. It is calculated as the logarithm of the number of websites belonging to the family disease divided by the number of websites containing the term. The TF-IDF for each term appearing on a website in family disease has been averaged across all websites of family disease to determine the average TF-IDF. The average TF-IDF measures the overall importance of terms in family disease, where higher values indicate greater significance.

Emotion Analysis Using 3 Models

For the emotion classification task, we selected 2 transformer models from the Hugging Face Transformers repository. The models are distilled versions of state-of-the-art models of RoBERTa [17] and BERT [18], which means that they have been compressed and made smaller in size while maintaining most of the original model's performance. These 2 models are the most downloaded among the distilled models for emotion detection. Based on diverse text classification data sets, these models have been carefully fine-tuned to capture and leverage contextual information, enabling them to effectively classify text.

1. j-hartmann/emotion-english-distilroberta-base: the model is a fine-tuned checkpoint of DistilRoBERTa-base. The model is available in the HuggingFace repository as "j-hartmann/emotion-english-distilroberta-base." Using 6 diverse data sets, the model predicts Ekman's 6 basic emotions, plus a neutral class: anger, disgust, joy, fear, sadness, surprise, and neutral [48].
2. bhadresh-savani/distilbert-base-uncased-emotion: we used a pretrained model for emotion detection. It is available on the HuggingFace repository under the name "bhadresh-savani/distilbert-base-uncased-emotion." This model [18] is a distilled version of the BERT base model, fine-tuned on the emotion data set with an accuracy of 93.8% [49] of English Twitter messages with 8 basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

Several state-of-the-art transformer models such as RoBERTa [17], Distilbert [18], and ALBERT [19] were trained on an emotion data set of a web-based health community, annotated with 8 fine-grained emotions [29]. This data set allows us to select the best model based on accuracy and F_1-scores that fit most commonly for web-based health-related information and then use the model for estimating fear emotions arising from text of new health-related data. The data set contains 5389 sentences annotated with fear or unfear emotions equally. We used these labels to measure the accuracy and F_1-score of each model in predicting the fear emotion and choose the model with the best performance. distillbert-based-uncased was the best performance model:

1. distillbert-medical-fear-emotion: the model is available on HuggingFace repository under the name distillbert-based-uncased (details in the "Results" section). We call our fine-tuned model distillbert-medical-fear-emotion.

Fear Level

Each sentence was classified into confidence score using the 2 emotion models (bhadresh-savani/distilbert-base-uncased-emotion and distillbert-medical-fear-emotion). The confidence score represents the probability that the model is confident that a given input belongs to a particular class, in our case, fear. The confidence score ranges from 0 to 1, the higher the score, the more confident it is that the input is related to fear. For example, a sentence classified with a score of 0.70 means that the likelihood that this sentence belongs to fear is 70%. From now on, we refer to the confidence score as the fear level. The fear level was determined using the 2 transformer models to determine consistency. For both models, we calculated the study independent variable, website level of fear.

We calculated the level of fear as follows:

1. Each website's text content was analyzed at the sentence level.
2. For each sentence, a pretrained language model was used to estimate its probability of expressing fear. The confidence score ranges from 0 to 1, with higher values indicating a higher level of fear.
3. We calculated the level of fear for a particular website based on the mean of the fear confidence scores across all sentences on that website.
4. We categorized the websites into 2 categories: formal and informal.
5. The final level of fear variable represents the average of the fear levels calculated for all websites within each group (formal or informal).

The model of j-hartmann/emotion-english-distilroberta-base was used to examine the distribution of emotions (anger, disgust, joy, fear, sadness, surprise, and neutral) for formal and informal websites.

Sentiment Analysis Using SiEBERT Model

For the purpose of sentiment analysis of each website content, we used the SiEBERT model [50]. SiEBERT model is a fine-tuned checkpoint of RoBERTa-large [17]. It is available on HuggingFace repository under the name siebert/sentiment-roberta-large-english. It provides reliable binary sentiment analysis for English language texts. It predicts either positive or negative sentiment for each instance.

XSL•FO
RenderX

Negativity Level

In the same way, as we did with the emotion analysis models, SiEBERT classified each sentence according to its likelihood of being negative.

We calculated the negativity level as follows:

1. The text content of each website was analyzed at the sentence level.
2. A pretrained language model was used to estimate the probability of each sentence expressing negativity. There is a range of confidence scores from 0 to 1, with higher values indicating a greater level of negativity.
3. We calculated the level of negativity for a particular website based on the mean of the negativity confidence scores across all sentences on that website.
4. The websites were categorized into 2 categories: formal and informal.
5. The final level of negativity variable represents the average level of negativity calculated for all websites within each group (formal or informal).

Statistical Analysis

Prior to calculating the averages of fear and negativity in each site type, we filtered repeated results to avoid duplication of results across different websites and content types.

We conducted a 2-tailed *t* test to evaluate the difference between the average fear detected on formal health-related websites and the average fear detected on informal health-related websites. For each result, $P<.05$ was considered statistically significant. SPSS (version 28; IBM Corp) software was used for statistical analyses.

Results

Description of the Created Database

After searching in Google Search all the selected symptoms, we received a database containing a total of 1448 unique websites. A total of 534 websites were classified as formal websites and a total of 914 websites were classified as informal websites. The data set has been cleaned by removing results containing unrelated content or errors caused by any restrictions from the websites. A total of 593 results were removed. In total, for the formal websites, there were 1820 pages containing 30,701 sentences, while for the informal websites, there were 1494 pages containing 25,961 sentences.

As a result of reducing duplicate results in which the same website returned with the same content in different search results, we were left with 1461 formal websites and 1294 informal websites.

Words Representing Each Disease Family

The top 1000 significant words from each disease family retrieved from the content of the websites are shown in Multimedia Appendix 1. In order to determine the most significant words for each disease family, the average TF-IDF was calculated separately for each disease family. It can be seen that the most significant words for cancer are breast, cough, and pain. In addition, the most significant words in infectious diseases are breath, muscle, fever, and shortness. The most significant words in cardiovascular diseases are heart, chest, and pain.

The Results of Bhadresh-Savani/Distilbert-Base-Uncased-Emotion Model for Classifying the Level of Fear

In the Methods section, the level of fear was defined. For this part, we used the pretrained model bhadresh-savani/distilbert-base-uncased-emotion. According to Multimedia Appendix 2, the fear level in formal websites is higher than that in informal websites based on this emotion detection model.

We used the *t* test to determine whether there was a significant difference between the fear level in formal and informal websites. The degree of freedom was 2753.

The results of the *t* test indicate a statistically significant difference between formal and informal websites in terms of the fear level in this model (Table 2 presents the *t* test table).

Table 2. The *t* test results of the comparison between formal and informal websites in terms of fear and negativity level.

The model	Groups mean		Groups SDs		*t* test values (*df*)	*P* value
Fear level						
bhadresh-savani/distilbert-base-uncased-emotion	0.388	0.366	0.177	0.168	3.331 (2753)	.001
distillbert-medical-fear-emotion	0.584	0.569	0.121	0.134	3.040 (2753)	.002
Negativity level						
SiEBERT	0.657	0.636	0.211	0.201	2.726 (2753)	.006

The Results of Distillbert-Medical-Fear-Emotion Model for Classifying the Level of Fear

The emotion data set from an web-based health community was used to train several transformer models [29]. As a result of the data set, we can train existing models that estimate fear levels arising from health-related information and find the best model based on accuracy and F_1-scores. Eighty percent of the data set was used for training, 10% for validation, and 10% for testing.

The distillbert-based-uncased model provided the highest performance with accuracy on the validation set of 0.944 and an F_1-score of 0.943. On the testing set, the model achieved an accuracy of 74.4%. There was 75% intercoder agreement on the same task with human annotation in the study by Sosea and Caragea [29]. We call our fine-tuned model

XSL•FO
RenderX

distillbert-medical-fear-emotion. Table 3 shows the accuracy and F_1-scores of transformer models on the testing set.

Table 3. Transformer model accuracy and F1-scores on the test set.

Models	Description	Accuracy	F_1-scores
distilbert-base-uncased	Distilbert is a distilled smaller and faster version [18] of the BERT base model.	0.744	0.771
ClinicalBERT	Bio_ClinicalBERT is a model trained on data from the Beth Israel Hospital's ICU patients [51].	0.729	0.762
bert-base-uncased	Pretrained model on English language using a masked language modeling objective [15].	0.736	0.757
roberta-large	RoBERTa is a transformers model pretrained on a large corpus of English data in a self-supervised fashion [17].	0.736	0.765
albert-base-v2	ALBERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion [19].	0.738	0.757

Our best model, distillbert-medical-fear-emotion, was then used to automatically classify each sentence by its fear level. For both formal and informal websites, fear levels are presented. Multimedia Appendix 3 shows the model result. It can be seen that formal websites have a higher fear level.

Another *t* test was performed and revealed a statistically significant difference between formal and informal websites in terms of the fear level of distillbert-medical-fear-emotion (Table 2 presents the *t* test table).

According to the results, both models (the bhadresh-savani/distilbert-base-uncased-emotion model and the distillbert-medical-fear-emotion model) were consistent in their results when comparing formal and informal websites.

The Results of the SiEBERT for Classifying the Level of Negativity

In the Methods section, the level of negativity was defined. Using the SiEBERT model, we analyzed the level of negative sentiment arising from formal and informal websites. Multimedia Appendix 4 shows that informal websites also exhibit a higher degree of negativity in the sentiment analysis. We used the *t* test to determine whether there was a significant difference between formal and informal websites. Our results indicate that there is a statistically significant difference between formal and informal websites in terms of negative sentiment reflected in the text (Table 2 presents the *t* test table).

Figure 2 presents the negativity level by disease family, including cancer, cardiovascular diseases, and infectious diseases. It can be observed that the negativity level in texts related to infectious diseases is higher than that in texts related to cancer and cardiovascular diseases.

XSL•FO
RenderX

Figure 2. Negativity level by disease family.

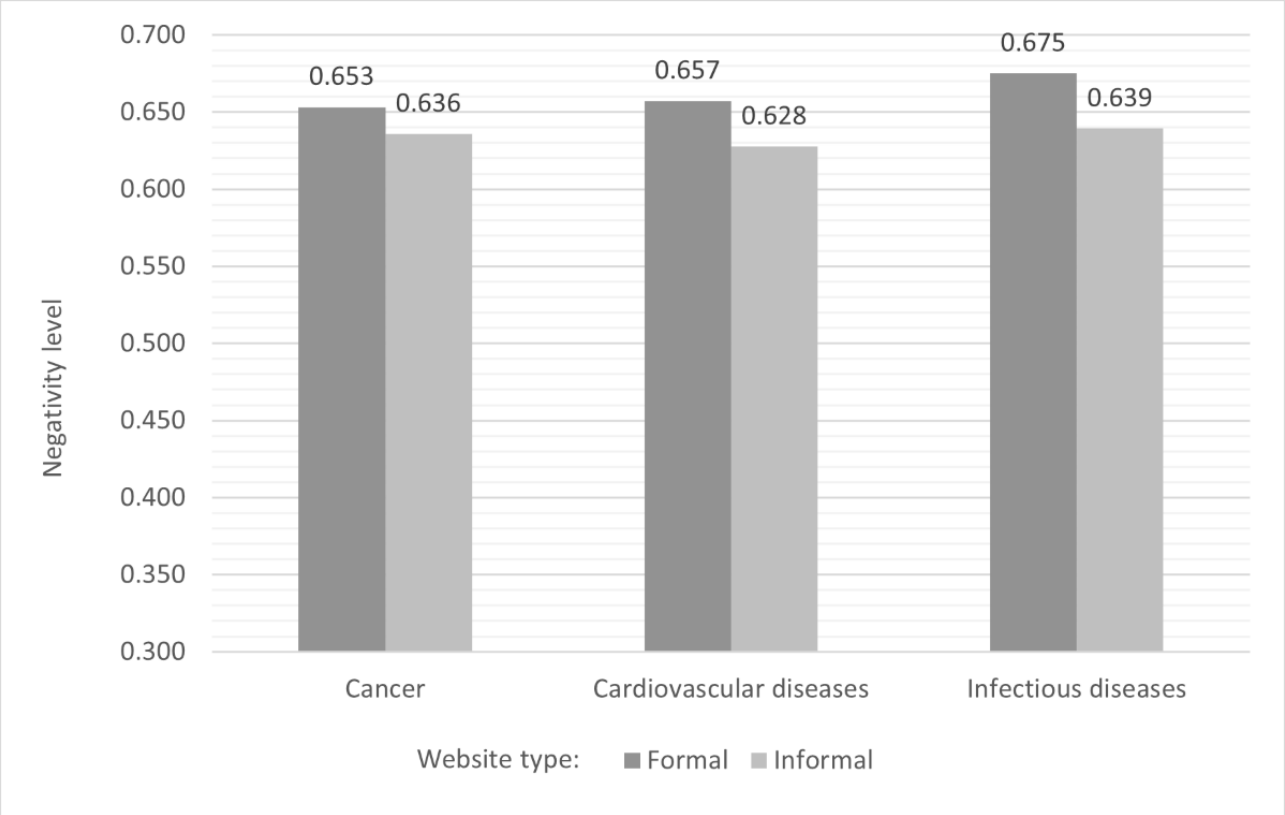

The Results of the J-Hartmann/Emotion-English-Distilroberta-Base for Emotion Detection

As part of the emotion analysis process, we ran the j-hartmann/emotion-english-distilroberta-base emotion analysis model across the website texts to gain a deeper understanding of the emotions expressed in these texts.

In Figure 3, we show the distribution of emotions across formal and informal websites. As can be seen in both cases, neutral emotion dominates, followed by fear, sadness, and disgust, indicating that both types of websites convey mostly negative emotions. A comparison between formal and informal websites reveals that fear is higher in formal websites, as was shown in previous models, but anger and sadness are also higher in formal websites.

Figure 3. Emotions (fear, surprise, sadness, anger, disgust, joy, and neutral) associated with formal and informal websites.

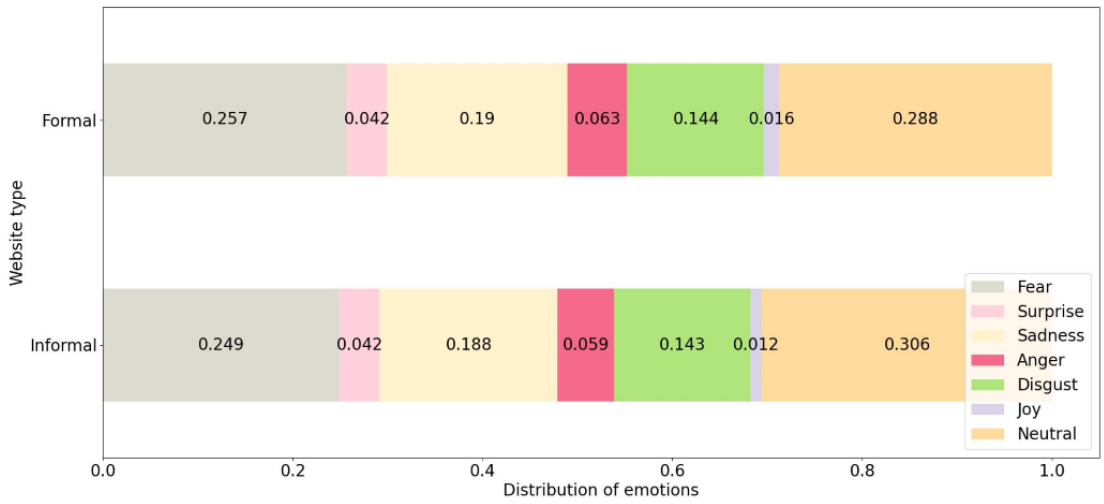

XSL•FO
RenderX

Discussion

Principal Findings

As part of our results, we observed that websites related to infectious diseases induced a higher level of negativity than websites related to cancer.

A possible explanation for this finding is that cancer is widely perceived by the public as a severe disease, and, as such, the public is more conformed with formal guidelines and recommendations regarding cancer prevention and treatment, and no fearful tactics and messages are required to achieve compliance. On contrary, the severity of infection diseases is underestimated by large parts of the public. Moreover, antivaxxers spread misinformation regarding various vaccines that are crucial to primary prevention of infectious diseases, including influenza, COVID-19, measles, smallpox, and more, and may prevent severe outcomes and mortality.

As a result, health professionals may include frightening text related to infectious diseases, in order to increase the public compliance with formal recommendations. Moreover, the recently emerging outbreaks of some infectious diseases may have been resulted in a more stressful and fear-inducing tone in formal messages.

Efforts may be made to obtain a balance between communicating diseases' severity and prevention measures and avoiding excessive scare tactics. This may be achieved by combining relevant facts with appropriate emotional appeals. Such a balance may increase the public awareness and compliance.

Furthermore, our results demonstrate that formal websites introducing health-related information induce more negative sentiments and fear emotions than informal websites.

A possible explanation for the higher negative sentiments and fear emotions found on formal websites could be that formal websites are more precise and provide more correct information on health-related issues [52]. As such, they may contain more clinical and scientific terminology regarding the searched medical issue than informal websites. Such details may include disease symptoms, the diagnosis process, prognosis, side effects, risks, possible treatments, and worst-case scenarios. These details may increase the negativity and fear levels induced by their provided text [52]. A formal website may also be associated with higher levels of negative sentiment and fear emotions since it is managed by government and health authorities, which may use fear appeals to motivate behavior changes such as vaccinations. These tactics may increase anxiety if they are not carefully framed [53].

On the other hand, informal websites are less informative and may contain inaccurate information and misinformation about health-related issues. For example, the claim that consuming apricot seeds will cure cancer is a misinformative non–evidence-based claim that can be found in some informal websites [54]. As such, informal websites may have decreased negativity and fear level in their provided texts, and, by that, they may appear more relaxing to the reader.

The public seeking for health-related information uses formal as well as informal websites. Formal health-related websites contain accurate, evidence-based information regarding different diseases, and it is desirable that the public will use formal health-related information [52,54]. On the other hand, informal health-related websites may include misinformation and may lead the public to erroneous decisions [54]. For example, misinformation regarding the measles, mumps, and rubella vaccine led to some severe measles outbreaks in some European countries and the United States [54]. However, the public tends to choose information sources that are not stressful and introduces the information in a manner that does not increase the reader's stress [1,55,56]. Therefore, the public may prefer to retrieve information from informal health-related websites, which, according to our results, induces less stress, regardless of misinformation which may exist.

This may lead to undesirable public health consequences. Selective focus on a specific type of medical content that is less stressful or anxiety-inducing can result in biased search strategies when retrieving health information [57]. Biased health information retrieval strategies can negatively affect public health and decision-making, resulting in inaccurate, incomplete, or misleading information [58,59].

Confirmation bias is a common bias when users search for information online. A confirmation bias is defined as the unconscious inclination to favor information that supports one's own viewpoint or belief [60]. As a result of confirmation bias, users are more likely to accept misleading information since it is consistent with their previous preconceptions [53]. Confirmation bias can be associated with the tendency to prefer positive information, in which users do not only seek information that confirms their existing beliefs but also seek to avoid negative information in order to maintain a positive emotional state.

In light of our recommendation to reduce levels of negativity and fear in the texts, it is important to balance fears raised by the text, while also providing accurate and comprehensible information to reduce the risk of biased information.

It is essential that health information is provided in a comprehensive manner [53], and some topics require a more serious tone of voice. For example, a serious tone may be necessary when discussing the full risks associated with a health condition, but a positive tone may be more effective in encouraging preventive health behaviors, such as vaccines. Therefore, the strategy for delivering information should include accurate, complete, and comprehensible information, as well as a focus on enhancing the positivity of the information.

The COVID-19 pandemic demonstrated that transmitting clear, precise, and evidence-based information to be used by the general public may be used as a strategy for controlling the disease, and it is one of the challenges of formal health organizations [55]. During the COVID-19 pandemic, an abundance of contradicting information existed, including misinformation regarding the disease severity and the required measures, including the vaccines that were developed during the outbreak. A study that examined the differences between formal and informal websites with regard to COVID-19

XSL•FO
RenderX

prevention measures found significant differences regarding some essential measures such as wearing masks [52]. Another study found an association between the source of information regarding the COVID-19 vaccines (formal or informal) and the willingness to receive the vaccine [56]. Hence, one of the challenges of public health professionals while facing the COVID-19 pandemic was to communicate challenging messages that build public trust [61]. Moreover, the COVID-19 pandemic was considered a unique and exceptional event worldwide [62]. As such, it created a new normal that we need to learn to live with [62]. One of the implications of living with the new normal is understanding that public health exists within a social context. Embracing public health as a social concept is essential for dealing with COVID-19, as well as with other global crises, such as climate change [62]. Therefore, a challenge for living with the new normal is achieving solidarity in health and public health and, by that, confronting cross-national and cross-cultural risks [62]. The use of positive texts by formal health-related websites, which can be achieved by reducing the negativity and fear levels in the texts, may increase the credibility of formal health websites, perceived by the public, and increase their usage by the general public, as well as the public's trust, engagement, and compliance with the measures during the pandemic. This may contribute to the solidarity, which is essential to living with the new normal.

Encouraging public professionals and public health organizations to use NLP tools before publishing health-related information, in order to identify negativity and fear levels in the text and reduce them, may be helpful to achieve a more positive and less stressful text to be transmitted to the public and may increase the public compliance to suggested measures.

Limitations

This study has some limitations. Using NLP tools as transformer models for estimating negativity and fear in the text comes with certain limitations. While these models have revolutionized the field of NLP and have accomplished remarkable feats in various tasks, they still have difficulty to capture the nuances of negative or fearful sentiments in the same manner as a human being. In addition, fear is a subjective feeling and may vary between people. Nevertheless, we do believe that our algorithm catches texts that appear negative and fearful to the majority of the population, since it is based on words that are likely to be perceived as negative and fearful by a large portion of the population, such as death, lethal, complications, and more. Further research is required to examine the correlation between negativity and fear levels received by our algorithm and negativity and fear levels observed by human participants.

Another limitation is that this study focused on 3 specific diseases: cancer, cardiovascular, and infectious diseases. We focused on these diseases since they are leading causes of death and as such, people are likely to search on the web for information about their symptoms [8]. We believe that focusing on these diseases provided a reasonable starting point for this initial exploratory study. Further research is required to expand the scope of this research by examining other diseases.

Furthermore, in this study, we examined the emotions and sentiments expressed in the text; however, people may have additional priorities or preferences when consuming web-based health information. Health information consumption should consider a broad range of user needs and preferences, such as how clear, easy to understand, and complete the information is. The differences between formal and informal health-related websites should be further studied to gain a deeper understanding of the diverse needs and priorities of health information users when accessing health information, for example, to examine the differences between the clarity of the information on the 2 types of websites.

Conclusions

We believe that improving the positivity of messages provided by formal health-related websites by reducing the negativity and fear levels may increase the public use of these websites, as well as the public's trust in them, and the public's compliance with recommended measures. It can be achieved by applying NLP tools to texts published by official public health websites. Public health decision makers may use existing NLP models, which include appropriate graphical user interface to assess the negativity and fear levels in a text, before publishing it to the public. An example of such a model can be found in SiEBERT—English-Language Sentiment Classification [63]. However, training public health professionals to use such tools and to reduce the negativity and fear levels in their published texts is essential to achieving this goal.

Data Availability

The data sets generated and analyzed during this study are not publicly available because they contain sensitive information that could compromise the privacy and consent of the websites. The transformed data are, however, available upon reasonable request from the authors.

Authors' Contributions

APV led the conceptualization (equal), software (lead), methodology (equal), data curation (lead), formal analysis (lead), visualization (lead), and writing (equal). AM led the conceptualization (equal), methodology (equal), visualization (equal), writing (equal), data curation (equal), and formal analysis (equal).

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Top 1000 words for each disease.
[XLSX File (Microsoft Excel File), 84 KB - jmir_v26i1e55151_app1.xlsx]

Multimedia Appendix 2
Fear level in formal and informal websites based on distilbert-base-uncased-emotion.
[PNG File , 40 KB - jmir_v26i1e55151_app2.png]

Multimedia Appendix 3
Fear level in formal and informal websites based on distilbert-medical-fear-emotion.
[PNG File , 42 KB - jmir_v26i1e55151_app3.png]

Multimedia Appendix 4
Negativity level in formal and informal websites.
[PNG File , 45 KB - jmir_v26i1e55151_app4.png]

References

1. White RW, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Trans Inf Syst 2009;27(4):1-37. [doi: 10.1145/1629096.1629101]
2. Kubb C, Foran HM. Online health information seeking by parents for their children: systematic review and agenda for further research. J Med Internet Res 2020;22(8):e19985 [FREE Full text] [doi: 10.2196/19985] [Medline: 32840484]
3. White RW, Horvitz E. Web to world: predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. AMIA Annu Symp Proc 2010;2010:882-886 [FREE Full text] [Medline: 21347105]
4. Peng RX. How online searches fuel health anxiety: investigating the link between health-related searches, health anxiety, and future intention. Comput Hum Behav 2022;136:107384. [doi: 10.1016/j.chb.2022.107384]
5. Lagoe C, Atkin D. Health anxiety in the digital age: an exploration of psychological determinants of online health information seeking. Comput Hum Behav 2015;52:484-491. [doi: 10.1016/j.chb.2015.06.003]
6. Baumgartner SE, Hartmann T. The role of health anxiety in online health information search. Cyberpsychol Behav Soc Netw 2011;14(10):613-618. [doi: 10.1089/cyber.2010.0425] [Medline: 21548797]
7. Benigeri M, Pluye P. Shortcomings of health information on the internet. Health Promot Int 2003;18(4):381-386. [doi: 10.1093/heapro/dag409] [Medline: 14695369]
8. Schneider PP, van Gool CJ, Spreeuwenberg P, Hooiveld M, Donker GA, Barnett DJ, et al. Using web search queries to monitor influenza-like illness: an exploratory retrospective analysis, Netherlands, 2017/18 influenza season. Euro Surveill 2020;25(21):1900221 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.21.1900221] [Medline: 32489174]
9. Abroms LC, Yom-Tov E. The role of information boxes in search engine results for symptom searches: analysis of archival data. JMIR Infodemiology 2022;2(2):e37286 [FREE Full text] [doi: 10.2196/37286] [Medline: 37113445]
10. Hochberg I, Allon R, Yom-Tov E. Assessment of the frequency of online searches for symptoms before diagnosis: analysis of archival data. J Med Internet Res 2020;22(3):e15065 [FREE Full text] [doi: 10.2196/15065] [Medline: 32141835]
11. Alm C, Roth D, Sproat R. Emotions from text: machine learning for text-based emotion prediction. 2005 Presented at: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing; 2005 Oct 10; Columbia, Canada p. 579-586. [doi: 10.3115/1220575.1220648]
12. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 2014;5(4):1093-1113. [doi: 10.1016/j.asej.2014.04.011]
13. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 2021;11(1):81 [FREE Full text] [doi: 10.1007/s13278-021-00776-6] [Medline: 34484462]
14. Ekman P. Basic emotions. In: Dalgleish T, Power MJ, editors. Handbook of Cognition and Emotion. New York, NY: Wiley; 1999:16.
15. Devlin J, Chang M, Lee K. Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv ;1810:04805 Preprint posted online October 2018 [FREE Full text]
16. Yang Z, Dai Z, Yang Y. Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst 2019;32.
17. Liu Y, Ott M, Goyal N. Roberta: a robustly optimized bert pretraining approach. ArXiv ;1907:11692 Preprint posted online July 2019 [FREE Full text]
18. SanhV, Debut L, Chaumond J. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv ;1910:1108 Preprint posted online October 2019 [FREE Full text]
19. Lan Z, Chen M, Goodman S. Albert: a lite bert for self-supervised learning of language representations. ArXiv ;1909:11942 Preprint posted online September 2019 [FREE Full text]

XSL•FO
RenderX

20. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 2020;8(1):e16023 [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
21. Portier K, Greer GE, Rokach L, Ofek N, Wang Y, Biyani P, et al. Understanding topics and sentiment in an online cancer survivor community. J Natl Cancer Inst Monogr 2013;2013(47):195-198. [doi: 10.1093/jncimonographs/lgt025] [Medline: 24395991]
22. Cabling ML, Turner JW, Hurtado-de-Mendoza A, Zhang Y, Jiang X, Drago F, et al. Sentiment analysis of an online breast cancer support group: communicating about tamoxifen. Health Commun 2018;33(9):1158-1165 [FREE Full text] [doi: 10.1080/10410236.2017.1339370] [Medline: 28678549]
23. Carrillo-de-Albornoz J, Rodríguez Vidal J, Plaza L. Feature engineering for sentiment analysis in e-health forums. PLoS One 2018;13(11):e0207996 [FREE Full text] [doi: 10.1371/journal.pone.0207996] [Medline: 30496232]
24. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
25. Ali T, Schramm D, Sokolova M. Can I hear you? Sentiment analysis on medical forums. 2013 Presented at: Proceedings of the Sixth International Joint Conference on Natural Language Processing; 2024 June 20; Nagoya, Japan p. 667-673.
26. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc 2014;21(6):1098-1103 [FREE Full text] [doi: 10.1136/amiajnl-2014-002711] [Medline: 24918109]
27. Hopper AM, Uriyo M. Using sentiment analysis to review patient satisfaction data located on the internet. J Health Organ Manag 2015;29(2):221-233 [FREE Full text] [doi: 10.1108/JHOM-12-2011-0129] [Medline: 25800334]
28. Khanpour H, Caragea C. Fine-grained emotion detection in health-related online posts. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 10; Brussels, Belgium p. 1160-1166. [doi: 10.18653/v1/d18-1147]
29. Sosea T, Caragea C. Canceremo: a dataset for fine-grained emotion detection. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 10; Toronto, Canada p. 8892-8904. [doi: 10.18653/v1/2020.emnlp-main.715]
30. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
31. Oksanen A, Garcia D, Sirola A, Näsi M, Kaakinen M, Keipi T, et al. Pro-anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. J Med Internet Res 2015;17(11):e256 [FREE Full text] [doi: 10.2196/jmir.5007] [Medline: 26563678]
32. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Gningaye Kengni FL, et al. Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of Facebook and Twitter posts. JMIR Public Health Surveill 2017;3(3):e51 [FREE Full text] [doi: 10.2196/publichealth.7004] [Medline: 28793981]
33. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. J Med Internet Res 2018;20(12):e11817 [FREE Full text] [doi: 10.2196/11817] [Medline: 30522991]
34. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards Covid-19 vaccines with Twitter data in the United States: spatiotemporal perspective. J Med Internet Res 2021;23(9):e30854 [FREE Full text] [doi: 10.2196/30854] [Medline: 34346888]
35. Kwok SWH, Vadde SK, Wang G. Tweet topics and sentiments relating to Covid-19 vaccination among Australian Twitter users: machine learning analysis. J Med Internet Res 2021;23(5):e26953 [FREE Full text] [doi: 10.2196/26953] [Medline: 33886492]
36. Basiri ME, Nemati S, Abdar M, Asadi S, Acharrya UR. A novel fusion-based deep learning model for sentiment analysis of Covid-19 tweets. Knowl Based Syst 2021;228:107242 [FREE Full text] [doi: 10.1016/j.knosys.2021.107242] [Medline: 36570870]
37. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS One 2021;16(2):e0245909 [FREE Full text] [doi: 10.1371/journal.pone.0245909] [Medline: 33630869]
38. Aygun I, Kaya B, Kaya M. Aspect based Twitter sentiment analysis on vaccination and vaccine types in Covid-19 pandemic with deep learning. IEEE J Biomed Health Inform 2022;26(5):2360-2369. [doi: 10.1109/JBHI.2021.3133103] [Medline: 34874877]
39. Alam KN, Khan MS, Dhruba AR, Khan MM, Al-Amri JF, Masud M, et al. Deep learning-based sentiment analysis of Covid-19 vaccination responses from Twitter data. Comput Math Methods Med 2021;2021:4321131 [FREE Full text] [doi: 10.1155/2021/4321131] [Medline: 34899965]
40. Bokaee Nezhad Z, Deihimi MA. Twitter sentiment analysis from Iran about Covid 19 vaccine. Diabetes Metab Syndr 2022;16(1):102367 [FREE Full text] [doi: 10.1016/j.dsx.2021.102367] [Medline: 34933273]

XSL•FO
RenderX

41. Zulfiker MS, Kabir N, Biswas AA, Zulfiker S, Uddin MS. Analyzing the public sentiment on Covid-19 vaccination in social media: Bangladesh context. Array (N Y) 2022;15:100204 [FREE Full text] [doi: 10.1016/j.array.2022.100204] [Medline: 35722449]
42. Elyashar A, Plochotnikov I, Cohen IC, Puzis R, Cohen O. The state of mind of health care professionals in light of the Covid-19 pandemic: text analysis study of Twitter discourses. J Med Internet Res 2021;23(10):e30217 [FREE Full text] [doi: 10.2196/30217] [Medline: 34550899]
43. Aduragba O, Yu J, Cristea A, Shi L. Detecting fine-grained emotions on social media during major disease outbreaks: health and well-being before and during the Covid-19 pandemic. AMIA Annu Symp Proc 2021;2021:187-196 [FREE Full text] [Medline: 35308991]
44. Christensen B, Laydon D, Chelkowski T, Jemielniak D, Vollmer M, Bhatt S, et al. Quantifying changes in vaccine coverage in mainstream media as a result of the Covid-19 outbreak: text mining study. JMIR Infodemiology 2022;2(2):e35121 [FREE Full text] [doi: 10.2196/35121] [Medline: 36348981]
45. Mueller J, Jay C, Harper S, Davies A, Vega J, Todd C. Web use for symptom appraisal of physical health conditions: a systematic review. J Med Internet Res 2017;19(6):e202 [FREE Full text] [doi: 10.2196/jmir.6755] [Medline: 28611017]
46. Akpan IJ, Aguolu OG, Kobara YM, Razavi R, Akpan AA, Shanker M. Association between what people learned about Covid-19 using web searches and their behavior toward public health guidelines: empirical infodemiology study. J Med Internet Res 2021;23(9):e28975 [FREE Full text] [doi: 10.2196/28975] [Medline: 34280117]
47. Hartmann J. emotion-english-distilroberta-base. 2022. URL: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base [accessed 2024-06-22]
48. Biyani P, Caragea C, Mitra P. Identifying emotional and informational support in online health communities. 2014 Presented at: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers; 2014 Aug 10; Dublin, Ireland p. 827-836.
49. Saravia E, Toby Liu HC, Huang YH, Wu J, Chen YS. Carer: contextualized affect representations for emotion recognition. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 10; Brussels, Belgium. [doi: 10.18653/v1/d18-1404]
50. Hartmann J, Heitmann M, Siebert C, Schamp C. More than a feeling: accuracy and application of sentiment analysis. Int J Res Mark 2023;40(1):75-87. [doi: 10.1016/j.ijresmar.2022.05.005]
51. Alsentzer E, Murphy J, Boag W. Publicly available clinical BERT embeddings. ArXiv ;1904:3323 Preprint posted online June 2019. [doi: 10.18653/v1/w19-1909]
52. Hernández-García I, Giménez-Júlvez T. Assessment of health information about Covid-19 prevention on the internet: infodemiological study. JMIR Public Health Surveill 2020;6(2):e18717 [FREE Full text] [doi: 10.2196/18717] [Medline: 32217507]
53. Ludolph R, Allam A, Schulz PJ. Manipulating Google's knowledge graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. J Med Internet Res 2016;18(6):e137 [FREE Full text] [doi: 10.2196/jmir.5430] [Medline: 27255736]
54. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
55. Abdekhoda M, Ranjbaran F, Sattari A. Information and information resources in Covid-19: awareness, control, and prevention. J Librarianship Inf Sci 2021;54(3):363-372. [doi: 10.1177/09610006211016519]
56. Yoda T, Suksatit B, Tokuda M, Katsuyama H. The relationship between sources of Covid-19 vaccine information and willingness to be vaccinated: an internet-based cross-sectional study in Japan. Vaccines (Basel) 2022;10(7):1041 [FREE Full text] [doi: 10.3390/vaccines10071041] [Medline: 35891205]
57. Russell-Rose T, Chamberlain J. Expert search strategies: the information retrieval practices of healthcare information professionals. JMIR Med Inform 2017;5(4):e33 [FREE Full text] [doi: 10.2196/medinform.7680] [Medline: 28970190]
58. Mansour RF, Fatouh AH. Measurement of bias in the contents of web search for health information retrieval. J Scientometric Res 2023;12(3):621-630. [doi: 10.5530/jscires.12.3.060]
59. Meppelink CS, Smit EG, Fransen ML, Diviani N. "I was right about vaccination": confirmation bias and health literacy in online health information seeking. J Health Commun 2019;24(2):129-140. [doi: 10.1080/10810730.2019.1583701] [Medline: 30895889]
60. Azzopardi L. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. 2021 Presented at: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval; 2021 Mar 14; United Kingdom p. 27-37. [doi: 10.1145/3406522.3446023]
61. Bashkin O, Otok R, Leighton L, Czabanowska K, Barach P, Davidovitch N, et al. Emerging lessons from the Covid-19 pandemic about the decisive competencies needed for the public health workforce: a qualitative study. Front Public Health 2022;10:990353 [FREE Full text] [doi: 10.3389/fpubh.2022.990353] [Medline: 36117595]
62. Boas H, Davidovitch N. Into the "new normal": the ethical and analytical challenge facing public health post-Covid-19. Int J Environ Res Public Health 2022;19(14):8385 [FREE Full text] [doi: 10.3390/ijerph19148385] [Medline: 35886236]
63. SiEBERT - English-Language Sentiment Classification. URL: https://huggingface.co/siebert/sentiment-roberta-large-english [accessed 2024-06-22]

XSL•FO
RenderX

Abbreviations

NLP: natural language processing
TF-IDF: term frequency—inverse document frequency

Edited by S Ma, T Leung; submitted 04.12.23; peer-reviewed by E Vashishtha, C Abbatantuono, TAR Sure; comments to author 12.04.24; revised version received 19.05.24; accepted 07.06.24; published 09.08.24.

Please cite as:
Paradise Vit A, Magid A
Differences in Fear and Negativity Levels Between Formal and Informal Health-Related Websites: Analysis of Sentiments and Emotions
J Med Internet Res 2024;26:e55151
URL: https://www.jmir.org/2024/1/e55151
doi:10.2196/55151
PMID:39120928

©Abigail Paradise Vit, Avi Magid. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study

Yueye Wang[1,2*], Dr med; Xiaotong Han[1*], PhD, Prof Dr Med; Cong Li[1], Dr med; Lixia Luo[1], PhD, Prof Dr Med; Qiuxia Yin[1], MPH; Jian Zhang[1], MD; Guankai Peng[3], BSc; Danli Shi[2,4], PhD, Dr med; Mingguang He[2,4,5], PhD, Prof Dr Med

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China

[2]School of Optometry, The Hong Kong Polytechnic University, Kowloon, China (Hong Kong)

[3]Guangzhou Vision Tech Medical Technology Co, Ltd, Guangzhou, China

[4]Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Kowloon, China (Hong Kong)

[5]Centre for Eye and Vision Research, Hong Kong, China (Hong Kong)

[*]these authors contributed equally

Corresponding Author:
Mingguang He, PhD, Prof Dr Med
School of Optometry
The Hong Kong Polytechnic University
Hung Hom
Kowloon
China (Hong Kong)
Phone: 852 34002795
Email: mingguang.he@polyu.edu.hk

Abstract

Background: For medical artificial intelligence (AI) training and validation, human expert labels are considered the gold standard that represents the correct answers or desired outputs for a given data set. These labels serve as a reference or benchmark against which the model's predictions are compared.

Objective: This study aimed to assess the accuracy of a custom deep learning (DL) algorithm on classifying diabetic retinopathy (DR) and further demonstrate how label errors may contribute to this assessment in a nationwide DR-screening program.

Methods: Fundus photographs from the Lifeline Express, a nationwide DR-screening program, were analyzed to identify the presence of referable DR using both (1) manual grading by National Health Service England–certificated graders and (2) a DL-based DR-screening algorithm with validated good lab performance. To assess the accuracy of labels, a random sample of images with disagreement between the DL algorithm and the labels was adjudicated by ophthalmologists who were masked to the previous grading results. The error rates of labels in this sample were then used to correct the number of negative and positive cases in the entire data set, serving as postcorrection labels. The DL algorithm's performance was evaluated against both pre- and postcorrection labels.

Results: The analysis included 736,083 images from 237,824 participants. The DL algorithm exhibited a gap between the real-world performance and the lab-reported performance in this nationwide data set, with a sensitivity increase of 12.5% (from 79.6% to 92.5%, $P<.001$) and a specificity increase of 6.9% (from 91.6% to 98.5%, $P<.001$). In the random sample, 63.6% (560/880) of negative images and 5.2% (140/2710) of positive images were misclassified in the precorrection human labels. High myopia was the primary reason for misclassifying non-DR images as referable DR images, while laser spots were predominantly responsible for misclassified referable cases. The estimated label error rate for the entire data set was 1.2%. The label correction was estimated to bring about a 12.5% enhancement in the estimated sensitivity of the DL algorithm ($P<.001$).

Conclusions: Label errors based on human image grading, although in a small percentage, can significantly affect the performance evaluation of DL algorithms in real-world DR screening.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52506) doi:10.2196/52506

KEYWORDS

artificial intelligence; diabetic retinopathy; diabetes; real world; deep learning

Introduction

In recent years, the application of artificial intelligence (AI) in identifying ocular lesions and diseases has gained increasing popularity [1-4]. One prominent example is the application of deep learning (DL) in fundus image–based diabetic retinopathy (DR) screening [5-8]. Despite the widely reported high accuracy in existing studies, as well as the advantages in efficiency and cost-effectiveness, the real-world implementation of DL-based DR-screening models still faces challenges [9]. A major concern is the gap observed between the high laboratory performance and the unsatisfying real-world performance. For example, although most DL algorithms have achieved a sensitivity of over 90% in previous research, the reported sensitivity ranges from 55% to 85% in real-world tasks [4,7,10]. A better understanding of the reasons for the performance gap so as to enhance the real-world performance of DL-based models could help promote the implementation of AI on the ground, but related evidence is scarce.

In the development of medical AI, human expert labels usually serve as the gold standard to compare with and evaluate model results [11]. The accuracy of AI model–based tasks critically depends on high-quality labels. Assembling large data sets with human expert labels can be a huge task, and ensuring the quality of the gold standard remains an open problem [12]. In real-world image data sets, the gold standard is usually generated by trained graders who manually assign grades to the data [13-15]. However, human error is inevitable in this process, leading to problems, such as bias, subjective judgement, inadequate repeatability, and consistency. Studies have reported significant discrepancies in DR grading among human graders, with sensitivities ranging from approximately 60% to 90% [16]. As a result, using manual grading outputs as an objective reference standard can be problematic when the accuracy of manual grading is in question [11,17,18].

We hypothesized that errors in human labels may exist in the real world and exert a nonnegligible impact on real-world DL model performance. Thus, in this study, we included over 0.7 million fundus images from 0.2 million individuals from a nationwide DR-screening program, based on which we aimed to assess the gap between lab-reported and real-world performance of a previously validated DL-based DR-screening model, identify the existence of label errors, and assess the impact of those label errors on the observed performance gaps.

Methods

Ethical Considerations

The study adhered to the tenets of the Declaration of Helsinki and received approval from the Institutional Review Board of the Zhongshan Ophthalmic Centre (2023KYPJ108) and the Chinese Foundation for Lifeline Express. All images were fully anonymized. Written informed consent was obtained from all Lifeline Express participants. Given the retrospective nature of this study using anonymized images, the ethics review board exempted this study from additional informed consent.

Real-World Data Set

The pipeline of this study is depicted in Figure 1. The overall design of this study comprised 3 sections. First, more than 0.8 million images were collected from a real-world data set, and all images were included except those that failed quality control or lacked human labels. Second, fundus images were reviewed by National Health Service (NHS)–certificated human graders and manually classified as nonreferable DR (R0 and R1) or referable DR (R2, R3s, and R3a) according to NHS DR-screening guidelines. Next, a previously validated DL model was deployed for identifying referable DR, and the performance of the DL model was evaluated against human labels. Third, a random sample of images, comprising false-negative (FN) and false-positive (FP) subsets, was extracted for adjudication by 2 ophthalmologists. The details of the process are given next.

When adjudication yielded consistent results with DL instead of human labels, the errors of the labels were documented. The extracted sample's error rates were used to recalculate the number of positive and negative cases in the entire data set, as postcorrection labels. The diagnostic performance of the DL algorithm was reassessed against the postcorrection labels.

We included 865,152 color fundus images of participants who participated in a nationwide real-world DR screening program, the Lifeline Express DR Screening Program, between 2014 and 2019. For each participant, nonmydriatic, 45° field color fundus images were captured using locally available imaging devices (including AFC-230, NIDEK; Canon CR-DGi; FundusVue, Cystalvue) with at least 1 image centered on the macula or optic disc. The size of the images was 512 × 512 pixels, at a resolution of 72 pixels per inch. Images were subsequently sent to 5 different central grading centers (Peking Union Medical College Hospital, Beijing Tongren Hospital, Peking University Third Hospital, Joint Shantou International Eye Center, and Zhongshan Ophthalmic Center) and graded for DR by human graders who possessed English NHS-certified qualifications. The following images were excluded: (1) those that failed the quality control procedure and (2) those without human labels. Details regarding the quality control and DR-grading criteria can be found in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Schematic pipeline of the study design. DL: deep learning; DR: diabetic retinopathy; FN: false negative; FP: false positive; NHS: National Health Service; TN: true negative; TP true positive.

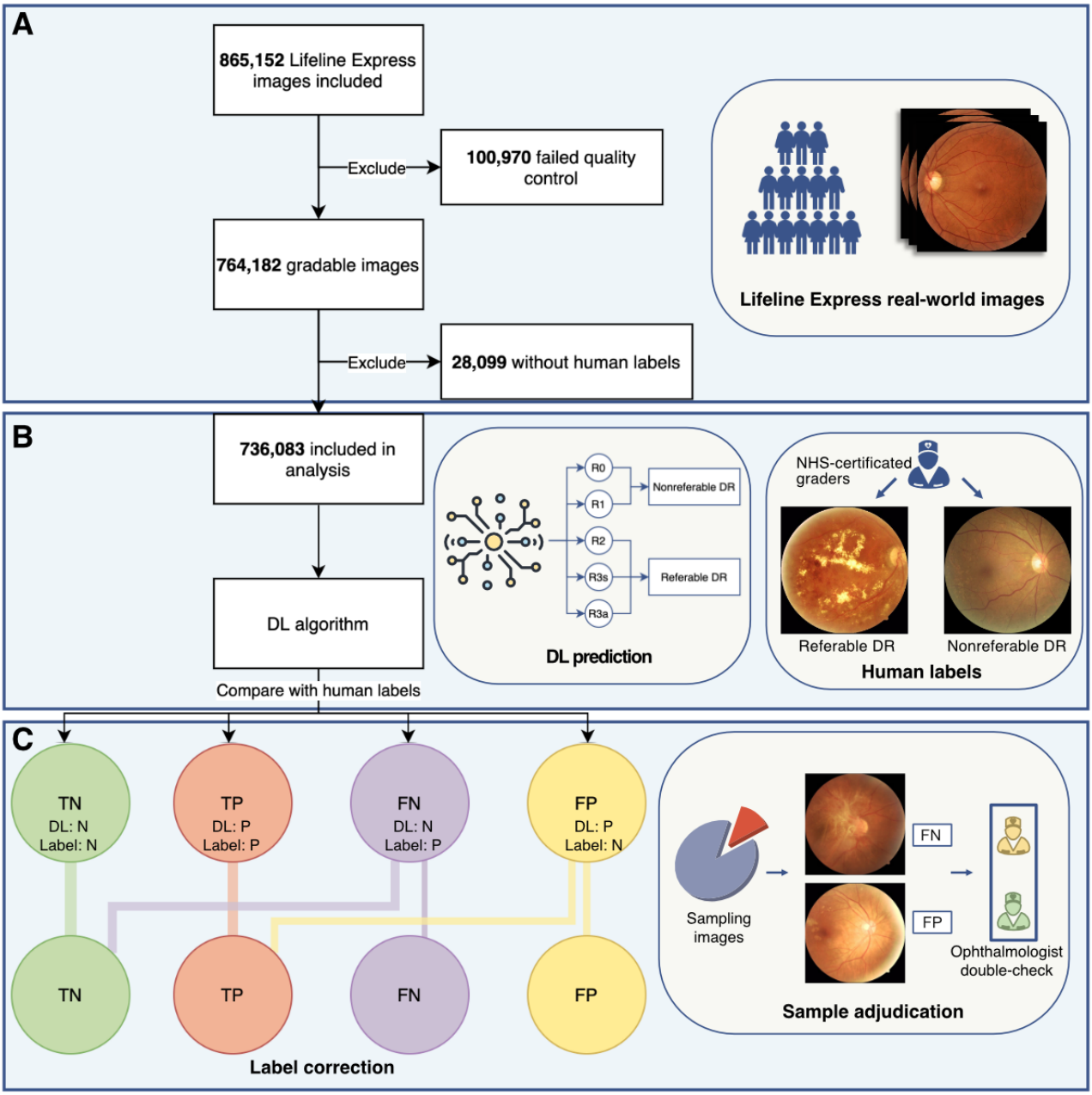

Grading Criteria

Based on NHS DR-screening guidelines, each fundus image was classified into 1 of 5 grades: R0 (no DR), R1 (background DR), R2 (preproliferative DR), R3s (static proliferative DR), or R3a (active proliferative DR). Referable DR was defined as preproliferative DR or worse (ie, R2, R3s, and R3a) and recorded as positive, whereas nonreferable DR (ie, R0 and R1) was recorded as negative. A participant was diagnosed as having referable DR if at least 1 of their gradable images was classified as positive.

The manual grading process involved the following stages. Initially, all images were assessed by a primary grader. Images classified as positive and a random sample of 15% of images classified as negative were then subjected to a second round of grading by another primary grader. Images with any discrepancy between the assessments of the 2 primary graders were referred to a secondary grader, whose conclusion served as a final determination. The resulting manual grading outputs formed human labels.

Development of the DR-Grading DL Algorithm

The DR-grading algorithm is based on a convolutional neural network. The algorithm was trained using 71,043 retinal images from 36 ophthalmic departments and validated using an additional 106,244 retinal images from 3 population-based studies. Image preprocessing involved scaling image pixel values within a range of 0-1 and subtracting the local-space average color. Images were resized to 299 × 299 pixels. We performed data augmentation techniques on training data. This involved randomly shifting each image horizontally by 0-3 pixels and rotating it by 90°, 180°, or 270°.

XSL•FO
RenderX

The original version of the algorithm has been reported previously [19]. The original DL algorithm achieved an area under the receiver operating characteristic curve (AUROC) of 0.955 (sensitivity: 92.5%, specificity: 98.5%) in validation for the detection of referable DR. The algorithm used in our study was retrained using the Xception architecture to improve model performance, while using the same training data set. Max pooling of each Inception module and global average pooling were used in the Xception architecture. For training, a minibatch gradient descent size of 32 was used. The learning rate adopted an inverse time decay schedule with an initial value of 0.001.

Adjudication and Human Label Correction

We adjudicated images that showed discrepancies between DL and human labels in referable DR classification, including FN (DL: negative, label: positive) and FP (DL: positive, label: negative). Adjudication was performed by 2 experienced ophthalmologists who were masked to the DL and human grading results; they used a back-to-back method in accordance with NHS DR-screening guidelines. Images with any discrepancy between the 2 ophthalmologists were sent to a senior ophthalmologist for reassessment, whose grading results were deemed as final.

Given the large number of images in this study, reviewing all images with discrepancies would have been impractical. Therefore, random samples were drawn for adjudication. To ascertain the representativeness of these random samples across the entire data set, a series of gradual samplings were conducted, varying in size from 100 to 500 FN images. Notably, a similar distribution of the estimated error rate was observed across all sample sizes (Figure S1 in Multimedia Appendix 1), suggesting that even a subset as small as 100 images can effectively represent the label distribution within the entire data set. Considering the total volume of FN and FP cases, we examined a 10% sample from FN cases and a 5% sample from FP cases, at both image and participant levels, aiming for a representative and manageable size of the sample. When adjudication yielded consistent results with DL instead of human labels, the label errors were documented in the FN and FP samples, respectively. Using the label error rates in the FN and FP samples, we recalculated the number of positive and negative cases in the entire data set, serving as postcorrection labels.

Statistical Analysis

We evaluated the performance of the DL algorithm using pre- and postcorrection labels as the gold standard. Primary evaluations metrics for diagnostic performance included accuracy, precision, sensitivity (recall), specificity, and the F_1-score. AUROC was calculated at the image level. These evaluations were conducted at both image and individual levels. Label error rates in adjudicated samples were calculated as follows:

where fnr and fpr represent label error rates in FN and FP samples, respectively. Label changes during label correction included:

$$pn = fnr \times FN \text{ cases},$$

$$np = fpr \times FP \text{ cases},$$

where pn is the number of labels changing from positive to negative, while np is the number of labels changing from negative to positive. The total positive and negative cases after label correction were calculated as follows:

$$\text{Total positive cases} = \text{Positive labels} + np - pn$$

$$\text{Total negative cases} = \text{Negative labels} + pn - np$$

Comparisons of diagnostic performance were conducted using chi-square tests. Adjudication was conducted on the open source LabelMe platform [20]. All statistical analyses were performed using R version 4.1.2 (Foundation for Statistical Computing). A 2-sided P value of <.05 was considered statistically significant.

Results

Precorrection Model Performance

Of the 865,152 images acquired from 251,535 participants, 736,083 (85.1%) images from 237,824 (94.5%) participants were included in this analysis. Table 1 shows the results of the DL algorithm and precorrection labels for detecting referable DR. At the image level, DL-based grading yielded 9339 (1.3%) FN, 57,886 (7.9%) FP, 632,457 (85.9%) true negative (TN), and 36,401 (5.0%) true positive (TP) images against precorrection labels. The DL algorithm yielded an AUROC of 0.927, a sensitivity (recall) of 79.6%, and a specificity of 91.6%, lower than the lab-reported performance values (AUROC: 0.955, sensitivity: 92.5%, specificity: 98.5%). The corresponding gaps in sensitivity and specificity were 12.9% and 6.9%, respectively. The precision and F_1-score at the image level were 38.6% and 52.0%, respectively. Evaluation at the individual level yielded 2062 (0.9%) FN, 33,878 (14.2%) FP, 186,648 (78.5%) TN, and 15,236 (6.4%) TP images. The sensitivity, specificity, and accuracy for identifying individuals with referable DR were 88.1%, 84.6%, and 84.9%, respectively. The precision and F_1-score at the individual level were 31.0% and 45.9%, respectively. Subgroup analysis for each DR grade indicated label discrepancies between the DL algorithm and human labels, mainly accumulated in R1, R2, and R3a (Figure S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Grading results of the DL[a] algorithm and the ground truth in identifying the presence of referable DR[b,c].

Performance	Images (N=736,083)		Participants (N=237,824)	
	DL model	Human labels[d]	DL model	Human labels
Positive, n (%)	98,127 (13.3)	45,740 (6.2)	49,114 (20.7)	17,298 (7.3)
Negative, n (%)	666,055 (90.5)	690,343 (93.8)	188,710 (79.4)	220,526 (92.7)
TP[e], n (%)	36,401 (5.0)	—[f]	15,236 (6.4)	—
FP[g], n (%)	57,886 (7.9)	—	33,878 (14.2)	—
TN[h], n (%)	632,457 (85.9)	—	186,648 (78.5)	—
FN[i], n (%)	9,339 (1.3)	—	2,062 (0.9)	—
Accuracy, %	90.9	—	84.9	—
Precision, %	38.6	—	31.0	—
Sensitivity (recall), %	79.6	—	88.1	—
Specificity, %	91.6	—	84.6	—
F_1-score, %	52.0	—	45.9	—

[a]DL: deep learning.
[b]DR: diabetic retinopathy.
[c]Referable DR was defined as preproliferative and proliferative DR.
[d]Human labels came from the final DR-grading results of human graders, based on the NHS DR-screening guidelines.
[e]TP: true positive.
[f]Not applicable.
[g]FP: false positive.
[h]TN: true negative.
[i]FN: false negative.

Adjudication and Label Correction

Table 2 presents the adjudication results of extracted samples with inconsistent grading results between DL and human labels. After adjudication, 560 (63.6%) of 880 images labeled positive were deemed negative, leading to a mean estimated FN error rate of 63.6% (SD 1.6%, 95% CI 60.4%-66.8%). However, 140 (5.2%) of 2710 images labeled negative were deemed positive, with a mean estimated FP error rate of 5.2% (SD 0.4%, 95% CI 4.4%-6.1%). The adjudication at the individual level yielded a similar finding (FN error rate: mean 65.3%, SD 3.4%, 95% CI 58.2%-72.0%; FP error rate: mean 5.1%, SD 0.5%, 95% CI 4.1%-6.3%). Detailed classifications are listed in Table S2 in Multimedia Appendix 1.

Table 2. Adjudicated results of sample images.

Sample adjudication	Image level		Participant level	
	Images labeled positive (n=880)	Images labeled negative (n=2710)	Images labeled positive (n=195)	Images labeled negative (n=1657)
Erroneous labels	560 (63.6%)	140 (5.2%)	128 (65.3%)	85 (5.1%)
Accurate labels	320 (36.4%)	2570 (94.8%)	68 (34.7%)	1572 (94.9%)

As shown in Table 3, the top reasons for misclassifying R0 images as referable DR images in precorrection labels included high myopia (n=106, 28.5%), a normal-appearing fundus (n=82, 22%), artifacts (n=40, 10.8%), and age-related macular degeneration (AMD; n=37, 10%). The predominant reasons for misclassified referable DR cases in precorrection labels included undetected laser spots (n=56, 41.8%), potential venous anomalies (n=25, 18.7%), and intraretinal microvascular anomalies (IRMAs; n=22, 16.4%). Typical images with erroneous labels are presented in Figure 2: Figures 2A and 2B show referable DR images misclassified as negative in human labels, with laser spots and retinal neovascularization, while Figures 2C and 2D show nonreferable DR images misclassified as positive in labels, probably due to a high-myopia fundus and artifacts.

XSL•FO
RenderX

Table 3. Potential reasons for errors[a] in precorrection labels.

Probable reasons	Count, n (%)
Images misclassified as positive	
High myopia	106 (28.5)
Normal-appearing fundus	82 (22.0)
Artifacts	40 (10.8)
AMD[b]	37 (10.0)
Underexposure	33 (8.9)
Other retinopathies	20 (5.4)
Poor quality	11 (3.0)
Hypertensive retinopathy	8 (2.2)
Media opacity	6 (1.6)
Images m isclassified as negative	
Laser spots	56 (41.8)
Venous anomalies	25 (18.7)
IRMAs[c]	22 (16.4)
Neovascularization	14 (10.5)
Multiple blot hemorrhages	11 (8.2)
Fibrous membrane	3 (2.2)
Vitreous hemorrhage	1 (0.8)

[a]Errors in precorrection labels were estimated against the results of the adjudication. Images misclassified as positive represented those graded as referable diabetic retinopathy (DR) in the labels but turned out to be adjudicated as R0; images misclassified as negative represented those graded as nonreferable DR in the labels but turned out to be referable DR after adjudication.

[b]AMD: age-related macular degeneration.

[c]IRMA: intraretinal microvascular anomaly.

Based on the FN and FP error rates in the sample, we assumed that approximately 8933 images (from 3084 individuals) were misclassified in the precorrection labels in the entire data set, resulting in a mean estimated image error rate of 1.2% (SD 0.01%, 95% CI 1.1%-1.3%; individual level: mean 1.3%, SD 0.02%, 95% CI 1.1%-1.5%). The postcorrection labels of the entire data set were estimated to include 42,787 truly positive images and 693,296 truly negative images (from n=17,689, 7.4%, referable DR participants and n=220,135, 92.6%, nonreferable DR participants, respectively).

XSL•FO
RenderX

Figure 2. Typical images with human label errors before correction.

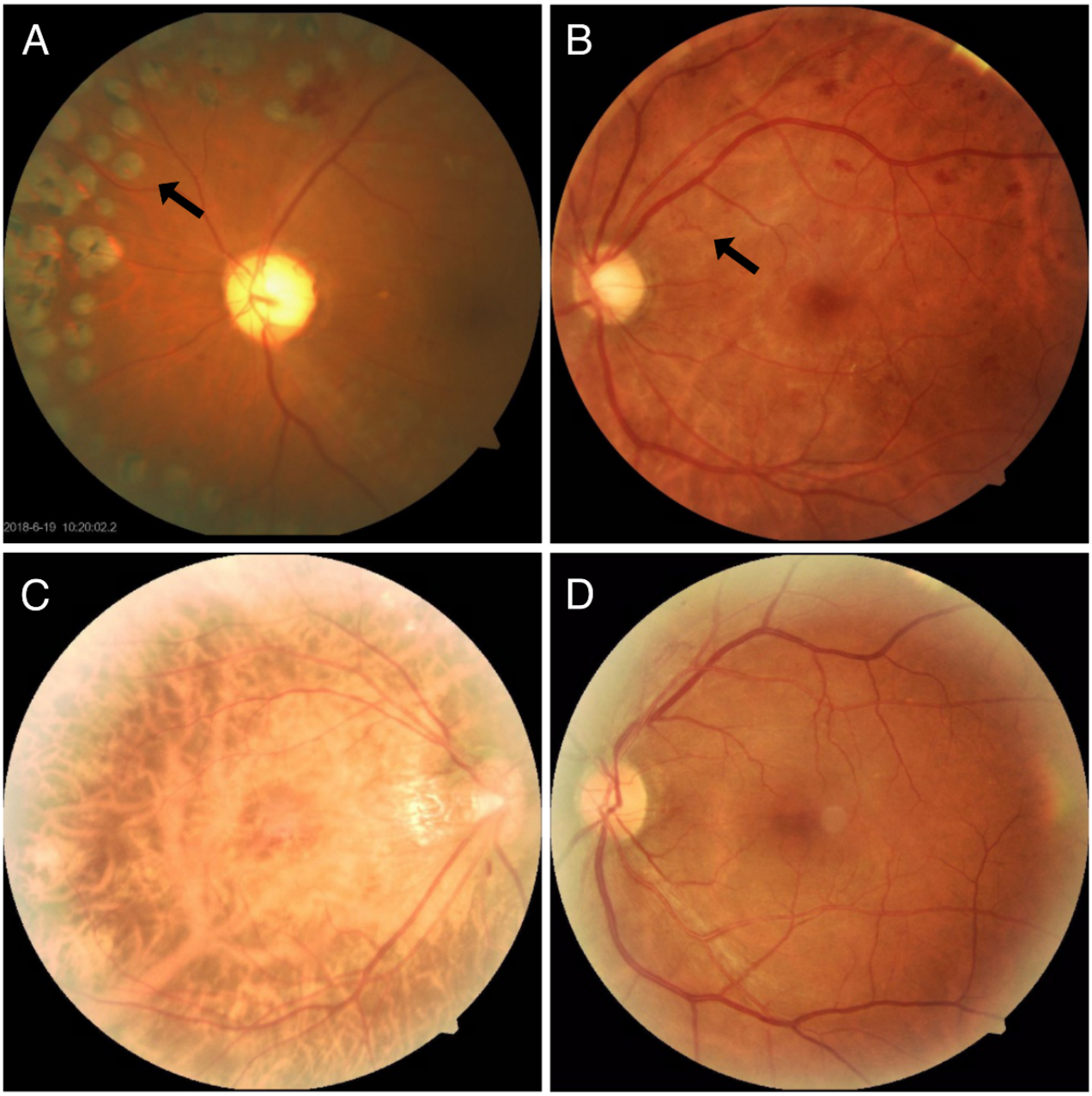

Postcorrection Model Performance

The confusion matrixes in Figure 3 show the comparison of DL model performance before and after label correction. After correction at the image level, sensitivity significantly improved by 12.5% (from 79.6% to 92.1%, $P<.001$), and the corresponding improvement in specificity and accuracy was 0.5% (from 91.6% to 92.1%, $P<.001$) and 1.2% (from 90.9% to 92.1%, $P<.001$), respectively. Similarly, postcorrection results at the individual level showed a 7.9% increase in sensitivity (from 88.1% to 96%, $P<.001$), a 0.8% increase in specificity (from 84.6% to 85.4%, $P<.001$), and a 1.3% increase in accuracy (from 84.9% to 86.2%, $P<.001$). The lab/real-world performance gap was significantly smaller after label correction, with a significantly reduced sensitivity gap from 12.9% to 0.4% ($P<.001$), as well as a reduced specificity gap from 6.9% to 6.4% ($P<.001$).

XSL•FO
RenderX

Figure 3. Performance of the DL model before and after label correction. DL: deep learning.

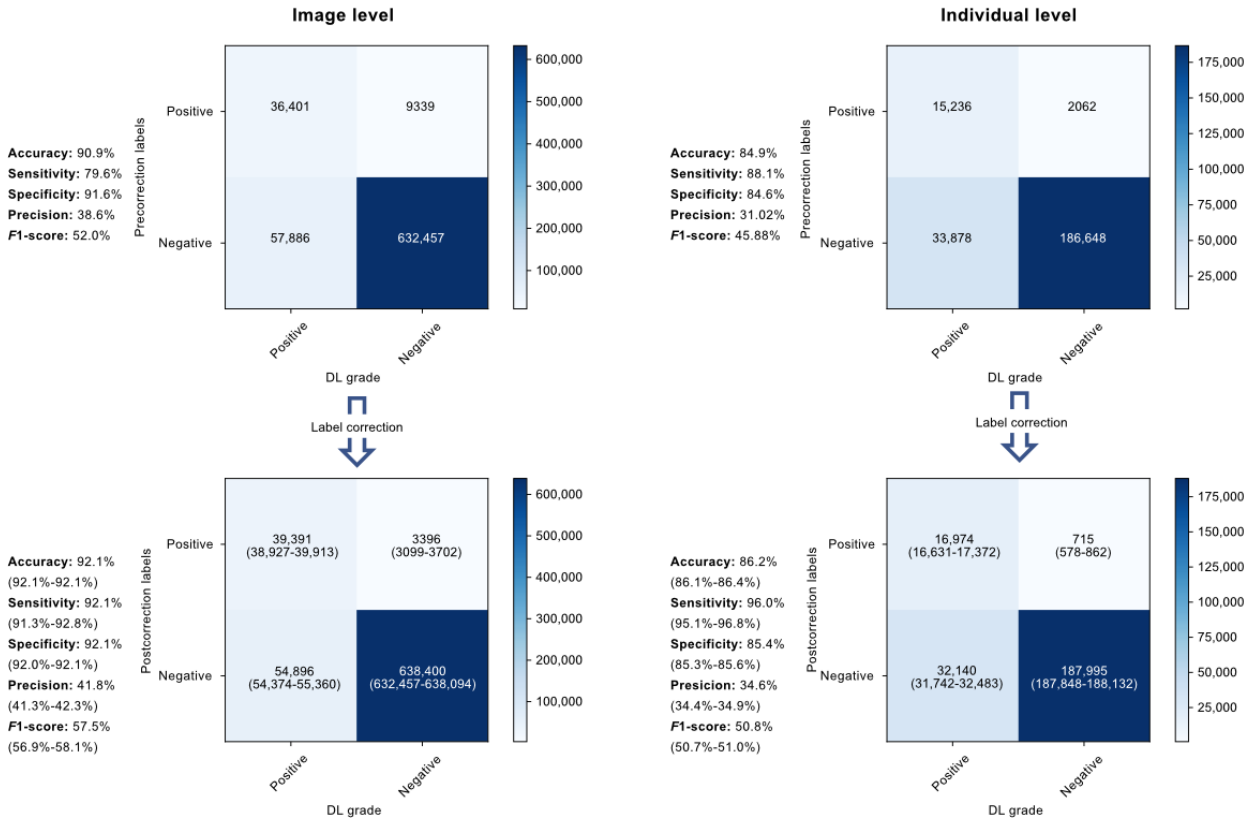

Discussion

Principal Findings

Based on a nationwide DR-screening program, we identified a significant gap between the lab-reported and real-world performance of a DL-based DR-screening model, which could be largely reduced by checking and correcting the errors in human labels. These results highlight the significance of acknowledging potential inaccuracies in human labels, which is often regarded as the gold standard. In this study, we proposed an approach of sampling adjudication to address this issue for a large data set. By evaluating the error rate within sampled subsets, we extrapolated and approximated the accurate distribution of labels across the entire data set. This approach enables a more precise evaluation of AI performance in real-world practice by providing estimates that consider and mitigate the impact of label errors.

Currently, there are 2 predominant approaches to generate a manually labeled gold standard in image-based real-world investigations [6-8,21-28]. The first approach involves a 2-stage grading process. In this process, 2 primary graders review all images, and a senior grader intervenes to make a final decision in the case of a disagreement between the primary graders [21]. The second method is a 1-stage grading process, wherein a single grader reviews all images to obtain human labels. In some instances, the 1-stage grading process might need additional verification for positive cases or a subsample of images [6]. Although accuracy may be somewhat compromised, the 1-stage grading method has advantages in terms of efficiency, convenience, time-effectiveness, and cost-effectiveness. Therefore, 1-stage grading is more commonly used in real-world practice, such as in the case of the Lifeline Express in this study. However, regardless of which grading method is used, we recommend double-checking human labels, even in a subsample of the data, before assessing the performance of the DL model in the real world.

Despite increasing endeavors aimed at enhancing the accuracy of human labels through measures such as standardized training for graders, the issue of mislabeling remains inevitable. Even if human graders make just 1 mistake in 100 cases, such as misclassifying different retinal conditions as DR or missing laser scars on the peripheral retina, it affects the gold standard. These small flaws can significantly influence the development and evaluation of AI algorithms. However, the amount of research available that evaluates the accuracy of manual grading outputs in DR screening is limited. In one study, 7379 images collected from 735 individuals were used to assess the performance of manual grading, which achieved a sensitivity of 82.2% and a specificity of 84.4% [7]. In another study involving 2384 images collected from 1208 participants in a nationwide DR-screening program in Thailand, the grading performance of human graders was deemed relatively reliable (accuracy: 93.5%, sensitivity: 84.8%, specificity: 95.5%) [28]. Moreover, Krause et al [29] compared DR-grading outcomes of a majority decision made by 3 ophthalmologists with the standard reference of adjudication by retinal specialists [29]. A discrepancy rate of around 0.8% can be drawn for identifying vision-threatening DR from their conclusion. These previous studies, in conjunction with our findings, support the need to estimate the error rate associated with manual grading in various real-world DR-screening settings. These estimates range from

XSL•FO
RenderX

approximately 1% to 7%. Variations in the DR prevalence in the target population, grading workflow, and grader certifications may serve as contributing factors.

The label error rate in our study was around 1%, primarily attributed to the misclassification of nonreferable cases as referable DR. This is not anticipated to impede prompt referral and intervention for patients truly requiring medical attention. However, even such a small error rate can largely deteriorate the performance of DL algorithms in the real world. To address this concern, 1 previously proposed solution is to establish a rigorous reference standard during the human labeling process—for example, incorporating adjudication consensus from retinal specialists, rather than relying on major decisions made by multiple graders [29]. However, implementing a rigorous grading workflow in large-scale real-world screening practices can be time-consuming and resource intensive. Additionally, this process cannot be executed after study completion. Our study introduced an optimization strategy based on sample adjudication and label correction. Importantly, this strategy does not require modifications to the grading workflow or retraining of the DL model. Despite addressing only around 1% of label errors in the entire data set, this optimization strategy has been shown to bring about 12% enhancement in the AI model's sensitivity.

The label correction strategy in our study can be considered a post hoc quality control procedure. It has been reported that label corrections for model tuning could effectively enhance the performance of AI algorithms [29]. Our findings further support the potential beneficial effect of adding this quality control procedure in existing and future DR-screening programs, especially when a large gap is observed between the real-world and lab-reported performance of the AI algorithm. However, implementation of this strategy in real-world practice may present challenges. Even if only a small subsample of images is selected for review, this would increase the workload of ophthalmologists, necessitating additional time and resources. Based on our study, we recommend that sampling approximately 5% of images with disagreement is reasonable for post hoc quality control. We recommend that future studies determine an appropriate sample size based on their expected error rate, resources, and other real-world considerations. We believe that this label correction strategy offers a standardized quality control pipeline across diverse image-based AI applications. Future investigations could explore the potential effect of this strategy in other AI-based models (eg, other diseases or imaging modalities) and other health care scenarios (eg, disease diagnosis or prognosis).

Strengths

This study possesses several notable strengths. First, it was embedded in a nationwide DR-screening program, ensuring a large and robust data set comprising over 0.7 million fundus images collected from more than 0.2 million participants. This considerable scale enhanced the reliability of our findings. Second, to the best of our knowledge, this study represents a pioneering effort in evaluating the impact of label errors on the performance of a DR-grading DL algorithm. In addition, the proposal and validation of an effective optimization strategy in this study contribute significantly to the understanding and advancement of DL implementation in real-world DR screening.

Limitations

This study has some limitations. In our evaluation, images with identical outputs between DL and labels (TP and TN cases) were considered correctly classified, without undergoing further adjudication. In smaller-scale trials, we adjudicated a limited subset of TP (200 images) and TN (1000 images) cases. These assessments revealed a minimal error rate in human labeling, at 2.5% for TP and 0.1% for TN (further details in Table S3 in Multimedia Appendix 1). Given the substantial quantity of TP and TN images within our data set, refraining from label correction in these subsets would unlikely cause a significant alteration in the estimated performance of the DL model. Additionally, during the adjudication process, we conducted random sampling instead of reviewing all images with discrepancies. To ensure the representativity of the random sample, we performed multiple samplings with varying sample sizes and validated consistent distributions of the obtained results. Moreover, the Lifeline Express program exclusively recruited Chinese participants. It is imperative to conduct additional validation studies encompassing diverse ethnicities to corroborate our findings. Lastly, our analysis focused solely on a single DR-screening model. Nevertheless, it is reasonable to infer that our conclusions are applicable to alternative models used in the screening of other ocular diseases. Therefore, further investigations are warranted to substantiate this discovery.

Conclusion

In conclusion, we found that even a small percentage of label errors could have a substantial impact on the performance of DL algorithms in real-world DR screening. Attention should be paid to minimizing and correcting label errors in future studies and implementations of AI-based DR-screening models in the real world.

Acknowledgments

We would like to acknowledge the Chinese Foundation for Lifeline Express and the Lifeline Express Hong Kong Foundation for authorizing access to the data set. We would also like to acknowledge Ms Nellie Fong and Ms Mingyue Wang for their kind assistance during data collection.

This study was funded by the National Natural Science Foundation of China (82000901, 82101171), the Global STEM Professorship Scheme (P0046113), the Fundamental Research Funds of the State Key Laboratory of Ophthalmology (303060202400201209), and the Outstanding PI Research Funds of the State Key Laboratory of Ophthalmology (3030902113083).

XSL•FO
RenderX

Authors' Contributions

Study concept and design were handled by YW, XTH, and MH; acquisition, analyses, interpretation, and critical revision of the manuscript for important intellectual content by all authors; drafting of the manuscript by YW, CL, XTH, and DLS; statistical analyses by YW and JZ; funding by MH and XTH; administrative, technical, and material support by GKP and DLS; and study supervision by MH.

Conflicts of Interest

The authors declare no proprietary interest in any aspect of the study. The Chinese Foundation for Lifeline Express and the Guangzhou Vision Tech Medical Technology Co hereby are not associated directly or indirectly with the subject or design of this study.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 749 KB - jmir_v26i1e52506_app1.docx]

References

1. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018 Sep;24(9):1342-1350. [doi: 10.1038/s41591-018-0107-6] [Medline: 30104768]
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
3. Milea D, Najjar RP, Jiang Z, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. N Engl J Med 2020 Apr 30;382(18):1687-1695. [doi: 10.1056/nejmoa1917130]
4. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
5. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. Lancet 2010 Jul 10;376(9735):124-136. [doi: 10.1016/S0140-6736(09)62124-3] [Medline: 20580421]
6. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
7. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021 May;44(5):1168-1175 [FREE Full text] [doi: 10.2337/dc20-1877] [Medline: 33402366]
8. Zhang Y, Shi J, Peng Y, Zhao Z, Zheng Q, Wang Z, et al. Artificial intelligence-enabled screening for diabetic retinopathy: a real-world, multicenter and prospective study. BMJ Open Diabetes Res Care 2020 Oct 21;8(1):e001596 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001596] [Medline: 33087340]
9. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]
10. Bellemo V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. Curr Diab Rep 2019 Jul 31;19(9):72. [doi: 10.1007/s11892-019-1189-3] [Medline: 31367962]
11. Lebovitz S, Levina N, Lifshitz-Assa H. Is AI ground truth really ‘true’? The dangers of training and evaluating AI tools based on experts’ know-what. MISQ 2021 Sep 01;45(3):1501-1526. [doi: 10.25300/misq/2021/16564]
12. Khattak F, Salleb-Aouissi A. Quality control of crowd labeling through expert evaluation. 2011 Presented at: NIPS 2011: 2nd Workshop on Computational Social Science and the Wisdom of Crowds; December 17, 2011; Sierra Nevada, Spain.
13. Jaton F. Assessing biases, relaxing moralism: on ground-truthing practices in machine learning design and application. Big Data Soc 2021 May 05;8(1):205395172110135. [doi: 10.1177/20539517211013569]
14. Lemoigne Y, Caner A. Molecular Imaging: Computer Reconstruction and Practice. Berlin: Springer Science & Business Media; 2008.
15. Søgaard A, Plank B, Hovy D. Selection bias, label bias, and bias in ground truth. 2014 Presented at: COLING 2014: 25th International Conference on Computational Linguistics: Tutorial Abstracts; August 2014; Dublin, Ireland.
16. Lin D, Xiong J, Liu C, Zhao L, Li Z, Yu S, et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. Lancet Digital Health 2021 Aug;3(8):e486-e495. [doi: 10.1016/s2589-7500(21)00086-8]

XSL•FO
RenderX

17. Kats E, Goldberger J, Greenspan H. A soft STAPLE algorithm combined with anatomical knowledge. 2019 Presented at: MICCAI 2019: 22nd International Conference Medical Image on Computing and Computer Assisted Intervention; October 13-17, 2019; Shenzhen, China. [doi: 10.1007/978-3-030-32248-9_57]
18. Jacob J, Ciccarelli O, Barkhof F, Alexander D, editors. Disentangling human error from the ground truth in segmentation of medical images. 2020 Presented at: NIPS 2020: 34th International Conference on Neural Information Processing Systems; December 6-12, 2020; Online.
19. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. Diabetes Care 2018 Dec;41(12):2509-2516. [doi: 10.2337/dc18-0147] [Medline: 30275284]
20. Kentaro WM, Martijn B, Zhang C, Kubovčík M, Myczko A, Zhu L, et al. wkentaro/labelme: v4.6.0. Zenodo. 2021 Nov 18. URL: https://zenodo.org/records/5711226 [accessed 2024-07-01]
21. Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. Diabetes Care 2019 Apr;42(4):651-656. [doi: 10.2337/dc18-0148] [Medline: 30765436]
22. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. Br J Ophthalmol 2021 May;105(5):723-728 [FREE Full text] [doi: 10.1136/bjophthalmol-2020-316594] [Medline: 32606081]
23. Ming S, Xie K, Lei X, Yang Y, Zhao Z, Li S, et al. Evaluation of a novel artificial intelligence-based screening system for diabetic retinopathy in community of China: a real-world study. Int Ophthalmol 2021 Apr 03;41(4):1291-1299. [doi: 10.1007/s10792-020-01685-x] [Medline: 33389425]
24. Scheetz J, Koca D, McGuinness M, Holloway E, Tan Z, Zhu Z, et al. Real-world artificial intelligence-based opportunistic screening for diabetic retinopathy in endocrinology and indigenous healthcare settings in Australia. Sci Rep 2021 Aug 04;11(1):15808 [FREE Full text] [doi: 10.1038/s41598-021-94178-5] [Medline: 34349130]
25. Li N, Ma M, Lai M, Gu L, Kang M, Wang Z, et al. A stratified analysis of a deep learning algorithm in the diagnosis of diabetic retinopathy in a real-world study. J Diabetes 2022 Feb;14(2):111-120 [FREE Full text] [doi: 10.1111/1753-0407.13241] [Medline: 34889059]
26. Malerbi FK, Andrade RE, Morales PH, Stuchi JA, Lencione D, de Paulo JV, et al. Diabetic retinopathy screening using artificial intelligence and handheld smartphone-based retinal camera. J Diabetes Sci Technol 2022 May;16(3):716-723 [FREE Full text] [doi: 10.1177/1932296820985567] [Medline: 33435711]
27. Mehra AA, Softing A, Guner MK, Hodge DO, Barkmeier AJ. Diabetic retinopathy telemedicine outcomes with artificial intelligence-based image analysis, reflex dilation, and image overread. Am J Ophthalmol 2022 Dec;244:125-132. [doi: 10.1016/j.ajo.2022.08.008] [Medline: 35970206]
28. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. Lancet Digit Health 2022 Apr;4(4):e235-e244 [FREE Full text] [doi: 10.1016/S2589-7500(22)00017-6] [Medline: 35272972]
29. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018 Aug;125(8):1264-1272. [doi: 10.1016/j.ophtha.2018.01.034] [Medline: 29548646]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
DL: deep learning
DR: diabetic retinopathy
FN: false negative
FP: false positive
NHS: National Health Service
TN: true negative
TP: true positive

XSL•FO
RenderX

Edited by B Puladi; submitted 11.09.23; peer-reviewed by M Liu, C Zheng, J Jaskari; comments to author 06.11.23; revised version received 30.12.23; accepted 22.03.24; published 14.08.24.

Please cite as:
Wang Y, Han X, Li C, Luo L, Yin Q, Zhang J, Peng G, Shi D, He M
Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study
J Med Internet Res 2024;26:e52506
URL: https://www.jmir.org/2024/1/e52506
doi:10.2196/52506
PMID:

©Yueye Wang, Xiaotong Han, Cong Li, Lixia Luo, Qiuxia Yin, Jian Zhang, Guankai Peng, Danli Shi, Mingguang He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Domain Adaptation and Inductive Transfer Learning to Improve Patient Outcome Prediction in the Intensive Care Unit: Retrospective Observational Study

Maruthi Kumar Mutnuri[1,2], MSc; Henry Thomas Stelfox[3,4], MD, PhD; Nils Daniel Forkert[5,6], PhD; Joon Lee[1,7,8,9], PhD

[1]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[2]Department of Biomedical Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB, Canada
[3]Department of Critical Care Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[4]O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[5]Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[6]Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[7]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[8]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[9]Department of Preventive Medicine, School of Medicine, Kyung Hee University, Seoul, Republic of Korea

Corresponding Author:
Joon Lee, PhD
Data Intelligence for Health Lab
Cumming School of Medicine
University of Calgary
CWPH 5E17
3280 Hospital Drive Northwest
Calgary, AB, T2N 4Z6
Canada
Phone: 1 4032202968
Email: joon.lee@ucalgary.ca

Abstract

Background: Accurate patient outcome prediction in the intensive care unit (ICU) can potentially lead to more effective and efficient patient care. Deep learning models are capable of learning from data to accurately predict patient outcomes, but they typically require large amounts of data and computational resources. Transfer learning (TL) can help in scenarios where data and computational resources are scarce by leveraging pretrained models. While TL has been widely used in medical imaging and natural language processing, it has been rare in electronic health record (EHR) analysis. Furthermore, domain adaptation (DA) has been the most common TL method in general, whereas inductive transfer learning (ITL) has been rare. To the best of our knowledge, DA and ITL have never been studied in-depth in the context of EHR-based ICU patient outcome prediction.

Objective: This study investigated DA, as well as rarely researched ITL, in EHR-based ICU patient outcome prediction under simulated, varying levels of data scarcity.

Methods: Two patient cohorts were used in this study: (1) eCritical, a multicenter ICU data from 55,689 unique admission records from 48,672 unique patients admitted to 15 medical-surgical ICUs in Alberta, Canada, between March 2013 and December 2019, and (2) Medical Information Mart for Intensive Care III, a single-center, publicly available ICU data set from Boston, Massachusetts, acquired between 2001 and 2012 containing 61,532 admission records from 46,476 patients. We compared DA and ITL models with baseline models (without TL) of fully connected neural networks, logistic regression, and lasso regression in the prediction of 30-day mortality, acute kidney injury, ICU length of stay, and hospital length of stay. Random subsets of training data, ranging from 1% to 75%, as well as the full data set, were used to compare the performances of DA and ITL with the baseline models at various levels of data scarcity.

Results: Overall, the ITL models outperformed the baseline models in 55 of 56 comparisons (all P values <.001). The DA models outperformed the baseline models in 45 of 56 comparisons (all P values <.001). ITL resulted in better performance than

XSL•FO
RenderX

DA in terms of the number of times and the margin with which it outperformed the baseline models. In 11 of 16 cases (8 of 8 for ITL and 3 of 8 for DA), TL models outperformed baseline models when trained using 1% data subset.

Conclusions: TL-based ICU patient outcome prediction models are useful in data-scarce scenarios. The results of this study can be used to estimate ICU outcome prediction performance at different levels of data scarcity, with and without TL. The publicly available pretrained models from this study can serve as building blocks in further research for the development and validation of models in other ICU cohorts and outcomes.

(J Med Internet Res 2024;26:e52730) doi:10.2196/52730

KEYWORDS

transfer learning; patient outcome prediction; intensive care; deep learning; electronic health record

Introduction

Electronic health records (EHRs) are databases that hospitals and health care providers use to record an individual's health history. There has been significant progress in using deep learning models for predicting patient outcomes using EHR data [1]. However, using deep learning models is not feasible in some settings, such as rural hospital ICUs, which have low patient volumes and limited computational capacity due to budget restrictions.

Transfer learning (TL) can be useful in these challenging scenarios. TL research using EHR data has been uncommon compared to medical image analysis and natural language processing. The basic idea of TL is to use the knowledge and representations learned while training the model on a source prediction task, and to improve prediction performance on a different, but potentially closely related target prediction task [2]. In practice, this is achieved by pretraining a model with data for the source prediction task and saving the trained weights. These weights capture the intrinsic knowledge of the data. This pretrained model is typically loaded without the final layers (usually the fully connected layers immediately preceding the output layer and the output layer itself), which are then replaced with new layers. Finally, this pretrained model is retrained to fine-tune it to the target prediction task (Figure 1).

XSL•FO
RenderX

Figure 1. An illustration of how DA and ITL were applied in this study. For ITL, the source and target prediction tasks were 30-day mortality and AKI, respectively, while the source and target domains were both eCritical. For DA, the source and target prediction tasks were both 30-day mortality, while the source and target domains were eCritical and MIMIC-III, respectively. AKI: acute kidney injury; DA: domain adaptation; ITL: inductive transfer learning; MIMIC-III: Medical Information Mart for Intensive Care III.

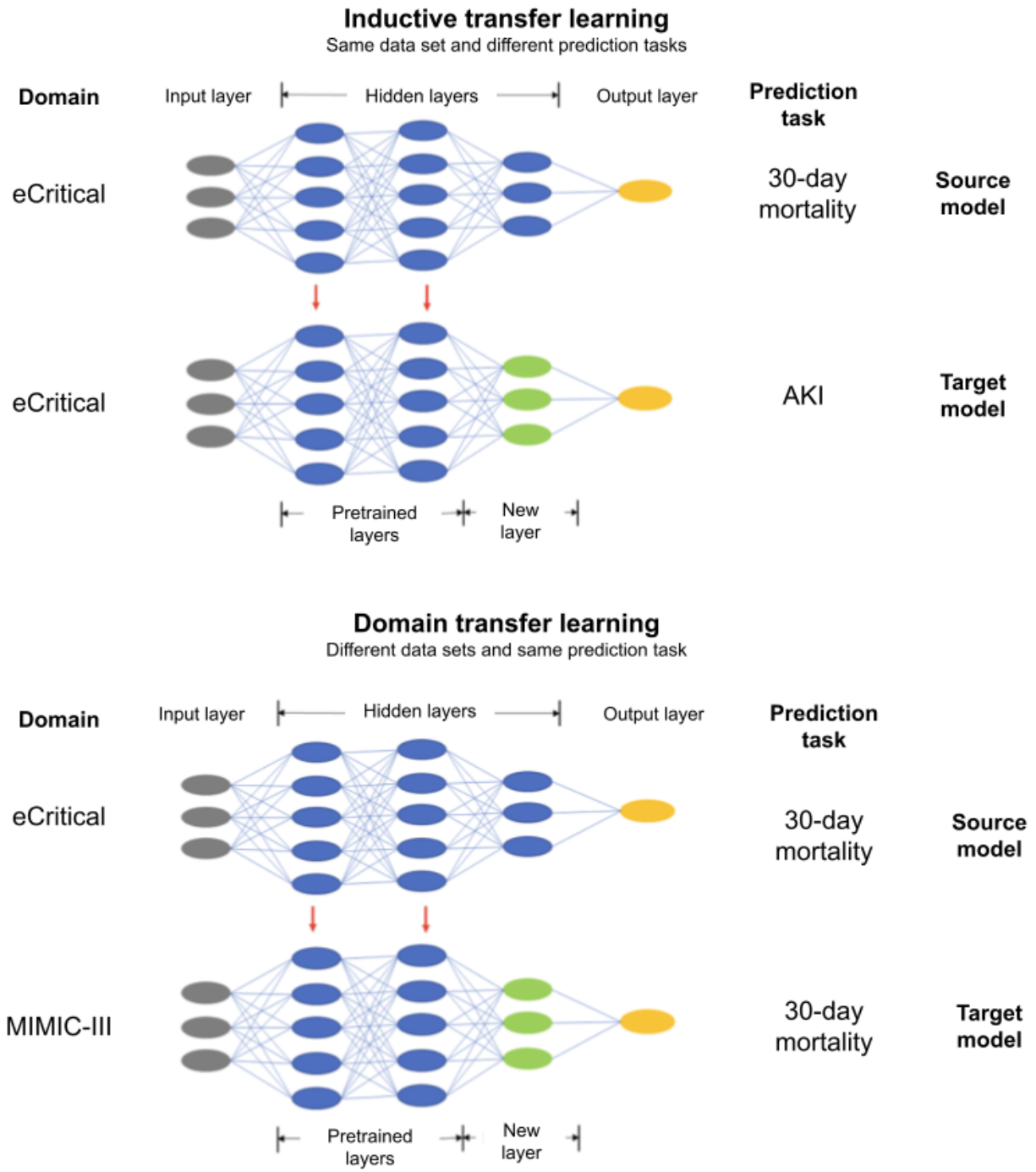

In this work, we considered 2 commonly used types of TL methods. First, inductive transfer learning (ITL) aims to improve performance on the target task after learning a different but related source task, usually from the same domain [3]. For example, in the study by Tokuoka et al [4], a model trained for brain tissue annotation label (source task) was adapted to the task of brain tumor segmentation (target task) using magnetic resonance imaging images. Second, transductive TL, also referred to as domain adaptation (DA), makes use of different domains but the same prediction task [3]. For example, in the study by Titoriya and Sachdeva [5], AlexNet [6] model, which was pretrained on ImageNet data (source data set) for object classification (source task), was retrained on the Breakhis data set (target data set) to classify medical images into malignant or benign (target task) to predict breast cancer.

TL can be useful in a data-scarce scenario where the target data set does not have a sufficient volume to train a deep learning model but there is a sufficiently large source data set to train a pretrained model or a relevant pretrained model is available.

XSL•FO
RenderX

For example, DA can be useful in a scenario where a rural intensive care unit (ICU) does not have sufficient data to train a model to predict a patient outcome, but an urban teaching hospital has a large data set to train a model for the same patient outcome. ITL can be useful when predicting new or rare patient outcomes at an ICU if that ICU has a sufficiently large data set for training a model to predict a different patient outcome.

TL research in medical image analysis is relatively comprehensive, as pretrained convolutional neural networks, such as AlexNet [6], ResNet [7], VGGNet [8], and GoogleNet [9], are publicly available and have been used widely for prediction problems such as image classification [10], image segmentation [11], object identification [12], disease categorization [13], and severity grading [14]. Most of these examples used DA rather than ITL.

Another field with significant previous TL research is natural language processing. Some of the established pretrained language models are Word2Vec [15], GloVe [16], Bidirectional Encoder Representations from Transformers [17], and fastText [18]. Some of the use cases of pretrained natural language processing models include text mining [19], word classification [20], and sentiment classification [21].

A recent trend has been to adapt the progress made in TL research in the natural language processing domain to EHR data analysis. Inspired by the pretrained Bidirectional Encoder Representations from Transformers model [17], Li et al [22] developed BEHRT using EHR data to improve future visit diagnosis prediction. This study used only 4 features, which included age, segment, position, and diagnosis. Longitudinal data were used, and a transformer-based model was trained. This study had inclusion criteria of patients with at least 5 visits and a diagnosis available in their EHR. One of the shortcomings of this study was not using all the available features in the EHR data such as demographics, laboratory results, vitals, and prescriptions. Another concern was the requirement of 5 previous visits with diagnosis, causing the model to be unable to predict the diagnosis for patients with fewer previous visits.

Liu et al [23] used TL to improve the prediction of acute kidney injury (AKI) using EHR data acquired at the University of Kansas Medical Center. Logistic regression (LR) was used as the global (baseline) model and the global TL model. Global TL and baseline models are the same LR models except the baseline model was trained on the original data set and the TL model was trained on the modified data set. To create the modified data set, data of each feature were multiplied by the corresponding feature coefficient, which was obtained from the global baseline model. The personalized model was then trained on a subset of the training sample with the highest similarity (nearest neighbors) with the selected test sample. For each test sample, a subset of the training data set with the highest k-nearest neighbor score was selected and a personalized model and personalized TL model were trained. The personalized TL model followed the same approach as the global TL model except that the training sample was selected from the modified data set using the k-nearest neighbor score (similarity) with the selected test sample. Although the source and target prediction tasks and domains were the same, this can be considered as DA since TL models use a modified data set. Here, deep learning models were not used, which are often assumed to be better at learning the general representations of the data. Since both these TL models (global and personalized) were not like the traditional pretrained models with saved weights, the transfer of knowledge happened by modifying the data set. Thus, the transferred knowledge is stored in a modified data set, not in the TL models. These TL models are strongly tied to this specific data set and their generalizability has not been well established by using external validation data.

In the study by Shickel et al [24], TL was used to improve hospital discharge prediction using a conventional ICU cohort acquired at the University of Florida Health as the source cohort and the intelligent ICU cohort also acquired at the University of Florida Health as the target cohort. It used a feed-forward neural network for TL. Here, the source cohort had 48,400 patient records whereas the target cohort had 51 patient records. This study used only 9 features and DA, even though not stated specifically in the paper. This semantic progression was observed in other studies as well; DA has become so prevalent that the terms TL and DA are being used interchangeably, with the former most commonly referring to the latter [24,25].

To the best of our knowledge, TL research has been rare in EHR-based ICU patient outcome prediction, particularly in terms of ITL, leading to a limited understanding of the effectiveness of TL in data-scarce ICU settings. Furthermore, there is a lack of publicly available EHR-based pretrained models for ICU patient outcome prediction. These are important gaps because EHR-based tabular data are one of the most widely used forms of data in predictive modeling studies in health, and the state-of-the-art TL methods from general computer vision and natural language processing are often not readily applicable to EHR data. Hence, this retrospective study aimed to compare the performances of DA, ITL, and baseline (without TL) models in predicting the following 4 ICU patient outcomes at varying levels of data scarcity: 30-day post-ICU admission mortality, AKI, hospital length of stay (H_LOS), and ICU length of stay (ICU_LOS).

Methods

Data Sources

EHR data from 2 patient cohorts were used in this retrospective study. The first cohort was eCritical, which has 55,689 unique admission records from 48,672 unique patients admitted to 15 ICUs in Alberta, Canada, between March 2013 and December 2019. The second cohort was the Medical Information Mart for Intensive Care III (MIMIC-III) database [26] version 1.4, which includes 61,532 unique admission records from 46,476 unique patients admitted to the ICUs at the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012.

Ethical Considerations

Since MIMIC-III is a publicly available database, the need to obtain research ethics approval to use it in this study was waived. However, eCritical contains patient identifying information, and approval from the Conjoint Health Research

XSL•FO
RenderX

Ethics Board, University of Calgary, was obtained (REB17-0389). Informed consent was waived due to the large number of patients involved in the study. All research was performed in accordance with relevant guidelines and regulations set by the University of Calgary and Alberta Health Services, the custodian of the eCritical data, as well as the Declaration of Helsinki.

Patient Cohorts

Two different sets of inclusion and exclusion criteria were applied. The first set was used to establish the base cohorts for the entire study, whereas the second set was specific to each patient outcome and applied to the base cohorts.

For the base cohorts, the following inclusion criteria were applied to both eCritical and MIMIC-III (Figure 2): (1) only the first ICU admission of each patient; (2) ICU_LOS greater than 24 hours; (3) only adult patients with an age 18 years and older; and (4) samples with data available of at least 80% of the features (missing values are imputed as discussed in data preprocessing section). As a result, the base cohorts for eCritical and MIMIC-III consisted of 39,317 and 31,446 patient records, respectively.

Figure 2. Patient cohort flowchart. AKI: acute kidney injury; H_LOS: hospital length of stay; ICU: intensive care unit; ICU_LOS: ICU length of stay; MIMIC: Medical Information Mart for Intensive Care.

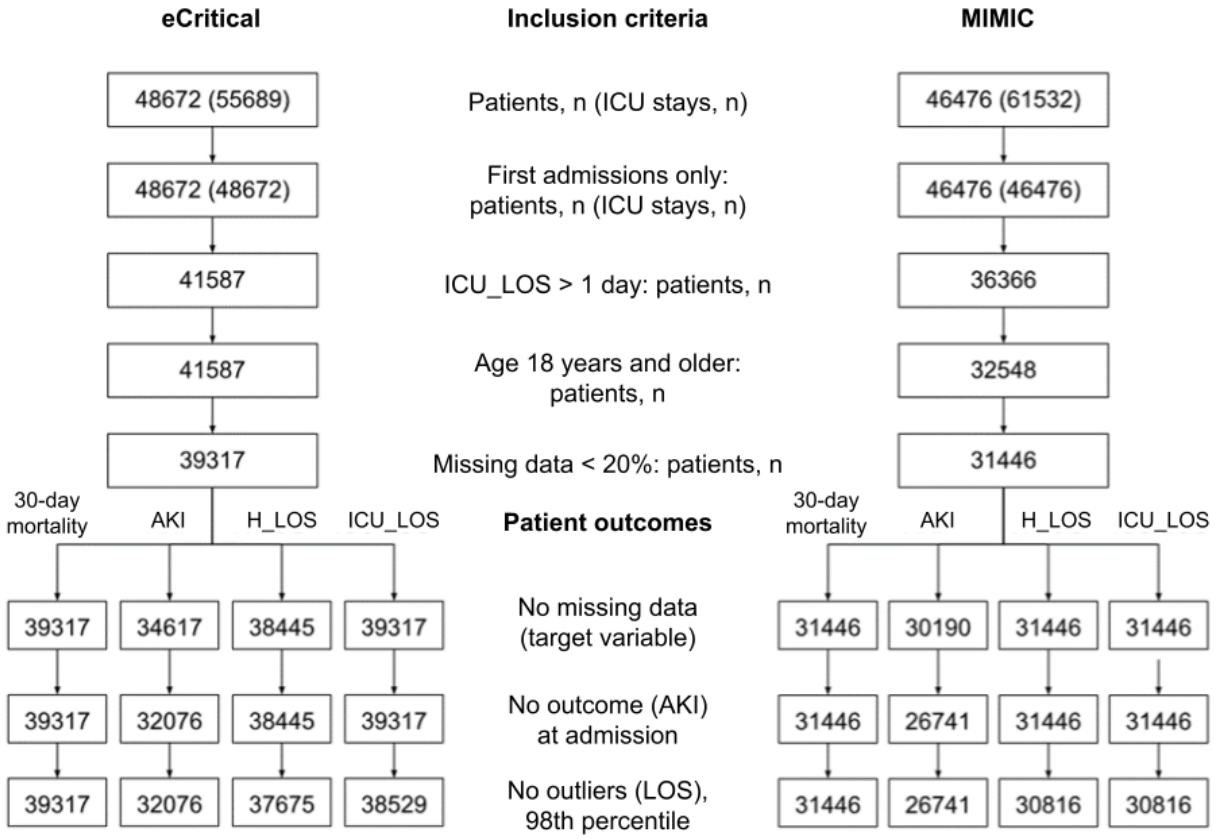

Further inclusion and exclusion criteria were applied to each patient outcome, resulting in different data sets for each outcome, as shown in Figure 2. The 30-day mortality did not require further inclusion criteria and was modeled using the base cohorts. For AKI, the following criteria were applied: (1) only patient records with sufficient data to determine the presence or absence of AKI, with one serum creatinine laboratory measurement within the first 24 hours of admission to be able to establish a baseline and another measurement after 24 hours of ICU admission; and (2) no AKI onset at or within 24 hours of admission. For ICU_LOS, the following inclusion criteria were applied: (1) the presence of ICU admission and discharge date times; and (2) to exclude outliers only the bottom 98th percentile values of ICU_LOS are included. For H_LOS, the following inclusion criteria were applied: (1) the presence of hospital admission and discharge date times; and (2) to exclude outliers only the bottom 98th percentile values of H_LOS were included.

In the end, the eCritical and MIMIC-III cohorts had 39,317 and 31,446 samples, respectively, for 30-day mortality. Similarly, AKI had 32,076 and 26,741 samples, H_LOS had 37,675 and 30,816 samples, and ICU_LOS had 38,529 and 30,816 samples, respectively. These cohorts were randomly split into 80% training, 10% validation, and 10% test data. We compared the prediction performance of TL with those of baseline models at different levels of training data scarcity with random subsets of 1%, 5%, 10%, 25%, 50%, 75%, and 100% of the training data.

Patient Outcomes

The primary patient outcomes for this study were 30-day post-ICU admission mortality and ICU_LOS. A patient was defined as deceased if he or she died after being admitted to the

XSL•FO
RenderX

ICU and within 30 days of ICU admission. ICU_LOS was defined as the time between ICU admission and discharge.

Furthermore, AKI after 24 hours of ICU admission and H_LOS were predicted as secondary patient outcomes. AKI was identified using the creatinine criteria of KDIGO [27]. H_LOS was defined as the time between hospital admission and discharge.

30-day mortality and AKI were predicted as classification problems, whereas H_LOS and ICU_LOS were predicted as regression problems.

Feature Set

Machine learning (ML) models were trained with the following predictor variables from the first 24 hours in the ICU that were common in both eCritical and MIMIC-III: demographics, vitals, laboratory test results, Glasgow Coma Scale, prescriptions, dialysis, and mechanical ventilation. Features with more than 30% missing data were excluded from the study. Table 1 shows a complete list of the predictor variables. Four statistical features (5th percentile, 95th percentile, IQR, and median) were extracted from the longitudinal variables such as vitals, laboratory results, and Glasgow Coma Scale. The maximum and minimum values of laboratory assessments and vitals carry crucial information regarding the health condition of the patient, but to minimize the influence of outliers, the 5 percentile and 95 percentiles were used instead. Other predictor variables, such as prescriptions, dialysis, and mechanical ventilation, were transformed into binary features to indicate the presence or absence. In the end, a total of 104 features were included in this study.

XSL•FO
RenderX

Table 1. Common feature set between eCritical and MIMIC-III[a].

Feature	Category	Unit of measurement
Age	Demographics	years
Weight	Demographics	kg
Sex	Demographics	Binary (M/F)
Eye-opening	GCS[b]	N/A[c]
Verbal response	GCS	N/A
GCS	GCS	N/A
Motor response	GCS	N/A
Urine volumes	Urine volumes	mL
Heart rate	Vitals	bpm
BP[d] systolic	Vitals	mm Hg
BP diastolic	Vitals	mm Hg
SpO_2[e]	Vitals	%
Respiratory rate	Vitals	breaths/min
Urea blood	Laboratory findings	mmol/L
CO_2 content blood	Laboratory findings	mmol/L
Creatinine blood	Laboratory findings	µmol/L
Glucose blood	Laboratory findings	mmol/L
Potassium blood	Laboratory findings	mmol/L
Sodium blood	Laboratory findings	mmol/L
pCO_2[f] arterial	Laboratory findings	mmHg
FiO_2	Laboratory findings	%
PH arterial	Laboratory findings	N/A
pO_2[g] arterial	Laboratory findings	mmHg
Hemoglobin	Laboratory findings	g/L
Hematocrit	Laboratory findings	%
RBC[h]	Laboratory findings	E+12 units/L
WBC[i]	Laboratory findings	E+9 units/L
Dialysis	Dialysis	Binary (1/0)
Mechanical ventilation	Mechanical ventilation	Binary (1/0)
Norepinephrine	Prescriptions	Binary (1/0)
Phenylephrine	Prescriptions	Binary (1/0)
Vasopressin	Prescriptions	Binary (1/0)
Dobutamine	Prescriptions	Binary (1/0)
Dopamine	Prescriptions	Binary (1/0)
Epinephrine	Prescriptions	Binary (1/0)

[a]MIMIC-III: Medical Information Mart for Intensive Care III.
[b]GCS: Glasgow Coma Scale.
[c]N/A: not applicable.
[d]BP: blood pressure.
[e]SpO_2: oxygen saturation.
[f]pCO_2: partial pressure of carbon dioxide.

XSL•FO
RenderX

[g]pO_2: partial pressure of oxygen.

[h]RBC: red blood cell.

[i]WBC: white blood cell.

Data Preprocessing

Differences in units of measurement between eCritical and MIMIC-III were handled by converting all features in MIMIC-III to the eCritical units of measurement.

The train set was used for training the models, whereas the validation set was used for tuning hyperparameters. The test set was used for model performance evaluation. Numerical features (eg, vitals and laboratory findings) were scaled (unit variance and zero mean) and categorical features (eg, sex, prescriptions, and mechanical ventilation) were transformed using one-hot encoding.

Missing data were present to varying degrees in both cohorts. Features with more than 30% missing data were excluded from the study. Patient records with missing data for more than 20% of the features were dropped. The remaining patient records with missing values were imputed using the IterativeImputer from the Scikit-learn Python package, which is similar to the multiple imputation by chained equations. Imputation was performed after splitting the data to avoid data leakage. Training, validation, and test data were imputed separately.

Both categorical patient outcomes, 30-day mortality and AKI, had varying degrees of class imbalance in both cohorts. The 30-day mortality had the highest class imbalance; the event rates were 16.79% and 12.24% in eCritical and MIMIC-III, respectively. The class imbalance was mitigated using the Synthetic Minority Oversampling Technique [28], by oversampling the minority class to 50% of the majority class and then undersampling the majority class to 100% of the minority class.

Baseline Models

Since logistic [29] and lasso [30] regression are widely used in medical research for classification and regression, respectively, they were used as baseline models. In addition, deep learning models (fully connected neural network [FCNN]) with random initialization of weights were used as baseline models.

Hyperparameter tuning for the LR and lasso models was done using grid search with 3-fold cross-validation. The searched hyperparameter space for LR included: solvers of newton-cg, liblinear, lbfgs, sag, and saga; penalties of L1, L2, and none; and C values ranging from 0.01 to 10 with a step of 0.01. The searched hyperparameter space for lasso included: α values ranging from .01 to 1 with a step of .01. FCNN baseline models used the same architecture and hyperparameters as the corresponding TL models so that we were comparing models that were trained the same way except how the weights were initialized.

Eight FCNN models were created for the 4 patient outcomes and 2 cohorts. Four LR models were created for 30-day mortality and AKI trained on eCritical and MIMIC-III. Similarly, 4 lasso regression models were created for H_LOS and ICU_LOS trained on eCritical and MIMIC-III.

TL Models

In DA, the source and target domains were eCritical and MIMIC-III, respectively. The source and target tasks were the same, and each DA model predicted one of the four patient outcomes. Each model was pretrained on the source domain data before being fine-tuned and evaluated on the target domain data. As a result, four pretrained DA models were created, one for each of the 4 patient outcomes.

In ITL, both the source and target domains were eCritical and the source task was 30-day mortality prediction whereas the target task was the prediction of one of the four patient outcomes. The ITL model where both the source and target tasks were 30-day mortality served as a benchmark for the other ITL models. In the end, 4 pretrained ITL models were created for each of the 4 patient outcomes.

The pretrained TL models were FCNNs trained on the training data set of the source domain. Hyperparameters were tuned using the validation data set. The searched hyperparameter space included: dropout rates of 0.5, 0.4, and 0.3; batch sizes of 32, 64, and 128; numbers of neurons per hidden layer of 100, 128, 256, and 200; learning rates of 0.001 and 0.0001; activation functions of ReLU, tanh, selu, elu, LeakyReLU, and PReLU; kernel initializers of He Uniform and He Normal; and kernel regularizers of L2 (l2=1e-3) and L1 (l1=0.001). In addition, different architectures were explored. The first one had 3 hidden layers with layer, layer/2, and layer/4 number of neurons. The second architecture had 7 hidden layers with layer, layer*2, layer*2, layer, layer, layer/2, and layer/4 neurons. Here, the layer had 100, 200, 128, and 256 neurons. Finally, these models were tested using the hold-out test set from the source domain to identify the best-performing model concerning balanced accuracy (to account for class imbalance) for classification tasks and mean absolute error (MAE) for regression tasks. Then, these best-performing models were used as the pretrained models.

For fine-tuning, the pretrained model was loaded and the last hidden layer was replaced with a new hidden layer with randomly initialized weights. Then, all pretrained model layers were frozen (preventing those layers from learning) except for the newly added hidden layer and the model was trained to allow the new hidden layer to adjust its weights. Then, all layers were unfrozen (allowing weights to update) and the model was trained for the final time.

Prediction Performance Comparisons at Varying Levels of Data Scarcity

To investigate prediction performance at varying levels of data scarcity, random subsets of 1%, 5%, 10%, 25%, 50%, and 75% were created from the full training data set (100%). To avoid selection bias [31], each subset was obtained 10 different times using 10 different random states. Models were trained on these 10 data subsets and then performance metrics from all models were then aggregated for each subset. As there were 6 subsets (1%, 5%, 10%, 25%, 50%, and 75%) and the full training data

XSL•FO
RenderX

set of 100%, 61 (6×10+1) models were trained for each outcome and each model. For example, for AKI, 61 LR, 61 FCNN, 61 ITL, and 61 DA models were trained.

To obtain the median and 95% CI of the performance metrics, 1000 bootstrap samples of the test set were obtained for the full data set (100%), and for the random subsets, 100 bootstrap samples for each of the 10 random states (1000 in total) were created and then tested using these bootstrapped test sets.

All classification models were assessed using balanced accuracy as the primary metric (to account for class imbalance) and the following 4 secondary metrics on the hold-out test set: area under the receiver operating characteristic curve, accuracy, precision (also known as positive predictive value), and recall (also known as sensitivity). All regression models were evaluated using MAE and mean squared error (MSE).

Finally, Wilcoxon rank sum tests were performed to compare the performance of TL models to the baseline models. Since there were repeated comparisons involved, a Bonferroni correction was applied. Because the classification tasks had 35 comparisons (7 data subsets and 5 metrics), statistical significance was indicated by $P<.001$ (.05/35). The regression tasks had 14 comparisons (7 data subsets and 2 metrics), leading to statistical significance set at $P<.001$ (.05/14).

Results

Patient Cohorts

Based on the inclusion and exclusion criteria, the final eCritical and MIMIC-III cohorts were different for each patient outcome (Figure 2). The 30-day mortality cohort had 39,317 and 31,446 samples in eCritical and MIMIC-III databases, respectively, whereas the AKI cohort had 32,076 and 26,741 samples, respectively. The H_LOS cohort had 37,675 and 30,816 samples, and the ICU_LOS cohort had 38,529 and 30,816 samples, respectively. In the eCritical cohort, there were 6713 (17.07%) 30-day mortalities, whereas, in the MIMIC-III database, there were 3900 (12.40%) 30-day mortalities. The eCritical cohort had 4524 (14.11%) AKI cases whereas MIMIC-III had 5789 (21.64%) AKI cases. The eCritical cohort had a median H_LOS of 11.48 (IQR 5.59-23.29) days, whereas the MIMIC-III cohort had a median (IQR) of 7.39 (4.67-12.32). Similarly, the median (IQR) ICU_LOS in the eCritical and MIMIC-III cohorts were 3.97 (2.2-7.67) and 2.47 (1.59-4.58) days, respectively. The descriptive statistics for the 2 cohorts are shown in Table 2.

XSL•FO
RenderX

Table 2. Descriptive statistics of the 2 patient cohorts.

Descriptor	eCritical	MIMIC-III[a]
Male, n (%)	22,957 (58.39)	17,900 (56.92)
Age (years), median (IQR)	60 (46-70)	66 (53-78)
Admission weight (kg), median (IQR)	80 (67.4-96.6)	79.3 (66.5-94)
30-day mortality, n (%)	6713 (17.07)	3900 (12.40)
AKI[b], n (%)	4524 (14.10)	5789 (21.64)
H_LOS[c] (days), median (IQR)	11.48 (5.59-23.29)	7.39 (4.67-12.32)
ICU_LOS[d] (days), median (IQR)	3.97 (2.2-7.67)	2.47 (1.59-4.58)
Blood creatinine (µmol/L), median (IQR)	90.5 (65.5-152.05)	79.56 (61.88-123.76)
Glasgow Coma Scale, median (IQR)	11 (7.2-14.25)	12 (8-15)
Blood glucose (mmol/L), median (IQR)	7.49 (6.09-9.45)	7.1 (5.83-8.99)
Blood potassium (mmol/L), median (IQR)	3.95 (3.6-4.4)	4.07 (3.7-4.5)
Blood sodium (mmol/L), median (IQR)	138.2 (135.15-141)	138.5 (136-141)
Arterial pH, median (IQR)	7.38 (7.32-7.43)	7.38 (7.33-7.43)
WBC[e] (/µL), median (IQR)	12.03 (8.55-16.65)	11.2 (8.1-14.9)
RBC[f] (/µL), median (IQR)	3.7 (3.15-4.24)	3.54 (3.15-3.99)
Systolic blood pressure (mm Hg), median (IQR)	118.5 (100.25-139.8)	117 (100.8-135.75)
Diastolic blood pressure (mm Hg), median (IQR)	61.5 (52-73.8)	59.5 (49.5-71)
SpO_2[g] (%), median (IQR)	97 (94-99)	97.7 (95-99.7)
Hemoglobin (g/L), median (IQR)	112.45 (95.75-129.2)	107 (94.5-121)
Hematocrit (%), median (IQR)	34 (29-39)	31.5 (27.95-35.5)
Mechanical ventilation, n (%)	34,073 (86.66)	16,429 (52.25)
Dialysis, n (%)	2173 (5.53)	6909 (21.97)
Norepinephrine, n (%)	15,797 (40.18)	3,339 (10.62)
Phenylephrine, n (%)	4384 (11.15)	6737 (21.42)
Vasopressin, n (%)	5087 (12.94)	712 (2.26)
Dobutamine, n (%)	754 (1.92)	429 (1.36)
Dopamine, n (%)	1051 (2.67)	1512 (4.81)
Epinephrine, n (%)	1320 (3.36)	1221 (3.88)

[a]MIMIC-III: Medical Information Mart for Intensive Care III.
[b]AKI: acute kidney injury.
[c]H_LOS: hospital length of stay.
[d]ICU_LOS: intensive care unit length of stay.
[e]WBC: white blood cell.
[f]RBC: red blood cell.
[g]SpO_2: oxygen saturation.

Pretrained Models

After hyperparameter tuning for the 30-day mortality source task, the pretrained model with 3 hidden layers of 128, 64, and 32 neurons was selected. This model had the highest balanced accuracy of 0.7810. This was the pretrained model for all 4 ITL target tasks and the 30-day mortality DA target task. Similarly, after hyperparameter tuning for the AKI, H_LOS, and ICU_LOS source tasks, pretrained models with 3 hidden layers of 256, 128, and 64 neurons, 7 hidden layers of 256, 512, 512, 256, 256, 128, and 64 neurons, and 7 hidden layers of 256, 512, 512, 256, 256, 128, and 64 neurons, with a balanced accuracy of 0.7199, an MAE of 11.8019, and an MAE of 3.0887, were selected, respectively. These were the pretrained models for the DA target tasks for AKI, H_LOS, and ICU_LOS, respectively.

These pretrained models are publicly available via GitHub [32].

XSL•FO
RenderX

Domain Adaptation

Multimedia Appendices 1-4 show the complete prediction performances of the DA and baseline models at varying levels of data scarcity represented by the data subsets for 30-day mortality, AKI, ICU_LOS, and H_LOS, respectively. Figures 3-6 pictorially compare the DA and baseline models for each patient outcome.

Figure 3. 30-day mortality prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% CIs. AUC: area under the receiver operating characteristic curve; DA: domain adaptation; FCNN: fully connected neural network; ITL: inductive transfer learning; LR: logistic regression.

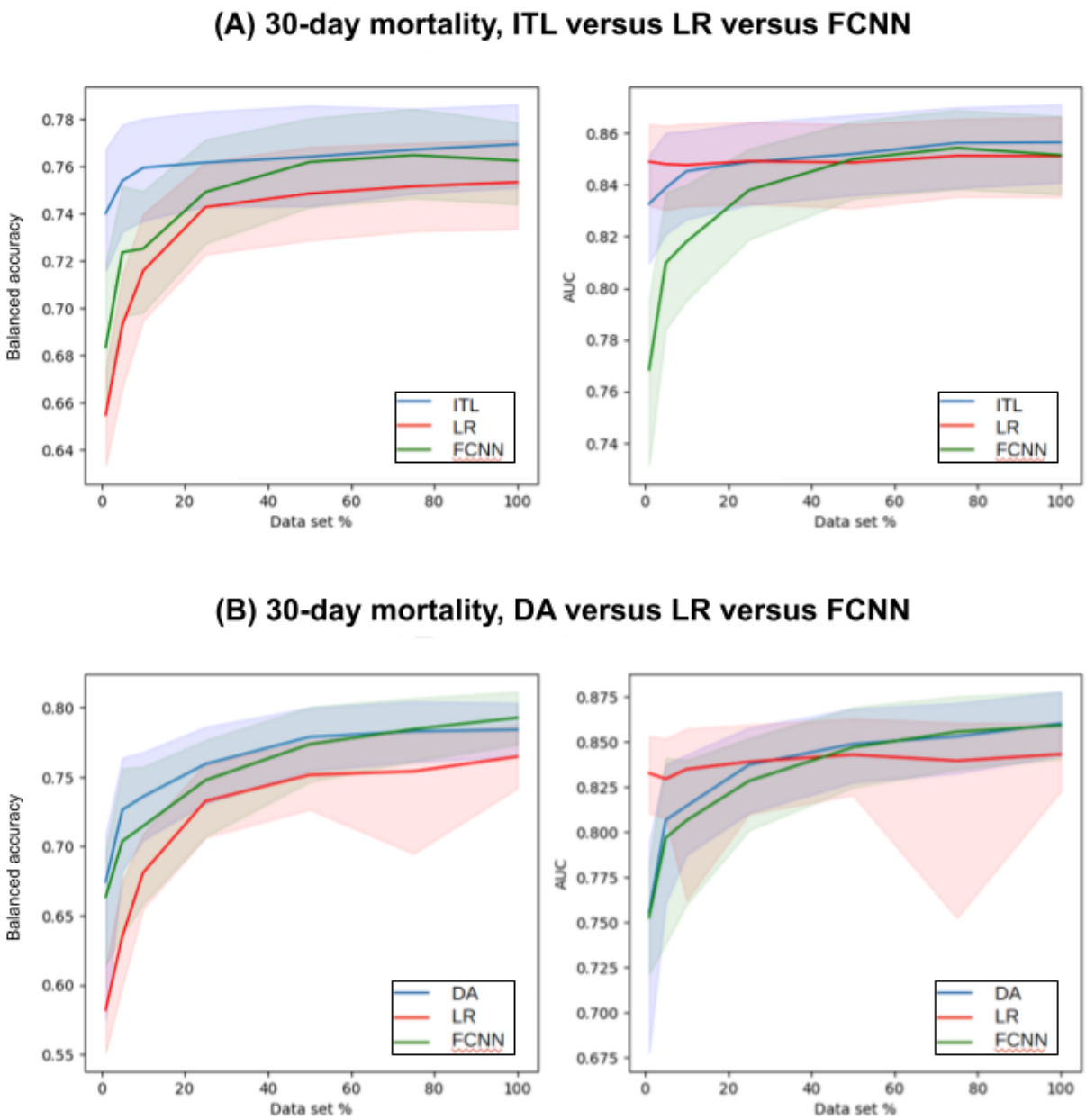

XSL•FO
RenderX

Figure 4. AKI prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% CIs. AKI: acute kidney injury; AUC: area under the receiver operating characteristic curve; DA: domain adaptation; FCNN: fully connected neural network; ITL: inductive transfer learning; LR: logistic regression.

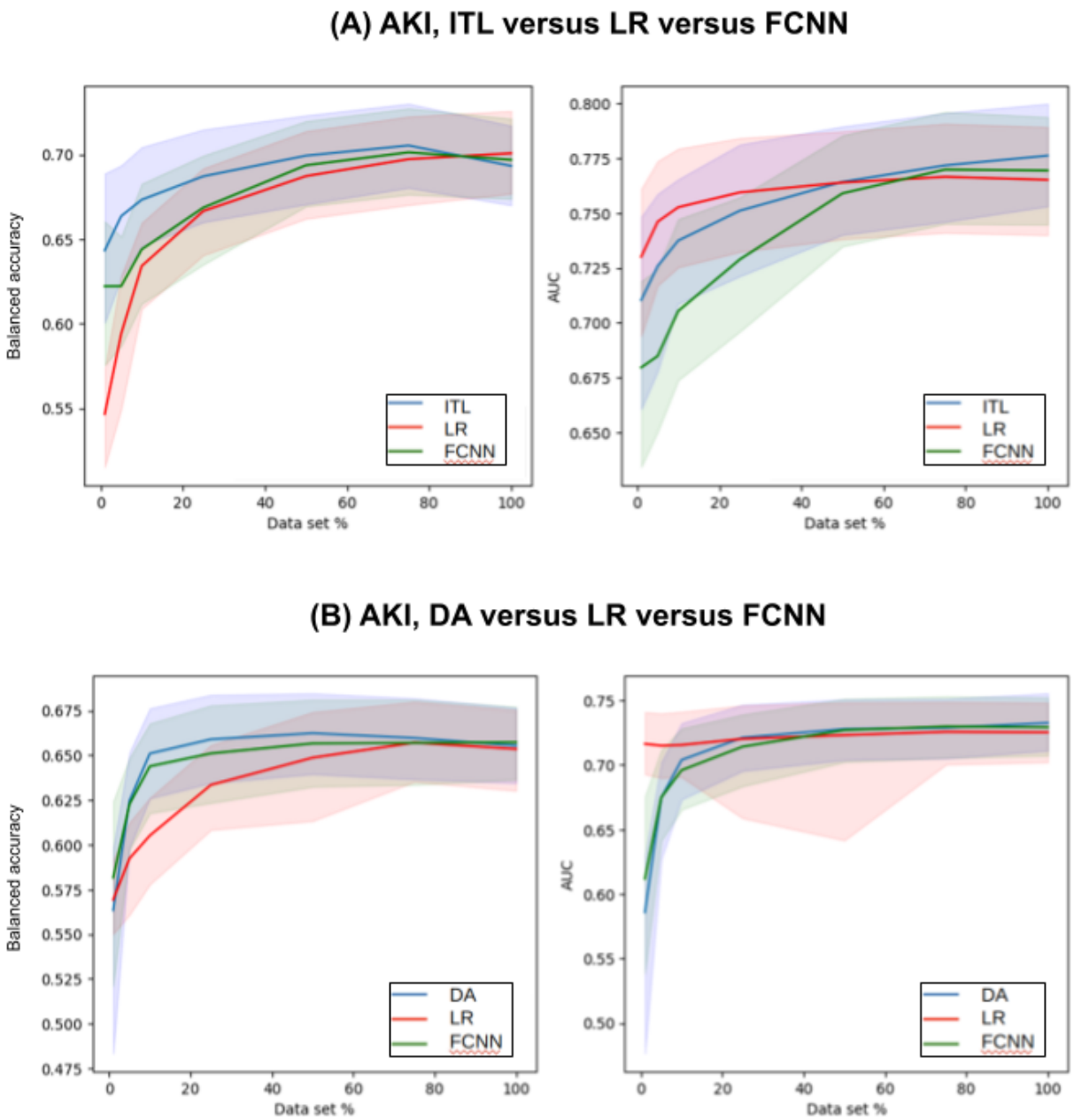

XSL•FO
RenderX

Figure 5. ICU_LOS prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% CIs. DA: domain adaptation; FCNN: fully connected neural network; ICU_LOS: intensive care unit length of stay; ITL: inductive transfer learning; MAE: mean absolute error; MSE: mean squared error.

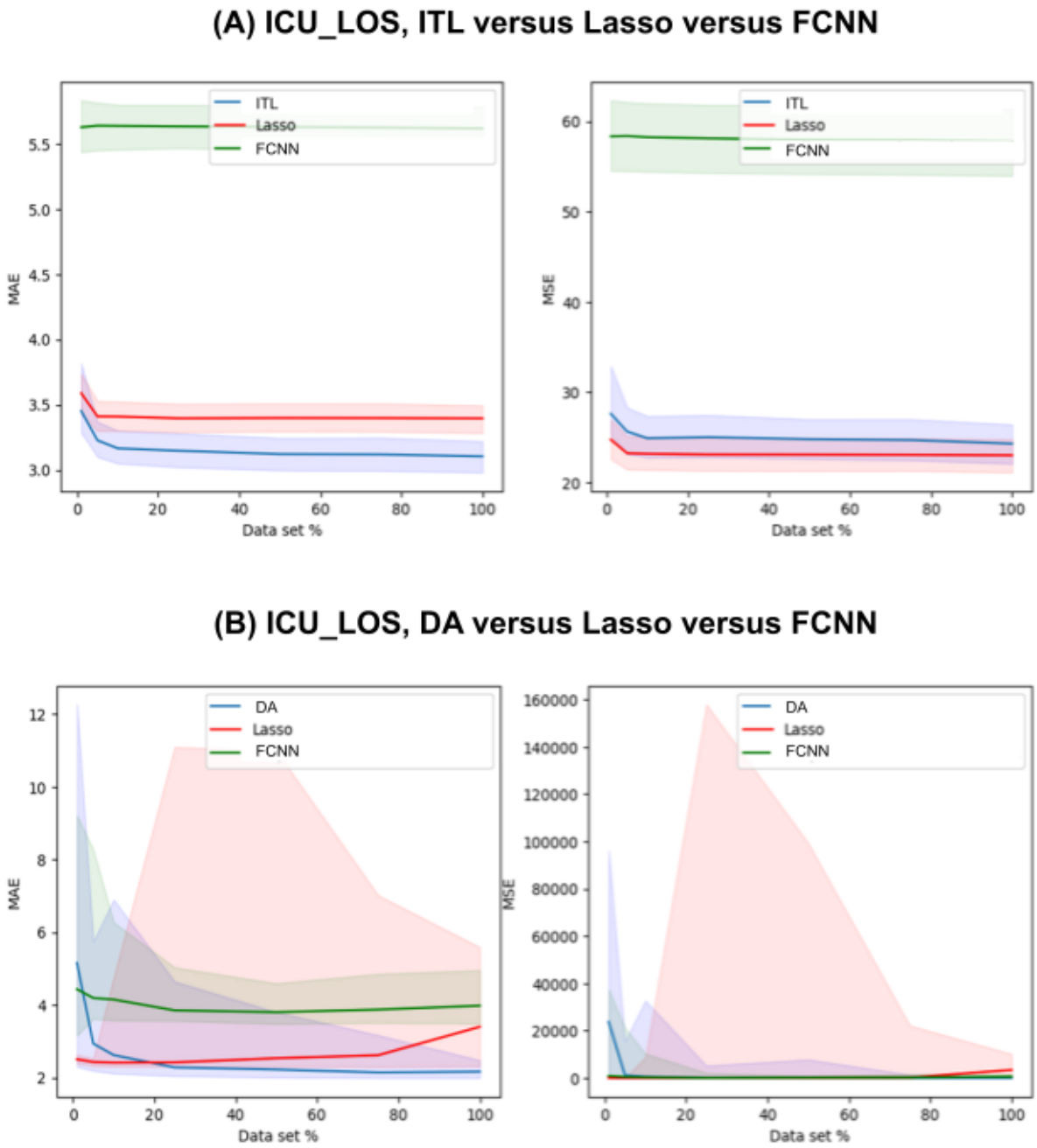

XSL•FO
RenderX

Figure 6. H_LOS prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% CIs. DA: domain adaptation; FCNN: fully connected neural network; H_LOS: hospital length of stay; ITL: inductive transfer learning; MAE: mean absolute error; MSE: mean squared error.

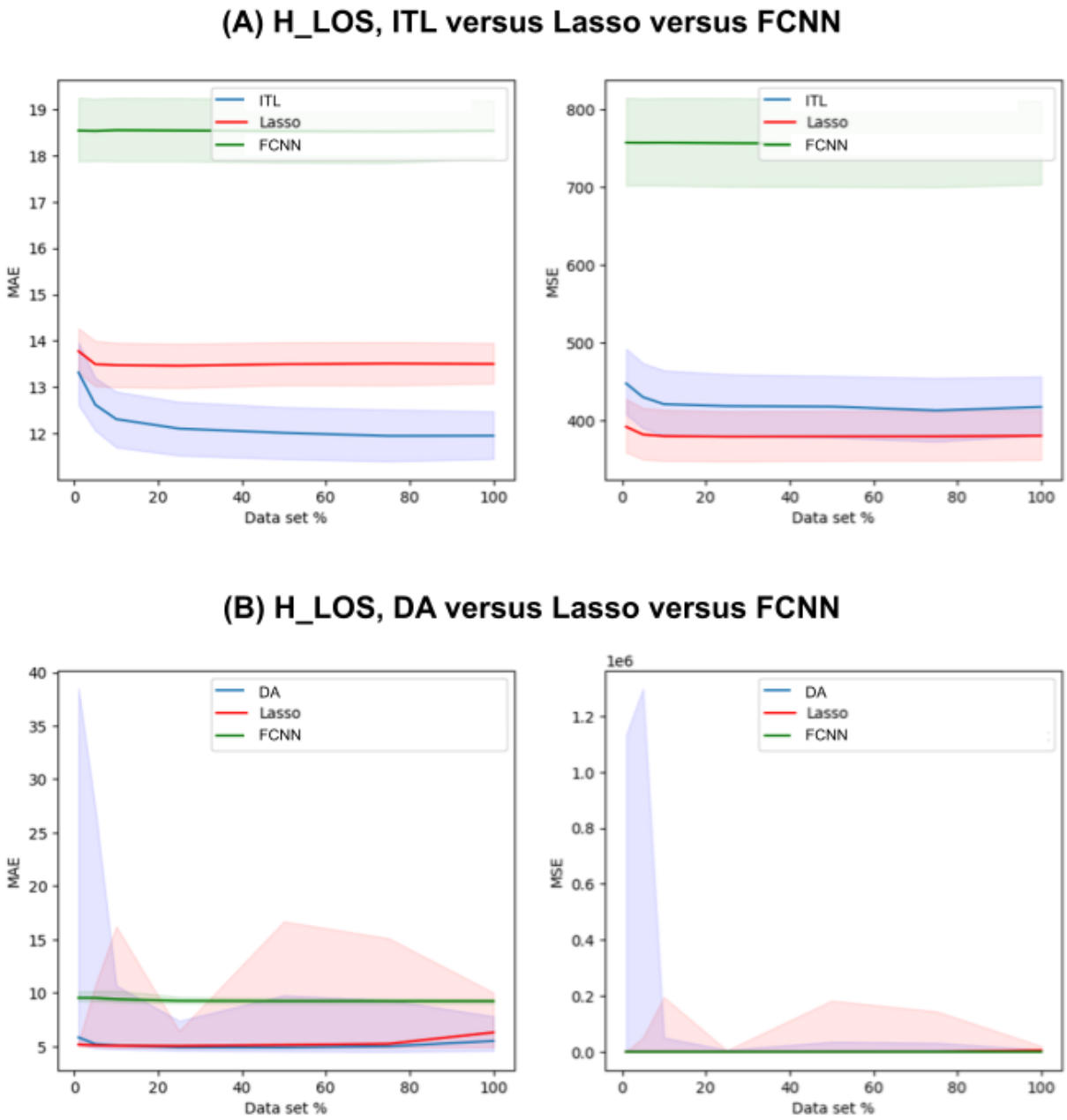

For 30-day mortality, DA models outperformed both the baseline models LR and FCNN for data subsets 1% to 50%. For data sets, 75% and 100% DA models outperformed the LR model but underperformed the FCNN model. For example, when 1% data set was used for training, the DA model had a median balanced accuracy of 0.6744 (95% CI 0.5758-0.7083), whereas LR had 0.5821 (95% CI 0.551-0.6134) and FCNN had 0.6636 (95% CI 0.6146-0.6971).

For AKI, the DA models outperformed both baseline models for some of the data subsets (75%, 50%, 25%, 10%, and 5%) and underperformed both baseline models for the data subset (1%). The DA model outperformed the LR model and underperformed the FCNN model for the 100% data set. When the 10% data subset was used, the DA model had a median balanced accuracy of 0.6511 (95% CI 0.626-0.6763), whereas the LR model had 0.6052 (95% CI 0.5779-0.6262) and the FCNN model had 0.6439 (95% CI 0.6177-0.6678).

For ICU_LOS, the DA models outperformed both baseline models for some of the data subsets (25%-100%). The DA models outperformed the FCNN models but underperformed the lasso models in some cases (5% and 10%). In addition, the results from the 1% data subset were not significant between DA and FCNN (P=.05). When the 25% training data subset was used for training, the DA model had a median MAE of 2.2781 (95% CI 2.0427-4.643), whereas the lasso model had 2.4165

XSL•FO
RenderX

(95% CI 2.2967-11.1001) and the FCNN model had 3.8481 (95% CI 3.5641-5.0331).

For H_LOS, the DA models outperformed both baseline models for some of the data subsets (25%-100%). The DA models outperformed the FCNN models but underperformed the lasso models in some cases (1% and 5%). The results from the 10% data subset were not significantly different between DA and lasso (P*P*.02), but the DA model outperformed the FCNN model. When the 25% data subset was used for training, the DA model had a median MAE of 4.9109 (95% CI 4.5982-7.389), whereas the lasso model had 5.0491 (95% CI 4.8903-6.457) and the FCNN model had 9.2677 (95% CI 9.0162-9.6332).

Inductive Transfer Learning

Multimedia Appendices 3-6 show the complete prediction performances of the ITL and baseline models at varying levels of data scarcity represented by the data subsets for 30-day mortality, AKI, ICU_LOS, and H_LOS, respectively. Figures 3-6 pictorially compare the ITL and baseline models for each patient outcome.

As mentioned in the Methods section, the 30-day mortality prediction results from ITL presented in Multimedia Appendix 5 serve as a benchmark, since the source and target domains and prediction tasks were the same. The fast convergence of ITL performance at very small data subsets shown in Figure 3 corroborates the limited learning taking place during fine-tuning.

For AKI, the ITL models outperformed both baseline models for all data subsets except 100%. For example, when the 1% data set was used for training the models, the ITL model had a median balanced accuracy of 0.6434 (95% CI 0.6006-0.6888), whereas the LR model had 0.5467 (95% CI 0.5154-0.5732) and the FCNN model had 0.6222 (95% CI 0.5757-0.6604).

For ICU_LOS, the ITL models outperformed both baseline models for all the data subsets. For example, when the 1% data set was used, the ITL model had a median MAE of 3.4519 (95% CI 3.2863-3.8158), whereas lasso had 3.5883 (95% CI 3.4255-3.7376) and FCNN had 5.626 (95% CI 5.4351-5.8329).

For H_LOS, the ITL models outperformed both baseline models for all data subsets. For example, when the 1% data set was used, the ITL model had a median MAE of 13.3182 (95% CI 12.6128-13.9609) whereas lasso had 13.7765 (95% CI 13.3118-14.2661) and FCNN had 18.5363 (95% CI 17.8711-19.243).

Discussion

Principal Results

Overall, the ITL models outperformed the baseline models in 55 of the 56 cases (7 data subsets × 4 outcomes × 2 baseline models). The DA models outperformed the baseline models 45 times in 56 cases. While TL is expected to yield better prediction performance than the baseline models, particularly with small data subsets, this assumption has not been confirmed in a comprehensive manner yet in the context of EHR-based ICU patient outcome prediction. In particular, the ITL prediction performances reported in this study are important contributions, given that ITL has seldom been investigated in EHR-based studies in general.

Moreover, the results from this study characterize DA, ITL, and baseline prediction performances as a function of target data volume. ICUs with limited data or computing resources can use our results as a guide to decide which would be better: fine-tune our pretrained models on their data or train new models from scratch using their small data set.

It is also worth noting that DA did not always outperform the baseline models even at very small data subsets (eg, ICU_LOS at the 1% subset; Multimedia Appendix 3). This finding implies that one should not blindly apply TL even when the target data set is small and expect performance improvement.

The ITL models performed better than the DA models in terms of the number of times and the margin with which they outperformed the baseline models. This speaks to the fact that the eCritical and MIMIC-III cohorts are quite different, and the knowledge learned from eCritical exhibited limited use in predicting the outcomes of the MIMIC-III patients. This is corroborated by the substantial differences in all 4 outcomes shown in Table 2. There seem to be more similarities between different outcomes within the same cohort than between different cohorts for the same outcome. This finding may be surprising to many researchers since the differences between different patient outcomes in terms of disease progression and risk factors are believed to be substantial, whereas even at different sites, the fundamentals of the diseases and ICU care should have many similarities.

Our pretrained models have been made available publicly which can be used at other ICUs or in future research. In many scenarios, TL was useful even at the 1% data subset, representing only about 200 samples. Hence, ICUs with their own data set containing just 200 samples can potentially benefit from our pretrained models. Given the paucity of public pretrained models in EHR-based ICU patient outcome prediction, our pretrained models are an invaluable contribution to the field.

Clinical Implications

While accurate patient outcome prediction in the ICU has the potential to enable early initiation of preventative care and improve clinical efficiency and care resource management, it may be infeasible for many ICUs to build their own predictive models due to a lack of digital data infrastructure. The pretrained models from this study address this barrier by serving as predictive models that can be used out-of-the-box (albeit with suboptimal performance) or fine-tuned with a small amount of local data. This study provides a pathway for a wider group of ICUs to consider bringing patient outcome prediction models to the point of care.

Comparisons With the Literature

This study has several strengths in comparison with previous EHR-based TL studies. First, this study is one of the few studies that explored ITL using EHR data. Second, this study used 2 large cohorts: eCritical with 55,689 ICU admissions from 48,672 patients as the source domain, and MIMIC-III with 61,532 ICU

XSL•FO
RenderX

admissions from 46,476 patients as the target domain. These are considerably larger than the cohorts used in previous DA studies. For example, Shickel et al [24] used the conventional ICU cohort at the University of Florida Health with 48,400 distinct ICU admissions as the source domain and the intelligent ICU cohort at the University of Florida Health with only 51 ICU admissions as the target cohort. Third, this study used a large feature set of 104. In comparison, Shickel et al [24] used only 9 features, while Li et al [22] used 4 features. Fourth, our pretrained models are able to predict outcomes for new patients without previous admissions, unlike the natural language processing models developed by Li et al [22].

Secondary Performance Metric Results

While the TL models outperformed the baseline models in general with respect to the primary performance metrics (balanced accuracy and MAE), the baseline models (particularly LR and lasso) often outperformed the TL models in terms of the secondary metrics. In mortality and AKI prediction, the LR models tended to show higher precisions and lower recalls than the TL models. Given that both mortality and AKI exhibited low event rates leading to substantial class imbalance, the precision and recall results indicate that the LR models were more biased toward the majority class than the TL models. The higher area under the receiver operating characteristic curves from the LR models indicates that they achieved higher specificities in general than the TL models, further corroborating the bias toward the majority class. This is why we chose balanced accuracy as our primary metric since it reflects the performances of both the majority and minority classes.

In the regression tasks of ICU_LOS and H_LOS prediction, the lasso models often yielded better results than the TL models in terms of MSE. This implies that the TL models often led to larger errors that were amplified by the squaring effect of MSE. Similarly, the large confidence intervals in Figures 5B and 6B are also likely due to occasional large errors caused by the fact that the model output is not upper-bounded.

Limitations

This study has limitations. First, we could not include all available features due to the constraint of having to use common features in both eCritical and MIMIC-III. Second, this study could not investigate other major ICU patient outcomes, such as sepsis, delirium, and acute respiratory distress syndrome, due to data unavailability in either or both eCritical and MIMIC-III. Third, only 2 baseline models were investigated per prediction task, and more advanced ML models (eg, XGBoost [XGBoost Contributors]) were not used. However, because the primary objective of this study was to demonstrate the benefits of TL, the most appropriate benchmark models were the FCNNs where everything was equivalent to the TL models except for the use of a pretrained model. While our focus was not to produce the best prediction performance, our results are comparable to the best MIMIC-based results in the literature as shown by the review study conducted by Syed et al [33]. Fourth, this study only examined the discrimination of the prediction models and did not investigate calibration. While many health ML studies focus only on discrimination and neglect calibration [34], this remains an important limitation of this study. Finally, even though some of the features, such as vitals and laboratory findings, were longitudinal, we performed a cross-sectional study via feature aggregation. More advanced recurrent deep learning models (eg, long short-term memory and gated recurrent unit) that can leverage longitudinal information may have led to different results.

Future Work

First, future work can include an investigation of DA and ITL on other ICU patient outcomes, such as sepsis and delirium, using data sets that can support such research. Second, the application of DA and ITL to more advanced recurrent deep learning models would be worthwhile. Third, the effectiveness of pretrained models in the combination of both DA and ITL (ie, both the domain and prediction task would change from source to target) remains to be studied. For example, a mortality prediction model pretrained on a source data set can be fine-tuned on a target data set from a different domain to predict AKI. Finally, the effectiveness of TL across patient subgroups with respect to demographics and socioeconomic status would be worthwhile investigating.

Conclusions

In this retrospective study, we found that TL can lead to improved prediction performance when compared to baseline models trained from scratch only using target data. This performance improvement was observed in a wide range of simulated data scarcity. In addition, the performance of ITL was superior to that of DA. This implies fine-tuning a pretrained model to predict a different patient outcome within the same domain would be a promising approach. We hope that the pretrained models from this study are useful to other researchers and ICUs.

Acknowledgments

MKM was supported by an Alberta Innovates Graduate Student Scholarship for Data-Enabled Innovation from Alberta Innovates and Alberta Advanced Education, as well as a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-02588). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The authors would like to thank Dr Filipe Lucini for assistance with data retrieval, and Alberta Health Services for granting access to the eCritical and administrative health databases. Finally, advanced research computing and medical advanced research computing clusters used in this project were provided by the University of Calgary.

XSL•FO
RenderX

Data Availability

The Medical Information Mart for Intensive Care (MIMIC) data sets analyzed during this study are available in the MIMIC-III repository [35]. The eCritical data sets analyzed during this study are not publicly available due to their patient-identifiable information but are available from the corresponding author upon reasonable request. The code that conducted the experiments, as well as the pretrained models, are available on GitHub [32].

Authors' Contributions

MKM designed the study, prepared the data, conducted all experiments, and wrote the manuscript. JL conceived and designed the study, and directly supervised MKM. HTS facilitated accessing the eCritical data. All authors contributed to revisions of the manuscript.

Conflicts of Interest

JL is a cofounder and a major shareholder of Symbiotic AI, Inc.

Multimedia Appendix 1
Thirty-day mortality prediction performances of domain adaptation and baseline models across all data subsets.
[DOCX File , 24 KB - jmir_v26i1e52730_app1.docx]

Multimedia Appendix 2
Acute kidney injury prediction performances of domain adaptation and baseline models across all data subsets.
[DOCX File , 24 KB - jmir_v26i1e52730_app2.docx]

Multimedia Appendix 3
Intensive care unit length of stay prediction performances of inductive transfer learning, domain adaptation, and baseline models across all data subsets.
[DOCX File , 24 KB - jmir_v26i1e52730_app3.docx]

Multimedia Appendix 4
Hospital length of stay prediction performances of inductive transfer learning, domain adaptation, and baseline models across all data subsets.
[DOCX File , 24 KB - jmir_v26i1e52730_app4.docx]

Multimedia Appendix 5
Thirty-day mortality prediction performances of inductive transfer learning and baseline models across all data subsets.
[DOCX File , 25 KB - jmir_v26i1e52730_app5.docx]

Multimedia Appendix 6
Acute kidney injury prediction performances of inductive transfer learning and baseline models across all data subsets.
[DOCX File , 25 KB - jmir_v26i1e52730_app6.docx]

References

1. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
2. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798-1828. [doi: 10.1109/TPAMI.2013.50] [Medline: 23787338]
3. Niu S, Liu Y, Wang J, Song H. A decade survey of transfer learning (2010-2020). IEEE Trans Artif Intell 2020;1(2):151-166. [doi: 10.1109/tai.2021.3054609]
4. Tokuoka Y, Suzuki S, Sugawara Y. An inductive transfer learning approach using cycle-consistent adversarial domain adaptation with application to brain tumor segmentation. New York, NY, USA: ACM; 2019 Presented at: Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering; November 13-15, 2019; Shanghai, China p. 44-48. [doi: 10.1145/3375923.3375948]
5. Titoriya A, Sachdeva S. Breast cancer histopathology image classification using AlexNet. 2019 Presented at: 2019 4th International Conference on Information Systems and Computer Networks (ISCON); November 21-22, 2019; Mathura, India p. 708-712. [doi: 10.1109/iscon47742.2019.9036160]

XSL•FO
RenderX

6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84-90. [doi: 10.1145/3065386]
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV, USA p. 770-778. [doi: 10.1109/cvpr.2016.90]
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv Preprint posted online on September 4, 2014 [FREE Full text]
9. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 7-12, 2015; Boston. [doi: 10.1109/cvpr.2015.7298594]
10. Kumar KA, Prasad A, Metan J. A hybrid deep CNN-Cov-19-Res-Net transfer learning architype for an enhanced brain tumor detection and classification scheme in medical image processing. Biomed Signal Process Control 2022;76:103631. [doi: 10.1016/j.bspc.2022.103631]
11. Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N, Marchiori E, Pesteie M, et al. Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. 2017 Presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; September 11-13, 2017; Quebec City, QC, Canada p. 516-524. [doi: 10.1007/978-3-319-66179-7_59]
12. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35(5):1285-1298 [FREE Full text] [doi: 10.1109/TMI.2016.2528162] [Medline: 26886976]
13. Diamant I, Bar Y, Geva O, Wolf L, Zimmerman G, Lieberman S, et al. Chest radiograph pathology categorization via transfer learning. In: Deep Learning for Medical Image Analysis. London: Academic Press; 2017:299-320.
14. Sugeno A, Ishikawa Y, Ohshima T, Muramatsu R. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. Comput Biol Med 2021;137:104795. [doi: 10.1016/j.compbiomed.2021.104795] [Medline: 34488028]
15. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv Preprint posted online on January 16, 2013 [FREE Full text] [doi: 10.3126/jiee.v3i1.34327]
16. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. USA: Association for Computational Linguistics; 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October, 2014; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]
17. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis p. 4171-4186. [doi: 10.18653/v1/N19-1423]
18. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText.zip: compressing text classification models. ArXiv Preprint posted online on December 12, 2016 [FREE Full text]
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep Learning-based text classification. ACM Comput Surv 2022;54(3):1-40. [doi: 10.1145/3439726]
21. Gao Z, Feng A, Song X, Wu X. Target-dependent sentiment classification with BERT. IEEE Access 2019;7:154290-154299. [doi: 10.1109/access.2019.2946594]
22. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep 2020;10(1):7155 [FREE Full text] [doi: 10.1038/s41598-020-62922-y] [Medline: 32346050]
23. Liu K, Zhang X, Chen W, Yu ASL, Kellum JA, Matheny ME, et al. Development and validation of a personalized model with transfer learning for acute kidney injury risk estimation using electronic health records. JAMA Netw Open 2022;5(7):e2219776 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.19776] [Medline: 35796212]
24. Shickel B, Davoudi A, Ozrazgat-Baslanti T, Ruppert M, Bihorac A, Rashidi P. Deep multi-modal transfer learning for augmented patient acuity assessment in the intelligent ICU. Front Digit Health 2021;3:640685 [FREE Full text] [doi: 10.3389/fdgth.2021.640685] [Medline: 33718920]
25. Harsono IW, Liawatimena S, Cenggoro TW. Lung nodule detection and classification from thorax CT-scan using RetinaNet with transfer learning. J King Saud Univ Comput Inf Sci 2022;34(3):567-577. [doi: 10.1016/j.jksuci.2020.03.013]
26. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
27. Kellum JA, Lameire N. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl 2012;2(1):1-138 [FREE Full text] [doi: 10.1038/sdata.2016.35]
28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]

XSL•FO
RenderX

29. Ge W, Huh J, Park YR, Lee J, Kim Y, Turchin A. An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units. AMIA Annu Symp Proc 2018;2018:460-469 [FREE Full text] [Medline: 30815086]
30. Hepp T, Schmid M, Gefeller O, Waldmann E, Mayr A. Approaches to regularized regression—a comparison between gradient boosting and the lasso. Methods Inf Med 2016;55(5):422-430. [doi: 10.3414/ME16-01-0033] [Medline: 27626931]
31. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. Nephron Clin Pract 2010;115(2):c94-c99 [FREE Full text] [doi: 10.1159/000312871] [Medline: 20407272]
32. Mutnuri M. ICU_outcome_prediction_Transfer_Learning. URL: https://github.com/data-intelligence-for-health-lab/ICU_outcome_prediction_Transfer_Learning [accessed 2023-09-12]
33. Syed M, Syed S, Sexton K, Syeda HB, Garza M, Zozus M, et al. Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: systematic review. Informatics (MDPI) 2021;8(1):16 [FREE Full text] [doi: 10.3390/informatics8010016] [Medline: 33981592]
34. Staartjes V, Kernbach J. Letter to the editor. Importance of calibration assessment in machine learning-based predictive analytics. J Neurosurg Spine 2020;32(6):985-987. [doi: 10.3171/2019.12.SPINE191503] [Medline: 32084640]
35. Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database. URL: https://physionet.org/content/mimiciii/1.4/ [accessed 2023-09-12]

Abbreviations

AKI: acute kidney injury
DA: domain adaptation
EHR: electronic health record
FCNN: fully connected neural network
H_LOS: hospital length of stay
ICU: intensive care unit
ICU_LOS: intensive care unit length of stay
ITL: inductive transfer learning
LR: logistic regression
MAE: mean absolute error
MIMIC-III: Medical Information Mart for Intensive Care III
ML: machine learning
MSE: mean squared error
TL: transfer learning

Edited by A Mavragani; submitted 14.09.23; peer-reviewed by C Sun, Z Su; comments to author 18.12.23; revised version received 22.03.24; accepted 08.07.24; published 21.08.24.

Please cite as:
Mutnuri MK, Stelfox HT, Forkert ND, Lee J
Using Domain Adaptation and Inductive Transfer Learning to Improve Patient Outcome Prediction in the Intensive Care Unit: Retrospective Observational Study
J Med Internet Res 2024;26:e52730
URL: https://www.jmir.org/2024/1/e52730
doi:10.2196/52730
PMID:

©Maruthi Kumar Mutnuri, Henry Thomas Stelfox, Nils Daniel Forkert, Joon Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predictors of Medical and Dental Clinic Closure by Machine Learning Methods: Cross-Sectional Study Using Empirical Data

Young-Taek Park[1], PhD; Donghan Kim[2], PhD; Ji Soo Jeon[3], MS; Kwang Gi Kim[3,4], PhD

[1]HIRA Research Institute, Health Insurance Review & Assessment Service, Wonju-si, Republic of Korea

[2]Center for Geospatially Enabled Society, Korea Research Institute for Human Settlements, Sejong-si, Republic of Korea

[3]Department of Biomedical Engineering, College of Medicine, Gil Medical Center, Gachon University, Inchon, Republic of Korea

[4]KMAIN Corp, Seongnam-si, Republic of Korea

Corresponding Author:
Kwang Gi Kim, PhD
Department of Biomedical Engineering
College of Medicine, Gil Medical Center
Gachon University
58-13 Docjemro
NamdongGum
Inchon, 21565
Republic of Korea
Phone: 82 324582770
Email: kimkg@gachon.ac.kr

Abstract

Background: Small clinics are important in providing health care in local communities. Accurately predicting their closure would help manage health care resource allocation. There have been few studies on the prediction of clinic closure using machine learning techniques.

Objective: This study aims to test the feasibility of predicting the closure of medical and dental clinics (MCs and DCs, respectively) and investigate important factors associated with their closure using machine running techniques.

Methods: The units of analysis were MCs and DCs. This study used health insurance administrative data. The participants of this study ran and closed clinics between January 1, 2020, and December 31, 2021. Using all closed clinics, closed and run clinics were selected at a ratio of 1:2 based on the locality of study participants using the propensity matching score of logistic regression. This study used 23 and 19 variables to predict the closure of MCs and DCs, respectively. Key variables were extracted using permutation importance and the sequential feature selection technique. Finally, this study used 5 and 6 variables of MCs and DCs, respectively, for model learning. Furthermore, four machine learning techniques were used: (1) logistic regression, (2) support vector machine, (3) random forest (RF), and (4) Extreme Gradient Boost. This study evaluated the modeling accuracy using the area under curve (AUC) method and presented important factors critically affecting closures. This study used SAS (version 9.4; SAS Institute Inc) and Python (version 3.7.9; Python Software Foundation).

Results: The best-fit model for the closure of MCs with cross-validation was the support vector machine (AUC 0.762, 95% CI 0.746-0.777; $P<.001$) followed by RF (AUC 0.736, 95% CI 0.720-0.752; $P<.001$). The best-fit model for DCs was Extreme Gradient Boost (AUC 0.700, 95% CI 0.675-0.725; $P<.001$) followed by RF (AUC 0.687, 95% CI 0.661-0.712; $P<.001$). The most significant factor associated with the closure of MCs was years of operation, followed by population growth, population, and percentage of medical specialties. In contrast, the main factor affecting the closure of DCs was the number of patients, followed by annual variation in the number of patients, year of operation, and percentage of dental specialists.

Conclusions: This study showed that machine running methods are useful tools for predicting the closure of small medical facilities with a moderate level of accuracy. Essential factors affecting medical facility closure also differed between MCs and DCs. Developing good models would prevent unnecessary medical facility closures at the national level.

(J Med Internet Res 2024;26:e46608) doi:10.2196/46608

XSL•FO
RenderX

KEYWORDS

machine learning; health facility closure; hospital closure; clinic closure; clinic bankruptcy; hospital bankruptcy; health clinic; prediction; healthcare resources; artificial intelligence; medical clinic; health insurance; health facility closure

Introduction

Small medical facilities, such as hospitals and clinics, are critical in providing local community residents with health care. Approximately 6100 hospitals operate in the United States, and 35% are located in rural areas [1]. In the United Kingdom, about 1900 hospitals exist, and many small hospitals provide medical care to half the population [2,3]. Korea has 1400 small hospitals, 34,000 medical clinics (MCs), and 18,800 dental clinics (DCs) as of October 2022 [4].

The closure of small medical facilities may result in a lack of health care provision and thus bring about a sustainability issue in the community [5,6]. Small hospitals could be more critically affected than large hospitals by an equal amount of external impacts [7]. According to a study in the United States, many hospitals have closed, averaging 21 hospitals annually between 2010 and 2015 [6], and the closure of those hospitals is continuing [8,9]. In Germany, 1 small hospital closes every month [10]. In Korea, the annual closure rate of small hospitals and MCs has been reported to be 5.8% and 3.4%, respectively [11]. For these reasons, preventing the closure of medical facilities is important for the sustainability of health care systems.

Meanwhile, owing to the advancement of information technologies and statistical methods, many artificial intelligence (AI) methods, such as machine learning, have been applied to predict events in health care. They have been used for the prediction of bankruptcy [12-14], timely attendance and no-shows in medical appointments [15,16], hospital admissions and discharges, and hospital length of stay [17-20]. Although there are several studies on hospital or clinic closure [21-24], no study predicts the closure of MCs or DCs using AI methods and empirical data.

Regarding this study on the closure of hospitals or clinics using machine learning methods, there was 1 case where machine learning methods were used to study the closure of small hospitals. A study conducted in the United States investigated the predictive factors related to hospital bankruptcy using AI methods and found that various factors such as facility age, market concentration, and Medicare percentage were associated with hospital bankruptcy. This study used AI techniques such as a linear support vector machine (SVM) model with a hinge function, a perceptron neural network model, and so on [24]. A recent study predicted an annual number of hospital patients using several machine learning techniques and found that the most important factor associated with a predicted number of inpatients was the number of beds, followed by the number of nurses; in contrast, the best predictive factor for the outpatient was several doctors followed by several local households [25]. When asked about the important factors for an annual number of hospital patients, most people would say “physicians” or “nurses.” However, machine learning suggests that the number of beds and nurses is important in the inpatient section because the number of patients is produced through beds and by nurses. In contrast, the number of doctors is important in the outpatient section because they sequentially care about patients. Machine learning techniques also pointed out that the number of households in local areas is important for predicting annual outpatients.

Various factors have been known to be associated with the closure of medical facilities, such as affiliation status and financial distress [22,26,27], profitability [28], size of hospital and locality [29], and market competition [30]. However, they used traditional statistical methods such as logistic regression (LR) rather than machine learning. We can develop a good model that predicts the closure of hospitals and MCs with advanced technologies. In that case, we can help health care policy makers improve the sustainability of health care systems and maintain healthy local medical communities by targeting those factors with appropriate policies.

This study is important from several perspectives. This study is the first to use machine learning to predict the closure of MCs and DCs and study closure-related factors. The results expand the scope of business closure research from current hospitals to clinics, DCs, pharmacies, nursing homes, and long-term care hospitals. Second, this study will be of particular interest to those who want to start a new business in the health care field. In many countries, every year, many medical professionals and pharmacists graduate from universities or colleges and strive to open their clinics. This study provides support for using AI to identify suitable areas for clinics and pharmacies in their business plan. Finally, the results obtained from this study help to fill the research gap between good prediction and reality, which will contribute to developing good methodological models predicting the closure of health care institutions.

Therefore, this study aimed to predict the closure of MCs and DCs, including selecting the best model and identifying significant factors critically affecting this process. The organizational characteristics and knowledge gained from this study could provide health care policy makers and international colleagues with useful information to build sustainable and healthy health care delivery systems.

Methods

Study Setting

This study is conducted in Korea, as it has many medical facilities, and closures occur frequently. Therefore, using AI methods, Korea serves as a good experimental setting to observe and predict the closure of medical facilities. Institutionally, Korea has adopted the national health insurance system. For the same types of medical care, each medical facility receives an equal amount of reimbursement because of the setting of the national health insurance systems. There are also many private and public medical facilities, such as small hospitals, MCs, and DCs. Patients can choose any primary care clinic without

XSL•FO
RenderX

constriction. Judging from these facts, the market is very competitive for survival, and dynamic factors may affect the closure of medical facilities. This study targeted MCs and DCs as study participants because there are many clinics and frequent occurrences of events such as clinic closures. Approximately 30,000 MCs and 25,000 DCs are operating as of December 31, 2022. Owing to the high competition, there are several closures.

Regarding the recruitment settings and procedures of study participants, this study used a research data set from the Health Insurance Review and Assessment Service (HIRA). The HIRA is a third-party administrator providing professional medical claim review and assessment services to Korea's National Health Insurance Service. Every health care organization, such as hospitals and clinics, should register or report their openness and closure of facility to the government through a Unified Portal Reporting Healthcare Resource [31]. This portal is run by the HIRA, which collects data for administrative work. In detail, this study took data from the HIRA on clinics that operated and closed for 2 years from 2020 to 2021. Then, the participants for this study were selected from operating and closed clinics based on region at a ratio of 2:1. The detailed method will be explained in the Study Design section. The second data set was from an open public data set on residential and household information [32] run by the Ministry of Interior and Safety. This study collected data such as area code, local population, population growth, and number of households from the Ministry of Interior and Safety. These data were analyzed by linking them with the region code to which the clinics belong. The last one was from information on subway stations provided by the National Geographic Information Institute [33]. Using the clinic's location information, we identified whether there was a subway station within a 1 km radius or not. National Geographic Information Institute provided this information.

Study Design

This study adopted a cross-sectional study design for over a 2-year observation period. The unit of analysis was individual MCs and DCs. For this study, the closure status of MCs and DCs during the period 2020-2021 was identified. By observing 2 time points on December 31, this study identified whether the clinics closed or not and coded them as a group of clinics closed (1) or not (0). Using the closed group, we selected the participants for this study from closed and operating medical facilities with a ratio of 1:2 in MCs and DCs, respectively, because the proportion of closed facilities was too low, constituting less than 5% of the total study population. Training and testing participants were processed with a 4:1 ratio and 5-fold cross-validations for 4 machine learning algorithms: LR, SVM, random forest (RF), and Extreme Gradient Boost (XGB) methods.

The amount of data was sufficient, so the model learning the characteristics of each variable would be suitable. Therefore, we used stratified cross-validation to maintain the same ratio of running and closed clinics for each fold to prevent bias instead of using simple K-fold cross-validation. This study selected the LR and SVM methods because the former is the most classical model among the various machine learning models, and the latter is considered the fundamental model of deep learning and performs well in dealing with complicated models. This study also selected the RF and XGB methods because they are most frequently used for machine learning, with moderate modeling performance. Using these methods, this study predicted the probability of closure of MCs and DCs. After developing the prediction model, we evaluated the prediction results. The results were expressed through cross-validation. Figure 1 demonstrates the data processing flow or diagram of this study.

XSL•FO
RenderX

Figure 1. Overall diagram for data processing steps: a cross-sectional study design over 2 years (2020-2021). AUC: area under the curve; CV: cross-validation; LR: logistic regression; RF: random forest; ROC: receiver operating characteristic; SVM: support vector machine; XGB: Extreme Gradient Boost.

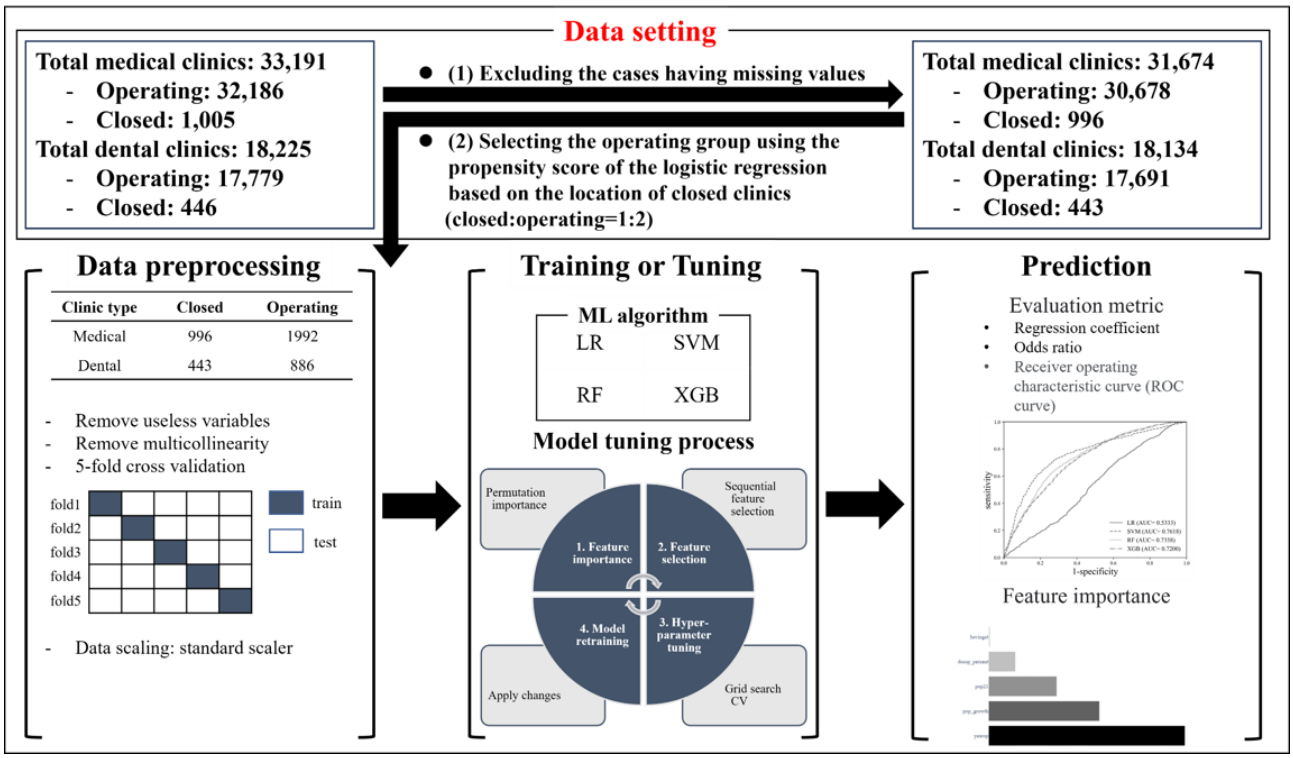

Outcome and Covariate Measures

This study predicted the closure or nonclosure of MCs and DCs, the major outcome variable. The other input variables were chosen based on previous studies but within the boundary of data availability: years of operation, number of patients, annual variation in the number of patients, having beds or not (for clinics), having computed tomography devices or not (for clinics), having magnetic resonance imaging devices or not (for clinics), number of physicians (for clinics) or dentists (for DCs), percentage of medical or dental specialists, percentage of nurses or hygienists, the local population of the year 2021, population growth rate (%), number of clinics or DCs within the local area, and having a subway station or not within 1 km.

The input variables were selected based on several previous studies. These studies used general statistical methods such as LR rather than machine learning. A study found that the closure of MCs was associated with years of operation, for-profit status, the proportion of specialists among all physicians, the number of clinics within the local area, the closure of DCs related to the proportion of specialized dentists among all dentists, the proportion of nurse and dental hygienists among all nursing staff, and several DCs within the local area [11]. Several studies on the closure of hospitals and business firms presented that closure was associated with factors such as facility age, financial status, affiliation status, and market factors [22,24,26,27]. This study used all available variables as input variables.

The annual number of patients was observed in 2020 for operating MCs and DCs. For closed MCs and DCs, the number of patients a year before closure was used. For the annual variation in the number of patients, the SD of the number of patients for 5 years was calculated. If clinics had less than 5 years of operation, we calculated the SD of the number of patients for the years of operation. The percentage of medical or dental specialists was calculated by (specialized physicians for clinics (dentists for DCs)/sum of physicians of clinics (dentists for DCs) × 100), and the percentage of nurses for MCs was calculated by (number of nurses/(number of nurses + nurse aids)× 100). The percentage of nurses for DCs was calculated by ((number of nurses + number of dental hygiene)/(number of nurses + number of dental hygiene + nurse aids)) × 100. Local area means an administrative district in Korea.

Statistical Analysis

Various machine learning techniques are used to predict patients' clinical behavior and decision-making [34,35]. This study used similar methods to those in previous studies. Before the main analysis, we took the following data preprocessing steps. First, by using the SAS program, we excluded all missing data. Second, as mentioned above, this study selected study participants from closed and running medical facilities with a ratio of 1:2 because the proportion of event groups, meaning closed facilities, was too low, being less than 5%. This process was conducted in MCs and DCs, respectively. This process was conducted with the SAS program, and this study used the propensity matching score. Third, for the cases of numeric valued data, this study rescaled them to have a normal distribution with a mean of 0 (SD 0) and a variance of 1, which is called the standard scaling process.

This study used MedCalc (version 14.8.1; MedCalc Software Ltd) to reanalyze the predicted results regarding the machine learning module. When we calculated the receiver operating characteristic curve comparison, Stata (version 11; StataCorp

XSL•FO
RenderX

LLC) was used. Regarding the machine learning algorithm, this study used the LR, SVM, and RF, which are provided by the library of scikit-learn (version 0.23.2; Python Software Foundation) and xgboost (version 1.4.0; Python Software Foundation). Regarding feature selection, this study used an algorithm that shows the permutation importance, which the library of Eli5 provides for computing feature importance. This study used 23 and 19 variables to predict the closure of MCs and DCs, respectively. Key variables were extracted using the permutation importance from those variables. The sequential feature selection technique was also used. Finally, this study selected 5 and 6 variables of MCs and DCs, respectively, for model learning.

Regarding the validation methods of the models, this study used sensitivity and specificity indices following the definition of a previous study [36]. Regarding the prediction capacity of the models, this study used area under the curve (AUC), which was calculated based on sensitivity and specificity. According to the *Standards for Reporting Diagnostic Accuracy Studies 2015* [37], it is suggested that specificity, sensitivity, and receiver operating characteristic curves be used for diagnostic evaluation. Thus, this study chose them and presented its study outputs. This study used Python (version 3.7.9; Python Software Foundation) to solve the main analysis.

Regarding AI methods and paper writing, this study discloses that no generative AI was used at any stage of writing this paper. Finally, the computer operating and hardware system were Intel Core i9-10900 CPU @ 2.80 GHz and Windows 10 OS.

Ethical Considerations

This study was approved by the institutional review board (IRB) of the HIRA on February 10, 2022 in Korea (2022-010-001). The IRB letter says that it is exempted from proceeding with the formal IRB review due to the use of the secondary health insurance administrative data. The letter also says that it has nothing to do with obtaining informed consent from research participants because the study participants are individual clinics.

Results

General Characteristics of the Study Participants

Table 1 presents the basic characteristics of clinics and DCs regarding closure status. Closed MCs are more likely to have longer operating years, fewer patients, and less variation in the number of annual patients compared to those operating. Closed DCs are less likely to be specialized and have fewer specialized nurses compared to the others.

XSL•FO
RenderX

Table 1. General characteristics of the study participants (medical and dental clinics) by clinic closure status.

Variables	Medical clinics				Dental clinics			
	Closure	Nonclosure	All	*P* value	Closure	Nonclosure	All	*P* value
Study participants, n	996	1992	2988	__[a]	443	886	1329	—
Operation (years), mean (SD)	15.6 (11.7)	13.4 (8.4)	14.1 (9.7)	*.001* [b]	15.7 (11.8)	14.6 (9.2)	14.9 (10.2)	*.08*
Patients, mean (SD)	12,028.6 (13,157.4)	14,465.3 (11,557.9)	13,653 (12,166.8)	*.001*	1178.3 (1164.7)	1518.9 (1162.1)	1405.4 (1173.6)	*.001*
Annual variation of the number of patients, mean (SD)	3104.1 (3715.7)	2665.6 (3670.8)	2957.9 (3706)	*.002*	260.9 (396.2)	228.6 (311.5)	239.4 (342.3)	*.10*
Having beds, n (%)				*.50*				—
Yes	158.364 (15.9)	334.656 (16.8)	493.02 (16.5)		—	—	—	—
No	837.636 (84.1)	1657.344 (83.2)	2494.98 (83.5)		—	—	—	—
Having computed tomography, n (%)				*.35*				—
Yes	12.948 (1.3)	35.856 (1.8)	47.808 (1.6)		—	—	—	—
No	983.052 (98.7)	1956.144 (98.2)	2940.192 (98.4)		—	—	—	—
Having magnetic resonance imaging, n (%)				*.25*				—
Yes	3.984 (0.4)	15.936 (0.8)	17.928 (0.6)		—	—	—	—
No	992.016 (99.6)	1976.064 (99.2)	2970.072 (99.4)		—	—	—	—
Doctors or dentists, mean (SD)	1.3 (1.2)	1.4 (1.1)	1.4 (1.2)	*.43*	1.4 (0.8)	1.5 (0.9)	1.5 (0.9)	*.09*
Medical or dental specialists, mean% (SD)	87.9 (31)	88.7 (29.8)	88.4 (30.2)	*.48*	10.5 (26.6)	14 (28.2)	12.8 (27.7)	*.03*
Nurse or hygienist specialized, mean% (SD)	9.6 (24)	10 (23.3)	9.9 (23.5)	*.65*	54.9 (40.8)	60.1 (37.6)	58.4 (38.8)	*.02*
Local population, mean (SD)	353,320 (179,791.1)	353,312.3 (179,759.9)	353,314.9 (179,740.2)	*.99*	360,496.1 (172,732.1)	360,496.1 (172,634.5)	360,496.1 (172,602)	*.99*
Population growth rate, mean (SD)	–0.112 (11)	–0.120 (11)	–0.117 (11)	*.98*	–1.299 (10.3)	–1.299 (10.3)	–1.299 (10.3)	*.99*
Clinics or dental clinics within a local area, mean (SD)	376.4 (471.2)	377.1 (471.3)	376.9 (471.2)	*.96*	193.2 (148.7)	191.7 (145.6)	192.2 (146.6)	*.85*
Subway station within 1 km, n (%)				*.58*				*.15*
Yes	445.212 (44.7)	868.512 (43.6)	1314.72 (44)		237.005 (53.5)	436.798 (49.3)	673.803 (50.7)	
No	550.788 (55.3)	1123.488 (56.4)	1673.28 (56)		205.995 (46.5)	449.202 (50.7)	655.197 (49.3)	

[a]Not applicable.

[b]Italicized values indicate the *P* value of *t* test for numerical variables, or the *P* value of chi-square test for categorical variables.

Model Performance Using Cross-Validation

Figure 2 presents the model performance results of 4 models in MCs and DCs. In MCs, model performance was highest in SVM for the AUC index, followed by RF, XGB, and LR. In DCs, model performance was highest in XGB, followed by RF, LR, and SVM.

XSL•FO
RenderX

Figure 2. Predicted model performance for medical and dental clinics. The ROC represents the receiver operating characteristic plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) for different classification thresholds. AUC: area under the curve; LR: logistic regression; RF: random forest; ROC: receiver operating characteristic; SVM: support vector machine; XGB: Extreme Gradient Boost.

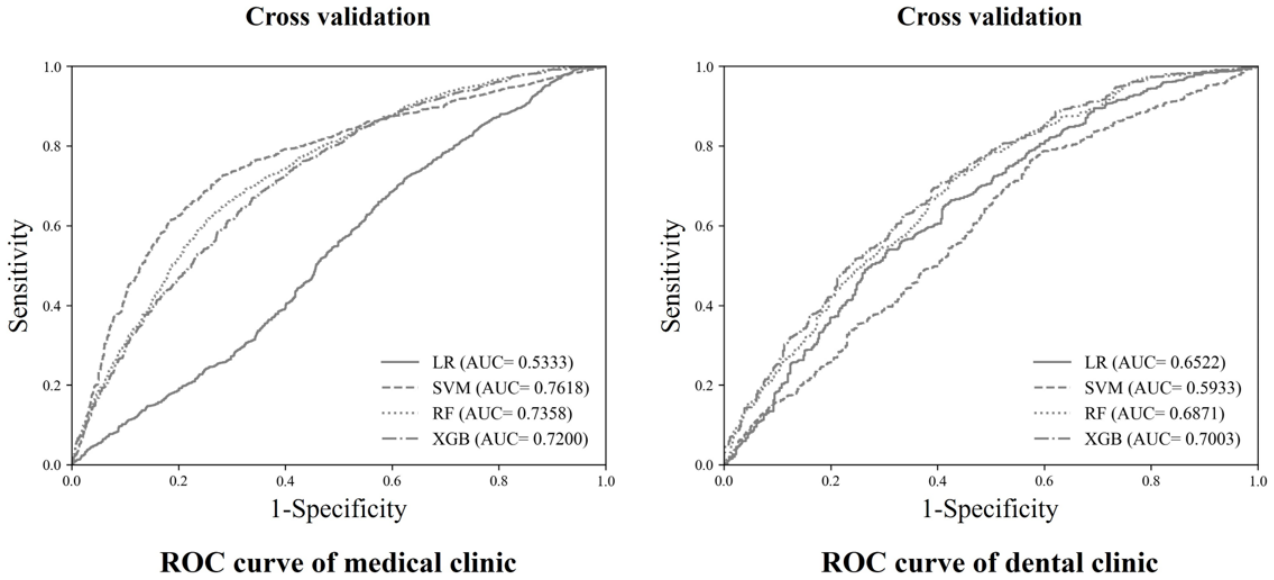

Model Cross-Validation

Table 2 presents the performance indicator for each model. In the cross-validation, the AUC of MCs was highest in SVM (0.762), followed by RF (0.736). The AUC of DCs was highest in XGB (0.700), followed by RF (0.687).

Table 2. Performance indicator for each model (cross-validation) by medical and dental clinics.

Clinic types and models	Sensitivity, AUC[a] (95% CI)	Specificity, AUC (95% CI)	AUC (95% CI)	*P* value (AUC)
Medical clinics				
LR[b]	0.996 (0.992-0.998)	0.033 (0.023-0.046)	0.533 (0.515-0.551)	.004
SVM[c]	0.896 (0.881-0.909)	0.326 (0.297-0.356)	0.762 (0.746-0.777)	<.001
RF[d]	0.939 (0.928-0.949)	0.278 (0.250-0.307)	0.736 (0.720-0.752)	<.001
XGB[e]	0.823 (0.805-0.839)	0.480 (0.448-0.511)	0.720 (0.704-0.736)	<.001
Dental clinics				
LR	0.990 (0.981-0.995)	0.034 (0.019-0.055)	0.652 (0.626-0.678)	<.001
SVM	0.998 (0.992-0.999)	0.005 (0.001-0.016)	0.593 (0.566-0.620)	<.001
RF	0.916 (0.896-0.934)	0.278 (0.236-0.322)	0.687 (0.661-0.712)	<.001
XGB	0.903 (0.882-0.922)	0.323 (0.279-0.369)	0.700 (0.675-0.725)	<.001

[a]AUC: area under the curve.
[b]LR: logistic regression.
[c]SVM: support vector machine.
[d]RF: random forest.
[e]XGB: Extreme Gradient Boost.

Major Factors Affecting Closure by Clinic Types

Years of operation, population growth, number of residents within the local area, and percentage of medical specialists are crucial factors for closing MCs. In contrast, number of patients, annual variation in the number of patients, years of operation, percentage of dental specialists, and number of residents within the local area are considered essential factors affecting DC closures.

Table 3 presents the crucial factors measured by permutation importance across 4 models. The year of operation for MCs had the highest score (0.194), followed by population growth (0.110). For DCs, the number of patients had the highest score (0.096), followed by annual variation in the number of patients (0.060).

XSL•FO
RenderX

Table 3. Selecting important factors based on the mean score of the permutation feature importance by clinic types (medical and dental clinics) and prediction models.

Clinic type and factors	LR[a]	SVM[b]	RF[c]	XGB[d]	Mean score (SD)
Medical clinics					
Having CT[e]	0.001	–0.004	0.005	0.003	0.001 (0.001)
Medical specialty (%)	0.004	0.046	0.034	0.021	0.026 (0.002)
Local population	–0.009	0.150	0.073	0.055	0.067 (0.004)
Population growth	–0.004	0.224	0.116	0.103	0.110 (0.004)
Operation (years)	0.043	0.284	0.244	0.207	0.194 (0.01)
Dental clinics					
Dentists (n)	–0.001	0	0.002	0	0 (0.009)
Local population	0	0.003	0.004	0.004	0.003 (0.004)
Dental specialty (%)	0.018	0.006	0.008	0.005	0.009 (0.004)
Operation (years)	–0.003	0.018	0.059	0.065	0.035 (0.015)
Annual variation in the number of patients (n)	0.056	0.014	0.078	0.094	0.060 (0.012)
Patients (n)	0.135	0.033	0.111	0.106	0.096 (0.01)

[a]LR: logistic regression.
[b]SVM: support vector machine.
[c]RF: random forest.
[d]XGB: Extreme Gradient Boost.
[e]CT: computed tomography.

Discussion

Principal Results

This study found that it is feasible to predict the closure of MCs and DCs using machine learning methods with a fair level of the validation score measured by the AUC. Interestingly, MCs and DCs have different best-fitting models. The best-fitting MC and DC closure models were in SVM and XGB, respectively. Finally, except for years of operation, different important features affected the closure of MCs and DCs. Years of operation were an important feature affecting the closure of both MCs and DCs. External environmental factors, such as population growth and local residential population, largely influenced the closure of clinics. However, the closure of DCs was influenced by the internal characteristics of the facility, such as the number of patients and annual variation in the number of patients.

Comparison With Prior Work

This study indicates the feasibility of predicting clinic closure with machine learning methods. This result aligns with previous studies in which many AI methods well predicted patients' clinical behaviors [34,35] and partially predicted hospital closure and general firm bankruptcy [12-14,24]. Other than that, we could not find any similar previous studies targeting MCs and DCs; thus, it was impossible to compare the results of this study with others. However, our study results suggest that the same machine learning technologies could be applied to predicting the closure of medical facilities, such as many long-term care facilities, pharmacies, and clinics with oriental medicine.

Regarding the model performance to predict closure, SVM and XGB were the best-fitting models to predict the closure of MCs and DCs, respectively. A very interesting finding is that each study group has its best-fitting model. While a study on bankruptcy prediction presented that XGB is the best performer, the best-fit model for predicting the annual number of inpatients and outpatients was RF and linear regressor, respectively [25,38]. These results suggest that future research should refer to the optimal models of previous studies and include such models as alternatives in their research methods.

For cross-validation using the AUC score, MCs had the highest score in SVM (0.762) followed by RF (0.736) and DCs had the highest score in XGB (0.700) followed by RF (0.687). These study results suggest 2 things. The first is about the models' cross-validation per se. Generally, an AUC score above 0.70 would be considered "fair" for validating the model [39]. Several previous studies dealing with various health care participants have a broad range of AUC and modeling accuracy ranging from 0.69 [24] to 0.85 AUC [40]. Compared to these previous validation scores, this study has a fair cross-validation score, and thus, there is ample potential to develop further good models based on this study. Second, this study suggests that the model performance of MCs and DCs needs to improve further. As evidenced by the results of this study, the overall cross-validation score of each model in MCs is slightly higher than that of DCs. Further search and analysis are necessary to investigate why these are different. This study suggests that methodological approaches that involve selecting different input variables by each model with their own are necessary.

XSL•FO
RenderX

This study found that years of operation were an important factor affecting the closure of MCs and DCs. Clinics performing medical practices over a longer period in 1 place might consider moving to other places, which is why factors such as years of operation are significant factors affecting clinic closure. Our study results showed that external environmental factors, such as population growth and local residential population, are essential to consider in the closure of clinics, and the internal characteristics of DCs, such as the number of patients and annual variation in the number of patients, critically affect the closure of DCs. The results of this study partially align with those of a previous study in which factors such as medical and nurse specialties (both in MCs and DCs), year of operation (in both), and market competition are closely associated with the closure of medical facilities [11]. Nevertheless, there are some differences in that while the previous study was based only on LR investigating factors associated with the closure of MCs, this study involved predicting and identifying critical factors affecting the closure of clinics with various AI methods.

These research results tell people who want to open a new MC or DC what factors are related to business closure. This study argues that these factors must be considered in countries where many MCs and DCs are operating. This study used several machine learning methods, but they needed more accuracy. This study presents the challenge of increasing the accuracy of the methodology, thus providing fundamental reference values for these analysis methods.

Regarding the results of this study concerning hospital closure, there are distinctive differences between hospitals and clinics. While factors such as affiliation status, financial distress, size of the hospital, locality, low-profit margin, and market competition were critical factors affecting hospital closure [22,26,27,29,30], features such as years of operation, local population, and number of patients were significant factors affecting MC and DC closure [11]. However, direct comparisons between hospitals and clinics are still unreasonable because the participants of this study are different.

This study is valuable in several respects. First, this study is the first to predict the closure of MCs and DCs. There have been several studies on the closure of hospitals but not clinics. Second, studies on clinic closure are very important because clinics play a dominant role in providing health care to the community residents. Identifying closure factors is crucial to prevent the breakdown of the health care delivery systems of local communities. By investigating their survival through this kind of study, we can obtain useful information to increase the sustainability of their business survival. Third, this study had highly valid data sets comprising national data. A study using a large cumulative data set would be invaluable because the validity of study results increases as the sample size of a study increases.

Limitations

This study has several limitations. First, although this study was based on a 2-year national data set, few clinics have closed. This study conducted predictions using 2-year aggregated data by selecting operating clinics and matching them with closed clinics. This process might lead to the loss of some valuable information. However, this study had a rigorous methodology in that it increased comparability by equally weighting the risk factors to those study participants within the same locality. Second, although financial factors are critical factors affecting hospital closure [22,27], this study did not include that variable. However, to minimize its impact, this study included the number of patients as a potential proxy variable of financial factors. Third, this study design is of closed clinics, and running clinics were selected at a ratio of 1:2 based on the locality of study participants using the propensity matching score of LR, which might result in an under-sampling issue and, thus, would have the risk of overfitting. We suggest that future research could consider a hypothesis-free approach, such as repeated undersampling, to overcome this study's limitations. Fourth, the data collection period is within the COVID-19 pandemic. The COVID-19 pandemic is very special, in which the closure of small clinics might be common. Small health care institutions are more greatly affected by external forces [7]. The data from the pandemic period would differ from those we see today. Thus, the interpretation of study results needs consideration of this background. However, although data were collected during this unique period, this study argues that the impact of COVID-19 would equally affect those clinics running and closed. This may minimize its negative impact on study data. Finally, the results of this study may be limited to Korea and countries with health care delivery systems similar to those in Korea. Suppose the operation or closure of MCs is dependent on governmental plans. In that case, the results of this study may not apply to those countries because this study is based on free business medical markets.

Conclusions

This study predicted the closure status of MCs and DCs and found that composing models that predict the closure status of clinics is feasible. Among the 4 models, SVM was the best-fit model for MCs and XGB for DCs. Years of operation were an important factor that critically affected the closure of MCs and DCs. Additionally, environmental factors critically affected the closure of MCs; in contrast, DC's internal features, such as the number of patients, were critical factors affecting the closure of DCs. These discoveries increase the range of our knowledge and understanding that machine learning techniques could be useful in predicting the closure of small medical institutions and applicable to long-term care hospitals, nursing homes, and pharmacies. This study also indirectly suggests that governments could conserve health care resources and prevent wastage if they develop good prediction models and provide information to medical or potential medical providers. We hope this study will inspire similar studies with further dynamic methods, leading to knowledge expansion and providing various insights to policy makers and researchers in related fields.

XSL•FO
RenderX

Acknowledgments

We would like to thank the research staff at the Health Insurance Review and Assessment Service (HIRA) Research Institute for providing us with valuable data and several important comments on this paper. We also thank Miss Seon-Min Lee, Registered Nurse, at the Medical Devices R&D Center, Gachon University Gil Medical Center, in Korea, for providing various technical support on AI methods. This work was supported by the Starting Growth Technological R&D Program (RS-2023-00284655) funded by the Ministry of SMEs and Startups (Korea). This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program (K_G012001187805, Development of Diagnostic Medical Devices with Artificial intelligence Based Image Analysis Technology) funded By the Ministry of Trade Industry & Energy (MOTIE, Korea).

Data Availability

The data sets used and analyzed during this study are available from the corresponding author upon reasonable request. In addition, the data sets on clinic closures may be obtained from the HIRA by reasonable request.

Authors' Contributions

Study design, institutional review board approval, funding acquisition, and data collection were carried out by YTP and KGK. Statistical analysis, interpretation of data, and writing of the original paper were conducted by YTP, KGK, and DK. The revision of this paper was performed by YTP, KGK, and JSJ. DK and JSJ significantly contributed to the improvement of the manuscript. All authors read and approved the final paper.

Conflicts of Interest

None declared.

References

1. Fast facts on U.S. hospitals, 2024. American Hospital Association. 2024. URL: https://www.aha.org/infographics/2024-01-18-fast-facts-us-hospitals-infographics [accessed 2023-08-02]
2. Statista. URL: https://www.statista.com/statistics/949580/hospitals-in-united-kingdom/ [accessed 2023-01-14]
3. Vaughan L. Why bigger isn't always better: caring for patients in smaller, rural and remote hospitals. Future Healthc J 2020;7(1):4-5 [FREE Full text] [doi: 10.7861/fhj.ed-7-1-2] [Medline: 32104756]
4. KOSIS. URL: https://tinyurl.com/428hedk9
5. Rural hospital closures have spillover effects for other nearby facilities, researchers warn. Fierce Healthcare. 2022. URL: https://www.fiercehealthcare.com/providers/rural-hospital-closures-have-spillover-effects-other-nearby-facilities-researchers-warn [accessed 2022-01-15]
6. Saghafian S, Song LD, Raja AS. Towards a more efficient healthcare system: opportunities and challenges caused by hospital closures amid the COVID-19 pandemic. Health Care Manag Sci 2022;25(2):187-190 [FREE Full text] [doi: 10.1007/s10729-022-09591-7] [Medline: 35292872]
7. Park Y, Lane C, Lee H, Lee J. Was size of healthcare institution a factor affecting changes in healthcare utilisation during the COVID-19 pandemic in Korea? A retrospective study design analysing national healthcare big data. BMJ Open 2022;12(12):e064537 [FREE Full text] [doi: 10.1136/bmjopen-2022-064537] [Medline: 36600350]
8. Kirn J. 35% of rural hospitals in Missouri at risk of closing, nonprofit says. St. Louis Business Journal. 2023. URL: https://www.bizjournals.com/stlouis/news/2023/01/04/rural-hospitals-at-risk-of-closing-nonprofit-says.html(accessed [accessed 2023-01-01]
9. Doug Richards. Atlanta hospital closing follows national trend. Alive. 2022. URL: https://www.11alive.com/article/news/health/atlanta-medical-center-hospital-closure/85-431c5311-0845-49e2-9394-e71eb8961568 [accessed 2024-08-21]
10. Pieper O. Germany's small hospitals are disappearing. Health Germany. 2022. URL: https://www.dw.com/en/germanys-small-hospitals-are-disappearing/a-64128586 [accessed 2023-01-18]
11. Park YT. Factors associated with the closure of small hospitals, clinics, and dental clinics. HIRA Res 2022;2(1):120-130. [doi: 10.52937/hira.22.2.1.120]
12. Narvekar A, Guha D. Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession[J]. Data Sci Finance Econ 2021;1(2):180-195. [doi: 10.3934/dsfe.2021010]
13. Wang H, Liu X. Undersampling bankruptcy prediction: Taiwan bankruptcy data. PLoS One 2021;16(7):e0254030 [FREE Full text] [doi: 10.1371/journal.pone.0254030] [Medline: 34197533]
14. Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction, expert systems with applications. Expert Syst Appl 2017;83:405-417 [FREE Full text] [doi: 10.1016/j.eswa.2017.04.006]
15. Xu X, Fairley CK, Chow EPF, Lee D, Aung ET, Zhang L, et al. Using machine learning approaches to predict timely clinic attendance and the uptake of HIV/STI testing post clinic reminder messages. Sci Rep 2022;12(1):8757 [FREE Full text] [doi: 10.1038/s41598-022-12033-7] [Medline: 35610227]

XSL•FO
RenderX

16. Liu D, Shin WY, Sprecher E, Conroy K, Santiago O, Wachtel G, et al. Machine learning approaches to predicting no-shows in pediatric medical appointment. NPJ Digit Med 2022;5(1):50 [FREE Full text] [doi: 10.1038/s41746-022-00594-w] [Medline: 35444260]
17. Feretzakis G, Karlis G, Loupelis E, Kalles D, Chatzikyriakou R, Trakas N, et al. Using machine learning techniques to predict hospital admission at the emergency department. J Crit Care Med (Targu Mures) 2022;8(2):107-116 [FREE Full text] [doi: 10.2478/jccm-2022-0003] [Medline: 35950158]
18. Saadatmand S, Salimifard K, Mohammadi R, Kuiper A, Marzban M, Farhadi A. Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients. Ann Oper Res 2023;328:1043-1071 [FREE Full text] [doi: 10.1007/s10479-022-04984-x] [Medline: 36196268]
19. Gao R, Cheng WX, Suganthan PN, Yuen KF. Inpatient discharges forecasting for Singapore hospitals by machine learning. IEEE J Biomed Health Inform 2022;26(10):4966-4975. [doi: 10.1109/JBHI.2022.3172956] [Medline: 35522640]
20. Fang C, Pan Y, Zhao L, Niu Z, Guo Q, Zhao B. A machine learning-based approach to predict prognosis and length of hospital stay in adults and children with traumatic brain injury: retrospective cohort study. J Med Internet Res 2022;24(12):e41819 [FREE Full text] [doi: 10.2196/41819] [Medline: 36485032]
21. Yun J, Cho B, Lim S. Risk factors of primary care clinic closure in Korea using nationwide data. Eur J Public Health 2018;28(suppl_4):cky214.267 [FREE Full text] [doi: 10.1093/eurpub/cky214.267]
22. Holmes GM, Kaufman BG, Pink GH. Predicting financial distress and closure in rural hospitals. J Rural Health 2017;33(3):239-249. [doi: 10.1111/jrh.12187] [Medline: 27500663]
23. Kaufman BG, Thomas SR, Randolph RK, Perry JR, Thompson KW, Holmes GM, et al. The rising rate of rural hospital closures. J Rural Health 2016;32(1):35-43. [doi: 10.1111/jrh.12128] [Medline: 26171848]
24. Beauvais B, Ramamonjiarivelo Z, Betancourt J, Cruz J, Fulton L. The predictive factors of hospital bankruptcy—an exploratory study. Healthcare (Basel) 2023;11(2):165 [FREE Full text] [doi: 10.3390/healthcare11020165] [Medline: 36673533]
25. Park Y, Lee S, Lee Y, Kim K. Performance evaluation of artificial intelligence methods predicting annual number of patients in hospitals. HIRA Res 2024;4(1):73-86 [FREE Full text] [doi: 10.52937/hira.24.4.1.e4]
26. Jiang HJ, Fingar KR, Liang L, Henke RM. Risk of closure among independent and multihospital-affiliated rural hospitals. JAMA Health Forum 2022;3(7):e221835 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.1835] [Medline: 35977220]
27. Enumah SJ, Chang DC. Predictors of financial distress among private U.S. hospitals. J Surg Res 2021;267:251-259 [FREE Full text] [doi: 10.1016/j.jss.2021.05.025] [Medline: 34161840]
28. Donnelly-DeRoven C. NC study finds three financial indicators that could predict the closure of a rural hospital. 2022. URL: https://www.northcarolinahealthnews.org/2022/01/24/nc-study-finds-three-financial-indicators-that-could-predict-the-closure-of-a-rural-hospital/ [accessed 2023-02-01]
29. Vaughan L, Edwards N. The problems of smaller, rural and remote hospitals: separating facts from fiction. Future Healthcare J 2020;7(1):38-45 [FREE Full text] [doi: 10.7861/fhj.2019-0066] [Medline: 32104764]
30. Hsia RY, Kellermann AL, Shen Y. Factors associated with closures of emergency departments in the United States. JAMA 2011;305(19):1978-1985 [FREE Full text] [doi: 10.1001/jama.2011.620] [Medline: 21586713]
31. Unified Portal Reporting Healthcare Resource. URL: https://www.hurb.or.kr/hira_sg/index.jsp?sso=ok [accessed 2024-08-02]
32. Resident registration population and household status by administrative district. Resident registration population and household status by administrative district.: Ministry of the Interior and Safety(South Korea) URL: https://jumin.mois.go.kr/index.jsp [accessed 2024-08-02]
33. National Geographic Information Institute. URL: https://www.ngii.go.kr/eng/main.do [accessed 2024-08-02]
34. Kim YJ, Ganbold B, Kim KG. Web-based spine segmentation using deep learning in computed tomography images. Healthcare Inf Res 2020;26(1):61-67 [FREE Full text] [doi: 10.4258/hir.2020.26.1.61] [Medline: 32082701]
35. Abhari S, Niakan Kalhori SR, Ebrahimi M, Hasannejadasl H, Garavand A. Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. Healthcare Inf Res 2019;25(4):248-261 [FREE Full text] [doi: 10.4258/hir.2019.25.4.248] [Medline: 31777668]
36. van Stralen KJ, Stel VS, Reitsma JB, Dekker FW, Zoccali C, Jager KJ. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. Kidney Int 2009;75(12):1257-1263 [FREE Full text] [doi: 10.1038/ki.2009.92] [Medline: 19340091]
37. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. URL: https://www.equator-network.org/reporting-guidelines/stard [accessed 2024-05-28]
38. Elhorst P, Abreu M, Amaral P, Bhattacharjee A, Bond-Smith S, Chasco C, et al. Raising the bar (20). Spat Econ Anal 2022;17(2):151-155. [doi: 10.1080/17421772.2022.2053402]
39. Muller MP, Tomlinson G, Marrie TJ, Tang P, McGeer A, Low DE, et al. Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia? Clin Infect Dis 2005;40(8):1079-1086 [FREE Full text] [doi: 10.1086/428577] [Medline: 15791504]
40. Shetty S, Musa M, Brédart X. Bankruptcy prediction using machine learning techniques. J Risk Financ Manage 2022;15(1):35. [doi: 10.3390/jrfm15010035]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DC: dental clinic
HIRA: Health Insurance Review and Assessment Service
IRB: institutional review board
LR: logistic regression
MC: medical clinic
RF: random forest
SVM: support vector machine
XGB: Extreme Gradient Boost

Edited by A Mavragani; submitted 17.02.23; peer-reviewed by W Yu Jen , U Kanike, HY Wang; comments to author 03.05.24; revised version received 12.06.24; accepted 01.07.24; published 30.08.24.

Please cite as:
Park YT, Kim D, Jeon JS, Kim KG
Predictors of Medical and Dental Clinic Closure by Machine Learning Methods: Cross-Sectional Study Using Empirical Data
J Med Internet Res 2024;26:e46608
URL: https://www.jmir.org/2024/1/e46608
doi:10.2196/46608
PMID:

©Young-Taek Park, Donghan Kim, Ji Soo Jeon, Kwang Gi Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancing Patient Selection in Sepsis Clinical Trials Design Through an AI Enrichment Strategy: Algorithm Development and Validation

Meicheng Yang[1*], PhD; Jinqiang Zhuang[2,3*], MD; Wenhan Hu[4], PhD; Jianqing Li[1,5], PhD; Yu Wang[3], MD; Zhongheng Zhang[6], MD; Chengyu Liu[1], PhD; Hui Chen[4], MD

[1]State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing, China

[2]Emergency Intensive Care Unit (EICU), The Affiliated Hospital of Yangzhou University, Yangzhou University, Yangzhou, China

[3]Key Laboratory of Big Data Analysis and Knowledge Services of Yangzhou City, Yangzhou University, Yangzhou, China

[4]Jiangsu Provincial Key Laboratory of Critical Care Medicine, Department of Critical Care Medicine, Zhongda Hospital, School of Medicine, Southeast University, Nanjing, China

[5]School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

[6]Department of Emergency Medicine, Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Hui Chen, MD
Jiangsu Provincial Key Laboratory of Critical Care Medicine, Department of Critical Care Medicine, Zhongda Hospital
School of Medicine
Southeast University
No 87, Dingjiaqiao Road, Gulou District
Nanjing, 210009
China
Phone: 86 15905162429
Email: 15905162429@163.com

Abstract

Background: Sepsis is a heterogeneous syndrome, and enrollment of more homogeneous patients is essential to improve the efficiency of clinical trials. Artificial intelligence (AI) has facilitated the identification of homogeneous subgroups, but how to estimate the uncertainty of the model outputs when applying AI to clinical decision-making remains unknown.

Objective: We aimed to design an AI-based model for purposeful patient enrollment, ensuring that a patient with sepsis recruited into a trial would still be persistently ill by the time the proposed therapy could impact patient outcome. We also expected that the model could provide interpretable factors and estimate the uncertainty of the model outputs at a customized confidence level.

Methods: In this retrospective study, 9135 patients with sepsis requiring vasopressor treatment within 24 hours after sepsis onset were enrolled from Beth Israel Deaconess Medical Center. This cohort was used for model development, and 10-fold cross-validation with 50 repeats was used for internal validation. In total, 3743 patients with sepsis from the eICU Collaborative Research Database were used as the external validation cohort. All included patients with sepsis were stratified based on disease progression trajectories: rapid death, recovery, and persistent ill. A total of 148 variables were selected for predicting the 3 trajectories. Four machine learning algorithms with 3 different setups were used. We estimated the uncertainty of the model outputs using conformal prediction (CP). The Shapley Additive Explanations method was used to explain the model.

Results: The multiclass gradient boosting machine was identified as the best-performing model with good discrimination and calibration performance in both validation cohorts. The mean area under the receiver operating characteristic curve with SD was 0.906 (0.018) for rapid death, 0.843 (0.008) for recovery, and 0.807 (0.010) for persistent ill in the internal validation cohort. In the external validation cohort, the mean area under the receiver operating characteristic curve (SD) was 0.878 (0.003) for rapid death, 0.764 (0.008) for recovery, and 0.696 (0.007) for persistent ill. The maximum norepinephrine equivalence, total urine output, Acute Physiology Score III, mean systolic blood pressure, and the coefficient of variation of oxygen saturation contributed the most. Compared to the model without CP, using the model with CP at a mixed confidence approach reduced overall prediction

XSL•FO
RenderX

errors by 27.6% (n=62) and 30.7% (n=412) in the internal and external validation cohorts, respectively, as well as enabled the identification of more potentially persistent ill patients.

Conclusions: The implementation of our model has the potential to reduce heterogeneity and enroll more homogeneous patients in sepsis clinical trials. The use of CP for estimating the uncertainty of the model outputs allows for a more comprehensive understanding of the model's reliability and assists in making informed decisions based on the predicted outcomes.

(J Med Internet Res 2024;26:e54621) doi:10.2196/54621

KEYWORDS

sepsis; enrichment strategy; disease progression trajectories; artificial intelligence; predictive modeling; conformal prediction

Introduction

Sepsis, defined as a dysregulated immune response to infection that leads to acute organ dysfunction, persistently stands as one of the leading causes of mortality worldwide [1,2]. Sepsis is a heterogeneous syndrome, underscored by various infection sites, pathogens, and a vast, multidimensional array of clinical and biological features [3], concomitant with the between-patient variability in response to treatment, which might account for the absence of benefit in most randomized controlled trials (RCTs) assessing various therapies in sepsis [4]. The targeted inclusion of more homogeneous patient populations shows great promise and is an emerging trend in clinical trials [5].

Clinical trial enrichment strategies typically encompass prognostic and predictive enrichment. PaO_2/FiO_2 ratio has been used to identify patients with high-risk acute respiratory distress syndrome or those more likely to respond to prone positioning [6,7]. Combined variables can enhance the accuracy of patient stratification, thereby improving the efficacy of clinical trial enrichment [8]. Several phenotypes of sepsis have been identified through various combinations of clinical features, with both prognostic and predictive enrichment [9,10]. However, these phenotypes brought another layer of complexity due to their heterogeneity and were not originally conceptualized to improve the stratification of homogeneous patient groups in sepsis clinical trials. Patients with sepsis are often accompanied by circulation dysfunction and require the use of vasopressors. Tracking how vasopressor administration changes could help identify those likely to benefit most from treatment. Some may get worse quickly and die, others might slowly improve, but some could have a longer disease progression and might gain more from clinical drug trials. Therefore, developing an artificial intelligence (AI) model to predict which patients with sepsis will have a longer course is key to making clinical trials more effective [11].

AI has now facilitated the identification of homogeneous subgroups, allowing rapid selection of target patients for potential enrollment in further RCTs [12]. However, a significant challenge in AI modeling lies in quantifying the reliability of the model predictions for new patients, particularly when data extend beyond the original training domain. In addition, most AI models provide binary point predictions, essentially yes or no, without assessing the reliability of these predictions. However, when implementing AI models in high-risk environments, it is critical to incorporate uncertainty quantification to minimize the occurrence of unanticipated model failures such as the risk of automation bias [13].

This study aimed to develop and validate a novel data-driven AI methodology for predicting patients with sepsis who are likely to have a sufficiently long disease course, enabling the enrollment of homogeneous patient populations in clinical trials. The AI model was expected to provide interpretable factors and confidence measures to monitor predictions and identify uncertain predictions at a customized confidence level for human review, so as to enroll reliable persistent ill patients.

Methods

Study Design and Setting

This study used data from 2 distinct databases: the Medical Information Mart for Intensive Care Database-IV (MIMIC-IV) [14] and the eICU Collaborative Research Database (eICU-CRD) [15]. MIMIC-IV contained critical care data for 73,181 patients admitted to the ICUs at Beth Israel Deaconess Medical Center between 2008 and 2019. As a large multicenter intensive care unit (ICU) database, eICU-CRD collected more than 200,000 ICU admissions from 335 units in 208 hospitals across the United States from 2014 to 2015. This study followed the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist (Multimedia Appendix 1).

Ethical Considerations

The MIMIC-IV and eICU-CRD, which had been made publicly accessible, received ethical approval from the institutional review boards at Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, in alignment with the principles outlined in the Declaration of Helsinki. The approval also encompassed a waiver of the need for informed consent, since all protected health information within the databases was deidentified [14,15]. Access to these databases was provided after completion of a training program in human research ethics and signing of a data use agreement with PhysioNet, earning a certification number of 27252652.

Study Population

For inclusion in our study, patients were required to meet the sepsis diagnosis and initiate vasopressor treatment within 24 hours after identifying sepsis. In MIMIC-IV, sepsis was defined according to the Sepsis-3 criteria, known or suspected infection, and Sequential Organ Failure Assessment (SOFA) score ≥2 points [1]. The following 2 time points were specified to define

XSL•FO
RenderX

the onset time of sepsis t_{sepsis} [16]: (1) $t_{suspicion}$: clinical suspicion of infection as determined by the earlier timestamp of intravenous (IV) antibiotics administration and cultures acquired within a specific timeframe. If IV antibiotics were given first, the cultures must have been obtained within 24 hours. If cultures were obtained first, then antibiotics must have been subsequently ordered within 72 hours. (2) t_{SOFA}: the occurrence of organ failure as identified by a 2-point increase in the SOFA score within a 24-hour period. t_{sepsis} is the earlier of $t_{suspicion}$ and t_{SOFA}, as long as t_{SOFA} occurs no more than 24 hours before or 12 hours after $t_{suspicion}$; otherwise, the patient will not be diagnosed as sepsis. Specifically, if $t_{suspicion}$–24 hours≤t_{SOFA}≤$t_{suspicion}$+12 hours, then t_{sepsis}=min ($t_{suspicion}$, t_{SOFA}). In eICU-CRD, patients with sepsis were identified based on the admission diagnosis, with the sepsis identification defined as the time of ICU admission. To be as close as possible to the sepsis onset defined in eICU-CRD, only patients with sepsis onset within 24 hours of ICU admission were included in MIMIC-IV. Exclusion criteria included patients younger than 18 years, those who died within the first 24 hours of sepsis onset, or ICU stay of less than 24 hours. For patients with multiple ICU admissions, only the first admission was considered.

Trajectories of Sepsis and Candidate Predictors

We hypothesized that patients with sepsis were recruited into RCTs within 24 hours of identification. Following the enrollment, all patients with sepsis were stratified based on their disease progression trajectories: “rapid death” included patients who expired within 48 hours after enrollment, “recovery” included patients who liberated from vasopressor support within 48 hours after enrollment and maintained for at least 24 hours, and “persistent ill” included patients who necessitated ongoing vasopressor support within 48 hours after enrollment.

The selection of candidate predictors incorporated both static and time-varying variables. Time-varying variables were extracted within the 24-hour period between identification and enrollment. Any outliers were identified and subsequently excluded according to the criteria defined in Multimedia Appendix 2. For time-varying variables with multiple measurements during the 24-hour duration, we included the maximum, minimum, median, and coefficient of variation (SD/mean) values for analysis. This resulted in a total of 148 features.

Predictive Model Development and Explanation

The MIMIC-IV data set was used for model development, and 10-fold cross-validation with 50 repeats was used for internal validation (Figure 1). The eICU-CRD was used for the retrospective external validation. A total of 12 combinations of model architectures and setups were used (Figure 1). The 4 specific architectures evaluated were gradient boosting machine (GBM) [17], neural decision forest (NDF) [18], random forest (RF), and logistic regression (LR; Multimedia Appendix 3). GBM, RF, and LR were constructed with reference to the research of Schwager et al [11] for predicting acute respiratory distress syndrome trajectories. In addition, these 3 models were commonly used in critical illness prediction tasks using electronic health record (EHR) data [19-21]. However, they were all traditional machine learning models, so we developed a deep learning model for comparison, that is the NDF, which has also been shown to be effective for disease prediction using EHR data [22].

XSL•FO
RenderX

Figure 1. Workflow of the study. For the multiple predictions in evaluation, it means that the prediction was uncertain, and the artificial intelligence model could not distinguish between the possible class labels, thus awaiting the human review in order to result in a single class prediction. APS: Acute Physiology Score; eICU-CRD: eICU Collaborative Research Database; GBM: gradient boosting machine; GCS: Glasgow Coma Scale; HR: heart rate; ICU: intensive care unit; IMV: invasive mechanical ventilation; IV: intravenous; LR: logistic regression; MAP: mean arterial pressure; MIMIC-IV: Medical Information Mart for Intensive Care Database-IV; NDF: neural decision forest; NEE: norepinephrine equivalence; RF: random forest; SBP: systolic blood pressure; SHAP: Shapley Additive Explanations; UOP: urine output.

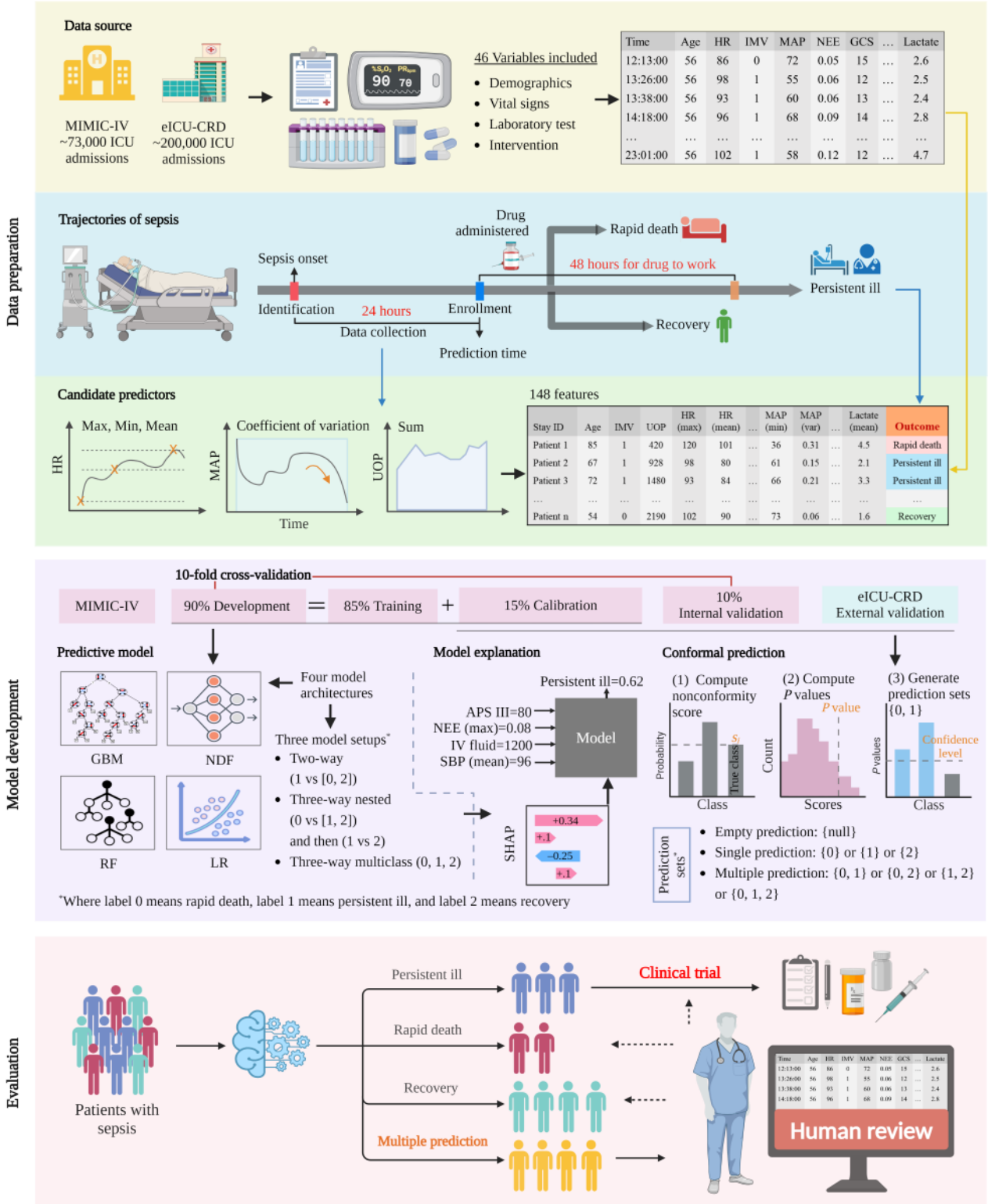

These model architectures were used in 3 different setups, namely, the 2-way (modeling persistent ill and rapid death or recovery), 3-way nested (first predicting rapid death versus persistent ill or recovery, then classifying persistent ill or recovery), and 3-way multiclass (generating probabilities for each trajectory simultaneously) configurations. For missing data, a mean imputation method was used for NDF, RF, and LR, while GBM did not require imputation.

XSL•FO
RenderX

We used Shapley Additive Explanations [23] values to explain the best model among the 12 constructed models. In addition, we examined the performance of the lightweight model using the feature subset in 3 ways (Multimedia Appendix 3): retraining the model using only the 15 most important features (measured by Shapley Additive Explanations), using the Boruta feature selection method, or using only the maximum norepinephrine equivalence (NEE), that is, NEE (max).

Conformal Prediction

To estimate the uncertainty of the model outputs, we used the conformal prediction (CP) framework built on top of the prediction algorithm (Multimedia Appendix 3). The CP is a method that can be mathematically guaranteed to make reliable predictions at a user-specified desired error rate for unknown samples that differ from the training data [13]. We specifically implemented a Mondrian CP to handle the prediction tasks with unbalanced data and worked on a class basis to ensure the desired error rate within each class. For the 3 trajectories of the patients with sepsis, the possible prediction sets are single predictions of {0}, {1}, {2}; multiple predictions of {0, 1}, {0, 2}, {1, 2}, {0, 1, 2}; and the empty prediction of {null}, where label 0 means rapid death, label 1 means persistent ill, and label 2 means recovery. We further split the development set into a training set to train the AI models and a calibration set (Figure 1) to develop the Mondrian CP and also to tune the model hyperparameters using a Bayesian optimizer [24].

Statistical Analysis

Univariate analysis was performed using logistic regression and was performed by associating each variable with three trajectories: (1) rapid death versus recovery or persistent ill, (2) persistent ill versus recovery or rapid death, and (3) recovery versus rapid death or persistent ill. Coefficients and CIs were determined on the logit scale and then transformed through exponentiation to provide estimates and 95% CIs for the odds ratios.

The discrimination performance of the model was assessed using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve. Calibration was performed using calibration curves and the Brier score. We mainly used the AUROC value to identify the model architecture and setup with the best performance for predicting persistent ill. Metrics such as true positive rate (TPR), positive predictive value (PPV), and the *F*-0.5 score (a composite of TPR and PPV that gives more weight to PPV) were also calculated. Cutoff values were chosen to maximize the F-0.5 score for the 2-way and 3-way nested setup. For the multiclass setup, the predicted outcome was obtained with the class that had the highest probability. These metrics were reported as mean (SD) for both the internal validation and external validation.

Since the conformal predictor could generate multiple or empty predictions, we could not calculate the TPR and PPV of the conformal predictor directly. To evaluate the effectiveness of the conformal predictor and assess its performance, we randomly select 1 model with the optimal architecture from the 10-fold cross-validation with 50 repeats as an example. We evaluated efficiency, which measures the proportion of all predictions that yield a single correct prediction, and the validity (error rate), representing the proportion of all predictions that did not surpass the predetermined significance level [13].

Results

Study Population

In the MIMIC-IV, 9135 patients met the inclusion and exclusion criteria. In the eICU-CRD, 3743 patients with sepsis were identified and classified as external validation cohort (Multimedia Appendix 4). The clinical characteristics of included patients are described in Table 1. In the MIMIC-IV, the median age was 68 (IQR 58-78) years, with a median SOFA score of 4 (IQR 3-5). The median NEE was 0.12 (IQR 0.07-0.25) µg/kg/minute, and the in-hospital mortality was 16.4% (n=1501). For the eICU-CRD, the median NEE was 0.15 (IQR 0.07-0.34) µg/kg/minute, and the in-hospital mortality was 22.9% (n=857). Univariate analysis was used to characterize the factors distinguishing the subpopulations of the 3 trajectories (Multimedia Appendix 5).

XSL•FO
RenderX

Table 1. Characterization of cohorts.

Characteristics	MIMIC-IV[a] (n=9135)	eICU-CRD[b] (n=3743)	*P* value
Age (years), median (IQR)	68 (58-78)	68 (58-79)	.48
Male, n (%)	5482 (60)	1887 (50.4)	<.001
BMI (kg/m^2), median (IQR)	27.7 (24.1-32.7)	27.3 (23.0-33.1)	<.001
Emergency admission, n (%)	4635 (50.7)	2426 (64.8)	<.001
Pulmonary infection, n (%)	3643 (39.9)	1308 (34.9)	<.001
Charlson comorbidity index, median (IQR)	6 (4-8)	__[c]	—
SOFA[d] score, median (IQR)	4 (3-5)	7 (5-10)	<.001
NEE[e] (max) (µg/kg/min), median (IQR)	0.12 (0.07-0.25)	0.15 (0.07-0.34)	<.001
Use of CRRT[f], n (%)	515 (5.6)	80 (2.1)	<.001
Trajectories of sepsis, n (%)			<.001
Recovery	5835 (63.9)	2151 (57.5)	
Persistent ill	2937 (32.1)	1283 (34.3)	
Rapid death	363 (4)	309 (8.2)	
In-hospital mortality, n (%)	1501 (16.4)	857 (22.9)	<.001
28-day mortality, n (%)	1386 (15.2)	—	—

[a]MIMIC-IV: Medical Information Mart for Intensive Care Database-IV.
[b]eICU-CRD: eICU Collaborative Research Database.
[c]Not available.
[d]SOFA: Sequential Organ Failure Assessment.
[e]NEE: norepinephrine equivalence.
[f]CRRT: continuous renal replacement therapy.

Trajectories of Sepsis

Patients in the MIMIC-IV exhibited the following outcomes within a 48-hour period after enrollment: 5835 (63.9%) patients experienced recovery, 2937 (32.1%) patients remained persistent ill, and 363 (4%) patients experienced rapid death. In the external validation cohort, that is, eICU-CRD, 2151 (57.5%) patients experienced recovery, 1283 (34.3%) patients remained persistent ill, and 309 (8.2%) patients experienced rapid death. In patients who recovered, the in-hospital mortality rates in the MIMIC-IV and eICU-CRD were 5.8% (n=338) and 8.5% (n=183), respectively. For patients with persistent ill, the in-hospital mortality rates in the MIMIC-IV and eICU-CRD were 27.2% (n=800) and 28.4% (n=365), respectively (Figure 2). Compared to patients with recovery or persistent ill, patients with rapid death had sustained high NEE, high heart rate, and low systolic blood pressure (SBP; Figure 3).

Figure 2. Trajectories of patients with sepsis. eICU-CRD: eICU Collaborative Research Database; MIMIC-IV: Medical Information Mart for Intensive Care Database-IV.

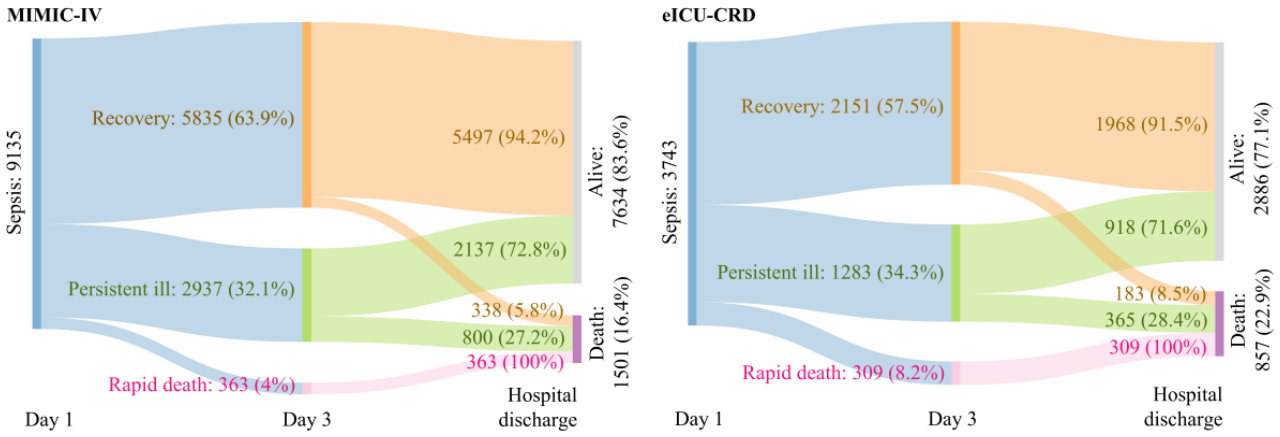

XSL•FO
RenderX

Figure 3. Trends of typical parameters over time according to sepsis trajectories 24 hours after identification. Points represent the median parameter values in an hourly interval. The intervals represent the 25th and 75th percentiles of the parameter values. Cubic polynomial smoothing splines highlight trends. NEE: norepinephrine equivalence; SBP: systolic blood pressure.

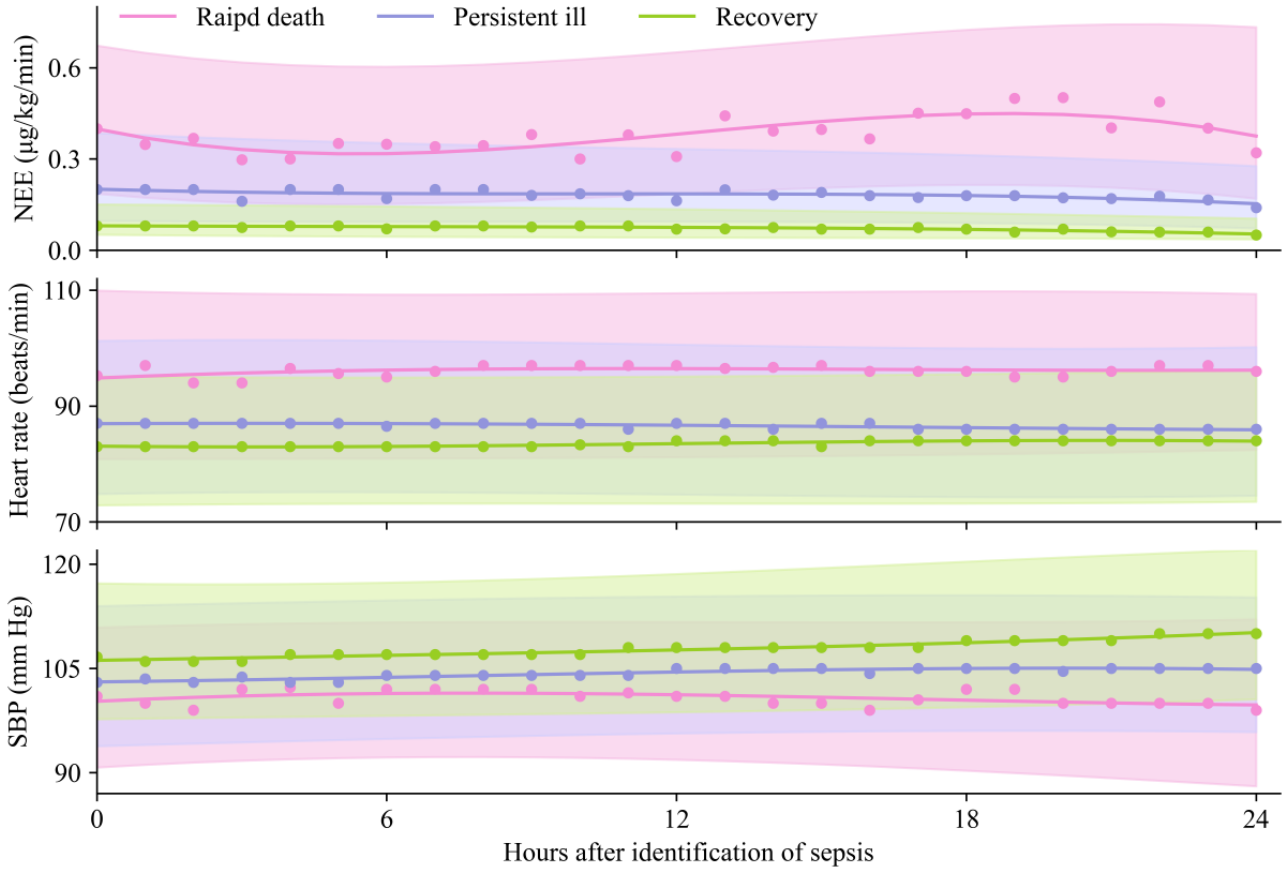

Predictive Performance of AI Model

Of the 4 machine learning architectures in 3 model setups for predicting persistent ill on internal validation, GBM had the highest consistent performance of mean AUROC (SD) ranging from 0.806 (0.010) to 0.807 (0.010) compared to NDF ranging from 0.779 (0.013) to 0.802 (0.011), RF ranging from 0.791 (0.013) to 0.795 (0.010), and LR ranging from 0.719 (0.031) to 0.790 (0.007; Multimedia Appendix 6). Of the 3 model setups, the 3-way multiclass with GBM performed best with the mean AUROC of 0.807 (SD 0.010). In addition, the multiclass setup could also provide a probability risk score for each trajectory, with a mean AUROC (SD) of 0.906 (0.018) for rapid death and 0.843 (0.008) for recovery. In the external validation cohort, the mean AUROCs (SDs) were 0.878 (0.003) for rapid death, 0.764 (0.008) for recovery, and 0.696 (0.007) for persistent ill (Multimedia Appendix 7).

When evaluating the discrimination performance of the multiclass GBM under 4 feature subsets in identifying persistent ill patients, on the internal validation cohort, the results showed that the model achieved AUROCs of 0.664-0.807 from using NEE (max) only to using all 148 features (Figure 4A). When applied to the external validation data set, the results were AUROCs of 0.561-0.696. Figure 4B shows the calibration plot. The full metric results in different feature subsets of all 3 trajectories are presented in Multimedia Appendix 8.

XSL•FO
RenderX

Figure 4. Performance results of the 3-way multiclass gradient boosting machine model. (A) Receiver operating characteristic curves for the model across different feature subsets in predicting persistent ill. (B) Calibration plots of each outcome for the model using all the features. The predicted probabilities were binned into deciles, and the mean and 95% CI were shown in each decile. AUROC: area under the receiver operating characteristic curve; NEE: norepinephrine equivalence.

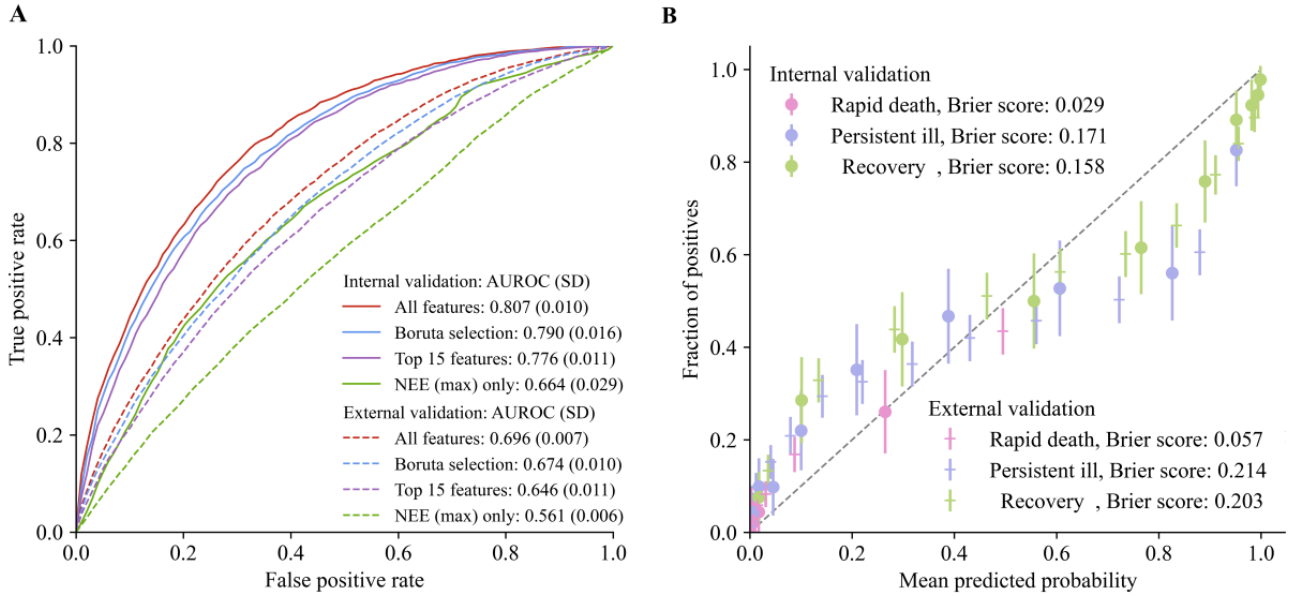

Model Explanation

The global feature importance for the multiclass GBM model using all the features is shown in Multimedia Appendix 9. The visible, individually interpretable summary of the impact of features across patients showed that oliguria, higher NEE (max), unstable oxygen saturation, higher lactate, and lower SBP were associated with a higher risk of rapid death; higher NEE (max), higher Acute Physiology Score III score, more IV fluid administrated, pulmonary infection, and lower SBP were associated with a higher risk of persistent illness. In contrast, patients with lower NEE (max), lower Acute Physiology Score III, higher SBP, more urine output, and less IV fluid administrated were more likely to recover (Figure 5A). For the example of explaining the individual prediction for different trajectories, see Figures 5B-5D.

XSL•FO
RenderX

Figure 5. Model explanations. (A) Individual explanation summary of the top 15 clinical features for each trajectory. Beeswax plots show the feature importance across patients for the top 15 features, where each dot represents the feature importance value for 1 patient sample. Where multiple points fall on the same x position, they are stacked to show density. Features with positive impact values push risk up, while negative impact values push risk down. Long tails indicate features that are extremely important for some patients. Explanation of the output of the risk score for (B) a patient who did not survive, (C) a patient whose trajectory was persistent ill, and (D) a patient who recovered. APS: Acute Physiology Score; aPTT: activated partial thromboplastin time; BUN: blood urea nitrogen; GCS: Glasgow Coma Scale; HCT: hematocrit; IV: intravenous; MAP: mean arterial pressure; NEE: norepinephrine equivalence; PT: prothrombin time; SBP: systolic blood pressure; SHAP: Shapley Additive Explanations; SpO2: oxygen saturation; UOP: urine output; vr: variation; WBC: white blood cell.

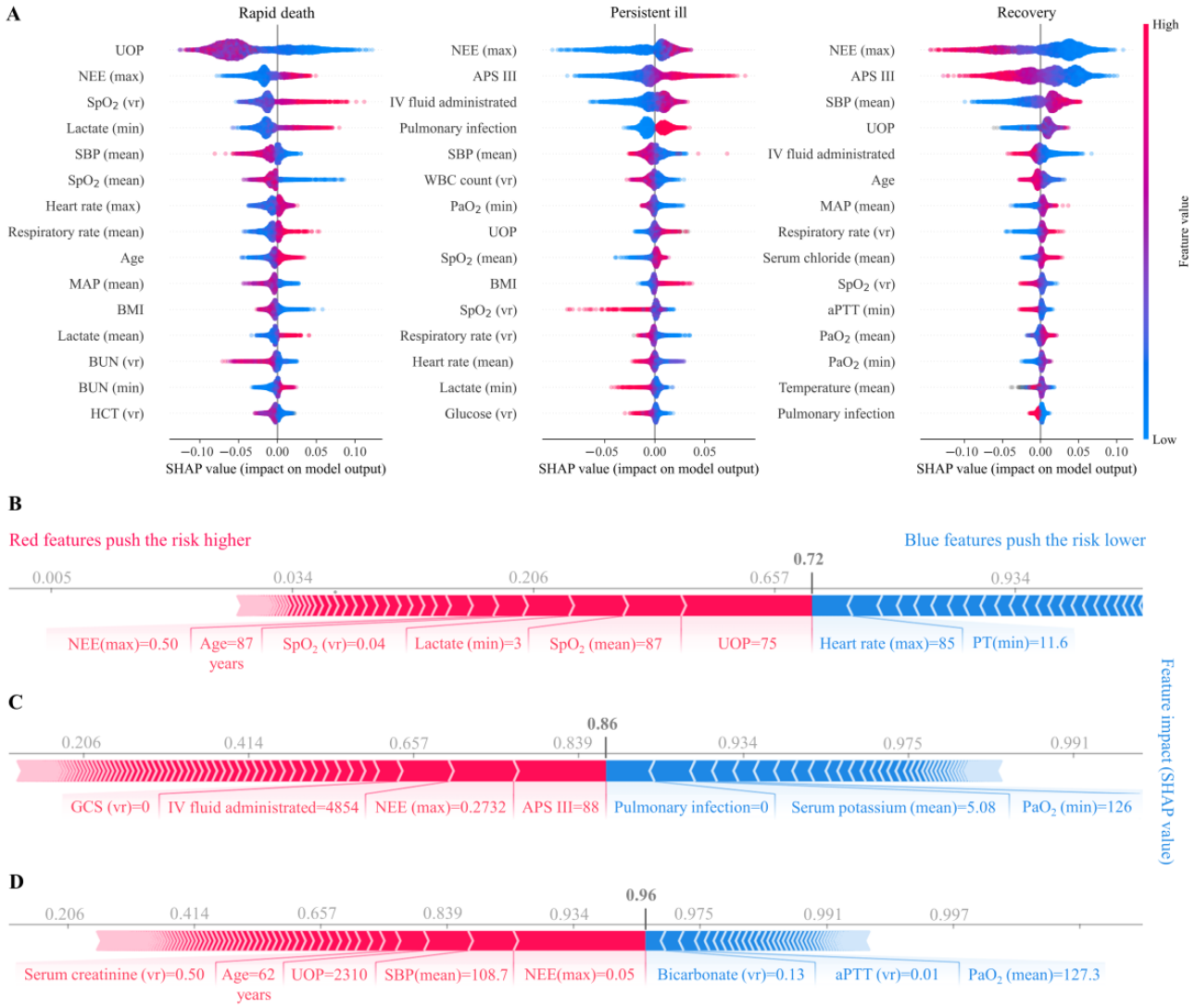

Conformal Prediction

The calibration curves of the observed prediction error were analyzed at different significance levels (1–confidence), ranging from 0% to 100% (Figure 6). When the model was evaluated using a mixed confidence approach, with a confidence level of 85% for recovery and a confidence level of 75% for rapid death and persistent ill, the model with CP reduced overall prediction errors by 27.6% (n=62) and 30.7% (n=412) in the internal and external validation cohorts, respectively, compared with the model without CP (Table 2). Specifically for the predictions in persistent ill patients, it made 69 (23.5%) errors in the internal validation cohort (compared with n=122, 41.5% errors without CP). In the external validation, it still produced significantly lower error rates compared with the model without CP (n=398, 31% vs n=689, 53.7%).

XSL•FO
RenderX

Figure 6. Plots of conformal prediction for the 2 validation cohorts. (A and B) The calibration curves depict the relationship between the observed prediction error (which represents the fraction of true labels not encompassed within the prediction region) on the y-axis and the prespecified significance level (the tolerated error rate) on the x-axis. The conformal predictor was well-calibrated for internal validation, while the calibration curve for external populations deviated from the ideal diagonal line but was relatively good for the persistent ill when the significance level was below 0.25. (C and D) Label distribution plots at different prespecified significance levels, with the model incorporating more multiple predictions at lower significance (higher confidence) levels. The blue dotted line indicates the corresponding significance level that produces the highest number of single-label predictions. The black dotted line indicates the corresponding significance level that first produces no multiple predictions. Single predictions output {rapid death} or {persistent ill} or {recovery}, multiple predictions output {rapid death, persistent ill}, {rapid death, recovery} or {persistent ill, recovery} or {rapid death, persistent ill, recovery}, and empty means {null}. eICU-CRD: eICU Collaborative Research Database; MIMIC-IV: Medical Information Mart for Intensive Care Database-IV.

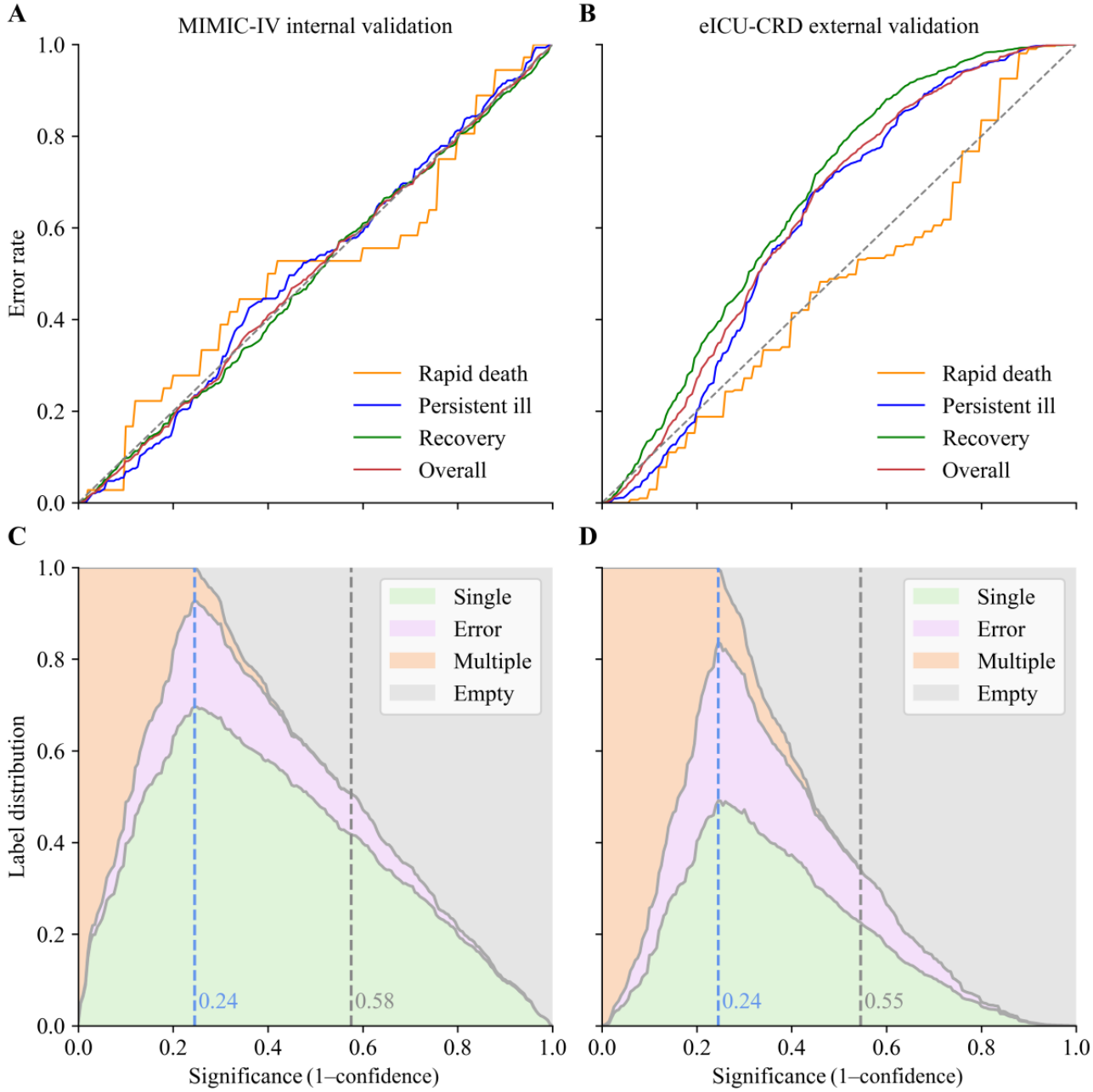

XSL•FO
RenderX

Table 2. Prediction regions on the internal validation and external validation cohort.

Prediction region	Internal validation cohort				External validation cohort			
	Rapid death (n=36), n (%)	Persistent ill (n=294), n (%)	Recovery (n=584), n (%)	Overall (n=914), n (%)	Rapid death (n=309), n (%)	Persistent ill (n=1283), n (%)	Recovery (n=2151), n (%)	Overall (n=3743), n (%)
Confidence level: 90%								
Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Error	6 (16.7)	20 (6.8)	57 (9.8)	83 (9.1)	9 (2.9)	83 (6.5)	292 (13.6)	384 (10.3)
Single	3 (8.3)	48 (16.3)	312 (53.4)	363 (39.7)	12 (3.9)	28 (2.2)	537 (25)	577 (15.4)
Multiple	27 (75)	226 (76.9)	215 (36.8)	468 (51.2)	288 (93.2)	1172 (91.3)	1322 (61.5)	2782 (74.3)
Confidence level: 85%								
Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Error	8 (22.2)	35 (11.9)	84 (14.4)	127 (13.9)	34 (11)	163 (12.7)	474 (22)	671 (17.9)
Single	4 (11.1)	107 (36.4)	369 (63.2)	480 (52.5)	31 (10)	184 (14.3)	853 (39.7)	1068 (28.5)
Multiple	24 (66.7)	152 (51.7)	131 (22.4)	307 (33.6)	244 (79)	936 (73)	824 (38.3)	2004 (53.5)
Confidence level: 80%								
Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Error	10 (27.8)	48 (16.3)	114 (19.5)	172 (18.8)	58 (18.8)	259 (20.2)	696 (32.4)	1013 (27.1)
Single	7 (19.4)	160 (54.4)	410 (70.2)	577 (63.1)	53 (17.2)	402 (31.3)	1053 (49)	1508 (40.3)
Multiple	19 (52.8)	86 (29.3)	60 (10.3)	165 (18.1)	198 (64.1)	622 (48.5)	402 (18.7)	1222 (32.6)
Confidence level: 75%								
Empty	0 (0)	1 (0.3)	2 (0.3)	3 (0.3)	1 (0.3)	11 (0.9)	16 (0.7)	28 (0.7)
Error	10 (27.8)	68 (23.1)	132 (22.6)	210 (23)	57 (18.4)	387 (30.2)	839 (39)	1283 (34.3)
Single	10 (27.8)	178 (60.5)	448 (76.7)	636 (69.6)	85 (27.5)	488 (38)	1263 (58.7)	1836 (49.1)
Multiple	16 (44.4)	47 (16)	2 (0.3)	65 (7.1)	166 (53.7)	397 (30.9)	33 (1.5)	596 (15.9)
Mixed confidence level[a]								
Empty	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Error	10 (27.8)	69 (23.4)	84 (14.4)	163 (17.8)	58 (18.8)	398 (31)	474 (22)	930 (24.8)
Single	9 (25)	134 (45.6)	450 (77)	593 (64.9)	76 (24.6)	294 (22.9)	1279 (59.5)	1649 (44.1)
Multiple	17 (47.2)	91 (31)	50 (8.6)	158 (17.3)	175 (56.6)	591 (46.1)	398 (18.5)	1164 (31.1)
AI[b] point prediction								
Error	22 (61.1)	122 (41.5)	81 (13.9)	225 (24.6)	198 (64.1)	689 (53.7)	455 (21.2)	1342 (35.9)
Correct	14 (38.9)	172 (58.5)	503 (86.1)	689 (75.4)	111 (35.9)	594 (46.3)	1696 (78.8)	2401 (64.1)

[a]85% for recovery, 75% for rapid death and persistent ill. The model demonstrates relatively improved performance in recognizing recovery due to a number of training examples for this class. As a result, when applying the conformal predictor, we can assign higher confidence levels within this class without generating excessively broad and uninformative prediction intervals.

[b]AI: artificial intelligence.

Although the number of accurately predicted persistent ill cases (efficiency) was lower compared to the conventional model without CP (n=134, 45.6% vs n=172, 58.5% in the internal cohort and n=294, 22.9% vs n=594, 46.3% in the external cohort), focusing on patients with multiple predictions of {rapid death, persistent ill} proved beneficial. Because among these predictions, 46 of 62 (74.2%) predictions were persistent ill in the internal validation cohort, and 351 of 496 (70.8%) predictions were persistent ill in the external validation cohort (Table 3).

XSL•FO
RenderX

Table 3. Multiple prediction analysis when set the mixed confidence level.

Multiple prediction[a]	Internal validation cohort				External validation cohort			
	Rapid death (n=17), n (%)	Persistent ill (n=91), n (%)	Recovery (n=50), n (%)	Total (n=158), n (%)	Rapid death (n=175), n (%)	Persistent ill (n=591), n (%)	Recovery (n=398), n (%)	Total (n=1164), n (%)
{0, 1}	16 (25.8)	46 (74.2)	—[b]	62 (39.2)	145 (29.2)	351 (70.8)	—	496 (42.6)
{0, 2}	1 (33.3)	—	2 (66.7)	3 (1.9)	15 (23.1)	—	50 (76.9)	65 (5.6)
{1, 2}	—	44 (48.9)	46 (51.1)	90 (57)	—	194 (40)	291 (60)	485 (41.7)
{0, 1, 2}	0 (0)	1 (33.3)	2 (66.7)	3 (1.9)	15 (12.7)	46 (39)	57 (48.3)	118 (10.1)

[a]Multiple prediction means that the prediction is uncertain, and the model could not distinguish between the possible class labels, thus awaiting the human review. Label 0 means rapid death, label 1 means persistent ill, and label 2 means recovery.

[b]Not applicable.

Additionally, a web calculator was developed to enable users to visually understand the outputs of our AI model (Figure 7). This web-based calculator is internally deployed and is currently being updated at the Department of Critical Care Medicine, Zhongda Hospital. The web calculator is only accessible using a private URL by clinicians within the hospital and is connecting to the hospital's EHR system. After a patient is identified as having sepsis and successfully collects 24-hour data after sepsis onset, the calculator would include this patient in the list. The users can then select 1 patient and view the relevant features that are calculated automatically based on data sourced from the hospital's EHR system. After that, users can view the risk score predicted by this calculator and enter the confidence level to get a reliable prediction. The confidence level could be adjusted according to the model performance and workload, as a number of uncertain predictions can lead to an unmanageable situation.

XSL•FO
RenderX

Figure 7. User interface of the web-based calculator. (A) Users can view the characteristics of the selected patient with sepsis and 24-hour collected feature values. (B) Generate risk scores for each sepsis trajectory by pressing the "Predict the trajectory" button. Then, press the "Conformal prediction" button, enter the confidence level for each sepsis trajectory, and press the "Confirm" button to output the final prediction under the conformal prediction framework. If the result was a multiple prediction, the users could press the "More information" button to go back to the original hospital electronic health record system to get more information for making a further decision, as the multiple prediction means that the prediction is uncertain as the model could not distinguish between the possible class labels and thus await human review. Users could also press "Quit" to leave the system. AI: artificial intelligence; APS: Acute Physiology Score; IV: intravenous; MAP: mean arterial pressure; NEE: norepinephrine equivalence; SBP: systolic blood pressure; SpO2: oxygen saturation; UOP: urine output.

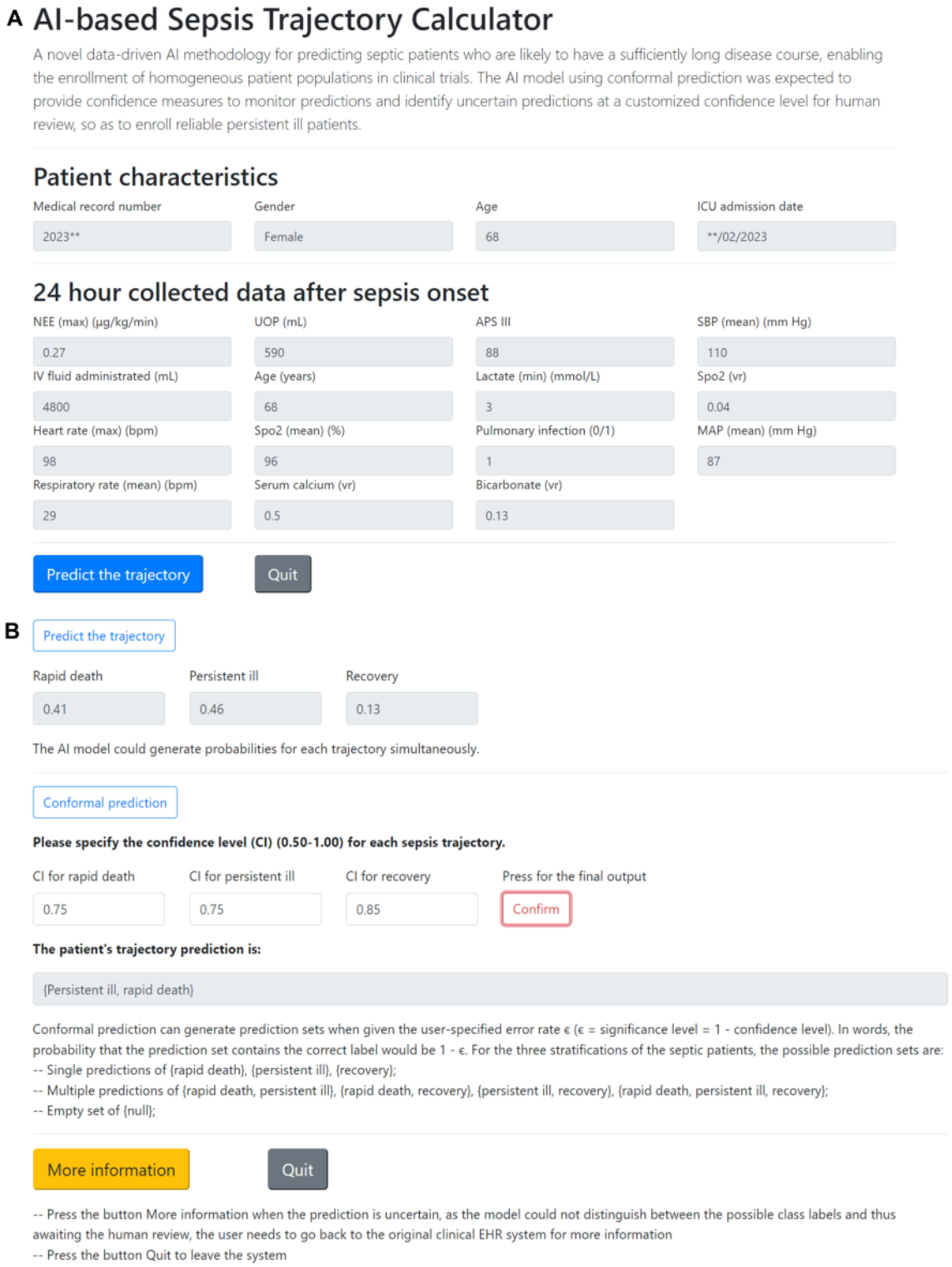

XSL•FO
RenderX

Discussion

Principal Findings

We developed a novel AI-based model to predict patients with sepsis likely to have a long disease course. Such a model improves the homogeneity of patients with sepsis enrolled in clinical trials and ensures that a patient recruited into a trial will still be persistently ill by the time the proposed therapy can impact patient outcome. Furthermore, we used CP to estimate the uncertainty of the model outputs, which allows for a more comprehensive understanding of the model's reliability and assists in making informed decisions based on the predicted outcomes. In addition, the developed web-based calculator may assist clinicians in comprehending the relationship between input data and predicted output. This could facilitate feedback from the clinicians and help improve the model predictions, which may be more effectively translated into medical practice for clinical decision-making.

To identify innovative pharmacological treatments for sepsis, enrolling patients who are most likely to benefit is essential. The enrichment strategy in this study differed from the prior prognostic and predictive enrichment strategies. In the Efficacy of Xuebijing Injection in Patients With Sepsis RCT, considering that the mortality rate was 74.3% in patients with a SOFA score of more than 13 in the Chinese population, Liu et al [25] used a prognostic enrichment strategy to enroll patients with sepsis with a SOFA score of 2 to 13 to evaluate the treatment effect of Xuebijing injection on 28-day mortality. However, this approach fails to consider patients with rapid recovery trajectories who may not derive significant benefits from Xuebijing injection treatment. Whether the targeted SOFA score can be effectively generalized in other trials remains uncertain. One important constraint with the predictive enrichment strategy is that different therapeutics necessitate different populations [26,27], and we lack the prior knowledge to accurately determine which patients are most likely to benefit from a specific therapy. Moreover, the limited availability of point-of-care devices for the identification of targeted patients, coupled with the lengthy measurement turnaround time, creates considerable obstacles for clinical trial enrichment [28]. These challenges are further exacerbated by the rapidly changing nature of diseases observed within clinical trials. From another perspective, we stratified patients based on the disease course of sepsis and suggested that future sepsis clinical trials should prioritize including patients with persistent ill, who would have a sufficiently long disease course to benefit from a therapy. Additionally, the current enrichment strategy does not necessarily take into account the different infection sites, pathogens, and pharmacological interventions.

Using real-time eHealth data, we developed an AI-based model for patient enrollment in sepsis clinical trials. Of the 12 machine learning models tested, a multiclass GBM model performed best and was able to provide a probability risk for each trajectory. The lightweight model using the top 15 features also showed acceptable performance and could be more suitable for future clinical applications. In addition, this model overcame the challenges of a heterogeneous patient population and tailored patient stratification to the needs of a particular drug candidate. Meanwhile, the predictive window, defined as the period between identification and enrollment, was 24 hours for the AI model, which was consistent with the previous sepsis RCTs [29]. It is reasonable to predict the disease course of patients within 24 hours after the identification of sepsis onset in future trials.

To mitigate the risk of automation bias in AI and evaluate model uncertainty, we implemented a CP framework on top of the prediction algorithm. This framework enabled the model to generate confidence measures, allowing for the monitoring of predictions and identification of uncertain outcomes based on adjustable confidence levels for human review. This is particularly crucial in different clinical scenarios where the AI system should be capable of acknowledging its limitations and saying "I don't know" [30]. This could help to reduce prediction errors, especially in external validation, where the model performance could not be guaranteed, as well as in internal validation due to factors such as population heterogeneity and differences in clinical practice [31]. In addition, this study demonstrated that flagging the multiple (uncertain) predictions that the model could not classify for human review proved valuable, as it enabled the identification of more potentially persistent ill patients while simultaneously identifying high-risk individuals. In this case, the initial screening of patients could be achieved, allowing clinicians to avoid having to manually screen each patient for eligibility in the trial. Instead, they only need to review and assess the uncertain results of the model. As discussed by Goodman et al [32], clinicians were supposed to integrate their clinical experience with AI model outputs for clinical decision-making, thus avoiding inappropriate reliance on algorithms alone.

Limitations

Our study faces some limitations. First, although this retrospective study demonstrates that the proposed AI-based enrichment strategy is effective in reducing patient heterogeneity in the design of sepsis clinical trials, we still need to include more patients to perform further prospective external validation using the developed web-based calculator. Second, the definition of sepsis onset was different between MIMIC-IV and eICU-CRD since the onset time of infection and organ failure in eICU-CRD were missing. Third, considering that our AI-based model needed a 24-hour predictive window after identifying sepsis onset, this model is not available for the mandated emergency therapies for sepsis, including the administration of antibiotics and IV fluid. The final limitation arises from the absence of criteria for balancing the trade-off between generating single-label and multiple predictions. Avoiding an excessive number of multiple (uncertain) predictions is important, as this can lead to an unmanageable situation.

Conclusions

Our interpretable AI-based model using clinical data accurately identifies patients with different disease courses, which can reduce the heterogeneity of patients with sepsis in future clinical trials. Additionally, we estimate the uncertainty of the model outputs and enable the performance of the model using a mixed

XSL•FO
RenderX

confidence approach. Application of this model can identify more homogeneous patient populations and precision pharmacological treatments for sepsis trials.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (62171123, 62211530112, and 62071241), Yilu "Gexin"-Fluid Therapy Research Fund Project (YLGX-ZZ-2020002), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0088), and the Open Project Program of Key Laboratory of Big Data Analysis and Knowledge Services, Yangzhou City, Yangzhou University (YBK 202208).

Data Availability

The data sets and code generated and analyzed during this study are available on the PhysioNet website [33,34] and GitHub repository [35], respectively.

Authors' Contributions

HC and MY had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. HC, ZZ, CL, JZ, and MY conceived and designed the study. HC, MY, WH, JL, and YW collected the data. HC, MY, and YW did the data analysis and interpretation. HC and MY drafted the manuscript. ZZ, CL, and JZ revised the manuscript. The final version of the manuscript was read and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist: prediction model development and validation.
[DOCX File , 22 KB - jmir_v26i1e54621_app1.docx]

Multimedia Appendix 2
Static and dynamic variables extracted and corresponding range.
[DOCX File , 33 KB - jmir_v26i1e54621_app2.docx]

Multimedia Appendix 3
Methods details.
[DOCX File , 103 KB - jmir_v26i1e54621_app3.docx]

Multimedia Appendix 4
Study population and exclusion criteria.
[DOCX File , 107 KB - jmir_v26i1e54621_app4.docx]

Multimedia Appendix 5
Top univariate associations with patient trajectory.
[DOCX File , 23 KB - jmir_v26i1e54621_app5.docx]

Multimedia Appendix 6
Model performance of the 12 machine learning model architecture and setup combinations.
[DOCX File , 44 KB - jmir_v26i1e54621_app6.docx]

Multimedia Appendix 7
Performance of the 3-way multiclass gradient boosting machine model using all the features for each outcome.
[DOCX File , 918 KB - jmir_v26i1e54621_app7.docx]

Multimedia Appendix 8
Results of the of multiclass gradient boosting machine in different feature selections in predicting different outcomes.
[DOCX File , 35 KB - jmir_v26i1e54621_app8.docx]

XSL•FO
RenderX

Multimedia Appendix 9
The overall global feature importance for the model using all the features.
[DOCX File , 121 KB - jmir_v26i1e54621_app9.docx]

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet 2020;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]
3. Reddy K, Sinha P, O'Kane CM, Gordon AC, Calfee CS, McAuley DF. Subphenotypes in critical care: translation into clinical practice. Lancet Respir Med 2020;8(6):631-643. [doi: 10.1016/S2213-2600(20)30124-7] [Medline: 32526190]
4. Santacruz CA, Pereira AJ, Celis E, Vincent JL. Which multicenter randomized controlled trials in critical care medicine have shown reduced mortality? A systematic review. Crit Care Med 2019;47(12):1680-1691. [doi: 10.1097/CCM.0000000000004000] [Medline: 31567349]
5. Maslove DM, Tang B, Shankar-Hari M, Lawler PR, Angus DC, Baillie JK, et al. Redefining critical illness. Nat Med 2022;28(6):1141-1148. [doi: 10.1038/s41591-022-01843-x] [Medline: 35715504]
6. Papazian L, Forel JM, Gacouin A, Penot-Ragon C, Perrin G, Loundou A, et al. Neuromuscular blockers in early acute respiratory distress syndrome. N Engl J Med 2010;363(12):1107-1116. [doi: 10.1056/NEJMoa1005372] [Medline: 20843245]
7. Villar J, Ferrando C, Martínez D, Ambrós A, Muñoz T, Soler JA, et al. Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. Lancet Respir Med 2020;8(3):267-276. [doi: 10.1016/S2213-2600(19)30417-5] [Medline: 32043986]
8. Villar J, Ambrós A, Mosteiro F, Martínez D, Fernández L, Ferrando C, et al. A prognostic enrichment strategy for selection of patients with acute respiratory distress syndrome in clinical trials. Crit Care Med 2019;47(3):377-385. [doi: 10.1097/CCM.0000000000003624] [Medline: 30624279]
9. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 2019;321(20):2003-2017 [FREE Full text] [doi: 10.1001/jama.2019.5791] [Medline: 31104070]
10. Bhavani SV, Semler M, Qian ET, Verhoef PA, Robichaux C, Churpek MM, et al. Development and validation of novel sepsis subphenotypes using trajectories of vital signs. Intensive Care Med 2022;48(11):1582-1592 [FREE Full text] [doi: 10.1007/s00134-022-06890-z] [Medline: 36152041]
11. Schwager E, Jansson K, Rahman A, Schiffer S, Chang Y, Boverman G, et al. Utilizing machine learning to improve clinical trial design for acute respiratory distress syndrome. NPJ Digit Med 2021;4(1):133 [FREE Full text] [doi: 10.1038/s41746-021-00505-5] [Medline: 34504281]
12. Maddali MV, Churpek M, Pham T, Rezoagli E, Zhuo H, Zhao W, et al. Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. Lancet Respir Med 2022;10(4):367-377 [FREE Full text] [doi: 10.1016/S2213-2600(21)00461-6] [Medline: 35026177]
13. Olsson H, Kartasalo K, Mulliqi N, Capuccini M, Ruusuvuori P, Samaratunga H, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nat Commun 2022;13(1):7761 [FREE Full text] [doi: 10.1038/s41467-022-34945-8] [Medline: 36522311]
14. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
15. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
16. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med 2020;48(2):210-217 [FREE Full text] [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
17. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA, United States p. 785-794. [doi: 10.1145/2939672.2939785]
18. Kontschieder P, Fiterau M, Criminisi A, Bulò SR. Deep neural decision forests. 2015 Presented at: IEEE International Conference on Computer Vision (ICCV); December 7, 2015; Santiago, Chile p. 7-13. [doi: 10.1109/iccv.2015.172]
19. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
20. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med 2019;47(11):1485-1492 [FREE Full text] [doi: 10.1097/CCM.0000000000003891] [Medline: 31389839]

XSL•FO
RenderX

21. Yang M, Liu C, Wang X, Li Y, Gao H, Liu X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med 2020;48(11):e1091-e1096. [doi: 10.1097/CCM.0000000000004550] [Medline: 32885937]
22. Li J, Tian Y, Li R, Zhou T, Li J, Ding K, et al. Improving prediction for medical institution with limited patient data: leveraging hospital-specific data based on multicenter collaborative research network. Artif Intell Med 2021;113:102024. [doi: 10.1016/j.artmed.2021.102024] [Medline: 33685587]
23. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
24. Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. Taking the human out of the loop: a review of Bayesian optimization. Proc IEEE 2016;104(1):148-175. [doi: 10.1109/jproc.2015.2494218]
25. Liu S, Yao C, Xie J, Liu H, Wang H, Lin Z, et al. Effect of an herbal-based injection on 28-day mortality in patients with sepsis: the EXIT-SEP randomized clinical trial. JAMA Intern Med 2023;183(7):647-655 [FREE Full text] [doi: 10.1001/jamainternmed.2023.0780] [Medline: 37126332]
26. Meyer NJ, Reilly JP, Anderson BJ, Palakshappa JA, Jones TK, Dunn TG, et al. Mortality benefit of recombinant human interleukin-1 receptor antagonist for sepsis varies by initial interleukin-1 receptor antagonist plasma concentration. Crit Care Med 2018;46(1):21-28 [FREE Full text] [doi: 10.1097/CCM.0000000000002749] [Medline: 28991823]
27. Panacek EA, Marshall JC, Albertson TE, Johnson DH, Johnson S, MacArthur RD, et al. Efficacy and safety of the monoclonal anti-tumor necrosis factor antibody F(ab')2 fragment afelimomab in patients with severe sepsis and elevated interleukin-6 levels. Crit Care Med 2004;32(11):2173-2182. [doi: 10.1097/01.ccm.0000145229.59014.6c] [Medline: 15640628]
28. Prescott HC, Calfee CS, Thompson BT, Angus DC, Liu VX. Toward smarter lumping and smarter splitting: rethinking strategies for sepsis and acute respiratory distress syndrome clinical trial design. Am J Respir Crit Care Med 2016;194(2):147-155 [FREE Full text] [doi: 10.1164/rccm.201512-2544CP] [Medline: 27244481]
29. Sevransky JE, Rothman RE, Hager DN, Bernard GR, Brown SM, Buchman TG, et al. Effect of vitamin C, thiamine, and hydrocortisone on ventilator- and vasopressor-free days in patients with sepsis: the VICTAS randomized clinical trial. JAMA 2021;325(8):742-750 [FREE Full text] [doi: 10.1001/jama.2020.24505] [Medline: 33620405]
30. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". NPJ Digit Med 2021;4(1):134 [FREE Full text] [doi: 10.1038/s41746-021-00504-6] [Medline: 34504260]
31. Yang M, Liu S, Hao T, Ma C, Chen H, Li Y, et al. Development and validation of a deep interpretable network for continuous acute kidney injury prediction in critically ill patients. Artif Intell Med 2024;149:102785. [doi: 10.1016/j.artmed.2024.102785] [Medline: 38462285]
32. Goodman KE, Rodman AM, Morgan DJ. Preparing physicians for the clinical algorithm era. N Engl J Med 2023;389(6):483-487. [doi: 10.1056/NEJMp2304839] [Medline: 37548320]
33. Accessing MIMIC-IV data. Medical Information Mart for Intensive Care. URL: https://mimic.mit.edu/ [accessed 2022-10-28]
34. Accessing eICU-CRD data. eICU Collaborative Research Database. URL: http://eicu-crd.mit.edu/ [accessed 2022-10-28]
35. Accessing open-source code. GitHub. URL: https://github.com/Betapower/AI-based-sepsis-enrichment-strategy [accessed 2023-11-10]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
CP: conformal prediction
EHR: electronic health record
eICU-CRD: eICU Collaborative Research Database
GBM: gradient boosting machine
ICU: intensive care unit
IV: intravenous
LR: logistic regression
MIMIC-IV: Medical Information Mart for Intensive Care Database-IV
NDF: neural decision forest
NEE: norepinephrine equivalence
PPV: positive predictive value
RCT: randomized controlled trial
RF: random forest
SBP: systolic blood pressure
SOFA: Sequential Organ Failure Assessment
TPR: true positive rate
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

XSL•FO
RenderX

Edited by B Puladi; submitted 16.11.23; peer-reviewed by S Nemati, X Tong; comments to author 27.02.24; revised version received 22.04.24; accepted 21.07.24; published 04.09.24.

Please cite as:
Yang M, Zhuang J, Hu W, Li J, Wang Y, Zhang Z, Liu C, Chen H
Enhancing Patient Selection in Sepsis Clinical Trials Design Through an AI Enrichment Strategy: Algorithm Development and Validation
J Med Internet Res 2024;26:e54621
URL: https://www.jmir.org/2024/1/e54621
doi:10.2196/54621
PMID:

©Meicheng Yang, Jinqiang Zhuang, Wenhan Hu, Jianqing Li, Yu Wang, Zhongheng Zhang, Chengyu Liu, Hui Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Advanced Machine Learning Model for a Web-Based Artificial Intelligence–Based Clinical Decision Support System Application: Model Development and Validation Study

Tai-Han Lin[1*], MD; Hsing-Yi Chung[1*], MS; Ming-Jr Jian[1], MS, PhD; Chih-Kai Chang[1], MS; Cherng-Lih Perng[1], MS; Guo-Shiou Liao[2], MD; Jyh-Cherng Yu[2], MD; Ming-Shen Dai[3], MD, PhD; Cheng-Ping Yu[4*], MD, PhD; Hung-Sheng Shang[1*], MD, PhD

[1]Division of Clinical Pathology, Department of Pathology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[2]Division of General Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[3]Division of Oncology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei City, Taiwan

[4]Department of Pathology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Hung-Sheng Shang, MD, PhD
Division of Clinical Pathology
Department of Pathology
Tri-Service General Hospital, National Defense Medical Center
No. 161, Sec. 6, Minquan E. Road, Neihu District
Taipei, 11490
Taiwan
Phone: 886 920713130
Fax: 886 87927226
Email: iamkeith001@gmail.com

Abstract

Background: Breast cancer is a leading global health concern, necessitating advancements in recurrence prediction and management. The development of an artificial intelligence (AI)–based clinical decision support system (AI-CDSS) using ChatGPT addresses this need with the aim of enhancing both prediction accuracy and user accessibility.

Objective: This study aims to develop and validate an advanced machine learning model for a web-based AI-CDSS application, leveraging the question-and-answer guidance capabilities of ChatGPT to enhance data preprocessing and model development, thereby improving the prediction of breast cancer recurrence.

Methods: This study focused on developing an advanced machine learning model by leveraging data from the Tri-Service General Hospital breast cancer registry of 3577 patients (2004-2016). As a tertiary medical center, it accepts referrals from four branches—3 branches in the northern region and 1 branch on an offshore island in our country—that manage chronic diseases but refer complex surgical cases, including breast cancer, to the main center, enriching our study population's diversity. Model training used patient data from 2004 to 2012, with subsequent validation using data from 2013 to 2016, ensuring comprehensive assessment and robustness of our predictive models. ChatGPT is integral to preprocessing and model development, aiding in hormone receptor categorization, age binning, and one-hot encoding. Techniques such as the synthetic minority oversampling technique address the imbalance of data sets. Various algorithms, including light gradient-boosting machine, gradient boosting, and extreme gradient boosting, were used, and their performance was evaluated using metrics such as the area under the curve, accuracy, sensitivity, and F_1-score.

Results: The light gradient-boosting machine model demonstrated superior performance, with an area under the curve of 0.80, followed closely by the gradient boosting and extreme gradient boosting models. The web interface of the AI-CDSS tool was effectively tested in clinical decision-making scenarios, proving its use in personalized treatment planning and patient involvement.

Conclusions: The AI-CDSS tool, enhanced by ChatGPT, marks a significant advancement in breast cancer recurrence prediction, offering a more individualized and accessible approach for clinicians and patients. Although promising, further validation in diverse clinical settings is recommended to confirm its efficacy and expand its use.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56022) doi:10.2196/56022

KEYWORDS

breast cancer recurrence; artificial intelligence–based clinical decision support system; machine learning; personalized treatment planning; ChatGPT; predictive model accuracy

Introduction

Breast cancer, the most diagnosed cancer among women, presents a global health crisis with an extensive impact across diverse populations [1]. In 2022, more than 40,000 lives were taken due to breast cancer in the United States [2]. Characterized by its complexity and varied patient responses, breast cancer underscores the critical need for accurate prediction and management, particularly for recurrences [3]. This challenge necessitates the development of advanced personalized treatment strategies [3]. The effective prediction of recurrence significantly influences treatment decisions and patient outcomes [4].

Despite the numerous breast cancer prediction models, there is a notable gap in their practical applications [5]. Although scientifically sound, these models are often inaccessible to the broader medical community and patients [6]. This disconnection undermines their use in clinical settings, particularly in aiding treatment decisions [7]. The complexity of choosing an appropriate treatment strategy, even with guidelines, remains a challenge [3,7]. Guidelines are used to recommend the type of procedure, antiestrogen therapy, aromatase inhibitors, chemotherapy, and radiotherapy; however, the decision-making process using the guidelines can be daunting and complicated for both doctors and patients [3,4,7,8].

To bridge this gap, recent technological advancements have aimed to make predictive tools more accessible and user-friendly [9]. The move toward intuitive medical applications signifies a leap forward in the demystification of advanced tools for both clinicians and patients [9]. The introduction of artificial intelligence (AI) technologies, notably ChatGPT, promises to revolutionize this area [10]. With AI, the creation of efficient and intuitive predictive models has become increasingly feasible [11]. These advancements lay the groundwork for developing an AI-based clinical decision support system (AI-CDSS) that transforms the way clinicians and patients navigate their treatment choices [10].

The primary goal of this study was to use ChatGPT to develop a practical AI-CDSS tool for predicting breast cancer recurrence. This tool was designed to assist clinicians and patients in navigating the complex landscape of treatment decisions. By leveraging ChatGPT, the model simplifies the intricate data processes and feature selection, thereby enhancing the decision-making process regarding various therapy options. This study explores the development of this AI-enhanced tool, focusing on its application in clinical settings and its potential to significantly impact breast cancer management and patient care.

Methods

Research Design and Overview

This study focused on developing an advanced machine learning model to predict breast cancer recurrence. Central to this endeavor is the use of ChatGPT, which aims to enhance the model's accuracy and make it user-friendly for clinical decision support. The hypothesis underpinning this research is that ChatGPT integration will not only improve the model's effectiveness but also simplify its application, offering tangible benefits to medical professionals and patients in making informed treatment decisions.

Ethical Considerations

This study was conducted in strict accordance with the ethical standards outlined in the Declaration of Helsinki and applicable local regulations for the protection of human subjects. It was approved by the institutional review board of Tri-Service General Hospital, Taipei, Taiwan (TSGHIRB B202005044). Given the retrospective nature of the research, the requirement for informed consent was waived. All data analyzed were anonymized to ensure confidentiality.

Data Collection and Source Description

This study used data from the Tri-Service General Hospital's breast cancer registry in Taiwan, spanning from 2004 to 2016, with the study's retrospective nature allowing the institutional review board (TSGHIRB B202005044) to waive the requirement for patient consent. The initial data set comprised 4503 patients who had undergone surgical treatment at our main medical center, which is part of a network, including 3 branches in the northern region and 1 branch on an offshore island in our country. These branches primarily manage chronic diseases but refer complex surgical cases, including breast cancer, to our main center, thus enriching the diversity of our study population across different health care settings. Essential parameters collected included age, estrogen receptor (ER) and progesterone receptor (PR) percentages and intensities, pathological stage, tumor size, nodal involvement, metastasis, and detailed treatment type, such as surgery type, antiestrogen therapy, aromatase inhibitor usage, chemotherapy, and radiotherapy. To refine the data set for AI-CDSS model development, exclusion criteria were applied: incomplete hormone receptor data (394 patients), missing detailed treatment records (292 patients), and human epidermal growth factor receptor 2 (HER2) overexpression (240 patients), focusing the study on non-HER2 mediated recurrence. These exclusions resulted in a final cohort of 3577 patients, as shown in Figure 1, ensuring data quality and homogeneity for effective analysis.

XSL•FO
RenderX

Figure 1. Patient selection flowchart and model development. This figure provides a flowchart outlining the process of developing the AI-CDSS model using data from the Tri-Service General Hospital breast cancer registry. It starts with the initial number of patients with breast cancer who received surgical intervention during 2004-2016 and follows through the stages of exclusion criteria application, data preprocessing, feature selection, model training, performance evaluation, and web interface design. The flowchart also notes the assistance of ChatGPT with Python scripting for machine learning while keeping sensitive data on the local hospital computer, highlighting the integration of AI in model development. AI: artificial intelligence; AI-CDSS: artificial intelligence–based clinical decision support system; ER: estrogen receptor; HER2: human epidermal growth factor receptor 2; PR: progesterone receptor.

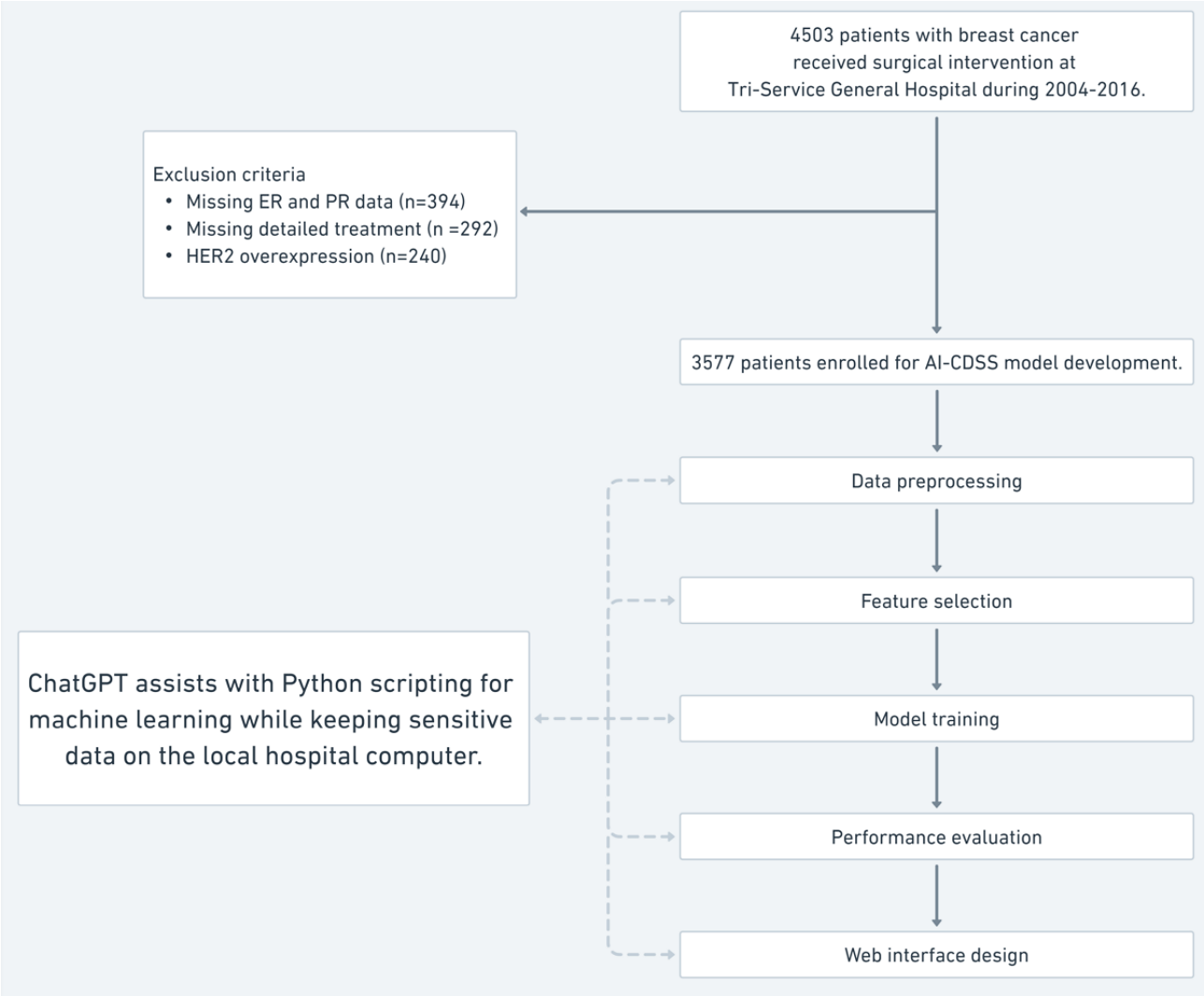

Integration of ChatGPT in AI-CDSS Model Development

Scripting and data processing were conducted on secure, local hospital systems using Python (version 3.9, Python Software Foundation) in Jupyter Notebooks (Project Jupyter), with no external data transfers or cloud processing to ensure the confidentiality and security of patients. ChatGPT's guidance was crucial in navigating the complexities of AI-CDSS development, enhancing task efficiency and accuracy on local systems. Below is a step-by-step breakdown of how ChatGPT supported each development process:

Data Preprocessing

ChatGPT assisted in crafting Python scripts for cleaning, normalizing numerical inputs, and encoding categorical variables, all of which were executed within the secure environment of local Jupyter Notebooks.

Feature Selection

Under ChatGPT's guidance, we developed scripts to generate heatmaps for visualizing data correlations and distributions, which aided effective feature selection. For feature engineering, hormone receptors were categorized into detailed intervals, age was segmented by a 50-year threshold, and treatment types and tumor characteristics were appropriately formatted for further analysis.

Model Training, Validation, and Performance Evaluation

Model training and performance evaluation scripting were supported by ChatGPT. It assisted in setting and optimizing parameters for various machine-learning algorithms through Jupyter Notebooks on local hospital computers. ChatGPT helped develop scripts for evaluating model performance and obtaining detailed reports and visualizations, such as confusion matrices and receiver operating characteristic (ROC) curves, ensuring comprehensive and precise analysis.

XSL•FO
RenderX

AI-CDSS Web Interface Design

Guidance from ChatGPT was instrumental in developing scripts for a secure web interface that allows clinicians and patients to interact with the AI-CDSS efficiently by using local computing resources.

Data Preprocessing

Data preprocessing was conducted using Python 3.9 in Jupyter Notebooks on secure local hospital systems, adhering to health care data protection regulations. We transformed the raw CSV data into Pandas DataFrames, enhancing data set manageability and facilitating accurate variable classification as categorical or continuous. ChatGPT helped create a function that appended a new column to the DataFrames to serve as the outcome variable for our models. This function calculated the time difference between the surgical resection date and any recorded recurrence date. If no recurrence was recorded or occurred more than 5 years post surgery, the outcome was classified as 0; otherwise, it was classified as 1.

Feature Selection

When developing our machine learning models, efficient feature selection was pivotal for optimizing performance. We developed Python scripts for generating and analyzing a heatmap, which significantly enhanced our understanding of the relationships between variables and 5-year breast cancer recurrence. The heatmap, shown in Figure 2, illustrates the correlations between various parameters such as ER percentage, PR percentage, surgical types, and treatment modalities with the occurrence of breast cancer recurrence within 5 years. The color gradients in the heatmap represent varying strengths of these correlations, providing a visual and quantitative method to identify and select the most predictive features for our models. This approach streamlined the feature selection process and ensured that our models were trained on variables most relevant to predicting early recurrence, enhancing their accuracy and clinical relevance.

Figure 2. Correlation heatmap of parameters with 5-year breast cancer recurrence. The heat map depicts the correlations between various parameters and the occurrence of 5-year breast cancer recurrence. This effectively shows the relevance of each parameter in predicting breast cancer recurrence and serves as a crucial tool for identifying and selecting key features for the machine learning model. The color gradients in the heatmap correspond to the varying strengths of these correlations. ER: estrogen receptor; PR: progesterone receptor.

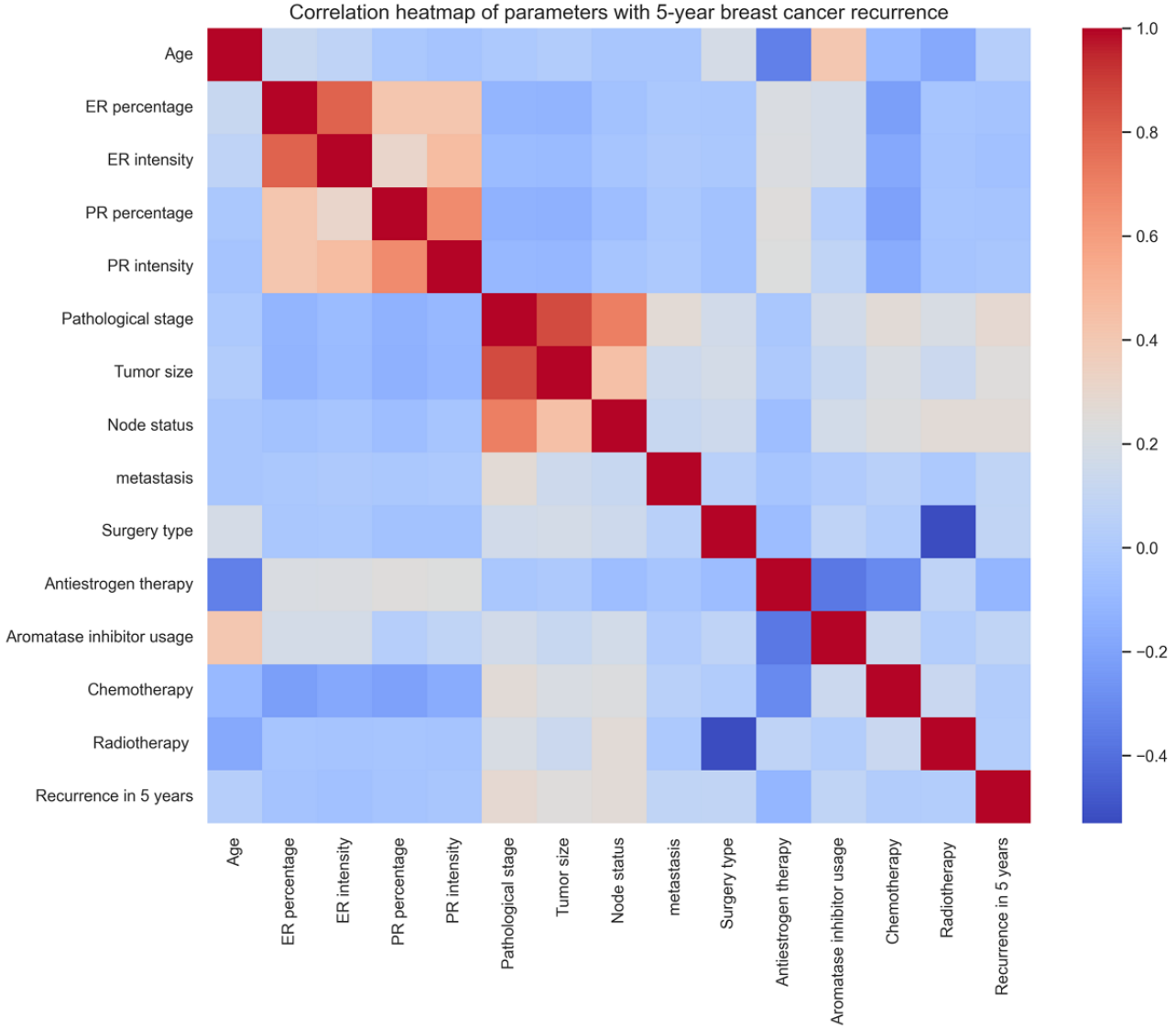

XSL•FO
RenderX

After visualizing relationships between variables and 5-year breast cancer recurrence through heatmaps, we used feature engineering for optimal data preparation for machine learning analysis. This process involved the intricate categorization and transformation of key variables. ER and PR receptor percentages were divided into negative and positive groups, with the latter further divided into 10 incremental categories at 10% intervals, enhancing the granularity for recurrence outcome prediction. Age was categorized into 2 groups using a 50-year threshold to differentiate risk levels associated with younger and older age groups. We applied one-hot encoding to categorical data, including treatment types and tumor characteristics, ensuring all variables were accurately represented in a machine-readable format. These feature engineering steps were crucial for developing robust predictive models, enabling the models to effectively use the most relevant and precisely formatted data.

Model Training, Validation, and Performance Evaluation

Model training used patient data in 2004-2012, with subsequent validation using data in 2013-2016, ensuring comprehensive assessment and robustness of our predictive models. Under ChatGPT's guidance, we selected a range of advanced machine learning algorithms to effectively tackle the complexities inherent in breast cancer recurrence data. The algorithms chosen included light gradient boosting machine (LGBM), gradient boosting (GB), extreme gradient boosting (XGB), random forest (RF), AdaBoost, and artificial neural networks, each selected for their unique strengths in addressing different aspects of the data set [12,13]. LGBM is a highly efficient and scalable gradient-boosting framework optimized for fast training, low memory usage, and high accuracy, capable of handling large data sets. GB is a powerful machine learning technique that combines multiple weak models, sequentially correcting their errors to improve accuracy but requires significant computational resources. XGB is an efficient, high-accuracy machine learning algorithm with early stopping and built-in regularization to prevent overfitting, well-suited for large data sets. RF is an ensemble machine learning model that builds multiple decision trees using random subsets of features and data to improve accuracy and reduce overfitting, effectively handling classification tasks. AdaBoost is an ensemble algorithm that sequentially trains weak classifiers, adjusting for errors by reweighting instances to improve accuracy and combining their weighted predictions for robust classification. Artificial neural networks are brain-inspired computational models with interconnected layers that effectively learn complex patterns from large data sets, suitable for classification tasks but require significant computational power and resources.

To assess the effectiveness of our machine learning models, we used several key performance metrics such as area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F_1-score. We implemented these metrics using the *scikit-learn* Python library, specifically using functions, such as roc_auc_score, accuracy_score, recall_score, precision_score, and f1_score from the sklearn.metrics module. This approach was crucial for estimating general accuracy and evaluating how effectively the models identified true positives and negatives, an essential factor in the clinical context of predicting breast cancer recurrence.

Due to a significant disparity in the number of cases between groups with and without recurrence, we used the synthetic minority oversampling technique (SMOTE) to balance the class distribution. This strategy enhanced the overall accuracy and generalizability of our predictive models. By mitigating bias toward the larger group of patients without recurrence, we were able to ensure with this proactive approach that our models could learn more effectively from a more equitably distributed data set. After correcting class imbalances with SMOTE and selecting appropriate algorithms, we conducted extensive hyperparameter tuning. This process involved systematically testing various combinations of hyperparameter values on local hospital computers to identify the optimal settings for each model. These settings are shown in Multimedia Appendix 1.

AI-CDSS Web Interface Design

After identifying optimal parameters for various advanced machine learning algorithms, we saved the best-performing models using the *joblib* library. This enabled seamless integration into a web-based interface, enhancing patient care and supporting clinicians in making informed decisions. With assistance from ChatGPT, we developed a web page built with HTML, CSS, and JavaScript, featuring dynamic forms where clinicians could input patient data and receive immediate predictions on breast cancer recurrence. The backend, powered by a Flask server, uses these serialized models to compute predictions based on user input. This setup not only ensures that the interface is a practical tool for clinical decision support but also aligns with the transformative goals of the AI-CDSS tool in managing breast cancer.

Furthermore, we provide ongoing support through webinars and one-on-one training sessions to accommodate different learning preferences, ensuring users feel confident using the AI-CDSS tool in their clinical practice. We collect regular feedback from users through surveys and interviews to continuously improve the tool and address emerging needs. By offering extensive training and support, we aim to maximize the tool's usability and integration into everyday clinical workflows, ultimately enhancing patient care.

Results

Characteristics of the Study Population

The data set, derived from the Tri-Service General Hospital breast cancer registry, encompassed 3577 patients treated between 2004 and 2016. Demographic and clinical variables such as hormone receptor status, tumor characteristics, and treatment modalities were meticulously collected. Most patients were older than 50 years of age, and the majority were ER and PR positives. A significant proportion of patients underwent conservative surgery, and adjuvant therapies such as chemotherapy, antiestrogen therapy, and radiotherapy were commonly administered (Table 1).

XSL•FO
RenderX

Table 1. Patient demographics, tumor characteristics, and treatment details.

	Overall (N=3577)	Training (n=2503)	Validation (n=1074)
Recurrence within 5 years, n (%)			
No	3152 (88.1)	2208 (88.2)	944 (87.9)
Yes	425 (11.9)	295 (11.8)	130 (12.1)
Age (years), mean (SD)	52.97 (11.29)	52.99 (11.18)	52.92 (11.55)
Estrogen receptor status, n (%)			
Negative	799 (22.3)	569 (22.7)	230 (21.4)
Positive	2778 (77.7)	1934 (77.3)	844 (78.6)
Progesterone receptor status, n (%)			
Negative	658 (18.4)	470 (18.8)	188 (17.5)
Positive	2919 (81.6)	2033 (81.2)	886 (82.5)
Estrogen receptor intensity, n (%)			
0	799 (22.3)	569 (22.7)	230 (21.4)
1	216 (6.0)	140 (5.6)	76 (7.1)
2	1789 (50.01)	1263 (50.5)	526 (48.98)
3	773 (21.6)	531 (21.2)	242 (22.5)
Progesterone receptor intensity, n (%)			
0	656 (18.3)	468 (18.7)	188 (17.5)
1	492 (13.8)	341 (13.6)	151 (14.1)
2	1887 (52.8)	1327 (53.01)	560 (52.1)
3	542 (15.2)	367 (14.7)	175 (16.3)
Pathological stage, n (%)			
Stage 0	997 (27.9)	707 (28.2%)	290 (27.0)
Stage 1	1097 (30.7)	750 (30.0)	347 (32.3)
Stage 2	1036 (28.96)	731 (29.2%)	305 (28.4)
Stage 3	395 (11.04)	278 (11.1)	117 (10.9)
Stage 4	52 (1.5)	37 (1.5)	15 (1.4)
Tumor size, n (%)			
T0	1003 (28.04)	711 (28.4)	292 (27.2)
T1	1432 (40.03)	985 (39.4)	447 (41.6)
T2	964 (26.9)	683 (27.3)	281 (26.2)
T3	122 (3.4)	86 (3.4)	36 (3.4)
T4	56 (1.6)	38 (1.5)	18 (1.7)
Nodal involvement, n (%)			
N0	2646 (73.97)	1834 (73.3)	812 (75.6)
N1	574 (16.04)	419 (16.7)	155 (14.4)
N2	231 (6.5)	162 (6.5)	69 (6.4)
N3	126 (3.5)	88 (3.5)	38 (3.5)
Metastasis, n (%)			
M0	3541 (98.99)	2475 (98.9)	1066 (99.3)
M1	36 (1.01)	28 (1.1)	8 (0.7)
Treatment details			

XSL•FO
RenderX

	Overall (N=3577)	Training (n=2503)	Validation (n=1074)
Surgery type			
Breast-conserving surgery	1363 (38.1)	966 (38.6)	397 (36.96)
Modified radical mastectomy	2214 (61.9)	1537 (61.4)	677 (63.03)
Antiestrogen therapy			
No	1382 (38.6)	978 (39.1)	404 (37.6)
Yes	2195 (61.4)	1525 (60.9)	670 (62.4)
Aromatase inhibitor usage			
No	2314 (64.7)	1608 (64.2)	706 (65.7)
Yes	1263 (35.3)	895 (35.8)	368 (34.3)
Chemotherapy			
No	897 (25.1)	607 (24.3)	290 (27)
Yes	2680 (74.9)	1896 (75.7)	784 (73.0)
Radiotherapy			
No	1910 (53.4)	1326 (52.98)	584 (54.4)
Yes	1667 (46.6)	1177 (47.02)	490 (45.6)

Model Performance Metrics, Validation Results, and ROC Curve Analysis

To ensure robustness, these models were validated using a separate validation data set comprising 1074 patients, with data from 2013 to 2016. After evaluating the performance of machine learning models using a suite of metrics, the LGBM model stood out, with an AUC of 0.80, demonstrating its robust predictive power in identifying breast cancer recurrence. The GB model followed closely with an AUC of 0.78, and the XGB model matched this performance, both reflecting a strong predictive capability. These models were further assessed for accuracy, sensitivity, specificity, PPV, NPV, and F_1-score. Notably, LGBM achieved an accuracy of 0.88, sensitivity of 0.33, specificity of 0.96, PPV of 0.51, NPV of 0.91, and an F_1-score of 0.40, underscoring a well-rounded performance.

The diagnostic proficiency of the models was also analyzed using the ROC curve analysis. This method provides a visual comparison of each model's true positive rate and false positive rate (Figure 3). The AUC obtained from the ROC curves reinforced the quantitative findings, illustrating that the LGBM, GB, and XGB models possess superior capabilities for discriminating between patients with and without breast cancer recurrence. The results from the ROC analysis and model metrics are shown in Table 2.

XSL•FO
RenderX

Figure 3. ROC curves of the evaluated machine learning models. This figure shows the ROC curves for the various machine learning models assessed in our study, including the LGBM, gradient boosting, XGB, random forest, AdaBoost, and ANN. These curves graphically represent the true positive rate versus the false positive rate for each model, elucidating their efficacy in differentiating recurrence from non-recurrence cases. The area under the curve values were also provided to gauge the predictive accuracy of each model. ANN: artificial neural networks; LGBM: light gradient boosting machine; ROC: receiver operating characteristic; XGB: extreme gradient boosting.

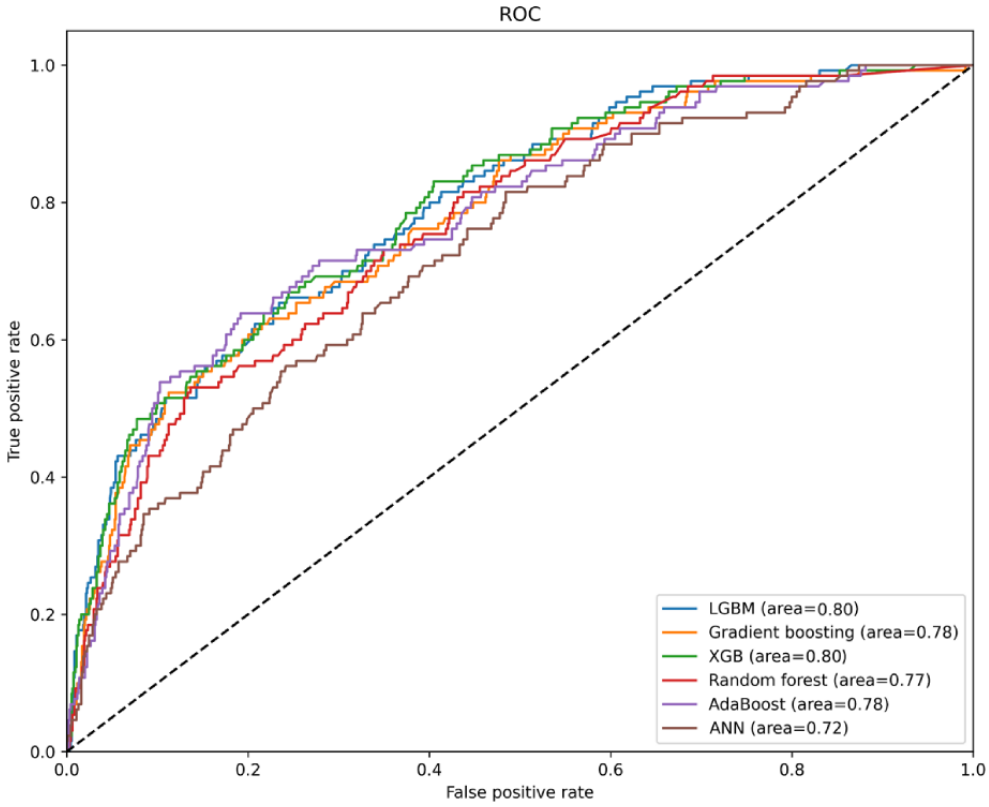

Table 2. Overview of machine learning models' performance.

Model	Training AUC[a]	Testing AUC	Accuracy	Sensitivity	Specificity	PPV[b]	NPV[c]	F_1-score
LGBM[d]	0.98	0.80	0.88	0.33	0.96	0.51	0.91	0.40
GB[e]	0.98	0.78	0.88	0.35	0.95	0.49	0.91	0.41
XGB[f]	0.98	0.80	0.88	0.39	0.94	0.49	0.92	0.44
RF[g]	0.99	0.77	0.87	0.24	0.95	0.42	0.90	0.30
AdaBoost	0.97	0.78	0.87	0.25	0.96	0.46	0.90	0.32
ANN[h]	0.99	0.72	0.84	0.22	0.93	0.30	0.90	0.25

[a]AUC: area under the curve.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]LGMB: light gradient boosting machine.
[e]GB: gradient boosting.
[f]XGB: extreme gradient boosting.
[g]RF: random forest.
[h]ANN: artificial neural network.

User Interface Effectiveness and Case Studies

The web interface of the AI-CDSS tool, depicted in Figure 4, was assessed for its efficacy in clinical decision-making. It offers a dynamic platform for clinicians and patients to enter critical patient data, such as age, hormone receptor status, and tumor characteristics, and explore various treatment options, including surgery, antiestrogen therapy, and radiotherapy. This web-based system enables the calculation of 5-year recurrence risks for different treatment choices, facilitating more informed and personalized decision-making. Its practicality in real-world scenarios has been underscored by case studies demonstrating its significant role in enhancing patient involvement and aiding clinicians in selecting appropriate treatment strategies.

XSL•FO
RenderX

Figure 4. User interface of the AI-CDSS tool. This figure displays the web-based user interface of the AI-CDSS for breast cancer recurrence, showcasing the input fields for patient information, including age, hormone receptor percentages and intensities, pathological stage, tumor size, node status, and metastasis. It also details the treatment choices such as surgery type, antiestrogen therapy, aromatase inhibitor usage, chemotherapy, and radiotherapy. The interface features a "Predict Recurrence" button that, when clicked, calculates and displays the 5-year recurrence probability, demonstrating the tool's capability for real-time, personalized recurrence risk assessment. AI-CDSS: artificial intelligence–based clinical decision support system.

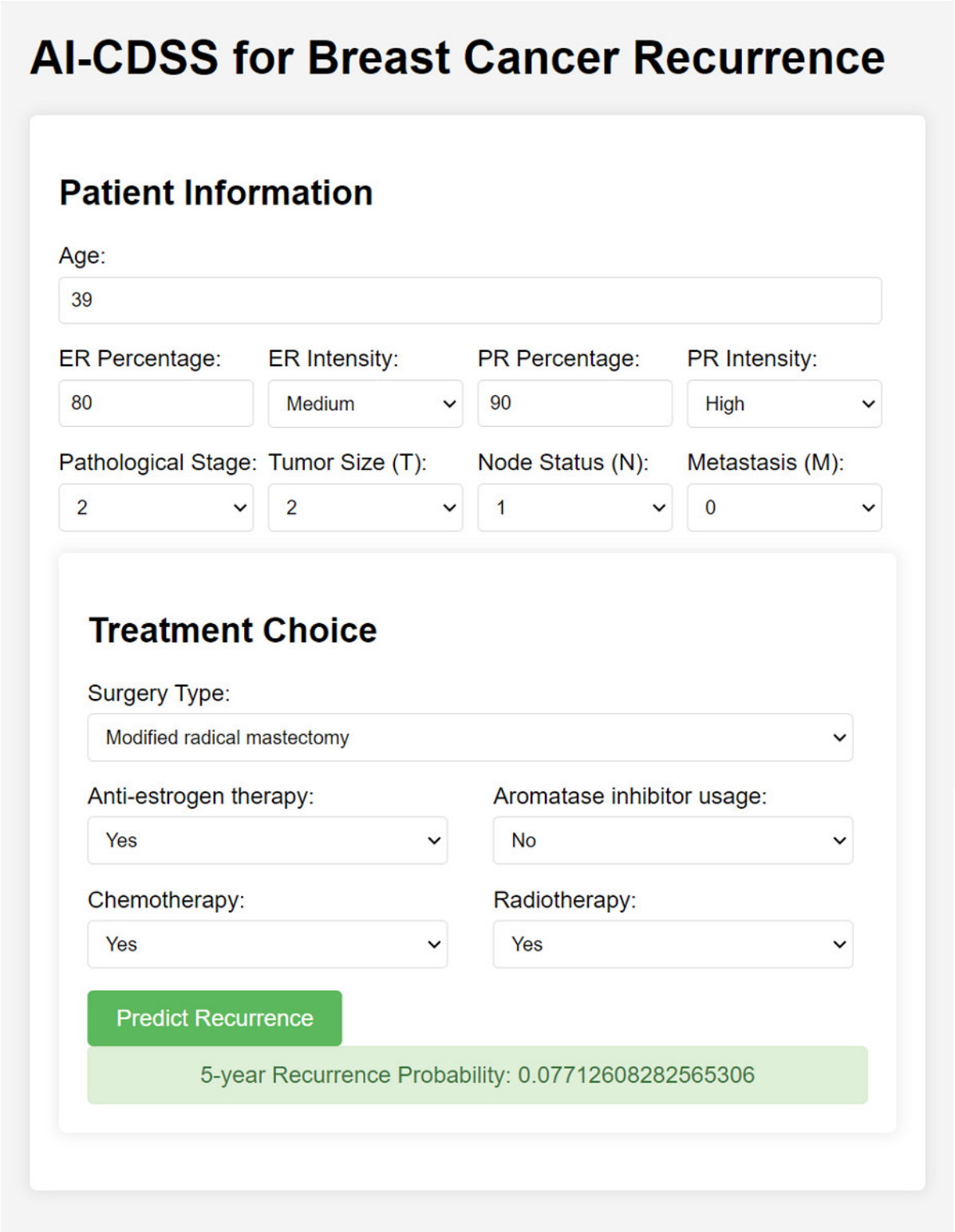

To evaluate the integration of the AI-CDSS tool into clinical workflows, we use surveys and interviews to comprehensively assess the tool from multiple dimensions. The survey assesses usability aspects such as ease of use, navigation, and accessibility throughout patient care stages. A key focus is on the tool's impact on decision-making, including how it influences treatment planning, personalizes options, and potentially improves outcomes. The survey also covers training adequacy, overall satisfaction, and the likelihood of recommendation. Open-ended questions are included to gather feedback on improvements and observations for a holistic perspective. The interviews align with the survey concepts,

XSL•FO
RenderX

offering qualitative insights that complement the structured data. This approach ensures a comprehensive assessment of usability, integration, effectiveness, decision support capabilities, training quality, and user satisfaction from the users' perspectives.

Discussion

Principal Results

The research marked a significant advancement in breast cancer management with the development of an AI-CDSS leveraging the capabilities of ChatGPT. This study used a comprehensive data set of 3577 patients from Tri-Service General Hospital, Taiwan. Key achievements include the meticulous categorization of hormone receptor status and age, which significantly enhances the precision of the data set for machine learning applications. Among the various models tested, the LGBM model showed superior performance, with an AUC of 0.80, a finding that was consistently supported by the ROC curve analysis. These results underscore the proficiency of the AI-CDSS tool for predicting breast cancer recurrence, thereby establishing a new benchmark in the field.

Interpretations

The reliability of the LGBM model in predicting breast cancer recurrence reflects a broader trend in personalized medicine [14,15]. Compared to traditional models, our AI-CDSS tool provides a more tailored prediction, aligning with the current shift in oncological research toward individualized patient care [14,16]. The refined categorization of hormone receptor status and age groups in our model concurred with the findings of other studies, emphasizing the heterogeneity of breast cancer [14-16]. However, our model distinguishes itself by leveraging AI to address these complexities more effectively.

Optimizing Model Selection for Clinical Contexts

In evaluating the optimal model to predict recurrence, each exhibited unique strengths and limitations suitable for specific clinical contexts. LGBM, GB, and XGB all showed a high accuracy of 0.88. LGBM excelled in rapid computation and achieved a highest AUC of 0.80, indicating a good balance between sensitivity (0.33) and specificity (0.96). GB offered high accuracy (0.88) and slightly better sensitivity (0.35), but its high computational demands limit its real-time practicality. XGB matched LGBM in AUC (0.80) and demonstrated the highest sensitivity (0.39) but required significantly more computational resources. Therefore, LGBM is preferable for rapid analyses prioritizing speed and balanced efficiency, while GB and XGB are ideal for resource-rich settings [12]. Selecting the appropriate model for each clinical context ensures optimal performance and reliability aligned with specific needs and resource constraints. If the clinical condition prioritizes early detection and aggressive treatment of recurrence, such as in more advanced stages of cancer, the sensitivity might be prioritized, making a model such as XGB preferred despite its higher computational requirements. However, if avoiding overtreatment and reducing patient anxiety are paramount, such as in early-stage cancer cases, which constituted 87.6% (N=3130) of the cases in our study (stages 0-2), specificity might be prioritized, making LGBM (with a specificity of 0.96) a better choice.

Implications

Current trends in breast cancer research value precision and accessibility [14,17]. Unlike existing models that primarily emphasize scientific robustness, our AI-CDSS tool uniquely bridges the gap between high-level scientific accuracy and user-friendliness [17]. This balance is pivotal in the current era of breast cancer management, which is witnessing a shift toward greater patient involvement and clinician accessibility to advanced diagnostic tools. By offering a tool that is both technically sound and easily navigable, our study responds to the growing demand for personalized care and empowers health care providers with practical solutions [16,18]. Thus, the AI-CDSS tool stands out in the landscape of breast cancer predictive models, offering a nuanced and patient-centric approach that aligns with the evolving needs and expectations of patients with cancer and medical practitioners [18].

Limitations

Although our findings were promising, this study had some limitations. Despite a robust initial data set of 4503 patients who had undergone surgical treatment at our main medical center, the study's retrospective design and focus on data primarily from a single center pose inherent limitations. Although this center is supported by branches that refer to complex surgical cases, including breast cancer, the final cohort used for analysis was relatively small (N=3577), and the exclusion of certain patient groups (such as those with HER2 overexpression) may have impacted the generalizability of our results. To ensure fairness and transparency, we used SMOTE to balance the data set and selected machine learning algorithms known for their accuracy and interpretability. All data processing was conducted on secure local systems to protect patient privacy, with the AI-CDSS tool designed to assist, not replace, clinical judgment. Notably, 87.6% of our cases were early-stage cancer (stages 0-2), making the LGBM model, with a specificity of 0.96, a suitable core model to avoid overtreatment and reduce patient anxiety. However, varying patient demographics across different institutions may necessitate the adoption of different models. Continuous surveys and interviews are essential to comprehensively evaluate user needs. Future studies should include a more diverse patient cohort and consider real-time data to enhance the applicability and robustness of this model.

Recommendations

We recommend a broader validation of the AI-CDSS tool across diverse clinical environments and patient populations to ascertain its efficacy and adaptability. Enriching the data set with a wider array of breast cancer cases would further improve the predictive accuracy of the model. Integrating patient feedback into the AI-CDSS interface could also enhance user-friendliness, potentially increasing its adoption and impact in clinical settings. Such efforts will not only validate our current findings but also pave the way for more advanced patient-centric cancer care solutions.

XSL•FO
RenderX

Conclusions

In this research, we developed a web-based AI-CDSS application, enhanced by ChatGPT's guidance, for predicting breast cancer recurrence. Using data from 3577 patients with breast cancer at Tri-Service General Hospital (2004-2016), the study applied advanced machine learning algorithms, including LGBM, GB, and XGB. ChatGPT significantly aided in data preprocessing, such as hormone receptor categorization, age binning, and one-hot encoding, and tackled data imbalance with SMOTE techniques. The models' performances were evaluated using AUC, accuracy, sensitivity, and F_1-score, with LGBM achieving the highest AUC of 0.80. The AI-CDSS tool's web interface proved effective in clinical decision-making, enhancing personalized treatment plans and patient engagement. This study underscores AI-CDSS's role, especially with ChatGPT's contribution, in advancing personalized medicine and integrating health care technology, offering a more tailored approach to breast cancer recurrence prediction.

Acknowledgments

Financial support for this research was provided by the Tri-Service General Hospital in Taipei, Taiwan (grants TSGH-D-113105 and TSGH-D-113106). The sponsors did not participate in the design of the study, the gathering and analysis of data, or the determination to publish this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Parameters of machine learning models for predicting 5-year breast cancer recurrence.
[DOCX File , 16 KB - jmir_v26i1e56022_app1.docx]

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022 Jan;72(1):7-33 [FREE Full text] [doi: 10.3322/caac.21708] [Medline: 35020204]
3. Sledge GW, Mamounas EP, Hortobagyi GN, Burstein HJ, Goodwin PJ, Wolff AC. Past, present, and future challenges in breast cancer treatment. J Clin Oncol 2014;32(19):1979-1986 [FREE Full text] [doi: 10.1200/JCO.2014.55.4139] [Medline: 24888802]
4. Lu H, Xie J, Gerido LH, Cheng Y, Chen Y, Sun L. Information needs of breast cancer patients: theory-generating meta-synthesis. J Med Internet Res 2020;22(7):e17907 [FREE Full text] [doi: 10.2196/17907] [Medline: 32720899]
5. Clift AK, Collins GS, Lord S, Petrou S, Dodwell D, Brady M, et al. Predicting 10-year breast cancer mortality risk in the general female population in England: a model development and validation study. Lancet Digit Health 2023;5(9):e571-e581 [FREE Full text] [doi: 10.1016/S2589-7500(23)00113-9] [Medline: 37625895]
6. Yaqoob A, Musheer Aziz R, verma NK. Applications and techniques of machine learning in cancer classification: a systematic review. Hum-Cent Intell Syst 2023;3(4):588-615. [doi: 10.1007/s44230-023-00041-3]
7. Kim MJ, Kim HJ, Kang D, Ahn HK, Shin S, Park S, et al. Preliminary attainability assessment of real-world data for answering major clinical research questions in breast cancer brain metastasis: framework development and validation study. J Med Internet Res 2023;25:e43359 [FREE Full text] [doi: 10.2196/43359] [Medline: 36951923]
8. Gradishar WJ, Moran MS, Abraham J, Abramson V, Aft R, Agnese D, et al. NCCN Guidelines® insights: breast cancer, version 4.2023. J Natl Compr Canc Netw 2023;21(6):594-608. [doi: 10.6004/jnccn.2023.0031] [Medline: 37308117]
9. Tang YM, Chau KY, Kwok APK, Zhu T, Ma X. A systematic review of immersive technology applications for medical practice and education—trends, application areas, recipients, teaching contents, evaluation methods, and performance. Educ Res Rev 2022;35:100429. [doi: 10.1016/j.edurev.2021.100429]
10. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Evaluations 2023;3(1):100105. [doi: 10.1016/j.tbench.2023.100105]
11. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
12. Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Expert Syst Appl 2024;244:122778. [doi: 10.1016/j.eswa.2023.122778]
13. Ding S, Li H, Su C, Yu J, Jin F. Evolutionary artificial neural networks: a review. Artif Intell Rev 2011;39(3):251-260. [doi: 10.1007/s10462-011-9270-6]

XSL•FO
RenderX

14. Ahn JS, Shin S, Yang S, Park EK, Kim KH, Cho SI, et al. Artificial intelligence in breast cancer diagnosis and personalized medicine. J Breast Cancer 2023;26(5):405-435 [FREE Full text] [doi: 10.4048/jbc.2023.26.e45] [Medline: 37926067]
15. Zuo D, Yang L, Jin Y, Qi H, Liu Y, Ren L. Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Med Inform Decis Mak 2023;23(1):276 [FREE Full text] [doi: 10.1186/s12911-023-02377-z] [Medline: 38031071]
16. Rezayi S, R Niakan Kalhori S, Saeedi S. Effectiveness of artificial intelligence for personalized medicine in neoplasms: a systematic review. Biomed Res Int 2022;2022:7842566 [FREE Full text] [doi: 10.1155/2022/7842566] [Medline: 35434134]
17. Zhang J, Wu J, Zhou XS, Shi F, Shen D. Recent advancements in artificial intelligence for breast cancer: image augmentation, segmentation, diagnosis, and prognosis approaches. Semin Cancer Biol 2023;96:11-25. [doi: 10.1016/j.semcancer.2023.09.001] [Medline: 37704183]
18. Xu F, Sepúlveda MJ, Jiang Z, Wang H, Li J, Liu Z, et al. Effect of an artificial intelligence clinical decision support system on treatment decisions for complex breast cancer. JCO Clin Cancer Inform 2020;4:824-838 [FREE Full text] [doi: 10.1200/CCI.20.00018] [Medline: 32970484]

Abbreviations

AI: artificial intelligence
AI-CDSS: artificial intelligence–based clinical decision support system
AUC: area under the curve
ER: estrogen receptor
GB: gradient boosting
HER2: human epidermal growth factor receptor 2
LGBM: light gradient-boosting machine
NPV: negative predictive value
PPV: positive predictive value
PR: progesterone receptor
RF: random forest
ROC: receiver operating characteristic
SMOTE: synthetic minority oversampling technique
XGB: extreme gradient boosting

Edited by G Eysenbach; submitted 03.01.24; peer-reviewed by G Pau, A Abd El-Aziz; comments to author 10.05.24; revised version received 17.05.24; accepted 16.07.24; published 04.09.24.

Please cite as:
Lin TH, Chung HY, Jian MJ, Chang CK, Perng CL, Liao GS, Yu JC, Dai MS, Yu CP, Shang HS
An Advanced Machine Learning Model for a Web-Based Artificial Intelligence–Based Clinical Decision Support System Application: Model Development and Validation Study
J Med Internet Res 2024;26:e56022
URL: https://www.jmir.org/2024/1/e56022
doi:10.2196/56022
PMID:

©Tai-Han Lin, Hsing-Yi Chung, Ming-Jr Jian, Chih-Kai Chang, Cherng-Lih Perng, Guo-Shiou Liao, Jyh-Cherng Yu, Ming-Shen Dai, Cheng-Ping Yu, Hung-Sheng Shang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using AI to Differentiate Mpox From Common Skin Lesions in a Sexual Health Clinic: Algorithm Development and Validation Study

Nyi Nyi Soe[1,2], MBBS, MPH; Zhen Yu[1,3], PhD; Phyu Mon Latt[1,2], MBBS, MPH; David Lee[4], BAppSc, GradDip (CCRN), MPH, DrPH; Ranjit Singh Samra[5], MBChB, MSc, MPH; Zongyuan Ge[3], PhD; Rashidur Rahman[4], MSc; Jiajun Sun[1,2], BE, MS; Jason J Ong[2,4], MBBS, MMed, PhD; Christopher K Fairley[2,4], MBBS, PhD; Lei Zhang[1,2,6], PhD

[1]Artificial Intelligence and Modelling in Epidemiology Program, Melbourne Sexual Health Centre, Alfred Health, Melbourne, Australia
[2]School of Translational Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia
[3]Augmented Intelligence and Multimodal analytics (AIM) for Health Lab, Faculty of Information Technology, Monash University, Melbourne, Australia
[4]Melbourne Sexual Health Centre, Alfred Health, Melbourne, Australia
[5]Department of Infectious Diseases, Alfred Hospital, Alfred Health, Melbourne, Australia
[6]Clinical Medical Research Center, Children's Hospital of Nanjing Medical University, Nanjing, China

Corresponding Author:
Lei Zhang, PhD
Artificial Intelligence and Modelling in Epidemiology Program
Melbourne Sexual Health Centre
Alfred Health
580 Swanston Street, Carlton
Melbourne
Australia
Phone: 61 0433452013
Email: Lei.Zhang1@monash.edu

Abstract

Background: The 2022 global outbreak of mpox has significantly impacted health facilities, and necessitated additional infection prevention and control measures and alterations to clinic processes. Early identification of suspected mpox cases will assist in mitigating these impacts.

Objective: We aimed to develop and evaluate an artificial intelligence (AI)–based tool to differentiate mpox lesion images from other skin lesions seen in a sexual health clinic.

Methods: We used a data set with 2200 images, that included mpox and non-mpox lesions images, collected from Melbourne Sexual Health Centre and web resources. We adopted deep learning approaches which involved 6 different deep learning architectures to train our AI models. We subsequently evaluated the performance of each model using a hold-out data set and an external validation data set to determine the optimal model for differentiating between mpox and non-mpox lesions.

Results: The DenseNet-121 model outperformed other models with an overall area under the receiver operating characteristic curve (AUC) of 0.928, an accuracy of 0.848, a precision of 0.942, a recall of 0.742, and an F_1-score of 0.834. Implementation of a region of interest approach significantly improved the performance of all models, with the AUC for the DenseNet-121 model increasing to 0.982. This approach resulted in an increase in the correct classification of mpox images from 79% (55/70) to 94% (66/70). The effectiveness of this approach was further validated by a visual analysis with gradient-weighted class activation mapping, demonstrating a reduction in false detection within the background of lesion images. On the external validation data set, ResNet-18 and DenseNet-121 achieved the highest performance. ResNet-18 achieved an AUC of 0.990 and an accuracy of 0.947, and DenseNet-121 achieved an AUC of 0.982 and an accuracy of 0.926.

Conclusions: Our study demonstrated it was possible to use an AI-based image recognition algorithm to accurately differentiate between mpox and common skin lesions. Our findings provide a foundation for future investigations aimed at refining the algorithm and establishing the place of such technology in a sexual health clinic.

(J Med Internet Res 2024;26:e52490) doi:10.2196/52490

XSL•FO
RenderX

KEYWORDS

mpox; sexually transmitted infections; artificial intelligence; deep learning; skin lesion

Introduction

Mpox is a zoonotic viral infection caused by the monkeypox virus [1]. The infection originated from remote areas of Central and West Africa and was first reported in 1970 in the Democratic Republic of Congo. The disease typically presents with initial symptoms including fever, headache, rash, and lymph node enlargement [1,2]. It can also present as lesions, sores, or ulcers in the face, mouth, extremities, and anogenital areas [3]. Mpox shares many similarities with other pox-like infections, such as chickenpox, smallpox, and other bacterial and viral infections. Mpox is transmitted through close contact with infected individuals and the most recent cases have occurred predominantly in men who have sex with men, especially those with multiple and anonymous sexual partners. Although mpox is generally self-limiting, delayed diagnosis could lead to transmission within communities and increase the risk of complications in immunocompromised individuals, children, and pregnant women [4].

The 2022 global mpox outbreak significantly impacted health facilities because it required additional infection prevention and control measures including personal protective equipment to prevent the spread of infection to staff [5,6]. Moreover, suspected mpox cases had to be separated from main clinic patients to mitigate the risk of infection to others, and so the processes in the clinic were modified and slowed [7]. If it were possible to identify suspected mpox cases before their visit, it would benefit the workflow of clinics.

Recently, artificial intelligence (AI)–assisted medical diagnosis has emerged as a promising research area in the health care sector [8]. AI has been used in medical imaging to assist radiologists in identifying COVID-19–related abnormalities in chest x-rays and computed tomographic scans [9]. AI algorithms can diagnose skin lesions such as melanoma with high accuracy [10-12] and have seen increasing use in risk prediction [13,14] and early diagnosis of sexually transmitted infections (STIs) [15]. To our knowledge, many studies [16-20] have been conducted to examine the performance of tools using AI to diagnose mpox. These studies were designed to differentiate mpox from chickenpox, smallpox, cowpox, and measles. While these studies provided valuable insights, their application may be limited in today's clinical setting where cowpox is rare and smallpox has been eliminated. In the context of a sexual health clinic, a diverse range of skin lesions, including STIs such as genital herpes, genital warts, syphilis, and molluscum contagiosum. Nonsexually transmitted skin conditions such as dermatosis and other inflammatory skin diseases are also commonly seen.

We aimed to develop an AI-assisted diagnostic tool and evaluate its ability to distinguish mpox from other common skin lesions (STIs and non-STIs) in a sexual health clinic setting. If proven accurate, such a tool could offer a preliminary screening capability through a web-based platform, potentially aiding in the early detection of mpox cases and timely initiation of treatment. Furthermore, its integration into clinical workflows could streamline the triage process and enhance patient management in the clinic.

Methods

Data Collection

Melbourne Sexual Health Centre (MSHC) conducted this study using the checklist for Artificial Intelligence in Medical Imaging (CLAIM) [21]. Since 2010, we have collected clinical images of STIs and non-STI skin lesions. In 2022 we added mpox lesion images to our database. The MSHC data set included 86 mpox and 1565 non-mpox images. We also implemented web scraping techniques on Google and Bing (Microsoft) search engines with different terms ("monkeypox," "Monkey Pox," "MPX," and "Mpox") to collect public domain or Creative Commons mpox images. This enabled us to collect 271 mpox images which required clinicians' validation. Both MSHC and internet-sourced data sets contained uncropped original images. Moreover, we included 278 mpox images from the Kaggle data set [22] used in previous studies [16-18] for model evaluation. This data set consists of anonymized skin lesion images without any associated patient information, precluding the comparison of patient characteristics across different data sets. The images from the Kaggle data set had already undergone preprocessing and their lesion areas were cropped for evaluation.

Inclusion and Exclusion Criteria

The study included images collected from the MSHC attendees who provided written informed consent for the use of their image for research purposes and who were diagnosed with mpox, genital herpes, genital warts, primary or secondary syphilis, molluscum contagiosum, dermatosis, or healthy skin. The web-scraped mpox images were included if they were open source and could be verified by MSHC clinicians. Images not meeting these criteria were excluded.

Data Labeling

Two experienced sexual health physicians independently labeled images that did not have diagnoses and ensured they were correct by reviewing the clinical notes, and laboratory results from the Clinical Patient Management System. For the images from the Kaggle data sets, the diagnoses were also verified visually by the experts. Any discrepancies in image diagnosis between the 2 clinicians were reviewed by a third clinician. If consensus on the diagnosis could not be reached between the 3 clinicians, the image was discarded. Our study focused on distinguishing mpox from other common skin lesions, thus, labeled images were organized into mpox and non-mpox folders. Table 1 shows the number of images for each disease class.

XSL•FO
RenderX

Table 1. Distribution of images across diagnostic classes.

Class and diseases	# Original images, n (%)	# Augmented images, n	# Total images, n
Mpox			1537
Mpox	635 (28.86)	902	
Non-mpox			1565
Dermatosis	215 (9.77)	0	
Genital herpes	250 (11.36)	0	
Molluscum contagiosum	200 (9.09)	0	
Healthy skin	250 (11.36)	0	
Primary syphilis	200 (9.09)	0	
Secondary syphilis	250 (11.36)	0	
Genital warts	200 (9.09)	0	
Total	2200 (100.00)	902	3102

Data Processing

We manually checked all images to ensure they had no identifiable features (eg, face, tattoos, and birthmarks). We then removed 14 duplicated and low-resolution images from the whole data set by using the difPy Python package (Python Software Foundation) [23]. As part of the image preprocessing pipeline, all original images were resized to a fixed resolution of 320×320 pixels and converted to JPEG format to ensure uniformity and facilitate subsequent processing.

Data Partition

We divided the main data set into 3 subsets for training, testing, and external validation (see Figure 1). Except for those from the Kaggle data set, all mpox images were randomly divided into the "training and validation" data set (n=274, 80% of images) and the "testing" data set (n=70, 20% of images). The 277 mpox images from the Kaggle data set were reserved for "external validation." Similarly, non-mpox images were also randomly divided into "training and validation," "testing," and "external validation," with 1172, 76, and 317 images, respectively.

Figure 1. Workflow of the AI-based mpox classification model. AI: artificial intelligence; AUC: area under the receiver operating characteristic curve; MPX: mpox; MSHC: Melbourne Sexual Health Centre.

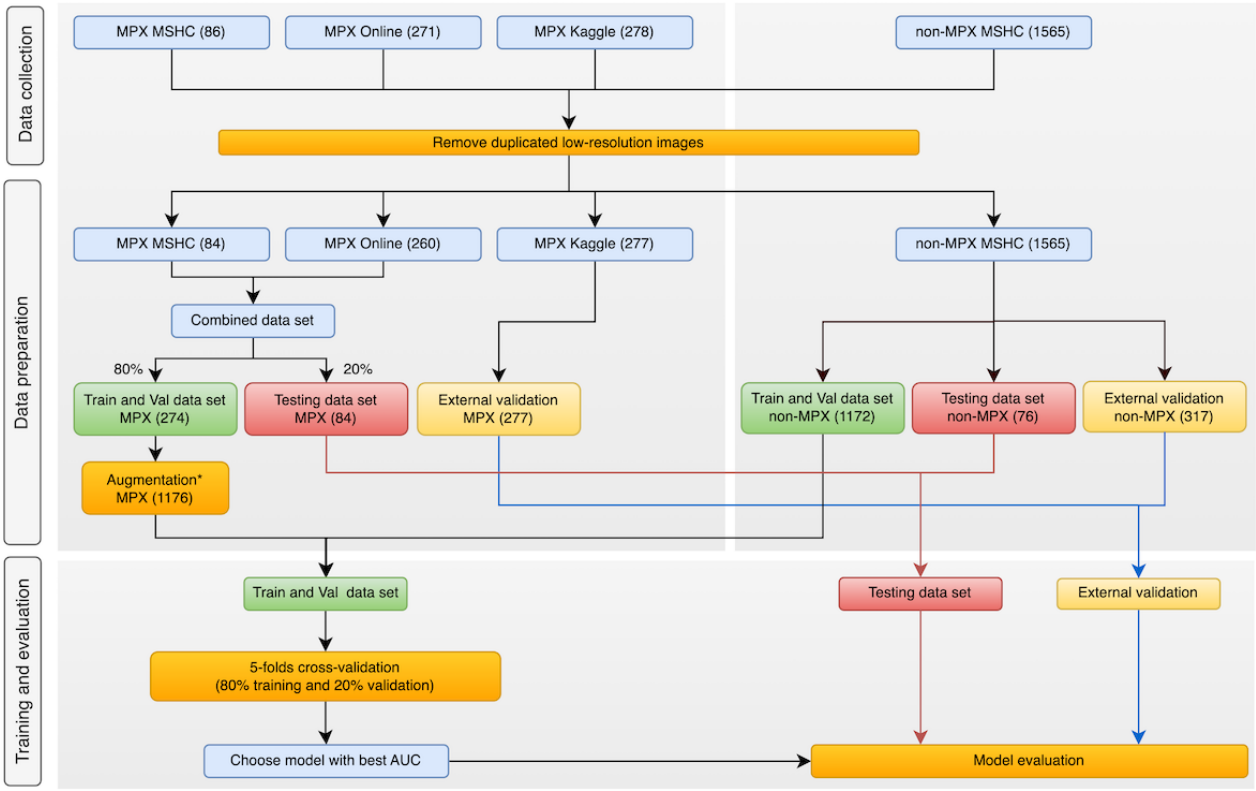

XSL•FO
RenderX

Image Augmentation

The training and validation data set was imbalanced, with 274 mpox images but 1172 non-mpox images. To address this imbalanced data set [24], we applied data augmentation to increase the number of images in the mpox data set [25]. This included rotating images, cropping areas, zooming in or out, flipping horizontally or vertically, and adjusting brightness and contrast (see Table S1 in Multimedia Appendix 1 for data augmentations). We generated 902 new images for mpox with augmentation, resulting in a final training and validation data set of 1176 mpox and 1172 non-mpox images. Testing and external validation data sets were not augmented and the final data sets contained 146 and 594 images, respectively (see Table 1 and Figure 1).

Model Training and Validation

We adopted a transfer learning approach to develop the binary image classification model. This approach involved using the pretrained model on a large data set and fine-tuning it on a smaller target data set, making it effective in training a deep model with fewer images [26,27]. During transfer learning, we froze the weights of all backbone layers and replaced the last fully connected layer with a new classification layer. We only trained it with random weights thereafter. We experimented with 6 pretrained deep neural network architectures with different-sized parameters—MobileNet-V2 [28], ShuffleNet-V2 [29], DenseNet-121 [30], ResNet-18 [31], ResNet-34 [31], and Swin-Transformer [32] (see Figure 2). The size of the parameters can be seen in Table S2 in Multimedia Appendix 1.

Figure 2. Overview of convolutional neural network (CNN) architecture.

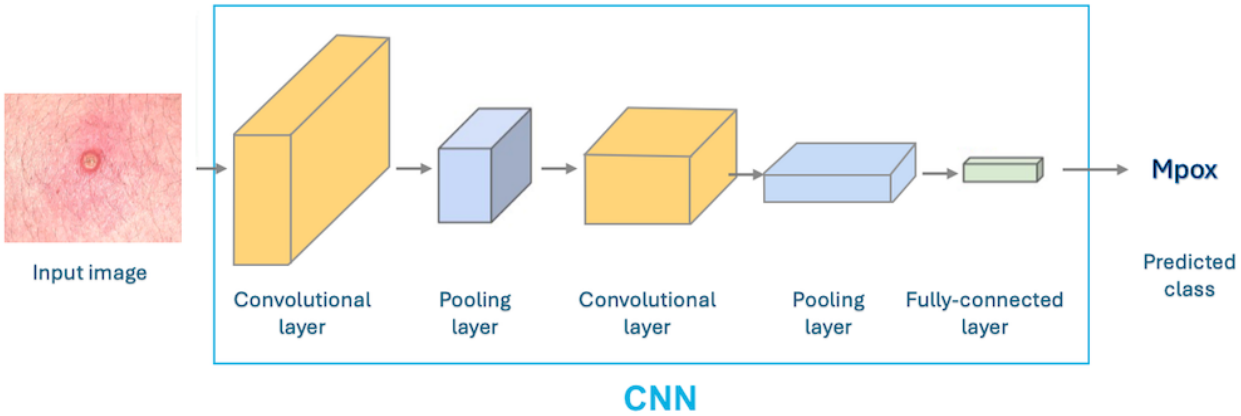

A 5-fold cross-validation was performed on the training and validation data sets, to enhance the model's robustness and generalizability. We trained and validated each pretrained model 5 times with different subsets, and the results were averaged for better performance evaluation. We trained an AI image classification model using a pretrained model backbone, Adam optimizer, cross-entropy loss function, a batch size of 72, a dropout rate of 0.2, an image size of 320×320 pixels, 150 epochs, and a learning rate of $3e^{-4}$. We implemented model training with PyTorch (Meta AI) on a Tesla T4 GPU machine. The model training time ranged from 3 to 4 hours for smaller models like MobileNet-V2 and ShuffleNet-V2, while larger models like Swin-Transformer took up to 5 to 6 hours.

Performance Evaluation Metrics

The model performance for distinguishing between mpox and non-mpox images was evaluated with various metrics including area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F_1-score. AUC measures the model's ability for binary classification, ranging from 0 to 1, where 1 indicates perfect classification [33-35]. The equation and definition of the performance metrics are shown in Table S3 in Multimedia Appendix 1. The scikit-learn Python package (version 1.2.0) was used to calculate all metrics [36].

Region of Interest Approach to Improve Model Performance

In this study, the original images from MSHC were used for the testing data set without any cropping of the lesion areas, whereas the mpox images from the Kaggle data set were precropped for the area of interest in the external validation data set. The performance of the model was compared between the testing and external validation data sets. Following this, a region of interest (ROI) approach was applied to the testing data set, whereby the lesion areas were cropped from the original images, and the model's performance was reevaluated. The efficacy of this approach was visualized using gradient-weighted class activation mapping (Grad-CAM), which enabled the assessment of the model's performance before and after the ROI approach.

Model External Validation

We first evaluated all 6 models with a testing data set that had never been used for model training and calculated the performance metrics. To further test the generalizability and robustness of those models, we evaluated their performance with an external validation data set containing unseen images from the Kaggle data set, which was commonly used in previous studies (Figure 1). Then, we computed the confusion matrices for the best models on testing and external validation data sets to analyze the model's correct and incorrect predictions. Finally, we conducted the visual analysis on a set of sample prediction outputs from the best model using Grad-CAM [33]. We generated the image with a heatmap where the most important

XSL•FO
RenderX

region for the predicted class was highlighted. In other words, the features of the images used for prediction were highlighted with color gradients from the most important to the least important areas, helping to understand the model's behavior and to identify the error for wrong predictions.

Ethical Considerations

The study protocol received ethics approval from the Alfred Hospital Ethics Committee (683/22). All research activities were carried out in compliance with relevant ethical principles and guidelines. To protect participant privacy, all data underwent a deidentification process before being used for model development and analysis.

Results

Overview

Table 1 shows the number of images with mpox and non-mpox conditions in the data set. The mpox images accounted for 29% (635/2200) of the original images while other non-mpox images included conditions that individually accounted for between 9% (200/2200) and 11% (250/2200) of the data set. Figure 1 shows the processes for how the final data sets were created and the source of images for the final data sets.

Table 2 shows key performance metrics for the 6 different models, including AUC, accuracy, precision, recall, and F_1-score. These metrics are presented for the testing data set both before cropping and after applying the ROI approach, as well as for the external validation data set. Additionally, Figure 3 shows the visual comparison of the AUC for 6 different models across different data sets.

Table 2. Model evaluation: performance metrics across 5 folds.

Model	AUC[a], mean (SD)	Accuracy, mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)
Model evaluation on testing data set (before cropping)					
MobileNet-V2	0.918 (0.011)	0.832 (0.022)	0.942 (0.025)	0.686 (0.054)	0.792 (0.033)
ShuffleNet-V2	0.908 (0.004)	0.820 (0.014)	0.964 (0.022)	0.618 (0.049)	0.752 (0.035)
DenseNet-121	0.928 (0.022)	0.848 (0.041)	0.942 (0.013)	0.742 (0.024)	0.834 (0.018)
ResNet-18	0.916 (0.005)	0.828 (0.008)	0.934 (0.026)	0.688 (0.016)	0.79 (0.010)
ResNet-34	0.876 (0.043)	0.798 (0.015)	0.900 (0.000)	0.618 (0.011)	0.734 (0.005)
Swin-Transformer	0.928 (0.016)	0.798 (0.008)	0.926 (0.019)	0.622 (0.022)	0.746 (0.011)
Model evaluation on testing data set (with region of interest approach)					
MobileNet-V2	0.963 (0.004)	0.907 (0.006)	0.949 (0.006)	0.851 (0.016)	0.898 (0.008)
ShuffleNet-V2	0.964 (0.002)	0.871 (0.003)	0.924 (0.008)	0.797 (0.012)	0.856 (0.005)
DenseNet-121	0.982 (0.002)	0.951 (0.003)	0.956 (0.000)	0.940 (0.006)	0.948 (0.003)
ResNet-18	0.963 (0.001)	0.901 (0.008)	0.922 (0.018)	0.869 (0.012)	0.894 (0.008)
ResNet-34	0.959 (0.004)	0.882 (0.009)	0.918 (0.007)	0.829 (0.017)	0.871 (0.011)
Swin-Transformer	0.967 (0.011)	0.862 (0.013)	0.931 (0.002)	0.769 (0.027)	0.842 (0.018)
Model evaluation on external validation data set					
MobileNet-V2	0.985 (0.002)	0.937 (0.007)	0.911 (0.019)	0.959 (0.017)	0.934 (0.007)
ShuffleNet-V2	0.963 (0.040)	0.923 (0.005)	0.925 (0.009)	0.910 (0.003)	0.917 (0.005)
DenseNet-121	0.982 (0.001)	0.926 (0.003)	0.906 (0.006)	0.939 (0.006)	0.922 (0.003)
ResNet-18	0.990 (0.001)	0.947 (0.006)	0.934 (0.009)	0.953 (0.009)	0.943 (0.006)
ResNet-34	0.974 (0.002)	0.908 (0.008)	0.887 (0.013)	0.920 (0.006)	0.903 (0.008)
Swin-Transformer	0.979 (0.003)	0.912 (0.007)	0.944 (0.005)	0.862 (0.012)	0.901 (0.009)

[a]AUC: area under the receiver operating characteristic curve.

XSL•FO
RenderX

Figure 3. AUC comparison of 6 different models on testing and external validation data set. AUC: Area Under the Receiver Operating Characteristic Curve; ROC: Receiver Operating Characteristics Curve.

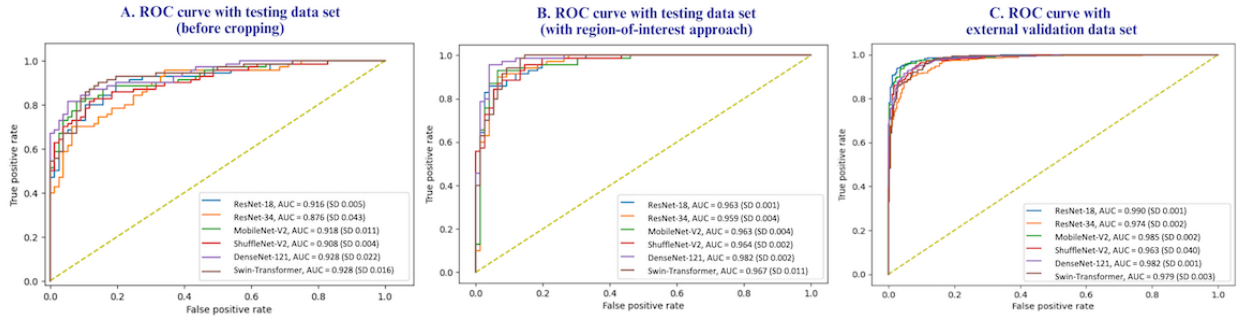

Evaluation of Model Performance on the Testing Data Set (Before Cropping)

Evaluation of the performance of 6 different models with the testing data set (before cropping) showed that the DenseNet-121 outperformed the other models in terms of overall AUC (0.928, SD 0.022), accuracy (0.848, SD 0.041), precision (0.942, SD 0.013), recall (0.742, SD 0.024), and F_1-score (0.834, SD 0.018; Table 2).

Evaluation of the Model With the ROI Approach on the Testing Data Set

We applied the ROI approach to the testing data set and conducted a reevaluation of the models (Table 2). This approach significantly improved the performance of all 6 models for image classification. The AUC score of MobileNet-V2 increased from a mean of 0.918 (SD 0.011) to a mean of 0.963 (SD 0.004), ShuffleNet-V2 from a mean of 0.908 (SD 0.004) to a mean of 0.964 (SD 0.002), ResNet-18 from a mean of 0.916 (SD 0.005) to a mean of 0.963 (SD 0.001), ResNet-34 from a mean of 0.876 (SD 0.043) to a mean of 0.959 (SD 0.004), and Swin-Transformer from a mean of 0.928 (SD 0.016) to a mean of 0.967 (SD 0.011). The AUC score of the best-performing DenseNet-121 model increased from a mean of 0.928 (SD 0.022) to mean of 0.982 (SD 0.002) on the testing data set following the adoption of the ROI approach.

Evaluation of Model Performance on the External Validation Data Set

We evaluated the performance of these models with the external validation data set. The findings in Table 2 showed that the ResNet-18 performed best on this data set, achieving an AUC of mean 0.990 (SD 0.001), an accuracy of 0.947 (SD 0.006), a precision of 0.934 (SD 0.009), a recall of 0.953 (SD 0.009), and an F-score of 0.943 (SD 0.006). DenseNet-121 also achieved high performance with an AUC of 0.982 (SD 0.001) on the external validation data set.

Confusion Matrix Results for DenseNet-121 and ResNet-18 Models

Figure 4 shows the confusion matrixes for 5 folds for the DenseNet-121 and ResNet-18 models. In the testing data set (before cropping), the confusion matrix of the best fold in DenseNet-121 showed that 55 out of 70 (79%) mpox and 74 out of 77 (96%) non-mpox images were correctly classified (Figure 4A). After applying the ROI approach, the confusion matrix of the best fold in DenseNet-121 showed that 66 out of 70 (94%) mpox images were correctly classified which represents a notable increase from 79% (55/70) before applying this approach (Figure 4A). In the external validation data set, the confusion matrix of the best fold in ResNet-18 showed that 283 out of 268 (95%) mpox and 302 out of 312 (97%) non-mpox images were correctly identified by the model (Figure 4B).

XSL•FO
RenderX

Figure 4. Confusion matrix for best models (5-fold cross-validations). ROI: region of interest.

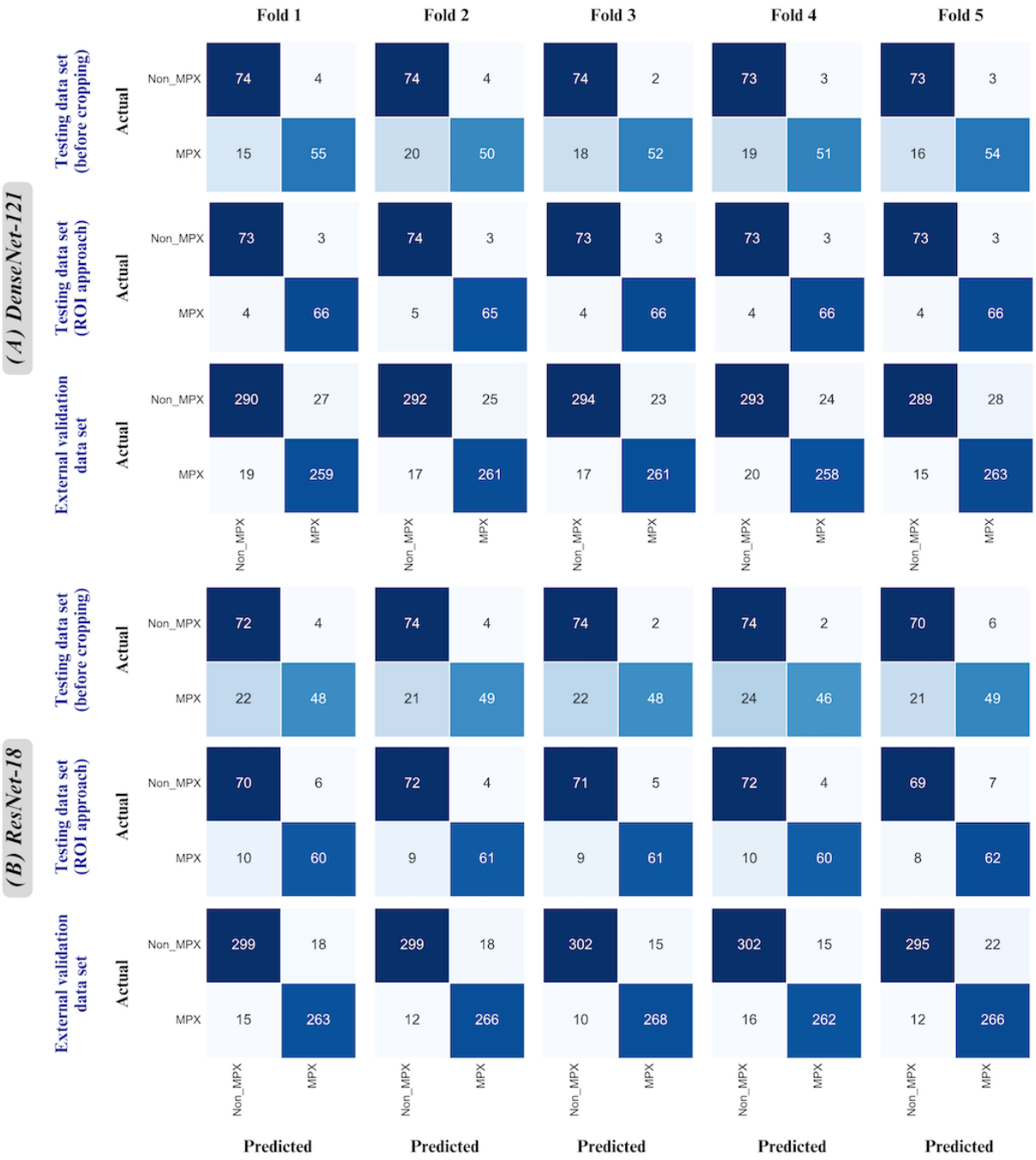

Visual Analysis With Grad-CAM

Figure 5 shows the visual analysis of the lesion images with Grad-CAM. The visual analysis using Grad-CAM on random images from the testing data set (before cropping) showed inaccuracies in the areas of importance for predicted diagnosis (see Figure 5B, first row). We observed false detections occurring in the background rather than the lesion areas of the images from the testing data set. After applying the ROI approach to the testing data set, the model's accuracy in detecting lesion areas showed a significant improvement (Figure 5B, second row).

XSL•FO
RenderX

Figure 5. Visual analysis on predicted image class from DenseNet-121. (A) The original image (left) and the corresponding heatmap image (right) generated by Grad-CAM were shown and the heatmap highlighted regions of the image that the model used to make its prediction. (B)The first row showed the original image, and the second row showed the same image after cropping ROI, showing the accuracy of the model improved after the ROI approach. Grad-CAM: gradient-weighted class activation mapping; ROI: region of interest.

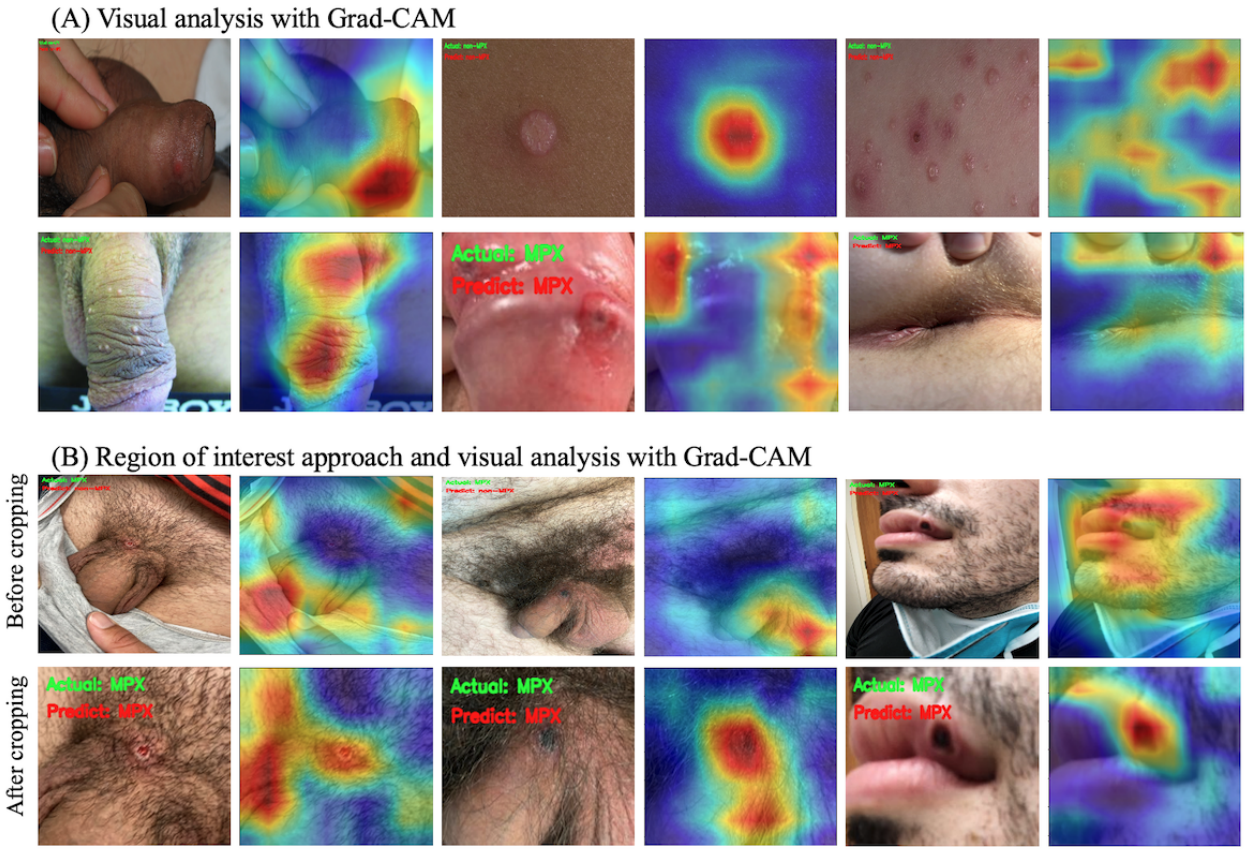

Discussion

Principal Findings

Previous studies [16-18] have shown AI tools can differentiate mpox from non-STI lesions such as measles, smallpox, and chickenpox. Our study added to this by demonstrating that AI can differentiate mpox from images of common skin lesions including STIs and non-STIs collected from a sexual health clinic. We showed that the DenseNet-121 model performed most accurately compared to other models although the differences were small. We also demonstrated that the accuracy of the image classification models could be significantly improved by cropping the main ROI lesion areas from the original image prior to feeding it to the model.

Our study found similar results to other published studies that used AI-based image recognition algorithms to detect mpox cases among symptomatic clinic attendees. The DenseNet-121 model outperformed the other models, achieving a 0.951 accuracy on the testing data set with 147 MSHC mpox images confirmed by laboratory diagnosis. In addition, our DenseNet-121 model also demonstrated a 0.926 accuracy on the external validation data set consisting of 544 mpox images from the Kaggle data set. This external validation data set was previously used in other studies [16-18]. Our model's performance was similar to those used in previous studies. The study by Akin et al [18] achieved a 0.980 accuracy with ResNet-18 for binary classification between mpox and other lesions. Similarly, Shams et al [16] achieved an 0.830 accuracy with ResNet-50 for the same classification task. This higher accuracy was possibly due to the limited variety of lesion images in the testing data sets, consisting of 45 and 56 images, respectively. Another study by Islam et al [17] achieved 0.790 accuracy with ShuffleNet-V2 for multiple classifications among 6 different lesions, with a testing data set of 158 images. However, it is important to note that our findings could not be directly compared with these studies because they used smaller testing data sets, different lesion types, and different model architectures.

There are a number of advantages to our study over other published studies. First, our study was the first one to demonstrate the potential for developing an AI tool to differentiate mpox from common skin lesions from images collected from a sexual health clinic. Second, we used a significantly larger data set compared to previous AI-based mpox classification studies [16-18]. Our data set included 2200 images, in contrast to previous studies' data sets, which included 804 and 228 images. Third, our study included the clinic images in addition to the open-source images from the web-scraping method. All diagnoses of the images from MSHC were confirmed with laboratory tests and clinically by well-experienced sexual health physicians. Finally, we used a 5-fold cross-validation process to improve the robustness and generalizability of our models.

XSL•FO
RenderX

Our study demonstrated the crucial role of preprocessing techniques, specifically focused on cropping the ROI lesion areas, within the context of differentiating mpox from non-mpox lesions. As exemplified by the incorrectly classified cases in Figure 5B, in the first row, confusing backgrounds such as fabrics or hairy skin patterns appeared to contribute to inaccurate model predictions prior to the ROI approach. Additionally, when the main lesion areas comprised very small regions relative to the entire image area, the model struggled to focus on the relevant features. By cropping the images to solely focus on the lesion areas, the model was able to better recognize the lesion features, resulting in improved accuracy. Our study highlights the importance of focusing on the lesion area with minimum background in achieving optimal performance of the AI model. This also underscores the importance of user education and training to ensure that users are able to apply them effectively when using the model in a clinical setting. Nonetheless, it could be a challenge when end users submit images without preprocessing in a real-world setting. It emphasizes the need to integrate image preprocessing functions into the user interface of the tool during deployment; however, it was out of the scope of our study. Overall, our findings provide important insights into the development and optimization of image classification models in medical applications.

There are substantial potential applications of this AI-assisted diagnosis tool. First, this AI-assisted diagnosis tool could be integrated with existing health care platforms, such as patient booking systems, to identify suspected mpox cases among clinic attendees even before they attend the service. Early identification of infectious cases could help manage clinic workflows through appropriate infection control processes. Moreover, the tool could be integrated into the clinic's web-based services, allowing patients to check their symptoms and we could prioritize health services for those identified as mpox and also warn individuals of their risk of transmission. Furthermore, this tool could be integrated into a mobile app for point-of-care services, especially in low-resource health care settings where access to medical professionals may be limited. The mobile app could potentially provide real-time assistance to rural health care providers, enabling more efficient and cost-effective diagnosis. However, it is important to consider the ethical implications associated with the use of AI-assisted diagnostic tools, including concerns about data privacy and unintended consequences. It is necessary to develop appropriate guidelines and frameworks to mitigate these concerns.

Our study had several limitations. First, we used augmentation methods to multiply mpox classes to solve the imbalanced data which might affect the model performance during the training and validation process. Second, we included web-scraped mpox images and this limitation of not being able to confirm their diagnosis with laboratory tests was addressed by having them reviewed and visually diagnosed by sexual health physicians. Third, although our models yielded promising results, they also highlighted opportunities for improvement. The evaluation of model performance showed that there were some misclassifications between classes and false detection areas during visual analysis with Grad-CAM. Given these findings, further research could explore whether increasing the data set size and developing more customized deep learning approaches rather than relying on fine-tuning and ROI cropping methods might improve the model performance. Furthermore, while our models demonstrated differentiation between mpox and non-mpox, it has not been validated in a real-world clinical setting due to the declining number of mpox cases. Our study's aim was primarily to demonstrate the potential use of AI tools to differentiate skin lesions and provide a foundation for future research to test the AI model in real-world scenarios. Finally, we did not use any epidemiological or clinical history data that may also improve the model. For example, currently, there are no new cases of mpox in Australia and so the low underlying incidence of mpox would be likely to influence the accuracy of the models. Moreover, the integration of demographic and sexual behavioral data, such as whether an individual identifies as men who have sex with men and number of sexual partners, could potentially improve the model's performance.

Conclusions

In this study, our study demonstrated the potential use of an AI-based image recognition algorithm to differentiate between mpox and common skin lesions. The model with a high accuracy of 95% indicates its potential benefit from clinic workflow management and prevention of the transmission of mpox infection. However, further research is needed to improve the performance of the model and validate its effectiveness in real-world clinical settings. Overall, the successful development and evaluation of this AI-based image recognition algorithm offer a promising approach to improving diagnostic accuracy and efficiency in sexual health clinics.

Acknowledgments

The authors thank Monash University for the PhD scholarship awarded to NNS. The authors thank all the clinicians at Melbourne Sexual Health Centre, who provided their valuable insights and contributions to the data collection and processing of this study. CKF is supported by the National Health and Medical Research Council (NHMRC) leadership investigator grant (GNT1172900). JJO is supported by the NHMRC investigator grant (GNT1193955). The generative artificial intelligence was not used in any portion of the paper writing.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are not publicly available due to privacy and consent limitations for images collected from the Melbourne Sexual Health Centre (MSHC), but are available from the corresponding author on reasonable request. The external validation data set used in this study is publicly available on Kaggle, titled "Monkeypox Skin Image Dataset."

Authors' Contributions

LZ and NNS conceptualized the study. NNS, ZY, and PML performed model training and analysis. DL, RSS, and CKF collected mpox and sexually transmitted infection image data. ZG, RR, JS, JJO, and CKF contributed to the study design. NNS and PML drafted the paper with input from all coauthors. LZ and CKF oversaw the study as guarantors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data augmentation to address the imbalanced dataset, comparison of pretrained models and their approximate trainable parameters and model size, and performance metrics for classification model: equation and definition.
[DOCX File , 31 KB - jmir_v26i1e52490_app1.docx]

References

1. Fact sheets on monkeypox. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/monkeypox [accessed 2024-08-02]
2. Monkeypox (Mpox). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/poxvirus/monkeypox/symptoms/index.html [accessed 2024-08-02]
3. Monkeypox fact sheets. NSW Health. 2022. URL: https://www.health.nsw.gov.au/Infectious/factsheets/Pages/monkeypox.aspx [accessed 2024-08-02]
4. Australian Government. Mpox (Previously Known as Monkeypox). Australia: Department of Health and Aged Care; 2024.
5. Thomas L. The Implications of Monkeypox for Healthcare Workers. Manchester: News Medical; 2022.
6. Monkeypox (mpox) control guideline. NSW Health. 2022. URL: https://www.health.nsw.gov.au/Infectious/controlguideline/Pages/monkeypox.aspx [accessed 2024-08-02]
7. Infection prevention and control of mpox in healthcare settings. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/poxvirus/monkeypox/clinicians/infection-control-healthcare.html#anchor_1653508896553 [accessed 2024-08-02]
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
9. Born J, Beymer D, Rajan D, Coy A, Mukherjee V, Manica M, et al. On the role of artificial intelligence in medical imaging of COVID-19. Patterns (NY) 2021;2(6):100269 [FREE Full text] [doi: 10.1016/j.patter.2021.100269] [Medline: 33969323]
10. Yu Z, Jiang X, Zhou F, Qin J, Ni D, Chen S, et al. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Trans Biomed Eng 2019 Apr;66(4):1006-1016. [doi: 10.1109/TBME.2018.2866166] [Medline: 30130171]
11. Yu Z, Jiang X, Wang T, Lei B. Aggregating deep convolutional features for melanoma recognition in dermoscopy images. In: Machine Learning in Medical Imaging. Cham: Springer International Publishing; 2017.
12. Jinnai S, Yamazaki N, Hirano Y, Sugawara Y, Ohe Y, Hamamoto R. The development of a skin cancer classification system for pigmented skin lesions using deep learning. Biomolecules 2020 Jul 29;10(8):1123 [FREE Full text] [doi: 10.3390/biom10081123] [Medline: 32751349]
13. Xu X, Ge Z, Chow EPF, Yu Z, Lee D, Wu J, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. J Clin Med 2022 Mar 25;11(7):1818 [FREE Full text] [doi: 10.3390/jcm11071818] [Medline: 35407428]
14. Latt PM, Soe NN, Xu X, Ong JJ, Chow EPF, Fairley CK, et al. Identifying individuals at high risk for HIV and sexually transmitted infections with an artificial intelligence-based risk assessment tool. Open Forum Infect Dis 2024;11(3):ofae011 [FREE Full text] [doi: 10.1093/ofid/ofae011] [Medline: 38440304]
15. Soe NN, Latt PM, Yu Z, Lee D, Kim CM, Tran D, et al. Clinical features-based machine learning models to separate sexually transmitted infections from other skin diagnoses. J Infect 2024 Apr;88(4):106128 [FREE Full text] [doi: 10.1016/j.jinf.2024.106128] [Medline: 38452934]
16. Shams NAM, Joydip P, Tasnim JS, Sani SMS, Noor N, Hasan T. Monkeypox skin lesion detection using deep learning models: a feasibility study. Senior Member 2022. [doi: 10.1109/cait56099.2022.10072140]

XSL•FO
RenderX

17. Islam T, Hussain MA, Chowdhury FUH, Islam BMR. Can artificial intelligence detect monkeypox from digital skin images? Comput Sci Eng 2022. [doi: 10.1101/2022.08.08.503193]
18. Akin KDG, Budak C, Karata A, Karataş H. Classification of monkeypox skin lesion using the explainable artificial intelligence assisted convolutional neural networks. Eur J Sci Technol 2022(40):106-110. [doi: 10.31590/ejosat.1171816]
19. Thieme AH, Zheng Y, Machiraju G, Sadee C, Mittermaier M, Gertler M, et al. A deep-learning algorithm to classify skin lesions from mpox virus infection. Nat Med 2023;29(3):738-747 [FREE Full text] [doi: 10.1038/s41591-023-02225-7] [Medline: 36864252]
20. Sahin VH, Oztel I, Yolcu Oztel G. Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application. J Med Syst 2022;46(11):79 [FREE Full text] [doi: 10.1007/s10916-022-01863-7] [Medline: 36210365]
21. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029 [FREE Full text] [doi: 10.1148/ryai.2020200029] [Medline: 33937821]
22. Monkeypox Skin Images Datset (MSID). Bangladesh: Kaggle; 2022.
23. difPy duplicate image finder - automated search for duplicate or similar images. difPy 4.1.0. 2023. URL: https://pypi.org/project/difPy/ [accessed 2024-08-02]
24. Kim Y, Lee Y, Jeon M. Imbalanced image classification with complement cross entropy. Pattern Recognit Lett 2021;151:33-40. [doi: 10.1016/j.patrec.2021.07.017]
25. Augmentor, Python package designed for augmentation and artificial genteration of image data. Augmentor. 2020. URL: https://augmentor.readthedocs.io/en/stable/ [accessed 2022-11-10]
26. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data 2022;9(1):102 [FREE Full text] [doi: 10.1186/s40537-022-00652-w] [Medline: 36313477]
27. Xu M, Yoon S, Jeong Y, Park DS. Transfer learning for versatile plant disease recognition with limited data. Front Plant Sci 2022;13:1010981 [FREE Full text] [doi: 10.3389/fpls.2022.1010981] [Medline: 36507376]
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: inverted residuals and linear bottlenecks. 2018 Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; USA p. 4510-4520. [doi: 10.1109/cvpr.2018.00474]
29. Ningning M, Xiangyu Z, Hai-Tao Z, Sun J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. China: ShuffleNet; 2018.
30. Huang G, Liu Z, Der Maaten LV, Weinberger KQ. Densely connected convolutional networks. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21; Hawaii, USA p. 4700-4708. [doi: 10.1109/cvpr.2017.243]
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 June 27; Nevada, USA p. 770-778. [doi: 10.1109/cvpr.2016.90]
32. Swin transformer: models and pretrained weights. PyTorch. 2022. URL: https://pytorch.org/vision/main/models/generated/torchvision.models.swin_b.html [accessed 2024-08-02]
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2013 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); 2017; Venice, Italy p. 618-626 URL: https://ieeexplore.ieee.org/document/8237336/authors#authors
34. Ling CX, Huang J, Zhang H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. Berlin, Heidelberg: Springer; 2003:329-341.
35. Fawcett T. ROC graphs: notes and practical considerations for researchers. Machine Learning 2004;31:1-38.
36. Scikit-learn machine learning in Python. scikit learn. 2022. URL: https://scikit-learn.org/stable/index.html [accessed 2022-12-10]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CLAIM: checklist for Artificial Intelligence in Medical Imaging
MSHC: Melbourne Sexual Health Centre
ROI: region of interest
STI: sexually transmitted infection

XSL•FO
RenderX

Edited by S Ma; submitted 06.09.23; peer-reviewed by V Singh, T Abd El-Hafeez; comments to author 08.02.24; revised version received 11.02.24; accepted 10.06.24; published 13.09.24.

Please cite as:
Soe NN, Yu Z, Latt PM, Lee D, Samra RS, Ge Z, Rahman R, Sun J, Ong JJ, Fairley CK, Zhang L
Using AI to Differentiate Mpox From Common Skin Lesions in a Sexual Health Clinic: Algorithm Development and Validation Study
J Med Internet Res 2024;26:e52490
URL: https://www.jmir.org/2024/1/e52490
doi:10.2196/52490
PMID:

©Nyi Nyi Soe, Zhen Yu, Phyu Mon Latt, David Lee, Ranjit Singh Samra, Zongyuan Ge, Rashidur Rahman, Jiajun Sun, Jason J Ong, Christopher K Fairley, Lei Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Trial Factors Associated With Completion of Clinical Trials Evaluating AI: Retrospective Case-Control Study

David Chen[1], BMSc; Christian Cao[1], BHSc; Robert Kloosterman[1], MSc; Rod Parsa[2], MSc; Srinivas Raman[3], MASc, MD

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[2]Michael G. DeGroote School of Medicine, McMaster University, Hamilton, ON, Canada

[3]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Srinivas Raman, MASc, MD
Department of Radiation Oncology
University of Toronto
610 University Avenue
Toronto, ON, M5G 2M9
Canada
Phone: 1 416 946 4501 ext 2320
Fax: 1 416 946 6566
Email: srinivas.raman@uhn.ca

Abstract

Background: Evaluation of artificial intelligence (AI) tools in clinical trials remains the gold standard for translation into clinical settings. However, design factors associated with successful trial completion and the common reasons for trial failure are unknown.

Objective: This study aims to compare trial design factors of complete and incomplete clinical trials testing AI tools. We conducted a case-control study of complete (n=485) and incomplete (n=51) clinical trials that evaluated AI as an intervention of ClinicalTrials.gov.

Methods: Trial design factors, including area of clinical application, intended use population, and intended role of AI, were extracted. Trials that did not evaluate AI as an intervention and active trials were excluded. The assessed trial design factors related to AI interventions included the domain of clinical application related to organ systems; intended use population for patients or health care providers; and the role of AI for different applications in patient-facing clinical workflows, such as diagnosis, screening, and treatment. In addition, we also assessed general trial design factors including study type, allocation, intervention model, masking, age, sex, funder, continent, length of time, sample size, number of enrollment sites, and study start year. The main outcome was the completion of the clinical trial. Odds ratio (OR) and 95% CI values were calculated for all trial design factors using propensity-matched, multivariable logistic regression.

Results: We queried ClinicalTrials.gov on December 23, 2023, using AI keywords to identify complete and incomplete trials testing AI technologies as a primary intervention, yielding 485 complete and 51 incomplete trials for inclusion in this study. Our nested propensity-matched, case-control results suggest that trials conducted in Europe were significantly associated with trial completion when compared with North American trials (OR 2.85, 95% CI 1.14-7.10; *P*=.03), and the trial sample size was positively associated with trial completion (OR 1.00, 95% CI 1.00-1.00; *P*=.02).

Conclusions: Our case-control study is one of the first to identify trial design factors associated with completion of AI trials and catalog study-reported reasons for AI trial failure. We observed that trial design factors positively associated with trial completion include trials conducted in Europe and sample size. Given the promising clinical use of AI tools in health care, our results suggest that future translational research should prioritize addressing the design factors of AI clinical trials associated with trial incompletion and common reasons for study failure.

(J Med Internet Res 2024;26:e58578) doi:10.2196/58578

KEYWORDS

artificial intelligence; clinical trial; completion; AI; cross-sectional study; application; intervention; trial design; logistic regression; Europe; clinical; trials testing; health care; informatics; health information

XSL•FO
RenderX

Introduction

The advent of artificial intelligence (AI) is expected to transform the practice and delivery of health care practices, including applications in clinical diagnosis, treatment, and management [1,2]. Adoption of these promising but often untested tools requires systematic evaluation through clinical trials, widely regarded as one of the highest forms of evidence to inform clinical practice [3,4].

It is well known that clinical trials fail to compete at different stages of the research and development process [5,6]. However, the evaluation of AI tools as interventions in clinical trials raises the potential for new modes of trial incompletion that have yet to be explored. Unique challenges in translation AI research can include poor patient cohort selection, ineffective patient monitoring during trials, and logistical difficulties for implementation [7,8].

Prioritizing research to address the common limitations of AI in trials is a critical step toward the validation and adoption of these tools. To address this issue, we performed a case-control study of AI trials in ClinicalTrials.gov to identify trial design factors associated with trial completion and catalog study-reported reasons for trial incompletion.

Methods

We queried ClinicalTrials.gov, the largest international registry of clinical trials, on December 23, 2023, using AI keywords based on previous methodology to identify AI-related trials (n=6738) [9]. Incomplete trials were defined as terminated, suspended, or withdrawn trials. We only included complete and incomplete trials testing AI technologies as a primary intervention, categorized by complete (n=485) or incomplete status (n=51; Figure S1 in Multimedia Appendix 1). Our study focused on clinical trials that evaluated AI technologies in at least 1 study arm to assess the characteristics of trial completion associated with primary AI interventions. We excluded exact duplicate trials but included separate trials evaluating the same AI intervention for different trial methods or targeted populations. We excluded ongoing trials given their unknown status of trial completion by the date of data collection needed for this study's case-control design. We excluded observational trials and studies with missing trial design elements such as allocation, intervention model, and masking.

Two reviewers (CC and RK) independently screened all studies for inclusion and data extraction after a pilot on 20 studies to improve interreviewer agreement. Discordance of screening and data extraction were resolved through discussion with a third reviewer (DC) to achieve full agreement across the reviewer team. We extracted trial design factors including the clinical area addressed by AI technology, the intended use population, and the intended role of AI technology (Table S1 in Multimedia Appendix 1). Trial factors with no available data were coded as "Unknown."

Chi-square tests with Benjamini-Hochberg correction were used to compare the distribution of trial factors between complete and incomplete trials. Logistic regression models were fit for univariable and multivariable analysis of trial factors associated with trial completion. Stepwise variable selection using the Akaike information criterion was used to identify the optimal set of trial factors useful for the multivariable regression model. The propensity score–matched, multivariable regression was conducted using 3 complete trials for each incomplete trial, matched based on all insignificant trial factors from the multivariable analysis. Odds ratio (OR) and 95% CI values were calculated for all factors observed in both complete and incomplete trials. A 2-sided *P* value threshold of .05 was used for statistical significance. This study was completed in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [10].

No ethics approval and informed consent were needed since this study analyzed publicly available data and did not include human subjects.

Results

The majority of AI trials were categorized as being completed (485/536, 90.5%). Trials primarily implemented diagnostic AI interventions (200/536, 37.3%), tested AI interventions intended for health care providers (432/536, 80.6%), and were conducted in adult and older adult populations (397/536, 74.6%). Furthermore, the most prevalent clinical areas addressed in trials included oncology (93/536, 17.4%), cardiovascular system (71/536, 13.2%), and generic health (60/536, 11.2%). Geographically, the majority of trials were conducted in Europe (174/536, 32.5%), North America (145/536, 27.1%), or Asia (142/536, 26.5%), and many trials recruited relatively larger sample sizes (1000 participants; 131/536, 24.4%). We found a paucity of reporting information concerning study allocation (399/536, 74.4% unknown), intervention model (321/536, 60.1% unknown), masking procedures (59.9% unknown), and funding source (438/536, 81.7% unknown) and excluded these factors from the univariable analysis. Multimedia Appendix 2 reports a summary of the included studies.

From the univariable logistic regression model, the role of AI (prediction: OR 3.93, 95% CI 1.32-11.72; *P*<.001), study type (interventional: OR 0.52, 95% CI 0.29-0.92; *P*=.03), continent (Asia: OR 16.0, 95% CI 3.73-68.77; *P*<.001 and Europe: OR 4.11, 95% CI 1.90-9.25; *P*<.001), and sample size (OR, 1.00; 95% CI 1.000-1.003; *P*<.001) were associated with completion of AI trials (Multimedia Appendix 3). All significant factors from univariable analysis were included in our multivariable model. From the multivariable logistic regression model, the role of AI (prediction: OR 4.55, 95% CI 1.44-14.36; *P*=.01), continent (Asia: OR 11.57, 95% CI 2.59-51.73; *P*=.001 and Europe: OR 4.44, 95% CI 1.91-10.3; *P*<.001), and sample size (OR 1.00, 95% CI 1.000232-1.00243; *P*=.02) were associated with completion of AI trials (Multimedia Appendix 3). The case-control, propensity-matched, multivariable logistic regression model found that continent (Europe: OR 2.85, 95% CI 1.14-7.10; *P*=.02) and sample size (OR 1.00, 95% CI 1.000202-1.00218; *P*=.02) were associated with completion of AI trials (Multimedia Appendix 3)

XSL•FO
RenderX

Common study-reported reasons for trial incompletion include poor accrual (13/51, 25.5%), poor results at interim (3/51, 5.9%), administration (22/51, 43.1%), and other (12/51, 23.5%; Figure 1). Among incomplete trials due to poor administration, reported reasons included logistical difficulties (8/22, 36.4%), lack of funding (7/22, 31.8%), COVID-19 pandemic (4/22, 18.2%), departure of investigator (3/22, 13.6%), and lack of ethics approval (1/22, 4.5%).

Figure 1. Reasons for failure of clinical trials that evaluated artificial intelligence. PI: principal investigator.

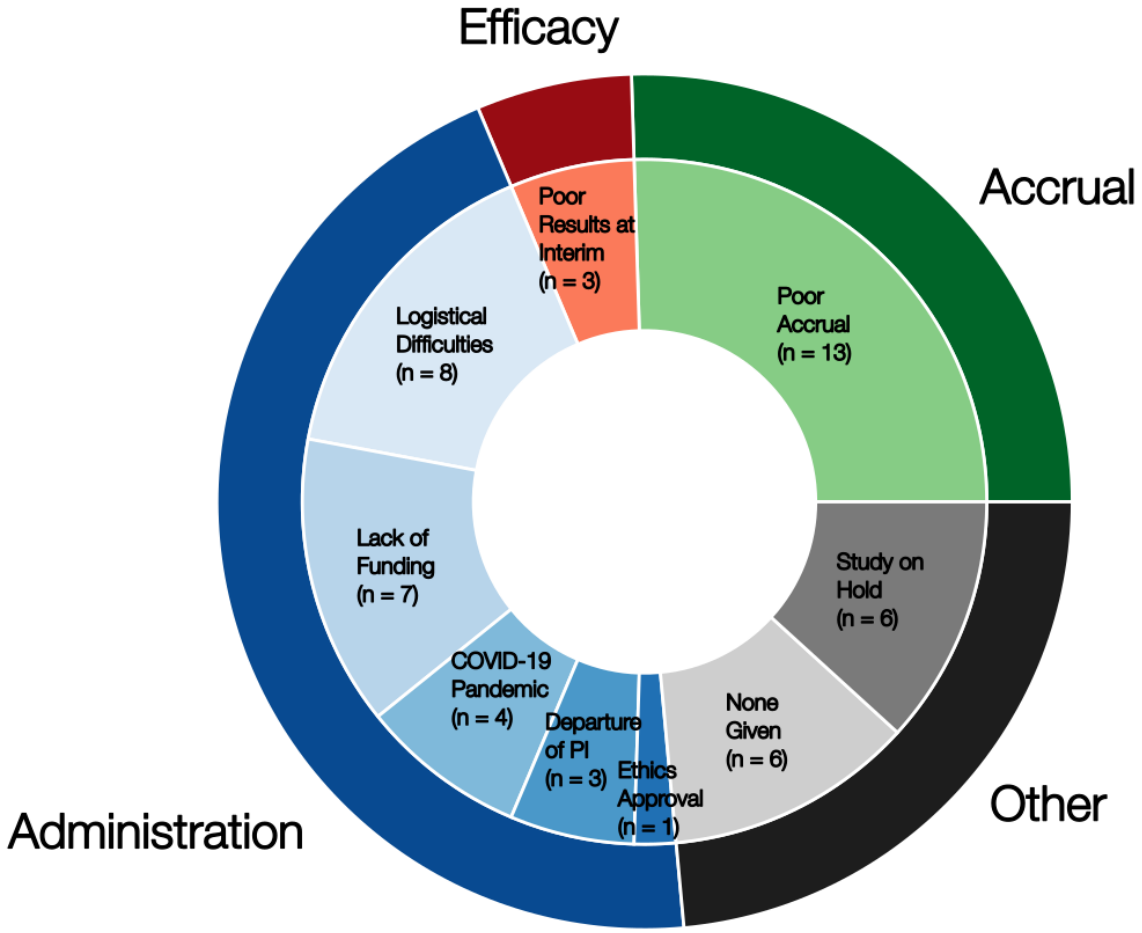

Discussion

Principal Findings

Our multivariable analysis found that clinical trials conducted in Europe were positively associated with trial completion compared with trials in North America. However, we note that trials conducted outside of North America are less likely to be registered on ClinicalTrials.gov [11], which may be due in part to different national and funding mandates, thus potentially resulting in geographic reporting bias. There remains a need for sound methodological design in AI model training to improve generalizability in validation cohorts [12]. We hypothesize that AI models trained on data from local, homogenous cohorts may be more likely to complete but could fail to generalize to external cohorts with increased heterogeneity and a lack of training representation. The design of AI training architectures should consider representation from data-poor sources and due diligence in external validation before clinical implementation [13].

Our findings also highlight the importance of trial design, noting that larger participant sizes correlate with successful trial completion, which may be unique to AI-based trials, where the performance of the tools are enhanced with larger datasets. We note that studies involving large sample sizes may lead to false positive discoveries due to inflation of *P* values [14], and administrative difficulties that can contribute to trial failure [15]. This finding aligns with recent reports on the need to consider appropriate sample size to ensure reliable estimates of AI intervention performance [16] as well as the association between sample size and the sensitivity of detecting differences in study outcomes [17,18]. The emergence of noninferiority trials evaluating the performance of AI interventions compared with standard-of-care controls should require consistent reporting and justification for sample size [19] and should consider the use and challenges of large-scale training and validation cohorts for AI models [20]. Randomized clinical trials (RCTs), including but not limited to trials evaluating AI interventions, should consider sample size with respect to type 1 error, power, effect size of clinical interest, and population variance, as well as justify the use of sample size calculations based on applicable assumptions [21].

Compared with a cross-sectional study of all trials reported in ClinicalTrials.gov [22], our results also demonstrated that administrative reasons made up a greater proportion (45.1% vs

XSL•FO
RenderX

29.8%) of reasons for AI trial failure compared with all trials, and further research is required to understand unique administrative challenges present in AI trials. There remain several key challenges that should be addressed to translate AI interventions in medicine, including the assessment of performance metrics in relation to clinical use, algorithmic biases that limit generalizability to new populations, and logistical difficulties in implementing AI systems into clinical workflows led by clinicians [8]. Broadly, the shift of AI interventions toward integration into compound systems with multiple inputs, outputs, and operators may present new administrative challenges in clinical workflows that should be addressed in future AI clinical trials.

Our study has several limitations. First, this study is limited by the lack of paired trials in the literature that could provide case-control comparisons between different trial design factors. Despite incorporating relevant trial design covariates and applying propensity-matching techniques, simplifying the intricacies of trial completion into its constituent design factors may overlook several other considerations, which could significantly influence trial outcomes. Second, there was a paucity of information reporting, where several study design factors were not consistently reported [23]. Researchers should adhere to reporting guidelines wherever possible to enhance scientific transparency and accountability [24].

Conclusion

This study suggests that clinical trials that recruited larger sample sizes and conducted in Europe, compared with North America, are associated with successful trial completion. The most common reasons for trial incompletion included poor participant accrual and administrative difficulties. Future research is needed to address the limitations of AI clinical trials associated with trial incompletion to improve the translation of AI into clinical practice.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inclusion and exclusion criteria and data extraction procedure.
[DOCX File , 203 KB - jmir_v26i1e58578_app1.docx]

Multimedia Appendix 2
Descriptive statistics of complete and incomplete interventional trials evaluating artificial intelligence.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v26i1e58578_app2.xlsx]

Multimedia Appendix 3
Predictive statistics of complete and incomplete interventional trials evaluating artificial intelligence.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v26i1e58578_app3.xlsx]

References

1. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126 [FREE Full text] [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
2. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
3. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA 2009;301(8):831-841. [doi: 10.1001/jama.2009.205] [Medline: 19244190]
4. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open 2022;5(9):e2233946 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33946] [Medline: 36173632]
5. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
6. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018;11:156-164 [FREE Full text] [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
7. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019;40(8):577-591 [FREE Full text] [doi: 10.1016/j.tips.2019.05.005] [Medline: 31326235]
8. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]

XSL•FO
RenderX

9. Pearce FJ, Cruz Rivera S, Liu X, Manna E, Denniston AK, Calvert MJ. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997-2022). Lancet Digit Health 2023;5(3):e160-e167 [FREE Full text] [doi: 10.1016/S2589-7500(22)00249-7] [Medline: 36828608]
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
11. Lindsley K, Fusco N, Li T, Scholten R, Hooft L. Clinical trial registration was associated with lower risk of bias compared with non-registered trials among trials included in systematic reviews. J Clin Epidemiol 2022;145:164-173 [FREE Full text] [doi: 10.1016/j.jclinepi.2022.01.012] [Medline: 35081449]
12. Eche T, Schwartz LH, Mokrane F, Dercle L. Toward generalizability in the deployment of artificial intelligence in radiology: role of computation stress testing to overcome underspecification. Radiol Artif Intell 2021;3(6):e210097 [FREE Full text] [doi: 10.1148/ryai.2021210097] [Medline: 34870222]
13. Celi LA, Cellini J, Charpignon M, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-a global review. PLOS Digit Health 2022;1(3):e0000022 [FREE Full text] [doi: 10.1371/journal.pdig.0000022] [Medline: 36812532]
14. Lin M, Lucas HC, Shmueli G. Too big to fail: larger samples and false discoveries. Robert H. Smith School Research Paper No. RHS 06-068 2011:37. [doi: 10.2139/ssrn.1336700]
15. Reith C, Landray M, Devereaux P, Bosch J, Granger CB, Baigent C, et al. Randomized clinical trials--removing unnecessary obstacles. N Engl J Med 2013;369(11):1061-1065. [doi: 10.1056/NEJMsb1300760] [Medline: 24024845]
16. Rajput D, Wang W, Chen C. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 2023;24(1):48 [FREE Full text] [doi: 10.1186/s12859-023-05156-9] [Medline: 36788550]
17. Gieraerts C, Dangis A, Janssen L, Demeyere A, de Bruecker Y, de Brucker N, et al. Prognostic value and reproducibility of AI-assisted analysis of lung involvement in COVID-19 on low-dose submillisievert chest CT: sample size implications for clinical trials. Radiol Cardiothorac Imaging 2020;2(5):e200441 [FREE Full text] [doi: 10.1148/ryct.2020200441] [Medline: 33778634]
18. Liu L, Parker KJ, Jung S. Design and analysis methods for trials with AI-based diagnostic devices for breast cancer. J Pers Med 2021;11(11):1150 [FREE Full text] [doi: 10.3390/jpm11111150] [Medline: 34834502]
19. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PPJ. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. BMJ Open 2016;6(10):e012594 [FREE Full text] [doi: 10.1136/bmjopen-2016-012594] [Medline: 27855102]
20. L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM. Machine learning with big data: challenges and approaches. IEEE Access 2017;5:7776-7797. [doi: 10.1109/access.2017.2696365]
21. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant 2010;25(5):1388-1393. [doi: 10.1093/ndt/gfp732] [Medline: 20067907]
22. Williams RJ, Tse T, DiPiazza K, Zarin DA. Terminated trials in the clinicaltrials.gov results database: evaluation of availability of primary outcome data and reasons for termination. PLoS One 2015;10(5):e0127242 [FREE Full text] [doi: 10.1371/journal.pone.0127242] [Medline: 26011295]
23. Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials.gov. N Engl J Med 2015;372(11):1031-1039 [FREE Full text] [doi: 10.1056/NEJMsa1409364] [Medline: 25760355]
24. Kwong JCC, Khondker A, Lajkosz K, McDermott MBA, Frigola XB, McCradden MD, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open 2023;6(9):e2335377 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.35377] [Medline: 37747733]

Abbreviations

AI: artificial intelligence
OR: odds ratio
RCT: randomized clinical trials
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.03.24; peer-reviewed by A Wani, J Mistry; comments to author 27.04.24; revised version received 02.05.24; accepted 11.07.24; published 23.09.24.

Please cite as:
Chen D, Cao C, Kloosterman R, Parsa R, Raman S
Trial Factors Associated With Completion of Clinical Trials Evaluating AI: Retrospective Case-Control Study
J Med Internet Res 2024;26:e58578
URL: https://www.jmir.org/2024/1/e58578
doi:10.2196/58578
PMID:39312296

©David Chen, Christian Cao, Robert Kloosterman, Rod Parsa, Srinivas Raman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Natural Language Processing (GPT-4) for Computed Tomography Image Analysis of Cerebral Hemorrhages in Radiology: Retrospective Analysis

Daiwen Zhang[1,2*], PhD; Zixuan Ma[1,2*], PhD; Ru Gong[1*], PhD; Liangliang Lian[3*], MD; Yanzhuo Li[4*], MD; Zhenghui He[1,2], PhD; Yuhan Han[1,2], PhD; Jiyuan Hui[1], PhD; Jialin Huang[2], PhD; Jiyao Jiang[1,2], PhD; Weiji Weng[1,2*], PhD; Junfeng Feng[1,2], PhD

[1]Brain Injury Centre, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]Shanghai Institute of Head Trauma, Shanghai, China

[3]Department of Radiology, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]Department of Radiology, Minhang Hospital, Fudan University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Junfeng Feng, PhD
Brain Injury Centre
Ren Ji Hospital
Shanghai Jiao Tong University School of Medicine
160 Pujian Road, Pudong New District
Shanghai, 200127
China
Phone: 86 136 1186 0825
Fax: 86 021 68383709
Email: fengjfmail@163.com

Abstract

Background: Cerebral hemorrhage is a critical medical condition that necessitates a rapid and precise diagnosis for timely medical intervention, including emergency operation. Computed tomography (CT) is essential for identifying cerebral hemorrhage, but its effectiveness is limited by the availability of experienced radiologists, especially in resource-constrained regions or when shorthanded during holidays or at night. Despite advancements in artificial intelligence–driven diagnostic tools, most require technical expertise. This poses a challenge for widespread adoption in radiological imaging. The introduction of advanced natural language processing (NLP) models such as GPT-4, which can annotate and analyze images without extensive algorithmic training, offers a potential solution.

Objective: This study investigates GPT-4's capability to identify and annotate cerebral hemorrhages in cranial CT scans. It represents a novel application of NLP models in radiological imaging.

Methods: In this retrospective analysis, we collected 208 CT scans with 6 types of cerebral hemorrhages at Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, between January and September 2023. All CT images were mixed together and sequentially numbered, so each CT image had its own corresponding number. A random sequence from 1 to 208 was generated, and all CT images were inputted into GPT-4 for analysis in the order of the random sequence. The outputs were subsequently examined using Photoshop and evaluated by experienced radiologists on a 4-point scale to assess identification completeness, accuracy, and success.

Results: The overall identification completeness percentage for the 6 types of cerebral hemorrhages was 72.6% (SD 18.6%). Specifically, GPT-4 achieved higher identification completeness in epidural and intraparenchymal hemorrhages (89.0%, SD 19.1% and 86.9%, SD 17.7%, respectively), yet its identification completeness percentage in chronic subdural hemorrhages was very low (37.3%, SD 37.5%). The misidentification percentages for complex hemorrhages (54.0%, SD 28.0%), epidural hemorrhages (50.2%, SD 22.7%), and subarachnoid hemorrhages (50.5%, SD 29.2%) were relatively high, whereas they were relatively low for acute subdural hemorrhages (32.6%, SD 26.3%), chronic subdural hemorrhages (40.3%, SD 27.2%), and intraparenchymal hemorrhages (26.2%, SD 23.8%). The identification completeness percentages in both massive and minor bleeding showed no

XSL•FO
RenderX

significant difference (*P*=.06). However, the misidentification percentage in recognizing massive bleeding was significantly lower than that for minor bleeding (*P*=.04). The identification completeness percentages and misidentification percentages for cerebral hemorrhages at different locations showed no significant differences (all *P*>.05). Lastly, radiologists showed relative acceptance regarding identification completeness (3.60, SD 0.54), accuracy (3.30, SD 0.65), and success (3.38, SD 0.64).

Conclusions: GPT-4, a standout among NLP models, exhibits both promising capabilities and certain limitations in the realm of radiological imaging, particularly when it comes to identifying cerebral hemorrhages in CT scans. This opens up new directions and insights for the future development of NLP models in radiology.

Trial Registration: ClinicalTrials.gov NCT06230419; https://clinicaltrials.gov/study/NCT06230419

(J Med Internet Res 2024;26:e58741) doi:10.2196/58741

KEYWORDS

GPT-4; natural language processing; NLP; artificial intelligence; AI; cerebral hemorrhage; computed tomography; CT

Introduction

Cerebral hemorrhage mainly encompasses intracranial bleeding resultant from trauma, hypertension, or cerebral vascular disorders. Typically, it manifests acutely, with severe symptoms and a prognosis that is often unfavorable, thereby imposing substantial economic burdens on individuals, families, and society [1-3]. Hence, the prompt and accurate diagnosis of cerebral hemorrhage is of paramount clinical and societal importance.

Computed tomography (CT) scanning stands as the quintessential diagnostic tool for cerebral hemorrhage. An expedient and precise diagnosis not only facilitates the comprehensive evaluation of patient conditions, the optimization of therapeutic strategies, and the prognostication of outcomes but also enhances communication between clinicians and patients or their respective families. Nonetheless, several challenges persist. First, CT diagnosis of cerebral hemorrhage often requires experienced radiologists. In low-income areas or countries, there is a significant shortage of these specialists [4,5]. Radiologists lacking sufficient experience will lead to missed or incorrect diagnosis, which can be potentially fatal in the clinical management of cerebral hemorrhage. Second, CT diagnosis of cerebral hemorrhage requires rapid reporting, but even in high-income areas and countries, the substantial workload makes it difficult for radiologists to provide prompt and comprehensive radiology reports [6]. Third, cultivating an experienced radiologist requires a mature medical education system and a substantial database of imaging resources, which is currently lacking in China [7].

Artificial intelligence (AI) has emerged as a formidable instrument in computer-aided diagnosis across various medical fields, enhancing classification [8], detection [9], and image segmentation [10]. AI proves particularly beneficial in supporting decision-making and rapid diagnosis, especially in underserved rural areas or densely populated regions with a scarcity of medical imaging experts [11]. Moreover, AI has substantially advanced medical education [12-14].

Currently, various AI models, including You Only Look Once and bespoke convolutional neural networks [15-17], have demonstrated efficacy in detecting cerebral hemorrhages in CT imagery. However, the construction and utilization of these AI models necessitate an understanding of computer algorithms. For most radiologists, the implementation of these models poses a significant challenge, indicating a need for more intuitive and accessible AI-driven diagnostic solutions.

The recent introduction of new natural language processing (NLP) models, such as ChatGPT (OpenAI) and the latest iteration of GPT-4, has enabled imaging analysis capabilities without the need for intricate algorithmic knowledge and extensive training [18,19]. Previously, large language models faced the challenge of hallucination, presenting presumably persuasive results that were not based on the input [20,21]. However, many studies suggest that GPT-4 can assist in generating, extracting, and interpreting radiology reports [22-24]. GPT-4 can play a role in the diagnosis and treatment of oral diseases, orthopedic diseases, neurological disorders, and pulmonary diseases using radiology reports [23,25-27]. GPT-4 is capable of accurately extracting lesion information from radiology reports, such as identifying metastatic diseases and generating correct labels for tumor progression [23,24]; it can integrate radiology reports and medical history to make diagnoses [27]; and it can even provide some treatment suggestions based on the radiology reports [26]. However, current research focuses only on GPT-4's capabilities with free text, and its potential for imaging recognition has not yet been explored.

This work aims to delineate our study by utilizing GPT-4 for the identification and annotation of cerebral hemorrhages in CT images, including various types of intracranial bleeding. It is the first attempt to directly use GPT-4 to identify CT images of cerebral hemorrhages. It broadens the horizon for GPT-4 applications in the medical field, offering fresh perspectives for the development of GPT-4 and other NLP models in health care.

Methods

Ethical Considerations

This retrospective study was registered at ClinicalTrials.gov (NCT06230419) and approved by the Ethics Committee of Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine (IIT-2024-0006). This study exclusively utilized CT images, which do not include personally identifiable information or sensitive individual data. There was no direct interaction with patients, complying with the principles of ethical conduct in research. Meanwhile, the Ethics Committee of Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, approved

XSL•FO
RenderX

the usage of CT images without any identifiable private information. All the data are securely stored and only accessed by personnel involved in the research team.

Study Design

As shown in Figure 1, raw CT images of different types of cerebral hemorrhages were collected between January and September 2023 from the radiology database of Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine. Since GPT-4 cannot recognize continuous CT images, we first preprocessed the CT images. We chose the horizontal cranial CT image with the largest volume of hemorrhage in the brain window (window width: 90, window level: 35) as the representative image. Representative CT images were exported to be unified as JPG format files with a size of 700×700 pixels to minimize the influence from the image format and the image size.

By employing a predefined question-and-answer mode, we guided GPT-4 to annotate the hemorrhagic lesions in the provided images (Multimedia Appendix 1). All CT images of different types of cerebral hemorrhages were mixed together and each CT image was sequentially numbered, so all 208 CT images had their own corresponding number. A random sequence from 1 to 208 was generated using Python (version 3.8; Python Software Foundation), and all CT images were inputted into GPT-4 for analysis in the order of the random sequence. Once GPT-4 finished the analyze process, all annotated pictures were downloaded and saved as JPG format files.

All processed CT images were both analyzed by Photoshop (version 24.3.0; Adobe) and given to 10 radiologists to evaluate identification completeness, accuracy, and success.

Figure 1. Flowchart of the study design. ASH: acute subdural hemorrhage; CH: complex hemorrhage; CSH: chronic subdural hemorrhage; CT: computed tomography; EH: epidural hemorrhage; IH: intraparenchymal hemorrhage; SH: subarachnoid hemorrhage.

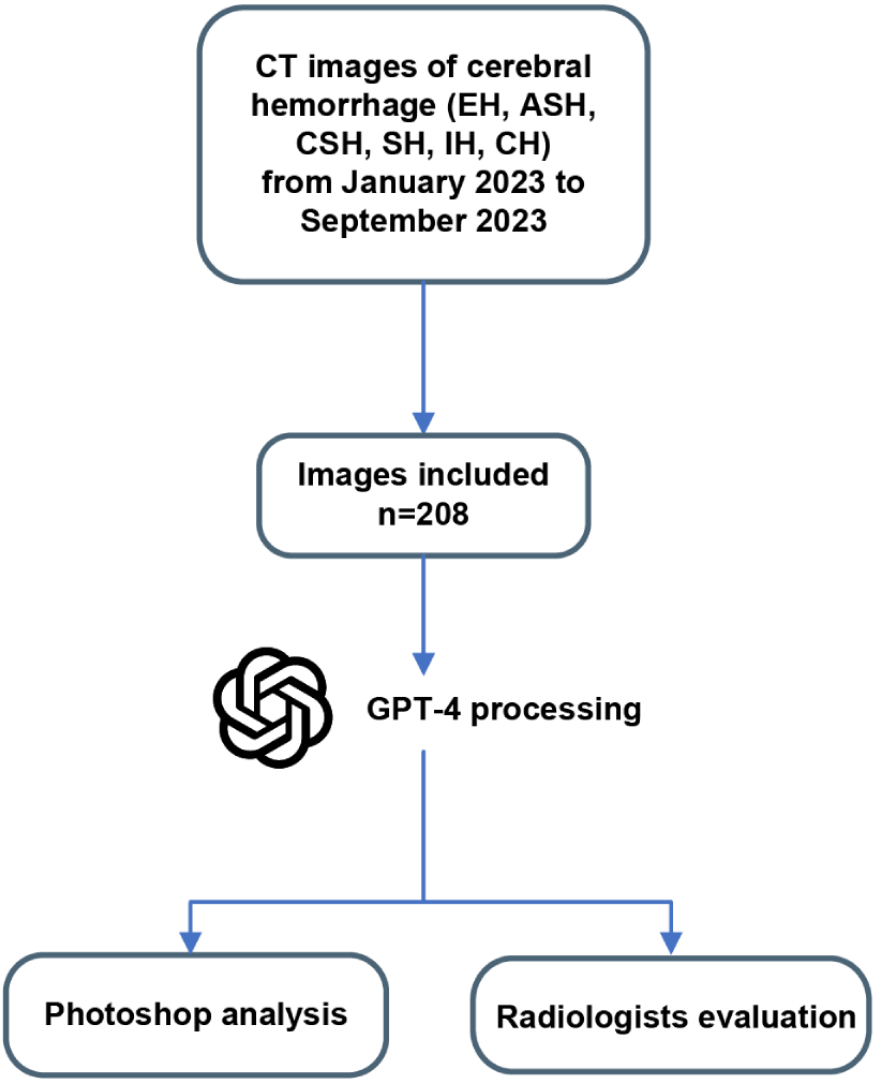

GPT-4 Update

We utilized the latest updated GPT-4 released by OpenAI in November 2023, which at present, empowers more powerful image analysis in GPT-4.

Photoshop Analysis

After all the processed images were collected, the hemorrhage areas and the annotation areas were isolated from the images using Photoshop [28]. According to previous studies, the ABC/2 method has been universally applied in clinical medicine to

XSL•FO
RenderX

estimate cerebral hemorrhage volumes swiftly [29,30]. The same method was used in the calculation of intraparenchymal hemorrhage volumes, given that intraparenchymal hemorrhages typically conform to the assumption of this formula: the hemorrhages approximate an elliptical shape. After tallying all intraparenchymal hemorrhage volumes, based on the corresponding bleeding areas measured using Photoshop, the hemorrhage volume per pixel was determined and the hemorrhage volume for each CT image in the other 5 types of irregular cerebral hemorrhages was calculated.

After the size of the exact image's bleeding and annotated part were counted, the identification completeness and misidentification percentages were calculated through the following formulas (Figure 2):

Identification completeness percentage = correct annotated area / total bleeding area × 100%

Misidentification percentage = (total annotated area – correct annotated area) / total annotated area × 100%

Figure 2. Calculation methods of identification completeness and misidentification percentage using Photoshop. The 3 pairs of pictures show the method of using Photoshop to isolate the total bleeding area, the correct annotated area, and the total annotated area.

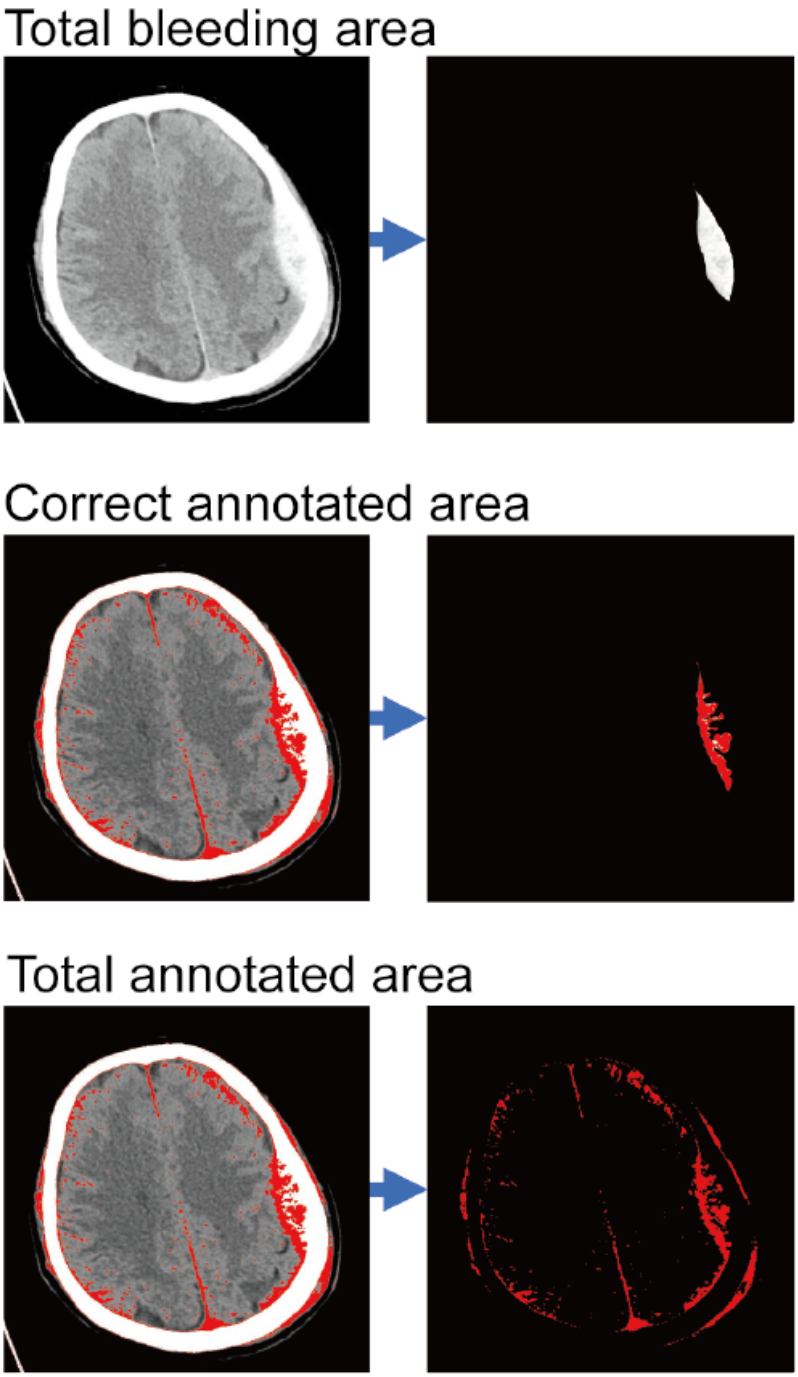

XSL•FO
RenderX

Radiologist Evaluation

For the purpose of evaluating the competence of GPT-4 in cerebral hemorrhage identification, 10 professional radiologists were invited to compare the original CT images and the annotated images. We designed a 4-point scale questionnaire. In this questionnaire, the original image and the annotated image were compared in pairs, and those pairs of CT images were shown to the radiologists. To avoid ambiguous and unclear outcomes, the choices of the questionnaire did not contain neutral responses. The questionnaire was completed by radiologists separately and alone. The criteria and groups used for rating are shown in Multimedia Appendix 2.

Statistical Analysis

The data in this study were all continuous and conformed to normal distribution and homogeneity of variance; therefore, we chose the 2-tailed Student *t* test and 1-way ANOVA for statistical analysis. SPSS (version 16.0; IBM Corp) was used to perform all statistical analysis. Summary data are presented as mean and SD, with statistical significance assessed by 2-tailed Student *t* test for 2-group comparisons or 1-way ANOVA for comparisons with more than 2 groups. $P<.05$ was considered statistically significant.

Results

A total of 208 CT scans of different types of cerebral hemorrhages were collected between January and September in 2023 from the radiology database. These images consisted of epidural hematomas (32 images); acute subdural hematomas (42 images); chronic subdural hematomas (30 images); subarachnoid hemorrhages (36 images); intraparenchymal hemorrhages (31 images); and complex hemorrhages, which included 2 or more different types of intracranial bleeding (37 images).

First and foremost, we were keen to assess whether GPT-4 could identify cerebral hemorrhages and the completeness of such detections. In Figure 3A, we present the performance of GPT-4 across 6 different types of hemorrhages, where red indicates the annotated hemorrhagic lesions identified by GPT-4. The overall identification completeness percentage for the 6 types of hemorrhages by GPT-4 reached 72.6% (SD 18.6%). Among these, GPT-4 demonstrated the highest identification completeness percentages for epidural and intraparenchymal hemorrhages, achieving 89.0% (SD 19.1%) and 86.9% (SD 17.7%), respectively. The identification completeness percentages for acute subdural hemorrhages (74.4%, SD 33.8%), subarachnoid hemorrhages (71.5%, SD 36.5%), and complex hemorrhages (76.4%, SD 32.9%) were also relatively high, whereas it was very low for chronic subdural hemorrhages (37.3%, SD 37.5%; Figure 3B). Nevertheless, the misidentification percentages for complex hemorrhages (54.0%, SD 28.0%), epidural hemorrhages (50.2%, SD 22.7%), and subarachnoid hemorrhages (50.5%, SD 29.2%) were high (>50%), whereas they were relatively low (<40%) for acute subdural hemorrhages (32.6%, SD 26.3%), chronic subdural hemorrhages (40.3%, SD 27.2%), and intraparenchymal hemorrhages (26.2%, SD 23.8%; Figure 3C). These results suggest that GPT-4 has a good capability to recognize some types of acute hemorrhages but lacks the ability to comprehensively identify chronic hemorrhages.

Further, we defined identification completeness percentages of >90%, from 50% to 90%, and ≤50% as high, moderate, and low identification, respectively. For epidural hemorrhages, the proportion of high identification reached 66% (21/32), and the proportion of moderate to high identification reached 91% (29/32), which further illustrated GPT-4's strong capability in identifying epidural hemorrhages (Figure 3D). For intraparenchymal hemorrhages, the proportion of high identification was 48% (15/31), and the proportion of moderate to high identification was 94% (29/31), indicating that GPT-4 also had a good ability to recognize intraparenchymal hemorrhages (Figure 3D). In the cases of acute subdural hemorrhages and complex hemorrhages, approximately one-third of each set of images were categorized into high (12/42, 29% and 13/37, 37%, respectively), moderate (17/42, 40% and 14/37, 38%), and low (13/42, 31% and 10/37, 27%) identification; for subarachnoid hemorrhages, there was a substantial polarization (high identification at 21/36, 58% and low identification at 12/36, 33%), suggesting that GPT-4's recognition of these 3 types of hemorrhages was not stable (Figure 3D). Finally, in chronic subdural hemorrhages, the proportion of low identification reached up to 73% (22/30), further showing GPT-4's challenges in accurately identifying chronic subdural hemorrhages (Figure 3D).

XSL•FO
RenderX

Figure 3. Identification completeness and misidentification in 6 types of cerebral hemorrhages. (A) Representative images of 6 types of cerebral hemorrhages, separated into the massive and minor bleeding groups using a benchmark of 5 mL for the volume of bleeding. (B and C) Identification completeness and misidentification percentage of different types of the cerebral hemorrhages (mean and SD); **P*<.05, ***P*<.01, ****P*<.001, and *****P*<.0001 by 1-way ANOVA with Tukey multiple comparison test. (D) The distribution of different identification completeness percentage groups in 6 types of cerebral hemorrhages.

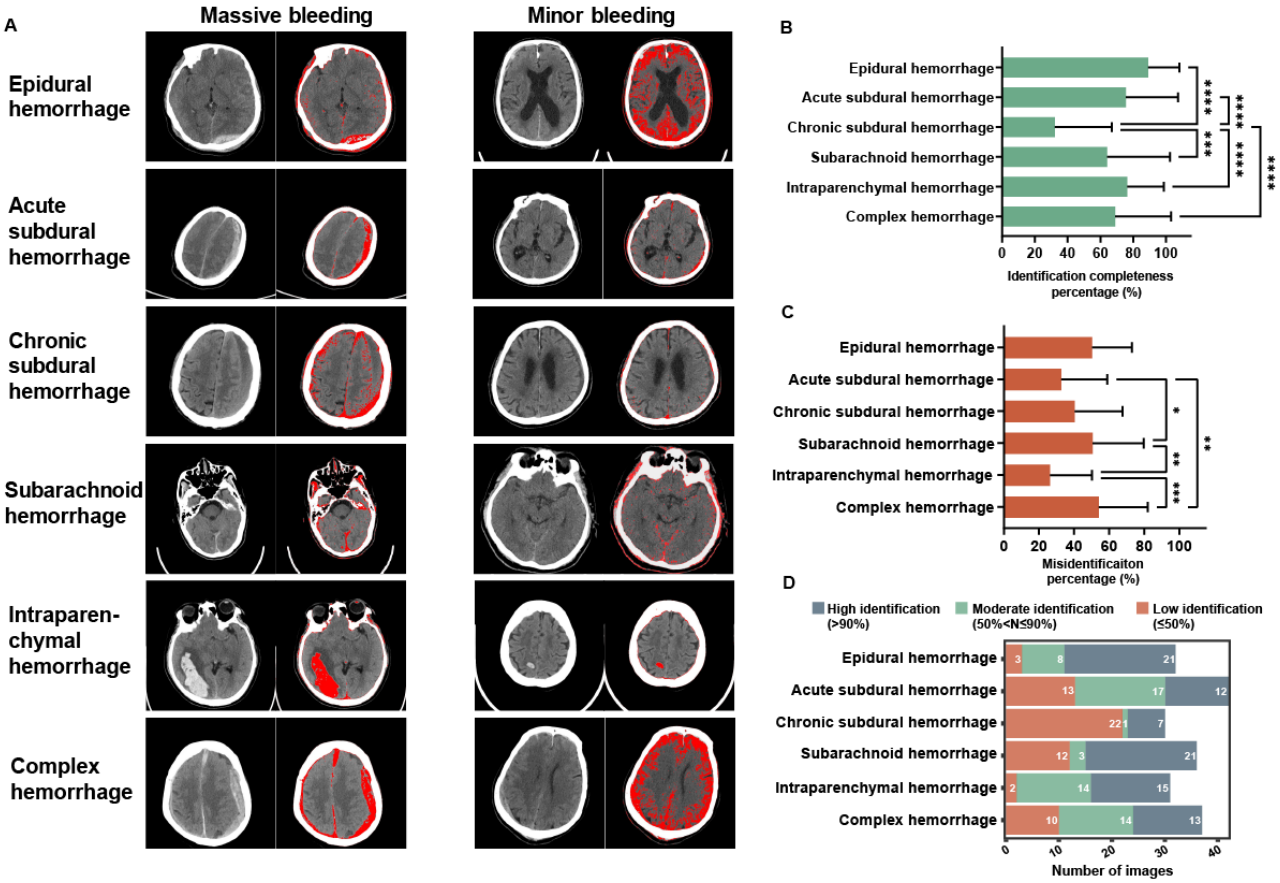

We classified cerebral hemorrhages into massive and minor bleeding, using a benchmark of 5 mL for the volume of bleeding. Surprisingly, there was no significant difference in the identification completeness percentages by GPT-4 between massive and minor bleeding (*P*=.06), indicating that the bleeding volume, within a certain range, did not affect GPT-4's ability to identify cerebral hemorrhages (Figure 4A). However, the misidentification percentage was higher for minor bleeding than for massive bleeding, suggesting that GPT-4 was more adept at recognizing substantial cerebral hemorrhages (*P*=.04; Figure 4B).

XSL•FO
RenderX

Figure 4. Identification completeness and misidentification percentage in the massive and minor bleeding groups. Cerebral hemorrhages were classified into massive and minor bleeding, using a benchmark of 5 mL for the volume of bleeding (mean and SD); *P<.05 and ns=P>.05 by 2-tailed t test. ns: no significance.

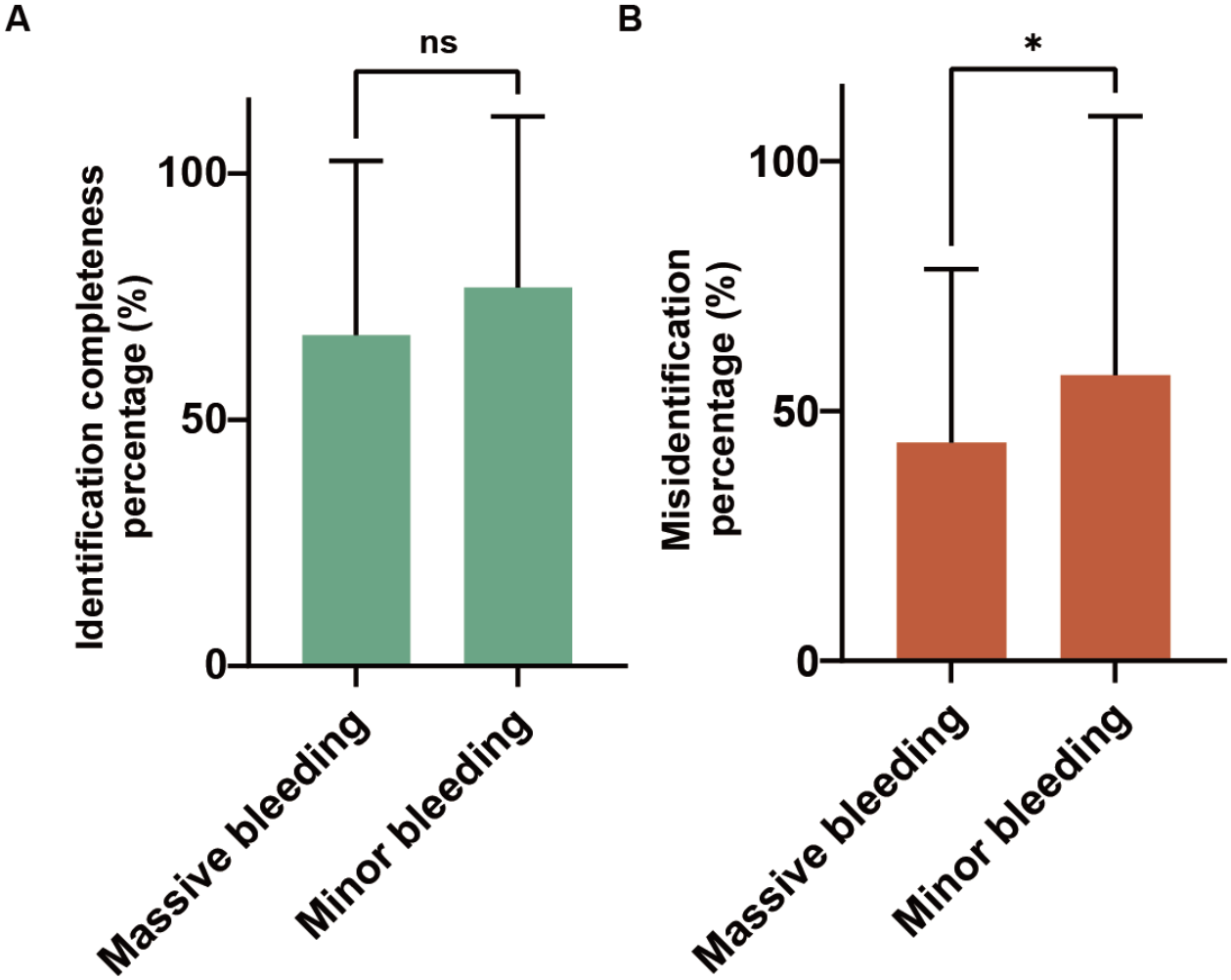

When we further performed analysis based on the location of hemorrhage within the brain, it was observed that GPT-4 demonstrated relatively higher identification completeness percentages and lower misidentification percentages in identifying hemorrhages in the cerebral ventricles and basal ganglia (Figure 5A and B). Conversely, the identification completeness percentage was comparatively lower for recognizing hemorrhages in the cerebral cisterns, and there was a higher misidentification percentage for hemorrhages in the occipitofrontal regions (Figure 5A and B). However, overall, there were no significant differences in the identification completeness percentage and misidentification percentage among different locations (all P>.05; Figure 5A and B). This indicated that GPT-4 did not show a marked preference for hemorrhages in specific areas but was relatively more proficient at identifying hemorrhages within the cerebral ventricles and basal ganglia.

XSL•FO
RenderX

Figure 5. Identification completeness and misidentification percentage in the different locations of hemorrhage within the brain. Cerebral hemorrhages were classified according to different hemorrhagic locations (mean and SD); by 1-way ANOVA with Tukey multiple comparison test.

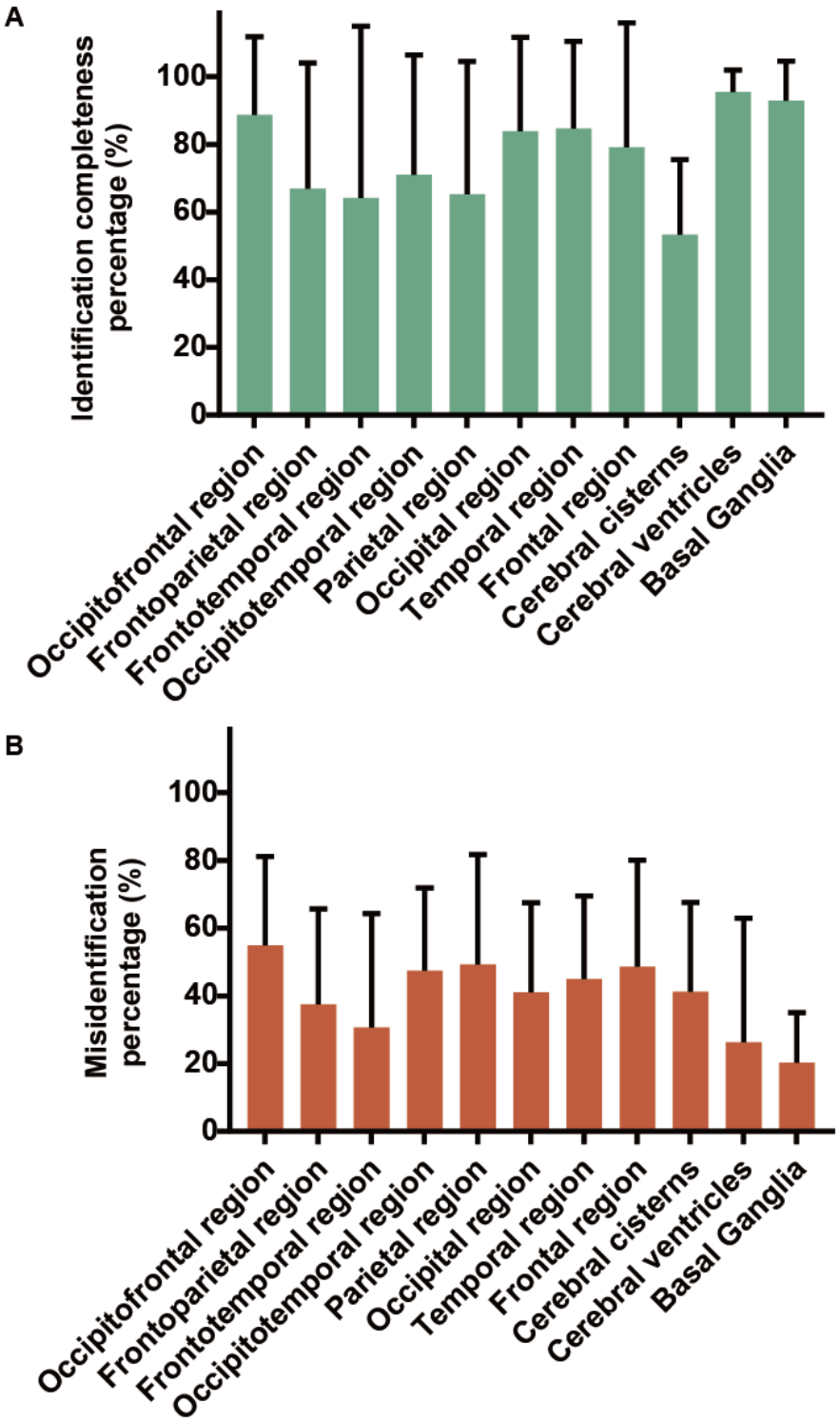

Based on the aforementioned results, we examined how radiologists viewed the performance of GPT-4 in identifying cerebral hemorrhages. Overall, radiologists expressed relative acceptance in terms of identification completeness (3.60, SD 0.54), accuracy (3.30, SD 0.65), and success ((3.38, SD 0.64; Figure 6A). Specifically, 89.9% (187/208) of CT images were accepted in terms of identification completeness; 66.3% (138/208) of CT images were accepted in terms of identification accuracy; and 75.5% (157/208) of CT images were accepted in terms of identification success (Figure 6B). Moreover, at the same level of identification success, the average rating for identification completeness was higher than that for identification accuracy, indicating that radiologists placed more importance on GPT-4's ability to fully identify hemorrhagic lesions and relative tolerance for false positives (Figure 6C). Generally, whether it was in terms of identification

XSL•FO
RenderX

completeness, accuracy, or success, the majority of CT scans were only deemed "relatively acceptable" (from 103/208, 49.5% to 111/208, 53.4%; Figure 6B). However, the proportion categorized as "completely unacceptable" was also very low (from 31/208, 14.9% to 76/208, 36.5%; Figure 6B). Further analysis was conducted based on different types of hemorrhages. The identification completeness, accuracy, and success of intraparenchymal hemorrhages and acute subdural hemorrhages received the highest acceptance from radiologists. Chronic subdural hemorrhages and subarachnoid hemorrhages had the lowest level of acceptance, with the identification completeness nearly reaching "relatively unacceptable" and both the identification accuracy and success being rated as "relatively unacceptable" (Figure 6D).

Figure 6. Evaluation on the CT identification by 10 radiologists. (A) The rating of the identification completeness, accuracy, and success (mean and SD). (B) The distribution of rating from identification completeness, accuracy and success. (C) The correlation between the identification completeness (green) or accuracy (red) and success. (D) The rating of identification completeness, accuracy, and success in 6 types of cerebral hemorrhages (mean and SD); *P<.05, **P<.01, ***P<.001, and ****P<.0001 by 1-way ANOVA with Tukey multiple comparison test. CT: computed tomography.

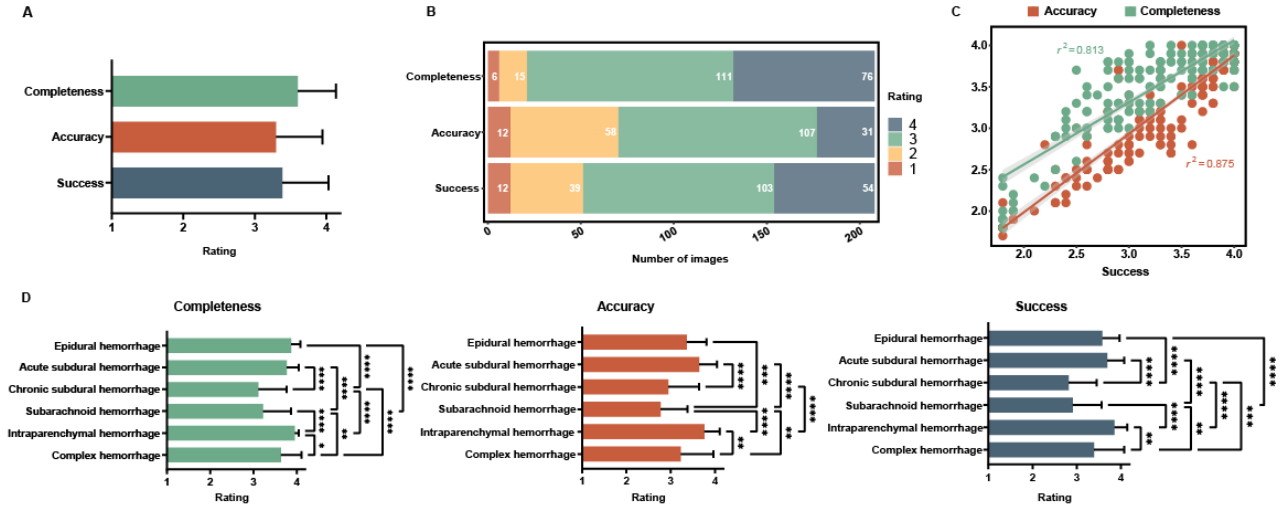

Discussion

In this study, we explored the possibility of applying GPT-4 to identify cerebral hemorrhages directly from cranial CT images, implying that NLP models like GPT-4 could be of clinical and commercial value in practice.

Overall, GPT-4 can identify most cerebral hemorrhage CT images after simple language-based training, achieving an overall identification completeness percentage of 72.6% (SD 18.6%). If chronic subdural hemorrhage is not considered, then the identification completeness percentage improves to 79.6% (SD 7.8%). This is significantly higher than our initial expectations prior to conducting this study, and radiologists have expressed an acceptable attitude toward this outcome (Figure 6A).

Specifically, in terms of different types of cerebral hemorrhages, GPT-4 demonstrates relatively high identification completeness percentages for epidural and intraparenchymal hemorrhages, with 89.0% (SD 19.1%) and 86.9% (SD 17.7%), respectively (Figure 3B). However, its identification completeness percentage for chronic subdural hemorrhage is very low, at only 37.3% (SD 37.5%; Figure 3B). The reason for this is that acute hematomas have a higher CT value compared to normal brain tissue, making the hematoma on CT images appear "whiter," thus easier for GPT-4 to differentiate from normal brain tissue. In contrast, chronic hematomas have CT values closer to that of normal brain tissue, making it challenging for GPT-4 to identify them based on image brightness variations. Radiologists diagnose chronic subdural hematomas more on the basis of changes in brain tissue symmetry, hematoma shape, and the absence of brain sulci and gyri structures in the hematoma, rather than just changes in CT values [31]. These are aspects that the current version of GPT-4 struggles to comprehend. Additionally, this may also be related to the inadequacy of the prompts provided to GPT-4. We attempted to instruct GPT-4 to recognize changes in the symmetry of brain tissue structures, but this resulted in even poorer identification completeness for chronic subdural hemorrhage (data not included). The current version of GPT-4 may not yet fully grasp human anatomical structures and their alterations. This suggests that if NLP models are to be applied in clinical scenarios, they may need enhanced training in fundamental medical knowledge.

In terms of identification accuracy, GPT-4's performance was not consistent. Compared to minor bleeding, GPT-4 was relatively more accurate in identifying massive bleeding (Figure 4B). This might be because it is challenging for GPT-4 to distinguish smaller hematomas from normal brain tissue due to their limited size. GPT-4 demonstrated relatively greater precision in recognizing hematomas in the cerebral ventricles and basal ganglia, possibly because these hematomas were located in the central part of the cranial CT scans, where their grayscale values were less likely to be negatively affected by the skull bones (Figure 5B). Therefore, for larger and relatively isolated hemorrhagic lesions, GPT-4 can identify them with more accuracy.

However, the current iteration of GPT-4 is not yet ready for clinical application and has the following limitations. First, the identification completeness and accuracy for some types of intracranial hemorrhages are relatively low. This directly prevents GPT-4 from being applied in clinical settings. Second,

XSL•FO
RenderX

at present, GPT-4 can only recognize a single CT image plane and cannot recognize a series of continuous CT images. The specific choice of which image plane to use still requires a doctor's experience to decide. Third, a simple language-based training is needed before using GPT-4; therefore, inappropriate prompt words can greatly affect the recognition effectiveness of GPT-4. Fourth, the use of GPT-4 to identify intracranial hemorrhage also involves ethical and legal issues, which may not be universally accepted.

Nevertheless, we believe that the future trend will see AI assisting clinicians and enhancing efficiency in their work. Imagine scenarios like this: in hospitals lacking radiologists, assistants without medical background could upload patients' head CT images to GPT-4 for preliminary screening; young radiologists could use GPT-4 to analyze cerebral hemorrhage CT images, helping to avoid missed diagnoses; and in extraordinary situations like wars, earthquakes, or maritime disasters, even laypeople without medical training would not be perplexed when facing head CT scans. From the evaluation of 10 radiologists, it appears that they are more concerned about GPT-4's ability to identify hematomas, showing more tolerance toward misidentifications (Figure 6C). After all, for NLP models, not missing a hematoma holds greater clinical significance. In addition to clinical work, GPT-4 could potentially play a more significant role in medical education. In the traditional learning of CT or other radiologic imaging, medical students typically only receive images and the correct answers. Identifying the exact part of the image that represents the correct answer often leaves students confused. This approach focuses more on the result than the learning process. However, GPT-4 can assist in identifying abnormalities in images, and we can even envisage a future where GPT-4 could provide the correct diagnosis step by step. This shift would make medical education more process oriented and efficient.

In conclusion, this study represents the first exploration of applying GPT-4 to the identification of cerebral hemorrhages in cranial CT images, as well as the first attempt to utilize GPT-4 in the identification of radiological images. The current version of GPT-4 is capable of identifying a small portion of cerebral hemorrhages, indicating that it has a foundation for computer-assisted diagnosis. We believe that in the future, NLP models like GPT-4 will demonstrate immense clinical and commercial value.

Acknowledgments

This study was supported by grants from the Program of Shanghai Academic Research Leader (21XD1422400 to JF), Shanghai Municipal Education Commission-Gaofeng Clinical Medicine Support (02.101005.001.30.30A to JF), and the Project of Shanghai Medical And Health Development Foundation (20224Z0012 to JF).

We extend our heartfelt gratitude to the other 8 dedicated radiologists from the Department of Radiology at Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, and Minhang Hospital, Fudan University. Their invaluable participation in evaluation significantly contributed to the depth and breadth of this study.

Authors' Contributions

Conceptualization – WW, JF
Data curation – DZ, ZM, RG
Formal analysis – DZ, ZM, RG, ZH
Funding acquisition – JF
Investigation – DZ, LL, YL
Methodology – DZ, ZM, RG, LL, YL, J Huang, WW, JF
Supervision – WW, JF
Validation – ZM, RG, YH, J Hui
Writing – original draft – DZ, JJ, WW, JF
Writing – review & editing – JJ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Representative steps of computed tomography (CT) images inputted into GPT-4. (A) The initial natural language prompt is inputted with the CT image into GPT-4, assisting GPT-4 to inspect the original picture and give the annotation back. Users' feedbacks help GPT-4 successfully annotate the bleeding area precisely. (B) The training CT image.
[PNG File , 638 KB - jmir_v26i1e58741_app1.png]

Multimedia Appendix 2
Criteria and categories used for rating of annotated images by radiologists.
[DOCX File , 24 KB - jmir_v26i1e58741_app2.docx]

XSL•FO
RenderX

References

1. Stocchetti N, Zanier ER. Chronic impact of traumatic brain injury on outcome and quality of life: a narrative review. Crit Care 2016 Jun 21;20(1):148 [FREE Full text] [doi: 10.1186/s13054-016-1318-1] [Medline: 27323708]
2. Bayen E, Jourdan C, Ghout I, Darnoux E, Azerad S, Vallat-Azouvi C, et al. Objective and subjective burden of informal caregivers 4 years after a severe traumatic brain injury: results from the PariS-TBI study. J Head Trauma Rehabil 2016;31(5):E59-E67. [doi: 10.1097/HTR.0000000000000079] [Medline: 24992640]
3. Humphreys I, Wood RL, Phillips CJ, Macey S. The costs of traumatic brain injury: a literature review. Clinicoecon Outcomes Res 2013 Jun 26;5:281-287 [FREE Full text] [doi: 10.2147/CEOR.S44625] [Medline: 23836998]
4. Lee WJ, Shah Y, Ku A, Patel N, Salvador M. Evaluating health disparities in radiology practices in New Jersey: exploring radiologist geographical distribution. Cureus 2023 Aug;15(8):e43474 [FREE Full text] [doi: 10.7759/cureus.43474] [Medline: 37583547]
5. Zhang J, Han X, Yang Z, Wang Z, Zheng J, Yang Z, et al. Radiology residency training in China: results from the first retrospective nationwide survey. Insights Imaging 2021 Feb 17;12(1):25 [FREE Full text] [doi: 10.1186/s13244-021-00970-2] [Medline: 33595737]
6. Bruls RJM, Kwee RM. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights Imaging 2020 Nov 23;11(1):121 [FREE Full text] [doi: 10.1186/s13244-020-00925-z] [Medline: 33226490]
7. Wang YE, Liu M, Jin L, Lungren MP, Grimm LJ, Zhang Z, et al. Radiology education in China. J Am Coll Radiol 2013 Mar;10(3):213-219. [doi: 10.1016/j.jacr.2012.11.006] [Medline: 23571062]
8. Mohsen H, El-Dahshan ESA, El-Horbaty ESM, Salem ABM. Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal 2018 Jun;3(1):68-71. [doi: 10.1016/j.fcij.2017.12.001]
9. Fernando T, Gammulle H, Denman S, Sridharan S, Fookes C. Deep learning for medical anomaly detection – a survey. ACM Comput Surv 2021 Jul 18;54(7):1-37. [doi: 10.1145/3464423]
10. Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. J Digit Imaging 2019 Aug 29;32(4):582-596 [FREE Full text] [doi: 10.1007/s10278-019-00227-x] [Medline: 31144149]
11. Guo J, Li B. The application of medical artificial intelligence technology in rural areas of developing countries. Health Equity 2018 Aug;2(1):174-181 [FREE Full text] [doi: 10.1089/heq.2018.0037] [Medline: 30283865]
12. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
13. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in medical education: global situation, effects and challenges. Educ Inf Technol 2023 Jul 10;29(4):4611-4633. [doi: 10.1007/s10639-023-12009-8]
14. Duong MT, Rauschecker AM, Rudie JD, Chen P, Cook TS, Bryan RN, et al. Artificial intelligence for precision education in radiology. Br J Radiol 2019 Nov;92(1103):20190389 [FREE Full text] [doi: 10.1259/bjr.20190389] [Medline: 31322909]
15. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV p. 779-788. [doi: 10.1109/cvpr.2016.91]
16. Ertuğrul Ö, Akıl MF. Detecting hemorrhage types and bounding box of hemorrhage by deep learning. Biomedical Signal Processing and Control 2022 Jan;71:103085. [doi: 10.1016/j.bspc.2021.103085]
17. Lee H, Huang C, Yune S, Tajmir SH, Kim M, Do S. Machine friendly machine learning: interpretation of computed tomography without image reconstruction. Sci Rep 2019 Oct 29;9(1):15540 [FREE Full text] [doi: 10.1038/s41598-019-51779-5] [Medline: 31664075]
18. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2024-06-10]
19. GPT-4. OpenAI. URL: https://openai.com/gpt-4 [accessed 2024-06-10]
20. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Mar 03;55(12):1-38. [doi: 10.1145/3571730]
21. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res 2024 May 22;26:e53164 [FREE Full text] [doi: 10.2196/53164] [Medline: 38776130]
22. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023 May 01;307(4):e230725. [doi: 10.1148/radiol.230725] [Medline: 37014240]
23. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology 2023 Sep;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
24. Santana LDM, Floresta L, Alves, dos Santos MAL, Barbosa BF, de Vasconcellos S, et al. Can GPT-4 be a viable alternative for discussing complex cases in digital oral radiology? a critical analysis. EXCLI J 2023 Aug 1;22:749-751 [FREE Full text] [doi: 10.17179/excli2023-6373] [Medline: 37662708]
25. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997 [FREE Full text] [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]

XSL•FO
RenderX

26. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023 Nov 17;13(1):20159 [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
27. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 2024 Jan;66(1):73-79. [doi: 10.1007/s00234-023-03252-4] [Medline: 37994939]
28. Tang XN, Berman AE, Swanson RA, Yenari MA. Digitally quantifying cerebral hemorrhage using Photoshop and Image J. J Neurosci Methods 2010 Jul 15;190(2):240-243 [FREE Full text] [doi: 10.1016/j.jneumeth.2010.05.004] [Medline: 20452374]
29. Kothari RU, Brott T, Broderick JP, Barsan WG, Sauerbeck LR, Zuccarello M, et al. The ABCs of measuring intracerebral hemorrhage volumes. Stroke 1996 Aug;27(8):1304-1305. [doi: 10.1161/01.str.27.8.1304] [Medline: 8711791]
30. Webb AJ, Ullman NL, Morgan TC, Muschelli J, Kornbluth J, Awad IA, MISTIECLEAR Investigators. Accuracy of the ABC/2 score for intracerebral hemorrhage: systematic review and analysis of MISTIE, CLEAR-IVH, and CLEAR III. Stroke 2015 Sep;46(9):2470-2476 [FREE Full text] [doi: 10.1161/STROKEAHA.114.007343] [Medline: 26243227]
31. Yadav Y, Parihar V, Namdev H, Bajaj J. Chronic subdural hematoma. Asian J Neurosurg 2016 Sep 20;11(4):330-342 [FREE Full text] [doi: 10.4103/1793-5482.145102] [Medline: 27695533]

Abbreviations

AI: artificial intelligence
CT: computed tomography
NLP: natural language processing

Edited by G Eysenbach; submitted 23.03.24; peer-reviewed by W Wu, J Mistry, A Wani, D Ghosh; comments to author 24.04.24; revised version received 19.06.24; accepted 31.08.24; published 26.09.24.

Please cite as:
Zhang D, Ma Z, Gong R, Lian L, Li Y, He Z, Han Y, Hui J, Huang J, Jiang J, Weng W, Feng J
Using Natural Language Processing (GPT-4) for Computed Tomography Image Analysis of Cerebral Hemorrhages in Radiology: Retrospective Analysis
J Med Internet Res 2024;26:e58741
URL: https://www.jmir.org/2024/1/e58741
doi:10.2196/58741
PMID:

©Daiwen Zhang, Zixuan Ma, Ru Gong, Liangliang Lian, Yanzhuo Li, Zhenghui He, Yuhan Han, Jiyuan Hui, Jialin Huang, Jiyao Jiang, Weiji Weng, Junfeng Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancing Performance of the National Field Triage Guidelines Using Machine Learning: Development of a Prehospital Triage Model to Predict Severe Trauma

Qi Chen[1*], PhD; Yuchen Qin[1*], PhD; Zhichao Jin[1*], PhD; Xinxin Zhao[2], PhD; Jia He[1], PhD; Cheng Wu[1*], PhD; Bihan Tang[3], PhD

[1]Department of Health Statistics, Naval Medical University, Shanghai, China
[2]School of Medicine, Tongji University, Shanghai, China
[3]Department of Health Management, Naval Medical University, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Bihan Tang, PhD
Department of Health Management
Naval Medical University
No 800 Xiangyin Road
Shanghai, 200433
China
Phone: 86 02181871425
Fax: 86 02181871425
Email: mangotangbihan@126.com

Abstract

Background: Prehospital trauma triage is essential to get the right patient to the right hospital. However, the national field triage guidelines proposed by the American College of Surgeons have proven to be relatively insensitive when identifying severe traumas.

Objective: This study aimed to build a prehospital triage model to predict severe trauma and enhance the performance of the national field triage guidelines.

Methods: This was a multisite prediction study, and the data were extracted from the National Trauma Data Bank between 2017 and 2019. All patients with injury, aged 16 years of age or older, and transported by ambulance from the injury scene to any trauma center were potentially eligible. The data were divided into training, internal, and external validation sets of 672,309; 288,134; and 508,703 patients, respectively. As the national field triage guidelines recommended, age, 7 vital signs, and 8 injury patterns at the prehospital stage were included as candidate variables for model development. Outcomes were severe trauma with an Injured Severity Score ≥16 (primary) and critical resource use within 24 hours of emergency department arrival (secondary). The triage model was developed using an extreme gradient boosting model and Shapley additive explanation analysis. The model's accuracy regarding discrimination, calibration, and clinical benefit was assessed.

Results: At a fixed specificity of 0.5, the model showed a sensitivity of 0.799 (95% CI 0.797-0.801), an undertriage rate of 0.080 (95% CI 0.079-0.081), and an overtriage rate of 0.743 (95% CI 0.742-0.743) for predicting severe trauma. The model showed a sensitivity of 0.774 (95% CI 0.772-0.776), an undertriage rate of 0.158 (95% CI 0.157-0.159), and an overtriage rate of 0.609 (95% CI 0.608-0.609) when predicting critical resource use, fixed at 0.5 specificity. The triage model's areas under the curve were 0.755 (95% CI 0.753-0.757) for severe trauma prediction and 0.736 (95% CI 0.734-0.737) for critical resource use prediction. The triage model's performance was better than those of the Glasgow Coma Score, Prehospital Index, revised trauma score, and the 2011 national field triage guidelines RED criteria. The model's performance was consistent in the 2 validation sets.

Conclusions: The prehospital triage model is promising for predicting severe trauma and achieving an undertriage rate of <10%. Moreover, machine learning enhances the performance of field triage guidelines.

(J Med Internet Res 2024;26:e58740) doi:10.2196/58740

XSL•FO
RenderX

KEYWORDS

severe trauma; field triage; machine learning; prediction model

Introduction

Trauma is a universal health challenge that places a massive burden on national economies. It causes 4.4 million deaths annually, and an estimated 10% of all years lived with disability [1,2]. The American College of Surgeons Committee on Trauma (ACS-COT) recommends that severe trauma be treated at levels 1 and 2 trauma care facilities [3]. Patients with severe trauma have approximately 25% lower mortality rates when treated at levels 1 or 2 trauma centers than when treated at lower-level or nontrauma centers [4]. Prehospital estimation of injury severity is essential for prehospital triage. It is a critical step for emergency medical service (EMS) providers in making decisions regarding patient destination. Under- and overtriage are incorrect triages. A low-sensitivity triage tool results in many false-negative cases, indicating undertriage and a possible failure in trauma first aid. Conversely, low specificity is associated with a high rate of false-positive cases, indicating overtriage [5].

The national field triage guidelines were initially developed by ACS-COT in 1976 and revised in 2011 and 2021 [6,7]. The national field triage guidelines have been widely implemented in the United States and represent 1 of the few standardized national protocols for EMS. It was developed based on peer-reviewed research, resulting in biased estimates and reduced generalizability [3,8]. A prospective national triage guidelines validation study for identifying high-risk trauma patients found that the guidelines were relatively insensitive in identifying severely injured patients and those requiring early critical resource use [9]. In addition, other triage tools, such as the Glasgow Coma Score (GCS), Prehospital Index (PHI), and revised trauma score (RTS), have not shown ideal predictive performance [5,10-12]. Therefore, it emphasizes the need to improve the prehospital triage tool [13].

Machine learning (ML) development has advanced rapidly in the medical field, notably in trauma medicine, and has demonstrated that the ML model's predictive ability is significantly better than that of the conventional trauma triage tools for mortality outcomes, hospitalization, and critical care admission [14,15]. Therefore, this study aimed to build a prehospital triage model to predict severe trauma (pTEST) and enhance the performance of the national field triage guideline.

Methods

Recruitment

This multisite prediction study was conducted to predict trauma severity during field triage. We developed, validated, and reported our triage model following the TRIPOD (Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis) statement [16], as shown in Multimedia Appendix 1.

Ethical Considerations

The Naval Medical University Ethics Committee approved the study protocol (reference number NMUEC2022-088). Our study consisted of secondary analyses using the National Trauma Data Bank (NTDB) with primary consent, and the data were anonymized.

Source of Data and Patients

Data from the NTDB, the largest aggregation of trauma registry data in the United States assembled by the ACS, were used in this study [17]. In 2017, the ACS Trauma Quality Program transitioned to a new technical vendor and redesigned the NTDB infrastructure. The 2017 and 2018 NTDB datasets comprising 2,041,706 patients were used for pTEST model development, hyperparameter tuning, and internal validation. The 2019 NTDB dataset comprising 1,097,190 patients was used for the external validation of the pTEST model.

Notably, all patients with injury, aged 16 years of age or older, and transported by ground or aerial ambulance from the injury scene to any trauma center were potentially eligible. Due to a lack of crucial information on outcomes and predictors, we excluded patients who died in the EMS, patients discharged from the emergency department (ED) to another hospital, and those without any EMS records or an Injury Severity Score (ISS). In total, 960,443 and 508,703 participants were included in the development and validation sets, respectively. Figure 1 shows the patient selection process.

XSL•FO
RenderX

Figure 1. The flowchart of model development and validation. ED: emergency department; EMS: emergency medical service; ISS: Injury Severity Score; NTDB: National Trauma Data Bank; pTEST: prehospital triage model to predict severe trauma.

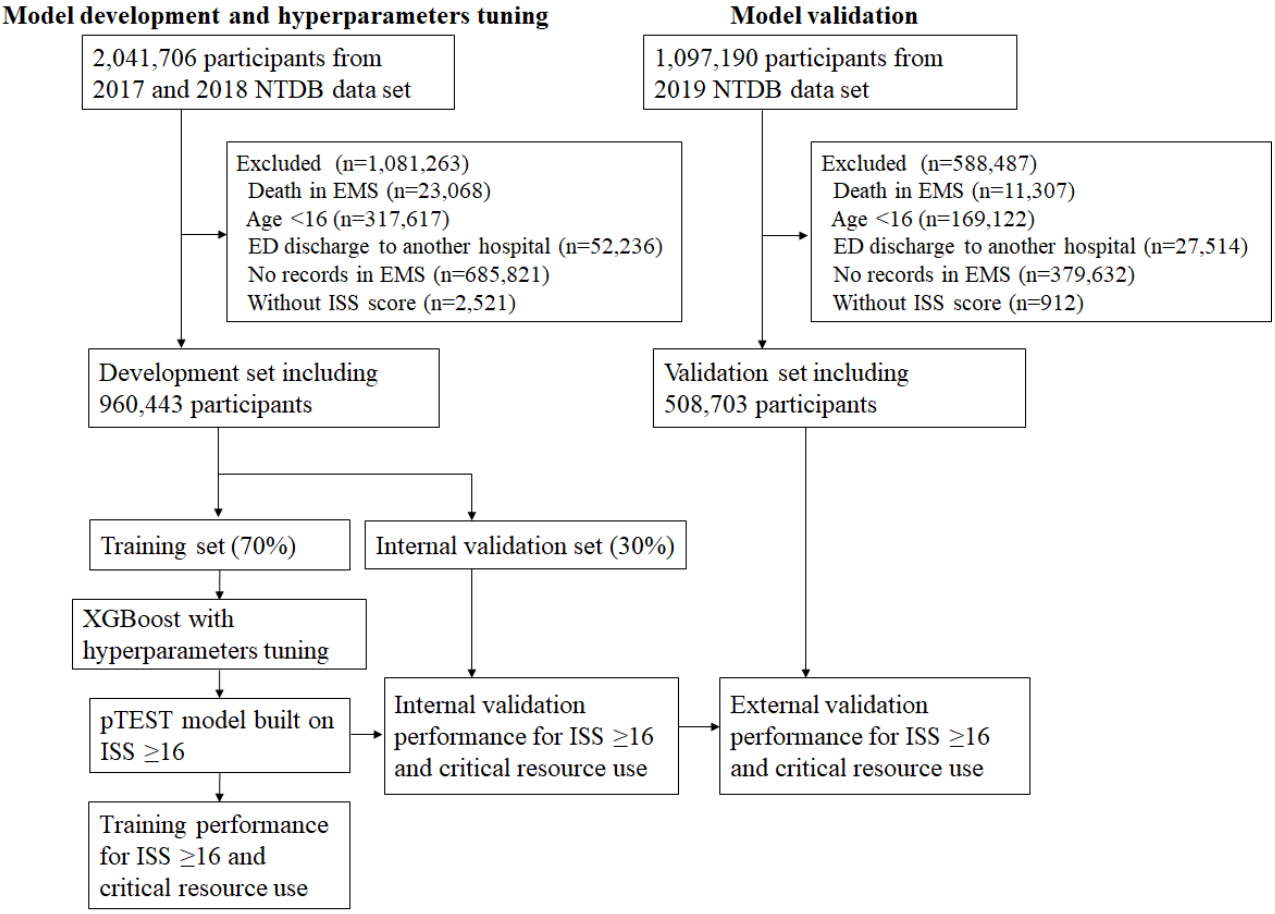

Outcome and Predictors

Defining "severe trauma" is a challenge in prehospital triage model development. It varies widely across studies. The reference standard (primary outcome) of "severe injury" was considered as an ISS ≥16 as the benchmark to evaluate triage accuracy recommended by the ACS [7,18,19]. The ISS calculated by anatomical criteria is assumed to be consistent with the patient status on the scene and is associated with high mortality [4,19]. However, it does not reflect resource use directly; therefore, we included a composite resource-based early critical resource use measure as the secondary outcome. According to similar studies [9,18], early critical resource use included intubation in the EMS or ED, discharge to the intensive care unit from the ED, surgery for hemorrhage control, interventional radiology procedures, cerebral monitoring, and in-hospital death, all within 24 hours. A detailed definition of severe trauma is provided in Multimedia Appendix 2.

According to a recent field triage protocol review in 2017 [10], the significant predictors of a severely injured patient were age, vital signs, injury patterns, and injury mechanism. In addition, in the US National Guidelines for the Field Triage of Injured Patients in 2011 and 2021 [6,7], age, vital signs, and injury pattern measurements were the field triage's top priorities. In the 2011 national field triage guidelines, severely injured patients who should be transported preferentially to the highest-level trauma center were identified using the RED criteria, which included 3 vital signs (GCS, systolic blood pressure, and respiratory rate) and 8 injury patterns. During field triage, time is essential, and the number and complexity of hand-collected variables must be limited. Therefore, as the US National Guideline for the Field Triage recommended and recorded in the NTDB, we incorporated 16 candidate variables in EMS for model development, including age at the time of injury (AGEYEARS), GCS eye (EMSGCSEYE), GCS motor (EMSGCSMOTOR), GCS verbal (EMSGCSVERBAL), systolic blood pressure (EMSSBP), oxygen saturation (EMSPULSEOXIMETRY), respiratory rate (EMSRESPIRATORYRATE), pulse rate (EMSPULSERATE), penetrating injuries (TCCPEN), chest wall instability (TCCCHEST), long-bone fractures (TCCLONGBONE), crushed extremity (TCCCRUSHED), amputation (TCCAMPUTATION), pelvic fracture (TCCPELVIC), skull fracture (TCCSKULLFRACTURE), and paralysis (TCCPARALYSIS). The detailed definitions of candidate variables are listed in Multimedia Appendix 3.

Model Development

The pTEST model was developed using the extreme gradient boosting model (XGBoost) and Shapley additive explanation analysis (SHAP). XGBoost is a novel boosting tree-based ensemble algorithm through which new models are created to predict residuals or errors of prior models and then combined to make a final prediction [20]. Recently, XGBoost has been widely used in ML due to its outstanding prediction

XSL•FO
RenderX

performance, ability to use continuous and categorical inputs, lack of data preprocessing, imbalanced data handling capacity, high internal optimization, and relatively modest computational costs [21].

Patients in the development set from the 2017 and 2018 NTDB datasets were randomly grouped into training (70%) and internal validation (30%) sets for model development. Using a grid search, a 10-fold cross-validation process was used on the training set for hyperparameter tuning. The goal of hyperparameter tuning is to find the values that lead to the best-predicted performance. The optimal values of hyperparameters were learning rate=0.04071151, minimum loss reduction required to make a further partition=20.36485, maximum tree depth=14, minimum sum of instance weights needed in a leaf node=39, maximum number of boosting iterations=1051, and subsample ratio of the training instance=0.7763707, with the other hyperparameters set to default. In addition to the training set, the model's reproducibility, transportability, and generalizability were evaluated using internal and external validation sets.

Missing values are an essential concern in trauma triage because there may not always be time to measure critical variables. The absent proportions of the training, internal validation, and external validation sets are shown in Multimedia Appendix 4, with all variables, except pulse oximetry, missing below 6%. XGBoost supports branch directions for predictors with missing values, creating an advantage in real-world situations where XGBoost can still achieve individual prediction without complete prehospital data.

To gain insight into the risk prediction model, we investigated different predictors' contributions based on Shapley values using SHAP, a game theory concept introduced in the 1950s [22,23]. A predictor's SHAP value can be positive or negative, suggesting an increased or decreased probability of severe trauma, respectively. In our study, the SHAP values were visualized at global (dataset level) and local (patient-specific) levels to investigate the predictors' impact. XGBoost and SHAP were implemented using the R (R Core Team) packages *tidy models* and *shapviz*.

Statistical Analysis Methods

For the sample size calculation, the prevalence of events was set at 17.8%, and the number of predictors was 16 based on the development set. The area under the curve (AUC) of the optimal prehospital triage model in a previous study was 0.68 [11], and our pTEST model was expected to achieve an AUC of 0.7. At least 1871 patients were required for model development or validation using the R package *pmsampsize* [24].

Continuous data are presented as mean and SD, and categorical data are presented as frequencies and percentages (%). The 2-tailed *t* test was used to evaluate the differences in continuous data, which followed a normal distribution and variance homogeneity; otherwise, the Wilcoxon rank-sum test was used. The differences in categorical data were evaluated using the Pearson chi-square test. The area under the receiver operating characteristic curve was calculated to assess model discrimination. The AUCs between models were compared using the DeLong test. The best thresholds of the models were determined by maximizing the Youden index, and performance metrics, including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV), were calculated. Performance metrics 95% CIs were calculated using 500 bootstrap replicates. In addition, the pTEST model is intended to identify severe trauma that requires high sensitivity and NPV to rule it out. Sensitivity and specificity are inversely proportional and a tradeoff needs to be made between sensitivity and specificity. Therefore, the sensitivities and NPV of the different models were compared using a 0.5 specificity as in the previous studies [25,26].

Our study defined the over- and undertriage rates. Overtriage rate (1-PPV) is equal to the number of patients with negative outcomes (ISS <16 or no critical resource use) predicted as positive outcomes divided by the total number of predicted positive outcomes. The undertriage rate (1-NPV) is equal to the number of patients with positive outcomes predicted as negative outcomes divided by the total number of predicted negative outcomes.

In addition, a calibration plot and Brier score were generated to assess how closely the predicted probability approximated the actual probability. The clinical benefit of the models was evaluated using a decision curve analysis method. The discrimination, calibration, and clinical benefit of the pTEST were compared with the GCS, PHI, RTS, and RED criteria of the 2011 National Field Triage Guidelines. Statistical significance was set at $P<.05$. All statistical analyses were performed using the R software (version 4.3.1).

Results

Patient Characteristics

Table 1 shows the patients' baseline characteristics in the training, internal validation, and external validation sets. In these 3 sets, severe trauma proportions were 17.80% (119,690/672,309), 17.80% (51,296/288,134), and 17.08% (86,902/508,703), respectively, and critical resource use proportions were 29.36% (177,570/604,806), 29.56% (76,604/259,148), and 28.17% (129,551/459,843), respectively. Notably, most variables showed statistically significant differences among the 3 sets for large sample sizes, but the differences were minimal. The demographic characteristics, vital signs, and injury patterns of nonsevere and severe trauma are listed in Multimedia Appendices 5-7, and the attributes of noncritical and critical resource users are listed in Multimedia Appendices 8-10. Severe trauma and critical resource users are usually male, air-transported, taken to higher-level trauma centers, and have extreme trauma patterns.

XSL•FO
RenderX

Table 1. Baseline characteristics of the patients from the training set, internal validation set, and external validation set.

Characteristics	Training set (n=672,309)	Internal validation set (n=288,134)	External validation set (n=508,703)	*P* value
Sex (male), n (%)	403,434 (60.01)	172,821 (59.98)	300,711 (59.12)	<.001
Transport mode, n (%)				<.001
Ground	622,489 (92.59)	266,622 (92.53)	474,645 (93.30)	
Helicopter	48,464 (7.21)	20,911 (7.26)	33,346 (6.56)	
Fixed-wing	1356 (0.20)	601 (0.21)	712 (0.14)	
Trauma center level, n (%)				<.001
Level 1	275,723 (55.75)	118,199 (55.72)	206,563 (54.39)	
Level 2	179,419 (36.28)	77,005 (36.30)	142,433 (37.50)	
Level 3	39,397 (7.97)	16,944 (7.99)	30,819 (8.11)	
TCCPEN[a] (yes), n (%)	26,987 (4.01)	11,319 (3.93)	17,658 (3.47)	<.001
TCCCHEST[b] (yes), n (%)	4295 (0.64)	1825 (0.63)	3216 (0.63)	.89
TCCLONGBONE[c] (yes), n (%)	4598 (0.68)	1867 (0.65)	3213 (0.63)	.002
TCCCRUSHED[d] (yes), n (%)	3188 (0.47)	1374 (0.48)	2712 (0.53)	<.001
TCCAMPUTATION[e] (yes), n (%)	832 (0.12)	372 (0.13)	613 (0.12)	.58
TCCPELVIC[f] (yes), n (%)	7495 (1.11)	3346 (1.16)	5852 (1.15)	.07
TCCSKULLFRACTURE[g] (yes), n (%)	5127 (0.76)	2237 (0.78)	4087 (0.80)	.04
TCCPARALYSIS[h] (yes), n (%)	4379 (0.65)	1887 (0.65)	3077 (0.60)	.003
ISS[i] score (≥16), n (%)	119,690 (17.80)	51,296 (17.80)	86,902 (17.08)	<.001
Surgery for hemorrhage control (yes), n (%)	14,714 (2.45)	6441 (2.50)	10,886 (2.37)	.002
Cerebral monitor (yes), n (%)	8527 (1.42)	3563 (1.38)	5819 (1.27)	<.001
Interventional radiology procedures (yes), n (%)	5154 (0.86)	2182 (0.85)	3848 (0.84)	.59
Discharge to the ICU[j] from ED[k] (yes), n (%)	142,422 (21.46)	61,384 (21.58)	105,670 (21.06)	<.001
In-hospital death within 24 hours (yes), n (%)	9303 (1.38)	3987 (1.38)	6371 (1.25)	<.001
Intubation in the EMS[l] or ED (yes), n (%)	72,755 (10.82)	31,080 (10.79)	49,779 (9.79)	<.001
Critical resource use (yes), n (%)	177,570 (29.36)	76,604 (29.56)	129,551 (28.17)	<.001
RED criteria (yes), n (%)	87,577 (13.03)	37,448 (13.00)	62,194 (12.23)	<.001
Age (years), mean (SD)	53.12 (21.86)	53.22 (21.84)	54.40 (21.77)	<.001
EMSSBP[m] (mm Hg), mean (SD)	139.89 (28.37)	139.92 (28.42)	140.74 (28.62)	<.001
EMSPULSERATE[n] (n/minute), mean (SD)	90.58 (20.31)	90.58 (20.29)	90.34 (20.41)	<.001
EMSRESPIRATORYRATE[o] (n/minute), mean (SD)	18.42 (4.73)	18.44 (4.79)	18.46 (4.82)	.08
EMSPULSEOXIMETRY[p], mean (SD)	96.26 (5.50)	96.26 (5.48)	96.19 (5.44)	<.001
EMSGCSEYE[q], mean (SD)	3.81 (0.65)	3.81 (0.65)	3.82 (0.63)	<.001
EMSGCSVERBAL[r], mean (SD)	4.60 (0.95)	4.60 (0.95)	4.61 (0.93)	.02
EMSGCSMOTOR[s], mean (SD)	5.73 (0.97)	5.73 (0.97)	5.75 (0.94)	<.001
EMSTOTALGCS[t], mean (SD)	14.12 (2.44)	14.12 (2.43)	14.17 (2.36)	<.001
Total time spent in ED (minutes), mean (SD)	189.94 (150.35)	207.75 (400.72)	207.68 (1355.03)	<.001

XSL•FO
RenderX

Characteristics	Training set (n=672,309)	Internal validation set (n=288,134)	External validation set (n=508,703)	*P* value
Length of stay (days), mean (SD)	6.20 (9.67)	6.21 (8.51)	6.21 (8.38)	.001
ISS score, mean (SD)	9.69 (8.37)	9.69 (8.32)	9.55 (8.20)	<.001
PHI[u] score, mean (SD)	1.32 (2.23)	1.32 (2.23)	1.31 (2.22)	.26
RTS[v] score, mean (SD)	11.73 (0.88)	11.73 (0.89)	11.74 (0.86)	.045

[a]TCCPEN: penetrating injuries.
[b]TCCCHEST: chest wall instability.
[c]TCCLONGBONE: long-bone fractures.
[d]TCCCRUSHED: crushed extremity.
[e]TCCAMPUTATION: amputation.
[f]TCCPELVIC: pelvic fracture.
[g]TCCSKULLFRACTURE: skull fracture.
[h]TCCPARALYSIS: paralysis.
[i]ISS: Injury Severity Score.
[j]ICU: intensive care unit.
[k]ED: emergency department.
[l]EMS: emergency medical service.
[m]EMSSBP: systolic blood pressure.
[n]EMSPULSERATE: pulse rate.
[o]EMSRESPIRATORYRATE: respiratory rate.
[p]EMSPULSEOXIMETRY: oxygen saturation.
[q]EMSGCSEYE: Glasgow Coma Score eye.
[r]EMSGCSVERBAL: Glasgow Coma Score verbal.
[s]EMSGCSMOTOR: Glasgow Coma Score motor.
[t]EMSTOTALGCS: Glasgow Coma Score total.
[u]PHI: Prehospital Index.
[v]RTS: revised trauma score.

Model Performance

For predicting severe trauma, we compared the performance metrics of the other models at the same specificity fixed at a moderate number of 0.5. The pTEST model showed a higher sensitivity of 0.799 (95% CI 0.797-0.801), a lower undertriage rate of 0.080 (95% CI 0.079-0.081), and a lower overtriage rate of 0.743 (95% CI 0.742-0.743) in the training set (Table 2). In addition, for critical resource use prediction fixed at a specificity of 0.5, the pTEST model showed a higher sensitivity of 0.774 (95% CI 0.772-0.776), lower undertriage rate of 0.158 (95% CI 0.157-0.159), and lower overtriage rate of 0.609 (95% CI 0.608-0.609) than the other models in the training set (Multimedia Appendix 11). We validated the pTEST model performance using 2 validation sets and obtained consistent results (Table 2 and Multimedia Appendix 11). The model performance metrics for predicting severe trauma and critical resource use at the best thresholds with the maximum Youden index are listed in Multimedia Appendices 12 and 13, demonstrating a higher pTEST Youden index than other models.

XSL•FO
RenderX

Table 2. Model performance metrics for predicting severe trauma fixed at a specificity of 0.5.

Prediction tool	Specificity	Sensitivity (95% CI)	Accuracy, AUC (95% CI)	Undertriage rate (1-NPV), AUC (95% CI)	Overtriage rate (1-PPV), AUC (95% CI)	Youden index
Training set						
pTEST[a]	0.500	0.799 (0.797-0.801)	0.553 (0.553-0.554)	0.080 (0.079-0.081)	0.743 (0.742-0.743)	1.299
GCS[b]	0.500	0.682 (0.680-0.684)	0.532 (0.532-0.532)	0.119 (0.119-0.120)	0.774 (0.774-0.775)	1.182
PHI[c]	0.500	0.711 (0.709-0.713)	0.536 (0.536-0.537)	0.107 (0.107-0.108)	0.772 (0.771-0.772)	1.211
RTS[d]	0.500	0.634 (0.632-0.636)	0.523 (0.523-0.523)	0.132 (0.132-0.133)	0.791 (0.791-0.792)	1.134
RED criteria	0.500	0.620 (0.619-0.621)	0.521 (0.521-0.522)	0.141 (0.141-0.142)	0.788 (0.788-0.789)	1.120
Internal validation set						
pTEST	0.500	0.794 (0.791-0.798)	0.552 (0.552-0.553)	0.082 (0.081-0.083)	0.744 (0.743-0.745)	1.294
GCS	0.500	0.682 (0.680-0.685)	0.532 (0.532-0.533)	0.120 (0.119-0.120)	0.774 (0.773-0.775)	1.182
PHI	0.500	0.710 (0.707-0.714)	0.536 (0.536-0.537)	0.108 (0.106-0.109)	0.772 (0.771-0.773)	1.210
RTS	0.500	0.633 (0.631-0.636)	0.523 (0.523-0.523)	0.132 (0.132-0.133)	0.791 (0.791-0.792)	1.133
RED criteria	0.500	0.620 (0.619-0.622)	0.521 (0.521-0.522)	0.141 (0.141-0.142)	0.788 (0.788-0.789)	1.120
External validation set						
pTEST	0.500	0.794 (0.792-0.797)	0.550 (0.550-0.551)	0.078 (0.077-0.079)	0.753 (0.753-0.754)	1.294
GCS	0.500	0.681 (0.679-0.683)	0.531 (0.530-0.531)	0.115 (0.115-0.116)	0.783 (0.782-0.783)	1.181
PHI	0.500	0.709 (0.706-0.711)	0.535 (0.534-0.535)	0.103 (0.103-0.104)	0.781 (0.780-0.782)	1.209
RTS	0.500	0.633 (0.631-0.635)	0.522 (0.522-0.522)	0.127 (0.126-0.128)	0.799 (0.799-0.800)	1.133
RED criteria	0.500	0.616 (0.615-0.618)	0.520 (0.520-0.520)	0.137 (0.136-0.137)	0.798 (0.797-0.798)	1.116

[a]pTEST: prehospital triage model to predict severe trauma.

[b]GCS: Glasgow Coma Score.

[c]PHI: Prehospital Index.

[d]RTS: revised trauma score.

In Figure 2, pTEST AUCs for severe trauma prediction were 0.755 (95% CI 0.753-0.757), 0.751 (95% CI 0.749-0.754), and 0.750 (95% CI 0.749-0.752) in training, internal validation, and external validation sets, respectively, and the AUCs for predicting critical resource use were 0.736 (95% CI 0.734-0.737), 0.732 (95% CI 0.730-0.734), and 0.733 (95% CI 0.732-0.735), respectively, demonstrating better discrimination ability than GCS, PHI, RTS, and RED criteria. Multimedia Appendix 14 depicts the pTEST model predicted outcome probability as a waterfall plot. The calibration curves in Multimedia Appendix 15 show that the severe trauma predicted probability and pTEST critical resource use agreed with the proportion observed using the smallest Brier score. In Multimedia Appendix 16, pTEST provides a consistently higher net benefit across a broad range of risk thresholds (10%-100%) than the 2 default strategies and other models.

XSL•FO
RenderX

Figure 2. The ROC curves of 5 models. (A) Predicting severe trauma in the training set. (B) Predicting severe trauma in the internal validation set. (C) Predicting severe trauma in the external validation set. (D) Predicting critical resource use in training set. (E) Predicting critical resource use in the internal validation set. (F) Predicting critical resource use in the external validation set. AUC: area under the curve; GCS: Glasgow Coma Score; PHI: Prehospital Index; pTEST: prehospital triage model to predict severe trauma; ROC: receiver operating characteristic; RTS: revised trauma score.

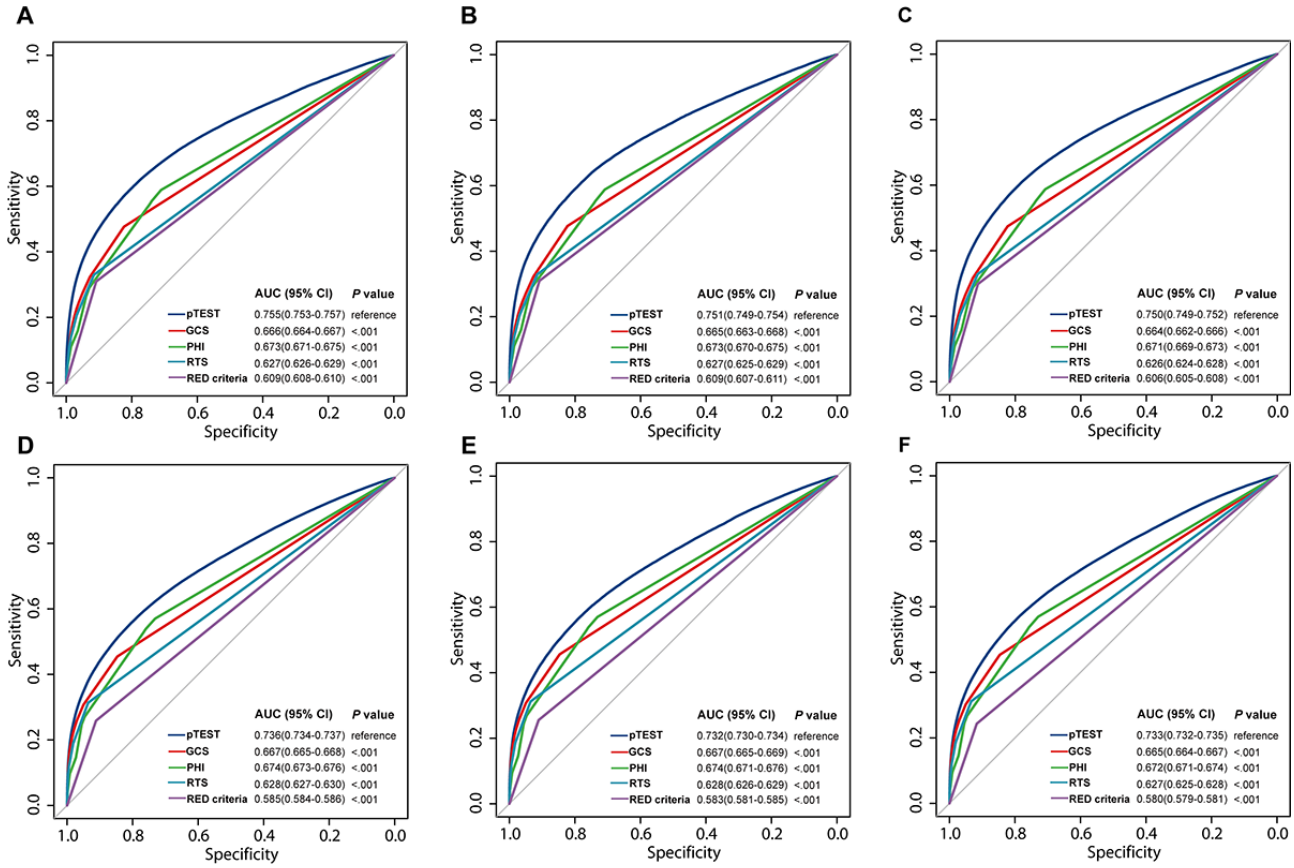

Model Interpretation

As shown in the SHAP summary plots (Figure 3A), the contributions of the variables to the pTEST model for severe trauma prediction were evaluated using the average absolute SHAP values; the top 5 important variables were EMSGCSVERBAL, EMSSBP, EMSRESPIRATORYRATE, EMSGCSMOTOR, and EMSPULSEOXIMETRY. Figure 3B lists the impact of the different variables illustrated by the SHAP values for severe trauma prediction. Figure 3C shows each variable's SHAP values versus measured values. Figure 3B, C shows that the higher the EMSGCSVERBAL score, the lower the probability of severe trauma ("negative" impact). Similarly, AGEYEARS, EMSGCSEYE, EMSGCSMOTOR, and EMSPULSEOXIMETRY negatively contributed to the predicted probability. In contrast, 8 injury patterns contributed positively to the predicted probability. The SHAP summary and dependence plots of the pTEST model for critical resource use prediction are shown in Multimedia Appendix 17. The personalized feature attributes for 2 representative patients with and without severe trauma in the training set are provided in Multimedia Appendix 18.

XSL•FO
RenderX

Figure 3. The global model explanation for predicting severe trauma by the SHAP method in the training set. (A) SHAP summary bar plot of the average SHAP value for each variable. (B) SHAP summary dot plot. In each variable, a dot is made for each single patient, representing the SHAP value of this variable. The colors of the dots demonstrate the actual values of the features, and the dots are stacked vertically to show density. (C) SHAP dependence plot. Each dependence plot shows the association between the actual value and the SHAP value of the variable, and each dot represents a single patient. AGEYEARS: age at the time of injury; EMSGCSEYE: Glasgow Coma Score eye; EMSGCSMOTOR: Glasgow Coma Score motor; EMSGCSVERBAL: Glasgow Coma Score verbal; EMSSBP: systolic blood pressure; EMSPULSEOXIMETRY: oxygen saturation; EMSRESPIRATORYRATE: respiratory rate; EMSPULSERATE: pulse rate; SHAP: Shapley additive explanation analysis; TCCPEN: penetrating injuries; TCCCHEST: chest wall instability; TCCLONGBONE: long-bone fractures; TCCCRUSHED: crushed extremity; TCCAMPUTATION: amputation; TCCPELVIC: pelvic fracture; TCCSKULLFRACTURE: skull fracture; TCCPARALYSIS: paralysis.

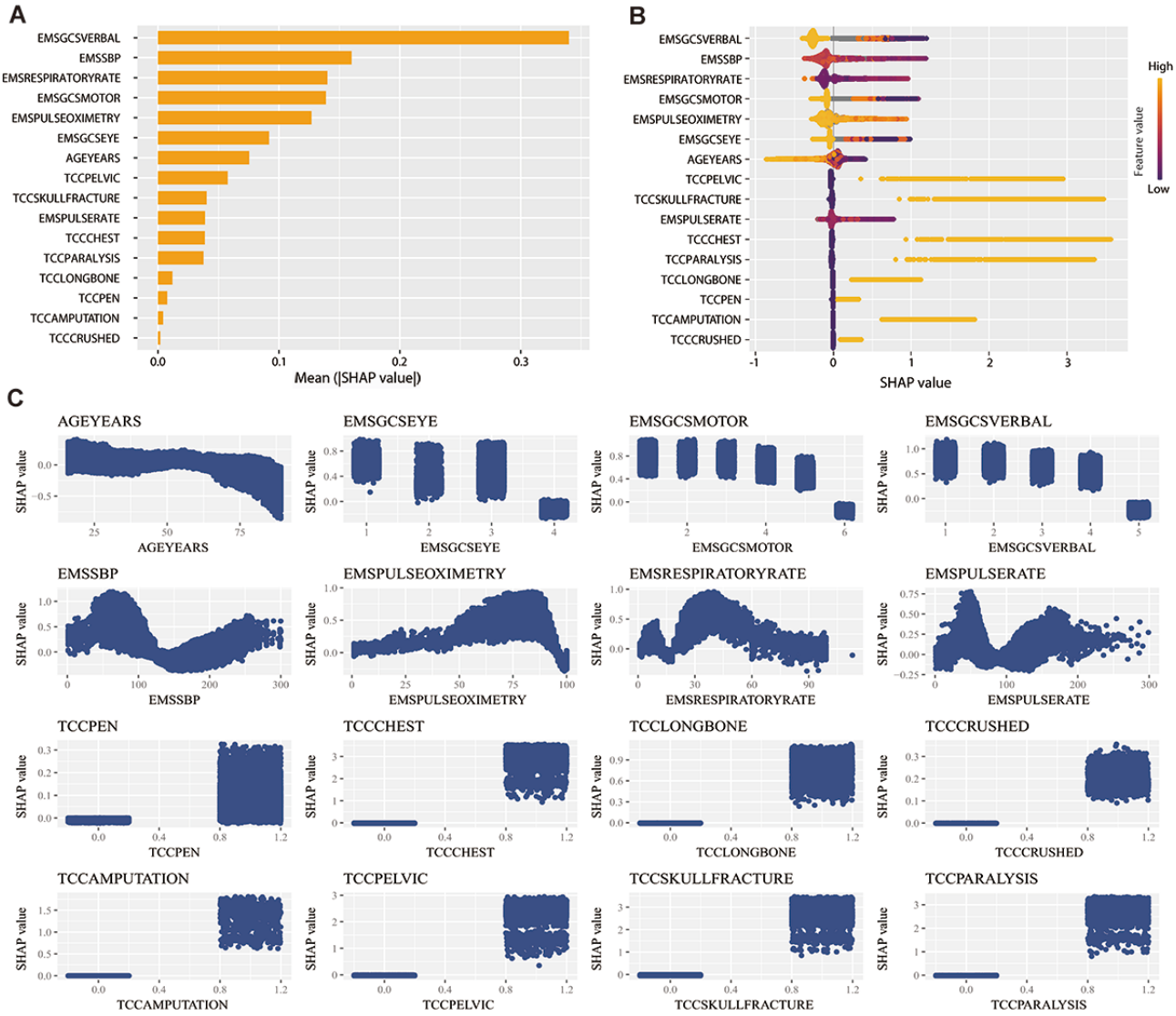

Subgroup Analysis

In Multimedia Appendix 19, subgroup analyses were performed according to age, sex, transport mode, trauma type, and prehospital time. The pTEST model AUCs for severe trauma prediction in patients 60 years of age or older were relatively low at 0.717 (95% CI 0.714-0.720), 0.712 (95% CI 0.708-0.717), and 0.710 (95% CI 0.707-0.714) in the 3 sets, respectively. The AUCs in patients with penetrating injuries were relatively high at 0.815 (95% CI 0.810-0.820), 0.810 (95% CI 0.802-0.817), and 0.810 (95% CI 0.805-0.816) in the 3 sets, respectively. The high proportion of severe trauma in helicopter-transported patients (approximately 41%) led to a high undertriage rate (>0.2) and a low overtriage rate (<0.5). Multimedia Appendix 20 illustrates the pTEST model performance in critical resource use prediction in different subgroups.

Discussion

Principal Findings

In this multisite, large-sample study, we present a prehospital trauma triage tool, pTEST, for severe trauma prediction in EMS. To our knowledge, this is the first study combining ML with national triage guidelines. The pTEST performed optimally in internal and external validations. In addition, its diagnostic accuracy was evaluated using anatomical and resource-based outcomes, and the resource-based outcome is a better alternative to determine the need for specialized trauma care [27]. Furthermore, the pTEST was developed based on national triage guidelines, a globally adopted standard in many organizations. Therefore, the pTEST model can be conveniently applied in EMS practice and may have global relevance.

XSL•FO
RenderX

Comparison With Previous Studies

Consistent with previous studies [8], we also demonstrated the poor performance of the RED criteria from the national triage guidelines. Globally, all triage guidelines are based on a criteria checklist, including vital signs, injury type, and mechanism of injury [7,28,29]. These guidelines are simplistic—patients meeting any 1 of the criteria should be transported to the highest-level trauma center. In reality, there was an interaction among variables and nonlinear effects of continuous variables (Figure 3C). XGBoost, as a nonlinear ensemble method, can train a more accurate classifier from several weak classifiers and has other benefits, such as dealing with missing values and interaction, avoiding overfitting, and accelerating the training speed by parallel calculation [30]. Our study incorporated XGBoost into national field triage guidelines and developed the pTEST model to enhance performance. The pTEST model included age, 7 vital signs, and 8 injury patterns. We did not perform further variable selection because these variables are the most important in the national field triage guidelines, and the overall number is moderate for field triage. The pTEST model did not achieve an undertriage rate of <5% or an overtriage rate of <35%, as the ACS Committee on Trauma targeted. However, 2 aspects must be noted. First, the definition of the undertriage rate by the ACS is different from that used in our study. For example, the undertriage rate in ACS is equal to the number of patients with ISS ≥16 transported to a low-level trauma center or nontrauma center divided by the total number of patients transported to a low-level trauma center or nontrauma center [19]. In contrast, the undertriage rate in our study is equal to the number of patients with ISS ≥16 predicted as ISS <16 divided by the total number of patients predicted ISS <16. In the national field triage guideline, the patients predicted as ISS <16 can be transported to a low-level or nontrauma center. If the patients with predicted ISS <16 and ≥16 are transferred to low and high-level trauma centers, respectively, then the undertriage rate in our study is the same as that in ACS, but this is unlikely in practice. Second, the under- and overtriage rates are affected by severe trauma proportion. Since most (>90%) of our study population came from level 1 and 2 trauma centers, the severe trauma proportion was approximately 18%, resulting in undertriage rate overestimation and overtriage rate underestimation. Based on previous studies with good sample representativeness [9], we assumed that the proportion of severe trauma was 3%. Keeping the sensitivity (0.794) and specificity (0.5) of the pTEST model and the sample size (n=508,703) in the external validation set unchanged, the under- and overtriage rates were 1.26% and 95.3%, respectively, in the external validation set (Multimedia Appendix 20), meeting the ACS undertriage rate target.

The pTEST model and the national field triage guidelines in subgroup analyses were particularly insensitive among older adults [9]. Possible explanations include different physiological responses to injury [31], medication use that potentially worsens injury [32], high prevalence of frailty, and comorbidities [33]. Previous studies have explored elderly-specific triage criteria [15,34,35]. The pTEST model performed better for penetrating traumas. Notably, several studies have found that penetration is a strong severe trauma predictor, and severely penetrated injured patients are more easily recognized [10]. The undertriage rate was high in patients transported using helicopters. Patients with a high proportion of severe trauma, such as those experiencing large-scale casualties, should be transferred to high-level trauma centers without field triage to reduce the undertriage rate.

Previous studies have reported controversial results regarding the use of ML in medical prediction issues [36]. Overall, in studies with a limited number of predictors, ML does not demonstrate advantages over traditional models, such as logistic regression [37], whereas, for studies with many predictors, advanced ML may have an advantage [38]. A recent review, including 14 studies, demonstrated that the predictive ability of ML-based models was significantly better than that of conventional trauma triage tools for outcomes of mortality, hospitalization, and critical care admission, and XGBoost was the most commonly used ML algorithm [14]. In this study, the relatively large number of predictors and sufficient amount of data tended to favor ML applications. We built the pTEST model using XGBoost but did not evaluate other ML methods. We believe an excellent model can be created using a large sample size, an advanced ML method, and robust hyperparameter tuning. In addition, we minimized the risk of chance findings and overfitting by avoiding exploring other modeling strategies.

Limitations

This study had some limitations. First, most of our study population were from level 1 and 2 trauma centers, and the proportion of patients with severe trauma (approximately 18%) was significantly higher than that of all prehospital trauma patients (approximately 3%) [9]. However, unlike PPV and NPV, the sensitivity, specificity, and AUC of the pTEST model were not affected by the proportion of severe trauma, and the high sensitivity, specificity, and AUC objectively reflect the pTEST model's good performance. In addition, some emergency resources may be unavailable in low-level trauma and nontrauma centers [39], and samples from high-level trauma centers make it possible to evaluate the pTEST model with critical resource use as the end point. Second, the pTEST model was not developed into a software application, as in other studies [40], because software development requires adaptation to existing information systems in EMS, which is a complex project. However, in the future, EMS providers can develop software based on available data and programs. Third, the 2017-2019 NTDB data followed the 2011 national field triage guidelines, and the latest guidelines have been revised in 2021. An additional "active bleeding" has been added to high-risk trauma types [7]. The new guidelines will take several years to be implemented, and our model must be further validated and updated as necessary.

Conclusions

We constructed a prehospital triage model, pTEST, to predict severe trauma and achieved an undertriage rate of <10%. Moreover, our study demonstrated that ML is a promising method for enhancing field triage guidelines performance. In the future, we will validate our pTEST model using populations from different countries and casualty backgrounds. In addition,

XSL•FO
RenderX

software must be developed to increase user convenience of the pTEST model in the EMS.

Acknowledgments

The study was funded by grants (2023YFC3107201 and 2023YFC3107202) from the National Key Research and Development Program of China, grant (2022-JCJQ-QT-012) from Young Elite Scientists Sponsorship Program by China Association for Science and Technology, and grant (2022JC011) from Shanghai Emerging Cross Disciplinary Research Project. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the study; and decision to submit the study for publication. BT had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Data Availability

The datasets generated and analyzed during this study are available in ACS Trauma Quality Programs.

Authors' Contributions

QC, CW, and BT discussed and developed the study question for this report. QC, YQ, and ZJ conducted the data extraction and statistical analysis, which was validated by BT. All authors were involved in the interpretation of the data and discussed the results. QC wrote the first draft of this paper. All authors agreed on the final draft of this study. The corresponding author has the right to grant on behalf of all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist.
[DOCX File , 32 KB - jmir_v26i1e58740_app1.docx]

Multimedia Appendix 2
Detailed definition of severe trauma and critical resource use.
[DOCX File , 23 KB - jmir_v26i1e58740_app2.docx]

Multimedia Appendix 3
The candidate variables for model development.
[DOCX File , 23 KB - jmir_v26i1e58740_app3.docx]

Multimedia Appendix 4
The missing proportion of the candidate variables for model development. (A) The missing proportion in training set. (B) The missing proportion in internal validation set. (C) The missing proportion in external validation set.
[DOCX File , 649 KB - jmir_v26i1e58740_app4.docx]

Multimedia Appendix 5
The demographic characteristics, vital signs, and injury patterns between nonsevere traumas and severe traumas in training set.
[DOCX File , 40 KB - jmir_v26i1e58740_app5.docx]

Multimedia Appendix 6
The demographic characteristics, vital signs, and injury patterns between nonsevere traumas and severe traumas in internal validation set.
[DOCX File , 41 KB - jmir_v26i1e58740_app6.docx]

Multimedia Appendix 7
The demographic characteristics, vital signs, and injury patterns between nonsevere traumas and severe traumas in external validation set.
[DOCX File , 40 KB - jmir_v26i1e58740_app7.docx]

Multimedia Appendix 8

XSL•FO
RenderX

The demographic characteristics, vital signs, and injury patterns between no critical resource users and critical resource users in training set.
[DOCX File , 37 KB - jmir_v26i1e58740_app8.docx]

Multimedia Appendix 9
The demographic characteristics, vital signs, and injury patterns between no critical resource users and critical resource users in internal validation set.
[DOCX File , 37 KB - jmir_v26i1e58740_app9.docx]

Multimedia Appendix 10
The demographic characteristics, vital signs, and injury patterns between no critical resource users and critical resource users in external validation set.
[DOCX File , 37 KB - jmir_v26i1e58740_app10.docx]

Multimedia Appendix 11
Model performance metrics for predicting critical resource use fixed at a specificity of 0.5.
[DOCX File , 25 KB - jmir_v26i1e58740_app11.docx]

Multimedia Appendix 12
Model performance metrics for predicting severe trauma at the best thresholds with maximum Youden index.
[DOCX File , 25 KB - jmir_v26i1e58740_app12.docx]

Multimedia Appendix 13
Model performance metrics for predicting critical resource use at the best thresholds with maximum Youden index.
[DOCX File , 26 KB - jmir_v26i1e58740_app13.docx]

Multimedia Appendix 14
Waterfall plot for the predicted probability of pTEST model related to the outcome. (A) Predicted probability of severe trauma in training set. (B) Predicted probability of severe trauma in internal validation set. (C) Predicted probability of severe trauma in external validation set. (D) Predicted probability of critical resource use in training set. (E) Predicted probability of critical resource use in internal validation set. (F) Predicted probability of critical resource use in external validation set. pTEST: prehospital triage model to predict severe trauma.
[DOCX File , 883 KB - jmir_v26i1e58740_app14.docx]

Multimedia Appendix 15
Calibration curves of 5 models. (A) Predicting severe trauma in training set. (B) Predicting severe trauma in internal validation set. (C) Predicting severe trauma in external validation set. (D) Predicting critical resource use in training set. (E) Predicting critical resource use in internal validation set. (F) Predicting critical resource use in external validation set.
[DOCX File , 300 KB - jmir_v26i1e58740_app15.docx]

Multimedia Appendix 16
DCA curves of 5 models. (A) Predicting severe trauma in training set. (B) Predicting severe trauma in internal validation set. (C) Predicting severe trauma in external validation set. (D) Predicting critical resource use in training set. (E) Predicting critical resource use in internal validation set. (F) Predicting critical resource use in external validation set. DCA: decision curve analysis.
[DOCX File , 496 KB - jmir_v26i1e58740_app16.docx]

Multimedia Appendix 17
Global model explanation for predicting critical resource use by the SHAP method in training set. (A) SHAP summary bar plot of the average SHAP value for each variable. (B) SHAP summary dot plot. (C) SHAP dependence plot. SHAP: Shapley additive explanation analysis.
[DOCX File , 523 KB - jmir_v26i1e58740_app17.docx]

Multimedia Appendix 18
Local model explanation for predicting severe trauma by the SHAP method in training set. (A) Waterfall plot of risks contributed by each variable for a patient at high risk of severe trauma. (B) Waterfall plot for a patient at low risk of severe trauma. SHAP: Shapley additive explanation analysis.

XSL•FO
RenderX

[DOCX File , 310 KB - jmir_v26i1e58740_app18.docx]

Multimedia Appendix 19
The performance of the pTEST model for predicting severe trauma in different subgroups. pTEST: prehospital triage model to predict severe trauma.
[DOCX File , 31 KB - jmir_v26i1e58740_app19.docx]

Multimedia Appendix 20
The performance of the pTEST model for predicting critical resource use in different subgroups. The undertriage and overtriage rates of the pTEST in external validation set with the fixed sensitivity, specificity, and sample size. pTEST: prehospital triage model to predict severe trauma.
[DOCX File , 30 KB - jmir_v26i1e58740_app20.docx]

References

1. GBD 2019 Diseases Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet 2020;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. Injuries and violence. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence [accessed 2024-02-22]
3. Morris RS, Karam BS, Murphy PB, Jenkins P, Milia DJ, Hemmila MR, et al. Field-triage, hospital-triage and triage-assessment: a literature review of the current phases of adult trauma triage. J Trauma Acute Care Surg 2021;90(6):e138-e145. [doi: 10.1097/TA.0000000000003125] [Medline: 33605709]
4. MacKenzie EJ, Rivara FP, Jurkovich GJ, Nathens AB, Frey KP, Egleston BL, et al. A national evaluation of the effect of trauma-center care on mortality. N Engl J Med 2006;354(4):366-378. [doi: 10.1056/NEJMsa052049] [Medline: 16436768]
5. Gianola S, Castellini G, Biffi A, Porcu G, Fabbri A, Ruggieri MP, Italian National Institute of Health guideline working group. Accuracy of pre-hospital triage tools for major trauma: a systematic review with meta-analysis and net clinical benefit. World J Emerg Surg 2021;16(1):31 [FREE Full text] [doi: 10.1186/s13017-021-00372-1] [Medline: 34112209]
6. Sasser SM, Hunt RC, Faul M, Sugerman D, Pearson WS, Dulski T, et al. Guidelines for field triage of injured patients: recommendations of the national expert panel on field triage, 2011. MMWR Recomm Rep 2012;61(RR-1):1-20 [FREE Full text] [Medline: 22237112]
7. Newgard CD, Fischer PE, Gestring M, Michaels HN, Jurkovich GJ, Lerner EB, et al. National guideline for the field triage of injured patients: recommendations of the national expert panel on field triage, 2021. J Trauma Acute Care Surg 2022;93(2):e49-e60 [FREE Full text] [doi: 10.1097/TA.0000000000003627] [Medline: 35475939]
8. Lupton JR, Davis-O'Reilly C, Jungbauer RM, Newgard CD, Fallat ME, Brown JB, et al. Under-triage and over-triage using the field triage guidelines for injured patients: a systematic review. Prehosp Emerg Care 2023;27(1):38-45. [doi: 10.1080/10903127.2022.2043963] [Medline: 35191799]
9. Newgard C, Fu R, Zive D, Rea T, Malveau S, Daya M, et al. Prospective validation of the national field triage guidelines for identifying seriously injured persons. J Am Coll Surg 2016;222(2):146-158.e2 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2015.10.016] [Medline: 26712244]
10. van Rein EAJ, Houwert R, Gunning A, Lichtveld R, Leenen L, van Heijl M. Accuracy of prehospital triage protocols in selecting severely injured patients: a systematic review. J Trauma Acute Care Surg 2017;83(2):328-339. [doi: 10.1097/TA.0000000000001516] [Medline: 28452898]
11. Malik NS, Chernbumroong S, Xu Y, Vassallo J, Lee J, Bowley DM, et al. The BCD triage sieve outperforms all existing major incident triage tools: comparative analysis using the UK national trauma registry population. EClinicalMedicine 2021;36:100888 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100888] [Medline: 34308306]
12. Bhalla MC, Frey J, Rider C, Nord M, Hegerhorst M. Simple triage algorithm and rapid treatment and sort, assess, lifesaving, interventions, treatment, and transportation mass casualty triage methods for sensitivity, specificity, and predictive values. Am J Emerg Med 2015;33(11):1687-1691. [doi: 10.1016/j.ajem.2015.08.021] [Medline: 26349777]
13. Voskens FJ, van Rein EAJ, van der Sluijs R, Houwert RM, Lichtveld RA, Verleisdonk EJ, et al. Accuracy of prehospital triage in selecting severely injured trauma patients. JAMA Surg 2018;153(4):322-327 [FREE Full text] [doi: 10.1001/jamasurg.2017.4472] [Medline: 29094144]
14. Adebayo O, Bhuiyan ZA, Ahmed Z. Exploring the effectiveness of artificial intelligence, machine learning and deep learning in trauma triage: a systematic review and meta-analysis. Digit Health 2023;9:20552076231205736 [FREE Full text] [doi: 10.1177/20552076231205736] [Medline: 37822960]
15. Garwe T, Newgard CD, Stewart K, Wan Y, Cody P, Cutler J, et al. Enhancing utility of interfacility triage guidelines using machine learning: development of the geriatric interfacility trauma triage score. J Trauma Acute Care Surg 2023;94(4):546-553 [FREE Full text] [doi: 10.1097/TA.0000000000003846] [Medline: 36404409]

XSL•FO
RenderX

16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594 [FREE Full text] [doi: 10.1136/bmj.g7594] [Medline: 25569120]
17. Hashmi ZG, Kaji AH, Nathens AB. Practical guide to surgical data sets: national trauma data bank (NTDB). JAMA Surg 2018;153(9):852-853. [doi: 10.1001/jamasurg.2018.0483] [Medline: 29617536]
18. van der Sluijs R, Lokerman RD, Waalwijk JF, de Jongh MAC, Edwards MJR, den Hartog D, et al. Accuracy of pre-hospital trauma triage and field triage decision rules in children (P2-T2 study): an observational study. Lancet Child Adolesc Health 2020;4(4):290-298. [doi: 10.1016/S2352-4642(19)30431-6] [Medline: 32014121]
19. Committee on Trauma. Resources for optimal care of the injured patient 2014. American College of Surgeons. 2014. URL: https://www.facs.org/media/yu0laoqz/resources-for-optimal-care.pdf [accessed 2024-09-20]
20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
21. Sharma V, Kulkarni V, Jess E, Gilani F, Eurich D, Simpson SH, et al. Development and validation of a machine learning model to estimate risk of adverse outcomes within 30 days of opioid dispensation. JAMA Netw Open 2022;5(12):e2248559 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.48559] [Medline: 36574245]
22. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
23. Fahmy AS, Csecs I, Arafati A, Assana S, Yankama TT, Al-Otaibi T, et al. An explainable machine learning approach reveals prognostic significance of right ventricular dysfunction in nonischemic cardiomyopathy. JACC Cardiovasc Imaging 2022;15(5):766-779 [FREE Full text] [doi: 10.1016/j.jcmg.2021.11.029] [Medline: 35033500]
24. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
25. He J, Wang B, Tao J, Liu Q, Peng M, Xiong S, et al. Accurate classification of pulmonary nodules by a combined model of clinical, imaging, and cell-free DNA methylation biomarkers: a model development and external validation study. Lancet Digit Health 2023;5(10):e647-e656 [FREE Full text] [doi: 10.1016/S2589-7500(23)00125-5] [Medline: 37567793]
26. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. Inhal Toxicol 2014;26(13):811-828 [FREE Full text] [doi: 10.3109/08958378.2014.955932] [Medline: 25264934]
27. Lerner EB, Cushman JT, Drendel AL, Badawy M, Shah MN, Guse CE, et al. Effect of the 2011 revisions to the field triage guidelines on under- and over-triage rates for pediatric trauma patients. Prehosp Emerg Care 2017;21(4):456-460. [doi: 10.1080/10903127.2017.1300717] [Medline: 28489471]
28. Hamada S, Gauss T, Duchateau F, Truchot J, Harrois A, Raux M, et al. Evaluation of the performance of French physician-staffed emergency medical service in the triage of major trauma patients. J Trauma Acute Care Surg 2014;76(6):1476-1483. [doi: 10.1097/TA.0000000000000239] [Medline: 24854319]
29. Dinh MM, Bein KJ, Oliver M, Veillard AS, Ivers R. Refining the trauma triage algorithm at an Australian major trauma centre: derivation and internal validation of a triage risk score. Eur J Trauma Emerg Surg 2014;40(1):67-74. [doi: 10.1007/s00068-013-0315-1] [Medline: 26815779]
30. Dupont T, Kentish-Barnes N, Pochard F, Duchesnay E, Azoulay E. Prediction of post-traumatic stress disorder in family members of ICU patients: a machine learning approach. Intensive Care Med 2024;50(1):114-124. [doi: 10.1007/s00134-023-07288-1] [Medline: 38112774]
31. Heffernan DS, Thakkar RK, Monaghan SF, Ravindran R, Adams JC, Kozloff MS, et al. Normal presenting vital signs are unreliable in geriatric blunt trauma victims. J Trauma 2010;69(4):813-820. [doi: 10.1097/TA.0b013e3181f41af8] [Medline: 20938267]
32. Nishijima DK, Shahlaie K, Sarkar K, Rudisill N, Holmes JF. Risk of unfavorable long-term outcome in older adults with traumatic intracranial hemorrhage and anticoagulant or antiplatelet use. Am J Emerg Med 2013;31(8):1244-1247 [FREE Full text] [doi: 10.1016/j.ajem.2013.04.035] [Medline: 23759685]
33. Galimberti S, Graziano F, Maas A, Isernia G, Lecky F, Jain S, et al. Effect of frailty on 6-month outcome after traumatic brain injury: a multicentre cohort study with external validation. Lancet Neurol 2022;21(2):153-162. [doi: 10.1016/S1474-4422(21)00374-4] [Medline: 35065038]
34. Ichwan B, Darbha S, Shah MN, Thompson L, Evans DC, Boulger CT, et al. Geriatric-specific triage criteria are more sensitive than standard adult criteria in identifying need for trauma center care in injured older adults. Ann Emerg Med 2015;65(1):92-100.e3. [doi: 10.1016/j.annemergmed.2014.04.019] [Medline: 24908590]
35. Newgard CD, Holmes JF, Haukoos JS, Bulger EM, Staudenmayer K, Wittwer L, et al. Improving early identification of the high-risk elderly trauma patient by emergency medical services. Injury 2016;47(1):19-25 [FREE Full text] [doi: 10.1016/j.injury.2015.09.010] [Medline: 26477345]
36. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]

XSL•FO
RenderX

37. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol 2020;122:95-107 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.03.005] [Medline: 32201256]
38. Larsson A, Berg J, Gellerfors M, Gerdin Wärnberg M. The advanced machine learner XGBoost did not reduce prehospital trauma mistriage compared with logistic regression: a simulation study. BMC Med Inform Decis Mak 2021;21(1):192 [FREE Full text] [doi: 10.1186/s12911-021-01558-y] [Medline: 34148560]
39. Soto JM, Zhang Y, Huang JH, Feng DX. An overview of the American trauma system. Chin J Traumatol 2018;21(2):77-79 [FREE Full text] [doi: 10.1016/j.cjtee.2018.01.003] [Medline: 29605432]
40. van Rein EAJ, van der Sluijs R, Voskens FJ, Lansink KWW, Houwert RM, Lichtveld RA, et al. Development and validation of a prediction model for prehospital triage of trauma patients. JAMA Surg 2019;154(5):421-429 [FREE Full text] [doi: 10.1001/jamasurg.2018.4752] [Medline: 30725101]

Abbreviations

ACS-COT: American College of Surgeons Committee on Trauma
AUC: area under the curve
ED: emergency department
EMS: emergency medical service
GCS: Glasgow Coma Score
ISS: Injury Severity Score
ML: machine learning
NPV: negative predictive value
NTDB: National Trauma Data Bank
PHI: Prehospital Index
PPV: positive predictive value
pTEST: prehospital triage model to predict severe trauma
RTS: revised trauma score
SHAP: Shapley additive explanation analysis
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XGBoost: extreme gradient boosting model

Edited by T de Azevedo Cardoso; submitted 24.03.24; peer-reviewed by L Zhang, D Peng; comments to author 19.04.24; revised version received 11.05.24; accepted 07.08.24; published 30.09.24.

Please cite as:
Chen Q, Qin Y, Jin Z, Zhao X, He J, Wu C, Tang B
Enhancing Performance of the National Field Triage Guidelines Using Machine Learning: Development of a Prehospital Triage Model to Predict Severe Trauma
J Med Internet Res 2024;26:e58740
URL: https://www.jmir.org/2024/1/e58740
doi:10.2196/58740
PMID:

©Qi Chen, Yuchen Qin, Zhichao Jin, Xinxin Zhao, Jia He, Cheng Wu, Bihan Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Based Prediction of Neurodegenerative Disease in Patients With Type 2 Diabetes by Derivation and Validation in 2 Independent Korean Cohorts: Model Development and Validation Study

Hyunji Sang[1,2*], MD, MSc; Hojae Lee[2,3*], MSc; Jaeyu Park[2,3], MSc; Sunyoung Kim[4], MD, PhD; Ho Geol Woo[5], MD, PhD; Ai Koyanagi[6], MD, PhD; Lee Smith[7], PhD; Sihoon Lee[8], MD, PhD; You-Cheol Hwang[9], MD, PhD; Tae Sun Park[10], MD, PhD; Hyunjung Lim[11], PhD; Dong Keon Yon[2,3,12,13], MD, PhD; Sang Youl Rhee[1,2,3,12], MD, PhD

[1]Department of Endocrinology and Metabolism, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[2]Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[3]Department of Regulatory Science, Kyung Hee University, Seoul, Republic of Korea

[4]Department of Family Medicine, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[5]Department of Neurology, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[6]Research and Development Unit, Parc Sanitari Sant Joan de Deu, Barcelona, Spain

[7]Centre for Health, Performance and Wellbeing, Anglia Ruskin University, Cambridge, United Kingdom

[8]Department of Internal Medicine, Gachon University College of Medicine, Incheon, Republic of Korea

[9]Division of Endocrinology and Metabolism, Department of Internal Medicine, Kyung Hee University Hospital at Gangdong and Kyung Hee University School of Medicine, Seoul, Republic of Korea

[10]Division of Endocrinology and Metabolism, Department of Internal Medicine, Research Institute of Clinical Medicine, Jeonbuk National University, Jeonbuk National University Medical School, Jeonju, Republic of Korea

[11]Department of Medical Nutrition, Graduate School of East-West Medical Science, Kyung Hee University, Yongin, Republic of Korea

[12]Department of Precision Medicine, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[13]Department of Pediatrics, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sang Youl Rhee, MD, PhD
Department of Endocrinology and Metabolism
Kyung Hee University Medical Center, Kyung Hee University College of Medicine
23 Kyungheedae-ro, Dongdaemun-gu
Seoul, 02447
Republic of Korea
Phone: 82 2 958 8200
Email: bard95@hanmail.net

Abstract

Background: Several machine learning (ML) prediction models for neurodegenerative diseases (NDs) in type 2 diabetes mellitus (T2DM) have recently been developed. However, the predictive power of these models is limited by the lack of multiple risk factors.

Objective: This study aimed to assess the validity and use of an ML model for predicting the 3-year incidence of ND in patients with T2DM.

Methods: We used data from 2 independent cohorts—the discovery cohort (1 hospital; n=22,311) and the validation cohort (2 hospitals; n=2915)—to predict ND. The outcome of interest was the presence or absence of ND at 3 years. We selected different ML-based models with hyperparameter tuning in the discovery cohort and conducted an area under the receiver operating characteristic curve (AUROC) analysis in the validation cohort.

XSL•FO
RenderX

Results: The study dataset included 22,311 (discovery) and 2915 (validation) patients with T2DM recruited between 2008 and 2022. ND was observed in 133 (0.6%) and 15 patients (0.5%) in the discovery and validation cohorts, respectively. The AdaBoost model had a mean AUROC of 0.82 (95% CI 0.79-0.85) in the discovery dataset. When this result was applied to the validation dataset, the AdaBoost model exhibited the best performance among the models, with an AUROC of 0.83 (accuracy of 78.6%, sensitivity of 78.6%, specificity of 78.6%, and balanced accuracy of 78.6%). The most influential factors in the AdaBoost model were age and cardiovascular disease.

Conclusions: This study shows the use and feasibility of ML for assessing the incidence of ND in patients with T2DM and suggests its potential for use in screening patients. Further international studies are required to validate these findings.

(J Med Internet Res 2024;26:e56922) doi:10.2196/56922

KEYWORDS

machine learning; neurodegenerative disease; diabetes mellitus; prediction; AdaBoost

Introduction

Neurodegenerative diseases (NDs) are characterized by the progressive dysfunction of synapses, neurons, glial cells, and their networks [1]. NDs include dementia, Parkinson disease (PD), multiple sclerosis, Huntington disease, and amyotrophic lateral sclerosis [1]. Risk factors for dementia, the most common type of ND, include older age, genetic risk factors (family history of dementia and Apolipoprotein E), cardiometabolic risk factors (diabetes mellitus, hypertension, dyslipidemia, obesity, and vascular disease), smoking, hearing impairment, depression, less education, physical inactivity, alcohol consumption, traumatic brain injury, and air pollution [2-7]. Known risk factors for Parkinson disease, the second most common ND, include advanced age, male sex, family history of Parkinson disease, environmental exposure (pesticides and air pollution), and comorbidities (obesity, metabolic syndrome, diabetes mellitus, traumatic brain injury, and a history of melanoma or prostate cancer) [8-13].

Type 2 diabetes mellitus (T2DM) is a significant health problem and it requires careful management because it can be accompanied by several complications. In addition to well-known diabetic complications such as retinopathy, neuropathy, and nephropathy, T2DM plays a vital role in the development of cardiovascular, peripheral vascular, and cerebrovascular diseases [14]. Recently, the link between T2DM and the development of NDs has gained attention [15]. The incidence of dementia and Alzheimer disease (AD) among people with diabetes is estimated to be 9.5 and 6.8 per 1000 person-years, respectively [16]. Parkinson disease has affected 31,577 people with T2DM as of 2016 [17]. Compared to healthy people, people with impaired fasting glucose and diabetes for less than 5 years and diabetes for more than 5 years have a 1.04-fold, 1.19-fold, and 1.62-fold higher risk of Parkinson disease, respectively [17].

In clinical practice, biomarkers are needed to accurately diagnose NDs and identify their underlying pathogenesis [18,19]. Current guidelines for ND prevention strategies state that prevention should be based on a plan to reduce modifiable risk factors [20]. To effectively prevent ND in primary care, where most cognitively impaired patients with suspected ND are observed, it is essential to identify good predictors of ND among the standard physical examinations and laboratory tests performed for health screening purposes. To date, there is a lack of information on the correlation and dominance of these predictors.

Recent developments in artificial intelligence have focused on applying novel techniques, such as machine learning (ML), to existing disease models [21]. ML is a powerful tool that can overcome existing limitations using clinical data to uncover hidden patterns and identify critical variables associated with disease development. Clinicians can efficiently detect early warning signs and risk factors of complications by integrating ML algorithms with clinical data [22]. Recently, several models using ML have been developed to predict ND in T2DM [23-25]. The purpose of this study was to identify the relationship between clinical factors and ND and to develop a predictive model for the occurrence of ND through intensive model training and validation by taking advantage of the strengths of ML technology in patients with T2DM in South Korea.

Methods

Study Population and Data Collection

This retrospective study used data from 2 independent longitudinal cohorts previously enrolled in an observational study. Hospital-based data were collected from January 1, 2008, to December 31, 2022. Eligible participants were selected from patients with T2DM, and those with type 1 diabetes or a history of ND were excluded. Finally, 22,311 patients from a tertiary hospital at the Kyung Hee University Medical Center were selected as the discovery cohort. Data for extra validation were collected from a retrospective dataset from the secondary hospitals, Kyung Hee University Medical Center at Gangdong and Gachon University Gil Hospital (validation cohort), and 2915 eligible patients were selected (Figure 1).

XSL•FO
RenderX

Figure 1. Study workflow. ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; GGT: gamma-glutamyl transferase; HbA1c: glycated hemoglobin; HDL: high-density lipoprotein; LDL: low-density lipoprotein; ND: neurodegenerative disease; NIA: National Information Society Agency.

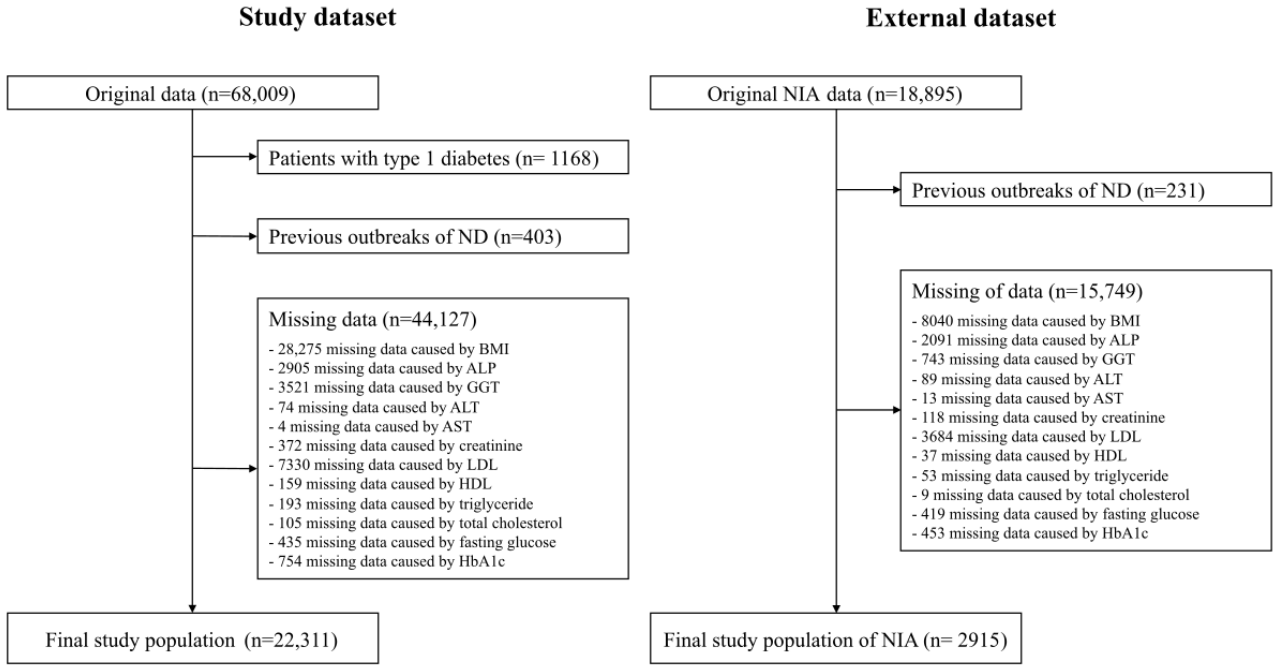

Input Variables

A comprehensive set of 56 variables is included in the model. Baseline patient demographics include age and sex, and medical histories include the presence of hypertension, dyslipidemia, macrovascular complications (cardiovascular and peripheral vascular diseases), microvascular complications (retinopathy, chronic kidney disease, end-stage renal disease, and neuropathy), and cancer. Medication history includes types of antidiabetic agents (metformin, sulfonylurea, dipeptidyl peptidase-4 inhibitor, meglitinide, thiazolidinedione, α-glucosidase inhibitor, insulin, glucagon-like peptide-1 receptor agonist, and sodium-glucose co-transporter 2 inhibitor), antihypertensive drugs (angiotensin II receptor blocker [ARB], angiotensin-converting enzyme inhibitor [ACEi], calcium channel blocker [CCB], diuretics, and beta-blocker), dyslipidemia drugs (statin, fibrate, and ezetimibe), and antiplatelet agents (aspirin, clopidogrel, cilostazol, and glycoprotein IIb/IIIa antagonist). The clinical parameters include BMI [26]. Blood tests included those for glycated hemoglobin (HbA_{1c}), serum glucose, total cholesterol, triglyceride, high-density lipoprotein (HDL), low-density lipoprotein (LDL), serum creatinine, aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma-glutamyl transferase (GGT), and alkaline phosphatase (ALP). For BMI and blood test results, we used the median and standard deviation of each parameter as input variables.

Identification of New ND Cases

New-onset ND in patients with T2DM was identified using the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes for dementia (F00.X–F03.X and G30.X) and Parkinson disease (G20.X). The primary end point was new-onset ND within 3 years.

Data Preprocessing

Missing data were excluded from the analysis. The covariates were divided into three sections: (1) demographics, (2) physical examination and blood tests, and (3) medication and comorbidities. Using the examination date, we used the physical examination and blood test data before the ND outbreak. The dataset was calculated for the entire study period and converted to a mean value before the onset of ND. Information on medications and comorbidities at the first visit was used as a covariate. Data are presented as number (%) or mean (SD).

Model Training and Validation

A common ML approach for prediction involves dividing data into training and test sets. In this study, the target value of the given data on the incidence of ND over 3 years was insufficient. Therefore, the model was trained on the entire dataset rather than splitting it for internal validation. Although including a separate test set is beneficial for providing an unbiased evaluation of the model performance on unseen data from the same distribution as the training data, using an external dataset for validation has its own merits. Validating the model with data from different distributions can verify its robustness and applicability in various real-world settings. A separate external dataset was used to assess the extent to which the model would have generalizability. This approach is essential for verifying whether the model performs well on new and previously unseen data.

XSL•FO
RenderX

Model Development

We chose decision tree–based ensemble models, such as AdaBoost, LightGBM, Random Forest, and XGBoost. Hyperparameter tuning was performed using GridSearchCV, and the area under the receiver operating characteristic curve (AUROC) was maximized to determine the best combination of hyperparameters for optimizing the performance of each model.

ML Analysis

Various tree-based and linear classification models have been used to determine AUROC scores for predicting ND occurrence. GridSearchCV was used to optimize the hyperparameters of models, using the AUROC score as a scoring metric. After determining the optimal hyperparameters, the model is trained for subsequent predictions. Given the class imbalance in our data, we used a synthetic minority oversampling technique to generate synthetic samples. We used various metrics, such as AUROC, accuracy, sensitivity, specificity, and balanced accuracy, to evaluate the model's performance. These metrics are calculated based on the probability predictions produced by the model. A 10-fold stratified cross-validation was performed to assess the model's ability to handle new data. The Youden index was used for each stratification to identify the optimal threshold [27]. Subsequently, we calculated the mean and 95% CIs for each performance metric to measure the average and variability of the model's performance.

A receiver operating characteristic (ROC) curve was plotted to illustrate the performance of the model. This was complemented by the mean ROC curve and SD within that range, demonstrating the distribution of the model performance. Because the AdaBoost model yielded the highest AUROC score among the various decision tree models tested, we selected this model to identify the most important features for predicting ND. The importance of each feature was extracted using the feature importance attributes of the AdaBoost model. We selected the top 15 features with the largest impact on the model and plotted them on a bar graph to visualize their influence on the model predictions. Logistic regression analysis was used to obtain odds ratios (ORs) to compare the effects of the different medications on ND occurrence.

Performance Metrics

To comprehensively understand the performance of our model, we selected 5 performance metrics: AUROC, accuracy, sensitivity, specificity, and balanced accuracy. The AUROC is a robust performance measure that assesses a model's ability to discriminate between classes across all possible thresholds. Its robustness originates from the fact that it considers both sensitivity and specificity, making it a preferred metric, particularly in situations where classes are imbalanced. Accuracy is a simple and intuitive performance metric that indicates the proportion of true results (both true positives and true negatives) from the total number of cases examined. However, accuracy alone can be misleading, particularly for unbalanced datasets; therefore, additional performance metrics are required. Sensitivity and specificity were used to assess how well the model identified the positive and negative cases, respectively. Sensitivity provides insight into a model's ability to detect positive cases by measuring the proportion of true positives correctly identified by the model. Specificity is a measure of the proportion of true negatives that are correctly identified and provides a sense of a model's ability to avoid false alarms. Finally, we include balanced accuracy to provide a more balanced view of the performance of our model, particularly in the face of class disparity. As an average of sensitivity and specificity, balanced accuracy assigns equal weights to both metrics, making it an excellent alternative to accuracy when addressing unbalanced datasets. Combining these metrics enables us to evaluate the performance of our model from different perspectives, thereby ensuring a more robust evaluation [28,29].

Software and Libraries

Data preprocessing, model development, and analyses were conducted using Python (version 3.9.16; Python Software Foundation). The main libraries used in our study include Scikit-learn 1.2.2, NumPy 1.23.5, and Pandas 1.5.3 for ML algorithms and data manipulation. Matplotlib 3.7.1 and Seaborn 0.12.2 were used for data visualization.

Ethical Considerations

This study was approved by the Institutional Review Board of the Kyung Hee University Hospital (KHSIRB-22-473(EA)). The requirement for informed consent was waived by the institutional review board because de-identified data were used in the analyses. This study followed the guidelines outlined in the TRIPOD (Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis) statement. Participants in this study did not receive compensation as the data were anonymized.

Results

Cohort Characteristics

A total of 22,311 patients were selected from the discovery cohort, of whom 133 (0.6%) had ND. Among the participants, 11,545 (51.8%) were male, and the mean age was 63.5 (SD 12.0) years. For additional validation, 2915 patients were included, including 15 (0.5%) patients with ND from the validation cohort. The validation cohort had 1625 (55.8%) men with a mean age of 57.8 (SD 11.8) years (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics of the discovery and validation datasets.

	Discovery dataset			Validation dataset		
	Total (N=22,311)	Control (N=22,178)	Case[a] (N=133)	Total (N=2915)	Control (N=2900)	Case[a] (N=15)
Age, mean (SD)	63.5 (12.0)	63.4 (11.9)	73.3 (7.6)	57.8 (11.8)	57.8 (11.8)	69.5 (10.6)
Male, n (%)	11,545 (51.8)	11,492 (51.8)	53 (39.9)	1625 (55.8)	1617 (55.8)	8 (53.3)
BMI, kg/m^2, mean (SD)	24.9 (3.6)	24.9 (3.6)	24.1 (3.6)	25.2 (3.4)	25.2 (3.4)	23.3 (2.7)
Blood test, mean (SD)						
HbA_{1c}[b], (%)	6.78 (1.00)	6.78 (1.00)	6.77 (0.84)	7.19 (1.14)	7.19 (1.14)	7.64 (1.37)
Fasting blood glucose, mg/dL	146.8 (44.9)	146.8 (44.9)	149.0 (45.9)	145.9 (45.7)	145.8 (45.7)	155.8 (38.5)
Total cholesterol, mg/dL	157.9 (33.9)	157.9 (34.0)	155.1 (28.3)	163.1 (33.0)	163.2 (32.9)	142.2 (34.4)
Triglyceride, mg/dL	141.8 (67.9)	141.9 (68.0)	129.9 (51.0)	147.8 (67.9)	147.7 (67.8)	148.9 (86.0)
HDL[c] cholesterol, mg/dL	47.4 (11.9)	47.4 (11.9)	48.0 (11.6)	45.2 (10.0)	45.3 (10.0)	37.6 (10.7)
LDL[d] cholesterol, mg/dL	90.3 (27.1)	90.4 (27.1)	87.6 (20.8)	92.5 (29.8)	92.6 (29.8)	76.9 (35.8)
Creatinine, mg/dL	0.90 (0.44)	0.90 (0.44)	0.96 (0.48)	1.08 (0.89)	1.08 (0.89)	1.42 (1.28)
AST[e], U/L	26.9 (13.5)	26.9 (13.5)	25.1 (9.4)	24.7 (8.4)	24.7 (8.4)	22.6 (9.2)
ALT[f], U/L	24.5 (14.2)	24.5 (14.2)	19.6 (9.6)	25.6 (13.2)	25.7 (13.2)	15.4 (6.7)
GGT[g], U/L	38.3 (35.8)	38.4 (35.8)	33.0 (30.4)	38.4 (32.4)	38.4 (32.5)	31.3 (18.6)
ALP[h], U/L	79.1 (25.6)	79.1 (25.6)	75.3 (21.9)	171.6 (89.0)	172.1 (89.0)	70.8 (23.3)
Comorbid conditions, n (%)						
Hypertension	9440 (42.3)	9,381 (42.3)	59 (44.4)	1670 (57.3)	1660 (57.2)	10 (66.7)
Dyslipidemia	10,389 (46.6)	10,334 (46.6)	55 (41.4)	1419 (48.7)	1418 (48.9)	1 (6.7)
Macrovascular complications, n (%)						
Cardiovascular disease[i]	8310 (37.3)	8257 (37.2)	53 (39.9)	980 (33.6)	968 (33.4)	12 (80.0)
Peripheral vascular disease	129 (0.6)	129 (0.6)	N/A	422 (14.5)	421 (14.5)	1 (6.7)
Microvascular complications, n (%)						
Retinopathy	1670 (7.5)	1664 (7.5)	6 (4.5)	488 (16.7)	487 (16.8)	1 (6.7)
Chronic kidney disease	2138 (9.6)	2131 (9.6)	7 (5.3)	582 (20.0)	580 (20.0)	2 (13.3)
ESRD[j]	145 (0.7)	144 (0.7)	1 (0.8)	178 (6.1)	176 (6.1)	2 (13.3)
Neuropathy	5061 (22.7)	5027 (22.7)	34 (25.6)	700 (24.0)	698 (24.1)	2 (13.3)
Cancer	3609 (16.2)	3604 (16.3)	5 (3.8)	303 (10.4)	302 (10.4)	1 (6.7)
Medication use, n (%)						
Diabetes mellitus						
Metformin	12,300 (55.1)	12,227 (55.1)	73 (54.9)	1219 (41.8)	1219 (42.0)	N/A
Sulfonylurea	7188 (32.2)	7135 (32.2)	53 (39.9)	462 (15.9)	462 (15.9)	N/A
DPP-4[k] inhibitor	5374 (24.1)	5347 (24.1)	27 (20.3)	195 (6.7)	195 (6.7)	N/A
Meglitinide	985 (4.4)	972 (4.4)	13 (9.8)	171 (5.9)	171 (5.9)	N/A
Thiazolidinedione	1,367 (6.1)	1,356 (6.1)	11 (8.3)	40 (1.4)	40 (1.4)	N/A
α-Glucosidase inhibitor	1,007 (4.5)	1,001 (4.5)	6 (4.5)	161 (5.5)	161 (5.6)	N/A
Insulin	7,188 (32.2)	7,135 (32.2)	53 (39.9)	N/A	N/A	N/A

XSL•FO
RenderX

	Discovery dataset			Validation dataset		
	Total (N=22,311)	Control (N=22,178)	Case[a] (N=133)	Total (N=2915)	Control (N=2900)	Case[a] (N=15)
GLP-1[l] receptor agonist	32 (0.1)	32 (0.1)	N/A	N/A	N/A	N/A
SGLT2[m] inhibitor	634 (2.8)	631 (2.9)	3 (2.3)	N/A	N/A	N/A
Hypertension						
Angiotensin II receptor blocker	10,063 (45.1)	9999 (45.1)	64 (48.1)	245 (8.4)	245 (8.5)	N/A
ACE inhibitor	1735 (7.8)	1724 (7.8)	11 (8.3)	125 (4.3)	125 (4.3)	N/A
Calcium channel blocker	10,485 (47.0)	10,420 (47.0)	65 (48.9)	603 (20.7)	603 (20.8)	N/A
Diuretics	7280 (32.6)	7228 (32.6)	52 (39.1)	229 (7.9)	229 (7.9)	N/A
Beta blocker	5981 (26.8)	5936 (26.8)	45 (33.8)	2 (0.1)	2 (0.1)	N/A
Dyslipidemia						
Statin	12,711 (57.0)	12,625 (56.9)	86 (64.7)	958 (32.9)	958 (33.0)	N/A
Fibrate	896 (4.0)	892 (4.0)	4 (3.0)	92 (3.2)	92 (3.2)	N/A
Ezetimibe	1542 (6.9)	1531 (6.9)	11 (8.3)	106 (3.6)	106 (3.7)	N/A
Antiplatelet						
Aspirin	9136 (41.0)	9064 (40.9)	72 (54.1)	762 (26.1)	762 (26.3)	N/A
Clopidogrel	6406 (28.7)	6359 (28.7)	47 (35.3)	309 (10.6)	309 (10.7)	N/A
Cilostazol	2707 (12.1)	2682 (12.1)	25 (18.8)	164 (5.6)	164 (5.7)	N/A
Glycoprotein IIb/IIIa antagonist	279 (1.3)	276 (1.2)	3 (2.3)	N/A	N/A	N/A

[a]Group of patients with newly developed neurodegenerative disease within 3 years.

[b]HbA_{1c}: glycated hemoglobin.

[c]HDL: high-density lipoprotein.

[d]LDL: low-density lipoprotein.

[e]AST: aspartate transaminase.

[f]ALT: alanine transaminase.

[g]GGT: gamma-glutamyl transferase.

[h]ALP: alkaline phosphatase.

[i]Cardiovascular diseases included ischemic heart disease, myocardial infarction, heart failure, atrial fibrillation, stroke, and other cerebrovascular diseases.

[j]ESRD: end-stage renal disease.

[k]DPP-4: dipeptidyl peptidase-4.

[l]GLP-1: glucagon-like peptide-1.

[m]SGLT2: sodium-glucose co-transporter 2.

Comparisons of Prediction Model Performance

The AdaBoost model performed well on the discovery set (AUROC 0.82, 95% CI 0.79-0.85; accuracy 74.2%, 95% CI 70.9-77.4; sensitivity 73.6%, 95% CI 69.9-77.3; specificity 74.2%, 95% CI 70.9-77.4; and balanced accuracy 73.9%, 95% CI 70.6-77.2). The LightGBM model performed next best (AUROC 0.791, 95% CI 0.756-0.825; accuracy 72.5%, 95% CI 68.1-76.8; sensitivity 71.8%, 95% CI 67.0-76.6; specificity 72.5%, 95% CI 68.1-76.8; and balanced accuracy 72.2%, 95% CI 68.0-76.3). The Random Forest model closely followed, yielding an AUROC of 0.79 (95% CI 0.76-0.82) and solid metrics (accuracy 69.3%, 95% CI 63.0-75.5; sensitivity 69.1%, 95% CI 63.6-74.6; specificity 69.3%, 95% CI 63.0-75.6; and balanced accuracy 69.2%, 95% CI 63.4-75.0). The XGBoost model also had similar performance (AUROC 0.79, 95% CI 0.77-0.81; accuracy 72.0%, 95% CI 68.5-75.6; sensitivity 69.1%, 95% CI 65.7-72.5; specificity 72.0%, 95% CI 68.5-75.6; and balanced accuracy 70.6%, 95% CI 67.4-73.8; Table 2).

XSL•FO
RenderX

Table 2. Performance metrics of 4 different ML[a] algorithms on the original and external validation datasets.

Model	AUROC[b]	Accuracy, (%)	Sensitivity, (%)	Specificity, (%)	Balanced accuracy, (%)
Original dataset					
ADB[c]	0.819 (0.786–0.851)	74.2 (70.9-77.4)	73.6 (69.9-77.3)	74.2 (70.9-77.4)	73.9 (70.6-77.2)
LGB[d]	0.791 (0.756-0.825)	72.5 (68.1-76.8)	71.8 (67.0-76.6)	72.5 (68.1-76.8)	72.2 (68.0-76.3)
RF[e]	0.788 (0.757-0.819)	69.3 (63.0-75.5)	69.1 (63.6-74.6)	69.3 (63.0-75.6)	69.2 (63.4-75.0)
XGB[f]	0.788 (0.766-0.810)	72.0 (68.5-75.6)	69.1 (65.7-72.5)	72.0 (68.5-75.6)	70.6 (67.4-73.8)
Validation dataset					
ADB	0.830	78.6	78.6	78.6	78.6
LGB	0.833	84.3	78.6	84.4	81.5
RF	0.820	77.0	78.6	76.9	77.8
XGB	0.786	74.3	71.4	74.3	72.9

[a]ML: machine learning.
[b]AUROC: area under the receiver operating characteristic curve.
[c]ADB: AdaBoost.
[d]LGB: LightGBM.
[e]RF: Random Forest.
[f]XGB: XGBoost.

Performance metrics of 4 machine learning models (AdaBoost, LightGBM, Random Forest, and XGBoost) were provided for the prediction of the onset of neurodegenerative disease within 3 years in patients, using both the original and an additional external validation dataset.

Upon applying these models to the external validation set, the AdaBoost and LightGBM models achieved high AUROCs of 0.83 and 0.83, respectively. The Random Forest and XGBoost models exhibited improved performance metrics with AUROC values of 0.82 and 0.79, respectively (Table 2).

Consequently, with excellent and consistent results in both independent datasets, the AdaBoost model emerged as the best predictor of ND development within 3 years among patients with diabetes (Figures 2 and 3).

Figure 2. Model architecture. An electronic medical record (EMR) dataset from Kyung Hee University Medical Center was used for model development processed by 10-fold cross-validation and AdaBoost. Extravalidation was executed by the National Information Society Agency (NIA) dataset from Kyung Hee University Hospital at Gangdong and Gachon University Gil Medical Center.

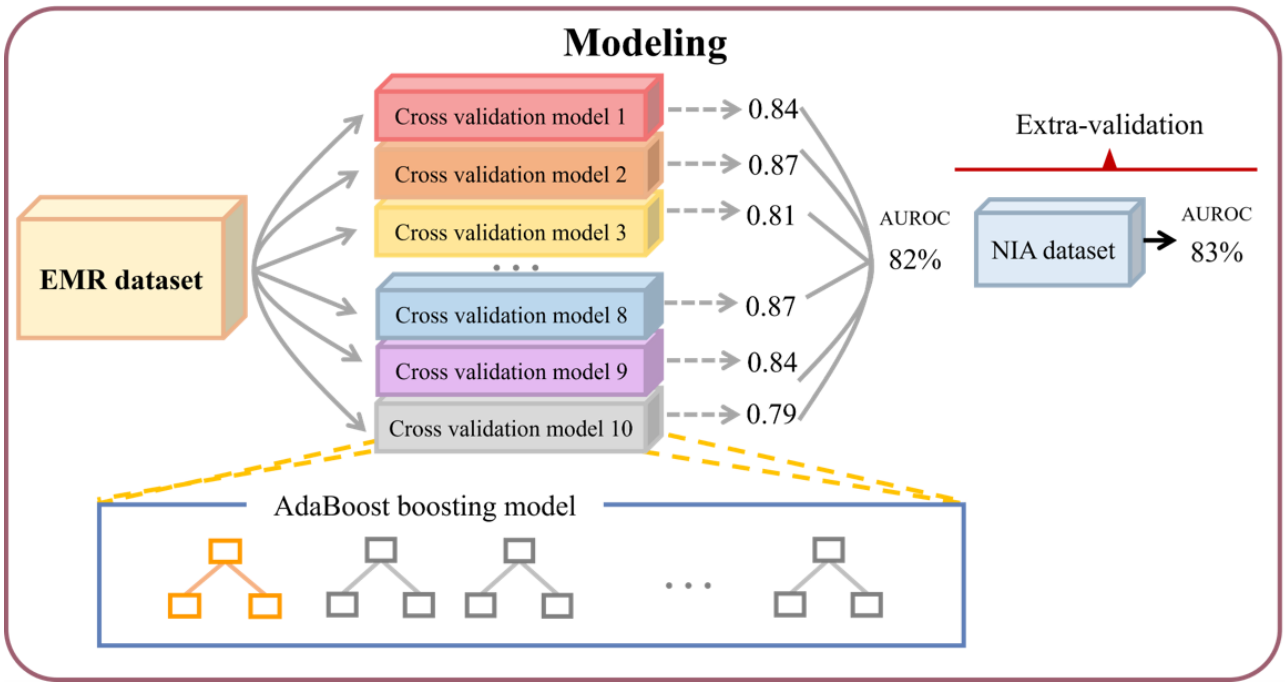

XSL•FO
RenderX

Figure 3. ROC curves of the AdaBoost model. Mean ROC curve from 10-fold cross-validation on the original dataset. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.

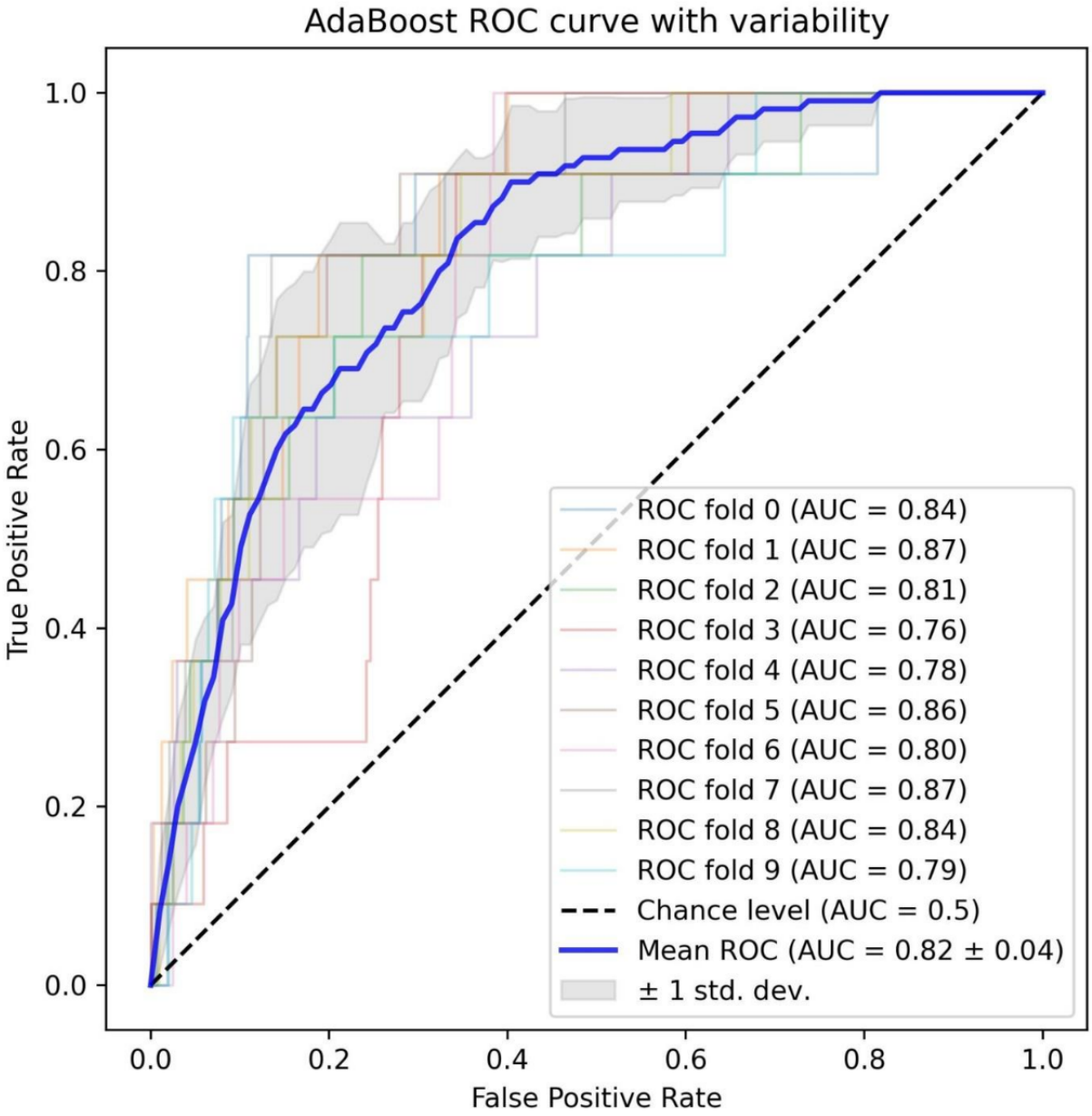

Feature Importance for Improving Interpretability of ML Models

The impacts of the contributing factors analyzed using the feature importance method are shown in Figure 4. Among the 56 variables considered in this study, age was the most important factor that contributed to the performance of the ND prediction model, followed by cardiovascular disease, cancer, neuropathy, and ALP levels. For other comorbidities, dyslipidemia, chronic kidney disease, and hypertension were among the top 10 feature importance, and for medications, metformin, calcium channel blockers, and meglitinide were among the top 15 feature importance.

XSL•FO
RenderX

Figure 4. Top 15 feature-importance of AdaBoost model. ALP: alkaline phosphatase; CCB: calcium channel blocker; CKD: chronic kidney disease; CVD: cardiovascular disease; LDL: low-density lipoprotein.

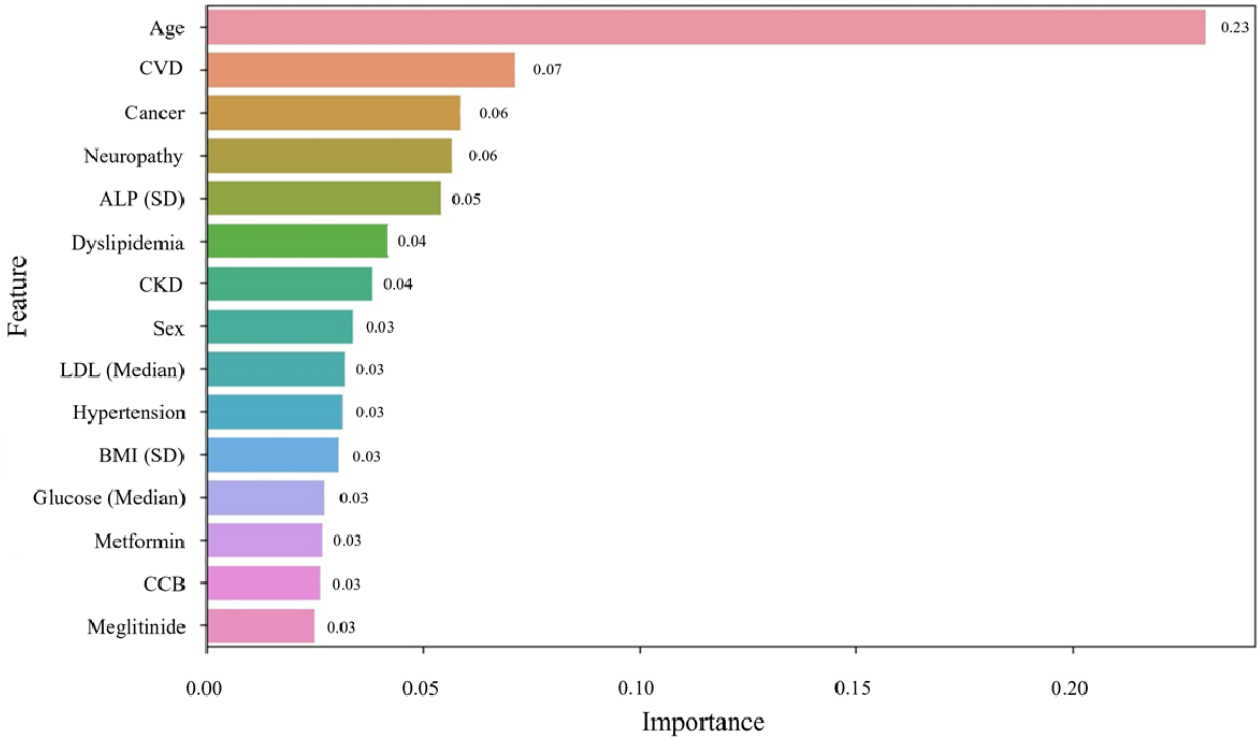

Comparison of the Impact of Different Drugs on ND Occurrence

We compared the risk of developing ND according to the type of medication previously administered by the participants by calculating ORs. Among antidiabetic drugs and antiplatelet drugs, meglitinide (OR 0.42, 95% CI 0.24-0.75), aspirin (OR 0.59, 95% CI 0.42-0.82), and cilostazol (OR 0.59, 95% CI 0.38-0.92) were associated with a significantly reduced risk of ND. The remaining antidiabetic, antihypertensive, antihyperlipidemic, and antiplatelet agents did not significantly increase or decrease the risk of ND development (Multimedia Appendix 1).

Discussion

This study emphasized the importance of developing a highly accurate ML-based ND prediction model that can be universally applied to adults with T2DM in South Korea. This study provided a simple and precise assessment of the future annual risk of ND in people with diabetes nationwide. The AdaBoost, LightGBM, Random Forest, and XGBoost ensemble models showed excellent performance with AUROC values ranging from 0.79 to 0.82 on the discovery dataset and 0.79 to 0.83 on the external validation dataset. Age and cardiovascular disease were the top 15 factors affecting the feature importance. The results of this study can potentially improve patient outcomes by enabling timely intervention, advancing the comprehension of contributing variables, and reducing the burden of neurodegenerative complications in patients with T2DM.

This study was based on a large cohort of the Korean population and used data from 3 university hospitals. Multiple variables, such as anthropometric variables, medical history, medication use, and laboratory tests, were used for model development. An advantage of this study is that long-term follow-up data of approximately 3 years were available for outcome evaluation. This ND prediction model is meaningful because it demonstrates sufficiently good performance, with a mean AUROC of 0.82, using only questionnaires, body measurements, and blood tests commonly conducted in clinical practice for patients with diabetes.

Our findings provide insights into the metrics that can be used in primary care for ND prediction. In addition to the risk factors considered by our ML model, other known risk factors for the development of ND include genetic risk factors, lifestyle factors, environmental exposure, and traumatic brain injury. Existing biomarkers for ND are related to amyloid beta or tau proteins, which are involved in the pathophysiology of ND and are measured using neuroimaging techniques, such as brain magnetic resonance imaging, single-photon emission computed tomography or positron emission tomography, or cerebrospinal fluid testing [30,31]. Additionally, biomarkers using blood samples exist, such as high-sensitivity C-reactive protein, GGT, homocysteine, apolipoprotein E, and uric acid [30,31]. Although they can be used as adjuncts to increase diagnostic confidence, most are expensive or invasive and are not recommended as routine diagnostic tests in clinical practice. As we aimed to predict ND risk for screening purposes in primary health care centers, we focused on constructing an ML model that can predict ND risk based on general physical measurements and blood tests without requiring specialized tests.

In this study, we performed a feature-importance analysis on the interpretability of the AdaBoost model, which performed the best among the models investigated. According to feature importance analysis, age, cardiovascular disease, cancer,

XSL•FO
RenderX

neuropathy, and ALP levels were among the top 5 predictors of ND. The association between age, cardiovascular disease, and neuropathy with ND was consistent with the results of previous studies. Age is a conventional risk factor for ND [32]. Cardiovascular disease is a known risk factor for ND [33]. Cardiometabolic risk factors such as diabetes, hypertension, and hyperlipidemia were also consistently associated with the risk of developing ND [34]. The association between peripheral neuropathy and ND in this study is consistent with its association with the development of mild cognitive impairment and dementia in the general population and people with diabetes [35]. Meanwhile, the relationship between cancer and ND is likely to be inverse according to previous studies. The incidence of cancer is reportedly lower in patients with ND [36]. It is important to note that aging also affects the occurrence of cancer [37], and this study did not adjust for the effect of aging on cancer; therefore, further research is needed to determine the causality between cancer itself and ND. The results related to ALP levels were consistent with previous reports showing that ALP levels were increased in patients with AD [38]. In contrast, some studies have found no significant association between ALP and PD [39]. This may be related to increased bone ALP, as PD is associated with an increased incidence of osteoporosis, falls, and fractures [40]. Moreover, the association between ALP variability and ND development has not been previously studied and warrants further investigation.

Although the AdaBoost model identified important features, such as the use of metformin, CCB, and meglitinide, the results of logistic regression analysis showed that only meglitinide significantly reduced the risk of ND. This difference arises because AdaBoost can capture complex patterns and nonlinear interactions among variables that logistic regression may not fully capture because of its linear assumptions. The ability of AdaBoost to highlight nonlinear relationships provides additional insights into the factors affecting ND risk.

While some studies have shown an increased incidence of ND with long-term exposure to metformin [41], conflicting studies have suggested that metformin has a therapeutic potential for ND [42]. Because metformin users may have more hyperglycemia than nonusers, it is difficult to conclude that metformin use worsens the risk of developing ND. However, few studies have investigated the association between meglitinide use and ND. In one study, meglitinide showed a significant protective effect against dementia in combination therapy rather than in monotherapy [43]. Because meglitinide is often used in combination with agents such as metformin rather than as a monotherapy, and in patients with diabetes who are not glycemically controlled despite multidrug therapy, there may be more meglitinide users among those who develop ND due to hyperglycemia [44,45]. However, the number of meglitinide users was too low to confirm this association.

Given that ARBs and CCBs are the first and second most prescribed drugs for hypertension in Korea as monotherapy, and the combination of ACEi/ARBs and CCBs is the first most prescribed drug in 2-drug therapy [46], CCBs are ranked higher in feature importance for the development of ND than ARBs. The preventive effect of CCBs on ND has been recognized in epidemiologic studies [47], and it is known that specific calcium channel subtypes are implicated in the pathogenesis of PD and that dihydropyridine CCBs with selectivity for these ion channels have a neuroprotective effect in animal models [48]. Although some conflicting studies have shown that antihypertensive drugs are not associated with ND [49], the results of this study show that CCB is effective in preventing ND.

Aspirin has previously been shown to reduce the incidence of AD and PD, as well as cardiovascular events and cancer. Aspirin-medicated acetylation prevents several neurodegenerative pathologies by interfering with protein aggregation [50]. Cilostazol has been shown to have a neuroprotective effect against vascular dementia in mice induced by L-methionine [51]. It is unclear whether the protective effects of aspirin and cilostazol against ND are due to an indirect lowering of the incidence of ND because of their pre-existing effects on reducing the risk of cardiovascular disease, another risk factor for ND, or whether they directly affect the pathological mechanisms of ND.

This study has several limitations. First, due to the retrospective nature of the study, obtaining accurate information from a dataset based on hospital medical records was difficult. Missing values, privacy regulations, and historical biases affect data availability for model training. The model was trained on data obtained from tertiary care centers, which may introduce selection bias because the patients may have different socioeconomic backgrounds from those of primary care patients and receive more comprehensive health care, thereby affecting ND risk and diabetes management. Information bias arising from inaccuracies in data-acquisition methods and the recording of medications and clinical parameters can result in the misclassification of both exposure and outcome, thus affecting the accuracy of the ML model's predictions for ND in patients with T2DM. Second, the performance of this prediction model was not compared with that of other existing prediction models for ND in T2DM, and we expect that future comparative analyses will provide insights into the added value of the ML approach. Additionally, this study was limited to a prediction period of 3 years because of the availability and robustness of the follow-up data in the cohort. In the future, including data from longer periods (eg, 5 or 10 years) would strengthen the generalizability of the model for assessing long-term ND risk.

Furthermore, this study, on its own, cannot prove a causal relationship between the predictors used in the model and the incidence of ND. Confounding factors in predicting ND in patients with T2DM include possible biases due to medications prescribed for conditions such as hypertension or dyslipidemia; the effects of unexplained variables such as the severity and duration of diabetes; and missing data pertaining to lifestyle factors (smoking status, physical activity, diet, and alcohol consumption), all of which can distort the true relationship between the risk factors and ND outcomes. Further experimental research is needed to clarify the biological pathways and demonstrate the mechanisms of interaction between variables related to ND and their impact on the development of ND. Additionally, the dataset used was derived from hospital data based on *ICD-10* codes for PD and dementia only, which we grouped and defined as ND. Subcategories based on the

XSL•FO
RenderX

pathology of ND were not considered, and other rare diseases such as multiple sclerosis, Huntington disease, and amyotrophic lateral sclerosis were not included in the ND outcome.

Also, integrating new models into existing health care systems can be challenging owing to compatibility issues with legacy software, potential disruption in clinical workflows, staff resistance, and the logistical and financial burdens of training. Hence, compatible software, user-friendly interfaces, and comprehensive training and support for health care professionals are necessary. Finally, the ML model used, which was trained and validated on data from Korean patients, may have limited generalizability to diverse populations owing to the different genetic, lifestyle, and environmental backgrounds. To enhance the applicability of the model, tone must incorporate data from diverse demographic groups and conduct external validation across different geographic locations and health care settings.

In conclusion, this study developed an ML-based prediction model using a representative national cohort. The model accurately predicted the risk of ND in all members of the Korean population with T2DM. We also demonstrate that the performance of several ML models is satisfactory. The AdaBoost model performed the best (AUROC 0.82 in the discovery dataset and AUROC 0.83 in the validation dataset). Our predictive model suggests that clinicians should consider age and cardiovascular disease, among other relevant variables, when assessing the risk of ND in patients with T2DM. This emphasizes the importance of comprehensive cardiovascular care and early intervention strategies to mitigate the risk of ND development in older patients with T2DM. This study is the first to apply an ML-based ND-prediction system to a national population with diabetes. In clinical practice, ML models enhance the prediction and management of ND by facilitating early intervention, personalizing treatment, optimizing resource allocation, and improving diagnostic accuracy. These models, when integrated with conventional diagnostics, can facilitate care for high-risk individuals and reduce long-term health care costs. The prediction model proposed in this study is expected to be competitive and cost-effective in preventing ND in Korean patients with T2DM and is expected to be widely used, especially in primary care settings. Future studies should focus on refining these models via longitudinal studies across diverse settings to address ethical concerns regarding data privacy and to promote multidisciplinary collaboration for advancing ND-prediction and treatment strategies.

Acknowledgments

This study will be expanded to include data collected by the National Information Society Agency, South Korea, and multicenter research associated with the Korean Diabetes Association and the Korea Disease Control and Prevention Agency.

This study was supported by the National Institutes of Health Research Project (2022-ER1102-02) of South Korea. The funders had no role in the study design, data collection, data analysis, data interpretation, or manuscript writing.

Data Availability

The datasets generated and analyzed during the current study are not publicly available to protect patient privacy but are available from the corresponding author on reasonable request.

Authors' Contributions

HS, DKY, and SYR conceptualized and designed the study. Hojae L and JP acquired, conducted, and interpreted the data. SK, HGW, AK, LS, SL, Y-CH, TSP, Hyunjung L, and SYR drafted and revised the manuscript. DKY and SYR approved the final version of the manuscript. All authors discussed the results and contributed to the final manuscript. HS and Hojae L contributed equally to this study. DKY and SYR contributed equally as corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Odds ratios with 95% CIs for neurodegenerative diseases in relation to medication use.
[DOCX File , 18 KB - jmir_v26i1e56922_app1.docx]

References

1. Pathak N, Vimal SK, Tandon I, Agrawal L, Hongyi C, Bhattacharyya S. Neurodegenerative disorders of alzheimer, parkinsonism, amyotrophic lateral sclerosis and multiple sclerosis: an early diagnostic approach for precision treatment. Metab Brain Dis 2022;37(1):67-104. [doi: 10.1007/s11011-021-00800-w] [Medline: 34719771]
2. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the lancet commission. Lancet 2020;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]

XSL•FO
RenderX

3. Niu H, Álvarez-Álvarez I, Guillén-Grima F, Aguinaga-Ontoso I. Prevalence and incidence of alzheimer's disease in Europe: a meta-analysis. Neurologia 2017;32(8):523-532 [FREE Full text] [doi: 10.1016/j.nrl.2016.02.016] [Medline: 27130306]
4. Castellano JM, Kim J, Stewart FR, Jiang H, DeMattos RB, Patterson BW, et al. Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. Sci Transl Med 2011;3(89):89ra57 [FREE Full text] [doi: 10.1126/scitranslmed.3002156] [Medline: 21715678]
5. Levine DA, Galecki AT, Langa KM, Unverzagt FW, Kabeto MU, Giordani B, et al. Trajectory of cognitive decline after incident stroke. JAMA 2015;314(1):41-51 [FREE Full text] [doi: 10.1001/jama.2015.6968] [Medline: 26151265]
6. Xu WL, Atti AR, Gatz M, Pedersen NL, Johansson B, Fratiglioni L. Midlife overweight and obesity increase late-life dementia risk: a population-based twin study. Neurology 2011;76(18):1568-1574 [FREE Full text] [doi: 10.1212/WNL.0b013e3182190d09] [Medline: 21536637]
7. Prince M, Acosta D, Ferri CP, Guerra M, Huang Y, Llibre Rodriguez JJ, et al. Dementia incidence and mortality in middle-income countries, and associations with indicators of cognitive reserve: a 10/66 dementia research group population-based cohort study. Lancet 2012;380(9836):50-58 [FREE Full text] [doi: 10.1016/S0140-6736(12)60399-7] [Medline: 22626851]
8. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of parkinson's disease, 1990-2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol 2018;17(11):939-953 [FREE Full text] [doi: 10.1016/S1474-4422(18)30295-3] [Medline: 30287051]
9. Noyce AJ, Bestwick JP, Silveira-Moriyama L, Hawkes CH, Giovannoni G, Lees AJ, et al. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. Ann Neurol 2012;72(6):893-901 [FREE Full text] [doi: 10.1002/ana.23687] [Medline: 23071076]
10. Elbaz A, Clavel J, Rathouz PJ, Moisan F, Galanaud J, Delemotte B, et al. Professional exposure to pesticides and parkinson disease. Ann Neurol 2009;66(4):494-504. [doi: 10.1002/ana.21717] [Medline: 19847896]
11. Jo S, Kim YJ, Park KW, Hwang YS, Lee SH, Kim BJ, et al. Association of NO2 and other air pollution exposures with the risk of parkinson disease. JAMA Neurol 2021;78(7):800-808 [FREE Full text] [doi: 10.1001/jamaneurol.2021.1335] [Medline: 33999109]
12. Nam GE, Kim SM, Han K, Kim NH, Chung HS, Kim JW, et al. Metabolic syndrome and risk of parkinson disease: a nationwide cohort study. PLoS Med 2018;15(8):e1002640 [FREE Full text] [doi: 10.1371/journal.pmed.1002640] [Medline: 30130376]
13. Gardner RC, Byers AL, Barnes DE, Li Y, Boscardin J, Yaffe K. Mild TBI and risk of parkinson disease: a chronic effects of neurotrauma consortium study. Neurology 2018;90(20):e1771-e1779 [FREE Full text] [doi: 10.1212/WNL.0000000000005522] [Medline: 29669907]
14. Aljerf L, Alhaffar I. Salivary distinctiveness and modifications in males with diabetes and behçet's disease. Biochem Res Int 2017;2017:9596202 [FREE Full text] [doi: 10.1155/2017/9596202] [Medline: 28321337]
15. Schernhammer E, Hansen J, Rugbjerg K, Wermuth L, Ritz B. Diabetes and the risk of developing parkinson's disease in Denmark. Diabetes Care 2011;34(5):1102-1108 [FREE Full text] [doi: 10.2337/dc10-1333] [Medline: 21411503]
16. Yu JH, Han K, Park S, Cho H, Lee DY, Kim JW, et al. Incidence and risk factors for dementia in type 2 diabetes mellitus: a nationwide population-based study in Korea. Diabetes Metab J 2020;44(1):113-124 [FREE Full text] [doi: 10.4093/dmj.2018.0216] [Medline: 31769236]
17. Rhee SY, Han KD, Kwon H, Park SE, Park YG, Kim YH, et al. Association between glycemic status and the risk of parkinson disease: a nationwide population-based study. Diabetes Care 2020;43(9):2169-2175 [FREE Full text] [doi: 10.2337/dc19-0760] [Medline: 32611610]
18. Ajong AB, Kenfack B, Ali IM, Yakum MN, Onydinma UP, Mangala FN, et al. Ionised and total hypocalcaemia in pregnancy: an analysis of prevalence and risk factors in a resource-limited setting, Cameroon. PLoS One 2022;17(5):e0268643 [FREE Full text] [doi: 10.1371/journal.pone.0268643] [Medline: 35584129]
19. Ajong AB, Kenfack B, Ali IM, Yakum MN, Ukaogo PO, Mangala FN, et al. Adverse maternofoetal outcomes associated with ionised calcaemia, total calcaemia, albuminaemia, and calcium supplementation in pregnancy: analysis from a resource-limited setting. PLoS One 2022;17(8):e0271525 [FREE Full text] [doi: 10.1371/journal.pone.0271525] [Medline: 35913943]
20. Risk reduction of cognitive decline and dementia: WHO guidelines. Geneva: World Health Organization; 2019.
21. Rhee SY, Sung JM, Kim S, Cho IJ, Lee SE, Chang HJ. Development and validation of a deep learning based diabetes prediction system using a nationwide population-based cohort. Diabetes Metab J 2021;45(4):515-525 [FREE Full text] [doi: 10.4093/dmj.2020.0081] [Medline: 33631067]
22. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
23. Kim J, Lee H, Lee J, Rhee SY, Shin JI, Lee SW, et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning: a post hoc analysis of a diagnostic trial and validation of an external additional trial. Alzheimers Res Ther 2023;15(1):127 [FREE Full text] [doi: 10.1186/s13195-023-01268-9] [Medline: 37481573]

XSL•FO
RenderX

24. Bron EE, Klein S, Papma JM, Jiskoot LC, Venkatraghavan V, Linders J, Alzheimer's Disease Neuroimaging Initiative, Parelsnoer Neurodegenerative Diseases study group. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of alzheimer's disease. Neuroimage Clin 2021;31:102712 [FREE Full text] [doi: 10.1016/j.nicl.2021.102712] [Medline: 34118592]
25. Bian J, Wang X, Hao W, Zhang G, Wang Y. The differential diagnosis value of radiomics-based machine learning in parkinson's disease: a systematic review and meta-analysis. Front Aging Neurosci 2023;15:1199826 [FREE Full text] [doi: 10.3389/fnagi.2023.1199826] [Medline: 37484694]
26. Eum S, Rhee SY. Age, ethnic, and sex disparity in body mass index and waist circumference: a bi-national large-scale study in South Korea and the United States. Life Cycle 2023;3:1-13.
27. Wacker C, Prkno A, Brunkhorst FM, Schlattmann P. Procalcitonin as a diagnostic marker for sepsis: a systematic review and meta-analysis. Lancet Infect Dis 2013;13(5):426-435. [doi: 10.1016/S1473-3099(12)70323-7] [Medline: 23375419]
28. Lee SW. Regression analysis for continuous independent variables in medical research: statistical standard and guideline of life cycle committee. Life Cycle 2022;2:1-8.
29. Kim J, Kim SC, Kang D, Yon DK, Kim JG. Classification of alzheimer's disease stage using machine learning for left and right oxygenation difference signals in the prefrontal cortex: a patient-level, single-group, diagnostic interventional trial. Eur Rev Med Pharmacol Sci 2022;26(21):7734-7741 [FREE Full text] [doi: 10.26355/eurrev_202211_30122] [Medline: 36394721]
30. Gasecka A, Siwik D, Gajewska M, Jaguszewski MJ, Mazurek T, Filipiak KJ, et al. Early biomarkers of neurodegenerative and neurovascular disorders in diabetes. J Clin Med 2020;9(9):2807 [FREE Full text] [doi: 10.3390/jcm9092807] [Medline: 32872672]
31. Pastorello Y, Carare RO, Banescu C, Potempa L, Di Napoli M, Slevin M. Monomeric C-reactive protein: a novel biomarker predicting neurodegenerative disease and vascular dysfunction. Brain Pathol 2023;33(6):e13164 [FREE Full text] [doi: 10.1111/bpa.13164] [Medline: 37158450]
32. Hou Y, Dan X, Babbar M, Wei Y, Hasselbalch SG, Croteau DL, et al. Ageing as a risk factor for neurodegenerative disease. Nat Rev Neurol 2019;15(10):565-581. [doi: 10.1038/s41582-019-0244-7] [Medline: 31501588]
33. Newman AB, Fitzpatrick AL, Lopez O, Jackson S, Lyketsos C, Jagust W, et al. Dementia and alzheimer's disease incidence in relationship to cardiovascular disease in the cardiovascular health study cohort. J Am Geriatr Soc 2005;53(7):1101-1107. [doi: 10.1111/j.1532-5415.2005.53360.x] [Medline: 16108925]
34. Zheng B, Su B, Price G, Tzoulaki I, Ahmadi-Abhari S, Middleton L. Glycemic control, diabetic complications, and risk of dementia in patients with diabetes: results from a large U.K. cohort study. Diabetes Care 2021;44(7):1556-1563. [doi: 10.2337/dc20-2850] [Medline: 34035076]
35. Hicks CW, Wang D, Schneider ALC, Johansen MC, Gottesman RF, Matsushita K, et al. Associations of peripheral neuropathy defined by monofilament insensitivity with mild cognitive impairment and dementia in older adults. Dement Geriatr Cogn Disord 2022;51(2):150-158 [FREE Full text] [doi: 10.1159/000523762] [Medline: 35344962]
36. Driver JA. Inverse association between cancer and neurodegenerative disease: review of the epidemiologic and biological evidence. Biogerontology 2014;15(6):547-557. [doi: 10.1007/s10522-014-9523-2] [Medline: 25113739]
37. DePinho RA. The age of cancer. Nature 2000;408(6809):248-254. [doi: 10.1038/35041694] [Medline: 11089982]
38. Kellett KA, Williams J, Vardy ER, Smith AD, Hooper NM. Plasma alkaline phosphatase is elevated in alzheimer's disease and inversely correlates with cognitive function. Int J Mol Epidemiol Genet 2011;2(2):114-121 [FREE Full text] [Medline: 21686125]
39. Meamar R, Maracy M, Chitsaz A, Ghazvini MRA, Izadi M, Tanhaei AP. Association between serum biochemical levels, related to bone metabolism and parkinson's disease. J Res Med Sci 2013;18(Suppl 1):S39-S42 [FREE Full text] [Medline: 23961283]
40. Abou-Raya S, Helmii M, Abou-Raya A. Bone and mineral metabolism in older adults with parkinson's disease. Age Ageing 2009;38(6):675-680. [doi: 10.1093/ageing/afp137] [Medline: 19684354]
41. Kuan Y, Huang K, Lin C, Hu C, Kao C. Effects of metformin exposure on neurodegenerative diseases in elderly patients with type 2 diabetes mellitus. Prog Neuropsychopharmacol Biol Psychiatry 2017;79(Pt B):77-83. [doi: 10.1016/j.pnpbp.2017.06.002] [Medline: 28583443]
42. Zhang Y, Zhang Y, Shi X, Han J, Lin B, Peng W, et al. Metformin and the risk of neurodegenerative diseases in patients with diabetes: a meta-analysis of population-based cohort studies. Diabet Med 2022;39(6):e14821. [doi: 10.1111/dme.14821] [Medline: 35213749]
43. Kim JY, Ku YS, Kim HJ, Trinh NT, Kim W, Jeong B, et al. Oral diabetes medication and risk of dementia in elderly patients with type 2 diabetes. Diabetes Res Clin Pract 2019;154:116-123. [doi: 10.1016/j.diabres.2019.07.004] [Medline: 31279960]
44. Ajong AB, Kenfack B, Ali IM, Yakum MN, Ukaogo PO, Mangala FN, et al. Calcium supplementation in pregnancy: an analysis of potential determinants in an under-resourced setting. PLoS One 2023;18(10):e0292303 [FREE Full text] [doi: 10.1371/journal.pone.0292303] [Medline: 37796953]
45. Ajong AB, Yakum MN, Aljerf L, Ali IM, Mangala FN, Onydinma UP, et al. Association of hypertension in pregnancy with serum electrolyte disorders in late pregnancy among Cameroonian women. Sci Rep 2023;13(1):20940 [FREE Full text] [doi: 10.1038/s41598-023-47623-6] [Medline: 38017060]

XSL•FO
RenderX

46. Kim HC, Lee H, Lee H, Lee G, Kim E, Song M, Korean Society of Hypertension (KSH)-Hypertension Epidemiology Research Working Group. Korea hypertension fact sheet 2022: analysis of nationwide population-based data with a special focus on hypertension in the elderly. Clin Hypertens 2023;29(1):22 [FREE Full text] [doi: 10.1186/s40885-023-00243-8] [Medline: 37580841]
47. Pasternak B, Svanström H, Nielsen NM, Fugger L, Melbye M, Hviid A. Use of calcium channel blockers and parkinson's disease. Am J Epidemiol 2012;175(7):627-635. [doi: 10.1093/aje/kwr362] [Medline: 22387374]
48. Swart T, Hurley MJ. Calcium channel antagonists as disease-modifying therapy for parkinson's disease: therapeutic rationale and current status. CNS Drugs 2016;30(12):1127-1135 [FREE Full text] [doi: 10.1007/s40263-016-0393-9] [Medline: 27826740]
49. Louis ED, Benito-León J, Bermejo-Pareja F, Neurological Disorders in Central Spain (NEDICES) Study Group. Antihypertensive agents and risk of parkinson's disease, essential tremor and dementia: a population-based prospective study (NEDICES). Neuroepidemiology 2009;33(3):286-292 [FREE Full text] [doi: 10.1159/000235641] [Medline: 19696520]
50. Ayyadevara S, Balasubramaniam M, Kakraba S, Alla R, Mehta JL, Shmookler Reis RJ. Aspirin-mediated acetylation protects against multiple neurodegenerative pathologies by impeding protein aggregation. Antioxid Redox Signal 2017;27(17):1383-1396 [FREE Full text] [doi: 10.1089/ars.2016.6978] [Medline: 28537433]
51. El-Dessouki AM, Galal MA, Awad AS, Zaki HF. Neuroprotective effects of simvastatin and cilostazol in l-methionine-induced vascular dementia in rats. Mol Neurobiol 2017;54(7):5074-5084. [doi: 10.1007/s12035-016-0051-8] [Medline: 27544235]

Abbreviations

ACEi: angiotensin-converting enzyme inhibitor
AD: Alzheimer disease
ALP: alkaline phosphatase
ALT: alanine aminotransferase
ARB: angiotensin II receptor blocker
AST: aspartate aminotransferase
AUROC: area under the receiver operating characteristic curve
CCB: calcium channel blocker
GGT: gamma-glutamyl transferase
HbA1c: glycated hemoglobin
HDL: high-density lipoprotein
ICD-10: International Statistical Classification of Diseases, Tenth Revision
LDL: low-density lipoprotein
ML: machine learning
ND: neurodegenerative disease
OR: odds ratio
PD: Parkinson disease
ROC: receiver operating characteristic
T2DM: type 2 diabetes mellitus
TRIPOD: Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis

Edited by T de Azevedo Cardoso; submitted 31.01.24; peer-reviewed by L Aljerf, J Mistry, S Venkatayogi; comments to author 08.04.24; revised version received 03.06.24; accepted 26.07.24; published 03.10.24.

Please cite as:
Sang H, Lee H, Park J, Kim S, Woo HG, Koyanagi A, Smith L, Lee S, Hwang YC, Park TS, Lim H, Yon DK, Rhee SY
Machine Learning–Based Prediction of Neurodegenerative Disease in Patients With Type 2 Diabetes by Derivation and Validation in 2 Independent Korean Cohorts: Model Development and Validation Study
J Med Internet Res 2024;26:e56922
URL: https://www.jmir.org/2024/1/e56922
doi:10.2196/56922
PMID:39361401

©Hyunji Sang, Hojae Lee, Jaeyu Park, Sunyoung Kim, Ho Geol Woo, Ai Koyanagi, Lee Smith, Sihoon Lee, You-Cheol Hwang, Tae Sun Park, Hyunjung Lim, Dong Keon Yon, Sang Youl Rhee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evolution of the "Internet Plus Health Care" Mode Enabled by Artificial Intelligence: Development and Application of an Outpatient Triage System

Lingrui Yang[1*], MSc; Jiali Pang[2*], MS; Song Zuo[1], ME; Jian Xu[1], ME; Wei Jin[1], BE; Feng Zuo[1], PhD; Kui Xue[2], ME; Zhongzhou Xiao[2], MSc; Xinwei Peng[2], MSc; Jie Xu[2], DHM; Xiaofan Zhang[2], PhD; Ruiyao Chen[2], ME; Shuqing Luo[2], BM; Shaoting Zhang[2*], PhD; Xin Sun[1*], MD

[1]Clinical Research and Innovation Unit, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Xin Sun, MD
Clinical Research and Innovation Unit
Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine
1665 Kongjiang Road
Shanghai
China
Phone: 86 02125077480
Email: sunxin@xinhuamed.com.cn

Abstract

Background: Although new technologies have increased the efficiency and convenience of medical care, patients still struggle to identify specialized outpatient departments in Chinese tertiary hospitals due to a lack of medical knowledge.

Objective: The objective of our study was to develop a precise and subdividable outpatient triage system to improve the experiences and convenience of patient care.

Methods: We collected 395,790 electronic medical records (EMRs) and 500 medical dialogue groups. The EMRs were divided into 3 data sets to design and train the triage model (n=387,876, 98%) and test (n=3957, 1%) and validate (n=3957, 1%) it. The triage system was altered based on the current BERT (Bidirectional Encoder Representations from Transformers) framework and evaluated by recommendation accuracies in Xinhua Hospital using the cancellation rates in 2021 and 2022, from October 29 to December 5. Finally, a prospective observational study containing 306 samples was conducted to compare the system's performance with that of triage nurses, which was evaluated by calculating precision, accuracy, recall of the top 3 recommended departments (recall@3), and time consumption.

Results: With 3957 (1%) records each, the testing and validation data sets achieved an accuracy of 0.8945 and 0.8941, respectively. Implemented in Xinhua Hospital, our triage system could accurately recommend 79 subspecialty departments and reduce the number of registration cancellations from 16,037 (3.83%) of the total 418,714 to 15,338 (3.53%) of the total 434200 ($P<.05$). In comparison to the triage system, the performance of the triage nurses was more accurate (0.9803 vs 0.9153) and precise (0.9213 vs 0.9049) since the system could identify subspecialty departments, whereas triage nurses or even general physicians can only recommend main departments. In addition, our triage system significantly outperformed triage nurses in recall@3 (0.6230 vs 0.5266; $P<.001$) and time consumption (10.11 vs 14.33 seconds; $P<.001$).

Conclusions: The triage system demonstrates high accuracy in outpatient triage of all departments and excels in subspecialty department recommendations, which could decrease the cancellation rate and time consumption. It also improves the efficiency and convenience of clinical care to fulfill better the usage of medical resources, expand hospital effectiveness, and improve patient satisfaction in Chinese tertiary hospitals.

(J Med Internet Res 2024;26:e51711) doi:10.2196/51711

XSL•FO
RenderX

KEYWORDS

artificial intelligence; triage system; all department recommendation; subspecialty department recommendation; “internet plus healthcare”; “internet plus health care”

Introduction

Recent advancements in technology, including the internet, the internet of things, artificial intelligence (AI), 5G, big data, and other emerging technologies, have made it possible to process medical data in order to enhance the effectiveness and convenience of patient treatment [1]. The emergence of “internet plus health care” allows online prehospital registration, diagnosis, and treatment [2-5]. Despite all the breakthroughs in health care technologies, billions of patients flock into tertiary hospitals for medical consultations every year, enduring the “3 long, 1 short” problem in China [6]. Using current online triage systems to help these patients during online registrations is not enough. As current triage systems in China mainly rely on manually maintained department templates, inaccurate department recommendations may occur when the input exceeds their scope or revisions are needed. Meanwhile, patients still struggle to choose the right department due to a lack of medical knowledge and the various specialist outpatient departments in tertiary hospitals [7]. Triage errors can waste time and medical resources, cause unpleasant medical experiences, and even result in misdiagnosis and treatment delays [8,9].

The universal usage of electronic medical records (EMRs) and AI has enabled the processing of medical data for better patient triage before clinical care, but there are still obstacles when applying AI-based triage systems to Chinese health care institutions [10-15]. As most current triage systems are designed for specialized departments, neglecting the numerous subspecialty departments in hospitals, achieving all-department recommendations remains a main challenge because Chinese patients do not rely on referrals before registrations [16]. Another significant challenge is using Chinese-based patient complaints as triage system inputs. Variations in word segmentations between English and Chinese [17,18] and the variability of patient descriptions significantly contribute to this challenge [19,20]. For example, “I feel pain in my head” could also be described as “headache.”

To further improve the “internet plus health care” mode and achieve accurate, efficient, and convenient medical services in China with better usage of clinical resources, this study proposed a precise triage system for Chinese tertiary hospitals that leverages patient-entered chief complaints as input to provide all-department recommendations.

Methods

Study Design

In this study, we designed an AI-based triage system for all-department triage in Chinese tertiary hospitals. All the clinical data were collected from Xinhua Hospital affiliated with the Shanghai Jiao Tong University School of Medicine, a comprehensive tertiary hospital that played a leading role throughout the process of medical digitization in China. We collected 500 medical dialogue groups from an online platform called Spring Rain and used them to design and train the triage system. After implementation of the system, we validated its performance by estimating cancellation rates and conducting a prospective observational study to compare the triage system with experienced triage nurses.

Construction of the Triage System

For the development of the triage system, we conducted a retrospective study and obtained EMRs from Xinhua Hospital, which encompasses 79 departments and specializes in pediatric diseases, with 1,178,694 outpatient visits from January 1 to December 31, 2021. A total of 395,790 (33.58%) EMRs, including chief complaints, present medical histories, and past medical histories, were selected after excluding records that contained physical examinations (n=145,817, 12.37%), vision tests (n=127,390, 10.81%), COVID-19 tests (n=89,657, 7.61%), medication refills (n=31,249, 2.65%), anesthesia records (n=17,231, 1.46%), and records with no extractable symptoms (n=371,560, 31.52%), as shown in Figure 1A. Meanwhile, 500 spoken dialogue groups were collected to extract patients’ oral expressions of symptoms. Sensitive information was removed, and EMRs were deidentified.

Our model was altered based on the current BERT (Bidirectional Encoder Representations from Transformers) [21] framework to enhance model adaptability to Chinese oral expressions. The construction of the triage system can be summarized in 3 steps (Figure 1B): First, a patient description extraction module was designed to extract entities (vital medical events), such as symptoms, anatomical location, and existence from patient descriptions (Multimedia Appendix 1). Second, a patient description normalization module was designed to convert nonstandard symptoms and anatomical locations into standardized medical terminologies according to the International Classification of Diseases, Injuries and Causes of Death, *10th Edition* (ICD-10), the most official code for diseases, symptoms, and anatomical locations acceptable for doctors, which eliminates uncertainty in patient oral expressions (Multimedia Appendix 1). Third, a department recommendation module was designed to recommend departments for patients based on standardized medical concepts, patients’ age, and patients’ gender. (Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Study overview of data processing for model training, the model framework, and the study design to compare the triage system with triage nurses: (A) processing of EMRs, (B) construction of the triage system and implementations, and (C) workflow of comparing the triage system with triage nurses. EMR: electronic medical record; GP: general physician.

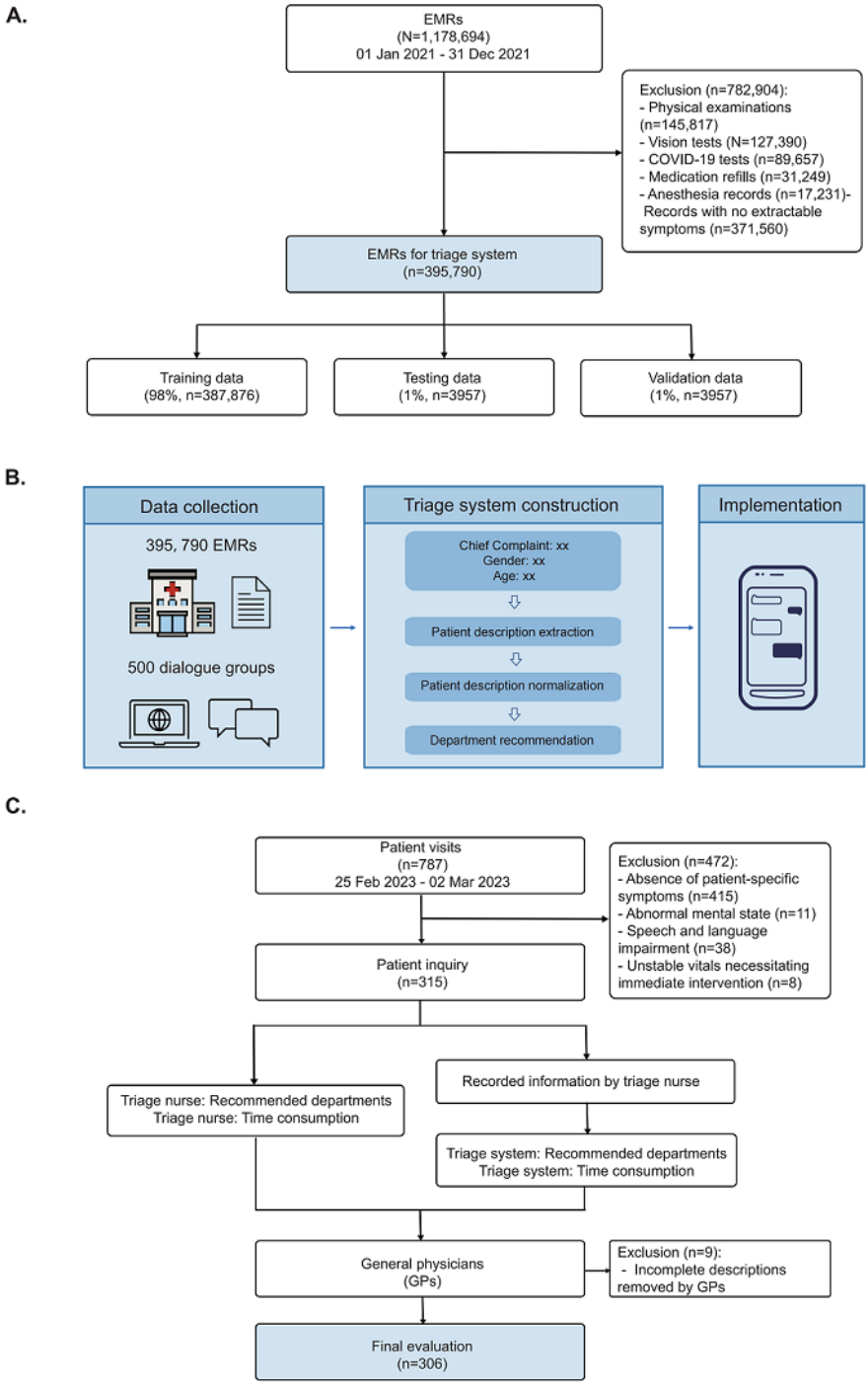

Model Training and Validation

We developed 2 strategies to train the model: first, training the model with all EMRs (original) and then training the model with downsampled chief complaints from EMRs to simulate patient complaints. Chief complaints in EMRs included more symptoms and were more comprehensive than patient descriptions. Specifically, the median number of symptoms in EMRs was 4 (IQR 2-7), while the median number of symptoms in patient oral expressions was 1 (IQR 0-1), as shown in Multimedia Appendix 2. Thus, for each chief complaint in the EMRs, we randomly selected symptoms to reduce symptoms contained in the EMRs. Specifically, for a sentence in chief complaints, if the existence of a certain extracted symptom was "yes," the probability of sampling it was 50%, otherwise the

XSL•FO
RenderX

probability of sampling it was 20% (Multimedia Appendix 3). We split the original and downsampled EMRs into training data (n=387,876, 98%), testing data (n=3957, 1%), and validation data (n=3957, 1%). The 500 dialogue groups were split accordingly and put into the model training stage (patient description extraction module) so that the model could understand the difference between medical terminologies and patient chief complaints to further normalize patient descriptions. Next, we evaluated the performance of our triage system by testing the 2 training strategies on both the original and downsampled testing and validation data sets. We calculated the accuracy of the top 1 and 3 recommended departments to evaluate department recommendation. Precision, recall, and the F_1-score were calculated to evaluate the patient description extraction module (Multimedia Appendix 1). These analyses were conducted using Python 3.8.

System Performance Validated by Cancellation Rate

The triage system was implemented as a mini program on mobile phones at Xinhua Hospital on October 27, 2022. Specifically, the triage system was linked to WeChat through its application programming interface (API), as shown in Multimedia Appendix 1. Records from October 29 to December 5, 2022, were collected to explore the ability of the triage system. The period was chosen to avoid a COVID-19 peak season in Shanghai, which helped reduce bias induced by fever clinics. The cancellation rate of this period was compared with that of the same period in 2021 by a proportional test using R version 4.2.1 (R Foundation for Statistical Computing). We tracked the number of triage system users from January 1 to May 6, 2023, and summarized the trend of user application.

System Performance Compared With Triage Nurses' Performance

A prospective observational study containing 306 (38.8%) samples was conducted at Xinhua Hospital from February 25 to March 2, 2023, to compare the triage system's performance against the triage nurses' performance. In total, 787 patients presented to the consultation center and 315 (40%) were included. Exclusion criteria included the absence of patient-specific symptoms, an abnormal mental state, speech and language impairment, and unstable vitals necessitating immediate intervention. The sample size was decided through the McNemar test. By setting the type I error at 0.05 and statistical power at 0.9, the required sample size was at least 152 (Multimedia Appendix 4).

During the comparison, departments recommended by 2 different triage nurses and the consultation time were recorded by the staff. The collected symptom information was entered into the triage system to obtain system recommendations and consultation time. The triage nurses had over 3 years of clinical practice, and the recommended departments were limited to 3. Next, 2 general physicians (GPs) with over 10 years of clinical practice evaluated the recommended departments based on the case report. A third GP was consulted when inconsistency occurred. The GPs needed to (1) evaluate whether the first recommended department was precise, (2) identify the recommended departments that made errors, and (3) provide supplementary recommended departments. During this process, 9 records were excluded as the case report was incomplete and 306 records were finally used for evaluation (Figure 1C). All GPs were blinded to the groups. The primary outcome was the precision of the first recommended department, the accuracy of the recommended departments, and recall of the top 3 recommended departments (recall@3) [22]. Accuracy was defined as the proportion of correct departments to recommended departments. Recall@3 was defined as the proportion of correct departments in the top 3 recommendations to all correct recommended departments. The calculation formulas are as follows:

The secondary outcome was the efficiency of the triage process, which was evaluated by the time consumption from the beginning of the triage consultation to receiving the recommended results.

We used the Wilcoxon rank-sum test to test differences in precision, accuracy, recall@3, and time consumption between triage nurses and our triage system. In addition, we conducted a subgroup analysis of adults and children (age<18 years) to evaluate system performance within the 2 subgroups. All analyses were conducted using SAS version 9.4 (SAS Institute) and R version 4.2.1. $P<.05$ was considered statistically significant.

Ethical Considerations

This retrospective study involved the collection of only basic patient information, such as gender, symptoms, department, diagnosis, and registration status. No patient samples were collected, and no serious adverse events were expected to occur in the subjects. The confidentiality of patients' identity information and their privacy were guaranteed, and no commercial interest was involved, thus meeting the criteria for exemption from ethical review. The Xinhua Hospital Research Ethics Committee has provided an official statement confirming this exemption (approval number XHEC-D-2024-119). The prospective observational study was approved by the Xinhua Hospital Research Ethics Committee, along with providing and ethical approval for the use of patient information (approval number XHEC-C-2023-013-1).

Results

Characteristics of Data Sets

A total of 395,790 (33.58%) EMRs were included in the study after excluding less meaningful records, such as physical examinations, vision tests, and COVID-19 tests. The median age for the training, testing, and validation data sets was 12 (IQR 4-47). Males accounted for 47.03% (182,420/387,876), 47.74% (1889/3957), and 46.75% (1850/3957) of the training, testing, and validation data sets, respectively, while children accounted for 54.85% (212,740/387,876), 54.76% (2167/3757), and 54.97% (2175/3957), respectively (Table 1).

The EMRs contained 79 subspecialty departments and over 16,000 symptoms (Figure 2). Based on the EMRs, we analyzed the range and frequency of outpatient visits. Overall, the

XSL•FO
RenderX

pediatric department has the most outpatient visits (n=117,667, 29.73%) and subspecialty departments (n=19, 24.1%), as shown in Figure 2 and Multimedia Appendix 5. In the obstetrics and gynecology department, genetics, gynecology, and obstetrics were the top 3 recorded departments, with over 15,000 cases in gynecology and obstetrics. In the surgery department, orthopedics, urology, and cardiothoracic surgery were the top 3 recorded departments. In the internal medicine department, gastroenterology, cardiology, and endocrinology were the top 3 recorded departments (Multimedia Appendix 5).

Table 1. Clinical characteristics of included data sets.

Characteristics	Training data set (n=387,876)	Testing data set (n=3957)	Validation data set (n=3957)	Implementation data set (n=36,683)
Age (years), median (IQR)	12 (4-47)	12 (5-47)	12 (4-46)	21 (5-35)
Gender, n (%)				
Male	182,420 (47.03)	1889 (47.74)	1850 (46.75)	17,777 (48.46)
Female	205,456 (52.97)	2068 (52.26)	2107 (53.25)	18,906 (51.54)
Age group, n (%)				
Children	212,740 (54.85)	2167 (54.76)	2175 (54.97)	17,444 (47.55)
Adults	175,136 (45.15)	1790 (45.24)	1782 (45.03)	19,239 (52.45)

Figure 2. Frequency of departments in EMRs: number of outpatient departments displayed in scale. EMR: electronic medical record; ENT: ear, nose, and throat; MDT: multidisciplinary team.

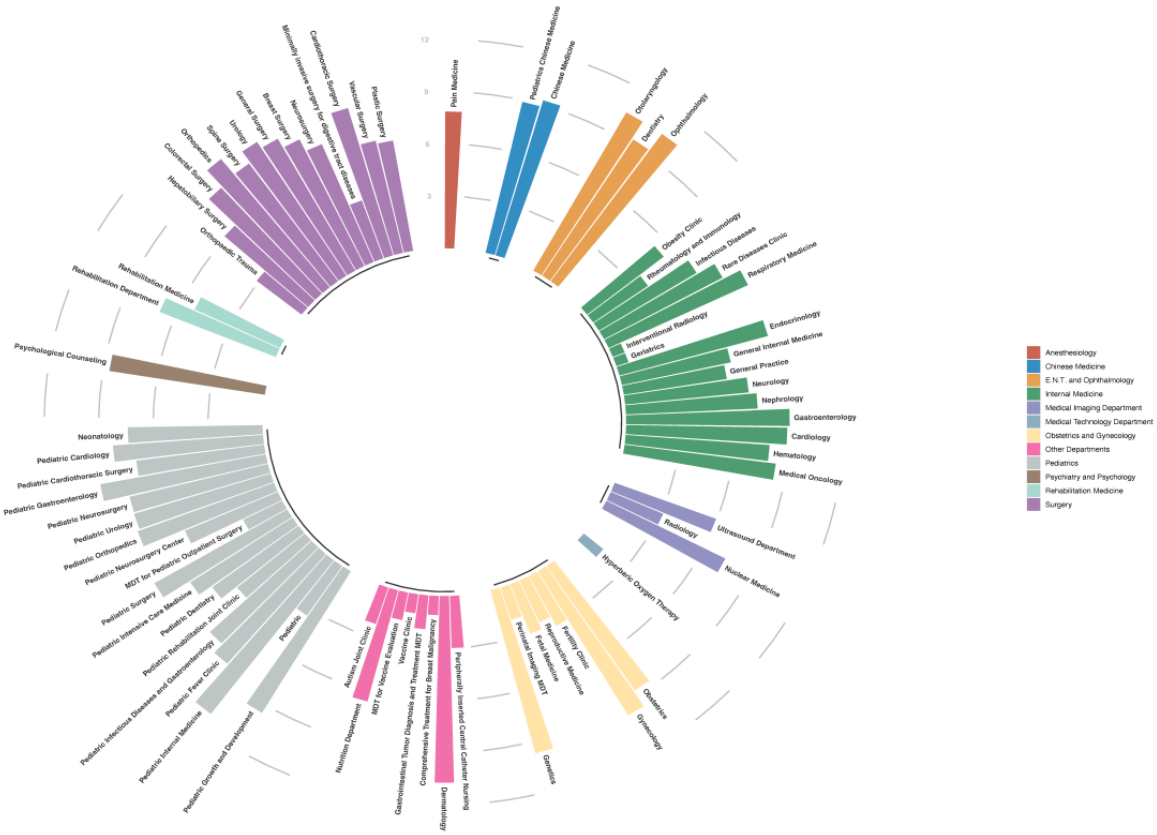

System Performance Validated Using Testing Data and Validation Data

There are 3 modules in the triage system: patient description extraction, patient description normalization, and department recommendation (Figure 1B). We accessed precision, recall, and the F_1-score based on testing data sets from EMRs and dialogues to evaluate whether the patient description extraction module can accurately extract medical events, such as symptoms and anatomical locations. In the EMR testing data set, the extraction module achieved a precision of 0.8420 for extracting symptoms and 0.7768 for extracting anatomical locations (Multimedia Appendix 6). In the dialogue testing data set, the extraction module achieved a precision of 0.6139 for extracting symptoms and 0.6346 for extracting anatomical locations (Multimedia Appendix 6).

XSL•FO
RenderX

Next, we calculated accuracies of top 1 and 3 recommendations when evaluating system performance. Overall, when training the model on the original data, the model achieved higher accuracies in the original testing and validation data sets. However, when training the model on downsampled data, the results were opposite. In this scenario, the top 3 accuracies were 0.8712 and 0.8641 for the original testing and validation data sets, respectively, and 0.8945 and 0.8941 for the downsampled testing and validation data sets, respectively (Table 2).

Table 2. Accuracy (%) of department recommendation in testing and validation data sets.

Data set and groups	Original data-trained model		Downsampled data-trained model	
	First	Top 3	First	Top 3
Testing				
Original	90.79	97.58	72.01	87.12
Downsampled	89.82	97.06	75.82	89.45
Validation				
Original	90.88	97.4	71.29	86.41
Downsampled	89.82	97.07	75.71	89.41

System Performance Validated Using the Cancellation Rate

We implemented the system on the WeChat mini program at the Xinhua Hospital WeChat official account for real-world applications. We observed that the patient cancellation rate reduced from 3.83% (16,037/418,714) in 2021 to 3.53% (15,338/434,200) in 2022 from October 29 to December 5, with $P<.05$. Of note, the age distribution of visits in the application period was similar to that in 2021, with the majority of visits being children (Multimedia Appendix 7). The number of people using the triage system for department recommendations increased from 3590 users in January 2023 to 6953 users in April 2023 according to a review of system usage trends (Multimedia Appendix 8).

System Performance Validated by Comparing With Triage Nurses

In the prospective observational study, we collected 306 samples to compare the performances of our triage system and triage nurses, of whom 149 (48.7%) were adults and 157 (51.3%) were children. After being evaluated by 3 GPs with over 10 years of experience, it was found that there was a relatively high proportion of patients with pediatric fever (n=69, 22.5%), pediatric patients (n=34, 11.1%), and cardiology patients (n=17, 5.6%). Table 3 summarizes the evaluations. Overall, triage nurses had higher accuracy (98.03%) and precision (92.13%) compared to the triage system (accuracy: 91.53%; precision: 90.49%), while our triage system significantly outperformed in recall@3 (62.30% vs 52.66%, $P<.001$) and time consumption (10.11s vs 14.33 seconds, $P<.001$).

After subgrouping by age, we found that regardless of whether the patients were adults or children, the system still performed better than the triage nurses in the recall@3 rate and average time consumption. In the adult subgroup, the system's first recommended department precision was 89.19%, which was higher than the precision of triage nurses (87.84%).

XSL•FO
RenderX

Table 3. Comparison between triage nurses and the triage system.

Groups and subgroups	Precision (%)	Accuracy (%)	Recall@3[a] (%)	Time consumption (seconds)
Triage nurses	92.13	98.03[b]	52.66	14.33[b]
Triage system	90.49	91.53	62.30[c]	10.11
Children (n=157, 51.3%)				
Triage nurses	96.18[c]	99.36[d]	51.72	13.79[b]
Triage system	91.72	93.63	60.04[e]	10.13
Adults (n=149, 48.7%)				
Triage nurses	87.84	96.62[f]	53.66	14.91[b]
Triage system	89.19	89.30	64.68[g]	10.09

[a]Recall@3: recall of the top 3 recommended departments.
[b]$P<.001$.
[c]$P=.031$.
[d]$P=.006$.
[e]$P=.002$.
[f]$P=.013$.
[g]$P=.001$.

Discussion

Principal Findings

This study developed an AI-powered triage system to generate accurate all-department recommendations for online prehospital registrations. Our main finding was that the triage system was able to make accurate outpatient department recommendations among 79 subspecialty departments in Xinhua Hospital and reduce the time consumption of triage.

In the construction of the triage system, its accuracy for the downsampled testing and validation data sets was 89.45% and 89.41%, respectively, as shown in Table 2, similar to the accuracy in the prospective observational study (91.53%). These results indicate that using downsampling EMRs to simulate patient chief complaints is reliable. During the training process, EMRs were used for accurate labels and substantial chief complaints, while medical dialogue groups were used for the conversion of nonstandard medical terminologies. Dialogue groups were important in extracting medical contents from patient descriptions to reduce the discrepancy between oral expressions and medical terminologies, thereby stabilizing the deep learning model's output and enhancing its accuracy. Additionally, unlike previous studies that have focused on directly extracting information from EMRs [23-25], we altered the BERT framework in the patient description extraction module to enhance the algorithm framework's adaptability to Chinese oral expressions [26].

One of the most significant contributions of the triage system is to fulfill all-department recommendations based on patients' chief complaints, which is suitable for Chinese tertiary hospitals. Most current triage systems focus on emergency department triage, hospitalization predictions, or specialized department triage, such as cardiology and oncology [12,27-30]. In addition, a survey revealed that patients complain about current online registration systems for not knowing which department to choose [31]. Our system can identify many subspecialty departments, while triage nurses or even GPs only recommend main departments. For example, when triage nurses recommend general departments such as pediatrics and cardiology, the proposed system will recommend departments such as pediatric fever and cardiovascular internal medicine. By achieving accurate subspecialty department recommendations, the triage system can improve precision prehospital diagnosis and convenient medical services as patients can precisely choose departments at home [7,11,31].

After implementing the triage system at Xinhua Hospital, we found that the cancellation rate decreased from 3.83% to 3.53% from 2021 to 2022 and the number of triage system users increased from 3590 users in January 2023 to 6953 users in April 2023 (Multimedia Appendix 8). Repeated registrations due to false registrations and long waiting times of the offline outpatient process can lead to patient no-shows, which may cause financial losses and underused medical resources [32,33]. The reduction in the cancellation rate and an increasing number of triage system users may help improve efficiency in the outpatient process and achieve better medical resource allocation, thereby expanding the hospital's service efficiency. Moreover, such a smartphone-based triage system may further drive the evolution of the "internet plus health care" mode [7], improving the convenience of patient care. With the wider usage of the internet and less cost of distant medical services, the internet hospital, where patients seek consultations over the internet, can further assist medical consultations from diagnosis to treatment recommendation entirely online [2,34].

Our prospective observational study demonstrated that our triage system is comparable to triage nurses, with 91.53% accuracy and 90.49% precision (Table 3). Additionally, our system

XSL•FO
RenderX

performs better than triage nurses in the recall@3 rate (62.3%) and average time consumption (10.11 seconds). A higher recall@3 rate indicates that our system recommends more relevant departments. Even though we did not include queue waiting times for comparison, our system's time consumption was still 3 seconds less than that of triage nurses. This shows the substantial potential of using triage systems to reduce patient waiting times in premedical consultations and improve medical service efficiency. It is possible that the proposed triage system, when integrated with internet hospitals, will fulfill the role of triage nurses in online hospitals and provide convenient medical services.

Limitations and Future Prospects

Our study has limitations. The first is the limited training data. EMRs were collected from a single center, Xinhua Hospital, in which the departments for pediatric diseases may be more specialized than those in other hospitals. We recognize that department settings vary among hospitals. Meanwhile, with more patient descriptions, our model may perform better for the identification and conversion of symptoms. Second, our triage system is based on the BERT model, which is extensively used for text classification, relation extraction, and entity recognition [26,35,36]. The recent release of Generative Pretrained Transformer 4 (GPT-4) enables inputs as sentences and outputs human-like responses, while BERT can only output word vectors. GPT-4 is better suited for generation tasks, such as medical dialogue generation, generative questions and answers, and automatic medical record filling [37,38]. This may further promote patients' experiences when they would like to ask more. Third, our system is currently in an early promotional phase, and there is an increasing trend of triage system users, so the influence of the triage system on antual convenient medical service is underestimated. As new systems always need time to become familiar to people, we need to promote the system to more people so that we can vividly see the actual benefit of such a system. Future study is need for long-term investigation. Finally, the comparison between the triage system and triage nurses contains inherent bias. During the experiment, triage nurses directly communicated with patients, while the input of the triage system was generated by rephrasing patients' chief complaints. In addition, both triage nurses and GPs recommended main departments. During evaluations, the recommended subspecialty department would be considered right if the recommended subspecialty department by the triage system falls under the main departments recommended by GPs. Therefore, recommending departments is more challenging for the triage system. Despite this unfairness, the system is equivalent to triage nurses and capable of triaging in internet hospitals. In the future, more precise labeling should be performed for the comparison to fairly evaluate the performance of the triage system.

Conclusion

In conclusion, the proposed AI-based triage system can provide more accurate all-department recommendations for a variety of departments, especially subspecialty ones. When implemented in internet hospitals, the system could serve as an effective tool for optimizing the outpatient process with less waiting time. This type of framework holds significant potential for achieving convenient medical service and improving the efficiency of patient care, bridging the gap of the "internet plus health care" mode in China. Although this impact may be seen in a single center now, such successful application may encourage more institutes to collaborate for a comprehensive and universal triage system, fulfilling better usage of medical resources, expanding hospital effectiveness, and providing patients with satisfactory medical services in China.

Acknowledgments

This study was supported by grants from the Hospital Development Strategy Project, Shanghai Jiao Tong University (HDSI-2022-B-004).

We thank Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (Shanghai, China) for providing data and the Shanghai Artificial Intelligence Laboratory (Shanghai, China) for computing resources.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request. Codes for the model are available in Multimedia Appendix 9.

Authors' Contributions

SZ and XS shared corresponding author responsibilities; SZ will respond to technical inquires, and XS will respond to inquires regarding clinical data. LY, JP, SZ, and XS contributed to conceptualization; LY, JP, and KX to methodology; LY, JP, SZ, JX, WJ, and FZ to data curation; LY, JP, KX, ZX, and XP to formal analysis; and LY, JP, ZX, JX, XZ, and SL to investigation; LY and JP wrote the original draft and contributed equally to the work. LY, JP, JX, XZ, SZ, and XS reviewed and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Detailed construction of the triage system algorithm and application examples.
[DOCX File , 1318 KB - jmir_v26i1e51711_app1.docx]

Multimedia Appendix 2
Number of symptoms in EMRs and patient oral expressions. EMR: electronic medical record.
[DOCX File , 18 KB - jmir_v26i1e51711_app2.docx]

Multimedia Appendix 3
Symptom extraction example.
[DOCX File , 17 KB - jmir_v26i1e51711_app3.docx]

Multimedia Appendix 4
Sample size calculation for prospective studies.
[DOCX File , 26 KB - jmir_v26i1e51711_app4.docx]

Multimedia Appendix 5
Outpatient visits and registrations in main departments and subspecialty departments.
[DOCX File , 183 KB - jmir_v26i1e51711_app5.docx]

Multimedia Appendix 6
Performance of patient description extraction in EMR testing data set and dialogue testing data set. EMR: electronic medical record.
[DOCX File , 16 KB - jmir_v26i1e51711_app6.docx]

Multimedia Appendix 7
Age distribution in EMRs and triage system users. EMR: electronic medical record.
[DOCX File , 215 KB - jmir_v26i1e51711_app7.docx]

Multimedia Appendix 8
Time trend of system users.
[DOCX File , 66 KB - jmir_v26i1e51711_app8.docx]

Multimedia Appendix 9
Main source code of the triage system.
[DOCX File , 29 KB - jmir_v26i1e51711_app9.docx]

References

1. Tian S, Yang W, Grange JML, Wang P, Huang W, Ye Z. Smart healthcare: making medical care more intelligent. Glob Health J 2019 Sep;3(3):62-65. [doi: 10.1016/j.glohj.2019.07.001]
2. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015 Aug;3(8):e445-e446 [FREE Full text] [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
3. Cheng T, Fu H, Xu D, Yip W. Technology platforms are revolutionizing health care service delivery in China. NEJM Catalyst 2022 Feb 23;3(1):1-11. [doi: 10.1056/CAT.21.0414]
4. Li X, Krumholz HM, Yip W, Cheng KK, De Maeseneer J, Meng Q, et al. Quality of primary health care in China: challenges and recommendations. Lancet 2020 Jun 06;395(10239):1802-1812 [FREE Full text] [doi: 10.1016/S0140-6736(20)30122-7] [Medline: 32505251]
5. Sun J, Lin Q, Zhao P, Zhang Q, Xu K, Chen H, et al. Reducing waiting time and raising outpatient satisfaction in a Chinese public tertiary general hospital-an interrupted time series study. BMC Public Health 2017 Aug 22;17(1):668 [FREE Full text] [doi: 10.1186/s12889-017-4667-z] [Medline: 28830400]
6. Dai J, Wang X, Ayala FJ. Medical informatics and the “three long, one short” problem of large urban hospitals in China. JAMA 2016 Jul 19;316(3):269-270. [doi: 10.1001/jama.2016.5487]
7. Xie W, Cao X, Dong H, Liu Y. The use of smartphone-based triage to reduce the rate of outpatient error registration: cross-sectional study. JMIR Mhealth Uhealth 2019 Nov 11;7(11):e15313 [FREE Full text] [doi: 10.2196/15313] [Medline: 31710300]
8. Nguyen H, Meczner A, Burslam-Dawe K, Hayhoe B. Triage errors in primary and pre–primary care. J Med Internet Res 2022 Jun 24;24(6):e37209. [doi: 10.2196/37209]

XSL•FO
RenderX

9. Kharbanda AB. Improving appendicitis care for all patients. JAMA Netw Open 2021 Aug 31;4(8):e2124523. [doi: 10.1001/jamanetworkopen.2021.24523]
10. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ Digit Med 2018 Mar 14;1(1):5 [FREE Full text] [doi: 10.1038/s41746-017-0012-2] [Medline: 31304291]
11. Desai AN. Artificial intelligence: promise, pitfalls, and perspective. JAMA 2020 Jun 23;323(24):2448-2449 [FREE Full text] [doi: 10.1001/jama.2020.8737] [Medline: 32492093]
12. Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. JAMA Netw Open 2019 Jan 04;2(1):e186937 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.6937] [Medline: 30646206]
13. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2020 Sep;2(9):e468-e474 [FREE Full text] [doi: 10.1016/S2589-7500(20)30185-0] [Medline: 33328114]
14. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv 2022 Aug 12;8(32):eabq6147 [FREE Full text] [doi: 10.1126/sciadv.abq6147] [Medline: 35960806]
15. Li JO, Liu H, Ting DS, Jeon S, Chan RP, Kim JE, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res 2021 May;82:100900 [FREE Full text] [doi: 10.1016/j.preteyeres.2020.100900] [Medline: 32898686]
16. Mao C, Zhu Q, Chen R, Su W. Medical department classification based on patient describing symptoms. arXiv Preprint posted online 2022. [doi: 10.21203/rs.3.rs-1797893/v1]
17. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc 2014 Sep;21(5):808-814 [FREE Full text] [doi: 10.1136/amiajnl-2013-002381] [Medline: 24347408]
18. Zhang Y, Wang X, Hou Z, Li J. Clinical named entity recognition from Chinese electronic health records via machine learning methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]
19. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. Expert Syst Appl 2019 Feb;116:494-520. [doi: 10.1016/j.eswa.2018.09.034]
20. Jones MM, Somerville C, Feder G, Foster G. Patients' descriptions of angina symptoms: a qualitative study of primary care patients. Br J Gen Pract 2010 Oct;60(579):735-741 [FREE Full text] [doi: 10.3399/bjgp10X532378] [Medline: 20883622]
21. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online 2012. [doi: 10.48550/arXiv.1810.04805]
22. Bui DDA, Jonnalagadda S, Del Fiol G. Automatically finding relevant citations for clinical guideline development. J Biomed Inform 2015 Oct;57:436-445 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.003] [Medline: 26363352]
23. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
24. Davenport TH, Glover WJ. Artificial intelligence and the augmentation of health care decision-making. NEJM Catalyst 2018 Jun 19;4(3). [doi: 10.1056/CAT.18.0151]
25. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern Med 2019 Mar 01;179(3):293-294 [FREE Full text] [doi: 10.1001/jamainternmed.2018.7117] [Medline: 30556825]
26. Xue K, Zhou Y, Ma Z, Ruan T, Zhang H, He P. Fine-tuning BERT for joint entity and relation extraction in Chinese Medical Text. 2019 Presented at: BIBM 2019: IEEE International Conference on Bioinformatics and Biomedicine; November 18-21, 2019; San Diego, CA p. 892-897. [doi: 10.1109/bibm47256.2019.8983370]
27. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019 May 28;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
28. Arnaud, Elbattah M, Gignon M, Dequen G. Deep learning to predict hospitalization at triage: integration of structured data and unstructured text. 2020 Presented at: 2020 IEEE International Conference on Big Data; December 10-13, 2020; Virtual p. 4836-4841. [doi: 10.1109/bigdata50022.2020.9378073]
29. Wolf JA, Moreau JF, Akilov O, Patton T, English JC, Ho J, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. JAMA Dermatol 2013 Apr;149(4):422-426 [FREE Full text] [doi: 10.1001/jamadermatol.2013.2382] [Medline: 23325302]
30. Kass-Hout T, Lee J, Tataris K, Richards CT, Markul E, Weber J, et al. Prehospital comprehensive stroke center vs primary stroke center triage in patients with suspected large vessel occlusion stroke. JAMA Neurol 2021 Oct 01;78(10):1220. [doi: 10.1001/jamaneurol.2021.2485]
31. Yu W, Yu X, Hu H, Duan G, Liu Z, Wang Y. Use of hospital appointment registration systems in China: a survey study. Glob J Health Sci 2013 Jul 22;5(5):193-201 [FREE Full text] [doi: 10.5539/gjhs.v5n5p193] [Medline: 23985121]
32. Marbouh D, Khaleel I, Al Shanqiti K, Al Tamimi M, Simsekler MCE, Ellahham S, et al. Evaluating the impact of patient no-shows on service quality. Risk Manag Healthc Policy 2020;13:509-517 [FREE Full text] [doi: 10.2147/RMHP.S232114] [Medline: 32581613]

XSL•FO
RenderX

33. Xie W, Yang X, Cao X, Liu P. Effects of a comprehensive reservation service for non-emergency registration on appointment registration rate, patient waiting time, patient satisfaction and outpatient volume in a tertiary hospital in China. BMC Health Serv Res 2019 Nov 01;19(1):782 [FREE Full text] [doi: 10.1186/s12913-019-4652-6] [Medline: 31675973]
34. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020 Jul 29;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
35. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
36. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
37. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
38. Arora A, Arora A. The promise of large language models in health care. Lancet 2023 Feb 25;401(10377):641. [doi: 10.1016/S0140-6736(23)00216-7] [Medline: 36841609]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
EMR: electronic medical record
GP: general physician
GPT-4: Generative Pretrained Transformer 4
recall@3: recall of the top 3 recommended departments

Edited by T de Azevedo Cardoso; submitted 10.08.23; peer-reviewed by A Ćirković, K Chen, Y Zhang; comments to author 21.11.23; revised version received 11.01.24; accepted 26.09.24; published 30.10.24.

Please cite as:
Yang L, Pang J, Zuo S, Xu J, Jin W, Zuo F, Xue K, Xiao Z, Peng X, Xu J, Zhang X, Chen R, Luo S, Zhang S, Sun X
Evolution of the "Internet Plus Health Care" Mode Enabled by Artificial Intelligence: Development and Application of an Outpatient Triage System
J Med Internet Res 2024;26:e51711
URL: https://www.jmir.org/2024/1/e51711
doi:10.2196/51711
PMID:

©Lingrui Yang, Jiali Pang, Song Zuo, Jian Xu, Wei Jin, Feng Zuo, Kui Xue, Zhongzhou Xiao, Xinwei Peng, Jie Xu, Xiaofan Zhang, Ruiyao Chen, Shuqing Luo, Shaoting Zhang, Xin Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automated Speech Analysis for Risk Detection of Depression, Anxiety, Insomnia, and Fatigue: Algorithm Development and Validation Study

Rachid Riad[1], PhD; Martin Denais[1], MS; Marc de Gennes[1], PhD; Adrien Lesage[1], PhD; Vincent Oustric[1], PhD; Xuan Nga Cao[1], PhD; Stéphane Mouchabac[2,3], MD; Alexis Bourla[2,3,4,5], MD

[1]Callyope, Paris, France

[2]Department of Psychiatry, Saint-Antoine Hospital, Sorbonne University, Assistance publique - Hôpitaux de Paris, Paris, France

[3]Infrastructure for Clinical Research in Neurosciences, Paris Brain Institute, Paris, France

[4]Medical Strategy and Innovation Department, Clariane, Paris, France

[5]NeuroStim Psychiatry Practice, Paris, France

Corresponding Author:
Rachid Riad, PhD
Callyope
5 Parvis Alan Turing
Paris, 75013
France
Phone: 33 666522141
Email: rachid@callyope.com

Abstract

Background: While speech analysis holds promise for mental health assessment, research often focuses on single symptoms, despite symptom co-occurrences and interactions. In addition, predictive models in mental health do not properly assess the limitations of speech-based systems, such as uncertainty, or fairness for a safe clinical deployment.

Objective: We investigated the predictive potential of mobile-collected speech data for detecting and estimating depression, anxiety, fatigue, and insomnia, focusing on other factors than mere accuracy, in the general population.

Methods: We included 865 healthy adults and recorded their answers regarding their perceived mental and sleep states. We asked how they felt and if they had slept well lately. Clinically validated questionnaires measuring depression, anxiety, insomnia, and fatigue severity were also used. We developed a novel speech and machine learning pipeline involving voice activity detection, feature extraction, and model training. We automatically modeled speech with pretrained deep learning models that were pretrained on a large, open, and free database, and we selected the best one on the validation set. Based on the best speech modeling approach, clinical threshold detection, individual score prediction, model uncertainty estimation, and performance fairness across demographics (age, sex, and education) were evaluated. We used a train-validation-test split for all evaluations: to develop our models, select the best ones, and assess the generalizability of held-out data.

Results: The best model was Whisper M with a max pooling and oversampling method. Our methods achieved good detection performance for all symptoms, depression (Patient Health Questionnaire-9: area under the curve [AUC]=0.76; F_1-score=0.49 and Beck Depression Inventory: AUC=0.78; F_1-score=0.65), anxiety (Generalized Anxiety Disorder 7-item scale: AUC=0.77; F_1-score=0.50), insomnia (Athens Insomnia Scale: AUC=0.73; F_1-score=0.62), and fatigue (Multidimensional Fatigue Inventory total score: AUC=0.68; F_1-score=0.88). The system performed well when it needed to abstain from making predictions, as demonstrated by low abstention rates in depression detection with the Beck Depression Inventory and fatigue, with risk-coverage AUCs below 0.4. Individual symptom scores were accurately predicted (correlations were all significant with Pearson strengths between 0.31 and 0.49). Fairness analysis revealed that models were consistent for sex (average disparity ratio [DR] 0.86, SD 0.13), to a lesser extent for education level (average DR 0.47, SD 0.30), and worse for age groups (average DR 0.33, SD 0.30).

Conclusions: This study demonstrates the potential of speech-based systems for multifaceted mental health assessment in the general population, not only for detecting clinical thresholds but also for estimating their severity. Addressing fairness and incorporating uncertainty estimation with selective classification are key contributions that can enhance the clinical utility and responsible implementation of such systems.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58572) doi:10.2196/58572

KEYWORDS

speech analysis; voice detection; voice analysis; speech biomarkers; speech-based systems; computer-aided diagnosis; mental health symptom detection; machine learning; mental health; fatigue; anxiety; depression

Introduction

Depression and anxiety disorders are recognized as the leading causes of disease burden [1], and their prevalences are high during the entire life span across the sexes and all around the globe [2]. This burden was aggravated by the COVID-19 pandemic [3]. In these disorders, early identification and evaluation of the severity of the symptoms are of prime importance since the incidence of suicide is associated with a diagnosis of depression more than 50% of the time [4]. Besides, measurement-based care, via the use of clinically valid scales, improves the follow-up and treatment of affected individuals with mental health disorders [5]. Timely interventions lead to better outcomes in mental health. This proactive approach can ensure early access to treatment and prevent significant complications. Yet, measuring mental health remains a challenge, since manifestations of depression and anxiety are heterogeneous [6] and co-occur with insomnia [7] and fatigue [8]. The exhaustive and objective assessment of these different mental health dimensions through validated assessment scales is long and fastidious for clinical staff and is particularly not adapted to primary care, which is at the forefront of handling mental health disorders [9]. The development of objective biomarkers, which are easy to collect without the synchronization of clinicians and patients, has the potential to overcome these limitations. These quantifiable measures, encompassing biological, genetic, or behavioral assessments, could revolutionize early detection, enabling timely and targeted interventions that ultimately improve the patient's outcomes and well-being. This is particularly significant for screening the general population across diverse mental health dimensions. Indeed, screening the first signs of mental health problems or symptoms could help to avoid escalation of symptoms, as different dimensions interact in time [7,10].

The study of speech biomarkers in mental health holds great potential, offering a noninvasive and easily accessible avenue to capture significant motor, cognitive, and behavioral changes due to mental health disorders such as depression [11-14]. Clinical evidence and research studies have increasingly linked specific automated extracted speech features, such as prosody, articulation, and fluency, with various mental health conditions, such as depression [11,15], anxiety [16], suicide-risk assessment [17], fatigue [18,19], or sleep deprivation [20]. The complexity of human speech extends beyond the intricate motor coordination involved. The speech production system within the brain relies on the synchronization of diverse cognitive, social, and motor processes [21,22]. This intricate interplay involves hundreds of muscles across the respiratory, phonatory, and supralaryngeal systems, working in concert with critical cognitive skills like attention, memory, and planning. Additionally, social skills such as theory of mind and emotional processing play a vital role. Importantly, disruptions in any of the aforementioned motor, cognitive, or social skills, as well as mental health states, can introduce perturbations in the resulting speech signal. Besides, beyond research evidence, clinical practitioners also use voice unconsciously when evaluating individuals, and these subjective evaluations could be complemented and refined with objective measures from automatic speech analysis.

Speech biomarkers emerge also as a promising avenue for mental health assessment due to their unique characteristics: they are noninvasive, cost-effective, and convenient tools. Recent hardware and software advancements have significantly simplified and reduced the cost of acquiring acoustic data, making it a more accessible option compared to traditional biological, imaging, or cognitive markers. In addition, speech data collection requires minimal effort from both patients and clinicians and can even be conducted remotely, further enhancing its feasibility in various settings.

However, despite its promises, the study of speech biomarkers remains largely fragmented, in laboratory settings, or not evaluated for deployment into clinical practice. This gap calls for more evidence to be integrated into clinical practice [23]. Research on speech in mental health in the general population often focuses on 1 isolated mental health dimension, even though there are proofs supporting the existence of networks of symptoms and syndromes in mental health that influence each other [24,25]. In addition, previous speech studies were limited to specific populations such as students [26] or older people [27]. However, before machine learning (ML) models can be used in clinical settings to make predictions for individuals, they must be "fair"—providing equally accurate predictions across all demographic groups [28]. Finally, speech-based systems should not be tested only for the simple classification of binary labels (eg, depressed or not depressed) but rather for the estimation of the severity of symptoms [29], and their ability to refrain from giving an output when uncertainty is too high, therefore deferring decisions to the health staff in practice [30].

In this study, the main objective was to assess the predictive potential of speech models in detecting and estimating the severity of depression, anxiety, fatigue, and insomnia within the general population using mobile-collected speech and mental health data. Besides, to prove that these models could be effectively implemented in diverse real-world settings, they are assessed for their fairness and uncertainty capabilities.

Methods

Participants

We recruited French healthy adult participants without any known severe psychiatric or neurological disorder (self-declaration) or speech pathologies such as stuttering or clutter.

XSL•FO
RenderX

Ethical Considerations

All participants signed an informed consent form to participate in the study, in line with the Declaration of Helsinki, current Good Clinical Practice guidelines, and local laws and regulations. All procedures were approved by the French National Institutional Review Board (identifier 23.00748.OOO2L7#I for the Committee for the Protection of Persons). All data were stored on secure health data servers, without any identifying information in the metadata. Participants received a €15 gift card as compensation for their time.

Study Procedure

Overview

The participants completed the protocol on smartphones through the Callyope research mobile app in a home environment. The participants completed self-assessment scales for different mental health dimensions and recorded different speech tasks. In this work, we only focused on 1 spontaneous and semistructured speech task where participants had to answer "Describe how you are feeling at the moment and how your nights' sleep have been lately" [31]. The participants were included by speech pathologist interns and recruited through social media platforms. Finally, self-reported symptoms were examined with clinically validated questionnaires (Figure 1A). Participants followed the instructions displayed on the Callyope app, and their vocal answers were recorded with the smartphone's microphone. The audio was sampled at 44.1 kHz with a 16-bit resolution. Each participant was asked to place his phone on a flat surface (eg, a table) with the microphone pointing toward the speaker, that is, himself. The session should take place in a quiet environment, whenever possible.

Figure 1. Overview of this study. (A) Overview of our Callyope General-Population (Callyope-GP) dataset with 865 included participants. (B) Flowchart of the pretraining phases of our speech encoders with an illustration of the pretraining speaker embedding process. (C) Graphical illustration of our speech and machine learning (ML) pipeline for a single individual. The pretrained speaker embedding is frozen, and only ML models on top are trained on training data. (D) Evaluation and comparison between true and predicted assessment on held-out participants (test set n=131). Si represents the speech turn vector embeddings obtained from the speech encoder model, colors represent speaker identities for each speech turn and embedding, and wi represents the words spoken in the given audio.

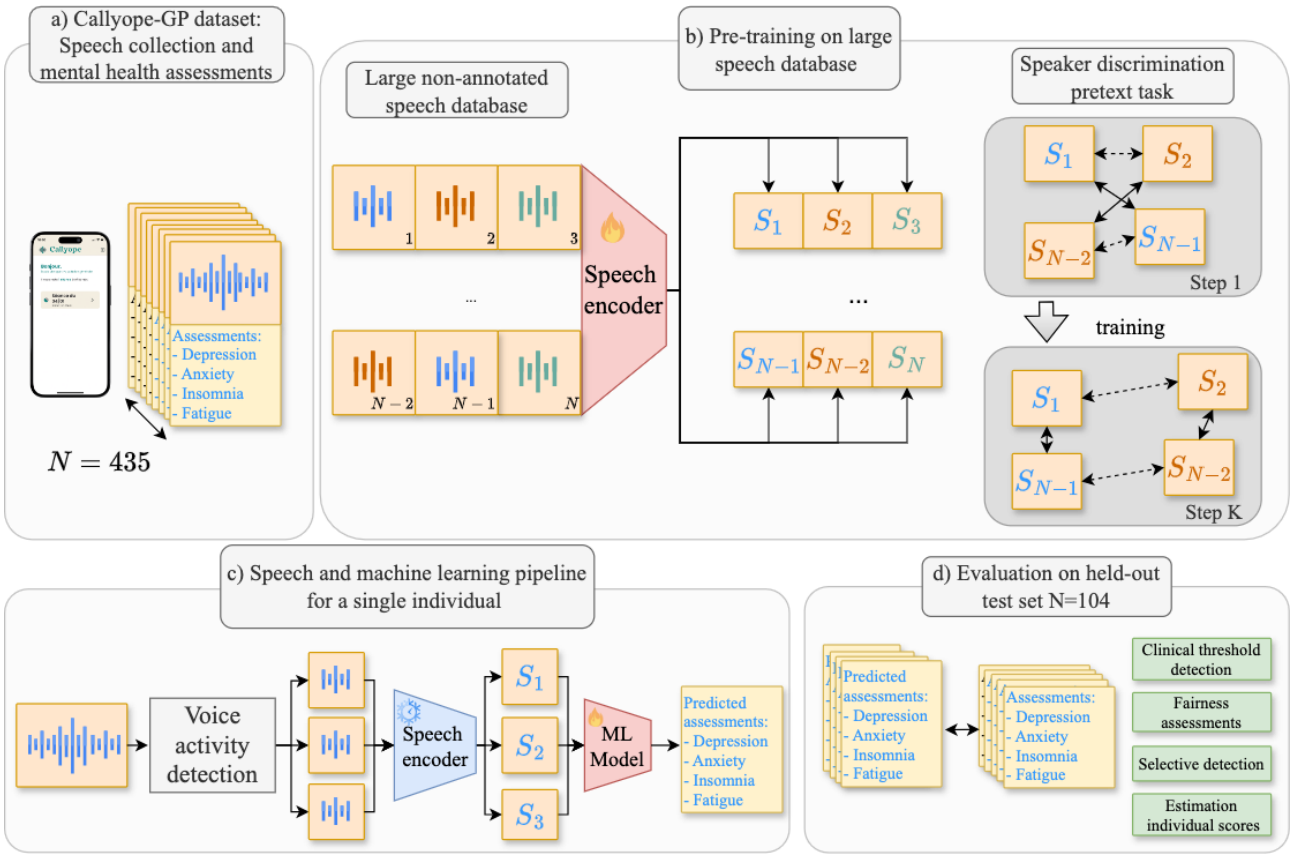

We refer to the dataset collected in this study as the Callyope General-Population (Callyope-GP) dataset. We split randomly the Callyope-GP dataset into 3 sets: training, validation, and testing. Demographic data, such as sex, age, and education level, were collected. We compared groups with adequate tests for their demographics and self-assessments to ensure that groups were consistent.

Measures of Depressive Symptoms

To allow broader use of our solutions, the severity of depression was assessed through the Beck Depression Inventory (BDI) [32] and Patient Health Questionnaire-9 (PHQ-9) [33] self-report questionnaires. In current clinical practices, it is common that different professionals interacting with a patient use different metrics to monitor depression. While depression assessment through these 2 measures exhibits a robust correlation at the group level [34], thus facilitating the development of an equational conversion for research uses, their limited efficacy

XSL•FO
RenderX

at the individual level impedes their reliable conversions to predict individual depressive status [35].

The PHQ-9 is a short, self-administered questionnaire mainly used to screen and measure the severity of depression [33] and is sensitive to potential changes [34]. It includes the 2 cardinal signs of depression: anhedonia and depressed mood. We considered a risk of depression if the total score for the PHQ-9 was more than 10 (PHQ-9≥10).

The BDI is a self-administered questionnaire with 21 items, each centered around a core theme [32,36]. Respondents are presented with statements for each item, and they are instructed to choose 1 statement, which is then associated with a score ranging from 0 to 3. The cumulative score for the scale can reach a maximum of 63 points. We considered the BDI threshold to be positive if the total score was more than 10 (BDI≥10), as it is above the normal range as defined by the authors of the BDI [36].

Measure of Anxiety

The Generalized Anxiety Disorder 7-item scale (GAD-7) questionnaire is to measure or assess the severity of GAD [37]. This is a self-administered questionnaire that takes less than 5 minutes to complete, and it was especially developed to be deployed efficiently in primary care. The optimal cutoff for the GAD-7 was found to be a cutoff for the total score of GAD-7≥10 [37].

Measure of Insomnia

The Athens Insomnia Scale (AIS) is a self-administered questionnaire to assess the patient's sleep difficulties according to the *International Classification of Diseases, Tenth Revision* (ICD-10) criteria [38]. The AIS-8 comprises 8 items (5 minutes) and is a good tool for general sleep assessment and insomnia screening and to measure the intensity of sleep-related problems but also as a screening tool in reliably establishing the diagnosis of insomnia. The optimal cutoff for diagnosis to detect insomnia troubles, for the AIS scale, is 6 [39,40].

Measures of Fatigue

We used the Multidimensional Fatigue Inventory (MFI) to assess the different dimensions of fatigue [41-43]. It is a short self-report questionnaire (5-10 minutes) based on 20 questions to determine 5 dimensions of fatigue: general fatigue, physical fatigue, reduced motivation, reduced activity, and mental fatigue. We also reported the total fatigue score as the sum of all subcomponents.

We used the normative data from Schwarz et al [42] and Hinz et al [44] to choose thresholds for each subcomponent. Individual subcomponents of fatigue in the 75% quantile in the studied populations are all above 10. Therefore, we aimed to predict individuals' scores, which are above or equal to 10, for each dimension. As mentioned also in Schwarz et al [42], the total score has clinical significance and validity, as it was observed to have the highest correlations with anxiety, depression, and quality of life. There is no consensus cutoff for the total sum fatigue score; yet, based on the Colombian normative data [44], we observed that the mean values for each studied subgroup were all above 40; therefore, we chose a clinical threshold of 40 for the total sum score.

ML Analyses

Overview

Our ML analyses can be decomposed into three main steps: (1) the pretraining of the speech encoder model (Figure 1B) (2) the fine-tuning of ML models for each mental health aspect considered in this study (Figure 1C), and (3) extensive evaluations of the clinical threshold detection, selective detection, fairness assessments, and severity estimations for each clinical scale (Figures 1D and 2).

Model Pretraining and Tuning

Audio intensity is normalized per sample, and we compared the three main approaches to obtain representations for large-scale speech models: (1) Speaker recognition is performed using a ThinResNet model with 34 layers. The model takes speech samples as input, which are encoded as 40-Mel spectrograms, with a hop length of 10 ms and a Hamming window. This architecture is based on the ResNet design introduced by He et al [45]. (2) We also considered a transformer [46] architecture adapted for speech (HuBERT [47]), trained in a self-supervision fashion, that is, to predict masked neighbor embeddings. (3) Finally, we evaluated a transformer architecture Whisper pretrained to tackle automatic speech recognition.

Speaker recognition as a pretraining task has proven great results in mental health and neurology [12,48]. The ThinResNetis pretrained on the VoxCeleb2 dataset [49], which is publicly available and contains over 1 million utterances from 6112 speakers, from 145 nationalities. The VoxCeleb2 dataset consists of almost only continuous speech. The pretraining learning forces the model to organize speech in terms of speaker characteristics, as we illustrated earlier in the right panel of Figure 1B. We used an additive-margin softmax loss for this speaker identification task [50].

We also compared these models to a self-supervised model, HuBERTXL, which exhibits great generalization for paralinguistic tasks such as emotion recognition [51]. HuBERT was trained on 960 hours of Librispeech [52]. Whisper is a recent robust automatic speech recognition system based on a transformer architecture trained on 680,000 hours of transcribed speech. Whisper training data are much bigger and more diverse with noisy labels than other speech models. We considered 3 versions of the model: small, medium, and large.

In this work, we did not fine-tune any of the speech encoder models on the Callyope-GP dataset, which we represented in Figure 1C by a frozen speech encoder model. For each speaker i, we obtained a vector representation, a speech vector embedding denoted S_i We extracted segments of 20 seconds with 10 seconds overlap, and we compared different pooling of predictions with mean and max pooling. Besides, as our data are imbalanced, we also compared classic sampling of examples to train models with undersampling of majority class and oversampling of minority classes. The default values of undersampling and oversampling were used. We found no differences with and without voice activity detection, so we

XSL•FO
RenderX

used windowing for simplicity. Extraction was performed using Python (version 3.9; Python Software Foundation), and the following packages were used to extract the acoustic features: *pytorch 2.0.1*, *imbalanced learn*, *torchaudio 2.0.2*, and *voxceleb_trainer project* [53].

For the fine-tuning of each task and each clinical score, different ML algorithms were compared on the validation set. For each task, once a model was selected, we retrained this final model on the concatenation of the training and validation sets and tested on the held-out test to avoid any inflated results. We used the scikit-learn implementation of each algorithm, splitting and evaluation [54].

For the speech collected in the Callyope-GP dataset, we applied the frozen speech encoder to each speech turn and propagated mental health assessment labels at the speech turn level to train and compare the final ML model. At inference, for final evaluation, we pool predictions at the speaker level, after we obtained varying speech turns from a specific speaker. We illustrated in Figure 2 each clinical end point, translated as an ML task based on the aforementioned procedure.

Figure 2. Schematic representation of the clinical tasks that are being assessed in this study. Each task is different in terms of the set and types of outputs. The different tasks were illustrated with the BDI clinical scale and its given threshold of 10. BDI: Beck Depression Inventory; ML: machine learning.

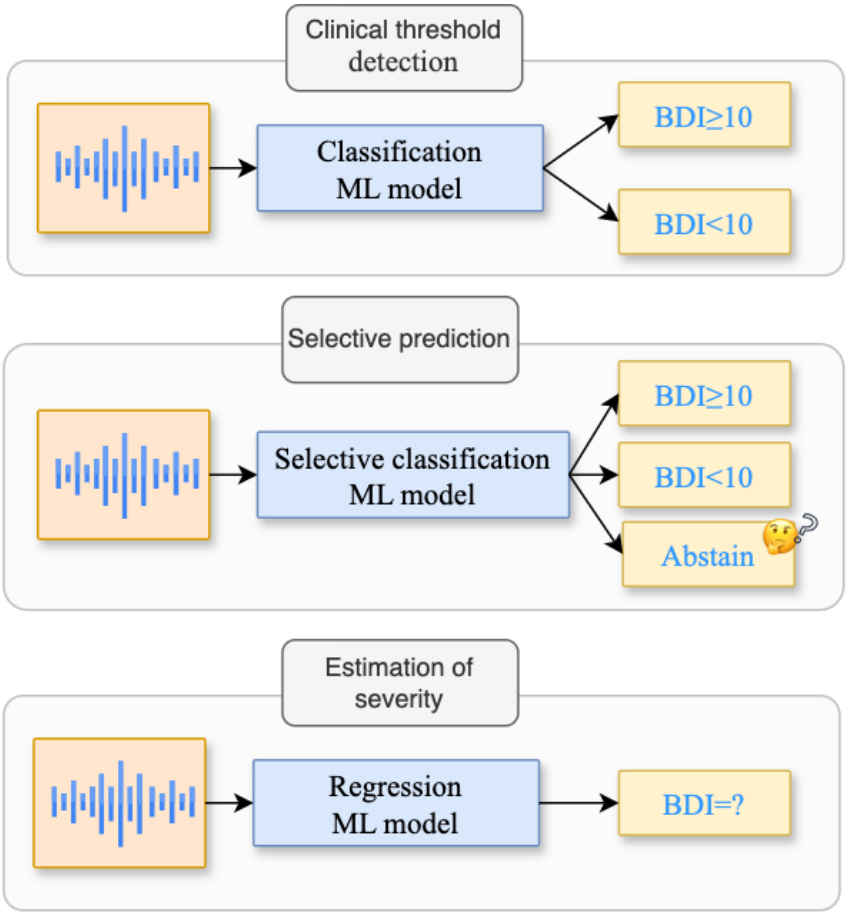

Clinical Threshold Detection (Classification)

We first compared the predictive power of the speech encoder to discriminate between individuals who are below or above the threshold for each clinical scale. The distributions of positive and negative labels vary across clinical dimensions, and to take into account imbalance, we reported the performances of the macro F_1-scores along the area under the curve (AUC) of the receiver operating characteristic on the test set.

We compared linear-based models (logistic regression for classification with L2 regularization and elastic net linear model for regression), tree-based models (random forests with 100 estimators), and gradient-boosting algorithms (histogram-based gradient boosting). Even though the pretraining phase captured information about the participants' mental health, it is important to build a final model for each mental health dimension to be more specific and more sensitive. In ML terms, the mental health characteristics in speech are not necessarily linearly separable in the last vector space of the speech encoder.

Estimation of Severity Through Predictions of Individual Scores (Regression)

The conventional approach in mental health assessment through speech analysis typically focuses on the group's statistical analyses or binary classifications of categorical outcomes, primarily discerning the presence or absence of a specific dimension. Yet, the risks of depression, anxiety, fatigue, and insomnia exist along a spectrum of severity levels that exert varying degrees of influence on an individual's well-being.

We go beyond traditional prediction and categorizations by integrating the estimation of severity through predictions of individual scores using regression ML models. This offers a

XSL•FO
RenderX

more nuanced and comprehensive understanding of mental health dynamics, allowing for a more refined and personalized assessment. We evaluated our estimation of the severity of each total score with the mean absolute error (MAE) between actual and predicted scores and Pearson correlations. MAE score directly measures how close the predicted scores are to the actual scores without considering the direction of error. We also reported the Pearson correlation and the *P* value between the actual test set and the predicted values.

Fairness Assessments: Quality of Services for Sex, Age, and Education Level Demographics

ML systems can behave unfairly for different reasons and in multiple ways [28]. In medicine, the use of ML and predictive models should be carefully evaluated, especially for potential quality-of-service harms, that is, it can occur when a system is not as performant for 1 specific group of people as it is for another group. We conducted a thorough analysis of the quality of service regarding potential harms for each clinical scale in the final predicted model across every dimension: sex, age, and education level. The disparity ratio (DR) was reported based on clinical threshold detection F_1-scores [55] to consider both false positives and false negatives and to be more stringent than equality of opportunity. The DR is computed as the fraction of the minimum F_1-score across subgroups divided by the average F_1-score on the full test set:

We did the same for Pearson correlation for regression. The higher the DR, the better it is, as it means that the model performs equally well across groups with the perfect DR being 1, that is, each group has the same level of performance. The *fairlearn* toolkit was used to perform our fairness evaluations [56].

Selective Clinical Threshold Detection (Selective Prediction)

ML approaches have made great strides in several domains; yet, apps to high-stakes settings remain challenging. In our case, in mental health assessments, communicating appropriately the uncertainty associated with the system predictions is critical [57]. Yet, the communication of probabilities to human users is hard [58], and a pragmatic approach is to determine if an artificial intelligence system is more likely to make erroneous predictions and defer these cases to clinicians. This approach can be viewed as a selective prediction task, where the ML system has the ability to withhold a prediction when it is too uncertain (essentially, the model saying “I don’t know”) [59,60]. In this work, we followed the method from Hendrycks and Gimpel [61], and we used the maximum output probabilities of the ML classification system as a way to measure uncertainty.

Based on a moving threshold, we can obtain a specific ML system to choose to abstain when its output probabilities are too low. This specific ML system is evaluated based on the predictions it chooses to make only; thus, there is a specific coverage and a specific accuracy or risk. There is a natural tradeoff between the coverage of the ML system and its accuracy. Therefore, the way to evaluate a selective prediction task is the AUC for the risk-coverage curve.

Results

Data Overview and Demographics of Participants

A total of 1150 participants were eligible and agreed to participate in our study. Among them, 865 completed the study, giving a recruitment yield of 75.2%. There was an equal split between Android (n=475, 54.9%) and iOS (n=390, 45.1%) devices used by participants. The reasons for which participants were not included in this analysis were the following: one missing speech task, missing demographic information, or one missing answer in the self-report questionnaires.

For our analyses, 605 participants were in the training set, 129 were in the validation set, and 131 were in the test set, and these groups did not differ in terms of demographics and mental health evaluations (Table 1). This yields a dataset sufficient (n>500) to evaluate error bars and predictive algorithms to avoid over-optimistic results [62]. Among the 865 participants, 275 (31.8%) were above the BDI screening threshold, 146 (16.9%) were above the PHQ-9 threshold, 133 (15.3%) were above the GAD-7 threshold, 371 (42.5%) were above the AIS threshold, 489 (56.5%) were above the MFI general fatigue threshold, 325 (37.5%) were above the MFI physical fatigue threshold, 283 (32.7%) were above the MFI reduced activity threshold, 379 (43.8%) were above the MFI mental fatigue threshold, 209 (24.1%) were above the MFI reduced motivation threshold, and 557 (64.4%) were above the threshold of the total score of the MFI. We reported a co-occurrence matrix of people at risk for each dimension in Figure S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of the training, validation, and test groups[a].

Characteristics[b]	Training (n=605)	Validation (n=129)	Test (n=131)	Group comparisons	
				F test (*df*)	*P* value
Demographics					
Sex, n (%)				9.0[c] (4)	.06
Female	386 (76.4)	85 (65.9)	69 (52.2)		
Male	214 (35.3)	44 (34.1)	63 (47.7)		
Other	5 (0.8)	0 (0)	0 (0)		
Education level, n (%)				5.2[c] (6)	.52
No diploma	19 (3.1)	7 (5.4)	8 (6.1)		
Secondary studies	107 (17.7)	21 (16.2)	24 (18.3)		
Short postbaccalaureate studies (Bac+2)	71 (11.7)	20 (15.5)	16 (12.2)		
Long postbaccalaureate studies (Bac+3 and above)	408 (67.4)	81 (62.8)	83 (63.3)		
Age (years)				0.78 (2)	.45
Mean (SD)	38.8 (18.2)	40.2 (18.7)	40.1 (20.0)		
Range	18-92	18-86	18-89		
Clinical evaluation					
BDI[d]				0.07 (2)	.92
Mean (SD)	7.7 (7.3)	7.9 (7.2)	7.9 (7.7)		
Range	0.0-32.0	0.0-29.0	0.0-36.0		
Negative, n (%)	413 (68.3)	88 (68.2)	89 (67.9)		
Positive[e], n (%)	192 (31.7)	41 (31.8)	42 (32.1)		
PHQ-9[f]				0.85 (2)	.42
Mean (SD)	5.6 (4.5)	5.2 (4.7)	5.1 (4.4)		
Range	0.0-23.0	0.0-25.0	0.0-22.0		
Negative, n (%)	499 (82.5)	108 (83.7)	112 (85.5)		
Positive[e], n (%)	106 (17.5)	21 (16.3)	19 (14.5)		
GAD-7[g]				0.50 (2)	.60
Mean (SD)	5.0 (4.5)	4.8 (4.7)	5.4 (5.0)		
Range	0.0-21.0	0.0-20.0	0.0-21.0		
Negative, n (%)	511 (84.5)	109 (84.5)	112 (85.5)		
Positive[e], n (%)	94 (15.5)	20 (15.5)	19 (14.5)		
AIS[h]				1.6 (2)	.19
Mean (SD)	5.6 (3.9)	5.0 (3.1)	5.2 (4.1)		
Range	0.0-24.0	0.0-16.0	0.0-19.0		
Negative, n (%)	338 (55.9)	75 (58.1)	81 (61.8)		
Positive[e], n (%)	267 (44.1)	54 (41.8)	50 (38.2)		
MFI[i] general fatigue				1.8 (2)	.17
Mean (SD)	10.4 (4.2)	11.1 (4.2)	10.4 (4.1)		
Range	4.0-20.0	4.0-20.0	4.0-20.0		

XSL•FO
RenderX

Characteristics[b]	Training (n=605)	Validation (n=129)	Test (n=131)	Group comparisons	
				F test (*df*)	*P* value
Negative, n (%)	335 (55.4)	80 (62)	74 (56.5)		
Positive[e], n (%)	270 (44.6)	49 (38)	57 (43.5)		
MFI physical fatigue				0.55 (2)	0.57
Mean (SD)	8.7 (3.9)	8.8 (4.0)	8.4 (3.9)		
Range	4.0-20.0	4.0-20.0	4.0-20.0		
Negative, n (%)	376 (62.1)	78 (60.5)	86 (65.6)		
Positive[e], n (%)	229 (37.9)	51 (39.5)	45 (34.4)		
MFI reduced activity				0.92 (2)	.40
Mean (SD)	8.3 (3.8)	7.9 (3.5)	8.1 (3.6)		
Range	3.0-20.0	4.0-20.0	4.0-18.0		
Negative, n (%)	400 (66.1)	92 (71.3)	90 (68.7)		
Positive[e], n (%)	205 (33.9)	37 (28.7)	41 (31.3)		
MFI mental fatigue				0.26 (2)	.77
Mean (SD)	9.2 (4.2)	9.5 (4.4)	9.3 (4.2)		
Range	4.0-20.0	4.0-20.0	4.0-20.0		
Negative, n (%)	340 (56.2)	69 (53.5)	77 (58.8)		
Positive[e], n (%)	265 (43.8)	60 (46.5)	54 (41.2)		
MFI reduced motivation				0.74 (2)	.47
Mean (SD)	7.5 (3.0)	7.2 (3.2)	7.7 (3.2)		
Range	4.0-20.0	4.0-20.0	4.0-18.0		
Negative, n (%)	458 (75.7)	102 (79.1)	96 (73.3)		
Positive[e], n (%)	147 (24.3)	27 (20.9)	35 (26.7)		
MFI total score				0.09 (2)	.91
Mean (SD)	44.2 (15.1)	44.6 (15.0)	43.9 (15.1)		
Range	19.0-99.0	20.0-90.0	20.0-83.0		
Negative, n (%)	338 (55.9)	79 (61.2)	77 (58.8)		
Positive[e], n (%)	267 (44.1)	50 (38.8)	54 (41.2)		

[a]Categorical variables are compared with Pearson chi-square test, and continuous variables are compared with 1-way ANOVA.
[b]Characteristics for the 3 splits of the dataset to ensure generalization.
[c]Chi-square test.
[d]BDI: Beck Depression Inventory.
[e]In addition to statistics for each clinical scale score, we report the participants below and above the cutoff (see the *Methods* section for each threshold).
[f]PHQ-9: Patient Health Questionnaire-9.
[g]GAD-7: General Anxiety Disorder 7-item scale.
[h]AIS: Athens Insomnia Scale.
[i]MFI: Multidimensional Fatigue Inventory.

ML Analyses

Overview

Models and pipelines were compared with the results on the validation set (Table S1 in Multimedia Appendix 1). Overall, all the Whisper models outperformed other approaches (Whisper M being the best with mean F_1-score=0.56, SD 0.09), while the speaker model performed the worst, even though still performing well. We found out that the performance of the pooling with the maximum prediction always outperformed the mean pooling, and undersampling and oversampling were helping. It was found out that the linear-based models were outperforming random forest and gradient boosting on the validation set. Thus, we retrained linear algorithms with max pooling and the Whisper

XSL•FO
RenderX

M frozen speech encoder on the combination of the training and validation sets and reported the final results on the held-out test set in Figures 3 and 4 and Tables 2-4.

Figure 3. ROC curves for the clinical threshold detection task on the held-out test set. (A) ROC curves to detect clinically relevant thresholds for depression (PHQ-9 and BDI), anxiety (GAD-7), and insomnia (AIS). (B) ROC curves to detect clinically relevant thresholds for fatigue components (MFI), general fatigue, physical fatigue, reduced activity, mental fatigue, reduced motivation, and total fatigue. AUC: area under the curve; AIS: Athens Insomnia Scale; BDI: Beck Depression Inventory; GAD-7: General Anxiety Disorder 7-item scale; MFI: Multidimensional Fatigue Inventory; PHQ-9: Patient Health Questionnaire-9; ROC: receiver operating characteristic curve.

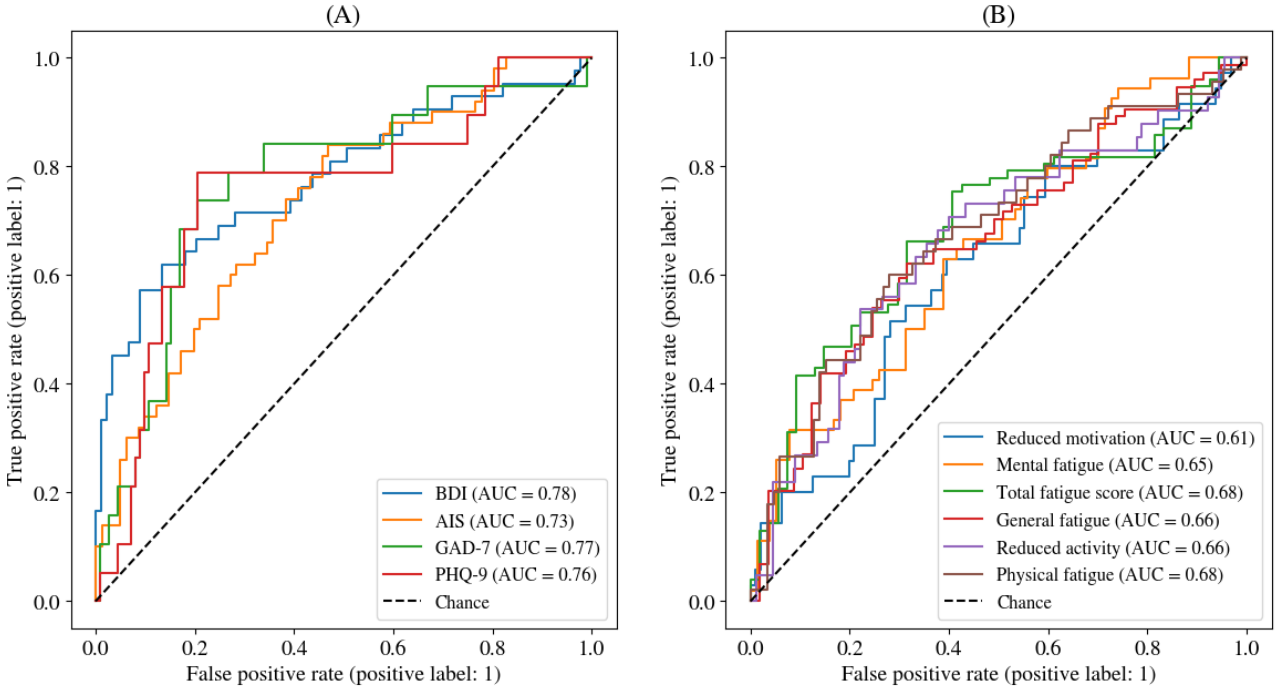

Figure 4. Risk-coverage curves for the selective clinical threshold detection task on the held-out test set illustrate the models' selective screening ability, that is, risk detection capabilities with the ability to abstain when too uncertain. Curves are smoothed for clarity with a Gaussian blur but not used to compute AUC. A lower AUC is better, and 0 is the perfect score. (A) Risk-coverage curves selectively detect clinically relevant thresholds for depression (PHQ-9 and BDI), anxiety (GAD-7), and insomnia (AIS). (B) Risk-coverage curves to detect clinically relevant thresholds for fatigue components (MFI), general fatigue, physical fatigue, reduced activity, mental fatigue, reduced motivation, and total fatigue. AIS: Athens Insomnia Scale; AUC: area under the curve; BDI: Beck Depression Inventory; GAD-7: General Anxiety Disorder 7-item scale, MFI: Multidimensional Fatigue Inventory; PHQ-9: Patient Health Questionnaire-9.

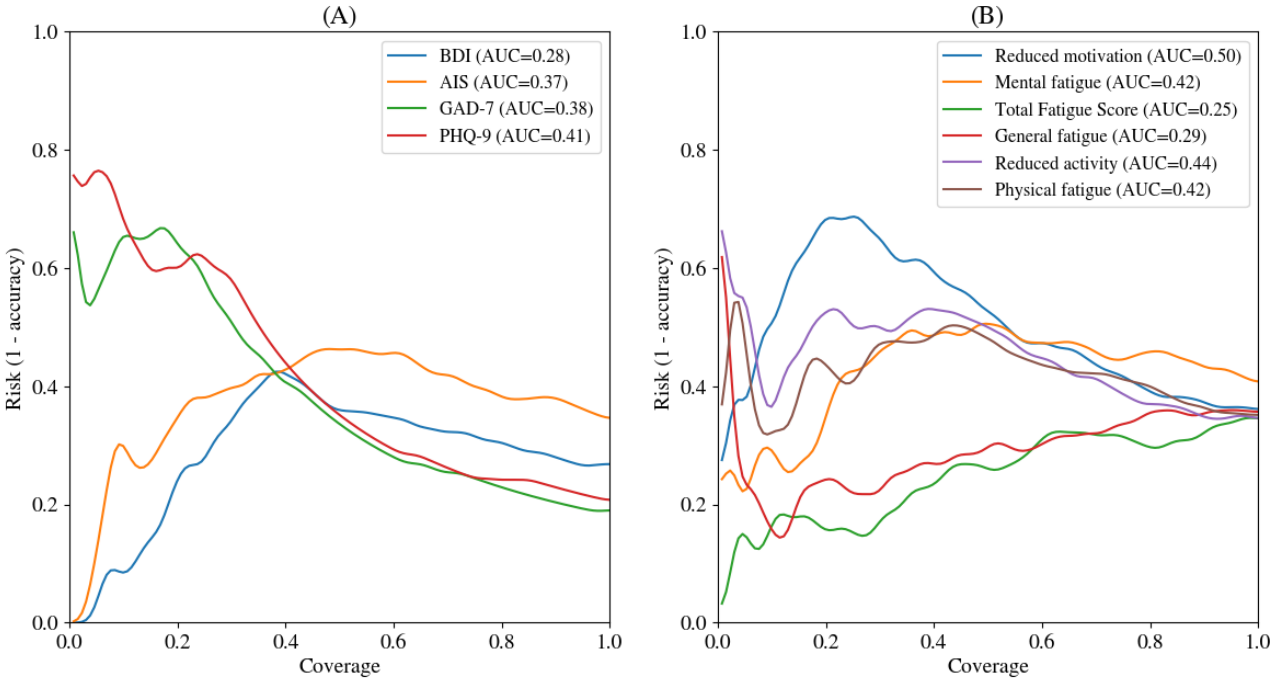

XSL•FO
RenderX

Table 2. Estimation of clinical threshold detection severity results on the held-out test set for the different considered dimensions of mental health (classification).

Dimensions of mental health	F_1-score[a]
PHQ-9[b]	0.49
GAD-7[c]	0.50
BDI[d]	0.65
AIS[e]	0.62
MFI[f] general fatigue	0.66
MFI physical fatigue	0.54
MFI reduced activity	0.52
MFI mental fatigue	0.53
MFI reduced motivation	0.43
MFI total fatigue score	0.69

[a]Higher F_1-score is better, and 1 is perfect.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: General Anxiety Disorder 7-item scale.

[d]BDI: Beck Depression Inventory.

[e]AIS: Athens Insomnia Scale.

[f]MFI: Multidimensional Fatigue Inventory.

XSL•FO
RenderX

Table 3. Disparity ratios (DRs) based on the F1-scores for sex, age, and education levels to assess fairness on the held-out test set for the clinical threshold detection (classification) for the different considered dimensions of mental health[a].

Dimensions of mental health	DR			Mean (SD)[b]
	Sex	Age[c]	Education level[d]	
PHQ-9[e]	0.61	0.00	0.00	0.20 (0.35)
GAD-7[f]	0.66	0.00	0.00	0.22 (0.38)
BDI[g]	0.85	0.25	0.57	0.56 (0.30)
AIS[h]	0.98	0.45	0.69	0.71 (0.27)
MFI[i] general fatigue	0.77	0.73	0.75	0.75 (0.02)
MFI physical fatigue	0.97	0.76	0.90	0.88 (0.11)
MFI reduced activity	0.95	0.27	0.56	0.59 (0.34)
MFI mental fatigue	0.91	0.00	0.58	0.50 (0.46)
MFI reduced motivation	0.94	0.25	0.30	0.50 (0.38)
MFI total fatigue score	0.91	0.62	0.35	0.63 (0.28)
Mean (SD)[j]	0.86 (0.13)	0.33 (0.30)	0.47 (0.30)	0.55 (0.33)

[a]Higher DR is better, and 1 is perfect DR.

[b]Mean and SD values per score.

[c]Participants are grouped into age categories to allow analysis: 18-30, 30-45, 45-65, and >65 years.

[d]Highest achieved study level: no diploma or secondary studies or short postbaccalaureate studies (Bac+2) or long postbaccalaureate studies (Bac+3 and above).

[e]PHQ-9: Patient Health Questionnaire-9.

[f]GAD-7: General Anxiety Disorder 7-item scale.

[g]BDI: Beck Depression Inventory.

[h]AIS: Athens Insomnia Scale.

[i]MFI: Multidimensional Fatigue Inventory.

[j]Mean and SD are reported per sensitive dimension.

XSL•FO
RenderX

Table 4. Estimation of severity results on the held-out test set for the different considered dimensions of mental health (regression)[a].

Dimensions of mental health	Pearson *r*	MAE[b]
PHQ-9[c]	0.47	3.1
GAD-7[d]	0.48	3.2
BDI[e]	0.49	4.9
AIS[f]	0.43	2.9
MFI[g] general fatigue	0.38	3.3
MFI physical fatigue	0.32	3.0
MFI reduced activity	0.31	2.9
MFI mental fatigue	0.34	3.1
MFI reduced motivation	0.32	2.5
MFI total fatigue score	0.44	11.3

[a]We reported mean absolute errors and Pearson correlations between actual and predicted values. Lower mean absolute error is better, and 0 is perfect. A higher Pearson correlation is better, and +1 is perfect. All correlations were significant ($P<1\times10^{-14}$).

[b]MAE: mean absolute error.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]GAD-7: General Anxiety Disorder 7-item scale.

[e]BDI: Beck Depression Inventory.

[f]AIS: Athens Insomnia Scale.

[g]MFI: Multidimensional Fatigue Inventory.

Clinical Threshold Detection (Classification)

The clinical threshold detection performed well based on the speech data and our developed system (Figure 3 and Table 2). All systems outperformed the chance levels. Based on the AUC, the classification results were the highest for the BDI score (AUC=0.78; F_1-score=0.65). Based on the F_1-score, it was the total MFI score (AUC=0.68; F_1-score=0.69). The lowest for both metrics was the MFI reduced motivation (AUC=0.61; F_1-score=0.43).

Fairness Assessments: Quality of Services for Sex, Age, and Education Level for Classification

We computed the sensitive attribute DRs (sex, age, and education level) and assessed the differences in the quality of the service made by the speech-based system (Table 3) for classification. Overall, the classification of the speech-based system had a better quality of service for sex (mean 0.86, SD 0.13), and the worst was for age (mean 0.33, SD 0.30). We also identified that the detection of PHQ-9 had the worst quality-of-service disparity (mean of DRs 0.20, SD 0.35), and the best quality-of-service was obtained with the AIS (mean of DRs 0.71, SD 0.27), the MFI general fatigue (mean of DRs 0.75, SD 0.02), and the MFI physical fatigue (mean of DRs 0.88, SD 0.11). We also observed that only the MFI general fatigue and MFI physical fatigue obtained a good performance for age (MFI general fatigue: DR=0.73 and MFI physical fatigue: DR=0.76), and, except for PHQ-9 and GAD-7, all DRs were satisfactory or high for sex.

Selective Clinical Threshold Detection (Selective Prediction)

The capabilities of speech-based models were also evaluated to selectively predict the different clinical thresholds. Great performances were observed for BDI, AIS, and MFI general fatigue (Figure 4). The model that selectively predicts the risk of depression based on the BDI score achieved the best result (BDI risk-coverage: AUC=0.28). The other scores could not achieve such feats of important coverage with no risk.

Estimation of Severity Through Predictions of Individual Scores (Regression)

The regression results for the estimation of severity are reported in Table 3. Speech-based models obtained significant results for all clinical variables based on the evaluation with the Pearson correlations (all $P<1\times10^{-14}$). The strongest correlations between the prediction and the actual scores on the held-out test were found for BDI score (r=0.49), GAD-7 (r=0.48), and PHQ-9 (r=0.47). The lowest correlation was found for the MFI reduced activity (r=0.31).

Speech-based models also obtained great results in terms of absolute errors. We observed less than 3 points of MAE for AIS and MFI reduced activity. All other scores were predicted on average with less than 5 points, except for the MFI total fatigue score, since its range is 13-88.

Discussion

Principal Findings

We aimed to explore the full capabilities and limitations of using speech data extracted from 865 participants in the general

XSL•FO
RenderX

population to predict the presence or absence of different mental health self-reported symptoms: depression, anxiety, and insomnia and the different dimensions of fatigue. We built a fully automated speech-based ML system that takes as input the audio waveform collected from 1 simple speech task performed on our smartphone app. The models were trained and calibrated on training and validation sets of participants, and we demonstrated the system's generalization on the held-out test set of 131 participants.

The results indicated that ML-based systems using speech only as input could identify participants above clinical thresholds for depression, insomnia, total fatigue components, but to a lesser extent, anxiety and fatigue subcomponents. All classification results were above chance levels for each clinical threshold.

This result was confirmed with an extensive fairness analysis of quality of service for age, sex, and education levels. Depression, insomnia, and different dimensions of fatigue clinical threshold detection results were particularly consistent for sex, slightly less for age, and to a lesser extent for education level. Anxiety risk identification fell behind in accuracy overall and was also unequal per group. The extension of our clinical threshold detection system to be able to abstain, with selective prediction, was conclusive, even for anxiety. Risk-coverage AUCs remained low for insomnia, total fatigue, and depression detection through BDI. Finally, we showed that speech-based models could also predict the severity, with the prediction of exact scores moving beyond binary interpretations of score thresholds. All correlations between the predicted scores and the actual scores given by participants were significant, exhibiting strengths ranging from 0.31 to 0.49.

Our study builds upon existing mental health research on speech analysis and extends the insights for deployment into clinical practice. For risk and anxiety depression assessments from speech in the general population, we found similar strong performances such as in previous studies using speech analyses [26,27,63,64]. Our recruitment and involvement of participants was in person. For medium-sized datasets with face-to-face recruitment (below 1000 participants, such as ours), it can be observed that crowdsourced recruitment of participants via web-based platforms needs more training data to yield the same level of model accuracy [63,65]. This was also observed in a large study, with over 6000 participants, with web-based data recruitment for risk detection in the general population in American English [64]. This discrepancy could be explained by the fact that data quality is variable on web crowdsourcing platforms, especially for participants' psychiatric evaluations [66], and also voice recording through laptops.

Our study revealed discrepancies in both clinical threshold detection and estimation of severity through self-reported depression scores between BDI and PHQ-9. This underlines the inherent limitations of score conversion and the crucial role of individual-level assessment in capturing the nuanced and different expressions of depression [35]. This reinforces the necessity of developing assessment tools and interpreting results with meticulous attention to individual variability, particularly by scrutinizing model performance at the individual level, mirroring real-world clinical scenarios.

Our study uniquely addresses the co-occurrence of perceived fatigue and reported insomnia, both prevalent mental health concerns, which could be detected simultaneously through speech analysis. While prior research, like the work by Thoret et al [20], has explored how sleep deprivation impacts specific vocal features like prosody and voice quality, no previous study has delved into the combined influence of fatigue and insomnia on speech. Addressing this gap is crucial because these conditions often co-occur and significantly impact the symptom trajectory and potential development of other mental health issues [67]. The prevention of recurrent sleep problems can prevent other mental health troubles or relapses. The observed co-occurrence of symptoms, particularly insomnia with other clinical dimensions, highlights the interconnected nature of these symptoms and syndromes.

Finally, to the best of our knowledge, this is the first study to assess the fairness and selective prediction capabilities in speech-based mental health assessments. This is of prime importance since speech signals can be heavily influenced by a multitude of factors such as age, sex, weight, and height [68]. Age factor was the least preserved in our grouped performances, this can be attributed to voice changes due to hormones [69], and normal aging affects the different parts of the vocal production system: larynx, respiratory system, resonators, saliva system, and the individual's global emotional status [70]. In addition to lower performances compared to other mental health dimensions, the anxiety risk detection performance collapsed for certain groups of demographics. This could be explained by the heterogeneity and low positive examples in our Callyope-GP dataset. Even though there are limits concerning some groups of individuals, selective classification offers an option to potentially remediate these variable quality of services, ensuring a deployment in clinical settings and still bringing overall clinical utility.

Limitations

While valuable, this study has some limitations. The French monolingual, medium-sized dataset (300<n<1000) needs more diverse data to achieve better generalizability, and the nonlongitudinal design misses insights on symptom evolution. Besides, the use of self-assessment scales introduces a potential bias because they rely heavily on the insight of participants. A limitation of our study is the use of a fixed train-dev-test split. This split approach, while convenient for comparing results across different tasks, can introduce bias and limit the generalizability of our findings. Future research with larger samples, longitudinal designs, and the inclusion of pathological data is crucial for exploiting the full potential of voice biomarkers in mental health. Studying various and diverse speech tasks and prompts also holds the potential for multiple benefits related to user adherence. Indeed, tailored tasks can address specific mental health needs, cater to individual preferences, and boost engagement. Varying prompts can reduce user fatigue and sustain interest, leading to more consistent system use and richer data collection.

XSL•FO
RenderX

Conclusions

This study demonstrates the potential of speech-based systems for detecting and predicting various mental health symptoms in the general population. While challenges remain regarding real-world apps and ensuring fairness across the population demographics, our findings pave the way for further development and responsible integration of such tools into clinical settings, advancing personalized mental health assessment and intervention. In future work, we will extend this study by including longitudinal data, adding more diverse linguistic and geographic data, and including more severely affected patients who are already followed by mental health practitioners. We will also look into fairness and uncertainty mitigation methods to improve the performance of our systems.

Acknowledgments

The authors are thankful to all the participants who volunteered for this research study. Without their active involvement, this study would not have been possible. The authors also would like to thank each of the speech pathology interns who helped with the participant recruitment and made sure that the protocol was completed successfully.

Conflicts of Interest

RR, XNC, AL, MDG, MD, and AB are shareholders of Callyope, and VO was a former employee of Callyope.

Multimedia Appendix 1
Co-occurrences of symptoms, and validation and comparison results for each speech modeling approach.
[DOCX File , 436 KB - jmir_v26i1e58572_app1.docx]

References

1. GBD 2019 Diseases Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
4. Henriksson MM, Aro HM, Marttunen MJ, Heikkinen ME, Isometsä ET, Kuoppasalmi KI, et al. Mental disorders and comorbidity in suicide. Am J Psychiatry 1993;150(6):935-940. [doi: 10.1176/ajp.150.6.935] [Medline: 8494072]
5. Guo T, Xiang Y, Xiao L, Hu C, Chiu H, Ungvari G, et al. Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. AJP 2015 Oct 01;172(10):1004-1013 [FREE Full text] [doi: 10.1176/appi.ajp.2015.14050652]
6. Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K. How many different ways do patients meet the diagnostic criteria for major depressive disorder? Compr Psychiatry 2015;56:29-34. [doi: 10.1016/j.comppsych.2014.09.007] [Medline: 25266848]
7. Soehner AM, Kaplan KA, Harvey AG. Prevalence and clinical correlates of co-occurring insomnia and hypersomnia symptoms in depression. J Affect Disord 2014;167:93-97 [FREE Full text] [doi: 10.1016/j.jad.2014.05.060] [Medline: 24953480]
8. Corfield EC, Martin NG, Nyholt DR. Co-occurrence and symptomatology of fatigue and depression. Compr Psychiatry 2016;71:1-10. [doi: 10.1016/j.comppsych.2016.08.004] [Medline: 27567301]
9. Park LT, Zarate CA. Depression in the primary care setting. N Engl J Med 2019;380(6):559-568 [FREE Full text] [doi: 10.1056/NEJMcp1712493] [Medline: 30726688]
10. Bard HA, O'Driscoll C, Miller CB, Henry AL, Cape J, Espie CA. Insomnia, depression, and anxiety symptoms interact and individually impact functioning: a network and relative importance analysis in the context of insomnia. Sleep Med 2023;101:505-514 [FREE Full text] [doi: 10.1016/j.sleep.2022.12.005] [Medline: 36527942]
11. Cummins N, Dineley J, Conde P, Matcham F, Siddi S, Lamers F, et al. Multilingual markers of depression in remotely collected speech samples: a preliminary analysis. J Affect Disord 2023;341:128-136 [FREE Full text] [doi: 10.1016/j.jad.2023.08.097] [Medline: 37598722]
12. Dumpala SH, Dikaios K, Rodriguez S, Langley R, Rempel S, Uher R, et al. Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. Sci Rep 2023;13(1):11155 [FREE Full text] [doi: 10.1038/s41598-023-35184-7] [Medline: 37429935]

XSL•FO
RenderX

13. Bone D, Gibson J, Chaspari T, Can D. Speech and language processing for mental health research and care. 2016 Presented at: 2016 50th Asilomar Conference on Signals, Systems and Computers; November 6-9, 2016; Pacific Grove, CA p. 831-835. [doi: 10.1109/acssc.2016.7869164]
14. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015;71:10-49 [FREE Full text] [doi: 10.1016/j.specom.2015.03.004]
15. Williamson JR, Quatieri TF, Helfer B. Vocal and facial biomarkers of depression based on motor incoordination and timing. 2014 Presented at: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge; November 7, 2014; Orlando, FL p. 65-72. [doi: 10.1145/2661806.2661809]
16. Teferra BG, Borwein S, DeSouza DD, Simpson W, Rheault L, Rose J. Acoustic and linguistic features of impromptu speech and their association with anxiety: validation study. JMIR Ment Health 2022;9(7):e36828 [FREE Full text] [doi: 10.2196/36828] [Medline: 35802401]
17. Stasak B, Epps J, Schatten HT, Miller IW, Provost EM, Armey MF. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. Speech Commun 2021;132:10-20 [FREE Full text] [doi: 10.1016/j.specom.2021.05.004]
18. Bayerl SP, Wagner D, Baumann I, Bocklet T, Riedhammer K. Detecting vocal fatigue with neural embeddings. J Voice 2023:11-15. [doi: 10.1016/j.jvoice.2023.01.012] [Medline: 36774263]
19. Elbéji A, Zhang L, Higa E, Fischer A, Despotovic V, Nazarov PV, et al. Vocal biomarker predicts fatigue in people with COVID-19: results from the prospective Predi-COVID cohort study. BMJ Open 2022;12(11):e062463 [FREE Full text] [doi: 10.1136/bmjopen-2022-062463] [Medline: 36414294]
20. Thoret E, Andrillon T, Gauriau C, Léger D, Pressnitzer D. Sleep deprivation detected by voice analysis. PLoS Comput Biol 2024;20(2):e1011849 [FREE Full text] [doi: 10.1371/journal.pcbi.1011849] [Medline: 38315733]
21. Levelt WJM. Speaking: From Intention to Articulation. Cambridge, MA: MIT Press; 1989.
22. Guenther FH. Neural Control of Speech. Cambridge, MA: MIT Press; 2016.
23. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol 2020;5(1):96-116 [FREE Full text] [doi: 10.1002/lio2.354] [Medline: 32128436]
24. Berta A, Ángel CM, Clara GS, Rubén H. A bibliometric analysis of 10 years of research on symptom networks in psychopathology and mental health. Psychiatry Res 2022;308:114380 [FREE Full text] [doi: 10.1016/j.psychres.2021.114380] [Medline: 34999293]
25. Borsboom D. A network theory of mental disorders. World Psychiatry 2017;16(1):5-13 [FREE Full text] [doi: 10.1002/wps.20375] [Medline: 28127906]
26. König A, Tröger J, Mallick E, Mina M, Linz N, Wagnon C, et al. Detecting subtle signs of depression with automated speech analysis in a non-clinical sample. BMC Psychiatry 2022;22(1):830 [FREE Full text] [doi: 10.1186/s12888-022-04475-0] [Medline: 36575442]
27. Lin D, Nazreen T, Rutowski T, Lu Y, Harati A, Shriberg E, et al. Feasibility of a machine learning-based smartphone application in detecting depression and anxiety in a generally senior population. Front Psychol 2022;13:811517 [FREE Full text] [doi: 10.3389/fpsyg.2022.811517] [Medline: 35478769]
28. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA: MIT Press; 2023.
29. Martin VP, Rouas JL. Estimating symptoms and clinical signs instead of disorders: the path toward the clinical use of voice and speech biomarkers in psychiatry. 2024 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); April 14-19, 2024; Seoul, South Korea. [doi: 10.1109/icassp48485.2024.10445888]
30. Bondi E, Koster R, Sheahan H, Chadwick M, Bachrach Y, Cemgil T, et al. Role of human-AI interaction in selective prediction. Proc AAAI Conf Artif Intell 2022;36(5):5286-5294 [FREE Full text] [doi: 10.1609/aaai.v36i5.20465]
31. Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. Front Psychol 2017;8:269 [FREE Full text] [doi: 10.3389/fpsyg.2017.00269] [Medline: 28321196]
32. Beck AT, Ward CH, Mendelson M, Mock J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561-574. [doi: 10.1001/archpsyc.1961.01710120031004]
33. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
34. Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract 2008;58(546):32-36 [FREE Full text] [doi: 10.3399/bjgp08X263794] [Medline: 18186994]
35. Hawley CJ, Gale TM, Smith PSJ, Jain S, Farag A, Kondan R, et al. Equations for converting scores between depression scales (MÅDRS, SRS, PHQ-9 and BDI-II): good statistical, but weak idiographic, validity. Hum Psychopharmacol 2013;28(6):544-551. [doi: 10.1002/hup.2341] [Medline: 24519690]
36. Beck AT, Steer RA, Garbin MG. Psychometric properties of the beck depression inventory: twenty-five years of evaluation. Clin Psychol Rev 1988;8:77-100 [FREE Full text] [doi: 10.1016/0272-7358(88)90050-5]
37. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]

XSL•FO
RenderX

38. Soldatos CR, Dikeos DG, Paparrigopoulos TJ. Athens Insomnia Scale: validation of an instrument based on ICD-10 criteria. J Psychosom Res 2000;48(6):555-560. [doi: 10.1016/s0022-3999(00)00095-7] [Medline: 11033374]
39. Soldatos CR, Dikeos DG, Paparrigopoulos TJ. The diagnostic validity of the Athens Insomnia Scale. J Psychosom Res 2003;55(3):263-267. [doi: 10.1016/s0022-3999(02)00604-9] [Medline: 12932801]
40. Okajima I, Miyamoto T, Ubara A, Omichi C, Matsuda A, Sumi Y, et al. Evaluation of severity levels of the Athens Insomnia Scale based on the criterion of Insomnia Severity Index. Int J Environ Res Public Health 2020;17(23):8789 [FREE Full text] [doi: 10.3390/ijerph17238789] [Medline: 33256097]
41. Smets EM, Garssen B, Bonke B, de Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39(3):315-325. [doi: 10.1016/0022-3999(94)00125-o] [Medline: 7636775]
42. Schwarz R, Krauss O, Hinz A. Fatigue in the general population. Onkologie 2003;26(2):140-144 [FREE Full text] [doi: 10.1159/000069834] [Medline: 12771522]
43. Gentile S, Delarozière JC, Favre F, Sambuc R, San Marco JL. Validation of the French 'Multidimensional Fatigue Inventory' (MFI 20). Eur J Cancer Care (Engl) 2003;12(1):58-64. [doi: 10.1046/j.1365-2354.2003.00295.x] [Medline: 12641557]
44. Hinz A, Barboza CF, Barradas S, Körner A, Beierlein V, Singer S. Fatigue in the general population of Colombia—normative values for the Multidimensional Fatigue Inventory MFI-20. Onkologie 2013;36(7-8):403-407. [doi: 10.1159/000353606] [Medline: 23921758]
45. He K, Zhang X, Ren S. Deep residual learning for image recognition. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV p. 770-778. [doi: 10.1109/cvpr.2016.90]
46. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L. Attention is all you need. 2017 Presented at: Advances in Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 6000-6010 URL: https://dl.acm.org/doi/10.5555/3295222.3295349 [doi: 10.5555/3295222.3295349]
47. Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021;29:3451-3461. [doi: 10.1109/taslp.2021.3122291]
48. Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR, Nöth E. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun 2018;101:11-25. [doi: 10.1016/j.specom.2018.05.007]
49. Chung JS, Nagrani A, Zisserman A. VoxCeleb2: deep speaker recognition. 2018 Presented at: Interspeech 2018; September 2-6, 2018; Hyderabad, India p. 1086-1090. [doi: 10.21437/interspeech.2018-1929]
50. Wang F, Cheng J, Liu W, Liu H. Additive margin softmax for face verification. IEEE Signal Process Lett 2018;25(7):926-930 [FREE Full text] [doi: 10.1109/lsp.2018.2822810]
51. Morais E, Hoory R, Zhu W. Speech emotion recognition using self-supervised features. 2022 Presented at: ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore. [doi: 10.1109/icassp43922.2022.9747870]
52. Panayotov V, Chen G, Povey D. Librispeech: an ASR corpus based on public domain audio books. 2015 Presented at: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); April 19-24, 2015; South Brisbane, Queensland, Australia p. 5206-5210. [doi: 10.1109/icassp.2015.7178964]
53. Chung S JS, Huh J, Mun S, Lee M, Heo HS, Choe S, et al. In defence of metric learning for speaker recognition. 2024 Presented at: Interspeech 2020; October 25-29, 2020; Shanghai, China URL: https://www.isca-archive.org/interspeech_2020/chung20b_interspeech.pdf
54. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
55. Agarwal A, Beygelzimer A, Dudik M, Langford J, Wallach H. A reductions approach to fair classification. 2018 Presented at: Proceedings of the 35th International Conference on Machine Learning; March 6, 2018; Stockholm, Sweden p. 66-69 URL: https://proceedings.mlr.press/v80/agarwal18a/agarwal18a.pdf
56. Weerts H, Dudík M, Edgar R, Jalali A, Lutz R. Fairlearn: assessing and improving fairness of AI systems. J Mach Learn Res 2023;24:1-8 [FREE Full text]
57. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021;4(1):4 [FREE Full text] [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
58. Zhang H, Maloney LT. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. Front Neurosci 2012;6:1 [FREE Full text] [doi: 10.3389/fnins.2012.00001] [Medline: 22294978]
59. Xin J, Tang R, Yu Y, Lin J. The art of abstention: selective prediction and error regularization for natural language processing. 2021 Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; August 28, 2021; Online p. 1040-1051. [doi: 10.18653/v1/2021.acl-long.84]
60. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. J Mach Learn Res 2010:1605-1641 [FREE Full text]

XSL•FO
RenderX

61. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. ArXiv. Preprint posted online on October 3, 2018 . [doi: 10.48550/arXiv.1610.02136]
62. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 2018;180(Pt A):68-77. [doi: 10.1016/j.neuroimage.2017.06.061] [Medline: 28655633]
63. Tasnim M, Ehghaghi M, Diep B, Novikova J. DEPAC: a corpus for depression and anxiety detection from speech. 2022 Presented at: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology; July 17, 2022; Seattle, WA p. 1-16. [doi: 10.18653/v1/2022.clpsych-1.1]
64. Rutowski T, Harati A, Shriberg E, Lu Y. Toward corpus size requirements for training and evaluating depression risk models using spoken language. 2022 Presented at: Interspeech 2022; September 18-22, 2022; Incheon, South Korea p. 3343-3347. [doi: 10.21437/interspeech.2022-10888]
65. Teferra BG, Rose J. Predicting generalized anxiety disorder from impromptu speech transcripts using context-aware transformer-based neural networks: model evaluation study. JMIR Ment Health 2023;10:e44325 [FREE Full text] [doi: 10.2196/44325] [Medline: 36976636]
66. Kolc KL, Tan YXK, Lo AZY, Shvetcov A, Mitchell PB, Perkes IE. Measuring psychiatric symptoms online: a systematic review of the use of inventories on Amazon Mechanical Turk (mTurk). J Psychiatr Res 2023;163:118-126 [FREE Full text] [doi: 10.1016/j.jpsychires.2023.05.027] [Medline: 37209617]
67. McCallum SM, Batterham PJ, Calear AL, Sunderland M, Carragher N, Kazan D. Associations of fatigue and sleep disturbance with nine common mental disorders. J Psychosom Res 2019;123:109727. [doi: 10.1016/j.jpsychores.2019.05.005] [Medline: 31376877]
68. Sorokowski P, Pisanski K, Frąckowiak T, Kobylarek A, Groyecka-Bernard A. Voice-based judgments of sex, height, weight, attractiveness, health, and psychological traits based on free speech versus scripted speech. Psychon Bull Rev 2024;31(4):1680-1689. [doi: 10.3758/s13423-023-02445-5] [Medline: 38238560]
69. Amir O, Biron-Shental T. The impact of hormonal fluctuations on female vocal folds. Curr Opin Otolaryngol Head Neck Surg 2004;12(3):180-184. [doi: 10.1097/01.moo.0000120304.58882.94] [Medline: 15167026]
70. Beton S, Yücel L, Başak H, Çiler Büyükatalay Z. The elderly voice: mechanisms, disorders and treatment methods. Turk Arch Otorhinolaryngol 2022;60(4):220-226 [FREE Full text] [doi: 10.4274/tao.2022.2022-8-1] [Medline: 37456599]

Abbreviations

AIS: Athens Insomnia Scale
AUC: area under the curve
BDI: Beck Depression Inventory
Callyope-GP: Callyope General-Population
DR: disparity ratio
GAD-7: General Anxiety Disorder 7-item scale
ICD-10: International Classification of Diseases, Tenth Revision
MAE: mean absolute error
MFI: Multidimensional Fatigue Inventory
ML: machine learning
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial

Edited by G Eysenbach, A Coristine; submitted 20.03.24; peer-reviewed by L Yang, V Martin; comments to author 08.04.24; revised version received 07.09.24; accepted 25.09.24; published 31.10.24.

Please cite as:
Riad R, Denais M, de Gennes M, Lesage A, Oustric V, Cao XN, Mouchabac S, Bourla A
Automated Speech Analysis for Risk Detection of Depression, Anxiety, Insomnia, and Fatigue: Algorithm Development and Validation Study
J Med Internet Res 2024;26:e58572
URL: https://www.jmir.org/2024/1/e58572
doi:10.2196/58572
PMID:

©Rachid Riad, Martin Denais, Marc de Gennes, Adrien Lesage, Vincent Oustric, Xuan Nga Cao, Stéphane Mouchabac, Alexis Bourla. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/),

XSL•FO
RenderX

which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use of Artificial Intelligence in Cobb Angle Measurement for Scoliosis: Retrospective Reliability and Accuracy Study of a Mobile App

Haodong Li[1*], MD; Chuang Qian[1*], MD; Weili Yan[1*], PhD; Dong Fu[1], MD; Yiming Zheng[1], MD; Zhiqiang Zhang[1], MD; Junrong Meng[1], MD; Dahui Wang[1], MD

Department of Orthopedics, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Dahui Wang, MD
Department of Orthopedics
Children's Hospital of Fudan University
National Children's Medical Center
Wanyuan Rd
Minhang District
Shanghai, 201102
China
Phone: 86 02164931101
Email: wangdahui@fudan.edu.cn

Abstract

Background: Scoliosis is a spinal deformity in which one or more spinal segments bend to the side or show vertebral rotation. Some artificial intelligence (AI) apps have already been developed for measuring the Cobb angle in patients with scoliosis. These apps still require doctors to perform certain measurements, which can lead to interobserver variability. The AI app (cobbAngle pro) in this study will eliminate the need for doctor measurements, achieving complete automation.

Objective: We aimed to evaluate the reliability and accuracy of our new AI app that is based on deep learning to automatically measure the Cobb angle in patients with scoliosis.

Methods: A retrospective analysis was conducted on the clinical data of children with scoliosis who were treated at the Pediatric Orthopedics Department of the Children's Hospital affiliated with Fudan University from July 2019 to July 2022. Three measurers used the Picture Archiving and Communication System (PACS) to measure the coronal main curve Cobb angle in 802 full-length anteroposterior and lateral spine X-rays of 601 children with scoliosis, and recorded the results of each measurement. After an interval of 2 weeks, the mobile AI app was used to remeasure the Cobb angle once. The Cobb angle measurements from the PACS were used as the reference standard, and the accuracy of the Cobb angle measurements by the app was analyzed through the Bland-Altman test. The intraclass correlation coefficient (ICC) was used to compare the repeatability within measurers and the consistency between measurers.

Results: Among 601 children with scoliosis, 89 were male and 512 were female (age range: 10-17 years), and 802 full-length spinal X-rays were analyzed. Two functionalities of the app (photography and photo upload) were compared with the PACS for measuring the Cobb angle. The consistency was found to be excellent. The average absolute errors of the Cobb angle measured by the photography and upload methods were 2.00 and 2.08, respectively. Using a clinical allowance maximum error of 5°, the 95% limits of agreement (LoAs) for Cobb angle measurements by the photography and upload methods were –4.7° to 4.9° and –4.9° to 4.9°, respectively. For the photography and upload methods, the 95% LoAs for measuring Cobb angles were –4.3° to 4.6° and –4.4° to 4.7°, respectively, in mild scoliosis patients; –4.9° to 5.2° and –5.1° to 5.1°, respectively, in moderate scoliosis patients; and –5.2° to 5.0° and –6.0° to 4.8°, respectively, in severe scoliosis patients. The Cobb angle measured by the 3 observers twice before and after using the photography method had good repeatability (P<.001). The consistency between the observers was excellent (P<.001).

Conclusions: The new AI platform is accurate and repeatable in the automatic measurement of the Cobb angle of the main curvature in patients with scoliosis.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50631) doi:10.2196/50631

KEYWORDS

scoliosis; photogrammetry; artificial intelligence; deep learning

Introduction

Scoliosis is identified as a spinal deformity characterized by lateral curvature and vertebral rotation, diagnosed through an X-ray indicating a Cobb angle greater than 10° [1]. The incidence rate of scoliosis in children is between 1% and 3% [2]. The causes of scoliosis are diverse and cannot be prevented [3]. Early screening during adolescence is crucial for identifying scoliosis, where most cases are mild but can progress rapidly in 10% to 20% of patients, necessitating regular quantitative monitoring [4]. Regular and quantitative monitoring is essential for these patients.

At present, the measurement of the Cobb angle is still the most commonly used quantitative index to evaluate scoliosis severity. It has important reference value for the diagnosis, choice of treatment strategy, and evaluation of the curative effect of scoliosis [5]. However, in clinical practice, Cobb angle measurement is different within and between observers [6,7]. Manual Cobb angle measurement has 3-5° of intraobserver variation and 5-7° of interobserver variation [8]. However, these values are reported for experienced raters, and thus, the variation could be higher for less experienced raters. This error may be caused by the different selection of the end vertebrae and the manual error during measurement [4]. When the child's positioning is not standardized during the radiography session, the quality of the full-length spinal X-ray may be compromised, failing to accurately reflect the true extent of the child's spinal curvature and the size of the Cobb angle. This is because the human spine itself has a certain degree of mobility. For instance, if the child's upper limbs are positioned asymmetrically during the X-ray, it will directly cause the spine to bend. This results in the measured Cobb angle being either larger or smaller compared to the child's actual Cobb angle, further leading to errors in the measurement of the Cobb angle. After these variations are compounded, some clinicians may have a measurement error of 5-10° in assessing the Cobb angle of a child, which is sufficient to lead clinicians to misjudge the progression of a patient's scoliosis, thereby causing the child to miss the optimal treatment opportunity [7,9].

Various methods for measuring the Cobb angle exist, ranging from traditional manual measurements with a pencil and protractor on X-ray films to digital methods like the Picture Archiving and Communication System (PACS) used increasingly in hospitals [10,11]. Despite advancements, challenges persist in tool specificity, usage conditions, and measurement repeatability.

In recent years, artificial intelligence (AI) models have been shown to be remarkably successful in the interpretation of medical images [12]. Due to advances in the computing power of smartphones, health care professionals can easily obtain AI-generated measurement data through a user-friendly mobile app system to assist in the diagnosis, prediction, and treatment of diseases. The mobile app can intelligently recognize and capture images for the assessment of body posture and measurement of the Cobb angle in kyphosis, delivering accurate results. Moreira et al [13] compared the NLMeasurer with a validated biophotogrammetry software for 6 posture measurements and found that the NLMeasurer, a mobile app based on the PoseNet deep learning algorithm, provides highly reliable measurements for frontal posture assessment.

In the realm of image-based mobile apps, the 2 commonly used smartphone apps for measuring the Cobb angle in scoliosis are the iPinPoint and Cobbmeter apps, which are available for download from the Apple iTunes store. Numerous researchers have investigated these apps [11,14,15]. The iPinPoint and Cobbmeter apps are mobile apps that have integrated spinal X-ray capture using a smartphone and an interface that allows evaluators to manually mark the upper and lower vertebrae of the full-length spinal X-ray using touchscreen functionality.

Despite the significant advancements that these mobile apps for measuring the Cobb angle represent in the scoliosis assessment process, desktop solutions still require the manual identification of the upper and lower vertebrae. After capturing the spinal X-ray with a mobile device, the examiner must mark the planes of the endplates of the terminal vertebrae within the software interface on the phone before the app can measure the Cobb angle. This process may lead to variability in the results, depending on the examiner's ability to accurately identify the terminal vertebrae.

The application of AI in radiology holds enormous potential in achieving more accurate diagnosis of diseases and improving treatment decision-making. Using the deep learning of AI, we developed an app (cobbAngle pro) that could automatically measure the Cobb angle of scoliosis. This app, designed for mobile devices, is user-friendly, requires minimal technical expertise in imaging and scene setup, and eliminates the need for the examiner to manually select the terminal vertebrae to measure the Cobb angle. This functionality offers flexibility during the assessment process and reduces the occurrence of human errors when selecting the terminal vertebrae.

This study attempts to evaluate the reliability of the new AI phone app based on deep learning.

Methods

General Information

We conducted a retrospective study on patients with scoliosis who came to our center from July 2019 to July 2022. The patients' spinal X-rays were all sourced from the PACS. The inclusion criteria were as follows: (1) age 10-18 years; (2) diagnosis of noncongenital scoliosis; (3) ability to complete photography in a standard posture (The patient stands with feet slightly apart, knees and hips naturally extended, torso without external force interference, both upper limbs naturally hanging

XSL•FO
RenderX

down, leaning on the support poles on both sides of the torso to reduce the impact of the upper limbs on the torso's balance, chin lifted, and eyes looking straight ahead).

The exclusion criteria were as follows: (1) congenital scoliosis and secondary scoliosis (unequal length of lower limbs, etc) and (2) nonstandard shooting positions (This refers to a posture that does not adhere to standard radiographic techniques. Common nonstandard postures for full-length anterior-posterior spinal radiographs primarily include asymmetrical positioning of the upper limbs on both sides and 1 lower limb positioned in front of the other).

Principles of Intelligent Application

In this study, an AI app was developed to automatically calculate the Cobb angle from spinal X-rays using deep learning and image processing techniques. The model employed by this mobile app uses a neural network framework based on the transformer mechanism, which directly predicts the Cobb angle and vertebral coordinates from the preprocessed training dataset. The neural network architecture comprises an encoder, a decoder, and a transformer module. The encoder is based on the ResNet network, with the final fully connected layer removed. The extracted feature blocks are passed to the transformer module. In this module, the feature maps extracted by the convolutional neural network are unfolded into 1-dimensional features, which are then combined with positional encoding. The transformer's inherent encoder and decoder extract spatial correlations, resulting in the output of 1-dimensional features. These features are subsequently converted into tensor blocks and fed into the decoder. The transformer module is essential for adjusting the dimensionality between the input and output layers, as well as for performing data denoising. The decoder receives the tensor blocks and queries feature blocks of the same size as those obtained during the encoding stage. It then performs deconvolution operations to ultimately generate a heatmap of the central point coordinates and offset vectors. This module is used to annotate the central points in the images, thereby enhancing the network's recognition efficiency. Finally, an angle offset loss function is incorporated into the network to ensure that the training results converge toward accurate predictions. The following text outlines the data processing procedures, AI model training procedures, and principles for confirming the Cobb angle using the AI model (Figure 1).

Figure 1. Network framework.

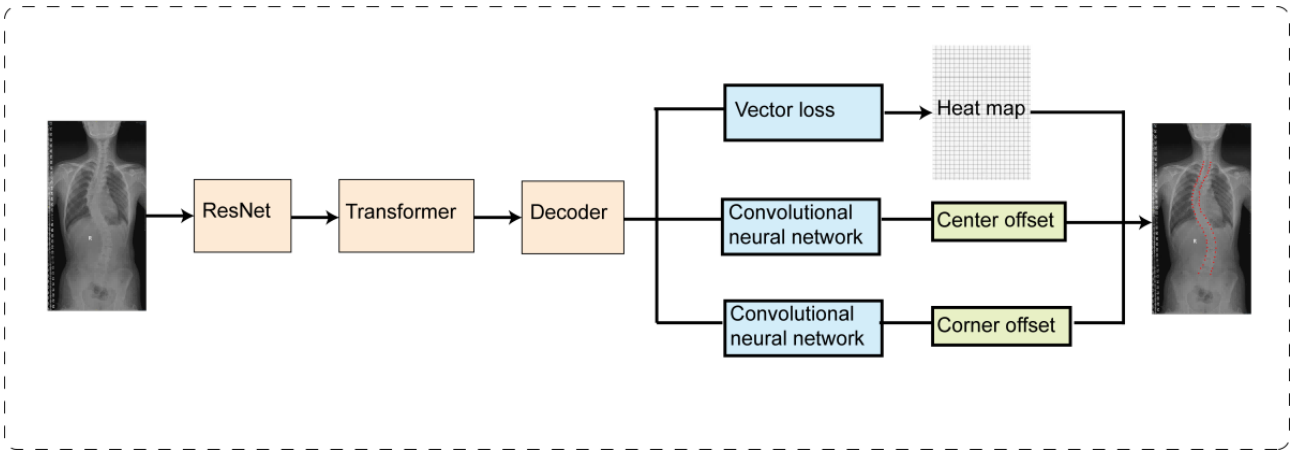

Data Collection

We collected 20,000 spinal X-rays from various hospitals. For each X-ray, 4 points on all vertebrae were manually annotated to ensure that each image was paired with the corresponding Cobb angle annotation.

Preprocessing

The images underwent preprocessing steps, such as normalization, grayscaling, and scaling, to meet the input requirements of ResNet. Data augmentation techniques (eg, rotation, scaling, and cropping) were applied to enhance data diversity and prevent model overfitting.

Data Splitting

The current dataset consists of 18,000 images for training and 2000 images for testing, with a 9:1 ratio. During the training process, 9-fold cross-validation was employed to mitigate overfitting. Based on hospital statistics, the ratio of mild-to-moderate-to-severe cases among pediatric orthopedic outpatients was 5:5:2. However, in the first phase of data collection, the ratio of mild-to-moderate-to-severe cases was 9:6:1. This does not align with the true distribution of severity levels, as there is a shortage of severe cases, leading to an imbalance in the dataset. To address this issue, we collected additional imaging data, including X-rays, from severe pediatric cases. During the training process, we also removed some data from mild cases to alleviate the data imbalance and better approximate the true distribution observed in our pediatric hospital.

Model Inference

Finally, a pretrained ResNet variant deep learning model was applied to the preprocessed images for feature extraction and key point detection, identifying the 4 key points on all vertebrae.

The program uses the moving least squares (MLS) method for curve fitting to obtain the overall shape of the spine. Polynomial fitting is then used to calculate the second derivative of the curve. Changes in the second derivative reflect the curvature and inflection points of the curve, indicating changes in the bending direction. Based on the bending direction, the spinal

XSL•FO
RenderX

curve is divided into multiple segments to calculate the Cobb angle for each segment. For each segment, 2 lines are fitted to the upper and lower edges of the vertebrae using key points, and a loop iterates through all possible line combinations to calculate the intersection angles. Angles below 5° are ignored.

The method for calculating the Cobb angle for each line combination is as follows:

First, the tangent of the angle between the 2 lines is calculated using the following formula:

Then, the angle is calculated using the arctangent function:

Among all the calculated angles, the largest Cobb angle is selected as the Cobb angle for the corresponding segment. After calculating the Cobb angles for all segments, the 2 largest Cobb angles are identified, with the slightly larger angle referred to as the major angle and the slightly smaller angle referred to as the minor angle.

Model Fairness and Code Privacy

Our trained model has demonstrated good fairness, as the calculation method for spinal vertebrae does not exhibit differences based on factors, such as gender, making the model broadly applicable. Due to data privacy concerns, however, we have not publicly shared the trained model or the code.

Research Methods

Selection of Observers

We selected 3 fellow doctors in pediatric orthopedics (with at least 2 years of working experience in spinal deformities) as observers. The developers equipped the 3 observers with the same type of terminal equipment and conducted mobile app operation training until they could accurately and skillfully use the app to measure the Cobb angle. When measuring the Cobb angle for scoliosis using the mobile app, it is essential to ensure that the photographed scene includes all spinal segments in the X-ray, covering the entire length of the spine. This is crucial for accurately measuring the primary Cobb angle of scoliosis patients. The plane of the photo should also be as parallel to the X-ray film as possible to avoid significant measurement errors due to improper use of the app method. Therefore, training for the 3 observers was necessary.

Study Setting

This study was conducted in the orthopedic doctors' office of the first ward of the inpatient department at the Children's Hospital of Fudan University. Three observers performed manual measurements using 3 PACS-equipped computers in the office. The process of measuring the Cobb angle with the mobile app was also carried out at the same location.

Measurement Methods

The X-ray films of children were disordered, and the identification information of each film was covered up. The person who selected the film and the observers were completely unaware of the patients' information. To reduce the impact of subjective and informational biases, the measurers first measured the main curvature Cobb angle using the PACS method, followed by measurements with the mobile app, with a specific time interval (2 weeks) between the 2 measurement methods. After measuring the main curvature Cobb angle with the PACS method, the measurers waited 2 weeks before continuing to measure the main curvature Cobb angle of the same set of spinal X-rays of patients using the mobile app by the photography and photo upload methods, further eliminating the impact of information bias. The measurement data were recorded on a separate data sheet. After 2 weeks, 3 observers took the second measurement to reduce their own deviation.

PACS Method

Three observers used the measuring tools of the PACS to determine the upper and lower vertebrae of the main curvature of scoliosis on the computer, marked the upper and lower endplates of the upper and lower vertebrae, and then measured their Cobb angles with the measuring tools and recorded the data.

Photography Method

In order to reduce subjective bias, the time of measuring the Cobb angle by using the 2 functions of the app was determined after the PACS method was completed. When shooting photos of the X-ray film on the computer, the full length of the spine was included in the shooting interface, and the mobile phone posture was parallel to the position of the X-ray film as much as possible, and was aligned and kept stable to reduce errors. After the successful shooting, the software automatically recognized and displayed the Cobb angle, and the 3 measurers recorded it individually (Figure 2).

XSL•FO
RenderX

Figure 2. Cobb angle measured by the photography method.

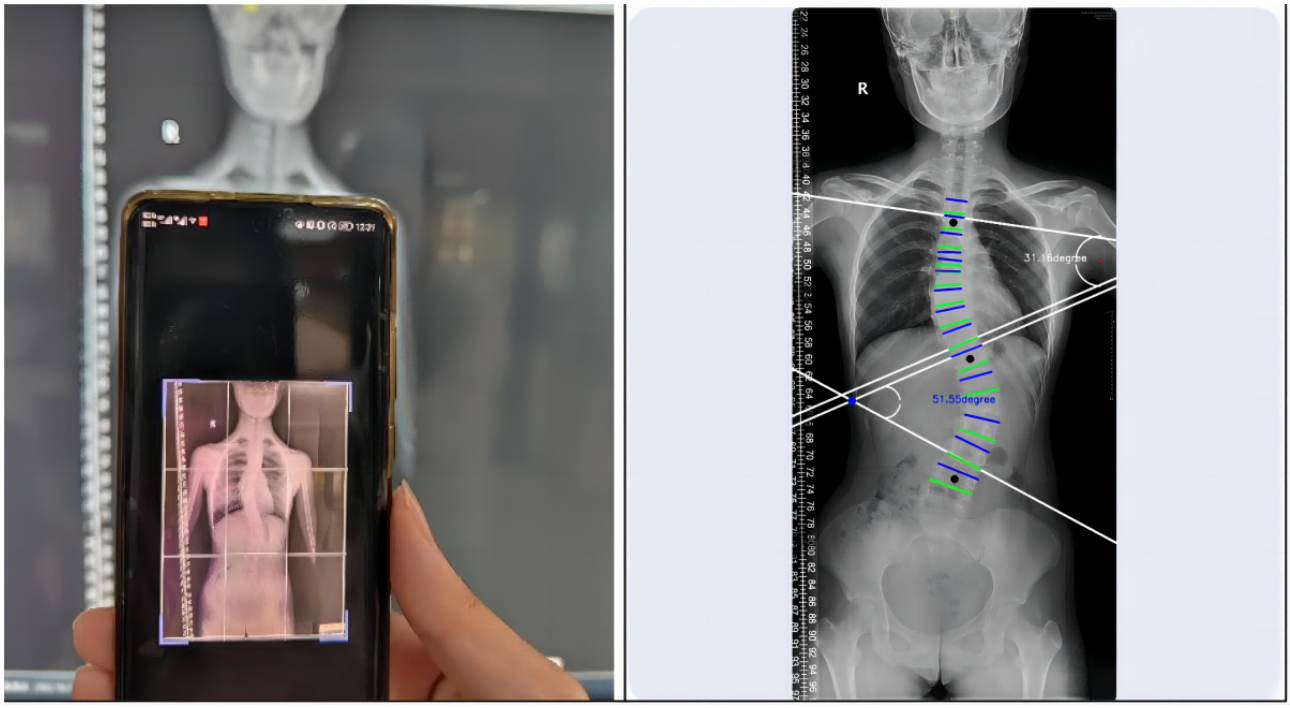

Photo Upload Method

We downloaded the X-ray films of all the included samples from the hospital system and saved them in the photo album of the test phone. Subsequently, we selected the "photo upload" function of the app to upload the original X-ray images of all samples. The app will automatically recognize the image and measure its Cobb angle. The data were recorded by 3 observers (Figure 3).

Figure 3. Cobb angle measured by the photo upload method.

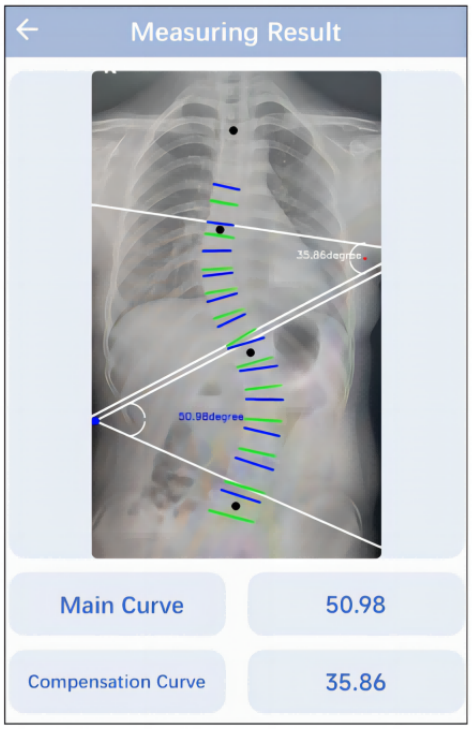

XSL•FO
RenderX

Statistical Analysis

Software SPSS v20 (IBM Corp) and MedCalc (MedCalc Software Ltd) were used for statistical analysis. The intraclass correlation coefficient (ICC) was used to compare the Cobb angle measurement results of the app method (photography and photo upload) and PACS method to judge the consistency of the measurement results. For ICC, 0.250-0.499 indicates poor consistency, 0.500-0.699 indicates moderate consistency, 0.700-0.899 indicates good consistency, and ≥0.900 indicates excellent consistency. A *P*-value <.05 was considered to indicate a statistically significant difference.

We divided the subjects into 3 groups according to the degree of scoliosis (mild: Cobb <25°, moderate: 25° ≤ Cobb ≤ 45°, severe: Cobb >45°) and used the Bland-Altman method to test the consistency of the Cobb angle measurement between the app and the PACS within the groups. The Bland-Altman test was used to obtain the 95% limits of agreement (LoAs) based on the average difference of ±1.96 SD between the 2 measurement methods. The measurement error of the 2 methods has been represented by drawing a scatter diagram.

We used the ICC for both reliability analysis (interobserver difference analysis) and repeatability analysis (intraobserver difference analysis). In these assessments, for ICC, 0.250-0.499 indicates poor consistency, 0.500-0.699 indicates moderate consistency, 0.700-0.899 indicates good consistency, and ICC ≥0.900 indicates excellent consistency. A *P*-value <.05 was considered to indicate a statistically significant difference.

Ethical Considerations

Human Subject Ethics Review Approvals

This study was approved by the Medical Ethics Committee of the Children's Hospital of Fudan University (reference number: 2019-184). The content and process of this study comply with the ethical requirements for biomedical research issued by international and national authorities.

Informed Consent Privacy and Confidentiality

The children included in the study or their guardians signed the informed consent form. In the initial informed consent form, participants were thoroughly informed that this study would involve secondary analysis of the Cobb angle data obtained from the collected spinal X-rays.

Privacy and Confidentiality

In the initial informed consent form, it was explicitly stated that the study would anonymize personal information, such as name, gender, birth date, and ID, when measuring the Cobb angle data from spinal X-rays. Access to the final summary tables containing original data with patient identifiers was restricted, allowing only authorized research team members to access sensitive information. At the outset of the study, clear data protection policies and standard operating procedures were established to ensure that all researchers are aware of and comply with data protection regulations.

Results

General Information

This study included 601 patients (aged from 10 to 17 years; mean 12.75, SD 1.54 years). Of these 601 patients, 89 were male (14.8%) and 512 were female (85.2%). Among the 802 full-length orthopedic films, 383 showed mild severity, 349 showed moderate severity, and 70 showed severe severity. The results of manually measuring the Cobb angle using the PACS method are shown in Table 1.

Table 1. Manual measurement of the Cobb angle by the Picture Archiving and Communication System (PACS) method.

Severity[a]	Cobb angle (°)	
	Mean (SD)	Maximum
Mild (n=383)	19.43 (3.35)	24.92
Moderate (n=349)	32.99 (5.52)	44.93
Severe (n=70)	52.04 (6.51)	75.02

[a]Mild: Cobb <25°, moderate: 25° ≤ Cobb ≤ 45°, and severe: Cobb >45°.

Intra- and Interobserver Consistency Test

The ICC values for the 3 measurers using the PACS method before and after 2 measurements were all greater than 0.9, indicating very good intra- and interobserver consistency. The ICC results for the 2 measurements using the mobile app method are shown in Table 2. Both methods demonstrated excellent intra- and interobserver consistency. The intraobserver ICC and interobserver ICC using the mobile photography method were superior to those of the PACS method.

XSL•FO
RenderX

Table 2. Intra- and interobserver agreement analysis.

Variable	PACS[a] method, ICC[b] (95% CI)	Photography method, ICC (95% CI)	*P* value
Intraobserver 1	0.989 (0.987-0.990)	0.996 (0.995-0.996)	<.001
Intraobserver 2	0.983 (0.981-0.986)	0.996 (0.995-0.997)	<.001
Intraobserver 3	0.984 (0.981-0.986)	0.996 (0.996-0.997)	<.001
Interobserver	0.992 (0.991-0.993)	0.997 (0.997-0.998)	<.001

[a]PACS: Picture Archiving and Communication System.

[b]ICC: intraclass correlation coefficient.

Reliability Analysis of the AI App

Based on the excellent interobserver and intraobserver consistency of the PACS method, we used the PACS method as a reference to evaluate the reliability of the AI app measurement capability. We compared 2 app measurements (photography and photo upload) with the PACS method for measuring Cobb angles, and evaluated the results using ICC tests, as shown in Table 3. In this test, the ICC values for the comparison of the photography method and PACS method by 3 observers were greater than 0.9 and statistically significant (*P*<.001). Similarly, the ICC values for the photo upload method and PACS method were greater than 0.9 and statistically significant (*P*<.001). This indicated that there is good consistency between the app measurement and manual measurement results, suggesting that these 2 methods of measuring the Cobb angle are equivalent.

Table 3. Consistency analysis of measurement results between applied measurement methods and the Picture Archiving and Communication System (PACS) method.

Observer	PACS[a] method, mean (SD)	Photography method		Photo upload method	
		Mean (SD)	ICC[b]	Mean (SD)	ICC
Observer 1	28.16 (10.90)	28.03 (11.21)	0.973	28.14 (11.33)	0.972
Observer 2	28.16 (10.93)	28.05 (11.22)	0.971	28.14 (11.33)	0.973
Observer 3	28.20 (10.95)	28.02 (11.18)	0.971	28.14 (11.33)	0.972

[a]PACS: Picture Archiving and Communication System.

[b]ICC: intraclass correlation coefficient.

When measuring the Cobb angle using the photography and photo upload methods, the mean absolute errors were very small (all less than 5°) compared to the PACS method, as shown in Table 4. It is generally considered that an increase of more than 5° in the Cobb angle between 2 consecutive X-ray examinations indicates the progression of scoliosis. Clinically, the allowable measurement error for the Cobb angle is not to exceed 5° [16]. Thus, we defined a measurement error of <5° as "accurate" and took the measurement results of the PACS method as the reference standard. As shown in Table 5, the smartphone app demonstrated a high accuracy rate in measuring the Cobb angle. The relationship between the mean signed differences of the 2 app methods and the PACS method, as shown in Figure 4, indicated that for all subjects, the 95% LoAs for the differences were all less than 5°.

Table 4. Cobb angle error of the main curvature from the smartphone app.

Function and severity	Error	
	Mean (SD)	Maximum
Photography	2.00 (1.43)	10.96
Mild (n=383)	1.87 (1.33)	7.43
Moderate (n=349)	2.12 (1.47)	10.96
Severe (n=70)	2.00 (1.64)	6.33
Photo upload	2.08 (1.38)	5.83
Mild (n=383)	1.93 (1.32)	5.57
Moderate (n=349)	2.21 (1.39)	5.83
Severe (n=70)	2.27 (1.63)	5.83

XSL•FO
RenderX

Table 5. Accuracy analysis of the smartphone app.

Function and severity	Error of <5°, n	Accuracy rate, %
Photography		
Mild (n=383)	375	97.9
Moderate (n=349)	339	97.1
Severe (n=70)	65	92.9
Photo upload		
Mild (n=383)	373	97.4
Moderate (n=349)	337	96.6
Severe (n=70)	63	90.0

Figure 4. The mean signed differences between the Picture Archiving and Communication System (PACS) method and the photography method (A) and photo upload method (B).

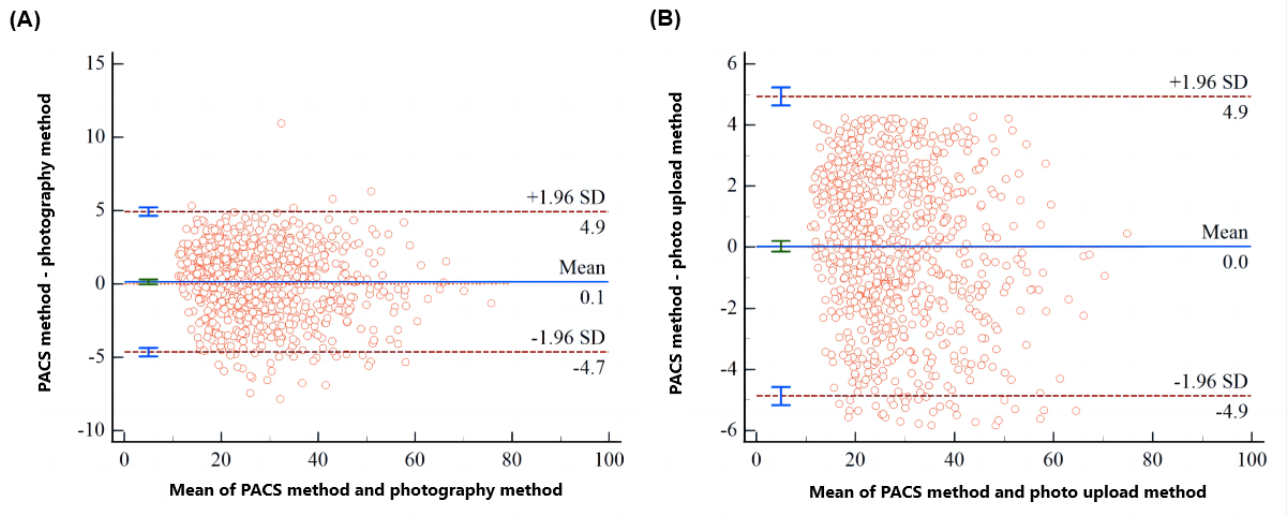

We categorized subjects into 3 subgroups based on the severity of spinal curvature in our study. The mean signed differences and 95% LoAs between the PACS method and photography method across the 3 subgroups are depicted in Figure 5A-C. The outcomes for the mean signed differences and 95% LoAs between the PACS method and photo upload method in these subgroups are presented in Figure 5D-F. Lastly, we focused on subjects with Cobb angles between 25° and 35° for manual measurement and analyzed the mean signed difference between the photography method and PACS method. The 95% LoAs for the difference were precisely –5° to 5°, as shown in Figure 5G. This indicates that AI-based app measurements of the Cobb angle in patients with mild scoliosis are more reliable, and the app measurements could be more accurate for patients with scoliosis of less than 35°.

XSL•FO
RenderX

Figure 5. The mean signed differences between the Picture Archiving and Communication System (PACS) method and photography method for mild scoliosis (A), moderate scoliosis (B), and severe scoliosis (C). The mean signed differences between the PACS method and photo upload method for mild scoliosis (D), moderate scoliosis (E), and severe scoliosis (F). The mean signed differences between the photography method and PACS method for moderate scoliosis with Cobb angles between 25° and 35° (G), and between 35° and 45° (H).

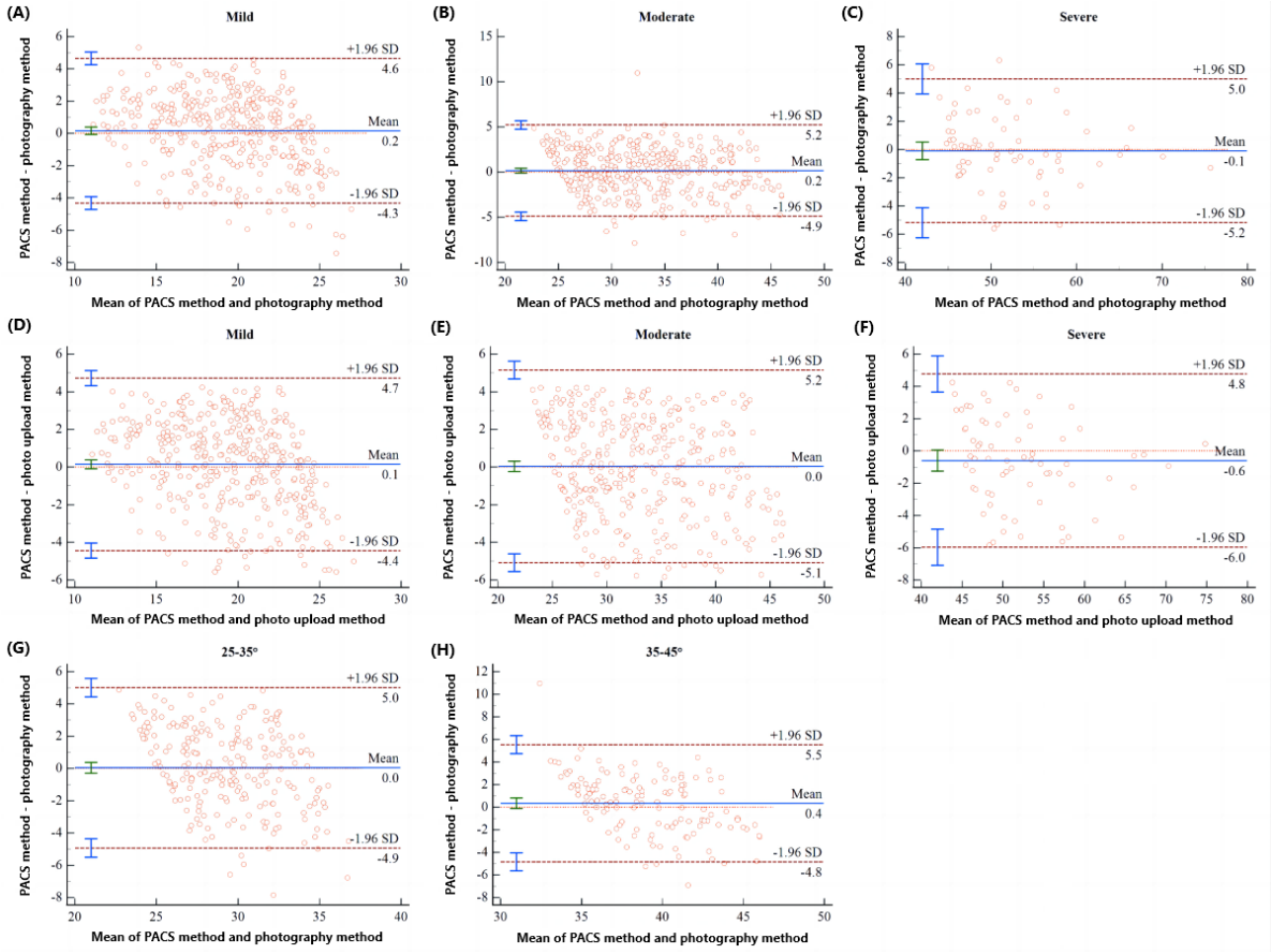

Endplate Selection Variability

Observer 1 used the PACS method to select the upper and lower end vertebrae on 670 full-length anteroposterior and lateral spine X-rays of 542 children and recorded the results. After an interval of 2 weeks, the end vertebrae were selected and recorded again using the mobile photo album upload function, comparing the consistency of end vertebrae selection. It is important to note that the mobile app does not directly provide the selection results of the upper and lower end vertebrae. Instead, it displays 2 white extension lines representing the major curve angle. The spinal segments that these white lines pass through are manually recorded and considered as the upper and lower end vertebrae determined by the mobile app's algorithm. Table 6 provides details of the difference in upper and lower endplate selection between successive pairs of phone upload method and PACS method measurements. As shown in Table 6, the selection of end vertebrae was not completely consistent between the 2 measurement methods (46.9% agreement in upper endplate selection vs 57.8% agreement in lower endplate selection).

Figure 6 shows the measurement of end vertebrae for a Lenke 5–type scoliosis patient using 2 methods, which revealed that the upper vertebral selection of the app method was 1 level lower than that of the PACS method (T11 to L4 for the PACS method vs T12 to L4 for the photo upload method).

XSL•FO
RenderX

Table 6. Comparison of differences in spinal end selection between the photo upload method and Picture Archiving and Communication System (PACS) method by observer 1.

Vertebral selection	Value (N=670), n (%)
Upper vertebral selection	
The upper vertebral selection of the app method is 1 level higher than that of the PACS[a] method	62 (9.3)
Select the same end vertebra level	314 (46.9)
The upper vertebral selection of the app method is 1 level lower than that of the PACS method	294 (43.9)
Lower vertebral selection	
The lower vertebral selection of the app method is 2 levels higher than that of the PACS method	11 (1.6)
The lower vertebral selection of the app method is 1 level higher than that of the PACS method	81 (12.1)
Select the same end vertebra level	387 (57.8)
The lower vertebral selection of the app method is 1 level lower than that of the PACS method	179 (26.7)
The lower vertebral selection of the app method is 2 levels lower than that of the PACS method	12 (1.8)

[a]PACS: Picture Archiving and Communication System.

Figure 6. Lenke 5 scoliosis (A), with the upper and lower end vertebrae measured by the Picture Archiving and Communication System (PACS) method being T11 and L4, respectively (B). (C, D) The scoliosis curve depicted by the photo upload method shows that the maximum scoliosis angle is between T12 and L4.

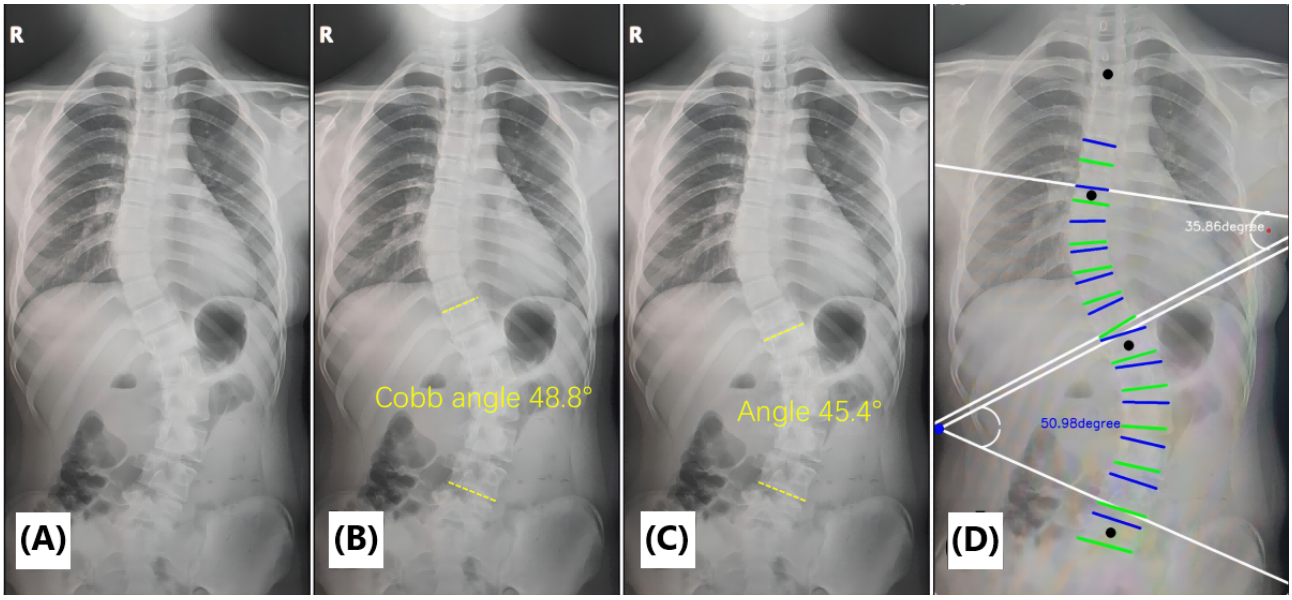

Discussion

Principal Findings

Our AI mobile app for measuring the Cobb angle can accurately and automatically mark the coordinates of the apex for each vertebra, fit the spinal curvature, and perform Cobb angle measurements. Evaluators only need to check that the mobile capture function of the software includes a complete X-ray image of the child's spine. Additionally, this AI mobile app automatically calculates the Cobb angle for children with scoliosis and marks the end vertebrae. Analysis of the Cobb angle measurement results indicates that this mobile app shows effectiveness when compared to the PACS.

Carman et al [9] pointed out that intraobserver variability is an important parameter in the clinical diagnosis and treatment of scoliosis, as intraobserver differences may lead to misdiagnosis of curve progression, thereby impacting clinical treatment decisions. However, Shaw et al [11] emphasized that interobserver consistency is equally important when measuring the Cobb angle, as in many cases the same patient may not be measured by the same physician. In this study, for the mobile app, the intraobserver ICC was 0.996 and the interobserver ICC was 0.997, both of which exceeded the values for the PACS method. The mobile app demonstrated significantly better intraobserver and interobserver reproducibility when measuring the Cobb angle. Good consistency can effectively prevent medical errors caused by measurement inaccuracies.

Compared to the PACS method, the mobile app method showed significantly better intra- and interobserver repeatability in measuring the Cobb angle. Good consistency can effectively avoid medical errors caused by measurement errors in patients.

In order to explore the accuracy of the software measurement, we took the Cobb angle results obtained by the PACS method

XSL•FO
RenderX

as a reference and compared the differences between the photography and upload methods and the PACS method. We found that the errors in the Cobb angle measured by the photography and upload methods and the PACS method were 2.00 and 2.08, respectively. The accuracy of the mobile phone app photography function in measuring patients with mild and moderate scoliosis was greater than 97%. Compared with the PACS method, the mean differences in the Cobb angle measured by the photography and upload methods were 0.1° and 0.03°, respectively, and the 95% consistency limits were –4.7° to 4.9° and –4.9° to 4.9°, respectively, which are less than the allowable error range (5°) [10]. Thus, the Cobb angle measured by the mobile phone app can be considered to be reliable compared with the traditional PACS manual measurement.

However, we believe that the current accuracy of the app will decrease with the severity of the patient's scoliosis. The results of subgroup analysis showed that there was a difference of >5° between mild and moderate severity scoliosis. We assume that the critical value of the Cobb angle with this difference may be in patients with moderate scoliosis. In order to find the critical value for the app, we attempted to use the photography method to measure the X-ray film of patients with a Cobb angle ≤35°, and the results showed that the 95% consistency limit was just –5° to 5°, which is in line with the maximum measurement error range allowed by clinical practice. The results suggest that the AI app is more accurate in measuring patients with scoliosis with a Cobb angle ≤35°. The reason for this is likely the insufficient deep learning data for X-ray images of severe scoliosis, and further data input is needed. It just fits the applicable population of the software: patients with scoliosis who need close and regular follow-up.

Comparison With Prior Work

We reviewed the intraobserver and interobserver variability and errors of other smart apps for measuring the Cobb angle, such as CobbMeter and Tiltmeter software. For CobbMeter, the intraobserver ICC was 0.983 and interobserver ICC was 0.973 [14]. In contrast, both the intraobserver and interobserver ICC for our app were greater than 0.99, indicating that the software program in this study has more considerable repeatability in measuring the Cobb angle compared to the CobbMeter app. The Tiltmeter software had an average absolute error of 2.1° (SD 1.7°) in measuring the Cobb angle, while our mobile app had an error of only 2.0° (SD 1.43°), showing that our mobile app has smaller errors in measuring the Cobb angle compared to similar apps and demonstrating superior performance in measuring the Cobb angle among apps of its kind [10].

Similar to this, the mobile intelligent app NLMeasurer uses a PoseNet solution based on computer vision and machine learning for human posture assessment and anatomical point identification. Research by Moreira et al [13] demonstrated that posture measurements calculated using NLMeasurer were consistent with validated biophotogrammetry software SAPO, proving it to be an effective tool for evaluating posture measurements in the frontal view. Consistent with our study's conclusions, these mobile apps show superior reliability and convenience in measuring body posture angles or Cobb angles for scoliosis compared to traditional computer-based methods like the PACS or SAPO. Additionally, the NLMeasurer app exhibited high interobserver and intraobserver reliability similar to our AI Cobb angle measurement app.

However, unlike the NLMeasurer mobile app, our developed AI app for measuring the Cobb angle has a significant advantage in that it requires no manual identification or annotation of key anatomical landmarks (end vertebrae) on the captured images to accurately measure the Cobb angle. While the NLMeasurer app can also perform posture measurements directly on unmarked photos, it benefits from using surface markers on specific anatomical landmarks (such as the ears, iliac crests, and ankle joints) to facilitate digital recognition of these points, thereby improving the reliability of posture measurements performed with NLMeasurer. Therefore, to enhance the accuracy of posture assessment using photos without surface anatomical markers, the NLMeasurer app would still need to optimize its related algorithms.

Strengths and Limitations

The AI Mobile App Can Fully Automate Cobb Angle Measurement, Reducing Errors Associated With Manually Marking End Vertebrae

With the rapid advancement of computer technology, computer-aided methods for measuring the Cobb angle have been widely applied in clinical practice and have proven to have reliability, good correlation, and measurement accuracy compared to traditional manual measurements [17-19]. Digital computer-aided measurement methods, such as Surgimap software [20] and the PACS, exemplify this trend. With the popularity of smartphones, researchers have developed numerous mobile apps for measuring the Cobb angle. These apps, such as Scoligauge, Tiltmeter, and CobbMeter, are considered to be more reliable and convenient for screening and measuring the Cobb angle in cases of scoliosis or kyphosis compared to other methods [11,21,22].

However, all the aforementioned methods require the examiner to manually identify and mark the upper and lower end vertebrae of the scoliosis patient before measuring the Cobb angle, which means the measurement results may be biased by the examiner's ability to accurately identify the end vertebrae. In contrast, our AI mobile app truly achieves fully automated Cobb angle measurement, allowing users to take a photo or upload an image to measure the Cobb angle without needing to premark the end vertebrae. This functionality provides flexibility during the assessment process and reduces the likelihood of human errors when placing markers.

The AI Mobile App Offers Flexible Multiscenario Usage Compared to Large Computers

The PACS has not yet been universally adopted in all hospitals, which limits the ability of clinical physicians to use it in certain situations. When clinicians consult with patients outside the hospital, the computer-based PACS cannot be accessed. If patients bring soft or hard copies of digital X-rays taken at other hospitals, doctors cannot retrieve the imaging data from their hospitals' computers, forcing them to either manually measure the Cobb angle or have the patient undergo repeat X-rays at

XSL•FO
RenderX

their facility. The former reduces result reliability, while the latter increases the patient's radiation risk and costs. In these scenarios, the mobile app can be used flexibly, allowing clinicians to measure the Cobb angle in children with scoliosis without needing a computer. When patients wish to have online consultations, they can send their spinal X-ray images via the internet, enabling doctors to upload and measure the Cobb angle directly using the app, thus assisting clinicians in determining whether the patient's scoliosis has progressed. When patients consult with X-rays taken at outside facilities, they do not need to undergo repeat imaging, and clinicians can quickly take pictures and measure the Cobb angle using the software.

The Use of the AI Mobile App in Guiding the Formulation of Surgical Strategies Holds Significant Potential Value

When measuring patients with moderate to severe spinal curvature, some discrepancies still exist between the results obtained by the app and the PACS. These differences may arise partly from insufficient data for deep learning, and we suspect that the algorithm itself also plays a role. Unlike manual Cobb angle measurements, the app does not first identify the upper and lower vertebrae of the scoliosis. Instead, it uses AI to plot the spinal curve in the coronal plane and subsequently determines the maximum angle of curvature on that curve. Therefore, we believe that the Cobb angle measured by the app is more accurate than that obtained via the PACS method.

To address these discrepancies, we reanalyzed the relationship between the segment of the Cobb angle line depicted by the app and the upper and lower vertebrae identified by the PACS method. Our findings indicated that, in some instances, the upper and lower vertebrae determined by the 2 methods differed by 1 to 2 vertebral bodies (Table 6). In cases of idiopathic scoliosis, this definition directly influences the surgical strategy for correcting spinal curvature, highlighting the significance of an app that can measure vertebrae more accurately for patients (Figure 6). The introduction of this app may help preserve the patient's range of motion by potentially avoiding the need to address the last segment of the spine during surgical planning. In future studies, we will explore orthotic correction and surgical treatment of patients based on the vertebrae and Cobb angle measurements provided by the app.

Limitations

Despite the encouraging results obtained from the use of the AI Cobb angle measurement app, it is essential to highlight the limitations of this study. First, when the app identifies the primary curve for measurement, it also provides a Cobb angle measurement for a compensatory curve. We need follow-up experiments to validate the app's reliability in measuring smaller compensatory curves. Second, we believe that there are still errors when measuring the Cobb angle with this app. However, since the app is based on deep learning algorithms, its accuracy can continuously improve. During the data model generation process, more images of abnormal spines can be incorporated, allowing the network to train on various boundary cases, thereby providing a traceable basis for testing.

Additionally, a Fourier transform module can be added to the transformer module to reduce computational load and enhance efficiency through the combination of the 2 modules. Furthermore, based on the practical situations encountered during the use of this model, qualified experts could provide direct feedback on the angles determined by the machine via the human-computer interface. This feedback could serve as part of the model's penalty function, further improving the model's accuracy and iterative optimization efficiency. We believe that through the deep learning capabilities of AI, the measurement accuracy of the app will continue to improve in the future.

Conclusion

This newly developed AI platform mobile app measures the main curvature Cobb angle with reliable results and high accuracy, which can provide convenience for clinicians and is worth promoting in clinical application.

Acknowledgments

This work was supported by the Ministry of Science and Technology of China Science and Technology Innovation 2030 (No. 2021ZD0113405).

Data Availability

Data will be made available upon reasonable request.

Authors' Contributions

HL contributed to methodology and investigation. CQ contributed to software, visualization, and writing – original draft. DF contributed to conceptualization and supervision. YZ contributed to visualization and writing – review & editing. ZZ contributed to methodology and software. JM contributed to conceptualization and writing – original draft. WY contributed to methodology and supervision. DW contributed to supervision, writing – review & editing, and funding acquisition.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Altaf F, Gibson A, Dannawi Z, Noordeen H. Adolescent idiopathic scoliosis. BMJ 2013 Apr 30;346(apr30 1):f2508-f2508. [doi: 10.1136/bmj.f2508]
2. Weinstein SL, Dolan LA, Cheng JC, Danielsson A, Morcuende JA. Adolescent idiopathic scoliosis. The Lancet 2008 May;371(9623):1527-1537. [doi: 10.1016/s0140-6736(08)60658-3]
3. Peng Y, Wang S, Qiu G, Zhang J, Zhuang Q. Research progress on the etiology and pathogenesis of adolescent idiopathic scoliosis. Chin Med J (Engl) 2020 Feb 20;133(4):483-493 [FREE Full text] [doi: 10.1097/CM9.0000000000000652] [Medline: 31972723]
4. Jin C, Wang S, Yang G, Li E, Liang Z. A review of the methods on Cobb angle measurements for spinal curvature. Sensors 2022 Apr 24;22(9):3258. [doi: 10.3390/s22093258]
5. Romano M, Minozzi S, Bettany-Saltikov J, Zaina F, Chockalingam N, Kotwicki T, et al. Exercises for adolescent idiopathic scoliosis. Cochrane Database Syst Rev 2012 Aug 15;2012(8):CD007837 [FREE Full text] [doi: 10.1002/14651858.CD007837.pub2] [Medline: 22895967]
6. Kuklo TR, Potter BK, Polly DW, O'Brien MF, Schroeder TM, Lenke LG. Reliability analysis for manual adolescent idiopathic scoliosis measurements. Spine 2005;30(4):444-454. [doi: 10.1097/01.brs.0000153702.99342.9c]
7. Morrissy RT, Goldsmith GS, Hall EC, Kehl D, Cowie GH. Measurement of the Cobb angle on radiographs of patients who have scoliosis. Evaluation of intrinsic error. The Journal of Bone & Joint Surgery 1990;72(3):320-327. [doi: 10.2106/00004623-199072030-00002]
8. Wong J, Reformat M, Parent E, Stampe K, Southon Hryniuk S, Lou E. Validation of an artificial intelligence-based method to automate Cobb angle measurement on spinal radiographs of children with adolescent idiopathic scoliosis. Eur J Phys Rehabil Med 2023 Aug;59(4):535-542 [FREE Full text] [doi: 10.23736/S1973-9087.23.08091-7] [Medline: 37746786]
9. Carman DL, Browne RH, Birch JG. Measurement of scoliosis and kyphosis radiographs. Intraobserver and interobserver variation. J Bone Joint Surg Am 1990 Mar;72(3):328-333. [Medline: 2312528]
10. Langensiepen S, Semler O, Sobottke R, Fricke O, Franklin J, Schönau E, et al. Measuring procedures to determine the Cobb angle in idiopathic scoliosis: a systematic review. Eur Spine J 2013 Feb 27;22(11):2360-2371. [doi: 10.1007/s00586-013-2693-9]
11. Shaw M, Adam CJ, Izatt MT, Licina P, Askin GN. Use of the iPhone for Cobb angle measurement in scoliosis. Eur Spine J 2011 Nov 9;21(6):1062-1068. [doi: 10.1007/s00586-011-2059-0]
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan 20;28(1):31-38. [doi: 10.1038/s41591-021-01614-0]
13. Moreira R, Fialho R, Teles AS, Bordalo V, Vasconcelos SS, Gouveia GPDM, et al. A computer vision-based mobile tool for assessing human posture: A validation study. Computer Methods and Programs in Biomedicine 2022 Feb;214:106565. [doi: 10.1016/j.cmpb.2021.106565]
14. Qiao J, Liu Z, Xu L, Wu T, Zheng X, Zhu Z, et al. Reliability analysis of a smartphone-aided measurement method for the Cobb angle of scoliosis. J Spinal Disord Tech 2012 Jun;25(4):E88-E92. [doi: 10.1097/BSD.0b013e3182463964] [Medline: 22237178]
15. Ketenci İ, Yanık HS, Erdoğan Ö, Adıyeke L, Erdem Ş. Reliability of 2 smartphone applications for Cobb angle measurement in scoliosis. Clin Orthop Surg 2021 Mar;13(1):67-70 [FREE Full text] [doi: 10.4055/cios19182] [Medline: 33747380]
16. Hughes J, Yaszay B, Bastrom T, Bartley C, Parent S, Cahill P, Harms Study Group. Long-term patient perception following surgery for adolescent idiopathic scoliosis if dissatisfied at 2-year follow-up. Spine (Phila Pa 1976) 2021 Apr 15;46(8):507-511. [doi: 10.1097/BRS.0000000000003828] [Medline: 33273434]
17. Kuklo TR, Potter BK, Schroeder TM, O'Brien M. Comparison of manual and digital measurements in adolescent idiopathic scoliosis. Spine (Phila Pa 1976) 2006 May 15;31(11):1240-1246. [doi: 10.1097/01.brs.0000217774.13433.a7] [Medline: 16688038]
18. Wills BPD, Auerbach JD, Zhu X, Caird MS, Horn BD, Flynn JM, et al. Comparison of Cobb angle measurement of scoliosis radiographs with preselected end vertebrae: traditional versus digital acquisition. Spine (Phila Pa 1976) 2007 Jan 01;32(1):98-105. [doi: 10.1097/01.brs.0000251086.84420.d1] [Medline: 17202899]
19. Rosenfeldt MP, Harding IJ, Hauptfleisch JT, Fairbank JT. A comparison of traditional protractor versus Oxford Cobbometer radiographic measurement: intraobserver measurement variability for Cobb angles. Spine (Phila Pa 1976) 2005 Feb 15;30(4):440-443. [doi: 10.1097/01.brs.0000153401.78638.cb] [Medline: 15706342]
20. Akbar M, Terran J, Ames CP, Lafage V, Schwab F. Use of Surgimap Spine in sagittal plane analysis, osteotomy planning, and correction calculation. Neurosurg Clin N Am 2013 Apr;24(2):163-172. [doi: 10.1016/j.nec.2012.12.007] [Medline: 23561555]
21. Franko O, Bray C, Newton P. Validation of a scoliometer smartphone app to assess scoliosis. J Pediatr Orthop 2012 Dec;32(8):e72-e75. [doi: 10.1097/BPO.0b013e31826bb109] [Medline: 23147635]
22. Jacquot F, Charpentier A, Khelifi S, Gastambide D, Rigal R, Sautet A. Measuring the Cobb angle with the iPhone in kyphoses: a reliability study. Int Orthop 2012 Aug;36(8):1655-1660 [FREE Full text] [doi: 10.1007/s00264-012-1579-5] [Medline: 22653103]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
ICC: intraclass correlation coefficient
LoAs: limits of agreement
PACS: Picture Archiving and Communication System

Edited by A Schwartz; submitted 07.07.23; peer-reviewed by A Teles, B Puladi, X Chen; comments to author 29.01.24; revised version received 01.03.24; accepted 04.10.24; published 01.11.24.

Please cite as:
Li H, Qian C, Yan W, Fu D, Zheng Y, Zhang Z, Meng J, Wang D
Use of Artificial Intelligence in Cobb Angle Measurement for Scoliosis: Retrospective Reliability and Accuracy Study of a Mobile App
J Med Internet Res 2024;26:e50631
URL: https://www.jmir.org/2024/1/e50631
doi:10.2196/50631
PMID:

©Haodong Li, Chuang Qian, Weili Yan, Dong Fu, Yiming Zheng, Zhiqiang Zhang, Junrong Meng, Dahui Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Based Prediction for Incident Hypertension Based on Regular Health Checkup Data: Derivation and Validation in 2 Independent Nationwide Cohorts in South Korea and Japan

Seung Ha Hwang[1,2*], MS; Hayeon Lee[1,2*], PhD; Jun Hyuk Lee[3*], BS; Myeongcheol Lee[2,4], MS; Ai Koyanagi[5], MD, PhD; Lee Smith[6], PhD; Sang Youl Rhee[2,4,7], MD, PhD; Dong Keon Yon[2,4,8], MD, PhD; Jinseok Lee[1], PhD

[1]Department of Biomedical Engineering, Kyung Hee University, Yongin, Republic of Korea
[2]Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[3]Health and Human Science, University of Southern California, Los Angeles, CA, United States
[4]Department of Regulatory Science, Kyung Hee University, Seoul, Republic of Korea
[5]Research and Development Unit, Parc Sanitari Sant Joan de Deu, Barcelona, Spain
[6]Centre for Health, Performance and Wellbeing, Anglia Ruskin University, Cambridge, United Kingdom
[7]Department of Endocrinology and Metabolism, Kyung Hee University School of Medicine, Seoul, Republic of Korea
[8]Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Jinseok Lee, PhD
Department of Biomedical Engineering
Kyung Hee University
1732 Deogyeong-daero
Yongin, 17104
Republic of Korea
Phone: 82 312012570
Fax: 82 312012571
Email: gonasago@khu.ac.kr

Abstract

Background: Worldwide, cardiovascular diseases are the primary cause of death, with hypertension as a key contributor. In 2019, cardiovascular diseases led to 17.9 million deaths, predicted to reach 23 million by 2030.

Objective: This study presents a new method to predict hypertension using demographic data, using 6 machine learning models for enhanced reliability and applicability. The goal is to harness artificial intelligence for early and accurate hypertension diagnosis across diverse populations.

Methods: Data from 2 national cohort studies, National Health Insurance Service-National Sample Cohort (South Korea, n=244,814), conducted between 2002 and 2013 were used to train and test machine learning models designed to anticipate incident hypertension within 5 years of a health checkup involving those aged ≥20 years, and Japanese Medical Data Center cohort (Japan, n=1,296,649) were used for extra validation. An ensemble from 6 diverse machine learning models was used to identify the 5 most salient features contributing to hypertension by presenting a feature importance analysis to confirm the contribution of each future.

Results: The Adaptive Boosting and logistic regression ensemble showed superior balanced accuracy (0.812, sensitivity 0.806, specificity 0.818, and area under the receiver operating characteristic curve 0.901). The 5 key hypertension indicators were age, diastolic blood pressure, BMI, systolic blood pressure, and fasting blood glucose. The Japanese Medical Data Center cohort dataset (extra validation set) corroborated these findings (balanced accuracy 0.741 and area under the receiver operating characteristic curve 0.824). The ensemble model was integrated into a public web portal for predicting hypertension onset based on health checkup data.

Conclusions: Comparative evaluation of our machine learning models against classical statistical models across 2 distinct studies emphasized the former's enhanced stability, generalizability, and reproducibility in predicting hypertension onset.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e52794) doi:10.2196/52794

KEYWORDS

machine learning; hypertension; cardiovascular disease; artificial intelligence; cause of death; cardiovascular risk; predictive analytics

Introduction

The World Health Organization (WHO) has identified cardiovascular diseases (CVDs) as the leading cause of mortality worldwide, with a staggering 17.9 million deaths recorded in 2019 [1]. This number is projected to rise to approximately 23 million by 2030. Of the multitude of CVDs, specific conditions such as myocardial infarction and ischemic stroke account for more than 85% of these CVD-related deaths [2]. The US Centers for Disease Control and Prevention (CDC) have highlighted that CVDs caused over US $216 billion in overall health care expenses and resulted in US $147 billion lost due to increased workplace absenteeism and corresponding productivity in the United States. As a result, CVDs impose a significant burden on the nation's economy [3].

Given the acknowledged biological and economic risks associated with CVDs, it is widely recognized that hypertension plays a significant role in these health complications, including myocardial infarction and stroke [4]. Predicting hypertension onset is notably challenging due to the disease's multifactorial origins, encompassing a wide range of genetic, environmental, and lifestyle factors. The subtle and often interrelated effects of these factors contribute to the complexity of early detection. For example, genetic predispositions may interact with lifestyle choices such as diet, exercise, and smoking habits, in ways that are not fully understood [5]. Environmental influences, including socioeconomic status and access to health care, further complicate the picture by affecting both the risk of developing hypertension and the ability to manage risk factors effectively [5,6]. Additionally, the asymptomatic nature of hypertension in its early stages means that it often goes unnoticed until more serious health issues arise, making timely and accurate prediction all the more difficult [7]. These challenges underscore the need for sophisticated predictive models that can integrate and analyze the myriad of contributing factors to identify individuals at risk of developing hypertension early in its progression. Considering the severe societal implications of hypertension across all nations, early diagnosis is crucial to mitigate its potential hazards. In this study, we propose a novel approach to predict the onset of hypertension using the population's regular health checkup and demographic factors. In recent years, machine learning models have emerged as powerful tools across many fields, particularly in medical applications [8]. Their ability to analyze complex patterns and make accurate predictions has revolutionized how we approach health care challenges.

However, ensuring this methodology's replicability and broad applicability in real-world settings presents an intricate challenge. To bolster the reliability of our hypertension projections, we conducted additional independent validation using distinct cohorts. This study investigated various machine learning approaches to strengthen the method's robustness, replicability, and real-world practicality. We delved into the hypertension landscape across Asian populations through machine learning optics, firmly anchoring our methodology within the burgeoning realm of artificial intelligence (AI)–driven disciplines. This research endeavors to amplify our comprehension of global hypertension trends by channeling multifaceted machine learning analyses, thereby catalyzing more timely and precise diagnostic efforts.

Methods

Data Source

We used 2 national, large-scale, and general population–based cohort studies: the National Health Insurance Service-National Sample Cohort (NHIS-NSC; N=973,303) and the Japanese Medical Data Center cohort (JMDC; N=12,143,715). This study was approved by the institutional review board of National Health Insurance Service, Kyung Hee University (KHSIRB-23-085[EA]), and the JMDC (PHP-00002201-04). The requirement for informed consent was waived as this study used deidentified administrative data.

NHIS-NSC (Discovery Cohort)

The NHIS-NSC [9], the population-based, nationwide, and large-scale cohort of South Korea, were from those aged ≥20 years who received general health checkups between January 1, 2002, and December 31, 2013. We used the NHIS-NSC to train, validate, and test the machine learning model to predict the presence or absence of hypertension within 5 years of a regular (yearly) health checkup. Hypertension was defined for patients who had received diagnoses with I10, I11, I12, I13, or I15 codes from the *ICD-10* (*International Classification of Disease*, *10th revision*) ≥2 times and were using antihypertensives [10].

During the data preprocessing phase, we transformed the cohort into a machine learning dataset by representing each eligible individual once, with all features recorded from their initial health check-up. The ground truth was determined by the occurrence of a hypertension event within the subsequent 5 years. We excluded participants with baseline hypertension or those lost to follow-up from this study. Individuals who developed hypertension after 5 years were classified as nonhypertensive for this study.

In this study, we excluded participant information that fulfilled one of the following criteria among the 973,303 registered participants: (1) those who had reported "yes" for hypertension in the questionnaire; (2) those who had a prior diagnosis of hypertension with I10, I11, I12, I13, or I15 codes of *ICD-10* before the health checkup; (3) those with missing data for information and questionnaire; (4) those who had died before the year 2013; and (5) those who have blood pressure over the criteria of hypertension (systolic blood pressure ≥140 mm Hg

XSL•FO
RenderX

or the diastolic blood pressure is ≥90 mm Hg). A graphical representation of the subject exclusion process of the NHIS-NSC is illustrated in Figure 1 [11].

Figure 1. Study population and data selection process in the NHIS-NSC (Korea) and JMDC (Japan). NHIS-NSC: National Health Insurance Service-National Sample Cohort; JMDC: Japanese Medical Data Center cohort.

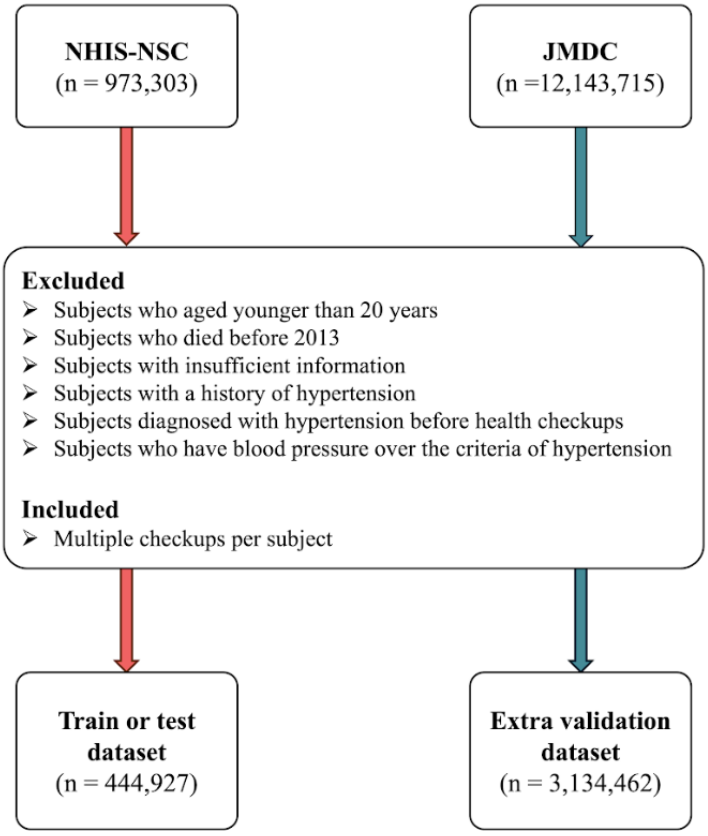

JMDC (Validation Cohort)

The JMDC dataset is the medical examination data from multiple health insurance associations in Japan since 2005 [12-14]. Given the machine learning model trained from the NHIS-NSC, we used the JMDC data for extra validation. We also applied the same exclusion criteria used in the NHIS-NSC to the JMDC data, resulting in the use of only 1,296,649 participant data of the total 12,143,715 data available. A graphical representation of the subject exclusion process of the JMDC dataset is also illustrated in Figure 1.

Study Design

To develop the machine learning model for predicting the presence or absence of hypertension within 5 years of a regular (yearly) health checkup, we used the following 18 available variables as the model's input: age, sex, region of residence, household income, systolic blood pressure, diastolic blood pressure, fasting blood glucose, serum total cholesterol, hemoglobin, aspartate transaminase (AST), alanine transaminase (ALT), γ-glutamyl transpeptidase (γ-GTP), BMI, history of diabetes mellitus, history of stroke (including ischemic stroke, hemorrhagic stroke, and/or transient ischemic attack), smoking status, alcohol intake, and physical activity [15]. The variables used in our machine learning model are summarized in Table 1. More specifically, the region of residence was categorized into rural and urban. Household income was categorized into 11 scales (0 to 10) based on basic livelihood recipient and decile (Table S1 in Multimedia Appendix 1); in 10 income deciles, the 5th decile is the reference median income. Compared to 5th decile (100%), 1st decile (the lowest income level) has an income of less than 30%, 2nd decile has an income of less than 50%, 3rd decile has an income of less than 70%, 4th decile has an income of less than 90%, 6th decile has an income of less than 130%, 7th decile has an income of less than 150%, 8th decile has an income of less than 200%, 9th decile has an income of less than 300%, and 10th decile (the highest income level) has an income of 300% or more. Basic livelihood recipients are individuals whose income falls within 1st decile (the lowest 30%) [12]. Smoking status was categorized into never, former, and current smokers. Alcoholic intakes were categorized into rare (less than one time per week), 1-2, 3-4, and more than four times per week [12-14]. Physical activity was categorized into never, 1-2, 3-4, and 5-6 times per week, and every day. The statistical characteristics of the variables for the NHIS-NSC and JMDC are summarized in Tables 1 and 2, respectively.

XSL•FO
RenderX

Table 1. Baseline characteristics of subjects in the discovery cohort (National Health Insurance Service-National Sample Cohort, N=244,814).

Variables	Values
Sex, n (%)	
Male	117,642 (48.05)
Female	127,172 (51.95)
Age (years), mean (SD)	47.03 (13.29)
Region of residence, n (%)	
Urban	111,640 (45.6)
Rural	133,174 (54.4)
Household income, n (%)	
Basic livelihood recipient	437 (0.18)
Income deciles (excluded basic livelihood recipients)	
D1 (the lowest income level; ≤30th percentile)	18,783 (7.67)
D2 (31st-50th percentile)	19,399 (7.92)
D3 (51st-70th percentile)	22,165 (9.05)
D4 (71st-90th percentile)	24,498 (10.01)
D5 (91st-100th percentile)	25,043 (10.23)
D6 (101st-130th percentile)	26,137 (10.68)
D7 (131st-150th percentile)	26,195 (10.7)
D8 (151st-200th percentile)	26,562 (10.85)
D9 (201st-300th percentile)	28,469 (11.63)
D10 (high income level, >300th percentile)	27,126 (11.08)
Systolic blood pressure (mm Hg), mean (SD)	111.6 (9.42)
Diastolic blood pressure (mm Hg), mean (SD)	68.69 (6.1)
Fasting blood glucose (mg/dL), mean (SD)	92 (23.48)
Serum total cholesterol (mg/dL), mean (SD)	187.91 (35.84)
Hemoglobin (g/dL), mean (SD)	13.78 (1.58)
Aspartate transaminase (U/L), mean (SD)	23.82 (15.12)
Alanine transaminase (U/L), mean (SD)	22.9 (21.32)
γ-glutamyl transpeptidase (U/L), mean (SD)	28.81 (37.6)
BMI (kg/m^2), mean (SD)	22.8 (3)
History of diabetes mellitus, n (%)	4596 (1.88)
History of stroke, n (%)	376 (0.15)
Smoking status, n (%)	
Nonsmoker	176,333 (72.03)
Ex-smoker	9148 (3.74)
Current smoker	59,333 (24.24)
Alcohol intake per week, n (%)	
Rarely	182,101 (74.38)
1-2	45,374 (18.53)
3-4	12,337 (5.04)
≥5	5002 (2.04)
Physical activity per week, n (%)	
Never	141,847 (57.94)

XSL•FO
RenderX

Variables	Values
1-2	63,046 (25.75)
3-4	23,426 (9.57)
5-6	5828 (2.38)
Every day	10,667 (4.36)

Table 2. Baseline characteristics of subjects in the validation cohort (Japanese Medical Data Center cohort; N=1,296,649).

Variables	Values
Sex, n (%)	
Male	754,055 (58.15)
Female	542,594 (41.85)
Age (years), mean (SD)	42.51 (10.24)
Systolic blood pressure (mm Hg), mean (SD)	111.65 (10.43)
Diastolic blood pressure (mm Hg), mean (SD)	67.88 (7.55)
Fasting blood glucose (mg/dL), mean (SD)	91.45 (14.31)
Serum total cholesterol (mg/dL), mean (SD)	200.05 (35.25)
Hemoglobin (g/dL), mean (SD)	14.19 (1.56)
Aspartate transaminase (U/L), mean (SD)	20.84 (8.89)
Alanine transaminase (U/L), mean (SD)	21.02 (15.86)
γ-glutamyl transpeptidase (U/L), mean (SD)	30.63 (33.48)
BMI (kg/m^2), mean (SD)	22.12 (3.22)
History of diabetes mellitus, n (%)	14,345 (1.11)
History of stroke, n (%)	3616 (0.28)
Smoking, n (%)	
No	978,245 (75.44)
Yes	318,404 (24.56)
Alcohol intake per week, n (%)	
Rarely	669,090 (51.6)
Sometimes	403,527 (31.12)
Every day	224,052 (17.28)
Physical activity, n (%)	
No	1,082,572 (83.49)
Yes	214,077 (16.51)

Proposed Machine Learning Models

In this study, we split the NHIS-NSC dataset (n=244,814) into train (n=195,851) and internal test (n=48,963) data with a ratio of 8:2 in a stratified fashion. The internal test set was used only for an independent test of our developed AI model and not for training or internal validation. The JMDC (n=1,296,649) was used as the external validation dataset in this study.

The data distribution was severely imbalanced: the ratio of hypertension and nonhypertension group was 1:15.32. To minimize the bias toward the majority group (nonhypertension) of the prediction model, we up-sampled the hypertension data using a synthetic minority oversampling technique during the model update [16]. In addition, in the preprocessing stage, we performed standard scaler normalization for all features: we calculated the mean and SD of each feature from the training dataset and then normalized all feature values from both the test dataset and external validation datasets to have a mean of 0 and a SD of 1.

To predict hypertension occurrence within 5 years based on regular health check-ups, we applied 6 machine learning models from 18 features: Extreme Gradient Boosting, random forest, gradient boosting machine (GBM), Light GBM, Adaptive Boosting (AdaBoost), and logistic regression (LR) [17,18]. Subsequently, we chose the best 3 among the 6 models and applied an ensemble approach by considering all possible

XSL•FO
RenderX

combinations [19]. Performance evaluations were based on 5-fold cross-validation using the train data following metrics: sensitivity, specificity, accuracy, balanced accuracy, and area under the receiver operating characteristics (AUROC) [17,18]. To compare the predictive performance of the models, we performed a Cochrane Q test on the model performance [20,21]. Due to the significant data imbalance, we used balanced accuracy as the primary model evaluation metric. Moreover, we also estimated additional metrics to comprehensively evaluate the performance of each model: precision, F_1-score, and area under the precision-recall curve. To address the issue of inappropriate precision and F_1-score under the severe data imbalance, we measured weighted average precision and weighted average F_1-score, accounting for the differences in class sizes. Finally, we presented its feature importance analysis, listing features in the order they contributed to hypertension prediction within 5 years of regular health checkups.

We implemented the models using Python (version 3.9.16; Python Software Foundation) with TensorFlow (version 2.9.1; Google LLC), Keras (version 2.9.0; Google LLC), NumPy (version 1.21.5; NumFOCUS, Inc), Pandas (version 1.4.4; NumFOCUS, Inc), Matplotlib (version 3.5.2; NumFOCUS, Inc), and Scikit-learn (version 1.0.2; NumFOCUS, Inc) [18,22]. All statistical analysis was performed using SAS (version 9.4, SAS Institute Inc) [22].

Feature Importance

To analyze the effect of each feature on predicting hypertension occurrence, we performed the feature importance analysis to confirm the contribution of each feature. For tree-based models, the mean decrease in impurity (MDI), which is also known as Gini importance, is used to assess feature importance [23,24]. The following equation represents MDI [24]:

where T is the total number of trees in the base estimator, is the set of all nodes in tree t, is the feature used for splitting the node m, is the decrease in impurity at node m, is the Gini impurity, is the number of samples at node m, and is an indicator function that is 1 if feature j is used for splitting at node m and 0 otherwise. Notation of l and r indicates left child node and right child node each. Those equations calculate feature importance by summing up the impurity reductions caused by each feature across all trees in the model. A higher MDI indicates greater feature importance. For LR, we used the regression coefficient to calculate feature importance. The following equation indicates the LR model [25]:

where d is the number of features. The regression coefficient describes the average change in the dependent variable for each 1-unit change in the independent variable for continuous independent variables or the expected difference versus a reference category for categorical independent variables. Further, for ensembled models, we calculated feature importance by averaging the standardized feature importance from each model used for the ensemble.

Risk Factors

We further investigated the association between the occurrence of hypertension and independent variables using univariate and multivariate LR analyses in the discovery and validation cohorts [26,27]. Predictor variables included categorical variables (sex, region of residence, history of diseases, smoking status, alcohol intake, and physical activity) and continuous variables transformed into categorical form (age, household income, BMI, blood pressure, fasting blood glucose, serum total cholesterol, hemoglobin, ALT, AST, and γ-GTP). Univariable and multivariable LR analyses were conducted for each variable to estimate the odds ratio and 95% CI for the occurrence of hypertension. All statistical analyses were conducted using SAS (version 9.4, SAS Institute) [28].

Ethical Considerations

The claims-based cohort data in South Korea and Japan were anonymous, and this study's protocol was approved by the Institutional Review Board of National Health Insurance Service, Kyung Hee University (KHSIRB-23-085(EA)), and the JMDC (PHP-00002201-04).

Results

K-Fold Cross-Validation

For the 6 machine learning models, we found the following optimized hyper-parameters using grid search with 5-fold cross-validation: For Extreme Gradient Boosting, we used booster type of gradient boosted tree, column subsample by tree 0.1, learning rate 0.2, maximum depth of 3, and number of estimators 100. For random forest, we used maximum depth of 3, maximum features of 3, minimum samples per leaf 3, minimum samples per split 3, number of estimators 50, and balanced class weight. For Light GBM, we used boosting parameter of gradient-based 1-sided sampling, objective function of binary classification objective, evaluation metrics of log loss function for binary classification, learning rate 0.002, number of estimators 70, and number of leaves 30. For GBM, we used learning rate 0.008, maximum depth of 2, minimum samples per leaf 3, minimum samples per split 3, and number of estimators 100. For AdaBoost, we used algorithm of stagewise additive modeling using a multiclass exponential loss function, real variant; number of tree estimators with 500; and learning rate with 0.02. For LR, the solver of the library for large linear classification; the penalty norm was with L2, inverse of regularization strength 0.1, and the maximum number of iterations was with 100. For ensemble models, we used the same hyperparameters as those used in the individual machine learning models. Additionally, no weights were applied when combining the models in the ensemble. After finding the optimal hyperparameters, we checked the performance of each model and ensemble model. To improve performance, we tuned the models to use optimal thresholds through AUROC [20,29]. The optimized threshold values for some models are as follows: 0.48

XSL•FO
RenderX

for AdaBoost, 0.46 for GBM and AdaBoost, 0.46 for LR and GBM, and 0.46 for the GBM, AdaBoost, and LR. Table 3 summarizes the 5-fold cross-validation accuracy comparison of each model and ensemble machine learning models using sensitivity, specificity, accuracy, balanced accuracy, and AUROC as evaluation metrics. Among the single models, GBM, LR, and AdaBoost demonstrated the best prediction performance per balanced accuracy and AUROC. To further improve the classification performance, we explored an ensemble approach using the top-3 single models: GBM, LR, and AdaBoost. The results show that the combination of LR and AdaBoost provides the highest performance with a sensitivity of 80.62%, specificity of 81.79%, balanced accuracy of 81.2%, and AUROC of 0.9012. In addition, we also summarize 3 additional metrics suitable for imbalanced data in Table S2 in Multimedia Appendix 1: weighted average precision, weighted average F_1-score, and area under the precision-recall curve.

Table 3. Comparison of the prediction performances of the prediction models on the training dataset in the discovery cohort[a].

Model	Sensitivity, mean (SD)	Specificity, mean (SD)	Accuracy, mean (SD)	Balanced accuracy, mean (SD)	AUROC[b], mean (SD)	*P* values[c]
AdaBoost[d]	0.8503 (0.0074)	0.7725 (0.0048)	0.7764 (0.0044)	0.8114 (0.0023)	0.9136 (0.0035)	<.001
LR[e]	0.8009 (0.0090)	0.8076 (0.0015)	0.8072 (0.0012)	0.8042 (0.0041)	0.8819 (0.0046)	<.001
XGBoost[f]	0.6208 (0.011)	0.9599 (0.0029)	0.943 (0.0029)	0.7904 (0.006)	0.8866 (0.0052)	<.001
Random forest	0.7328 (0.012)	0.8642 (0.0098)	0.8577 (0.009)	0.7985 (0.0052)	0.8875 (0.0056)	<.001
Light GBM[g]	0.8295 (0.0052)	0.7649 (0.0033)	0.7681 (0.0032)	0.7972 (0.0038)	0.8743 (0.0058)	<.001
GBM	0.7853 (0.0065)	0.8194 (0.0027)	0.8176 (0.0029)	0.8023 (0.0046)	0.8942 (0.0051)	<.001
GBM and AdaBoost	0.8194 (0.0063)	0.8006 (0.0065)	0.8016 (0.0061)	0.81 (0.0035)	0.9063 (0.0044)	<.001
LR and GBM	0.8221 (0.0081)	0.7934 (0.0012)	0.7949 (0.0012)	0.8078 (0.0041)	0.9009 (0.0048)	<.001
GBM, AdaBoost, and LR	0.8373 (0.0076)	0.7795 (0.0014)	0.7824 (0.0011)	0.8084 (0.0034)	0.9065 (0.0047)	<.001
AdaBoost and LR	0.8062 (0.0072)[h]	0.8179 (0.0015) [h]	0.8173 (0.0012) [h]	0.8120 (0.0030) [h]	0.9012 (0.0046) [h]	Reference

[a]All outcomes are averaged over 5-fold cross-validation.

[b]AUROC: area under receiver operating characteristic.

[c]To compare the predictive performance of the models, we performed a Cochrane Q test on the model performance.

[d]AdaBoost: Adaptive Boosting.

[e]LR: logistic regression.

[f]XGBoost: Extreme Gradient Boosting.

[g]GBM: gradient boosting machine.

[h]Indicates machine learning model with best performance of prediction.

Feature Importance Analysis

The ranked normalized feature importance is from the ensemble model combining AdaBoost and LR. According to the results, age had the highest importance value among the features, followed by diastolic blood pressure, BMI, systolic blood pressure, and fasting blood glucose. Feature importances are as follows: age, 1.00; diastolic blood pressure, 0.93; BMI, 0.75; systolic blood pressure, 0.58; fasting blood glucose, 0.35; γ-GTP, 0.24; serum total cholesterol, 0.18; ALT, 0.10; AST, 0.097; history of diabetes mellitus, 0.087; household income, 0.77; hemoglobin, 0.025; sex, 0.021; history of stroke, 0.014; physical activity, 0.010; alcohol intake per week, 0.0077; region of residence, 0.0065; and smoking, 0.0055.

Ablation Study

Table S3 in Multimedia Appendix 1 summarizes the ablation study results when one or some top 5-contribution features were excluded: age, diastolic blood pressure, BMI, systolic blood pressure, and fasting blood glucose. Without age, the model provides poor prediction performance: balanced accuracy dropped from 0.812 to 0.782, and AUROC dropped from 0.901 to 0.864. Without diastolic blood pressure, balanced accuracy dropped to 0.784, and AUROC dropped to 0.871. Without BMI, balanced accuracy dropped to 0.811, and AUROC dropped to 0.898. Without systolic blood pressure, balanced accuracy dropped to 0.801, and AUROC dropped to 0.890. Without fasting blood glucose, balanced accuracy dropped to 0.8118, and AUROC dropped to 0.9010. Additionally, we analyzed the model performance when the 2 blood pressure features (systolic and diastolic) were excluded. The results show that the performance significantly degraded across all accuracy metrics: balanced accuracy from 0.812 to 0.725 and AUROC from 0.901 to 0.797.

Test Data Results and External Validation Results

Table 4 summarizes the test data results from the test dataset from the NHIS-NSC and the external validation data results from the JMDC. The test data results also showed that the ensemble model combining AdaBoost and LR provides the highest value of balanced accuracy (0.8147). The similarity between the cross-validation and test data results denotes minimal overfitting or underfitting. The external validation data

XSL•FO
RenderX

results also showed the ensemble model combining AdaBoost and LR provided the highest value of balanced accuracy (0.7406). The results confirmed that our ensemble model combining AdaBoost and LR could provide an accurate prediction of hypertension within 5 years based on the regular health checkup data.

Table 4. Comparison of the prediction performances of the prediction models on the test dataset (discovery cohort) and the external validation dataset (validation cohort).

Model	Sensitivity	Specificity	Accuracy	Balanced accuracy	AUROC[a]
Test dataset (discovery cohort)					
AdaBoost[b]	0.8573	0.7677	0.7722	0.8125	0.9123
LR[c]	0.8093	0.8064	0.8066	0.8078	0.8832
XGBoost[d]	0.6257	0.9629	0.9461	0.7943	0.8886
Random forest	0.8913	0.6417	0.6542	0.7665	0.8598
Light GBM[e]	0.8230	0.7707	0.7733	0.7968	0.8714
GBM	0.7839	0.8379	0.8352	0.8109	0.8966
GBM and AdaBoost	0.8403	0.7884	0.7910	0.8143	0.9080
LR and GBM	0.8297	0.7988	0.8004	0.8143	0.9039
GBM, AdaBoost, and LR	0.8443	0.7841	0.7871	0.8142	0.9087
AdaBoost and LR	0.8129[f]	0.8165[f]	0.8163[f]	0.8147[f]	0.9022[f]
External validation dataset (validation cohort)					
AdaBoost	0.6724	0.7906	0.7840	0.7315	0.8148
LR	0.6378	0.8352	0.8242	0.7365	0.8134
XGBoost	0.5253	0.8724	0.8530	0.6989	0.7906
Random forest	0.7109	0.6715	0.6737	0.6912	0.7324
Light GBM	0.5446	0.8241	0.8084	0.6843	0.7402
GBM	0.4995	0.8869	0.8652	0.6932	0.7875
GBM and AdaBoost	0.5906	0.8349	0.8212	0.7127	0.8052
LR and GBM	0.6428	0.8379	0.8270	0.7404	0.8241
GBM, AdaBoost, and LR	0.6240	0.8546	0.8417	0.7393	0.8271
AdaBoost and LR	0.6354[f]	0.8458[f]	0.8341[f]	0.7406[f]	0.8242[f]

[a]AUROC: area under receiver operating characteristic.

[b]AdaBoost: Adaptive Boosting.

[c]LR: logistic regression.

[d]XGBoost: Extreme Gradient Boosting.

[e]GBM: gradient boosting machine.

[f]Indicates machine learning model with best performance of prediction.

Association Between Risk Factors and Occurrence of Hypertension

The association between the occurrence of hypertension and potential risk factors is presented in Tables S4 and S5 in Multimedia Appendix 1. In both the discovery and validation cohorts, consistently, the multivariable model revealed that older age, female sex, urban residence, high income, high blood pressure, high serum total cholesterol, high hemoglobin, high AST, high γ-GTP, high BMI, history of diabetes mellitus, history of stroke, frequent alcohol intake, and insufficient physical activity were significantly associated with an increased risk of hypertension.

AI-Driven Web Application

Our proposed ensemble model was deployed on our own public website [30] so that hypertension onset within 5 years can be predicted based on regular health checkup data. The deployed web application, which provides results for prediction of hypertension onset, is shown in Figure S1 in Multimedia Appendix 2. The web interface for entering information on 18 features from regular health checkup data is shown in Figure S1(a) in Multimedia Appendix 2. After entering the information in the web application, a user can immediately obtain the results for prediction of hypertension onset with its probability, as shown in Figure S1(b) in Multimedia Appendix 2. In the web application, the features input by a user are encoded to the

XSL•FO
RenderX

website server, and immediately deleted upon generation of the prediction result, so that there is no risk of exposing information. In addition, there is no need to enter any information that would be regarded as private. Furthermore, we have open-sourced the Python code for the proposed ensemble model as publicly available in a GitHub repository [31].

Discussion

Main Findings

Given the significant health and economic consequences of CVDs, particularly myocardial infarction and stroke, it is essential to examine hypertension, a principal contributing factor to these conditions. This study uses data from 244,814 South Korean participants, obtained from the NHIS-NSC over a 12-year study period, and data from 1,296,649 Japanese participants, collected by the JMDC from various health insurance associations in Japan since 2005.

Our findings indicated that an ensemble of AdaBoost and LR models provided superior performance, achieving a sensitivity of 80.62%, specificity of 81.79%, balanced accuracy of 81.2%, and AUROC of 90.12%, suggesting that quantifying the occurrence of hypertension using feature importance analysis with ensemble machine learning (AdaBoost and LR) can enhance generalizability and reproducibility.

Using our knowledge of the machine learning model, our study has analyzed the occurrences of hypertension. Using feature importance analysis, our study has indicated the top 5-contribution features of hypertension, which were age, diastolic blood pressure, BMI, systolic blood pressure, and fasting blood glucose. Following the feature importance analysis, to measure the impact of such contributing features of hypertension, through an ablation study, we have excluded some contribution features among the top 5 contributing features of hypertension.

From our further investigation into the association between hypertension and independent variables, we analyzed various risk factors. Our analysis revealed that older age, female sex, urban residence, high income, elevated blood pressure, high serum total cholesterol, elevated hemoglobin, high AST, high γ-GTP, high BMI, history of diabetes mellitus, history of stroke, frequent alcohol intake, and insufficient physical activity were significantly associated with an increased risk of hypertension.

After obtaining test data results, through extra validation using the JMDC dataset, we have validated that our ensemble model combining AdaBoost and LR could provide an accurate prediction of hypertension within 5 years based on the regular health checkup data (balanced accuracy 0.741 and AUROC 0.824). Using such analysis of both the NHIS-NSC and JMDC as original and extra validation, our study has established a web application allowing diagnosis of hypertension [32-34].

Comparison With Previous Studies

Similar to our study, past research efforts have worked on developing hypertension risk prediction models using variables akin to our study, including age, sex, BMI, blood pressure metrics, parental hypertension history, smoking habits, and in certain cases, additional markers such as C-reactive protein, apolipoprotein A, and uric acid [35] (United States, n=1717 [36], n=1130 [37], n=15,732 [38], n=876 [39], and n=23,095 [40]; United Kingdom, n=10,308 [41]; and Iran, n=380 [42]).

Although there were several prior studies to find the occurrence of hypertension and establish web applications, many of these studies presented limitations, presenting problems such as producing low levels of reliability and yielding conflicting results. These constraints can be attributed to smaller sample sizes, short follow-up durations, and inadequate study designs such as nonrepresentative or nonrandom selection of populations [37,40,42]. Additionally, most of the studies have not held web application-fortifying processes, such as extra validation studies.

Our research stands apart in this context. We used a longitudinal approach using extensive datasets from both South Korean and Japanese health insurance databases, encompassing a comprehensive range of hypertension-related data spanning over 10 years. By leveraging datasets from 244,814 individuals in South Korea and 1,296,649 in Japan, we implemented a 5-fold cross-validation for optimizing an ensemble machine learning model. This was followed by a feature importance analysis to identify the top 5 determinants of hypertension, an ablation study to gauge the significance of each contributing factor, and an additional validation procedure. As a result, our work culminated in the development of a robust machine learning-powered web application, a milestone that many preceding studies fell short of achieving.

Possible Explanations for Our Results

This study harnesses real-world data where conventional statistical methods often struggle to guarantee generalizability and reproducibility in real-life situations. However, such challenges can be surmounted with AI-powered machine learning techniques such as variable pruning and group optimization.

By integrating machine learning methodologies, specifically AdaBoost and LR, our research can perpetually evaluate potential features linked to hypertension onset. This translates to a resilient system adept at observing the correlation between standardized traits and hypertension episodes, which include age, diastolic blood pressure, BMI, systolic blood pressure, and fasting blood glucose concentrations. Notably, this strategy offers considerable benefits, ensuring dependable data on hypertension prevalence across a wide demographic, even if the analysis encompasses merely a fraction of the overall populace. Additionally, our pioneering methodology offers a distinct advantage by ensuring accessibility even for individuals who may be illiterate or disinclined to participate in hypertension-specific screenings. This capacity for rapid diagnostic evaluation equips health care professionals with the tools to offer more targeted and accurate services to patients facing hypertension risks.

Policy Implication

Our diagnostic method's validation, achieved via an ensemble machine learning strategy integrating AdaBoost and LR, consistently upholds accuracy in hypertension identification, even among newly discerned populations potentially susceptible

XSL•FO
RenderX

to hypertension. This tool not only paves the way for preemptive hypertension identification but also extends its reach to individuals distant from conventional health care infrastructure, such as hospitals and regional health centers. Our study is keen on transitioning our web-based platform to a mobile app [18], addressing and eliminating any accessibility barriers. Such an evolution positions our tool as a universally accessible resource, irrespective of an individual's socioeconomic status, domicile, or the developmental index of their nation. Worldwide, national administrations can advocate for our tool, empowering citizens to independently gauge their hypertension risk and pursue timely medical interventions. The distinct advantage of our platform is its avoidance of potential diplomatic sensitivities, given its nonreliance on any personal or confidential data.

Strengths and Limitations

An astute examination of this study's findings calls for recognizing inherent limitations. To elaborate, even though our research draws on data from 2 distinct cohorts—the NHIS-NSC (n=244,814) for training or testing and the JMDC (n=1,296,649) for extra-validation—these datasets encompass but a marginal segment of the overarching Asian demographic, and an even lesser representation of the worldwide populace. This fact accentuates the imperative for our conclusions to be subjected to broader international validation studies and exhaustive longitudinal investigations. Furthermore, although the sample size of the JMDC is larger, its somewhat limited set of variables led us to develop a model using the comprehensive set of variables available in the NHIS-NSC [43]. Despite the smaller size of the NHIS-NSC, our proposed ensemble model showed stable and consistent performance when validated with the JMDC. Moreover, it is crucial to acknowledge that our study tested a limited array of model types, excluding machine learning models such as k-nearest neighbors and support vector machines. Including these models would have provided a comprehensive comparison and potentially strengthened our findings. Additionally, our analysis did not include certain hypertension-related features, such as family history, dietary habits, and salt consumption. As it is well known that these factors play a significant role in the development and progression of hypertension, their absence may have influenced the predictive power of the models and the holistic understanding of hypertension risk factors. Lastly, segmenting related variables such as systolic and diastolic pressure, ALT, and AST can capture diverse aspects and reduce data loss, but it may dilute significance due to their correlation [44].

While the primary aim of our study has been to identify predictors for the onset of hypertension, we acknowledge that predicting the magnitude of blood pressure increases offers an invaluable perspective on the complex interplay between initial blood pressure levels and their changes over time. This area, although not explored within the current scope of our research, holds significant potential for advancing our understanding of hypertension. Future investigations that include baseline blood pressure measurements could yield profound insights into the risk factors and dynamics of blood pressure changes. Such research would enrich our predictive models and refine management strategies for hypertension, marking a crucial step forward in the field.

Yet, amid these confines, one must not undermine this study's significance. Our endeavor capitalizes on data meticulously gathered for over a decade from South Korea and Japan. In a methodical exercise of comparing a spectrum of 6 machine learning models and subsequently analyzing ensemble variations, we astutely pinpointed the critical determinants closely aligned with hypertension onset, ensuring commendable reproducibility and applicability. Furthermore, the genesis of a user-responsive web tool, facilitating individuals to input personal health metrics, epitomizes our groundbreaking stride toward expeditious, precision-driven, and worldwide accessible diagnostic avenues for hypertension.

Conclusions

In a pioneering endeavor, this research uniquely integrates both machine learning and conventional statistical frameworks to prognosticate the emergence of hypertension. A notable outcome of this exploration is the institution of a digital platform adept at forecasting a 5-year onset of hypertension, using data sourced from the NHIS-NSC and JMDC. Our empirical outcomes, extrapolated from 2 autonomous studies, substantiate that machine learning paradigms—particularly an amalgamation of AdaBoost and LR—eclipsed the traditional statistical methodologies in preempting hypertension onset. A meticulous inquiry was undertaken to ascertain the hierarchical significance of determinants linked with hypertension. The investigation earmarked age as the paramount factor, trailed by diastolic blood pressure, BMI, systolic blood pressure, and fasting blood glucose concentrations. This sequence, pivotal in curating the most efficacious machine learning model and subsequent hypertension emergence, was further corroborated via supplementary validation harnessing the JMDC datasets. Emerging from these discernments is an AI-infused digital interface, proficient in envisioning a quintennial likelihood of hypertension based on routine health assessment metrics. Such an innovation positions itself as an instrumental diagnostic conduit for individuals predisposed to hypertension.

Acknowledgments

No generative AI was used in any portion of this study. This research was supported by grants from the National Research Foundation of Korea, funded by the Korean Government (Ministry of Science and ICT: MSIT, ICT: Information and Communication Technology; RS-2023-00248157) and the MSIT, Korea, under the Information Technology Research Center (ITRC) support program (IITP-2024-RS-2024-00438239) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

XSL•FO
RenderX

Data Availability

Data are available on reasonable request to the corresponding authors and with appropriate ethical approvals.

Authors' Contributions

JL and DKY had full access to all of the data in this study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final version before submission. ML, AK, and LS reviewed this paper. SHH, HL, JHL, DKY, and JL did the study concept and design; worked on the acquisition, analysis, or interpretation of data; drafted this paper; and handled the statistical analysis. All authors critically revised this paper for important intellectual content. JL supervised this study and is the guarantor for this study. DKY and JL contributed equally to this study as corresponding authors. SHH, HL, and JHL contributed equally to this work as first authors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analysis.
[DOCX File , 52 KB - jmir_v26i1e52794_app1.docx]

Multimedia Appendix 2
AI-driven web application. AI: artificial intelligence.
[DOCX File , 376 KB - jmir_v26i1e52794_app2.docx]

References

1. Ghani MAA, Ugusman A, Latip J, Zainalabidin S. Role of terpenophenolics in modulating inflammation and apoptosis in cardiovascular diseases: a review. Int J Mol Sci 2023;24(6):5339 [FREE Full text] [doi: 10.3390/ijms24065339] [Medline: 36982410]
2. Kim HC. Epidemiology of cardiovascular disease and its risk factors in Korea. Glob Health Med 2021;3(3):134-141 [FREE Full text] [doi: 10.35772/ghm.2021.01008] [Medline: 34250288]
3. Arbogast JW, Moore-Schiltz L, Jarvis WR, Harpster-Hagen A, Hughes J, Parker A. Impact of a comprehensive workplace hand hygiene program on employer health care insurance claims and costs, absenteeism, and employee perceptions and practices. J Occup Environ Med 2016;58(6):e231-e240 [FREE Full text] [doi: 10.1097/JOM.0000000000000738] [Medline: 27281645]
4. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet 2021;398(10304):957-980 [FREE Full text] [doi: 10.1016/S0140-6736(21)01330-1] [Medline: 34450083]
5. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014;129(Suppl 2):19-31 [FREE Full text] [doi: 10.1177/00333549141291S206] [Medline: 24385661]
6. Lee M, Lee H, Park J, Kim HJ, Kwon R, Lee SW, et al. Trends in hypertension prevalence, awareness, treatment, and control in South Korea, 1998-2021: a nationally representative serial study. Sci Rep 2023;13(1):21724 [FREE Full text] [doi: 10.1038/s41598-023-49055-8] [Medline: 38066091]
7. Jang W, Kim S, Son Y, Kim S, Kim HJ, Jo H, et al. Prevalence, awareness, treatment, and control of type 2 diabetes in South Korea (1998 to 2022): nationwide cross-sectional study. JMIR Public Health Surveill 2024;10:e59571 [FREE Full text] [doi: 10.2196/59571] [Medline: 39190907]
8. Chung H, Ko Y, Lee IS, Hur H, Huh J, Han SU, et al. Prognostic artificial intelligence model to predict 5 year survival at 1 year after gastric cancer surgery based on nutrition and body morphometry. J Cachexia Sarcopenia Muscle 2023;14(2):847-859 [FREE Full text] [doi: 10.1002/jcsm.13176] [Medline: 36775841]
9. Woo A, Lee SW, Koh HY, Kim MA, Han MY, Yon DK. Incidence of cancer after asthma development: 2 independent population-based cohort studies. J Allergy Clin Immunol 2021;147(1):135-143. [doi: 10.1016/j.jaci.2020.04.041] [Medline: 32417133]
10. Shin YH, Shin JI, Moon SY, Jin HY, Kim SY, Yang JM, et al. Autoimmune inflammatory rheumatic diseases and COVID-19 outcomes in South Korea: a nationwide cohort study. Lancet Rheumatol 2021;3(10):e698-e706 [FREE Full text] [doi: 10.1016/S2665-9913(21)00151-X] [Medline: 34179832]
11. Choi EK. Cardiovascular research using the Korean national health information database. Korean Circ J 2020;50(9):754-772 [FREE Full text] [doi: 10.4070/kcj.2020.0171] [Medline: 32725984]

XSL•FO
RenderX

12. Nagai K, Tanaka T, Kodaira N, Kimura S, Takahashi Y, Nakayama T. Data resource profile: JMDC claims database sourced from health insurance societies. J Gen Fam Med 2021;22(3):118-127 [FREE Full text] [doi: 10.1002/jgf2.422] [Medline: 33977008]
13. Kim S, Lee H, Lee J, Lee SW, Kwon R, Kim MS, et al. Short- and long-term neuropsychiatric outcomes in long COVID in South Korea and Japan. Nat Hum Behav 2024;8(8):1530-1544. [doi: 10.1038/s41562-024-01895-8] [Medline: 38918517]
14. Choi Y, Kim HJ, Park J, Lee M, Kim S, Koyanagi A, et al. Acute and post-acute respiratory complications of SARS-CoV-2 infection: population-based cohort study in South Korea and Japan. Nat Commun 2024;15(1):4499 [FREE Full text] [doi: 10.1038/s41467-024-48825-w] [Medline: 38802352]
15. Lee SW, Yang JM, Yoo IK, Moon SY, Ha EK, Yeniova, et al. Proton pump inhibitors and the risk of severe COVID-19: a post-hoc analysis from the Korean nationwide cohort. Gut 2021;70(10):2013-2015. [doi: 10.1136/gutjnl-2020-323672] [Medline: 33303566]
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
17. Kim J, Lee H, Lee J, Rhee SY, Shin JI, Lee SW, et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning: a post hoc analysis of a diagnostic trial and validation of an external additional trial. Alzheimers Res Ther 2023;15(1):127 [FREE Full text] [doi: 10.1186/s13195-023-01268-9] [Medline: 37481573]
18. Kwon R, Lee H, Kim MS, Lee J, Yon DK. Machine learning-based prediction of suicidality in adolescents during the COVID-19 pandemic (2020-2021): derivation and validation in two independent nationwide cohorts. Asian J Psychiatr 2023;88:103704. [doi: 10.1016/j.ajp.2023.103704] [Medline: 37541104]
19. Asghari Varzaneh Z, Shanbehzadeh M, Kazemi-Arpanahi H. Prediction of successful aging using ensemble machine learning algorithms. BMC Med Inform Decis Mak 2022;22(1):258 [FREE Full text] [doi: 10.1186/s12911-022-02001-6] [Medline: 36192713]
20. Liang Q, Qi Z, Li Y. Machine learning to predict the occurrence of thyroid nodules: towards a quantitative approach for judicious utilization of thyroid ultrasonography. Front Endocrinol (Lausanne) 2024;15:1385836 [FREE Full text] [doi: 10.3389/fendo.2024.1385836] [Medline: 38774231]
21. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024;14(1):6086 [FREE Full text] [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]
22. Lee H, Cho JK, Park J, Lee H, Fond G, Boyer L, et al. Machine learning-based prediction of suicidality in adolescents with allergic rhinitis: derivation and validation in 2 independent nationwide cohorts. J Med Internet Res 2024;26:e51473 [FREE Full text] [doi: 10.2196/51473] [Medline: 38354043]
23. Li X, Wang Y, Basu S, Kumbier K, Yu B. A debiased MDI feature importance measure for random forests. 2019 Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 08; United States p. 8049-8059.
24. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. 2013 Presented at: Proceedings of the 26th International Conference on Neural Information Processing Systems; 2013 Dec 05; United States p. 421-439.
25. Mathew TE. A logistic regression with recursive feature elimination model for breast cancer diagnosis. Int J Emerging Technol 2019;10(3):55-63.
26. Kawasoe S, Kubozono T, Salim AA, Yoshimine H, Mawatari S, Ojima S, et al. Development of a risk prediction score and equation for chronic kidney disease: a retrospective cohort study. Sci Rep 2023;13(1):5001 [FREE Full text] [doi: 10.1038/s41598-023-32279-z] [Medline: 36973534]
27. Woo HG, Park S, Yon H, Lee SW, Koyanagi A, Jacob L, et al. National trends in sadness, suicidality, and COVID-19 pandemic-related risk factors among South Korean adolescents from 2005 to 2021. JAMA Netw Open 2023;6(5):e2314838 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.14838] [Medline: 37223902]
28. Lee SW. Regression analysis for continuous independent variables in medical research: statistical standard and guideline of life cycle committee. Life Cycle 2022;2:e3. [doi: 10.54724/lc.2022.e3]
29. Bai Q, Su C, Tang W, Li Y. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep 2022;12(1):8377 [FREE Full text] [doi: 10.1038/s41598-022-12316-z] [Medline: 35589908]
30. Predicting hypertension with health checkup. URL: http://ai-wm.khu.ac.kr/Hypertension/ [accessed 2024-10-16]
31. seunghahh / PredictingHypertension. GitHub. URL: https://github.com/seunghahh/PredictingHypertension [accessed 2024-10-05]
32. Sáez C, Ferri P, García-Gómez JM. Resilient artificial intelligence in health: synthesis and research agenda toward next-generation trustworthy clinical decision support. J Med Internet Res 2024;26:e50295 [FREE Full text] [doi: 10.2196/50295] [Medline: 38941134]
33. Cai YQ, Gong DX, Tang LY, Cai Y, Li HJ, Jing TC, et al. Pitfalls in developing machine learning models for predicting cardiovascular diseases: challenge and solutions. J Med Internet Res 2024;26:e47645 [FREE Full text] [doi: 10.2196/47645] [Medline: 38869157]

XSL•FO
RenderX

34. Kim H, Son Y, Lee H, Kang J, Hammoodi A, Choi Y, et al. Machine learning-based prediction of suicidal thinking in adolescents by derivation and validation in 3 independent worldwide cohorts: algorithm development and validation study. J Med Internet Res 2024;26:e55913 [FREE Full text] [doi: 10.2196/55913] [Medline: 38758578]
35. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One 2013;8(7):e67370 [FREE Full text] [doi: 10.1371/journal.pone.0067370] [Medline: 23861760]
36. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham heart study. Ann Intern Med 2008;148(2):102-110. [doi: 10.7326/0003-4819-148-2-200801150-00005] [Medline: 18195335]
37. Pearson TA, LaCroix AZ, Mead LA, Liang KY. The prediction of midlife coronary heart disease and hypertension in young adults: the Johns Hopkins multiple risk equations. Am J Prev Med 1990;6(2 Suppl):23-28. [Medline: 2383409]
38. Kshirsagar AV, Chiu Y, Bomback AS, August PA, Viera AJ, Colindres RE, et al. A hypertension risk score for middle-aged and older adults. J Clin Hypertens (Greenwich) 2010;12(10):800-808 [FREE Full text] [doi: 10.1111/j.1751-7176.2010.00343.x] [Medline: 21029343]
39. Fitriyani NL, Syafrudin M, Alfian G, Rhee J. Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. IEEE Access 2019;7:144777-144789. [doi: 10.1109/access.2019.2945129]
40. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
41. Kivimäki M, Batty GD, Singh-Manoux A, Ferrie JE, Tabak AG, Jokela M, et al. Validating the Framingham hypertension risk score: results from the Whitehall II study. Hypertension 2009;54(3):496-501 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.109.132373] [Medline: 19597041]
42. Gorbani F, Mahmoodi H, Sarbakhsh P, Shaghaghi A. Predictive performance of Pender's health promotion model for hypertension control in Iranian patients. Vasc Health Risk Manag 2020;16:299-305 [FREE Full text] [doi: 10.2147/VHRM.S258458] [Medline: 32764950]
43. Watada H, Kawakita Y, Tanaka K, Oh A, Nishigaki N, Nakajima T, et al. A claims-based cohort study on the treatment patterns of Japanese patients with type 2 diabetes mellitus and the association of early first physician visit with time to prescription of oral hypoglycemic agents. Diabetes Ther 2021;12(7):2035-2047 [FREE Full text] [doi: 10.1007/s13300-021-01090-2] [Medline: 34151415]
44. Yoo HW, Jin HY, Yon DK, Effenberger M, Shin YH, Kim SY, et al. Non-alcoholic fatty liver disease and COVID-19 susceptibility and outcomes: a Korean nationwide cohort. J Korean Med Sci 2021;36(41):e291 [FREE Full text] [doi: 10.3346/jkms.2021.36.e291] [Medline: 34697932]

Abbreviations

AI: artificial intelligence
ALT: alanine transaminase
AST: aspartate transaminase
AUROC: area under the receiver operating characteristic
CDC: Centers for Disease Control and Prevention
CVD: cardiovascular disease
GBM: gradient boosting machine
ICD-10: International Classification of Disease, 10th revision
JMDC: Japanese Medical Data Center cohort
LR: logistic regression
MDI: mean decrease in impurity
NHIS-NSC: National Health Insurance Service-National Sample Cohort
WHO: World Health Organization
γ-GTP: γ-glutamyl transpeptidase

Edited by Y Li; submitted 15.09.23; peer-reviewed by A Higaki, C Jeong; comments to author 09.02.24; revised version received 02.04.24; accepted 17.09.24; published 05.11.24.

Please cite as:
Hwang SH, Lee H, Lee JH, Lee M, Koyanagi A, Smith L, Rhee SY, Yon DK, Lee J
Machine Learning–Based Prediction for Incident Hypertension Based on Regular Health Checkup Data: Derivation and Validation in 2 Independent Nationwide Cohorts in South Korea and Japan
J Med Internet Res 2024;26:e52794
URL: https://www.jmir.org/2024/1/e52794
doi:10.2196/52794
PMID:

XSL•FO
RenderX

©Seung Ha Hwang, Hayeon Lee, Jun Hyuk Lee, Myeongcheol Lee, Ai Koyanagi, Lee Smith, Sang Youl Rhee, Dong Keon Yon, Jinseok Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of Deep Learning–Based Infectivity Prediction in Pulmonary Tuberculosis Through Chest Radiography: Retrospective Study

Wou young Chung[1*], MD, MS; Jinsik Yoon[2*], BS; Dukyong Yoon[3,4,5*], MD, PhD; Songsoo Kim[3], MD, MS; Yujeong Kim[3], PhD; Ji Eun Park[1], MD, PhD; Young Ae Kang[6], MD, PhD

[1]Department of Pulmonary and Critical Care Medicine, Ajou University School of Medicine, Suwon, Republic of Korea
[2]Department of Integrative Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[3]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea
[4]Institute for Innovation in Digital Healthcare, Severance Hospital, Seoul, Republic of Korea
[5]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea
[6]Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Young Ae Kang, MD, PhD
Department of Internal Medicine
Yonsei University College of Medicine
50-1 Yonsei-ro, Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 1954
Email: mdkang@yuhs.ac

Abstract

Background: Pulmonary tuberculosis (PTB) poses a global health challenge owing to the time-intensive nature of traditional diagnostic tests such as smear and culture tests, which can require hours to weeks to yield results.

Objective: This study aimed to use artificial intelligence (AI)–based chest radiography (CXR) to evaluate the infectivity of patients with PTB more quickly and accurately compared with traditional methods such as smear and culture tests.

Methods: We used DenseNet121 and visualization techniques such as gradient-weighted class activation mapping and local interpretable model-agnostic explanations to demonstrate the decision-making process of the model. We analyzed 36,142 CXR images of 4492 patients with PTB obtained from Severance Hospital, focusing specifically on the lung region through segmentation and cropping with TransUNet. We used data from 2004 to 2020 to train the model, data from 2021 for testing, and data from 2022 to 2023 for internal validation. In addition, we used 1978 CXR images of 299 patients with PTB obtained from Yongin Severance Hospital for external validation.

Results: In the internal validation, the model achieved an accuracy of 73.27%, an area under the receiver operating characteristic curve of 0.79, and an area under the precision-recall curve of 0.77. In the external validation, it exhibited an accuracy of 70.29%, an area under the receiver operating characteristic curve of 0.77, and an area under the precision-recall curve of 0.8. In addition, gradient-weighted class activation mapping and local interpretable model-agnostic explanations provided insights into the decision-making process of the AI model.

Conclusions: This proposed AI tool offers a rapid and accurate alternative for evaluating PTB infectivity through CXR, with significant implications for enhancing screening efficiency by evaluating infectivity before sputum test results in clinical settings, compared with traditional smear and culture tests.

(J Med Internet Res 2024;26:e58413) doi:10.2196/58413

XSL•FO
RenderX

KEYWORDS

pulmonary tuberculosis; chest radiography; artificial intelligence; tuberculosis; TB; smear; smear test; culture test; diagnosis; treatment; deep learning; CXR; PTB; management; cost effective; asymptomatic infection; diagnostic tools; infectivity; AI tool; cohort

Introduction

Pulmonary tuberculosis (PTB) is a global health concern that continues to challenge public health systems. In 2022, it affected approximately 10.6 million people, with an incidence rate of 133 per 100,000 people. Notably, it has claimed the lives of approximately 1.13 million patients who are HIV negative and 167,000 patients who are HIV positive, underscoring its high mortality rate [1].

Many people are infected with the tuberculosis bacteria, but not all develop active tuberculosis. Identifying this asymptomatic infection is a major challenge in public health [2]. Traditional tuberculosis diagnostic methods include smear and culture tests, which may be time-consuming or exhibit low sensitivities, resulting in missed early diagnosis. As tuberculosis bacteria are airborne, quickly identifying and isolating patients with active tuberculosis is crucial for preventing further spread. In addition, rapid diagnosis allows for the timely and appropriate treatment of patients, increasing the success rate and minimizing complications [3].

The sensitivity of traditionally used smear tests is highly variable, ranging from 50% to 60% [4]. Thus, these tests may not detect PTB, particularly when there are low bacterial loads, and are unreliable for early diagnosis. This problem can increase the risk of missing PTB in its early stages and result in patients with asymptomatic or mild symptoms not receiving appropriate diagnosis and treatment [5].

Although culture tests exhibit a high sensitivity of approximately 80% [6], they require several weeks and sophisticated laboratory facilities, which may not be readily available, especially in resource-limited settings. However, these delays can hinder the proper initiation of treatment for people with active PTB, which can eventually lead to poorer health and increased mortality. When a rapid and accurate diagnosis is not made, PTB can easily spread to others, imposing a significant burden on public health and complicating control and management efforts [7]. Furthermore, traditional sputum testing has exhibited a sensitivity of 52% for patients undergoing treatment [8].

Computed tomography (CT) is an important diagnostic tool that provides detailed information of lung structures for lung disease evaluations, particularly PTB [9]. However, despite its diagnostic use, using CT for PTB diagnosis has several critical issues, including limited availability, high costs, and radiation exposure concerns [10]. For example, a lack of imaging facilities necessary for CT scans may exist in resource-limited countries. This limitation restricts access to crucial diagnostic imaging, delaying the diagnosis and treatment of patients in these regions. Consequently, health crises may be exacerbated by allowing diseases to spread more widely without timely intervention. Recent advancements in medical imaging, specifically chest radiography (CXR), have contributed significantly to the early detection and management of PTB. The use of CXR as a primary PTB screening tool has been reaffirmed by its high sensitivity despite its specificity limitations. Various studies have demonstrated that CXR is instrumental for identifying PTB, especially in vulnerable populations that exhibit a notably higher disease prevalence than the general population [11,12].

The use of CXR for assessing PTB infectivity offers several key advantages over CT as x-ray equipment is more widely available and simple to operate. In addition, it enables swift diagnosis and timely treatments, which is particularly advantageous for regions with limited access to advanced medical equipment. Moreover, CXR involves significantly less radiation exposure than CT, which is crucial for sensitive groups such as pregnant women and children, who are more susceptible to the adverse effects of radiation. CXR is also highly suitable for large-scale screening. Thus, its rapid processing and cost efficiency make it ideal for mass PTB screening initiatives that are vital for curbing the spread of tuberculosis.

The advent of deep learning technologies has opened new avenues for enhancing the diagnosis and management of patients with PTB [13]. Our study leverages these advancements to differentiate infectious and noninfectious PTB cases using CXR, which is critical for determining appropriate quarantine measures and controlling the spread of tuberculosis [14]. Conventional PTB detection approaches are primarily used to identify the presence or absence of the disease. However, the infectious status of PTB plays a crucial role in public health, particularly for making quarantine decisions and preventing its spread [15].

This study was motivated by the need for diagnostic tools that can provide insights into the infectivity of PTB cases beyond mere detection. The aim of our artificial intelligence (AI) tool is to bridge the gap in PTB management in resource-limited countries by using CXR to evaluate the infectivity of patients with PTB more quickly and accurately than when using smear and culture tests. The proposed tool can significantly impact public health strategies and offer a more targeted approach for quarantine decisions.

Methods

Patients and Infectivity Labeling

The cohort comprised patients who met the following criteria: (1) first PTB diagnosis between January 2004 and June 2023, (2) underwent sputum examinations, (3) CXR performed within –60 to +180 days from the initial diagnosis date, (4) at least 1 positive sputum sample (smear or culture), and (5) received a successful 6-month treatment course (no treatment after the seventh month).

For positive labels, infectious CXR images were defined as those obtained from the initial diagnosis date to the last positive sputum test result. For negative labels, we used CXR images obtained after the conversion date, which was the day when the sputum test result changed from positive to negative.

XSL•FO
RenderX

Specifically, a 1-month gray zone was established following the next sputum test date after the conversion date, that is, the CXR image obtained after the second negative sputum test result post conversion was classified as noninfectious.

We used data from Severance Hospital for the training dataset. Internal validation was performed using temporal validation, and external validation was conducted using data from Yongin Severance Hospital, which has different geographical and scale characteristics compared with Severance Hospital. The data of patients with PTB who visited Severance Hospital, which is a 2437-bed tertiary teaching hospital in Seoul, between 2004 and 2020 (31,260 CXR images) were used for model training. Subsequently, data from patients who visited in 2021 (1855 CXR images) were used for testing, while data from those who visited between 2022 and 2023 (3027 CXR images) were used for internal validation. The total data included 4492 patients and 36,142 CXR images (18,557 negative and 17,585 positive). In addition, data from Yongin Severance Hospital, which is a separate 708-bed general hospital located in Yongin, were used for external validation. These included 299 patients and 1978 CXR images, with 863 negative and 1115 positive cases. Figure 1 illustrates the selection process, highlighting the distinct contributions of each institution.

We also used a lung segmentation dataset comprising 704 CXR images, sourced from the National Library of Medicine, National Institutes of Health, United States [16], and the Third People's Hospital of Shenzhen, China [17]. We divided the dataset into 3 subsets, allocating 60% for training, 20% for testing, and 20% for validation.

Figure 1. Systematic flowchart of the patient selection criteria and process used in this study. CXR: chest radiography.

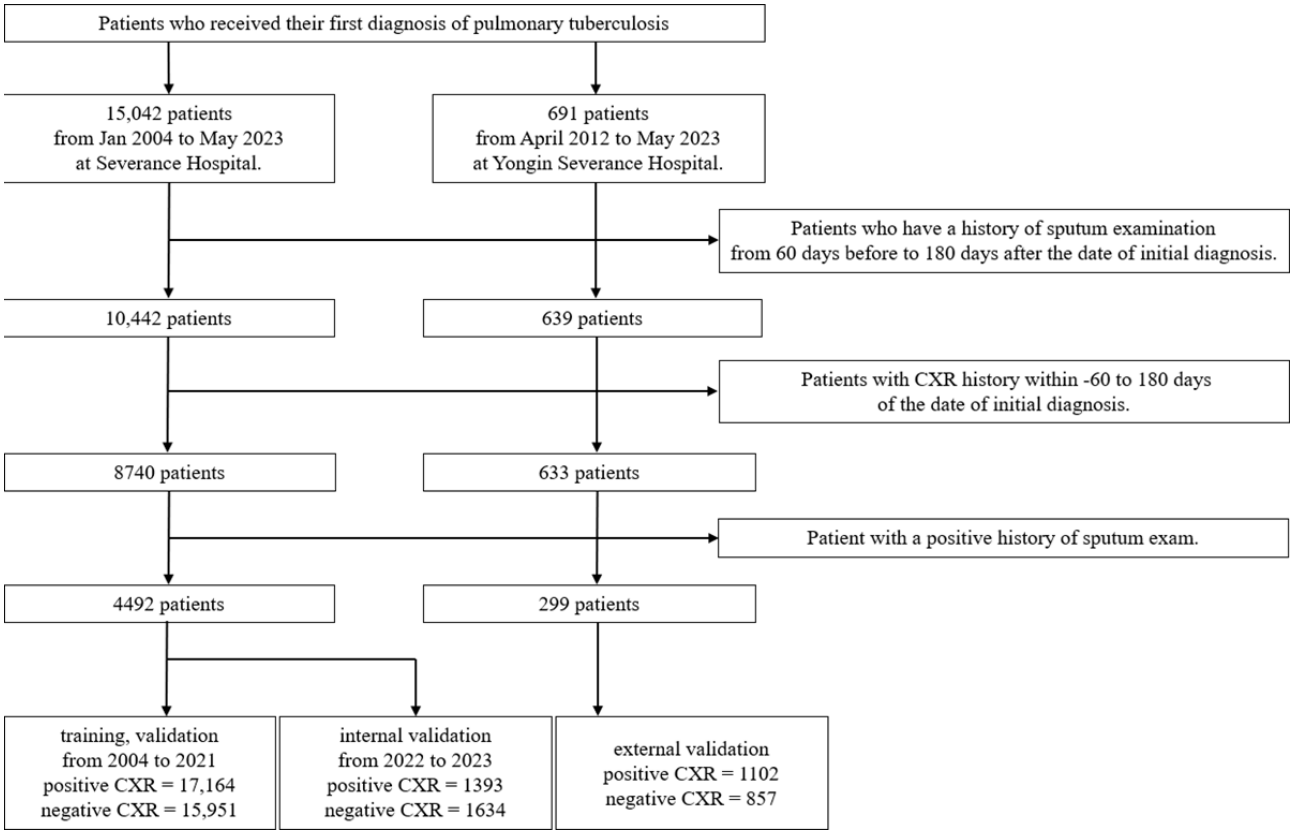

Segmentation Model and Cropping Lung Region

We developed a lung segmentation model based on TransUNet [18] to focus the analysis on lung parenchyma in the CXR images. Figure 2 shows an example CXR image, where only the lung region is segmented. Following segmentation, only the cropped lung regions were used in the analysis. Further details regarding the segmentation model are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Lung area segmentation and cropping in chest radiographs to enhance tuberculosis infectivity assessment. Left: original chest radiographs; center: segmented and cropped lung area; right: chest radiographs used for analysis.

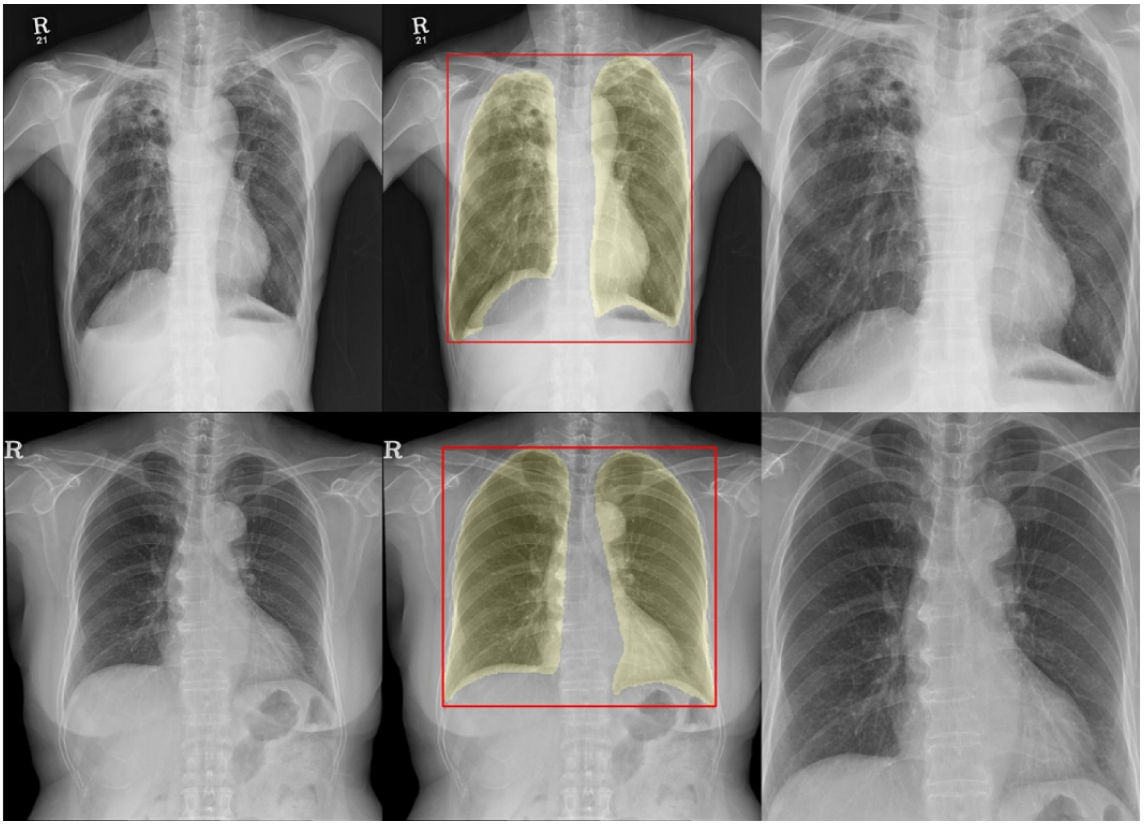

Convolutional Neural Network and Its Decision-Making Process

We selected the DenseNet121 [19] architecture, which was pretrained using the CheXNet [20] dataset, for developing an infectivity evaluation model. Detailed information regarding the infectivity evaluation model is provided in Multimedia Appendix 2.

In addition, we incorporated gradient-weighted class activation mapping [21]. It visually illustrates the CXR sections on which the neural network focuses during the prediction process through a heat map, highlighting the most influential areas for the predictions. Furthermore, we used local interpretable model-agnostic explanations [22] to segment the CXR images and elucidate the regions that contributed most significantly to its predictions, thereby providing additional insights into its decision-making process.

Statistical Analysis

This study presents the quantitative data as mean values with SD and categorical data as frequencies and percentages.

In the baseline characteristics, for the analysis of categorical variables, the Fisher exact test was used for “Leukemia” only, whereas the remaining variables were statistically tested using the chi-square test. In addition, “Age” was tested through the Mann-Whitney *U* test. In all statistical tests, a *P* value of <.05 was considered statistically significant. The accuracy, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were used to evaluate the performance of the AI model. We used 1000 bootstrap iterations to compute the 95% CIs for the above metrics.

All analyses were performed using Python (version 3.10.9; Python Software Foundation), TensorFlow-GPU (version 2.10.0; Google), and R (version 4.3.1; R Foundation for Statistical Computing).

Ethical Consideration

This study adhered to the Declaration of Helsinki guidelines and was approved by the institutional review board of Severance Hospital (4-2023-0679). Owing to its retrospective design, the requirement for individual patient consent was waived by the institutional review board. In addition, we adhered to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View [23] throughout our study.

Results

Baseline Characteristics

The baseline characteristics of the patients from both hospitals are summarized in Table 1. The proportion of male patients was slightly higher at Severance Hospital (2718/4492, 60.51%) than at Yongin Severance Hospital (178/299, 59.53%), although the difference was not statistically significant (*P*=.91). In addition, the average ages of patients at Severance Hospital and Yongin Severance Hospital were 65 (SD 18.95) years and 66.14 (SD 18.84) years, respectively. This difference in age distribution was not statistically significant (*P*=.22). However, there were

XSL•FO
RenderX

significant differences in the prevalence of certain conditions. Chronic obstructive pulmonary disease was more prevalent in the Yongin Severance Hospital group (89/299, 29.77%) than in the Severance Hospital group (914/4492, 20.34%), with a *P* value of <.001.

Similarly, the prevalence of lung cancer and all cancer types was significantly higher in the Severance Hospital group (lung cancer: 255/4492, 5.68%; all cancers: 1200/4492, 26.71%) than in the Yongin Severance Hospital group (lung cancer: 4/299, 1.34%; all cancers: 28/299, 9.36%), both with *P* values of <.001. Other conditions such as drug-induced interstitial lung disease, connective tissue diseases, end-stage renal disease, and diabetes showed no statistically significant differences between the 2 groups (all *P*>.05). In addition, the prevalence of leukemia and pulmonary edema was similar between the 2 groups (leukemia: Severance Hospital 73/4492, 1.63%; Yongin Severance Hospital 6/299, 2%; pulmonary edema: Severance Hospital 78/4492, 1.74%; Yongin Severance Hospital 4/299, 1.34%), with *P* values of .78. Overall, these findings suggest that although there were some statistically significant differences in the prevalence of certain conditions between the patients from the 2 hospitals, other baseline characteristics such as sex, age, and the prevalence of several other conditions were similar.

Table 1. Baseline characteristics.

Characteristics	Severance Hospital (n=4492)	Yongin Severance Hospital (n=299)	*P* value
Male, n (%)	2718 (60.51)	178 (59.53)	.91
Age (years), mean (SD)	65 (18.95)	66.14 (18.84)	.22
Chronic obstructive pulmonary disease, n (%)	914 (20.35)	89 (29.77)	<.001
Drug-induced interstitial lung disease, n (%)	109 (2.43)	6 (2.01)	<.001
Lung cancer, n (%)	255 (5.68)	4 (1.34)	<.001
Connective tissue disease, n (%)	170 (3.78)	5 (1.67)	.09
End-stage renal disease, n (%)	560 (12.47)	38 (12.71)	.99
Diabetes, n (%)	1057 (23.53)	88 (29.43)	.09
Leukemia, n (%)	73 (1.63)	6 (2.01)	.78
All cancers, n (%)	1200 (26.71)	28 (9.36)	<.001
Pulmonary edema, n (%)	78 (1.74)	4 (1.34)	.78

Infectivity Evaluation

The performance of the infectivity evaluation model was assessed using both internal and external validation sets. The results are summarized in Figure 3 and Table 2. In the internal validation set, the model exhibited an accuracy of 73.27% and a sensitivity of 67.55%, indicating its ability to identify true positive cases correctly, and a specificity of 78.15%, reflecting its effectiveness in identifying true negative cases. Its PPV and NPV were 72.50% and 73.86%, respectively. Its AUROC, which indicates the overall diagnostic ability, was 0.79, whereas its AUPRC was 0.77 (Figure 4A). In the external validation set, the model exhibited an accuracy of 70.29%, a sensitivity of 72.87%, a specificity of 66.98%, and PPV and NPV of 73.94% and 65.75%, respectively. In addition, it achieved AUROC and AUPRC values of 0.77 and 0.8 (Figure 4B), respectively, demonstrating robust performance in differentiating between positive and negative cases on an external validation. We performed calibration plotting to assess how well the predicted probabilities matched the actual outcomes and further evaluate the model. The Brier score, which measures the accuracy of probabilistic predictions, was 0.18 for internal validation and 0.2 for external validation, indicating high levels of calibration and reliability of the probability estimates of the model (Figure 4C). We performed a decision curve analysis to demonstrate clinical use. At nearly all thresholds, the net benefit of using the AI model was higher than that of other strategies. This indicates that the AI model operates effectively across various thresholds, with the highest net benefit observed in the midrange probability thresholds (Figure 4D). The results of comparing the model to other architectures are presented in Multimedia Appendices 3 and 4.

XSL•FO
RenderX

Figure 3. Confusion matrices for (A) internal and (B) external validation of the infectivity evaluation model at a threshold of 0.5.

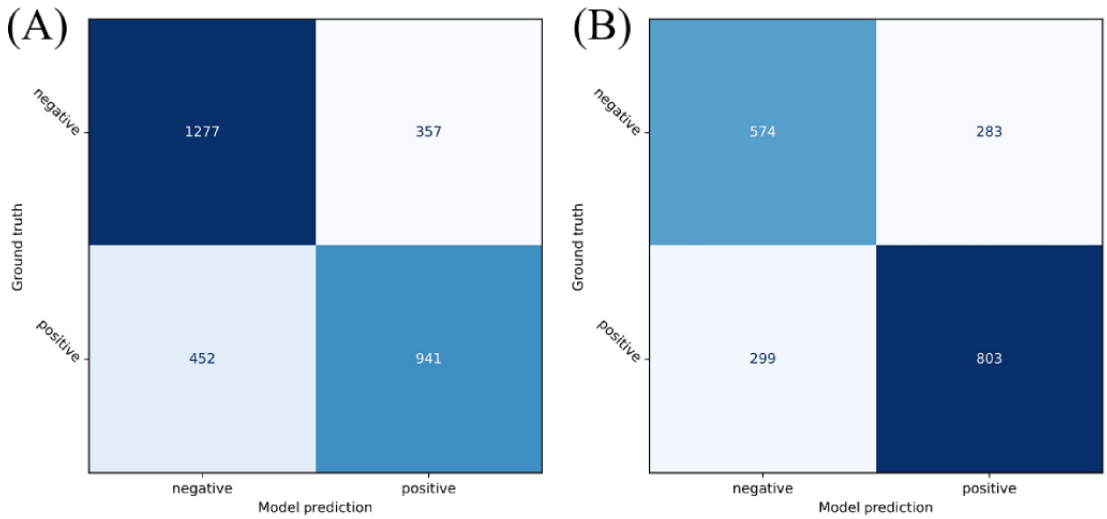

Table 2. Infectivity evaluation performance of the artificial intelligence model.

	Accuracy	AUROC[a]	AUPRC[b]	Sensitivity	Specificity	PPV[c]	NPV[d]
Severance Hospital (internal validation), mean (95% CI)	0.73 (0.72-0.75)	0.79 (0.78-0.81)	0.77 (0.75-0.80)	0.68 (0.65-0.70)	0.78 (0.76-0.80)	0.73 (0.70-0.75)	0.74 (0.72-0.76)
Yongin Severance Hospital (external validation), mean (95% CI)	0.7 (0.68-0.73)	0.77 (0.75-0.79)	0.8 (0.77-0.82)	0.73 (0.70-0.76)	0.67 (0.64-0.70)	0.74 (0.72-0.77)	0.66 (0.63-0.69)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

XSL•FO
RenderX

Figure 4. Performance evaluation of the model: internal and external validation results. (A) ROC curve of the model, illustrating the trade-off between sensitivity and specificity, with AUROC values for internal and external validation; (B) PR curve of the model, highlighting precision-recall trade-offs with AUPRC values; (C) calibration plot with Brier scores, showing the agreement between predicted probabilities and observed outcomes; and (D) decision curve analysis, evaluating the net benefit across different threshold probabilities. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; PR: precision-Recall; ROC: receiver operating characteristics.

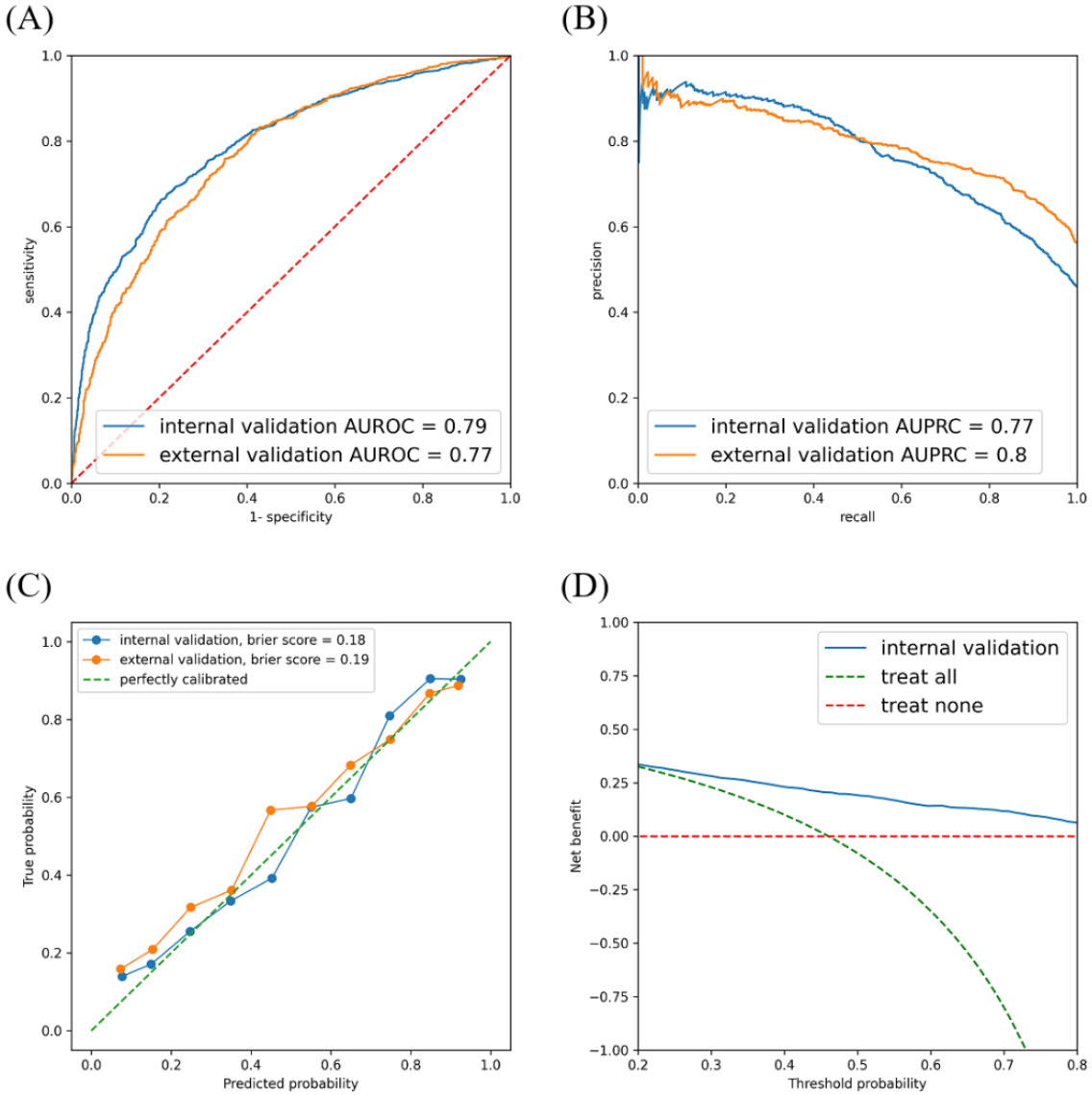

Visualizing the Decision-Making Process of the AI Model

In the visualization of the decision-making process of the AI model, it was observed that the model focuses particularly on the left and right lung apex regions. This is consistent with previous findings that the high oxygen concentration and poor lymphatic drainage in the lung apex provide a suitable environment for PTB growth [24]. Thus, it helped in understanding the features that were most indicative of infectivity within the lung regions (Figure 5). These visualization techniques made the model predictions transparent, ensuring that the decision-making process aligned with clinical expectations and knowledge.

XSL•FO
RenderX

Figure 5. Left: chest radiographs used for the analysis; middle: results of GradCAM; and right: results of LIME. Both the upper and lower panels display chest radiographs that were accurately identified as positive by the model. The analysis in the upper figure focuses on the left lung apex, while the analysis in the lower figure concentrates on the right lung apex, visualizing the areas of frequent tuberculosis occurrences. GradCAM: gradient-weighted class activation mapping; LIME: local interpretable model-agnostic explanations.

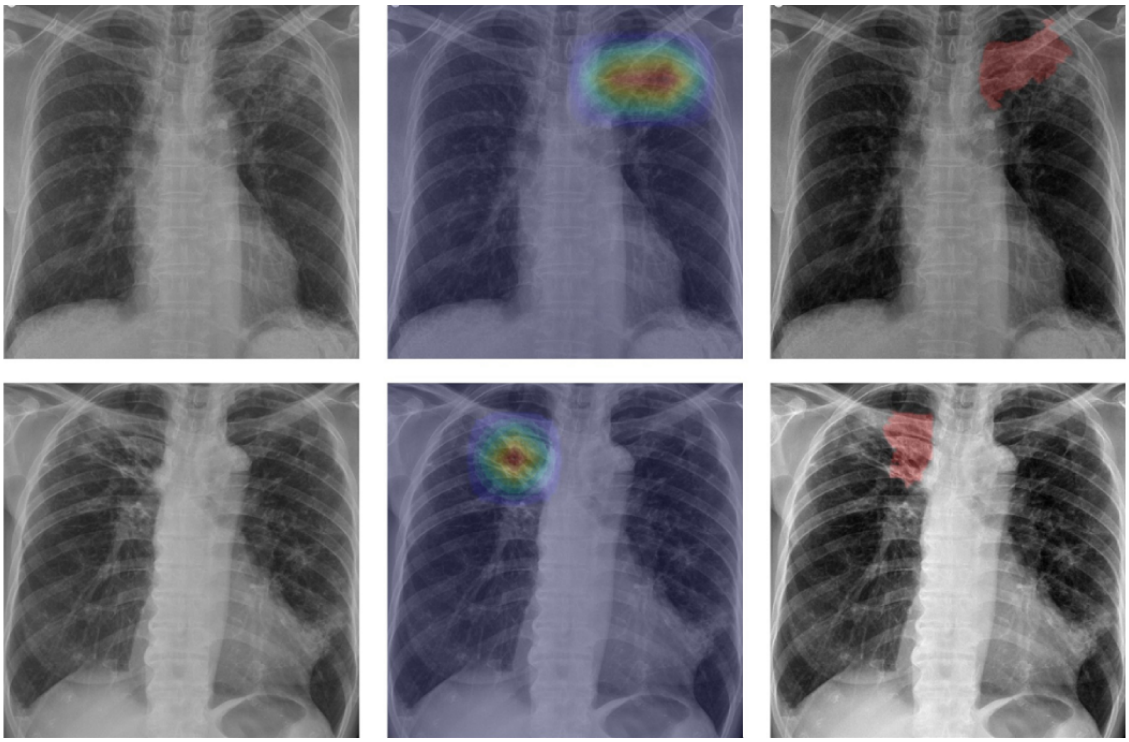

AI Performance in Evaluating Pulmonary Tuberculosis Infectivity Across Subgroups and Hospitals

Subgroup analyses were conducted among disease groups that showed differences in prevalence in the baseline characteristics, such as chronic obstructive pulmonary disease (COPD) and all cancers between the 2 hospitals, to ascertain whether the model was affected by a particular disease group. The drug-induced interstitial lung disease and lung cancer groups could not be analyzed owing to the presence of only negative cases and the small number of patients in the external validation set. Detailed results for the COPD and all cancers groups between hospitals are presented in Multimedia Appendix 5.

In the analysis between the COPD and non-COPD groups, the AI model exhibited lower sensitivity for both hospitals in the COPD group compared with the non-COPD group. In the comparison between the all cancers and noncancer groups, the all cancers group exhibited higher scores than the noncancer group in several metrics, except for specificity and NPV, at both hospitals.

Discussion

Principal Findings

The proposed deep learning model that uses CXR images to evaluate PTB infectivity can facilitate clinical decisions regarding quarantine or discharge as well as assist in evaluating the success of antituberculosis treatments. Furthermore, the ability to differentiate active lesions from healed, old lesions or those being healed can help to avoid unnecessary treatments or evaluations of inactive lesions, which is often an issue in countries with high PTB incidence rates [25,26].

CXR is an important tool for PTB screening, and the World Health Organization recommends its use for this purpose. However, many countries with a high PTB burden have a shortage of trained CXR specialists [27-29]. Therefore, AI-based screening of active PTB could provide these regions with a useful tool to protect public health and enhance PTB control. In addition, according to Kazemzadeh et al [27], deep learning–based, automatic, active PTB screening tools have the potential to offer significant cost savings of 40% to 80% in settings with a prevalence of 1% to 10%.

Furthermore, in countries with moderate to high PTB prevalence, it is common to encounter patients who have been treated or who have spontaneously healed from PTB [28]. Our study focused on distinguishing active, infective PTB, thereby equipping such countries with a more usefully adjusted tool for better PTB control.

Patients who are smear or culture positive for PTB are isolated until their sputum microbiological tests return negative, which can take several weeks for confirmation [30]. Our infectivity-targeted model offers the potential for immediate assessment, which could facilitate quicker decisions regarding discharge from the hospital or return to the workplace, potentially benefiting the socioeconomic situation of the region.

The proposed model exhibited consistent and reliable performance in both internal and external validations, highlighting its clinical applicability for determining PTB infectivity. By achieving a crucial balance between sensitivity and specificity, it effectively differentiated infective and

XSL•FO
RenderX

noninfective cases, which holds significant clinical importance. Moreover, the integration of visualization techniques improved its transparency by elucidating the specific areas that influenced its decisions. Thus, the proposed model can aid in PTB management and assessment by leveraging broadly accessible CXR images.

Traditionally used smear tests have low sensitivity, which can lead to missed diagnoses, especially in cases of early infections with low bacterial loads. Although culture tests are considered standard, they can be slow, requiring several weeks to produce results. Thus, diagnostic decision-making must be improved through more efficient methods. Table 3 lists the benefits and limitations of the smear, culture, and CXR tests, providing insights into their performance, speed, and overall efficacy for diagnosing PTB. Evidently, the proposed CXR test can produce diagnosis results within several seconds, enabling rapid decision-making for quarantine. Moreover, it is noninvasive, cost-effective, and widely accessible and features minimal patient discomfort. In addition, advances in AI have enabled automated analyses, which reduce clinician workloads and improve diagnostic accuracy.

Table 3. Comparison of smear, culture, and chest radiography tests.

	Culture [31]	Smear [32]	Chest radiography (ours)
Sensitivity	80.8%	46.2%	67.6%
Specificity	99.0%	99.3%	78.2%
Result time	Within weeks	Within hours	Within seconds
Advantages	High sensitivity	High specificity	Automated, rapid test results
Limitations	Long test results, contamination	Low sensitivity, low bacterial load	Low specificity

CT has also been used to differentiate PTB cases. Gao et al [33] developed and evaluated TBINet, a deep learning model that uses CT images to determine the infectivity of patients with PTB. It achieved AUROC scores of 0.82 and 0.75 for the validation and external validation sets, respectively, thereby outperforming traditional deep learning methods. However, the limitations of CT include high costs, limited facilities, and radiation exposure [34]. Our approach overcomes these limitations by using the more accessible and safer CXR, which is performed similarly to CT-based models. Choi et al [35] and Lee et al [28] have demonstrated the potential of CXR to effectively distinguish between patients with active PTB and healthy patients. Their findings, highlighted by AUROC scores ranging from 0.83 to 0.98, underscored the high accuracy of CXR-based techniques for identifying PTB. Existing studies have primarily focused on using AI for the diagnosis of PTB using CXR [36,37], but the potential of AI to monitor the progress of patients undergoing treatment has been relatively less explored. Unlike the studies mentioned above, which compared healthy patients against patients with PTB, our study differentiated between infectious and noninfectious cases to assess infectivity. Our study proposes the potential use of AI models to analyze CXR images of patients undergoing treatment, enabling the monitoring of treatment progress. In addition, our model can aid in determining the appropriate timing for lifting unquarantined measures. By assessing whether a patient remains infectious, health care providers can make more informed decisions about when it is safe to release patients from quarantine. This not only helps prevent the disease transmission but also reduces the psychological and social burden on patients who may otherwise face longer periods of quarantine than necessary.

Our model facilitates faster decision-making by capitalizing on the quick and accessible nature of CXR, which allows imaging within minutes, and the capability of the proposed model to evaluate CXR images in less than a second. This rapid turnaround time can lead to faster patient discharge or quarantine decisions, emphasizing its efficiency and effectiveness in clinical settings. Given the distinct challenges addressed in this study, comparing our AUROC scores directly with those of other studies may not fully capture its unique value and contributions, which lie in its innovative focus on distinguishing between PTB infectivity stages, rather than merely identifying active cases.

In addition, this study focused on the lung regions in CXR images and used a refined method for segmentation and cropping to accurately target the areas of interest. Thus, it not only improves the diagnostic accuracy by providing a clearer view of the lung regions but also significantly minimizes the impact of irrelevant factors on the decision-making process of the AI model. Building on the groundwork laid by Zaidi et al [38], who showed that lung segmentation can enhance the classification results of CXR analysis by reducing noise and improving efficiency, this study achieved further performance enhancements by exclusively concentrating on segmented and cropped lung areas. Thus, it demonstrated the substantial potential of focused image analysis for improving the accuracy and reliability of AI-driven diagnostics using medical images.

Moreover, our study used large-scale data that were collected over a broad time span, from 2004 to 2023, which was advantageous for analyzing temporal changes and trends in PTB diagnoses and treatments. In addition, data from almost 2 decades allowed us to draw more accurate and reliable conclusions, providing a robust foundation for the AI model to learn and predict with precision. Thus, it offers significant advancements in AI applications to medical imaging, particularly for PTB diagnosis and management, and provides a reliable and clinically applicable tool.

Furthermore, we established a 1-month gray zone following a change in the sputum test results of the patient for labeling purposes. This was necessary to enhance the certainty of the infection status of the patient after the sputum test result became negative because a negative sputum test result does not immediately indicate that the patient is no longer infected. We

XSL•FO
RenderX

excluded CXR images obtained during this period to enhance the accuracy for determining when the likelihood of the contagion decreased to aid in patient treatment and management. This 1-month period was selected based on its relevance compared with 2, 4, 6, and 8 weeks. However, it is important to recognize that using this gray zone could potentially reduce the discriminative power of the model owing to the overlap of continuous and minor changes in images throughout the treatment process. Although extending the gray zone could improve the discrimination power, we elected not to extend it beyond 1 month, which is the minimum interval for follow-ups to assess the infectivity reduction during treatment.

The results of the decision curve analysis demonstrated that the AI model maintained the highest net benefit across most threshold probabilities. This indicates that the AI model is effective in accurately identifying patients with infectious PTB while reducing unnecessary treatments. These findings suggest that the model can significantly contribute to public health efforts in reducing the spread of PTB.

According to the subgroup analysis comparing the model evaluation for the COPD and non-COPD groups, the sensitivity of the model was lower in patients with COPD than in the non-COPD group. This may be because COPD has different pathological features in CXR, including various degrees of inflammation of the lungs, hyperinflation and emphysema, and consolidations and bronchiolitis, which are causes of acute exacerbation episodes [39]. When these features are overlaid with typical CXR findings of PTB infectivity, it may be difficult for AI models to evaluate PTB infectivity, which means that the presence of COPD may disrupt markers of PTB infectivity [40,41], leading to reduced sensitivity.

In the comparison of the all cancers and noncancer groups, we found that the model's evaluation for the noncancer group was lower than that for all cancers for most metrics, except for specificity and NPV. This is because most patients with cancer had been treated with, or were receiving, chemotherapeutic agents. These treatments may impair their cellular and humoral immune systems. Moreover, lung cancer could adversely affect the immune response to PTB [42,43], potentially exacerbating PTB [44], which leads to pronounced changes on CXR [45]. Thus, the AI model may have been better equipped to evaluate signs of infectivity in the CXR images of these patients. In addition, active tuberculosis can increase the risk of cancer, and a cancer diagnosis can also heighten the risk of tuberculosis owing to a mutual interaction [46]. Consequently, the higher sensitivity of AI in detecting PTB in patients with cancer could assist in the screening of tuberculosis in these patients.

Limitations

Although the proposed AI model exhibited notable performance, it has certain limitations. First, although interpretation techniques aid in understanding the decisions of the AI model and provide valuable insights, their interpretations can be complex. While these methods effectively highlight regions significant to the AI decision-making process, such as the right and left upper lobes commonly affected by PTB, they sometimes focus on areas that pose challenges for straightforward clinical correlation. This indicates that although the areas of focus of the proposed AI model were aligned with known PTB-affected regions, these techniques underscore the need for further research to elucidate specific infectivity biomarkers in images. In addition, in some CXR images, the model focused on regions with foreign objects, such as sternotomy suture material, Electrocardiogram leads, and pacemakers, interpreting them as positive cases. These objects may act as confounders. Figure 6 depicts some of these CXRs.

Second, although a 1-month gray period was used to evaluate changes in the infection status, it remains uncertain whether this timeframe is universally applicable to all patients or offers adequate flexibility for diverse scenarios.

Third, although this study emphasized the PTB diagnosis accuracy of the model, acknowledging the inherent constraints of retrospective research, it did not assess the long-term effects on patient outcomes following the application of the AI tool, indicating the necessity for future prospective studies.

Fourth, it included only patients who were successfully treated through a 6-month short-course treatment regimen; this methodology omits a range of treatment outcomes, including delayed response, treatment failure, and multidrug-resistant tuberculosis, and may therefore fail to encompass all phases of the tuberculosis treatment continuum.

Finally, the model may have a bias toward Koreans as the study patients were only Korean. Further experiments with multicountry data, including different ethnicities, are required to address this.

XSL•FO
RenderX

Figure 6. The upper, middle, and lower cases are all positive. The AI evaluated the upper case as positive due to the sternotomy suture material, the middle case due to the pacemaker region, the lower left case due to the EKG leads, and the lower right case due to the chemoport catheter. AI: artificial intelligence; EKG: electrocardiogram.

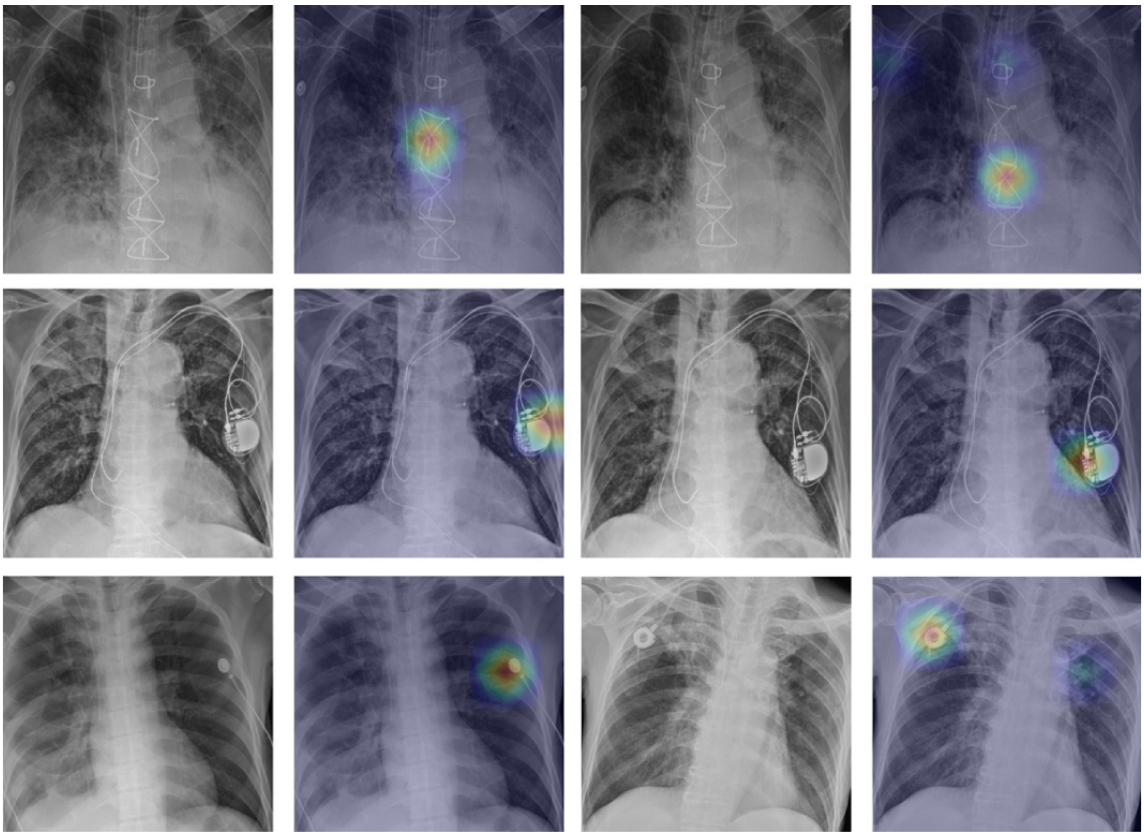

Conclusion

In conclusion, this study represents a step forward in tackling the challenges associated with PTB diagnosis by introducing an AI tool that uses CXR images to assess PTB infectivity. The proposed AI model and interpretation techniques can evaluate PTB infectivity and elucidate its decision-making process, thereby aiding in decisions regarding patient discharge or quarantine. Therefore, the proposed approach aims to minimize the spread of PTB by identifying individuals who pose a risk of infection. Importantly, it offers a more accurate and expedited screening process than conventional smear tests and provides quicker results than culture methods, with the potential to significantly enhance the efficiency and effectiveness of PTB diagnosis and management.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant HI22C0452).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. The developed infectivity evaluation and segmentation models are available on our GitHub repository [47].

Authors' Contributions

All authors have approved this manuscript and contributed to the interpretation of the results and revision of the paper. WYC, DY, and YK contributed to conceptualization. WYC, JY, SK, Kim Y, and Kang Y performed data curation. WYC, JY, DY, and SK performed an investigation. JY, SK, and Kim Y handled the formal analysis. WYC, JY, and Kim Y contributed to the methodology. WYC and DY handled project administration. DY and Kang Y managed resources. WYC, JY, and DY contributed to the writing.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Detailed optimization strategies and hyperparameters in the TransUNet training process.
[DOCX File , 14 KB - jmir_v26i1e58413_app1.docx]

Multimedia Appendix 2
Detailed optimization strategies and hyperparameters in the DenseNet121 training process.
[DOCX File , 43 KB - jmir_v26i1e58413_app2.docx]

Multimedia Appendix 3
Results of comparative experiments with different artificial intelligence (AI) model architectures.
[DOCX File , 16 KB - jmir_v26i1e58413_app3.docx]

Multimedia Appendix 4
Results from each artificial intelligence (AI) model.
[DOCX File , 573 KB - jmir_v26i1e58413_app4.docx]

Multimedia Appendix 5
Performance comparison for COPD, non-COPD, all cancer, and noncancer groups across internal and external validations. COPD: chronic obstructive pulmonary disease.
[DOCX File , 17 KB - jmir_v26i1e58413_app5.docx]

References

1. TB disease burden. World Health Organization. 2024. URL: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023/tb-disease-burden [accessed 2024-10-16]
2. Jilani TN, Avula A, Gondal AZ, Siddiqui AH. Active tuberculosis. StatPearls. 2023 Jan 26. URL: https://www.ncbi.nlm.nih.gov/books/NBK513246/ [accessed 2024-10-16]
3. Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. Consensus statement. global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA 1999;282(7):677-686. [doi: 10.1001/jama.282.7.677] [Medline: 10517722]
4. Siddiqi K, Lambert ML, Walley J. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. Lancet Infect Dis 2003;3(5):288-296. [doi: 10.1016/s1473-3099(03)00609-1] [Medline: 12726978]
5. Min J, Chung C, Jung SS, Park HK, Lee S, Lee KM. Clinical profiles of subclinical disease among pulmonary tuberculosis patients: a prospective cohort study in South Korea. BMC Pulm Med 2020;20(1):316 [FREE Full text] [doi: 10.1186/s12890-020-01351-z] [Medline: 33267859]
6. Kurdula T, Sravan Kumar M, Sumalatha C, Phani Kumar B, Sunitha P. SAS J Med 2022;8(3):225-230. [doi: 10.36347/sasjm.2022.v08i03.022]
7. El Khéchine A, Drancourt M. Diagnosis of pulmonary tuberculosis in a microbiological laboratory. Med Mal Infect 2011;41(10):509-517. [doi: 10.1016/j.medmal.2011.07.012] [Medline: 21917391]
8. Menon LJB, Feliciano CS, de Campos MR, Bollela VR. Decision making to discharge patients from airborne infection isolation rooms: The role of a single GeneXpert MTB/RIF strategy in brazil. Infect Control Hosp Epidemiol 2020;41(7):784-788. [doi: 10.1017/ice.2020.96] [Medline: 32299529]
9. Lee KS, Hwang JW, Chung MP, Kim H, Kwon OJ. Utility of CT in the evaluation of pulmonary tuberculosis in patients without AIDS. Chest 1996;110(4):977-984. [doi: 10.1378/chest.110.4.977] [Medline: 8874255]
10. Roy IN, Chan TY, Czanner G, Wallace S, Vallabhaneni SR. Prospective, single UK centre, comparative study of the predictive values of contrast-enhanced ultrasound compared to time-resolved CT angiography in the detection and characterisation of endoleaks in high-risk patients undergoing endovascular aneurysm repair surveillance: a protocol. BMJ Open 2018;8(4):e020835 [FREE Full text] [doi: 10.1136/bmjopen-2017-020835] [Medline: 29615450]
11. Paquette K, Cheng MP, Kadatz MJ, Cook VJ, Chen W, Johnston JC. Chest radiography for active tuberculosis case finding in the homeless: a systematic review and meta-analysis. Int J Tuberc Lung Dis 2014;18(10):1231-1236. [doi: 10.5588/ijtld.14.0105] [Medline: 25216838]
12. Park S, Sung C, Choi H, Lee YW, Kang Y, Kim H, et al. Comparison of active tuberculosis case finding strategies for immigrants in South Korea: epidemiology and cost-effectiveness analysis. PLoS One 2023;18(4):e0283414 [FREE Full text] [doi: 10.1371/journal.pone.0283414] [Medline: 37079530]
13. Meraj SS, Yaakob R, Azman A, Mohd Rum SN, Ahmad Nazri AS. Artificial intelligence in diagnosing tuberculosis: a review. Int J Adv Sci Eng Inf Technol 2019;9(1):81-91. [doi: 10.18517/ijaseit.9.1.7567]
14. Nachiappan AC, Rahbar K, Shi X, Guy ES, Mortani Barbosa EJ, Shroff GS, et al. Pulmonary tuberculosis: role of radiology in diagnosis and management. Radiographics 2017;37(1):52-72. [doi: 10.1148/rg.2017160032] [Medline: 28076011]

XSL•FO
RenderX

15. Huang F, Zhao Y. Global control of tuberculosis: current status and future prospects. Zoonoses 2022;2(1). [doi: 10.15212/zoonoses-2021-0021]
16. Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, et al. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 2014;33(2):233-245. [doi: 10.1109/TMI.2013.2284099] [Medline: 24108713]
17. Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, et al. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans Med Imaging 2014;33(2):577-590. [doi: 10.1109/TMI.2013.2290491] [Medline: 24239990]
18. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv Preprint posted online on February 8, 2021 [FREE Full text] [doi: 10.48550/arXiv.2102.04306]
19. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv Preprint posted online on August 25, 2016 [FREE Full text] [doi: 10.48550/arXiv.1608.06993]
20. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv Preprint posted online on November 14, 2017 [FREE Full text] [doi: 10.48550/arXiv.1711.05225]
21. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian V. Grad-CAM++: improved visual explanations for deep convolutional networks. arXiv Preprint posted online on October 30, 2017 [FREE Full text] [doi: 10.48550/arXiv.1710.11063]
22. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You"?: explaining the predictions of any classifier. arXiv Preprint posted online on February 16, 2016 [FREE Full text] [doi: 10.48550/arXiv.1602.04938]
23. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary View. J Med Internet Res 2016;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
24. Andreu J, Cáceres J, Pallisa E, Martinez-Rodriguez M. Radiological manifestations of pulmonary tuberculosis. Eur J Radiol 2004;51(2):139-149. [doi: 10.1016/j.ejrad.2004.03.009] [Medline: 15246519]
25. Yu Q, Yan J, Tian S, Weng W, Luo H, Wei G, et al. A scoring system developed from a nomogram to differentiate active pulmonary tuberculosis from inactive pulmonary tuberculosis. Front Cell Infect Microbiol 2022;12:947954 [FREE Full text] [doi: 10.3389/fcimb.2022.947954] [Medline: 36118035]
26. Agustin M, Yamamoto MT, Manaloto E, Chua LA. Black bronchoscopy in healed and active pulmonary tuberculosis. American Thoracic Society 2023:A2055. [doi: 10.1164/ajrccm-conference.2023.207.1_meetingabstracts.a2055]
27. Kazemzadeh S, Yu J, Jamshy S, Pilgrim R, Nabulsi Z, Chen C, et al. Deep learning detection of active pulmonary tuberculosis at chest radiography matched the clinical performance of radiologists. Radiology 2023;306(1):124-137. [doi: 10.1148/radiol.212213] [Medline: 36066366]
28. Lee S, Yim J, Kwak N, Lee YJ, Lee J, Lee JY, et al. Deep learning to determine the activity of pulmonary tuberculosis on chest radiographs. Radiology 2021;301(2):435-442. [doi: 10.1148/radiol.2021210063] [Medline: 34342505]
29. Ting DSW, Tan T, Lim CCT. Development and validation of a deep learning system for detection of active pulmonary tuberculosis on chest Radiographs: clinical and technical considerations. Clin Infect Dis 2019;69(5):748-750. [doi: 10.1093/cid/ciy969] [Medline: 30418534]
30. Higashiguchi M, Nishioka K, Kimura H, Matsumoto T. Prediction of the duration needed to achieve culture negativity in patients with active pulmonary tuberculosis using convolutional neural networks and chest radiography. Respir Investig 2021;59(4):421-427. [doi: 10.1016/j.resinv.2021.01.004] [Medline: 33707161]
31. Horita N, Yamamoto M, Sato T, Tsukahara T, Nagakura H, Tashiro K, et al. Sensitivity and specificity of cobas TaqMan MTB real-time polymerase chain reaction for culture-proven mycobacterium tuberculosis: meta-analysis of 26999 specimens from 17 studies. Sci Rep 2015;5:18113 [FREE Full text] [doi: 10.1038/srep18113] [Medline: 26648113]
32. Chadha VK, Anjinappa SM, Rade K, Baskaran D, Narang P, Kolappan C, et al. Sensitivity and specificity of screening tools and smear microscopy in active tuberculosis case finding. Indian J Tuberc 2019;66(1):99-104. [doi: 10.1016/j.ijtb.2018.05.015] [Medline: 30797292]
33. Gao Y, Zhang Y, Hu C, He P, Fu J, Lin F, et al. Distinguishing infectivity in patients with pulmonary tuberculosis using deep learning. Front Public Health 2023;11:1247141 [FREE Full text] [doi: 10.3389/fpubh.2023.1247141] [Medline: 38089031]
34. Cao CF, Ma KL, Shan H, Liu T, Zhao S, Wan Y, et al. CT scans and cancer risks: a systematic review and dose-response meta-analysis. BMC Cancer 2022;22(1):1238 [FREE Full text] [doi: 10.1186/s12885-022-10310-2]
35. Choi YR, Yoon SH, Kim J, Yoo JY, Kim H, Jin KN. Chest radiography of tuberculosis: determination of activity using deep learning algorithm. Tuberc Respir Dis (Seoul) 2023;86(3):226-233 [FREE Full text] [doi: 10.4046/trd.2023.0020] [Medline: 37183400]
36. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574-582. [doi: 10.1148/radiol.2017162326] [Medline: 28436741]
37. Hwang EJ, Park S, Jin K, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 2019;69(5):739-747 [FREE Full text] [doi: 10.1093/cid/ciy967] [Medline: 30418527]

XSL•FO
RenderX

38. Zaidi SZY, Akram MU, Jameel A, Alghamdi NS. Lung segmentation-based pulmonary disease classification using deep neural networks. IEEE Access 2021;9:125202-125214. [doi: 10.1109/access.2021.3110904]
39. Rangelov BA, Young AL, Jacob J, Cahn AP, Lee S, Wilson FJ, et al. Thoracic imaging at exacerbation of chronic obstructive pulmonary disease: a systematic review. Int J Chron Obstruct Pulmon Dis 2020;15:1751-1787 [FREE Full text] [doi: 10.2147/COPD.S250746] [Medline: 32801677]
40. Tan D, Singh D, Walters H. Mechanisms of impaired anti-bacterial Th1 responses in patients with chronic obstructive pulmonary disease. Eur Respir J 2018;52(suppl 62):PA4999 [FREE Full text] [doi: 10.1183/13993003.congress-2018.pa4999]
41. Ikechukwu AV, Murali S. xAI: an explainable AI model for the diagnosis of COPD from CXR images. 2023 Presented at: IEEE 2nd International Conference on Data, Decision and Systems (ICDDS); December 1-2, 2023; Mangaluru, India p. 1-6. [doi: 10.1109/icdds59137.2023.10434619]
42. Chan ED. Are there common threads in the susceptibility to cancer and tuberculosis? J Thorac Dis 2020;12(5):1776-1780 [FREE Full text] [doi: 10.21037/jtd.2020.03.114] [Medline: 32642083]
43. Wu CY, Hu HY, Pu CY, Huang N, Shen HC, Li CP, et al. Aerodigestive tract, lung and haematological cancers are risk factors for tuberculosis: an 8-year population-based study. Int J Tuberc Lung Dis 2011;15(1):125-130. [Medline: 21276308]
44. Ye MF, Su S, Huang ZH, Zou JJ, Su DH, Chen XH, et al. Efficacy and safety of concurrent anti-tuberculosis treatment and chemotherapy in lung cancer patients with co-existent tuberculosis. Ann Transl Med 2020;8(18):1143 [FREE Full text] [doi: 10.21037/atm-20-5964] [Medline: 33240992]
45. Wetscherek MTA, Sadler TJ, Lee JYJ, Karia S, Babar JL. Active pulmonary tuberculosis: something old, something new, something borrowed, something blue. Insights Imaging 2022;13(1):3 [FREE Full text] [doi: 10.1186/s13244-021-01138-8] [Medline: 35001143]
46. Chen GL, Guo L, Yang SE, Ji DM. Cancer risk in tuberculosis patients in a high endemic area. BMC Cancer 2021;21(1):679 [FREE Full text] [doi: 10.1186/s12885-021-08391-6] [Medline: 34107921]
47. AI and Segmentation Model for PTB Evaluation. GitHub. URL: https://github.com/mean0128/PTB-infectivity-evaluation

Abbreviations

AI: artificial intelligence
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
COPD: chronic obstructive pulmonary disease
CT: computed tomography
CXR: chest radiography
NPV: negative predictive value
PPV: positive predictive value
PTB: pulmonary tuberculosis

Edited by A Mavragani; submitted 18.03.24; peer-reviewed by S Yoo, A Wani, S Mao, J Mistry; comments to author 13.04.24; revised version received 29.05.24; accepted 12.09.24; published 07.11.24.

Please cite as:
Chung WY, Yoon J, Yoon D, Kim S, Kim Y, Park JE, Kang YA
Development and Validation of Deep Learning–Based Infectivity Prediction in Pulmonary Tuberculosis Through Chest Radiography: Retrospective Study
J Med Internet Res 2024;26:e58413
URL: https://www.jmir.org/2024/1/e58413
doi:10.2196/58413
PMID:

©Wou young Chung, Jinsik Yoon, Dukyong Yoon, Songsoo Kim, Yujeong Kim, Ji Eun Park, Young Ae Kang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Pioneering Klebsiella Pneumoniae Antibiotic Resistance Prediction With Artificial Intelligence-Clinical Decision Support System–Enhanced Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: Retrospective Study

Ming-Jr Jian[1*], MS, PhD; Tai-Han Lin[1*], MD; Hsing-Yi Chung[1,2], MS; Chih-Kai Chang[1], MS; Cherng-Lih Perng[1], MS; Feng-Yee Chang[3], MD, PhD; Hung-Sheng Shang[1], MD, PhD

[1]Division of Clinical Pathology, Department of Pathology, Tri-Service General Hospital, National Defense Medical Center, Taipei City, Taiwan

[2]Graduate Institute of Medical Science, National Defense Medical Center, Taipei City, Taiwan

[3]Division of Infectious Diseases and Tropical Medicine, Department of Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Hung-Sheng Shang, MD, PhD
Division of Clinical Pathology, Department of Pathology
Tri-Service General Hospital
National Defense Medical Center
No. 161, Sec. 6, Minquan E. Rd., Neihu Dist.
Division of Clinical Pathology
Taipei City, 11490
Taiwan
Phone: 886 920713130
Email: iamkeith001@gmail.com

Abstract

Background: The rising prevalence and swift spread of multidrug-resistant gram-negative bacteria (MDR-GNB), especially *Klebsiella pneumoniae* (KP), present a critical global health threat highlighted by the World Health Organization, with mortality rates soaring approximately 50% with inappropriate antimicrobial treatment.

Objective: This study aims to advance a novel strategy to develop an artificial intelligence-clinical decision support system (AI-CDSS) that combines machine learning (ML) with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), aiming to significantly improve the accuracy and speed of diagnosing antibiotic resistance, directly addressing the grave health risks posed by the widespread dissemination of pan drug-resistant gram-negative bacteria across numerous countries.

Methods: A comprehensive dataset comprising 165,299 bacterial specimens and 11,996 KP isolates was meticulously analyzed using MALDI-TOF MS technology. Advanced ML algorithms were harnessed to sculpt predictive models that ascertain resistance to quintessential antibiotics, particularly levofloxacin and ciprofloxacin, by using the amassed spectral data.

Results: Our ML models revealed remarkable proficiency in forecasting antibiotic resistance, with the random forest classifier emerging as particularly effective in predicting resistance to both levofloxacin and ciprofloxacin, achieving the highest area under the curve of 0.95. Performance metrics across different models, including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F_1-score, were detailed, underlining the potential of these algorithms in aiding the development of precision treatment strategies.

Conclusions: This investigation highlights the synergy between MALDI-TOF MS and ML as a beacon of hope against the escalating threat of antibiotic resistance. The advent of AI-CDSS heralds a new era in clinical diagnostics, promising a future in which rapid and accurate resistance prediction becomes a cornerstone in combating infectious diseases. Through this innovative approach, we answered the challenge posed by KP and other multidrug-resistant pathogens, marking a significant milestone in our journey toward global health security.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58039) doi:10.2196/58039

KEYWORDS

Klebsiella pneumoniae; multidrug resistance; AI-CDSS; quinolone; ciprofloxacin; levofloxacin

Introduction

The World Health Organization (WHO) has highlighted antibiotic resistance as a grave global health threat, particularly emphasizing the challenge posed by multidrug-resistant *Klebsiella pneumoniae* (MDR-KP) [1]. Bacteremia caused by MDR-KP is linked to mortality rates as high as 47%, underlining the critical impact of multidrug resistance [2-4]. Initially, quinolones were highly effective against such infections due to their unique mechanism of action, distinct from β-lactams, targeting different bacterial processes [3,4]. However, increasing quinolone resistance among KP, driven by mutations in key enzyme regions, efflux pump overexpression, and plasmid-mediated mechanisms, undermines quinolone effectiveness, complicating infection management and underscoring the adaptability of bacterial resistance [3,4].

The emergence of quinolone resistance in KP strains within clinical settings indicates the critical demand for novel diagnostic techniques [2-7]. Rapid and accurate detection of antibiotic-resistant pathogens in patients is essential for controlling resistance spread [8,9]. Thus, the slow pace of culture-based diagnostics in addressing antibiotic resistance required shifting toward new, more agile strategies, emphasizing the need for quicker, more effective treatment options to navigate the challenges of rapidly evolving infections [10].

The dynamic battle against antibiotic resistance necessitates quick-to-adapt diagnostic tools for effective management [10]. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is indispensable for the rapid and accurate identification of microorganisms [11,12]. Its application extends to the detection of virulence factors in antibiotic-resistant strains yet often transcends standard procedural bounds [11,12]. Combining machine learning (ML) with MALDI-TOF MS ushers in a novel paradigm for predicting antibiotic resistance [13], specifically targeting quinolone-resistant KP.

Our study introduces the development of an artificial intelligence-clinical decision support system (AI-CDSS) designed to enhance the clinical decision-making process regarding the use of quinolones as a potential antibiotic option. Integrating the precision of MALDI-TOF MS technology with advanced ML algorithms, our system aims to equip health care professionals with quick and evidence-based recommendations for the optimal use of antibiotics tailored to individual patient needs. This innovative approach not only improves the rapidity and precision of identifying antibiotic resistance but also offers clinicians valuable insights, enabling prompt and well-informed treatment choices. Consequently, this facilitates improved patient outcomes and more effective combat against the proliferation of antibiotic resistance.

Methods

Study Designs and Data Collection

From January 2021 to December 2023, a comprehensive research project was undertaken by the Tri-Service General Hospital alongside 4 district hospitals, examining 165,299 bacterial specimens using advanced MALDI-TOF MS technology provided by bioMérieux, France. This study identified 11,996 specimens as KP, which were further subjected to antibiotic sensitivity testing (AST) for levofloxacin (LEV) and ciprofloxacin (CIP) using the VITEK 2 system (bioMérieux). This system is claimed to be precise in such assessments. The analysis of the AST results strictly adhered to the clinical breakpoints defined by the Clinical and Laboratory Standards Institute, ensuring that the study aligned with the established microbiological norms [14,15].

The objective of this extensive analysis was to develop 2 predictive models: one to ascertain resistance to LEV and the other to CIP. The exclusion of 869 specimens from the LEV resistance analysis was necessary because of intermediate AST results, which refined the focus to 11,127 specimens for thorough investigation. Similarly, for CIP resistance, 373 specimens were excluded because of intermediate results, concentrating the analysis on 11,623 specimens. This detailed and systematic process emphasizes the rigor of the study and its commitment to deliver precise and reliable evaluations of antibiotic resistance among microbial specimens. The dataset was strategically divided as follows: samples collected from January to September were designated for training the models, whereas those collected from October to December served as the validation set. This temporal split was performed to simulate the application of the models in real-world settings.

Data Preprocessing and Feature Extraction

Our research was initiated by extracting mass-to-charge (m/z) ratios and intensity values from MALDI-TOF MS data to establish an essential foundation. We addressed data imbalance through a resampling strategy using sklearn.utils, applying downsampling to over-represented classes and upsampling for underrepresented classes, thus balancing the dataset to diminish model bias and enhance generalization. In feature engineering, we segmented m/z ranges into 1-unit intervals, allowing a ± 2-unit tolerance for grouping similar observations. We then extracted and logarithmically transformed the peak intensity values for both the antibiotic-resistant and antibiotic-susceptible strains within these segments. This crucial step normalizes the data, improves strain comparability, and enables precise identification of resistance pattern variations. We filtered out segments with a feature importance value below 0.01 as background noise, focusing our model on significant attributes to optimize performance and achieve our research objectives.

XSL•FO
RenderX

ML Model Development and Training

To detect LEV and CIP resistance in KP, we selected an array of sophisticated algorithms including Logistic Regression, Linear Discriminant Analysis, Random Forest, Gradient Boosting Classifier, AdaBoost Classifier, XGBoost, and light gradient boosting machine (LGBM). This diverse selection was aimed at leveraging their distinct strengths for a more precise and comprehensive analysis. Our goal was to thoroughly explore antibiotic resistance patterns to improve treatment strategies and provide valuable insights into the management of microbial resistance. A grid search was performed for each model to determine the optimal parameters, enhancing accuracy and performance. The best parameters identified for each model are detailed in Multimedia Appendix 1.

Model Evaluation

To assess the accuracy of our predictive models, we used several key metrics, including receiver operating characteristic (ROC) curves, area under the curve (AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F_1-score. These metrics provide a detailed evaluation of the effectiveness of the model and ensure that the findings are reliable and robust. Such a comprehensive evaluation method is vital for predicting antibiotic resistance.

Clinical Validation

To affirm the applicability of our models in real-world clinical settings, we conducted clinical validation at Tri-Service General Hospital using the validation dataset. During the validation period, we specifically validated 3339 cases of LEV resistance and 3487 cases of CIP resistance against traditional culture-based AST. We evaluated the predictive accuracy by analyzing sensitivity, specificity, PPV, NPV, and F1-score.

AI-CDSS Deployment

We created a web-based AI-CDSS specifically designed for health care professionals. This system leverages cutting-edge ML algorithms to quickly determine the resistance to LEV and CIP. The process began with the collection of a culture sample, which the AI-CDSS then analyzed to predict resistance patterns. For a clear understanding of our process, The flowchart in Figure 1 represents the systematic process behind the development of the AI-CDSS, which commenced with a comprehensive data collection phase, acquiring 165,299 bacterial specimens from January 2021 to December 2023. Post identification of *Klebsiella pneumoniae* within the samples, they were subjected to AST. Specimens yielding intermediate AST results were excluded, resulting in 11,127 for LEV and 11,623 for CIP resistance analyses, thus forming a training dataset. Subsequent data preprocessing involved meticulous feature extraction and log transformation of mass spectrometry data to eliminate noise and standardize the dataset. This dataset was then used to train robust ML models, such as Random Forest, LGBM, and XGBoost, to accurately predict antibiotic resistance. The AI-CDSS, by leveraging these models, facilitates rapid and precise predictions of LEV and CIP resistance, ultimately aiding health care professionals in informed clinical decision-making.

Figure 1. AI-CDSS development workflow. AI-CDSS: artificial intelligence-clinical decision support system; AST: antibiotic susceptibility testing; LEV: levofloxacin; CIP: ciprofloxacin; RF: random forest; LGBM: light gradient boosting machine; XGB: XGBoost.

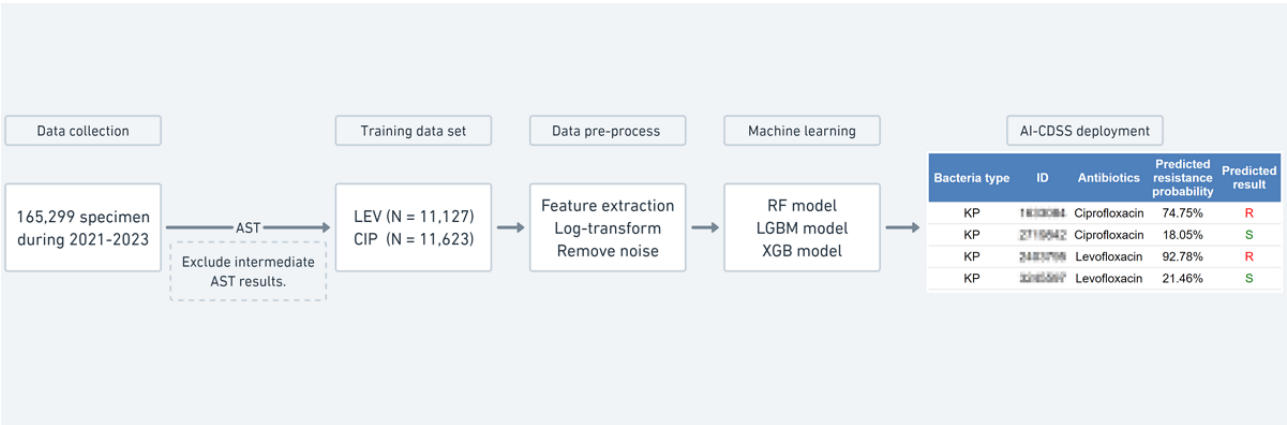

Ethical Considerations

This retrospective study adhered to the ethical standards outlined in the Declaration of Helsinki and relevant local regulations for human subject protection. The institutional review board of Tri-Service General Hospital, Taipei, Taiwan, approved the research (TSGHIRB No. C202305073). Due to the study's retrospective design, the requirement for informed consent was waived. All analyzed data were anonymized to ensure patient confidentiality.

Results

ML Model Development Using MS Data

MS analyses revealed distinct spectral differences that differentiated antibiotic-resistant KP strains from their susceptible counterparts. Figure 2 shows the average intensity distributions across m/z segments for KP strains, with CIP resistance shown in blue and susceptibility in red. The top graph (A) highlights the m/z segments that differ between CIP-resistant and susceptible strains, whereas the bottom graph (B) depicts analogous distributions for LEV resistance. These graphs provide a stark visual comparison of the spectral differences, emphasizing the distinct spectral markers associated with

XSL•FO
RenderX

antibiotic resistance that can be instrumental for rapid resistance detection.

Figure 2. Differential mass spectrometry profiles for CIP and LEV resistance in Klebsiella pneumoniae. (A) Average intensity distribution for CIP-resistant and susceptible Klebsiella pneumoniae. (B) Average intensity distribution for LEV-resistant and susceptible Klebsiella pneumoniae. CIP: ciprofloxacin; LEV: levofloxacin; m/z: mass-to-charge.

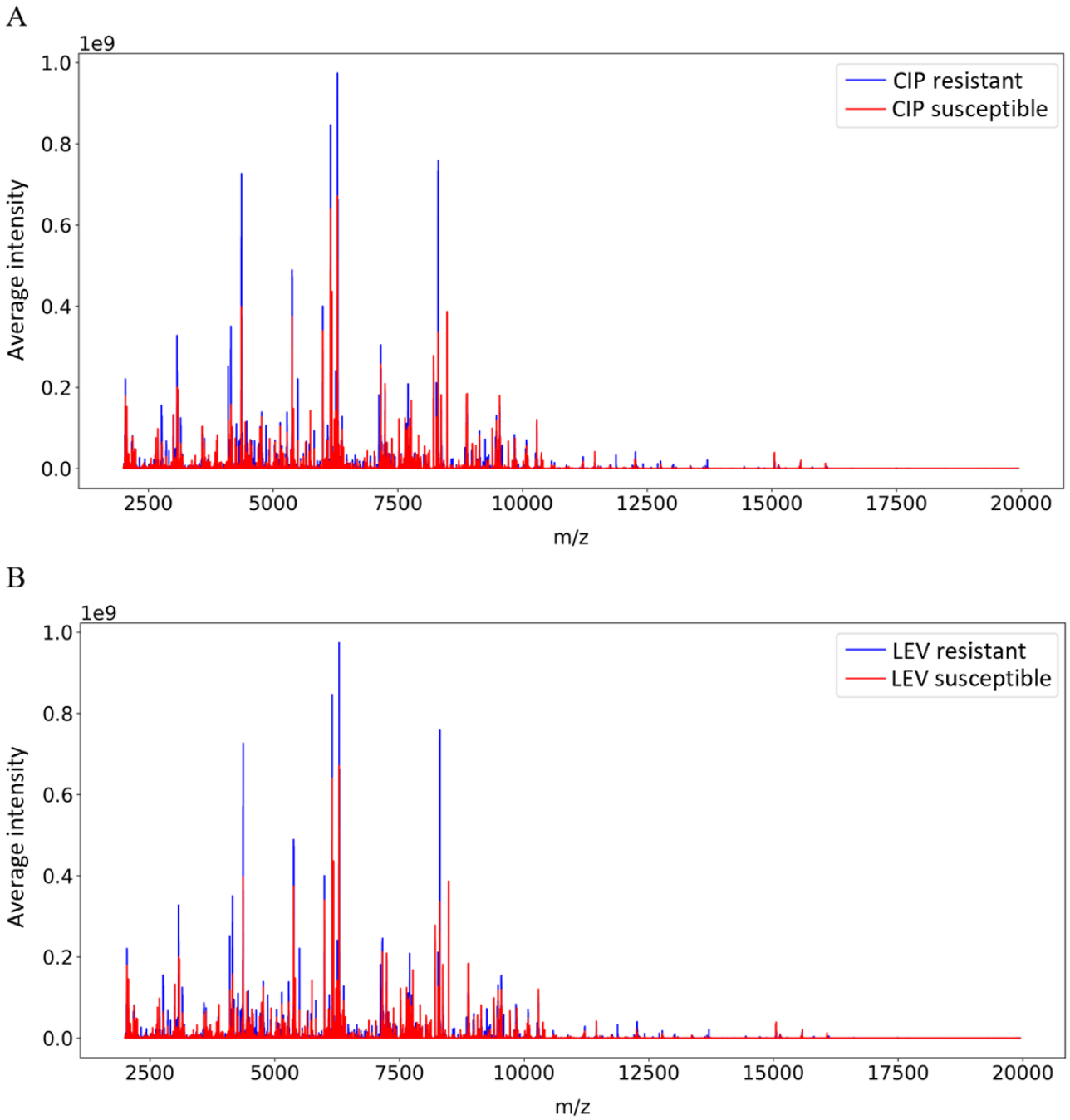

MS data for KP showed a consistent array of the top 15 spectral features in strains resistant to both CIP and LEV. These features span m/z ratio ranges of 2035-2040, 2067-2072, 2166-2171, 2180-2185, 2688-2695, 2759-2764, 3143-3152, 4363-4368, 4567-4572, 5278-5283, 5379-5384, 7238-7253, 7703-7708, 9136-9145, and 9532-9544. The shared features across CIP- and LEV-resistant strains suggest a universal mechanism of resistance despite the different modes of action of antibiotics.

This is graphically represented in the heat maps in Figure 3A for ciprofloxacin and Figure 3B for levofloxacin, where the correlation between each m/z ratio range and resistance status are color-coded for clarity. The intensity of the color correlates with the strength of the association; the redder the hue, the stronger the correlation with resistance. These heat maps provide a visual summary of the data, emphasizing the features most strongly associated with resistance, thus informing future molecular studies and aiding in the development of new approaches to combat antibiotic resistance. In Figure 3, The top graph (A) shows the correlation between the top 15 m/z segments and CIP resistance, whereas the bottom graph (B) illustrates these correlations for LEV resistance. The heatmaps transition from blue to red, indicating an increasing strength of correlation from weaker to stronger, respectively. These patterns

XSL•FO
RenderX

highlighted the top 15 differential m/z segments that were most indicative of antibiotic resistance.

Figure 3. Correlation heatmaps of Top 15 m/z features for CIP and LEV resistance in Klebsiella pneumoniae. (A) Correlation heatmap for CIP resistance. (B) Correlation heatmap for LEV resistance. m/z: mass-to-charge; CIP: ciprofloxacin; LEV: levofloxacin.

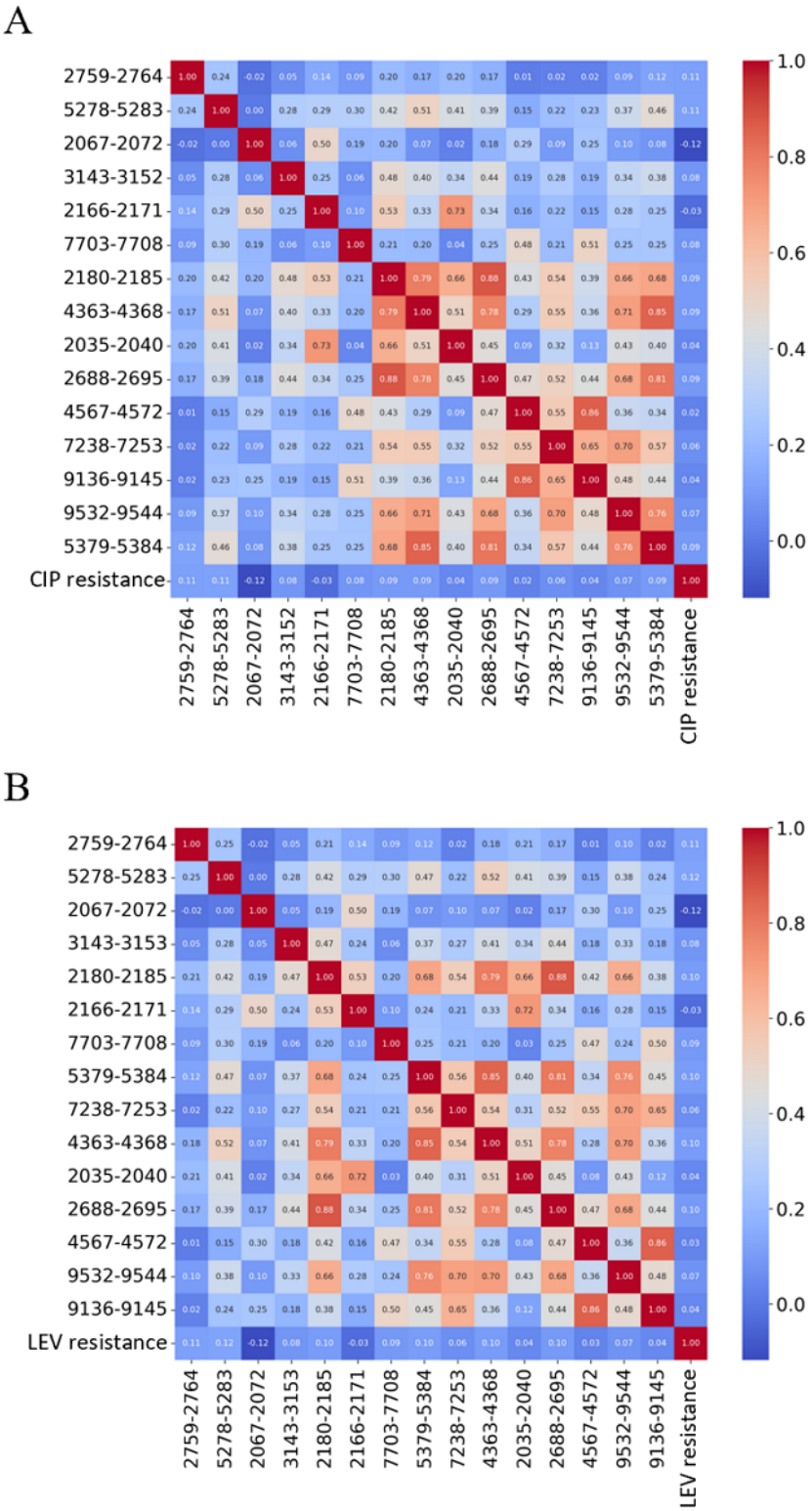

Model Performance Metrics, Validation Results, and ROC Curve Analyses

To assess the efficacy of ML models in predicting antibiotic resistance in KP, we evaluated several algorithms across 2 different drug resistances. The model performance metrics, including the AUC, accuracy, sensitivity, specificity, PPV, NPV, and F_1-score, are detailed in Tables 1 and 2 for CIP and LEV resistance, respectively. For CIP resistance, the Random Forest demonstrated the highest testing AUC at 0.95, indicative of its robust predictive power, with Gradient Boosting Classifier and XGBoost also performing notably well with AUCs of 0.95 and 0.94, respectively. These models are particularly useful in

XSL•FO
RenderX

identifying CIP-resistant KP strains with high accuracy and precision.

Table 1. Performance metrics of machine learning models for ciprofloxacin resistance.

Models	Training AUC[a]	Testing AUC	Accuracy	Sensitivity	Specificity	PPV[b]	NPV[c]	F_1-score
RF[d]	0.99	0.95	0.87	0.88	0.86	0.81	0.91	0.85
LGBM[e]	0.99	0.95	0.90	0.88	0.91	0.87	0.92	0.88
GBC[f]	0.99	0.95	0.87	0.92	0.84	0.80	0.93	0.86
XGBoost	0.99	0.95	0.87	0.89	0.86	0.82	0.92	0.85
AdaBoost	0.99	0.87	0.77	0.81	0.74	0.69	0.85	0.74
LR[g]	0.91	0.84	0.78	0.77	0.78	0.71	0.83	0.74
LDA[h]	0.96	0.80	0.77	0.80	0.75	0.69	0.84	0.74

[a]AUC: area under the curve.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]RF: random forest.
[e]LGBM: light gradient boosting machine.
[f]GBC: gradient boosting classifier.
[g]LR: logistic regression.
[h]LDA: linear discriminant analysis.

Table 2. Performance metrics of machine learning models for Levofloxacin resistance.

Models	Training AUC[a]	Testing AUC	Accuracy	Sensitivity	Specificity	PPV[b]	NPV[c]	F_1-score
RF[d]	0.99	0.95	0.86	0.89	0.82	0.86	0.86	0.87
GBC[e]	0.99	0.95	0.87	0.88	0.86	0.89	0.85	0.88
XGBoost	0.99	0.93	0.85	0.86	0.85	0.87	0.83	0.86
LGBM[f]	0.99	0.92	0.85	0.84	0.85	0.87	0.82	0.86
AdaBoost	0.99	0.90	0.85	0.85	0.84	0.87	0.83	0.86
LR[g]	0.77	0.73	0.66	0.74	0.57	0.68	0.64	0.70
LDA[h]	0.72	0.68	0.64	0.81	0.43	0.63	0.65	0.71

[a]AUC: area under the curve.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]RF: random forest.
[e]GBC: gradient boosting classifier.
[f]LGBM: light gradient boosting machine.
[g]LR: logistic regression.
[h]LDA: linear discriminant analysis.

The LEV resistance models displayed a similar trend, with the Random Forest classifier again showing exceptional performance, with an AUC of 0.95. The gradient boosting classifier and XGBoost showed similar effectiveness, with AUCs of 0.95 and 0.93. These algorithms proved effective in distinguishing LEV-resistant and LEV-susceptible strains with considerable accuracy. The ROC curves, as presented in Figures 4A and B, visually encapsulate the model performance, with the Random Forest, XGBoost, and Gradient Boosting classifiers exhibiting steep ascents toward the upper left corner, denoting high true-positive rates and low false-positive rates. Although logistic regression and Linear Discriminant Analysis showed lower AUC values, their inclusion offered a broad perspective and contributed to the ensemble approach in model evaluation. In Figure 4, the 2 charts display the performance of various ML models in classifying *Klebsiella pneumoniae* strains as resistant or susceptible to 2 antibiotics: CIP and LEV. The top panel (A) displays the ROC curves for models predicting CIP resistance, with the Random Forest classifier achieving an AUC of 0.95, indicating its excellent predictive ability. The bottom panel (B) displays the ROC curves for models predicting LEV resistance, where the random forest (RF) classifier also achieves a high

XSL•FO
RenderX

AUC of 0.95, demonstrating consistently high performance across different antibiotic resistances. The curves plot each model's true positive rate (sensitivity) against its false positive rate (1-specificity), with the diagonal line representing a random chance. The closer a curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

Figure 4. ROC curves for predicting antibiotic resistance in Klebsiella pneumoniae. (A) ROC curves for CIP resistance prediction. (B) ROC curves for LEV resistance prediction. ROC: receiver operating characteristic; CIP: ciprofloxacin; LEV: levofloxacin; RF: random forest; GBC: gradient boosting classifier; LGBM: light gradient boosting machine; LR: logistic regression; LDA: linear discriminant analysis.

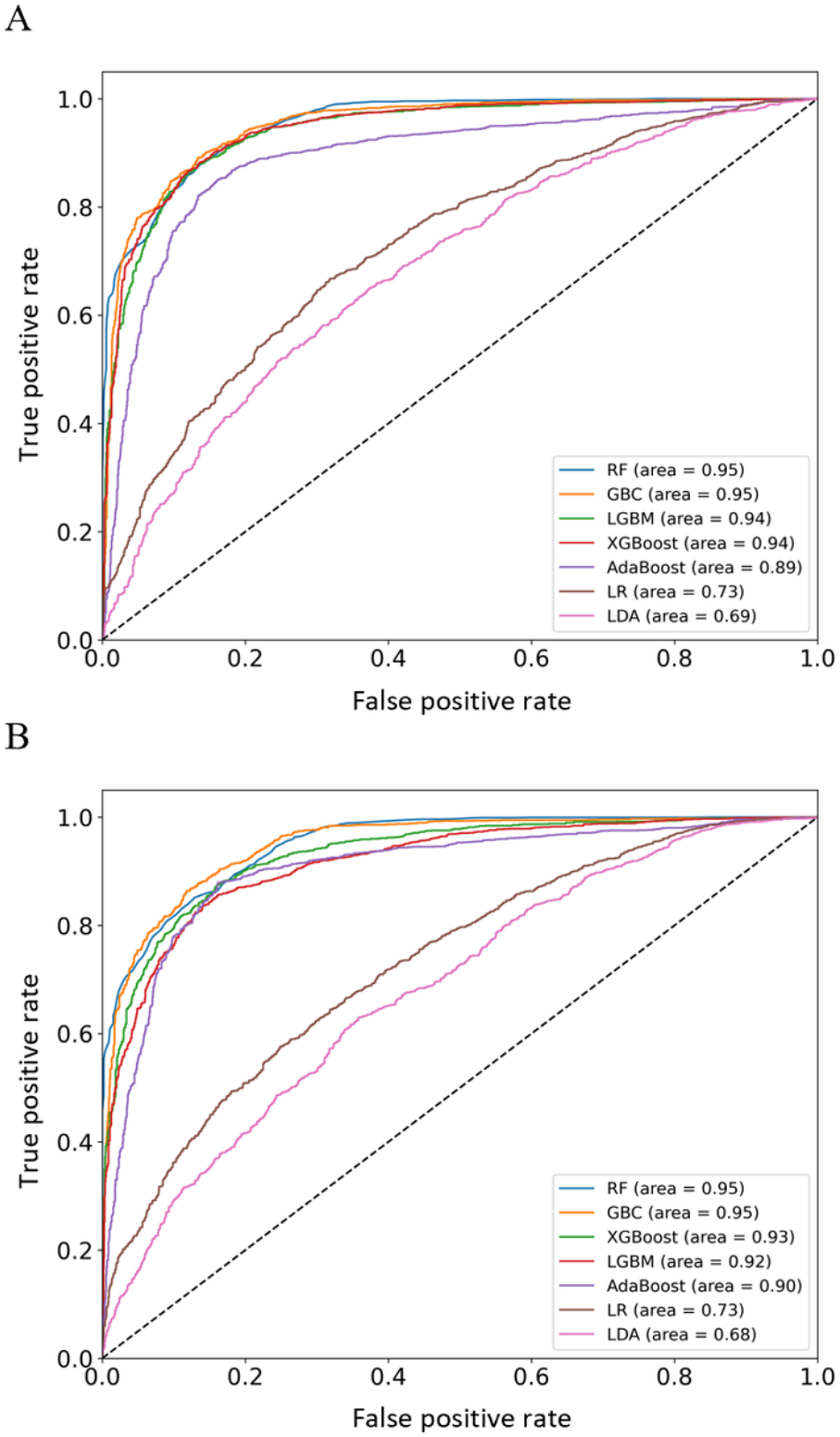

Despite their range of performance metrics, our ensemble of models collectively provided a robust method for assessing KP resistance to CIP and LEV. This comprehensive approach, which leverages the strengths of multiple predictive algorithms, offers valuable insights into antibiotic resistance patterns, aiding the development of precision treatment strategies using our AI-CDSS.

XSL•FO
RenderX

Integration of AI-CDSS in the Clinical Process

The integration of our AI-CDSS platform into the antibiotic resistance–testing workflow represents a significant advancement in the treatment of infectious diseases. This system streamlines the evaluation process from the initial culture of samples to the final decision-making in clinical settings. Using advanced ML algorithms, the platform rapidly predicts resistance to antibiotics, such as LEV and CIP, providing clinicians with precise and efficient analyses that inform treatment strategies. This is particularly critical for managing cases of KP, for which timely intervention is crucial. The platform's predictive capabilities allow for resistance determination for approximately 3 days, a notable improvement over the conventional 4-day timeline required for standard AST. By accelerating the decision-making process, the AI-CDSS contributes to improved patient outcomes and enhances antibiotic stewardship efforts, ensuring that treatments are both appropriate and effective. The workflow diagram clearly contrasts the AI-CDSS with traditional AST methods, highlighting the reduction in time to clinical decisions and underscoring the benefits of real-time data in the management of infectious diseases.

Discussion

Principal Findings

This investigation of the predictive capabilities of ML and MALDI-TOF MS for identifying quinolone-susceptible KP and quinolone-resistant KP signifies a groundbreaking development in the expedited diagnosis of infectious diseases. The synergy between MALDI-TOF MS and ML provides an innovative approach to detecting and managing antibiotic-resistant pathogens. This advancement is particularly relevant for low- and middle-income countries (LMICs), where the WHO has noted a rapid escalation in antibiotic use, presenting challenges to access and stewardship [1,16]. Our findings are of significant value in LMICs, as they propose a strategic balance between antibiotic availability and stewardship, which could influence policy adaptations aimed at fulfilling WHO objectives and strengthening global health security against antimicrobial resistance.

The increasing prevalence of antibiotic resistance, notably in quinolone-resistant KP strains, necessitates novel diagnostic methods [2-9]. Traditional methods fail to swiftly identify these bacteria, thereby prolonging the initiation of suitable treatments and containment actions [10]. Previous research on ML with MALDI-TOF for discerning quinolone-resistant KP has used a modest number of isolates. Our study compiles an unprecedentedly large dataset for a ML model to detect quinolone-resistant KP, demonstrating the potential of advanced technologies in predicting resistance with high precision, thereby expediting accurate diagnoses and enabling effective treatments. The incorporation of algorithms such as Random Forest and XGBoost has shown high efficacy in differentiating between resistant and susceptible KP strains, playing a pivotal role in informed antibiotic selection and combating the spread of resistance.

Comparative Advantages of MALDI-TOF MS and Next-Generation Sequencing in Clinical Settings

In clinical practice, KP was identified using MALDI-TOF MS and subsequently tested for antibiotic susceptibility with the VITEK 2 system. In our study, we used artificial intelligence to analyze MALDI-TOF MS spectra and predict KP as resistant or susceptible phenotypes. Although next-generation sequencing (NGS) is commonly used for genotype analysis in other studies, genotypic data may not always align with phenotypic expressions. Antibiotic resistance typically develops through genetic alterations, either by the acquisition of resistance genes or mutations in elements crucial for antibiotic activity. However, resistance can also occur without any genetic changes, a phenomenon known as phenotypic resistance illustrating the challenges of predicting antibiotic resistance based solely on genetic information [17-19]. Discrepancies may arise from pseudogenes in low antibiotic environments, where nonlethal mutations persist in contrast to the rapid elimination of ineffective genes in high antibiotic environments [18]. In addition, gene interactions, regulatory mechanisms, and expression levels significantly impact resistance gene functionality [19-21]. Our study emphasizes the ability of MALDI-TOF MS to quickly provide crucial, actionable resistance profiles for urgent clinical decisions [22,23]. NGS can provide a detailed analysis of resistance genes; however, its longer processing times limit its use in emergencies [19-21]. In contrast, MALDI-TOF MS can detect these mechanisms through its protein expression analysis, proving invaluable in clinical settings that require rapid decision-making [22,23].

Speeding Up Pathogen Diagnostics With AI and MALDI-TOF MS

Our integration of AI-CDSS with MALDI-TOF MS technology offers a significant improvement in diagnostic speed and efficiency that answers the need for rapid diagnostics for infectious diseases, particularly those involving drug-resistant pathogens such as KP. Our approach reduces the time needed to determine antibiotic resistance to approximately 72 hours from the initial culture. Traditional culturing methods, while reliable, require up to 96 hours—72 hours for culturing and identification and an additional 24 hours for AST interpretation, potentially delaying critical treatment decisions [10,24]. In comparison, polymerase chain reaction (PCR) and sequencing methods still require 72 hours for the initial pathogen identification by MALDI-TOF MS, followed by 1-2 days for PCR and 3-4 days for sequencing, depending on the complexity of the analysis [20,21,25,26]. Comparisons of the timelines for each method were provided in Multimedia Appendix 2. Therefore, our study demonstrates that the integration of AI-CDSS with MALDI-TOF MS not only significantly speeds up the diagnostic process but also improves the timeliness of clinical responses, crucial for managing severe infections.

Impact and Future Directions

In addition to the advancements detailed above, the integration of an AI-CDSS is a groundbreaking enhancement of our diagnostic capabilities. The AI-CDSS, which leverages the predictive power of ML models, offers clinicians real-time, data-driven insights, significantly improving the

XSL•FO
RenderX

decision-making process in the treatment of infections caused by LEV- and CIP-resistant KP. This system can rapidly interpret complex diagnostic data, recommend personalized treatment options, and predict potential resistance patterns, thereby streamlining the diagnostic workflow and facilitating a more targeted approach to antibiotic therapy. The introduction of AI-CDSS into our methodology underscores our commitment to harness cutting-edge technology to combat antibiotic resistance, representing a vital step forward in the optimization of clinical outcomes and advancement of antibiotic stewardship.

Limitations

This study underscores the intricacies involved in dataset balancing and the pivotal role of feature selection within the domain of ML algorithms. A comprehensive approach to data preprocessing, feature extraction, and the subsequent training of models delineates the multifaceted nature of constructing dependable predictive frameworks. In addition, intermediate results often present challenges in predictive models due to the uncertainty they introduce, which can potentially skew the model's outcomes. Consequently, the exclusion of specimens with intermediate AST results (869 for LEV and 373 for CIP) allowed for a more focused study but potentially omitted significant data, affecting model generalizability for these intermediate cases. Future research should focus on enhancing these models by incorporating intermediate cases, as well as diverse data modalities, to broaden their applicability and encompass a wider array of antibiotic-resistant organisms. Furthermore, the significance of this investigation transcends the boundaries of immediate clinical use and serves as a cornerstone for informing public health policies and shaping infection control methodologies. This study makes a substantial contribution to international efforts aimed at combating antimicrobial resistance by providing a swift, precise, and economically viable mechanism to identify antibiotic resistance. This issue remains a global public health priority and requires continued attention and action.

Comparison With Previous Work

Previous studies used conventional PCR-based methods to identify quinolone resistance genes among KP clinical isolates [27,28]. ML techniques combined with MALDI-TOF of discrimination of carbapenem-resistant KP with limited sample isolates [29,30]. To date, there are no studies discussing the use of ML combined with MALDI-TOF for the rapid identification of quinolone resistance in KP. Our research establishes the most extensive collection of isolates to date for constructing a ML model aimed at identifying LEV and CIP resistance. The methodological innovation of this study, through the analysis of approximately 12,000 KP cultures, demonstrates the feasibility and accuracy of using advanced technologies to predict resistance patterns. This approach not only enhances the precision of diagnostics but also significantly reduces the time required to identify resistant strains, facilitating timely and targeted therapeutic interventions.

Conclusions

MALDI-TOF MS is a swift, accurate, and cost-efficient method for identifying bacteria, including its application in the detection of antibiotic resistance. This study highlights the successful AI-CDSS with the application of a ML-enhanced MALDI-TOF approach for predicting quinolone-resistant KP. However, certain constraints exist, particularly when resistance mechanisms are not mediated at the protein-peptide level or fall outside the m/z ratio range of 2000-20,000. Future studies are required to validate the results further.

Acknowledgments

We would like to thank Editage for its English language editing service.

The grants awarded by the hospital (grant numbers TSGH-D-113105 and TSGH-D-113106) were pivotal in facilitating our research.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

M-JJ contributed to the conceptualization, formal analysis, investigation, data curation, and writing of the original draft. T-HL was involved in formal analysis, data interpretation, and writing – review and editing. H-YC contributed to data curation and data interpretation. C-KC was responsible for conceptualization and investigation. C-LP focused on data curation. F-YC contributed to data interpretation. H-SS was involved in conceptualization, supervision, methodology, investigation, project administration, and writing – both the original draft and review & editing. All authors contributed to the manuscript preparation and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Parameters of machine learning models for predicting Levofloxacin and Ciprofloxacin resistance.

XSL•FO
RenderX

[DOCX File , 22 KB - jmir_v26i1e58039_app1.docx]

Multimedia Appendix 2
Workflow comparison of diagnostic techniques for antibiotic resistance detection.
[PNG File , 112 KB - jmir_v26i1e58039_app2.png]

References

1. WHO. WHO Publishes List of Bacteria For Which New Antibiotics Are Urgently Needed. 2017. URL: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed [accessed 2017-02-27]
2. Capone A, Giannella M, Fortini D, Giordano A, Meledandri M, Ballardini M, et al. High rate of colistin resistance among patients with carbapenem-resistant Klebsiella pneumoniae infection accounts for an excess of mortality. Clin Microbiol Infect 2013;19(1):E23-E30. [doi: 10.1111/1469-0691.12070] [Medline: 23137235]
3. Navon-Venezia S, Kondratyeva K, Carattoli A. Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev 2017;41(3):252-275. [doi: 10.1093/femsre/fux013] [Medline: 28521338]
4. Lautenbach E, Metlay JP, Bilker WB, Edelstein PH, Fishman NO. Association between fluoroquinolone resistance and mortality in escherichia coli and klebsiella pneumoniae infections: the role of inadequate empirical antimicrobial therapy. Clin Infect Dis 2005;41(7):923-929. [doi: 10.1086/432940] [Medline: 16142655]
5. Hooper DC, Jacoby GA. Topoisomerase inhibitors: fluoroquinolone mechanisms of action and resistance. Cold Spring Harb Perspect Med 2016;6(9):a025320. [doi: 10.1101/cshperspect.a025320] [Medline: 27449972]
6. Du D, Wang-Kan X, Neuberger A, van Veen HW, Pos KM, Piddock LJV, et al. Multidrug efflux pumps: structure, function and regulation. Nat Rev Microbiol 2018;16(9):523-539. [doi: 10.1038/s41579-018-0048-6] [Medline: 30002505]
7. Hooper DC, Jacoby GA. Mechanisms of drug resistance: quinolone resistance. Ann N Y Acad Sci 2015;1354(1):12-31. [doi: 10.1111/nyas.12830] [Medline: 26190223]
8. Müller L, Srinivasan A, Abeles SR, Rajagopal A, Torriani FJ, Aronoff-Spencer E. A Risk-Based clinical decision support system for Patient-Specific antimicrobial therapy (iBiogram): design and retrospective analysis. J Med Internet Res 2021;23(12):e23571. [doi: 10.2196/23571] [Medline: 34870601]
9. Tseng YJ, Wu JH, Ping XO, Lin HC, Chen YY, Shang RJ, et al. A Web-based multidrug-resistant organisms surveillance and outbreak detection system with rule-based classification and clustering. J Med Internet Res 2012;14(5):e131. [doi: 10.2196/jmir.2056] [Medline: 23195868]
10. Garcia E, Diep JK, Sharma R, Rao GG. Model-based learn and confirm: designing effective treatment regimens against multidrug resistant Gram-negative pathogens. Int J Antimicrob Agents 2024;63(4):107100. [doi: 10.1016/j.ijantimicag.2024.107100] [Medline: 38280574]
11. Weis C, Cuénod A, Rieck B, Dubuis O, Graf S, Lang C, et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nat Med 2022;28(1):164-174. [doi: 10.1038/s41591-021-01619-9] [Medline: 35013613]
12. Yu J, Tien N, Liu YC, Cho DY, Chen JW, Tsai YT, et al. Rapid identification of Methicillin-Resistant staphylococcus aureus using MALDI-TOF MS and machine learning from over 20,000 clinical isolates. Microbiol Spectr 2022;10(2):e0048322. [doi: 10.1128/spectrum.00483-22] [Medline: 35293803]
13. Jang KS, Kim YH. Rapid and robust MALDI-TOF MS techniques for microbial identification: a brief overview of their diverse applications. J Microbiol 2018;56(4):209-216. [doi: 10.1007/s12275-018-7457-0] [Medline: 29492868]
14. Wayne P, CLSI. Performance Standards for Antimicrobial Susceptibility Testing. 28th ed. Pennsylvania: Clinical and Laboratory Standards Institute; 2018.
15. Humphries RM, Abbott AN, Hindler JA. Understanding and addressing CLSI breakpoint revisions: a primer for clinical laboratories. J Clin Microbiol 2019;57(6):e00203 [FREE Full text] [doi: 10.1128/JCM.00203-19] [Medline: 30971460]
16. Klein EY, Van Boeckel TP, Martinez EM, Pant S, Gandra S, Levin SA, et al. Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. Proc Natl Acad Sci U S A 2018;115(15):E3463-E3470. [doi: 10.1073/pnas.1717295115] [Medline: 29581252]
17. Hughes D, Andersson DI. Environmental and genetic modulation of the phenotypic expression of antibiotic resistance. FEMS Microbiol Rev 2017;41(3):374-391. [doi: 10.1093/femsre/fux004] [Medline: 28333270]
18. Davis MA, Besser TE, Orfe LH, Baker KNK, Lanier AS, Broschat SL, et al. Genotypic-phenotypic discrepancies between antibiotic resistance characteristics of escherichia coli isolates from calves in management settings with high and low antibiotic use. Appl Environ Microbiol 2011;77(10):3293-3299. [doi: 10.1128/AEM.02588-10] [Medline: 21421795]
19. Deekshit VK, Srikumar S. 'To be, or not to be'-The dilemma of 'silent' antimicrobial resistance genes in bacteria. J Appl Microbiol 2022;133(5):2902-2914. [doi: 10.1111/jam.15738] [Medline: 35882476]
20. Johansen Taber KA, Dickinson BD, Wilson M. The promise and challenges of next-generation genome sequencing for clinical care. JAMA Intern Med 2014;174(2):275-280. [doi: 10.1001/jamainternmed.2013.12048] [Medline: 24217348]
21. Reis-Filho JS. Next-generation sequencing. Breast Cancer Res 2009;Suppl 3(Suppl 3):S12. [doi: 10.1186/bcr2431] [Medline: 20030863]

XSL•FO
RenderX

22. Sow D, Fall B, Ndiaye M, Ba BS, Sylla K, Tine R, et al. Usefulness of MALDI-TOF mass spectrometry for routine identification of candida species in a Resource-Poor setting. Mycopathologia 2015;180(3-4):173-179. [doi: 10.1007/s11046-015-9905-2] [Medline: 26016846]
23. Duncan MW, Nedelkov D, Walsh R, Hattan SJ. Applications of MALDI mass spectrometry in clinical chemistry. Clin Chem 2016;62(1):134-143. [doi: 10.1373/clinchem.2015.239491] [Medline: 26585930]
24. Chang KW, Cheng HW, Shiue J, Wang JK, Wang YL, Huang NT. Antibiotic susceptibility test with Surface-Enhanced raman scattering in a microfluidic system. Anal Chem 2019;91(17):10988-10995. [doi: 10.1021/acs.analchem.9b01027] [Medline: 31387345]
25. Martineau F, Picard FJ, Grenier L, Roy PH, Ouellette M, Bergeron MG. Multiplex PCR assays for the detection of clinically relevant antibiotic resistance genes in staphylococci isolated from patients infected after cardiac surgery. the ESPRIT trial. J Antimicrob Chemother 2000;46(4):527-534. [doi: 10.1093/jac/46.4.527] [Medline: 11020248]
26. Li B, Yan T. Next generation sequencing reveals limitation of qPCR methods in quantifying emerging antibiotic resistance genes (ARGs) in the environment. Appl Microbiol Biotechnol 2021;105(7):2925-2936. [doi: 10.1007/s00253-021-11202-4] [Medline: 33738553]
27. El-Badawy MF, Tawakol WM, El-Far SW, Maghrabi IA, Al-Ghamdi SA, Mansy MS, et al. Molecular identification of aminoglycoside-Modifying enzymes and plasmid-Mediated quinolone resistance genes among clinical isolates recovered from Egyptian patients. Int J Microbiol 2017;2017:8050432 [FREE Full text] [doi: 10.1155/2017/8050432] [Medline: 28638412]
28. Swedan S, Alabdallah EA, Ababneh Q. Resistance to aminoglycoside and quinolone drugs among clinical isolates from northern Jordan. Heliyon 2024;10(1):e23368 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e23368] [Medline: 38163217]
29. Yu J, Lin YT, Chen WC, Tseng KH, Lin HH, Tien N, et al. Direct prediction of carbapenem-resistant, carbapenemase-producing, and colistin-resistant klebsiella pneumoniae isolates from routine MALDI-TOF mass spectra using machine learning and outcome evaluation. Int J Antimicrob Agents 2023;61(6):106799. [doi: 10.1016/j.ijantimicag.2023.106799] [Medline: 37004755]
30. Zhang YM, Tsao MF, Chang CY, Lin KT, Keller JJ, Lin HC. Rapid identification of carbapenem-resistant klebsiella pneumoniae based on matrix-assisted laser desorption ionization time-of-flight mass spectrometry and an artificial neural network model. J Biomed Sci 2023;30(1):25 [FREE Full text] [doi: 10.1186/s12929-023-00918-2] [Medline: 37069555]

Abbreviations

AI-CDSS: artificial intelligence-clinical decision support system
AST: antibiotic sensitivity testing
AUC: area under the receiver operating characteristic curve
CIP: ciprofloxacin
KP: Klebsiella pneumoniae
LEV: levofloxacin
LGBM: light gradient boosting machine
LMIC: low- and middle-income country
m/z: mass-to-charge ratio
MALDI-TOF MS: matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MDR-GNB: multidrug-resistant gram-negative bacteria
MDR-KP: multidrug-resistant Klebsiella pneumoniae
ML: machine learning
NGS: next-generation sequencing
NPV: negative predictive value
PCR: polymerase chain reaction
PPV: positive predictive value
ROC: receiver operating characteristic
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Schwartz; submitted 04.03.24; peer-reviewed by M Jani, E Lin; comments to author 14.04.24; revised version received 06.05.24; accepted 17.09.24; published 07.11.24.

Please cite as:
Jian MJ, Lin TH, Chung HY, Chang CK, Perng CL, Chang FY, Shang HS
Pioneering Klebsiella Pneumoniae Antibiotic Resistance Prediction With Artificial Intelligence-Clinical Decision Support System–Enhanced Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: Retrospective Study
J Med Internet Res 2024;26:e58039
URL: https://www.jmir.org/2024/1/e58039
doi:10.2196/58039
PMID:

©Ming-Jr Jian, Tai-Han Lin, Hsing-Yi Chung, Chih-Kai Chang, Cherng-Lih Perng, Feng-Yee Chang, Hung-Sheng Shang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

EDAI Framework for Integrating Equity, Diversity, and Inclusion Throughout the Lifecycle of AI to Improve Health and Oral Health Care: Qualitative Study

Samira Abbasgholizadeh Rahimi[1,2,3,4*], PhD; Richa Shrivastava[5*], PhD; Anita Brown-Johnson[1,2], MDCM; Pascale Caidor[6], PhD; Claire Davies[7], PhD; Amal Idrissi Janati[2,8], DMD; Pascaline Kengne Talla[2], PhD; Sreenath Madathil[2], PhD; Bettina M Willie[2,9], PhD; Elham Emami[2*], PhD

[1]Department of Family Medicine, McGill University, Montreal, QC, Canada
[2]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada
[3]Mila-Quebec AI Institute, Montreal, QC, Canada
[4]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada
[5]Postgraduate Institute of Medical Education and Research, Chandigarh, India
[6]Department of Communication, Université de Montréal, Montreal, QC, Canada
[7]Department of Mechanical and Materials Engineering, Queen's University, Kingston, ON, Canada
[8]The Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[9]Research Centre, Shriners Hospital for Children-Canada, Montreal, QC, Canada
[*]these authors contributed equally

Corresponding Author:
Samira Abbasgholizadeh Rahimi, PhD
Department of Family Medicine
McGill University
5858 Chemin de la Côte-des-Neiges
Montreal, QC, H3S 1Z1
Canada
Phone: 1 514 399 9218
Email: samira.rahimi@mcgill.ca

Abstract

Background: Recent studies have identified significant gaps in equity, diversity, and inclusion (EDI) considerations within the lifecycle of artificial intelligence (AI), spanning from data collection and problem definition to implementation stages. Despite the recognized need for integrating EDI principles, there is currently no existing guideline or framework to support this integration in the AI lifecycle.

Objective: This study aimed to address this gap by identifying EDI principles and indicators to be integrated into the AI lifecycle. The goal was to develop a comprehensive guiding framework to guide the development and implementation of future AI systems.

Methods: This study was conducted in 3 phases. In phase 1, a comprehensive systematic scoping review explored how EDI principles have been integrated into AI in health and oral health care settings. In phase 2, a multidisciplinary team was established, and two 2-day, in-person international workshops with over 60 representatives from diverse backgrounds, expertise, and communities were conducted. The workshops included plenary presentations, round table discussions, and focused group discussions. In phase 3, based on the workshops' insights, the EDAI framework was developed and refined through iterative feedback from participants. The results of the initial systematic scoping review have been published separately, and this paper focuses on subsequent phases of the project, which is related to framework development.

Results: In this study, we developed the EDAI framework, a comprehensive guideline that integrates EDI principles and indicators throughout the entire AI lifecycle. This framework addresses existing gaps at various stages, from data collection to implementation, and focuses on individual, organizational, and systemic levels. Additionally, we identified both the facilitators and barriers to integrating EDI within the AI lifecycle in health and oral health care.

XSL•FO
RenderX

Conclusions: The developed EDAI framework provides a comprehensive, actionable guideline for integrating EDI principles into AI development and deployment. By facilitating the systematic incorporation of these principles, the framework supports the creation and implementation of AI systems that are not only technologically advanced but also sensitive to EDI principles.

(J Med Internet Res 2024;26:e63356) doi:10.2196/63356

KEYWORDS

equity, diversity, and inclusion; EDI; health care; oral health care; machine learning; artificial intelligence; AI

Introduction

Artificial intelligence (AI) is a branch of engineering and computer science that creates systems capable of performing tasks that typically require human intelligence such as visual perception, speech recognition, and language translation, while also analyzing complex datasets to identify patterns, make predictions, and solve problems across various domains [1-3]. AI technologies are increasingly being applied in all sectors of our society including health care systems [4-8]. These systems could be helpful in improving the effectiveness of health care delivery and facilitate achieving quadruple aims of care, that is, improved patient experience, better outcomes, lower costs, and improved clinical experience [9]. The introduction of AI in medicine and dentistry has shown promising potential [8,10,11] in screening, detection, and treatment of diseases [12,13]. AI enhances the use of electronic health records, supports clinical decision-making, and enables the development of sophisticated robots and devices to enhance care delivery and alleviate workload in health care systems [8,14].

The deployment of AI across various sectors has underscored the importance of ethical governance and the integration of equity, diversity, and inclusion (EDI) principles. Drawing on prior research from various industries, it is evident that AI technologies can both mitigate and exacerbate existing social biases, depending on their design and implementation [6,15,16]. This dual potential of AI, as highlighted in studies like Obermeyer et al [17], emphasizes the necessity for incorporating EDI principles from the onset of development to ensure technologies promote inclusivity and prevent social inequality. Moreover, research on algorithmic accountability underscores the need for transparency and explainability in AI systems, requiring designs that provide traceability of decisions and clear data use to comply with international regulations [18]. Additionally, insights into data privacy stress the importance of robust data protection practices to prevent technology misuse and protect personal information, ensuring AI applications uphold individual rights and freedoms [19]. These principles, drawn from a broad array of previous research, constitute a comprehensive framework essential for the ethical deployment of AI technologies across all domains.

Serious concerns have been expressed about how AI might undermine the basis and essence of humanity [15]. Hypothetical syllogisms, which are arguments containing at least 1 hypothetical proposition as a premise, suggest a dystopian future with unfairness enacted by AI [16]. Recent studies also revealed that EDI concepts are not adequately implemented in the building of AI systems for health and oral health care [8,20-22]. For example, the use of AI in health care raises concerns about the privacy and security of sensitive medical information, particularly for patients who are vulnerable and may have a history of discrimination and mistrust in the health care system [23]. EDI is a conscious effort to include a diverse range of social identities and perspectives, including those typically marginalized in decision-making on social and health issues that impact their lives.

By definition, equity strives to treat everyone fairly with the recognition that each person has different circumstances and allocates the exact resources and opportunities needed to reach an equal outcome [20]. Diversity means including a range of people and perspectives such as those from different racial and ethnic backgrounds, genders, sexual orientations, ages, abilities, socioeconomic statuses, religions, and cultural backgrounds. Inclusion means the act of social inclusion of groups that are normally marginalized as well as the social values they subscribe to [20]. Paradoxically, some evidence suggests that AI can lead to fairness and efficiency by reducing the likelihood of human bias, but our decisions will determine how the future unfolds [16].

In this study, we argue that ethical and fair AI can be achieved only by integrating the EDI concept in the development and application of AI and machine learning models [15,16,20,24]. However, it is critically important that the building of AI systems be participatory in nature and includes diverse experts and stakeholders like system developers, health professionals, researchers, and policy makers, among other end users [2]. Unfortunately, this research shows that studies on the application of AI in health care have rarely included relevant stakeholders in the design, development, and implementation of AI systems [8,25,26].

The distinction between AI ethics and a framework concerning AI and EDI is important, emphasizing different facets of responsible AI development. AI ethics broadly encompasses moral values, guiding the development and application of AI technologies in a manner that respects human dignity and societal norms. It encompasses principles like transparency, accountability, and fairness, aiming to prevent AI systems from perpetuating harm or injustice [27]. In contrast, the framework focuses specifically on integrating these considerations as well as EDI principles throughout the AI system's lifecycle—from design and development through deployment and monitoring. This lifecycle approach ensures that EDI principles are central, actively promoting inclusivity and mitigating biases [26]. This clarifies how AI ethics provides the foundational values, while the framework offers a structured approach to operationalize these values concretely and continuously across all stages of AI implementation.

XSL•FO
RenderX

Although studies have identified limitations, to the best of our knowledge, no studies to date have explored the feasibility and challenges of incorporating EDI variables, concepts, practices, and outcomes within AI technologies in the field of health and oral health care. There is a need to explore how EDI principles can be integrated into the lifecycle of AI technologies and gain an understanding of the relevant stakeholders' perspectives on societal challenges and implementation of EDI in AI technologies for health care, oral health care, and related research. This study aimed to co-design the EDAI framework for integrating EDI throughout the lifecycle of AI in partnership with representatives of socio- or ethnocultural communities and end users, as well as with an international or intersectoral or interprofessional team of policy makers, researchers, clinicians, and industry representatives.

Methods

Study Design

Our research project began with a gap analysis and a comprehensive systematic scoping review, which explored the integration of EDI principles into the life cycle of AI within health and oral health care settings. The results of this initial systematic scoping review have been published separately [22]. In this paper, we focus on the other phases of the project, which is directly related to the framework development stage. This study followed a qualitative exploratory study design with a constructive grounded theory methodology approach and was presented according to the SRQR (Standards for Reporting Qualitative Research) guideline (Multimedia Appendix 1). Grounded theory enables "the generation of theories of process, sequence, and change pertaining to organizations, positions, and social interaction" [28]. This theory is a systematic methodology in which the researcher is not bound to select an existing theoretical framework. Instead, the findings may serve as the foundation for the development of new theories [29,30]. Therefore, for developing conceptual frameworks from diverse texts like the one presented in this study, the grounded theory technique is appropriate and insightful [31]. Constructivist grounded theory is a type of grounded theory that is based on constructivism [32]. It is centered on the coconstruction of experiences and meanings of the area of inquiry with the participants [32].

Ethical Considerations

This study received ethics approval from the McGill University Institutional Ethics Committee (A03-E10-23A). Informed consent was obtained from all participants. Participants were also informed of their right to opt out at any time. Data collected in this study were deidentified; personal identifiers such as names were not recorded. Participants were not compensated monetarily; however, food was provided during the workshop.

Study Phases

This research was conducted in 3 phases, and each phase is explained below in detail.

Phase 1: Gap Analysis

A comprehensive scoping review has been performed to explore what and how EDI principles and practices have been integrated into the conception, design, development, and implementation of AI studies conducted in health and oral health care settings [22]. The results of the scoping review indicate the lack of a comprehensive framework in health care, in which the principles of EDI are systematically deliberated and applied to the AI lifecycle [22]. Therefore, an expert group workshop was planned to develop a comprehensive framework, that is, EDAI, for health and oral health care settings. The review results informed the EDAI framework's development.

Phase 2: Expert Group Workshops

Establishing a Multidisciplinary Team Involving Stakeholders and Collaborators

The framework development process was structured around 2 strategic workshops. The initial workshop in Montreal, Canada, centered on collaborative creation, bringing together diverse stakeholders to codevelop the framework's foundational elements. Building on this groundwork, the second workshop in Geneva, Switzerland, was dedicated to refining and validating the preliminary version. This 2-phase approach ensured a comprehensive, iterative process that incorporated broad expertise while allowing for critical assessment and enhancement of the framework's applicability. The purposeful sampling with maximum variation was used to enroll members from an intersectoral or multidisciplinary background (engineering, dentistry, medicine, AI systems, ethics in AI, communication, and social sciences) including researchers, clinicians, knowledge users, patient representatives, community partners, policy makers, public health agencies, and industry representatives. The selection of stakeholders started 1 year before the first workshop to build an expert research group. The team members were consciously selected to represent excellence in research activities or leadership and diversity in socio- or ethnocultural and geographical contexts. The team's composition is in line with the participatory approach and the concept of codevelopment, which are at the core of this project. Most of the team members had a long-established working relationship and had successfully collaborated on various projects previously.

Data Collection

Overview

The study follows a qualitative research approach and includes 2-day, by-invitation, in-person workshops with focus group discussions to co-design the EDAI framework [21]. According to Ahmed and Mohd Asraf [33] and Lain [34], workshops are essential for prolonged engagement and are regarded as being one of the primary ways of establishing the credibility of the results of a qualitative study and upholding the trust between the researchers and participants. The workshops continued for 2 days and involved representatives from diverse backgrounds including government and nongovernment organizations, visible minority community leaders including Black communities, representatives from the World Health Organization, and the participation of health and oral health care providers, academia, patients, industry partners, and policy makers [21].

XSL•FO
RenderX

Workshop Engagement Exercises

The workshops included plenary presentations, round table working groups, and focused group discussions as a research methodology [21,35]. At the preworkshop phase, to facilitate the exchange of ideas, research team members prepared a draft document relevant to the area of research and distributed it to meeting participants 2-3 weeks before the workshops [21].

- Workshop presentations: Individuals were selected from the attendees to provide presentations on their previous experiences in EDI to enhance the discussions. A semistructured focus group discussion guide was developed before the workshop based on a review of the literature and previous work of team members [21]. A panel presentation linked the presentations to focus groups and included expert opinions on these questions: (1) What are the visions and values that guide the integration of EDI throughout the lifecycle of AI? (2) What are the facilitators and barriers to the integration of EDI throughout the lifecycle of AI within health and oral health care? (3) What should be the indicators of EDI in health and oral health care data and how can we measure the EDI variables throughout the lifecycle of AI? and (4) How a conceptual framework should be framed and what are the main elements?
- Round table working groups and focused group discussions: The presentation session was followed by focused group discussion groups based on qualitative methodology and inductive reasoning [21]. The discussion groups answered the same questions that had been asked of the panel participants.

Each focus group had a moderator and notetaker with 5-6 purposefully selected participants with maximum variation and lasted for 45-60 minutes. Afterward, there was a large group discussion among the workshop attendees to further explore the subject and reach a consensus together. All proceedings notes were written by the notetakers. All the notes and memos were compiled for data analysis.

Workshop Conclusion

The workshops were concluded with a summary of the workshop by restating the main points and identifying components for preliminary data analysis.

Data Analysis

Data analysis was followed immediately after each workshop. An inductive thematic analysis was done manually (SAR, EE, and RS). As per Braun and Clarke [36], the analysis began with the generation and collation of initial codes, which will then be sorted into multiple potential themes. These potential themes were then combined to form overarching key themes. During this process, working groups were created across team members to ensure data triangulation.

We followed the 8 phases of “conceptual framework analysis” within the grounded theory approach to systematically identify and interpret concepts, categories, and their relationships as explained by Jabareen [31]. Phase 1 is the mapping of data sources consisting of a literature review and workshop with focus group discussions [31]. Since EDI in AI is an unexplored concept, firsthand knowledge and experience using primary data collection was deemed important [37]. In phase 2, categorizing the selected data, and phase 3, identifying and naming concepts, the first set of concepts was inductively developed using line-by-line coding of focused group discussions [31,38]. After the identification of the patterns within the concepts, focused coding was initiated to establish initial categories [31,39]. Phase 4 is deconstructing and categorizing concepts, and phase 5 is integrating concepts and refining the coding by iterative assessment [31,37]. It included deconstructing and organizing the concepts and then integrating similar concepts together [31]. Constant comparative technique at each level of coding was used to assess similarities and differences for refining the concepts [32]. Phase 6, synthesis and resynthesis, included organizing the EDAI conceptual framework by inductively sorting the categories and grouping them under 3 main dimensions [31,37]. Phase 7, validation, and phase 8, rethinking the conceptual framework, were performed via triangulation, reflexivity, and participant feedback during the second workshop and via emails and digital meetings during data analysis [31,37].

Phase 3: Framework Development and Stakeholder Reviews

Based on the workshops’ key impressions and group discussions, the EDAI framework was developed and discussed with all participants electronically and in the second workshop. Their feedback was sought and iteratively incorporated to develop the final framework. Table 1 illustrates the key steps and the details of the process across the different phases of the study.

XSL•FO
RenderX

Table 1. Summary of study phases and steps.

Phase and step	Details
Phase 1: gap analysis	
1. Scoping review	A comprehensive scoping review was performed to explore how EDI[a] principles are integrated into AI[b] studies in health and oral health care settings.
Phase 2: expert group workshops	
2. Establishing a multidisciplinary team	A multidisciplinary team was built through purposeful sampling with maximum variation, including stakeholders from diverse backgrounds (eg, engineering, dentistry, medicine, AI, ethics, and social sciences). The selection began 1 year before the first workshop. The team composition emphasized codevelopment, participatory approaches, and diversity in socio- or ethnocultural and geographical contexts.
3. Data collection	Both 2-day, in-person workshops involved focus group discussions to co-design the EDAI framework. Representatives from diverse backgrounds, including government, nongovernment organizations, minority communities, academia, and industry, participated. The workshop included plenary presentations, round table working groups, and focused group discussions, with preworkshop preparations such as draft documents distributed to participants. Focus groups addressed key questions on integrating EDI in the AI lifecycle, identifying facilitators and barriers, and defining EDI indicators and conceptual framework elements. Discussions were moderated, and notetakers documented all proceedings for data analysis.
4. Workshop conclusions	The workshops concluded with a summary and identification of components for preliminary data analysis.
5. Data analysis	Inductive thematic analysis was conducted manually. Initial codes were generated and sorted into potential themes, which were then refined and organized into key themes through collaborative efforts. The 8 phases of "conceptual framework analysis" (eg, mapping data sources, categorizing data, identifying concepts, and integrating concepts) were used. Constant comparative techniques were used to refine the coding, and the EDAI conceptual framework was organized into 3 main dimensions.
6. Validation	Validation and refinement of the conceptual framework were achieved through triangulation, reflexivity, and participant feedback via emails and digital meetings as well as during the second workshop.
Phase 3: framework development and stakeholder reviews	
7. Framework development	Based on workshop findings, the EDAI framework was developed and discussed with participants electronically. Feedback was iteratively incorporated to finalize the framework.

[a]EDI: equity, diversity, and inclusion.

[b]AI: artificial intelligence.

Results

Overview

We identified the knowledge gaps through a past systematic scoping review, the results of which were published in another paper (phase 1) [22]. The participants involved in phase 2 and phase 3 of the study included over 60 representatives from diverse backgrounds and expertise from government and nongovernment organizations, visible minority community leaders including Black communities, public health agencies, health and oral health care providers, AI and health services researchers, patients, industry partners, and policy makers. This diversity ensured balanced representation not only in terms of sex, gender, and expertise but also included other aspects such as race, ethnicity, socioeconomic background, age, and disability status.

We also included international participants and people from minority groups and marginalized populations. By actively recruiting from varied demographics, we aimed to address historical disparities and support equitable participation, thereby enriching the quality and impact of our research findings (phase 2 and phase 3). Five themes emerged from the data collected in phases 2 and 3. Table 2 illustrates the coding data structure of the collected data. This table illustrates initial, focused, and theoretical codes to classify the factors affecting EDI throughout the lifecycle of AI at micro (individual), meso (organization), and macro (system) levels to reach the final outcome of quadruple aims of health care by ensuring equitable, diversified, and inclusive care.

XSL•FO
RenderX

Table 2. Coding data structure.

Initial codes	Focused codes	Theoretical codes	Core category	Final outcome
• Awareness about AIa and EDIb • Inclusion of diverse perspectives and voices • Addressing cultural sensitivities while respecting individual identities • Incorporating principles of social justice and decolonization	• Enhancing equitable access to health and oral health care • Incorporating diverse perspective • Inclusion of cultural safety and social justice	• Role of EDI and AI in health and oral health care settings	Based on theoretical codes, classifying the factors affecting EDI throughout the lifecycle of AI (planning, data collection, designing, training, deployment, and decision-making) at micro (individual), meso (organization), and macro (system) levels	To achieve quadruple aims of health care by ensuring equitable, diversified, and inclusive care
• Transparent use of data • Trust, confidentiality, and privacy • Political will	• Responsible and ethical use of AI • Mutual respect among stakeholders • Supportive policy and regulation	• Vision and values that will guide the integration of EDI in the life cycle of AI	Based on theoretical codes, classifying the factors affecting EDI throughout the lifecycle of AI (planning, data collection, designing, training, deployment, and decision-making) at micro (individual), meso (organization), and macro (system) levels	To achieve quadruple aims of health care by ensuring equitable, diversified, and inclusive care
• Awareness about AI and EDI • Access to health and oral health care • Demographic factors	• Factors affecting at the levels of patients, health care providers, AI developers, policy makers, and stakeholders	• The facilitators and barriers to the integration of EDI • Indicators of EDI in health and oral health care data	Based on theoretical codes, classifying the factors affecting EDI throughout the lifecycle of AI (planning, data collection, designing, training, deployment, and decision-making) at micro (individual), meso (organization), and macro (system) levels	To achieve quadruple aims of health care by ensuring equitable, diversified, and inclusive care

[a]EDI: equity, diversity, and inclusion.

[b]AI: artificial intelligence.

The current iteration of the EDAI framework is recognized as a preliminary version, designed to set the foundational elements necessary for integrating EDI principles within the AI lifecycle. Given the iterative nature of grounded theory, a second iteration of framework development is done during the second workshop. This session aimed to engage a broader sample and incorporate extensive data analysis and stakeholder input. The objective was to further refine and enhance the framework, thus improving its robustness and applicability in diverse health care settings.

Theme 1: The Role of EDI and AI in Health and Oral Health Care Settings

Participants proposed leveraging AI to enhance access to health and oral health care, particularly in remote areas, by improving referral pathways. They emphasized that AI could address obstacles like cost and distance by offering alternative options such as telehealth. The discussions also highlighted AI's potential in studying human behavior within health care and predicting health outcomes for diverse populations and the diverse use of AI in decision-making. They recommended organizing education, awareness, and training workshops to improve AI literacy among health care providers and the public, including marginalized communities. Furthermore, participants stressed the significance of incorporating EDI in patient data collection and analysis to reduce health and oral health disparities.

Theme 2: Vision and Values That Will Guide the Integration of EDI in the Life Cycle of AI

During the discussions, participants stressed the importance of addressing biases and discrimination in AI algorithms and ensuring fair representation in the AI algorithms including the decision-making process. They emphasized the need to achieve this by including diverse perspectives and voices, empowering marginalized communities, and addressing cultural sensitivities while respecting individual identities. Additionally, the participants highlighted the significance of incorporating principles of social justice and decolonization at various stages of the AI life cycle, including problem definition, data collection, designing, training, deployment, and decision-making. Responsible and ethical use of AI was a key aspect that participants stressed throughout the conversations.

Theme 3: The Facilitators and Barriers to the Integration of EDI Throughout the Lifecycle of AI Within Health and Oral Health Care

Table 3 depicts the facilitators and barriers to integrating the EDI within AI that were discussed within the workshop at the levels of patients, health care providers, AI developers, policy makers, and stakeholders. Participants stressed the importance of establishing an EDI committee for each stage of the AI lifecycle with a mandate to define fairness from an EDI perspective and ensure context-specific implementation.

XSL•FO
RenderX

Table 3. Facilitators and barriers to integration of EDIa in AIb.

Level and facilitators	Barriers
At patients' level	
• Enhancing education and awareness regarding AI and EDI to alleviate patient resistance and fears. • Ensuring clear and transparent communication with patients about the purpose and process of data collection for AI model development. • Providing access and training of AI to patients.	• Resistance to AI integration can arise due to fear and intimidation stemming from complex technological jargon, concerns about potential job loss, and negative portrayals in popular media. Marginalized communities may exhibit heightened apprehension toward AI. • Limited technology training and lack of familiarity with AI applications, often exacerbated by low socioeconomic conditions. • Generational disconnect among users and technology knowledge gap. • Privacy concerns regarding patient consent.
At health care providers' level	
• Implementation of training programs and knowledge enhancement initiatives for health care providers. • Creation of ethical guidelines and policy frameworks concerning the collection and use of patient data.	• Lack of EDI and AI literacy and understanding among health care providers.
At AI developers' level	
• Fostering the inclusion of diverse perspectives in AI development and implementation and involving affected communities during the conceptualization stage—ensuring AI solutions address their specific needs and concerns. • Identifying inclusive data and adopting participatory, community-oriented approaches in data collection—ensuring inclusivity and representation. • Continuous feedback and evaluation of AI algorithms to correct biases and improve reliability. • Adopting a transparent approach in the AI development stage to build trust and facilitate EDI integration. • Asking the right questions focused on the problem rather than the technology can lead to more relevant and impactful AI solutions.	• Limited knowledge and research on dimensions of EDI result in challenges in defining and categorizing diverse groups. • Bias and discrimination embedded in human data and missing data can lead to biased AI predictions. • Inadequate representation of diverse groups in the development process can lead to biased AI solutions. • Misconduct in responsible research and data misuse. • Limited resources for developing AI tools with integrated EDI can hinder progress.
At policy makers' level	
• Having political will and financial support from policy makers can facilitate the identification of AI needs and the development of inclusive and equitable AI. • Supportive policy and regulation frameworks for responsible and fair AI development and implementation.	• Lack of policy frameworks around integration of EDI through the lifecycle of AI. • The lack of regulatory authority and governance can create challenges in assessing AI integration from an EDI perspective, potentially leading to ethical and fairness concerns. • Politicians exerting pressure about how society is using or will use AI, often for political gain. • The current education system may limit diversity in AI developer groups.
At all stakeholders' level	
• Connecting end users' needs with stakeholders to facilitate EDI integration. • AI literacy and educating stakeholders at all levels, including users, providers, and the community, about AI and its implications can build a better understanding and acceptance of AI solutions.	• The lack of common vocabulary among different professionals, such as researchers, clinicians, AI developers, and consumers, can create obstacles in integrating EDI within the lifecycle of AI. • AI's potential to generate wealth may exacerbate societal disparities, with power dynamics among stakeholders influencing AI and EDI control. • Privacy concerns related to data ownership.

[a]EDI: equity, diversity, and inclusion.

[b]AI: artificial intelligence.

Theme 4: The Indicators of EDI in Health and Oral Health Care Data and Their Measurement

We identified the EDI indicators through a past scoping review, the results of which were published in another paper [22]. During the workshop, participants categorized the EDI indicators in health and oral health care into 3 levels: individual (micro level), community (meso level), and society (macro level). Table 4 details the identified EDI indicators under categorized levels.

XSL•FO
RenderX

Table 4. Refinement of the conceptual EDAI framework.

Levels and indicators	Details
Micro level (individual level)	
Sociodemographic	Age, gender, residence, education level, occupation, income, marital status, socioeconomic status, race or ethnicity, religion, literacy, health literacy (including for artificial intelligence), cultural or traditional practices, immigration or acculturation, and family or peer or community-based influences
Health and functional factors	Physical health, mental health, mobility, and cognitive function
Psychobiological factors	Health beliefs, attitude, emotions, self-esteem, perceived health status, and genetics
Behavioral	Lifestyle factors and habits (eg, tobacco and alcohol)
Meso level (organization level)	
Health care system	Existing health system at the place of inquiry
Education system	Exiting education system at the place of inquiry
Community infrastructure	Physical infrastructure and rurality or remoteness
Network and community support	Teamwork and shared ownership
Macro level (system level)	
Colonialism	Political, socioeconomic, and cultural domination by foreign power
Intersectionality	Racism, intersectional discrimination, and social exclusion
Governance	Decision-making, accountability, and leadership (eg, health care structure and funding agency)
Social belief	Social values and beliefs

To measure these EDI indicators and variables throughout the lifecycle of AI, participants stated that various tools and methods can be used including quantitative methodological analysis, surveys, interviews, anonymous self-reported questionnaires, image-based data collection, tracking location, and soliciting user feedback. Additionally, qualitative evaluation methods, such as reflective texts or live feedback from users, were mentioned as another source for providing valuable insights into EDI and the user experience with AI-enabled technologies and tools.

Theme 5: Framing the EDAI Framework

Integration of EDI in AI follows a multifactorial approach. This approach is shown in Figure 1. AI-based health and oral health practices should underpin all relevant factors at micro, meso, and macro levels. Therefore, the final framework is structured over these 3 levels: micro (individual level), meso (organization level), and macro levels (system level) with the indicators underpinning the whole lifecycle of AI. The micro level comprises smaller entities like individuals or families and their interactions. The meso level involves groups or communities or organizations and their interactions. The macro level encompasses broader systems such as policies or government structures at institutional, national, or international levels and their interactions. These 3 levels are interconnected and interrelated and significantly influence each other.

At the micro level, EDI encourages diverse representation and involvement of people of all sociodemographic factors (ie, age, gender, residence, socioeconomic status, and marital status) who may have diverse capabilities, skills, expertise, and values. It inspires people from various sociocultural factors such as race or ethnicity, religion, health literacy, cultural or traditional practices, immigration or acculturation, and family or peer or community-based influences. Moreover, respecting cultural diversities can widen the scope of AI-based health care. Integration of EDI within AI can positively impact an individual's psychobiological factors including their health beliefs, attitude, emotions, self-esteem, and perceived health status. Furthermore, EDI-informed approaches are helpful in inculcating healthy lifestyle behaviors and habits by fostering an inclusive environment that respects and values diverse cultural perspectives.

At the meso level, health systems comprising both public and private health organizations and institutes should foster the integration of EDI within the lifecycle of AI in health and oral health care by emphasizing on the provision of a culturally safe environment for patients and care providers and ensuring the presence of proficient professionals. The education system should ensure that all in the school community—students, their families, teachers, staff, and administrators—feel valued and heard. Community infrastructure needs to be considered while working on AI-EDI to have better representation of all community members irrespective of physical location, rurality or remoteness, representativity, and accessibility. Furthermore, networks and community support such as social networks, community groups, and peer support should also focus on diverse representation with teamwork and shared ownership.

XSL•FO
RenderX

Figure 1. The proposed EDAI framework. AI: artificial intelligence.

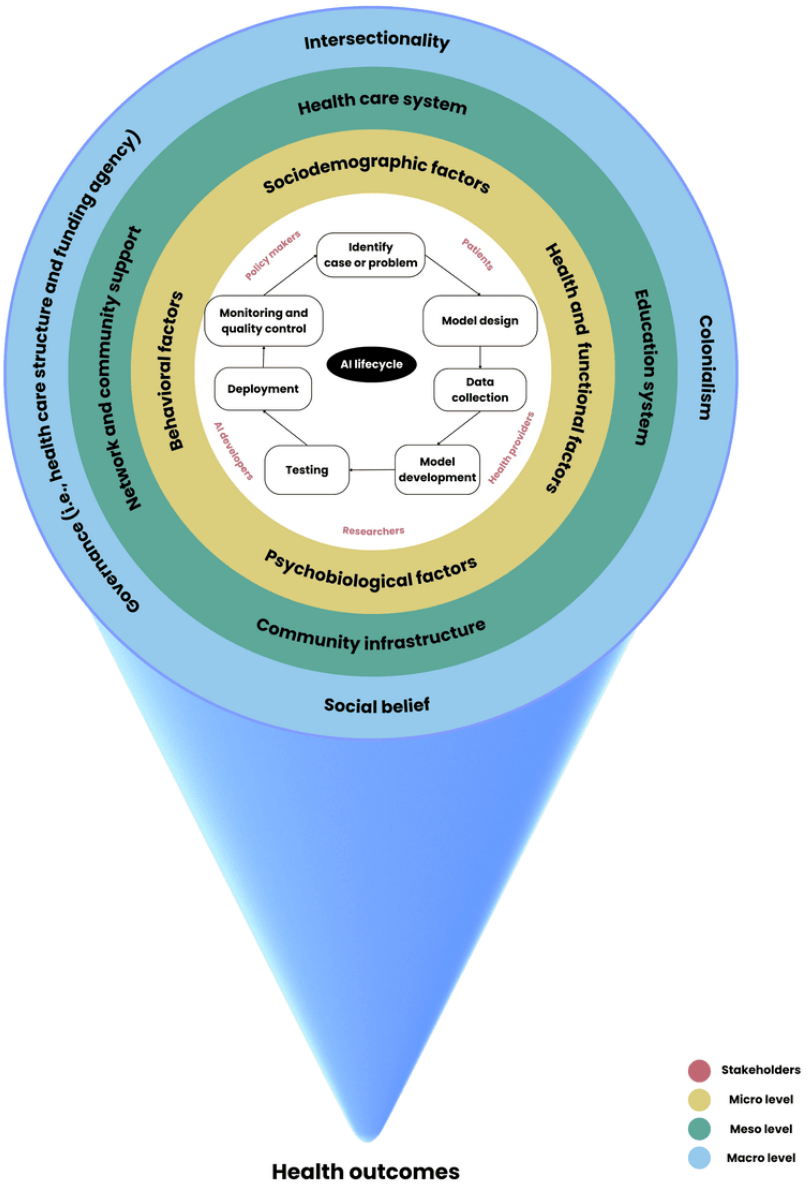

At the macro level, focusing on intersectionality can aid in understanding the complex interactions of various factors at micro and meso levels, thereby striving to achieve social justice. Integration of EDI within the lifecycle of AI should emphasize decolonization by minimizing the impact of settler colonialism, racism, and discrimination and acknowledging the historical injustices and systemic inequalities that have impacted marginalized communities. Furthermore, social beliefs play a significant role in shaping attitudes and actions toward EDI since positive social beliefs can foster environments where individuals from all backgrounds feel valued and empowered. In the context of EDI in the AI lifecycle, governance plays a critical role in setting policies, implementing practices, and enforcing regulations that promote EDI. Finally, in our workshops, participants also emphasized the importance of responsible AI including reliability and safety, privacy and security, transparency, and accountability [40], which should be considered during development and implementation.

Discussion

Principal Findings

Our findings reveal significant gaps in EDI considerations at various stages of the AI lifecycle, from problem definition and data collection to implementation. In response, this study developed a guiding framework—the co-designed EDAI framework—that integrates EDI principles and indicators comprehensively throughout the AI lifecycle. The field of AI ethics, which emerged alongside the evolution of AI technologies, has grown in importance, as AI systems have become more advanced. Focused initially on issues like algorithmic bias, transparency, privacy, and fairness, AI ethics seeks to align AI technologies with moral principles and values.

XSL•FO
RenderX

However, it is distinct from EDI in AI, which extends beyond ethical considerations to emphasize the integration of EDI throughout the AI lifecycle. Consideration of EDI not only aims to prevent biases but also ensures diverse representation and targets social equity and inclusivity.

Despite the advantages of AI, there are some concerns that underscore the necessity of consciously integrating EDI principles into the lifecycle of AI systems to ensure they are socially responsible and equitable. These concerns include discrimination, unfair treatment, and risks of developing or intensifying health-related inequalities [41]. Recent systematic reviews of EDI in AI found a lack of EDI considerations in the dataset and AI-based academic literature and a lack of EDI principles and indicators to be integrated into the AI lifecycle [22,41]. The studies also highlighted the need to integrate EDI and racial justice principles and practices in AI; however, there is no guideline or framework to promote this integration. This study is addressing this knowledge gap by identifying EDI principles and indicators to be integrated into the AI lifecycle and developing a guiding framework to guide the next-generation AI system development and implementation.

Our results highlight the importance of an EDI-sensitive approach in the life cycle of AI, reducing inequities and inequalities in health and oral health. The EDI concepts should be integrated not only during AI system development but also throughout its lifecycle, including AI planning, data collection, designing, training, deployment, and decision-making. Moreover, the study results emphasize the fair representation of all stakeholders including patients, health care providers, AI developers, and policy makers. Therefore, we developed an EDAI framework in this study encompassing several EDI indicators in health and oral health care data at the individual (micro), community (meso), and society (macro) levels underpinning the whole lifecycle of AI.

In other studies, the AI or Machine Learning Consortium to Advance Health Equity and the Researcher Diversity Ethics and Equity Workgroup has established ethics and equity principles. These principles focus on fostering trust with communities, intentionally designing and implementing AI systems, creating solutions, building capacity, and re-evaluating institutional mechanisms to reset the rules [42]. The group has also developed a glossary containing 12 terms, with particular emphasis on AI development in health equity research [42]. Furthermore, with the help of the Delphi method, Rokhshad et al [43] developed a checklist of 11 ethical principles specific to AI-based dentistry including equity and diversity as major principles. However, these studies are providing worthy groundwork by focusing on ethical principles and health equity, the current studies lack identifying and integrating EDI principles and indicators in the AI lifecycle. The proposed EDAI framework offers a structured approach to exploring EDI concepts and the factors influencing them throughout the AI lifecycle. It addresses the need to identify and integrate EDI principles and indicators within AI systems.

A recent scoping review conducted by Singhal et al [44] emphasized the crucial integration of fairness, accountability, transparency, and ethics across all phases of AI system development from conceptualization to deployment. The review also suggests using diverse datasets to ensure a more accurate reflection of the world's population [44]. Additionally, they advocate the active involvement of all stakeholders at every stage. Likewise, the EDAI framework outlined in this paper aims to address these recommendations, with particular attention to the factors influencing AI-driven health care technology across all stakeholder groups and throughout the lifecycle of AI system. In this line, Dankwa-Mullan et al [26] have also proposed a framework centered on integrating health equity and racial justice principles within each stage of the AI lifecycle to ensure ethical and moral behavior in AI system development. Similar to the proposed EDAI framework, various AI ethics frameworks have emphasized interdisciplinary stakeholders' collaboration as one of the foundations of ethical and responsible AI [45,46]. This multifaceted interdisciplinary stakeholders' collaboration can involve active participation from policy makers, health care providers, AI experts, and patients in AI-based ethical decision-making in health and oral health care [45,46].

A recent review of AI guidelines related to EDI by Cachat-Rosset and Klarsfeld [41] analyzed 46 published guidelines pinpointing 14 relevant principles and 18 practices. They reported fairness, justice, and nondiscrimination in a limited compliance approach within these guidelines. Their findings revealed a partial adherence to fairness, justice, and nondiscrimination within these guidelines, suggesting a constrained approach to compliance [41]. The authors recommended the need for holistic guidelines over the more technical ones [41]. Considering these recommendations, the EDAI framework endeavors to create a holistic structure that takes into account multiple factors at the individual, organizational, and societal levels.

EDI can be challenging to measure because definitions of variables, for example, gender may evolve, reflecting societal changes. Additionally, the measurement of EDI may vary depending on the specific context in which AI models are applied, making it challenging to establish a standardized set of indicators. Furthermore, fairness is context-dependent and relies on data collection methods, which can vary based on factors like self-identification of the data versus government documents. However, having a ready-to-use tool can help all the stakeholders assess the integration of EDI in health or oral health-based AI systems. Therefore, based on the findings of this study, the authors plan to create and validate an evaluation-based checklist and guidelines for the integration of EDAI in health and oral health care.

The developed framework can help AI developers, who are integral to the model, by increasing their awareness of EDI principles, thereby reducing inherent biases and ensuring that development teams are diverse and inclusive, leading to more equitable and effective AI solutions in health care. For instance, the EDAI framework could facilitate the development of AI-driven diagnostic tools that better account for demographic variations, ensuring that diagnoses are accurate across diverse populations. This not only improves patient outcomes but also addresses historical biases in medical research and treatment. Another example is health care management in which AI

XSL•FO
RenderX

systems designed and implemented under the EDAI framework can optimize resource allocation and patient scheduling to prioritize accessibility and minimize disparities in care delivery. Furthermore, this framework can guide the development of AI tools for patient education and engagement, ensuring that materials are culturally relevant and accessible to all, thereby improving health literacy and patient autonomy across diverse communities.

Strengths and Limitations

One of the strengths of this research work is its systematic approach and the intersectoral and multidisciplinary team. The team members were consciously selected to represent excellence in research activities and leadership and diversity in socio- or ethnocultural and geographical contexts. The team's composition is in line with the participatory approach and the concept of codevelopment, which are at the core of this project. Moreover, the EDAI framework also attempts to understand the linkages among various indicators at individual, organizational, and system levels. However, this study has a few limitations. First, while we were unable to record the workshop sessions due to logistical constraints, detailed notes were diligently taken by designated notetakers to ensure the accuracy and richness of the data. Second, despite our efforts, we were not able to engage First Nations and Aboriginal people in this initial phase. However, we did include a diverse population from different communities, ensuring a wide range of perspectives, and we are committed to incorporating First Nations and Aboriginal voices in future iterations of the framework.

Conclusions

Recent studies have found that there are significant gaps in EDI considerations within the AI lifecycle, from the problem definition and data collection stage to implementation. However, there is currently no existing guideline or framework to support this integration. This study aimed to address this gap by identifying EDI principles and indicators to be integrated into the AI lifecycle and developing a guiding framework to steer the development and implementation of future AI systems. The co-designed EDAI framework is comprehensive, encompassing a detailed compilation of EDI indicators across the AI lifecycle, spanning from individual to organizational and system levels for all stakeholders. These insights have the potential to reshape perspectives on integrating EDI into AI and offering valuable guidance to policy makers, service providers, and AI developers to implement change in AI practices.

Acknowledgments

This study was funded by the Canadian Institutes of Health Research and the Network for Oral and Bone Health Research (*Réseau de Recherche en Santé Buccodentaire et Osseuse*). The authors acknowledge support from these institutions. The authors also acknowledge the support from the Fondation de L'ordre des Dentistes du Quebec. SAR is Canada Research Chair (tier II) in Artificial Intelligence and Advanced Digital Primary Health Care, received salary support from a Research Scholar Junior 1 Career Development Award from the Fonds de Recherche du Québec-Santé during part of this project, and her research program is supported by the Natural Sciences Research Council Discovery (grant 2020-05246).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization was led by SAR and EE, who established the study's goals, framework design, and research questions. The methodology was developed by SAR, EE, and RS, who crafted the qualitative approach and data collection strategies. Data curation was managed by SAR, EE, and RS. Formal analysis was conducted by RS. The original draft was written by SAR, EE, and RS. Reviewing and editing were a collaborative effort with all authors. Supervision and overall project leadership were provided by SAR and EE.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SRQR (Standards for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 163 KB - jmir_v26i1e63356_app1.pdf]

References

1. Artificial intelligence and gender equality: key findings of UNESCO's Global Dialogue. UNESCO. 2020. URL: https://unesdoc.unesco.org/ark:/48223/pf0000374174 [accessed 2022-09-17]
2. The Topol review: preparing the healthcare workforce to deliver the digital future. NHS. 2019 Feb. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2022-09-16]
3. Szolovits P. Artificial Intelligence and Medicine. 1st Edition. New York, NY: Routledge; 1982.

XSL•FO
RenderX

4. Panda P, Bhatia V. Role of artificial intelligence (AI) in public health. Indian J Community Fam Med 2018;4(2):60. [doi: 10.4103/2395-2113.251442]
5. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
6. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. J Gen Intern Med 2019;34(8):1626-1630 [FREE Full text] [doi: 10.1007/s11606-019-05035-1] [Medline: 31090027]
7. Tvermosegaard M, Rønn PF, Pedersen ML, Bjerregaard P, Dahl Pedersen ID, Larsen CL, et al. Validation of cardiovascular diagnoses in the Greenlandic Hospital Discharge Register for epidemiological use. Int J Circumpolar Health 2018;77(1):1422668 [FREE Full text] [doi: 10.1080/22423982.2017.1422668] [Medline: 29384037]
8. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]
9. What is the quadruple aim and what role does it play in healthcare? Privis Health. 2019. URL: https://www.privishealth.com/what-is-the-quadruple-aim-and-what-role-does-it-play-in-healthcare/ [accessed 2022-09-15]
10. Aung YM, Wong CS, Ting SW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
11. Chen YW, Stanley K, Att W. Artificial intelligence in dentistry: current applications and future perspectives. Quintessence Int 2020;51(3):248-257. [doi: 10.3290/j.qi.a43952] [Medline: 32020135]
12. Chamberlain C, McNamara B, Williams ED, Yore D, Oldenburg B, Oats J, et al. Diabetes in pregnancy among indigenous women in Australia, Canada, New Zealand and the United States. Diabetes Metab Res Rev 2013;29(4):241-256 [FREE Full text] [doi: 10.1002/dmrr.2389] [Medline: 23315909]
13. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6] [Medline: 31894144]
14. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
15. Budhwar P, Malik A, de Silva MTT, Thevisuthan P. Artificial intelligence—challenges and opportunities for international HRM: a review and research agenda. Int J Hum Resour Manag 2022;33(6):1065-1097. [doi: 10.1080/09585192.2022.2035161]
16. Charlwood A, Guenole N. Can HR adapt to the paradoxes of artificial intelligence? Hum Resour Manag J 2022 07;32(4):729-742 [FREE Full text] [doi: 10.1111/1748-8583.12433]
17. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
18. Kossow N, Windwehr S, Jenkins M. Algorithmic transparency and accountability. Transparency International. 2022. URL: https://knowledgehub.transparency.org/assets/uploads/kproducts/Algorithmic-Transparency_2021.pdf [accessed 2024-11-06]
19. Yu S, Carroll F, Bentley BL. Insights into privacy protection research in AI. IEEE Access 2024;12:41704-41726. [doi: 10.1109/ACCESS.2024.3378126]
20. Nyariro M, Emami E, Abbasgholizadeh Rahimi S. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence in health. 2022 Presented at: AH2022: 13th Augmented Human International Conference; May 26-27, 2022; Winnipeg, MB URL: https://doi.org/10.1145/3532530.3539565 [doi: 10.1145/3532530.3539565]
21. Emami E, Rahimi SA, Nyariro M. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence for better health and oral health care: a workshop summary. J Can Dent Assoc 2023 Nov;89:n10 [FREE Full text] [Medline: 38350015]
22. Nyariro M, Emami E, Caidor P, Abbasgholizadeh Rahimi S. Integrating equity, diversity and inclusion throughout the lifecycle of AI within healthcare: a scoping review protocol. BMJ Open 2023;13(9):e072069 [FREE Full text] [doi: 10.1136/bmjopen-2023-072069] [Medline: 37751956]
23. Ojikutu BO, Bogart LM, Dong L. Mistrust, empowerment, and structural change: lessons we should be learning from COVID-19. Am J Public Health 2022;112(3):401-404. [doi: 10.2105/AJPH.2021.306604] [Medline: 35196050]
24. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866-872 [FREE Full text] [doi: 10.7326/M18-1990] [Medline: 30508424]
25. Clark CR, Wilkins CH, Rodriguez JA, Preininger AM, Harris J, DesAutels S, et al. Health care equity in the use of advanced analytics and artificial intelligence technologies in primary care. J Gen Intern Med 2021;36(10):3188-3193 [FREE Full text] [doi: 10.1007/s11606-021-06846-x] [Medline: 34027610]
26. Dankwa-Mullan I, Scheufele EL, Matheny M, Quintana Y, Chapman WW, Jackson G, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. J Health Care Poor Underserved 2021;32(2S):300-317. [doi: 10.1353/hpu.2021.0065]
27. Floridi L, Cowls J. A unified framework of five principles for AI in society. In: Machine Learning and the City: Applications in Architecture and Urban Design. New York, NY: Wiley; 2022:535-545.
28. Glaser BG, Strauss AL, Strutzel E. The discovery of grounded theory; strategies for qualitative research. Nurs Res 1968;17(4):364. [doi: 10.1097/00006199-196807000-00014]

XSL•FO
RenderX

29. Charmaz K. The genesis, grounds, and growth of constructivist grounded theory. In: Developing Grounded Theory: The Second Generation Revisited. 2nd Edition. New York, NY: Routledge/Taylor & Francis Group; 2021.
30. Ralph N, Birks M, Chapman Y. The methodological dynamism of grounded theory. Int J Qual Meth 2015;14(4):160940691561157. [doi: 10.1177/1609406915611576]
31. Jabareen Y. Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Meth 2009;8(4):49-62. [doi: 10.1177/160940690900800406]
32. Chun Tie Y, Birks M, Francis K. Grounded theory research: a design framework for novice researchers. SAGE Open Med 2019;7:2050312118822927 [FREE Full text] [doi: 10.1177/2050312118822927] [Medline: 30637106]
33. Ahmed S, Mohd Asraf R. The workshop as a qualitative research approach: lessons learnt from a "critical thinking through writing" workshop. 2018 Presented at: 4th International Postgraduate Conference on Arts, Education, and Business (IPCAEB 2018); July 3-5, 2018; Langkawi, Malaysia URL: https://www.researchgate.net/publication/330526938_The_Workshop_as_a_Qualitative_Research_Approach_Lessons_Learnt_from_a_critical_Thinking_Through_Writing'_Workshop
34. Lain S. Show, don't tell: reading workshop fosters engagement and success. Texas J Literacy Educ 2017;5(2):160-167 [FREE Full text]
35. Ørngreen R, Levinsen K. Workshops as a research methodology. Electron J e-Learn 2017;15:70-81 [FREE Full text]
36. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
37. Graser K, Walzer A, Hunhevicz J, Jähne R, Seiler F, Wüst R, et al. Qualitative technology evaluation of digital fabrication with concrete: conceptual framework and scoreboard. Autom Constr 2023 Oct;154:104964 [FREE Full text] [doi: 10.1016/j.autcon.2023.104964]
38. Charmaz K, Thornberg R. The pursuit of quality in grounded theory. Qual Res Psychol 2020;18(3):305-327. [doi: 10.1080/14780887.2020.1780357]
39. Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research: notes on the Gioia methodology. Organ Res Methods 2012;16(1):15-31. [doi: 10.1177/1094428112452151]
40. Putting principles into practice: how we approach responsible AI at Microsoft. Microsoft. URL: https://www.microsoft.com/cms/api/am/binary/RE4pKH5 [accessed 2024-10-17]
41. Cachat-Rosset G, Klarsfeld A. Diversity, equity, and inclusion in artificial intelligence: an evaluation of guidelines. Appl Artif Intell 2023;37(1):2176618. [doi: 10.1080/08839514.2023.2176618]
42. Hendricks-Sturrup R, Simmons M, Anders S, Aneni K, Wright Clayton E, Coco J, et al. Developing ethics and equity principles, terms, and engagement tools to advance health equity and researcher diversity in AI and machine learning: modified Delphi approach. JMIR AI 2023;2:e52888 [FREE Full text] [doi: 10.2196/52888] [Medline: 38875540]
43. Rokhshad R, Ducret M, Chaurasia A, Karteva T, Radenkovic M, Roganovic J, et al. Ethical considerations on artificial intelligence in dentistry: a framework and checklist. J Dent 2023;135:104593. [doi: 10.1016/j.jdent.2023.104593]
44. Singhal A, Neveditsin N, Tanveer H, Mago V. Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review. JMIR Med Inform 2024;12:e50048. [doi: 10.2196/50048]
45. Tahri Sqalli M, Aslonov B, Gafurov M, Nurmatov S. Humanizing AI in medical training: ethical framework for responsible design. Front Artif Intell 2023;6:1189914 [FREE Full text] [doi: 10.3389/frai.2023.1189914] [Medline: 37261331]
46. Prakash S, Balaji JN, Joshi A, Surapaneni KM. Ethical conundrums in the application of artificial intelligence (AI) in healthcare—a scoping review of reviews. JPM 2022;12(11):1914. [doi: 10.3390/jpm12111914]

Abbreviations

AI: artificial intelligence
EDI: equity, diversity, and inclusion
SRQR: Standards for Reporting Qualitative Research

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 17.06.24; peer-reviewed by F Iniesto, E Ray Chaudhuri; comments to author 09.07.24; revised version received 16.08.24; accepted 05.09.24; published 15.11.24.

Please cite as:
Abbasgholizadeh Rahimi S, Shrivastava R, Brown-Johnson A, Caidor P, Davies C, Idrissi Janati A, Kengne Talla P, Madathil S, Willie BM, Emami E
EDAI Framework for Integrating Equity, Diversity, and Inclusion Throughout the Lifecycle of AI to Improve Health and Oral Health Care: Qualitative Study
J Med Internet Res 2024;26:e63356
URL: https://www.jmir.org/2024/1/e63356
doi:10.2196/63356
PMID:39546793

XSL•FO
RenderX

©Samira Abbasgholizadeh Rahimi, Richa Shrivastava, Anita Brown-Johnson, Pascale Caidor, Claire Davies, Amal Idrissi Janati, Pascaline Kengne Talla, Sreenath Madathil, Bettina M Willie, Elham Emami. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Accuracy of Machine Learning in Discriminating Kawasaki Disease and Other Febrile Illnesses: Systematic Review and Meta-Analysis

Jinpu Zhu[1], MD; Fushuang Yang[2], MD; Yang Wang[3], MD; Zhongtian Wang[1], MD; Yao Xiao[1], MD; Lie Wang[2*], MD; Liping Sun[2*], MD

[1]College of Chinese Medicine, Changchun University of Chinese Medicine, Changchun, China

[2]Center of Children's Clinic, The Affiliated Hospital to Changchun University of Chinese Medicine, Changchun, China

[3]Beijing Jishuitan Hospital, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Liping Sun, MD
Center of Children's Clinic
The Affiliated Hospital to Changchun University of Chinese Medicine
No. 185, Shenzhen Street
Economic and Technological Development Zone, Jilin, P.R.C.
Changchun, 130022
China
Phone: 86 15948000551
Email: slpcczyydx@sina.com

Abstract

Background: Kawasaki disease (KD) is an acute pediatric vasculitis that can lead to coronary artery aneurysms and severe cardiovascular complications, often presenting with obvious fever in the early stages. In current clinical practice, distinguishing KD from other febrile illnesses remains a significant challenge. In recent years, some researchers have explored the potential of machine learning (ML) methods for the differential diagnosis of KD versus other febrile illnesses, as well as for predicting coronary artery lesions (CALs) in people with KD. However, there is still a lack of systematic evidence to validate their effectiveness. Therefore, we have conducted the first systematic review and meta-analysis to evaluate the accuracy of ML in differentiating KD from other febrile illnesses and in predicting CALs in people with KD, so as to provide evidence-based support for the application of ML in the diagnosis and treatment of KD.

Objective: This study aimed to summarize the accuracy of ML in differentiating KD from other febrile illnesses and predicting CALs in people with KD.

Methods: PubMed, Cochrane Library, Embase, and Web of Science were systematically searched until September 26, 2023. The risk of bias in the included original studies was appraised using the Prediction Model Risk of Bias Assessment Tool (PROBAST). Stata (version 15.0; StataCorp) was used for the statistical analysis.

Results: A total of 29 studies were incorporated. Of them, 20 used ML to differentiate KD from other febrile illnesses. These studies involved a total of 103,882 participants, including 12,541 people with KD. In the validation set, the pooled concordance index, sensitivity, and specificity were 0.898 (95% CI 0.874-0.922), 0.91 (95% CI 0.83-0.95), and 0.86 (95% CI 0.80-0.90), respectively. Meanwhile, 9 studies used ML for early prediction of the risk of CALs in children with KD. These studies involved a total of 6503 people with KD, of whom 986 had CALs. The pooled concordance index in the validation set was 0.787 (95% CI 0.738-0.835).

Conclusions: The diagnostic and predictive factors used in the studies we included were primarily derived from common clinical data. The ML models constructed based on these clinical data demonstrated promising effectiveness in differentiating KD from other febrile illnesses and in predicting coronary artery lesions. Therefore, in future research, we can explore the use of ML methods to identify more efficient predictors and develop tools that can be applied on a broader scale for the differentiation of KD and the prediction of CALs.

(J Med Internet Res 2024;26:e57641) doi:10.2196/57641

XSL•FO
RenderX

KEYWORDS

machine learning; artificial intelligence; Kawasaki disease; febrile illness; coronary artery lesions; systematic review; meta-analysis

Introduction

Background

Kawasaki disease (KD) is a systemic vasculitis of small and medium-sized vessels, which has nonspecific symptoms and predominantly afflicts children aged 5 years or younger [1]. The incidence of KD worldwide varies greatly among different regions and ethnic groups, ranging from 46.3 to 359 cases per 100,000 people in East Asia, compared with 4.5 to 20 cases per 100,000 children aged 5 years or younger in Europe and North America [2]. KD is regarded as the most frequent cause of pediatric-acquired heart disease in developed countries. Its primary sequelae are related to the coronary system and may have a huge impact on the myocardium, valves, and arterial walls in some pediatric patients, which is less commonly affected in other febrile illnesses [3]. Early recognition of KD and timely administration of gammaglobulin significantly reduces the risk of coronary artery lesions (CALs), promotes the regression of coronary aneurysms, and decreases the likelihood of long-term health complications in affected children with KD [3-5].

Currently, the etiology and exact causative agent of KD remain unknown. Its clinical diagnosis is made mainly on the basis of the patient's clinical manifestations, including fever, skin and mucosal changes, swollen lymph nodes, extremity changes, rash, and cardiac manifestations, such as coronary artery dilatation or aneurysm [3,6]. Early and accurate recognition of KD is challenging due to several factors, its clinical symptoms do not appear in a specific sequence and may not manifest simultaneously; the characteristic symptoms and laboratory findings often overlap with those of other febrile illnesses, such as respiratory infections, sepsis, and multisystem inflammatory syndrome; and its self-limited nature in the acute phase can cause symptoms to diminish or disappear before a definitive diagnosis is made [3,7,8]. People with KD who do not respond adequately to initial intravenous immunoglobulin (IVIG) therapy are at a higher risk of developing CALs. Therefore, classical scoring systems for predicting IVIG nonresponsiveness, such as the Kobayashi score [9], Harada score [10], and Egami score [11], are also considered valuable for assessing the risk of CALs. However, the sensitivity of these established scoring systems outside Japan is not good. In addition, since infectious agents and immune regulatory dysfunction in the body may be the causative factors of KD, stress-related inflammatory markers occurring after exogenous infections, such as the erythrocyte sedimentation rate (ESR), C-reactive protein (CRP) level, white blood cell (WBC) count, and platelet (PLT) count, are thought to have the ability to predict CALs in people with KD, and cardiac markers including N-terminal pro-B-type natriuretic peptide and markers of cardiomyocyte damage are also considered useful for predicting CALs. However, the predictive performance of single predictors is limited [12-14]. Therefore, at present, how to efficiently diagnose KD and how to early predict CALs in people with KD remain urgent.

Since Alan Turing proposed the theoretical foundation of artificial intelligence (AI) in the 1930s, the field has undergone over 70 years of development. Following the turn of the millennium, with the proliferation of computer technology and the internet, the surge in data availability and enhanced computational power have led to the widespread adoption of machine learning (ML) techniques [15]. Alongside advancements in medical testing technologies, vast amounts of medical imaging, laboratory monitoring data, and genetic sequencing results now require interpretation by physicians. The heavy load of medical reports and the complex communication processes between doctors and patients have contributed to diagnostic errors, inefficiencies in medical practice, and physician burnout. As a result, there is a growing clinical need for automated, intelligent, and reasonably accurate ML tools to assist in these tasks [16-18]. Classic ML techniques are broadly categorized into supervised and unsupervised learning. Compared with the latter, the former is more frequently used in the medical field because it can generate clinically applicable conclusions by learning from and analyzing large volumes of labeled data (such as laboratory data, medical images, and genetic data). This makes supervised learning particularly valuable for tasks such as disease prediction, diagnosis, and image segmentation in clinical practice [19]. ML can analyze complex data from a wide range of sources, such as clinical manifestations, laboratory results and imaging data. Unlike traditional approaches that focus on single biochemical indicators, ML can detect subtle correlations within large datasets, helping to address challenges that traditional statistical methods cannot solve. This capability enhances both predictive and diagnostic accuracy [20]. In the field of disease prediction, ML can forecast disease progression and adverse outcomes, which is crucial for developing targeted treatment and rehabilitation regimens. While, in disease diagnosis, ML is primarily applied to the diagnosis of complex diseases and the assessment of disease states. This can significantly aid in formulating specialized diagnostic strategies, thereby reducing the time and economic costs associated with diagnosis [21-24].

Objective

In recent years, some researchers have attempted to apply ML to differentiate KD from other febrile illnesses and to predict CALs in people with KD. However, there is still a lack of systematic evidence supporting the effectiveness of ML methods in these areas, which presents challenges for the development or refinement of AI diagnostic tools. Therefore, we have conducted the first systematic review and meta-analysis to evaluate the accuracy of ML in differentiating KD from other febrile illnesses and in predicting CALs in people with KD, so as to provide evidence-based support for the future development and application of AI in these fields.

XSL•FO
RenderX

Methods

Study Registration

This study was executed according to the operating guidelines for systematic reviews and meta-analyses and prospectively registered on PROSPERO (CRD42023481662).

Eligibility Criteria

The inclusion and exclusion criteria are shown in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- To differentiate Kawasaki disease (KD) from other febrile illnesses, people with KD and people with febrile illnesses suspected to be KD were included as study participants. To predict coronary artery lesions (CALs) in KD, people with confirmed KD were included as study subjects.
- The types of studies included were case-control studies, cohort studies, nested case-control studies, and case-cohort studies.
- A predictive model for identifying KD or predicting CALs in people with KD was constructed entirely.
- Some studies lacked an independent validation set and only adopted k-fold cross-validation or Bootstrap. Their contribution cannot be denied. Therefore, they were also included in our systematic review.
- A small number of studies may be based on distinct machine learning (ML) models published on the same dataset, but they were also incorporated into our systematic review.
- Included studies are published in English.

Exclusion criteria

- Meta-analyses, reviews, guidelines, expert opinions, and similar types of studies.
- Only risk factor analysis was executed, and a full ML model was not constructed.
- Studies that did not include any of the following outcome measures for assessing the predictive accuracy of ML models: Receiver Operating Characteristic, Concordance statistic, Concordance index, sensitivity, specificity, accuracy, recovery, precision, confusion matrix, fourfold table for estimating the quality of a diagnostic test, F_1-score, and calibration curve.
- Studies on the validation of mature scale.
- Studies on the predictive accuracy of single-factor predictors.

Data Sources and Search Strategy

PubMed, Cochrane Library, Embase, and Web of Science were systematically searched until September 26, 2023. The search terms were designed based on a combination of subject headings and free-text words, without geographical restriction or time limit (year). The search strategy is detailed in Table S1 in Multimedia Appendix 1.

Study Selection and Data Extraction

The retrieved records were imported into Endnote, and duplicate publications were excluded. Then we read the titles and abstracts to screen out the original studies that did not meet the requirements. Next, full texts were downloaded and read to select eligible studies.

A standard electronic data extraction spreadsheet was created before data extraction. The following data were extracted: titles of the original studies, the name of the first author, the year of publication, the author's nationality, the study type, the data source, the number of people with KD, the total number of study participants, the numbers and total number of people with KD in the training set and validation set when a model was trained to identify KD, the numbers and total number of patients developing CALs in the training set and validation set when a model was trained for early prediction of CALs in KD, model validation methods, methods for handling missing values, methods for variable screening, types of models used, and predictors used in the final model.

Literature screening and data extraction were carried out independently by 2 researchers (JZ and FY), and then, the results were cross-checked. Discrepancies were resolved by the third researcher (YW).

Risk of Bias in Studies

The risk of bias in eligible studies was appraised using the PROBAST (Prediction model Risk Of Bias Assessment Tool), which consisted of numerous questions across 4 different domains: participants, predictors, outcome, and analysis. This approach provides a comprehensive evaluation of the overall risk of bias and overall usability [25]. The 4 domains covered 2, 3, 6, and 9 specific questions, respectively. There are 3 answers (yes [Y]/probably yes [PY], no [N]/probably no [PN], and no information [NI]) to each question. If at least 1 question in a domain was answered as N or PN, it was rated at high risk. To be considered at low risk, all questions in the domain should be answered as Y or PY. The overall risk of bias was graded as low when all domains were considered at low risk, and the overall risk of bias was rated high risk when at least one domain was rated at high risk.

Risk of bias assessment was conducted independently by 2 researchers [JZ and ZW] based on the PROBAST, and then, their results were cross-checked. Discrepancies were resolved by the third researcher [LS].

XSL•FO
RenderX

Outcome Measures

The primary outcome measures were the concordance index (C-index), sensitivity, and specificity reflecting the predictive accuracy of ML models. The secondary outcome measure was the frequency of use of a predictor to construct a ML model.

Meta-Analysis Methods

The meta-analysis was performed using Stata (version 15.0; Stata Corp). A meta-analysis was conducted of the measure (C-index) to assess the overall accuracy of ML models. If 95% CIs and standard errors of the C-index were missing in some original studies, its standard errors were estimated with reference to a study by Debray et al [26]. A random-effects model was leveraged for meta-analysis of the C-index given the differences in variables and inconsistent parameters included in various ML models.

In addition, a bivariate mixed-effects model was leveraged for meta-analysis of sensitivity and specificity, which required that meta-analysis of sensitivity and specificity should be based on the fourfold table (2 × 2 table) for estimating the quality of a diagnostic test. However, the fourfold table was not reported in some original studies. Therefore, the fourfold table was calculated by the following 2 methods, calculation of sensitivity, specificity, and precision combined with the number of patients; and calculation of the best sensitivity and specificity determined based on the Youden index and then combined with the number of patients. When conducting a meta-analysis of the C-index, we used the “metan” program package, whereas for the meta-analysis of the fourfold table (2 × 2 table), we used the “midas” program package.

Results

Study Selection

A total of 815 records were searched from the databases. Of them, 196 were removed due to duplication, and 569 were excluded after screening of titles and abstracts. The full texts of the remaining 50 articles were searched and read. Of them, 2 were excluded without peer review of abstracts, and 19 were excluded for the reasons described in Figure 1. In the end, a total of 29 studies were included [27-55].

XSL•FO
RenderX

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram for study selection.

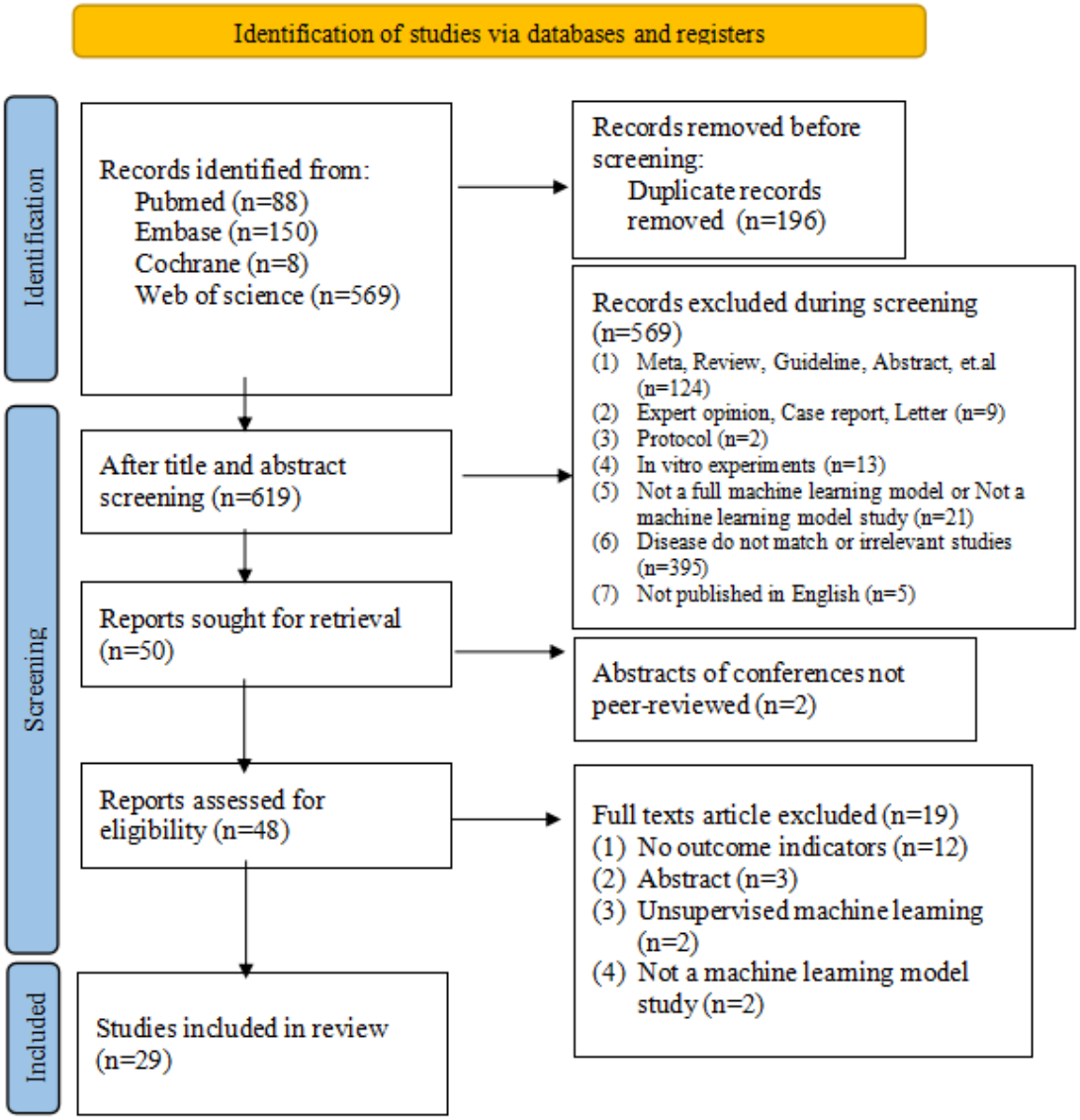

Study Characteristics

This study included a total of 29 publications between 2001 and 2023, involving 29 ML models with variables all derived from relatively common clinical characteristics.

Twenty studies involved a total of 20 models. They, published from 2013 to 2023, focused on the differential diagnosis of KD from Febrile controls (including respiratory infections, gastrointestinal infections, urinary infections, encephalitis, multisystem inflammatory syndrome, sepsis and cervical lymphadenopathy with fever due to bacterial infections, viral infections, or indeterminate pathogens). Of them, 11 were conducted in China [27-37], 5 in the United States [38-42], 1 in Egypt [43], 1 in Germany [44], 1 in Italy [45], and 1 in South Korea [46]. These studies involved a total of 103,882 study participants, including 12,541 people with KD.

The types of models used in these 20 studies included logistic regression (9/20, 45%), support vector machine (3/20, 15%), least absolute shrinkage and selection operator (2/20, 10%), artificial neural network (2/20, 10%), extreme gradient boosting (1/20, 5%), decision tree (1/20, 5%), random forest (1/20, 5%), and linear discriminant analysis (1/20, 5%).

The other 9 studies (involving 9 models) focused on the prediction of the risk of CALs in KD children. Of them, 5 were conducted in China [47-51], 2 in Japan [52,53], 1 in Spain [54], and 1 in the United States [55]. The types of models used in these 9 studies included multivariate logistic regression (8/9, 88.89%) and an artificial neural network (1/9, 11.11%). These 9 studies involved a total of 6503 people with KD, including 986 ones with CALs. Characteristics of including studies are presented in Table S2 and Table S3 in Multimedia Appendix 1.

Risk of Bias in Studies

In the differential diagnosis of KD from other febrile illnesses, 9 models [27-30,32,33,35,36,46] were based on data from retrospective studies, and therefore had a high risk of participant selection bias; 4 models [31,37,40,45] were constructed without data sources reported, so they were at unclear risk of bias. In

XSL•FO
RenderX

the domain of predictors, 20 models were constructed without quality control measures reported for predictor assessment, so they were at an unclear risk of bias. In terms of outcomes, 20 models were all at low risk of bias. Regarding statistical analysis, 14 models included in 13 studies [27,29-31,33,35,37,40-43,45,46] were constructed with poor sample size design, missing data handled inappropriately or predictors screened irrationally, so they were at high risk of bias, as shown in Figure 2. The risk of bias in including studies is presented in Table S4 and Table S5 in Multimedia Appendix 1.

Figure 2. Risk assessment of machine learning models to identify Kawasaki disease from febrile controls.

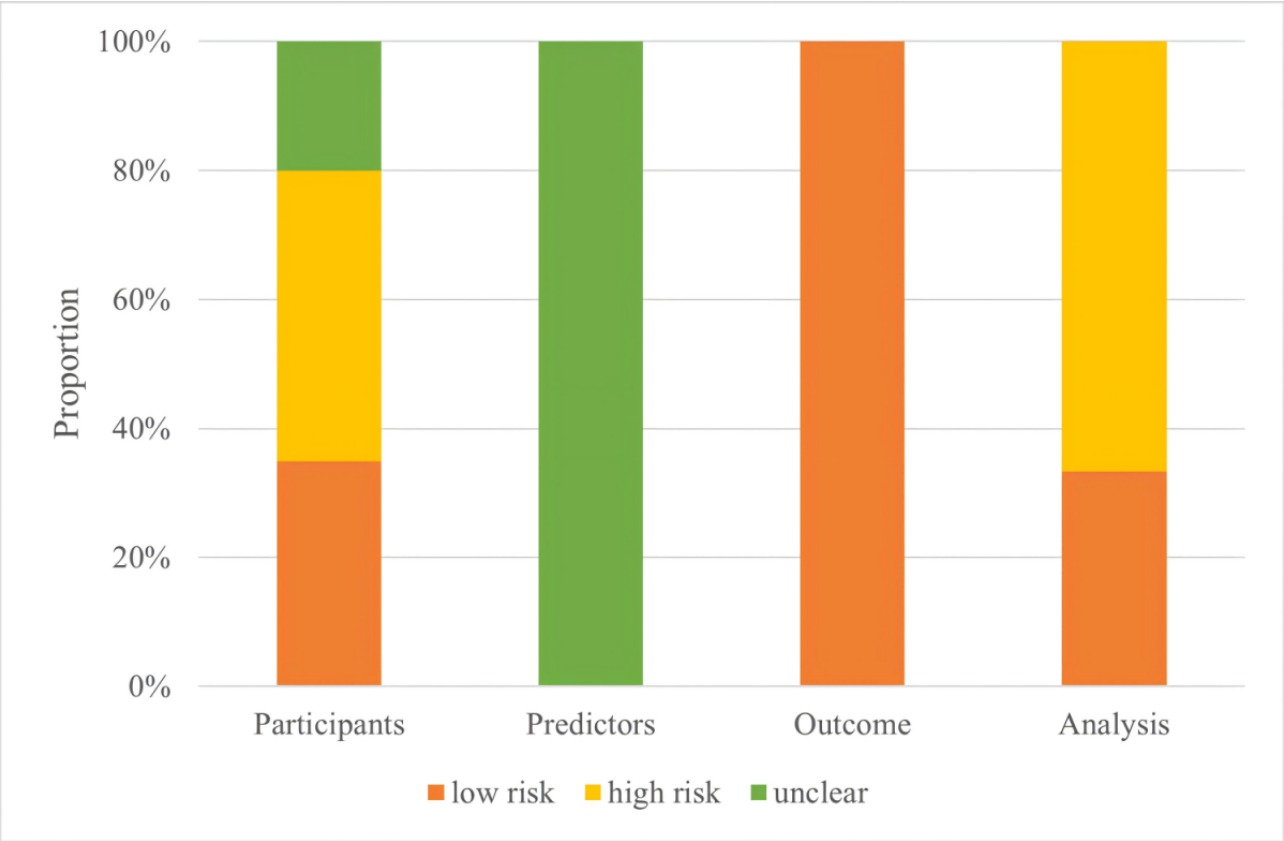

In the early prediction of CALs in people with KD, 5 models [48-52] were based on data from retrospective case-control studies, and therefore had a high risk of participant selection bias. In predictors and outcomes as domains, 9 models were all at low risk of bias. In statistical analysis as a domain, 6 models [47,51-55] were at high risk of bias because of irrational sample size design, and 3 models [48-50] were at high risk of bias owing to incorrect use of internal validation methods as shown in Figure 3.

XSL•FO
RenderX

Figure 3. Risk assessment of machine learning models to predict coronary artery lesions in people with Kawasaki disease.

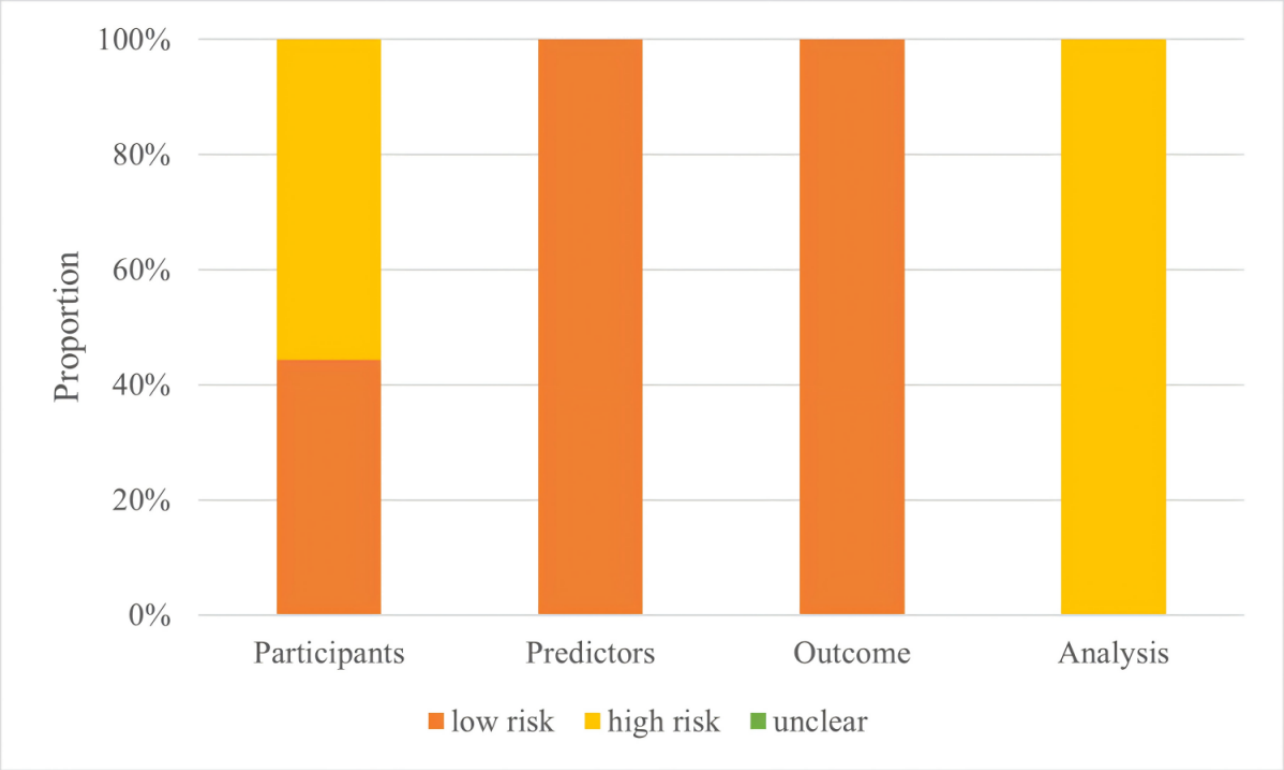

Meta-Analysis

Differentiation of Kawasaki Disease From Other Febrile Illnesses

In the training set, a random-effects model was leveraged for meta-analysis of the C-index, which was 0.898 (95% CI 0.874-0.922), as shown in Figure 4 [29-36,40-46]. The funnel plot showed no significant publication bias among the included studies. The results of meta-analysis for the 2 × 2 table showed a sensitivity of 0.89 (95% CI 0.83-0.92) and a specificity of 0.84 (95% CI 0.80-0.87) in the training set, as shown in Figure 5 [27,29-36,38,40-46].

XSL•FO
RenderX

Figure 4. Concordance index forest plots of machine learning to identify Kawasaki disease and febrile controls in training and validation sets [29-36,40-46].

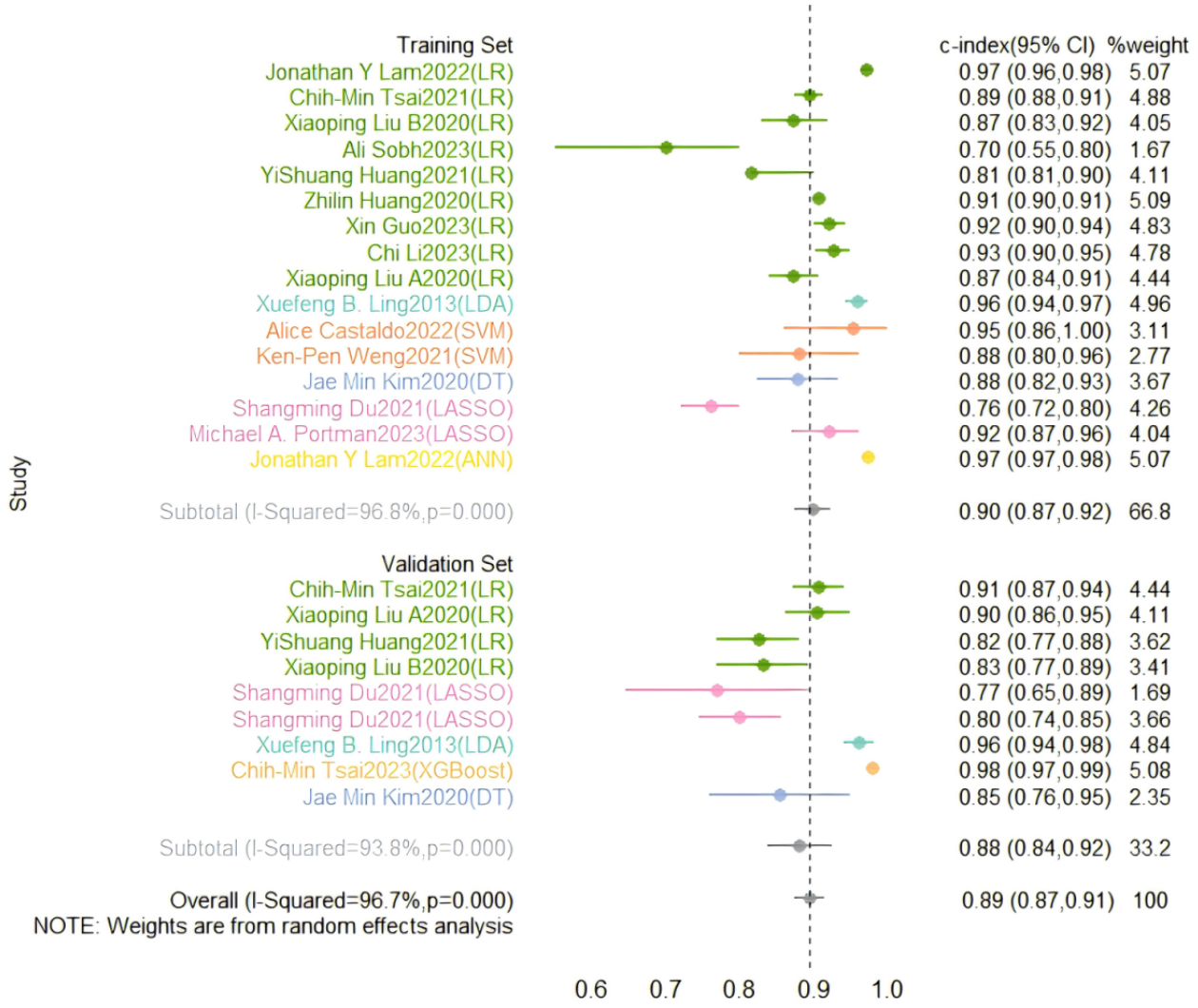

XSL•FO
RenderX

Figure 5. Sensitivity and specificity of machine learning to identify Kawasaki disease and febrile controls in training sets [27,29,30-36,38,40-46].

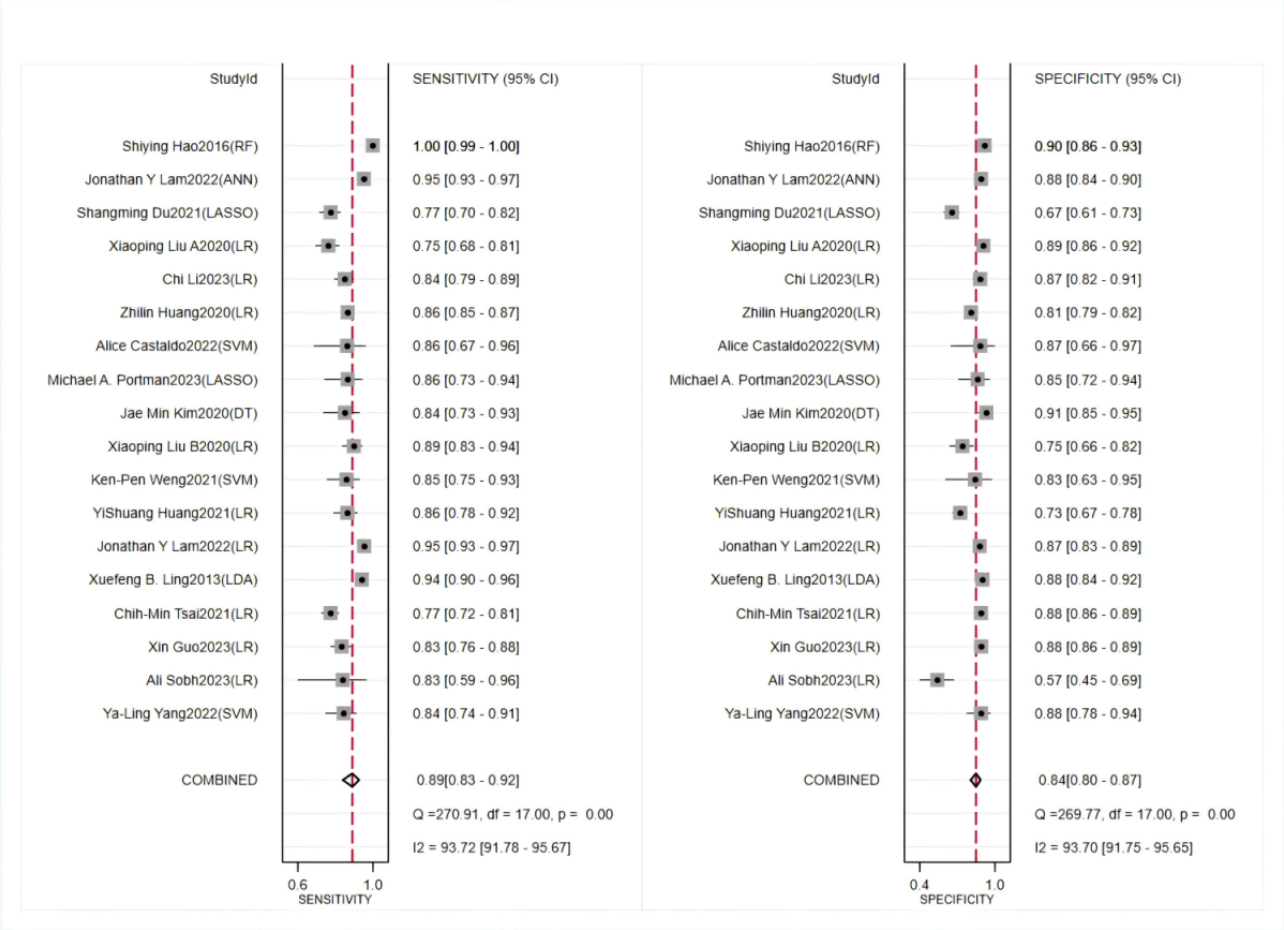

In the validation set, a random-effects model was used for meta-analysis of the C-index, which was 0.881 (95% CI 0.837-0.925), as shown in Figure 4 [29-36,40-46]. The funnel plot showed that there appeared to be publication bias among the included studies. The results of meta-analysis for the 2 × 2 table showed a sensitivity of 0.91 (95% CI 0.83-0.95) and a specificity of 0.86 (95% CI 0.80-0.90) in the validation set, as shown in Figure 6 [28,32-35,37-39,42,44,46].

XSL•FO
RenderX

Figure 6. Sensitivity and specificity of machine learning to identify Kawasaki disease and febrile controls in validation sets [28,32-35,37-39,42,44,46].

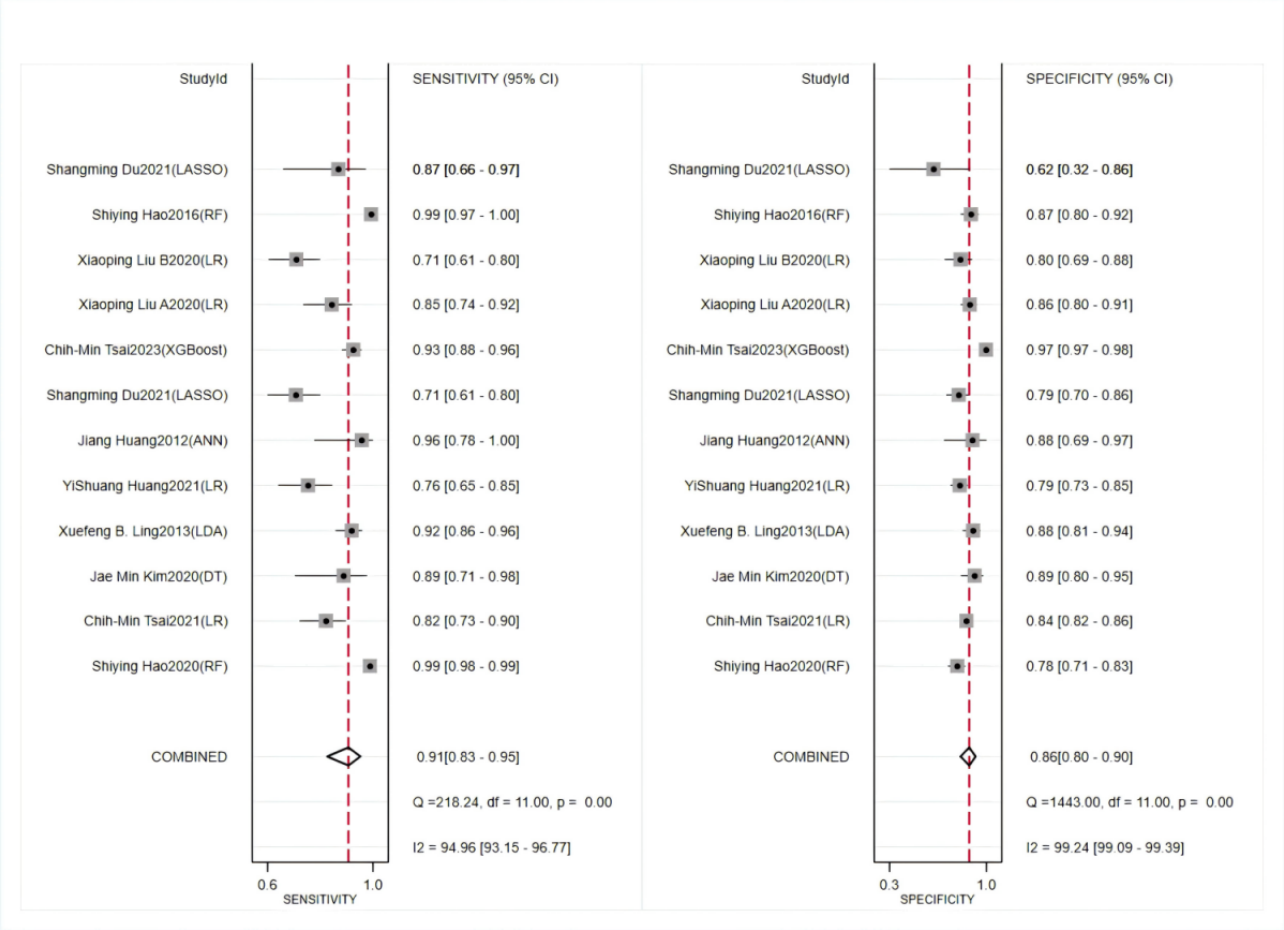

Prediction of Coronary Artery Lesions in Kawasaki Disease

In the training set, a random-effects model was used for meta-analysis of the C-index, which was 0.809 (95% CI 0.761-0.857), as shown in Figure 7 [47-55]. The funnel plot showed no significant publication bias among the included studies. The results of meta-analysis for the 2 × 2 table showed a sensitivity of 0.79 (95% CI 0.68-0.86) and a specificity of 0.83 (95% CI 0.65-0.93) in the training set, as shown in Figure 8 [48-54].

XSL•FO
RenderX

Figure 7. Concordance index forest plots of machine learning to predict coronary artery lesions in Kawasaki disease in training and validation sets [47-55].

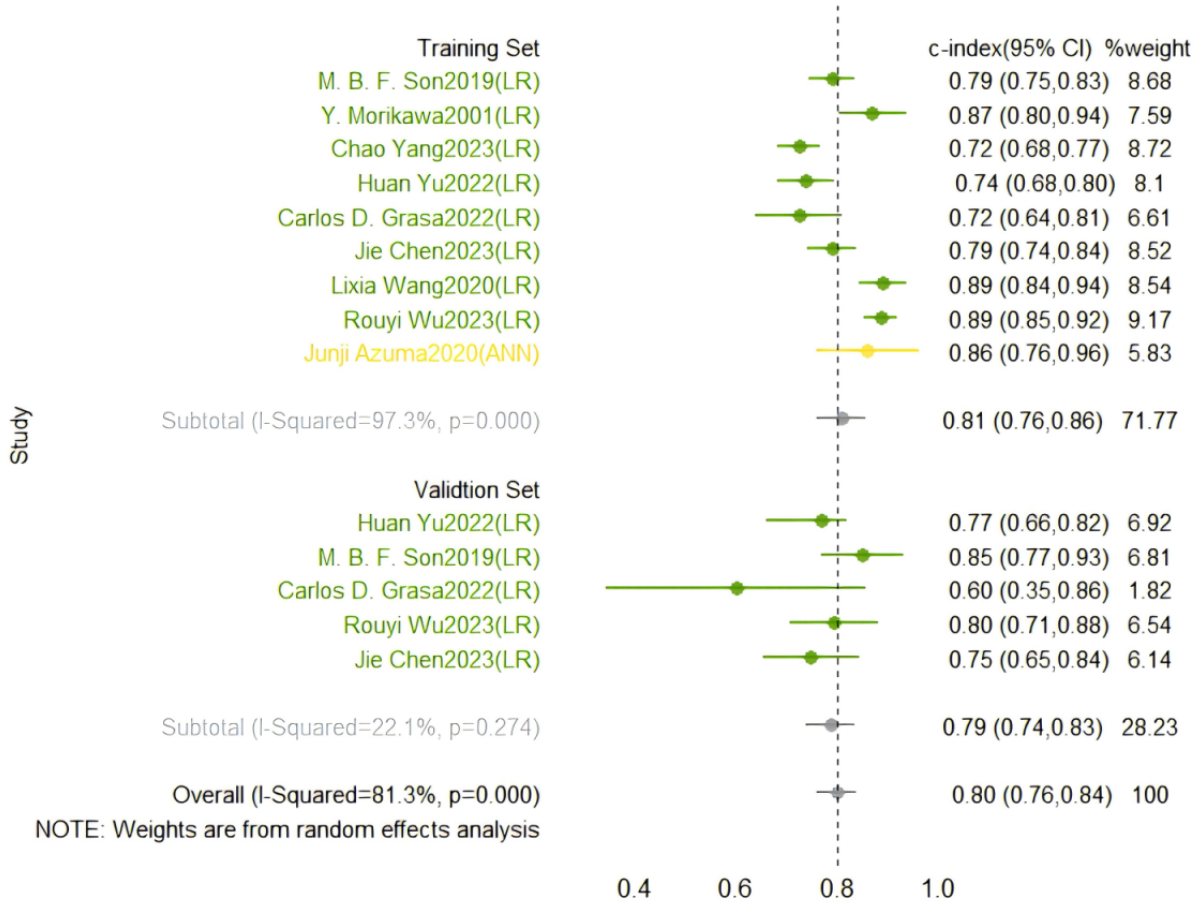

XSL•FO
RenderX

Figure 8. Sensitivity and specificity of machine learning to predict coronary artery lesions in Kawasaki disease in training sets [48-54].

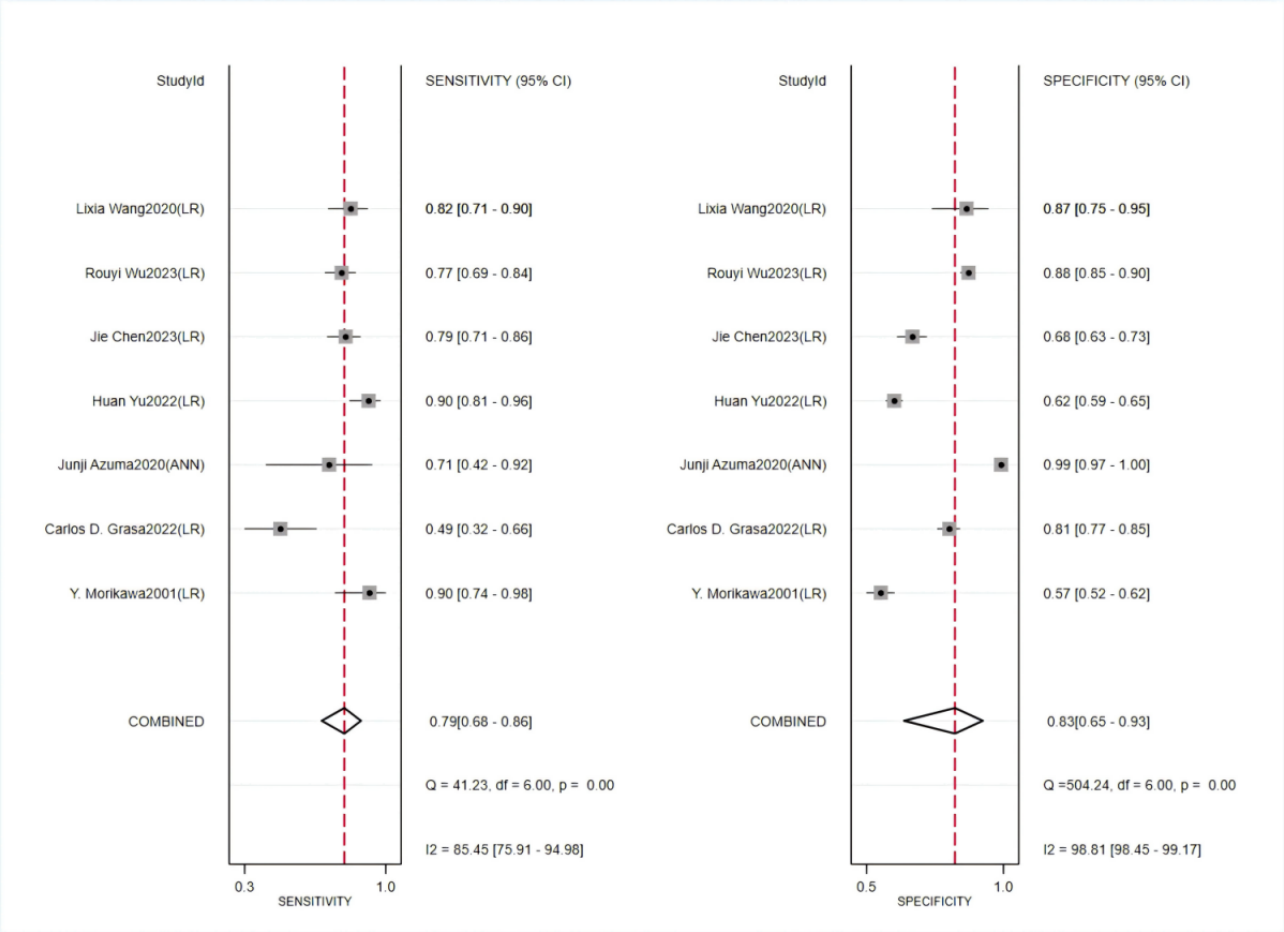

In the validation set, a random-effects model was used for meta-analysis of the C-index, which was 0.787 (95% CI 0.738-0.835), as shown in Figure 7 [47-55]. The funnel plot showed that there appeared to be publication bias among the included studies. Since the number of models in the validation set was less than 4 and did not reach the lower limit for meta-analysis, we only described the ranges of sensitivity and specificity in the validation set, which were 0.220-0.952 and 0.644-0.755, respectively.

Predictors

Frequencies of Use of Diagnostic Factors to Construct ML Models for Differentiation of KD From Other Febrile Illnesses

Regarding the use of diagnostic factors for differentiation of KD, there were a total of 39 diagnostic factors including CRP (n=11), neutrophils (NEUT; n=11), WBC (n=10), PLT (n=8), lymphocytes (n=8), hemoglobin (n=7), alanine aminotransferase (n=7), gamma-glutamyl transferase (n=7), eosinophils (n=7), monocytes (n=6), ESR (n=5), oral changes (n=5), conjunctival injection (n=5), extremity changes (n=4), age (n=3), cervical lymphadenopathy (n=3), rash (n=3), albumin (n=2), prognostic nutritional index (n=2), prealbumin (n=2), and duration (days) of fever (n=2). The following diagnostic factors were each used with a frequency of 1: Plasma Hepcidin, Pyuria, Urinary Leukocytes, Vomiting, Abdominal Pain, N-terminal Pro B-type Natriuretic Peptide, Thyroid Hormone Uptake, Height, complement C3, Procalcitonin, Neck Abscesses, Blood Phosphorus, Uric Acid, Chloride Compounds in serum, lactate dehydrogenase, aspartate aminotransferase, globulin, and temperature (Figure 9).

XSL•FO
RenderX

Figure 9. Frequencies of use of diagnostic factors to construct machine learning models to identify Kawasaki disease from febrile controls.

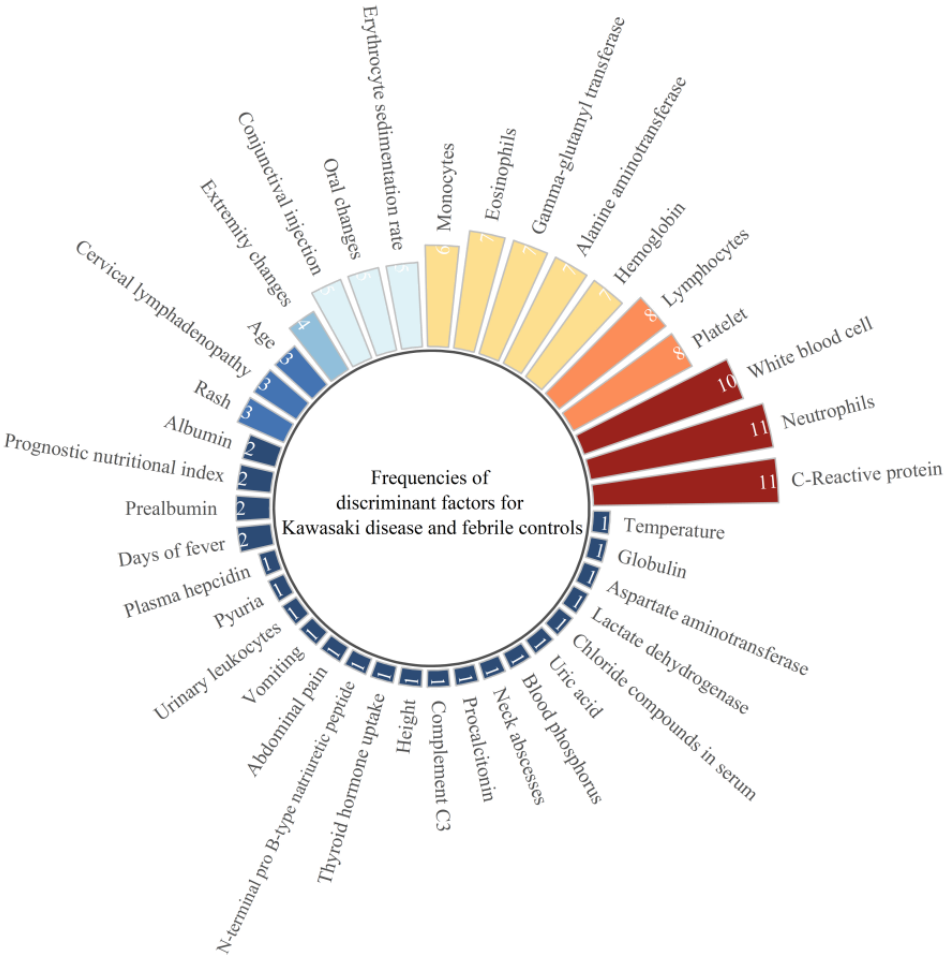

Frequencies of Use of Predictors to Construct ML Models for Prediction of Coronary Artery Lesions in KD

Regarding the use of diagnostic factors for prediction of CALs, there were a total of 26 predictors including CRP (n=7), IVIG resistance (n=4), duration of fever before IVIG infusion (n=3), age (n=3), baseline Z score (n=2), high-density Lipoprotein (n=2), albumin (n=2), PLT (n=2), ESR (n=2), duration of fever after IVIG infusion (n=1), monocytes (n=1), serum amyloid A (n=1), NEUT (n=1), lymphocytes (n=1), rash (n=1), oral changes (n=1), cervical lymphadenopathy (n=1), sex (n=1), race (n=1), creatinine (n=1), number of Steroid Pulse therapy sessions (n=1), Urinary β2-microglobulin (n=1), Hgb (n=1), Immunoglobulin A (n=1), immunoglobulin G (n=1), and matrix metalloproteinase 9 (n=1; Figure 10).

XSL•FO
RenderX

Figure 10. Frequencies of use of predictors to construct machine learning models to predict coronary artery lesions in Kawasaki disease.

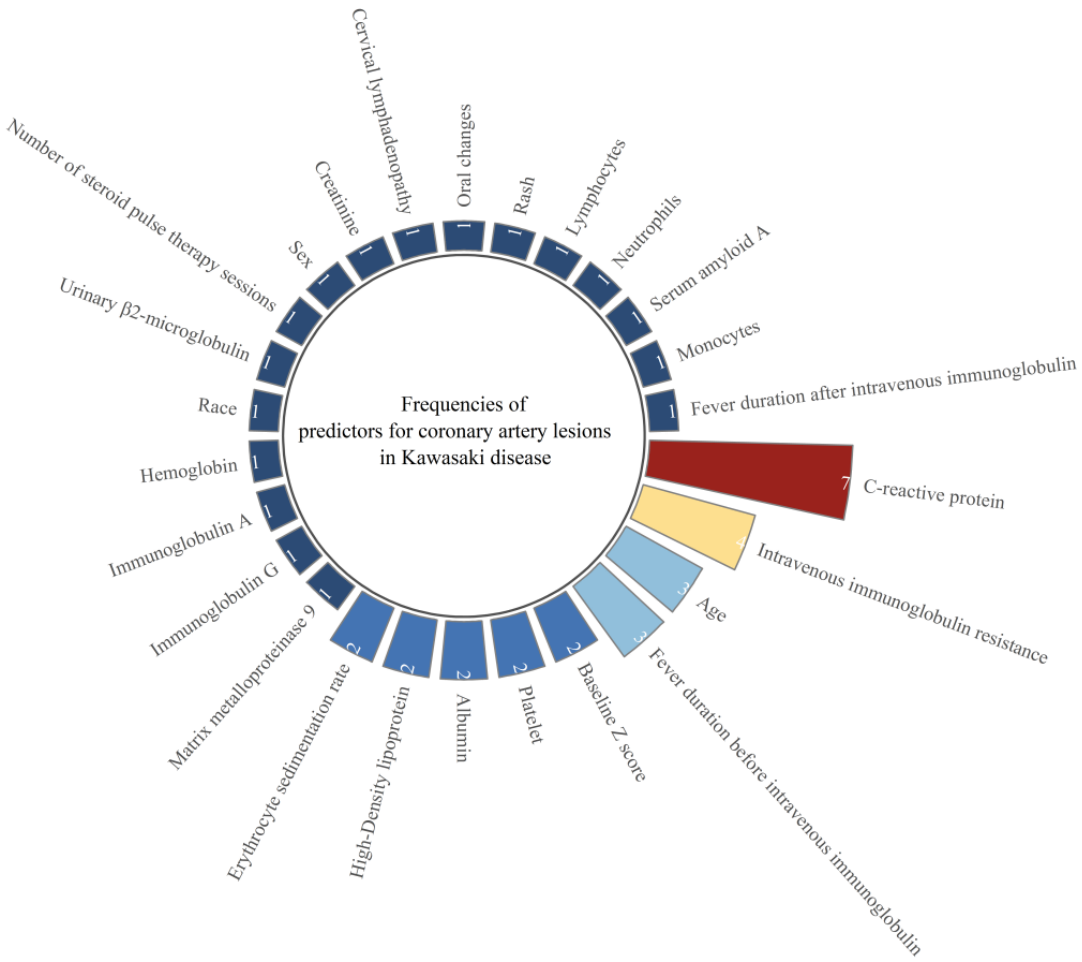

Discussion

Principal Findings

Our study included a total of 29 ML prediction models, of which 20 models focused on differentiating KD from other febrile illnesses, and 9 models focused on predicting the risk of CALs in children with KD. This study unraveled that ML achieved a favorable performance in differentiating KD from other febrile illnesses and predicting the risk of CALs in children with KD. In the validation set for ML to differentiate KD from other febrile illnesses, the C-index, sensitivity, and specificity were 0.881 (95% CI 0.837-0.925), 0.91 (95% CI 0.83-0.95), and 0.86 (95% CI 0.80-0.90), respectively. In the validation set for early prediction of the risk of CALs in children with KD, the C-index was 0.787 (95% CI 0.738-0.835), with a sensitivity range of 0.220-0.952 and a specificity range of 0.644-0.755. All the results in the validation set versus the training set showed no overfitting.

CRP, NEUT, and WBC were the most frequently used diagnostic factors in the diagnostic model for KD, and CRP was also the most frequent predictor in the predictive model for CALs in people with KD. CRP is a very common laboratory test item. All the models we included were constructed based on interpretable ML models and routine clinical or laboratory variables, suggesting that desirable predictive results can be obtained if common clinical characteristics and laboratory variables are combined with appropriate ML modeling methods.

Comparison With Previous Reviews

In recent years, ML has been extensively applied in the diagnosis and early prediction of diseases to enhance precision and efficiency in health care, but the value of its application in the diagnosis of KD and early prediction of the risk of CALs in KD children is still unknown. A systematic review of existing studies is necessary to guide the development and application of ML in this field. To our knowledge, our study is the first systematic review evaluating the accuracy of ML for differentiation of KD from other febrile illnesses and for early prediction of CALs in children with KD.

In clinical practice, there is an emphasis on using rapid and cost-effective diagnostic methods for children with fever and at risk of coronary artery disease. Most of the studies included in our review are based on common and routine clinical data. This indicated that when faced with patients with fever, clinicians can use ML methods and the most routine tests to differentiate and diagnose KD.

Over the past few years, many teams have attempted to diagnose KD using genomic technology in addition to common clinical characteristics and laboratory variables. A meta-analysis by Zheng et al [56] showed that total mixed miRNAs detection for diagnosing KD had a sensitivity of 0.7 (0.66-0.74), a specificity

XSL•FO
RenderX

of 0.87 (0.83-0.90), and an Summary Receiver Operating Characteristic of 0.8302. In contrast, our analysis showed that ML had a higher diagnostic accuracy. Another meta-analysis [57] showed that compared with ML, ncRNA had a similar diagnostic accuracy with an AUC of 0.90 (0.87-0.92) in differentiating KD from febrile illnesses. Some studies have found the same genetic association with risk for CAL in KD children. Xie et al [58] reviewed 164 previous studies and found that *BTNL2*, *CASP3*, *FCGR2A*, *FGF23*, *FG*β, *GRIN3A*, *HLA-E*, *IL10*, *ITPKC*, and *TGFBR2* were linked to a high risk for CALs in children with KD. Another meta-analysis [59] showed that Allele A at the functional SNP rs72689236 of the gene encoding caspase-3 was a genetic marker for susceptibility to KD with CALs. Although genomic technology has demonstrated some value and significance in the diagnosis of KD and the prediction of KD with CALs, genetic testing is time consuming and not applicable to the diagnosis of the acute phase of the disease. Considering the higher cost of genomic technology, the need for high-quality control in the laboratory and the complexity of genomic data analysis and interpretation, genomic technology is less clinically applicable to the early diagnosis of KD and the prediction of concomitant CALs. In contrast, common clinical variables are more applicable and convenient in most current clinical settings, while being more conducive to reducing the economic burden on people with KD.

During the application of ML in the field of medicine, model type selection is critical to balance the accuracy and interpretability of a model. Interpretable models, such as logistic regression and the decision tree, tend to have a worse accuracy in identifying outcome measures. However, less interpretable models, such as Random Forest, XGboost, and Artificial Neural Networks, tend to have a higher accuracy. Logistic regression was the main ML model included in our study, suggesting that early diagnosis of KD and early prediction of CALs in people with KD were feasible based on routinely interpretable models.

In addition, the selection of model variables is crucial for the construction of ML models, because it determines the testing accuracy of ML models. In current clinical practice, model variables can be classified into several types below, including common interpretable clinical characteristics, image-based radiomic features, and pathologic characteristics. Considering that it is difficult to avoid excessive imaging equipment configuration and bias in image segmentation during obtainment of variables through imaging and that pathologic characteristics have limitations including difficulty in quantification and limited data interpretation, we should prefer to use common clinical characteristics as model variables among these variables. The ML model variables included in our study were dominated by common interpretable clinical characteristics, which suggests that easy-to-use clinical tools can be developed in the future based on common clinical characteristics.

Advantages and Limitations of the Study

Our study is the first systematic review evaluating the accuracy of ML for the differentiation of KD from other febrile illnesses and for early prediction of CALs in children with KD. It was aimed at providing medical evidence for this field. However, our study has the following limitations. First, despite an extensive search, the number of original studies included is still limited, thereby resulting in a failure to conduct effective subgroup analysis based on ML model types in our study. Second, most of the ML models included in our review were constructed using random sampling for internal validation, with only a few studies using an independent external validation set to verify the accuracy of the models. This limitation somewhat restricts the generalizability and interpretation of our results. Third, we did not quantitatively describe the associations between the predictors (model variables) and KD or CALs in KD.

Conclusions

ML is desirably effective in differentiating KD from other febrile illnesses, and it is also able to fairly predict the occurrence of CALs. However, the development of ML models in the related field is still limited by a small amount of training data, its low quality and a single validation method. Thus, more subsequent multicenter studies are needed to provide more diverse data for ML models, enhance data quality, improve the validation method and outcome measures for ML models, refine the existing models, and develop new high-quality models. In future work, a quantitative characterization of the association between predictors (model variables) and both KD and CALs in KD is needed. In the meantime, more multicenter studies are desired to validate our conclusions. Furthermore, we plan to explore how to further validate and apply ML technologies in clinical practice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82405483), the Department of Science and Technology of Jilin Province (20200603008SF; YDZJ202401673ZYTS), and the Jilin Provincial Administration of Traditional Chinese Medicine (2023129).

Authors' Contributions

All authors contributed to the study conception and design. JZ and YX wrote the original draft preparation. LW and LS conducted review and editing. JZ contributed to conceptualization. JZ, FY and ZW handled methodology. JZ, FY, YW, ZW, and LS managed formal analysis and investigation. LW and LS handled supervision. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. LW and LS contributed equally as the corresponding authors of this manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File , 48 KB - jmir_v26i1e57641_app1.docx]

Multimedia Appendix 2
PRISMA checklist.
[DOCX File , 32 KB - jmir_v26i1e57641_app2.docx]

References

1. Agarwal S, Agrawal DK. Kawasaki disease: etiopathogenesis and novel treatment strategies. Expert Rev Clin Immunol 2017;13(3):247-258 [FREE Full text] [doi: 10.1080/1744666X.2017.1232165] [Medline: 27590181]
2. Watts RA, Hatemi G, Burns JC, Mohammad AJ. Global epidemiology of vasculitis. Nat Rev Rheumatol 2022;18(1):22-34. [doi: 10.1038/s41584-021-00718-8] [Medline: 34853411]
3. McCrindle BW, Rowley AH, Newburger JW, Burns JC, Bolger AF, Gewitz M, American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee of the Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; Council on Cardiovascular Surgery and Anesthesia; and Council on Epidemiology and Prevention. Diagnosis, treatment, and long-term management of Kawasaki disease: a scientific statement for health professionals from the American heart association. Circulation 2017;135(17):e927-e999. [doi: 10.1161/CIR.0000000000000484] [Medline: 28356445]
4. Galeotti C, Bayry J, Kone-Paut I, Kaveri SV. Kawasaki disease: aetiopathogenesis and therapeutic utility of intravenous immunoglobulin. Autoimmun Rev 2010;9(6):441-448. [doi: 10.1016/j.autrev.2009.12.004] [Medline: 20004744]
5. Gordon JB, Kahn AM, Burns JC. When children with kawasaki disease grow up: myocardial and vascular complications in adulthood. J Am Coll Cardiol 2009;54(21):1911-1920 [FREE Full text] [doi: 10.1016/j.jacc.2009.04.102] [Medline: 19909870]
6. Ayusawa M, Sonobe T, Uemura S, Ogawa S, Nakamura Y, Kiyosawa N, Kawasaki Disease Research Committee. Revision of diagnostic guidelines for Kawasaki disease (the 5th revised edition). Pediatr Int 2005;47(2):232-234. [doi: 10.1111/j.1442-200x.2005.02033.x] [Medline: 15771703]
7. Verdoni L, Mazza A, Gervasoni A, Martelli L, Ruggeri M, Ciuffreda M, et al. An outbreak of severe Kawasaki-like disease at the Italian epicentre of the SARS-CoV-2 epidemic: an observational cohort study. Lancet 2020;395(10239):1771-1778 [FREE Full text] [doi: 10.1016/S0140-6736(20)31103-X] [Medline: 32410760]
8. Godfred-Cato S, Bryant B, Leung J, Oster ME, Conklin L, Abrams J, California MIS-C Response Team. COVID-19-associated multisystem inflammatory syndrome in children - United States, March-July 2020. MMWR Morb Mortal Wkly Rep 2020;69(32):1074-1080. [doi: 10.15585/mmwr.mm6932e2] [Medline: 32790663]
9. Kobayashi T, Inoue Y, Takeuchi K, Okada Y, Tamura K, Tomomasa T, et al. Prediction of intravenous immunoglobulin unresponsiveness in patients with Kawasaki disease. Circulation 2006;113(22):2606-2612. [doi: 10.1161/CIRCULATIONAHA.105.592865] [Medline: 16735679]
10. Harada K. Intravenous gamma-globulin treatment in Kawasaki disease. Acta Paediatr Jpn 1991;33(6):805-810. [doi: 10.1111/j.1442-200x.1991.tb02612.x] [Medline: 1801561]
11. Egami K, Muta H, Ishii M, Suda K, Sugahara Y, Iemura M, et al. Prediction of resistance to intravenous immunoglobulin treatment in patients with Kawasaki disease. J Pediatr 2006;149(2):237-240. [doi: 10.1016/j.jpeds.2006.03.050] [Medline: 16887442]
12. de Zorzi A, Colan SD, Gauvreau K, Baker AL, Sundel RP, Newburger JW. Coronary artery dimensions may be misclassified as normal in Kawasaki disease. J Pediatr 1998;133(2):254-258. [doi: 10.1016/s0022-3476(98)70229-x] [Medline: 9709715]
13. Chaiyarak K, Durongpisitkul K, Atta T, Soongswang J, Laohaprasitiporn D, Nana A. Clinical manifestations of Kawasaki disease: what are the significant parameters? Asian Pac J Allergy Immunol 2009;27(2-3):131-136 [FREE Full text] [Medline: 19839499]
14. Lin KH, Chang SS, Yu CW, Lin SC, Liu SC, Chao HY, et al. Usefulness of natriuretic peptide for the diagnosis of Kawasaki disease: a systematic review and meta-analysis. BMJ Open 2015;5(4):e006703 [FREE Full text] [doi: 10.1136/bmjopen-2014-006703] [Medline: 25872939]
15. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
16. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013;173(6):418-425 [FREE Full text] [doi: 10.1001/jamainternmed.2013.2777] [Medline: 23440149]

XSL•FO
RenderX

17. Kassam AF, Cortez AR, Winer L, Conzen K, El-Hinnawi A, Jones C, et al. Extinguishing burnout: national analysis of predictors and effects of burnout in abdominal transplant surgery fellows. Am J Transplant 2021;21(1):307-313 [FREE Full text] [doi: 10.1111/ajt.16075] [Medline: 32463950]
18. Hodkinson A, Zhou A, Johnson J, Geraghty K, Riley R, Zhou A, et al. Associations of physician burnout with career engagement and quality of patient care: systematic review and meta-analysis. BMJ 2022;378:e070442 [FREE Full text] [doi: 10.1136/bmj-2022-070442] [Medline: 36104064]
19. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
20. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019;188(12):2222-2239. [doi: 10.1093/aje/kwz189] [Medline: 31509183]
21. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017;69(21):2657-2664 [FREE Full text] [doi: 10.1016/j.jacc.2017.03.571] [Medline: 28545640]
22. Dorado-Díaz PI, Sampedro-Gómez J, Vicente-Palacios V, Sánchez PL. Applications of artificial intelligence in cardiology. The future is already here. Rev Esp Cardiol (Engl Ed) 2019;72(12):1065-1075. [doi: 10.1016/j.rec.2019.05.014] [Medline: 31611150]
23. Kim IB, Park SC. Machine learning-based definition of symptom clusters and selection of antidepressants for depressive syndrome. Diagnostics (Basel) 2021;11(9):1631 [FREE Full text] [doi: 10.3390/diagnostics11091631] [Medline: 34573974]
24. Shao J, Ma J, Zhang Q, Li W, Wang C. Predicting gene mutation status via artificial intelligence technologies based on multimodal integration (MMI) to advance precision oncology. Semin Cancer Biol 2023;91:1-15 [FREE Full text] [doi: 10.1016/j.semcancer.2023.02.006] [Medline: 36801447]
25. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group†. Probast: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
26. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019;28(9):2768-2786 [FREE Full text] [doi: 10.1177/0962280218785504] [Medline: 30032705]
27. Yang YL, Kuo HC, Chen KD, Chu CH, Kuo K, Guo MM, et al. Combination of hemoglobin-for-age Z-score and plasma hepcidin identified as a novel predictor for Kawasaki disease. Children (Basel) 2022 Jun 18;9(6):93 [FREE Full text] [doi: 10.3390/children9060913] [Medline: 35740850]
28. Tsai CM, Lin CR, Kuo HC, Cheng FJ, Yu HR, Hung TC, et al. Use of machine learning to differentiate children with Kawasaki disease from other febrile children in a pediatric emergency department. JAMA Netw Open 2023;6(4):e237489 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.7489] [Medline: 37040115]
29. Li C, Liu Y, Zhang D, Han Y, Chen B, Long Y, et al. A machine learning model for distinguishing Kawasaki disease from sepsis. Sci Rep 2023;13(1):12553 [FREE Full text] [doi: 10.1038/s41598-023-39745-8] [Medline: 37532772]
30. Guo X, Liao J, Fan X, Xu M. Exploring the diagnostic value of eosinophil-to-lymphocyte ratio to differentiate Kawasaki disease from other febrile diseases based on clinical prediction model. Sci Rep 2023;13(1):3399 [FREE Full text] [doi: 10.1038/s41598-023-30463-9] [Medline: 36854770]
31. Weng KP, Li SC, Chien KJ, Tsai KW, Kuo HC, Hsieh KS, et al. Prediction model for diagnosis of Kawasaki disease using iTRAQ-Based analysis. Children (Basel) 2021;8(7):576 [FREE Full text] [doi: 10.3390/children8070576] [Medline: 34356555]
32. Tsai CM, Chu CH, Liu X, Weng KP, Liu SF, Huang YH, et al. A novel score system of blood tests for differentiating Kawasaki disease from febrile children. PLoS One 2021;16(1):e0244721 [FREE Full text] [doi: 10.1371/journal.pone.0244721] [Medline: 33481812]
33. Huang YS, Liu XP, Xia HB, Cui LN, Lang XL, Liu CY, et al. Complement 3 and the prognostic nutritional index distinguish Kawasaki disease from other fever Illness with a nomogram. Children (Basel) 2021;8(9):825 [FREE Full text] [doi: 10.3390/children8090825] [Medline: 34572257]
34. Liu XP, Huang YS, Xia HB, Sun Y, Lang XL, Li QZ, et al. A nomogram model identifies eosinophilic frequencies to powerfully discriminate Kawasaki disease from febrile infections. Front Pediatr 2020;8:559389 [FREE Full text] [doi: 10.3389/fped.2020.559389] [Medline: 33363059]
35. Liu XP, Huang YS, Kuo HC, Xia HB, Yi-Sun, Huang WD, et al. A novel nomogram model for differentiating Kawasaki disease from sepsis. Sci Rep 2020;10(1):13745 [FREE Full text] [doi: 10.1038/s41598-020-70717-4] [Medline: 32792679]
36. Huang Z, Tan XH, Wang H, Pan B, Lv T, Tian J. A new diagnostic model to distinguish Kawasaki disease from other febrile illnesses in Chongqing: a retrospective study on 10,367 patients. Front Pediatr 2020;8:533759 [FREE Full text] [doi: 10.3389/fped.2020.533759] [Medline: 33304865]
37. Huang J, Chen J. Bp neural network model for early diagnosis of Kawasaki disease. Advanced materials research 2012;468:723-726 [FREE Full text] [doi: 10.4028/www.scientific.net/amr.468-471.723]
38. Hao S, Jin B, Tan Z, Li Z, Ji J, Hu G, Pediatric Emergency Medicine Kawasaki Disease Research Group. A classification tool for differentiation of Kawasaki disease from other febrile illnesses. J Pediatr 2016;176:114-120.e8 [FREE Full text] [doi: 10.1016/j.jpeds.2016.05.060] [Medline: 27344221]

XSL•FO
RenderX

39. Hao S, Ling XB, Kanegaye JT, Bainto E, Dominguez SR, Heizer H, Pediatric Emergency Medicine Kawasaki Disease Research Group. Multicentre validation of a computer-based tool for differentiation of acute Kawasaki disease from clinically similar febrile illnesses. Arch Dis Child 2020;105(8):772-777. [doi: 10.1136/archdischild-2019-317980] [Medline: 32139365]
40. Portman MA, Magaret CA, Barnes G, Peters C, Rao A, Rhyne R. An artificial intelligence derived blood test to diagnose Kawasaki disease. Hosp Pediatr 2023;13(3):201-210. [doi: 10.1542/hpeds.2022-006868] [Medline: 36775804]
41. Lam JY, Shimizu C, Tremoulet AH, Bainto E, Roberts SC, Sivilay N, Pediatric Emergency Medicine Kawasaki Disease Research Group, CHARMS Study Group. A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the USA: a retrospective model development and validation study. Lancet Digit Health 2022;4(10):e717-e726 [FREE Full text] [doi: 10.1016/S2589-7500(22)00149-2] [Medline: 36150781]
42. Ling XB, Kanegaye JT, Ji J, Peng S, Sato Y, Tremoulet A, et al. Point-of-care differentiation of Kawasaki disease from other febrile illnesses. J Pediatr 2013;162(1):183-188.e3 [FREE Full text] [doi: 10.1016/j.jpeds.2012.06.012] [Medline: 22819274]
43. Sobh A, Mosa DM, Khaled N, Korkor MS, Noureldin MA, Eita AM, et al. How multisystem inflammatory syndrome in children discriminated from Kawasaki disease: a differentiating score based on an inception cohort study. Clin Rheumatol 2023;42(4):1151-1161 [FREE Full text] [doi: 10.1007/s10067-022-06444-0] [Medline: 36409406]
44. Du S, Mansmann U, Geisler BP, Li Y, Hornung R. A diagnostic model for Kawasaki disease based on immune cell characterization from blood samples. Front Pediatr 2021;9:769937 [FREE Full text] [doi: 10.3389/fped.2021.769937] [Medline: 35071130]
45. Castaldo A, D'Anna C, Gelzo M, Giannattasio A, Maglione M, Muzzica S, et al. Immunophenotyping of peripheral blood cells allows to discriminate MIS-C and Kawasaki disease. Transl Med Commun 2022;7(1):22 [FREE Full text] [doi: 10.1186/s41231-022-00128-2] [Medline: 36093039]
46. Kim JM, Kim J. Prediction model for the differential diagnosis of Kawasaki disease and acute cervical lymphadenitis in patients initially presenting with fever and cervical lymphadenitis. J Pediatr 2020;225:30-36.e2. [doi: 10.1016/j.jpeds.2020.05.031] [Medline: 32450069]
47. Yang C, Yang Y, Cao S, Ma Z, Du H, Li J, et al. Kawasaki disease coronary artery lesions prediction with monocyte-to-high-density lipoprotein ratio. Pediatr Res 2023;94(1):246-251. [doi: 10.1038/s41390-022-02401-4] [Medline: 36446921]
48. Wu R, Jiang W, Sun Y, Wu L, Di Y, Wang J, et al. Indicators of oxidative stress in the prediction of coronary artery lesions in patients with Kawasaki disease. J Clin Rheumatol 2023;29(3):126-131. [doi: 10.1097/RHU.0000000000001925] [Medline: 36730421]
49. Chen J, Li J, Yue Y, Liu Y, Xie T, Peng JQ, et al. Nomogram for predicting coronary artery lesions in patients with Kawasaki disease. Clin Cardiol 2023;46(11):1434-1441 [FREE Full text] [doi: 10.1002/clc.24113] [Medline: 37540643]
50. Yu H, Sun W, Wu H, Zhang S, Xu Z, Wu R, et al. The significance of symptoms in predicting coronary artery aneurysms of Kawasaki disease, especially in female patients. Front Cardiovasc Med 2022;9:823862 [FREE Full text] [doi: 10.3389/fcvm.2022.823862] [Medline: 35571186]
51. Wang L, Yang Y, Cui Q, Cui Y, Li Q, Che X, et al. Evaluating the added predictive ability of MMP-9 in serum for Kawasaki disease with coronary artery lesions. J Investig Med 2021;69(1):13-19. [doi: 10.1136/jim-2020-001281] [Medline: 33004469]
52. Azuma J, Yamamoto T, Nitta M, Hasegawa Y, Kijima E, Shimotsuji T, et al. Structure equation model and neural network analyses to predict coronary artery lesions in Kawasaki disease: a single-centre retrospective study. Sci Rep 2020;10(1):11868 [FREE Full text] [doi: 10.1038/s41598-020-68657-0] [Medline: 32681105]
53. Morikawa Y, Ohashi Y, Harada K, Asai T, Okawa S, Nagashima M, et al. Coronary risks after high-dose gamma-globulin in children with Kawasaki disease. Pediatr Int 2000;42(5):464-469. [doi: 10.1046/j.1442-200x.2000.01288.x] [Medline: 11059532]
54. Grasa CD, Fernández-Cooke E, Domínguez-Rodríguez S, Aracil-Santos J, Barrios Tascon A, Sánchez-Manubens J, Spanish group for the study of Kawasaki Disease (KAWA-RACE). Risk scores for Kawasaki disease, a management tool developed by the KAWA-RACE cohort. Clin Rheumatol 2022;41(12):3759-3768. [doi: 10.1007/s10067-022-06319-4] [Medline: 35939163]
55. Son MBF, Gauvreau K, Tremoulet AH, Lo M, Baker AL, de Ferranti S, et al. Risk model development and validation for prediction of coronary artery aneurysms in Kawasaki disease in a North American population. J Am Heart Assoc 2019;8(11):e011319 [FREE Full text] [doi: 10.1161/JAHA.118.011319] [Medline: 31130036]
56. Zheng X, Li Y, Yue P, Ma F, Zhang Y, Wu G. Diagnostic significance of circulating miRNAs in Kawasaki disease in China: current evidence based on a meta-analysis. Medicine (Baltimore) 2021;100(6):e24174 [FREE Full text] [doi: 10.1097/MD.0000000000024174] [Medline: 33578520]
57. Zhong X, Jia X, Wang H, Chen G, Li H, Li P, et al. Diagnostic significance of noncoding RNAs in kawasaki disease: a systematic review and meta-analysis. Front Pediatr 2022;10:1071434 [FREE Full text] [doi: 10.3389/fped.2022.1071434] [Medline: 36727009]
58. Xie X, Shi X, Liu M. The roles of genetic factors in Kawasaki disease: a systematic review and meta-analysis of genetic association studies. Pediatr Cardiol 2018;39(2):207-225. [doi: 10.1007/s00246-017-1760-0] [Medline: 29098351]

XSL•FO
RenderX

59. Peng Q, Chen C, Wu Q, Yang Y. Association of new functional SNP rs72689236 of CASP3 with Kawasaki disease: a meta-analysis. Zhongguo Dang Dai Er Ke Za Zhi 2013;15(6):477-483 [FREE Full text] [Medline: 23791066]

Abbreviations

AI: artificial intelligence
C-index: concordance index
CAL: coronary artery lesion
CRP: C-reactive protein
ESR: erythrocyte sedimentation rate
IVIG: intravenous immunoglobulin
KD: Kawasaki disease
ML: machine learning
NEUT: neutrophils
PLT: platelet
PROBAST: the Prediction Model Risk of Bias Assessment Tool
WBC: white blood cell

Edited by A Schwartz; submitted 22.02.24; peer-reviewed by Q Xia, N Goyal, N Goyal; comments to author 24.06.24; revised version received 25.07.24; accepted 20.09.24; published 18.11.24.

Please cite as:
Zhu J, Yang F, Wang Y, Wang Z, Xiao Y, Wang L, Sun L
Accuracy of Machine Learning in Discriminating Kawasaki Disease and Other Febrile Illnesses: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e57641
URL: https://www.jmir.org/2024/1/e57641
doi:10.2196/57641
PMID:

©Jinpu Zhu, Fushuang Yang, Yang Wang, Zhongtian Wang, Yao Xiao, Lie Wang, Liping Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Implementation of Machine Learning Applications in Health Care Organizations: Systematic Review of Empirical Studies

Luigi M Preti[1], MSc; Vittoria Ardito[1], MSc; Amelia Compagni[1,2], PhD; Francesco Petracca[1], MSc; Giulia Cappellaro[1,2], PhD

[1]Center for Research on Health and Social Care Management (CERGAS), SDA Bocconi School of Management, Milan, Italy

[2]Department of Social and Political Sciences, Bocconi University, Milan, Italy

Corresponding Author:
Giulia Cappellaro, PhD
Department of Social and Political Sciences
Bocconi University
Via Sarfatti 25
Milan, 20136
Italy
Phone: 39 02 58365267
Email: giulia.cappellaro@unibocconi.it

Abstract

Background: There is a growing enthusiasm for machine learning (ML) among academics and health care practitioners. Despite the transformative potential of ML-based applications for patient care, their uptake and implementation in health care organizations are sporadic. Numerous challenges currently impede or delay the widespread implementation of ML in clinical practice, and limited knowledge is available regarding how these challenges have been addressed.

Objective: This work aimed to (1) examine the characteristics of ML-based applications and the implementation process in clinical practice, using the Consolidated Framework for Implementation Research (CFIR) for theoretical guidance and (2) synthesize the strategies adopted by health care organizations to foster successful implementation of ML.

Methods: A systematic literature review was conducted based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The search was conducted in PubMed, Scopus, and Web of Science over a 10-year period (2013-2023). The search strategy was built around 4 blocks of keywords (artificial intelligence, implementation, health care, and study type). Only empirical studies documenting the implementation of ML applications in clinical settings were considered. The implementation process was investigated using a thematic analysis and coding procedure.

Results: Thirty-four studies were selected for data synthesis. Selected papers were relatively recent, with only 9% (3/34) of records published before 2019. ML-based applications were implemented mostly within hospitals (29/34, 85%). In terms of clinical workflow, ML-based applications supported mostly prognosis (20/34, 59%) and diagnosis (10/34, 29%). The implementation efforts were analyzed using CFIR domains. As for the inner setting domain, access to knowledge and information (12/34, 35%), information technology infrastructure (11/34, 32%), and organizational culture (9/34, 26%) were among the most observed dimensions influencing the success of implementation. As for the ML innovation itself, factors deemed relevant were its design (15/34, 44%), the relative advantage with respect to existing clinical practice (14/34, 41%), and perceived complexity (14/34, 41%). As for the other domains (ie, processes, roles, and outer setting), stakeholder engagement (12/34, 35%), reflecting and evaluating practices (11/34, 32%), and the presence of implementation leaders (9/34, 26%) were the main factors identified as important.

Conclusions: This review sheds some light on the factors that are relevant and that should be accounted for in the implementation process of ML-based applications in health care. While the relevance of ML-specific dimensions, like trust, emerges clearly across several implementation domains, the evidence from this review highlighted that relevant implementation factors are not necessarily specific for ML but rather transversal for digital health technologies. More research is needed to further clarify the factors that are relevant to implementing ML-based applications at the organizational level and to support their uptake within health care organizations.

Trial Registration: PROSPERO 403873; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=403873

International Registered Report Identifier (IRRID): RR2-10.2196/47971

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55897) doi:10.2196/55897

KEYWORDS

artificial intelligence; machine learning; implementation; health care organization; barriers; facilitators

Introduction

Background

Artificial intelligence (AI) has been unquestionably acknowledged as a game changer in health care [1], even more so after technological advances in the field of machine learning (ML) have contributed to further expand the frontiers of its possible applications [2]. Compared to knowledge- or rule-based systems that automate established human clinical reasoning methods through a series of "if-then" statements [3], ML encompasses all nonknowledge-based models that automatically (or semiautomatically) learn from the exposure to abundant quantities of data and detect patterns through explicit or latent recognition rather than conventional programming. ML is expected to serve primarily as a decision support tool to enhance human work rather than replace it [4], thereby providing health care professionals (HCPs) with improved predictions and rendering their decision-making process more accurate [5]. Although some AI systems have already been shown to be equal or even superior in performance to HCPs [6], full automation of a broad range of human tasks is expected to occur only at later stages.

Irrespective of whether ML is intended to provide inputs to human decision-making or to act autonomously, these technological advancements do not automatically translate into clinical practice. The road to implementing ML applications in patient care has several challenges, creating an inevitable chasm between ML and its clinical integration [7,8].

Challenges for the implementation of AI systems, without an exclusive focus on ML, have been previously outlined, with various interdependent factors at different stakeholder group levels [9,10]. For HCPs, core considerations pertain to the need for ML outputs to be meaningful inputs in their decision-making and be explainable. ML algorithms are often associated with the so-called "black box" effect [11,12]. The lack of transparency in data and outputs can be a significant concern for HCPs, as it hampers model interpretability (ie, the possibility to understand or interpret how a given output has been produced) and explainability (ie, the capacity of a model to be explained, even if not totally interpretable) [13]. ML applications and outputs are therefore likely to clash with the principles of evidence-based medicine, which instead involve the highest possible standards of interpretability and explainability. Concerns about the potential implications for accountability and personal responsibility regarding mistakes or computational misdiagnosis by ML applications present additional implementation challenges.

At the patient level, fair implementation of ML applications necessitates continuous supply of standardized data to train, validate, and incessantly improve performance and prevent algorithmic bias [9]. Notions of patient confidentiality and privacy should be reimagined entirely as data must be shared across multiple institutions to maximize their value and allow for improved algorithms [14].

Lastly, distinctive implementation challenges have been identified at the level of health care organizations, which are associated with financial challenges and funding mechanisms, as well as issues related to the computational resources that are necessary to support the implementation of ML.

Several implementation frameworks for health care technologies are on hand, but no widely recognized model addresses all the specific issues that are relevant to ML applications [15-17]. To date, research on ML implementation has been predominantly conceptual in nature, with an underreporting of empirical investigations into the specifics and consequences of implementation processes in real-life settings [18,19]. Available studies have primarily focused on the quantitative impact of ML algorithms on health outcomes or accuracy, without examining the corresponding implementation processes [20]. Recently, Chomutare et al [21] conducted a scoping review to identify barriers and facilitators to the implementation of ML from empirical studies, while Tricco et al [22] focused on the strategies adopted to implement ML tools in hospital settings. However, additional inquiry is needed to determine whether the literature on the implementation of ML applications in health care adequately acknowledges the unique challenges encountered along the implementation process, as well as the strategies adopted to overcome them.

Research Objectives

This systematic literature review primarily aims to identify studies on the real-life implementation of ML applications in clinical practice and to synthesize insights about the features of these innovations and the processes deployed to facilitate their effective implementation. We set out to address the following research questions:

1. What are the characteristics of ML applications implemented in clinical practice as reported in the scientific literature?
2. What processes and strategies do health care organizations employ to foster the successful implementation of ML applications in clinical practice? Which factors are recognized as more relevant for the unsuccessful implementation of ML applications?

Methods

Overview

This systematic review adopted the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Multimedia Appendix 1) [23]. The review was previously registered within the International Prospective Register of Systematic Reviews (PROSPERO) with registration number 403873. All methodological details are provided in the published research protocol [24]. The most relevant aspects are

XSL•FO
RenderX

summarized hereafter, with any deviations from the protocol duly noted.

Positionality of the Research Team

Positionality refers to how individuals identify with and relate to different social dimensions such as gender, race, and ethnicity [25], and as such is a relevant aspect to consider in qualitative research. To that end, the research team comprised 5 Italian white researchers (LMP, VA, AC, FP, and GC). Broadly speaking, the team as a whole included 60% females and 40% males and shared a common background in management studies with a focus on health care management. LMP is a PhD student who is working in the areas of AI and ML under the perspective of the organizational implementation of AI tools in health care organizations. VA is a PhD student who has conducted prior research at the intersection between digital health and implementation science. AC has multi-annual experience in organizational studies and qualitative research focusing on issues related to innovations in health care and professional dynamics in health care organizations. FP is a PhD student who is an expert in digital health technologies, focusing on their regulation and value assessment. GC has multi-annual experience in organizational studies and qualitative research focusing on institutional dynamics, novel technologies, and professions.

Eligibility Criteria

This review focused on empirical studies investigating aspects related to the implementation of ML applications within health care organizations. We adopted the definition of implementation as an "active and planned effort to mainstream innovation within an organization" [26], while health care organizations encompassed all entities delivering health services, including hospitals, outpatient centers, primary care facilities, and public health institutions. Studies were selected based on the eligibility criteria defined in the research protocol [24] and summarized in Textbox 1. The recently updated version of the Consolidated Framework for Implementation Research (CFIR), a commonly used model to assess factors influencing implementation and to explain barriers and facilitators to implementation effectiveness [27,28], was used as a criterion for inclusion. Specifically, only studies that explicitly reported factors related to the CFIR domain of inner setting or process were considered eligible for inclusion (Textbox 2).

Textbox 1. Eligibility criteria.

Inclusion criteria

Study design: Empirical studies illustrating the implementation of machine learning (ML)-based applications (eg, experimental/quasiexperimental, observational, hybrid, or simulation study designs, qualitative designs, case studies, etc)

Intervention: Analysis of the implementation of ML-based applications by at least covering factors related to the inner setting or process domain based on the Consolidated Framework on Implementation Research (CFIR)

Stakeholder groups: ML-based applications used at least by health care professionals (HCPs)

Setting: Hospitals, outpatients, and other community care settings

Timeframe: Studies published from 2013 until March 2023

Exclusion criteria

Study design: Effectiveness research study designs, literature reviews, commentaries, editorials, opinion articles, study protocols, studies collecting perceptions on implementation, and studies unrelated to specific ML-based applications

Intervention: Analysis of the implementation of logic- or knowledge-based applications (eg, expert systems) or ML-based applications with no considerations related to the inner setting or process domain

Stakeholder groups: ML-based applications targeting patients and other nonclinical stakeholders (eg, caregivers, policy makers, and regulators) only

Setting: All other settings, including home care

Timeframe: Studies published before 2013

Textbox 2. Domains of the Consolidated Framework for Implementation Research (CFIR).

Innovation: Domain that collects the characteristics of the implemented object from a multi-faceted point of view.

Outer setting: Domain designed to capture factors that are inherent in the context where the organization exists.

Inner setting: Domain that encompasses the characteristics of the organization in which the innovation is implemented. It includes both structural attributes, which characterize the inner setting regardless of the implementation, and features, which are specific to the implementation.

Roles: Domain that refers to the individuals who have significantly contributed to the implementation and their characteristics.

Implementation process: Domain that collects all the information on the activities and strategies adopted to concretely implement the innovation.

Information Sources

Literature searches were conducted in MEDLINE (PubMed), Scopus, and Web of Science and replicated in top-tier management journal databases. In addition, the reference lists of all included studies and of the reviews identified were scanned to ensure comprehensive coverage of relevant literature. Grey literature was not considered.

XSL•FO
RenderX

Search Strategy

The search strategy was developed by the research team through an iterative process and was based on 4 main concepts: (1) AI; (2) implementation; (3) health care; and (4) study design. Multimedia Appendix 2 presents the search strings used for each database. The general term "artificial intelligence" was used broadly to encompass studies that address AI and ML as synonymous terms. The search was performed in April 2023.

Study Selection and Data Collection Process

Two researchers (VA and LMP) screened the first 100 retrieved studies based on titles and abstracts. Once alignment over the inclusion and exclusion criteria was reached, the remaining records were independently screened by the 2 reviewers in equal parts based on the title and abstract. Disagreements over final inclusion were solved by a third researcher (FP). Studies deemed eligible for full-text reading were assessed in-depth (VA, LMP, and FP). Disagreements were resolved by dialogue with 2 additional researchers (GC and AC). The entire research team read all the studies included in the analysis. The data collection process was performed by 3 reviewers (VA, LMP, and FP) who extracted data using an ad hoc Microsoft Excel sheet preliminarily developed by the research team. To ensure consistency across reviewers, the extraction sheet was tested by each reviewer and recalibrated before starting the data collection process. Any disagreements were resolved by discussion with the research team, with final decisions reached by consensus.

Data Items

Data items were extracted based on established classifications or schemes, when applicable [24]. These encompassed information on the paper (eg, journal of publication and publication year), ML application (eg, name, brief description, main practice of use, level of autonomy, and degree of integration with other technologies), and implementation process (eg, stage of implementation, geographical location, care setting, and specific unit of implementation). Furthermore, factors influencing the implementation process were assessed following the 5 domains of the updated version of the CFIR.

Quality Assessment

Critical appraisal of the studies selected for data synthesis was performed using the Mixed Method Appraisal Tool (MMAT [29]), which has been designed specifically for systematic reviews that include heterogeneous studies, as it allows to assess the methodological quality of 5 types of study designs (ie, qualitative studies, randomized controlled trials, nonrandomized studies, quantitative descriptive studies, and mixed methods studies). Quality appraisal was performed by 2 researchers (VA and LMP), and disagreements were discussed and solved. The quality assessment represents a deviation from the protocol, which did not include this step.

Data Synthesis

Given the significant heterogeneity across study designs, research objectives, and outcomes observed, as well as the expected predominance of qualitative studies, we opted for a thematic synthesis approach to capture and synthesize the salient attributes of the implementation process based on the CFIR constructs [30,31]. The analysis considered findings from the data extraction process as qualitative data and included summaries and interpretation of findings from the authors of the reviewed studies. Hence, direct quotes from participants were excluded in cases where the study employed qualitative data collection methods (eg, interviews).

We used both an inductive and a deductive approach. Following the 3 thematic synthesis steps, we initially reviewed each paper and highlighted relevant aspects through line-by-line coding to capture and collect key data. The coding process involved 3 reviewers (VA, LMP, and FP). To identify recurring topics, primary codes were then compared, organized, and labeled to derive descriptive themes reflecting their meaning. Descriptive themes were used to develop higher-level analytical themes. The formulation of descriptive themes and the following assignation to analytical themes were initially proposed by a researcher (LMP) and iteratively refined through discussions with 2 other researchers (VA and FP).

The higher-level analytical themes were subsequently deductively redefined by the entire research team within the constructs of the CFIR, which served as the final theoretical framework guiding our analysis.

Results

Study Selection

We retrieved 3520 unique records that were initially screened based on the titles and abstracts. A total of 67 records were deemed eligible for full-text screening (67/3520, 1.9%). Additionally, we identified 36 eligible records from a manual search of reference lists of excluded literature reviews and full-text screened records. Out of the 103 papers analyzed in full text, 69 were excluded and 34 were included in the review (34/103, 33.0%). The primary reason for exclusion was the focus of the intervention analyzed in the papers (53/69, 77%), as they either had a clinical or technical purpose without addressing factors related to implementation in an organizational setting or involved non-ML–based applications. Figure 1 provides an overview of the selection process and the reasons for exclusion.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

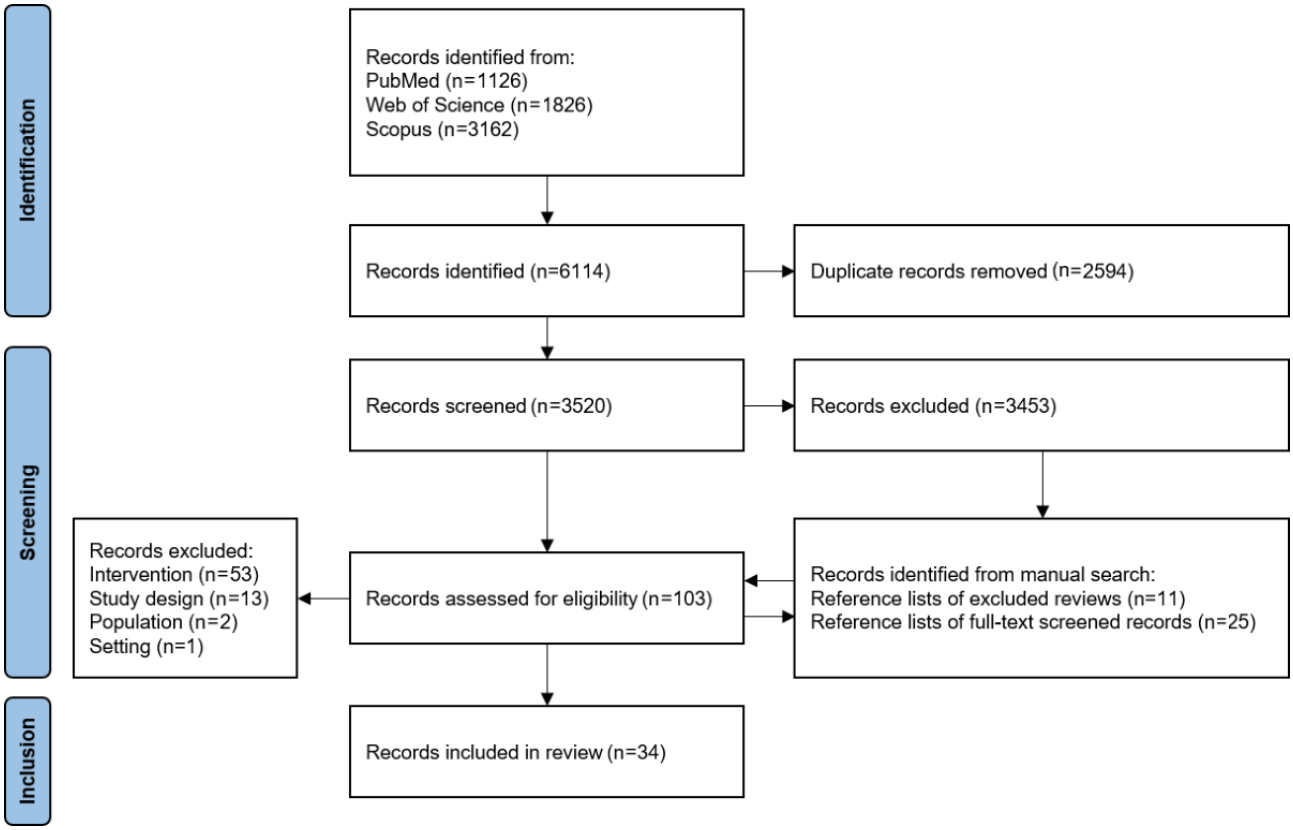

Study Characteristics

Most of the studies documenting the implementation of ML-based applications were set in the United States (18/34, 53%). Other locations included China (4/34, 12%); Canada, Brazil, and the Netherlands (each 2/34, 6%); and Italy, Spain, Norway, Korea, India, and Austria (each 1/34, 3%). Papers selected for data synthesis were relatively recent, with only 3 out of 34 (9%) published before 2019. Outlets were mostly clinical or in the field of information technology (IT) (30/34, 88%), while the remaining (4/34, 12%) focused on managerial or organizational studies. Most of the selected studies followed qualitative or mixed methods designs (22/34, 65%), often relying on methods such as interviews and case studies.

Quality Assessment in Studies

Quality appraisal of the selected studies was performed using the MMAT. The studies were heterogeneous in terms of study design, and different MMAT questions were used to assess their quality. Overall, 18 studies leveraged the questions of qualitative studies, 7 of quantitative nonrandomized studies, 5 of quantitative descriptive studies, 3 of mixed methods studies, and 1 of quantitative randomized studies. Overall, the quality assessment suggested a medium-good quality of the studies, with only 12.5% of the assessment questions uncertain or unclear ("Can't tell"). The detailed output of the quality appraisal is provided in Multimedia Appendix 3.

Characteristics of ML Applications

Table 1 provides a general description of the ML-based applications implemented in the selected studies [32-65], while Multimedia Appendix 4 provides more detailed information on the characteristics of these applications. The most recurrent applications comprised predictive modeling algorithms, visualization tools, and alert-delivering mechanisms. All the applications identified by our search were clinical practice applications, according to the definitions from the European Parliamentary Research Service [66]. Moreover, none of the applications had decisional autonomy; therefore, all systems could be classified as clinical decision support systems (CDSSs).

XSL•FO
RenderX

Table 1. Overview of machine learning applications.

Authors	Year	Application name	Application output description	Implementation setting (unit)	Clinical workflow activity
Lee et al [32]	2015	—[a]	Prediction of patient characteristics, complaint types, and admission and readmission patterns	Hospital (ED[b])	Prognosis
Hengstler et al [33]	2016	IBM Watson	Building hypotheses and evidence on cancer diagnosis	Hospital (oncology)	Diagnosis
McCoy & Das [34]	2017	InSight	Prediction of the risk of developing severe sepsis	Community hospital (ED, ICU[c])	Prognosis
Bhattacharya et al [35]	2019	Niramai Thermalytix and iBreastExam (iBE)	Earlier detection of breast cancer	Hospital (radiology)	Diagnosis
Cruz et al [36]	2019	Savana	Recommendations for improving adherence to health care pathways	Primary care	Clinical/organizational workflow
Ginestra et al [37]	2019	EWS 2.0	Prediction of the risk of developing sepsis	Teaching hospital (non-ICU setting)	Prognosis
Gonçalves et al [38]	2020	Laura	Prediction of the risk of developing sepsis	Hospital (several units)	Prognosis
Sun & Medaglia [39]	2019	IBM Watson for Oncology	Decision-making support for personalized treatment planning	Hospital (oncology)	Treatment
Baxter et al [40]	2020	—	Prediction of unplanned readmission	Teaching and research hospital (unspecified)	Prognosis
Cho et al [41]	2020	DEWS (Deep-Learning-based Early Warning System)	Prediction of in-hospital cardiac events	Hospital (cardiology)	Prognosis
Frontoni et al [42]	2020	—	Production of indicators for quality-of-care processes of T2D[d]	Primary care	Clinical/organizational workflow
Hassan et al [43]	2020	Viz.ai	Detection of large vessel occlusions	Hospital (stroke unit)	Diagnosis
Romero-Brufau et al [44]	2020	—	Prediction of hospital readmission and formulation of targeted recommendations	Hospital (all units)	Prognosis and treatment
Sandhu et al [45]	2020	Sepsis Watch	Prediction of the risk of developing sepsis	Teaching hospital (ED)	Prognosis
Sendak et al [46]	2020	Sepsis Watch	Prediction of the risk of developing sepsis	Teaching hospital (ED)	Prognosis
Strohm et al [47]	2020	BoneXpert	Assessment of child maturation and bone age, and prediction of adult height	Hospital (radiology)	Diagnosis and prognosis
Xu et al [48]	2020	SensEcho	Classification of sleep stage, detection of sleep apnea, and recognition of abnormal ECG[e] signals from a multi-sensor wearable device	Hospital (general and respiratory)	Diagnosis
Jauk et al [49]	2021	—	Prediction of the risk of developing delirium	Hospital (surgery, internal medicine)	Prognosis
Morales et al [50]	2021	Laura Digital ER	Detection of COVID-19 symptoms	Community	Diagnosis
Murphree et al [51]	2021	—	Treatment optimization and identification of likely-to-benefit patients for palliative care	Hospital (all inpatient units)	Treatment
Yao et al [52]	2021	3D CSAC-Net	Detection of mild COVID-19 pneumonia	Hospital (unspecified)	Diagnosis
Davis et al [53]	2022	Aidoc	Prediction of the risk of developing intracranial hemorrhage	Research hospital (radiology)	Prognosis

XSL•FO
RenderX

Authors	Year	Application name	Application output description	Implementation setting (unit)	Clinical workflow activity
Henry et al [54]	2022	TREWS	Prediction of the risk of developing sepsis	Acute care hospital (inpatient acute units and ED)	Prognosis
Joshi et al [55]	2022	—	Prediction of the risk of developing sepsis	Community and teaching hospitals (several units)	Prognosis
Lebovitz et al [56]	2022	—	Image processing, segmentation, and classification for imaging diagnostics	Teaching hospital (radiology)	Diagnosis
Rushlow et al [57]	2022	—	Prediction of the risk of low left ventricular ejection fraction	Primary care	Prognosis
Schwartz et al [58]	2022	CONCERxN	Prediction of the risk of in-hospital deterioration	Teaching hospital (acute units and ICU)	Prognosis
Sibbald et al [59]	2022	Isabel	Differential diagnosis	Teaching hospital (ED)	Diagnosis
Singer et al [60]	2022	Low Bed Tool and Readmission Risk Tool	Prediction of reduced bed availability and prediction of the risk of readmission	Hospital (ICU, surgery, pediatrics)	Clinical/organizational workflow and prognosis
Wijnhoven [61]	2022	Sepsis Identification Speed	Prediction of the risk of developing sepsis	Teaching hospital (neonatology)	Prognosis
Zhai et al [62]	2022	Nu-CDSS	Formulation of recommendations for nurses' diagnoses, interventions, and outcome evaluations	Teaching hospital (unspecified)	Clinical/organizational workflow
Pou-Prom et al [63]	2022	CHARTwatch	Early warning system designed to predict patient risk of clinical deterioration	Teaching hospital (general internal medicine)	Prognosis
Hinson et al [64]	2022	—	Estimation of the short-term risk for clinical deterioration in patients with or under investigation for COVID-19	Teaching hospital (ED)	Prognosis
Berge et al [65]	2023	Information System for Clinical Concept-based Search	Detection and classification of patient allergies	Hospital (anesthesia, ICU)	Diagnosis

[a]Not applicable.

[b]ED: emergency department.

[c]ICU: intensive care unit.

[d]T2D: type 2 diabetes.

[e]ECG: electrocardiogram.

In terms of settings, ML-based applications were mostly implemented within hospitals (29/34, 85%), including general, university, or teaching hospitals, academic medical centers, and research centers. A few studies (4/34, 12%) were based in a community or primary care setting. Within hospital settings, the most recurring implementation units were emergency departments (EDs) (11/34, 32%) and critical care units such as intensive care units (ICUs) (4/34, 12%), while in some studies, implementation occurred in multiple units or at the hospital level (5/34, 15%).

The clusters identified by Rajkomar et al [67] were used as a theoretical guide to classify the clinical workflow activities in which the ML-based applications were used. In 20 studies (59%), the ML tools supported prognosis. Many of these applications were designed to predict the risk of developing specific conditions such as sepsis (8/34, 24%), in-hospital deterioration (3/34, 9%), intracranial hemorrhage (1/34, 3%), or heart failure (1/34, 3%). Other applications predicted the risk of unplanned hospital admission or readmission (4/34, 12%). Ten papers (29%) illustrated applications for diagnosis, either as standalone computer vision tools to detect diseases from diagnostic imaging (eg, pneumonia from computed tomography [CT] scans, large vessel occlusions from CT angiograms, and child maturation from x-rays) or as diagnostic supports in emergency physician triage. Three papers (9%) illustrated applications for treatment optimization and personalization.

ML capabilities relate to clinical workflow activities, with forecasting (ie, the ability to find complex patterns in data and make predictions) being the most prevalent capability (19/34, 56%), as this function is typical of tools that predict the risk of an adverse event (12 of 34 forecasting tools were for prognosis). Computer vision was exclusively included in algorithms for diagnostic purposes, with all 6 computer vision tools intended for diagnosis.

As for the level of integration with existing technologies, 17 ML-based algorithms (50%) were embedded in electronic health

XSL•FO
RenderX

records (EHRs) or similar platforms (ie, add-ons to the EHR software in use). Fourteen algorithms (41%) were standalone applications, fed either with internal or external data, including images or text. One application (3%) was embedded in hospital hardware technology, namely scanner machines [34]. Computer vision applications were always standalone applications provided as software to be installed within existing hardware (ie, hospital computers) and integrated with local picture archiving and communication systems (PACS).

The ownership of the algorithms was also assessed, revealing a division between applications purchased from commercial vendors (14/34, 41%) and those developed internally (12/34, 35%). The latter algorithms were often linked with the organizational setting, as 6 of these studies were carried out in teaching hospitals, academic medical centers, or research centers. Externally purchased applications were more common in other settings and exhibited greater diversity in terms of purposes, while homegrown tools were generally intended for prognostic purposes. In 8 studies, information on the name or the development process of the application was irretrievable, preventing the determination of algorithm ownership.

Details on the specific ML models employed were often missing, although it was possible to infer that 20 of the analyzed studies (59%) were based on supervised learning models such as random forest, decision tree, and logistic regression.

Implementation Process Characteristics

This section presents the results of the thematic analysis, discussed following the 5 domains of the CFIR, namely innovation, outer setting, inner setting, roles, and implementation process. From the 34 studies analyzed, 222 quotes were extracted. Quotes were organized in 167 descriptive themes and 42 analytical themes. Analytical themes were finally embedded into 23 CFIR constructs. The detailed results of the coding process are presented in Multimedia Appendix 5. To provide a simplified overview of the coding process, Table 2 summarizes the analytical themes, their correspondence with CFIR constructs, and relative frequencies. The results are reported according to the frequency of information extracted on CFIR domains. The relative importance of CFIR constructs is presented in Figure 2.

XSL•FO
RenderX

Table 2. Analytical themes, constructs, and domains of the Consolidated Framework for Implementation Research (CFIR).

Construct	Analytical themes	Papers, n (%)	References
Inner setting domain (n=25, 74%)			
A. Structural characteristics (A.2 IT infrastructure)	Integration with existing IT; Data governance; System infrastructure	11 (32)	[33,36,38,39,47,48,50,57,58,60,61]
D. Culture	Professional habits; User perceptions	9 (26)	[32,37-40,45,47,59,62]
F. Compatibility	Local workflow adaptation	7 (21)	[40,44,47,49,62,63,65]
H. Incentive systems	Economic incentives; Organizational incentives	2 (6)	[42,56]
I. Mission alignment	Organizational strategy; Organizational support	4 (12)	[39,47,55,62]
J. Available resources	Resource reallocation	1 (3)	[32]
K. Access to knowledge & information	Skills	12 (35)	[32,34,38-40,45,48,49,55,57,63,64]
Innovation domain (n=22, 65%)			
A. Innovation source	Trust in the innovation source	3 (9)	[33,39,44]
B. Innovation evidence base	Empirical evidence on added value	2 (6)	[47,49]
C. Innovation relative advantage	Performance trust; Perceived cons; Perceived benefits	14 (41)	[37,39,40,43-45,47,55,56,58,59,61,62,65]
E. Innovation trialability	Testing period	1 (3)	[63]
F. Innovation complexity	Explainability	14 (41)	[32,33,37,39,44,45,49,54-56,58,59,61,65]
G. Innovation design	Complementarity; Ease of use; Risks	15 (44)	[32,33,39,45,49,52-56,58,59,62,64,65]
Process domain (n=22, 65%)			
E. Tailoring strategies	Framing; Tailoring	9 (26)	[33,38,45-47,54,55,57,62]
F. Engaging	Early involvement of end-users; Professional buy-in; Iterative development	12 (35)	[33,38,45,46,51,55,57,58,60-63]
H. Reflecting & evaluating	Feedback	11 (32)	[34,36,38,45,46,48,49,60-62,65]
I. Adapting	Local data; Adaptability	6 (18)	[36,39,41,44,51,58]
Individuals domain – Roles subdomain (n=11, 32%)			
E. Implementation leads	Implementation lead; Implementation team	9 (26)	[34,40,45-47,55,61-63]
F. Implementation team members	Interdisciplinary teams	7 (21)	[45,46,51,55,61,62,64]
Outer setting domain (n=9, 26%)			
B. Local attitudes	Patient acceptance; Public attitude	4 (12)	[33,35,39,50]
D. Partnership & connections	Interinstitutional partnerships; Public-private partnerships	4 (12)	[42,47,50,61]
E. Policies & laws	Medicolegal issues; Medical device regulation; Guidelines; Data protection	6 (18)	[33,39,46,47,50,61]
G. External pressure	Peer influence	1 (3)	[54]

XSL•FO
RenderX

Figure 2. Relative importance of Consolidated Framework for Implementation Research (CFIR) constructs. The percentages represent the proportions of papers in which each construct and domain appears out of the 34 included in the review. The size of the bubbles corresponds to the frequency of occurrence of each construct.

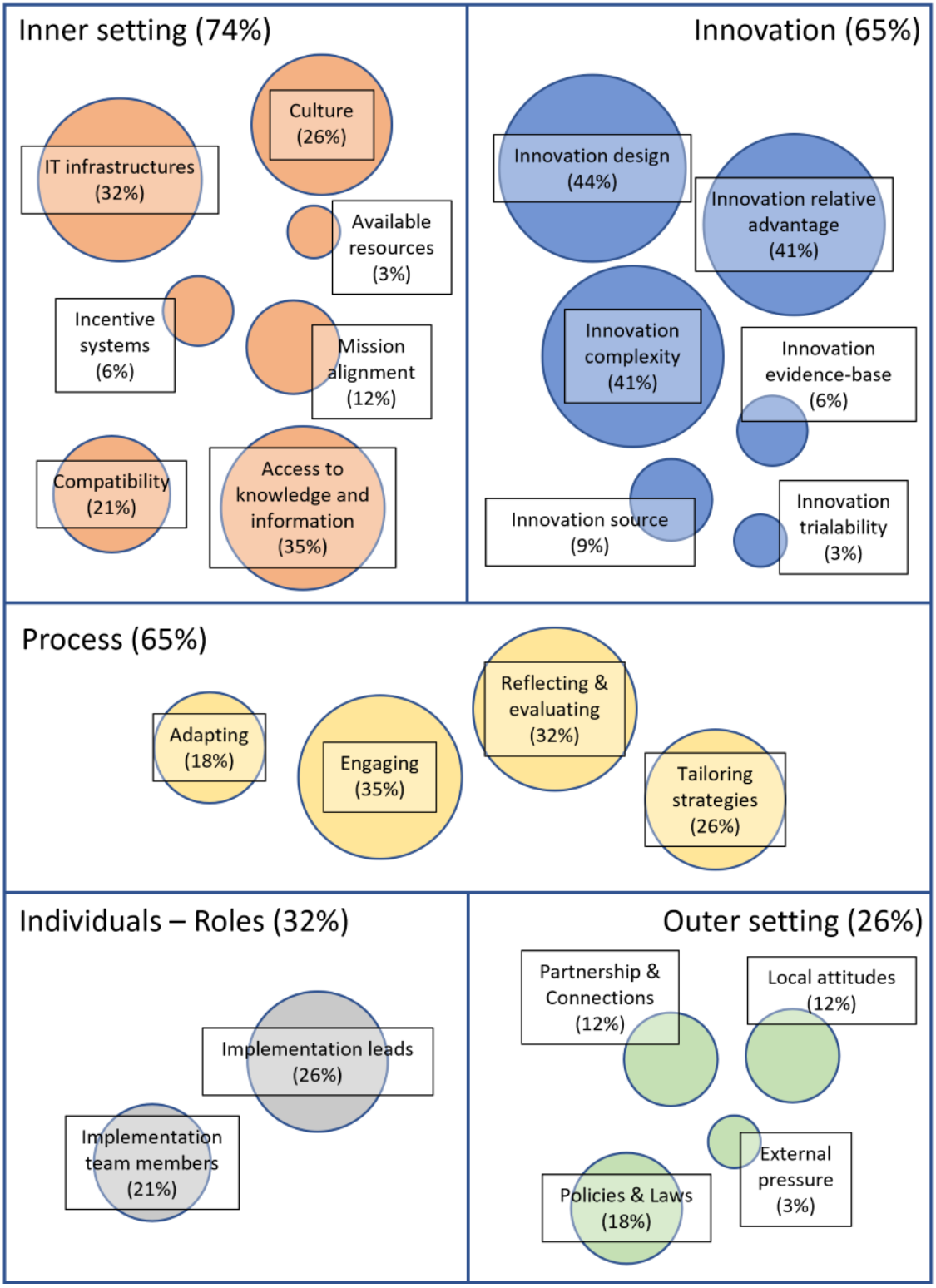

Inner Setting

The inner setting domain was the most frequently described, with 25 studies mentioning at least one construct from this domain as relevant to explaining the implementation process of the ML application. The most recurrent constructs were access to knowledge and information (12/34, 35%), IT infrastructure (11/34, 32%), and culture (9/34, 26%).

First, the *access to knowledge* construct aligned with the topic of skills. Studies emphasized the importance of providing end users with access to training programs on both hard and soft skills before implementation [48,49,55,64], including computer and technical literacy linked with the complexity of the application's functioning [57,63], and the medical domain that the application addresses [45]. The latter referred to dimensions,

XSL•FO
RenderX

such as communication, empathy, and ability to listen, especially when different HCPs were involved in the implementation process [38,39,45].

Second, the *IT infrastructure* construct encompassed 2 prominent themes. The first broadly concerned data management and data governance. Themes, such as data collection and quality [39,58,59], security [33], availability [36,38], and sharing [39,61], were highly described as challenges for the adoption of the application. There were also significant references to building IT infrastructure [61] and to the need to integrate new technologies with existing IT systems (eg, EHRs). While integration promoted ease of use by reducing the need for manual inputs [47,48,50], some argued for the ML application not to directly populate EHRs in order to preserve HCP autonomy and prevent medicolegal accountability [59].

Finally, the construct of *culture* was articulated into the themes of professional habits and alignment of perceptions among stakeholders. The impact of introducing ML applications on professional habits was significant as this affected how professionals work, interact, and make decisions [62]. For instance, the habit of working without technological support was considered a barrier to implementation [39,40]. The need for gradual changes in professional habits was seen as a factor that could hinder the adoption of ML applications in settings with high job rotations (eg, teaching hospitals) [32]. Additionally, ML applications often are not tailored to local workflows and do not consider the different approaches of professionals in diverse contexts [39]. The other theme related to culture was that of perceptions and expectations among different internal stakeholders (eg, management, physicians, nurses, and technical staff). Misalignments among these stakeholders were common, particularly regarding trust in ML in general [45,47] or the expected target users (eg, residents vs expert physicians) [45,59].

Innovation

Innovation and its characteristics were among the most frequently mentioned domains (22/34, 65%), with 3 constructs absorbing a significant portion of relevant descriptive themes: innovation design (15/34, 44%), relative advantage (14/34, 41%), and complexity (14/34, 41%).

First, *innovation design* encompassed themes related to the applications' design and functioning, including the types of human-machine interactions, as well as the associated risks. The most recurrent themes within this construct revolved around ease of use and intuitive design [33,45,49,62]. The former was often linked to minimizing manual interventions, such as data input [49,52], and was also associated with dimensions of trust in the applications, such as trust in the process and the cognitive burden for HCPs, in the form of fatigue from overalerting [33,62], which could be a barrier to professional buy-in [55]. Some studies explicitly cited the theme of human-centered design as a development framework that starts with the assessment of end users' needs and the environment in which the ML application will be used [54,64,68]. Another recurrent theme was the human-machine complementarity. For HCPs, it was often important to maintain a sense of control over the application and not perceive it as an attempt at uncontrolled substitution and automation [32,33,56,65]. Human-machine complementarity was also associated with fewer disruptions to established workflows, enhancing the overall benefits associated with the use of ML applications [45,53]. Moreover, complementarity could increase trust in the application from both a micro-perspective (eg, its functioning) [33,54] and macro-perspective (eg, the purpose of the application and the reasons for choosing to integrate ML within a clinical context) [56,58,65]. The risks of ML use in decision-making processes also emerged. These included the risk of automation, in terms of overreliance on ML recommendations [39], and the risk of bias, tied to the underlying data and training model of the ML application [59]. Moreover, the potential negative consequences of automation risk on clinical ability were mentioned [54].

Second, the *relative advantage* revolved around the perception of benefits and costs associated with the use of ML, as well as factors influencing trust in its performance. The most frequently perceived benefits were related to the organizational dimension, in terms of optimization of the workflow resulting from the elimination of unnecessary steps [43], increased attention from end users to all cases managed by the application [45], and enhanced interactions among physicians and other HCPs [62]. Conversely, references to the economic impact were ambiguous. On the one hand, faster decision-making could be considered a potential advantage [47], and on the other hand, human-machine interaction could lead to a loss of efficiency compared to human intervention only [40,56]. Another barrier to professional buy-in is that the perceived poor ability of the application to take contextual factors into account calls into question its clinical relevance. Among the perceived advantages, trust in the application's performance and its determinants were often commented on. For the analysis, we adopted the concept of trust as defined by Hengstler et al [33] who distinguished between trust in technology and trust in those who produce it (ie, the source of innovation). This definition further divided trust in technology into 3 dimensions: trust in performance, focusing on the accuracy and consistency of the output; trust in the process, concerning the understanding of the reasoning behind a given output; and trust in the purpose of the innovation to be implemented [16,44]. Concordance significantly influences trust in performance, with a greater difference between human judgment and machine recommendation associated with a lower level of trust in the recommendation [44,47,58]. Similarly, recommendations that did not arrive in a timely or adequate manner negatively influenced trust in performance [37,44]. Additionally, trust in performance could be fueled by experience, the application's ability to identify cases missed by humans, and the consistency over time of recommendations [47].

Third, *innovation complexity* highlighted the concepts of explainability and opacity as distinctive features of ML models. Many studies were consistent in identifying algorithm complexity as the primary barrier to trust in the process underlying the generation of an ML output. This is even more true when nonmedical professionals (eg, nurses) interact directly with the ML application [45]. Facilitating interpretability, explainability, or cognitive compatibility was mentioned as a

XSL•FO
RenderX

way to promote transparency, trust among HCPs, and professional buy-in [49,55,56,58,59,65].

Implementation Process

The reviewed articles often mentioned the characteristics of the implementation process (22/34, 65%), with a particular emphasis on the constructs of stakeholders engaging (12/34, 35%), reflecting and evaluating (11/34, 32%), and tailoring strategies (9/34, 26%).

Attracting and encouraging the *participation of different stakeholders* in the implementation process emerged as a recurring theme. The practice of early involvement of end users was frequently cited not only during the implementation process but also throughout the development phase [33,38,45,46,58,61,63]. This was positively associated with trust in the innovation's purpose [46], the application's functioning [58,61,62], and the ease of use of its design [45]. During the implementation phase, stakeholder engagement was linked to evident benefits, such as improvement in the implementation climate [46], greater willingness to adopt the role of implementation leader [38], greater professional buy-in [51,55], and better iterative collection of information and feedback [60]. Conversely, the absence of engagement was seen as a barrier to successful implementation, potentially leading to increased resistance toward the innovation among end users [62].

In the construct of *reflecting and evaluating*, feedback and feedback loops emerged as recurring topics, with many studies underscoring the importance for both ML developers and implementation teams to incorporate end users' feedback on either technical issues, system design, or clinical needs [34,36,38,49,62,65]. Some studies noted that feedback collection extended beyond implementation, with structured feedback loop processes integrated into routine use [46,48]. Regardless, feedback collection was described as an iterative activity [46,48,60,61], which also positively influenced professional buy-in [47]. However, a critical point raised was that end-users may lack the necessary technical skills to provide feedback conducive to improvement [62].

Two additional recurring constructs were tailoring strategies and adapting. The former referred to actions addressing barriers and leveraging facilitators, while the latter involved modifying the innovation itself to best fit the context in which it was inserted.

Among *tailoring strategies*, the importance of effectively communicating the implementation efforts was often highlighted. Some works referred to the need for clearly framing communication around the expected benefits, positively affecting trust in ML-based innovations [33,38,45-47,54] and trust in the innovation source [33], and fostering greater professional buy-in [55]. Another aspect of framing was related to the terminology used, asserting that using terms supporting concepts, such as "assistant" and "support" had a favorable impact on end users' trust toward ML-based innovations [46,47] and the innovation source [33].

In terms of *adapting*, the first theme involved the need for collected feedback to be effectively incorporated into the application, adapting systems to the local context of implementation [36,58]. The second involved the issue of data, emphasizing the importance that the model is effectively trained and adapted to the cases treated in the clinical context in which the application will be used before deployment. The absence of this aspect was perceived as a barrier to trust in the ML application's performance [39,41,44,51].

Individuals: Roles

The subdomain of roles was less frequently observed (11/34, 32%) and encompassed 2 constructs: implementation leads (9/34, 26%) and implementation teams (7/34, 21%).

The former referred to the individual or group that guided and oversaw the implementation process, and their presence was generally considered a positive factor for implementation as it contributed to establishing a favorable implementation climate [47]. Individual implementation leaders were often referred to as champions. Although it may theoretically involve figures that emerge from bottom-up processes, all works referring to this role mentioned a top-down identification [45,47,62,63]. Implementation teams were observed as well in the form of quality improvement teams [34], AI governance committees [40,46,63], or interdisciplinary teams of HCPs, software engineers, developers, IT specialists, and other figures [46,51,55,64].

Outer Setting

The outer setting domain emerged poorly in the reviewed studies (9/34, 26%), particularly in the form of 3 constructs: policy and laws (6/34, 18%), local attitudes (4/34, 12%), and partnership and connections (4/34, 12%).

In the *policy and laws* construct, 3 main themes emerged. The first concerned the medicolegal responsibility for decisions made using an ML application [39,61]. The second pertained to regulatory and certification aspects, with recognition of the application as a medical device seen both as a factor positively influencing trust in the application [33] and as a barrier to utilization [46]. Regulations on personal data protection were also considered implementation challenges [47]. Regarding policies, the only theme mentioned was the relevance of national policies and guidelines to create a common framework for the implementation of ML applications [50].

Local attitudes were societal expectations and beliefs on the use of ML applications. Cultural aspects, innovation attitudes, and public expectations could influence the acceptability of ML [35,39,50]. Equally relevant for acceptance was the visibility of the application (ie, how noticeable and observable an innovation is to the public), which influences how organizations foster innovation trust [33,69].

Within the *partnership and connections* construct, building partnerships with scientific societies and professional communities was considered a facilitator for implementation, as these can act as knowledge platforms or hubs [42,47]. Professional communities and peers could also trigger external pressure that may positively impact the willingness to implement ML applications [54]. Establishing development networks across hospitals and health care facilities was a relevant factor for the

XSL•FO
RenderX

increased reliability of the application, providing the opportunity to leverage larger datasets, which are known to end users [61]. Moreover, forging public-private partnerships was deemed a useful step for implementation, to leverage expertise not always available within public health care organizations [50,61].

Discussion

Review of the Main Findings

This work aimed at synthetizing extant academic knowledge on the implementation of ML-based applications in clinical practice, focusing specifically on the characteristics of the innovation and on the processes and strategies employed by health care organizations to ensure their successful implementation.

We identified 34 studies reporting on the implementation process of ML applications, all of which were CDSSs frequently based on supervised learning models in the form of predictive algorithms, visualizations, and alert-delivering tools. Overall, half of the observed applications were integrated into hospital information systems as add-ons to the EHR infrastructure. ML-based applications were mainly implemented in hospital settings and supported prognostic activities, although a relevant portion was intended for diagnosis. Among the diagnostic applications, those based on computer vision were either standalone software or embedded in the hospital hardware technology. Algorithms could be clustered into 2 groups: those internally developed, prevalently by university hospitals and academic medical centers and typically with a prognostic purpose, and those purchased from commercial vendors, which are more heterogeneous in terms of purposes and functions.

Furthermore, our analysis enabled us to scrutinize the characteristics of the implementation processes of ML-based applications, gathering pertinent insights relevant to their successful integration within health care organizations. Through the theoretical lens of the CFIR, we identified a predominant emphasis on 3 key domains: inner setting, innovation characteristics, and process dimension. First, evidence from the inner setting domain highlighted the importance of addressing IT infrastructure and data management challenges, as well as the necessity of fostering an organizational culture that favors the implementation of ML-based applications. Second, in terms of innovation design, the concept of human-machine complementarity was recurrent, highlighting the importance of integrating ML-based applications into existing workflows to enhance overall benefits and foster trust by ensuring HCPs maintain a sense of overall control. In the process domain, studies emphasized the importance of fostering early stakeholder engagement during the development and preimplementation phases, adapting strategies to local contexts, and initiating reflection and evaluation activities to support continuous improvement based on feedback loops. Conversely, while the complexity inherent in ML models in terms of algorithm opacity was largely acknowledged, we found limited investigation into effective mitigation strategies to tackle these challenges.

Comparison With Prior Work

Different from prior work encompassing logic-based and rule-based applications [10,70,71], our study focused exclusively on ML-based applications. While the frequency and relative significance of various application types are not directly comparable with those observed in the cited works, other recent reviews have adopted a similar approach to ours. In their scoping review, Chomutare et al [21] identified 19 studies on the implementation of AI applications powered by ML, highlighting a variety of solutions across medical fields and tasks within the clinical workflow. Similarly, Tricco et al [22] explored how implementation science strategies can facilitate the implementation of ML tools, but their work also included studies with effective research designs, thereby adopting a partially different approach from that of this work. Our review expanded the number of included studies, confirming the multitude of diverse applications of ML in clinical practice. The only condition for which we observed a conspicuous number of studies was sepsis, a dysfunction accounting for around 20% of deaths worldwide [72], for which ML-based applications are proliferating [73], although no definitive causal link with reduced mortality has been demonstrated to date [74]. Our search identified 8 studies on sepsis, showcasing the potential attributed to ML-based applications in supporting the timely identification of hospital-acquired conditions. On a similar note, a recent review encompassing over 10,000 ML applications in health care settings corroborated the relevance of prognostic algorithms among those in use [75].

Consistent with previous research [21,70], most of the included papers presented cases of real-world implementation rather than proper implementation studies on the later phases of rollout, often covering only a few aspects of the implementation process.

While we hypothesized that distinct implementation strategies would be prevalent based on the characteristics of ML-based applications, we only observed limited distinctions based on the types of clinical applications (prognostic, diagnostic, or therapeutic purposes) or their development process (internal development vis-à-vis external acquisition and adaptation).

For instance, the integration with existing IT infrastructure introduced ambiguity in the context of diagnostic applications, where such integration may be perceived as a risk with medicolegal implications [50,59]. On the other hand, for applications with nondiagnostic purposes, integration with existing IT systems was viewed as a positive factor for ease of use [47,62].

Other elements appeared relatively more pronounced in applications provided by external providers. This included perceived risks associated with application design (eg, overreliance, automation, and bias) [39,54,59], considerations regarding complementarity with HCPs [33,53,54,56], and aspects related to explainability. As such, exploring whether and how different application types entail different implications for their effective integration into clinical practice might be a valuable suggestion for future research.

Just like the report by Chomutare et al [21], our work confirmed that the outer setting domain was largely overlooked, although

XSL•FO
RenderX

prior studies have highlighted the importance of external factors, such as data privacy and security laws, ethical issues, regulatory frameworks, and medical liability, in implementing ML applications in clinical practice [5,76-78]. The limited relevance of such a domain in our sample may stem from 2 reasons. First, due to the nature of the included studies, only a few frameworks that were used accounted for elements beyond the organizational setting in which the implementation occurred. Factors associated with the outer setting may be more frequently highlighted in implementation processes perceived as unsuccessful, which are less often reported in the scientific literature. Second, since the primary studies predominantly involved HCPs, they did not incorporate managerial and policymaker perspectives. In fact, when the outer domain perspective was explored, nonclinical stakeholders were often involved [39,47,61]. Furthermore, Hogg et al [10] suggested prioritizing the perspective of non-HCP stakeholders in primary studies to enhance the understanding of implementation processes at a broader level, which may serve as a further valuable suggestion for future studies.

Implications for the Implementation of ML-Based Applications: A Focus on Trust

The importance of trust, particularly within the physician-patient relationship, has been heightened by the advent of digital health, especially with innovations, such as ML applications that heavily rely on data [79]. ML applications based on computational models are often characterized as opaque (ie, black boxes), introducing an extra layer of complexity to the trust relationship between end users and technological innovations [80]. A recent review by Adjekum et al [79] categorized factors influencing trust in digital health systems into personal, technological, and institutional elements. Building upon the concept of trust as articulated by Hengstler et al [33], our work contributes to understanding the determinants of trust in facilitating the implementation of ML-based applications in health care organizations.

We observed that the characteristics of the innovation itself significantly challenged trust in the performance of ML-based applications. The complexity and opacity of the underlying models constitute primary barriers to trust, with trust in performance further influenced by system design elements such as ease of use, the nature of the HCP-machine interaction, and the timeliness and consistency over time of recommendations. Additionally, considerations regarding data governance for internally developed applications and the reputation of the technology provider for procured solutions further influence trust in the performance of these applications. However, as trust primarily remains a human-led process, factors beyond mere technical and mechanical characteristics influence trust in ML.

While most of the observed implementation strategies were essentially ML agnostic, addressing the issue of clinician trust should theoretically require dedicated ML-specific processes. Our review highlights potential ways to enhance the application-perceived reliability of ML applications. On the one hand, tailoring and adaptation strategies, early end-user engagement, and appropriate framing of ML-based applications as decision-support tools might favor HCP trust in both the application's performance and its purpose [21]. On the other hand, specific tailoring strategies should be adopted to increase the explainability of nontotally interpretable models [13]. For instance, Jauk et al [49] enhanced clinical reasoning using a web application presenting relevant features from ML modeling, Davis et al [53] allowed radiologists to interact with the ML system by showing the types and locations of the abnormalities identified by the algorithm, and Henry et al [54] decided to delay alerts until the first verifiable symptoms were present in an attempt to increase acceptance.

However, these tailoring strategies may not be practicable when ML systems reach opacity levels that render the interpretation of their outputs impracticable. In such cases, other contributions have emphasized the need to highlight the level of actionability of ML models, in terms of their ability to enhance medical decision-making compared to clinical judgment alone, to power trust [81].

An additional contribution to enhancing trust may be achieved through continuous HCP involvement. This involvement, which generates engagement and professional buy-in, is equally significant for the successful implementation of these innovations. In the realm of digital health interventions, while there is frequent emphasis on patient engagement in the design of solutions, the empowerment of HCPs is often overlooked [82,83]. Active involvement of HCPs and frequent communications to raise awareness have been unambiguously identified among the most common enablers of trust in previous reviews on the implementation of ML applications [21,22,74,84]. This may facilitate the implementation of innovations by improving the implementation climate for reducing resistance to change and mitigating specific barriers associated with the complexity of ML models and the reliability of the recommendations they produce.

Limitations

This study has some limitations that should be considered when interpreting our findings. First, the rapidly evolving nature of the field of ML and the exponential growth of newly published studies posed challenges in managing the vast volume of retrieved records. To address this, our search strategy incorporated a supplementary block of keywords focused on "study designs," which may have excluded certain relevant articles. Additionally, our emphasis on peer-reviewed studies introduced a potential bias, as ML-based applications reported in the scientific literature may only represent a subset of implemented systems. This could impact the generalizability of our findings, as acknowledged in similar studies such as the study by Sharma et al [70]. Lastly, the decision to include only papers published in English might have led to the exclusion of valuable sources published in other languages, limiting the comprehensiveness of our review.

Conclusions

Despite a relative dearth of primary studies on the implementation of ML applications in health care organizations, the available evidence reveals the abundance and heterogeneity of factors involved when ML applications are introduced in routine clinical practice. While certain elements, such as complexity and trust, tend to emerge as distinctive factors for

XSL•FO
RenderX

ML applications, many other aspects reflect what is already known about the implementation of digital technologies, particularly traditional CDSSs.

Further research is needed to bridge the gap between the theoretical potential of ML and its actual use in health care organizations. Identifying the distinctive factors that can facilitate its implementation will build theoretical and practical knowledge for health care practitioners, ultimately promoting the uptake of ML in routine clinical practice.

Acknowledgments

This systematic review was performed within the MUSA (Multilayered Urban Sustainability Action) project funded by European Union – NextGenerationEU under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems," set up of "territorial leaders in R&D."

Data Availability

Data collected as part of this review will be made available by the corresponding author upon reasonable request.

Authors' Contributions

GC and AC conceived the initial concept for the review. VA and LMP carried out the search strategy and initial screening of the records. VA, LMP, and FP performed the full-text analysis and data extraction. VA, LMP, and FP generated the first draft of the manuscript. All authors contributed to the analytical process and approved the final version of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 87 KB - jmir_v26i1e55897_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 21 KB - jmir_v26i1e55897_app2.docx]

Multimedia Appendix 3
Quality assessment.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v26i1e55897_app3.xlsx]

Multimedia Appendix 4
Characteristics of the studies and applications.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v26i1e55897_app4.xlsx]

Multimedia Appendix 5
Coding process.
[XLSX File (Microsoft Excel File), 45 KB - jmir_v26i1e55897_app5.xlsx]

References

1. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Apr 10;21(1):125 [FREE Full text] [doi: 10.1186/s12911-021-01488-9] [Medline: 33836752]
2. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
3. Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training Validation and Evaluation. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240038462 [accessed 2024-11-04]
4. Mittermaier M, Raza M, Kvedar JC. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. NPJ Digit Med 2023 Aug 05;6(1):137 [FREE Full text] [doi: 10.1038/s41746-023-00889-6] [Medline: 37543707]
5. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019 Jul 10;21(7):e13659 [FREE Full text] [doi: 10.2196/13659] [Medline: 31293245]

XSL•FO
RenderX

6. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/s2589-7500(19)30123-2]
7. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. The Lancet 2022 Feb;399(10325):620. [doi: 10.1016/s0140-6736(22)00235-5]
8. Coiera E. The last mile: Where artificial intelligence meets reality. J Med Internet Res 2019 Nov 08;21(11):e16323 [FREE Full text] [doi: 10.2196/16323] [Medline: 31702559]
9. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
10. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston A, Kelly C, et al. Stakeholder perspectives of clinical artificial intelligence implementation: Systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
11. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019 Mar 12;364:l886 [FREE Full text] [doi: 10.1136/bmj.l886] [Medline: 30862612]
12. Castelvecchi D. Can we open the black box of AI? Nature 2016 Oct 06;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
13. Barragán-Montero A, Bibal A, Dastarac MH, Draguet C, Valdés G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Phys Med Biol 2022 May 27;67(11):11TR01 [FREE Full text] [doi: 10.1088/1361-6560/ac678a] [Medline: 35421855]
14. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - Addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
15. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: Scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
16. Nilsen P, Reed J, Nair M, Savage C, Macrae C, Barlow J, et al. Realizing the potential of artificial intelligence in healthcare: Learning from intervention, innovation, implementation and improvement sciences. Front Health Serv 2022 Sep 15;2:961475 [FREE Full text] [doi: 10.3389/frhs.2022.961475] [Medline: 36925879]
17. Nilsen P, Svedberg P, Neher M, Nair M, Larsson I, Petersson L, et al. A framework to guide implementation of AI in health care: Protocol for a cocreation research project. JMIR Res Protoc 2023 Nov 08;12:e50216 [FREE Full text] [doi: 10.2196/50216] [Medline: 37938896]
18. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: Bibliometric analysis. J Med Internet Res 2020 Jul 29;22(7):e18228 [FREE Full text] [doi: 10.2196/18228] [Medline: 32723713]
19. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020 Aug 21;3(1):107 [FREE Full text] [doi: 10.1038/s41746-020-00318-y] [Medline: 32885053]
20. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: Review of the literature. J Med Internet Res 2019 Apr 05;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]
21. Chomutare T, Tejedor M, Svenning TO, Marco-Ruiz L, Tayefi M, Lind K, et al. Artificial intelligence implementation in healthcare: A theory-based scoping review of barriers and facilitators. Int J Environ Res Public Health 2022 Dec 06;19(23):16359 [FREE Full text] [doi: 10.3390/ijerph192316359] [Medline: 36498432]
22. Tricco AC, Hezam A, Parker A, Nincic V, Harris C, Fennelly O, et al. Implemented machine learning tools to inform decision-making for patient care in hospital settings: a scoping review. BMJ Open 2023 Feb 07;13(2):e065845 [FREE Full text] [doi: 10.1136/bmjopen-2022-065845] [Medline: 36750280]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Ardito V, Cappellaro G, Compagni A, Petracca F, Preti LM. Implementation of machine learning applications in health care organizations: Protocol for a systematic review of empirical studies. JMIR Res Protoc 2023 Sep 12;12:e47971 [FREE Full text] [doi: 10.2196/47971] [Medline: 37698910]
25. Secules S, McCall C, Mejia JA, Beebe C, Masters AS, L. Sánchez - Peña M, et al. Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community. J of Engineering Edu 2021 Feb 05;110(1):19-43. [doi: 10.1002/jee.20377]
26. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629 [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
27. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]

XSL•FO
RenderX

28. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
29. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
32. Lee EK, Atallah HY, Wright MD, Post ET, Thomas C, Wu DT, et al. Transforming hospital emergency department workflow and patient care. Interfaces 2015 Feb;45(1):58-82. [doi: 10.1287/inte.2014.0788]
33. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. Technological Forecasting and Social Change 2016 Apr;105:105-120. [doi: 10.1016/j.techfore.2015.12.014]
34. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. BMJ Open Qual 2017 Oct 25;6(2):e000158 [FREE Full text] [doi: 10.1136/bmjoq-2017-000158] [Medline: 29450295]
35. Bhattacharya S, Sharma N, Singh A. Designing culturally acceptable screening for breast cancer through artificial intelligence-two case studies. J Family Med Prim Care 2019 Feb;8(2):760-762 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_391_18] [Medline: 30984711]
36. Cruz N, Canales L, Muñoz J, Pérez B, Arnott I. Improving adherence to clinical pathways through natural language processing on electronic medical records. Stud Health Technol Inform 2019 Aug 21;264:561-565. [doi: 10.3233/SHTI190285] [Medline: 31437986]
37. Ginestra JC, Giannini HM, Schweickert WD, Meadows L, Lynch MJ, Pavan K, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med 2019 Nov;47(11):1477-1484 [FREE Full text] [doi: 10.1097/CCM.0000000000003803] [Medline: 31135500]
38. Gonçalves L, Amaro MLDM, Romero ADLM, Schamne FK, Fressatto JL, Bezerra CW. Implementation of an Artificial Intelligence Algorithm for sepsis detection. Rev Bras Enferm 2020;73(3):e20180421 [FREE Full text] [doi: 10.1590/0034-7167-2018-0421] [Medline: 32294705]
39. Sun TQ, Medaglia R. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly 2019 Apr;36(2):368-383. [doi: 10.1016/j.giq.2018.09.008]
40. Baxter SL, Bass JS, Sitapati AM. Barriers to implementing an artificial intelligence model for unplanned readmissions. ACI open 2020 Jul 19;4(2):e108-e113 [FREE Full text] [doi: 10.1055/s-0040-1716748] [Medline: 33274314]
41. Cho K, Kwon O, Kwon J, Lee Y, Park H, Jeon K, et al. Detecting patient deterioration using artificial intelligence in a rapid response system. Crit Care Med 2020 Apr;48(4):e285-e289. [doi: 10.1097/CCM.0000000000004236] [Medline: 32205618]
42. Frontoni E, Romeo L, Bernardini M, Moccia S, Migliorelli L, Paolanti M, et al. A decision support system for diabetes chronic care models based on general practitioner engagement and EHR data sharing. IEEE J Transl Eng Health Med 2020;8:1-12. [doi: 10.1109/jtehm.2020.3031107]
43. Hassan AE, Ringheanu VM, Rabah RR, Preston L, Tekle WG, Qureshi AI. Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. Interv Neuroradiol 2020 Oct 26;26(5):615-622 [FREE Full text] [doi: 10.1177/1591019920953055] [Medline: 32847449]
44. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: A pre-post study of providers' experience with artificial intelligence-based clinical decision support. Int J Med Inform 2020 May;137:104072. [doi: 10.1016/j.ijmedinf.2019.104072] [Medline: 32200295]
45. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a machine learning system into clinical workflows: Qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421 [FREE Full text] [doi: 10.2196/22421] [Medline: 33211015]
46. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. JMIR Med Inform 2020 Jul 15;8(7):e15182 [FREE Full text] [doi: 10.2196/15182] [Medline: 32673244]
47. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct 26;30(10):5525-5532 [FREE Full text] [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
48. Xu H, Li P, Yang Z, Liu X, Wang Z, Yan W, et al. Construction and application of a medical-grade wireless monitoring system for physiological signals at general wards. J Med Syst 2020 Sep 04;44(10):182 [FREE Full text] [doi: 10.1007/s10916-020-01653-z] [Medline: 32885290]
49. Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. J Med Syst 2021 Mar 01;45(4):48 [FREE Full text] [doi: 10.1007/s10916-021-01727-6] [Medline: 33646459]

XSL•FO
RenderX

50. Morales H, Guedes M, Silva J, Massuda A. COVID-19 in Brazil-preliminary analysis of response supported by artificial intelligence in municipalities. Front Digit Health 2021;3:648585 [FREE Full text] [doi: 10.3389/fdgth.2021.648585] [Medline: 34713121]
51. Murphree D, Wilson P, Asai S, Quest D, Lin Y, Mukherjee P, et al. Improving the delivery of palliative care through predictive modeling and healthcare informatics. J Am Med Inform Assoc 2021 Jun 12;28(6):1065-1073 [FREE Full text] [doi: 10.1093/jamia/ocaa211] [Medline: 33611523]
52. Yao J, Wang T, Hou G, Ou D, Li W, Zhu Q, et al. AI detection of mild COVID-19 pneumonia from chest CT scans. Eur Radiol 2021 Sep 18;31(9):7192-7201 [FREE Full text] [doi: 10.1007/s00330-021-07797-x] [Medline: 33738595]
53. Davis MA, Rao B, Cedeno PA, Saha A, Zohrabian VM. Machine learning and improved quality metrics in acute intracranial hemorrhage by noncontrast computed tomography. Curr Probl Diagn Radiol 2022 Jul;51(4):556-561. [doi: 10.1067/j.cpradiol.2020.10.007] [Medline: 33243455]
54. Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit Med 2022 Jul 21;5(1):97 [FREE Full text] [doi: 10.1038/s41746-022-00597-7] [Medline: 35864312]
55. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Jul;5(2):ooac022 [FREE Full text] [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
56. Lebovitz S, Lifshitz-Assaf H, Levina N. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. Organization Science 2022 Jan;33(1):126-148. [doi: 10.1287/orsc.2021.1549]
57. Rushlow DR, Croghan IT, Inselman JW, Thacher TD, Friedman PA, Yao X, et al. Clinician adoption of an artificial intelligence algorithm to detect left ventricular systolic dysfunction in primary care. Mayo Clin Proc 2022 Nov;97(11):2076-2085 [FREE Full text] [doi: 10.1016/j.mayocp.2022.04.008] [Medline: 36333015]
58. Schwartz JM, George M, Rossetti SC, Dykes PC, Minshall SR, Lucas E, et al. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: Qualitative descriptive study. JMIR Hum Factors 2022 May 12;9(2):e33960 [FREE Full text] [doi: 10.2196/33960] [Medline: 35550304]
59. Sibbald M, Abdulla B, Keuhl A, Norman G, Monteiro S, Sherbino J. Electronic diagnostic support in emergency physician triage: Qualitative study with thematic analysis of interviews. JMIR Hum Factors 2022 Sep 30;9(3):e39234 [FREE Full text] [doi: 10.2196/39234] [Medline: 36178728]
60. Singer SJ, Kellogg KC, Galper AB, Viola D. Enhancing the value to users of machine learning-based clinical decision support tools: A framework for iterative, collaborative development and implementation. Health Care Manage Rev 2021 Sep 9;47(2):E21-E31. [doi: 10.1097/hmr.0000000000000324]
61. Wijnhoven F. Organizational learning for intelligence amplification adoption: Lessons from a clinical decision support system adoption project. Inf Syst Front 2021 Oct 09;24(3):731-744. [doi: 10.1007/s10796-021-10206-9]
62. Zhai Y, Yu Z, Zhang Q, Zhang Y. Barriers and facilitators to implementing a nursing clinical decision support system in a tertiary hospital setting: A qualitative study using the FITT framework. Int J Med Inform 2022 Oct;166:104841. [doi: 10.1016/j.ijmedinf.2022.104841] [Medline: 36027798]
63. Pou-Prom C, Murray J, Kuzulugil S, Mamdani M, Verma AA. From compute to care: Lessons learned from deploying an early warning system into clinical practice. Front Digit Health 2022 Sep 5;4:932123 [FREE Full text] [doi: 10.3389/fdgth.2022.932123] [Medline: 36133802]
64. Hinson JS, Klein E, Smith A, Toerper M, Dungarani T, Hager D, et al. Multisite implementation of a workflow-integrated machine learning system to optimize COVID-19 hospital admission decisions. NPJ Digit Med 2022 Jul 16;5(1):94 [FREE Full text] [doi: 10.1038/s41746-022-00646-1] [Medline: 35842519]
65. Berge GT, Granmo OC, Tveit TO, Munkvold BE, Ruthjersen AL, Sharma J. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. BMC Med Inform Decis Mak 2023 Jan 10;23(1):5 [FREE Full text] [doi: 10.1186/s12911-023-02101-x] [Medline: 36627624]
66. Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts. European Parliament. URL: https://data.europa.eu/doi/10.2861/568473 [accessed 2024-11-04]
67. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
68. Polhemus AM, Novák J, Ferrao J, Simblett S, Radaelli M, Locatelli P, et al. Human-centered design strategies for device selection in mHealth programs: Development of a novel framework and case study. JMIR Mhealth Uhealth 2020 May 07;8(5):e16043 [FREE Full text] [doi: 10.2196/16043] [Medline: 32379055]
69. Rogers EM. Diffusion of Innovations. New York, NY: Simon and Schuster; 2003.
70. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: Scoping review. J Med Internet Res 2022 Oct 05;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
71. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: Systematic review. J Med Internet Res 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]

XSL•FO
RenderX

72. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet 2020 Jan;395(10219):200-211. [doi: 10.1016/s0140-6736(19)32989-7]
73. Islam MM, Nasrin T, Walther BA, Wu C, Yang H, Li Y. Prediction of sepsis patients using machine learning approach: A meta-analysis. Comput Methods Programs Biomed 2019 Mar;170:1-9. [doi: 10.1016/j.cmpb.2018.12.027] [Medline: 30712598]
74. van der Vegt A, Scott I, Dermawan K, Schnetler R, Kalke V, Lane P. Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework. J Am Med Inform Assoc 2023 Jun 20;30(7):1349-1361 [FREE Full text] [doi: 10.1093/jamia/ocad075] [Medline: 37172264]
75. Kolasa K, Admassu B, Hołownia-Voloskova M, Kędzior KJ, Poirrier J, Perni S. Systematic reviews of machine learning in healthcare: a literature review. Expert Rev Pharmacoecon Outcomes Res 2024 Jan 24;24(1):63-115. [doi: 10.1080/14737167.2023.2279107] [Medline: 37955147]
76. Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022 Jul 01;22(1):850 [FREE Full text] [doi: 10.1186/s12913-022-08215-8] [Medline: 35778736]
77. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
78. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
79. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: Scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
80. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. ANNALS 2020 Jul;14(2):627-660. [doi: 10.5465/annals.2018.0057]
81. Ehrmann DE, Joshi S, Goodfellow SD, Mazwi ML, Eytan D. Making machine learning matter to clinicians: model actionability in medical decision-making. NPJ Digit Med 2023 Jan 24;6(1):7 [FREE Full text] [doi: 10.1038/s41746-023-00753-7] [Medline: 36690689]
82. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
83. Mesko B, Győrffy Z. The rise of the empowered physician in the digital health era: Viewpoint. J Med Internet Res 2019 Mar 26;21(3):e12490 [FREE Full text] [doi: 10.2196/12490] [Medline: 30912758]
84. Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: A systematic review. Informatics (MDPI) 2020 Sep 25;7(3):25 [FREE Full text] [doi: 10.3390/informatics7030025] [Medline: 33274178]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
CFIR: Consolidated Framework for Implementation Research
CT: computed tomography
ED: emergency department
EHR: electronic health record
HCP: health care professional
ICU: intensive care unit
IT: information technology
ML: machine learning
MMAT: Mixed Method Appraisal Tool

Edited by N Cahill; submitted 30.12.23; peer-reviewed by KY Jung, PH Liao, B Mesko; comments to author 22.04.24; revised version received 07.07.24; accepted 03.10.24; published 25.11.24.

Please cite as:
Preti LM, Ardito V, Compagni A, Petracca F, Cappellaro G
Implementation of Machine Learning Applications in Health Care Organizations: Systematic Review of Empirical Studies
J Med Internet Res 2024;26:e55897
URL: https://www.jmir.org/2024/1/e55897
doi:10.2196/55897
PMID:

XSL•FO
RenderX

©Luigi M Preti, Vittoria Ardito, Amelia Compagni, Francesco Petracca, Giulia Cappellaro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Taxonomy and Archetypes of AI-Based Health Care Services: Qualitative Study

Marlene Blaß[1*], MSc; Henner Gimpel[1*], Prof Dr, PhD; Philip Karnebogen[2*], PhD

[1]FIM Research Center for Information Management, University of Hohenheim, Branch Business & Information Systems Engineering of the Fraunhofer FIT, Stuttgart, Germany

[2]FIM Research Center for Information Management, Branch Business & Information Systems Engineering of the Fraunhofer FIT, University of Applied Sciences Augsburg, Augsburg, Germany

[*]all authors contributed equally

Corresponding Author:
Marlene Blaß, MSc
FIM Research Center for Information Management
University of Hohenheim
Branch Business & Information Systems Engineering of the Fraunhofer FIT
Schloss Hohenheim 1
Stuttgart, 70599
Germany
Phone: 49 0711 459 24051
Email: marlene.blass@fit.fraunhofer.de

Abstract

Background: To cope with the enormous burdens placed on health care systems around the world, from the strains and stresses caused by longer life expectancy to the large-scale emergency relief actions required by pandemics like COVID-19, many health care companies have been using artificial intelligence (AI) to adapt their services. Nevertheless, conceptual insights into how AI has been transforming the health care sector are still few and far between. This study aims to provide an overarching structure with which to classify the various real-world phenomena. A clear and comprehensive taxonomy will provide consensus on AI-based health care service offerings and sharpen the view of their adoption in the health care sector.

Objective: The goal of this study is to identify the design characteristics of AI-based health care services.

Methods: We propose a multilayered taxonomy created in accordance with an established method of taxonomy development. In doing so, we applied 268 AI-based health care services, conducted a structured literature review, and then evaluated the resulting taxonomy. Finally, we performed a cluster analysis to identify the archetypes of AI-based health care services.

Results: We identified 4 critical perspectives: agents, data, AI, and health impact. Furthermore, a cluster analysis yielded 13 archetypes that demonstrate our taxonomy's applicability.

Conclusions: This contribution to conceptual knowledge of AI-based health care services enables researchers as well as practitioners to analyze such services and improve their theory-led design.

(J Med Internet Res 2024;26:e53986) doi:10.2196/53986

KEYWORDS

healthcare; artificial intelligence; AI; taxonomy; services; cluster analysis; archetypes

Introduction

We are entering a new health era—an era of digital health defined by innovative medical solutions and an explosion of health-related data [1]. Fueled by these amounts of data, forecasts indicate that artificial intelligence (AI) will soon exceed today's capabilities in supporting and transforming health care systems [2-4]. There is reason to believe that AI will improve health care at various stages, ranging from disease detection to treatment [5], so much so that it promises to facilitate a new generation of health care services [6]. The proliferation of AI promises to be a panacea for our era's multiple health care issues, including the ongoing fight against the COVID-19 pandemic (eg, studies by Chung et al [7] and Krämer et al [8]), the long-term health complaints of an aging population (eg, study by Velazquez et al [9]), and the ever-increasing rates of chronic diseases (eg, study by Sekandi

XSL•FO
RenderX

et al [10]). It is unlikely that AI will live up to its hype of being a miracle cure for all health care–related problems, as its implementation is not without hurdles. Especially, privacy concerns due to the sensitive nature of health-related data and ethical considerations like the risk of biased algorithms leading to unequal health care outcomes pose significant challenges and require careful navigation.

Nevertheless, it has the significant potential of not only disrupting old and dysfunctional services but also providing new means of managing and perhaps resolving various health care challenges [2,11,12]. Already, the increasing reliance on AI is changing the nature of health care services. Indeed, AI is joining the established actors in health care services [13] and transforming the ways in which physical, mental, or social well-being is maintained or restored [14]. When we speak of AI in this context, it serves as an umbrella term for a broad set of algorithms [15-17] that use data to perform cognitive functions previously limited to humans [18,19].

Health care companies have included AI in their service provisions, which has led to ever-increasing investments in the development of AI-based health care services [20]. Likewise, the number of published research papers debating the role of AI in health care has gone up significantly, and it has done so in a broad range of disciplines. Despite this multilateral fascination with AI-based health care services, however, there is a notable and rather lamentable lack of conceptual insights into the design characteristics of AI-based health care services, which is why there is still no clear understanding of its general properties and potential.

To date, research has mainly focused on investigating the potential of individual AI applications in the health care sector [12,21-23]. Yet to be derived from these, however, is an overarching structure with which to classify the various real-world phenomena. In the absence of a clear and comprehensive taxonomy, there has been no consensus on AI-based health care service offerings, what they have in common, and how they differ. This lack of a shared understanding of AI services limits any further analysis of how AI may affect the health care sector, which in turn limits the ability of researchers to stay abreast of the life-changing and, indeed, life-saving developments in this domain [5,17]. Thus, this study sets out to answer the following research question: What are the design characteristics of AI-based health care services?

To answer this, we propose a multilayered taxonomy created in accordance with both the well-established method of taxonomy development by Nickerson et al [24] and the recent extension by Kundisch et al [25]. By exploring both empirical-to-conceptual (E2C) and conceptual-to-empirical (C2E) iterations, we pursued a bilateral development for the structure of our taxonomy, deductively from real-world examples and inductively from a structured literature review. Through 4 iterations, we identified 10 key dimensions and their respective characteristics, all of which we structured into 4 perspectives: agents, data, AI, and health impact. We also performed a thorough review of the current literature, particularly the 28 most relevant studies, and considered their insights with regard to 268 AI-based health care services. We then conducted a careful evaluation of our taxonomy, as proposed by Kundisch et al [25].

The taxonomy and the archetypes of AI-based health care services are of theoretical as well as practical value. With regard to the former, this study can provide a helpful frame of reference for the emerging field of digital transformation in health care by outlining and structuring the general features of AI-based health care services. With regard to the latter, our taxonomy makes it possible to map the field of AI-based health care services, assess their possibilities in a competitive field, and offer guidance to health care entrepreneurs.

Methods

Overview

To identify the defining features of AI-based health care services as well as their corresponding archetypes, we formulated a bilateral methodological approach to develop our taxonomy in accordance with both the guidelines of Nickerson et al [24] and their extension by Kundisch et al [25]. In doing so, we achieved our dual aim of creating a versatile set of tools for the analysis of both conceptual and empirical observations. Taxonomies comprise various layers, which in turn include multiple dimensions, which can again contain diverse characteristics [24]. They are frequently used (eg, Fischer et al [26] and Gimpel et al [27] to show how different concepts are connected or to investigate their relationships [28].

Defining Preliminary Conditions

As proposed by Kundisch et al [25], we began our taxonomy development by specifying the central phenomenon under investigation, that is, AI-based health care services. To pay due attention to all of their various design characteristics as well as those of our 2 target groups—researchers examining the field of AI-based health care services and practitioners in the health care sector who are already integrating or aim to integrate AI in their service offerings—we define the meta-characteristic [24], that is, the design characteristics of AI-based health care services.

In line with the best practices of Nickerson et al [24], we then define our objective and subjective ending conditions, both of which are checked after each iteration to assess whether the taxonomy development process is completed [24]. Nickerson et al [24] identified a set of ending conditions that can be applied or adapted to the taxonomy development process, four of which we selected for this study: (1) all objects of our dataset have been examined, (2) no new dimension or characteristics were added in the last iteration, (3) every dimension is unique and not repeated, and (4) at least one object from our dataset is classified as having every characteristic of every dimension [24]. Furthermore, we concur with Nickerson et al [24] regarding the subjective ending conditions in that the taxonomy should be concise, robust, comprehensive, extendible, and explanatory [24].

XSL•FO
RenderX

Taxonomy Development Process

Our taxonomy development process consists of 4 iterations. Nickerson et al [24] propose 2 development approaches to this development. First, inductive, which is to say E2C, and second, deductive, which is to say C2E [24]. These 2 approaches are to be executed consecutively in several iterations. Table 1 depicts details on this taxonomy development process, including the development approach chosen for each iteration, the objects in question, the significant changes, and the evaluation of the ending conditions.

Table 1. Details of the iterative taxonomy development process.

Iteration	App[a]	Objects	Major changes	Ending conditions
1	E2C[b]	25 AI[c]-based health care services randomly chosen from our sample	Identification of 3 perspectives with 10 dimensions and 35 characteristics in total	Objective and subjective ending conditions not met: • not all objects were examined • not robust because it is not grounded in theory, and not concise because of the many dimensions
2	C2E[d]	19 papers from a structured literature review	Addition of 1 perspective with 2 new dimensions and 5 new characteristics; reordering of the dimensions to match the perspectives	Objective and subjective ending conditions not met: • not all objects were examined • not all dimensions are unique • not concise due to the many dimensions
3	E2C	25 AI-based health care services randomly chosen from our sample	Abandonment of 1 dimension; replacement of 2 dimensions due to significant overlap; modification of multiple characteristics	Objective ending conditions not met: • not all objects were examined
4	E2C	The remaining sample of 218 AI-based health care services	No further modification	All ending conditions were met: • objective ending conditions (1)-(4) are met • the authors have achieved consensus and agree that the subjective ending conditions are met

[a]App: development approach.

[b]E2C: empirical-to-conceptual.

[c]AI: artificial intelligence.

[d]C2E: conceptual-to-empirical.

In our first iteration, we chose an E2C approach since this is generally agreed to be the best practice when a significant amount of data can be used to examine a specific research field. To collect the necessary empirical data when making our selection of AI-based health care services, we followed a similar approach to that of Labes et al [29] and Fischer et al [26]. In doing so, we were aware of the challenge of keeping up with the fast-changing field of AI in health care, which is why we identified so-called "top lists" that consolidate the most promising AI companies in health care. After compiling an initial list with the help of several internet search engines, using terms such as "top AI health care companies," "ranking AI companies in health care," and "top list health care and AI," we synthesized our results into a single integrated list of AI-based health care services derived from 6 data sources. Our sample only included services that describe an activity involving at least 2 entities with different roles, for example, a patient and a health care professional, where both use their resources, be it information or technology [30], in a collaborative process that serves their mutual benefit. Furthermore, it was not our aim to compile a complete list of all existing health care start-ups that use AI in their portfolio. Instead, we wanted to collect a sample that was as representative and generalizable as possible to serve as a basis for the taxonomy development. For this reason, we used 6 different top lists as a data source, which are based on various evaluation criteria such as growth, funding, and global impact. Furthermore, we use top lists from the years 2020 and 2023 to also cover a temporal variance. This ensures that our sample reflects sufficient variance and can be considered representative. The final list comprised 268 AI-based health care services. Distinct subsamples were used in iterations 1, 3, and 4. Tables 2 and 3 illustrate the data sources as well as exemplary services.

XSL•FO
RenderX

Table 2. Empirical-to-conceptual data used for our taxonomy development.

Considered top lists	Exemplary firms (services)
Forbes [31]	Bioformis (Biovitals); Karius (Karius-Bloodtest); Viz.ai (Stroke Monitoring)
Medical startups [32]	Babylon Health (Ask Babylon, Healthcheck); Suki (Digital Assistant; Speech Service)
Startus [33]	Roam Analytics (Roam Health Knowledge Graph); Sopris Health (Sopris Assistant); Artelus (DRISTi, Hansanet)
CBInsights [34]	Mindstrong (Mindstrong Service); Freenome (Cancer detection); PathAI (Imaging Service)
QWayHealth [35]	Arterys (Cardio AI, Lung AI) Caption Health (Caption AI); Corti (Audia)
Medical futurist [36]	Google Deepmind (Eye Disease Scan); Oncora Medical (Patient Care, Analytics); CloudMedX (AskSophie; Decision Point)
Total services, n	268

Table 3. Conceptual-to-empirical data used for our taxonomy development.

Activity	Details
Search string	(“artificial intelligence” OR “AI” OR “machine learning” OR “deep learning”) AND (“healthcare” OR “health care” OR “clinical” OR “medicine”) AND (“service” OR “application”)
Databases sources	Web of Science; forward and backward search
Results	20 papers found in the initial search; 8 papers found in the forward and backward searches [37]
Papers found	Agrawal et al [15]; Ågerfalk [17]; Reis et al [23]; Fischer et al [26]; Gimpel et al [27]; Bao et al [37]; Baars and Kemper [38]; Bardhan et al [39]; Brynjolfsson and Mitchell [40]; Chang [41]; Hansen and Baroody [42]; Hofmann et al [43]; Hunke et al [44]; Iansiti and Lakhani [45]; Mantzana et al [46]; Reddy et al [47]; Schuetz and Venkatesh [48]; Tan et al [49]; Thompson et al [50]; Wani et al [51]; Weglarz [52]; Antoniadi et al [53]; Shamshirband et al [54]; Ali et al [55]; Qadri et al [56]; Karatas et al [57]; Shatte et al [58]; Bindra and Jain [59]
Total papers, n	28

In the first iteration, we analyzed a random sample of 25 AI-based health care services to derive relevant dimensions and characteristics. In the second iteration, we followed a C2E approach informed by the insights of a structured literature review [60]; Tables 2 and 3 contain the details. With this 2D approach to data collection from practical and academic data sources, we were able to account for the multiple developments in the rapidly expanding field of AI-based health care services. To broaden our understanding accordingly and heed the lessons learned from the research already conducted, we performed a literature review on the Web of Science. We used an iterative search process for the literature review. Initially, we had a very narrow search string to gain an initial understanding of the topic, which we later expanded again to ensure a more comprehensive literature review. For the final search, we limited our search to papers published between 2015 and 2023 in the English language and excluded book chapters or proceeding papers. By using the exact search string “(“artificial intelligence” OR “AI” OR “machine learning” OR “deep learning”) AND (“healthcare” OR “health care” OR “clinical” OR “medicine”) AND (“service” OR “application”),” we were able to exclude any literature irrelevant to our study. This left us with 9318 relevant results. Furthermore, 2 members of the research team (MB and PK) then analyzed the title and abstract of each paper. After that, all those who did not directly address the topic under investigation were excluded. Since reliability was a guiding principle of this process, the team screened and classified a random sample of 30 papers in close collaboration. After that, 30 papers were screened independently, and their classifications were compared to certify the analytical adequacy. All subsequent papers were screened individually by one of the team members. Speaking in numbers, this meant that from the initial set of 9318 papers, 276 had their full text analyzed, and 20 of these were considered for our taxonomy. In addition, 8 papers were included through forward and backward searches, which brought the total to 28 papers.

For the third and fourth iterations, we again followed the E2C approach with 25 randomly selected AI-based health care services and 218 remaining from our initial sample. After the fourth iteration, the team agreed that the objective and subjective ending conditions were all met.

Demonstration and Evaluation Process

As Kundisch et al [25] pointed out, a rigorous demonstration and evaluation of a newly developed taxonomy require a 2-step approach. In the first step, 13 researchers with a domain-specific background in AI or health care, or indeed both, classified multiple real-world services by using our taxonomy. These demographics were selected because they constitute the intended future users of our taxonomy. We measured the agreement among the individual classifications by means of the percentage agreement and Fleiss κ [61]. The latter was chosen by virtue of the fact that it can be used to determine the degree of agreement between the assessments of more than 2 raters, which is to say that it works as an extension to Cohen κ [62], provided that the number of raters per assessed unit remains constant (described in detail in “Taxonomy Evaluation” in Multimedia Appendix 1). The unweighted κ [61] is determined, that is, each nonmatch is weighted equally. The calculation of κ [61] ( ) results from

XSL•FO
RenderX

the calculation of the relative agreement of the raters ( ) and the calculation of the probability of random agreement ( ).

n describes the number of raters (n=13), N is the number of considered service offerings per rater, and k is the number of possible evaluation categories. Aside from being a well-established procedure to evaluate all manner of taxonomies, it allowed us to ensure that classifications made with our taxonomy are not too susceptible to personal bias [27,63].

In the second step, we classified all available services to learn more about the occurrence rate of each characteristic in our dataset. This was followed by cluster analysis, a statistical method of similarity analysis performed to identify groups of similar objects, so-called clusters, within a dataset [64-66]. The matching objects are grouped in such a way as to achieve high homogeneity within clusters and high heterogeneity between clusters [67]. In our case, this cluster analysis allowed us to identify archetypes of AI-based health care services by observing the three necessary specifications: (1) a proximity measure to calculate similarities or distances between objects, (2) a clustering method with underlying grouping methods or fusion algorithms, and (3) the order in which the clusters are split or merged. It is worth noting, however, that the choice of the clustering procedure is closely related to the question of the optimal number of clusters. To determine an appropriate number of clusters for AI-based health care services, we calculated several measures discussed in the literature (described in detail in “Clustering” in Multimedia Appendix 1). For the various numbers of clusters, each analysis is performed with an agglomerative hierarchical method that creates a corresponding cluster solution for any number of clusters between 1 and n, where n provides the number of services considered in the sample [68,69]. To this end, we chose the Ward algorithm [70] as our clustering algorithm and used the Manhattan distance as our distance measure, seeing as the latter can be applied to nominally scaled data and has proven to perform well in connection with the Ward algorithm [70,71]. To ensure the correct application of the distance measure, we first dichotomized and then standardized the underlying data by accounting for the number of characteristics per dimension, thus also ensuring that all dimensions are equally weighted.

Ethical Considerations

This study did not require approval from an ethics review board, as it does not constitute human participants research. In accordance with the guidelines of the University of Hohenheim, ethical review is mandated only for research involving human participants, biological samples, or identifiable personal data. The focus of this study was on identifying the design characteristics of AI-based health care services, relying primarily on a review of existing literature and analysis of established AI-based health care services. Data from individuals were collected solely for the purpose of evaluating the proposed taxonomy, and as such, the study does not meet the criteria for human participants research as defined by the relevant policies of the University of Hohenheim (Satzung der Ethikkommission der Universität Hohenheim).

Results

A Taxonomy of AI-Based Health Care Services

Overview

Following the guidelines of Nickerson et al [24], we developed the taxonomy of AI-based health care services shown in Table 4. The taxonomy is predicated on the 4 perspectives of agents, data, AI, and health impact, and it comprises 10 essential dimensions structured in accordance with these 4 perspectives. However, whereas Nickerson et al [24] suggest that one should design all characteristics as mutually exclusive in order for a single characteristic to be observed per dimension at any one time [24], we instead decided to adhere to the design principles of common taxonomies [26,27,44]. Nonexclusive dimensions could be replaced with binary dimensions for each characteristic. However, this would drastically increase the size of the taxonomy and, thus, impair conciseness, which is why we decided to include nonexclusive dimensions. As a result, we are able to reflect on the complexity of AI-based services more appropriately while keeping the taxonomy parsimonious. Whether or not a dimension is mutually exclusive is also indicated in Table 4, where we introduce the dimensions and characteristics in detail.

XSL•FO
RenderX

Table 4. Taxonomy of AI[a]-based health care services.

Perspective and dimension	Characteristic	E[b] or N[c]
Agents		
Service recipient	• Patient • Health care professional	N
Mode of interaction	• AI-driven • Recipient-driven	E
Data		
Data generator	• Patient • Health care professional • Object	N
Data target	• Patient • Health care professional • Environment	N
Data type	• Structured • Unstructured	N
AI		
Hardware agnostic	• Yes • No	E
AI portfolio integration	• Wrapped around product • Wrapped around service • Stand-alone solution	E
AI capability	• Recognizing • Reasoning • Predicting • Decision-making • Generating • Acting	N
Health impact		
Application area	• Prevention • Diagnostics • Treatment and care • Patient enablement	N
Health benefit	• Triage • Physical • Mental • Social	N

[a]AI: artificial intelligence.

[b]E: mutually exclusive.

[c]N: mutually nonexclusive.

Agents Perspective

In this perspective, one looks at the entities involved in the provision of a service and the ways in which they interact during this service provision [46,71]. Therefore, this perspective has 2 dimensions, that is, the service recipient and the mode of interaction. The service recipient is the primary beneficiary of the service. Even though most health services ultimately aim to improve or maintain the health of patients, this category refers to the direct recipient of a given service [37]. Health care services can have various stakeholders [49], but during the development process of this study, we learned from our empirical iterations that only 2 potential entities profit directly from AI-based health care services, that is, the patient and the health care professional. How these entities are involved in the service communication and who triggers the information exchange is determined in the dimension mode of interaction. The interaction can be either AI-driven, meaning the service itself actively initiates interaction and requests information, or

XSL•FO
RenderX

recipient-driven, meaning the exchange of information is triggered by the service user [27].

Data Perspective

In this perspective, one examines the data that are generated during a service. This perspective, then, is predicated on the foundation of the underlying AI algorithm. It comprises 3 dimensions: the data generator, the data target, and the data type. As Hunke et al [44] noted, analytics-based services are characterized by a data generator and a data target. A data generator describes the entity that actively generates the data, while the data target describes the source from which the data emerges [44]. Since this also applies to our topic of investigation, we incorporated these 2 dimensions into our taxonomy once we made the necessary adaptations to the new context of our study. For AI-based health care services, data can be generated either by the patient, the health care professional, or an object involved in the service (eg, a diagnostic device). These data contain information about the patient, the health care professional, or the patient's environment, and it can take the shape of different data types, being either structured or unstructured [38,52]. Structured data are characterized by enforced composition to anatomic data types and can be directly processed, for example, into laboratory values in electronic health records [42]. In contrast, unstructured data have neither a structured format nor a conceptual definition, such as image or audio data.

AI Perspective

From this perspective, one can analyze how AI is integrated into health care services. It comprises 3 dimensions: hardware agnostic, AI portfolio integration, and applied AI capabilities. The dimension of hardware agnostic is based on our E2C iterations. It allows one to ascertain whether or not a service relies on specific hardware or specific platform resources, making the AI-based service either device-independent ("yes") or device-dependent ("no"). AI portfolio integration indicates whether the AI is integrated into a new stand-alone solution, linked to an existing product, or connected to a current service that has already been provided [44]. The dimension of AI capabilities refers to the idiosyncratic characteristics of AI algorithms that enable cognitive tasks formerly performed by humans [15,17,18,40,43,45,48]. The identified tasks include recognizing patterns or concepts, reasoning about relationships among distinct variables, predicting future outcomes or conditions, deciding between discrete options, generating something new, or even just acting.

Health Impact Perspective

In this final perspective, one focuses on the delivery of an AI-based service and its impact on health care [47]. There are 2 dimensions to this: one must identify the exact application area in the health care sector as well as the health benefit achieved by the service. After all, the incorporation of AI in the provision of health care services can impact different application areas in the health care sector. By fostering healthy lifestyles or avoiding unhealthy habits, AI-based services can help humans prevent new diseases. They can also help medical professionals refine their disease diagnosis and improve their treatment and care of patients [50]. Furthermore, our empirical investigation of AI-based health care services has shown that AI techniques can empower patients or professionals to actively handle a disease or reduce the adverse impact of a disability. In addition, this positive impact is scalable to a broad range of health care domains, where AI-based health care services can leverage various health benefits. In accordance with the World Health Organization's definition of health [14], AI-based health care services can impact the physical and psychological as well as the social aspects of health. Indeed, AI techniques can even refine medical triage to facilitate a more targeted treatment.

Demonstration and Evaluation

To validate the robustness of classifications made with our taxonomy, we recruited 13 researchers in the fields of AI, health care, or both. The overall demonstration phase took 3 weeks and was analyzed by 2 of the authors (MB and PK). We tasked the 13 researchers to classify 4 service offerings from our dataset anonymously. In choosing the service offerings for the application of our taxonomy, we made sure to cover a wide range of characteristics to provide a holistic evaluation of the taxonomy. Each service provider supplied the participating researchers with a brief description of their service along with a few images to illustrate how it works. As a further quality control measure, we made sure that there could be no confusion about the task assigned to the participants by giving them descriptions of all the dimensions and characteristics that they would have to consider when using the taxonomy, and we tested the comprehensibility of the questions as well as the appropriateness of the supplementary texts in advance. As a result of these measures, the percentage agreement among all authors and concerning all dimensions was an excellent 93.4%. Complementarily, the overall percentage agreement among all participants was 88% [72]. Fleiss κ is 62%, indicating a "substantial" agreement among participants [73]. For a detailed overview of percentage agreement and Fleiss κ across all individual dimensions, see "Taxonomy Evaluation" in Multimedia Appendix 1 [61]. It is suffice it to say that we conclude the taxonomy to be concise, robust, comprehensive, extendible, and explanatory [24].

In the following phase, we classified all 268 health care services to reveal the absolute and relative frequencies of all characteristics. When analyzing the relative frequencies, as presented in Table 5, some significant observations can be made, such as the health care professional is a direct recipient in 70% (n=186) of services, and the service recipient triggers the exchange of information in 79% (n=211) of services. Also noteworthy is that 74% (n=198) of data generators are objects, such as wearable devices. In 95% (254) of services, the primary data target is the patient, while 70% (n=188) of the data underlying the service are unstructured. As for the role of AI, 84% (n=223) of services are hardware agnostic, which is to say that they do not require a specific device to deploy the service. Also of interest in this context is the statistic that 83% (n=222) of the services in our sample were offered as a stand-alone solution.

XSL•FO
RenderX

Table 5. Absolute and relative frequencies of the characteristics of 268 AI[a]-based health care services.

Perspective and dimension	Characteristic	Frequency (N=268), n[b] (%[c])
Agents		
Service recipient	• Patient • Health care professional	• 96 (36) • 186 (70)
Mode of interaction	• AI-driven • Recipient-driven	• 56 (21) • 211 (79)
Data		
Data generator	• Patient • Health care professional • Object	• 72 (27) • 26 (10) • 198 (74)
Data target	• Patient • Health care professional • Environment	• 254 (95) • 7 (3) • 23 (9)
Data type	• Structured • Unstructured	• 95 (36) • 188 (70)
AI		
Hardware agnostic	• Yes • No	• 223 (84) • 44 (16)
AI portfolio integration	• Wrapped around product • Wrapped around service • Stand-alone solution	• 22 (8) • 23 (9) • 222 (83)
AI capability	• Recognizing • Reasoning • Predicting • Decision-making • Generating • Acting	• 257 (96) • 34 (13) • 28 (10) • 75 (28) • 89 (33) • 30 (11)
Health impact		
Application area	• Prevention • Diagnostics • Treatment and care • Patient enablement	• 57 (21) • 163 (61) • 83 (31) • 16 (6)
Health benefit	• Triage • Physical • Mental • Social	• 25 (9) • 231 (87) • 18 (7) • 4 (1)

[a]AI: artificial intelligence.

[b]Absolute frequency.

[c]Relative frequency.

A deep dive into AI capabilities revealed that 96% (n=257) of services can recognize patterns within audio, video, or other health-related data. At 33% (n=89), the second most frequent AI capability is to generate something new based on available input data. Examples are structured consultation protocols based on sound recordings or descriptions of injuries based on image files. Approximately 28% (n=75) of observed services include the AI capability to perform decision-making, for instance, to decide between multiple treatment alternatives. Another 13% (n=34) of observed services have the AI capability of reasoning, which offers the medical professional the significant benefit of delegating the task of explaining the link between bad habits and potential diseases. The rarest observed AI capabilities were acting (n=30, 11%), for example, calling emergency staff for help, and predicting (n=28, 10%), for example, being able to tell in advance whether people with specific characteristics will develop a particular disease. The final statistic of note is that most of our sample services are used for diagnostics (n=163, 61%). Second in demand (n=83, 31%) is the type of service used in treatment and care. At 87% (n=231), the vast majority

XSL•FO
RenderX

of our sample services are used to treat physical illness or injury. In comparison, 9% (n=25) are used for medical triage. Only a minority of our sample services use AI to improve mental (n=18, 7%) or social (n=4, 1%) health.

Archetypes and Interpretation

To ensure that our taxonomy can identify separable archetypes despite its extensive range, we performed the cluster analysis analogous to Gimpel et al [27], which is to say that we performed it for each of the identified dimensions separately. This cluster analysis revealed 13 archetypes of AI-based health care services that differ in their dominant design characteristics. We identified 2 archetypes in the agent perspective, 5 archetypes in the data perspective, 2 archetypes in the AI perspective, and 4 archetypes in the health impact perspective. Table 6 presents the respective number of services that can be assigned to each archetype as well as the absolute and relative frequencies of their characteristics.

First, the agent perspective includes the 2 archetypes, “AI-initiated” and “recipient-initiated.” Service offerings of the AI-initiated archetype actively request information from the user. They are designed for use by both the patient (27/57, 47%) and the health care professional (35/57, 61%). In contrast, the recipient-initiated archetype targets the medical professional in the majority of cases (152/211, 72%).

Second, the data perspective includes 5 archetypes. These mainly differ from one another in the origin and the type of the used data. What the archetypes have in common, however, is that their data are primarily related to the patient. The “structured object” archetype is characterized by the use of structured data (100%) that are provided by an object (100%). Almost one-quarter of services also use the patient (8/44, 18%) or medical professionals (3/44, 7%) as a source of information. In contrast, services in the “unstructured object” archetype work with unstructured data (100%) that is primarily available in audio or image formats. Occasionally, this data include information about the patient’s environment (13/145, 9%) but rarely about the responsible medical personnel (1/145, 1%). The “structured nonobject” archetype contains structured data concerning the patient (100%). These data are primarily provided by the patient (32/34, 94%) and less so by the treating physician (6/34, 18%). Services of the “unstructured nonobject” archetype work exclusively with unstructured data (100%). However, compared to the “structured nonobject” archetype, they are almost twice as likely (9/28, 32%) to be generated by the responsible health care professional. As for the “structured-unstructured” archetype, this one is characterized by the simultaneous processing of structured and unstructured data (100%) that are obtained from objects (12/17, 71%), health care professionals (8/17, 47%), and patients (7/17, 41%).

Third, the AI perspective comprises the 2 archetypes, “device-dependent” and “device-independent.” The former includes the majority of services in the sample (222/268, 82.8%) and operates with generic hardware. Services belonging to this archetype are primarily offered as stand-alone solutions (206/222, 93%). The underlying AI can recognize patterns and other correlations (215/222, 97%), decide between separate alternatives in approximately one-third of the cases (64/222, 29%), or generate output (79/222, 36%). To a lesser extent, it is also able to provide reasons for correlations (27/222, 12%), make predictions (23/222, 10%), and execute actions on its own (24/222, 11%). Services of the “device-dependent” archetype require the use of application-specific hardware. If supported by such hardware, this archetype can recognize contexts in a similar number of cases (44/46, 96%) as the previous archetype, but it makes decisions (10/46, 22%) or generates objects (10/46, 36%) in far fewer cases.

Fourth, the health impact perspective comprises 4 archetypes: “physical,” “diagnostic-physical,” “treatment,” and “triage.” Health care applications of the “physical” archetype provide material benefits to the patient (30/40, 75%). These are often achieved by means of preventive measures (28/40, 70%) or measures that help the patient to be more independent and self-determined in everyday life (11/40, 28%). Services of the “diagnostic-physical” archetype also focus on physical complaints (100%). In contrast to the previous archetype, however, they almost exclusively aim at disease diagnosis (131/132, 99%). Typically, services classified as belonging to the “treatment” archetype are used after the patient has been diagnosed (75/76, 99%). Another typical feature of these services is that they have a physical benefit in as much as 87% (66/76) of cases and a psychological benefit in as little as 13% (10/76). As for services of the “triage” archetype, these address patient referral to the appropriate health care specialist (100%). Given the comparatively small number of services in our sample (n=20), the underlying data may not be representative of real-world services. However, in three-quarters (15/20, 75%) of the cases, the occurrence of this archetype is closely linked to the diagnostic process. Other use cases are found mainly in the prevention (4/20, 20%) and treatment (9/20, 45%) of previously diagnosed diseases.

XSL•FO
RenderX

Table 6. Archetypes of 268 AI[a]-based health care services.

Archetype and dimension	Characteristics	Frequency (N=268), n[b] (%[c])
Agents		
AI-initiated (n=57)		
Service recipient	• Patient • Health care professional	• 27 (47.4) • 35 (61.4)
Mode of interaction	• AI-driven • Recipient-driven	• 57 (100) • 0 (0)
Recipient-initiated (n=211)		
Service recipient	• Patient • Health care professional	• 69 (32.7) • 152 (72)
Mode of interaction	• AI-driven • Recipient-driven	• 0 (0) • 211 (100)
Data		
Structured object (n=44)		
Data generator	• Patient • Health care professional • Object	• 8 (18.2) • 3 (6.8) • 44 (100)
Data target	• Patient • Health care professional • Environment	• 42 (95.5) • 2 (4.5) • 1 (2.3)
Data type	• Structured • Unstructured	• 44 (100) • 0 (0)
Unstructured object (n=145)		
Data generator	• Patient • Health care professional • Object	• 4 (2.8) • 1 (0.7) • 145 (100)
Data target	• Patient • Health care professional • Environment	• 137 (94.5) • 3 (2.1) • 13 (9)
Data type	• Structured • Unstructured	• 0 (0) • 145 (100)
Structured nonobject (n=34)		
Data generator	• Patient • Health care professional • Object	• 32 (94.1) • 6 (17.6) • 0 (0)
Data target	• Patient • Health care professional • Environment	• 34 (100) • 0 (0) • 5 (14.7)
Data type	• Structured • Unstructured	• 34 (100) • 0 (0)
Unstructured nonobject (n=28)		
Data generator	• Patient • Health care professional • Object	• 22 (78.6) • 9 (32.1) • 0 (0)

XSL•FO
RenderX

Archetype and dimension	Characteristics	Frequency (N=268), n[b] (%[c])
Data target	• Patient • Health care professional • Environment	• 26 (92.9) • 2 (7.1) • 3 (10.7)
Data type	• Structured • Unstructured	• 0 (0) • 28 (100)
Structured-unstructured (n=17)		
Data generator	• Patient • Health care professional • Object	• 7 (41.2) • 8 (47.1) • 12 (70.6)
Data target	• Patient • Health care professional • Environment	• 17 (100) • 0 (0) • 0 (0)
Data type	• Structured • Unstructured	• 17 (100) • 17 (100)
AI		
Device-independent (n=222)		
Hardware agnostic	• Yes • No	• 222 (100) • 0 (0)
AI portfolio integration	• Wrapped around product • Wrapped around service • Stand-alone solution	• 1 (0.5) • 15 (6.8) • 206 (92.8)
AI capability	• Recognizing • Reasoning • Predicting • Decision-making • Generating • Acting	• 215 (96.8) • 27 (12.2) • 23 (10.4) • 64 (28.8) • 79 (35.6) • 24 (10.8)
Device-dependent (n=46)		
Hardware agnostic	• Yes • No	• 0 (0) • 46 (100)
AI portfolio integration	• Wrapped around product • Wrapped around service • Stand-alone solution	• 9 (19.6) • 0 (0) • 37 (80.4)
AI capability	• Recognizing • Reasoning • Predicting • Decision-making • Generating • Acting	• 44 (95.7) • 6 (13) • 5 (10.9) • 10 (21.7) • 10 (21.7) • 7 (15.2)
Health Impact		
Physical (n=40)		

XSL•FO
RenderX

Archetype and dimension	Characteristics	Frequency (N=268), n[b] (%[c])
Application area	• Prevention • Diagnostics • Treatment and care • Patient enablement	• 28 (70) • 1 (2.5) • 1 (2.5) • 11 (27.5)
Health benefit	• Triage • Physical • Mental • Social	• 1 (2.5) • 30 (75) • 6 (15) • 3 (7.5)
Diagnostic-physical (n=132)		
Application area	• Prevention • Diagnostics • Treatment and care • Patient enablement	• 18 (13.6) • 131 (99.2) • 0 (0) • 0 (0)
Health benefit	• Triage • Physical • Mental • Social	• 5 (3.8) • 132 (100) • 2 (1.5) • 0 (0)
Treatment (n=76)		
Application area	• Prevention • Diagnostics • Treatment and care • Patient enablement	• 5 (6.6) • 17 (22.4) • 75 (98.7) • 2 (2.6)
Health benefit	• Triage • Physical • Mental • Social	• 0 (0) • 66 (86.8) • 10 (13.2) • 1 (1.3)
Triage (n=20)		
Application area	• Prevention • Diagnostics • Treatment and care • Patient enablement	• 4 (20) • 15 (75) • 9 (45) • 3 (15)
Health benefit	• Triage • Physical • Mental • Social	• 20 (100) • 4 (20) • 0 (0) • 0 (0)

[a]AI: artificial intelligence.

[b]Absolute frequency.

[c]Relative frequency.

In summary, the agent archetypes differ mainly in how and by whom information is requested. It is worth noting that approximately 8 (80%) out of 10 services require manual process initiation, which means they need human interaction. This reliance on the interaction between humans and AI is also predicated on the fact that both archetypes are directed at trained health care professionals in more than half the cases. When it comes to the “data” perspective, we found the distinguishing characteristics to be the type of data and the data generator. Potentially, this can be explained by the fact that all 5 archetypes focus predominantly on patients as data targets. This is logical since information about a patient’s complaints, symptoms, or medical history concerns that patient directly. It is questionable, however, whether AI-based applications in the health care sector are still making insufficient use of some of the available data, for example, environmental information about a patient’s working or living conditions. As for the archetypes of the “AI” perspective, they distinguish between health care service offerings according to their dependency on specific hardware. This is consistent with the aforementioned high proportion of applications for image file interpretation, as these typically do not involve hardware components. Another reason for the high proportion of device-independent applications could be the complexity and cost of product development. This brings us to the final point—the archetypes in the “health impact” perspective. Their analysis focuses on patient benefit and shows that, to date, AI-based applications are primarily used for the diagnosis of physical illnesses. Given the increasing life

XSL•FO
RenderX

expectancy of most industrialized nations, it is conceivable that the importance of AI in treatment and care, especially in geriatric care, might grow considerably in the near future. As for the other seismic shift in the technologically advanced West—our rapidly increasing awareness of mental health as a medical, social, and indeed political issue, it remains to be seen whether the number of services for the diagnosis and treatment of mental illnesses will also increase in the future.

Discussion

Principal Findings

To account for the phenomenon of AI-based health care services, we developed a new taxonomy based on the latest science [24], a structured literature review, and an extensive sample of real-world instantiations. As our rigorous demonstration and evaluation confirmed, this due diligence in the development of our taxonomy ensured its broad applicability and central positioning in the intended research field. As a result, this study contributes to the knowledge of how to incorporate AI into health care services successfully. Furthermore, it provides a clear and comprehensive structure to a fast-developing research field. Although there may be questions about the usefulness of taxonomy in such a dynamic field, this study indicates that a phenomenon does not have to be static for it to benefit from a taxonomy. As we have seen in similarly fast-developing research areas, such as the Internet of Things [26], FinTechs [27], analytics-based services [44], or Blockchain [74], there have been considerable benefits as a result of taxonomy development.

Meanwhile, our research results have theoretical as well as managerial implications. From a theoretical point of view, our research is the first to investigate the structure of AI-based health care services based on perspectives, dimensions, respective characteristics, and overall archetypes. The taxonomy supports academics and practitioners alike in that it enables them to recognize and re-evaluate the multiple forms of AI-based health care services, either to use those services to their full practical potential or to use the taxonomy for further theorizing. The archetypes we have identified in each perspective reveal various service offerings to have recurring patterns. These patterns advance the current understanding of how digital technologies transform the health care sector with the use of AI. Most importantly, this study incentivizes more extensive research on application areas in which AI has not yet been used to its full potential. Doing so encourages scholars to explore the multiple opportunities to apply AI in the ever more significant sector of digital health care.

From a practical point of view, our taxonomy and archetypes provide multiple use cases for practitioners. First, it is worth noting that practitioners who have already integrated AI into their health care service may find our taxonomy helpful in classifying this service and comparing it to any other offerings by their competitors. Second, practitioners who wish to introduce new AI-based health care services can use our taxonomy to gain an overview of possible applications, be it to find suitable options that are already out there or to find inspiration for potential new applications in their working fields. By analyzing numerous real-world examples, we were able to shed light on the intricacies of AI-based health care services and the general distribution of their design characteristics. By using our archetypes along with our insights into their frequencies, practitioners can identify which potential AI adoption fields are already widely used and might, therefore, be easier and more accessible areas in which to implement their services. While this focus provides limited potential for future innovations, our taxonomy facilitates the assessment of alternative possibilities in less common adoption fields, where it can be used for the dual benefits of orientation and structure. Third, the insights derived from our taxonomy on the characteristics of AI-based health care services carry substantial implications for health policy and delivery. By categorizing AI services based on their unique features, policy makers can make informed decisions on resource allocation, regulation, and integration into existing health care systems. This taxonomy facilitates the identification of targeted interventions for improving patient care, such as personalized treatment plans and enhanced diagnostic accuracy. Fourth, understanding the characteristics of AI-based services aids in the development of policies that address ethical considerations, privacy concerns, and the need for transparency in health care AI implementation. Fifth, in the context of population health, the taxonomy enables the strategic deployment of AI technologies to address specific health challenges, reduce disparities, and optimize overall health outcomes. This discussion underscores the practical relevance of our taxonomy, extending beyond theoretical considerations to directly influence health care policy making and health care delivery.

Of course, this study is also subject to certain limitations that require careful consideration to ensure the correct interpretation of our results. The first of those limitations to address is that our classification, as to whether or not a service is AI-driven, is based on the external appearance given by the service provider's website or information supplied by other media. It cannot, therefore, be guaranteed that all of the services considered in this study meet the academic criteria to be classified as AI-driven. There is a hidden benefit in this putative limitation, however. Our research approach does not focus on how AI is seen in research but rather on how AI is perceived and applied in practice. The second limitation to discuss is that AI is a rapidly evolving field, and as new developments overtake the research presented in these pages, our taxonomy may need to be adapted and extended accordingly. For example, the sample of service offerings that we analyzed has a noticeably high frequency of diagnostics-related services due to our definition of the term service and AI algorithms that were primarily used for making decisions and classifying symptoms. However, current developments in the topics of generative AI or robotics might shift this trend toward treatment and care or patient enablement. With that in mind, we were especially thorough in our structured analysis of current AI adoption and emphasized the extensibility of our taxonomy to provide the necessary guidance for future adaptations and extensions in research and practice. The third limitation worth considering is that we based our research on empirical insights from top lists. While our data collection was rigorous, our selection of the publicly available AI-based health care services only represents a subset of the current services. As for the fourth limitation of

XSL•FO
RenderX

this study, while we are confident that our taxonomy offers a good representation of the current AI landscape in health care, it will undoubtedly need to be changed and adapted due to rapid developments at a technical level [75]. This could particularly affect the dimensions of AI capability and the combination with the application area. However, the most significant change will probably be seen in the archetypes. A comparison between the current and future status would undoubtedly be interesting for future research.

Conclusions and Outlook

Being a relatively new paradigm in health care, AI has been heralded as a panacea, yet only the future will tell if it can deliver on those lofty expectations. In the meantime, we have developed a taxonomy that serves as a structuring tool for researchers in the field of AI and as an innovation map for practitioners. It provides a simple yet potent method of classifying current and future AI-based health care services. Such classification allows the user to better situate and evaluate these services by attributing their distinct characteristics to 10 potential dimensions, such as service recipient, mode of interaction, data generator, data target, data type, hardware agnostic, AI portfolio integration, AI capability, application area, and health benefit. Having tested our taxonomy with representatives of its intended future users in order to classify 268 AI-based health care services and identify all relevant archetypes, we were able to ascertain its applicability and assess the potential for AI integration in current health care services.

By way of conclusion, then, it remains only to be said that this study was conducted in the intellectual tradition of Nickerson et al [24] and Kundisch et al [25]. As such, it contributes to the expert literature by providing a common understanding of AI-based health care services and offering a taxonomy that promises to assist in the development of AI-based health care services and the assessment of their potential in this competitive field. It bears repeating, however, that the implementation of AI in health care is a young and rapidly developing research field, so although we are convinced that this taxonomy can serve as a useful analytical tool in the present, its future usefulness will depend on its adaptation to relevant advances in the field. We hope that our work here will inspire the future research required to identify those advances and make the necessary adaptations to our taxonomy.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Taxonomy evaluation and clustering.
[DOCX File , 93 KB - jmir_v26i1e53986_app1.docx]

References

1. Maier E, Reimer U, Wickramasinghe N. Digital healthcare services. Electronic Markets 2021;31(4):743-746 [FREE Full text] [doi: 10.1007/s12525-021-00513-z] [Medline: 35602115]
2. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc Journal 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
3. Mehta N, Pandit A, Shukla S. Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study. Journal of Biomedical Informatics 2019;100:103311 [FREE Full text] [doi: 10.1016/j.jbi.2019.103311] [Medline: 31629922]
4. Nordling L. A fairer way forward for AI in health care. Nature 2019;573(7775):S103-S105. [doi: 10.1038/d41586-019-02872-2] [Medline: 31554993]
5. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature Biomedical Engineering 2018;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
6. Agarwal R, Gao G, DesRoches C, Jha K. Research commentary—the digital transformation of healthcare: current status and the road ahead. Information Systems Research 2010 Dec;21(4):796-809. [doi: 10.1287/isre.1100.0327]
7. Chung H, Ko H, Kang WS, Kim KW, Lee H, Park C, et al. Prediction and feature importance analysis for severity of COVID-19 in South Korea using artificial intelligence: model development and validation. Journal of Medical Internet Research 2021;23(4):e27060 [FREE Full text] [doi: 10.2196/27060] [Medline: 33764883]
8. Krämer D, Brachem E, Schneider-Reuter L, D'Angelo I, Vollmann J, Haltaufderheide J. Smartphone apps for containing the COVID-19 pandemic in Germany: qualitative interview study with experts based on grounded theory. J Med Internet Res 2023;25:e45549 [FREE Full text] [doi: 10.2196/45549] [Medline: 37862068]
9. Velazquez-Diaz D, Arco JE, Ortiz A, Pérez-Cabezas V, Lucena-Anton D, Moral-Munoz JA, et al. Use of artificial intelligence in the identification and diagnosis of frailty syndrome in older adults: scoping review. J Med Internet Res 2023;25:e47346 [FREE Full text] [doi: 10.2196/47346] [Medline: 37862082]
10. Sekandi JN, Shi W, Zhu R, Kaggwa P, Mwebaze E, Li S. Application of artificial intelligence to the monitoring of medication adherence for tuberculosis treatment in Africa: algorithm development and validation. JMIR AI 2023;2(1):e40167 [FREE Full text] [doi: 10.2196/40167] [Medline: 38464947]

XSL•FO
RenderX

11. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
12. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
13. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Service Research 2018;18(1):545 [FREE Full text] [doi: 10.1186/s12913-018-3359-4] [Medline: 30001717]
14. Basic documents - forty-ninth edition: constitution of the world health organization. World Health Organization. 1948. URL: https://apps.who.int/gb/bd/ [accessed 2024-08-27]
15. Agrawal A, Gans J, Goldfarb A. Prediction Machines: The Simple Economics of Artificial Intelligence. Boston, MA: Harvard Business Review Press; 2018.
16. Mitchell TM. Machine Learning. New York, NY: McGraw-Hill; 1997.
17. Ågerfalk PJ. Artificial intelligence as digital agency. European Journal of Information Systems 2020;29(1):1-8. [doi: 10.1080/0960085x.2020.1721947]
18. Rai AP, Constantinides P, Sarker S. Next generation digital platforms : toward human-AI hybrids. MIS Q 2019;43:iii-iix [FREE Full text]
19. Hofmann P, Oesterle S, Rust P, Urbach N. Machine learning approaches along the radiology value chain - rethinking value propositions. 2019 Presented at: Proceedings of the 27th European Conference on Information Systems (ECIS); August 27, 2019; Stockholm-Uppsala, Sweden URL: https://www.fim-rc.de/Paperbibliothek/Veroeffentlicht/845/wi-845.pdf
20. Transforming healthcare with AI: the impact on the workforce and organizations. McKinsey & Company. 2020. URL: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/transforming-healthcare-with-ai [accessed 2024-08-27]
21. Das TK, Chowdhary CL, Gao X. Chest x-ray investigation: a convolutional neural network approach. Journal of Biomimetics, Biomaterials and Biomedical Engineering 2020;45:57-70. [doi: 10.4028/www.scientific.net/jbbbe.45.57]
22. Meier P, Beinke JH, Fitte C, Behne A, Teuteberg F. FeelFit – design and evaluation of a conversational agent to enhance health awareness. 2019 Presented at: ICIS 2019; Dec 15-18, 2019; Munich, Germany URL: https://aisel.aisnet.org/icis2019/is_health/is_health/22
23. Reis L, Maier C, Mattke J, Weitzel T. Chatbots in healthcare: Status quo, application scenarios for physicians and patients and future directions. 2020 Presented at: Proceedings of the 28th European Conference on Information Systems (ECIS 2020); June 15-17, 2020; Online URL: https://aisel.aisnet.org/ecis2020_rp/163
24. Nickerson RC, Varshney U, Muntermann J. A method for taxonomy development and its application in information systems. Eur J Inf Syst 2012;22(3):336-359. [doi: 10.1057/ejis.2012.26]
25. Kundisch D, Muntermann J, Oberländer AM, Rau D, Röglinger M, Schoormann T, et al. An update for taxonomy designers. Bus Inf Syst Eng 2021;64(4):421-439. [doi: 10.1007/s12599-021-00723-x]
26. Fischer M, Heim D, Hofmann A, Janiesch C, Klima C, Winkelmann A. A taxonomy and archetypes of smart services for smart living. Electron Markets 2020;30(1):131-149. [doi: 10.1007/s12525-019-00384-5]
27. Gimpel H, Rau D, Röglinger M. Understanding FinTech start-ups – a taxonomy of consumer-oriented service offerings. Electron Markets 2017;28(3):245-264. [doi: 10.1007/s12525-017-0275-0]
28. Glass RL, Vessey I. Contemporary application-domain taxonomies. IEEE Softw 1995;12(4):63-76. [doi: 10.1109/52.391837]
29. Labes S, Erek K, Zarnekow R. Common patterns of cloud business models. 2013 Presented at: Proceedings of the 19th Americas Conference on Information Systems (AMCIS); August 15-17, 2013; Chicago, Illinois URL: https://www.semanticscholar.org/paper/Common-Patterns-of-Cloud-Business-Models-Labes-Erek/a2b172c953967e3dde8d3fc7b39230c1aa91b729
30. Maglio PP, Lim CH. Innovation and big data in smart service systems. jim 2016;4(1):11-21. [doi: 10.24840/2183-0606_004.001_0003]
31. AI 50 2020: America's most promising artificial intelligence companies. Forbes. 2020. URL: https://www.forbes.com/sites/alanohnsman/2020/07/03/ai-50-americas-most-promising-artificial-intelligence-companies/?sh=2b21fca55c99 [accessed 2024-08-27]
32. Top 203 AI startups in healthcare. Medicalstartups. 2020. URL: https://www.medicalstartups.org/top/ai/ [accessed 2024-08-27]
33. 6 Top artificial intelligence start-ups. Startus. 2019. URL: https://www.startus-insights.com/innovators-guide/6-top-artificial-intelligence-startups-out-of-1038/ [accessed 2024-08-27]
34. From drug R & D to diagnostics: 90+ artificial intelligence startups in healthcare. CBInsights. 2019 Sep 12. URL: https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/ [accessed 2024-08-27]
35. Top 10 healthcare AI companies. QWayHealth. 2020. URL: https://qwayhealthcare.com/blog/top-10-healthcare-ai-companies/ [accessed 2024-08-27]
36. Top artificial intelligence companies in healthcare to keep an eye on. Medicalfuturist. 2020. URL: https://medicalfuturist.com/top-artificial-intelligence-companies-in-healthcare/ [accessed 2024-08-27]
37. Bao C, Bardhan IR, Singh H, Meyer BA, Kirksey K. Patient–provider engagement and its impact on health outcomes: a longitudinal study of patient portal use. MISQ 2020;44(2):699-723. [doi: 10.25300/misq/2020/14180]

XSL•FO
RenderX

38. Baars H, Kemper H. Management support with structured and unstructured data—an integrated business intelligence framework. Information Systems Management 2008;25(2):132-148. [doi: 10.1080/10580530801941058]
39. Bardhan I, Oh J, Zheng Z, Kirksey K. Predictive analytics for readmission of patients with congestive heart failure. Information Systems Research 2015;26(1):19-39. [doi: 10.1287/isre.2014.0553]
40. Brynjolfsson E, Mitchell T. What can machine learning do? Workforce implications. Science 2017;358(6370):1530-1534. [doi: 10.1126/science.aap8062] [Medline: 29269459]
41. Chang A. The role of artificial intelligence in digital health. In: Meyers A, editor. Digital Health Entrepreneurship. Cham, Switzerland: Springer International; 2023:75-85.
42. Hansen S, Baroody AJ. Electronic health records and the logics of care: complementarity and conflict in the U.S. healthcare system. Information Systems Research 2020;31(1):57-75. [doi: 10.1287/isre.2019.0875]
43. Hofmann P, Jöhnk J, Protschky D, Stähle P, Urbach N, Buck C. KI-Anwendungsfälle zielgerichtet identifizieren. Wirtschaftsinformatik & Management 2020 1;12(3):184-193. [doi: 10.1365/s35764-020-00257-z]
44. Hunke F, Engel C, Schüritz R, Ebel P. Understanding the anatomy of analytics-based services - a taxonomy to conceptualize the use of data and analytics in services. 2019 Presented at: Proceedings of the 27th European Conference on Information Systems (ECIS); June 8-14, 2019; Stockholm and Uppsala, Sweden URL: https://aisel.aisnet.org/ecis2019_rp/25
45. Iansiti M, Lakhani KR. Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World. Boston, MA: Harvard Business Review Press; 2020.
46. Mantzana V, Themistocleous M, Irani Z, Morabito V. Identifying healthcare actors involved in the adoption of information systems. European Journal of Information Systems 2017;16(1):91-102. [doi: 10.1057/palgrave.ejis.3000660]
47. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. Journal of the Royal Society of Medicine 2019;112(1):22-28 [FREE Full text] [doi: 10.1177/0141076818815510] [Medline: 30507284]
48. Schuetz S, Venkatesh V. "Research perspectives: the rise of human machines: how cognitive computing systems challenge assumptions of user-system interaction". Journal of the Association for Information Systems 2020;21(2):460-482. [doi: 10.17705/1jais.00608]
49. Tan C, Sun L, Liu K. Big data architecture for pervasive healthcare: a literature review. 2015 Presented at: Proceedings of the 23th European Conference on Information Systems (ECIS); 2015 May 26-29; Munster, Germany URL: https://aisel.aisnet.org/ecis2015_cr/177
50. Thompson S, Whitaker J, Kohli R, Jones C. Chronic disease management: how IT and analytics create healthcare value through the temporal displacement of care. MISQ 2020;44(1):227-256. [doi: 10.25300/misq/2020/15085]
51. Wani SUD, Khan NA, Thakur G, Gautam SP, Ali M, Alam P, et al. Utilization of artificial intelligence in disease prevention: diagnosis, treatment, and implications for the healthcare workforce. Healthcare (Basel) 2022;10(4):608 [FREE Full text] [doi: 10.3390/healthcare10040608] [Medline: 35455786]
52. Weglarz G. Two worlds of data unstructured and structured. DM Review 2004;14(9):19 [FREE Full text]
53. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences 2021;11(11):5088. [doi: 10.3390/app11115088]
54. Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H. A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. Journal of Biomedical Informatics 2021;113:103627 [FREE Full text] [doi: 10.1016/j.jbi.2020.103627] [Medline: 33259944]
55. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. Journal of Innovation & Knowledge 2023;8(1):100333. [doi: 10.1016/j.jik.2023.100333]
56. Qadri YA, Nauman A, Zikria YB, Vasilakos AV, Kim SW. The future of healthcare internet of things: a survey of emerging technologies. IEEE Commun. Surv. Tutorials 2020;22(2):1121-1167. [doi: 10.1109/comst.2020.2973314]
57. Karatas M, Eriskin L, Deveci M, Pamucar D, Garg H. Big data for healthcare industry 4.0: applications, challenges and future perspectives. Expert Systems with Applications 2022;200:116912. [doi: 10.1016/j.eswa.2022.116912]
58. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychological medicine 2019;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
59. Bindra S, Jain R. Artificial intelligence in medical science: a review. Irish Journal of Medical Science 2024;193(3):1419-1429. [doi: 10.1007/s11845-023-03570-9] [Medline: 37952245]
60. Webster J, Watson RT. Analyzing the past to prepare for the future: writing a literature review. MISQ 2002;26(2):xiii-xxiii [FREE Full text]
61. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971;76(5):378-382. [doi: 10.1037/h0031619]
62. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
63. Oberländer AM, Röglinger M, Rosemann M. Digital opportunities for incumbents – a resource-centric perspective. The Journal of Strategic Information Systems 2021;30(3):101670. [doi: 10.1016/j.jsis.2021.101670]
64. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. Harlow, United Kingdom: Pearson; 2014.

XSL•FO
RenderX

65. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience; Blackwell Publishing; 2005.
66. Romesburg HC. Cluster Analysis for Researchers. Morrisville, NC: Lulu Press; 2004.
67. Cormack RM. A review of classification. Journal of the Royal Statistical Society. Series A (General) 1971;134(3):321. [doi: 10.2307/2344237]
68. Bacher J. Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren. München: Oldenbourg; 2010.
69. Rokach L. A survey of clustering algorithms. In: Maimon Z, editor. Data Mining and Knowledge Discovery Handbook. Springer: New York, Heidelberg; 2010.
70. Ward JH. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 1963;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]
71. Sharma D, Aujla GS, Bajaj R. Evolution from ancient medication to human-centered healthcare 4.0: a review on health care recommender systems. Int J Commun Syst 2019;36(12):e4058. [doi: 10.1002/dac.4058]
72. Nahm AY, Rao SS, Solis-Galvan LE, Ragu-Nathan TS. The Q-sort method: assessing reliability and construct validity of questionnaire items at a pre-testing stage. Journal of Modern Applied Statistical Methods 2002;1(1):114-125. [doi: 10.22237/jmasm/1020255360]
73. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159. [doi: 10.2307/2529310]
74. Tönnissen S, Beinke JH, Teuteberg F. Understanding token-based ecosystems – a taxonomy of blockchain-based business models of start-ups. Electronic Markets 2020;30(2):307-323. [doi: 10.1007/s12525-020-00396-6]
75. Berente N, Gu B, Recker J, Santhanam R. Managing artificial intelligence. MIS Q 2021;45(3):1433-1450 [FREE Full text]

Abbreviations

AI: artificial intelligence
C2E: conceptual-to-empirical
E2C: empirical-to-conceptual

Edited by A Mavragani; submitted 28.10.23; peer-reviewed by C Rose, W LaMendola, T Ramesh, N Domingues; comments to author 19.01.24; revised version received 08.02.24; accepted 18.04.24; published 27.11.24.

Please cite as:
Blaß M, Gimpel H, Karnebogen P
A Taxonomy and Archetypes of AI-Based Health Care Services: Qualitative Study
J Med Internet Res 2024;26:e53986
URL: https://www.jmir.org/2024/1/e53986
doi:10.2196/53986
PMID:

©Marlene Blaß, Henner Gimpel, Philip Karnebogen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Prediction Model Using Sella Magnetic Resonance Imaging–Based Radiomics and Clinical Parameters for the Diagnosis of Growth Hormone Deficiency and Idiopathic Short Stature: Cross-Sectional, Multicenter Study

Kyungchul Song[1], MD; Taehoon Ko[2,3,4], PhD; Hyun Wook Chae[1], MD, PhD; Jun Suk Oh[5], MD; Ho-Seong Kim[1], MD, PhD; Hyun Joo Shin[6], MD, PhD; Jeong-Ho Kim[7], MD, PhD; Ji-Hoon Na[1], MD, PhD; Chae Jung Park[6*], MD, PhD; Beomseok Sohn[8*], MD, PhD

[1]Department of Pediatrics, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[3]Department of Medical Sciences, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[4]The Catholic Medical Center Institute for Basic Medical Science, The Catholic Medical Center of The Catholic University of Korea, Seoul, Republic of Korea

[5]Deparment of Pediatrics, Konyang University College of Medicine, Daejeon, Republic of Korea

[6]Department of Radiology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin-si, Gyeonggi-do, Republic of Korea

[7]Department of Laboratory Medicine, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin-si, Gyeonggi-do, Republic of Korea

[8]Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Beomseok Sohn, MD, PhD
Department of Radiology
Samsung Medical Center
Sungkyunkwan University School of Medicine
81, Irwon-ro, Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 1049045034
Email: Beomseoksohn@gmail.com

Abstract

Background: Growth hormone deficiency (GHD) and idiopathic short stature (ISS) are the major etiologies of short stature in children. For the diagnosis of GHD and ISS, meticulous evaluations are required, including growth hormone provocation tests, which are invasive and burdensome for children. Additionally, sella magnetic resonance imaging (MRI) is necessary for assessing etiologies of GHD, which cannot evaluate hormonal secretion. Recently, radiomics has emerged as a revolutionary technique that uses mathematical algorithms to extract various features for the quantitative analysis of medical images.

Objective: This study aimed to develop a machine learning–based model using sella MRI–based radiomics and clinical parameters to diagnose GHD and ISS.

Methods: A total of 293 children with short stature who underwent sella MRI and growth hormone provocation tests were included in the training set, and 47 children who met the same inclusion criteria were enrolled in the test set from different hospitals for this study. A total of 186 radiomic features were extracted from the pituitary glands using a semiautomatic segmentation process for both the T2-weighted and contrast-enhanced T1-weighted image. The clinical parameters included auxological data, insulin-like growth factor-I, and bone age. The extreme gradient boosting algorithm was used to train the prediction models. Internal validation was conducted using 5-fold cross-validation on the training set, and external validation was conducted on the test set. Model performance was assessed by plotting the area under the receiver operating characteristic curve. The mean absolute Shapley values were computed to quantify the impact of each parameter.

XSL•FO
RenderX

Results: The area under the receiver operating characteristic curves (95% CIs) of the clinical, radiomics, and combined models were 0.684 (0.590-0.778), 0.691 (0.620-0.762), and 0.830 (0.741-0.919), respectively, in the external validation. Among the clinical parameters, the major contributing factors to prediction were BMI SD score (SDS), chronological age–bone age, weight SDS, growth velocity, and insulin-like growth factor-I SDS in the clinical model. In the combined model, radiomic features including maximum probability from a T2-weighted image and run length nonuniformity normalized from a T2-weighted image added incremental value to the prediction (combined model vs clinical model, *P*=.03; combined model vs radiomics model, *P*=.02). The code for our model is available in a public repository on GitHub.

Conclusions: Our model combining both radiomics and clinical parameters can accurately predict GHD from ISS, which was also proven in the external validation. These findings highlight the potential of machine learning–based models using radiomics and clinical parameters for diagnosing GHD and ISS.

(J Med Internet Res 2024;26:e54641) doi:10.2196/54641

KEYWORDS

dwarfism; pituitary; idiopathic short stature; child; adolescent; machine learning; magnetic resonance imaging; MRI

Introduction

Short stature, a height below the third percentile or more than 2 SDs below the corresponding mean height for those of the same sex, age, and race, is associated with psychosocial problems and medical conditions, such as diet, genes, physical activity, and underlying diseases [1-3]. Although short stature often represents a normal variation among the general population, negative social stereotypes associated with short stature still exist, resulting in poorer psychosocial performance in short children who are actually healthy [3,4]. As children with short stature can achieve average height with treatment with human recombinant growth hormone (GH), proper assessment and screening of short stature is very important for the physical and mental well-being of children. Moreover, proper treatment with human recombinant GH can reduce the cardiovascular risk of GH deficiency (GHD) [5]. Among etiologies of short stature, GHD and idiopathic short stature (ISS) account for the most common causes [2]. GH, a polypeptide hormone produced by the pituitary gland, stimulates linear bone growth and cell reproduction. GHD is defined as a condition induced by insufficient secretion of GH [1,4], whereas ISS is defined as short stature without evidence of systemic, endocrine, nutritional, or chromosomal abnormalities [1,4].

For the diagnosis of GHD, meticulous evaluation, including the measurement of anthropometric data, bone age, insulin-like growth factor-I (IGF-I), and GH provocation tests, is required, among which the GH provocation test is considered the gold standard [6,7]. GHD can be diagnosed in children with short stature who show insufficient GH levels after at least 2 GH provocation tests. However, the GH provocation test is extremely invasive and burdensome to patients and requires hospitalization and multiple blood samplings; therefore, investigations on noninvasive screening methods to replace the GH provocation test are required [8].

Etiologies of GHD include pathological causes, such as brain tumors and hypoxic brain damage; therefore, sella magnetic resonance imaging (MRI) is required for the evaluation of GHD [2]. Several studies investigated the difference in pituitary volume in sella MRI according to etiologies of short stature, and Kessler et al [9] reported that pituitary volume is different between children with GHD and ISS and that it increases with older age [10-12]. However, the SD score (SDS) of height, diameter, and volume of the pituitary gland was not different between GHD and ISS among Korean children in our previous study [12]. Moreover, hormonal secretion cannot be assessed in sella MRI.

Meanwhile, artificial intelligence (AI) is increasingly being leveraged as a novel approach in medical imaging research and diagnosis. Supervised machine learning serves as the cornerstone of radiological AI, wherein algorithms undergo training to identify pathologies, such as tumors, in computed tomography or MRI scans based on the gold standard [13,14]. These algorithms refine their diagnostic capabilities by learning from numerous cases and subsequently applying this acquired knowledge to identify such markers within new test cohorts containing unseen images. However, traditional AI in imaging analysis presents limitations, with diagnostic information often remaining obscured within the computational “black box,” offering merely simplistic outcomes, such as the presence of a lesion [15].

Thus, radiomics, a method that extracts various features using mathematical algorithms, has emerged as a revolutionary technique addressing these shortcomings by offering a quantitative image analysis framework [14,16]. Radiomics can be used to determine molecular profiles and disease characteristics that cannot be detected by the human eye [15,17]. Based on the concept of information in biomedical images that reflects the underlying pathophysiology, radiomics converts digital medical images into mineable high-dimensional data [15]. Although quantitative analyses of medical images have been performed in adults as numerous radiomic features can be extracted and analyzed using radiomics, investigations of the pituitary gland using radiomics in pediatrics are limited [15,18,19].

Notably, clinical parameters associated with GHD diagnosis have been investigated in several studies that included anthropometric data, such as height and BMI, and laboratory tests, such as IGF-1 [20-23]. In addition, a prediction model for the screening of GHD and ISS was suggested in a few studies [22,24]. However, the predictability of the previous studies using clinical parameters was limited. Moreover, literature regarding prediction models for the differential diagnosis of

XSL•FO
RenderX

GHD and ISS using both radiomics and clinical parameters or validation with the external set is lacking.

Recently, there has been a lot of research in the medical field that uses machine learning to create models to aid in clinical diagnosis. Neural networks, such as graph neural networks, are being used to diagnose Alzheimer disease by using structural MRI and positron emission tomography scans [25-27]. There are also studies on diagnosing respiratory diseases, such as COVID-19 and interstitial lung diseases, by converting respiratory or pulmonary sounds into spectrograms and classifying them with neural networks [28,29]. In contrast to unstructured data, such as medical images and sounds, gradient-boosting machines have been mainly used for structured data, such as electronic health records and vital signs. Oh et al [30] showed that extreme gradient boosting (XGBoost) can precisely estimate low-density lipoprotein cholesterol, a therapeutic target for dyslipidemia, using large-scale electronic health records. A light gradient boosting machine could predict cardiac arrest within 24 hours by training on heart rate variability calculated from electrocardiograms in the intensive care unit [31]. In the field of pediatrics, various machine learning models, such as random forests and support vector machines, have been used to predict early neonatal early-onset sepsis [32], and XGBoost has been used to identify children with Kawasaki disease in the pediatric emergency department [33]. However, there are not yet many studies in pediatrics that use machine learning models that consider both structured and unstructured data to aid in diagnosis.

Therefore, we aimed to develop a machine learning–based prediction model for the diagnosis of GHD and ISS using radiomics and clinical parameters, thereby overcoming the limitation of the clinical model with radiomics feature. In addition, we aimed to increase the reliability of the model with external validation. Our objectives were to (1) extract radiomic features using a T2-weighted image (T2WI) and contrast-enhanced T1-weighted image (T1C) in sella MRI; (2) develop a prediction model using both radiomic features and clinical parameters; (3) compare predictability among the models using radiomics, clinical parameters, and both parameters; (4) estimate the accuracy of the predictive models with external validation; and (5) evaluate the contribution of each clinical parameter and radiomic feature from the prediction models. To achieve this goal, we investigated the following contents: (1) baseline characteristics of the participants; (2) receiver operating characteristic (ROC) curve analyses of clinical, radiomics, and combined models; and (3) Shapley value of clinical parameters and radiomic features.

The chapters of this paper are organized as follows. First, the *Methods* chapter describes the study population and corresponding dataset, data preprocessing, machine learning methods, and interpretation. In the *Results* chapter, we describe the baseline characteristics of the study population, the performance of the machine learning models in diagnosing GHD and ISS, and our interpretation of the results. We then discuss the implications and limitations of our findings and finally summarize our findings in the *Discussion* chapter.

Methods

Study Population

Figure 1 shows the flowchart of this retrospective study. To develop a prediction model for the diagnosis of GHD and ISS, electronic records of children aged 18 years or younger with short stature who underwent GH provocation test and sella MRI between March 2011 and July 2020 were retrieved from the Clinical Data Repository System of Severance Hospital. Among these, participants with endocrinological or systemic pathology or those with pituitary lesions were excluded from the final derivation set. For the external validation set, electronic records of children aged 18 years or younger with short stature who underwent GH provocation test and sella MRI between September 2020 and November 2022 were retrieved from the Clinical Data Repository System of Yongin Severance Hospital. The exclusion criteria for the final external validation set were the same as those for the derivation set. Finally, a total of 293 children with MRI findings in the training set and 47 children in the test set from different hospitals were enrolled.

XSL•FO
RenderX

Figure 1. Study flowchart. *Endocrinological or systemic pathology: hypopituitarism; adrenal insufficiency; hyperthyroidism or hypothyroidism except euthyroid state; skeletal dysplasia; small for gestational age; genetic disease including chromosomal abnormalities; Russel-Silver syndrome; Prader-Willi syndrome; and chronic diseases, including a history of brain irradiation, congenital heart disease, and systemic lupus erythematosus. **Pituitary lesion: pituitary tumor or empty sella. EHR: electronic health record; GH: growth hormone; GHD: growth hormone deficiency; ISS: idiopathic short stature; MRI: magnetic resonance imaging.

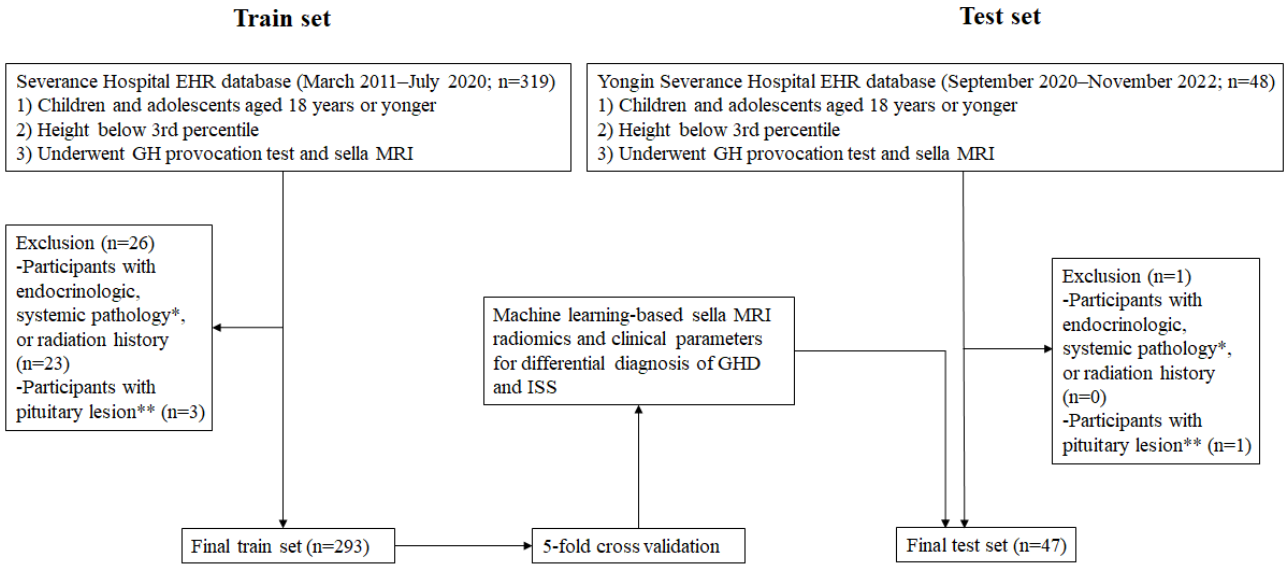

Ethical Considerations

This study conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Institutional Review Board of Yonsei University Severance Hospital (4-2022-1258), which waived the requirement for informed consent due to strict measures were implemented to protect the privacy and confidentiality of the research participants. All personal identifiers were removed from the dataset prior to analysis, ensuring that all data used were anonymized and deidentified. No compensation was provided to participants involved in this study as it involved minimal risk and was primarily based on the analysis of existing medical records and imaging data. Additionally, no identifiable images of participants are included in the study or supplementary materials.

Definition of GHD and ISS

GHD was defined as follows: (1) height below the third percentile for age, sex, and race based on the 2017 Korean National Growth Charts [34]; (2) peak GH level below 10 ng/mL after stimulation in two types of GH provocation tests using insulin, arginine, or L-dopa; and (3) children without genetic, endocrine, or systemic abnormalities [2,35].

ISS was defined as height below the third percentile for individuals of the same age, sex, and race, with no other identifiable causes, including genetic, endocrine, or systemic pathologies [2,34,35].

Clinical Parameters

Height was recorded with an accuracy of 0.1 cm, whereas body weight was measured using an electronic load with a precision of 0.01 kg. BMI was calculated by dividing body weight in kilograms by the square of height in meters (kg/m^2). Height, weight, and BMI were expressed as SDS using the 2017 Korean National Growth Charts [34]. Children were categorized based on their BMI into 3 groups: normal (<85th percentile), overweight (85th-95th percentile), or obese (≥95th percentile). Midparental height (MPH) was determined by calculating the average height of the parents and adjusting it by subtracting 6.5 cm for girls and adding 6.5 cm for boys. Puberty was considered at any pubertal development with Tanner stage ≥2 [36,37].

The detailed method of laboratory evaluation is provided in Multimedia Appendix 1.

SDS values of IGF-Ⅰ and IGF binding protein 3 (IGFBP-3) were calculated based on reference data for the Korean population [38]. Bone age was assessed according to the Greulich-Pyle method by experienced pediatric endocrinologists [39]. In addition, we calculated chronological age–bone age (CA-BA).

Image Acquisition

The detailed image acquisition parameters from both the training and test sets are provided in Multimedia Appendix 2.

Image Processing and Radiomic Feature Extraction

The T2WI and T1C from the sella MRI were examined, and the entire pituitary gland was identified within the region of interest. The outermost boundary of the sliced pituitary gland was outlined.

Following the conversion of the T2WI and T1C from the sella MRI, which were in Digital Imaging and Communication in Medicine format, into NIfTI files, the images were resampled to a resolution of 1×1×1 mm. Additionally, a correction for low-frequency intensity nonuniformity was applied using N4 bias correction [40]. The images were performed by a radiologist (BS) with 10 years of experience, who was unaware of the participants' clinical information. An open-source software (Medical Image Processing, Analysis, and Visualization; Center for Information Technology, National Institutes of Health) was used for the analysis. Segmentation of the pituitary gland in each image slice was performed semiautomatically using

XSL•FO
RenderX

techniques such as region growing, signal intensity thresholding, and edge detection. To ensure the reliability of the segmentation, another radiologist (CJP) with 10 years of experience independently conducted the segmentation of 10% of the final images selected from the dataset, which were chosen randomly. The Dice coefficient was calculated to assess the agreement between the segmentation masks generated by the two radiologists. Next, the radiomic features were extracted using Pyradiomics 2.1.0 with 128 fixed bin counts [41,42]. In total, 14 shapes, 18 first-order, 24 gray-level co-occurrence matrix (GLCM), 16 gray-level run length matrix (GLRLM), 16 gray-level size zone matrix (GLSZM), and 5 neighborhood gray tone difference matrix were extracted from the region of interests on T2WI and T1C, constituting a total of 186 radiomic features.

Machine Learning and Statistical Analysis

Figure 2 shows the machine learning pipeline. We trained and compared 3 models that classified GHD and ISS according to the following parameters: radiomic features, clinical parameters, and both of these parameters. This comprehensive approach aimed to assess the combined predictive ability of radiomics and clinical parameters for diagnosis [43,44]. The XGBoost algorithm was used to train the models. XGBoost is an ensemble of decision trees with high predictive and explanatory ability [45]. In particular, XGBoost can learn datasets with missing values. The XGBoost hyperparameters were optimized using Bayesian optimization with the Gaussian process. Internal validation was conducted using repeated 5-fold cross-validation. The stability of the model increased by repeating the cross-validation multiple times. Tuned hyperparameters and corresponding candidates are described in Multimedia Appendix 3. Through Bayesian optimization, we found the best hyperparameter sets for XGBoost models with clinical parameters, radiomics features, and all features, as represented in Multimedia Appendix 4. The evaluation metrics used were accuracy, sensitivity, specificity, precision, and area under the ROC curve (AUC). The bootstrap method was used for pairwise comparison of the AUC, and the prediction models were externally validated using the Yongin Severance Hospital dataset. All analyses were performed using Python (version 3.9; Python Software Foundation). Significance was determined as $P<.05$.

XSL•FO
RenderX

Figure 2. Machine learning pipeline. ROC: receiver operating characteristics; T1C: contrast-enhanced T1-weighted image; T2WI: T2-weighted image; SHAP: Shapley additive explanations; XGBoost: extreme gradient boosting.

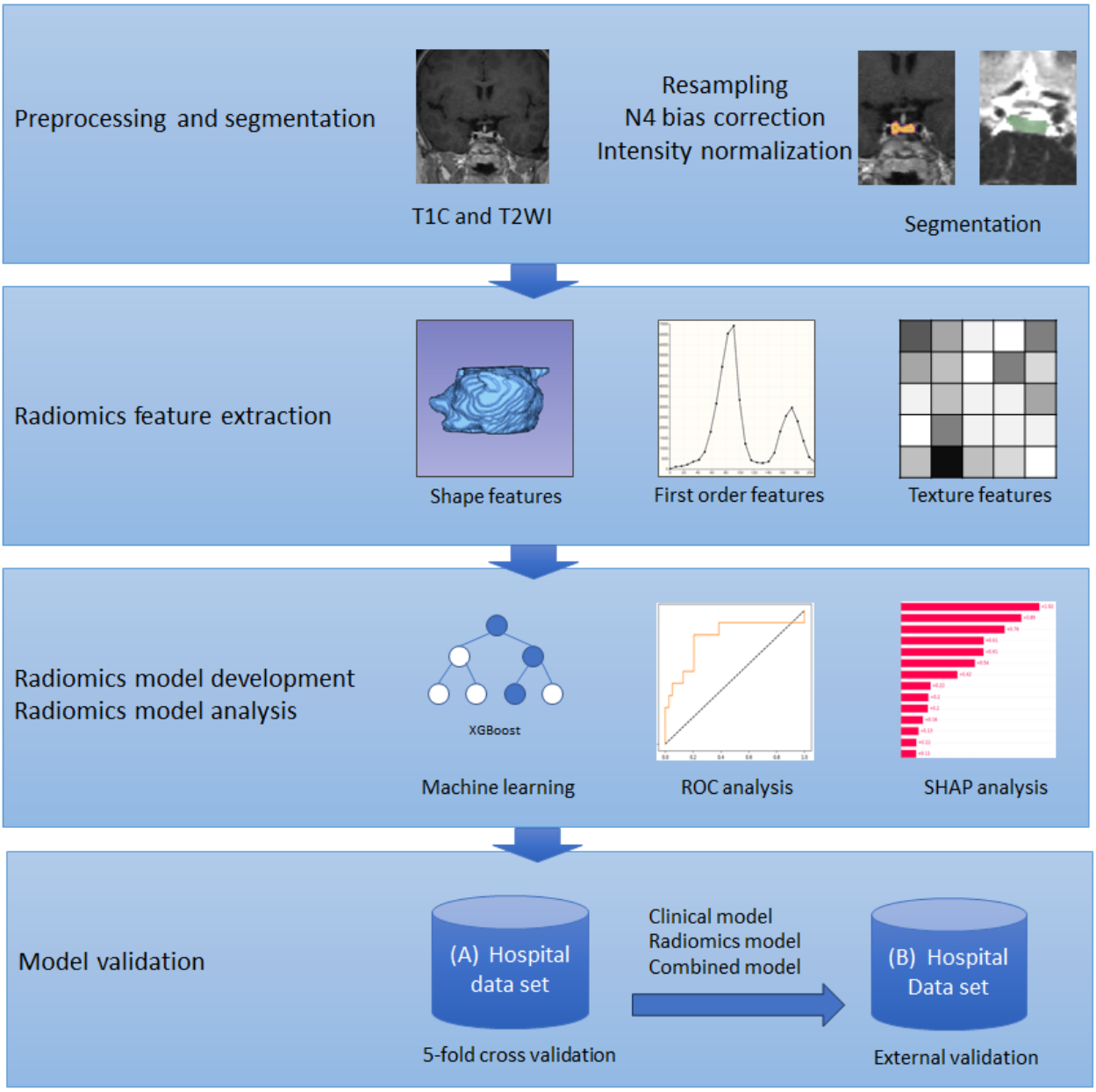

Model Interpretability with Shapley Additive Explanations

Shapley additive explanations (SHAP) was used to interpret and evaluate the significance of each clinical parameter and radiomic feature from the prediction models [46]. SHAP measured the contribution of each feature, called the Shapley value, to the prediction of GHD. This analysis allowed us to visualize and understand the significance of each feature in contributing to the performance of the model. This study used three perspectives to interpret the models: feature importance plots, dot summary plots, and waterfall plots. Importance was calculated by averaging the Shapley values per feature. The dot summary plot is a scatter plot of the feature importance based on the magnitude of each feature value. The waterfall plot shows the impact of the features on the machine-learning models for each case. This study sampled true-positive and true-negative cases for GHD classification and examined a machine-learning model using waterfall plots.

Results

Baseline Characteristics of the Participants

Table 1 summarizes the baseline characteristics of the study participants according to the etiology of their short stature. BMI and the proportions of underweight, overweight, and obese participants were higher in participants with GHD than in those with ISS. IGF-Ⅰ and IGF-Ⅰ SDS were lower in participants with GHD than in those with ISS, whereas CA-BA was higher in those with GHD.

XSL•FO
RenderX

Table 1. Baseline characteristics of participants according to etiology of short stature[a].

	GHD[b] (n=248)	ISS[c] (n=96)	*P* value[d]
Sex (male), n (%)	150 (60.5)	53 (55)	.37
Age (years), mean (SD)	7.24 (2.81)	7.21 (2.73)	.92
Height (cm), mean (SD)	112.53 (14.88)	111.68 (14.34)	.63
Height SDS[e], mean (SD)	–2.54 (0.56)	–2.67 (0.67)	.08
Weight (kg), mean (SD)	20.94 (7.73)	19.37 (6.17)	.07
Weight SDS, mean (SD)	–2.10 (2.31)	–2.42 (1.02)	.20
BMI (kg/m^2), mean (SD)	16.06 (2.50)	15.16 (1.72)	.001
BMI SDS, mean (SD)	–0.73 (2.53)	–1.05 (1.00)	.22
BMI percentile, n (%)			.02
Underweight	189 (76.2)	69 (72)	
Normal	40 (16.1)	24 (25)	
Overweight	11 (4.4)	2 (2)	
Obesity	8 (3.2)	1 (1)	
Growth velocity (cm/year), mean (SD)	4.44 (1.63)	4.21 (1.56)	.25
Pubertal status, n (%)			.29
Prepuberty	213 (85.9)	78 (81)	
Puberty	35 (14.1)	18 (19)	
MPH[f] SDS, mean (SD)	–0.09 (0.08)	–0.09 (0.09)	.51
MPH SDS—height SDS, mean (SD)	2.63 (0.57)	2.76 (0.67)	.08
IGF-Ⅰ[g] (ng/mL), mean (SD)	137.55 (58.38)	153.58 (70.16)	.03
IGF-Ⅰ SDS, mean (SD)	–0.79 (0.63)	–0.69 (0.71)	.02
IGFBP-3[h] (ng/mL), mean (SD)	2344.02 (1127.25)	2159.60 (786.76)	.32
IGFBP-3 SDS, mean (SD)	0.82 (0.83)	0.68 (0.76)	.001
Bone age (years), mean (SD)	6.69 (2.76)	6.72 (2.72)	.94
CA-BA[i] (years), mean (SD)	0.61 (0.95)	0.34 (0.97)	.03

[a]Continuous variables are presented as mean (SD) and categorical variables as numbers (percentages).
[b]GHD: growth hormone deficiency.
[c]ISS: idiopathic short stature.
[d]*P* value was assessed using an independent 2-tailed *t* test for continuous variables and the chi-square test for categorical variables.
[e]SDS: SD score.
[f]MPH: midparental height.
[g]IGF-Ⅰ: insulin-like growth factor-Ⅰ.
[h]IGFBP-3: insulin-like growth factor binding protein-3.
[i]CA-BA: chronological age–bone age.

Regarding the baseline characteristics of participants in the training and test sets, the proportions of boys, underweight, prepuberty, and ISS were higher in the training set than in the test set (Multimedia Appendix 5). Age, height, MPH SDS, and BA were higher in the test set than in the training set, whereas the MPH SDS—height, SDS, and CA-BA were higher in the training set.

Among the training set, MPH, CA-BA, and proportion of the participants with underweight were higher in the GHD group compared to the ISS group (Multimedia Appendix 6). Among the test set, MPH and IGFBP-3 were lower in the GHD group compared to those in the ISS group.

ROC Curve Analyses of Clinical, Radiomics, and Combined Models

Table 2 and Figure 3 summarize the results of the ROC curve analyses and present the AUCs with corresponding 95% CIs for GHD prediction using the clinical, radiomics, and combined models. Among the clinical parameters, age, sex, height SDS, weight SDS, BMI SDS, growth velocity, pubertal state, MPH

XSL•FO
RenderX

SDS, MPH SDS—height SDS, IGF-I SDS, and CA-BA were assessed using clinical and combined models. IGFBP-3 was excluded from the parameters as the value was substantially different between the two centers owing to different assays and reagents.

Table 2. AUCs[a] of each model for predicting GHD[b,c].

	Accuracy	Sensitivity	Specificity	Precision	AUC (95% CI)
Clinical model					
Internal validation	0.717	0.738	0.667	0.838	0.690 (0.628-0.753)
External validation	0.702	0.707	0.667	0.936	0.684 (0.590-0.778)
Radiomics model					
Internal validation	0.678	0.691	0.685	0.578	0.674 (0.609-0.738)
External validation	0.698	0.643	0.667	0.831	0.691 (0.620-0.762)
Combined model					
Internal validation	0.817	0.857	0.722	0.878	0.835 (0.776-0.896)
External validation	0.813	0.810	0.833	0.971	0.830 (0.741-0.919)

[a]AUC: area under the receiver operating characteristics curve.

[b]GHD: growth hormone deficiency.

[c]*P* value was determined using the receiver operating characteristics curve for AUC.

Figure 3. ROC curves from the clinical model, radiomics model, and combined model. (A) ROC curves of the clinical, radiomics, and combined models for internal validation. (B) ROC curves of the clinical, radiomic, and combined models for external validation. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristics.

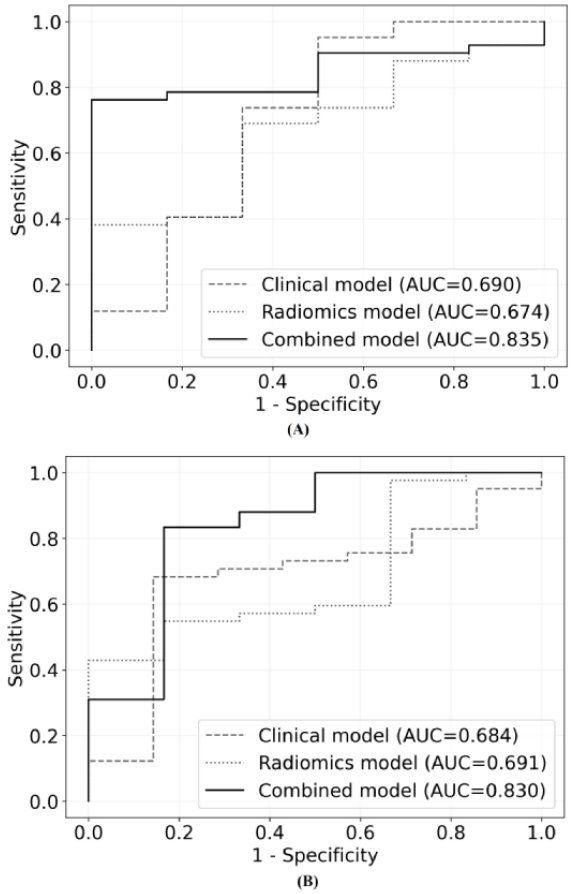

The accuracy and AUC (95% CI) of the clinical model were 0.717 and 0.690 (0.628-0.753) and 0.702 and 0.684 (0.590-0.778) for internal and external validations, respectively. In the radiomics model, the corresponding values were 0.668 and 0.674 (0.609-0.738) for internal validation and 0.698 and 0.691 (0.620-0.762) for external validation. In the combined model, the corresponding values were 0.817 and 0.835

XSL•FO
RenderX

(0.776-0.896) for internal validation and 0.813 and 0.830 (0.741-0.919) for external validation.

In pairwise comparison, the combined model was significantly superior to both the clinical and radiomics models in internal validation (combined model vs clinical model, *P*=.01; combined model vs radiomic model, *P*=.03) and external validation (combined model vs clinical model, *P*=.03; combined model vs radiomic model, *P*=0.02; Multimedia Appendix 7). The AUC was not statistically different between the clinical and radiomics models.

Shapley Value of Clinical Parameters and Radiomics Features

We computed the mean absolute Shapley values for each clinical variable and radiomics feature to illustrate their importance in the predictive models for external validation. Among the clinical parameters, the SHAP value of BMI SDS was the highest, followed by those of CA-BA, weight SDS, growth velocity, IGF-I SDS, MPH SDS, and height SDS (Figure 4A). Among the radiomics features, the SHAP value of inverse variance from T2WI (GLCM) was the highest, followed by energy from T1C (first order) and sum entropy from T2WI (GLCM; Figure 4B). In the combined model, the SHAP value of CA-BA was the highest, followed by weighted SDS, maximum probability from T2WI (GLCM), and run length nonuniformity normalized from T2WI (GLRLM; Figure 4C).

XSL•FO
RenderX

Figure 4. Feature importance according to mean absolute SHAP values in the prediction models for the diagnosis of GHD from external validation. (A) Mean absolute SHAP values in the clinical model. (B) Mean absolute SHAP values in the radiomics model. (C) Mean absolute SHAP values in the combined model. CA-BA: chronological age–bone age; GHD: growth hormone deficiency; IGF-Ⅰ: insulin-like growth factor-I; MPH: midparental height; SDS: SD score; SHAP: Shapley additive explanations.

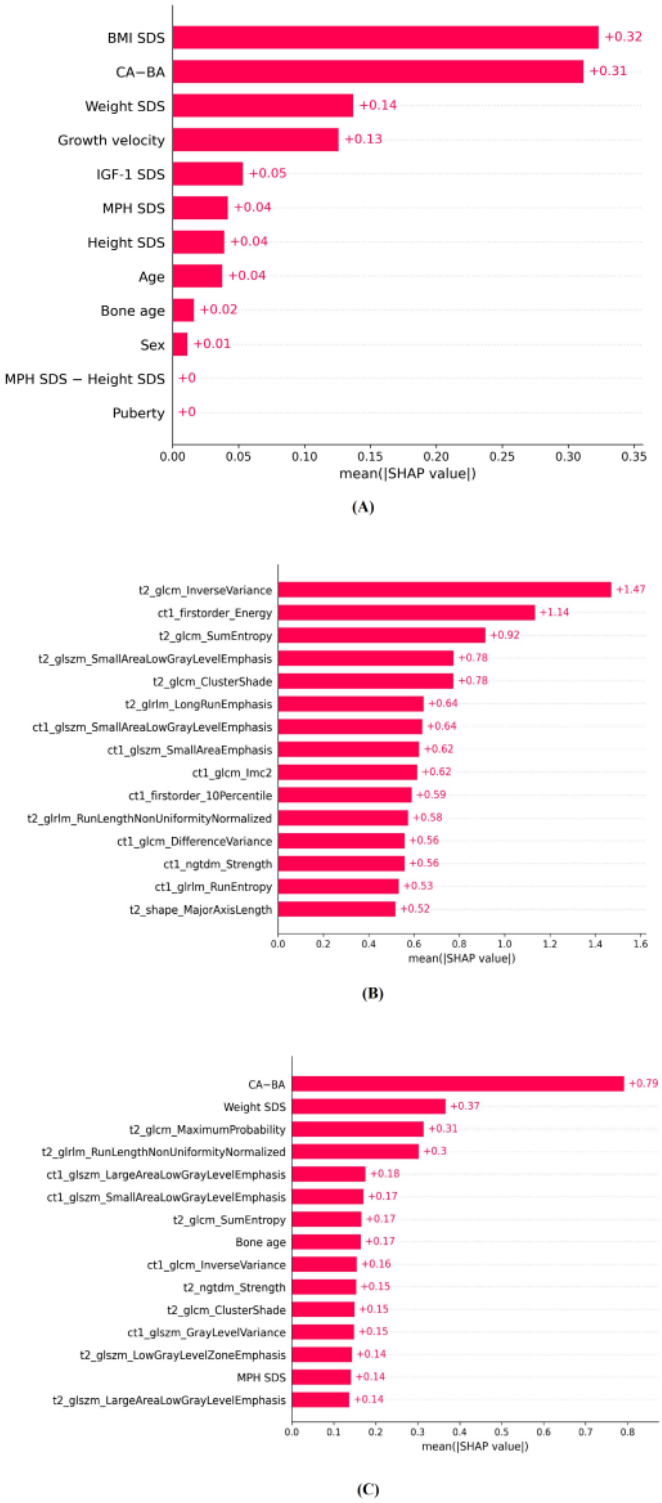

Analysis of the dot summary plots revealed that high CA-BA values and low value of IGF-Ⅰ SDS values influenced the prediction of GHD in the clinical model (Figure S1 in Multimedia Appendix 8). In the radiomics model, the high value of inverse variance from T2WI (GLCM) influenced the prediction of the ISS, and low values of sum entropy from T2WI (GLCM) and small area low gray level emphasis from T2WI (GLSZM) influenced the prediction of GHD (Figure S2 in Multimedia Appendix 8). In the combined model, low values of the CA-BA influenced the prediction of GHD, whereas weight SDS, maximum probability from T2WI (GLCM), and run length nonuniformity normalized from T2WI (GLRLM) contributed highly to the model (Figure S3 in Multimedia Appendix 8).

XSL•FO
RenderX

By conducting SHAP analysis, waterfall plots were generated for each patient, and an example of such a waterfall plot using the clinical model is shown in Figure S1 in Multimedia Appendix 9. The clinical model predicted the participant with ISS as ISS. In this case, the contribution of the CA-BA was the highest, followed by the BMI SDS and IGF-I SDS. Figure S2 in Multimedia Appendix 9 shows a waterfall plot in which the combined model predicts a participant with GHD as having GHD. In this case, the contribution of CA-BA was the highest, followed by T1C GLSZM.

Model Code

The code for our model is available in a public repository on GitHub [47].

Discussion

Principal Findings

In this study, the combined model using both clinical parameters and radiomic features accurately predicted GHD. The combined model was superior to the clinical and radiomics models. Among the clinical parameters, the BMI SDS, CA-BA, weight SDS, and growth velocity were the major contributing factors to the clinical model. Among the radiomics features, inverse variance from T2WI and energy from T1C were the major factors contributing to the radiomics model. In the combined model, CA-BA, weighted SDS, maximum probability from T2WI, and run length nonuniformity normalized from T2WI were the major contributing factors.

Owing to the invasiveness and limitations of the GH provocation test, some studies have investigated the prediction models using clinical parameters for GHD diagnosis. A single-center study from Argentina assessed clinical parameters including pituitary abnormalities, such as pituitary dysgenesis, midline abnormalities, and pituitary hormone deficiencies, in children and developed a GHD prediction model using a decision tree with internal validation only [22]. The sensitivity, specificity, and accuracy of the validation model were 55.6%, 99.2%, and 89.4%, respectively. However, that study focused on children with brain pathology. A study from China developed a predictive model of GHD and ISS using clinical parameters, including IGF-1 and IGFBP-3, and MRI texture [24]. The AUC of the clinical and MRI texture predictive models were 0.607 and 0.852, respectively, although only limited clinical parameters and T1-weighted images were considered and external validation was not performed. We aimed to develop a clinical model for diagnosing GHD in children without pituitary abnormalities, systemic pathology, or endocrinological pathology, excluding GHD and ISS. We assessed various clinical parameters that can be easily obtained in local clinics and developed a machine-learning model with external validation; the results were significant. Therefore, this model can be used to assess the etiology of short stature in real-world clinical settings.

To date, investigations of radiomics models in pediatric endocrinology including the diagnosis of short stature have been limited. A Chinese study attempted to predict central precocious puberty using radiomics in a relatively small number of patients, reporting an AUC of 0.759 [48]. Another technical study focused on the details of computer-aided diagnosis and proved its predictive potential that it can predict GHD; however, the study lacked clinical information [49]. Our previous study analyzed T2-weighted sella MRI images of children with short stature and developed a radiomics-based model to differentiate between GHD and ISS with internal validation, in which the AUC and accuracy were 0.705 and 70.6%, respectively [17]. However, clinical parameters were not considered, and only a single series of MRIs was analyzed in the study without external validation. In this study, the accuracy and AUC of the radiomics model were 0.698 and 0.691, respectively, for external validation. To improve the predictability of radiomics and clinical models, we combined both parameters using a machine learning classifier, XGBoost, to build the prediction models in this study. XGBoost is well-known for handling numerous features for model development with good performance, which is suitable for radiomics studies [50,51]. The pure radiomics model did not yield high predictive performance in external validation; however, the combined clinical and radiomics model could accurately predict GHD with an AUC of 0.830 in external validation. Furthermore, the combined clinical and radiomics model yielded superior predictive performance compared with the clinical model. The added value of radiomics for predicting GHD was validated using an independent test set. Therefore, we believe that radiomics may have a predictive potential for differentiating between GHD and ISS.

To interpret the selected radiomic features and clinical parameters, we performed a SHAP analysis. SHAP analysis enables quantification of the impact of radiomic features and clinical parameters on the prediction of GHD. SHAP estimates the importance and value of each feature in the built model and facilitates informed clinical decision-making. We provided several SHAP plots to visualize the power of each selected feature on global (in the overall study population) and local (one patient) levels. This provides an intuitive visualization of how clinical and radiomic features contribute to the prediction of GHD. In both the radiomics and combined models, we found that the radiomic features extracted from both T1C and T2WI contributed to the prediction. Texture features and first-order features were used in the radiomics model. In the combined model, texture features were used for the prediction. Shape features, including volume, were not used for the prediction, which is consistent with the fact that distinguishing GHD from ISS based on simple pituitary gland volume alone was not successful in previous studies. The maximum probability feature, a GLCM feature, was the most powerful predictor of GHD among the radiomic features, followed by the run length nonuniformity normalized, a GLRLM feature. GLCM measures the spatial distribution of gray-level intensities within an image, which is a biomarker for heterogeneity [16]. Particularly, as the maximum probability represents occurrences of the most predominant pair of neighboring intensity values [41], it may capture the different intensities of the pituitary gland between GHD and ISS, which cannot be detected by visual comparison. The GLRLM quantifies the gray-level runs, which are defined as the length of the number of pixels, of consecutive pixels that have the same gray-level value. A run length nonuniformity normalized, one of GLRLM features, measures the similarity of run lengths throughout the image, with a lower value

XSL•FO
RenderX

indicating more homogeneity among run lengths in the image [41]. As higher values of run length nonuniformity normalized showed a significant association with GHD in this study, we can infer that more heterogeneous pituitary glands can be observed in GHD than in ISS.

Among the clinical parameters, BMI SDS, CA-BA, weight SDS, growth velocity, IGF-I SDS, MPH SDS, and height SDS were the major contributing factors to the prediction model in this study. This result is consistent with those of previous studies. The clinical parameters related to the diagnosis of GHD have been investigated in several studies. A retrospective study reported that height velocity and IGF-1 could be used for screening GHD [52]. A cohort study reported that BMI was negatively related to peak GH level on the GH provocation test [21]. In addition, pubertal maturation is delayed in children with GHD, which is associated with delayed bone age [53,54]. In a cohort study, bone age delay was higher in children with GHD than in those with ISS [35]. In a cohort study, MPH was different according to the etiologies of short stature [2]. Summary Statement of the Growth Hormone Research Society recommends considering height SDS and height velocity when deciding whether or not to perform a GH provocation test [1].

Limitations

This study had some limitations. First, this was a retrospective study limited to a single ethnicity. Second, we could not consider IGFBP-3 since the values from both centers were significantly different owing to the different methods and reagents used. Third, a genetic evaluation was not performed. Fourth, the hypothalamus was not included in this analysis since sella MRI focuses on pituitary glands. As the MRI protocol centers the field of view on the sella or suprasellar area, T2WI often fails to include the entire hypothalamus. In addition, the pituitary gland has relatively clear anatomical boundaries, making segmentation an easy task. However, the hypothalamus lacks clear anatomical boundaries, leading to difficulties in setting the region of interest. Consequently, the segmentation process itself is likely to be biased. MRI is still burdensome for children although it is less burdensome than the GH provocation test, which requires multiple sampling and hospitalization. As sella MRI is performed for patients who have endocrinological problems, further studies investigating radiomics using various protocols of brain MRI are required for the incrementing practical value of radiomics for the prediction of GHD and ISS.

Conclusions

In conclusion, our research strongly emphasizes the potential of combining radiomics-based diagnostic models with clinical parameters for the differentiation between GHD and ISS in children. This study meticulously analyzed both T2WI and T1C in sella MRI, alongside a comprehensive range of clinical parameters, such as puberty status and bone age, and scrutinized the individual contributions of these parameters to the predictive model. Our model combining both radiomics and clinical parameters can accurately predict GHD from ISS, which was also proven in the external validation, thereby proving its predictive potential. Subsequently, we may expect an individualized treatment strategy with our radiomics model combined with machine learning. The code for our model can be assessed in a public repository on GitHub [47]. Further studies with larger samples, including various ethnicities and various brain MRI series, are required to overcome the limitations of this study. In addition, we hope to develop a robust model using genetic information, as well as radiomics and clinical parameters, to replace the GH provocation test in the future.

Acknowledgments

This work was supported by the Korea Health Technology R&D Projects through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grants HG22C0024 and HI23C0040). This work was also supported by a grant from the Central Medical Service Co, Ltd. Research Fund. This work was also supported by the Basic Medical Science Facilitation Program through the Catholic Medical Center of the Catholic University of Korea, funded by the Catholic Education Foundation. No generative artificial intelligence was used in any portion of the manuscript writing.

Data Availability

The datasets generated during or analyzed during this study are available in Zenodo [55] with the permission of the corresponding authors. The code for our model is available in a public repository on GitHub [47].

Authors' Contributions

KS and TK have contributed equally to this work and share first authorship. CJP and BS have contributed equally to this work and share corresponding authorship. KS, TK, CJP, and BS conceptualized and designed this study, undertook the analyses for this study, drafted the initial manuscript, and reviewed and revised the manuscript; HWC provided input in the study design and critically reviewed the manuscript for important intellectual content; JSO, HSK, and JHK contributed to data acquisition; HJS and JHN critically reviewed the manuscript for important intellectual content. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Laboratory evaluation.
[DOCX File , 18 KB - jmir_v26i1e54641_app1.docx]

Multimedia Appendix 2
Image acquisition parameters from both training and test sets.
[DOCX File , 18 KB - jmir_v26i1e54641_app2.docx]

Multimedia Appendix 3
Hyperparameters and corresponding candidates used for tuning XGBoost. XGBoost: extreme gradient boosting.
[DOCX File , 19 KB - jmir_v26i1e54641_app3.docx]

Multimedia Appendix 4
Best hyperparameter set for each XGBoost model. XGBoost: extreme gradient boosting.
[DOCX File , 17 KB - jmir_v26i1e54641_app4.docx]

Multimedia Appendix 5
Baseline characteristics of training set and test set.
[DOCX File , 23 KB - jmir_v26i1e54641_app5.docx]

Multimedia Appendix 6
Baseline characteristics of training set and test set according to etiology of short stature.
[DOCX File , 26 KB - jmir_v26i1e54641_app6.docx]

Multimedia Appendix 7
Comparison of AUCs of the predictive models. AUC: area under the receiver operating characteristic curve.
[DOCX File , 19 KB - jmir_v26i1e54641_app7.docx]

Multimedia Appendix 8
Feature importance presented by mean absolute SHAP value dot summary plot using the prediction models in external validation.
[DOCX File , 701 KB - jmir_v26i1e54641_app8.docx]

Multimedia Appendix 9
Representative waterfall plot of the prediction model cases.
[DOCX File , 399 KB - jmir_v26i1e54641_app9.docx]

References

1. Growth Hormone Research Society. Consensus guidelines for the diagnosis and treatment of growth hormone (GH) deficiency in childhood and adolescence: summary statement of the GH Research Society. J Clin Endocrinol Metab 2000 Nov;85(11):3990-3993. [doi: 10.1210/jcem.85.11.6984] [Medline: 11095419]
2. Song KC, Jin SL, Kwon AR, Chae HW, Ahn JM, Kim DH, et al. Etiologies and characteristics of children with chief complaint of short stature. Ann Pediatr Endocrinol Metab 2015 Mar;20(1):34-39 [FREE Full text] [doi: 10.6065/apem.2015.20.1.34] [Medline: 25883925]
3. Song K, Lee J, Lee S, Jeon S, Lee HS, Kim H, et al. Height and subjective body image are associated with suicide ideation among Korean adolescents. Front Psychiatry 2023;14:1172940 [FREE Full text] [doi: 10.3389/fpsyt.2023.1172940] [Medline: 37377472]
4. Voss LD. Short normal stature and psychosocial disadvantage: a critical review of the evidence. J Pediatr Endocrinol Metab 2001;14(6):701-711. [doi: 10.1515/jpem.2001.14.6.701] [Medline: 11453518]
5. Park SH, Lee YJ, Cheon J, Shin CH, Jung HW, Lee YA. The effect of hypothalamic involvement and growth hormone treatment on cardiovascular risk factors during the transition period in patients with childhood-onset craniopharyngioma. Ann Pediatr Endocrinol Metab 2023;28(2):107-115 [FREE Full text] [doi: 10.6065/apem.2244046.023] [Medline: 35798299]
6. Hanew K, Utsumi A. The role of endogenous GHRH in arginine-, insulin-, clonidine- and l-dopa-induced GH release in normal subjects. Eur J Endocrinol 2002;146(2):197-202. [doi: 10.1530/eje.0.1460197] [Medline: 11834428]
7. Martha PM, Gorman KM, Blizzard RM, Rogol AD, Veldhuis JD. Endogenous growth hormone secretion and clearance rates in normal boys, as determined by deconvolution analysis: relationship to age, pubertal status, and body mass. J Clin Endocrinol Metab 1992;74(2):336-344. [doi: 10.1210/jcem.74.2.1730812] [Medline: 1730812]

XSL•FO
RenderX

8. Bidlingmaier M. Problems with GH assays and strategies toward standardization. Eur J Endocrinol 2008;159 Suppl 1:S41-S44. [doi: 10.1530/EJE-08-0284] [Medline: 18819944]
9. Kessler M, Tenner M, Frey M, Noto R. Pituitary volume in children with growth hormone deficiency, idiopathic short stature and controls. J Pediatr Endocrinol Metab 2016;29(10):1195-1200. [doi: 10.1515/jpem-2015-0404] [Medline: 27710916]
10. Jiménez ABA, Ollero MJMA, Siguero JPL. Differences between patients with isolated GH deficiency based on findings in brain magnetic resonance imaging. Endocrinol Diabetes Nutr (Engl Ed) 2020;67(2):78-88. [doi: 10.1016/j.endinu.2019.09.001] [Medline: 31734177]
11. Nagel BHP, Palmbach M, Petersen D, Ranke MB. Magnetic resonance images of 91 children with different causes of short stature: pituitary size reflects growth hormone secretion. Eur J Pediatr 1997;156(10):758-763. [doi: 10.1007/s004310050707] [Medline: 9365063]
12. Oh JS, Sohn B, Choi Y, Song K, Suh J, Kwon A, et al. The influence of pituitary volume on the growth response in growth hormone-treated children with growth hormone deficiency or idiopathic short stature. Ann Pediatr Endocrinol Metab 2024;29(2):95-101 [FREE Full text] [doi: 10.6065/apem.2346052.026] [Medline: 37946439]
13. Wagner MW, Bilbily A, Beheshti M, Shammas A, Vali R. Artificial intelligence and radiomics in pediatric molecular imaging. Methods 2021;188:37-43. [doi: 10.1016/j.ymeth.2020.06.008] [Medline: 32544594]
14. Madhogarhia R, Haldar D, Bagheri S, Familiar A, Anderson H, Arif S, et al. Radiomics and radiogenomics in pediatric neuro-oncology: a review. Neurooncol Adv 2022;4(1):vdac083 [FREE Full text] [doi: 10.1093/noajnl/vdac083] [Medline: 35795472]
15. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563-577 [FREE Full text] [doi: 10.1148/radiol.2015151169] [Medline: 26579733]
16. Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev 2016;1(2):207-226 [FREE Full text] [doi: 10.1080/23808993.2016.1164013] [Medline: 28042608]
17. Lee T, Song K, Sohn B, Eom J, Ahn SS, Kim H, et al. A radiomics-based model with the potential to differentiate growth hormone deficiency and idiopathic short stature on sella MRI. Yonsei Med J 2022;63(9):856-863. [doi: 10.3349/ymj.2022.63.9.856] [Medline: 36031786]
18. Rui W, Wu Y, Ma Z, Wang Y, Wang Y, Xu X, et al. MR textural analysis on contrast enhanced 3D-SPACE images in assessment of consistency of pituitary macroadenoma. Eur J Radiol 2019;110:219-224. [doi: 10.1016/j.ejrad.2018.12.002] [Medline: 30599863]
19. Zhang Y, Chen C, Tian Z, Cheng Y, Xu J. Differentiation of pituitary adenoma from rathke cleft cyst: combining MR image features with texture features. Contrast Media Mol Imaging 2019;2019:6584636 [FREE Full text] [doi: 10.1155/2019/6584636] [Medline: 31741657]
20. Stanley T. Diagnosis of growth hormone deficiency in childhood. Curr Opin Endocrinol Diabetes Obes 2012;19(1):47-52 [FREE Full text] [doi: 10.1097/MED.0b013e32834ec952] [Medline: 22157400]
21. Yang A, Cho SY, Kwak MJ, Kim SJ, Park SW, Jin D, et al. Impact of BMI on peak growth hormone responses to provocative tests and therapeutic outcome in children with growth hormone deficiency. Sci Rep 2019;9(1):16181 [FREE Full text] [doi: 10.1038/s41598-019-52644-1] [Medline: 31700044]
22. Clément F, Grinspon RP, Yankelevich D, Benítez SM, de la Ossa Salgado MC, Ropelato MG, et al. Development and validation of a prediction rule for growth hormone deficiency without need for pharmacological stimulation tests in children with risk factors. Front Endocrinol (Lausanne) 2020;11:624684 [FREE Full text] [doi: 10.3389/fendo.2020.624684] [Medline: 33613456]
23. Song K, Jung MK, Oh JS, Kim SJ, Choi HS, Lee M, et al. Comparison of growth response and adverse reaction according to growth hormone dosing strategy for children with short stature: LG growth study. Growth Hormone IGF Res 2023;69-70:101531. [doi: 10.1016/j.ghir.2023.101531] [Medline: 36906505]
24. Cong M, Qiu S, Li R, Sun H, Cong L, Hou Z. Development of a predictive model of growth hormone deficiency and idiopathic short stature in children. Exp Ther Med 2021;21(5):494 [FREE Full text] [doi: 10.3892/etm.2021.9925] [Medline: 33791003]
25. Cao G, Zhang M, Wang Y, Zhang J, Han Y, Xu X, et al. End-to-end automatic pathology localization for Alzheimer's disease diagnosis using structural MRI. Comput Biol Med 2023;163:107110 [FREE Full text] [doi: 10.1016/j.compbiomed.2023.107110] [Medline: 37321102]
26. Huang L, Ye X, Yang M, Pan L, Zheng SH. MNC-Net: Multi-task graph structure learning based on node clustering for early Parkinson's disease diagnosis. Comput Biol Med 2023;152:106308. [doi: 10.1016/j.compbiomed.2022.106308] [Medline: 36462371]
27. Zhang Y, He X, Chan YH, Teng Q, Rajapakse JC. Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans. Comput Biol Med 2023;164:107328. [doi: 10.1016/j.compbiomed.2023.107328] [Medline: 37573721]
28. Choi Y, Choi H, Lee H, Lee S, Lee H. Lightweight skip connections with efficient feature stacking for respiratory sound classification. IEEE Access 2022;10:53027-53042. [doi: 10.1109/access.2022.3174678]

XSL•FO
RenderX

29. Dianat B, La Torraca P, Manfredi A, Cassone G, Vacchi C, Sebastiani M, et al. Classification of pulmonary sounds through deep learning for the diagnosis of interstitial lung diseases secondary to connective tissue diseases. Comput Biol Med 2023;160:106928. [doi: 10.1016/j.compbiomed.2023.106928] [Medline: 37156223]
30. Oh GC, Ko T, Kim J, Lee MH, Choi SW, Bae YS, et al. Estimation of low-density lipoprotein cholesterol levels using machine learning. Int J Cardiol 2022;352:144-149. [doi: 10.1016/j.ijcard.2022.01.029] [Medline: 35065153]
31. Lee H, Yang H, Ryu HG, Jung C, Cho YJ, Yoon SB, et al. Real-time machine learning model to predict in-hospital cardiac arrest using heart rate variability in ICU. NPJ Digital Med 2023;6(1):215 [FREE Full text] [doi: 10.1038/s41746-023-00960-2] [Medline: 37993540]
32. Stocker M, Daunhawer I, van Herk W, El Helou S, Dutta S, Schuerman FABA, et al. Machine learning used to compare the diagnostic accuracy of risk factors, clinical signs and biomarkers and to develop a new prediction model for neonatal early-onset sepsis. Pediatr Infect Dis J 2022;41(3):248-254. [doi: 10.1097/INF.0000000000003344] [Medline: 34508027]
33. Tsai C, Lin CR, Kuo H, Cheng F, Yu H, Hung T, et al. Use of machine learning to differentiate children with Kawasaki disease from other febrile children in a pediatric emergency department. JAMA Network Open 2023;6(4):e237489. [doi: 10.1001/jamanetworkopen.2023.7489] [Medline: 37040115]
34. Kim JH, Yun S, Hwang S, Shim JO, Chae HW, Lee YJ, et al. The 2017 Korean National Growth Charts for children and adolescents: development, improvement, and prospects. Korean J Pediatr 2018;61(5):135-149 [FREE Full text] [doi: 10.3345/kjp.2018.61.5.135] [Medline: 29853938]
35. Yoon JY, Cheon CK, Lee JH, Kwak MJ, Kim H, Kim YJ, et al. Response to growth hormone according to provocation test results in idiopathic short stature and idiopathic growth hormone deficiency. Ann Pediatr Endocrinol Metab 2022;27(1):37-43 [FREE Full text] [doi: 10.6065/apem.2142110.055] [Medline: 35038835]
36. Huynh QTV, Ho BT, Le NQK, Trinh TH, Lam LHT, Nguyen NTK, et al. Pathological brain lesions in girls with central precocious puberty at initial diagnosis in Southern Vietnam. Ann Pediatr Endocrinol Metab 2022;27(2):105-112 [FREE Full text] [doi: 10.6065/apem.2142146.073] [Medline: 35592901]
37. Chotipakornkul N, Onsoi W, Numsriskulrat N, Aroonparkmongkol S, Supornsilchai V, Srilanchakon K. The utilization of basal luteinizing hormone in combination with the basal luteinizing hormone and follicle-stimulating hormone ratio as a diagnostic tool for central precocious puberty in girls. Ann Pediatr Endocrinol Metab 2023;28(2):138-143 [FREE Full text] [doi: 10.6065/apem.2346072.036] [Medline: 37401058]
38. Hyun SE, Lee BC, Suh BK, Chung SC, Ko CW, Kim HS, et al. Reference values for serum levels of insulin-like growth factor-I and insulin-like growth factor binding protein-3 in Korean children and adolescents. Clin Biochem 2012;45(1-2):16-21. [doi: 10.1016/j.clinbiochem.2011.10.003] [Medline: 22032863]
39. Greulich WW, Pyle SI. Radiographic Atlas of Skeletal Development of the Hand and Wrist. 2nd Edition. Redwood City, CA: Stanford University Press; 1959.
40. Avants B, Tustison N, Song G. Advanced normalization tools (ANTS). Insight J 2009:1-35 [FREE Full text]
41. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104-e107 [FREE Full text] [doi: 10.1158/0008-5472.CAN-17-0339] [Medline: 29092951]
42. PY-RADIOMICS. Computational Imaging & Bioinformatics Lab. URL: https://www.radiomics.io/pyradiomics.html [accessed 2024-04-04]
43. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
44. Du N, Song L, Gomez-Rodriguez M, Zha H. Scalable influence estimation in continuous-time diffusion networks. Adv Neural Inf Process Syst 2013;26:3147-3155 [FREE Full text] [Medline: 26752940]
45. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.29397]
46. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:1-10 [FREE Full text]
47. taehoonko / Prediction_GHD_ISS. GitHub. URL: https://github.com/taehoonko/Prediction_GHD_ISS [accessed 2024-11-19]
48. Jiang H, Shu Z, Luo X, Wu M, Wang M, Feng Q, et al. Noninvasive radiomics-based method for evaluating idiopathic central precocious puberty in girls. J Int Med Res 2021;49(2):300060521991023 [FREE Full text] [doi: 10.1177/0300060521991023] [Medline: 33596690]
49. Qiu S, Jin Y, Feng S, Zhou T, Li Y. Dwarfism computer-aided diagnosis algorithm based on multimodal pyradiomics. Inf Fusion 2022;80:137-145. [doi: 10.1016/j.inffus.2021.11.012]
50. Nazari M, Shiri I, Zaidi H. Radiomics-based machine learning model to predict risk of death within 5-years in clear cell renal cell carcinoma patients. Comput Biol Med 2021;129:104135 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.104135] [Medline: 33254045]
51. Wang X, You X, Zhang L, Huang D, Aramini B, Shabaturov L, et al. A radiomics model combined with XGBoost may improve the accuracy of distinguishing between mediastinal cysts and tumors: a multicenter validation analysis. Ann Transl Med 2021;9(23):1737 [FREE Full text] [doi: 10.21037/atm-21-5999] [Medline: 35071431]

XSL•FO
RenderX

52. Lemaire P, Brauner N, Hammer P, Trivin C, Souberbielle J, Brauner R. Improved screening for growth hormone deficiency using logical analysis data. Med Sci Monit 2009;15(1):MT5-M10. [Medline: 19114977]
53. Zadik Z, Chalew S, Zung A, Landau H, Leiberman E, Koren R, et al. Effect of long-term growth hormone therapy on bone age and pubertal maturation in boys with and without classic growth hormone deficiency. J Pediatr 1994;125(2):189-195. [Medline: 8040760]
54. Rhie Y, Yoo J, Choi J, Chae H, Kim JH, Chung S, et al. Long-term safety and effectiveness of growth hormone therapy in Korean children with growth disorders: 5-year results of LG growth study. PLoS One 2019;14(5):e0216927 [FREE Full text] [doi: 10.1371/journal.pone.0216927] [Medline: 31095622]
55. Development and validation of a prediction model using sella magnetic resonance imaging-based radiomics and clinical parameters for diagnosis of growth hormone deficiency and idiopathic short stature: a multicenter, cross-sectional study. Zenodo. URL: https://zenodo.org/records/13875769 [accessed 2024-11-19]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CA-BA: chronological age–bone age
GH: growth hormone
GHD: growth hormone deficiency
GLCM: gray-level co-occurrence matrix
GLRLM: gray-level run length matrix
GLSZM: gray-level size zone matrix
IGF-Ⅰ: insulin-like growth factor-Ⅰ
IGFBP-3: IGF binding protein 3
ISS: idiopathic short stature
MPH: midparental height
MRI: magnetic resonance imaging
ROC: receiver operating characteristic
SDS: SD score
SHAP: Shapley additive explanations
T1C: contrast-enhanced T1-weighted image
T2WI: T2-weighted image
XGBoost: extreme gradient boosting

Edited by A Mavragani; submitted 16.11.23; peer-reviewed by Y Zhang, H Chen, V Vakharia; comments to author 28.02.24; revised version received 15.04.24; accepted 02.10.24; published 27.11.24.

Please cite as:
Song K, Ko T, Chae HW, Oh JS, Kim HS, Shin HJ, Kim JH, Na JH, Park CJ, Sohn B
Development and Validation of a Prediction Model Using Sella Magnetic Resonance Imaging–Based Radiomics and Clinical Parameters for the Diagnosis of Growth Hormone Deficiency and Idiopathic Short Stature: Cross-Sectional, Multicenter Study
J Med Internet Res 2024;26:e54641
URL: https://www.jmir.org/2024/1/e54641
doi:10.2196/54641
PMID:

©Kyungchul Song, Taehoon Ko, Hyun Wook Chae, Jun Suk Oh, Ho-Seong Kim, Hyun Joo Shin, Jeong-Ho Kim, Ji-Hoon Na, Chae Jung Park, Beomseok Sohn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

The Triage and Diagnostic Accuracy of Frontier Large Language Models: Updated Comparison to Physician Performance

Michael Joseph Sorich[1], BPharm, GradDipMedStat, PhD; Arduino Aleksander Mangoni[1,2], MD, PhD; Stephen Bacchi[3], MBBS, PhD; Bradley Douglas Menz[1], BPharm; Ashley Mark Hopkins[1], BPharm, PhD

[1]College of Medicine and Public Health, Flinders University, Adelaide, Australia

[2]Department of Clinical Pharmacology, Southern Adelaide Local Health Network, Adelaide, Australia

[3]Department of Neurology and the Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States

Corresponding Author:
Michael Joseph Sorich, BPharm, GradDipMedStat, PhD
College of Medicine and Public Health
Flinders University
GPO Box 2100
Adelaide, 5001
Australia
Phone: 61 82013217
Email: michael.sorich@flinders.edu.au

(J Med Internet Res 2024;26:e67409) doi:10.2196/67409

KEYWORDS

generative artificial intelligence; large language models; triage; diagnosis; accuracy; physician; ChatGPT; diagnostic; primary care; physicians; prediction; medical care; internet; LLMs; AI

Introduction

The medical capabilities of large language models (LLMs) are progressing rapidly [1-3]. Benchmarking LLMs against human performance with clinically relevant tasks enables tracking current capabilities and progress. The triage (level/urgency of care to seek) and diagnostic accuracy of the GPT-3 model were recently compared with 5000 lay individuals using the internet and 21 practicing primary care physicians [4]. The triage ability of GPT-3 was significantly inferior to that of physicians, having similar accuracy to lay individuals. The diagnostic ability was close to but below that of physicians [4]. It is uncertain whether more recent frontier LLMs are still inferior to physicians on this benchmark.

Methods

Overview

The 48 case vignettes—including both common and severe conditions—validated by Levine and colleagues [4] were evaluated using three LLMs that are typically highly ranked across diverse benchmarks: GPT-4o-2024-05-13 (OpenAI), Claude-3.5-Sonnet (Anthropic), and Gemini-1.5-Pro-001 (Google) via a Python application programming interface. The LLMs were instructed to identify potential diagnoses and provide step-by-step reasoning. Subsequently, they reflected on the reasoning and selected the top three diagnoses in order of likelihood. For triage prediction, the LLM was supplied with the vignette and the three diagnoses it predicted. It was instructed to identify the urgency of the required medical care, including its step-by-step reasoning.

A multi-agent workflow involving collaboration between the three distinct LLMs was also evaluated (Figure 1). Each LLM was provided with its initial analysis (decision plus reasoning) and the analyses of the two other LLMs. Each LLM was instructed to reflect on all analyses and update its proposed diagnoses/triage as appropriate. The consensus decision (majority vote) was identified by an independent frontier LLM (Llama-3.1-405B; Meta) to avoid preferencing the output of a specific LLM.

XSL•FO
RenderX

Figure 1. Large language model (LLM) collaboration: a triage/diagnosis workflow involving initial analysis (the LLM's initial decision and step-by-step reasoning), updated analysis (reflecting on all LLM initial analyses and updating decision if appropriate), and consensus decision (majority vote of the individual LLM's updated decisions).

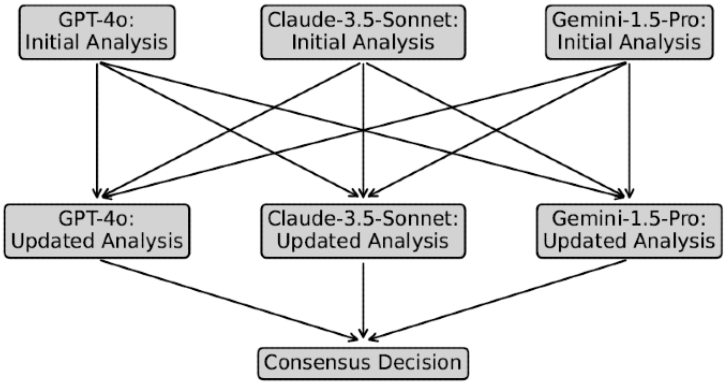

Diagnostic accuracy was evaluated by whether the correct diagnosis was one of the three proposed by the LLM (top 3) [4]. Additionally, the accuracy of the first-ranked diagnosis (top 1) was assessed. Triage was assessed as urgent (emergency department or seeing a doctor within a day) versus nonurgent (seeing a doctor within a week or self-care) [4]. The prompts and LLM settings are provided in Multimedia Appendix 1.

Ethical Considerations

This study involved a secondary analysis of publicly available synthetic case vignettes. No data on human participants were used. The research was undertaken with approval from the Flinders University Human Research Ethics Committee (project ID 7800).

Results

The correct diagnosis was among the top three proposed diagnoses for 98.6% (142/144; frontier LLMs) and 100% (48/48; LLM collaboration) of cases. Individually, the performance of GTP-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro was 98% (47/48), 100% (48/48), and 98% (47/48), respectively.

The most likely diagnosis prediction was correct for 86.8% (125/144; frontier LLMs) and 98% (47/48; LLM collaboration) of cases. Individually, the performance of GTP-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro was 94% (45/48), 96% (46/48), and 71% (34/48), respectively.

Triage was correct for 92.4% (133/144; frontier LLMs) and 92% (44/48; LLM collaboration) of cases. The most common error was overestimating the urgency. Individually, the performance of GTP-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro was 92% (44/48), 94% (45/48), and 92% (44/48), respectively.

Discussion

Contemporary frontier LLMs have substantially improved performance compared to GPT-3 for diagnosis (top three: 142/144, 98.6% vs 42/48, 88%; top one: 125/144, 86.8% vs 31/48, 65%) and triage (133/144, 92.4% vs 34/48, 71%) [4], highlighting the rapid progress in generative artificial intelligence performance. For diagnosis of these clinical vignettes, frontier LLMs performed similarly to physicians (top three: 142/144, 98.6% vs 637/666, 95.6%) [4].

In triaging these clinical vignettes, frontier LLMs (133/144, 92.4%) now perform substantially better than lay individuals (3706/5000, 74.1%) who could use the internet (before the availability of LLMs) and similarly to primary care physicians (608/666, 91.3%) [4]. This capability is consistent with recent evaluations of modern LLMs for emergency department triage [5,6]. A limitation of this study is the relatively small sample size of cases evaluated. Given the encouraging performance of contemporary LLMs for triage assessment, future studies should assess whether LLMs allow lay individuals to make better triage decisions regarding the urgency of care they require.

The rapid progress in LLM capabilities poses challenges for tracking their current capability for health-related tasks. This includes challenges for traditional peer-reviewed publications, which can become outdated by the time of publication.

Additionally, we show that newer techniques involving collaboration between multiple distinct LLMs may improve diagnostic performance. However, this comes at the cost of adding operational complexity. Other methods, such as fine-tuning and in-context learning (eg, integrating search functionality and demonstrations of how to work through complex cases), offer opportunities to improve the performance of LLMs [1,2].

Acknowledgments

MJS is supported by a Beat Cancer Research Fellowship from the Cancer Council South Australia. AMH holds an Emerging Leader Investigator Fellowship from the National Health and Medical Research Council, Australia (APP2008119). The PhD scholarship of BDM is supported by the National Health and Medical Research Council, Australia (APP2030913). The funders

XSL•FO
RenderX

had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

The case vignettes used are publicly available [4]. The prompts used are available in Multimedia Appendix 1.

Conflicts of Interest

MJS reported receiving grants from Pfizer, AstraZeneca, Boehringer Ingelheim, and the National Health and Medical Research Council of Australia outside the submitted work. AMH reported receiving grants from Boehringer Ingelheim, Hospital Research Foundation, Tour De Cure, and Flinders Foundation outside the submitted work. No other disclosures were reported.

Multimedia Appendix 1
Settings and prompts used for large language models.
[DOCX File , 18 KB - jmir_v26i1e67409_app1.docx]

References

1. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv Preprint posted online on November 28, 2023. [doi: 10.48550/arXiv.2311.16452]
2. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of Gemini models in medicine. arXiv Preprint posted online on April 29, 2024. [doi: 10.48550/arXiv.2404.18416]
3. Sorich MJ, Menz BD, Hopkins AM. Quality and safety of artificial intelligence generated health information. BMJ 2024 Mar 20;384:q596. [doi: 10.1136/bmj.q596] [Medline: 38508683]
4. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health 2024 Aug;6(8):e555-e561 [FREE Full text] [doi: 10.1016/S2589-7500(24)00097-9] [Medline: 39059888]
5. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open 2024 May 01;7(5):e248895 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.8895] [Medline: 38713466]
6. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res 2024 Jun 14;26:e53297 [FREE Full text] [doi: 10.2196/53297] [Medline: 38875696]

Abbreviations

LLM: large language model

Edited by A Mavragani; submitted 10.10.24; peer-reviewed by C Williams, Q Shi; comments to author 04.11.24; revised version received 11.11.24; accepted 13.11.24; published 06.12.24.

Please cite as:
Sorich MJ, Mangoni AA, Bacchi S, Menz BD, Hopkins AM
The Triage and Diagnostic Accuracy of Frontier Large Language Models: Updated Comparison to Physician Performance
J Med Internet Res 2024;26:e67409
URL: https://www.jmir.org/2024/1/e67409
doi:10.2196/67409
PMID:

©Michael Joseph Sorich, Arduino Aleksander Mangoni, Stephen Bacchi, Bradley Douglas Menz, Ashley Mark Hopkins. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Early Detection of Dementia in Populations With Type 2 Diabetes: Predictive Analytics Using Machine Learning Approach

Phan Thanh Phuc[1,2], MBA; Phung-Anh Nguyen[3,4,5,6], PhD; Nam Nhat Nguyen[7], MD, PhD; Min-Huei Hsu[3,8], PhD; Nguyen Quoc Khanh Le[9], PhD; Quoc-Viet Tran[3], PhD; Chih-Wei Huang[6,10], PhD; Hsuan-Chia Yang[6,10,11], PhD; Cheng-Yu Chen[12,13], MD, PhD; Thi Anh Hoa Le[2], MHM; Minh Khoi Le[2], MD, PhD; Hoang Bac Nguyen[2], MD, PhD; Christine Y Lu[14,15,16], PhD; Jason C Hsu[1,4,5,6], PhD

[1]College of Management, Taipei Medical University, New Taipei, Taiwan
[2]University Medical Center, University of Medicine and Pharmacy, Ho Chi Minh City, Vietnam
[3]Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
[4]Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan
[5]Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
[6]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan
[7]College of Medicine, Taipei Medical University, Taipei, Taiwan
[8]Office of Data Science, Taipei Medical University, Taipei, Taiwan
[9]Research Center for Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[10]International Center for Health Information Technology, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
[11]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
[12]Department of Radiology, College of Medicine, Taipei Medical University, Taipei, Taiwan
[13]Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan
[14]School of Pharmacy, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[15]Kolling Institute, Faculty of Medicine and Health, The University of Sydney and the Northern Sydney Local Health District, Sydney, Australia
[16]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, United States

Corresponding Author:
Jason C Hsu, PhD
College of Management
Taipei Medical University
11F. Biomedical Technology Building, No. 301, Yuantong Rd., Zhonghe Dist
New Taipei, 235
Taiwan
Phone: 886 2 66202589 ext 16119
Fax: 886 2 66202589
Email: jasonhsu@tmu.edu.tw

Abstract

Background: The possible association between diabetes mellitus and dementia has raised concerns, given the observed coincidental occurrences.

Objective: This study aimed to develop a personalized predictive model, using artificial intelligence, to assess the 5-year and 10-year dementia risk among patients with type 2 diabetes mellitus (T2DM) who are prescribed antidiabetic medications.

Methods: This retrospective multicenter study used data from the Taipei Medical University Clinical Research Database, which comprises electronic medical records from 3 hospitals in Taiwan. This study applied 8 machine learning algorithms to develop prediction models, including logistic regression, linear discriminant analysis, gradient boosting machine, light gradient boosting machine, AdaBoost, random forest, extreme gradient boosting, and artificial neural network (ANN). These models incorporated a range of variables, encompassing patient characteristics, comorbidities, medication usage, laboratory results, and examination data.

Results: This study involved a cohort of 43,068 patients diagnosed with type 2 diabetes mellitus, which accounted for a total of 1,937,692 visits. For model development and validation, 1,300,829 visits were used, while an additional 636,863 visits were

XSL•FO
RenderX

reserved for external testing. The area under the curve of the prediction models range from 0.67 for the logistic regression to 0.98 for the ANNs. Based on the external test results, the model built using the ANN algorithm had the best area under the curve (0.97 for 5-year follow-up period and 0.98 for 10-year follow-up period). Based on the best model (ANN), age, gender, triglyceride, hemoglobin A_{1c}, antidiabetic agents, stroke history, and other long-term medications were the most important predictors.

Conclusions: We have successfully developed a novel, computer-aided, dementia risk prediction model that can facilitate the clinical diagnosis and management of patients prescribed with antidiabetic medications. However, further investigation is required to assess the model's feasibility and external validity.

(J Med Internet Res 2024;26:e52107) doi:10.2196/52107

KEYWORDS

diabetes; dementia; machine learning; prediction model; TMUCRD; Taipei Medical University Clinical Research Database

Introduction

Diabetes mellitus (DM) is a predominant chronic disease worldwide, with type 2 DM (T2DM) accounting for most cases. The prevalence of DM is drastically increasing, with approximately 537 million patients being diagnosed with DM globally in 2021, according to the International Diabetes Federation [1]. The complications of DM are among the leading causes of morbidity and mortality in affected people [2]. Therefore, timely detection and appropriate complication management have recently gained increasing attention [3]. One of the most effective approaches is identifying those at higher risk of developing these complications early using risk prediction models, which could support clinical decision-making [4].

In addition to well-known macro- and microvascular complications, dementia is currently considered a devastating health problem in patients with DM, especially older adults. With the global increase in the aging population, dementia is one of the primary causes of disability and dependency [5-7]. Alzheimer disease is the most common cause among the subtypes of dementia, followed by vascular dementia (also known as cerebrovascular dementia) and Lewy body disease [8]. Concern regarding a potential association between DM and dementia has been raised based on observations of their coincidence. Evidence has illustrated that DM is associated with adverse effects on cognitive function and that patients with DM have a 73% higher risk of all-cause dementia than those without DM [9]. Despite this established relationship, identifying individuals more likely to develop dementia is challenging. The dementia phenotype seen in DM (ie, diabetes dementia) is often due to a combination of vascular and degenerative etiologies [10]. Dementia, thus, may progress slowly and be characterized by a long asymptomatic phase before diagnosis [11]. Consequently, there is a compelling need to optimize risk stratification.

Artificial intelligence (AI) is a robust developing area. AI can improve medicine by better leveraging big data for algorithm development [3]. It has been suggested that AI may restructure how diseases are diagnosed and managed [12,13]. Machine learning (ML) standards have been exploited to construct algorithms to develop predictive models for assessing the risk of DM or its subsequent complications [14,15]. ML learns all composite and nonlinear interactions between features by diminishing inaccuracies between predicted and observed targets [16]. ML is a complementary method that can serve as a benchmark for prediction modeling that may tackle existing drawbacks [14,17]. ML is anticipated to outperform traditional analytical methods due to its ability to identify intricate patterns and relationships within complex datasets. Its predictive skills and flexibility enable it to excel in diverse domains, such as health care [18-23].

Few studies have used ML to assess and construct prediction models for determining the risk of dementia in patients with T2DM. To address these knowledge gaps, this study aims to establish computer-aided risk prediction models for incident dementia within the 5- and 10-year follow-up period among patients with T2DM having complications who are on antidiabetic medications.

Methods

Data Source

This investigation was a retrospective observational cohort study. We used data between January 1, 2008, and December 31, 2020, from the Taipei Medical University Clinical Research Database (TMUCRD) [24]. This dataset amalgamates a rich array of information from 3 distinct medical centers, comprising both structured data (such as essential patient particulars, visit records, laboratory assessments, diagnostic outcomes, treatments, surgical history, and medication details) and unstructured data (including physician notes, pathology and radiological reports, as well as discharge notes). The use of such a diverse and extensive dataset holds significant potential for insightful data-driven analyses and exploration within the realm of medical research and informatics [25]. All data were anonymized before the analysis.

Study Population

We identified potential individuals who sought medical care in the outpatient department or were admitted to the inpatient department with a documented diagnosis of DM. The diagnosis was based on specific *International Classification of Disease, Ninth Revision, Clinical Modification* (*ICD-9-CM*) and *International Classification of Disease, Tenth Revision, Clinical Modification* (*ICD-10-CM*) codes, namely 250xx and E11xx, respectively, during the period between January 1, 2009, and December 31, 2019.

XSL•FO
RenderX

One study from Taiwan found that middle-aged individuals (aged 45-65 years) with diabetes had a higher risk of dementia compared with older patients with diabetes (aged >65 years) [26]. Since most dementia occurs in middle-aged and older people aged >40 years rather than the young and middle-aged group, we selected patients with T2DM older than 40 years as the target [26]. To ensure the homogeneity and reliability of the study population, we excluded patients with type 1 DM (T1DM) and individuals with T2DM without associated complications, such as neurological, renal, or cardiovascular diseases. We also excluded patients with no previous visit histories or those already diagnosed with dementia before their DM diagnosis, as well as those who did not receive any specific treatments for DM. As a result of these stringent criteria, our study focused exclusively on the subset of patients who were accurately classified as having DM and had received multiple prescriptions for antidiabetic drugs, classified by the Anatomical Therapeutic Chemical code A10. This selection process is to maximize the validity and relevance of our findings within the context of the research objective [27].

The Rationale for Prescription Sampling in Risk Prediction Modeling

We conducted simulations to replicate clinically relevant scenarios where physicians aim to assess the probability of dementia occurrence within the 5- and 10-year follow-up for specific patients with T2DM. These simulations were based on clinical features collected during the clinic visit or admission, along with the prescription of antidiabetic medications [28].

Outcomes

In this study, we meticulously examined a cohort of patients with T2DM who received antidiabetic medications. Our investigation followed them from the initiation of their first prescription of antidiabetic medication to a new dementia diagnosis. Dementia cases were identified through an extensive analysis of comprehensive data, involving specific diagnostic (*ICD-9-CM* codes 290, 294, and 311.2, as well as *ICD-10-CM* codes F00-F03, F05.1, G30, and G31.1) and antidementia medication codes (Anatomical Therapeutic Chemical codes N06DA and N06DX). The chronological sequence for defining a dementia case was the diagnosis of dementia followed by the prescription of antidementia medication. We tracked patients from the inception of antidiabetic medication until they were prescribed their first antidementia medication, confirming the dementia diagnosis.

Notably, patients who did not undergo any dementia-related treatments were intentionally excluded from the study group. This deliberate exclusion aimed to emphasize our focus on actively managed or treated dementia cases, as opposed to solely those diagnosed. Such an approach significantly bolsters the connection between antidiabetic medication and subsequent dementia management or progression.

In the context of the 5-year dementia incidence, it pertains to individuals in whom a validated diagnosis of dementia was established within a period of 5 years subsequent to the index date. Concerning the 10-year dementia incidence, it signifies the potential for these individuals to have received a confirmed diagnosis of dementia at any point during an extended time span of 10 years following the index date.

Previous studies have identified several risk factors for developing mild cognitive impairment over a median of 5 years, including the presence of diabetes, diabetes duration, and poor diabetes control [29,30]. Other longitudinal studies with follow-ups ranging from 10 to over 20 years have reported that diabetes increases the risk of developing dementia [31]. Consequently, based on the literature, our study established 5-year and 10-year time-at-risk windows.

To ensure data accuracy, follow-up was conducted until December 31, 2020, and data censoring was applied in cases of patients lost to follow-up or mortality. Comprehensive details can be found in Table S1 in Multimedia Appendix 1.

Predictors or Features

We identified various features associated with outcomes from both outpatient and inpatient datasets. These predictors included various factors sourced from outpatient and admission datasets, encompassing diagnoses, medications, and laboratory tests. Specifically, the predictors included:

1. Demographic factors such as gender, age, and BMI
2. Preexisting comorbidities before starting antidiabetic medication, including cardiovascular disease, chronic obstructive pulmonary disease, rheumatic diseases, and the Charlson Comorbidity Index (CCI) score
3. Long-term medication usage in the 6 months before prescribing antidiabetic drugs, such as antacids, medications for gastroesophageal reflux disease, and gastrointestinal disorder agents
4. Laboratory test results from the 12 months preceding the prescription of antidiabetic drugs, including hemoglobin A_{1c} (HbA_{1c}), fasting blood glucose levels, and albumin levels

This thorough approach allowed us to account for a wide range of factors that could potentially influence the outcomes being studied.

Specifically, we established the index date as the prescription of the antidiabetic drug. The collected variables included patient characteristics such as age and gender, preexisting comorbidities, and other long-term medications taken before the index date. In addition, within the 12 months preceding this date, we gathered relevant laboratory test results. This 12-month window was chosen to minimize missing laboratory data, and these results will serve as reference features in our prediction models.

The TMUCRD dataset contained missing data, primarily related to laboratory test data. Specifically, fasting glucose, creatinine, HbA_{1c}, total cholesterol, and triglyceride data were often found to be missing at the first visit. To address this issue, we used the multiple imputation by chained equations method, a statistical technique used to handle missing data in datasets [32].

Multiple imputations by chained equations uses an iterative approach to impute missing values by creating multiple imputed datasets, thereby preserving the uncertainty associated with the imputations. By generating multiple plausible imputations, we

XSL•FO
RenderX

were able to obtain more accurate results and perform a comprehensive analysis while effectively addressing the challenge of missing data in our study. This method ensures that the potential biases and limitations resulting from missing data are appropriately addressed, enhancing the validity and robustness of our findings [32].

Model Development

In this study, various ML algorithms were used for the development and validation of predictive models. The ML algorithms used included logistic regression, linear discriminant analysis, gradient boosting machine (GBM), light GBM (LGBM), AdaBoost, random forest (RF), extreme gradient boosting, and artificial neural network (ANN) [33-39]. The specific details of each ML model, including parameter settings, can be found in "S1. Methods" in Multimedia Appendix 1.

Model Evaluation

To ensure the robustness of the developed models and account for potential sample selection bias [40], the dataset was divided into training and test sets. The training set consisted of data from Taipei Medical University Hospital and Wang Fang Hospital and was used for the learning process. Stratified 10-fold cross-validation was applied within the training set to assess the performance of different ML models and determine generalization errors.

In addition, an external test set, comprising data from Shuang Ho Hospital, was used to validate the models' performances. The models were evaluated based on various metrics, including the area under the curve (AUC), sensitivity (recall), specificity, positive predictive value (or precision), negative predictive value, and F_1-score. These metrics provide a comprehensive assessment of the models' predictive capabilities, aiding in the selection and validation of the most effective ML algorithms for the given dataset [39,41], We identified the best-performing model by comparing different models based on the external testing set, selecting the one with the highest AUC. To gain deeper insights into this best model, we conducted an analysis of the features' impacts using Shapley additive explanations (SHAP) values. SHAP is a critical feature attribution method in the domain of interpretability and explainability for ML models. Its purpose is to facilitate the understanding of predictions made by complex models, including deep learning algorithms, RFs, GBMs, and others. By using SHAP values, we were able to assign importance scores to individual features, effectively revealing their respective contributions to specific predictions.

This comprehensive approach allowed us to unravel the factors influencing the model's performance and provided valuable insights into the significance of each feature in predicting outcomes. Understanding the relative importance of different features is essential in our research, where accurate and interpretable models are pivotal for making informed decisions and improving patient care [42].

Statistical Analysis

Descriptive analysis of the study population, including the frequency (%) and mean (SD) for categorical and numerical variables, respectively, were evaluated. Univariate and multivariate analyses were used to investigate the significant correlations between risk predictors and outcome variables.

We performed statistical analysis using R version 4.1.3 (R Core Team; R Project for Statistical Computing). The ML algorithms (logistic regression, linear discriminant analysis, LGBM, GBM, extreme gradient boosting, AdaBoost, and RF) were generated using *scikit-learn* library (version 1.0.2), and the ANN model was developed with *TensorFlow* (version 2.9.0; Google Brain Team) in Python programing language (version 3.9; Python Software Foundation).

Ethical Considerations

This study was approved by the Joint Institutional Review Board of Taipei Medical University, Taipei, Taiwan (approval N202208033).

The use of data from the TMUCRD for research purposes is exempt from review by the Institutional Review Board in Taiwan because the data used are public and aggregated patient-level information.

Results

Baseline Characteristics of the Participants

In our investigation, we identified a cohort of 177,009 patients who were newly diagnosed with DM between the years 2008 and 2021 in the TMUCRD dataset. To ensure the homogeneity and validity of the study population, we excluded various subgroups. Specifically, we excluded 43,097 patients with T1DM; 29,450 patients with T2DM but without any associated complications; and 46,166 individuals who had not received any antidiabetic medications. Besides, we excluded 2390 patients below the age of 40 years and 3423 participants who were diagnosed with dementia before their DM diagnosis. In addition, 9415 patients with incomplete medical histories or lacking follow-up treatments recorded at hospitals were excluded from the study (Figure 1). As a result of these rigorous selection criteria, we established a final cohort comprising 43,068 patients (corresponding to 1,937,692 visits) with T2DM. This carefully curated cohort was used to develop and validate our predictive models. Specifically, a total of 1,300,829 visits were used for the development and validation of the models, while the external testing set consisted of 636,863 visits. We analyzed a sizable cohort of patients, consisting of 46.4% (12,686/27,360) female patients in the training set and 46.1% (7243/15,708) female patients in the testing set. The mean age and BMI of patients in the training set were 62.9 (SD 14.0) years and 26.1 (SD 4.66) kg/m^2, respectively. For the testing set, the mean age and BMI were 62.1 (SD 13.7) years and 26.0 (SD 4.73) kg/m^2, respectively.

XSL•FO
RenderX

Figure 1. Overview of the cohort population. DM: diabetes mellitus; T1DM: type 1 diabetes mellitus; T2DM: type 2 diabetes mellitus; TMUCRD: Taipei Medical University Clinical Research Database.

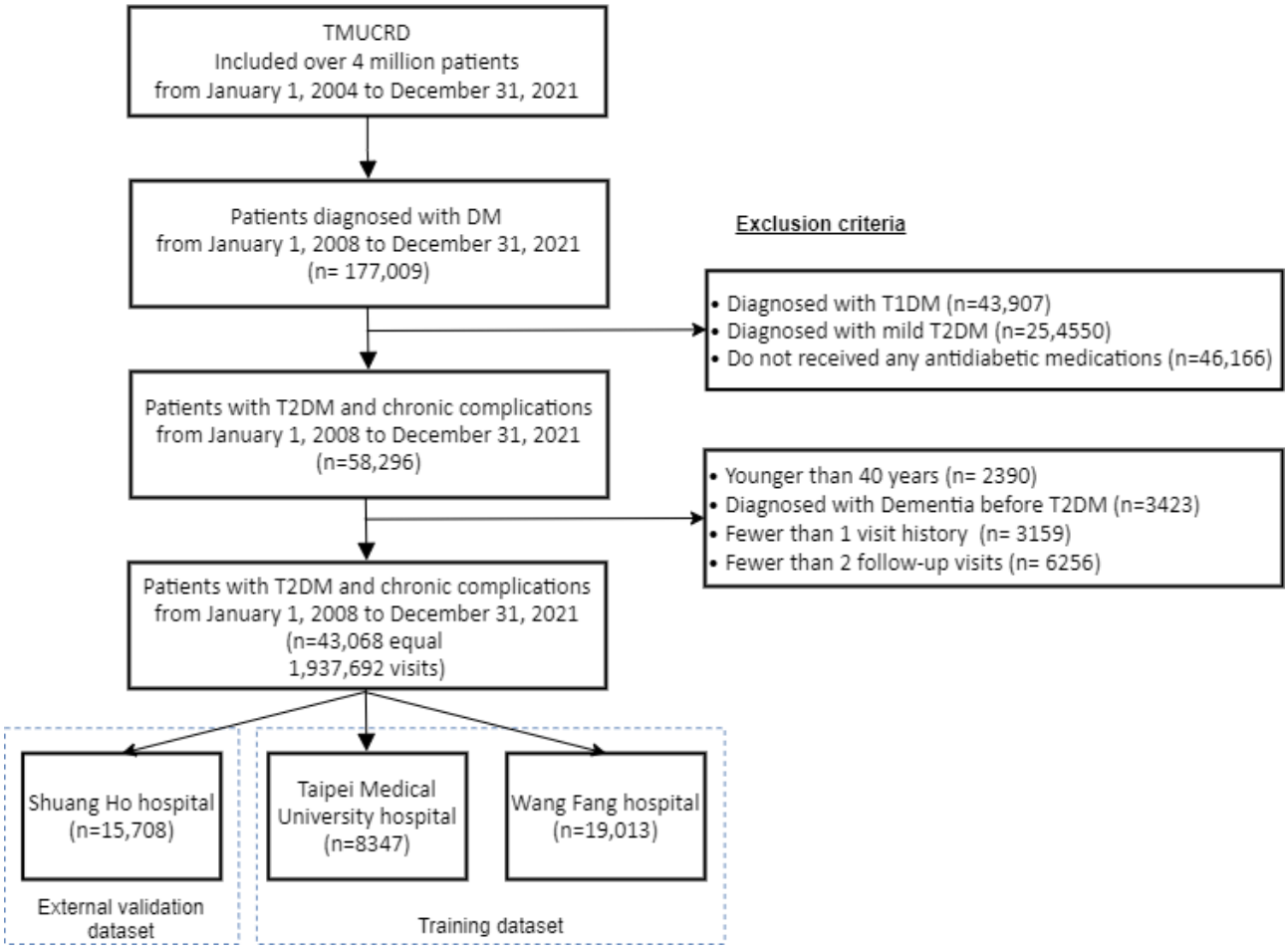

Furthermore, we evaluated the CCI score, which represents the burden of comorbidities in patients. In the training set, the average CCI score was 5.16 (SD 1.58), while in the testing set, it was 4.92 (SD 1.42; refer to Table 1 and Table S2 in Multimedia Appendix 1 for detailed results).

XSL•FO
RenderX

Table 1. Baseline characteristics of the training and testing cohorts.

Characteristic	Training cohort[a] (n=27,360)	Testing cohort[b] (n=15,708)
Age (years)		
Mean (SD)	62.9 (14.0)	62.1 (13.7)
Median (IQR)	62.7 (53.5-73.1)	61.9 (53.2-71.7)
Sex, n (%)		
Male	14,674 (53.6)	8465 (53.9)
Female	12,686 (46.4)	7243 (46.1)
BMI (kg/m^2)		
Mean (SD)	26.1 (4.7)	26.0 (4.7)
Median (IQR)	25.7 (23-29)	25.7 (22.9-29)
Number of visits for antidiabetic drugs, n (%)	1,300,829 (76.1)	636,863 (32.9)
Insulin and analogues	130,938 (10.1)	77,397 (12.2)
Metformin	207,513 (16)	93,685 (14.7)
Sulfonylureas	38,770 (3)	26,220 (4.1)
α-glucosidase inhibitors	9198 (0.7)	3717 (0.6)
Thiazolidinediones	2617 (0.2)	1088 (0.2)
Dipeptidyl peptidase 4 inhibitors	55,064 (4.2)	19,493 (3.1)
Glucagon-like peptide-1 analogues	1195 (0.1)	195 (0.03)
Sodium-glucose cotransporter 2 inhibitors	2800 (0.2)	1066 (0.2)
Other blood glucose–lowering drugs, excluding insulins	13,665 (1.1)	13,089 (2.1)
Combinations of oral blood glucose–lowering drugs	839,069 (64.5)	400,913 (63.0)
5-year dementia, n (%)	193 (0.7)	125 (0.8)
10-year dementia, n (%)	328 (1.2)	185 (1.2)
Comorbidities, n (%)		
Hyperlipidemia	7769 (28.4)	1494 (9.5)
Hypertension	8410 (30.7)	2625 (16.7)
Previous stroke	732 (2.7)	457 (2.9)
Heart problem[c]	2454 (9)	705 (4.5)
Chronic pulmonary disease	805 (2.9)	93 (0.6)
Renal disease	1075 (3.9)	508 (3.2)
Rheumatic disease	67 (0.2)	15 (0.1)
Peptic ulcer disease	832 (3)	131 (0.8)
Any malignancy	332 (1.2)	75 (0.5)
Liver disease	2085 (7.6)	216 (1.4)
Anemias	601 (2.2)	101 (0.6)
Depressive disorder	1075 (3.9)	144 (0.9)
Parkinson	118 (0.4)	50 (0.3)
Charlson Comorbidity Index (CCI)		
Mean (SD)	5.16 (1.58)	4.92 (1.42)
Median (IQR)	5 (4-6)	5 (4-6)
Other medications (ATC[d]), n (%)		
Antacids (A02AA and A02AX)	620 (2.3)	84 (0.5)

XSL•FO
RenderX

Characteristic	Training cohort[a] (n=27,360)	Testing cohort[b] (n=15,708)
Drugs for peptic ulcer and gastroesophageal reflux disease (A02BA and A02BC)	215 (0.8)	91 (0.6)
Gastrointestinal disorders (A03AX and A03FA)	310 (1.1)	54 (0.3)
Liver therapy (A05BA)	207 (0.8)	33 (0.2)
Laxatives (A06AB and A06AD)	869 (3.2)	140 (0.9)
Antithrombotic (B01AA and B01AC)	2888 (10.6)	631 (4)
Antianemic agents (B03BA, B03BB, and B03XA)	592 (2.2)	189 (1.2)
Cardiac therapy (C01AA, C01BD, C01DA, and C01DX)	1051 (3.8)	271 (1.7)
Antihypertensives (C02CA and C02DB)	203 (0.7)	63 (0.4)
Diuretics (C03AA, C03BA, C03CA, and C03DA)	1565 (5.7)	278 (1.8)
Purine derivatives (C04AD)	382 (1.4)	169 (1.1)
β-blocking agents (C07AA, C07AB, and C07AG)	2495 (9.1)	450 (2.9)
Calcium channel blockers (C08CA and C08DB)	2552 (9.3)	378 (2.4)
Renin angiotensin (C09AA, C09CA, C09DB, and C09DX)	3850 (14.1)	518 (3.3)
Lipid modifying agents (C10AA, C10AB, C10AX, and C10BA)	3925 (14.3)	525 (3.3)
α-adrenoreceptor antagonists (G04CA)	325 (1.2)	44 (0.3)
Glucocorticoids (H02AB)	74 (0.3)	17 (0.1)
Thyroid hormones (H03AA)	231 (0.8)	33 (0.2)
Anti-inflammatory and antirheumatic, nonsteroids (M01AB, M01AC, and M01AH)	238 (0.9)	77 (0.5)
Antigout (M04AA, M04AB, and M04AC)	892 (3.3)	109 (0.7)
Nervous system (N02AJ, N02BE, N03AE, N03AX, N04BA, N05AH, N05BA, N05BB, N05CD, N05CF, N06AA, N06AX, N06BX, N07AB, and N07CA)	1956 (7.1)	428 (2.7)
Antihistamines (R06AE and R06AX)	156 (0.6)	42 (0.3)
Laboratory test, mean (SD)		
HbA_{1c}[e], %	8.07 (2)	8.63 (2.3)
Fasting glucose (mg/dL)	163 (88.6)	172 (76.1)
Creatinine (mg/dL)	1.19 (1.2)	1.31 (1.6)
Total cholesterol (mg/dL)	187 (43.6)	199 (49.6)
Triglyceride (mg/dL)	167 (169)	193 (267)

[a]The training cohort consisted of data from Taipei Medical University and Wan Fang Hospital.

[b]The testing cohort consisted of data from Shuang Ho Hospital.

[c]Heart problem included heart failure, myocardial infarction, cerebrovascular disease, and peripheral vascular disease.

[d]ATC: Anatomical Therapeutic Chemical.

[e]HbA_{1c}: hemoglobin A_{1c}.

Regarding the prescribed antidiabetic agents, we observed a predominant use of a combination of oral blood glucose–lowering drugs, accounting for 64.5% (839,069/1,300,829) of visits in the training set and 63% (400,913/636,863) of visits in the testing set. In addition, in the testing set, we noted 14.7% (93,685/636,863) of visits for metformin and 12.2% (77,397/636,863) of visits for insulin and analogs. These findings shed light on the medication patterns and treatment strategies used in the management of diabetes within our study population.

By presenting comprehensive demographic and medication use data, we aim to provide a detailed understanding of the patient characteristics and therapeutic choices pertinent to our medical informatics investigation. These insights are crucial for assessing the generalizability and clinical implications of our research findings. The associations between clinical features and 5-year and 10-year dementia incidence at baseline were shown in Table S3 in Multimedia Appendix 1, respectively.

Prediction Models Performance of 5-Year and 10-Year Dementia Incidence

The performance metrics of diverse prediction models are summarized in Table 2. Specifically, for the 5-year dementia incidence prediction, the ANN model achieved the highest AUC

XSL•FO
RenderX

value of 0.97, with a recall of 0.7, precision of 0.01, and F_1-score of 0.03. Following closely, the GBM and voting models attained AUCs of 0.82 each.

Table 2. Performance of various prediction models.

Model	Training AUC[a]	Testing AUC	Precision	Recall	F_1-score
5-year follow-up					
LR[b]	0.72	0.67	0.01	0.70	0.02
LDA[c]	0.80	0.81	0.02	0.88	0.04
LGBM[d]	0.94	0.82	0.02	0.90	0.04
GBM[e]	0.86	0.82	0.02	0.95	0.05
RF[f]	0.95	0.81	0.02	0.88	0.04
XGBoost[g]	0.98	0.80	0.02	0.86	0.04
AdaBoost	0.82	0.81	0.02	0.85	0.04
Voting	0.86	0.82	0.02	0.91	0.05
ANN[h,i]	0.98	0.97	0.01	0.70	0.03
10-year follow-up					
LR	0.68	0.67	0.01	0.73	0.03
LDA	0.78	0.78	0.02	0.85	0.04
LGBM	0.92	0.80	0.02	0.88	0.05
GBM	0.84	0.80	0.02	0.90	0.05
RF	0.94	0.80	0.02	0.85	0.05
XGBoost	0.97	0.78	0.02	0.83	0.05
AdaBoost	0.81	0.80	0.02	0.84	0.05
Voting	0.84	0.80	0.02	0.89	0.05
ANN[i]	0.98	0.98	0.02	0.79	0.03

[a]AUC: area under the curve.
[b]LR: logistic regression.
[c]LDA: linear discriminant analysis.
[d]LGBM: light gradient boosting machine.
[e]GBM: gradient boosting machine.
[f]RF: random forest.
[g]XGBoost: extreme gradient boosting.
[h]ANN: artificial neural network.
[i]Best model based on area under the curve values.

Similarly, for the 10-year dementia incidence prediction, the ANN model exhibited the most favorable performance with an AUC of 0.98, accompanied by a recall of 0.79, precision of 0.02, and F_1-score of 0.03. In addition, the GBM, LGBM, RF, and voting models achieved AUCs of 0.8.

For a comprehensive analysis of the models' performance, including various measurement metrics, please refer to Table S4 in Multimedia Appendix 1. Furthermore, Figure 2 visually represents the receiver operating characteristic curves for the different prediction models at both 5-year and 10-year dementia incidences. The loss function of curves of the ANN model are showed in Figure S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. The performance of the prediction models in the testing dataset. (A) 5-year follow-up of dementia incidence using machine learning models, (B) 10-year follow-up of dementia incidence using machine learning models, (C) 5-year follow-up of dementia incidence using the ANN model, and (D) 10-year follow-up of dementia incidence using the ANN model. ANN: artificial neural network; AUC: area under the curve; LGBM: light gradient boosting machine; XGB: extreme gradient boosting.

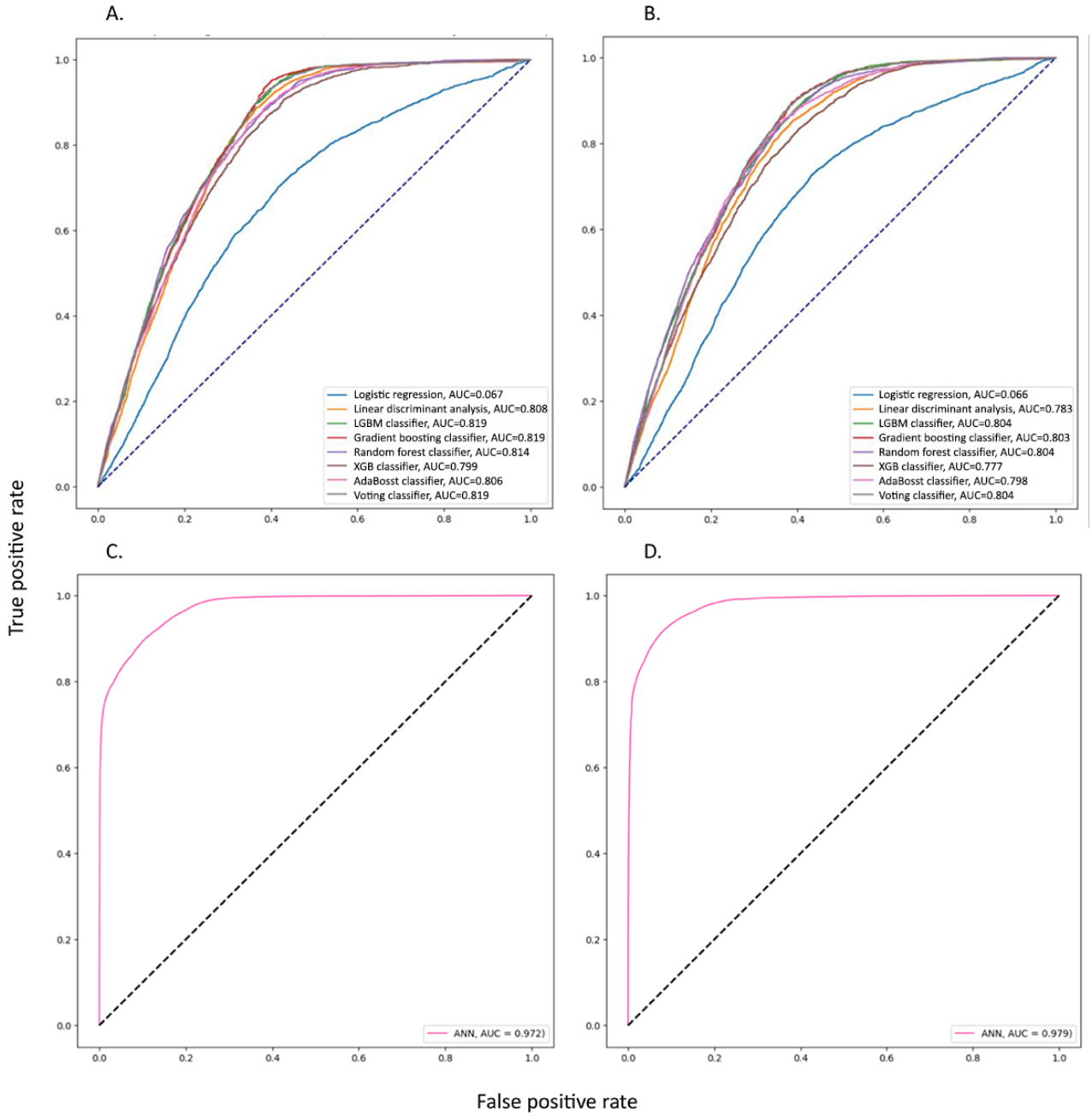

Feature Importance

In Figure 3, we present the top 20 crucial features that significantly influenced the performance of prediction models for both 5-year and 10-year dementia incidence. For the 5-year follow-up model, essential features included age, renin-angiotensin system medications, hyperlipidemia, triglyceride levels, CCI score, gender, antithrombotic drugs, stroke history, hypertension, and calcium-channel blockers.

XSL•FO
RenderX

Figure 3. Feature importance of the artificial neural network prediction model. (A) 5-year follow-up of dementia incidence, (B) 10-year follow-up of dementia incidence. DM: diabetes mellitus; Dx: disease; Lab: laboratory exam; Rx: medication; SHAP: Shapley additive explanations.

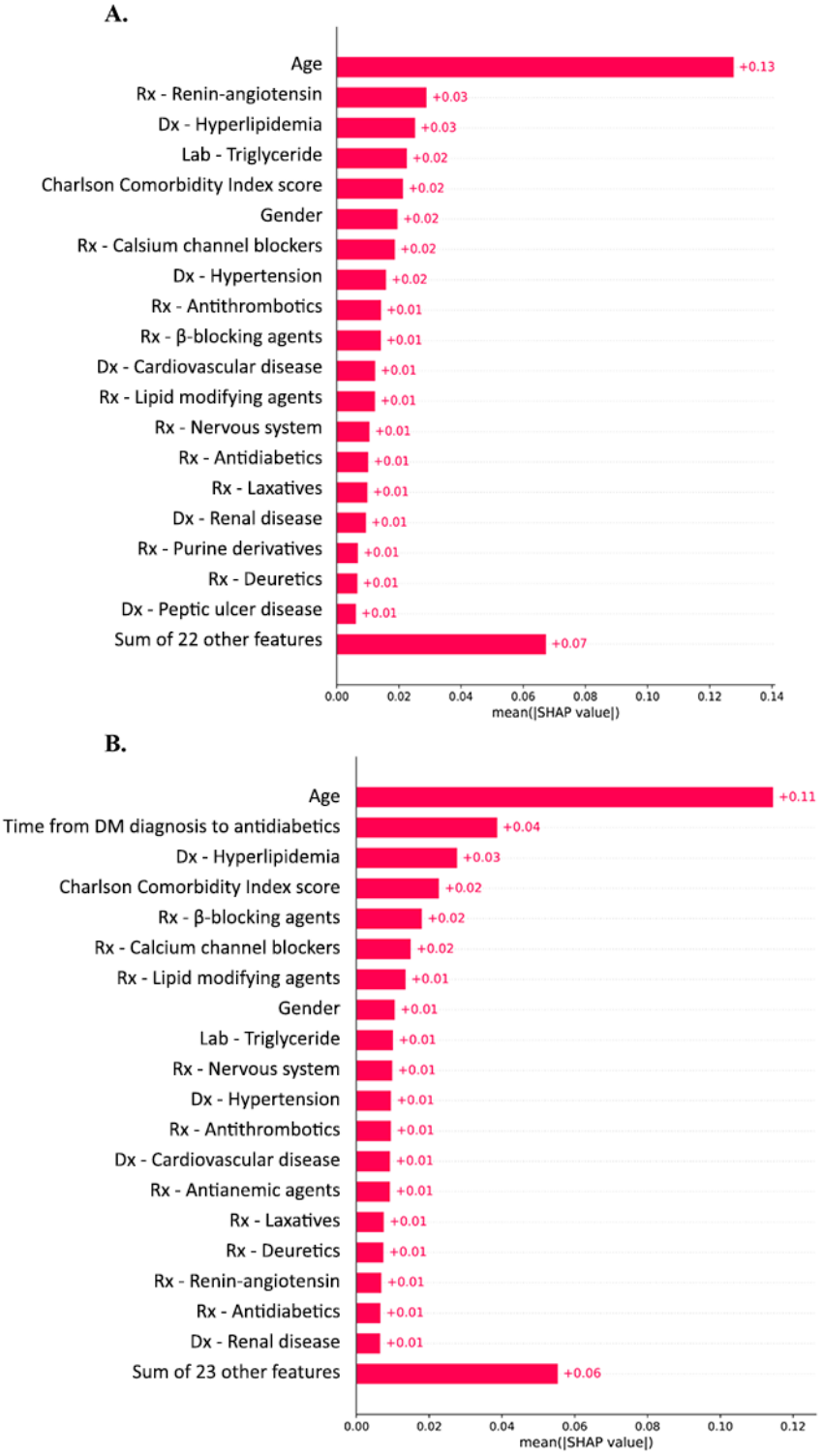

Conversely, the top features identified for the 10-year follow-up model included age; duration of diabetes; hyperlipidemia; CCI score; and the usage of long-term medications such as calcium β-blockers, channel blockers, and lipid-modifying drugs. These analyses shed light on the critical variables that significantly impacted the capabilities of our models in predicting 5-year and 10-year dementia incidence.

By identifying and analyzing these important features, we enhance our understanding of the underlying factors associated with dementia risk, enabling more accurate and informed predictions within the context of medical informatics. These insights contribute to the development of precision medicine approaches and targeted interventions for patients at risk of developing dementia.

XSL•FO
RenderX

Discussion

Main Findings

This study investigates the application of a computer-aided risk prediction algorithm to identify risk factors for dementia in patients with T2DM. The algorithm has the potential to streamline clinical assessment and improve resource allocation in health care settings. The study also explores the modulatory effects of various medications on dementia risk, including not only antidiabetic drugs but also medications for other conditions. This comprehensive investigation offers valuable insights into the potential for ML to improve dementia risk stratification and early detection.

Our findings showed that age was the most critical feature that needs to be considered. It was in line with the findings of previous publications reporting age as the most decisive risk factor for dementia [43,44]. Aging is related to both physiological and pathological changes in the brain. At the cellular level, many types of neural cells are involved in aging. For example, older patients had significantly fewer neocortical oligodendrocytes than younger ones [43]. Another possible explanation is that older age is often accompanied by several comorbidities that may predispose individuals to dementia, such as cerebrovascular disease, stroke, or depressive disorder. Findings reconfirmed that the comorbidity burden, reflected by the CCI, was also associated with a 66% greater risk of dementia in our model.

A gender difference has been reported in dementia, regardless of the cause. Generally, female individuals were more likely to develop dementia than male individuals [45,46]. The role of gender in each subtype, however, may not be similar. For example, inflammatory dysregulation, a risk factor for Alzheimer disease, was reported to be more common in females [47]. Many women's health initiatives have shed light on the gender disparity in the risk of Alzheimer disease, including the difference in brain structure and loss of the neuroprotective effect of estrogen postmenopausal, resulting in neurobiological vulnerability in these individuals [48].

Among antidiabetic medications, insulin therapy was linked to an increased risk of dementia. This result was in accordance with the findings of a meta-analysis showing the harmful effects of insulin in patients with T2DM. Some authors hypothesized that this deleterious impact was driven by severe hypoglycemia events, which were more often seen in patients treated with insulin than those treated with other therapies [49]. Another possible explanation is that an indication for insulin is a surrogate indicator of long-lasting and more severe T2DM, which already increases the possibility of dementia. Therefore, the association between insulin therapy and dementia risk should be cautiously interpreted. The relationship between dipeptidyl peptidase-4 inhibitors and dementia was more complex. This class is generally considered to have a neutral effect on the hypoglycemic rate, making it impossible to explain the increased odds ratio in our model. Instead, dipeptidyl peptidase-4 inhibitors were preferably used in older patients. In the study, we also observed the performance of the ANN model that determined the risk of dementia stratified by different antidiabetic drugs (Table S5 and Figure S1 in Multimedia Appendix 1).

In contrast, metformin was potentially beneficial in reducing dementia development. Metformin was superior to placebo and sulfonylureas in preventing dementia [50-52]. Treatment with metformin has been found to alleviate neuroinflammation, enhance memory, and improve the survival ability of neural cells in animal experiments [53]. As an insulin sensitizer, metformin could act principally to ameliorate neuronal insulin resistance, a characteristic of Alzheimer disease [54]. More directly, metformin could prevent tau phosphorylation, which plays a central role in the pathogenesis of this disease [55].

Regarding blood pressure-lowering medications, although we found that β-blockers and renin-angiotensin system–acting agents may reduce the risk of dementia, previous publications have been inconsistent [56-58]. Therefore, no firm conclusions can be drawn to date. Further studies are warranted to investigate their actual effects.

The association between lipid-lowering agents and dementia risk should be considered distinctly for each class of drugs because of their different mechanisms of action. In our model, lipid-lowering drugs were found to lower the risk of dementia. This adjudication was, at least in part, driven by the benefit of statins, which are the most commonly used medications to treat hyperlipidemia. A meta-analysis confirmed the protective effect of statins on reducing all-cause dementia [59,60]. Hence, the reduction in the odds of dementia in patients with hyperlipidemia in our model might be explained by the fact that they were prescribed and benefited from statin use, regardless of their lipid profiles.

Our study exhibits numerous strengths that contribute to its robustness and clinical relevance. First, we adopted a comprehensive approach that thoroughly evaluated not only the primary disease, that is, T2DM, but also considered other pertinent health conditions and medications used by the patients. This approach aligns well with the clinical reality, as patients with DM often present with at least 1 comorbidity. Therefore, our evaluation strategy is consistent with the prevailing clinical guidelines' emphasis on conducting a comprehensive medical assessment.

Second, the use of a large-scale electronic clinical research database sourced from 3 prominent teaching hospitals provided us with a substantial sample size, encompassing a diverse population. This sizeable dataset enhances the statistical power of our analysis and augments the generalizability of our findings to broader patient populations. Also, this database contains important variables offering in-depth clinical characteristics that are often missed by studies using large health insurance claims data.

Third, the adoption of a longitudinal design allowed the investigation of dementia incidence rather than focusing solely on its prevalence. This longitudinal perspective allows us to observe disease development over time and provides a more accurate understanding of the temporal relationship between T2DM and the onset of dementia.

XSL•FO
RenderX

Fourth, the inclusion of an adequate follow-up period enabled us to detect potentially significant results, particularly in the context of the slow progression of dementia. By allowing sufficient observation time, we enhance the likelihood of capturing clinically relevant associations between T2DM and dementia.

Finally, our study incorporated a computer-aided risk prediction model to augment the classification performance. By using this sophisticated model, we could accurately identify individuals at a higher risk of developing dementia, aiding in early intervention and personalized care management.

Overall, these strengths collectively enhance the credibility and clinical utility of our study, providing valuable insights into the association between T2DM and dementia, and paving the way for improved patient care and management strategies.

Despite many strengths, this study still has inevitable limitations. First, this was a retrospective cohort study. Therefore, missing data or selection bias could be present. Second, the health care system allows patients to conveniently have laboratory or imaging tests at a different location before seeing a physician. Hence, some parts of the patients' health records could not be accessed on synchronized virtual platforms. Finally, despite incorporating a variety of medications into our model, we were unable to evaluate dose-response effects because of the complexity of the treatment regimens. This issue is beyond the scope of our topic, and further well-controlled prospective studies are needed to examine actual drug effects.

Conclusions

By implementing a computer-aided risk prediction algorithm, we have successfully developed a highly effective predictive model for identifying specific risk factors associated with dementia among individuals diagnosed with T2DM. The model's success was evident in capturing significant clinical features, including age, gender, the types of antidiabetic agents prescribed, comorbidities, and other long-term medications used. This predictive model holds immense potential in supporting health care professionals in the clinical diagnosis and management of patients with diabetic complications. With the provision of valuable insights, the model enables personalized and targeted care for this vulnerable population, thereby contributing to improved patient outcomes and overall health care efficacy.

Acknowledgments

This work was supported by National Science and Technology Council grants (grant NSTC113-2314-B-038-087) and Pfizer (grant A-0001582).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

PTP, JCH, and APN conceptualized and designed the study. NNN, APN, LMK provided suggestions for the research design from a clinical perspective. PTP collected data, performed the analyses, and drafted the manuscript (introduction, methods, results and discussion). NNN, KNQL, QVT drafted the manuscript (introduction, discussion, and conclusion). APN, NHB, LNQK reviewed all data and revised the manuscript critically for intellectual content. JCH, MHH, CL reviewed all data and revised the manuscript critically for intellectual content. All authors approved the final version for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 289 KB - jmir_v26i1e52107_app1.docx]

References

1. Carracher AM, Marathe PH, Close KL. International Diabetes Federation 2017. J Diabetes 2018;10(5):353-356. [doi: 10.1111/1753-0407.12644] [Medline: 29345068]
2. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. Phys Ther 2008;88(11):1254-1164 [FREE Full text] [doi: 10.2522/ptj.20080020] [Medline: 18801858]
3. Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020;133(8):895-900. [doi: 10.1016/j.amjmed.2020.03.033] [Medline: 32325045]
4. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9:103 [FREE Full text] [doi: 10.1186/1741-7015-9-103] [Medline: 21902820]

XSL•FO
RenderX

5. Lee DY, Kim J, Park S, Park SY, Yu JH, Seo JA, et al. Fasting glucose variability and the risk of dementia in individuals with diabetes: a nationwide cohort study. Diabetes Metab J 2022;46(6):923-935 [FREE Full text] [doi: 10.4093/dmj.2021.0346] [Medline: 35609876]
6. Cichosz SL, Jensen MH, Hejlesen O. Cognitive impairment in elderly people with prediabetes or diabetes: a cross-sectional study of the NHANES population. Prim Care Diabetes 2020;14(5):455-459. [doi: 10.1016/j.pcd.2019.11.005] [Medline: 31831376]
7. Xue M, Xu W, Ou YN, Cao XP, Tan MS, Tan L, et al. Diabetes mellitus and risks of cognitive impairment and dementia: a systematic review and meta-analysis of 144 prospective studies. Ageing Res Rev 2019;55:100944. [doi: 10.1016/j.arr.2019.100944] [Medline: 31430566]
8. DeTure MA, Dickson DW. The neuropathological diagnosis of alzheimer's disease. Mol Neurodegener 2019;14(1):32 [FREE Full text] [doi: 10.1186/s13024-019-0333-5] [Medline: 31375134]
9. Gudala K, Bansal D, Schifano F, Bhansali A. Diabetes mellitus and risk of dementia: a meta-analysis of prospective observational studies. J Diabetes Investig 2013;4(6):640-650 [FREE Full text] [doi: 10.1111/jdi.12087] [Medline: 24843720]
10. Cholerton B, Baker LD, Montine TJ, Craft S. Type 2 diabetes, cognition, and dementia in older adults: toward a precision health approach. Diabetes Spectr 2016;29(4):210-219 [FREE Full text] [doi: 10.2337/ds16-0041] [Medline: 27899872]
11. Arvanitakis Z, Shah RC, Bennett DA. Diagnosis and management of dementia: review. JAMA 2019;322(16):1589-1599 [FREE Full text] [doi: 10.1001/jama.2019.4782] [Medline: 31638686]
12. Vu GT, Tran BX, McIntyre RS, Pham HQ, Phan HT, Ha GH, et al. Modeling the research landscapes of artificial intelligence applications in diabetes (GAP). Int J Environ Res Public Health 2020;17(6):1982 [FREE Full text] [doi: 10.3390/ijerph17061982] [Medline: 32192211]
13. Cichosz SL, Johansen MD, Hejlesen O. Toward big data analytics: review of predictive models in management of diabetes and its complications. J Diabetes Sci Technol 2015;10(1):27-34 [FREE Full text] [doi: 10.1177/1932296815611680] [Medline: 26468133]
14. Kerr D, Klonoff DC. Digital diabetes data and artificial intelligence: a time for humility not hubris. J Diabetes Sci Technol 2019;13(1):123-127 [FREE Full text] [doi: 10.1177/1932296818796508] [Medline: 30182736]
15. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
16. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. Behav Ther 2020;51(5):675-687. [doi: 10.1016/j.beth.2020.05.002] [Medline: 32800297]
17. Bellazzi R, Dagliati A, Sacchi L, Segagni D. Big data technologies: new opportunities for diabetes management. J Diabetes Sci Technol 2015;9(5):1119-1125. [doi: 10.1177/1932296815583505] [Medline: 25910540]
18. Gast KB, Tjeerdema N, Stijnen T, Smit JWA, Dekkers OM. Insulin resistance and risk of incident cardiovascular events in adults without diabetes: meta-analysis. PLoS One 2012;7(12):e52036. [doi: 10.1371/journal.pone.0052036] [Medline: 23300589]
19. Ceriello A, de Cosmo S, Rossi MC, Lucisano G, Genovese S, Pontremoli R, et al. Variability in HbA1c, blood pressure, lipid parameters and serum uric acid, and risk of development of chronic kidney disease in type 2 diabetes. Diabetes Obes Metab 2017;19(11):1570-1578. [doi: 10.1111/dom.12976] [Medline: 28432733]
20. Larsson SC, Scott RA, Traylor M, Langenberg CC, Hindy G, Melander O, et al. Type 2 diabetes, glucose, insulin, BMI, and ischemic stroke subtypes: mendelian randomization study. Neurology 2017;89(5):454-460. [doi: 10.1212/WNL.0000000000004173] [Medline: 28667182]
21. Chen W, Pan Y, Jing J, Zhao X, Liu L, Meng X, et al. Recurrent stroke in minor ischemic stroke or transient ischemic attack with metabolic syndrome and/or diabetes mellitus. J Am Heart Assoc 2017;6(6):e005446 [FREE Full text] [doi: 10.1161/JAHA.116.005446] [Medline: 28572281]
22. Iwase M, Komorita Y, Ohkuma T, Fujii H, Ide H, Yoshinari M, et al. Incidence of stroke and its association with glycemic control and lifestyle in Japanese patients with type 2 diabetes mellitus: the fukuoka diabetes registry. Diabetes Res Clin Pract 2021;172:108518. [doi: 10.1016/j.diabres.2020.108518] [Medline: 33096191]
23. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
24. Nguyen PA, Hsu MH, Chang TH, Yang HC, Huang CW, Liao CT, et al. Taipei medical university clinical research database: a collaborative hospital EHR database aligned with international common data standards. BMJ Health Care Inform 2024;31(1):e100890. [doi: 10.1136/bmjhci-2023-100890] [Medline: 38749529]
25. Hsu JC, Nguyen PA, Phuc PT, Lo TC, Hsu MH, Hsieh MS, et al. Development and validation of novel deep-learning models using multiple data types for lung cancer survival. Cancers (Basel) 2022;14(22):5562. [doi: 10.3390/cancers14225562] [Medline: 36428655]
26. Wang K, Zhao S, Lee EKP, Yau SZM, Wu Y, Hung CT, et al. Risk of dementia among patients with diabetes in a multidisciplinary, primary care management program. JAMA Netw Open 2024;7(2):e2355733. [doi: 10.1001/jamanetworkopen.2023.55733] [Medline: 38345817]

XSL•FO
RenderX

27. Sheen YJ, Hsu CC, Jiang YD, Huang CN, Liu JS, Sheu WHH. Trends in prevalence and incidence of diabetes mellitus from 2005 to 2014 in Taiwan. J Formos Med Assoc 2019;118 Suppl 2:S66-S73 [FREE Full text] [doi: 10.1016/j.jfma.2019.06.016] [Medline: 31300322]
28. Ioannou GN, Tang W, Beste LA, Tincopa MA, Su GL, Van T, et al. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. JAMA Netw Open 2020;3(9):e2015626. [doi: 10.1001/jamanetworkopen.2020.15626] [Medline: 32870314]
29. Savelieff MG, Chen KS, Elzinga SE, Feldman EL. Diabetes and dementia: clinical perspective, innovation, knowledge gaps. J Diabetes Complications 2022;36(11):108333 [FREE Full text] [doi: 10.1016/j.jdiacomp.2022.108333] [Medline: 36240668]
30. Rawlings AM, Sharrett AR, Albert MS, Coresh J, Windham BG, Power MC, et al. The association of late-life diabetes status and hyperglycemia with incident mild cognitive impairment and dementia: the ARIC study. Diabetes Care 2019;42(7):1248-1254. [doi: 10.2337/dc19-0120] [Medline: 31221696]
31. Sadahiro R, Sawada N, Matsuoka YJ, Mimura M, Nozaki S, Shikimoto R, et al. Midlife cancer/diabetes and risk of dementia and mild cognitive impairment: a population-based prospective cohort study in Japan. Psychiatry Clin Neurosci 2019;73(9):597-599. [doi: 10.1111/pcn.12905] [Medline: 31226234]
32. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40-49 [FREE Full text] [doi: 10.1002/mpr.329] [Medline: 21499542]
33. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35(5-6):352-359 [FREE Full text] [doi: 10.1016/s1532-0464(03)00034-0] [Medline: 12968784]
34. Vaibhaw, Sarraf J, Pattnaik PK. Brain-Computer Interfaces and Their Applications. Cambridge, MA: Academic Press, Elsevier Inc; 2020:31-54.
35. Friedman JH. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 2001;29(5):1189-1232.
36. LightGBM. Microsoft. 2022. URL: https://www.microsoft.com/en-us/research/project/lightgbm/ [accessed 2024-10-30]
37. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaBoost. Statistics and Its Interface 2022;2(3):349-360. [doi: 10.4310/SII.2009.v2.n3.a8]
38. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology 2007;88(11):2783-2792. [doi: 10.1890/07-0539.1] [Medline: 18051647]
39. Kumar NM, Manjula R. Smart Intelligent Computing and Applications. Singapore: Springer Nature Singapore Pte Ltd; 2019:689.
40. Kim D, You S, So S, Lee J, Yook S, Jang DP, et al. A data-driven artificial intelligence model for remote triage in the prehospital environment. PLoS One 2018;13(10):e0206006. [doi: 10.1371/journal.pone.0206006] [Medline: 30352077]
41. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006;27(8):861-874. [doi: 10.1016/j.patrec.2005.10.010]
42. Ning Y, Ong MEH, Chakraborty B, Goldstein BA, Ting DSW, Vaughan R, et al. Shapley variable importance cloud for interpretable machine learning. Patterns (N Y) 2022;3(4):100452 [FREE Full text] [doi: 10.1016/j.patter.2022.100452] [Medline: 35465224]
43. Fabricius K, Jacobsen JS, Pakkenberg B. Effect of age on neocortical brain cells in 90+ year old human females--a cell counting study. Neurobiol Aging 2013;34(1):91-99. [doi: 10.1016/j.neurobiolaging.2012.06.009] [Medline: 22878165]
44. Streit WJ, Braak H, Xue QS, Bechmann I. Dystrophic (senescent) rather than activated microglial cells are associated with tau pathology and likely precede neurodegeneration in Alzheimer's disease. Acta Neuropathol 2009;118(4):475-485. [doi: 10.1007/s00401-009-0556-6] [Medline: 19513731]
45. Jia L, Du Y, Chu L, Zhang Z, Li F, Lyu D, et al. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. Lancet Public Health 2020;5(12):e661-e671 [FREE Full text] [doi: 10.1016/S2468-2667(20)30185-7] [Medline: 33271079]
46. Seshadri S, Wolf PA, Beiser A, Au R, McNulty K, White R, et al. Lifetime risk of dementia and alzheimer's disease. The impact of mortality on risk estimates in the Framingham study. Neurology 1997;49(6):1498-1504. [doi: 10.1212/wnl.49.6.1498] [Medline: 9409336]
47. Hall JR, Wiechmann AR, Johnson LA, Edwards M, Barber RC, Winter AS, et al. Biomarkers of vascular risk, systemic inflammation, and microvascular pathology and neuropsychiatric symptoms in Alzheimer's disease. J Alzheimers Dis 2013;35(2):363-371 [FREE Full text] [doi: 10.3233/JAD-122359] [Medline: 23403534]
48. Snyder HM, Asthana S, Bain L, Brinton R, Craft S, Dubal DB, et al. Sex biology contributions to vulnerability to Alzheimer's disease: a think tank convened by the women's alzheimer's research Initiative. Alzheimers Dement 2016;12(11):1186-1196 [FREE Full text] [doi: 10.1016/j.jalz.2016.08.004] [Medline: 27692800]
49. McMillan JM, Mele BS, Hogan DB, Leung AA. Impact of pharmacological treatment of diabetes mellitus on dementia risk: systematic review and meta-analysis. BMJ Open Diabetes Res Care 2018;6(1):e000563. [doi: 10.1136/bmjdrc-2018-000563] [Medline: 30487973]
50. Samaras K, Makkar S, Crawford JD, Kochan NA, Wen W, Draper B, et al. Metformin use is associated with slowed cognitive decline and reduced incident dementia in older adults with Type 2 diabetes: the Sydney memory and ageing study. Diabetes Care 2020;43(11):2691-2701. [doi: 10.2337/dc20-0892] [Medline: 32967921]

XSL•FO
RenderX

51. Orkaby AR, Cho K, Cormack J, Gagnon DR, Driver JA. Metformin vs sulfonylurea use and risk of dementia in US veterans aged ≥65 years with diabetes. Neurology 2017;89(18):1877-1885 [FREE Full text] [doi: 10.1212/WNL.0000000000004586] [Medline: 28954880]
52. Campbell JM, Stephenson MD, de Courten B, Chapman I, Bellman SM, Aromataris E. Metformin use associated with reduced risk of dementia in patients with diabetes: a systematic review and meta-analysis. J Alzheimers Dis 2018;65(4):1225-1236. [doi: 10.3233/JAD-180263] [Medline: 30149446]
53. Oliveira WH, Nunes AK, França MER, Santos LA, Lós DB, Rocha SW, et al. Effects of metformin on inflammation and short-term memory in streptozotocin-induced diabetic mice. Brain Res 2016;1644:149-160. [doi: 10.1016/j.brainres.2016.05.013] [Medline: 27174003]
54. Gupta A, Bisht B, Dey CS. Peripheral insulin-sensitizer drug metformin ameliorates neuronal insulin resistance and Alzheimer's-like changes. Neuropharmacology 2011;60(6):910-920. [doi: 10.1016/j.neuropharm.2011.01.033] [Medline: 21277873]
55. Kickstein E, Krauss S, Thornhill P, Rutschow D, Zeller R, Sharkey J, et al. Biguanide metformin acts on tau phosphorylation via mTOR/protein phosphatase 2A (PP2A) signaling. Proc Natl Acad Sci U S A 2010;107(50):21830-21835 [FREE Full text] [doi: 10.1073/pnas.0912793107] [Medline: 21098287]
56. Holm H, Ricci F, Di Martino G, Bachus E, Nilsson ED, Ballerini P, et al. Beta-blocker therapy and risk of vascular dementia: a population-based prospective study. Vascul Pharmacol 2020;125-126:106649. [doi: 10.1016/j.vph.2020.106649] [Medline: 31958512]
57. den Brok MGHE, van Dalen JW, Abdulrahman H, Larson EB, van Middelaar T, van Gool WA, et al. Antihypertensive medication classes and the risk of dementia: a systematic review and network meta-analysis. J Am Med Dir Assoc 2021;22(7):1386-1395.e15. [doi: 10.1016/j.jamda.2020.12.019] [Medline: 33460618]
58. Ding J, Davis-Plourde KL, Sedaghat S, Tully PJ, Wang W, Phillips C, et al. Antihypertensive medications and risk for incident dementia and Alzheimer's disease: a meta-analysis of individual participant data from prospective cohort studies. Lancet Neurol 2020;19(1):61-70. [doi: 10.1016/S1474-4422(19)30393-X] [Medline: 31706889]
59. Poly TN, Islam MM, Walther BA, Yang HC, Wu CC, Lin MC, et al. Association between use of statin and risk of dementia: a meta-analysis of observational studies. Neuroepidemiology 2020;54(3):214-226. [doi: 10.1159/000503105] [Medline: 31574510]
60. Chu CS, Tseng PT, Stubbs B, Chen TY, Tang CH, Li DJ, et al. Use of statins and the risk of dementia and mild cognitive impairment: a systematic review and meta-analysis. Sci Rep 2018;8(1):5804. [doi: 10.1038/s41598-018-24248-8] [Medline: 29643479]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
AUC: Area Under the Curve
CCI: Charlson Comorbidity Index
DM: diabetes mellitus
GBM: gradient boosting machine
HbA1c: hemoglobin A1c
ICD-9-CM: International Classification of Disease, Ninth Revision, Clinical Modification
ICD-10-CM: International Classification of Disease, Tenth Revision, Clinical Modification
LGBM: light gradient boosting machine
ML: machine learning
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus
TMUCRD: Taipei Medical University Clinical Research Database
RF: random forest
SHAP: Shapley additive explanation

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 23.08.23; peer-reviewed by X Yang, PH Liao, L Aljerf; comments to author 22.04.24; revised version received 06.07.24; accepted 21.10.24; published 11.12.24.

Please cite as:
Thanh Phuc P, Nguyen PA, Nguyen NN, Hsu MH, Le NQK, Tran QV, Huang CW, Yang HC, Chen CY, Le TAH, Le MK, Nguyen HB, Lu CY, Hsu JC
Early Detection of Dementia in Populations With Type 2 Diabetes: Predictive Analytics Using Machine Learning Approach
J Med Internet Res 2024;26:e52107
URL: https://www.jmir.org/2024/1/e52107
doi:10.2196/52107
PMID:39434474

©Phan Thanh Phuc, Phung-Anh Nguyen, Nam Nhat Nguyen, Min-Huei Hsu, Nguyen Quoc Khanh Le, Quoc-Viet Tran, Chih-Wei Huang, Hsuan-Chia Yang, Cheng-Yu Chen, Thi Anh Hoa Le, Minh Khoi Le, Hoang Bac Nguyen, Christine Y Lu, Jason C Hsu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model

Xiaolan Chen[1*], MD; Ziwei Zhao[1*], MD; Weiyi Zhang[1], MS; Pusheng Xu[1], MD; Yue Wu[1], MD; Mingpu Xu[1], MD; Le Gao[2], MD; Yinwen Li[3,4], MD; Xianwen Shang[1], PhD; Danli Shi[1,5], MD, PhD; Mingguang He[1,5,6], MD, PhD

[1]School of Optometry, The Hong Kong Polytechnic University, Hong Kong, China

[2]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China

[3]Department of Ophthalmology, Shanghai General Hospital (Shanghai First People's Hospital), School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[4]National Clinical Research Center for Eye Diseases, Shanghai, China

[5]Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Hong Kong, China

[6]Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong, China

[*]these authors contributed equally

Corresponding Author:
Danli Shi, MD, PhD
School of Optometry
The Hong Kong Polytechnic University
11 Yuk Choi Road
Hung Hom, KLN
Hong Kong, 999077
China
Phone: 852 27664825
Email: danli.shi@polyu.edu.hk

Abstract

Background: Large language models (LLMs) have the potential to enhance clinical flow and improve medical education, but they encounter challenges related to specialized knowledge in ophthalmology.

Objective: This study aims to enhance ophthalmic knowledge by refining a general LLM into an ophthalmology-specialized assistant for patient inquiries and medical education.

Methods: We transformed Llama2 into an ophthalmology-specialized LLM, termed EyeGPT, through the following 3 strategies: prompt engineering for role-playing, fine-tuning with publicly available data sets filtered for eye-specific terminology (83,919 samples), and retrieval-augmented generation leveraging a medical database and 14 ophthalmology textbooks. The efficacy of various EyeGPT variants was evaluated by 4 board-certified ophthalmologists through comprehensive use of 120 diverse category questions in both simple and complex question-answering scenarios. The performance of the best EyeGPT model was then compared with that of the unassisted human physician group and the EyeGPT+human group. We proposed 4 metrics for assessment: accuracy, understandability, trustworthiness, and empathy. The proportion of hallucinations was also reported.

Results: The best fine-tuned model significantly outperformed the original Llama2 model at providing informed advice (mean 9.30, SD 4.42 vs mean 13.79, SD 5.70; $P<.001$) and mitigating hallucinations (97/120, 80.8% vs 53/120, 44.2%, $P<.001$). Incorporating information retrieval from reliable sources, particularly ophthalmology textbooks, further improved the model's response compared with solely the best fine-tuned model (mean 13.08, SD 5.43 vs mean 15.14, SD 4.64; $P=.001$) and reduced hallucinations (71/120, 59.2% vs 57/120, 47.4%, $P=.02$). Subgroup analysis revealed that EyeGPT showed robustness across common diseases, with consistent performance across different users and domains. Among the variants, the model integrating fine-tuning and book retrieval ranked highest, closely followed by the combination of fine-tuning and the manual database, standalone fine-tuning, and pure role-playing methods. EyeGPT demonstrated competitive capabilities in understandability and empathy when compared with human ophthalmologists. With the assistance of EyeGPT, the performance of the ophthalmologist was notably enhanced.

XSL•FO
RenderX

Conclusions: We pioneered and introduced EyeGPT by refining a general domain LLM and conducted a comprehensive comparison and evaluation of different strategies to develop an ophthalmology-specific assistant. Our results highlight EyeGPT's potential to assist ophthalmologists and patients in medical settings.

(J Med Internet Res 2024;26:e60063) doi:10.2196/60063

KEYWORDS

large language model; generative pretrained transformer; generative artificial intelligence; ophthalmology; retrieval-augmented generation; medical assistant; EyeGPT; generative AI

Introduction

Ophthalmic diseases pose significant concerns for public health [1]. However, shortages of professionals and inefficiencies in primary eye care systems often funnel patients into overcrowded tertiary centers. This results in extended wait times and unaddressed postconsultation questions, frequently requiring additional face-to-face appointments [2]. These challenges can be attributed to the limited ophthalmic knowledge among patients and the limited experience in eye care among primary health care providers [3]. Therefore, there is a pressing need to enhance ophthalmic health education for both patients and primary health care providers. However, relying solely on manpower to address these issues presents further challenges, particularly as the rate of population aging continues to outpace the growth rate of ophthalmologists.

Large language models (LLMs) have recently emerged as powerful tools to alleviate these burdens and streamline clinical flow with the capability of understanding and generating human-like text [4]. In ophthalmology, LLMs show promise both for ophthalmic certification exams [5] and interpreting imaging reports across various linguistic environments [6,7]. However, there are several limitations to existing LLMs. First, there are challenges with addressing specialized ophthalmology knowledge for general LLMs. Previous research has demonstrated the suboptimal performance of ChatGPT in ophthalmology, with only 15.4% of the responses graded as completely accurate in vitreoretinal disease [8]. Even with GPT-4, which currently exhibits the greatest capability, nonnegligible instances of misinformation occur, with only 30.6%, 21.5%, and 55.6% of responses about ocular multimodal images considered accurate, highly usable, and harmless, respectively [9]. A critical factor underlying these shortcomings is the model's insufficient grasp of specialized knowledge, particularly in handling medical abbreviations and jargon within highly specialized domains [5]. Therefore, there is a need to design a dedicated model trained on clinically relevant domain data. Second, it is widely recognized that LLMs occasionally generate inaccurate and misleading statements (hallucinations), which can potentially lead to medical errors. Fine-tuning with professional data can somewhat mitigate hallucinations, but the model can still produce them when faced with unfamiliar input [10]. Therefore, additional solutions are required. Third, there is a noticeable absence of comprehensive evaluations for LLMs in ophthalmology. Although previous studies have explored the ophthalmic question-answering (QA) capabilities of LLMs, the majority have been limited to multiple-choice formats [5,11-13]. Although a few studies have used open-ended questions to evaluate the performance of LLMs, they lack detailed categorization of the questions and primarily focus on scattered aspects such as accuracy, comprehensiveness, or safety [14,15]. Consequently, a comprehensive evaluation framework is urgently needed to test ophthalmology-related LLMs and compare their responses with those provided by certified ophthalmologists.

Recognizing this, we aimed to develop an artificial intelligence (AI) assistant, namely EyeGPT, to meet the specific informational needs in ophthalmic clinical and educational scenarios. By leveraging Llama2, a flexible and scalable open-source LLM known for its impressive performance in medicine [16-18], we infused the model with a granular level of ophthalmic expertise through role-playing, fine-tuning, and retrieval-augmented generation (RAG). The resultant model, EyeGPT, was evaluated for its efficacy in patient consultations and medical education. This work provides valuable insights into building and evaluating ophthalmic assistants, paving the way for the next generation of AI-assisted ophthalmic practice.

Methods

Ethical Considerations

The study overview is presented in Figure 1. Our research protocol adhered to the principles of the Helsinki Declaration. The study was approved by the Institutional Review Board of the Hong Kong Polytechnic University (number: HSEARS20240202004). This research involves publicly available data. We ensured that the data were deidentified and all private information was removed. Informed consent was unnecessary as the publicly available data do not contain identifiable information.

XSL•FO
RenderX

Figure 1. Overview of this study. GPT: generative pre-trained transformer; MCQA: multiple-choice question answering; RAG: retrieval-augmented generation; USMLE=United States Medical Licensing Examination.

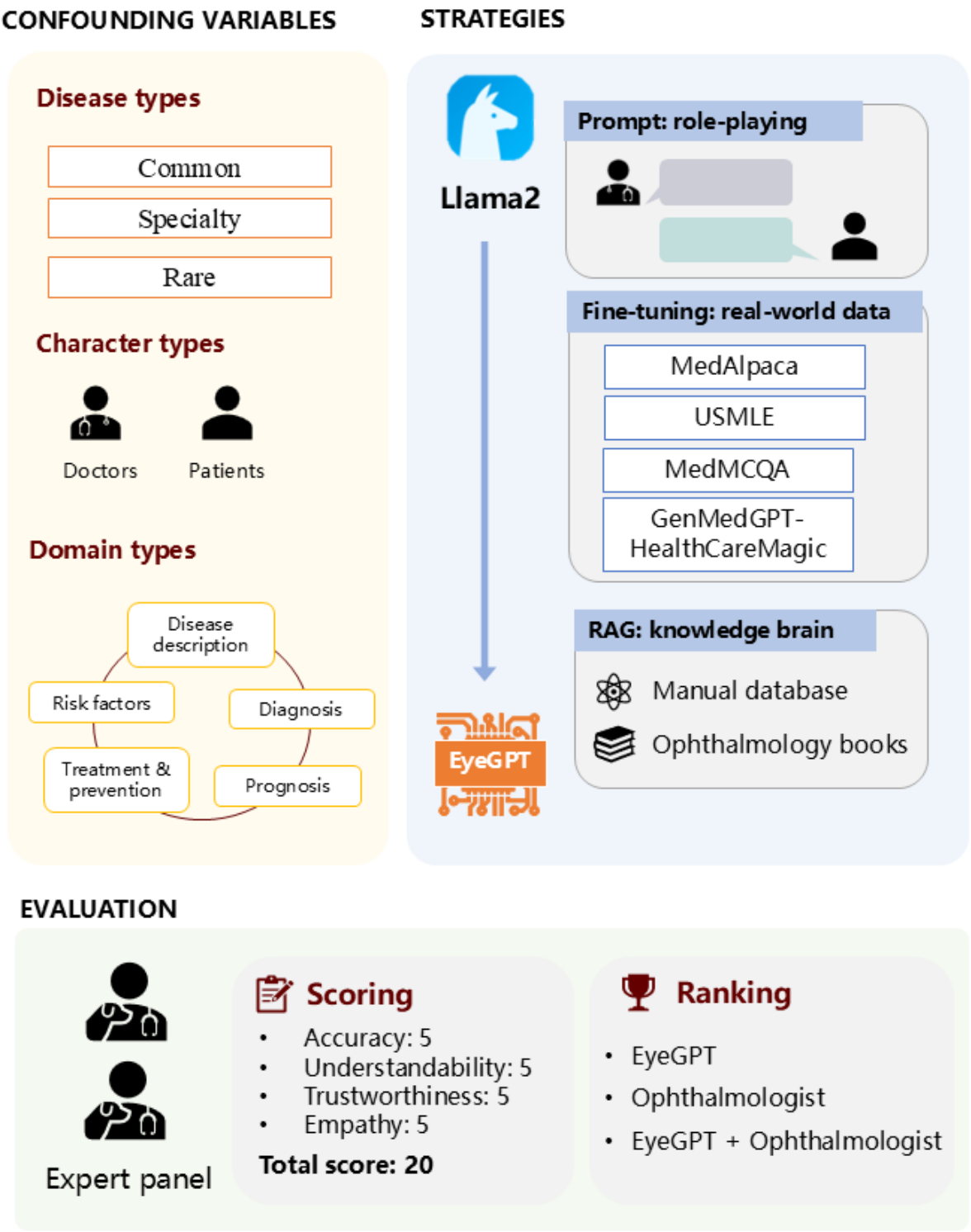

Development of EyeGPT

Base Model

We used Meta's Llama2 as the base model in our study, which was trained on 2 trillion tokens from publicly accessible data [19]. We used the Llama2-7b-chat model, which was additionally fine-tuned on publicly available instruction data sets and over 1 million human annotations, thus having basic conversation skills [20]. To inject professional ophthalmic knowledge into the model, we did experiments successively under the scenarios described in the following paragraphs.

Role-Playing

In generative AI, the engineering technique known as "role-playing" involves directing LLMs to "embody" or "imitate" specific roles for improved results [21]. To enable the

XSL•FO
RenderX

LLM to generate more relevant and empathetic responses, we assigned it the role of an "ophthalmologist" and the user the dual roles of a "patient" and "medical student." This was achieved by giving the following instructions: "Suppose you are an ophthalmologist, and you need to answer the patient's question with care/student's question with patience."

Fine-Tuning

To inject domain-specific knowledge and make Llama2 more proficient in capturing ophthalmic terminologies and logical reasoning, we trained it on domain-specific data sets, including MedAlpaca [22], GenMedGPT-HealthCareMagic [23], MedMCQA [24], and the United States Medical Licensing Examination (USMLE). Processing of the USMLE data followed the method proposed by Jin et al [25]. The data sets were filtered to remove conversations of little practical significance and responses with errors. We used instruction tuning [26] to align the model with task-specific user objectives, enhance model controllability, and ensure rapid domain-specific adaptation. For data sets initially designed for multiple-choice QA, we automatically added an instruction at the beginning: "Answer the multiple-choice question." For our specific task, we filtered out nonophthalmology data with eye-related keywords. Multimedia Appendix 1 presents the characteristics of the filtered data sets, and Multimedia Appendix 2 lists the keywords we used.

The final data set comprised 83,919 samples, with 81,919 used for training and 2000 used for validation. We used low-rank adaptation (LoRA) [27] to fine-tune the Llama2-7B model by adding a low-rank matrix while keeping the original parameters frozen, aiming to complement the original weight matrices of the model. The models were fine-tuned using 3*V100 GPUs with a batch size of 24, learning rate of 0.00003, maximum sequence length of 512 tokens, and warm-up ratio of 0.03. For LoRA-specific hyperparameters, the rank of low-rank factorization was 8, the scaling factor for the rank was 16, and the dropout was 0.05. Specifically, we performed 3 types of fine-tuning: Fine-tune 1 (2000 iterations), Fine-tune 2 (3500 iterations), and Fine-tune 3 (10,000 iterations). The entire training process took approximately 11 hours to complete.

Retrieval-Augmented Generation

LLMs may produce potential inaccuracies responses (hallucinations) to questions [28], which is unacceptable in the medical field. However, the accuracy of these models could be significantly improved if they could generate responses based on a reliable knowledge database. Here, to further improve the performance of EyeGPT, we introduced the external knowledge corpus of medical books and a manual database.

For the medical books, we used 14 specialized ophthalmology textbooks that cover a wide range of comprehensive ophthalmic knowledge, including general ophthalmology, optometry, retinal diseases, and more [29-31]. Please refer to Multimedia Appendix 3 for the specific textbook list.

We manually built a database (sample shown in Multimedia Appendix 4) containing information on diseases, symptoms, medical tests and treatment procedures, and potential medications. This database, sourced from the open-access web and research papers, serves as an external and offline knowledge corpus for EyeGPT. It can be continually updated without model retraining and may provide more up-to-date information than textbooks.

To leverage external knowledge, we adopted the LangChain framework's information retrieval techniques. The "all-MiniLM-L6-v2" [32] open-source embedding model was used to map text into vector space. We used the "RecursiveCharacterTextSplitter" [33] to segment the text for efficient retrieval, with a chunk size set to 1024 characters. Roughly 2 segments are retrieved from the vector storage for each response. In addition, we constructed a retriever with Facebook AI Similarity Search (FAISS) [34] based on the segmented documents and established a conversational retrieval chain that seamlessly integrated our EyeGPT with the external database through LangChain.

Evaluation

Overview of the Evaluation

To assess the professional performance of various EyeGPT variants, namely (1) original (Llama2), (2) role-play (original plus role-play), (3) fine-tune 1-3 (fine-tuned model versions 1-3 plus role-play), (4) role-play+book (role-play plus book retrieval), (5) role-play+database (role-play plus manual database retrieval), (6) best fine-tune+book (the best fine-tuned model plus book retrieval), (7) best fine-tune+database (the best fine-tuned model plus manual database retrieval), our ophthalmology expert panel curated a set of 120 ophthalmic care-related questions based on their clinical expertise. We followed the user-centered evaluation approach proposed by Abbasian et al [35], considering the following 3 key factors: disease type, character type, and domain type. Disease type covered a wide range of medical conditions from various subspecialties, including common, specialty, and rare diseases, resulting in 12 disease categories such as myopia, retinal detachment, and Stickler syndrome (refer to Multimedia Appendix 5 for the detailed disease list). Character types included patients and medical students representing potential EyeGPT users. Domain types were divided into 5 topics: disease description, risk factors, diagnosis, treatment and prevention, and prognosis. We conducted the evaluations manually, including an independent evaluation of different EyeGPT variants, best-ranked comparisons for evaluating human-machine performance, and error analysis of the machine.

Independent Evaluation

This evaluation was designed to compare the performance of various optimization strategies of the EyeGPT variants and identify the best-performing one. Two board-certified ophthalmologists independently conducted manual assessment using a 5-point scale to assess the responses of each variant. The evaluation focused on the following 4 aspects: accuracy, understandability, trustworthiness, and empathy [35]. The detailed grading scale is presented in Multimedia Appendix 6. The scale ranged from 1 (strongly disagree) to 5 (strongly agree), with the average score from the 2 evaluators recorded as the score for each response aspect. The maximum score for each

XSL•FO
RenderX

aspect was 5, and these scores were summed to obtain the final score for each response, with a maximum possible score of 20.

To evaluate the effectiveness of different optimization strategies in mitigating hallucinations, we defined answers with accuracy scores below 4 as containing hallucinations in our study. To ensure the evaluators could not identify the source of the responses, all generated responses were formatted as plain text, concealing any model-specific features. These responses were then randomly shuffled and mixed before being presented to the evaluators.

The evaluation was conducted in 2 rounds with a 1-month washout period to mitigate residual effects [36,37]. In the first round, we compared models using different fine-tuning approaches, including original, role-play, and fine-tune 1-3. The goal was to determine the best fine-tuning model for the subsequent RAG. In the second round, we compared models using different RAG strategies based on the best-performing fine-tuned model selected from the first round. These models included best fine-tune (the best fine-tuned model from round 1), role-play+database, best fine-tune+database, role-play+book, and best fine-tune+book.

Best-Ranked Comparison

After independently evaluating the different EyeGPT variants, we identified the best-performing system. To assess if EyeGPT can match ophthalmologists' expertise and offer them assistance, we conducted a human-machine best-ranked comparison. This evaluation method, inspired by that of Tu et al [38], aimed to efficiently assess answers comprehensively, reducing the need for assessors to delve into every detail and thereby minimizing subjectivity.

We invited 2 junior ophthalmologists (with 1-3 years of clinical experience) to answer the 120 questions with and without the aid of EyeGPT. The answers from different groups (EyeGPT, unassisted ophthalmologist, and EyeGPT+ophthalmologist) were evaluated by 2 senior ophthalmologists (with over 3 years of clinical experience) who were unaware of the sources, and the presentation order was randomized. Raters were asked to rank the 3 answers based on their clinical judgment across 4 dimensions, without the option of declaring a tie. In cases of disagreement, an ophthalmology expert (with over 10 years of clinical experience) reviewed the case until consensus was reached. The final result was recorded as the proportion of responses from different sources ranked as the best.

Error Analysis

To further investigate the quality of EyeGPT answers and identify areas for improvement, we conducted an error analysis on the best-performing EyeGPT model. The quality of the EyeGPT-generated QA pairs was evaluated by 2 board-certified ophthalmologists based on their expert judgment. The analysis focused on identifying occurrences of unrelated information, factual errors, incomplete information, and faulty logic [39].

Statistical Analysis

Statistical analyses were conducted using R (Version 4.3.1). The Mann-Whitney *U* test was used to compare the scores of the 2 models in the independent evaluation. When creating the bar chart, we compared the performance of the base model (Llama 2 or best fine-tune) with the most competitive optimization model in the same round to display statistically significant differences on the chart. The score for each answer in the independent evaluation was based on the average score from 2 raters. The scoring criteria used in the bar chart were as follows: strongly disagree (1 to <2), agree (2 to <3), neutral (3 to <4), approve (4 to <5), strongly agree (5). For subgroup analysis based on different confounding variables, the Kruskal-Wallis test and Mann-Whitney *U* test were used, depending on the number of comparison groups. Cohen kappa was calculated to determine the agreement among raters [40]. *P* values <.05 were considered statistically significant.

Results

Comparative Study of Model Construction Strategies

Overall Performance

In the first round of evaluation, the total scores for the original, role-play, and fine-tune 1-3 models were 9.30, 12.79, 12.95, 12.83, and 13.79, respectively. All optimized models significantly outperformed the original model in accuracy, understandability, trustworthiness, and empathy, with fine-tune 3 performing the best. For the different fine-tuning variants, we observed that, as the number of iterations increased, the evaluation loss on the test data decreased (refer to Multimedia Appendix 7) and the model performance improved. In the subsequent comparison of RAG strategies, the best fine-tune+book model scored the highest, at 15.14, outperforming other strategies, as elaborated in Multimedia Appendix 8**.** To ensure reliability, we compared the scores of fine-tune 3 (named best fine-tune in round 2) across 2 rounds. We found no statistically significant difference between the scores of the 2 rounds (*P*=.11). Inter-rater reliability in 2 rounds of independent evaluation was confirmed, with kappa values ranging from 0.611 to 0.872, indicating substantial agreement among raters (Multimedia Appendix 9). For illustrative examples of the varied grades of responses from the independent evaluation, see Multimedia Appendix 10.

Figure 2 demonstrates that more than one-half (accuracy: 67/120, 55.8%; understandability: 74/120, 61.7%; trustworthiness: 75/120, 62.5%; empathy: 74/120, 61.7%) of responses from the best fine-tune model were considered "good" responses (rated 4 or above) across all 4 dimensions. Compared with the original model (with an 80.8% [97/120] hallucination rate), the role-play and best fine-tune models mitigated hallucinations, by 30% (36/120) and 36.7% (44/120), respectively. Figure 3 shows that the best fine-tune+book model further enhanced the proportion of "good" responses to the maximum. We compared the performance of the best model in round 1 (fine-tune 3) with the most competitive modified model to check for statistically significant differences. The scores and scoring criteria are the same as in Figure 2. Compared with the best fine-tune model, the best fine-tune+database and best fine-tune+book models further reduced hallucinations by 3.3% (4/120) and 11.7% (14/120), respectively.

XSL•FO
RenderX

Figure 2. Performance in terms of (A) accuracy, (B) understandability, (C) trustworthiness, and (D) empathy of the different models in round 1 of the human evaluation, with the percentage of good responses (strongly agree and agree) indicated by the black numbers, the percentage of hallucinations indicated by the blue numbers, and significance determined using Mann-Whitney *U* tests.

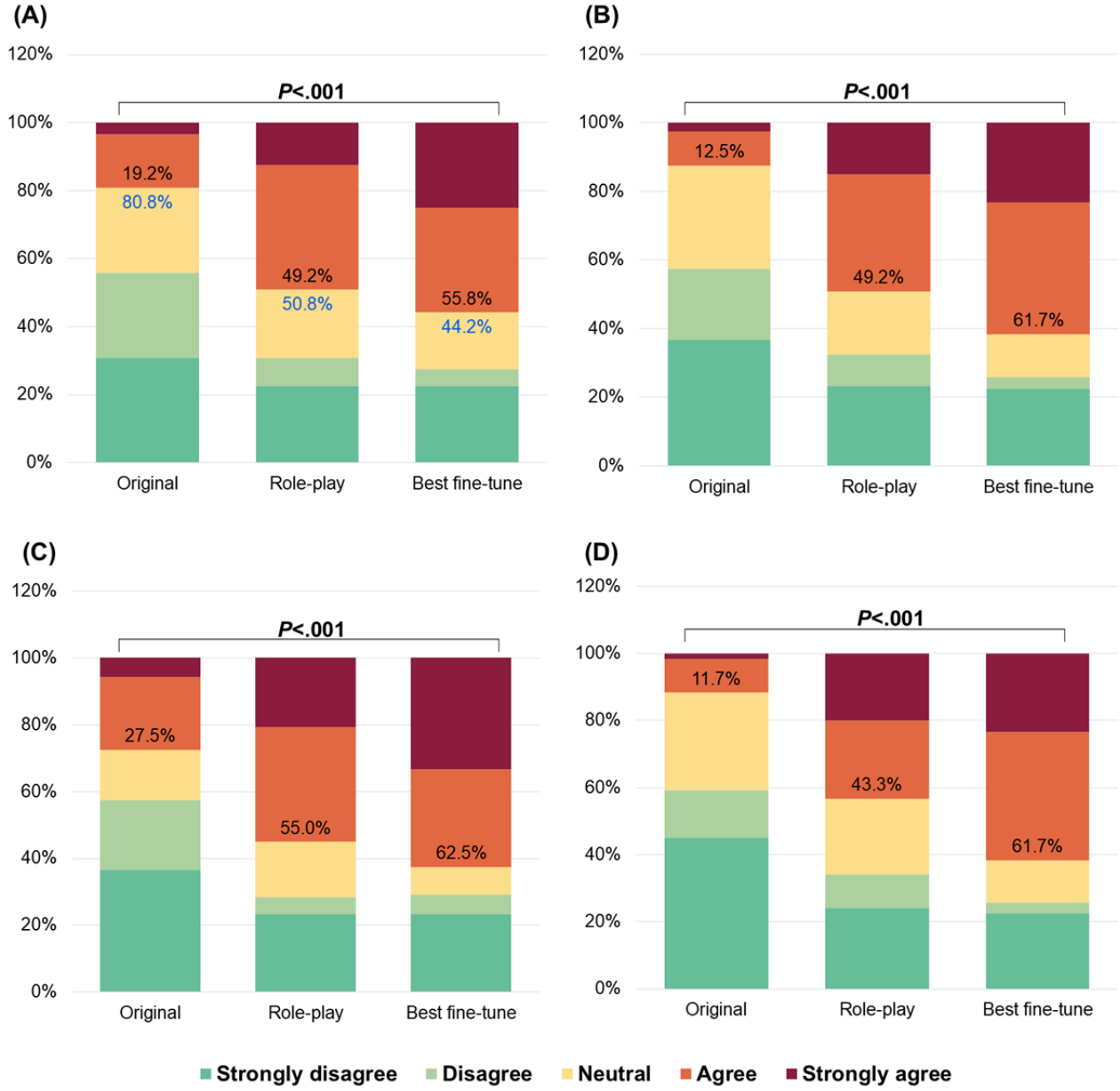

XSL•FO
RenderX

Figure 3. Performance in terms of (A) accuracy, (B) understandability, (C) trustworthiness, and (D) empathy of the different models in round 2 of the human evaluation, with the percentage of good responses (strongly agree and agree) indicated by the black numbers, the percentage of hallucinations indicated by the blue numbers, and significance determined using Mann-Whitney *U* tests.

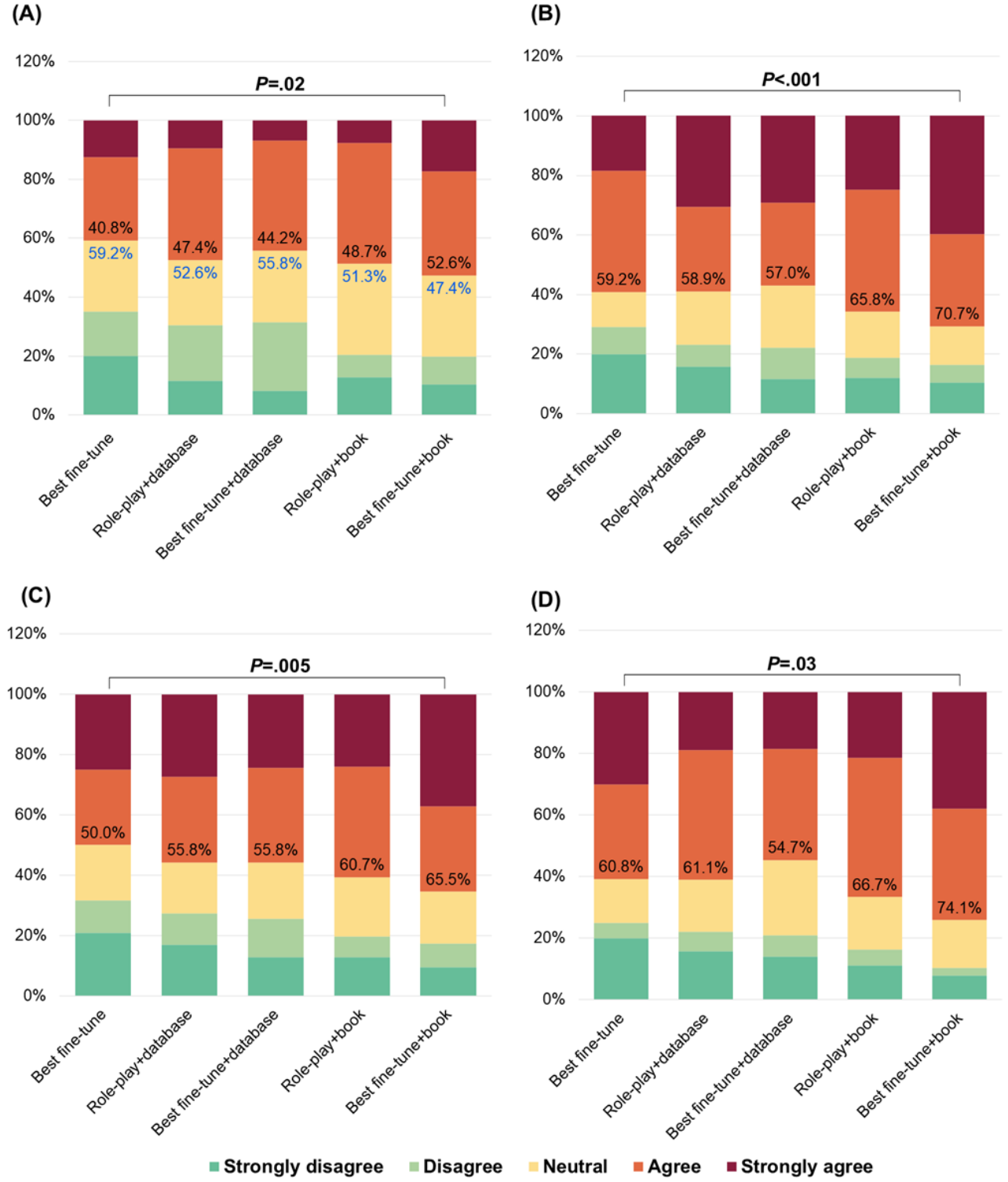

Subgroup Analysis

We also performed subgroup analysis to further evaluate the model performance under different confounding factors, including subspecialty questions of varying difficulty levels, questions raised by different characters, and question domains.

Different Subspecialties

Across all RAG strategies, the models scored higher for common diseases than for specialty and rare conditions (Table 1). For common ophthalmic conditions, the RAG models delivered more precise and contextually relevant information. For more specialized conditions like central serous chorioretinopathy, the best fine-tune model provided general information about its treatment options, while the RAG models offered more

XSL•FO
RenderX

specialized responses concerning laser treatment and photodynamic therapy depending on the specific circumstances. For rare conditions like morning glory syndrome, although the best fine-tune model could not generate responses as it mistakenly identified it as "bilateral posterior superior temporal arcade spikes," the RAG model was able to retrieve relevant information from the external knowledge database and make accurate responses. The best fine-tune model accurately recognized 43% (13/30) of ophthalmic abbreviations. RAG strategies improved this recognition rate, ranging from 60% (18/30) to 83% (25/30) for different models.

Table 1. Subgroup analysis of the performance of EyeGPT by subspecialty.

EyeGPT model	Common diseases, mean (SD)[a]	Specialty diseases, mean (SD)[a]	Rare diseases, mean (SD)[a]	*P* value
Best fine-tune[b]	15.79 (4.15)	13.11 (5.40)	10.33 (5.31)	<.001
Role-play+database[c]	15.28 (5.17)	14.18 (5.11)	12.18 (4.88)	.01
Best fine-tune+database[d]	15.45 (4.57)	14.29 (4.89)	12.17 (4.58)	.01
Role-play+book[e]	15.70 (3.94)	12.89 (5.32)	14.66 (4.53)	.02
Best fine-tune+book[f]	17.24 (3.02)	14.08 (5.49)	14.23 (4.42)	.003

[a]Overall response score (the sum of 4 rating dimensions, with a maximum score of 20 representing the best performance).
[b]The fine-tuned model with 10,000 iterations.
[c]Role-play plus manual database retrieval.
[d]The best fine-tuned model plus manual database retrieval.
[e]Role-play plus book retrieval.
[f]The best fine-tuned model plus book retrieval.

Different Role-Play Characters

When comparing the influence of the questioner's assumed identity—patient versus medical student—on model performance, responses to patients consistently scored higher than those of medical students (Table 2). This difference reached statistical significance in the best fine-tune and role-play+database models. However, no significant differences were observed with the best fine-tune+database, role-play+book, and best fine-tune+book models, suggesting that these adjusted models can answer both general patient questions and more specialized queries from medical students.

Table 2. Subgroup analysis of the performance of EyeGPT by role-play character.

EyeGPT model	Patients, mean (SD)[a]	Medical students, mean (SD)[a]	*P* value
Best fine-tune[b]	13.45 (5.79)	10.99 (6.32)	.03
Role-play+database[c]	14.67 (5.07)	11.62 (6.55)	.03
Best fine-tune+database[d]	14.52 (4.99)	12.84 (5.34)	.08
Role-play+book[e]	14.85 (4.78)	12.65 (5.85)	.06
Best fine-tune+book[f]	14.44 (6.13)	13.38 (5.83)	.07

[a]Overall response score (the sum of 4 rating dimensions, with a maximum score of 20 representing the best performance).
[b]The fine-tuned model with 10,000 iterations.
[c]Role-play plus manual database retrieval.
[d]The best fine-tuned model plus manual database retrieval.
[e]Role-play plus book retrieval.
[f]The best fine-tuned model plus book retrieval.

Different Domains

In the subgroup analysis of EyeGPT's performance across different domains, there were no statistically significant differences in the scores of disease description, risk factors, diagnosis, treatment and prevention, and prognosis across all models (Table 3).

XSL•FO
RenderX

Table 3. Subgroup analysis of the performance of EyeGPT by domain.

EyeGPT model	Disease description, mean (SD)[a]	Risk factors, mean (SD)[a]	Diagnosis, mean (SD)[a]	Treatment and prevention, mean (SD)[a]	Prognosis, mean (SD)[a]	*P* value
Best fine-tune[b]	12.92 (6.51)	12.67 (5.38)	11.81 (6.45)	13.21 (5.12)	9.90 (6.60)	.35
Role-play+database[c]	12.98 (6.57)	12.73 (6.00)	13.73 (5.38)	13.53 (4.55)	11.48 (6.26)	.49
Best fine-tune+database[d]	15.14 (4.76)	12.08 (5.64)	14.70 (2.98)	12.78 (4.82)	11.17 (6.10)	.06
Role-play+book[e]	12.19 (6.15)	13.78 (5.61)	13.91 (5.92)	14.60 (3.31)	13.00 (5.18)	.80
Best fine-tune+book[f]	11.70 (7.46)	13.33 (6.53)	15.15 (4.64)	13.64 (5.79)	12.17 (6.14)	.36

[a]Overall response score (the sum of 4 rating dimensions, with a maximum score of 20 representing the best performance).

[b]The fine-tuned model with 10,000 iterations.

[c]Role-play plus manual database retrieval.

[d]The best fine-tuned model plus manual database retrieval.

[e]Role-play plus book retrieval.

[f]The best fine-tuned model plus book retrieval.

Performance Comparison: AI Model Versus Human Ophthalmologists

In the human-machine best-ranked comparison, EyeGPT showed competitive capabilities, particularly in understandability and empathy. With the assistance of EyeGPT, human ophthalmologists' performance was notably improved. Figure 4 summarizes the frequencies of answers generated by EyeGPT, unassisted ophthalmologists, or EyeGPT-assisted ophthalmologists, ranked as the best among the 3 candidate answers across 4 dimensions. Regarding understandability and empathy, the EyeGPT answers ranked best for 23 (19.2%) and 41 (34.2%) of the 120 questions, respectively, which were higher than those of ophthalmologists, which ranked best for 12 (10%) and 8 (6.7%) of the 120 questions. The answers provided by EyeGPT-assisted ophthalmologists were most frequently ranked as the best, at 85 (85/120, 70.8%) and 71 (71/120, 59.2%) for understandability and empathy, respectively; however, the accuracy and trustworthiness of EyeGPT answers were slightly lower than those by the ophthalmologists (accuracy: 12/120, 10% vs 14/120, 11.7%; trustworthiness: 12/120, 10% vs 15/120, 12.5%), highlighting areas for improvement. With the assistance of EyeGPT, the answers provided by the ophthalmologists excelled, ranking highest in accuracy and trustworthiness in 94 (78.3%) and 93 (77.5%) of the 120 questions, respectively. For illustrative examples of the best-ranked comparison, see Multimedia Appendix 11.

Figure 4. Percentage of answers ranked best by EyeGPT answers and ophthalmologists' answers. EyeGPT(best): best fine-tune+book model.

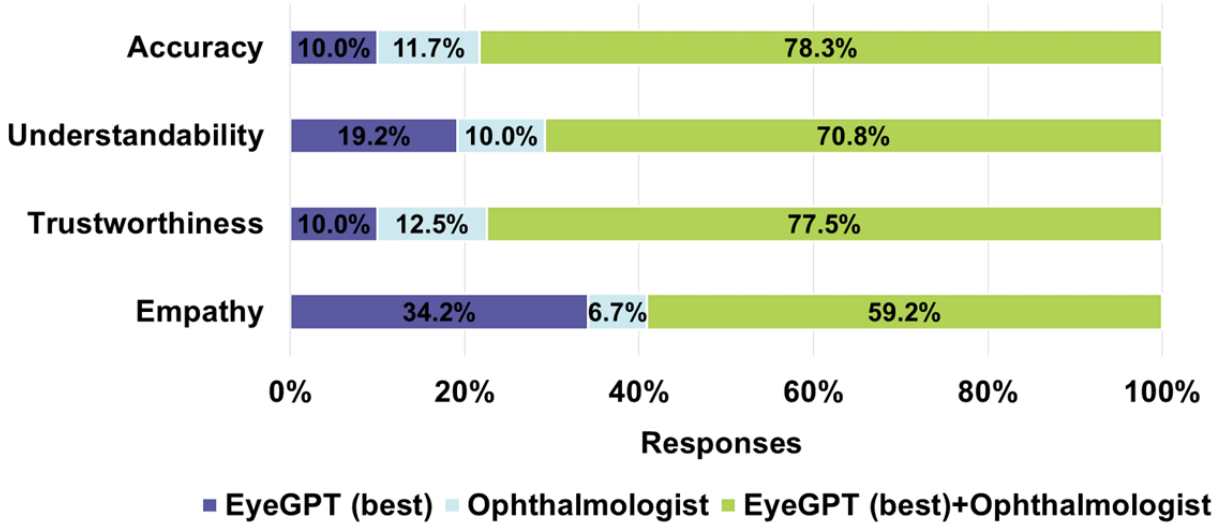

Error Analysis: Areas for Improvement

The results of the error analysis are shown in Multimedia Appendix 12. Rater 1 identified 5 (5/120, 4.2%) QA pairs as containing unrelated information, 35 (35/120, 29.2%) as containing apparent factual errors, 23 (23/120, 19.2%) as having incomplete information, and 6 (6/120, 5%) exhibiting faulty logic. Rater 2 found 6 (6/120, 5%) QA pairs with unrelated information, 30 (30/120, 25%) with factual errors, 22 (22/120,

XSL•FO
RenderX

18.3%) with incomplete information, and 4 (4/120, 3.3%) demonstrating faulty logic. The inter-rater reliability, assessed using kappa values, was 0.905, 0.895, 0.699, and 0.792, respectively.

Discussion

Principal Findings

In this study, we integrated specialized ophthalmic knowledge into a general LLM using role-play, fine-tuning, and RAG methods, resulting in the development of EyeGPT for ophthalmology. In terms of accuracy, understandability, trustworthiness, and empathy, all fine-tuned models showcased remarkable improvements compared with the original model. Among them, the best fine-tune model exhibited the highest efficacy. Among the RAG strategies, the best fine-tune+book model emerged as the most capable. Subgroup analysis revealed that EyeGPT performed well in the category of common diseases and showed consistent performance across different users and domains. EyeGPT demonstrated competitive capabilities in understandability and empathy when compared with a human ophthalmologist. With the assistance of EyeGPT, the performance of the ophthalmologists was notably enhanced.

Comparison With Prior Work

LLMs in health care raise concerns about inaccurate recommendations and fabricated information (hallucinations), which could lead to severe consequences. Previous studies have assessed the QA capabilities of existing LLMs in ophthalmology [11-13,15], highlighting the significance of augmenting LLMs with ophthalmic expertise. Our study achieved this by using 3 optimization methods: role-play, fine-tuning, and RAG. Role-playing helped position EyeGPT as an ophthalmologist, resulting in more professional responses, evidenced by the significantly increased accuracy, understandability, and trustworthiness. By setting the input role as the patient or student, the LLM's response tended to be more compassionate and preaching, as reflected in the higher empathy scores than those with the original model. "To Cure Sometimes, To Relieve Often, To Comfort Always" is a well-known saying in medicine reminding us that providing care involves not only treating ailments but also offering relief and comfort to patients. Similarly, AI models should also embody empathy when assisting users, underscoring the importance of role-playing in developing medical AI assistants. Fine-tuning with publicly available real-world patient-doctor interactions further enhanced EyeGPT's knowledge and performance. In addition, we observed that the reduction in evaluation loss with the validation set was consistent with the improvement in the model's performance as evaluated by ophthalmologists. RAG is another way to make the LLM knowledgeable, particularly to reduce hallucinations. In previous studies, Zakka et al [10] developed Almanac, an LLM framework augmented with retrieval capabilities from curated medical resources for medical guidelines and treatment recommendations. In ophthalmology, Singer et al [41] used verified ophthalmology textbooks as source material, providing citations to address the trustworthiness and accuracy gaps in LLM responses to Ophthalmic Knowledge Assessment Program style queries. In this study, hallucination mitigation was also observed in the model enhanced by the manual database or books, reducing it by over 3.4% compared with the best fine-tune model. Among them, the best fine-tune+book model demonstrated the highest proportion of mitigating hallucinations and outperformed best fine-tune+database in all 4 aspects, which could potentially be attributed to the fact that the books surpassed the self-manual database regarding content richness and reference value.

Interestingly, we found no significant difference in performance between RAG and fine-tuned models. Fine-tuning is a popular approach but has limitations. One limitation is its dependence on specific formats of medical dialogue data, which are scarce and require validation and curation by medical professionals [2]. RAG overcomes these issues by directly leveraging authoritative external resources like textbooks, medical literature, or professional websites [42]. However, it is important to note that these optimization methods are not mutually exclusive. Our results demonstrated the combined effectiveness of fine-tuning and RAG, with the best-performing EyeGPT model obtained through integration. Furthermore, the data used for fine-tuning are publicly available and reliable, and the enhanced ophthalmic books are also openly accessible, rendering these strategies valuable references for future specific LLMs.

The health care environment is complex; therefore, it is essential to assess the performance of health care AI models in different scenarios [28]. Current research has primarily focused on evaluation for general questions [15], with limited studies on specific and rare diseases. Our study validated EyeGPT by analyzing its performance across various disease categories, demonstrating strong performance in common diseases but indicating room for improvement in special and rare diseases. Future improvements can be achieved by using high-quality data sets, specialized external knowledge resources, and exploring low-shot or few-shot learning. Additionally, we found that solely fine-tuned and RAG models were less informative for specialized medical student inquiries than simpler patient inquiries. The best-performing EyeGPT performed equally well for patient and student inquiries, suggesting that combining fine-tuning and RAG enhances LLM's expertise in meeting the needs of both groups. Importantly, our evaluation set covers a wide range of question categories, from common diseases to rare diseases, and user roles encompassing patients and medical students, including disease descriptions, examinations, treatments, and more. By establishing multiple evaluation dimensions, including accuracy, understandability, trustworthiness, empathy, and hallucination, we aimed to provide a comprehensive reference framework for future ophthalmic specialized models.

Despite a growing global ophthalmologist workforce, limited-resource countries face a severe shortage of specialists [43]. EyeGPT has the potential to address this gap. Although its accuracy and trustworthiness are lower than those of human ophthalmologists, our findings show competitive capabilities in terms of understandability and empathy. This finding aligns with another study demonstrating the potential advantages of LLMs at enhancing efficiency and empathy in outpatient environments [44]. We attribute this to EyeGPT's ability to patiently process large amounts of information and initiate and

XSL•FO
RenderX

conclude conversations with consistent courtesy, unaffected by fatigue, emotions, or other factors. Although this may not be genuine in the human sense in high-demand scenarios, EyeGPT received higher empathy scores than human doctors who may, at times, provide brief responses or use complex medical jargon due to their level of medical knowledge, potentially leading to issues with poor understandability and empathy. However, the LLMs' simplified expressions may overlook certain nuanced yet crucial medical information, leading to decreased accuracy and trustworthiness. The error analysis revealed that the main gap lies in factual inaccuracies and incomplete responses, highlighting the need to integrate more ophthalmic knowledge into the model and combine it with the professional expertise and experience of human doctors for comprehensive decision-making. Although LLMs cannot replace human professionals, they could serve as an auxiliary tool to enhance physicians' performance. In our ideal scenario, EyeGPT acts as a continuous, personalized assistant, providing guidance and clarification to patients throughout their care journey, without relying on physical queues or multiple face-to-face interactions with health care personnel. Additionally, EyeGPT can serve as an educational tool for medical students seeking immediate clarification on complex subjects. For example, EyeGPT may help primary doctors improve their decision-making ability and reduce diagnosis time.

Limitations

Our study has several limitations. First, the current version of the model focuses on augmenting ophthalmic knowledge at the textual level. Future iterations should prioritize enhancing the model's image interpretation capabilities [7,39,45], crucial for ophthalmology given its heavy reliance on multimodal imaging. Second, assessing the appropriateness of medical advice may be subjective and biased by grader opinion. More efforts could be achieved in the future, for example, by incorporating a broader spectrum of ophthalmic data and real-world feedback from users including medical students and patients. Last, a more secure application at this stage is using LLMs to assist physicians in their face-to-face consultations. This pilot study has initially validated its potential, and forthcoming research should aim to disseminate the findings more widely among the population.

Conclusions

In conclusion, through role-playing, fine-tuning, and RAG, EyeGPT can potentially improve accuracy and efficiency in patient consultation and medical education. It may also be expected to increase access to high-quality medical consultations, especially for patients in underprivileged regions. We hope our study can make a good contribution to the current literature on ophthalmic AI assistants to provide an effective tool for enhancing health care.

Acknowledgments

The study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (P0048623) from Hong Kong Special Administrative Region (HKSAR), the Global STEM Professorship Scheme (P0046113), and Henry G Leong Endowed Professorship in Elderly Vision Health. The sponsors or funding organizations had no role in the design or conduct of this research.

We thank the InnoHK HKSAR Government for providing valuable supports.

Data Availability

The data sets generated during this study are available in the figshare repository at [46]. EyeGPT is available for use via [47].

Authors' Contributions

DS and XC contributed to conceptualization. DS and WZ developed the methodology. DS, XC, ZZ, PX, YW, MX, LG, and YL were responsible for data curation, formal analysis of the data, and validation. XC, ZZ, and PX wrote the original draft of the manuscript. All authors commented on the manuscript and approved the current version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Public datasets used in fine-tuning EyeGPT.
[PDF File (Adobe PDF File), 93 KB - jmir_v26i1e60063_app1.pdf]

Multimedia Appendix 2
The specific list of keywords used in the public datasets filtering process.
[PDF File (Adobe PDF File), 82 KB - jmir_v26i1e60063_app2.pdf]

Multimedia Appendix 3
The specific list of textbooks used in knowledge enhancement.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 119 KB - jmir_v26i1e60063_app3.pdf]

Multimedia Appendix 4
Sample of our manual database.
[PDF File (Adobe PDF File), 427 KB - jmir_v26i1e60063_app4.pdf]

Multimedia Appendix 5
Specific diseases of question lists.
[PDF File (Adobe PDF File), 97 KB - jmir_v26i1e60063_app5.pdf]

Multimedia Appendix 6 [PDF File (Adobe PDF File), 109 KB - jmir_v26i1e60063_app6.pdf]

Multimedia Appendix 7
Tensorboard training logs of Finetune 3.
[PDF File (Adobe PDF File), 118 KB - jmir_v26i1e60063_app7.pdf]

Multimedia Appendix 8
Statistical analysis of independent evaluations of 120 questions on the test set along four dimensions.
[PDF File (Adobe PDF File), 129 KB - jmir_v26i1e60063_app8.pdf]

Multimedia Appendix 9
Inter-rater reliability analysis of 120 questions on the test set along four dimensions in independent evaluation.
[PDF File (Adobe PDF File), 102 KB - jmir_v26i1e60063_app9.pdf]

Multimedia Appendix 10
Examples of generated answers with difference grating in independent evaluation.
[PDF File (Adobe PDF File), 152 KB - jmir_v26i1e60063_app10.pdf]

Multimedia Appendix 11
Examples in best-ranked comparison.
[PDF File (Adobe PDF File), 131 KB - jmir_v26i1e60063_app11.pdf]

Multimedia Appendix 12
Error analysis of the EyeGPT.
[PDF File (Adobe PDF File), 91 KB - jmir_v26i1e60063_app12.pdf]

References

1. Burton MJ, Ramke J, Marques AP, Bourne RRA, Congdon N, Jones I, et al. The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. Lancet Glob Health 2021 Apr;9(4):e489-e551 [FREE Full text] [doi: 10.1016/S2214-109X(20)30488-5] [Medline: 33607016]
2. Betzler BK, Chen H, Cheng C, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. The Lancet Digital Health 2023 Dec;5(12):e917-e924. [doi: 10.1016/s2589-7500(23)00201-7]
3. Liu Y, Swearingen R. Diabetic eye screening: knowledge and perspectives from providers and patients. Curr Diab Rep 2017 Aug 31;17(10):94 [FREE Full text] [doi: 10.1007/s11892-017-0911-2] [Medline: 28856510]
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
6. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. J Med Internet Res 2024 Jan 22;26:e51926 [FREE Full text] [doi: 10.2196/51926] [Medline: 38252483]
7. Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, Shi D, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit Med 2024 May 03;7(1):111 [FREE Full text] [doi: 10.1038/s41746-024-01101-z] [Medline: 38702471]

XSL•FO
RenderX

8. Caranfa JT, Bommakanti NK, Young BK, Zhao PY. Accuracy of vitreoretinal disease information from an artificial intelligence chatbot. JAMA Ophthalmol 2023 Sep 01;141(9):906-907. [doi: 10.1001/jamaophthalmol.2023.3314] [Medline: 37535363]
9. Xu P, Chen X, Zhao Z, Shi D. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. Br J Ophthalmol 2024 Sep 20;108(10):1384-1389. [doi: 10.1136/bjo-2023-325054] [Medline: 38789133]
10. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 2024 Feb 25;1(2):1 [FREE Full text] [doi: 10.1056/aioa2300068] [Medline: 38343631]
11. Lin JC, Younessi DN, Kurapati SS, Tang OY, Scott IU. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. Eye (Lond) 2023 Dec 08;37(17):3694-3695. [doi: 10.1038/s41433-023-02564-2] [Medline: 37156862]
12. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Jun 01;141(6):589-597 [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.1144] [Medline: 37103928]
13. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023 Oct;254:141-149. [doi: 10.1016/j.ajo.2023.05.024] [Medline: 37339728]
14. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 02;6(10):e2336997 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
15. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience 2023 Nov 17;26(11):108163 [FREE Full text] [doi: 10.1016/j.isci.2023.108163] [Medline: 37915603]
16. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024 Nov 01;80(5):1158-1168. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
17. Civettini I, Zappaterra A, Granelli BM, Rindone G, Aroldi A, Bonfanti S, et al. Evaluating the performance of large language models in haematopoietic stem cell transplantation decision-making. Br J Haematol 2024 Apr 09;204(4):1523-1528 [FREE Full text] [doi: 10.1111/bjh.19200] [Medline: 38070128]
18. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024 Mar 06;15(1):2050 [FREE Full text] [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
19. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y. Llama 2: Open foundation and fine-tuned chat models. arXiv Preprint posted online on July 19. [doi: 10.48550/arXiv.2307.09288]
20. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C. tatsu-lab / stanford_alpaca. GitHub. URL: https://github.com/tatsu-lab/stanford_alpaca [accessed 2024-11-26]
21. Kong A, Zhao S, Chen H, Li Q, Qin Y, Sun R, et al. Better Zero-Shot Reasoning with Role-Play Prompting. 2024 Presented at: Annual Conference of the North American Chapter of the Association for Computational Linguistics; June 16–21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.228]
22. Han T, Adams L, Papaioannou J, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca -- an open-source collection of medical conversational AI models and training data. arXiv Preprint posted online on October 4.
23. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895 [FREE Full text] [doi: 10.7759/cureus.40895] [Medline: 37492832]
24. Pal A, Umapathi L, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. Proceedings of Machine Learning Research 2022;174:248-260 [FREE Full text]
25. Jin D, Pan E, Oufattole N, Weng W, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences 2021 Jul 12;11(14):6421. [doi: 10.3390/app11146421]
26. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 2022:35-44 [FREE Full text]
27. Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. 2021 Presented at: The Ninth International Conference on Learning Representations; May 3-7, 2021; Virtual meeting.
28. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models in medical applications: a survey. arXiv Preprint posted online on May 13.
29. Basic and Clinical Science Course: 2014-2015. San Francisco, CA: American Academy of Ophthalmology; 2014.
30. Denniston A, Murray P. Oxford handbook of ophthalmology, third edition. Oxford, United Kingdom: Oxford University Press; 2014.
31. Ryan S, Sadda S, Hinton D, Schachat A, Sadda S, Wilkinson C, et al, editors. Retina. Philadelphia, PA: Saunders Elsevier; 2013.

XSL•FO
RenderX

32. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019 Presented at: Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing; November 3–7, 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1410]
33. RecursiveCharacterTextSplitter. URL: https://apipythonlangchaincom/en/latest/character/langchain_text_splitterscharacterRecursiveCharacterTextSplitterhtml [accessed 2024-12-02]
34. Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al. The faiss library. arXiv Preprint posted online on Sep 06.
35. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82 [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
36. Lim DSW, Makmur A, Zhu L, Zhang W, Cheng AJL, Sia DSY, et al. Improved productivity using deep learning-assisted reporting for lumbar spine MRI. Radiology 2022 Oct;305(1):160-166. [doi: 10.1148/radiol.220076] [Medline: 35699577]
37. Hu B, Shi Z, Lu L, Miao Z, Wang H, Zhou Z, China Aneurysm AI Project Group. A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study. Lancet Digit Health 2024 Apr;6(4):e261-e271 [FREE Full text] [doi: 10.1016/S2589-7500(23)00268-6] [Medline: 38519154]
38. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. NEJM AI 2024 Feb 22;1(3):1. [doi: 10.1056/aioa2300138]
39. Chen X, Xu P, Li Y, Zhang W, Song F, He M, et al. ChatFFA: An ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. iScience 2024 Jul 19;27(7):110021 [FREE Full text] [doi: 10.1016/j.isci.2024.110021] [Medline: 39055931]
40. Mandrekar JN. Measures of interrater agreement. Journal of Thoracic Oncology 2011 Jan;6(1):6-7. [doi: 10.1097/jto.0b013e318200f983]
41. Singer MB, Fu JJ, Chow J, Teng CC. Development and evaluation of Aeyeconsult: a novel ophthalmology chatbot leveraging verified textbook knowledge and GPT-4. J Surg Educ 2024 Mar;81(3):438-443. [doi: 10.1016/j.jsurg.2023.11.019] [Medline: 38135548]
42. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: A survey. arXiv Preprint posted online on Mar 27. [doi: 10.2139/SSRN.4895062]
43. Resnikoff S, Lansingh VC, Washburn L, Felch W, Gauthier T, Taylor HR, et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? Br J Ophthalmol 2020 Apr;104(4):588-592 [FREE Full text] [doi: 10.1136/bjophthalmol-2019-314336] [Medline: 31266774]
44. Wan P, Huang Z, Tang W, Nie Y, Pei D, Deng S, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nat Med 2024 Oct;30(10):2878-2885. [doi: 10.1038/s41591-024-03148-7] [Medline: 39009780]
45. Chen X, Zhang W, Zhao Z, Xu P, Zheng Y, Shi D, et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. Br J Ophthalmol 2024 Sep 20;108(10):1450-1456. [doi: 10.1136/bjo-2023-324446] [Medline: 38508675]
46. EyeQA: Evaluation Dataset for Ophthalmic Assistant with Large Language Models. Figshare. URL: https://figshare.com/s/cb0525c72ef467d2d809 [accessed 2024-11-26]
47. EyeGPT. Hugging Face. URL: https://huggingface.co/spaces/spaceis42/EyeGPT [accessed 2024-11-26]

Abbreviations

AI: artificial intelligence
FAISS: Facebook AI Similarity Search
LLM: large language model
LoRA: low-rank adaptation
QA: question-answering
RAG: retrieval-augmented generation
USMLE: United States Medical Licensing Examination

XSL•FO
RenderX

Edited by G Eysenbach, A Coristine; submitted 01.05.24; peer-reviewed by K Jin, ML Chee, W Rojas-Carabali, A Saxena, R Scherer; comments to author 22.08.24; revised version received 01.10.24; accepted 01.11.24; published 11.12.24.

Please cite as:
Chen X, Zhao Z, Zhang W, Xu P, Wu Y, Xu M, Gao L, Li Y, Shang X, Shi D, He M
EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model
J Med Internet Res 2024;26:e60063
URL: https://www.jmir.org/2024/1/e60063
doi:10.2196/60063
PMID:39661433

©Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Yue Wu, Mingpu Xu, Le Gao, Yinwen Li, Xianwen Shang, Danli Shi, Mingguang He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Accuracy of Machine Learning in Detecting Pediatric Epileptic Seizures: Systematic Review and Meta-Analysis

Zhuan Zou[1,2], MD; Bin Chen[1,2], MD; Dongqiong Xiao[1,2], PhD; Fajuan Tang[1,2], MD; Xihong Li[1,2], PhD

[1]Department of Emergency, West China Second University Hospital, Sichuan University, Chengdu, China

[2]Key Laboratory of Birth Defects and Related Diseases of Women and Children, Ministry of Education, Sichuan University, Chengdu, China

Corresponding Author:
Xihong Li, PhD
Department of Emergency
West China Second University Hospital
Sichuan University
No 20, Section 3, Renmin South Road
Wuhou District
Chengdu, 610000
China
Phone: 86 13551089846
Email: lixihonghxey@163.com

Abstract

Background: Real-time monitoring of pediatric epileptic seizures poses a significant challenge in clinical practice. In recent years, machine learning (ML) has attracted substantial attention from researchers for diagnosing and treating neurological diseases, leading to its application for detecting pediatric epileptic seizures. However, systematic evidence substantiating its feasibility remains limited.

Objective: This systematic review aimed to consolidate the existing evidence regarding the effectiveness of ML in monitoring pediatric epileptic seizures with an effort to provide an evidence-based foundation for the development and enhancement of intelligent tools in the future.

Methods: We conducted a systematic search of the PubMed, Cochrane, Embase, and Web of Science databases for original studies focused on the detection of pediatric epileptic seizures using ML, with a cutoff date of August 27, 2023. The risk of bias in eligible studies was assessed using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies–2). Meta-analyses were performed to evaluate the C-index and the diagnostic 4-grid table, using a bivariate mixed-effects model for the latter. We also examined publication bias for the C-index by using funnel plots and the Egger test.

Results: This systematic review included 28 original studies, with 15 studies on ML and 13 on deep learning (DL). All these models were based on electroencephalography data of children. The pooled C-index, sensitivity, specificity, and accuracy of ML in the training set were 0.76 (95% CI 0.69-0.82), 0.77 (95% CI 0.73-0.80), 0.74 (95% CI 0.70-0.77), and 0.75 (95% CI 0.72-0.77), respectively. In the validation set, the pooled C-index, sensitivity, specificity, and accuracy of ML were 0.73 (95% CI 0.67-0.79), 0.88 (95% CI 0.83-0.91), 0.83 (95% CI 0.71-0.90), and 0.78 (95% CI 0.73-0.82), respectively. Meanwhile, the pooled C-index of DL in the validation set was 0.91 (95% CI 0.88-0.94), with sensitivity, specificity, and accuracy being 0.89 (95% CI 0.85-0.91), 0.91 (95% CI 0.88-0.93), and 0.89 (95% CI 0.86-0.92), respectively.

Conclusions: Our systematic review demonstrates promising accuracy of artificial intelligence methods in epilepsy detection. DL appears to offer higher detection accuracy than ML. These findings support the development of DL-based early-warning tools in future research.

Trial Registration: PROSPERO CRD42023467260; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023467260

(J Med Internet Res 2024;26:e55986) doi:10.2196/55986

KEYWORDS

epileptic seizures; machine learning; deep learning; electroencephalogram; EEG; children; pediatrics; epilepsy; detection

XSL•FO
RenderX

Introduction

Epilepsy is defined as a transient sign or symptom arising from abnormal, excessive, or synchronous neuronal activity in the brain [1]. It is the second most common neurological disorder [2]. Over 65 million people worldwide suffer from this mental disorder, equating to 1 in every 26 individuals [3]. In pediatric intensive care units, the prevalence of epileptic seizures among all hospitalized children is estimated to be 0.8% [4]. Hence, research into the prediction of epileptic seizures is particularly imperative.

Electroencephalography (EEG) has been established as an electrographic recording technique of brain activity, capable of timely predicting the occurrence of epileptic seizures from scalp EEG signals. This allows for more proactive and effective intervention for patients, making it an effective tool for the evaluation and diagnosis of epilepsy [5]. EEG is currently the gold standard for diagnosing neonatal epilepsy [6]. However, the interpretation of EEG primarily relies on the clinician's previous experience, which can impact the interpretation of some critical signals. Therefore, monitoring epilepsy, especially in real time, remains a challenging task [7].

In the current epileptic seizure research, there is a lack of in-depth understanding of the mechanisms of epileptic seizures and accurate predictive models. Current research is still exploring the biological and neurological mechanisms of epileptic seizures, and there is a lack of models that can accurately predict and assess individual risk. In this study, most of the data came from publicly available datasets, including but not limited to the Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) [8] and the University of Helsinki [9], which used the international 10-20 system configuration for the 19-channel EEG systems with a sampling rate of 256 Hz. With the rapid development of computer systems and ongoing advancements in statistical theories, artificial intelligence (AI) has gradually demonstrated a significant applicative value in clinical practice. This is particularly evident in the diagnosis and risk stratification of prognosis for some refractory diseases. In recent years, deep learning (DL) with its advantage of automatic feature extraction from images, has been extensively applied in image processing to assist in the auxiliary diagnosis of a wide variety of diseases.

For instance, Wei et al [10] conducted a study using machine learning (ML) to predict depression and anxiety of epilepsy patients in China, and Yossofzai et al [11] developed and validated an ML model to predict the outcome of epilepsy surgery in children. ML encompasses conventional ML and DL methods. In the past decade, DL has been used as a promising alternative to traditional ML and has been widely applied in various research fields. Truong et al [12] applied convolutional neural networks (CNNs) to different EEG datasets and demonstrated the effectiveness of DL. Daoud and Bayoumi [13] used cellular neural networks to extract meaningful features and then used recursive neural networks to classify them. Ozcan and Erturk [14] constructed 3D patterns based on the electrode positions and applied an image-based 3D CNN to predict epileptic seizures. Using directed transfer functions to explore the special information exchange between brain electrical channels, and then using cellular neural networks to predict epilepsy seizures, achieving satisfactory performance [15]. Yang et al [16] proposed a dual self-attention residual network to classify the short-time Fourier transform features of brain electrical signals.

In this context, multiple studies have also attempted to construct ML models for forecasting epileptic seizures based on different modeling variables, even DL models based on EEG to differentiate epileptic seizures. To date, both ML and DL lack a systematic understanding of their predictive accuracy. As such, this study aims to discern the accuracy of ML methods, including classical ML and DL models, in detecting seizures in children, and provide evidence-based recommendations for the development of AI in this field.

Methods

Study Registration

Our systematic review was implemented following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [17] in Multimedia Appendix 1 and prospectively registered with PROSPERO (ID CRD42023467260).

Eligibility Criteria

We developed detailed inclusion and exclusion criteria for our systematic review based on population, modeling, study type, language, and outcome measures, as detailed in Textbox 1.

XSL•FO
RenderX

Textbox 1. The inclusion and exclusion criteria.

Inclusion criteria

- Population
  - Minor
- Modeling
  - Comprehensive construction of machine learning (ML) models for predicting epileptic seizures, including both traditional ML and deep learning approaches
- Study type
  - Case-control, cohort, nested case-control, cross-sectional studies
- Language
  - Studies published in English
- Outcome measures
  - Receiver operating characteristic, C-index, sensitivity, specificity, accuracy, recall, precision, confusion matrix,F_1-score

Exclusion criteria

- Population
  - Studies that did not strictly distinguish between adults and minors
- Modeling
  - Studies focused solely on risk factors for epileptic seizures;
  - Studies that only assessed the predictive accuracy of single factors for seizures;
  - Studies focused solely on seizure image segmentation
- Study type
  - Reviews, guidelines, expert opinions, and non–peer-reviewed conference abstracts
- Language
  - Studies published in non-English languages
- Outcome measures
  - Studies missing any outcome measures assessing ML accuracy

Data Sources and Search Strategy

PubMed, Cochrane, Embase, and Web of Science databases were thoroughly retrieved as of August 27, 2023. Both MeSH (Medical Subject Headings) terms and free-text keywords were used, without restrictions on publication region or year. Details of the search strategy are available in Table S1 in Multimedia Appendix 2.

Study Selection and Data Extraction

All identified articles were imported into EndNote software (Clarivate). After deleting duplicates, titles and abstracts were screened to rule out irrelevant studies. The remaining articles were reviewed in full for inclusion. A spreadsheet was used to extract data, including title, first author, year of publication, type of study, patient origin, dataset source, number of epileptic seizure cases, total number of epileptic seizure cases in the training set, overfitting methods, number of cases in the validation set, missing value handling methods, variable selection, types of models used, and modeling variables.

In total, 2 independent investigators (ZZ and BC) implemented the literature screening and data extraction, followed by cross-checking. In case of any disagreements, a third investigator (DX) was consulted for resolution.

Risk of Bias Assessment

The QUADAS-2 tool (Quality Assessment of Diagnostic Accuracy Studies–2) [18] was leveraged to appraise the risk of bias and applicability of the included studies. This tool evaluates 4 aspects, such as patient selection, index test, reference standard, and flow and timing. Each domain contains specific questions answered as “yes,” “no,” or “uncertain,” corresponding to a bias risk of “low,” “high,” or “uncertain,” respectively. Studies were considered at low risk of bias if all key questions in each domain were answered with “yes.” Any

XSL•FO
RenderX

"no" response indicated potential bias, requiring evaluators to assess the risk level according to established guidelines. An "uncertain" rating signified insufficient information for a definitive judgment.

Outcomes

This systematic review assessed diagnostic accuracy through sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, diagnostic odds ratio, and the summary receiver operating characteristic curve. Calculating these estimates required diagnostic 4-fold tables. For ML, given the risk of overfitting, the diagnostic 4-fold tables in both training and validation sets were extracted from studies on ML. Funnel plots were used to analyze the publication bias of the C-index and then Egger test was performed for statistical evaluation of the publication bias. For DL, only the results of validation or test sets were considered.

Statistical Analysis

Data analysis was executed using Stata 15.0 (StataCorp LLC). A bivariate mixed-effects model was used for the meta-analysis of sensitivity and specificity. However, if original studies lacked diagnostic 4-fold tables, they were calculated using two methods: (1) one based on sensitivity, specificity, precision, and case numbers, and (2) the other based on sensitivity and specificity calculated using the optimal Youden index and case numbers. Pooled estimates of sensitivity, specificity, likelihood ratios, diagnostic odds ratio, and their 95% CIs were calculated. The summary receiver operating characteristic curve area was also estimated. Publication bias assessment was conducted using funnel plots. Statistical significance was determined at $P<.05$. The studies included were all based on EEG-based ML methods, and we know that traditional ML and DL have certain differences in their image-processing capabilities and modeling efficiency. Therefore, to reduce heterogeneity, we conducted a subgroup analysis by model type for traditional ML and DL.

Ethical Considerations and Consent to Participate

All analyses were based on previously published studies; thus, no ethical approval and patient consent are required.

Results

Study Selection

Our systematic search identified 15,389 related articles, of which 3130 duplicates (2975 identified by software and 155 manually) were excluded. After title and abstract screening, we excluded 12,014 additional articles. Of the remaining 44 articles, 6 unpublished conference abstracts and 2 articles without full-text access were excluded. Among the remaining 36 articles, we excluded 5 for missing outcome measures, 1 for data overlap, and 2 for including adult participants. After these exclusions, 28 studies met the eligibility criteria [19-46]. Figure 1 illustrates the literature screening process.

Figure 1. Literature selection flow chart.

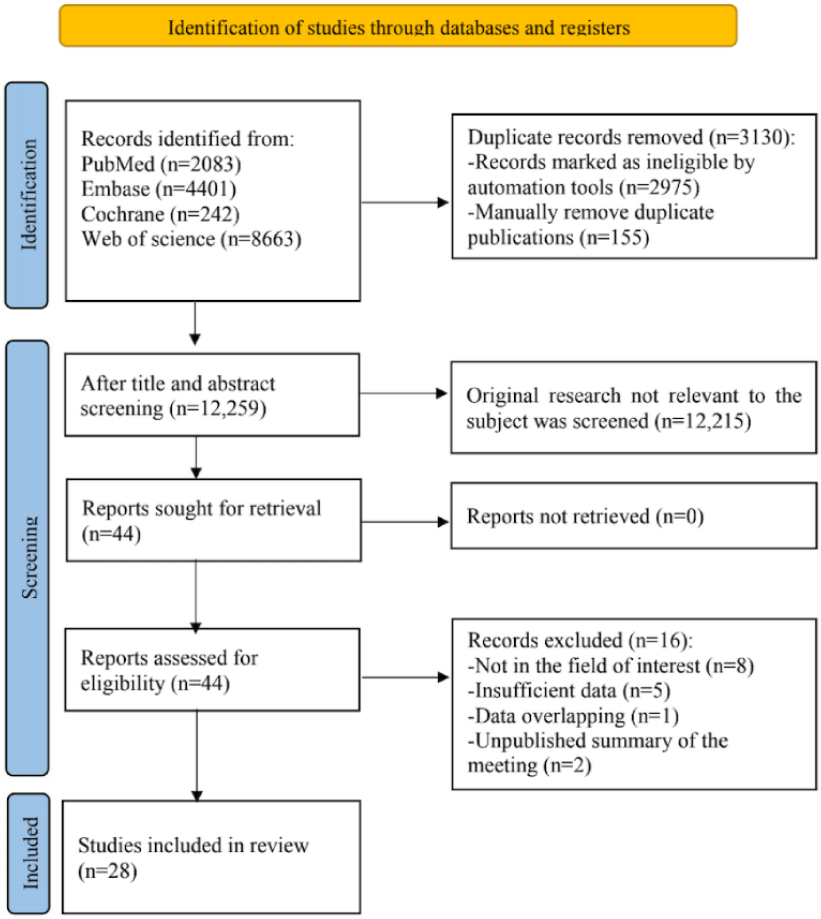

XSL•FO
RenderX

Study Characteristics

Table 1 presents the detailed characteristics of the 28 included studies, with 15 on ML [32-46] and 13 on DL [19-31]. Of these, 6 were multicenter studies [23,26,27,32,34,39], while the remaining 22 were single-center investigations. Furthermore, 5 studies were prospective [37,40,41,43,44], with the remaining 23 being retrospective. A total of 6 studies incorporated prospective external validations [26,27,30,40,43,44], 2 implemented multicenter external validations [26,27], and the remaining 22 used internal validations. Among studies with internal validation, 3 used 10-fold cross-validation [25,32,34], 4 used 5-fold cross-validation [21,35,37,41], 3 used the leave-one-out method [22,32,45], 1 used the Bootstrap method [33], and 10 used random sampling [19,20,23,28,31,36,38,39,42,46]. Of the total, 1 study did not specify the generation method of validation set [24]. Across the 28 studies, 5742 participants were included, with 1814 confirmed patients with epilepsy, ranging from 9 to 1117 patients per individual study. Furthermore, 10 studies were carried out in Asian countries, while 18 were conducted in Western countries. Out of total, 8 studies used EEG signals from the CHB-MIT dataset [19,22,24,31,34-36,39], another 8 used signals from The University of Helsinki dataset [20,21,23,25,28,29,38,42], and 3 used data from Cork University Maternity Hospital [26,27,45]. In addition, 1 ML study used 3 distinct models, each with separate training and validation sets [38]. All participants in the included studies were children, with 19 studies focusing specifically on neonates [19-21,23,25-30,32,33,36,38,42-46]. Studies on DL predominantly used CNN models. Of the 13 studies on DL detection of pediatric epileptic seizures included in this paper, 7 studies mainly applied the CNN model [21-23,25,27,29,30]. The DL models also included adaptive grey wolf optimizer (AGWO) [19], graph convolutional neural network [20], convolutional gated recurrent neural network [31], cross-feature fusion stream convolutional neural network [24], and pretrained deep convolution neural networks [28], and these models are mostly evolved from the CNN. ML studies used models including random forest (RF), k-nearest neighbors (KNN), support vector machine (SVM), extreme gradient boosting, adaptive boosting, and decision tree (DT). RF is an ensemble learning algorithm used for classification and regression [47], and 1 of its primary advantages is its robustness to noise. In our study, 4 original studies using this model were included [32,33,37,39]. KNN is the simplest ML tool with supervision for classification [48], and it depends on the predefined nearest number or the K value. One of the included original studies used this model [36]. DT predicts the correct classification by recursively partitioning the real space, and it consists of 2 types of multiple nodes, that is, leaf nodes and decision nodes. Of the total, 2 included original studies used this model [33,36]. SVM is an advanced algorithm mainly used for pattern recognition and feature reduction, using discriminative techniques to classify the input, and is best suited for binary classification [49]. Up to 5 original studies on ML in this study applied this model [34,38,42,44,45]. With the exception of 1 study that used electrocardiography as a variable [35], all other studies used EEG data. Clinical characteristics of patients were included as modeling variables in 6 studies [32,37,40,41,43,46].

XSL•FO
RenderX

Table 1. The detailed characteristics of the included studies.

No.	Author	Year	Country	Study design	Patient source	Task (prediction diagnosis)	Age	Dataset source	Events - P	Sample size - P	Training set - P	Verification mode	Validation set - P	Model classification	Modeling variable	Model classification (ML[a] / DL[b])
1	Pavel et al [32]	2023	Ireland	Retrospective cohort	Multicenter	Prediction	Neonatal	NCT02160171 and NCT02431780	53	162	162	10-fold cross-validation	None	RF[c] and GB-DT[d]	Clinical features + EEG[e]	ML
2	McKee et al [33]	2023	United States	Retrospective cohort	Single center	Prediction	Neonatal	Philadelphia, United States	150	1117	1117	Bootstrap	None	LR[f], DT[g], and RF	EEG	ML
3	Azriel et al [35]	2022	Israel	Retrospective cohort	Single center	Prediction	5.1 (0.8-11.5) y	CHB-MIT[h]	19	176	111	5-fold cross-validation	55	META[i] + HRV[j] + MOR[k]	ECG[l]	ML
4	Hu et al [37]	2021	United States	Prospective cohort	Single center	Prediction	<18 y	NCT03419260	184	719	719	5-fold cross-validation	None	RF	Clinical features + EEG	ML
5a	Elakkiya [38]	2021	India	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	49	Random	13	SVM[m]	EEG	ML
5b	Elakkiya [38]	2021	India	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	37	Random	25	ANN[n]	EEG	ML
5c	Elakkiya [38]	2021	India	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	33	Random	25	CNN[o]	EEG	DL
6	Aayesha et al [39]	2020	Pakistan	Retrospective cohort	Multicenter	Diagnosis	1.5-22 y	CHB-MIT	23	23	23	Random	5	RF, SVM, KNN[p], DT, and MLP[q]	EEG	ML
7	Fung et al [40]	2020	United States	Prospective cohort	Single center	Prediction	5.5 (1.43-13.2) y	NCT03419260	252	1033	719	External verification	314	LR	Clinical features + EEG	ML
8	Fung et al [41]	2020	United States	Prospective cohort	Single center	Prediction	5.5 (1.35-12.9) y	NCT03419260	184	719	719	5-fold cross-validation	719	LR	Clinical features + EEG	ML

XSL•FO
RenderX

No.	Author	Year	Country	Study design	Patient source	Task (prediction diagnosis)	Age	Dataset source	Events - P	Sample size - P	Training set - P	Verification mode	Validation set - P	Model classification	Modeling variable	Model classification (ML[a] / DL[b])
9	Sansevere et al [43]	2020	United States	Prospective cohort	Single center	Prediction	Neonatal	Boston Children's Hospital	73	210	210	External verification	None	LR	Clinical features + EEG	ML
10	Pan et al [44]	2020	China	Prospective cohort	Single center	Diagnosis	Neonatal	Peking University First Hospital	310	526	526	External verification	None	SVM, LR, AdaBoost[r], XGBoost[s], RF, and GBDT	EEG	ML
11	Temko et al [45]	2011	Ireland	Retrospective cohort	Single center	Diagnosis	Neonatal (term)	Cork University Maternity Hospital	17	55	55	the leave-one-out method	None	SVM	EEG	ML
12	Açikoğlu and Tuncer [42]	2019	Turkey	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	79	Random	39	SVM	EEG	ML
13	Sethy et al [36]	2021	India	Retrospective cohort	Single center	Diagnosis	Neonatal	CHB-MIT	23	23	23	Random	23	DT, KNN, LDA[t], LR. NB[u], SVM	EEG	ML
14	Perez-Sanchez et al [34]	2022	Mexico	Retrospective cohort	Multicenter	Prediction	2-19 y	CHB-MIT	14	22	None	10-fold cross-validation	22	SVM and MLP	EEG	ML
15	Karayiannis et al [46]	2006	United States	Retrospective cohort	Single center	Diagnosis	Neonatal	The Clinical Research Centers for Neonatal Seizures	9	9	None	Random	9	QNNs[v] and FFNN[w]	Clinical features + EEG	ML
16	Jaishankar et al [19]	2023	India	Retrospective cohort	Single center	Prediction	Neonatal	CHB-MIT	23	23	None	Random	8	AGWO[x]	EEG	DL
17	Raeisi et al [20]	2022	Italy	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	None	None	79	GCNN[y]	EEG	DL
18	Gramacki A and Gramacki J [21]	2022	Poland	Retrospective cohort	Single center	Prediction	Neonatal	The University of Helsinki	57	79	None	5-fold cross-validation	57	CNN	EEG	DL

XSL•FO
RenderX

No.	Author	Year	Country	Study design	Patient source	Task (prediction diagnosis)	Age	Dataset source	Events - P	Sample size - P	Training set - P	Verification mode	Validation set - P	Model classification	Modeling variable	Model classification (ML[a] / DL[b])
19	Gao et al [22]	2022	China	Retrospective cohort	Single center	Prediction	3-12 y	CHB-MIT	16	16	None	the leave-one-out method	16	CNN	EEG	DL
20	Borovac et al [23]	2022	Iceland	Retrospective cohort	Multicenter	Prediction	Neonatal	The University of Helsinki	38	79	79	Random	28	CNN	EEG	DL
21	Wang et al [24]	2021	China	Retrospective cohort	Single center	Prediction	9.89 y	CHB-MIT	23	23	None	none	23	CFS-CNN[z]	EEG	DL
22	Tanveer et al [25]	2021	Pakistan	Retrospective cohort	Single center	Prediction	Neonatal	The University of Helsinki	39	79	None	10-fold cross-validation	39	CNN	EEG	DL
23	O'Shea et al [26]	2021	Ireland	Retrospective cohort	Multicenter	Prediction	Preterm infants (GA<32 w)	CUMH[aa], Parma University Hospital	23	33	17	External verification	10	DL	EEG	DL
24	Daly et al [27]	2021	Ireland	Retrospective cohort	Multicenter	Diagnosis	Neonatal	CUMH, the University of Helsinki	80	229	72	External verification	78	CNN	EEG	DL
25	Caliskan and Rencuzogullari [28]	2021	Turkey	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	None	Random	79	p-DC-NN[ab]	EEG	DL
26	Frassineti et al [29]	2020	Italy	Retrospective cohort	Single center	Diagnosis	Neonatal	The University of Helsinki	39	79	None	the leave-one-out method	39	CNN	EEG	DL
27	Ansari et al [30]	2019	Netherlands	Retrospective cohort	Single center	Diagnosis	Neonatal (term)	Sophia Children's Hospital	48	48	26	External verification	22	CNN	EEG	DL

XSL•FO
RenderX

No.	Author	Year	Country	Study design	Patient source	Task (prediction diagnosis)	Age	Dataset source	Events - P	Sample size - P	Training set - P	Verification mode	Validation set - P	Model classification	Modeling variable	Model classification (ML[a] / DL[b])
28	Affes et al [31]	2019	Tunisia	Retrospective cohort	Single center	Diagnosis	1.5-22 y	CHB-MIT	23	23	None	Random	13	CGRNN[ac]	EEG	DL

[a]ML: machine learning.
[b]DL: deep learning.
[c]RF: random forest.
[d]GBDT: gradient boosting decision tree.
[e]EEG: electroencephalography.
[f]LR: logistic regression.
[g]DT: decision tree.
[h]CHB-MIT: Children's Hospital Boston-Massachusetts Institute of Technology.
[i]META: metadata.
[j]HRV: heart rate variability.
[k]MOR: morphological features.
[l]ECG: electrocardiography.
[m]SVM: support vector machine.
[n]ANN: artificial neural network.
[o]CNN: convolutional neural networks.
[p]KNN: k-nearest neighbors.
[q]MLP: multilayer perceptron.
[r]AdaBoost: adaptive boosting.
[s]XGBoost: extreme gradient boosting.
[t]LDA: linear discriminant analysis.
[u]NB: naive Bayes.
[v]QNNs: Quantum Neural Net.
[w]FFNN: feedforward neural network.
[x]AGWO: adaptive grey wolf optimizer.
[y]GCNN: graph convolutional neural network.
[z]CFS-CNN: the cross-feature fusion stream convolutional neural network.
[aa]CUMH: Cork University Maternity Hospital.
[ab]p-DCNN: pretrained deep convolution neural networks.
[ac]CGRNN: convolutional gated recurrent neural network.

Risk of Bias in Studies

In the 28 included studies, all cases were consecutive case series, eliminating concerns of case selection bias. Notably, the type of study design had minimal influence on the results of image-based DL studies [21,26,27], leading to a low risk of bias for this category. Conversely, conventional ML methods often involve manual segmenting regions of interest segmentation and extracting texture features, potentially introducing significant bias in case-control studies. Consequently, this study was categorized as having an unclear risk of bias [35]. There was a study [32] in which we doubted its results were interpreted without the outcome of the gold standard trial of epileptic seizures, the risk of bias of the study was classified as unclear. Further details are shown in Figure 2.

XSL•FO
RenderX

Figure 2. Risk of bias (A) graph and (B) summary.

Meta-Analysis

Machine Learning

The 15 studies on ML used multi-arm predictive diagnostic experiments, generating 25 training sets and 13 independent validation sets. A total of 18 training sets and 6 validation sets provided C-index values. A random-effects model revealed pooled C-indexes of 0.76 (95% CI 0.69-0.82) and 0.73 (95% CI 0.67-0.79) in training and validation sets, respectively (Figures 3A and 3B). The funnel plots of the pooled training and pooled validation sets of ML were analyzed, respectively. The funnel plot generated from the training set suggested an uneven but symmetric distribution of the C-indexes of the included studies, and Egger test indicated a possible publication bias (P=.001 and P<.05). On the other hand, the funnel plot from the validation set showed that distribution of the C-indexes tended to be symmetrical, and Egger test suggested no significant publication bias (P=.47 and P>.05; Figures 3C and 3D).

In addition, 12 training sets and 11 validation sets provided sensitivity, specificity, and accuracy. A random-effects model was leveraged. The analysis showed a pooled sensitivity of 0.77 (95% CI 0.73-0.80) and specificity of 0.74 (95% CI 0.70-0.77) for training sets (Figure 4A). For validation sets, the pooled sensitivity was 0.88 (95% CI 0.83-0.91) and specificity was 0.83 (95% CI 0.71-0.90; Figure 4B). The accuracy for the pooled training set and validation set in ML were 0.75 (95% CI 0.72-0.77) and 0.78 (95% CI 0.73-0.82), respectively (Figure 4C). No significant publication bias was detected in these results.

XSL•FO
RenderX

Figure 3. Forest and funnel plots of machine learning models for detecting seizures in children. The presence of repeated authors in the literature arises from the development of multiple machine learning models. (A) Forest plot illustrating the C-index summarization for the training set. (B) Forest plot illustrating the C-index summarization for the validation set. (C) Funnel plot illustrating the C-index for training set. (D) Funnel plot illustrating the C-index for the validation set. AdaBoost: adaptive boosting; ANN: artificial neural network; GBDT: gradient boosting decision tree; LDA: linear discriminant analysis; LR: logistic regression; NB: naive Bayes; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

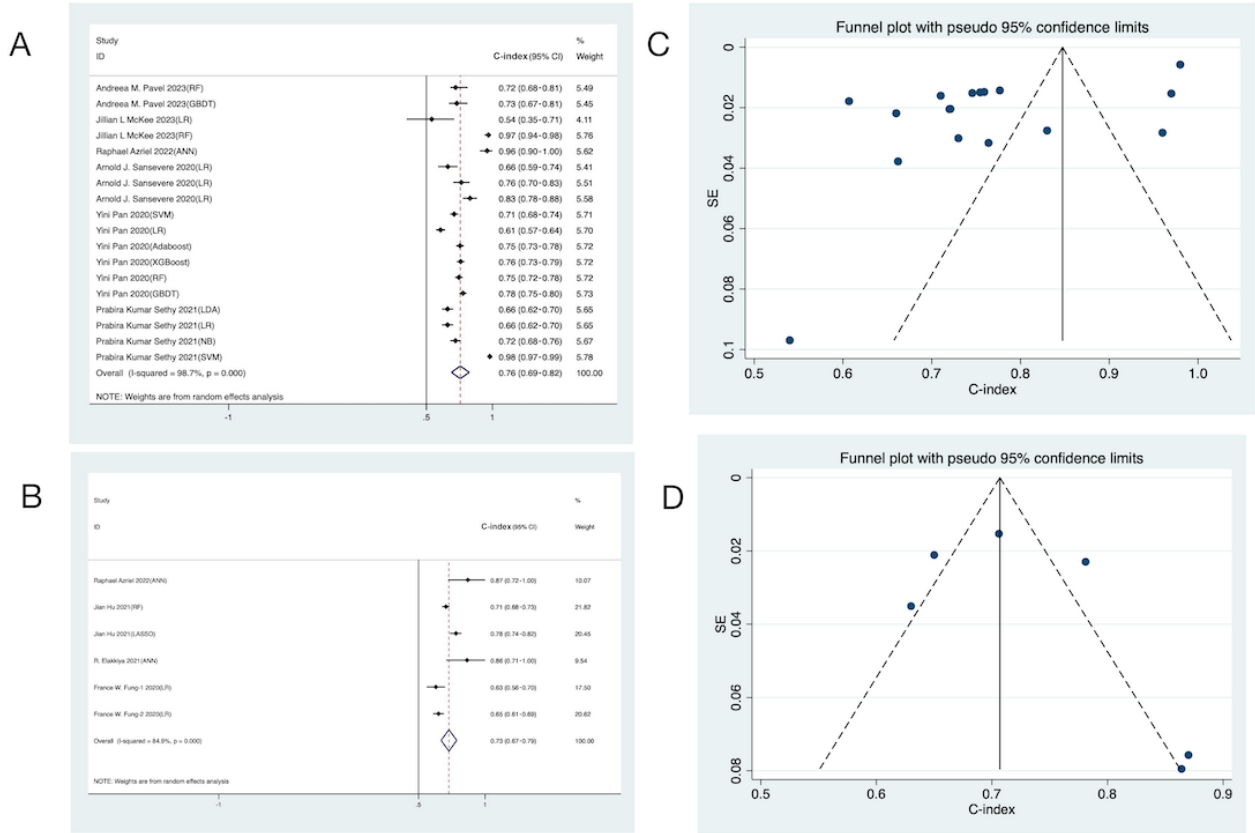

Figure 4. The forest plot shows the sensitivity, specificity, and accuracy of machine learning models in detecting seizures in children. The presence of repeated authors in the literature arises from the development of multiple machine learning models. (A) Sensitivity and specificity of the training set, (B) Sensitivity and specificity of the validation set, and (C) Accuracy of the machine learning models for both the training set and the validation set post summarization. AdaBoost: adaptive boosting; ANN: artificial neural network; DT: decision tree; GBDT: gradient boosting decision tree; KNN: k-nearest neighbors; LR: logistic regression; SVM: support vector machine; RF: random forest; XGBoost: extreme gradient boosting.

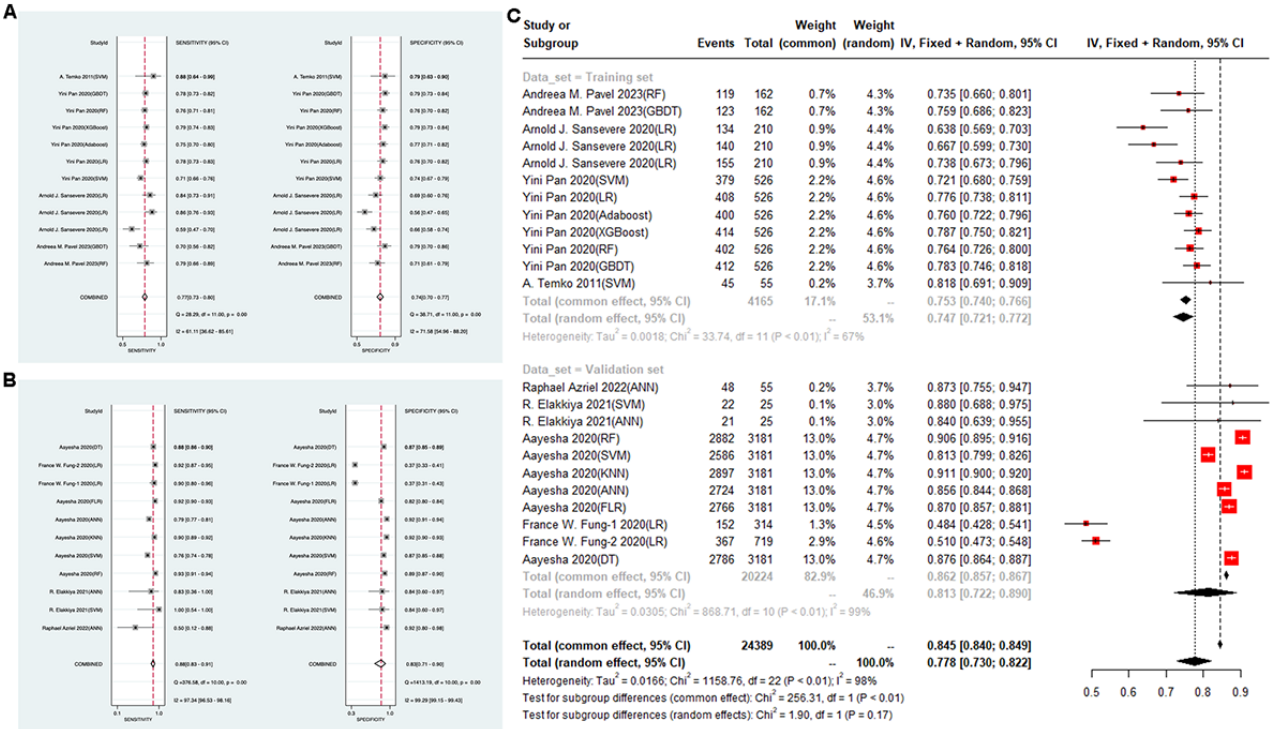

XSL•FO
RenderX

Deep Learning

In total, 13 DL studies focused on DL for diagnosing pediatric epileptic seizures. Since DL is less susceptible to overfitting, only 26 validation sets were analyzed. Among the validation sets, 4 provided only the C-index, 10 provided only sensitivity, specificity, and accuracy without the C-index value, and the remaining 11 validation sets provided both the C-index as well as sensitivity, specificity, and accuracy. Therefore, a total of 15 validation sets were included in the C-index analysis for DL, while 21 validation sets were included in the analysis of sensitivity, specificity, and accuracy for DL. Of these, 15 validation sets provided C-index values. A random-effects model revealed a pooled C-index of 0.91 (95% CI 0.88-0.94; Figure 5A). The funnel plot of the C-indexes in the DL validation set also showed a generally symmetric distribution, and Egger test suggested no significant publication bias (*P*=.75 and *P*>.05; Figure 5B).

Figure 5. Forest and funnel plots of the C-index for the validation set for deep learning models for detecting seizures in children. The presence of repeated authors in the literature arises from the development of multiple deep learning models. (A) Forest plot illustrating the C-index summarization for the validation set. (B) Funnel plot illustrating the C-index for the validation set.

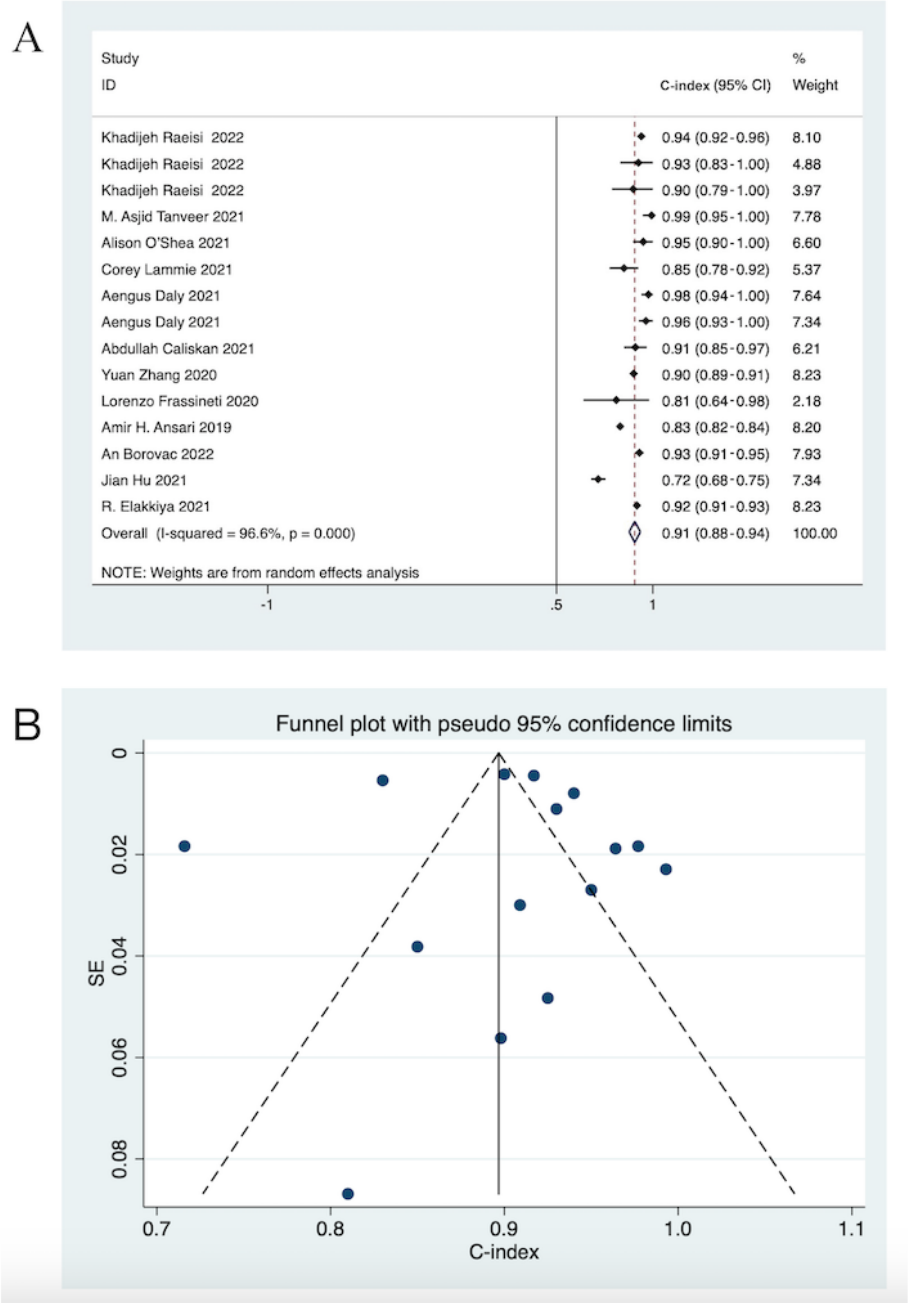

Furthermore, 21 validation sets provided sensitivity, specificity, or precision values. A random-effects model was leveraged for data analysis. The analysis showed a pooled sensitivity of 0.89 (95% CI 0.85-0.91), specificity of 0.91 (95% CI 0.88-0.93; Figure 6A), and an accuracy rate of 0.89 (95% CI 0.86-0.92; Figure 6B). Similarly, no significant publication bias was observed in these results.

XSL•FO
RenderX

Figure 6. The forest plot shows the sensitivity, specificity, and accuracy of deep learning models in detecting seizures in children. The presence of repeated authors in the literature arises from the development of multiple deep learning models. (A) The application of deep learning for seizure detection in children demonstrates the sensitivity and specificity of ensemble methods and forest plots; (B) Accuracy of the deep learning validation set post summarization.

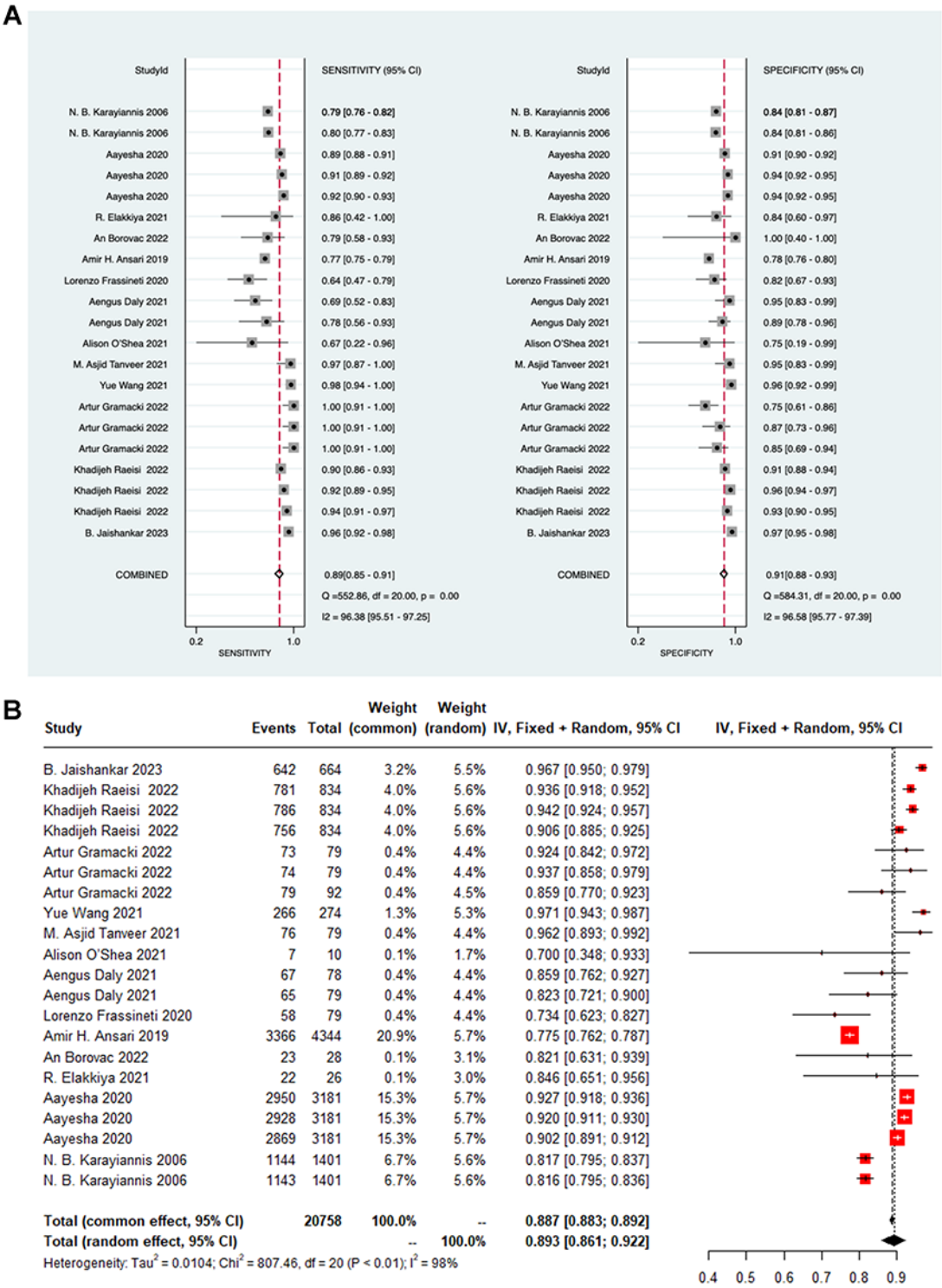

Discussion

Summary of Main Findings

Real-time detection of epileptic seizures has profound clinical significance. Studies have shown that if a child, who has a normal nervous system without a history of neurological diseases, experienced unprovoked epileptic seizures without obvious acute causes, then the risk of recurrence is about 25% within the following 1 year, and 45%-50% within the following 3 years [50-52]. In this study, the results showed that AI-based methods have high sensitivity, specificity, and accuracy in detecting epileptic seizures. The overall C-index, sensitivity, and specificity of ML are 0.73 (95% CI 0.67-0.79), 0.88 (95%

XSL•FO
RenderX

CI 0.83-0.91), and 0.83 (95% CI 0.71-0.90) in the validation set, respectively. DL exhibited even higher accuracy, with an overall C-index, sensitivity, and specificity of 0.91 (95% CI 0.88-0.94), 0.89 (95% CI 0.85-0.91), and 0.91 (95% CI 0.88-0.93) in the validation set. It was also found that EEG-based DL appears capable of detecting SE in a timely manner. This suggests that this study has the potential to be used in the development of portable devices for real-time monitoring in future research. Such devices could enhance the timely detection of epileptic seizures, enabling prompt clinical intervention and thereby reducing the risk of recurrence.

Comparison With Other Reviews

Several noninvasive methods for epilepsy detection are actively investigated. In recent years, researchers have also explored other noninvasive means for detecting epileptic seizures besides EEG, such as magnetoencephalography (MEG). MEG measures the magnetic fields generated by the brain to provide information about neural activity during epileptic seizures. Similar to EEG, MEG can also be used to detect signal changes before epileptic seizures. Some studies have shown that MEG has specific advantages in detecting extratemporal epileptic spikes, especially those from epileptic lesions located on the surface of the brain [53-55]. A meta-analysis by Brændholt and Jensen [56] showed that the sensitivity and specificity of MEG for detecting epileptic seizures were 0.77 (95% CI 0.60-0.90) and 0.75 (95% CI 0.53-0.90), respectively. While this finding supports MEG as an accurate method for detecting SE, compared with the results of this study, it suggested that both EEG-based ML and DL may offer superior performance in detecting SE. Another method is electrical source imaging (ESI), which is a model-based imaging technique that integrates the spatial and temporal components of EEG to identify the sources of abnormal electrical activity related to epileptic seizures [57]. In a study by Ricci et al [58], the application of ESI in pediatric patients with refractory epilepsy was evaluated. The results showed that the sensitivity, specificity, and accuracy of ESI in predicting seizures in children were 0.57 (95% CI 0.34-0.78), 0.86 (95% CI 0.57-0.98), and 0.69 (95% CI 0.51-0.84), respectively (P=.01). Furthermore, 1 study showed that the diagnostic performance of ESI is comparable with that of magnetic resonance imaging and positron emission computed tomography. The sensitivity of these 3 methods for predicting epileptic seizures was 0.88, 0.71, and 0.66, respectively, while their specificity was 0.47, 0.71, and 0.59, respectively [59]. While these detection methods have shown promise, real-time monitoring of epileptic seizures remains a challenge, requiring further attention in clinical practice. Our study demonstrates that DL-based approaches offer promising results with a sensitivity of 0.89 (95% CI 0.85-0.91) and a specificity of 0.91 (95% CI 0.88-0.93) for pediatric epileptic seizure detection. Integrating DL based on EEG or electrocardiography into portable smart devices could significantly enhance real-time monitoring capabilities.

We have also noticed that there are permanent differences between DL and ML, and both have their own advantages. For ML, interpretable clinical features and corresponding ML models can be developed. For example, one of the original studies we included based on ML showed that the RF, SVM, and KNN models had a sensitivity of 0.93, 0.76, and 0.90, a specificity of 0.89, 0.76, and 0.90, and an accuracy of 0.91, 0.81, and 0.91 in predicting epileptic seizures, respectively [39]. Interpretable models in the clinical setting are still a direction that we are very interested in. However, for certain imaging applications, ML requires manual preprocessing, like segmenting regions of interest and extracting features. During this process, it is difficult to avoid any heterogeneity or bias due to manual experience, therefore, for image analysis, we prefer to use the DL to intelligently train the differences between them to develop intelligent detection tools. The primary advantage of CNN is that it can automatically detect important robust features without any manual intervention. The point is to estimate and determine the number of layers and the size of the filters in each layer, and the depth of the structure is the key to deep CNN. The network depth change will affect the convolutional receptive field and the corresponding learning feature complexity [60]. For instance, one of the original studies based on DL showed that the model had a sensitivity of 0.96, a specificity of 0.97, and an accuracy of 0.97 in predicting epileptic seizures, and was compared with common ML methods, such as RF, DT, and SVM [19].

In clinical practice, we need real-time monitoring for some diseases, such as epileptic seizures. However, ML still faces serious challenges in real-time disease monitoring, as the ML modeling is based on manual encoding, making it difficult to process images intelligently. Compared with ML, DL has significant advantages in real-time monitoring. We also found that researchers have done further work in other fields, developing wearable devices for real-time disease monitoring based on DL methods, such as monitoring heart-related diseases [61,62] or automatically assessing the severity of knee osteoarthritis [63].

Advantages and Limitations

We found that current meta-analyses related to epilepsy detection are usually based on the direct diagnosis by clinical neurologists through the reading of EEG signals [64,65]. This study is the first study to discuss the use of AI-based methods, including ML and DL, for the prediction and diagnosis of epileptic seizures. Particularly, DL has shown very desirable accuracy in the diagnosis of SE, which may provide some theoretical support for the subsequent development of intelligent reading tools or wearable devices. However, several limitations should be considered. First, the number of cases used for model construction is limited, whereas ML algorithms require a larger sample size to build robust models. This may influence the generalizability of these results. Second, some included studies lack independent validation. Third, the detection performance of DL across different age groups remains uncertain. Third, we only searched for literature published in English in this study. For this, there are 2 primary reasons, one is there are certain search barriers in this study and the other is that considering most readers would need to review the relevant original studies while reading this article, we developed this search strategy to improve the readability for the readers. Of course, it is also a limitation of this study, and we hope that future research can cover more detection tools developed by diverse populations

XSL•FO
RenderX

from different countries and ethnicities to validate the results of this study.

Conclusions

Our systematic review demonstrates promising accuracy of AI methods in epilepsy detection. DL appears to offer higher detection performance compared with ML, this finding supports our initiative to further research and develop early warning tools using DL.

Acknowledgments

We would like to thank the researchers and study participants for their contributions. The authors did not receive specific funding.

Data Availability

All data generated or analyzed during this study are included in the article and supplementary material.

Authors' Contributions

ZZ contributed to conceptualization, investigation, and writing-original draft. BC contributed to visualization, supervision, and writing-review and editing. DX contributed to investigation and writing-review and editing. FT contributed to investigation and writing-review and editing. XL contributed to resources, project administration, and writing-review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 445 KB - jmir_v26i1e55986_app1.pdf]

Multimedia Appendix 2
Literature search strategy.
[DOCX File , 18 KB - jmir_v26i1e55986_app2.docx]

References

1. Fisher RS, van Emde Boas W, Blume W, Elger C, Genton P, Lee P, et al. Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia 2005;46(4):470-472 [FREE Full text] [doi: 10.1111/j.0013-9580.2005.66104.x] [Medline: 15816939]
2. Zhang Y, Yang S, Liu Y, Zhang Y, Han B, Zhou F. Integration of 24 feature types to accurately detect and predict seizures using scalp EEG signals. Sensors (Basel) 2018;18(5):1372 [FREE Full text] [doi: 10.3390/s18051372] [Medline: 29710763]
3. Tran LV, Tran HM, Le TM, Huynh TTM, Tran HT, Dao SVT. Application of machine learning in epileptic seizure detection. Diagnostics (Basel) 2022;12(11):2879 [FREE Full text] [doi: 10.3390/diagnostics12112879] [Medline: 36428941]
4. Valencia I, Lozano G, Kothare SV, Melvin JJ, Khurana DS, Hardison HH, et al. Epileptic seizures in the pediatric intensive care unit setting. Epileptic Disord 2006;8(4):277-284 [FREE Full text] [Medline: 17150441]
5. Bandarabadi M, Teixeira CA, Rasekhi J, Dourado A. Epileptic seizure prediction using relative spectral power features. Clin Neurophysiol 2015;126(2):237-248. [doi: 10.1016/j.clinph.2014.05.022] [Medline: 24969376]
6. Nagarajan L, Palumbo L, Ghosh S. Classification of clinical semiology in epileptic seizures in neonates. Eur J Paediatr Neurol 2012;16(2):118-125. [doi: 10.1016/j.ejpn.2011.11.005] [Medline: 22185688]
7. van Leeuwen KG, Sun H, Tabaeizadeh M, Struck AF, van Putten MJAM, Westover MB. Detecting abnormal electroencephalograms using deep convolutional networks. Clin Neurophysiol 2019;130(1):77-84 [FREE Full text] [doi: 10.1016/j.clinph.2018.10.012] [Medline: 30481649]
8. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
9. Stevenson NJ, Tapani K, Lauronen L, Vanhatalo S. A dataset of neonatal EEG recordings with seizure annotations. Sci Data 2019;6:190039 [FREE Full text] [doi: 10.1038/sdata.2019.39] [Medline: 30835259]
10. Wei Z, Wang X, Ren L, Liu C, Liu C, Cao M, et al. Using machine learning approach to predict depression and anxiety among patients with epilepsy in China: a cross-sectional study. J Affect Disord 2023;336:1-8. [doi: 10.1016/j.jad.2023.05.043] [Medline: 37209912]

XSL•FO
RenderX

11. Yossofzai O, Fallah A, Maniquis C, Wang S, Ragheb J, Weil AG, et al. Development and validation of machine learning models for prediction of seizure outcome after pediatric epilepsy surgery. Epilepsia 2022;63(8):1956-1969. [doi: 10.1111/epi.17320] [Medline: 35661152]
12. Truong ND, Nguyen AD, Kuhlmann L, Bonyadi MR, Yang J, Ippolito S, et al. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Netw 2018;105:104-111. [doi: 10.1016/j.neunet.2018.04.018] [Medline: 29793128]
13. Daoud H, Bayoumi MA. Efficient epileptic seizure prediction based on deep learning. IEEE Trans Biomed Circuits Syst 2019;13(5):804-813. [doi: 10.1109/TBCAS.2019.2929053] [Medline: 31331897]
14. Ozcan AR, Erturk S. Seizure prediction in scalp EEG using 3D convolutional neural networks with an image-based approach. IEEE Trans Neural Syst Rehabil Eng 2019;27(11):2284-2293. [doi: 10.1109/TNSRE.2019.2943707] [Medline: 31562096]
15. Wang G, Wang D, Du C, Li K, Zhang J, Liu Z, et al. Seizure prediction using directed transfer function and convolution neural network on intracranial EEG. IEEE Trans Neural Syst Rehabil Eng 2020;28(12):2711-2720. [doi: 10.1109/TNSRE.2020.3035836] [Medline: 33147147]
16. Yang X, Zhao J, Sun Q, Lu J, Ma X. An effective dual self-attention residual network for seizure prediction. IEEE Trans Neural Syst Rehabil Eng 2021;29:1604-1613. [doi: 10.1109/TNSRE.2021.3103210] [Medline: 34370668]
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
19. Jaishankar B, A AM, Vidyabharathi D, Raja L. A novel epilepsy seizure prediction model using deep learning and classification. Healthcare Analytics 2023;4:100222. [doi: 10.1016/j.health.2023.100222]
20. Raeisi K, Khazaei M, Croce P, Tamburro G, Comani S, Zappasodi F. A graph convolutional neural network for the automated detection of seizures in the neonatal EEG. Comput Methods Programs Biomed 2022;222:106950 [FREE Full text] [doi: 10.1016/j.cmpb.2022.106950] [Medline: 35717740]
21. Gramacki A, Gramacki J. A deep learning framework for epileptic seizure detection based on neonatal EEG signals. Sci Rep 2022;12(1):13010 [FREE Full text] [doi: 10.1038/s41598-022-15830-2] [Medline: 35906248]
22. Gao Y, Chen X, Liu A, Liang D, Wu L, Qian R, et al. Pediatric seizure prediction in scalp EEG using a multi-scale neural network with dilated convolutions. IEEE J Transl Eng Health Med 2022;10:4900209 [FREE Full text] [doi: 10.1109/JTEHM.2022.3144037] [Medline: 35356539]
23. Borovac A, Gudmundsson S, Thorvardsson G, Moghadam SM, Nevalainen P, Stevenson N, et al. Ensemble learning using individual neonatal data for seizure detection. IEEE J Transl Eng Health Med 2022;10:4901111 [FREE Full text] [doi: 10.1109/JTEHM.2022.3201167] [Medline: 36147876]
24. Wang Y, Wang Y, Piao Y. Automatic seizure prediction based on cross-feature fusion stream convolutional neural network. In: Mahmud M, Kaiser MS, Kasabov N, Iftekharuddin K, Zhong N, editors. Applied Intelligence and Informatics. AII 2021. Communications in Computer and Information Science. Switzerland: Springer, Cham; 2021.
25. Tanveer MA, Khan MJ, Sajid H, Naseer N. Convolutional neural networks ensemble model for neonatal seizure detection. J Neurosci Methods 2021;358:109197. [doi: 10.1016/j.jneumeth.2021.109197] [Medline: 33864835]
26. O'Shea A, Ahmed R, Lightbody G, Pavlidis E, Lloyd R, Pisani F, et al. Deep learning for EEG seizure detection in preterm infants. Int J Neural Syst 2021;31(8):2150008. [doi: 10.1142/S0129065721500088] [Medline: 33522460]
27. Daly A, O'Shea A, Lightbody G, Temko A. Towards deeper neural networks for neonatal seizure detection. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:920-923. [doi: 10.1109/EMBC46164.2021.9629485] [Medline: 34891440]
28. Caliskan A, Rencuzogullari S. Transfer learning to detect neonatal seizure from electroencephalography signals. Neural Comput & Applic 2021;33(18):12087-12101. [doi: 10.1007/s00521-021-05878-y]
29. Frassineti L, Ermini D, Fabbri R, Fabbri C. Neonatal seizures detection using stationary wavelet transform and deep neural networks: preliminary results. 2020 Presented at: IEEE 20th Mediterranean Electrotechnical Conference (MELECON); 2020 June 16-18; Palermo, Italy.
30. Ansari AH, Cherian PJ, Caicedo A, Naulaers G, De Vos M, Van Huffel S. Neonatal seizure detection using deep convolutional neural networks. Int J Neural Syst 2019;29(4):1850011. [doi: 10.1142/S0129065718500119] [Medline: 29747532]
31. Affes A, Mdhaffar A, Triki C, Jmaiel M, Freisleben B. A convolutional gated recurrent neural network for epileptic seizure prediction. 2019 Presented at: 17th International Conference on Smart Living and Public Health (ICOST); 2019 October 14; Berlin, Heidelberg.
32. Pavel AM, O'Toole JM, Proietti J, Livingstone V, Mitra S, Marnane WP, ANSeR Consortium. Machine learning for the early prediction of infants with electrographic seizures in neonatal hypoxic-ischemic encephalopathy. Epilepsia 2023;64(2):456-468 [FREE Full text] [doi: 10.1111/epi.17468] [Medline: 36398397]
33. McKee JL, Kaufman MC, Gonzalez AK, Fitzgerald MP, Massey SL, Fung F, et al. Leveraging electronic medical record-embedded standardised electroencephalogram reporting to develop neonatal seizure prediction models: a retrospective cohort study. Lancet Digit Health 2023;5(4):e217-e226 [FREE Full text] [doi: 10.1016/S2589-7500(23)00004-3] [Medline: 36963911]

XSL•FO
RenderX

34. PEREZ-SANCHEZ AV, VALTIERRA-RODRIGUEZ M, PEREZ-RAMIREZ CA, DE-SANTIAGO-PEREZ JJ, AMEZQUITA-SANCHEZ JP. Epileptic seizure prediction using wavelet transform, fractal dimension, support vector machine, and eeg signals. Fractals 2022;30(07):2250154. [doi: 10.1142/s0218348x22501547]
35. Azriel R, Hahn CD, De Cooman T, Van Huffel S, Payne ET, McBain KL, et al. Machine learning to support triage of children at risk for epileptic seizures in the pediatric intensive care unit. Physiol Meas 2022;43(9). [doi: 10.1088/1361-6579/ac8ccd] [Medline: 36007520]
36. Sethy PK, Panigrahi M, Vijayakumar K, Behera SK. Machine learning based classification of EEG signal for detection of child epileptic seizure without snipping. Int J Speech Technol 2021;26(3):559-570. [doi: 10.1007/s10772-021-09855-7]
37. Hu J, Fung FW, Jacobwitz M, Parikh DS, Vala L, Donnelly M, et al. Machine learning models to predict electroencephalographic seizures in critically ill children. Seizure 2021;87:61-68 [FREE Full text] [doi: 10.1016/j.seizure.2021.03.001] [Medline: 33714840]
38. Elakkiya R. Machine learning based intelligent automated neonatal epileptic seizure detection. J Intell Fuzzy Syst 2021;40(5):8847-8855. [doi: 10.3233/JIFS-200800] [Medline: 23367031]
39. Aayesha, Qureshi MB, Afzaal M, Qureshi MS, Fayaz M. Machine learning-based EEG signals classification model for epileptic seizure detection. Multimed Tools Appl 2021;80(12):17849-17877. [doi: 10.1007/s11042-021-10597-6]
40. Fung FW, Parikh DS, Jacobwitz M, Vala L, Donnelly M, Wang Z, et al. Validation of a model to predict electroencephalographic seizures in critically ill children. Epilepsia 2020;61(12):2754-2762. [doi: 10.1111/epi.16724] [Medline: 33063870]
41. Fung FW, Jacobwitz M, Parikh DS, Vala L, Donnelly M, Fan J, et al. Development of a model to predict electroencephalographic seizures in critically ill children. Epilepsia 2020;61(3):498-508. [doi: 10.1111/epi.16448] [Medline: 32077099]
42. Açıkoğlu M, Tuncer SA. Incorporating feature selection methods into a machine learning-based neonatal seizure diagnosis. Med Hypotheses 2020;135:109464. [doi: 10.1016/j.mehy.2019.109464] [Medline: 31731060]
43. Sansevere AJ, Kapur K, Peters JM, Fernández IS, Loddenkemper T, Soul JS. Seizure prediction models in the neonatal intensive care unit. J Clin Neurophysiol 2019;36(3):186-194 [FREE Full text] [doi: 10.1097/WNP.0000000000000574] [Medline: 30882530]
44. Pan Y, Li H, Liu L, Li Q, Hou X, Dong B. aEEG signal analysis with ensemble learning for newborn seizure detection. 2019 Presented at: Multiscale Multimodal Medical Imaging: First International Workshop, MMMI 2019; 2019 Oct 13; Berlin, Heidelberg.
45. Temko A, Thomas E, Marnane W, Lightbody G, Boylan G. EEG-based neonatal seizure detection with support vector machines. Clin Neurophysiol 2011;122(3):464-473 [FREE Full text] [doi: 10.1016/j.clinph.2010.06.034] [Medline: 20713314]
46. Karayiannis NB, Mukherjee A, Glover JR, Frost JD, Hrachovy JRA, Mizrahi EM. An evaluation of quantum neural networks in the detection of epileptic seizures in the neonatal electroencephalogram. Soft Comput 2005;10(4):382-396. [doi: 10.1007/s00500-005-0498-4] [Medline: 16822243]
47. Breiman L. Random forests. Machine Learning 2001;45(1):5-32.
48. Agarwal S. Data mining: data mining concepts and techniques. 2013 Presented at: International Conference on Machine Intelligence and Research Advancement; 2013 December 21-23; Katra, India.
49. Cherkassky V. The nature of statistical learning theory~. IEEE Trans Neural Netw 1997;8(6):1564. [doi: 10.1109/TNN.1997.641482] [Medline: 18255760]
50. Shinnar S, O'Dell C, Berg AT. Mortality following a first unprovoked seizure in children: a prospective study. Neurology 2005;64(5):880-882. [doi: 10.1212/01.WNL.0000152893.72146.92] [Medline: 15753427]
51. Garcia Pierce J, Aronoff S, Del Vecchio M. Systematic review and meta-analysis of seizure recurrence after a first unprovoked seizure in 815 neurologically and developmentally normal children. J Child Neurol 2017;32(13):1035-1039. [doi: 10.1177/0883073817726461] [Medline: 28879801]
52. Pellino G, Faggioli R, Madrassi L, Falsaperla R, Suppiej A. Operational diagnosis of epilepsy in children at undetermined risk: a meta-analysis of prognostic factors for seizure recurrence. Epilepsy Behav 2022;127:108498. [doi: 10.1016/j.yebeh.2021.108498] [Medline: 34954509]
53. Knowlton RC, Laxer KD, Aminoff MJ, Roberts TP, Wong ST, Rowley HA. Magnetoencephalography in partial epilepsy: clinical yield and localization accuracy. Ann Neurol 1997;42(4):622-631. [doi: 10.1002/ana.410420413] [Medline: 9382474]
54. Shiraishi H, Watanabe Y, Watanabe M, Inoue Y, Fujiwara T, Yagi K. Interictal and ictal magnetoencephalographic study in patients with medial frontal lobe epilepsy. Epilepsia 2001;42(7):875-882 [FREE Full text] [doi: 10.1046/j.1528-1157.2001.042007875.x] [Medline: 11488887]
55. de Jongh A, de Munck JC, Gonçalves SI, Ossenblok P. Differences in MEG/EEG epileptic spike yields explained by regional differences in signal-to-noise ratios. J Clin Neurophysiol 2005;22(2):153-158. [doi: 10.1097/01.wnp.0000158947.68733.51] [Medline: 15805816]
56. Brændholt M, Jensen M. Evidence from meta-analysis supports ictal magnetoencephalographic source imaging as an accurate method in presurgery evaluation of patients with drug-resistant epilepsy. Clin EEG Neurosci 2020;51(6):403-411. [doi: 10.1177/1550059420921534] [Medline: 32437218]

XSL•FO
RenderX

57. Mégevand P, Seeck M. Electric source imaging for presurgical epilepsy evaluation: current status and future prospects. Expert Rev Med Devices 2020;17(5):405-412. [doi: 10.1080/17434440.2020.1748008] [Medline: 32223351]
58. Ricci L, Tamilia E, Alhilani M, Alter A, Perry MS, Madsen JR, et al. Source imaging of seizure onset predicts surgical outcome in pediatric epilepsy. Clin Neurophysiol 2021;132(7):1622-1635 [FREE Full text] [doi: 10.1016/j.clinph.2021.03.043] [Medline: 34034087]
59. Lascano AM, Perneger T, Vulliemoz S, Spinelli L, Garibotto V, Korff CM, et al. Yield of MRI, high-density electric source imaging (HD-ESI), SPECT and PET in epilepsy surgery candidates. Clin Neurophysiol 2016;127(1):150-155. [doi: 10.1016/j.clinph.2015.03.025] [Medline: 26021550]
60. OrShea A, Lightbody G, Boylan G, Temko A. Investigating the impact of CNN depth on neonatal seizure detection performance. Annu Int Conf IEEE Eng Med Biol Soc 2018;2018:5862-5865. [doi: 10.1109/EMBC.2018.8513617] [Medline: 30441669]
61. Xie L, Li Z, Zhou Y, He Y, Zhu J. Computational diagnostic techniques for electrocardiogram signal analysis. Sensors (Basel) 2020;20(21):6318 [FREE Full text] [doi: 10.3390/s20216318] [Medline: 33167558]
62. Liu J, Li Z, Jin Y, Liu Y, Liu C, Zhao L, et al. A review of arrhythmia detection based on electrocardiogram with artificial intelligence. Expert Rev Med Devices 2022;19(7):549-560. [doi: 10.1080/17434440.2022.2115887] [Medline: 35993248]
63. Yang J, Ji Q, Ni M, Zhang G, Wang Y. Automatic assessment of knee osteoarthritis severity in portable devices based on deep learning. J Orthop Surg Res 2022;17(1):540 [FREE Full text] [doi: 10.1186/s13018-022-03429-2] [Medline: 36514158]
64. Falsaperla R, Scalia B, Giaccone F, Suppiej A, Pulvirenti A, Mailo J, et al. aEEG vs cEEG's sensivity for seizure detection in the setting of neonatal intensive care units: a systematic review and meta-analysis. Acta Paediatr 2022;111(5):916-926. [doi: 10.1111/apa.16251] [Medline: 35006632]
65. Trujillo MH, Castillo A, España J, Manzo A, Zerpa R. Impact of intensive care management on the prognosis of tetanus. Analysis of 641 cases. Chest 1987;92(1):63-65. [doi: 10.1378/chest.92.1.63] [Medline: 3595250]

Abbreviations

AI: artificial intelligence
CHB-MIT: Children's Hospital Boston-Massachusetts Institute of Technology
CNN: convolutional neural network
DL: deep learning
DT: decision tree
EEG: electroencephalography
ESI: electrical source imaging
KNN: k-nearest neighbors
MEG: magnetoencephalography
MeSH: Medical Subject Headings
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies–2
RF: random forest
SVM: support vector machine

Edited by A Mavragani; submitted 02.01.24; peer-reviewed by A Jafarizadeh, C Ordun, S Hansun; comments to author 01.04.24; revised version received 13.05.24; accepted 28.10.24; published 11.12.24.

Please cite as:
Zou Z, Chen B, Xiao D, Tang F, Li X
Accuracy of Machine Learning in Detecting Pediatric Epileptic Seizures: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e55986
URL: https://www.jmir.org/2024/1/e55986
doi:10.2196/55986
PMID:

©Zhuan Zou, Bin Chen, Dongqiong Xiao, Fajuan Tang, Xihong Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leveraging Large Language Models for Improved Understanding of Communications With Patients With Cancer in a Call Center Setting: Proof-of-Concept Study

Seungbeom Cho[1*], PhD; Mangyeong Lee[2,3*], PhD; Jaewook Yu[1], BS; Junghee Yoon[4], PhD; Jae-Boong Choi[1,5], PhD; Kyu-Hwan Jung[6,7], PhD; Juhee Cho[2,4,8], PhD

[1]School of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Republic of Korea

[2]Department of Digital Health, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea

[3]Center for Clinical Epidemiology, Samsung Medical Center, Seoul, Republic of Korea

[4]Department of Clinical Research Design and Evaluation, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea

[5]Service Convergence Design (Interdisciplinary), Sungkyunkwan University, Suwon-si, Republic of Korea

[6]Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea

[7]Smart Healthcare Research Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea

[8]Cancer Education Center, Samsung Comprehensive Cancer Center, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Kyu-Hwan Jung, PhD
Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology
Sungkyunkwan University
81 Irwon - ro, Gangnam
Seoul, 06355
Republic of Korea
Phone: 82 02 3410 3632
Email: kyuhwanjung@gmail.com

Abstract

Background: Hospital call centers play a critical role in providing support and information to patients with cancer, making it crucial to effectively identify and understand patient intent during consultations. However, operational efficiency and standardization of telephone consultations, particularly when categorizing diverse patient inquiries, remain significant challenges. While traditional deep learning models like long short-term memory (LSTM) and bidirectional encoder representations from transformers (BERT) have been used to address these issues, they heavily depend on annotated datasets, which are labor-intensive and time-consuming to generate. Large language models (LLMs) like GPT-4, with their in-context learning capabilities, offer a promising alternative for classifying patient intent without requiring extensive retraining.

Objective: This study evaluates the performance of GPT-4 in classifying the purpose of telephone consultations of patients with cancer. In addition, it compares the performance of GPT-4 to that of discriminative models, such as LSTM and BERT, with a particular focus on their ability to manage ambiguous and complex queries.

Methods: We used a dataset of 430,355 sentences from telephone consultations with patients with cancer between 2016 and 2020. LSTM and BERT models were trained on 300,000 sentences using supervised learning, while GPT-4 was applied using zero-shot and few-shot approaches without explicit retraining. The accuracy of each model was compared using 1,000 randomly selected sentences from 2020 onward, with special attention paid to how each model handled ambiguous or uncertain queries.

Results: GPT-4, which uses only a few examples (a few shots), attained a remarkable accuracy of 85.2%, considerably outperforming the LSTM and BERT models, which achieved accuracies of 73.7% and 71.3%, respectively. Notably, categories such as “Treatment,” “Rescheduling,” and “Symptoms” involve multiple contexts and exhibit significant complexity. GPT-4 demonstrated more than 15% superior performance in handling ambiguous queries in these categories. In addition, GPT-4 excelled in categories like “Records” and “Routine,” where contextual clues were clear, outperforming the discriminative models. These

XSL•FO
RenderX

findings emphasize the potential of LLMs, particularly GPT-4, for interpreting complicated patient interactions during cancer-related telephone consultations.

Conclusions: This study shows the potential of GPT-4 to significantly improve the classification of patient intent in cancer-related telephone oncological consultations. GPT-4's ability to handle complex and ambiguous queries without extensive retraining provides a substantial advantage over discriminative models like LSTM and BERT. While GPT-4 demonstrates strong performance in various areas, further refinement of prompt design and category definitions is necessary to fully leverage its capabilities in practical health care applications. Future research will explore the integration of LLMs like GPT-4 into hybrid systems that combine human oversight with artificial intelligence–driven technologies.

(J Med Internet Res 2024;26:e63892) doi:10.2196/63892

KEYWORDS

large language model; cancer; supportive care; LLMs; patient communication; natural language processing; NLP; self-management; teleconsultation; triage services; telephone consultations

Introduction

Patients with cancer face various physical, psychosocial, and financial problems throughout their diagnosis, treatment, and recovery [1-4]. These multifactorial problems often increase patient distress and make it difficult for patients to commit themselves to their treatment [5]. Missed opportunities for timely intervention can lead to an impaired quality of life, reduced treatment effectiveness, poor prognosis, and even death [6-9]. Therefore, timely and accurate identification of problems raised by patients is important [10].

Operating a call center in a hospital is an effective way for patients to access information and support to ensure quality care. One-third of patients receive at least one telephone consultation [11,12]. In addition to informational support, patients can also advance to scheduled clinical visits due to worsening conditions and receive brief instructions for self-management. Moreover, one study reported that incorporating nurse practitioners into telephone triage services was associated with 30% fewer subsequent health care encounters [13].

Despite the clinical contributions of telephone consultations, there are still challenges relating to their operational efficiency and standardization. Patient satisfaction with call centers is associated with the average speed of answers [14]. In addition, poor knowledge among nurses in identifying and prioritizing patients' problems leads to patient dissatisfaction and slows clinical workflows [15-17]. In another study with nurses, 97% felt constant pressure to minimize call length, but only 27% agreed that there was standardized guidance supporting clinical decisions during calls [18]. Given that most call center workers are not medical professionals, these challenges may be held.

To address these concerns, there is growing interest in leveraging natural language processing, particularly to understand patient intents [19]. Because most existing consultation records lack data with accurate annotations about patient intents, fine-tuning of the pretrained language model is required. There have been well-documented studies on the performance of pretrained language model, such as bidirectional encoder representations from transformers (BERT), in clinical contexts [20,21]. However, this approach is slightly limited in that it may still depend on the quality and quantity of annotation, which is time-consuming and costly to acquire in real-world clinical settings. In contrast, large language models (LLMs), such as the GPT-3, which leverages billions of words and parameters, are known to better understand contextual information without explicit fine-tuning by leveraging in-context learning capabilities [22-24]. Therefore, this study evaluated whether LLMs can effectively understand complex and diverse queries during telephone consultations with patients with cancer. We also assessed the impact of uncertainty in the annotations of consultation data on the performance of the LLM models, with a particular focus on how such uncertainties affect model accuracy.

Methods

Overall Process

The overall process of applying discriminative deep learning models and generative LLM to classification tasks is presented in Figure 1. The purpose of phone calls is based on the transcripts of calls of patients with cancer as described. Initially, phone call transcripts were processed by converting them into a corpus through tokenization postremoval of stop words. Subsequently, each text segment within the transcripts underwent conversion into token sequences. These sequences were used to train and test discriminant deep learning models such as long short-term memory (LSTM) and BERT using supervised learning.

XSL•FO
RenderX

Figure 1. Process of applying discriminative deep learning models for classifying the purpose of telephone consultations with patients with cancer. BERT: bidirectional encoder representations from transformers; LSTM: long short-term memory.

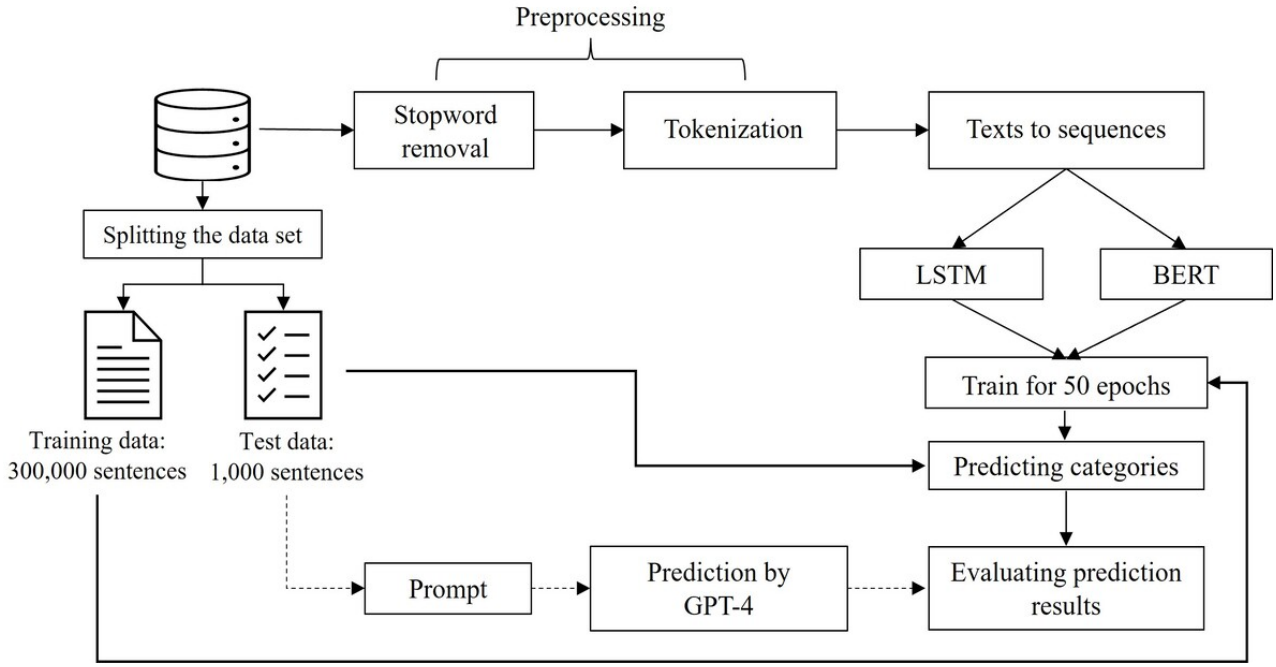

Despite the effectiveness of supervised deep learning models in discriminative tasks when a high-quality annotated dataset is available, the subjective and uncertain characteristics of phone call purpose categorization can lead to degraded performance. To address this issue, this study explored the potential of using GPT-4, a chatbot powered by a generative LLM to categorize phone calls without explicitly training the model with annotated data. Finally, we compared the performances of the two approaches with respect to the categorization accuracy.

Data Preparation

Datasets

This study used real-world data collected between 2016 and 2020 from the customer service center of a university hospital in Seoul, Korea. Specifically, we analyzed 430,355 instances documenting the initial requirements of patients when they contacted the hospital. All incoming calls were initially handled by the customer service center, after which counselors redirected them to the appropriate departments based on patients' unmet needs, categorized into 17 prespecified call purposes. Counselors are typically trained to write the exact words of patients verbatim when hearing their unmet needs, ensuring clear communication and understanding when the information is passed on to other departments from the customer service center. The records were all documented in Korean.

Approximately 300,000 of these sentences from data up to 2019 were used as training data, whereas 1,000 randomly selected sentences from the remaining 2020 data were used as the test dataset. The training dataset for the deep learning model was divided at an 80:20 ratio, with 20% used for validation during the training phase. The test data were categorized into 2 types: "Initial Categories," classified by counselors in the customer service center, and "Refined Categories," redefined by an oncology nurse and two researchers experienced in educating cancer patients or supportive care study for patients with cancer.

Definition and Refinement of Categories

The "Treatment" category encompasses consultations related to diagnosis, surgery, procedures, medication, dental treatment, and includes both therapeutic and procedural aspects of patient care. "Records" focuses on inquiries related to medical documents, including access to, interpretation of, and correction of medical records and opinions. The "Routine" category includes vaccination advice, such as flu shots, as well as lifestyle questions that emphasize the value of preventative care and routine health care. "Rescheduling" encompasses all aspects of organizing and coordinating medical appointments, surgeries, and treatments. The "Symptom" category includes issues related to medication side effects and symptoms occurring after an illness or treatment. A detailed description of each category is provided in Table S1 in Multimedia Appendix 1.

Model Development and Training

Training of the Discriminative Deep Learning Models

Two popular models widely used for processing sequence data in deep learning, LSTM, and BERT, were applied for classification. LSTMs, which are variants of recurrent neural networks (RNNs), are structurally more complex than RNNs, resulting in higher computational expense but superior capability in handling complex sequence data. BERT uses a different approach from RNNs, LSTMs, and gated recurrent units; it is a pretrained language model that leverages a large text corpus and is constructed on a bidirectional transformer with a self-attention mechanism [25].

Owing to the complexity of tokenizing Korean, which has a word order and contains multiple morphemes in a single clause, the LSTM model uses the Kiwi tokenizer, which is suitable for generating a Korean corpus [26]. For the BERT model, its tokenizer was applied for pretraining with special tokens ([CLS] and [SEP]) to denote the beginning and end of a sentence.

XSL•FO
RenderX

In this study, the BERT model accessible through the Keras library was used, with all layers of the preexisting model being retrained. For the LSTM model, a single LSTM layer with 256 units was used to capture the long-term dependencies in the sequence data. The Softmax activation function was used in the final layer of both models to generate a probability distribution across each category for the final classification task. Training of the models spanned 50 epochs and incorporated an EarlyStopping callback to avoid overfitting. This callback was configured to halt training if no improvement in validation loss was noted across 15 consecutive epochs. The model configuration that achieved the minimum validation loss during training was preserved and used to make predictions on the test dataset and conduct subsequent evaluations.

In Context Learning Using GPT-4

As a novel approach for classifying the purpose of phone calls, this study leveraged the capabilities of GPT-4, extending beyond traditional deep learning models such as LSTM and BERT. Initially, five categories—Treatment, Records, Routine, Rescheduling, and Symptom—were used for classification. These categories were expanded to 9 by incorporating the chain of thought. This refined categorization approach was applied to the prompts used with GPT-4 to enable a more precise classification within context learning. For comprehensive reference, the original and refined prompts used across all models are thoroughly documented in Table S2 in Multimedia Appendix 1.

Figure 2 shows the use of GPT-4 for categorizing both the initial and refined categories. The categories were restructured using the test dataset, and the prediction outcomes were compared by applying GPT-4 without an example (zero-shot) and with a few examples (few-shot). In GPT-4 zero-shot, specific prompts for categories such as “Treatment,” “Records,” “Routines,” “Rescheduling,” and “Symptoms” were used to determine the purpose of the consultation. In contrast, the GPT-4 few-shot approach aims to improve the accuracy of the model by adding more examples of correct answers to the zero-shot approach. In this process, the LSTMs were used to predict the validation data, and example sentences were generated from sentences that were incorrectly predicted. Example sentences are presented in Table S3 in Multimedia Appendix 1. Various settings, including temperature and role assignments in the system prompt, were experimented with in an attempt to achieve more consistent responses from GPT-4. However, no notable improvements or distinctive patterns were observed as a result of these adjustments. Consequently, GPT-4’s default web interface settings were applied for all experiments in this study.

Figure 2. Classification process of initial and refined categories using GPT-4. CoT: chain of thought; LSTM: long short-term memory.

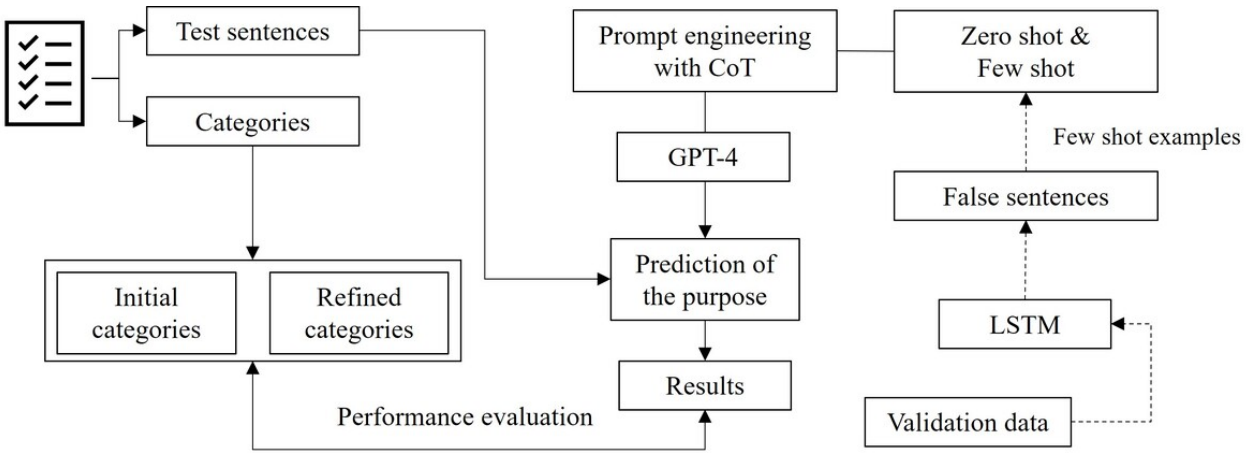

Performance Evaluation Methods

For the evaluation, 1000 randomly selected sentences from the 2020 dataset were used as the evaluation data. Each sentence was assigned both the existing labels and new labels reviewed by professional nurses. While 9 categories were used in the prompt construction, accuracy and recall scores were analyzed for 5 categories to evaluate the model’s performance. This was because the goal was to assess the model’s performance and accuracy based on categories commonly used in the actual hospital call center.

Ethical Considerations

The study was approved by the institutional review board of the Samsung Medical Center (SMC 2021-10-001). Our research uses retrospective data containing anonymized personal information and has received a waiver of informed consent from the subjects.

Results

Classification Results With Initial Categories

Figure 3 compares the prediction results of LSTM, BERT, GPT-4 zero-shot, and GPT-4 few-shot approaches based on the initial categories. According to these preliminary results, LSTM and BERT outperformed GPT-4 in most categories. In the “Treatment” category, all models showed relatively low performance, while GPT-4 slightly outperformed the others in the “Records” category. In the “Routine” category, GPT-4 few-shot achieved the highest accuracy; however, LSTM showed the best performance in the “Rescheduling” and “Symptoms” categories.

XSL•FO
RenderX

Figure 3. Comparison of long short-term memory (LSTM), bidirectional encoder representations from transformers (BERT), and GPT-4 in consultation call purpose classification with initial categories.

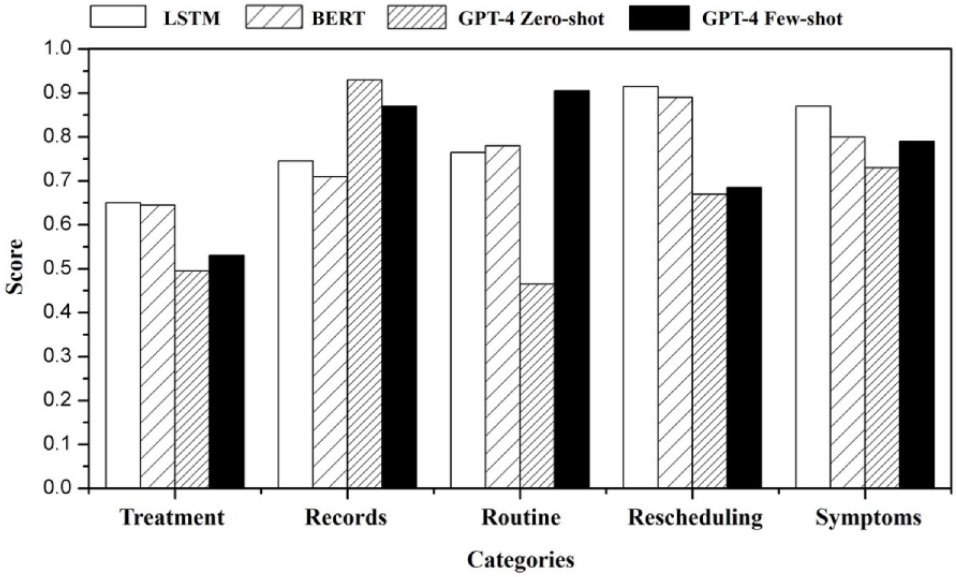

However, these results should be interpreted with caution. The differences may be due to uncertainty in the initial category definitions rather than variations in model capability. The boundaries between categories are often ambiguous and overlapping, potentially distorting the evaluation of model performance.

Thus, these preliminary results suggest that the need lies not only in comparing model performance but also in reviewing and improving the category definitions.

Confusion Matrix Results Between Initial and Refined Categories

To address the uncertainties present within the test dataset, the predefined categories were thoroughly reassessed by multiple researchers. Figure 4 shows confusion matrices, a standard metric for evaluating the efficacy of classification models that show the differences between the original and expert-refined categories.

Figure 4. Comparison of confusion matrix results between initial and refined categories.

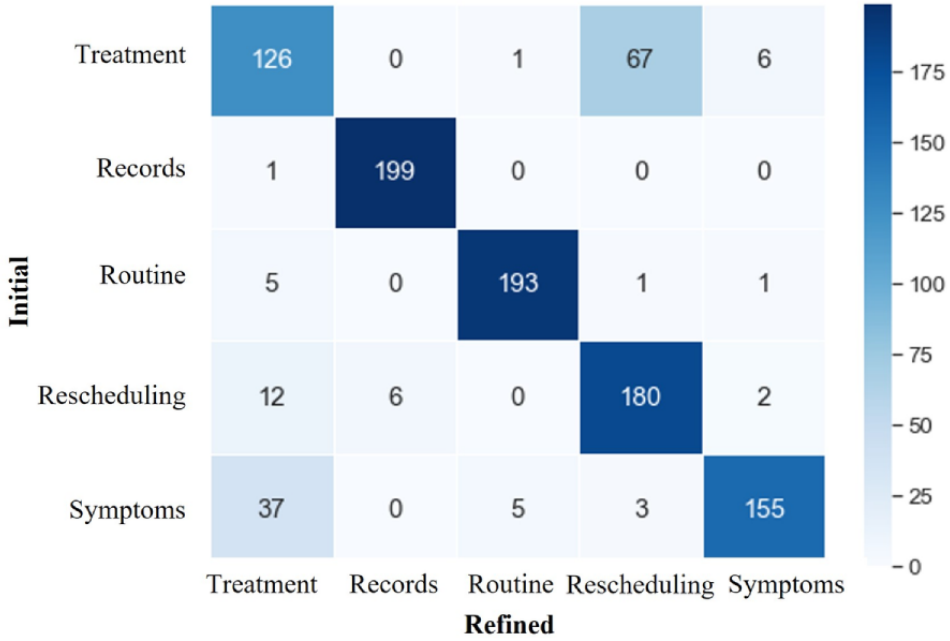

In the "Treatment" category, approximately 63% of the sentences matched correctly, but 33.5% were incorrectly classified as rescheduling and 3% as symptoms, indicating a significant overlap and misunderstanding among these categories. For the "Rescheduling" category, a precise correlation with the intended classification was evident in 90% of the sentences; however, approximately 6% were misclassified as treatment. The "Symptom" category had 77.5% correct matches, but 18.5% were confused about the treatment. In contrast, the "Records" and "Routine" categories demonstrated exceptional precision with accuracies of 99.5% and 96.5%, respectively, indicating a well-defined lexical and semantic

XSL•FO
RenderX

distinction within these categories, which is conducive to accurate classification. This analysis highlighted considerable ambiguity, primarily among the "Treatment," "Symptoms," and "Rescheduling" categories, leading to a strategic refinement of the classification prompts to resolve these issues.

Classification Results With Refined Categories

When constructing the GPT-4 prompts, the performance improved when the 5 categories were further subdivided based on the chain of thought. Figure 5 and Table 1 display the results evaluated using the recall score for each category as well as the overall accuracy. The GPT-4 few-shot model outperformed the LSTM and zero-shot models in most categories, achieving the highest recall for records at 0.9024, routine at 0.8995, and symptoms at 0.878. The zero-shot model exhibits the best performance in the "Rescheduling" and "Treatment" categories, with recall scores of 0.8964 and 0.7238, respectively. The LSTM model obtained the highest recall in the "Symptoms" category at 0.9268, but had the lowest performance in all other categories. Considering the total accuracy, the GPT-4 few-shot model achieved the highest score of 0.852, followed by the zero-shot model at 0.828 and the LSTM model at 0.737.

Figure 5. Performance metrics of long short-term memory (LSTM), bidirectional encoder representations from transformers (BERT), and GPT-4 models in consultation call purpose classification with refined categories.

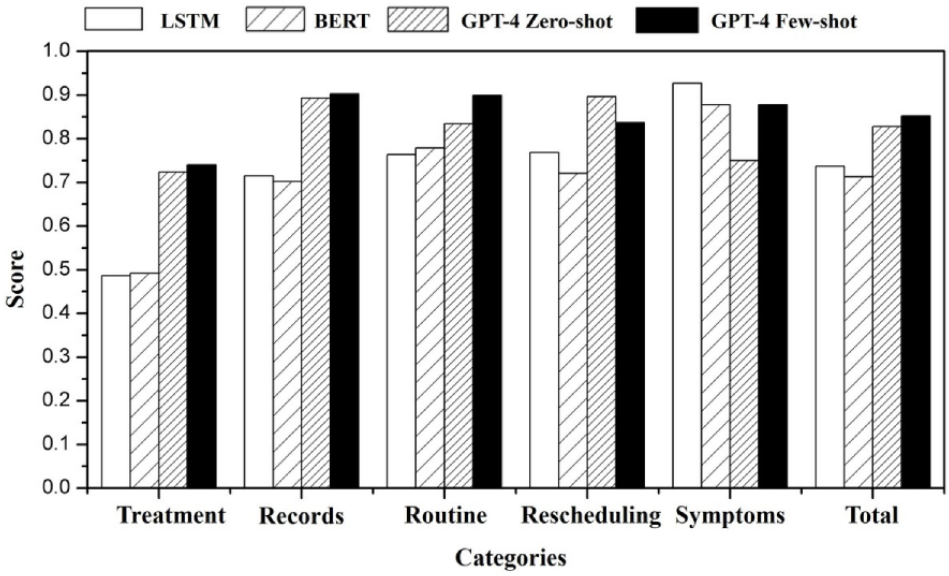

Table 1. Performance comparison of GPT-4 zero-shot and few-shot on refined categories.

	Discriminative models		GPT-4 (with CoT[a])	
	LSTM[b]	BERT[c]	Zero-shot	Few-shot
Recall score				
Treatment	0.486	0.492	0.724	0.740
Records	0.715	0.702	0.893	0.902
Routine	0.764	0.779	0.834	0.899
Rescheduling	0.769	0.721	0.896	0.837
Symptoms	0.927	0.878	0.750	0.878
Total				
Accuracy	0.737	0.713	0.828	0.852

[a]CoT: chain of thought.

[b]LSTM: long short-term memory.

[c]BERT: bidirectional encoder representations from transformers.

Discussion

Principal Results

The findings of this study highlight the potential challenges of using deep learning models, specifically LSTM and BERT, as well as the novel application of GPT-4 for categorizing phone calls in a medical context. Analyzing a dataset that includes transcripts from consultations of a patient with cancer has provided valuable perspectives on the efficiency of various learning approaches and the influence that the way categories are defined can have on classification accuracy.

XSL•FO
RenderX

In particular, GPT-4 few-shot exhibited an impressive capacity to use a limited set of specific examples, leading to notable enhancements in classification efficacy compared with discriminative models. With refined categories, few-shot GPT-4 achieved an accuracy of 85.2%, outperforming LSTM (73.7%) and BERT (71.3%). These results demonstrate that GPT-4 few-shot is effective in overcoming issues arising from ambiguous category boundaries within datasets characterized by complex and subjective features, which are common in telephone consultations with patients with cancer. This is comparable to other LLM studies that reported an accuracy of more than 90% in organizing structured data from medical records [27,28].

The evaluation of GPT-4's predictive capabilities yielded encouraging results but encountered difficulties. Notably, errors in classification were observed, specifically within medical documentation contexts, such as the removal of surgical adhesives, administration of topical medications, and symptom consultations related to medication and treatment references. To delve deeper into these findings, GPT-4 few-shots were used to predict multiple categories. Subsequently, a comparative analysis was performed between the predicted outcomes and categories revised by the experts. The examination revealed 55 instances of misclassification, representing a performance enhancement of more than 10% compared with single-category predictions.

Table 2 shows the representative sentences among the misclassified sentences. Sentences related to medical documentation, particularly those concerning prescription issuance, were predominantly misidentified as inquiries related to medical appointments. Moreover, questions about the removal of surgical tape or the application of ointment were incorrectly classified as related to daily living activities rather than consultations on symptoms or surgical and medical procedures. Certain sentences were misclassified as consultations about symptoms due to mentions of drugs and treatments rather than being recognized as related to routine daily activities. However, we need to consider the prevailing assumption that there is often a discrepancy between the patient's surface questions and the actual problems that physicians must address [29,30]. This is because doctors and nurses refer not only to patients' comments but also to medical records to determine patients' actual needs. Therefore, it may be difficult for LLMs to make in-depth inferences based solely on the patient's comments.

Table 2. Example of a sentence that was inaccurately predicted when compared with multicategories.

Sentences	Initial categories	Multicategories
The patient inquires if they can remove the tape applied after the prosthetic surgery on their own and how to apply the ointment. Please confirm. 환자, 보형물 수술후 테잎 붙여진것을 임의로 떼도 될지와 연고 바르는 것은 어떻게 하는지 상담 원하십니다. 확인부탁드립니다.	Routine	Treatment
TF2 general patient GS is currently receiving treatment from the Department of Breast Surgery and wishes to get a prescription—directed to the Gastroenterology Center to inquire about making an appointment for tomorrow; no issues related to COVID-19. TF2 일반환자GS 유방외과 진료보고 있는데 처방전을 발급 받을려고함-> 소화기 센터로 안내해줘서 내일 진료예약할수 있는지 문의-코로나 관련 이상없음	Records	Treatment, Rescheduling
Professor's patient, currently undergoing chemotherapy, reports discomfort in the prostate area, and is taking supplements. They are inquiring if it is safe to take the supplements along with chemotherapy. Please advise. 교수님환자, 항암제 복용중인데 전립선쪽이 안좋은셔서 영양제 복용중이라고 합니다. 항암제랑 영양제 함께 복용가능한지 문의합니다 상담부탁드립니다..	Routine	Treatment, Symptom
The patient has been taking tamoxifen and experienced menstruation during the month of surgery but has not had any periods since then, until they resumed on February 22. Please confirm if it would be appropriate to schedule an appointment directly with the gynecology department, or if the patient should attend another outpatient consultation first. 타목시펜 복용중으로 수술하신 달에는 생리하셨고 그이후로는 생리를 안하시다가 2/22일부터 생리를 하신다고 합니다. 바로 산부인과쪽으로 예약을 해드려도 될지, 아니면 외래진료를 한번 더 받으셔야 할지 확인부탁드립니다.	Symptom	Treatment, Rescheduling
The daughter mentioned that she had inquired about the possibility of hospitalization after speaking with the nurse at IM6 (hemato-oncology) and was told that they would consult and call back. She now wishes to be admitted today and requests to speak again with the nurse involved in the previous consultation. Thank you. 딸아까 im6 (혈액종양내과) 간호사와 통화후 입원여부 여쭤봤을때 상의하고 다시 전화드리겠다고 했었다고 합니다. 오늘 입원하겠다며 해당 상담했던 간호사와 재통화원하십니다. 감사합니다.	Treatment	Rescheduling Records

Limitations

This study had several limitations that require further methodological consideration. The best performance was observed when prompts were constructed based on more detailed categories derived from the initial categories, suggesting that the results may not be generalizable. For the LSTM model, we used the Kiwi tokenizer, a tool specifically designed for Korean, which differs from tools for other languages in its handling of syntactic properties and morpheme normalization. This

XSL•FO
RenderX

difference can impact the generalizability of the model performance across different languages. Furthermore, the data used in this study were not directly sourced from the content provided by the patients but rather based on the information summarized by the consultants. However, consultants (or counselors) are typically trained to write the exact words spoken by patients verbatim. Consequently, there are clear limitations for other researchers in obtaining the same level of performance as the results of this study.

Comparison With Prior Work

Despite encountering challenges in inferring the contextual purpose of calls, this study significantly contributes to broadening the application of LLMs in the field of oncology. Currently, LLMs play versatile roles in clinical settings, such as extracting meaningful outcomes from medical records or supporting clinical decision-making related to treatment options [27,28,31,32]. Some studies have attempted to use LLM to support clinician workflows in emergency department triage [33,34]. Our study suggests that LLMs can help both clinical and nonclinical workers develop a standardized perspective of problem identification in hospital call centers for patients with cancer. Although the accuracy and repeatability of LLMs may be controversial depending on the tasks performed by them, our study also suggests that LLMs are an efficient tool for using enormous and varied real-world data in the clinical environment [35,36].

Conclusions

This study demonstrated the potential of leveraging LLMs, particularly GPT-4, to understand the complex and diverse intentions of patients with cancer during telephone consultations. The superior performance of GPT-4 over retrained discriminative models highlights its ability to consistently identify patient intent without laborious annotation. However, challenges in classification accuracy underscore the importance of carefully defining categories and refining prompts to fully harness the power of LLMs. As more advanced LLMs emerge, exciting opportunities will emerge to further enhance the efficiency and standardization of telephone consultations for better cancer care. In the future, further study should be conducted to confirm whether an LLM-based telephone consultation system can enhance the work productivity and efficiency of hospital staff in identifying the unmet needs of patients and providing appropriate support.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A2C2011083).

Data Availability

Data supporting the findings of this study are available upon request from the corresponding authors. The codes supporting the findings of this study are available upon request from the corresponding authors.

Authors' Contributions

ML, KH, and JC performed the conceptualization of the study. KH designed the methodology, and ML carried out the investigation. ML and JY managed data curation, and SB and JW conducted formal analysis. SB led the visualization efforts. SB and ML prepared the original draft of the manuscript, with all authors contributing to the review and editing process. KH and JC provided supervision throughout the study. JC also provided the necessary resources and acquired funding for the project.
JC is co-corresponding author and can be reached at jcho@skku.edu.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GPT-4 prompt applied in this study.
[DOCX File , 26 KB - jmir_v26i1e63892_app1.docx]

References

1. Lentz R, Benson AB, Kircher S. Financial toxicity in cancer care: prevalence, causes, consequences, and reduction strategies. J Surg Oncol 2019;120(1):85-92. [doi: 10.1002/jso.25374] [Medline: 30650186]
2. Roseleur J, Edney LC, Jung J, Karnon J. Prevalence of unmet supportive care needs reported by individuals ever diagnosed with cancer in Australia: a systematic review to support service prioritisation. Support Care Cancer 2023;31(12):676. [doi: 10.1007/s00520-023-08146-y] [Medline: 37934313]
3. Smith A, Hyde YM, Stanford D. Supportive care needs of cancer patients: a literature review. Pall Supp Care 2014;13(4):1013-1017. [doi: 10.1017/s1478951514000959]
4. Hoekstra RA, Heins MJ, Korevaar JC. Health care needs of cancer survivors in general practice: a systematic review. BMC Fam Pract 2014;15:94. [doi: 10.1186/1471-2296-15-94] [Medline: 24885266]

XSL•FO
RenderX

5. Riba MB, Donovan KA, Ahmed K, Andersen B, Braun I, Breitbart WS, et al. NCCN guidelines® insights: distress management, version 2.2023. J Natl Compr Canc Netw 2023;21(5):450-457. [doi: 10.6004/jnccn.2023.0026] [Medline: 37156476]
6. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
7. Basch E, Mody GN, Dueck AC. Electronic patient-reported outcomes as digital therapeutics to improve cancer outcomes. JCO Oncol Pract 2020;16(9):541-542. [doi: 10.1200/OP.20.00264] [Medline: 32484724]
8. Basch E, Schrag D, Henson S, Jansen J, Ginos B, Stover AM, et al. Effect of electronic symptom monitoring on patient-reported outcomes among patients with metastatic cancer: a randomized clinical trial. JAMA 2022;327(24):2413-2422. [doi: 10.1001/jama.2022.9265] [Medline: 35661856]
9. Dai W, Feng W, Zhang Y, Wang XS, Liu Y, Pompili C, et al. Patient-reported outcome-based symptom management versus usual care after lung cancer surgery: a multicenter randomized controlled trial. J Clin Oncol 2022;40(9):988-996. [doi: 10.1200/jco.21.01344]
10. Johnson KD, Punches BE, Smith CR. Perceptions of the essential components of triage: a qualitative analysis. J Emerg Nurs 2021;47(1):192-197 [FREE Full text] [doi: 10.1016/j.jen.2020.08.009] [Medline: 33097241]
11. Pimentel LE, Yennurajalingam S, Chisholm G, Edwards T, Guerra-Sanchez M, De La Cruz M, et al. The frequency and factors associated with the use of a dedicated supportive care center telephone triaging program in patients with advanced cancer at a comprehensive cancer center. J Pain Symptom Manage 2015;49(5):939-944 [FREE Full text] [doi: 10.1016/j.jpainsymman.2014.10.020] [Medline: 25666520]
12. Paro A, Rice DR, Hyer JM, Palmer E, Ejaz A, Shaikh CF, et al. Telehealth utilization among surgical oncology patients at a large academic cancer center. Ann Surg Oncol 2022;29(12):7267-7276 [FREE Full text] [doi: 10.1245/s10434-022-12259-9] [Medline: 35896926]
13. Kaboli PJ, Augustine MR, Rose DE, Chawla N, Bouchard M, Hebert P. Call center remote triage by nurse practitioners was associated with fewer subsequent face-to-face healthcare visits. J Gen Intern Med 2021;36(8):2315-2322. [doi: 10.1007/s11606-020-06536-0] [Medline: 33501532]
14. Griffith KN, Li D, Davies ML, Pizer SD, Prentice JC. Call center performance affects patient perceptions of access and satisfaction. Am J Manag Care 2019;25(9):e282-e287 [FREE Full text] [Medline: 31518100]
15. Bijani M, Khaleghi AA. Challenges and barriers affecting the quality of triage in emergency departments: a qualitative study. Galen Med J 2019;8:e1619 [FREE Full text] [doi: 10.31661/gmj.v8i0.1619] [Medline: 34466538]
16. Reisi Z, Saberipour B, Adienh M, Hemmatipour A, Abdolahi Shahvali E. The level of awareness of the emergency department nurses of the triage principles in teaching hospitals. J Nurs Midwifery Sci 2018;5(1):32-37.
17. Rahmani F, Sepehri Majd P, Ebrahimi Bakhtavar H, Rahmani F. Evaluating the accuracy of emergency nurses in correct triage using emergency severity index triage in Sina Hospital of Tabriz: a cross-sectional analysis. J Emerg Pract Trauma 2017;4(1):9-13. [doi: 10.15171/jept.2017.19]
18. Stacey D, Graham ID, O'Connor AM, Pomey M. Barriers and facilitators influencing call center nurses' decision support for callers facing values-sensitive decisions: a mixed methods study. Worldviews Evid Based Nurs 2005;2(4):184-195. [doi: 10.1111/j.1741-6787.2005.00035.x] [Medline: 17040526]
19. Asti G, Sauta E, Curti N. Clinical text reports to stratify patients affected with myeloid neoplasms using natural language processing. Blood 2023;1:142. [doi: 10.1182/blood-2023-188292]
20. Possibilities and pitfalls of generative pre-trained transformers in healthcare. : IEEE; 2023 Presented at: International Conference on Advanced Computing & Communication Technologies (ICACCTech);; 2023 December 24; Banur, India. [doi: 10.1109/icacctech61146.2023.00016]
21. Shukla AD, Agarwal L, Mein J, Agarwal R. Catch me if you can: identifying fraudulent physician reviews with large language models using generative pre-trained transformers. ArXiv. Preprint posted on April 19, 2023 2023.
22. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening. medRxiv 2024:24302376 [FREE Full text] [doi: 10.1101/2024.02.08.24302376] [Medline: 38370719]
23. Chiang CC, Luo M, Dumkrieger G, Trivedi S, Chen YC, Chao CJ, et al. A large language model-based generative natural language processing framework finetuned on clinical notes accurately extracts headache frequency from electronic health records. medRxiv 2023:23296403. [doi: 10.1101/2023.10.02.23296403] [Medline: 37873417]
24. Brown T, Mann B, Ryder N. Language models are few-shot learners. Advances in neural information processing systems. arXiv:2005.14165 2005:33-901.
25. Eom G, Yun S, Byeon H. Predicting the sentiment of South Korean Twitter users toward vaccination after the emergence of COVID-19 Omicron variant using deep learning-based natural language processing. Front Med (Lausanne) 2022;9:948917 [FREE Full text] [doi: 10.3389/fmed.2022.948917] [Medline: 36186808]
26. Park D, Lim S, Choi Y, Oh H. Depression emotion multi-label classification using everytime platform with DSM-5 diagnostic criteria. IEEE Access 2023;11:89093-89106. [doi: 10.1109/access.2023.3305477]

XSL•FO
RenderX

27. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023;41(3):209-216 [FREE Full text] [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
28. Kehl KL, Xu W, Lepisto E, Elmarakeby H, Hassett MJ, Van Allen EM, et al. Natural language processing to ascertain cancer outcomes from medical oncologist notes. JCO Clinical Cancer Informatics 2020(4):680-690. [doi: 10.1200/cci.20.00020]
29. Sakai H, Umeda M, Okuyama H, Nakamura S. Differences in perception of breast cancer treatment between patients, physicians, and nurses and unmet information needs in Japan. Support Care Cancer 2020;28(5):2331-2338 [FREE Full text] [doi: 10.1007/s00520-019-05029-z] [Medline: 31482403]
30. Shin DW, Kim SY, Cho J, Sanson-Fisher RW, Guallar E, Chai GY, et al. Discordance in perceived needs between patients and physicians in oncology practice: a nationwide survey in Korea. J Clin Oncol 2011;29(33):4424-4429. [doi: 10.1200/JCO.2011.35.9281] [Medline: 22010016]
31. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6(11):e2343689. [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
32. Liu Z, Wang P, Li Y. Radonc-gpt: a large language model for radiation oncology. arXiv. Preprint posted on November 6, 2023 2023.
33. Franc JM, Cheng L, Hart A, Hata R, Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. CJEM 2024;26(1):40-46. [doi: 10.1007/s43678-023-00616-w] [Medline: 38206515]
34. Kim JH, Kim SK, Choi J, Lee Y. Reliability of ChatGPT for performing triage task in the emergency department using the Korean triage and acuity scale. Digit Health 2024;10:20552076241227132. [doi: 10.1177/20552076241227132] [Medline: 38250148]
35. Caglayan A, Slusarczyk W, Rabbani RD, Ghose A, Papadopoulos V, Boussios S. Large language models in oncology: revolution or cause for concern? Curr Oncol 2024;31(4):1817-1830 [FREE Full text] [doi: 10.3390/curroncol31040137] [Medline: 38668040]
36. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol 2023;13:1268915 [FREE Full text] [doi: 10.3389/fonc.2023.1268915] [Medline: 37731643]

Abbreviations

LLM: large language model
BERT: bidirectional encoder representations from transformers
LSTM: long short-term memory
RNN: recurrent neural network

Edited by G Eysenbach, A Coristine; submitted 02.07.24; peer-reviewed by B Bhasuran, L Wreyford; comments to author 09.10.24; revised version received 16.10.24; accepted 19.10.24; published 11.12.24.

Please cite as:
Cho S, Lee M, Yu J, Yoon J, Choi JB, Jung KH, Cho J
Leveraging Large Language Models for Improved Understanding of Communications With Patients With Cancer in a Call Center Setting: Proof-of-Concept Study
J Med Internet Res 2024;26:e63892
URL: https://www.jmir.org/2024/1/e63892
doi:10.2196/63892
PMID:

©Seungbeom Cho, Mangyeong Lee, Jaewook Yu, Junghee Yoon, Jae-Boong Choi, Kyu-Hwan Jung, Juhee Cho. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of a Literature Screening Tool: Few-Shot Learning Approach in Systematic Reviews

Phongphat Wiwatthanasetthakarn[1], BEng; Wanchana Ponthongmak[1], PhD; Panu Looareesuwan[1], PhD; Amarit Tansawet[2], MD, PhD; Pawin Numthavaj[1], MD, PhD; Gareth J McKay[3], PhD; John Attia[4], MD, PhD; Ammarin Thakkinstian[1], PhD

[1]Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[2]Department of Research and Medical Innovation, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand

[3]Centre for Public Health, Queen's University Belfast, Belfast, United Kingdom

[4]Centre for Clinical Epidemiology and Biostatistics, School of Medicine and Public Health, University of Newcastle, New South Wales, Australia

Corresponding Author:
Wanchana Ponthongmak, PhD
Department of Clinical Epidemiology and Biostatistics
Faculty of Medicine Ramathibodi Hospital
Mahidol University
4th Floor, Sukho Place Building
Sukhothai Road, Dusit
Bangkok, 10300
Thailand
Phone: 66 022010833
Fax: 66 022011284
Email: wanchana.pon@mahidol.edu

Abstract

Background: Systematic reviews (SRs) are considered the highest level of evidence, but their rigorous literature screening process can be time-consuming and resource-intensive. This is particularly challenging given the rapid pace of medical advancements, which can quickly make SRs outdated. Few-shot learning (FSL), a machine learning approach that learns effectively from limited data, offers a potential solution to streamline this process. Sentence-bidirectional encoder representations from transformers (S-BERT) are particularly promising for identifying relevant studies with fewer examples.

Objective: This study aimed to develop a model framework using FSL to efficiently screen and select relevant studies for inclusion in SRs, aiming to reduce workload while maintaining high recall rates.

Methods: We developed and validated the FSL model framework using 9 previously published SR projects (2016-2018). The framework used S-BERT with titles and abstracts as input data. Key evaluation metrics, including workload reduction, cosine similarity score, and the number needed to screen at 100% recall, were estimated to determine the optimal number of eligible studies for model training. A prospective evaluation phase involving 4 ongoing SRs was then conducted. Study selection by FSL and a secondary reviewer were compared with the principal reviewer (considered the gold standard) to estimate the false negative rate.

Results: Model development suggested an optimal range of 4-12 eligible studies for FSL training. Using 4-6 eligible studies during model development resulted in similarity thresholds for 100% recall, ranging from 0.432 to 0.636, corresponding to a workload reduction of 51.11% (95% CI 46.36-55.86) to 97.67% (95% CI 96.76-98.58). The prospective evaluation of 4 SRs aimed for a 50% workload reduction, yielding numbers needed to screen 497 to 1035 out of 995 to 2070 studies. The false negative rate ranged from 1.87% to 12.20% for the FSL model and from 5% to 56.48% for the second reviewer compared with the principal reviewer.

Conclusions: Our FSL framework demonstrates the potential for reducing workload in SR screening by over 50%. However, the model did not achieve 100% recall at this threshold, highlighting the potential for omitting eligible studies. Future work should focus on developing a web application to implement the FSL framework, making it accessible to researchers.

(J Med Internet Res 2024;26:e56863) doi:10.2196/56863

XSL•FO
RenderX

KEYWORDS

few shots learning; deep learning; natural language processing; S-BERT; systematic review; study selection; sentence-bidirectional encoder representations from transformers

Introduction

The evidence generated from systematic reviews (SRs) and meta-analyses of the published literature is considered to reflect the pinnacle of the evidence hierarchy pyramid [1,2], which in part explains the approximately 20-fold increase in published SRs between 2000 and 2019 [3]. However, conducting an SR requires significant time and human resources, particularly for the screening and selection of potentially eligible studies, data extraction, and bias assessment [4]. Furthermore, some SRs can be outdated by the time of publication, especially in highly progressive medical fields [5].

Multiple automated artificial intelligence (AI) tools have been developed using natural language processing techniques (eg, Abstrackr [6], DistillerSR [7], EPPI-reviewer [8], Rayyan [9], and Covidence [10]) to facilitate the SR processes, particularly study screening. Although the use of automated tools remains controversial [11], the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guideline [12] considers them a valuable tool given their potential to significantly reduce screening time or workload [13-18].

These tools [6-10] have largely been developed using supervised machine learning approaches with various document representation techniques and active learning frameworks, which typically require a considerable number of annotated studies as a training set. For example, Abstrackr typically requires 100 studies or more that have been manually screened by a reviewer before making a prediction [13]. During the model training process, iterative annotation of eligible and ineligible studies is required for improved model performance, with the AI model ranking or reordering studies according to their relevance. Subsequently, reviewers can choose between excluding all ineligible AI-predicted studies or simply using the predictions as a guide. The drawback of this approach is that reviewers must annotate studies without previous knowledge of the sufficient number of studies required for model training.

“Few-shot learning” (FSL) is a supervised machine learning approach that can learn from a small number of samples for model training and generalize from limited data [19]. Unlike traditional machine learning, it typically requires large datasets for high accuracy. FSL relies on metric learning during training to measure the similarity between new samples (unseen data) and known samples. Recently, FSL approaches have been successfully applied in many research areas, including computer vision, robotics, and natural language processing [19]. In addition, FSL has been used for concept extraction in health care, such as named entity recognition and text classification [20]. Therefore, the FSL approach is potentially useful for the development of SRs, particularly for study eligibility screening. Only a small number of studies need to be identified for training the FSL framework; theoretically, this approach should make machine learning faster at identifying the most relevant studies for SR compared with traditional approaches.

To the best of our knowledge, the use of FSL has yet to be applied as an automated tool for screening studies for SRs. Therefore, this study aimed to develop a new automated framework using FSL to facilitate the SR screening process with similarly high performance to traditional approaches. The model was trained and evaluated using previously published SRs completed within our institute and was prospectively validated.

Methods

Model Development and Validation Phase

Data Source

A total of 9 SRs [21-29], hereafter called SR1-9, were used to develop and validate the FSL framework. These SRs covered a range of SR topics including therapy (n=4), prognosis or risk (n=2), genetic association (n=2), and economic evaluation (n=1). All were conducted and published by researchers from the Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University between 2016 and 2018. This diverse set of 9 SRs was chosen to encompass various types of SRs, providing a comprehensive dataset to evaluate the flexibility and robustness of the FSL framework. The total number of identified studies for the 9 SRs ranged between 426 and 7341, of which 9 to 48 studies met the eligibility criteria for individual SRs (Table 1). Titles and abstracts of the studies identified were used as input for model training; the median number of tokens (words) per individual study ranged between 244 and 305. For each SR, the principal and second reviewers selected studies based on individual eligibility criteria. These reviewers included experts in specific areas, such as general physicians, surgeons, pharmacists, and clinical epidemiologists.

XSL•FO
RenderX

Table 1. Individual systematic review characteristics for model development and prospective evaluation.

Study	Project name	Study type	Number of studies	Total words (Vocabulary size)	Words per study median (IQR) (range)	Principal reviewer or secondary reviewer background
SR[a]1	Mesh Position for Hernia Prophylaxis After Midline Laparotomy: A Systematic Review and Network Meta-Analysis of Randomized Clinical Trials [21]	Therapeutic studies	3966	1,120,513 (37,777)	285 (206-343) (32-1629)	Surgeon
SR2	The Efficacy of Antibiotic Treatment versus Surgical Treatment of Uncomplicated Acute Appendicitis: Systematic Review and Network Meta-Analysis of Randomized Controlled Trial [22]	Therapeutic studies	1702	411,157 (18,174)	244 (164-307) (37-877)	Epidemiologist or surgeon
SR3	Efficacy and Safety of Urate-Lowering Agents in Asymptomatic Hyperuricemia: Systematic Review and Network Meta-Analysis of Randomized Controlled Trials [23]	Therapeutic studies	7341	2,205,482 (71,645)	297 (214-364) (20-1385)	Internist or pharmacist
SR4	Efficacy and Safety of Antiviral Agents in the Prophylaxis and Pre-Emptive Strategies for Cytomegalovirus Infection on Kidney Transplantation: A Systematic Review and Network Meta-Analysis [24]	Therapeutic studies	3144	874,261 (30,226)	272 (195-337) (33-3509)	Pharmacist or internist
SR5	Association Between Vitamin D and Uric Acid in Adults: A Systematic Review and Meta-Analysis [25]	Prognostic or risk studies	699	191,189 (15,298)	274 (189-340) (20-821)	Physician
SR6	Prognostic Model of Complications in Type 2 Diabetes: Systematic Review and Meta-Analysis [26]	Prognostic or risk studies	426	125,362 (10,843)	292 (232-339) (17-1112)	Epidemiologist
SR7	The Association Between Genetic Polymorphisms in ABCG2 and SLC2A9 and Urate: An Updated Systematic Review and Meta-Analysis [27]	Genetic association studies	1708	444,383 (26,883)	259 (204-312) (29-635)	Pharmacist
SR8	AHSG Gene Polymorphisms, Serum Fetuin-A Levels and Association with Type 2 Diabetes and Cardiovascular Diseases: A Systematic Review and Meta-Analysis [28]	Genetic association studies	1053	318,339 (18,787)	305 (255-346) (70-647)	Physician
SR9	Evaluation of the Cost Utility of Phosphate Binders as a Treatment Option for Hyperphosphatemia in Chronic Kidney Disease Patients: A Systematic Review and Meta-Analysis of the Economic Evaluations [29]	Economic evaluation studies	1653	463,892 (30,602)	244 (176-322) (18-2542)	Pharmacist

XSL•FO
RenderX

Study	Project name	Study type	Number of studies	Total words (Vocabulary size)	Words per study median (IQR) (range)	Principal reviewer or secondary reviewer background
Prospective evaluation SR1	Efficacy of EGFR-TKIs Targeted Therapy as Adjuvant Systemic Treatment for Non-Small Cell Lung Cancer: A Systematic Review and Meta-Analysis [30]	Therapeutic studies	1061	386,456 (21,281)	314 (225-442) (48-2553)	Pulmonologist or oncologist
Prospective evaluation SR2	Effect of Repetitive Peripheral Magnetic Stimulation on Upper Extremity Function After Stroke: A systematic review and meta-analysis [31]	Therapeutic studies	1699	561,904 (22,733)	307 (254-362) (19-4422)	Rehabilitation physician or obstetrician
Prospective evaluation SR3	Regular versus As-needed treatments for mild asthma in children, adolescents, and adults: A systematic review and meta-analysis [32]	Therapeutic studies	2136	752,467 (26,540)	324 (271-384) (20-2539)	Oncologist or pulmonologist
Prospective evaluation SR4	The association between cervical sonographic and successful induction of labor: A Systematic review and meta-analysis [33]	Prognostic or risk studies	1646	522,366 (22,494)	301 (243-361) (44-6122)	Obstetrician or rehabilitation physician

[a]SR: systematic review.

Data Splitting

In each SR, identified studies were dichotomized as either eligible or ineligible, denoted as S_+ and S_-, respectively. Studies were subdivided into training, validation (for model tuning), and test pools according to the following steps (Figure S1 in Multimedia Appendix 1). First, each eligible study was randomly assigned to the training, validation, or test pools with a split ratio of 50%:25%:25%. The eligible studies in the data pools were denoted as for the training pool, for the validation pool, and for the test pool, where n is the number of eligible studies in the pool. Second, training and validation pools included the integration of ineligible studies and, which were randomly selected from the overall S_-. The number of ineligible studies and included in the training and validation pools was ten times higher than the eligible studies and, that is, in a ratio of 1:10 representing 1 eligible per 10 ineligible studies. Finally, the remaining ineligible studies were assigned as and then combined with to form the test pool. The ineligible studies in training, validation, and test pools could be denoted as, , and, respectively, where m is the number of ineligible studies in each pool.

Experimental Scenarios

Multiple scenarios for each SR were tested by varying the number of studies included in and as follows:

- Using all studies in both training and validation pools (scenario 1).
- Reducing S_+ iteratively by 10% in the training and validation pools (scenario 2, 3, …, i) to a minimum number of 2 in S_+ (as training data requires a minimum of one positive paired sample in the model framework [refer to Data pairing section]). For and selection in scenario 2 to scenario i, the Euclidean distance between the centroid (ie, the estimated center of all S_+ in the vector space) and individual S_+ was used as a criterion for the iterative exclusion of S_+ from the training and validation pools. The S_+ farthest from the centroid was initially excluded for each iteration.

Furthermore, the S_- number for both pools was reduced to maintain a constant ratio (ie, 1:10) of S_+ and S_-. In the test pools, the numbers of and remained the same as scenario 1 for all scenarios to determine the optimal number of eligible studies required for model training and validation.

Data Pairing

The model was trained using the Sentence-Bidirectional Encoder Representation for Transformer (S-BERT) [34] which requires paired samples for input data that were generated from each data pool (ie, training, validation, and test pools). For instance, each was paired with and; and pairs were labeled as 1 (ie, positive pairs), whereas and pairs were labeled as 0 (ie, negative pairs). The same pairing process was also performed in the validation set and used for model tuning. All potential combinations of positive paired samples were represented by, where is a combination of positive paired

XSL•FO
RenderX

samples, and the number of eligible studies. For instance, the scenario with 6 eligible studies will have 60 ineligible studies. A training dataset will consist of 4 and 40 eligible and ineligible studies, respectively. These could be paired as 6 positive pairs and 160 negative pairs with a total of 166 paired samples (Figure S2 in Multimedia Appendix 1). A validation dataset includes 2 eligible and 20 ineligible studies with a total of 41 paired samples. The remaining studies were used as test data.

Model Architecture and Word Embedding

This study adopted an FSL framework that required several samples for initial model training. We used S-BERT, a Siamese networks architecture model [35] (Figure S3 in Multimedia Appendix 1), for model training with pretrained weighting of "all-mpnet-base-v2," available on Hugging Face [36]. Each paired sample represented 2 individual studies, study A and B. Raw text data (ie, title and abstract) from each study were fed into S-BERT and transformed into a numerical vector represented by a size of 1 × 384 dimensions. The cosine similarity score was computed from the vector representations of studies A and B, normalized to a scale between 0 and 1 to align contrastive loss in model training. This training approach learns vector representations that bring similar data points closer together and push dissimilar data points further apart. The raw text for each study included was truncated to a maximum length of 384 tokens (words) for S-BERT.

Experiment

The FSL model framework was trained separately for each SR project, with training tailored to the specific number of paired samples for each scenario. For instance, in a scenario involving 6 eligible and 60 ineligible studies, the dataset was structured as follows:

- Training data: 4 eligible and 40 ineligible studies were used, resulting in 6 positive pairs and 160 negative pairs (Figure S2 in Multimedia Appendix 1).
- Validation data: 2 eligible and 20 ineligible studies were used, resulting in a total of 41 paired samples.
- Test data: data from the remaining studies were used for testing.

This ensured models were trained and evaluated on datasets specifically tailored to the characteristics of each SR project. The experiments consisted of the following steps:

Determination of the Optimal Number of Eligible Studies for Model Training

For each experimental scenario, the number needed to screen (NNS) and %Reduced workload, also known as work, saved oversampling [37], were estimated using the following formulae:

where TP and FP are true and false positive studies (predicted eligible studies), and N is the total number of studies for each SR. The lower the value for the NNS, the higher the %Reduced workload. The %Reduced workload was estimated by fixing the recall rate at 100% (ie, sensitivity), plotted against the number of eligible studies used for training (N positive) for each scenario. The optimal N positive was estimated using the Kneedle algorithm [38], which identifies the knee point of the graph. This method was chosen because it automatically detects the point where the trend in the data significantly changes, providing a more objective and accurate estimation compared with visual inspection or the elbow method.

A CI [39] for the model performance metric (eg, %Reduced workload, precision, recall, F_1-score) was estimated as follows:

where is the model performance metric, N is the number of studies identified for each SR, Z is a standardized normal distribution, and α is a type I error of .05.

Similarity Threshold

Similarity between studies was assessed using a cosine similarity threshold that represented the distance between 2 vector representations for each study within a paired sample. For example, a positive paired sample in a training set consists of 2 eligible studies and . Each study was transformed into a vector representation with a dimension of 1 × 384. The cosine similarity score [40] was calculated using the following equation:

where are vector representations of the first and second studies within a paired sample.

The cosine similarity score ranges from 1 to –1, where 1 represents perfect similarity between both studies and –1 represents complete dissimilarity enabling quantification of the degree of similarity between pairs of studies and the identification of potentially relevant studies based on their vector representations.

Identification of the optimal similarity threshold was based on the %Reduced workload as described in the previous phase, and the feasibility and information available from each of the SR projects that included 4 to 6 eligible studies and a 10 times greater number of ineligible studies. The data from the remaining studies were used as a query set (ie, test pool).

Following the completion of model training, a support set was retrieved from . Each was again paired with all studies from the query set (ie, Q_1, Q_2, Q_3, …, and Q_n; Figure S4 in Multimedia Appendix 1). The average cosine similarity score for all pairs from the first query study (Q_1) for each was calculated. For instance, the similarity score of Q_1 represented an average cosine similarity score of paired samples between Q_1 and each supporting study set , , , and . The average cosine similarity score for each query study set (ie, Q_1, Q_2, Q_3, …, Q_n) was ordered from highest to lowest, and an optimal threshold was selected to achieve 100% recall (ie, the lowest average similarity score for the eligible studies included).

XSL•FO
RenderX

Prospective Evaluation Phase

Our study prospectively evaluated the FSL framework on four ongoing SR projects (Prospective evaluation SR [PESR] 1-4; Table 1). For each PESR, the principal reviewer initially selected 6 eligible and 60 ineligible studies (4 eligible and 40 ineligible studies for training, and 2 eligible and 20 ineligible studies for validation), leaving the remaining studies in the query set for testing. This initial selection was checked and confirmed by a third senior reviewer and served as the training basis for our FSL framework. The principal reviewer later completed the entire study selection process, providing the "ground truth" for our evaluation.

The FSL framework was applied to rank studies based on average similarity scores. The top-ranked 50% were positive studies representing a potential 50% workload reduction, in contrast to the bottom-ranked 50% which were predicted as negative studies. Positive studies were then allocated to a secondary reviewer for independent screening.

The performance was evaluated with the following 2 steps (Table S1 in Multimedia Appendix 1): (1) comparing the FSL framework model's output (positive and negative studies) with the principal reviewer's selections, demonstrating model performance to identify relevant studies and (2) comparing performance between the secondary reviewer and the principal reviewer's screening based on positive studies by FSL framework. A confusion matrix provided estimates of NNS, recall, precision, and F_1-scores (the harmonic means of precision and recall) for the query set. In addition, false negative rates (FNR), that is, the number of eligible studies misidentified by the FSL framework (not included in the top-ranked 50%, FNR_{FSL}) and those not selected by the secondary reviewer (FNR_{R2}) were also estimated. As the model was trained on abstract and title data only, the evaluation focused solely on the title and abstract screening.

Results

Optimal Number of Eligible Studies (N positive)

Of the 9 SRs, the optimal number of eligible studies (N positive) included in the model training varied between 4 and 12 studies, corresponding to a %Reduced workload of between 80.87% and 99.37% (Figures S5-S8 in Multimedia Appendix 1). A trend for a higher %Reduced workload associated with a higher N positive was observed in SR 1, 3, 4, 5, and 6. In contrast, a higher N positive did not significantly improve %Reduced workload for SRs 2, 7, 8, and 9. However, the optimum N positive ranged between 4 and 12 studies according to the Kneedle algorithm, as represented by the dashed line (Figures S5-S8 in Multimedia Appendix 1). Our findings suggest that the median (range) for the optimum N positive was 9 (4-12) studies, resulting in a median (range) of 95.78% (88.87%-99.37%) %Reduced workload. The ideal N positive for feasible model training ranged from 4 to 6, as indicated by the similarity threshold, which is considered acceptable by most reviewers.

Similarity Threshold

Of the 4 therapeutic SRs (ie, SR1-4), the maximum similarity threshold ranged from 0.439 to 0.617, in line with 100% recall. These threshold values corresponded to a %Reduced workload between 64.81% (95% CI 63.14-66.48) and 96.94% (95% CI 96.12-97.76; Table 2 and Figures S9-S12 in Multimedia Appendix 1). For the prognostic or risk SR5, a similarity threshold of 0.578 was reported equating to a %Reduced workload of 84.16% (95% CI 81.45-86.87). A lower %Reduced workload was observed for prognostic or risk SR6 (51.11%, 95% CI 46.36-55.86 at the similarity threshold of 0.432). For genetic association SRs, the similarity thresholds were 0.546 (SR7) and 0.635 (SR8), corresponding to %Reduced workload of 69.11% (95% CI 66.92-71.30) and 97.67% (95% CI 96.76-98.58), respectively. The similarity threshold and %Reduced workload for the sole economic evaluation SR (SR9) were 0.636 and 95.34% (95% CI 94.32-96.36), respectively. Accordingly, the overall median (range) similarity thresholds and %Reduced workload from our study findings were 0.546 (0.432-0.636) and 84.16% (51.11%-97.67%), respectively.

XSL•FO
RenderX

Table 2. Reduced workload and similarity threshold for each systematic review included.

Type of study and project	Optimal similarity score	Reduced workload, % (95% CI)
Therapeutic study		
SR[a]1 [21]	0.439	65.79 (64.31-67.27)
SR2 [22]	0.617	96.94 (96.12-97.76)
SR3 [23]	0.544	87.46 (86.70-88.22)
SR4 [24]	0.539	64.81 (63.14-66.48)
Prognostic or risk study		
SR5 [25]	0.578	84.16 (81.45-86.87)
SR6 [26]	0.432	51.11 (46.36-55.86)
Genetic association study		
SR7 [27]	0.546	69.11 (66.92-71.30)
SR8 [28]	0.635	97.67 (96.76-98.58)
Economic evaluation study		
SR9 [29]	0.636	95.34 (94.32-96.36)

[a]SR: systematic review.

Prospective Evaluation

The prospective evaluation included four PESRs, with PESR1-3 representing therapeutic studies and PESR4 a prognostic or risk study. The total number of studies included ranged from 1061 to 2136, with 49 to 129 of these considered eligible. For model training, 66 studies (6 eligible and 60 ineligible) were initially selected, leaving 995 to 2070 studies for testing, with 43 to 123 eligible studies. Our model reduced the number of overall studies by approximately 50% equating to a workload reduction of 46.84% to 48.46% (ie, NNS divided by total number of studies). The NNS ranged from 497 to 1035, with 40 to 108 eligible studies (Table S2 in Multimedia Appendix 1). With the aspect of the comparison between the FSL framework and principal reviewer, the FNR_{FSL} varied between 1.87% (95% CI 0-4.44) and 12.2% (95% CI 6.42-17.98), producing a recall rate (REC_{FSL}) of 87.8% (95% CI 82.02-93.58) to 98.13% (95% CI 95.56-100; Table 3 and Table S2 in Multimedia Appendix 1). The comparison of screening results between the secondary reviewer and the principal reviewer based on positive studies by FSL indicated that PESR1 and PESR4 achieved high recall rates (REC_{R2}) of 95% (95% CI 88.25-100) and 88.57% (95% CI 82.48-94.66), respectively, along with FNR_{R2} of 5% (95% CI 0-11.75) and 11.43% (95% CI 5.34-17.52). In contrast, PESR2 and PESR3 indicated high disagreement between both reviewers, with corresponding recall rates (REC_{R2}) of 80.88% (95% CI 71.53-90.23) and 43.52% (95% CI 34.17-52.87), with corresponding FNR_{R2} of 19.12% (95% CI 9.77-28.47) and 56.48% (95% CI 47.13-65.83; Table 3 and Tables S2-S6 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 3. Performance of prospective systematic review evaluation in the test dataset.

Study type and ID	Few-shot learning versus principal reviewer[a]			Secondary reviewer versus principal reviewer[b]		
	$FNR_{Few\text{-}shot\ learning}$[c] (95% CI)	$REC_{Few\text{-}shot\ learning}$[d] (95% CI)	$F_{1Few\text{-}shot\ learning}$[e] (95% CI)	$FNR_{secondary\ reviewer}$[c] (95% CI)	$REC_{secondary\ reviewer}$[d] (95% CI)	$F_{1secondary\ reviewer}$[e] (95% CI)
Therapeutic study						
PESR1[f] [30]	6.98 (0-14.60)	93.02 (85.40-100)	14.83 (11.71-17.95)	5.00 (0-11.75)	95.00 (88.25-100)	95 (88.41-100)
PESR2[f] [31]	6.85 (1.06-12.64)	93.15 (87.36-98.94)	15.29 (12.83-17.75)	19.12 (9.77-28.47)	80.88 (71.53-90.23)	51.64 (43.85-59.43)
PESR3[f] [32]	12.20 (6.42-17.98)	87.80 (82.02-93.58)	18.63 (16.28-20.98)	56.48 (47.13-65.83)	43.52 (34.17-52.87)	51.65 (43.22-60.08)
Prognostic or risk study						
PESR4[f] [33]	1.87 (0-4.44)	98.13 (95.56-100)	23.46 (20.51-26.41)	11.43 (5.34-17.52)	88.57 (82.48-94.66)	88.57 (82.80-94.34)

[a]Evaluation metrics of few-shot learning framework versus principal reviewer, the model evaluation based on the test data.

[b]Evaluation metrics of secondary reviewer versus principal reviewer, the model evaluation based on number needed to screen.

[c]FNR: false negative rate.

[d]REC: recall.

[e]F_1: F_1-score.

[f]PESR: prospective evaluation systematic review.

Discussion

Principal Findings

This study applied an FSL framework to create an automated SR screening tool. The use of 4 to 6 eligible studies for the purpose of model training was sufficient to provide a %Reduced workload between 51.11% and 97.67%, while maintaining 100% recall efficiency. Optimal similarity thresholds varied between 0.432 and 0.636. Paradoxically, increasing the number of eligible studies for model training did not always improve %Reduced workload.

In practice, the principal reviewers undertaking an SR usually perform a preliminary search to identify potentially eligible studies before undertaking a full SR. Considering several eligible studies (4-6 studies) are likely to be identified at this early stage, an FSL approach offers the potential to substantially reduce the subsequent labor-intensive effort required to undertake the SR. The model training within the FSL framework is required once and is generally less time consuming compared with other AI algorithms; for example, existing SR screening tools that use supervised machine learning algorithms with an active learning framework require multiple iterations compared with FSL, which only requires a single training iteration with a small annotated dataset. Subsequently, the trained models are based on those annotated data to predict the remaining studies.

Another aspect of this approach that differs from the existing automated tools involves the use of word embedding (ie, text representation) and classifiers. Several existing tools use text representation techniques that lack context consideration (eg, term frequency), and inverse document frequency with simple classifiers (eg, support vector machine) where performance depends significantly on text representation [6-10]. Instead of using context-free embedding and support vector machine classifiers, this study investigated the feasibility of the use of S-BERT which considers the semantic relationship patterns between a target word and its context before its transformation into a text representation using cosine similarities as a classifier. Thus, FSL frameworks offer improved model performance over currently available SR screening tools.

Among the tools currently available, the performance of Abstrackr has been widely considered with a reported workload reduction between 9.5% and 88.4%, and recall ranging between 79% and 96% [15]. Gates and colleagues also reported a 15%-43% workload reduction with 0%-14% FNR in the subsequent study [18]. Another Abstrackr study [13] reported a 9%-57% workload reduction with associated 0%-0.13% FNR. Abstrackr has also been compared with DistillerSR and RobotAnalyst [16], showing a 40%, 49%, and 35% median workload reduction, respectively. With a 100% principal reviewer recall aim (ie, 0% FNR), Tsou et al [17] found a %Reduced workload of 3.99% to 48.41%, and 8.68% to 60.11% for Abstrackr and EPPI-reviewer, respectively. In this context, the FSL framework used in our study improved performance through a %Reduced workload between 51.11% and 97.67%, under the constraints of 100% recall.

We also further prospectively evaluated and validated our FSL framework in 4 additional PESRs. Our FSL framework achieved a 50% workload reduction, with NNS of 497 to 1035 from a total of 995 to 2070 studies after removing 66 studies for training from the initial identified studies. The FNR_{FSL}, representing eligible studies missed by the model framework, ranged between 1.87% and 12.2%. This performance was assessed in a prospective evaluation where the principal reviewers initially selected 6 eligible and 60 ineligible studies for training and validation. However, this process differed from the model development phase in several key ways. All studies identified

XSL•FO
RenderX

in the development phase were already annotated, whereas only 66 studies were initially annotated when the prospective evaluation experiment began. Guidelines recommend cross-validation between independent reviewer groups for robust SR. Therefore, the potential misidentification of eligible studies by the FSL approach can be mitigated through cross-checking and validation of the principal reviewer's selection. When comparing the performance of secondary reviewers (ie, evaluating based on the top-ranked 50% of studies by FSL) with the principal reviewer, FNR_{R2} ranged from 5% to 56.48%, which is higher than the corresponding FNR_{FSL}, indicating that errors from secondary reviewers contributed significantly to the overall error rate. This emphasizes the importance of training reviewers before and during screening to reduce selection errors. In addition, the involvement of a third senior reviewer is required to resolve these conflicts.

SRs typically require at least 2 independent reviewers for study selection, minimizing potential selection bias as recommended by the PRISMA guidelines [12]. Disagreements between both reviewers are resolved and adjudicated by a third senior reviewer. The aim of our FSL framework is to significantly reduce the workload of the secondary reviewers while still retaining a high recall rate, enabling the secondary reviewer to focus on the studies selected by the FSL framework, rather than having to evaluate all of the studies identified. However, the balance between maximizing workload reduction and minimizing missed eligible studies (ie, FNR_{FSL}) must be carefully considered.

When considering the desired workload reduction, researchers should deliberate the complexity of the review topic, including the type of patients, number of treatments, number of genes, number of exposures, and outcomes of interest, which in turn impacts the number of studies identified, which may be subject to personnel resource constraints. For SRs with many identified studies and limited reviewers, a higher workload reduction may be necessary. Conversely, if multiple reviewers are available, the decreased workload reduction required may be less in order to still maintain a reasonable FNR_{FSL}. Nevertheless, good practice guidelines recommend cross-validation between independent reviewer groups for robust SR. Therefore, the potential misidentification of eligible studies by the FSL approach can be mitigated through cross-checking and validation of the principal reviewer's selection.

The trade-off between workload reduction and FNR_{FSL} can vary depending on the type of SR. To assess this variability, our study included a broad range of SRs including therapeutic, risk or prognostic, genetic association, and economic evaluation. While these SRs share a common process, they often involve different patient populations, interventions (eg, treatments, exposures, genes, and costs), and outcomes of interest. For instance, therapeutic studies typically rely on randomized controlled trials with stringent inclusion criteria, while other study types may use cohort data with more flexible inclusion criteria. Model performance is therefore influenced by the complexity of the subject SRs; for example, therapeutic SRs often have clearer eligibility criteria compared with prognostic or risk SRs. Both conceptual approaches resulted in a workload reduction of 64.81%-96.94% for therapeutic SRs and 51.11%-84.16% for prognostic or risk SRs while maintaining 100% recall. Nevertheless, these thresholds are not absolute and depend on the specific complexity of the SR as highlighted.

Strengths and Limitations

A strength of this study includes the evaluation of the FSL model performance in multiple SR types including therapy, risk or prognosis, genetic association, and economic evaluation. Our FSL framework has the potential to accelerate SR processes during urgent situations, such as public health crises, or in resource-limited settings where efficient resource use is crucial. Furthermore, our framework might be integrated with digital information sources, including eHealth, uHealth, or internet-based medical research, to facilitate study selection for SRs. However, some limitations should be acknowledged. First, the study did not provide model performance in terms of time-saving capabilities. Second, the factors contributing to the worsening recall in the prospective phase were not fully investigated. Both these recognized limitations will be examined in future studies to evaluate the robustness of this approach. Third, the pretrained model requirement was limited to a maximum of 384 words (although, most studies identified across all SRs did not exceed this limit), considering the full texts for each study may not be feasible. Fourth, this study did not evaluate user satisfaction or ease of use. Finally, more SR scenarios are required to prospectively confirm the potential benefits of FSL, with future full-text evaluation compared with the limitations associated with title and abstract screening, although this would likely require the use of high-performance computer processing.

Directions for Future Work

Our future work will focus on the development of a web application with a simplified usable interface to leverage FSL framework algorithms to streamline literature screening for researchers undertaking SRs to reduce user workload and maintain high-quality study selection using AI-assisted solutions. This will include a more comprehensive prospective evaluation across a wider range of SR types. Furthermore, comparisons of FSL approaches with the integration of large language models to automate study selection and data extraction may reduce the most time-consuming and labor-intensive aspects of SRs and will offer estimates for accuracy in both study selection and data extraction. Automation of these key processes will accelerate the SR workflow and enable researchers to complete more comprehensive and robust reviews with potentially more informative translation into clinical and public health practice.

Conclusion

In conclusion, the application of FSL approaches for title and abstract screening for undertaking SRs is clearly feasible. The findings from the retrospective evaluation and validation offer promise, although the balance between workload reduction and FNRs for the identification of eligible studies is one that warrants careful consideration.

XSL•FO
RenderX

Acknowledgments

This manuscript is a part of PW's thesis within the international MSc program in Data Science for Healthcare, conducted at the Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand. This research received financial support from the National Research Council of Thailand (N42A640323). The funding body was not involved in the study's design or conducting.

Authors' Contributions

PW contributed to methodology, software, data curation, formal analysis, and writing the original draft. WP managed conceptualization, data curation, methodology, software, formal analysis, writing the original draft, review and editing, and supervision. PL managed conceptualization, data curation, methodology, software, formal analysis, writing the original draft, review and editing, and supervision. ATa contributed to conceptualization, methodology, software, formal analysis, writing the original draft, review and editing, and supervision. PN conducted conceptualization and supervision. GJM and JA managed review and editing. AT handled conceptualization, methodology, formal analysis, review and editing, supervision, and funding acquisition.

Conflicts of Interest

This study was funded by the National Research Council of Thailand (N42A640323), Thailand. However, the funding agency had no role in the design or conduct of the study.

Multimedia Appendix 1
Additional figures and tables.
[DOCX File , 2089 KB - jmir_v26i1e56863_app1.docx]

Multimedia Appendix 2
Guidelines for developing and reporting machine learning predictive models in biomedical research.
[XLSX File (Microsoft Excel File), 15 KB - jmir_v26i1e56863_app2.xlsx]

References

1. Paul M, Leibovici L. Systematic review or meta-analysis? their place in the evidence hierarchy. Clin Microbiol Infect 2014;20(2):97-100 [FREE Full text] [doi: 10.1111/1469-0691.12489] [Medline: 24354996]
2. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med 2016;21(4):125-127 [FREE Full text] [doi: 10.1136/ebmed-2016-110401] [Medline: 27339128]
3. Hoffmann F, Allers K, Rombey T, Helbach J, Hoffmann A, Mathes T, et al. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol 2021;138:1-11 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.05.022] [Medline: 34091022]
4. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7(2):e012545 [FREE Full text] [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
5. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? a survival analysis. Ann Intern Med 2007;147(4):224-233 [FREE Full text] [doi: 10.7326/0003-4819-147-4-200708210-00179] [Medline: 17638714]
6. Wallace B, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. : Association for Computing Machinery; 2012 Presented at: IHI '12: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; January 28-30, 2012; Miami, FL p. 819-824. [doi: 10.1145/2110363.2110464]
7. Smarter reviews: trusted evidence.: DistillerSR URL: https://www.distillersr.com/ [accessed 2024-10-16]
8. EPPI-reviewer. URL: https://eppi.ioe.ac.uk/EPPIReviewer-Web/home [accessed 2024-08-24]
9. Faster systematic reviews.: Rayyan URL: https://www.rayyan.ai/ [accessed 2024-10-16]
10. The world's #1 systematic review tool.: Covidence URL: https://www.covidence.org/ [accessed 2024-10-16]
11. Gartlehner G, Wagner G, Lux L, Affengruber L, Dobrescu A, Kaminski-Hartenthaler A, et al. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. Syst Rev 2019;8(1):277 [FREE Full text] [doi: 10.1186/s13643-019-1221-3] [Medline: 31727159]
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
13. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev 2015;4:80 [FREE Full text] [doi: 10.1186/s13643-015-0067-6] [Medline: 26073974]

XSL•FO
RenderX

14. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
15. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018;7(1):45 [FREE Full text] [doi: 10.1186/s13643-018-0707-8] [Medline: 29530097]
16. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev 2019;8(1):278 [FREE Full text] [doi: 10.1186/s13643-019-1222-2] [Medline: 31727150]
17. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev 2020;9(1):73 [FREE Full text] [doi: 10.1186/s13643-020-01324-7] [Medline: 32241297]
18. Gates A, Gates M, Sebastianski M, Guitard S, Elliott SA, Hartling L. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. BMC Med Res Methodol 2020;20(1):139 [FREE Full text] [doi: 10.1186/s12874-020-01031-w] [Medline: 32493228]
19. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples. ACM Comput Surv 2020;53(3):1-34. [doi: 10.1145/3386252]
20. Ge Y, Guo Y, Das S, Al-Garadi M, Sarker A. Few-shot learning for medical text: a review of advances, trends, and opportunities. J Biomed Inform 2023 Aug;144:104458 [FREE Full text] [doi: 10.1016/j.jbi.2023.104458] [Medline: 37488023]
21. Tansawet A, Numthavaj P, Techapongsatorn S, Wilasrusmee C, Attia J, Thakkinstian A. Mesh position for hernia prophylaxis after midline laparotomy: a systematic review and network meta-analysis of randomized clinical trials. Int J Surg 2020;83:144-151 [FREE Full text] [doi: 10.1016/j.ijsu.2020.08.059] [Medline: 32927135]
22. Poprom N, Numthavaj P, Wilasrusmee C, Rattanasiri S, Attia J, McEvoy M, et al. The efficacy of antibiotic treatment versus surgical treatment of uncomplicated acute appendicitis: systematic review and network meta-analysis of randomized controlled trial. Am J Surg 2019;218(1):192-200 [FREE Full text] [doi: 10.1016/j.amjsurg.2018.10.009] [Medline: 30340760]
23. Sapankaew T, Thadanipon K, Ruenroengbun N, Chaiyakittisopon K, Ingsathit A, Numthavaj P, et al. Efficacy and safety of urate-lowering agents in asymptomatic hyperuricemia: systematic review and network meta-analysis of randomized controlled trials. BMC Nephrol 2022;23(1):223 [FREE Full text] [doi: 10.1186/s12882-022-02850-3] [Medline: 35739495]
24. Ruenroengbun N, Numthavaj P, Sapankaew T, Chaiyakittisopon K, Ingsathit A, Mckay GJ, et al. Efficacy and safety of conventional antiviral agents in preventive strategies for cytomegalovirus infection after kidney transplantation: a systematic review and network meta-analysis. Transpl Int 2021;34(12):2720-2734 [FREE Full text] [doi: 10.1111/tri.14122] [Medline: 34580930]
25. Isnuwardana R, Bijukchhe S, Thadanipon K, Ingsathit A, Thakkinstian A. Association between vitamin D and uric acid in adults: a systematic review and meta-analysis. Horm Metab Res 2020;52(10):732-741 [FREE Full text] [doi: 10.1055/a-1240-5850] [Medline: 33049785]
26. Saputro SA, Pattanaprateep O, Pattanateepapon A, Karmacharya S, Thakkinstian A. Prognostic models of diabetic microvascular complications: a systematic review and meta-analysis. Syst Rev 2021;10(1):288 [FREE Full text] [doi: 10.1186/s13643-021-01841-z] [Medline: 34724973]
27. Lukkunaprasit T, Rattanasiri S, Turongkaravee S, Suvannang N, Ingsathit A, Attia J, et al. The association between genetic polymorphisms in ABCG2 and SLC2A9 and urate: an updated systematic review and meta-analysis. BMC Med Genet 2020;21(1):210 [FREE Full text] [doi: 10.1186/s12881-020-01147-2] [Medline: 33087043]
28. Bassey P, Numthavaj P, Rattanasiri S, Sritara P, McEvoy M, Ongphiphadhanakul B, et al. Causal association pathways between fetuin-A and kidney function: a mediation analysis. J Int Med Res 2022;50(4):3000605221082874 [FREE Full text] [doi: 10.1177/03000605221082874] [Medline: 35435033]
29. Chaiyakittisopon K, Pattanaprateep O, Ruenroengbun N, Sapankaew T, Ingsathit A, Mckay GJ, et al. Evaluation of the cost-utility of phosphate binders as a treatment option for hyperphosphatemia in chronic kidney disease patients: a systematic review and meta-analysis of the economic evaluations. Eur J Health Econ 2021;22(4):571-584 [FREE Full text] [doi: 10.1007/s10198-021-01275-3] [Medline: 33677736]
30. Sa-nguansai S, Rattanasiri S, Pornsuriyasak P, Numthavaj P, McKay GJ, Attia J, et al. Efficacy of targeted therapy or immunotherapy as adjuvant treatment for non-small cell lung cancer: a systematic review and network meta-analysis. SSRN. Preprint posted online on January 8, 2024 2024 [FREE Full text]
31. Keesukphan A, Suntipap M, Thadanipon K, Boonmanunt S, Thakkinstian A. Effects of repetitive peripheral magnetic stimulation on upper extremity function after stroke: a systematic review and meta-analysis. 2023 Apr 28 Presented at: RSU International Research Conference 2023; Pathum Thani, Thailand p. 126-134 URL: https://rsucon.rsu.ac.th/files/proceedings/RSUSCI2023/IN23-038.pdf
32. Pornsuriyasak P, Thadanipon K, Sa-nguansai S, Numthavej P, McKay GJ, Attia J, et al. Regular versus as-needed treatments for mild asthma in children, adolescents, and adults: a systematic review and network meta-analysis. Research Square. Preprint posted online on September 2, 2024 2024 [FREE Full text]

XSL•FO
RenderX

33. Suntipap M, Keesukphan A, Rattnasiri S, Numthavaj P, Thakkinstian A. The association between cervical sonography and successful induction of labor: a systematic review and meta-analysis. 2023 Presented at: 10th National and International Virtual Conference on Multidisciplinary Research; November 27, 2023; Dire Dawa University, Ethiopia URL: https://so09.tci-thaijo.org/index.php/PMR/article/view/3443
34. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on October 11, 2018 2018:4805 [FREE Full text]
35. Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for One-shot image recognition. Ontario, Canada: University of Toronto; 2015. URL: https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf [accessed 2024-10-22]
36. Models.: Hugging Face URL: https://huggingface.co/models?library=sentence-transformers [accessed 2024-10-16]
37. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13(2):206-219 [FREE Full text] [doi: 10.1197/jamia.M1929] [Medline: 16357352]
38. Satopää V, Albrecht J, Irwin D, Raghavan B. Finding a "Kneedle" in a haystack: detecting knee points in system behavior. : IEEE; 2011 Presented at: 31st International Conference on Distributed Computing Systems Workshops; June 20-24, 2011; Minneapolis, USA. [doi: 10.1109/icdcsw.2011.20]
39. Edwardes M. Confidence intervals for a binomial proportion by S. E. Vollset, Statistics in Medicine, 12, 809-824 (1993). Stat Med 1994;13(16):1693-1698. [doi: 10.1002/sim.4780131609] [Medline: 7973244]
40. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM 1975;18(11):613-620 [FREE Full text] [doi: 10.1145/361219.361220]

Abbreviations

AI: artificial intelligence
FNR: false negative rate
FSL: few-shot learning
NNS: number needed to screen
PESR: prospective evaluation systematic review
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
S-BERT: sentence-bidirectional encoder representations from transformers
SR: systematic review

Edited by N Cahill, T Leung; submitted 29.01.24; peer-reviewed by L Guo, N Nama; comments to author 30.04.24; revised version received 23.06.24; accepted 03.10.24; published 11.12.24.

Please cite as:
Wiwatthanasetthakarn P, Ponthongmak W, Looareesuwan P, Tansawet A, Numthavaj P, McKay GJ, Attia J, Thakkinstian A
Development and Validation of a Literature Screening Tool: Few-Shot Learning Approach in Systematic Reviews
J Med Internet Res 2024;26:e56863
URL: https://www.jmir.org/2024/1/e56863
doi:10.2196/56863
PMID:39662894

©Phongphat Wiwatthanasetthakarn, Wanchana Ponthongmak, Panu Looareesuwan, Amarit Tansawet, Pawin Numthavaj, Gareth J McKay, John Attia, Ammarin Thakkinstian. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating Older Adults' Perceptions of AI Tools for Medication Decisions: Vignette-Based Experimental Survey

Sarah E Vordenberg[1], PharmD, MPH; Julianna Nichols[1], BS; Vincent D Marshall[1], MS; Kristie Rebecca Weir[2,3], PhD; Michael P Dorsch[1], PharmD, MS

[1]College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

[2]Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia

[3]Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland

Corresponding Author:
Sarah E Vordenberg, PharmD, MPH
College of Pharmacy
University of Michigan
428 Church St
Ann Arbor, MI, 48109
United States
Phone: 1 734 763 6691
Email: skelling@med.umich.edu

Abstract

Background: Given the public release of large language models, research is needed to explore whether older adults would be receptive to personalized medication advice given by artificial intelligence (AI) tools.

Objective: This study aims to identify predictors of the likelihood of older adults stopping a medication and the influence of the source of the information.

Methods: We conducted a web-based experimental survey in which US participants aged ≥65 years were asked to report their likelihood of stopping a medication based on the source of information using a 6-point Likert scale (scale anchors: 1=not at all likely; 6=extremely likely). In total, 3 medications were presented in a randomized order: aspirin (risk of bleeding), ranitidine (cancer-causing chemical), or simvastatin (lack of benefit with age). In total, 5 sources of information were presented: primary care provider (PCP), pharmacist, AI that connects with the electronic health record (EHR) and provides advice to the PCP ("EHR-PCP"), AI with EHR access that directly provides advice ("EHR-Direct"), and AI that asks questions to provide advice ("Questions-Direct") directly. We calculated descriptive statistics to identify participants who were extremely likely (score 6) to stop the medication and used logistic regression to identify demographic predictors of being likely (scores 4-6) as opposed to unlikely (scores 1-3) to stop a medication.

Results: Older adults (n=1245) reported being extremely likely to stop a medication based on a PCP's recommendation (n=748, 60.1% [aspirin] to n=858, 68.9% [ranitidine]) compared to a pharmacist (n=227, 18.2% [simvastatin] to n=361, 29% [ranitidine]). They were infrequently extremely likely to stop a medication when recommended by AI (EHR-PCP: n=182, 14.6% [aspirin] to n=289, 23.2% [ranitidine]; EHR-Direct: n=118, 9.5% [simvastatin] to n=212, 17% [ranitidine]; Questions-Direct: n=121, 9.7% [aspirin] to n=204, 16.4% [ranitidine]). In adjusted analyses, characteristics that increased the likelihood of following an AI recommendation included being Black or African American as compared to White (Questions-Direct: odds ratio [OR] 1.28, 95% CI 1.06-1.54 to EHR-PCP: OR 1.42, 95% CI 1.17-1.73), having higher self-reported health (EHR-PCP: OR 1.09, 95% CI 1.01-1.18 to EHR-Direct: OR 1.13 95%, CI 1.05-1.23), having higher confidence in using an EHR (Questions-Direct: OR 1.36, 95% CI 1.16-1.58 to EHR-PCP: OR 1.55, 95% CI 1.33-1.80), and having higher confidence using apps (EHR-Direct: OR 1.38, 95% CI 1.18-1.62 to EHR-PCP: OR 1.49, 95% CI 1.27-1.74). Older adults with higher health literacy were less likely to stop a medication when recommended by AI (EHR-PCP: OR 0.81, 95% CI 0.75-0.88 to EHR-Direct: OR 0.85, 95% CI 0.78-0.92).

Conclusions: Older adults have reservations about following an AI recommendation to stop a medication. However, individuals who are Black or African American, have higher self-reported health, or have higher confidence in using an EHR or apps may be receptive to AI-based medication recommendations.

(J Med Internet Res 2024;26:e60794) doi:10.2196/60794

XSL•FO
RenderX

KEYWORDS

older adults; survey; decisions; artificial intelligence; vignette; drug; pharmacology; pharmaceutic; medication; decision-making; geriatric; aging; surveys; attitude; perception; perspective; recommendation; electronic heath record

Introduction

Approximately 90% of adults 65 years and older ("older adults") in the United States take at least 1 prescription medication, two-thirds take at least 1 dietary supplement, and one-third take at least 1 over-the-counter (OTC) medication regularly [1,2]. Medications are an important tool to prevent and treat diseases. However, up to one-half of older adults take at least 1 inappropriate medication, which can lead to drug-drug interactions, adverse events, additional medical appointments, emergency department visits, functional decline, reduced quality of life, and increased health care costs [3].

Older adults frequently report that they prefer to participate in shared decision-making [4]. Therefore, they may discuss prescription or OTC medications with physicians or pharmacists as they trust the information provided [5,6]. While these conversations allow patients to receive personalized recommendations, many health care professionals are experiencing high workload and burnout, which interferes with their ability to engage in patient-centered conversations about medications [7,8]. Older adults are increasingly using technology such as smartphone apps and telehealth [9]. Therefore, technology may be an alternative or complementary strategy for older adults to receive personalized medication advice, particularly given the limited time that health care professionals have available with each patient [10].

SEV and MPD sought to address the gap in technological tools for patients related to OTC medications by creating an electronic decision algorithm prototype to help adults identify safe and effective treatments for common ailments [11]. The tool required that participants answer questions about themselves and their symptoms and were given an OTC product recommendation or encouraged to seek advice from a health care professional. Given the complexity and time-intensiveness of manually building and maintaining a comprehensive database, it did not contain detailed information about individual medications (eg, dosage forms, directions for use, and side effects). When the prototype was shared with adults during semistructured interviews, participants reported that they would likely use the technology. However, they simultaneously raised concerns about the accuracy and comprehensiveness of the information, data storage, and accessibility while aligning with broader social and ethical concerns about using technology [12].

While adults reported being interested in using technology to support medication decisions, SEV and MPD quickly identified that creating and maintaining a knowledge base for more than 300,000 OTC medications in the United States would not be feasible [13]. However, the public release of a large language model (LLM) in 2022 revealed how artificial intelligence (AI) can revolutionize how diseases are diagnosed, develop patient-centered treatment plans, and increase patient engagement [14,15]. For example, it is able to provide high-quality, empathetic responses to patient questions [16]. It can also readily provide clarifying information and adjust the health literacy level of the information. Therefore, we shifted our work to identify predictors of the likelihood of older adults starting or stopping a medication and the influence of the source of the information.

Methods

Study Design and Sample

We developed a 15-minute web-based experimental survey informed by our previous research, which focused on factors important to older adults when making medication decisions [17-19]. We used Qualtrics panels, which include more than 3 million people, to distribute the survey. We initially pilot-tested the survey by distributing it to 50 older adults. We did not make any changes to the survey. Overall, we aimed to recruit 1200 US adults aged 65 years and older to participate between December 2023 and January 2024. We used quotas for gender (50% female), age (50% aged 65-69 years and 50% aged 70 years and older), race (15% Black, 5% Asian, and 80% White or another race), ethnicity (18% Hispanic), and education (70% less than a bachelor's degree). We aimed to align our quotas with the prevalence of these groups in the US population. The sampling algorithm continued to invite eligible participants to complete the survey until all quotas were achieved. Survey invitations did not include the study topic to decrease self-selection bias, and strategies such as checking IP addresses were used to prevent individuals from submitting multiple responses.

Intervention

First, we defined 5 potential sources of information for learning about OTC and prescription medications (Table 1). These included 2 health care professionals (ie, primary care provider [PCP] or pharmacist at a drugstore) and 3 versions of an AI app (ie, connect with electronic health record [EHR] and sends medication recommendations to PCP for approval before sending it to the patient ["EHR-PCP"], connects with EHR and sends a medication recommendation directly to the patient ["EHR-Direct"], or collects information by asking the patient a question and then provides a medication recommendation directly to the patient ["Questions-Direct"]). We developed the descriptions of the AI apps based on semistructured interviews that we had previously conducted with adults to explore their interest in using technology when making OTC medication decisions [11].

XSL•FO
RenderX

Table 1. Study design[a].

Variable	Text description
Source information	There are many ways that you can learn about *over-the-counter and prescription medications*. You may receive a recommendation from a health care professional: • You might have an appointment with or message a primary care provider (PCP) who has access to your electronic health record. • You might ask a pharmacist at a drugstore who does not have access to your electronic health record. [Page break] Alternatively, you may receive a recommendation from an application ('app') that uses *artificial intelligence* (AI) and all of the available *research evidence* to make personalized recommendations for you. • One type of app may gather information from your electronic health record such as your health conditions and medications to create a personalized recommendation. The recommendation would be shared with your doctor for approval before your receive the information. [EHR-PCP] • A second type of app may gather information from your electronic health record such as your health conditions and medications to create a personalized recommendation. You would receive the recommendation immediately and a copy of the recommendation would be stored in your electronic health record. [EHR-Direct] • A third type of app may gather information from you as you answer a series of questions about your health conditions and medications to create a personalized recommendation. You would receive the recommendation immediately. [Questions-Direct]
Medication	Imagine that you are interested in simplifying your medication regimen to avoid unnecessary medications and decrease the risk of side effects. Consider the following three situations and identify how likely you would be to *stop* the medication if recommended by each of the following sources of information. *[randomized order with page breaks]* • Imagine that you have taken low-dose aspirin once daily for several years to prevent you from having a heart attack or stroke. You have never had a heart attack or stroke. However, you recently learned that aspirin can increase your risk of bleeding. • Imagine that you have taken ranitidine once daily for several years to treat your heartburn symptoms. Your symptoms are well controlled. However, you recently learned that ranitidine contains a cancer-causing chemical. • Imagine that you have taken prescription simvastatin once daily for several years to decrease your risk of a heart attack or stroke. Your cholesterol lab values are within the normal range. However, you recently learned that it may not provide as much benefit as you age. You are looking for advice as to whether this information means you should stop taking [medication] How likely would you be to stop the [medication] if recommended by [source of information]? 1 = Not at all likely 2 3 4 5 6 = Extreme likely

[a]Participants received 3 vignettes (ie, aspirin, ranitidine, and simvastatin) in a randomized order. For each vignette, participants indicated how likely they would be to stop the medication if recommended by 5 different sources of information (ie, primary care provider, a pharmacist at a drugstore, "EHR-PCP," "EHR-Direct," and "Questions-Direct").

We presented 3 medication vignettes (ie, aspirin, ranitidine, and simvastatin) to each participant in a randomized order. Aspirin was selected given it is widely accessible as an OTC medication, taken by approximately one-third of older adults without cardiovascular disease, yet not recommended by clinical guidelines given the risks of bleeding [20,21]. Ranitidine was an OTC histamine-2 receptor antagonist, which was withdrawn from the US market in 2020 after being a best-selling medication in the United States, given that it contained a probable carcinogen [22,23]. Simvastatin is a prescription hydroxymethylglutaryl-CoA reductase inhibitor recommended for the primary prevention of cardiovascular events [24,25]. Participants were not able to go backward in the study due to the randomization of the vignettes.

Participants were asked to rate the likelihood that they would stop each medication based on the source of information (6-point Likert scale with "Not at all likely (1)" and "Extremely likely (6)" as scale anchors). We subsequently asked participants to think about the AI app that connects with the EHR and provides immediate, personalized advice about OTC medications with a recommendation stored in the EHR. Participants were asked to rate their level of agreement with statements related to access, trust, nervousness, confusion and discomfort, and preferring to use the app before reaching out to the PCP (5-point Likert scale with "1=Strongly disagree" and "2=Strongly agree (5)" as scale anchors).

Demographic Characteristics

We collected information about participants' age, gender, race, ethnicity, and highest level of education. Participants reported

XSL•FO
RenderX

their health status using a 5-point Likert scale with response options of “Poor (1),” “Fair (2),” “Good (3),” “Very good (4),” and “Excellent (5),” and their health literacy level was measured by their level of confidence filling out medical forms with response options of “Not at all (1),” “A little bit (2),” “Somewhat (3),” “Quite a bit (4),” and “Excellent (5)” [26-28]. We also asked about their confidence using various forms of technology such as the internet, EHR, and apps (5-point Likert scale with “Not at all confident (1)” and “Very confident (5)” as scale anchors). We included 5 items modified from the technology acceptance model (ie, whether the AI app increases access, trustworthiness, nervousness, confusion and discomfort, and is preferred to contacting a PCP) [29,30].

Ethical Considerations

The University of Michigan Institutional Review Board deemed this study exempt (HUM00246349). The informed consent information was included on the first page of the survey. All questions were required; however, participants could exit the survey anytime, and all data were collected anonymously. Prospective participants were told the anticipated duration of the survey, where and how long data would be stored, the investigator, and the purpose of the survey. Participants were compensated based on the terms of their panel agreement.

Statistical Analysis

We calculated descriptive statistics and reported the percentage of participants who reported being extremely likely to stop a medication (score of 6) based on the source of information. High Cronbach α scores indicated that participants responded consistently across medications by the source of information. Therefore, we created an average medication score by the source of information per participant. We used logistic regression to explore demographic characteristics that predict a high likelihood of stopping medications (scores 4-6) compared to those with low scores (scores 1-3). We used linear regression with continuous integer values between “Not at all confident (1)” and “Very confident (5)” to explore demographic characteristics that predict attitudes toward using the Questions-Direct-OTC app. All regression models included the demographic characteristics of age; gender; race; ethnicity; education; health status; health literacy; and confidence using the internet, EHR, and apps. We used a statistical significance level of $P<.05$. All analyses were conducted with R statistical software (version 4.2.2). We reported our study using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist and Checklist for Reporting Results of Internet E-Surveys (CHERIES) [31,32].

Results

The survey was opened by 1318 individuals, of whom 73 participants were excluded due to not meeting the inclusion criteria (eg, aged younger than 65 years) or demographic quotas being filled. The final sample for analysis included 1245 respondents (94% completion rate).

Demographic Characteristics

Half (n=625, 50%) of the 1245 participants were female, and the median age was 70 (IQR 67-74 years; Table 2). Participants most often reported earning a high school diploma or less (n=391, 32%) or attending trade school, some college, or earning an associate’s degree (n=484, 39%). Participants most frequently reported having good health (n=595, 48%) and being extremely comfortable filling out medical forms. Most participants were confident making phone calls (n=1070, 86%), sending and receiving emails (n=1025, 82%) and SMS text messages (n=992, 80%), and using the internet (n=906, 73%). Fewer participants were confident using an EHR (n=790, 63%), apps (n=769, 62%), social media (n=607, 49%), or wearable devices (n=536, 43%).

XSL•FO
RenderX

Table 2. Demographic characteristics of older adult participants (n=1245).

Demographic characteristics	Value
Age (years), m edian (IQR)	70 (67-74)
Gender, n (%)	
Female	625 (50)
Male	619 (50)
Transgender	1 (0)
Race, n (%)	
White	894 (72)
Black or African American	195 (16)
Asian or Asian American	61 (5)
Another race	95 (8)
Ethnicity, n (%)	
Hispanic	229 (18)
Education, n (%)	
High school diploma or less	391 (32)
Trade school, some college, or associate's degree	484 (39)
Bachelor's degree	240 (19)
Master's degree or higher	130 (10)
Health status, n (%)	
Excellent	55 (4)
Very good	282 (23)
Good	595 (48)
Fair	267 (21)
Poor	46 (4)
Confidence filling out medical forms, n (%)	
Extremely	662 (53)
Quite a bit	376 (30)
Somewhat	142 (11)
A little bit	45 (4)
Not at all	20 (2)
Confidence using technology[a], n (%)	
Phone calls	1070 (86)
Emails	1025 (82)
SMS text messages	992 (80)
Internet	906 (73)
Electronic health record	790 (63)
Apps	769 (62)
Social media	607 (49)
Wearable devices	536 (43)

[a]Participants who selected scores of 4 or 5 on a 5-point Likert scale with scale anchors 1=not at all confident and 5=very confident.

XSL•FO
RenderX

Likelihood of Stopping Medications

Older adults (n=1245) more frequently reported being extremely likely to stop a medication when the recommendation came from a PCP (range from n=748, 60.1% for aspirin due to a risk of bleeding to n=858, 68.9% for ranitidine due to a cancer-causing chemical; Table 3). In contrast, older adults were less likely to say that they would follow the recommendation from a pharmacist at a drugstore (extremely likely ranging from n=227, 18.2% for simvastatin due to a lack of benefit with age to n=361, 29% for ranitidine).

Older adults were least likely to agree with stopping medication when the recommendation came from an AI app. Between 14.6% (n=182; aspirin) and 23.2% (n=289; ranitidine) of older adults were extremely likely to stop the medication when recommended by the EHR-PCP app. Regardless of integration within the EHR, when the AI app did not include approval by a PCP, only ~10% (n=130, 10.4% for aspirin and n=132, 10.6% simvastatin) and 17% (n=212; ranitidine) of older adults were extremely likely to stop the medication.

In adjusted analyses, older adults who were female as opposed to male were more likely to stop a medication when the recommendation came from a PCP (odds ratio [OR] 1.55, 95% CI 1.18-2.03) or pharmacist (OR 1.27, 95% CI 1.10-1.47; Table 4).

Compared to older adults who were White, older adults who were Black or African American were less likely to stop a medication when the recommendation came from a PCP (OR 0.51, 95% CI 0.37-0.71); however, they were more likely to stop a medication when recommended by the EHR-PCP app (OR 1.42, 95% CI 1.17-1.73), EHR-Direct app (OR 1.38, 95% CI 1.14-1.66), or Questions-Direct app (OR 1.28, 95% CI 1.06-1.54).

Compared to older adults who were White, older adults who identified as another race were less likely to stop a medication when the recommendation came from a pharmacist (OR 0.71, 95% CI 0.57-0.88). However, they were more likely to stop a medication when the recommendation came from the EHR-direct app (OR 1.23, 95% CI 1.00-1.51).

Compared with older adults with less than a bachelor's degree, older adults with a bachelor's degree were more likely to stop a medication when the recommendation came from a PCP (OR 1.71, 95% CI 1.23-2.38) or the EHR-PCP app (OR 1.36, 95% CI 1.17-1.58).

There was no difference in the likelihood of stopping a medication by health status when the recommendation came from a PCP. However, higher health status was associated with an increased likelihood of stopping the medication for the other 4 sources, ranging from the EHR-PCP app (OR 1.09, 95% CI 1.01-1.18) to the EHR-Direct app (OR 1.13, 95% CI 1.05-1.23). In contrast, as health literacy increased, older adults were more likely to stop a medication when recommended by a PCP (OR 1.19, 95% CI 1.04-1.36) and less likely to stop a medication when it was recommended by the EHR-PCP app (OR 0.81, 95% CI 0.75-0.88), EHR-Direct app (OR 0.85, 95% CI 0.78-0.92), or Questions-Direct app (OR 0.83, 95% CI 0.77-0.90).

There was no difference in the likelihood of stopping a medication by confidence using the internet except for recommendations by a pharmacist at a drugstore (OR 1.31, 95% CI 1.10-1.56). However, older adults who were more confident using the EHR were more likely to stop a medication when it was recommended by the EHR-PCP app (OR 1.55, 95% CI 1.33-1.80), EHR-Direct app (OR 1.44, 95% CI 1.24-1.67), or Questions-Direct app (OR 1.36, 95% CI 1.16-1.58). Similarly, older adults who were more confident using apps were more likely to stop a medication when it was recommended by the EHR-PCP app (OR 1.49, 95% CI 1.27-1.74), EHR-Direct app (OR 1.38, 95% CI 1.18-1.62), or Questions-Direct app (OR 1.39, 95% CI 1.19-1.64).

XSL•FO
RenderX

Table 3. Older adult's likelihood of stopping a medication based on the source of the recommendation and score (n=1245)[a].

Source of information	Low-dose aspirin (risk of bleeding), n (%)	Ranitidine (cancer-causing chemical), n (%)	Simvastatin (lack of benefit with age), n (%)
PCP[b]			
1=Not at all likely	42 (3.4)	21 (1.7)	20 (1.6)
2	22 (1.8)	8 (0.6)	15 (1.2)
3	51 (4.1)	35 (2.8)	43 (3.5)
4	104 (8.4)	78 (6.3)	100 (8)
5	278 (22.3)	245 (19.7)	279 (22.4)
6=Extremely likely	748 (60.1)	858 (68.9)	788 (63.3)
Pharmacist at a drugstore			
1=Not at all likely	137 (11)	104 (8.4)	122 (9.8)
2	105 (8.4)	68 (5.5)	105 (8.4)
3	190 (15.3)	156 (12.5)	218 (17.5)
4	294 (23.6)	261 (21)	292 (23.5)
5	286 (23)	295 (23.7)	281 (22.6)
6=Extremely likely	233 (18.7)	361 (29)	227 (18.2)
AI[c] app that connects with EHR[d] and provides personalized advice to PCP for approval prior to sharing with you			
1=Not at all likely	270 (21.7)	230 (18.5)	266 (21.4)
2	122 (9.8)	95 (7.6)	117 (9.4)
3	200 (16.1)	178 (14.3)	193 (15.5)
4	237 (19)	214 (17.2)	252 (20.2)
5	234 (18.8)	239 (19.2)	217 (17.4)
6=Extremely likely	182 (14.6)	289 (23.2)	200 (16.1)
AI app with EHR access that provides personalized advice directly to you			
1=Not at all likely	323 (25.9)	270 (21.7)	310 (24.9)
2	151 (12.1)	135 (10.8)	175 (14.1)
3	261 (21)	218 (17.5)	242 (19.4)
4	231 (18.6)	237 (19)	243 (19.5)
5	149 (12)	173 (13.9)	157 (12.6)
6=Extremely likely	130 (10.4)	212 (17)	118 (9.5)
AI app that collects information by asking you questions and provides personalized advice directly to you			
1=Not at all likely	337 (27.1)	281 (22.6)	318 (25.5)
2	169 (13.6)	151 (21.1)	165 (13.3)
3	248 (19.9)	222 (17.8)	271 (21.8)
4	228 (18.3)	208 (16.7)	240 (19.3)
5	142 (11.4)	179 (14.4)	119 (9.6)
6=Extremely likely	121 (9.7)	204 (16.4)	132 (10.6)

[a]This table shows the percentage of participants who selected each score on a 6-point Likert scale with 1=not at all likely and 6=extremely likely to stop the medication as scale anchors.

[b]PCP: primary care provider.

[c]AI: artificial intelligence.

[d]EHR: electronic health record.

XSL•FO
RenderX

Table 4. Demographic predictors of being likely to stop a medication based on the source of the recommendation using logistic regression (n=1245)[a].

	PCP[b]	Pharmacist at a drugstore	AI[c] app with EHR[d] access and PCP approval[e]	AI app with EHR access and direct recommendation[f]	AI app without EHR access and direct recommendation[g]
Age, OR[h] (95% CI)	1.03 (1.00-1.06)	0.99 (0.98-1.00)	1.00 (0.99-1.01)	1.01 (1.00-1.03)	1.01 (1.00-1.02)
Gender, OR (95% CI)					
Male	REF[i]	REF	REF	REF	REF
Female	1.55 (1.18-2.03)	1.27 (1.10-1.47)	1.14 (0.99-1.31)	1.01 (0.88-1.16)	1.04 (0.91-1.20)
Race, OR (95% CI)					
White	REF	REF	REF	REF	REF
Black of African American	0.51 (0.37-0.71)	0.90 (0.74-1.10)	1.42 (1.17-1.73)	1.38 (1.14-1.66)	1.28 (1.06-1.54)
Another race	0.93 (0.60-1.41)	0.71 (0.57-0.88)	1.21 (0.98-1.49)	1.23 (1.00-1.51)	1.17 (0.95-1.44)
Ethnicity, OR (95% CI)					
Non-Hispanic	REF	REF	REF	REF	REF
Hispanic	0.88 (0.61-1.27)	1.15 (0.95-1.40)	0.98 (0.82-1.17)	0.96 (0.81-1.15)	1.04 (0.87-1.24)
Education, OR (95% CI)					
<Bachelor's degree	REF	REF	REF	REF	REF
≥Bachelor's degree	1.71 (1.23-2.38)	0.97 (0.83-1.14)	1.36 (1.17-1.58)	1.14 (0.98-1.32)	1.05 (0.91-1.22)
Health status	1.03 (0.89-1.21)	1.13 (1.04-1.22)	1.09 (1.01-1.18)	1.13 (1.05-1.23)	1.13 (1.05-1.22)
Health literacy	1.19 (1.04-1.36)	1.02 (0.94-1.11)	0.81 (0.75-0.88)	0.85 (0.78-0.92)	0.83 (0.77-0.90)
Confidence using internet[j], OR (95% CI)					
Low	REF	REF	REF	REF	REF
High	1.36 (1.00-1.85)	1.31 (1.10-1.56)	1.16 (0.98-1.37)	1.06 (0.90-1.26)	1.13 (0.96-1.34)
Confidence using electronic health record[j], OR (95% CI)					
Low	REF	REF	REF	REF	REF
High	1.22 (0.92-1.63)	1.13 (0.96-1.32)	1.55 (1.33-1.80)	1.44 (1.24-1.67)	1.36 (1.16-1.58)
Confidence using apps[j], OR (95% CI)					
Low	REF	REF	REF	REF	REF
High	1.30 (0.95-1.78)	1.13 (0.96-1.32)	1.49 (1.27-1.74)	1.38 (1.18-1.62)	1.39 (1.19-1.64)

[a]6-point Likert scale: 1=not at all likely, 6=extremely likely; participants with scores 4-6 were classified as "likely."
[b]PCP: primary care provider.
[c]AI: artificial intelligence.
[d]EHR: electronic health record.
[e]AI app that connects with your EHR and provides personalized advice to your PCP for approval before sharing the recommendation with you.
[f]AI app that connects with your EHR and provides personalized advice directly to you.
[g]AI app that collects information by asking you questions and provides personalized advice directly to you.
[h]OR: odds ratio.
[i]REF: reference category.
[j]5-Point Likert scale: 1=not at all confident, 5=very confident; participants were categorized as no (scores 1-3) or yes (scores 4-5).

Attitudes Toward Using an AI App

We asked participants about their attitudes toward using an AI app with EHR access to provide immediate, personalized advice about OTC medications (Table 5). Compared to participants who were White, participants who were Black or African American were more likely to report that this technology would improve access to information (=0.32, 95% CI 0.13-0.51), they would trust the personalized advice (=0.24, 95% CI 0.06-0.43), and it would not lead to confusion or discomfort (=0.21, 95% CI 0.00-0.42) as it relates to advice about OTC medications. Compared to White participants, participants of another race were likelier to report that the app would not lead to confusion or discomfort (=0.27, 95% CI 0.04-0.49). There were no

XSL•FO
RenderX

differences by age, gender, or ethnicity in attitudes toward using an AI app.

Compared to participants with less than a bachelor's degree, participants with a bachelor's degree or higher were more likely to report that the app would improve access to medication information (=0.21, 95% CI 0.06-0.35), they would trust it (=0.20, 95% CI 0.06-0.35), and it would not lead to confusion or discomfort (=0.20, 95% CI 0.04-0.37). In contrast, participants who had higher health literacy, as evidenced by higher confidence in filling out medical forms, were less likely to believe that the app would improve access (=–0.10, 95% CI –0.18 to –0.02), they were less likely to trust the advice (=–0.14, 95% CI –0.22 to –0.07), and they were less likely to report that they would prefer to use the app before reaching out to their PCP (=–0.17, 95% CI –0.24 to –0.09). There were no differences by self-reported health status.

Compared to participants who reported low confidence, participants who reported high confidence using the internet were likelier to report that an app would increase their access to information about OTC medications (=0.22, 95% CI 0.06-0.39). Participants who reported high confidence using an EHR reported more positive perceptions of the app across all domains, while those who reported high confidence using apps reported more positive perceptions across all domains except nervousness, which showed negative perceptions.

XSL•FO
RenderX

Table 5. Demographic predictors of attitudes towards the use of an app that uses artificial intelligence (AI) along with the electronic health record to provide immediate, personalized advice about over-the-counter medications with a recommendation stored in the electronic health record using linear regression (n=1245).

	Access: using an AI app that provides personalized advice would improve my access to information about medications that are right for me	Trust: I would trust an AI app that provides personalized advice about medications that are right for me	(Not) nervous: using an AI app that provides personalized advice would make me feel (not) nervous[a]	(Not) uncomfortable: using an AI app that provides personalized advice would (not) make me confused and uncomfortable[a]	Prefer: whenever I need advice about medications, I would prefer to use an AI app that provides personalized advice before reaching out to my primary care provider with the same question
Age, OR[b] (95% CI)	0.00 (–0.01 to 0.01)	0.00 (–0.01 to 0.01)	–0.01 (–0.03 to 0.00)	–0.01 (–0.02 to 0.00)	–0.01 (–0.02 to 0.01)
Gender, OR (95% CI)					
Male	REF[c]	REF	REF	REF	REF
Female	–0.04 (–0.18 to 0.09)	–0.07 (–0.20 to 0.06)	–0.10 (–0.26 to 0.05)	–0.19 (–0.34 to –0.03)	–0.08 (–0.22 to 0.05)
Race, OR (95% CI)					
White or Caucasian	REF	REF	REF	REF	REF
Black of African American	0.32 (0.13 to 0.51)	0.24 (0.06 to 0.43)	0.20 (–0.01 to 0.40)	0.21 (0.00 to 0.42)	0.10 (–0.09 to 0.28)
Another race	0.19 (–0.02 to 0.39)	0.05 (–0.15 to 0.25)	0.03 (–0.20 to 0.26)	0.27 (0.04 to 0.49)	0.16 (–0.04 to 0.37)
Ethnicity, OR (95% CI)					
Non-Hispanic	REF	REF	REF	REF	REF
Hispanic	–0.09 (–0.27 to 0.09)	0.08 (–0.09 to 0.25)	0.10 (–0.09 to 0.30)	0.19 (–0.01 to 0.38)	–0.14 (–0.31 to 0.04)
Education, OR (95% CI)					
<Bachelor's degree	REF	REF	REF	REF	REF
≥Bachelor's degree	0.21 (0.06 to 0.35)	0.20 (0.06 to 0.35)	0.13 (–0.04 to 0.29)	0.20 (0.04 to 0.37)	0.10 (–0.04 to 0.25)
Health status	–0.02 (–0.10 to 0.06)	0.04 (–0.04 to 0.11)	0.02 (–0.07 to 0.10)	0.04 (–0.05 to 0.13)	0.01 (–0.06 to 0.09)
Health literacy	–0.10 (–0.18 to –0.02)	–0.14 (–0.22 to –0.07)	0.03 (–0.05 to 0.12)	0.04 (–0.04 to 0.13)	–0.17 (–0.24 to –0.09)
Confidence using internet[d], OR (95% CI)					
Low	REF	REF	REF	REF	REF
High	0.22 (0.06 to 0.39)	0.16 (0.00 to 0.32)	0.14 (–0.04 to 0.32)	0.01 (–0.18 to 0.19)	0.05 (–0.11 to 0.22)
Confidence using electronic health record[d], OR (95% CI)					
Low	REF	REF	REF	REF	REF
High	0.25 (0.10 to 0.40)	0.21 (0.07 to 0.36)	0.23 (0.07 to 0.40)	0.32 (0.15 to 0.48)	0.18 (0.04 to 0.33)
Confidence using apps[d], OR (95% CI)					
Low	REF	REF	REF	REF	REF
High	0.53 (0.37 to 0.69)	0.47 (0.31 to 0.62)	0.17 (0.00 to 0.35)	0.35 (0.18 to 0.53)	0.38 (0.22 to 0.53)

[a]Ratings for "nervous" and "uncomfortable" were reverse-coded for analysis so that higher values represent positive outcomes in all regressions.

[c]OR: odds ratio.

[c]REF: reference category.

[d]5-Point Likert scale: 1=not at all confident, 5=very confident; participants were categorized as low (scores 1-3) or high (scores 4-5).

Discussion

Principal Findings

Older adults are increasingly using technology, such as smartphone apps and virtual medical appointments, as part of managing their health [9]. In our study, we asked older adults to imagine interacting with an AI app that provides personalized medication recommendations. In this study using hypothetical vignettes, most older adults preferred to follow a recommendation to stop a medication when provided by a PCP

XSL•FO
RenderX

rather than a pharmacist at a drugstore or via an AI app. However, older adults who were Black or African American reported a lower likelihood of following a recommendation from a PCP and a higher likelihood of following a medication recommendation from an AI app. AI-based tools are quickly being integrated into health care systems to diagnose diseases, suggest evidence-based treatment plans, and help health care professionals make medical decisions [33-36]. However, older adults have expressed reservations about their use as part of the medical decision-making process [37]. To our knowledge, this is the first quantitative study to explore older adults' attitudes about using a patient-facing AI-app to help make medication decisions.

Comparison With Prior Work

Previous studies identified that older adults frequently report being willing to stop a medication if their doctor said it was possible [38,39]. Tools such as the Anticholinergic Burden Calculator, Beers Criteria, and MedStopper provide evidence-based recommendations to clinicians about potentially inappropriate medications [40-42]. Furthermore, clinical decision support systems are increasingly used within health systems to assist clinicians in making medication decisions [43-45]. Rao et al [46] recently tested whether a publicly available LLM could provide appropriate medication discontinuation recommendations and concluded that specially trained LLMs may be useful for clinicians.

Doctors are among the most trusted professions in the United States [47]. Older adults expect their doctors to provide appropriate medication recommendations and trust that information when they make decisions [5,48]. Increased patient trust is associated with improved satisfaction with treatment, health outcomes, and quality of life [49]. Unfortunately, engaging in shared decision-making about potentially stopping medication in routine clinical practice is difficult due to limited time, competing priorities, and the increasing lack of continuity of care, leading to a breakdown in relationships between patients and health care professionals [50-52]. While adults in the United States have expressed concerns about clinicians relying on AI when making health care decisions, we sought to explore older adults' perceptions of using AI tools specifically to support medication discontinuation recommendations [53]. Across the 3 AI app hypothetical vignettes that we presented, in 47% of vignettes, older adults agreed (scores 4-6) with a recommendation to stop the medication, of which 14% of ratings were extremely likely (score 6). This indicates that some older adults may be willing to use this technology but have reservations about automatically following its recommendations. This caution is warranted, given the lack of regulations on the use of AI in health care [54].

We observed that older adults' likelihood of stopping a medication varied by source of information; however, there was little variation across different therapeutic classes and rationales for discontinuing a medication. This aligns with a previous web-based survey of older adults using hypothetical vignettes that found that characteristics of older adults, as opposed to medication-specific characteristics, predicted willingness to stop a medication [55]. In contrast, a larger survey of older adults in 4 countries using hypothetical vignettes identified that the medication type and rationale for discontinuation were important factors in the decision-making process [17]. Additional research is needed to assess older adults' willingness to accept recommendations provided by an AI app with a broader range of prescription and OTC medications and rationales for discontinuation.

We were surprised that individuals with low health literacy were more likely to accept a recommendation for a medication when it was provided by an AI tool. It is possible that individuals with low health literacy were more interested in using technology, given the stigma they may experience when interacting with the traditional health care system [56,57]. However, older adults with lower health literacy may need additional support to use this type of web-based tool effectively. We observed that older adults with high confidence in using the EHR and apps were more likely to have positive perceptions of using an AI app to make medication decisions and to follow its recommendation to stop a medication. This is an important area for continuing research, given the potential for increased inequalities due to inadequate digital health literacy [58,59].

In our study, older adults who were Black or African American (as compared to White) were less likely to agree with stopping a medication when a PCP recommended it. This aligns with literature that reports that Black or African American adults are less likely to trust their doctor [60,61]. This may be due to the negative experiences and discrimination that Black or African American adults report continuing to occur in health care settings [62]. Health care providers should engage more frequently in shared decision-making conversations with all patients, especially those from minoritized communities, taking into consideration cultural values, preferences, and what matters most to them, as these may be important factors that play a role in the decision-making process for initiating or discontinuing prescribed treatment strategies. In contrast, we found that older adults who were Black or African American were more likely to agree with stopping a medication when an AI app recommended it. While Black or African American individuals are less likely to have internet access, those who do have access use social media, which is often available via apps, at higher rates than other racial groups [63-65]. AI apps may have the potential to improve accessibility to personalized health care information, but they also include biases and stereotypes that could amplify existing health care disparities [66].

Limitations

The primary limitation of our study is that we asked participants to share their perceptions about hypothetical AI apps with limited written descriptions. Further research is needed to determine whether older adults who reported having an interest in using an AI app would use the tool when making real-world medication decisions. Second, we acknowledge that our study was conducted as a web-based survey, which may have resulted in participation from older adults who were relatively comfortable with technology or who responded positively to questions about technology due to social desirability bias. Third, our findings may not be generalizable to all medications, given that agreement with stopping a medication is influenced by

XSL•FO
RenderX

factors such as the type of medication, rationale for recommending that the medication be discontinued, and contextual factors related to the individual patient [17-19,67-70]. Finally, this study was limited to adults aged 65 years and older; additional research is needed to explore to what extent younger adults prefer to use AI apps when making medication decisions.

Conclusions

It is important to understand older adults' perceptions of AI tools for health care to inform the development of user-friendly, reliable apps that will be useful in the real-world setting. Our study findings suggest that older adults have reservations about stopping a medication when an AI app recommends it. However, individuals who are Black or African American, have higher self-reported health, or have higher confidence in using an EHR or apps may be receptive to AI-based medication recommendations. Future research is needed to explore whether patient-facing, AI-based clinical decision support systems may complement the traditional health care system and increase access to accurate, personalized medication advice to improve clinical and patient-centered outcomes.

Acknowledgments

The University of Michigan College of Pharmacy provided financial support for this study.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

SV and MD conceptualized the study. All authors contributed to the methodology. VM conducted the formal analysis. SV wrote the original draft and all authors participated in the reviewing and editing of the manuscript. SV obtained funding.

Conflicts of Interest

None declared.

References

1. Kirzinger A, Neuman T, Cubanksi J, Brodie M. Data note: prescriptions drugs and older adults. Kaiser Family Foundation. URL: https://www.kff.org/affordable-care-act/issue-brief/data-note-prescription-drugs-and-older-adults/ [accessed 2024-05-02]
2. Qato DM, Wilder J, Schumm LP, Gillet V, Alexander GC. Changes in prescription and over-the-counter medication and dietary supplement use among older adults in the United States, 2005 vs 2011. JAMA Intern Med 2016;176(4):473-482 [FREE Full text] [doi: 10.1001/jamainternmed.2015.8581] [Medline: 26998708]
3. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. Expert Opin Drug Saf 2014;13(1):57-65 [FREE Full text] [doi: 10.1517/14740338.2013.827660] [Medline: 24073682]
4. Wolff JL, Boyd CM. A look at person- and family-centered care among older adults: results from a national survey [corrected]. J Gen Intern Med 2015;30(10):1497-1504 [FREE Full text] [doi: 10.1007/s11606-015-3359-6] [Medline: 25933625]
5. Donohue JM, Huskamp HA, Wilson IB, Weissman J. Whom do older adults trust most to provide information about prescription drugs? Am J Geriatr Pharmacother 2009;7(2):105-116 [FREE Full text] [doi: 10.1016/j.amjopharm.2009.04.005] [Medline: 19447363]
6. Paliwal Y, Gendron TL, Jones RM, Moczygemba L, Nadpara PA, Slattum PW. A qualitative study to understand over-the-counter medication use and decision-making among residents of senior-living communities. Res Social Adm Pharm 2019;15(6):730-737. [doi: 10.1016/j.sapharm.2018.09.004] [Medline: 30253975]
7. Ilardo ML, Speciale A. The community pharmacist: perceived barriers and patient-centered care communication. Int J Environ Res Public Health 2020;17(2):536 [FREE Full text] [doi: 10.3390/ijerph17020536] [Medline: 31952127]
8. Ortega MV, Hidrue MK, Lehrhoff SR, Ellis DB, Sisodia RC, Curry WT, et al. Patterns in physician burnout in a stable-linked cohort. JAMA Netw Open 2023;6(10):e2336745 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36745] [Medline: 37801314]
9. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022;12(12):1102-1105 [FREE Full text] [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
10. Martin-Hammond AM, Abegaz T, Gilbert JE. Designing an over-the-counter consumer decision-making tool for older adults. J Biomed Inform 2015;57:113-123 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.006] [Medline: 26184058]
11. Vordenberg SE, Maitah Y, Khan Z, DeBacker K, Dorsch MP. Consumer perceptions of the OTC coach: a clinical decision support system aimed at improving the safe use of over-the-counter medications. J Am Pharm Assoc 2023;63(1):135-143. [doi: 10.1016/j.japh.2022.09.008] [Medline: 36243654]

XSL•FO
RenderX

12. Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices 2023;8:1-8 [FREE Full text] [doi: 10.1007/s44174-023-00063-2] [Medline: 36785697]
13. Over-the-Counter (OTC) Drugs Branch: The OTC Drug Review. U.S. Food and Drug Administration. URL: https://www.fda.gov/drugs/enforcement-activities-fda/over-counter-otc-drugs-branch-otc-drug-review [accessed 2024-05-02]
14. OpenAI. ChatGPT. URL: https://openai.com/index/chatgpt [accessed 2024-05-02]
15. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
17. Vordenberg SE, Weir KR, Jansen J, Todd A, Schoenborn N, Scherer AM. Harm and medication-type impact agreement with hypothetical deprescribing recommendations: a vignette-based experiment with older adults across four countries. J Gen Intern Med 2023;38(6):1439-1448 [FREE Full text] [doi: 10.1007/s11606-022-07850-5] [Medline: 36376636]
18. Weir KR, Vordenberg SE, Scherer AM, Jansen J, Schoenborn N, Todd A. Exploring different contexts of statin deprescribing: a vignette-based experiment with older adults across four countries. J Gen Intern Med 2024;39(9):1773-1776 [FREE Full text] [doi: 10.1007/s11606-024-08698-7] [Medline: 38514582]
19. Vordenberg SE, Kirch M, Singer D, Solway E, Roberts JS, Zikmund-Fisher BJ, et al. Interest in medication deprescribing among US adults aged 50-80. J Gen Intern Med 2024;39(13):2631-2633. [doi: 10.1007/s11606-024-08945-x] [Medline: 39020227]
20. US Preventive Services Task Force, Davidson KW, Barry MJ, Mangione CM, Cabana M, Chelmow D, et al. Aspirin use to prevent cardiovascular disease: us preventive services task force recommendation statement. J Am Med Assoc 2022;327(16):1577-1584. [doi: 10.1001/jama.2022.4983] [Medline: 35471505]
21. Gupta M, Gulati S, Nasir K, Sarraju A. Aspirin use prevalence for cardiovascular disease prevention among U.S. adults from 2012 to 2021. Ann Intern Med 2024;177(8):1139-1141. [doi: 10.7326/M24-0427] [Medline: 38914002]
22. Wright R. How Zantac became the best-selling drug in history. J Health Care Mark 1996;16(4):24-29. [Medline: 10169076]
23. U.S. Food and Drug Administration. FDA requests removal of all ranitidine products (Zantac) from the market. URL: https://www.fda.gov/news-events/press-announcements/fda-requests-removal-all-ranitidine-products-zantac-market [accessed 2024-09-04]
24. US Preventive Services Task Force, Mangione CM, Barry MJ, Nicholson WK, Cabana M, Chelmow D, et al. Statin use for the primary prevention of cardiovascular disease in adults: US preventive services task force recommendation statement. J Am Med Assoc 2022;328(8):746-753. [doi: 10.1001/jama.2022.13044] [Medline: 35997723]
25. Jacobs JA, Addo DK, Zheutlin AR, Derington CG, Essien UR, Navar AM, et al. Prevalence of statin use for primary prevention of atherosclerotic cardiovascular disease by race, ethnicity, and 10-year disease risk in the US: national health and nutrition examination surveys, 2013 to March 2020. JAMA Cardiol 2023;8(5):443-452 [FREE Full text] [doi: 10.1001/jamacardio.2023.0228] [Medline: 36947031]
26. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. Fam Med 2004;36(8):588-594. [Medline: 15343421]
27. DeSalvo KB, Bloser N, Reynolds K, He J, Muntner P. Mortality prediction with a single general self-rated health question. A meta-analysis. J Gen Intern Med 2006;21(3):267-275 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.00291.x] [Medline: 16336622]
28. Wallace LS, Rogers ES, Roskos SE, Holiday DB, Weiss BD. Brief report: screening items to identify patients with limited health literacy skills. J Gen Intern Med 2006;21(8):874-877 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00532.x] [Medline: 16881950]
29. Davis FD, Davis F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-339. [doi: 10.2307/249008]
30. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018;9(3):604-634 [FREE Full text] [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
31. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
32. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
33. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
34. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]

XSL•FO
RenderX

35. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discov Artif Intell 2023;3(1):5. [doi: 10.1007/s44163-023-00049-5]
36. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863 [FREE Full text] [doi: 10.1177/23333928241234863] [Medline: 38449840]
37. AI in health care: thoughts and opinions among the 50-plus. AARP. 2024. URL: https://www.aarp.org/content/dam/aarp/research/topics/health/conditions-treatment/artificial-intelligence-health-care.doi.10.26419-2fres.00802.001.pdf [accessed 2024-09-07]
38. Reeve E, Low L, Shakib S, Hilmer SN. Development and validation of the revised patients' attitudes towards deprescribing (rPATD) questionnaire: versions for older adults and caregivers. Drugs Aging 2016;33(12):913-928. [doi: 10.1007/s40266-016-0410-1] [Medline: 27785734]
39. Crutzen S, Abou J, Smits SE, Baas G, Hugtenburg JG, Heringa M, et al. Older people's attitudes towards deprescribing cardiometabolic medication. BMC Geriatr 2021;21(1):366 [FREE Full text] [doi: 10.1186/s12877-021-02249-z] [Medline: 34134649]
40. ACB Calculator. URL: https://www.acbcalc.com/ [accessed 2024-05-15]
41. By the 2023 American Geriatrics Society Beers Criteria® Update Expert Panel. American geriatrics society 2023 updated AGS beers criteria® for potentially inappropriate medication use in older adults. J Am Geriatr Soc 2023;71(7):2052-2081. [doi: 10.1111/jgs.18372] [Medline: 37139824]
42. MedStopper. 2024. URL: https://medstopper.com/ [accessed 2024-05-15]
43. Roumeliotis N, Sniderman J, Adams-Webber T, Addo N, Anand V, Rochon P, et al. Effect of electronic prescribing strategies on medication error and harm in hospital: a systematic review and meta-analysis. J Gen Intern Med 2019;34(10):2210-2223 [FREE Full text] [doi: 10.1007/s11606-019-05236-8] [Medline: 31396810]
44. Keine D, Zelek M, Walker JQ, Sabbagh MN. Polypharmacy in an elderly population: enhancing medication management through the use of clinical decision support software platforms. Neurol Ther 2019;8(1):79-94 [FREE Full text] [doi: 10.1007/s40120-019-0131-6] [Medline: 30900186]
45. Nanji KC, Garabedian PM, Langlieb ME, Rui A, Tabayoyong LL, Sampson M, et al. Usability of a perioperative medication-related clinical decision support software application: a randomized controlled trial. J Am Med Inform Assoc 2022;29(8):1416-1424 [FREE Full text] [doi: 10.1093/jamia/ocac035] [Medline: 35575780]
46. Rao A, Kim J, Lie W, Pang M, Fuh L, Dreyer KJ, et al. Proactive polypharmacy management using large language models: opportunities to enhance geriatric care. J Med Syst 2024;48(1):41. [doi: 10.1007/s10916-024-02058-y] [Medline: 38632172]
47. Doctors are the most trusted profession in the U.S. and across the world. Ipsos. URL: https://www.ipsos.com/en-us/news-polls/global-trustworthy-index-US-version [accessed 2024-08-07]
48. Belcher VN, Fried TR, Agostini JV, Tinetti ME. Views of older adults on patient participation in medication-related decision making. J Gen Intern Med 2006;21(4):298-303 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00329.x] [Medline: 16686804]
49. Birkhäuer J, Gaab J, Kossowsky J, Hasler S, Krummenacher P, Werner C, et al. Trust in the health care professional and health outcome: a meta-analysis. PLoS One 2017;12(2):e0170988 [FREE Full text] [doi: 10.1371/journal.pone.0170988] [Medline: 28170443]
50. Bynum JPW, Barre L, Reed C, Passow H. Participation of very old adults in health care decisions. Med Decis Making 2014;34(2):216-230 [FREE Full text] [doi: 10.1177/0272989X13508008] [Medline: 24106235]
51. Anderson K, Stowasser D, Freeman C, Scott I. Prescriber barriers and enablers to minimising potentially inappropriate medications in adults: a systematic review and thematic synthesis. BMJ Open 2014;4(12):e006544 [FREE Full text] [doi: 10.1136/bmjopen-2014-006544] [Medline: 25488097]
52. Prochaska MT, Zhang H, Meltzer DO, Arora VM. Patient characteristics associated with and changes over time in trust in inpatient physicians. J Gen Intern Med 2022;37(1):266-268 [FREE Full text] [doi: 10.1007/s11606-021-06649-0] [Medline: 33634381]
53. Tyson A, Pasquini G, Spencer A, Funk C. 60% of Americans would be uncomfortable with providers relying on AI in their own health care. Pew Research Center. URL: https://www.pewresearch.org/wp-content/uploads/sites/20/2023/02/PS_2023.02.22_AI-health_REPORT.pdf [accessed 2024-05-15]
54. McKee M, Wouters OJ. The challenges of regulating artificial intelligence in healthcare comment on "clinical decision support and new regulatory frameworks for medical devices: are we ready for it? - a viewpoint paper". Int J Health Policy Manag 2023;12:7261 [FREE Full text] [doi: 10.34172/ijhpm.2022.7261] [Medline: 36243948]
55. Vordenberg SE, Zikmund-Fisher BJ. Characteristics of older adults predict concern about stopping medications. J Am Pharm Assoc (2003) 2020;60(6):773-780. [doi: 10.1016/j.japh.2020.01.019] [Medline: 32115393]
56. Easton P, Entwistle VA, Williams B. How the stigma of low literacy can impair patient-professional spoken interactions and affect health: insights from a qualitative investigation. BMC Health Serv Res 2013;13:319 [FREE Full text] [doi: 10.1186/1472-6963-13-319] [Medline: 23958036]
57. Mackert M, Mabry-Flynn A, Donovan EE, Champlin S, Pounders K. Health literacy and perceptions of stigma. J Health Commun 2019;24(11):856-864. [doi: 10.1080/10810730.2019.1678705] [Medline: 31630662]

XSL•FO
RenderX

58. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
59. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
60. Black Americans' views about health disparities, experiences with health care. Pew Research Center. URL: https://www.pewresearch.org/science/2022/04/07/black-americans-views-about-health-disparities-experiences-with-health-care/ [accessed 2024-09-07]
61. Musa D, Schulz R, Harris R, Silverman M, Thomas SB. Trust in the health care system and the use of preventive health services by older black and white adults. Am J Public Health 2009;99(7):1293-1299. [doi: 10.2105/AJPH.2007.123927] [Medline: 18923129]
62. Hamel L, Lopes L, Muñana C, Artiga S, Brodie M. Race, health, and COVID-19: the views and experiences of Black Americans. Kaiser Family Foundation. URL: https://files.kff.org/attachment/Report-Race-Health-and-COVID-19-The-Views-and-Experiences-of-Black-Americans.pdf [accessed 2024-05-16]
63. Aging connected: exposing the hidden connectivity crisis for older adults. Older Adult Technology Services from AARP. URL: https://agingconnected.org/wp-content/uploads/2021/05/Aging-Connected_Exposing-the-Hidden-Connectivity-Crisis-for-Older-Adults.pdf [accessed 2024-05-16]
64. Stamps DL. Black audiences' identity-focused social media use, group vitality, and consideration of collective action. J Mass Commun Q 2022;99(3):660-675. [doi: 10.1177/10776990221104152]
65. Auxier B. Social media outlets continue to be important political outlets for Black Americans. Pew Research Center. URL: https://pewresearch.org/short-reads/2020/12/11/social-media-continue-to-be-important-political-outlets-for-black-americans/ [accessed 2024-05-16]
66. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019;9(2):010318 [FREE Full text] [doi: 10.7189/jogh.09.020318] [Medline: 31788229]
67. Weir KR, Shang J, Choi J, Rana R, Vordenberg SE. Factors important to older adults who disagree with a deprescribing recommendation. JAMA Netw Open 2023;6(10):e2337281 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.37281] [Medline: 37819657]
68. Vordenberg SE, Rana R, Shang J, Choi J, Scherer AM, Weir KR. Reasons why older adults in three countries agreed with a deprescribing recommendation in a hypothetical vignette. Basic Clin Pharmacol Toxicol 2023;133(6):673-682 [FREE Full text] [doi: 10.1111/bcpt.13857] [Medline: 36894739]
69. Maitah Y, Rana R, Vordenberg S. Older adults experiencing polypharmacy preferred to continue their non-prescription products during a card sorting activity. Int J Pharm Pract 2023;31(2):257-260. [doi: 10.1093/ijpp/riac100] [Medline: 36535005]
70. Weir KR, Scherer AM, Vordenberg SE, Streit S, Jansen J, Jungo KT. The patient typology about deprescribing and medication-related decisions: a quantitative exploration. Basic Clin Pharmacol Toxicol 2024;134(1):39-50 [FREE Full text] [doi: 10.1111/bcpt.13911] [Medline: 37300477]

Abbreviations

AI: artificial intelligence
CHERIES: Checklist for Reporting Results of Internet E-Surveys
EHR: electronic health record
EHR-Direct: artificial intelligence app that connects with the electronic health record and directly provides advice to the consumer
EHR-PCP: artificial intelligence app that connects with the electronic health record and provides advice to the primary care provider for approval before sharing the recommendation with the consumer
LLM: large language mode
OR: odds ratio
OTC: over-the-counter
PCP: primary care provider
Questions-Direct: artificial intelligence app that asks questions and provides advice directly to the consumer
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by N Cahill; submitted 21.05.24; peer-reviewed by U Sinha, HO Abu, ES Obi, E Oluokun; comments to author 21.08.24; revised version received 07.09.24; accepted 18.11.24; published 16.12.24.

Please cite as:
Vordenberg SE, Nichols J, Marshall VD, Weir KR, Dorsch MP
Investigating Older Adults' Perceptions of AI Tools for Medication Decisions: Vignette-Based Experimental Survey
J Med Internet Res 2024;26:e60794
URL: https://www.jmir.org/2024/1/e60794
doi:10.2196/60794
PMID:

©Sarah E Vordenberg, Julianna Nichols, Vincent D Marshall, Kristie Rebecca Weir, Michael P Dorsch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence–Aided Diagnosis System for the Detection and Classification of Private-Part Skin Diseases: Decision Analytical Modeling Study

Wei Wang[1], PhD; Xiang Chen[2], Prof Dr Med; Licong Xu[3], MM; Kai Huang[2], MD; Shuang Zhao[2], Prof Dr Med; Yong Wang[1], Prof Dr

[1]School of Automation, Central South University, Changsha, China
[2]Department of Dermatology, Xiangya Hospital, Central South University, Changsha, China
[3]Jinhua Fifth Hospital, Jinhua, China

Corresponding Author:
Yong Wang, Prof Dr
School of Automation
Central South University
932 South Lushan Road
Changsha, 410083
China
Phone: 86 18507313729
Email: ywang@csu.edu.cn

Abstract

Background: Private-part skin diseases (PPSDs) can cause a patient's stigma, which may hinder the early diagnosis of these diseases. Artificial intelligence (AI) is an effective tool to improve the early diagnosis of PPSDs, especially in preventing the deterioration of skin tumors in private parts such as Paget disease. However, to our knowledge, there is currently no research on using AI to identify PPSDs due to the complex backgrounds of the lesion areas and the challenges in data collection.

Objective: This study aimed to develop and evaluate an AI-aided diagnosis system for the detection and classification of PPSDs: aiding patients in self-screening and supporting dermatologists' diagnostic enhancement.

Methods: In this decision analytical modeling study, a 2-stage AI-aided diagnosis system was developed to classify PPSDs. In the first stage, a multitask detection network was trained to automatically detect and classify skin lesions (type, color, and shape). In the second stage, we proposed a knowledge graph based on dermatology expertise and constructed a decision network to classify seven PPSDs (condyloma acuminatum, Paget disease, eczema, pearly penile papules, genital herpes, syphilis, and Bowen disease). A reader study with 13 dermatologists of different experience levels was conducted. Dermatologists were asked to classify the testing cohort under reading room conditions, first without and then with system support. This AI-aided diagnostic study used the data of 635 patients from two institutes between July 2019 and April 2022. The data of Institute 1 contained 2701 skin lesion samples from 520 patients, which were used for the training of the multitask detection network in the first stage. In addition, the data of Institute 2 consisted of 115 clinical images and the corresponding medical records, which were used for the test of the whole 2-stage AI-aided diagnosis system.

Results: On the test data of Institute 2, the proposed system achieved the average precision, recall, and F_1-score of 0.81, 0.86, and 0.83, respectively, better than existing advanced algorithms. For the reader performance test, our system improved the average F_1-score of the junior, intermediate, and senior dermatologists by 16%, 7%, and 4%, respectively.

Conclusions: In this study, we constructed the first skin-lesion–based dataset and developed the first AI-aided diagnosis system for PPSDs. This system provides the final diagnosis result by simulating the diagnostic process of dermatologists. Compared with existing advanced algorithms, this system is more accurate in identifying PPSDs. Overall, our system can not only help patients achieve self-screening and alleviate their stigma but also assist dermatologists in diagnosing PPSDs.

(J Med Internet Res 2024;26:e52914) doi:10.2196/52914

XSL•FO
RenderX

KEYWORDS

artificial intelligence-aided diagnosis; private parts; skin disease; knowledge graph; dermatology; classification; artificial intelligence; AI; diagnosis

Introduction

Skin diseases affect 1.9 billion people worldwide and place a burden on patients' mental health and quality of life [1-3]. Private-part skin diseases (PPSDs) are a group of high-incidence skin diseases that occur in the private parts of the human body (such as breasts, genitals, or anus), including syphilis, genital herpes, Paget disease, and condyloma acuminatum [4-7]. In clinical practice, a patient needs to expose the affected area sufficiently to dermatologists for visual inspection. Under this condition, patients with PPSDs may become nervous and embarrassed due to stigma and may be reluctant to see a doctor, which hinders the early diagnosis of PPSDs [8-11]. Patients may also be unaware of PPSDs, which is another important reason for the delay of the PPSD diagnosis. For example, sometimes patients easily mistake Paget disease for eczema, which may lead to cancer metastasis. Moreover, patients with PPSDs may hide some medical records because PPSDs may be related to sexual infidelity. Therefore, in order to promote the early diagnosis of PPSDs, it is necessary to develop a novel and private artificial intelligence (AI)–aided diagnosis technology.

Among many advanced AI algorithms, convolutional neural networks (CNNs) have developed rapidly and have shown remarkable performance on many computer vision–related tasks [12-16]. CNNs can learn meaningful and robust features directly from data and have been widely used in AI-aided diagnosis of skin diseases [17-22]. For example, Esteva et al [23] trained an end-to-end CNN with a dataset of 129,450 clinical images. The performance of this CNN is comparable to that of dermatologists in two binary classification tasks (keratinocyte carcinoma vs benign seborrheic keratosis and melanoma vs benign nevus) [23]. It is a seminal work in applying AI to skin disease diagnosis. Fink et al [24] used a pretrained GoogleNet Inception_v4 architecture to classify combined naevi and melanomas and achieved better performance than 11 trained dermatologists [24].

Although CNN-based algorithms have shown promising results in the field of skin disease diagnosis, most of them focus on the classification of skin tumors, and none has been developed for the diagnosis of PPSDs. It may be a feasible solution to introduce existing CNN-based algorithms in skin diseases to assist diagnosis of PPSDs. However, directly applying them to PPSDs has the following two issues. First, the skin representation in clinical images of private parts is complex. For example, a large area in clinical images of private parts is occupied by irregular tissues or other visual obstacles (eg, hair and skinfolds) that are irrelevant to disease recognition, which has a negative effect on the recognition performance. Second, the data on PPSDs is scarce because it is difficult to collect the data, which limits the application of CNN-based algorithms in PPSD diagnosis. Because of these two issues, the classification of the clinical images of PPSDs is usually more difficult than that of images in natural scenes [25]. Therefore, it is necessary to develop more effective techniques to improve the classification performance of PPSDs.

In this paper, we make the first attempt at the assistant diagnosis of PPSDs and develop a two-stage AI-aided diagnosis system by simulating the diagnostic process of dermatologists. This system builds a dermatological knowledge graph from skin lesions and medical records to disease diagnosis, aiming to reduce the model's dependence on data. Therefore, the system can render the diagnosis of PPSDs with only a small amount of data. Unlike other algorithms that directly learn the mapping from original images to diseases, this system recognizes the type, color, and shape of all skin lesions in original images, and combines the recognition results of skin lesions with the medical records to determine diseases. This system simplifies the disease classification problem into the skin lesion classification problem, which eliminates irregular tissues and visual obstacles in original images to a certain degree, thus alleviating the issue of complex skin representation in clinical images.

When our system is applied to dermatological assistant diagnosis, patients with PPSDs can avoid face-to-face diagnosis with dermatologists, thereby relieving their sense of shame and embarrassment. In addition, patients with PPSDs can carry out disease screening in time and fill in the medical records truthfully.

Methods

Ethical Considerations

This study is a secondary analysis based on the data generated during the diagnosis and treatment of PPSDs, all of which were approved by the Ethics Committee of Xiangya Hospital (review number 202308636). Before taking photos, all patients signed informed consent forms for imaging data shooting and scientific research analysis. Of course, all of these data were anonymized and the information that exposed the patients' identities was removed. As it is a necessary step in the diagnosis and treatment of PPSDs, patients who provide data have no additional financial compensation. We ensure that all the images in the written materials published by this study do not contain the identity information of individual participants.

Dataset Collection and Annotation

For this study, ethics review and institutional review board approval were obtained from 2 participating institutes (Institute 1: Xiangya Hospital of Central South University, and Institute 2: Wuhan No.1 Hospital). This study followed the Helsinki protocol and good clinical practice guidelines [26,27]. For the development and evaluation of our AI-aided diagnosis system, we retrospectively collected clinical images of PPSDs from the two institutes between July 2019 and April 2022. The collection criteria include (1) private parts skin diseases, (2) images do not affect judgment, and (3) informed consent holder. Exclusion criteria include (1) the inability to capture clear images and (2) those who do not agree to disclose images of skin lesions. These

XSL•FO
RenderX

processes were reported in accordance with the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) guidelines. The data collection process can be divided into 2 steps.

The first step was the collection of clinical images and their corresponding medical records for 7 categories of PPSDs: condyloma acuminatum, Paget disease, eczema, pearly penile papules, genital herpes, syphilis, and Bowen disease. Most of the collected cases had pathological confirmation and the remaining cases were unanimously recognized by 3 dermatologists from Institute 1. The medical records included the location of skin lesions, subjective symptoms, medication history, and history of high-risk sexual activities. Through data collection, we have collected 635 clinical images, of which 520 clinical images from Institute 1 were used for system development, and 115 clinical images and their corresponding medical records from Institute 2 were used for system evaluation. Detailed information on the PPSD dataset is provided in Table S1 in Multimedia Appendix 1.

In the second step, we conducted data labeling and cropping on the collected clinical images. To simulate the visual diagnosis process of dermatologists, three dermatologists from Institute 1 were asked to label the clinical images at the bounding box level, where each bounding box represents a skin lesion (see examples later). These labels characterize the appearance of the skin lesions, including type (papule, plaque, nodule, erosion, and vesicle), color (red, brown, and skin color), and shape (rice shape, round, irregular, papillary, and cauliflower). By cropping the bounding boxes from the original images (each may have multiple bounding boxes), we finally obtained 2701 skin lesion images. It is worth noting that among the five types of skin lesions, the amount of data for erosion and vesicle is relatively small, and we added some images of these 2 types of skin lesions from other human body parts.

AI-Aided Diagnosis System Overview

In the first stage, the multitask detection network consists of a detection model and a multitask classification model. For an input original image, the detection model first uses the feature extraction module together with the region proposal network to extract the region of interests in the image, that is, the skin lesions. Then, we perform the cropping operation. Note that an image may have multiple bounding boxes of skin lesions. Under this condition, the results of the detection model include multiple cropped images of skin lesions. Then, these cropped images are fed into the multitask classification model to further predict the type, color, and shape of skin lesions, together with their confidence scores (probabilities). Afterward, we count the number of skin lesions in each type. The detailed architecture of the multitask detection network can be seen in Section S1 in Multimedia Appendix 1.

In the second stage, to imitate the dermatologists' diagnostic process (visual diagnosis and inquiry diagnosis), we designed a decision network based on a dermatological knowledge graph (Figure S1 in Multimedia Appendix 1), which was constructed by dermatologists referencing medical guidelines and professional textbooks. The dermatological knowledge graph is composed of 30 decision paths. For each decision path, the decision network encodes the detection results of the first stage with medical records into a vector using predefined encoding rules. Subsequently, the inference result for each decision path is derived by multiplying the elements within the vector. Finally, the probabilities of all decision paths belonging to the same disease are summed and normalized to infer the final decision result. The inference process is shown in Figure S2 and Section S2 in Multimedia Appendix 1.

Inspired by how dermatologists diagnose disease, we proposed a 2-stage AI-aided diagnosis system capable of processing multimodal information, as shown in Figure 1. In the first stage, we trained a multitask detection network to jointly learn the type, color, and shape of skin lesions. In the second stage, we designed a decision network embedded with dermatology expertise, which combined the detection results of the first stage with medical records to obtain the final decision result. Overall, our two-stage system leverages multitask learning to explore skin lesion representations at multiple granularities and introduce dermatology expertise to alleviate the data dependency of deep learning models. The details of the two stages are presented as follows.

XSL•FO
RenderX

Figure 1. Flowchart of the proposed two-stage AI-aided diagnosis system for PPSDs. AI: artificial intelligence; PPSD: private-part skin disease; ROI: region of interest; RPN: region proposal network.

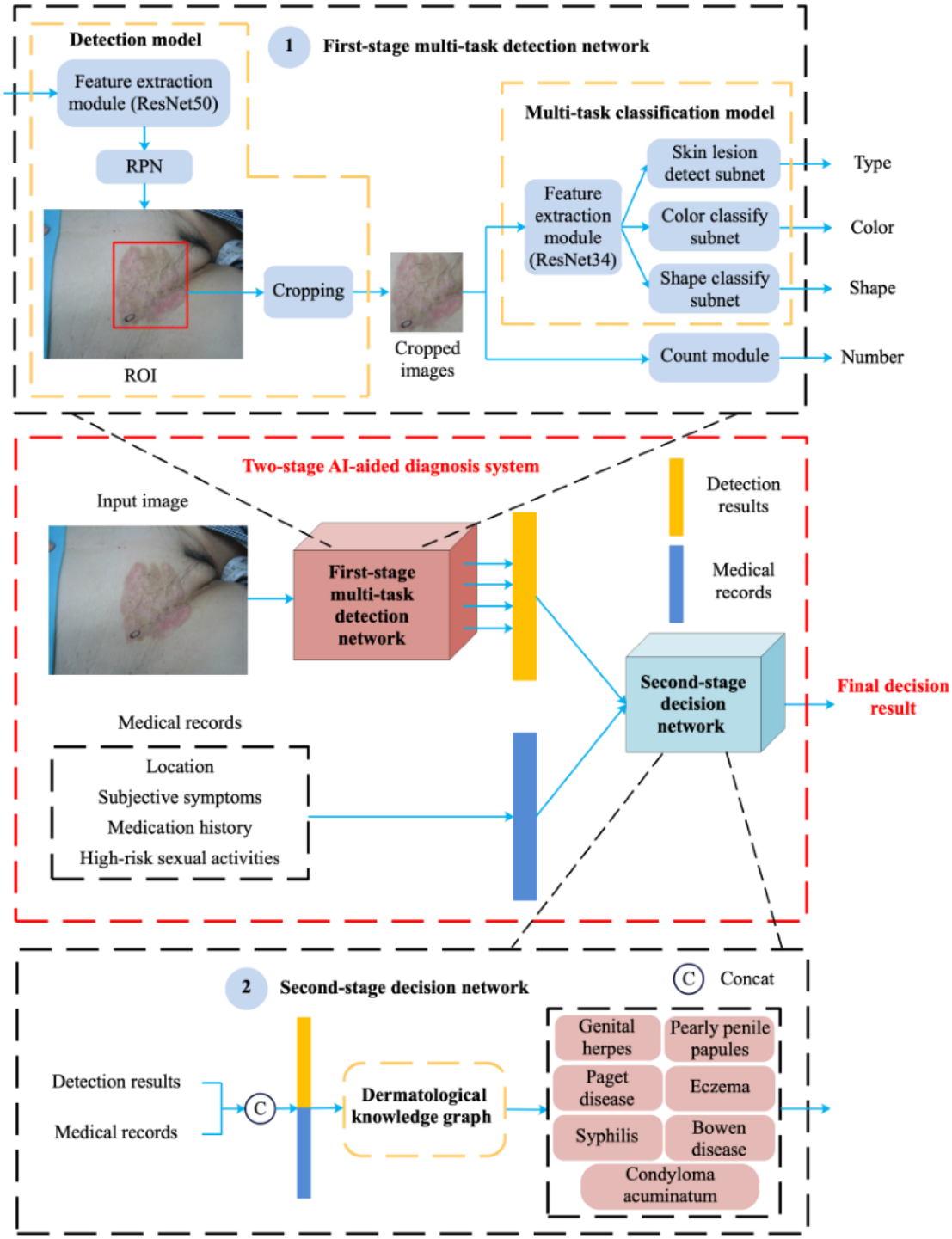

Reader Study Protocol

In order to test whether our system can assist dermatologists in diagnosing PPSDs, we conducted a reader study. The participants in this study included 3 senior dermatologists, 5 intermediate dermatologists, and 5 junior dermatologists. Senior dermatologists refer to dermatologists who have been in medical practice for more than 10 years. Intermediate dermatologists refer to those who have been licensed for more than 3 years. Junior dermatologists are those who have been in medical practice for less than 3 years. The participants were provided with 115 pairs of clinical images and their corresponding medical records as the test set to evaluate the system. They were asked to give the final disease diagnosis for the test set in 2 scenarios. In scenario one, the participants made diagnoses based on the pairs of clinical images and medical records. In

XSL•FO
RenderX

scenario two, the participants were additionally given the skin lesion detection results and disease prediction results (with PPSD confidence score) provided by our system, and they were asked to remake their diagnoses. We recorded the performance of the participants in both scenarios and calculated the performance difference between these two scenarios for each PPSD.

Performance Evaluation Index

To evaluate the first-stage multitask detection network, we used the confusion matrix, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUROC) to test the performance of skin lesion classification. The confusion matrix is used to observe the performance of the multitask classification model in various categories. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. The ROC curve is the plot of the true positive rate against the false positive rate, at various threshold settings. AUROC measures the entire 2D area underneath the entire ROC curve, and a larger AUROC value represents better classification performance. In addition, for system evaluation, we calculated precision, recall, and F_1-score to measure the PPSD classification performance. These indices are defined as: precision=TP/(TP+FP), recall=TP/(TP+FN), and F_1-score=2×precision×recall/(precision+recall), where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively. During the evaluation process, for a positive sample, if the classifier predicts it as a positive sample, it is marked as true positive; otherwise, it is marked as false negative. For a negative sample, if the classifier predicts it as a negative sample, it is marked as true negative; otherwise, it is marked as false positive.

Results

Performance Report of the Multitask Detection Network

To verify the performance of the multitask detection network in the first stage, we carried out the classification of the type, color, and shape of skin lesions. From Figure 2A, the results of the confusion matrices show that the network has the highest accuracy on both plaque (0.93) and nodule (0.93) in lesion type recognition. In color and shape recognition, our network achieved the highest accuracy on red (0.93) and cauliflower (0.84), respectively. The results in Figure 2B suggest that the AUROCs of all categories in type and color recognition are higher than 0.9. In particular, the AUROCs of nodule in type recognition and skin color in color recognition are 0.996 and 0.955, respectively. In shape recognition, our network performed well on cauliflower (0.993), rice shape (0.976), and papillary (0.956), but performed relatively poorly on round (0.840) and irregular (0.753). The reason for this observation is elaborated in the discussion.

Figure 2C visualizes the output results of the first-stage multitask detection network in some representative cases. Here, the blue bounding boxes and texts represent the ground truths of the locations of the skin lesions and the ground truths of the labels of the skin lesion type, color, and shape, respectively. The red bounding boxes and texts represent the detection results of this network, and the numbers in brackets denote the confidence scores of the predicted categories.

XSL•FO
RenderX

Figure 2. Results of multitask classification model. (A) The confusion matrix. (B) The ROC curve. (C) The visualization results. ROC: receiver operating characteristic.

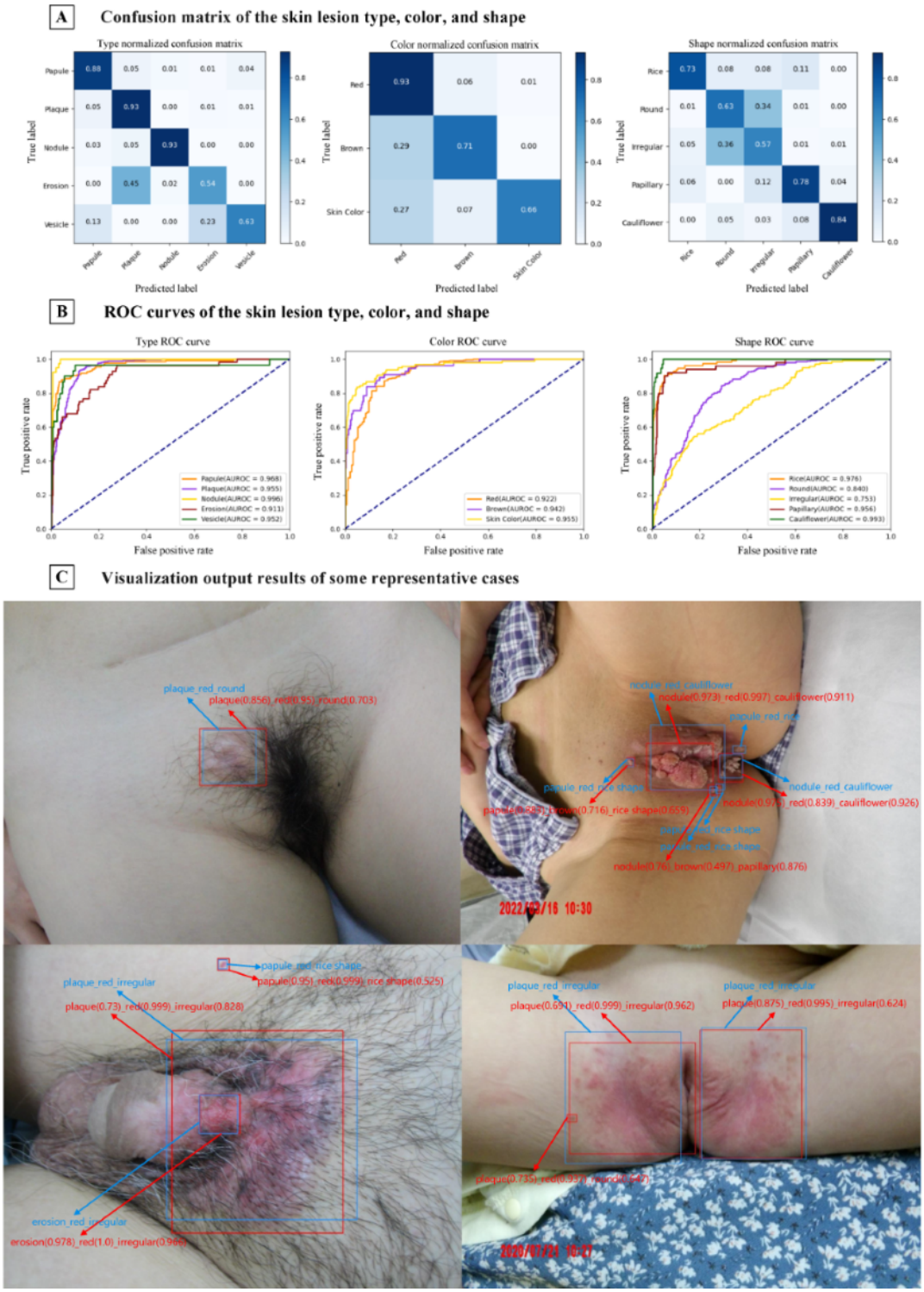

Comparison of Existing Algorithms and Proposed System

To demonstrate the effectiveness of the system, we compared it with seven existing advanced deep learning algorithms. Among the 7 algorithms, ResNet101, SENet, SKNet, and Convnext are CNN-based networks, and Deit, Levit, and Swin-twins are transformer-based networks. The results in Table 1 show that the average precision, recall, and F_1-score of our system on seven PPSDs are 0.81, 0.86, and 0.83, respectively, better than those of all the competitors. The system performance during training and validation is shown in Table S2 in Multimedia Appendix 1. In addition, our system achieved the best F_1-score among all the compared algorithms on condyloma acuminatum (0.90), genital herpes (0.94), Bowen disease (0.63), eczema (0.82), Paget disease (0.82), and syphilis (0.78). The comparison of the seven existing methods and the proposed system in computation time is given in Table S3 and Section S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Precision, recall, and F_1-score of our system and some advanced algorithms for each PPSD[a].

PPSDs	ResNet101	SKNet	Convnext	Deit	SENet	Levit	Swin-twins	Ours
Average								
F1[b]	0.52	0.59	0.57	0.59	0.59	0.60	0.69	*0.83* [c]
R[d]	0.57	0.62	0.59	0.60	0.63	0.61	0.69	*0.86*
P[e]	0.55	0.62	0.72	0.62	0.64	0.64	0.72	*0.81*
SP[f]								
F1	0.21	0.51	0.22	0.57	0.39	0.40	0.44	*0.78*
R	0.13	0.38	0.13	0.50	0.25	0.29	0.33	*0.73*
P	0.60	0.82	*1.00*	0.67	0.86	0.64	0.67	*0.84*
PD[g]								
F1	0.64	0.64	0.77	0.59	0.66	0.71	0.71	*0.82*
R	0.74	*0.91*	0.87	0.65	0.87	0.87	0.74	0.78
P	0.57	0.49	0.69	0.54	0.53	0.61	0.68	*0.86*
EZ [h]								
F1	0.32	0.17	0.36	0.48	0.36	0.34	0.67	*0.82*
R	0.35	0.12	0.35	0.41	0.35	0.35	0.82	*0.94*
P	0.29	0.29	0.38	0.58	0.38	0.33	0.56	*0.73*
BD[i]								
F1	0.10	0.31	0.45	0.13	0.19	0.33	0.38	*0.63*
R	0.06	0.25	0.44	0.13	0.13	0.25	0.38	*0.56*
P	0.20	0.40	0.47	0.14	0.40	0.50	0.38	*0.71*
PPP[j]								
F1	0.91	*1 .00*	*1 .00*	0.91	*1.00*	*1 .00*	0.89	0.91
R	1.00	1.00	1.00	1.00	1.00	1.00	0.80	*1 .00*
P	0.83	*1.00*	*1 .00*	0.83	*1 .00*	*1 .00*	*1 .00*	0.83
GH[k]								
F1	0.80	0.78	0.55	0.77	0.78	0.71	0.93	*0.94*
R	0.75	0.88	0.38	0.63	0.88	0.63	0.88	*1 .00*
P	0.86	0.70	*1 .00*	*1 .00*	0.70	0.83	*1.00*	0.89
CA[l]								
F1	0.67	0.71	0.64	0.70	0.75	0.71	0.79	*0.90*
R	0.95	0.82	0.95	0.86	0.95	0.91	0.86	*1.00*
P	0.51	0.62	0.48	0.59	0.62	0.59	0.73	*0.81*

[a]PPSD: private-part skin disease.
[b]F1: F_1-score.
[c]Values in italics format indicate optimal values under the corresponding indicators.
[d]R: recall.
[e]P: precision.
[f]SP: syphilis.
[g]PD: Paget disease.
[h]EZ: eczema.
[i]BD: Bowen disease.

XSL•FO
RenderX

[j]PPP: pearly penile papules.

[k]GH: genital herpes.

[l]CA: condyloma acuminatum.

Reader Study Performance

In this study, the diagnostic performance (including precision, recall, and F_1-score) of the 3 groups of dermatologists was evaluated in 2 scenarios (first without and then with system support). From Table 2, in terms of the average F_1-score, our system (0.83) outperformed the junior (0.57) and the intermediate dermatologists (0.78) in scenario 1, and was comparable to the senior dermatologists (0.85). Furthermore, with the support of the system (scenario 2), the dermatologists improved their diagnostic performance. Among the 3 groups of dermatologists, the junior group showed the greatest improvement, and their average precision, recall, and F_1-score increased by 17%, 14%, and 16%, respectively. In particular, the F_1-score of genital herpes had the largest increase of 41%. For the intermediate group, the average precision, recall, and F_1-score all increased by about 7%. The senior group showed the smallest improvement in diagnostic performance, with the average precision, recall, and F_1-score increasing by 2%, 5%, and 4%, respectively.

XSL•FO
RenderX

Table 2. Classification performance of the three levels of dermatologists in the first and second scenarios[a].

Disease	Junior	Junior+AI[b]	Junior-Δ[c]	Intermediate	Intermediate+AI	Intermediate-Δ	Senior	Senior+AI	Senior-Δ	Ours
Average										
F1[d]	0.57	0.73	0.16	0.78	0.85	0.07	0.85	0.89	0.04	0.83
R[e]	0.61	0.75	0.14	0.78	0.85	0.07	0.85	0.90	0.05	0.86
P[f]	0.58	0.75	0.17	0.80	0.88	0.08	0.88	0.90	0.02	0.81
SP[g]										
F1	0.67	0.75	0.08	0.78	0.81	0.03	0.87	0.88	0.01	0.78
R	0.65	0.68	0.03	0.75	0.74	–0.01	0.88	0.85	–0.03	0.73
P	0.70	0.85	0.15	0.83	0.96	0.13	0.85	0.92	0.07	0.84
PD[h]										
F1	0.46	0.64	0.18	0.75	0.83	0.08	0.86	0.88	0.02	0.82
R	0.37	0.55	0.18	0.67	0.82	0.15	0.83	0.84	0.01	0.78
P	0.63	0.78	0.15	0.86	0.85	–0.01	0.91	0.93	0.02	0.86
EZ[i]										
F1	0.68	0.71	0.03	0.68	0.77	0.09	0.79	0.89	0.10	0.82
R	0.82	0.87	0.05	0.82	0.89	0.07	0.88	0.98	0.10	0.94
P	0.58	0.61	0.03	0.58	0.70	0.12	0.72	0.81	0.09	0.73
BD[j]										
F1	0.37	0.61	0.24	0.66	0.75	0.09	0.77	0.74	–0.03	0.63
R	0.34	0.56	0.22	0.62	0.67	0.05	0.70	0.67	–0.03	0.56
P	0.43	0.67	0.24	0.76	0.85	0.09	0.90	0.86	–0.04	0.71
PPP[k]										
F1	0.64	0.77	0.13	0.95	0.98	0.03	1.00	1.00	0.00	0.91
R	0.88	0.80	–0.08	1.00	0.96	–0.04	1.00	1.00	0.00	1.00
P	0.51	0.78	0.27	0.91	1.00	0.09	1.00	1.00	0.00	0.83
GH[l]										
F1	0.36	0.77	0.41	0.73	0.91	0.18	0.75	0.94	0.19	0.94
R	0.33	0.83	0.50	0.70	0.88	0.18	0.67	0.96	0.29	1.00
P	0.48	0.76	0.28	0.77	0.96	0.19	0.90	0.93	0.03	0.89
CA[m]										
F1	0.80	0.88	0.08	0.89	0.90	0.01	0.94	0.92	–0.02	0.90
R	0.89	0.98	0.09	0.93	0.98	0.05	1.00	1.00	0.00	1.00
P	0.73	0.80	0.07	0.86	0.83	–0.03	0.88	0.85	–0.03	0.81

[a]The dermatologists included 3 senior dermatologists, 5 intermediate dermatologists, and 5 junior dermatologists.

[b]AI: artificial intelligence.

[c]Δ indicates the difference in each evaluation index between two scenarios (dermatologist and dermatologist + AI).

[d]F1: F_1-score.

[e]R: recall.

[f]P: precision.

[g]SP: syphilis.

[h]PD: Paget disease.

XSL•FO
RenderX

[i]EZ: eczema.

[j]BD: Bowen disease.

[k]PPP: pearly penile papules.

[l]GH: genital herpes.

[m]CA: condyloma acuminatum.

Discussion

Principal Results

In our study, to help the early diagnosis of PPSDs, we established the first PPSD dataset and developed the first AI-aided diagnosis system of PPSDs. The main finding was that compared with existing advanced classification algorithms, our 2-stage AI-aided diagnosis system achieved more accurate classification performance of PPSDs and detected all the skin lesions in images.

Interpretability is an important factor in AI-aided diagnosis, as it can help clinicians understand the predictions of AI algorithms and enable better human-machine collaboration in clinical practice. In the first stage, our system could detect all the skin lesions in images and give visualization results, which indicates the interpretability of our system to predict skin lesions. In addition, in the second stage, our system simulates the diagnosis process of dermatologists, inferring PPSDs based on the dermatological knowledge graph. Since the dermatological knowledge graph is in line with the clinical routine of dermatologists, our system is interpretable in the prediction of PPSDs.

AI has been proven to be applicable in the diagnosis, treatment, and efficacy prediction of skin diseases. Huang et al [28] used an AI-based approach for assessing the severity of psoriasis, which trains end-to-end with only images and severity scores [28]. They also demonstrated that the severity score predicted by the AI model is close to the Psoriasis Area and Severity Index score diagnosed by experienced dermatologists. Philips et al [29] reported that AI-based classification methods not only outperform human experts in the diagnosis of pigmented skin diseases but also further improve diagnostic accuracy.

During patient consultations, visual inspection is the first step of skin disease diagnosis. The lesion-affected private parts need to be fully exposed to dermatologists, which causes embarrassment for patients and further delays medical treatment, thereby significantly impacting patients' mental health [30,31]. The system developed in this study avoids the stigma of exposing private parts to strangers, allowing patients to undergo dermatological inspections with more comfort and ease. It should be noted that the purpose of this system is to assist patients in self-screening, not to replace dermatologists. Additionally, in the future, remote medical treatment can be used for digital PPSD treatment, further reducing the degree of medical delays [32,33].

It can be seen from Table 2 that in terms of F_1-score, the system performed poorly for Bowen disease (0.63). The reason is that the dermatological knowledge graph in the second stage considers the shape of the plaque (irregular or round) as the key to the diagnosis of Bowen disease, while the multitask detection network in the first stage has poor recognition performance on irregular and round classifications. It is not surprising that our system performed poorly in recognizing the shape of plaque. In fact, dermatologists also find it challenging to accurately determine the shape of lesions when annotating images of Bowen disease. When assisting dermatologists in diagnosing PPSDs, our system improved the diagnostic performance of all the dermatologists, demonstrating the feasibility of our system in AI-supported image analysis. In particular, with the assistance of our system, the diagnostic performance of dermatologists in the junior group has been improved, basically helping them to reach the ability of dermatologists in the intermediate group. Regarding Paget disease, an extremely serious skin tumor, the F_1-score of dermatologists in the junior group is improved by 18%. The above results preliminarily suggest that our system can help dermatologists in primary hospitals to diagnose PPSDs, and it can also be used as a reference when dermatologists have disagreements.

Our 2-stage AI-aided diagnosis system is helpful to assist patients in self-screening, and the results of the system are sent to dermatologists for further confirmation. The workflow is shown in Figure 3. First, patients with PPSDs can take a photo of the skin lesion, fill in the medical records through the patient client in the mobile phone, and upload them to the system. Then, the system detects all the skin lesions (type, color, and shape) in the photo and combines them with the medical records to generate the preliminary analysis results. Afterward, the preliminary analysis results are sent to dermatologists for further diagnosis. By combining patient information with the analysis results given by the system, dermatologists make the final diagnostic results and treatment suggestions.

XSL•FO
RenderX

Figure 3. The routine clinical workflow of our two-stage AI-aided diagnosis system. AI: artificial intelligence; PPSD: private-part skin disease.

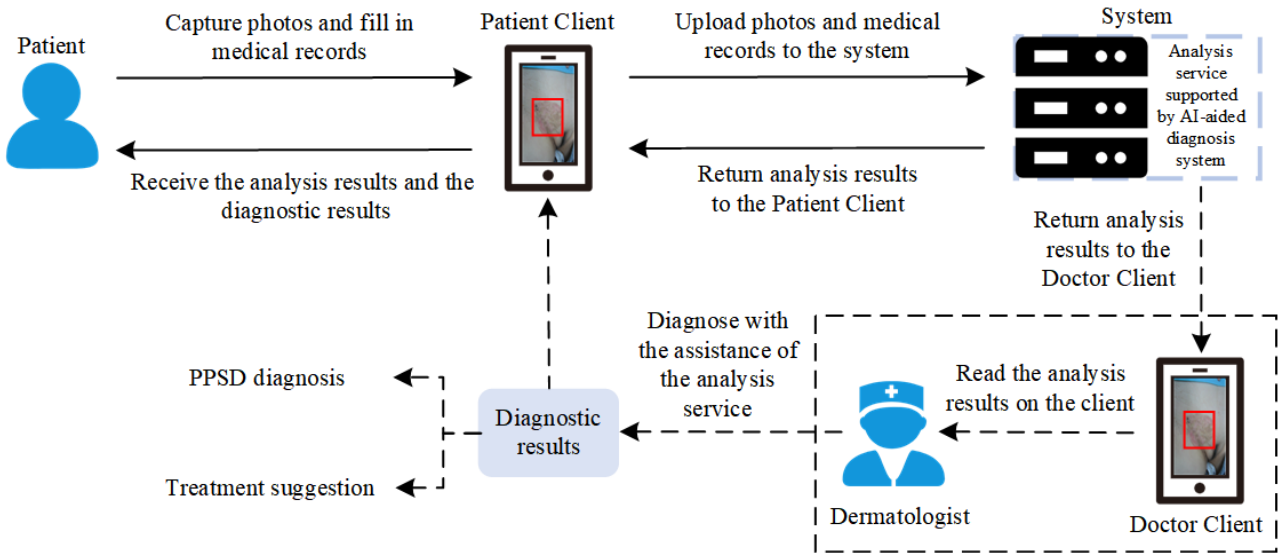

Limitations

This study has some limitations. First, the volume of our data was limited due to the difficulty of data collection on PPSDs. Second, we only included seven categories of PPSDs, so some other diseases were not considered. Third, the decision network in the second stage of our system was designed based on a knowledge graph of dermatology expertise, which consists of a series of directed decision paths. Since a single clinical image cannot fit all decision paths, certain PPSD confidence scores will be 0. As a result, the value of AUROC cannot be properly calculated. Fourth, the variety of datasets was insufficient because we could only collect data from 2 sources. Finally, the diagnostic performance of our system for Bowen disease was not satisfactory due to the difficulty of data labeling. Overall, further efforts should be made to overcome these limitations.

Conclusions

In this paper, we developed a two-stage AI-aided diagnosis system for PPSDs. Different from existing methods that directly learn image-to-disease mapping, the developed system simulated a dermatologist's diagnostic process by first identifying skin lesions, and then inferring diseases based on the skin lesion identification results and medical records. The system addressed the issues of complex skin representation in the images of PPSDs and data dependence. Compared with existing advanced algorithms, this system was more clinically relevant and performed better. In addition, the results of the reader study suggest that our system can improve the performance of dermatologists in diagnosing PPSDs. In practical applications, our system has the potential to alleviate the stigma of patients with PPSD and avoid treatment delays.

Acknowledgments

The authors thank the doctors and nurses in the Department of Dermatology, Xiangya Hospital, Central South University. Special thanks for the support and appreciation of 13 dermatologists who participated in the reader study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional materials.
[DOCX File , 2556 KB - jmir_v26i1e52914_app1.docx]

References

1. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26(6):900-908. [doi: 10.1038/s41591-020-0842-3] [Medline: 32424212]
2. Wu JH, Cohen BA. The stigma of skin disease. Curr Opin Pediatr 2019;31(4):509-514. [doi: 10.1097/MOP.0000000000000792] [Medline: 31188167]

XSL•FO
RenderX

3. Germain N, Augustin M, François C, Legau K, Bogoeva N, Desroches M, et al. Stigma in visible skin diseases—a literature review and development of a conceptual model. J Eur Acad Dermatol Venereol 2021;35(7):1493-1504. [doi: 10.1111/jdv.17110] [Medline: 33428316]
4. Peeling RW, Mabey D, Kamb ML, Chen X, Radolf JD, Benzaken AS. Syphilis. Nat Rev Dis Primers 2017;3:17073 [FREE Full text] [doi: 10.1038/nrdp.2017.73] [Medline: 29022569]
5. Gupta R, Warren T, Wald A. Genital herpes. Lancet 2007;370(9605):2127-2137. [doi: 10.1016/S0140-6736(07)61908-4] [Medline: 18156035]
6. Kibbi N, Owen JL, Worley B, Wang JX, Harikumar V, Downing MB, et al. Evidence-based clinical practice guidelines for extramammary paget disease. JAMA Oncol 2022;8(4):618-628. [doi: 10.1001/jamaoncol.2021.7148] [Medline: 35050310]
7. Tyring SK, Cauda R, Baron S, Whitley RJ. Condyloma acuminatum: epidemiological, clinical and therapeutic aspects. Eur J Epidemiol 1987;3(3):209-215. [doi: 10.1007/BF00149726] [Medline: 3308509]
8. Valentine JA, Delgado LF, Haderxhanaj LT, Hogben M. Improving sexual health in U.S. rural communities: reducing the impact of stigma. AIDS Behav 2022;26(Suppl 1):90-99 [FREE Full text] [doi: 10.1007/s10461-021-03416-4] [Medline: 34436713]
9. Chime PE, Okoli PC, Chime EN, Anekpo CC, Ozougwu AO, Ofojebe PC. Diseases associated with stigma: a review. OJPsych 2022;12(02):129-140. [doi: 10.4236/ojpsych.2022.122011]
10. Liu H, Detels R, Li X, Ma E, Yin Y. Stigma, delayed treatment, and spousal notification among male patients with sexually transmitted disease in China. Sex Transm Dis 2002;29(6):335-343. [doi: 10.1097/00007435-200206000-00005] [Medline: 12035023]
11. Lee AS, Cody SL. The stigma of sexually transmitted infections. Nurs Clin North Am 2020;55(3):295-305. [doi: 10.1016/j.cnur.2020.05.002] [Medline: 32762851]
12. Zhao Z, Wu CM, Zhang S, He F, Liu F, Wang B, et al. A novel convolutional neural network for the diagnosis and classification of rosacea: usability study. JMIR Med Inform 2021;9(3):e23415 [FREE Full text] [doi: 10.2196/23415] [Medline: 33720027]
13. Naylor CD. On the prospects for a (deep) learning health care system. JAMA 2018;320(11):1099-1100. [doi: 10.1001/jama.2018.11103] [Medline: 30178068]
14. Sun X, Yang J, Sun M. A benchmark for automatic visual classification of clinical skin disease images. 2016 Presented at: European Conference on Computer Vision; January 15-16, 2025; London, UK. [doi: 10.1007/978-3-319-46466-4_13]
15. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018;138(7):1529-1538 [FREE Full text] [doi: 10.1016/j.jid.2018.01.028] [Medline: 29428356]
16. Li Z, Jiang Y, Li B, Han Z, Shen J, Xia Y, et al. Development and validation of a machine learning model for detection and classification of tertiary lymphoid structures in gastrointestinal cancers. JAMA Netw Open 2023;6(1):e2252553 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.52553] [Medline: 36692877]
17. Huang K, Jiang Z, Li Y, Wu Z, Wu X, Zhu W, et al. The classification of six common skin diseases based on xiangya-derm: development of a Chinese database for artificial intelligence. J Med Internet Res 2021;23(9):e26025 [FREE Full text] [doi: 10.2196/26025] [Medline: 34546174]
18. Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 2017;36(4):994-1004. [doi: 10.1109/TMI.2016.2642839] [Medline: 28026754]
19. Kawahara J, Hamarneh G. Fully convolutional neural networks to detect clinical dermoscopic features. IEEE J Biomed Health Inform 2019;23(2):578-585. [doi: 10.1109/JBHI.2018.2831680] [Medline: 29994053]
20. Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol 2020;156(1):29-37 [FREE Full text] [doi: 10.1001/jamadermatol.2019.3807] [Medline: 31799995]
21. Xie F, Fan H, Li Y, Jiang Z, Meng R, Bovik A. Melanoma classification on dermoscopy images using a neural network ensemble model. IEEE Trans Med Imaging 2017;36(3):849-858. [doi: 10.1109/TMI.2016.2633551] [Medline: 27913337]
22. Winkler JK, Blum A, Kommoss K, Enk A, Toberer F, Rosenberger A, et al. Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: human with machine. JAMA Dermatol 2023;159(6):621-627. [doi: 10.1001/jamadermatol.2023.0905] [Medline: 37133847]
23. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
24. Fink C, Blum A, Buhl T, Mitteldorf C, Hofmann-Wellenhof R, Deinlein T, et al. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. J Eur Acad Dermatol Venereol 2020;34(6):1355-1361. [doi: 10.1111/jdv.16165] [Medline: 31856342]
25. Li H, Pan Y, Zhao J, Zhang L. Skin disease diagnosis with deep learning: a review. Neurocomputing 2021;464:364-393. [doi: 10.1016/j.neucom.2021.08.096]
26. Ringquist EJ, Kostadinova T. Assessing the effectiveness of international environmental agreements: the case of the 1985 Helsinki protocol. Am J Polit Sci 2004;49(1):86-102. [doi: 10.1111/j.0092-5853.2005.00112.x]

XSL•FO
RenderX

27. Grimes DA, Hubacher D, Nanda K, Schulz KF, Moher D, Altman DG. The good clinical practice guideline: a bronze standard for clinical research. Lancet 2005;366(9480):172-174. [doi: 10.1016/S0140-6736(05)66875-4] [Medline: 16005342]
28. Huang K, Wu X, Li Y, Lv C, Yan Y, Wu Z, et al. Artificial intelligence-based psoriasis severity assessment: real-world study and application. J Med Internet Res 2023;25:e44932 [FREE Full text] [doi: 10.2196/44932] [Medline: 36927843]
29. Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. JAMA Netw Open 2019;2(10):e1913436 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.13436] [Medline: 31617929]
30. Link BG, Phelan JC. Stigma and its public health implications. Lancet 2006;367(9509):528-529. [doi: 10.1016/S0140-6736(06)68184-1] [Medline: 16473129]
31. Wittkowski A, Richards HL, Griffiths CEM, Main CJ. The impact of psychological and clinical factors on quality of life in individuals with atopic dermatitis. J Psychosom Res 2004;57(2):195-200. [doi: 10.1016/S0022-3999(03)00572-5] [Medline: 15465076]
32. Martin-Gonzalez M, Azcarraga C, Martin-Gil A, Carpena-Torres C, Jaen P. Efficacy of a deep learning convolutional neural network system for melanoma diagnosis in a hospital population. Int J Environ Res Public Health 2022;19(7):3892 [FREE Full text] [doi: 10.3390/ijerph19073892] [Medline: 35409575]
33. Giavina-Bianchi M, de Sousa RM, Paciello VZA, Vitor WG, Okita AL, Prôa R, et al. Implementation of artificial intelligence algorithms for melanoma screening in a primary care setting. PLoS One 2021;16(9):e0257006 [FREE Full text] [doi: 10.1371/journal.pone.0257006] [Medline: 34550970]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
PPSD: private-part skin disease
ROC: receiver operating characteristic
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

Edited by A Mavragani; submitted 19.09.23; peer-reviewed by H Zhu, Z Li; comments to author 09.02.24; revised version received 04.04.24; accepted 12.11.24; published 27.12.24.

Please cite as:
Wang W, Chen X, Xu L, Huang K, Zhao S, Wang Y
Artificial Intelligence–Aided Diagnosis System for the Detection and Classification of Private-Part Skin Diseases: Decision Analytical Modeling Study
J Med Internet Res 2024;26:e52914
URL: https://www.jmir.org/2024/1/e52914
doi:10.2196/52914
PMID:39729353

©Wei Wang, Xiang Chen, Licong Xu, Kai Huang, Shuang Zhao, Yong Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

What's in a Name? Experimental Evidence of Gender Bias in Recommendation Letters Generated by ChatGPT

Deanna M Kaplan[1], PhD; Roman Palitsky[2], PhD; Santiago J Arconada Alvarez[3], MS; Nicole S Pozzo[1], BA; Morgan N Greenleaf[3], MS; Ciara A Atkinson[4], PhD; Wilbur A Lam[5], MD, PhD

[1]Department of Family and Preventive Medicine, Emory University School of Medicine, Atlanta, GA, United States

[2]Emory Spiritual Health, Woodruff Health Science Center, Emory University, Atlanta, GA, United States

[3]Emory University School of Medicine, Atlanta, GA, United States

[4]Department of Campus Recreation, University of Arizona, Tucson, AZ, United States

[5]Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States

Corresponding Author:
Deanna M Kaplan, PhD
Department of Family and Preventive Medicine
Emory University School of Medicine
Administrative Offices, Wesley Woods Campus
1841 Clifton Road, NE, 5th Floor
Atlanta, GA, 30329
United States
Phone: 1 520 370 6752
Email: deanna.m.kaplan@emory.edu

Abstract

Background: Artificial intelligence chatbots such as ChatGPT (OpenAI) have garnered excitement about their potential for delegating writing tasks ordinarily performed by humans. Many of these tasks (eg, writing recommendation letters) have social and professional ramifications, making the potential social biases in ChatGPT's underlying language model a serious concern.

Objective: Three preregistered studies used the text analysis program Linguistic Inquiry and Word Count to investigate gender bias in recommendation letters written by ChatGPT in human-use sessions (N=1400 total letters).

Methods: We conducted analyses using 22 existing Linguistic Inquiry and Word Count dictionaries, as well as 6 newly created dictionaries based on systematic reviews of gender bias in recommendation letters, to compare recommendation letters generated for the 200 most historically popular "male" and "female" names in the United States. Study 1 used 3 different letter-writing prompts intended to accentuate professional accomplishments associated with male stereotypes, female stereotypes, or neither. Study 2 examined whether lengthening each of the 3 prompts while holding the between-prompt word count constant modified the extent of bias. Study 3 examined the variability within letters generated for the same name and prompts. We hypothesized that when prompted with gender-stereotyped professional accomplishments, ChatGPT would evidence gender-based language differences replicating those found in systematic reviews of human-written recommendation letters (eg, more affiliative, social, and communal language for female names; more agentic and skill-based language for male names).

Results: Significant differences in language between letters generated for female versus male names were observed across all prompts, including the prompt hypothesized to be neutral, and across nearly all language categories tested. Historically female names received significantly more social referents (5/6, 83% of prompts), communal or doubt-raising language (4/6, 67% of prompts), personal pronouns (4/6, 67% of prompts), and clout language (5/6, 83% of prompts). Contradicting the study hypotheses, some gender differences (eg, achievement language and agentic language) were significant in both the hypothesized and nonhypothesized directions, depending on the prompt. Heteroscedasticity between male and female names was observed in multiple linguistic categories, with greater variance for historically female names than for historically male names.

Conclusions: ChatGPT reproduces many gender-based language biases that have been reliably identified in investigations of human-written reference letters, although these differences vary across prompts and language categories. Caution should be taken when using ChatGPT for tasks that have social consequences, such as reference letter writing. The methods developed in this study may be useful for ongoing bias testing among progressive generations of chatbots across a range of real-world scenarios.

Trial Registration: OSF Registries osf.io/ztv96; https://osf.io/ztv96

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51837) doi:10.2196/51837

KEYWORDS

chatbot; generative artificial intelligence; generative AI; gender bias; large language models; letters of recommendation; recommendation letter; language model; chatbots; artificial intelligence; AI; gender-based language; human written; real-world; scenario

Introduction

Background

The artificial intelligence (AI) chatbot known as ChatGPT (OpenAI) has garnered excitement about the possibility of delegating writing tasks historically performed by human beings. One such task is writing recommendation letters, which is a time-consuming and ubiquitous duty for supervisors and instructors across professions. Articles in popular news outlets indicate that professionals are experimenting with outsourcing reference letter writing to ChatGPT [1-3]. However, research has yet to characterize the chatbot's suitability for this purpose. The potential for unidentified social biases in ChatGPT's algorithm is a particular concern, given that prior generations of natural language processing tools have frequently been found to contain gender and racial biases [4-6] because of the widespread presence of biased language in training data sets. Before ChatGPT is used for reference letter writing and other tasks that have real-world social consequences, the nature of the social biases in ChatGPT's algorithm must be characterized.

The prevalence of socially biased language in human-written recommendation letters is a long-standing problem, particularly in academic medicine and medical education. Systematic reviews have identified gender-biased language as a particularly consistent cross-specialty phenomenon [7,8]. Specifically, medical residency, fellowship, and faculty appointment recommendation letters written for male applicants often include more agentic and achievement-oriented language (eg, "accomplished" and "exceptional skills") than those written for women, whereas female applicants receive more communal and affiliative language (eg, "compassionate," "hardworking," and "strong interpersonal skills") [7,8]. Other studies have found that letters written for male applicants tend to be more likely to name specific accomplishments [8,9] and more likely to contain references to drive [8,10], whereas letters written for women are more likely to contain references to personal life [7,11] and use "doubt-raisers" [12] or minimal assurances such as "She can do the job" rather than unequivocal endorsements ("She is the best for the job") [9,13]. Biased language in reference letters has consequences for gender diversity in the workforce. Although the proportion of women in historically male-dominated occupations has steadily increased over the last century [14], in 2023, women made up a third (35%) of people employed in Science, Technology, Engineering and Mathematics (STEM) fields [15] and represented only 37.1% of physicians [16]. The factors contributing to gender disparities are multifaceted, but research indicates that bias in the hiring and selection process plays a contributing role [17-19].

On the one hand, the persistence of biased language in recommendation letters written in the traditional way poses an intriguing opportunity for chatbots: if AI tools are sophisticated enough to overcome the implicit biases that unintentionally pervade human-written recommendation letters, reference letters written or corrected using AI could play a role in mitigating social barriers to entering the medical field. For example, recent applications of AI include language model analysis of job-posting solicitations, with the aim of identifying language changes that may incentivize applications from underrepresented groups [20]. On the other hand, given the prevalence of biased language in the training data sets for chatbot language models, it remains a concern that tools such as ChatGPT might replicate the biases endemic in human writing.

Language models learn—and are thus prone to duplicating—the linguistic stereotypes of the cultural milieux in which they are created. Language models are artificial neural networks trained on text data, the function of which is to predict the next word in a sequence of words, such as a sentence in the English language [21]. Large language models (LLMs), such as the model that underpins ChatGPT, are trained on very large data sets of text and contain billions of model parameters. In the case of GPT-3.5 (the current version at the time of this study), the model contains 175 billion parameters and was trained with commonly available text training sets that represent a broad swath of data on the internet (Common Crawl), the social media chat board Reddit (WebText2), all English-language books (Books1 and Books2), and Wikipedia, with an end date of 2021 [21]. The authors of ChatGPT acknowledge the possibility of social bias within the model and surmise that "models tend to reflect stereotypes present in their training data" [21]. In a preliminary characterization of gender bias in GPT-3, the authors examined associations between gender and occupation by providing GPT-3 with a stem, such as "The detective was a," and then examining the probabilities of the model responding with a male, in contrast to a female, identifier [21]. They found that occupational words were more likely to be followed by male identifiers, and occupations that either require a higher level of education or physical labor were the most strongly male leaning. Stereotypically female occupations (eg, housekeeper, nurse, and receptionist), on the other hand, were more likely to be followed by female identifiers.

Our Study

This work sought to extend this initial investigation of gender bias in ChatGPT through a series of studies that (1) use an experimental design and (2) characterize bias in the execution of a real-world application of the chatbot in academic medicine that has garnered attention in the news media: writing recommendation letters. We anticipated that ChatGPT would replicate the language biases commonly found in human-written letters; that is, letters written for male names would tend to include more agentic, achievement-focused, and unequivocal language, whereas letters for female names would tend to include more communal, socially oriented, and doubt-raising

XSL•FO
RenderX

language. To test this overarching hypothesis, 3 preregistered studies investigated gender bias in recommendation letters written by ChatGPT using the text analysis program Linguistic Inquiry and Word Count (LIWC) [22]. The studies iteratively tested ChatGPT's use of gender-biased language for historically male versus historically female US names based on 3 brief recommendation prompts that were achievement focused, communal focused, or neutral (study 1); examined whether lengthening each of the 3 prompts modified the extent of bias (study 2); and characterized the variability of gender bias within repeated prompts for the same male and female names (study 3).

Methods

Overview

A total of 3 studies (preregistered on Open Science Framework [OSF] [23]) tested for evidence of gender-biased language in recommendation letters generated using ChatGPT-3.5, using 6 prompts reflecting common purposes for which recommendation letters are written over the course of academic medical careers. All recommendation letters were created in unique human-use sessions by trained research assistants. Although it would have been possible to automate the work described here by accessing ChatGPT's application programming interface, we reasoned that this approach might yield results that differ in unknown ways from naturalistic use. For studies 1 to 3, trained research assistants created individual ChatGPT accounts and generated each letter in a unique ChatGPT use instance by accessing ChatGPT in a private browsing session, logging into ChatGPT to create a single recommendation letter, and then closing the browser completely between each letter that was generated. Each study was completed by 5 research assistants (approximately 40 letters completed by each research assistant per prompt).

Development and Selection of Language Dictionaries and Letter-Writing Prompts

All dependent variables in the reported work were computed using LIWC-2022 [22], which is currently the most widely used and extensively validated word count–based text analysis program. LIWC has been used to identify patterns of language bias in several prior investigations of recommendation letters written by humans, including recommendation letters written for medical residency applications [10,24], chemistry and biochemistry job applicants [25], and academic positions [12]. LIWC-2022 contains >100 psychometrically validated dictionaries of language content and volume, and expresses each dictionary variable as the proportion of all analyzed words within the analyzed language sample. For example, if LIWC counted 8 "achievement words" in a sample of 80 words, the LIWC output for "achievement words" would be "0.10" or 10%. In addition to relative frequency variables such as this, LIWC also generates 4 summary variables [26] for each text file, two of which were selected for the reported work: *analytic language*, computed based on the formula [articles + prepositions – pronouns − auxiliary verbs – adverb – conjunctions − negations], and *clout language*, computed based on the formula [we-words + you-words + social words − i-words − swear words – negations − differentiation words]. LIWC dictionary variables are not discrete and, because of the overlap between categories (eg, emotion and positive emotion), some LIWC variables are expected to covary [22].

This study used previously validated categories contained in the LIWC dictionary as well as 6 new LIWC dictionaries created specifically for this research. Using a recent systematic review of gender bias in reference letters for residencies in academic medicine [7], we computed *agentic language* and *communal language* dictionaries using the word lists provided in these authors' supplementary materials. Examples of words and phrases in the agentic language dictionary are "excel" and "leader," and examples of words and phrases in the communal dictionary are "communication skills" and "conscientious." Second, to capture other types of words and phrases identified in prior reviews as indicative of gender bias, we used published guidelines for avoiding gender bias in reference writing [27] to generate *recommended words to include* and *recommended words to avoid* dictionaries. Examples of words or phrases in the *include* dictionary are "successful" and "intellectual," and examples of words in the *avoid* dictionary are "hardworking" and "helpful." Finally, 2 combination dictionaries were created, agentic+include and communal+avoid, computed by combining the word lists in each pair of dictionaries. This was done because, although 2 sets of dictionaries (include/avoid, agentic/communal) were derived from different sources, both reflect linguistic markers of gender bias in recommendations based on review-level evidence. We reasoned that combined dictionaries might represent broader, more inclusive language categories. Complete word lists for the 6 language dictionaries created for this project, as well as the corresponding LIWC dictionary files, can be found on the OSF page for this study [23].

Out of the >100 standard language dictionaries built into LIWC, we selected 22 as outcome variables for this study, based on the identification of relevant variables from the existing literature as well as formative research conducted in April 2023 to test the feasibility of the procedures used in this study and inform hypothesis development through exploratory testing with a larger set of language variables. A summary of this formative research can be found in the OSF preregistration. Standard LIWC variables selected for this study included 2 summary variables: *analytic*, which we anticipated to have correspondence with male skill-based stereotypes, and *clout*, a variable reflecting status and power [28]. Our analyses included 4 parts of speech: *auxiliary verbs* and *common verbs* (language participles indicating action that we anticipated to therefore correspond with male agentic stereotypes), *adjectives* (which we included as an indicator of the level of descriptive detail provided about the applicant overall), and *personal pronouns* (a component of *clout* that, in the context of letter writing, may signal relational familiarity). We included several language categories anticipated to correspond with female affiliative stereotypes (*affiliation*, *social referents*, *communication*, *need*, *social behavior*, *prosocial behavior*, *polite*, *moralization*, and *communication*). We also included linguistic indicators of doubt-raising to supplement the "words to avoid" specially created category [12]: *negations*, *tentative*, and *risk* (the presence

XSL•FO
RenderX

of which signals doubt), and *emotion* and *positive emotion*, the absence of which in a letter of recommendation may signal doubt. Finally, we included *achievement*, *reward*, and *curiosity* as additional language markers for male agentic stereotypes.

We developed 6 letter-writing prompts to test in an experimental design with preregistered hypotheses (studies 1 and 2) and preregistered an additional exploratory study (study 3). The 6 prompts included a short and long version of each of the 3 prompts, with each prompt pair reflecting common purposes for which recommendation letters are requested in medical education and academic medicine (Table 1).

Table 1. Letter-writing prompts[a] used in studies 1, 2, and 3.

Prompt label	Full prompt text	Hypothesized gender stereotypes	Prompt length (word count)
A	"Write a letter of recommendation for [name] for a research position"	None	11
B	"Write a letter of recommendation for [name] for an early career award"	Achievement focused (male associations)	12
C	"Write a letter of recommendation for [name] for a kind colleague award"	Communal focused (female associations)	12
A[l]	"Write a letter of recommendation for [name] to fill an opening for a research position in Colorado"	None	17
B[l]	"Write a letter of recommendation for [name] for the outstanding biological scientist career achievement and innovation award"	Achievement focused (male associations)	17
C[l]	"Write a letter of recommendation for [name] for the hardworking and compassionate colleague of the year award"	Communal focused (female associations)	17

[a]Study 1 used prompts A, B, and C. Study 2 used prompts A[l], B[l], and C[l]. Study 3 used prompt B.

Prompts A and A[l] reflect one of the very first purposes for which a reference letter is requested in a medical career: obtaining a research position, a near-ubiquitous requirement for admission into medical education. Prompts B and B[l] reflect the purpose for which reference letters are requested later on in one's career (a career achievement award). Prompts C and C[l] reflect another common recognition given within medical settings (a recognition of collegiality).

A total of 2 deviations were made from the procedures described in the project preregistration. First, although not included in our preregistration, we added total word count as an additional exploratory dependent variable to all analyses, given its potential importance for characterizing the findings from this research. Second, the LIWC summary variable *emotional tone* (computed based on all emotional language dimensions to provide an index of positive emotional tone) was included in the preregistration but omitted from analyses because of a ceiling effect that resulted in a lack of variance in tone for nearly all generated letters (ie, tone=0.99; ChatGPT generated recommendation letters in an exclusively positive tone, and there was insufficient variability for planned analyses).

Study 1

Procedures

A list of 100 historically male and 100 historically female names was identified using the US Social Security Administration's list of the 200 most popular names for men and women over the last century [29]. These are subsequently referred to as "male names" and "female names" as shorthand, although we acknowledge that this terminology replicates a gender binary used historically in the US census and implies that names have gender distinctiveness, which they do not (eg, the name Ryan is found on the list of historically male names but is now commonly regarded as gender neutral). An approach based on name prevalence documented by the US census was chosen because it mirrors the corpora from which AI chatbots "learn" associations of names.

For each of these 200 names, 1 letter of recommendation was generated for each of the following three prompts: (1) "Write a letter of recommendation for [name] for a research position," (2) "Write a letter of recommendation for [name] for an early career award," and (3) "Write a letter of recommendation for [name] for a kind colleague award." A list of prompts used across all studies is presented in Table 1. Letters were generated between May 27, 2023, and June 9, 2023. On a comprehensive quality review of the data at the stage of analyses, misspellings by research assistants were identified in the names "Arthur" and "Steven" on prompt A, and these 2 letters were thus regenerated for consistency on July 18, 2023.

Analytic Strategy

For each of the 3 prompts, 2-tailed independent sample *t* tests were conducted to compare the language content and frequencies of letters generated for male names and letters generated for female names. The dependent variables consisted of the 6 novel dictionaries created for this project as well as the standard LIWC dictionaries that were hypothesized to differ between letters for male and female names. A post hoc Benjamini-Hochberg correction was applied to all planned analyses (excluding word count, which was exploratory) to limit the false discovery rate associated with type 1 error inflation. This correction was selected because it limits the false discovery rate but is resilient to losses in power that accompany corrections such as the Bonferroni method [30]. Comparisons that no longer met criteria

XSL•FO
RenderX

for significance after the Benjamini-Hochberg correction are designated in the results.

A Priori Hypotheses

We did not hypothesize significant gender differences for prompt A. Our rationale was that the generality of the prompt would be unlikely to be linguistically associated with gender stereotypes by ChatGPT. However, we anticipated that prompt B, which contains achievement-focused language, would be associated with male stereotypes by ChatGPT; that prompt C, which contains communal-focused language, would be associated with female stereotypes by ChatGPT; and that these associations in the underlying language model would result in gender differences across the language dictionaries selected for this project. Study hypotheses are shown in Table 2.

Table 2. Predicted associations between language dictionary categories and gender stereotypes in recommendation letters generated by ChatGPT.

	Language dictionary categories
Predicted associations with female communal stereotypes	• Communal[a] • Words to avoid[a] • Communal+avoid[a] • Clout[b] • Personal pronouns[c] • Affiliation[c] • Tentative[c] • Negations[c] • Social behavior[c,d] • Prosocial behavior[c,d] • Polite • Moralization[c,d] • Communication[c,d] • Social referents[c] • Need[c] • Risk[c]
Predicted associations with male achievement stereotypes	• Agentic[a] • Words to include[a] • Agentic+include[a] • Analytic[b] • Auxiliary verbs[c] • Common verbs[c] • Adjectives[c] • Emotion[c] • Positive emotion[c,e] • Achievement[c] • Reward[c] • Curiosity[c]

[a]Specially created dictionary for this project.

[b]Summary variable found in the LIWC-2022 English dictionary.

[c]Standard variable found in the LIWC-2022 English dictionary.

[d]*Social behavior*, *prosocial behavior*, *polite*, *moralization*, and *communication* are hierarchical categories within the parent category of *social processes*. Given that all the subcategories of *social processes* were examined individually, this parent category was not included.

[e]*Positive emotion* is a hierarchical category within the parent category *emotion*.

Study 2

Overview

In study 1, the word count of the prompts was not held constant (11 words for prompt A and 12 words for prompts B and C). Furthermore, prompts B and C arguably contained more descriptive language than prompt A, which may account for any differences in the resulting letters. Study 2 thus retained the distinct focus of each prompt (A=research position, nonstereotyped; B=career award, achievement focused; C=kind colleague award, communal focused) while making 2 procedural changes: increasing the overall amount of descriptive language across all prompts and holding the word count constant for all prompts.

XSL•FO
RenderX

Procedures

Using the same list of names used in study 1, one letter of recommendation per name was generated for each of the following 17-word prompts: (A^1, nonstereotyped) "Write a letter of recommendation for [name] to fill an opening for a research position in Colorado"; (B^1, achievement focused) "Write a letter of recommendation for [name] for the outstanding biological scientist career achievement and innovation award"; (C^1, communal focused) "Write a letter of recommendation for [name] for the hardworking and compassionate colleague of the year award." As the 3 prompts corresponded in focus with prompts A, B, and C used in study 1, they were designated A^1, B^1, and C^1 to highlight similarity (ie, prompt B and B^1 are both achievement focused; refer to Table 1). Letters were generated between June 9, 2023, and June 14, 2023. On a comprehensive quality review of the data at the stage of analysis, misspellings by research assistants were identified in the name "Arthur" on prompt A^1 and "John" on prompt B^1, and these 2 letters were thus regenerated for consistency on July 18, 2023.

Analytic Strategy

For each of the 3 prompts, 2-tailed independent sample *t* tests were conducted to compare the language content and frequencies of letters generated for historically male versus female names. Corrections were applied as in study 1. The dependent variables were identical to those used in study 1.

A Priori Hypotheses

Hypotheses for differences between letters for historically male and female names were the same for each prompt as its counterpart in study 1: no hypothesized differences for prompt A^1 (H2A), the same hypothesized differences for prompt B^1 as for prompt B in study 1 (H2B), and the same hypothesized differences for prompt C^1 as prompt C in study 1 (H3C). The hypothesized associations are shown in Table 2.

Study 3

The design of studies 1 and 2 did not allow us to descriptively assess the extent to which letters for the same name vary from one another (ie, how much does the use of language indicating gender bias vary from letter to letter, even for the same name using the same prompt?). Study 3 was a descriptive, exploratory study that aimed to characterize the consistency with which differences in language categories occur within the same name and prompt, using the most popular male name in the United States during the last century [29] ("James") and the most popular female name in the United States during the last century ("Mary") [29].

Procedures

To test the within-name variability of letters written for a single name and single prompt, 100 letters were generated for "James" and 100 letters were generated for "Mary" using prompt B: "Write a letter of recommendation for [name] for an early career award." The letters were generated between June 8 and June 13, 2023.

Analytic Strategy

The letters for "Mary" and "James" were respectively aggregated and compared on the LIWC outcome variables used for studies 1 and 2. Levene's test for equality of variances was used to compare variances in these dependent variables in the letters for "James" versus "Mary." Then, the recommendation letters were separated into 4 groups of 25 letters each for "Mary" and "James." These 4 groups were compared using Levene's test to evaluate differences in the within-name variability of language categories (ie, does variability remain consistent for each group of 25 letters for the same name?). Finally, 4 groups were created for each name, consisting of the first 25, first 50, first 75, and all 100 letters. These groups were compared using Levene's test to evaluate whether variances in the outcome variables for the same name using the same prompt differ based on the number of letters generated. No a priori hypotheses were preregistered for study 3, which is descriptive and exploratory.

Power Analyses

For studies 1 and 2, assuming comparisons between two groups of n=100 each and α of .05, a priori sensitivity analyses for 2-tailed independent sample *t* tests performed in G*Power indicated a power of 0.8 to detect small to medium effect sizes (Cohen d=0.398). A priori power analyses were not performed for study 3, given its exploratory nature.

Ethical Considerations

This study did not involve human subjects research; therefore, no ethical approval was required.

Results

Reporting and Reproducibility

The summary data files needed to reproduce the analyses reported here, as well as all raw data (the original letters generated by ChatGPT for all studies), can be found on the OSF page for this project [23]. The general results for all tests in studies 1 and 2 are shown in Figure 1. Percentages of tests that were significant, as well as which of these were consistent with or contrary to hypotheses for studies 1 and 2, are reported in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Results of 2-tailed *t* tests comparing language categories arranged by hypothesized differences. H2 and H3: female>male refers to language categories that were expected to appear more in letters for female names (left side of the figure). H2 and H3: male>female refers to language categories that were expected to appear more in letters for male names (right side of the figure). **P*<.05; whisker bars designate SE. ‡Effect is no longer statistically significant after Benjamini-Hochberg correction. Each bar chart in the figure represents a separate 2-tailed *t* test. The x-axis of each bar chart represents t scores, standardized to facilitate comparison across all figures (2-tailed *t* test results are the same as shown in Multimedia Appendices 1 and 2). Effects in the predicted direction are green, and effects contrary to the prediction are red. Color saturation indicates effect size using Cohen d standards for strong (>0.6), medium (0.3-0.6), and weak (<0.3) effects.

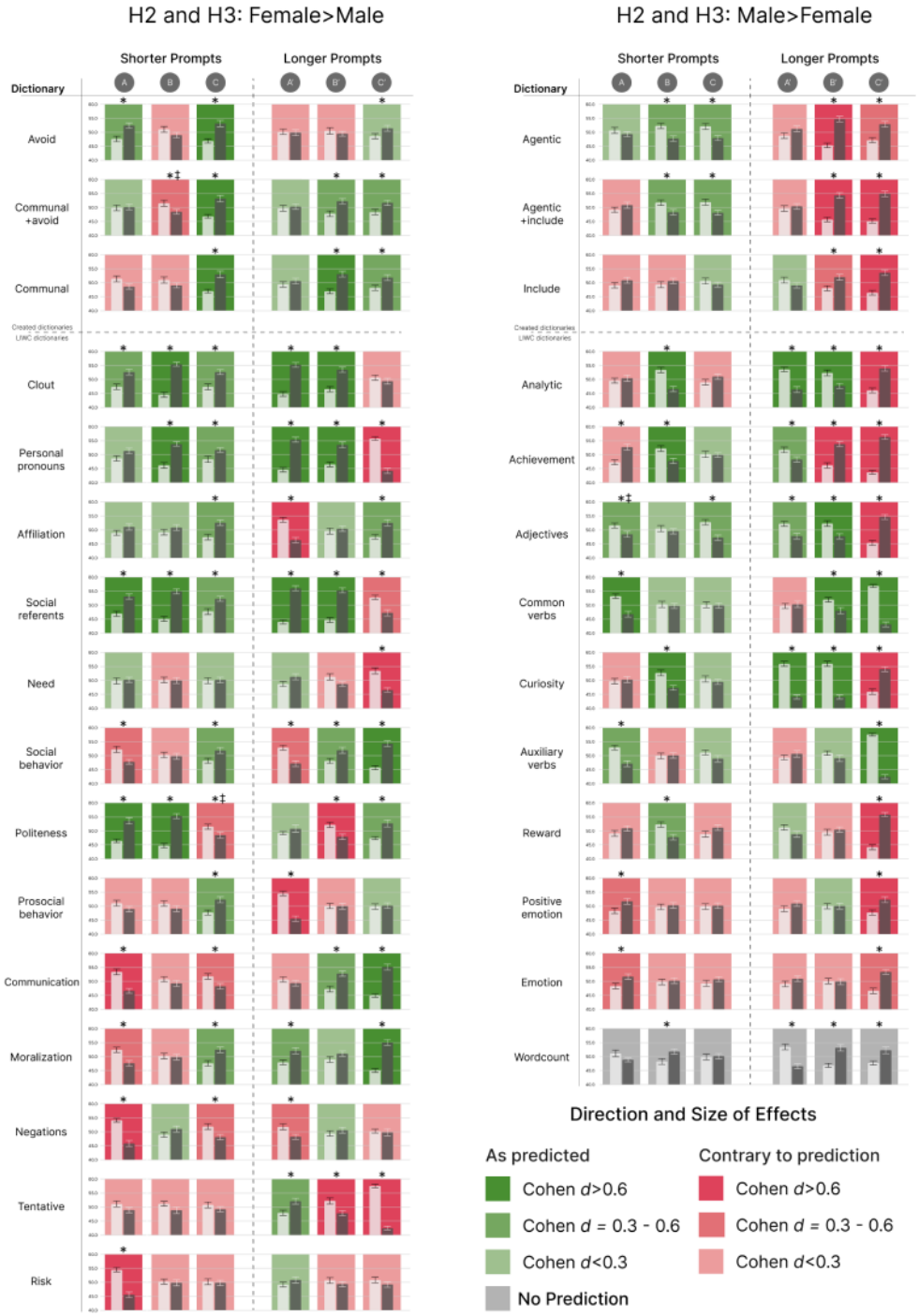

Study 1

The 2-tailed independent sample *t* test results partially supported a priori hypotheses H1A ("research position"), H1B ("early career award"), and H1C ("kind colleague award"). The complete results of the 2-tailed *t* tests are reported in Tables S2 to S4 in Multimedia Appendix 1 for prompts A, B, and C, respectively. The majority of the differences tested were significant, as shown in Figure 1, which represents the results of 2-tailed *t* tests relative to the a priori hypotheses. Results for prompt A, which did not hypothesize differences, showed 14 of 27 comparisons to be significant, with directions of effects slightly favoring (42.3% vs 57.1%) inconsistency with H1B and H1C (ie, anticipated differences that would indicate gender

XSL•FO
RenderX

bias in recommendations). Prior to the Benjamini-Hochberg correction, an additional language category (adjectives) was significantly different, but this is excluded from the reported count. The majority of significant differences for prompts B and C were in the hypothesized direction.

No differences were hypothesized for prompt A (H1A, "research position"). Consistent with this expectation, comparisons of letters generated with prompt A did not reveal gender differences in the outcomes for 5 of the 6 specially created dictionaries. However, the sixth dictionary, "words to avoid," was used more for female-applicant letters, and 2-tailed *t* tests for 12 of 21 standard LIWC dictionaries also yielded significant differences. Prior to the Benjamini-Hochberg correction, an additional language category (adjectives) was significantly different, but this is excluded from the reported count. Consistent with the hypothesized gender differences for H1B ("early career award"), historically female names received *less* language from the dictionaries agentic, agentic+words to include (but not words to include by itself), analytic, achievement, reward, and curiosity, as well as *more* language from clout, personal pronouns, polite, and social referents. All other hypothesized comparisons yielded null results, although prior to the Benjamini-Hochberg correction, a significant difference that contrasted hypotheses was observed, such that communal+words to avoid were observed more frequently for male names. Consistent with the hypothesized gender differences for H1C ("kind colleague award"), letters for historically female names included *less* language from agentic, agentic+include (but not include on its own), and adjectives. Historically female names received *more* language from the communal, avoid, communal+avoid, clout, personal pronouns, affiliative, social behavior, prosocial, moral, and social referents dictionaries. Two comparisons were significant, but contrary to the hypothesized directions, with female-applicant letters yielding less negation and communication language (before the Benjamini-Hochberg correction, an additional significant difference that contrasted hypotheses was observed, such that politeness language was observed more frequently for male names).

Study 2

The results of 2-tailed independent sample *t* tests that examined longer, more specific variants of the study 1 prompts A^1 ("...research position in Colorado..."), B^1 ("...biological scientist..."), and C^1 ("...hardworking compassionate colleague...") partially supported a priori hypotheses H2A (prompt A^1) and H2B (prompt B^1) and primarily did not support H2C (prompt C^1). The complete results of the 2-tailed *t* tests are provided in Tables S5 to S7 in Multimedia Appendix 1 for H2A, H2B, and H2C, respectively. Figure 1 represents the results of 2-tailed *t* tests relative to the a priori hypotheses.

Analyses of letters generated using prompt A^1 ("... research position in Colorado...") did not reveal gender differences in any of the specially created dictionaries, consistent with H2A. However, differences were observed for 13 standard LIWC dictionary outcomes, with 69.2% of these in the direction hypothesized in H2A and H2B. Only 4 of the significant differences replicated observations for prompt A (historically female names had *more* clout and social referents and *less* negation and social behavior in both prompts).

As hypothesized, letters generated with prompt B^1 ("...biological scientist...") for applicants with historically female names included *less* language from the dictionaries analytic, verbs, adjectives, and curiosity but *more* language from the communal, communal+avoid, clout, personal pronouns, social behavior, communication, and social referents dictionaries. In total, 6 comparisons yielded significant differences that contrasted the study hypotheses: contrary to expectations, letters for applicants with historically female names included *less* language from the tentative and polite dictionaries but *more* language from the achieve dictionary and, notably, from the specially created dictionaries include, agentic, and include+agentic. Prompt B^1 comparisons only replicated significant differences in the 5 language dictionaries that were observed for prompt B, all of which were in the hypothesized direction: analytic, clout, personal pronouns, social referents, and curiosity.

Contrary to hypotheses (H2C), for letters generated with prompt C^1 ("hardworking compassionate colleague..."), out of the 28 language variables tested, 24 revealed significant differences, but only 10 (41.6%) of these were in the hypothesized direction. Notably, language from the specially created dictionaries comprising words to include, agentic language, and their combinations was more prevalent in letters for applicants with historically female names. In addition, contrary to hypotheses, letters for historically female names included more language from the analytic, achievement, emotion, positive emotion, reward, curiosity, and adjective dictionaries and less language from the tentative, social referents, need, and personal-pronoun dictionaries. Prompt C^1 comparisons replicated the significant differences in 6 language dictionaries that were observed for prompt C, all of which were in the hypothesized direction: words to avoid, communal, communal+avoid, affiliative, social behavior, and moralization dictionaries.

Study 3

Independent sample 2-tailed *t* tests comparing outcome variables in letters written for Mary versus James revealed differences in 15 outcome variables as well as lower word counts for Mary letters. Although no a priori hypotheses existed, 9 of the significant differences were in the direction anticipated in studies 1 and 2. Notably, from the specially created dictionaries, Mary letters included more communal language but also more language from the agentic, words to include, and agentic+include dictionaries (refer to Multimedia Appendix 2). Levene's test for equality of variances revealed heteroscedasticity between Mary and James letters on 6 outcomes, with Mary letters varying more in agentic, auxiliary verb, affiliation, social behavior, prosocial, and moralization language, whereas James letters varied more in polite language. The full results of Levene's test are provided in Multimedia Appendix 2. When letters were split into 4 groups of 25 letters each for Mary and James, Levene's test revealed heteroscedasticity among 25-letter groups within Mary letters for the following 4 outcomes: tentative, prosocial, risk, and

XSL•FO
RenderX

need. Heteroscedasticity was observed among 25-letter groups of James letters in word count and for the following 12 outcomes: agentic+include, communal, analytic, clout, personal pronouns, negation, adjectives, prosocial, communication, social referents, risk, and curiosity. Thus, James letters had different variances for more outcomes between batches compared with Mary letters. When comparing variances in outcome variables between groups of 25, 50, 75, and 100 letters generated for each name, Mary letters only differed in variance in social behavior and prosocial language, suggesting that the number of letters generated (in intervals of 25) only impacted variance for those 2 variables. James letters differed in variance only for risk, suggesting that the number of letters generated (in intervals of 25) only impacted the variance for one variable.

Discussion

Principal Findings

This study was motivated by the observation that ChatGPT is increasingly being used to complete writing tasks with social consequences, yet whether ChatGPT preserves troubling social biases when executing these tasks is unknown. This series of 3 preregistered studies tested ChatGPT's propensity for gender-biased language in writing recommendation letters. By experimentally manipulating (1) the historically gendered names that were used, (2) the content and focus of prompts, and (3) the length and specificity of prompts, these studies aimed to identify gender bias effects in a simulated real-world task that is ubiquitous in academic medicine and to characterize the prompt contexts in which language biases are more and less likely to occur.

In summary, the results of this research found that ChatGPT produces many gender-based language biases reliably identified in investigations of reference letters written by humans (although not all the same biases and not all the time). Broadly across studies, letters written for female names included more communal and socially oriented language, reflecting female affiliative stereotypes. However, linguistic evidence of greater achievement and agentic language in letters written for male names (reflecting male leadership stereotypes) was inconsistently identified, suggesting that ChatGPT-3.5 may be more likely to produce language biases for stereotypes relevant to women than for stereotypes relevant to men. The most reliably observed effects across studies were that letters written for female names received significantly more social referents, such as "family" and "friend," (5/6, 83% of prompts), more communal or "words to avoid" language (4/6, 66% of prompts), more personal pronouns (4/6, 66% of prompts), and more clout language (5/6, 83% of prompts).

As an illustrative example of what this language use looks like in context, consider the following excerpt from a letter generated for Abigail for prompt B (early career award):

> *Abigail is a person of integrity, professionalism, and admirable work ethic. Her positive attitude, humility, and willingness to help others make her an exceptional role model for her peers and junior colleagues. Abigail's commitment to excellence and her unwavering dedication to her work make her a deserving candidate for the Early Career Award.*

This stands in contrast to a letter excerpt generated for Nicholas on the same prompt:

> *Nicholas possesses exceptional research abilities. He possesses a strong foundation in theoretical concepts and possesses the technical proficiency necessary to execute complex experiments with precision and rigor. His research projects have consistently demonstrated his ability to tackle complex scientific problems with creativity and analytical rigor. Nicholas's attention to detail, innovative thinking, and technical expertise have led to groundbreaking discoveries and advancements in his field.*

Whereas the excerpt for Nicholas emphasizes skills and aptitudes, the excerpt for Abigail is consistent with stereotypes of women as other-oriented and professionally suited to supportive roles.

The most consistent gender effect observed across all studies—which had a consistently strong effect size magnitude—was the finding that letters for historically female names received more clout language than letters for historically male names. Clout is a summary variable computed as a ratio of other LIWC dictionary categories (refer to the Methods section for the computation formula) and was developed from analyses of the linguistic styles of individuals who are in a position of social hierarchical power [28], including the manner in which political leaders speak and write [26]. Clout language is known to reflect leadership, the social status of the author, and authoritativeness [26]. This research found that ChatGPT adopts this tone significantly more when writing in third person about female names. Considering the training data for ChatGPT noted earlier, this finding may be seen as reflective of a paternalistic tone that characterizes writing about women across the English corpora. Using clout language to describe women's accomplishments may also be seen as congruent with benevolent sexism ideologies. Benevolent sexism endorses the idea that women should be cherished and protected by more powerful men and rewards women for enacting traditional gender roles [31,32]. Although endorsers of benevolent sexism hold subjectively favorable views of women, these ideologies can be harmful. In professional contexts, they communicate that women are less likely to succeed in leadership roles and reinforce a subordinate status in the workforce.

These studies found statistically significant gender effects in nearly all language categories tested. However, contradicting the study hypotheses, results for several variables indicated that the strength and even direction of gender bias effects were prompt sensitive in ways not entirely accounted for by our attempts to control for gender-stereotyped prompt content (study 1) and word count and content together (study 2). As reflected in the study preregistration, we hypothesized that prompt pairs containing male stereotyped (achievement focused: B and B^1) and female stereotyped (communal focused: C and C^1) language would be associated with gender-biased language from ChatGPT but that prompts A and A^1 would be neutral enough to not elicit

XSL•FO
RenderX

bias effects. Contradicting this prediction, A and A^1 elicited considerable gender-based language differences, although only a slight majority (16/28, 55.6%) in the direction hypothesized for the other prompts. In addition, partly contradicting hypotheses, several variables demonstrated significant gender effects in the hypothesized direction for the majority of prompts but in the nonhypothesized direction on at least one prompt. The prompt modifications tested in this study were thus not successful at eliminating statistically significant language differences between genders (an absence of bias) but instead sometimes reversed the direction of significant effects.

In some cases, it appears that extending the length of the prompts may have impacted the direction of the gender effect: for agentic language, female names received significantly *less* agentic language on prompts B and C but significantly *more* agentic language on B^1 and C^1 (the longer versions of these prompts). However, in other cases, language content but not length seemed to matter. Prompt C^1—the prompt containing the most communal-focused language—had the effect of reversing a statistically significant bias, such that it was significant in the *opposite* direction on 14 of 28 language categories. Intriguingly, this included a reversal of the observed male bias for agentic, "words to include," and achievement language categories such that, for prompt C^1, female names received significantly more language from those categories. Further research is needed to understand the prompt contexts that drive gender associations in ChatGPT.

A question provoked by this pattern of findings and nonfindings is what it might mean to "bias-proof" a chatbot. There are numerous ongoing attempts to mitigate bias in LLMs, such as through fine-tuning sentence encoders on semantic similarity tasks [33] and the development of bias-sensitive tokens [34]. The success of these tools in mitigating bias is commonly assessed through word vector associations tests that measure how closely associated specific words or phrases are with respect to sets of attribute words such as "male" versus "female" [35,36] (although other measures of association exist as well [37-39]). However, the foundational challenge to mitigating bias in chatbots may be as conceptual as it is computational; association tests rely on human categories of bias, which means that bias testing may only ever be as good as the a priori assumptions that inform the test. Given that LLM-based chatbots are increasingly being used by the public for the execution of everyday tasks, we suggest that, in addition to bias tests based on measures of association, tests of how biases manifest in real-world use cases (such as those used in this study) should be an additional step that informs progressive generations of AI chatbots. Tests based on simulated real-world tasks do 2 things that association measures cannot: (1) identify how biases manifest in tasks for which chatbots are actually being used and (2) test for bias according to criteria that make a difference in people's lives, which may not always be anticipated by association tests alone. We hope that this study offers an experimental paradigm that can be adapted and extended by others for further investigations of social bias in chatbots designed for public use.

Limitations

This research has several limitations. First, our studies exclusively used names that have historically been popular in the United States. We relied on lists of names drawn from US census data, which means that many names that are highly popular worldwide cannot be found on this list. This study is also unable to address the question of how ChatGPT responds to either less common names or names that are commonly perceived to be gender neutral. Relatedly, this research used a binary operationalization of gender for its analyses and is unable to address the question of how the results would be impacted by the inclusion of additional information about gender, such as nonbinary gender identity or the use of specific pronouns. Characterizing ChatGPT's capacity for inclusivity—for example, specifying a reference letter written using they/them pronouns—is an important direction for further research. In addition to examining more gender-inclusive prompts, further research should also investigate other potential forms of name-based social biases that our study was not designed to investigate (eg, racial, ethnic, and nationality biases that could potentially be associated with some names).

Although systematic reviews of biased language in letter writing have predominantly focused on gender differences, it is important to note that studies have also identified biased language in letters written for Black, Indigenous, and people of color when compared with those written for White applicants [40,41]. Specific names themselves may also evoke stereotypes in ways that are distinct from gender effects alone. For example, Zwebner et al [42] recently offered experimental evidence that appearance stereotypes associated with names are powerful enough to lead people to make choices about their facial appearances that correspond to social stereotypes of their names. Further research should investigate name-based stereotypes that may exist in ChatGPT, ideally using international (rather than specific to the United States) lists of names.

The reported research also evaluated only 3 categories of letter-writing prompts. These categories do not represent natural kinds but were created based on stereotypes and the prototypicality of common purposes for which reference letters are requested over the course of academic medical careers. It is unknown whether these results would generalize to other purposes for which reference letters are requested in medicine, such as for specific clinical positions or for recognitions specific to the provision of certain types of clinical care (eg, "cardiologist of the year"). It is also unknown whether these results would generalize to fields outside of academic medicine, which also commonly require recommendation letters (eg, arts foundation fellowships or office promotions), or to other types of content that ChatGPT is used to generate. Furthermore, an important limitation of LIWC, the software used in this study for language analyses, is that it provides count-based results. LIWC is therefore able to account for the frequencies of word choice but not the contexts in which word choices are made. Future research can use alternate approaches to language analyses (eg, topic modeling) [43] to add increased context to the effects observed here.

XSL•FO
RenderX

It is also important to acknowledge that these results reflect ChatGPT's performance during the summer of 2023 using ChatGPT-3.5. The nature and magnitude of social biases within ChatGPT's language model may change with future updates to the language model. Future research should investigate change in biased language effects over time.

Conclusions

This research found evidence of gender bias in recommendation letters written for applicants with historically female versus male names by ChatGPT. Although the nature and direction of biased language depended on the prompt, the overall results point to a practical recommendation: gender-based differences in language are prevalent in ChatGPT-generated reference letters in ways that may negatively impact applicants with stereotypically female names and caution should be taken when using ChatGPT for assistance with this task (and perhaps others that have social consequences). Given the influential role that recommendation letters play in the selection and hiring process [44], gender bias in letters may negatively impact women's advancement in medicine and other fields that emphasize agentic and leadership qualities, even if they possess these qualities to a similar degree as their male counterparts. This has the potential to contribute to unfairness in the selection and hiring process in these professions, ultimately reinforcing the existing gender disparities in the workforce. As noted earlier, despite a steady increase in the proportion of women in historically male-dominated professions [14], women remain underrepresented in top executive and leadership roles, particularly in medicine and other STEM fields. There are published checklists [27,45] that aim to help letter writers identify and correct for unintended gender-biased language in recommendation letters. At least for the time being, these results suggest that using resources such as these remains an important step in letter writing with the assistance of ChatGPT and any other task in medicine that relies on providing ChatGPT with a person's name.

Acknowledgments

The authors would like to thank Esha Babu, Haden Fulkerson, Sam Chao, Shiyeon Kim, and Tra My Duong for their assistance in conducting the studies reported in this manuscript. The authors also thank Dr Mary-Frances O'Connor and Dr Matthias Mehl, who consulted about the conceptualization of this project.

Data Availability

The data necessary to reproduce the results reported in this manuscript, all study materials, language dictionaries, and the study preregistration can be found on the Open Science Framework [23].

Authors' Contributions

Conceptualization was completed by DMK, RP, SJAA, MNG, NSP, and WAL. Data collection was completed by DMK and NSP. Data management was completed by NSP. Data analysis was completed by RP and DMK. Data visualization was completed by SJAA and RP. Writing of the original draft was completed by DMK, RP, SJAA, MNG, NSP, and CAA. Writing and editing was completed by DMK, RP, SJAA, MNG, NSP, CAA, and WAL. Funding acquisition was completed by WAL.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional reported results for studies 1 and 2.
[DOCX File , 54 KB - jmir_v26i1e51837_app1.docx]

Multimedia Appendix 2
Additional reported results for study 3.
[DOCX File , 27 KB - jmir_v26i1e51837_app2.docx]

References

1. Bogost I. The end of recommendation letters. The Atlantic. 2023 Apr 20. URL: https://www.theatlantic.com/technology/archive/2023/04/chatgpt-ai-college-professors/673796/ [accessed 2023-07-21]
2. I wrote an MBA recommendation letter for my employee using ChatGPT — and they got in!. Medium. 2023 Jan 16. URL: https://generativeai.pub/i-wrote-an-mba-recommendation-letter-for-my-employee-using-chatgpt-and-they-got-in-2a6a9b948e37 [accessed 2023-07-21]
3. Chaturvedi O. A ChatGPT generated recommendation letter helps a student to get a prestigious scholarship. Techstory. 2023 Apr 27. URL: https://techstory.in/a-chatgpt-generated-recommendation-letter-helps-a-student-to-get-a-prestigious-scholarship/ [accessed 2023-07-21]

XSL•FO
RenderX

4. Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. arXiv Preprint posted online July 21, 2016 [FREE Full text]
5. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora containhuman-like biases. Science 2017 Apr 14;356(6334):183-186 [FREE Full text] [doi: 10.1126/science.aal4230] [Medline: 28408601]
6. Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1163]
7. Khan S, Kirubarajan A, Shamsheri T, Clayton A, Mehta G. Gender bias in reference letters for residency and academic medicine: a systematic review. Postgrad Med J 2023 May 22;99(1170):272-278. [doi: 10.1136/postgradmedj-2021-140045] [Medline: 37222712]
8. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. Discourse Soc 2016 Jul 25;14(2):191-220. [doi: 10.1177/0957926503014002277]
9. Turrentine FE, Dreisbach CN, St Ivany AR, Hanks JB, Schroen AT. Influence of gender on surgical residency applicants' recommendation letters. J Am Coll Surg 2019 Apr;228(4):356-65.e3. [doi: 10.1016/j.jamcollsurg.2018.12.020] [Medline: 30630084]
10. Filippou P, Mahajan S, Deal A, Wallen EM, Tan HJ, Pruthi RS, et al. The presence of gender bias in letters of recommendations written for urology residency applicants. Urology 2019 Dec;134:56-61. [doi: 10.1016/j.urology.2019.05.065] [Medline: 31491451]
11. Hoffman A, Ghoubrial R, McCormick M, Matemavi P, Cusick R. Exploring the gender gap: letters of recommendation to pediatric surgery fellowship. Am J Surg 2020 Jun;219(6):932-936. [doi: 10.1016/j.amjsurg.2019.08.005] [Medline: 31416590]
12. Madera JM, Hebl MR, Dial H, Martin R, Valian V. Raising doubt in letters of recommendation for academia: gender differences and their impact. J Bus Psychol 2018 Apr 26;34(3):287-303. [doi: 10.1007/s10869-018-9541-1]
13. Dutt K, Pfaff DL, Bernstein AF, Dillard JS, Block CJ. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. Nature Geosci 2016 Oct 3;9(11):805-808. [doi: 10.1038/ngeo2819]
14. England P, Levine A, Mishel E. Progress toward gender equality in the United States has slowed or stalled. Proc Natl Acad Sci U S A 2020 Mar 31;117(13):6990-6997 [FREE Full text] [doi: 10.1073/pnas.1918891117] [Medline: 32229559]
15. Diversity and STEM: women, minorities, and persons with disabilities. National Center for Science and Engineering Statistics (NCSES). 2023. URL: https://ncses.nsf.gov/pubs/nsf23315/ [accessed 2024-01-19]
16. Active physicians by sex and specialty, 2021. Association of American Medical Colleges. URL: https://www.aamc.org/data-reports/workforce/data/active-physicians-sex-specialty-2021 [accessed 2023-08-13]
17. Eaton AA, Saunders JF, Jacobson RK, West K. How gender and race stereotypes impact the advancement of scholars in STEM: professors' biased evaluations of physics and biology post-doctoral candidates. Sex Roles 2019 Jun 3;82(3-4):127-141. [doi: 10.1007/s11199-019-01052-w]
18. Koch AJ, D'Mello SD, Sackett PR. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. J Appl Psychol 2015 Jan;100(1):128-161. [doi: 10.1037/a0036734] [Medline: 24865576]
19. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci U S A 2012 Oct 09;109(41):16474-16479 [FREE Full text] [doi: 10.1073/pnas.1211286109] [Medline: 22988126]
20. Hire and retain a diverse team. Textio. URL: https://textio.com/ [accessed 2023-08-13]
21. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020 Presented at: NIPS '20; December 6-12, 2020; Vancouver, BC p. 1877-1901 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883
22. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. University of Texas at Austin. 2022. URL: https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf [accessed 2024-01-19]
23. Kaplan DM, Palitsky R. Investigating evidence of gender bias in letters of recommendation generated by ChatGPT. Open Science Framework. 2023 Jun 12. URL: https://osf.io/g6b8y/ [accessed 2024-01-22]
24. Girgis MY, Qazi S, Patel A, Yu D, Lu X, Sewards J. Gender and racial bias in letters of recommendation for orthopedic surgery residency positions. J Surg Educ 2023 Jan;80(1):127-134. [doi: 10.1016/j.jsurg.2022.08.021] [Medline: 36151044]
25. Schmader T, Whitehead J, Wysocki VH. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex Roles 2007 Aug 17;57(7-8):509-514 [FREE Full text] [doi: 10.1007/s11199-007-9291-4] [Medline: 18953419]
26. Jordan KN, Sterling J, Pennebaker JW, Boyd RL. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. Proc Natl Acad Sci U S A 2019 Feb 26;116(9):3476-3481 [FREE Full text] [doi: 10.1073/pnas.1811987116] [Medline: 30808741]
27. Avoiding gender bias in reference writing. University of Arizona Commission on the Status of Women. URL: https://csw.arizona.edu/sites/default/files/avoiding_gender_bias_in_letter_of_reference_writing.pdf [accessed 2024-01-19]
28. Kacewicz E, Pennebaker JW, Davis M, Jeon M, Graesser AC. Pronoun use reflects standings in social hierarchies. J Lang Soc Psychol 2013 Sep 19;33(2):125-143. [doi: 10.1177/0261927x13502654]

XSL•FO
RenderX

29. Top names over the last 100 years. Social Security Administration. URL: https://www.ssa.gov/oact/babynames/decades/century.html [accessed 2024-01-19]
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 2018 Dec 05;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
31. Connor RA, Glick P, Fiske ST. Ambivalent sexism in the twenty-first century. In: Sibley SG, Barlow FK, editors. The Cambridge Handbook of the Psychology of Prejudice. Cambridge, MA: Cambridge University Press; 2016.
32. Glick P, Fiske ST. The ambivalent sexism inventory: differentiating hostile and benevolent sexism. J Pers Soc Psychol 1996 Mar;70(3):491-512. [doi: 10.1037/0022-3514.70.3.491]
33. Dolci T. Fine-tuning language models to mitigate gender bias in sentence encoders. 2022 Presented at: IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService); August 15-18, 2022; Newark, CA. [doi: 10.1109/bigdataservice55688.2022.00036]
34. Liang PP, Wu C, Morency LP, Salakhutdinov R. Towards understanding and mitigating social biases in language models. arXiv Preprint posted online June 24, 2021.
35. Chaloner K, Maldonado A. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. 2019 Presented at: First Workshop on Gender Bias in Natural Language Processing; August 2, 2019; Florence, Italy. [doi: 10.18653/v1/w19-3804]
36. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, St. John R, et al. Universal sentence encoder. arXiv Preprint posted online March 29, 2018.
37. Brunet ME, Alkalay-Houlihan C, Anderson A, Zemel R. Understanding the origins of bias in word embeddings. arXiv Preprint posted online October 8, 2018 [FREE Full text]
38. Wolfe R, Caliskan A. VAST: the valence-assessing semantics test for contextualizing language models. arXiv Preprint posted online March 14, 2022. [doi: 10.1609/aaai.v36i10.21400]
39. Ethayarajh K, Duvenaud D, Hirst G. Understanding undesirable word embedding associations. arXiv Preprint posted online August 18, 2019. [doi: 10.18653/v1/p19-1166]
40. Akos P, Kretchmar J. Gender and ethnic bias in letters of recommendation: considerations for school counselors. Prof School Counsel 2018 Feb 15;20(1). [doi: 10.5330/1096-2409-20.1.102]
41. Grimm LJ, Redmond RA, Campbell JC, Rosette AS. Gender and racial bias in radiology residency letters of recommendation. J Am Coll Radiol 2020 Jan;17(1 Pt A):64-71. [doi: 10.1016/j.jacr.2019.08.008] [Medline: 31494103]
42. Zwebner Y, Sellier AL, Rosenfeld N, Goldenberg J, Mayo R. We look like our names: the manifestation of name stereotypes in facial appearance. J Pers Soc Psychol 2017 Apr;112(4):527-554. [doi: 10.1037/pspa0000076] [Medline: 28240942]
43. Vayansky I, Kumar SA. A review of topic modeling methods. Inf Syst 2020 Dec;94:101582. [doi: 10.1016/j.is.2020.101582]
44. Marwan Y, Waly F, Algarni N, Addar A, Saran N, Snell L. The role of letters of recommendation in the selection process of surgical residents in canada: a national survey of program directors. J Surg Educ 2017 Jul;74(4):762-767. [doi: 10.1016/j.jsurg.2017.01.006] [Medline: 28126378]
45. Avoiding unintended gender bias in letters of recommendation. National Center for Women & Information Technology. 2010. URL: https://tinyurl.com/2c8eh8v7 [accessed 2024-01-11]

Abbreviations

LIWC: Linguistic Inquiry and Word Count
LLM: large language model
OSF: Open Science Framework
STEM: Science, Technology, Engineering and Mathematics

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 14.08.23; peer-reviewed by A Mihalache, K Wolsiefer; comments to author 02.10.23; revised version received 12.11.23; accepted 27.11.23; published 05.03.24.

Please cite as:
Kaplan DM, Palitsky R, Arconada Alvarez SJ, Pozzo NS, Greenleaf MN, Atkinson CA, Lam WA
What's in a Name? Experimental Evidence of Gender Bias in Recommendation Letters Generated by ChatGPT
J Med Internet Res 2024;26:e51837
URL: https://www.jmir.org/2024/1/e51837
doi:10.2196/51837
PMID:38441945

©Deanna M Kaplan, Roman Palitsky, Santiago J Arconada Alvarez, Nicole S Pozzo, Morgan N Greenleaf, Ciara A Atkinson, Wilbur A Lam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.03.2024. This is an

XSL•FO
RenderX

open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Parents' Perceptions of Their Parenting Journeys and a Mobile App Intervention (Parentbot—A Digital Healthcare Assistant): Qualitative Process Evaluation

Joelle Yan Xin Chua[1], BSc; Mahesh Choolani[2], PhD; Cornelia Yin Ing Chee[3], MBBS; Huso Yi[4], PhD; Yiong Huak Chan[5], PhD; Joan Gabrielle Lalor[6], PhD; Yap Seng Chong[2], PhD; Shefaly Shorey[1], PhD

[1]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2]Department of Obstetrics and Gynaecology, National University Hospital, Singapore, Singapore
[3]Department of Psychological Medicine, National University Hospital, Singapore, Singapore
[4]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore
[5]Biostatistics Unit, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[6]School of Nursing and Midwifery, Trinity College Dublin, Dublin, Ireland

Corresponding Author:
Shefaly Shorey, PhD
Alice Lee Centre for Nursing Studies
Yong Loo Lin School of Medicine
National University of Singapore
Level 2, Clinical Research Centre
Block MD11 10 Medical Drive
Singapore, 117597
Singapore
Phone: 65 66011294
Fax: 65 6776 7135
Email: nurssh@nus.edu.sg

Abstract

Background: Parents experience many challenges during the perinatal period. Mobile app–based interventions and chatbots show promise in delivering health care support for parents during the perinatal period.

Objective: This descriptive qualitative process evaluation study aims to explore the perinatal experiences of parents in Singapore, as well as examine the user experiences of the mobile app–based intervention with an in-built chatbot titled Parentbot—a Digital Healthcare Assistant (PDA).

Methods: A total of 20 heterosexual English-speaking parents were recruited via purposive sampling from a single tertiary hospital in Singapore. The parents (control group: 10/20, 50%; intervention group: 10/20, 50%) were also part of an ongoing randomized trial between November 2022 and August 2023 that aimed to evaluate the effectiveness of the PDA in improving parenting outcomes. Semistructured one-to-one interviews were conducted via Zoom from February to June 2023. All interviews were conducted in English, audio recorded, and transcribed verbatim. Data analysis was guided by the thematic analysis framework. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used to guide the reporting of data.

Results: Three themes with 10 subthemes describing parents' perceptions of their parenting journeys and their experiences with the PDA were identified. The main themes were (1) new babies, new troubles, and new wonders; (2) support system for the parents; and (3) reshaping perinatal support for future parents.

Conclusions: Overall, the PDA provided parents with informational, socioemotional, and psychological support and could be used to supplement the perinatal care provided for future parents. To optimize users' experience with the PDA, the intervention could be equipped with a more sophisticated chatbot, equipped with more gamification features, and programmed to deliver personalized care to parents. Researchers and health care providers could also strive to promote more peer-to-peer interactions among users. The provision of continuous, holistic, and family-centered care by health care professionals could also be emphasized. Moreover, policy changes regarding maternity and paternity leaves, availability of infant care centers, and flexible work arrangements could be further explored to promote healthy work-family balance for parents.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56894) doi:10.2196/56894

KEYWORDS

perinatal; parents; mobile app; chatbot; qualitative study; interviews; experiences; mobile phone

Introduction

Background

Many parents feel stressed during the perinatal period (duration from conception to 1 year postpartum), putting them at risk for developing psychological problems such as perinatal anxiety and depression [1,2]. Studies have reported that approximately between 26% and 15% of mothers [3,4] and 8% to 14% and 11% of fathers [5,6] experience perinatal depression and anxiety, respectively. Moreover, a previous study conducted in Taiwan reported an increasing trend of stress, anxiety, and depression levels among women from around 24 gestational weeks to 1 month post partum, indicating that parents would benefit from more support during this crucial transition to the parenting period [7]. Moreover, the signs of postpartum depression and anxiety tend to be evident within 1 to 3 months after childbirth [4,8-10]. Poor mental health among parents can cause adverse effects on their child's physical, social, emotional, cognitive, and psychological development [11-14]. These effects are not only evident during infancy but also in their childhood and adolescent years [11,15,16]. In Singapore, approximately 40% of mothers have reported experiencing perinatal depressive symptoms [17], and fathers have also expressed feelings of distress [18]. In addition, parents in Singapore have found it especially difficult and stressful to adjust to life with their new babies within the first month post partum [19]. Moreover, a recent study reported that Singaporean fathers are negatively affected by their partners' poor mental health during the first 3 months post partum [20]. Given that parents in Singapore are at risk for mental health problems, there is an increased need to understand parents' experiences with perinatal care and provide them with needed support.

As mobile health apps can provide health care support swiftly and opportunely to many individuals concurrently [21-23], the World Health Organization has promoted their development and use to encourage self-care [24]. Since the functions of mobile apps can be tailored to meet users' specific needs, researchers have developed mobile apps to help individuals with various medical conditions such as cancer, depression, anxiety, and hypertension [25-27]. A recent review also found that specially developed perinatal mobile apps were promising in providing informational, social, and mental health support to parents during the perinatal period [28]. Considering that most of Singapore's population are smartphone users [29], conducting a mobile app–based intervention for parents in Singapore is viable.

Previous mobile app–based interventions provided parents with multimedia educational resources and discussion forums to connect with peers and health care professionals [30-32]. Parents appreciated the convenient and trustworthy information about perinatal care provided by these mobile app–based interventions [19,33,34]. In addition, they valued receiving helpful parenting tips from peers and having their queries answered by local health care professionals [19,33,34]. However, parents wanted faster answers to their queries [33,34], and some suggested using chatbots in future mobile apps [19].

Chatbots are digital programs that communicate with humans using voice, text, or animation and can provide instant support to individuals in real time [35]. They have been reported to be capable of providing parents with valuable educational, socioemotional, and mental health support during the perinatal period [36]. The anonymous and nonjudgmental nature of chatbots could also help to connect with individuals who may be hesitant to interact with their peers or health care professionals, even via a digital anonymous platform [37]. Therefore, a mobile app–based intervention with an in-built chatbot titled Parentbot–a Digital Healthcare Assistant (PDA) was developed and implemented to provide more efficient perinatal support to parents [38]. The PDA is a multicomponent mobile app–based intervention that provides parents with informational, psychological, and social support during pregnancy (>24 weeks of gestation) till the first month post partum. The quantitative evaluation of the PDA's effectiveness on parenting outcomes including parenting self-efficacy, stress, anxiety, depression, and social support has been conducted and reported under the trial outcomes (article under review), and this study aimed to present the in-depth exploration of parental views about the intervention via process evaluation. A process evaluation would allow the thorough examination of the various aspects of intervention delivery and receipt as well as the mechanisms through which the intervention can bring about change and identify the barriers and facilitators that affected intervention delivery and experience [39]. A qualitative approach was selected to better collect descriptive insights about the parents' experiences and allow subsequent intervention methods to be tailored to the parents' specific needs [40,41].

Aims

This study aimed to explore the perinatal experiences of parents and examine the user experiences of the PDA over the same period (from >24 gestational weeks, age of viability in Singapore until 1 month post partum). The perinatal experiences of both the control and intervention groups were included to compare and differentiate their experiences to gain holistic insights related to the PDA. The research questions are as follows: (1) What are the perceptions of parents who received standard perinatal care in Singapore? (2) What are the experiences of parents who received the PDA in addition to standard perinatal care in Singapore?

Methods

Research Design

A descriptive qualitative study design was used to analyze and interpret the data that closely captured the participants' raw experiences [42]. This study's findings were reported according to the COREQ (Consolidated Criteria for Reporting Qualitative

XSL•FO
RenderX

Research) checklist [43]. This study took place in a tertiary hospital in Singapore from November 2022 to August 2023. The study was part of a randomized controlled trial (RCT) that investigated the effectiveness of the PDA on 7 perinatal parenting outcomes: parenting self-efficacy, stress, anxiety, depression, social support, parent-child bonding, and parenting satisfaction (article under review). The PDA is a multicomponent mobile app–based intervention that provides both first-time and experienced parents (mothers and fathers) with educational and psychosocial support from >24 gestational weeks (age of viability in Singapore) until 1 month post partum. It includes multimedia educational resources, a rule-based chatbot, a discussion forum, expert advice, and gamification features such as digital badges. Most of the content in the PDA is relevant for both first-time and experienced parents. However, some resources specifically catered to experienced parents (eg, managing sibling rivalry and engaging older children to care for the newborn sibling). More details of the PDA can be found in the published development paper by Chua et al [38].

Participants

A total of 118 heterosexual couples were recruited from a tertiary public hospital offering maternity care to the central region of Singapore in the main RCT. The hospital has >1200 beds with >50 clinical specialties. After obtaining written informed consent, couples were randomly assigned to the intervention or control group. Couples in the control group received standard perinatal care provided by the hospital in the form of hospital visits and appointments with their respective obstetricians. By contrast, couples in the intervention group received access to the PDA from the point of study recruitment (>24 weeks of gestation) until 1 month post partum, in addition to the standard perinatal care provided by the hospital. Data for the main RCT were collected at baseline (at the time of recruitment and consent taking) and at the first- and third-month postpartum period. Immediately after the PDA intervention at 1 month post partum, fathers or mothers from different couples were invited from both the intervention and control groups to explore their experiences with the PDA and standard perinatal care received. Parents were interviewed one-on-one without their partners to elicit more honest responses from them [44], and fathers and mothers from different couples were interviewed to obtain more variation in their perinatal experiences.

The eligibility criteria for parents to be included in the qualitative study were if they (1) were aged ≥21 years, (2) were able to read and speak English, (3) gave birth to a newborn without serious medical complications (eg, congenital anomalies or genetic disorders), (4) stayed in Singapore for the 3 months after childbirth, and had varied mean parenting self-efficacy scores as measured by the 10-item Parenting Efficacy Scale (PES) at baseline (>24 gestational weeks) [45]. Items of the PES were rated on a 4-point Likert scale ranging from 1 to 4, with higher total scores indicating a higher level of perceived parenting self-efficacy [45]. An example of an item is “How good do you feel you are at feeding your baby?” [45]. A high internal consistency with a Cronbach α of 0.91 was reported for the PES at the baseline time point among all parents recruited in the main RCT. A purposive sample of 20 parents with varied PES scores (mothers: n=10, 50%; fathers: n=10, 50%) participated in this study. Data saturation was achieved at the 18th participant when no new findings emerged [46], and 2 additional interviews were conducted to confirm the findings.

Data Collection

Parents’ sociodemographic characteristics and childbirth details were collected using web-based surveys. A semistructured interview guide was developed based on literature [19,33] and the multidisciplinary study team’s expertise (eg, nurses, obstetricians, and midwives). To ensure the comparability of data, parents from the intervention and control groups were asked the same questions regarding their experience with standard perinatal care. In addition, parents from the intervention group were asked about their personal experience with using the PDA, while parents from the control group were asked about their views regarding telehealth services for perinatal care. Open-ended questions, which allowed parents to freely share their views with little restrictions (eg, What were the challenges faced during the perinatal period? What were your experiences with using the PDA?), guided the main structure of the interview [47,48], while follow-up and probing questions were asked to allow researchers to elicit clarity and further insights from parents [49]. The one-on-one interviews were conducted via the web-based Zoom platform (Zoom Video Communications; as per participants’ preferences) at the end of the intervention (ie, after 1 month post partum) and were audio recorded. All interviews were conducted by the primary researcher (JYXC), a female, year-2 PhD nursing student who had been trained in qualitative interviewing techniques by an experienced qualitative nurse researcher and midwife (SS). Field notes of the parents’ nonverbal cues were also documented after each interview. The interviewer (JYXC), with a maternity care background, was trained by the experienced qualitative researcher (SS) to document information regarding parents’ tone of voice, facial expressions, movements, hand gestures, and emotional expressions during the interview. Data collection lasted over a period of 5 months from February to June 2023. The average duration of interviews was approximately 53 (SD 14) minutes for the control group and 49 (SD 12) minutes for the intervention group. All the recordings were transcribed verbatim by research assistants and checked for accuracy by the primary researcher.

Data Analysis

Descriptive statistics (mean and SD for continuous variables and frequency and percentage for categorical outcomes) were used to summarize parents’ sociodemographic characteristics and childbirth details. The 6-step qualitative thematic analysis framework proposed by Braun and Clarke [50] guided the data analysis process: (1) familiarizing oneself with the collected data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report [50]. All transcripts and field notes were read numerous times to gain familiarity with the data set. Next, a manual color-coding method was used to generate the initial set of codes before organizing them into common themes. Codes from the intervention and control groups were compared and tallied to triangulate the data. The initial coding process on all the collected data was conducted by the primary researcher (JYXC) independently, and frequent meetings were held with

XSL•FO
RenderX

the research team to discuss the emerging themes. During team meetings, discrepancies in researchers' judgments were resolved, and the themes were confirmed to ensure that the analysis process was credible.

Rigor

Rigor was maintained throughout the study by ensuring credibility, dependability, transferability, and confirmability [51,52]. To ensure credibility, transcripts were read and reread numerous times to confirm that the derived themes represented parents' experiences [51,52]. The provision of thick, vivid descriptions of parents' experiences using verbatim quotes helped to facilitate the transferability of data [51,52]. Copies of all transcripts, authors' reflections, and any supporting documents were kept to maintain an audit trail and ensure the confirmability and dependability of the study [51,52]. All authors kept their critical reflections in individual web-based journals throughout the analysis process to enhance and ensure the reflexivity of this study's findings [53,54].

Ethical Considerations

Ethics approval for the main RCT with this nested study was obtained from the Domain Specific Review Board (reference number 2021/00227). Voluntary participation among participants was reinforced, and written informed consent was obtained before the study recruitment. Interviews were coded to maintain participants' confidentiality. This unique code number was attached to the respective participant's transcript. All consent forms were kept in a locked cupboard in the principal investigator's office, while all audio recordings, survey responses, and transcripts were stored in a password-protected stand-alone computer.

Results

Participants' Characteristics

Married fathers (10/20, 50%) and mothers (10/20, 50%) took part in this study. All parents except 1 (19/20, 95%) were employed. Parents from the control group had a mean age of 33.6 (SD 3.03) years, while parents in the intervention group had a mean age of 34.1 (SD 5.99) years. Of the 20 participants, most were Chinese (n=9, 45%), followed by Indian (n=5, 25%), Malay (n=4, 20%), Vietnamese (n=1, 5%), and White (n=1, 5%). In both groups, half of the parents (n=10, 50%) were first-time parents, while the rest (n=10, 50%) were experienced parents. Their mean parenting self-efficacy baseline scores ranged from 1.1 to 3.5 (total score range 1-4). More details of the parents' characteristics and their PES scores are presented in Tables 1 and 2, respectively. The dropout rates of the main RCT were 16.9% (20/118) and 6.8% (8/118) for the intervention and control groups, respectively, at 1 month post partum and 8.5% (10/118) and 13.6% (16/118) for the intervention and control groups, respectively, at 3 months post partum.

Parents' perceptions of the perinatal care received and their experiences with the PDA were captured in 3 main themes and 10 subthemes (Figure 1). The main themes were (1) new babies, new troubles, and new wonders; (2) support system for the parents; and (3) reshaping perinatal support for future parents.

XSL•FO
RenderX

Table 1. Participant characteristics (n=20).

Baseline characteristics	Intervention group (n=10)	Control group (n=10)
Age (y), mean (SD)	34.1 (5.99)	33.6 (3.03)
Ethnicity, n (%)		
Chinese	5 (50)	4 (40)
Indian	3 (30)	2 (20)
Malay	1 (10)	3 (30)
Vietnamese	0 (0)	1 (10)
White	1 (10)	0 (0)
Marital status, n (%)		
Married	10 (100)	10 (100)
Length of marriage (y), mean (SD)	6.0 (3.97)	4.9 (1.91)
Education level, n (%)		
Secondary school	1 (10)	1 (10)
ITE[a], polytechnic, or junior college	1 (10)	0 (0)
University degree	6 (30)	7 (70)
Postgraduate degree	2 (20)	2 (20)
Employment status, n (%)		
Employed	9 (90)	10 (100)
Unemployed	1 (10)	0 (0)
Monthly household income (S GD), n (%)		
<3000	1 (10)	1 (10)
3000-7000	5 (50)	0 (0)
>7000	4 (40)	9 (90)
Baby's gender, n (%)		
Male	6 (60)	5 (50)
Female	4 (40)	5 (50)
Childbirth method, n (%)		
Normal vaginal delivery	5 (50)	7 (70)
LSCS[b] (elective)	3 (30)	1 (10)
LSCS (emergency)	1 (10)	2 (20)
Assisted vaginal delivery using vacuum extraction	1 (10)	0 (0)
Attended antenatal preparation courses, n (%)	3 (30)	5 (50)
Did the mother follow the confinement period? n (%)		
Yes	4 (40)	5 (50)
No	1 (10)	0 (0)
Method of feeding baby, n (%)		
Breastmilk only (refers to both direct breastfeeding and expressed breastmilk)	5 (50)	4 (40)
A mixture of breastmilk and formula feeds	5 (50)	6 (60)
Length of maternity or paternity leave, n (%)		
<1 month	4 (40)	5 (50)
1-4 months	4 (40)	5 (50)
Did not take leave	1 (10)	0 (0)

XSL•FO
RenderX

Baseline characteristics	Intervention group (n=10)	Control group (n=10)
Not applicable (not currently working)	1 (10)	0 (0)
Total number of children (including new baby), n (%)		
1	5 (50)	5 (50)
≥2	5 (50)	5 (50)

[a]ITE: Institute of Technical Education.

[b]LSCS: lower segment cesarean section.

Table 2. Participants' parenting self-efficacy scores at baseline.

Study participant code	Total parenting self-efficacy score (range 10-40)	Mean parenting self-efficacy score (range 1-4)
Control group		
Mother control 1	28	2.8
Mother control 2	20	2
Mother control 3	30	3
Mother control 4	28	2.8
Mother control 5	29	2.9
Father control 1	32	3.2
Father control 2	20	2
Father control 3	23	2.3
Father control 4	11	1.1
Father control 5	35	3.5
Intervention group		
Mother intervention 1	27	2.7
Mother intervention 2	30	3
Mother intervention 3	29	2.9
Mother intervention 4	31	3.1
Mother intervention 5	35	3.5
Father intervention 1	35	3.5
Father intervention 2	26	2.6
Father intervention 3	31	3.1
Father intervention 4	20	2
Father intervention 5	27	2.7

XSL•FO
RenderX

Figure 1. Summary of themes and subthemes. PDA: Parentbot–a Digital Healthcare Assistant.

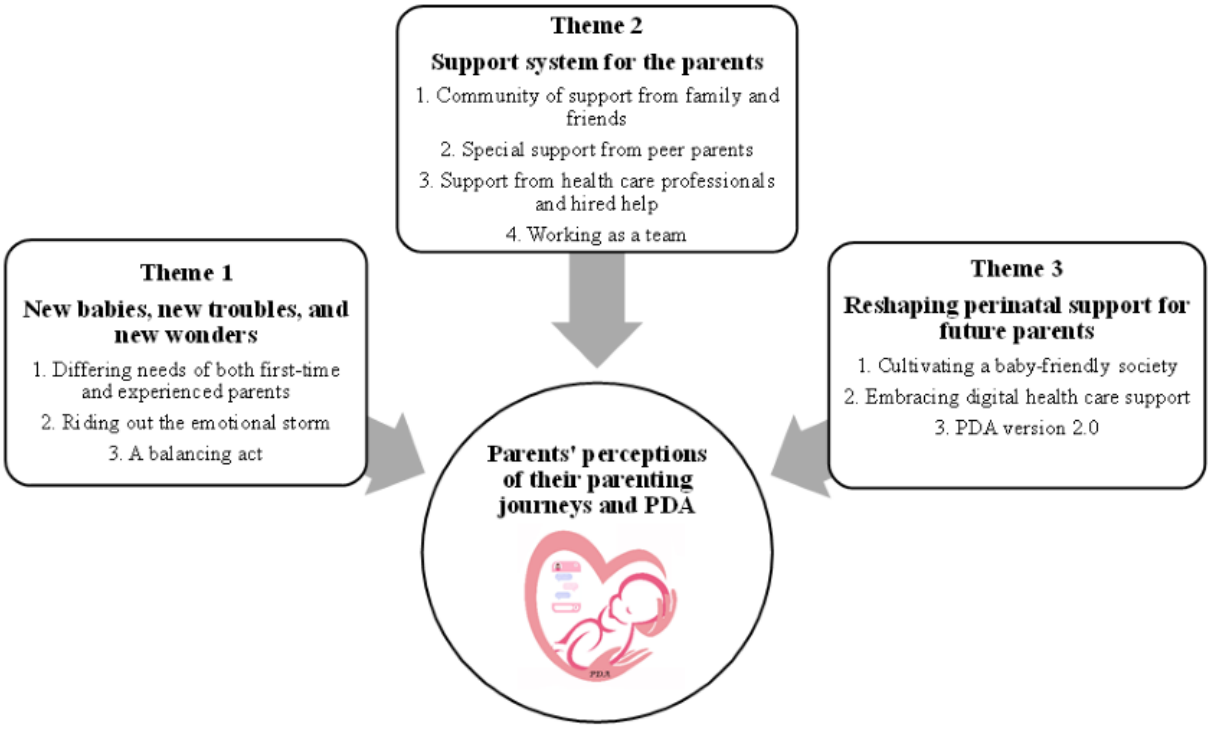

New Babies, New Troubles, and New Wonders

Parents' experiences and challenges faced while caring for their new baby were categorized into 3 subthemes: differing needs of both first-time and experienced parents, riding out the emotional storm, and a balancing act.

Differing Needs of Both First-Time and Experienced Parents

First-time parents from both the control and intervention groups experienced a "steep learning curve" during the first few weeks post partum as they learned the skills needed to care for their newborn (eg, breastfeeding and bathing). The parents from the control group felt confused by the conflicting advice received from websites on the internet, family and friends, and health care providers. They also often used "trial-and-error" means to identify the best methods to care for their child. Parents from the intervention group used the information provided by the PDA to improve their baby care skills and resolve the conflicting advice they received from other sources:

> *You don't know what she wants whenever she cries...really you just have to figure it out each time...just trial and error.* [Mother control 2]

> *[T]he information provided in the app [PDA]...maybe it gives me 80% confidence to take care of my baby.* [Father intervention 2]

Although some experienced parents needed to refresh their memory of baby care skills, these parents had more confidence in caring for their newborn. However, they had to learn how to care for a new baby with unique needs and manage sibling rivalry. Those in the intervention group used the PDA to help them meet their evolving learning needs:

> *[M]y [first] child didn't experience any like colic symptoms, so for the second one...I think experience colic all the time so I have to search like what to do...* [Mother control 4]

> *It's quite helpful even for second time parents because like we will forget how to let's say, take care of the umbilical cord and there's some on sibling rivalry all that, that we can read on which is helpful for us.* [Mother intervention 3]

Riding Out the Emotional Storm

Parents from both groups experienced many positive and negative emotions within the first few weeks of postpartum. Their moods were closely tied to their abilities to care for the newborn; they felt stressed, helpless, and frustrated when they were unable to meet their newborn's needs and felt joyous after learning how to care for their newborn sufficiently:

> *Like when the baby won't stop crying, sometimes I get frustrated.... [But when] the problem is solved then it makes you feel better, gives you a boost that hey that you are capable, you are doing things right...* [Father control 4]

Parents from both groups managed their emotions by engaging in personal leisure activities (eg, watching Netflix, cooking, or cycling); planning their daily schedule carefully; and seeking professional help when needed. Those from the intervention group used the PDA's activities (journal exercises, mindfulness exercises, and making and sharing of positive posters) to engage in self-reflection and regulate their thoughts and emotions:

> *Because people always post something positive and...you read, you think about it, ah, like that's what exactly what you're feeling now, and you feel happy*

XSL•FO
RenderX

about it. Yeah. This is very positive. I like this [the poster board]. [Father intervention 5]

A Balancing Act

Parents from both groups highlighted that they had to balance their work demands with baby care responsibilities. They worried about baby care arrangements after mothers' maternity leave ended and applying for leave if their babies were to fall sick. Many parents wanted to forge a strong bond with their babies and were concerned that returning to work could cause them to miss their babies' developmental milestones:

> *Maybe I'll miss out on her milestones and all that...maybe at the end when she grows older, then I'll regret not spending more time with her.* [Father control 2]

All parents also balanced their personal life and baby care by taking turns or engaging their family members (eg, mothers-in-law or mothers) to care for their baby:

> *Sometimes when my mother-in-law is free, she will help to take care of the baby. Me and my husband will go out, maybe just have a meal together.* [Mother control 3]

In addition, experienced parents from both groups had to balance the time spent with each of their children so that no child felt "neglected." Experienced fathers often adopted the role of being the "main caregiver" for their older children, while experienced mothers mainly cared for the newborn due to their breastfeeding duties:

> *I had to take care more of the other kid...because she's the one breastfeeding the other one...* [Father control 5]

Support System for the Parents

The varied sources of support that parents had were organized into 4 subthemes: community of support from family and friends, special support from peer parents, support from health care professionals and hired help, and working as a team.

Community of Support From Family and Friends

Parents from both groups received valuable support from their family members (eg, parents, parents-in-law, and siblings) in terms of teaching them baby care skills, helping mothers recover after labor, and helping to care for their newborns. However, some parents from both groups had limited family support as their family members were busy with other responsibilities or unfit to care for newborns due to poor health. Moreover, some parents from both groups felt stressed when their older family members had different "outdated" methods of caring for babies that clashed with their views:

> *My mommy in-law is stubborn with her ways and plus that she also tends to be very critical in our processes at home...* [Father control 5]

> *[T]hey're [own parents] quite old already, so I don't really want to stress them too much, put too much strain on them.* [Father control 3]

In addition, parents from both groups valued hearing about their friends' first-hand parenting experiences as it helped them to prepare for their baby's birth and care for their baby:

> *We have a few friends who also just given birth...so we get to hear from them, their first-hand experience, what the husband has to do, what the wife has to do, what is the good to have things to be prepared...* [Father control 3]

Special Support From Peer Parents

Many mothers, mostly from the control group (5/5, 100%) and a few from the intervention group (1/5, 20%), sought support and advice from other local parents using web-based forums and WhatsApp groups, while fathers did not adopt this practice. However, parents were aware that these peers could sometimes provide "funny" advice:

> *I felt more comforted because a lot of people will just say that yeah I had my baby like hiccupping almost around the clock and they're totally fine and so many people have that same experience.* [Mother control 5]

The poster board and support forum in the PDA provided parents in the intervention group with a sense of community. The simple interactions of making and sharing digital posters containing positive messages allowed parents to feel "less alone." Although the support forum was underused, parents felt that the forum was a helpful and useful feature to have, especially if there was more activity on it:

> *If more people use [the support forum]...I think this will be a good way to find out from other parents or get some feedback from other parents on certain issues.* [Father intervention 4]

Support From Health Care Professionals and Hired Help

Parents from both groups appreciated the support provided by health care professionals at the hospital and clinics. Health care professionals answered parents' queries thoroughly, taught them how to care for their babies patiently, and provided them with emotional support during their labor and inpatient stay. Some parents (mostly from the control group, 5/10, 50% and a few from the intervention group, 2/10, 20%) wanted more in-depth explanations from the physicians at the polyclinics to allay their anxieties:

> *The nurses were very nice...it was my first baby, so they actually taught me how to breastfeed, how to latch...When my son was crying, then I get overwhelmed...the nurse actually came by to help...* [Mother intervention 1]

> *Whatever questions we asked, the doctor [from polyclinic] would just say oh its normal normal normal, so we felt that it wasn't really assuring.* [Mother control 2]

Confinement nannies helped parents from both groups by cooking nutritious meals for mothers to facilitate their postchildbirth recovery, taking care of newborns, and teaching

XSL•FO
RenderX

parents how to care for their new babies. After the nannies left, some parents were sufficiently confident to care for their babies while, others (mostly from the control group and a few from the intervention group) were lost and stressed:

> *For the first month I hired a confinement nanny...Help to take care of the baby, feed the baby, uh, clothe the baby...She will slowly teach us la gradually.* [Mother intervention 2]

Domestic helpers mainly helped parents from both groups with household chores and meal preparations. While some parents from both groups also trained their helpers to care for their babies, the cultural differences between Singapore and the helpers' own country made it challenging for helpers to assist with infant care:

> *We're worried because they are from a different country...their ways of taking care and developing a baby might be more basic, more rudimentary.* [Father control 4]

Working as a Team

To support mothers, fathers from both groups spent more time at home by leaving work on time and seeking additional leave days. Fathers from both groups also supported mothers by doing household chores, caring for older children, and purchasing household and baby care necessities (eg, nursing pillows). Conversely, mothers from both groups took on more responsibility for their newborns during maternity leave so that fathers could have sufficient rest and focus on their work. In addition, all parents provided emotional support to each other and reminded one another that they were "both getting to be better parents day by day":

> *[A]t the end of the working day, I just try to drop all my stuff and just run away, you know? Don't care if work is incomplete, will just rush home.* [Father control 2]

Parents from the intervention group mostly used the PDA independently on their respective mobile phones and discussed the content they read to improve baby care:

> *I have missed out maybe example with the breastfeeding part...then [husband] come and share to me eh I have read it in Parentbot [PDA] they have stated like this, so you can try in this position, you can do it in this way.* [Mother intervention 4]

Reshaping Perinatal Support for Future Parents

The different suggestions to improve perinatal care for parents were sorted into 3 subthemes: cultivating a baby-friendly society, embracing digital health care support, and PDA version 2.0.

Cultivating a Baby-Friendly Society

Parents from both groups would like to receive more financial assistance in the form of subsidies for baby supplies (eg, diapers, milk bottles, and milk powder) and infant care fees. To help parents balance their work demands and baby care, flexible work arrangements, longer and equal maternity and paternity leave (eg, 6 months), and more infant care slots were highlighted:

> *I think like a six month phase [leave] may even encourage more people to breastfeed at least up to that six month mark.* [Mother control 5]

> *That also calls for some changes where mother and father both are given equal leave. Then, the whole family gets involved in baby care because we get to leave, we are more in touch with the newborn.* [Mother intervention 5]

Parents from both groups also wanted to receive continuity of care during the perinatal period by having the same clinician during pregnancy, the same maternity team in the postnatal ward, and the same clinician caring for their baby at affordable subsidized rates. They also wanted babysitting and confinement nanny services to be subsidized and well regulated. Moreover, parents from both groups wanted to receive more holistic family-centered care by having their informational and emotional needs met and having their opinions respected by health care professionals:

> *[A]fter giving birth, if they could attach us to a family doctor...from the beginning, they will understand the history of the children more.... If we have continuity of care from the same doctor, it will be better.* [Mother control 4]

> *I had to insist to bring in the midwife earlier to check how dilated I am.... I had to ask three times before the nurse brought in the midwife.* [Mother control 3]

Embracing Digital Health Care Support

All parents in this study found it more convenient to search online for perinatal care information or use the PDA, as they could do their learning while commuting. Parents from the control group had difficulties searching for online validated information that applies to Singapore's context, whereas those from the intervention group appreciated the contextualized trustworthy information provided by the PDA:

> *I found myself using this kind of on the go, like when I was sitting on the MRT [public train system in Singapore], I just scroll through and look for things.* [Father intervention 3]

> *Sometimes when I try to find things, it will show other country stuff.... I try to find just related to Singapore stuff and it's scattered everywhere.* [Mother control 1]

Parents from the control group were interested in using telemedicine services and were open to the idea of communicating with chatbots as long as the computer-generated answers were prevetted by health care professionals. Parents from the intervention group who tried the chatbot in the PDA also had a pleasant experience interacting with it:

> *I think it's quite interactive...there are one or two times I actually type in something, then the bot will actually reply to the questions.... I believe this one is already done with a lot of machine learning.... So*

XSL•FO
RenderX

> *answers are actually by the professional.* [Father intervention 1]

All the parents were interested in the PDA; when informed about the intervention, those from the control group wanted to try it, and all the parents from the intervention group would recommend the PDA to others. Parents from the intervention group indicated that the PDA was helpful for both fathers and mothers and both first-time and experienced parents. Those who received the PDA also wanted to have access to the PDA for a longer duration (eg, 6 months post partum) and wanted the PDA to provide more information about raising children throughout infancy and toddlerhood:

> *I think everybody should have it [PDA]...many of my friends who are becoming first-time mothers keep asking me, oh, what happens to this?... then I impart the same information, which I get from here [PDA]. [Mother intervention 5*

PDA Version 2.0

Parents from the intervention group provided various suggestions to improve the PDA. They suggested using more "eye-catching" elements (eg, animations and brighter colors) to make the application more stimulating. Better organization and a more intuitive interface could help parents locate and use the PDA's features more easily. Parents also found it "fun" to complete their assigned weekly activities in the PDA and see the corresponding colors of a digital rainbow badge light up (eg, green band will be filled in after completing a gratitude journal entry). However, they suggested having flexibility in the weekly activities and allowing users to customize their activities according to personal preferences to sustain their motivation to use the PDA:

> *So you can still have a rainbow badge, but maybe the violet is for if you complete one activity in the app, maybe the blue is...you can attach points to specific actions and then have them all add up until you get the full rainbow [badge].* [Father intervention 3]

To improve parents' ease of access to information, the chatbot and search function in the PDA could be improved to locate more accurate answers to parents' queries. The educational videos could be categorized into different sections for parents to easily access their segments of interest. Parents also requested to have more information on their emotional and mental health and possible pregnancy and infant complications:

> *For general questions, the chatbot is okay, but for more specific questions [e.g. How much weight should my baby gain in the first month?] it doesn't answer...it keeps giving you the same answer.* [Mother intervention 5]

Parents also wanted to receive immediate replies (within 15 minutes) to their queries from health care experts (the standard reply was within 24 hours), feedback on their journal entries, and more encouragement on the poster board:

> *If possible, maybe you can have a live agent or somebody that we can talk to during working hours.* [Mother intervention 3]

Other features that parents would like to have in the PDA include baby care skills simulation games (eg, diaper changing and baby bathing) possibly using virtual reality technology, games to help parents with emotional regulation, and short quizzes on perinatal care knowledge with answers provided. Parents also wanted the PDA to track their pregnancy and children's growth milestones and provide them with appropriate care tips at the respective time points:

> *[M]aybe the baby is crying and you have to figure out why it's upset...So to have a game like that, it would help to put a lot of peoples' fears at peace I think...* [Father intervention 3]

Discussion

Principal Findings

This study conducted a process evaluation of the PDA trial to explore the perinatal experiences and the user experiences of the PDA offered to parents in Singapore. Parents from different ethnic groups were interviewed, with Chinese, Malay, and Indian parents constituting most of the sample, which corresponds to the ethnic profile of Singapore [55]. There was also an equal mix of first-time and experienced parents from both the control and intervention groups who had a varied range of parenting self-efficacy scores at baseline.

All parents experienced positive and negative emotions during the first few weeks post partum, which is similar to previous research on local parents [19]. Although parents from both groups found ways to manage their negative emotions, parents in the intervention group had more options such as engaging in mindfulness-based and positive psychology (gratitude journal and poster making) activities in the PDA. Since these activities helped some parents to improve their psychological well-being as reported in previous research conducted in the United States and Europe [56,57], future perinatal interventions could include psychotherapeutic components to address parents' psychological well-being [28]. However, as the PDA activities did not appeal to some parents, alternative activities such as games on emotional regulation were suggested by parents and could be explored by future studies, given that games could improve mental health outcomes among adults [58,59].

All parents in this study used web-based methods of learning perinatal care information (eg, websites, forums, and message groups) due to its convenience and accessibility. This ties in with the current generation of Singaporeans being more educated and encouraged to use digital technology in their daily lives [60,61]. Given that cultural norms affect perinatal care practices [62], parents benefited from receiving information that was relevant in the local context, a sentiment echoed by a previous local study [19]. Suggestions to improve parents' PDA learning experience include quizzes, baby care simulation games (using virtual reality or digital animations), and personalized care tips being provided by push notifications at significant milestones during the perinatal period. Considering the prior success in using games to teach baby care skills (eg, breastfeeding) [63], future studies could develop more games to deliver perinatal education, especially in countries with good technological infrastructure such as Singapore [64]. Moreover, the tailoring

XSL•FO
RenderX

of mobile app–based interventions' content using individualized push notifications could improve participants' engagement and the effectiveness of the intervention [65].

Parents from both groups expressed interest in communicating with chatbots, which ties in with the global trend of young adults being technology savvy and satisfied with interacting with artificial intelligence (AI) for their health needs [66,67]. Therefore, chatbots could supplement the current perinatal care by providing more in-depth information about parents' health conditions to meet their needs [36], thereby addressing the manpower shortage faced by health care providers in Singapore and worldwide [68,69]. However, parents from the intervention group pointed out that the chatbot in the PDA could not sufficiently address all their queries, due to its lack of AI. This was a common problem reported by parents in a recent review, and chatbots were recommended to be programmed with natural language processing and AI to improve their communication skills [36]. Moreover, as parents in this study and previous research valued chatbots that were trustworthy and applicable to one's culture [36], future chatbots could be preapproved by local health care professionals.

Parents from both groups received childcare support from older family members, especially their mothers and mothers-in-law, which is a common practice in Singapore and Asia [19,70]. While some parents valued the support provided by their family, others experienced distress due to their family's differing views on childcare and hence preferred to seek advice from their friends who have had children recently and could provide up-to-date information. A similar finding was reported in a previous local study on parents [19]. Moreover, grandparents could also experience caregiver strain, especially if they have poor health conditions [71,72]. Some parents in this study were aware of this problem and decided to enroll their babies in infant care centers when they returned to work. Considering the current trend of couples delaying childbirth in Singapore [73], more grandparents might be older and have poorer health. Hence, more affordable and conveniently located infant care centers might be needed to meet the childcare needs of future local parents.

In addition to speaking with their friends, parents from both groups (especially mothers) sought peer support from web-based sources (eg, forums and messaging groups) to improve their socioemotional well-being. Those from the intervention group also benefited from their additional access to the PDA. Similar benefits of interacting with other parents have been noted in previous local studies [19,74]. According to the concept analysis of peer support by Dennis [75], peer support could effectively improve parents' perinatal well-being by encouraging one's help-seeking behaviors, allowing one to learn realistic coping methods, and inspiring one to adopt constructive and proper coping strategies undertaken by peers [76]. In the PDA, parents interacted with one another via a discussion forum and shared digital posters. While the poster board was generally enjoyed by the parents, the discussion forum was underused possibly due to the lack of moderation on the platform by researchers. To increase the forum's activity, weekly topics could be introduced, and notifications or regular messages or posts could be initiated by researchers to encourage participants' engagement [77].

While parents from both groups were generally satisfied with the current health care services, they would like to receive more continuity of care and family-centered care, which could lead to improved maternal and neonatal health outcomes [78,79]. In Singapore's current public health system, parents do not receive care from the same team of health care professionals throughout the perinatal period. Hence, the provision of continuity of care could be enhanced by keeping detailed shared electronic records of mothers and babies and using collaborative care models [80]. Additional training of maternity health care professionals could be conducted to deliver respectful and holistic family-centered care [79]. Moreover, the implementation of midwifery-led care models to provide continuous high-quality maternal care throughout the perinatal period could be explored and promoted in Singapore and on a global scale [81-84].

Current findings highlighted the importance of parents' support for each other. The PDA promoted partner support by encouraging parents to discuss the information they read and learn tips on supporting one another, which was similar to the findings from previous local studies [19,33]. As receiving support from one's partner can improve one's psychological well-being and parenting behaviors [85,86], future perinatal interventions delivered to both parents should continue to focus on this aspect to optimize their benefits. Moreover, many fathers in this study showed interest in being involved with childcare. The shift in fathers' mindset from the traditional breadwinner role to taking on more active roles in childcare was also reported in other parts of Asia (eg, Thailand, Japan, and Taiwan) [87,88] possibly due to couples adopting more egalitarian relationships [89].

With fathers wanting to be more involved in childcare duties, mothers and fathers from both groups in this study were concerned about balancing their work demands and baby care responsibilities. This is a common worry among parents with young children [90], and employers in Singapore have been looking into establishing flexible work arrangements especially after the COVID-19 pandemic to help parents achieve a healthy work-family balance [91,92]. Parents also requested longer maternity and paternity leaves for up to 6 months to encourage exclusive breastfeeding for the first 6 months of an infant's life [93], especially since only 40% of mothers in Singapore exclusively breastfeed their children for the first 6 months post partum [94]. An equal duration of paternity and maternity leaves was also suggested, which align with the current generation of parents wanting to assume more equal responsibility for their children's well-being [89]. As paternal involvement significantly contributes to better maternal and child health outcomes [95], Singapore has recently doubled paternity leave from 2 to 4 weeks to encourage more father-infant bonding [96]. Plans to further lengthen paternity and maternity leaves and its influence on parenting and newborn outcomes could be explored.

Strengths and Limitations

This study provided valuable insight into the parenting experiences of first-time and experienced parents with perinatal care in the current digital age and their user experiences of the

XSL•FO
RenderX

newly developed PDA intervention. It explored how mobile app–based support and chatbots could enhance perinatal care. Current findings are important in guiding future research on the provision of technology-based perinatal support for parents in Asia, especially using mobile apps and chatbots.

However, this study has some limitations. As only married English-speaking heterosexual parents who had a newborn without serious medical complications were recruited from a single tertiary hospital, the findings of this study may not represent the views of all parents in Singapore. Although data saturation was achieved, the relatively small sample size for this study could have hindered the transferability of its findings to all parents in Singapore. Since this study was conducted on multiracial parents in Singapore, its results might not be transferable to parents in other contexts. Moreover, with all study participants identifying as heterosexual, the findings of this study might not be representative of parents from the lesbian, gay, bisexual, transgender, and queer (LGBTQ) community. While some gender-specific and parity-specific differences in the perinatal experiences of parents were reported in this study, more in-depth gender and parity differences could not be explored due to the nature of this study with small sample size. In addition, the use of only personal interviews in this study could introduce self-reported bias and impair the confirmability of the findings. As this study only recruited a convenient sample of parents who had smartphones and were willing to participate, other parents who did not own smartphones (rarely seen in the Singapore context) or who had actual needs but were not keen to participate in research studies could be underrepresented in this study. Furthermore, since all study participants were only interviewed once between 1 and 3 months after childbirth, this study's findings do not reflect parents' experiences across the entire perinatal period. Finally, due to logistical limitations (limited grant funding and resources), the research team could not feasibly develop the PDA to provide support for parents until 1 year post partum, and hence this study could not explore parents' needs and experiences throughout the perinatal period.

Implications for Future Research and Practice

Considering the unmet needs of new parents and that the PDA can provide all parents with informational, socioemotional, and psychological benefits, mobile app–based perinatal interventions with in-built chatbots could be used to supplement the perinatal care provided for parents in Singapore. To optimize performance, future chatbots could have natural language processing and AI capabilities and have their answers vetted by local health care professionals. Future interventions could include more peer-to-peer interactions, games on emotional regulation and baby care skills, quizzes on perinatal care information, and personalized information for parents at significant milestones. Health care professionals could adopt midwife-led continuity of care models to deliver family-centered care. To address the parents' childcare needs, longer and equal paternity and maternity leaves could be explored, and flexible work arrangements could be promoted. Future research on more diverse parents (eg, from the LGBTQ community; single parents; unmarried parents; parents who had newborns with medical complications; and non–English-speaking parents) could be conducted to obtain a more holistic understanding of local parents' perinatal experiences. Educational resources in the PDA would also have to undergo revisions to meet the needs of varied family compositions and parents whose babies have medical complications. For instance, advice on bonding with a nonbiological child and tips on caring for children with medical complications could be provided. Future longitudinal mixed methods studies on the perinatal experiences of parents could also be conducted to triangulate the findings and obtain more in-depth data. These mixed methods studies could consider conducting quantitative surveys among larger sample of parents from both control and intervention groups to further compare their experiences and satisfaction in using the PDA. Future research could recruit larger sample sizes, explore gender- and parity-specific differences more in depth, and use more methods to collect qualitative data from participants such as observations. Furthermore, future trials could aim to develop and evaluate similar technology-based interventions for parents up to 1 year post partum to obtain a more holistic understanding of their experiences throughout the perinatal period. Finally, future research could engage grandparents to explore their views regarding childcare since they are also involved in caring for children in Singapore.

Conclusions

Overall, this study showed that the PDA can provide parents with informational, socioemotional, and psychological support during the perinatal period. Despite its limitations, this study has paved the way for the development of better mobile app–based interventions and chatbots to meet the needs of Asian parents during the perinatal period. Moving forward, researchers would need to overcome the problems posed by the limited available resources (funds, technological capabilities, and manpower) to design future perinatal supportive interventions with a more sophisticated chatbot, more gamification features, personalized and engaging care tips, and more peer-to-peer interactions. Health care providers could focus on delivering more family-centered, holistic, and continuous care for parents and babies. Policy makers and employers could be encouraged to help parents achieve a healthy work-family balance. Finally, collaboration between health care providers, researchers, and technical experts would be needed to develop appropriate and effective chatbot- and mobile app–based interventions to supplement perinatal care shortly.

Acknowledgments

The authors would like to thank the contributions of the following research assistants who assisted with the transcription of interviews and data management for this study: Ms Heather Pang Jing Wen, Ms Thong Kai Wen Charmaine, and Ms Sivaraman Suvetha. The authors would like to thank the National University Health System for their generous funding for this project. This work was supported by a Ministry of Education Tier 1, National University Health System Reimagine Research Fund—Type 1

XSL•FO
RenderX

(STEM) grant. Finally, the authors would like to thank the parents who participated in this study and generously shared their perinatal experiences with the authors.

Conflicts of Interest

None declared.

References

1. Woody CA, Ferrari AJ, Siskind DJ, Whiteford HA, Harris MG. A systematic review and meta-regression of the prevalence and incidence of perinatal depression. J Affect Disord 2017 Sep;219:86-92. [doi: 10.1016/j.jad.2017.05.003] [Medline: 28531848]
2. Philpott LF, Leahy-Warren P, FitzGerald S, Savage E. Stress in fathers in the perinatal period: a systematic review. Midwifery 2017 Dec;55:113-127. [doi: 10.1016/j.midw.2017.09.016] [Medline: 28992554]
3. Al-Abri K, Edge D, Armitage CJ. Prevalence and correlates of perinatal depression. Soc Psychiatry Psychiatr Epidemiol 2023 Nov 16;58(11):1581-1590 [FREE Full text] [doi: 10.1007/s00127-022-02386-9] [Medline: 36646936]
4. Dennis CL, Falah-Hassani K, Shiri R. Prevalence of antenatal and postnatal anxiety: systematic review and meta-analysis. Br J Psychiatry 2017 May 25;210(5):315-323 [FREE Full text] [doi: 10.1192/bjp.bp.116.187179] [Medline: 28302701]
5. Leiferman JA, Farewell CV, Jewell J, Rachael Lacy, Walls J, Harnke B, et al. Anxiety among fathers during the prenatal and postpartum period: a meta-analysis. J Psychosom Obstet Gynaecol 2021 Jun 25;42(2):152-161. [doi: 10.1080/0167482X.2021.1885025] [Medline: 33632067]
6. Rao WW, Zhu XM, Zong QQ, Zhang Q, Hall BJ, Ungvari GS, et al. Prevalence of prenatal and postpartum depression in fathers: a comprehensive meta-analysis of observational surveys. J Affect Disord 2020 Feb 15;263:491-499. [doi: 10.1016/j.jad.2019.10.030] [Medline: 31757623]
7. Cheng CY, Chou YH, Chang CH, Liou SR. Trends of perinatal stress, anxiety, and depression and their prediction on postpartum depression. Int J Environ Res Public Health 2021 Sep 03;18(17):9307 [FREE Full text] [doi: 10.3390/ijerph18179307] [Medline: 34501906]
8. Dikmen-Yildiz P, Ayers S, Phillips L. Depression, anxiety, PTSD and comorbidity in perinatal women in Turkey: a longitudinal population-based study. Midwifery 2017 Dec;55:29-37. [doi: 10.1016/j.midw.2017.09.001] [Medline: 28917088]
9. Cohen LS, Wang B, Nonacs R, Viguera AC, Lemon EL, Freeman MP. Treatment of mood disorders during pregnancy and postpartum. Psychiatr Clin North Am 2010 Jun;33(2):273-293. [doi: 10.1016/j.psc.2010.02.001] [Medline: 20385337]
10. Craig M, Howard L. Postnatal depression. BMJ Clin Evid 2009 Jan 26;2009:1407 [FREE Full text] [Medline: 19445768]
11. Charrois J, Côté SM, Paquin S, Séguin JR, Japel C, Vitaro F, et al. Maternal depression in early childhood and child emotional and behavioral outcomes at school age: examining the roles of preschool childcare quality and current maternal depression symptomatology. Eur Child Adolesc Psychiatry 2020 May 13;29(5):637-648. [doi: 10.1007/s00787-019-01385-7] [Medline: 31410578]
12. Dowse E, Chan S, Ebert L, Wynne O, Thomas S, Jones D, et al. Impact of perinatal depression and anxiety on birth outcomes: a retrospective data analysis. Matern Child Health J 2020 Jun 17;24(6):718-726. [doi: 10.1007/s10995-020-02906-6] [Medline: 32303935]
13. Ashraf S, Shah K, Vadukapuram R, Shah B, Jaiswal S, Mansuri Z, et al. Impact of paternal depression on child neurodevelopmental outcomes and disorders. Prim Care Companion CNS Disord 2023 Feb 07;25(1):22r03303 [FREE Full text] [doi: 10.4088/PCC.22r03303] [Medline: 36763820]
14. D'Souza S, Waldie KE, Peterson ER, Underwood L, Morton SM. Antenatal and postnatal determinants of behavioural difficulties in early childhood: evidence from growing up in New Zealand. Child Psychiatry Hum Dev 2019 Feb 2;50(1):45-60. [doi: 10.1007/s10578-018-0816-6] [Medline: 29860616]
15. Rogers A, Obst S, Teague SJ, Rossen L, Spry EA, Macdonald JA, et al. Association between maternal perinatal depression and anxiety and child and adolescent development: a meta-analysis. JAMA Pediatr 2020 Nov 01;174(11):1082-1092 [FREE Full text] [doi: 10.1001/jamapediatrics.2020.2910] [Medline: 32926075]
16. Gentile S, Fusco ML. Untreated perinatal paternal depression: effects on offspring. Psychiatry Res 2017 Jun;252:325-332. [doi: 10.1016/j.psychres.2017.02.064] [Medline: 28314228]
17. Teo J. Budget debate: new task force to boost women and children's health. The Straits Times. 2021 Mar 17. URL: https://www.straitstimes.com/singapore/health/helping-children-get-a-head-start-even-before-they-are-born [accessed 2022-03-28]
18. Koh YLM, Tan LCA, Shorey S. Perceptions of distressed fathers in the early postpartum period: a descriptive qualitative study. J Fam Issues 2020 Dec 21;42(10):2397-2417. [doi: 10.1177/0192513x20980042]
19. Shorey S, Tan TC, Thilagamangai, Mathews J, Lim SH, Shi L, et al. Perinatal support for couples during COVID-19: a descriptive qualitative study. J Clin Nurs 2023 Jul 30;32(13-14):3528-3542. [doi: 10.1111/jocn.16432] [Medline: 35773956]
20. Ng KL, Buvanaswari P, Loh LW, Chee CY, Teng JY, Wang W, et al. A descriptive qualitative study exploring the experiences of fathers with partners suffering from maternal perinatal depression. Midwifery 2021 Nov;102:103075. [doi: 10.1016/j.midw.2021.103075] [Medline: 34237515]

XSL•FO
RenderX

21. Ngai FW, Chan PS. Perception of family sense of coherence during parental transition: a qualitative study. J Health Psychol 2021 Nov;26(13):2435-2449. [doi: 10.1177/1359105320914062] [Medline: 32306765]
22. Jusoh S. A survey on trend, opportunities and challenges of mHealth apps. Int J Interact Mob Technol 2017 Nov 27;11(6):73. [doi: 10.3991/ijim.v11i6.7265]
23. Osma J, Barrera AZ, Ramphos E. Are pregnant and postpartum women interested in health-related apps? Implications for the prevention of perinatal depression. Cyberpsychol Behav Soc Netw 2016 Jun;19(6):412-415. [doi: 10.1089/cyber.2015.0549] [Medline: 27327069]
24. Digital health. World Health Organization. URL: https://www.who.int/health-topics/digital-health/#tab=tab_1 [accessed 2023-02-16]
25. Charbonneau DH, Hightower S, Katz A, Zhang K, Abrams J, Senft N, et al. Smartphone apps for cancer: a content analysis of the digital health marketplace. Digit Health 2020 Feb 11;6:2055207620905413 [FREE Full text] [doi: 10.1177/2055207620905413] [Medline: 32110428]
26. Litvin CB, Ornstein SM. Development and initial use of an evidence-based home blood pressure monitoring mHealth platform. J Med Syst 2023 Apr 29;47(1):53. [doi: 10.1007/s10916-023-01955-y] [Medline: 37118616]
27. Wu A, Scult MA, Barnes ED, Betancourt JA, Falk A, Gunning FM. Smartphone apps for depression and anxiety: a systematic review and meta-analysis of techniques to increase engagement. NPJ Digit Med 2021 Feb 11;4(1):20 [FREE Full text] [doi: 10.1038/s41746-021-00386-8] [Medline: 33574573]
28. Chua JY, Shorey S. Effectiveness of mobile application-based perinatal interventions in improving parenting outcomes: a systematic review. Midwifery 2022 Nov;114:103457 [FREE Full text] [doi: 10.1016/j.midw.2022.103457] [Medline: 35985142]
29. Smartphone market in Singapore - statistics and facts. Statista. URL: https://www.statista.com/topics/5842/smartphones-in-singapore/ [accessed 2023-02-16]
30. Shorey S, Lau Y, Dennis CL, Chan YS, Tam WW, Chan YH. A randomized-controlled trial to examine the effectiveness of the 'Home-but not Alone' mobile-health application educational programme on parental outcomes. J Adv Nurs 2017 Sep 07;73(9):2103-2117. [doi: 10.1111/jan.13293] [Medline: 28276086]
31. Shorey S, Ng YP, Ng ED, Siew AL, Mörelius E, Yoong J, et al. Effectiveness of a technology-based supportive educational parenting program on parental outcomes (part 1): randomized controlled trial. J Med Internet Res 2019 Feb 13;21(2):e10816 [FREE Full text] [doi: 10.2196/10816] [Medline: 30758289]
32. Shorey S, Law E, Thilagamangai, Mathews J, Lim SH, Shi L, et al. Evaluating the effectiveness of the supportive parenting app on parental outcomes: randomized controlled trial. J Med Internet Res 2023 Jan 16;25:e41859 [FREE Full text] [doi: 10.2196/41859] [Medline: 36645699]
33. Shorey S, Ng ED. Evaluation of mothers' perceptions of a technology-based supportive educational parenting program (part 2): qualitative study. J Med Internet Res 2019 Feb 13;21(2):e11065 [FREE Full text] [doi: 10.2196/11065] [Medline: 30758295]
34. Shorey S, Yang YY, Dennis CL. A mobile health app-based postnatal educational program (Home-but not Alone): descriptive qualitative study. J Med Internet Res 2018 Apr 19;20(4):e119 [FREE Full text] [doi: 10.2196/jmir.9188] [Medline: 29674314]
35. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
36. Chua JY, Choolani M, Chee CY, Chan YH, Lalor JG, Chong YS, et al. Insights of parents and parents-to-be in using chatbots to improve their preconception, pregnancy, and postpartum health: a mixed studies review. J Midwifery Womens Health 2023 Feb 03;68(4):480-489. [doi: 10.1111/jmwh.13472] [Medline: 36734375]
37. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
38. Chua JY, Choolani M, Chee CY, Yi H, Chan YH, Lalor JG, et al. 'Parentbot - a digital healthcare assistant (PDA)': a mobile application-based perinatal intervention for parents: development study. Patient Educ Couns 2023 Sep;114:107805. [doi: 10.1016/j.pec.2023.107805] [Medline: 37245443]
39. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: medical research council guidance. BMJ 2015 Mar 19;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
40. Agius SJ. Qualitative research: its value and applicability. Psychiatrist 2018 Jan 02;37(6):204-206. [doi: 10.1192/pb.bp.113.042770]
41. Tenny S, Brannan JM, Brannan GD. Qualitative study. StatPearls. 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK470395/ [accessed 2024-04-29]
42. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]

XSL•FO
RenderX

43. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
44. Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research: interviews and focus groups. Br Dent J 2008 Mar 22;204(6):291-295. [doi: 10.1038/bdj.2008.192] [Medline: 18356873]
45. Leerkes EM, Crockenberg SC. The development of maternal self‐efficacy and its impact on maternal behavior. Infancy 2010 Jan 18;3(2):227-247. [doi: 10.1207/s15327078in0302_7]
46. Guest G, Bunce A, Johnson L. How many interviews are enough?: an experiment with data saturation and variability. Field methods 2016 Jul 21;18(1):59-82. [doi: 10.1177/1525822X05279903]
47. Doody O, Noonan M. Preparing and conducting interviews to collect data. Nurse Res 2013 May;20(5):28-32. [doi: 10.7748/nr2013.05.20.5.28.e327] [Medline: 23687846]
48. McGrath C, Palmgren PJ, Liljedahl M. Twelve tips for conducting qualitative research interviews. Med Teach 2019 Sep 28;41(9):1002-1006. [doi: 10.1080/0142159X.2018.1497149] [Medline: 30261797]
49. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019 Mar 08;7(2):e000057 [FREE Full text] [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
50. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
51. Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry. Int J Intercult Relat 1985 Jan;9(4):438-439. [doi: 10.1016/0147-1767(85)90062-8]
52. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 02;16(1):160940691773384. [doi: 10.1177/1609406917733847]
53. Buetow S. Apophenia, unconscious bias and reflexivity in nursing qualitative research. Int J Nurs Stud 2019 Jan;89:8-13. [doi: 10.1016/j.ijnurstu.2018.09.013] [Medline: 30316055]
54. Barrett A, Kajamaa A, Johnston J. How to … be reflexive when conducting qualitative research. Clin Teach 2020 Feb 23;17(1):9-12. [doi: 10.1111/tct.13133] [Medline: 31970937]
55. Population trends 2020. Department of Statistics, Singapore. URL: https://www.singstat.gov.sg/-/media/files/publications/population/population2020.pdf [accessed 2023-08-14]
56. Corno G, Espinoza M, Maria Baños R. A narrative review of positive psychology interventions for women during the perinatal period. J Obstet Gynaecol 2019 Oct 09;39(7):889-895. [doi: 10.1080/01443615.2019.1581735] [Medline: 31179814]
57. Sbrilli MD, Haigler K, Laurent HK. The indirect effect of parental intolerance of uncertainty on perinatal mental health via mindfulness during COVID-19. Mindfulness (N Y) 2021 Jun 02;12(8):1999-2008 [FREE Full text] [doi: 10.1007/s12671-021-01657-x] [Medline: 34093889]
58. Kim HJ, Kim SM, Shin H, Jang JS, Kim YI, Han DH. A mobile game for patients with breast cancer for chemotherapy self-management and quality-of-life improvement: randomized controlled trial. J Med Internet Res 2018 Oct 29;20(10):e273 [FREE Full text] [doi: 10.2196/jmir.9559] [Medline: 30578205]
59. Kim Y, Hong S, Choi M. Effects of serious games on depression in older adults: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Sep 06;24(9):e37753 [FREE Full text] [doi: 10.2196/37753] [Medline: 36066964]
60. Ong J. S'pore population better educated across age, ethnicity; women make greater strides. The Straits Times. URL: https://www.straitstimes.com/singapore/spore-population-better-educated-across-age-ethnicity-women-make-greater-strides [accessed 2023-08-30]
61. Wong C. National movement launched to encourage digitalisation, President's challenge to provide $2.5 million in seed funding. The Straits Times. URL: https://tinyurl.com/ytsm4f3h [accessed 2023-08-30]
62. Withers M, Kharazmi N, Lim E. Traditional beliefs and practices in pregnancy, childbirth and postpartum: a review of the evidence from Asian countries. Midwifery 2018 Jan;56:158-170. [doi: 10.1016/j.midw.2017.10.019] [Medline: 29132060]
63. White B, Giglia RC, White JA, Dhaliwal S, Burns SK, Scott JA. Gamifying breastfeeding for fathers: process evaluation of the milk man mobile app. JMIR Pediatr Parent 2019 Jun 20;2(1):e12157 [FREE Full text] [doi: 10.2196/12157] [Medline: 31518324]
64. Here's how Singapore is building a tech ecosystem that extends beyond its shores. Singapore Economic Development Board. URL: https://tinyurl.com/3h853n74 [accessed 2023-08-30]
65. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
66. Staines R. 'Generation Z' happy to be diagnosed by chatbots. Pharmaphorum – An EVERSANA Company. URL: https://pharmaphorum.com/news/generation-z-happy-to-be-diagnosed-by-chatbots/ [accessed 2022-10-06]
67. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-39. [doi: 10.1145/3485874]

XSL•FO
RenderX

68. Khalik S. Budget debate: more healthcare workers needed in coming years. The Straits Times. URL: https://www.straitstimes.com/singapore/politics/budget-debate-more-healthcare-workers-needed-in-coming-years [accessed 2023-08-15]
69. Nursing and midwifery. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/nursing-and-midwifery [accessed 2023-08-15]
70. Aubel J. Grandmothers - a neglected family resource for saving newborn lives. BMJ Glob Health 2021 Feb 15;6(2):e003808 [FREE Full text] [doi: 10.1136/bmjgh-2020-003808] [Medline: 33589417]
71. Tang F, Li K, Jang H, Rauktis MB. Depressive symptoms in the context of Chinese grandparents caring for grandchildren. Aging Ment Health 2022 Jun 10;26(6):1120-1126. [doi: 10.1080/13607863.2021.1910788] [Medline: 33843385]
72. Komonpaisarn T, Loichinger E. Providing regular care for grandchildren in Thailand: an analysis of the impact on grandparents' health. Soc Sci Med 2019 May;229:117-125. [doi: 10.1016/j.socscimed.2018.05.031] [Medline: 29804635]
73. Tan T. First-time mums in Singapore are older, more educated. The Straits Times. URL: https://www.straitstimes.com/singapore/community/first-time-mums-in-singapore-are-older-more-educated [accessed 2023-08-17]
74. Shorey S, Chee CY, Ng ED, Lau Y, Dennis CL, Chan YH. Evaluation of a technology-based peer-support intervention program for preventing postnatal depression (part 1): randomized controlled trial. J Med Internet Res 2019 Aug 29;21(8):e12410 [FREE Full text] [doi: 10.2196/12410] [Medline: 31469084]
75. Dennis CL. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321-332. [doi: 10.1016/S0020-7489(02)00092-5]
76. Rice C, Ingram E, O'Mahen H. A qualitative study of the impact of peer support on women's mental health treatment experiences during the perinatal period. BMC Pregnancy Childbirth 2022 Sep 06;22(1):689 [FREE Full text] [doi: 10.1186/s12884-022-04959-7] [Medline: 36068490]
77. de Lima DP, Gerosa MA, Conte TU, de M. Netto JF. What to expect, and how to improve online discussion forums: the instructors' perspective. J Internet Serv Appl 2019 Dec 04;10(1):22. [doi: 10.1186/s13174-019-0120-0]
78. Barimani M, Vikström A. Successful early postpartum support linked to management, informational, and relational continuity. Midwifery 2015 Aug;31(8):811-817. [doi: 10.1016/j.midw.2015.04.009] [Medline: 25982847]
79. Rigg E, Dahlen HG. Woman centred care: has the definition been morphing of late? Women Birth 2021 Feb;34(1):1-3. [doi: 10.1016/j.wombi.2020.12.013] [Medline: 33500123]
80. D'haenens F, Van Rompaey B, Swinnen E, Dilles T, Beeckman K. The effects of continuity of care on the health of mother and child in the postnatal period: a systematic review. Eur J Public Health 2020 Aug 01;30(4):749-760. [doi: 10.1093/eurpub/ckz082] [Medline: 31121019]
81. Bradford BF, Wilson AN, Portela A, McConville F, Fernandez Turienzo C, Homer CS. Midwifery continuity of care: a scoping review of where, how, by whom and for whom? PLOS Glob Public Health 2022 Oct 5;2(10):e0000935 [FREE Full text] [doi: 10.1371/journal.pgph.0000935] [Medline: 36962588]
82. Edmonds J, Ivanof J, Kafulafula U. Midwife led units: transforming maternity care globally. Ann Glob Health 2020 Apr 28;86(1):44 [FREE Full text] [doi: 10.5334/aogh.2794] [Medline: 32377509]
83. Liew K, Kimball C. Continuity of care in pregnancy and labour – the EMMa care experience. Int J Evid Based Healthc 2014;12(3):209 [FREE Full text] [doi: 10.1097/01.xeb.0000455223.48433.44]
84. Voon ST, Lay JT, San WT, Shorey S, Lin SK. Comparison of midwife-led care and obstetrician-led care on maternal and neonatal outcomes in Singapore: a retrospective cohort study. Midwifery 2017 Oct;53:71-79. [doi: 10.1016/j.midw.2017.07.010] [Medline: 28778037]
85. Antoniou E, Tzanoulinou MD, Stamoulou P, Orovou E. The important role of partner support in women's mental disorders during the perinatal period. A literature review. Maedica (Bucur) 2022 Mar 18;17(1):194-200 [FREE Full text] [doi: 10.26574/maedica.2022.17.1.194] [Medline: 35733735]
86. Malinen K, Rönkä A, Sevón E, Schoebi D. The difficulty of being a professional, a parent, and a spouse on the same day: daily spillover of workplace interactions on parenting, and the role of spousal support. J Prev Interv Community 2017 Jun 22;45(3):156-167. [doi: 10.1080/10852352.2016.1198121] [Medline: 28641062]
87. Sriyasak A, Almqvist AL, Sridawruang C, Neamsakul W, Häggström-Nordin E. The new generation of Thai fathers: breadwinners involved in parenting. Am J Mens Health 2018 Sep 23;12(5):1368-1378 [FREE Full text] [doi: 10.1177/1557988316651062] [Medline: 27215945]
88. Ho HZ, Lam YW. Father involvement in East Asia. In: Sheldon SB, Turner-Vorbeck TA, editors. The Wiley Handbook of Family, School, and Community Relationships in Education. Hoboken, NJ: John Wiley & Sons; 2018:333-356.
89. Xue WL, Shorey S, Wang W, He HG. Fathers' involvement during pregnancy and childbirth: an integrative literature review. Midwifery 2018 Jul;62:135-145. [doi: 10.1016/j.midw.2018.04.013] [Medline: 29680523]
90. Tohme P, Abi-Habib R. Correlates of work productivity and maternal competence after having a baby: the roles of mother-infant bonding and maternal subjective experiences. BMC Womens Health 2022 Sep 27;22(1):392 [FREE Full text] [doi: 10.1186/s12905-022-01958-w] [Medline: 36167585]
91. Goh YH. Flexi-work, fertility health part of holistic framework to encourage couples to have kids: indranee. The Straits Times. URL: https://tinyurl.com/5e3fzcya [accessed 2023-08-14]

XSL•FO
RenderX

92. Chin SF. Staff requests for flexi-work arrangements must be considered by employers by 2024. The Straits Times. URL: https://www.straitstimes.com/singapore/staff-requests-for-flexi-work-arrangements-must-be-considered-by-employers-by-2024 [accessed 2023-08-14]
93. Breastfeeding. World Health Organization. URL: https://www.who.int/health-topics/breastfeeding#tab=tab_1 [accessed 2023-08-15]
94. Ang S. Workplaces should support mums as more breastfeed their children: HPB. The Straits Times. URL: https://www.straitstimes.com/singapore/breastfeeding-on-the-rise-in-singapore-maternity-hospitals [accessed 2023-08-31]
95. Tokhi M, Comrie-Thomson L, Davis J, Portela A, Chersich M, Luchters S. Involving men to improve maternal and newborn health: a systematic review of the effectiveness of interventions. PLoS One 2018 Jan 25;13(1):e0191620 [FREE Full text] [doi: 10.1371/journal.pone.0191620] [Medline: 29370258]
96. Ang HM. Budget 2023: government-paid paternity leave to double to 4 weeks; mandatory for employers in due course. Channel New Asia. URL: https://www.channelnewsasia.com/singapore/budget-2023-paternity-leave-doubled-4-weeks-government-paid-working-mothers-child-relief-3276511 [accessed 2023-08-16]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
LGBTQ: lesbian, gay, bisexual, transgender, and queer
PDA: Parentbot–a Digital Healthcare Assistant
PES: Parenting Efficacy Scale
RCT: randomized controlled trial

Edited by A Mavragani; submitted 29.01.24; peer-reviewed by K Haßdenteufel, R Otte; comments to author 28.02.24; revised version received 16.03.24; accepted 18.04.24; published 21.06.24.

Please cite as:
Chua JYX, Choolani M, Chee CYI, Yi H, Chan YH, Lalor JG, Chong YS, Shorey S
Parents' Perceptions of Their Parenting Journeys and a Mobile App Intervention (Parentbot—A Digital Healthcare Assistant): Qualitative Process Evaluation
J Med Internet Res 2024;26:e56894
URL: https://www.jmir.org/2024/1/e56894
doi:10.2196/56894
PMID:

©Joelle Yan Xin Chua, Mahesh Choolani, Cornelia Yin Ing Chee, Huso Yi, Yiong Huak Chan, Joan Gabrielle Lalor, Yap Seng Chong, Shefaly Shorey. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of the Quality of Discharge Letters Written by Large Language Models and Junior Clinicians: Single-Blinded Study

Joshua Yi Min Tung[1,2], MPH; Sunil Ravinder Gill[1], MBBS; Gerald Gui Ren Sng[2,3], MMed; Daniel Yan Zheng Lim[2,4], MTech; Yuhe Ke[2,5], MMed; Ting Fang Tan[6], MBBS; Liyuan Jin[7], BEng; Kabilan Elangovan[8], BEng; Jasmine Chiat Ling Ong[9], PharmD; Hairil Rizal Abdullah[2,5], PhD; Daniel Shu Wei Ting[6,7,8], PhD; Tsung Wen Chong[1], DPhil

[1]Department of Urology, Singapore General Hospital, Singapore, Singapore
[2]Data Science and Artificial Intelligence Laboratory, Singapore General Hospital, Singapore, Singapore
[3]Department of Endocrinology, Singapore General Hospital, Singapore, Singapore
[4]Department of Gastroenterology and Hepatology, Singapore General Hospital, Singapore, Singapore
[5]Department of Anesthesiology, Singapore General Hospital, Singapore, Singapore
[6]Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore
[7]Duke-NUS Medical School, Singapore, Singapore
[8]Artificial Intelligence Office, Singapore Health Services, Singapore, Singapore
[9]Division of Pharmacy, Singapore General Hospital, Singapore, Singapore

Corresponding Author:
Joshua Yi Min Tung, MPH
Department of Urology
Singapore General Hospital
16 College Road, Block 4 Level 1
Singapore, 169854
Singapore
Phone: 65 62223322
Email: joshua.tung@gmail.com

Abstract

Background: Discharge letters are a critical component in the continuity of care between specialists and primary care providers. However, these letters are time-consuming to write, underprioritized in comparison to direct clinical care, and are often tasked to junior doctors. Prior studies assessing the quality of discharge summaries written for inpatient hospital admissions show inadequacies in many domains. Large language models such as GPT have the ability to summarize large volumes of unstructured free text such as electronic medical records and have the potential to automate such tasks, providing time savings and consistency in quality.

Objective: The aim of this study was to assess the performance of GPT-4 in generating discharge letters written from urology specialist outpatient clinics to primary care providers and to compare their quality against letters written by junior clinicians.

Methods: Fictional electronic records were written by physicians simulating 5 common urology outpatient cases with long-term follow-up. Records comprised simulated consultation notes, referral letters and replies, and relevant discharge summaries from inpatient admissions. GPT-4 was tasked to write discharge letters for these cases with a specified target audience of primary care providers who would be continuing the patient's care. Prompts were written for safety, content, and style. Concurrently, junior clinicians were provided with the same case records and instructional prompts. GPT-4 output was assessed for instances of hallucination. A blinded panel of primary care physicians then evaluated the letters using a standardized questionnaire tool.

Results: GPT-4 outperformed human counterparts in information provision (mean 4.32, SD 0.95 vs 3.70, SD 1.27; *P*=.03) and had no instances of hallucination. There were no statistically significant differences in the mean clarity (4.16, SD 0.95 vs 3.68, SD 1.24; *P*=.12), collegiality (4.36, SD 1.00 vs 3.84, SD 1.22; *P*=.05), conciseness (3.60, SD 1.12 vs 3.64, SD 1.27; *P*=.71), follow-up recommendations (4.16, SD 1.03 vs 3.72, SD 1.13; *P*=.08), and overall satisfaction (3.96, SD 1.14 vs 3.62, SD 1.34; *P*=.36) between the letters generated by GPT-4 and humans, respectively.

Conclusions: Discharge letters written by GPT-4 had equivalent quality to those written by junior clinicians, without any hallucinations. This study provides a proof of concept that large language models can be useful and safe tools in clinical documentation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57721) doi:10.2196/57721

KEYWORDS

artificial intelligence; AI; discharge summaries; continuity of care; large language model; LLM; junior clinician; letter writing; single-blinded; ChatGPT; urology; primary care; fictional electronic record; consultation note; referral letter; simulated environment

Introduction

Large language models (LLMs) are artificial intelligence (AI) constructs with the capability to parse and generate human-like text based on pretraining with a vast corpus of unstructured text data. LLMs have shown promise in parsing large corpi of text data and generating human-like text across various domains, including health care [1].

One of the most promising applications of LLMs in health care is for automating tasks traditionally performed by health care professionals, such as writing discharge summaries and letters [2]. Despite the fact that such documents serve as a critical communication tool between hospital specialists and primary care providers (PCPs) in ensuring continuity of care, they are time-consuming to write, are often tasked to junior doctors, and are underprioritized when balanced against direct clinical care [3,4].

Prior studies have shown that inpatient hospital discharge summary quality is inadequate in many of the domains/elements required by The Joint Commission and endorsed at the Transitions of Care Consensus Conference (TOCCC) [3]. Many elements, including the provision of follow-up care plans, are of particular importance to patients being discharged to primary care. Improving discharge summary timeliness was shown to have positive effects on the quality of content and the transmission rate to PCPs [5,6].

The automation of inpatient discharge summaries has previously been explored using electronic discharge summary (EDS) programs. Maslove et al [7] showed no differences in PCP-reported overall quality between summaries generated by an EDS program and conventional dictated summaries. However, these studies assessed communication between inpatient and outpatient providers for single episodes of hospital admissions rather than summative information regarding a patient's clinical condition across a longer follow-up, as would be the case with specialist outpatient care.

Moreover, the applications of LLMs in outpatient clinical documentation remain largely unexplored. Unlike inpatient admissions, which are typically self-contained and time-limited, outpatient specialist care can have more complex clinical documentation with longitudinal health records that may span across years. It remains to be seen if LLMs are able to parse clinical information with a strong temporal dimension and produce high-quality output.

There is currently no consensus on the single best way to objectively evaluate clinically relevant LLM output. However, approaches comparing LLM output against that generated by clinicians can help with assessment of a benchmark of "human-like" or "clinician-like" performance, which have been described in the related literature [8]. In addition, independent expert evaluation of LLM output has also been used for physician evaluation of LLM output [9]. We sought to combine both approaches in this study.

The aims of this study were to (1) evaluate the feasibility and quality of an LLM in writing discharge letters for patients discharged from a specialist outpatient clinic and (2) compare the discharge letters generated against similar letters written by human junior clinicians.

Methods

Ethical Considerations

This study was conducted in a simulated environment using only fictional patient data. As the use of fictional data does not fall under local Human Biomedical Research Act regulations [10], ethical approval was not required.

Data Generation

Fictional patient data were generated by clinicians from the Singapore General Hospital's Department of Urology, mimicking typical electronic medical records (EMRs) of 5 common patient groups seen at the outpatient clinic in a tertiary center. The time horizon for follow-up ranged from 6 months to 6 years. Data included initial referral letters from PCPs or emergency departments, initial and follow-up urology consultation notes, referrals to other departments, correspondence notes, and discharge summaries from relevant admissions. Information on the fictional cases is presented in Table 1. Each case varied in complexity, ranging between 4 and 10 documents, and included 1238-3009 characters (424-1110 tokens).

XSL•FO
RenderX

Table 1. Fictional case-mix data used for assessment.

Simulated case	Initial referral	Clinical condition	Follow-up duration	Additional/incidental clinical information
1	Elevated PSA[a]	Benign prostatic hyperplasia	6 years	Initially declined biopsy and opted for PSA monitoring, but ultimately underwent a transperineal prostate biopsy and developed acute retention of urine postoperatively
2	LUTS[b]	Overactive bladder	3 years	Failed behavioral modification measures and experienced adverse effects from multiple lines of anticholinergic medications
3	Asymptomatic microhematuria	Renal calculus	1.5 years	Incidental pancreatic lesion on imaging and was referred to general surgery for further evaluation
4	Ureteric colic	Distal ureteric stone	6 months	Failed medical expulsive therapy and underwent ureteroscopy and lithotripsy; also noted to have hypertension perioperatively
5	Erectile dysfunction	Erectile dysfunction	2.5 years	Noted to have abnormal ECG[c] and was referred to cardiology for further assessment prior to initiation of PDE-5[d] inhibitors

[a]PSA: prostate-specific antigen.

[b]LUTS: lower urinary tract symptoms.

[c]ECG: electrocardiography.

[d]PDE-5: phosphodiesterase-5.

Discharge Letter Generation

GPT-4 is a state-of-the-art LLM developed by OpenAI with a parameter count of 1.76 trillion, and has exhibited human-level performance on professional and academic benchmarks [11]. GPT-4, with an 8000-token context length, was provided with the fictional EMRs and tasked to write discharge letters to PCPs. The prompt instructed GPT-4 to assume the role of a physician assistant in a urology clinic, specifying the context of the task (“this patient is being discharged”) and the target audience (“meant to be read by the general practitioner”). To reduce the risk of hallucination, prompts for safety included specific instructions to include only information provided in the fictional EMR [12]. Desired content guidelines were provided to standardize the structure of the discharge letter. Prompts for style included instructions to write in prose, in a cordial and concise manner, and not to exceed half a page of text unless necessary. To evaluate the generative capacity of the LLM, only the above instructions (but no examples) were provided to GPT-4, which is a technique known as “zero-shot” prompting [13].

Three separate discharge letters were generated for each fictional case. One was generated by GPT-4 and the comparators were two letters that were written concurrently by junior clinicians from the Department of Urology using the same set of fictional patients. For the comparator letters, use of generative AI tools or other automated summarization methods was disallowed.

Full prompt instructions are shown in Multimedia Appendix 1. Sample medical records and discharge letters are shown in Multimedia Appendix 2.

Comparison of Discharge Letters

All letters underwent an initial independent screening by two study team members (JYMT and SRG) for factual inaccuracy and grammatical errors. For letters generated by GPT-4, this included screening for instances of hallucination by the language model. For cases 3 and 4, letters were also assessed with respect to their attention to secondary problems beyond the prompted surgical issue based on the presence or absence of follow-up recommendations for other medical issues found during the course of the patient’s treatment.

A panel of 5 senior primary care physicians was presented with the discharge letters written by GPT-4 and junior clinicians in a blinded fashion. The participating physicians on the expert panel had an average clinical practice experience of 34.4 years. They compared the letters produced by GPT-4 and junior clinicians using a standardized rubric, but they were not informed that one of the letters was LLM-generated. Letters were evaluated on a 5-point Likert scale for completeness of information, conciseness, clarity, collegiality, whether follow-up care plans were articulated, and a single overall satisfaction question (Figure 1). Since there are no validated tools for assessment of outpatient discharge letter quality, questionnaire items were selected based on elements endorsed by the TOCCC for inpatient summaries.

XSL•FO
RenderX

Figure 1. Scoring questionnaire.

Question	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1. Information about the patient's diagnosis and clinical course is satisfactory.	☐	☐	☐	☐	☐
2. The discharge summary is appropriately concise.	☐	☐	☐	☐	☐
3. Information in the discharge summary is presented with clarity.	☐	☐	☐	☐	☐
4. The discharge summary is collegial.	☐	☐	☐	☐	☐
5. Recommendations for follow-up are clear.	☐	☐	☐	☐	☐
6. Overall, I would be happy to receive this discharge summary as the patient's next managing physician.	☐	☐	☐	☐	☐

Data Analysis

SPSS version 26.0 (IBM Corp) was used for the statistical analysis of quantitative data. Total scores and arithmetic means with SDs were calculated for all responses as a whole and for responses to each individual item. As the data were nonnormally distributed (Shapiro-Wilk test $P<.001$), differences in means were compared using the Mann-Whitney U test.

Results

Discharged letters generated by GPT-4 averaged 276 words (1462 characters), in comparison to 155 words (737 characters) for letters written by junior clinicians. No instances of hallucination and no grammatical errors were present in the letters generated by GPT-4. While no factual inaccuracies were found in the letters written by the junior clinicians, they made an average of 1.7 grammatical errors per letter.

Each panel member scored all 15 letters, for a total of 75 responses. The mean total score per letter was 22.99 (SD 6.26). GPT-4 scored higher on question items 1 (information), 3 (clarity), 4 (collegiality), 5 (follow-up plans), and 6 (overall satisfaction), and received higher total mean scores as compared to those of human subjects, although this difference was only statistically significant for item 1. GPT-4 received lower scores on item 2 (conciseness). The complete results are shown in Table 2.

Table 2. Comparison of mean scores for each evaluation item of discharge letters written by GPT-4 and junior clinicians.

Item	GPT-4 (n=25), mean (SD)	Junior clinicians (n=50), mean (SD)	*P* value
Information	4.32 (0.95)	3.70 (1.27)	.03
Conciseness	3.60 (1.12)	3.64 (1.27)	.71
Clarity	4.16 (0.94)	3.68 (1.24)	.12
Collegiality	4.36 (1.00)	3.84 (1.22)	.05
Recommendations	4.16 (1.03)	3.72 (1.13)	.08
Overall satisfaction	3.96 (1.14)	3.62 (1.34)	.36
Total	24.56 (5.13)	22.20 (6.66)	.20

GPT-4 also had some unanticipated strengths when compared against the letters written by the junior clinicians. In particular, GPT-4 demonstrated attention to secondary problems that were not explicitly included in the original prompt. In cases 3 and 4, GPT-4 provided follow-up recommendations based on the given fictional records for the incidental pancreatic lesion ("we recommend annual imaging studies to monitor the renal stone and the pancreatic lesion, in collaboration with the General Surgery team") and perioperative hypertension ("he is recommended to continue with Amlodipine 5mg daily for his hypertension"). In contrast, although the junior clinicians also made mention of the pancreatic lesion and subsequent general surgery referral, they did not provide follow-up recommendations to the PCP.

XSL•FO
RenderX

Discussion

Principal Findings

To our knowledge, this study represents the first blinded comparison of discharge letters written by an advanced LLM versus human physicians.

Our analysis showed that GPT-4 performed comparably against junior clinicians from an inpatient specialist department, with equivalent scores in clarity, collegiality, and ability to articulate follow-up recommendations. GPT-4 also demonstrated the ability to generate a holistic patient summary with attention to secondary medical issues and made fewer grammatical errors than human comparators. The higher score in information provision highlights the capability of language models with strong embeddings to capture and present important details from patient EMRs. However, this appears to have come at the expense of a relative lack of conciseness, as indicated by GPT-4 receiving a lower score in this domain. More specific metaprompting with instructions for word limits could overcome this limitation. Nevertheless, the importance of an LLM's ability to rapidly digest and summarize a large free-text corpus cannot be overstated, which offers the potential for more consistent information relay between health care providers.

These findings echo prior work in the field comparing EDSs against traditional dictated summaries. While earlier digitization efforts were limited by technological capabilities of the era, advances in natural language processing and linguistic ability allow modern language models to write fluently and mimic human prose convincingly. While the potential of language models in this area of health care has been discussed [2], these first results prove that LLMs do have the capability to automate time-consuming tasks for health care professionals. This is a transformative opportunity that could be a useful tool in bridging communications between PCPs and inpatient specialists by ensuring the quality and accuracy of discharge letters. Clinicians burdened with documentation in addition to clinical duties stand to benefit immensely from automation of these tasks, allowing more time for patient contact and clinical care. Likewise, patients could benefit from consistent, timely handovers and a more robust system of continuity of care, instead of the vagaries of the junior clinicians writing their discharge letters. With appropriate refinement of prompt instructions, we postulate that this transformative benefit can extend beyond automating the documentation work of physicians to other health care professionals (eg, nurses and pharmacists).

Limitations and Prospects

The exclusive reliance on fictional patient data in this study, while necessary for ethical and patient data considerations, might not capture the real-world complexities in patient EMRs. We attempted to overcome this limitation by having the fictional EMRs written by physicians who also regularly see patients in the urology outpatient clinic on a day-to-day basis and enter clinical notes into actual patient EMRs. The small number of panel assessors as well as the use of clinical cases from a single surgical specialty may reduce the generalizability of these findings. No evaluation was performed as to whether the panel assessors inferred that any of the letters were LLM-generated. The time taken for junior clinicians to write discharge letters was not measured, precluding comparison in terms of time savings against an LLM. In addition, despite the advantages of LLM tools, misinformation phenomena such as hallucination and concept drift remain areas of concern. We propose that such an AI tool should be used to augment, not replace, human-written discharge letters, and that physician oversight should still be required before discharge documents are handed to patients. Users of LLMs as productivity aids in clinical medicine must also be aware of prevailing privacy protection policies, and closed-access LLM implementations (eg, on a private server) may be needed for regulatory compliance. While automating discharge letter writing may undermine the clinical reasoning process that physicians undertake [14], we believe that LLMs implemented with a "human in the loop" can eventually exert a long-term training effect, reinforcing these cognitive skill sets [15].

Future research using actual patient EMRs will be necessary to confirm these preliminary findings. Studies may be expanded to care across different medical specialties to ascertain the ability of language models to comprehend abbreviations and concepts from different disciplines. Fine-tuning a language model on medical information and patient EMRs and refining metaprompts are avenues to further improve the quality of discharge letters generated by LLMs and to balance conciseness with information density.

Conclusion

The emergence of AI in health care promises a paradigm shift in the way clinical medicine is practiced. Our study provides insight into AI's capacity to optimize the continuity of patient care. As evaluation of language models for clinical applications continues to advance, we propose benchmarking their performance against human counterparts to determine feasibility and assess output quality.

Data Availability

The data sets generated and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JYMT, GGRS, DYZL, and TWC contributed to the conceptualization of the study. JYMT performed the data curation and formal analysis. SRG contributed to the writing of the original draft. HRA, DSWT, and TWC supervised the study. YK, TFT, LJ, KE, and JCLO contributed to revisions and editing of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt instructions.
[PNG File , 117 KB - jmir_v26i1e57721_app1.png]

Multimedia Appendix 2
Sample fictional clinical records and discharge letters.
[DOCX File , 56 KB - jmir_v26i1e57721_app2.docx]

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
3. Horwitz LI, Jenq GY, Brewster UC, Chen C, Kanade S, Van Ness PH, et al. Comprehensive quality of discharge summaries at an academic medical center. J Hosp Med 2013 Aug 22;8(8):436-443 [FREE Full text] [doi: 10.1002/jhm.2021] [Medline: 23526813]
4. Rao P, Andrei A, Fried A, Gonzalez D, Shine D. Assessing quality and efficiency of discharge summaries. Am J Med Qual 2005 Sep 03;20(6):337-343. [doi: 10.1177/1062860605281078] [Medline: 16280397]
5. Weetman K, Spencer R, Dale J, Scott E, Schnurr S. What makes a "successful" or "unsuccessful" discharge letter? Hospital clinician and general practitioner assessments of the quality of discharge letters. BMC Health Serv Res 2021 Apr 15;21(1):349 [FREE Full text] [doi: 10.1186/s12913-021-06345-z] [Medline: 33858383]
6. Cresswell A, Hart M, Suchanek O, Young T, Leaver L, Hibbs S. Mind the gap: improving discharge communication between secondary and primary care. BMJ Qual Improv Rep 2015 Sep 02;4(1):u207936.w3197 [FREE Full text] [doi: 10.1136/bmjquality.u207936.w3197] [Medline: 26734391]
7. Maslove DM, Leiter RE, Griesman J, Arnott C, Mourad O, Chow C, et al. Electronic versus dictated hospital discharge summaries: a randomized controlled trial. J Gen Intern Med 2009 Sep 16;24(9):995-1001 [FREE Full text] [doi: 10.1007/s11606-009-1053-2] [Medline: 19609623]
8. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 01;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
10. Human Biomedical Research Act. Ministry of Health Singapore. URL: https://www.moh.gov.sg/policies-and-legislation/human-biomedical-research-act [accessed 2024-07-19]
11. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. arXiv Preprint posted online on March 15, 2023 2024 [FREE Full text]
12. Bender E, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery; 2021:610-623.
13. Zhou C, He J, Ma X, Berg-Kirkpatrick T, Neubig G. Prompt consistency for zero-shot task generalization. arXiv Preprint posted online on April 29, 2022 2024. [doi: 10.18653/v1/2022.findings-emnlp.192]
14. Preiksaitis C, Sinsky CA, Rose C. ChatGPT is not the solution to physicians' documentation burden. Nat Med 2023 Jun;29(6):1296-1297. [doi: 10.1038/s41591-023-02341-4] [Medline: 37169865]
15. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]

Abbreviations

AI: artificial intelligence
EDS: electronic discharge summary
EMR: electronic medical record
LLM: large language model
PCP: primary care provider
TOCCC: Transitions of Care Consensus Conference

XSL•FO
RenderX

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 25.02.24; peer-reviewed by L Yong, G Tofler, CS Wu; comments to author 08.04.24; revised version received 18.04.24; accepted 04.07.24; published 24.07.24.

Please cite as:
Tung JYM, Gill SR, Sng GGR, Lim DYZ, Ke Y, Tan TF, Jin L, Elangovan K, Ong JCL, Abdullah HR, Ting DSW, Chong TW
Comparison of the Quality of Discharge Letters Written by Large Language Models and Junior Clinicians: Single-Blinded Study
J Med Internet Res 2024;26:e57721
URL: https://www.jmir.org/2024/1/e57721
doi:10.2196/57721
PMID:

©Joshua Yi Min Tung, Sunil Ravinder Gill, Gerald Gui Ren Sng, Daniel Yan Zheng Lim, Yuhe Ke, Ting Fang Tan, Liyuan Jin, Kabilan Elangovan, Jasmine Chiat Ling Ong, Hairil Rizal Abdullah, Daniel Shu Wei Ting, Tsung Wen Chong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness and User Experience of a Smoking Cessation Chatbot: Mixed Methods Study Comparing Motivational Interviewing and Confrontational Counseling

Linwei He[1], MSc; Erkan Basar[2], MSc; Emiel Krahmer[1], PhD; Reinout Wiers[3], PhD; Marjolijn Antheunis[1], PhD

[1]Department of Communication and Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, Netherlands

[2]Behavioral Science Institute, Radboud University, Nijmegen, Netherlands

[3]Addiction Development and Psychopathology (ADAPT)-lab, Department of Psychology and Centre for Urban Mental Health, University of Amsterdam, Amsterdam, Netherlands

Corresponding Author:
Linwei He, MSc
Department of Communication and Cognition
Tilburg School of Humanities and Digital Sciences
Tilburg University
Dante Building, D407
Warandelaan 2
Tilburg, 5037AB
Netherlands
Phone: 31 644911989
Email: l.he_1@tilburguniversity.edu

Abstract

Background: Cigarette smoking poses a major public health risk. Chatbots may serve as an accessible and useful tool to promote cessation due to their high accessibility and potential in facilitating long-term personalized interactions. To increase effectiveness and acceptability, there remains a need to identify and evaluate counseling strategies for these chatbots, an aspect that has not been comprehensively addressed in previous research.

Objective: This study aims to identify effective counseling strategies for such chatbots to support smoking cessation. In addition, we sought to gain insights into smokers' expectations of and experiences with the chatbot.

Methods: This mixed methods study incorporated a web-based experiment and semistructured interviews. Smokers (N=229) interacted with either a motivational interviewing (MI)–style (n=112, 48.9%) or a confrontational counseling–style (n=117, 51.1%) chatbot. Both cessation-related (ie, intention to quit and self-efficacy) and user experience–related outcomes (ie, engagement, therapeutic alliance, perceived empathy, and interaction satisfaction) were assessed. Semistructured interviews were conducted with 16 participants, 8 (50%) from each condition, and data were analyzed using thematic analysis.

Results: Results from a multivariate ANOVA showed that participants had a significantly higher overall rating for the MI (vs confrontational counseling) chatbot. Follow-up discriminant analysis revealed that the better perception of the MI chatbot was mostly explained by the user experience–related outcomes, with cessation-related outcomes playing a lesser role. Exploratory analyses indicated that smokers in both conditions reported increased intention to quit and self-efficacy after the chatbot interaction. Interview findings illustrated several constructs (eg, affective attitude and engagement) explaining people's previous expectations and timely and retrospective experience with the chatbot.

Conclusions: The results confirmed that chatbots are a promising tool in motivating smoking cessation and the use of MI can improve user experience. We did not find extra support for MI to motivate cessation and have discussed possible reasons. Smokers expressed both relational and instrumental needs in the quitting process. Implications for future research and practice are discussed.

(J Med Internet Res 2024;26:e53134) doi:10.2196/53134

KEYWORDS

chatbot; smoking cessation; counseling; motivational interviewing; confrontational counseling; user experience; engagement

XSL•FO
RenderX

Introduction

Background

Cigarette smoking is one of the most preventable causes of premature death and illness [1], and supporting smokers in their attempts to quit is a public health priority. While the use of cessation support (eg, pharmacological treatment and behavioral support) can effectively increase the successful quitting rate [2], such services are facing substantial staff and resource shortages [3,4], which have been aggravated by the COVID-19 pandemic. Moreover, people tend to express resistance to using professional support when it comes to quitting smoking [5]. To combat these challenges, digital tools that can complement traditional cessation services have become increasingly popular. Conversational agents, or chatbots (ie, computer programs that use artificial intelligence to simulate a conversation with users through natural language [6]), are one of the recent innovations that bring several benefits: accessibility, scalability, personalization, and synchronicity, to name a few [7]. These opportunities have sparked interest to test the use of chatbots in health care settings, and initial effectiveness has been found in various domains, such as healthy lifestyle promotion [8], sexual health [9], mental health [10], and smoking cessation [11]. However, it is still unclear when and how chatbots are acceptable and effective. Some studies in human-computer interaction have shown that the positive effect of chatbots tends to decrease over time [12], and a few barriers that hinder human-chatbot interaction have been identified, including a lack of engagement caused by counseling styles that users do not prefer [13]. As long-term support is essential for smoking cessation, which usually requires sustained effort, and in view of the one-to-one nature of chatbots that allows for individualized conversation, it is important to find the optimal counseling style to overcome potential resistance to effectively motivate smoking cessation.

Within the field of individualized smoking cessation counseling, a chatbot can use various counseling styles, such as confrontational counseling (CC) and motivational interviewing (MI). CC originated in alcoholism treatment and emphasizes a hard-hitting and directive style intended to break through patients' defense mechanisms [14]. CC counselors are trained to confront patients with the consequences of their unhealthy behaviors to counter self-exempting beliefs and use direct, unsolicited advice to increase patients' risk perception [15]. In the field of smoking cessation, CC techniques include direct education about the risks associated with smoking, challenging smokers' minimization or denial of the problem, and urging for abstinence by providing a quitting plan [16-18]. A typical example situation is the counselor coercing the client to *face up to reality* by emphasizing their problematic smoking behavior and the associated risks [14]. However, this does not necessarily entail an aggressive approach but rather attempts to raise clients' awareness of the likely severe consequences of their smoking behavior [19]. Indeed, contrary to the common assumption that confrontation leads to denial [15], CC has been found to be predictive of higher patient involvement and higher long-term abstinence [20,21]. For several decades, CC has been regarded as the only way to get patients to listen [14], and there is positive evidence of health care providers using directive confrontational tactics to advise smokers to quit [22]. However, recent research suggests that the effectiveness and acceptability of CC depends on the counselor's legitimacy and a rather empathetic manner [23], whereas resistance may escalate if the client feels threatened by the confrontational tone [24].

In response to the potential resistance, an alternative approach has been developed where patients' autonomy and intrinsic motivation are highlighted [14,25]. Miller and Rollnick [25] introduced MI, a nonconfrontational counseling style for eliciting behavior change by helping clients explore and resolve ambivalence. The strategies of MI are more subtle than coercive, prioritizing encouragement over contention, with the ultimate goal of eliciting the client's intrinsic motivation rather than imposing or desiring behavior change [25,26]. MI counselors use the principles of expressing empathy, avoiding arguing and confrontation, and supporting the individual's self-efficacy. A typical example is "What you choose to do is up to you. My role is to help you figure out what you want to do and support you in accomplishing your goals" [27]. In the context of smoking cessation, MI counselors use techniques such as asking open-ended questions to elicit personally relevant reasons to quit, reflecting on the client's words to encourage disclosure, affirming the client's expressions toward quitting, and summarizing to help the client have a better overview of their thoughts [25,28]. The MI counselor assists the client in realizing the discrepancy between their values and their behaviors, thereby eliciting intrinsic motivation to resolve this discrepancy by changing their behavior [29,30].

Over the past decades, a rich amount of research on the effects of MI and CC on smoking cessation and other health behaviors has spawned, yielding mixed and inconclusive results. While CC has been found effective in increasing smokers' risk perceptions and, in some cases, led to smoking cessation in the long run [20,31], there is also evidence suggesting that CC may lead to defense mechanisms among the clients (eg, denial and dismissal of the information) [32]. Similar inconclusiveness has been observed regarding the effectiveness of MI; while successful with other behavior domains (eg, alcohol use and physical exercise), the treatment tends to have weaker effects on smoking cessation [33,34]. Despite the mixed evidence regarding the 2 styles, modern addiction treatment research believes that MI is more likely to be successful than CC [14]. In general, MI-style interventions have been found to motivate a larger proportion of individuals than CC-style interventions, making them a more applicable and efficacious approach [14,35]. Moreover, the effectiveness of CC depends on the legitimacy and authority of the counselor [36]. In the context of chatbot-delivered interventions, CC might not be as effective as chatbots are usually not regarded as a medical authority [37]. On a broader note, it has been suggested that people prefer chatbots that offer emotional support (ie, MI) more than chatbots that predominantly offer informational support (ie, CC) [38]. In summary, MI seems to be a better approach for a smoking cessation chatbot.

XSL•FO
RenderX

Objectives

However, it remains unclear whether the hypothesized advantage of MI over CC can be experimentally sustained in a chatbot setting. To date, only a few studies in human-delivered counseling have directly compared the 2 styles and found mixed results such that MI (vs CC) did not significantly reduce drinking and drug use but yielded less resistance and higher client satisfaction [14,18]. Moreover, no study has explicitly compared the 2 styles in a chatbot setting, and it is yet unknown whether the effect of MI is also applicable in automated settings. Human-computer interaction research in health care has emphasized the chatbots' relational strategies to improve their effectiveness and acceptability [39,40], which are keenly matched to the principles of MI [25]. However, despite the theoretical advantages, the empirical evidence of automated MI is inconclusive [41]. Such insignificance of MI could be explained by the methodological disparities in the line of research: the comparison groups varied considerably, and the robustness of MI cannot be concluded. In a recent review of technology-delivered MI interventions [41], various studies compared MI with other types of interventions (eg, self-help booklet and assessment only) but not with an alternative counseling style (eg, CC). Therefore, the effectiveness of MI as a chatbot counseling style cannot be ascertained. In fact, regardless of the therapeutic counseling style, people might find a chatbot acceptable and even persuasive when presented nonintrusively [42], which, therefore, shadows the effect of chatbot counseling styles. Indeed, in an earlier study [43], an MI-style chatbot was compared with a neutral-style chatbot in motivating smokers to quit, expecting MI to generate a higher intention to quit and better therapeutic experience among smokers. However, no significant difference between the MI style and the neutral style was found. To follow up on previous research and further understand the role of the chatbot and better ascertain the usefulness of MI as a chatbot counseling style, this study compared the MI and CC therapeutic counseling styles in a chatbot setting. To sum up, this study hypothesized that an MI-style chatbot, compared to a CC-style chatbot, results in higher self-efficacy, stronger motivation to quit, more engagement, a stronger therapeutic alliance, more perceived empathy, and higher interaction satisfaction.

Moreover, to have a more comprehensive picture of whether and how an MI (vs CC) chatbot works and understand how users experience the chatbot interaction, we explored users' perspectives on the chatbots through semistructured interviews. Specifically, we examined the following aspects: (1) What is the overall user experience with the smoking cessation chatbots? (2) How does the chatbot influence smokers' perceptions of smoking and intention to quit? (3) How can we improve the chatbots for future use?

Methods

Study Design

This mixed methods study combined a quantitative part using a web-based experiment and a qualitative part of semistructured interviews. For the experiment, participants were randomly assigned to interact with a chatbot using an MI style or a chatbot in a CC style. Among a smaller set of participants, semistructured interviews were conducted after the chatbot interaction to gain further insights into user experience.

Ethical Considerations

Ethics approval was obtained from the Research Ethics and Data Management Committee of the Tilburg School of Humanities and Digital Sciences (identification code REDC 2021.18ab), and the study was conducted in compliance with the ethical and data management regulations of the school. Informed consent was obtained from participants via the Qualtrics (Qualtrics International Inc) web form. Participants were recruited via the participant pools from the Tilburg School of Humanities and Digital Sciences and Tilburg School of Social and Behavioral Sciences and were compensated with credits. The study design, raw materials, and analysis plan were preregistered at the Open Science Framework.

Participants and Procedure

Power calculations were conducted for multivariate ANOVA (MANOVA) using G*Power (version 3.1) [44], indicating that a sample size of 226 participants is adequate to uncover medium effects (effect size: f^2=0.0625; power=0.8) in accordance with previous meta-analyses on the effects of MI on smoking cessation [45]. Eligible participants were smokers (ie, had smoked at least one cigarette in the week before participation) aged ≥18 years with a competent proficiency in English reading and writing.

In the CONSORT (Consolidated Standards of Reporting Trials) flow diagram in Figure 1, the study procedure is visualized. Upon starting the study, participants completed the pretest questionnaire assessing demographics and baseline motivation to quit smoking, after which they were randomly assigned to interact with either the MI or the CC chatbot for 2 consecutive sessions, simulating a typical intake session and a first consultation session in smoking cessation interventions [46,47]. The randomization was double blinded and was carried out automatically by the Qualtrics software. In the first session, the chatbots carried out a conversational assessment of the participants' smoking behavior and initial motivation and barriers to quitting smoking. After approximately 5 to 10 minutes, which resembles the waiting room setting, the second session started, in which the chatbot discussed with participants their previous quit attempts, aimed to strengthen their self-efficacy, and encouraged the participants to form future quit plans. Relevant outcome measures were assessed after each session. Upon completion, all participants were debriefed. In addition, interview invitations were sent out at random to participants in both conditions, and the final interview sample consisted of 16 participants, 8 (50%) from each condition.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. CC: confrontational counseling; MI: motivational interviewing.

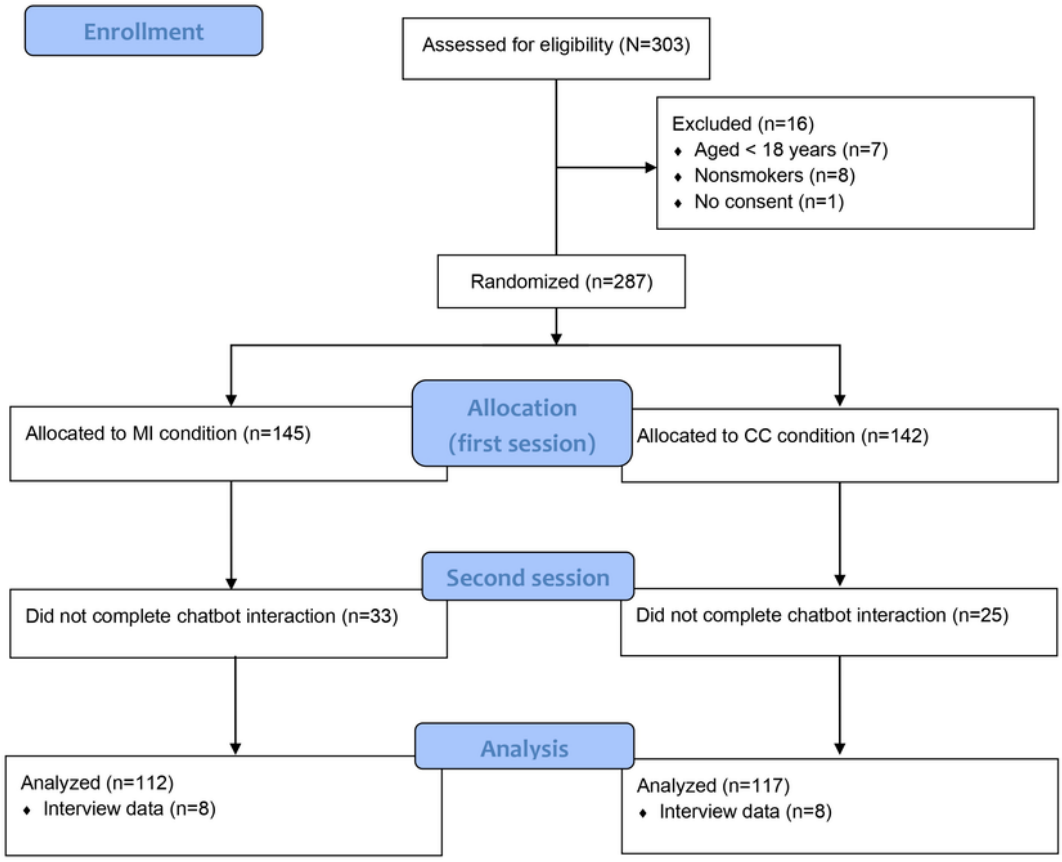

The Chatbot Interaction

Overview

The 2 chatbots (MI and CC) were both named Roby and were operated on the Rocket.Chat web interface. They were equipped with a natural language understanding module and a response generation module, where the former was trained using conversational data from an earlier study on an MI chatbot for smoking cessation [43]. A set of human-authored response utterances was devised by the author team and reviewed by a clinical psychology expert experienced in addiction treatment to ensure appropriateness. The chatbot scripts adhered to the Dutch national guideline for smoking cessation in primary care [48] except for the offering of pharmacological and behavioral support, in which case the chatbot referred participants to a public health website [49] for further information. The chatbot system recorded user data and conversation history to understand the "context" and selected the best-fitting utterance for the user at each turn based on the conversation context. A detailed description of the technical infrastructure is provided elsewhere [50]. The first session took approximately 5 to 7 minutes, and the second session took approximately 8 to 10 minutes.

The MI Chatbot Condition

The practice of MI involves a relational component and a technical component. The relational component is expressed through partnership, acceptance, compassion, and evocation [25]. The technical component includes skills such as asking open questions, reflecting on client input, affirming, summarizing, and asking for permission before providing information [14,21].

In the first session, the chatbot introduced itself and shared the agenda of the session, after which it asked about the participants' smoking behavior and potential motivation to quit. The chatbot reflected on the participants' input to show active attention and empathy. For example, when a participant indicated concerns about coloring teeth from smoking, the chatbot responded with the following: "You want to look and smell good, and you're seeing that smoking might impact you on that." The chatbot provided personalized normative feedback (ie, the percentage of smokers in their age group) after asking for permission, asked open questions, and reflected on participants' answers. At the end of the first session, the chatbot summarized their conversation, presenting participants' own thoughts on quitting, aiming to strengthen intrinsic motivation to quit smoking. In the second session, the chatbot asked whether the participant had made a quit attempt before and invited the participant to reflect on the experience. The chatbot discussed with participants their reasons to quit and their earlier approaches to quitting and asked the participants to think of personal strengths that helped them in their previous quit attempts. If the participant had not attempted to quit before, the chatbot asked for another challenging experience they had accomplished, aiming to elicit personal strengths and self-efficacy. Following the discussion, the chatbot summarized the conversation, let the participant review their strengths and experience, encouraged the participant to form a plan and a date for the next quit attempt, or emphasized autonomy when the participant was not ready to

XSL•FO
RenderX

quit. To ensure fidelity, the chatbot dialogues were adapted from communication strategies that have been extensively tested and verified in human-delivered counseling practice [25,28].

The CC Chatbot Condition

Therapeutic confrontation involves a process in which the counselor provides unsolicited, direct, reality-oriented, and confrontational feedback regarding the clients' behaviors and thoughts [51]. The main aim is for the client to realize the issues that their smoking behavior is causing. CC counselors work through identifying certain conceptions about smoking such as self-exempting beliefs [16], raising risk perception, and helping the clients realize the issue that their smoking behavior is causing.

In the first session, the chatbot asked about the participants' smoking behavior while expressing concerns about their smoking severity [14,52]. While the MI chatbot elicited participants' own motivation to quit, the CC chatbot provided factual information to confront the minimization of the issue or resistance and urge for readiness [18]. In the second session, the chatbot discussed with the participants their previous quit attempts, reminded them of their reasons to quit, and restated the risks of smoking. If the participant had not had quit attempts before, the chatbot confronted the participants' beliefs for not quitting. An example of a typical self-exempting thought of a smoker would be the following: "I only smoke in social situations, so I don't think it's so bad for my health." The chatbot then provided unsolicited factual information to challenge the smoker's beliefs. After the discussion, the chatbot provided direct advice on quitting approach and asked for a plan and date for the next quit attempt. Except for the aforementioned manipulations, the content of the dialogues (eg, questions, topics, and information provided) and the length of the interaction were as similar as possible across the conditions. Figure 2 shows an example of the final conversation. The complete chatbot scripts can be found in the Open Science Framework preregistration.

Figure 2. An illustration of the chat interface of the motivational interviewing condition.

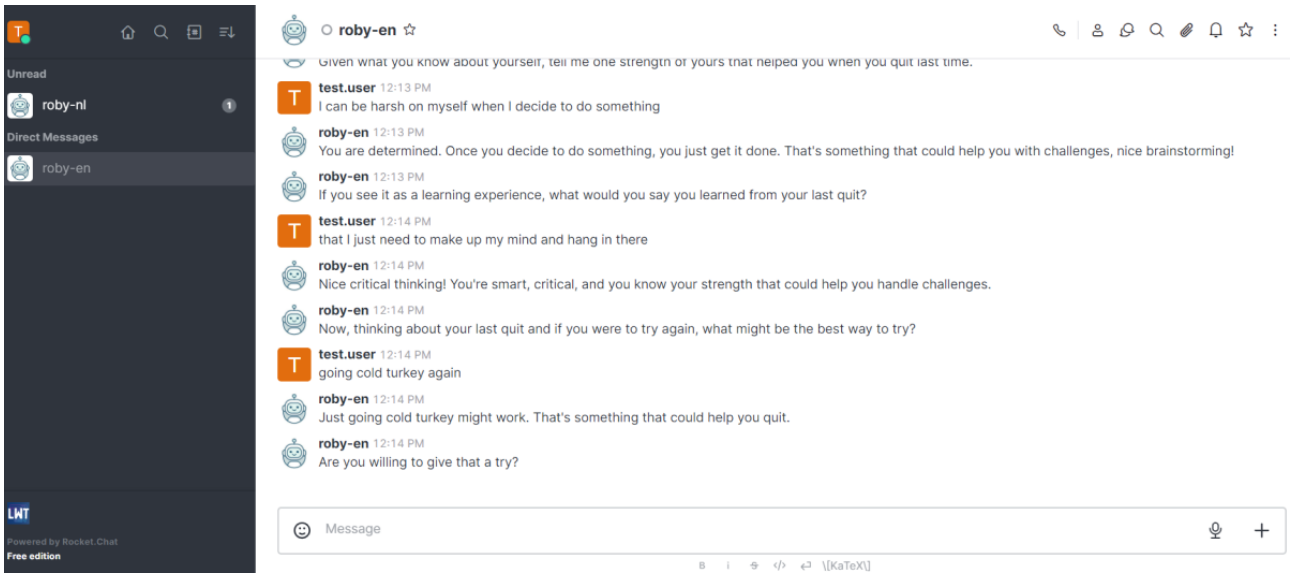

Measures

Demographics and smoking-related baseline information were collected at the pretest time point. Age and gender were assessed using single items. Baseline motivation to quit was measured using the Contemplation Ladder [53], where participants indicated their motivational status ranging from 0 (*no thoughts on quitting*) to 10 (*taking action to quit*). Level of nicotine dependence was measured using the Fagerström Test for Nicotine Dependence [54], and average daily cigarette consumption was measured using a single item.

For outcome measures, we assessed both cessation-related variables and user experience–related variables, and all variables were measured both after the first session (T1) and after the second session (T2). All scale variables were measured using 5-point scales. The measurement scheme followed that of the earlier study [43] to facilitate comparison.

Intention to quit was measured using the Contemplation Ladder [53]. Participants indicated where they identified themselves on a single item, with responses ranging from 0 (*no thoughts on quitting*) to 10 (*taking action to quit*).

Self-efficacy was assessed using the Smoking Self-Efficacy Questionnaire [55]. On a 5-point scale, participants indicated whether they thought they would be able to refrain from smoking in various difficult situations. An example question is as follows: "how sure are you that you could refrain from smoking when you feel nervous."

Engagement with the chatbot interaction was assessed using 9 items from subscales of the short form of the User Engagement Scale [56]. The esthetic appeal subscale was removed as this study focused on the communication process instead of the interface design. An example item is as follows: "I was absorbed in this experience." An additional question was included asking about participants' endorsement for future use.

Therapeutic alliance was measured using the Working Alliance Inventory–Short Revised [57], a 12-item self-report measure used to assess the relationship between participants and the

XSL•FO
RenderX

chatbot. An example item is as follows: “We agree on what is important for me to work on.”

Perceived empathy was measured using a 3-item (eg, Roby seemed to understand me) scale based on research on interpersonal communication by Rubin and Martin [58].

Interaction satisfaction was measured using 5 items on a 5-point scale ranging from “Not at all” to “Very much” following studies on telephone smoking cessation counseling [59]. An example item is as follows: “how was your counsellor in terms of being a good listener?”

As a manipulation check question, perception of MI was measured using an 8-item scale based on the Client Evaluation of Motivational Interviewing scale [60]. Example items are “Roby argued with you to change your behavior” and “Roby helped you feel confident in your ability to change your behavior.”

Statistical Analysis

Experiment data were analyzed using R (version 4.2.1; R Foundation for Statistical Computing). Independent-sample 2-tailed *t* tests and a chi-square test were conducted to check for equal distribution of background variables (ie, gender, age, nicotine dependence level, and baseline motivation to quit) across conditions. Variables that were not equally distributed across conditions were included as covariates in hypothesis testing.

To test the preregistered hypothesis regarding the effect of MI chatbot conversations on the outcomes, a 1-way MANOVA was performed using the 6 outcome variables (ie, self-efficacy, intention to quit, engagement, therapeutic alliance, perceived empathy, and interaction satisfaction). Data measured at T2 were used for this hypothesis.

Qualitative Data Collection and Analysis

Qualitative data were gathered at the postmeasurement time point through semistructured video interviews with 16 participants, 8 (50%) each randomly selected from the 2 conditions. The interviews lasted 20 to 30 minutes each and were conducted by the first author, who was familiar with the chatbot content. Before study commencement, an interview guide was developed by the author team and pilot-tested. The interview topics were guided by the theoretical framework of acceptability of health care interventions [61], which consists of several constructs (eg, affective attitude, burden, intervention coherence, and perceived effectiveness) that closely relate to user experience with health care interventions.

The interviews were audio recorded and transcribed. The transcripts were analyzed using thematic analysis [62]. We adopted a hybrid approach combining deductive and inductive coding [63] in which coding was mainly informed by themes that could potentially answer the research questions and inform chatbot improvement but also allowed for novel codes to emerge for additional insights. A pilot coding round was carried out by 2 independent reviewers, and consensus regarding the codes was reached through discussion. Coding was performed in an iterative process over multiple reads of the transcripts. Multiple group meetings were organized aiming to identify potential biases in the coding and interpretation process. One author (LH) carried out the coding of the remaining interviews. The analysis procedure was performed using the ATLAS.ti (ATLAS.ti Scientific Software Development GmbH) program.

Results

Quantitative Results

Sample Characteristics and Descriptive Results

A total of 303 participants were recruited, with 287 (94.7%) being eligible for participation. Among the eligible participants, 50.5% (145/287) were allocated to interact with the MI chatbot, 49.5% (142/287) were allocated to the CC chatbot, and 20.2% (58/287) were excluded from the analysis for not completing the chatbot interaction and questionnaire, resulting in a final analyzed sample of 229 participants. Of these 229 participants, 153 (66.8%) were female, and the average age of the participants was 21.1 (SD 2.75) years. The average Fagerström Test for Nicotine Dependence score was 1.17 (SD 1.71), indicating overall low dependence on nicotine. The 2 groups did not differ significantly on any of the background variables at the pretreatment time point. Table 1 shows the sample characteristics and the *P* values.

All measures used demonstrated high reliability. Table 2 shows the reliability and descriptive results for the outcome variables.

XSL•FO
RenderX

Table 1. Demographic and tobacco use characteristics of participants at baseline by condition (n=229).

Characteristic	Total	MI[a] (n=112)	CC[b] (n=117)	*P* value
Age (y), mean (SD)	21.06 (2.75)	21.17 (3.04)	20.95 (2.44)	.55
Sex , n (%)				.51
Female	153 (66.8)	71 (63.4)	82 (70.1)	
Male	73 (31.9)	39 (34.8)	34 (29.1)	
Other	3 (1.3)	2 (1.8)	1 (0.9)	
Daily cigarette consumption, mean (SD)	4.90 (5.01)	5.06 (4.95)	4.74 (5.08)	.64
FTND[c] score, mean (SD)	1.17 (1.71)	1.19 (1.73)	1.15 (1.71)	.88
Score of baseline motivation to quit, mean (SD)	5.33 (2.82)	5.31 (2.84)	5.34 (2.81)	.94

[a]MI: motivational interviewing.
[b]CC: confrontational counseling.
[c]FTND: Fagerström Test for Nicotine Dependence.

Table 2. Descriptive and reliability results of the outcomes.

	T1 (first session)			T2 (second session)		
	MI[a], mean (SD)	CC[b], mean (SD)	Cronbach α	MI, mean (SD)	CC, mean (SD)	Cronbach α
Intention to quit	5.71 (2.80)	5.69 (2.81)	__[c]	6.26 (2.85)	5.91 (2.85)	—
Self-efficacy	2.62 (0.76)	2.79 (0.88)	0.85	2.74 (0.76)	2.82 (0.92)	0.87
Engagement	3.41 (0.67)	3.38 (0.66)	0.83	3.40 (0.73)	3.14 (0.77)	0.87
Therapeutic alliance	3.47 (0.65)	3.18 (0.75)	0.90	3.61 (0.67)	3.22 (0.85)	0.92
Perceived empathy	3.51 (0.84)	3.00 (0.92)	0.81	3.59 (0.92)	2.94 (1.10)	0.91
Interaction satisfaction	3.81 (0.67)	3.50 (0.77)	0.73	3.85 (0.78)	3.41 (0.80)	0.78

[a]MI: motivational interviewing.
[b]CC: confrontational counseling.
[c]Not applicable.

Main Effects of MI

We hypothesized that the MI-style chatbot would lead to better outcomes in terms of both motivating cessation and user experience. We performed a manipulation check on participants' perception of MI using the Client Evaluation of Motivational Interviewing scale [60], which showed that participants in the MI condition perceived the chatbot as significantly more MI-like (mean score 3.77, SD 0.62) than participants in the CC condition (mean score 3.21, SD 0.60; t_{227}=6.89; *P*<.001), representing a large effect size (*d*=0.91). Thus, the manipulation was deemed successful. For this preregistered hypothesis, the 2 sessions were regarded as 1 interaction, and we used data measured after the entire interaction (ie, at T2). Results of a subsequent MANOVA using T2 data revealed significant overall differences between the 2 groups (Pillai trace=0.12; $F_{6,222}$=5.00; *P*<.001; η^2=0.12). We performed a follow-up discriminant analysis to examine which outcomes contributed the most to the overall group differences [64]. Results showed that user experience–related outcomes (ie, engagement, perceived empathy, and interaction satisfaction) had the highest correlation coefficients with the discriminant function, indicating that they mostly differentiated between the MI and CC conditions, whereas cessation-related outcomes (ie, intention to quit and self-efficacy) practically did not differ between the 2 groups. These results are presented in Table 3. Overall, the results of the MANOVA and discriminant analysis suggest that chatbot counseling style (MI vs CC) has a significant effect on the outcome variables and that the effect can be mostly explained by differences in user experience.

XSL•FO
RenderX

Table 3. Results of discriminant function analysis of the outcome variables by condition.

Variable	Discriminant function	
	Correlation coefficients with discriminant function	Standardized coefficient with discriminant function
Intention to quit	0.00	0.01
Self-efficacy	–0.19	–0.16
Engagement	*–0.80*[a]	*–0.60*
Therapeutic alliance	0.23	0.18
Perceived empathy	*0.69*	*0.71*
Interaction satisfaction	*0.85*	*0.67*

[a]Italics indicate variables that mostly discriminated the 2 conditions.

Exploratory Analyses

As the chatbot interaction involved 2 sessions, we explored potential changes in participants' experience over the 2 sessions (ie, the multi-session effect). We performed a mixed ANOVA for each of the outcome variables, including condition (MI vs CC) as a factor and time (T1 vs T2) as a repeated measure. Test statistics can be found in Table 4, and means and SDs can be found in Table 2. For intention to quit, there was a significant effect of time ($F_{1,227}$=39.02; P<.001). Participants' intention to quit increased after the interaction, and such increase was significantly more profound in the MI condition ($F_{1,227}$=7.43; P=.007). For *self-efficacy*, an increase was found over time ($F_{1,227}$=6.89; P=.009), and the increase was only significant in the MI condition ($F_{1,227}$=4.34; P=.04). For *engagement*, participants had a significant overall decrease in their engagement with the chatbot ($F_{1,227}$=10.07; P=.002). A significant interaction effect between time and condition was also found, and follow-up simple-effect analysis per condition showed that the CC condition significantly contributed to the overall decrease ($F_{1,227}$=8.55; P=.004). For the remaining 3 user experience–related outcomes, the MI chatbot was rated significantly better in terms of *therapeutic alliance* ($F_{1,227}$=13.61; P<.001), *perceived empathy* ($F_{1,227}$=25.90; P<.001), and *interaction satisfaction* ($F_{1,227}$=18.14; P<.001). Finally, a significant overall increase over time in therapeutic alliance was found in both conditions ($F_{1,227}$=6.64; P=.01). The results are visualized in Figure 3.

Table 4. Results of repeated-measure ANOVA on the outcomes.

	Condition		Time		Condition × time[a]		Simple-effect post hoc analysis of the interaction effect			
							Effect of time in MI[b]		Effect of time in CC[c]	
	F test (*df*)	*P* value	*F* test (*df*)	*P* value	*F* test (*df*)	*P* value	*F* test (*df*)	*P* value[d]	*F* test (*df*)	*P* value
Intention to quit	0.26 (1, 227)	.61	39.02 (1, 227)	<.001	7.43 (1, 227)	.007	31.30 (1, 111)	<.001	8.41 (1, 116)	.008
Self-efficacy	1.14 (1, 227)	.29	6.89 (1, 227)	.009	4.34 (1, 227)	.04	8.03 (1, 111)	.01	0.23 (1, 116)	>.99
Engagement	2.74 (1, 227)	.10	10.07 (1, 227)	.002	8.55 (1, 227)	.004	0.03 (1, 111)	>.99	17.2 (1, 116)	<.001
Therapeutic alliance	13.61 (1, 227)	<.001	6.64 (1, 227)	.01	2.12 (1, 227)	.15	__[e]	—	—	—
Perceived empathy	25.90 (1, 227)	<.001	0.03 (1, 227)	.87	1.54 (1, 227)	.22	—	—	—	—
Interaction satisfaction	18.14 (1, 227)	<.001	0.24 (1, 227)	.63	2.28 (1, 227)	.13	—	—	—	—

[a]The interaction effect between time and condition.

[b]MI: motivational interviewing.

[c]CC: confrontational counseling.

[d]Bonferroni-adjusted *P* value.

[e]Not applicable.

XSL•FO
RenderX

Figure 3. Changes over time in individual tracings and group means for the outcome variables. CC: confrontational counseling; MI: motivational interviewing.

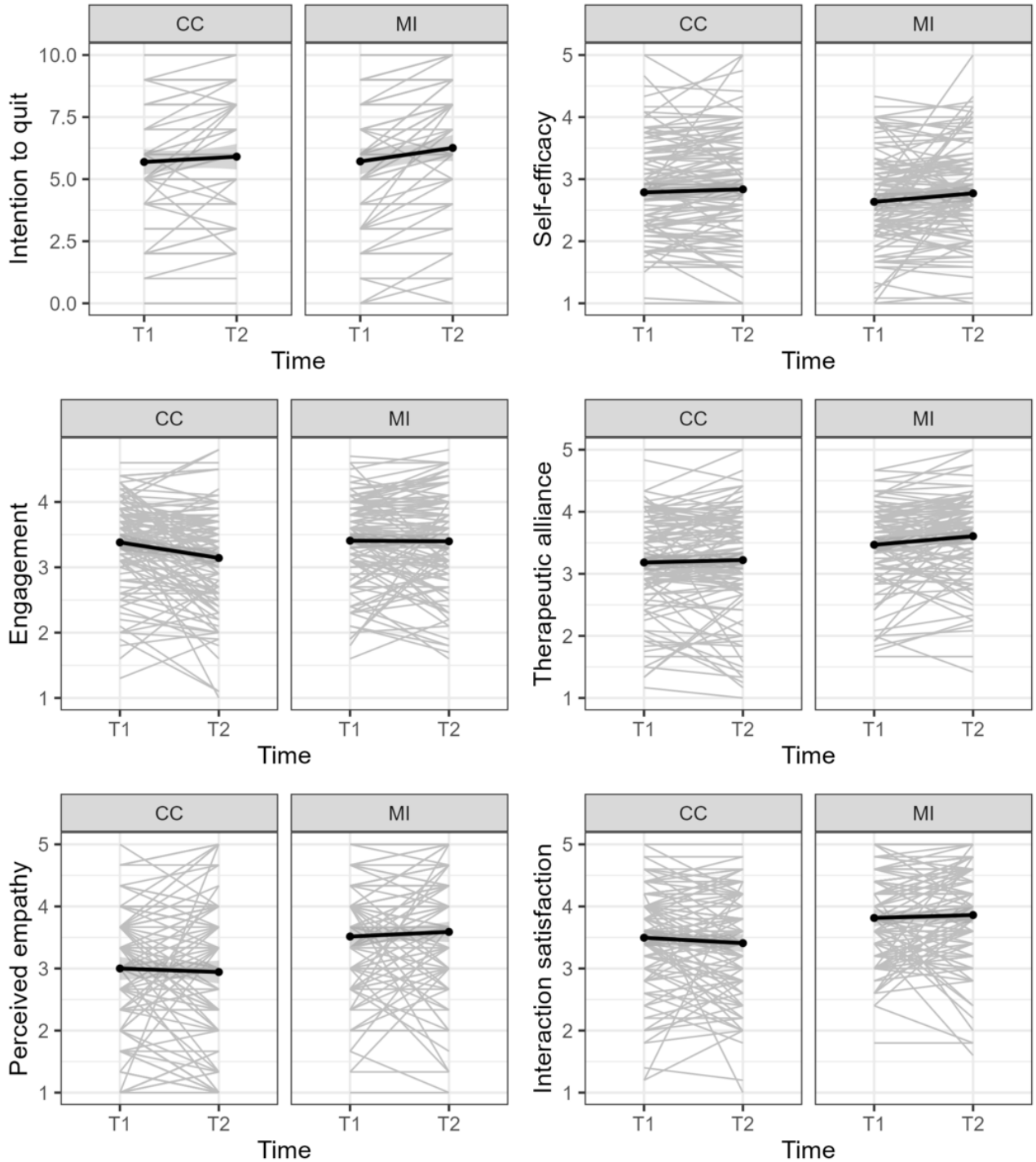

Qualitative Results

Overview

The qualitative data provided insights into the user experience with the chatbots, participants' perceptions of the motivating aspect of the chatbots, and suggestions for future design. Moreover, the results supported and expanded on the experimental comparison between the MI- and CC-style chatbots. The findings centered on 3 main themes with accompanying subthemes. The themes were *prior expectation*, with the subthemes *authoritative confrontation* and *generic interaction*; *timely experience*, with the subthemes *affective attitude*, *engagement*, and *usability*; and *postinteraction reflection*, with the subthemes *perceived effectiveness*, *perception of health chatbots*, and *ideal intervention*. Table 5 presents the themes and illustrative quotes.

XSL•FO
RenderX

Table 5. Illustrative quotes for the qualitative findings.

Main theme and subtheme	Participant quotes
Prior expectation	
Authoritative confrontation	• "I have the expectation that it was going to be some sort of, gaslighting situation, you need to stop smoking because this and this." [Participant 5; female, aged 19 years; MIa condition]
Generic interaction	• "I thought it was just going to be generic for everybody the same, not regarding the answers the person gave. I didn't think it was going to be personalized." [Participant 2; female, aged 22 years; CC[b] condition]
Timely experience	
Affective attitude	• "I had fun chatting with the chatbot. And in my opinion, the chatbot was really friendly and understanding." [Participant 10; female, aged 18 years; CC condition] • "It didn't feel like I was being put into a position where I can feel guilty about my smoking habits. He actually made sure that I felt safe." [Participant 5; female, aged 19 years; CC condition]
Engagement	• "In the conversation, I lost my engagement for a moment in the situation with finding correct answers that the chatbot will understand." [Participant 7; female, aged 24 years; MI condition] • "It was just enough to keep the conversation going but not extremely involved." [Participant 3; female, aged 19 years; CC condition] • "I found it really interesting because I couldn't foresee what this conversation would bring to me, so it's always nice to keep engaged in the conversation and see what will come out of it." [Participant 10; female, aged 18 years; CC condition] • "I knew it was a chatbot and everything, so if it was a person, I would be more engaged." [Participant 2; female, aged 22 years; CC condition]
Usability	• "It's easier to make a chatbot than to find a therapist who would guide you through the process, and I also think it's good to have a computer because you can always text him at any time of the day." [Participant 13; female, aged 19 years; MI condition] • "The amount of information was a bit, not really enough, I feel like it could have given me more information about certain things." [Participant 3; female, aged 19 years; CC condition]
Postinteraction reflection	
Perceived effectiveness	• "I really like the question 'what did you feel when you stopped smoking?' I really asked myself how I felt, why did I stop. And it was really good, looking at that time, it was good for me to see that I felt good, maybe I should do it again." [Participant 13; female, aged 19 years; MI condition] • "I'm actively thinking about it [quitting] and actively attempting at sometimes, but it didn't bring out any new techniques or methods or reasons to." [Participant 4; male, aged 29 years; MI condition]
Perception of health chatbots	• "As long as I know that it would be destroyed after it's not needed anymore, I'd be fine with it. I think maybe it's even easier to tell your feelings to a bot, which you know is programmed to not be judgmental and just want to help, and also it'll be gone." [Participant 9; male, aged 21 years; CC condition] • "It's quite scary because technology is evolving so much that you could actually have a conversation with the bot about something so subjective and personal." [Participant 11; female, aged 18 years; MI condition] • "I wasn't fully comfortable because like I said in the back of my mind it was a computer, and I thought it was not going to understand my emotions, so I didn't really feel like telling them." [Participant 13; female, aged 19 years; MI condition]
Ideal intervention	• "I'm interested in people's thoughts and why they do things, like what's the inner reason for them to do that. Because it's the most effective way to find a solution. Just need the why's. If a bot could that, it would be better." [Participant 6; female, aged 21 years; CC condition] • "Because you gave the bot all that information and it's stored somewhere, you can also use that information to make [progress] charts available for the user." [Participant 4; male, aged 29 years; MI condition]

[a]MI: motivational interviewing.

[b]CC: confrontational counseling.

Theme 1: Prior Expectation

Almost all participants had some expectations regarding the chatbot interaction when they signed up for participation. Most people referred to their previous experience with other chatbots as baseline expectations and expected nonpersonalized *generic interaction*. Customer service chatbots were the most frequently mentioned comparison, and participants were expecting similar general task-oriented conversations. One participant mentioned the following:

> *I just thought it would be a conversation the same as when you text customer service or something. And then they have a few answers there and they send it*

XSL•FO
RenderX

> *to you no matter what question you asked.* [Participant 12; female, aged 20 years; MI condition]

In terms of conversation content, many participants expected to passively receive persuasive information, such as risks of smoking and benefits of quitting. The chatbot was imagined as an *authoritative* figure in a formal setting (eg, a physician) giving advice. Most people expressed positive expectations regarding factual information and guidance, which could be motivating to quit. A few other participants mentioned that the anticipated confrontation from an authoritative figure could be intimidating and induce resistance at the beginning. Most participants later noted that the chatbot's nonjudgmental, supportive, and personalized approach was much better than expected. One participant put it as follows:

> *I was kind of, I wouldn't say scared, but I wasn't really willing to start the conversation because I was like "Oh, this is going to be, you have to quit, you should not smoke." But then I was surprised because it wasn't like that.* [Participant 1; female, aged 25 years; CC condition]

Theme 2: Timely Experience

The second theme was related to people's timely experience during the chatbot interaction. Participants had an overall positive *affective attitude* toward the interaction, and it was largely driven by the hedonic and therapeutic experience. Almost all participants found the chatbot interaction fun and enjoyable, attributing it to the friendly and knowledgeable feeling of the chatbot. Participants appreciated the safe space that the chatbot created for them to freely discuss their smoking behavior without being judged. Many participants expressed that they felt respected and supported by the chatbot, and they acknowledged that such a safe space is important for counseling and reducing resistance. However, a small number of participants from the CC condition mentioned that they felt coerced and confronted. One participant mentioned the following (more illustrative quotes can be found in Table 5):

> *He was actually pretty understanding, he wasn't judging. And I think that was the best part of it because I know what it's like to be judged by people and it felt really good not to. He was actually being helpful without being judgmental.* [Participant 5; female, aged 19 years; MI condition]

Engagement with the interaction was frequently discussed by participants both positively and negatively and was largely driven by people's technical experience. When the interaction went smoothly without many technical errors (eg, chatbot misunderstanding and asking for rephrasing), people felt engaged with the interaction, and disengagement occurred when the conversation was interrupted by technical glitches. Despite being engaged in the conversation, a few participants mentioned that their engagement level was limited to task completion (ie, completing their research participation) and they were not emotionally involved. Supporting the experimental findings, more interviewees in the MI condition mentioned that they felt engaged with the chatbot compared to participants in the CC condition. In addition, some people compared chatbot interaction with human-human interaction and generally preferred engaging with humans for more flexibility and true understanding. One participant stated the following:

> *A human can understand whatever you say, a chatbot sometimes has issues processing information. It's just for me, I'm more emotionally involved when talking to a person than talking to a computer.* [Participant 13; female, aged 19 years; MI condition]

In terms of *usability*, almost all participants felt that the chatbot was easy to use, and some participants extrapolated such comments to health chatbots in general, acknowledging the highly accessible benefit of chatbots. While some participants thought that the chatbot provided the right amount of information, which was easy to process and not overwhelming, more people felt that the amount of information was not sufficient to motivate quitting in and of itself, and they emphasized that more less known information is needed. One participant suggested the following:

> *The fact that most people don't smoke and that 1/3 of smokers decided to quit was a new fact for me, and it affected me. But I needed some more facts, so the quantity is, I think, not enough. It could give more things.* [Participant 6; female, aged 21 years; CC condition]

Theme 3: Postinteraction Reflection

The third theme captured people's reflection on how the chatbot influenced their motivation and suggestions for improvement. Most participants discussed their *perceived effectiveness* of the chatbot interaction in a positive way, which parallels the experiment finding that both chatbots led to an increased intention to quit. Many people acknowledged that the conversation encouraged them to reflect on their behavior and brought them new perspectives on their thoughts about smoking. Novel factual information (eg, number of smokers at a certain age) was another highly appreciated component of the interaction, identified as a motivating factor by almost all participants. However, a few people explicitly expressed that the chatbot did not motivate them. Part of this ineffectiveness stemmed from participants' low readiness to quit, whereas a number of people perceived the conversation as reiterating already known information (eg, health benefits of quitting). Substantial differences were found between interviewees from the 2 experimental conditions such that it was mentioned more in the MI condition that people appreciated the chatbot encouraging self-reflection and more participants in the CC condition found the factual information useful:

> *I think it touched my soul a little bit. I tried to think about my habits and what I don't like. And I think that considering what I don't like helped me the most to realize even more why I should quit.* [Participant 7; female, aged 24 years; MI condition]

When discussing their *perception of health chatbots*, most participants felt comfortable talking about smoking and health-related topics in general with the chatbot. The chatbot was perceived as an accessible conversational partner that is nonjudgmental with helping intentions. Several participants compared chatbots with humans and noted that chatbots are

XSL•FO
RenderX

particularly useful because they require less logistic and mental effort. However, a number of participants also made a clear distinction between chatbots and humans, and they found it hard to anthropomorphize them. They felt that the chatbot conversation has a fixed flow and chatbots have no real emotional capability, which differentiated them from real humans. While many people reported being aware of and appreciated the anonymous nature of the interaction and the temporary storage of data, a few participants expressed concerns about technology in general. They felt uncanny that technology was advancing at a speed that it could analyze human emotions and were worried that such technology was premature in terms of true understanding. In addition, there was a perception of risk in data sharing and privacy. For example, one participant commented the following:

> *If I were to talk to a human counselor, they are bounded by professional secrecy, and the bot, even though it is used by a professional and that professional is bounded to secrecy, the bot's data is stored somewhere. Data leaks happen all the time.* [Participant 4; male, aged 29 years; MI condition]

Finally, participants discussed the *ideal intervention* they would like for smoking cessation. The most frequently mentioned suggestion for intervention content was that more background stories should be explored (eg, why one started smoking and what is the triggering situation for one to smoke) to increase personal relevance. Many participants also mentioned that they would like to discuss quitting methods in more detail to feel confident. Specifically, regarding the use of chatbots, most participants preferred them as a standby tool that can respond to users' questions and provide timely advice. Users would like to be in contact with the chatbot when lacking motivation or having cravings and would like to receive practical tips from the chatbot. In terms of long-term use, several participants suggested that the chatbot could generate progress reports and have adjusted conversations at different progress points.

Discussion

Principal Findings

In this mixed methods study, we aimed to investigate the effectiveness and user experience of MI strategies delivered by a chatbot. Results showed support for our hypothesis regarding the advantage of the MI chatbot in terms of user experience (ie, engagement, therapeutic alliance, perceived empathy, and interaction satisfaction), but no significant effect was found for the cessation-related outcomes (ie, intention to quit and self-efficacy). Importantly, exploratory analyses showed increases in participants' intention to quit, self-efficacy, and therapeutic alliance over the sessions, suggesting the usefulness of chatbots for this purpose. These increases were more profound in the MI condition (except for therapeutic alliance, for which the increase over time did not differ between conditions), suggesting the potential of MI in the long run. Finally, participants reported an overall decrease in their engagement with the chatbot. Qualitative data supported and expanded the experiment findings. We identified constructs that relate to users' prospective acceptability, timely experience, and retrospective acceptability of our smoking cessation chatbot, with several insights applicable to health chatbots in general.

Our findings regarding user experience are in line with those of previous research not involving chatbots demonstrating clients' preferences for an empathetic and collaborative approach. Previous research in health counseling has highlighted patients' preferences regarding an empathic and collaborative approach [14,18], and our findings echo such preferences and extend their value to the chatbot setting. Our qualitative results provided further possible explanations for this. Participants in the MI condition mentioned more frequently than participants in the CC condition that they felt understood by the chatbot and that they were more engaged in the conversation. In the CC condition, participants more frequently felt coerced and confronted by the chatbot, and they indicated that they would like more personal conversation and that the provided information was not sufficient. These findings show that participants value the relational aspect of the chatbot, suggesting that it is essential for the chatbot to have a person-centered approach and have a supportive attitude. Moreover, several participants mentioned having previous expectations of an authoritative and confrontational figure sending generic information, whereas the MI chatbot exceeded their expectations with its personalized, understanding, and empathetic approach. Client expectations regarding treatment are known as a key factor in counseling outcomes and satisfaction [65]. Expectancy violation theory suggests that positive expectation violation (eg, the chatbot was more personalized than expected) predicts more favorable outcomes than negative expectation confirmation (eg, the chatbot was confrontational as expected) [66]. Indeed, despite recent research favoring MI, many addiction counselors may still use confrontational techniques. Additionally, smokers often face criticism from their social environment, which may influence their expectations of how a chatbot will interact with them [16,51]. In other words, the previous expectation of a CC-like chatbot interaction may account for the positive user experience with the MI chatbot.

In terms of the motivating effect on smokers' intention to quit and self-efficacy, our results did not demonstrate significant differences between the 2 chatbot styles. The MI chatbot did not evoke more motivation or self-efficacy than the CC chatbot. One plausible explanation for why we did not find differences regarding the different styles could be that our participants were recruited with established motivation at baseline. On average, participants in both conditions placed themselves on the baseline contemplation ladder between "I think I should quit but not quite ready" and "I am starting to think about how to change my smoking patterns." Smokers at this stage of change might appreciate both encouragement and concrete guidance on quitting [67]. In other words, participants may value the intrinsic motivation-evoking aspect of MI, whereas the direct advice in CC is perceived as useful at the same time. Together, these effects may have resulted in nonsignificant differences in motivating intention to quit between the 2 styles. Another interpretation stems from the particular chatbot setting. While we hypothesized the MI chatbot to be helpful in evoking motivation and encouraging personally achievable quit plans, its effectiveness might depend on clients' actual perception of

XSL•FO
RenderX

it. The fidelity of MI requires true understanding and empathy, and it remains unclear whether the chatbot encompasses these qualities. Research in human-computer interaction has suggested that inaccurate artificial empathy might be less favorable than no empathy [68,69]. This is supported by our interview findings such that, despite the clearly appreciated nonjudgmental and supportive tone of the MI chatbot, people preferred discussing emotional experiences with real humans because chatbots do not have emotional capacities. On the other hand, while CC was expected to have less positive effects, using a chatbot as the portrayed counselor might have mitigated the anticipated resistance. As noted by both our interview participants and previous research, health chatbots are predominantly perceived as an instrumental tool that answers questions and provides information [70], and therefore, people might value the informational aspect more than the therapeutic aspect. However, it is important to highlight that, regardless of the condition, participants overall reported a small yet significant increase in their motivation to quit and self-efficacy after the chatbot interaction. This is consistent with previous research suggesting that minimal conversation about smoking cessation can affect quitting intentions and behaviors [71], and chatbots might be particularly useful due to their high accessibility. It is conceivable that the use of a chatbot alone is effective, and the specific communication strategies used may not have a major impact on the outcomes [43]. To better realize their potential, future research is needed to understand the impact of different automatic conversation strategies and the role of users' communication preferences regarding chatbots and design chatbots that combine effective strategies from different counseling approaches.

Another notable finding is that people's engagement with the chatbot decreased over time. Despite the fact that our experimental setting consisted of only 2 sessions, this provides a stepping stone for the long-term use of chatbots, which is particularly important for smoking cessation, which requires sustained effort. Research in human-computer interaction shows that user engagement tends to decline over time [72-74]. In addition to the novelty effect (ie, people engage with new technology due to curiosity, which declines gradually [75]) that has been proposed to explain declining engagement [12], our study found that users' technical experience was an active determinant in the engagement process. The most frequently mentioned reason for participants to disengage from the chatbot interaction was the encountered technical errors (eg, chatbot misunderstanding and repairing). Being able to input free text was one of the aspects that participants found engaging. This provides a plausible explanation for the finding that the MI chatbot was perceived as more engaging than the CC chatbot such that the open-ended questions asked by the MI chatbot offered users active involvement in the conversation. To increase people's engagement with the chatbot and the intervention, future research is needed to identify more factors that influence the engagement process. One area worthy of investigation relates to the automatic detection of engagement or disengagement from user utterances so that the chatbot can respond in time to keep the users engaged. For example, He et al [76] explored textual cues indicative of user engagement and suggested using cognitive strategies (eg, quizzes and reflective questions) to keep the user engaged.

Limitations

A number of methodological limitations should be borne in mind when interpreting the results. First, the length of the interaction was fairly short. We aimed to simulate a brief intake session and a first counseling session, and our results were in line with research on the positive effects of brief interventions for smoking cessation [77]. Even though we found a significant increase in smokers' motivation and self-efficacy to quit, this increase was relatively small, which is not unexpected given the briefness of the interaction. Our exploratory results suggest that the small but positive effects increased over time. The promising outlook calls for studies that involve prolonged use to seek more meaningful changes and investigate whether the effects can be sustained in the longer term. We believe that chatbots are well suited for addressing this long-term objective as they facilitate extended and personalized interactions, and we plan to delve deeper into this potential in a follow-up longitudinal study. In addition, we measured intention rather than actual cessation as a primary outcome, which provided initial support for the usefulness of the chatbot interaction. To further this line of research, behavioral outcomes are needed to ascertain the effectiveness of chatbot interventions. Moreover, objective use data (eg, interaction length and sentence length) can be used to complement self-reported engagement. It is recommended that future research include more objective instruments such as biochemical and behavioral measures to validate the effects of chatbot interaction. In addition, although our sample represents the main user group of new technologies such as chatbots, their smoking behaviors might not be representative of the general smoking population. Their familiarity with technology might have also led to results that are less applicable to other user groups. Considering the potential impact of demographics and smoking behaviors, it is important to design chatbots that can cater to a broader range of populations.

Implications for Future Research and Practice

Our findings, coupled with previous research, suggest that chatbots have the potential to motivate smokers to quit, and equipping them with MI skills could improve user experience. However, the insignificant results on cessation-related outcomes call for more research to further identify strategies that can effectively motivate cessation. For example, previous research has attempted to understand the working mechanism of MI in an automated setting by comparing different types of open-ended questions and reflections and suggested the advantage of incorporating multiple-choice questions particularly in an automated setting, expanding on the general guideline of using open-ended questions in traditional MI [28,78]. Therefore, it is important to disentangle the MI components and translate active ingredients to a chatbot setting. Moreover, our findings suggest that the directive informational aspect of the CC chatbot was appreciated by the participants, and future research is encouraged to combine such aspects with the nonjudgmental spirit of MI to capture the best of both worlds. In addition, in view of the person-centered approach and the potential of

XSL•FO
RenderX

personalization, it is important to explore individual factors that can influence people's preferences and experience with the chatbots. For instance, previous research has found that users' need for autonomy and their self-efficacy was influential in their responses to an MI or a CC chatbot [79], highlighting the importance of addressing users' unique needs in designing future chatbot interventions. As noted in the qualitative findings, technology-delivered counseling has both positive (eg, easy access and facilitated disclosure) and negative (eg, lack of true understanding) effects. Considering the interpersonal nature of health counseling, it is essential to understand the role of the human touch in chatbot-delivered interventions. Our study provided some initial insights into this, such as that the chatbots facilitated health-related disclosure but people are skeptical of their ability to understand human feelings. Future work is needed to further understand how people perceive chatbots in an interpersonal setting; for example, how do people view the relational communication delivered by a chatbot? How do people disclose their emotions and understand chatbots' responses? This line of research will help improve the design of chatbot-delivered MI and health counseling in general.

On a practical note, researchers and practitioners should be cautious in humanizing chatbots. Even though research has pointed to the potential of empathetic agents [40], overly imitating human emotions could result in an uncanny feeling [80]. Our findings suggest that, while humanlikeness was generally favored, participants felt hesitant discussing emotions with the chatbot when they were aware of its mindless nature. To best equip chatbots with social cues, future work needs to understand users' perceptions of different facets of emotions. For example, cognitive empathy (ie, understanding others' perspectives and emotions) from a chatbot is appreciated, whereas affective empathy (ie, emotional response to others' emotions) induces perceived eeriness [38]. Moreover, we argue that chatbots should be considered a supplementary tool rather than a replacement for human caregivers, especially in health-related contexts where inaccurate and invalid information can have a harmful impact on people's well-being. Several of our interview participants suggested that the chatbot should be a standby conversational tool that provides timely support. Functions supporting long-term use (eg, quitting progress report) should be considered as well.

An important implication of this study is that previous expectations play a crucial role in people's experiences with and responses to chatbot interventions. While people used to have expectations of a customer service–like generic chatbot, such expectations might be drastically changed by recent rapid technological advancements. In particular, ChatGPT, one of the latest large language models, has captured the attention of millions since its release. It might reshape people's expectations of chatbots with its large amount of knowledge and natural text generation and its capability to respond to any conceivable input. However, it should be made clear to the public that such technologies should not be regarded as a replacement for medical authorities and are not ideal for long-term support as it requires personalized care and tailored interventions, whereas ChatGPT's capabilities are limited to the general information that it was trained on. In addition, it is well established that chatbots relying on large language models may give socially desirable answers, may suffer from undesired biases, and are not necessarily truthful [81]. Despite the extensive amount of information available, such technologies simply answer to requests and often overlook user background and conversational contexts. We are aiming to develop chatbots that can understand and engage with users in the long term while ensuring accuracy and safety. Health care chatbots need to balance safety measures (eg, using prescripted utterances to avoid harmful content) with engagement features (eg, flexibility and variety in texts). Technologies such as ChatGPT could bring additional support by providing coherently written information, but human control should be involved to ensure safe, responsible, and accountable use.

Conclusions

This study adds to the body of research regarding the effectiveness and user experience of chatbots using 2 different counseling styles. The results point in the direction that chatbots are a promising tool in motivating cessation and the use of MI can improve user experience. Several constructs such as affective attitude and engagement were identified to understand users' previous expectations, timely experience, and retrospective acceptability of the chatbot interaction. These findings highlight the potential of chatbots for smoking cessation and suggest a few avenues for future research.

Acknowledgments

This work is supported by the Look Who's Talking project (official project code 406.D1.19.054), funded by the Dutch Research Council. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank the project members of Look Who's Talking for their valuable insights during the discussions they had. They would like to thank Divyaa Balaji in particular for her contribution in coding 2 of the interviews.

Authors' Contributions

LH, EK, MA, and RW conceived the presented idea and designed the experiment. LH and EB developed the chatbot used in the experiment. LH carried out the experiment and collected and analyzed the data. LH drafted the main manuscript text and prepared the tables and figures. All authors contributed to and approved the final version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. Lancet 2014 Jul 05;384(9937):45-52. [doi: 10.1016/S0140-6736(14)60648-6] [Medline: 24996589]
2. Stead L, Lancaster T. Combined pharmacotherapy and behavioural interventions for smoking cessation. Cochrane Database Syst Rev 2012 Oct 17;10(3):CD008286 [FREE Full text] [doi: 10.1002/14651858.CD008286.pub2] [Medline: 23076944]
3. Li IC, Lee SY, Chen CY, Jeng YQ, Chen YC. Facilitators and barriers to effective smoking cessation: counselling services for inpatients from nurse-counsellors' perspectives--a qualitative study. Int J Environ Res Public Health 2014 May 06;11(5):4782-4798 [FREE Full text] [doi: 10.3390/ijerph110504782] [Medline: 24806190]
4. Everett K, Odendaal HJ, Steyn K. Doctors' attitudes and practices regarding smoking cessation during pregnancy. S Afr Med J 2005 May;95(5):350-354. [Medline: 15931451]
5. Vogt F, Hall S, Marteau TM. Examining why smokers do not want behavioral support with stopping smoking. Patient Educ Couns 2010 May;79(2):160-166. [doi: 10.1016/j.pec.2009.10.007] [Medline: 19932580]
6. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
7. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
8. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. JMIR Mhealth Uhealth 2020 May 19;8(5):e15085 [FREE Full text] [doi: 10.2196/15085] [Medline: 32427114]
9. Balaji D, He L, Giani S, Bosse T, Wiers R, de Bruijn GJ. Effectiveness and acceptability of conversational agents for sexual health promotion: a systematic review and meta-analysis. Sex Health 2022 Jul 22;19(5):391-405. [doi: 10.1071/sh22016]
10. Daley K. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. OSF Preprints Preprint posted online July 27, 2020 [FREE Full text] [doi: 10.31219/osf.io/6vne3]
11. Kato A, Tanigawa T, Satake K, Nomura A. Efficacy of the Ascure smoking cessation program: retrospective study. JMIR Mhealth Uhealth 2020 May 14;8(5):e17270 [FREE Full text] [doi: 10.2196/17270] [Medline: 32406856]
12. Croes EA, Antheunis ML. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. J Soc Pers Relatsh 2020 Sep 25;38(1):279-300. [doi: 10.1177/0265407520959463]
13. Perski O, Blandford A, Ubhi HK, West R, Michie S. Smokers' and drinkers' choice of smartphone applications and expectations of engagement: a think aloud and interview study. BMC Med Inform Decis Mak 2017 Feb 28;17(1):25 [FREE Full text] [doi: 10.1186/s12911-017-0422-8] [Medline: 28241759]
14. Miller W, Benefield RG, Tonigan JS. Enhancing motivation for change in problem drinking: a controlled comparison of two therapist styles. J Consult Clin Psychol 1993 Jun;61(3):455-461. [doi: 10.1037//0022-006x.61.3.455] [Medline: 8326047]
15. Davis TD. Beliefs about confrontation among substance abuse counselors. J Evid Based Soc Work 2005 May 09;2(1-2):1-17. [doi: 10.1300/j394v02n01_01]
16. Kotz D, Wesseling G, Huibers MJ, van Schayck OC. Efficacy of confrontational counselling for smoking cessation in smokers with previously undiagnosed mild to moderate airflow limitation: study protocol of a randomized controlled trial. BMC Public Health 2007 Nov 15;7(1):332 [FREE Full text] [doi: 10.1186/1471-2458-7-332] [Medline: 18005415]
17. Kotz D, Huibers MJ, Vos R, van Schayck CP, Wesseling G. Principles of confrontational counselling in smokers with chronic obstructive pulmonary disease (COPD). Med Hypotheses 2008 Jan;70(2):384-386. [doi: 10.1016/j.mehy.2007.05.012] [Medline: 17604568]
18. Schneider RJ, Casey J, Kohn R. Motivational versus confrontational interviewing: a comparison of substance abuse assessment practices at employee assistance programs. J Behav Health Serv Res 2000 Feb;27(1):60-74. [doi: 10.1007/bf02287804]
19. Polcin DL. Reexamining confrontation and motivational interviewing. Addict Disord Their Treat 2006;5(4):201-209. [doi: 10.1097/01.adt.0000205048.44129.6a]
20. Kotz D, Wesseling G, Huibers MJ, van Schayck OC. Efficacy of confronting smokers with airflow limitation for smoking cessation. Eur Respir J 2009 Apr 01;33(4):754-762 [FREE Full text] [doi: 10.1183/09031936.00116308] [Medline: 19129277]
21. Moyers TB, Miller WR, Hendrickson SM. How does motivational interviewing work? Therapist interpersonal skill predicts client involvement within motivational interviewing sessions. J Consult Clin Psychol 2005 Aug;73(4):590-598. [doi: 10.1037/0022-006x.73.4.590]
22. Ahluwalia JS, Okuyemi K, Nollen N, Choi WS, Kaur H, Pulvers K, et al. The effects of nicotine gum and counseling among African American light smokers: a 2 x 2 factorial design. Addiction 2006 Jun 12;101(6):883-891. [doi: 10.1111/j.1360-0443.2006.01461.x] [Medline: 16696632]
23. Moeseneder L, Figlioli P, Caspar F. Confronting patients: therapists' model of a responsiveness based approach. J Contemp Psychother 2017 Nov 1;48(2):61-67. [doi: 10.1007/s10879-017-9371-x]

XSL•FO
RenderX

24. Francis N, Rollnick S, McCambridge J, Butler C, Lane C, Hood K. When smokers are resistant to change: experimental analysis of the effect of patient resistance on practitioner behaviour. Addiction 2005 Aug 15;100(8):1175-1182. [doi: 10.1111/j.1360-0443.2005.01124.x] [Medline: 16042648]
25. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change and Grow. New York, NY: Guilford Publications; 2023.
26. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312 [FREE Full text] [Medline: 15826439]
27. Toriello PJ, Strohmer DC. Addictions counselors' credibility: the impact of interactional style, recovery status, and nonverbal behavior. J Addict Offender Couns 2011 Dec 23;25(1):43-57. [doi: 10.1002/j.2161-1874.2004.tb00192.x]
28. Apodaca TR, Longabaugh R. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. Addiction 2009 May 07;104(5):705-715 [FREE Full text] [doi: 10.1111/j.1360-0443.2009.02527.x] [Medline: 19413785]
29. Borrelli B. Motivational interviewing for smoking cessation. In: Polosa R, editor. Advances in Smoking Cessation. London, UK: Future Medicine; 2013:128-141.
30. Markland D, Ryan RM, Tobin VJ, Rollnick S. Motivational interviewing and self–determination theory. J Soc Clin Psychol 2005 Sep;24(6):811-831. [doi: 10.1521/jscp.2005.24.6.811]
31. Kotz D, Huibers MJ, West RJ, Wesseling G, van Schayck OC. What mediates the effect of confrontational counselling on smoking cessation in smokers with COPD? Patient Educ Couns 2009 Jul;76(1):16-24. [doi: 10.1016/j.pec.2008.11.017] [Medline: 19150590]
32. Pilnick A, Coleman T. "I'll give up smoking when you get me better": patients' resistance to attempts to problematise smoking in general practice (GP) consultations. Soc Sci Med 2003 Jul;57(1):135-145. [doi: 10.1016/s0277-9536(02)00336-2] [Medline: 12753822]
33. Burke BL, Arkowitz H, Menchola M. The efficacy of motivational interviewing: a meta-analysis of controlled clinical trials. J Consult Clin Psychol 2003;71(5):843-861. [doi: 10.1037/0022-006x.71.5.843]
34. Heckman CJ, Egleston BL, Hofmann MT. Efficacy of motivational interviewing for smoking cessation: a systematic review and meta-analysis. Tob Control 2010 Oct 30;19(5):410-416 [FREE Full text] [doi: 10.1136/tc.2009.033175] [Medline: 20675688]
35. Chatzisarantis NL, Hagger MS, Kamarova S, Kawabata M. When effects of the universal psychological need for autonomy on health behaviour extend to a large proportion of individuals: a field experiment. Br J Health Psychol 2012 Nov 19;17(4):785-797. [doi: 10.1111/j.2044-8287.2012.02073.x] [Medline: 22515432]
36. Harrits GS, Larsen LT. Advice not safely ignored: professional authority and the strength of legitimate complexity. Sociology 2021 Apr 29;55(5):1015-1034. [doi: 10.1177/0038038521994029]
37. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019 Apr 05;21(4):e12887 [FREE Full text] [doi: 10.2196/12887] [Medline: 30950796]
38. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):625-636. [doi: 10.1089/cyber.2018.0110] [Medline: 30334655]
39. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. J Med Internet Res 2020 Sep 30;22(9):e22845 [FREE Full text] [doi: 10.2196/22845] [Medline: 32996892]
40. Bickmore T, Gruber A. Relational agents in clinical psychiatry. Harv Rev Psychiatry 2010;18(2):119-130. [doi: 10.3109/10673221003707538] [Medline: 20235777]
41. Shingleton RM, Palfai TP. Technology-delivered adaptations of motivational interviewing for health-related behaviors: a systematic review of the current research. Patient Educ Couns 2016 Jan;99(1):17-35 [FREE Full text] [doi: 10.1016/j.pec.2015.08.005] [Medline: 26298219]
42. Chen J, Chen C, Walther JB, Shyam Sundar S. Do you feel special when an AI doctor remembers you? Individuation effects of AI vs. human doctors on user experience. In: Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI EA '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411763.3451735]
43. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022 Apr 12;22(1):726 [FREE Full text] [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
44. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Method 2009 Nov;41(4):1149-1160. [doi: 10.3758/brm.41.4.1149]
45. Lindson-Hawley N, Thompson TP, Begh R. Motivational interviewing for smoking cessation. Cochrane Database Syst Rev 2015 Mar 02(3):CD006936. [doi: 10.1002/14651858.CD006936.pub3] [Medline: 25726920]
46. Zhu SH, Stretch V, Balabanis M, Rosbrook B, Sadler G, Pierce JP. Telephone counseling for smoking cessation: effects of single-session and multiple-session interventions. J Consult Clin Psychol 1996 Feb;64(1):202-211. [doi: 10.1037//0022-006x.64.1.202] [Medline: 8907100]

XSL•FO
RenderX

47. Reiter E, Robertson R, Osman LM. Lessons from a failure: gnerating tailored smoking cessation letters. Artif Intell 2003 Mar;144(1-2):41-58. [doi: 10.1016/s0004-3702(02)00370-3]
48. Verbiest M, Brakema E, van der Kleij R, Sheals K, Allistone G, Williams S, et al. National guidelines for smoking cessation in primary care: a literature review and evidence analysis. NPJ Prim Care Respir Med 2017 Jan 20;27(1):2 [FREE Full text] [doi: 10.1038/s41533-016-0004-8] [Medline: 28108747]
49. ROKENinfo. URL: https://www.rokeninfo.nl/ [accessed 2024-04-05]
50. Basar E, Balaji D, He L, Hendrickx I, Krahmer E, de Bruijn GJ, et al. HyLECA: a framework for developing hybrid long-term engaging controlled conversational agents. In: Proceedings of the 5th International Conference on Conversational User Interfaces. 2023 Presented at: CUI '23; July 19-21, 2023; Eindhoven, The Netherlands. [doi: 10.1145/3571884.3604404]
51. White WL, Miller WR. The use of confrontation in addiction treatment history, science, and time for change. Counselor 2007;8(4):12-30 [FREE Full text]
52. Pan D, Huey SJJ, Heflin LH. Ethnic differences in response to directive vs. non-directive brief intervention for subsyndromal depression. Psychother Res 2019 Feb 22;29(2):186-197. [doi: 10.1080/10503307.2017.1325023] [Medline: 28532259]
53. Biener L, Abrams DB. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. Health Psychol 1991;10(5):360-365. [doi: 10.1037//0278-6133.10.5.360] [Medline: 1935872]
54. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. Br J Addict 1991 Sep 13;86(9):1119-1127. [doi: 10.1111/j.1360-0443.1991.tb01879.x] [Medline: 1932883]
55. Etter JF, Bergman MM, Humair JP, Perneger TV. Development and validation of a scale measuring self-efficacy of current and former smokers. Addiction 2000 Jun 03;95(6):901-913. [doi: 10.1046/j.1360-0443.2000.9569017.x] [Medline: 10946439]
56. O'Brien HL, Cairns P, Hall M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. Int J Hum Comput Stud 2018 Apr;112:28-39. [doi: 10.1016/j.ijhcs.2018.01.004]
57. Hatcher RL, Gillaspy JA. Development and validation of a revised short version of the working alliance inventory. Psychother Res 2006 Jan;16(1):12-25. [doi: 10.1080/10503300500352500]
58. Rubin RB, Martin MM. Development of a measure of interpersonal communication competence. Commun Res Rep 1994 Jun;11(1):33-44. [doi: 10.1080/08824099409359938]
59. Rogers ES, Smelson DA, Gillespie CC, Elbel B, Poole S, Hagedorn HJ, et al. Telephone smoking-cessation counseling for smokers in mental health clinics: a patient-randomized controlled trial. Am J Prev Med 2016 Apr;50(4):518-527. [doi: 10.1016/j.amepre.2015.10.004] [Medline: 26711163]
60. Madson MB, Mohn RS, Zuckoff A, Schumacher JA, Kogan J, Hutchison S, et al. Measuring client perceptions of motivational interviewing: factor analysis of the Client Evaluation of Motivational Interviewing scale. J Subst Abuse Treat 2013 Mar;44(3):330-335. [doi: 10.1016/j.jsat.2012.08.015] [Medline: 22999814]
61. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
62. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
63. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Method 2016 Nov 29;5(1):80-92. [doi: 10.1177/160940690600500107]
64. Huberty CJ, Morris JD. Multivariate analysis versus multiple univariate analyses. Psychol Bull 1989;105(2):302-308. [doi: 10.1037//0033-2909.105.2.302]
65. Greenberg RP, Constantino MJ, Bruce N. Are patient expectations still relevant for psychotherapy process and outcome? Clin Psychol Rev 2006 Oct;26(6):657-678. [doi: 10.1016/j.cpr.2005.03.002] [Medline: 15908088]
66. Burgoon JK. Expectancy violations theory. In: The International Encyclopedias of Communication. Hoboken, NJ: John Wiley & Sons; Dec 11, 2015.
67. O'Neill HK, Gillispie MA, Slobin K. Stages of change and smoking cessation: a computer-administered intervention program for young adults. Am J Health Promot 2000 Aug 25;15(2):93-6, iii. [doi: 10.4278/0890-1171-15.2.93] [Medline: 11194701]
68. Eagle T, Blau C, Bales S, Desai N, Li V, Whittaker S. "I don't know what you mean by `I am anxious'": a new method for evaluating conversational agent responses to standardized mental health inputs for anxiety and depression. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-23. [doi: 10.1145/3488057]
69. Urakami J, Moore BA, Sutthithatip S, Park S. Users' perception of empathic expressions by an advanced intelligent system. In: Proceedings of the 7th International Conference on Human-Agent Interaction. 2019 Presented at: HAI '19; October 6-10, 2019; Kyoto, Japan. [doi: 10.1145/3349537.3351895]
70. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019 Aug 21;5:2055207619871808 [FREE Full text] [doi: 10.1177/2055207619871808] [Medline: 31467682]
71. Jeong M, Tan AS, Brennan E, Gibson L, Hornik RC. Talking about quitting: interpersonal communication as a mediator of campaign effects on smokers' quit behaviors. J Health Commun 2015 Jul 06;20(10):1196-1205 [FREE Full text] [doi: 10.1080/10810730.2015.1018620] [Medline: 26147367]

XSL•FO
RenderX

72. de Haas M, Vogt P, van den Berghe R, Leseman P, Oudgenoeg-Paz O, Willemsen B, et al. Engagement in longitudinal child-robot language learning interactions: disentangling robot and task engagement. Int J Child Comput Interact 2022 Sep;33:100501. [doi: 10.1016/j.ijcci.2022.100501]
73. Ahmad MI, Mubin O, Orlando J. Adaptive social robot for sustaining social engagement during long-term children–robot interaction. Int J Hum Comput Interact 2017 Mar 03;33(12):943-962. [doi: 10.1080/10447318.2017.1300750]
74. Bickmore T, Schulman D, Yin L. Maintaining engagement in long-term interventions with relational agents. Appl Artif Intell 2010 Jul 01;24(6):648-666 [FREE Full text] [doi: 10.1080/08839514.2010.492259] [Medline: 21318052]
75. Wells JD, Campbell DE, Valacich JS, Featherman M. The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective. Decis Sci 2010 Nov 24;41(4):813-843. [doi: 10.1111/j.1540-5915.2010.00292.x]
76. He L, Braggaar A, Basar E, Krahmer E, Antheunis M, Wiers R. Exploring user engagement through an interaction lens: what textual cues can tell us about human-chatbot interactions. In: Proceedings of the 6th International Conference on Conversational User Interfaces. 2024 Presented at: CUI '24; July 8-10, 2024; Luxembourg City, Luxembourg URL: https://programs.sigchi.org/cui/2024/program/content/166851 [doi: 10.1145/3640794.3665536]
77. Aveyard P, Begh R, Parsons A, West R. Brief opportunistic smoking cessation interventions: a systematic review and meta-analysis to compare advice to quit and offer of assistance. Addiction 2012 Jun 28;107(6):1066-1073. [doi: 10.1111/j.1360-0443.2011.03770.x] [Medline: 22175545]
78. Friederichs SA, Oenema A, Bolman C, Guyaux J, Van Keulen HM, Lechner L. Motivational interviewing in a web-based physical activity intervention: questions and reflections. Health Promot Int 2015 Sep 06;30(3):803-815. [doi: 10.1093/heapro/dat069] [Medline: 24101160]
79. Leeuwis L, He L. Hi, I'm Cecil(y) the smoking cessation chatbot: the effectiveness of motivational interviewing and confrontational counseling chatbots and the moderating role of the need for autonomy and self-efficacy. In: Proceedings of the 6th International Workshop on Chatbot Research and Design. 2022 Presented at: CONVERSATIONS 2022; November 22-23, 2022; Amsterdam, The Netherlands. [doi: 10.1007/978-3-031-25581-6_1]
80. Gray K, Wegner DM. Feeling robots and human zombies: mind perception and the uncanny valley. Cognition 2012 Oct;125(1):125-130. [doi: 10.1016/j.cognition.2012.06.007] [Medline: 22784682]
81. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual Event, Canada. [doi: 10.1145/3442188.3445922]

Abbreviations

CC: confrontational counseling
CONSORT: Consolidated Standards of Reporting Trials
MANOVA: multivariate ANOVA
MI: motivational interviewing

Edited by SR Ramos; submitted 28.09.23; peer-reviewed by M Kapitány-Fövény, E Meijer; comments to author 24.02.24; revised version received 18.04.24; accepted 02.05.24; published 06.08.24.

Please cite as:
He L, Basar E, Krahmer E, Wiers R, Antheunis M
Effectiveness and User Experience of a Smoking Cessation Chatbot: Mixed Methods Study Comparing Motivational Interviewing and Confrontational Counseling
J Med Internet Res 2024;26:e53134
URL: https://www.jmir.org/2024/1/e53134
doi:10.2196/53134
PMID:39106097

©Linwei He, Erkan Basar, Emiel Krahmer, Reinout Wiers, Marjolijn Antheunis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Health Professionals' Views on the Use of Conversational Agents for Health Care: Qualitative Descriptive Study

A Luke MacNeill[1,2], PhD; Lillian MacNeill[1,2], PhD; Alison Luke[1,2], PhD; Shelley Doucet[1,2], PhD

[1]Centre for Research in Integrated Care, University of New Brunswick, Saint John, NB, Canada

[2]Department of Nursing and Health Sciences, University of New Brunswick, Saint John, NB, Canada

Corresponding Author:
A Luke MacNeill, PhD
Centre for Research in Integrated Care
University of New Brunswick
355 Campus Ring Road
Saint John, NB, E2L 4L5
Canada
Phone: 1 506 648 5777
Email: luke.macneill@unb.ca

Abstract

Background: In recent years, there has been an increase in the use of conversational agents for health promotion and service delivery. To date, health professionals' views on the use of this technology have received limited attention in the literature.

Objective: The purpose of this study was to gain a better understanding of how health professionals view the use of conversational agents for health care.

Methods: Physicians, nurses, and regulated mental health professionals were recruited using various web-based methods. Participants were interviewed individually using the Zoom (Zoom Video Communications, Inc) videoconferencing platform. Interview questions focused on the potential benefits and risks of using conversational agents for health care, as well as the best way to integrate conversational agents into the health care system. Interviews were transcribed verbatim and uploaded to NVivo (version 12; QSR International, Inc) for thematic analysis.

Results: A total of 24 health professionals participated in the study (19 women, 5 men; mean age 42.75, SD 10.71 years). Participants said that the use of conversational agents for health care could have certain benefits, such as greater access to care for patients or clients and workload support for health professionals. They also discussed potential drawbacks, such as an added burden on health professionals (eg, program familiarization) and the limited capabilities of these programs. Participants said that conversational agents could be used for routine or basic tasks, such as screening and assessment, providing information and education, and supporting individuals between appointments. They also said that health professionals should have some oversight in terms of the development and implementation of these programs.

Conclusions: The results of this study provide insight into health professionals' views on the use of conversational agents for health care, particularly in terms of the benefits and drawbacks of these programs and how they should be integrated into the health care system. These collective findings offer useful information and guidance to stakeholders who have an interest in the development and implementation of this technology.

(J Med Internet Res 2024;26:e49387) doi:10.2196/49387

KEYWORDS

conversational agents; chatbots; health care; health professionals; health personnel; qualitative; interview

Introduction

Background

Global health spending has reached record levels and is expected to increase in the coming years [1,2]. Despite this enormous investment, health care systems around the world are under pressure as they struggle to cope with health worker shortages [3,4], health system overuse [5,6], and a growing burden of noncommunicable diseases [7,8], among other challenges. These challenges and the pressure that they place on health care systems can have consequences for people providing and accessing health services. For instance, health professionals

XSL•FO
RenderX

face heavy workloads and high levels of exhaustion and burnout [9-12]. In addition, individuals who are seeking care often experience long wait times that can negatively impact their health and well-being [13,14].

To reduce strain on health services and ensure the safe and efficient delivery of care, health care systems need to find accessible and cost-effective ways to supplement and support traditional forms of health care. Conversational agents are an emerging technology that might help. Broadly speaking, conversational agents are automated software programs that are designed to engage in humanlike conversation with people [15,16]. Common examples include text-based chatbots [17], voice-activated virtual assistants [18], and "embodied" conversational agents that appear in virtual or animated bodies [19,20]. Whereas standard computer interfaces support a somewhat rudimentary and impersonal interaction with users, conversational agents provide users with the impression of interacting with a real human. This quality would presumably make these programs well suited for dealing with sensitive issues or topics, such as those that arise within a health context.

In recent years, conversational agents that are used for health care have become more common in clinical and commercial settings [16,21-23]. Although the literature on these programs is still growing, current evidence suggests that they are an effective tool for health promotion and service delivery across a range of health areas [24-26]. For instance, they have been used to promote physical activity [27,28] and weight loss [29,30], support chronic disease management [31,32], provide preconception care [33,34], and improve mental health [35,36]. Research also shows that people who use these programs tend to have positive perceptions and opinions about them, as well as a desire or intention to continue using them in the future [16,37,38].

One aspect of conversational agents that tends to get overlooked in the literature is the views of health professionals toward these programs. Health professionals have considerable experience with the management and delivery of health care and may be able to offer unique insight into the use of this technology for health purposes. To date, only a handful of studies have considered health professionals' views in this area [39-41]. These studies had a narrow scope in terms of the specific types of health professionals that were assessed (eg, mental health professionals). In addition, much of this research captured participants' views using structured questionnaires with predefined questions and response scales [40,41]. These questionnaires allowed for standardized and easily aggregated data, but they may not have covered topics or points that would have emerged organically through other means, such as qualitative interviews. Due to these limitations, there is currently a fragmented and incomplete understanding of health professionals' views surrounding this technology.

Objectives

The purpose of this study was to gain a more comprehensive understanding of health professionals' views on the use of conversational agents for health care. We recruited health professionals with a variety of professional backgrounds and asked them to share their views via semistructured open-ended interviews. We were interested in their views on 2 specific subjects. First, we wanted to learn more about the potential benefits and drawbacks of using conversational agents for health care. Second, we wanted to gain a better understanding of how to best integrate conversational agents into the health care system. Capturing health professionals' views on these subjects was expected to provide useful insight into whether and how these programs should be implemented in practical settings.

Methods

Study Design

We used a qualitative descriptive design to explore health professionals' views on the use of conversational agents for health care. The goal of qualitative description is to provide a rich, detailed description of a phenomenon, emphasizing surface versus more interpretive readings of the available data in an effort to preserve participants' voices and perspectives [42,43]. It is viewed as a useful approach for answering questions relevant to practitioners and policy makers [42] and therefore well suited for this study. All data for this study were collected using cross-sectional semistructured interviews (details in the Procedure section). The study is reported in accordance with Standards for Reporting Qualitative Research guidelines [44].

Participants and Recruitment

We recruited Canadian health professionals for the study, specifically physicians, nurses, and regulated mental health professionals (eg, clinical psychologists). No prior experience with conversational agents was necessary for participation. Participants were recruited through social media posts, classified websites, e-newsletters, and emails to relevant organizations. The initial recruitment target was 24 participants, divided evenly between physicians, nurses, and regulated mental health professionals. Once these participants were recruited and interviewed, we evaluated the interview data to determine whether additional participants should be recruited. We decided that no further participants were required, as no new themes were identified in the last few interviews (ie, thematic saturation was reached) and we felt that the data set was sufficiently rich to address the research objectives.

Procedure

Prospective participants contacted the research team to learn more about the study and set up a web-based interview. Prior to the interview, participants were emailed a PDF document that provided information about conversational agents and illustrated several conversational agents that are used for health care. The purpose of the PDF document was to familiarize participants with these types of programs before the interview. The document covered conversational agents that support the prevention and treatment of specific health conditions, as well as conversational agents that promote healthy lifestyle habits (physical activity, etc). It included screenshots of both embodied and disembodied conversational agents. Screenshots were drawn from several publications in this area [34,45-50]. In some cases, the images from these publications were replaced with higher-quality versions of the same or similar images that were available from web-based sources (eg, researcher websites).

XSL•FO
RenderX

Participants were interviewed individually using the Zoom (Zoom Video Communications, Inc) videoconferencing platform. Interviews took place between March and September 2021. The same general procedure was used for all interviews. The interviews began with personal introductions and a review of the study information. Next, verbal consent was obtained from participants and the interview questions were administered. Question administration lasted between 10 and 35 minutes (depending on the participant) and followed a semistructured format. Participants were asked about the potential benefits and drawbacks of using conversational agents for health care, as well as the best way to integrate conversational agents into the health care system. See Multimedia Appendix 1 for a full list of the interview questions. At the end of the interviews, participants were given an opportunity to ask any questions that they had.

Data Analysis

A research assistant (SY) transcribed the interview recordings verbatim, and the transcripts were reviewed by the first author for accuracy. The transcripts were uploaded to NVivo (version 12; QSR International, Inc) for thematic analysis. The thematic analysis followed the 6 phases described by Braun and Clarke [51,52], as follows. The first and second authors familiarized themselves with the data set by reading and rereading the transcripts and making preliminary notes (phase 1). Next, they coded the data set via an iterative process (phase 2). They started by reviewing the transcripts of the first 3 interviews to generate preliminary codes and working definitions. This initial analysis guided the coding of the full data set, which was performed by the first author. The second author independently coded 25% of the transcripts and found a high level of agreement with the first author (Cohen κ=0.81; agreement weighted by source size=99.74%). The 2 coders refined the coding guidelines to address the few disagreements and applied the refined guidelines to the full data set. Next, they worked together to generate initial themes (phase 3); further develop and review the themes (phase 4); and refine, define, and name the themes (phase 5). Finally, the first author wrote the initial draft of the manuscript and all authors performed manuscript revisions (phase 6).

Trustworthiness

Trustworthiness is a broad concept that refers to the credibility, confirmability, dependability, and transferability of research [53,54]. We sought to establish trustworthiness in our study through the use of several strategies, including investigator triangulation (eg, interrater coding), the creation of an audit trail, purposive sampling of different participant groups, the use of verbatim quotes to support the themes, and peer debriefing with digital health researchers in both the academic and commercial spaces. To enhance methodological rigor, the data analysis was conducted by 2 researchers with different professional backgrounds, specifically the first and second authors. The first author is a health researcher with a focus on digital health research. He has led several studies on the use of conversational agents for health care, including an evaluation study and a scoping review. The second author is a health researcher with a more general focus on access to health services. Her knowledge of digital health and digital health research is more limited. Whereas the second author is primarily a qualitative researcher, the first author is a quantitative researcher with some qualitative experience.

Ethical Considerations

This study was approved by the research ethics board at the University of New Brunswick (009-2021). The research ethics board provided a delegated review due to the low-risk nature of the study. Informed consent was obtained from all study participants (details in the Procedure section). Transcripts were deidentified during the transcription process to protect participant confidentiality, and interview files were stored on a secure Microsoft SharePoint site. Participants received an entry into a draw for a CAD $100 (US $72.61) Amazon.ca gift card for their participation in the study.

Results

Participant Characteristics

A total of 24 health professionals completed the study. Participants included 8 physicians, 8 nurses, and 8 regulated mental health professionals (2 clinical psychologists, 2 psychotherapists, 2 counselors, and 2 clinical social workers). The sample consisted of 19 women and 5 men with a mean age of 42.75 (SD 10.71) years. Participants were from 7 Canadian provinces: Alberta, British Columbia, New Brunswick, Nova Scotia, Ontario, Prince Edward Island, and Saskatchewan. All participants reported that they had used conversational agents in the past, although only 2 had used a conversational agent for health care specifically.

Benefits of Conversational Agents

We identified 2 themes related to benefits of using conversational agents for health care: *increased access to care* and *workload support for health professionals* (see Table 1). More details on these themes are provided in the following subsections.

Table 1. Outline of themes and subthemes grouped by research objective.

Research objective	Theme (subthemes)
Benefits of conversational agents	• Increased access to care (convenience and availability, rural access to care) • Workload support for health professionals
Drawbacks of conversational agents	• Added burden on health professionals • Limited capabilities of conversational agents (technological limitations, inadequate care)
Integration into the health care system	• Health professional oversight • Routine or basic tasks (screening and assessment, information and education, support between appointments)

XSL•FO
RenderX

Increased Access to Care

Participants said that conversational agents could benefit patients or clients by providing increased access to care. Most participants who discussed this theme focused on the convenience and availability of these programs, although some participants discussed rural access to care as well.

Convenience and Availability

According to participants, conversational agents are able to provide patients or clients with convenient access to care. In other words, people can access these programs whenever they want, regardless of the day or time. For instance, one participant pointed out that conversational agents are available "anytime … 24/7" (Participant 18, nurse), and another participant stated: "To have access to something 24 hours a day is certainly beneficial" (Participant 1, nurse). Some participants contrasted conversational agents with health professionals, many of whom adhere to scheduled office hours. For instance, one participant said that conversational agents "would be always on, not dependent on the hours that health care professionals keep" (Participant 13, physician). By providing convenient access to care, these programs would allow patient or clients to avoid the wait times that are often experienced when seeking an appointment with a care provider: "Wait times are certainly an issue for people, being able to access a provider…. Perhaps these particular types of programs … would improve access [to care] for patients that might have difficulty connecting with a health care provider" (Participant 11, nurse).

Rural Access to Care

Participants noted that people who live in rural areas often have more limited access to care than people who live in urban areas, and that conversational agents could be one means to address this disparity. For instance, one participant said, "I think in certain ways it [the use of conversational agents] can be helpful for people if they live in a rural area and they don't have access to a specialist or they don't have a health care provider that they need" (Participant 4, psychotherapist). Other participants reinforced this statement, saying that these programs "might be able to reach more geographical locations that we can't reach" (Participant 5, clinical social worker), and that they would be useful "if you live way out in the boondocks and can't get into town" (Participant 21, clinical psychologist). Notably, most of the participants who discussed rural access to care were regulated mental health professionals.

Workload Support for Health Professionals

Participants said that conversational agents could provide health professionals with workload support. Although this point was discussed by all 3 health professional groups, it was particularly common among physicians and nurses. Participants noted that health professionals are often overworked and exhausted, and that conversational agents may be able to ease their workload to a certain extent. As one participant said, "It can take some of the pressure off of actual providers, if people are using these chatbots and whatnot. Maybe some of the demand goes down" (Participant 4, psychotherapist). By easing demand for health services, these programs would allow health professionals to devote more time and resources to people who have greater or more urgent needs. For example, one participant said, "[You] may be able to filter some of the more basic health care concerns, if you will, to this type of platform and then allow for health care providers to spend time with patients that have greater needs or need more time in-office or in-clinic" (Participant 11, nurse). One participant illustrated this point by sharing a story about a patient who was seriously injured after failing to get an appointment for a mental health medication refill:

> *[The patient] said "I call my family doctor, I don't get ahold of him. My psychiatrist left." You know, here's a guy where, maybe if we have AI taking the burden off the family doctor... maybe the guy calls, and the family doctor's able to see him and this problem never happens.* [Participant 3, physician]

Drawbacks of Conversational Agents

We identified 2 themes related to drawbacks of using conversational agents for health care: *added burden on health professionals* and *limited capabilities of conversational agents*. More details on these themes are provided in the following subsections.

Added Burden on Health Professionals

Although many participants thought that conversational agents could help ease their existing workload, some were wary that these programs could create new burdens that would strain their time and attention. For instance, participants anticipated that they would be receiving questions about conversational agents from their patients or clients, and that answering these questions would take time out of their workday. As one participant said, she expected "more questions from the patients about it, which ends up taking a lot of time for me" (Participant 13, physician). Participants also recognized that they would need to familiarize themselves with these programs so that they could answer patient or client questions, which would require a further time investment: "People may have questions about it, and if you don't take the time to familiarize with everything, you may not be able to answer them" (Participant 12, physician). Integrating conversational agents into their daily workflow was a concern as well. For instance, one participant worried about keeping on top of patient-related reports that might be supplied by these programs: "It would be an extra layer of my daily work that is time sensitive, you know? Like, every day at three, I have to make sure that I've looked at the last 24 hours of these things" (Participant 23, physician). Finally, some participants were concerned that they would have to spend time and resources addressing health issues that arise from the improper or incorrect use of a conversational agent: "If they [patients or clients] self-selected inappropriately to be seen by a virtual interface, and they delayed care and deteriorated … that would make my job a lot more complex" (Participant 13, physician).

Limited Capabilities of Conversational Agents

Participants spent a great deal of time discussing the capabilities of conversational agents. They had many concerns about the technological limitations of these programs, as well as the consequences of these technological limitations, particularly the potential for inadequate care.

XSL•FO
RenderX

Technological Limitations

Participants highlighted several technological limitations of conversational agents. Some participants focused on the emotional limitations of these programs, noting that they are unable to experience or convey genuine empathy or compassion: “When we’re at our most vulnerable moments, we expect compassionate care, and there's a limit to what you can get from a … conversational agent” (Participant 19, nurse). Others touched on this same topic from an interpersonal standpoint, commenting on “that loss of human connectedness, that human element” (Participant 11, nurse). Participants also discussed the sensory limitations of conversational agents, particularly their inability to detect important nonverbal cues from patients or clients: “I would wonder about a chatbot’s ability to notice nuance, to hear a voice choking … or to see eyes welling with tears” (Participant 6, clinical social worker). These nonverbal cues include cues that signal deception or dishonesty: “They [patients or clients] say they’re fine, but they actually look like they just dragged themselves out of bed and they haven’t showered in six days … I don’t know how good computers are at picking up that kind of ‘faking it’ behavior” (Participant 21, clinical psychologist). Finally, participants discussed the conversational limitations of these programs, doubting their ability to handle the complex question and answer exchanges that characterize health care interactions: “There’s no actual ability to have context-driven questioning, follow-up, reflection on what the person is hearing” (Participant 8, clinical psychologist).

Inadequate Care

Participants were concerned that the technological limitations of conversational agents would result in inadequate care for patients or clients. They were especially worried about patients or clients who are dealing with serious health issues. One participant provided an example within a mental health context:

> *If someone has really mild anxiety or mild depression, I think something like that could be a really great resource for them…. Whereas on the other hand, someone who is dealing with complex trauma or really significant mental illnesses, there’s no possibility that something automated could provide the care for that.* [Participant 4, psychotherapist]

A further point of concern was the possibility that patients or clients might not recognize that they are receiving inadequate care, which could compound their health issues. As one participant said, “They feel like what they're getting is enough and it may not actually be enough” (Participant 14, psychotherapist). One participant speculated about a patient’s thought processes in these situations: “‘Robot nurse Becky told me that I’m totally fine. So I’m just not going to go in to the doctor. I’m not going to call telehealth’ … I think it can have the potential of, I guess, false reassurance” (Participant 10, nurse).

Integration Into the Health Care System

We identified 2 themes related to the integration of conversational agents into the health care system: *health professional oversight* and *routine or basic tasks*. More details on these themes are provided in the following subsections.

Health Professional Oversight

Participants said that health professionals should have some oversight of conversational agents that are used for health care. For instance, many participants suggested that health professionals should have input into the development process so that there is some assurance that the program content is up-to-date and correct. As one participant commented, “It needs to have the signoff of a health care professional, somebody that tests it and can ensure all the information is accurate” (Participant 10, nurse). Participants also said that health professionals should have clinical oversight of these programs, which would allow them to evaluate whether a particular conversational agent is suitable for a particular patient or client: “The patient or client should come and see the health care professional and be referred to the program as opposed to just self-referral…. That way, at least someone has evaluated ‘Is this appropriate for that person?’” (Participant 12, physician). Clinical oversight would also allow health professionals to evaluate whether the information supplied by a program is consistent with the information they themselves provide to a patient or client. For example, one participant stated, “I would want to see what kind of information they [patients or clients] were getting, and whether it fit with the way I saw things” (Participant 21, clinical psychologist).

Routine or Basic Tasks

Participants said that conversational agents would be most appropriate for conducting routine or basic tasks. In the words of participants, these tasks would include “tedious jobs” (Participant 15, nurse), “basic things” (Participant 19, nurse), and “routine care” (Participant 13, physician). Although participants mentioned several tasks that would be appropriate for conversational agents, their reports tended to center on 3 general areas: screening and assessment, information and education, and support between appointments.

Screening and Assessment

Participants said that conversational agents could be used for screening and assessment purposes. For instance, these programs could screen patients or clients who have reached a specific service to see if the service is aligned with their needs: “I see it as a front-end thing where … you’re doing some screening, like, ‘Is this person in the right place?’” (Participant 21, clinical psychologist). In a similar manner, conversational agents would be able to screen individuals who are trying to access a particular program to see whether they meet the program qualifications: “In the program that I work with, we … really have to scope it down to those clients who qualify. So I think it might help us a little bit with filtering clients” (Participant 7, nurse). Once patients or clients have reached an appropriate program or service, conversational agents could perform an initial assessment to determine the severity of their health issues and the urgency of their needs. As one participant said, “It’s like that initial step to gather information and assess sort of the immediacy of response” (Participant 23, physician).

XSL•FO
RenderX

Information and Education

Participants said that conversational agents could be used to provide basic health information and education to patients or clients. For example, one participant said, "I can see the role it can play in basic patient education" (Participant 19, nurse), and another participant stated, "I think if people were seeking general health information, then that might be certainly appropriate" (Participant 11, nurse). Several participants provided examples of health areas where conversational agents could be applied. Many of these participants thought that these programs could be used to educate people on the prevention and management of chronic diseases, particularly diabetes, and to provide information to patients prior to and after surgery. One participant touched on both of these points: "In terms of basic health care stuff like diabetic education, pre- and post-op teaching … I think that it would be okay there" (Participant 1, nurse). Participants also thought that conversational agents could be used for psychoeducation and for educating people on health promotion topics such as physical activity and smoking cessation.

Support Between Appointments

Participants said that conversational agents could be used to support patients or clients between appointments. This point was especially common among regulated mental health professionals, who thought that these programs would be useful for "between-session maintenance" (Participant 4, psychotherapist) and "outside-of-session work" (Participant 14, psychotherapist). One participant provided a detailed example of how they might be used:

> *They'd be great for homework things, I think. So often, you check in with people and they're like, "Well, I kind of forgot what you told me" or "Oh, I didn't do it" … Having something that would have reminders, different pieces of homework that would pop up, I think that would be great.* [Participant 9, counselor]

In addition to addressing the topic of between-session work, this participant's quote highlights the importance of patient or client accountability. Participants thought that conversational agents could be used to remind individuals about any necessary work to be done between appointments with a care provider. For instance, one physician said, "I think that would be helpful, as like that little at-home reminder of 'This is what I was supposed to do for my exercise,' or whatever …. Something that keeps them accountable at home, once they leave my office" (Participant 24, physician).

Discussion

Principal Findings

The results of this study provide insight into health professionals' views on the use of conversational agents for health care, particularly in terms of the benefits and drawbacks of these programs and how they should be integrated into the health care system. Participants said that conversational agents would be useful for increasing access to care for the public, including rural populations that often face access issues. They also said that these programs could provide health professionals with some much-needed workload support. However, participants were wary of any added burden on health professionals, such as the need for program familiarization and having to integrate this technology into their workflow. They were also mindful of the technological limitations of these programs and the implications of these limitations for patients or clients. To help safeguard patient or client welfare, participants said that health professionals should have some oversight in terms of the development and implementation of these programs. They also suggested that the use of conversational agents should be limited to routine or basic health care tasks, such as screening and assessment, providing information and education, and supporting individuals between appointments.

Comparison With Prior Work

The current results reinforce many findings from past studies in this area, despite important differences in the scope and methodologies of these studies. For instance, similar insights were reported by Palanica et al [40], who administered a web-based survey to general practitioners; Sweeney et al [41], who distributed a web-based survey to mental health professionals; and Barnett et al [39], who conducted in-person focus groups with substance use counselors. Across these studies, participants said that conversational agents could benefit patients or clients by improving access to care. They also touched on the limited capabilities of these programs, particularly their inability to understand or convey genuine human emotions. With respect to implementation, the participants in these studies suggested that conversational agents would be best suited for routine tasks, such as providing information or education and supporting people between appointments (eg, sending reminders). All of these topics were major discussion points among the participants in our study.

There were some differences between the current results and past findings as well. Participants in our study emphasized several points that received little attention in previous studies, such as the impact of conversational agents on health professionals' workloads and the importance of health professional oversight of this technology. At the same time, our participants failed to mention certain other points that emerged in past studies, particularly the studies by Palanica et al [40] and Sweeney et al [41]. For instance, participants in these latter studies expressed concern that patients or clients could abuse conversational agents by self-diagnosing too often, and that they might not feel connected to their actual health care providers if they use these programs. Neither of these points were raised by participants in our study. These divergent findings are likely due to methodological factors. Both Palanica et al [40] and Sweeney et al [41] assessed health professionals' views using structured surveys with predefined questions and response scales. These surveys may have prompted participants to consider topics that they otherwise might not have thought about, while also limiting the scope and substance of their responses. By comparison, our study used semistructured interviews with open-ended questions, which captured participants' thoughts more organically.

XSL•FO
RenderX

Differences Among Health Professionals

Although the results of this study were fairly consistent across physicians, nurses, and regulated mental health professionals, there were some differences between these participant groups. For instance, mental health professionals were more likely than other participants to discuss how conversational agents could improve rural access to care, an outcome that might reflect challenges in delivering conventional mental health services to rural populations [55-57]. Mental health professionals were also more likely than others to say that conversational agents have the potential to support patients or clients between appointments. Most of their comments focused on the topic of between-session work, which speaks to the importance of this work in mental health treatment [58-60]. Meanwhile, discussion of workload support for health professionals was more common among physicians and nurses than regulated mental health professionals. This outcome may reflect the considerable occupational demands of physicians and nurses and the importance of managing time pressures within these professions [61-63]. Taken together, these collective differences suggest that there are profession-specific uses and benefits with respect to conversational agents, in addition to the more general uses and benefits that apply across professions.

Practical Implications

The results of this study offer insight to stakeholders who have an interest in the adoption and use of conversational agents for health care. For instance, policy makers, institutional leaders, and health professionals will be able to draw on these results when weighing the potential benefits and drawbacks of adopting this technology in practice. These individuals can also look to our results for guidance on how to safely implement this technology; namely, by employing programs that perform routine or basic tasks and by ensuring that these programs have appropriate oversight. Researchers and companies that develop conversational agents for health care might find these results useful as well. Based on our findings, developers should include health professionals in the development process and focus their efforts on programs that pose minimal safety risk to users. They should also clearly communicate the limitations of these programs to users to help prevent instances of inadequate care. By promoting the safe and responsible implementation of conversational agents for health care, developers and other stakeholders can increase the adoption and use of this technology, which will ultimately benefit those individuals who are either accessing or providing care.

Limitations and Future Directions

Participants in this study worked within the Canadian health care system, and so their views on the benefits and drawbacks of using conversational agents for health care, as well as the best way to implement this technology in practice, were shaped by their experiences within this particular setting. It is unclear whether these results are transferable to people working in other health care systems. It should be noted that many of the current results are consistent with findings from previous studies involving health professionals from other countries [39-41], although these studies had a narrower scope and may not be indicative of broader views on conversational agents within these health care systems. In any case, researchers may want to replicate this study with health professionals in other countries to see whether their views differ.

Another limitation with this study is the fact that many participants had never used a conversational agent for health care. The views shared by these participants could reflect their inexperience and may lack a certain degree of depth or insight. To their credit, all of these participants did have experience with conversational agents in other areas of their lives (eg, customer service chatbots), and so they likely had a reasonable understanding of how this technology would function and perform in a health context. In addition, study participants were shown several examples of conversational agents that provide health care before their interviews to ensure that they had knowledge of these programs and their capabilities. Regardless, future studies should focus on health professionals who have more direct experience with conversational agents in this domain so that there is some assurance that their understanding of these programs is accurate and comprehensive.

A further limitation with this study concerns the demographics of the sample. Although the participants in this study covered a wide range of ages and professions, most of the sample (19/24, 79%) identified as women. This gender imbalance is not particularly surprising, as women occupy the majority of health care positions in Canada [64,65]. However, a sample with a greater number of men may have produced additional or alternate insights. In the future, researchers in this area may want to emphasize men in their recruitment strategies so that they achieve a more gender-balanced sample of participants.

Conclusions

The results of this study provide insight into how health professionals view the use of conversational agents for health care. The health professionals in this study saw distinct benefits and drawbacks to the use of conversational agents for health care, and they shared several suggestions for safely integrating these programs into the health care system. This collective feedback provides a useful indicator of how conversational agents will be received in health settings in the coming years, and it also offers some guidance on how these programs should be implemented to provide the most benefit to people who are either accessing or providing care.

Acknowledgments

The authors thank Sungmin Yi for her assistance with transcription. This research was supported in part by a Research Professionals Initiative grant from the New Brunswick Innovation Foundation (RPI2020-010).

XSL•FO
RenderX

Data Availability

The data for this study (eg, interview transcripts) are not publicly available as per guidelines from the host institution's research ethics board. Data access is restricted to help safeguard the privacy and confidentiality of the participants.

Authors' Contributions

ALM, AL, and SD conceived and designed the study. ALM collected the interview data. ALM and LM conducted the analysis. ALM wrote the original draft of the manuscript and all authors contributed to manuscript revisions. No generative artificial intelligence tools were used in the preparation of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview questions.
[DOCX File , 15 KB - jmir_v26i1e49387_app1.docx]

References

1. Global Burden of Disease Health Financing Collaborator Network. Trends in future health financing and coverage: future health spending and universal health coverage in 188 countries, 2016-40. Lancet 2018;391(10132):1783-1798 [FREE Full text] [doi: 10.1016/S0140-6736(18)30697-4] [Medline: 29678341]
2. Global expenditure on health: public spending on the rise? 2021. World Health Organization. URL: https://apps.who.int/iris/bitstream/handle/10665/350560/9789240041219-eng.pdf [accessed 2022-08-18]
3. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? BMJ Glob Health 2022;7(6):e009316 [FREE Full text] [doi: 10.1136/bmjgh-2022-009316] [Medline: 35760437]
4. Crisp N, Chen L. Global supply of health professionals. N Engl J Med 2014;370(10):950-957. [doi: 10.1056/NEJMra1111610] [Medline: 24597868]
5. Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, et al. Evidence for overuse of medical services around the world. Lancet 2017;390(10090):156-168 [FREE Full text] [doi: 10.1016/S0140-6736(16)32585-5] [Medline: 28077234]
6. Ellen ME, Wilson MG, Vélez M, Shach R, Lavis JN, Grimshaw JM, Synthesis working group. Addressing overuse of health services in health systems: a critical interpretive synthesis. Health Res Policy Syst 2018;16(1):48 [FREE Full text] [doi: 10.1186/s12961-018-0325-x] [Medline: 29907158]
7. Bloom DE, Cafiero ET, Jané-Llopis E, Abrahams-Gessel S, Bloom LR, Fathima S, et al. The Global Economic Burden of Non-communicable Diseases. Switzerland: World Economic Forum; 2011.
8. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease study 2019. Lancet 2020;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
9. Humphries N, Morgan K, Conry MC, McGowan Y, Montgomery A, McGee H. Quality of care and health professional burnout: narrative literature review. Int J Health Care Qual Assur 2014;27(4):293-307. [doi: 10.1108/IJHCQA-08-2012-0087] [Medline: 25076604]
10. O'Connor K, Muller Neff D, Pitman S. Burnout in mental health professionals: a systematic review and meta-analysis of prevalence and determinants. Eur Psychiatry 2018;53:74-99. [doi: 10.1016/j.eurpsy.2018.06.003] [Medline: 29957371]
11. Shen X, Xu H, Feng J, Ye J, Lu Z, Gan Y. The global prevalence of burnout among general practitioners: a systematic review and meta-analysis. Fam Pract 2022;39(5):943-950. [doi: 10.1093/fampra/cmab180] [Medline: 35089320]
12. Woo T, Ho R, Tang A, Tam W. Global prevalence of burnout symptoms among nurses: a systematic review and meta-analysis. J Psychiatr Res 2020;123:9-20. [doi: 10.1016/j.jpsychires.2019.12.015] [Medline: 32007680]
13. McIntyre D, Chow CK. Waiting time as an indicator for health services under strain: a narrative review. Inquiry 2020;57:46958020910305 [FREE Full text] [doi: 10.1177/0046958020910305] [Medline: 32349581]
14. Organisation for Economic Co-operation and Development. Waiting times for health services: next in line. 2020. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/waiting-times-for-health-services_242e3c8c-en [accessed 2022-10-11]
15. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
16. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]

XSL•FO
RenderX

17. Rapp A, Curti L, Boldi A. The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. Int J Hum Comput Stud 2021;151:102630. [doi: 10.1016/j.ijhcs.2021.102630]
18. de Barcelos Silva A, Gomes MM, da Costa CA, da Rosa Righi R, Barbosa JLV, Pessin G, et al. Intelligent personal assistants: a systematic literature review. Expert Syst Appl 2020;147(C):113193. [doi: 10.1016/j.eswa.2020.113193]
19. Loveys K, Sebaratnam G, Sagar M, Broadbent E. The effect of design features on relationship quality with embodied conversational agents: a systematic review. Int J of Soc Robotics 2020;12(6):1293-1312. [doi: 10.1007/s12369-020-00680-7]
20. ter Stal S, Kramer LL, Tabak M, op den Akker H, Hermens H. Design features of embodied conversational agents in eHealth: a literature review. Int J Hum Comput Stud 2020;138:102409. [doi: 10.1016/j.ijhcs.2020.102409]
21. Jovanovic M, Baez M, Casati F. Chatbots as conversational healthcare services. IEEE Internet Comput 2021;25(3):44-51. [doi: 10.1109/mic.2020.3037151]
22. Montenegro JLZ, da Costa CA, da Rosa Righi R. Survey of conversational agents in health. Expert Syst Appl 2019;129(11):56-67. [doi: 10.1016/j.eswa.2019.03.054]
23. Parmar P, Ryu J, Pandya S, Sedoc J, Agarwal S. Health-focused conversational agents in person-centered care: a review of apps. NPJ Digit Med 2022;5:21 [FREE Full text] [doi: 10.1038/s41746-022-00560-6] [Medline: 35177772]
24. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020;22(7):e16021 [FREE Full text] [doi: 10.2196/16021] [Medline: 32673216]
25. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
26. Oh YJ, Zhang J, Fang ML, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18:160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
27. Bickmore TW, Silliman RA, Nelson K, Cheng DM, Winter M, Henault L, et al. A randomized controlled trial of an automated exercise coach for older adults. J Am Geriatr Soc 2013;61(10):1676-1683. [doi: 10.1111/jgs.12449] [Medline: 24001030]
28. Ellis T, Latham NK, DeAngelis TR, Thomas CA, Saint-Hilaire M, Bickmore TW. Feasibility of a virtual exercise coach to promote walking in community-dwelling persons with Parkinson disease. Am J Phys Med Rehabil 2013;92(6):472-481 [FREE Full text] [doi: 10.1097/PHM.0b013e31828cd466] [Medline: 23552335]
29. Graham SA, Pitter V, Hori JH, Stein N, Branch OH. Weight loss in a digital app-based diabetes prevention program powered by artificial intelligence. Digit Health 2022;8:20552076221130619 [FREE Full text] [doi: 10.1177/20552076221130619] [Medline: 36238752]
30. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. JMIR Diabetes 2017;2(2):e28 [FREE Full text] [doi: 10.2196/diabetes.8590] [Medline: 30291087]
31. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. Acceptability of an embodied conversational agent for type 2 diabetes self-management education and support via a smartphone app: mixed methods study. JMIR Mhealth Uhealth 2020;8(7):e17038 [FREE Full text] [doi: 10.2196/17038] [Medline: 32706734]
32. Guhl E, Althouse AD, Pusateri AM, Kimani E, Paasche-Orlow MK, Bickmore TW, et al. The atrial fibrillation health literacy information technology trial: pilot trial of a mobile health app for atrial fibrillation. JMIR Cardio 2020;4(1):e17162 [FREE Full text] [doi: 10.2196/17162] [Medline: 32886070]
33. Gardiner P, Hempstead MB, Ring L, Bickmore T, Yinusa-Nyahkoon L, Tran H, et al. Reaching women through health information technology: the Gabby preconception care system. Am J Health Promot 2013;27(3 Suppl):eS11-eS20 [FREE Full text] [doi: 10.4278/ajhp.1200113-QUAN-18] [Medline: 23286652]
34. Jack BW, Bickmore T, Yinusa-Nyahkoon L, Reichert M, Julce C, Sidduri N, et al. Improving the health of young African American women in the preconception period using health information technology: a randomised controlled trial. Lancet Digit Health 2020;2(9):e475-e485 [FREE Full text] [doi: 10.1016/S2589-7500(20)30189-8] [Medline: 33328115]
35. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
36. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018;5(4):e64 [FREE Full text] [doi: 10.2196/mental.9782] [Medline: 30545815]
37. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
38. Gabarron E, Larbi D, Denecke K, Årsand E. What do we know about the use of chatbots for public health? Stud Health Technol Inform 2020;270:796-800. [doi: 10.3233/SHTI200270] [Medline: 32570492]

XSL•FO
RenderX

39. Barnett A, Savic M, Pienaar K, Carter A, Warren N, Sandral E, et al. Enacting 'more-than-human' care: clients' and counsellors' views on the multiple affordances of chatbots in alcohol and other drug counselling. Int J Drug Policy 2021;94:102910 [FREE Full text] [doi: 10.1016/j.drugpo.2020.102910] [Medline: 33059955]
40. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019;21(4):e12887 [FREE Full text] [doi: 10.2196/12887] [Medline: 30950796]
41. Sweeney C, Potts C, Ennis E, Bond R, Mulvenna MD, O'Neill S, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthc 2021;2(3):25. [doi: 10.1145/3453175]
42. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
43. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
44. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
45. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, O'Donnell J. Response to a relational agent by hospital patients with depressive symptoms. Interact Comput 2010;22(4):289-298 [FREE Full text] [doi: 10.1016/j.intcom.2009.12.001] [Medline: 20628581]
46. Bickmore TW, Schulman D, Sidner C. Automated interventions for multiple health behaviors using conversational agents. Patient Educ Couns 2013;92(2):142-148 [FREE Full text] [doi: 10.1016/j.pec.2013.05.011] [Medline: 23763983]
47. Horsch CHG, Lancee J, Griffioen-Both F, Spruit S, Fitrianie S, Neerincx MA, et al. Mobile phone-delivered cognitive behavioral therapy for insomnia: a randomized waitlist controlled trial. J Med Internet Res 2017;19(4):e70 [FREE Full text] [doi: 10.2196/jmir.6524] [Medline: 28400355]
48. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
49. Lisetti C, Amini R, Yasavur U, Rishe N. I can help you change! an empathic virtual agent delivers behavior change health interventions. ACM Trans Manage Inf Syst 2013;4(4):19. [doi: 10.1145/2544103]
50. Richards D, Caldwell P. Improving health outcomes sooner rather than later via an interactive website and virtual specialist. IEEE J Biomed Health Inform 2018;22(5):1699-1706. [doi: 10.1109/JBHI.2017.2782210] [Medline: 29990229]
51. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
52. Braun V, Clarke V. Thematic Analysis: A Practical Guide. New York, NY: SAGE Publications; 2022.
53. Lincoln YS, Guba EG. Naturalistic Inquiry. New York, NY: SAGE Publications; 1985.
54. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018;24(1):120-124 [FREE Full text] [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]
55. Friesen E. The landscape of mental health services in rural Canada. Univ Tor Med J 2019;96(2):47-52.
56. Mental Health Commission of Canada. Rural and Remote Mental Health in Canada: Evidence Brief on Best and Promising Practices. Ottawa, Canada: Mental Health Commission of Canada; 2020.
57. Novik N, McKee B, Pearce K. Social work practice and mental health services outside of urban settings. In: Jeffery B, Novik N, editors. Rural and Northern Social Work Practice: Canadian Perspectives. Regina, Canada: University of Regina; 2022:154-171.
58. Kazantzis N, Dattilio FM. Definitions of homework, types of homework, and ratings of the importance of homework among psychologists with cognitive behavior therapy and psychoanalytic theoretical orientations. J Clin Psychol 2010;66(7):758-773. [doi: 10.1002/jclp.20699] [Medline: 20527055]
59. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. Behav Ther 2016;47(5):755-772. [doi: 10.1016/j.beth.2016.05.002] [Medline: 27816086]
60. Strunk DR. Homework. Cogn Behav Pract 2022;29(3):560-563. [doi: 10.1016/j.cbpra.2022.02.009]
61. Goldsby E, Goldsby M, Neck CB, Neck CP. Under pressure: time management, self-leadership, and the nurse manager. Adm Sci 2020;10(3):38. [doi: 10.3390/admsci10030038]
62. Gordon CE, Borkan SC. Recapturing time: a practical approach to time management for physicians. Postgrad Med J 2014;90(1063):267-272. [doi: 10.1136/postgradmedj-2013-132012] [Medline: 24599633]
63. Leis SJ, Anderson A. Time management strategies for new nurses. Am J Nurs 2020;120(12):63-66. [doi: 10.1097/01.NAJ.0000724260.01363.a3] [Medline: 33214377]
64. Khanam F, Langevin M, Savage K, Uppal S. Women Working in Paid Care Occupations. Canada: Statistics Canada; 2022.
65. Health workforce in Canada: Overview. Canadian Institute for Health Information. 2022. URL: https://www.cihi.ca/en/health-workforce-in-canada-overview [accessed 2024-03-01]

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.05.23; peer-reviewed by Y Xi, S Baptista, S Wei; comments to author 05.01.24; revised version received 01.03.24; accepted 01.06.24; published 25.09.24.

Please cite as:
MacNeill AL, MacNeill L, Luke A, Doucet S
Health Professionals' Views on the Use of Conversational Agents for Health Care: Qualitative Descriptive Study
J Med Internet Res 2024;26:e49387
URL: https://www.jmir.org/2024/1/e49387
doi:10.2196/49387
PMID:39320936

©A Luke MacNeill, Lillian MacNeill, Alison Luke, Shelley Doucet. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis

Michelle Hoang Nguyen[1,2], BS; João Sedoc[3], PhD; Casey Overby Taylor[1,2,4], PhD

[1]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, United States

[2]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, United States

[3]Department of Technology, Operations and Statistics, Stern School of Business, New York University, New York, NY, United States

[4]Division of General Internal Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Corresponding Author:
Michelle Hoang Nguyen, BS
Institute for Computational Medicine
Johns Hopkins University
Hackerman Hall
Baltimore, MD, 21218
United States
Phone: 1 (410) 516 4116
Email: mnguye79@jhmi.edu

Abstract

Background: Family health history (FHx) is an important predictor of a person's genetic risk but is not collected by many adults in the United States.

Objective: This study aims to test and compare the usability, engagement, and report usefulness of 2 web-based methods to collect FHx.

Methods: This mixed methods study compared FHx data collection using a flow-based chatbot (KIT; the curious interactive test) and a form-based method. KIT's design was optimized to reduce user burden. We recruited and randomized individuals from 2 crowdsourced platforms to 1 of the 2 FHx methods. All participants were asked to complete a questionnaire to assess the method's usability, the usefulness of a report summarizing their experience, user-desired chatbot enhancements, and general user experience. Engagement was studied using log data collected by the methods. We used qualitative findings from analyzing free-text comments to supplement the primary quantitative results.

Results: Participants randomized to KIT reported higher usability than those randomized to the form, with a mean System Usability Scale score of 80.2 versus 61.9 ($P<.001$), respectively. The engagement analysis reflected design differences in the onboarding process. KIT users spent less time entering FHx information and reported more conditions than form users (mean 5.90 vs 7.97 min; $P=.04$; and mean 7.8 vs 10.1 conditions; $P=.04$). Both KIT and form users somewhat agreed that the report was useful (Likert scale ratings of 4.08 and 4.29, respectively). Among desired enhancements, personalization was the highest-rated feature (188/205, 91.7% rated medium- to high-priority). Qualitative analyses revealed positive and negative characteristics of both KIT and the form-based method. Among respondents randomized to KIT, most indicated it was easy to use and navigate and that they could respond to and understand user prompts. Negative comments addressed KIT's personality, conversational pace, and ability to manage errors. For KIT and form respondents, qualitative results revealed common themes, including a desire for more information about conditions and a mutual appreciation for the multiple-choice button response format. Respondents also said they wanted to report health information beyond KIT's prompts (eg, personal health history) and for KIT to provide more personalized responses.

Conclusions: We showed that KIT provided a usable way to collect FHx. We also identified design considerations to improve chatbot-based FHx data collection: First, the final report summarizing the FHx collection experience should be enhanced to provide more value for patients. Second, the onboarding chatbot prompt may impact data quality and should be carefully considered. Finally, we highlighted several areas that could be improved by moving from a flow-based chatbot to a large language model implementation strategy.

(J Med Internet Res 2024;26:e55164) doi:10.2196/55164

XSL•FO
RenderX

KEYWORDS

family health history; chatbots; conversational agents; digital health tools; usability; engagement; report usefulness; evaluation; crowdsourcing; mixed methods

Introduction

Background

Comprehensive and high-quality family health history (FHx) is a valuable tool for research and patient care, even as the genomic medicine landscape evolves. As next-generation sequencing costs decline, genotype-first approaches are being explored as an alternative to traditional phenotypic ascertainment or clinical informatics approaches. Even in this era of genomic advancement, FHx remains a valuable risk predictor, especially for complex conditions not fully explained by genomic factors alone [1,2]. For example, in a population genetics screening study by Bylstra et al [3], participants with higher cancer risk based on FHx assessment had a higher prevalence of clinically actionable variants in American College of Medical Genetics secondary findings (version 2.0) cancer genes than both participants with average FHx-derived cancer risk and those who had no documented FHx.

Although FHx is widely recognized as a crucial step in genetic risk assessment, less than 50% of Americans report actively collecting their FHx [4]. Furthermore, most patients do not have complete FHx in their electronic health record (EHR) [5].

Barriers to collecting FHx have persisted for both patients and providers. Providers face challenges in collecting FHx due to a lack of time during patient visits and limitations of current tools to document FHx comprehensively [6-8]. Patient challenges have included a lack of knowledge of the value of FHx to assess disease risk, the great effort required to confirm conditions present among family members, the need to regularly update FHx to maintain accuracy, and insufficient encouragement from clinicians to complete FHx documentation [4,6,9,10]. Innovative digital health strategies may help to address patient and provider challenges by providing mechanisms to reduce clinician burden and improve FHx completion rate by empowering patients to electronically complete their FHx at a time and place that is convenient for them. Furthermore, for patients, digital health strategies can facilitate access to relevant educational materials and mechanisms to save and update FHx over time.

In recent years, there have been increasing efforts to advance FHx collection with the development, implementation, and use of web-based tools [11-14]. Many of the digital health strategies to date are patient-facing and involve patient-entered information. These strategies are promising. For example, in the Cancer Health Assessments Reaching Many study, Mittendorf et al [15] adapted a web-based survey application for Lynch syndrome risk assessment from a provider-facing to a patient-facing format that supported computer and mobile use. Adaptations were designed to be patient-friendly by presenting survey questions about family history one relative at a time (or a small group of relatives at a time), including literacy aids, and to computationally determine side of the family instead of relying on the patient. In another study designed to assess the effectiveness of a cancer genetic service delivery among a diverse patient population, Mittendorf et al [16] found that patients preferred FHx self-assessment and data entry using an electronic history tool over clinician-gathered history. Factors favorable among study participants included being self-paced, private, convenient, and allowing them to gather more detailed information from relatives without time constraints [16].

While electronic tools may have their benefits, a subset of individuals from the Cancer Health Assessments Reaching Many study who preferred clinician-gathered history noted that clinicians could provide more tailored responses and educational information [16]. Those findings suggest a need to support tailoring and information retrieval features that may be difficult to implement using standard form-based methods.

A growing number of studies explore more advanced interfaces, such as conversational agents for health data collection and risk assessment, that may support features standard form-based methods cannot [17-25]. For example, Soni et al [26] compared the usability of a health data collection chatbot, Dokbot, against a standard REDCap (Research Electronic Data Capture; Vanderbilt University) form. Dokbot is a Health Insurance Portability and Accountability Act–compliant application accessible on both desktop and mobile devices that allows users to complete health data collection by chatting with a customizable chatbot avatar through multiple-choice button responses [27]. Soni et al [26] found that a collection chatbot is more engaging and interactive than form-based methods but does not appear more usable than a traditional web-based form [26]. However, more work is needed to understand chatbot-specific usability, the impact of chatbot-specific features on usability, and to determine what features users would prioritize for future development.

Objectives

To address the challenges of collecting FHx and better understand the benefits and drawbacks between different web-based modalities, this paper studied the potential for a flow-based chatbot approach to collect FHx. The primary objective of this study was to understand the usability of a chatbot compared to a form-based method and observe any chatbot-specific user interface characteristics that may explain this usability. To do this, we created an interactive web-embedded chatbot to administer a 3-generational family history survey (KIT, the curious interactive test). We studied its overall usability, perceived usefulness in collecting FHx, and how engaging it was. This research is motivated by recommendations from prior work to improve the effectiveness of FHx tools and address gaps surrounding patient's FHx knowledge [4,6]. New web-based methods, such as chatbots for FHx collection, hold promise to address these issues.

To address our research objective, we developed KIT, the FHx data collection chatbot, and a baseline form-based FHx data collection method. We compared usability measures,

XSL•FO
RenderX

engagement with FHx data collection methods, and FHx summary report usefulness across these 2 methods.

Methods

Intervention Design

Both FHx data collection tools were developed by adapting the National Institutes of Health *All of Us* Research Program Family Health History Questionnaire, which is currently being administered on the web to *All of Us* research participants [28]. For both FHx tools, we enabled users to learn more about health conditions they might not have been familiar with by providing options to access National Institutes of Health National Library of Medicine (NLM) consumer–focused definitions from MedlinePlus [29].

Study participants were randomized to 1 of 2 FHx data collection interventions: a form-based and a chatbot-based intervention. The form-based FHx data collection method was hosted on JHMI Qualtrics. The question structure and presentation of the form remained true to the *All of Us* survey, aside from adding an information button for health conditions (Figure 1). Participants were asked to complete data entry for all 11 condition categories. A diagram is provided in Figure S1A in Multimedia Appendix 1. Data entry for FHx collection consisted of multiple-answer and multiple-choice questions. There were no free-response questions. At the end of data collection, participants were shown a copy of their responses and invited to download it to serve as a report of their FHx. The responses were in the same visual format as the survey and were available as a PDF file. More details about the form intervention questionnaire format and questions can be found in Figure S2A in Multimedia Appendix 1.

Figure 1. Form-based family health history tool user interface. Condition definition displayed in the red box.

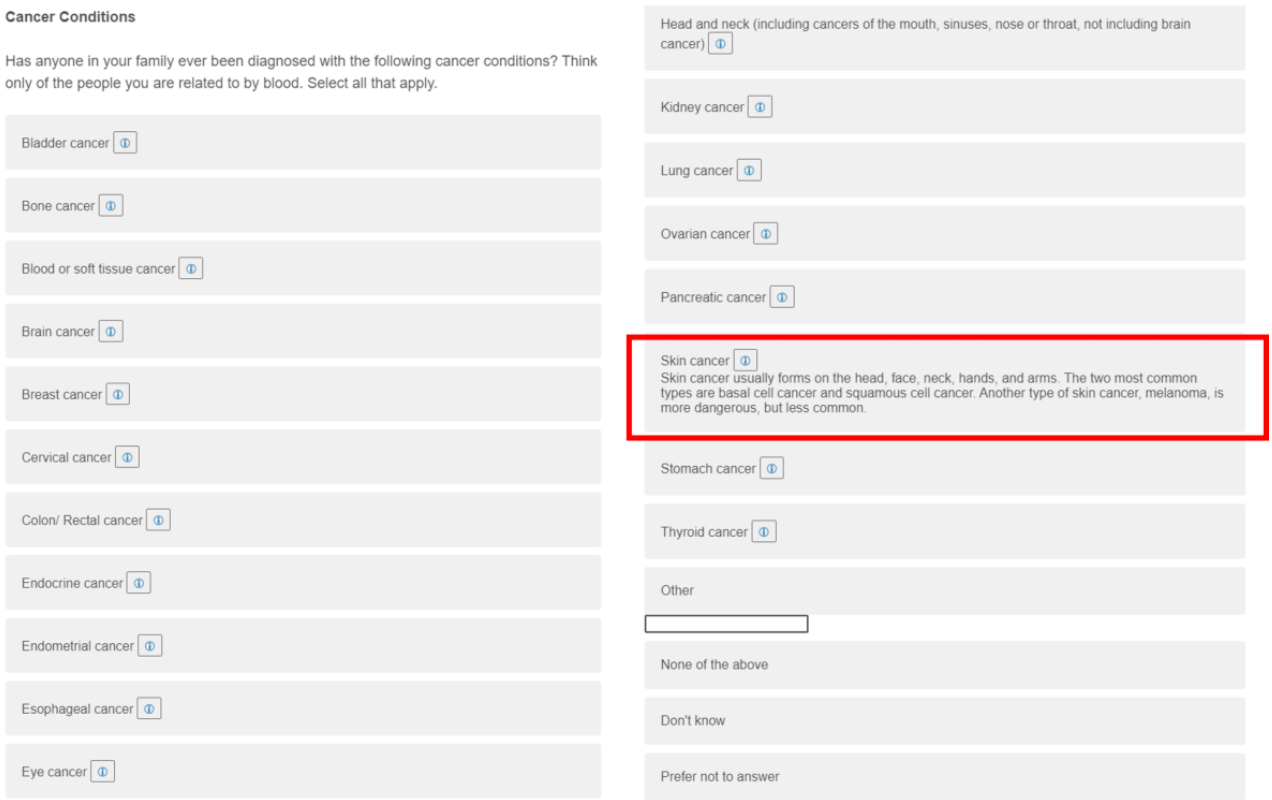

The chatbot intervention, KIT, was developed on Juji (Juji, Inc) [30]. Juji is a low-code chatbot implementation platform that reduces the technical barrier to creating chatbots. Chatbots created with Juji can be instantiated with a chatbot avatar selected from a library of available default avatars, such as the nurse figure selected for KIT. Although Juji does allow customization for more specialized chatbot behavior, such as media responses, we limited customization only when functionally necessary to move the data collection conversation forward without halting the user. Such customization included a custom topic called "Asking a question about medical information" to prevent the Question and Answer knowledge bank (which contains all the NLM MedlinePlus condition definitions) from being triggered at inappropriate instances and to allow users to input information without being incorrectly rerouted in their conversation (Figure 2). Participants were asked to select condition categories for which they knew family history among the 11 condition categories. Then, they completed data entry for the condition categories they preselected. A diagram is provided in Figure S3A in Multimedia Appendix 1 with a sample interaction video (Video S5A in Multimedia Appendix 1). The primary data entry for KIT FHx collection consisted of multiple-answer, multiple-choice questions. Users had the option to message KIT (free-response text entry) to respond (in lieu of clicking the multiple-choice button) if they had a question or if they selected "other" at any point during data collection. When users selected "other," KIT asked the user for additional details about their data entry (eg, "Can you describe what kind of hormone/endocrine condition your father has had?"). At the end of the interaction, all participants were invited to download a

XSL•FO
RenderX

copy of their conversation to serve as a report of their FHx. The conversation is a transcript in a text file format. More details about the KIT intervention format and questions appear in Multimedia Appendix 1.

The intervention designs described above reflect the improvements made after initial pilot testing of the FHx data collection tools with crowdsourced participants. From this pilot study, we tested the crowdsourced recruitment workflow to ensure acceptable response data quality and monitored user comments to improve the FHx interventions. From the comments, we reduced the latent time between KIT's responses and implemented the FHx intervention format described earlier and in Figure S3A in Multimedia Appendix 1. More details about procedures and modifications from the pilot study are available in Multimedia Appendix 2.

Figure 2. Curious interactive test (KIT), family health history chatbot, user interface on Juji. Condition question and response from KIT with KIT Avatar in the red box.

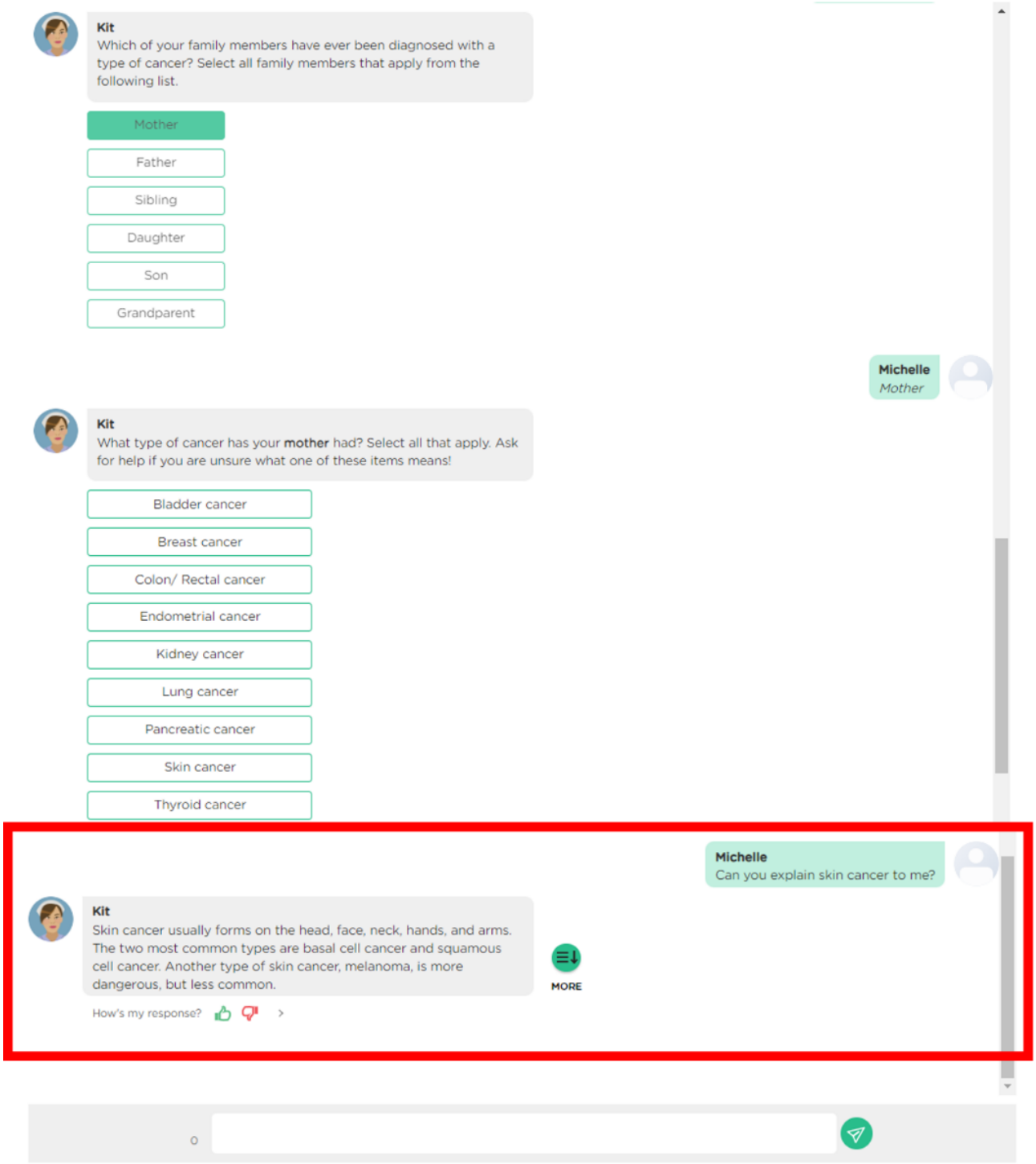

XSL•FO
RenderX

Study Design and Population

This mixed methods study compared 2 modalities to collect FHx: chatbot-based and form-based tools. All study participants completed their FHx before completing a web-based survey via Qualtrics (Qualtrics). The survey gathered both quantitative and qualitative data. Qualitative results were collected and analyzed to validate and expand on the quantitative findings related to the primary outcomes studied, that is, usability, usefulness, and engagement. A between-subjects approach was chosen to reduce participant fatigue and potential learning effects. Our approach was built on the findings of a 25-subject pilot. We drew several lessons from the pilot and applied them in this study. In summary, we improved our recruitment strategy, survey workflow, and FHx intervention. Our findings also helped us to better estimate the sample size for this study.

Eligibility for this study was limited to individuals in the United States aged >18 years who knew at least 2 first-degree relatives with at least 1 condition each. We recruited from 2 separate nonclinical study source populations, 1 from Amazon Mechanical Turk (MTurk; Amazon Inc) and 1 from Qualtrics Panels. Amazon MTurk is a web-based crowdsourcing marketplace in which businesses and researchers post tasks for registered MTurk workers to complete. Qualtrics Panels is a market research panel platform managed by Qualtrics that recruits participants to complete surveys.

Qualtrics Panel participants were recruited from various web-based sources, including website intercept recruitment, member referrals, targeted email lists, gaming sites, customer loyalty web portals, permission-based networks, and social media. Crowdsourcing has been a growing source of participant recruitment for health studies [31-33]. However, previous work has demonstrated mixed data quality when recruiting crowdsourced participants for behavioral studies [30,31]. Because of this, we aimed to recruit participants from 2 separate crowdsourced populations to validate findings and increase confidence in survey results.

The MTurk source population consisted of 580 MTurk workers who had demonstrated high-quality data for past behavioral studies. We created an MTurk Human Intelligence Task and released this incrementally, allowing 9 workers to complete the study per batch. Once deemed eligible, workers who passed the screening criteria could view the consent form. We programmed randomization through Qualtrics to assign participants to equal-sized groups at an equal rate, with a quota to keep track of assignments to either of the FHx data collection tools. We randomized to reduce intervention selection bias and ensure adequate participant counts in both arms for a fully powered study. We recruited Qualtrics Panel participants with demographic quotas approximating the US population as limited by Qualtrics Panel project managers.

All recruited participants followed similar study processes (Figure 3). After determining their eligibility, we showed potential study participants contextual introductory information about genetics and family history. This was followed by a knowledge question, “Which of the following family members are not first-degree relatives?” (choices: father, sister, brother, niece) to ensure participants had thoroughly read the introductory information. If participants correctly answered “niece,” they were shown the rest of the FHx intervention. They were then redirected to proceed to the consent form. One difference between the MTurk and Qualtrics Panels procedures was that MTurk workers needed to verify a valid MTurk Worker ID before being randomized to one of the 2 FHx data collection tools. In contrast, Qualtrics Panel participants were randomized upon accepting the study invitation. After data collection, participants could download the report or chat transcript and were redirected to complete the assessment survey (Multimedia Appendix 3). Within the assessment survey, there was one attention check. If users did not correctly respond to the question, they were redirected and could not complete the survey.

Figure 3. Study design flow diagram consisting of stages: eligibility, consent, intervention, and assessment. CUQ: Chatbot Usability Questionnaire; FHx: family health history; KIT: curious interactive test; MTurk: Mechanical Turk; SUS: System Usability Scale.

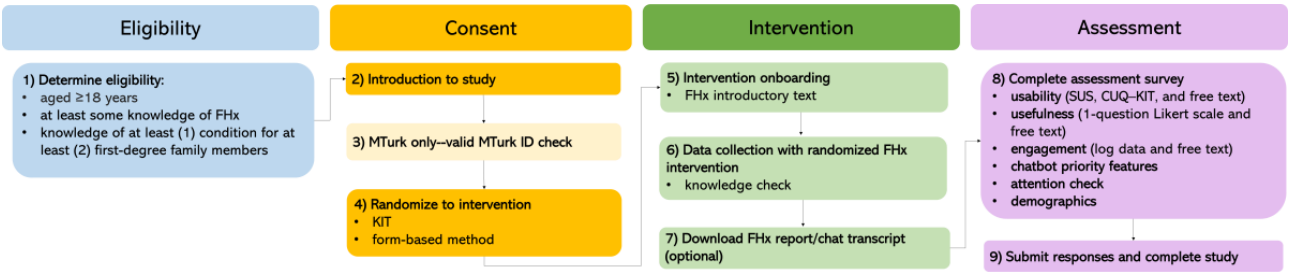

Ethical Considerations

The study was reviewed under the Expedited designation and approved by the institutional review board at Johns Hopkins University under protocol IRB00276421. Because this study was of minimal risk, participants consented with a waiver of documentation of consent that covered key study information, study purpose, procedures, risk and benefits, voluntary participation, payment, and study personnel contact information. Study data were anonymous. Amazon MTurk worker participants were paid US $3 for completing the survey. Qualtrics Panels’ participants were compensated based on prior compensation preferences (for example, participants can indicate that they would like to be compensated for SkyMiles, retail points, cash, or gift cards). Qualtrics Panels participants were informed of their compensation when invited to take the survey. Data collection was conducted from June 2022 through March 2023.

XSL•FO
RenderX

We completed a STARE-HI (Statement on Reporting of Evaluation Studies in Health Informatics) checklist and included the completed checklist in Multimedia Appendix 4 [34,35].

Data Collection Procedures

We were interested in studying the user experience (UX) of the 2 FHx tools. The importance of UX (including usability) in technology acceptance is well recognized according to existing theoretical models (eg, Technology Acceptance Model and Task-Technology Fit) [36,37]. The assessment survey used to study UX described in Multimedia Appendix 3 collected quantitative data used to study the usability of the FHx data collection tools, the usefulness of the reports downloaded from those tools, and preferences to prioritize proposed chatbot features. It also collected qualitative data via free-text comments in response to questions about FHx tool features they liked or disliked. To study engagement, we passively collected log data exported from Juji and Qualtrics (for KIT and the form-based FHx data collection intervention, respectively).

First, to assess the usability of KIT and the form-based method, we analyzed assessment survey items based on the standardized System Usability Scale (SUS) [36] and the Chatbot Usability Questionnaire (CUQ) [38]. The SUS is a 10-question survey with a 5-point Likert response scale, ranging from "strongly disagree" to "strongly agree." We adapted the SUS for our use case by removing 1 question (question 1: "I think I would like to use this system frequently") that was not relevant to our context; the SUS score calculation was modified accordingly [39]. To capture chatbot users' feedback, we included 16 items from CUQ in the assessment survey only for those participants randomized to the KIT intervention.

Second, we used 1 assessment survey item to assess the usefulness of reports downloaded from the FHx data collection tools for primary care visits. That item was "Do you think the final report from KIT would be useful to share with your primary care provider?" This was measured on a 5-point Likert scale.

Third, we aimed to assess participant interactions with the FHx data collection tools to understand if there were differences in UX and duration of experience between the tools (ie, engagement measures). We used log data on the duration of tool use, resource use (ie, button clicks and questions asked), and several conditions reported to measure engagement quantitatively.

Fourth, to supplement the quantitative data collected to study tool usability, report usefulness, and engagement, we included 2 free-text questions in the assessment survey about which features they liked or disliked in the collection tool they used. The free text was intended to capture a more comprehensive understanding of usability, report usefulness, and engagement that could not be captured through quantitative data alone. For example, for engagement, we could collect data on feelings associated with the FHx tool (such as interest and involvement) that went beyond the log data in the engagement quantitative data.

Finally, to assess priorities to enhance KIT in future iterations, we asked users to rate 3 chatbot-specific features that, if included in future chatbot design, would improve FHx collection. These features could potentially improve chatbot UX and value to users in health care contexts, as demonstrated by previous studies [40-45]. The included chatbot-specific features (with potential examples) were chatbot personalization (providing more education or tailored responses based on user entries), media elements (photos, videos, and graphics interchange formats as chatbot responses), and gamification (point reward system for engaging with the chatbot, complete FHx collection, and sharing results with family). For each feature, the participants picked 1 of 3 response options (high, medium, or low priority).

Mixed Methods Data Analyses

For the quantitative analyses, we compared differences in mean *usability*, *engagement*, and *usefulness* measures between the 2 FHx data collection tools using data from all study participants who passed data quality checks. Participants were removed from the final analysis if they had low-quality UX comments data. The significance of the differences was assessed with the Welch 2-sample 2-tailed *t* test with a 95% CI. For individuals randomized to KIT, we reported the number and percentage of individuals who agreed ("strongly agree" or "agree"), disagreed ("strongly disagree" or "disagree"), and responded neutrally to each item from the CUQ. As an exploratory data analysis, we performed multiple linear regression to explore whether participants' engagement with the FHx tools (as measured from log data) influenced their reported SUS usability scores. For 3 proposed chatbot feature enhancements (gamification, media elements, and personalization), we assessed *enhancement prioritization* based on the frequency of individuals indicating a high or medium priority from the largest (rank 1) to the smallest.

For the qualitative data analyses of *usability*, *engagement*, and *usefulness*, the 3 research team members (MHN, JS, and COT) used a set of definitions for constructs relevant to deductive coding. The definitions were aligned with the quantitative metrics captured for this study. They were further adapted for qualitative analysis by drawing from factors in measurement tools in human-computer interaction and information technology design settings [38,46,47]. For example, the definition used for usability-related comments captured "effectiveness (ability of users to complete tasks using the system, and the quality of the output of the tasks), efficiency (level of resource consumed in performing tasks), and satisfaction (users' subjective reactions to using the system)," derived from the SUS definition of usability [38]. In addition, 2 coders mapped free-text comments shared by study participants on usability, engagement, and usefulness constructs. When concordance between the 2 coders was not achieved, the third research team member served as arbiter. We also noted comments on *desired features* made by both intervention groups (KIT and form). After we reached a consensus on theme categorization, author MHN reviewed comments within each theme to organize the data for subthemes and key concepts. Counts were reported for comments categorized under each subtheme and concept.

All quantitative analyses were conducted using R (version 4.2.0; R Foundation for Statistical Computing), and qualitative data were analyzed in Microsoft Excel

XSL•FO
RenderX

(Microsoft Corp). In the last stage of the analysis, the qualitative responses relevant to usability, engagement, and report usefulness were compared with findings from the quantitative analyses. Doing this provided additional context for our quantitative findings.

Results

Study Population

A total of 138 MTurk workers and 445 Qualtrics Panel participants were randomized to an FHx data collection intervention. Of these, 38 (27.5%) MTurk workers and 332 (74.6%) Qualtrics Panel participant responses were not included because they had duplicate responses, knowledge check failures, or failed to complete the study. Details on the Qualtrics Panels sample exclusions can be found in the in Multimedia Appendix 5. Table 1 summarizes self-reported demographic information. There were no significant differences between the form and KIT populations. The total cohort was primarily composed of female participants (109/213, 51.2%), aged ≥45 years (127/213, 59.6%), non-Hispanic or Latino White (164/213, 77%), and with a bachelor degree or higher (108/213, 50.4%). The largest share of participants was located in the Southeast region of the United States (69/213, 32.4%).

Table 1. Participant characteristics of the total cohort—Amazon Mechanical Turk and Qualtrics panels.

	Total (N=213), n (%)	Form based (n=112), n (%)	KIT[a] (n=101), n (%)	*P* value
Gender				.30
Women	109 (51.2)	52 (46.4)	57 (56.4)	
Men	100 (46.9)	57 (50.9)	43 (42.6)	
Nonbinary	2 (0.9)	1 (0.9)	1 (1)	
Prefer not to say	2 (0.9)	2 (1.8)	0 (0)	
Age (y)				.23
≤44	84 (39.4)	39 (34.8)	45 (44.6)	
≥45	127 (59.6)	71 (63.4)	56 (55.4)	
Race				.44
Non-Hispanic or Latino White	164 (77)	86 (76.8)	78 (77.2)	
Other ethnicities or race	44 (20.7)	22 (19.6)	22 (21.8)	
Missing	5 (2.3)	4 (3.6)	1 (1)	
Region				.51
Midwest (Iowa, Illinois, Indiana, Kansas, Michigan, Montana, Missouri, North Dakota, Nebraska, Ohio, South Dakota, and Wisconsin)	41 (19.2)	25 (22.3)	16 (15.8)	
Northeast (Connecticut, District of Columbia, Delaware, Massachusetts, Maryland, Maine, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont)	36 (16.9)	19 (17)	17 (16.8)	
Southeast (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia)	69 (32.4)	32 (28.6)	37 (36.6)	
Southwest (Arizona, New Mexico, Oklahoma, and Texas)	26 (12.2)	16 (14.3)	10 (9.9)	
West (Alaska, California, Colorado, Hawaii, Idaho, Montana, Nevada, Oregon, Utah, Washington, and Wyoming)	41 (19.2)	20 (17.9)	21 (20.8)	
Education				.64
Bachelor's or higher	108 (50.7)	57 (50.9)	51 (50.5)	
Lower than a bachelor's degree	101 (47.4)	53 (47.3)	48 (47.5)	
Prefer not to say	1 (0.5)	1 (0.9)	0 (0)	

[a]KIT: the curious interactive test.

Table 2 provides a quantitative analysis of summary statistics of mean and median usability, engagement, and usefulness measures according to the FHx data collection intervention type. All following results tables exclude participants from analysis

XSL•FO
RenderX

who had low-quality UX comments results –4/101 KIT participants and 4/112 form-based participants.

Table 2. Combined cohort results for usability, engagement, and usefulness outcomes for KIT[a] as compared with the form-based family health history tool.

Outcome and measure	Intervention type		*P* value
	Form based (n=108)	KIT (n=97)	
Usability			
SUS[b] scores			<.001[c]
Mean (SD)	61.9 (9.16)	80.2 (17.6)	
Median (IQR)	61.1 (58.3-72.2)	86.1 (66.6-94.4)	
CUQ[d] scores			N/A[f]
Mean (SD)	—[e]	79.4 (13.6)	
Median (IQR)	—	81.3 (1.50-4.70)	
Engagement			
Interaction duration			.04[g]
Mean (SD)	7.97 (8.30)	5.90 (5.79)	
Median (IQR)	5.90 (3.85-8.38)	4.00 (3.00-7.00)	
Responses per minute			.85
Mean (SD)	3.26 (2.26)	3.21 (1.91)	
Median (IQR)	2.93 (1.29-4.66)	3.14 (1.50-4.70)	
Number of reported conditions			.04[g]
Mean (SD)	10.1 (7.53)	7.77 (8.55)	
Median (IQR)	9.00 (5.00-14.00)	6.00 (2.00-9.00)	
Character length of comments about liked features			.40
Mean (SD)	62.7 (41.3)	67.5 (40.6)	
Median (IQR)	55.5 (35.0-82.0)	59.5 (39.0-86.5)	
Character length of comments about disliked features			.88
Mean (SD)	65.4 (50.9)	66.4 (42.1)	
Median (IQR)	51.5 (33.0-83.0)	53.0 (37.0-84.0)	
Usefulness			
Perceived usefulness of final report			.13
Mean (SD)	4.29 (0.865)	4.08 (1.05)	
Median (IQR)	4.00 (4.00-5.00)	4.00 (4.00-5.00)	

[a]KIT: the curious interactive test.
[b]SUS: System Usability Score.
[c]*P*<.001.
[d]CUQ: Chatbot Usability Questionnaire.
[e]Not applicable for the form-based group.
[f]N/A: not applicable because the form-based group did not complete the Chatbot Usability Questionnaire.
[g]*P*<.05.

Tables 3 and 4 summarize the qualitative analysis of positive and negative comments mapped to usability, engagement, and usefulness concepts for each FHx data collection intervention. Table 3 summarizes form-based qualitative results, and Table 4 summarizes KIT qualitative results.

XSL•FO
RenderX

Table 3. Form: themes and subthemes from qualitative analysis of form-based user experience comments[a].

Interventions, themes, and subthemes, and concepts (+/–; n)	(+/–) Example comments
Usability	
Ease of use (53)	
• Ease of use (+; 44) • Ease of understanding (+; 7) • Difficult to navigate (–; 1) • Too technical (–; 1)	• +: "It was easy and fast. I felt like I was able to quickly and accurately input the info as someone who already was confident in their family medical history." • –: "The method to select the health issue could be more explicit; use a check box."
Specific functionality (5)	
• Unclear or inconvenient functionality (–; 5)	• –: "On each page I could still select 'None of the above' even if I selected a disease"
Usefulness	
Report-related comments (8)	
• Can be shared with family or family would use (+; 2) • Report allows you to see family history in total (+; 2) • Report is not useful (–; 1) • Report is too long (–; 1) • Report may not be accurate (–; 1) • Report download does not work (–; 1)	• +: "I liked that the tool was simple to use. It was also easy to read, and I could see my other family members using it as well." • +: "The report should summarize the results in a nice concise format, as opposed to the lengthy questionnaire." • –: "Tried to download collection tool and it didn't work."
Engagement	
Overall design and organization (41)	
• Comprehensive (+; 4) • Conditions are provided (+; 3) • General positive comments about interface (+; 10) • Awkward user interface (–; 1) • Condition list is not comprehensive enough (–; 10) • Intrusive (–; 2) • Too long (–; 3) • Quick (+; 8)	• +: "They cover a very comprehensive range of content that is easy to use" • +: "The answers were already there. I just had to choose from the available options instead of thinking of each possible disease myself. It was straightforward. Very easy and unobtrusive." • +: "I like the feature that each disease is asked separately instead of being cluttered in a table." • –: "I kept having to go back and forth to make my answers correct because I kept realizing how I was clicking it incorrectly. I'm highly computer savvy and this was awkward."
Comments about specific functionality (43)	
• Information buttons (+; 13) • Multiple-choice format (+; 6) • Desire for additional information (–; 5) • Desire for additional functionality (ie, to add more detailed information; –; 7) • No self-reporting included (–; 12)	• +: "I liked that it was easy and there wasn't any ambiguity. I also liked the info buttons so you could read about the disease if you weren't sure what it was." • –: "The person completing the survey tool could not insert information about themselves." • +: "No typing required; mouse click only was easy." • –: "I didn't like that there were no ways to answer further." • –: "I would have liked to report more information when reporting the diseases my grandparents had. I felt like I needed to put in more information or would have liked to explained more, such as some of the conditions were the result of having surgery."

[a]Count is not mutually exclusive (individual likes and dislikes can be related to multiple themes).

XSL•FO
RenderX

Table 4. KIT[a]: Themes and subthemes from qualitative analysis of KIT user experience comments.

Interventions, themes, subthemes, and concepts (+ or –; n)	(+ or –) Example comments
Usability	
Ease of use (44)	
• General ease of use (+; 30) • Ease of understanding (+; 4) • Quick (+; 10)	• +: "I liked that it listed out the different conditions to click, rather than having me type them in since medical terminology can be difficult to spell. Overall, I just liked how easy it was to use." • +: "I liked the speed of it. I liked how easy it was to answer her questions and some of the suggestions to her questions were nice also."
Usefulness	
Report-related comments (10)	
• Transcript is difficult to read (–; 7) • Download functionality issues (–; 1) • Unclear purpose (–; 1) • Desire for personalized response (–; 1)	• –: "The final report was a bunch of gibberish and unreadable." • –: "I was unsure how exactly this information would be used relative to my family history." • –: "I thought the questions were somewhat generic, and no information based on my responses was provided."
Engagement	
Chatbot personality (27)	
• Conversational (+; 1) • Engaging (+; 1) • Nonpersonal (+; 1) • Understanding (+; 1) • Able to elicit honest responses (+; 1) • Informative (+; 6) • General positive comments about personality (+; 6) • Robotic personality (–; 10)	• +: "...I like that it was generally 'friendly' acting, but still had the distance of being a chat bot, so I felt less awkward about my answers/potential for mistakes, etc" • +: "Engaging, thorough clarification by Kit and very easy to ask questions" • –: "The bot seem very impersonal and formal."
Overall design and organization (56)	
• Conditions are provided (+; 17) • Multiple-choice buttons (+; 8) • FHxc collection is guided (+; 5) • Not comprehensive enough (–; 1) • Functionality issues (–; 3) • Desire for additional information (–; 10) • Lack of functionality to add more details (–; 7) • Desire for personalized response (–; 5)	• +: "I liked to see all the options for the health problems. I was able to recall family problems that I initially didn't think of." • +: "I felt that the multiple-choice questions were good and provided a safe space to be truthful about personal family business." • +: "I felt it guided me through collecting a family history quite well; it was easy to focus on one relative and disease at a time." • –: "KIT kept freezing and I had to refresh 3 or 4 times to reactivate her." • –: "I think descriptions of what each disease was on hover would be useful, or more options to choose from." • –: "I wish there would have been a little more info about the health conditions and my risks provided while we talked."
Conversation pace (19)	
• Quick (+; 10) • Too fast (–; 6) • Too slow (–; 3)	• +: "Able to answer my questions quickly" • – "It felt like it moved fast, and even though I could have scrolled up/down, delayed answering, etc. I still felt like I was being rushed through the process." • – "There was a lot of information to read at once that was sent quickly and I had to scroll up to start reading at the beginning." • – "It seemed a little slow to respond."

[a]KIT: the curious interactive test.

Usability

The quantitative usability analyses showed a significant difference ($P<.001$) in median SUS scores, with a median score of 61.1 for form respondents and 86.1 for KIT participants (Table 2). Both the quantitative CUQ questionnaire results and the qualitative findings gave insights into the positive features contributing to SUS scores.

Aligning with KIT participants' high SUS scores, KIT users reported a mean CUQ score of 79.4 (SD 13.6). Findings from analyzing CUQ questionnaires (Table 5) indicated the high (positive) scores for questions related to the ease of use (90/97, 93% agreed with P8 and 81/97, 84% disagreed with N8), ease of navigation (90/97, 93% agreed with P4 and 81/97, 84% disagreed with N4), ability to answer (75/97, 77% agreed with P6 and 81/97, 84% disagreed with N6), and understanding user

XSL•FO
RenderX

prompts (78/97, 80% agreed with P5 and 82/97, 85% disagreed with N5). Similarly, the qualitative analyses (Form: Table 3 and KIT: Table 4) indicated that both KIT and form users valued the ease of use, citing ease of understanding and speed as positive attributes of the 2 methods. For example, one respondent noted:

> *I liked the speed of it. I liked how easy it was to answer her questions, and some of the suggestions to her questions were nice also.* [Table 4]

Qualitative findings clarified negative issues affecting usability, showing that form users noted the form being unclear or having inconvenient functionality (5 respondents; eg, "The method to select the health issue could be more explicit; use a check box"), being difficult to navigate (1 respondent), and being too technical (1 respondent; Table 3). For the KIT method, no negative features were noted as comments for usability.

Table 5. CUQ[a] responses by item[b] (n=97).

CUQ item number	CUQ item	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
Positive constructs				
P1	KIT's[c] personality was realistic and engaging	56 (58)	32 (33)	9 (9)
P2	KIT was welcoming during the initial setup	77 (79)	16 (17)	4 (4)
P3	KIT explained its scope and purpose well	51 (53)	14 (14)	32 (33)
P4	KIT was easy to navigate	90 (93)	6 (6)	1 (1)
P5	KIT understood me well	78 (80)	16 (17)	3 (3)
P6	KIT's responses were useful, appropriate, and informative	75 (77)	17 (18)	5 (5)
P7	KIT coped well with any errors or mistakes	44 (45)	42 (43)	11 (11)
P8	KIT was very easy to use	90 (93)	3 (3)	4 (4)
Negative constructs				
N1	KIT seemed too robotic	60 (62)	20 (21)	17 (18)
N2	KIT seemed very unfriendly	46 (47)	7 (7)	44 (45)
N3	KIT gave no indication as to its purpose	92 (95)	3 (3)	2 (2)
N4	It would be easy to get confused when using KIT	81 (83)	7 (7)	9 (9)
N5	KIT failed to recognize a lot of my inputs	82 (85)	8 (8)	7 (7)
N6	KIT's responses were not relevant	81 (84)	13 (13)	3 (3)
N7	KIT seemed unable to handle any errors	63 (65)	27 (28)	7 (7)
N8	KIT was very complex	81 (84)	9 (9)	7 (7)

[a]CUQ: Chatbot Usability Questionnaire.

[b]Items were measured on a 5-point Likert scale. Agree and disagree responses are grouped.

[c]KIT: the curious interactive test.

Engagement

Findings from analyzing engagement were mixed. A subset of the CUQ items used to assess KIT captured engagement-related concepts (Table 5).

In particular, we found that users responded more neutrally or negatively to CUQ questions related to KIT's onboarding (14/97, 14% were neutral and 32/97, 33% disagreed with P3; less neutral or negative sentiment with N3), error management (42/97, 43% were neutral and 11/97, 11% disagreed with P7; 27/97, 28% were neutral and 7/97, 7% agreed with N7), and KIT's personality (33/97, 33% were neutral and 9/97, 9% disagreed with P1; 20/97, 21% were neutral and 17/97, 18% agreed with N1; 7/97, 7% were neutral and 44/97, 45% agreed with N2), as presented in Table 5. User comments about error handling and personality were also present in the qualitative data. Users mentioned functionality issues (3 respondents; Table 4), such as KIT crashing, which might add context to the CUQ error handling questions (CUQ items P7 and N7).

In addition, participants shared a mix of positive and negative sentiments about KIT's personality. In total, 10 KIT respondents identified KIT's robotic personality in the qualitative findings, for example, remarking that "The bot seemed very impersonal and formal," which aligns with the neutral or negative CUQ finding for personality questions (CUQ items P1 and N1). However, most (17/27, 63%) user comments related to KIT's personality were positive, contrasting with the CUQ finding. One such comment mentioned, "...I like that it was generally 'friendly' acting, but still had the distance of being a chat bot, so I felt less awkward about my answers/potential for mistakes, etc.," in which they found the nonpersonal nature of KIT to be a positive feature. In total, 2 form users felt that the questions

XSL•FO
RenderX

were intrusive, which was not a comment found in KIT users' responses.

We quantitatively compared engagement log data between KIT and the form-based method and found significant differences in 2 measures, interaction duration (mean of form 7.97 minutes, KIT 5.90 minutes; *P*=.04) and the number of reported conditions (mean of form 10.1, mean of KIT 7.77; *P*=.04), as presented in Table 2. We examined user comments for features that influenced interaction time or speed, contributing to the significant quantitative findings. From the qualitative data, we found that 6 (6%) of the 97 KIT participants felt the conversational pace of KIT was too fast with 1 respondent noting, "It felt like it moved fast, and even though I could have scrolled up/down, delayed answering, etc. I still felt like I was being rushed through the process" (Table 4). The conversational pace of KIT, as identified from the qualitative data, may have contributed to the significant quantitative finding in interaction duration.

Both KIT and form users said they wanted additional features or functionality within the FHx data collection tools. In total, 5 (4.6%) of the 108 form participants and 10 (10%) of the 97 KIT participants wanted the FHx tool to present additional information (example from a KIT participant, "I think descriptions of what each disease was on hover would be useful, or more options to choose from"). In addition, 7 (7%) form participants and 7 (7%) KIT participants expressed the desire for additional functionality. For form users, examples of this functionality included adding other family members to the history record (ie, aunts and uncles, "...I think we should add Father and Mother's Siblings") and adding more details surrounding a family member or condition, that is, specify which side of the family for a grandparent, specify how the condition occurred, or was diagnosed:

> *I would have liked to report more information when reporting the diseases my grandparents had. I felt like I needed to put in more information or would have liked to explained more, such as some of the conditions were the result of having surgery.*

KIT users wanted the functionality to add more detailed and open-ended information about their conditions (eg, "I wish the choice of other allowed the consumer to type in that box") and to add more details about their family members, like form users (eg, "there could have been more specifics on which family member").

Some desired features were unique to form or KIT participants. In total, 12 (11%) of the 108 form participants expressed the desire to include themselves in the FHx record:

> *I thought it would have me listed too, which it didn't. I was surprised I couldn't put down what things I have or have had.*

In total, 5 (5%) of the 97 KIT users expressed a desire for a personalized response from KIT, with 1 respondent mentioning, "I wish there would have been a little more info about the health conditions and my risks provided while we talked."

As an exploratory analysis, to understand whether participants' engagement with the FHx tools influenced their reported SUS usability scores, we performed multiple linear regression, including responses per minute, UX comments, likes and dislikes, length, duration, cohort (Qualtrics Panels and MTurk), and intervention type (KIT and form based) as covariates. Responses per minute were significantly correlated with SUS scores, β=1.193; t_{198}=2.509; *P*=.02. Additional details are in Table S2 in Multimedia Appendix 5.

Report Usefulness

Analysis of the usefulness of the final reports showed that both form and KIT respondents gave a median score of 4 in response to the Likert scale question regarding report usefulness, indicating that participants somewhat agreed that the final reports were useful. There was no significant difference found between groups, with mean scores of 4.29 for form-based users and 4.08 for KIT users (*P*=.13). The qualitative user comments from form users highlighted 2 features that may have made the form-based FHx report useful: the report could be shared with family members (2 respondents), and the report allowed you to see family history in total (2 respondents). There were no positive comments from KIT users in the qualitative data.

For both groups, the qualitative data provided more specific reasons why participants did not "strongly agree" (median score of 5) to the report usefulness question in the quantitative findings. Both user groups noted common issues with the functionality and general use of both FHx reports. For functionality, 1 participant from each FHx tool commented on the difficulty of downloading the reports. For general use, some participants desired FHx report features beyond a simple copy of their interaction. For example, 1 form user suggested a summary report format (eg, "The report should summarize the results in a nice concise format, as opposed to the lengthy questionnaire "; Table 3). KIT respondents had more negative sentiments about the report's readability (7 user comments), which did not appear in form participant responses. However, 1 form user commented that the report was too long. Unlike form users, KIT users also desired a personalized response and clearer purpose for the final report (2 comments in Table 4, eg, "somewhat generic and no information based on my responses was provided"), suggesting this would improve upon the current usefulness of the report. While KIT users had no positive comments about the report's usefulness, some form users expressed that the report provided them with a snapshot of their family history and that they could imagine their family members using it as well (4 comments in Table 4).

Enhancement Prioritization of Proposed Chatbot Features

Both KIT and form-based users reported their opinions on proposed chatbot feature enhancements to prioritize for chatbot design. On the basis of the responses in Figure 4, they were ranked as follows: 1-personalization, 2-media elements, and 3-gamification. Notably, for all features, most respondents reported that they would prioritize them, with 55.6% (114/205), 60% (121/205), and 91.7% (188/205), indicating high or medium priority for gamification, media elements, and personalization, respectively.

XSL•FO
RenderX

Figure 4. Prioritization of proposed chatbot feature enhancements: gamification, media elements, and personalization.

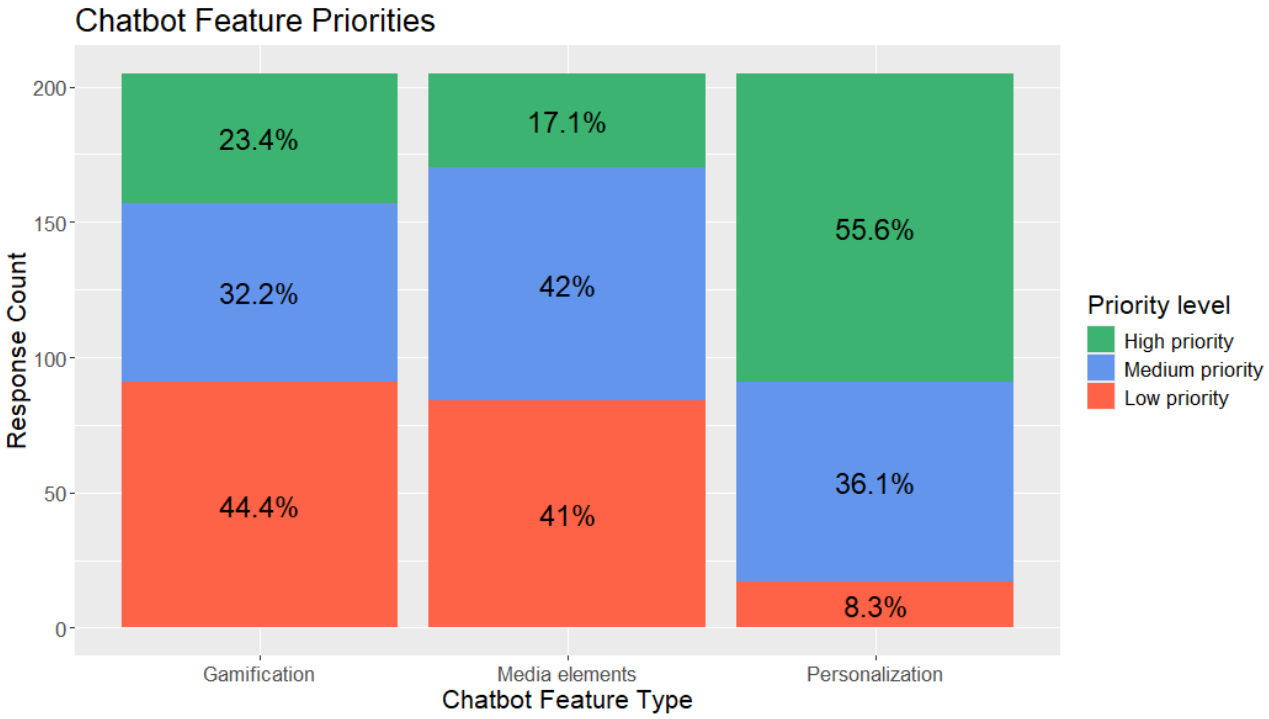

Discussion

Principal Findings

Patient-facing digital health approaches to collect FHx are gaining acceptance and adoption due to their advantages over paper-based collection methods and their potential to support more convenient data collection, EHR integration, and automated risk assessment [13,48]. Currently, chatbots are being explored for clinical use as new ways to collect patient data and deliver care asynchronously, with use cases ranging from nutrition journaling to cognitive behavioral therapy delivery [43,44,49]. We purposefully designed our study to enable a fair comparison of KIT to the form-based FHx data collection. Both used the same data collection items and knowledge base for educational content (refer to the Intervention Design section). We focus our discussion on areas where our research provides insight into chatbots' potential benefits over standard form-based methods and also discuss areas for enhancement. Where benefits exist, we further describe how specific features may improve chatbot usability and thus should be prioritized for future development.

Overall, we found that KIT was more usable than the form-based method (Table 2). There were some significant differences in 2 engagement metrics between the form-based method and KIT; the form elicited more reported conditions, and form users had longer interaction durations on average. We found no differences between KIT and the form-based method of FHx data collection for 3 engagement metrics (response per minute, character length of comments about liked features, and character length of comments about disliked features) and for the usefulness of the final report. Qualitative findings indicated that mismatches between what was implemented and expected functionalities may be an important factor distinguishing KIT from form-based FHx collection, with more respondents randomized to the form having more negative comments about this mismatch (eg, "On each page, I could still select 'None of the above,' even if I selected a disease"; Table 3). Further analyses of KIT indicated that participants felt it was easy to use and navigate, and they were able to understand user prompts and provide appropriate responses (Table 5). However, respondents were more neutral or showed a negative sentiment for KIT's personality, onboarding process, and its management of errors. Despite this finding, qualitative findings indicated that when there were comments about KIT's personality, they were positive, and in fact, its perceived nonhuman nature may be an added advantage (eg, "...I like that it was generally 'friendly' acting, but still had the distance of being a chat bot, so I felt less awkward about my answers/potential for mistakes, etc.") We also proposed 3 areas to enhance KIT and found that all 3 (gamification, media elements, and personalization) were highly endorsed, with >50% (114/205, 121/205, 188/205, respectively) of the respondents indicating a high or moderate ranking for each of the features (Figure 4). According to our findings, the priority ranking of features was 1-personalization, 2-media elements, and 3-gamification.

Together, our findings provide valuable insight into design considerations for implementing a chatbot for FHx collection, including the potential benefit of its use over form-based methods. In the subsequent sections, we combine our findings with the broader literature exploring the use of conversational agents for FHx data collection.

What Conversational Agents Can Add to FHx Data Collection

This study shows that conversational agents or chatbots offer specific functionality that may enhance FHx data collection and emphasize features to prioritize for future FHx tool development. We intentionally did not introduce many chatbot-specific

XSL•FO
RenderX

features into KIT's design to observe participants' perceptions of a chatbot UX at a baseline level. Still, participants found KIT more usable than the form-based method ($P<.001$).

This distinction may be partly due to KIT's conversational way of collecting FHx from users. Soni et al [26], for example, found that a chatbot used to complete a standardized health assessment was perceived as conversational, interactive, and intuitive, which may be important characteristics contributing to its usability. Most participants in that study reported a preference for a chatbot over a form-based strategy. In contrast to our work, however, they found no statistically significant differences in SUS scores when comparing the 2 groups [26]. In another study comparing a chatbot-based to a form-based survey format, Xiao et al [50] found that the chatbot elicited clearer, more relevant, and more specific responses than the form. In particular, the authors found higher levels of self-disclosure from users of the chatbot in comparison to form users, suggesting that chatbots may support higher-quality data collection by promoting interactive user engagement [50].

Another potential value of chatbots over other formats is their humanness. Despite limited chatbot-specific additions, a small portion of KIT users anthropomorphized KIT in their free-text comments, with about 18 participants referring to KIT as "she," "her," or by name in their comment (as opposed to "it," or "the chatbot," etc). This finding may reflect the KIT avatar we selected (feminine-presenting nurse). Participant self-disclosure may increase when chatbots have more humanization features applied to their design [51]. From a study of the 4th *All of Us* data set release, Sulieman et al [52] demonstrated that patient self-reported surveys contained data not found in the EHR and could help potentially fill in missing EHR data gaps [52]. Thus, we believe that designing FHx tools that encourage higher self-disclosure, such as chatbots, may further enhance the utility of patient-reported information to fill in gaps in EHR-documented information.

Comparing the form-based method with KIT, there are differences in the pathway to satisfy the users' health information seeking. Chatbots allow for further questioning beyond a single info button. In a previous study comparing a FHx collection virtual conversational agent to the Surgeon General's My Family History Portrait web tool, study participants preferred the dialogue or chat-based guidance of the chatbot over the standard web tool [53]. Chat-based guidance may be better equipped to support additional just-in-time health information seeking than an info button would. Health information seeking has been linked to more frequent discussions of FHx between patients and their families and with their primary care providers [54]. Better support for patients' health information seeking needs within the FHx collection tool could encourage patients to be more likely to discuss FHx inside and outside the clinic.

Both form and KIT users responded positively to the multiple-choice button format, in which users were presented with answer options and could simply click the relevant answer options for data entry. This suggests that despite the chatbot format, users may still appreciate being shown response options, and this could reduce some working memory requirements for respondents and may be useful for medical history-taking for more accurate patient recall, drawing principles from cognitive load theory [55]. In a previous study of chatbot input mechanism (buttons vs free-text responses) on perceived pragmatic and hedonic interaction quality, chatbots with buttons were associated with both high usability (pragmatic quality) and engagement (hedonic quality) in comparison to chatbots accepting free-response answers only [56]. Moreover, button-input chatbots were not associated with less perceived anthropomorphism or social presence of the chatbot [56].

For both formats, users wanted the ability to answer further depending on the options they selected and to record medical history for themselves as well. These limitations stemmed from implementing the *All of Us* Family History Questionnaire outside of the *All of Us* Program context, as this family history questionnaire was designed to be administered in addition to other basic history surveys and EHR data integration [57]. However, for a stand-alone FHx tool, supporting users' additional information would be beneficial. Chatbots may be especially capable of supporting further open-ended questions based on user input [50].

Both KIT and form users appreciated the information provided in the FHx tools. This adaptation to the *All of Us* questionnaire was not present in the original. By providing users with educational information about unfamiliar diseases, we sought to support their health literacy, potentially leading to more accurate data collection. Although there is limited literature on the effect of low health literacy on the accuracy of patient-reported health data, there is a great deal of evidence that low health literacy is linked to poorer health outcomes [58-60].

The Final Report Is an Opportunity to Add Value to Patients

User feedback about the value of the FHx report was mixed. Although quantitative findings were positive, from qualitative user comments, we have found that the usefulness of the reports was limited by their awkward formatting, unclear purpose, and lack of value added beyond the scope of the data collection intervention (eg, summary table and personalized response). From this, we discuss motivations and design considerations for developing future FHx reports to add value for patients.

We were motivated to explore FHx report usefulness because of information gaps for both patients and providers. We sought to understand the value of FHx and to support the actionable use of FHx in care and condition management. In addition, we sought to better understand user desires for reports based on evidence that other FHx tools have developed patient-facing or clinician-facing reports. From a review of 62 generic and cancer-specific FHx tools, Cleophat et al [11] found that about one-fourth of FHx tools provided patient-facing recommendations after data collection, and about one-tenth of tools provided clinician-facing management recommendations. While some available FHx tools only provide clinician-facing decision support or no guidance at all, many provide patient-facing decision support through recommendations or risk assessments based on collected FHx data [11,12]. However,

XSL•FO
RenderX

few studies highlight what specific features of FHx reports are valuable to patients.

Our work adds to the literature to identify FHx report features to improve usefulness for a patient's primary care visit. For form users, their final report was a summary of their responses that they could choose to download. For KIT users, their final report was a time-stamped transcript of their responses as a text file they could download. We relied on the host interface's (Qualtrics and Juji) baseline capabilities for both web-based tools without custom formatting.

In future FHx tool design, it is important to introduce what the use of the final report might be at the beginning of the intervention, as well as to customize it for consumer use intentionally and to understand patient preferences and attitudes toward features, such as personalized FHx risk assessment. Previous work has designed patient-facing genetic test reports to maximize patient understanding [61,62], and many of the lessons learned in the structure, content, communication style, and visual design of these reports can be applied to patient-facing FHx reports.

Balance Between Participant Burden and Data Quality With FHx Collection

Variation in the onboarding process implemented in this study highlighted a balance between participant burden and the quality of FHx data collected. For the form, the only onboarding step was to review instructions. Participants were then presented with each of the 11 condition groups. Each condition group was asked to indicate whether any conditions within that group were present in their family, and if so, for what family members. They were asked this information regardless of whether they had any family members with a relevant history (Multimedia Appendix 1). For KIT, there were 2 onboarding steps. Following the instructions, users were presented with the initial list of 11 condition groups, from which they were prompted to select only those present (or known to be present) in their family. This process was modified based on KIT user comments during the pilot phase of this study (Multimedia Appendix 2) to eliminate the need to see irrelevant condition groups.

The number of reported conditions was significantly different between the 2 intervention groups, with form-based users reporting more conditions on average than KIT users, respectively (10.1 vs 7.77; P=.04). This finding suggests that differences in onboarding may have impacted the number of reported health conditions. While the onboarding strategy used in KIT may have reduced the burden for the users by decreasing the number of data items to collect, there was also a potential impact on the quality of FHx data collected. For example, there may have been instances where a specific condition name was known, but the condition group name was unknown. If the group name was not selected, the user would never see the specific condition found in their family. Therefore, the occurrence of missing FHx data could be higher. In future work, the chatbot onboarding strategy should be an important design consideration influencing data quality, and it could be particularly important when implementing chatbots for data collection.

Opportunities for Large Language Models in FHx Collection

We learned that KIT's flow-based nature may limit its interaction style. Large language models (LLMs) may help address some of the deficits identified in user comments. Personalization was the highest ranked among the 3 potential enhancements presented to participants. In addition, participants negatively responded to KIT's personality, ability to explain its purpose, and ability to manage errors. The flow-based chatbot strategy we used did not customize error handling messages or its personality features; however, these issues may be addressed by integrating LLMs into future chatbot design. In a study exploring the use of LLMs to support chatbot collection of health self-reported data, Wei et al [63] present potential benefits of using an LLM-driven chatbot due to its abilities to ask specific follow-up questions, include social attributes, recover from errors, and track conversational context. For our work, we relied on a flow-based chatbot framework to compare the interfaces as directly as possible and observe any baseline chatbot characteristics that differentiate the usability of the 2 modalities. As LLMs and generative artificial intelligence continue to advance, they are being explored for use in health care contexts. However, there are still essential implementation considerations, such as Health Insurance Portability and Accountability Act compliance, model bias, and rigorous safety evaluation before widely deploying these tools for patient care [64].

Additional Comparisons to Prior Work

This study assessed KIT's usability among a crowdsourced cohort and not among in-clinic users. For health-specific contexts, prior work has not reached a consensus on whether chatbots are more usable or preferred compared to standard web-based forms in a clinical setting.

For clinician-facing use, Iftikhar et al [65] conducted a study to assess the comparative usability of 3 web-based forms (single-page form, multipage form, and chatbot [conversational form]) to record patient referrals to a cardiology department as replacement options for paper-based forms. In their study, the chatbot was the least preferred choice and had a below-average SUS score among participating nurses, commenting that the chatbot was difficult to understand and use. For patient-facing use, Te Pas et al [66] compared a chatbot with a standard web-based form to administer a preoperative questionnaire in an outpatient clinic setting [66]. They found that chatbot users reported higher User Experience Questionnaire scores than the standard form [66]. Chatbots are becoming more prevalent in modern life; therefore, the public may perceive chatbots to be more usable as they become more familiar. Although our crowdsourced findings suggest that KIT may be a more usable method than a standard form, further work is needed to understand how patients and care providers would perceive a FHx chatbot like KIT.

Limitations

This study has some limitations. First, engagement metrics based on duration alone are not wholly informative of user involvement with a tool because they may measure inactive

XSL•FO
RenderX

time. This limitation was partly addressed by including rate-based engagement measurements (responses per minute). Using responses per minute, we did not find any significant differences between form-based and KIT users.

Second, we introduced potential chatbot enhancements (personalization, media elements, and gamification) to all study participants. For form-based users, this question was more hypothetical than for KIT users, who responded to the question after chatbot interaction. Therefore, the question may measure different constructs between the groups. However, when conducting Welch *t* tests for response differences between groups, we observed no significant statistical differences (Multimedia Appendix 5).

Third, as described earlier, there were differences in the onboarding process between the KIT and form groups. These differences may have introduced some measurement bias, leading to the observed difference in the number of reported conditions between KIT and form users.

Fourth, we did not require participants to enter their true FHx, which limited a closer comparison of the quality of collected data between the 2 tools. Although an investigation into FHx data quality collected by web-based tools was not in the scope of this study, this is an important consideration for future FHx tool development and refinement.

Finally, some researchers have raised concerns about using crowdsourcing platforms for behavioral studies and health-related research, citing poor data quality, lack of research transparency, generalizability to other population groups, and potentially inflated outcomes [61-63]. In our study, we followed best practices for crowdsourcing study design and recruited from a vetted pool of MTurk participants. We also used Qualtrics Panels as a validation cohort. There may still, however, be a bias toward a sample with more familiarity with web-based data collection methods than a general population.

Conclusions

We compared a flow-based chatbot (KIT) to a form-based method and observed that chatbot-specific features may contribute to a more usable experience. To support our comparison of the 2 methods, both incorporated FHx data collection items from the *All of Us* research program and the NLM MedlinePlus knowledge base for educational content. Our work contributes to the existing literature on how chatbots may be used for FHx collection and adds evidence of its potential benefit over form-based methods. In addition, findings from our study highlight opportunities to improve upon a final report summarizing the FHx collection experience, show that chatbot question prompts and question flow have an essential role in ultimately how complete the FHx data are, and show several areas that can be improved upon with a move to an LLM implementation strategy.

Acknowledgments

The authors would like to acknowledge support for this research. MHN was partially supported by the National Institutes of Health (NIH) and NCATS TL1 Predoctoral Clinical Research Training Program (TL1 TR003100). This work was partly supported by a NIH NHGRI Genomic Innovator award (R35 HG010714 to COT). In addition, the authors would like to thank Michelle Zhou, PhD, and Wenxi Chen, MS, from Juji, for their support during KIT's development process on Juji. The authors did not use generative artificial intelligence tools, such as ChatGPT to generate manuscript content in either the ideation or writing process.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to potentially identifiable information but are available from the corresponding author on reasonable request. The data generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MHN contributed to conceptualization, methodology, software use and development, investigation, data curation, formal analysis, writing the original draft, writing, reviewing, and editing the manuscript, and visualization. JS contributed to conceptualization, methodology, data curation, writing, reviewing, and editing the manuscript, and supervision. COT contributed to conceptualization, methodology, data curation, writing, reviewing, and editing the manuscript, supervision, project administration, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Intervention design.
[DOCX File , 131 KB - jmir_v26i1e55164_app1.docx]

Multimedia Appendix 2
Pilot results and modifications.

XSL•FO
RenderX

[DOCX File , 19 KB - jmir_v26i1e55164_app2.docx]

Multimedia Appendix 3
Assessment survey instrument.
[PDF File (Adobe PDF File), 275 KB - jmir_v26i1e55164_app3.pdf]

Multimedia Appendix 4
STARE-HI (Statement on Reporting of Evaluation Studies in Health Informatics) checklist.
[PDF File (Adobe PDF File), 497 KB - jmir_v26i1e55164_app4.pdf]

Multimedia Appendix 5
Additional results.
[DOCX File , 165 KB - jmir_v26i1e55164_app5.docx]

References

1. Moonesinghe R, Beckles GL, Liu T, Khoury MJ. The contribution of family history to the burden of diagnosed diabetes, undiagnosed diabetes, and prediabetes in the United States: analysis of the National Health and Nutrition Examination Survey, 2009-2014. Genet Med 2018 Oct;20(10):1159-1166 [FREE Full text] [doi: 10.1038/gim.2017.238] [Medline: 29369292]
2. Ginsburg GS, Wu RR, Orlando LA. Family health history: underused for actionable risk assessment. Lancet 2019 Aug 17;394(10198):596-603 [FREE Full text] [doi: 10.1016/S0140-6736(19)31275-9] [Medline: 31395442]
3. Bylstra Y, Lim WK, Kam S, Tham KW, Wu RR, Teo JX, et al. Family history assessment significantly enhances delivery of precision medicine in the genomics era. Genome Med 2021 Jan 07;13(1):3 [FREE Full text] [doi: 10.1186/s13073-020-00819-1] [Medline: 33413596]
4. Welch BM, O'Connell N, Schiffman JD. 10 years later: assessing the impact of public health efforts on the collection of family health history. Am J Med Genet A 2015 Sep;167A(9):2026-2033. [doi: 10.1002/ajmg.a.37139] [Medline: 25939339]
5. Polubriaginof F, Tatonetti NP, Vawdrey DK. An assessment of family history information captured in an electronic health record. AMIA Annu Symp Proc 2015 Nov 5;2015:2035-2042 [FREE Full text] [Medline: 26958303]
6. Wildin RS, Messersmith DJ, Houwink EJ. Modernizing family health history: achievable strategies to reduce implementation gaps. J Community Genet 2021 Jul;12(3):493-496 [FREE Full text] [doi: 10.1007/s12687-021-00531-6] [Medline: 34028705]
7. Rich EC, Burke W, Heaton CJ, Haga S, Pinsky L, Short MP, et al. Reconsidering the family history in primary care. J Gen Intern Med 2004 Mar;19(3):273-280 [FREE Full text] [doi: 10.1111/j.1525-1497.2004.30401.x] [Medline: 15009784]
8. Taber P, Ghani P, Schiffman JD, Kohlmann W, Hess R, Chidambaram V, et al. Physicians' strategies for using family history data: having the data is not the same as using the data. JAMIA Open 2020 Oct 08;3(3):378-385 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa035] [Medline: 34632321]
9. Weiner S, Amini E, Koeppe E, Resnicow K, Stoffel EM, Griggs JJ. Perceived provider barriers to collecting and documenting a complete family history in a statewide oncology consortium. J Clin Oncol 2021 Sep 21;39(28_suppl):221. [doi: 10.1200/JCO.2020.39.28_suppl.221]
10. Madhavan S, Bullis E, Myers R, Zhou CJ, Cai EM, Sharma A, et al. Awareness of family health history in a predominantly young adult population. PLoS One 2019 Oct 25;14(10):e0224283 [FREE Full text] [doi: 10.1371/journal.pone.0224283] [Medline: 31652289]
11. Cleophat JE, Nabi H, Pelletier S, Bouchard K, Dorval M. What characterizes cancer family history collection tools? A critical literature review. Curr Oncol 2018 Aug;25(4):e335-e350 [FREE Full text] [doi: 10.3747/co.25.4042] [Medline: 30111980]
12. Welch BM, Wiley K, Pflieger L, Achiangia R, Baker K, Hughes-Halbert C, et al. Review and comparison of electronic patient-facing family health history tools. J Genet Couns 2018 Apr;27(2):381-391 [FREE Full text] [doi: 10.1007/s10897-018-0235-7] [Medline: 29512060]
13. Miroševič Š, Klemenc-Ketiš Z, Peterlin B. Family history tools for primary care: a systematic review. Eur J Gen Pract 2022 Dec;28(1):75-86 [FREE Full text] [doi: 10.1080/13814788.2022.2061457] [Medline: 35510897]
14. Li X, Kahn RM, Wing N, Zhou ZN, Lackner AI, Krinsky H, et al. Leveraging health information technology to collect family cancer history: a systematic review and meta-analysis. JCO Clin Cancer Inform 2021 Jun;5:775-788 [FREE Full text] [doi: 10.1200/CCI.21.00004] [Medline: 34328789]
15. Mittendorf KF, Ukaegbu C, Gilmore MJ, Lindberg NM, Kauffman TL, Eubanks DJ, et al. Adaptation and early implementation of the PREdiction model for gene mutations (PREMM™) for lynch syndrome risk assessment in a diverse population. Fam Cancer 2022 Apr;21(2):167-180 [FREE Full text] [doi: 10.1007/s10689-021-00243-3] [Medline: 33754278]
16. Mittendorf KF, Lewis HS, Duenas DM, Eubanks DJ, Gilmore MJ, Goddard KA, et al. Literacy-adapted, electronic family history assessment for genetics referral in primary care: patient user insights from qualitative interviews. Hered Cancer Clin Pract 2022 Jun 10;20(1):22 [FREE Full text] [doi: 10.1186/s13053-022-00231-3] [Medline: 35689290]

XSL•FO
RenderX

17. Ritchie JB, Frey LJ, Lamy JB, Bellcross C, Morrison H, Schiffman JD, et al. Automated clinical practice guideline recommendations for hereditary cancer risk using chatbots and ontologies: system description. JMIR Cancer 2022 Jan 31;8(1):e29289 [FREE Full text] [doi: 10.2196/29289] [Medline: 35099392]
18. Hong G, Smith M, Lin S. The AI will see you now: feasibility and acceptability of a conversational AI medical interviewing system. JMIR Form Res 2022 Jun 27;6(6):e37028 [FREE Full text] [doi: 10.2196/37028] [Medline: 35759326]
19. Welch BM, Allen CG, Ritchie JB, Morrison H, Hughes-Halbert C, Schiffman JD. Using a chatbot to assess hereditary cancer risk. JCO Clin Cancer Inform 2020 Sep;4:787-793 [FREE Full text] [doi: 10.1200/CCI.20.00014] [Medline: 32897737]
20. Wang C, Paasche-Orlow MK, Bowen DJ, Cabral H, Winter MR, Norkunas Cunningham T, et al. Utility of a virtual counselor (VICKY) to collect family health histories among vulnerable patient populations: a randomized controlled trial. Patient Educ Couns 2021 May;104(5):979-988 [FREE Full text] [doi: 10.1016/j.pec.2021.02.034] [Medline: 33750594]
21. Khoury R, Dawood S, Khoury M, Khan F, Al Khatib F, Bello M, et al. Hereditary cancer syndromes and development of tele genetics program in the Asia-Pacific, Middle East region. Ann Oncol 2018 Nov;29(Supplement 9):ix114. [doi: 10.1093/annonc/mdy441.003]
22. Sato A, Haneda E, Suganuma N, Narimatsu H. Preliminary screening for hereditary breast and ovarian cancer using a chatbot augmented intelligence genetic counselor: development and feasibility study. JMIR Form Res 2021 Feb 05;5(2):e25184 [FREE Full text] [doi: 10.2196/25184] [Medline: 33544084]
23. Ponathil A, Ozkan F, Welch B, Bertrand J, Chalil Madathil K. Family health history collected by virtual conversational agents: an empirical study to investigate the efficacy of this approach. J Genet Couns 2020 Dec;29(6):1081-1092. [doi: 10.1002/jgc4.1239] [Medline: 32125052]
24. Ireland D, Bradford D, Szepe E, Lynch E, Martyn M, Hansen D, et al. Introducing Edna: a trainee chatbot designed to support communication about additional (secondary) genomic findings. Patient Educ Couns 2021 Apr;104(4):739-749. [doi: 10.1016/j.pec.2020.11.007] [Medline: 33234441]
25. Allen CG, Neil G, Halbert CH, Sterba KR, Nietert PJ, Welch B, et al. Barriers and facilitators to the implementation of family cancer history collection tools in oncology clinical practices. J Am Med Inform Assoc 2024 Feb 16;31(3):631-639. [doi: 10.1093/jamia/ocad243] [Medline: 38164994]
26. Soni H, Ivanova J, Wilczewski H, Bailey A, Ong T, Narma A, et al. Virtual conversational agents versus online forms: patient experience and preferences for health data collection. Front Digit Health 2022 Oct 13;4:954069 [FREE Full text] [doi: 10.3389/fdgth.2022.954069] [Medline: 36310920]
27. Dokbot - a better way to collect data from patients. Dokbot. URL: https://dokbot.io [accessed 2024-04-11]
28. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" research program. N Engl J Med 2019 Aug 15;381(7):668-676 [FREE Full text] [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
29. Welcome to MedlinePlus. MedlinePlus. URL: https://medlineplus.gov/ [accessed 2023-04-18]
30. Introduction: the Juji platform. Juji Documentation. URL: https://juji.io/docs/ [accessed 2023-04-18]
31. Créquit P, Mansouri G, Benchoufi M, Vivot A, Ravaud P. Mapping of crowdsourcing in health: systematic review. J Med Internet Res 2018 May 15;20(5):e187 [FREE Full text] [doi: 10.2196/jmir.9330] [Medline: 29764795]
32. Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, et al. Crowdsourcing--harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med 2014 Jan;29(1):187-203 [FREE Full text] [doi: 10.1007/s11606-013-2536-8] [Medline: 23843021]
33. Wang C, Han L, Stein G, Day S, Bien-Gund C, Mathews A, et al. Crowdsourcing in health and medical research: a systematic review. Infect Dis Poverty 2020 Jan 20;9(1):8 [FREE Full text] [doi: 10.1186/s40249-020-0622-9] [Medline: 31959234]
34. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI--statement on reporting of evaluation studies in health informatics. Int J Med Inform 2009 Jan;78(1):1-9. [doi: 10.1016/j.ijmedinf.2008.09.002] [Medline: 18930696]
35. Brender J, Talmon J, de Keizer N, Nykänen P, Rigby M, Ammenwerth E. STARE-HI - statement on reporting of evaluation studies in health informatics: explanation and elaboration. Appl Clin Inform 2013 Jul 24;4(3):331-358 [FREE Full text] [doi: 10.4338/ACI-2013-04-RA-0024] [Medline: 24155788]
36. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb 01;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
37. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Q 1995 Jun;19(2):213-236. [doi: 10.2307/249689]
38. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
39. Lewis JR, Sauro J. Can I leave this one out? The effect of dropping an item from the SUS. J Usability Stud 2017 Nov;13(1):38-46 [FREE Full text]
40. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The personalization of conversational agents in health care: systematic review. J Med Internet Res 2019 Nov 07;21(11):e15360 [FREE Full text] [doi: 10.2196/15360] [Medline: 31697237]

XSL•FO
RenderX

41. Liu YL, Yan W, Hu B, Li Z, Lai YL. Effects of personalization and source expertise on users' health beliefs and usage intention toward health chatbots: evidence from an online experiment. Digit Health 2022 Oct 02;8:20552076221129718 [FREE Full text] [doi: 10.1177/20552076221129718] [Medline: 36211799]
42. Bui TA, Pohl M, Rosenfelt C, Ogourtsova T, Yousef M, Whitlock K, et al. Identifying potential gamification elements for a new chatbot for families with neurodevelopmental disorders: user-centered design approach. JMIR Hum Factors 2022 Aug 19;9(3):e31991 [FREE Full text] [doi: 10.2196/31991] [Medline: 35984679]
43. Han R, Todd A, Wardak S, Partridge SR, Raeside R. Feasibility and acceptability of chatbots for nutrition and physical activity health promotion among adolescents: systematic scoping review with adolescent consultation. JMIR Hum Factors 2023 May 05;10:e43227 [FREE Full text] [doi: 10.2196/43227] [Medline: 37145858]
44. Goonesekera Y, Donkin L. A cognitive behavioral therapy chatbot (Otis) for health anxiety management: mixed methods pilot study. JMIR Form Res 2022 Oct 20;6(10):e37877 [FREE Full text] [doi: 10.2196/37877] [Medline: 36150049]
45. El Hefny W, El Bolock A, Herbert C, Abdennadher S. Chase away the virus: a character-based chatbot for COVID-19. In: Proceedings of the IEEE 9th International Conference on Serious Games and Applications for Health. 2021 Presented at: SeGAH 2021; August 4-6, 2021; Dubai, United Arab Emirates. [doi: 10.1109/segah52098.2021.9551895]
46. Brooke J. SUS: a retrospective. J Usability Stud 2013 Feb;8(2):29-40 [FREE Full text]
47. O'Brien HL, Cairns P, Hall M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. Int J Hum Comput Stud 2018 Apr;112:28-39. [doi: 10.1016/j.ijhcs.2018.01.004]
48. Wu RR, Myers RA, Sperber N, Voils CI, Neuner J, McCarty CA, et al. Implementation, adoption, and utility of family health history risk assessment in diverse care settings: evaluating implementation processes and impact with an implementation framework. Genet Med 2019 Feb;21(2):331-338 [FREE Full text] [doi: 10.1038/s41436-018-0049-x] [Medline: 29875427]
49. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
50. Xiao Z, Zhou MX, Liao QV, Mark G, Chi C, Chen W, et al. Tell me about yourself: using an AI-powered chatbot to conduct conversational surveys with open-ended questions. ACM Trans Comput Hum Interact 2020 Jun 13;27(3):1-37. [doi: 10.1145/3381804]
51. Rhim J, Kwak M, Gong Y, Gweon G. Application of humanization to survey chatbots: change in chatbot perception, interaction experience, and survey data quality. Comput Hum Behav 2022 Jan;126:107034. [doi: 10.1016/j.chb.2021.107034]
52. Sulieman L, Cronin RM, Carroll RJ, Natarajan K, Marginean K, Mapes B, et al. Comparing medical history data derived from electronic health records and survey answers in the All of Us research program. J Am Med Inform Assoc 2022 Jun 14;29(7):1131-1141 [FREE Full text] [doi: 10.1093/jamia/ocac046] [Medline: 35396991]
53. Ponathil A, Ozkan F, Bertrand J, Agnisarman S, Narasimha S, Welch B, et al. An empirical study investigating the user acceptance of a virtual conversational agent interface for family health history collection among the geriatric population. Health Informatics J 2020 Dec;26(4):2946-2966 [FREE Full text] [doi: 10.1177/1460458220955104] [Medline: 32938275]
54. Kaphingst KA, Goodman M, Pandya C, Garg P, Stafford J, Lachance C. Factors affecting frequency of communication about family health history with family members and doctors in a medically underserved population. Patient Educ Couns 2012 Aug;88(2):291-297 [FREE Full text] [doi: 10.1016/j.pec.2011.11.013] [Medline: 22197261]
55. Sweller J. Chapter two - cognitive load theory. In: Mestre JP, Ross BH, editors. Psychology of Learning and Motivation. Cambridge, MA: Academic Press; 2011:37-76.
56. Haugeland IK, Følstad A, Taylor C, Bjørkli CA. Understanding the user experience of customer service chatbots: an experimental study of chatbot interaction design. Int J Hum Comput Stud 2022 May;161:102788. [doi: 10.1016/j.ijhcs.2022.102788]
57. Cronin RM, Halvorson AE, Springer C, Feng X, Sulieman L, Loperena-Cortes R, et al. Comparison of family health history in surveys vs electronic health record data mapped to the observational medical outcomes partnership data model in the All of Us research program. J Am Med Inform Assoc 2021 Mar 18;28(4):695-703 [FREE Full text] [doi: 10.1093/jamia/ocaa315] [Medline: 33404595]
58. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res 2022 Sep 12;22(1):1148 [FREE Full text] [doi: 10.1186/s12913-022-08527-9] [Medline: 36096793]
59. Wolf MS, Feinglass J, Thompson J, Baker DW. In search of 'low health literacy': threshold vs. gradient effect of literacy on health status and mortality. Soc Sci Med 2010 May;70(9):1335-1341. [doi: 10.1016/j.socscimed.2009.12.013] [Medline: 20167411]
60. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
61. Farmer GD, Gray H, Chandratillake G, Raymond FL, Freeman AL. Recommendations for designing genetic test reports to be understood by patients and non-specialists. Eur J Hum Genet 2020 Jul;28(7):885-895 [FREE Full text] [doi: 10.1038/s41431-020-0579-y] [Medline: 32024982]

XSL•FO
RenderX

62. Dwyer AA, Au MG, Smith N, Plummer L, Lippincott MF, Balasubramanian R, et al. Evaluating co-created patient-facing materials to increase understanding of genetic test results. J Genet Couns 2021 Apr;30(2):598-605 [FREE Full text] [doi: 10.1002/jgc4.1348] [Medline: 33098367]
63. Wei J, Kim S, Jung H, Kim YH. Leveraging large language models to power chatbots for collecting user self-reported data. Proc ACM Hum Comput Interact 2024 Apr 26;8(CSCW1):1-35 [FREE Full text] [doi: 10.1145/3637364]
64. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022 Feb 10;10(2):e32875 [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
65. Iftikhar A, Bond RR, McGilligan V, Leslie SJ, Rjoob K, Knoery C, et al. Comparing single-page, multipage, and conversational digital forms in health care: usability study. JMIR Hum Factors 2021 May 26;8(2):e25787 [FREE Full text] [doi: 10.2196/25787] [Medline: 34037531]
66. Te Pas ME, Rutten WG, Bouwman RA, Buise MP. User experience of a chatbot questionnaire versus a regular computer questionnaire: prospective comparative study. JMIR Med Inform 2020 Dec 07;8(12):e21982 [FREE Full text] [doi: 10.2196/21982] [Medline: 33284125]

Abbreviations

CUQ: Chatbot Usability Questionnaire
EHR: electronic health record
FHx: family health history
KIT: curious interactive test
LLM: large language model
MTurk: Mechanical Turk
NLM: National Library of Medicine
REDCap: Research Electronic Data Capture
STARE-HI: Statement on Reporting of Evaluation Studies in Health Informatics
SUS: System Usability Scale
UX: user experience

Edited by A Mavragani; submitted 04.12.23; peer-reviewed by K Kaphingst, K Wiley, N Sperber; comments to author 29.02.24; revised version received 31.05.24; accepted 25.07.24; published 30.09.24.

Please cite as:
Nguyen MH, Sedoc J, Taylor CO
Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis
J Med Internet Res 2024;26:e55164
URL: https://www.jmir.org/2024/1/e55164
doi:10.2196/55164
PMID:39348188

©Michelle Hoang Nguyen, João Sedoc, Casey Overby Taylor. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Models May Help Patients Understand Peer-Reviewed Scientific Articles About Ophthalmology: Development and Usability Study

Reza Kianian[1*], MD; Deyu Sun[1*], PhD; William Rojas-Carabali[2,3,4*], MD; Rupesh Agrawal[2,3,4*], MD; Edmund Tsui[1*], MD

[1]Stein Eye Institute, Department of Ophthalmology, David Geffen School of Medicine, Los Angeles, CA, United States
[2]Nanyang Technological University, Lee Kong Chian School of Medicine, Singapore, Singapore
[3]Tan Tock Seng Hospital, National Healthcare Group Eye Institute, Singapore, Singapore
[4]National Healthcare Group, Programme for Ocular Inflammation & Infection Translational Research, Singapore, Singapore
[*]all authors contributed equally

Corresponding Author:
Edmund Tsui, MD
Stein Eye Institute
Department of Ophthalmology
David Geffen School of Medicine
200 Stein Plaza
Los Angeles, CA, 90095
United States
Phone: 1 310 825 5440
Email: etsui@mednet.ucla.edu

Abstract

Background: Adequate health literacy has been shown to be important for the general health of a population. To address this, it is recommended that patient-targeted medical information is written at a sixth-grade reading level. To make well-informed decisions about their health, patients may want to interact directly with peer-reviewed open access scientific articles. However, studies have shown that such text is often written with highly complex language above the levels that can be comprehended by the general population. Previously, we have published on the use of large language models (LLMs) in easing the readability of patient-targeted health information on the internet. In this study, we continue to explore the advantages of LLMs in patient education.

Objective: This study aimed to explore the use of LLMs, specifically ChatGPT (OpenAI), to enhance the readability of peer-reviewed scientific articles in the field of ophthalmology.

Methods: A total of 12 open access, peer-reviewed papers published by the senior authors of this study (ET and RA) were selected. Readability was assessed using the Flesch-Kincaid Grade Level and Simple Measure of Gobbledygook tests. ChatGPT 4.0 was asked “I will give you the text of a peer-reviewed scientific paper. Considering that the recommended readability of the text is 6th grade, can you simplify the following text so that a layperson reading this text can fully comprehend it? - Insert Manuscript Text -”. Appropriateness was evaluated by the 2 uveitis-trained ophthalmologists. Statistical analysis was performed in Microsoft Excel.

Results: ChatGPT significantly lowered the readability and length of the selected papers from 15th to 7th grade ($P<.001$) while generating responses that were deemed appropriate by expert ophthalmologists.

Conclusions: LLMs show promise in improving health literacy by enhancing the accessibility of peer-reviewed scientific articles and allowing the general population to interact directly with medical literature.

(J Med Internet Res 2024;26:e59843) doi:10.2196/59843

KEYWORDS

uveitis; artificial intelligence; ChatGPT; readability; peer review; large language models; LLMs; health literacy; patient education; medical information; ophthalmology

XSL•FO
RenderX

Introduction

Health literacy is pivotal for empowering individuals to make informed health decisions, navigate the health care system, and manage their well-being. It bridges the gap between complex medical information and patient understanding, thus playing a crucial role in enhancing public health outcomes and reducing health disparities [1]. Readability in clinical medicine refers to the reading level required to fully comprehend the information presented in a body of medical text. The health literacy of the United States' general population is poor, as the average person cannot comprehend text beyond an eighth-grade level, a limitation exacerbated among patients with Medicare or Medicaid [2]. Contributing to the rapid and unsustainable expansion of annual health care costs in the United States, poor health literacy has been shown to be detrimental to the nation's general health [3]. A wealth of published work exists on the association between poor health literacy and more hospitalizations, inadequate follow-up, underuse of preventative care, poor medication compliance, and increased mortality [3,4].

To address these disparities, the American Medical Association (AMA) and the National Institutes of Health (NIH) recommend that the readability of patient-targeted health information be equivalent to a sixth- to eighth-grade level [2,5]. This is particularly crucial in the age of the internet, as over half of the US population browses health information on the internet to learn about their conditions and their management [6].

Numerous studies have demonstrated that the readability of health information found on the internet, including content published by academic institutions, is often poor and notably more complex than the recommended levels [7-10]. With the recent surge and growing popularity of large language models (LLMs), such as ChatGPT (OpenAI), some investigators have begun exploring the possibility of using this tool to enhance the readability and accessibility of health information for patients [11-18]. Our team recently conducted a study showcasing ChatGPT's ability to rewrite existing health information with poor readability into documents with reading levels that align with the recommendations of the AMA and NIH [13].

Recent studies have shown the readability of scientific texts to be getting worse over time with a study showing the average reading level required to understand papers and their abstracts to be as high as 17th grade (college graduate level) [19,20]. Building on our previous research, which demonstrated the efficacy of ChatGPT in addressing health literacy challenges, our objective was to assess whether this technology could successfully transform the complex language found in scientific papers into more accessible health information. This evaluation specifically aimed to cater to laypersons seeking medical advice on topics related to uveitis.

Methods

Overview

To conduct this study, we identified 12 open access papers on various uveitis-related topics published by the 2 senior authors of this study (ET and RA). The papers covered several available study types, such as case reports, imaging studies, prospective studies, retrospective studies, and review papers. The readability of these articles was assessed using 2 validated and commonly used tools in the literature, that is, Flesch-Kincaid Grade Level (FKGL) and Simple Measure of Gobbledygook (SMOG). There are other readability assessment tools, such as the Gunning Fog index, Coleman-Liau index, and automated readability index, available. However, the FKGL and the SMOG tools are the 2 most commonly used readability assessment tools that we identified in our literature review. Total words, sentences, syllables, and polysyllabic words are used in the formulas underpinning these 2 readability tools. A calculator (readabilityformulas website) was used. We then asked ChatGPT 4.0 the following prompt: "I will give you the text of a peer-reviewed scientific paper. Considering that the recommended readability of the text is 6th grade, can you simplify the following text so that a layperson reading this text can fully comprehend it? - Insert Manuscript Text -". Appropriateness was evaluated similarly to criteria previously published by 2 fellowship-trained uveitis physicians (ET and RA) who authored the papers selected for this study [13,21]. Responses were marked as "appropriate" or "inappropriate" based on the authors' clinical experiences and knowledge of the literature. An appropriate response was one that accurately simplified the entire text of the article without including any information that would be deemed as false or inaccurate. An inappropriate response was one that either included false or inaccurate information or included information that was not intended by the authors of the study. If there were disagreements between the 2 authors on the appropriateness of a generated response, a third fellowship-trained uveitis physician would be asked for an independent opinion. Statistical analysis was performed using Microsoft Excel (version 2401). Paired-sample *t* test was used to compare the average readability and the word count obtained from the original text of the included articles with the generated responses by ChatGPT with the recommended sixth-grade readability levels. Descriptive statistics were used to represent the rest of the data.

Ethical Considerations

Human participants and their associated data were not used in this study. Therefore, no informed consents, statements on language waivers, and privacy, and confidentiality statements are necessary.

Results

ChatGPT was able to rewrite scientific articles into an appropriate body of text with significantly improved reading levels ($P<.001$). The average reading level of the original text was around 15th grade, and it was reduced to a seventh-grade reading level. This improvement was present across all study types analyzed. The output generated by ChatGPT was also found to be appropriate and accurately represent the data presented in the original text in a simple format without any medical mistakes. The average word count for the original papers was 2882 (SD 1348) words, and ChatGPT responses had an average count of 417 (SD 78) words ($P<.001$). The readability and appropriateness of each response, along with

XSL•FO
RenderX

the original readability of the paper and study types are documented in Table 1.

Table 1. Selected open access research articles and their corresponding readability and appropriateness.

Title	Type	Original FKGL[a]	ChatGPT FKGL	Original SMOG[b]	ChatGPT SMOG	Appropriateness (yes or no)
Delayed onset anterior uveitis and macular edema after cessation of pembrolizumab	Case report	16.17	7.4	14.65	6.57	Yes
Post typhoid fever neuroretinitis with serous retinal detachment and choroidal involvement-A case report	Case report	16.41	4.72	14.84	4.57	Yes
Choroidal Vascularity Index (CVI)—A Novel Optical Coherence Tomography Parameter for Monitoring Patients with Panuveitis?	Imaging study	16.34	8.3	14.79	8.21	Yes
Quantification of Anterior Chamber Cells in Children with Uveitis Using Anterior Segment Optical Coherence Tomography	Imaging study	16.92	8.38	15.25	8.9	Yes
Evaluation of Retinal Vascularity Index in Patients with COVID-19: A Case–Control Study	Prospective study	15.19	7.26	14.05	7.31	Yes
Choroidal structural changes in preterm children with and without retinopathy of prematurity	Prospective study	12.12	7.33	11.96	7.2	Yes
Implementation of a vision-screening program in rural northeastern United States	Prospective study	14.91	7.37	14.11	6.62	Yes
Choroidal vascularity index as a measure of vascular status of the choroid: Measurements in healthy eyes from a population-based study	Prospective study	14.26	7.23	12.72	7.27	Yes
Factors affecting final functional outcomes in open - globe injuries and use of ocular trauma score as a predictive tool in Nepalese population	Retrospective study	14.04	8.29	13.65	8.41	Yes
Prognostic factors for vision outcome after surgical repair of open globe injuries	Retrospective study	12.9	7.57	13.07	8.01	Yes
Multimodal imaging in pediatric uveitis	Review paper	17.83	9.17	16.12	9.59	Yes
Recent advances in the treatment of juvenile idiopathic arthritis–associated uveitis	Review paper	24.11	7.08	20.41	7.17	Yes
Average (SD)		15.93 (3.07)	7.51 (1.08)	14.64 (2.15)	7.49 (1.29)	
P value		<.001	<.001	<.001	<.001	

[a]FKGL: Flesch-Kincaid Grade Level.

[b]SMOG: Simple Measure of Gobbledygook.

Discussion

Principal Findings

In summary, ChatGPT lowered the readability of open access and peer-reviewed science articles by 8 grade points from 15th grade to 7th grade ($P<.001$). Since the introduction of LLMs like ChatGPT just over a year ago, there has been increasing interest in the use of such artificial intelligence (AI) models in health care and clinical education. In addition to generating patient-targeted health information, investigators have explored the use of LLMs in medical education, guiding patients postoperatively, and answering common questions [22-26]. The authors of this study were some of the first to examine the role of ChatGPT in producing readable health care information or rewriting existing patient-targeted information. This analysis is an extension of such a study.

Previously, we found mixed results when asking ChatGPT to produce patient-targeted health information regarding uveitis and surgical management of glaucoma. In 1 study, we illustrated the superiority of ChatGPT to Bard, Google's AI, and demonstrated that ChatGPT produced educational material for uveitis at the recommended sixth-grade reading level [13]. ChatGPT was also able to convert already existing health information found on the internet with a readability of 11th grade into significantly more readable content with a reading level of eighth grade. In another study, however, while Kianian et al [12] demonstrated the reliability of ChatGPT in assessing the quality of health information targeted to the layperson, they found that ChatGPT was not able to produce highly readable information on the surgical management of glaucoma.

The findings of this study align more with our previous study on ChatGPT and uveitis. When asked to rewrite peer-reviewed

XSL•FO
RenderX

scientific papers with complex texts and a readability of 15th grade into easier-to-understand information for a layperson, ChatGPT was able to produce appropriate responses with a readability of seventh grade. These findings are promising since this can potentially improve the comprehension of peer-reviewed open access scientific articles for the general population while maintaining accurate information. Before the introduction of LLMs to the population, patients may have had difficulty comprehending highly technical and difficult-to-understand scientific articles. Previous studies support the positive influence of improved medical literature comprehension and health knowledge on trust in scientific bodies and the enhanced management of conditions by patients [27,28]. Therefore, the use of ChatGPT by patients may contribute meaningfully to public trust in the scientific process.

Patients may benefit from directly engaging with peer-reviewed research as their ability to comprehend research may enhance their care on multiple fronts [29]. Extensive research has identified that understanding the rationale behind medical decisions is a significant factor in improving patients' adherence to their medical regimen across specialties [30]. Furthermore, patients may become more well-informed when faced with inaccurate and misleading controversies, such as the link between vaccines and autism or the vast amount of misinformation present on the COVID-19 pandemic [29]. To combat health misinformation, Swire-Thompson and Lazer [31] argue for improved health literacy, using the internet collaboratively with physicians and stronger signal of source quality. The ability of the general public to directly interact with clinical trials and peer-reviewed scientific articles may aid in empowering patients in making better-informed decisions regarding their health.

Limitations

There are limitations to this analysis and its application. Although we checked for the appropriateness of the information generated by ChatGPT, studies have shown that ChatGPT can sometimes produce inaccurate information [32,33]. Therefore, we urge patients to not solely base their health education on ChatGPT rewrite of scientific articles, rather we recommend that they consult their physicians regarding their health conditions. Second, it is possible that 1 article may not by itself provide context or provide all the information needed to make sound decisions regarding care. Hence, patients must study various trusted sources of information to make informed decisions. Furthermore, we are not aware if the appropriateness of the responses would be affected negatively if patients ask for a follow-up to make the generated responses even shorter. Third, although readable content is essential for comprehension, future studies may recruit patients in well-designed prospective studies to assess the true potential of ChatGPT in the comprehension of medical education and their decision-making. Fourth, the version of ChatGPT (version 4.0), which was used in this study, requires a monthly subscription fee. However, previous studies have demonstrated the ability of the free version of ChatGPT (version 3.5) to generate and also rewrite patient-targeted health information in a language with significantly improved readability scores [34,35]. Therefore, we suspect that ChatGPT-3.5 could also be helpful to patients. Fifth, only peer-reviewed scientific articles on topics mostly around uveitis written by the 2 senior authors of this study were analyzed in this investigation. It is important for the readers to avoid making generalizations from the findings of our study toward all published scientific articles in the field of ophthalmology or, more broadly, medicine. Finally, although we were careful to choose open access articles that are coauthored by the senior authors of this paper, we do recognize that further evaluation and oversight are needed to determine the legality and the ethics of entering copyrighted material, or any form of intellectual property, into LLMs such as ChatGPT.

Future Work

In this investigation, we only chose peer-reviewed open access articles that have been coauthored by the senior authors of this study for several reasons. First, given concern regarding the ethics of inputting text from peer-reviewed papers written by other authors not involved in this study, we decided to use papers that are both open access as well as written by the senior authors of this study. This choice, we believe, also allowed for a better judgment when assessing the appropriateness of ChatGPT's generated summary of the inputted text.

As the popularity of LLMs, such as ChatGPT, rises among health care professionals and patients across the world, the application of AI, such as ChatGPT, in simplifying the text of additional highly impactful articles for the general population should be investigated. A recent study by Sener et al [36] has identified the top 50 most-cited articles within the field of uveitis, and another article by Ohba et al [37] in the *Journal of American Medical Association* (*JAMA*) has identified the 100 most frequently cited articles in the field of ophthalmology. Sener et al [36] use a unique approach in identifying articles that are not only highly cited in scientific journals, but also those that are cited frequently across social media platforms [36]. Given the exposure of the general public to such articles, the methodology of our study could be applied to the identified peer-reviewed papers.

Conclusions

In conclusion, our study demonstrates the exciting potential of LLMs, demonstrated by ChatGPT, in giving the general population the power and the tools needed to tackle the challenge of engaging directly with peer-reviewed open access scientific articles. In this study, ChatGPT was able to appropriately rewrite poorly readable scientific papers into an appropriate seventh-grade response. We advocate caution, however, as LLMs are integrated more and more into the daily life of the general population, and we emphasize the importance of consultation with health care providers before making health decisions.

XSL•FO
RenderX

Acknowledgments

This work is supported in part by the National Eye Institute of the National Institutes of Health under award K23EY032990 (ET). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, and an unrestricted grant from Research to Prevent Blindness, Inc (New York) to the UCLA Stein Eye Institute for research. The funding organizations had no role in the design or conduction of the study.

Our study primarily studied the ability of generative AI in summarizing peer-reviewed scientific articles in easy-to-understand text. However, generative AI was not used in writing any portion of this manuscript.

Authors' Contributions

Conceptualization was contributed by RK, DS, and ET. Data curation and formal analysis was performed by RK and DS. Funding acquisition was managed by RA and ET. Investigation, methodology, validation, writing—original draft, writing—review and editing, and project administration were performed by all authors. Software was handled by RK and DS. Resources were contributed by all authors. Supervision was handled by RA and ET. Visualization was performed by RK and DS.

Conflicts of Interest

ET is a consultant for Kowa, Cylite, Oculis, and Eyepoint Pharmaceuticals.

References

1. Benjamin RM. Improving health by improving health literacy. Public Health Rep 2010;125(6):784-785 [FREE Full text] [doi: 10.1177/003335491012500602] [Medline: 21121221]
2. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 2021;8:2374373521998847 [FREE Full text] [doi: 10.1177/2374373521998847] [Medline: 34179407]
3. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
4. Coughlin SS, Vernon M, Hatzigeorgiou C, George V. Health literacy, social determinants of health, and disease prevention and control. J Environ Health Sci 2020;6(1):1-10 [FREE Full text] [Medline: 33604453]
5. Abu-Heija A, Shatta M, Ajam M, Abu-Heija U, Imran N, Levine D. Quantitative readability assessment of the internal medicine online patient information on Annals.org. Cureus 2019 Mar 06;11(3):e4184 [FREE Full text] [doi: 10.7759/cureus.4184] [Medline: 31106084]
6. Hall AK, Bernhardt JM, Dodd V, Vollrath MW. The digital health divide: evaluating online health information access and use among older adults. Health Educ Behav 2015 Apr;42(2):202-209 [FREE Full text] [doi: 10.1177/1090198114547815] [Medline: 25156311]
7. Kianian R, Belarmino A, Finkelshtein I, Mills JN, Eleswarapu SV, Donin NM. Should I freeze my sperm?-readability and quality of health resources for sperm banking. Transl Androl Urol 2023 Oct 31;12(10):1561-1567 [FREE Full text] [doi: 10.21037/tau-23-120] [Medline: 37969777]
8. Kianian R, Hu MY, Lavold AJ, Andino JJ, Morrison JC, Eleswarapu SV, et al. Patient-directed vasectomy information: how readable is it? World J Mens Health 2024 Apr;42(2):408-414 [FREE Full text] [doi: 10.5534/wjmh.230033] [Medline: 37853530]
9. Ayoub S, Tsui E, Mohammed T, Tseng J. Readability assessment of online uveitis patient education materials. Ocul Immunol Inflamm 2019;27(3):399-403. [doi: 10.1080/09273948.2017.1413396] [Medline: 29286863]
10. Tran J, Tsui E. Assessment of the Readability, Availability, and Quality of Online Patient Education Materials Regarding Uveitis Medications. Ocul Immunol Inflamm 2021 Nov 17;29(7-8):1507-1512. [doi: 10.1080/09273948.2020.1737144] [Medline: 32275173]
11. Golan R, Ripps S, Reddy R, Loloi J, Bernstein A, Connelly Z, et al. ChatGPT's ability to assess quality and readability of online medical information: evidence from a cross-sectional study. Cureus 2023 Jul;15(7):e42214 [FREE Full text] [doi: 10.7759/cureus.42214] [Medline: 37484787]
12. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? J Glaucoma 2024 Feb 01;33(2):94-100. [doi: 10.1097/IJG.0000000000002338] [Medline: 38031276]
13. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina 2024 Feb;8(2):195-201 [FREE Full text] [doi: 10.1016/j.oret.2023.09.008] [Medline: 37716431]
14. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV information on Google versus AI (ChatGPT). Otolaryngol Head Neck Surg 2024 Jun;170(6):1504-1511. [doi: 10.1002/ohn.506] [Medline: 37622581]
15. Hurley ET, Crook BS, Lorentz SG, Danilkowicz RM, Lau BC, Taylor DC, et al. Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. Arthroscopy 2024 Mar;40(3):726-731.e6. [doi: 10.1016/j.arthro.2023.07.048] [Medline: 37567487]

XSL•FO
RenderX

16. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS. Evaluation of online artificial intelligence-generated information on common hand procedures. J Hand Surg Am 2023 Nov;48(11):1122-1127. [doi: 10.1016/j.jhsa.2023.08.003] [Medline: 37690015]
17. Tepe M, Emekli E. Decoding medical jargon: the use of AI language models (ChatGPT-4, BARD, microsoft copilot) in radiology reports. Patient Educ Couns 2024;126:108307. [doi: 10.1016/j.pec.2024.108307] [Medline: 38743965]
18. Park J, Oh K, Han K, Lee YH. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. Sci Rep 2024;14(1):13218 [FREE Full text] [doi: 10.1038/s41598-024-63824-z] [Medline: 38851825]
19. Plavén-Sigray P, Matheson GJ, Schiffler BC, Thompson WH. The readability of scientific texts is decreasing over time. Elife 2017;6:e27725 [FREE Full text] [doi: 10.7554/eLife.27725] [Medline: 28873054]
20. Yeung AWK, Goto TK, Leung WK. Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae. Front Hum Neurosci 2018;12:308 [FREE Full text] [doi: 10.3389/fnhum.2018.00308] [Medline: 30158861]
21. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 2023;7(10):862-868. [doi: 10.1016/j.oret.2023.05.022] [Medline: 37277096]
22. Wong RS, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023;9:e47274 [FREE Full text] [doi: 10.2196/47274] [Medline: 37988149]
23. Bhattamisra S, Banerjee P, Gupta P, Mayuren J, Patra S, Candasamy M. Artificial intelligence in pharmaceutical and healthcare research. Big Data and Cognitive Computing 2023;7(1):10 [FREE Full text] [doi: 10.3390/bdcc7010010]
24. Capelleras M, Soto-Galindo GA, Cruellas M, Apaydin F. ChatGPT and rhinoplasty recovery: an exploration of AI's role in postoperative guidance. Facial Plast Surg 2024;40(5):628-631. [doi: 10.1055/a-2219-4901] [Medline: 38029791]
25. Giorgino R, Alessandri-Bonetti M, Luca A, Migliorini F, Rossi N, Peretti GM, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg 2023;10:1284015 [FREE Full text] [doi: 10.3389/fsurg.2023.1284015] [Medline: 38026475]
26. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, Amoro Sanchez R, Muyolema Arce V, Mustafa A, et al. The future of patient education: AI-driven guide for type 2 diabetes. Cureus 2023;15(11):e48919 [FREE Full text] [doi: 10.7759/cureus.48919] [Medline: 38024047]
27. Tsai T, Yu W, Lee SD. Is health literacy associated with greater medical care trust? Int J Qual Health Care 2018;30(7):514-519. [doi: 10.1093/intqhc/mzy043] [Medline: 29608676]
28. Chen X, Hay JL, Waters EA, Kiviniemi MT, Biddle C, Schofield E, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734 [FREE Full text] [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
29. Smith R. Do patients need to read research? BMJ 2003;326(7402):1307 [FREE Full text] [doi: 10.1136/bmj.326.7402.1307] [Medline: 12805166]
30. Miller TA. Health literacy and adherence to medical treatment in chronic and acute illness: a meta-analysis. Patient Educ Couns 2016;99(7):1079-1086 [FREE Full text] [doi: 10.1016/j.pec.2016.01.020] [Medline: 26899632]
31. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
32. Emsley R. ChatGPT: these are not hallucinations - they're fabrications and falsifications. Schizophrenia (Heidelb) 2023;9(1):52 [FREE Full text] [doi: 10.1038/s41537-023-00379-4] [Medline: 37598184]
33. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
34. Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, et al. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med 2024;39(4):573-577 [FREE Full text] [doi: 10.1007/s11606-023-08469-w] [Medline: 37940756]
35. Kianian R, Carter M, Finkelshtein I, Eleswarapu SV, Kachroo N. Application of artificial intelligence to patient-targeted health information on kidney stone disease. J Ren Nutr 2024;34(2):170-176. [doi: 10.1053/j.jrn.2023.10.002] [Medline: 37839591]
36. Sener H, Evereklioglu C, Horozoglu F. An analysis of the 50 most-cited "Uveitis" articles published between 2010-2020 from a bibliographic and altmetric perspective. Cureus 2022;14(10):e29930 [FREE Full text] [doi: 10.7759/cureus.29930] [Medline: 36348932]
37. Ohba N, Nakao K, Isashiki Y, Ohba A. The 100 most frequently cited articles in ophthalmology journals. Arch Ophthalmol 2007;125(7):952-960. [doi: 10.1001/archopht.125.7.952] [Medline: 17620577]

Abbreviations

AI: artificial intelligence
AMA: American Medical Association
FKGL: Flesch-Kincaid Grade Level
LLM: large language model

XSL•FO
RenderX

NIH: National Institutes of Health
SMOG: Simple Measure of Gobbledygook

Edited by A Coristine; submitted 23.04.24; peer-reviewed by C Ganjavi, E Emekli; comments to author 26.06.24; revised version received 10.07.24; accepted 05.11.24; published 24.12.24.

Please cite as:
Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E
Large Language Models May Help Patients Understand Peer-Reviewed Scientific Articles About Ophthalmology: Development and Usability Study
J Med Internet Res 2024;26:e59843
URL: https://www.jmir.org/2024/1/e59843
doi:10.2196/59843
PMID:

©Reza Kianian, Deyu Sun, William Rojas-Carabali, Rupesh Agrawal, Edmund Tsui. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Identification of Ethical Issues and Practice Recommendations Regarding the Use of Robotic Coaching Solutions for Older Adults: Narrative Review

Cécilia Palmier[1,2*], MSc; Anne-Sophie Rigaud[1,2*], Prof Dr Med; Toshimi Ogawa[3], PhD; Rainer Wieching[4], Prof Dr; Sébastien Dacunha[1,2*], MSc; Federico Barbarossa[5], MEng; Vera Stara[5], PhD; Roberta Bevilacqua[5], MSc; Maribel Pino[1,2*], PhD

[1]Maladie d'Alzheimer, Université de Paris, Paris, France
[2]Service de Gériatrie 1 & 2, Hôpital Broca, Assistance Publique - Hôpitaux de Paris, Paris, France
[3]Smart-Aging Research Center, Tohoku University, Sendai, Japan
[4]Institute for New Media & Information Systems, University of Siegen, Siegen, Germany
[5]Scientific Direction, Istituto Nazionale di Ricovero e Cura per Anziani, Ancona, Italy
[*]these authors contributed equally

Corresponding Author:
Anne-Sophie Rigaud, Prof Dr Med
Service de Gériatrie 1 & 2
Hôpital Broca
Assistance Publique - Hôpitaux de Paris
54 rue Pascal
Paris, 75013
France
Phone: 33 144083503
Fax: 33 144083510
Email: anne-sophie.rigaud@aphp.fr

Abstract

Background: Technological advances in robotics, artificial intelligence, cognitive algorithms, and internet-based coaches have contributed to the development of devices capable of responding to some of the challenges resulting from demographic aging. Numerous studies have explored the use of robotic coaching solutions (RCSs) for supporting healthy behaviors in older adults and have shown their benefits regarding the quality of life and functional independence of older adults at home. However, the use of RCSs by individuals who are potentially vulnerable raises many ethical questions. Establishing an ethical framework to guide the development, use, and evaluation practices regarding RCSs for older adults seems highly pertinent.

Objective: The objective of this paper was to highlight the ethical issues related to the use of RCSs for health care purposes among older adults and draft recommendations for researchers and health care professionals interested in using RCSs for older adults.

Methods: We conducted a narrative review of the literature to identify publications including an analysis of the ethical dimension and recommendations regarding the use of RCSs for older adults. We used a qualitative analysis methodology inspired by a Health Technology Assessment model. We included all article types such as theoretical papers, research studies, and reviews dealing with ethical issues or recommendations for the implementation of these RCSs in a general population, particularly among older adults, in the health care sector and published after 2011 in either English or French. The review was performed between August and December 2021 using the PubMed, CINAHL, Embase, Scopus, Web of Science, IEEE Explore, SpringerLink, and PsycINFO databases. Selected publications were analyzed using the European Network of Health Technology Assessment Core Model (version 3.0) around 5 ethical topics: benefit-harm balance, autonomy, privacy, justice and equity, and legislation.

Results: In the 25 publications analyzed, the most cited ethical concerns were the risk of accidents, lack of reliability, loss of control, risk of deception, risk of social isolation, data confidentiality, and liability in case of safety problems. Recommendations included collecting the opinion of target users, collecting their consent, and training professionals in the use of RCSs. Proper data management, anonymization, and encryption appeared to be essential to protect RCS users' personal data.

XSL•FO
RenderX

Conclusions: Our analysis supports the interest in using RCSs for older adults because of their potential contribution to individuals' quality of life and well-being. This analysis highlights many ethical issues linked to the use of RCSs for health-related goals. Future studies should consider the organizational consequences of the implementation of RCSs and the influence of cultural and socioeconomic specificities of the context of experimentation. We suggest implementing a scalable ethical and regulatory framework to accompany the development and implementation of RCSs for various aspects related to the technology, individual, or legal aspects.

(J Med Internet Res 2024;26:e48126) doi:10.2196/48126

KEYWORDS

robotic coaching solutions; ethical issues; ethical recommendations; older adults; geriatrics; guidelines

Introduction

Challenges Associated to Population Aging

Technological and medical advances have led to a demographic shift in the population, with the number of older adults constantly increasing. According to the United Nations [1], older adults (aged 60-65 years) will represent 16% of the world's population in 2050. In addition, life expectancy is increasing, from 64.2 years in 1990 to 72.6 years in 2019, and is expected to reach 77.1 years in 2050 [1]. However, there is a wide diversity of health conditions among older adults. The health status of older adults is dependent on multiple factors, including nonmodifiable genetic factors and environmental factors, such as lifestyle [2]. Thus, older adults represent a very heterogeneous population with multiple and diverse needs and desires. With advancing age, the loss of functional independence; frailty; and other health diseases such as cardiovascular problems, cancers, osteoarthritis, osteoporosis, or major neurocognitive disorders may appear [3-5]. Among age-related conditions, major neurocognitive disorders (eg, Alzheimer disease) receive particular attention due to the increasing prevalence of these diseases [6].

The aging population is not only a public health issue but also a socioeconomic one. To face this challenge, it is important to develop preventive measures to support active and healthy aging and to preserve the independent functioning and quality of life of older adults. The adoption of healthy behaviors can help prevent or delay the onset of pathologies or treat them if detected early [7].

The Use of Technologies for Older Adults

Preventive health measures can be supported through new technologies, such as robotic coaching solutions (RCSs) that promote healthy aging among older adults [8,9]. RCSs have been defined as personalized systems that continuously monitor the activities and environment of the user and provide them with timely health-related advice and interventions [10-12]. These systems can help users define and achieve different health-oriented goals [12].

RCSs may encompass artificial intelligence (AI) technologies that can analyze user data, personalize coaching programs, and adapt recommendations based on each individual's needs [1,13-19]. RCSs can involve robots equipped with sensors such as cameras, microphones, or motion sensors to collect real-time data about the user, AI, and programming that enables their interaction with users [20,21]. These technologies are often equipped with voice and visual recognition and learning capabilities [20,21]. They can benefit from advanced natural language processing techniques, which allow for understanding of the user's input, facilitating natural and effective communication [22]. RCSs can offer guidance, support, and feedback based on preprogrammed information or real-time data analysis. These data can inform coaching strategies and allow RCSs to provide users with relevant feedback [8].

RCSs can also encompass a virtual agent, which refers to a computer program or an AI system that interacts with users in a manner that simulates human conversation [14,18,23]. A virtual agent is an animated character capable of adopting a social behavior mimicking that of humans to encourage the users to make changes in their habits [14]. Virtual agents might take the form of a chatbot, voice assistant, or other AI-driven communication system [14]. Biometric monitoring devices to track physiological data such as heart rate, sleep patterns, or stress levels can also be included in RCSs [8,20,21]. These data can contribute to the configuration of personalized coaching plans. RCSs can also encompass advanced data analytics that can process large data sets generated by users' interactions and behaviors. This functionality helps in identifying patterns, trends, and areas for improvement in coaching strategies [24]. Integrating Internet-of-Things devices in RCSs can provide additional data points about a user's environment, lifestyle, or habits, thus contributing to a personalized coaching approach [25].

Health-oriented RCSs could enable users to lead a healthy lifestyle, by identifying needs and goals and providing appropriate risk predictions and individualized recommendations [12,26-28]. There are RCSs dedicated to a particular domain, such as physical activity or motor rehabilitation [9,16]. Others may have the objective of promoting independent and healthy aging [29].

Promoting active and healthy aging can allow older adults to maintain their independence and continue to live at home [4,30], which is a wish of many [3]. This intervention could also help to reduce the need for assistance, usually provided by informal caregivers and health professionals [4,19,30-33]. Furthermore, RCSs could lead to a reduction in individual and collective health care expenses [4,32,34] by easing access to health and social care interventions to a wide population, including hard-to-reach (eg, geographically isolated) individuals. However, although the use of health-related RCSs could have many benefits, several ethical issues arise with their

XSL•FO
RenderX

development and implementation in human environments [3,35-38].

An Ethical Framework for the Use of Technologies for Older Adults

For RCSs to contribute to active and healthy aging, it is important that all the stakeholders (engineers, geriatricians, psychologists, etc) involved in their design and implementation refer to an ethical framework [3,38]. It is also important to inform society (politicians and legal experts) about such an extension of technology in people's lives (private, professional, medicosocial, and commercial context), so that we can create a legal framework for the use of these technologies. An analysis of the way in which ethical and legal dimensions have been addressed by studies, in the field of RCSs for health care, seems useful to support the key actors in their development and implementation. The growing interest in the ethical questions associated with the use of social and assistive robots is evidenced by the volume of literature reviews [3,12,18,31,32,37,39-51] on the topic.

Now, it appears appropriate to systematically examine this body of work, focusing on the ethical analysis, and provide an overview of the literature. Therefore, we performed a review of the literature on RCSs for older adults using the European Network of Health Technology Assessment (EUnetHTA Core Model; version 3.0) model [52] for analysis. This Health Technology Assessment (HTA) model makes it possible to assess the intended and unintended consequences of the use of a specific technology regarding multiple domains (eg, technological, ethical, clinical, and organizational), providing methods and concepts for this analysis [53]. Therefore, HTA is a process that informs decision-making about the introduction of new technologies such as RCSs in health care. It also seems necessary to issue guidelines for the development and implementation of health-oriented RCSs [54].

The objective of this study was to highlight the main ethical questions and corresponding recommendations linked to the use of RCSs for older adults for engineers, researchers, and health professionals in this field. For this purpose, we conducted a narrative literature review using the ethical dimension of the EUnetHTA Core Model to guide the analysis. To the best of our knowledge, such a study has not been conducted so far.

Methods

Overview

A thematic analysis of the literature was performed to identify publications that describe RCSs for supporting older adults in health care and prevention and those that address ethical issues and recommendations regarding their development and implementation. The methodology used for the narrative review was inspired by the study by Green et al [55].

Inclusion and Exclusion Criteria

The review encompassed papers focusing on all populations, with particular attention to older adults. It focused on the concept of RCSs for health, while also incorporating publications discussing other health technologies for older adults if the authors have delved into relevant ethical considerations for their development or implementation.

The context of the review revolved around the use of RCSs (or related technologies), especially for older adults, across diverse living environments such as homes, hospitals, and nursing homes. Publications addressing RCSs and related ethical issues within the health care domain were considered, whereas those focusing solely on technical aspects (eg, AI and deep learning) or those outside the health care domain were excluded.

Various types of publications, including theoretical papers, research studies, and reviews, were included if they offered ethical reflections or recommendations for RCS use in health care. These reflections and recommendations were expected to align with the topics and issues of the ethical dimension of the EUnetHTA Core Model.

All publications, regardless of language (English or French), were eligible if published after 2011. This time frame was chosen considering the technological advancements over the past decade, which may have influenced the evolution of ethical issues and recommendations in the field of remote care systems and related technologies. Textbox 1 summarizes the inclusion and exclusion criteria adopted for the selection of papers in this review.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for the selection of publications about ethical issues regarding the use of robotic coaching solutions (RCSs) for older adults.

Inclusion criteria

- Types of participants: all populations
- Interventions or phenomena of interest: RCSs or other technologies used in health care, if ethical issues are discussed
- Context: the use of RCSs in the health care sector
- Paper type: all paper types (theoretical papers, research studies, and reviews) that discuss ethical issues
- Language: English or French
- Date of publication: after 2011

Exclusion criteria

- Types of participants: not applicable
- Interventions or phenomena of interest: RCSs or all other types of technology outside the health care sector
- Context: the use of RCSs in non–health care sectors
- Paper type: papers about RCSs and other technologies that are not dealing with ethical issues
- Language: all other languages
- Date of publication: before 2011

Search Strategy and Study Selection

The review was conducted using the following keywords: “seniors,” “older adults,” “social robots,” “assistive robots,” “assistive technology,” “robots,” “virtual coach,” “e-coaching,” “coaching system,” “coaching device,” “ethics,” and “recommendations.”

The review was performed between August 2021 and December 2021 using the PubMed, CINAHL, Embase, Scopus, Web of Science, IEEE Explore, SpringerLink, and PsycINFO databases. This search allowed us to find 4928 initial publications. Then, secondary research using references from other articles and the same inclusion criteria was conducted. This search allowed us to find 13 additional papers.

In total, 4943 papers were analyzed. The selection of the final publications was performed after reading the title and abstract first and, then, the full article. This selection process helped us to exclude irrelevant papers and duplicates (Figure 1). In total, 0.51% (25/4943) of the papers were included in our review.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing a summary of the screening and selection process of publications about ethical issues regarding the use of robotic coaching systems for older adults.

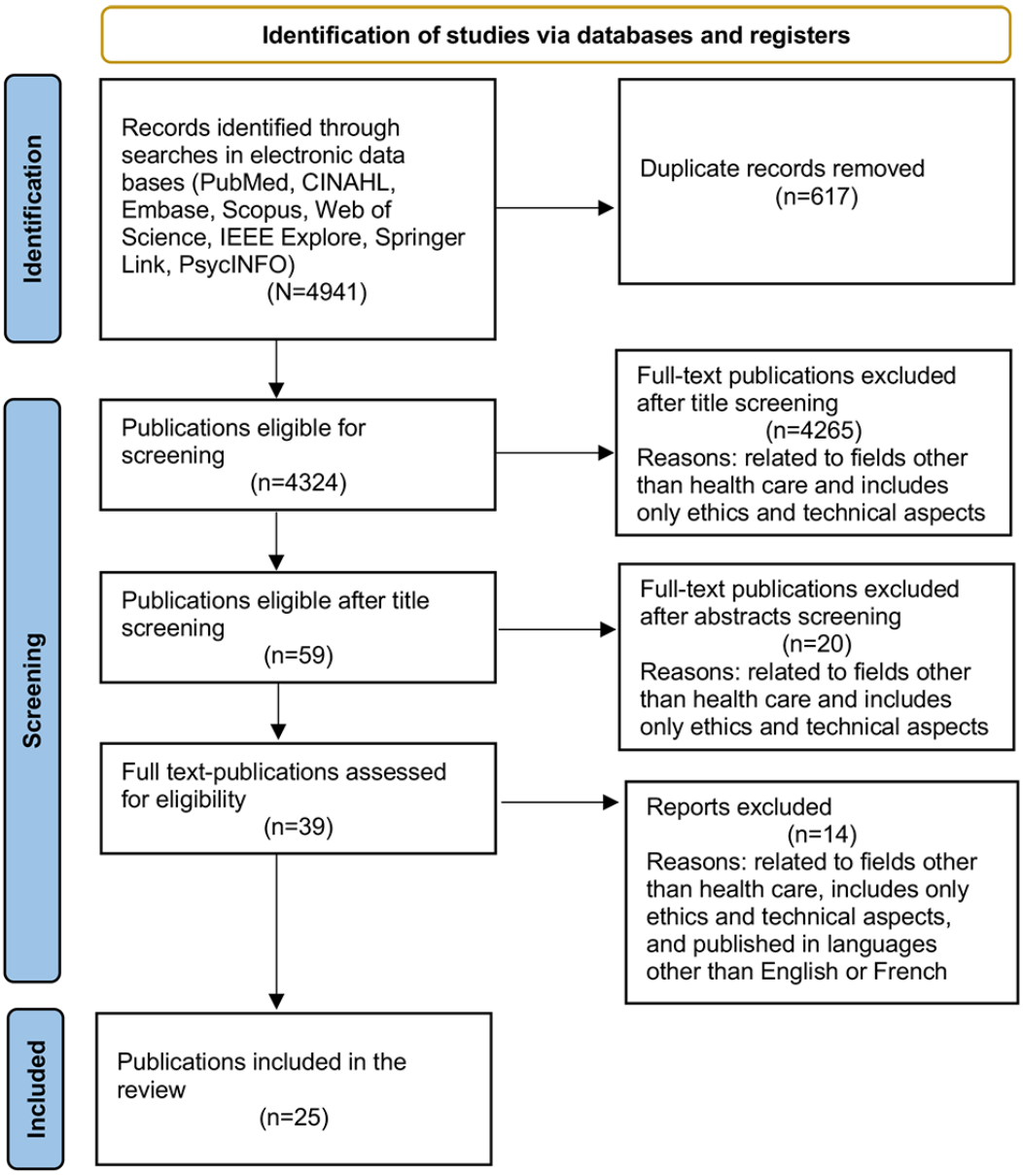

Data Analysis Criteria

The selected papers were analyzed using the ethical domain of the EUnetHTA Core Model [52]. Proper registration of the use of EUnetHTA Core Model for the purpose of this review was made on the HTA Core Model website [52].

The model was developed for the production and sharing of HTA information, allowing for the support of evidence-based decision-making in health care, but it can also be customized to other research needs. The EUnetHTA Core Model is composed of 9 domains, each including several topics. Each topic also includes different issues (ie, questions that should be considered for the evaluation of health technologies). Thus, the model is structured into 3 levels: *domain* (level 1), *topic* (level 2), and *issue* (level 3). The combination of a domain, topic, and issue is linked to an assessment element ID, which can be identified using a specific code for standardization purposes (B0001, B0002, etc).

The main EUnetHTA model domains include the following: (1) health and current use of the technology, (2) description and technical characteristics of the technology, (3) safety, (4) clinical effectiveness, (5) costs and economic evaluation, (6) ethical aspects, (7) organizational aspects, (8) patient and social aspects, and (9) legal aspects.

The ethical domain (level 1) in the EUnetHTA Core Model [52] includes 5 topics (level 2): “benefit-harm balance,” “autonomy,” “respect for people,” “justice and equity,” and “legislation.” Each of these topics includes several issues (level 3) [52].

In this study, 2 authors (CP and ASR) independently analyzed the 25 selected articles. First, they read the articles several times to improve familiarity with the ideas addressing the ethical

XSL•FO
RenderX

aspects of RCSs. Then, in each publication (methods, results, and discussion sections), they identified segments of data that were relevant or captured an idea linked to the "ethical" domain of the model. A subsequent exploration of the coded data (sentences or set of statements) was performed to get a more precise classification at the topic level (level 2) and at the issue level (level 3). Then, the coding was performed using the HTA nomenclature. The 2 experts (CP and ASR) compared their results. In a few cases, the coding results showed a lack of consensus between the 2 coding authors, which was resolved through a subsequent discussion between them. Interrater correlation was not calculated.

A thematic analysis using the EUnetHTA framework for conducting a literature review has been described in other studies [56,57]. Furthermore, the use of EUnetHTA to perform an ethical analysis of health technologies has already been proposed [58]. The 25 selected articles were all coded using this methodology. Some authors have previously emphasized the possibility of overlapping issues between topics in the HTA analysis. They have suggested to assess the overlapping issues in the most relevant topic section [59].

This review was not registered, and a protocol for the review was not prepared.

Results

Selected articles are presented in Multimedia Appendix 1 [3,12,18,31,32,37-51,60-64]. For each topic, we have presented our findings in terms of questions and recommendations according to the EUnetHTA Core Model, wherever possible.

Ethical Issues and Recommendations for the Use of New Technologies

This section aims to summarize the ethical analysis performed regarding the use of RCSs with older adults and to provide recommendations for ethical use of these devices. Table 1 presents a synthetic summary of the elements presented in this section.

XSL•FO
RenderX

Table 1. Summary of the main ethical issues and recommendations on the use of robotic coaching solution in older adults following the Health Technology Assessment–inspired analysis.

Topic and ethical issues (European Network of Health Technology Assessment Core Model)	Ethical concerns	Recommendations
Topic 1: benefit-harm balance		
What are the known and estimated benefits and harms for patients when implementing or not implementing the technology?	• Risk of social isolation • Risk of deception • Bias of algorithms	• To develop devices that promote human interaction • To provide documentation and to explain in a clear and understandable way the functioning of the devices but also their benefits in a clear and understandable manner • To explain the ins and out of the algorithms used in the device in detail
What are the benefits and harms of the technology for relatives, other patients, organizations, commercial entities, society, etc?	• Tensions with proxy • Ecological impact	• To evaluate the impact of technological devices on users • To evaluate the impact on the environment • To encourage material recycling • To develop energy-efficient technologies
Are there any unintended consequences of the technology and its application for patients?	• Quickly evolving technology • Unsuitability of technology	• To offer user simplified training to the user • To create devices adapted to the needs and preferences of end users (user-centered design)
Topic 2: autonomy		
Is the technology used for individuals who are especially vulnerable?	• Informed consent	• To ensure obtain consent from users before the use of a technological device • To offer advance directives and a proxy to sign consent • To ensure regular collection of consent and reminders of information about the technology
Does the implementation or use of the technology affect the patient's capability and possibility to exercise autonomy?	• Dependence on the technology • Loss of freedom • Creating a new source of authority	• To continually reassess the trust placed in technological devices • To regularly check that the system is adapted to the user and serves their interests
Topic 3: respect for persons		
Does the implementation or use of the technology affect human dignity?	• Stigmatization	• To promote positive information
Does the technology invade the sphere of privacy of the patient or user?	• Privacy • Security of data	• Users must have control over the technology and know where their data will be stored • The systems must also comply with the relevant data protection framework
Topic 4: justice and equity		
How does implementation or withdrawal of the technology affect the distribution of health care resources?	• Societal pressure • Digital divide • Inequalities in resources	• Ethical approval should be required obtained before technologies can be used • To ensure access to all, open-source development is preferred to ensure access to all people • To involve communities as much as possible
How are technologies with similar ethical issues treated in the health care system?	• Replacement of professionals	• Devices should only be used to enhance the work of caregivers, not to replace them • Caregivers must have the opportunity to train on the use of these devices
Topic 5: legislation		
Can the use of the technology pose ethical challenges that have not been considered in the existing legislations and regulations?	• Safety of devices • Regulation of technology	• To develop specific legal frameworks • An emergency stop button is required in these devices • An ethics committee must be created to help researchers develop ethical devices • Researchers need to keep abreast of regulatory developments

XSL•FO
RenderX

Topic 1: Benefit-Harm Balance

RCSs should be developed according to the principles of beneficence (ie, to promote the interest of users) and nonmaleficence (ie, to avoid inflicting harm) [39,60,64].

What Are the Known and Estimated Benefits and Harms for Patients When Implementing or Not Implementing the Technology?

Risk of Social Isolation

According to Sharkey and Sharkey [50], technological devices, when used appropriately, could benefit older adults by promoting social interaction and connection with their loved ones [4,31,40]. Broadbent et al [19] have discussed the potential of robots to reduce older adults' social isolation. However, other authors reported the negative influence of the use of robotic devices on human contact [31,32,65]. The use of robots (eg, telepresence robots) to make some cost savings (eg, reducing travel costs and time spent on trips for family and professionals to visit older adults) would reduce face-to-face interactions [3,36,39,40]. Moreover, according to Körtner [47], the more people become accustomed to communicating with robots, the less they will be used to communicating with humans. The use of social robots could lead to a reduction of interactions with humans and thus to social isolation and emotional dependence [39]. However, the influence of technological devices, such as RCSs, on social isolation is still under debate, and the impact of technology would depend on the manner in which it is used.

To avoid exacerbating the users' social isolation, Portacolone et al [38] advocate that social robots and similar technologies should be designed with the objective of fostering interactions with other humans, for instance, keeping users informed about the entertainment and socializing activities near their home, connecting them with their loved ones, and so on.

Risk of Deception

Another major risk for users is deception [39,64,66]. Portacolone et al [38] described 3 types of deception that people with neurocognitive disorders may face when interacting with social robotic systems but which may also apply to all users. The first type involves the user's misconception of what is driving the technological device [51]. Users may be misled if they think that behind a medical chatbot, there is a real physician who communicates and reads their messages [44] or, alternatively, if they are not aware that, at some point, there are real humans guiding the technological device [38]. The second type refers to robotic devices programmed to express feelings or other types of affective communication, which may lead the user to believe that the system's emotions are authentic. Related to this issue, Körtner [47] discussed how some older adults may fear that their social robot will forget them during their absence from home. The resemblance with the living in terms of affective behavior (eg, crying, laughing, or expressing concern) can make the user believe that there is a reciprocity between human and robot feelings [43]. The last type of deception is related to the inadequate interpretations that older adults may have regarding the nature of the robot, for example, thinking that an animal-shaped robot is a real animal or a pet [38]. Some current developments of social robots tend to make them resemble a living being, in terms of their verbal and nonverbal behaviors [34,60] or by highly anthropomorphizing their design [47], which may blur the boundary between the real and the artificial [45,60]. These design choices can also impact users' dignity by infantilizing them as they are led to believe in something that is false [50].

However, according to some researchers [51,63,64], the notion of deception should be considered in terms of the gradation between what is morally acceptable and what is not. Deception would be morally acceptable when it aims to improve a person's health or quality of life, for example, the use companion robots to calm a person experiencing behavioral disorders linked to dementia [51].

According to Danaher [43] and Vandemeulebroucke et al [40], to avoid deception, it is essential to be transparent to users about the design and operation of devices. As the information given to the participants is the basis for obtaining consent to use the technology, it is essential to offer them documents explaining how the device is built and its advantages and limitations in a clear manner adapted to the user's knowledge and experience. It is also important to inform users on how to behave with technology [12]. Researchers should also answer users' questions, pay attention to their feedback, and use it to improve the device and its documentation [60]. During experiments with RCSs, it is also important that researchers regularly remind participants of the nature of the technological device to reduce the risk of misinterpretation and to ensure that they still consent to participate in the study [38].

Biases of Algorithms

An autonomous device does not work without AI or algorithms that allow it to make decisions. However, these technologies are created by humans, and programming biases can be incorporated into them and lead to failures [44]. A technological device can, for instance, misread a situation and react accordingly, leading to a safety risk for the user [18]. Thus, it is essential that the researcher scrutinizes the algorithms used in RCSs before their implementation [44]. Fiske et al [44] also suggest providing the users with detailed explanations about the algorithms present in the technological device they are using.

What Are the Benefits and Harms of the Technology for Relatives, Other Patients, Organizations, Commercial Entities, Society, Etc?

At the society level, Boada et al [39] mentioned an ethical consideration related to the ecological impact of robotic devices in the current context of climate crisis and the lack of natural resources. The construction of RCSs requires raw materials, high energy consumption, and the management of their waste. Therefore, it is important for developers to design technologies that consume less energy and can be recycled.

Are There Any Unintended Consequences of the Technology and Its Application for Patients?

Technologies Evolving Very Quickly

For some older adults, technologies evolve very quickly, which makes it difficult for them to keep up with [62]. Denning et al [67] encourage designers to develop products that are intuitive

XSL•FO
RenderX

to use or to offer users a simplified training. However, although some technologies are progressing quickly, technological limitations are still present, especially regarding social robotic systems, impacting their performance [68] and generating frustration among some users [69].

Unsuitability of Technology

The lack of experience with the technologies and the fact that the systems are not suitable to everyone can reduce the usability and acceptability of RCSs among older adults [3,60,62]. Frennert and Östlund [62] highlighted that some older adults were not confident in their ability to handle a robot because of previous complicated experience with technology. Peek et al [70] also reported that users had doubts about their ability to use the technology and feared that they would easily forget how to use it. They may also fear false alarms generated by monitoring technologies. For example, a person may decide to sit on the floor, but this behavior can be considered as a fall by the technology, and it could call for an ambulance to be sent to the person's home in vain [70].

To promote acceptability and usability of RCSs, it is essential to develop them considering the capabilities, needs, and wishes of various users [31,47]. "User-centered design" approaches should be used for this purpose [71]. This methodology must be performed in a continuous manner to consider the development, new preferences, and experiences of the users. Technology assessment should also be conducted before deployment in ecological environments to improve the predictability of RCSs and decrease the risk of confusion and accidents [40,47].

Topic 2: Autonomy

According to Anderson and Kamphorst [42], the notion of autonomy implies the recognition of people, for instance, users of RCSs, as thinking individuals who have their own perspective on matters and are able to judge what is best for them.

Is the Technology Used for Individuals Who Are Especially Vulnerable?

Free and informed consent is a prerequisite for the involvement of an individual in research, regardless of the domain. This aspect is mentioned in numerous codes and declarations such as the Declaration of Helsinki (1964-2008) [72]. In the context of studies of the use of RCSs, this principle ensures that the person has freely chosen to use a device. However, some older adults, particularly those with cognitive disorders, may have difficulties in understanding and evaluating information related to RCSs and therefore in making appropriate choices [3]. Moreover, the person may not remember that the RCS is in their environment or how it works [38,44]. The question of how to ensure that the older adult has understood the purpose of RCS and that their choice of using the technology is based solely on their own decision and not that of a relative, caregiver, or institution has also been discussed [46].

Researchers in the field of RCS should adapt to the cognitive abilities of the populations they work with to facilitate communication and decision-making [46]. Thus, the observation of the person's behavior is necessary to identify potential reservations regarding the use of RCSs. When the person is very vulnerable to respond, informed consent could be sought by proxy such as from children, spouse, or partner [46,64]. However, according to Diaz-Orueta et al [37], the final decision of using RCSs lies with the user. To prevent loss of capacity and to guard against any risk of inducement to participate, advance directives [46,64] or implementation of an advance power of attorney [46] can be proposed.

Does the Implementation or Use of the Technology Affect the Patient's Capability and Possibility to Exercise Autonomy?

Dependence on the Technology

Although the main interest of RCSs for older adults is the maintenance of functional independence, it has been claimed that these devices could make people dependent on them. By replacing users in tasks that they can still perform, the use of RCSs could create new forms of vulnerability [3,31,39,41,51].

People could rely entirely on autonomous technological devices, such as RCSs, to guide their behaviors, goals, and actions [12,73]. A questioning of the authenticity of users' actions has been mentioned by Anderson and Kamphorst [42]. Users might not feel responsible for the success of their actions if they feel they are completely driven by the guidance of the RCS. People could also develop emotional and psychological feelings toward the technology. This may have negative consequences for the individuals [38,49] and lead to new vulnerabilities [39].

Loss of Freedom

Another ethical issue relates to the conflict between the user's safety, encouraged by the technology guidance, and a loss of freedom. The RCS could impose constraints on the user under the pretext that the user's actions are not good for them [39,40,74]. Sharkey and Sharkey [50] explained that to promote home care, RCS could act as a supervisor (ie, programmed to ensure that no danger is present and, if there is a danger, to implement procedures to stop it and avoid it in the future). For instance, the RCS could prevent the person from eating fatty and high-caloric food because it is harmful to them. To protect users and ensure that they live in good health, individuals using RCSs could end up being deprived of certain actions or being under some type of "house arrest" [50].

One of the goals of using such RCSs is to support older adults' independence; therefore, it is essential that developers and researchers in the field take measures to preserve the person's autonomy [75]. Furthermore, RCS users must have the opportunity to evaluate and re-evaluate the role given to the device, to assess whether the system is reliable and whether it is serving their interests [12,42].

Creating a New Source of Authority

The use of RCSs could alter human relationships, for example, by creating tensions between older adults and their informal caregivers. Their use could also create some tensions with health care professionals by creating a new source of authority [12]. Monitoring older adults through RCSs can generate anger in the user, for example, when the device insists that the older adult should take a medication that they do not want to take [41,75].

XSL•FO
RenderX

Topic 3: Respect for Persons

Does the Implementation or Use of the Technology Affect Human Dignity?

Human dignity may be affected by the use of RCSs as these technologies may be perceived as "problem evocators" [41]. Some RCSs are used to compensate for impaired capacities. However, according to Körtner [47], their use can make older adults aware of their limitations and lead to negative feelings, anxiety, or exhaustion. RCS use can also lead to a form of stigmatization by making one's own inabilities visible to others [3,70]. It is important to have positive communication regarding RCSs, to provide a less stigmatizing view of their use.

Does the Technology Invade the Sphere of Privacy of the Patient or User?

Privacy

To continue living at home, users are increasingly willing to tolerate intrusion in their privacy [70], but they are not always aware of when and how they are being monitored by RCSs [61]. Portacolone et al [38] provided the example of an animal-shaped companion robot, for which the older adults can signal that they no longer wish to interact with it by putting the robot to sleep. However, the animal-shaped robot can record data even when it is sleeping, but users are not always aware of this information. Forgetfulness and the lack of understanding of the device can lead to the risk of manipulation and coercion [44]. The person who is vulnerable may forget that they are being monitored and reveal personal information [50].

Technological devices, such as RCSs, must remain under the control of the users [47]. Users should have the ability to define when and where the device is used—when it collects data—to maintain their privacy, especially in intimate or private care settings.

Security of Data

According to Portacolone et al [38], remote monitoring technologies are usually controlled by third parties, sometimes even operating in another country, which can lead to cultural biases during the interaction between the older adult and the RCS. This context involves the risk that the person controlling the device (third party) takes advantage of the older adult's vulnerability to steal their personal information or exposes the user to financial abuse [38]. Older adults are not always aware or vigilant about the sharing and use of data, which may be personal and sensitive [73]. Furthermore, RCSs can be connected to internet services that collect, store, and transfer these sensitive data [47] for commercial use [49,61].

In addition, the use of technologies connected to digital networks involves the risk of hacking and unauthorized surveillance [34,51], which can make people vulnerable [62]. Denning et al [67] found that home robots could not only be remotely located and identified but also hacked and controlled. First, users may have either preconceived and erroneous ideas about the capabilities of the device or a lack of knowledge to evaluate the safety, especially regarding data protection [3]. Second, users do not always configure their technological device correctly or update them [67].

Encryption or security systems must be put in place to protect users' personal data captured by the devices at every stage: during collection, storage, transmission, and processing [3]. Researchers must also give particular attention to data security. In Europe, for instance, researchers and technology providers are required to comply with the General Data Protection Regulation [40,76]. Data collection must be performed legally or approved by the local relevant ethical committees.

To address data security challenges, 3 principles are recommended by Ienca et al [46] when developing technological devices: transparency, legitimate purpose, and proportionality. Transparency refers to the fact that the user knows that the system is collecting data and has consented to it. The user must also have precise information about when and what type of data are recorded and who has access to them [47]. Legitimate purpose refers to the notion that the monitoring and collection of data is performed for a specific purpose, (ie, in the best interest of the user or, if applicable, a relative who has consented to it). Finally, the principle of proportionality refers to the fact that the data collected are not disproportionate to the user's needs.

Topic 4: Justice and Equity

The consequences of the technology implementation on the distribution of health care resources was discussed in the literature.

How Does Implementation or Withdrawal of the Technology Affect the Distribution of Health Care Resources?

Societal Pressure

Socioeconomic issues are also linked to the development and use of RCSs can also be raised. Individual freedom may be hindered by the "incentive" of certain stakeholders or authorities to enforce the use of RCSs [37]. The use of RCSs and similar systems may also lead to a lesser involvement of relatives, caregivers, and institutions that provide care to older adults and to the reduction of care costs; these perceived economic benefits may pressurize older adults to consent to use these devices [40,46]. It is also possible that older adults may have to agree to use the technological device to receive other health care benefits (eg, aids and subsidies) [42].

Digital Divide

Different opportunities to access RCSs can result in digital divide, defined by the Organisation for Economic Cooperation and Development [77] as a gap between those who have access to information and communication technologies and those who do not. This difference can create educational, economic, social, and even health-related disparities among citizens. Some citizens would be able to use these devices and, therefore, could benefit from their advantages, whereas others will not be able to use them and will not enjoy their benefits. The use of technologies in the health care context, through public or private institutions, should be subject to previous authorization by independent ethical committees to ensure that the use of these devices will not harm users in any way.

XSL•FO
RenderX

Inequalities in Resources

Questions about justice, equity, and equality among all citizens also arise [12,40,46]. RCSs have relatively high costs [64] and can generate additional expenses such as an internet subscription [3] that only a part of the population can afford, and this may be owing to the lack of research allowing to measure the cost-to-benefit ratio of these technologies on health [32]. It is important to ensure the access to RCSs among different living areas (ie, urban and rural). Therefore, involving municipalities and neighborhood associations seems an interesting way of raising awareness about the opportunities offered by RCSs for older adults and reaching a wider range of people.

To promote justice, equity, and fair distribution, Ienca et al [46] and Wangmo et al [64] recommend reducing the development costs of RCSs by promoting an open dissemination of source codes. In addition, RCSs should be distributed in priority to those in greatest need; therefore, measures to ensure access to RCSs under fair conditions should be established [51]. Joachim [78] also suggests to cover some of the costs of these health care–oriented technologies through health insurance.

Recommendations have been published by researchers to improve equality of access to technologies, such as using open-source software, providing priority access for individuals with low income, or relying on certain collective financing systems such as retirement or health insurance [46,51,78]. Discussions must be conducted among developers, legislators, and private and public organizations to identify viable financing solutions that allow for fair distribution of RCSs.

Replacement of Professionals

Researchers have also reported fears expressed by older adults and caregivers about how the use of technological devices could eliminate care-related jobs or replace humans [17,34,48,61]. There are also concerns about the use of these technological tools to reduce health care costs by decreasing the number of available health care resources and services, thereby exacerbating social inequalities [44]. The introduction of health-oriented RCSs requires adapting the contexts of care practices, which may threaten their quality [39]. Their incorporation into the care work environment can be difficult because the devices are automated and some care situations are unpredictable [17,62]. Furthermore, the gestion of certain tasks by technological devices requires a restructuring of the roles and responsibilities of caregivers [39]. Fiske et al [44] highlight that there are currently no recommendations or training to enable health care professionals to adopt RCSs, even though these professionals are increasingly confronted with technological devices in their practice.

The incorporation of RCSs must always be accompanied by a discussion with concerned care professionals regarding the advantages and limits of the technology. Professionals must also be supported in the use of these devices through effective training. Structured training and supervision will contribute to the development of a controlled framework of practice around the use of RCSs and thus avoid potential abuse [44]. Moreover, to encourage their use among professionals, it is essential to clearly define the role of RCSs as an additional resource for professionals and not a replacement of human care services [44].

Topic 5: Legislation

The ethical challenges linked to the lack of existing legislations and regulations dedicated to the use of the technology were discussed in the literature.

Can the Use of the Technology Pose Ethical Challenges That Have Not Been Considered in the Existing Legislations and Regulations?

Safety of Devices

The use of RCSs by older adults can result in damage and harm to their environment [79], especially when the device is still at the prototype stage [47]. Safety risks linked to the use of RCSs (eg, malfunctioning of the technology and incorrect decisions made by the coaching system) arise when they share a common space with humans and interact with them [39]. The following questions must be considered: Who is responsible in case of an accident, and who pays for the damages [39,40,48,62,80]? Is it the designer, the device, or the user himself? Currently, the civil code favors the cascade system (ie, first, the liability falls on the designer of the product; then, on the developer; and finally, on the user who has not followed the rules of use) [74]. However, the more the machine becomes autonomous, the less the existing legal frameworks can answer these questions [80]. This is a key legal issue regarding the implementation of RCSs in real settings because the person responsible for damage to the user or the environment may incur legal or even penal proceedings.

Damage and prejudice can also be caused by a failure to share authority [45,49,60]. Who between the human and the technological device holds the power to make decisions and control a functionality [81]? According to Grinbaum et al [45], it is important to specify the circumstances in which the human must take control over the technological device (RCS) and those in which the device should decide autonomously. According to Riek and Howard [49], it is preferable that in certain cases, the technological device, although autonomous, requires a human validation of its actions to keep the user in control of the device. In addition, Bensoussan and Puigmal [80] suggested the idea that technological devices must have an emergency stop button, so that the human can switch off the technology at any time.

Regulation of Technology

Currently, there is a gray area between the capabilities of RCSs, the reality of the field, and the regulations in force [38]. To accompany the researcher during the whole process of development and diffusion of RCSs, an ethical framework should be established [18,60]. Specifically, this can be in the form of an ethical code of conduct illustrating the expectations to all the employees of a company [18]. The researcher must regularly inform themselves about the ethics to be consistent with the evolution of the regulatory framework [60]. However, according to Nevejans [82], these ethical recommendations have no legal value and cannot protect humans from the damage caused by new technologies. Thus, it is necessary to think about a new legal framework to protect the users of RCSs [37].

XSL•FO
RenderX

Discussion

Summary

The use of technologies, such as RCSs, in the health care field has grown significantly in recent years [17,18]. RCSs are increasingly being used for older adults with the aim of promoting healthy behaviors, quality of life, and well-being. However, the use of RCSs also raises several ethical challenges regarding the cost-to-benefit balance of these new care practices, respect for the autonomy of users, respect for privacy, justice and equity linked to their access, or need for a suitable legal framework. Such challenges could be addressed by establishing relevant recommendations for the development and use of RCSs. Some guidelines regarding the use of robotic systems have been published [49,83]. Moreover, in April 2021, the European Commission unveiled the first legal framework about AI [84]. However, to the best of our knowledge, no recommendations have been proposed in this field directly linked to an analysis of the literature dealing specifically with these ethical issues and potential solutions to address them.

This narrative review identified 25 articles in which authors highlighted ethical issues and recommendations related to the use of RCSs and similar technologies. The use of the EUnetHTA Core Model for the analysis of these articles made it possible to classify the information retrieved in the publications according to 5 main ethical topics—"benefit-harm balance," "autonomy," "respect for persons," "justice and equity," and "legislation"—and to provide a detailed analysis of RCS-related ethical issues. Our review also aimed to identify recommendations for better development, diffusion, and use of RCSs by a population of older adults.

Technology devices, such as RCSs, are used with older adults to enable them to live independently; to enhance their quality of life and well-being; and, therefore, to cope with the increasing care demands for older populations. RCSs may be used to encourage a range of health-related goals: physical, cognitive, nutritional, social, and emotional domains. To be effective, RCSs must be able to motivate the user by providing highly personalized care programs [85,86]. However, studies have shown that not all potential target users are included in the development of these devices [37,87,88]. Therefore, RCSs design might fail to meet a wide range of users' needs, capabilities, and wishes. Thus, it is essential to apply "user-centered design" approaches and involve target users with various sociodemographic characteristics and technology experience throughout the development process. A strong involvement of the intended users of these systems in their design process would also improve the quality of the information provided to potential users of RCSs regarding their operation, type of data collected, and potential benefits of the technology. In this way, the involvement of the users would improve the quality of the process of obtaining the consent required from older adults to use the technology.

Another ethical challenge related to the use of RCSs is the fact that their wide implementation for older adults' care may affect the distribution of health care resources. For instance, it has been found that for some older adults and informal and formal caregivers, the use of RCSs could replace humans in many caregiving tasks, eventually leading to a suppression of jobs or to a degradation of the quality of health care services [17,34,48,61]. In this regard, the participation of a third person (professional, volunteer, or family member) as a "human coach" could be considered when implementing RCSs in the older adults' environment. This "human coach" could help build a "chain of trust" by being an intermediary between the RCS and the user. On the one hand, the involvement of a real person in the use of the RCS could reduce the risk of replacement of human assistance by technological assistance. On the other hand, the "human coach" could help enhance the acceptability and usability of the device, while at the same time, reassuring the user and providing recommendations to the developers, so that the RCS is consistent with users' needs and desires. However, the benefits of involving a "human coach" in the RCS service provision has yet to be evaluated by scientific studies.

According to some studies [3,39,41,51,65], the use of RCSs can have an impact on social relationships, reducing human contact and even altering social relationships by creating tension between older adults and their caregivers. Thus, it would be interesting to identify the repercussions and implications of these devices in older adults' daily life and in the life of the members of their social environment through new studies. It also seems necessary to evaluate the organizational impact of the implementation of RCSs and to identify potential obstacles to their use in the care professionals' work context.

Our analysis also confirmed that for RCSs to provide personalized health-related recommendations, the collection of sensitive data is necessary. Data collection in this context also raises several ethical issues. For instance, personal data can be exposed to hacking and misuse. Proper data management, anonymization, and encryption are essential to protect the personal data of RCS users [86]. In addition, researchers and developers in this field must evaluate RCSs before implementation to ensure that they do not cause physical or moral harm to users. Thus, it has been suggested that stakeholders refer to local and regional regulatory and safety standards to guide their development and use.

Finally, our analysis also discussed how legal and ethical frameworks regarding the use of RCSs need to be adapted to cope with the constant development of new technologies. So far, existing legal frameworks are not yet adequate to respond effectively to the question of liability in case of damage caused by RCSs, particularly because these devices are becoming increasingly autonomous [80]. The establishment of "operational ethics committees in digital sciences and technologies" could help in the development and conduct of projects in this area [60]. Guidelines should be established to identify the types of applications and technological devices that require regulatory review and approval [44]. Research projects and working groups involving users, researchers, and lawyers should be set up to further investigate the legal and ethical issues related to the use of RCSs.

Some countries and regions, such as Europe and Japan have initiated the work of structuring relevant legal and ethical frameworks; however, their orientations and measures may

XSL•FO
RenderX

differ culturally [78]. Future studies in the area of RCSs could consider the influence of cultural and socioeconomic specificities of the contexts of experimentation (countries and regions) regarding the acceptance and use of RCSs by older adults and formal and informal caregivers and regarding the definition of ethical and legal frameworks governing their uses. Therefore, the use of validated and widely applied analysis frameworks, for example, the Western, Educated, Industrialized, Rich and Democratic framework [89], formulated to measure countries' commonalities in their approaches to the interpretation of behavioral research findings (eg, regarding technology adoption) could be interesting. The Western, Educated, Industrialized, Rich and Democratic framework [89] could help not only to explore the differences among countries regarding the validation and adoption of new technologies for older adult care but also to seek greater cultural and demographic diversity in technology research.

This dimension of cross-cultural comparison has received particular attention in the framework of a current international research partnership between Europe and Japan, such as the EU-Japan Virtual Coach for Smart Ageing (e-VITA) project. This project aims to develop a cross-cultural RCS that can be tailored to the needs of healthy older adults to promote aging well. The e-VITA RCS will be made available to older adults in their homes, which raises many of the ethical questions discussed in this paper. Therefore, the study will require the researchers to set up procedures adapted not only to the users but also to the 2 cultures (European and Japanese), respecting the corresponding ethical and legal regulations. Thus, it would be interesting to perform an analysis of the ethical issues raised by users from different countries and cultures within the framework of the e-VITA project.

Limitations

A narrative review of the literature was conducted to provide a nonexhaustive synthesis of the various ethical concerns and recommendations when using RCSs for older adults. This review has some limitations. Only articles in French and English were included. Some articles indicating ethical concerns or recommendations may not have been included when this information was not mentioned in the keywords or abstract.

Conclusions

The use of RCSs in the context of health care, particularly with an older adult population, tends to show many benefits. RCSs have the potential to improve the quality of life of older adults and their independence. When used in an ethical and appropriate manner, RCSs can help improve older adults' emotions and cognitive and physical abilities and promote social relationships. By helping older adults to continue living at home for as long as possible, the use of health-oriented RCSs could help to address some of the challenges resulting from demographic aging. However, the use of these new health care technologies involves some ethical concerns, with the most cited issues being not only the risk of accidents, lack of reliability, loss of control, risk of deception, and risk of social isolation but also the confidentiality of data and liability in case of safety problems.

Some recommendations have been made in the past regarding the use of social and assistive robotic technologies for older adults, such as considering the opinion of target users; collecting their consent; training the care professionals to use them; and ensuring proper data management, anonymization, and encryption. However, the integration of RCSs in current health practices and, particularly, in the private homes of older adults can be disruptive. It requires the establishment of scalable and adapted ethical and regulatory frameworks that follow the technology progress and the social and digital change of society Thus, studies are needed to identify new ethical concerns arising from the organizational impact of the implementation of RCSs in different contexts, especially in the homes of older adults. The influence of cultural and socioeconomic specificities of the contexts of experimentation (countries and regions) regarding the acceptance and use of RCSs by older adults and formal and informal caregivers is also an area of interest for future studies.

Acknowledgments

This paper is a part of the EU-Japan Virtual Coach for Smart Ageing (e-VITA) project, which aims to develop a robotic coaching system for older adults [90]. The authors thank the collaborators who made this project possible: European Commission and Assistance Publique–Hôpitaux de Paris (Délégation à la Recherche Clinique et à l'Innovation). This review was based on data collected within the e-Vita project, funded by the European Union H2020 Program (grant 101016453) and the Japanese Ministry of Internal Affairs and Communication (Ministry of Internal Affairs and Communication; grant JPJ000595).

Data Availability

Data sharing is not applicable to this paper as no data sets were generated or analyzed during this review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The selected relevant articles on ethical issues regarding the use of robotic coaching solutions for older adults.
[DOCX File , 20 KB - jmir_v26i1e48126_app1.docx]

XSL•FO
RenderX

References

1. United Nations. Croissant à un rythme plus lent, la population mondiale devrait atteindre 9,7 milliards d'habitants en 2050 et pourrait atteindre près de 11 milliards vers 2100 : Rapport de l'ONU. United Nations, Population Division.: United Nations; 2019. URL: https://population.un.org/wpp/Publications/Files/WPP2019_PressRelease_FR.pdf [accessed 2022-03-04]
2. Rapport mondial sur le vieillissement et la santé. Organisation Mondiale de la Santé. 2016. URL: https://apps.who.int/iris/handle/10665/206556 [accessed 2022-03-04]
3. Chung J, Demiris G, Thompson HJ. Ethical considerations regarding the use of smart home technologies for older adults: an integrative review. Annu Rev Nurs Res 2016;34:155-181. [doi: 10.1891/0739-6686.34.155] [Medline: 26673381]
4. Ganesan B, Gowda T, Al-Jumaily A, Fong KN, Meena SK, Tong RK. Ambient assisted living technologies for older adults with cognitive and physical impairments: a review. Eur Rev Med Pharmacol Sci 2019 Dec;23(23):10470-10481 [FREE Full text] [doi: 10.26355/eurrev_201912_19686] [Medline: 31841201]
5. Palazzolo J, Baudu C, Quaderi A. Psychothérapies du Sujet Âgé: Prise en Charge des Pathologies du Vieillissement. 2nd edition. Amsterdam, The Netherlands: Elsevier Masson; 2016.
6. Principaux repères de l'OMS sur la démence. World Health Organization. URL: https://www.who.int/fr/news-room/fact-sheets/detail/dementia [accessed 2022-03-04]
7. Guilbaud A, Mailliez A, Boulanger É. Vieillissement: une approche globale, multidimensionnelle et préventive. Med Sci (Paris) 2020 Dec 09;36(12):1173-1180 [FREE Full text] [doi: 10.1051/medsci/2020224] [Medline: 33296634]
8. op den Akker H, Klaassen R, op den Akker R, Jones VM, Hermens HJ. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013 Presented at: CBMS '13; June 20-22, 2013; Porto, Portugal p. 546-547 URL: https://ieeexplore.ieee.org/document/6627870 [doi: 10.1109/cbms.2013.6627870]
9. Görer B, Salah AA, Akın HL. A robotic fitness coach for the elderly. In: Augusto JC, Wichert R, Collier R, Keyson D, Salah AA, Tan AH, editors. Ambient Intelligence: Lecture Notes in Computer Science. Volume 8309. Cham, Switzerland: Springer; 2013.
10. Lete N, Beristain A, García-Alonso A. Survey on virtual coaching for older adults. Health Informatics J 2020 Dec 01;26(4):3231-3249 [FREE Full text] [doi: 10.1177/1460458220926894] [Medline: 32744137]
11. Siewiorek D, Smailagic A, Dey A. Architecture and applications of virtual coaches. Proc IEEE 2012 Aug;100(8):2472-2488. [doi: 10.1109/JPROC.2012.2200560]
12. Kamphorst BA. E-coaching systems: what they are, and what they aren't. Pers Ubiquit Comput 2017 May 23;21(4):625-632. [doi: 10.1007/s00779-017-1020-6]
13. Kyriazakos S, Schlieter H, Gand K, Caprino M, Corbo M, Tropea P, et al. A novel virtual coaching system based on personalized clinical pathways for rehabilitation of older adults-requirements and implementation plan of the vCare project. Front Digit Health 2020 Oct 5;2:546562 [FREE Full text] [doi: 10.3389/fdgth.2020.546562] [Medline: 34713034]
14. Wrobel J, Pino M, Wargnier P, Rigaud AS. Robots et agents virtuels au service des personnes âgées : une revue de l'actualité en gérontechnologie. NPG Neurol Psychiatr Geriatr 2014 Aug;14(82):184-193. [doi: 10.1016/j.npg.2014.02.007]
15. Pérez PJ, Garcia-Zapirain B, Mendez-Zorrilla A. Caregiver and social assistant robot for rehabilitation and coaching for the elderly. Technol Health Care 2015;23(3):351-357. [doi: 10.3233/THC-150896] [Medline: 25669209]
16. Fasola J, Mataric MJ. A socially assistive robot exercise coach for the elderly. J Hum-Robot Interact 2013;2(2):3-32. [doi: 10.5898/jhri.2.2.fasola]
17. Cresswell K, Cunningham-Burley S, Sheikh A. Health care robotics: qualitative exploration of key challenges and future directions. J Med Internet Res 2018 Jul 04;20(7):e10410 [FREE Full text] [doi: 10.2196/10410] [Medline: 29973336]
18. Winfield AF, Jirotka M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos Trans A Math Phys Eng Sci 2018 Oct 15;376(2133):20180085 [FREE Full text] [doi: 10.1098/rsta.2018.0085] [Medline: 30323000]
19. Broadbent E, Stafford R, MacDonald B. Acceptance of healthcare robots for the older population: review and future directions. Int J of Soc Robotics 2009 Oct 3;1(4):319-330. [doi: 10.1007/s12369-009-0030-6]
20. Abou Allaban A, Wang M, Padır T. A systematic review of robotics research in support of in-home care for older adults. Information 2020 Jan 30;11(2):75. [doi: 10.3390/info11020075]
21. Cortellessa G, De Benedictis R, Fracasso F, Orlandini A, Umbrico A, Cesta A. AI and robotics to help older adults: revisiting projects in search of lessons learned. Paladyn J Behav Robot 2021;12(1):356-378. [doi: 10.1515/pjbr-2021-0025]
22. Lima MR, Wairagkar M, Gupta M, Rodriguez y Baena F, Barnaghi P, Sharp DJ, et al. Conversational affective social robots for ageing and dementia support. IEEE Trans Cogn Dev Syst 2022 Dec;14(4):1378-1397. [doi: 10.1109/tcds.2021.3115228]
23. Sidner CL, Bickmore T, Nooraie B, Rich C, Ring L, Shayganfar M, et al. Creating new technologies for companionable agents to support isolated older adults. ACM Trans Interact Intell Syst 2018 Jul 24;8(3):1-27. [doi: 10.1145/3213050]
24. Pepito JA, Locsin RC, Constantino RE. Caring for older persons in a technologically advanced nursing future. Health 2019;11(05):439-463. [doi: 10.4236/health.2019.115039]
25. Pinheiro PR, Pinheiro PG, Filho RH, Barrozo JP, Rodrigues JJ, Pinheiro LI, et al. Integration of the mobile robot and internet of things to monitor older people. IEEE Access 2020;8:138922-138933. [doi: 10.1109/access.2020.3009167]
26. Chatterjee A, Gerdes MW, Martinez S. eHealth initiatives for the promotion of healthy lifestyle and allied implementation difficulties. In: Proceedings of the 2019 International Conference on Wireless and Mobile Computing, Networking and

XSL•FO
RenderX

Communications. 2019 Presented at: WiMOB '19; October 21-23, 2019; Barcelona, Spain p. 1-8 URL: https://ieeexplore.ieee.org/document/8923324 [doi: 10.1109/WiMOB.2019.8923324]
27. Yousuf H, Reintjens R, Slipszenko E, Blok S, Somsen GA, Tulevski II, et al. Effectiveness of web-based personalised e-Coaching lifestyle interventions. Neth Heart J 2019 Jan 28;27(1):24-29 [FREE Full text] [doi: 10.1007/s12471-018-1200-7] [Medline: 30488381]
28. Bevilacqua R, Casaccia S, Cortellessa G, Astell A, Lattanzio F, Corsonello A, et al. Coaching through technology: a systematic review into efficacy and effectiveness for the ageing population. Int J Environ Res Public Health 2020 Aug 15;17(16):5930 [FREE Full text] [doi: 10.3390/ijerph17165930] [Medline: 32824169]
29. Justo R, Ben Letaifa L, Palmero C, Gonzalez-Fraile E, Torp Johansen A, Vázquez A, et al. Analysis of the interaction between elderly people and a simulated virtual coach. J Ambient Intell Human Comput 2020 May 22;11(12):6125-6140. [doi: 10.1007/s12652-020-01983-3]
30. Fiorini L, De Mul M, Fabbricotti I, Limosani R, Vitanza A, D'Onofrio G, et al. Assistive robots to improve the independent living of older persons: results from a needs study. Disabil Rehabil Assist Technol 2021 Jan 22;16(1):92-102. [doi: 10.1080/17483107.2019.1642392] [Medline: 31329000]
31. Sriram V, Jenkinson C, Peters M. Informal carers' experience of assistive technology use in dementia care at home: a systematic review. BMC Geriatr 2019 Jun 14;19(1):160 [FREE Full text] [doi: 10.1186/s12877-019-1169-0] [Medline: 31196003]
32. Pilotto A, Boi R, Petermans J. Technology in geriatrics. Age Ageing 2018 Nov 01;47(6):771-774. [doi: 10.1093/ageing/afy026] [Medline: 29546366]
33. Koceska N, Koceski S, Beomonte Zobel P, Trajkovik V, Garcia N. A telemedicine robot system for assisted and independent living. Sensors (Basel) 2019 Feb 18;19(4):834 [FREE Full text] [doi: 10.3390/s19040834] [Medline: 30781647]
34. Broadbent E. Interactions with robots: the truths we reveal about ourselves. Annu Rev Psychol 2017 Jan 03;68(1):627-652. [doi: 10.1146/annurev-psych-010416-043958] [Medline: 27648986]
35. Suwa S, Tsujimura M, Ide H, Kodate N, Ishimaru M, Shimamura A, et al. Home-care professionals' ethical perceptions of the development and use of home-care robots for older adults in Japan. Int J Hum Comput Interact 2020 Mar 13;36(14):1295-1303. [doi: 10.1080/10447318.2020.1736809]
36. Nordgren A. How to respond to resistiveness towards assistive technologies among persons with dementia. Med Health Care Philos 2018 Sep 6;21(3):411-421 [FREE Full text] [doi: 10.1007/s11019-017-9816-8] [Medline: 29214555]
37. Diaz-Orueta U, Hopper L, Konstantinidis E. Shaping technologies for older adults with and without dementia: reflections on ethics and preferences. Health Informatics J 2020 Dec;26(4):3215-3230 [FREE Full text] [doi: 10.1177/1460458219899590] [Medline: 31969045]
38. Portacolone E, Halpern J, Luxenberg J, Harrison KL, Covinsky KE. Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment: a call for interdisciplinary collaborations. J Alzheimers Dis 2020;76(2):445-455 [FREE Full text] [doi: 10.3233/JAD-190952] [Medline: 32250295]
39. Boada JP, Maestre BR, Genís CT. The ethical issues of social assistive robotics: a critical literature review. Technol Soc 2021 Nov;67:101726. [doi: 10.1016/j.techsoc.2021.101726]
40. Vandemeulebroucke T, Dierckx de Casterlé B, Gastmans C. The use of care robots in aged care: a systematic review of argument-based ethics literature. Arch Gerontol Geriatr 2018 Jan;74:15-25. [doi: 10.1016/j.archger.2017.08.014] [Medline: 28926749]
41. Zafrani O, Nimrod G. Towards a holistic approach to studying human-robot interaction in later life. Gerontologist 2019 Jan 09;59(1):e26-e36. [doi: 10.1093/geront/gny077] [Medline: 30016437]
42. Anderson J, Kamphorst B. Ethics of e-coaching: implications of employing pervasive computing to promote healthy and sustainable lifestyles. In: Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communication Workshops. 2014 Presented at: PerComW '14; March 24-28, 2014; Budapest, Hungary p. 351-356. [doi: 10.1109/percomw.2014.6815231]
43. Danaher J. Robot betrayal: a guide to the ethics of robotic deception. Ethics Inf Technol 2020 Jan 04;22(2):117-128. [doi: 10.1007/s10676-019-09520-3]
44. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
45. Grinbaum A, Chatila R, Devillers L, Ganascia JG, Tessier C, Dauchet M. Ethics in robotics research: CERNA mission and context. IEEE Robot Automat Mag 2017 Sep;24(3):139-145. [doi: 10.1109/mra.2016.2611586]
46. Ienca M, Jotterand F, Vică C, Elger B. Social and assistive robotics in dementia care: ethical recommendations for research and practice. Int J of Soc Robotics 2016 Jun 22;8(4):565-573. [doi: 10.1007/s12369-016-0366-7]
47. Körtner T. Ethical challenges in the use of social service robots for elderly people. Z Gerontol Geriatr 2016 Jun 25;49(4):303-307. [doi: 10.1007/s00391-016-1066-5] [Medline: 27220734]
48. Operto F. Ethics in advanced robotics. IEEE Robot Automat Mag 2011 Mar;18(1):72-78. [doi: 10.1109/mra.2010.940155]

XSL•FO
RenderX

49. Riek LD, Howard DA. A Code of Ethics for the Human-Robot Interaction Profession. In: Proceedings of 2014 Conference on We Robot. 2014 Presented at: WeRobot '14; April 4-5, 2014; Coral Gables, FL p. 1 URL: https://robots.law.miami.edu/2014/wp-content/uploads/2014/03/a-code-of-ethics-for-the-human-robot-interaction-profession-riek-howard.pdf
50. Sharkey A, Sharkey N. Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf Technol 2010 Jul 3;14(1):27-40. [doi: 10.1007/s10676-010-9234-6]
51. Yew GC. Trust in and ethical design of carebots: the case for ethics of care. Int J Soc Robot 2021 May 23;13(4):629-645 [FREE Full text] [doi: 10.1007/s12369-020-00653-w] [Medline: 32837630]
52. HTA Core Model ® version 3. EUnetHTA. URL: http://www.corehta.info/model/HTACoreModel3.0.pdf [accessed 2022-03-04]
53. Health technology assessment process: fundamentals. EUPATI Toolbox. 2015. URL: https://toolbox.eupati.eu/resources/health-technology-assessment-process-fundamentals/ [accessed 2022-03-04]
54. Tuli TB, Terefe TO, Rashid MM. Telepresence mobile robots design and control for social interaction. Int J Soc Robot 2021 Jul 13;13(5):877-886 [FREE Full text] [doi: 10.1007/s12369-020-00676-3] [Medline: 32837631]
55. Green BN, Johnson CD, Adams A. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. J Chiropr Med 2006 Sep;5(3):101-117. [doi: 10.1016/s0899-3467(07)60142-6]
56. Isabet B, Pino M, Lewis M, Benveniste S, Rigaud AS. Social telepresence robots: a narrative review of experiments involving older adults before and during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Mar 30;18(7):3597 [FREE Full text] [doi: 10.3390/ijerph18073597] [Medline: 33808457]
57. Naudé B, Rigaud AS, Pino M. Video calls for older adults: a narrative review of experiments involving older adults in elderly care institutions. Front Public Health 2021 Jan 14;9:751150 [FREE Full text] [doi: 10.3389/fpubh.2021.751150] [Medline: 35096731]
58. Saarni S, Hofmann B, Lampe K, Lühmann D, Mäkelä M, Velasco-Garrido M, et al. Ethical analysis to improve decision-making on health technologies. Bull World Health Organ 2008 Aug 01;86(8):617-623 [FREE Full text] [doi: 10.2471/blt.08.051078] [Medline: 18797620]
59. Lampe K, Mäkelä M, Garrido MV, Anttila H, Autti-Rämö I, Hicks NJ, et al. The HTA core model: a novel method for producing and reporting health technology assessments. Int J Technol Assess Health Care 2009 Dec 23;25(S2):9-20. [doi: 10.1017/s0266462309990638]
60. Collectif C. Éthique de la recherche en robotique. ALLISTENE. 2014. URL: http://cerna-ethics-allistene.org/digitalAssets/38/38704_Avis_robotique_livret.pdf [accessed 2024-04-02]
61. Tisseron S. Introduction. In: Tisseron S, Tordo F, editors. Robots, de Nouveaux Partenaires de Soins Psychiques. Fairfield CA: Érès Publication; 2018.
62. Frennert S, Östlund B. How do older people think and feel about robots in health- and elderly care? In: Proceedings of the 2018 INBOTS Conference on Inclusive Robotics for a Better Society. 2018 Presented at: INBOTS '18; October 16-18, 2018; Pisa, Italy p. 167-174 URL: https://link.springer.com/chapter/10.1007/978-3-030-24074-5_28 [doi: 10.13140/RG.2.2.13289.13928]
63. van Maris A, Zook N, Caleb-Solly P, Studley M, Winfield A, Dogramadzi S. Designing ethical social robots-a longitudinal field study with older adults. Front Robot AI 2020 Jan 24;7:1 [FREE Full text] [doi: 10.3389/frobt.2020.00001] [Medline: 33501170]
64. Wangmo T, Lipps M, Kressig RW, Ienca M. Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders. BMC Med Ethics 2019 Dec 19;20(1):98 [FREE Full text] [doi: 10.1186/s12910-019-0437-z] [Medline: 31856798]
65. Portet F, Vacher M, Rossato S. Les technologies de la parole et du TALN pour l'assistance à domicile des personnes âgées : un rapide tour d'horizon (quick tour of NLP and speech technologies for ambient assisted living). In: Proceedings of the 2012 Workshop on Interactions Langagières pour personnes Agées Dans les habitats Intelligents. 2012 Presented at: ILADI '12; June 4-8, 2012; Grenoble, France URL: https://aclanthology.org/W12-1402.pdf
66. Bradwell HL, Winnington R, Thill S, Jones RB. Ethical perceptions towards real-world use of companion robots with older people and people with dementia: survey opinions among younger adults. BMC Geriatr 2020 Jul 14;20(1):244 [FREE Full text] [doi: 10.1186/s12877-020-01641-5] [Medline: 32664904]
67. Denning T, Matuszek C, Koscher K, Smith JR, Kohno T. A spotlight on security and privacy risks with future household robots: attacks and lessons. In: Proceedings of the 11th international conference on Ubiquitous computing. 2009 Presented at: UbiComp '09; September 30-October 3, 2009; Orlando, FL p. 105-114 URL: https://dl.acm.org/doi/10.1145/1620545.1620564 [doi: 10.1145/1620545.1620564]
68. Papadopoulos I, Koulouglioti C, Lazzarino R, Ali S. Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: a systematic review. BMJ Open 2020 Jan 09;10(1):e033096 [FREE Full text] [doi: 10.1136/bmjopen-2019-033096] [Medline: 31924639]
69. Pripfl J, Kortner T, Batko-Klein D. Results of a real world trial with a mobile social service robot for older adults. In: Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction. 2016 Presented at: HRI '16; March 7-10, 2016; Christchurch, New Zealand p. 497-498 URL: https://ieeexplore.ieee.org/document/7451824 [doi: 10.1109/hri.2016.7451824]

XSL•FO
RenderX

70. Peek ST, Wouters EJ, van Hoof J, Luijkx KG, Boeije HR, Vrijhoef HJ. Factors influencing acceptance of technology for aging in place: a systematic review. Int J Med Inform 2014 Apr;83(4):235-248 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.01.004] [Medline: 24529817]
71. Lespinet-Najib V, Roche A, Chibaudel Q. Santé et handicap: d'une conception centrée «utilisateur» à la conception universelle. Ann Mines Réal Ind 2017;2:25-27. [doi: 10.3917/rindu1.172.0025]
72. Déclaration d'Helsinki de l'AMM – Principes éthiques applicables à la recherche médicale impliquant des êtres humains. L'Association Médicale Mondiale. URL: https://www.wma.net/fr/policies-post/declaration-dhelsinki-de-lamm-principes-ethiques-applicables-a-la-recherche-medicale-impliquant-des-etres-humains/ [accessed 2022-02-03]
73. Adams S, Niezen M. Digital 'solutions' to unhealthy lifestyle 'problems': the construction of social and personal risks in the development of eCoaches. Health Risk Soc 2016 Feb 06;17(7-8):530-546. [doi: 10.1080/13698575.2015.1136409]
74. Saerens P. Le droit des robots, un droit de l'homme en devenir ? Commun Technol Dév 2020 Jun 30(8). [doi: 10.4000/ctd.2877]
75. Jenkins S, Draper H. Care, monitoring, and companionship: views on care robots from older people and their carers. Int J of Soc Robotics 2015 Sep 15;7(5):673-683. [doi: 10.1007/s12369-015-0322-y]
76. Règlement (UE) 2016/679 du parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE). Le Parlement Européen Et Le Conseil De L'union Européenne. 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj/fra [accessed 2022-02-03]
77. Understanding the digital divide. OECD Digital Economy Papers. 2001. URL: https://econpapers.repec.org/paper/oecstiaab/49-en.htm [accessed 2024-04-02]
78. Joachim C. Silver economy, robotique et droit - Comparaison franco-japonaise. Centre d'études et de Coopération Juridique Interdisciplinaire – Université de Poitier. 2020 Jan. URL: https://hal.science/hal-03147937/document [accessed 2024-04-05]
79. Kernaghan K. The rights and wrongs of robotics: ethics and robots in public organizations. Can Public Adm 2014 Dec 12;57(4):485-506. [doi: 10.1111/capa.12093]
80. Bensoussan A, Puigmal L. Le droit des robots ? Quelle est l'autonomie de décision d'une machine ? Quelle protection mérite-t-elle ? Arch Philos Droit 2017;59(1):165-174. [doi: 10.3917/apd.591.0202]
81. Tessier C. Robots autonomy: some technical issues. In: Lawless WF, Mittu R, Sofge D, Russell S, editors. Autonomy and Artificial Intelligence: A Threat or Savior?. Cham, Switzerland: Springer; 2017:179-194.
82. Nevejans N. Comment protéger l'homme face aux robots ? Arch Phil Droit 2017;59(1):131-163. [doi: 10.3917/apd.591.0168]
83. Tambornino L, Lanzerath D, Rodrigues R, Wright D. SIENNA D4.3: survey of REC approaches and codes for artificial intelligence and robotics. Zenodo. 2019 Aug 19. URL: https://zenodo.org/records/4067990 [accessed 2024-04-04]
84. Ethics guidelines for trustworthy AI. Shaping Europe's Digital Future. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [accessed 2022-07-15]
85. Kamali ME, Angelini L, Caon M, Carrino F, Rocke C, Guye S, et al. Virtual coaches for older adults' wellbeing: a systematic review. IEEE Access 2020;8:101884-101902. [doi: 10.1109/access.2020.2996404]
86. Banos O, Nugent C. E-coaching for health. Computer 2018 Mar;51(3):12-15. [doi: 10.1109/mc.2018.1731070]
87. Gustafson DH, McTavish F, Gustafson DH, Mahoney JE, Johnson RA, Lee JD, et al. The effect of an information and communication technology (ICT) on older adults' quality of life: study protocol for a randomized control trial. Trials 2015 Apr 25;16(1):3-8. [doi: 10.1186/S13063-015-0713-2]
88. Sakaguchi-Tang DK, Cunningham JL, Roldan W, Yip J, Kientz JA. Co-design with older adults: examining and reflecting on collaboration with aging communities. Proc ACM Hum Comput Interact 2021 Oct 18;5(CSCW2):1-28. [doi: 10.1145/3479506]
89. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun 15;33(2-3):61-83. [doi: 10.1017/s0140525x0999152x]
90. EU-Japan virtual coach for smart ageing. e-VITA Consortium. URL: https://www.e-vita.coach/ [accessed 2024-04-05]

Abbreviations

AI: artificial intelligence
EUnetHTA: European Network of Health Technology Assessment
e-VITA: EU-Japan Virtual Coach for Smart Ageing
HTA: Health Technology Assessment
RCS: robotic coaching solution

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.04.23; peer-reviewed by J Sedlakova, S Liu; comments to author 20.08.23; revised version received 22.12.23; accepted 12.03.24; published 18.06.24.

Please cite as:
Palmier C, Rigaud AS, Ogawa T, Wieching R, Dacunha S, Barbarossa F, Stara V, Bevilacqua R, Pino M
Identification of Ethical Issues and Practice Recommendations Regarding the Use of Robotic Coaching Solutions for Older Adults: Narrative Review
J Med Internet Res 2024;26:e48126
URL: https://www.jmir.org/2024/1/e48126
doi:10.2196/48126
PMID:38888953

©Cécilia Palmier, Anne-Sophie Rigaud, Toshimi Ogawa, Rainer Wieching, Sébastien Dacunha, Federico Barbarossa, Vera Stara, Roberta Bevilacqua, Maribel Pino. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Advanced Technology in a Real-World Rehabilitation Setting: Longitudinal Observational Study on Clinician Adoption and Implementation

Louise Michelle Nettleton Pearce[1,2,3], PhD; Julie Pryor[1,4], PhD; Jason Redhead[1], BPhty; Catherine Sherrington[2,5], PhD; Leanne Hassett[2,3,6], PhD

[1]Royal Rehab Group, Sydney, Australia
[2]Institute for Musculoskeletal Health, Sydney Local Health District, Sydney, Australia
[3]Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[4]Susan Wakil School of Nursing and Midwifery, The University of Sydney, Sydney, Australia
[5]Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[6]Implementation Science Academy, Sydney Health Partners, Sydney, Australia

Corresponding Author:
Louise Michelle Nettleton Pearce, PhD
Institute for Musculoskeletal Health
Sydney Local Health District
Level 10N, King George V Building
Missenden Road, Camperdown
Sydney
Australia
Phone: 61 8627 6447
Email: louise.pearce@sydney.edu.au

Abstract

Background: Advanced technologies are becoming increasingly accessible in rehabilitation. Current research suggests technology can increase therapy dosage, provide multisensory feedback, and reduce manual handling for clinicians. While more high-quality evidence regarding the effectiveness of rehabilitation technologies is needed, understanding of how to effectively integrate technology into clinical practice is also limited. Current implementation of rehabilitation technology is inconsistent, with low uptake among clinicians and frequent reports of technology abandonment. An Australian rehabilitation provider opened a new technology therapy center in 2022, offering a unique opportunity to generate practice-based evidence to inform future technology implementation and research.

Objective: This study aimed to investigate the implementation and adoption of advanced technology within a real-world rehabilitation setting.

Methods: This study was a longitudinal observational study in a rehabilitation organization with inpatient, outpatient, and community settings. Allied health clinicians (n=119) within the organization had access to advanced technologies, with patients receiving neurological, spinal cord injury, brain injury, or general rehabilitation. Interventions included 21 advanced technologies, including robotic, virtual reality (VR), sensor-based, and functional electrical stimulation devices. Clinicians received training for devices in a staged approach by external and internal trainers. Data were collected from patient electronic medical records from July 1, 2022, to June 30, 2023. Outcomes included frequency of advanced technology use, patient demographics (age, gender, and primary health condition), clinician discipline, rehabilitation service (inpatient, outpatient, or community), goals of technology therapy sessions, and therapy dosage achieved (minutes active, number of repetitions, and meters walked).

Results: Clinicians used advanced technology 4208 times with 269 patients over 12 months; specifically, physiotherapists (2716/4208, 65%), occupational therapists (1396/4208, 33%), and allied health assistants (96/4208, 2%). The majority of patients had stroke, spinal cord injury, or brain injury diagnoses (188/269, 70%). Devices were typically used to target impairment and activity limitation–related goals. Frequently used devices included gait training body-weight support (VR treadmill and overground), overground robotic exoskeletons, and upper limb robotic VR devices. Outpatient services were the dominant users of advanced technology (3940/4208, 94%). Clinicians most commonly used devices for patients with stroke (1973/4208, 47%) and the greatest variety of devices for patients with stroke and spinal cord injury. The relative use of lower limb robotic devices was greater in

XSL•FO
RenderX

inpatient services (91/178, 51%, vs outpatient services, 963/2335, 41%) (χ^2_1=6.6, *P*=.01) and for patients with spinal cord injury (48/95, 51%, vs all other conditions, between 24%-31%; χ^2_5=16.8, *P*=.005).

Conclusions: The type and amount of advanced technology use differed between patient populations and rehabilitation settings. To support clinician use of advanced technology, devices should match the rehabilitation context. Tailored strategies are important, such as clinician training. Further practice-based research is required to provide guidance on implementation and to establish the effectiveness of advanced technology use.

(J Med Internet Res 2024;26:e60374) doi:10.2196/60374

KEYWORDS

rehabilitation; technology; digital health; virtual reality; robotics; exoskeleton device; implementation science; physiotherapy; physical therapy; occupational therapy; mobile phone

Introduction

There is burgeoning interest in harnessing the power of digital interventions to enhance health care services [1]. Such interventions have potential to help address the growing need for rehabilitation worldwide [2,3]. Within rehabilitation, this may involve devices such as virtual reality, robotics, smartphone apps and activity trackers [4]. While evidence is emerging, research has shown digital interventions can improve patient outcomes (eg, in mobility and upper limb function) [5-9], improve patient engagement in rehabilitation [10-13], and increase therapy dosage [6,14,15].

As technologies advance and become more affordable [16-18], they are increasingly accessible in rehabilitation [10,19,20]. Despite this, low device uptake among clinicians and frequent reports of technology abandonment persist [19,21-24]. Addressing these issues is imperative to avoid wasting health care resources on unsuccessful technology implementation efforts [24,25]. With a paucity of research conducted within rehabilitation settings, key issues include a poor understanding in selecting devices for clinical settings, who to use devices with and how to effectively integrate technology with current rehabilitation practices [1,19,26-28].

A new technology therapy center was opened by an Australian-based rehabilitation provider in 2022. This study is based within the center, which incorporates a range of advanced technologies, defined by the provider as “new technological interventions used in therapy to empower physical, cognitive, social or communicative skill development.” Advanced features include robotic componentry to maneuver limbs (eg, powered and nonpowered exoskeleton and end-effector devices) [29,30], immersive and nonimmersive virtual reality (VR) [6,31], multichannel functional electrical stimulation (FES), built-in sensors which provide real-time feedback, or a combination of these features.

Current evidence for advanced technologies is in its infancy, demonstrating potential benefits for additional therapy dosage [6,8,15,32], multisensory feedback and reducing physical demands on therapists, especially when working with patients who have little to no active motor control in major muscle groups [4,13,22,30]. In the landscape of rapidly evolving technology, research cannot afford to wait for effectiveness to be definitively established before examining implementation within clinical settings. For complex interventions, evidence of effectiveness alone is not adequate to facilitate integration into practice due to the interplay of factors such as device novelty, substantial learning requirements, and extensive organizational adaptation [24,33]. Investigating the real-world implementation of a range of advanced technologies presents a unique opportunity to advance this research field, generating practice-based evidence to directly inform technology use within clinical rehabilitation and future large-scale trials [34].

The aim of this study was to investigate the implementation and adoption of advanced technology within a real-world clinical rehabilitation setting. Specific research questions regarding this rehabilitation setting were as follows:

1. Which advanced technologies did clinicians use?
2. Who used the advanced technology?
3. Why did clinicians use advanced technology?
4. What therapy dosage did patients achieve with advanced technology?

Methods

Study Design

This longitudinal observational study of usual care is reported according to the Reporting of Studies Conducted using Observational Routinely-collected health Data (RECORD) statement [35].

Setting and Participants

The study site is a combined private-public rehabilitation setting in Australia which serves adults with neurological and nonneurological conditions, specializing in spinal cord and brain injury. Rehabilitation services include 5 combined outpatient and community services, 3 inpatient units, and 2 community only services. The technology therapy center opened in July 2022 and all allied health clinicians in the organization (n=119) had access to devices.

Clinicians received training on devices before and after center opening. Training requirements for each device varied and are outlined in Multimedia Appendix 1; training timeframes ranged from 2 hours to more than 35 hours. Various models of training were used, including external trainers, web-based modules, internal train-the-trainer, and communities of practice. Clinicians had clinical support, ongoing training, and technical support

XSL•FO
RenderX

through an internal advanced technology clinical lead and ad hoc support from technology suppliers.

Intervention: Advanced Technologies

Between September and October 2020, the advanced technology clinical lead, senior allied health clinicians and management collaborated to finalize a list of technologies for purchase, based on current Australian Therapeutic Goods Administration–approved rehabilitation devices and using a selection criteria tool [36]. Another element of device selection was data security. Devices were only selected if data could be stored locally or on secure cloud-based software in Australia. Devices were acquired between November 2020 and February 2022. At center opening, 25 advanced technologies were available, 21 of which were physical rehabilitation devices. Multimedia Appendix 1 contains details of these devices, categorized in Textbox 1.

Textbox 1. Device classifications categorized by lower limb, upper limb and other devices.

Lower limb devices (n=8)

- Body-weight support treadmill with virtual reality (BWS-T-VR)
- Body-weight support overground (BWS-OG)
- Robotic overground (Robotic-OG)
- Sensor-based virtual reality (Sensors-VR)
- Robotic treadmill body-weight support virtual reality (Robotic-T-BWS-VR)
- Robotic functional electrical stimulation (robotic functional electrical stimulation [FES])

Upper limb devices (n=9)

- Robotic-virtual reality (VR)
- Sensors-VR

Other devices (n=4)

- Augmented VR
- Immersive VR
- Advanced FES

Data Sources and Variables

Data were recorded by clinicians and extracted weekly through a Business Intelligence report, which searched all patient electronic medical records (Kyra Clinical, Telstra Health, Melbourne, Australia) organization-wide from July 1, 2022, to June 30, 2023. Progress note entries which included an “advanced technology therapy session” template were eligible for inclusion. This template includes device-specific subheadings to prompt clinicians to record device data (refer to Multimedia Appendix 2) and was introduced as usual practice before commencing data collection. Our only exclusion criterion was any progress note entry which detailed use of a device not outlined in Multimedia Appendix 1 (eg, single or dual-channel electrical stimulation devices), resulting in exclusion of 54 progress note entries. Data extracted included: session date, patient details (medical record number, age, gender, primary health condition, and clinical service), clinician discipline, device used, location, technology session goals, duration (total time, active time [eg, walk time]) and steps, repetitions, or meters walked. In addition to medical record data, nonidentifiable service data regarding number of staff trained per device were collected. Variables used for analysis are detailed in Textbox 2.

XSL•FO
RenderX

Textbox 2. Details of variables used for analysis.

Types of advanced technologies

- Advanced technologies used categorized into lower limb, upper limb, or other, with subcategories outlined in Textbox 1

Patient data

- Age, gender, and primary health condition

Clinician and rehabilitation service

- Clinician discipline and rehabilitation service type categorized into inpatient (onsite), outpatient (onsite; including combined outpatient and community services, or community services who saw patients onsite), or community (offsite)

Goals of advanced technology therapy session

- Clinician-reported therapy session goals coded and categorized according to the International Classification of Functional, Disability and Health (ICF) framework [37], as addressing "impairment" (in body function or structure), "activity limitation," or "participation restriction"

Therapy dosage achieved

- Therapy dosage was calculated based on active therapy time (minutes spent active during the session), number of repetitions completed (number of repetitions completed of an exercise or number of steps, where one step is equivalent to one repetition [38]), or number of meters walked (if applicable)

Statistical Analysis

Data analysis for this study was conducted in R (version 4.3.1; R Core Team) and consisted of summary statistics to describe device use by device category, patient characteristics, clinician discipline, and clinical service, each month and across the 12-month period. Missing data within included progress note entries were also recorded and summarized descriptively. Chi-square tests of independence were conducted to investigate associations between type of device used and patient primary health condition, as well as rehabilitation service. A significance level of .05 was used for chi-square tests.

Ethical Considerations

Ethics approval for this study was provided by Northern Sydney Local Health District Human Research Ethics Committee on May 19, 2022 (2022/ETH00364). A waiver of consent was granted to collect data from patients' electronic medical records, including patients' medical record numbers to allow analysis of the number of sessions and variety of different technologies used by each patient. However, no personal information regarding patients or clinicians was collected and no individual participants are identifiable in any findings presented in this study.

Results

Overall Advanced Technology Use and Staff Training

Across the 12 months, clinicians used advanced technologies 4208 times with 269 patients (14% of all patients). All rehabilitation services within the organization accessed devices for patient therapy sessions. In total, 80 allied health clinicians (80/119, 67%) were trained across devices (n=41 outpatient and community clinicians, n=29 inpatient clinicians and n=10 clinicians working across inpatient, outpatient, and community services); including 35 physiotherapists, 33 occupational therapists, and 4 allied health assistants.

Which Advanced Technologies Did Clinicians Use?

Our data captured clinicians using 20 of 21 (95%) physical rehabilitation devices. Monthly device use gradually increased, excluding holiday months (Figure 1).

Lower limb devices were most frequently used (2513/4208, 60%). The most used device was an augmented VR treadmill with optional BWS (580/4208, 14%). Devices for upper limb training were also frequently used (1270/4208, 30%); most commonly, a robotic VR device for distal upper limb joints (374/4208, 9%). Table 1 and Table 2 detail device use across patient populations and clinical services.

XSL•FO
RenderX

Figure 1. Monthly device use across the 12-month period (July 2022 to June 2023).

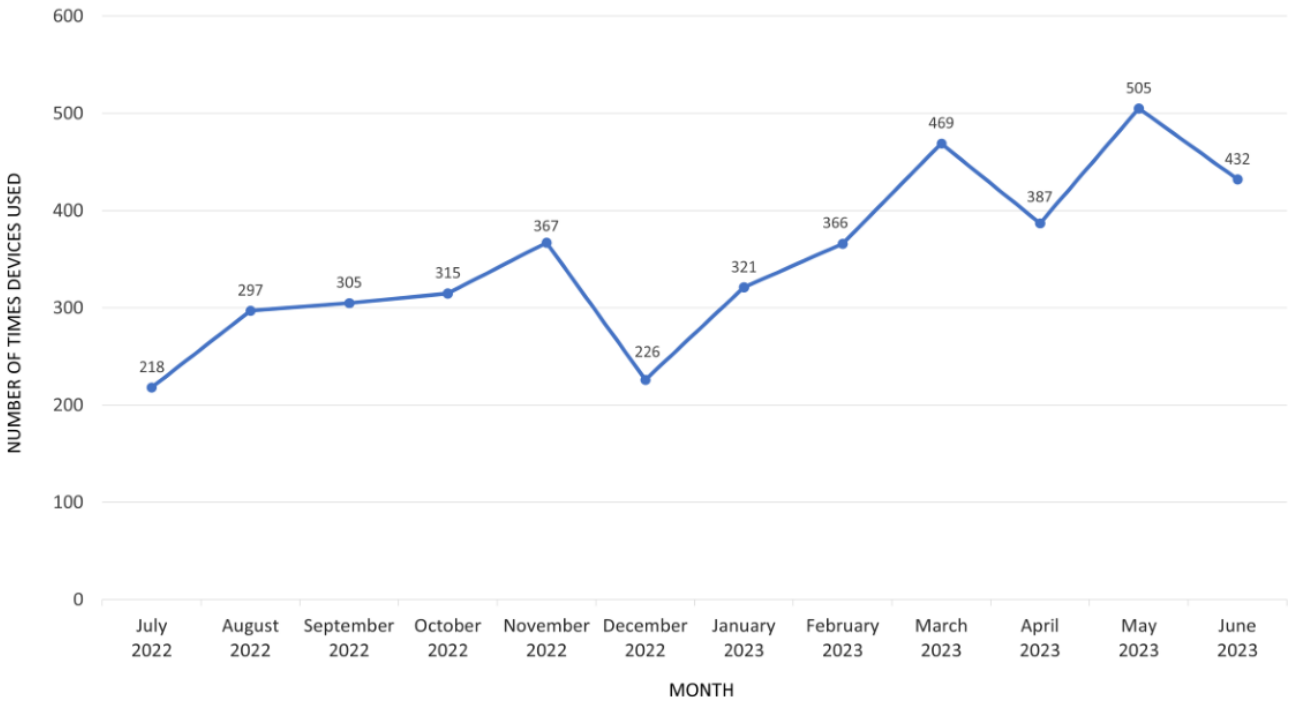

XSL•FO
RenderX

Table 1. Technology use across the 12-month period by patient primary health condition.

Device category (Number of devices)	Patient primary health condition							Total
	Stroke	SCI[a]	Brain injury	Other acquired neurological	Progressive neurological	Congenital neurological	Other	
Lower limb, n (%)[b]								
BWS[c]-T[d]-VR[e] (1)	254 (13)	84 (10)	125 (26)	14 (13)	55 (9)	9 (7)	39 (42)	580 (14)
BWS-OG[f] (2)	225 (11)	69 (8)	117 (24)	14 (13)	95 (16)	23 (19)	12 (13)	555 (13)
Robotic-OG (1)	239 (12)	151 (18)	45 (9)	7 (7)	90 (15)	5 (4)	0 (0)	537 (13)
Sensors-VR (2)	119 (6)	64 (8)	31 (6)	16 (15)	68 (11)	12 (10)	14 (15)	324 (8)
Robotic-T-BWS-VR (1)	106 (5)	71 (9)	19 (4)	3 (3)	61 (10)	4 (3)	0 (0)	264 (6)
Robotic-FES[g] (1)	40 (2)	127 (15)	19 (4)	1 (1)	63 (10)	3 (3)	0 (0)	253 (6)
Total lower limb (8)	__[h]	—	—	—	—	—	—	2513 (60)
Upper limb, n (%)[h]								
Robotic-VR (4)	562 (28)	106 (13)	32 (7)	6 (6)	53 (9)	26 (21)	1(1)	786 (19)
Sensors-VR (4)	251 (13)	77 (9)	23 (5)	28 (27)	53 (9)	32 (26)	20 (22)	484 (12)
Total upper limb (8)	—	—	—	—	—	—	—	1270 (30)
Other, n (%)[h]								
Augmented VR (1)	116 (6)	33 (4)	43 (9)	4 (4)	38 (6)	3 (2)	2 (2)	239 (6)
Advanced FES (2)	47 (2)	39 (5)	33 (7)	2 (2)	16 (3)	4 (3)	4 (4)	145 (3)
Immersive VR (1)	14 (1)	5 (1)	0 (0)	10 (10)	12 (2)	0 (0)	0 (0)	41 (1)
Total other (4)	—	—	—	—	—	—	—	425 (10)
Total, n	1973	826	487	105	604	121	92	4208

[a]SCI: spinal cord injury.
[b]Percentages are based on column counts.
[c]BWS: body weight support.
[d]T: treadmill.
[e]VR: virtual reality.
[f]OG: overground.
[g]FES: functional electrical stimulation.
[h]Not applicable.

XSL•FO
RenderX

Table 2. Technology use across the 12-month period by clinical service.

Device category (number of devices)	Clinical service			Total
	Inpatient (onsite)	Outpatient (on-site)	Community (offsite)	
Lower limb, n (%)[a]				
BWS[b]-T[v]-VR[d] (1)	13 (5)	567 (14)	0 (0)	580 (14)
BWS-OG[e] (2)	74 (28)	481 (12)	0 (0)	555 (13)
Robotic-OG (1)	50 (19)	487 (12)	0 (0)	537 (13)
Sensors-VR (2)	0 (0)	324 (8)	0 (0)	324 (8)
Robotic-T-BWS-VR (1)	9 (3)	255 (6)	0 (0)	264 (6)
Robotic-FES[f] (1)	32 (12)	221 (6)	0 (0)	253 (6)
Total lower limb (8)	__[g]	—	—	2513 (60)
Upper limb, n (%)[a]				
Robotic-VR (4)	47 (18)	739 (19)	0 (0)	786 (19)
Sensors-VR (4)	8 (3)	472 (12)	4 (100)	484 (12)
Total upper limb (8)	—	—	—	1270 (30)
Other, n (%)[a]				
Augmented VR (1)	17 (6)	222 (6)	0 (0)	239 (6)
Advanced FES (2)	13 (5)	132 (3)	0 (0)	145 (3)
Immersive VR (1)	1 (0)	40 (1)	0 (0)	41 (1)
Total other (4)	—	—	—	425 (10)
Total	264	3940	4	4208

[a]Percentages are based on column counts.
[b]BWS: body weight support.
[c]T: treadmill.
[d]VR: virtual reality.
[e]OG: overground.
[f]FES: functional electrical stimulation.
[g]Not applicable.

Who Used the Advanced Technology?

Clinicians

Physiotherapists were the highest users of advanced technology (2716/4208, 65%), followed by occupational therapists (1396/4208, 33%), and allied health assistants (96/4208, 2%). Occupational therapists gradually increased device use each month, relative to physiotherapists (from 25% in the first month to 48% by the twelfth month, Figure 2).

XSL•FO
RenderX

Figure 2. Monthly device use across the 12-month period (July 2022 to June 2023) by clinician discipline, demonstrating an increasing proportion of device use by occupational therapists.

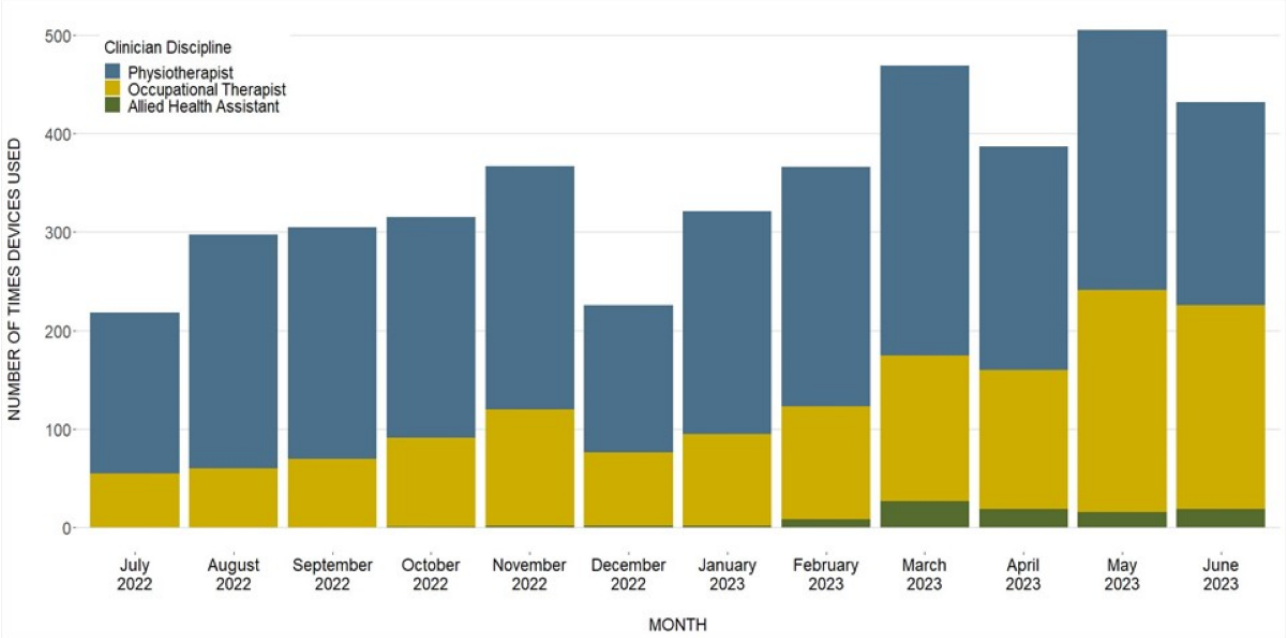

Patients

Over 12 months, 445 inpatients and 1502 outpatient and community patients received rehabilitation services. Out of the 269 patients with whom clinicians used advanced technology, the average age of patients was 52.3 years (SD 14.1, range 14-83 years), 95 were female (35%) and 255 had neurological diagnoses (95%); predominantly stroke, spinal cord injury, or brain injury (71%). Table 3 displays patient demographics.

The majority of patients (178/269, 66%) used more than 1 device, including 20% of whom used 5 or more devices. Approximately 45% of patients had 5 or less sessions with advanced technology (120/269), 23% had 6 to 10 sessions (62/269) and 32% had more than 10 sessions (87/269). Data on repeated device use (ie, the same patient using a given device more than once) demonstrated a median of 5 (IQR 1-7) repeated uses overall across devices. The Robotic-T-BWS-VR device had the highest repeated uses (median 7, IQR 5-10), where 89% (32/36) of patients who used this device did so more than once. This was followed closely by the Robotic-OG device (median 7, IQR 4-12), where 81% (52/64) of patients who used this device did so more than once. Details regarding repeated use across each device are summarized in Multimedia Appendix 3.

There is strong evidence that use of robotic versus nonrobotic lower limb devices is associated with patient primary health condition (χ^2_5=16.8, *P*=.005). For patients with spinal cord injury, clinicians used robotic and nonrobotic lower limb devices equally. For all other conditions, clinicians typically used nonrobotic lower limb devices (between 69% and 76%). Figure 3 depicts these differences.

For upper limb devices, there is no evidence of an association between type of device used (robotic vs nonrobotic) and primary health condition (χ^2_4=1.1, *P*=.78). Almost all patients with nonneurological conditions only used sensor and BWS devices, with only 1 having used a robotic upper limb device.

XSL•FO
RenderX

Table 3. Demographics of patients who clinicians used advanced technology with (n=269).

Patient demographics	Values
Patient age (years) mean (SD), range	52.3 (14.1), 14-83
Patient sex, n (%)	
Male	174 (65)
Female	95 (35)
Patient primary health condition, n (%)	
Stroke	88 (33)
Spinal cord injury	66 (25)
Brain injury	34 (13)
Multiple sclerosis	15 (6)
Neuro-oncology	9 (3)
Cerebral palsy	8 (3)
Musculoskeletal	8 (3)
Parkinson disease	7 (3)
Muscular dystrophy	6 (2)
Other neurological conditions	22 (8)
Other	6 (2)
Patient primary health condition by group[a]; n (%)	
Acquired	216 (80)
Progressive	42 (16)
Congenital	11 (4)

[a]Acquired conditions include stroke, spinal cord injury, brain injury, musculoskeletal, neuro-oncology, functional neurological disorder, amputee, polio, autoimmune disease, spinal cord injury and brain injury (combined), acute transverse myelitis, deconditioning, cauda equina, and burns. Progressive conditions include multiple sclerosis, Parkinson disease, muscular dystrophy, Multiple systems atrophy, ataxia, spinal muscular atrophy, autosomal dominant adrenoleukodystophy, motor neurone disease, Charcot Marie tooth disease, Parkinson syndrome, neuromyelitis optica, and myositis. Congenital conditions include cerebral palsy, spinal bifida, and heredity spastic paraplegia.

Figure 3. Proportion of patients using robotic versus nonrobotic lower limb devices, grouped by patient primary health condition.

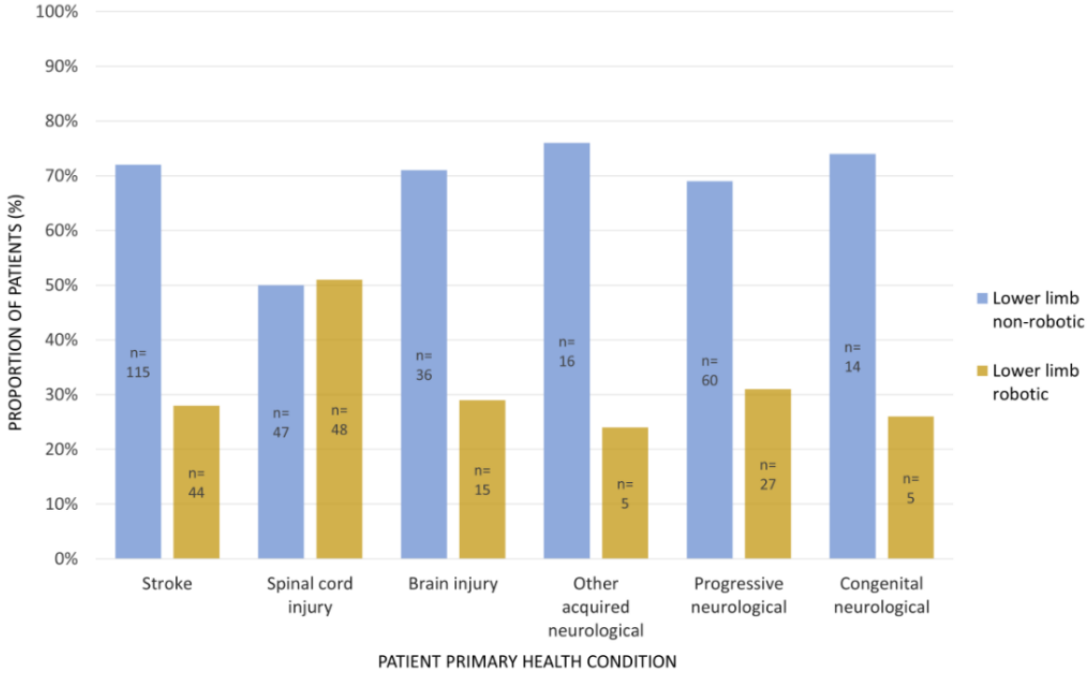

XSL•FO
RenderX

Rehabilitation Service Type

Outpatient rehabilitation services were dominant users of advanced technology (3936/4208, 94%). Inpatient services used advanced technology 264 times (6%), while offsite community services used advanced technology 4 times (0.1%). Inpatient clinicians used advanced technology with 49 of 445 patients (11%), while onsite outpatient and community clinicians used advanced technology with 226 of 1502 patients (15%).

There is strong evidence of an association between type of lower limb device used (robotic vs nonrobotic) and rehabilitation service (χ^2_1=6.6, *P*=.01). Outpatient clinicians used more nonrobotic lower limb devices (1372/2335, 59%), whereas inpatient clinicians used robotic (91/178, 51%) and nonrobotic (87/178, 49%) lower limb devices equally. There is also very strong evidence of an association between type of upper limb device used (robotic vs nonrobotic) and rehabilitation service (χ^2_1=13.3, *P*<.001). Inpatient clinicians used robotic upper limb devices (47/55, 85%) considerably more than nonrobotic upper limb devices (8/55, 15%). Meanwhile, outpatient clinicians also used robotic upper limb devices (739/1211, 61%) more than nonrobotic upper limb devices (472/1211, 39%), however the difference was not as prominent. Figure 4 portrays these differences. The relative use of augmented or immersive VR and advanced FES devices was similar across all onsite services. Offsite community services only used portable upper limb sensor devices (4/4, 100%).

Figure 4. Proportion of sessions in which robotic versus nonrobotic lower and upper limb devices were used, grouped by rehabilitation service.

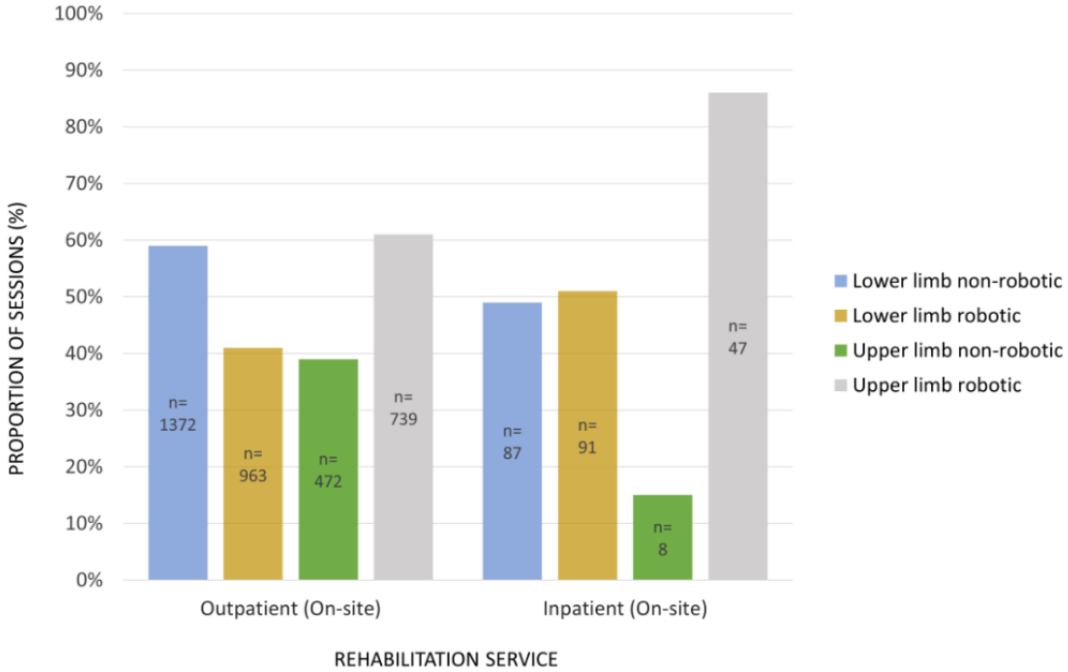

Why Did Clinicians Use Advanced Technology?

Clinicians reported 144 different goals for using advanced technology in patient therapy sessions. On average, clinicians reported 1.8 goals per session (SD 1.2, range 1-5), with only 41 sessions (41/4208, 1%) missing goal documentation. Goals were predominantly at the impairment (5063/8499, 60%) or activity-limitation level (2600/8499, 31%). Although more participation-level goals were reported as the year progressed, these only formed 1% of goals across the 12 months (most commonly return to employment [16/8499, 0.2%]). Clinicians frequently cited goals related to gait training (1959/8499, 23%), lower limb strengthening (555/8499, 7%) and balance (537/8499, 6%). Upper limb goals, particularly fine motor control (942/8499, 11%), range of motion (827/8499, 10%), and strengthening (690/8499, 8%) were also frequently reported. Table 4 details advanced technology therapy session goals.

XSL•FO
RenderX

Table 4. Advanced technology therapy session goals.

Type of goal	Sessions, n (%)	ICF[a] categories (sessions), n (%)			
		Impairment	Activity limitations	Participation restriction	Other
Lower limb	3821 (45)	2581 (30)	987 (12)	__[b]	253 (3)
Gait training	1959 (23)	1211 (14)	498 (6)	—	250 (3)
Strengthening	555 (7)	555 (7)	—	—	—
Balance	537 (6)	537 (6)	—	—	—
Lower limb–specific activities[c]	489 (6)	—	489 (6)	—	—
Range of motion	127 (1)	127 (1)	—	—	—
Spasticity	92 (1)	92 (1)	—	—	—
Coordination	44 (1)	44 (1)	—	—	—
Sensory	13 (0.15)	13 (0.15)	—	—	—
Other	5 (0.06)	2 (0.02)	—	—	3 (0.04)
Upper limb	3092 (36)	2000 (24)	853 (10)	—	239 (3)
Fine motor control	942 (11)	283 (3)	659 (8)	—	—
Range of motion	827 (10)	824 (10)	3 (0.04)	—	—
Strengthening	690 (8)	690 (8)	—	—	—
Spasticity	150 (2)	150 (2)	—	—	—
Gross motor control	97 (1)	—	97 (1)	—	—
Upper limb–specific activities[d]	94 (1)	—	94 (1)	—	—
Sensory	52 (0.6)	52 (0.6)	—	—	—
Unspecified or other	240 (3)	1 (0.01)	—	—	239 (3)
ADLs[e] and participation	650 (8)	—	598 (7)	52 (1)	—
Personal ADL	497 (6)	—	497 (6)	—	—
Domestic ADL	71 (1)	—	71 (1)	—	—
Occupation-related	23 (0.3)	—	—	23 (0.3)	—
Recreation-related	21 (0.2)	—	—	21 (0.2)	—
Other	38 (0.5)	—	30 (0.4)	8 (0.09)	—
Other impairments and activities	519 (6)	410 (5)	90 (1)	—	19 (0.2)
Fitness	148 (2)	148 (2)	—	—	—
Cognition	119 (1)	113 (1)	6 (0.1)	—	—
Trunk strength or control	86 (1)	60 (0.7)	26 (0.3)	—	—
Psychological	37 (0.5)	13 (0.2)	24 (0.3)	—	—
Transfers	33 (0.4)	—	33 (0.4)	—	—
Bone density	30 (0.4)	30 (0.4)	—	—	—
Pain	15 (0.2)	15 (0.2)	—	—	—
Visual scanning	15 (0.2)	15 (0.2)	—	—	—
Other	36 (0.4)	16 (0.2)	1 (0.01)	—	19 (0.2)
Trial device or equipment	273 (3)	—	—	—	273 (3)
Clinical assessment	144 (2)	72 (1)	72 (1)	—	__[b]
Total[f]	8499 (100)	5063 (60)	2600 (31)	52 (1)	784 (9)

[a]ICF: International Classification of Functional, Disability and Health framework.

XSL•FO
RenderX

[b]Not applicable.

[c]Lower limb–specific activities: sit-to-stand training, standing tolerance, running training, stair training, high level mobility (jumping, hopping, skipping), dual tasking

[d]Upper limb–specific activities: typing, handwriting, using the computer, using the computer mouse, texting, phone use.

[e]ADL: activities of daily living.

[f]Total number of observations (n) is large as clinicians reported more than one goal for most therapy sessions. Percentages may not add up to 100 due to rounding.

What Therapy Dosage Did Patients Achieve With Advanced Technology?

Therapy dosage data were collected from progress note entries, in which clinicians reported data from devices. Therapy dosage data from each device are summarized in Multimedia Appendix 4, displaying median active therapy time for 16 devices (80%), median total time on device for 18 devices (90%) and median number of repetitions for 5 devices (25%). However, due to inconsistent device measuring and reporting of therapy dosage data, there were large amounts of missing data and results should be interpreted with caution.

Overall, median total time on devices was 30 minutes (IQR 28-45 minutes), while median device active therapy time was 20 minutes (IQR 13-28 minutes). Active time for lower limb devices was highest for the robotic treadmill BWS device with VR (median 30 minutes, IQR 23-37 minutes), followed by the robotic device with integrated FES (median 27 minutes, IQR 21-33 minutes), and the VR treadmill with optional BWS (median 24 minutes, IQR 16-32 minutes). Drawing conclusions regarding active therapy time relative to total therapy time for lower limb devices was particularly challenging due to inadequate recording of either total time or active time. An OG robotic exoskeleton was the only lower limb device which consistently recorded both active therapy time and total therapy time. This showed approximately 50% of the time spent on the device was active, where median active time achieved in a session was 17 minutes (IQR 13-22 minutes) and total time on the device was 35 minutes (IQR 28-40 minutes).

Active time for upper limb devices was highest for a robotic device with VR used to train the proximal upper limb (median 24 minutes, IQR 16-32 minutes), followed by a sensor-based VR upper limb device (median 20 minutes, IQR 17-20 minutes). There were no marked differences in median active therapy time between robotic and sensor-based upper limb devices. Generally, active therapy time relative to total time on the device was roughly similar across upper limb devices, from between 53% and 67% of time spent active.

A more accurate measure of therapy dosage is number of repetitions or steps, and meters walked [38]. However, this was only reported by four lower limb devices and 1 advanced FES device. No upper limb devices consistently reported number of repetitions. Across the 4 lower limb devices, the range of median repetitions was 356 to 1292 repetitions per session. The highest repetitions and meters walked per session were achieved through the robotic treadmill BWS device with VR (median 1292 repetitions, IQR 994-1674 repetitions; median 756 meters, IQR 573-983 meters).

Discussion

Principal Results

This longitudinal observational study investigated the implementation of advanced technologies in a clinical rehabilitation setting. A total of 20 devices were used by clinicians 4208 times, with 269 patients in the first year of the center opening. Frequently used devices included gait training BWS (VR treadmill and OG), OG robotic exoskeletons, and upper limb robotic VR devices. Device use differed considerably between rehabilitation settings, with the majority (3940/4208, 94%) from outpatient services. Clinicians used advanced technologies with a wide range of patients, aged 14 to 83 years, the majority diagnosed with neurological conditions (255/269, 95%). Devices were most often used to target impairment and activity limitation–related goals. Conclusions regarding therapy dosage were limited due to inconsistent recording across devices. These insights derived within a rehabilitation setting can directly inform clinical practice and future research in the field.

Our study demonstrates that while clinicians can successfully incorporate advanced technology into practice, this differs between inpatient, outpatient, and community settings. Notably, inpatient services only accounted for 15% of device use and offsite community services accounted for less than 1%. Reasons for this are likely multifactorial, including marked differences in the available number of inpatients (n=445) compared with outpatient and community patients (n=1502). Coupled with the need to address other goals such as equipment prescription and discharge planning, inpatient clinicians used advanced technology less frequently. Community therapists based offsite showed little use of advanced technologies despite availability of portable devices (n=8). Logistical challenges to accessing and planning device use, and differing service models (eg, focus on hospital-to-home transitions) are likely reasons. These factors collectively influenced inpatient and community-based clinicians' opportunities to use advanced technology and familiarity with devices. Organizational factors (eg, clinician training and support personnel) also likely contributed to variations between settings, which are being investigated in a separate qualitative study.

Advanced technology use also differed across patient populations and rehabilitation settings. Clinicians most commonly used advanced technologies for patients with stroke (88/269, 47%), and the greatest variety of devices (n=20) for patients with stroke and spinal cord injury. The relative use of lower limb robotic devices was greater for patients with spinal cord injury. Meanwhile, the relative use of both upper limb and lower limb robotic devices was greater in inpatient services. One possible explanation is the nature of impairments in spinal cord injury and acute or subacute neurological conditions in

XSL•FO
RenderX

inpatient settings. Clinicians often seek to promote neuroplasticity and address impairments such as muscle weakness through high repetitions, which robotic devices can facilitate [29,30]. Furthermore, the manual handling benefits and opportunities for therapy with robotic devices are potentially greater in these settings and populations, where muscle weakness can be significant [22,39,40]. Although a key argument for using technology in rehabilitation is to increase therapy dosage [6,14,15], our study has highlighted that in practice, advanced technology prescription is highly variable and measuring actual therapy dosage is challenging. Digital health developers should address this as a priority through designing devices which report and standardize dosage output. Devices should include measures regarding the number of repetitions and time spent active, and ensure this information is visible and easily accessible to clinicians. This has important implications for facilitating rigorous research to guide rehabilitation technology prescription and recommended dosage. Overall, further research is required to provide guidance on device selection and prescription across rehabilitation settings and populations. Importantly, future research should seek to establish the effectiveness of using advanced technology alongside conventional interventions to improve patient outcomes in rehabilitation.

Despite the range of devices in the study site, not all devices were used equally. While devices targeting gait training and upper limb impairments were used frequently, advanced functional electrical stimulation and immersive VR devices were used least. In addition, 1 upper limb sensor-VR rehabilitation device was not used at all by clinicians. Reasons for this are likely related to clinician preference and device functionality, both of which warrant investigation in future research. Another possible explanation is the level and reach of clinician training on these devices, which was lower relative to other devices. There was no formal training process for using the upper limb sensor-VR device, and immersive VR and advanced FES were 2 of 5 devices with no increase in number of clinicians trained between baseline and 12 months (n=13 and 11, respectively). With a busy clinical workload and many devices to learn, there are fundamental complexities to providing clinician training within clinical practice which merits further research. Interestingly, there was no evidence in our study of clinicians preferring devices with easier or quicker training processes, reflected in the fact that among the highest used devices was the robotic OG exoskeleton which had the longest training requirements (30-35 hours). A scoping review conducted by our team found training and educating clinicians were the most common strategies used to implement digital interventions in rehabilitation [27]. With evidence to show this is an important element for technology implementation, future research should focus on how training should be conducted, the impact of training on clinician uptake of technology, and determining models of training that are efficient, cost-effective, and sustainable.

Comparison With Previous Work

Other studies have investigated clinical use of advanced technologies, mostly focusing on OG robotic exoskeleton devices [39,41,42]. These studies also found competing priorities in inpatient settings, where discharge planning often took precedence. In contrast to our study, Gillespie et al [41] found patients with brain injury used the robotic exoskeleton more than patients with stroke or spinal cord injury. However, it was also reported that patients with brain injury had the most variability in therapy dosage achieved with the robotic device [41]. Swank et al [42] also highlighted differences in OG robotic exoskeleton use between patients with spinal cord injury and stroke. Our study extends these findings across a broad spectrum of advanced technologies and neurological conditions. While our study found advanced technologies were primarily used to address impairments and activity limitations, Putrino and Krakauer [43] argue that neurological rehabilitation technologies are best used to address impairments, and conventional rehabilitation is better at targeting activity and participation-level goals. Therefore, it is important that use of technology is combined with conventional rehabilitation methods, rather than using technology in isolation [6,20,22]. Finally, our study found clinicians using advanced technologies with only 11% of inpatients and 15% of outpatient and community patients. This suggests that current advanced technologies may not be suitable for everyone and is consistent with other literature in the field demonstrating 15%-20% patient eligibility for studies evaluating rehabilitation technologies [5,6]. This emphasizes the need for a better understanding of who technology is appropriate and effective for, and how to design technologies to better meet the needs of both patients and clinicians.

Limitations

This study has a few limitations. First, a limitation was the inability to capture patient outcome measure data due to a lack of standardized recording and reporting of outcome measures clinically. As such, this study focuses on clinician use and uptake of advanced technology in a real-world rehabilitation setting, rather than the effectiveness of using advanced technology to improve patient outcomes. Inconsistent measuring and recording of outcome measures is a common issue in rehabilitation and should be prioritized for clinical practice and practice-embedded research [44-47]. Second, a multisite study was not feasible given the novelty of the technology center. We sought to address generalizability by collecting data on a wide range of patients, clinical disciplines, and rehabilitation services. However, the authors of this study acknowledge that the study site is a well-resourced, urban, private-public rehabilitation setting within a high-income country. Noteworthy factors which likely contributed to substantial device uptake are the construction of a purpose-built technology therapy center within the existing site, availability of a wide variety of advanced technologies, appointment of an advanced technology lead to support multiple aspects of implementation including staff training, and patient access to long-term rehabilitation services through public funding models. Therefore, lower resourced settings would likely face challenges in applying the findings of this research. Current evidence has shown low-cost devices such as commercially available VR, activity trackers and mobile phone apps are cost-effective in rehabilitation [48]. Future research should investigate the cost-effectiveness of high-cost advanced technologies. Research in the field should also continue to place equal emphasis on affordable and low-cost rehabilitation technologies. Third, data in this study were reliant

XSL•FO
RenderX

on clinicians using note templates for documentation. As such, device use is likely underreported, particularly for disciplines such as speech pathology or recreational therapy who were not captured in our data. However, given the large number of advanced technology sessions captured, the data likely adequately reflect clinical practice for physiotherapists and occupational therapists in the setting, currently the main users of advanced technologies.

Conclusions

This is one of the largest longitudinal observational studies conducted within a real-world rehabilitation setting which successfully implemented numerous advanced robotic and virtual reality technologies for patients with a wide range of health conditions. This study provides valuable insights into the types of advanced technologies most likely to be used in rehabilitation, by which clinical disciplines and rehabilitation service types, for which kind of patients, and for what purpose. Findings are important to inform rehabilitation technology implementation, aiding clinicians in device selection, outlining the likely uptake of devices, and providing recommendations for future research in the field.

Acknowledgments

The authors of this study would like to sincerely thank all patients, clinicians, and managers at the study site who made this study possible. The authors of this study would also like acknowledge the statistical support received from the Clinical Research Centre within Sydney Local Health District. This study was not funded and was undertaken as a part of the first author's PhD, which is supported by a Research Training Program Stipend Scholarship awarded by the University of Sydney. Author CS receives salary support from the Australian National Health and Medical Research Council.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to privacy reasons. Deidentified data sets are available from the corresponding author on reasonable request and with approval from the ethics committee and study site.

Authors' Contributions

All authors (LMNP, JP, JR, CS, and LH) were involved in conceptualization, methodology, validation, and writing (review and editing) for this study. With oversight from LH and JP, LMNP conducted data curation, formal analysis, investigation, project administration, visualization, and writing (original draft). LH, JP, and CS were all involved in supervision for this study. All authors have approved the final version of this manuscript for publication and agree to be accountable for the accuracy and integrity of the work.

Conflicts of Interest

Three members of the authorship team (LMNP, JP, and JR) are employed by the study site organization. However, this research was undertaken independently of the organization as a part of the first author's PhD at the University of Sydney.

Multimedia Appendix 1
Description of devices in the advanced technology therapy center.
[DOCX File , 29 KB - jmir_v26i1e60374_app1.docx]

Multimedia Appendix 2
Advanced technology therapy session note template (lower limb device example).
[DOCX File , 21 KB - jmir_v26i1e60374_app2.docx]

Multimedia Appendix 3
Average repeated use per device, by number of sessions and by number of patients (ie. the same patient using a given device more than once).
[DOCX File , 21 KB - jmir_v26i1e60374_app3.docx]

Multimedia Appendix 4
Summary of therapy dosage data collected from progress note entries, in which clinicians documented data reported on devices.
[DOCX File , 22 KB - jmir_v26i1e60374_app4.docx]

References

1. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2024-11-03]

XSL•FO
RenderX

2. Rehabilitation 2030 initiative. World Health Organization. 2019. URL: https://www.who.int/initiatives/rehabilitation-2030 [accessed 2024-11-03]
3. The need to scale up rehabilitation. World Health Organization. 2017. URL: https://www.who.int/docs/default-source/documents/health-topics/rehabilitation/call-for-action/need-to-scale-up-rehab-july2018.pdf?sfvrsn=f627c34c_5 [accessed 2024-11-03]
4. Hassett L, Allen N, van den Berg M. Feedback-based technologies for adult physical rehabilitation. In: Hayre CM, Muller D, Scherer M, editors. Everyday Technologies in Healthcare. Boca Raton, FL: CRC Press; 2020:143-174.
5. Hassett L, van den Berg M, Lindley RI, Crotty M, McCluskey A, van der Ploeg HP, et al. Digitally enabled aged care and neurological rehabilitation to enhance outcomes with Activity and MObility UsiNg Technology (AMOUNT) in Australia: A randomised controlled trial. PLoS Med 2020 Feb;17(2):e1003029 [FREE Full text] [doi: 10.1371/journal.pmed.1003029] [Medline: 32069288]
6. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2017 Nov 20;11(11):CD008349 [FREE Full text] [doi: 10.1002/14651858.CD008349.pub4] [Medline: 29156493]
7. Leow XRG, Ng SLA, Lau Y. Overground robotic exoskeleton training for patients with stroke on walking-related outcomes: a systematic review and meta-analysis of randomized controlled trials. Arch Phys Med Rehabil 2023 Oct;104(10):1698-1710. [doi: 10.1016/j.apmr.2023.03.006] [Medline: 36972746]
8. Yang X, Shi X, Xue X, Deng Z. Efficacy of robot-assisted training on rehabilitation of upper limb function in patients with stroke: a systematic review and meta-analysis. Arch Phys Med Rehabil 2023 Sep;104(9):1498-1513. [doi: 10.1016/j.apmr.2023.02.004] [Medline: 36868494]
9. Mirelman A, Rochester L, Maidan I, Del Din S, Alcock L, Nieuwhof F, et al. Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial. Lancet 2016;388(10050):1170-1182. [doi: 10.1016/S0140-6736(16)31325-3] [Medline: 27524393]
10. Hamilton C, McCluskey A, Hassett L, Killington M, Lovarini M. Patient and therapist experiences of using affordable feedback-based technology in rehabilitation: a qualitative study nested in a randomized controlled trial. Clin Rehabil 2018;32(9):1258-1270. [doi: 10.1177/0269215518771820] [Medline: 29696990]
11. Barry G, Galna B, Rochester L. The role of exergaming in Parkinson's disease rehabilitation: a systematic review of the evidence. J Neuroeng Rehabil 2014;11:33 [FREE Full text] [doi: 10.1186/1743-0003-11-33] [Medline: 24602325]
12. Vaughan-Graham J, Brooks D, Rose L, Nejat G, Pons J, Patterson K. Exoskeleton use in post-stroke gait rehabilitation: a qualitative study of the perspectives of persons post-stroke and physiotherapists. J Neuroeng Rehabil 2020;17(1):123 [FREE Full text] [doi: 10.1186/s12984-020-00750-x] [Medline: 32912215]
13. Klaic M, Fong J, Crocher V, Davies K, Brock K, Sutton E, et al. Application of the extended technology acceptance model to explore clinician likelihood to use robotics in rehabilitation. Disabil Rehabil Assist Technol 2024;19(1):52-59. [doi: 10.1080/17483107.2022.2060356] [Medline: 35400278]
14. Allen NE, Song J, Paul SS, Smith S, O'Duffy J, Schmidt M, et al. An interactive videogame for arm and hand exercise in people with Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2017;41:66-72. [doi: 10.1016/j.parkreldis.2017.05.011] [Medline: 28528804]
15. Zhang B, Wong KP, Kang R, Fu S, Qin J, Xiao Q. Efficacy of robot-assisted and virtual reality interventions on balance, gait, and daily function in patients with stroke: a systematic review and network meta-analysis. Arch Phys Med Rehabil 2023;104(10):1711-1719. [doi: 10.1016/j.apmr.2023.04.005] [Medline: 37119954]
16. Cresswell K, Cunningham-Burley S, Sheikh A. Health care robotics: qualitative exploration of key challenges and future directions. J Med Internet Res 2018;20(7):e10410 [FREE Full text] [doi: 10.2196/10410] [Medline: 29973336]
17. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: methodological issues and recommendations for practice. Pharmacoeconomics 2022;40(4):367-378 [FREE Full text] [doi: 10.1007/s40273-022-01130-0] [Medline: 35132606]
18. Morone G, Cocchi I, Paolucci S, Iosa M. Robot-assisted therapy for arm recovery for stroke patients: state of the art and clinical implication. Expert Rev Med Devices 2020;17(3):223-233. [doi: 10.1080/17434440.2020.1733408] [Medline: 32107946]
19. Pilli K, Worne B, Simpson G. Clinician experiences with using assistive technology in brain injury rehabilitation: a survey of clinician capability, attitudes, and barriers. Brain Impair 2023;24(2):185-203. [doi: 10.1017/BrImp.2023.5] [Medline: 38167181]
20. Langan J, Subryan H, Nwogu I, Cavuoto L. Reported use of technology in stroke rehabilitation by physical and occupational therapists. Disabil Rehabil Assist Technol 2018;13(7):641-647. [doi: 10.1080/17483107.2017.1362043] [Medline: 28812386]
21. Glegg SMN, Levac DE. Barriers, facilitators and interventions to support virtual reality implementation in rehabilitation: a scoping review. PM R 2018;10(11):1237-1251.e1 [FREE Full text] [doi: 10.1016/j.pmrj.2018.07.004] [Medline: 30503231]
22. Lo K, Stephenson M, Lockwood C. Adoption of robotic stroke rehabilitation into clinical settings: a qualitative descriptive analysis. JBI Evid Implement 2020;18(4):376-390. [doi: 10.1097/XEB.0000000000000231] [Medline: 33570321]

XSL•FO
RenderX

23. Mitchell J, Shirota C, Clanchy K. Factors that influence the adoption of rehabilitation technologies: a multi-disciplinary qualitative exploration. J Neuroeng Rehabil 2023;20(1):80 [FREE Full text] [doi: 10.1186/s12984-023-01194-9] [Medline: 37340496]
24. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
25. Gallagher R, Zhang L. Evaluating mobile health technologies: does the traditional randomized controlled trial serve our needs? Eur J Cardiovasc Nurs 2021;20(6):623-626. [doi: 10.1093/eurjcn/zvab053] [Medline: 34269387]
26. Charette C, Déry J, Blanchette AK, Faure C, Routhier F, Bouyer LJ, et al. A systematic review of the determinants of implementation of a locomotor training program using a powered exoskeleton for individuals with a spinal cord injury. Clin Rehabil 2023;37(8):1119-1138 [FREE Full text] [doi: 10.1177/02692155231164092] [Medline: 37036438]
27. Pearce L, Costa N, Sherrington C, Hassett L. Implementation of digital health interventions in rehabilitation: a scoping review. Clin Rehabil 2023;37(11):1533-1551. [doi: 10.1177/02692155231172299] [Medline: 37132030]
28. Jarvis K, Thetford C, Turck E, Ogley K, Stockley RC. Understanding the barriers and facilitators of digital health technology (DHT) implementation in neurological rehabilitation: an integrative systematic review. Health Serv Insights 2024;17:11786329241229917 [FREE Full text] [doi: 10.1177/11786329241229917] [Medline: 38690403]
29. Gassert R, Dietz V. Rehabilitation robots for the treatment of sensorimotor deficits: a neurophysiological perspective. J Neuroeng Rehabil 2018;15(1):46 [FREE Full text] [doi: 10.1186/s12984-018-0383-x] [Medline: 29866106]
30. Molteni F, Gasperini G, Cannaviello G, Guanziroli E. Exoskeleton and end-effector robots for upper and lower limbs rehabilitation: narrative review. PM R 2018;10(9 Suppl 2):S174-S188. [doi: 10.1016/j.pmrj.2018.06.005] [Medline: 30269804]
31. Lohse KR, Hilderman CGE, Cheung KL, Tatla S, Van der Loos HFM. Virtual reality therapy for adults post-stroke: a systematic review and meta-analysis exploring virtual environments and commercial games in therapy. PLoS One 2014;9(3):e93318 [FREE Full text] [doi: 10.1371/journal.pone.0093318] [Medline: 24681826]
32. Miller LE, Zimmermann AK, Herbert WG. Clinical effectiveness and safety of powered exoskeleton-assisted walking in patients with spinal cord injury: systematic review with meta-analysis. Med Devices (Auckl) 2016;9:455-466 [FREE Full text] [doi: 10.2147/MDER.S103102] [Medline: 27042146]
33. Turchetti G, Vitiello N, Trieste L, Romiti S, Geisler E, Micera S. Why effectiveness of robot-mediated neurorehabilitation does not necessarily influence its adoption. IEEE Rev Biomed Eng 2014;7:143-153. [doi: 10.1109/RBME.2014.2300234] [Medline: 24803207]
34. Horn SD, DeJong G, Deutscher D. Practice-based evidence research in rehabilitation: an alternative to randomized controlled trials and traditional observational studies. Arch Phys Med Rehabil 2012;93(8 Suppl):S127-S137. [doi: 10.1016/j.apmr.2011.10.031] [Medline: 22840879]
35. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, RECORD Working Committee. The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015;12(10):e1001885. [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
36. Aprile I, Cruciani A, Germanotta M, Gower V, Pecchioli C, Cattaneo D, et al. Upper limb robotics in rehabilitation: an approach to select the devices, based on rehabilitation aims, and their evaluation in a feasibility study. Appl Sci 2019;9(18):3920. [doi: 10.3390/app9183920]
37. International classification of functioning, disability and health: ICF. World Health Organization. URL: https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health [accessed 2024-11-03]
38. Scrivener K, Sherrington C, Schurr K. Exercise dose and mobility outcome in a comprehensive stroke unit: description and prediction from a prospective cohort study. J Rehabil Med 2012 Oct;44(10):824-829 [FREE Full text] [doi: 10.2340/16501977-1028] [Medline: 22878434]
39. Heinemann AW, Jayaraman A, Mummidisetty CK, Spraggins J, Pinto D, Charlifue S, et al. Experience of robotic exoskeleton use at four spinal cord injury model systems centers. J Neurol Phys Ther 2018 Oct;42(4):256-267. [doi: 10.1097/NPT.0000000000000235] [Medline: 30199518]
40. Postol N, Barton J, Wakely L, Bivard A, Spratt NJ, Marquez J. "Are we there yet?" expectations and experiences with lower limb robotic exoskeletons: a qualitative evaluation of the therapist perspective. Disabil Rehabil 2024;46(5):1023-1030. [doi: 10.1080/09638288.2023.2183992] [Medline: 36861846]
41. Gillespie J, Arnold D, Trammell M, Bennett M, Ochoa C, Driver S, et al. Utilization of overground exoskeleton gait training during inpatient rehabilitation: a descriptive analysis. J Neuroeng Rehabil 2023;20(1):102 [FREE Full text] [doi: 10.1186/s12984-023-01220-w] [Medline: 37542322]
42. Swank C, Trammell M, Bennett M, Ochoa C, Callender L, Sikka S, et al. The utilization of an overground robotic exoskeleton for gait training during inpatient rehabilitation-single-center retrospective findings. Int J Rehabil Res 2020;43(3):206-213. [doi: 10.1097/MRR.0000000000000409] [Medline: 32282573]
43. Putrino D, Krakauer JW. Neurotechnology's prospects for bringing about meaningful reductions in neurological impairment. Neurorehabil Neural Repair 2023;37(6):356-366. [doi: 10.1177/15459683221137341] [Medline: 36384334]

XSL•FO
RenderX

44. Pohl J, Held JPO, Verheyden G, Alt Murphy M, Engelter S, Flöel A, et al. Consensus-based core set of outcome measures for clinical motor rehabilitation after stroke-a delphi study. Front Neurol 2020;11:875 [FREE Full text] [doi: 10.3389/fneur.2020.00875] [Medline: 33013624]
45. Tyson S, Watson A, Moss S, Troop H, Dean-Lofthouse G, Jorritsma S, Greater Manchester Outcome Measures Project. Development of a framework for the evidence-based choice of outcome measures in neurological physiotherapy. Disabil Rehabil 2008;30(2):142-149. [doi: 10.1080/09638280701216847] [Medline: 17852285]
46. Skinner A, Turner-Stokes L. The use of standardized outcome measures in rehabilitation centres in the UK. Clin Rehabil 2006;20(7):609-615. [doi: 10.1191/0269215506cr981oa] [Medline: 16894804]
47. Colquhoun HL, Lamontagne M, Duncan EA, Fiander M, Champagne C, Grimshaw JM. A systematic review of interventions to increase the use of standardized outcome measures by rehabilitation professionals. Clin Rehabil 2017;31(3):299-309. [doi: 10.1177/0269215516644309] [Medline: 27090265]
48. Pinheiro MB, Hassett L, Sherrington C, Hayes A, van den Berg M, Lindley RI, et al. Economic evaluation of digitally enabled aged and neurological rehabilitation care in the activity and mobility using technology (AMOUNT) trial. Clin Rehabil 2023;37(5):651-666. [doi: 10.1177/02692155221138920] [Medline: 36408722]

Abbreviations

BWS: body-weight support
BWS-T: body-weight support treadmill
FES: functional electrical stimulation
ICF: International Classification of Functional, Disability and Health
OG: overground
RECORD: Reporting of Studies Conducted using Observational Routinely-collected health Data
VR: virtual reality

Edited by A Coristine; submitted 09.05.24; peer-reviewed by V Crocher, M Bourgain; comments to author 20.06.24; revised version received 04.07.24; accepted 21.10.24; published 30.12.24.

Please cite as:
Pearce LMN, Pryor J, Redhead J, Sherrington C, Hassett L
Advanced Technology in a Real-World Rehabilitation Setting: Longitudinal Observational Study on Clinician Adoption and Implementation
J Med Internet Res 2024;26:e60374
URL: https://www.jmir.org/2024/1/e60374
doi:10.2196/60374
PMID:

©Louise Michelle Nettleton Pearce, Julie Pryor, Jason Redhead, Catherine Sherrington, Leanne Hassett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing A Blockchain-Empowered Telehealth Artifact for Decentralized Identity Management and Trustworthy Communication: Interdisciplinary Approach

Xueping Liang[1]; Nabid Alam[2]; Tahmina Sultana[2]; Eranga Bandara[3]; Sachin Shetty[3]

[1]Florida International University, Miami, FL, United States
[2]Troy University, Troy, AL, United States
[3]Old Dominion University, Suffolk, VA, United States

Corresponding Author:
Xueping Liang
Florida International University
11200 S.W. 8th Street
Miami, FL, 33199
United States
Phone: 1 305 348 2830
Email: xuliang@fiu.edu

Abstract

Background: Telehealth played a critical role during the COVID-19 pandemic and continues to function as an essential component of health care. Existing platforms cannot ensure privacy and prevent cyberattacks.

Objective: The main objectives of this study are to understand existing cybersecurity issues in identity management and trustworthy communication processes in telehealth platforms and to design a software architecture integrated with blockchain to improve security and trustworthiness with acceptable performance.

Methods: We improved personal information security in existing telehealth platforms by adopting an innovative interdisciplinary approach combining design science, social science, and computer science in the health care domain, with prototype implementation. We used the design science research methodology to implement our overall design. We innovated over existing telehealth platforms with blockchain integration that improves health care delivery services in terms of security, privacy, and efficiency. We adopted a user-centric design approach and started with user requirement collection, followed by system functionality development. Overall system implementation facilitates user requirements, thus promoting user behavior for the adoption of the telehealth platform with decentralized identity management and an access control mechanism.

Results: Our investigation identified key challenges to identity management and trustworthy communication processes in telehealth platforms used in the current health care domain. By adopting distributed ledger technology, we proposed a decentralized telehealth platform to support identity management and a trustworthy communication process. Our design and prototype implementation using a smart contract–driven telehealth platform to provide decentralized identity management and trustworthy communication with token-based access control addressed several security challenges. This was substantiated by testing with 10,000 simulated transactions across 5 peers in the Rahasak blockchain network. The proposed design provides resistance to common attacks while maintaining a linear time overhead, demonstrating improved security and efficiency in telehealth services. We evaluated the performance in terms of transaction throughput, smart contract execution time, and block generation time. To create a block with 10,000 transactions, it takes 8 seconds on average, which is an acceptable overhead for blockchain-based applications.

Conclusions: We identified technical limitations in current telehealth platforms. We presented several design innovations using blockchain to prototype a system. We also presented the implementation details of a unique distributed architecture for a trustworthy communication system. We illustrated how this design can overcome privacy, security, and scalability limitations. Moreover, we illustrated how improving these factors sets the stage for improving and standardizing the application and for the wide adoption of blockchain-enabled telehealth platforms.

(J Med Internet Res 2024;26:e46556) doi:10.2196/46556

XSL•FO
RenderX

KEYWORDS

telehealth; blockchain; security; software; proof of concept; implementation; privacy

Introduction

Since the COVID-19 pandemic, the world has seen a dramatic shift toward social distancing efforts with the help of technological tools. Technology-enabled social distancing efforts have led to the wide adoption of work-from-home solutions, virtual meeting places including classrooms, financial service delivery, and health service delivery. A recent guidance regarding operation management from the US Centers for Disease Control and Prevention (CDC) indicates that telehealth services should be optimized when available and appropriate. Government health authorities from other developed and developing countries have issued similar directives. Moreover, health service organizations, such as hospital systems and insurance providers, have started to promote telehealth services to minimize physical traffic in medical facilities. With these recommendations and advancements, health care service use via telehealth services has seen a radical upshift, and it is expected that this trend will be sustained in the postpandemic era. The large-scale adoption of telehealth services within a short time comes with several social and technical concerns. Telehealth services overall are understudied, with complex, highly person-oriented, and often knowledge-intensive services, which pose difficulty in generating successful and sustainable business models [1]. Moreover, telehealth services pose privacy and identity protection threats to users that hinder telehealth's successful adoption.

Personal privacy is a significant research stream [2,3]. Earlier studies have shown that user's privacy concerns can be handled by reducing computer anxiety and increasing the incentive to disclose [4-6]. Later studies have shown more focus on protecting digital identity and trust to reduce privacy concerns, which can be managed well by increasing the trust of online services and increasing the protection of digital identity through anonymity. Sutanto et al [7] proposed an IT solution that delivers a personalized service but avoids the third-party use of personal information and thus increases the process and content gratification. As a decentralized solution to digital identity management and identity verification operations, blockchain technology, a consensus-based public or private ledger system, has emerged [8]. By considering the benefits of blockchain, such as user's control over information and decentralized security features, various organizations (ie, the Government of Canada, the United Nations, etc) have started looking into the use case to enhance user privacy by decentralizing user information control [9-11].

Considering that blockchain integration into the health care industry could significantly transform traditional services with its security features and capabilities, previous research [12] has investigated blockchain adoption in the tracking of medical supplies and infected patients. The data can be handled instantly without any intermediary processing, thus improving efficiency, and data manipulation is impossible owing to the transparent nature of blockchain networks, which can further prevent misinformation [13] and preserve privacy. With the adoption of blockchain, data consistency and immutability can be assured along the data-sharing process, thus providing the capability of self-validation for each data record collected and uploaded. Most importantly, patients can limit personal information exposure under their own control. Fragmented records are easy to merge and integrate without any third-party involvement. Patients can own their health records, such as telehealth visits and immunization records, for future reference or verification.

State-of-the-art telehealth has emerged as a solution to overcome the challenges of an increasing demand for health care, an increasing demand for care outside of the hospital, a demand for efficiency and individualization, and a shortage of health care personnel to attend to an increasing number of patients [14]. Studies have shown that there is a positive relationship between telehealth and clinical care outcomes [15]. Although current research mostly focuses on the vital sign parameter monitoring solutions of telehealth [14], extant literature provides insights into which technology of telehealth is more efficacious. According to Dellifraine Jami and Dansky [15], video technology is more efficient among telehealth technologies because of its real-time interactivity. Moreover, studies have shown a relationship between different disease categories and telehealth technology. In a meta-analysis, Dellifraine Jami and Dansky [15] revealed that telehealth plays a positive role in improving the condition of psychiatric patients and plays a moderately positive role in heart failure outcomes. Additionally, no relationship was found between telehealth and diabetes outcomes. However, there are still challenges in combining incompatible information systems used by different stakeholders.

After 2010, telehealth solutions started to grow rapidly [16]. Telehealth is now addressing issues of chronic and episodic conditions along with acute conditions. Moreover, studies have shown that telehealth is an effective solution for mental health patients [17]. It has been found that telehealth solutions help in reducing health care costs, and the migration of telehealth has been seen from hospitals and satellite clinics to homes and mobile devices [16]. However, there are limitations of the current telehealth solutions that need to be addressed by future solutions. These limitations include limited insurance coverage of telehealth, lower quality of patient-physician relationships, different legal issues for different states, and a lack of trust from patients and unwillingness from clinicians [16]. Moreover, telehealth can be an effective solution during emergencies. Through telehealth, it is possible to respond quickly to any disaster. However, it is a matter of regret that still telehealth solutions are underutilized in emergency situations [18]. In the recent COVID-19 pandemic, we saw an emerging demand for telehealth in emergencies. There were restrictions on face-to-face interactions because of the highly infectious COVID-19. Therefore, telemedicine can provide medical advice and solutions in this regard. However, there is still training needed from the clinician's side [19]. Moreover, a lack of trust remains a challenge. To overcome this challenge, the literature suggests that telehealth should be implemented proactively rather than

XSL•FO
RenderX

reactively [19]. However, studies have also reported that there are risks of fake providers of telehealth in an emergency [18].

In information systems, discipline, collection, unauthorized secondary use, improper access, and errors are considered as the dimensions of information privacy [20]. Malhotra et al [21] have indicated that control, awareness, and collection of personal information can reflect the privacy concerns of an individual. It is identified that privacy concerns are dependent on computer anxiety [6]. This computer anxiety construct indicates that users are experiencing discomfort about the digitization of their personal life and society. Even though privacy protection is associated with positive practices, users are also motivated to share information about themselves if there are financial gains and conveniences offered by online service providers [4]. An experimental study found that (1) a clear privacy statement can reduce privacy concerns, but a security seal cannot; (2) monetary incentive influences information disclosure; and (3) information requests negatively influence information disclosure [5]. It is also found that perceived anonymity of self can reduce privacy concerns [22].

An IT solution [7] has been proposed that delivers a personalized service but avoids the third-party usage of personal information and thus increases the process and content gratification. Another study has shown that to balance the beneficial use of data and individual privacy in the era of big data, policymakers must address some of the most fundamental concepts of privacy, such as reducing the collection of personally identifiable information, increasing individual control, and implementing the principles of data minimization and purpose limitation policies [23]. An enhanced Antecedent-Privacy Concern-Outcome model [24] has shown that privacy experiences, awareness, personality, demographic differences, culture, and climate are the antecedents of privacy concern. Two recent studies have reflected on the data sharing and privacy aspects surrounding this [25,26]. According to Son and Kim [27], a major consequence of privacy concerns is the user's privacy-protective behavior. Privacy-protective behavior can be defined as information provision (refusal and misinterpretation), private action (removal and negative word-of-mouth), and public action (direct complaint and indirect complaint).

On the other hand, the term "digital identity" has evolved through the process of digitization and the internet. Users, by their internet browsing history and footprint, educate others about who they are, what they do, and especially what they think. With the construction process, users' digital footprints create their "digital identity." In other literature, the term digital personhood is also used. Digital personhood refers mainly to the projected status in the online environment [28]. Windley [29] argues that the management and control of user's digital identity, trust, and privacy are the fundamental concerns for both users and service providers in the current internet economy.

Prior telehealth platforms are exposed to certain levels of data privacy risks, lack of traceability, centralized control, and lack of data provenance. The research objective of this study is to develop a blockchain-assisted, smart contract–driven telehealth platform that is trustworthy and capable of preventing various cyberattacks. To achieve this objective, we followed the design science research (DSR) paradigm to guide the development of the proposed IT artifact [30] and adopted a user-centric design methodology. Based on that, we further analyzed system requirements and backend requirements to retrieve the design methodology. This combination of analysis helps us retrieve the design principles (DPs) and design features provided by blockchain integration. Experimental results reveal that the proposed design and implementation are baked with security considerations at an acceptable level of performance in terms of scalability and throughput characteristics. Moreover, we summarize a list of generalizable and actionable guidelines for future blockchain-assisted projects and integration models.

Methods

Capabilities of Blockchain to Support Identity Management

According to the draft on Blockchain Technology Overview by the National Institute of Standards and Technology (NIST), blockchain is an immutable digital ledger system implemented in a distributed fashion (ie, without a central repository) and usually without a central authority. There are 2 major forms of blockchain-based systems: public blockchain and private blockchain [31]. In the public blockchain, the ledger is completely decentralized among the users, and anyone can participate in the system by storing the full ledger in their own custody or by transacting data using the blockchain. The private blockchain preserves some key features of the public blockchain while compromising on the fully unregulated feature. The private blockchain is mostly tested by financial institutions, firms engaged in supply chain and logistics, health care organizations, and government entities. In both the public and private blockchain cases, users and organizations want to leverage the unique capabilities provided by blockchain technology. Previous developments and implementations of blockchains and smart contracts in telehealth were built on top of Ethereum, including teleconsultation, drug administration, and medical testing [32]. The BlockHeal framework is validated by demonstrating several use cases, including user registration and verification, patient online consultation with doctors, health care and laboratory facilities, drug counterfeiting, pharmacovigilance, language translations, and emergency facilities [33]. There are issues with scalability and security features when faced with multiserver architectures [34].

Decentralization

Decentralization in blockchain refers to the situation where the ledger can be maintained without the need for any regulator (in the public blockchain context) and without the need for any trusted third-party arbitrator (in the private blockchain context) [35,36]. Ideally, the public blockchain is peer-to-peer and censorship-resistant. As the public blockchain can be operated without the need for any centralized authority, any person, with the help of web-based or mobile-based technologies, can join the blockchain as a user. A person or organization can participate in the public blockchain as a full node, which means they can keep a copy of the entire blockchain and take part in making the entire system fully decentralized across the globe. A private blockchain's decentralization is mainly for running

XSL•FO
RenderX

intraorganizational or interorganizational processes without the need for a third-party intermediary. Business processes can be operated in the form of smart contracts in the blockchain that require the participating departments or organizations to set the terms and conditions upfront. Private blockchain ledgers can be stored within the organization or across multiple organizations to provide a higher level of trust in the transactions.

Consensus Mechanism

A consensus mechanism is basically a way to mitigate the double-spending problem and to create a single truth among all the participating nodes or organizations of the blockchain [35-37]. The consensus mechanism helps the blockchain to remain decentralized across the participating entities. The main benefit of such a mechanism is that it makes sure the current state of the blockchain is the same for all participants, each of the blocks in the blockchain is properly linked with the previous block, and all participating entities can verify the block. Popular consensus mechanisms are Proof of Work, Proof of Stake, Round Robin, etc. In the private blockchain, the consensus mechanism largely depends on the participating organizations. As the private blockchain is mainly useful to store information about operations that can impact multiple organizations, the consensus can also be based on a defined contract.

Immutability

The data stored in the blockchain are tamper-proof and immutable [35,37]. With the help of the consensus mechanism, the blockchain creates a single truth across all the nodes, and thus, it is computationally impossible and infeasible to alter the data that have already been stored in the earlier blocks. Each block contains some transactions, and based on those transactions, a cryptographic hash is created. That cryptographic hash is used to connect the block with the most recently verified block. When a piece of data on a block is altered, the hash gets changed and the block containing the altered data gets isolated from the blockchain. The tamper-proof and immutable feature of the blockchain can be found in both public and private blockchain-based solutions. This benefit of the blockchain increases the security of the data and digital assets on the ledger.

Privacy Requirements in Identity Management

The technology literature tries to provide some technical solutions that can be implemented within the organization that is collecting digital identity data from users. Digital identity management can be a pervasive infrastructure that helps an organization handle user account provisioning, authorization, and authentication. Windley [29] emphasized the inclusion of 5 building blocks (process architecture, data architecture, technical reference architecture, policies, and interoperability framework) in the identity management system to provide better protection against any breach. However, these technical solutions cannot single-handedly solve privacy breach incidents. Thus, there are some external policies created by governments and standardization bodies in place for organizations. There are 3 categories of regulations and policies an organization might consider following. The first category involves global-level policies such as the recent General Data Protection Regulation (GDPR) and OECD's Data Protection and User Control for Identity. The second category involves country-level policies such as The United States Privacy Act, The Swiss Federal Law on Personal Data Protection, and the CSA Model Code for the Protection of Personal Information. The third category involves general business-specific policies such as the Health Insurance Portability and Accountability Act.

Most of these regulations help an organization to define the boundary of the operation needing users' private data and to synchronize the technical solution with regulatory compliance requirements. Ayed [38] stated that increasing users' control over their digital identity is a key factor in decreasing the chance of privacy-related breaches such as privacy invasion and fraud. Ayed [38] further posited that an increase in control will give users ownership over their digital identity and thus will allow users to see what attributes are stored online, how those can be used further, and what steps to take for reducing the risk of becoming an unfortunate victim. To achieve this feature, Ayed [38] proposed a digital identity–related privacy requirement (DIPR), which can be used in the process of designing digital identity management systems. The DIPR includes 10 requirements as shown in Multimedia Appendix 1.

DSR Methodology

DSR involves the construction of a wide range of sociotechnical artifacts [39], and the impact of this research is witnessing a more striking influence with the appropriate and effective consumption and production of knowledge throughout the research process. This paper follows the workflow toward a research contribution that can benefit software developers and technology innovators and most importantly the researchers working in this field in terms of the embedded phenomena in blockchain-assisted telehealth. The application domain of this paper is telehealth, and we argue that the solution maturity is low, while the problem space of telehealth is relatively high. Thus, based on the DSR paradigm identified in [39], we are designing a solution to improve current technical adoption. According to the Fogg Behavior Model (FBM) [40], there are 3 factors that control behavior change, namely, motivation, ability, and triggers, which can be applied to the health care domain. Specifically, to design a telehealth system and platform and promote the adoption of such a system or platform, users should be sufficiently motivated and have the ability to perform the desired behavior expected by the system or platform, and there are triggers to prompt users to perform the behavior. All 3 factors, according to the FBM, must be present at the same instant for the behavior to occur. We have adopted this behavior design model and proposed a user-centric model for the blockchain-assisted telehealth platform. Combined with the FBM theory, we present this DSR methodology to reflect the user's motivation and ability to perform telehealth behavior and receive telehealth services triggered by both the blockchain-assisted platform and the user's medical condition, as well as the health care service provider's technical capability. Considering the differences between traditional, centralized, and blockchain-supported telehealth and telemedicine systems [41], we illustrate that the motivation for technology innovation adoption plays an important role in terms of user experience (UX) and acceptance aspects (Figure 1).

XSL•FO
RenderX

Figure 1. Behavior changes from traditional health care to telehealth and blockchain-supported telehealth.

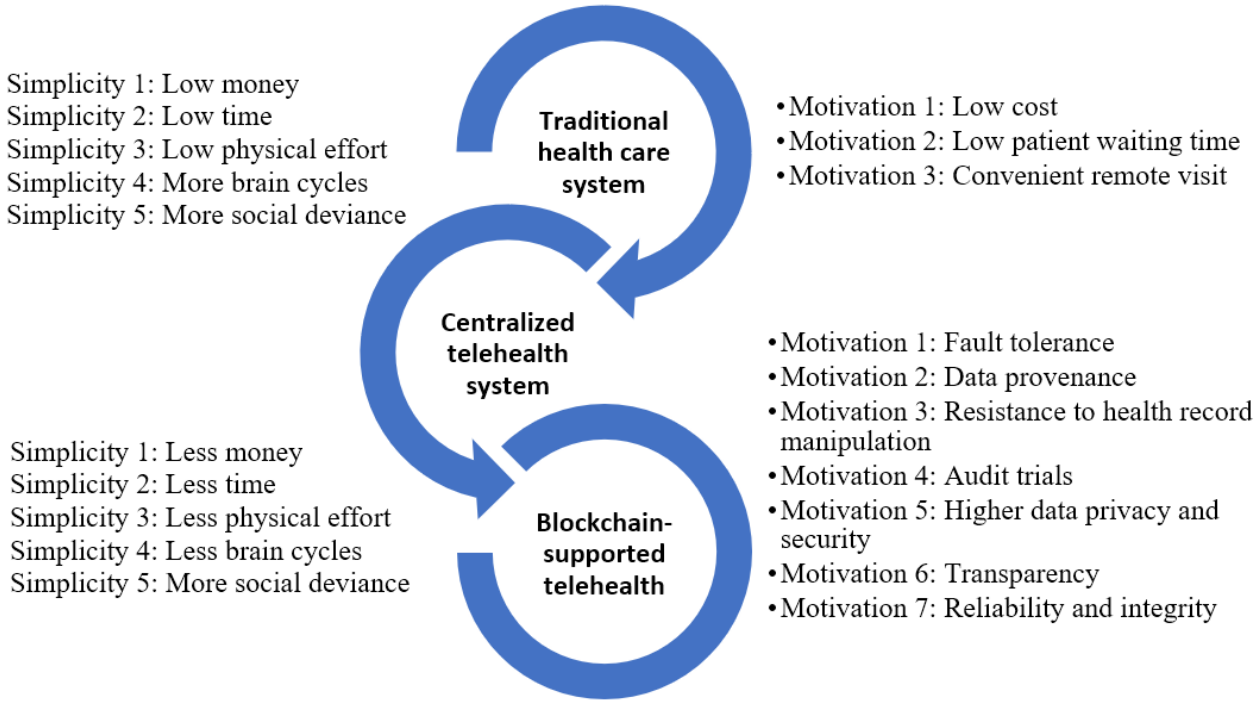

We have presented both the motivation aspect and simplicity/ability aspect of the behavior change from traditional health care to telehealth and blockchain-supported telehealth. With added motivation and added simplicity, our blockchain-supported telehealth platform design adopts a user-centric methodology and provides a straightforward process for the technology transformation of the underlying architecture implementation. The following section will illustrate how we designed each development cycle and present the main contributions from the design science perspective.

Design Cycle Description

Following the modeling design process theory [42], the overall design cycle adopted in this paper starts with problem awareness and scope, followed by suggestions from both academic and industrial practices (Figure 2). Considering the user-centric design paradigm, more than purely technical factors, this paper adds a new dimension in the second stage, which is user requirement collection. Following that, we have developed innovative IT artifacts to provide the functions and platform implementation. As such, we have presented overall DSR outcomes and contributions as follows, which distinguish this paper from previous work.

Figure 2. Our design science research approach.

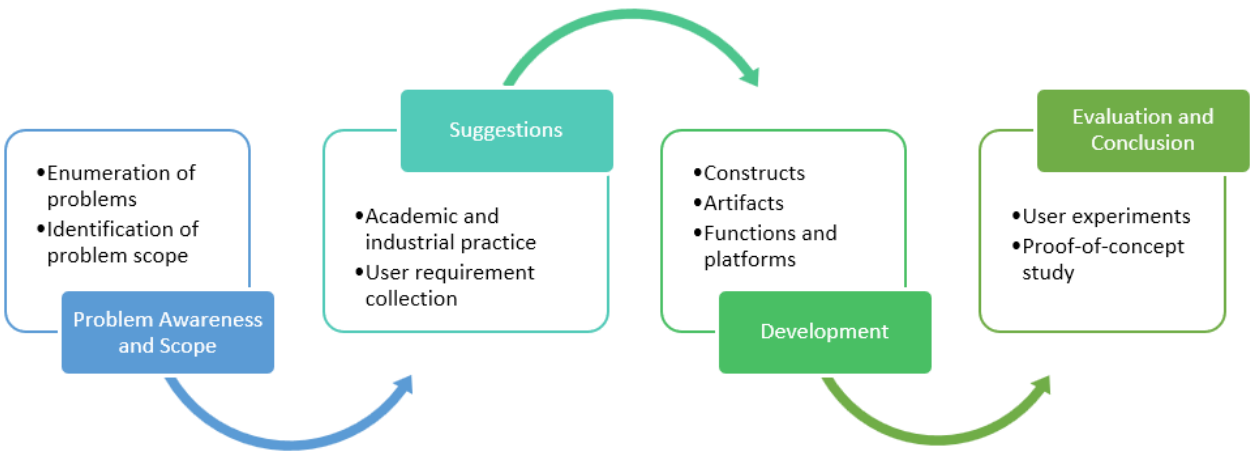

While there are applications available to support telehealth services, there are few guidelines for decentralized implementation. Based on blockchain characteristics and the needs associated with telehealth services, we have proposed a set of smart contract–driven telehealth DPs from the design requirements collected from user experiments and academic or industrial initiatives.

We have designed and presented 2 artifacts that can facilitate both user end (frontend) and backend system integration with

XSL•FO
RenderX

blockchain technology, namely, the user credential management function and the access control function, which can be driven by underlying smart contracts. The proof-of-concept project demonstrates the feasibility and verifies the usable potential of the proposed concept and design theory. We have followed the reporting guidelines suggested by [43] and have provided a checklist in Multimedia Appendix 2.

From the design of the artifacts, we have presented a reflection of the design and performed an evaluation of the proposed design in terms of security properties and functionality performances, followed by a summary of the lessons learned and contributions to DSR theories.

Organizational Setting

In our examination of the current landscape of telehealth, we analyzed both academic literature and practical applications across the industry. Our findings indicate that the telehealth sector is in the early stages of both development and adoption. A range of functionalities have been explored in prior research, including access to medical data, processing of medical services, video conferencing capabilities, epidemiological reporting, support for diagnostics and treatment, aggregation of patient data, scheduling of visits, ordering of medications, and integration of payment and fundraising systems [44]. This study focuses on the identity management and communication aspects of telehealth services.

System Details and System in Use

This study selected 2 features, which patients care the most about according to our user study related to security and privacy, including the user credential management functionality and access control functionality. The user credential management functionality has a design that incorporates user motivation as well as the ability to adapt to new technologies or platforms. Meanwhile, the access control functionality follows the principle that users are respected in terms of their privacy and data security.

Participants

There are 2 categories of participants interviewed during the design requirement collection process, namely, general users (patient side) and subject matter expert users (health provider side). The research included data from 20 general users and 5 expert users within the United States and Europe, and the demographic characteristics of these respondents are shown in Multimedia Appendix 3. On average, the interview with general users took 71.06 minutes and that with expert users took 66 minutes.

Outcome and Evaluation Criteria

From the design of the artifacts, we have presented a reflection of the design and performed an evaluation of the proposed design in terms of security properties and functionality performances, followed by a summary of the lessons learned and contributions to DSR theories. Specifically, we have assessed performance metrics, such as transaction throughput, smart contract execution time, and block generation time. Our security analysis evaluated the resilience of the system against common attacks prevalent in telehealth systems.

Methods for Data Acquisition and Measurement

We have adopted a user-centric design approach and have started with user requirement collection. We have used the information in the report by Chanson et al [45] for developing the requirements. With the analysis of the collected data, we have moved forward to design the system functionalities that can facilitate user requirements, thus promoting user behaviors toward the telehealth platform. The purpose of the interview study is to inform the design process in terms of user requirements from both general users and expert users for a privacy-preserving telehealth solution. The interview questions are related to prior experience of use or potential functionality expectations from a telehealth artifact.

Methods for Data Analysis

In the development of design requirements, we have employed qualitative methods to analyze responses from our interviewees. We have used a thematic analysis approach to interpret the interview data. Our first step involved transcribing the interviews and reading the transcripts multiple times to gain a deep understanding of the content. We performed initial coding in an inductive manner, where codes were generated from the data without trying to fit them into a pre-existing coding frame. This allowed emerging themes to be grounded directly in the interview data. After coding, we grouped related codes into potential themes. These themes were continually reviewed and refined to ensure they accurately represented the data. This part of the analysis was iterative, as themes were defined and redefined through discussions among the research team members. Each theme was then reviewed in the context of the coded extracts and the entire data set. This was to ensure that themes were coherent, consistent, and distinct. We finalized themes by defining and refining their specifics and how they are related to the user requirements in telehealth. This step involved a detailed analysis of each user requirement captured.

In the system implementation section, following the design evaluation guidelines proposed in [30], we have evaluated our artifacts using both the descriptive evaluation method and analytical evaluation method for the security domain, and the experimental evaluation method for the performance domain.

Ethical Considerations

We did not submit an application for ethics review board assessment because our research methodology has been classified as “Quality assurance activities designed to continuously improve the quality or performance of a department or program (results will not be shared with public)” according to the Office of Research Integrity (ORI) at Florida International University [46]. The interview data were collected anonymously and contained no information that could be linked to specific individuals. The data collected in this study were primarily used to design a software architecture integrated with blockchain technology, following a quality improvement approach. The raw data were analyzed to extract user requirements, which informed the design of the system architecture and the implementation of system functions. We used the raw data solely for the purpose of designing system functions. This paper presents the resulting system architecture

XSL•FO
RenderX

and functions, rather than the original data, ensuring that specific details from the interviews remain confidential and are used solely for system development purposes. The results presented in this paper are derived from our analysis and not the original data.

Results

Developing Design Requirements for Telehealth Artifacts

User Requirements (Frontend)

From the patient or user-centric standpoint, we retrieved the following 4 features from the survey as design requirements: data ownership, data sharing, self-sovereign identity (SSI) capability, and system monitoring with user notifications (Figure 3).

Figure 3. User-centric approach for requirement collection.

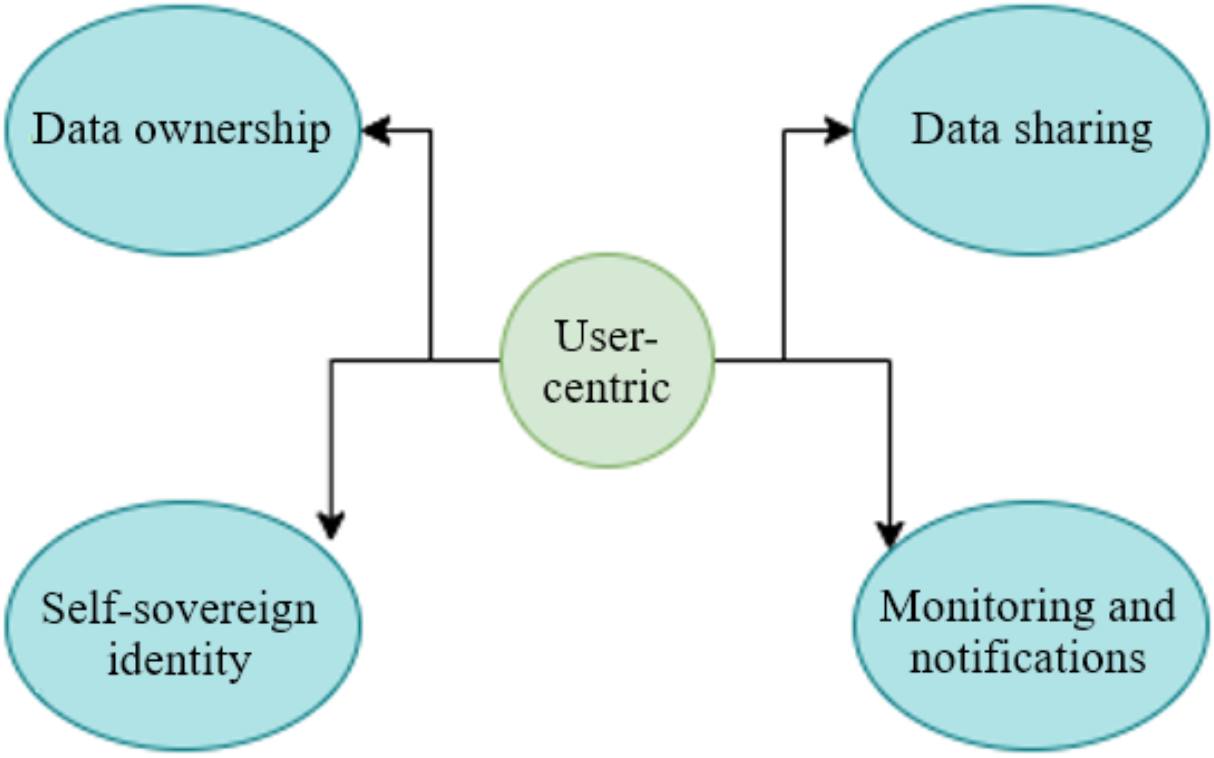

Data Ownership

Ownership means that everyone using the platform owns their data. They can limit personal information exposure at their own control. Fragmented records are easy to merge and integrate without any third-party involvement. Patients can own their health records, such as telehealth visits and immunization records, for future reference or verification. We have provided snippets of interview comments that helped us generate the data ownership requirements.

One current telehealth user mentioned:

> *Telehealth should make sure, each time before the user who logs into the system, the system itself should at least ask the person one or two personal questions, which have been asked before, so that it could verify if that is it the very same person who is logging in.*

Another telehealth user commented:

> *Perhaps telehealth systems should send me regular updates regarding the security of my personal data and should allow me to have control over how secure my data is and what security measures are put in place to protect it.*

One expert user mentioned:

> *I would want patients to be able to access their data if they wanted to, so that they can see what is recorded about them. I would want there to be a clear way for patients to review the data and communicate with the service if they felt any information was incorrect or needed amending. I would want patients to be offered clear choices about how their data is managed.*

Data Sharing

The patient can decide what information gets shared and to whom the information gets shared. The permission for data sharing does not rely on health care providers or medical institutions, which means that the reliance on institutions is minimized. Most importantly, the patient can only share the necessary information without revealing personal or sensitive

XSL•FO
RenderX

information. We have provided several interview comments on the data-sharing requirement of telehealth.

A user commented:

> *In my opinion, information should only be shared between doctors - obviously - but also only when completely necessary. For example, if someone is visiting a mental health professional, I think a big concern is that the information shared will be shared with their family doctor for example.*

Another user mentioned:

> *Data that is shared between the medical services should only be found on the patient history, only with his or her permission, so that they know that they were there ones who allowed it to happen or to take place.*

A medical professional stated:

> *There is a vast security network in place to allow the safe sharing of data between medical service providers. I would expect the telehealth system to have similar security systems in place. For instance, there should be levels of clearance given to different members of staff as to which data they are able to access and who is able to share it.*

SSI Capability

Patients are required to verify their identity each time they request a telehealth visit. The identification and verification processes are launched on top of the blockchain-supported credential authentication platform. The verified credentials will be hashed and recorded in the blockchain ledger and the patient's local storage for immutability and protection purposes. We have provided interview comments that helped us generate the SSI requirements.

One current telehealth user mentioned:

> *Telehealth systems can verify user identity by asking questions only the specific user would know the answers to. Identity verification software can be researched and utilized.*

Another telehealth user commented:

> *I think it would be beneficial for video calling to be the main method of telehealth services, and then the patient would be able to display their health card.*

An expert user agreed:

> *The person could also be asked to input some verification details when logging on for the first time - for example, they could be asked to put in their name, date of birth and address, and these could be checked against the recorded details - if these did not match, then the person would not be able to log on.*

Monitoring and Notifications

The patient has the option to enable the function of monitoring risks according to their health conditions after the telehealth visit, with assisting wearable devices or other medical equipment. The patient can also be notified about safety measures or potential risks of certain activities, especially in the pandemic era, with their permission and initiative to share the intended activities. We have provided several interview comments on the monitoring and notification requirements of telehealth.

A user commented:

> *Telehealth systems can monitor my progress by allowing me to enter symptoms, information, personal notes, and observations, and so forth into an online diary or sorts.*

Another telehealth user mentioned:

> *I think the best option would be to follow up via the same method as the initial consultation that allows for either user to communicate whenever they are available rather than having to be connected at the same time.*

An expert user stated:

> *I would like optional questionnaires on the system, that the patient could be asked to complete during or after each session. I would like there to be a broad range of options available, so that the patient and I could pick the questionnaires most suited to them and their health conditions, and we could pick the frequency of monitoring this.*

System Functions (Backend)

For system functions in the telehealth scenario, we have considered 4 aspects, including architecture, data operation, integration, and interoperation among systems, which will be improved for the adoption of the blockchain component. Moreover, for the security of user accounts and access management, the system functions will include authentication and access control. Detailed descriptions of each function are presented below (Figure 4).

XSL•FO
RenderX

Figure 4. System functionality analysis.

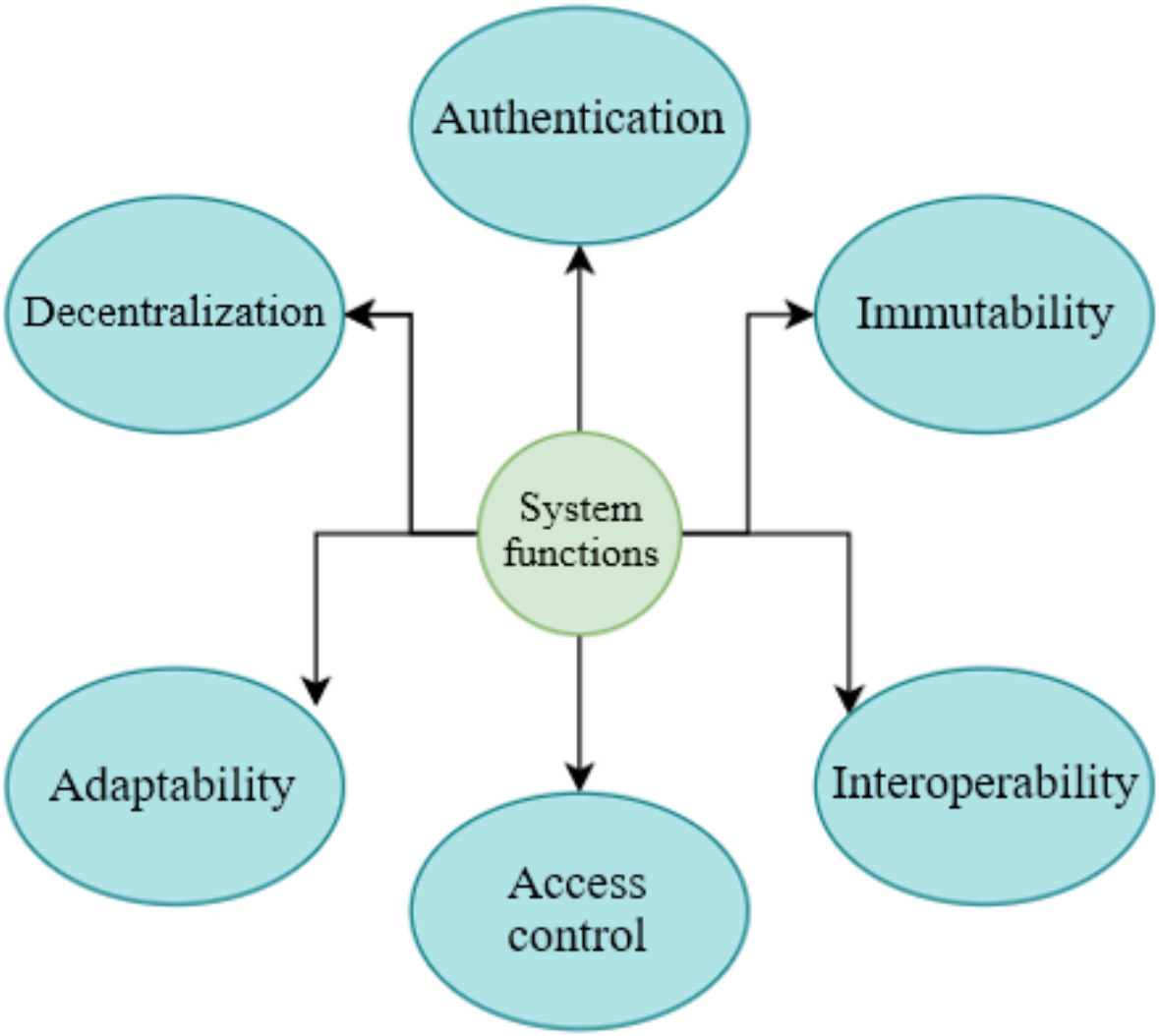

Decentralization

The system will be built on top of decentralized architecture, which removes the vulnerability of a single point of failure and removes the reliability on a trusted third party. With decentralized architecture, there is extra assurance of the resilience of the system, especially when faced with potential attacks from both insiders and outsiders. The decentralized governance of the data and system operations assures that the system still functions normally even when some portion of the system is compromised. For telehealth visits, it is essential to maintain the business uninterrupted during the communication process between patients and health care providers, and maintain after-visit accessibility. The decentralized server architecture will provide seamless communication and connection whenever the patient needs access to either service or data remotely.

Immutability

The blockchain structure provides the function of immutability and tamper resistance in terms of data protection and makes sure that any data records or access logs cannot be modified by malicious users. Both the data records and access records are hashed and stored in the blockchain and local storage. This technique can prevent both insider and outsider attacks. Further, it is essential to provide the authenticity of system logs when auditing or digital forensics is requested.

Authentication and Access Control

As we mentioned in the user requirements, the SSI mechanism has been implemented with blockchain technology, so the authentication process in the system is based on the identity and user credentials provided when requested. Delegated access can be designed for users at different levels so that only authorized users can access the data, and the data can only be accessed by users authorized with delegated access and valid tokens [47]. Once the tokens expire, access is revoked. Appropriate access control mechanisms should be designed and implemented to protect data security and defend against attacks such as email phishing and social engineering attacks.

Adaptability

This system function is required at 3 different levels. First, the telehealth system should adapt to user groups (pediatrics, adults, and seniors). The sign-up process should be easy and transparent. There should not be any technical barriers for users to adopt the system. Moreover, system access cannot be slow and costly. Users should be capable of accessing their comprehensive medical records at their fingertips [48]. Second,

XSL•FO
RenderX

the telehealth system was demanded by the pandemic situation, but it should also adapt to postpandemic use cases when the telehealth need will be further developed and improved [49]. Third, for compliance with regulations, policies, and guidelines from federal, state, and local governments, the system should be accompanied by a function to adapt well to future updates and upgrades by continuous system iterations.

Interoperability

The artifact considers and integrates the system upgrade with the function and capability of interoperability, considering the current health care system separation and data silos. The blockchain component serves as an underlayer of the telehealth system and should be designed as an open application programming interface for easy adoption and interoperability. Users should be capable of accessing their medical records from other systems, which can be easily imported. The transfer of records should be seamless, saving money and time.

Deriving Blockchain DPs and Design Features for Privacy Protection

Following the existing body of knowledge and the design requirements concluded from the previous step, we have derived the following basic DPs that can be fulfilled with design features [50] summarized in the next step.

First, considering the objectives of this research, the UX and DSR paradigms can be combined to design a user-centric IT system [51] and can offer telehealth services to potential users. By applying the DSR and UX principles and exploring how these 2 paradigms can individually and jointly strengthen the design and development processes for user-centric systems, we can identify their value to both theory and practice. Thus, we have derived the following DP (DP1): Focus on a user-centered business process before, during, and after the telehealth service is provisioned. The second philosophy during the design was a minimalist design, and it is summarized as the “less is more” principle, due to the reality that irrelevant information distracts [52]. This DP reflects the system design requirements of adaptability and interoperability and is summarized as follows (DP2): Redesign the business process with minimal change to the current working system in telehealth services.

Owing to the business logic and nature of health care service providers, the telehealth platform needs coordination among multiple health care providers and relevant stakeholders in the telehealth ecosystem to serve many patients on demand [53]. Thus, a decentralized design pattern allows a real-time patient response in terms of workload balance. A health care member in the blockchain network can join and access the telehealth platform on demand while the number of patient requests varies. This will lead to the maximization of business value and patient benefit. Thus, we summarize the DP as follows (DP3): Increase business value and reduce costs associated with telehealth integration and adaptation. Another DP is as follows (DP4): Apply automated business process design to telehealth system configuration and maintenance. Considering the user requirements of data ownership, automation can be used to implement and accomplish data sharing, identity management, and monitoring. Automation can be realized by a blockchain smart contract where a business process can be decomposed into single steps that are performed repeatedly and follow specific rules [50]. Specifically, in smart contract execution, the business process can be translated into algorithms, which a computer can execute automatically and repeatedly, in near real-time patterns. Meanwhile, automation can reduce the cognitive efforts required from human attention, thus improving the system efficiency and removing human errors. Another DP is as follows **(**DP5): Security-oriented design baked into the entire data lifecycle. Security is always a crucial component, especially when faced with sensitive patient data. The user requirements for frontend design include patient authentication and identification verification processes, which can ensure that the patient is the same person they claim to be. Meanwhile, patient data can be considered a critical asset, and the platform should strictly enforce designed access control policies so that only authorized entities can access patient data and should adopt auditing components.

Existing blockchain-supported telehealth platforms have various features that can benefit both users and health care providers, as well as other stakeholders. A summary of the features is presented in Table 1.

Based on the DPs, we have proposed the following design features in this paper to provide a blockchain-assisted telehealth solution (Figure 5). For the system functions in the telehealth scenario, we will consider 4 aspects, including architecture, data operation, integration, and interoperation among systems, which will be improved for the adoption of the blockchain component. Besides, for the security of user accounts and access management, the system functions will include authentication and access control. The system will be built on top of the decentralized architecture, which removes the vulnerability of a single point of failure and removes the reliability on a trusted third party. With decentralized architecture, there is extra assurance of the resilience of the system, especially when faced with potential attacks from both insiders and outsiders. The decentralized governance of the data and system operations assures that the system will function normally even when some part of the system is compromised.

For telehealth visits, it is essential to maintain the business uninterrupted during the communication process between patients and health care providers, and maintain after-visit accessibility. The decentralized server architecture will provide seamless communication and connection whenever the patient needs access to either service or data remotely. The SSI mechanism will be implemented with blockchain technology, and thus, the authentication process in the system will be based on the identity and user credentials provided when requested. Delegated access can be designed for users at different levels so that only authorized users can access the data, and the data can only be accessed by users authorized with delegated access and valid tokens. Once the tokens expire, access is revoked [47]. Appropriate access control mechanisms should be designed and implemented to protect data security and defend against attacks such as email phishing and social engineering attacks. Overall, the security design is baked into the entire data lifecycle, from data collection and data storage to data sharing.

XSL•FO
RenderX

Table 1. Existing blockchain solutions and design features.

Design features	Blockchain capability and components required	Benefits and improvements
Medical data access; medical records for distant consultations; patient data and prescription	• Decentralized network • Smart contract • Tamper resistance	• Patient outreach: extend the reach and expand health care services to patients • Data provenance: clinical data; patient data
Medical services processing	• Transparency • Peer-to-peer communication • Smart contract	• Convenience: conveniently monitor and manage long-term patient care • Workforce limitations: minimize patient access and workforce limitation issues
Video conferencing	• Peer-to-peer communication • Data provenance • Smart contract	• Effectiveness: present more effective use of health care provider's time slots • Throughput: assist physicians to serve more patients • Diagnosis time: prevent delay of diagnosis for certain deadly conditions or diseases • Social distance: prevent the spread of airborne infections • Revenue: provide new streams of revenue to physicians and hospitals
Epidemiology reporting and remote monitoring	• Peer-to-peer communication • Smart contract • Data provenance	• Convenience: conveniently monitor and manage long-term patient care • Throughput: assist physicians to serve more patients
Diagnostics support; emergency service and translation support	• Peer-to-peer communication • Data provenance • Smart contract	• Convenience: conveniently monitor and manage long-term patient care • Data provenance
Treatment support and doctor recommendation	• Immutability • Smart contract • Data provenance	• Data provenance • Throughput: assist physicians to support more patients
Patient data aggregation	• Immutability • Peer-to-peer communication • Data provenance • Smart contract	• Identity management • Access control: authorize participants without a central authority
Visit arrangement	• Peer-to-peer communication • Smart contract	• Convenience: conveniently arrange patient care • Throughput: assist physicians to serve more patients
Medicine ordering and drug tracking	• Immutability • Data provenance • Smart contract	• Convenience: conveniently monitor and manage long-term patient care • Data provenance
Automated payments for billing and claims	• Immutability • Data provenance • Smart contract	• Automated execution: payment settlement • Data provenance • Immutability: ledger service
Fundraising	• Data provenance • Smart contract	• Identity management: manage user identities • Anonymity: preserve patient anonymity • Immutability: ledger service

XSL•FO
RenderX

Figure 5. Proposed design features.

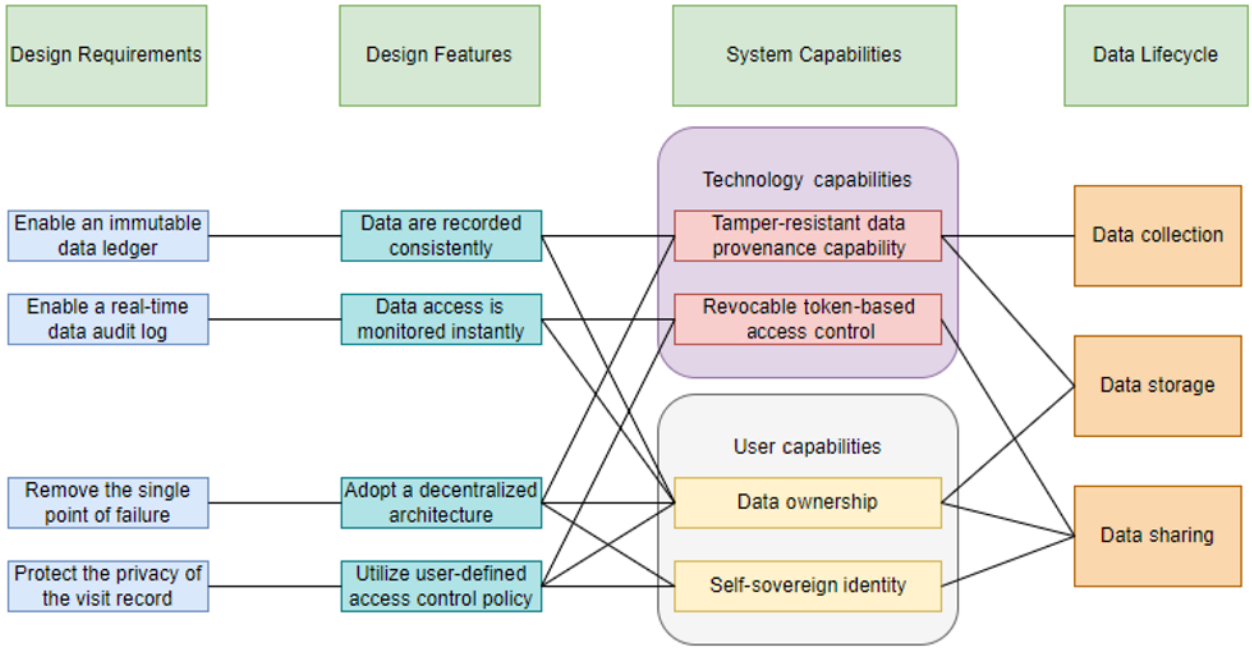

Developing and Implementing the Blockchain Platform for Privacy Protection

This paper discusses 2 features that are implemented and that users directly interact with according to our user study, namely, user account/credential management functionality and token-based access control functionality, both of which are closely related to security and privacy.

To provide smart contract–driven functionalities, we have built our platform, Dokka (Figure 6), with the Rahasak blockchain [54]. Rahasak is a permissioned blockchain platform where participating institutions can register and enroll their identity through a membership management service. The platform has been designed using a microservices architecture. Dokka is a blockchain SSI-based credentialing engine for telehealth platforms. Multiple health care companies (eg, Sentara and Cigna) collaborate with blockchain platforms (eg, IBM HUN). There will be a separate blockchain node for each health care company. Smart contracts will be running on each blockchain node. The data on the platform can be shared among multiple hospitals in the network.

Figure 6. Dokka platform's microservice-based architecture.

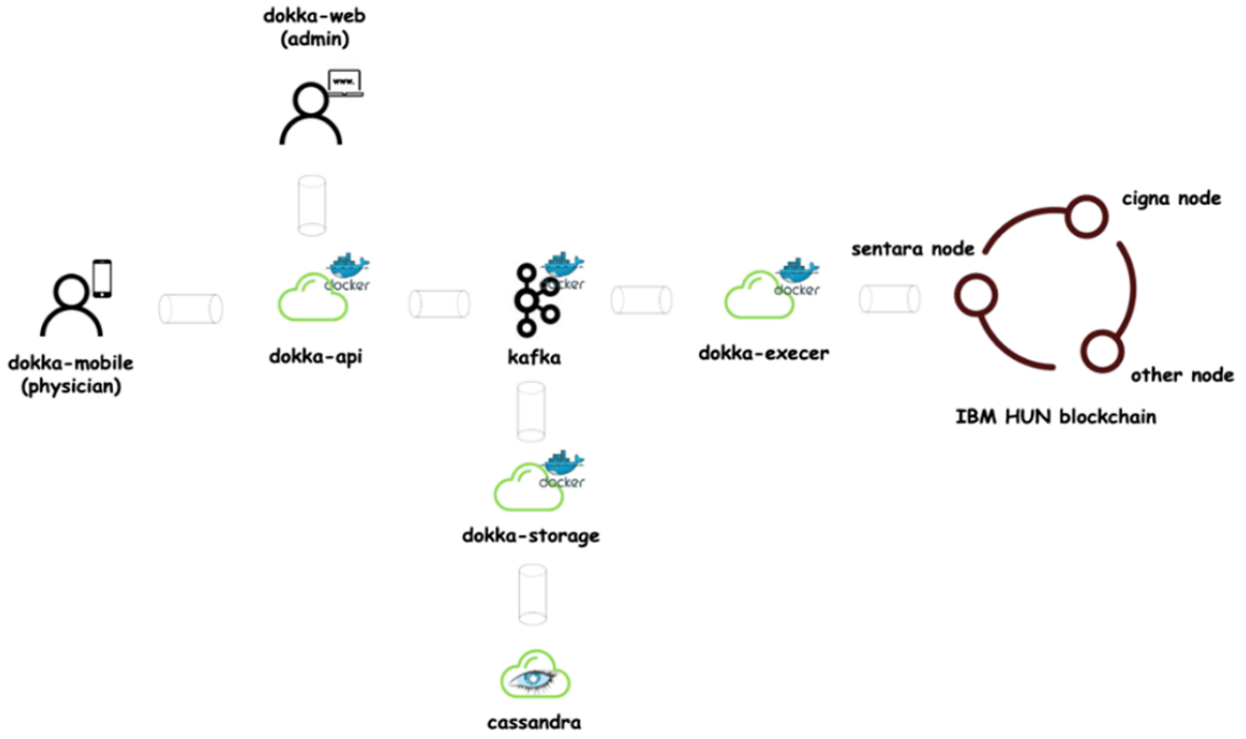

XSL•FO
RenderX

Smart Contract–Supported User Credential Management

Physicians can use this application to manage their physician credentials (Figure 7). Patients can use the application to manage their telehealth functions such as appointment booking and prescription handling (Figure 8). An SSI-based identity wallet has been implemented to capture or verify user identity proofs. It stores user data on the blockchain platform by using SSI proofs. All services are dockerized and available to deploy with Kubernetes [55]. The Dokka platform is built on the microservice architecture [56] to support high scalability and high transaction load following the single responsibility principle. To cope with high transaction loads and back-pressure [57] operations in practical systems, we have followed a reactive stream–based approach in the implementation process. All microservice communications are handled via the Apache Kafka message broker. The functionalities of the blockchain have been implemented with Aplos smart contracts. There are 3 main smart contracts: account contracts, device contracts, and credentialing contracts. The Dokka mobile wallet is the client application on the Dokka platform. The functions implemented in the blockchain smart contracts will be invoked by mobile clients. The requests generated from mobile apps will be directed to blockchain smart contracts via the Dokka gateway service. There is a peer-to-peer communication channel between Dokka mobile wallets to exchange credential data.

Figure 7. Physician interface for credential management.

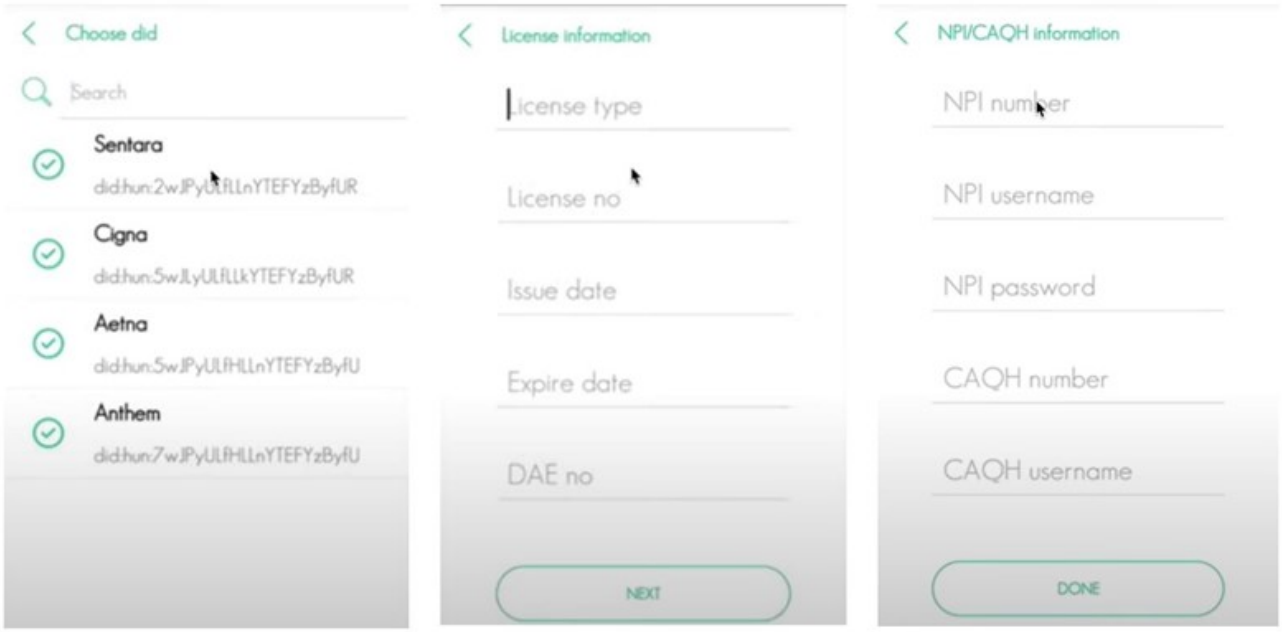

Figure 8. Patient interface for credential management.

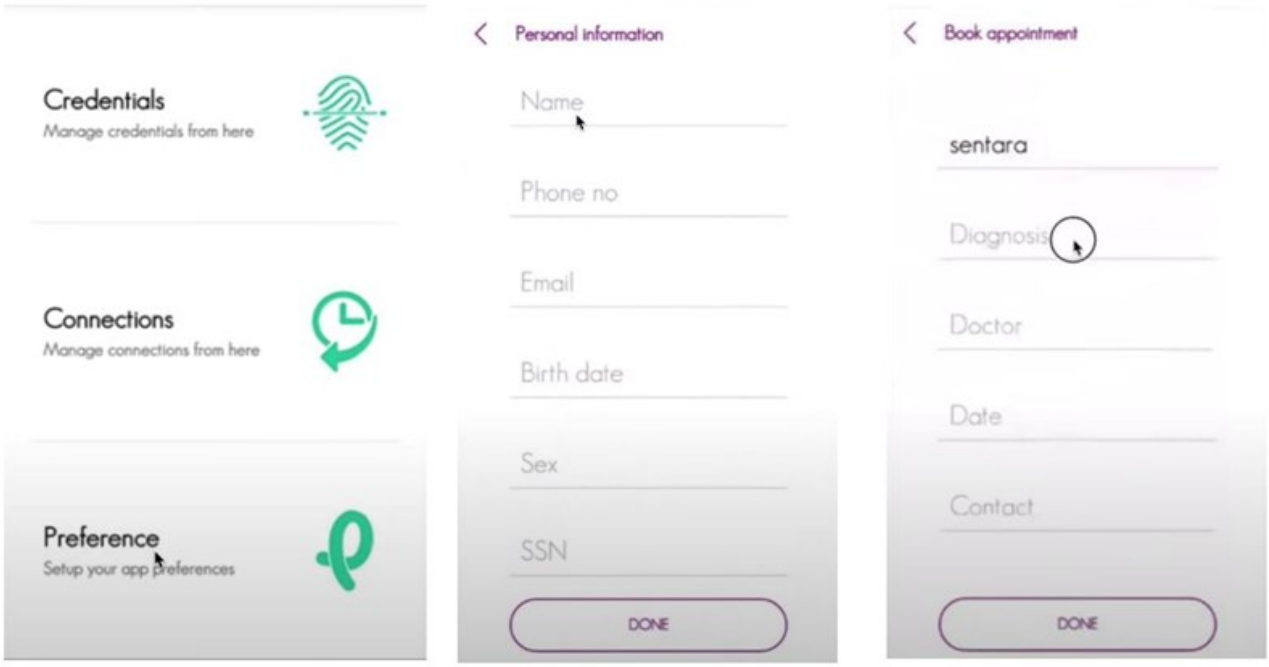

XSL•FO
RenderX

Token-Based Access Control

A missing design component in existing studies is the access control mechanism in telehealth platforms. This project adopts token-based access control to secure personal health data access and exposure. The access control policies are enforced by blockchain smart contracts in a decentralized way to ensure integrity and remove the necessity of a trusted third party. Various stakeholders can request data access from the data owner, that is, the registered user in the system. Users have the option to grant, deny, and revoke access from other participants. User registration is based on U-Prove technology [58], which includes 3 entities, namely, issuer, prover, and verifier. In our implementation, the issuer and verifier represent the same entity. Parameter definitions for both the prover and issuer in the token generation phase are listed in Table 2.

With the above information provided, we have used issuance protocol version number 0x01 and the user platform identification key to generate a private key in the token generation protocol (Figure 9). For each telehealth request, the user will generate an access token for the requester. The algorithm supports the generation of multiple keys for a single user for better performance. The Dokka platform nodes issue tokens to users with a digital signature. For privacy concerns, the application attributes are hashed for the generation of U-Prove–based tokens. During certain circumstances, the issuer can generate multiple tokens at one time for better performance.

Table 2. Parameter definitions for both the prover and issuer in the token generation phase.

Parameter	Parameter definition
*(TI): TI ∈ (0, 1)**	The value of the token information field
*(PI): PI ∈ (0, 1) **	The value of the prover information field
(AA): (A1, ..., An), TI	Application attributes
UID $_H$	Identifier for the secure hash algorithm
(g0, g1, ..., gn, gt)	Issuer's public key
(e1, ..., en)	Format of each application attribute
(IP): UIDp, desc(Gp),UID $_H$ *,(g0, g1, ..., gn, gt),(e1, ..., en)*	Issuer parameters
P = H(IP)	The hash of the *(IP)*
(DP): gd, xd, hd	Device parameters

Figure 9. Access token generation algorithm.

Protocol 1. User Registration

Input:

$x_t = Hash(0x01, P, TI)$, $x_i = Hash(A_i)$, $\gamma = g_0 g_1^{x_1} ... g_n^{x_n} h_d$

UID_p, random $\alpha, \beta_1, \beta_2, \omega$, and issuer private key y_0

Compute:

$h = \gamma^{\alpha}$, $\sigma_z = \gamma^{y_0}$, $\sigma_z^1 = \gamma^{y_0}$, $\sigma_a^1 = g_0^{\beta_1} g^{\beta_2} g^{\omega}$, $\sigma_b^1 = (\sigma_z^1)^{\beta_1} h^{\beta_2} \gamma^{\omega\alpha}$

$\sigma_c^1 = Hash(h, PI, \sigma_z^1, \sigma_a^1, \sigma_b^1)$, $\sigma_r^1 = (\sigma_c^1 + \beta_1 \mod q) y_0 + \omega \mod q + \beta_2 \mod q$

Output:

U-Prove token T: $UID_P, h, TI, PI, \sigma_z^1, \sigma_c^1, \sigma_r^1, d$

prover private key: α^{-1}

Evaluation of Security and Design Performance

Security Analysis

From the analytical perspective, our security design is baked into the overall telehealth architecture. Based on the attack surface analysis of the telehealth platform, we have considered the authentication, communication, and auditing processes. According to the design blueprint for a large-scale telehealth platform [53], 5 layers are taken into account. Using smart contract–driven methodology, this paper improves the security level for each layer, including user authentication, user credential verification, access control, data protection, and auditing (layers 1-5). Users interact with layer 1 directly, and this is the initial step for the telehealth service. During this phase, users are authenticated by the blockchain network via an SSI-enabled identity verification process. Layer 2 determines the appropriate access control policy, and the tokens are issued in this step for both patients and stakeholders. Specifically, the tokens are periodically refreshed to defend against replay attacks, man-in-the-middle attacks, stolen verifier attacks, and impersonation attacks [34]. Layer 3 functions on top of the issuance of the tokens, and access is granted automatically via smart contracts. Users can perform data-sharing operations with other parties in layer 4 with appropriate cryptographic settings.

XSL•FO
RenderX

The auditing process in layer 5 monitors all the activities in this process and automatically generates alerts once a malicious behavior is detected. The auditing process is captured by smart contracts so that the log is tamper-resistant.

As pointed out in [53], the user registration and identification verification layer is the frontend layer, which is a representative artifact of user-centered design. Thus, we have analyzed the security properties provided by the user registration and identification verification process here from a descriptive perspective. The critical identity information could be accessed by attackers via technical approaches such as key compromise and identity theft. Impersonation attacks are typical cases of identity manipulation attacks, which can be launched via identity theft or credential compromise. If attackers get access to previous account tokens, they could recognize or identify a pattern from these tokens, and thus, valid tokens could be compromised eventually.

For the token-based access control component, there are potential attacks such as replay attacks. Smart contracts will provide preventive measures against token-related attacks by the enforcement of automatic program execution and peer validation of tokens. In this way, the replay attack will fail due to unauthorized access and the consensus protocols designed for peer validation. For the video conferencing component, there are potential attacks, such as man-in-the-middle attacks and impersonation attacks, but the peer-to-peer communication process is well protected by the underlying smart contracts and cryptographic algorithms.

In summary, the proposed design provides resistance against common attacks targeting general user communication processes with multiple stakeholders involved. Specifically, the design provides resistance against replay attacks, collusion attacks, man-in-the-middle attacks, stolen verifier attacks, and impersonation attacks [34].

Performance Analysis

Technically, the blockchain can support telehealth platform functionalities; however, previous studies have mentioned that the blockchain's scalability is currently limited [59]. We evaluated the performance in terms of transaction throughput, smart contract execution time, and block generation time. Besides, the scalability and throughput of blockchain transactions have always been concerns and bottlenecks for system processing capability. In the following experiments, we will investigate these different aspects of system performance and compare them with existing solutions. There are 3 Kafka broker nodes for consensus management, with 3 Zookeeper nodes representing service providers in Dokka with multiple Rahasak blockchain clusters in AWS 2xlarge instances (16 GB RAM and 8 CPUs). The simulation platform uses Apache Cassandra as the state database and is implemented with Scala functional programming and the actor-based Aplos smart contract platform. The evaluation results are obtained with a varying number of blockchain peers in different evaluations.

For transaction throughput, we recorded the number of digital identity proof create transactions and digital identity proof query transactions that can be executed in each peer in the Dokka platform. Digital identity proof query transactions and invoke transactions generate transaction records that update the status of the ledger, while query transactions only perform search operations without updating the existing transactions. We obtained consistent throughput in each peer on the Dokka platform, flooded with concurrent transaction load (Figure 10). Similarly, we recorded the time cost to execute and validate different sets of transactions (Figure 11), which indicates an acceptable and linear time overhead.

Figure 12 shows the transaction scalability aspect. As the number of nodes in the cluster increases, the transaction throughput increases linearly. Query transactions have higher scalability than invoke transactions since the query operation avoids ledger status updates. To evaluate the transaction execution rate, we captured the number of transactions executed using different numbers of blockchain peers. Figure 13 shows that when the number of peers increases, the transaction execution rate increases proportionally.

Figure 14 shows the transaction execution rate and transaction submission rate in a single peer. The results are satisfying with regard to the stable performance over time. Moreover, the average time to generate new blocks is recorded against the number of transactions in a block. Block generation time depends on the message propagation time, block hash time, and transaction verification time. When the number of transactions increases in a single block, these factors will increase. Therefore, when the number of transactions increases, block generation time also increases correspondingly. As shown in Multimedia Appendix 4, it takes 8 seconds on average to create a block with 10,000 transactions.

XSL•FO
RenderX

Figure 10. Invoke and query transaction throughput.

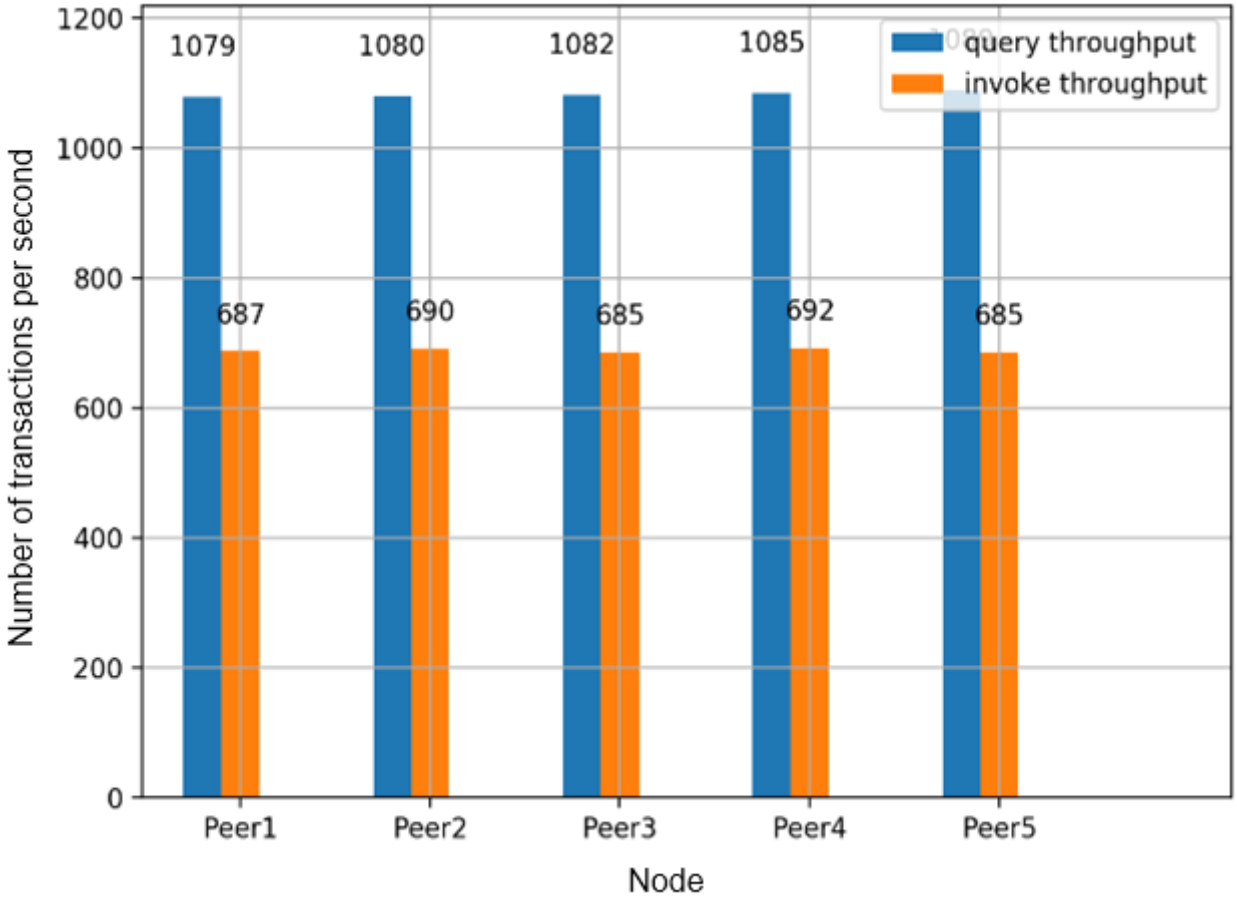

Figure 11. Transaction execution time and validation time.

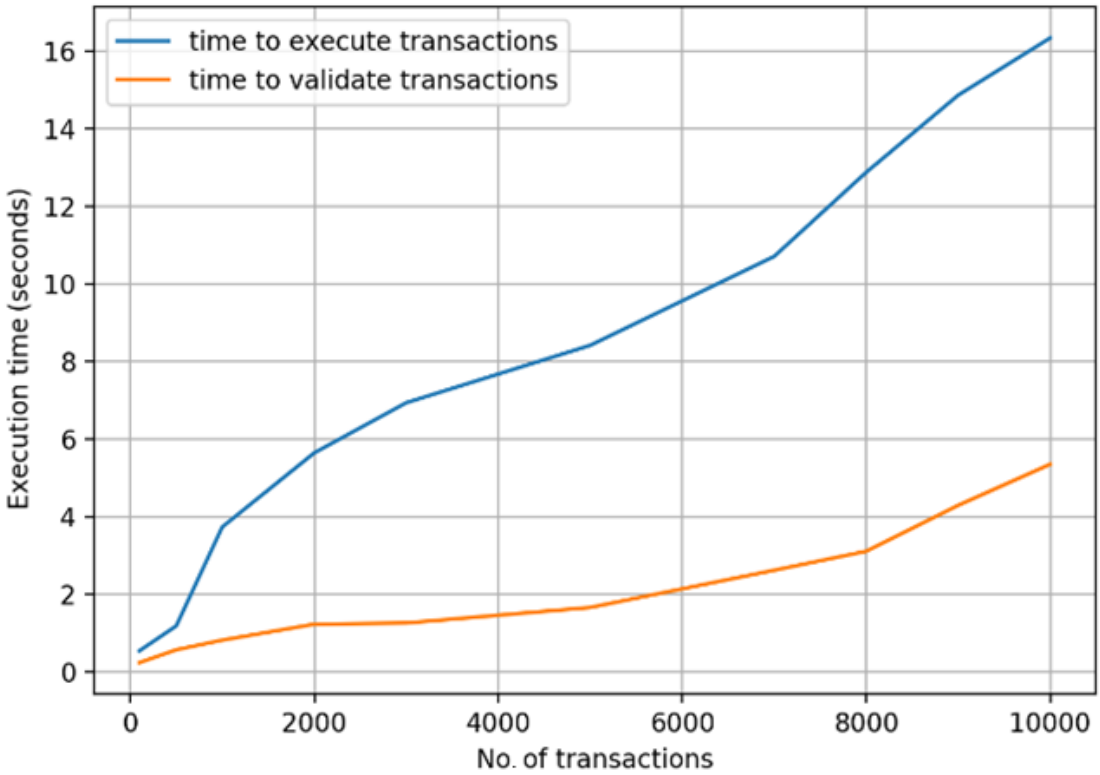

XSL•FO
RenderX

Figure 12. Transaction scalability.

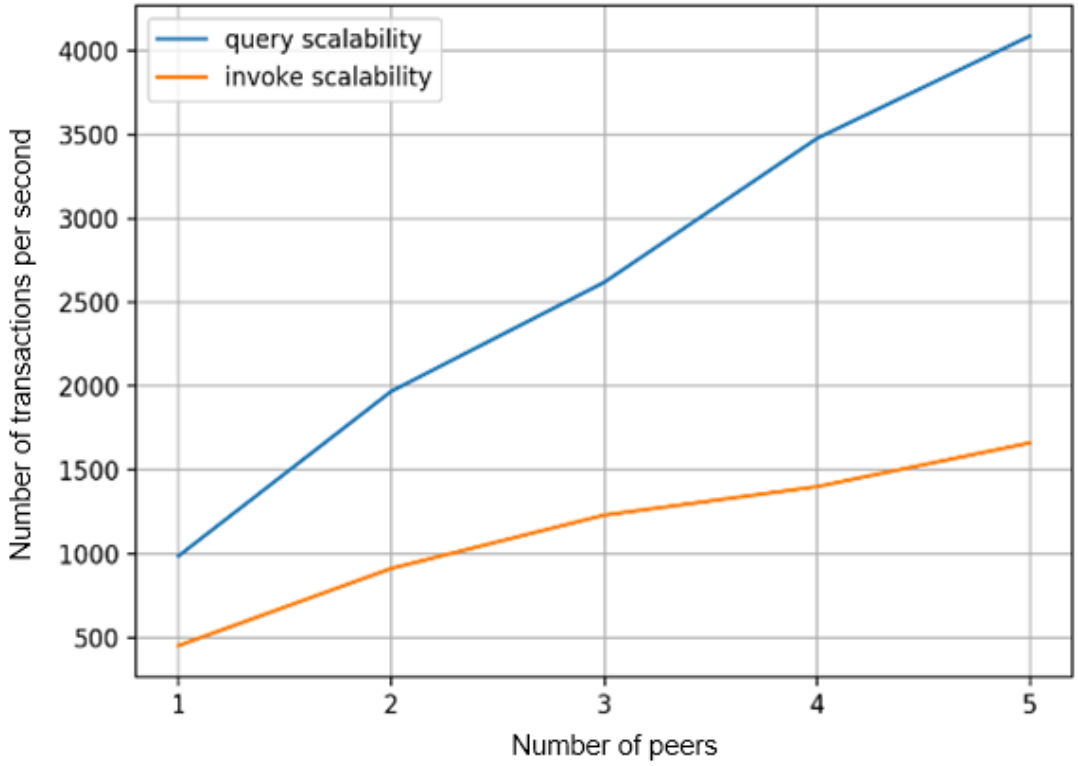

Figure 13. Transaction execution rate.

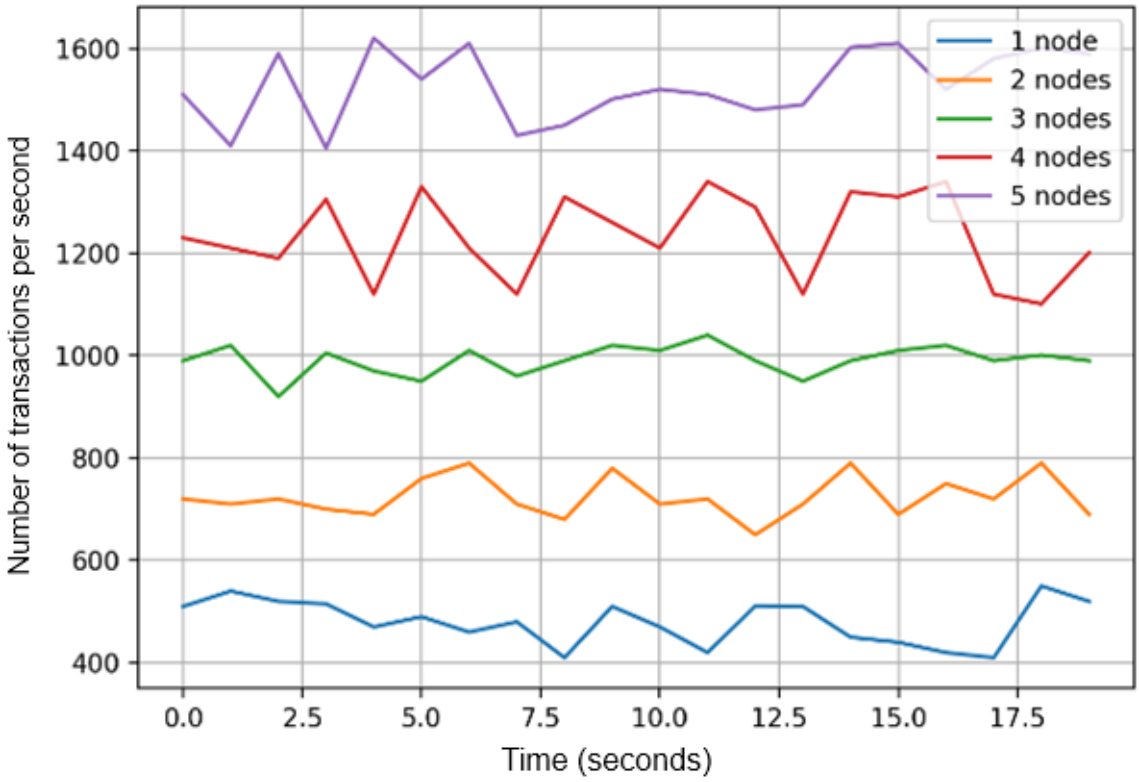

XSL•FO
RenderX

Figure 14. Transaction execution rate and submission rate in a single blockchain peer.

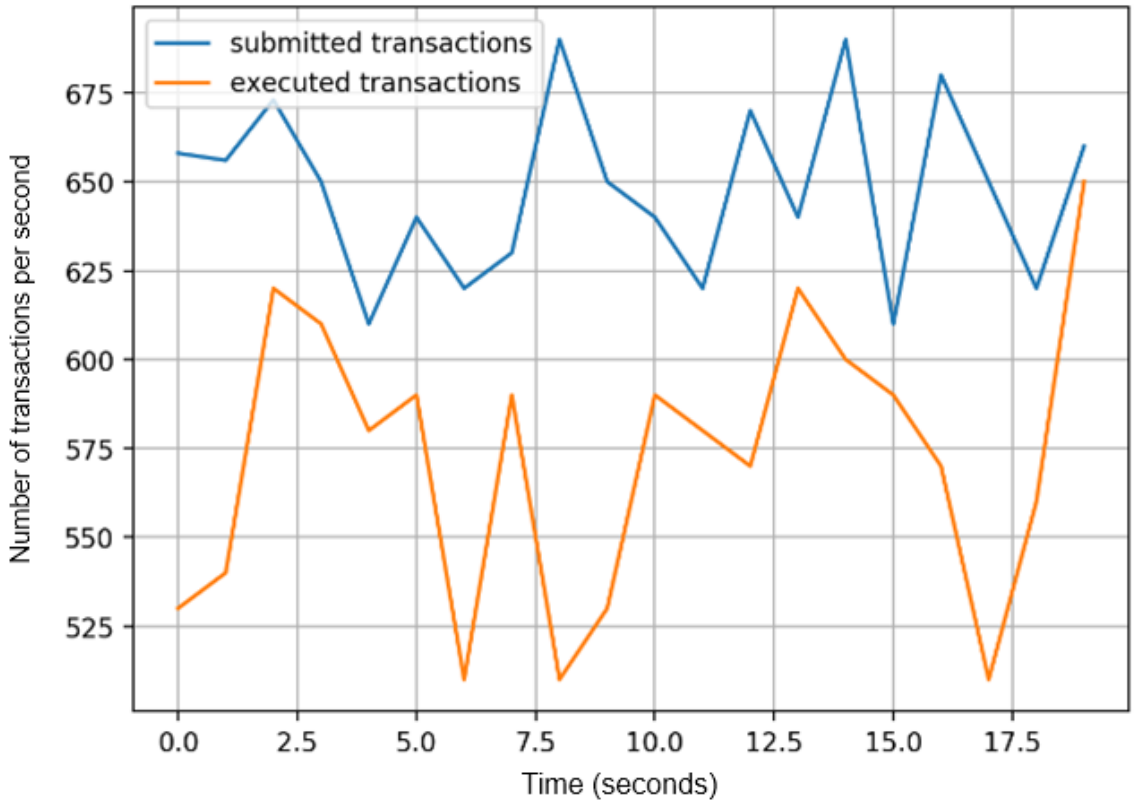

Discussion

Principal Findings

According to the matrix approach [39] proposed for evaluating the knowledge contribution in DSR from the information system discipline, the overall contribution in this paper falls into the advancement quadrant introducing an innovation. We presented a decentralized telehealth platform design and development approach that can help guide future blockchain integration and new technology adoption. We have summarized the blockchain telehealth design implications (DIs) as our knowledge contribution to the design theory for blockchain integration strategies.

DI1: Prioritize User Requirements in the Business Process Design

Try to focus on the human experience and consider the actions and needs of patients who will use the portal and not just focus on the technical aspects of the system design. Recognize the importance of collaboration with patients to ultimately deliver an impactful experience. User-centric design will benefit every stage of the research, from testing and validation to solutions, in order to meet user needs. Consumer preferences will play an important role in the transition from the pandemic period to the postpandemic period [60]. The blockchain integration process is hidden from users, and most of the time, users are not required or not aware of their actions to interact with blockchain networks. Thus, when designing smart contracts, try to optimize the workflow with user-centered principles and provide maximized simplicity.

DI2: Redesign the Business Process With Minimal Change

Business processes are not standalone; thus, it is very important to reuse the components of blockchain integration and share the common design artifacts across the system development lifecycle. For health care providers, there is always a need to minimize costs without reducing the quality of the services provided to customers [61].

DI3: Increase Security and Business Value

Security design should be baked into the entire lifecycle of telehealth design. Specifically, data management and access control policies should be enforced across the system components involved in patient data operations. Along with deployability, scalability, reliability, and ease of use, patient trust should be taken care of so that users are motivated to adopt the telehealth platform at their comfort level. Do not oversimplify the process, and try to understand the business value up front and focus on long-term benefits.

DI4: Pilot Project Experiment and Iteration

End users can provide feedback about the current iteration of the portal to inform future improvements that will meet their needs. Considering complicated patient demographics and enabling various functionalities in multiple rounds of iterations will boost patient enrollment and engagement [62]. Evaluate different platform options and product features to best suit the business use cases for telehealth innovations.

XSL•FO
RenderX

Limitations

Blockchain as a new technology is not yet fully developed in the health care domain [63] or framed by government rules and regulations [64]. There are some technical limitations of our prototype, especially related to the health care domain. First, generalized and standardized mechanisms for institutional collaboration are required to address the gaps between different medical procedures before, during, and after the telehealth process. Second, the sharing of health care data can lead to unintended intellectual information disclosure if the deployment of the system is not done correctly. Using thorough planning and negotiation between institutions can address this limitation. Third, the incentive mechanism for institutional collaboration could also be explored to address this limitation in future work. Full participation from all stakeholders [65] is extremely essential for promoting the adoption of this innovative architectural design to achieve better health care delivery.

Comparison With Prior Work

Prior work has provided an overview of the adoption of digital technologies in the health care domain, which highlighted blockchain as an innovative solution to provide assistance to diagnosis during a crisis [66]. The necessity of security regulations and privacy policies is indicated to adopt blockchain technology [41], which calls for studies like ours to protect patient data and especially the communication process in the telehealth platform. Previous developments and implementations of smart contracts were built on top of Ethereum for 3 telehealth services, including teleconsultation, drug administration, and medical testing [32]. The BlockHeal framework has been validated by demonstrating several use cases, including user registration and verification, patient online consultation with doctors, health care and laboratory facilities, drug counterfeiting, pharmacovigilance, language translations, and emergency facilities [33]. There are issues with scalability and security features when faced with multiserver architectures [34]. We have compared blockchain-assisted telehealth solutions in Table 3 and have highlighted our contribution.

In summary, while there are applications available to support telehealth services, there are few guidelines for decentralized implementation. Based on blockchain characteristics and the needs associated with telehealth services, we have proposed a set of smart contract–driven telehealth DPs from the design requirements collected from user experiments and academic or industrial initiatives. We have designed and presented 2 artifacts that can facilitate both user end (frontend) and backend system integration with blockchain technology, namely, the user credential management function and access control function, driven by underlying smart contracts. The proof-of-concept project demonstrates the feasibility and verifies the usable potential of the proposed concept and design theory. From the artifact design, we have presented a reflection of the design and have performed an evaluation of the proposed design in terms of security properties and functionality performances, followed by a summary of the lessons learned and contributions to DSR theories.

Table 3. A comparison of blockchain-assisted telehealth solutions.

Solution	Storage	Blockchain role	Smart contract	Blockchain type
Authentication for multiserver edge computing [34]	Edge and cloud	Authentication and key agreement	Practical byzantine fault tolerance (PBFT)	A lightweight consortium blockchain
MEC[a] real-time authentication between IoT, MEC, and cloud [67]	Storage offloading capability	Rewarding scheme	Access control	Permissioned, Hyperledger Fabric
Blockchain-based telehealth framework, BlockHeal [33]	Decentralized Hyperledger Fabric	Decentralized applications (DApps)	Central health regulatory authority	Permissioned, Hyperledger Fabric
Remote patient monitoring [68]	EHR[b] data management using Ethereum	Blockchain-oriented software (BOS) engineering	Solidity	Permissioned, Ethereum and Hyperledger Fabric
MedHypChain [69]	Cloud blockchain	Identity-based broadcast group signcryption mechanism	PBFT chaincode	Permissioned, Hyperledger Fabric
Three telehealth services [32]	IPFS[c], cloud, and blockchain	Tracing, tracking, alert, and provenance	Public	Permissioned, Ethereum
Our solution	IPFS and blockchain	Tracing, tracking, access control, identity management, alert, and provenance	Actor-based smart contract	Permissioned, Rahasak

[a]MEC: multiaccess edge computing.

[b]EHR: electronic health record.

[c]IPFS: InterPlanetary File System.

Conclusion and Future Directions

We presented the design, development, and implementation of a blockchain-assisted and smart contract–driven telehealth platform that can help users communicate remotely and improve medical services efficiently. The system adopts a user-centric approach and follows the behavior design model to promote new technology adoption among different groups of users. We further evaluated the security and performance of the platform, which exhibited the feasibility and efficiency of the proposed platform. Moreover, we concluded with a set of actionable

XSL•FO
RenderX

guidelines and DIs to facilitate the design of a practical blockchain-based telehealth platform. Owing to the limited adoption of blockchain technology in the health care sector, there is a lack of fundamental design theory for blockchain technology integration in the current stage [44]. It is important to address who gets to see what data and when [70]. Future research can dive deep into each service model and analyze the market segmentation as a new ecosystem. The application of nonfungible tokens to health care [71] is another application scenario of blockchain technology that will significantly transform the innovation of the health care industry. For example, telehealth record data can be shared directly with the patient after the televisit and with the health care provider or insurance provider. This might lead to incentive plans that lower insurance premiums paid by patients out-of-pocket when they can demonstrate cost-efficient and flexible health care management.

This study extends the existing body of knowledge and develops a set of actionable guidelines to facilitate the design of a practical blockchain-based telehealth platform, reflecting the rigor of design research contributions. To the best of our knowledge, few studies have focused on this topic. The extension of our work will benefit several new directions for new technology integration, blockchain adoption, data privacy, and trust in multiple aspects of smart and connected health. Adapting to the sharing economy approach, the telehealth platform should explore the potential as a one-stop service application for patients in need and a promising solution for stakeholders in terms of a new business model.

Acknowledgments

The authors thank all the reviewers and editors for their helpful suggestions and comments during the review process. The authors also thank all the colleagues who provided valuable feedback during the manuscript preparation process.

Data Availability

Data will be provided upon reasonable request.

Authors' Contributions

XL, NA, TS, EB, and SS designed the study and developed the study architecture. XL, NA, and TS drafted the manuscript. XL and EB implemented the architecture. SS reviewed the manuscript for intellectual content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital identity-related privacy requirements.
[DOCX File , 17 KB - jmir_v26i1e46556_app1.docx]

Multimedia Appendix 2
Reporting guideline checklist.
[DOCX File , 18 KB - jmir_v26i1e46556_app2.docx]

Multimedia Appendix 3
Demographic characteristics of interview respondents.
[DOCX File , 17 KB - jmir_v26i1e46556_app3.docx]

Multimedia Appendix 4
Block creation time against the number of transactions in the block.
[PNG File , 18 KB - jmir_v26i1e46556_app4.png]

References

1. Peters C, Blohm I, Leimeister JM. Anatomy of successful business models for complex services: Insights from the telemedicine field. Journal of Management Information Systems 2015 Dec 16;32(3):75-104. [doi: 10.1080/07421222.2015.1095034]
2. Bélanger F, Crossler RE. Privacy in the digital age: A review of information privacy research in information systems. MIS Quarterly 2011;35(4):1017. [doi: 10.2307/41409971]
3. Smith HJ, Dinev T, Xu H. Information privacy research: An interdisciplinary review. MIS Quarterly 2011;35(4):989. [doi: 10.2307/41409970]

XSL•FO
RenderX

4. Hann I, Hui K, Lee ST, Png IP. Overcoming online information privacy concerns: An information-processing theory approach. Journal of Management Information Systems 2014 Dec 08;24(2):13-42. [doi: 10.2753/mis0742-1222240202]
5. Hui KL, Teo HH, Lee SYT. The value of privacy assurance: An exploratory field experiment. MIS Quarterly 2007;31(1):19. [doi: 10.2307/25148779]
6. Stewart KA, Segars AH. An empirical examination of the concern for information privacy instrument. Information Systems Research 2002 Mar;13(1):36-49. [doi: 10.1287/isre.13.1.36.97]
7. Sutanto J, Palme E, Tan C, Phang CW. Addressing the personalization-privacy paradox: An empirical assessment from a field experiment on smartphone users. MISQ 2013 Apr 4;37(4):1141-1164. [doi: 10.25300/misq/2013/37.4.07]
8. Mattila J. The Blockchain Phenomenon: The Disruptive Potential of Distributed Consensus Architectures. ETLA Working Papers. 2016. URL: https://www.etla.fi/wp-content/uploads/ETLA-Working-Papers-38.pdf [accessed 2024-09-01]
9. Jacobovitz O. Blockchain for identity management. Beer Sheva: The Lynne and William Frankel Center for Computer Science Department of Computer Science, Ben-Gurion University; 2016.
10. Patel V. Airport passenger processing technology: a biometric airport journey. Embry-Riddle Aeronautical University. 2018. URL: https://commons.erau.edu/cgi/viewcontent.cgi?article=1384&context=edt [accessed 2024-09-01]
11. Zyskind G, Nathan O, Pentland A. Decentralizing privacy: Using blockchain to protect personal data. 2015 Presented at: 2015 IEEE Security and Privacy Workshops; May 21-22, 2015; San Jose, CA. [doi: 10.1109/SPW.2015.27]
12. Khurshid A. Applying blockchain technology to address the crisis of trust during the COVID-19 pandemic. JMIR Med Inform 2020 Sep 22;8(9):e20477 [FREE Full text] [doi: 10.2196/20477] [Medline: 32903197]
13. Chang MC, Park D. How can blockchain help people in the event of pandemics such as the COVID-19? J Med Syst 2020 Apr 16;44(5):102 [FREE Full text] [doi: 10.1007/s10916-020-01577-8] [Medline: 32300906]
14. Koch S. Home telehealth--current state and future trends. Int J Med Inform 2006 Aug;75(8):565-576. [doi: 10.1016/j.ijmedinf.2005.09.002] [Medline: 16298545]
15. Dellifraine JL, Dansky KH. Home-based telehealth: a review and meta-analysis. J Telemed Telecare 2008 Mar 01;14(2):62-66. [doi: 10.1258/jtt.2007.070709] [Medline: 18348749]
16. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/nejmra1601705]
17. Shigekawa E, Fix M, Corbett G, Roby DH, Coffman J. The current state of telehealth evidence: A rapid review. Health Aff (Millwood) 2018 Dec;37(12):1975-1982. [doi: 10.1377/hlthaff.2018.05132] [Medline: 30633674]
18. Lurie N, Carr BG. The role of telehealth in the medical response to disasters. JAMA Intern Med 2018 Jun 01;178(6):745-746. [doi: 10.1001/jamainternmed.2018.1314] [Medline: 29710200]
19. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Mar 20;26(5):309-313. [doi: 10.1177/1357633x20916567]
20. Smith HJ, Milberg SJ, Burke SJ. Information privacy: Measuring individuals' concerns about organizational practices. MIS Quarterly 1996 Jun;20(2):167. [doi: 10.2307/249477]
21. Malhotra NK, Kim SS, Agarwal J. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information Systems Research 2004 Dec;15(4):336-355. [doi: 10.1287/isre.1040.0032]
22. Jiang Z, Heng CS, Choi BCF. Research note—Privacy concerns and privacy-protective behavior in synchronous online social interactions. Information Systems Research 2013 Sep;24(3):579-595. [doi: 10.1287/isre.1120.0441]
23. Tene O, Polonetsky J. Big data for all: Privacy and user control in the age of analytics. Nw J Tech Intell Prop 2013;11(5):Article 1 [FREE Full text]
24. Dinev T, McConnell AR, Smith HJ. Research commentary—Informing privacy research through information systems, psychology, and behavioral economics: thinking outside the “APCO” box. Information Systems Research 2015 Dec;26(4):639-655. [doi: 10.1287/isre.2015.0600]
25. Gopal RD, Hidaji H, Patterson RA, Rolland E, Zhdanov D. How much to share with third parties? User privacy concerns and website dilemmas. MISQ 2018 Jan 1;42(1):143-164. [doi: 10.25300/misq/2018/13839]
26. Menon S, Sarkar S. Privacy and big data: Scalable approaches to sanitize large transactional databases for sharing. MIS Quarterly 2016;40(4):963-982 [FREE Full text] [doi: 10.1201/b19694-18]
27. Son JY, Kim SS. Internet users' information privacy-protective responses: A taxonomy and a nomological model. MIS Quarterly 2008;32(3):503. [doi: 10.2307/25148854]
28. Rundle M, Blakley B, Broberg J. At a crossroads: personhood and digital identity in the information society. Analysis & Policy Observatory. 2008. URL: https://apo.org.au/node/16588 [accessed 2024-09-01]
29. Windley P. Digital Identity: Unmasking Identity Management Architecture (IMA). Sebastopol, CA: O'Reilly Media; 2005.
30. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Quarterly 2004;28(1):75. [doi: 10.2307/25148625]
31. Yaga D, Mell P, Roby N, Scarfone K. Blockchain technology overview. arXiv. 2019. URL: https://arxiv.org/abs/1906.11078 [accessed 2024-09-01]
32. Hasan HR, Salah K, Jayaraman R, Yaqoob I, Omar M, Ellahham S. Blockchain-enabled telehealth services using smart contracts. IEEE Access 2021;9:151944-151959. [doi: 10.1109/access.2021.3126025]

XSL•FO
RenderX

33. Bawany NZ, Qamar T, Tariq H, Adnan S. Integrating healthcare services using blockchain-based telehealth framework. IEEE Access 2022;10:36505-36517. [doi: 10.1109/access.2022.3161944]
34. Wang W, Huang H, Xue L, Li Q, Malekian R, Zhang Y. Blockchain-assisted handover authentication for intelligent telehealth in multi-server edge computing environment. Journal of Systems Architecture 2021 May;115:102024. [doi: 10.1016/j.sysarc.2021.102024]
35. Lin IC, Liao TC. A survey of blockchain security issues and challenges. Int J Netw Secur 2017;19(5):653-659 [FREE Full text]
36. Walsh C, OReilly P, Gleasure R, Feller J, Li S, Cristoforo J. New kid on the block: a strategic archetypes approach to understanding the Blockchain. 2016 Presented at: Thirty Seventh International Conference on Information Systems; December 11-14, 2016; Dublin, Ireland URL: https://core.ac.uk/reader/301370270
37. Christiansen R. Blockchain: An Exploded View. Academia. 2016. URL: https://www.academia.edu/26216202/Blockchain_An_Exploded_View [accessed 2024-09-01]
38. Ayed GB. Architecting User-Centric Privacy-as-a-Set-of-Services: Digital Identity-Related Privacy Framework. Cham: Springer; 2014.
39. Gregor S, Hevner AR. Positioning and presenting design science research for maximum impact. MISQ 2013 Feb 2;37(2):337-355. [doi: 10.25300/misq/2013/37.2.01]
40. Fogg B. A behavior model for persuasive design. In: Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology. 2009 Presented at: 4th International Conference on Persuasive Technology; April 26-29, 2009; Claremont, CA. [doi: 10.1145/1541948.1541999]
41. Ahmad RW, Salah K, Jayaraman R, Yaqoob I, Ellahham S, Omar M. The role of blockchain technology in telehealth and telemedicine. Int J Med Inform 2021 Apr;148:104399 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104399] [Medline: 33540131]
42. Takeda H, Veerkamp P, Yoshikawa H. Modeling design process. AI Mag 1990;11(4):37. [doi: 10.1609/aimag.v11i4.855]
43. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI -Statement on reporting of evaluation studies in health informatics. Yearb Med Inform 2018 Mar 07;18(01):23-31. [doi: 10.1055/s-0038-1638633]
44. Koshechkin K, Lebedev G, Radzievsky G, Seepold R, Martinez NM. Blockchain technology projects to provide telemedical services: Systematic review. J Med Internet Res 2021 Aug 18;23(8):e17475 [FREE Full text] [doi: 10.2196/17475] [Medline: 34407924]
45. Chanson M, Bogner A, Bilgeri D, Fleisch E, Wortmann F. Blockchain for the IoT: Privacy-preserving protection of sensor data. JAIS 2019:1272-1307. [doi: 10.17705/1jais.00567]
46. Determining if IRB Review is Needed. Florida International University. URL: https://research.fiu.edu/irb/determine-irb-review/ [accessed 2024-06-10]
47. Rathee T, Singh P. MedSecureChain: Applying Blockchain for Delegated Access in Health Care. In: Kountchev R, Mironov R, Li S, editors. New Approaches for Multidimensional Signal Processing. Smart Innovation, Systems and Technologies, vol 216. Singapore: Springer; 2021:153-163.
48. Patel PD, Cobb J, Wright D, Turer R, Jordan T, Humphrey A, et al. Rapid development of telehealth capabilities within pediatric patient portal infrastructure for COVID-19 care: barriers, solutions, results. J Am Med Inform Assoc 2020 Jul 01;27(7):1116-1120 [FREE Full text] [doi: 10.1093/jamia/ocaa065] [Medline: 32302395]
49. Bakken S. Telehealth: Simply a pandemic response or here to stay? J Am Med Inform Assoc 2020 Jul 01;27(7):989-990 [FREE Full text] [doi: 10.1093/jamia/ocaa132] [Medline: 32692845]
50. Meth H, Mueller B, Maedche A. Designing a requirement mining system. JAIS 2015 Sep;16(9):799-837. [doi: 10.17705/1jais.00408]
51. Djamasbi S, Strong D, Wilson EV, Ruiz C. Designing and testing user-centric systems with both user experience and design science research principles. AMCIS Proc 2016:11 [FREE Full text]
52. Holzinger A. User-Centered Interface Design for Disabled and Elderly People: First Experiences with Designing a Patient Communication System (PACOSY). In: Miesenberger K, Klaus J, Zagler W, editors. Computers Helping People with Special Needs. ICCHP 2002. Lecture Notes in Computer Science, vol 2398. Berlin, Heidelberg: Springer; 2002:33-40.
53. Poonsuph R. The design blueprint for a large-scale telehealth platform. Int J Telemed Appl 2022 Jan 5;2022:8486508-8486515 [FREE Full text] [doi: 10.1155/2022/8486508] [Medline: 35035477]
54. Bandara E, Liang X, Foytik P, Shetty S, Ranasinghe N, De Zoysa K. Rahasak—Scalable blockchain architecture for enterprise applications. Journal of Systems Architecture 2021 Jun;116:102061. [doi: 10.1016/j.sysarc.2021.102061]
55. Lukša M. Kubernetes in Action. Munich, Germany: Carl Hanser Verlag; 2017.
56. Thönes J. Microservices. IEEE Softw 2015 Jan;32(1):116-116. [doi: 10.1109/ms.2015.11]
57. Destounis A, Paschos GS, Koutsopoulos I. Streaming big data meets backpressure in distributed network computation. 2016 Presented at: 35th Annual IEEE International Conference on Computer Communications; April 10-14, 2016; San Francisco, CA. [doi: 10.1109/infocom.2016.7524388]
58. Paquin C. U-Prove Technology Overview V1.1 (Revision 2). Microsoft. URL: https://www.microsoft.com/en-us/research/publication/u-prove-technology-overview-v1-1-revision-2/ [accessed 2024-09-01]

XSL•FO
RenderX

59. Zawada S. Poster: Blockchain Application Feasibility in Telehealth for DSCSA Compliance. 2020 Presented at: 2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE); December 16-18, 2020; Crystal City, VA. [doi: 10.1145/3384420.3431772]
60. Possible Futures: Telehealth During and After the Pandemic. YouTube. URL: https://www.youtube.com/watch?v=D4FnZZNnsDA [accessed 2022-09-30]
61. Jain N, Gupta V, Dass P. Chapter 3 - Blockchain: A novel paradigm for secured data transmission in telemedicine. In: Jude HD, Gupta D, Khanna A, Khamparia A, editors. Wearable Telemedicine Technology for the Healthcare Industry: Product Design and Development. Cambridge, MA: Academic Press; 2022:33-52.
62. Integrated Patient Portals and Improving the Virtual Experience. YouTube. URL: https://www.youtube.com/watch?v=-kddMUrrie4 [accessed 2022-09-30]
63. Yeung K. The health care sector's experience of blockchain: A cross-disciplinary investigation of its real transformative potential. J Med Internet Res 2021 Dec 20;23(12):e24109 [FREE Full text] [doi: 10.2196/24109] [Medline: 34932009]
64. Dubovitskaya A, Baig F, Xu Z, Shukla R, Zambani PS, Swaminathan A, et al. ACTION-EHR: Patient-centric blockchain-based electronic health record data management for cancer care. J Med Internet Res 2020 Aug 21;22(8):e13598 [FREE Full text] [doi: 10.2196/13598] [Medline: 32821064]
65. Mackey TK, Miyachi K, Fung D, Qian S, Short J. Combating health care fraud and abuse: Conceptualization and prototyping study of a blockchain antifraud framework. J Med Internet Res 2020 Sep 10;22(9):e18623 [FREE Full text] [doi: 10.2196/18623] [Medline: 32909952]
66. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: Systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
67. Hewa T, Braeken A, Ylianttila M, Liyanage M. Multi-Access Edge Computing and Blockchain-based Secure Telehealth System Connected with 5G and IoT. 2020 Presented at: GLOBECOM 2020 - 2020 IEEE Global Communications Conference; December 07-11, 2020; Taipei, Taiwan. [doi: 10.1109/globecom42002.2020.9348125]
68. Hossain FM, Shahriar H, Valero M, Sneha S, Ahamed S, Rahman M. Towards Blockchain-Based Secure Data Management for Remote Patient Monitoring. 2021 Presented at: 2021 IEEE International Conference on Digital Health (ICDH); September 05-10, 2021; Chicago, IL. [doi: 10.1109/icdh52753.2021.00054]
69. Kumar M, Chand S. MedHypChain: A patient-centered interoperability hyperledger-based medical healthcare system: Regulation in COVID-19 pandemic. J Netw Comput Appl 2021 Apr 01;179:102975 [FREE Full text] [doi: 10.1016/j.jnca.2021.102975] [Medline: 34173429]
70. Chatterjee S, Byun J, Dutta K, Pedersen RU, Pottathil A, Xie H. Designing an Internet-of-Things (IoT) and sensor-based in-home monitoring system for assisting diabetes patients: iterative learning from two case studies. European Journal of Information Systems 2018 Jul 03;27(6):670-685. [doi: 10.1080/0960085x.2018.1485619]
71. Gambril J, Boyd C, Egbaria J. Application of nonfungible tokens to health care. Comment on "Blockchain Technology Projects to Provide Telemedical Services: Systematic Review". J Med Internet Res 2022 May 30;24(5):e34276 [FREE Full text] [doi: 10.2196/34276] [Medline: 35635749]

Abbreviations

DI: design implication
DIPR: digital identity–related privacy requirement
DP: design principle
DSR: design science research
FBM: Fogg Behavior Model
SSI: self-sovereign identity
UX: user experience

Edited by T de Azevedo Cardoso, T Leung; submitted 15.02.23; peer-reviewed by W Chen, Y Zhuang, W Chen; comments to author 02.01.24; revised version received 17.01.24; accepted 18.06.24; published 25.09.24.

Please cite as:
Liang X, Alam N, Sultana T, Bandara E, Shetty S
Designing A Blockchain-Empowered Telehealth Artifact for Decentralized Identity Management and Trustworthy Communication: Interdisciplinary Approach
J Med Internet Res 2024;26:e46556
URL: https://www.jmir.org/2024/1/e46556
doi:10.2196/46556
PMID:39320943

XSL•FO
RenderX

©Xueping Liang, Nabid Alam, Tahmina Sultana, Eranga Bandara, Sachin Shetty. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automatic Recommender System of Development Platforms for Smart Contract–Based Health Care Insurance Fraud Detection Solutions: Taxonomy and Performance Evaluation

Rima Kaafarani[1], MSc; Leila Ismail[2], PhD; Oussama Zahwe[3], PhD

[1]Intelligent Computing and Communication Systems Laboratory, Computer Science Department, American University of Culture and Education, Beirut, Lebanon

[2]Intelligent Distributed Computing and Systems Laboratory, Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

[3]Vectrawave Device, Lannion, France

Corresponding Author:
Leila Ismail, PhD
Intelligent Distributed Computing and Systems Laboratory, Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University
15551 Al Maqam Campus
Al Ain, Abu Dhabi, 15551
United Arab Emirates
Phone: 971 37673333 ext 5530
Email: leila@uaeu.ac.ae

Abstract

Background: Health care insurance fraud is on the rise in many ways, such as falsifying information and hiding third-party liability. This can result in significant losses for the medical health insurance industry. Consequently, fraud detection is crucial. Currently, companies employ auditors who manually evaluate records and pinpoint fraud. However, an automated and effective method is needed to detect fraud with the continually increasing number of patients seeking health insurance. Blockchain is an emerging technology and is constantly evolving to meet business needs. With its characteristics of immutability, transparency, traceability, and smart contracts, it demonstrates its potential in the health care domain. In particular, self-executable smart contracts are essential to reduce the costs associated with traditional paradigms, which are mostly manual, while preserving privacy and building trust among health care stakeholders, including the patient and the health insurance networks. However, with the proliferation of blockchain development platform options, selecting the right one for health care insurance can be difficult. This study addressed this void and developed an automated decision map recommender system to select the most effective blockchain platform for insurance fraud detection.

Objective: This study aims to develop smart contracts for detecting health care insurance fraud efficiently. Therefore, we provided a taxonomy of fraud scenarios and implemented their detection using a blockchain platform that was suitable for health care insurance fraud detection. To automatically and efficiently select the best platform, we proposed and implemented a decision map–based recommender system. For developing the decision-map, we proposed a taxonomy of 102 blockchain platforms.

Methods: We developed smart contracts for 12 fraud scenarios that we identified in the literature. We used the top 2 blockchain platforms selected by our proposed decision-making map–based recommender system, which is tailored for health care insurance fraud. The map used our taxonomy of 102 blockchain platforms classified according to their application domains.

Results: The recommender system demonstrated that Hyperledger Fabric was the best blockchain platform for identifying health care insurance fraud. We validated our recommender system by comparing the performance of the top 2 platforms selected by our system. The blockchain platform taxonomy that we created revealed that 59 blockchain platforms are suitable for all application domains, 25 are suitable for financial services, and 18 are suitable for various application domains. We implemented fraud detection based on smart contracts.

Conclusions: Our decision map recommender system, which was based on our proposed taxonomy of 102 platforms, automatically selected the top 2 platforms, which were Hyperledger Fabric and Neo, for the implementation of health care insurance fraud detection. Our performance evaluation of the 2 platforms indicated that Fabric surpassed Neo in all performance metrics, as depicted by our recommender system. We provided an implementation of fraud detection based on smart contracts.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e50730) doi:10.2196/50730

KEYWORDS

blockchain; blockchain development platform; eHealth; fraud detection; fraud scenarios; health care; health care insurance; health insurance; machine learning; medical informatics; recommender system; smart contract; taxonomy

Introduction

Rationale

Health care insurance fraud presents a significant challenge for both the medical industry and government bodies. It represents a serious concern for the insurance industry due to fraud's financial impact on policyholders and insurance companies. According to the National Health Care Anti-Fraud Association, health care insurance fraud leads to the loss of tens of billions of dollars annually [1]. In 2020, according to the US Department of Justice, a noteworthy accomplishment in combating health care insurance fraud, recovering US $2.7 billion through settlements and judgment, was announced; however, it represented a significant 50% increase compared to the previous year [2]. Furthermore, the global health care insurance fraud analytics market demonstrates substantial growth, rising from US $2.43 billion in 2022 to US $3.09 billion in 2023, reflecting a compound annual growth rate of 27% [3]. On the other hand, health insurance is crucial to ensure people's lives due to the high cost of medical treatments. The advantages of health insurance are being threatened by theft and fraudulent claims. With the increasing number of patient demands for health insurance, manual auditing for validating and pinpointing fraud is no longer efficient. Therefore, it is essential to create an automatic and efficient system that detects fraud.

Hence, machine learning solutions for detecting fraud that rely on data sets to train models for fraud detection have been introduced [4,5]. However, they raise ethical concerns as the trained models could be biased toward the majority [6], and there are also privacy and security issues [7] due to the potential compromise of sensitive personally identifiable information of patients. These considerations would have severe consequences, including reputational damage to insurance firms. Machine learning models should rely on high-quality data [8]. Therefore, they are not trustworthy so far. Recently, blockchain has emerged as a decentralized technology to implement secure transactions in a peer-to-peer network. It consists of a series of interconnected blocks of transactions. Each block contains data and is secured through cryptographic measures, such as hash functions and asymmetric encryption [9]. Transactions occur between nodes in a peer-to-peer network without the need for a central authority. All transactions are recorded in an immutable ledger, and peers can only add to the ledger, not alter or delete any previously recorded information [10]. When a new node joins the network, it downloads a copy of the ledger. Before adding a block to the blockchain, a consensus is reached among peers. In addition, blockchain can execute smart contracts [11].

Blockchain has demonstrated its potential in various domains, including the health care system [12,13]. In particular, smart contracts in a blockchain were introduced as self-executing agents based on the transactions being executed [14]. However, there is a proliferation of blockchain development platforms in the literature with various characteristics, imposing challenges for software developers to determine suitable platforms that include the functionalities needed to implement insurance fraud detection solutions based on smart contracts. In this paper, we propose an automated decision map recommender system specifically designed to select the most suitable blockchain platform among the proposed platforms in the literature. We exemplify the use of our proposed recommender system by implementing smart contract–based solutions for insurance fraud detection on the selected platform. The main contributions of this research are as follows:

- We proposed and developed an innovative, adaptive, and automated recommender system based on our proposed decision map. The map evaluates blockchain platforms, considering selected and categorized blockchain features. to suggest the most suitable platform. The system is flexible and responsive to changes, ensuring that, if a platform becomes unavailable or gains new features, it will generate updated results accordingly.
- We introduced a decision-making map recommender system that allows us to identify the best blockchain platform that is adequate for the implementation of health care insurance fraud detection. The decision map is generic and can be applied to any other domain.
- We developed a taxonomy of blockchain development platforms, used to determine the characteristics of the platforms that are available for implementing applications in the health insurance field. The platform taxonomy is based on the investigation of 102 blockchain platforms and their applications domains in the literature.
- We exemplified the applicability of our automated decision map recommender system by developing and implementing blockchain smart contracts for the detection of 12 fraud scenarios.
- We evaluated the implementation of our recommender system by applying it to 42 blockchain platforms. Consequently, we developed and implemented the detection of fraud on the top 2 platforms recommended by our decision-making map recommender system and evaluated their performances.
- We made the recommender system toolkit and source code available on GitHub for blockchain developers.

Related Works

To our knowledge, no work in the literature has automated the selection of a suitable blockchain development platform for a specific use case, such as health insurance fraud detection. In the study by Farshidi et al [15], the authors divided blockchain features into four categories: (1) *must-have*, which indicates that the platform needed to include the specified blockchain feature to be deemed suitable; (2) *should-have*, which implies that the defined blockchain features are highly recommended; (3) *could-have*, which represent optional blockchain features;

XSL•FO
RenderX

and (4) *won't-have*, which meant to list the features that are not required by the developer. However, this method may not be precise as blockchain platforms often possess multiple features; for instance, some platforms offer various consensus mechanisms. Thus, classifying a single consensus into the *won't-have* category could unjustly disqualify a blockchain platform that might otherwise be suitable for the use case. In addition, our system implements software that can be used by any clinic or hospital interested in adopting blockchain platforms. Moreover, the system is adaptive as it allows for adding both blockchain platforms and features as well as the modification of existing ones.

Some works have introduced machine learning and deep learning models for identifying fraud and overcoming the constraints of manual detection methods. Learning models automate the detection process and enhance the analysis of patterns. As shown in Table 1, the study by Lu et al [16] proposes a deep learning graph model, which relies on an attributed heterogeneous information network with a hierarchical attention mechanism. The study by Sowah et al [17] develops a decision support system using Genetic Support Vector Machines to enhance the detection and classification of health insurance fraud in Ghana. The study by Settipalli et al [18] proposes an unsupervised multivariate analysis model named Weighted MultiTree Density-Based Clustering. However, the use of artificial intelligence for detecting health care insurance fraud has raised security concerns, largely due to the sensitive client data used in training the models, consequently suffering from privacy and security issues. In addition, these works do not consider the bias introduced by the use of machine learning or deep learning algorithms. As a result, our emphasis will be on solutions that leverage smart contracts, which are self-executing agreements with predefined rules that activate when conditions are fulfilled. These contracts are immutable, meaning that they cannot be altered once deployed, providing a secure and privacy-preserving blockchain solution for detecting health care insurance fraud. Moreover, throughput, latency, and Central Processing Unit (CPU) and memory use have not been taken into account in the aforementioned works.

Therefore, researchers and developers are turning toward privacy-preserving and secure blockchain-based solutions that incorporate smart contracts for the detection of health insurance fraud. These contracts execute automatically under set conditions once deployed on the blockchain, benefiting from the platform's immutability, decentralization, and transparency, and cannot be changed after they are set up. The study by Mackey et al [19] focuses on determining whether a claim adheres to the applicable provisions of the health care insurance policy. The study by Saldamli et al [20] proposes a solution for preventing health insurance fraud by using 2 fraud scenarios. The study by Liu et al [21] uses the Ethereum blockchain to develop a framework for recording claim data and transaction patients as validators to assist in the detection of fraud. However, none of these works takes into account all possible fraud scenarios; the quality of service of fraud detection in terms of throughput and latency; or computing resource use, such as CPU and memory. In addition, the use of blockchain platforms is unjustified, and the choice of the development platform is not justified. Our recommender system is adaptive to the evolution of blockchain platforms, offering a comprehensive approach. Furthermore, smart contracts are portable and can operate across different platforms. In this study, we implemented smart contracts based on a blockchain development platform that is selected by our adaptive automatic decision map recommender system. On the basis of these fraud scenarios, we implemented smart contracts for insurance fraud detection on the top 2 blockchain development platforms selected by our recommender system. The strengths and weaknesses of recent works using blockchain development platforms for detecting health care insurance fraud are summarized in Table 2.

XSL•FO
RenderX

Table 1. Summary of related works on fraud detection machine learning (ML) and deep learning (DL) algorithms in health insurance claims.

Algorithm under study	Number of fraud scenarios detected	Considering privacy and security	Considering bias issue	ML or DL	Throughput	Latency	CPU[a] use	Memory use	Data set	Metrics
MHAMFD[b] [16]	NR[c]	X[d]	X	DL	X	X	X	X	• Medical-1: balanced data set with a ratio of positive to negative samples of 1:2 • Medical-2: unbalanced data set with a ratio of positive to negative samples of approximately 1:70	• Medical-1: • Accuracy: 0.8961 • F1-score: 0.8694 • Medical-2: • F1-score: 0.8361 • Recall: 0.8764 • Precision: 0.9194
GSVMs[e] [17]	NR	X	X	ML	X	X	X	X	• 100-claim data set • 300-claim data set • 500-claim data set • 750-claim data set • 1000-claim data set	• 100-claim data set accuracy: 71.43% • 300-claim data set accuracy: 95.45% • 500-claim data set accuracy: 99.18% • 750-claim data set accuracy: 82.56% • 1000-claim data set accuracy: 90.91%
WMTDBC[f] [18]	NR	X	X	ML	X	X	X	X	• The data set used in the study was the claims data submitted by health care providers under the US Medicare CMS[g] Part B health care program	• Overall accuracy ranged from 0.857 to 0.946.

[a]CPU: Central Processing Unit.

[b]MHAMFD: Multilevel Hierarchical Attention Mechanism for Fraud Detection.

[c]NR: not reported.

[d]X: not considered.

[e]GSVM: Genetic Support Vector Machine.

[f]WMTDBC: Weighted MultiTree Density-Based Clustering.

[g]CMS: Centers for Medicare & Medicaid Services.

XSL•FO
RenderX

Table 2. Summary of related works on blockchain-based health care insurance fraud detection.

Study	Throughput	Latency	CPU[a] use	Memory use	Fraud scenarios considered, N	Smart contract	Recommender system	Platform	Reason for choosing the platform
Mackey et al [19]	X[b]	X	X	X	1	✓[c]	X	Ethereum	NR[d]
Saldamli et al [20]	X	X	X	X	2	✓	X	BigchainDB	NR
Liu et al [21]	X	X	X	X	3	✓	X	NR	NR
Our work	✓	✓	✓	✓	12	✓	✓	Hyperledger Fabric and Neo	On the basis of our proposed decision-making map recommender system tailored to health care insurance fraud detection

[a]CPU: Central Processing Unit.

[b]X: not considered.

[c]✓: considered.

[d]NR: not reported.

Methods

Overview

The taxonomy of blockchain platforms was based on reviewing published research articles and white papers that mentioned blockchain platforms. Our study revealed 102 blockchain platforms that we classified according to the application domains they were developed for, such as financial services, social media, Internet of Things, and platforms that can be used across several domains. In addition, we gathered information on various features, such as whether the platform is open source, the consensus mechanism used, the type of blockchain used, and the availability of smart contracts. For the detection of health care insurance fraud, the fraud scenarios were based on the study by Ismail and Zeadally [22], which proposes a taxonomy of 12 fraud scenarios that are divided into 7 categories, as shown in Figure 1. The first category is *commission-based*, which includes 3 fraud scenarios. The first scenario involves a health care provider directing patients to specific hospitals, clinics, pharmacies, medications, or equipment suppliers in return for a commission. The second fraud scenario involves pharmacies dispensing specific brands of medicines in exchange for commissions from pharmaceutical companies. The third fraud scenario involves pharmaceutical companies offering incentives to physicians to recommend unapproved or off-label drugs. The second category is *Pinning the System*, which involves health care providers guiding patients to internal entities such as laboratories or pharmacies to keep profits within the organization. The third category, *Waiving Copayments*, is where the physician regularly waives patients' copayments and overcharges the health care provider. The fourth category, *Managed Care*, consists of organizations limiting costs by denying necessary care, providing substandard treatment, and creating administrative barriers for patients. The fifth category, *Billing Manipulation*, consists of 4 fraud scenarios. The first involves unlicensed hospitals and physicians billing patients for care. The second scenario occurs when a physician alters a diagnosis on a claim without the patient's knowledge. The third scenario involves health care providers offering unnecessary care, inflating service hours, submitting duplicate claims, phantom billing, or substituting diagnosis codes for higher reimbursements. The final scenario involves medical equipment providers inflating prices for insured patients or claiming expensive equipment while supplying cheaper alternatives. The sixth category, *Physician Shopping*, involves patients consulting multiple health care providers to obtain prescriptions for nonmedical use. Finally, the seventh category, *Self-referral*, occurs when physicians direct patients to clinics or health care facilities in which they have a financial interest, potentially leading to conflicts of interest.

XSL•FO
RenderX

Figure 1. Taxonomy of health care insurance fraud scenarios [22].

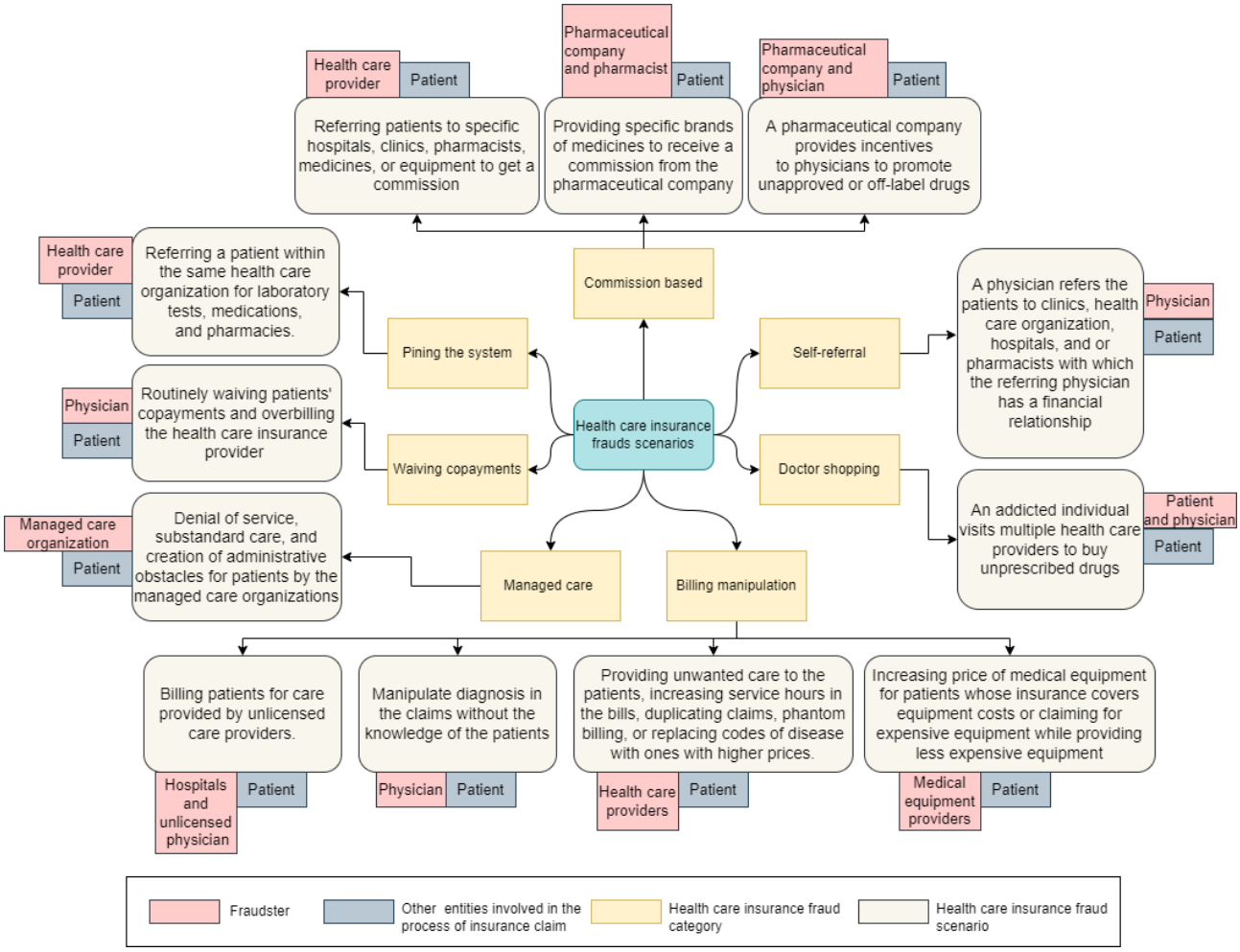

Ethical Considerations

The study does not involve any personal or patient-related data, focusing solely on blockchain platforms features. As no human subjects are included and no identifiable information is used, the research does not require ethics review board assessment, in accordance with institutional and regional guidelines for nonhuman subjects research.

Results

Decision-Making Map Recommender System for Selecting the Best Blockchain Platform for Health Care Insurance Fraud Detection

Overview

The proliferation of blockchain platforms has led to a multitude of choices for developers. However, it is important to note that the various blockchain platforms available today have different features, capabilities, and use cases [23,24]. Therefore, developers need to evaluate the available options and select the platform that best fits their specific needs. In this section, we provide an overview of our taxonomy, which encompasses 102 blockchain platforms. Subsequently, we present our feature-based decision map recommender system to select the best platform.

Taxonomy of Blockchain Platforms

In 2008, Bitcoin [25] made its debut, and the subsequent addition of smart contract technology by Ethereum [26] contributed significantly to the rapid growth and development of blockchain technology. As a result, >100 distinct blockchain platforms were developed for various purposes. To provide a comprehensive understanding of these platforms, we present a taxonomy of 102 blockchain platforms, which we organized based on their respective application domains. Along with the application domain, our classification takes into account the open-source nature of the platform, the consensus mechanism used, the type of blockchain, and the platform's capability to support smart contract development. The taxonomy of blockchain platforms is presented as a graph. Figure 2 shows an overview of the different generic blockchain platforms that can be used to build a wide range of applications.

Figure 3 presents the blockchain platforms that have been specifically designed for financial services, whereas Figure 4 presents the platforms that are tailored to meet the needs of a particular application domain. These platforms offer specialized features and functionality to cater to the specific needs of their respective industries or sectors, thus providing a more specialized and customized solution for these specific use cases.

XSL•FO
RenderX

Figure 2. Graphical representation of the taxonomy of blockchain platforms—generic blockchain platforms. aBFT: Asynchronous Byzantine Fault Tolerant; BFT: Byzantine Fault Tolerance; dBFT: Delegated Byzantine Fault Tolerance; DPoS: Delegated Proof of Stake; FPC: Fast Probabilistic Consensus; IBFT: Istanbul Byzantine Fault Tolerance; pBFT: Practical Byzantine Fault Tolerance; PoA: Proof of Authority; PoET: Proof of Elapsed Time; PoP: Proof of Pledge; PoS: Proof of Stake; PoW: Proof of Work; PoX: Proof of Transfer; PPoS: Pure Proof of Stake; RAFT: Reliable, Replicated, Redundant, and Fault-Tolerant; RoR: Return on Reputation; SBFT: Smilo Byzantine Fault Tolerance; XDPoS: Extended Delegated Proof of Stake; YAC: Yet Another Consensus.

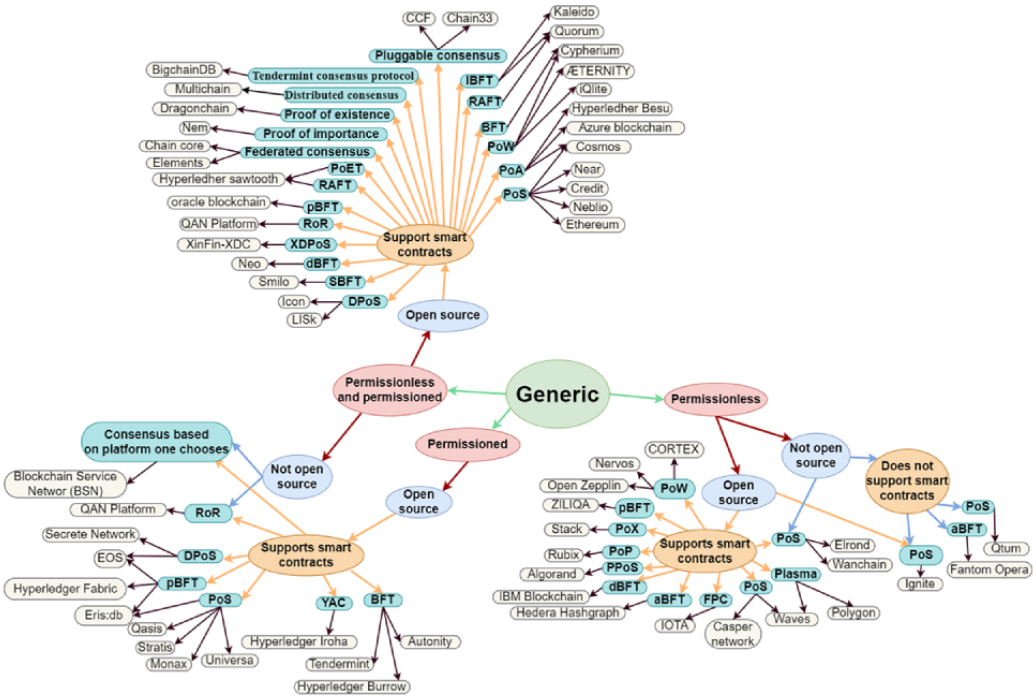

Figure 3. Graphical representation of the taxonomy of blockchain platforms—blockchain platforms dedicated to financial services. aBFT: Asynchronous Byzantine Fault Tolerant; BFT: Byzantine Fault Tolerance; DAG: Directed Acyclic Graph; dBFT: Delegated Byzantine Fault Tolerance; DeFi: Decentralized Finance; DPoS: Delegated Proof of Stake; FPC: Fast Probabilistic Consensus; IBFT: Istanbul Byzantine Fault Tolerance; LibraBFT: Libra Byzantine Fault Tolerance; NFT: None Fungible Tokens; pBFT: Practical Byzantine Fault Tolerance; PoA: Proof of Authority; PoC: Proof of Capacity; PoET: Proof of Elapsed Time; PoH: Proof of History; PoO: Proof of Ownership; PoP: Proof of Pledge; PoS: Proof of Stake; PoSA: Proof of Staked Authority; PoW: Proof of Work; PoX: Proof of Transfer; PPoS: Pure Proof of Stake; RAFT: Reliable, Replicated, Redundant, and Fault-Tolerant; SBFT: Smilo Byzantine Fault Tolerance; SCP: Stellar Consensus Protocol; XDPoS: Extended Delegated Proof of Stake; XRP: Ripple; YAC: Yet Another Consensus.

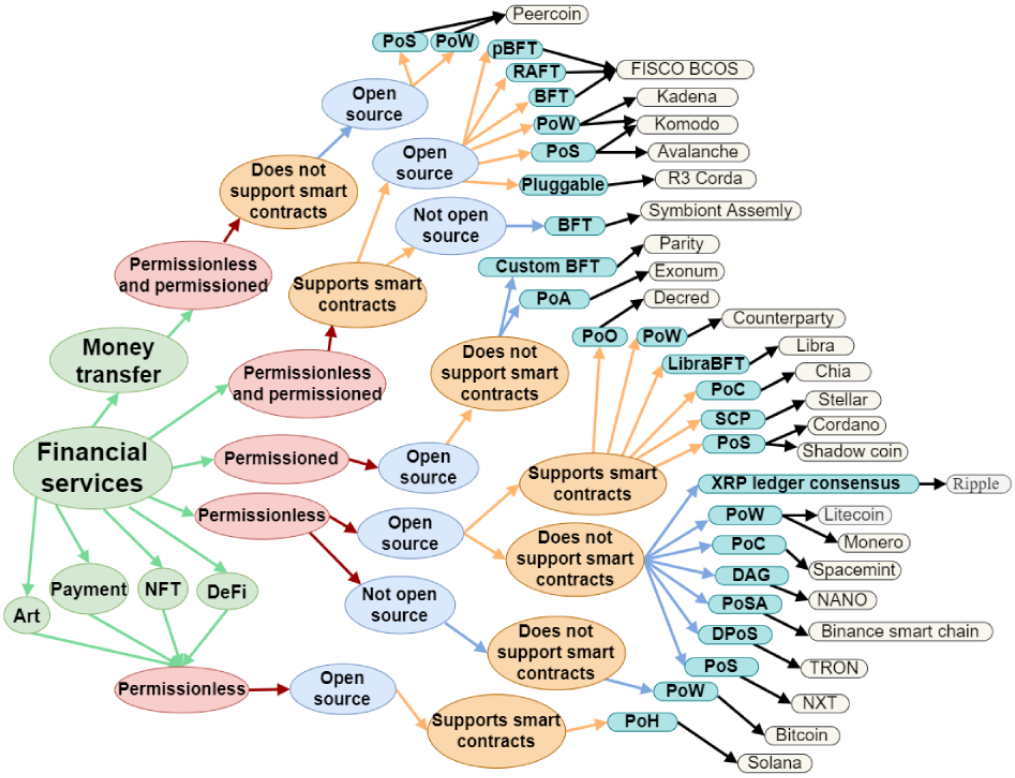

XSL•FO
RenderX

Figure 4. Graphical representation of the taxonomy of blockchain platforms—blockchain platforms mapped to specific application domains. BFT: Byzantine Fault Tolerance; DAG: Directed Acyclic Graph; DPoS: Delegated Proof of Stake; IoT: Internet of Things; PoA: Proof of Authority; PoC: Proof of Capacity; PoR: Proof of Randomness; PoS: Proof of Stake; PoW: Proof of Work; RBFT: Redundant Byzantine Fault Tolerance; ZAB: ZooKeeper Atomic Broadcas.

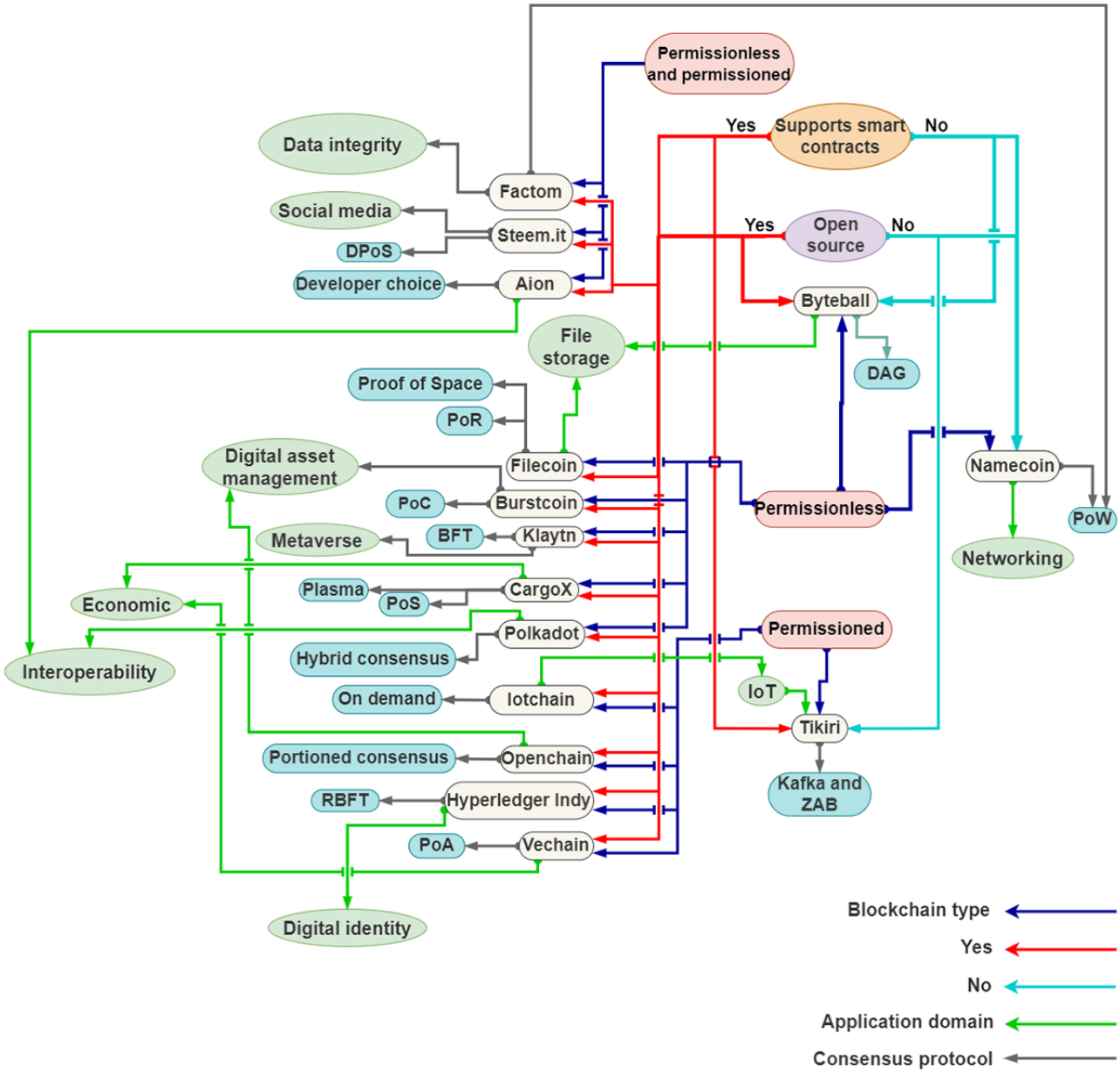

Decision-Making Map–Based Recommender System

While a blockchain platform selection method was proposed in the study by Farshidi et al [15], it included unnecessary categories of features and did not specifically focus on the detection of insurance fraud. Therefore, we propose a decision-making map that is tailored specifically to health care insurance fraud detection solutions. It classifies blockchain features into 3 main categories: *compulsory features*, which are essential to the platform; *mandatory features*, which are sufficient; and *possible features*, which are desirable but not necessary. These categories differ in weight, which determines the value of one feature over another. As shown in Figure 5, our map offers a targeted approach to selecting a blockchain platform for developing health insurance fraud detection mechanisms.

In the health care insurance domain, privacy is a crucial aspect as insurance companies deal with sensitive patient data [12]. Several research works in health care have implemented blockchain technology to ensure integrity, accountability, and nonrepudiation in the claim process [19,22]. The study by Ismail and Zeadally [22] proposes a blockchain system for health care insurance antifraud that ensures trusted medical process information entry and reading as well as a data privacy protection scheme. In the study by Mackey et al [19], a blockchain system is proposed and implemented to prevent counterfeiting in health care insurance, providing a secure and private system.

To implement health care insurance fraud detection using blockchain, we should select features that ensure privacy, such as on-chain transactions and permissioned platforms. This is in addition to other technical features that should be available in

XSL•FO
RenderX

the platform, such as the smart contract and user interface development tool features. In summary, we determined the most suitable features in terms of both their relevance to the task of health care insurance fraud detection and the technical capabilities of the platforms. We divided these features into 3 categories, which are compulsory, mandatory, and possible features (Textbox 1).

On the basis of the aforementioned selected features (compulsory, mandatory, and possible), our enforced decision-making map selects 42 platforms out of 102. This extraction of 42 platforms is derived from our proposed taxonomy of blockchain platforms. This taxonomy maps the blockchain platforms into their corresponding application domains and blockchain features. To ensure the privacy and security of patient files, the decision map recommender system selects the blockchain platforms that meet these specific criteria. Therefore, the selection process excludes platforms that are based on permissionless blockchain type, which is open to the public and may compromise data confidentiality. Instead, the recommender system prioritizes platforms that are suited for generic application domains and financial services, as per our taxonomy. In addition, the recommender system focuses on platforms that support the development of smart contracts.

After identifying the relevant blockchain features for health care insurance fraud detection, the recommender system initiates a mapping process to match each feature with the platforms that support it. Figure 5 illustrates the outcomes of this mapping. Initially, after organizing the blockchain features into categories, the recommender system proceeds to map each feature with its corresponding functionality. Next, the recommender system maps the features to the blockchain platforms. On the basis of this, it determines the suitability of each platform. Only the platforms that have all the compulsory features are considered suitable. As shown in Figure 5, platforms R3 Corda and BigchainDB were eliminated from consideration due to their lack of some of the compulsory features. Our mapping process revealed that Hyperledger Fabric [27] was the most optimal platform, followed by Neo [28], XinFin XDC [29], Quorum [30], and Ethereum. These results demonstrate the effectiveness of our mapping process in identifying the ideal blockchain platform for this specific use case.

Table 3 streamlines the mapping process for health care insurance fraud detection by listing the top 5 platforms and highlighting the selected features. The table is designed to simplify the decision-making map by providing a concise and easy-to-read format for comparing the features.

Figure 5. Proposed decision-making map for selecting a platform for health care insurance fraud detection. BFT: Byzantine Fault Tolerance; VM: Virtual Machine.

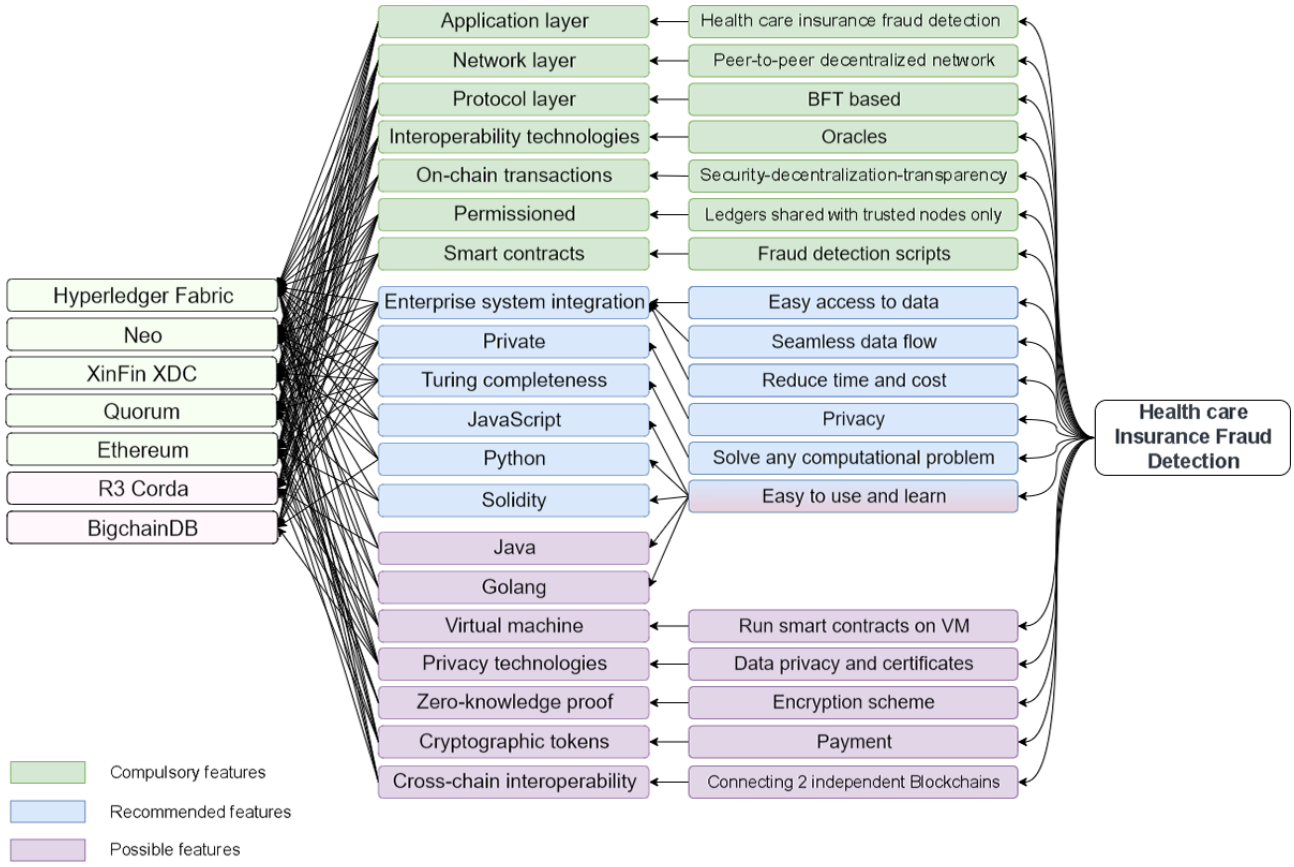

XSL•FO
RenderX

Textbox 1. Definition of compulsory, mandatory, and possible features.

Compulsory features

- Application layer: this capability enables the creation of a user interface and the execution of smart contracts for health care insurance.
- Network layer: enables the establishment of a peer-to-peer decentralized network.
- Protocol layer: enables the selection of a consensus protocol. We used Byzantine-based consensus protocols because they prevent the case of a failing or malicious node [11].
- Interoperability technologies: technologies such as Oracle that facilitate the integration of data from off-chain resources into smart contracts.
- On-chain transaction: the transaction is conducted on the main blockchain for increased security, decentralization, and transparency.
- Permissioned blockchain: this type of blockchain limits access to the ledger to a select group of trusted nodes.
- Smart contracts: enables the development of algorithms that can identify health care insurance fraud.

Mandatory features

- Enterprise system interrogation: provides easy access to data, seamless data flow, and time and cost savings.
- Private: this type of blockchain network is only accessible to authenticated users.
- Turing completeness: the virtual machine of the blockchain platform is capable of solving any computational problem.
- JavaScript, Python, and Solidity: these languages are specifically mentioned because they are intuitive and easily learned by programmers.

Possible features

- Java and Golang: these languages, similar to the 3 mentioned in the mandatory features list, are intuitive and easily learned by programmers.
- Virtual machine: it is used to execute smart contracts.
- Privacy technology: ensures data privacy and certifies the eligibility of peers to participate in the network, particularly when handling sensitive patient data.
- Zero-knowledge proof: this encryption scheme allows one party (the prover) to assure another party (the verifier) that they know a certain value (X) without revealing the value itself.
- Cryptographic token: these tokens have the potential to be used as a means of payment.
- Cross-chain interoperability: this feature enables the connection of 2 separate blockchains to facilitate information exchange.

XSL•FO
RenderX

Table 3. Decision-making map results simplified.

Category and feature name	Hyperledger Fabric	Neo	Ethereum	Quorum	XinFin XDC
Compulsory features					
Application layer	✓	✓	✓	✓	✓
Interoperability technology	✓	✓	✓	✓	✓
Network layer	✓	✓	✓	✓	✓
On-chain transaction	✓	✓	✓	✓	✓
Permissioned blockchain	✓	✓	✓	✓	✓
Protocol layer	✓	✓	✓	✓	✓
Smart contract	✓	✓	✓	✓	✓
Mandatory features					
Enterprise system integration	✓	✓	✓	✓	✓
JavaScript	✓	✓			✓
Private	✓	✓	✓	✓	✓
Python	✓	✓			✓
Solidity	✓		✓	✓	
Turing completeness	✓	✓	✓	✓	✓
Possible features					
Zero-knowledge proof	✓		✓	✓	
Virtual machine	✓	✓	✓	✓	✓
Java	✓	✓			
Golang	✓	✓		✓	
Cryptographic token	✓	✓	✓		✓
Cross-chain interoperability	✓	✓	✓		✓
Privacy technology	✓	✓	✓	✓	✓

Use Case Diagram for the Recommender System

Figure 6 illustrates the use case diagram of our recommender system. Users can perform actions such as adding, editing, and deleting blockchain platforms and features. Following that, they are required to select their desired features, categorize them, and assign weights to mandatory and possible features. Ultimately, users will receive the outcome of the most suitable blockchain platform for their specific use case based on the chosen features.

XSL•FO
RenderX

Figure 6. Recommender system use case diagram.

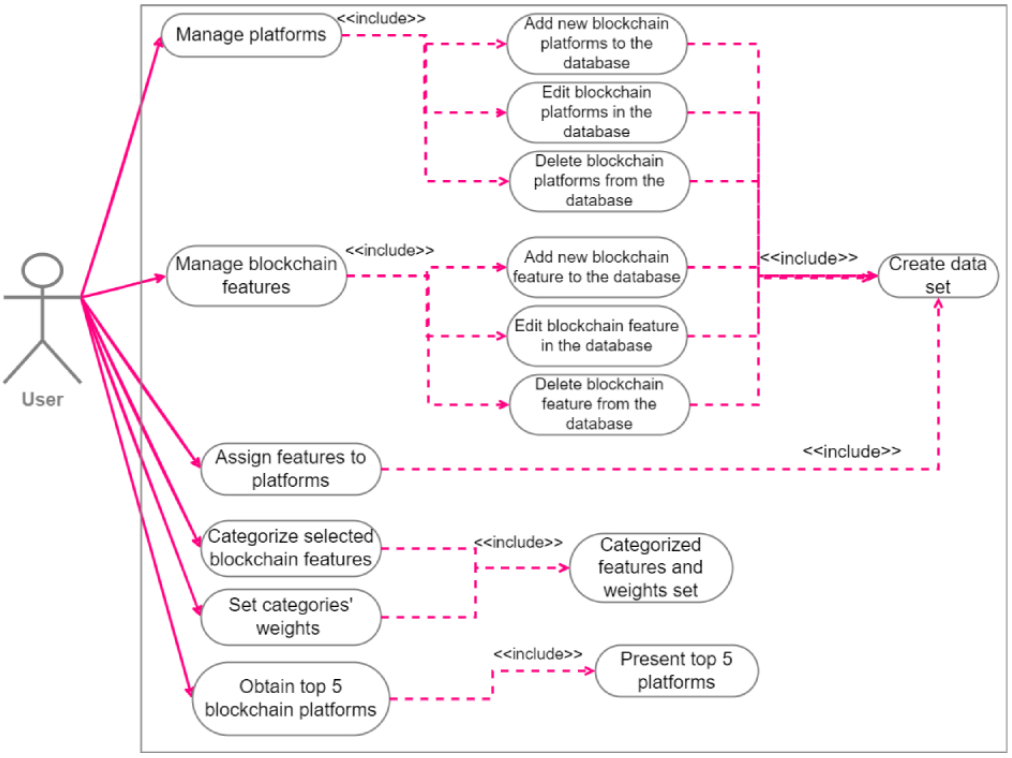

Decision-Making Map Recommender System Implementation

In this subsection, we present our implementation of the decision map recommender system, which is a desktop software solution that provides a streamlined and efficient method to select the most suitable blockchain platform for a specific use case. Our software uses WinForms C# technology (.NET Foundation) and SQL as the database to deliver a user-friendly experience and recommend the top blockchain platforms. To demonstrate the effectiveness of our software, we used it to identify the top 5 blockchain platforms that are most suitable for health care insurance fraud detection.

Textbox 2 defines each function of the recommender system. As previously discussed, the blockchain feature selection process involves dividing features into 3 categories: compulsory, mandatory, and possible. Compulsory features are those that must be present in the blockchain platform for it to be considered. These features are typically critical to the platform's functionality. Mandatory features, on the other hand, are those that are essential for a specific use case or application. They are not necessarily required for the platform to function, but they are necessary for the platform to be suitable for a particular purpose. Finally, possible features are those that provide additional functionality or value to the platform. They are not necessary for the platform to function, but they can enhance its performance or provide additional benefits.

Textbox 2. Decision map recommender system functions and their definitions.

Function and definition

- Create the data set: blockchain platform and blockchain feature names are initially entered. Subsequently, the platforms are associated with their corresponding features.
- Select features and their categories and set weights: specify the category of the blockchain feature by selecting 1 of the 3 options, namely, compulsory, mandatory, or possible. Then, assign the selected feature to the designated category. In addition, assign weights to the mandatory and possible features.
- Obtain top platforms: retrieve blockchain platforms with *compulsory features* and count the number of mandatory and possible features found for each platform. Calculate a score for each platform based on its features and weights and add it to an array. Sort the array based on the calculated score to display the top-performing blockchain platforms.

Creating the Data Set

This section illustrates the data set creation process, including user interactions with the recommender system. The blue annotations in the figures represent instructions. The pink annotations indicate the textboxes for input, buttons for actions, and grid controls for displaying the added platforms and features. In this step, we focused on adding the necessary platforms and blockchain features to build a comprehensive data set. Figure 7 shows the user's process of entering the platform name and selecting *Add Platform* to populate a table showing the added platforms. The same procedure applies to adding blockchain features, resulting in a comprehensive table showcasing both platforms and features. After that, users are

XSL•FO
RenderX

provided with the capability to edit platform names or delete them, as well as modify the names or choose to delete blockchain features. As shown in Figure 8, by double-clicking on a platform and single-clicking on a feature, users can select and make changes to the respective names according to their preferences. After that, users should establish the association between each blockchain platform and its corresponding blockchain features. They can begin by selecting a platform by double-clicking on the row corresponding to the platform name and subsequently choosing the blockchain features that apply to that particular platform, which is done by double-clicking on the rows that correspond to the blockchain features that should be mapped to the selected blockchain platform (Figure 9). Subsequently, a table will be populated with the IDs of the selected blockchain platform, the chosen blockchain feature, and the name of the blockchain feature.

Figure 7. Data set creation—step 1: adding platforms and blockchain features that will be used to create a data set.

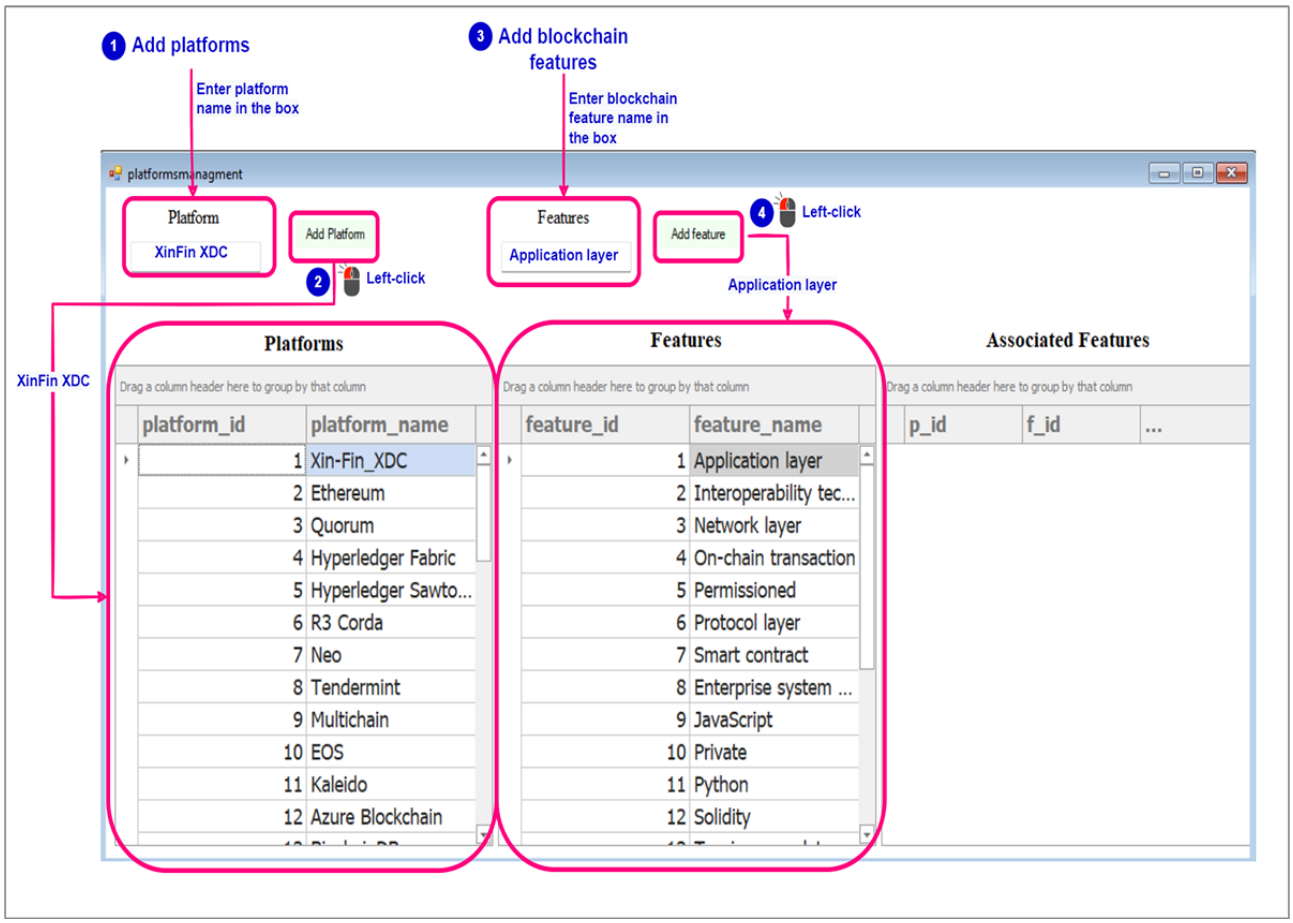

XSL•FO
RenderX

Figure 8. Data set creation—modifying and removing platforms and features from the data set.

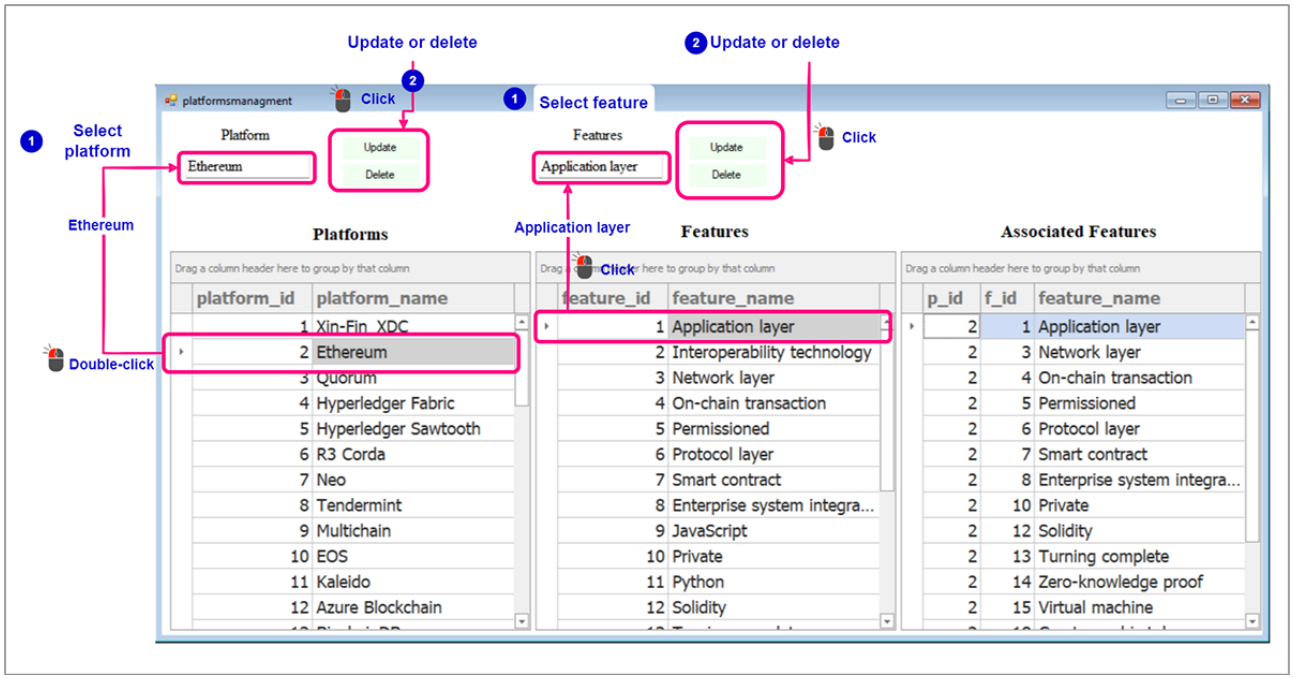

Figure 9. Data set creation—linking specific features with corresponding blockchain platforms.

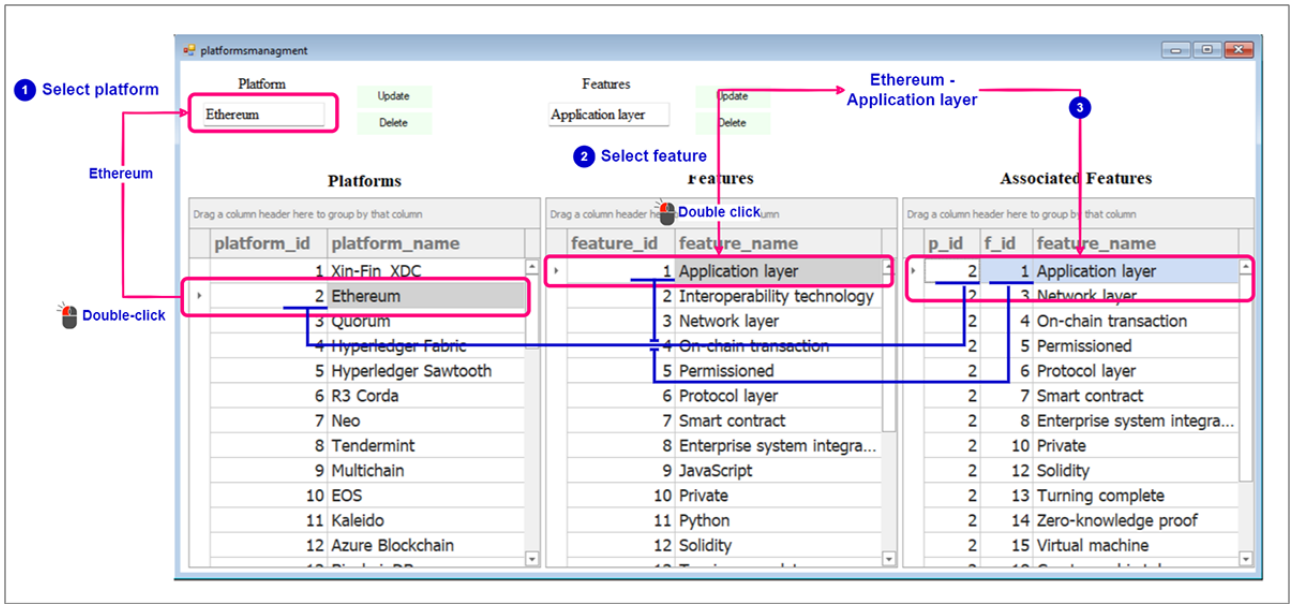

Select Features and Their Categories and Set Weights

Figure 10 shows a screenshot illustrating user interaction during the process of selecting features, assigning them to their respective categories, and assigning weights to those categories. In the initial step, users will choose a category, followed by selecting the desired feature to be assigned to that category. This selection process involves double-clicking on the feature name in the table. Subsequently, a table will display the categorized features, providing a clear overview of the features that have been assigned to their respective categories. Once the categorization of features is complete, users can proceed to set the weights for the mandatory and possible feature categories. Afterward, by clicking on the *Get Platforms* button, users can view the resulting platforms based on the assigned weights and feature categorization.

XSL•FO
RenderX

Figure 10. Allowing for specification of preferred blockchain features followed by categorization and weight assignment.

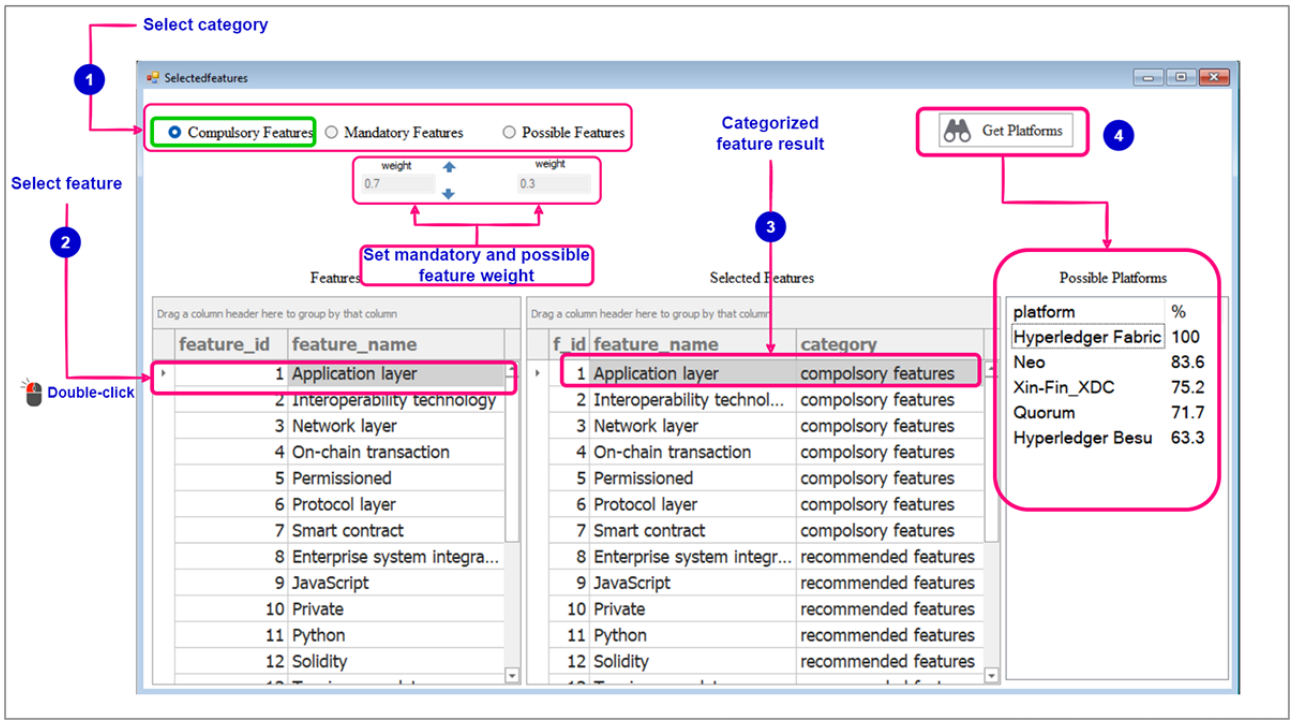

Obtaining the Top Platforms

Figure 11A and B show the sequence diagram to obtain the suitability percentage of each possible blockchain platform. The initial step involves creating a list of blockchain platforms that meet the requirements of the compulsory features. Once that is done, we determine the total number of mandatory and possible features that have been chosen (Figure 11A).

After that, we iterate through the list, and for each platform, we calculate the number of mandatory and possible features (Figure 11B). Finally, using equation 1, we calculate the suitability percentage of each platform (ρ). The first part of the formula calculates the contribution of the mandatory features to the suitability percentage. It takes the number of mandatory features found for the platform (M_{found}), multiplies it by 100 to convert it to a percentage, and then divides it by the number of mandatory features selected (M_{total}) multiplied by the weight assigned to mandatory features (ω_M), which is 0.7.

The second part of the formula calculates the contribution of the possible features to the suitability percentage. It takes the number of possible features found for the platform (P_{found}), multiplies it by 100 to convert it to a percentage, and then divides it by the number of possible features selected (P_{total}) multiplied by the weight assigned to possible features (ω_P), which is 0.3.

By combining these 2 contributions, the suitability percentage provides an overall assessment of how well a blockchain platform meets the selected features, with a higher percentage indicating a better match.

(1)

Once we have calculated the suitability percentage (ρ) for each platform, we sort the list of platforms in descending order based on their scores. Figure 12 shows the flowchart of the recommender system's algorithm, which consists of the different functions involved along with their corresponding input and output parameters.

The platform with the highest score will be at the top of the list, whereas the one with the lowest score will be at the bottom. Finally, we display the top 5 platforms in the list, which are the ones that have the highest scores and, therefore, are the most suitable based on the selected features.

XSL•FO
RenderX

Figure 11. Diagram to obtain the fitness of the possible blockchain platforms for the use case.

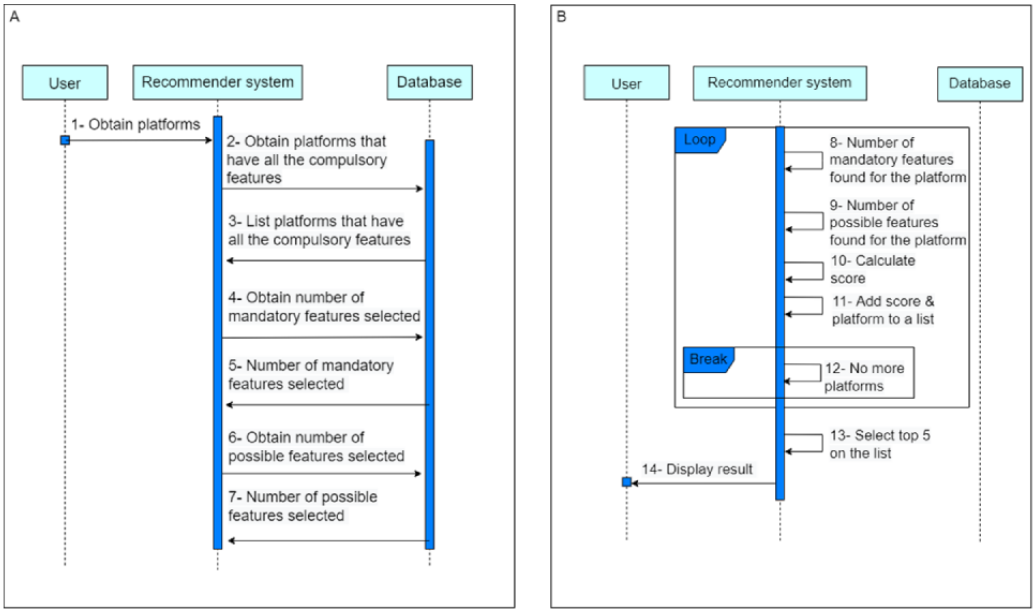

Figure 12. Flowchart to obtain the fitness of the possible blockchain platforms for the use case.

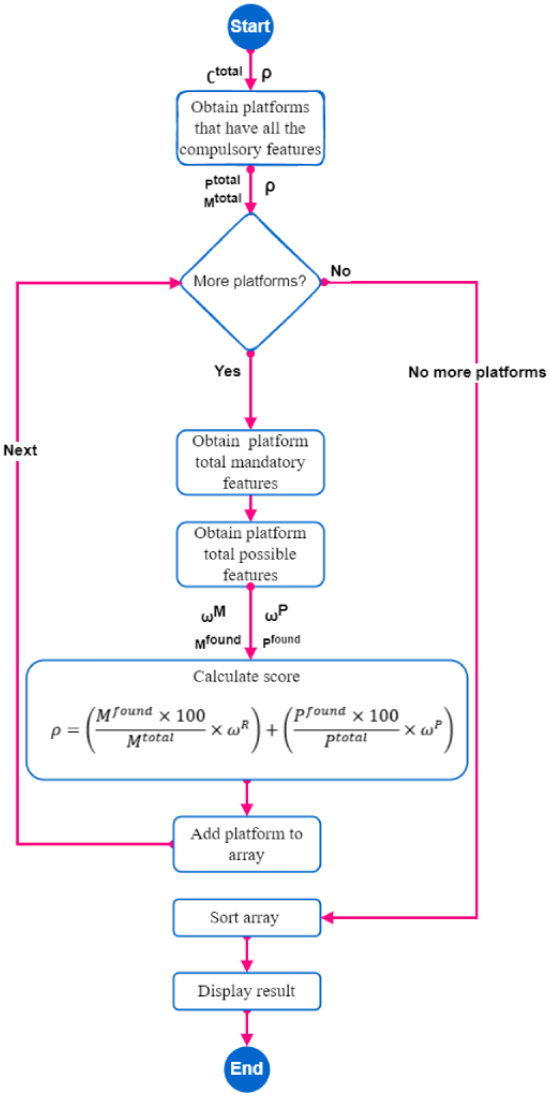

XSL•FO
RenderX

Smart Contracts for Health Care Insurance Fraud Detection

Ismail and Zeadally [22] identified the fraud scenarios used for detecting health care insurance fraud, as shown in Figure 1. The network for detecting health care insurance fraud is made up of 9 participants, as illustrated in Figure 13.

Figure 13. Participants in the health care insurance fraud detection network.

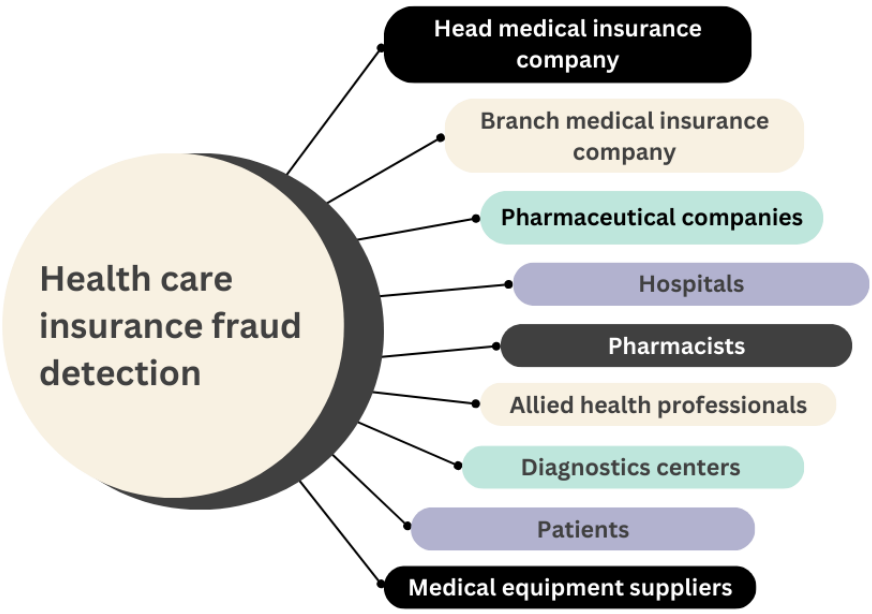

Algorithms for Fraud Scenarios

Overview

For certain fraud scenarios, we need to discover a detectable pattern, whereas for others, data from off-chain sources may be required. The required data for processing claims consist of detailed records of patient visits, including the dates in which they occurred, the departments involved, the services rendered, and patients' information. Consequently, they are on-chain. However, documentation of billed services, detailed service invoices, and pharmacy records are off-chain in the database.

3 Referral Fraud Scenarios

As shown in Figure 14, we use an algorithm to recognize 3 fraud scenarios that have the same pattern—the referral. We then check for the first scenario, in which the fraudster refers patients within the same health care organization. If this is confirmed, the fraud type is *pinning the system*. If not, we investigate whether a financial relationship exists between the fraudster and the other organizations. If such a relationship is detected, it is self-referral fraud. If no financial relationship is found, we investigate whether the fraudster received a commission from the organization; if so, it is a commission-based fraud.

XSL•FO
RenderX

Figure 14. The 3 fraud scenarios related to referrals.

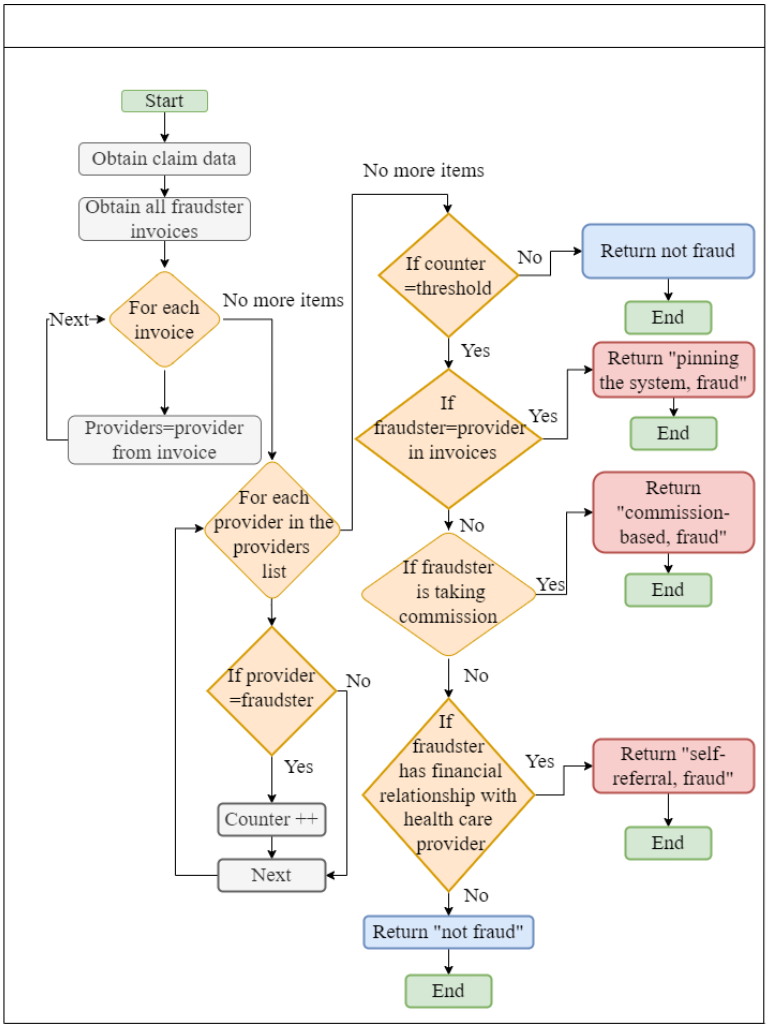

Commission-Based Fraud

In this fraud scenario (Figure 15A), we obtain all the medication that the possible fraudster has prescribed. We then check whether a specific medication is prescribed more frequently and determine whether the fraudster is receiving a commission. If that is the case, it is a fraud. In Figure 15B, we should obtain a list from the minister of health containing the approved and labeled drugs and then compare it to the ones prescribed by the fraudster; if we find a drug that does not exist on the list from the minister of health, there is a fraud.

XSL•FO
RenderX

Figure 15. (A) Providing specific brands of medicines to receive a commission from the pharmaceutical company. (B) A pharmaceutical company provides incentives to physicians to promote unapproved or off-label drugs.

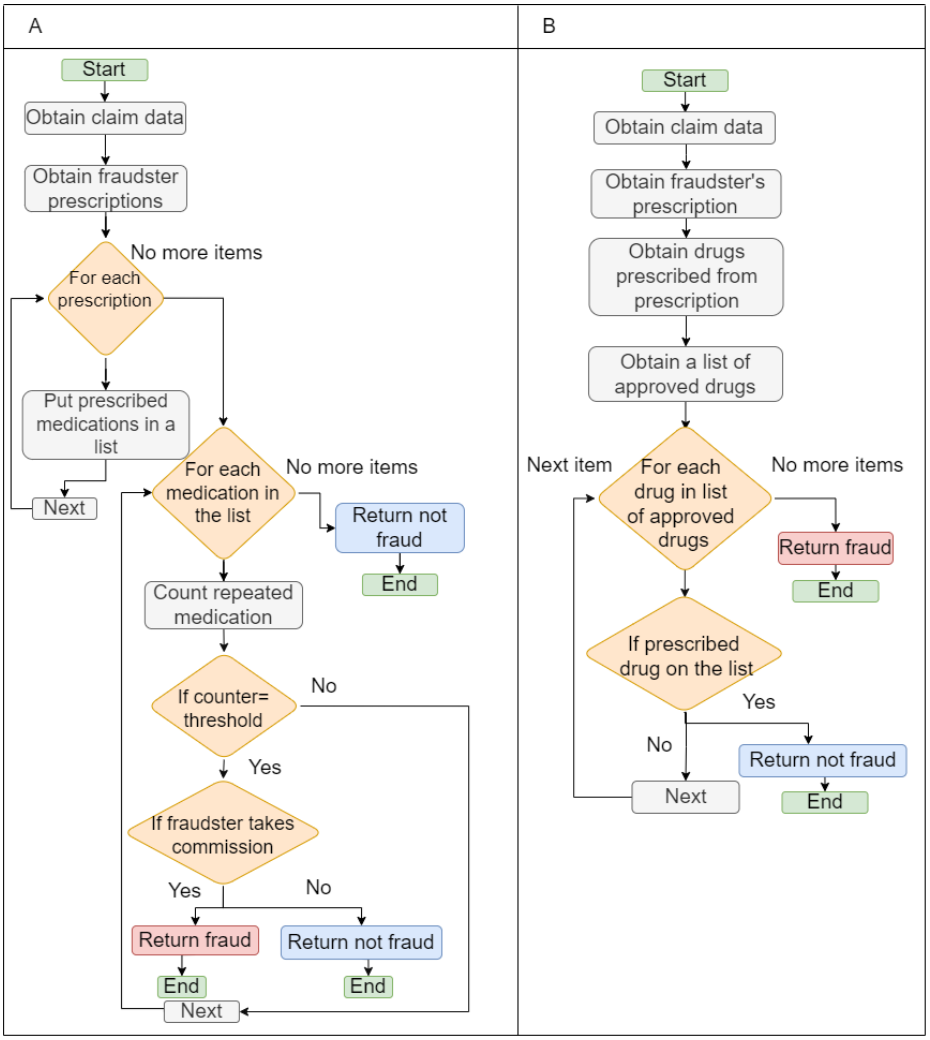

Managed Care and Waiving Copayment Fraud

In Figure 16A, we investigate whether other patients on that date received the same service that the patient requested, and if the number of patients reaches a certain threshold, we are able to demonstrate that the managed care scenario is occurring.

The code for detecting waiving copayment fraud in Figure 16B involves comparing the price listed in the claim with the price mentioned in the corresponding invoice. If there is a mismatch between the 2 prices, it indicates a potential instance of waiving copayment fraud. The code performs a comparison operation to check whether the claim price and the invoice price are equal. If they are not, it raises an alert or triggers further actions to investigate the possibility of fraud.

XSL•FO
RenderX

Figure 16. (A) Managed care and (B) waiving copayment fraud.

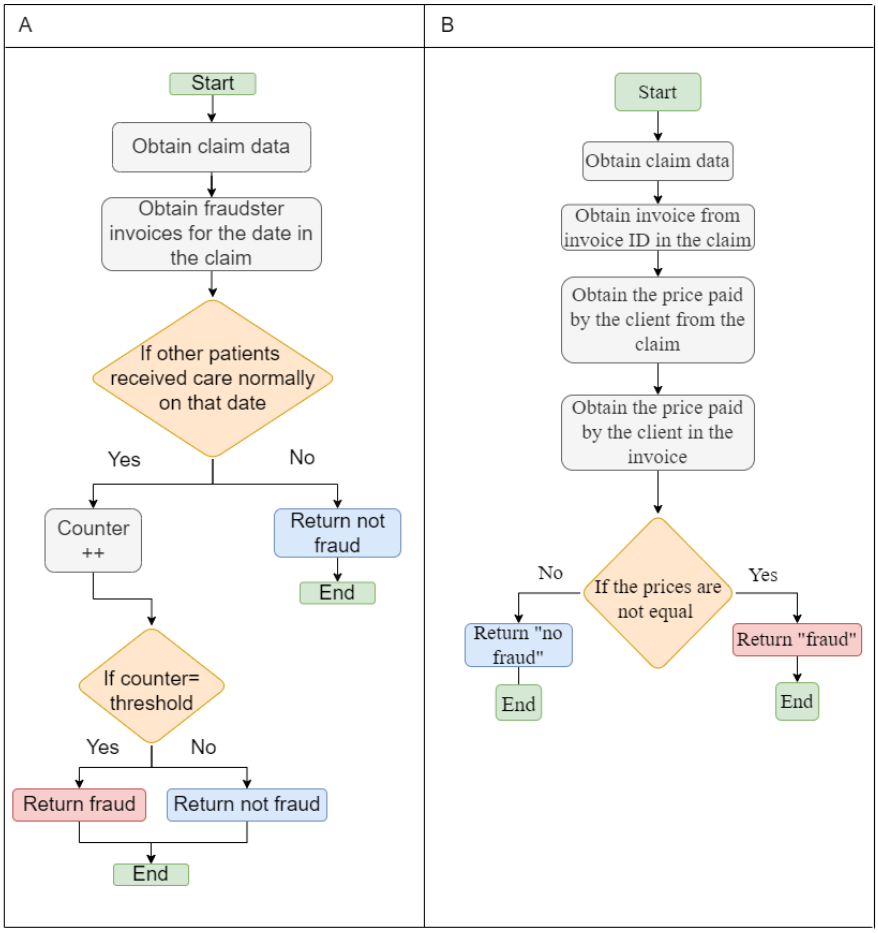

Billing Manipulation Fraud

In Figure 17A, we obtain a list of licensed health care providers to determine whether the suspected fraudster is listed. If not, there is a case of fraud. In Figure 17B, we compare the diagnosis code on the claim to the one on the patient files; if they do not match, we will assume fraud.

In this fraud scenario (Figure 18A), we may require the opinion of another physician, so after determining whether the claim has been duplicated, we gather all the necessary data to be reviewed by another physician, and based on the physician's response, we determine whether the claim is fraudulent. In Figure 18B, we obtain a price list from other equipment suppliers and then compare it to the price paid by the patient; if it is higher than the price on the price list, the patient paid more than necessary for the equipment, and hence, it is a fraud.

XSL•FO
RenderX

Figure 17. (A) Billing patients for care provided by an unlicensed care provider. (B) Manipulation of diagnosis in the claims without the knowledge of patients.

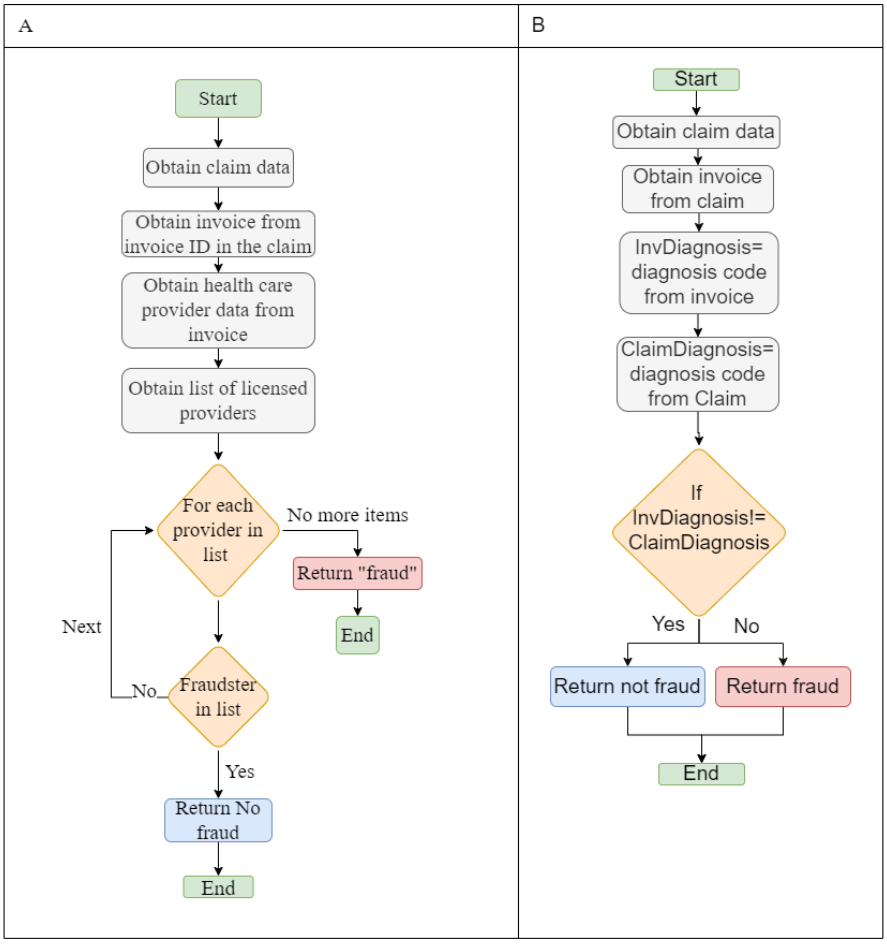

XSL•FO
RenderX

Figure 18. (A) Providing unwanted care to patients, increasing service hours in the bill, duplicating claims, phantom billing, or replacing codes of diseases with ones with higher prices. (B) Billing manipulation in equipment prices.

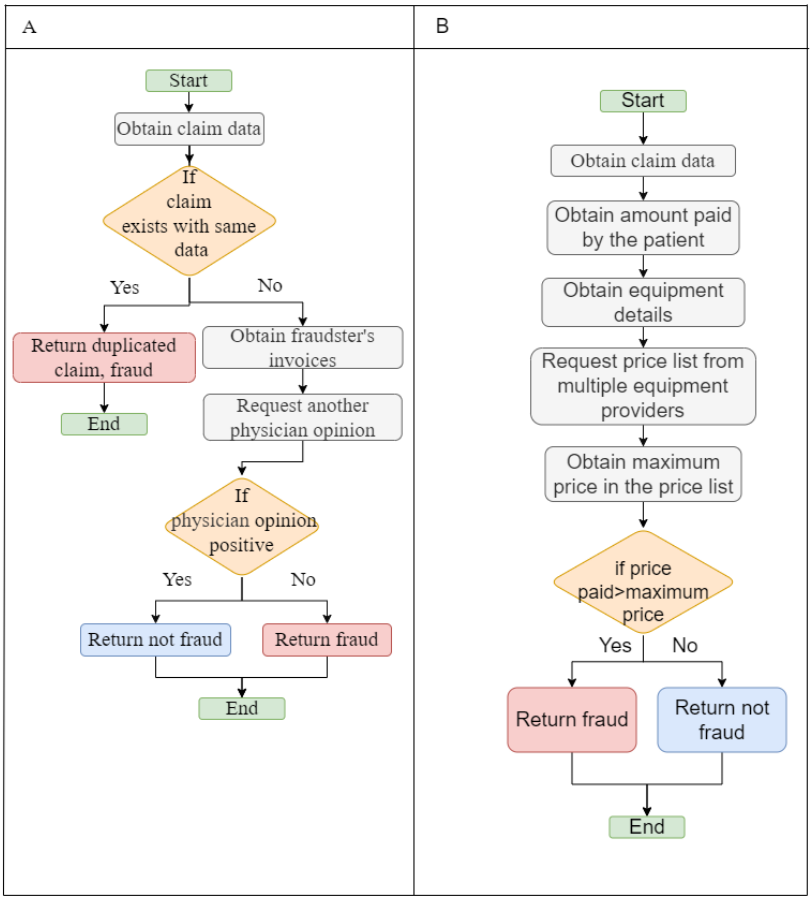

Physician Shopping Fraud

Figure 19 illustrates physician shopping fraud, in which an addicted individual visits multiple health care providers to obtain unprescribed drugs. To detect this fraud, we must examine 5 invoices. We check whether the patient visits the provider regularly based on the dates from the invoices and whether the visits are not to the same provider. If this is confirmed, it is a case of fraud.

XSL•FO
RenderX

Figure 19. Physician shopping fraud.

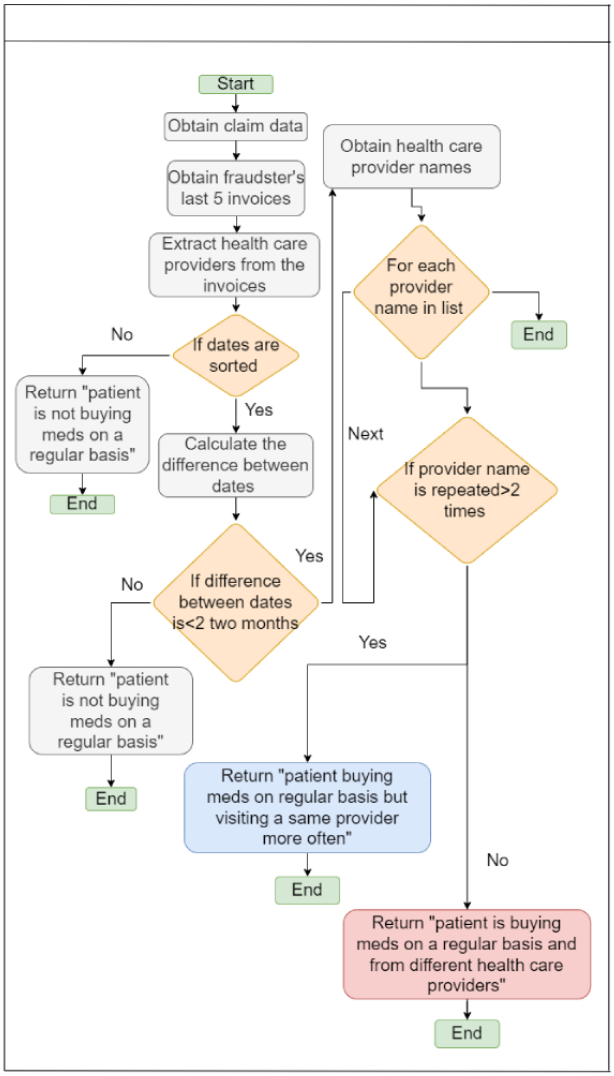

Discussion

Experiments

We created algorithms for smart contracts that address the fraudulent situations mentioned in the study by Ismail and Zeadally [22]. We used the 2 platforms that were selected as the top 2 options according to our decision-making recommender system. Each transaction contains a single claim record and requires supporting files such as invoices for verification. We assessed the performance of the platforms based on several metrics, including throughput, latency, CPU use, and memory use. Transaction throughput reflects the blockchain network's efficiency in terms of the number of transactions processed per second. A transaction is considered successful once it has been included in a block and committed to the ledger. Transaction latency measures the time it takes to send a transaction request and receive a transaction response, indicating the network's responsiveness. CPU and memory use are essential for determining infrastructure requirements and maintaining reliable performance under varying loads [31]. It is essential to ensure that a platform maintains high throughput, low latency, and minimal CPU and memory use [32]. This means that quality of service is maintained, and consequently, fraud is detected more quickly; costs are reduced; and, in some cases, patient lives are saved [33]. To evaluate the platforms' effectiveness, we established 2 testing scenarios. The first scenario involved peers sending a consistent number of transactions over a period ranging from 30 to 120 seconds. In the second scenario, we progressively increased the number of transactions transmitted over the network from 1000 to 10,000.

Experimental Environment

Our experimental setup involved using Ubuntu Windows Subsystem for Linux 2 (version 20.04; Canonical Ltd) on Windows 11 operating systems (Microsoft Corp). In the case of Fabric, we used Docker (Docker, Inc) to run the platform, and all peers were connected to a single channel representing an insurance company. The Fabric version we used was 2.2, with 4 organizations and 1 orderer. A batch size of 500 was set for processing transactions. We used Golang as the programming language for developing smart contracts.

For the Neo private blockchain, we used the N3 Neo Visual DevTracker extension on Visual Studio Code (Microsoft Corp). The programming language recommended in the Neo documentation, which is C#, was used. Similar to Fabric, a batch size of 500 was used. Both platforms used LevelDB as the key-value data storage.

XSL•FO
RenderX

In terms of hardware, we used a system with 16 GB of RAM and an 11th-generation Intel Core i7 processor running at 2.80 GHz. To conduct the benchmarking, we used Hyperledger Caliper [34] for Fabric and Neo-bench [35] for Neo.

Security and Privacy Concerns

Health care data comprise sensitive patient information, making their security and privacy crucial. These data are typically stored in local or cloud databases. However, in such architecture, data face several issues regarding access and are subject to be deleted or modified and to cybersecurity attacks [14].

Using a blockchain-based approach addresses these concerns. Blockchain stores data in an immutable ledger where data can only be added after reaching a consensus and cannot be altered or modified. Changing a block or transaction impacts all subsequent blocks, and revalidating all subsequent blocks requires enormous computational power, making it nearly impossible for a malicious node. Furthermore, access control rights in blockchain can be defined in a smart contract, ensuring trustless and secure data access for network participants. Textbox 3 provides a concise overview of the challenges encountered in local on-premise or cloud database systems and demonstrates how blockchain technology addresses these issues.

Textbox 3. Comparison of issues in local on-premise or cloud databases versus blockchain for managing medical data.

Local on-premise or cloud database

- Traditional record-keeping methods may not provide reliable auditing.
- Locally stored medical data might be unavailable in critical real-time situations.
- Local on-premise or cloud databases often lack proper authentication, leading to potential misuse of medical information.
- Local on-premise or cloud databases typically have less secure data access control.
- In a local on-premise or cloud database, unauthorized users can impersonate legal users to access sensitive medical data.
- Medical data in local databases can be easily altered or deleted.
- Without robust security, users could deny accessing or modifying data in local on-premise or cloud database systems.

Blockchain

- Blockchain's replicated, time-stamped ledger facilitates efficient and trusted auditing.
- Data replication ensures real-time availability from the local copy of the ledger.
- Blockchain's encryption and digital signature techniques ensure user authenticity for accessing and uploading medical information.
- Access control rights can be defined in smart contracts, ensuring secure and trustless data access for network participants.
- The private blockchain network restricts data access to authorized participants based on access control rights.
- Medical data are stored as transactions in blocks linked cryptographically to ensure immutability.
- Each operation's authenticity is recorded in an immutable ledger, preventing repudiation.

Results Analysis

Regarding the performance of Neo and Fabric in terms of throughput during the first test scenario, Fabric had 789 transaction per second (TPS), 445 TPS, 409 TPS, and 329 TPS, and Neo had 438 TPS, 629 TPS, 304 TPS, and 329 TPS during 30-second interval, 60-second interval, 90-second interval, and 120-second interval respectively. It is interesting to note that Neo initially experienced an increase in throughput, reaching a peak of 629 transactions per second (TPS) during the 60-second interval. However, it subsequently declines to 304 TPS and remains at that level for the remainder of the test. Similarly, Fabric follows a comparable pattern, starting with a high throughput of approximately 800 TPS and gradually decreasing to around approximately 450 TPS by the end of the test. This indicates that both platforms exhibited fluctuations in their transaction processing speeds throughout the test.

Regarding latency, it is worth noting that Neo consistently took approximately 14 to 24 seconds to confirm a transaction throughout the test duration.

The latency of Fabric in the first test scenario was 0.05 seconds, 0.12 seconds, 0.1 seconds, and 0.1 seconds during 30-second interval, 60-second interval, 90-second interval, and 120-second interval respectively. Here, we can observe that, as more transactions were submitted, the latency slightly increased. This implies that, as the workload on Fabric intensified with a higher number of transactions, the time taken to process and confirm each transaction also increased.

Regarding the throughput in the second test scenario, in which we increased the number of transactions sent over the network by 1000 up to 10,000 transactions. Fabric had 426 TPS, 470 TPS, 630 TPS, 577 TPS, 676 TPS, 707 TPS, 718 TPS, 732 TPS, 740 TPS, and 754 TPS. Neo had 373 TPS, 427 TPS, 489 TPS, 422 TPS, 534 TPS, 543 TPS, 434 TPS, 424 TPS, 409 TPS, and 474 TPS. Fabric outperformed Neo once again. Fabric demonstrated an upward trend in throughput as the number of transactions increased, whereas Neo exhibited more fluctuations in its performance.

Regarding the latency in the second test scenario, Fabric had 0.03 seconds for 1000 TX sent, 0.04 for 2000 TX, 3000 TX, 4000 TX and 5000 TX, 0.03 seconds for 6000 TX, 7000 TX,

XSL•FO
RenderX

8000 TX and 9000 TX, and 0.04 for 10,000 TX. The latency for Fabric remained relatively stable, with a slight increase observed. The latency values consistently ranged between 0.03 and 0.04 seconds throughout the test. This indicates that Fabric can maintain low and consistent latency even as the number of transactions increases.

Delays in fraud detection can slow the identification of fraudulent activities, causing financial losses and putting patients' health at risk. High latency impedes the prompt discovery of fraud, giving wrongdoers the chance to persist in their schemes unchecked, possibly resulting in more harm and greater financial damage. Thus, reducing latency is essential to improve the efficiency and precision of fraud detection, ultimately protecting health care resources, patients' health, and the credibility of health care services [36]. Throughput and latency can be significantly impacted by large transaction sizes, which are driven by extensive file requirements and block size. In private blockchain networks, the computational complexity and energy consumption of encryption and decryption operations add to this burden. Furthermore, replicating the ledger across all nodes increases computational and network overhead, resulting in high energy consumption, low transaction throughput, and limited scalability. As the number of nodes increases, so does the volume of data transferred, leading to longer processing times. In addition, the choice of consensus mechanism affects scalability; for instance, proof of work is particularly known for its high energy consumption, further exacerbating these challenges.

The CPU and memory use comparison between Fabric and Neo reveals that Fabric used fewer resources than Neo. Fabric utilized 191 MB of memory and 47% of the CPU, while Neo used 515 MB of memory and 39% of the CPU. Throughout both conducted tests, both platforms achieved a 100% success rate, indicating their reliability for securely sharing sensitive health care insurance data. Fabric consistently outperformed Neo in both tests, showcasing its superiority in creating health care insurance fraud detection systems. The lower resource consumption by Fabric suggests that it offers more efficient resource use, making it an optimal choice for health care insurance fraud detection applications.

Limitations

Validating the system using real-world health care data presents several challenges and ethical considerations. Obtaining access to real-world health care data can be difficult due to stringent regulations and privacy concerns. Ensuring that the data are accurate, complete, and representative of the broader population can be challenging as inconsistent or incomplete data can affect the validity of the results. Integrating diverse data sources and formats into a cohesive system requires significant technical expertise and resources. From an ethical standpoint, protecting patient confidentiality is paramount, necessitating robust measures to ensure that data are anonymized and secure. In addition, there is a risk of introducing bias if the data are not representative of the entire population, potentially leading to skewed results and harmful recommendations.

In addition to the technical and ethical considerations, the successful implementation of this system requires the cooperation and acceptance of various participants in the health care system [12]. Key stakeholders such as hospitals, clinics, insurance companies, and health care providers must be willing to contribute their data and support the integration of a blockchain-based solution. Each participant has unique requirements for data security, privacy, and interoperability that must be addressed to ensure their cooperation. Furthermore, patient data must be governed and secured to ensure privacy and controlled access. Only authorized personnel should be able to access sensitive information, and all data handling procedures should comply with relevant regulations [37].

Conclusions and Future Work

Health care insurance fraud detection is crucial for the health care industry. This is due to the high costs incurred from health care insurance fraud. In addition, some frauds pose risks to patient health. In this study, we designed and implemented smart contracts to detect health care insurance fraud. This is based on our proposed taxonomy of fraud scenarios. Furthermore, we used a blockchain platform that is specifically suited for health care insurance fraud detection. To improve the selection of a suitable platform, we designed and implemented a decision map–based recommender system, which automates and streamlines the platform selection process. To feed the recommender system with suitable candidates, we proposed a taxonomy of 102 blockchain development platforms. Through these efforts, we aimed to improve the efficiency and accuracy of health care insurance fraud detection by leveraging the capabilities of blockchain technology. The recommender system revealed Fabric and Neo as the top 2 platform candidates for the development of health care insurance fraud detection solutions, with Fabric having the highest rank. Our experimental numerical evaluation of the 2 selected platforms showed that Fabric outperformed Neo, demonstrating a more suitable network structure and features than Neo. Furthermore, based on our experiments, Fabric offers greater configurability, enabling further performance improvements. Machine and deep learning algorithms are alternative promising approaches for detecting patterns of fraud in large-scale environments such as health care insurance. However, these algorithms suffer from security and privacy issues and are prone to bias. In future work, we aim to integrate machine learning techniques for health care insurance fraud detection into blockchain. This integration would provide trust in the machine learning techniques, and aid their traceability and understanding, leading to privacy-preserving machine learning models. In particular, we will explore how blockchain can improve the privacy and security of machine learning models, investigate the most effective ways to integrate federated learning with blockchain to ensure that data remain decentralized and secure, and design smart contracts to automate and verify the training and deployment of machine learning models.

XSL•FO
RenderX

Acknowledgments

We would like to thank the United Arab Emirates University for supporting this work.

Conflicts of Interest

None declared.

References

1. Home page. National Health Care Anti-Fraud Association (NHCAA). URL: https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/ [accessed 2024-04-29]
2. Home page. U.S. Department of Justice. URL: https://www.justice.gov/ [accessed 2024-04-29]
3. Healthcare fraud analytics global market report. ReportLinker. 2023. URL: https://www.reportlinker.com/p06318507/Healthcare-Fraud-Analytics-Global-Market-Report.html?utm_source=GNW [accessed 2024-04-29]
4. Alanazi A. Using machine learning for healthcare challenges and opportunities. Inform Med Unlocked 2022;30:100924. [doi: 10.1016/j.imu.2022.100924]
5. Gupta RY, Mudigonda SS, Baruah PK. A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes. Int J Eng Trends Technol 2021 Mar 25;69(3):96-102. [doi: 10.14445/22315381/ijett-v69i3p216]
6. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021 Jul 20;4(1):123-144 [FREE Full text] [doi: 10.1146/annurev-biodatasci-092820-114757] [Medline: 34396058]
7. Kapadiya K, Patel U, Gupta R, Alshehri MD, Tanwar S, Sharma G, et al. Blockchain and AI-empowered healthcare insurance fraud detection: an analysis, architecture, and future prospects. IEEE Access 2022;10:79606-79627. [doi: 10.1109/access.2022.3194569]
8. Ding J, Li X. An approach for validating quality of datasets for machine learning. In: Proceedings of the 2018 IEEE International Conference on Big Data. 2018 Presented at: BigData '18; December 10-13, 2018; Seattle, WA p. 2795-2803 URL: https://ieeexplore.ieee.org/document/8622640 [doi: 10.1109/bigdata.2018.8622640]
9. Ismail L, Materwala H, Khan MA. Performance evaluation of a patient-centric blockchain-based healthcare records management framework. In: Proceedings of the 2nd International Electronics Communication Conference. 2020 Presented at: IECC '20; July 8-10, 2020; Singapore, Singapore p. 39-50 URL: https://dl.acm.org/doi/abs/10.1145/3409934.3409941 [doi: 10.1145/3409934.3409941]
10. Ismail L, Materwala H. BlockHR: a blockchain-based framework for health records management. In: Proceedings of the 12th International Conference on Computer Modeling and Simulation. 2020 Presented at: ICCMS '20; June 22-24, 2020; Brisbane, Australia p. 164-168 URL: https://dl.acm.org/doi/abs/10.1145/3408066.3408106 [doi: 10.1145/3408066.3408106]
11. Ismail L, Materwala H. A review of blockchain architecture and consensus protocols: use cases, challenges, and solutions. Symmetry 2019 Sep 24;11(10):1198. [doi: 10.3390/sym11101198]
12. Ismail L, Materwala H, Karduck AP, Adem A. Requirements of health data management systems for biomedical care and research: scoping review. J Med Internet Res 2020 Jul 07;22(7):e17508 [FREE Full text] [doi: 10.2196/17508] [Medline: 32348265]
13. Ismail L, Materwala H, Sharaf Y. BlockHR – a blockchain-based healthcare records management framework: performance evaluation and comparison with client/server architecture. In: Proceedings of the 2020 International Symposium on Networks, Computers and Communications. 2020 Presented at: ISNCC '20; October 20-22, 2020; Montreal, QC p. 1-8 URL: https://ieeexplore.ieee.org/document/9297216 [doi: 10.1109/isncc49221.2020.9297216]
14. Ismail L, Materwala H, Hennebelle A. A scoping review of integrated blockchain-cloud (BCC) architecture for healthcare: applications, challenges and solutions. Sensors (Basel) 2021 May 28;21(11):3753 [FREE Full text] [doi: 10.3390/s21113753] [Medline: 34071449]
15. Farshidi S, Jansen S, Espana S, Verkleij J. Decision support for blockchain platform selection: three industry case studies. IEEE Trans Eng Manage 2020 Nov;67(4):1109-1128. [doi: 10.1109/tem.2019.2956897]
16. Lu J, Lin K, Chen R, Lin M, Chen X, Lu P. Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism. BMC Med Inform Decis Mak 2023 Apr 06;23(1):62 [FREE Full text] [doi: 10.1186/s12911-023-02152-0] [Medline: 37024897]
17. Sowah RA, Kuuboore M, Ofoli A, Kwofie S, Asiedu L, Koumadi KM, et al. Decision support system (DSS) for fraud detection in health insurance claims using Genetic Support Vector Machines (GSVMs). J Eng 2019 Sep 02;2019:1-19. [doi: 10.1155/2019/1432597]
18. Settipalli L, Gangadharan G. WMTDBC: an unsupervised multivariate analysis model for fraud detection in health insurance claims. Expert Syst Appl 2023 Apr;215:119259. [doi: 10.1016/j.eswa.2022.119259]
19. Mackey TK, Miyachi K, Fung D, Qian S, Short J. Combating health care fraud and abuse: conceptualization and prototyping study of a blockchain antifraud framework. J Med Internet Res 2020 Sep 10;22(9):e18623 [FREE Full text] [doi: 10.2196/18623] [Medline: 32909952]

XSL•FO
RenderX

20. Saldamli G, Reddy V, Bojja KS, Gururaja MK, Doddaveerappa Y, Tawalbeh L. Health care insurance fraud detection using blockchain. In: Proceedings of the 7th International Conference on Software Defined Systems. 2020 Presented at: SDS '20; April 20-23, 2020; Paris, France p. 145-152 URL: https://ieeexplore.ieee.org/document/9143900 [doi: 10.1109/sds49854.2020.9143900]
21. Liu W, Yu Q, Li Z, Li Z, Su Y, Zhou J. A blockchain-based system for anti-fraud of healthcare insurance. In: Proceedings of the 5th International Conference on Computer and Communications. 2019 Presented at: ICCCSP '19; December 6-9, 2019; Chengdu, China p. 1264-1268 URL: https://ieeexplore.ieee.org/abstract/document/9064274 [doi: 10.1109/iccc47050.2019.9064274]
22. Ismail L, Zeadally S. Healthcare insurance frauds: taxonomy and blockchain-based detection framework (Block-HI). IT Prof 2021 Jul;23(4):36-43. [doi: 10.1109/mitp.2021.3071534]
23. Ismail L, Hameed H, AlShamsi M, AlHammadi M, AlDhanhani N. Towards a blockchain deployment at UAE university: performance evaluation and blockchain taxonomy. In: Proceedings of the 2019 International Conference on Blockchain Technology. 2019 Presented at: ICBCT '19; March 15-18, 2019; Honolulu, HI p. 30-38 URL: https://dl.acm.org/doi/abs/10.1145/3320154.3320156 [doi: 10.1145/3320154.3320156]
24. Kaafarani R, Ismail L, Zahwe O. An adaptive decision-making approach for better selection of blockchain platform for health insurance frauds detection with smart contracts: development and performance evaluation. Procedia Comput Sci 2023;220:470-477. [doi: 10.1016/j.procs.2023.03.060]
25. Bitcoin: a peer-to-peer electronic cash system. United States Sentencing Commission. URL: https://www.ussc.gov/sites/default/files/pdf/training/annual-national-training-seminar/2018/Emerging_Tech_Bitcoin_Crypto.pdf [accessed 2024-04-29]
26. Home page. Ethereum. URL: https://ethereum.org/en/whitepaper/ [accessed 2024-04-29]
27. Home page. Hyperledger Fabric. URL: https://www.hyperledger.org/wp-content/uploads/2020/03/hyperledger_fabric_whitepaper.pdf [accessed 2024-04-29]
28. Home page. NEO. URL: https://neo.org/ [accessed 2024-04-29]
29. XinFin Network-XDC consensus algorithm white paper (updated). XinFin Hybrid Blockchain Network. 2021. URL: https://xinfin.org/docs/whitepaper-tech.pdf [accessed 2024-04-29]
30. Home page. Quorum Blockchain Service. URL: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RWRLFs [accessed 2024-04-29]
31. Mazzoni M, Corradi A, Di Nicola V. Performance evaluation of permissioned blockchains for financial applications: the ConsenSys Quorum case study. Blockchain Res Appl 2022 Mar;3(1):100026. [doi: 10.1016/j.bcra.2021.100026]
32. Ismail L, Mills B, Hennebelle A. A formal model of dynamic resource allocation in grid computing environment. In: Proceedings of the 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. 2008 Presented at: SNPD '08; August 6-8, 2008; Phuket, Thailand p. 685-693 URL: https://ieeexplore.ieee.org/document/4617452 [doi: 10.1109/snpd.2008.136]
33. Buyya R, Srirama SN, Mahmud R, Goudarzi M, Ismail L, Kostakos V. Quality of service (QoS)-driven edge computing and smart hospitals: a vision, architectural elements, and future directions. In: Proceedings of the 2023 NIELIT's International Conference on Communication, Electronics and Digital Technology. 2023 Presented at: NICE-DT '23; July 26-31, 2023; New Delhi, India p. 1-23 URL: https://link.springer.com/chapter/10.1007/978-981-99-1699-3_1 [doi: 10.1007/978-981-99-1699-3_1]
34. Home page. Caliper Hyperledger. URL: https://hyperledger.github.io/caliper/v0.5.0/getting-started/ [accessed 2024-04-29]
35. Neo Blockchain benchmarking tool. GitHub. URL: https://github.com/nspcc-dev/neo-bench [accessed 2024-04-29]
36. Settipalli L, Gangadharan GR. Healthcare fraud detection using primitive sub peer group analysis. Concurr Comput 2021 Mar 18;33(23):e6275. [doi: 10.1002/cpe.6275]
37. Lee HA, Kung HH, Udayasankaran JG, Kijsanayotin B, B Marcelo A, Chao LR, et al. An architecture and management platform for blockchain-based personal health record exchange: development and usability study. J Med Internet Res 2020 Jun 09;22(6):e16748 [FREE Full text] [doi: 10.2196/16748] [Medline: 32515743]

Abbreviations

CPU: Central Processing Unit
TPS: transactions per second

XSL•FO
RenderX

Edited by Y Zhuang; submitted 11.07.23; peer-reviewed by H Subramanian, C Zaki, K Danach; comments to author 21.02.24; revised version received 16.05.24; accepted 01.07.24; published 18.10.24.

Please cite as:
Kaafarani R, Ismail L, Zahwe O
Automatic Recommender System of Development Platforms for Smart Contract–Based Health Care Insurance Fraud Detection Solutions: Taxonomy and Performance Evaluation
J Med Internet Res 2024;26:e50730
URL: https://www.jmir.org/2024/1/e50730
doi:10.2196/50730
PMID:39423005

©Rima Kaafarani, Leila Ismail, Oussama Zahwe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Redefining Health Care Data Interoperability: Empirical Exploration of Large Language Models in Information Exchange

Dukyong Yoon[1,2,3], MD, PhD; Changho Han[1], MD; Dong Won Kim[1], MS; Songsoo Kim[1], MD; SungA Bae[3,4], MD, PhD; Jee An Ryu[1], BS; Yujin Choi[1], BS

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea
[2]Institute for Innovation in Digital Healthcare (IIDH), Severance Hospital, Seoul, Republic of Korea
[3]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea
[4]Department of Cardiology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea

Corresponding Author:
Dukyong Yoon, MD, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
50-1 Yonsei-ro Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 31 5189 8450
Email: dukyong.yoon@yonsei.ac.kr

Abstract

Background: Efficient data exchange and health care interoperability are impeded by medical records often being in nonstandardized or unstructured natural language format. Advanced language models, such as large language models (LLMs), may help overcome current challenges in information exchange.

Objective: This study aims to evaluate the capability of LLMs in transforming and transferring health care data to support interoperability.

Methods: Using data from the Medical Information Mart for Intensive Care III and UK Biobank, the study conducted 3 experiments. Experiment 1 assessed the accuracy of transforming structured laboratory results into unstructured format. Experiment 2 explored the conversion of diagnostic codes between the coding frameworks of the *ICD-9-CM* (*International Classification of Diseases, Ninth Revision, Clinical Modification*), and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) using a traditional mapping table and a text-based approach facilitated by the LLM ChatGPT. Experiment 3 focused on extracting targeted information from unstructured records that included comprehensive clinical information (discharge notes).

Results: The text-based approach showed a high conversion accuracy in transforming laboratory results (experiment 1) and an enhanced consistency in diagnostic code conversion, particularly for frequently used diagnostic names, compared with the traditional mapping approach (experiment 2). In experiment 3, the LLM showed a positive predictive value of 87.2% in extracting generic drug names.

Conclusions: This study highlighted the potential role of LLMs in significantly improving health care data interoperability, demonstrated by their high accuracy and efficiency in data transformation and exchange. The LLMs hold vast potential for enhancing medical data exchange without complex standardization for medical terms and data structure.

(J Med Internet Res 2024;26:e56614) doi:10.2196/56614

KEYWORDS

health care interoperability; large language models; medical data transformation; data standardization; text-based

Introduction

Efficient health care data exchange is essential in medicine, particularly in facilitating continuous care [1]. Such data exchange becomes crucial when a patient uses multiple health care facilities or receives concurrent care, significantly influencing accurate treatment strategies. The emergence of personalized health care, becoming a cornerstone of modern medicine, necessitates the use of personal health records. This shift complicates data exchange processes as it demands the

XSL•FO
RenderX

integration of data from multiple health care institutions, thereby posing substantial challenges [2,3]. Additionally, health care is increasingly including patient-generated health data (PGHD) from a diverse range of devices, including wearable technology, given the heterogeneity of products from different vendors [4-6].

Globally, health care systems contend with varying medical record formats and disparate coding systems. In the globalized health care paradigm, the mobility of patients across international boundaries introduces an added layer of complexity. The necessity for efficiently leveraging consolidated information from multiple nations escalates as international collaborative research broadens [7,8]. The *International Classification of Diseases* (*ICD*) has served as a global standard for diagnostic nomenclature, whereas the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) presents a detailed, structured, and multiaxial medical terminology system, gaining adoption worldwide, including in the United States and Europe. Divergent drug coding systems also continue to exist between the United States and Europe, with the RxNorm system adopted in the United States and the ATC system used across Europe. These discrepancies underscore the urgent need for robust and effective health care data exchange pipelines.

Over the years, significant attempts have been made toward the standardization of health care data amid notable challenges and limitations. Protocols, such as Health Level Seven International and Fast Healthcare Interoperability Resources (FHIR), have been introduced to enhance data exchange between medical devices and electronic health records [9]. However, despite their use, these standards often meet with noncompliance or suboptimal implementation. Specifically, FHIR has received criticism for its inherent complexity, obstructing its widespread adoption [10]. Moreover, a key obstacle in the exchange of health care data lies in the initial state of medical records, many of which are not stored following a universal standard. This inconsistency creates a significant challenge even before leveraging exchange protocols like Health Level Seven International and FHIR, designed to facilitate data sharing. The presence of standards does not automatically solve the issue of initiating the exchange when the starting point involves aligning diverse data formats.

The Observational Health Data Sciences and Informatics initiative represents one of the most robust efforts toward data standardization for research purposes. This initiative has developed a common data model and promoted data standardization across various institutions in accordance with this format, significantly accelerating data analysis across institutions [8,11]. Nonetheless, the standardization process has its limitations. One is a notable risk of information loss from the original data during standardization [12,13]. Despite sustained global efforts to transition data into standardized formats, the inherent challenges of standardization inhibit complete conversion and representation of the finer details in the original data. Therefore, effective data standardization remains a pervasive challenge in health care data exchange.

To address the challenges associated with data standardization, we attempted to explore alternatives beyond traditional approaches. A potential solution might be a system that supports flexible communication of raw data, for example, in natural language, permitting the end user to process and interpret data as required, thereby reducing the necessity for strict standardization. Large language models (LLMs), such as ChatGPT, which are designed to produce contextually relevant and coherent natural language responses based on input data, might be promising tools in this regard. Leveraging the capabilities of LLMs can enhance natural human interaction and streamline the management and summarization of extensive language-based data sets. Multiple studies have reported these potential applications of LLMs in the medical field; for example, mining medical text data for relevant clinical information, summarizing patient records and research findings, inferring medical outcomes from complex case histories, and reviewing medical literature to identify trends and validate clinical practices [14-17]. Consequently, if LLMs can proficiently transcribe patient data into text format and the receiving end can efficiently structure the resultant text data, then the intricate stages of data standardization may become redundant. This paradigm shift could significantly alter health care data exchange, heralding a future of seamless and universal data interoperability.

This study tests the hypothesis that text-based conversion and integration of hospital data in different databases would be more effective than current methods. To prove this, we focused on 3 aspects: accuracy of numerical data transformation into text and back, fidelity of text-based transformation for semantic data using *ICD* codes (ie, *ICD-9-CM* [*International Classification of Diseases, Ninth Revision, Clinical Modification*]), and effectiveness of extracting specific information, such as intensive care unit (ICU) medication details, during the transfer of text-format data. This study aims to demonstrate the potential of natural language–based systems for future health care data exchange.

Methods

Ethical Considerations

This study was approved by the institutional review board of Yongin Severance Hospital (9-2023-0037), and conducted in accordance with the Declaration of Helsinki, and the requirement for written informed consent was waived due to its retrospective nature.

Data Sources

This study used 2 comprehensive public health care data sets, namely, the UK Biobank and the Medical Information Mart for Intensive Care III (MIMIC-III). The UK Biobank serves as a notable national and international health resource, monitoring the lives of 500,000 voluntary participants aged between 40 and 69 years across the United Kingdom from 2006 to 2010. This resource aims to bolster the prevention, diagnosis, and treatment of a wide range of serious and life-threatening diseases. The data set includes genotypic and phenotypic data, covering medical, lifestyle, and environmental aspects. The UK Biobank contains structured data from diverse diagnostic tests, medical and family histories, and various physical measures. The MIMIC-III database, crafted by the Lab for Computational

XSL•FO
RenderX

Physiology at MIT, is a broad, publicly available resource containing the deidentified health data of approximately 40,000 critical care patients [18]. This data set includes demographic information, vital signs, laboratory tests, and medications, among other features. It is valued for its over 2 million free-text clinical notes, presenting a rich source of natural language medical data.

This study used ChatGPT (version 3.5; OpenAI), an artificial intelligence model recognized for its exceptional performance among universally applicable models [19-21]. Given that our primary aim was to assess the ability of LLMs to facilitate health care data exchange in general scenarios, we opted against fine-tuning the model to prevent overspecialization to specific data sets. As a result, we used ChatGPT (version 3.5) in its original form, without any modifications. Furthermore, our focus was on testing the accuracy of information extraction and transformation rather than the creativity of the language model. Therefore, in all experiments, we set the temperature to 0 to ensure a deterministic output from the model.

The objectives of our study required the conduct of multiple trials featuring a range of prompts, a process termed prompt engineering. This process carries the potential risk of introducing an overfitting bias, which could boost the performance on specific data sets. Hence, we differentiated between the data used for prompt engineering experiments and those used to assess the performance of our experiments (Figure 1). Given the absence of a standardized methodology for prompt engineering, researchers often carry out this process manually, relying on trial-and-error approaches based on experience.

Figure 1. Comparison of traditional standardization-based and proposed text–based flexible data exchange processes. This figure illustrates the difference between the conventional process of data exchange, which relies on standardization, and our suggested method of flexible data exchange leveraging unstructured, text-based data. The traditional approach necessitates standardization, potentially leading to the loss or distortion of original information, diminished adaptability in new settings, and an increase in the cost and effort required for data exchange. Our proposed text-based, flexible data exchange process avoids these issues by reducing the loss of original information and boosting adaptability. This method is expected to cut down both the cost and effort involved in data exchange. At the bottom, we have delineated the 3 stages that our experiments aimed to validate. ETL: Extract, Transform, Load; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven International; LLM: large language model.

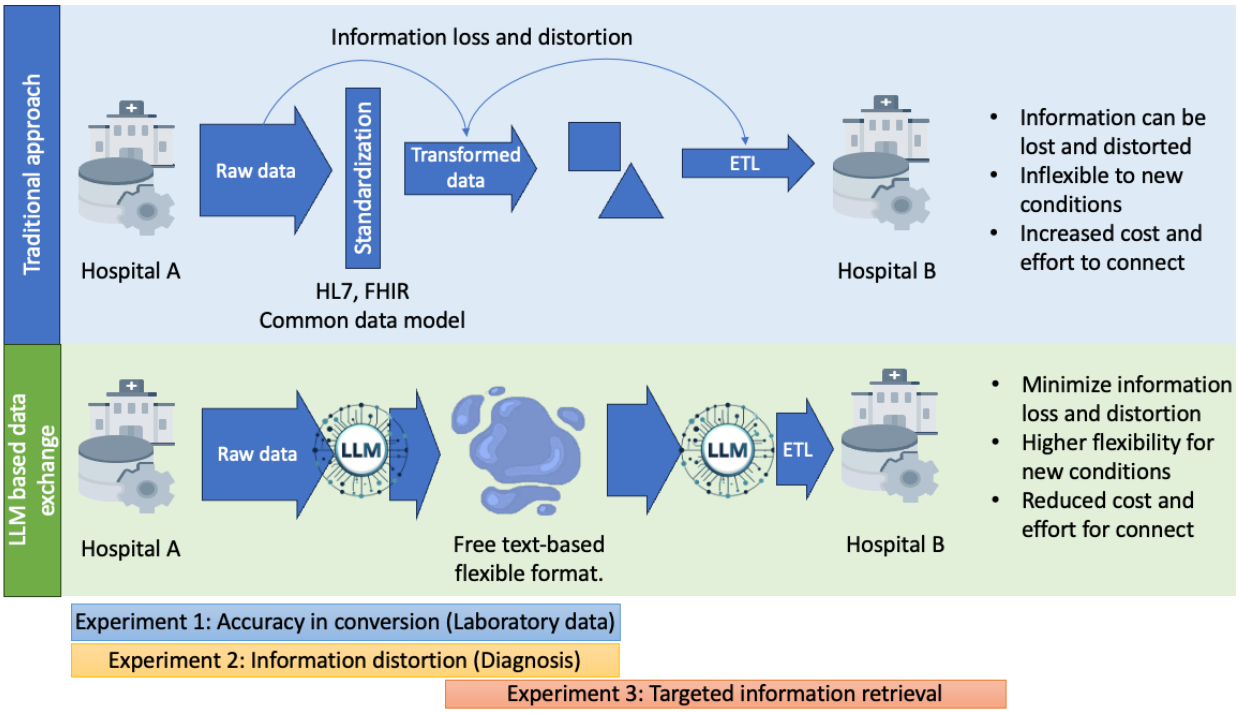

Overview of the Experimental Design

We hypothesized that converting a hospital's data into text format and then integrating such data in another hospital's database can be more accurate and comprehensive compared with other data transformation methods. To prove this, we tested 3 key aspects (Figure 2). First, we investigated whether the original data could be accurately conveyed when transformed into text (experiment 1). This involved converting numerical data into text and back into numerical form to check for any deviations from the original data. Second, we sought to validate that text-based transformation of information with numerical and semantic meaning would result in less distortion compared with rule-based transformations (experiment 2). To this end, we experimented with converting *ICD*-based diagnostic codes into text and back, comparing this with the results of converting them to and from the SNOMED-CT coding system. Finally, we evaluated whether the receiving institution could accurately extract specific desired information during the transmission of complex medical information in text form to another institution (experiment 3). In this experiment, we assumed that the content would resemble a discharge summary when all aspects of a patient's hospital stay were compiled into a text format. Therefore, we aimed to test whether specific data, such as medication information prescribed in the ICU, could be accurately extracted from these summaries. In this experiment, we specifically worked under the assumption that the information to be extracted would be medication information

XSL•FO
RenderX

prescribed in the ICU. From the 3 experiments, we aimed to evaluate the possibility of our hypothesis: a potential solution for health care data exchange in the future might be a system that supports flexible communication of raw data for example, in natural language (experiments 1 and 2), permitting the end user to process and interpret such data as required (experiment 3).

Figure 2. Overview of experimental approaches to evaluate LLM performance in data extraction and transformation. This figure outlines the methodologies used in experiments 1, 2, and 3. (A) In experiment 1, lab data from the UK Biobank data set was presented in natural language format and then restructured back into the MIMIC-III format using the LLM. The restructured data were then compared with the original laboratory data. (B) Experiment 2 evaluated the hypothesis that expressing diagnosis names in natural language might be more efficient than mapping them between varying coding systems. Diagnoses recorded in ICD-9-CM were rendered into natural language or SNOMED-CT, and then reverted into ICD-9-CM codes to examine the degree of information distortion. (C) In experiment 3, the LLM was assigned to extract targeted information concerning medications prescribed in the ICU from discharge summaries in the MIMIC-III database. The extracted information was then compared with the actual prescription records to assess the LLM's accuracy in identifying and extracting details from unstructured text. ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification; LLM: large language model; MIMIC-III: Medical Information Mart for Intensive Care III; SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

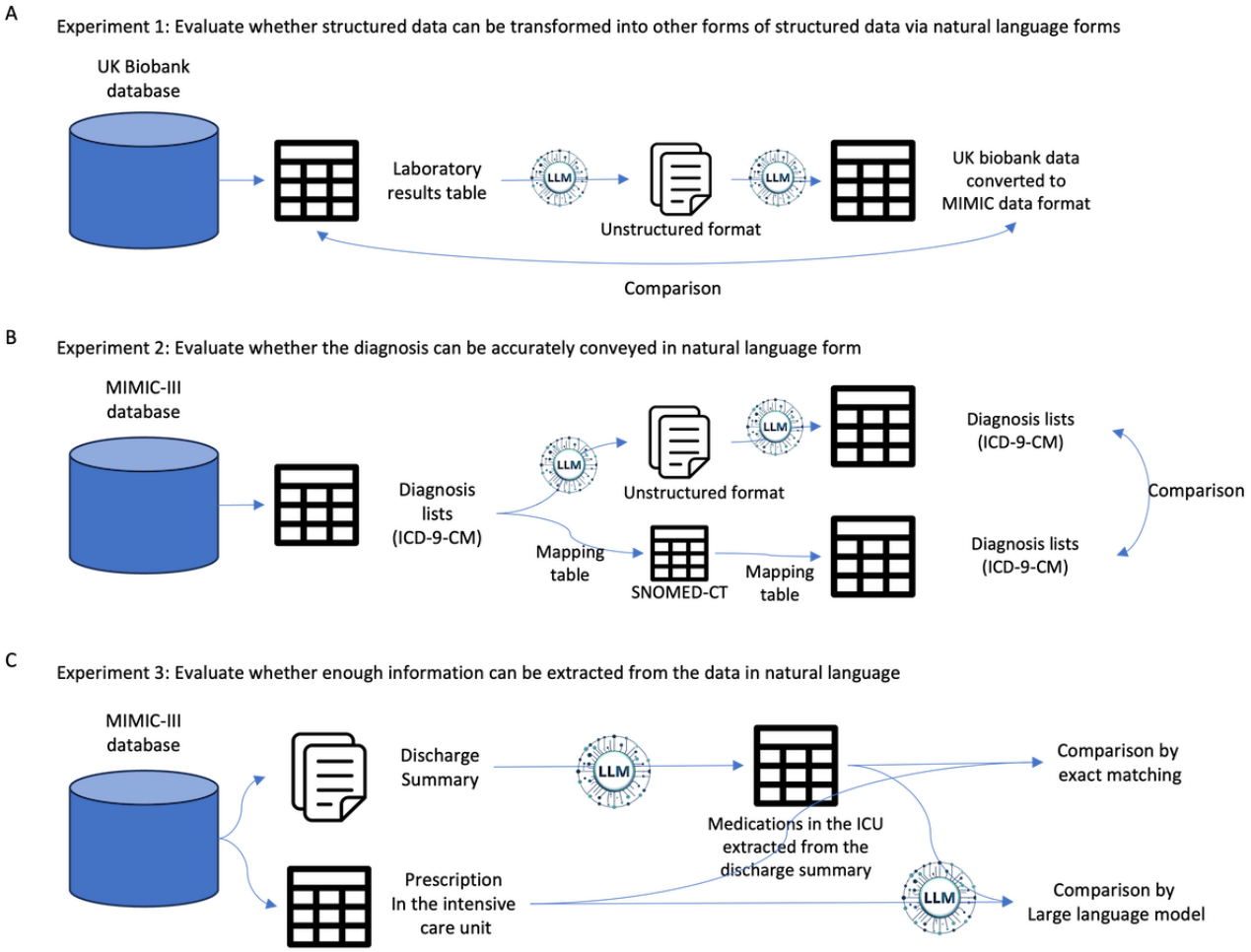

Experiment 1: Evaluating Accuracy in Data Exchange via an LLM

To evaluate the feasibility of data exchange using an LLM, we randomly selected laboratory test result data from 1000 individuals from the UK Biobank data set. For each individual, we gathered laboratory test results and converted them into an unstructured format. Subsequently, the data were restructured to comply with the MIMIC-III data architecture. The prompts used throughout this process are detailed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Summary of prompts used in experiments 1, 2, and 3.

Experiment 1

Step 1: Translating laboratory test results into free text

"I have the following patient. Based on this information, summarize the patient's condition in natural language. Make sure to include all the information presented. The values of the lab results should remain numerical. (For the Sex variable, 0 = female and 1 = male.)"

{List of lab results}

Step 2: Transforming free text data into the structured format

"I have the following patient."

{Generated text from the above step}

"Extract and organize information on the following items."

(Add the value next to the variable name with no further explanation.)

{Defined result extraction format}

Experiment 2

Step 1: Translating diagnosis codes to natural language text

"I have a diagnosis called {Diagnosis code}.

Describe it in natural language used by doctors and other health care professionals.

Write it as a single phrase of only a few words (less than 15 words but do not use abbreviations).

All semantics must be included."

Step 2: Translating natural language text to diagnosis codes

"Where does {Descriptions on diagnosis} fit in the following categories?

{Categories according to International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)}

Provide the most appropriate ICD-9-CM code directly or choose one of the categories above.

Choose only one answer that seems the most relevant and answer in the following format.

The corresponding code: [Code (without periods): a description of the code]."

Experiment 3

Step 1: Extracting medication list from discharge summary

role: "system," "content:" Your role is to interpret medical records.

role: "assistant," "content:" I only need prescriptions from the ICU, not from the general ward or not from outside our hospital.

Organize by ingredient name, not generic name.

Never include medications on admission and discharge medications.

Exclude information before ICU admission or after ICU discharge, even if it is for a hospital stay.

In other words, exclude prescriptions that were written in a regular ward or emergency room.

Exclude any medications that may not have been prescribed in the ICU.

Finally, exclude all prescriptions for procedures, and tests. that are not prescriptions for medication.

role: "user," "content:" Observing the following patient record, organize a list of medications prescribed during the ICU visit.

Organize them in the following format (Provide only the name, not the dose)

drug name 1

drug name 2

If any information on the medications prescribed in the ICU is unavailable, simply answer "None."

Step 2: Converting drug names to ingredient names

{extracted drug list from the above step}

Organize the above medications by ingredient name.

If the drug is recorded by trade name, replace it with the ingredient name.

In the case of multiple ingredient names, record a representative one.

XSL•FO
RenderX

The format should be a single line of ingredient names with no further explanation, like this

List: Ingredient 1, Ingredient 2,...

Step 3: Comparing extracted drug information with actual prescription records

Here is the medication information extracted from the discharge summary.

{extracted drug list from the above step}

These are the medication details actually recorded in the prescription record.

{Ingredient list from the above step}

Organize the medication information extracted from the discharge summary by its actual inclusion in the prescription record.

Medications not mentioned in the discharge summary should not be listed.

The exact name of the medication may not be recorded, or a synonym may be used.

In these cases, mark the medication as actually prescribed.

For example, warfarin might be described as coumadin.

Record the same medication under different names as the one that was prescribed.

Match the same ingredient even if the added bases differ.

For example, the ingredient name of Lopressor is Metoprolol tartrate, but the ingredient must be confirmed as “true” even if it is Metoprolol.

Ingredient names may be written as abbreviations. For example, acetaminophen may be written as APAP.

Exclude P.R.N. prescriptions.

Exclude simple fluid prescriptions.

Provide only “true” or “false” information for each drug.

Do not provide Python code. Provide only the results in an array.

Fill in the blanks with a “true” or “false” result in the following format

{Defined result extraction format}

After the conversion to MIMIC-III data format via the LLM, we checked for potential omissions of information and any discrepancies in numerical values. We assessed the absence or presence of data omissions using sensitivity, specificity, and positive and negative predictive values. To assess the accuracy of the conversion, we used values transformed manually as the reference standard. Sensitivity indicated whether information from the original data set also existed in the transformed data. Conversely, specificity pertained to whether data absent in the original were also absent in the transformed data. The positive predictive value (PPV) referred to whether data present in the transformed data also existed in the original, whereas the negative predictive value determined whether data absent in the transformed data were also absent in the original. Numerical discrepancies were calculated only for test results presented in numerical format. They were assessed via the computation of the mean squared error between the original and transformed values.

Experiment 2: Evaluating Possible Information Distortion During Conversion of Diagnosis Codes

In this experiment, we explored a scenario of diagnostic codes from the primary data set undergoing transformation for sharing across different institutions or to diverse end users. We aimed to clarify potential discrepancies emerging from transitions between the original and an alternate coding framework. Initially, we used a code-mapping table to facilitate the transition from one coding system to another. Subsequently, we reverted the transformed codes to the original coding framework, and then quantified discrepancies by comparing the reverted data against the primary data set. Using the MIMIC-III database, we converted diagnoses encoded in *ICD-9-CM* to SNOMED-CT, and subsequently reverted the same to *ICD-9-CM*. This conversion was based on a mapping table from a previous study [22]. Our proposed approach primarily leveraged the capabilities of the LLM, converting the primary coding structure into a natural text format. For a comparative analysis with the traditional approach, we recoded the text-converted diagnoses into the primary coding system (*ICD-9-CM*) using the LLM, as illustrated in Figure 2B. However, for this experiment, we excluded E and V codes (supplementary classifications for external causes of injury).

In assessing the accuracy of the restoration of diagnostic codes, we conducted evaluations based on the depth of the *ICD-9-CM* coding system. The highest level was labeled level 1, with each subsequent, more specific layer labeled level 2, level 3, and so forth. For instance, if the original data had been coded as “401.1 Hypertension, benign” but the restored data were denoted as “401.9 Hypertension, unspecified,” then the evaluation would be a mismatch at level 3. However, at level 2 granularity (ie, “401. Hypertension”), the codes were considered matching.

XSL•FO
RenderX

Experiment 3: Assessing the Efficacy of LLMs in Extracting Targeted Information From Unstructured Medical Records

To evaluate the capability of our model in extracting targeted medical information from unstructured text, we selected narrative-style discharge summaries from the EVENTNOTES section of the MIMIC-III database, based on the assumption that they would reflect the comprehensive format typical of patient summaries transmitted between hospitals. These summaries provide a comprehensive account of a patient's stay in the ICU, including clinicians' assessments, patient medical history, laboratory results, interpretations of medical imaging, prescriptions, and ensuing care plans. This data set presents a detailed array of narrative insights that illustrate the complexities of patient care, diagnostics, and therapeutic strategies within the ICU context.

For this experiment, we specifically extracted discharge summaries documented by clinicians. These summaries encapsulated patient diagnoses, vital sign readings, current medication regimens, and other relevant status updates, all expressed in natural language. The prompts used in this process are presented in Textbox 1.

To evaluate the performance, we compared the information extracted from natural language with the information stored in structured tables. For this assessment, we made a random selection of 1000 discharge summaries, and we used structured data—prescription records—to verify the accuracy of the information retrieved through the LLM. Our focus was on assessing the PPV, representing the precision of the information extracted by the LLM. The extracted information was considered correct if it was also present within the structured data; otherwise, it was classified as incorrect. Notably, not all prescriptions are routinely documented in natural language by clinicians. Generally, only therapeutics significantly influencing the patient's clinical status would be transcribed in the notes. As such, calculating the negative predictive value (ie, the number of medications not mentioned in the narrative notes that were actually not administered) was deemed impracticable. Similarly, sensitivity (ie, the degree to which prescribed medications are documented in narrative notes) and specificity (ie, the extent to which nonprescribed medications are not mentioned in narrative notes) could not be reliably estimated.

Used Software

We accessed ChatGPT (version 3.5) via its API interface. We used Google BigQuery to manage and deploy the MIMIC-III and UK Biobank data sets. We used Python for certain tasks, such as assessing model performance.

Results

Features of the Extracted Data

In experiment 1, we used the lab test results of 1000 individuals randomly selected from the UK Biobank data set. For experiment 2, we used all diagnosis codes recorded within the MIMIC-III database. Finally, we used 1000 discharge summaries extracted randomly from the MIMIC-III database for experiment 3. Table 1 presents a detailed summary of the data used across all experiments.

Table 1. Summary of data used in each experiment.

	Experiment 1	Experiment 2	Experiment 3
Database	UK Biobank	MIMIC-III[a]	MIMIC-III
Data type	Laboratory test results	Diagnosis code (*ICD-9-CM*[b])	Discharge summary
Number of records	502,396	651,047	59,652
Number of patients	502,396	46,520	41,127
Age (years), mean (SD)	56.53 (8.09)	64.43 (57.20)	58.35 (53.63)
Sex (male), n (%)	229,079 (45.6)	26,121 (56.2)	23,199 (56.4)
Length of text (number. of characters), mean (SD)	N/A[c]	N/A	9618.92 (5539.64)
Number of tests	11,973	N/A	N/A
Number of diagnosis codes	N/A	6984	N/A

[a]MIMIC-III: Medical Information Mart for Intensive Care III.

[b]ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification.

[c]N/A: not applicable.

Results of Experiment 1: Efficiency of the LLM in Data Transformation and Retrieval

In experiment 1, our objective was to assess the capability of the LLM in transforming and extracting laboratory results. We randomly selected the laboratory results of 1000 individuals from an initial data set of 502,396 individuals. This resulted in 11,996 data points spanning 13 distinct test items (excluding tests with null results). These data points were subsequently translated into natural language. Remarkably, only 23 items were lost during the transformation process, with 11,973 (99.8%) being successfully converted. Among the transformed data, 24 items did not match their original values perfectly. However, upon closer examination of these discrepancies, all inconsistencies were found to stem from the rounding off of decimal values. For instance, an original BMI value of 24.4383

XSL•FO
RenderX

was translated as 24.44. Consequently, the calculated mean squared error was a minimal 1.76e-07. Table 2 provides a comprehensive summary of errors for each laboratory test.

Table 2. Summary of experimental results from data transformation and extraction using LLM in experiment 1.

Variable	Raw data		After transformation		Number of data not transferred during the transformation process	Number of data with changed values during the transformation process	MSE[a]
	n, (%)	Mean (SD)	n, (%)	Mean (SD)			
Age	1000 (100)	56.94 (8.03)	1000 (100)	56.94 (8.03)	0	0	0
Sex	1000 (100)	0.47 (0.5)	1000 (100)	0.47 (0.5)	0	0	0
BMI	994 (100)	27.04 (4.78)	994 (100)	27.04 (4.78)	0	24	2.12×10^{-6}
ALT[b]	919 (100)	23.55 (15.32)	919 (100)	23.55 (15.32)	0	0	0
AST[c]	918 (100)	26.13 (11.21)	918 (100)	26.13 (11.21)	0	0	0
Bilirubin	772 (100)	1.84 (0.81)	772 (100)	1.84 (0.81)	0	0	0
Creatinine	920 (100)	72.94 (18.65)	907 (98.6)	72.85 (18.69)	13	0	0
GGT[d]	920 (100)	39.06 (46.96)	920 (100)	39.06 (46.96)	0	0	0
HbA_{1c}[e]	930 (100)	35.88 (5.66)	930 (100)	35.88 (5.66)	0	0	0
HDL[f]	846 (100)	1.46 (0.38)	846 (100)	1.46 (0.38)	0	0	0
LDL[g]	915 (100)	3.54 (0.88)	915 (100)	3.54 (0.88)	0	0	0
Platelet count	943 (100)	255.39 (59.79)	943 (100)	255.39 (59.79)	0	0	0
Triglycerides	919 (100)	1.73 (1.04)	909 (98.9)	1.74 (1.05)	10	0	0
Total	11996 (100)	42.9 (69.66)	11973 (99.8)	42.9 (69.71)	23	24	1.76×10^{-7}

[a]MSE: mean squared error.
[b]ALT: alanine transaminase.
[c]AST: aspartate transaminase.
[d]GGT: gamma-glutamyl transferase.
[e]HbA_{1c}: hemoglobin A_{1c}.
[f]HDL: high-density lipoprotein.
[g]LDL: low-density lipoprotein.

Results of Experiment 2: Analysis of Diagnostic Code Conversion (Mapping Table vs Text-Based Methods)

In the conversion, diagnostic codes were adapted based on a mapping table. Specifically, the original *ICD-9-CM* codes transitioned through SNOMED-CT before being remapped to *ICD-9-CM*. During this procedure, 5748 diagnostic codes expanded to 218,088 codes. This expansion may be attributed to the fact that specific mapping codes do not always allow for a direct 1:1 representation, leading to a 1:n relationship owing to challenges in semantic translation. As an illustration, the *ICD-9-CM* code for “Malignant pleural effusion: Malignant pleural effusion (51181)” was mapped as 2 distinct codes in SNOMED-CT: “Malignant pleural effusion (363346000)” and “Pleural effusion owing to malignant neoplastic disease (disorder) (860792009).” However, when converting through text, the mapping was nearly direct with a 1:1 ratio, ensuring that the 5748 original codes corresponded to 5748 records.

Assessing the results before and after the conversion, we found that the mapping table achieved the following consistency values: 0.096 (21,000/218,088), 0.248 (54,068/218,088), and 0.626 (136,431/218,088) at levels 3, 2, and 1, respectively. Conversely, when relying on text-based methods, the consistency was higher, with corresponding values of 0.597 (3430/5748), 0.844 (4850/5,748), and 0.904 (5197/5748) for the same levels. An important observation pertained to the accuracy of conversion in relation to frequency use is that as

XSL•FO
RenderX

the frequency increased, accuracy followed suit. Specifically, the top 1000 diagnostic names, based on their frequency, achieved values of 0.733, 0.896, and 0.918 at levels 3, 2, and 1, respectively, outperforming less common names. This observed relation was linear, as demonstrated in Figure 3. These results suggested that the frequent use of diagnostic names may provide better precision when shared between different databases.

Figure 3. Results of converting diagnoses from ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) coding to natural language and back to ICD-9-CM. At the highest level, Level 1, most information aligned closely with the original data. However, accuracy decreased as the categories became more specific at levels 2 and 3. Notably, more frequently used diagnoses (toward the left on the x-axis) showed higher conversion accuracy.

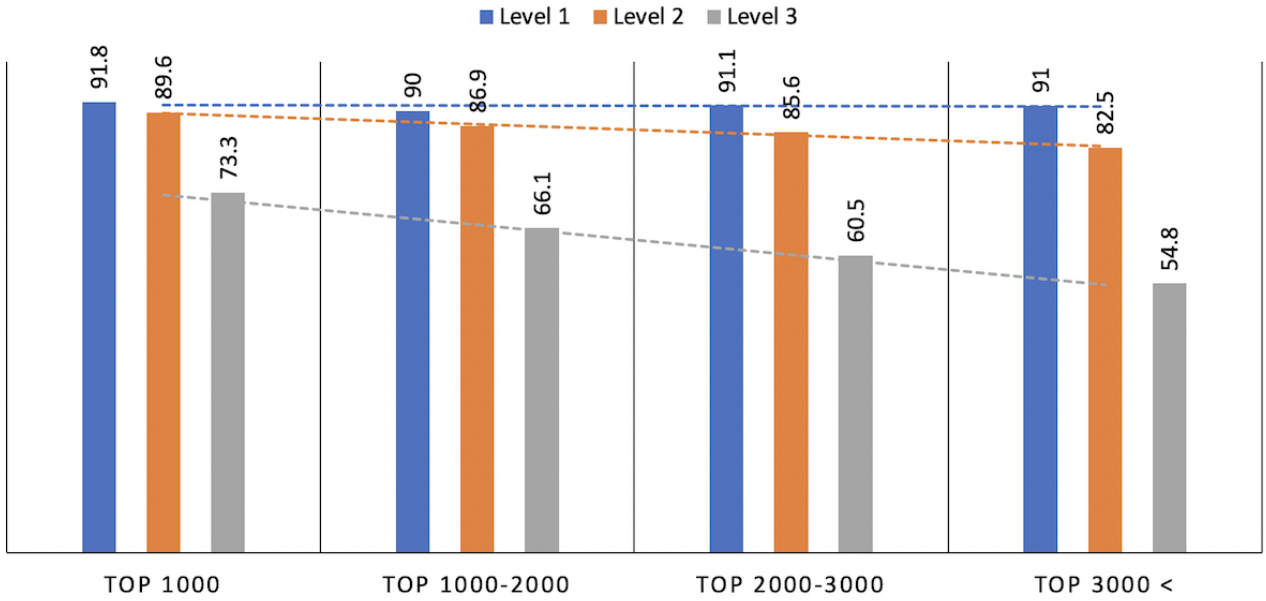

During a review of the misclassified instances, we identified several cases as errors based on our evaluation standards. Notably, the semantic core of the original and converted phrases remained largely consistent. For example, we observed a transformation from “51881: Acute respiratory failure” to “78609: Respiratory abnorm NEC: Other respiratory abnormalities.” A comprehensive list of these misclassifications is provided in Table S1 in Multimedia Appendix 1.

Results of Experiment 3: Effectiveness of the LLM in Extracting Relevant Information From Medical Records

In reviewing 1000 discharge summaries, the LLM identified a total of 5604 instances of medication prescriptions within the ICU setting. Of these, 2483 perfectly matched the entries in the prescription table, resulting in a PPV of 44.3%. When evaluated based on the shared active ingredient, we found a higher level of agreement, with 5055 out of the 5604 (90.2%) prescriptions showing alignment (Table 3). These findings, as exemplified by instances where “Acetaminophen” in the prescription information was referred to as “Paracetamol” in the discharge summaries and cases where “Metoprolol Tartrate” was simply documented as “Metoprolol,” underscore the tendency of physicians to note down familiar medication names. This behavior occurs instead of strictly adhering to the terminology prescribed in the prescription database. These examples highlight a preference for more universally recognized or familiar terms over the precise terminology listed in medical records. Despite this inherent variability in naming conventions, the LLM showed significant effectiveness in identifying and extracting the necessary information.

Table 3. Comparison of drug information extracted from natural language discharge summaries with prescription records.

	Number of medications, n (%)
Medications in the intensive care unit extracted from the discharge summary	5604 (100)
Medications that exactly matched the prescription name	2483 (44.3)
Medications semantically matched by large language model, including synonyms	5055 (90.2)[a]

[a]A total of 2572 medications were described using different terminology than the prescription.

Discussion

Our research highlighted a new direction in health care by demonstrating the effective use of LLMs in medical data exchange. We aimed to overcome the current challenges related to data sharing among health care institutions, particularly owing to the unstructured nature of several medical records. We successfully validated all the key aspects we aimed to investigate, demonstrating the efficacy of our approach in enhancing health care data interoperability. The experiments

XSL•FO
RenderX

revealed that converting hospital data into text format and subsequently integrating the converted data into another hospital's database was not only feasible but also more accurate and comprehensive compared with traditional data transformation methods. Notably, our findings confirmed that the original data retained their accuracy and integrity when transformed into and back from the text format, a crucial factor in health care where precision is paramount. Moreover, our results indicated that text-based transformation, particularly for semantically rich information such as *ICD*-based diagnostic codes, resulted in significantly less distortion compared with rule-based methods. Finally, our method effectively enhanced medical data exchange by enabling precise extraction of specific information, such as ICU medication details, from text-transmitted data, thus, bolstering health care systems' efficiency in integrating such data.

Our study highlights the significant role of LLMs in the field of health care informatics, demonstrating their transformative ability to manage, interpret, and share large volumes of medical data. Traditional data standardization methods, while important, have often been slow and challenging, creating barriers to fast and efficient data exchange. Our results showed that LLMs can not only interpret unstructured data but also convert it into easily understandable formats, greatly reducing the need for time-consuming standardization and allowing for faster data transfer.

Furthermore, the impact of our research extends beyond the clinical or institutional settings, affecting the broader area of personal health records. Integrating data from multiple providers into a single, unified record has always been a complex task. Different institutions often use varied formats, terminologies, and standards. Our work with LLMs suggested that these models can simplify this integration process. By understanding, transforming, and combining different data sources, LLMs can improve data sharing and enrich the information available.

LLMs' adaptability in processing and interpreting structured and unstructured data hints at their potential to significantly enhance the handling of PGHD. Given the variety and unstructured nature of PGHD, from health diaries to wearable technology outputs, our findings suggest a promising avenue for applying LLMs to integrate and understand these diverse data sources more effectively. This capability aligns with our current results. Moreover, it opens up new pathways for creating more personalized and comprehensive approaches to patient care, leveraging the vast and untapped resources of PGHD.

Our study also provided significant insights into the process of converting diagnostic codes between standard coding systems, such as *ICD-9-CM* and SNOMED-CT. The higher number of diagnostic codes produced through this conversion process highlights the detailed and comprehensive nature of code capture enabled by the LLM. However, the approximate 1:1 ratio achieved in text-based conversions points to a more accurate and straightforward method. Importantly, these text-based conversions emphasize the major advantage of keeping the accuracy of the original data. For frequently used diagnostic terms, this method ensured that the core information from the original data remained consistent. Our examination of misclassifications revealed that, although identified as errors based on our criteria, several converted codes maintained similarity in their underlying meaning. Thus, despite "errors" in conversion, the core medical information is typically retained. Moreover, the direct relationship between the accuracy of conversion and frequency of diagnostic names hints at a possible inherent alignment of standard coding systems with commonly used terms. Our findings highlighted the critical importance of preserving data accuracy when moving between detailed medical coding systems. This aligns with the findings of previous studies, which suggest that using LLMs can lead to more accurate phenotype extraction from medical data [23,24].

Our findings have implications beyond individual health care systems and emphasize the potential for a significant change in the global health care landscape. Our data revealed that using LLMs can enhance international health information exchanges. Such improved communication can lead to better collaboration between countries, potentially benefiting patient care worldwide by ensuring that medical knowledge and practices are more consistently applied. Furthermore, our research points to a new direction in the design and operation of electronic medical record systems. The ability of LLMs to efficiently process and structure natural language data can make extracting, analyzing, and presenting medical data more straightforward. This not only allows for immediate analyses using the latest data but also promotes a more adaptable environment within electronic medical record systems to meet the dynamic needs of the health care sector, as illustrated in Figure 3.

While our study demonstrates the promising capabilities of LLMs in medical data processing, it is not without limitations. In this study, we used the GPT-3.5 model. Notably, using the newer GPT-4 might lead to better results, given that the efficiency of LLMs is continually improving. Comparative studies have demonstrated that GPT-4 performs better than its predecessors in various domains [25,26]. This progress in language model capabilities indicates the ongoing advancements we can expect. In addition to technological considerations, our reliance on specific data sets such as MIMIC-III and the UK Biobank, while providing valuable insights, introduces limitations regarding representativeness across diverse health care environments and languages. These data sets, representing particular health care settings and populations, may not fully encapsulate the complexity and diversity of global medical practices, especially in non–English speaking countries. This aspect underscores the necessity for broader research in applying LLMs across more varied data sets to ensure generalizability and applicability to different health care contexts. Regarding technological improvements, on-premise solutions can be expected to continue to improve in capabilities. Hence, our research serves as a foundation, showing the feasibility of data exchange based on LLMs. The accuracy and use of these transformations will be enhanced further in future versions. For institutions concerned with security implications, transitioning from externally provided models, such as ChatGPT, to an on-premise, self-built language model is a recommended strategy. Custom-built models can match the performance of GPT-3.5 for specific tasks [27,28]. Our choice to evaluate performance using the 3.5 version in this research provides a

XSL•FO
RenderX

reference point and offers guidance for users considering the use of their custom language models.

Our research focused on specific data sets, and more extensive studies involving a wider range of data would be needed to confirm our initial observations. Moreover, the ability of LLMs to handle different types of unstructured data, each with its unique challenges, requires thorough assessment. Nevertheless, with ongoing advancements in artificial intelligence and machine learning, we expect that these challenges will be addressed, and the efficiency of LLMs in managing medical data will continue to improve. Future versions of LLMs, combined with careful validation, can bring significant improvements to health care informatics.

Conclusions

In conclusion, our in-depth study provides important insights into the potential transformation of health care data exchange in the near future. The LLMs have a significant role in enhancing medical data sharing, ensuring both precision and efficiency. As technology advances and these language models become more refined, their role in health care data management and communication is anticipated to expand. Their potential goes beyond merely simplifying processes; they might also play a key role in minimizing errors, guaranteeing that medical professionals worldwide can access accurate and timely data. Ultimately, our findings suggest that with the incorporation of LLMs, the global health care landscape could become more unified, facilitating seamless knowledge transfer and collaboration among health care providers everywhere.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D (Research and Development) Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number HI22C0452). This study was conducted using data from UK Biobank (application number 85037).

Data Availability

The data sets analyzed during this study are available in the UK biobank [29] and MIMIC-III [30] repository.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A comprehensive list of misclassifications in experiment 2.
[XLSX File (Microsoft Excel File), 38 KB - jmir_v26i1e56614_app1.xlsx]

References

1. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294 [FREE Full text] [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
2. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: a scoping review. J Am Med Inform Assoc 2011;18(4):515-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000105] [Medline: 21672914]
3. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
4. Celik Y, Godfrey A. Bringing it all together: wearable data fusion. NPJ Digit Med 2023;6(1):149 [FREE Full text] [doi: 10.1038/s41746-023-00897-6] [Medline: 37591989]
5. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. J Med Internet Res 2015;17(2):e40 [FREE Full text] [doi: 10.2196/jmir.3703] [Medline: 25707035]
6. Patel MS, Asch DA, Volpp KG. Wearable devices as facilitators, not drivers, of health behavior change. JAMA 2015;313(5):459-460. [doi: 10.1001/jama.2014.14781] [Medline: 25569175]
7. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
8. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
9. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]

XSL•FO
RenderX

10. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
11. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019;394(10211):1816-1826 [FREE Full text] [doi: 10.1016/S0140-6736(19)32317-7] [Medline: 31668726]
12. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
13. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
14. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology 2023;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
15. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023;6(1):158 [FREE Full text] [doi: 10.1038/s41746-023-00896-7] [Medline: 37620423]
16. Guo E, Gupta M, Deng J, Park Y, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res 2024;26:e48996 [FREE Full text] [doi: 10.2196/48996] [Medline: 38214966]
17. Han C, Kim DW, Kim S, Chan You S, Park JY, Bae S, et al. Evaluation of GPT-4 for 10-year cardiovascular risk prediction: insights from the UK Biobank and KoGES data. iScience 2024;27(2):109022 [FREE Full text] [doi: 10.1016/j.isci.2024.109022] [Medline: 38357664]
18. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
19. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
20. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng 2023;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
21. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
22. OHDSI vocabularies repository. ATHENA. URL: https://athena.ohdsi.org/search-terms/start [accessed 2023-12-02]
23. Cunningham JW, Singh P, Reeder C, Claggett B, Marti-Castellote PM, Lau ES, et al. Natural language processing for adjudication of heart failure in a multicenter clinical trial: a secondary analysis of a randomized clinical trial. JAMA Cardiol 2024;9(2):174-181. [doi: 10.1001/jamacardio.2023.4859] [Medline: 37950744]
24. Alsentzer E, Rasmussen MJ, Fontoura R, Cull AL, Beaulieu-Jones B, Gray KJ, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. NPJ Digit Med 2023;6(1):212 [FREE Full text] [doi: 10.1038/s41746-023-00957-x] [Medline: 38036723]
25. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023;13(1):18562 [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]
26. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
27. Latif E, Zhai X. Fine-tuning ChatGPT for automatic scoring. ArXiv. Preprint posted online on October 16, 2023 [FREE Full text]
28. Soman K, Rose PW, Morris JH, E Akbas R, Smith B, Peetoom B, et al. Biomedical knowledge graph-enhanced prompt generation for large language models. ArXiv. Preprint posted online on November 29, 2023 [FREE Full text]
29. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015 Mar;12(3):e1001779 [FREE Full text] [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
30. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]

Abbreviations

FHIR: Fast Healthcare Interoperability Resources
ICD: International Classification of Diseases
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
ICU: intensive care unit
LLM: large language model
MIMIC-III: Medical Information Mart for Intensive Care III

XSL•FO
RenderX

PGHD: patient-generated health data
PPV: positive predictive value
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms

Edited by G Tsafnat; submitted 22.01.24; peer-reviewed by J Lee, S Lee; comments to author 22.03.24; revised version received 22.04.24; accepted 27.04.24; published 31.05.24.

Please cite as:
Yoon D, Han C, Kim DW, Kim S, Bae S, Ryu JA, Choi Y
Redefining Health Care Data Interoperability: Empirical Exploration of Large Language Models in Information Exchange
J Med Internet Res 2024;26:e56614
URL: https://www.jmir.org/2024/1/e56614
doi:10.2196/56614
PMID:38819879

©Dukyong Yoon, Changho Han, Dong Won Kim, Songsoo Kim, SungA Bae, Jee An Ryu, Yujin Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Contextual Barriers to Implementing Open-Source Electronic Health Record Systems for Low- and Lower-Middle-Income Countries: Scoping Review

Sarah Bostan[1,2,3], PhD; Owen A Johnson[3]; Lena J Jaspersen[1], PhD; Rebecca Randell[4,5], PhD

[1]Leeds University Business School, University of Leeds, Leeds, United Kingdom
[2]School of Healthcare, University of Leeds, Leeds, United Kingdom
[3]School of Computing, University of Leeds, Leeds, United Kingdom
[4]Faculty of Health Studies, University of Bradford, Bradford, United Kingdom
[5]Wolfson Centre for Applied Health Research, Bradford, United Kingdom

Corresponding Author:
Sarah Bostan, PhD
Leeds University Business School
University of Leeds
Maurice Keyworth Building
Woodhouse
Leeds, LS2 9JT
United Kingdom
Email: s.bostan@leeds.ac.uk

Abstract

Background: Low- and lower-middle-income countries account for a higher percentage of global epidemics and chronic diseases. In most low- and lower-middle-income countries, there is limited access to health care. The implementation of open-source electronic health records (EHRs) can be understood as a powerful enabler for low- and lower-middle-income countries because it can transform the way health care technology is delivered. Open-source EHRs can enhance health care delivery in low- and lower-middle-income countries by improving the collection, management, and analysis of health data needed to inform health care delivery, policy, and planning. While open-source EHR systems are cost-effective and adaptable, they have not proliferated rapidly in low- and lower-middle-income countries. Implementation barriers slow adoption, with existing research focusing predominantly on technical issues preventing successful implementation.

Objective: This interdisciplinary scoping review aims to provide an overview of contextual barriers affecting the adaptation and implementation of open-source EHR systems in low- and lower-middle-income countries and to identify areas for future research.

Methods: We conducted a scoping literature review following a systematic methodological framework. A total of 7 databases were selected from 3 disciplines: medicine and health sciences, computing, and social sciences. The findings were reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. The Mixed Methods Appraisal Tool and the Critical Appraisal Skills Programme checklists were used to assess the quality of relevant studies. Data were collated and summarized, and results were reported qualitatively, adopting a narrative synthesis approach.

Results: This review included 13 studies that examined open-source EHRs' adaptation and implementation in low- and lower-middle-income countries from 3 interrelated perspectives: socioenvironmental, technological, and organizational barriers. The studies identified key issues such as limited funding, sustainability, organizational and management challenges, infrastructure, data privacy and protection, and ownership. Data protection, confidentiality, ownership, and ethics emerged as important issues, often overshadowed by technical processes.

Conclusions: While open-source EHRs have the potential to enhance health care delivery in low- and lower-middle-income-country settings, implementation is fraught with difficulty. This scoping review shows that depending on the adopted perspective to implementation, different implementation barriers come into view. A dominant focus on technology distracts from socioenvironmental and organizational barriers impacting the proliferation of open-source EHRs. The role of local implementing organizations in addressing implementation barriers in low- and lower-middle-income countries remains unclear.

XSL•FO
RenderX

A holistic understanding of implementers' experiences of implementation processes is needed. This could help characterize and solve implementation problems, including those related to ethics and the management of data protection. Nevertheless, this scoping review provides a meaningful contribution to the global health informatics discipline.

(J Med Internet Res 2024;26:e45242) doi:10.2196/45242

KEYWORDS

implementation; open source; electronic health records; digital health; low- and lower-middle-income countries; barriers; global health care; scoping; review

Introduction

Background

Low- and lower-middle-income countries (LMICs) are challenging contexts that lack robust infrastructure, technical expertise, and other key resources [1-3]. In most LMICs, there is limited access to information about the health of individuals considered vulnerable, making it difficult to improve health care systems because these settings often require additional funding and maintenance support [4,5]. Furthermore, individuals considered vulnerable seldom have a platform to express their views on health care development and implementation strategies within their communities [6-8]. The resource-scarce settings of LMICs account for a higher percentage of global epidemics and chronic diseases in comparison to the Global North because of limited access to health care, particularly for individuals considered vulnerable [9-11]. This persistent problem of limited access to health care exacerbates inequalities in LMIC settings [12,13], and this calls for innovative and sustainable interventions.

Open-source electronic health records (EHRs) can enhance health care delivery in LMICs by improving the collection, management, and analysis of health data needed to inform health care delivery, policy, and planning. Open source is broadly defined as free software that includes a flexible source code [14-16] that can be modified for various settings [17]. An EHR system is a computerized version of a patient's (longitudinal) medical records maintained by a given health care provider [18-20]. There is demand for open-source EHR systems in LMICs because they provide implementers with the flexibility to customize the system to meet context-specific needs [21]. An implementer is a member of a local implementing organization who understands the technology, context, and end users. An implementing organization can be a social enterprise or a nongovernment organization that is concerned with the implementation of software. Local implementers and software developers implement open-source EHRs to enhance the delivery of health care for local health facilities [22-25].

The process of implementing open-source EHRs often requires implementing organizations to systematize and conduct an initial analysis of the context, reinvent the software to meet local requirements, lead configuration and installation, and provide user training [4,18,22]. While open-source EHR systems are cost-effective and adaptable, they have not proliferated rapidly in LMICs. Implementation barriers slow adoption, with existing research focusing predominantly on technical issues preventing successful implementation. This scoping review provides an overview of barriers impacting the implementation of open-source EHR systems in LMICs, identifies gaps in the existing literature, and points to opportunities for future research.

Digital Transformation of Global Health Care

The transformation and innovation of technology over the last few decades have shaped how the global health care industry operates [26]. The demand for universal access to quality health care is rising, putting pressure on governments to develop sustainable solutions for the effective delivery of health care [27]. While digital technologies have the potential to enable equal delivery of better health care [28], they also raise important questions about ethics and data protection [9,29-31], particularly in settings characterized by stark power imbalances. A focus on the development of better technical solutions can distract from the underlying values and ethical concerns associated with health care technologies.

Goal 3 of the United Nations sustainable development goals, "good health and wellbeing," highlights the importance of improving access to quality health care and managing global health risks [32]. The World Health Organization advocates the use of electronic health tools [33] to enhance the monitoring of patient care [34,35]. An EHR is "a computer based patient records system designed mainly for the use of doctors [or other clinicians that have direct contact with the patient]" [19]. EHR systems can help ease the burden of existing paper record processes and provide better management of patient care electronically [24]. The sustainable development and implementation of EHR systems is a challenge in the contemporary environment [21,36], specifically from political, economic, social, technological, and ethical perspectives [19,37]. Data ownership, informed consent, data protection, and confidentiality are concerns that influence the implementation of many health technologies, including EHR systems [31,38]. Evans [39] and Manders-Huits [40] assert the importance of acknowledging and integrating human values responsibly in health care technology. Therefore, if the context of implementing EHR systems is better understood, it could help address key challenges and barriers from a comprehensive perspective.

Evolution of Open-Source EHR Systems

There has been a rapid growth of open-source software, notably in the health care sector [18,41]. The phrase "open-source software" was coined in 1998 [17]. Open-source software includes an adaptable source code; when the source code is made publicly accessible under a free license [42], it can be customized by health care providers to meet context-specific requirements [14,15,43,44]. There are various open-source software solutions used for distinct purposes in LMIC settings

XSL•FO
RenderX

[45,46], such as the Open Enterprise-level Laboratory Information System (OpenELIS Foundation); District Health Information Software 2 (HISP Centre at the University of Oslo), a web-based platform communicating health data across several levels of a given health care system; and GNU Health, a hospital information system [24,47-49].

An open-source EHR system provides an adaptable and digitalized version of a patient's medical history, a record that comprises identifiable and personal health information such as demographics, allergies, medication, medical episodes, and health facility visits [18-20]. Open-source EHR systems can enhance health care delivery, inform the development and delivery of health care at the policy level, and lower costs for LMIC settings [21-23,50,51]. A good example of an open-source EHR system, adopted primarily for LMIC settings, is the Open Medical Record System (OpenMRS; OpenMRS Inc) platform [52-55]. The OpenMRS platform is perceived as a collaborative project [25,56-58] aiming to serve a moral purpose by "bringing people together to write code and save lives" [59].

Two early examples of open-source EHRs were the Computer Stored Ambulatory Record and the Veterans Health Information Systems and Technology Architecture, both developed nearly 5 decades ago in a high-income country (the United States) [20,41,60,61]. Since then, there has been an increasing interest in developing and using open-source EHR solutions [17,18,24]. Today, open-source EHR systems are used in many countries, but predominantly in LMICs, where they are seen as helping to address the high cost and inflexibility associated with proprietary EHR systems [4,23,62-65]. Open-source EHR systems can, however, introduce different tensions when implemented in LMIC settings [15,18]. Despite the promises of open-source EHR systems, they are not proliferating as expected. It remains unclear what barriers inhibit their implementation, whether these barriers vary according to different contexts, and how they can be addressed.

Open-Source Software Versus Proprietary Software

Open-source software offers publicly available source code that can be modified and distributed without incurring licensing fees [14,42,66]. By contrast, proprietary software has copyright restrictions on source code that is not publicly available [16]. Proprietary software such as Microsoft Windows or Office can be perceived as an out-of-the-box solution, and any adaptations required must be completed by the proprietor of the software, resulting in additional fees [17]. Open-source software can be tailored to the specifics of a given context, but this often requires technical expertise and adequate funding for implementation [29,67-69]. Moreover, Reynolds and Wyatt [15] contend that opening the source code compels developers to carefully examine and craft the quality of their code, making bug patching easier, which strengthens the security aspect of open-source systems [41,43]. Proprietary software can be more costly to develop in comparison to open-source software [51], where the source code can be adapted and shared, particularly if there is a need to customize certain system aspects or add additional system features [44,70].

Open-source software does not miraculously address the inadequacies of existing health care systems [17,66,68,71-73]. It still requires a level of expertise and human competence for software design and developing effective systems for end users [74]. Nevertheless, open-source software offers the potential for communities to collaborate effectively, build stronger networks, develop new skills, and transform policy and practice where required [75,76]. Therefore, there are several benefits of using open-source software for health care in LMIC settings [22,51,77].

Open-source EHR systems provide implementers with greater flexibility in building customized systems for a given context and can ease suffering from vendor lock-in often found with proprietary EHR systems [16,44,65,69,77]. For example, vendors of proprietary EHR systems are restrictive in what they share with users, adopt surveillance measures, and impose upgrades [14,15], thus inhibiting freedom and flexibility for end users [18,61]. Conversely, a perceived challenge with open-source EHR systems is that they require specialized skills for implementation and maintenance support [4,17]. Nevertheless, open-source EHR systems allow implementers to adapt a given system to meet context-specific needs [51,70,78].

Context-Specific Barriers

For LMIC settings, the adoption of open-source EHR systems requires context-specific adaptations [4].

Context can be defined as "the place where an intervention is delivered...or unique factors surrounding an implementation effort." [79]. There is extensive literature on EHR implementations [4,24,28]; however, existing research has focused on technical perspectives [30,31,40] and factors such as technology infrastructure, power supply, and backups as well as a lack of financial resources [22,65,70,80,81]. However, local and regional context-specific barriers inhibiting the adaptation and implementation of open-source EHRs and issues inhibiting implementing organizations from adopting such technology in LMICs have not yet been researched.

IT implementation requires diverse stakeholders "taking a design and translating it into a working system" [19]. The process of implementing an open-source EHR system involves different stakeholders conducting initial analysis and adapting the software to local requirements, software development, configuration, and installation, as well as providing user training and support [22,82,83]. Open-source EHR implementation requires stakeholder engagement and participation to cocreate a solution that results in a change that generates true value for a given context [79]. Consequently, there is a need for various stakeholders, such as local software developers, implementers, IT providers, and health care practitioners, to work collaboratively when designing and adopting open-source EHR systems for LMIC settings [21,39,84].

Therefore, this scoping review provides an overview of the context-specific barriers and facilitators impacting the adaptation and implementation of open-source EHR systems in LMIC settings. No previous scoping review has explored the perceived contextual barriers impacting the adaptation and implementation of open-source EHRs for LMICs.

XSL•FO
RenderX

Methods

Overview

This scoping review aims to provide an overview of the contextual barriers impacting the adaptation and implementation of open-source EHR systems in LMIC settings and outline opportunities for future research. A scoping review methodology was chosen because it provides an understanding of the potential breadth of literature available, lends itself to the identification of relevant concepts and research gaps, and enables the researcher to assess whether a full systematic review is needed or indeed possible [85,86]. The following subsections describe the methodological framework and approach used to undertake the scoping review.

Stage 1: Identifying the Research Question

To address the interdisciplinary nature and scope of the gap in our understanding, a broad review question was chosen to map the breadth of literature available and identify key concepts and related themes for further exploration. The following research question was formulated: "What are the perceived key contextual barriers impacting the adaptation and implementation of open-source EHRs in LMICs?"

This review follows the 5 stages of the systematic methodological framework for conducting scoping studies by Arksey and O'Malley [87] and follows relevant guidance from the JBI [88], Levac et al [89], and Davis et al [90].

Stage 2: Identifying Relevant Studies

The 7 most relevant electronic databases were searched from January 1960 to September 2021 (Textbox 1), as EHR developments initially started in the early 1960s [20]. The first reviewer (SB) performed a comprehensive search, which was supported using a population, intervention, and outcome framework [91] that identified four key terms: (1) LMIC, (2) open source, (3) EHR, and (4) adaptation and implementation. The key terms and synonyms are detailed in Multimedia Appendix 1. The selection of databases covered literature from medicine and health sciences, computing, and social sciences. In addition, Open Grey was used to identify relevant gray literature using the key terms, but no results were retrieved. A backward and forward citation search was performed using Google Scholar from reference lists of selected papers to ensure potentially relevant articles were not overlooked [91]. The searches concluded on September 27, 2021. The searches for MEDLINE and MEDLINE In-Process, CINAHL, and Web of Science are included in Multimedia Appendix 2.

Textbox 1. Literature search.

Electronic databases

- MEDLINE and MEDLINE In-Process
- Embase
- CINAHL
- EBSCO Business Source Premier
- Web of Science
- Cochrane Library
- IEEE Xplore

Stage 3: Study Selection

Studies were included in the review if they were published in English, presented empirical data (including systematic reviews), or nonempirical accounts of experiences and system descriptions on the adaptation and implementation of open-source EHRs in all variations (associated synonyms and phrases) in LMICs. No restrictions were placed on the study design or the format of the publication, and both published and unpublished (gray) literature were included. Papers were reviewed at the title, abstract, and full-text level to exclude articles that did not address the adaptation and implementation of open-source EHRs or were not concerned with LMICs. No other criteria were applied; this was to maximize the search results for an initial inquiry. Multimedia Appendix 3 [3] shows the list of LMICs used for the inclusion criteria.

The following terms were used in the search for relevant material: LMIC, open source, EHR, adaptation, and implementation. These were defined before the search to ensure consistency across the range of databases. The definitions of these key terms can vary; however, for this review, the following definitions were adopted: the World Bank's definition of LMICs was used [3]; an *EHR* system is a computerized version of a patient's (longitudinal) medical records, specifically designed for clinicians who have direct contact with patients [19,20]; *open-source software* includes publicly available source code that can be adapted to meet context-specific requirements without incurring any license fees [14-17,42,43]; *implementation* is a complex process that requires organized and deliberated effort to put a given innovation or intervention into practice in such a way that it results in better outcomes for an identified context [79,92]; and *adaptation* is an element of the implementation process. It is the construction of different processes (eg, analysis, customization, installation and configuration, training, and support) that support a given innovation or intervention [93].

Overall, the search yielded 3504 articles, which were exported into EndNote X9 reference software (Clarivate). After removing 893 (25.49%) duplicates from the 3504 articles, the first reviewer (SB) screened articles via a 2-level process: 2611 (74.51%) titles and abstracts were screened for relevance. Thus,

XSL•FO
RenderX

of the 2611 studies, 170 (6.51%) full-text articles were included for review. In addition, a team of 3 reviewers (RR, LJ, and OJ) independently screened 10% (17/170) of the articles to ensure consistency using the inclusion and exclusion algorithm for screening titles and abstracts and then full-text reviews (Multimedia Appendix 4). Any disagreements were discussed and resolved among the reviewers. Finally, 12 studies met all criteria of the inclusion and exclusion algorithm. An additional article was discovered through the forward and backward reference list checks. In total, 13 studies were found to satisfy the inclusion and exclusion criteria and were included in the review. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [94] flow diagram was used to report the results of the study selection process (Figure 1).

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram detailing the study selection process. EHR: electronic health record; LMIC: low- and lower middle–income country

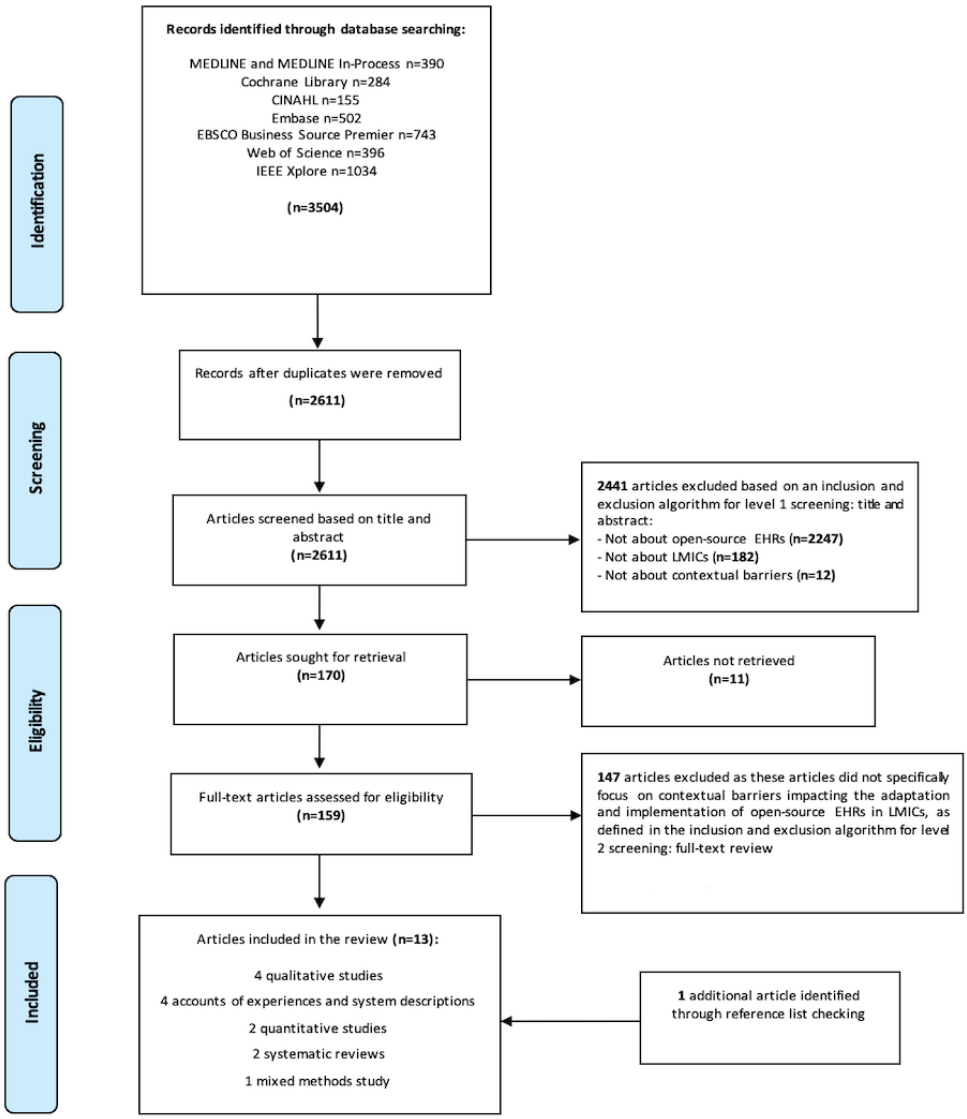

Stage 4: Charting the Data

A standard data extraction form was used to obtain an overview of the 13 selected studies. For each study, the following information was extracted: authors publication year, country of origin (where the study was conducted), aims or purpose, study design, study population and sample size (if applicable), methods, intervention type (open-source EHRs), and key findings that relate to the scoping review question. Any inconsistencies were discussed and resolved among the reviewers.

Stage 5: Collating, Summarizing, and Reporting the Results

After charting the key data from the 13 studies, a qualitative thematic analysis [95,96] and a synthesis approach was adopted [89]. The first reviewer coded, categorized, and grouped the results into key themes to address the scoping review question and identify the implications for future research. The thematic analysis was inspired by existing frameworks presented by Jawhari et al [84] and Muinga et al [21], who synthesized key findings using relevant categories to illustrate the advantages and barriers to implementing electronic medical record systems in a given geographically bounded space. However, this scoping

XSL•FO
RenderX

review focused on the global level, examining contextual barriers impacting the adaptation and implementation of open-source EHR systems for LMIC settings (local and regional). Therefore, the categories and key themes identified in this review were generated from the analysis of the 13 studies.

Quality Assessment

The Mixed Methods Appraisal Tool and the Critical Appraisal Skills Programme checklists were adopted to assess the quality, where relevant, of the included studies, as the review comprises a broad range of study designs and methodologies [97-99].

Results

Characteristics of Studies

Table 1 briefly summarizes the characteristics of the included studies. All studies were published between 2002 and 2021. Of the 13 studies included in this review, geographically, 9 (69%) report research conducted in Sub-Saharan Africa, Uganda, Ethiopia, Kenya, Ghana, and Sierra Leone [9,21,24,28,44,84,100-102]. A total of 2 studies were conducted in South Asia: India and Nepal [103,104]. The type of research varied across the spectrum, with the most common following a qualitative design, with the use of interviews, surveys including qualitative questions, participatory techniques and observations used to address the design, and barriers and facilitators with implementing open-source EHRs [9,21,84,101].

Table 1. A summary of the characteristics of the included studies.

Study and country	Study design	Urban or rural
Mohammed-Rajput et al [9], 2011; Kenya, Rwanda, Lesotho, Tanzania, Uganda, and Malawi	Qualitative	Rural
Syzdykova et al [24], 2017; Ethiopia	Systematic review	Rural
Muinga et al [21], 2018; Kenya	Qualitative	Rural
Oza et al [28], 2017; Sierra Leone	Quantitative	Rural
Akanbi et al [44], 2012; Sub-Saharan Africa	Systematic review	Rural
Fish and Guha [68], 2020; Haiti	Descriptive	Rural
Verma et al [83], 2021; Kenya, Nepal, Liberia, Lesotho, Haiti, Uganda, Sierra Leone, Rwanda, Nigeria, Mozambique, Malawi, Kazakhstan, India, Ethiopia, Democratic People's Republic of Korea, and Peru	Mixed methods	Rural
Jawhari et al [84], 2016; Kenya	Qualitative	Urban
Gainer et al [100], 2012; Ethiopia	Descriptive	Rural
Gyamfi et al [101], 2017; Ghana	Qualitative	Urban
Were et al [102], 2010; Uganda	Quantitative	Urban
Anantraman et al [103], 2002; India	Descriptive	Rural
Raut et al [104], 2017; Nepal	Descriptive	Rural

Of the 2 quantitative studies, one used controlled observations of clinicians and patients before implementation and postimplementation and the other used a survey with standardized measures [28,102]. One study adopted a mixed methods design inclusive of quantitative and qualitative aspects [83]. The systematic reviews looked at various open-source EHRs in LMICs and the challenges inhibiting implementation [24,44]. Other studies were more accounts of experiences and system descriptions [68,100,103,104]. Most of the studies (10/13, 77%) addressed rural areas [9,21,24,28,44,68,83,100,103,104], and other studies (3/13, 23%) addressed urban settings [84,101,102]. Multimedia Appendix 5 [9,21,24,28,44,68,83,84,100-104] illustrates the detailed characteristics of each included study.

Quality Assessment

The Mixed Methods Appraisal Tool checklist was used for 7 studies [9,21,28,83,84,101,102], and the Critical Appraisal Skills Programme Systematic Review checklist was used for 2 systematic reviews [24,44]. A total of 4 studies [68,100,103,104] were nonempirical accounts of experiences and system descriptions of open-source EHR implementations, which provided interesting insights but were not suitable for quality assessment. In total, 2 qualitative studies in this review were of moderate quality, as it is unclear what data were included in the analysis [9], and there is no information about the analysis [21]. A total of 5 studies were assessed as good quality: 2 qualitative [84,101], 2 quantitative [28,102], and the mixed methods study [83]. Moreover, 2 systematic reviews were assessed as moderate quality [24,44] with less scientific rigor, as there was no information about the quality of the included studies. The data table reporting the quality assessments is included in Multimedia Appendix 6 [9,21,24,28,44,83,84,101,102].

Contextual Barriers Impacting Open-Source EHR Implementations

Overview

The 13 included studies provide a broad overview of the perceived contextual barriers impacting the adaptation and implementation of open-source EHRs in LMICs. Three distinct but interrelated perspectives emerged from the thematic review:

XSL•FO
RenderX

(1) *socioenvironmental barriers* draw attention to issues surrounding the relationship between humans in a given society and their external environments, such as social capital, social cohesion, infrastructure (local and regional), culture, values, languages, institutions, and stakeholders; (2) *technological barriers* emphasize the issues surrounding the software and hardware used in open-source EHR implementations; and (3) *organizational barriers* draw attention to the operational practices in organizational structures of open-source EHR implementations. In Textbox 2, we organize our findings around these 3 themes. While some issues can be assigned to one thematic barrier, others are cross-cutting barriers, as summarized in Textbox 2. Some subcategories are addressed within >1 barrier, where they are interpreted through different lenses. For example, issues such as infrastructure, ethical practices, or finance are not only of importance from a socioenvironmental perspective but are also raised, albeit in different ways, by researchers who adopt a technological or organizational perspective.

Textbox 2. Thematic framework (contextual barriers and subcategories [subsumed issues]) for the analysis of adaptation and implementation barriers impacting open-source electronic health record systems in low- and lower-middle-income countries.

Socioenvironmental barriers

- Lack of social cohesion: voice and trust [9,21,24,28,44,84,100,102-104]
- Require stakeholder engagement and participation as well as political support [9,21,24,28,44,68,83,84,100-104]
- Sustainability: co-design and collaboration [9,21,24,28,44,68,83,84,100-104]
- Social capital and lack of funding [9,21,24,28,44,68,83,84,100-104]
- Language barriers: reliance on the local language [44,68]
- Epidemic diseases (health emergencies) [9,21,28,44,68,84,100,103,104]
- Environment—lack of resources (poverty) [9,21,24,28,44,68,83,84,100-104]
- Infrastructure: access to electricity, local network coverage, medical facilities, and rural and urban health [9,21,24,28,44,68,83,84,100-104]

Technological barriers

- Infrastructure (power, network, and technology) [9,21,24,28,44,68,83,84,100,101,104]
- Data security, privacy and confidentiality, storage, quality, and ethics [9,21,24,28,44,68,83,84,100-102,104]
- Software and hardware suitability (context specific) [9,21,28,44,68,84,100-102,104]
- Interoperability [21,24,28,68,84,100]
- Sustainability of systems [9,21,28,44,68,83,84,100-102,104]
- User interface: not supporting different clinical roles [9,21,24,28,44,68,83,84,100,101]
- Patient-centered design (lack of end user [clinicians, health facility administrators, and patients] involvement) [9,21,28,68,83,84,100,103,104]

Organizational barriers

- Finance and benefactors (costs for context-specific implementations and maintenance) [9,21,24,28,44,68,83,84,100,101,103,104]
- Human resource development (training, IT skills or expertise, support, and staff shortage) [9,21,24,28,44,68,83,84,100-102,104]
- Resistance to change [21,28,44,68,83,84,100,101]
- Organizational culture and change management [9,21,24,28,44,68,83,84,100,101]
- Strategic, agile planning (context specific) [9,21,24,28,68,84,100,101,104]
- Lack of leadership [21,24,28,44,68,83,84,101,104]
- Documentation and resources [9,24,28,68,83,84,100,101,104]
- Self-sufficiency [9,21,68,84,101]
- Communication and transparency [9,21,24,28,83,84,100,101,103]
- Trust, social cohesion, and ownership [9,21,68,83,84,101,104]
- Workflow pressure, staff morale, and ethical practices [9,21,24,28,44,84,100-102,104]
- More reactive than proactive [9,21,28,44,68,84,101,103]
- Deep-rooted habits (relative to culture or specific context) [28,44,68,83,84,100,101]

Adaptation of open-source EHR software is needed because of contextual factors such as limited resources, infrastructure, organizational setup, and workflows. LMIC settings have unique requirements. There are different user realities to consider, and

XSL•FO
RenderX

specific issues seem to determine the success or failure of open-source EHR implementations [105]. These contextual barriers have been found to include political, economic, social, and technological issues [31,106]. To better understand these barriers to implementing open-source EHRs in LMICs, one has to first acknowledge the given setting, understand end-user realities, and consider the availability of resources [107].

Socioenvironmental Barriers Impacting Open-Source EHR Adaptation and Implementation

The socioenvironmental barriers include diverse issues: resource scarcity, limited political support, and socioeconomic difficulties. These issues impact open-source EHR adaptation and implementation in LMICs. Nearly all studies [9,21,24,28,44,68,83,84,100,102-104] show that a lack of resources, such as technical expertise, power and connectivity, investment, workstations, human resource, and support, impedes successful adaptations and implementations.

The lack of effective stakeholder engagement and participation may impact long-term sustainability and health care delivery in LMICs [102]. Anantraman et al [103] state that the adaptation and implementation of open-source EHRs are best supported through the effective involvement of various stakeholders, such as end users, IT providers, clinicians, and developers [21]. Moreover, Akanbi et al [44], Syzdykova et al [24], and Muinga et al [21] highlighted the importance of government intervention in supporting open-source EHR projects, both before implementation and postimplementation. This is confirmed by Raut et al [104]. They describe the success of 2 open-source EHR implementations in rural Nepal, primarily because of the commitment, cohesion, and support provided by the government of Nepal (at the local and regional level) as a key stakeholder.

Jawhari et al [84] reported the importance of addressing social challenges and health inequalities within communities considered marginalized in urban-poor contexts as a key issue, but it is one that is often overlooked in adaptation and implementation. Moreover, how implementers enact the implementation process and how they respond to the perceived challenges in local settings remain unclear. An important finding by Jawhari et al [84] illustrates that socially constructed stigmas associated culturally with certain diseases can impact the acceptance and effective use of open-source EHRs in slums and other urban-poor settings, where some patients use multiple identities or show resistance to their data being recorded electronically. Resistance from patients may be related to “general mistrust of systems that might track identities” [84], as often these individuals reside in insecure tenures and lack rights and legal status. These issues are perceived to be influenced simultaneously by the *technological barriers*, such as poor patient-centered design and data privacy and confidentiality issues, and *organizational barriers*, such as lack of trust, communication issues, limited funds for context-specific implementations, and maintenance support [68].

Technological Barriers Impacting Open-Source EHR Adaptation and Implementation

Our review suggests that technological barriers include power outages, network failure, interoperability, hardware suitability, data privacy, and system sustainability [9,68,83,100]. Data security, privacy, and confidentiality issues emerged as a critical need in terms of addressing adequate security features before adaptation and implementation, with concerns that patient data are too broadly accessible to health care professionals not directly involved in a patient’s care [21,84,101].

Several studies [9,21,24,28,44,84,100,101,103,104] reported on the limited interoperability of their open-source EHRs with legacy health systems and how that can hinder successful adaptations and implementations. These are additionally influenced by socioenvironmental (lack of resources and stakeholder intervention) and organizational (lack of expertise and finance) barriers [68]. Unreliable infrastructure at local and regional levels is perceived to be a major challenge encountered in urban and rural settings, often because of limited funding, poor stakeholder intervention, and limitations of key resources [83,101], issues that are perceived to be influenced by socioenvironmental and organizational barriers. Moreover, there is little information on how implementers address the identified challenges responsively.

Organizational Barriers Impacting Open-Source EHR Adaptation and Implementation

There are broad issues, particularly relating to the management of organizations and human resources, that are categorized within the organizational barriers theme. For example, finance, human resource development, strategic planning, change management, data ownership, social cohesion, trust, and ethical practice influence open-source EHR adaptation and implementation in LMICs. The ethical issues intersect across multiple perspectives. From an organizational perspective, ethical issues comprise ownership, trust, management, and organizational culture [9,21,101]. From a technological perspective, ethical issues raise concerns for data privacy and confidentiality, protection, and infrastructure [24,100]. From a socioenvironmental perspective, the ethical issues emphasize sustainability and context, stakeholder intervention, and socioeconomic factors [68,84].

Several studies identified hybrid interventions where open-source EHR systems were used alongside paper processes and described how this approach appeared to impede successful adaptations and implementations [28,44,84,100]. This may be interpreted in terms of inadequate stakeholder intervention, unreliable local and regional infrastructure, lack of proficiency and training, and inherent habits impeding open-source EHR adaptation and implementation [68,84,104]. The practice of such hybrid interventions lowers the true value of open-source EHR success and constructs a dysfunctional organizational environment in LMIC settings, overburdening staff, increasing workload pressure, hindering morale, and influencing resistance to change [28,83,100]. Furthermore, there is limited information on how exactly open-source EHR systems are adapted and implemented in LMICs and the roles of implementers in the implementation process.

From a socioenvironmental and technological perspective, finance can be seen as an important factor for resources and sustainability. However, from an organizational perspective, finance is perceived to be a major challenge that impacts the

XSL•FO
RenderX

adaptation and implementation of open-source EHR systems in LMICs. This can be observed in terms of the inability to source proficient software developers and IT providers, inadequate staff training and support, limited funding from potential stakeholders, and overreliance on funders [9,44,101]. The lack of adequate training and support from stakeholders, such as implementers, developers, IT providers, funders, and government, can adversely influence open-source EHR ownership [21,83]. This is understood in terms of inefficient organizational responsibility of system management, quality data, and conflicting stakeholder relations: the lack of leadership, inability to take responsibility, risking patient data, raising confidentiality concerns, poor communication, and trust leading to an insecure organizational environment [28,68,84,100]. End users report a lack of leadership, motivation, and communication and suggest the need for a reliable organizational culture and human resources that provide adequate support and training [101].

Discussion

Principal Findings

This scoping review revealed thematic sets of socioenvironmental, technological, and organizational barriers to the adaptation and implementation of open-source EHR systems in LMICs. Specific issues, which were mentioned often, include organizational and management, limited funding, local and regional infrastructure, data privacy, confidentiality and protection, ownership, and sustainability, which appear to influence the adaptation and implementation of open-source EHRs in LMICs [9,21,83,101]. From an organizational and management perspective, data protection and confidentiality, ownership, and ethics emerge as important issues that are context dependent. The process of implementation is a key topic to explore because it is an issue often overshadowed by technical processes, with less emphasis on the social perspective [30,31,40,68], which requires the application of a relational lens to gain an in-depth understanding [108]. There is a need for a holistic understanding to explore how implementing organizations in LMIC settings addresses socioenvironmental, technical, and organizational barriers.

This scoping review shows that shortfalls in funding, leadership, and organizational and human resources also give rise to serious data protection and confidentiality issues [44,68,84,101]. It has been argued that there is a need for implementing organizations to develop self-sufficiency and take responsibility for data protection and ownership, establish local support and training initiatives, and build relations with key stakeholders for long-term sustainability [102,104]. This notion shows that implementing organizations could better engage with key stakeholders and develop better ethical practices in terms of taking responsibility, managing organizational culture, and implementing change responsibly. Furthermore, the adaptation and implementation of open-source EHR systems can be supported through collaborative actions such as ensuring local sites are sufficiently equipped with access to power, considering data protection and confidentiality, ensuring that networks and workstations are integrated efficiently, and considering alternative measures in case of emergencies [21,28,84]. Nevertheless, strategic and agile organizational planning are also perceived as essential for successful implementations in LMICs.

There is a need to examine the given context; encourage meaningful change; involve end users such as clinicians, health facility administrators, and patients in implementation design; and collaborate with a diverse range of stakeholders cohesively for successful implementations [21,68,84,101]. Key stakeholders, particularly government, need to collaborate, support, and develop sustainable context-specific open-source EHR implementations that offer robust functionality. Unreliable technical infrastructure at the local and regional levels can have negative implications for open-source EHR adaptations and implementations in LMICs. Open-source EHR systems can enable better health care access, have a positive impact on medical record quality, and enhance the potential of patient care [23,77]. However, the challenge with adapting and implementing open-source EHR systems is finding local expertise, technical skills, and sophisticated support for implementation and maintenance, which are key ingredients needed to make it work meaningfully in LMIC settings [4,15,17,83]. Therefore, there is a disconnect between the understanding of the context-specific barriers, the implementation process, and stakeholder relations.

These findings show that, depending on the thematic perspective adopted, different issues stand out. Each of the 3 perspectives brings into view some issues and obscures others. Hence, a holistic view, inclusive of all 3 perspectives, is needed to better understand the implementation process and how challenges are addressed locally. Context-specific barriers and issues in open-source EHR implementations can be better understood through the voices of key stakeholders (implementers and end users) on the ground [21,68,83,101,106].

Some challenges need to be addressed systematically, such as considering the socioenvironmental and organizational perspectives, understanding change from a reflexive perspective, conducting responsible operations, integrating values, and including stakeholders in design before adapting and implementing open-source EHR systems in LMICs [84,103]. For example, Oza et al [28] highlight that adapting and implementing an open-source EHR system during a health emergency, the Ebola outbreak, was a major problem as the outbreak (fortunately) started to decline, which limited the usefulness of the system (postimplementation) in a resource-scarce environment. Furthermore, designing open-source EHR systems while epidemic cases are increasing is not a sustainable action; intervention needs to be planned earlier and in retrospect to address the true value of what open-source EHR systems can offer in health emergencies [28]. This perception shows that different socioenvironmental barriers to adaptation and implementation can arise, depending on the situation. The findings indicate that understanding and collectively addressing the perceived contextual barriers in LMICs before implementation is of paramount importance. There is a gap in our understanding of how open-source EHR systems are implemented and the importance of implementers in the implementation process.

XSL•FO
RenderX

Literature, which was not included in the study selection but was considered useful, provided supplementary information. For example, it indicates the presence of a wider discourse on how stakeholders need to reflect on ownership, data confidentiality, protection of information, patient dignity, and addressing unethical operations [15,30,37,39,63]. Were and Meslin [31] contend that many ethical frameworks exist (relating to issues of research ethics); however, there are no such frameworks that evaluate how well open-source EHR implementations address ethical issues in LMIC settings. While the level of discussion in high-income countries focuses on "privacy, confidentiality, data security, informed consent, data ownership, and secondary use of data" [31], little research has been conducted to address similar ethical issues existing with open-source EHR implementations in LMICs. Therefore, if the open-source EHR implementation process is better understood, it could help implementers address implementation readiness issues effectively and improve outcomes for LMICs.

There is also the issue of being responsible and accountable for data quality and negotiating during implementation. This points to the broader issue of whether responsibility is socially assumed or coconstructed. Manders-Huits [40] highlights that "[health] technologies can promote or undermine specific human values...technology shapes our practices and institutions in important ways." This shows that developers and implementers, as key stakeholders, have a level of responsibility in acting purposefully in a role to embed human values and ethical considerations within open-source EHR system adaptations and implementations. There is potential for research to explore the roles of local implementing organizations that play a fundamental role in adopting open-source EHRs on the ground.

The findings in this review highlight that more funding is required to achieve better EHR implementation readiness in LMIC settings [9,68,84,101]. There are several eHealth readiness assessment (eHRA) frameworks and related tools within the existing literature that have identified 8 readiness types: organizational, technology and infrastructure, health care providers, engagement, social, core, government, and public or patient [109-111]. The 3 thematic barriers and the subcategories identified in Textbox 2 resonate with some aspects of the 8 eHRA types. However, existing eHRA frameworks are found to be inadequate to support implementation readiness in the context of LMIC settings [112-114]. It is key to better understand the implementation process from a holistic perspective. There are limitations when one looks at adaptation and implementation from one of the 3 perspectives (socioenvironmental, technological, and organizational barriers). Hence, implementation needs to be clearly understood holistically, through implementers' perspectives, to better navigate barriers encountered in LMICs.

Strengths and Limitations

The results of this review may be limited because only studies published in English were included, and hence, they are subject to cultural selection bias. Nevertheless, no other restrictions were placed on the study design or the publication format to maximize the search results. The identified key search terms, LMIC, open source, EHR, and adaptation and implementation, were defined before the search to ensure consistency across the range of databases. In addition, using an inclusion and exclusion algorithm to screen titles, abstracts, and full-text reviews helped to ensure consistency. This scoping review identified a small number of helpful studies exploring the contextual barriers impacting open-source EHR implementations in LMICs. However, existing research does not examine how implementers understand and navigate the implementation process more closely and how they respond to barriers in a given context. Therefore, the identified opportunity in the literature highlights a need to conduct further research in this area.

We acknowledge that the use of picture archiving and communication systems in LMIC settings could be a useful area for future research. Furthermore, research opportunities could explore the acceptability of international standards and compare alternatives such as Health Level Seven, Fast Health care Interoperability Resources, and Digital Imaging and Communications in Medicine in LMIC settings. The comparison of open-source EHRs with web applications and mobile apps or focusing evaluation research is also a useful direction for future research.

Conclusions

Open-source EHRs have the potential to facilitate enhanced health care and encourage sustainable development in LMICs, where designed effectively and responsibly within country-specific requirements. This scoping review provides an overview of the contextual barriers impacting the adaptation and implementation of open-source EHR systems in LMIC settings. It shows that depending on the adopted perspective to implementation, different implementation barriers come into view. A dominant focus on technology distracts from socioenvironmental and organizational barriers impacting the proliferation of open-source EHRs. Each of the 3 perspectives (socioenvironmental, technological, and organizational) draws attention to key implementation issues and highlights the important role implementers may play in addressing these issues. However, by itself, none of these 3 perspectives enable us to appreciate more fully the many interlocking challenges associated with implementing open-source EHRs in LMIC settings.

It is vital to consider the more specific context in which open-source EHRs are to be adopted and to address the need for effective implementation through a better understanding and collaboration with all stakeholders. A lack of empirical evidence limits our understanding of how exactly open-source EHR systems are adapted and implemented. Research is required to explore the roles of local implementing organizations in addressing implementation barriers in LMIC settings. A holistic understanding of implementers' experiences of implementation processes is needed. This could help characterize and solve implementation problems, including those related to ethics and the management of data protection. Nevertheless, this scoping review provides a meaningful contribution to the global health informatics discipline. We hope that the review results will inform areas for future research and enhance implementation.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge and appreciate the support from the OpenMRS community and Dr Hamish Fraser, who provided an opportunity to present and discuss the preliminary findings of this review at the 2019 OpenMRS Implementers' Conference (December 6, 2019) in Maputo, Mozambique.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Identification of the key terms and synonyms using the population, intervention, and outcome framework.
[DOCX File , 19 KB - jmir_v26i1e45242_app1.docx]

Multimedia Appendix 2
Search terms and search strategy for the scoping review.
[DOCX File , 29 KB - jmir_v26i1e45242_app2.docx]

Multimedia Appendix 3
List of low- and lower middle–income countries.
[DOCX File , 25 KB - jmir_v26i1e45242_app3.docx]

Multimedia Appendix 4
The inclusion and exclusion algorithm for screening articles.
[DOCX File , 35 KB - jmir_v26i1e45242_app4.docx]

Multimedia Appendix 5
Characteristics of the included studies.
[DOCX File , 57 KB - jmir_v26i1e45242_app5.docx]

Multimedia Appendix 6
Quality assessment tools table.
[XLSX File (Microsoft Excel File), 19 KB - jmir_v26i1e45242_app6.xlsx]

Multimedia Appendix 7
Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.
[DOCX File , 85 KB - jmir_v26i1e45242_app7.docx]

References

1. Puchalski Ritchie LM, Khan S, Moore JE, Timmings C, van Lettow M, Vogel JP, et al. Low- and middle-income countries face many common barriers to implementation of maternal health evidence products. J Clin Epidemiol 2016 Aug;76:229-237 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.02.017] [Medline: 26931284]
2. Shuvo T, Islam R, Hossain S, Evans JL, Khatun F, Ahmed T, et al. eHealth innovations in LMICs of Africa and Asia: a literature review exploring factors affecting implementation, scale-up, and sustainability. Innov Entrep Health 2015 Oct;2:95-106 [FREE Full text] [doi: 10.2147/ieh.s88809]
3. World Bank country and lending groups. World Bank. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2021-09-01]
4. Jawhari B, Ludwick D, Keenan L, Zakus D, Hayward R. Benefits and challenges of EMR implementations in low resource settings: a state-of-the-art review. BMC Med Inform Decis Mak 2016 Sep 06;16(1):116 [FREE Full text] [doi: 10.1186/s12911-016-0354-8] [Medline: 27600269]
5. Mbuthia D, Molyneux S, Njue M, Mwalukore S, Marsh V. Kenyan health stakeholder views on individual consent, general notification and governance processes for the re-use of hospital inpatient data to support learning on healthcare systems. BMC Med Ethics 2019 Jan 08;20(1):3 [FREE Full text] [doi: 10.1186/s12910-018-0343-9] [Medline: 30621693]

XSL•FO
RenderX

6. Indicators of urban poor communities and their accessibility. The United Nations Human Settlements Programme (UN-Habitat). 2015. URL: https://unhabitat.org/indicators-of-urban-poor-communities-and-their-accessibility [accessed 2024-04-29]
7. Mahabir R, Crooks A, Croitoru A, Agouris P. The study of slums as social and physical constructs: challenges and emerging research opportunities. Reg Stud Reg Sci 2016 Sep 22;3(1):399-419. [doi: 10.1080/21681376.2016.1229130]
8. Lilford RJ, Oyebode O, Satterthwaite D, Melendez-Torres GJ, Chen YF, Mberu B, et al. Improving the health and welfare of people who live in slums. Lancet 2017 Feb 04;389(10068):559-570. [doi: 10.1016/S0140-6736(16)31848-7] [Medline: 27760702]
9. Mohammed-Rajput NA, Smith DC, Mamlin B, Biondich P, Doebbeling BN, Open MRS Collaborative Investigators. OpenMRS, a global medical records system collaborative: factors influencing successful implementation. AMIA Annu Symp Proc 2011;2011:960-968 [FREE Full text] [Medline: 22195155]
10. Ellis JA. Leveraging mobile phones for monitoring risks for noncommunicable diseases in the future. J Med Internet Res 2017 May 05;19(5):e137 [FREE Full text] [doi: 10.2196/jmir.7622] [Medline: 28476721]
11. McMaster D, Clare G. Designing a mobile eye hospital to support health systems in resource-scarce environments. Disaster Med Public Health Prep 2022 Aug 14;16(4):1618-1622. [doi: 10.1017/dmp.2021.108] [Medline: 33988494]
12. Peters DH, Garg A, Bloom G, Walker DG, Brieger WR, Rahman MH. Poverty and access to health care in developing countries. Ann N Y Acad Sci 2008;1136:161-171. [doi: 10.1196/annals.1425.011] [Medline: 17954679]
13. Gibson DG, Pariyo GW, Wosu AC, Greenleaf AR, Ali J, Ahmed S, et al. Evaluation of mechanisms to improve performance of mobile phone surveys in low- and middle-income countries: research protocol. JMIR Res Protoc 2017 May 05;6(5):e81 [FREE Full text] [doi: 10.2196/resprot.7534] [Medline: 28476729]
14. Stallman R. ViewpointWhy "open source" misses the point of free software. Commun ACM 2009 Jun;52(6):31-33. [doi: 10.1145/1516046.1516058]
15. Reynolds CJ, Wyatt JC. Open source, open standards, and health care information systems. J Med Internet Res 2011 Feb 17;13(1):e24 [FREE Full text] [doi: 10.2196/jmir.1521] [Medline: 21447469]
16. Dhir S, Dhir S. Adoption of open‐source software versus proprietary software: an exploratory study. Strateg Change 2017 Jul 24;26(4):363-371. [doi: 10.1002/jsc.2137]
17. Blaya J, Otzoy D. Buy, build, or adapt – how to decide? A guide to open source electronic health records (EHRs). Inter-American Development Bank. 2019. URL: https://socialdigital.iadb.org/en/sph/resources/research-publications/838 [accessed 2024-04-29]
18. Maglogiannis I. Towards the adoption of open source and open access electronic health record systems. J Healthc Eng 2000 Jan;3(1):141-161. [doi: 10.1260/2040-2295.3.1.141]
19. Coiera E. Guide to Health Informatics, 3rd Edition. Boca Raton, FL: CRC Press; 2015.
20. Atherton J. Development of the electronic health record. Virtual Mentor 2011 Mar 01;13(3):186-189 [FREE Full text] [doi: 10.1001/virtualmentor.2011.13.3.mhst1-1103] [Medline: 23127323]
21. Muinga N, Magare S, Monda J, Kamau O, Houston S, Fraser HS, et al. Implementing an open source electronic health record system in Kenyan health care facilities: case study. JMIR Med Inform 2018 Apr 18;6(2):e22 [FREE Full text] [doi: 10.2196/medinform.8403] [Medline: 29669709]
22. Fraser HS, Blaya J. Implementing medical information systems in developing countries, what works and what doesn't. AMIA Annu Symp Proc 2010 Nov 13;2010:232-236 [FREE Full text] [Medline: 21346975]
23. Blaya JA, Fraser HS, Holt B. E-health technologies show promise in developing countries. Health Aff (Millwood) 2010 Feb;29(2):244-251. [doi: 10.1377/hlthaff.2009.0894] [Medline: 20348068]
24. Syzdykova A, Malta A, Zolfo M, Diro E, Oliveira JL. Open-source electronic health record systems for low-resource settings: systematic review. JMIR Med Inform 2017 Nov 13;5(4):e44 [FREE Full text] [doi: 10.2196/medinform.8131] [Medline: 29133283]
25. Mamlin BW, Biondich PG, Wolfe BA, Fraser HS, Jazayeri D, Allen C, et al. Cooking up an open source EMR for developing countries: OpenMRS - a recipe for successful collaboration. AMIA Annu Symp Proc 2006;2006:529-533 [FREE Full text] [Medline: 17238397]
26. Dinesen B, Nonnecke B, Lindeman D, Toft E, Kidholm K, Jethwani K, et al. Personalized telehealth in the future: a global research agenda. J Med Internet Res 2016 Mar 01;18(3):e53 [FREE Full text] [doi: 10.2196/jmir.5257] [Medline: 26932229]
27. Williams F, Boren SA. The role of electronic medical record in care delivery in developing countries. Int J Inf Manage 2008 Dec;28(6):503-507 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2008.01.016] [Medline: 30774175]
28. Oza S, Jazayeri D, Teich JM, Ball E, Nankubuge PA, Rwebembera J, et al. Development and deployment of the OpenMRS-Ebola electronic health record system for an Ebola treatment center in Sierra Leone. J Med Internet Res 2017 Aug 21;19(8):e294 [FREE Full text] [doi: 10.2196/jmir.7881] [Medline: 28827211]
29. Grodzinsky FS, Miller K, Wolf MJ. Ethical issues in open source software. J Inf Commun Ethics Soc 2003;1(4):193-205. [doi: 10.1108/14779960380000235]
30. Spriggs M, Arnold MV, Pearce CM, Fry C. Ethical questions must be considered for electronic health records. J Med Ethics 2012 Sep;38(9):535-539. [doi: 10.1136/medethics-2011-100413] [Medline: 22573881]

XSL•FO
RenderX

31. Were MC, Meslin EM. Ethics of implementing electronic health records in developing countries: points to consider. AMIA Annu Symp Proc 2011;2011:1499-1505 [FREE Full text] [Medline: 22195214]
32. About the sustainable development goals. United Nations. 2019. URL: https://www.un.org/sustainabledevelopment/sustainable-development-goals/ [accessed 2024-04-29]
33. National eHealth strategy toolkit. World Health Organization and International Telecommunication Union. 2012. URL: https://apps.who.int/iris/bitstream/handle/10665/75211/9789241548465_eng.pdf?sequence=1&isAllowed=y [accessed 2024-04-29]
34. Patient engagement: technical series on safer primary care. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511629 [accessed 2024-04-29]
35. Electronic health records: manual for developing countries. World Health Organization. 2006. URL: https://apps.who.int/iris/bitstream/handle/10665/207504/9290612177_eng.pdf?sequence=1&isAllowed=y [accessed 2024-04-29]
36. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
37. Barrows RC, Clayton PD. Privacy, confidentiality, and electronic medical records. J Am Med Inform Assoc 1996 Mar 01;3(2):139-148 [FREE Full text] [doi: 10.1136/jamia.1996.96236282] [Medline: 8653450]
38. Amato F, Mazzocca N, De Pietro G, Esposito M. A system for semantic-based access control. In: Proceedings of the 2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. 2013 Presented at: 3PGCIC '13; October 28-30, 2013; Compiegne, France p. 442-446 URL: https://ieeexplore.ieee.org/document/6681268 [doi: 10.1109/3pgcic.2013.74]
39. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2018 Mar 06;25(S 01):S48-S61. [doi: 10.15265/iys-2016-s006]
40. Manders-Huits N. What values in design? The challenge of incorporating moral values into design. Sci Eng Ethics 2011 Jun 12;17(2):271-287 [FREE Full text] [doi: 10.1007/s11948-010-9198-2] [Medline: 20224927]
41. Webster PC. The rise of open-source electronic health records. Lancet 2011 May 14;377(9778):1641-1642. [doi: 10.1016/s0140-6736(11)60659-4] [Medline: 21591284]
42. The open source definition (annotated). Opensource. 2007. URL: https://opensource.org/osd.html [accessed 2021-09-10]
43. Cowan C. Software security for open-source systems. IEEE Secur Privacy 2003 Jan;1(1):38-45. [doi: 10.1109/msecp.2003.1176994]
44. Akanbi MO, Ocheke AN, Agaba PA, Daniyam CA, Agaba EI, Okeke EN, et al. Use of electronic health records in sub-Saharan Africa: progress and challenges. J Med Trop 2012;14(1):1-6 [FREE Full text] [Medline: 25243111]
45. Curto-Millet D, Shaikh M. The emergence of openness in open-source projects: the case of OpenEhR. J Inf Technol 2017 Dec 01;32(4):361-379. [doi: 10.1057/s41265-017-0042-x]
46. Raymond E. The cathedral and the bazaar. Knowl Technol Policy 1999 Sep;12(3):23-49. [doi: 10.1007/S12130-999-1026-0]
47. What is OpenELIS global? OpenELIS. 2021. URL: https://openelis-global.org/about/ [accessed 2021-10-22]
48. About DHIS2. DHIS2. URL: https://dhis2.org/about/ [accessed 2021-10-22]
49. Welcome to GNU health!. GNU Health. 2020. URL: https://www.gnuhealth.org/about-us.html [accessed 2021-10-22]
50. OpenMRS releases 2015 annual report. OpenMRS Inc. 2015. URL: https://openmrs.org/openmrs-releases-2015-annual-report/ [accessed 2024-04-29]
51. Seebregts CJ, Mamlin BW, Biondich PG, Fraser HS, Wolfe BA, Jazayeri D, OpenMRS Implementers Network. The OpenMRS implementers network. Int J Med Inform 2009 Nov;78(11):711-720. [doi: 10.1016/j.ijmedinf.2008.09.005] [Medline: 19157968]
52. Open medical record system. OpenMRS Inc. URL: https://openmrs.org [accessed 2021-08-20]
53. Introduction to OpenMRS. OpenMRS Wiki. 2021. URL: https://wiki.openmrs.org/display/docs/Introduction+to+OpenMRS [accessed 2024-04-29]
54. Atlas. Where we are: OpenMRS Atlas. OpenMRS Inc. 2022. URL: https://atlas.openmrs.org/ [accessed 2022-01-21]
55. Introduction: what is OpenMRS 3? OpenMRS Wiki. URL: https://wiki.openmrs.org/display/projects/Designer+Onboarding+to+OpenMRS+3#DesignerOnboardingtoOpenMRS3-VisualHistoryoftheOpenMRSFrontend [accessed 2022-04-18]
56. Who we are: OpenMRS community. OpenMRS Inc. URL: https://openmrs.org/who-we-are/ [accessed 2022-01-21]
57. OpenMRS community. OpenMRS Inc. 2022. URL: https://wiki.openmrs.org/display/RES/OpenMRS+Community [accessed 2024-04-29]
58. Guide for the new and curious. OpenMRS Inc. URL: https://wiki.openmrs.org/display/docs/Guide+for+the+New+and+Curious [accessed 2022-04-22]
59. Thengo L. “write code! save lives!”. Medium. URL: https://medium.com/amplify/write-code-save-lives-784a3a6abd09 [accessed 2024-04-29]
60. McDonald CJ, Schadow G, Barnes M, Dexter P, Overhage JM, Mamlin B, et al. Open source software in medical informatics--why, how and what. Int J Med Inform 2003 Mar;69(2-3):175-184 [FREE Full text] [doi: 10.1016/s1386-5056(02)00104-1] [Medline: 12810121]

XSL•FO
RenderX

61. Alsaffar M, Yellowlees P, Odor A, Hogarth M. The state of open source electronic health record projects: a software anthropology study. JMIR Med Inform 2017 Feb 24;5(1):e6 [FREE Full text] [doi: 10.2196/medinform.5783] [Medline: 28235750]
62. Bagayoko CO, Dufour JC, Chaacho S, Bouhaddou O, Fieschi M. Open source challenges for hospital information system (HIS) in developing countries: a pilot project in Mali. BMC Med Inform Decis Mak 2010 Apr 16;10(1):22 [FREE Full text] [doi: 10.1186/1472-6947-10-22] [Medline: 20398366]
63. Flores Zuniga AE, Win KT, Susilo W. Functionalities of free and open electronic health record systems. Int J Technol Assess Health Care 2010 Oct;26(4):382-389. [doi: 10.1017/S0266462310001121] [Medline: 20974022]
64. Fraser HS, Thomas D, Tomaylla J, Garcia N, Lecca L, Murray M, et al. Adaptation of a web-based, open source electronic medical record system platform to support a large study of tuberculosis epidemiology. BMC Med Inform Decis Mak 2012 Nov 07;12(1):125 [FREE Full text] [doi: 10.1186/1472-6947-12-125] [Medline: 23131180]
65. Millard PS, Bru J, Berger CA. Open-source point-of-care electronic medical records for use in resource-limited settings: systematic review and questionnaire surveys. BMJ Open 2012;2(4):e000690 [FREE Full text] [doi: 10.1136/bmjopen-2011-000690] [Medline: 22763661]
66. Singh A, Bansal R, Jha N. Open source software vs proprietary software. Int J Comput Appl 2015 Mar 18;114(18):26-31. [doi: 10.5120/20080-2132]
67. Gropper A. Open-source health care software. Virtual Mentor 2011 Sep 01;13(9):632-636 [FREE Full text] [doi: 10.1001/virtualmentor.2011.13.9.stas1-1109] [Medline: 23137477]
68. Fish KE, Guha GS. Towards an electronic medical record system for a rural Haitian medical clinic. J Strateg Innov Sustain 2020 Aug 11;15(4):15-20. [doi: 10.33423/jsis.v15i4.2957]
69. Boulanger A. Open-source versus proprietary software: is one more reliable and secure than the other? IBM Syst J 2005;44(2):239-248. [doi: 10.1147/sj.442.0239]
70. Aminpour F, Sadoughi F, Ahamdi M. Utilization of open source electronic health record around the world: a systematic review. J Res Med Sci 2014 Jan;19(1):57-64 [FREE Full text] [Medline: 24672566]
71. Holck J, Larsen MH, Pedersen MK. Managerial and technical barriers to the adoption of open source software. In: Proceedings of the 4th International Conference on COTS-Based Software Systems. 2005 Presented at: ICCBSS '05; February 7-11, 200; Bilbao, Spain p. 289-300 URL: https://link.springer.com/chapter/10.1007/978-3-540-30587-3_38 [doi: 10.1007/978-3-540-30587-3_38]
72. Brown AW, Booch G. Reusing open-source software and practices: the impact of open-source on commercial vendors. In: Proceedings of the 7th International Conference on Software Reuse: Methods, Techniques, and Tools. 2002 Presented at: ICSR '02; April 15-19, 2002; Austin, TX p. 123-136 URL: https://link.springer.com/chapter/10.1007/3-540-46020-9_9 [doi: 10.1007/3-540-46020-9_9]
73. Canonical and Ubuntu. Ubuntu. 2021. URL: https://ubuntu.com/community/canonical [accessed 2021-09-10]
74. Shaw N, Pepper D, Cook T, Houwink P, Jain N, Bainbridge M. Open source and international health informatics: placebo or panacea? J Innov Health Inform 2002;10(1):39-43. [doi: 10.14236/jhi.v10i1.240]
75. Edwards K. Epistemic communities, situated learning and open source software development. In: Proceedings from the 2001 Conference on Epistemic Cultures and the Practice of Interdisciplinarity. 2001 Presented at: ECPI '01; January 1-4, 2001; Trondheim, Norway p. 1-33 URL: https://backend.orbit.dtu.dk/ws/portalfiles/portal/2976336/Epistemic_com_Sit_Learn+and+OSS+v074.pdf
76. Jin L, Robey D, Boudreau MC. Beyond development: a research agenda for investigating open source software user communities. Inf Resour Manag J 2007;20(1):68-80. [doi: 10.4018/irmj.2007010105]
77. Purkayastha S, Allam R, Maity P, Gichoya JW. Comparison of open-source electronic health record systems based on functional and user performance criteria. Healthc Inform Res 2019 Apr;25(2):89-98 [FREE Full text] [doi: 10.4258/hir.2019.25.2.89] [Medline: 31131143]
78. Allen C, Jazayeri D, Miranda J, Biondich PG, Mamlin BW, Wolfe BA, et al. Experience in implementing the OpenMRS medical record system to support HIV treatment in Rwanda. Stud Health Technol Inform 2007;129(Pt 1):382-386. [Medline: 17911744]
79. Pfadenhauer LM, Mozygemba K, Gerhardus A, Hofmann B, Booth A, Lysdahl KB, et al. Context and implementation: a concept analysis towards conceptual maturity. Z Evid Fortbild Qual Gesundhwes 2015;109(2):103-114 [FREE Full text] [doi: 10.1016/j.zefq.2015.01.004] [Medline: 26028447]
80. Zaidan AA, Zaidan BB, Al-Haiqi A, Kiah ML, Hussain M, Abdulnabi M. Evaluation and selection of open-source EMR software packages based on integrated AHP and TOPSIS. J Biomed Inform 2015 Feb;53:390-404 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.012] [Medline: 25483886]
81. Akhlaq A, McKinstry B, Muhammad KB, Sheikh A. Barriers and facilitators to health information exchange in low- and middle-income country settings: a systematic review. Health Policy Plan 2016 Nov 16;31(9):1310-1325. [doi: 10.1093/heapol/czw056] [Medline: 27185528]
82. Implementer documentation. OpenMRS Inc. 2017. URL: https://wiki.openmrs.org/display/docs/Implementer+Documentation#ImplementerDocumentation-Overview [accessed 2019-07-25]

XSL•FO
RenderX

83. Verma N, Mamlin B, Flowers J, Acharya S, Labrique A, Cullen T. OpenMRS as a global good: impact, opportunities, challenges, and lessons learned from fifteen years of implementation. Int J Med Inform 2021 May;149:104405. [doi: 10.1016/j.ijmedinf.2021.104405] [Medline: 33639327]
84. Jawhari B, Keenan L, Zakus D, Ludwick D, Isaac A, Saleh A, et al. Barriers and facilitators to electronic medical record (EMR) use in an urban slum. Int J Med Inform 2016 Oct;94:246-254 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.07.015] [Medline: 27573333]
85. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009 Jun 27;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
86. Peterson J, Pearce PF, Ferguson LA, Langford CA. Understanding scoping reviews: definition, purpose, and process. J Am Assoc Nurse Pract 2017 Jan;29(1):12-16. [doi: 10.1002/2327-6924.12380] [Medline: 27245885]
87. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
88. Peters M, Mcinerney P, Godfrey CM, Soares CB. Methodology for JBI scoping reviews. In: Micah P, Christina D, Patricia M, Cassia S, Hanan K, Deborah P, editors. The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews. Adelaide, Australia: The Joanna Briggs Institute; 2016.
89. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
90. Davis K, Drey N, Gould D. What are scoping studies? A review of the nursing literature. Int J Nurs Stud 2009 Oct;46(10):1386-1400. [doi: 10.1016/j.ijnurstu.2009.02.010] [Medline: 19328488]
91. Gough D, Oliver S, Thomas J. An Introduction to Systematic Reviews. 2nd edition. Thousand Oaks, CA: Sage Publications; 2017.
92. Rycroft-Malone J, Seers K, Chandler J, Hawkes CA, Crichton N, Allen C, et al. The role of evidence, context, and facilitation in an implementation trial: implications for the development of the PARIHS framework. Implement Sci 2013 Mar 09;8(1):28 [FREE Full text] [doi: 10.1186/1748-5908-8-28] [Medline: 23497438]
93. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
94. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
95. Saldaña J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: Sage Publications; 2015.
96. Castleberry A, Nolen A. Thematic analysis of qualitative research data: is it as easy as it sounds? Curr Pharm Teach Learn 2018 Jun;10(6):807-815. [doi: 10.1016/j.cptl.2018.03.019] [Medline: 30025784]
97. Pluye P, Hong QN. Combining the power of stories and the power of numbers: mixed methods research and mixed studies reviews. Annu Rev Public Health 2014;35:29-45. [doi: 10.1146/annurev-publhealth-032013-182440] [Medline: 24188053]
98. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Department of Family Medicine. 2018 Aug 01. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf [accessed 2024-04-29]
99. Systematic review checklist. Critical Appraisal Skills Programme. 2018. URL: https://casp-uk.b-cdn.net/wp-content/uploads/2018/01/CASP-Systematic-Review-Checklist_2018.pdf [accessed 2021-04-05]
100. Gainer A, Roth M, Strong P, Davis J. A standards-based open source application to gather health assessment data in developing countries. In: Proceedings of the 2012 IEEE Global Humanitarian Technology Conference. 2012 Presented at: GHTC '12; October 21-24, 2012; Seattle, WA p. 293-298 URL: https://ieeexplore.ieee.org/document/6387064 [doi: 10.1109/ghtc.2012.78]
101. Gyamfi A, Mensah KA, Oduro G, Donkor P, Mock CN. Barriers and facilitators to electronic medical records usage in the emergency centre at Komfo anokye teaching hospital, Kumasi-Ghana. Afr J Emerg Med 2017 Dec;7(4):177-182 [FREE Full text] [doi: 10.1016/j.afjem.2017.05.002] [Medline: 30456135]
102. Were MC, Emenyonu N, Achieng M, Shen C, Ssali J, Masaba JP, et al. Evaluating a scalable model for implementing electronic health records in resource-limited settings. J Am Med Inform Assoc 2010;17(3):237-244 [FREE Full text] [doi: 10.1136/jamia.2009.002303] [Medline: 20442140]
103. Anantraman V, Mikkelsen T, Khilnani R, Kumar VS, Pentland A, Ohno-Machado L. Open source handheld-based EMR for paramedics working in rural areas. Proc AMIA Symp 2002:12-16 [FREE Full text] [Medline: 12463777]
104. Raut A, Yarbrough C, Singh V, Gauchan B, Citrin D, Verma V, et al. Design and implementation of an affordable, public sector electronic medical record in rural Nepal. J Innov Health Inform 2017 Jun 23;24(2):862 [FREE Full text] [doi: 10.14236/jhi.v24i2.862] [Medline: 28749321]
105. Hawari AA, Heeks R. Explaining ERP failure in a developing country: a Jordanian case study. J Enterp Inf Manag 2010;23(2):135-160 [FREE Full text] [doi: 10.1108/17410391011019741]

XSL•FO
RenderX

106. Odekunle FF, Odekunle RO, Shankar S. Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. Int J Health Sci (Qassim) 2017;11(4):59-64 [FREE Full text] [Medline: 29085270]
107. Heeks R. Health information systems: failure, success and improvisation. Int J Med Inform 2006 Feb;75(2):125-137. [doi: 10.1016/j.ijmedinf.2005.07.024] [Medline: 16112893]
108. Dépelteau F. The Palgrave Handbook of Relational Sociology. Cham, Switzerland: Springer; 2018.
109. Khoja S, Scott RE, Casebeer AL, Mohsin M, Ishaq AF, Gilani S. e-Health readiness assessment tools for healthcare institutions in developing countries. Telemed J E Health 2007 Aug;13(4):425-431. [doi: 10.1089/tmj.2006.0064] [Medline: 17848110]
110. Rezai-Rad M, Vaezi R, Nattagh F. E-health readiness assessment framework in iran. Iran J Public Health 2012;41(10):43-51 [FREE Full text] [Medline: 23304661]
111. Mauco KL, Scott RE, Mars M. Critical analysis of e-health readiness assessment frameworks: suitability for application in developing countries. J Telemed Telecare 2018 Feb;24(2):110-117. [doi: 10.1177/1357633X16686548] [Medline: 28008790]
112. Oliver DR, Demiris G. An assessment of the readiness of hospice organizations to accept technological innovation. J Telemed Telecare 2004;10(3):170-174. [doi: 10.1258/135763304323070832] [Medline: 15165444]
113. Yusif S, Hafeez-Baig A, Soar J. e-Health readiness assessment factors and measuring tools: a systematic review. Int J Med Inform 2017 Nov;107:56-64 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.08.006] [Medline: 29029692]
114. Mauco KL, Scott RE, Mars M. Development of an eHealth readiness assessment framework for Botswana and other developing countries: interview study. JMIR Med Inform 2019 Aug 22;7(3):e12949 [FREE Full text] [doi: 10.2196/12949] [Medline: 31441429]

Abbreviations

EHR: electronic health record
eHRA: eHealth readiness assessment
LMIC: low- and lower-middle-income country
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Mavragani; submitted 21.12.22; peer-reviewed by Y Chu, M Mars, I Mircheva; comments to author 06.03.23; revised version received 29.09.23; accepted 09.02.24; published 01.08.24.

Please cite as:
Bostan S, Johnson OA, Jaspersen LJ, Randell R
Contextual Barriers to Implementing Open-Source Electronic Health Record Systems for Low- and Lower-Middle-Income Countries: Scoping Review
J Med Internet Res 2024;26:e45242
URL: https://www.jmir.org/2024/1/e45242
doi:10.2196/45242
PMID:

©Sarah Bostan, Owen A Johnson, Lena J Jaspersen, Rebecca Randell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Maturity as a Predictor of Quality and Safety Outcomes in US Hospitals: Cross-Sectional Observational Study

Anne Snowdon[1*], BSCN, MSc, PhD; Abdulkadir Hussein[1*], PhD; Melissa Danforth[2*], BA; Alexandra Wright[1*], BPR, MPA, PhD; Reid Oakes[3*], BSc

[1]Department of Mathematics & Statistics, University of Windsor, Windsor, ON, Canada

[2]Leapfrog, Washington, DC, United States

[3]Healthcare Information and Management Systems Society, Chicago, IL, United States

[*]all authors contributed equally

Corresponding Author:
Alexandra Wright, BPR, MPA, PhD
Department of Mathematics & Statistics
University of Windsor
Windsor, ON
Canada
Phone: 1 4164001956
Email: alexandra.wright@uwindsor.ca

Abstract

Background: This study demonstrates that digital maturity contributes to strengthened quality and safety performance outcomes in US hospitals. Advanced digital maturity is associated with more digitally enabled work environments with automated flow of data across information systems to enable clinicians and leaders to track quality and safety outcomes. This research illustrates that an advanced digitally enabled workforce is associated with strong safety leadership and culture and better patient health and safety outcomes.

Objective: This study aimed to examine the relationship between digital maturity and quality and safety outcomes in US hospitals.

Methods: The data sources were hospital safety letter grades as well as quality and safety scores on a continuous scale published by The Leapfrog Group. We used the digital maturity level (measured using the Electronic Medical Record Assessment Model [EMRAM]) of 1026 US hospitals. This was a cross-sectional, observational study. Logistic, linear, and Tweedie regression analyses were used to explore the relationships among The Leapfrog Group's Hospital Safety Grades, individual Leapfrog safety scores, and digital maturity levels classified as advanced or fully developed digital maturity (EMRAM levels 6 and 7) or underdeveloped maturity (EMRAM level 0). Digital maturity was a predictor while controlling for hospital characteristics including teaching status, urban or rural location, hospital size measured by number of beds, whether the hospital was a referral center, and type of hospital ownership as confounding variables. Hospitals were divided into the following 2 groups to compare safety and quality outcomes: hospitals that were digitally advanced and hospitals with underdeveloped digital maturity. Data from The Leapfrog Group's Hospital Safety Grades report published in spring 2019 were matched to the hospitals with completed EMRAM assessments in 2019. Hospital characteristics such as number of hospital beds were obtained from the CMS database.

Results: The results revealed that the odds of achieving a higher Leapfrog Group Hospital Safety Grade was statistically significantly higher, by 3.25 times, for hospitals with advanced digital maturity (EMRAM maturity of 6 or 7; odds ratio 3.25, 95% CI 2.33-4.55).

Conclusions: Hospitals with advanced digital maturity had statistically significantly reduced infection rates, reduced adverse events, and improved surgical safety outcomes. The study findings suggest a significant difference in quality and safety outcomes among hospitals with advanced digital maturity compared with hospitals with underdeveloped digital maturity.

(J Med Internet Res 2024;26:e56316) doi:10.2196/56316

XSL•FO
RenderX

KEYWORDS

digital health; readiness; cross sectional; observational; regression; digital maturity; association; associations; correlation; correlations; quality and safety; hospital performance; workforce; health outcomes; safety; service; services; healthcare system; healthcare systems; hospital; hospitals

Introduction

Digital health is widely viewed as an opportunity to improve patient safety outcomes [1-5]. Digitally enabled health care has been proposed as an enabler for hospitals and health systems to advance and strengthen performance outcomes linked to the quadruple aim (ie, reduced costs, improved patient experience, improved workforce satisfaction, and better quality outcomes) [6-8].

Digital health technologies enable the collection of data to inform decisions to improve the quality and safety of care delivery, automate care processes to reduce errors, and improve the patient experience by engaging patients in their own health and care [1,2]. Although digital health tools and technologies have been proposed to enhance the quality and safety of health care, there is limited empirical evidence demonstrating the benefits of digital health technologies [1,2]. Specifically, research that has examined the benefits of digital health technologies on effectiveness and quality of care is skewed toward mental health services and behavioral therapies [2]. There are considerable gaps in the evidence of the impact of digital health transformation on cost, equitable care outcomes, and patient-centered care for hospitals or health systems [2].

The progress of digital transformation in health systems is often assessed using measures of digital maturity, which quantify the degree to which a health care organization has advanced digital capability, information infrastructure, and health data exchange, to provide digitally enabled health care services [9]. As health systems advance digital transformation efforts, the impact on organizational performance, workforce and workflows, and quality and safety outcomes for patients is required to document the return on investment to advance health system performance. However, evidence of the role of advanced digital maturity in strengthening these performance outcomes has not been well established empirically [10].

The progress of digital transformation in health systems has accelerated in recent years, fueled initially by incentives such as the Health Information Technology for Economic and Clinical Health (HITECH) act of 2009 [11,12]. The HITECH program provided financial incentives to US hospitals to adopt and implement electronic medical records (EMRs) with the goal of improving quality and safety of health care [11,12]. Over the life of the HITECH program, 98% of US hospitals participated in the program and received US $21.8 billion in federal incentives, and 60% of eligible office-based providers participated in the program and received US $16.2 billion in federal incentives. “Meaningful use” was defined in the HITECH program in terms of achieving required processes such as data collection requirements, adoption of secure email, and tracking the volume of messages, resulting in what some described as a “check the box” compliance program [12]. Although the adoption of EMR technologies has accelerated since the HITECH program, the actual use of advanced EMR functions such as physician order entry [13] and the use of data in clinical settings remain more limited [14].

Digital maturity model assessments have been used widely to document the adoption of EMR technologies and functions. One maturity model adopted widely in the United States is the Electronic Medical Record Assessment Model (EMRAM), developed by the Health Information Management Systems Society (HIMSS), a professional, not-for-profit organization with the mission of reforming the global health ecosystem through the power of information and technology [15]. The EMRAM model has been available in the United States for the last 15 years. Hospitals complete the online EMRAM survey of approximately 200 indicator statements that measure key dimensions of advanced EMR infrastructure and maturity including capacity for health information exchange, clinician adoption of digital tools, data privacy and security, and governance. Hospital teams receive a detailed assessment and report that indicate their level of digital maturity, strengths, and opportunities to advance their digital transformation strategy.

To date, there is clear evidence that the HITECH program accelerated EMR adoption across US health systems [13]. However, empirical research on performance outcomes such as quality, safety, and patient experience has been limited and may be influenced by the degree to which the full range of the functional features of EMR technologies has been adopted and scaled across provider organizations [16]. In addition, very few quality and safety programs in health organizations consider the use of digital technologies and associated risks of technologies used in health care, which further limits the evidence on how quality and safety outcomes are influenced by digital maturity [13]. More recently, significant attention to advances in digital transformation has emerged as an outcome of the COVID-19 pandemic, when a massive surge in demand for virtual care accelerated digitally enabled care services, which has continued to persist to varying degrees for specific patient populations [1]. Since the pandemic, quality and safety outcomes in US hospitals have reportedly declined, despite advances in digital transformation [17,18]. Since 2019, quality and safety outcomes for US hospitals have resulted in lower Leapfrog Group's Hospital Safety Grades (2023) [19]. Significant workforce shortages are common [20], which have further challenged hospitals to deliver quality and safe care. There is evidence that the adoption of EMR technologies is associated with added workload and burnout for the health workforce [21], which further underscores the importance of documenting the impact and value of advancing digital maturity on workforce, quality, and safety outcomes.

Numerous scholars and global agencies have examined quality and safety in hospitals and have documented gaps in care delivery whereby some patients receive health care services that are well below health and quality standards [19,22-26]. One in

XSL•FO
RenderX

10 patients seeking care in high-income countries are "adversely affected during care and treatment," suggesting the need for significant improvements in quality of care across the globe [25]. Patient harm has become the 14th leading cause of global disease burden [27], which has resulted in significant harm for patients and contributed to substantial health system costs. Globally, 12% of national health expenditures are the result of managing the harmful outcomes resulting from unsafe care [28], which has fueled a renewed urgency for strengthening quality and safety outcomes among health care systems as a strategy to reduce health system costs. Effective interventions associated with improvements in quality and safety of care delivery include communication between clinicians and their patients [29,30], prevention of central line–associated blood infections (CLABSI) [31-33], use of hand hygiene practices to prevent infections [34,35], quality and safety training, established leadership priorities that focus on quality and safety [36], and improvements in clinical pathways [26,37,38]. More recently, there is evidence of the link between implementation of EMRs, improved patient outcomes [39], and cost benefits [40]. Notably, EMR-based interventions optimize digitally enabled care delivery (eg, digital communication between patients and clinicians; early detection of infections using predictive algorithms) and have all been proposed to improve quality and safety outcomes, particularly for hospital settings [40]. However, empirical evidence of the contribution of advances in digital transformation toward quality and safety outcomes is limited, which presents a challenge for health system leaders to make the case for investing in digital transformation efforts to advance quality and safety outcomes. The purpose of this research was to empirically examine quality and safety outcomes in US hospitals associated with advances in digital maturity.

Methods

Ethical Considerations

An ethics review board application was not submitted for this study as it did not meet the criteria for research ethics review, as this research does not involve human subjects or human biological materials (eg, human embryos, fetuses, fetal tissue, reproductive materials, and stem cells) [41]. This research used data available in the public domain and included only organizational level data associated with hospitals in the United States.

Study Design

This research used a cross-sectional observational study design to examine the association between digital maturity and quality and safety outcomes in US hospitals. The analysis relied on organizational-level data that are publicly reported. The Leapfrog Group partnered with HIMSS on this research and provided The Leapfrog Group's Hospital Safety Grades data set from the spring 2019 report (Table 1).

The Leapfrog Group is a national, not-for-profit organization in the United States that collects, analyzes, and publishes data on safety and quality outcomes in hospitals, intended to inform decisions to strengthen quality, improve safety, and reduce cost in US health care systems [42]. Leapfrog's quality and safety measures include the Leapfrog Hospital survey; Leapfrog Hospital Safety Grade; Leapfrog Ambulatory Surgery Center Survey; and Leapfrog Value-Based Purchasing Program [42]. The Leapfrog Group administers 2 publicly reported hospital ratings programs annually. Publicly available data (eg, Centers for Medicare and Medicaid Services [CMS], the Leapfrog Hospital Survey, and secondary data sources such as the American Hospital Association's Annual Survey and IT Supplement) are weighted and then combined to produce a composite score that is published as an A, B, C, D, or F letter grade for almost 3000 hospitals in the United States [43]. Austen et al [44] further described this methodology for arriving at a composite patient safety score.

A Leapfrog Group's Hospital Safety Grade of "A" represents advanced patient safety outcomes achieved by acute care hospitals, and "F" represents safety outcomes that are lower than would be expected for an acute care hospital. The overall letter grades reported by Leapfrog are based on 28 individual quality and safety measures. In this study, we included 22 of the Leapfrog individual measures (Table 1). Variation in response rates was reported by Leapfrog for each safety measure resulting in different sample sizes for each Leapfrog measure included in this analysis (Table 1). The 2019 data set was selected for this analysis as it was the most complete data set compared with either 2020 or 2021, due to the impact of the COVID-19 pandemic on hospital participation.

XSL•FO
RenderX

Table 1. Leapfrog quality and safety variables included in the study.

Variable name (effective sample size)	Description	Analysis scale
Hospital safety grade (n=1026)	An overall letter grade based on a score calculated from 28 measures (see details in [6])	Categorical: A, B, C, D, or F
Process/structural measures		
CPOE[a] (n=911)	CPOE measures a hospital's progress toward implementing a CPOE system to reduce medication ordering errors.	Fully meets the standard (Yes=1, No=0)
BCMA[b] (n=902)	It measures (1) hospitals' progress toward implanting BCMA throughout the hospital, (2) compliance with patient and medication scans during administration, (3) use of decision support, and (4) structures to monitor and reduce workarounds.	Fully meets the standard (Yes=1, No=0)
IPS[c] (n=988)	It measures a hospital's use of intensivists in ICUs[d].	Fully meets the standard (Yes=1, No=0)
SP[e]		
SP1: Culture of Leadership Structures and Systems (n=615); SP2: Culture Measurement, Feedback & Intervention (n=615); SP4: Identification and Mitigation of Risks and Hazards (n=615); SP9: Nursing Workforce (n=615); SP19: Hand Hygiene (n=615)	These measure a hospital's progress toward implementing NQF[f]-endorsed processes and protocols to reduce and prevent adverse events.	These variables are on continuous scales that are 0-120 for SP1 and SP2, 0-100 for SP4 and SP9, and 0-60 for SP19. Most of the hospitals achieved full scores. Therefore, our unit of measurement in this analysis is whether a hospital achieved a full score (Yes=1, No=0).
Outcomes		
Hospital-acquired conditions		
FOR[g] (n=1026); FAT[h] (n=1026)	Three hospital-acquired condition measures are calculated by CMS[i] through the DRA HAC[j] Reporting Program: FOR after Surgery, Air Embolism, and FAT. Hospital-acquired condition rates are calculated by CMS based on Medicare Fee-for-Service claims.	Rate per 1000 inpatient CMS discharges on a scale of 0 to 1000
Hospital-acquired infections		
CLABSI[k] (n=759); CAUTI[l] (n=882); SSI[m]: Colon (n=687); MRSA[n] (n=641); C. Diff.[o] (n=1026)	Hospitals were required to (1) join Leapfrog's NHSN[p] group, (2) provide an accurate NHSN ID in the Profile section of the Online Survey Tool, and (3) submit Section 7 Managing Serious Errors by November 30 for the purposes of calculating the Safety Grade.	These are measured as SIRs[q] on a scale from 0 and unbounded above.
PSI[r]		
PSI 3: Pressure ulcer rate (n=1026); PSI 4: Death rate among surgical inpatients with serious treatable conditions (n=661); PSI6: Collapsed lung due medical treatment (n=1026); PSI 11: Postoperative respiratory failure rate (n=969); PSI 12: Perioperative PE/DVT[s] rate (n=1026); PSI 14: Postoperative wound dehiscence rate (n=1026); PSI 15: Unrecognized abdominopelvic accidental puncture/laceration rate (n=1026)	PSI rates are calculated by CMS based on Medicare Fee-for-Service claims and publicly reported as a rate per 1000 inpatient discharges.	Rate per 1000 CMS eligible discharges on a scale of 0 to 1000

[a]CPOE: computerized physician order entry.

[b]BCMA: bar code medication administration.

[c]IPS: intensive care unit physician staffing.

[d]ICUs: intensive care units.

[e]SP: safe practices.

[f]NQF: National Quality Forum.

[g]FOR: foreign object retained.

XSL•FO
RenderX

[h]FAT: falls and trauma.

[i]CMS: Centers for Medicare and Medicaid Services.

[j]DRA HAC: Deficit Reduction Act Hospital-Acquired Condition.

[k]CLABSI: central line–associated bloodstream infection.

[l]CAUTI: catheter-associated urinary tract infection.

[m]SSI: surgical site infection.

[n]MRSA: methicillin-resistant *Staphylococcus aureus*.

[o]C. Diff.: *Clostridioides difficile*.

[p]NHSN: National Healthcare Safety Network.

[q]SIR: standardized incidence rates.

[r]PSI: patient safety indicators.

[s]PE/DVT: pulmonary embolism/deep vein thrombosis.

The EMRAM maturity model data were made available by HIMSS for this analysis, as EMRAM has been widely available in the US market for the past 15 years. Hospitals complete an online assessment that has an embedded algorithm that generates a digital maturity level score on a 7-point scale of 0 through 7. Hospitals that score at a 6 or 7 digital maturity level are further assessed by an onsite HIMSS team to validate achievement of advanced digital maturity. Hospitals that do not meet the minimum EMRAM requirements for digital maturity are scored at level 0. For example, to achieve level 1 digital maturity, hospitals must have laboratory, imaging, pharmacy, and cardiology information systems fully integrated into the EMR to enable clinicians to access diagnostic reports and medication profiles, as well as resilience management plans to manage downtime of IT systems. For the purposes of this study, hospitals at level 0 on EMRAM were classified as having underdeveloped digital maturity. To be assessed at EMRAM level 0, data from laboratory, pharmacy, imaging, and interventional cardiology have not been integrated into the EMR, resulting in significant data silos. Underdeveloped digital maturity may indicate that the organization has not yet started their digital transformation journey or that the organization has not progressed its digital maturity beyond basic digital functionality. Hospitals assessed to have achieved stage 6 or 7 level of maturity were classified as achieving advanced digital maturity, which includes features such as ≥95% staff adoption and use of advanced EMR functions, including computerized physician order entry (CPOE), bar code medication administration (BCMA), integration of data from multiple internal and external sources and clinical settings, use of alerts to identify risk of errors, and automated digital tools to track progress of patient outcomes. Each of these advanced functions is designed to improve quality and safety, such as reduced errors related to physician orders or reduced errors during medication administration. Achievement of a level 7 EMRAM digital maturity indicates that the organization has reached advanced digital maturity. Requirements to achieve each subsequent level of digital maturity for the EMRAM model are described in Table 2. For this analysis, we focused only on hospitals at digital maturity levels of 0 (underdeveloped digital maturity) and level 6 or 7 (advanced digital maturity), as there were very few hospitals (eg, <20 hospitals) at maturity levels 1 through 5; these hospitals were excluded from the analysis due to a small sample size. The small number of hospitals scoring between levels 1 and 5 is viewed as the outcome of a number of factors, including lack of financial incentives to measure or advance digital maturity (eg, completion of HITECH program), limited financial and staff resources to complete assessments, competing priorities among leaders such as financial pressures, and workforce shortages, which have escalated since the pandemic. Hospitals that complete the EMRAM assessment and score level 0 may also be less inclined to repeat the assessment until such time as they have achieved significant progress in their digital maturity, whereby assessments of progress serve as motivation to profile their accomplishments when they reach level 6 or 7. The final study sample consisted of 1026 hospitals with complete Leapfrog Group's Hospital Safety Grades (letter grades); complete EMRAM maturity level 0, 6, or 7; and complete CMS data describing hospital characteristics.

XSL•FO
RenderX

Table 2. Description of the Electronic Medical Record Assessment Model (EMRAM) level of digital maturity.

EMRAM level	Requirement to achieve stage
Level 0	The organization has not installed all the key ancillary department systems (eg, laboratory, pharmacy, cardiology, radiology).
Level 1	Laboratory, imaging, pharmacy, and cardiology systems produce patient-centric reports and results. Resilience management plans are in place.
Level 2	A CDR[a] provides access to results and reports, governance and policy control, clinical decision support opportunities, training records, and IT security.
Level 3	Electronic clinical documentation is accessed remotely through the CDR. Role-based access controls are in place.
Level 4	Computerized practitioner order entry and electronic prescribing are available within an electronic medicines administration record. Clinical and information governance is well defined. Monitoring of clinical outcome and patient satisfaction targets occurs.
Level 5	Data are integrated from extremal sources. Changes in clinical parameters are continuously monitored by alerts and warnings. Telehealth and virtual care services are available. Intruder prevention systems manage unauthorized access. Technology supports bedside processes.
Level 6	Medical devices are integrated. Health information exchange supports data sharing. Service users submit self-reported outcomes data. Wearables and implants support remote monitoring and patient management of health and care. Online services improve access and health literacy.
Level 7	Data are integrated from multiple external sources. Service users receive alerts and reminders to support self-managed care and use automated tools to measure patient outcomes. Digital infrastructure tools enable dynamic patient engagement in managing personal health and care.

[a]CDR: clinical data repository.

Statistical Analysis

Descriptive statistics were used to summarize the demographic features of hospitals included in this study. The number of beds was used to categorize hospital size as small (0-149 beds), medium (150-499 beds), and large (≥400 beds) hospitals. The outcome variables in this study differed in their scales of measurement and included binary variables, ordered categories, and continuous variables. The details of the analysis scales are reported in Table 1. Consequently, different types of regression modeling approaches were required to examine the associations between digital maturity levels (underdeveloped versus advanced) and The Leapfrog Group's Hospital Safety Grades, practices, and outcomes. The regression methodologies used in each case are described in the following paragraphs.

The association between EMRAM maturity and Leapfrog Safety (letter) grades, which is an ordinal categorical variable, was analyzed using a cumulative logit regression model, whereby the odds of achieving higher Leapfrog Safety Grades were estimated. EMRAM maturity levels were used as predictors while controlling for hospital characteristics (eg, bed size, ownership, rural or urban location, and teaching status).

The association between EMRAM maturity and the Process and Structural Measures of Leapfrog safety practices (CPOE, ICU, SP1, SP2, SP4, SP9, and SP19, described in Table 1) were analyzed using logistic regression analysis to model the odds of meeting the standard or achieving a full score with respect to EMRAM maturity levels and hospital characteristics.

The hospital-acquired infection (HAI) outcomes described in Table 1 had a substantial number of 0s, which makes ordinary regression models invalid. Therefore, the association between EMRAM maturity levels and HAI measures were analyzed using Tweedie regression, which is designed for outcomes that have nonnegative values with a large number of 0s [45]. The results of this analysis are presented as the multiplicative effects of digital maturity (ie, the predictor) on the average rate of the HAI outcome variables. For example, if the effect of EMRAM maturity on CLABSI is 1.2, it is to be interpreted as hospitals with underdeveloped maturity having 20% higher rates of CLABSI on average than digitally advanced (EMRAM 6, 7) hospitals.

Patient safety indicators (PSI3, 4, 6, 11, 12, 14, 15) and hospital-acquired conditions (“foreign object retention” and “falls and trauma,” described in Table 1), are reported by Leapfrog as the rate per 1000 CMS eligible discharges. These measures vary between 0 and 1000. The associations of these measures with EMRAM maturity and other hospital characteristics (eg, bed size, ownership, rural or urban location, teaching status) were analyzed using ordinary linear regression analysis. The results of this regression analysis are presented as the additive increases or decreases in average rates due to changes in the levels of the predictor variable (advanced versus underdeveloped maturity levels).

Results

There were 1026 hospitals included in this study. Their characteristics are described in Table 3. The majority of hospitals in this study had not advanced their digital maturity, evidenced by EMRAM level 0 digital maturity.

The results revealed that, after controlling for hospital characteristics, the odds of achieving a higher Leapfrog Group's Hospital Safety Grade were statistically significantly higher, by 3.25 times, for hospitals with advanced digital maturity (odds ratio [OR] 3.25, 95% CI 2.33-4.55). When examining the role of hospital characteristics, nonteaching hospitals had better odds of achieving a higher Leapfrog Group’s Hospital Safety Grade, and governmental hospitals had lower odds of achieving a higher

XSL•FO
RenderX

Leapfrog Group's Hospital Safety Grade than both nonprofit and privately owned hospitals (Table 4). There was no statistical evidence that hospital size (eg, bed size) influenced Leapfrog Group's Hospital Safety Grade outcomes.

The odds of achieving a full score on the culture of leadership structures and systems (SP1) was statistically significantly higher for hospitals with advanced digital maturity (OR 3.45, 95% CI 1.64-7.26). Similarly, such odds were 3.54-fold greater for digitally advanced hospitals with respect to nursing workforce (SP9; OR 3.54, 95% CI 1.53-8.18; Table 5). Although not statistically significant, the odds were also higher for hospitals with advanced digital maturity relative to identification and mitigation of risks and hazards (SP4) and hand hygiene practices (SP19; Table 5). Digital maturity was significantly associated with CPOE, intensive care unit physician staffing, and BCMA, whereby digitally mature hospitals had higher odds of meeting the standard for these 3 measures (Table 4). Advanced digital maturity was statistically significantly associated with a number of patient safety indicators, including the incidence of pressure ulcers (PSI3), episodes of collapsed lung (PSI6), and unrecognized abdominopelvic accidental puncture/laceration (PSI15; Tables 6 and 7). The majority of PSI outcomes for digitally mature hospitals indicated reduced rates of these adverse events.

CLABSI rates were an average 21% lower for hospitals with advanced digital maturity (Table 8), which was the only statistically significant association between digital maturity and HAIs. Large hospitals had significantly lower rates of CLABSI when compared with smaller hospitals, and nonprofit hospitals had significantly lower rates of CLABSI compared with either government or proprietary hospitals. Digitally advanced hospitals had, on average, lower rates of methicillin-resistant *staphylococcal aureus*, *Clostridioides difficile* infection, and surgical site infections, but the differences were not statistically significant.

Table 3. Demographic characteristics of hospitals.

Digital maturity level	Results, n (%)
Advanced (EMRAM[a] 6,7)	162 (15.8)
Underdeveloped (EMRAM 0)	864 (84.2)
Teaching status	
Teaching	475 (46.3)
Nonteaching	551 (53.7)
Rural vs urban status	
Rural	433 (42.2)
Urban	593 (57.8)
Ownership type	
Government	129 (12.6)
Nonprofit	666 (64.9)
Proprietary	231 (22.5)
Hospital size (number of beds)	
Small (1-149)	229 (22.3)
Medium (150-399)	180 (17.7)
Large (≥400)	615 (59.9)
Referral status	
Referral center	289 (28.2)
Nonreferral center	737 (71.8)

[a]EMRAM: Electronic Medical Record Assessment Model.

XSL•FO
RenderX

Table 4. Association between digital maturity and Leapfrog Safety Grade, Structural, and Process Measures reported as odds ratios (ORs) and 95% CIs using a separate logistic regression model on each outcome (columns).

Independent variables (predictors)	Leapfrog Safety Grade, OR (95% CI)	CPOE[a], OR (95% CI)	BCMA[b], OR (95% CI)	IPS[c], OR (95% CI)
Digital maturity (advanced vs underdeveloped)	3.25 (2.33-4.55)	2.29 (1.57-3.35)	1.51 (1.00-2.27)	1.75 (1.18-2.60)
Teaching (no vs yes)	1.38 (1.05-1.83)	0.98 (0.70-1.38)	0.96 (0.67-1.39)	0.78 (0.55-1.11)
Rural vs urban	1.2 (0.92-1.57)	0.78 (0.56-1.09)	0.62 (0.43-0.91)	0.77 (0.53-1.12)
Government vs nonprofit	0.67 (0.47-0.95)	0.89 (0.57-1.38)	1.31 (0.82-2.09)	1.05 (0.65-1.71)
Government vs proprietary	0.58 (0.39-0.86)	0.45 (0.27-0.73)	0.88 (0.52-1.47)	0.86 (0.50-1.48)
Nonprofit vs proprietary	0.86 (0.65-1.15)	0.51 (0.36-0.71)	0.67 (0.46-0.97)	0.81 (0.56-1.19)
Hospital size				
0-149 beds vs 150-399 beds	0.98 (0.73-1.31)	0.62 (0.43-0.89)	0.73 (0.50-1.09)	0.22 (0.14-0.34)
0-149 beds vs ≥400 beds	1.05 (0.71-1.56)	0.40 (0.25-0.66)	0.90 (0.53-1.53)	0.17 (0.10-0.29)
150-399 beds vs ≥400 beds	1.08 (0.77-1.49)	0.65 (0.44-0.96)	1.22 (0.80-1.87)	0.78 (0.53-1.14)
Referral (no vs yes)	1.54 (1.14-2.09)	1.41 (0.97-2.05)	0.99 (0.66-1.49)	1.29 (0.86-1.93)

[a]CPOE: computerized physician order entry.

[b]BCMA: bar code medication administration.

[c]IPS: intensive care unit physician staffing.

Table 5. Association between digital maturity and safety practices (SP) reported as odds ratios (ORs) and 95% CIs using a separate logistic regression model on each outcome (columns).

Independent variables (predictors)	Culture of leadership (SP1), OR (95% CI)	Culture of measurement (SP2), OR (95% CI)	Risk mitigation (SP4), OR (95% CI)	Nursing workforce (SP9), OR (95% CI)	Hand hygiene (SP19), OR (95% CI)
Digital maturity (advanced vs underdeveloped)	3.45 (1.64-7.26)	1.46 (0.84-2.52)	0.76 (0.45-1.30)	3.54 (1.53-8.18)	1.34 (0.72-2.50)
Teaching (no vs yes)	0.61 (0.37-0.99)	0.89 (0.55-1.44)	0.77 (0.47-1.26)	0.53 (0.30-0.92)	0.82 (0.46-1.46)
Rural vs urban	0.52 (0.31-0.85)	1.05 (0.64-1.71)	0.96 (0.58-1.60)	0.81 (0.46-1.44)	0.94 (0.53-1.67)
Government vs nonprofit	0.41 (0.23-0.74)	0.59 (0.34-1.03)	0.49 (0.27-0.86)	1.57 (0.72-3.43)	0.52 (0.28-0.96)
Government vs proprietary	0.37 (0.20-0.70)	0.34 (0.18-0.65)	0.38 (0.20-0.74)	0.88 (0.38-2.05)	0.17 (0.07-0.38)
Nonprofit vs proprietary	0.91 (0.57-1.46)	0.58 (0.35-0.95)	0.79 (0.47-1.33)	0.56 (0.32-0.96)	0.32 (0.16-0.65)
Hospital size					
0-149 beds vs 150-399 beds	0.94 (0.57-1.55)	0.65 (0.40-1.07)	0.71 (0.42-1.18)	0.94 (0.53-1.65)	0.64 (0.35-1.15)
0-149 beds vs ≥400 beds	0.60 (0.28-1.31)	0.68 (0.34-1.35)	0.61 (0.29-1.27)	0.65 (0.27-1.57)	0.68 (0.30-1.56)
150-399 beds vs ≥400 beds	0.65 (0.34-1.24)	1.04 (0.59-1.84)	0.86 (0.47-1.58)	0.69 (0.32-1.48)	1.07 (0.54-2.13)
Referral (no vs yes)	0.84 (0.48-1.47)	1.20 (0.71-2.04)	1.33 (0.77-2.29)	1.89 (1.03-3.46)	1.15 (0.62-2.14)

XSL•FO
RenderX

Table 6. Association between digital maturity and patient safety indicators (PSI) reported as mean change and 95% CIs using a separate linear regression model on each outcome (columns).

Independent variables (predictors)	Pressure ulcer rate (PSI3), mean change[a] (95% CI)	Death rate serious conditions (PSI4), mean change[a] (95% CI)	Collapsed lung (PSI6), mean change[a] (95% CI)
Digital maturity (advanced vs underdeveloped)	0.06 (0.00 to 0.12)	1.87 (–1.39 to 5.13)	0.01 (0.00 to 0.02)
Teaching (yes vs no)	0.01 (–0.04 to 0.06)	–0.59 (–3.49 to 2.31)	0.01 (–0.00 to 0.01)
Urban (yes vs no)	0.00 (–0.05 to 0.05)	3.18 (–0.07 to 6.43)	0.00 (–0.00 to 0.01)
Ownership (vs government owned)			
Proprietary ownership	–0.07 (–0.14 to 0.01)	–8.05 (–12.91 to –3.19)	–0.01 (–0.02 to 0.00)
Nonprofit ownership	–0.03 (–0.09 to 0.04)	–4.77 (–9.11 to –0.44)	–0.01 (–0.02 to –0.00)
Hospital size (vs small hospitals [0-149 beds])			
Large hospitals (≥400 beds)	0.09 (0.02 to 0.16)	–0.38 (–5.13 to 4.38)	–0.01 (–0.02 to 0.00)
Medium hospitals (150-399 beds)	0.03 (–0.02 to 0.09)	0.60 (–3.49 to 4.70)	–0.00 (–0.01 to 0.00)
Referral (yes vs no)	0.00 (–0.06 to 0.06)	5.95 (2.68 to 9.23)	0.00 (–0.01 to 0.01)

[a]Average additive change in the mean of the measure with respect to a reference group.

Table 7. Association between digital maturity and patient safety indicators (PSI) reported as mean change and 95% CIs using a separate linear regression model on each outcome (columns).

Independent variables (predictors)	Postoperative respiratory failure (PSI11), mean change[a] (95% CI)	Postoperative DVT/PE[b] rate (PSI12), mean change[a] (95% CI)	Postoperative wound dehiscence (PSI14), mean change[a] (95% CI)	Accidental puncture/laceration (PSI15), mean change[a] (95% CI)
Digital maturity (advanced vs underdeveloped)	0.46 (–0.02 to 0.94)	–0.05 (–0.21 to 0.11)	0.01 (–0.04 to 0.05)	0.07 (0.03 to 0.12)
Teaching (yes vs no)	0.25 (–0.16 to 0.67)	0.16 (0.02 to 0.29)	0.01 (–0.03 to 0.04)	0.01 (–0.03 to 0.05)
Urban (yes vs no)	0.14 (–0.26 to 0.55)	0.03 (–0.10 to 0.16)	–0.01 (–0.05 to 0.02)	–0.01 (–0.04 to 0.03)
Ownership (vs government owned)				
Proprietary ownership	0.40 (–0.21 to 1.02)	–0.22 (–0.42 to –0.03)	–0.04 (–0.09 to 0.01)	–0.03 (–0.08 to 0.02)
Nonprofit ownership vs government	–0.09 (–0.62 to 0.45)	–0.14 (–0.31 to 0.03)	–0.00 (–0.05 to 0.04)	–0.00 (–0.05 to 0.04)
Hospital size (vs small hospitals [0-149 beds])				
Large hospitals (≥400 beds)	0.37 (–0.22 to 0.97)	0.19 (–0.01 to 0.38)	–0.04 (–0.09 to 0.01)	0.01 (–0.03 to 0.07)
Medium-sized hospitals (150-399 beds)	0.38 (–0.05 to 0.81)	–0.11 (–0.25 to 0.03)	–0.00 (–0.04 to 0.04)	0.01 (–0.03 to 0.04)
Referral (yes vs no)	0.41 (–0.04 to 0.87)	–0.01 (–0.15 to 0.14)	–0.04 (–0.08 to –0.00)	–0.00 (–0.04 to 0.04)

[a]Average additive change in the mean of the measure with respect to a reference group.

[b]DVT/PE: deep vein thrombosis/pulmonary embolism.

XSL•FO
RenderX

Table 8. Association between digital maturity and hospital-associated infections and hospital-acquired conditions, reported as multiplicative mean change and 95% CIs using a separate Tweedie regression model on each outcome (columns).

Independent variables (predictors)	CLABSI[a], mean change[b] (95% CI)	MRSA[c], mean change[b] (95% CI)	CAUTI[d], mean change[b] (95% CI)	CDI[e], mean change[b] (95% CI)	SSI[f], mean change[b] (95% CI)	FOR[g], mean change[b] (95% CI)	FAT[h], mean change[b] (95% CI)
Digital maturity (advanced vs underdeveloped)	1.21 (1.02-1.42)	1.08 (0.91-1.29)	0.9 (0.78-1.05)	1.1 (0.99-1.22)	1.12 (0.94-1.33)	0.73 (0.39-1.35)	0.98 (0.83-1.16)
Teaching (yes vs no)	1.04 (0.9-1.19)	0.98 (0.84-1.14)	1.04 (0.91-1.18)	1.1 (1.01-1.2)	1.04 (0.89-1.22)	3.16 (1.51-6.62)	1.07 (0.93-1.24)
Urban vs rural	1.22 (1.04-1.43)	1.18 (0.99-1.4)	1.04 (0.91-1.2)	1.02 (0.94-1.11)	1.2 (1.01-1.43)	1.82 (0.92-3.56)	0.82 (0.71-0.95)
Ownership (vs government owned)							
Proprietary vs government	0.81 (0.66-1.0)	0.83 (0.66-1.05)	0.78 (0.64-0.95)	1.03 (0.91-1.17)	0.65 (0.51-0.84)	0.33 (0.11-0.98)	1.09 (0.88-1.35)
Nonprofit vs government	0.81 (0.67-0.97)	0.78 (0.64-0.96)	0.82 (0.69-0.97)	1.07 (0.96-1.2)	0.9 (0.73-1.1)	1.28 (0.53-3.06)	1.09 (0.91-1.32)
Hospital size (vs small hospitals [0-149 beds])							
Large hospitals (≥400 beds)	1.24 (1.0-1.54)	1.1 (0.83-1.44)	1.06 (0.88-1.27)	0.98 (0.87-1.11)	1.33 (1.05-1.69)	27.11 (9.58-76.71)	1.30 (1.06-1.59)
Medium hospitals (150-399 beds)	1.13 (0.95-1.35)	1.06 (0.83-1.35)	0.99 (0.85-1.14)	1.03 (0.94-1.12)	1.13 (0.92-1.38)	6.23 (2.44-16.12)	1.16 (0.99-1.35)
Referral (yes vs no)	1.12 (0.95-1.31)	1.14 (0.95-1.35)	1.08 (0.94-1.25)	0.99 (0.9-1.09)	1.21 (1.02-1.43)	1.51 (0.75-3.00)	0.89 (0.76-1.04)

[a]CLABSI: central line–associated bloodstream infection.

[b]Average multiplicative change in mean of the measure with respect to a reference group.

[c]MRSA: methicillin-resistant *Staphylococcus aureus*.

[d]CAUTI: catheter-associated urinary tract infection.

[e]CDI: *Clostridioides difficile* infection.

[f]SSI: surgical site infection.

[g]FOR: foreign object retained.

[h]FAT: falls and trauma.

Discussion

This cross-sectional study included a large and diverse group of US hospitals (n=1026), whereby advanced digital maturity was a significant predictor of The Leapfrog Group's Hospital Safety Grades for hospitals of all sizes and types in this study. However, only 15.8% (162/1026) of these hospitals had achieved advanced digital maturity (assessed as EMRAM 6 or 7), with the majority of hospitals having underdeveloped digital maturity (scoring EMRAM 0). This finding is consistent with evidence of the impact of the HITECH policy in 2009, known as the “meaningful use” policy, which accelerated the adoption of EMR technologies in US hospitals but failed to stimulate adoption of the advanced EMR technology that is required to strengthen hospital performance, such as quality and safety [11,12]. The achievements of the 15.8% of hospitals with advanced digital maturity in this study offer insights into the key features of advanced digital maturity that are associated with stronger quality and safety outcomes. Hospitals validated as having advanced digital maturity are able to mobilize data from multiple external sources, have automated tracking of patient progress and health outcomes, and engage patients directly in accessing their health data to manage their health and care [15]. In contrast, the 84.2% of hospitals with underdeveloped digital maturity (EMRAM level 0) had not completed installation of all of the key ancillary information systems such as laboratory, pharmacy, cardiology, and imaging systems, which suggests that leaders and clinicians working in these hospital settings cannot readily track quality and safety outcomes or assess patient health status. When digital maturity is at a very basic level, data from these critical information systems cannot be readily accessed or mobilized, and significant limitations in the flow of quality and safety data result in lack of data-informed strategies and decision-making needed to advance quality, safety, and performance outcomes more generally.

A key question emerging from these findings is why so few hospitals have advanced their adoption of the advanced features of EMR technology. One potential explanation is the limited resources available to invest in advancing maturity without policy incentives such as the HITECH program to stimulate continued progress. Another possible explanation may be related to the profound impact of the pandemic on hospitals that experienced exceptional financial and workforce challenges from 2020 to 2023, whereby progress toward digital maturity likely halted or slowed down given the many competing demands during this challenging period. Finally, a third reason may be that hospital leaders in the United States could be

XSL•FO
RenderX

hesitant to invest in digital maturity assessments as they progress their digital transformation effort until such time as they have made significant progress due to the concern that such results will be viewed as "poor or limited" progress among key stakeholders such as patients and funders. The US system is highly competitive, and leaders may be limited in their tolerance for maturity assessment outcomes of 0 to 4 that may have the potential to impact revenues and market reach.

In addition, hospitals with advanced digital maturity were significantly more likely to have a strong leadership safety culture, more advanced safety structures and systems, and stronger quality and safety outcomes that are sensitive to nursing workforce capacity (eg, risk assessments related to staffing, board accountability for safety-related staffing, budget resources that prioritize optimal staffing, and actionable strategies focused on performance improvement) [46]. Evidence of hospital strategies associated with advanced EMR adoption identify 3 key leadership and systems integration requirements [46]. Specifically, digitally mature hospitals have achieved advanced systems integration that offers the critical infrastructure capacity to support data-driven decision-making to advance quality and safety performance. A leadership culture that prioritizes safety may also be a critical driver of systems integration and advances in digital maturity that strengthen workforce capacity by integrating safe practices into workflows and establishing accountability structures to measure performance outcomes that are supported by a robust digital infrastructure for tracking and reporting safety outcomes.

The lack of integration of ancillary information systems evident in the majority of hospitals in this study may be a contributing factor to the negative impact of digital maturity on clinician workflows and clinician burnout, described in a number of studies [46,47]. Specifically, if lab, pharmacy, cardiology and imaging data do not flow automatically to an individual patient's health record, then clinicians have few options but to search for these critical data points from multiple and disconnected software systems, which requires significant clinician time and attention. With the increased volume and complexity of patient demand, clinician workload is high, and the lack of system integration may further heighten the workload burden that has been reported relative to EMR technologies [47]. In addition, without the integration of data flows across the organization, there is very limited capacity for these hospitals to track organizational performance, track progress of patient outcomes, or proactively identify risks to enable interventions to mitigate risks to patients, in order to strengthen quality and safety outcomes.

The results of this study profile a number of key features of digitally advanced hospitals, including a strong leadership culture of quality and safety that leverages advanced digital maturity and EMR technologies as a strategy to achieve quality and safety performance, effective clinical workflows, workforce strategies that strengthen quality and safety (eg, staffing levels associated with accountability for safety outcomes), and data-rich work environments that automate data-driven decision-making and enable real-time tracking of performance outcomes to inform strategic decisions among hospital leaders.

Efforts to advance digital maturity in hospitals as a strategic priority have heightened since the pandemic, particularly given current workforce shortages and the decline in quality and safety performance [19]. The findings of this study support current evidence that advanced digital maturity strengthens quality and safety hospital performance outcomes [6-8], which previously had not been empirically well documented [10]. These findings may also offer policymakers and hospital leaders a strategic road map for advancing digital maturity as a strategy to improve quality and safety performance well into the future [17,18]. At a time when global workforce shortages are so challenging [20], these findings may also inform organizational strategies that advance quality of digital work environments that enable greater automation in flow of data to advance data-driven decision-making, reduce workload burden among clinicians, and support quality outcomes for patients [21]. Future policy levers may be effective in accelerating progress toward advanced digital maturity, focused on prioritizing quality and safety outcomes and reducing workforce burden, rather than focusing more narrowly on adoption of EMR technology [12].

New digital and sociotechnical trends in health care continue to emerge rapidly, including progress toward learning health systems [41], artificial intelligence technologies such as machine learning, and large language models such as generative artificial intelligence tools. All of these technologies rely heavily on a robust and mature digital infrastructure that is well integrated and able to capture, securely store, and flow data within and across clinical settings to inform health and care decisions. Yet, if the majority of US hospitals have not achieved digital maturity beyond level 0, then a significant number of hospitals in the United States may not have the digital maturity required to consider analytics strategies focused on adoption of these new technologies that hold so much potential for advancing quality care for patients. In addition, hospitals and health systems that have not advanced beyond level 0 maturity are likely to be profoundly limited in their capacity to advance learning health systems in which performance outcomes and quality of patient care are tracked and monitored to inform the design and development of new, digitally enabled models of care that ensure that every person in every health care setting has access to the highest quality of care possible.

In order to achieve the aspirational promises of digital transformation for US health systems, the findings from this study suggest that hospitals must accelerate progress toward advanced digital maturity to not only strengthen quality and safety outcomes for patients and quality work environments that support a robust and sustainable health workforce but also, perhaps most importantly, ensure that every hospital is well positioned to readily adopt and scale the many impressive digital tools and technologies such as artificial intelligence to ensure equitable and highly effective care delivery focused on sustaining and strengthening health and wellness for the US population.

There are a number of limitations in this study. The sampling frame of hospitals for this study was limited to organizations with a completed EMRAM assessment and Leapfrog survey in 2019, which may not represent all US hospitals. The second limitation is the lack of representation of hospitals at EMRAM

XSL•FO
RenderX

maturity levels 1 through 5, which precluded examination of quality and safety advances at each level of EMRAM maturity. Very few hospitals assessed at EMRAM levels 1 through 5 are represented in the HIMSS US data set, which is a consistent pattern in other countries including Canada, Germany, and Australia. Hospitals typically assess their EMRAM maturity, initially at level 0, and then do not reassess again until they reach level 6, at which point they are validated with an onsite assessment by the HIMSS team.

The results of this study provide many opportunities for future research to document the role of digital maturity to better understand how it advances quality and safety. Specifically, future research may examine the role of data-driven decision-making, mobilizing data from multiple and diverse sources, automation of data flow, and digitally mature work environments that support clinicians to deliver quality care and enable active participation of patients in their care decisions. Future research that documents the complexity of care environments and the many features of digitally mature hospital care environments is needed to better understand how digital maturity advances quality and safe care that is equitable and offers every patient the opportunity to achieve their health and wellness goals. Future research may also examine the specific factors that have limited progress in advanced digital maturity for the majority of hospitals in this study. Examination of the key factors that preclude advancement of digital transformation may inform hospital leaders about strategies to overcome these barriers and challenges. The key feature of leadership expertise and strategy associated with learning health systems that create a culture of learning warrants further research with specific focus on how data-rich environments create the capacity for learning in health systems such as advances in quality of work environments, engagement of patients in their health and care, and effective management of the many competing demands to ensure every patient has the opportunity to access a strengthened health system performance that supports the health and wellness of every US citizen.

Acknowledgments

We gratefully acknowledge the contributions of Olubisi Ajetunmobi for technical support of this research.

Conflicts of Interest

None declared.

References

1. Borycki EM, Kushniruk AW. Health technology, quality and safety in a learning health system. Healthc Manage Forum 2023 Mar;36(2):79-85 [FREE Full text] [doi: 10.1177/08404704221139383] [Medline: 36562483]
2. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: A systematic mapping of review studies. Digit Health 2022 Mar 18;8:20552076221085810 [FREE Full text] [doi: 10.1177/20552076221085810] [Medline: 35340904]
3. Alotaibi YK, Federico F. The impact of health information technology on patient safety. Saudi Med J 2017 Dec 03;38(12):1173-1180 [FREE Full text] [doi: 10.15537/smj.2017.12.20631] [Medline: 29209664]
4. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS Med 2011 Jan 18;8(1):e1000387 [FREE Full text] [doi: 10.1371/journal.pmed.1000387] [Medline: 21267058]
5. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-06-16]
6. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12(6):573-576 [FREE Full text] [doi: 10.1370/afm.1713] [Medline: 25384822]
7. Rahimi K. Digital health and the elusive quest for cost savings. The Lancet Digital Health 2019 Jul;1(3):e108-e109. [doi: 10.1016/s2589-7500(19)30056-1]
8. Woods L, Eden R, Canfell OJ, Nguyen K, Comans T, Sullivan C. Show me the money: how do we justify spending health care dollars on digital health? Med J Aust 2023 Feb 06;218(2):53-57 [FREE Full text] [doi: 10.5694/mja2.51799] [Medline: 36502453]
9. Teixeira F, Li E, Laranjo L, Collins C, Irving G, Fernandez MJ, et al. Digital maturity and its determinants in General Practice: A cross-sectional study in 20 countries. Front Public Health 2022 Jan 13;10:962924 [FREE Full text] [doi: 10.3389/fpubh.2022.962924] [Medline: 36711349]
10. Duncan R, Eden R, Woods L, Wong I, Sullivan C. Synthesizing Dimensions of Digital Maturity in Hospitals: Systematic Review. J Med Internet Res 2022 Mar 30;24(3):e32994 [FREE Full text] [doi: 10.2196/32994] [Medline: 35353050]
11. Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. Health Aff (Millwood) 2017 Aug 01;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]
12. Halamka JD, Tripathi M. The HITECH Era in Retrospect. N Engl J Med 2017 Sep 07;377(10):907-909. [doi: 10.1056/nejmp1709851]

XSL•FO
RenderX

13. Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a "basic" EHR. J Am Med Inform Assoc 2020 Aug 01;27(8):1198-1205 [FREE Full text] [doi: 10.1093/jamia/ocaa090] [Medline: 32585689]
14. Walker DM, Tarver WL, Jonnalagadda P, Ranbom L, Ford EW, Rahurkar S. Perspectives on Challenges and Opportunities for Interoperability: Findings From Key Informant Interviews With Stakeholders in Ohio. JMIR Med Inform 2023 Feb 24;11:e43848 [FREE Full text] [doi: 10.2196/43848] [Medline: 36826979]
15. Electronic Medical Record Adoption Model (EMRAM). HIMSS. URL: https://www.himss.org/what-we-do-solutions/digital-health-transformation/maturity-models/electronic-medical-record-adoption-model-emram [accessed 2023-11-16]
16. Lin Y, Lin M, Chen H. Do Electronic Health Records Affect Quality of Care? Evidence from the HITECH Act. Information Systems Research 2019 Mar;30(1):306-318 [FREE Full text] [doi: 10.1287/isre.2018.0813]
17. Brborović O, Brborović H, Hrain L. The COVID-19 Pandemic Crisis and Patient Safety Culture: A Mixed-Method Study. Int J Environ Res Public Health 2022 Feb 16;19(4):2237 [FREE Full text] [doi: 10.3390/ijerph19042237] [Medline: 35206429]
18. Rangachari P, L Woods J. Preserving Organizational Resilience, Patient Safety, and Staff Retention during COVID-19 Requires a Holistic Consideration of the Psychological Safety of Healthcare Workers. Int J Environ Res Public Health 2020 Jun 15;17(12):4267 [FREE Full text] [doi: 10.3390/ijerph17124267] [Medline: 32549273]
19. Vincent C, Staines A. Enhancing the Quality and Safety of Swiss Healthcare. Bundesamt für Gesundheit (BAG). 2019 Jun 25. URL: https://tinyurl.com/4usdc2rz [accessed 2024-06-16]
20. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? BMJ Glob Health 2022 Jun 27;7(6):e009316 [FREE Full text] [doi: 10.1136/bmjgh-2022-009316] [Medline: 35760437]
21. Singh N, Lawrence K, Sinsky C, Mann DM. Digital Minimalism — An Rx for Clinician Burnout. N Engl J Med 2023 Mar 30;388(13):1158-1159. [doi: 10.1056/nejmp2215297]
22. Dhalla IA, Tepper J. Improving the quality of health care in Canada. CMAJ 2018 Oct 01;190(39):E1162-E1167 [FREE Full text] [doi: 10.1503/cmaj.171045] [Medline: 30274993]
23. Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. National Academies Press. 2001. URL: https://nap.nationalacademies.org/read/10027/chapter/1 [accessed 2024-06-16]
24. McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003 Jun 26;348(26):2635-2645. [doi: 10.1056/NEJMsa022615] [Medline: 12826639]
25. Delivering quality health services: A global imperative for universal health coverage. World Health Organization. 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/272465/9789241513906-eng.pdf [accessed 2024-06-16]
26. Vincent C, Amalberti R. Safer Healthcare: Strategies for the Real World. Cham, Switzerland: Springer; 2016.
27. Patient safety. World Health Organization. URL: https://www.who.int/teams/integrated-health-services/patient-safety [accessed 2024-06-16]
28. OECD. The economics of patient safety: from analysis to action. OECD. 2020. URL: https://www.oecd.org/health/health-systems/Economics-of-Patient-Safety-October-2020.pdf [accessed 2024-06-16]
29. Hannawa AF, Spitzberg BH, Childress MD, Frankel R, Pham JC, Wu AW. Communication science lessons for patient safety and quality care. Journal of Patient Safety and Risk Management 2020 Oct 27;25(5):197-204. [doi: 10.1177/2516043520926424]
30. Staines A, Baralon C, Farin A. TeamSTEPPS : des outils de travail en équipe pour la sécurité des patients (TeamSTEPPS: teamwork tools for patient safety). Risques & Qualité 2016;XIII(3):159-164 [FREE Full text]
31. Redstone CS, Zadeh M, Wilson M, McLachlan S, Chen D, Sinno M, et al. A Quality Improvement Initiative to Decrease Central Line–Associated Bloodstream Infections During the COVID-19 Pandemic: A “Zero Harm” Approach. J Patient Saf 2023 Feb 24;19(3):173-179. [doi: 10.1097/pts.0000000000001107]
32. Pronovost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, et al. An Intervention to Decrease Catheter-Related Bloodstream Infections in the ICU. N Engl J Med 2006 Dec 28;355(26):2725-2732. [doi: 10.1056/nejmoa061115]
33. Blot K, Bergs J, Vogelaers D, Blot S, Vandijck D. Prevention of central line-associated bloodstream infections through quality improvement interventions: a systematic review and meta-analysis. Clin Infect Dis 2014 Jul 01;59(1):96-105 [FREE Full text] [doi: 10.1093/cid/ciu239] [Medline: 24723276]
34. Vermeil T, Peters A, Kilpatrick C, Pires D, Allegranzi B, Pittet D. Hand hygiene in hospitals: anatomy of a revolution. J Hosp Infect 2019 Apr;101(4):383-392. [doi: 10.1016/j.jhin.2018.09.003] [Medline: 30237118]
35. Song X, Stockwell DC, Floyd T, Short BL, Singh N. Improving hand hygiene compliance in health care workers: Strategies and impact on patient outcomes. Am J Infect Control 2013 Oct;41(10):e101-e105. [doi: 10.1016/j.ajic.2013.01.031] [Medline: 23643451]
36. Buljac-Samardzic M, Doekhie KD, van Wijngaarden JDH. Interventions to improve team effectiveness within health care: a systematic review of the past decade. Hum Resour Health 2020 Jan 08;18(1):2 [FREE Full text] [doi: 10.1186/s12960-019-0411-3] [Medline: 31915007]

XSL•FO
RenderX

37. Rotter T, de Jong RB, Lacko SE, Ronellenfitsch U, Kinsman L. Clinical pathways as a quality strategy. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving healthcare quality in Europe: Characteristics, effectiveness and implementation of different strategies. Copenhagen, Denmark: European Observatory on Health Systems and Policies; 2019:309-330.
38. Maharaj AD, Holland JF, Scarborough RO, Evans SM, Ioannou LJ, Brown W, et al. The Upper Gastrointestinal Cancer Registry (UGICR): a clinical quality registry to monitor and improve care in upper gastrointestinal cancers. BMJ Open 2019 Sep 30;9(9):e031434 [FREE Full text] [doi: 10.1136/bmjopen-2019-031434] [Medline: 31575580]
39. Lee J, Kuo Y, Goodwin J. The effect of electronic medical record adoption on outcomes in US hospitals. BMC Health Serv Res 2013 Feb 01;13:39 [FREE Full text] [doi: 10.1186/1472-6963-13-39] [Medline: 23375071]
40. Nguyen K, Wright C, Simpson D, Woods L, Comans T, Sullivan C. Economic evaluation and analyses of hospital-based electronic medical records (EMRs): a scoping review of international literature. NPJ Digit Med 2022 Mar 08;5(1):29 [FREE Full text] [doi: 10.1038/s41746-022-00565-1] [Medline: 35260765]
41. Government of Canada. TCPS 2 (2022) – Chapter 2: Scope and Approach. Panel on Research Ethics.: Government of Canada URL: https://ethics.gc.ca/eng/tcps2-eptc2_2022_chapter2-chapitre2.html#a [accessed 2024-06-18]
42. About. The Leapfrog Group. URL: https://www.leapfroggroup.org/about [accessed 2024-06-16]
43. Leapfrog hospital safety grade scoring methodology. Leapfrog Hospital Safety Grade. 2022 Apr 18. URL: https://www.hospitalsafetygrade.org/media/file/Safety-Grade-Methodology-Spring-20222.pdf [accessed 2024-06-16]
44. Austin JM, D'Andrea G, Birkmeyer JD, Leape LL, Milstein A, Pronovost PJ, et al. Safety in numbers: the development of Leapfrog's composite patient safety score for U.S. hospitals. J Patient Saf 2014 Mar;10(1):64-71. [doi: 10.1097/PTS.0b013e3182952644] [Medline: 24080719]
45. Wolf H, Snowdon AW. People, process and technology: A model for digital transformation of healthcare. Management in Healthcare 2022;6(4):316-334 [FREE Full text]
46. Holmgren AJ, Phelan J, Jha AK, Adler-Milstein J. Hospital organizational strategies associated with advanced EHR adoption. Health Serv Res 2022 Apr;57(2):259-269 [FREE Full text] [doi: 10.1111/1475-6773.13655] [Medline: 33779993]
47. Wu DTY, Xu C, Kim A, Bindhu S, Mah KE, Eckman MH. A Scoping Review of Health Information Technology in Clinician Burnout. Appl Clin Inform 2021 May;12(3):597-620 [FREE Full text] [doi: 10.1055/s-0041-1731399] [Medline: 34233369]

Abbreviations

BCMA: bar code medication administration
CLABSI: central line–associated blood infection
CMS: Centers for Medicare and Medicaid Services
CPOE: computerized physician order entry
EMR: electronic medical record
EMRAM: Electronic Medical Record Assessment Model
HAI: hospital-acquired infection
HIMSS: Health Information Management Systems Society
HITECH: Health Information Technology for Economic and Clinical Health
PSI: patient safety indicator

Edited by Y Zhuang; submitted 12.01.24; peer-reviewed by GA Silvera, J Rubin, W LaMendola; comments to author 02.04.24; revised version received 16.04.24; accepted 15.05.24; published 06.08.24.

Please cite as:
Snowdon A, Hussein A, Danforth M, Wright A, Oakes R
Digital Maturity as a Predictor of Quality and Safety Outcomes in US Hospitals: Cross-Sectional Observational Study
J Med Internet Res 2024;26:e56316
URL: https://www.jmir.org/2024/1/e56316
doi:10.2196/56316
PMID:39106100

©Anne Snowdon, Abdulkadir Hussein, Melissa Danforth, Alexandra Wright, Reid Oakes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing Opportunities and Barriers to Improving the Secondary Use of Health Care Data at the National Level: Multicase Study in the Kingdom of Saudi Arabia and Estonia

Janek Metsallik[1], MSc; Dirk Draheim[2], PhD; Zlatan Sabic[3], PhD; Thomas Novak[4], BA; Peeter Ross[1,5], MD, PhD

[1]E-Medicine Centre, Department of Health Technologies, School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia
[2]Information Systems Group, School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia
[3]Health, Nutrition and Population Global, The World Bank Group, Washington, DC, United States
[4]Office of the National Coordinator for Health Information Technology, Department of Health and Human Services, Washington, DC, United States
[5]Research Department, East Tallinn Central Hospital, Tallinn, Estonia

Corresponding Author:
Janek Metsallik, MSc
E-Medicine Centre, Department of Health Technologies
School of Information Technologies
Tallinn University of Technology
Akadeemia 15a
Tallinn, 12616
Estonia
Phone: 372 56485978
Email: janek.metsallik@taltech.ee

Abstract

Background: Digitization shall improve the secondary use of health care data. The Government of the Kingdom of Saudi Arabia ordered a project to compile the National Master Plan for Health Data Analytics, while the Government of Estonia ordered a project to compile the Person-Centered Integrated Hospital Master Plan.

Objective: This study aims to map these 2 distinct projects' problems, approaches, and outcomes to find the matching elements for reuse in similar cases.

Methods: We assessed both health care systems' abilities for secondary use of health data by exploratory case studies with purposive sampling and data collection via semistructured interviews and documentation review. The collected content was analyzed qualitatively and coded according to a predefined framework. The analytical framework consisted of data purpose, flow, and sharing. The Estonian project used the Health Information Sharing Maturity Model from the Mitre Corporation as an additional analytical framework. The data collection and analysis in the Kingdom of Saudi Arabia took place in 2019 and covered health care facilities, public health institutions, and health care policy. The project in Estonia collected its inputs in 2020 and covered health care facilities, patient engagement, public health institutions, health care financing, health care policy, and health technology innovations.

Results: In both cases, the assessments resulted in a set of recommendations focusing on the governance of health care data. In the Kingdom of Saudi Arabia, the health care system consists of multiple isolated sectors, and there is a need for an overarching body coordinating data sets, indicators, and reports at the national level. The National Master Plan of Health Data Analytics proposed a set of organizational agreements for proper stewardship. Despite Estonia's national Digital Health Platform, the requirements remain uncoordinated between various data consumers. We recommended reconfiguring the stewardship of the national health data to include multipurpose data use into the scope of interoperability standardization.

Conclusions: Proper data governance is the key to improving the secondary use of health data at the national level. The data flows from data providers to data consumers shall be coordinated by overarching stewardship structures and supported by interoperable data custodians.

(J Med Internet Res 2024;26:e53369) doi:10.2196/53369

XSL•FO
RenderX

KEYWORDS

health data governance; secondary use; health information sharing maturity; large-scale interoperability; health data stewardship; health data custodianship; health information purpose; health data policy

Introduction

Background

Governments seek guidance and strategic directions for deploying effective, efficient, and reliable mechanisms for the secondary use of data collected in health care provision. While the primary use of digital data in health care institutions has developed well during the last decades, health care systems look to improve their practice for secondary use. The secondary use of data controls the burden of data capture by enabling the reuse of already collected data for alternative purposes. Among others, the categories of secondary data use include improving the patient experience, health care facility management, service planning and benchmarking, policy development, public health, health care financing, research, and business support [1]. The categories above exploit data traditionally collected in separate data streams and silos. For instance, public health registries or health insurance claims are managed in most countries by dedicated organizations within their databases using specific data collection processes. The siloed approach has led to the duplication of data collection and the waste of health care resources. A report by the Open Data Institute from 2021 concludes that initiatives for health data ecosystems for data reuse are still fragmented in Europe [2].

Digital data and digitalized processes allow for a change in these practices, making data capture universal and allowing digital health care data sharing for different purposes.

From 2019 to 2021, we conducted projects in the Kingdom of Saudi Arabia and Estonia, assessing health and health care data analytics and developing context-specific recommendations. The governments of both countries were looking to advance their decision-making capabilities due to the digitalization of the flow of health data.

The Project in the Kingdom of Saudi Arabia

The Saudi Health Council (SHC), in cooperation with the World Bank, developed the National Master Plan for Health Data Analytics to guide and provide strategic direction for the deployment of effective, efficient, and reliable mechanisms to share data from the health sector for policy and decision-making [3].

The government sought to boost the regulatory, institutional, and technical infrastructure, allowing for efficient data collection from health care systems to process and provide appropriate data analytics and business intelligence for policy and decision makers. The project assessed the existing situation and conceptualized the harmonized national health data analytics operational model and the logical architecture, including core elements such as the Health Data Analytics Framework, actors and their roles, and critical processes.

The initial driver for the development was perceived inefficiency and observable delays in producing analytical data products about the country's health care system. Indirectly, the existing data flow was limiting the ability to produce accurate and timely information for decision-making on many levels of the health care system. The project focused on the requirements of the significant national-level decision makers, including the SHC and management of the health care sectors, namely, Ministry of Health (MoH) Medical Services, National Guard Medical Services, Ministry of Interior Medical Services, and King Faisal Specialist Hospital & Research Center.

The Project in Estonia

The analysis of health and health care data management was part of the Structural Reform Support Service mission of the European Commission, to provide support for the preparation and implementation of growth-enhancing administrative and structural reforms by mobilizing European Union funds and technical expertise. Estonia requested support from the European Commission under Regulation (EU) 2017/825 on the establishment of the Structural Reform Support Program ("SRSP Regulation") to prepare the Person-Centered Integrated Hospital Master Plan [4].

The master plan targeted to (1) provide a map of the current hospital system, its ability to supply health care in different specialties, distribution of its physical and human resources, its financial flows, and its mechanisms of governance and information sharing; (2) provide evidence-based estimates of the population needs and the supply of health workforce and health care services and infrastructures; and (3) propose a hospital master plan of sound reforms in the hospital sector in the midterm.

The planning included an assessment of the data sharing mechanisms and governance. Current and future organizations delivering the data for decision-making throughout the hospital network were analyzed. The scope of the analysis included data for the national-level health care management and policy (Ministry of Social Affairs [MoSA]), public health (National Institute for Health Development [NIHD]), health care financing (Estonian Health Insurance Fund [EHIF]), and health care service management and clinical decision-making (hospitals and family health centers).

The Tension Between Demand and Supply of Information

In both cases, the digitalization of the health and medical data flows should improve the quality of the decisions. Data in health care are often produced and consumed by different stakeholders. They need to cross the boundaries of specialties, institutions, regions, and sectors to deliver informational value to data consumers so that they can make decisions. From the point of view of decision makers, the place of capture has a surplus, and the place of decision-making has a shortage of information. The tension of disbalance generates the need for data flow: data consumers need data from data producers to extract information for decision-making.

XSL•FO
RenderX

The stakeholders in health care that need data for decision-making, such as governmental organizations, payers, policy makers, and others, feel the tension and try to resolve it. They request data providers to collect and deliver data for each type of consumption. As the providers cannot always align the requirements, they often capture the same information multiple times. The uncoordinated design of data flows has manifested in duplication, gaps, and delays.

The projects analyzed health care data supply and demand for data for different health care decisions. The analysis aimed to provide a better basis for planning data management organization and infrastructure.

In both cases, the client saw issues producing proper analytical data products. In the Kingdom of Saudi Arabia case, the focus was on the reports on public health and health care system indicators. The Estonian terms underlined person centricity and efficacy, which introduced the requirement to study data sharing for clinical decision-making, patient engagement, hospital management, health care system planning, health care policy, and health care funding. To assess the situation and plan for better data sharing, we found it essential to map the providers and consumers and evaluate the usability and use of data for decision-making. The assessment of health care data systems focused on the data purpose, flow, and sharing (Figure 1).

Figure 1. Health care decisions are the purpose of data consumption. Health care decisions also provide new data. Often, the source location of the provided data does not match the desired location of the consumed data. This disbalance between supply and demand creates tension that forces data to flow. Data management resolves the tension by sharing data to required locations.

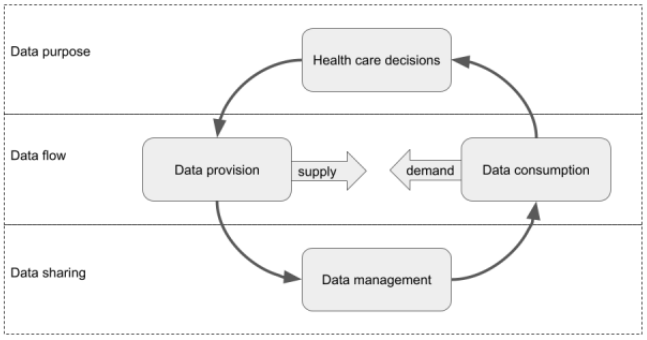

The Aim of the Study

This study aimed to report problems and outcomes from the 2 distinct projects that assess the potential of secondary use of health care data and support of governmental decisions and to map the common thread of thought to apply in similar circumstances. We looked for the matching elements of the problems and the results of the 2 assessments.

Methods

Overview

The respective terms of the projects regulated the work conducted in the Kingdom of Saudi Arabia and Estonia. In general, both projects had to deliver an initial assessment and recommendations for improvement. In the Kingdom of Saudi Arabia, the project concluded with the National Master Plan for Health Data Analytics [3]. In Estonia, the results were integrated into the Person-Centered Integrated Hospital Master Plan [4].

This section describes the framework of definitions and research methods that we developed for the projects.

Mapping Data Sharing Purposes

Digitalization shall enable gains in the effectiveness of decision-making (better inputs, better decisions) and increase the efficiency of data processing (timely and cost-effective delivery of data), resiliency against missing or erratic data, and sustainability of the data management (agility of the data models and infrastructure). One can assess digital health effectiveness by its ability to generate data (inputs) for decision-making. The World Health Organization (WHO) lists various health care performance indicators for management and policy decisions. The WHO has divided the core health indicators into 4 domains: health status, risk factors, service coverage, and health systems [5]. The system of indicators supports rationalized alignment of priorities and harmonization of investments for various levels of health care systems. We used the system of indicators to model health data completeness. The indicators allowed us to cross-check if the needs of the decision makers were fully met. We analyzed the ability of a health care system to coordinate the data required for the indicators.

The purpose of the collected data is to support health care activities (Figure 1). The activities depend on input data and generate new data, including clinical, managerial, financial, and others. The organizational or human actors of studied ecosystems perform these activities. For example, a hospital manager preparing the financial plans consumes data about the average cost per patient case. On the basis of the WHO indicator domains, the studies searched for evidence of data consumption in public health status and risks, health care activity, resources, funding, and clinical decision-making and research. In the Kingdom of Saudi Arabia case, we paid less attention to the clinical side, mainly focusing on the national public health indicators and the health care system. In the Estonian case, in addition to clinical decision-making, we investigated patient-side decision-making, and patient engagement was considered a separate health care activity.

Mapping Data Flow

Digital health ecosystems facilitate data flows to resolve data provision and consumption tension. The projects in Kingdom

XSL•FO
RenderX

of Saudi Arabia and Estonia mapped the roles, procedures, structures, terminologies, and master data involved in coordinating the flow. We built catalogs of organizations that capture health care data (data providers) and organizations that receive health care data (data consumers). The interoperability between the sources and targets may be organized in many ways, either via bilateral point-to-point agreements or multilateral standards-based agreements. The parties share registries of identifiable objects, such as persons, legal entities, locations, services, and others. We checked the availability of data standards and master data registries. For greater secondary use, multilateral data-sharing agreements shall be in place. If we identified standards-based data sharing, we also examined the governing organization around the standards. Depending on the data governance setup, 1 or more entities would coordinate the data requirements between data consumers and providers, govern information assets, design and enforce data standards, and monitor the continuity of the data flow. The responsibility for the coordination is called data stewardship. The data policy's task is to regulate the distribution of data governance responsibilities and enable control over them. For example, a data policy may state that licensed health care institutions shall follow the data collection standards set by a single stewardship organization; this may enforce a data flow that makes a log of activities and resources spent on those activities available to health care management and funding stakeholders.

Mapping Data Sharing

Data sharing is based on the organizational and technical capability that transports data between decision-making locations. We gathered information about the data management platforms, the organizations running the platforms, and the standards supporting data exchange. A data manager or custodian is responsible for maintaining a technical environment and a database structure for data sharing. Regardless of the topology of a data sharing system, centralized or distributed, the data shall be delivered to the correct location at the right time for decision-making. For example, a public registry of laboratory results may transport data between the laboratories, health care providers, and researchers.

The discipline of enterprise information management defines the elements of data flow and sharing. Data governance roles, namely, data policy, data stewardship, and data custodianship, have been used for structuring enterprise information management [6,7]. In the case of the studied projects, we looked for the data governance structures in the national-level health care data organizations.

Assessing the Maturity of Capability

Various capability maturity models support assessing health care information systems [8]. Many maturity models focus on a specialty; a type of organization; or an area of function (hospital management, diagnostic images, and more). Some models focus on the digital health system's ability to connect sources and targets of data sharing. The Health Information Sharing Maturity Model (HISMM) from the Mitre Corporation suggests assessing a digital health system from the perspective of 11 capability areas, which cover technology, use, and governance perspectives [9].

The HISMM model has 2 dimensions: 11 characteristics and 5 maturity levels. The maturity level reflects the level of development or goal achievement regarding a capability. If each launch of a data flow requires the creation of a new organization, the flow has a project-based (1) capability level. At the expert-based (2) capability level, existing organizations (experts) can process a data flow. At the standard-based (3) capability level, data flow can be initiated by involving several organizations providing the same level of service. The data flow at the performance-based (4) level constantly produces indicators of the success of its activities. The data flow is at the learning-based or optimizing (5) level if the performance indicators trigger continuous improvement [9]. For the assessment, we enhanced the HISMM levels with the maturity criteria from ISO 33020 and Capability Maturity Model Integration (CMMI), which guide the assessment of capability maturity of processes related to information systems [10,11]. For example, when level 3 of ISO 33020 requires that "a standard process, including appropriate tailoring guidelines, is established and maintained," we looked for such evidence in our desktop research and in the interview notes.

The characteristics dimension of the HISMM model contains 11 characteristics. The 11 characteristics form a checklist for developing data flows. According to these 11 characteristics, analyzing which additions must be added to the use, technology, or governance organization to increase capacity is possible.

On the basis of our experience with the Kingdom of Saudi Arabia, we introduced the HISMM as an additional tool for assessment in Estonia. Despite our interest, it was economically unreachable for us to redo the assessment in the Kingdom of Saudi Arabia only to compare the HISMM assessment results.

In Estonia, we structured maturity evidence based on the stakeholders' purpose. The structuring allowed the study to analyze the variation in the inputs collected from different decision makers. For example, we were looking for the differences in the maturity of health care management, clinical decision-making, and patient engagement, where all stakeholders may need data about health care resources.

The Framework of the Assessments

The complete framework provides categories for mapping the capabilities and assessing the maturity of those capabilities. Textbox 1 below provides a summary of the categories.

The framework drove the capture and analysis of the findings in both projects. We asked the interviewees about the purposes of using health data and the means they used to manage the data. Together with the interview participants, we investigated the stakeholders, information systems, standards, technologies, and platforms on which their data flows were based. For example, we asked the hospitals' management about the indicators they used in decision-making. Then, we asked the statisticians and IT specialists to describe the sources of the data and the data processing activities.

XSL•FO
RenderX

Textbox 1. The assessment framework includes categories for mapping capabilities from the data and stakeholder purpose perspectives. The purpose is satisfied via data flow and sharing capabilities, which indicate details of the implementation's maturity level.

Data purpose

- Public health status
- Public health risks
- Health care activity
- Health care resources
- Health care funding
- Clinical decision-making
- Clinical research

Data flow

- Data providers
- Data consumers
- Data stewardship
- Data standards
- Data policy
- Master data

Data sharing

- Data platforms
- Data custodianship
- Data exchange standards

Stakeholder purpose

- Patient engagement
- Clinical decisions
- Health care management
- Research and monitoring
- Health care policy
- Health care funding

Capability maturity

- Level 5. Optimizing
- Level 4. Performance
- Level 3. Standards
- Level 2. Experts
- Level 1. Projects

Capability characteristic

- Technology

1. Data quality
2. Data transport
3. Data security
4. Interoperability

- Use

XSL•FO
RenderX

5. Usability

6. Alignment

7. Participation

8. Consent

- Governance

9. Data governance

10. Stakeholder governance

11. Sustainability

Project Activity in the Kingdom of Saudi Arabia

For the situational analysis in the Kingdom of Saudi Arabia, we completed a comprehensive institutional review of the current data systems in the health sector, with an emphasis on how these data were collected and could be routinely made available to the responsible authorities.

We assessed the processes at the institutional, operational, and technical levels. To understand the architectural options for integrated data management, we analyzed the current development plans and statuses, including a rapid assessment of existing computerized information systems, services, and tools. Specifically, we assessed the critical processes for data management and use, system architectures, database architectures, key relevant data sets, data exchange capabilities, geospatial tools, and system platforms available in the health system. The assessment identified the health system's critical information systems, data sets, and exchange capabilities.

The assessment methodology included primary and secondary sources, including interviews with stakeholders. The research team interviewed policy makers and stakeholders during a sequence of missions in 2019. The stakeholders included the national-level health care coordination (SHC) and management of the sectors, namely, MoH Medical Services, National Guard Medical Services, Ministry of Interior Medical Services, and King Faisal Specialist Hospital & Research Center. The project included an in-depth web search of written information and web-based resources on digital health tools and systems for data exchange and analytics. To improve the primary stakeholders' capacity and achieve a common understanding of master plan goals, a seminar about global experiences and examples of technical solutions took place during the first technical mission.

Project Activity in Estonia

The team evaluated Estonian hospitals' health data and information exchange levels. The method combined interviews and desk research. The aim was to understand the value of health and medical information sharing capabilities to stakeholders, identify gaps in funding, and relate governmental activities to strategies and frameworks.

This rapid assessment used semistructured interviews. Institutional, operational, and technical experts described their view on Estonian digital health care data governance; health data flows; information security; and existing computerized information systems, services, and tools. The analysis considered the hospitals part of a more comprehensive data sharing network. Hence, participants provided inputs about both internal and external data sharing. Specifically, interviews with the stakeholders touched on the critical processes for data management and use, system architectures, database architectures, key relevant data sets, data exchange capabilities, and system platforms available in the health system, considering the current use of the EHIF, the Estonian Health Information System (EHIS), and the NIHD databases.

The data sharing network under discussion included health care institutions, public authorities, and other data users, for example, researchers and patients (from the point of view of hospitals). As the interviews covered the involved participants in both data provider and data consumer roles, the captured evidence also touches on the existing and potential use of data for primary and secondary purposes. On multiple occasions, the interviewed stakeholders were able to share insights into the integration of health data exchange and services with social and labor market services. The assessment covered vital information systems, data sets, and data exchange capabilities of the Estonian health care system.

Altogether 9 stakeholders were interviewed, including hospitals (North Estonia Medical Centre, Tartu University Clinic, Pärnu Hospital, Viljandi Hospital, Põlva Hospital, and East Viru Central Hospital) and specialists from the NIHD, the EHIF, and the Estonian Society of Cardiology. Before the interview, we provided the interviewees with a comprehensive set of questions divided into 11 categories according to the capability attributes of the HISMM. The length of the interviews ranged from 1.5 to 2 hours. Usually, the group consisted of 4 to 5 persons, including the head or vice-head of the institution; chief specialists of clinical, IT, service development departments; health accounting; and statistics.

Ethical Considerations

This study compiled the framework, methods, and findings from the past project deliverables, which were available publicly or per request from the respective owners. The projects in the Kingdom of Saudi Arabia and Estonia assessed material available from public sources and interviews. The included organizations-appointed interview participants. The projects did not compensate for the participation. We informed the participants about the purpose of the assessment and used the interview results anonymously. The study team never recorded any health data during the interviews or site visits. This study

XSL•FO
RenderX

includes statements on the possible limitations of the conclusions.

Summary of the Methodology

Table 1 summarizes the methods used by the projects in the Kingdom of Saudi Arabia and Estonia.

Table 1. Methodology of the case projects.

Methodology element	The Kingdom of Saudi Arabia	Estonia
Research type	Evaluation research	Evaluation research
Research design	Exploratory case study	Exploratory case study
Sampling method	Purposive sampling	Purposive sampling
Data collection method 1	Personal semistructured interviews	Personal semistructured interviews
Data collection method 2	Documentation review	Documentation review
Data analysis method	Qualitative content analysis	Qualitative content analysis
Data coding 1	Data purpose, flow, and sharing	Data purpose, flow, and sharing
Data coding 2	__[a]	Health Information Sharing Maturity Model
Target application	National Master Plan for Health Data Analytics	Person-Centered Integrated Hospital Master Plan, and Information-Sharing Capability Maturity Assessment
Research question	What are the gaps and critical elements for the national-level improvement of the secondary use of health care data?	What are the gaps and critical elements for the national-level improvement of the secondary use of health care data?

[a]Not applicable.

Results

Digital Health Landscape in the Kingdom of Saudi Arabia

The Kingdom of Saudi Arabia health care information system encompasses several stakeholders, including primary health care (PHC), hospitals under different jurisdictions, the SHC, the MoH, public health research, quality management, and others. The project looked at the health care system as a whole.

The Kingdom of Saudi Arabia has a population of 35 million, divided between 21.4 million Saudis and 13.6 million non-Saudis. The annual population growth rate was 2.38 in 2020, which dropped slightly from 3.19 in 2010. Part of it can be accounted for by the lowered fertility rate of 1.9 in 2018 and 2.98 in 2010 [12]. The population aged >65 years was 3.4% in 2019, and it is expected to grow to 6% by 2030, which makes it a country with a relatively young population compared with its neighbors in West Asia [13].

The health care system in the Kingdom of Saudi Arabia is mainly funded via the MoH of the Kingdom of Saudi Arabia, which covers 287 hospitals with 45,180 beds, 2257 PHC centers, and 973 specialized medical facilities. In addition to MoH, the governmental health care sector includes providers under the Armed Forces Medical Services, National Guard Medical Services, Ministry of Interior Medical Services, King Faisal Specialist Hospital & Research Center, Royal Commission Hospitals, ARAMCO Hospitals, and Ministry of Education. The total number of hospital beds in other government sectors is 13,989, divided among 50 hospitals. The private sector in the Kingdom of Saudi Arabia runs 167 hospitals with 19,427 beds [14].

In 2000, the Kingdom of Saudi Arabia institutionalized the development of electronic health care as a governmental committee. In 2005, the government established the Saudi Association for Health Informatics, which focused on the growing awareness of electronic health among health care professionals [15]. The effort put into awareness and education has supported the adoption of health IT. Still, the adoption could have been more cohesive, and the use of electronic health systems has been limited [16]. A multiple-case study based on a survey (conducted in 2010) of 6 of the seemingly most advanced medical cities of the Kingdom of Saudi Arabia concluded that inadequate data management policies and procedures, resistance to change, the low analysis of data, and lack of accreditation impact the health IT adoption. The study revealed a need to introduce a national regulator and establish a data exchange plan through a national health information network [17]. The MoH of the Kingdom of Saudi Arabia has invested in the growth of health information exchange on all health care levels. Some researchers have found that the MoH's focus on information exchange between the health care system participants supports the greater adoption of electronic health records. The sharing adds more value to the data and increases the motivation for quality data capture and improvement of health IT tooling [18].

The SHC, established in 2014, is a successor to the Health Services Council, established in 2002. The role of the SHC is to coordinate and integrate health care stakeholders regardless of the type of ownership or the sector of governance. At the time of the project, the SHC included 16 representatives from several ministries, national health care agencies, education institutions, and health care institutions. The SHC governs some national health centers, including the National Center for Health Information [19].

The National Health Information Center (NHIC) was established in 2013, with the mission to organize health information exchange among all health sectors and related parties, to develop

XSL•FO
RenderX

and customize terminology and data exchange standards, to create and supervise telehealth networks, to create national disease registries, and to provide health information to the beneficiaries [20].

Digital Health Landscape in Estonia

Estonia has a population of 1.3 million, which has declined since 1990. The annual population growth rate has been approaching 0 (from the negative side) in the past years. However, the growth rate has been lifted by migration as the fertility rate of 1.6 per woman is less than that required for reproduction [21]. The population aged >65 years was 20% in 2019 and is expected to grow to 23.5% by 2030, slightly above the average of 18.8% and 21.8% in the region of Northern Europe [22].

All health care institutions operate under private law in Estonia. General practitioners are private entrepreneurs or limited companies. At the same time, hospitals are joint-stock companies or not-for-profit foundations licensed by the Health Board and provide various inpatient or outpatient medical or nursing care. In total, 1428 health care institutions were covered by the National Institute of Health Development statistics in 2019. There are 52 hospitals with 6788 beds, 436 family health centers, 490 dental care providers, 317 specialized outpatient medical care, and others [23]. The health care system in Estonia is governed by the MoSA. The system's structure includes agencies of the MoSA (eg, State Agency of Medicines, Health Board, NIHD, and Center of Health and Welfare Information Systems); independent public bodies (EHIF); (mainly publicly owned) hospitals under private regulation; private PHC units; and various nongovernmental organizations and professional associations. The financing is organized chiefly through an independent single public-payer EHIF, including ambulance services [24]. The government regulation establishes a list of regional, central, general, local, and rehabilitation hospitals, a total of 19 hospitals, to ensure uniform access to health care services. These hospitals are entitled to receive the necessary construction, renovation, and reprofiling investments from the government budget. With the hospitals mentioned in the list, EHIF concludes treatment financing contracts for at least 5 years based on the type of hospital listed and the corresponding operating license.

The MoSA of Estonia covers public health from the state budget. Private, primarily out-of-the-pocket spending was 22.7% in 2016 [21].

Health care data are divided between 14 primary national-level sources, in-house sources of health care service providers, and databases of research institutions. In addition to inherent health care data sources, the e-government platform enables the secondary use of public registers for health care needs. For example, the Population Register, managed by the Estonian Police and Border Guard Board, is the source of personal data for patient management. The Health Board manages public registers of health care professionals and health care institutions. The State Agency of Medicines manages registers of drugs, medicinal products, and pharmacies. The EHIF collects reimbursement-related health data, registers the status of insured persons, and manages digital prescriptions. The Center of Health and Welfare Information Systems maintains a significant platform for health data sharing, the EHIS [25] that encompasses the whole country, registers all residents' health history from birth to death, and is based on the e-government infrastructure.

Since 2008, the Digital Health Platform (DHP) has been operating in Estonia, which shares the health care data of the entire country's residents in a secure e-government environment, both between authorized health care workers and between a health care worker and a natural person. The DHP, whose official name is the EHIS, aims, among other things, to process the data related to the area of health care for entry into and performance of contracts for the provision of health services; for ensuring the quality of health services and the rights of patients; and for the protection of public health, including for the upkeep of registers and the organization of health statistics and the management of health care [24].

Assessment Results in the Kingdom of Saudi Arabia

In the case of the Kingdom of Saudi Arabia, the project delivered 2 consecutive components that led to the recommendations for decision makers to boost the regulatory, institutional, and technical infrastructure within the country's health care system, thereby allowing for the efficient collection of health care data.

First, we performed the situational analysis of health and health care data management. The delivered assessment report provided a brief institutional review of the Kingdom of Saudi Arabia health care system's data management and identified vital information systems, analytical data sets, and data exchange capabilities.

The assessment report revealed that data reporting and analytics procedures, standards, and forms in health care should be coordinated and coherent across the ministries with health care institutions under their jurisdiction and with the MoH and SHC.

On the basis of the review, the report gave recommendations for the long-term institutional, organizational, architectural, and technical redesign of health care data analytics to move from static, fragmented, and incomplete data sets to rapid, reliable, and dynamic data processing, exchange, extraction, and consolidation. We used these recommendations as the basis for the development of the second deliverable, the Master Plan for National Health Data Analytics; it is a policy document that describes the regulatory, institutional, and technical infrastructure that allows the efficient collection of digital health and health care data from health care systems to process and provide appropriate data analytics and business intelligence for policy and decision makers. The SHC of the Kingdom of Saudi Arabia approved the Master Plan in 2020.

The Master Plan outlines a framework for data, roles, and processes. The framework considers analytical data sets, health care indicators, reports, metadata, and catalogs as parts of the data dimension. The framework of roles supports the governance of the data and information flows and endorses the strategic value of the analytical data. The Master Plan defines the specific organizations that shall fulfill the defined roles. The process dimension outlined a set of workstreams for analytical data. It

XSL•FO
RenderX

described a path to produce needed policies, objectives, data definitions, analytical product definitions, and standards.

The Master Plan evaluated multiple options for assigning the data governance roles. The final recommendation was to share the responsibilities between the units of the SHC: the policy and strategy management to the SHC Board, data stewardship and analytics to the National Health Analytics Department, data custodianship to the National Health Observatory at the NHIC, and standardization to the Data Standardization Unit at the NHIC.

To operationalize the framework, the Master Plan proposes 3 years to transition responsibilities and develop institutional capacity on all levels. After that, all actors at the national and subnational levels shall adapt to the general and national data analytics frameworks.

Assessment Results in Estonia

In Estonia, the study used interviews to collect inputs for the assessment. The researchers adopted the HISMM and reorganized the notes using capability attributes and stakeholder purpose. The purpose dimension aimed to summarize the capabilities or flows that the interviews covered. The interview content covered the purpose, flow, and sharing. We asked the interviewees to cover the topics for all the data governance roles of their institutions. For example, depending on a specific capability discussed, a hospital can be a data provider, consumer, custodian, and steward. The notes were analyzed for evidence of maturity and labeled accordingly. The method resulted in a 3D matrix with dimensions for stakeholder purpose, capability attributes, and maturity level (Figure 2).

The researchers estimated the maturity of the information sharing to be on levels 2 to 4. Level 2 represents a situation where the information flow stands on the existing expertise, and the flow outcomes are repeatable. Level 3 indicates the existence of standards and procedures that allow new providers to enter the market. On level 4, the assessed ecosystem shall demonstrate an ability to measure the achievement of the information sharing goals.

The assessment suggests, also visualized in Figure 3, that the improvement focus should be on data quality, transport, and stakeholder governance. From the perspective of stakeholder purpose, the flows that feed decisions on health care policy show the lowest maturity. The maturity matrix indicates that the data and information may circulate in silos of governance; experts are needed (level 2) to support the data to reach the policy makers. The below-average estimates on data and stakeholder governance hint that the coordination of the data flows is mostly implicit. The interviewees missed the explicit rules and coordination of the secondary use of data. The above-average estimates on data for clinical purposes indicate the success of standardizing patient data flows via the EHIS.

The project in Estonia combined the HISMM into the framework of analytics. The role of the HISMM was to provide insight into the improvement potential of the established flow of data. The findings of the maturity assessment allowed the stakeholders' viewpoints to be drawn to and the specific characteristics of the flows to be analyzed. The HISMM is a valuable tool for cases where the primary data policy, governance structure, and platform are already in place. Analysis of the specific characteristics provides a basis for targeted improvements. When a data flow misses expectations, an assessment may reveal a specific characteristic that limits the flow. The summary of the HISMM results in Estonia shows the need to improve the focus on data quality, data transport, data and stakeholder governance, and process alignment (Figure 3).

The study in Estonia concluded with a recommendation to align the roles of data providers, consumers, stewards, and custodians for the expanded multipurpose data flows. The current document-based health information sharing model shall transform into a shared space of Integrated Care Records. We reported the assessment results as part of the Person-Centered Integrated Hospital Master Plan, which also included reports of teams of other specialists.

XSL•FO
RenderX

Figure 2. The figure summarizes the Health Information Sharing Maturity Model assessment findings from Estonia. The matrix's cells depict the maturity levels grouped by the 11 capability characteristics in the vertical dimension and by the 6 stakeholder purposes in the horizontal dimension. The number in the cell shows the corresponding level of maturity, colored red for level 2, yellow for level 3, and green for level 4.

Perspective	Capability characteristic	Stakeholder purpose					
		1. Engagement	2. Clinical	3. Management	4. Research	5. Policy	6. Funding
Technology	1. Data quality	2	2	2	3	2	3
	2. Data transport	2	3	2	2	2	3
	3. Data security	4	4	4	4	4	4
	4. Interoperability	3	3	3	2	2	3
Use	5. Usability	3	3	3	3	2	3
	6. Alignment	2	3	3	3	2	2
	7. Participation	4	3	4	3	3	2
	8. Consent	3	3	3	3	2	2
Governance	9. Data governance	3	3	2	2	2	2
	10. Stakeholder governance	2	3	3	2	2	2
	11. Sustainability	4	2	2	4	3	4

Figure 3. The figure visualizes the maturity levels of Health Information Sharing Maturity Model characteristics in aggregation. It shows that data quality and transport, data and stakeholder governance, and process alignment maturity are lower than other characteristics.

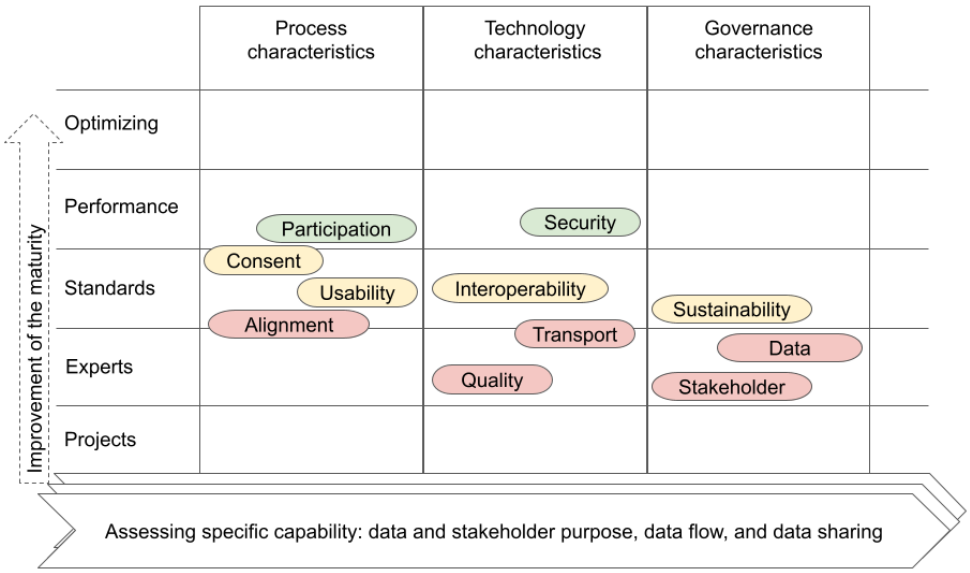

Comparison of the Assessments

We conducted the assessments according to the framework discussed earlier in this study. To map the elements of data sharing, we interviewed stakeholders using the shared data for the purposes defined by the scope of the projects. We also captured the evidence of the stakeholders interacting with specific data sharing platforms.

The data collection in the Kingdom of Saudi Arabia took place 1 year before the project in Estonia. We mostly replicated the methodological experience from the Kingdom of Saudi Arabia in Estonia, except that we introduced an additional assessment tool, the HISMM (Table 2).

The analysis of the projects in the Kingdom of Saudi Arabia and Estonia demonstrates the challenges of coordinating data flows on a distributed data sharing system. Even if the health care systems in the 2 countries are coordinated differently, in both cases, the conclusions focus on the need to strengthen data flow stewardship (Table 3).

The conclusions advised the governments to introduce governance policies, which would clarify the responsibilities of the stakeholders. Proper management of the responsibilities of data stewards would increase the value of data providers' contributions and the value of the custodians' data. Figure 4 illustrates our conceptual understanding of the data governance roles and their relationships, which we used as a tool to map the roles of the existing or future organizations in our recommendation.

Current data governance in the studied countries follows the vertical model of stewardship, where the data consumers coordinate the information flows for their own needs. Most

XSL•FO
RenderX

consumers have also established their own data management or custodian organizations. Notable exceptions are the national databases and message exchange platforms, which support data consumption by multiple institutions. For example, the EHIS manages data consumed by the network of health care providers, patients, and health care registries. A pervasive stewardship function shall increase the secondary use of data.

Table 2. The assessments share comparable attributes of scope. The only exception is that the project in Estonia conducted a maturity assessment, which was not part of the project's scope in the Kingdom of Saudi Arabia.

Attributes of scope	The Kingdom of Saudi Arabia	Estonia
Assessed stakeholder purposes	• Clinical, population health, and health care management	• Clinical, population health, patient engagement, health care funding, health care management, and health care policy
Assessed data sharing platforms	• Hospital EMRs[a] and analytical data sets	• Hospital EMRs, the national EHR[b], disease registries, and insurance claims registry
Assessed capability maturity characteristics	—[c]	• HISMM[d] technology, use, and governance
Stakeholders interviewed	• National-level health care coordination (Saudi Health Council) • Management of the sectors (MoH[e] Medical Services, National Guard Medical Services, Ministry of Interior Medical Services, and King Faisal Specialist Hospital & Research Center)	• National-level health care management and policy (Ministry of Social Affairs, North Estonia Medical Centre, Tartu University Clinic, Pärnu Hospital, Viljandi Hospital, Põlva Hospital, East Viru Central Hospital, the National Institute of Health Development, the Estonian Health Insurance Fund, and the Estonian Society of Cardiology)

[a]EMR: electronic medical record.

[b]EHR: electronic health record.

[c]Not applicable.

[d]HISMM: Health Information Sharing Maturity Model.

[e]MoH: Ministry of Health.

Table 3. The findings from the 2 countries demonstrate similarities in the expected achievements, identified barriers, opportunities, and principal conclusions. Regardless of the digitization of workplaces in both cases and sophisticated data integration solutions in the Estonian case, siloed data stewardship limits the multiuse of health data.

Attributes of findings	The Kingdom of Saudi Arabia	Estonia
Expected achievement	Timely and efficient delivery of health care system and public health indicators, and standard and special reports	Timely and efficient decision support for clinical, management, and financial decisions
Barriers	Lack of interoperability standards and siloed sectoral stewardship	Siloed vertical stewardship
Opportunities	Digitized workplaces in health care and cross-sectoral health care governance structures (SHC[a])	Digitized workplaces in health care, secure integration platform (XRoad), national EHR[b] (EHIS[c]), data and data exchange standards, and national-level clinical decision support
Principal conclusions	Align the roles of the stakeholders and engage the participants in a standardized data flow	Align the roles of stakeholders and standardize the event-driven sharing of EMRs[d]

[a]SHC: Saudi Health Council.

[b]EHR: electronic health record.

[c]EHIS: Estonian Health Information System.

[d]EMR: electronic medical record.

XSL•FO
RenderX

Figure 4. A conceptual overview of the data governance roles and relationships used as a base for mapping the actual organizational structure of the countries. A policy for multiple data use supports resolving demand and supply between data providers and consumers. The policy establishes authority and responsibility for pervasive data stewardship and custodianship.

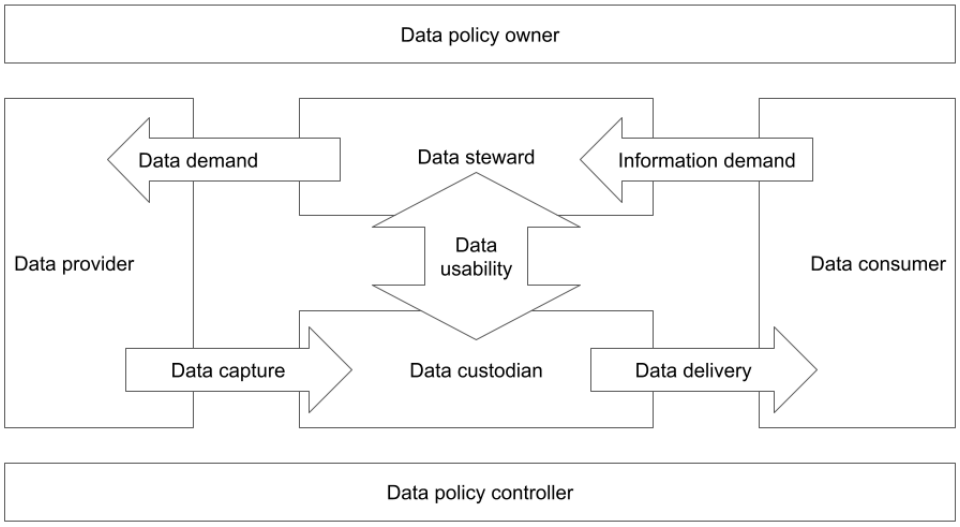

Discussion

Principal Findings

The study maps the common thread of thought from 2 distinct projects in the Kingdom of Saudi Arabia and Estonia, assessing the potential of secondary use of health care data and supporting governmental decisions. The projects apply comparable frameworks and methods, allowing the comparison of the barriers, opportunities, and conclusions. The findings include both the frameworks used and the conclusions made. The framework of assessment defines 2 dimensions of analysis. First, there is the need to identify and improve data sharing flows between data providers and consumers. The analysis investigates multiple purposes of data, data and exchange standards, stakeholders' governance, and shared data management platforms. Second, the framework considers the maturity of the data sharing implementation. The maturity assessment measures the level of institutionalization of health data sharing. The second part of the framework was included and applied only to the project in Estonia.

The assessment revealed opportunities and barriers in the secondary use of health data. Starting with the opportunities, the included institutions demonstrated high levels of digitization in the workplace. In the Estonian case, we also experienced advanced integration platforms and interoperability standards implemented nationally. The latter has supported the development of sophisticated solutions for national electronic health record and clinical decision support. However, the countries have maintained a fragmented organization of data stewardship, which has not been able to coordinate the need for data. In both cases, the assessments concluded with a recommendation to implement pervasive data stewardship to align the need for data.

While many countries have digitalized information necessary for clinical work and described the data relatively well, especially in most European countries and North America, unified routines and applications for the secondary use of digital data in health care are largely still being planned.

In Sweden, more than 100 health care quality registries collect individual-based clinical data for research and improvement of health care delivery [26]. In Estonia, 6 medical registries and databases collect, process, and distribute data about health and medicine [27]. Studies propose that clinical quality registers can be cost-effective and yield significant investment returns [28]. The number and quality of the registries indicate success in the secondary use of health care data. The registries also introduce data capture, integration, and delivery costs for secondary use. These professional specialty or national quality registries are often developed and managed in silos, leading to high maintenance costs and challenges in interoperability.

Regardless of the advanced information systems in hospitals, the health care system in the Kingdom of Saudi Arabia spends considerable time and resources collecting statistical data. There is much manual processing due to the lack of standards for integration and semantics. The same applies to Estonia; despite the common health data interoperability standards and transport system, secondary use of health care data is often in silos and needs additional effort. In the Kingdom of Saudi Arabia and Estonia, the data consumers coordinate their needs directly with the data providers, reinforcing the traditional model of form-based reporting. The form-based approach introduces duplication at the data capture; one may call it secondary capture. To avoid resource wasting and duplication, collecting the data consumers' need shall be part of the standards of primary data capture.

In Estonia, the advanced semantic interoperability of the clinical documents shared via the EHIS enables the automation of clinical decision support [29-32]. Such features include drug-drug interaction alerting, context-driven suggestions of clinical guidelines, and automatic patient summaries based on clinical documents. These features increase the use of data but only inside the vertical of clinical decision-making. The EHIS could also facilitate data flows for public health, health care management, and clinical research decision-making.

XSL•FO
RenderX

The stakeholders of health care data need to cooperate through a strategic digitalization process. Often, the participants are not aware of the discontinuity of the data flow. The study in Estonia indicated that the participants were relatively satisfied with the data management tools and their engagement in the flow. Instead, they reported problems with data quality and governance. The users expressed their frustration regarding duplicate data capture but could not relate it to the low alignment of the processes. We hypothesize that the interview results indicate disruptions in the data flow. The respondents struggled to find source data to fill in the data entry forms for secondary use. Designing and managing a flow that connects data capture with a single consumer is relatively easy, ensuring satisfaction with the tools and participation. Only a helicopter view of the landscape of data needs shows the shortcomings of governance and the chronic waste of capturing the same data repeatedly. Efficiency in the secondary use of data starts from the health care policy establishing clear goals and management.

Single-purpose capture of data and single-purpose databases are indicators of the low secondary use of health care data. The data flow design should follow the principle of "collect once and use multiple times." The studied cases reveal the barriers built between the domains of information purpose. The health care system extends over 6 ministries in the Kingdom of Saudi Arabia. It takes a long time and heavy work to combine data across the borders of the governance verticals. In Estonia, where the organization is more straightforward, data collection for different purposes is still split between different data consumers, resulting in independent data flow designs without proper interoperability. For example, health care providers must simultaneously record the exact data for clinical decisions, management, funding, statistics, and research. The new policy shall require the unification of the demand of data consumers into a single standard of data capture.

The analysis of health data sharing challenges in the Kingdom of Saudi Arabia and Estonia demonstrates that the digitization of the workplaces, integration of information systems, and advanced semantic interoperability are insufficient for secondary use on a large scale. A prerequisite for secondary use is a health care policy that emphasizes the need for the continuity of the data flow. The health policy should address governance of the data and stakeholders without introducing central bottlenecks for innovation. The policy should guide parties to map the impacts of their data processing and increase the value of their data through greater secondary use. A health data sharing system shall reward the measurable secondary use of data assets.

Advances in the digitization of health data and integration of information systems open the way to the digitalization of health care processes. Shared data enables the coordination of activities of a digital process. Stewards and custodians must govern health care data through the diverse organizations, workplaces, and information systems landscape. Data governance conceptualizes and carries out stewardship responsibilities based on data access, custodianship, and use policies [33]. The conceptual framework for data governance by Abraham et al [6] suggests structural, procedural, and relational mechanisms. The structural and relational mechanisms include establishing clear organizational responsibility and communication. The governments in demand for greater secondary use of health care data shall establish data policy with precise coordinating mechanisms.

We saw that digitalizing data providers and consumers is insufficient for efficient secondary use of data. There is a need for a DHP that enables data and information sharing. However, having a DHP only for clinical data is insufficient. For secondary use, the stewardship must include the requirements of all targeted consumers. This recommendation is also very much in line with the observation from a 2021 report that calls for more substantial public-patient participation in the secondary use of health data [2].

There is an ever-growing demand for better data and information for decision-making. Modern health care and research depend on data from various domains, including education, environment, and social care. It is an ongoing effort to analyze and integrate the new demand for data.

Limitations

The study reports the findings from 2 projects from 2 countries. The findings present certain commonalities but still have a limited generalizability for different contexts. Countries or regions searching for advice may present circumstances that demand noticeably different strategies for their digital health improvement. It is also essential to understand the role of the frameworks when trying to replicate the results. For practical reasons, a solid framework is essential for such projects, as effective planning, execution, and analysis require a rigid structure. However, the choice of a framework indicates the researcher's focus and may lead to a limited space of findings. The studied projects develop policy suggestions for health data governance on the national level. Controlled empirical validation of the suggested policies is nearly impossible. The conclusions mainly depend on the internal validity of the research, where we build on the experience of the involved stakeholders and findings from similar experiments.

Conclusions

In this study, we have analyzed 2 projects that assessed and provided advice for the national-level improvement of the secondary use of health care data. The study provided an overview of the projects' backgrounds, frameworks, methods, and results. Finally, we discussed the main advice from the projects. The study shows that 2 high-income countries with very different health care systems have comparable issues with the secondary use of health care data. National-level secondary use shall build on an overarching data policy that enables horizontal stewardship to coordinate requirements of a diverse landscape of health care data consumers.

XSL•FO
RenderX

Acknowledgments

The project's client in the Kingdom of Saudi Arabia was the Saudi Health Council, and the project funding came from the World Bank. The Ministry of Social Affairs ordered the project in Estonia and received funding from the European Union Structural Reform Support Program. The authors would like to thank all the participants in this study.

Authors' Contributions

JM and PR were the main contributors to both studied projects, and ZS was a leading contributor in the case of the Kingdom of Saudi Arabia. DD worked on the structure and methodology of the study. TN, one of the leading developers of Health Information Sharing Maturity Model (HISMM), contributed to the required details and reviewed the results concerning HISMM adoption.

Conflicts of Interest

None declared.

References

1. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
2. Boyd M, Zimeta DM, Tennison DJ, Alassow M. Secondary use of health data in Europe. Open Data Institute. 2021. URL: https://secondary-use-health-data.theodi.org/assets/pdf/Report-2021.pdf [accessed 2024-04-23]
3. National master plan for health data analytics: harmonizing the collection and use of data in Saudi Arabia. General Directorate for National Health Economics and Policy, Saudi Health Council, Kingdom of Saudi Arabia. 2020. URL: https://shc.gov.sa/Arabic/NewStrategy/DepartmentAchievements/Pages/StrategyReports.aspx [accessed 2024-04-29]
4. Person-centred integrated hospital master plan in Estonia. The Ministry of Social Affairs, Estonia. URL: https://www.sm.ee/haiglavork-ja-haiglavorgu-arengukavad [accessed 2022-07-18]
5. Global reference list of 100 core health indicators. World Health Organization. URL: https://iris.who.int/handle/10665/259951 [accessed 2021-07-20]
6. Abraham R, Schneider J, vom Brocke J. Data governance: a conceptual framework, structured review, and research agenda. Int J Inf Manage 2019 Dec;49:424-438. [doi: 10.1016/j.ijinfomgt.2019.07.008]
7. Data governance: driving value in healthcare. Klynveld Peat Marwick Goerdeler (KPMG). URL: https://kpmg.com/xx/en/home/insights/2018/06/data-governance-driving-value-in-health.html [accessed 2023-07-27]
8. Carvalho JV, Rocha Á, Abreu A. Maturity models of healthcare information systems and technologies: a literature review. J Med Syst 2016 Jun 15;40(6):131. [doi: 10.1007/s10916-016-0486-5] [Medline: 27083575]
9. Hamilton G, Novak T. Health information sharing maturity model. MITRE Health. 2018. URL: https://health.mitre.org/blog/health-information-sharing-maturity-model-2/ [accessed 2020-06-15]
10. ISO/IEC 33020:2019 information technology- process assessment- process measurement framework for assessment of process capability. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/78526.html [accessed 2024-04-14]
11. CMMI levels of capability and performance. CMMI Institute. URL: https://cmmiinstitute.com/learning/appraisals/levels [accessed 2024-04-19]
12. Statistical yearbook. Ministry of Health of Saudi Arabia. 2010. URL: https://www.moh.gov.sa/en/Ministry/Statistics/book/Pages/default.aspx [accessed 2024-04-29]
13. World population ageing. United Nations. 2019. URL: https://digitallibrary.un.org/record/3846855?v=pdf [accessed 2024-04-29]
14. Statistical yearbook. Ministry Of Health Saudi Arabia. 2020. URL: https://www.moh.gov.sa/en/Ministry/Statistics/book/Pages/default.aspx [accessed 2021-05-11]
15. Alharbi M. The status quo of health information technology and health information management efficiency in Saudi Arabia: a narrative review. The University of New England. 2018. URL: https://hdl.handle.net/1959.11/23228 [accessed 2021-05-06]
16. Almalki M, Fitzgerald G, Clark M. Health care system in Saudi Arabia: an overview. East Mediterr Health J 2011 Oct;17(10):784-793 [FREE Full text] [doi: 10.26719/2011.17.10.784] [Medline: 22256414]
17. Alkraiji AI, Jackson T, Murray I. Factors impacting the adoption decision of health data standards in tertiary healthcare organisations in Saudi Arabia. J Enterp Inf Manag 2016 Sep 12;29(5):650-676. [doi: 10.1108/jeim-11-2014-0111]
18. Attallah N, Gashgari H, Al Muallem Y, Al Dogether M, Al Moammary E, Almeshari M, et al. A literature review on health information exchange (HIE). Stud Health Technol Inform 2016;226:173-176. [Medline: 27350496]
19. Vision and mission. Saudi Health Council. URL: https://shc.gov.sa/en/AboutCouncil/Pages/Vision.aspx [accessed 2021-05-18]
20. Home page. National Health Information Center. URL: https://nhic.gov.sa/ [accessed 2021-05-19]
21. Habicht T, Reinap M, Kasekamp K, Sikkut R, Aaben L, van Ginneken E. Estonia: health system review. Health Syst Transit 2018 Mar;20(1):1-189 [FREE Full text] [Medline: 30277217]
22. World population ageing. United Nation. URL: https://digitallibrary.un.org/record/3846855 [accessed 2021-05-12]

XSL•FO
RenderX

23. Health statistics and health research database. Statistika. URL: https://statistika.tai.ee/index_en.html [accessed 2021-05-17]
24. Health services organisation act. Riigi Teataja. URL: https://www.riigiteataja.ee/en/eli/ee/508012018001/consolide/current [accessed 2023-07-27]
25. Situation analysis on evidence-informed health policy-making: Estonia. World Health Organization. URL: https://www.who.int/europe/publications/i/item/WHO-EURO-2019-3513-43272-60649 [accessed 2021-07-14]
26. Emilsson L, Lindahl B, Köster M, Lambe M, Ludvigsson JF. Review of 103 Swedish healthcare quality registries. J Intern Med 2015 Jan;277(1):94-136 [FREE Full text] [doi: 10.1111/joim.12303] [Medline: 25174800]
27. Registers. Tervise Arengu Instituut. URL: https://www.tai.ee/et/statistika-ja-registrid [accessed 2023-06-15]
28. Lee P, Chin K, Liew D, Stub D, Brennan AL, Lefkovits J, et al. Economic evaluation of clinical quality registries: a systematic review. BMJ Open 2019 Dec 15;9(12):e030984 [FREE Full text] [doi: 10.1136/bmjopen-2019-030984] [Medline: 31843824]
29. The decision support of the Estonian Health Insurance Fund. TEHIK. URL: https://www.tehik.ee/en/the-decision-support-of-the-estonian-health-insurance-fund [accessed 2024-04-29]
30. Patient-centered decision support. Estonian Health Insurance Fund. URL: https://www.tervisekassa.ee/en/node/166395/clinical-decision-supports/patient-centered-decision-support [accessed 2024-04-29]
31. Drug interaction decision support. Estonian Health Insurance Fund. URL: https://www.tervisekassa.ee/en/node/166395/clinical-decision-supports/drug-interaction-decision-support [accessed 2024-04-29]
32. Marandi T, Orav L, Kändmaa A, Nahkur S. Drug-drug interaction database SFINX – first results from north Estonia Medical Centre, Estonia. Clin Ther 2015 Aug;37(8):e5-e6. [doi: 10.1016/j.clinthera.2015.05.026]
33. Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. Health Serv Res 2010 Oct 02;45(5 Pt 2):1442-1455 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01140.x] [Medline: 21054365]

Abbreviations

CMMI: Capability Maturity Model Integration
DHP: Digital Health Platform
EHIF: Estonian Health Insurance Fund
EHIS: Estonian Health Information System
HISMM: Health Information Sharing Maturity Model
MoH: Ministry of Health
MoSA: Ministry of Social Affairs
NHIC: National Health Information Center
NIHD: National Institute for Health Development
PHC: primary health care
SHC: Saudi Health Council
WHO: World Health Organization

Edited by S Ma, T Leung; submitted 05.10.23; peer-reviewed by A Civit, A Gangadhara Rao; comments to author 15.03.24; revised version received 03.05.24; accepted 03.05.24; published 08.08.24.

Please cite as:
Metsallik J, Draheim D, Sabic Z, Novak T, Ross P
Assessing Opportunities and Barriers to Improving the Secondary Use of Health Care Data at the National Level: Multicase Study in the Kingdom of Saudi Arabia and Estonia
J Med Internet Res 2024;26:e53369
URL: https://www.jmir.org/2024/1/e53369
doi:10.2196/53369
PMID:

©Janek Metsallik, Dirk Draheim, Zlatan Sabic, Thomas Novak, Peeter Ross. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptance of Electronic Labeling for Medicinal Product Information Among Malaysian Hospital Patients: Cross-Sectional Study

Xin Yee Loh[1], PharmB; Ai Ling Woo[2], PharmB; Azwa Haris[3], PharmB; Cheryl Shajini Pereira[3], PharmB; Bee Kim Tan[1,4], PharmB, MSc, PhD

[1]School of Pharmacy, Faculty of Health and Medical Sciences, Taylor's University, Subang Jaya, Malaysia
[2]AstraZeneca Limited, Petaling Jaya, Malaysia
[3]Department of Pharmacy, University Malaya Medical Centre, Kuala Lumpur, Malaysia
[4]Digital Health and Medical Advancement Impact Lab, Subang Jaya, Malaysia

Corresponding Author:
Bee Kim Tan, PharmB, MSc, PhD
School of Pharmacy
Faculty of Health and Medical Sciences
Taylor's University
1 Jalan Taylors
Subang Jaya, 47500
Malaysia
Phone: 60 3 5629 5000
Fax: 60 3 5629 5001
Email: beekim.tan@taylors.edu.my

Abstract

Background: While perceptions of electronic labeling (e-labeling) in developed countries have been generally positive, existing data primarily come from studies involving hospital pharmacists, community pharmacy customers who may not be frequent medication users, and individuals receiving COVID-19 vaccines.

Objective: This study aims to assess e-labeling acceptance, perceptions of its benefits, challenges with its implementation, and preferences among hospital ambulatory care patients in Malaysia. Additionally, the study investigates the factors influencing patients' acceptance of e-labeling.

Methods: A cross-sectional study using a 28-item questionnaire was conducted at the outpatient pharmacy department of a quaternary hospital in Kuala Lumpur, Malaysia, from May to June 2023. The questionnaire was developed based on a review of published literature related to e-labeling and was guided by the Unified Theory of Acceptance and Use of Technology, second version (UTAUT2). Patients aged 18 years and above were recruited using a stratified sampling method to ensure representative age-related medication usage. A mobile tablet was provided to patients for self-completion of the e-survey in their preferred language (English, Malay, or Mandarin). Categorical data on e-labeling acceptance, perceptions, and preferences were analyzed using descriptive statistics. Qualitative content analysis was performed to characterize participants' responses to open-ended questions. Univariate and multivariate binomial logistic regression analyses were conducted to identify predictors of e-labeling acceptance.

Results: Out of 462 patients approached, 387 (83.8%) participated in the survey, with 283 (73.1%) accepting e-labeling. Most participants perceived the electronic version of the package insert as beneficial, particularly for understanding their medication better through the choice of language (352/387, 91.0%). However, around half of the participants (197/387, 50.9%) expressed concerns about the potential risks of obtaining illegal medication information via e-labeling. Most participants (302/387, 78.0%) preferred to access electronic leaflets through government websites. However, 221/387 (57.1%) still wanted the option to request printed leaflets. Significant predictors of e-labeling acceptance included perceived benefits such as better understanding of medication (adjusted odds ratio [AOR] 8.02, 95% CI 2.80-22.97, $P<.001$), environmental protection (AOR 7.24, 95% CI 3.00-17.51, $P<.001$), and flexibility in information retrieval (AOR 2.66, 95% CI 1.11-6.35, $P=.03$). Conversely, being of Chinese ethnicity compared with Malay (AOR 0.28, 95% CI 0.13-0.60, $P=.005$) and perceived lack of self-efficacy in browsing electronic leaflets (AOR 0.25, 95% CI 0.11-0.56, $P<.001$) were associated with lower acceptance.

XSL•FO
RenderX

Conclusions: The acceptance rate for e-labeling among hospital ambulatory care patients was moderately high and was significantly influenced by ethnicity as well as patients' perceived benefits and challenges related to its implementation. Future strategies to enhance e-labeling uptake should address patient concerns regarding the challenges of using the digital platform and emphasize the benefits of e-labeling.

(J Med Internet Res 2024;26:e56591) doi:10.2196/56591

KEYWORDS

electronic labeling; e-labeling; electronic medication information; patient preference; acceptance; hospital ambulatory care patient

Introduction

Pouliot et al [1] defined medication literacy as an individual's ability to make safe decisions regarding medications and health based on the processing of patient-centered medication information (eg, written, oral, and visual). Patients with limited medication literacy often struggle with reading medication labels, understanding printed care instructions and health advice, and tend to use medications inappropriately. They are also less adherent to therapy [2,3]. The National Health Morbidity Survey Malaysia 2019 reported that 1 in 3 adult Malaysians has poor health literacy [4]. The high prevalence of poor health literacy among the public is concerning, as those with low health literacy are more likely to incur higher health care costs, placing a tremendous burden on the health care system. Despite health care professionals (HCPs) conveying medication information verbally, most patients have limited cognitive ability to retain orally transmitted information [5]. Therefore, medicinal product information leaflets can serve as a useful aid, in addition to verbal counseling, to address postconsultation gaps [6]. In Malaysia, there are 2 categories of health authority–approved medicinal product information: Package Inserts (PIs) and Consumer Medication Information Leaflets (RiMUP) [7]. RiMUP is written in layman's terms and is available in both English and Malay. While PIs are legally required to be printed and enclosed with all products containing scheduled poisons and injectable over-the-counter medicines, distributing RiMUP with the product is optional.

Patients who receive and read medicinal product information leaflets are more likely to discuss their medications with their health care providers. This creates an opportunity for providers to use the leaflets as educational material to improve treatment knowledge and facilitate shared decision-making [8]. A majority (64.9%) of participants in a Saudi Arabian study reported that their medication adherence improved after reading medicinal product information leaflets [9]. However, 55% of respondents in a survey conducted by the European Association of Hospital Pharmacists revealed that hospital patients do not receive medicinal product information leaflets [10]. Furthermore, patients reported several issues with paper PIs that hinder their effective use, such as undersized fonts, medical jargon, and the lack of options in local languages [9,11,12].

Electronic labeling (e-labeling) is an emerging trend in disseminating legally approved medicinal product information in dynamic formats, such as XML. By leveraging digital advancements, e-labeling systems enable the use of personalized medication information to meet the specific needs of both patients and HCPs. In today's digital age, as the public increasingly seeks online health information, e-labeling offers a convenient way to access regulatory-approved medicinal product information through a trusted channel [13]. E-labeling not only streamlines the updating process, enabling prompt dissemination of medicinal product information to a wide range of HCPs and patients, but also creates opportunities for integration into digital health services. This can support e-prescribing by reducing the risk of medication incompatibilities and enhancing patient safety. From the industry's perspective, electronic provision of medicinal product information reduces logistical challenges in label updates, lowers printing costs for PIs, and improves the efficiency of the global pharmaceutical supply chain through shared labeling between countries. These advantages of e-labeling collectively contribute to achieving the United Nations Sustainable Development Goals (UNSDG) 3 (Good health and well-being) and 12 (Responsible consumption and production) [14].

To date, e-labeling is regulated at varying levels across different regions worldwide. The United States Food and Drug Administration (FDA) has mandated the electronic distribution of PIs since 2015 [15]. By contrast, only selected hospital medicinal products in the Baltic countries have been granted marketing authorization through e-labeling [12]. In Asia, Japan's Pharmaceuticals and Medical Devices Agency (PMDA) officially enforced the removal of paper PIs for prescription drugs and medical devices in July 2023, aiming to transition toward a paperless system [16]. Malaysia's National Pharmaceutical Regulatory Agency (NPRA) released the Guideline on Electronic Labelling (E-labeling) for Pharmaceutical Products, which came into effect on May 1, 2023. According to this guideline, approved PIs, RiMUPs, or both must be provided electronically via a machine-readable QR code on the product's outer carton or inner label, linking to the NPRA QUEST system. The implementation of e-labeling is voluntary and applies to newly registered pharmaceutical products, biologics, and generic products containing scheduled poisons. An extension to other categories of products is still under review [17].

Understanding patients' acceptance, perception, and preferences regarding e-labeling can help implement a more patient-centric approach to foster engagement. Such data are lacking in Malaysia, a developing country in Southeast Asia with a multiethnic population and a unique socioeconomic context that could influence the public's readiness to adopt new digital health services. While perceptions of e-labeling in developed countries are generally positive, concerns have been reported, particularly among older adults and those with low digital literacy. Moreover, data representing patients were primarily obtained

XSL•FO
RenderX

from HCPs [10], customers visiting community pharmacies who may not be frequent medication users [11], and individuals receiving COVID-19 vaccines [18]. These data cannot be generalized to hospital ambulatory care patients, who are primarily managing chronic diseases and require ongoing medication information for self-administration. Therefore, this study aimed to assess e-labeling acceptance, perceptions of its benefits, challenges with local implementation in Malaysia, and preferences among hospital ambulatory care patients. Additionally, the factors influencing patients' acceptance of e-labeling were investigated.

Methods

Study Design and Population

A cross-sectional study was conducted at the University Malaya Medical Centre (UMMC), a quaternary teaching hospital with 1617 beds and multidisciplinary clinics spanning 40 clinical specialties. Established in 1962 and located in Kuala Lumpur, the capital of Malaysia, UMMC served as the study site. The study population consisted of a convenience sample of patients who visited the outpatient pharmacy department of UMMC. The inclusion criteria were patients aged 18 years and older; collecting prescription medications; and capable of reading English, Malay, or Mandarin. Exclusion criteria were patients with limited cognitive abilities, those collecting medication on behalf of others, those who were unwell, or those who refused to participate.

The calculated minimum sample size was 386, based on government hospital outpatient statistics from 2020, which totaled 16,635,350 [19]. This calculation used a 5% margin of error, a confidence level of 95%, and a response distribution of 50% [20]. A stratified sampling method was used based on the estimated proportions of prescription drug users: 18.0% under 40 years of age, 46.0% between 40 and 64 years of age, and 85.0% over 65 years old, to ensure age representativeness [21].

Questionnaire

The study instrument was a 28-item questionnaire developed based on a review of the published literature related to e-labeling and informed by the Unified Theory of Acceptance and Use of Technology, second version (UTAUT2) [22]. The questionnaire consisted of 3 sections/pages: (A) demographics (4 items) and utility of medicinal product PIs (6 items), (B) perceptions of the benefits (6 items) and challenges with e-labeling implementation (5 items), and (C) acceptance (1 item) and preferences regarding e-labeling (6 items). Acceptance, as defined by Adell et al [23], is the willingness to use a system based on theoretical knowledge or experience.

The demographic information of the participants included age, gender, ethnicity, and education level. Age was categorized into 3 groups: 18-39, 40-64, and 65 and older. Ethnicity was categorized into Malay, Chinese, Indian, and other, while education level was classified as university/college, secondary school, primary school, and no formal schooling. The utility characteristics of medicinal product PIs captured included the sources of written medicine information, reasons for choosing these sources, and the frequency and reasons for reading PIs. Participants' practices regarding the frequency of using PIs were rated as follows: always, sometimes, only when receiving new medication, or never. Acceptance, perceived benefits (such as ease of retrieval, medication understanding, personalization, up-to-date information, and environmental protection), perceived challenges (including issues with electronic gadgets, digital literacy, internet access, and label security), and preferences regarding e-labeling (such as format, access, and options) were assessed using a 5-point Likert scale, ranging from 1=strongly disagree to 5=strongly agree. The questions were presented in a choice format, except for 2 open-ended questions designed to elicit reasons for not reading PIs and to gather additional views from participants regarding e-labeling.

The content validity of the questionnaire was assessed by 6 subject matter experts using a 4-point Likert scale to evaluate the relevance of each survey item. The experts included 2 regulatory pharmacists, 2 hospital pharmacists, and 2 academic pharmacists. The degree of relevance was categorized into 2 groups: "not relevant" and "somewhat relevant" were considered as "0=irrelevant," while "quite relevant" and "very relevant" were considered as "1=relevant." The scale-level content validity index based on the average method and the universal agreement method were 0.97 and 0.85, respectively, meeting the satisfactory level (≥0.83). The wording of some questions and choices was modified following a discussion within the research team based on the feedback received.

The English version of the questionnaire was translated into Malay and Mandarin using forward and backward translation methods. The translation was performed by 2 native Malay speakers proficient in English. The translations were reviewed by the research team and reconciled into an optimal version based on the appropriateness of the wording. This reconciled version was then back-translated into English by 2 additional native Malay speakers with a strong command of English. The original English version of the questionnaire and its translations were compared by the research team, and any discrepancies were discussed (Multimedia Appendix 1). Revisions were made to the Malay version of the questionnaire as needed (Multimedia Appendix 2). A similar translation method was used for the Chinese version of the questionnaire (Multimedia Appendix 3).

To ensure the feasibility of the recruitment procedure and the face validity of the questionnaire, a pilot study was conducted with 30 participants (10 for each language: English, Malay, and Mandarin) through cognitive debriefing to assess clarity and understanding. Cronbach α coefficients were calculated, resulting in .73 for section B on perceived benefits and .79 for challenges with e-labeling implementation. The usability and technical functionality of the electronic questionnaire were tested before the pilot study.

Data Collection

Data collection took place from May to June 2023 at the outpatient pharmacy department of UMMC. Potential participants were approached by XYL and invited to participate while waiting for their prescriptions to be filled. Those who agreed to participate were briefed on the study objectives and the estimated time required to complete the survey. A tablet was provided to each participant to indicate their informed

XSL•FO
RenderX

consent, followed by the self-completion of the e-survey form in their preferred language (English, Malay, or Mandarin). The survey was designed to be open, allowing participants to review and change their answers using a back button. Completeness was ensured before survey submission through mandatory items in the e-survey form. Participation was voluntary, and participants could opt out without facing any negative consequences. No incentives were provided. Participants' data were anonymized and stored in a password-protected file.

Data Analysis

Statistical analysis was performed using the SPSS software (version 27; IBM Corp.). Descriptive statistics, including frequencies and percentages, were generated for all categorical variables. To facilitate analysis, responses for acceptance, perceived benefits, perceived challenges, and preferences were dichotomized into 2 categories: "strongly disagree," "disagree," and "neutral" were classified as "No," while "agree" and "strongly agree" were classified as "Yes." Univariate logistic regression was used to test the effect of each independent variable (demographic characteristics, utility of medicine PI, perceived benefits, and perceived challenges) on the probability of acceptance of e-labeling. Covariates with $P<.25$ were selected [24] and subsequently tested in a multivariate logistic regression model to identify significant predictors of e-labeling acceptance. P values <.05 in the multivariate logistic regression model were considered statistically significant.

A qualitative content analysis was performed following the 8 steps outlined by Zhang and Wildemuth [25] to characterize participants' responses to the open-ended questions. The procedures included the following: (1) Importing participants' response text data into qualitative data analysis software (NVivo version 10; Lumivero). (2) Coding data related to participants' reasons for not reading the PI and their opinions on e-labeling implementation. (3) A coding scheme and a list of initial categories were developed using the constant comparison method. (4) To validate the coding scheme and ensure consistency, 2 researchers (XYL and BKT) independently coded the data from the first 5 participants. The coding by both researchers was found to be in agreement. (5) XYL then coded the remaining data and added new categories as needed. (6) BKT assessed the coding consistency against the raw data. (7) The categories/themes were refined based on the patterns observed in the coded data. Homogeneity of codes within each category and heterogeneity of codes across categories were reviewed to ensure there was no overlap; and finally, (8) the inductive content analysis process and results were reported descriptively.

Ethics Approval

The study was granted ethics approval by the Medical Ethics Committee of UMMC (MREC ID Number: 2023214-12138, dated April 3, 2023), conducted in accordance with the Declaration of Helsinki, and reported according to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [26].

Results

Participant Demographics and Characteristics of Package Insert Use

Out of the 462 patients approached, 387 agreed to participate and completed the e-survey, resulting in a response rate of 83.8%. Participant demographics and characteristics of PI use are summarized in Tables 1 and 2. Participants were predominantly male (n=202, 52.2%), Chinese (n=185, 47.8%), aged 40-64 years (n=173, 44.7%), and had a university/college education (n=289, 74.7%). Among the 387 participants, more than three-quarters (n=312, 80.6%) reported seeking written information about their medication. PI was the second most popular source of written medicine information, with a utility rate of 34.6% (n=108), following the internet (n=188, 60.3%). By contrast, only 1.6% (n=5) of participants read the RiMUP published on the NPRA website. The internet was perceived as more readily accessible (162/299, 54.2%, vs 76/176, 43.2%) and easier to understand (98/299, 32.8%, vs 48/176, 27.3%) compared with PI, but was considered less trustworthy (16/299, 5.4%, vs 47/176, 26.7%). Most participants read the PI only when they received a new medication (n=125, 40.1%).

Among the 48 responses to an open-ended question, reasons for not reading the PI included the leaflets being voluminous and containing too much information (n=15, 31.2%), small font size that is hard to read (n=10, 20.8%), difficulty understanding medical terms (n=9, 18.7%), preference for Google due to convenience (n=6, 12.5%), the paper being too small (n=5, 10.4%), and already being well-informed by doctors (n=3, 6.2%). Conversely, side effects (228/824, 27.7%) and information on the medication's purpose and how it works (224/824, 27.2%) were the main reasons for reading the PI (Tables 1 and 2).

XSL•FO
RenderX

Table 1. Participant demographic (N=387) and characteristics of package insert use (Ns=312).

Demographic characteristics and package insert utility	Values, n (%)
Age (years)	
18-39	108 (27.9)
40-64	173 (44.7)
65 and above	106 (27.4)
Gender	
Male	202 (52.2)
Female	185 (47.8)
Ethnicity	
Malay	109 (28.2)
Chinese	185 (47.8)
Indian and others	93 (24.0)
Highest level of education	
University/college	289 (74.7)
Secondary school	92 (23.8)
Primary school	5 (1.3)
No formal schooling	1 (0.3)
Obtain or seek medicine written information	
Yes	312 (80.6)
No	75 (19.4)
Frequency of package insert use	
Only when I receive a new medication	125 (40.1)
Sometimes	119 (38.1)
Always	40 (12.8)
Never	28 (9.0)
Reason for package insert use (n=824)	
Side effects	228 (27.7)
Medication purpose and how it works	224 (27.2)
Dosage or administration	190 (23.1)
Drug interactions or precaution with other diseases	143 (17.4)
Safety in pregnancy and breastfeeding	34 (4.1)
Others	5 (0.6)

XSL•FO
RenderX

Table 2. Source of written information about medicine (N=312).

Source of written information about medicine	Total, n (%)	Reasons for the chosen source[a], n/N (%)				
		Trustworthy	Easy to understand	Readily accessible	Recommended by others	Other reasons
Internet (eg, Google)	188 (60.3)	16/299 (5.4)	98/299 (32.8)	162/299 (54.2)	11/299 (3.7)	12/299 (4.0)
Package insert	108 (34.6)	47/176 (26.7)	48/176 (27.3)	76/176 (43.2)	4/176 (2.3)	1/176 (0.6)
Leaflets from health care professionals	8 (2.6)	4/12 (33.3)	7/12 (58.3)	1/12 (8.3)	0/12 (0)	0/12 (0)
RiMUP[b] on the NPRA[c] website	5 (1.6)	2/8 (25.0)	2/8 (25.0)	4/8 (50.0)	0/8 (0)	0/8 (0)
Others	3 (1.0)	3/7 (42.9)	1/7 (14.3)	2/7 (28.6)	1/7 (14.3)	0/7 (0)

[a]Participants can choose more than 1 reason for the chosen source of written information about medicine.

[b]RiMUP: Consumer Medication Information Leaflets.

[c]NPRA: National Pharmaceutical Regulatory Agency.

Perceived Benefits and Challenges With e-Labeling Implementation

Most participants strongly agreed or agreed that the electronic version of the PI is beneficial, with 352/387 (91.0%) appreciating the ability to understand their medication better through their preferred language, 348/387 (89.9%) valuing the inclusion of images and videos, and 344/387 (88.9%) benefiting from advanced features such as adjustable font size and keyword search. Participants also agreed that e-labeling could help protect the environment by reducing paper use (340/387, 87.9%); provide the most up-to-date medication information (325/387, 84.0%); and allow access to information anywhere, anytime, without the fear of losing it (325/387, 84.0%; Figure 1).

Figure 1. Participants' perceived benefits toward electronic labeling implementation.

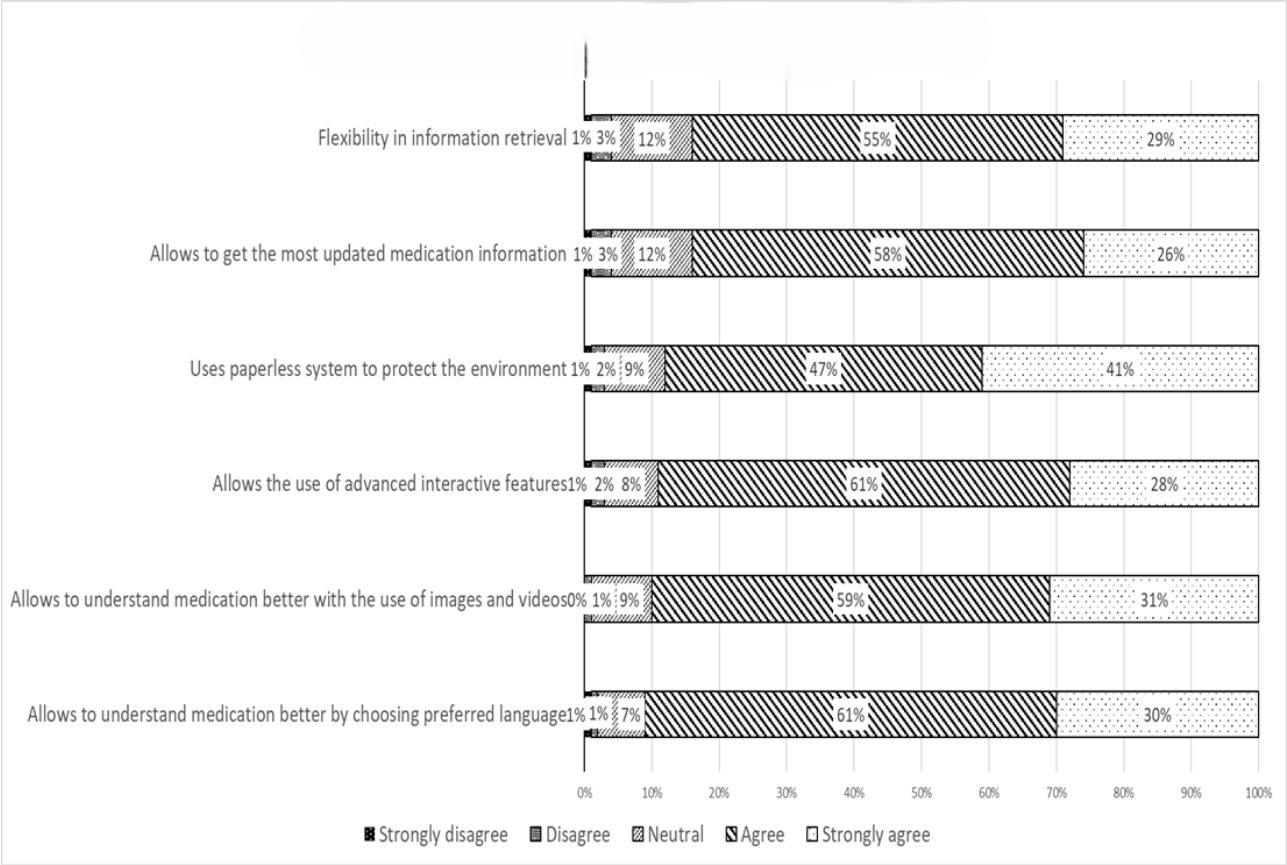

At the same time, around half of the participants (197/387, 50.9%) were concerned about obtaining potentially illegal medication information via e-labeling. A minority of participants expressed concerns about limited skills in browsing electronic medicinal product information (70/387, 18.1%), limited skills in using electronic gadgets (39/387, 10.1%), limited internet access (27/387, 7.0%), and not owning electronic gadgets (23/387, 5.9%; Figure 2).

XSL•FO
RenderX

Figure 2. Participants' perceived challenge towards electronic labeling implementation.

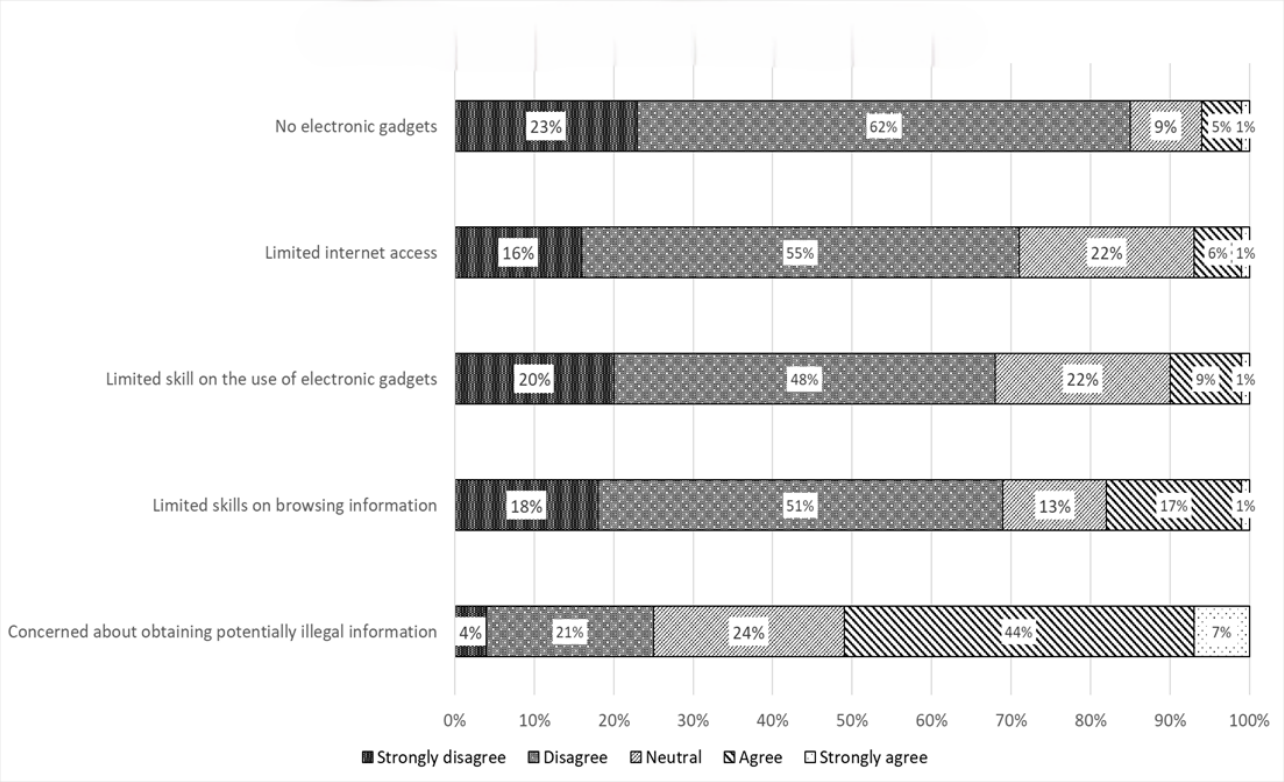

Acceptance of e-Labeling and Influencing Factors

Overall, the participants' acceptance rate of e-labeling was moderately high at 283/387 (73.1%; Figure 3). Univariate regression analysis revealed that all independent variables—including demographic characteristics, utility of PI, perceived benefits, and perceived challenges with e-labeling implementation—were potential factors associated with e-labeling acceptance (P<.25; Table 3).

Using a forward stepwise elimination method, multivariate regression analysis (Table 3) showed that participants who perceived a benefit in understanding medication better through images and videos were 8 times more likely to accept e-labeling (adjusted odds ratio [AOR] 8.02, 95% CI 2.80-22.97, P<.001). Those who perceived a benefit in using a paperless system to protect the environment had a 7 times higher probability of acceptance (AOR 7.24, 95% CI 3.00-17.51, P<.001) and those who perceived a benefit in being able to retrieve information anywhere, anytime, and without fear of losing it had 2 times the likelihood of accepting e-labeling (AOR 2.66, 95% CI 1.11-6.35, P=.03). By contrast, Chinese ethnicity was associated with a 72% lower probability of accepting e-labeling compared with Malay ethnicity (AOR 0.28, 95% CI 0.13-0.60, P=.005). Participants who perceived limited skills in browsing electronic medicinal product information were 75% less likely to accept e-labeling (AOR 0.25, 95% CI 0.11-0.56, P<.001). The binary logistic regression model was statistically significant (χ^2_6=49.285, P<.001). The model explained 39.3% of the variance in e-labeling acceptance (Nagelkerke R^2). The Hosmer and Lemeshow test indicated that the model was a good fit for the data (P=.21, >0.05). Overall, the model had a good accuracy rate of 84% and exhibited excellent sensitivity (96.6%) in predicting e-labeling acceptance.

XSL•FO
RenderX

Figure 3. Participant's acceptance of electronic labeling.

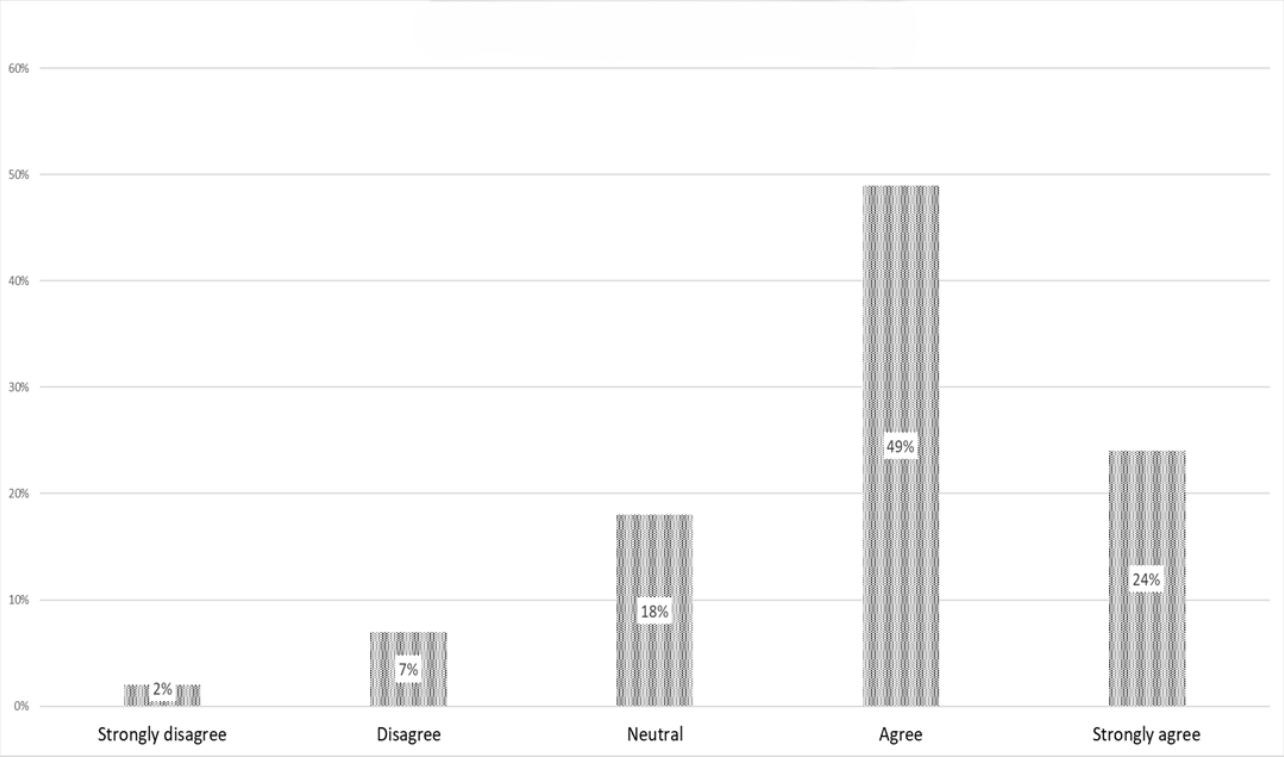

XSL•FO
RenderX

Table 3. Factors associated with participants' acceptance of e-labeling.

Independent variables	Crude odds ratio (95% CI)	*P* value	Adjusted odds ratio (95% CI)	*P* value
Demographics and characteristics of package insert use				
Age (years)				
18-39	Reference	Reference	N/A[a]	N/A
40-64	0.62 (0.34-1.15)	.13[b]	N/A	N/A
65 and above	0.29 (0.16-0.56)	<.001[b]	N/A	N/A
Gender				
Male	Reference	Reference	N/A	N/A
Female	0.66 (0.42-1.04)	.08[b]	N/A	N/A
Ethnicity				
Malay	Reference	Reference	Reference	Reference
Chinese	0.38 (0.21-0.68)	.001[b]	0.28 (0.13-0.60)	.001[c]
Indian and others	0.88 (0.43-1.80)	.73	1.61 (0.58-4.51)	.36
Highest level of education				
Secondary school and below	Reference	Reference	N/A	N/A
University/college	1.69 (1.03-2.78)	.04[b]	N/A	N/A
Obtain or seek written information about medicine				
No	Reference	Reference	N/A	N/A
Yes	1.39 (0.80-2.40)	.24[b]	N/A	N/A
Source of written information about medicine				
Other sources	Reference	Reference	N/A	N/A
Product inserts	0.71 (0.42-1.20)	.20[b]	N/A	N/A
Frequency of reading the medicinal product package insert				
Never	Reference	Reference	N/A	N/A
Always	0.36 (0.11-1.16)	.09[b]	N/A	N/A
Sometimes	1.01 (0.35-2.98)	.98	N/A	N/A
Only when I received a new medication	0.50 (0.18-1.41)	.19[b]	N/A	N/A
Perceived benefits				
Flexibility in information retrieval				
No	Reference	Reference	Reference	Reference
Yes	4.73 (2.70-8.26)	<.001[b]	2.66 (1.11-6.35)	.03[c]
Allows to understand medication better with images and videos				
No	Reference	Reference	Reference	Reference
Yes	8.88 (4.32-18.25)	<.001[b]	8.02 (2.80-22.97)	<.001[c]
Allows to understand medication better by choosing preferred language				
No	Reference	Reference	N/A	N/A
Yes	4.06 (1.98-8.33)	<.001[b]	N/A	N/A
Allows the use of advanced interactive features				
No	Reference	Reference	N/A	N/A
Yes	2.99 (1.52-5.86)	.001[b]	N/A	N/A

XSL•FO
RenderX

Independent variables	Crude odds ratio (95% CI)	*P* value	Adjusted odds ratio (95% CI)	*P* value
Allows to get the most updated medication information				
No	Reference	Reference	N/A	Reference
Yes	3.61 (2.04-6.41)	<.001[b]	N/A	N/A
Uses a paperless system to protect the environment				
No	Reference	Reference	Reference	Reference
Yes	8.08 (4.15-15.74)	<.001[b]	7.24 (3.00-17.51)	<.001[c]
Perceived challenges				
No electronic gadgets				
No	Reference	Reference	N/A	N/A
Yes	0.37 (0.16-0.87)	.02[b]	N/A	N/A
Limited skill in the use of electronic gadgets				
No	Reference	Reference	N/A	N/A
Yes	0.40 (0.20-0.80)	.009[b]	N/A	N/A
Limited internet access				
No	Reference	Reference	N/A	N/A
Yes	0.34 (0.15-0.77)	.01[b]	N/A	N/A
Limited skills in browsing information				
No	Reference	Reference	Reference	Reference
Yes	0.25 (0.15-0.44)	<.001[b]	0.25 (0.11-0.56)	<.001[c]
Concerned about obtaining potentially illegal information				
No	Reference	Reference	N/A	N/A
Yes	0.54 (0.34-0.85)	.008[b]	N/A	N/A

[a]N/A: not applicable.
[b]*P*<.25.
[c]*P*<.05.

Preference Toward e-Labeling

Most participants preferred accessing electronic medicinal product information through official or government websites (302/387, 78.0%). Participants also showed interest in scanning a digital code, such as a QR code printed on the outer medication package (282/387, 72.9%), or accessing information through digital patient services, such as medication apps (282/387, 72.9%), compared with receiving a link via SMS text message or email (194/387, 50.1%). However, 221/387 (57.1%) of participants still preferred the option to request a printed copy of the medicinal product information (Figure 4).

In response to the open-ended question about views on e-labeling implementation for medicinal product information, most participants (33/83, 40%) emphasized that the e-labeling platform should consistently provide updated medication information that is neutral and free from product advertisements. They also highlighted the importance of the platform being easily accessible, user-friendly, easy to understand, and compatible with various electronic devices. Some participants also expressed that the content of the electronic label (e-label) must be reliable and protected from third-party modifications or cybersecurity attacks to ensure it is safe for patient use (7/83, 8%). Suggestions included accessing e-labeling through hospital websites verified by competent authorities. Additionally, participants recommended features for the e-labeling platform, such as the ability to compare information across medications for the same indication, separate sections for medication information on different diseases to facilitate easy location, links to journal or research articles, a section for user feedback, and a notification function to alert patients about new updates (6/83, 7%). As e-labeling for medicinal product information is a new initiative, participants also felt that a helpline should be available for patients needing assistance (3/83, 4%).

XSL•FO
RenderX

Figure 4. Participant's preference toward electronic labeling.

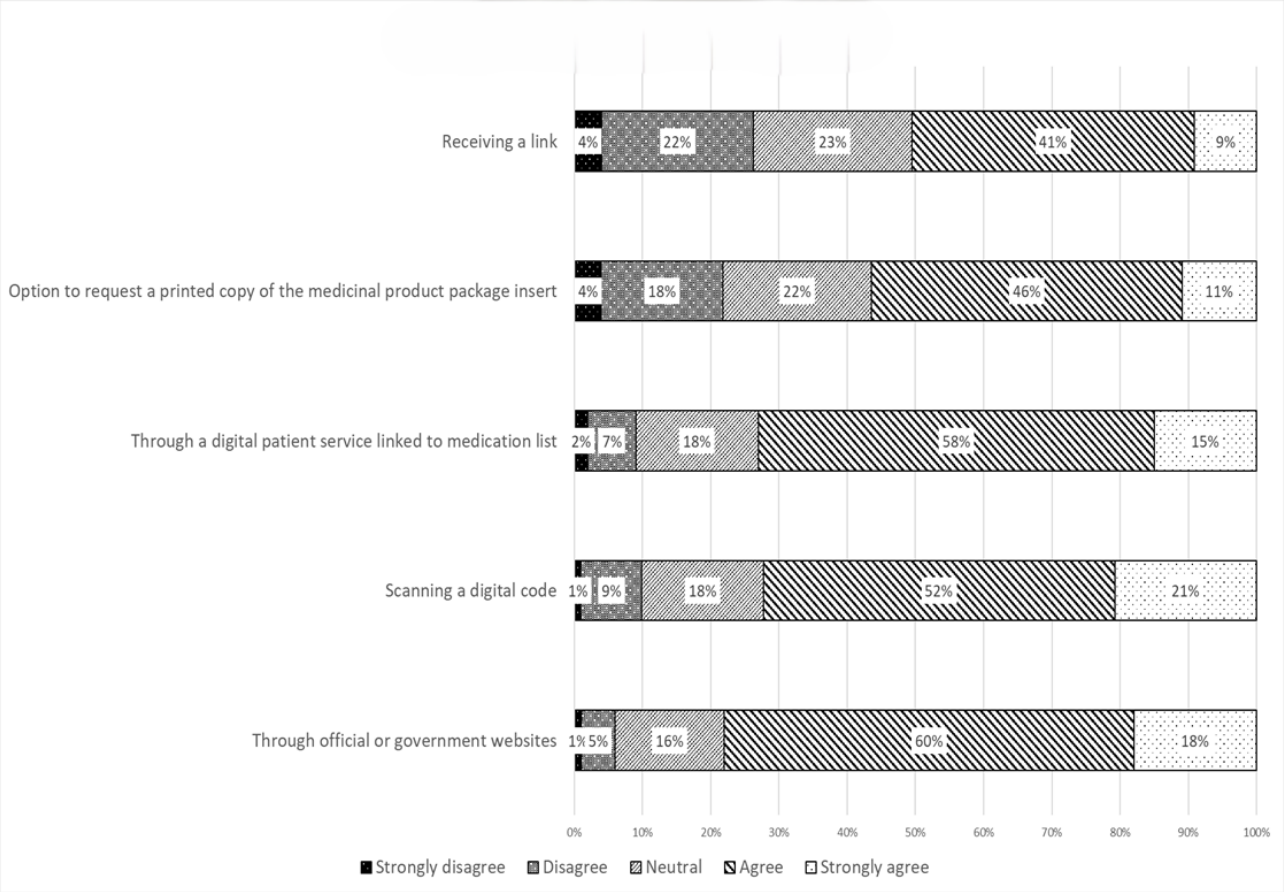

Despite the generally positive perception of e-labeling for medicinal product information, some participants expressed concerns about certain populations, including older adults, individuals with low digital literacy, those without internet access, and those without electronic devices (23/83, 28%). They suggested that it might be necessary to provide both paper and electronic inserts and recommended that authorities implement the e-labeling initiative in phases to allow the public time to adapt to the new platform (10/83, 12%).

Discussion

Principal Findings and Comparison With Prior Work

We found a moderately high acceptance rate (283/387, 73.1%) for e-labeling among hospital patients, with more than half (221/387, 57.1%) preferring to retain the option to request a printed copy. Most participants viewed the electronic version of the PI as beneficial, especially for understanding their medication better through language choices (352/387, 91.0%). However, around half of the participants (197/387, 50.9%) were concerned about the potential risk of accessing illegal medication information via e-labeling. Most participants preferred accessing e-labels from trusted sources such as government websites (302/387, 78.0%). Acceptance of e-labeling was significantly influenced by patients' perceptions of benefits, including a better understanding of medication, environmental protection, and flexibility in information retrieval. By contrast, patients of Chinese ethnicity and those who perceived limited skills in using electronic inserts were less likely to accept e-labeling.

Compared with older studies, the acceptance rate for e-labeling in our study was higher. For example, a study conducted in Sweden before the pandemic reported that only 41% of 406 customers surveyed in community and hospital pharmacies were interested in using electronic medicinal product information. Additionally, 54% of respondents indicated they would request a printed version from the pharmacy if the paper leaflet was not included in the package [11]. During the pandemic, a survey of 2518 vaccine recipients or their parents across 4 European countries (Belgium, Italy, Bulgaria, and France) reported an acceptance rate for electronic leaflets ranging from 55% to 82%, with an overall acceptability of 64% when a printed leaflet option was available [18].

Our patients' perception of e-labeling as enhancing their understanding of medication aligns with findings from a Saudi Arabian study, where patients reported that reading medicinal product information leaflets positively impacted their knowledge about medicines and medication adherence [9]. However, only 1.6% (5/312) of participants in our study who obtained written medicine information used RiMUP, in contrast to the 91.1% utility rate of patient information leaflets observed in the Saudi Arabian study. This discrepancy may be attributed to the fact that RiMUPs are not distributed with products but are instead available as PDFs on the Malaysian NPRA website. HCPs are responsible for retrieving, printing, and disseminating them to patients if needed. Similar to experiences in Australia, this

XSL•FO
RenderX

practice has not led to widespread dissemination of RiMUP as intended [27]. In our study, patients' perceptions of the convenience of accessing e-labeling anytime and anywhere, as well as the ease of information retrieval, align with their primary source of medical information—the internet. Studies have shown that the availability of the internet has increased the use of online sources for medication information [11,18,28].

Malaysian patients have shown support for transitioning from paper medicinal product PIs to e-labeling for several reasons. First, there is widespread awareness of the negative impact of paper consumption associated with printing paper inserts [29]. This awareness is likely influenced by frequent media reports on extreme weather events and the broader effects of deforestation on climate change, which have heightened public concern about environmental issues. Second, the public adopted new health behaviors during the pandemic, which required transitioning many occupational and social activities to online platforms as a preventive measure against COVID-19 transmission [30]. Malaysians adapted to paperless systems such as QR codes and mobile apps, which explains the high preference for digital code scanning and medication apps among patients [31]. As socioeconomic activities resumed in the postpandemic period, this practice has become the new norm. By contrast, receiving a link to electronic medication information was the least favored option among patients. This reluctance may be attributed to the rising incidence of scams in Malaysia in recent years, which has made patients wary of clicking on links [32]. Additionally, several nationwide digitalization programs, such as the paperless road tax and online passport renewal policies recently introduced by the Malaysian government, have increased public acceptance of digital services [33]. This aligns with the mission of Malaysia's National Fourth Industrial Revolution (4IR) policy, which aims to leverage digital technology to transform the economy in line with the Shared Prosperity Vision of creating a fair, equitable, and inclusive society by 2030 [34]. Additionally, the low perceived challenges related to digital gadget ownership, usage, and internet access may be due to ongoing income tax exemptions on laptops and the incentives promoting smartphone and laptop ownership. These measures have contributed to the public's high readiness to adopt the e-labeling platform [35].

Paper leaflets for medicinal product information have an unavoidable environmental footprint, and shifting to electronic versions can significantly reduce production costs [12,16,28,36]. Additionally, features such as zooming and search functions on electronic devices make it easier and faster for patients to locate information [11,28,36,37]. These functionalities address the limitations of paper inserts and enhance the overall patient experience in managing their medication. This shift to e-labeling could potentially encourage patients who were previously hesitant to use medicinal product information leaflets to view electronic formats as a reliable source of information. Additionally, participants in our study suggested that the e-labeling system should present information in a comparative format across different drugs with the same indications and link to credible sources such as journals or research articles. This indicates that Malaysian patients are eager to learn about their medications and take an active role in managing their treatment. Providing patient-centric medicinal information in local languages can enhance medication literacy. Ultimately, e-labeling has the potential to improve medication use and lead to better health outcomes.

In our study, patients considered the legitimacy of the e-labels as a crucial aspect of the e-labeling system. Most patients preferred accessing e-labels through trusted platforms, such as government or official websites. This preference is likely due to the prevalent cybersecurity issues in Malaysia [38]. Consequently, patients emphasized the importance of maintaining system security to mitigate the risk of biased information that could impact patient safety. Our findings suggest that patients' perceived limited skills in browsing e-labels correlate with lower acceptance of e-labeling, a phenomenon explained by Bandura's Theory of Self-Efficacy [39]. According to this theory, individuals who feel confident in their ability to use the e-labeling platform are more likely to engage with and accept the technology. Consequently, providing a helpline for patients could facilitate their adaptation to the e-labeling system. Despite the generally high acceptance of e-labeling, 221/387 (57.1%) participants preferred not to completely eliminate paper inserts, a preference consistent with previous studies [11,18]. Therefore, it is important to implement procedures that support patients with limited digital skills until the e-labeling platform is fully established and effective.

Our study found that acceptance of e-labeling was lower among Chinese patients. Currently, the RiMUP is not available in Chinese or Tamil, the 2 major languages in our region of Malaysia, which may have impacted Chinese patients' perceptions of e-labeling. Further research is needed to explore the underlying reasons for ethnic discrepancies in e-labeling acceptance among Malaysians. In our study, age was a significant factor in the univariate analysis but not in the multivariate analysis. This finding contrasts with Hammar et al [11], which suggested that older age might hinder the adoption of electronic patient information leaflets. The perceived lack of digital literacy skills among patients, which could be a more relevant factor, may have been reflected in our study, thereby minimizing age as a potential confounder. Similar results were reported in a recent European vaccine study that focused on individuals aged over 60 years [18].

Implications for Policies and Strategies

Future strategies to enhance patient uptake of e-labeling should address concerns about the challenges associated with using digital platforms for medical information. First, flexibility should be provided to allow patients to request a printed copy of medicinal product information leaflets when necessary. Second, public awareness campaigns could encourage individuals who have not previously utilized medicinal product information leaflets to start using e-labeling, thereby increasing overall engagement. Patients should be informed about the benefits of e-labeling, including personalized information, enhanced medication safety, improved supply chain efficiency, and environmental protection. Educational materials should be provided in common local languages to ensure accessibility for patients from diverse backgrounds. Third, to help patients adapt to the e-labeling platform, practical demonstration videos with

XSL•FO
RenderX

simple, clear instructions and visual aids can be displayed in pharmacy waiting areas to create a positive learning experience. Fourth, periodic reviews of the e-labeling system should be conducted to ensure it remains user-friendly and compatible with various electronic devices. Features such as linking e-labels to research studies and journals, enabling medication comparisons for the same indications, and including a section for public feedback can enhance the platform's patient-centric approach. Fifth, a robust process should be established to update information in a centralized database. The responsible authority must verify and ensure the accuracy of the data before they are made available to the public. Lastly, a helpline should be provided to offer patients assistance whenever needed.

Strength and Limitations

To the best of our knowledge, this is the first study assessing patient acceptance of e-labeling for medicinal product information during the postpandemic transition. The study also identified factors influencing patients' acceptance of e-labeling. A stratified sampling method was used to ensure the sample accurately represents the patient population's distribution in terms of age-related medication usage.

This study has several limitations. Conducted in a single hospital pharmacy located in Kuala Lumpur, a highly urbanized and densely populated city in Malaysia, the findings may not be generalizable to suburban or rural populations elsewhere in the country. Patients in rural areas may have limited internet access, which could influence their acceptance of e-labeling. Currently, the Malaysian government is working with the industry to enhance internet connectivity as part of the Madani Economy framework, aiming to provide stable and affordable internet access to Malaysians across all regions [40]. The proportion of participants with higher education levels in this study was greater than that observed in the national population census, which may introduce bias, as higher education is often linked with better economic status, ownership of digital devices, and proactive health information–seeking behavior [41]. However, the results from the multivariate analysis indicated that education level was not a significant predictor of e-labeling acceptance.

Conclusions

Malaysian hospital patients demonstrated a moderately high level of acceptance of e-labeling of medicinal product information. Key factors predicting high acceptance included perceived benefits such as improved understanding of medication, environmental protection, and flexibility in information retrieval. By contrast, lower acceptance was associated with being of Chinese ethnicity and having perceived limitations in digital self-efficacy. Future strategies to enhance e-labeling uptake should focus on addressing patients' concerns about digital platform challenges and emphasizing the advantages of e-labeling.

Availability of Data and Materials

The data sets used and analyzed in this study are available from the corresponding author on reasonable request.

Acknowledgments

The authors express their gratitude to all UMMC patients who have participated in the study and appreciate Ms Wan Lee Chow, Dr Jason Loo Siau Ee, and Mr Selvarajah Seeragam who were involved in the validation of the questionnaire. We also thank AstraZeneca Sdn Bhd for funding this study in accordance with the Good Publication Practice Guideline 2022. We acknowledge AstraZeneca Sdn Bhd and Taylor's University for open-access funding of this publication.

Authors' Contributions

BKT, XYL, ALW, AH, and CSP conceived and designed the study. XYL collected the data. XYL and BKT analyzed the data. XYL drafted the manuscript. BKT, ALW, AH, and CSP critically revised the manuscript for important intellectual content and approved the final version for publication. AstraZeneca did not have any role in decision-making or in the preparation of consensus guidelines, if any.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire: Acceptance of electronic labeling among hospital ambulatory patients (English).
[PDF File (Adobe PDF File), 97 KB - jmir_v26i1e56591_app1.pdf]

Multimedia Appendix 2
Questionnaire: Acceptance of electronic labeling among hospital ambulatory patients (Malay).
[PDF File (Adobe PDF File), 98 KB - jmir_v26i1e56591_app2.pdf]

Multimedia Appendix 3
Questionnaire: Acceptance of electronic labeling among hospital ambulatory patients (Mandarin).
[PDF File (Adobe PDF File), 203 KB - jmir_v26i1e56591_app3.pdf]

XSL•FO
RenderX

References

1. Pouliot A, Vaillancourt R, Stacey D, Suter P. Defining and identifying concepts of medication literacy: an international perspective. Res Social Adm Pharm 2018 Sep;14(9):797-804. [doi: 10.1016/j.sapharm.2017.11.005] [Medline: 29191647]
2. Pourhabibi N, Mohebbi B, Sadeghi R, Shakibazadeh E, Sanjari M, Tol A, et al. Determinants of poor treatment adherence among patients with type 2 diabetes and limited health literacy: a scoping review. J Diabetes Res 2022;2022:2980250 [FREE Full text] [doi: 10.1155/2022/2980250] [Medline: 35832786]
3. Caruso R, Magon A, Baroni I, Dellafiore F, Arrigoni C, Pittella F, et al. Health literacy in type 2 diabetes patients: a systematic review of systematic reviews. Acta Diabetol 2018 Jan;55(1):1-12. [doi: 10.1007/s00592-017-1071-1] [Medline: 29129000]
4. Ganapathy S, Aris H, Ahmad N. National Health and Morbidity Survey. Shah Alam, Selangor, Malaysia: National Institute of Health; 2019.
5. Bekker CL, Mohsenian Naghani S, Natsch S, Wartenberg NS, van den Bemt BJF. Information needs and patient perceptions of the quality of medication information available in hospitals: a mixed method study. Int J Clin Pharm 2020 Dec;42(6):1396-1404. [doi: 10.1007/s11096-020-01125-x] [Medline: 32857256]
6. Young A, Tordoff J, Leitch S, Smith A. Patient-focused medicines information: general practitioners' and pharmacists' views on websites and leaflets. Health Education Journal 2018 Nov 14;78(3):340-351. [doi: 10.1177/0017896918811373]
7. National Pharmaceutical Regulatory Agency (NPRA). Drug Registration Guidance Document (DRGD). NPRA Official Portal. 2023. URL: https://www.npra.gov.my/easyarticles/images/users/1047/DRGD%20Jan%202023/Complete-Drug-Registration-Guidance-Document-DRGD-3rd-Edition-4th-Revision-January-2023.pdf [accessed 2024-09-04]
8. Wongtaweepkij K, Corlett S, Krska J, Pongwecharak J, Jarernsiripornkul N. Patients' experiences and perspectives of receiving written medicine information about medicines: a qualitative study. Patient Prefer Adherence 2021;15:569-580 [FREE Full text] [doi: 10.2147/PPA.S298563] [Medline: 33727802]
9. Al Jeraisy M, Alshammari H, Albassam M, Al Aamer K, Abolfotouh MA. Utility of patient information leaflet and perceived impact of its use on medication adherence. BMC Public Health 2023 Mar 14;23(1):488 [FREE Full text] [doi: 10.1186/s12889-023-15346-y] [Medline: 36918823]
10. European Association of Hospital Pharmacists (EAHP). EAHP survey on the future potential of electronic product information (ePI). EAHP. 2022 Feb 22. URL: https://www.eahp.eu/practice-and-policy/ehealth-and-mhealth/ePIsurvey [accessed 2024-09-04]
11. Hammar T, Nilsson A, Hovstadius B. Patients' views on electronic patient information leaflets. Pharm Pract (Granada) 2016;14(2):702 [FREE Full text] [doi: 10.18549/PharmPract.2016.02.702] [Medline: 27382423]
12. Sirkas K, Juppo A, Miettinen M, Siven M. Could paper package leaflet be left out from hospital products? Explor Res Clin Soc Pharm 2022 Sep;7:100176. [doi: 10.1016/j.rcsop.2022.100176] [Medline: 36106157]
13. Use of electronic labeling. IFPMA. 2022 Feb 7. URL: https://www.ifpma.org/publications/position-paper-use-of-electronic-labeling/ [accessed 2024-09-04]
14. United Nations. Sustainable development goals. UNDP. URL: https://www.undp.org/sustainable-development-goals [accessed 2024-09-04]
15. Chaudhary P, Shetty V. E-labeling: change is underway. PharmExec. URL: https://www.pharmexec.com/view/e-labeling-change-underway [accessed 2024-04-09]
16. Matsui R, Yamaguchi K, Lee JJV, Ting I, Khairilisani D, Chang J, et al. Survey result for E-labeling initiatives in Asia. Ther Innov Regul Sci 2023 Mar;57(2):251-260 [FREE Full text] [doi: 10.1007/s43441-022-00462-5] [Medline: 36215028]
17. National Pharmaceutical Regulatory Agency (NPRA). Guideline on Electronic Labelling (E-labeling) for Pharmaceutical Products in Malaysia. NPRA. 2023 Apr. URL: https://npra.gov.my/easyarticles/images/users/1047/Lampiran-A-Guideline-on-E-labeling-for-Pharmaceutical-Products-in-Malaysia.pdf [accessed 2024-09-04]
18. Bamberger M, De Loof H, Marstboom C, Oury S, Bonanni P, Launay O, et al. Replacing vaccine paper package inserts: a multi-country questionnaire study on the acceptability of an electronic replacement in different target groups. BMC Public Health 2022 Jan 24;22(1):156 [FREE Full text] [doi: 10.1186/s12889-022-12510-8] [Medline: 35073891]
19. Ministry of Health Malaysia. Health facts 2021. Ministry of Health Malaysia. 2021 Nov. URL: https://www.moh.gov.my/index.php/pages/view/58?mid=19 [accessed 2024-09-04]
20. Sample size calculator by Raosoft, Inc. Raosoft. URL: http://www.raosoft.com/samplesize.html [accessed 2024-09-04]
21. Martin CB, Hales CM, Gu Q, Ogden CL. Prescription drug use in the United States, 2015-2016 (NCHS data brief, no 334). CDC. Hyattsville, MD: National Center for Health Statistics; 2019. URL: https://www.cdc.gov/nchs/products/databriefs/db334.htm [accessed 2024-09-04]
22. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
23. Adell E, Várhelyi A. Driver Acceptance of New Technology. Theory, Measurement and Optimisation. London, UK: Taylor and Francis Group; Jun 12, 2014.
24. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med 2008 Dec 16;3:17 [FREE Full text] [doi: 10.1186/1751-0473-3-17] [Medline: 19087314]

XSL•FO
RenderX

25. Zhang Y, Wildemuth BM. Qualitative Analysis of Content. In: Wildemuth BM, editor. Applications of Social Research Methods to Questions in Information and Library Science. London, UK: Libraries Unlimited; 2017:318-329.
26. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
27. Hamrosi KK, Raynor DK, Aslani P. Pharmacist, general practitioner and consumer use of written medicine information in Australia: are they on the same page? Res Social Adm Pharm 2014;10(4):656-668. [doi: 10.1016/j.sapharm.2013.10.002] [Medline: 24239213]
28. Bechini A, Chiesi F, Giammarco B, Gori E, Di Tommaso M, Strambi N, et al. Electronic package leaflets for vaccines: what are people's perceptions in Italy? Vaccines (Basel) 2022 Jul 04;10(7):1075 [FREE Full text] [doi: 10.3390/vaccines10071075] [Medline: 35891238]
29. Mei NS, Wai CW, Ahamad R. Environmental Awareness and Behaviour Index for Malaysia. Procedia - Social and Behavioral Sciences 2016 Jun;222:668-675. [doi: 10.1016/j.sbspro.2016.05.223]
30. Yadegaridehkordi E, Nilashi M, Nizam Bin Md Nasir MH, Safie Bin Mohd Satar N, Momtazi S. Evaluating COVID-19 infection prevention measures in Malaysia: a fuzzy DEMATEL approach. Digit Health 2023;9:20552076231211670 [FREE Full text] [doi: 10.1177/20552076231211670] [Medline: 38074341]
31. Ang ZY, Cheah KY, Shakirah MS, Fun WH, Anis-Syakira J, Kong Y, et al. Malaysia's health systems response to COVID-19. Int J Environ Res Public Health 2021 Oct 22;18(21):11109 [FREE Full text] [doi: 10.3390/ijerph182111109] [Medline: 34769629]
32. Wilson S, Hassan N, Khor K, Sinnappan S, Abu Bakar A, Tan S. A holistic qualitative exploration on the perception of scams, scam techniques and effectiveness of anti-scam campaigns in Malaysia. JFC 2023 Sep 27:0151 (forthcoming) [FREE Full text] [doi: 10.1108/JFC-06-2023-0151]
33. Khalid NA, Yang Y. Advancing democracy: the digital Government of Malaysia. ScienceOpen Posters. 2024 Apr 8. URL: https://www.scienceopen.com/hosted-document?doi=10.14293/P2199-8442.1.SOP-.PWXQD9.v1 [accessed 2024-09-04]
34. Economic Planning Unit, Prime Minister's Department. Putrajaya, Malaysia. National Fourth Industrial Revolution (4IR) Policy. Economic Planning Unit, Prime Minister's Department. URL: https://www.ekonomi.gov.my/sites/default/files/2021-07/National-4IR-Policy.pdf [accessed 2024-09-04]
35. Edrak B, Nor ZM, Shaik AR. The readiness of Malaysia digital economy: a study of three government policies from 1991 to 2020. IJEF 2022 Nov 30;14(12):84. [doi: 10.5539/ijef.v14n12p84]
36. Bolislis WRR, Mortazavi C, Riccioni R, Schaeffer P, Kühler TC. From print to screen: regulatory considerations to adopting innovative approaches for patient information and safety. Ther Innov Regul Sci 2020 Jul;54(4):831-838 [FREE Full text] [doi: 10.1007/s43441-019-00018-0] [Medline: 32557303]
37. Fung EWT, Au-Yeung GTF, Tsoi LM, Qu L, Cheng TKW, Chong DW, et al. Pharmacists' perceptions of the benefits and challenges of electronic product information system implementation in Hong Kong: mixed-method study. J Med Internet Res 2020 Nov 10;22(11):e20765 [FREE Full text] [doi: 10.2196/20765] [Medline: 33170130]
38. Ministry of Communications and Digital Malaysia. Cybersecurity Malaysia. Ministry of Communications and Digital Malaysia. URL: https://www.cybersecurity.my/en/media_centre/media_releases/2023/main/detail/2782/index.html [accessed 2024-09-04]
39. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [doi: 10.1037//0033-295x.84.2.191] [Medline: 847061]
40. Madani economy to boost Malaysian economy, improve quality of life – PM Anwar. Kementerian Kewangan Malaysia. 2023 Jul 27. URL: https://www.mof.gov.my/portal/en/news/press-citations/madani-economy-to-boost-malaysian-economy-improve-quality-of-life-pm-anwar [accessed 2024-09-04]
41. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]

Abbreviations

4IR: Fourth Industrial Revolution
AOR: adjusted odds ratio
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
E-labeling: electronic labeling
E-labels: electronic labels
E-prescribing: electronic prescribing
FDA: Food and Drug Administration
HCP: health care professional
NPRA: National Pharmaceutical Regulatory Agency
PI: package insert
PMDA: Pharmaceuticals and Medical Devices Agency
RiMUP: Consumer Medication Information Leaflets

XSL•FO
RenderX

UMMC: University Malaya Medical Centre
UNSDG: United Nations Sustainable Development Goals
UTAUT2: Unified Theory of Acceptance and Use of Technology, second version

Edited by T de Azevedo Cardoso; submitted 21.01.24; peer-reviewed by N Li, BF Ababneh; comments to author 04.06.24; revised version received 29.06.24; accepted 30.07.24; published 18.09.24.

Please cite as:
Loh XY, Woo AL, Haris A, Pereira CS, Tan BK
Acceptance of Electronic Labeling for Medicinal Product Information Among Malaysian Hospital Patients: Cross-Sectional Study
J Med Internet Res 2024;26:e56591
URL: https://www.jmir.org/2024/1/e56591
doi:10.2196/56591
PMID:

©Xin Yee Loh, Ai Ling Woo, Azwa Haris, Cheryl Shajini Pereira, Bee Kim Tan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Construction and Application of a Private 5G Standalone Medical Network in a Smart Health Environment: Exploratory Practice From China

Baozhan Chen[1,2], MSc; Xiaobing Shi[1,2], MSc; Tianyi Feng[3], MSc; Shuai Jiang[1,4], PhD; Yunkai Zhai[5], PhD; Mingxing Ren[1], MSc; Dongqing Liu[1,2], PhD; Chengzeng Wang[1,4,6*], PhD; Jinghong Gao[1,4*], PhD

[1]The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[2]National Engineering Laboratory for Internet Medical Systems and Applications, Zhengzhou, China
[3]China Academy of Information and Communications Technology, Beijing, China
[4]Institute for Hospital Management of Henan Province, Zhengzhou, China
[5]School of Management Engineering, Zhengzhou University, Zhengzhou, China
[6]Henan Institute of Interconnected Intelligent Health Management, Zhengzhou, China
[*]these authors contributed equally

Corresponding Author:
Jinghong Gao, PhD
The First Affiliated Hospital of Zhengzhou University
Number 1, Longhu Middle Ring Road
Jinshui District
Zhengzhou, 450000
China
Phone: 86 371 66964536
Email: fccgaojh@zzu.edu.cn

Abstract

Background: To date, the differentiated requirements for network performance in various health care service scenarios—within, outside, and between hospitals—remain a key challenge that restricts the development and implementation of digital medical services.

Objective: This study aims to construct and implement a private 5G (the 5th generation mobile communication technology) standalone (SA) medical network in a smart health environment to meet the diverse needs of various medical services.

Methods: Based on an analysis of network differentiation requirements in medical applications, the system architecture and functional positioning of the proposed private 5G SA medical network are designed and implemented. The system architecture includes the development of exclusive and preferential channels for medical use, as well as an ordinary user channel. A 3-layer network function architecture is designed, encompassing resource, control, and intelligent operation layers to facilitate management arrangements and provide network open services. Core technologies, including edge cloud collaboration; service awareness; and slicing of access, bearer, and core networks, are employed in the construction and application of the 5G SA network.

Results: The construction of the private 5G SA medical network primarily involves system architecture, standards, and security measures. The system, featuring exclusive, preferential, and common channels, supports a variety of medical applications. Relevant standards are adhered to in order to ensure the interaction and sharing of medical service information. Security is achieved through mechanisms such as authentication, abnormal behavior analysis, and dynamic access control. Three typical medical applications that rely on the 5G network in intrahospital, interhospital, and out-of-hospital scenarios—namely, mobile ward rounds, remote first aid, and remote ultrasound—were conducted. Testing of the 5G-enabled mobile ward rounds showed an average download rate of 790 Mbps and an average upload rate of 91 Mbps. Compared with 4G, the 5G network more effectively meets the diverse requirements of various business applications in prehospital emergency scenarios. For remote ultrasound, the average downlink rate of the 5G network is 4.82 Mbps, and the average uplink rate is 2 Mbps, with an average fluctuation of approximately 8 ms. The bandwidth, performance, and delay of the 5G SA network were also examined and confirmed to be effective.

XSL•FO
RenderX

Conclusions: The proposed 5G SA medical network demonstrates strong performance in typical medical applications. Its construction and application could lead to the development of new medical service models and provide valuable references for the further advancement and implementation of 5G networks in other industries, both in China and globally.

(J Med Internet Res 2024;26:e52404) doi:10.2196/52404

KEYWORDS

5G; medical private network; construction; application; performance test

Introduction

Medical institutions are densely populated, with unpredictable business volumes, fluctuating personnel flow, complex demands, and diverse medical service scenarios, all of which pose significant challenges to the network capacity required for their operation and maintenance. In light of the increasingly advanced and rapidly evolving medical services, traditional networking technology is encountering bottlenecks in areas such as the digital transformation of indoor networks, service distribution in transmission networks, intelligent network operations, and end-to-end safety guarantees [1,2]. Particularly during a major public health emergency like the COVID-19 pandemic, when hospitals exceed their normal capacity, health care professionals grappling with this unprecedented crisis face the challenge of effectively coordinating medical resources across widely dispersed locations [3]. Not only have patients been marginalized, but many clinicians have also had limited access to the treatment consultations and guidelines they need from hospitals during the COVID-19 outbreak [4]. As the pandemic continues, relying solely on conventional networking technologies (eg, 3G and 4G) may lead to significant costs and health risks. Studies have emphasized the need to design and implement an indoor network coverage system for hospitals that can meet the multifaceted business requirements and growing traffic demands [3,5]. This system should include features such as capacity redundancy, on-demand configurability, and scalability to ensure adaptability to dynamic changes.

Compared with 4G (the 4th generation mobile communication technology), the emerging 5G (the 5th generation mobile communication technology) mobile communication technology offers higher speeds, lower latency, enhanced capacity, and increased bandwidth. These improvements enable 5G to connect a large number of devices and support massive data processing with greater reliability, making it a game changer in health care services [6,7]. The architecture of the 5G network is divided into non-standalone (NSA) and standalone (SA) configurations [8]. NSA serves as a transitional framework from 4G to 5G, providing 5G mobile bandwidth services by utilizing the existing 4G infrastructure for 5G deployment. It upgrades the current 4G evolved packet core (EPC) network and uses dual connectivity technologies to access 5G New Radio. NSA has become the initial rapid commercial solution for some operators. However, the existing 4G EPC architecture and 4G air interface cannot meet 5G's requirements for low latency and transmission reliability [9]. SA represents the latest advancements in 5G networks, incorporating state-of-the-art base stations, backhaul infrastructure, and an upgraded EPC. In addition, SA introduces new network elements and interfaces, while adopting advanced technologies such as the Internet of Things (IoT), cloud computing, network virtualization, software-defined networks, and big data [9,10]. Compared with the NSA network architecture, the SA network architecture offers more advantages.

In recent years, advancements in 5G technology have driven the rapid development of health care services. The integration of IoT, big data, and cloud computing has enabled greater interconnectivity among patients, health care professionals, medical devices, and institutions. This has led to the emergence of new eHealth applications, such as mobile health, telemedicine, and internet-based health care [6,11]. These advancements encompass various aspects of disease management, including diagnosis, treatment, rehabilitation, prevention, and health monitoring, effectively enhancing the accessibility and capacity for delivering high-quality medical services [6,12]. However, diverse network performance requirements across different health care service scenarios—both within and between hospitals—continue to pose significant challenges that hinder the advancement of digital and intelligent medical services.

To effectively address the aforementioned challenges and meet the growing demand for network differentiation in medical services, the development of a new 5G-based network is urgently needed. Therefore, this study aims to design and implement a private 5G SA medical network within a smart health environment to accommodate diverse medical service scenarios. First, based on a requirements analysis, this study introduces the overall architecture, functional design, core technologies, and security mechanisms involved in the design and construction of the 5G SA network. Second, the construction and functionality of the 5G network are implemented to achieve the objectives outlined in the requirements analysis. Finally, the performance of the proposed 5G SA network is evaluated through an investigation of 3 representative applications. Lastly, the findings of this study undergo rigorous discussion, leading to the derived conclusions. To the best of our knowledge, this is the first comprehensive account of constructing and applying a private 5G SA medical network within a smart health environment in China. The findings may not only contribute to the further development and implementation of 5G medical networks in China but also provide valuable insights for other regions worldwide.

XSL•FO
RenderX

Methods

Requirements Analysis

The 5G network equipment generates a significant volume of diverse data across multiple layers, including the physical, virtual, link, network, and user layers. The current centralized data processing model struggles to keep pace with the low latency and high bandwidth demands of computing loads. Additionally, the traditional network operation and maintenance approach involves numerous manual tasks, which have proven inadequate for meeting the requirements of comprehensive cloud and intelligent network deployment for 5G.

Given the demand for network differentiation in medical applications, the following challenges and requirements must be addressed to construct and implement a private 5G SA medical network within a smart health environment. First, the design of the 5G network should comprehensively consider network coverage, capacity, and cost, taking into account the complexity of medical applications and the variations in transmission loss across different frequency bands and spectrum bandwidths. Second, when building a 5G network, health care facilities should carefully consider the diverse types of medical network application services, as well as the associated network requirements and standards. Third, health care facilities face significant challenges in effectively managing the privacy of information derived from medical devices, health care information systems, and medical networks [13]. Lastly, the concurrent development of a visualization platform for the intelligent operation and maintenance of the 5G network is typically required.

System Design

Overview

The private 5G medical network focuses on 3 major service scenarios: medical monitoring and nursing applications that utilize wireless data collection from hospital medical equipment; medical diagnosis and guidance applications based on video and medical data interaction; and remote-control applications for intelligent medical equipment [5,14]. Based on the specific needs of medical institutions—such as intrahospital mobile medical services, interhospital telemedicine, and out-of-hospital internet medical services—network slicing techniques, mobile edge computing, big data, and artificial intelligence (AI) are used to design and deploy the 5G SA medical industry–specific network. The 5G SA network can meet the diverse security requirements of users in 5G medical services by enabling functions such as network security isolation between the medical private network and public users, flexible deployment of network slices, and dynamic network awareness.

Following the service-based architecture of the 5G Core (5GC) network as defined by the 3rd Generation Partnership Project (3GPP) standard and leveraging cloud-native technology, the 5G medical network divides the application scenarios of medical services into network slices. It designs their functions and deploys them at different levels based on the service-based microservice architecture [15]. The private network is constructed by defining the combination of network functional components to cater to diverse medical application scenarios [16]. This approach offers enhanced ubiquitous access, increased flexibility in control and forwarding, and greater network openness for medical institutions [15,17]. Considering the various scenarios of intrahospital, interhospital, and out-of-hospital medical services, the 5G network can be categorized into medical-exclusive and preferential channels, as well as ordinary user channels (Figure 1).

Figure 1. Overall architecture of the proposed private 5G SA medical network. 5G: 5th generation mobile communication technology; CT: computed tomography; HIS: hospital information system; LIS: laboratory information system; MEC: mobile edge computing; MRI: magnetic resonance imaging; PACS: picture archiving and communication system; PC: personal computer; SA: standalone.

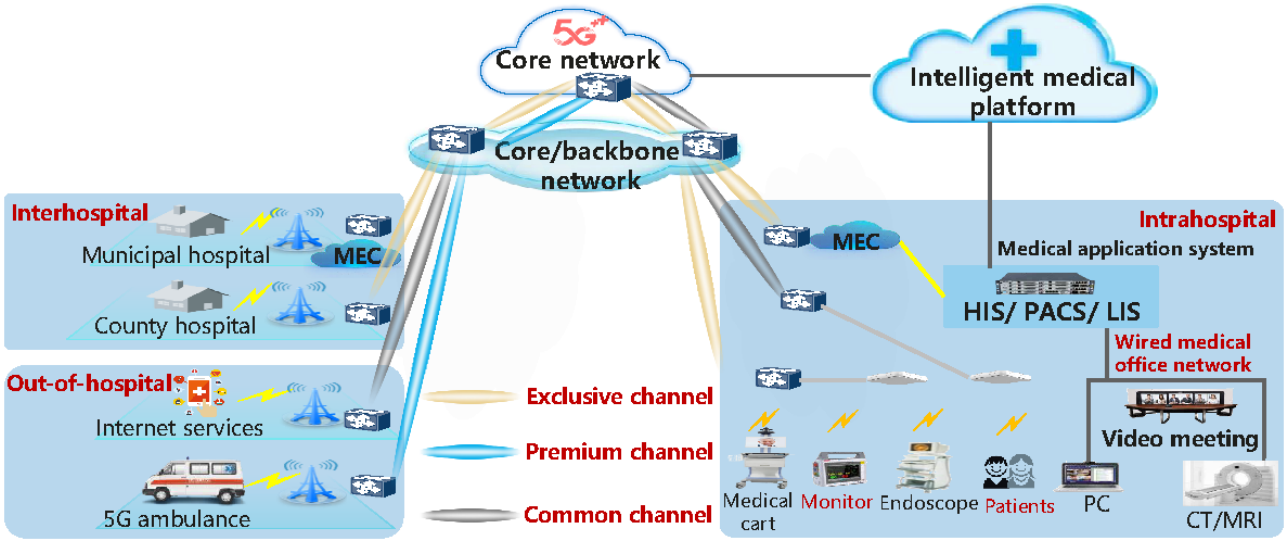

XSL•FO
RenderX

Exclusive Channel

The Flexible Ethernet (FlexE) slicing technology is used to partition dedicated network resources for transmitting fixed services within the hospital and collaborative services between hospitals, such as mobile ward rounds, appointment registration, and teleconsultation. This approach ensures guaranteed capabilities regarding delay, speed, reliability, and connection capacity [18]. Additionally, historical data on concurrent business and traffic statistics are combined to plan network resources, which in turn optimizes the allocation of these resources.

Premium Channel

It dynamically creates slices to support temporary medical services based on business needs, such as remote first aid, online consultation, and chronic disease management. This approach ensures a seamless experience for temporary medical services and facilitates the support of multiple nested subslices within a main slice for prioritized allocation and meticulous management. Additionally, it strictly isolates traffic between subslices to prevent any interference. It will not disrupt the operational sequence of other services due to sudden traffic. Furthermore, the subslice can be deleted upon completion of the business process, freeing up bandwidth resources, reducing network bandwidth consumption, and achieving cost savings.

Common Channel

It primarily supports public user services, allowing patients to access the internet and obtain public services through their mobile phones, tablets, and other mobile devices. These public user services are segregated from both the exclusive and preferential channels to meet the internet needs of patients within the hospital while facilitating isolated operations for medical services across other channels.

Function Design

Network Architecture

The proposed private 5G SA network, centered around intelligent operation, is capable of providing resource support for network access and data forwarding functions [19]. It offers an upward management arrangement and open network services, establishing a 3-layer network function architecture (Figure 2).

Figure 2. The functional architecture of the proposed private 5G SA medical network. 5G: 5th generation mobile communication technology; SA: standalone.

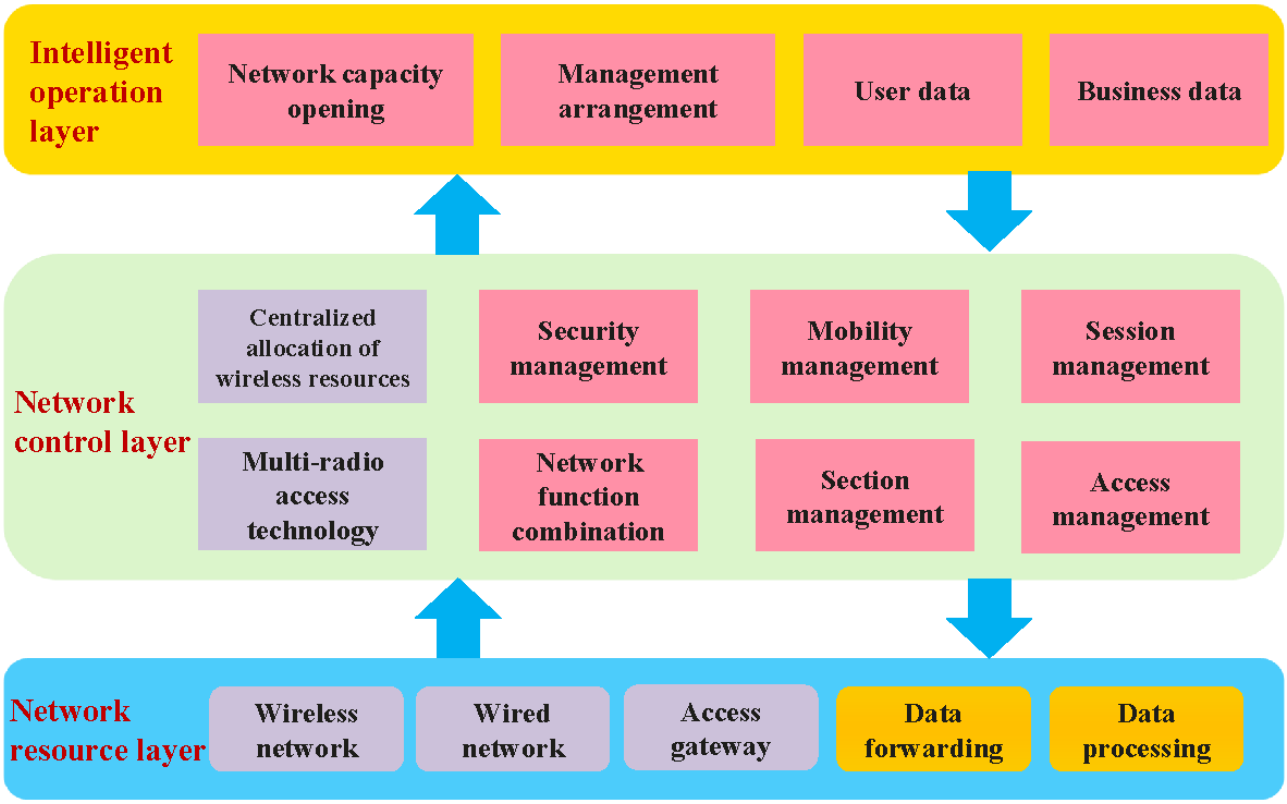

Network Resource Layer

This layer includes functions on both the access and network sides. The access side comprises a central unit and a distribution unit. The central unit primarily provides service aggregation functions, while the distribution unit offers data access points for terminals, including radiofrequency capabilities and some signal processing functions. The network side focuses on functions such as data forwarding, traffic optimization, and content services. Utilizing distributed anchor points and flexible forwarding path settings, data packets are directed to the appropriate processing nodes to achieve efficient forwarding and comprehensive data processing, including deep packet inspection, content billing, and traffic compression.

Network Control Layer

This layer allows the 5G SA medical network to restructure and modularize its control functions. The primary function modules include centralized allocation of wireless resources; unified management and control of multiple access points; traffic

XSL•FO
RenderX

grooming; and management of mobility, sessions, and security. According to the instructions provided by the 5G network function view of the management arrangement layer, the aforementioned functional components are integrated within the network control layer, allowing for flexible scheduling of resources.

Intelligent Operation Layer

It consists of 4 functional components: business data, user data, management arrangement, and network capacity opening. The business data function records information related to business processes and objects, and traffic statistics. The user data function stores information concerning user subscriptions, business policies, and network status. The management arrangement function utilizes network function virtualization technology to enable dynamic orchestration of network functions and the on-demand creation of network slices. The capacity opening function facilitates the centralized collection and encapsulation of network information, while also providing third parties access through application programming interfaces (APIs).

Network Slicing Technologies

Advanced Technologies in 5G SA Medical Networks

To meet the needs of various medical institutions, doctors, patients, and other stakeholders involved in medical services, advanced technologies (Table 1) such as network slice integration, edge cloud collaboration, and business awareness are used to construct the 5G SA medical network. This approach facilitates innovative applications such as on-demand deployment, high-speed connections, intelligent computing, and data security (Figure 3).

Table 1. Advanced technologies used to construct the private 5G[a] SA[b] medical network.

Technologies	Description
Network slice	The key feature of network function virtualization applied in 5G, which involves slice selection and management. Slice selection establishes access mapping between the user terminal and the network slice. Slice management also provides a safe isolation and highly automatic dedicated logical network for different medical services, hospital departments, and doctors.
Edge cloud collaboration technology	A technology used to achieve business linkage and resource sharing between edge computing and medical cloud systems; also used to realize the collaboration between resources, data, and applications.
Service-aware technology	A technology used to achieve the security, visualization, control, and management of the medical network and services.

[a]5G: 5th generation mobile communication technology.

[b]SA: standalone.

Figure 3. Technical architecture of the 5G SA medical private network. 5G: 5th generation mobile communication technology; AMF: access and mobility management function; NF: network element function; QCI: QoS class identifier; QoS: quality of service; RB: resource block; SMF: session management function; UPF: user plane function.

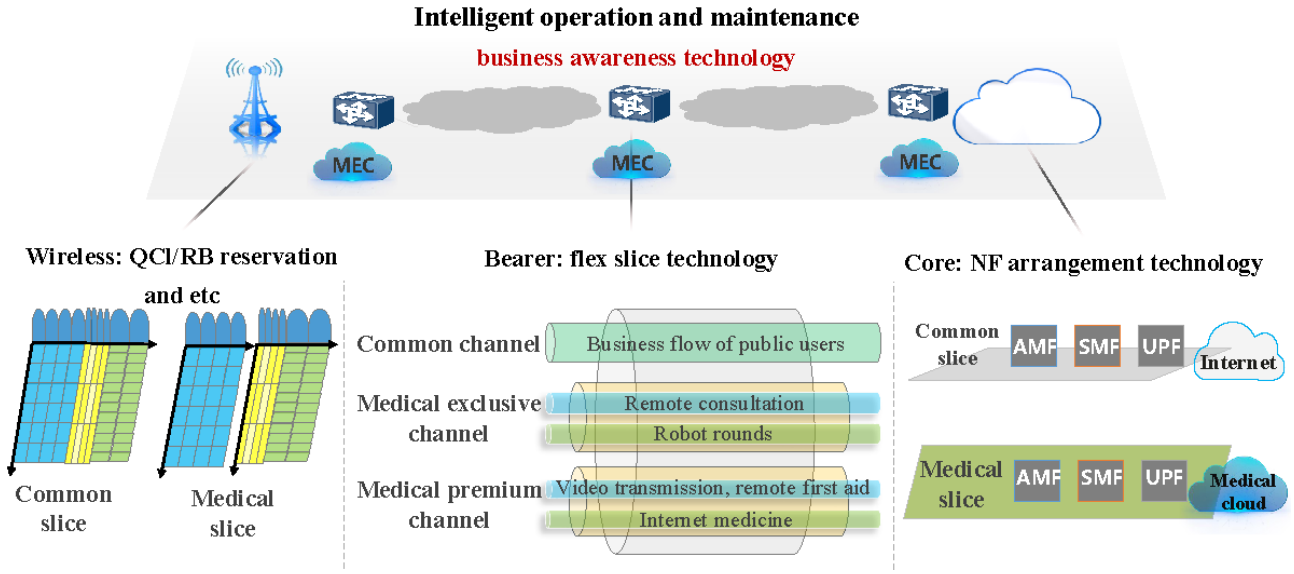

Access Network Slicing Technology

The 5G base station access network, as a crucial component of the medical system, can utilize 3 specialized network slicing technologies to meet layering and grading requirements: quality of service (QoS) technology, resource block (RB) reservation technology, and independent private network technology (Figure 4).

XSL•FO
RenderX

Figure 4. Three special network slicing techniques for the 5G base station access. 5G: 5th generation mobile communication technology; QoS: quality of service; RB: resource block.

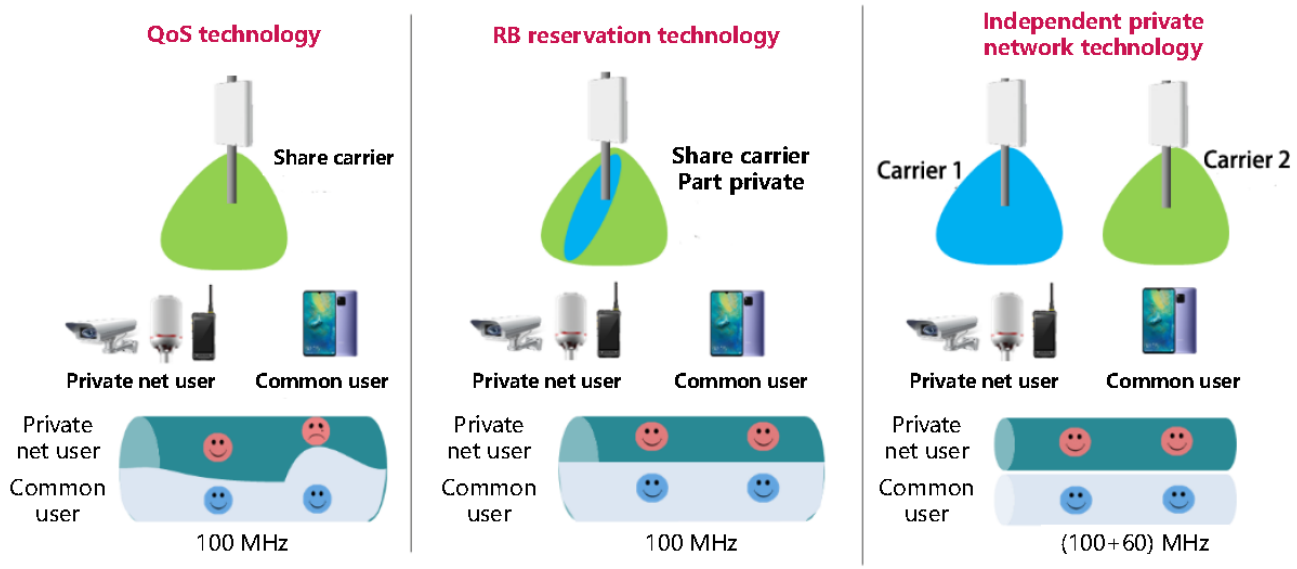

QoS technology primarily allows mobile phone users, as well as small medical institutions and clinics, to access network services anytime and anywhere. For ordinary user terminal business, the access point name uses an ordinary priority default bearer, and the QoS class identifier (QCI) is set to 9, with the SIM card account selecting the QCI. For the dedicated mobile terminals of doctors, an access point name is used to establish a special bearer on demand, with the QCI set to 3. QoS is configured through control program facilities or policy and charging rule functions based on the terminal access controller. This QoS technology prioritizes services for doctor-specific mobile terminals, thereby enhancing the end-user experience when accessing the medical private network.

RB reservation technology is primarily designed to provide network services with business security for small and medium-sized medical institutions. RBs serve as the scheduling resources on the 5G access network side, and all services must request RB resources to be forwarded. This technique involves reserving RB resources and binding slice IDs based on medical service types and industry users, ensuring that only medical services with matching slice IDs can access the reserved RB resources. This approach enhances the security of medical service data and improves QoS guarantee capabilities.

Independent private network technology primarily offers exclusive network services for large medical institutions. Public and private network base stations can be deployed in the same physical area to ensure complete physical isolation. This technology utilizes the industry-specific 4.9 GHz band, while ordinary users operate on the 2.6 GHz band.

Bearing Network Slicing Technology

As illustrated in Figure 5, the 5G bearer network slicing primarily uses the international standard FlexE technology, introducing a FlexE Shim layer to achieve decoupling of the media access control (MAC) and physical layers. Additionally, it utilizes the FlexE time slot multiplexing technology to divide the physical port of the operator's bearer network into multiple subchannel ports. This allows for complete isolation of services between each slice at the forwarding level through hardware time slot multiplexing, offering superior isolation compared with other forwarding isolation technologies. The 5G SA medical network slice, built on the original FlexE technology, uses time slot separation technology to enhance the granularity of the 5 Gbps slice to a more refined 1 Gbps granularity for medical-specific network slices. This solution offers an expanded virtual slice network for medical services, enabling efficient management and support of diverse medical business scenarios while simultaneously reducing the costs associated with constructing a private medical network.

XSL•FO
RenderX

Figure 5. The FlexE Shim layer. FlexE: Flexible Ethernet; IP: internet protocol; MAC: media access control; MPLS: multiprotocol label switching; NP: network processor; PHY: physical layer; QoS: quality of service; TM: traffic management.

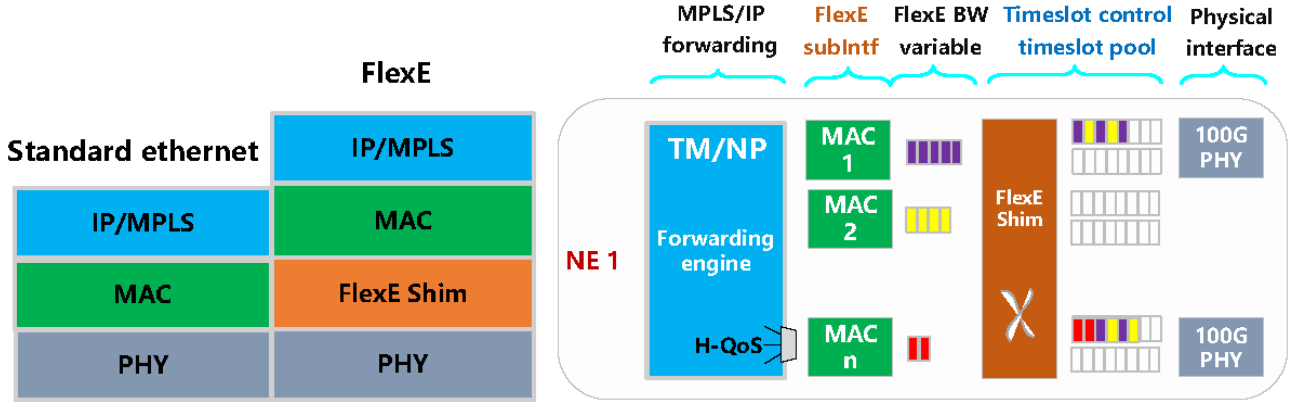

Core Network Slicing Technology

The operation of core network subslicing relies on the collaborative functioning of various network element functions (NFs), including logical function entities defined by the 3GPP standards, such as the access and mobility management function, session management function, user plane function (UPF), unified data management, and policy control function [20]. The 5GC network supports a flexible combination of standard NF services alongside common services defined by the 3GPP. The visual interface enables the flexible arrangement of NFs through a drag-and-drop feature, facilitating the creation of customized network slices. Core network slice isolation technology primarily refers to the combination of varying degrees of isolation for each NF within the slice, allowing each NF to select its level of isolation based on the specific application scenario.

The NF isolation in the core network primarily adopts the following methods based on the requirements of service-level agreements (SLAs), cost considerations, and security isolation. First, the partial control method utilizes a specific user interface, offering moderate safety isolation while supporting multiple slices for the same terminal. This approach balances general costs with ease of deployment. Second, the specific user interface method focuses on isolating particular scenarios, which has the lowest cost but can be challenging to deploy. Currently, the medical field is well-suited for user-specific solutions. Lastly, the full isolation method separates the control interface from the user interface, making it suitable for scenarios with high-security isolation or customization demands. However, this approach entails the highest cost and moderate deployment difficulty. To address the diverse needs of various smart medical applications, the core network slice isolation adopts a dedicated scheme for the user plane. The UPF is deployed separately for different industries, relying on open and reliable connections, as well as computing and storage resources, to support the flexible delivery of multiecological services at the access edge.

Edge Cloud Collaboration Technology

Edge cloud collaboration targets different levels of medical institutions and business-to-business medical application scenarios. It relies on provincial-, municipal-, and county-level medical facilities to establish a distributed medical service system, including stationed edge clouds and prefectural and municipal medical clouds. This approach fully leverages edge computing capabilities, such as real-time or accelerated data processing and analysis, efficient network traffic management, and enhanced local data security. The distributed medical service system facilitates collaboration among multilevel medical clouds, edge networks, resources, and services. Specifically, the provincial medical cloud primarily serves large medical institutions at the provincial level, deploying industry-level collaborative and massive computing operations. It acts as the core for business release, management, and supervision, overseeing collaboration across all edge networks. The municipal center cloud functions as a city-level medical collaboration center, managing the business of the provincial medical cloud and overseeing the release, management, and supervision of city-level medical operations. Coordinating with the provincial medical cloud, the municipal center cloud oversees the management of all stationed edge clouds within the city. The stationed edge cloud primarily supports the business endpoints and data storage centers of county-level medical institutions, enhancing the response efficiency of real-time operations by diverting local business and implementing real-time processing. For collaborative tasks or those requiring significant computational power, data are uploaded to the city-level central cloud for calculation and processing.

Service Aware Technology

Medical Network Management and Control Platform

It is imperative to establish a medical network management and control platform to support diverse medical services. Leveraging advanced networking technologies, including visualization, manageability, control, and security monitoring, the platform can create a visual, manageable, and controllable environment for medical networks, business operations, and security.

Network Visualization

By utilizing the millisecond-following flow detection technique of in-band flow information telemetry (IFIT), alongside telemetry reporting and intelligent analysis from the network management system, network visualization can monitor service SLA in real time. This capability allows for quick fault delimitation and localization, enabling rapid recovery of services within minutes. SLA visualization technology can be effectively applied in the operation and maintenance of large-scale medical private networks, significantly enhancing efficiency.

XSL•FO
RenderX

Network Coordination

Based on the Route and Area Matching (ROAM) intelligent routing algorithm, in conjunction with the original routing algorithm, the system selects the optimal path according to various network requirements, including bandwidth and service time delay. This approach addresses the limitations of traditional routing algorithms, which typically focus solely on the shortest path. By enabling path planning and dynamic optimization at the service level, it enhances the overall utilization rate of the medical network.

Network Control

The management and control platform monitors slice bandwidth utilization and historical service concurrency in real time, allowing for reasonable planning of slice sizes. In the event of network congestion, the platform automatically detects the utilization rate of network slices and dynamically adjusts their bandwidth to ensure the smooth operation of medical services.

Security Detection

Security in 5G medical networks encompasses the protection of terminal access, network transmission, and system applications. Terminal access security ensures that 5G terminals utilize a special network card and implement security measures through native protocols, such as the 5G Authentication and Key Agreement mechanism, as well as 2-way authentication, to safeguard access to the network. To ensure visibility, manageability, and controllability of network security in the medical industry, security measures such as deploying security equipment or software—including the UPF security gateway, NF isolation, IFIT, telemetry, firewalls, and advanced persistent threat intrusion detection—are implemented. Additionally, system application security focuses on delivering network and business security services to users within the medical sector. To ensure the security of terminals, data, and systems in medical businesses, various security mechanisms are used, including system log-in verification, face and fingerprint recognition, certification authority signature authentication, management of system permissions, and data desensitization for medical information sharing.

Ethical Considerations

Ethical approval of this study was granted by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (No.2020-KY-381).

Results

Construction of a Private 5G SA Medical Network

Framework Overview

The private 5G SA medical network is constructed on key 5G technologies and tailored to meet the specific needs of medical services. This innovative network offers a range of services, including bandwidth, latency, connection capacity, and security isolation, addressing diverse network requirements. It utilizes various slicing technologies in the wireless access, transport, and core networks. This facilitates the transformation of traditional network architecture, allowing for the selection and management of medical slices tailored to different users in health care settings. This ensures prompt responses to various medical services while promoting the safe, standardized, and efficient operation of health care operations. The proposed 5G SA medical network within a smart health environment is illustrated in Figure 6. The construction of the 5G network primarily encompasses the medical application system, information communication standards, and network security mechanisms.

Figure 6. The proposed 5G SA medical network in a smart health environment. 5G: 5th generation mobile communication technology; IoT: Internet of Things; LTE: Long Term Evolution; SA: standalone.

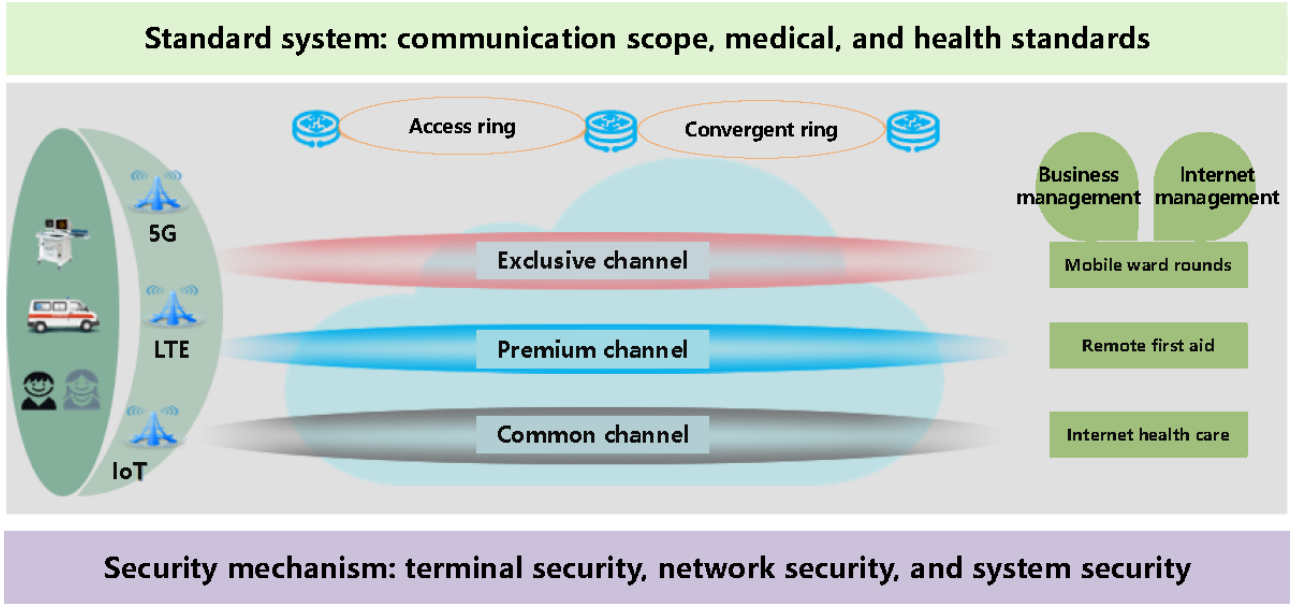

As depicted in Figure 6, the private 5G SA medical network offers customized virtual network slicing services categorized into 3 types: exclusive, premium, and common networks. This structure meets the high-reliability network requirements for various application scenarios, including mobile ward rounds, remote first aid, and internet health care, while also optimizing

XSL•FO
RenderX

the user's network experience. The medical exclusive channel supports services accessed by fixed service sites of medical institutions and allows connectivity to various networks, including 3G/4G Long Term Evolution (LTE), Narrowband IoT, and Wi-Fi. The premium channel is designated for out-of-hospital mobile terminal access scenarios. In cities and counties that support these mobile services, a dedicated network virtual slice is preallocated for the access ring in the region, with real-time synchronization of the routing tables across all access rings in the entire network. During actual business operations, multiple subnetwork slices are connected to the same premium channel to ensure that out-of-hospital mobile business data remain within the hospital, thereby protecting patient privacy and information security. After allocating bandwidth to the exclusive and premium channels, the remaining bandwidth is automatically assigned to the common channel, with network bandwidth resources dynamically adjusted based on actual business needs.

Deployment Framework

Through end-to-end network slicing technology, the 5G SA medical network allocates independent resources across the "province-city-county" hierarchy, ensuring rapid and comprehensive coverage of medical institutions at all levels. This approach achieves isolation between medical and public services, safeguarding medical operations from any potential impact caused by public services. The 5G SA medical network supports access through 5G, 4G, Wi-Fi, the internet, and dedicated virtual private network lines. The iGW-S gateway is deployed in the hospital's machine room to facilitate access to fixed devices and edge cloud services. In large- and medium-sized hospitals or operator data centers, the iGW-M and UPF are utilized to manage the diversion of private network services. Small hospitals can lease the operator's iGW-C gateway, which supports access for mobile devices within the hospital and for vehicle-mounted devices outside of it. Mobile and internet medical users access the regional medical cloud through the operator's iGW-C, significantly expanding the coverage of smart medical services. Medical alliances and communities connect to the iGW-M via either the iGW-C or iGW-S to access the edge or medical cloud. The iGW-M then uniformly distributes interhospital collaborative traffic to the edge or medical cloud, facilitating business collaboration between fixed and mobile access units. This centralized access point ensures network connectivity across the entire system, simplifies network connections at the business layer, and guarantees the efficient and high-quality execution of medical services.

Physical Components of the Hardware System

To implement the 5G SA network, various hardware components are utilized, including 5G base stations, 5GC networks, transmission networks, edge computing systems, antennas, power and environmental monitoring devices, network management tools, and user terminals. The main physical components are outlined in Table 2. Together, these hardware devices form the infrastructure of the 5G SA network, enabling the delivery of high-bandwidth, low-latency, and highly reliable communication services.

Table 2. The main physical components of hardware used in the 5G[a] SA[b] network.

Hardware	Physical components	Description
5G base station equipment	Antenna, radiofrequency module, baseband processing unit	Wireless access points of the 5G network and communicating with user terminals
5G Core network equipment	Active antenna unit, distributed and centralized unit, access and aggregation interface, user plane function, session management, and access and mobility management	Supporting key 5G features such as high-speed data transmission, low-latency communications, and large-scale device connectivity
Transmission network	Optical transmission equipment, IP/MPLS[c] routers, and switches	Connecting the 5G base stations and core network
Edge computing equipment	Smart cameras, Internet of Things gateways, edge servers, and micro data centers	To handle delay-sensitive services and reduce the burden on the core network
Antennas equipment	Massive MIMO[d] antenna system	Improving network capacity and coverage
Equipment of power and environmental monitoring	Two-circuit mains power supply or UPS[e] power supply, and intelligent power detector	Providing power support for base stations and core networks, and monitoring the operating status of hardware
Network management equipment	Software-defined networking controller, fiber optic switches, and virtual private network gateways	Managing and monitoring the entire 5G SA network
User terminals	Mobile phones, 5G portable Wi-Fi, customer terminal equipment, and Internet of Things	Communicate directly with the 5G base stations and support the 5G SA network

[a]5G: 5th generation mobile communication technology.

[b]SA: standalone.

[c]IP/MPLS: Internet Protocol/Multi-Protocol Label Switching.

[d]MIMO: Multiple Input Multiple Output.

[e]UPS: uninterruptible power supply.

XSL•FO
RenderX

Medical Application System

The 5G SA network is compatible with existing network modes, including 3G, 4G LTE, and IoT. To address the diverse network requirements of medical services, it utilizes network slicing and virtualization technology to construct exclusive, preferential, and common channels. This approach supports medical business operations and network management, facilitating seamless integration of mobile ward rounds, remote first aid, and internet-based health care services across intrahospital, interhospital, and out-of-hospital scenarios (Figure 6).

Information Communication Standards

The network adheres to the relevant standards of 3GPP R15 for wireless and transmission networks, as well as core network slices. It provides precise definitions for slices, message queuing telemetry transport protocol, differentiated scheduling processing, and resource mapping, ensuring the standardization of network construction. Additionally, it supports Health Level Seven (HL7), the 10th revision of the International Classification of Diseases (ICD-10), and other medical and health information communication standards, facilitating the interaction and sharing of various medical service information.

Network Security Mechanism

The security of the 5G SA network is underpinned by technologies such as business slice isolation, software-defined networking, certificate authority authentication, hierarchical decentralization, and blockchain. These technologies effectively manage terminal access and identity while ensuring secure data access and application processing. The 5G network provides security authentication, abnormal behavior analysis, and dynamic access control for medical terminals, networks, and application systems, thereby safeguarding the integrity of the medical private network.

Applications of the Private 5G SA Medical Network

Robot-Assisted Mobile Ward Round

Robot-assisted mobile ward rounds extend medical care services directly to the patient's bedside. This system is built on the IoT, edge computing, data fusion, and robotics, integrating various medical information systems within hospitals. It enhances daily nursing ward rounds by enabling medical professionals to utilize the 5G network to access medical data acquisition terminals, audio and video interaction systems, electronic medical records, and robot control systems. This integration supports efficient medical data transmission and nondestructive data interaction. Consequently, the homogenization of mobile ward round business services can be achieved, ultimately enhancing the quality and efficiency of ward rounds and nursing services (Figure 7). Additionally, in specialized areas such as isolation wards and environments with electromagnetic radiation, health care professionals can maneuver medical service robots to designated beds for conducting remote ward rounds and providing nursing services. This further ensures the personal safety of medical staff.

Figure 7. The 5G-enabled robot-assisted mobile ward rounds. 5G: 5th generation mobile communication technology.

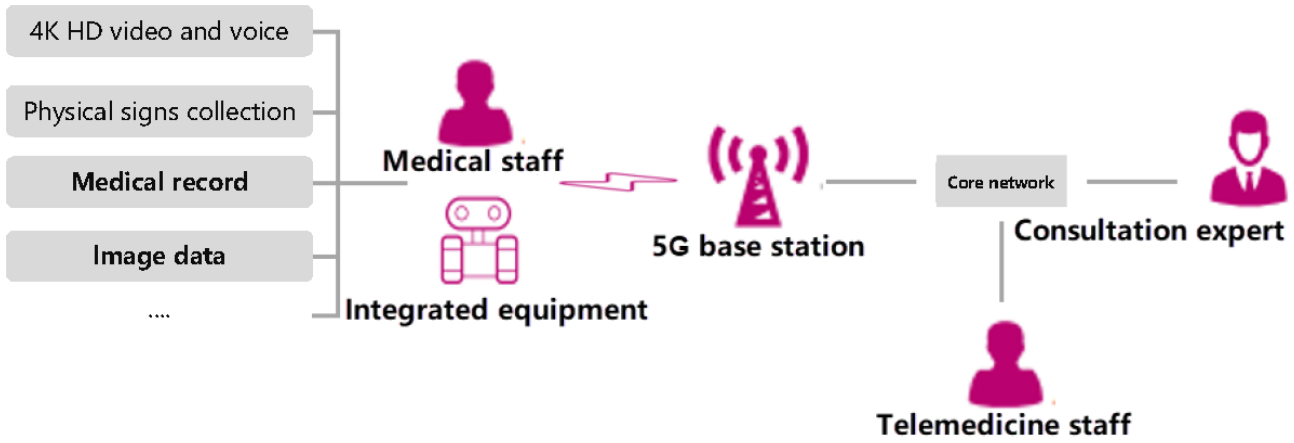

The robot-assisted mobile ward rounds primarily encompass application scenarios such as IoT-based data collection, remote control of clinical examinations and tests, and audio-video interaction for telemedicine. To meet the actual needs of medical staff for clinical diagnosis and treatment—including real-time control of smart terminals, stable and reliable mobile terminals, and secure data exchange—it is both necessary and feasible to establish a 5G SA medical private network. This network will provide differentiated and secure support for medical services. As shown in Table 3, the performance of 4G is inadequate to meet the transmission requirements for 4K high-definition video, control information, and medical data. Regarding terminal mobility, under conditions of network security and stability, the actual terminal mobile speed for 4G is 200 km/h, which is 250 km/h lower than the 450 km/h achievable by 5G terminals. Additionally, 4G lacks dedicated transmission channels for industry users, making it challenging to ensure adequate security. The proposed 5G SA network supports network slicing, allowing for the allocation of multiple slices specifically for smart medical networks. It provides dedicated wireless air interface resources and control management for medical services. This ensures the independence and isolation of these services, thereby enhancing the feasibility, real-time capabilities, reliability, and security of medical services.

During the COVID-19 epidemic, medical staff could remotely control a mobile ward inspection robot to enter the isolation ward via an exclusive 5G medical service channel. This enabled them to gather vital signs, such as body temperature and electrocardiograms, from patients, as well as access electronic medical records. The system also facilitated collaboration with

XSL•FO
RenderX

external experts for multidisciplinary consultations and provided remote guidance to medical staff in the isolation ward for patient rescue. The medical exclusive channel creates a clear separation between medical services and public affairs, mitigating the impact of sudden influxes of public activity during critical periods. It establishes a seamless connection between health care professionals and patients. To conduct network testing and service applications during the COVID-19 epidemic, mobile ward rounds utilizing the proposed 5G SA medical network were implemented in the isolation ward area of the First Affiliated Hospital of Zhengzhou University. The field test results showed a peak 5G download rate of 918 Mbps and an average download rate of 790 Mbps. The peak upload rate was 94.15 Mbps, with an average upload rate of 91 Mbps.

Table 3. The feasibility of the 5G[a]-enabled robot-assisted mobile ward rounds.

Items and requirements of typical applications	Indicators	Basic needs	4G[b]	5G SA[c]
Expert side				
High-definition lossless transmission of doctor's 4K video	Downlink	36 Mbps	20 Mbps	91 Mbps
No-sense real-time transmission of control information	Latency	10 ms	20-30 ms	1 ms
Medical data	Downlink	200 Mbps	100 Mbps	790 Mbps
Robot side				
High-definition lossless transmission of patient's 4K video	Uplink	36 Mbps	20 Mbps	91 Mbps
No-sense real-time transmission of control information	Latency	10 ms	20-30 ms	1 ms
Medical data	Uplink	50 Mbps	20 Mbps	91 Mbps
Mobile terminal				
Stable and reliable network	Speed	300 km/h	200 km/h	450 km/h
Security				
Security interaction of medical privacy data	Data security	Data security access	Network reuse and no security guarantee	FlexE[d] logical slicing, hard isolation, and high security

[a]5G: 5th generation mobile communication technology.

[b]4G: 4th generation mobile communication technology.

[c]SA: standalone.

[d]FlexE: Flexible Ethernet.

Remote Prehospital First Aid

Prehospital first aid is a crucial component of emergency medicine, aimed at improving the rescue efficiency of patients before they reach the hospital, as well as reducing disability and mortality rates. By implementing the 5G SA medical network, the digital prehospital first aid system at the emergency center can equip first aid vehicles with essential medical devices such as monitors, electrocardiographs, emergency breathing machines, defibrillation monitors, and cardiopulmonary resuscitators. By adopting technologies such as mobile communication, IoT, and audio and video communication, and centering on the prehospital first aid electronic medical record system, we can integrate multidimensional systems for vital signs monitoring, mobile vehicles, vehicle navigation, GPS positioning, and emergency command and dispatch. With this design and construction, remote prehospital first aid can support the transmission of video images and inspection results to emergency command centers and hospitals through 5G premium channels, whether ambulances are stationary or in motion. It also enables real-time transmission of information related to medical equipment monitoring, vehicle positioning, and video footage both inside and outside the vehicle. Furthermore, it facilitates the collection, processing, storage, transmission, and sharing of information pertinent to remote prehospital first aid. If needed, clinical experts can be connected through the prehospital emergency system for remote rescue guidance.

To verify the effectiveness of the 5G SA network in the prehospital emergency system, the First Affiliated Hospital of Zhengzhou University established a network test environment in the Longzihu area of Zhengzhou City. They conducted both static and dynamic service tests on ambulances equipped with the remote first aid system. The network bandwidth, jitter, delay, and other performance parameters of the system were recorded during operation under various conditions. These data were then used to compare and analyze network changes during actual application. The test results demonstrated that the 5G SA network effectively meets the requirements of diverse business applications in prehospital emergency scenarios (Figure 8). The upstream bandwidth of 4G was clearly insufficient compared with 5G, exhibiting poor delay stability and a significant range of fluctuations, especially during regional boundary switches. Additionally, in high-load networks with concurrent multiservice usage, the performance of the 4G network deteriorated more rapidly than that of 5G.

XSL•FO
RenderX

Figure 8. Comparison of uplink and downlink average service bandwidth of the 4G/5G-enabled remote first aid service at static/dynamic status. 4G: 4th generation mobile communication technology; 5G: 5th generation mobile communication technology; CT: computed tomography; DR: digital radiography; MR: magnetic resonance.

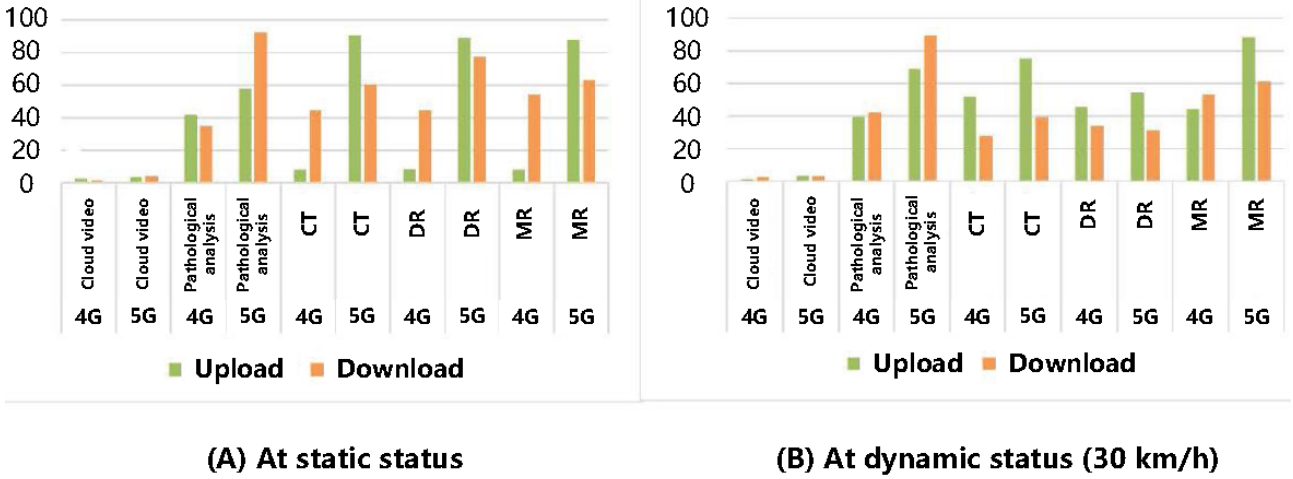

Remote Ultrasound

The noninvasive nature of medical ultrasonography greatly contributes to its widespread clinical use, which largely depends on the scanning techniques employed by the physician. Selecting an incorrect scanning area can lead to misdiagnosis and inappropriate treatment of patients. Remote ultrasound experts can leverage the low-latency 5G network to control the robot arm in real time using a handheld simulation ultrasound device. This capability enables them to perform ultrasound examinations for patients in county-level hospitals and provide diagnostic recommendations to local doctors. The full automation of real-time remote ultrasound diagnosis across regions and hospitals minimizes the medical and nursing operations required at the patient's side. As a result, the quality of diagnosis and treatment, as well as the level of ultrasound imaging services in primary hospitals, can be significantly improved.

Compared with 4G, the latest 5G technology offers a transmission rate that is 20 times higher, mobile speeds that are 1.5 times faster, and latency that is 10 times lower. In addition to its high bandwidth, low latency, and extensive connectivity, 5G can support multichannel transmissions of high-definition video and precise ultrasound images. It also enables the simultaneous networking of a significant number of medical devices both inside and outside of hospitals. Thus, the quality of images transmitted and received by the 5G-powered remote ultrasound diagnostic system surpasses the limitations of the 4G network and meets the requirements for remote ultrasound services. In this study, to verify the feasibility of remote ultrasound, a 5G-enabled ultrasound examination system was implemented between the First Affiliated Hospital of Zhengzhou University and Luoning County People's Hospital (Figure 9). The test results indicated that the average downlink rate of the 5G network was 4.82 Mbps, while the average uplink rate reached 2 Mbps, with an average fluctuation of approximately 8 ms. The system delay of 83 ms was significantly lower than the perceptible artificial delay of 200 ms, thereby meeting the practical requirements for remote ultrasound examinations.

Figure 9. The 5G-enabled remote ultrasonic examination. 5G: 5th generation mobile communication technology.

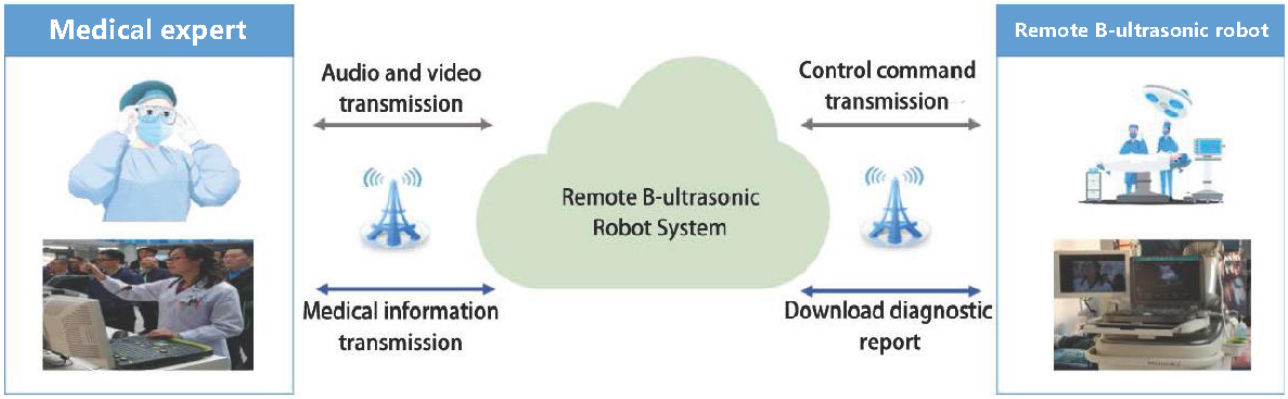

Performance Tests

Given the specific requirements of medical applications in intrahospital, interhospital, and out-of-hospital scenarios, this study conducted static and dynamic business tests under both 4G and 5G conditions to evaluate the performance of the proposed 5G SA medical private network. The tools used during the performance tests are listed in Table 4. The testing environment for the 5G network includes a 5G macro station, 64T64R, customer premise equipment (CPE) access, a core network SA, and bearer network packet transport network equipment. By contrast, the 4G environment comprises a 4G macro station, 8T8R, mobile phone or CPE access, a core network EPC, and bearer network packet transport network

XSL•FO
RenderX

equipment. The tests record performance parameters such as network bandwidth, jitter, and latency during system operation under different conditions. These data are then used to compare and analyze network changes during the actual use of medical services under both 4G and 5G conditions, aiming to support the implementation of 5G smart medical applications.

Table 4. Details of the testing tools.

Testing tools	Description
Flow instrument (Sunrise Telecom)	Simulates network congestion
5G[a] New Radio CPE[b]	5G network switches to wired broadband
HUAWEI mate 20 X (5G packet filling test)	Mobile testing
HUAWEI mate 20 X (4G[c] card)	Pathology, consultation, and electrocardiogram tests
4G CPE	Cloud video testing in a 4G environment
GENEX Probe 5.2 (testing software; Informer Technologies, Inc.)	Provides upload/download/ping services for both server side and client side

[a]5G: 5th generation mobile communication technology.

[b]CPE: customer premise equipment.

[c]4G: 4th generation mobile communication technology.

The 2R2T CPE at the test site of the First Affiliated Hospital of Zhengzhou University is connected to the 4.9G macro station and accesses the SA core network via an exclusive network channel. The doctor's workstation connects to the CPE using a standard 6-network cable, while traffic and delay changes through the CPE are monitored using GENEX Probe software. During the bandwidth testing process, data packets were continuously sent to the core network server using user datagram protocol (UDP), with the CPE connected to the nearest base station for relevant testing. The measured uplink rate of the 5G SA medical network, as shown in Figure 10A, remained consistently stable at 90 Mbps. The download rate of the network is depicted in Figure 10B, with the physical layer's rate remaining stable at 800 Mbps. The MAC layer and the radio link control layer are components of the data link layer. The compression and encryption processes that occur when information transitions from the physical layer to the data link layer can lead to attenuation and fluctuations in the download rate, which is considered a normal occurrence.

Figure 10. Bandwidth test of the private 5G SA medical network. 5G: 5th generation mobile communication technology; DL: download; MAC: media access control; PHY: physical layer; RLC: radio link control; SA: standalone; UL: uplink.

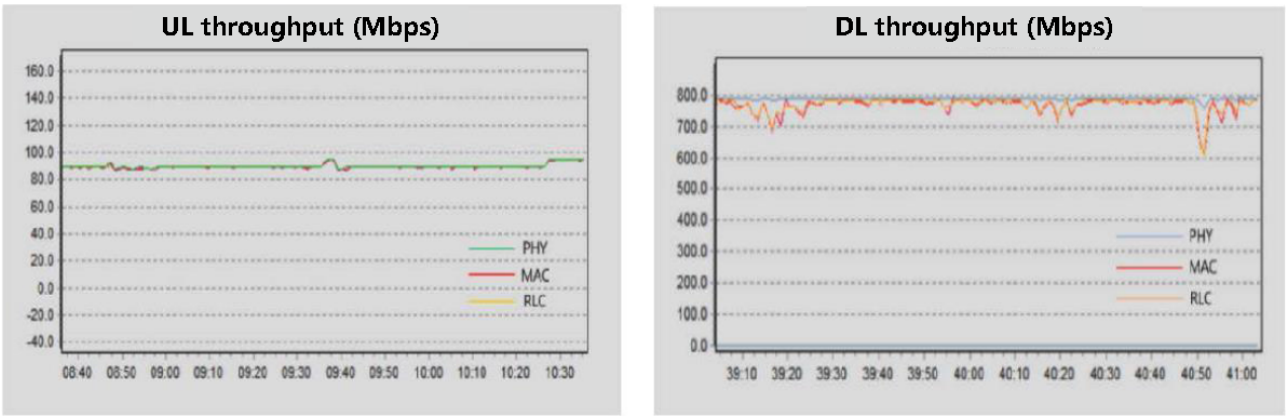

(A) Uplink rate (B) Download rate

The background flow is created using a flow meter, which occupies the physical pipeline to simulate a congested scenario. The 5G medical network slice has a bandwidth of 1 Gbps, while the ordinary user slice offers a bandwidth of 3 Gbps. The background traffic measures 2.935 Gbps, with the pulse signal entering the ordinary user slice. The medical special network service, ordinary user service, and background flow service share the network port. The test results are shown in Figure 11.

Before the background stream was introduced, the downlink rate of the medical private network remained stable at over 700 Mbps, while the download bandwidth of the ordinary user service fluctuated around 600 Mbps. After the background stream was input, the download bandwidth of the 5G medical network service remained unaffected, whereas the download rate of the ordinary user service dropped to below 100 Mbps.

XSL•FO
RenderX

Figure 11. Comparison between the medical private network and the common network before and after the flow. DL: download; MAC: media access control; PHY: physical layer; RLC: radio link control.

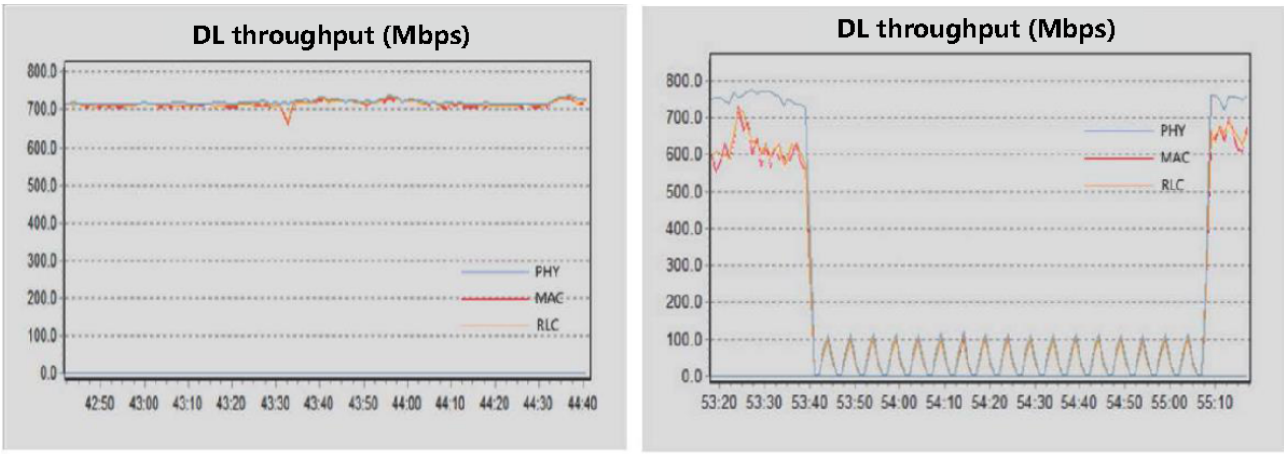

(A) Medical private network (B) General network

Before and after the network congestion, ordinary user services were significantly impacted, as background stream activities consumed a substantial portion of the bandwidth allocated for normal operations. This resulted in a significant decrease in download speeds for regular users. By contrast, the user services of the medical private network remained largely unaffected after the congestion, fully meeting the demands for high concurrency in medical services.

Interhospital cross-regional services, such as remote ultrasound, require stricter time delay standards. This study established a test environment to conduct delay jitter tests both within and between hospitals. During in-hospital testing, ping packets were sent directly to the doctor at the remote ultrasound workstation, using a batch size of 1000 messages, each 32 bytes in length. In 180 Ping tests, 178 results fell between 8 and 10 ms, with an average delay of 8.86 ms. During the interhospital tests, Luoning County Hospital, Xinghua Town Health Center, and Dongsi Village were all directly connected to the local 64T64R 5G macro station via 2R2T CPE. Additionally, the doctor workstation at the First Affiliated Hospital of Zhengzhou University sent Ping packets to each county hospital's workstation. The test results are presented in Figure 12. Compared with the NSA network, the SA network demonstrated a significant reduction in delay, with the average interhospital latency measured at 14 ms.

Figure 12. Delay test of the private 5G SA medical network. 5G: 5th generation mobile communication technology; NSA: non-standalone; SA: standalone.

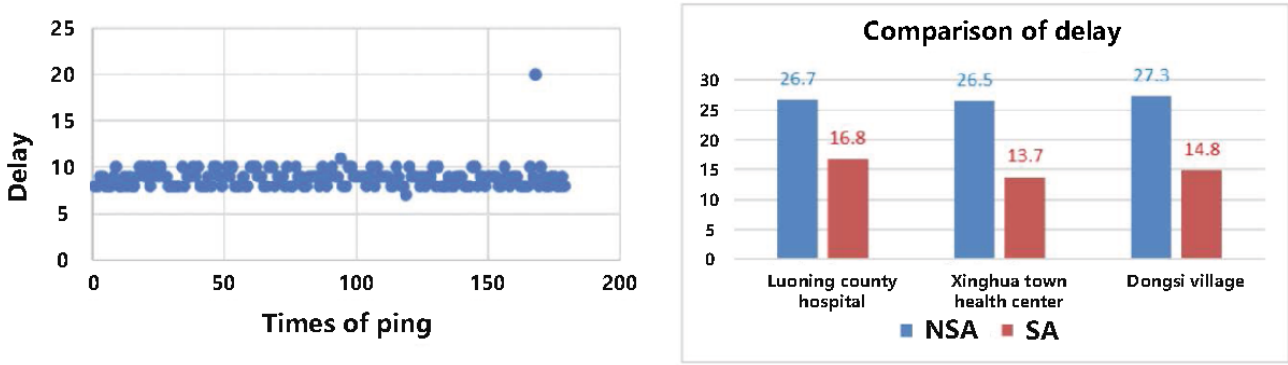

(A) In-hospital delay testing (B) Interregional delay testing

Discussion

Principal Findings

Based on a comprehensive requirements analysis and the design of the system and functional architecture, this study developed and launched a private 5G SA medical network using key technologies such as end-to-end elastic network slicing, edge cloud collaboration, and service awareness. The construction and functionality of the network were explored, along with 3 typical applications and performance tests of the proposed 5G SA medical network. To our knowledge, this is the first comprehensive account of a private 5G SA medical network explored in a smart health environment based on real-world practice in China. The findings can serve as a valuable reference for the further development and implementation of 5G medical networks in China and other regions worldwide.

The characteristics and advantages of 5G are precisely what current medical services urgently require. Challenges that remained unresolved during the 4G era—such as the interconnection of intrahospital equipment, the implementation of interhospital medical services, and the effective integration of out-of-hospital emergency treatment with regional health

XSL•FO
RenderX

care systems—can be swiftly addressed in the 5G era [5,6,12,21]. The advent of 5G facilitates the interconnection of all things, empowers mobile health care, enhances AI performance, and accelerates the industrialization of smart medical care [22]. Compared with NSA, SA offers 2 distinct advantages. First, network slicing in SA supports differentiated SLAs for medical services, providing varying QoS guarantees for medical applications. This improves network service quality and reduces deployment costs [23,24]. Second, SA facilitates the implementation of distributed cloud data centers. The mobile edge computing platform on the hospital side allows users to connect directly with the edge node, enhancing the efficiency and effectiveness of processes and operations. With extremely low network latency, SA enables applications in medical service robots, remote surgery, and other medical services [25,26]. Although the deployment costs for SA are higher than those for NSA, its operational expenses are comparatively lower. Therefore, while NSA may serve as an interim solution, it is expected to be replaced by SA in the future [9].

In terms of the proposed system design in this study, the advantages of a private 5G SA medical network with cloud-native characteristics are primarily reflected in the following aspects [27,28]. First, network compatibility: building on traditional medical networks, it supports multiple access modes such as satellite, Wi-Fi, wired networks, and cellular data networks. This versatility facilitates large-scale medical data collection and sharing, as well as collaboration in medical services. Second, the network slicing design utilizes different slicing technologies based on the functional positioning, business scenarios, and 5G characteristics of medical institutions. This allows for the hierarchical deployment of 5G networks and facilitates the flexible combination of medical private networks through decoupled microservices. As a result, diverse network requirements for various medical service scenarios can be promptly addressed. The isolation of medical services from public user services enhances the security of medical operations. Third, edge cloud collaboration involves deploying edge cloud systems at various levels of medical institutions. Real-time interactive tasks are offloaded to the edge cloud for processing, while non–real-time and computationally intensive tasks are uploaded to the medical cloud for execution. This approach improves the utilization of IT resources and enhances the real-time performance of medical services. Fourth, intelligent operation and maintenance leverage the types of medical businesses and historical data to analyze concurrent 5G network traffic, ensuring service-level commitments for bandwidth and latency. Minute-level fault location is achieved through the use of IFIT and telemetry, while medical application–level SLA guarantees are provided by implementing SLA slice selection and management functions.

The “Overall Plan for 5G Technology R&D and Testing” [29] released by IMT-2020 outlines clear performance index requirements. Specifically, the user experience rate should reach 100 Mbps in wide-area coverage scenarios and 1 Gbps in hotspot coverage areas, while the end-to-end delay must remain within 10 ms [30,31]. Based on the test results of the proposed 5G SA medical network, the current bandwidth, concurrency, and delay all meet the performance index requirements set by IMT-2020. Meanwhile, the logical bandwidth of the 5G SA medical network slicing at the First Affiliated Hospital of Zhengzhou University is 1G, with enhanced granularity, representing only one-fifth of the previous network slicing. This means that the proposed solution for the 5G SA medical network allows for the allocation of additional network resources, thereby promoting the development and implementation of differentiated medical applications [10,24].

For the construction, application, and performance of the 5G SA medical network in a smart health environment, this study makes several notable contributions. First, the proposed 5G SA network integrates network slicing and edge computing technologies, which effectively address the requirements for high bandwidth, wide coverage, low latency, and visualization in medical applications. Second, this study introduces a security mechanism tailored for the SA medical network in 5G to protect user privacy. Third, 3 typical applications of the 5G SA medical network were explored: robot-assisted mobile ward rounds, remote ultrasonic examinations, and virtual reality–enabled remote emergencies. Lastly, the 5G SA medical network underwent a series of experimental tests to accurately assess its performance, providing valuable references through specific indicators.

Based on the test results, the quality of the 5G SA medical network slicing can be assured and remains unaffected by public business emergencies. However, due to current network capacity limitations, several challenges persist. First, the level of integration between the medical industry’s private network and medical business operations is currently suboptimal and requires further optimization. To date, the application of network slicing has only achieved end-to-end network slicing and facilitated the construction of the 5G SA network. However, the current integration with medical business operations is insufficient. A crucial focus of future research on medical network slicing should be the integration between medical business microservices and medical slice microservices [1,7]. Second, achieving efficient deployment of end-to-end network slices is essential for the practical implementation of medical private networks. The end-to-end slice manager facilitates the rapid creation, deployment, and deletion of network slices. Additionally, slice deployment and upper-level applications in the medical industry can be automatically connected through relevant APIs, expediting the instantiation process and enabling large-scale deployment of the 5G medical network [20,32]. Third, the medical industry currently lacks sufficient 5G-enabled application scenarios, highlighting the need to optimize business models. Current 5G medical network application scenarios are predominantly confined to laboratory testing due to insufficient support from the 5G industry chain, limited wireless coverage, and a singular 5G application mode [3,21,26]. This situation presents several notable issues, including limited diversity in business types, unclear business demands, and ambiguous business models [32,33]. The next step should involve investigating and verifying additional medical applications, such as remote surgery and virtual medical teaching, along with other relevant scenarios. It is essential to focus on researching the business traffic model of the medical network that integrates

XSL•FO
RenderX

5G and wired networks, while further expanding and optimizing the private 5G medical network [21,22,34].

Conclusions

The 5G network is capable of meeting diverse requirements for intelligent medical equipment access, patient privacy security, the complexity of medical scenarios, and the mobility of medical services. In this study, a private 5G SA medical network was proposed for a smart health environment. On the one hand, the construction and application of the 5G SA medical network could give rise to a series of innovative medical service models. It provides a robust network foundation for the rapid development of new service models, including medical service robots, AI-assisted diagnosis, and virtual doctors. Simultaneously, the incubation of related industries will enhance the efficiency of medical services and improve the interactive experience for patients. On the other hand, the 5G SA medical network serves as a significant reference for the design and implementation of specialized networks in other industries. The objective of this study is to establish a differentiated 5G SA private medical network that caters to various levels of medical institutions and their corresponding services. This network not only meets the requirements for the development and implementation of medical services such as mobile medical care, telemedicine, and internet-based health care, but also offers construction ideas and reference plans for 5G deployment in other sectors, including agriculture, households, transportation, and power.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant 72374183); the Key Research and Development and Promotion Special Project of Henan Province (scientific and technological research; grants 232102310062 and 222102310678); the Soft Science Project of the Henan Province Medical Science and Technology Research Plan (grant RKX202202022); the Joint Construction Project of the Henan Province Medical Science and Technology Research Plan (grant LHGJ20210323); and the Key Scientific Research Project Plan for Higher Education Institutions of Henan Province (grant 24A320024). The funders played no role in the design, development, or interpretation of this work. The views expressed in the article are those of the authors and do not necessarily reflect the position of the funding bodies.

Conflicts of Interest

None declared.

References

1. Lanza M, Sebastian A, Lu WD, Le Gallo M, Chang M, Akinwande D, et al. Memristive technologies for data storage, computation, encryption, and radio-frequency communication. Science 2022 Jun 03;376(6597):eabj9979. [doi: 10.1126/science.abj9979] [Medline: 35653464]
2. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on Wearable Technology for Connected Health: Scoping Review of Research Trends, Advances, and Barriers. J Med Internet Res 2019 Sep 05;21(9):e14017 [FREE Full text] [doi: 10.2196/14017] [Medline: 31489843]
3. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020 Jul 14;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
4. Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? Int J Epidemiol 2020 Feb 22;49(3):717-726. [doi: 10.1093/ije/dyaa033] [Medline: 32086938]
5. Moglia A, Georgiou K, Marinov B, Georgiou E, Berchiolli RN, Satava RM, et al. 5G in healthcare: from COVID-19 to future challenges. IEEE J Biomed Health Inform 2022 Aug;26(8):4187-4196. [doi: 10.1109/JBHI.2022.3181205] [Medline: 35675255]
6. Xiang T, Zhang PY, Zhuo GY, Dai H. Contribution of the 5G smart first-aid care platform to achieving high-quality prehospital care. J Med Internet Res 2023 May 22;25:e43374 [FREE Full text] [doi: 10.2196/43374] [Medline: 37213169]
7. Chen H, Sun H, Chen Y. Application of 5G technology to conduct real-time teleretinal laser photocoagulation for the treatment of diabetic retinopathy-reply. JAMA Ophthalmol 2022 Feb 01;140(2):205-206. [doi: 10.1001/jamaophthalmol.2021.5391] [Medline: 34913946]
8. Ksentini A, Nikaein N. Toward enforcing network slicing on RAN: flexibility and resources abstraction. IEEE Commun Mag 2017;55(6):102-108. [doi: 10.1109/mcom.2017.1601119]
9. Fehmi H, Amr MF, Bahnasse A, Talea M. 5G network: analysis and compare 5G NSA /5G SA. Procedia Computer Science 2022;203:594-598. [doi: 10.1016/j.procs.2022.07.085]
10. Mohamed R, Zemouri S, Verikoukis C. Performance evaluation and comparison between SA and NSA 5G networks in indoor environment. New York, NY: IEEE; 2021 Presented at: 2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom); September 7-10, 2021; Athens, Greece p. 112-116. [doi: 10.1109/meditcom49071.2021.9647621]

XSL•FO
RenderX

11. Yi C, Cai J. A priority-aware truthful mechanism for supporting multi-class delay-sensitive medical packet transmissions in e-health networks. IEEE Trans on Mobile Comput 2017 Sep 1;16(9):2422-2435. [doi: 10.1109/tmc.2016.2628034]
12. Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine during the COVID-19 pandemic: experiences from Western China. J Med Internet Res 2020 May 08;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]
13. Porambage P, Okwuibe J, Liyanage M, Ylianttila M, Taleb T. Survey on Multi-Access Edge Computing for Internet of Things Realization. IEEE Commun. Surv. Tutorials 2018 Jun 21;20(4):2961-2991. [doi: 10.1109/comst.2018.2849509]
14. Zhai Y, Xu X, Chen B, Lu H, Wang Y, Li S, et al. 5G-network-enabled smart ambulance: architecture, application, and evaluation. IEEE Network 2021 Jan;35(1):190-196. [doi: 10.1109/mnet.011.2000014]
15. Halabian H. Distributed resource allocation optimization in 5G virtualized networks. IEEE J Select Areas Commun 2019 Mar 5;37(3):627-642. [doi: 10.1109/jsac.2019.2894305]
16. Mei J, Wang X, Zheng K. Intelligent network slicing for V2X services toward 5G. IEEE Network 2019 Nov 28;33(6):196-204. [doi: 10.1109/mnet.001.1800528]
17. Chekired DA, Togou MA, Khoukhi L, Ksentini A. 5G-slicing-enabled scalable SDN core network: toward an ultra-low latency of autonomous driving service. IEEE J Select Areas Commun 2019 Aug 10;37(8):1769-1782. [doi: 10.1109/jsac.2019.2927065]
18. Vilalta R, Martínez R, Casellas R, Muñoz R, Lee Y, Fei L. Network slicing using dynamic flex ethernet over transport networks. 2017 Sep 17 Presented at: 2017 European Conference on Optical Communication (ECOC); September 17-21, 2017; Gothenburg, Sweden. [doi: 10.1109/ecoc.2017.8346065]
19. Gutierrez-Estevez DM, Wang Y, Gramaglia M, Domenico AD, Dandachi G, Khatibi S, et al. Artificial intelligence for elastic management and orchestration of 5G networks. IEEE Wireless Commun 2019 Oct 21;26(5):134-141. [doi: 10.1109/mwc.2019.1800498]
20. Sattar D, Matrawy A. Optimal slice allocation in 5G Core networks. IEEE Netw Lett 2019 Mar 29;1(2):48-51. [doi: 10.1109/lnet.2019.2908351]
21. Zhou B, Wu Q, Zhao X, Zhang W, Wu W, Guo Z. Construction of 5G all-wireless network and information system for cabin hospitals. J Am Med Inform Assoc 2020 Jun 01;27(6):934-938 [FREE Full text] [doi: 10.1093/jamia/ocaa045] [Medline: 32447373]
22. Guo C, Li H. Application of 5G network combined with AI robots in personalized nursing in China: a literature review. Front Public Health 2022 Aug 24;10:948303 [FREE Full text] [doi: 10.3389/fpubh.2022.948303] [Medline: 36091551]
23. Zhang H, Liu N, Chu X, Long K, Aghvami A, Leung VCM. Network slicing based 5G and future mobile networks: mobility, resource management, and challenges. IEEE Commun Mag 2017 Aug 8;55(8):138-145. [doi: 10.1109/mcom.2017.1600940]
24. Mallikarjun S, Schellenberger C, Hobelsberger C, Schotten H, editors. Performance analysis of a private 5G SA campus network. Berlin, Germany: VDE; 2022 May 18 Presented at: Communication - Technologies and Applications; 26th ITG-Symposium; May 18-19, 2022; Osnabrueck, Germany p. 1-5.
25. Guo H, Liu J, Zhang J. Computation offloading for multi-access mobile edge computing in ultra-dense networks. IEEE Commun Mag 2018 Aug 14;56(8):14-19. [doi: 10.1109/mcom.2018.1701069]
26. Naruse K, Yamashita T, Onishi Y, Niitaka Y, Uchida F, Kawahata K, et al. High-quality transmission of cardiotocogram and fetal information using a 5G system: pilot experiment. JMIR Med Inform 2020 Sep 08;8(9):e19744 [FREE Full text] [doi: 10.2196/19744] [Medline: 32897237]
27. Chergui H, Verikoukis C. Offline SLA-constrained deep learning for 5G networks reliable and dynamic end-to-end slicing. IEEE J Select Areas Commun 2020 Feb 1;38(2):350-360. [doi: 10.1109/jsac.2019.2959186]
28. Hua Y, Li R, Zhao Z, Chen X, Zhang H. GAN-powered deep distributional reinforcement learning for resource management in network slicing. IEEE J Select Areas Commun 2020 Feb 1;38(2):334-349. [doi: 10.1109/jsac.2019.2959185]
29. IMT-2020. IMT-2020. 2022. URL: https://www.imt2020.org.cn/html/ /default/zh/chengguofabu/huiyiwenjian/1538696064364744705.html?index=3 [accessed 2024-10-13]
30. Fuentes M, Carcel JL, Dietrich C, Yu L, Garro E, Pauli V, et al. 5G new radio evaluation against IMT-2020 key performance indicators. IEEE Access 2020 Jun 11;8:110880-110896. [doi: 10.1109/access.2020.3001641]
31. Romano G. IMT-2020 Requirements and Realization. Wiley 5G Ref (eds R. Tafazolli, C.-L. Wang and P. Chatzimisios) 2020 Dec 29;5:1-28 [FREE Full text] [doi: 10.1002/9781119471509.w5gref216]
32. Wen M, Li Q, Kim KJ, Lopez-Perez D, Dobre O, Poor HV, et al. Private 5G networks: concepts, architectures, and research landscape. IEEE J Sel Top Signal Process 2022 Jan 1;16(1):7-25. [doi: 10.1109/jstsp.2021.3137669]
33. Yang Y, Geng S, Zhang B, Zhang J, Wang Z, Zhang Y, et al. Long term 5G network traffic forecasting via modeling non-stationarity with deep learning. Commun Eng 2023 Jun 06;2(1):33. [doi: 10.1038/s44172-023-00081-4]
34. Yan W, Shu Q, Gao P. Security risk prevention and control deployment for 5G private industrial networks. China Commun 2021 Oct 4;18(9):167-174. [doi: 10.23919/jcc.2021.09.013]

Abbreviations

3GPP: 3rd Generation Partnership Project
4G: 4th generation mobile communication technology

XSL•FO
RenderX

5G: 5th generation mobile communication technology
5GC: 5G Core
AI: artificial intelligence
API: application programming interface
CPE: customer premise equipment
EPC: evolved packet core
FlexE: Flexible Ethernet
HL7: Health Level Seven
ICD-10: 10th revision of the International Classification of Diseases
IFIT: in-band flow information telemetry
IoT: Internet of Things
LTE: Long Term Evolution
MAC: media access control
NF: network element function
NSA: non-standalone
QCI: QoS class identifier
QoS: quality of service
RB: resource block
ROAM: Route and Area Matching
SA: standalone
SLA: service-level agreement
UPF: user plane function

Edited by A Coristine; submitted 01.09.23; peer-reviewed by A Martian, R Gutierrez-Castrejon, Y Niu; comments to author 01.07.24; revised version received 31.07.24; accepted 03.09.24; published 24.10.24.

Please cite as:
Chen B, Shi X, Feng T, Jiang S, Zhai Y, Ren M, Liu D, Wang C, Gao J
Construction and Application of a Private 5G Standalone Medical Network in a Smart Health Environment: Exploratory Practice From China
J Med Internet Res 2024;26:e52404
URL: https://www.jmir.org/2024/1/e52404
doi:10.2196/52404
PMID:39446419

©Baozhan Chen, Xiaobing Shi, Tianyi Feng, Shuai Jiang, Yunkai Zhai, Mingxing Ren, Dongqing Liu, Chengzeng Wang, Jinghong Gao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

In-Depth Examination of the Functionality and Performance of the Internet Hospital Information Platform: Development and Usability Study

Guang-Wei Zhang[1,2*], MD, PhD; Bin Li[3*], SE; Zheng-Min Gu[3*], SE; Wei-Feng Yang[3*], SE; Yi-Ran Wang[4,5], BD; Hui-Jun Li[4,5], BD; Han-Bing Zheng[6], BD; Ying-Xu Yue[7], BD; Kui-Zhong Wang[8], BD; Mengchun Gong[9], MD; Da-Xin Gong[1,2], MD, PhD

[1]Department of Smart Hospital, The First Hospital of China Medical University, Shenyang, China
[2]The Internet Hospital Branch of the Chinese Research Hospital Association, Beijing, China
[3]Department of Information Center, The First Hospital of China Medical University, Shenyang, China
[4]Shenyang Medical & Film Science and Technology Co Ltd, Shenyang, China
[5]Enduring Medicine Smart Innovation Research Institute, Shenyang, China
[6]ZuoYi Technology Co Ltd, Beijing, China
[7]YLZ Ruitu Information Technology Co Ltd, Guangzhou, China
[8]iMEDWAY Technology Co Ltd, Beijing, China
[9]Digital Health China Co Ltd, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Da-Xin Gong, MD, PhD
Department of Smart Hospital
The First Hospital of China Medical University
155 Nanjingbei Street
Shenyang, 110001
China
Phone: 86 02483283350
Fax: 86 02483283350
Email: ydyyzhyy@163.com

Abstract

Background: Internet hospitals (IHs) have rapidly developed as a promising strategy to address supply-demand imbalances in China's medical industry, with their capabilities directly dependent on information platform functionality. Furthermore, a novel theory of "Trinity" smart hospital has provided advanced guidelines on IH constructions.

Objective: This study aimed to explore the construction experience, construction models, and development prospects based on operational data from IHs.

Methods: Based on existing information systems and internet service functionalities, our hospital has built a "Smart Hospital Internet Information Platform (SHIIP)" for IH operations, actively to expand online services, digitalize traditional health care, and explore health care services modes throughout the entire process and lifecycle. This article encompasses the platform architecture design, technological applications, patient service content and processes, health care professional support features, administrative management tools, and associated operational data.

Results: Our platform has presented a set of data, including 82,279,669 visits, 420,120 online medical consultations, 124,422 electronic prescriptions, 92,285 medication deliveries, 6,965,566 prediagnosis triages, 4,995,824 offline outpatient appointments, 2025 medical education articles with a total of 15,148,310 views, and so on. These data demonstrate the significant role of IH as an indispensable component of our physical hospital services, with deep integration between online and offline health care systems.

Conclusions: The upward trends in various data metrics indicate that our IH has gained significant recognition and usage among both the public and healthcare workers, and may have promising development prospects. Additionally, the platform construction approach, which prioritizes comprehensive service digitization and the 'Trinity' of the public, healthcare workers, and managers,

XSL•FO
RenderX

serves as an effective means of promoting the development of Internet Hospitals. Such insights may prove invaluable in guiding the development of IH and facilitating the continued evolution of the Internet healthcare sector.

(J Med Internet Res 2024;26:e54018) doi:10.2196/54018

KEYWORDS

internet hospital; smart hospital; mobile applications; operational data; information system; online service; patient service; management tool; electronic prescriptions; medical education; integration

Introduction

Internet hospitals (IHs) are an innovative health care service model pioneered in the Chinese medical industry, which have experienced rapid development in recent years, especially after the outbreak of the COVID-19 pandemic [1-6]. As of May 2023, the number of approved IHs has exceeded 3000, with an overwhelming number and recognition in the public ones [7]. They have been regarded as a promising strategy to alleviate the supply-demand imbalance and shape the future development of the medical industry, due to their potential to offer improved accessibility, efficient resource utilization, enhanced doctor-patient communication, enable personalized medicine, support public health emergencies, and facilitate medical innovation [8-10]. However, the current operational conditions of many IHs are far from ideal, even in a "zombie state" [11]. The reasons that may account for this situation, include low enthusiasm among health care professionals, limited functionalities and service offerings of IHs, as well as insufficient construction experience among managers without sufficient successful case references [2].

It has been widely acknowledged that internet-based medical platforms are pivotal elements in the construction of IHs, as they form the fundamental infrastructure for the operation of IHs and directly influence the successful implementation and development of IHs [4,8]. Since the launch of the first internet-based diagnosis and treatment platform in 2014 [1], IHs in China have undergone a decade of exploration and a clear and improved evolution in terms of their concept, functional positioning, construction models, and management regulations. Strictly speaking, IHs are no longer confined to online consultations alone, rather they are medical institutions that use the internet to provide medical and related services. In an ideal state, IHs should maximize the networked realization of physical hospital functions, even encompassing all the functions of physical hospitals online. Moreover, a theory of "Trinity" smart hospital construction has been recently developed in China, which emphasizes an intelligence and collaborative development of "smart service" for the people, "smart medicine" for medical workers, and "smart management" for hospital managers [12]. As pioneers in smart hospital initiatives, IHs will be further enhanced in the theoretical framework and functional positioning depending on this new theory [12].

Following these guiding principles, The First Hospital of China Medical University has pioneered the planning, design, and implementation of a comprehensive platform for IHs, called the "Smart Hospital Intelligent Internet Platform (SHIIP)." By leveraging existing information systems (eg, hospital information system and picture archiving and communication system) and internet service functionalities, the platform has undergone progressive development and integration, demonstrating initial effectiveness. This article comprehensively presents our construction experience of this platform, covering its conceptual framework, objectives, overall planning, technological applications, functional achievements, construction process, and application outcomes.

Methods

Overview

The SHIIP provides a comprehensive, humanized, patient-centered medical service system for patients, and also provides comprehensive support for medical staff members and hospital administrators. The details are elaborated in Multimedia Appendix 1, including Objective and Concept, Blueprint of the SHIIP, Technical Architecture, Network Structure, and Data Service Architecture. The functionalities will be described from the perspectives of different users as follows.

First, for the public (as illustrated in Figure 1), the general public is roughly categorized into 4 groups based on their health or medical service requirements, including healthy individuals, routine patients, emergency patients, and referral patients. Addressing the distinct needs of these different groups, SHIIP offers 8 major categories of online services, such as health management services and routine medical services, which also serve as references for the functional partitioning of the interface implementation of the public-side (Figure 2). The ultimate goal is to establish a comprehensive medical and health service ecosystem, encompassing the entire process and lifecycle of patient diagnosis and treatment. Within each service category, numerous specific service projects have been already implemented or are currently in progress, with the platform aiming to continuously expand and introduce additional projects in the future, although Figure 1 provides 3-6 main projects for each service category.

XSL•FO
RenderX

Figure 1. Service processes and functionalities of smart hospital intelligent internet platform for the public. SHIIP: smart hospital intelligent internet platform.

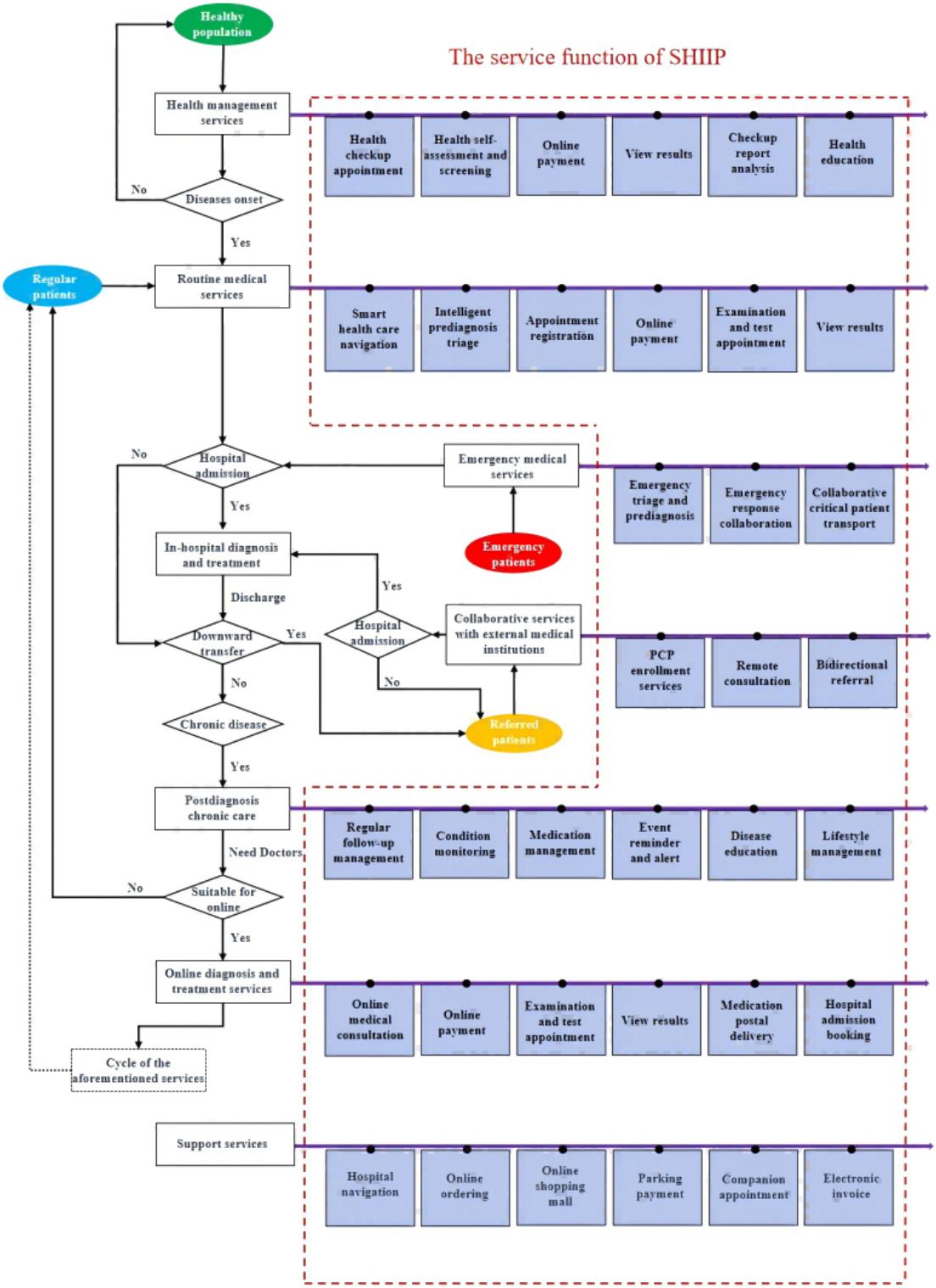

XSL•FO
RenderX

Figure 2. Interface implementation of the public-side. Q&A: question and answer.

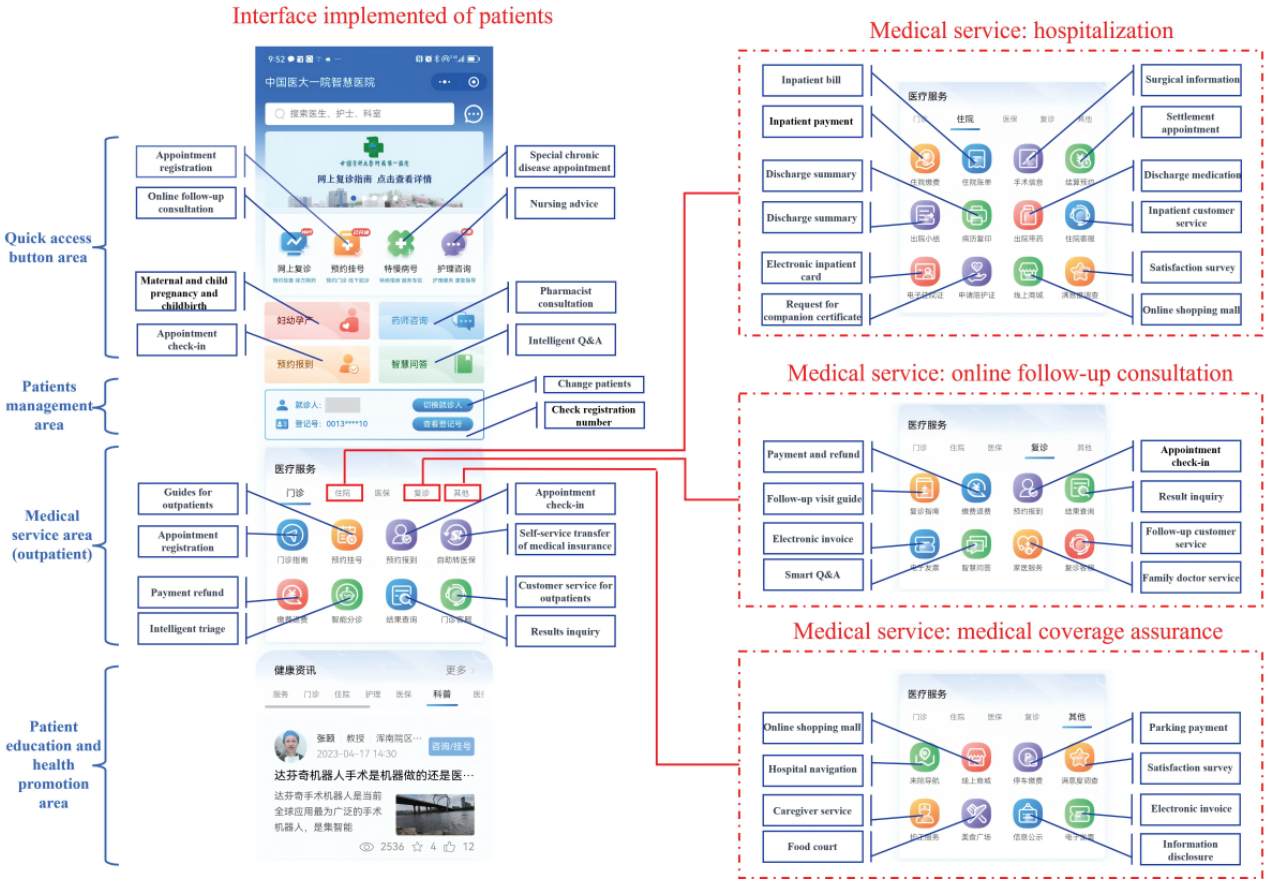

The public involved in the service process is divided into 4 types, each in different colors elliptical box, green for healthy population, red for emergency patients, yellow for referred patients, and blue for regular patients. Starting from the respective colors, the service processes can be identified based on the directions indicated by the arrows (left side), and the corresponding functional items of SHIIP for the public are displayed in the red dashed box (right side). All functions are described in Table S7 in Multimedia Appendix 1.

This figure depicts the home page of SHIIP for the public-side. The functional zones of the home page are described by blue words and black words indicate translation for buttons. The red dashed boxes show the switching pages for 3 medical service objects.

Second, for health care workers (as shown in Figure 3 and Figure 4), the SHIIP provides comprehensive assistance to doctors in various aspects of their medical practice, enabling them to deliver optimal patient care and streamline their operations, such as (1) assistance in medical operations; the platform supports doctors with visit management, antibiotic review, consultation management, publication of medical education, offline clinic appointment, and offline system login; (2) assistance in patient management; the platform offers unified patient management, patient grouping, patient labeling, message sending, research follow-up, and patient recruitment; (3) assistance in browsing and querying; doctors can access historical medical records, prescription information, test and laboratory results, surgery progress information, receive critical value alerts, and share information with colleagues; and (4) assistance in doctor consultation; the platform facilitates doctor consultations with features such as preconsultation data collection, refusal of consultation, prescription recommendations, quick replies, private phone calls, and medical guidance.

XSL•FO
RenderX

Figure 3. The supporting functionalities for health care workers and administrators. SHIIP: smart hospital intelligent internet platform.

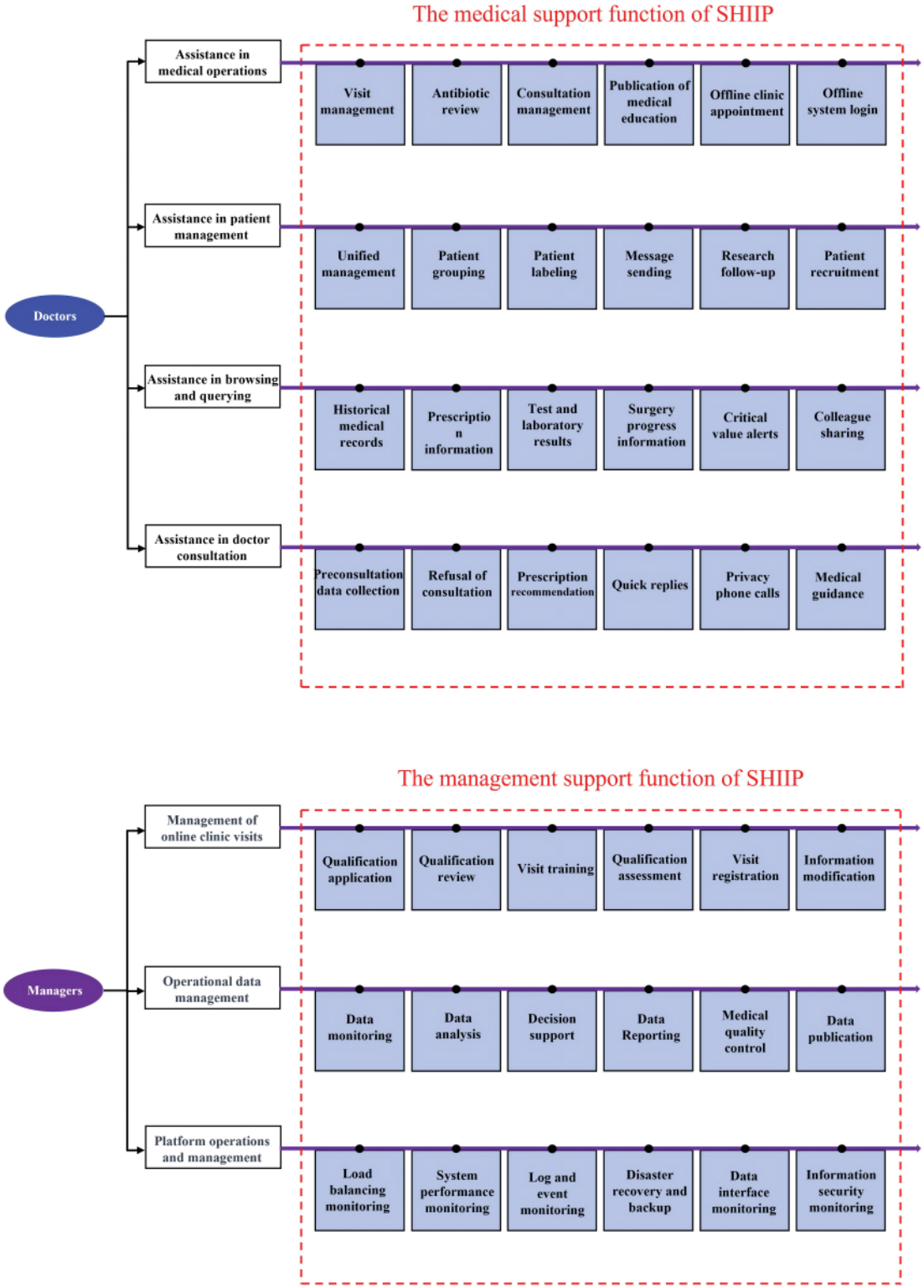

XSL•FO
RenderX

Figure 4. Interface implementation of the health care workers-side.

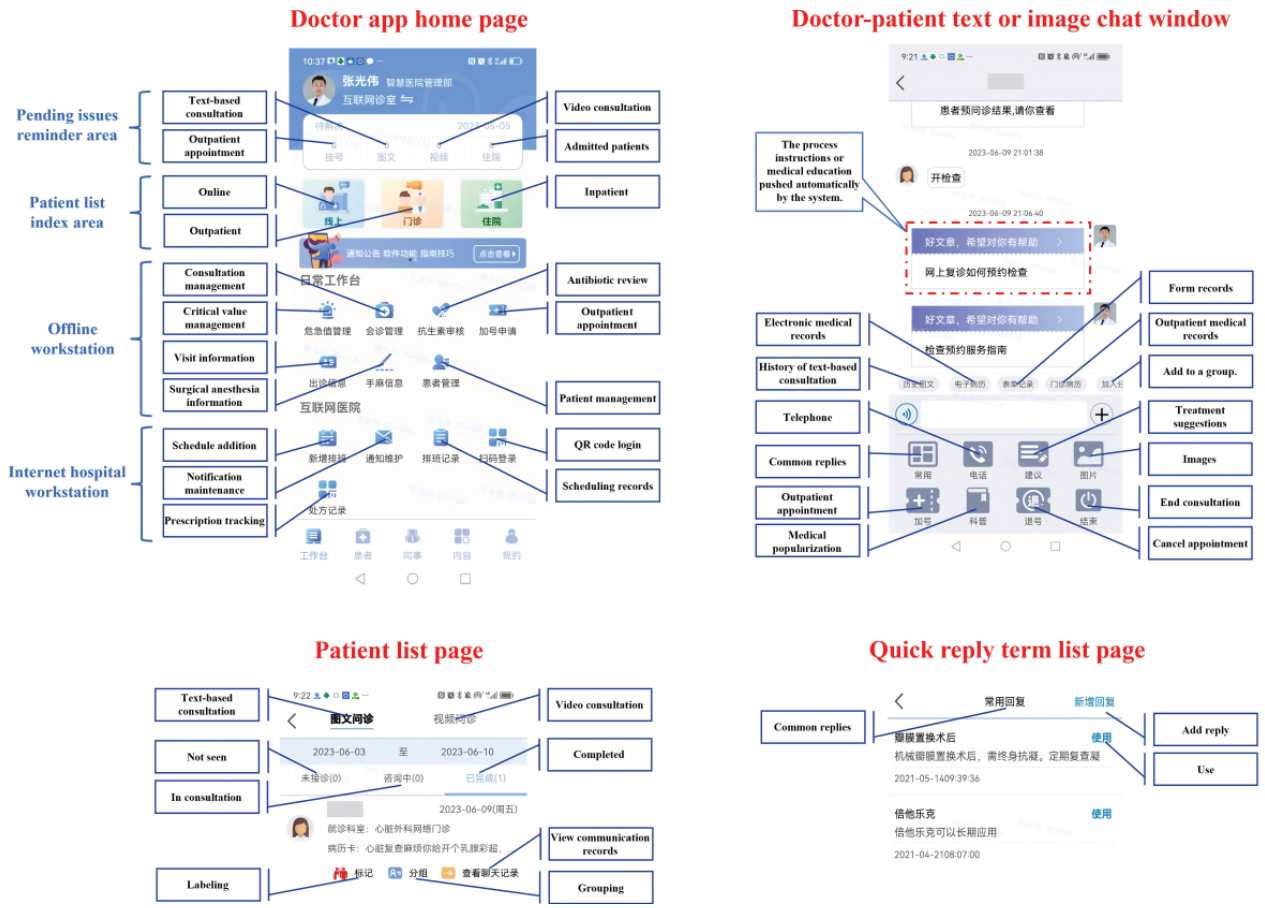

The 2 red dashed boxes represent the main functional items of SHIIP for health care workers and administrators. The functions of health care workers are described in Table S8 in Multimedia Appendix 1. The functions of administrators are described in Table S9 in Multimedia Appendix 1.

The main interfaces of the app for health care workers are shown, including the Doctor app home page, Doctor-patient text or image chat window, Patient list page, and Quick reply term list page (red words). Blue words indicate functional zones and black words present translations for buttons or explanations for regions.

Third, for administrators, the SHIIP offers a range of features and functionalities to support their management responsibilities (Figure 3). These include (1) management of online clinic visits; the SHIIP facilitates the management of online clinic visits through various features, including qualification application, qualification review, visit training, qualification assessment, visit registration, and information modification; (2) operational data management; the SHIIP supports administrators in managing operational data by offering features such as data monitoring, data analysis, decision support, data reporting, medical quality control, and data publication; and (3) platform operations and management; administrators have access to features for platform operations and management, including load balancing monitoring, system performance monitoring, log and event monitoring, disaster recovery and backup, data interface monitoring, and information security monitoring.

Ethical Considerations

In accordance with the authors' institutional ethical guidelines, this study does not involve any experiments on human beings or animals, nor does involve any private data that expose personal information about patients. Therefore, it does not seek an ethics review board assessment.

Results

Overall Operational Data of the Platform's Public-Side

From its launch until December 31, 2023, the SHIIP recorded a total of 82,279,669 visits. The average user visit frequency was 31, and it accumulated 3.6 million user favorites. Further analysis showed a clear upward trend in both visit volume and new user numbers, with growth rates of approximately 73% ($P<.001$) and 29% ($P=.002$), respectively (Figures 5A and 5B). A device source analysis indicated that over 99% (351,877/354,965) of the visits originated from mobile devices (Figure 5C), with less than 1% (3123/354,965) from desktop computers, indicating smartphones as the predominant means of accessing the IH services. Temporal analysis of platform visits showed that over 95% (325,821/343,506) lasted for more than 10 seconds, and nearly 70% (237,961/343,506) exceeded 100 seconds (Figure 5D), identifying a majority of meaningful engagements. Further analysis of the sources of new patient user traffic revealed that the 5 primary methods are link redirection, mobile search, public account menu, one-on-one WeChat (Tencent Holdings Limited) sharing, and offline QR

XSL•FO
RenderX

code scanning (Figure 5E). The foremost 3 of these methods, combined with taskbar and miniprogram messages, also emerged as the principal sources of system access (Figure 5F). These findings underscore the significance of the 5 identified pathways for the analysis, development, and managerial decision-making of the SHIIP platform, delineating critical user traffic trends essential for optimizing engagement strategies.

It presents a comprehensive overview of the SHIIP's usage, user growth, drug delivery distribution, traffic sources, access channels, user engagement duration, and platform accessibility. The growth trend of the number of visits to the platform is shown in Figure 5A. The continuous growth in new patient user numbers over time is displayed in Figure 5B. Figures 5C-5F show, respectively, the traffic diversion sources of new users, distribution of access sources for the platform, distribution of users' access duration, and the distribution of platforms accessed by users. Figure 5G showcases the visual representation of the distribution and volume of medicine deliveries across different regions and darker blue indicates more medicine delivery orders in the area. Figure 5H shows the curve of consultation volume.

Figure 5. Usage statistics and service data of smart hospital intelligent internet platform.

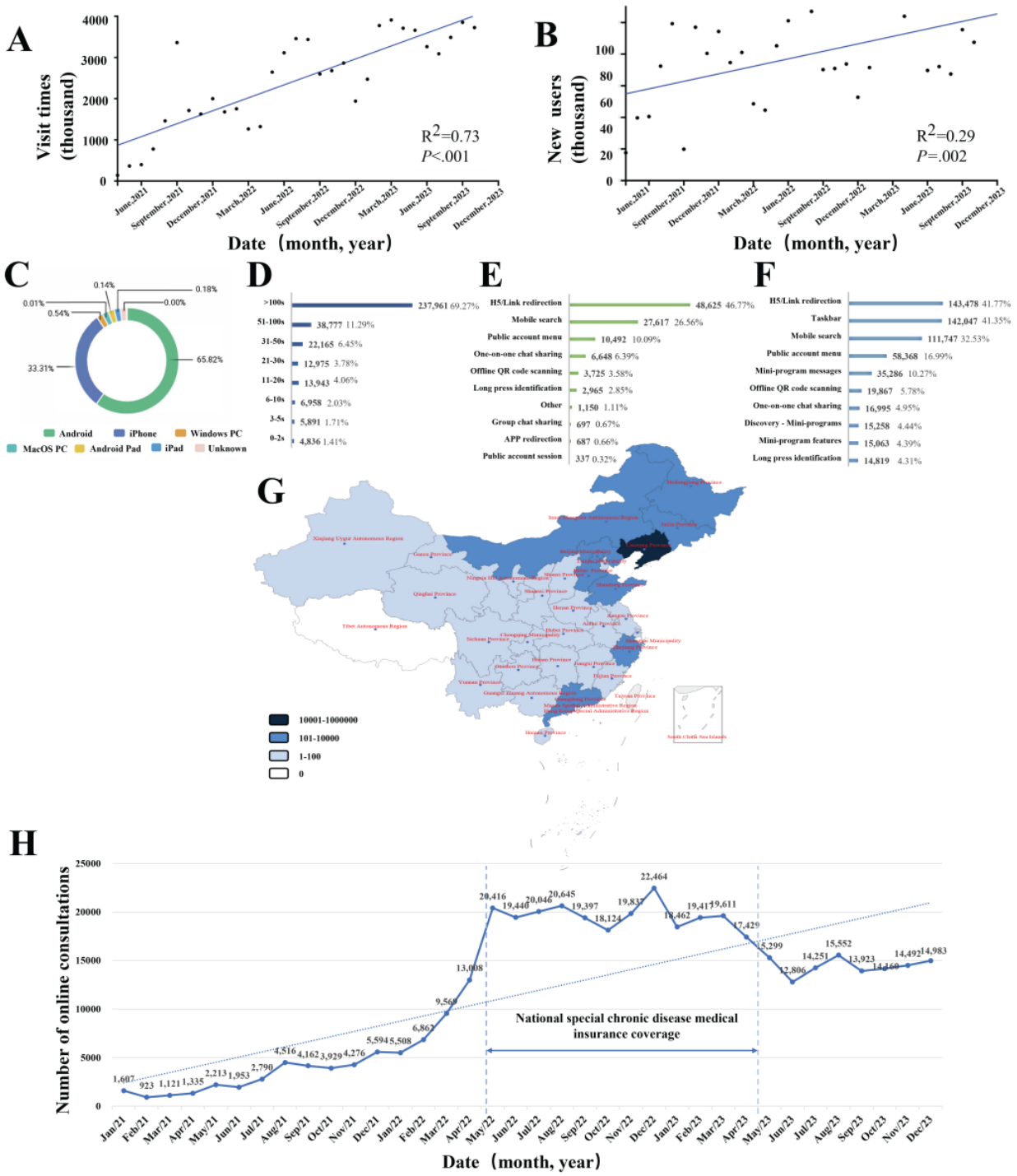

XSL•FO
RenderX

The Overview of Online Consultations

Under the autonomous decision-making mechanism for online services, 574 doctors (out of 797), representing 72% of all qualified physicians, engaged in online consultations, signifying substantial acceptance of online medical practices among the physician community. These practitioners spanned 64 departments and represented 44 distinct medical subdisciplines, encompassing 96% (64/66) and 94% (44/47) of all medical departments and subdisciplines in our hospital, respectively. Although there was an uneven distribution among departments and subdisciplines, with disparities such as 22 doctors in cardiovascular medicine compared with only 1 in general geriatric (Table S4 in Multimedia Appendix 1), the extensive participation across diverse medical fields and specialties has been fundamental in enabling the platform to support a significant volume of medical services throughout its 3 years of operation. The number of online medical consultations reached 420,120 from 180,020 patients, with 124,422 electronic prescriptions, 24,862 examination and test orders, and 92,285 medication deliveries. The delivery range covered 30 provinces, with the farthest delivery distance exceeding 3500 kilometers, demonstrating truly nationwide services (Figure 5G). The total volume of refusals was 64,306, constituting an overall refusal rate of 16%. The primary reason for these refusals was the physicians' assessments that the patients' conditions were unsuitable for online consultations. A quantitative analysis of online medical consultations indicates a consistent upward trend, except for a significant fluctuation between April 2022 and May 2023 (Figure 5H), which was explicitly affected by health insurance policies and pandemic management measures. This underscores the critical influence of policy factors on the development of IH.

The Stratified Analysis of Online Consultations

The annual variations in consultation and refusal volumes across different departments are summarized in Table S3 in Multimedia Appendix 1. Breast surgery I, gastroenterology, rheumatology and immunology, endocrinology, and cardiovascular emerged as the top 5 departments, with total consultation numbers of 26,636, 28,039, 24,152, 18,984, and 16,407, respectively. Furthermore, many departments exhibited a year-over-year growth trend consistent with the overall trajectory, such as endocrinology (increasing from 1417 in 2021 to 14,147 in 2023) and thoracic surgery (from 1117 in 2021 to 11,676 in 2023). Conversely, some departments experienced a decline, for instance, cardiovascular medicine (from 333 in 2021 to 122 in 2023) and endocrinology (from 44 in 2021 to 22 in 2023). Similar patterns of change were also observed at the level of subdisciplines (Table S5 in Multimedia Appendix 1). Beyond policy factors, the reasons for these changes warrant further analysis, which could contribute to regulating departmental consultation volumes.

Furthermore, the analysis at the physician level revealed an average annual consultation volume of approximately 189 per doctor, with an average annual patient-to-doctor ratio of about 1:3. However, significant disparities in this ratio were observed across different subdisciplines, with the largest being 1:1654 in traditional Chinese medicine in 2023, and the smallest at 1:2 in rehabilitation medicine and geriatrics in the same year. In addition, certain departments experienced considerable fluctuations in their annual patient-to-doctor ratios, such as traditional Chinese medicine, which shifted from 1:4 in 2021 to 1:295 in 2022, and then to 1:1654 in 2023. These findings highlight 2 critical insights. First, the development of our IH remains uneven, with many departments needing enhancement; and second, some departments or subdisciplines have already demonstrated strong capabilities in internet-based medical services, affirming the significant role and potential for growth of IHs.

The Disease Spectrum Analysis of Online Consultations

In our disease spectrum analysis, we identified 3191 diseases involved in online consultations, constituting 8.86% (3191/36,016) of all diagnoses currently used in our hospital. The most prevalent condition was health check-ups (n=39,361), followed by common and chronic internal medicine conditions such as nodular goiter (n=6112) and coronary atherosclerotic heart disease (n=3544). Notably, there were also many surgical conditions, including abdominal pain (n=3443) and breast malignancy (n=3190), and even emergency conditions in their chronic phase, like hypokalemia (n=71) and acute pancreatitis (n=31), demonstrating a promising capacity of online consultations for broad disease coverage. A comparative analysis with traditional in-person outpatient data for the top 20 conditions treated (Table S2 in Multimedia Appendix 1) revealed that more than 10% (17,301/126,235) of patients, for certain diseases, opted for online consultations, with the highest being chronic hepatitis B (1523/5919, 25.73%) and systemic lupus erythematosus (1418/5636, 25.16%), followed by breast malignancy (3190/21,961, 14.53%), thyroid malignancy (3118/23,054, 13.52%), anxiety state (2614/24,447, 10.69%), and pregnancy (1712/16,292, 10.51%). However, common conditions like diabetes mellitus (1935/72,378, 2.67%) and hypertension (1890/110,401, 1.71%) showed lower online consultation rates. These results not only reveal a development imbalance similar to that observed at the departmental and subspecialty levels but also suggest that over 25% of patients could be adequately managed through IH-derived services.

Smart Hospital Intelligent Internet Platform Support for Traditional Offline Services

Beyond online consultations, SHIIP has conducted a total of 6,965,566 prediagnosis triages, encompassing 5,895,867 outpatient, 689,576 emergency, 324,271 internet medical, 50,471 fever clinic, and 5381 artificial intelligence (AI) screenings for COVID-19. Over the past 3 years, SHIIP has facilitated 4,995,835 offline outpatient appointments, accounting for 84% (4,995,835/5,928,673) of all bookings, with a notable increasing trend from 67% (1,431,544/2,135,093) in 2021 to 92% (1,483,753/1,606,790) in 2022 and reaching 95% (2,080,538/2,186,790) in 2023. Detailed annual appointment data by department are summarized in Table S6 in Multimedia Appendix 1. In addition, the total number of online examination and test appointments was 187,356. There were 9617 cases of collaborative emergency assistance. In terms of remote consultations, 1311 cases of multidisciplinary remote consultations were completed. Chronic disease services were

XSL•FO
RenderX

provided to 55,148 individuals. In terms of health education, 113 online live consultation sessions were conducted, with 2025 medical education videos and articles published, and a total of 15,148,310 views. There were 2711 appointments for physical examinations, with 486 interpretations of physical examination results. In terms of smart support services, 329,704 orders were completed for online delivery of life support products, 246,677 copies of electronic medical records were printed, and 328,111 applications were made for electronic hospitalization support cards.

Discussion

Principal Results

Despite the widespread attention to IHs, there are absent comprehensive and detailed investigations of their platforms. This study presents a pioneering single-center experience, offering a comprehensive and systematic exposition of the development of IHs within the framework of a "Trinity" smart hospital for the first time [1,7-9]. It encompasses diverse dimensions, including platform architecture design, technological applications, service content and processes for the public, auxiliary functionalities for health care professionals, management tools for administrators, and corresponding operational data and outcomes. Our SHIIP diverges significantly from the narrow perception held by many individuals, who merely associate IHs with the provision of online diagnosis and drug delivery [13]. Through practical implementation, we have demonstrated a new model for the construction and management of IHs, which provides concrete examples to elucidate the conceptual definition, functional positioning, and future development direction of IHs.

Our platform has presented a remarkable set of data on online diagnosis and treatment, medication delivery, remote consultations, and health consultations, demonstrating the significant role of an IH as an indispensable component of our physical hospital services. These exciting results have provided robust evidence of the efficacy of our IH construction and management paradigm, substantiating the viability of health care institutions developing IHs. This success not only validates the operational model but also establishes a benchmark for the strategic integration of digital and traditional health care services. Our data indicate that online consultations encompass a vast majority of departments and prevalent illnesses, suggesting that telemedicine is broadly applicable at the disease level. Furthermore, numerous departments and specialties within our hospital have significant potential for growth in their online consultation ratios, heralding substantial prospects for the development of internet-based health care facilities. More importantly, a consistent increase in monthly traffic and the number of new users indicates a progressive rise in public acceptance and recognition of online health care services, which have also gained widespread recognition from most doctors. These growing trends suggest that IHs will play an increasingly important role in the health care landscape [4,14].

In addition, benefiting from the intelligent and user-friendly features on both the patient and physician sides, it takes an average of only 6 minutes to complete an online medical visit on the platform, significantly shorter than the time required for traditional outpatient visits involving transportation, waiting, consultation, and medication retrieval. Combining these actual business data of our platform and the subjective reports collected from users through surveys conducted by Zhejiang Hospital [15], there is an unequivocal conclusion that IHs will undoubtedly bring convenience and improve work efficiency for users, which stands as a key determinant for the limitless growth potential of IHs in the future. In fact, the internet is precipitating a global transformation in health care service models. In Europe, some countries, like Sweden, have integrated telehealth into their national health system, resulting in significant improvements in health care efficiency and patient satisfaction [16,17]. In the United States, some IHs have elevated health care standards through innovative technological efforts [18]. In New England, the exploration of digital health technologies, particularly the lens of telemedicine's feasibility for patients, provides critical insights into the complexities of digital health integration and the universal challenges faced in ensuring the efficacy and acceptance of such technologies [19]. Therefore, IHs are emerging as a critically important new model in medical services, poised to reshape the landscape of health care delivery globally.

A large amount of noninternet medical data on the platform, including outpatient appointments, result viewing, and online shopping, signify the deep integration of our IH with the physical hospital business. Drawing from similar successful experiences in other hospitals [9,20], it is evident that the digitization of traditional medical services and their incorporation into IH platforms is not only feasible but also highly significant, even essential. This assertion has also been well-supported by the forementioned analogous international examples [21,22]. Furthermore, IHs have showcased their significant potential in prevention and control against the COVID-19 pandemic [3], which is confirmed by our substantial business data again, including online consultations, intelligent screening, and prediagnosis of fever clinics. Therefore, they are poised to play a crucial role in future responses to similar unforeseen events in the realm of public health [6]. Furthermore, it is indicated that our internet platform has become a crucial avenue for delivering health education and promoting public education and awareness, underscoring the important role of IHs in improving the health literacy of residents. Recently, several Chinese hospitals have also reported outstanding achievements in various aspects of their respective IHs. These encouraging results demonstrate the increasingly vital role of IHs in China's health care service system, and the upward trends in our multiple data metrics suggest a promising outlook for its sustained and positive development in the future.

Our use data indicate that smartphones have become the mainstream device for accessing IH services, emphasizing the need for the platform to prioritize the development of mobile apps to meet the demands of primary users. WeChat and Alipay (Alibaba group), each with over 1 billion users, are widely used mobile social media and mobile payment platforms in China. They have emerged as common support systems for building IH platforms, due to their convenience, flexible development frameworks, diverse development technologies, open interfaces,

XSL•FO
RenderX

unified identity authentication, and secure reliability [23,24]. Within their frameworks, the SHIIP was designed to provide services through various forms for catering to the diverse user preferences and usage habits, such as miniprograms and web pages, resulting in significant recognition and usage from a large number of users. Similar outcomes were achieved by multiple IHs, such as Sichuan University West China Second University Hospital [25], providing strong evidence that this strategy will become the mainstream approach for the development of IHs.

Innovation in Concepts, Methods, and Technologies

Unlike traditional physical hospitals, the capabilities of IHs are directly determined by the functionality of information platforms, as their operations rely entirely on the latter [9]. Regarding this aspect, various medical institutions have been actively exploring, from the Second People's Hospital of Guangdong Province sponsoring the first public internet medical platform (Guangdong IH) for online consultation [1], to the First Affiliated Hospital of Zhejiang University attempting family physician services and management of chronic diseases based on online diagnosis and treatment platform [9,26], followed by many ones exploring further various functionalities in terms of patient services or assistance for health care professionals and managers [8,10,12,27]. Until recently, it has been proposed by the Affiliated Second Hospital of Zhejiang University School of Medicine that the IH platform has begun to integrate partially and irreversibly into traditional health care systems [4]. This progressive evolution has led to a clearer and more refined understanding of their functionality and construction direction, serving as a valuable reference for platform development. Furthermore, the success of various third-party commercial IH platforms with their highly flexible and practical features, including "Ping An Good Doctor," "Haodaifu Online," "Jingdong Health," and others [8,28], has also provided invaluable experience for the establishment of IH platforms in public health care institutions.

Combining these experiences and the aforementioned novel theories in this field, we have developed the comprehensive SHIIP, which incorporates several innovative features crucial to its functionality. First, by actively expanding online services and digitalizing traditional health care as illustrated in Figure 2, it further enhances the "patient-centered" service model, standing out with greater comprehensiveness and systematization compared to previously reported online medical platforms [29]. The aim is to minimize time and financial burdens benefiting from a maximally convenient approach that relies on the internet to overcome time and geographic constraints [7,27], such as "single visit" or "zero visit" seeking medical care, an impenetrable challenge for traditional offline outpatient services in China. Second, it serves as an efficient tool for health care professionals and managers, embodying the balanced development of the "Trinity". As summarized in Figure 3, the platform not only incorporates diverse, convenient, flexible, and highly practical functionalities to assist doctors in their work but also continuously achieves digitization in management, which will contribute to the improvement of work and management efficiency [30]. Third, it establishes a data platform that facilitates data interoperability among systems through various means, such as data interface transmission, supporting our platform as a central hub for integrating online and offline functionalities. This innovative approach is expected to further promote the deeper integration of online and offline health care services, building upon existing reports on IHs [4,8], to ensure a smooth flow of health care business processes. Furthermore, based on its ability of data and health care business interoperability, our SHIIP has acted as a pioneer of comprehensive health care services mode throughout the entire process and lifecycle, attempting to cover various health care needs from various patients at various situations as shown in Figure 1. It presents a dramatic extension in the availability of services beyond traditional offline settings, although there are so many aspects that need to be improved.

Due to the complexity and comprehensiveness of our IHs, there is a tremendous demand for information platform development, comparable to the informatization efforts of traditional physical hospitals. Accomplishing this formidable task solely relying on a single system or software vendor is nearly impossible, similar to how the information systems of traditional physical hospitals require hospital information systems, picture archiving and communication systems, laboratory information systems, and other systems from different software vendors. Our platform currently integrates 16 systems from 9 software vendors, which were built before, during, or after the implementation of the internet diagnosis and treatment system, leading to a complex and extensive project in the construction of SHIIP. Therefore, meticulous considerations are crucial for overall planning, including functional coordination, goal achievement, technological advancement, anticipated iterative updates, system scalability, integration, and server allocation. The technological roadmap cannot be achieved instantly, necessitating a step-by-step construction or transformation approach, gradually advancing the integration strategy through various means. In fact, some other IHs have already validated the feasibility of this modular construction approach [4,8]. However, our exploratory work lacks traces and successful cases for reference in many aspects, necessitating continuous adjustments and revisions, despite meticulous planning. This platform construction experience emphasizes that the development of IH information systems should be a process characterized by comprehensiveness, planning, complexity, and continuity, extending far beyond the establishment of a diagnosis and treatment system. Considering that each hospital has different foundational conditions, we recommend formulating strategies tailored to their own characteristics based on the existing state of informatization when constructing IHs.

Data Privacy and Security

Throughout the SHIIP building process, we were fully aware that data privacy and security were our top priorities. Specifically, we have implemented advanced technologies and measures in information security. Our measures include hiding sensitive information, encrypting personal and medical data, and using firewalls, authentication, and security audits to strengthen our defenses. Furthermore, our database servers are located within a core network, isolated from external networks, providing additional protection against unauthorized access and cyberattacks. Despite facing more than 10 cyber threats daily,

XSL•FO
RenderX

our strict security protocols have prevented data breaches. This demonstrates our commitment to information security and ensures our platform integrates online and offline functions without compromising data privacy.

Policies and Norms

In navigating the complex landscape of IHs, adherence to and alignment with established policies and standards have been paramount. As delineated in Table S1 in Multimedia Appendix 1, key regulations such as the "Administration Regulations on Internet Diagnoses and Treatments (Trial)" and "Administration Regulations on Internet Hospital (Trial)" have provided a foundational legal framework that has informed our platform's development, ensuring compliance with quality and safety standards for online diagnoses and treatments. Furthermore, the "Smart Service Scoring System" and "Grading Evaluation Standard for The Application Level of Electronic Medical Record System" have served as benchmarks for assessing the integration of medical resources and the application level of our electronic medical records systems, respectively. These standards not only guided the technological development of our SHIIP but also ensured that our services aligned with the expectations for smart services for the public, smart medicine for medical workers, and smart management for hospital administrators. The "Smart Management Scoring System," in particular, has been instrumental in evaluating the effectiveness of our management practices, reflecting our commitment to maintaining a high standard of smart management. The convergence of these policies and standards has been instrumental in shaping a conducive operational environment for our SHIIP, underscoring the pivotal role of regulatory frameworks in fostering innovation while ensuring patient safety and service quality.

Development Planning and Prospect

In the future, the platform will continue to adhere to the concept of continuous optimization and improvement through the adoption of advanced technologies, aiming to better serve patients, doctors, and administrators. First, AI technology is recognized as playing a crucial role in health care, especially with the emergence of generative AI models like ChatGPT (OpenAI) [31,32]. Building upon the existing AI apps, our platform plans to further expand the scope and methods of AI use, making it more intelligent, personalized, and seamless. Second, extensive research has revealed that the application of medical Internet of Things (IoT) devices is another important direction for enhancing the functionality and capabilities of health care information platforms [33]. In the next phase, the platform intends to gradually integrate IoT devices such as blood pressure monitors and glucose meters through data interface connections, achieving a more comprehensive IoT integration. In addition, the platform will leverage blockchain technology to further enhance information security, as it is considered an effective means to ensure secure sharing and interaction of medical data [34]. Furthermore, innovative health care models like virtual wards and remote surgeries have shown vast potential and will be areas of exploration for our platform [35].

Limitations

Despite the progressive strides taken by our SHIIP, it remains in an embryonic phase of development, characterized by pronounced imbalances in the breadth of services provided, the engagement of various medical departments and professionals, and the spectrum of diseases covered. The foundational structure and administrative frameworks of the SHIIP are yet to be optimized, reflecting a need for ongoing refinement. Furthermore, the lack of comprehensive online quality control protocols tailored to specific ailments underscores a gap in our current operational model. This deficiency points to the imperative for methodical advancements and the establishment of stringent quality assurance measures in future endeavors to ensure the delivery of high-caliber online health care services.

Conclusions

The theoretical framework of IHs in China has undergone significant advancements, with our practical implementation providing substantial evidence of its feasibility and innovation. The extensive construction experience and operational effectiveness benefited from SHIIP, strongly underscore the pivotal role of information platforms in the development of IHs, and their significance in traditional physical hospitals as well. Therefore, it is very important to select an appropriate construction model. While certain aspects still require further validation and refinement, our platform has already achieved considerable scale and will continue to undergo continuous optimization and enhancement. This pioneering exploration holds tremendous significance and serves as a valuable guiding reference for IHs construction and the progressive development of the internet health care sector.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020YFC2006406 and 2020YFC2006401). No generative AI was used during the manuscript generation process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional files.
[DOCX File , 10204 KB - jmir_v26i1e54018_app1.docx]

XSL•FO
RenderX

References

1. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015;3(8):e445-e446 [FREE Full text] [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
2. Xie X, Zhou W, Lin L, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]
3. Gong K, Xu Z, Cai Z, Chen Y, Wang Z. Internet hospitals help prevent and control the epidemic of COVID-19 in China: multicenter user profiling study. J Med Internet Res 2020;22(4):e18908 [FREE Full text] [doi: 10.2196/18908] [Medline: 32250962]
4. Ge F, Qian H, Lei J, Ni Y, Li Q, Wang S, et al. Experiences and challenges of emerging online health services combating COVID-19 in China: retrospective, cross-sectional study of Internet hospitals. JMIR Med Inform 2022;10(6):e37042 [FREE Full text] [doi: 10.2196/37042] [Medline: 35500013]
5. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
6. Sun S, Yu K, Xie Z, Pan X. China empowers internet hospital to fight against COVID-19. J Infect 2020;81(1):e67-e68 [FREE Full text] [doi: 10.1016/j.jinf.2020.03.061] [Medline: 32251688]
7. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021;23(7):e22330 [FREE Full text] [doi: 10.2196/22330] [Medline: 34283025]
8. Zhang Q. The internet hospital: how to combine with traditional healthcare model. Hepatobiliary Surg Nutr 2022;11(2):273-275 [FREE Full text] [doi: 10.21037/hbsn-2022-10] [Medline: 35464285]
9. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021;23(9):e18307 [FREE Full text] [doi: 10.2196/18307] [Medline: 34342267]
10. Qi H, Ying Y, Zhu L, Li Q, Wang T, Chen B, et al. Exploration on the development of public hospital-sponsored telemedicine platform: a case study in China. J Telemed Telecare 2023:1357633X231176871. [doi: 10.1177/1357633X231176871] [Medline: 37309129]
11. Gu JW. The "Zombie State" of 90% of Internet hospitals needs to be broken. NPC & CPPCC. 2022. URL: http://www.rmzxb.com.cn/c/2022-03-03/3061899.shtml [accessed 2024-09-11]
12. Zhang GW, Gong M, Li HJ, Wang S, Gong DX. The "Trinity" smart hospital construction policy promotes the development of hospitals and health management in China. Front Public Health 2023;11:1219407 [FREE Full text] [doi: 10.3389/fpubh.2023.1219407] [Medline: 37546298]
13. Internet hospital. Baidu. URL: https://baike.baidu.com/item/%E4%BA%92%E8%81%94%E7%BD%91%E5%8C%BB%E9%99%A2/23450749?fr=aladdin#reference-[3]-23982017-wrap [accessed 2024-09-11]
14. Cui W, Zhu W, Li X, Wu D, He P, Yu G. Attitudes and perspectives of 534 Chinese pediatricians toward internet hospitals. Front Pediatr 2022;10:948788 [FREE Full text] [doi: 10.3389/fped.2022.948788] [Medline: 36245742]
15. Chen Y, Wang J, Gao W, Yu D, Shou X. Construction and clinical application effect of general surgery patient-oriented nursing information platform using cloud computing. J Healthc Eng 2022;2022:8273701 [FREE Full text] [doi: 10.1155/2022/8273701] [Medline: 35368952]
16. Ellegård LM, Kjellsson G. Telemedicine services associated with higher use of regular primary care in a Swedish region. Lakartidningen 2019;116:FSWP [FREE Full text] [Medline: 31638708]
17. Fonda F, Galazzi A, Chiappinotto S, Justi L, Frydensberg MS, Boesen RL, et al. Healthcare system digital transformation across four European countries: a multiple-case study. Healthcare (Basel) 2023;12(1):16 [FREE Full text] [doi: 10.3390/healthcare12010016] [Medline: 38200922]
18. Fagherazzi G. Towards a European "Cohort Moonshot": revisiting the long-term strategy to support health research of tomorrow. Eur J Epidemiol 2023;38(1):121-122 [FREE Full text] [doi: 10.1007/s10654-022-00939-5] [Medline: 36598706]
19. Boyle T, Boggs K, Gao J, McMahon M, Bedenbaugh R, Schmidt L, et al. Hospital-Level implementation barriers, facilitators, and willingness to use a new regional disaster teleconsultation system: cross-sectional survey study. JMIR Public Health Surveill 2023;9:e44164 [FREE Full text] [doi: 10.2196/44164] [Medline: 37368481]
20. Gao J, Fan C, Chen B, Fan Z, Li L, Wang L, et al. Telemedicine is becoming an increasingly popular way to resolve the unequal distribution of healthcare resources: evidence from China. Front Public Health 2022;10:916303 [FREE Full text] [doi: 10.3389/fpubh.2022.916303] [Medline: 35874991]
21. Sun HY, Ko YM, Lee SW, Lee B, Kim JH. The application rate for urology specialty compared with other specialties from 2007 to 2014 in Korea: is it influenced by social interest manifested by internet trends? BMC Urol 2018;18(1):65 [FREE Full text] [doi: 10.1186/s12894-018-0375-y] [Medline: 30041640]
22. Lee S, Kang WS, Seo S, Kim DW, Ko H, Kim J, et al. Model for predicting in-hospital mortality of physical trauma patients using artificial intelligence techniques: nationwide population-based study in Korea. J Med Internet Res 2022;24(12):e43757 [FREE Full text] [doi: 10.2196/43757] [Medline: 36512392]

XSL•FO
RenderX

23. Chen S, Sun G, Cen X, Liu J, Ye J, Chen J, et al. Characteristics and requirements of hypertensive patients willing to use digital health tools in the Chinese community: a multicentre cross-sectional survey. BMC Public Health 2020;20(1):1333 [FREE Full text] [doi: 10.1186/s12889-020-09462-2] [Medline: 32873273]
24. Liang F. COVID-19 and health code: how digital platforms tackle the pandemic in China. Soc Media Soc 2020;6(3):2056305120947657 [FREE Full text] [doi: 10.1177/2056305120947657] [Medline: 34192023]
25. Li Y, Zhou Y, Chen M, Fu MR, Luo B, Yu P, et al. A WeChat-based rehabilitation platform for children and adolescents with congenital heart disease to promote cardiac FITness (HeartFIT): protocol for a mixed-methods strategy from evidence-based design to pilot study. J Multidiscip Healthc 2022;15:907-920 [FREE Full text] [doi: 10.2147/JMDH.S349519] [Medline: 35519154]
26. Zheyi Internet hospital: how to "+" the Internet and medical care? Zhejiang University. URL: https://www.zju.edu.cn/2017/0217/c41532a445549/pagem.htm [accessed 2024-09-11]
27. Yu T, Jin C, Wu X, Yue D. Implementation of shared decision-making within internet hospitals in China based on patients' needs: feasibility study and content analysis. JMIR Form Res 2023;7:e39965 [FREE Full text] [doi: 10.2196/39965] [Medline: 36607710]
28. Jiang X, Xie H, Tang R, Du Y, Li T, Gao J, et al. Characteristics of online health care services from China's largest online medical platform: cross-sectional survey study. J Med Internet Res 2021;23(4):e25817 [FREE Full text] [doi: 10.2196/25817] [Medline: 33729985]
29. Zhang Q, Zhang R, Lu X, Zhang X. What drives the adoption of online health communities? An empirical study from patient-centric perspective. BMC Health Serv Res 2023;23(1):524 [FREE Full text] [doi: 10.1186/s12913-023-09469-6] [Medline: 37221504]
30. Lv Q, Jiang Y, Qi J, Zhang Y, Zhang X, Fang L, et al. Using mobile apps for health management: a new health care mode in China. JMIR Mhealth Uhealth 2019;7(6):e10299 [FREE Full text] [doi: 10.2196/10299] [Medline: 31162131]
31. Bu F, Sun H, Li L, Tang F, Zhang X, Yan J, et al. Artificial intelligence-based internet hospital pharmacy services in China: perspective based on a case study. Front Pharmacol 2022;13:1027808 [FREE Full text] [doi: 10.3389/fphar.2022.1027808] [Medline: 36438784]
32. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
33. Bhardwaj V, Joshi R, Gaur AM. IoT-based smart health monitoring system for COVID-19. SN Comput Sci 2022;3(2):137 [FREE Full text] [doi: 10.1007/s42979-022-01015-1] [Medline: 35079705]
34. Xie Y, Zhang J, Wang H, Liu P, Liu S, Huo T, et al. Applications of blockchain in the medical field: narrative review. J Med Internet Res 2021;23(10):e28613 [FREE Full text] [doi: 10.2196/28613] [Medline: 34533470]
35. Chen X, Huang X, Yin M. Implementation of hospital-to-home model for nutritional nursing management of patients with chronic kidney disease using artificial intelligence algorithm combined with CT Internet. Contrast Media Mol Imaging 2022;2022:1183988 [FREE Full text] [doi: 10.1155/2022/1183988] [Medline: 35414801]

Abbreviations

AI: artificial intelligence
IH: internet hospital
IoT: Internet of Things
SHIIP: smart hospital intelligent internet platform

Edited by A Coristine; submitted 26.10.23; peer-reviewed by Z Smutny, P Wang, Y Jiang; comments to author 30.01.24; revised version received 30.03.24; accepted 21.08.24; published 08.11.24.

Please cite as:
Zhang GW, Li B, Gu ZM, Yang WF, Wang YR, Li HJ, Zheng HB, Yue YX, Wang KZ, Gong M, Gong DX
In-Depth Examination of the Functionality and Performance of the Internet Hospital Information Platform: Development and Usability Study
J Med Internet Res 2024;26:e54018
URL: https://www.jmir.org/2024/1/e54018
doi:10.2196/54018
PMID:39168813

©Guang-Wei Zhang, Bin Li, Zheng-Min Gu, Wei-Feng Yang, Yi-Ran Wang, Hui-Jun Li, Han-Bing Zheng, Ying-Xu Yue, Kui-Zhong Wang, Mengchun Gong, Da-Xin Gong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Investigating the Norwegian eHealth Governance Model: Document Study

Line Helen Linstad[1,2], MSc; Hilde Bjørnå[3], PhD; Anne Moen[2,4], PhD; Truls Tunby Kristiansen[3], PhD; Anne Helen Hansen[1,5], PhD

[1]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[2]Norwegian Centre for E-health Research, Tromso, Norway
[3]Department of Social Sciences, Faculty of Humanities, Social Sciences and Education, UiT the Arctic University of Norway, Tromsø, Norway
[4]Department of Public Health Sciences, Institute of Health and Society, University of Oslo, Oslo, Norway
[5]University Hospital of North Norway, Tromsø, Norway

Corresponding Author:
Line Helen Linstad, MSc
Department of Community Medicine
Faculty of Health Sciences
UiT The Arctic University of Norway
Postboks 6050 Stakkevollan
Tromsø, 9037
Norway
Phone: 47 41682342
Email: line.h.linstad@uit.no

Abstract

Background: Governments and policy makers struggle to achieve a balance between hierarchical steering and horizontal governance in systems characterized by fragmented decision authority and multiple interests. To realize its One Citizen–One Journal eHealth policy vision, the Norwegian government established a special eHealth board of stakeholders to create an inclusive governance model that aligned stakeholders' interests with the government's ambitions through coordination and consensus. Little empirical knowledge exists on how countries realize inclusive governance models.

Objective: This study aims to investigate how the Norwegian inclusive eHealth governance model was developed as a tool to align the government's policy ambitions with stakeholders' concerns from January 2012 to December 2022.

Methods: This document study used a thematic analysis based on a constructivist research approach. We included 16 policy documents and 175 consultation response documents issued between January 2012 and December 2022 related to the Norwegian One Citizen–One Journal policy implementation process. The themes were constructed deductively from a review of governance models and public governance theory and were applied as our analytical lens to each document. The findings were interpreted, analyzed, and synthesized.

Results: The national policy implementation process progressed through 3 phases, with changes in stakeholder inclusion and perceived influence on the decision-making process characterizing transitions from phase to phase. Tension developed between 2 contrasting views regarding top-down government authority and stakeholders' autonomy. The view of the regional health trusts, municipalities, health care professional organizations, and industry actors contrasted with that of the patient organizations. Governmental insensitivity to participation, lack of transparency, and decreasing trust by stakeholder groups challenged the legitimacy of the inclusive governance model.

Conclusions: We illustrated that Norway's One Citizen–One Journal policy trajectory was characterized by a process that unfolded across 3 distinct phases. The process was characterized by 2 contrasting stakeholder perspectives. Finally, it was shaped by diminishing trust in the inclusive governance model. The National eHealth Governance Board faced challenges in establishing legitimacy as a top-down inclusive governance model, primarily attributed to its addressing of participation, transparency, and trust dilemmas. Such dilemmas represent significant obstacles to inclusive governance models and require ongoing governmental vigilance and responsiveness from governmental entities.

(J Med Internet Res 2024;26:e59717) doi:10.2196/59717

XSL•FO
RenderX

KEYWORDS

eHealth policy; fragmented decision authority; top-down governance; bottom-up network; participation; electronic health record

Introduction

Governance Models in a Multistakeholder Field

The successful implementation of national eHealth programs requires countries to be sensitive to the dynamics of governance and participation [1]. Countries struggle to achieve a balance between hierarchical steering and horizontal governance to create inclusive governance models that nurture synergy among all stakeholders involved in the adoption of eHealth in national health care systems [1,2]. The inclusiveness of the model relates to how it enables stakeholder participation in the policy implementation process [3].

eHealth is a multistakeholder field in which numerous self-regulated actors participate, representing government bodies, health care services, health care professionals' labor unions, industry vendors, and citizens (patients, next of kin, and their interest organizations) [4,5]. These stakeholders have diverse interests, fragmented decision-making authority, and varying operating logics, but they all endorse the same health policy goal, that is, to realize patient-oriented, digitally supported health services across organizational borders [6,7].

Few empirical studies have considered how countries realize inclusive governance models to align stakeholders' interests [2]. Through a recent literature review of eHealth governance models, we found that policy processes and goal attainment often result from negotiations between governments and self-regulated actors [2,8]. The review also revealed that such negotiations generate governance dynamics among hierarchical steering, medical bottom-up governance, market governance, and patients' concerns about their health data and participation in health-related procurement decisions [9].

Insights from Denmark and New Zealand have indicated that top-down policies are often met with bottom-up reactions, prompting negotiation processes [1,2,10,11]. The United Kingdom's and Denmark's experiences have demonstrated that eHealth processes are complex and unpredictable, involving various stakeholders [1,2,11,12]. To maintain progress, researchers have proposed stronger hierarchical steering of the sector to ensure increasing digitalization and innovation in health care [13]. In Scandinavia, eHealth policies are characterized by unstable governance structures, as illustrated by the Danish government's back-and-forth changes in policy, from top-down governance to horizontal governance [1,2,11].

Norwegian Context

The challenges of governing the fragmented eHealth sector, comprising self-regulated actors' diverse interests, are evident in Norway [4,14]. In 2012, the government introduced its eHealth vision, encapsulated in the concept of One Citizen–One Journal (OCOJ) as the overarching objective. The OCOJ initiative aims to address three distinct sets of stakeholder needs: (1) ensuring that health care personnel have easy and secure access to patient data, (2) providing citizens with easy and secure access to digital health services, and (3) making data accessible for quality improvement, health monitoring, governance, and research purposes [4].

The government described eHealth as a domain with a plethora of stakeholders exhibiting different decision authorities and interests. From the government's point of view, the OCOJ was seen as a response to the stakeholders' articulated need for a stronger top-down approach in eHealth [4]. The white paper mandated the Ministry of Health and Social Care to outline the implementation of a national electronic health record (EHR) in a report [4,15].

Norway's strategy was to create the National eHealth Governance Board (NEGB) to implement an inclusive national governance model. It comprised regional health trusts, municipalities, general practitioners (GPs), patient organizations, labor unions for health care professionals, and governmental bodies [15] (Figure 1). The board aimed to ensure that EHR policy implementation was realized via "a strong national governance model" grounded in consensus and coordination that could solve problems involving different interests; procurement regimes; variations in requirements; local adaptations and adoptions; and different versions, configurations, and EHR technological platforms [4,14]. Industry vendors were not included in the NEGB [4].

In 2016, the government established the Directorate of eHealth, the main purpose of which was to focus efforts on realizing and strengthening the OCOJ vision. The Directorate of eHealth acted as a secretariat for the NEGB [15]. Alongside the NEGB, the government proposed new eHealth regulations (eHealth Act 2019-2020) and a national funding scheme [14,16]. The proposed eHealth Act was retracted from the Norwegian Parliament by the end of 2020 following significant discussion of its implications [16].

The NEGB members were self-regulated actors bringing multiple voices and interests into the process that included the voices of the regional health trusts, the municipalities, the health care professionals, the patient organizations, and governmental bodies. In addition, the Norwegian Association of Local and Regional Authorities (KS) represented Norwegian municipalities in the NEGB (Figure 1).

Under the Norwegian constitution, municipalities are self-governed entities that align with the European Charter of Local Self-Government [17,18]. The municipalities are responsible for primary health care services, including managing agreements with self-governed GPs, while the state owns and operates specialist health care services in regional health care trusts [19-21]. Municipalities, GPs, and regional health trusts are responsible for investing in and implementing their EHR systems. The government and the municipalities meet twice a year under a consultation scheme to discuss policy implications for municipalities regarding the national budget and other policy issues. They have a meeting specifically related to digitalization policy once a year [22].

XSL•FO
RenderX

Figure 1. The Norwegian eHealth Governance Board and the stakeholder groups wanting to be board members and participate in the policy implementation process. EHR: electronic health record.

Theoretical Perspectives: From Government to Governance

The term government refers to the formal institutions of the state that have a monopoly on legitimate hierarchical steering power [23,24]. In the 1970s and 1980s public governance literature, the perspective changed from a "hierarchical government paradigm" to a "horizontal governance paradigm" [23]. In the horizontal governance paradigm, the government "is not the only actor that attempts to influence societal developments, and government interventions are interventions in policy networks, in which power, resource dependency, and strategic behavior are vital elements" [23]. This implies that government interventions must negotiate multiple interests and adjust to stakeholders' actions and reactions [24,25]. Successful government interventions based on inclusive national governance models depend on how the models facilitate alignment and foster inclusion of stakeholders' concerns and top-down government ambitions in a multistakeholder environment [26,27].

These processes, in which the government and stakeholders collaborate to achieve a common goal within an inclusive governance model, may progress through distinct phases [28]. Governance models are constructions characterized by a mix of top-down governance, bottom-up network, and market governance [29]. In one phase, the model is predominantly characterized by an emphasis on the network approach. Conversely, in another phase, both the government and stakeholders may recognize the need for models that incorporate hierarchical structures or market-oriented strategies [29]. Recently, researchers in governance theory have questioned how the governance models must adapt to a global context of increased turbulence through a combination of network, market, and top-down governance [29-31].

We aligned our perspective on governance with "the nonhierarchical process whereby public and private actors and resources are coordinated and given common sense and meaning" [26]. Political goals result from negotiated compromises between the government and stakeholders [23,32,33]. Such negotiations are consequences of the inclusive governance model's ability to address governance dilemmas and balance stakeholder concerns and top-down government ambitions [23,28]. These dilemmas may relate to issues such as participation, transparency, trust, mutual rules, governance processes, and stakeholders' autonomy versus top-down government steering [2,23,27,34]. Governments may create inclusive top-down horizontal governance models to ensure that stakeholders' interests and government policy goals are aligned [3,23]. However, inclusive horizontal governance may also emerge from bottom-up networks based on user needs [2,35].

XSL•FO
RenderX

To establish a robust model for trust management, stakeholders must perceive that their involvement in policy-making processes is meaningful, their concerns are acknowledged and addressed effectively, and they can influence decision-making outcomes [26,27,31]. Participation and transparency are vital for building and maintaining trust in governance models, and stakeholders' participation in public governance models should be voluntary [23,27,36,37]. When stakeholders perceive that the government is using a hierarchical approach characterized by stringent, directive policies, trust levels are likely to diminish. Consequently, concerns may arise about inclusive governance models' legitimacy, prompting questions about their efficacy and ability to accommodate diverse stakeholder perspectives [3,23,24]. This may lead to stakeholders withdrawing from the model or reducing their participation. However, if the government adopts a hands-off strategy, the model may be viewed as a loose network that allows members to cherry-pick policy themes and avoid difficult negotiations [23]. An inclusive horizontal governance model must address this dilemma to ensure stakeholder alignment and policy goal attainment [3,23]. To balance an inclusive horizontal governance model, the government may also increase its trustworthiness through mutually agreed-upon conflict-resolution procedures [3,23].

Through this study on the Norwegian eHealth governance model, we develop knowledge that is valuable for eHealth policy makers and stakeholders in practice. Furthermore, we aspire to stimulate increased multistakeholder debates on governance models in eHealth policy. We also aspire to contribute knowledge to eHealth policy research.

Aims

This study aimed to investigate how the inclusive Norwegian eHealth governance model served as a policy tool to realize the Norwegian EHR policy, as outlined in the vision for OCOJ. Specifically, we analyzed how the governance model addressed public governance dilemmas by asking the following research questions: (1) What characterized national governance in the process of implementing the OCOJ policy in Norway from January 2012 to December 2022? and (2) How did the national governance model address public governance dilemmas by aligning multiple interests during different phases of the policy process in a fragmented field?

Methods

A Case Study With a Longitudinal Approach

This paper is based on a descriptive and interpretative case study conducted to analyze a major national eHealth policy implementation process [38,39]. We decided to perform a document study with a thematic constructivist approach [40]. This technique is commonly used in political science when analyzing policy processes and how they evolve [41]. By including policy documents from the government and consultation response documents we were able to uncover the most prominent policy themes. This approach facilitated a comparison of the government's policy ambitions and the stakeholders' perspectives on the different themes. Each document was treated as an "informant," the same way as documents derived from transcribed qualitative interviews [42]. Analysis was performed between May 1, 2022, and June 1, 2023. Through a thick description of what the government and the stakeholders presented as their ambitions and perspectives, we established a detailed description of their social reality. The aim of applying a thick description may simplify other researchers' assessment of our study's compatibility with their context [40].

This paper provides an in-depth analysis; discusses its empirical findings and concepts, such as horizontal governance, hierarchical steering, and eHealth governance models; and presents analytical and conceptual generalizations [43].

Search Strategy to Identify Empirical Material

To identify relevant documents from the OCOJ process, such as government policy documents and different stakeholders' consultation response documents, we searched various websites to locate publicly available information issued by the Norwegian Parliament, the Ministry of Health and Social Care, and the Directorate of eHealth. We conducted our document search from May 1 to June 30, 2022. The search started on the Directorate of eHealth's website, where we found a chronological list of all documents related to the OCOJ policy process. To identify relevant consultation response documents, we then searched the websites of the NEGB members and 2 non-NEGB umbrella organizations for industry vendors. We extended our search to the websites of the Ministry of Health and Social Care and the Norwegian Parliament, through which we accessed all consultation response documents from official hearings involving stakeholders on our topic of interest. We confirmed our search findings and inclusions by email correspondence with the Ministry of Health and Social Care, the Norwegian Parliament, the Directorate of eHealth, and KS. All documents are included in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

We resolved any disagreements through dialogue and consideration of the research questions. All documents directly related to OCOJ were included, along with those addressing policy issues related to national eHealth governance and consultancy reports commissioned by the government.

We excluded documents not written in Norwegian, policy documents on national health care reforms and general health policy, and documents on policy topics not concerned with horizontal eHealth governance and network governance. Interviews in newspapers, position documents, and policy notes on different stakeholders' websites were not included. Furthermore, such documents were often comments on already-included documents. All included documents are listed in Table S1 in Multimedia Appendix 1.

Policy documents represent the government's future visions. They are political tools to enable policy implementation processes that entail policy discussions, negotiations, and compromises between the government and the stakeholders. The consultation response documents reflect the policy perspectives of various stakeholders, including patients, medical professionals, industry representatives, and municipal authorities.

XSL•FO
RenderX

We have listed the report from the Office of the Auditor General of Norway under policy documents. This document may differ slightly from other policy documents as it aims to present an audit of the OCOJ policy implementation. We have chosen to list it under policy documents because we define the general auditor of Norway as a governmental actor and not as a stakeholder. Table 1 lists all policy documents included in the study, and Table 2 lists all included consultation documents.

Prop 3 L: Proposition to the Parliament for a decision on a legal act.

OCOJ: One Citizen–One Journal.

Table 1. Policy documents included in the study.

Period	Type of document	Author	Policy documents (N=16), n (%)
2012-2013	White Paper Number 9 (2012-2013): "One Citizen–One Journal" (OCOJ) [4]	Ministry of Health and Social Services	1 (6)
2013-2019	Reports on OCOJ [15,44-51]	Directorate of eHealth	9 (56)
2018-2020	Consultancy reports on OCOJ [52,53]	Holte Consulting	2 (12)
2019	eHealth Act, Prop.[a] 65 L (2019-2020) A proposal to Parliament [16]	Ministry of Health and Social Services	1 (6)
2021	Office of the Auditor General of Norway [14]	An investigation conducted by the Office of the Auditor General of Norway on the Ministry of Health and Care Services' governance of the work on OCOJ	1 (6)
2021	Changes in the Patient Journal Act and access to and payment for eHealth solutions and more: Prop 3 L (2021-2022), Innst[b] 47 L (2021-2022), and Lovvedtak 26 (2021-2022) [54].	Ministry of Health and Social Services	1 (6)
2022	KS[c] policy report [55]	The municipal sector's eHealth ambitions	1 (6)

[a]Prop.: Proposition to the Norwegian Parliament for a decision on a legal act.

[b]Innst: A legal act proposal from one committee in the Norwegian Parliament to a plenary decision in the Parliament.

[c]KS: Norwegian Association of Local and Regional Authorities.

Table 2. Consultation response documents included in the study.

Date	Consultation response documents linked to [4,56]	Consultation response documents (N=175), n (%)
November 2012	White Paper Number 9 (2012-2013): OCOJ[a]	5 (3)
April 2019	eHealth Act, Proposal 65 law (2019-2020) Proposals to Parliament	91 (52)
October 2020	eHealth Act hearings in Norwegian Parliament	20 (11)
October 2021	Changes in the Patient Journal Act and access to and payment for eHealth solutions and more: Proposal 3 L (2021-2022), Innst[b] 47 L (2021-2022), and Lovvedtak 26 (2021-2022)	59 (34)

[a]OCOJ: One Citizen–One Journal.

[b]Innst: a legal act proposal from one committee in the Parliament to a plenary decision in the Parliament.

Search Terms

We constructed our search terms from public governance theory and a review of eHealth governance models. Textbox 1 presents an overview of search terms and their source (theory or review).

XSL•FO
RenderX

Textbox 1. Search terms and their source.

Ekeland and Linstad [2]

- Top-down governance model
- Hierarchical governance model
- Bottom-up network governance model
- Bottom-up medical network governance model

Ansell and Torfing [23]

- Inclusive governance model
- Participation

Klijn [25]

- Network governance or horizontal governance

Christensen [34]

- Transparency

Lane and Bachmann [37]

- Trust

Ansell and Torfing [23]

- Governance dilemma

Analysis

We analyzed the collected documents using a thematic document analysis from an analytical perspective based on governance perspectives from political science and a review of eHealth governance models (Textbox 1). The thematic analysis seeks to identify, analyze, organize, describe, and report themes found within a dataset [57]. This explorative thematic study entailed an iterative process of moving back and forth between the data and the theoretical perspective.

The review process was conducted manually by LHL, TTK, and AM. AHH and HB reviewed the final table and provided their input. LHL individually reviewed each included document (PDF files) and applied all search terms from Textbox 1 to each document. Relevant text matching the search terms was extracted and copied into a Microsoft Excel file alongside the corresponding document. Upon completion of this initial data extraction, TTK reviewed the data and engaged in discussions with LHL. Subsequently, LHL synthesized the findings from each document and compiled them into a table (Microsoft Word file). TTK reviewed the table and provided feedback. AM then reviewed both the data and the synthesized findings from each document. LHL and AM discussed and reached a consensus on the text and the level of synthesis. The table of documents is included in Table S1 in Multimedia Appendix 1.

Through this data extraction and subsequent synthesis of data, we were able to elicit the thematic OCOJ governance policy themes that covered eHealth and EHR implementation within municipalities and specialist care. We abstracted thematic issues related to governance, top-down steering, bottom-up governance, horizontal governance, hierarchical steering, network governance, legal requirements, and funding mechanisms. Issues related to trust, the model's legitimacy, stakeholders' autonomy, transparency, and stakeholder participation in the national governance model were also of particular interest.

This data synthesizing process and extraction of policy themes from the OCOJ policy process unfolded the perspectives of the different stakeholders for each theme.

Ethical Considerations

This project was approved by the data protection officer of the University Hospital North Norway on April 13, 2021 (number 02705; Governance of eHealth: models and strategies to obtain health policy goals—Implementation capacity in a multistakeholder environment).

Results

A National Policy Process Characterized by 3 Phases

Overview

We found that the OCOJ policy process unfolded as a consultation process in 3 phases, characterized by distinct empirical turning points (Figure 2). We briefly describe these phases and subsequently categorize the available documents accordingly. Table 3 provides an overview of the stakeholders' main policy concerns related to policy issues during the various phases (Figure 2).

XSL•FO
RenderX

Figure 2. Different phases of the One Citizen–One Journal's policy process. Each phase is a result of the tension and negotiations between the government and the stakeholders.

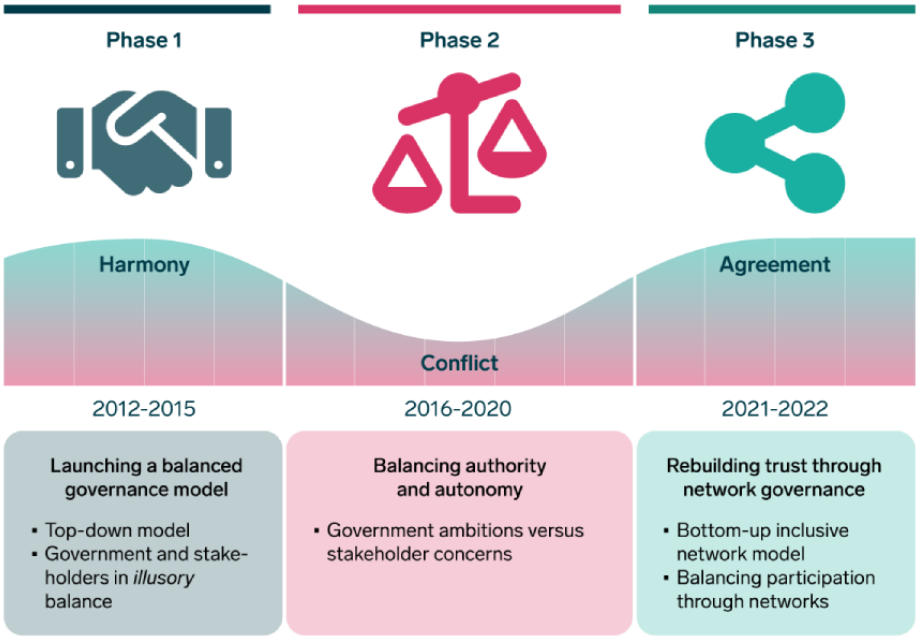

Table 3. Phases, government ambitions, and stakeholders' concerns.

Policy aim stakeholders	Phase 1: launching a balanced governance model	Phase 2: balancing authority and autonomy	Phase 3: rebuilding trust through network governance
Regional health trusts	Initial support. The regional health trusts lost trust in the NEGB[a] when the government planned for the Directorate of eHealth to lead, posing a threat to their autonomy.	Ran their own regional EHR[b] implementation processes independent of the NEGB but promised to coordinate with the NEGB and remain NEGB members	Ran their own regional EHR processes and promised to coordinate with the network governance model
Municipalities or KS[c]	Supported a balanced national governance model and the need for national funding. The municipalities wanted to participate in the process.	Supported the NEGB until they realized that the government was not funding a national EHR. The NEGB did not fulfill its participation and transparency responsibilities. The Directorate of eHealth leadership threatened their autonomy.	Wanted a national governance model based on bottom-up networks that coordinated and governed the municipal "voice" in the market dialogue and in consultations with the government.
Labor unions or organizations of health care professionals	Supported an inclusive national governance model. They expressed a need for national funding and new regulations. They wanted to participate in the national EHR process and the NEGB.	They viewed the implementation of 1 system as too risky. The new EHR systems must be ecosystems that satisfy users' needs. The NEGB did not contribute to an inclusive horizontal governance model that also included medical governance.	The bottom-up network model would be in line with realizing a national ecosystem that is based on health care workers' needs and align with bottom-up medical trust needs.
Patient organizations	Supported a top-down inclusive national governance model to increase patients' access to equal services. Wanted an EHR system that would facilitate easy access to patients' data.	Supported the NEGB and a strong role for the Directorate of eHealth. This would ensure equal access and quality and strengthen patients' self-governance of their health data in national health data systems, such as EHRs.	No data
Industry vendors	Supported a model to create bottom-up processes to implement a national network of regional EHRs.	The national governance model did not include all stakeholders to build trust, depriving stakeholders of the opportunity to realize innovation.	No data

[a]NEGB: National eHealth Governance Board.

[b]EHR: electronic health record.

[c]KS: Norwegian Association of Local and Regional Authorities.

XSL•FO
RenderX

Phase 1: Launching a Balanced Governance Model (January 2012 to November 2015)

The first phase started with widespread anticipation and support for an inclusive horizontal governance model, but latent tension simmered between the government and stakeholders' ideas of what an inclusive horizontal governance model would entail in practice [4]. The tension involved stakeholders' concerns about how the model would balance their autonomy and government ambitions [14,15]. According to the consultation documents, the stakeholders expected the governance model to facilitate participation in a transparent process, along with national funding schemes and legal requirements to support EHR implementation (Textbox 1). The Norwegian Pharmacy Association emphasized the need for both public and private actors' participation and involvement in the project [58]. The Norwegian Medical Association highlighted the importance of a national funding scheme [56].

When the regional health trusts withdrew from the national project (led by the Directorate of eHealth), the process changed significantly, and phase 2 began [15]. The regional health trusts "did not want a common national EHR solution that included both primary and secondary care because this direction pointed toward a big standard solution" [14]. They withdrew from the national OCOJ project and continued their regional EHR implementation processes [14]. Stakeholders perceived sustained engagement in the national EHR implementation process, overseen by the Directorate of eHealth, as posing challenges to the balance between government authority and stakeholders' autonomy. This stemmed from the fact that both the regional health trusts and the Directorate of eHealth reported directly to the Ministry of Health and Social Care. Notably, the ministry, as proprietor of the regional health trusts, wielded overarching authority, while the directorate operated as a ministerial agency devoid of direct formal jurisdiction over the regional health trusts [14] (Table 3).

Phase 2: Balancing Authority and Autonomy (December 2015 to June 2021)

The government had to scale back its OCOJ ambitions (the NEGB and national EHR project) for phase 2 [15], emphasizing that it is "not appropriate to focus on the choice of one concept alternative in the traditional sense, but one direction of development" [15]. Thus, the government's main ambition for phase 2 was to realize a national municipal EHR journal and coordinate the process with regional health trusts to attain OCOJ goals [15].

Although the regional health trusts left the national EHR project, they remained members of the NEGB and continued to inform NEGB members about their regional processes [44]. The general feeling toward the NEGB at the beginning of this phase was positive [14]. However, over time, the latent tension between hierarchical and inclusive horizontal governance became evident [14]. The NEGB's aim was challenged, significantly impacting the goal of realizing an inclusive horizontal governance model based on consensus [4]. The stakeholders became increasingly concerned about the government's choice of policy actions and how the NEGB balanced top-down authority and stakeholders' autonomy [14,45,46,52,59,60]. Both the regional health trusts and municipalities acknowledged that "it is demanding to achieve a consensus-based governance model with different member concerns, experiences, and control lines" [14]. To achieve consensus and increase trust in the model, several stakeholders requested that "official hearings should be used to a greater extent than current practice suggests" [14]. No official hearings to elicit feedback on the policy documents issued by the Directorate of eHealth were held between March 2013 and April 2019 [14,46,59] (Tables 1 and 2).

By the end of phase 2, the government had introduced the eHealth Act proposal [16]. Stakeholders' (eg, regional health trusts, municipalities, and industry organizations) principal concerns focused on how this legislation would formalize the Directorate of eHealth's dominant position and the NEGB's function as the steering mechanism for implementing OCOJ policy.

Patient organizations supported the Directorate of eHealth's dominant role and the NEGB's role as the steering group. They argued that today's EHR systems are too fragmented, and patients' health information would be scattered among different systems without centralized access [61]. They supported the proposed eHealth Act, which, in their view, would function as a tool to improve the digitalization of the domain and strengthen patients' self-governance of their health data [61].

The Ministry of Health and Social Care stated that by the end of phase 2, cooperation between the government and municipalities was facilitated mainly by traditional hierarchical governance structures outside of the NEGB, such as the Consultation Scheme [55].

Phase 2 ended in June 2021 because of the report from the Office of the Auditor General of Norway that spotlighted deficiencies in the Ministry of Health and Social Care's governance of the process dating back to January 2012 [14]. The report emphasized that the Ministry of Health and Social Care "has not effectively fulfilled its responsibility for follow-up, quality assurance, and reporting" [14]. Following this critical report and the municipalities' bypassing of the NEGB, the Ministry of Health and Social Care initiated an evaluation of the governance model [47]. This evaluation resulted in a new inclusive governance model based on a bottom-up network [47].

Phase 3: Rebuilding Trust Through Network Governance (July 2021 to December 2022)

Phase 3 was characterized by government efforts to restore confidence in the inclusive governance model. The emphasis shifted to a governance model that would facilitate network governance, to align eHealth governance networks horizontally and hierarchically, with an emphasis on bottom-up processes [48,55]. The government's focus was primarily on how the revised governance model could be adapted to municipal needs by building on municipal networks [47].

On the basis of the municipalities' concerns, the new model in phase 3 combined bottom-up network governance with hierarchical governance through the consultation scheme [55]. This combination would ensure that contested topics could be escalated from the network model to the consultation scheme.

XSL•FO
RenderX

The network model was accompanied by new procedures to reduce conflict, build trust, and address governance dilemmas [47].

Two Contrasting Views Regarding Stakeholders' Autonomy and Top-Down Government Authority

The second main finding is the tension that developed between the 2 contrasting views regarding stakeholders' autonomy and top-down government authority. Regional health trusts, municipalities, health care professional organizations, and industry actors became increasingly concerned about the model's ability to balance stakeholders' autonomy concerns with top-down government authority. In contrast, patient organizations wanted a hierarchical model to ensure equal access to care and quality of care through coherent digital solutions [14,57].

Furthermore, we found that the government and the municipalities disagreed on whether municipal participation in the national EHR project should be mandatory. According to the municipalities, compulsory participation diminished their autonomy while accentuating the hierarchical aspect of the NEGB, rather than fostering its horizontal inclusivity and cooperation potential [60].

Conversely, patient organizations advocated for a "strong governance model" along a different trajectory [61]. The National Umbrella Organization for Patient Organizations supported the proposed eHealth Act and emphasized that "it is necessary for the Directorate of eHealth to have a clear authority role" [59]. From the patient organizations' perspective, the proposed eHealth Act elucidated and mandated top-down national governance and development requirements to achieve national coherent eHealth solutions [61,62].

Furthermore, our study also found that the patient organizations emphasized the following: "The Norwegian healthcare sector exhibits a complex and fragmented organizational structure. Patients, relatives, and citizens experience firsthand the consequences of inadequate coordination, deficient information flow, and disjointed services and treatment processes. A national governance model and the development of digital services are imperative for patients and citizens, both in terms of patient safety, treatment quality, and the efficient utilization of time and resources" [61]. In the context of a fragmented health care service, ensuring access to up-to-date information is imperative for maintaining a secure patient trajectory [61].

Our findings also show that the municipalities' apprehension regarding the Directorate of eHealth and NEGB's commitment to respecting their autonomy prompted them to circumvent the NEGB as a governance tool for the OCOJ initiative, reverting instead to traditional hierarchical governance through the consultation scheme [14].

In its report, the Office of the Auditor General of Norway claimed that the policy became increasingly challenging for "the actors' autonomy" [14]. The municipalities were concerned that a dominant Directorate of eHealth would impose increased costs for EHR implementation processes [60,63]. However, the municipalities stayed with the process longer than the regional health trusts and supported the NEGB because they expected a national EHR funding scheme [14].

Participation, Transparency, and Trust as Key Aspects of a Legitimate Inclusive Governance Model

Third, our research revealed that the ambiguity surrounding a national funding scheme and the proposed eHealth Act seemed to escalate tensions between the government and various stakeholders, including municipalities, the Norwegian Medical Association, and industry organizations. Consequently, this heightened tension resulted in greater scrutiny of the transparency and legitimacy of the top-down defined inclusive governance model [14,64]. This was evident in the Norwegian Medical Association's concerns about the eHealth Act. The Association asserted that the existing governance model and its decision-making structure had become excessively "bureaucratized," characterized by limited transparency in decision-making processes. Furthermore, the association highlighted a proliferation of issues being raised without full consideration of substantive input [59].

The municipalities and the Norwegian Medical Association increasingly criticized the NEGB's lack of legitimacy as a governance tool, particularly its lack of ability to facilitate participation, transparency, and trust through an inclusive horizontal governance model [59,60]. Hence, the municipalities circumvented the NEGB based on arguments related to transparency and trust [14].

The lack of official hearings on the reports produced by the Directorate of eHealth contributed to a decrease in their trust in the NEGB [14]. The Norwegian Medical Association stated that the model engendered a perception of "mock processes for numerous individuals within the healthcare sector, contributing to an eHealth environment marked by reduced trust among healthcare personnel" [59]. The association viewed bottom-up network governance as a more suitable way of aligning medical concerns and technological developments in an inclusive model [60,65]. Moreover, our investigation uncovered that the regional health trust of Northern Norway perceived the NEGB as a mechanism wielded by the Directorate of eHealth to justify their top-down policy approach, rather than as a legitimate inclusive horizontal governance model [14].

We found that in its evaluation report, the Directorate of eHealth emphasized that the future governance model (phase 3) "is an advisory arena where the sector gives its recommendations on strategic eHealth issues" [47]. It also emphasized that the model should include all key stakeholders [48]. The bottom-up, inclusive network governance model (in phase 3) combined horizontal and hierarchical governance through the consultation scheme [47]. According to the Directorate of eHealth, the hierarchical aspect of the new inclusive network governance model would bolster its resilience in cases of disputes between stakeholders and the government [47].

Our study showed that in the autumn of 2020, KS assumed leadership and issued a mandate for the national EHR project following consultations with the Directorate of eHealth, the Ministry of Health and Social Care, and participating municipalities [14,55]. The regional health trusts continued their

XSL•FO
RenderX

regional implementation processes and promised to align them with the national municipal EHR project [14].

Discussion

Principal Findings

The empirical analysis revealed 3 main findings. First, the national policy process evolved across 3 phases, with changes in stakeholder inclusion, participation in the process, and perceived influence on the decision-making process characterizing the transitions. The process commenced with the initial endorsement of an inclusive horizontal governance model, which subsequently was accompanied by stakeholders' reduced trust in the model. This erosion of trust led to heightened polarization between the government and a pivotal group of stakeholders. Ultimately, the culmination of these developments prompted a re-evaluation of the governance model in favor of a more bottom-up, inclusive network approach. Our second primary finding was the presence of 2 contrasting perspectives regarding stakeholders' autonomy and top-down government authority related to EHR implementation. Regional health care trusts, municipalities, health care professional organizations, and industry stakeholders expressed growing apprehension about the top-down defined governance model's capacity to align stakeholder concerns regarding their autonomy to implement their EHRs. Conversely, patient organizations welcomed a top-down hierarchical model to guarantee equitable access, and quality of care through coherent national digital solutions. Our third finding pertained to stakeholder perceptions of government insensitivity concerning participation in the inclusive governance model, the transparency of the model's decision structure, and its ability to build trust between government and stakeholders. The distrust that emerged between the government and numerous stakeholders posed a challenge to the top-down defined, inclusive governance model. Tension and criticism, fueled by perceived deficiencies in participation, transparency, and trust, made the government change its approach to a bottom-up network model that integrated both inclusive horizontal and hierarchical decision-making.

A Policy Process Characterized by 3 Phases

This first finding relates to the process unfolding across 3 phases and how this development characterized the process and influenced the inclusive governance model's ability to address governance dilemmas. The turning point for each phase resulted from stakeholders' perceptions of participation in and influence on the decision-making process and the governance model's ability to align these concerns with government ambitions. The phases were characterized by increased stakeholder concerns about lack of participation, transparency, and trust in the OCOJ decision-making process (Figure 2).

The transitions between phases are in line with the findings from New Zealand and Denmark [8,10,11]. A recent review suggested that transitions between phases could be described as "negotiation processes" between government and stakeholders [2]. However, international findings neither delineate the specific stakeholders involved in various phases nor elucidate how governance models in Denmark and New Zealand addressed governance dilemmas regarding stakeholders' participation to align government objectives with stakeholders' concerns. In Denmark, a decision was made to adopt a bottom-up municipal network governance approach aimed at fostering dialogue among municipalities. This approach was taken to ensure an equitable process across all municipalities [11]. The government opted for a hands-off approach, diverging from the hands-on approach observed during earlier phases of national EHR projects [11]. This development resembled the Norwegian process, in which municipalities resumed responsibility for the inclusive network governance model following a period characterized by a top-down government approach [11,55].

It is a commonplace observation in public governance research that policy processes unfold in distinct phases [23,28,29]. When policy themes emerge as hot topics, they trigger the mobilization of stakeholders and heighten their demands for participation in the policy process to safeguard their interests [3,23,31]. The tension in policy discussions and negotiations influences how governance models evolve and lead to new phases [23,28,29].

Given that our empirical findings, theoretical perspectives, and international experiences illustrated that policy processes, such as the OCOJ initiative, typically progress through phases, it is reasonable to expect that stakeholders will articulate demands for a more robust and inclusive governance model in the Norwegian eHealth domain in the future, in line with the findings from Denmark and New Zealand [10,11].

Two Contrasting Views Regarding Stakeholders' Autonomy and Top-Down Government Authority

The second finding pertains to the 2 divergent stakeholder views regarding how the inclusive governance model would reconcile stakeholders' autonomy with top-down government authority [14,59]. Patient organizations endorsed a governance model that curtailed regional health trusts' and municipalities' autonomy to implement EHRs [61], whereas the regional health trusts, municipalities, and the Norwegian Medical Association supported a bottom-up network model to protect their autonomy and their own EHR implementation processes [14,48] (Figure 1).

We found similar contrasting views in a study from the United Kingdom on radiology networks and a national eHealth project. Radiology stakeholders in the United Kingdom argued that national scaling must balance their medical autonomy against top-down government authority regarding the national standardization of services [66]. They anticipated that the new national EHR project—the National Program for Information in the National Health Service would build on their bottom-up governed medical network [67]. The government did not meet these expectations, leaving radiologists to question how their medical autonomy was met by a top-down hierarchical approach leading to reduced medical trust in national policy [67].

Studies from Denmark are in line with our findings [1,11]. In these studies, the 2 contrasting views were represented by the government and the regions. Evidence revealed government apprehension regarding potential divergence in regional trajectories of EHR implementation, which could result in disparate levels of care access and quality. Consequently, the Danish government opted to use a top-down approach in

XSL•FO
RenderX

response to these concerns [2,11]. During different phases, regions of Denmark sought a model that would support their greater autonomy, while at other junctures, the government found it imperative to bolster its top-down governance to align regional and national EHR policies in an inclusive model [11]. Diverging from our comprehensive analysis, which meticulously explored the multifaceted viewpoints of stakeholders, the 2 Danish studies predominantly focus on elucidating the perspectives solely of the regions and the government [1,2,11].

A third study on the Danish eHealth policy process had a similar focus to our study: exploring the national eHealth policy process in a fragmented system. A comprehensive examination of eHealth policy within the Danish context reveals the paramount significance attributed to the national eHealth policy among diverse stakeholders. The dynamics characterizing the interplay between these stakeholders and governmental entities are akin to a strategic power struggle, underscored by varying degrees of power and influence wielded by each stakeholder group [5]. In contrast to our study, this study fails to elucidate whether disparate stakeholder groups harbor divergent perspectives that may be antithetical and how these potential divergent views may have influenced the Danish inclusive governance model [5].

Our study shows that the patient organizations supported a governance model that curtailed the autonomy of regional health trusts and municipalities to ensure easy access to and overview of their health data [59,62]. There is little knowledge of how patient organizations, based on their members' production of health data, collectively mobilize to exert influence upon the national governance model [2]. A study on mobile apps and sleep monitoring showed that the growth of mobile apps designed to facilitate health and wellness self-governance has generated large amounts of health data, which individuals subsequently may use as a foundation for self-governance to guide their actions [9]. This study does not elaborate on how patient's ownership of health data may represent a bargaining power in future governance models. However, the study questions how this development may change the patient's role and influence health care systems [9]. This is in line with the health data governance principles, which aim to create both regulations and international principles that empower the patient's ownership of their health data in health systems built on human rights [68].

In the "government paradigm," the government is not the only actor that attempts to influence policy processes and must find ways to govern through stakeholder networks' dynamics [3,23-25]. In the realm of public governance research, it is commonly observed that governments use various top-down or bottom-up governance strategies to engage in negotiations with heterogeneous stakeholders holding divergent perspectives to cultivate inclusive horizontal governance models [23-25]. In multistakeholder fields, such as eHealth, participating in negotiations and governance models is optional [23,25,28]. If the stakeholders perceive that their goal of preserving their autonomy is overruled or ignored, they may circumvent the governance model or withdraw to their autonomous path to realize their EHR policies. This is in line with our findings. The regional health trusts withdrew from the process, the municipalities circumvented the NEGB, and the Norwegian Medical Association aligned themselves with the KS to increase their medical autonomy, leaving the patient organizations as the strongest supporters of a top-down approach [14].

Our empirical findings, theoretical perspectives, and international experiences illustrate that inclusive governance models must accommodate and align varying concerns among stakeholder groups, for example, regional health trusts, municipalities, health professions, and patient groups regions, to become legitimate, inclusive governance models. Furthermore, our findings elucidate personal health data's evolving nature and their prospective amalgamation with EHR data, engendering novel dynamics at the nexus of bottom-up and top-down governance mechanisms, thereby shaping future inclusive governance models. A potential implication of this progression is the augmentation of citizens' perspectives on inclusive governance models in the eHealth realm.

Participation, Transparency, and Trust as Key Aspects of a Legitimate Inclusive Governance Model

Our third finding pertains to how the government's insensitivity to stakeholder participation, the governance model's lack of decision transparency, and decreasing trust between the government and stakeholder groups challenged the inclusive horizontal governance model and its legitimacy. The government changed its approach to an inclusive bottom-up network model that combined horizontal and hierarchical decision-making [14].

The previously mentioned study from New Zealand showed how the government implemented a national inclusive governance model called the National Health IT Board to facilitate national data exchange standards and increased adoption of eHealth [10]. The establishment of this board directly, subordinate to the National Health Board, conferred inclusivity upon the model, encompassing all pertinent stakeholders. This proximity to the national health policy level was deliberately chosen to enhance stakeholder trust in the model as a legitimate policy instrument [10]. This aligns with the findings of our study. The Norwegian government endeavored to establish the NEGB as an inclusive board situated closely to the national health policy level, with the overarching objective of fostering trust in the model and affirming its legitimacy [4].

The second study from Denmark showed that the Danish government in collaboration with regions took a similar approach [11]. Their inclusive governance model aimed to facilitate horizontal dialogue among the regions through a governmental hands-off approach [11]. This model exhibits comparable characteristics to the Norwegian bottom-up inclusive network governance model observed in phase 3, as identified in our examination of the OCOJ policy process. The study from Denmark highlights how this dialogue may increase the transparency on the implementation of EHR standards at both regional, state, and European Union levels [11].

In our examination of the OCOJ process, there was no exploration or discourse regarding the conceptualization of the NEGB as a governance model intended to foster network collaboration with other European countries to augment

XSL•FO
RenderX

participation and transparency concerning various transborder eHealth policy matters [15]. The Directorate of eHealth investigated how other countries operationalized their eHealth policy [15]. However, this investigation was focused on understanding the national models, rather than evaluating their potential application as transborder policy instruments in turbulent times [15,31]. The Directorate of eHealth's work on the OCOJ reports was conducted before the COVID-19 crisis [15,46]. Consequently, their focus on the realization of the NEGB did not account for external turbulent disruptions associated with pandemics and conflicts.

Our study showed that the government aimed to increase the medical trust in the NEGB by including the Norwegian Medical Association as a member of the NEGB. A study from New Zealand showed similar findings [8]. The government of New Zealand aimed at increasing the medical trust in the inclusive national governance model by including medical stakeholders in the model [8]. This finding is in line with our findings. At the outset of phase 1, the government did not see the need to include the Norwegian Medical Association as a member of the NEGB. However, this changed by the end of phase 2 when both the Norwegian Medical Association and the Norwegian Nursing Association became members of the NEGB.

A study from Australia pointed out how the implementation of telehealth systems may facilitate immediate communication between patients and health care providers and increase interprofessional trust [66]. Furthermore, the study pointed out how scaling telehealth in Australia might increase the quality of health care and contribute to an inclusive governance model with increased uptake of evidence-based care [66]. This is in line with our findings on medical trust in the NEGB. The Norwegian Medical Association argued that its membership in the NEGB and participation would ensure the inclusion of the medical perspective and thereby increase health personnel's trust in the NEGB and the OCOJ policy [56,60].

There is little knowledge of governance models and transparency in eHealth decision processes [2]. However, a study identified transparency as a policy issue related to health information and the sale of pharmaceutical products on the internet [69]. The study questions how the quality assurance of the health information related to the sale of pharmaceutical products on the internet is governed. The authors suggested that increased transparency may be achieved through a global internet governance model that involves the World Health Organization and other relevant stakeholders [69].

The review by Ekeland and Linstad [2] questioned how future national governance models in eHealth may balance global market governance and national top-down or bottom-up approaches. Our investigation and analysis on the Norwegian governance model in eHealth did not find that the Directorate of eHealth planned to explicitly use the NEGB as a tool to balance global market governance and national top-down or bottom-up approaches. We found that the Directorate of eHealth after assessing the market concluded that the only vendors suited for the Norwegian system were global market actors.

On the basis of public governance theory, we suggested that constructions of legitimate governance models require balancing top-down authority and bottom-up autonomy to create trust [23,25]. We also illustrated that inclusive governance models in EHR policy processes must facilitate participation by different stakeholder groups and include transparent policy decision processes to cultivate trust, particularly medical trust [2,3,23]. A possible implication from these findings is that future inclusive horizontal governance models must address dilemmas regarding participation and transparency to build trust by combining network and top-down governance [3,25,27]. Achieving success in establishing an inclusive governance model may entail the incorporation of structured procedures and regulations aimed at addressing misperceptions and enhancing resilience in governing stakeholder trust [31,47,55]. Future developments in health data governance may influence how governments rethink and choose to balance top-down and bottom-up strategies with market-oriented governance approaches to create robust inclusive governance models [29-31].

Implications and Further Research

The research findings in this case study align with reports from other health systems, underscoring the necessity for governmental attentiveness to policy issues such as participation, transparency, and trust to realize national inclusive horizontal governance models. Additional insights from case studies conducted across diverse contexts are imperative to advance our understanding. Further research is needed to elucidate how different contextual conditions influence governments' abilities to balance authority and autonomy in a policy field such as eHealth with its fragmented decision authority and plural interests.

Strengths and Limitations

The research team comprises experienced professionals in eHealth and public governance, drawing on expertise from political science, sociology, medical science, and diverse roles in eHealth research. However, field insights provide a unique starting point for our study. We are aware that our roles as experts may bring biases to our research. Long experience may influence our perception and "worldview" [70]. In contrast, policy studies require a broad overview of the field at the outset of the study [41]. We believe that our awareness of our positionality in the field and the interdisciplinary composition of our research team balance our views and potential biases.

A notable strength of our study lies in its comprehensive incorporation of all publicly available government policy documents and stakeholder consultation materials, which addressed policy proposals and considerations. This enabled us to discern the active participation of stakeholders across different phases and to evaluate how their perspectives influenced the government's capacity to establish an inclusive horizontal governance model.

Determining the optimal search strategy to identify pertinent documents poses a challenge unless researchers possess a comprehensive understanding and overview of the field from the outset. This aspect may have represented a weakness attributable to the researchers' focus, as the in-depth longitudinal case study exclusively examined the Norwegian context,

XSL•FO
RenderX

potentially limiting its generalizability to the processes and challenges encountered by other countries when implementing EHR policies. Nevertheless, this study may offer valuable insights as a meticulously documented case, providing lessons learned and actionable recommendations for countries facing analogous circumstances.

Conclusions

We conclude that Norway's OCOJ policy trajectory was characterized by a process that unfolded across 3 distinct phases. Furthermore, the process was characterized by 2 contrasting views regarding stakeholders' autonomy to govern their EHR implementation processes and the need for top-down government authority to create a national journal. Regional health trusts, municipalities, health care professional organizations, and industry actors became increasingly concerned about the model's ability to balance stakeholders' autonomy and top-down government authority. In contrast, patient organizations wanted an inclusive hierarchical model to ensure equal access to care and quality of care through coherent digital solutions. Finally, the policy process was characterized by diminishing trust in the inclusive governance model.

Our study found that measures were taken to include a variety of voices in the NEGB through consultations. However, the different phases illustrated that participation, transparency, and trust dilemmas can occur in a top-down defined inclusive governance model. Such dilemmas require ongoing addressing through vigilance and responsiveness from governmental entities. The responses made the governance model evolve from being a top-down defined inclusive model into a bottom-up network model.

Acknowledgments

The authors would like to thank the Governance and Democracy research group at the Arctic University of Norway for reviewing this paper and sharing their comments. The authors would also like to thank Professor Kari Kværner at Oslo University Hospital and Oslo Business School, Personal eHealth Director Siri Bjørvig at the Norwegian Center for eHealth Research, and Professor Anne Granstrøm Ekeland for their fruitful discussions on this study's design and results. The authors would like to express their gratitude to senior advisor Jarl-Stian Olsen and senior researcher Meghan Corin Bradway for their assistance with the figures.

Data Availability

The data used in this paper are publicly available on the websites of the Norwegian Parliament [71] and the Directorate of eHealth [72].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Policy and consultation response documents, phases, and main contributions.
[DOCX File , 81 KB - jmir_v26i1e59717_app1.docx]

References

1. Kierkegaard P. Governance structures impact on eHealth. Health Policy Technol 2015 Mar;4(1):39-46. [doi: 10.1016/j.hlpt.2014.10.016]
2. Ekeland AG, Linstad LH. Elaborating models of eHealth governance: qualitative systematic review. J Med Internet Res 2020 Oct 28;22(10):e17214 [FREE Full text] [doi: 10.2196/17214] [Medline: 33112247]
3. Ansell C, Sørensen E, Torfing J. The democratic quality of co-creation: a theoretical exploration. Public Policy Adm 2023 Apr 20;39(2):149-170. [doi: 10.1177/09520767231170715]
4. Én innbygger - én journal. Digitale tjenester i helse- og omsorgssektoren. Stortinget. 2012. URL: https://www.stortinget.no/no/Saker-og-publikasjoner/Saker/Sak/?p=55703 [accessed 2024-10-29]
5. Benneke M, Charalambous G, Jensen O, Andrioti D. Policy analysis for eHealth in Denmark. Int J Community Fam Med 2018 May 03;3:1-6. [doi: 10.15344/2456-3498/2018/140]
6. Fra stykkevis til helt - en sammenhengende helsetjeneste. Regjeringen.no. 2005. URL: https://www.regjeringen.no/no/dokumenter/nou-2005-03/id152579/ [accessed 2024-10-29]
7. Smaradottir BF, Severinsen GH, Steinsbekk A, Berntsen GK. User-centred design of a digital care plan for patients and professionals in cross-organisational teams. Stud Health Technol Inform 2021 May 27;281:901-905. [doi: 10.3233/SHTI210309] [Medline: 34042804]
8. Atalag K. Using a single content model for eHealth interoperability and secondary use. Stud Health Technol Inform 2013;193:282-296. [Medline: 24018523]
9. Williams SJ, Coveney C, Meadows R. 'M-apping' sleep? Trends and transformations in the digital age. Sociol Health Illn 2015 Sep;37(7):1039-1054. [doi: 10.1111/1467-9566.12283] [Medline: 26182959]

XSL•FO
RenderX

10. Park YT, Atalag K. Current national approach to healthcare ICT standardization: focus on progress in New Zealand. Healthc Inform Res 2015 Jul;21(3):144-151 [FREE Full text] [doi: 10.4258/hir.2015.21.3.144] [Medline: 26279950]
11. Kierkegaard P. Interoperability after deployment: persistent challenges and regional strategies in Denmark. Int J Qual Health Care 2015 Apr;27(2):147-153. [doi: 10.1093/intqhc/mzv009] [Medline: 25699549]
12. Asthana S, Jones R, Sheaff R. Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors. BMC Health Serv Res 2019 Dec 21;19(1):984 [FREE Full text] [doi: 10.1186/s12913-019-4790-x] [Medline: 31864370]
13. Garmann-Johnsen NF, Eikebrokk TR. Dynamic capabilities in e-health innovation: implications for policies. Health Policy Technol 2017 Sep;6(3):292-301 [FREE Full text] [doi: 10.1016/j.hlpt.2017.02.003]
14. Undersøkelser av IT-satsingen Én innbygger — én journal; styring og anskaffelser. Riksrevisjonen. URL: https://www.riksrevisjonen.no/rapporter-mappe/no-2020-2021/undersokelser-av-en-innbygger---en-journal-styring-og-anskaffelser/ [accessed 2024-10-29]
15. Utredning av «en innbygger – én journal». Riksrevisjonen. 2014 Sep. URL: https://www.regjeringen.no/contentassets/355890dd2872413b838066702dcdad88/ikt_utfordringsbilde_helse_omsorgssektoren.pdf [accessed 2024-10-29]
16. Prop. 65 L (2019–2020): lov om e-helse (e-helseloven). Regjeringen.no. 2019. URL: https://www.regjeringen.no/no/dokumenter/prop.-65-l-20192020/id2696053/ [accessed 2024-10-29]
17. Knutsen O. The Nordic Models in Political Science: Challenged, But Still Viable?. Bergen, Norway: Fagbokforlaget; 2017.
18. Pejanovic M. Local self-government: a must for democracy, civil society and EU integration. In: Fischer M, editor. Peacebuilding and Civil Society in Bosnia-Herzegovina. Ten Years after Dayton. Münster, Germany: Lit-Verlag; 2006.
19. Laegreid P, Opedal S, Stigen IM. The Norwegian hospital reform: balancing political control and enterprise autonomy. J Health Polit Policy Law 2005 Dec;30(6):1027-1064. [doi: 10.1215/03616878-30-6-1027] [Medline: 16481307]
20. Tjerbo T. The politics of local hospital reform: a case study of hospital reorganization following the 2002 Norwegian hospital reform. BMC Health Serv Res 2009 Nov 20;9:212 [FREE Full text] [doi: 10.1186/1472-6963-9-212] [Medline: 19930553]
21. Opedal S, Rommetvedt H. From politics to management – or more politics? Public Manag Rev 2010 Mar 03;12(2):191-212. [doi: 10.1080/14719031003616115]
22. Consultations between Central Government and KS. Regjeringen.no. 2021 Aug 10. URL: https://www.regjeringen.no/en/topics/kommuner-og-regioner/municipal-law-and-municipal-organisation/forholdet-kommune-stat/consultations-between-central-government/id544787/ [accessed 2024-10-30]
23. Ansell CK, Torfing J. Handbook on Theories of Governance. Cheltenham, UK: Edward Elgar Publishing; 2016.
24. Jessop B. Capitalism and its future: remarks on regulation, government and governance. Rev Int Polit Econ 1997;4(3):561-581. [doi: 10.1080/096922997347751]
25. Klijn E. Governance and governance networks in Europe. Public Manag Rev 2008 Aug 15;10(4):505-525. [doi: 10.1080/14719030802263954]
26. Røiseland A, Vabo SI. Governance på norsk. Samstyring som empirisk og analytisk fenomen. Norsk Statsvitenskapelig Tidsskrift 2008 Jul 23;24(1-2):86-107. [doi: 10.18261/issn1504-2936-2008-01-02-05]
27. Osborne SP. The new public governance? Public Manag Rev 2006;8(3):377-387. [doi: 10.1080/14719030600853022]
28. Morse RS, Stephens JB. Teaching collaborative governance: phases, competencies, and case-based learning. J Public Aff Educ 2018 Apr 18;18(3):565-583. [doi: 10.1080/15236803.2012.12001700]
29. Bouckaert G, Im T. Neo weberian state: from theory to practice? J Policy Stud 2024;39(2):1-3. [doi: 10.52372/jps39201]
30. Torfing J. Rethinking Public Governance. Cheltenham, UK: Edward Elgar Publishing; 2023.
31. Ansell C, Sorensen E, Torfing J, Trondal J. Robust Governance in Turbulent Times. Cambridge, UK: Cambridge University Press; 2024.
32. Egeberg M, Trondal J. An Organizational Approach to Public Governance: Understanding and Design. Oxford, UK: Oxford University Press; 2018.
33. Stoker G. Governance as theory: five propositions. Int Soc Sci J 2019 Mar 19;68(227-228):15-24. [doi: 10.1111/issj.12189]
34. Christensen T. Post-NPM and changing public governance. Meiji J Polit Sci Econ 2012;1:1-11 [FREE Full text]
35. Koppenjan E, Klijn EH. Governance Networks in the Public Sector. Milton Park, UK: Taylor & Francis; 2015.
36. Braithwaite V, Levi M. Trust and Governance. New York, NY: Russell Sage Foundation; 1998.
37. Lane C, Bachmann R. Trust Within and Between Organizations: Conceptual Issues and Empirical Applications. Oxford, UK: Oxford University Press; 1998.
38. Rossman GB, Rallis SF. An Introduction to Qualitative Research: Learning in the Field. Thousand Oaks, CA: SAGE Publications; 2016.
39. Walsham G. Interpretive case studies in IS research: nature and method. Eur J Inf Syst 2017 Dec 19;4(2):74-81. [doi: 10.1057/ejis.1995.9]
40. Appleton JV, King L. Journeying from the philosophical contemplation of constructivism to the methodological pragmatics of health services research. J Adv Nurs 2002 Dec;40(6):641-648. [doi: 10.1046/j.1365-2648.2002.02424.x] [Medline: 12473042]
41. Brinkmann S, Tanggaard L. Kvalitative Metoder: Empiri og Teoriutvikling. Oslo, Norway: Gyldendal Akademisk; 2012.

XSL•FO
RenderX

42. Basit T. Manual or electronic? The role of coding in qualitative data analysis. Educ Res 2010 Dec 02;45(2):143-154. [doi: 10.1080/0013188032000133548]
43. Yin RK. Validity and generalization in future case study evaluations. Evaluation 2013 Jul 10;19(3):321-332. [doi: 10.1177/1356389013497081]
44. Én innbygger – én journal. ehelse. URL: https://www.ehelse.no/strategi/en-innbygger-en-journal [accessed 2024-10-30]
45. Veikart for realiseringen av målbildet for Én innbygger – én journal. Direktoratet for e-helse. 2018 Jan 12. URL: https://tinyurl.com/bdeerw35 [accessed 2024-10-30]
46. Konseptvalgutredning: nasjonal løsning for kommunal helse- og omsorgstjeneste. Direktoratet for e-helse. URL: https://tinyurl.com/mtdn5fcz [accessed 2024-10-30]
47. Videreutvikling av nasjonal styringsmodell for e-helse: evaluering og anbefalinger. Direktoratet for e-helse. 2022 Jan. URL: https://tinyurl.com/ms55hufu [accessed 2024-10-30]
48. Nasjonal styringsmodell for e-helse. ehelse. URL: https://www.ehelse.no/horinger/horing-nasjonal-styringsmodell-for-e-helse [accessed 2024-10-30]
49. Nasjonal e-helsestrategi 2017–2022. Direktoratet for e-helse. 2019. URL: https://tinyurl.com/e6e8pyu3 [accessed 2024-11-19]
50. Sentralt styringsdokument: Akson: Helhetlig samhandling og felles kommunal journalløsning. Direktoratet for e-helse. 2020. URL: https://tinyurl.com/ydzj2rr6 [accessed 2024-11-19]
51. Veikart for utvikling og innføring av nasjonale e-helseløsninger 2021 - 2025. Direktoratet for e-helse. 2020 Nov. URL: https://tinyurl.com/mw6f9926 [accessed 2024-11-19]
52. Kvalitetsrapport: KS1 av nasjonal løsning for kommunal helse- og omsorgstjeneste. Utarbeidet for Finansdepartementet og Helse- Og Omsorgsdepartementet. 2018 Dec 20. URL: https://www.ntnu.no/documents/1261860271/1261975586/KS1+av+nasjonal+l%C3%B8sning+for+kommunal+helse-+og+omsorgstjeneste.pdf/556c0998-4fd8-4b7e-9b2e-9323bf62b9c6?version=1.0 [accessed 2024-10-30]
53. Kvalitetsrapport KS2 av Akson: helhetlig samhandling og felles kommunal journal. Utarbeidet for Finansdepartementet og Helse-og omsorgsdepartementet. 2020. URL: https://www.regjeringen.no/contentassets/697dd17c89d24b1890d8eb3c511942f7/rapport-ks2-akson.pdf [accessed 2024-11-19]
54. Innstilling fra helse- og omsorgskomiteen om Endringer i pasientjournalloven (tilgjengeliggjøring av og betaling for nasjonale e-helseløsninger m.m.). Stortinget. URL: https://www.stortinget.no/no/Saker-og-publikasjoner/Publikasjoner/Innstillinger/Stortinget/2021-2022/inns-202122-047l/?all=true [accessed 2024-11-19]
55. Kommunal sektors ambisjoner på e-helseområdet. Kommunesektorens Organisasjon. URL: https://www.ks.no/fagomrader/digitalisering/felleslosninger/digitalisering-i-helse-og-omsorgsektoren-e-helse/kommunal-sektors-ambisjoner-pa-e-helseomradet/ [accessed 2024-10-30]
56. St. Meld. 9 (2012-2013) Én innbygger - én journal. Legeforeningen. 2013 Jan 22. URL: https://www.legeforeningen.no/hoeringer/stortinget/2013/5114/hoeringsuttalelse/ [accessed 2024-11-18]
57. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 02;16:1-13. [doi: 10.1177/1609406917733847]
58. Høringsuttalelser og -innspill. Apotekforeningen. URL: https://www.apotek.no/dokumentarkiv/h%C3%B8ringer [accessed 2024-11-18]
59. Videokonferansehøring: lov om e-helse (e-helseloven). Stortinget. 2020. URL: https://www.stortinget.no/no/Hva-skjer-pa-Stortinget/Horing/horing/?h=10004172 [accessed 2024-10-30]
60. Felles uttalelse: "en innbygger - en journal". Kommunesektorens Organisasjon and Den Norske Legeforening. 2019 Feb 15. URL: https://www.ks.no/globalassets/Felles-uttalelse-med-Legeforeningen.pdf [accessed 2024-11-18]
61. Høringsinnspill fra Funksjonshemmedes Fellesorganisasjon (FFO). Stortinget. 2020 Sep 28. URL: https://www.stortinget.no/no/Hva-skjer-pa-Stortinget/Horing/horingsinnspill/?dnid=12318&h=10004172 [accessed 2024-10-30]
62. Én innbygger - én journal. Digitale tjenester i helse- og omsorgssektoren. Stortinget. URL: https://www.stortinget.no/no/Saker-og-publikasjoner/Saker/Sak/?p=55703 [accessed 2024-10-30]
63. Høringsuttalelse En innbygger ? en journal. The Norwegian Association of Local and Regional Authorities (KS). 2013. URL: https://www.stortinget.no/no/Saker-og-publikasjoner/Saker/Sak/?p=55703 [accessed 2024-10-30]
64. Høring ny e-helselov med endring av forskrift - KS' høringsttalelse. Kommunesektorens Organisasjon. 2019 Dec 13. URL: https://www.regjeringen.no/contentassets/2cc94248f15b4838b469a05190941cb1/kommunesektorens-interesse--og-arbeidsgiverorganisasjon.pdf?uid=Kommunesektorens_interesse-_og_arbeidsgiverorganisasjon [accessed 2024-10-30]
65. Høring - ny e-helselov og endringer i IKT-standardforskriften. Regjeringen.no. 2019 Oct 28. URL: https://www.regjeringen.no/no/dokumenter/horing---ny-e-helselov-og-endringer-i-ikt-standardforskriften/id2675404/?showSvar=true&consterm=&page=1&isFilterOpen=true [accessed 2024-10-30]
66. Wade VA, Eliott JA, Hiller JE. A qualitative study of ethical, medico-legal and clinical governance matters in Australian telehealth services. J Telemed Telecare 2012 Mar;18(2):109-114. [doi: 10.1258/jtt.2011.110808] [Medline: 22267306]
67. Crocker M, Cato-Addison WB, Pushpananthan S, Jones TL, Anderson J, Bell BA. Patient safety and image transfer between referring hospitals and neuroscience centres: could we do better? Br J Neurosurg 2010 Aug;24(4):391-395. [doi: 10.3109/02688697.2010.508847] [Medline: 20726748]

XSL•FO
RenderX

68. The principles: health data governance principles. Health Data Governance Principles. URL: https://healthdataprinciples.org/principles [accessed 2024-10-30]
69. Mackey TK, Eysenbach G, Liang BA, Kohler JC, Geissbuhler A, Attaran A. A call for a moratorium on the .health generic top-level domain: preventing the commercialization and exclusive control of online health information. Global Health 2014 Sep 26;10(1):62 [FREE Full text] [doi: 10.1186/s12992-014-0062-z] [Medline: 25283176]
70. Darwin Holmes AG. Researcher positionality - a consideration of its influence and place in qualitative research - a new researcher guide. Int J Educ 2020 Sep 01;8(4):1-10. [doi: 10.34293/education.v8i4.3232]
71. Stortinget homepage. Stortinget. URL: https://www.stortinget.no/ [accessed 2024-11-19]
72. e-helse homepage. e-helse. URL: https://www.ehelse.no/ [accessed 2024-11-19]

Abbreviations

EHR: electronic health record
GP: general practitioner
KS: Norwegian Association of Local and Regional Authorities
NEGB: National eHealth Governance Board
OCOJ: One Citizen–One Journal

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by A Garcia, JF Menard; comments to author 22.06.24; revised version received 05.09.24; accepted 08.10.24; published 04.12.24.

Please cite as:
Linstad LH, Bjørnå H, Moen A, Tunby Kristiansen T, Hansen AH
Investigating the Norwegian eHealth Governance Model: Document Study
J Med Internet Res 2024;26:e59717
URL: https://www.jmir.org/2024/1/e59717
doi:10.2196/59717
PMID:

©Line Helen Linstad, Hilde Bjørnå, Anne Moen, Truls Tunby Kristiansen, Anne Helen Hansen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multifactor Quality and Safety Analysis of Semaglutide Products Sold by Online Sellers Without a Prescription: Market Surveillance, Content Analysis, and Product Purchase Evaluation Study

Amir Reza Ashraf[1], PharmD, PhD; Tim Ken Mackey[2], MAS, PhD; Róbert György Vida[1], PharmD, PhD; Győző Kulcsár[3], MSc, PhD; János Schmidt[4], MA; Orsolya Balázs[1], PharmD; Bálint Márk Domián[1], PharmD; Jiawei Li[5], MS; Ibolya Csákó[6], PharmD; András Fittler[1], PharmD, PhD

[1]Department of Pharmaceutics, Faculty of Pharmacy, University of Pécs, Pécs, Hungary
[2]Global Health Program, Department of Anthropology, University of California San Diego, La Jolla, CA, United States
[3]Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Pécs, Pécs, Hungary
[4]Institute of Biochemistry and Medical Chemistry, Medical School, University of Pécs, Pécs, Hungary
[5]S-3 Research, San Diego, CA, United States
[6]Criminal Department, Criminal Directorate, Hungarian National Police Headquarters, Budapest, Hungary

Corresponding Author:
Amir Reza Ashraf, PharmD, PhD
Department of Pharmaceutics
Faculty of Pharmacy
University of Pécs
Rókus str. 4.
Pécs, H-7624
Hungary
Phone: 36 (72) 503 650 ext 28841
Email: ashraf.amir.reza@pte.hu

Abstract

Background: Over the past 4 decades, obesity has escalated into a global epidemic, with its worldwide prevalence nearly tripling. Pharmacological treatments have evolved with the recent development of glucagon-like peptide 1 agonists, such as semaglutide. However, off-label use of drugs such as Ozempic for cosmetic weight loss has surged in popularity, raising concerns about potential misuse and the emergence of substandard and falsified products in the unregulated supply chain.

Objective: This study aims to conduct a multifactor investigation of product quality and patient safety risks associated with the unregulated online sale of semaglutide by examining product availability and vendor characteristics and assessing product quality through test purchases.

Methods: We used a complex risk and quality assessment methodology combining online market surveillance, search engine results page analysis, website content assessment, domain traffic analytics, conducting targeted product test purchases, visual quality inspection of product packaging, microbiological sterility and endotoxin contamination evaluation, and quantitative sample analysis using liquid chromatography coupled with mass spectrometry.

Results: We collected and evaluated 1080 links from search engine results pages and identified 317 (29.35%) links belonging to online pharmacies, of which 183 (57.7%) led to legal pharmacies and 134 (42.3%) directed users to 59 unique illegal online pharmacy websites. Web traffic data for the period between July and September 2023 revealed that the top 30 domains directly or indirectly affiliated with illegal online pharmacies accumulated over 4.7 million visits. Test purchases were completed from 6 illegal online pharmacies with the highest number of links offering semaglutide products for sale without prescription at the lowest price range. Three injection vial purchases were delivered; none of the 3 Ozempic prefilled injection pens were received due to nondelivery e-commerce scams. All purchased vials were considered probable substandard and falsified products, as visual inspection indicated noncompliance in more than half (59%-63%) of the evaluated criteria. The semaglutide content of samples substantially exceeded labeled amounts by 28.56%-38.69%, although no peptide-like impurities were identified. The lyophilized peptide samples were devoid of viable microorganisms at the time of testing; however, endotoxin was detected in all samples

XSL•FO
RenderX

with levels ranging between 2.1645 EU/mg and 8.9511 EU/mg. Furthermore, the measured semaglutide purity was significantly low, ranging between 7.7% and 14.37% and deviating from the 99% claimed on product labels by manufacturers.

Conclusions: Glucagon-like peptide 1 agonist drugs promoted for weight loss, similar to erectile dysfunction medications more than 2 decades ago, are becoming the new blockbuster lifestyle medications for the illegal online pharmacy market. Protecting the pharmaceutical supply chain from substandard and falsified weight loss products and raising awareness regarding online medication safety must be a public health priority for regulators and technology platforms alike.

(J Med Internet Res 2024;26:e65440) doi:10.2196/65440

KEYWORDS

semaglutide; Ozempic; Wegovy; search engines; online pharmacies; patient safety; medication safety; nondelivery schemes; counterfeit; substandard and falsified medical products

Introduction

Global Obesity Epidemic

Over the past 4 decades, obesity has escalated into a global epidemic, with its worldwide prevalence nearly tripling. In 2016, over 1.9 billion adults had a BMI between 25 kg/m^2 and 29.9 kg/m^2, classifying them as overweight, while over 650 million had a BMI of ≥30 kg/m^2, falling into the obese category, accounting for 39% and 13% of the adult population, respectively [1]. If current trends continue, it is predicted that by 2035, more than half of the global population, exceeding 4 billion people, will be overweight, and 1 in 4 individuals, nearly 2 billion people, will be obese [2]. This alarming projection highlights the urgent need for the development of effective interventions, including new pharmacological treatments, given that overweight and obesity are key contributors to the global disease burden, ultimately leading to a decrease in life expectancy and quality of life [3-6].

Traditionally, weight loss management focuses on lifestyle modifications, diet, and exercise. However, these interventions alone are often inadequate for achieving and maintaining substantial weight reduction [7], prompting the development of surgical and pharmacological weight loss treatment options. The history of pharmacological weight loss treatments dates back over a century, with various medications being developed and subsequently withdrawn due to their low efficacy and severe adverse effects.

In the 1930s, amphetamines were introduced as a weight-loss treatment after their appetite-suppressing effects were discovered during narcolepsy studies [8]. However, their high abuse potential led pharmaceutical chemists to search for nonaddictive alternatives for appetite suppression. Around the same time, 2,4-dinitrophenol, a compound used in explosive manufacturing, was marketed as a weight loss drug and was subsequently banned by the United States Food and Drug Administration (FDA) due to its narrow therapeutic index and severe adverse effects [9,10]. Despite its ban from medical prescriptions, dinitrophenol has recently regained popularity among bodybuilders and extreme dieters seeking rapid weight loss, with several documented fatalities linked to dinitrophenol weight loss capsules obtained from illegal online vendors [11-14].

Noteworthy anti-obesity treatments developed in the 1990s, such as sibutramine and orlistat, faced their own challenges. Sibutramine was withdrawn from the market in 2010 due to increased cardiovascular adverse events, while orlistat's labeling had to be revised to include a warning about the potential for severe liver injury [15]. Other anti-obesity drugs, such as Belviq (lorcaserin HCI), a controlled substance approved by the FDA and recalled from the US market in 2020, have also been detected as illegally sold drugs by online vendors even before they were approved for sale [16].

Emergence of Glucagon-Like Peptide 1 Receptor Agonists and Concerns About Illegal Trade

In recent years, the landscape of weight-loss pharmacotherapy has evolved significantly, with the emergence of new incretin-based therapies targeting glucose and appetite regulation through glucose-dependent insulinotropic polypeptide and glucagon-like peptide 1 (GLP-1) receptor agonists [17]. Although research into the long-term safety profile and possible adverse effects of these drugs is still evolving, current evidence supports their use as effective treatment options for managing overweight and obesity in individuals with and without diabetes [18]. Notable new incretin-based therapies include tirzepatide (Mounjaro and Zepbound), liraglutide (Victoza and Saxenda), and semaglutide marketed under the trade names Ozempic, Wegovy, and Rybelsus. Among these preparations, as of early 2024, only Zepbound, Saxenda, and Wegovy have been FDA-approved for chronic weight management in individuals who have at least 1 weight-related medical condition (eg, type 2 diabetes, hypertension, or high cholesterol). For instance, Ozempic is only approved for the management of type 2 diabetes, and its use by individuals without diabetes for weight loss is considered off-label use.

Despite a limited labeled indication, Ozempic has surged in mainstream popularity as an off-label cosmetic weight loss treatment due to widespread media attention and celebrity endorsements across various platforms [19]. This heightened popularity is reflected in a 2023 nationwide survey in the United States, with 22% of individuals reporting to have asked their health care providers about using Ozempic for weight loss, 15% have used it for weight loss themselves, and nearly half (47%) know someone who has used Ozempic [20]. However, this rapid rise in popularity has not been without concerns; 76% of medical practitioners worry about its potential misuse, 59% fear it will limit access for patients with diabetes, and 54% fear impending shortages [20].

XSL•FO
RenderX

The popularity of the off-label use of Ozempic for weight loss, paired with capacity limitations at several manufacturing sites, has resulted in ongoing shortages throughout the European Union and the United States [21-23] causing tangible difficulties for patients with diabetes and creating fertile conditions for the proliferation of illegal online pharmacies that aim to capitalize on the heightened demand to sell counterfeit or substandard versions of the medication. In fact, major regulators such as the FDA in the United States and the European Medicines Agency, as well as international organizations such as the World Health Organization (WHO) have issued statements warning about counterfeit versions of Ozempic.

Study Objectives

In response to the rapidly evolving landscape of weight loss drug development, this study aimed to perform an in-depth multifactor investigation of the unregulated online trade of semaglutide through online pharmacy websites, examining online product availability and vendor characteristics, and assessing product quality through test purchases. This multiphase study expands on the results reported in a previously published short research letter by completely describing the methodology and characteristics of that study [24]. The aim of this study is to enable the development of effective strategies to combat this growing threat to patient safety, public health, and the integrity of the pharmaceutical supply chain.

Methods

Overview

This multiphase study was conducted in three phases: (1) manual and computer-assisted web market surveillance and content analysis for semaglutide products offered for sale online from July to September 2023; (2) completing a set of targeted test purchases from August to September 2023; and (3) completing a set of tests for pharmaceutical quality assessment, including examining logistics, visual package inspection, microbiological contamination, and chemical analysis of purchased products (Figure 1).

Figure 1. Summary of study methodology, including online market availability, website monitoring, packaging analysis, and complex product quality control testing. EU: European Union; FIP: International Pharmaceutical Federation.

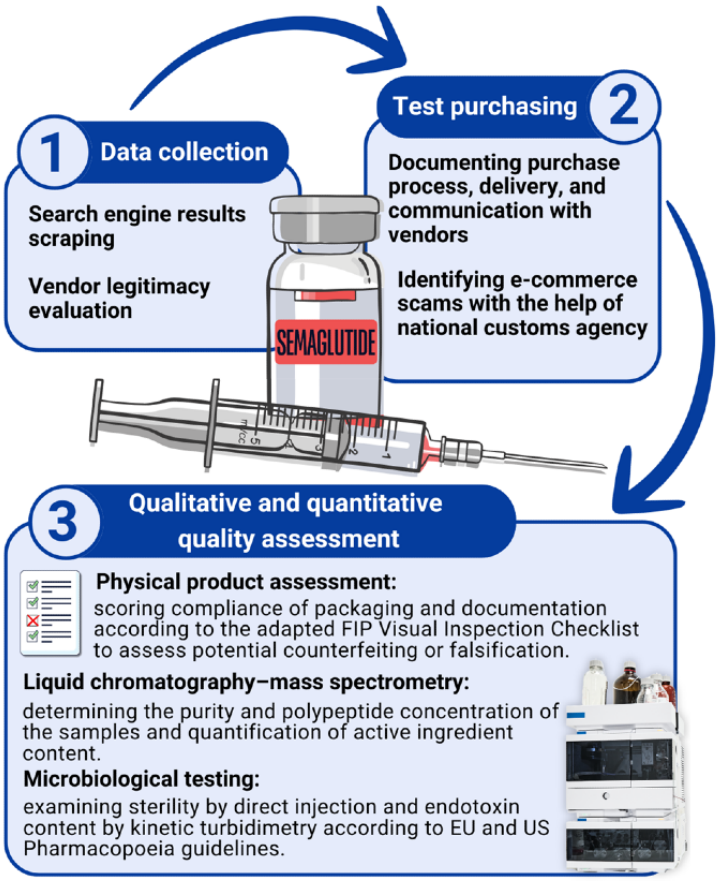

Online Investigation and Search Engine Results Data Collection

In this first phase of the study, we aimed to identify the presence of illegal online sales of semaglutide, specifically from highly ranked search engine results of online pharmacies. We chose online pharmacies, rather than other potential sources of online sales (eg, social media advertisements, communication platforms, and dark web markets), as our unit of analysis because these sites are more conducive to content analysis related to patient safety concerns and pharmaceutical advertising and generally have e-commerce tools (eg, product menus and shopping carts) that facilitate test purchasing processes. We developed a market surveillance methodology, combining automated web crawling and search engine scraping with manual

XSL•FO
RenderX

website evaluation similar to that used in prior published studies [25].

Google (Google LLC) and Bing (Microsoft Corporation) search engines were queried for "buy [Proprietary Name/API]" and "buy [Proprietary Name/API] without prescription" keyword combinations used to collect links from search engine results pages (SERPs) for Ozempic, Wegovy, and semaglutide. Data collection was performed using virtual machines hosted on Amazon Web Services and an automated web crawling program developed in Python (Python Software Foundation) using the Selenium package. Automated web crawling generated search engine results for both Google and Bing with all web browser cookie data, cached data, and search history deleted before conducting each search. To ensure that the links collected from SERPs were regionally relevant, geolocation and IP addresses were configured for study countries before data collection (Multimedia Appendix 1 [26-29]).

The following four categories were used to categorize the links: (1) legal internet pharmacies, (2) illegal vendors, (3) telemedicine websites, and (4) a collective *other* category for websites not providing pharmacy or telemedicine services. Websites were categorized as illegal pharmacies if the domain was classified as "rogue" by LegitScript (a private company that monitors and classifies online pharmacies based on their compliance with applicable laws and regulations and classifies them as "rogue" if they are operating out of compliance, illegally, or fraudulently) [30] or categorized as "not recommended" according to the National Association of Boards of Pharmacy (NABP) Safe Pharmacy verification database. NABP is a nonprofit organization that supports and works with the US state boards of pharmacy and maintains a list of websites that are reviewed for possible fraudulent and unsafe prescription medication sales online [31].

In addition, unverified pharmacy websites, whose domains were not listed in these databases, underwent a thorough manual evaluation to determine legitimacy. Two authors, ARA and AF, both licensed pharmacists, independently manually inspected every website, which was then followed by a joint discussion to reach a consensus on the legitimacy rating of all unverified websites not included in official databases. The manual website assessment was performed based on the following criteria: (1) indication of prescription requirement; (2) presence of sufficient patient guidance and product information; (3) presence of vendor's address and location on the website; and (4) most importantly, whether a relevant and valid regulatory body certification, logo, seal of approval, or registration information could be found on the website. For the purpose of website content analysis and vendor selection, we selected only those domains that were featured >3 times in the aggregated search engine results.

Vendor Selection Criteria and Test Purchasing

Websites included for content evaluation and test purchase were assessed based on the following variables: availability of formulation (prefilled injection pens or injection vial of lyophilized powder) of the parenteral semaglutide product, shipping information and geographical limitations, price, prescription requirement, comprehensiveness of product description (medication use, side effect information, promotion of off-label or unauthorized use, precautions, etc), vendor's address and location, regulatory body logo, and seal of approval. We aimed to document potential medication and patient safety issues originating from information provided by online vendors, in line with methods used in our previously published studies that have examined aspects of safety risks associated with online pharmacies [32-35]. Websites inaccessible during our evaluation, referral sites (ie, affiliate sites) not offering the sale of products directly to customers, duplicate content websites (different domains with identical content), sites requiring submission of prescription documents before purchase, vendors not delivering products to Hungary or United States, and sites offering products above the previously defined price of US $200/single vial or pen were not included for detailed content analysis and test purchase. This price cutoff was chosen based on the average online price of semaglutide products that consumers would encounter at the time of the investigation, assuming they would opt for affordable and accessible choices when purchasing semaglutide products from online sources.

The overall rationale for this selection approach was to enable an exploratory study that assessed the highest-risk vendors based on previously known attributes (eg, low pricing and product availability) associated with illegal online sellers and the sourcing of substandard and falsified medicines. For the purpose of this study, we adopted the definitions and characterization of products as "substandard" or "falsified" consistent with the WHO definition and characterization of products as "counterfeit" consistent with both Hungarian and US regulations, given that these were the 2 destination countries for test purchases (further information on definitions is given in Multimedia Appendix 1). Each step of the online ordering procedure was photographed and simultaneously video recorded for future reference.

Credit cards, PayPal (PayPal Holdings, Inc), and bank transfers were the preferred payment options. Because chemical and microbiological analyses were to be performed in Hungary, the preferred shipping location was Hungary; however, if a seller did not offer direct delivery to this location, parcels were shipped to California, United States, and then forwarded to Hungary by express courier service.

Assessment of Labeling and Packaging

The International Pharmaceutical Federation's (FIP's) checklist for visual inspection of medicines [36] was initially designed to aid health care workers in assessing whether a product is substandard or falsified. This evaluation encompasses a thorough assessment of the integrity of the container, which is essential to protect the product from external contamination and ensure its stability throughout its intended shelf life. The checklist also focuses on legal aspects, such as registration with the relevant drug regulatory authority and proper marketing authorization. Attention is given to the accuracy of information on labels, including the correct spelling of active ingredients and consistency between trade names and registrations. The checklist also addresses the authenticity of track and trace labeling, the explicit indication of storage conditions, and the inclusion of comprehensive informative leaflets. These elements are critical

XSL•FO
RenderX

for maintaining consumer safety and ensuring compliance with applicable regulatory standards.

The authors have adapted the FIP checklist for the evaluation of packaging and labeling of delivered pharmaceutical products, based on the methodology outlined in the published work of Schiavetti et al [37], as well as our own preceding research [33]. This 22-item list is tailored to the specific dosage form of pharmaceutical products and accounts for significant advancements in protective measures that have been developed since the FIP list's initial publication, such as serialization and implementation of data matrix or 2D codes on product packaging (Multimedia Appendix 1). The primary objective of this checklist was to rigorously evaluate the quality, safety, and regulatory compliance of medical products. The visual inspection involved observation and photo documentation of the primary and secondary packaging, leaflets, and labeling.

Microbiological Testing

Sterility and microbiological contamination assessment tests were performed on lyophilized peptide samples obtained from sellers who successfully delivered purchased products to assess product quality and cleanliness of the manufacturing environment. Testing was performed at the International Organization for Standardization (ISO) 14644-1–certified microbiology laboratory of PharmaValid Ltd in Budapest, Hungary. Sterility testing was performed using direct injection technique according to European Pharmacopoeia (European Pharmacopoeia 11.0 2023 2.6.1) and United States Pharmacopeia (United States Pharmacopeia-National Formulary 2023 Issue 2, Chapter 71) guidelines, and bacterial endotoxin content measurement was performed via kinetic turbidimetry technique according to the European Pharmacopoeia (European Pharmacopoeia 11.0 2023 2.6.14) and US Pharmacopeia (US Pharmacopeia-National Formulary 2023 Issue 2, Chapter 85) guidelines. Both United States Pharmacopeia and European Pharmacopoeia standards were used to ensure that test-purchased product testing encompassed standards that applied to shipping destinations in both the United States and Hungary. Endosafe KTA2 Limulus Amebocyte Lysate reagent and *Escherichia coli* O55:B5 Control Standard Endotoxin were used. Samples were stored at 20 to 25 °C during analysis. In the case of 2 products, the content of each vial was dissolved in 2 ml of bacteriostatic water for injection provided by each vendor. One of the online vendors we purchased the product from did not provide the solvent necessary for reconstitution; therefore, endotoxin-free water was used during laboratory testing.

Liquid Chromatography Mass Spectrometry Analysis

Stock solutions of the standard and the polypeptide samples were prepared in methanol. The working solutions were diluted using water/acetonitrile/formic acid (49/49/2, v/v/v). The estimated concentration of the polypeptide samples after dilution was 5 µg/mL. The final concentrations of the standard used for the calibration were 5, 1, 0.5, 0.25, and 0.1 µg/mL. To prevent polypeptide adsorption, every dilution step was carried out using low-protein binding Eppendorf microcentrifuge tubes. The prepared calibration standards and samples were transferred into 0.5 mL low protein-binding Eppendorf tubes, and 5 µL of each sample were injected from a special Eppendorf carrier plate. Chromatographic separation was performed using a Thermo Ultimate 3000 UHPLC system (Thermo Fisher Scientific) with a Luna Omega PS-C18 reversed-phase column (1.6 µm, 2.1 mm×150 mm inner diameter) from Phenomenex. For the multistep gradient-based separation method, we used 2 different solvents. Solvent A consisted of water/formic acid (99.9/0.1, v/v), while solvent B was composed of acetonitrile/formic acid (99.9/0.1, v/v). The gradient program included steps outlined in Multimedia Appendix 2. Data-dependent mass spectrometric acquisition was performed using a Bruker Maxis 4G UltraHigh Resolution Quadrupole Time of Flight instrument (Bruker Daltonics). The mass spectrometer was operated in positive ion mode, and the scanning range was set to 300 to 2200 m/z. The flow of nebulizer gas was 6 L/min at a pressure of 2 bar, and the temperature was set to 180 °C. The capillary voltage was 3.8 kV, and the 10 most intensive compounds were selected for collision-induced dissociation fragmentation. All data were processed using the Data Analysis software package (version 4.4, Bruker Daltonics).

Ethical Considerations

Our study protocol was reviewed by the institutional review board at the University of Pécs (9710-PTE 2023) and was granted a waiver of ethics approval and consent as no human participants were participating in this study, and all information evaluated in our study was publicly available online and accessible via search engines and websites.

Results

Online Vendors Offering Semaglutide for Sale

Upon evaluating the 1080 links extracted from SERPs, we identified 317 (29.35%) links belonging to internet pharmacies, out of which 183 (57.7%) links directed users to legal pharmacies, and 134 links (42.3%) were identified directing users to 59 unique illegal online pharmacies and vendor websites. Notably, out of 59 illegal online pharmacies 21 (36%) were listed multiple times across the SERPs, with SemaSpace emerging as the most prevalent domain. In addition to parenteral products such as the Ozempic pen, we were able to identify offerings of unbranded injection vials of semaglutide and oral dosage form formulated tablets. Notably, oral formulations for semaglutide were not widely available through legal channels at the time market surveillance was conducted for this study.

Using the NABP's Safe Pharmacy database to investigate the legitimacy of websites, we identified 47% (28/59) online pharmacies that were flagged as not recommended, whereas 53% (31/59) the remainder were not listed in the database, highlighting the challenge organizations such as NABP face in maintaining a comprehensive and current registry of illegal online pharmacies. Similarly, an evaluation through LegitScript's database found that 47% (28/59) online pharmacies were classified as rogue, 24% (14/59) as unapproved, and 29% (17/59) as unlisted. Further emphasizing the elusive nature of these unregulated websites, it is important to highlight that 19% (11/59) of the domains were not listed in either database, which indicates that it is often challenging to identify and keep an up-to-date list of illegal online pharmacies, as these websites

XSL•FO
RenderX

constantly cycle through being removed or shut down and new sites easily created and indexed.

A geographic analysis of the vendors revealed that nearly one-third (19/59, 32%) did not provide any address or contact information on their website, a clear indication of risk for counterfeit or illegitimate products. Among those disclosing their alleged locations, Canada and the United States were the most frequently listed, at 31% (18/59) and 22% (13/59), respectively. This was corroborated by WHOIS (a public directory service providing access to domain name registration information, eg, domain ownership and contact) registration data, which also indicated that most (26/59, 44%) of the domains were registered in these countries. Specifically, WHOIS registration data indicated that 31% (18/59) of the domains were registered to individuals based in the United States, 14% (8/59) were based in Canada, and 8% (5/59) were based in Iceland, while registrant location information for 24% (14/59) of the domains was withheld by the registrar, citing General Data Protection Regulation privacy laws as the reason. Investigating the hosting service provider data for the illegal online pharmacies revealed that a significant portion (45/59, 76%) of the websites were hosted by companies in the United States and Canada, with Cloudflare Inc being a preferred Canadian service provider, hosting more than a quarter (16/59, 27%) of the illegal websites detected.

Analysis of the web traffic data provided by SimilarWeb for the period between July and September 2023 revealed that the top 30 domains affiliated with illegal online pharmacies (Table 1) collectively accumulated an estimated 4.7 million visits, with a major share of traffic distributed among the top 5 websites, capturing more than 58% (2,730,848/4,705,502, 58.04%) of the total visits, indicating a substantial concentration of website traffic going to the top ranked websites indexed in structured search results. There was a considerable skew in visitor distribution, as toward the bottom of the top 30 list, the last 5 websites only accounted for a little over 1.55% (73,166/4,705,502) of the total visits, evidencing how website traffic drops for sites not prominently appearing in search engine results. The top sites included both conventional illegal pharmacies as well as newly emerging peptide-focused sellers, reflecting the increasing interest in purchasing peptide-based pharmaceuticals.

XSL•FO
RenderX

Table 1. Top 30 online vendor and intermediary website analytics and market surveillance data between July and September 2023.

Domain	LegitScript	Safe Pharmacy database (NABP[a])	Location listed on the website	Domain registrant country (WHOIS)	Host	Server location	Total estimated visits between July and September 2023
muscleandbrawn.com	Not in the database	Not in the database	United States	Iceland	DigitalOcean LLC	United States	842,045
empowerpharmacy.com	Unapproved	Not in the database	United States	United States	Amazon Inc	United States	591,844
aminoasylum.shop	Rogue	Not in the database	__[b]	United States	Cloudflare, Inc	Canada	496,859
northwestpharmacy.com	Rogue	Not recommended	Canada	—	Cloudflare, Inc	Canada	447,587
canadapharmacy.com	Unapproved	Not recommended	Canada	United States	Amazon Inc	United States	352,513
canadadrugsdirect.com	Unapproved	Not recommended	Canada	United States	Amazon Inc	United States	215,126
corepeptides.com	Rogue	Not in the database	United States	Canada	Cloudflare London, LLC	United States	194,704
oddwayinternational.com	Rogue	Not in the database	India	India	Cloudflare, Inc	Canada	171,865
pharmaserve.com	Unapproved	Not recommended	Canada	—	Pretecs Networks Inc	Canada	143,664
insulinoutlet.com	Unapproved	Not recommended	Canada	United States	Google LLC	United States	132,143
biotechpeptides.com	Rogue	Not in the database	United States	Canada	Cloudflare London, LLC	United States	131,539
canadianpharmacyking.com	Rogue	Not recommended	Canada	Turks and Caicos Islands	Ntirety, Inc	United States	124,872
canshipmeds.com	Unapproved	Not recommended	Canada	—	Cloudflare, Inc	Canada	121,423
canadianinsulin.com	Unapproved	Not recommended	Canada	Canada	Cloudflare, Inc	Canada	101,742
israelpharm.com	Rogue	Not recommended	Israel	Israel	Cloudflare, Inc	Canada	93,549
semaspace.com	Not in the database	Not recommended	United States	United States	Cloudflare London, LLC	United States	75,686
buycanadianinsulin.com	Unapproved	Not recommended	Canada	Canada	Google LLC	United States	70,756
kiwidrug.com	Rogue	Not recommended	New Zealand	Panama	Solar Communications GmbH	Switzerland	56,360
oxygenpharm.cc	Not in the database	Not in the database	Canada	The Bahamas	Cloudflare Inc	Canada	48,222
polarbearmeds.com	Not in the database	Not recommended	Canada	Canada	BigScoots	United States	43,668
canpharm.com	Unapproved	Not recommended	Canada	—	Amazon Inc	United States	41,154
mydietdoc.com	Unapproved	Not in the database	United States	United States	Google LLC	United States	38,605
genx.bio	Not in the database	Not in the database	United States	Canada	Vultr Holdings LLC	United States	37,133
insulin.store	Not in the database	Not recommended	Canada	Iceland	Namecheap Inc	United States	32,185
australiaweightloss.com	Not in the database	Not in the database	Australia	—	Cloudflare, Inc	Canada	27,092
nevergiveupteam.com	Not in the database	Not in the database	Australia	United States	Cloudflare, Inc	Canada	21,358
help-pharm.com	Not in the database	Not in the database	—	Russia	Cloudflare, Inc	Canada	19,440

XSL•FO
RenderX

Domain	LegitScript	Safe Pharmacy database (NABP[a])	Location listed on the website	Domain registrant country (WHOIS)	Host	Server location	Total estimated visits between July and September 2023
uschemlabs.com	Rogue	Not in the database	—	United States	Cloudflare, Inc	Canada	14,649
rxsavemeds.com	Rogue	Not in the database	United States	Malaysia	FiberState LLC	United States	9226
thebodyfollows.com	Unapproved	Not in the database	—	United States	Network Solutions LLC	United States	8493

[a]NABP: National Association of Boards of Pharmacy.

[b]Not applicable.

Semaglutide Vendor Selection for Test Purchasing

A total of 6 online vendors offering parenteral semaglutide products were included in our content evaluation and test purchase phase (Table 2). To highlight the potential real-world issues of unregulated online semaglutide sales, we aimed to perform test purchases from what we consider higher-risk unapproved websites that were most likely visited by customers at the time of our evaluation. Consequently, these 6 rogue domains had the highest number of links (high prevalence) in all SERPs and offered semaglutide products for direct-to-consumer sale without prescription (easy access) at the lowest price range (affordability).

Table 2. Rogue online vendor characteristics and the semaglutide products purchased for testing in August 2023.

Vendor and product name	In the top 10 search results	LegitScript Verification	NABP[a] category	Product form and dosage	Instructions (storage, administration, side effects, etc)	Communicated health-related benefits	Product price[b] (+shipping fee)	Payment options	Shipped from (courier)
BiotechPeptides "Semaglutide (GLP-1[c]) 3mg"	Yes	Rogue	Not recommended	Semaglutide vial (2 mg)	No	Appetite suppression and cardiovascular	US $113/vial (no shipping fee)	ACH[d], CashApp, Venmo, and credit card	San Diego, California, United States (USPS[e])
GeniusPharmacy "Ozempic- 0.25mg, 2Pens×1"	No	Rogue	Not recommended	Ozempic pen (0.25 mg)	Yes	Weight loss and diabetes	US $360/2 pens (+US $ 50)	Bitcoin and Zelle	Nondelivery scam (MailingOnTime)
PureMedsOnline "Ozempic Online - 0.25 mg 2 pens"	No	Rogue	Not recommended	Semaglutide vial (1 mg)	Yes	Diabetes, weight loss, and cardiovascular	US $300/2 pens (+US $30)	PayPal, Zelle, and bitcoin	Nondelivery scam (AsianaLogisticsExpress)
SemaSpace "Weeks 1-6 Bundle: 2 mg vial and bacteriostatic water 30 ml vial"	Yes	Not in the database	Not recommended	Semaglutide vial (3 mg)	Yes	Obesity	US $199/vial (+US $30)	PayPal only	New York, New York, United States (USPS)
USChemLabs "Semaglutide 1mg 5 Vials"	Yes	Rogue	Not in the database	Ozempic pen (0.25 mg)	No	Weight loss and blood sugar regulation	US $148.9/5 vials (+US $25)	Credit card, CashApp, and crypto currency	Camarillo, California, United States (FedEx)
WeightCrunchShop "Ozempic 0.25mg 1Pen"	Yes	Rogue	Not recommended	Ozempic pen (0.25 mg)	Yes	Weight loss	US $190/1 pen (+US $30)	Apple and Google pay, Zelle, bank transfer, Osko, and Bitcoin	Nondelivery scam (GLSCourierServices)

[a]NABP: National Association of Boards of Pharmacy.

[b]Smallest quantity offered for sale.

[c]GLP-1: glucagon-like peptide 1.

[d]ACH: payments through the Automated Clearing House network.

[e]USPS: United States Postal Service.

XSL•FO
RenderX

All included online vendors were categorized as illegitimate (eg, “not recommended,” “rogue,” or “unapproved”) by the LegitScript or NABP verification databases. In total, 3 (50%) of the 6 websites offered Ozempic injection pens for sale, while another 3 (50%) sold vials of lyophilized semaglutide powder to be reconstituted into a solution for injection. The price for the smallest dose and quantity of product available on the websites ranged from US $113 to US $360/purchase (mean $218.5, SD $93.6). One domain allowed only 1 payment method, while the rest offered numerous payment options, with 66% (4/6) vendors also offering cryptocurrency payments. None of the vendors requested prior medical prescriptions or any health-related information from patients before or during the purchase, and semaglutide pen sellers explicitly highlighted “no prescription” availability as marketing messages. All online vendors directly or indirectly referred to weight loss and obesity on the product page; thus, all the websites included in our study explicitly promoted unauthorized or off-label use of Ozempic or semaglutide-containing products. Instructions for use, storage, and administration were provided on 4 domains; however, 2 out of 3 lyophilized semaglutide peptide sellers did not communicate how to reconstitute, use, or administer the product.

The product descriptions provided by the lyophilized semaglutide peptide sellers presented their products ambiguously. On one hand, the ampules and the company logos suggested a research and laboratory focus, listing chemical properties, formulas, etc, with product labels indicating “research use only.” By contrast, most product descriptions provided by the peptide dealers emphasized the effects of GLP-1 agonists, describing preclinical and clinical studies on potential health-related benefits (eg, beta cell protection, weight loss, diabetes, appetite reduction, and neurodegenerative disease). Both BiotechPeptides and USChemLabs listed numerous scientific publications and references to support the clinical benefit claims listed in their product descriptions.

Regardless of the product formulation, all 6 websites permitted free access to semaglutide vials or prefilled injection pens without requiring any official documentation or professional information from the buyer. Thus, no prescriptions or patient health-related data were requested for Ozempic injections by the illegal vendors nor was any proof of personal qualification or corporate information required for research chemicals sold as injection vials by the peptide dealers. Although 2 lyophilized semaglutide peptide sellers seemingly discouraged human use of their “lab chemicals” and did not mention instructions on how to administer the product, the SemaSpace website, the other illegal pharmacy selling lyophilized semaglutide peptide, explicitly provided instructions on how to mix semaglutide with bacteriostatic water, along with syringes, alcohol prep pads, and dosing and injection guidance for customers (Figure 2). Prefilled Ozempic injection pen sellers listed information on product administration, precautions, and other pertinent information for consumer use. One website, WeightCrunchShop, contained numerous elements that are common among illegal internet pharmacy practice, including an emphasis on “without prescription” availability, “discrete delivery,” discount and “lowest price promise,” “money-back guarantee,” and related access to prescription-only products, such as ivermectin or ephedrine (Multimedia Appendix 1).

All online test purchases were completed within minutes from product selection to checkout. Most online vendors (4/6, 66%) offered untraceable payment methods. Furthermore, payment with cryptocurrency was encouraged and incentivized by offering a 5% discount or free shipping during the checkout process. On the BiotechPeptides website, customers had to sign the following statement before placing their order: “I agree to the terms and conditions. I am at least 21 years of age and understand that purchases are limited to licensed researchers and qualified professionals. I understand that the products listed on the site are not for human or animal use.” Similarly, the USChemLabs website’s terms and conditions also mention that “These products should only be used by Qualified Professionals. Bodily introduction into Human or Animals of any kind is strictly forbidden by law. All product information on this website is for educational purpose only.”

XSL•FO
RenderX

Figure 2. Screenshot of SemaSpace demonstrating common characteristics of illegal online pharmacies: off-label use promotion, no prescription requirements, unverified before-and-after photos to suggest efficacy, and misleading marketing strategies, such as risk-free trial and money-back guarantees. Personal identifiers have been blurred.

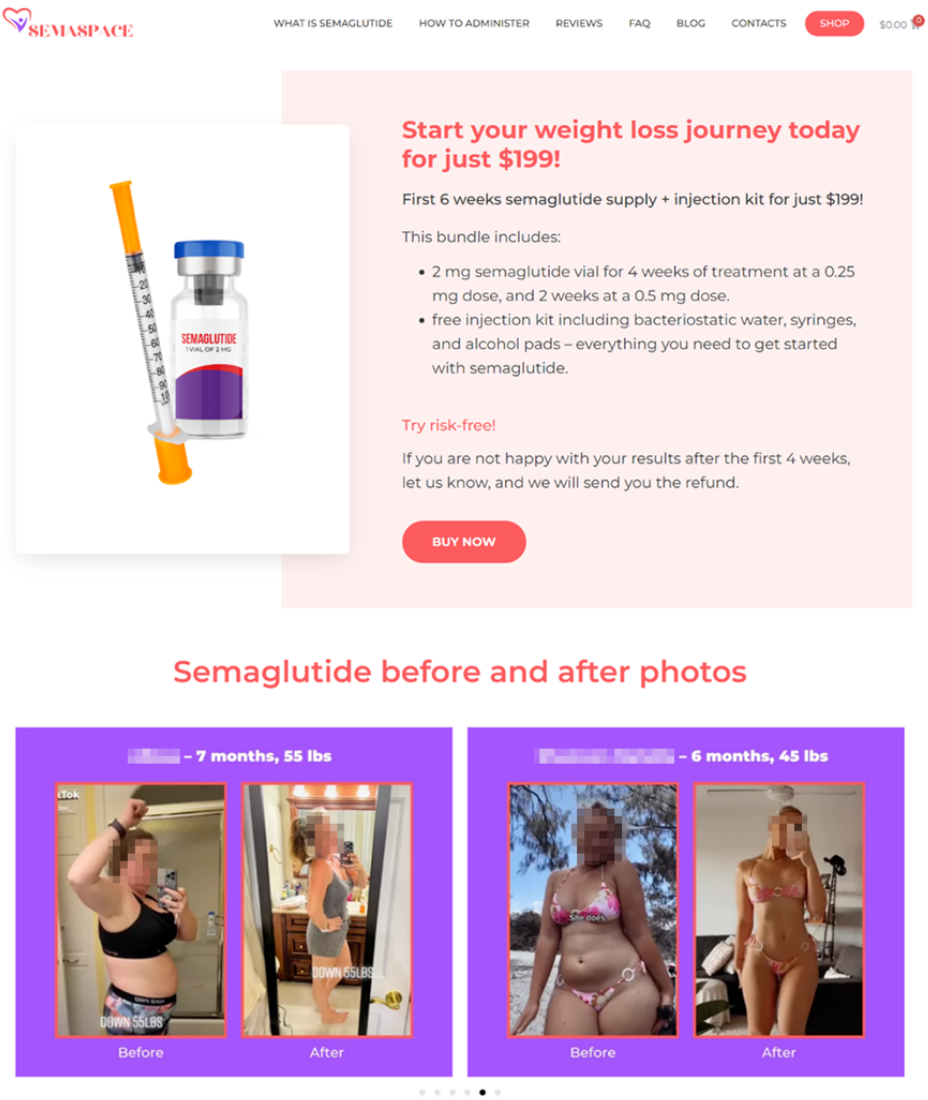

Payment, Delivery, and Communication With Sellers

All purchases were confirmed by email. The automatic confirmation messages sent by the 3 Ozempic pen sellers were highly similar in layout and content and were created using the same open-source e-commerce platform (WooCommerce WordPress plugin). Only 2 (33%) of the 6 sellers (USChemLabs and BiotechPeptides) provided real-time credit card payment, while the rest (n=4, 66%) requested customers to complete their payment after receiving instructions by email following the online purchase.

Although purchases were confirmed via email, all further communication regarding payment and customs fees was initiated by sellers via WhatsApp Messenger (Meta Platforms, Inc). As cryptocurrency was the most preferred payment option, vendors provided guidance for customers on “How to pay with Bitcoin” using a credit card or cash app. However, because cryptocurrency payment is not available to all consumers and could be time consuming, we preferred and used more accepted payment solutions, such as PayPal or credit card. PayPal payments followed a similar process for all sellers. First, the vendors sent the PayPal username as an email associated with the vendor. Second, detailed instructions were sent by email, provided on the website, and communicated via WhatsApp on how to send the required payment amount. The payment instructions specified that the payment description field on PayPal could not contain anything except the customer’s name, and the payment type must be “For Friends and Family” (not for goods and services), or the order would be canceled and the money refunded, suggesting that vendors were requesting buyers to help conceal the nature of the payment.

All semaglutide injection vial products purchased were delivered; however, none of the Ozempic prefilled injection pens were received, as the illegal online pharmacies selling Ozempic pens used nondelivery e-commerce scams and the so-called customs clearance advance fee scam. After making the online purchases, vendors involved in nondelivery schemes referred us to third-party courier websites (AsianaLogisticsExpress, GLSCourierServices, and MailingOnTime) and provided fake tracking numbers that would

XSL•FO
RenderX

show the packages were held up at customs. The sellers then demanded substantial sums of money to assist in releasing the purchased semaglutide medication, which they claimed was being held by local customs. These requested sums included US $1200 as an “insurance fee,” €450 (US $481.5) for “x-ray custom stamps,” and US $650 for “insurance and prescription stamps” to clear the supposedly stuck product through customs in Hungary. They assured us that the money would be refunded upon successful delivery. Upon investigation by the National Tax and Customs Administration, we confirmed that the packages listed as stuck in customs on the tracking website did not exist. Although the courier services, tracking codes, and delivery details could appear valid initially, they were entirely fraudulent and part of an elaborate e-commerce scam. For instance, GLSCourierServices clearly imitated the official GLS Group domain and misappropriated its logo to deceive consumers who fell victim to the e-commerce scam (Figure 3 [38]).

Figure 3. Screenshots of legitimate GLS Group on the left, and fraudulent GLSCourierServices website on the right. The legitimate website features authentic corporate branding, while the fraudulent website, which imitates the official GLS site, provides misleading information intended to deceive consumers. Note that GLS Group has since updated its logo, removing the yellow arrow [38].

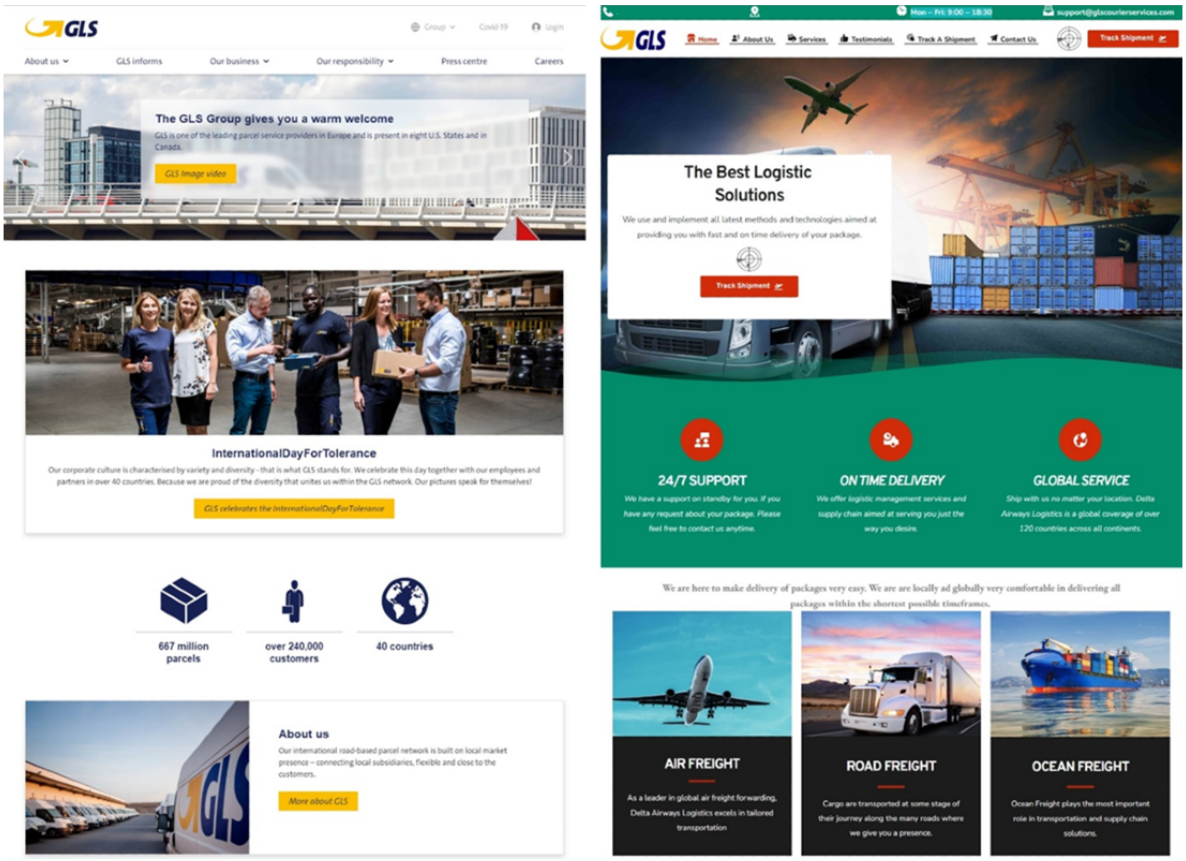

Visual Inspection of Delivered Products

We adopted the FIP tool for visual inspection of medicines checklist for inspection of the delivered products, which we tailored to align with project-specific requirements, allowing us to evaluate semaglutide products against criteria essential for maintaining high standards of quality of a legitimate manufacturer (Table 3).

Genuine Ozempic achieved a perfect score of 22 points, demonstrating full compliance with all assessed criteria. In contrast, products from SemaSpace, BiotechPeptides, and USChemLabs scored considerably lower, with 9, 8, and 8 points, respectively. The most notable discrepancies were observed in areas such as regulatory registration and oversight, accurate and legitimate labeling, and the provision of product-related information. Further visual inspection of the samples provided evidence that the products were unregistered or unlicensed. The FIP tool for visual inspection of medicines checklist revealed noncompliance in 13 (59%) and 14 (64%) out of 22 evaluated criteria, signifying a lack of compliance with quality, safety, and authenticity standards. Consequently, all test-purchased samples were classified as substandard, falsified, and counterfeit medical products (Figure 4).

XSL•FO
RenderX

Table 3. Results of visual inspection of products and corresponding scores for each semaglutide product.

Criterion	Genuine Ozempic pen	SemaSpace vial	BiotechPeptides vial	USChemLabs vial
Authentic track and trace labeling	Yes	No	No	No
Trade name legally registered	Yes	No	No	No
Trade name correctly spelled and includes ® symbol	Yes	No	No	No
Active ingredient name correctly spelled	Yes	Yes	Yes	Yes
Manufacturer's name and logo legible and correct	Yes	Yes	Yes	Yes
Trade name and active ingredient match registered product	Yes	No	No	No
Product registered by manufacturer or agent	Yes	No	No	No
Dosage form registered and authorized for sale	Yes	No	No	No
Packaging and container protect from external environments	Yes	Yes	Yes	Yes
Container and closure appropriate for the product	Yes	Yes	Yes	Yes
Container safely and securely sealed	Yes	Yes	Yes	Yes
Container maintains product quality throughout shelf life	Yes	No	No	No
Carton and container labels match	Yes	Yes	No	No
Label information is legible and indelible	Yes	Yes	Yes	No
Manufacturer information and expiry dates indicated on label	Yes	No	Yes	No
Strength clearly stated on label	Yes	Yes	Yes	Yes
Dosage form clearly indicated on label	Yes	No	No	No
Dosage number clearly indicated on label	Yes	No	No	No
Labeled dosage form matches actual product	Yes	No	No	No
Manufacturer's full address legible and correct	Yes	No	No	No
Storage conditions indicated on the label	Yes	Yes	No	Yes
Leaflet with dosage and use information included	Yes	No	No	Yes
Score and degree of compliance with the criteria (n=22), n (%)	22 (100)	9 (41)	8 (36)	8 (36)

Figure 4. Injection vials containing lyophilized semaglutide peptide purchased from illegal internet pharmacies. USChemLabs and BiotechPeptides included purity information and warnings on the labels, while SemaSpace labeling did not. SemaSpace provided a complete injection kit with the semaglutide vials, BiotechPeptides provided bacteriostatic water with the semaglutide vials, while USChemLabs supplied semaglutide vials only.

XSL•FO
RenderX

Microbiological Testing

Bacterial endotoxin content was measured using a kinetic turbidimetry assay at a detection limit of <0.01 EU/mL. Following sample dilution and the measurement of temporal turbidity changes, endotoxin levels for the USChemLabs and BiotechPeptides samples were found to be <2.8658 EU/mg and <2.1645 EU/mg, respectively. In the case of the sample obtained from SemaSpace, a more elevated endotoxin content of 8.9511 EU/mg was detected, indicating potential environmental contamination during production. The presence of endotoxin indicates residuals and remnants from the cell walls of Gram-negative bacteria. Despite the presence of endotoxin, all 3 lyophilized peptide samples were found to be devoid of viable microorganisms at the time of testing. The bacteriostatic water for injection supplied with the products was an off-the-shelf standard product and was not manufactured by the online vendor; therefore, the contaminants most likely originated from the water or nonsterile components used during the manufacturing of the lyophilized peptide. However, the exact source of the endotoxin cannot be determined.

Quantitative and Qualitative Analysis

Semaglutide polypeptide can be identified and effectively quantified using mass spectrometry, as indicated by its characteristic peak at 1029.3 Da $[M+4H]^{4+}$ in the mass spectrum, which serves as a critical reference point for evaluating the analytical measurements (Multimedia Appendix 2). The impurities found in products produced using automated peptide synthesis are highly variable. They can be unwanted by-products, solvent residues, or residues of chemicals used to protect the peptide chain. Due to their low molar mass and poor ionization capabilities, mass spectrometric identification of these substances can be challenging. To ensure the integrity of a pharmaceutical product, manufacturers must carefully monitor the manufacturing processes, which include monitoring the active pharmaceutical ingredient, impurities, and any potential contaminants that may arise from cross contamination from previous manufacturing processes. During our measurements, sample purity and quantity of the active ingredients were examined. The samples were tested for protein impurities; the chromatogram only displayed 1 signal for each delivered sample, indicating the exclusive presence of semaglutide (Multimedia Appendix 2). This indicates that the samples do not contain peptide-like impurities. However, a discrepancy emerged while comparing the purity levels provided by suppliers with the actual content of the active substance determined through our analysis. Documentation, available on vendors' websites or shipped with the product, claims a purity of at least 99%. Contrarily, our analysis of products obtained from SemaSpace, USChemLabs, and BiotechPeptides revealed significantly lower polypeptide concentrations of 14.37%, 8.97%, and 7.7%, respectively. The quantity of semaglutide content in each sample also substantially exceeded the labeled amount, with excesses ranging from +28.56% to +38.69% (Table 4).

Table 4. Labeling accuracy and purity analysis of purchased lyophilized semaglutide samples compared with the original product obtained from the legitimate supply chain.

Sample	Total weight of powder in vial (mg)	Semaglutide content indicated on the label (mg)	Measured semaglutide content (mg), mean (SD)	Labeling accuracy	Purity indicated on label or website (%)	Measured purity (%), mean (SD)
Ozempic 1 mg solution for injection in prefilled pen (reference)[a]	__[b]	1	1.05 (0.02)	Semaglutide content exceeded the labeled amount by+5.05%	—	—
Biotech Peptides lyophilized semaglutide powder vial	50.1	3	3.86 (0.14)	Semaglutide content exceeded the labeled amount by+28.56%	99	7.70 (0.28)
SemaSpace lyophilized semaglutide powder vial	19.3	2	2.77 (0.12)	Semaglutide content exceeded the labeled amount by+38.69%	99	14.37 (0.63)
USChemLabs lyophilized semaglutide powder vial	14.9	1	1.34 (0.07)	Semaglutide content exceeded the labeled amount by+33.58%	99	8.97 (0.51)

[a]On the basis of the official Ozempic European public assessment report product information document published by the European Medicines Agency [39], 1 prefilled Ozempic 1 mg pen contains 4 mg semaglutide in a 3 mL solution. Pen is designed to deliver 4 doses of 1 mg (0.74 mL/dose).

[b]Not applicable.

These findings suggest the presence of low-mass, weakly ionizing impurities, although further studies using additional analytical techniques are necessary to conclusively identify contaminants. It is plausible that the observed impurities are an indication of the omission or incomplete execution of purification steps typically required at the end of the polypeptide manufacturing process, as they are time-consuming and costly, an indicative sign of substandard manufacturing practices used by unregulated or illegal online pharmacies [40,41]. The chemical analysis provides evidence for the online sale of out-of-specification (substandard) products along with falsified medical product classification based on the misleading labeling information.

Discussion

Principal Findings

Our study used a multifactor approach to assess risks associated with purchasing semaglutide online, identifying >300 links to online pharmacies, with close to half (n=134, 42.3%) directing users to 59 unique illegal online pharmacies, many of which sold semaglutide without requiring a prescription or any clinician oversight. On the basis of web analytics data, we

XSL•FO
RenderX

estimate that the top 30 domains affiliated with these illegal online pharmacies accumulated over 4.7 million visits during the data collection phase of the study, between July and September 2023, a period when excitement and demand for semaglutide were accelerating [19,42]. Our test purchases from a selection of high-risk and easily accessible illegal online pharmacy websites revealed that half (3/6, 50%) of the transactions were fraudulent, with sellers engaging in e-commerce nondelivery scams, and the other half (3/6, 50%) of the transactions yielded clearly substandard and falsified products that were not manufactured to specifications and misrepresented their constituency based on labeled information. In fact, 2 of the websites evaluated by test purchases conducted in this study have subsequently been sent FDA warning letters for engaging in unlawful sales of unapproved and misbranded semaglutide [43,44].

Specifically, we identified 2 distinct online marketing approaches used by illegal online pharmacies for this medication, which is trending globally, commands high prices and is impacted by ongoing drug shortages. The first strategy involved e-commerce fraud coupled with customs clearance advance fee scams, also known as "nondelivery schemes." These scams are a form of financial crime, where online actors set up fake internet pharmacy websites to entice buyers with uncontrolled access to prescription-only medications. The catch is that these illegal online pharmacies likely do not actually have any products in stock, which is why shortages do not affect them, and, of course, the purchased products are never shipped. Furthermore, these financial scams include referrals to fake third-party courier service providers that request additional fees to release packages from customs, an effort to scam the customers who fall for the original nondelivery scam a second time.

The second strategy involves selling dubious laboratory chemicals, such as lyophilized peptides shipped in glass vials, or fake generic products while promoting potential health benefits and off-label use. On the basis of the results of our test purchases, these vendors ship products as the purchased vials were delivered. However, through our extensive laboratory evaluation of these products, we determined that although they contained semaglutide, they were substandard and falsified products with considerably low purity levels and elevated presence of endotoxin sold through illegal online sales, corroborating findings of a recent NABP report and concerns raised by global regulators and public health organizations such as the WHO [45]. Furthermore, assay values that indicated the quantity of semaglutide content in tested samples substantially exceeded labeled amounts raise concerns about possible incorrect dosing or overdose of semaglutide, particularly when self-administering using the reconstituted product as purchased in this study. In fact, the FDA has issued similar warnings due to the detection of adverse events and hospitalizations associated with compounded semaglutide injections that were self-administered [46,47].

Opportunities for drug counterfeiting, diversion, financial fraud, and patient harm are likely to continue to grow as global demand for and widespread off-label use of semaglutide continue to be promoted. For example, social media posts by celebrities have facilitated the spread of misinformation [48-50] and hype news diffusion [51] related to the benefits of semaglutide. New technologies, including generative artificial intelligence and search engine chatbots, may also recommend illegal online pharmacies for purchasing semaglutide, as recently reported in a separate study by Ashraf et al [52], further exacerbating potential consumer exposure and harm. The business model of rogue online pharmacies takes an opportunistic approach triggered by this increased demand; popularity driven by online content and drug promotion; and restricted access due to prescription requirements, shortages, or high prices. This "same old recipe" was previously used by illegal online pharmacies for sildenafil in the past 2 decades and more recently for ivermectin during the COVID-19 pandemic [26,53].

The drug supply chain has faced increasing challenges over the past 15 years, with multiple drug shortages affecting legitimate supply chains worldwide [54]. The shortages of antidiabetic and obesity drugs, which began in 2022, continue to be a problem 2 years later. We have entered an era where short viral videos and endorsements by celebrities about these products can have a long-lasting effect on consumers and the entire health care system and drug supply chain [55,56]. This new phenomenon should be investigated in depth from a public health and psychological perspective, and the results should be incorporated into the education of health care professionals and supply chain security tools to detect and interdict counterfeit versions of medications.

In our previous research, we drew attention to the potential risks faced by consumers searching for drug-related information about medicines in shortage, as we identified multiple illegal online vendors offering unavailable medicines. We hypothesized that there is no shortage in the illegal online market; however, we did not purchase medicines subject to a shortage at that time due to safety, ethical, and legal concerns in 2011, 2014, and 2016 [57-59]. In this study, we encountered a similar experience, as illegal internet vendors did not communicate that the medicines were out of stock. In parallel, by completing online purchases, we can strongly assume that an offline shortage coexists with online shortages, which increases the financial risk for patients as rogue online pharmacies engage in nondelivery schemes.

In our previous research, when our research group test purchased the growth hormone somatropine, we identified low amounts of the active ingredient and out-of-specification substandard medications due to inappropriate shipping conditions. In the case of semaglutide vials purchased in this study, analytical and microbiological testing revealed sterile, lyophilized products without extensive degradation. Peptide sellers offer active ingredients of popular prescription-only medicines, communicate the potential health-related benefits, and provide instructions for the reconstitution of freeze-dried powders. However, they simultaneously claim that their unbranded "laboratory products" are only for research purposes and are not suitable for human use. Consumers are likely to be misled and confused by quality and purity-related statements, as research-grade 99% purity and pharmaceutical-grade product quality are distinctly different.

XSL•FO
RenderX

Hence, complex instrumental analysis is required to investigate the quality of online-ordered medical products, including pharmaceutical technology tests; Raman spectroscopy to differentiate between genuine and falsified or substandard products; and qualitative, quantitative, and microbiological tests determined by pharmacopeias. As shown in the study by Ahmed et al [60], substandard and falsified medicines obtained online may pass pharmacopoeial requirements, and the active ingredient can be identified in the products. Due to the complexity and rise in the use of biotechnology medicines, illegal online pharmacies have also turned to these products. It must be emphasized that the quality of an online-ordered product is only one potential source of risk; the potential misuse and the incorrect use of prescription-only drug products without the supervision of a medical practitioner are also of great concern and a source of potential patient harm [60-62].

The seemingly unrestricted proliferation and access to illegal online pharmacies pose significant global public health risks, exacerbated by the public's lack of knowledge on pharmaceutical quality assessment, the distinctions between off-label and labeled use, and the increasing trend of online purchases. To mitigate these risks, it is crucial to integrate training on health products and the unique characteristics of medicines into educational curricula and implement public awareness campaigns on the safe and effective procurement of medical products. Such interventions can help reduce the demand for medications outside the regulated supply chain. One potential tool is the development and dissemination of an international consumer version of the FIP checklist, which could guide consumers in evaluating the purchased medicines and understanding the associated risks. Moreover, there is a notable gap in research analyzing this complex phenomenon with few studies using multifactor methodologies, such as the evaluation of search trends, website metrics, and the health risks associated with products purchased online, as conducted in this study. This study underscores the urgent need for comprehensive strategies to educate consumers, improve content moderation among search engines and service providers, and better enforce regulations to safeguard public health against the dangers posed by illegal online pharmacies and rogue marketplaces of medical products [40,41,62].

Limitations

Comprehensiveness is a major strength and simultaneously a limitation of our study. Incorporation of market surveillance approaches (search engine scraping and website content analysis) and pharmaceutical quality control (visual inspection, microbiological, and chemical analysis) provides a general overview of aspects and concerns associated with online pharmacies and falsified or counterfeit medical products. At the same time, each element of our research methodology could have been elaborated more extensively (eg, larger sample size, assessment of additional verification databases, and toxicological analyses) to provide additional details and identify specific safety and risk characteristics. Major limitations include a relatively small sample size of test purchases based on risk assessment due to the high costs and complexities of conducting test purchases, which limits the generalizability of our findings to all online pharmacies selling semaglutide illegally. While our search methodology has been proven effective in previous studies, not all illegal online pharmacy operations may have been captured in our analysis. The ≤US $200 per item price cutoff, while reflecting average online prices at the time of investigation, may have excluded higher-priced illicit semaglutide products. In addition, our analysis only included online pharmacy vendors shipping to the United States and Hungary, which may not be representative of the global market for illicit semaglutide that likely includes sellers using social media platforms, poorly regulated e-commerce sites, and notorious illegal markets such as dark web marketplaces. Therefore, our findings may not capture the full range of counterfeit semaglutide products available, although the identification across highly visited search engine results is still a serious concern. Lack of reproducibility another limitation as selected illegal online pharmacies may have different production runs or not be accessible online at all for future studies. Although the possibility that product quality might have been compromised during courier delivery could be a concern, this also reflects the real-world conditions consumers experience when purchasing these products from online vendors.

Conclusions

GLP-1 agonists represent an exciting new class of drugs with great potential to improve health outcomes for diabetes, weight loss management, and other health conditions, but are not without their own concerns regarding appropriateness for use, safety, and now quality and accessibility. This study identified a significant presence of illegal online pharmacies selling semaglutide products from online searches, often without requiring prescriptions and promoting off-label use. These illicit vendors pose serious risks to consumer safety, as evidenced by the substandard quality of their products revealed through physical, chemical, and microbiological assessments. Moreover, the popularity of these illegal websites, attracting millions of visits, underscores the urgent need for enhanced regulatory strategies and consumer education. Illegal and unauthorized access to these drugs from an evolving array of online sources may exacerbate such concerns, particularly in the context of postmarket surveillance and identifying causes for an increasing number of adverse events being reported for semaglutide-containing products. Multifactor approaches, such as the ones pursued in this study, have the potential to better triangulate factors related to online risk while also helping to develop targeted interventions (eg, patient education, online content moderation, and law enforcement action against illegal online pharmacies) aimed at ensuring safe and appropriate use of GLP-1s as they become more mainstream for patient care. Future research should continue to monitor consumer demand and online purchasing behavior, examine the role of emerging digital platforms and social media in the distribution of illegal pharmaceutical products, and propose solutions to improve the reliability and safety of search engine recommendations in the context of GLP-1s and online pharmacies.

XSL•FO
RenderX

Acknowledgments

The research was supported by the Hungarian Scientific Research Fund of the National Research, Development, and Innovation Office (NKFIH 143684). The research was performed in collaboration with the Mass Spectrometry Core Facility at the Szentágothai Research Centre of the University of Pécs.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ARA contributed to conceptualization, methodology, investigation, data curation, formal analysis, validation, visualization, writing the original draft, and reviewing and editing the manuscript. TKM contributed to conceptualization, methodology, writing the original draft, and reviewing and editing the manuscript. GK contributed to methodology, analysis, investigation, and writing the original draft. JS contributed to methodology, investigation, data curation, writing the original draft, and reviewing and editing the manuscript. OB and BMD contributed to the methodology, writing the original draft, and reviewing and editing the manuscript. JL contributed to data curation. IC contributed to reviewing and editing the manuscript. AF contributed to conceptualization, methodology, investigation, funding acquisition, project administration, supervision, writing the original draft, and reviewing and editing the manuscript.

Conflicts of Interest

TKM and JL are employees of the startup company S-3 Research LLC. TKM is also a cofounder and co-owner of the company. S-3 Research is a small business that was originally created through a National Institute on Drug Abuse startup business award and has been subsequently funded through government contracts with federal agencies to develop research tools and services for data science in public health, including social listening, data mining, and machine learning approaches. TKM has also received funding from the World Health Organization (WHO) to attend a round table of subject matter experts for the WHO Member State Mechanism for Substandard and Falsified Medical Products Working Group on Informal Markets. TKM is also the editor-in-chief of JMIR Infodemiology. The authors report no other conflicts of interest associated with this manuscript.

Multimedia Appendix 1
Supplementary methods, supplementary additional results, adapted International Pharmaceutical Federation Checklist, additional study figures and examples.
[DOCX File , 7661 KB - jmir_v26i1e65440_app1.docx]

Multimedia Appendix 2
Liquid chromatography–mass spectrometry analysis.
[DOCX File , 46 KB - jmir_v26i1e65440_app2.docx]

References

1. Obesity and overweight. World Health Organization. 2024 Mar 1. URL: https://tinyurl.com/3mw8bsu3 [accessed 2024-10-20]
2. Obesity atlas 2023. World Obesity Federation Global Obesity Observatory. URL: https://tinyurl.com/4kd6xdfs [accessed 2023-12-27]
3. Chen YQ, Pottanat TG, Siegel RW, Ehsani M, Qian Y, Zhen EY, et al. Angiopoietin-like protein 8 differentially regulates ANGPTL3 and ANGPTL4 during postprandial partitioning of fatty acids. J Lipid Res 2020 Aug;61(8):1203-1220. [doi: 10.1194/jlr.ra120000781]
4. González-Muniesa P, Mártinez-González MA, Hu FB, Després JP, Matsuzawa Y, Loos RJ, et al. Obesity. Nat Rev Dis Primers 2017 Jun 15;3:17034. [doi: 10.1038/nrdp.2017.34] [Medline: 28617414]
5. Mendrick DL, Diehl AM, Topor LS, Dietert RR, Will Y, La Merrill MA, et al. Metabolic syndrome and associated diseases: from the bench to the clinic. Toxicol Sci 2018 Mar 01;162(1):36-42 [FREE Full text] [doi: 10.1093/toxsci/kfx233] [Medline: 29106690]
6. Colditz GA, Peterson LL. Obesity and cancer: evidence, impact, and future directions. Clin Chem 2018 Jan;64(1):154-162. [doi: 10.1373/clinchem.2017.277376] [Medline: 29038151]
7. Bessesen DH, Van Gaal LF. Progress and challenges in anti-obesity pharmacotherapy. Lancet Diabetes Endocrinol 2018 Mar;6(3):237-248. [doi: 10.1016/s2213-8587(17)30236-x]
8. Bray GA. Medical treatment of obesity: the past, the present and the future. Best Pract Res Clin Gastroenterol 2014 Aug;28(4):665-684. [doi: 10.1016/j.bpg.2014.07.015] [Medline: 25194183]
9. Colman E. Dinitrophenol and obesity: an early twentieth-century regulatory dilemma. Regul Toxicol Pharmacol 2007 Jul;48(2):115-117. [doi: 10.1016/j.yrtph.2007.03.006] [Medline: 17475379]

XSL•FO
RenderX

10. Phillips L, Singer MA. Peripheral neuropathy due to dinitrophenol used for weight loss: something old, something new. Neurology 2013 Feb 19;80(8):773-774. [doi: 10.1212/wnl.0b013e3182825367]
11. Dufayet L, Gorgiard C, Vayssette F, Barbet JP, Hoizey G, Ludes B. Death of an apprentice bodybuilder following 2,4-dinitrophenol and clenbuterol intake. Int J Legal Med 2020 May 03;134(3):1003-1006. [doi: 10.1007/s00414-020-02268-2] [Medline: 32125503]
12. Jung M, Lee SJ, Yoo SH, Kim HS. Death from 2,4-dinitrophenol poisoning: an autopsy case. Korean J Leg Med 2020 Aug;44(3):140-142. [doi: 10.7580/kjlm.2020.44.3.140]
13. Holborow A, Purnell RM, Wong JF. Beware the yellow slimming pill: fatal 2,4-dinitrophenol overdose. BMJ Case Rep 2016 Apr 04;2016:27045052 [FREE Full text] [doi: 10.1136/bcr-2016-214689] [Medline: 27045052]
14. Zack F, Blaas V, Goos C, Rentsch D, Büttner A. Death within 44 days of 2,4-dinitrophenol intake. Int J Legal Med 2016 Sep 11;130(5):1237-1241. [doi: 10.1007/s00414-016-1378-4] [Medline: 27166706]
15. Khorassani FE, Misher A, Garris S. Past and present of antiobesity agents: focus on monoamine modulators. Am J Health Syst Pharm 2015 May 01;72(9):697-706. [doi: 10.2146/ajhp140034] [Medline: 25873617]
16. Liang BA, Mackey TK, Archer-Hayes AN, Shinn LM. Illicit online marketing of lorcaserin before DEA scheduling. Obesity (Silver Spring) 2013 May 20;21(5):861-864 [FREE Full text] [doi: 10.1002/oby.20429] [Medline: 23512816]
17. Idrees Z, Cancarevic I, Huang L. FDA-approved pharmacotherapy for weight loss over the last decade. Cureus 2022 Sep;14(9):e29262 [FREE Full text] [doi: 10.7759/cureus.29262] [Medline: 36277516]
18. Tan Q, Akindehin SE, Orsso CE, Waldner RC, DiMarchi RD, Müller TD, et al. Recent advances in incretin-based pharmacotherapies for the treatment of obesity and diabetes. Front Endocrinol (Lausanne) 2022 Mar 1;13:838410 [FREE Full text] [doi: 10.3389/fendo.2022.838410] [Medline: 35299971]
19. Han SH, Safeek R, Ockerman K, Trieu N, Mars P, Klenke A, et al. Public interest in the off-label use of glucagon-like peptide 1 agonists (Ozempic) for cosmetic weight loss: a Google Trends analysis. Aesthet Surg J 2023 Dec 14;44(1):60-67. [doi: 10.1093/asj/sjad211] [Medline: 37402640]
20. O'Neill ES, Wiegmann AL, Parrella N, Pittman T, Hood K, Kurlander D. Injectable weight loss medications in plastic surgery: what we know, perioperative considerations, and recommendations for the future. Plast Reconstr Surg Glob Open 2024 Jan;12(1):e5516 [FREE Full text] [doi: 10.1097/GOX.0000000000005516] [Medline: 38268718]
21. Shortage of Ozempic (semaglutide). European Medicines Agency. 2022 Oct 19. URL: https://tinyurl.com/3mzyt54m [accessed 2024-10-21]
22. Drug shortage detail: semaglutide injection. American Society of Health-System Pharmacists. URL: https://tinyurl.com/ynbb2rrb [accessed 2024-02-04]
23. FDA drug shortages. U.S. Food and Drug Administration. 2022 Mar 31. URL: https://tinyurl.com/5yc5rjst [accessed 2024-02-04]
24. Ashraf AR, Mackey TK, Schmidt J, Kulcsár G, Vida RG, Li J, et al. Safety and risk assessment of no-prescription online semaglutide purchases. JAMA Netw Open 2024 Aug 01;7(8):e2428280 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.28280] [Medline: 39093567]
25. Mackey TK, Jarmusch AK, Xu Q, Sun K, Lu A, Aguirre S, et al. Multifactor quality and safety analysis of antimicrobial drugs sold by online pharmacies that do not require a prescription: multiphase observational, content analysis, and product evaluation study. JMIR Public Health Surveill 2022 Dec 23;8(12):e41834 [FREE Full text] [doi: 10.2196/41834] [Medline: 36563038]
26. Fittler A, Paczolai P, Ashraf AR, Pourhashemi A, Iványi P. Prevalence of poisoned Google search results of erectile dysfunction medications redirecting to illegal internet pharmacies: data analysis study. J Med Internet Res 2022 Nov 08;24(11):e38957 [FREE Full text] [doi: 10.2196/38957] [Medline: 36346655]
27. Council of Europe Convention on the counterfeiting of medical products and similar crimes involving threats to public health (CETS No. 211). Council of Europe. URL: https://tinyurl.com/3kt3xkcs [accessed 2024-08-12]
28. 21 USC chapter 9, subchapter III: prohibited acts and penalties. United States Code. URL: https://tinyurl.com/8rjp5cuc [accessed 2024-08-11]
29. FDASIA title VII drug supply chain provisions. U.S. Food & Drug Administration. 2018 Mar 28. URL: https://tinyurl.com/4c43e9h7 [accessed 2024-08-11]
30. Safety and transparency for the commercial internet. LegitScript. URL: https://www.legitscript.com/ [accessed 2024-02-09]
31. Safe pharmacy. National Association of Boards of Pharmacy. URL: https://safe.pharmacy/ [accessed 2024-02-09]
32. Vida RG, Merczel S, Jáhn E, Fittler A. Developing a framework regarding a complex risk based methodology in the evaluation of hazards associated with medicinal products sourced via the internet. Saudi Pharm J 2020 Dec;28(12):1733-1742 [FREE Full text] [doi: 10.1016/j.jsps.2020.10.018] [Medline: 33424264]
33. Merczel S, Vida RG, Tasi T, Fittler A, Botz L. Quality of dorzolamide hydrochloride and timolol maleate containing eye drops distributed online. Saudi Pharm J 2023 Jun;31(6):921-928 [FREE Full text] [doi: 10.1016/j.jsps.2023.04.018] [Medline: 37250359]
34. Vida RG, Strauss LV, Bajtel Á, Kiss T, Csupor D, Fittler A. Safety and risks of CBD oils purchased online: unveiling uncertain quality and vague health claims. Front Pharmacol 2023 Dec 14;14:1273540 [FREE Full text] [doi: 10.3389/fphar.2023.1273540] [Medline: 38192407]

XSL•FO
RenderX

35. Fittler M, Rozmer Z, Fittler A. Rampant online marketing of teeth whitening products: evaluation of online information, labelling accuracy and quantitative analysis of high peroxide content gels. Heliyon 2023 Sep;9(9):e19463 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e19463] [Medline: 37809774]
36. Tool for visual inspection of medicines. International Pharmaceutical Federation. URL: https://tinyurl.com/bdhmrbva [accessed 2024-10-21]
37. Schiavetti B, Wynendaele E, Melotte V, Van der Elst J, De Spiegeleer B, Ravinetto R. A simplified checklist for the visual inspection of finished pharmaceutical products: a way to empower frontline health workers in the fight against poor-quality medicines. J Pharm Policy Pract 2020 May 01;13(1):9 [FREE Full text] [doi: 10.1186/s40545-020-00211-9] [Medline: 32377348]
38. GLS introduces refreshed brand identity. GLS. 2021 Oct 4. URL: https://tinyurl.com/2ex3vcyn [accessed 2024-08-13]
39. Annex I: summary of product characteristics. European Medicines Agency. URL: https://tinyurl.com/587haax2 [accessed 2024-10-21]
40. Janvier S, De Spiegeleer B, Vanhee C, Deconinck E. Falsification of biotechnology drugs: current dangers and/or future disasters? J Pharm Biomed Anal 2018 Nov 30;161:175-191. [doi: 10.1016/j.jpba.2018.08.037] [Medline: 30165334]
41. Janvier S, Cheyns K, Canfyn M, Goscinny S, De Spiegeleer B, Vanhee C, et al. Impurity profiling of the most frequently encountered falsified polypeptide drugs on the Belgian market. Talanta 2018 Oct 01;188:795-807. [doi: 10.1016/j.talanta.2018.06.023] [Medline: 30029448]
42. Scannell C, Romley J, Myerson R, Goldman D, Qato DM. Prescription fills for semaglutide products by payment method. JAMA Health Forum 2024 Aug 02;5(8):e242026 [FREE Full text] [doi: 10.1001/jamahealthforum.2024.2026] [Medline: 39093590]
43. US Chem labs. U.S. Food and Drug Administration. 2024 Feb 7. URL: https://tinyurl.com/bdhhuwdu [accessed 2024-10-21]
44. Warning letter: www.semaspace.com. U.S. Food and Drug Administration. 2023 Oct 2. URL: https://tinyurl.com/ycxnwcdt [accessed 2024-01-04]
45. RogueRx activity report: injectable weight loss drugs: how illegal online drug sellers are taking advantage of patients. National Association of Boards of Pharmacy. 2024. URL: https://tinyurl.com/yaezzmf4 [accessed 2024-10-21]
46. FDA alerts health care providers, compounders and patients of dosing errors associated with compounded injectable semaglutide products. U.S. Food and Drug Administration. 2024 Jul 26. URL: https://tinyurl.com/35dc2shu [accessed 2024-09-16]
47. Kindelan K. Some people are overdosing on semaglutide, FDA warns. ABC News. 2024 Jul 30. URL: https://tinyurl.com/27tantyu [accessed 2024-09-16]
48. Whitley HP, Trujillo JM, Neumiller JJ. Special report: potential strategies for addressing GLP-1 and dual GLP-1/GIP receptor agonist shortages. Clin Diabetes 2023;41(3):467-473 [FREE Full text] [doi: 10.2337/cd23-0023] [Medline: 37456085]
49. Massy-Westropp M. What we can learn from the Ozempic shortage. The University of New South Wales, Sydney. 2023 Feb 13. URL: https://tinyurl.com/mr45cmsw [accessed 2024-05-27]
50. Arillotta D, Floresta G, Guirguis A, Corkery JM, Catalani V, Martinotti G, et al. GLP-1 receptor agonists and related mental health issues; insights from a range of social media platforms using a mixed-methods approach. Brain Sci 2023 Oct 24;13(11):1503 [FREE Full text] [doi: 10.3390/brainsci13111503] [Medline: 38002464]
51. Shi Z, Liu X, Srinivasan K. Hype news diffusion and risk of misinformation: the Oz effect in health care. J Market Res 2021 Dec 16;59(2):327-352. [doi: 10.1177/00222437211044472]
52. Ashraf AR, Mackey TK, Fittler A. Search engines and generative artificial intelligence integration: public health risks and recommendations to safeguard consumers online. JMIR Public Health Surveill 2024 Mar 21;10:e53086 [FREE Full text] [doi: 10.2196/53086] [Medline: 38512343]
53. Fittler A, Adeniye L, Katz Z, Bella R. Effect of infodemic regarding the illegal sale of medications on the internet: evaluation of demand and online availability of ivermectin during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Jul 13;18(14):7475 [FREE Full text] [doi: 10.3390/ijerph18147475] [Medline: 34299920]
54. Link MP, Hagerty K, Kantarjian HM. Chemotherapy drug shortages in the United States: genesis and potential solutions. J Clin Oncol 2012 Mar 01;30(7):692-694. [doi: 10.1200/jco.2011.41.0936]
55. Bhutada NS, Rollins BL. Disease-specific direct-to-consumer advertising of pharmaceuticals: an examination of endorser type and gender effects on consumers' attitudes and behaviors. Res Social Adm Pharm 2015 Nov;11(6):891-900. [doi: 10.1016/j.sapharm.2015.02.003] [Medline: 25797861]
56. The rise of GLP-1 influencers: celebrity endorsements and social media's role in the Ozempic weight loss craze. HealthEconomics.com. 2024 Feb 5. URL: https://tinyurl.com/5xedfu6y [accessed 2024-08-15]
57. Liang BA, Mackey TK. Vaccine shortages and suspect online pharmacy sellers. Vaccine 2012 Jan 05;30(2):105-108. [doi: 10.1016/j.vaccine.2011.11.016] [Medline: 22094281]
58. Liang BA, Mackey TK. Online availability and safety of drugs in shortage: a descriptive study of internet vendor characteristics. J Med Internet Res 2012 Feb 09;14(1):e27 [FREE Full text] [doi: 10.2196/jmir.1999] [Medline: 22321731]

XSL•FO
RenderX

59. Fittler A, Vida RG, Rádics V, Botz L. A challenge for healthcare but just another opportunity for illegitimate online sellers: dubious market of shortage oncology drugs. PLoS One 2018 Aug 28;13(8):e0203185 [FREE Full text] [doi: 10.1371/journal.pone.0203185] [Medline: 30153304]
60. Ahmed J, Modica de Mohac L, Mackey TK, Raimi-Abraham BT. A critical review on the availability of substandard and falsified medicines online: incidence, challenges and perspectives. J Med Access 2022 Feb 15;6:23992026221074548 [FREE Full text] [doi: 10.1177/23992026221074548] [Medline: 36204527]
61. Vanhee C, Janvier S, Desmedt B, Moens G, Deconinck E, De Beer JO, et al. Analysis of illegal peptide biopharmaceuticals frequently encountered by controlling agencies. Talanta 2015 Sep 01;142:1-10. [doi: 10.1016/j.talanta.2015.04.022] [Medline: 26003685]
62. Vida RG, Fittler A, Mikulka I, Ábrahám E, Sándor V, Kilár F, et al. Availability and quality of illegitimate somatropin products obtained from the internet. Int J Clin Pharm 2017 Feb;39(1):78-87. [doi: 10.1007/s11096-016-0398-y] [Medline: 27888454]

Abbreviations

FDA: Food and Drug Administration
FIP: International Pharmaceutical Federation
GLP-1: glucagon-like peptide 1
NABP: National Association of Boards of Pharmacy
SERP: search engine results page
WHO: World Health Organization

Edited by A Mavragani; submitted 15.08.24; peer-reviewed by J Hertig, A Ashenef; comments to author 12.09.24; revised version received 18.09.24; accepted 19.09.24; published 07.11.24.

Please cite as:
Ashraf AR, Mackey TK, Vida RG, Kulcsár G, Schmidt J, Balázs O, Domián BM, Li J, Csákó I, Fittler A
Multifactor Quality and Safety Analysis of Semaglutide Products Sold by Online Sellers Without a Prescription: Market Surveillance, Content Analysis, and Product Purchase Evaluation Study
J Med Internet Res 2024;26:e65440
URL: https://www.jmir.org/2024/1/e65440
doi:10.2196/65440
PMID:

©Amir Reza Ashraf, Tim Ken Mackey, Róbert György Vida, Győző Kulcsár, János Schmidt, Orsolya Balázs, Bálint Márk Domián, Jiawei Li, Ibolya Csákó, András Fittler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Big 5 Personality Traits and Individual- and Practice-Related Characteristics as Influencing Factors of Digital Maturity in General Practices: Quantitative Web-Based Survey Study

Lisa Weik[1], MSc; Leonard Fehring[2,3], MD; Achim Mortsiefer[4], Prof Dr; Sven Meister[1,5], Prof Dr

[1]Health Care Informatics, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[2]Helios University Hospital Wuppertal, Department of Gastroenterology, Witten/Herdecke University, Wuppertal, Germany

[3]Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[4]General Practice II and Patient-Centredness in Primary Care, Institute of General Practice and Primary Care, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[5]Department Healthcare, Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany

Corresponding Author:
Sven Meister, Prof Dr
Health Care Informatics
Faculty of Health, School of Medicine
Witten/Herdecke University
Pferdebachstr. 11
Witten, 58448
Germany
Phone: 49 230292678629
Email: sven.meister@uni-wh.de

Abstract

Background: Various studies propose the significance of digital maturity in ensuring effective patient care and enabling improved health outcomes, a successful digital transformation, and optimized service delivery. Although previous research has centered around inpatient health care settings, research on digital maturity in general practices is still in its infancy.

Objective: As general practitioners (GPs) are the first point of contact for most patients, we aimed to shed light on the pivotal role of GPs' inherent characteristics, especially their personality, in the digital maturity of general practices.

Methods: In the first step, we applied a sequential mixed methods approach involving a literature review and expert interviews with GPs to construct the digital maturity scale used in this study. Next, we designed a web-based survey to assess digital maturity on a 5-point Likert-type scale and analyze the relationship with relevant inherent characteristics using ANOVAs and regression analysis.

Results: Our web-based survey with 219 GPs revealed that digital maturity was overall moderate (mean 3.31, SD 0.64) and substantially associated with several characteristics inherent to the GP. We found differences in overall digital maturity based on GPs' gender, the expected future use of digital health solutions, the perceived digital affinity of medical assistants, GPs' level of digital affinity, and GPs' level of extraversion and neuroticism. In a regression model, a higher expected future use, a higher perceived digital affinity of medical assistants, a higher digital affinity of GPs, and lower neuroticism were substantial predictors of overall digital maturity.

Conclusions: Our study highlights the impact of GPs' inherent characteristics, especially their personality, on the digital maturity of general practices. By identifying these inherent influencing factors, our findings support targeted approaches to drive digital maturity in general practice settings.

(J Med Internet Res 2024;26:e52085) doi:10.2196/52085

KEYWORDS

digital health; eHealth; digital maturity; maturity assessment; general practitioners; primary care physicians; primary care; family medicine; personality; digital affinity; digital health adoption

XSL•FO
RenderX

Introduction

Background

Over the past decade, digital health solutions have garnered increasing attention owing to their transformative impact on health care systems worldwide [1]. Today, these have emerged as powerful tools in health care, revolutionizing the market by improving health outcomes [2,3], reducing costs [4], and improving the effectiveness and efficiency of health care delivery [4-6]. General practitioners (GPs) can choose various digital health solutions for their practice, ranging from remote monitoring of physiological parameters [7], video consultations [8], and mobile health apps [9] to digital appointment booking [10]. Thus, GPs must decide more than ever on their digital health agenda to successfully navigate the ever-changing landscape of digital health solutions and improve the quality of care [11].

To strategically develop a digital health agenda, it is crucial to evaluate the status quo [12]. Digital maturity assessment is a promising approach that allows evaluating health care facilities' digital status across various technological and organizational dimensions to move toward an enhanced maturity stage [13,14]. As such, digital maturity is not a stable construct but should be considered as evolving toward a theoretical end point of maturity within the current digital health landscape, describing a successful development path toward digitalization along several stages. It is not a unidimensional construct but encompasses various technological and organizational dimensions beyond digital health adoption [13]. In essence, digital maturity provides a broader and more holistic perspective compared with digital health adoption, considering not only the adoption of specific digital tools but also an organization's overall preparedness and capability to navigate the digital landscape. Accordingly, digital maturity can be considered a prerequisite to ensuring a successful digital transformation and optimized service delivery [12]. Although extensive research has studied digital maturity, its underlying dimensions [13-17], influencing factors [18], and associated outcomes [3,19] in inpatient care settings, only a few studies have undertaken similar investigations in general practice settings [20,21].

However, assessing digital maturity and its influencing factors is important in general practice settings. In European health care systems, GPs have a unique and central role in providing comprehensive and continuous health care services [22], as they are most patients' first point of contact with the health care system [23]. This pivotal position places GPs at the forefront of digital transformation. Thus, the integration of digital technologies in general practices holds the promise of positively impacting a vast number of patients. Although the benefits of digital technologies in health care are evident [1], the full realization of these advantages hinges upon the level of digital maturity achieved by health care providers such as GPs.

Objectives

Therefore, understanding the factors that drive digital maturity among GPs is vital for optimizing health care service delivery and harnessing the full potential of digital health solutions. To date, only a few studies have adopted a GP-focused approach and analyzed the influence of GP-related characteristics on digital maturity. Of these, the focus was on demographics and practice- and digital health use–related characteristics [21]. Although studies link technology adoption to personality [24-26], no study has investigated its influence on digital maturity more broadly. However, as GPs play a pivotal role in the digital maturity of their practices, we believe that analyzing the impact of their personality enables a comprehensive understanding of GP-related drivers of digital maturity, the development of targeted and effective measures to support GPs in their digitalization efforts, and the development of a dedicated and comprehensive maturity model for general practice settings. Thus, the objective of this study is to extend previous literature on influencing factors of digital maturity by shedding light on the role of GPs' inherent characteristics in the digital maturity of general practices. By inherent characteristics, we mean those qualities that describe them personally and their practice. This study specifically focused on demographics, practice-related characteristics, digital health use, digital affinity, and personality. We aimed to evaluate the relationship between these inherent characteristics and the digital maturity of general practices.

Methods

Study Design

This study used a sequential approach involving qualitative and quantitative data collection and analysis. First, we conducted a literature review following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline [27] to identify relevant constituting dimensions and indicators of digital maturity across health care settings. We then conducted expert interviews with GPs following the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [28] to validate and extend our literature review results, ensuring the relevance of the extracted indicators in general practice settings. In the next step, we designed a web-based survey following the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guideline [29] for internet surveys to assess digital maturity and relevant characteristics inherent to the GP and ultimately answer the following research question core to this study: Which demographic and practice-related characteristics, use-related variables, and personality traits substantially influence overall digital maturity?

Ethical Considerations

All research project steps were approved by the Ethics Committee of Witten/Herdecke University (number S-242/2022). GPs willing to participate in the expert interviews or the web-based survey were required to provide informed consent before participation. They did not receive any incentive for participation. All data generated during the interviews were factually anonymized, meaning the data were presented and analyzed in aggregated or enlarged forms, which prevents conclusions from being drawn about individuals or only with disproportionate effort. Furthermore, all personal data collected in our web-based survey were anonymized by applying the

XSL•FO
RenderX

k-anonymity principle, so that responses were deleted if they did not allow our anonymity criterion to apply.

Literature Review

Our literature review followed the PRISMA-ScR guideline [27] (refer to Multimedia Appendix 1 for the completed PRISMA-ScR checklist). We searched the PubMed and PsycINFO databases for our literature review to gather relevant citations. As we aimed to identify and synthesize relevant indicators of digital maturity that could be adapted to general practice settings, we developed our search terms following other literature reviews (scoping and systematic) on digital maturity in the digital health context [13,16,17,30]. Thus, our search term included common variants of *digital maturity* combined using the Boolean OR operator (details on search terms are provided in Multimedia Appendix 1, item 8).

For our abstract and full-text screening, we excluded articles if they (1) were not related to digital health; (2) were not focused on digital maturity; (3) did not report on indicators, frameworks, or assessments of digital maturity, as extracting indicators was the main focus of our literature review; (4) were focused on national health care systems or inpatient care delivery, as these contexts essentially differ from general practice settings; (5) were focused on low-income countries and were thus not comparable to the German health care context; and (6) were not original, peer-reviewed, and published full-text articles. We decided not to limit our literature review to GPs because we aimed to validate the extracted indicators in qualitative expert interviews with GPs. Evidence from the included studies was synthesized by extracting and grouping potentially relevant indicators of digital maturity in line with the framework proposed in a recent review [13].

Expert Interviews

As only limited research on digital maturity in general practice settings has been conducted to date [20], we conducted expert interviews with GPs to validate the indicators and dimensions identified in our literature review, ensure their relevance for general practices, and identify additional indicators relevant to digital maturity in general practices. The expert interviews followed the COREQ checklist for qualitative research [28] (refer to Multimedia Appendix 2 for the completed COREQ checklist). On the basis of the results of our literature review, we developed a semistructured interview guide to capture GPs' perspectives on relevant indicators of digital maturity in general practices and their evaluation of the relevance of the proposed dimensions (refer to Multimedia Appendix 3 for the translated interview guide). The questions were designed to explore four main topics: (1) experience with digital health solutions, (2) perspectives on indicators of digital maturity, (3) perceived barriers toward the adoption and use of digital health solutions, and (4) preferable strategies to improve digital health adoption. This study focuses on the first 2 topics, whereas the latter will be part of subsequent analyses published elsewhere.

We recruited participants via email. The first interviewees were 1 colleague and 4 personal contacts of the authors. Next, the participants were purposively sampled via their publicly available email addresses to represent GPs from different regions and age groups. The interviewees were neither acquainted with each other nor were they in direct contact. Before the interview, participants received information about the overarching research design, the research question, and a broad description of the topics to be covered in the interview. They were also asked to provide informed consent. We then conducted the interviews in an internet-based one-on-one setting, video recorded them, and transcribed them verbatim to allow for subsequent qualitative analysis.

Data saturation was achieved after 10 interviews with the GPs. On average, the interviews lasted 45 minutes. Interview transcripts were then coded and analyzed using VERBI GmbH MAXQDA 2022 [31]. For our content analysis [32], we derived our coding scheme deductively based on the literature review results to allow for subsequent comparison. In addition, we derived themes inductively from the interview material, if mentioned by >1 interviewee, to provide additional insights. On the basis of this, we extracted a numeric score representing the number of interviews in which specific indicators of digital maturity were mentioned. These results were then compared with the results of our literature review to develop items for assessing digital maturity as part of our web-based survey.

Web-Based Survey

Survey Design

The final and core part of our study was a cross-sectional survey among a convenience sample of GPs in Germany, investigating inherent characteristics and their digital maturity. The survey followed the CHERRIES checklist for internet surveys [29] (refer to Multimedia Appendix 4 for the completed CHERRIES; refer to Multimedia Appendix 5 for the survey questionnaire). Overall, the survey consisted of six sections: (1) demographics, practice-related characteristics, and digital health use; (2) GPs' digital affinity; (3) big 5 personality traits; (4) digital maturity of the practice; (5) perceived barriers to digital health adoption and use; and (6) potential measures to support digital health adoption. This paper discusses only the results of sections 1 to 4, as our goal was to explore the association between overall digital maturity and personality characteristics. An introductory page informed the participants about the research purpose, goals, target population, length of the survey, and institutional review board approval. On the next page, we informed the participants about the data security and storage policies and the researchers involved. The participants had to provide informed consent to continue with the survey.

Measures

To capture the participants' demographics and practice-related characteristics, we used single-choice questions concerning their gender, age range, practice location, professional experience, patient population, and type of practice. We further assessed digital health use (current use and expected future use of digital health solutions) and the perceived digital affinity of medical assistants using 5-point Likert-type scales.

We used the German version of the Affinity for Technology Interaction scale [33] to capture GPs' digital affinity. The 9-item scale captures a person's tendency to actively engage in intensive technology interaction on a 6-point Likert-type scale.

XSL•FO
RenderX

We assessed the GPs' personality using a German-language scale [34] (Big Five Inventory–Short Version). The 21-item scale assesses the big 5 personality traits of extraversion, agreeableness, conscientiousness, neuroticism, and openness on a 5-point Likert-type scale.

The 28 items for our digital maturity assessment were developed based on the results of our literature review and expert interviews. We used items for all digital maturity indicators proposed by >3 theoretical maturity models in our literature review or mentioned by >1 interviewee in our expert interviews to ensure theoretical and expert consensus. For all other indicators, we consulted an expert in the field of digital maturity before excluding them from the survey to ensure that all key indicators were covered. The items were then developed based on the theoretical descriptions of the corresponding digital maturity indicators identified in our literature review. Although some indicators could be covered in their entirety by 1 item (eg, *Change Management*), others were more complex in nature, thus requiring more items to cover them fully (eg, the indicator *Systems and Services* was captured by 5 items to cover various digital health solutions). All items were discussed and iteratively refined between the authors. In total, the items covered 7 constituting dimensions of digital maturity: *Governance and Management*, *IT Capability*, *People, Skills and Behavior*, *Interoperability*, *Strategy*, *Data Analytics*, and *Patient-centered Care*. Participants were asked to rate their agreement with the items on a 5-point Likert-type scale.

Pretest and Recruitment

To test the final questionnaire for clarity, comprehensiveness, usability, and technical functionality, we conducted a pretest with 15 physicians working in ambulatory care settings. On the basis of the pretest, we refined our welcome message and the wording of some items. The survey was then conducted between April and August 2023. We disseminated the final survey via various recruitment channels. These included interview participants and personal contacts of the researchers, teaching practices of affiliated universities, social networks such as LinkedIn, physician networks, research practice networks, and mailing lists. Personal contacts and social networks were addressed directly by the authors. For teaching practices, physician networks, research practice networks, and mailing lists, we initially contacted the respective person responsible via mail. The person responsible then disseminated the survey access link within the network via their communication channels and announcements. The survey was conducted in an open-access mode, meaning anyone with an access link could participate. In addition, we did not track which invited participants had started or completed the survey, thus limiting our ability to use reminders. We also did not provide incentives to participants. The final survey took approximately 10 to 15 minutes to complete.

Data Cleaning and Analysis

Before our statistical analysis, we performed thorough cleaning of the data obtained. Following common procedures [35], our data cleaning included the removal of responses without informed consent, incomplete responses, and duplicate entries, leading to 111 of the initial 354 responses (3.1%) being excluded. In addition, we applied data quality control procedures in line with practical recommendations [36]. Therefore, we excluded responses with very low completion times below our threshold of the fastest 5.3% (12/243) of respondents. In addition, we excluded respondents who showed careless answer behavior across multiple survey pages and thus in >1 item battery, that is, who chose the very same answer option for all items in >1 item battery, as this might indicate careless responding as opposed to straightlining owing to respondents' actual views [36]. We further removed all responses violating our k≥5 anonymity criterion to comply with data privacy regulations as part of the institutional review board approval. Consequently, we removed all questionnaires with <5 respondents per demographic item answer option. Ultimately, from the 354 participants who initially clicked on the survey link, responses from 219 (61.9%) respondents were included in our analysis (Figure 1).

We first computed the mean value for respondents' digital affinity, personality traits, and digital maturity dimensions to analyze the association between digital maturity and GPs' inherent characteristics. As we focused our statistical analysis on digital maturity more broadly instead of individual dimensions, we additionally computed an overall digital maturity score as the average score across all items. All subsequent statistical analyses were conducted with SPSS version 29.0 (IBM Corp) for Macintosh [37].

For all scales, Cronbach α [38] was used to assess the internal consistency of the scales. An overview can be found in Table 1. The internal consistency (Cronbach α=.92) for the Affinity for Technology Interaction scale is in line with previous research [33] and can be considered excellent. Internal consistencies for the big 5 personality traits can be regarded as acceptable or good. In comparison with previous literature [34], the internal consistencies in our sample differed. In particular, the internal consistency for conscientiousness was lower than that in the original study [34]. The overall high conscientiousness and low variability of the score in this sample might explain this. GPs constitute a relatively homogeneous sample for conscientiousness, which tends to lower the Cronbach α estimates. As we derived our digital maturity scale based on our literature review and qualitative interviews, no comparison with previous literature is possible. Internal consistencies can be considered acceptable for all subscales with >2 items. Only the Interoperability subscale had a Cronbach α that was well below the .70 threshold, but as this subscale consists of only 2 items, an interpretation of the internal consistency cannot be made [38].

XSL•FO
RenderX

Figure 1. Overview of the data cleaning approach. Straightlining was monitored as part of our data quality control procedure, excluding respondents who showed a straight line across >2 survey pages.

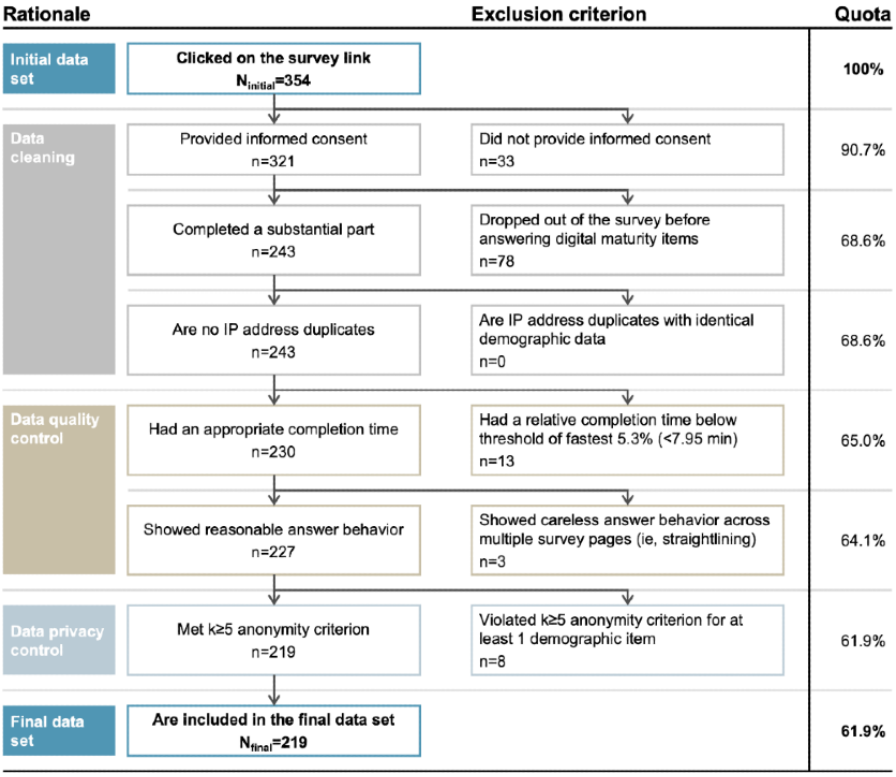

Table 1. Sample size, mean, SD, number of items, and internal consistencies of the scales used.

Scale	Values, n (%)	Values, mean (SD)	Items, n	Cronbach α (original)	Cronbach α (current)
Digital affinity (ATI[a] scale)	219 (100)	3.64 (1.20)	9	.89	.92
Personality (BFI-K[b] scale)					
Extraversion	219 (100)	3.62 (0.80)	4	.86	.81
Agreeableness	219 (100)	3.53 (0.76)	4	.64	.68
Conscientiousness	219 (100)	4.10 (0.58)	4	.70	.63
Neuroticism	219 (100)	2.42 (0.70)	4	.74	.71
Openness	219 (100)	3.81 (0.69)	5	.66	.74
Digital maturity					
Governance and management	218 (99.5)	3.97 (0.74)	6	__[c]	.87
IT capability	219 (100)	2.76 (0.73)	9	__[c]	.70
People, skills and behavior	219 (100)	4.00 (0.63)	4	__[c]	.71
Interoperability	219 (100)	3.60 (0.92)	2	__[c]	.54[d]
Strategy	219 (100)	3.51 (0.94)	2	__[c]	.69[d]
Data analytics	219 (100)	3.17 (1.22)	2	__[c]	.89[d]
Patient-centered care	219 (100)	2.18 (0.94)	3	__[c]	.71

[a]ATI: Affinity for Technology Interaction.

[b]BFI-K: Big Five Inventory–Short Version.

[c]All digital maturity subscales were developed based on the literature review and expert interviews. Thus, Cronbach α for the original studies cannot be shown.

[d]As the subscales *Interoperability*, *Strategy*, and *Data Analytics* only comprise 2 items, an interpretation of their internal consistencies should not be made.

XSL•FO
RenderX

We conducted separate ANOVAs with 2-tailed significance ($P<.05$) to assess differences in overall digital maturity, given the several inherent characteristics relevant to our study. For some of our independent variables (GPs' inherent characteristics), Q-Q plots and Shapiro-Wilk test showed that the overall digital maturity score did not follow a normal distribution for all levels. Levene test further showed heteroscedasticity in some instances. Following practical recommendations [39], we used Welch F [40] for all ANOVAs to adjust for the heteroscedasticity in the mentioned instances. Where ANOVAs revealed a significant difference ($P<.05$) in overall digital maturity, we investigated post hoc procedures. As sample sizes vary across groups, we used Hochberg GT2 (homogeneity of variance met) or Games-Howell (homogeneity of variance not met) as post hoc procedures[39]. Although we aimed to assess differences in overall digital maturity based on digital affinity and personality, we additionally clustered participants into 3 categories (low, moderate, and high) for each variable based on the underlying Likert-type scales. Thus, participants scoring in the bottom third of the Likert-type scale were categorized as low, those in the second third as moderate, and those in the upper third as high. As this categorization does not cover the whole spectrum of the continuous underlying variable, it was only used in our ANOVAs as an initial indicator for differences in these variables between GPs. These differences were then analyzed more specifically in a linear regression model using the continuous variables without categorization.

We conducted a linear regression analysis to deepen our understanding of the relationship between overall digital maturity and the independent variables. We followed a hierarchical approach for entering variables into our model to determine the influence of our demographic and practice-related variables on overall digital maturity and to separate this from the influence of digital health use, digital affinity, and personality. All nominal or ordinal variables—age, practice location size, professional experience, practice type, and current use—were dummy coded before being entered into the model. Before entering the predictors into the model, we assessed their potential multicollinearity using variance inflation factor and tolerance values following practical recommendations [39]. As all variance inflation factor values were <10 and tolerance values were >0.1, multicollinearity does not seem to flaw our analysis. In the first stage of our approach, we only included demographics and practice-related characteristics, that is, gender, age, practice location size, professional experience, and practice type. The second stage additionally included variables related to digital health use, that is, the current use and the expected future use. In model 3, we added digital affinity–related variables, that is, the perceived digital affinity of medical assistants and GPs. For the final model, we also included personality traits, that is, extraversion, agreeableness, conscientiousness, neuroticism, and openness. We chose the described sequence of insertion based on prior research and theoretical reasoning, with variables that were analyzed in previous research entered earlier into the model and the blocks, each covering different categories of inherent variables, that is, demographics and practice-related characteristics, digital health use, digital affinity, and personality.

Results

Characterizing Digital Maturity in General Practices (Literature Review and Expert Interview Results)

We narrowed the 554 initially identified citations to studies published in English or German between January 2018 and December 2022 to account for more recent research findings. For these, we conducted abstract screening based on the predefined inclusion criteria, leading to 55 potentially relevant articles being retained. Next, we performed a full-text review to determine eligibility, resulting in 21 papers [3,13-16,41-56] being included after the screening. A detailed overview of the screening process is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Flowchart for the literature review following PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Records removed for other reasons shows records removed based on our language and publication date criteria.

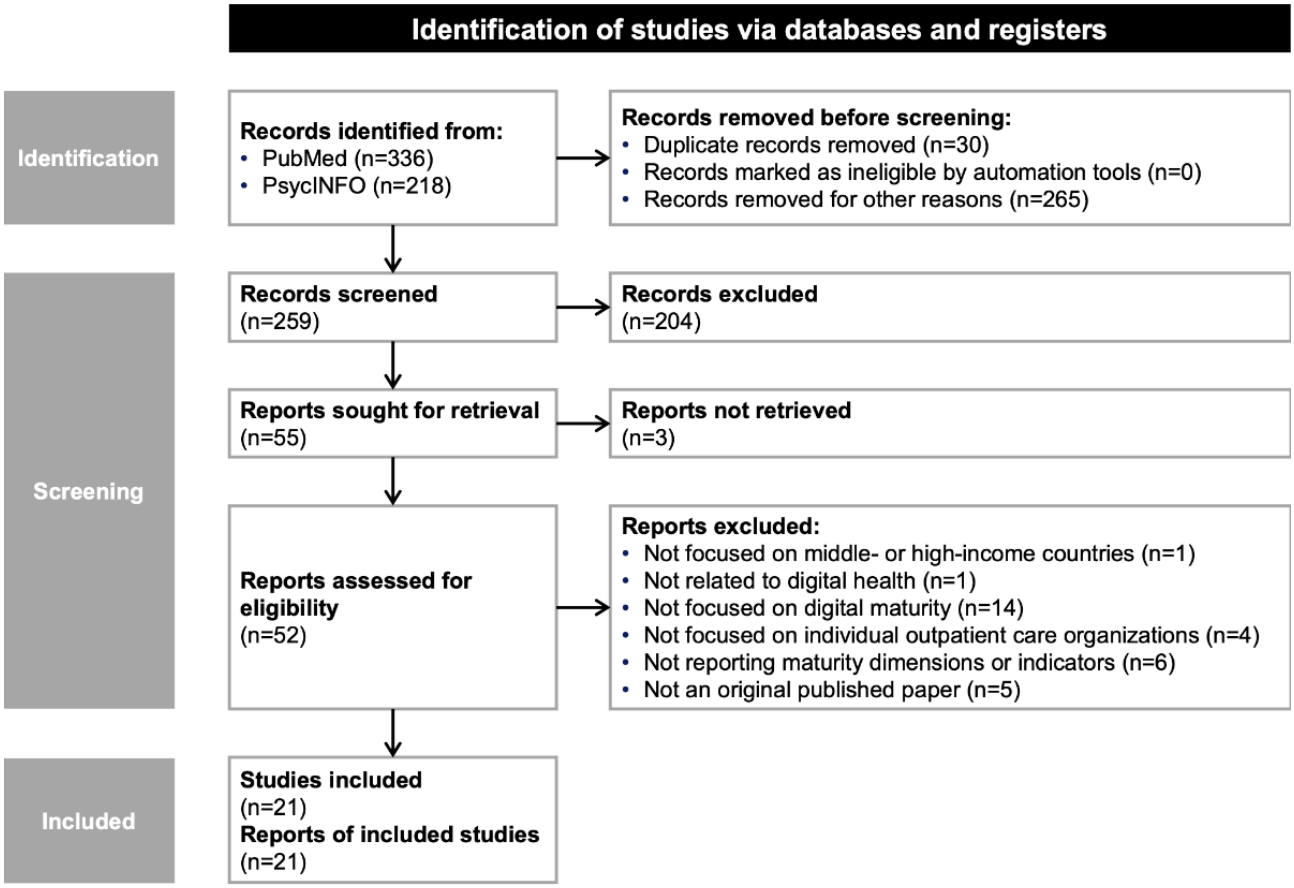

A total of 17 distinct maturity models were presented and used in the studies, with only around half looking at people- and skill-related indicators of digital maturity [13,16,43,44,46-48,50,53,56] (n=10, 58%). Of these, technology use (8/10, 80%) was the most cited indicator of digital maturity [13,16,43,47,48,50,53,56]. All other indicators in this dimension were only present in 3 (18%) of the 17 maturity models, covering the aspects of education and training [13,43,44], knowledge management [13,43,46], and individual competence [13,43,47]. The literature review identified the following 7 constituting dimensions of digital maturity: *Governance and Management*, *IT Capability*, *People*, *Skills and Behavior*, *Interoperability*, *Strategy*, *Data Analytics*, and *Patient-centered Care*.

Participants in our expert interviews were aged between 32 and 68 (mean 53.1, SD 12.7) years and worked as GPs for 1 to 36 (mean 17.7, SD 12.1) years in cities with approximately 4000 to 600,000 inhabitants. Four GPs worked in a single practice, 5 in a group practice, and 1 in a medical care center. The results of our expert interviews validated the relevance of all 7 constituting dimensions in general practice settings. The interviews revealed that GPs especially perceived the availability of digital systems and tools and their quality as relevant indicators of digital maturity. In line with a high estimated relevance (8.6/10.0), people- and skill-related indicators were mentioned by almost all interviewees (9/10, 90%). The dimension of *Patient-centered Care* was rated as least relevant concerning the digital maturity of practices (6.5/10.0). In our subsequent web-based survey, we included items for 22 digital maturity indicators based on the aforementioned inclusion criteria. The *Patient Empowerment* indicator was included after expert consensus, as it was considered a core value proposition of digital health solutions. An overview of the included digital maturity dimensions and indicators based on the synthesized results of our literature review and expert interviews is presented in Figure 3.

XSL•FO
RenderX

Figure 3. Overview of the constituting digital maturity dimensions and indicators based on the literature review and expert interview results. nMM represents the number of maturity models proposing the respective indicator; nEI shows the number of expert interviews in which the indicator was mentioned. All indicators proposed in >3 maturity models or mentioned in >1 interview were included in the survey. Light grey boxes with italic text show indicators not included in the subsequent web-based survey. Individual competence was not included as we assessed general practitioners' digital affinity separately; Patient empowerment was included owing to expert consensus.

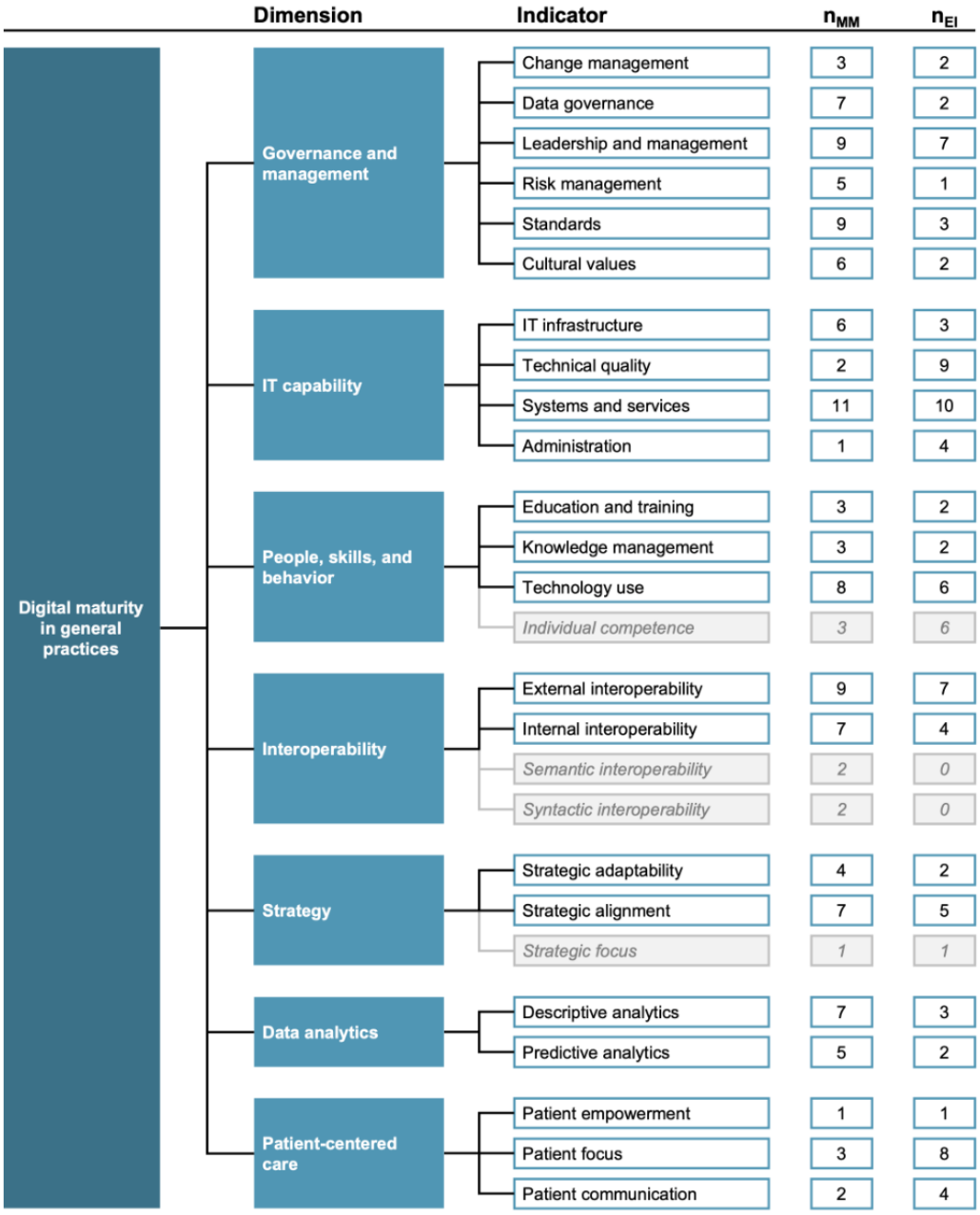

Exploring Influencing Factors of Overall Digital Maturity in General Practices (Web-Based Survey Results)

As we aimed to analyze potentially relevant influencing factors inherent to GPs, our web-based survey specifically assessed four areas of inherent characteristics: (1) demographics and practice-related characteristics, (2) digital health use, (3) digital affinity, and (4) personality.

GPs' Inherent Characteristics and Digital Maturity

Demographics, practice-related characteristics, and the digital health use of the participating GPs are summarized in Table 2. After data cleaning, quality control, and privacy control, the final sample consisted of 219 GPs, covering approximately 0.4% (219/55,112) of the German GP population [57]. The respondents were predominantly women (122/219, 55.7%), with a median age of 46 to 55 years. Approximately half of the respondents worked in relatively small town areas, either in towns with 5000 to 20,000 inhabitants (67/219, 30.6%) or in smaller villages with <5000 inhabitants (45/219, 20.5%).

XSL•FO
RenderX

Table 2. Characteristics of the participating general practitioners (N=219).

Characteristic and category	Values, n (%)
Gender	
Man	97 (44.3)
Woman	122 (55.7)
Nonbinary	0 (0)
No Answer	0 (0)
Age (y)	
<26	0 (0)
26-35	21 (9.6)
36-45	60 (27.4)
46-55	55 (25.1)
56-65	73 (33.3)
>65	10 (4.6)
Practice location size (number of inhabitants)	
<5000	45 (20.5)
5000-20,000	67 (30.6)
20,001-100,000	44 (20.1)
100,001-500,000	28 (12.8)
>500,000	35 (16)
Professional experience (y)	
<1	0 (0)
1-5	20 (9.1)
6-10	35 (16)
11-20	59 (26.9)
21-30	60 (27.4)
>30	45 (20.5)
Patient population	
Only privately health-insured	0 (0)
Only statutory health-insured	0 (0)
Both	219 (100)
Practice type	
Single practice	102 (46.5)
Practice sharing	12 (5.5)
Group practice	93 (42.5)
Practice clinic	0 (0)
Practice network	0 (0)
Medical care center	12 (5.5)
Collaborative laboratory	0 (0)
Current use of digital health solutions	
Never	18 (8.2)
Less than once per month	48 (21)
Monthly	24 (11)
Weekly	32 (14.6)

XSL•FO
RenderX

Characteristic and category	Values, n (%)
Daily	99 (45.2)
Expected future use of digital health solutions	
Very unlikely	10 (4.6)
Rather unlikely	30 (13.7)
Neither unlikely nor likely	15 (6.8)
Rather likely	48 (21.9)
Very likely	116 (53)
Perceived digital affinity of medical assistants	
Not at all digitally savvy	3 (1.4)
Rather not digitally savvy	43 (19.6)
Neither not digitally savvy nor digitally savvy	55 (25.1)
Rather digitally savvy	93 (42.5)
Very digitally savvy	25 (11.4)

Most respondents (164/219, 74.9%) had >10 years of professional experience . They predominantly worked in either single practices (102/219, 46.5%) or group practices (93/219, 42.5%). All respondents (219/219, 100%) treated privately and statutory health-insured patients .

Although approximately half of the respondents (99/219, 45.2%) used digital health solutions daily, almost one-third (64/219, 29.2%) did not use them at all or seldom. However, most respondents (116/219, 53%) were very likely to use digital health solutions in the future.

With regard to digital affinity, most respondents perceived their medical assistants to be moderately (55/219, 25.1%) or rather digitally savvy (93/219, 42.5%) and had a relatively moderate digital affinity [33] (mean 3.64, SD 1.20) themselves.

Regarding personality [34], on average, respondents showed high conscientiousness (mean 4.10, SD 0.58) and openness (mean 3.81, SD 0.69), moderate extraversion (mean 3.62, SD 0.80) and agreeableness (mean 3.53, SD 0.76), and low to moderate neuroticism (mean 2.42, SD 0.70).

Our sample's overall digital maturity was moderate (mean 3.31, SD 0.64) based on the 5-point Likert-type scale used. Among the 7 constituting dimensions analyzed in this study, *People, Skills, and Behavior* (mean 4.00, SD 0.63) and *Governance and Management* (mean 3.97, SD 0.74) registered the highest average scores, followed by moderate scores for *Interoperability* (mean 3.60, SD 0.92), *Strategy* (mean 3.51, SD 0.94), *Data Analytics* (mean 3.17, SD 1.22), and *IT Capability* (mean 2.76, SD 0.73). Interestingly, *Patient-centered Care* received the lowest average value (mean 2.18, SD 0.94). For most individual digital maturity items, scores were moderate to high, with the highest ratings for *data security and confidentiality* (mean 4.44, SD 0.72; median 5.00, IQR 1.00) and *hardware and network resources* (mean 4.24, SD 0.97; median 5.00, IQR 1.00). A detailed overview of the individual item scores is provided in Figure 4.

XSL•FO
RenderX

Figure 4. Frequencies, mean, and SD for digital maturity items along the constituting dimensions assessed (N=219). The bar chart shows the number of respondents per answer category. DA: data analytics; GM: governance and management; IO: interoperability; ITC: IT capability; PCC: patient-centered care; PSB: people, skills, and behavior; ST: strategy.

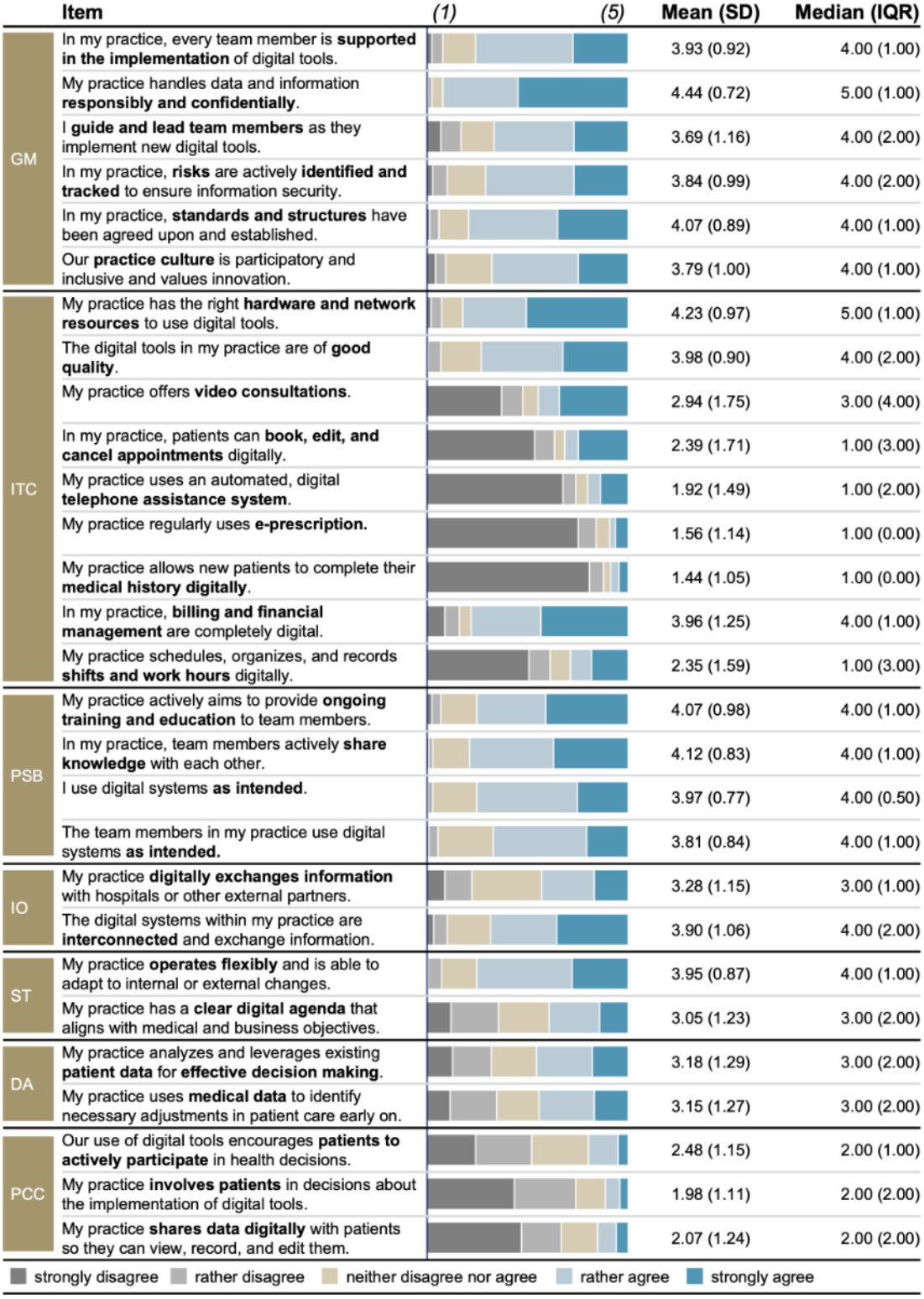

Comparing Differences in Overall Digital Maturity Based on GPs' Inherent Characteristics (ANOVA Results)

We found significant omnibus differences when comparing overall digital maturity scores based on the different areas of GPs' inherent characteristics (refer to Table 3 for ANOVA results and Table 4 for post hoc test results).

XSL•FO
RenderX

Table 3. ANOVAs for overall digital maturity (N=219).

Variable	Welch F (df_1, df_2)	P value
Demographics and practice-related characteristics		
Gender[a]	2.362 (217)	.01
Age	1.341 (4,46.40)	.27
Practice location size	0.256 (4,94.10)	.91
Professional experience	1.557 (4,80.48)	.19
Practice type	0.511 (3,29.31)	.68
Digital health use		
Current use	2.200 (4,65.41)	.08
Future use	2.843 (4,39.40)	.04
Digital affinity		
MAs'[b] digital affinity	4.813 (4,14.39)	.01
GPs'[c] level of digital affinity	8.504 (2,103.66)	.001
Personality		
Level of extraversion	4.753 (2,62.41)	.01
Level of agreeableness	0.087 (2,42.89)	.92
Level of conscientiousness	0.537 (2,2.63)	.64
Level of neuroticism	7.552 (2,30.04)	.002
Level of openness	0.073 (2,16.56)	.93

[a]As gender is a dichotomous variable, we conducted a 2-tailed t test instead of an ANOVA, with the results showing the t statistic (in the column *Welch F*), df, and P value.

[b]MA: medical assistant.

[C]GP: general practitioner.

In our sample, overall digital maturity differed significantly between men and women (t_{217}=2.362; P=.01), with men reporting a higher digital maturity (mean 3.42, SD 0.66; Cohen d=0.32). We did not find substantial differences in overall digital maturity for age, practice location size, professional experience, or practice types.

In addition, there was no substantial difference in overall digital maturity based on the current use of digital health solutions. We found significant differences in overall digital maturity based on the expected future use of digital health solutions (Welch $F_{4,39.40}$=2.843; P=.04). Respondents who expected to very likely use digital health solutions in the future had a higher overall digital maturity (mean 3.45, SD 0.66) compared with respondents reporting a rather low likelihood of future use (mean 3.10, SD 0.70). However, this difference was not statistically significant (P=.07; Cohen d=−0.52).

Furthermore, we found significant differences in overall digital maturity based on both the perceived digital affinity of medical assistants (Welch $F_{4,14.39}$=4.813; P=.01) and the level of respondents' digital affinity (Welch $F_{2,103.66}$=8.504; P=.001). Respondents who perceived their medical assistants to be rather digitally savvy (mean 3.37, SD 0.65; P=.01; Cohen d=−0.59) or very digitally savvy (mean 3.69, SD 0.70; P=.001; Cohen d=−1.12) had a significantly higher overall digital maturity compared with respondents who perceived their medical assistants to be somewhat not digitally savvy (mean 3.00, SD 0.56). In addition, the overall digital maturity was significantly higher for respondents with a high level of digital affinity (mean 3.57, SD 0.71) compared with those with moderate (mean 3.26, SD 0.54; P=.004; Cohen d=−0.50) or low levels (mean 3.07, SD 0.57; P=.001; Cohen d=−0.76).

Regarding personality, we found significant differences in overall digital maturity based on the level of extraversion (Welch $F_{2,62.41}$=4.753; P=.01) and neuroticism (Welch $F_{2,30.04}$=7.552; P=.002). Hochberg GT2 post hoc tests revealed that the overall digital maturity was significantly higher for respondents with high levels of extraversion (mean 3.43, SD 0.65) compared with respondents with moderate levels of extraversion (mean 3.17, SD 0.62; P=.01; Cohen d=−0.41). In addition, respondents with low levels of neuroticism had a significantly higher overall digital maturity (mean 3.45, SD 0.62) compared with those with moderate (mean 3.20, SD 0.60; P=.009; Cohen d=0.41) or high levels of neuroticism (mean 2.82, SD 0.70; P=.002; Cohen d=1.00).

XSL•FO
RenderX

Table 4. Post hoc tests for significant ANOVAs for digital maturity (N=219).

Variable and category	Values, n (%)	Values, mean (SD)	Group comparison	*P* value	Cohen *d*
Gender					
Men	97 (44.3)	3.42 (0.66)	Vs women	.01	0.32
Women	122 (55.7)	3.22 (0.61)	N/A[a]	N/A	N/A
Future use					
Very unlikely	10 (4.6)	3.09 (0.68)	Vs rather unlikely	.99	−0.01
Very unlikely	10 (4.6)	3.09 (0.68)	Vs neither unlikely nor likely	.99	−0.05
Very unlikely	10 (4.6)	3.09 (0.68)	Vs rather likely	.99	−0.24
Very unlikely	10 (4.6)	3.09 (0.68)	Vs very likely	.57	−0.54
Rather unlikely	30 (13.7)	3.10 (0.70)	Vs neither unlikely nor likely	.99	−0.03
Rather unlikely	30 (13.7)	3.10 (0.70)	Vs rather likely	.99	−0.21
Rather unlikely	30 (13.7)	3.10 (0.70)	Vs very likely	.07	−0.52
Neither unlikely nor likely	15 (6.8)	3.12 (0.54)	Vs rather likely	.99	−0.20
Neither unlikely nor likely	15 (6.8)	3.12 (0.54)	Vs very likely	.45	−0.20
Rather likely	48 (21.9)	3.22 (0.50)	Vs very likely	.33	−0.37
Very likely	116 (53)	3.45 (0.66)	N/A	N/A	N/A
MA[b] digital affinity					
Not at all digitally savvy	3 (1.4)	2.72 (0.86)	Vs rather not digitally savvy	.99	−0.49
Not at all digitally savvy	3 (1.4)	2.72 (0.86)	Vs neither not digitally savvy nor digitally savvy	.66	−1.14
Not at all digitally savvy	3 (1.4)	2.72 (0.86)	Vs rather digitally savvy	.51	−0.99
Not at all digitally savvy	3 (1.4)	2.72 (0.86)	Vs very digitally savvy	.09	−1.36
Rather not digitally savvy	43 (19.6)	3.00 (0.56)	Vs neither not digitally savvy nor digitally savvy	.14	−0.59
Rather not digitally savvy	43 (19.6)	3.00 (0.56)	Vs rather digitally savvy	.01	−0.59
Rather not digitally savvy	43 (19.6)	3.00 (0.56)	Vs very digitally savvy	<.001	−1.12
Neither not digitally savvy nor digitally savvy	55 (25.1)	3.31 (0.50)	Vs rather digitally savvy	1.00	−0.10
Neither not digitally savvy nor digitally savvy	55 (25.1)	3.31 (0.50)	Vs very digitally savvy	.10	−0.67
Rather digitally savvy	93 (42.5)	3.37 (0.65)	Vs very digitally savvy	.20	−0.48
Very digitally savvy	25 (11.4)	3.69 (0.70)	N/A	N/A	N/A
GPs'[c] level of digital affinity					
Low digital affinity	44 (20.1)	3.07 (0.57)	Vs moderate digital affinity	.21	−0.33
Low digital affinity	44 (20.1)	3.07 (0.57)	Vs high digital affinity	<.001	−0.76
Moderate digital affinity	112 (51.1)	3.26 (0.57)	Vs high digital affinity	.004	−0.50
High digital affinity	63 (28.8)	3.57 (0.71)	N/A	N/A	N/A
Level of extraversion					
Low level of extraversion	21 (9.6)	3.17 (0.48)	Vs moderate level of extraversion	1.00	0.00
Low level of extraversion	21 (9.6)	3.17 (0.48)	Vs high level of extraversion	.22	−0.41
Moderate level of extraversion	82 (37.4)	3.17 (0.62)	Vs high level of extraversion	.01	−0.41
High level of extraversion	116 (53)	3.43 (0.65)	N/A	N/A	N/A
Level of neuroticism					
Low level of neuroticism	114 (52.1)	3.45 (0.62)	Vs moderate level of neuroticism	.009	0.41

XSL•FO
RenderX

Variable and category	Values, n (%)	Values, mean (SD)	Group comparison	*P* value	Cohen *d*
Low level of neuroticism	114 (52.1)	3.45 (0.62)	Vs high level of neuroticism	.002	1.00
Moderate level of neuroticism	93 (42.5)	3.20 (0.60)	Vs high level of neuroticism	.13	0.62
High level of neuroticism	12 (5.5)	2.82 (0.70)	N/A	N/A	N/A

[a]N/A: not applicable.

[b]MA: medical assistant.

[c]GP: general practitioner.

Exploring the Relationship Between Overall Digital Maturity and GPs' Inherent Characteristics (Regression Model Results)

An overview of our regression model can be found in Table 5.

When only including demographics and practice-related characteristics, our model was not significant ($F_{16,202}$=0.865; P=.61; R^2=0.064). Including digital health use in our model yielded a significant improvement, but the model remained insignificant ($F_{21,197}$=1.274; P=.20; R^2=0.120; ΔR^2 =0.056; ΔP=.03). However, when including digital affinity–related variables in stage 3, our model reached statistical significance ($F_{23,195}$=2.676; P=.001). Adding these variables to the model significantly increased the proportion of the explained criterion variance to 24% (ΔP=.001). Including personality-related variables in our final model led to a significant increase in R^2 of approximately 7% (ΔP=.003) to an overall R^2 of 31% ($F_{28,190}$=3.029; P=.001).

In our final regression model, 4 variables were significantly associated with overall digital maturity (refer to Figure 5 for a visualization of the final regression model; a detailed overview of the coefficients for all models can be found in Multimedia Appendix 6). We found a significant association between overall digital maturity and the expected future use of digital health solutions (b=0.105, SE 0.051; β=.205; P=.04), the perceived digital affinity of medical assistants (b=0.1147, SE 0.044; β=.226; P=.001), respondents' digital affinity (b=0.145, SE 0.041; β=.250; P=.001), and neuroticism (b=−0.216, SE 0.065; β=−0.238; P=.001). Overall, a higher expected likelihood of future digital health use, a higher digital affinity of medical assistants, a higher digital affinity of GPs, and lower neuroticism were associated with higher levels of overall digital maturity.

Table 5. Model parameters of the regression model predicting overall digital maturity (N=219).

Model number	Included variables[a]	F (df_M,df_R)	*P* value	R^2	ΔR^2	ΔP value
1	Demographics and practice-related characteristics	0.865 (16,202)	.61	0.064	0.064	.61
2	Demographics, practice-related characteristics, and digital health use	1.274 (21,197)	.20	0.120	0.056	.03
3	Demographics, practice-related characteristics, digital health use, and digital affinity	2.676 (23,195)	.001	0.240	0.120	.001
4	Demographics, practice-related characteristics, digital health use, digital affinity, and personality	3.029 (28,190)	.001	0.309	0.069	.003

[a]The models with their respective variables that were included in the different stages of our approach are as follows: model 1 (gender, age dummy coded, practice location size dummy coded, professional experience dummy coded, and practice type dummy coded), model 2 (model 1 variables, current use dummy coded, and expected future use); model 3 (model 2 variables, perceived digital affinity of medical assistants, and general practitioners' digital affinity) and model 4 (model 3 variables, extraversion, agreeableness, conscientiousness, neuroticism, and openness).

XSL•FO
RenderX

Figure 5. Standardized coefficients for the final regression model predicting overall digital maturity (N=219). Age, practice location size, professional experience, practice type, and current use were dummy coded for the analysis. Age: 36-45, 46-55, 56-65, >65 versus 26-35 years (reference category). Practice location size: 5000-20,000, 20,001-100,000, 100,001-500,000, >500,000 versus 5000 inhabitants (reference category). Professional experience: 6-10, 11-20, 21-30, >30 versus 1-5 years of experience (reference category). Practice type: Practice sharing, group practice, medical care center versus single practice (reference category). Current use of digital health solutions: Less than once per month, monthly, weekly, daily versus never (reference category).

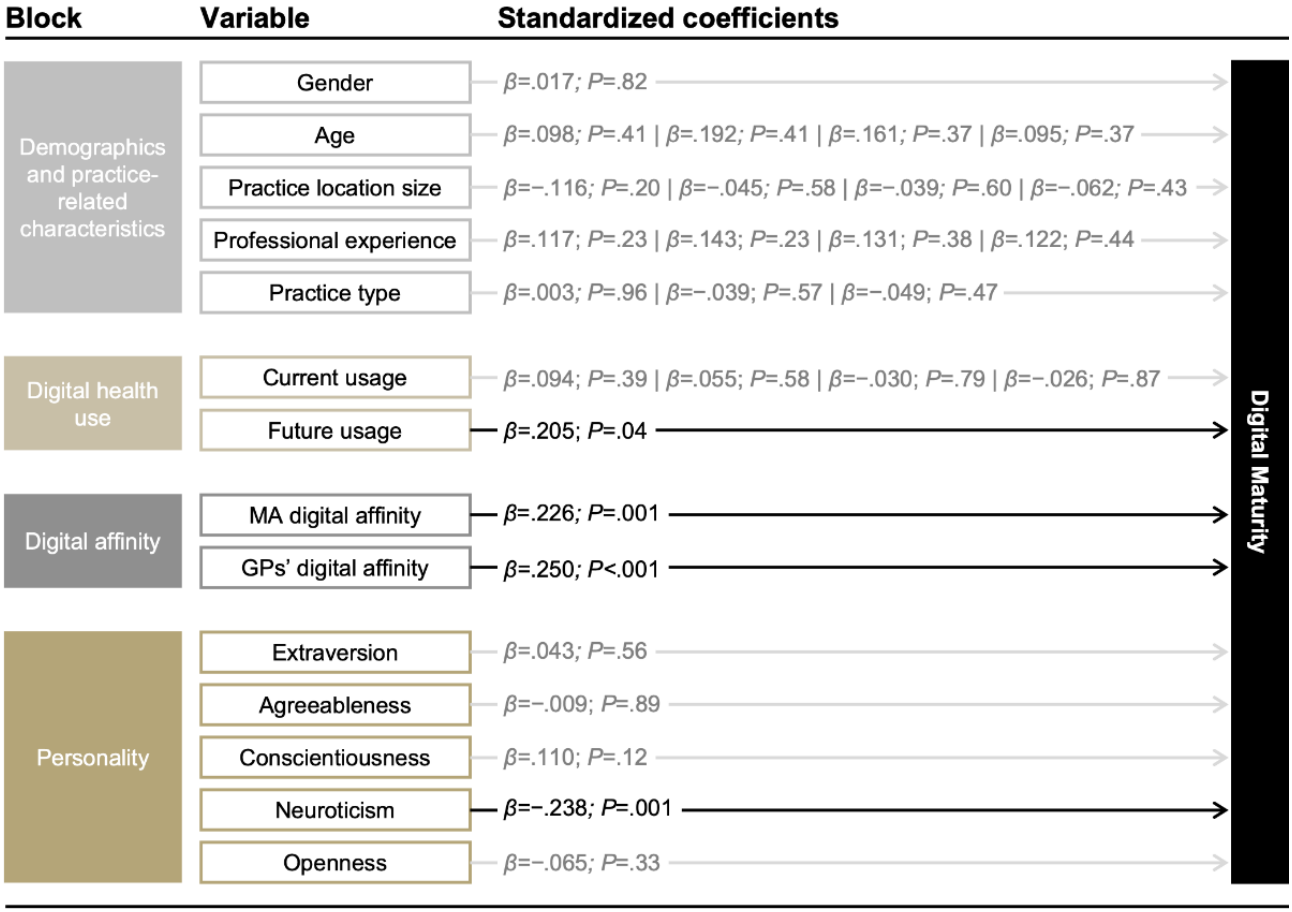

Discussion

Principal Findings

The significance of digital maturity in ensuring a successful digital transformation and optimized service delivery has been proposed in numerous maturity models and corresponding studies[13,16,43,47,55]. However, limited research focuses on digital maturity in general practice settings [20,21]. Thus, this study sought to investigate digital maturity in general practices and specifically evaluate the relationship with GPs' inherent characteristics. In our study, digital maturity was moderate and associated with several inherent characteristics. In our ANOVAs, we found differences in digital maturity based on gender, the expected future use of digital health solutions, the perceived digital affinity of medical assistants, GPs' level of digital affinity, the level of extraversion, and the level of neuroticism. In addition, the expected future use of digital health solutions, the perceived digital affinity of medical assistants, GPs' digital affinity, and neuroticism were significant predictors of digital maturity in the regression model.

Comparison With Prior Work

Digital Maturity in General Practices

In line with previous findings investigating digital maturity in general practices across Europe [21], we found a moderate digital maturity of general practices (mean 3.31, SD 0.64) on our 5-point Likert-type scale that ranged from 1.60 to 4.94 in our sample. As pointed out by a study analyzing the readiness for digital health technologies in general practices in England [58], this variation highlights the need to investigate the various constituting dimensions of digital maturity and understand the factors influencing digital maturity in general practices.

Digital maturity is a multifaceted construct encompassing various technological and organizational dimensions beyond digital systems use [13]. Thus, there are oftentimes not only differences in the overall digital maturity between practices but also between the various underlying dimensions [21]. A recent study found the highest scores for overall digital systems use, followed by collective and individual resources and ability [21]. Likewise, our study's subscales of *People, Skills, and Behavior* and *Governance and Management* received the highest scores. Both dimensions center around GPs' capabilities and knowledge in using digital health solutions and leadership in digitalization. Interestingly, our study's subscale *IT Capability*, which focuses

XSL•FO
RenderX

on hardware and network resources and the actual use of digital health solutions, received a lower score, contrasting with previous findings [21]. A comparison between both studies should be interpreted cautiously, given the different underlying foci of digital maturity assessments. Although Teixeira et al [21] used a tool evaluating the digital maturity of electronic health record systems, our study assessed digital maturity more broadly. In addition, the difference in the findings might be attributable to geographical differences in health care delivery. In the United Kingdom, for example, digitalization in general practices is oftentimes managed and taken care of by dedicated practice managers [59]. In Germany, GPs mostly deal with digitalization-related topics themselves, which tends to be perceived as an add-on to the actual medical work [60,61], ultimately limiting adoption.

Given the underlying digital maturity indicators, the low *IT Capability* scores highlight a lack of digital health adoption in general practice. In line with this finding, the adoption of digital health solutions has been slow and cumbersome, as health care professionals perceive several technological, social, and organizational barriers to adoption [60,61], ultimately leading to resistance toward adoption in some cases [62]. However, as GPs' adoption is one of the critical drivers for the success of digital health solutions [63], there is a potential for improvement regarding the use of digital health solutions.

The low scores in the *Patient-centered Care* dimension are especially surprising, as patient centeredness is proposed to be one of the leading quality indicators of digital health [64], benefiting patients' disease knowledge, treatment adherence, self-management, and self-efficacy [65]. Consequently, digital health solutions were considered essential for empowering patients to participate actively in their health decisions [66]. Indeed, some studies found an association between patient-centered digital health solutions and patient empowerment and participation [67]. Accordingly, patient centeredness can somewhat be considered an outcome of successful digital health adoption. This aligns with our sample's relatively low use of digital health solutions. As a result, patient centeredness might be a feature of digitally mature practices, but it is limited, given our sample's moderate maturity. In contrast, our finding might be in line with the perceived additional workload required for digital health adoption and the resulting potential negative impact on the interaction between GPs and patients, which is oftentimes cited as a barrier to adoption [61].

Influencing Factors of Overall Digital Maturity in General Practices

Currently, only a few studies have investigated the factors that influence digital maturity in general practice settings [21]. Regarding gender and sex, evidence regarding digital maturity [21] and digital health use [24,68] are mixed. In line with our findings, some studies found that being male was associated with a higher digital maturity [21] and using digital health technology [24]. Studies also link higher electronic health record use to being female [68]. These mixed findings highlight that, in fact, there might be no difference in overall digital maturity based on gender or sex, that this difference is limited to some facets of digital maturity, or that covariates substantially influenced the effects found.

Although age is commonly identified as a factor promoting digital health adoption [24,69], our finding aligns with that of another study focusing on digital maturity in general practices [21]: in a multivariable linear regression model, age was not a significant predictor of digital maturity. The difference in findings compared with studies focusing on digital health adoption might be explained by the scope of research: digital maturity is a more faceted construct encompassing various dimensions [13]. Although adoption-related dimensions of digital maturity may be associated with age, other dimensions may not. This is in line with the study by Teixeira et al [21], who showed a higher odds ratio for respondents who were younger compared with respondents who were older only for some dimensions of digital maturity.

Besides the demographic variables mentioned, we did not find a significant difference in overall digital maturity based on various practice-related characteristics. Although the professional experience findings align with previous research [21] and mimic our results concerning age, the results regarding practice location are in contrast with previous literature. In their cross-country study, Teixeira et al [21] found that practicing in a rural setting was inversely associated with digital maturity (B=−0.25, 95% CI −0.43 to −0.08), but there was no substantial association with urban practice settings. One potential reason might be the underlying geographical scope of the studies and the corresponding broadband coverage. In 2021, the disparity in broadband uptake between urban and rural households was relatively small in Germany (1.8 percentage points difference) and below the sample average of Organisation for Economic Co-operation and Development (OECD) countries (3.83 percentage points) [70]. The slight difference in broadband uptake between urban and rural households in Germany and the higher OECD average might explain this contrast in findings. An Australian study further underlines this hypothesis, as Australia has an equally low disparity in broadband coverage below the OECD average [70]: they similarly found no substantial difference in telehealth use based on the rurality of the practice location [71].

Surprisingly, we did not find substantial differences in overall digital maturity based on the current use of digital health solutions. This contrasts with studies linking digital maturity to more frequent and long-term access to electronic health records [21] and with many digital maturity models proposing the implementation of digital health solutions as an essential dimension of digital maturity [13,14,16,45]. The findings might be attributed to our sample's relatively low use of digital health solutions, as indicated by the relatively low scores for items relating to the actual use of individual digital health solutions in the *IT Capability* dimension of our digital maturity scale (eg, concerning e-prescription or digital appointment booking). As we found substantial differences in overall digital maturity based on the expected future use of digital health solutions, it might be beneficial to provide GPs with information highlighting the importance of digital health solutions, the latest advancements, and outlooks in digital health. This would act on the perceived lack of information and need for further information highlighted

XSL•FO
RenderX

in studies on digital health adoption [60,72]. As our research design does not allow for causal inferences, another plausible explanation for this result may lie in the opposite direction: respondents with higher overall digital maturity are already more experienced with digital health solutions and might expect to continue using these in the future. However, the results of our linear regression analysis point toward the former explanation of the results, proposing the expected future use of digital health solutions as a significant predictor of overall digital maturity beyond demographics and practice-related characteristics.

GPs' digital affinity in our study was comparable to a quota sample from the general public in larger German cities [33]. As our literature review pointed out, digital literacy is essential in maturity models [13,14,16]. In line with this, almost all interviewees in our expert interviews mentioned people- and skill-related indicators of technology use and perceived their relevance to be high. Taken together, our qualitative and quantitative results highlight the importance of digital skills for GPs and medical assistants. As digital affinity describes a person's tendency to actively engage in intensive technology interaction and provides a first indication of the actual use of technical systems in everyday use settings [33], it is not surprising that we found a substantial difference in overall digital maturity based on both the perceived digital affinity of medical assistants and the respondents' level of digital affinity. This association also holds in our regression model. In line with our findings, previous studies found a lack of digital skills to be a substantial barrier to digital health adoption [60,61,69]. Our finding further highlights the need for sufficient training of GPs and medical assistants regarding technology use, specifically digital health solutions [60,61], to enable the efficient use and management of these tools.

Although previous studies investigated the relationship between personality and digital health adoption [24], no study has examined the association between personality and digital maturity. This might be partially because of the inpatient focus of past research on digital maturity [13,20], where an individual physician has a rather subordinate role. Nevertheless, past research has already recognized that personality is an essential predictor of technology adoption and continued use of apps [25,26]. However, it has a relatively weak association with physicians' digital health use [24]. Looking at the personality traits associated with the personality variables, our results on the association with overall digital maturity seem plausible. Extroverted individuals can be characterized as talkative, energetic, assertive, outgoing, and enthusiastic [73], and thus, they often take on leadership positions. This association is plausible, as digital maturity is, among others, characterized by transparent management and leadership [13]. Neuroticism represents tendencies such as being insecure, anxious, and hostile [73]. Respondents with higher neuroticism might also be more anxious and insecure about adopting digital health solutions, leading to a lower overall digital maturity. For conscientiousness, our findings might be owing to the overall high average in our sample. Given that we examined GPs, it is important to note that their profession inherently demands a significant amount of discipline, organization, exceptional diligence, and reliability. Thus, conscientiousness may not be associated with overall digital maturity, as this is a characteristic inherent to most physicians.

Limitations

Although our study reveals important findings, it has some limitations. First, it needs to be noted that digital maturity encompasses various technological and organizational dimensions and focuses on the potential of general practices in the context of digitalization [13]. As there is a lack of evidence linking digital maturity to health care quality [1], a high level of digital maturity is not necessarily equivalent to a high quality of care. Thus, the findings of our study do not provide guidance on improving health care quality but can rather support GPs in their digitalization efforts by highlighting potential areas of improvement.

Second, as our research on the association of GPs' inherent characteristics and digital maturity was exploratory, we did not use a psychometrically constructed and validated questionnaire to assess digital maturity. To our knowledge, no validated measure for digital maturity exists for general practice settings. Therefore, we based our questionnaire on a literature review of relevant digital maturity indicators in inpatient care settings and a validation of their relevance in expert interviews. This economical approach allows us to assess digital maturity based on existing models in inpatient care settings. As our findings are aligned with previous literature using different measures of digital maturity [21], we are confident that the measure used in our study is appropriate. Compared with the single technology-focused assessment in previous studies [21], a holistic maturity assessment covering all relevant digital maturity dimensions identified in the current literature [13] provides a more thorough picture. However, the measure used in our study should be validated in future studies. In addition, given the lack of a unified scale for digital maturity in general practices and thus different underlying measures, the comparison of our study with previous work is limited. Nevertheless, the consistency in findings might point to the reliability of the patterns identified and the validity of the scale used.

In addition, our literature review was limited to peer-reviewed articles published in German or English in the last 5 years and excluded gray literature. Thus, we might not have identified all the literature relevant to our research question and potentially encountered a publication bias. Nevertheless, as the COVID-19 pandemic has accelerated the adoption of digital health, we aimed at capturing more recent evolvements. To counter a potential publication bias and use an economic approach, we validated the findings of our literature review externally via expert interviews.

Although we found statistically significant differences in digital maturity based on various inherent characteristics, their practical relevance should be discussed. Although the absolute difference in digital maturity was small, the corresponding effect sizes were moderate to large. Thus, we are confident that besides the small absolute difference in digital maturity based on inherent characteristics, our findings still provide a basis to inform future research and practice, that is, the inclusion of GPs' personal qualities when considering the digital maturity of their practice

XSL•FO
RenderX

and designing interventions aimed at increasing digital maturity in general practice settings.

Owing to our cross-sectional study design, we cannot draw causal inferences regarding the relationship between GPs' inherent characteristics and digital maturity. Although we found significant associations for some variables, we cannot assume a causal direction of the underlying effects. As personality can be considered a relatively stable characteristic, we can carefully interpret our findings so that the characteristics influence digital maturity. In addition, although inherent characteristics were associated with the current overall maturity level, it might be interesting to uncover the association with the advancement of digital maturity over time. Future research could take a longitudinal approach, following GPs along their journey and uncovering the associations between the advancement in digitalization and inherent characteristics. Specifically, it could be worthwhile to explore whether GPs' inherent characteristics account for differences in the long-term digital maturity development of a practice. Although the proposed longitudinal approach is more effortful, its results provide insights into digital maturity development and could potentially provide a strong argument for tailoring approaches that support the digital agenda of GPs to their inherent characteristics.

Finally, our research focused on GPs in Germany, thus potentially limiting the generalizability of our findings to other countries. As GPs have a unique and central role in providing comprehensive and continuous health care services in European countries [22] and our findings are aligned with previous cross-country research [21], the findings from our study could be applied to health care systems with a similarly central role of GPs, for example, to the United Kingdom, the Netherlands, Scandinavian countries, Australia, New Zealand, or Canada. However, the findings might not apply to countries where GPs do not play a central role in health care provision. In addition, a potential application and replication in other health care systems must consider the different clinical practice features, health care delivery models, funding mechanisms, and specific organizational structures within the respective health care system as well as cultural differences that potentially shape GPs' personalities differently.

Overall, we propose further research among a larger sample of GPs in Germany and other health care systems to validate the questionnaire used and extend the findings obtained. This could further enhance the value and practical relevance of our findings.

Conclusions

As the research on digital maturity in general practices is still in its infancy, this study provides valuable insights into relevant influencing factors. It highlights that GPs have a pivotal role in the digital maturity of general practices and that there are significant differences in overall digital maturity based on GPs' inherent characteristics. To our knowledge, this is the first study to specifically evaluate the association between personality characteristics and digital maturity. The results emphasize the relevance of personality characteristics for digital maturity in general practices by showing a substantial association between overall digital maturity, GPs' digital affinity, and neuroticism. This association highlights the need for dedicated, targeted approaches based on inherent characteristics to support GPs in their digital agenda and ultimately drive their digital maturity. On the basis of our findings, regulators and other health care stakeholders could consider several measures to drive digital maturity and digital health adoption in general practices: (1) providing an outlook regarding ongoing trends and advancements in digital health, ultimately highlighting the potential and ubiquity of digital health in patient care in the future; (2) offering more dedicated training to GPs and medical assistants, ultimately enhancing their digital literacy; (3) prioritizing support for GPs who would particularly benefit from it owing to their inherent characteristics, for example, GPs who expect a low future use of digital health solutions and have a low digital affinity themselves and within their practice team; and (4) targeting approaches to meet the needs of different types of GPs' rather than a one-size-fits-all approach based on GPs' inherent characteristics associated with lower digital maturity, for example, by providing digital literacy training to GPs with low digital affinity or by providing campaigns that convey confidence and a feeling of trust in the digital transformation process to increase the digital maturity of GPs with high neuroticism.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LW, LF, and SM developed the overarching research question and study design. LW performed the literature review; conducted, transcribed, and qualitatively analyzed the expert interviews; and performed statistical analysis of the data obtained in the web-based survey. SM, LF, and AM supported the recruitment of participants for the interviews and the web-based survey. LW drafted the manuscript. SM, LF, and AM provided a critical review. SM coordinated the project. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Completed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 574 KB - jmir_v26i1e52085_app1.docx]

Multimedia Appendix 2
Completed COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 728 KB - jmir_v26i1e52085_app2.docx]

Multimedia Appendix 3
Translated semistructured interview guide.
[DOCX File , 16 KB - jmir_v26i1e52085_app3.docx]

Multimedia Appendix 4
Completed CHERRIES (Checklist for Reporting Results of Internet E-Surveys).
[DOCX File , 578 KB - jmir_v26i1e52085_app4.docx]

Multimedia Appendix 5
Translated survey questionnaire for general practitioners.
[DOCX File , 31 KB - jmir_v26i1e52085_app5.docx]

Multimedia Appendix 6
Overview of regression coefficients for the regression models predicting digital maturity.
[DOCX File , 74 KB - jmir_v26i1e52085_app6.docx]

References

1. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Aust Health Rev 2018 Sep;42(5):568-578. [doi: 10.1071/AH17255] [Medline: 29986809]
2. Amarasingham R, Plantinga L, Diener-West M, Gaskin DJ, Powe NR. Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med 2009 Jan 26;169(2):108-114. [doi: 10.1001/archinternmed.2008.520] [Medline: 19171805]
3. Martin G, Clarke J, Liew F, Arora S, King D, Aylin P, et al. Evaluating the impact of organisational digital maturity on clinical outcomes in secondary care in England. NPJ Digit Med 2019 May 16;2:41 [FREE Full text] [doi: 10.1038/s41746-019-0118-9] [Medline: 31304387]
4. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752 [FREE Full text] [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]
5. Lingg M, Lütschg V. Health system stakeholders' perspective on the role of mobile health and its adoption in the Swiss health system: qualitative study. JMIR Mhealth Uhealth 2020 May 11;8(5):e17315 [FREE Full text] [doi: 10.2196/17315] [Medline: 32391802]
6. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. J Am Med Inform Assoc 2005;12(5):505-516 [FREE Full text] [doi: 10.1197/jamia.M1700] [Medline: 15905487]
7. Li P, Lee GH, Kim SY, Kwon SY, Kim HR, Park S. From diagnosis to treatment: recent advances in patient-friendly biosensors and implantable devices. ACS Nano 2021 Feb 23;15(2):1960-2004. [doi: 10.1021/acsnano.0c06688] [Medline: 33534541]
8. Wanderås MR, Abildsnes E, Thygesen E, Martinez SG. Video consultation in general practice: a scoping review on use, experiences, and clinical decisions. BMC Health Serv Res 2023 Mar 30;23(1):316 [FREE Full text] [doi: 10.1186/s12913-023-09309-7] [Medline: 36997997]
9. Ernsting C, Dombrowski SU, Oedekoven M, O Sullivan JL, Kanzler M, Kuhlmey A, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017 Apr 05;19(4):e101 [FREE Full text] [doi: 10.2196/jmir.6838] [Medline: 28381394]
10. Paré G, Raymond L, Castonguay A, Grenier Ouimet A, Trudel MC. Assimilation of medical appointment scheduling systems and their impact on the accessibility of primary care: mixed methods study. JMIR Med Inform 2021 Nov 16;9(11):e30485 [FREE Full text] [doi: 10.2196/30485] [Medline: 34783670]
11. Burmann A, Tischler M, Faßbach M, Schneitler S, Meister S. The role of physicians in digitalizing health care provision: web-based survey study. JMIR Med Inform 2021 Nov 11;9(11):e31527 [FREE Full text] [doi: 10.2196/31527] [Medline: 34545813]

XSL•FO
RenderX

12. Johnston DS. Digital maturity: are we ready to use technology in the NHS? Future Healthc J 2017 Oct;4(3):189-192 [FREE Full text] [doi: 10.7861/futurehosp.4-3-189] [Medline: 31098469]
13. Duncan R, Eden R, Woods L, Wong I, Sullivan C. Synthesizing dimensions of digital maturity in hospitals: systematic review. J Med Internet Res 2022 Mar 30;24(3):e32994 [FREE Full text] [doi: 10.2196/32994] [Medline: 35353050]
14. Vidal Carvalho J, Rocha Á, Abreu A. Maturity of hospital information systems: most important influencing factors. Health Informatics J 2019 Sep;25(3):617-631 [FREE Full text] [doi: 10.1177/1460458217720054] [Medline: 28720012]
15. Greenhalgh T, Rosen R, Shaw SE, Byng R, Faulkner S, Finlay T, et al. Planning and evaluating remote consultation services: a new conceptual framework incorporating complexity and practical ethics. Front Digit Health 2021 Aug 13;3:726095 [FREE Full text] [doi: 10.3389/fdgth.2021.726095] [Medline: 34713199]
16. Kolukısa Tarhan A, Garousi V, Turetken O, Söylemez M, Garossi S. Maturity assessment and maturity models in health care: a multivocal literature review. Digit Health 2020 Apr 1;6:2055207620914772 [FREE Full text] [doi: 10.1177/2055207620914772] [Medline: 32426151]
17. Gomes J, Romão M. Information system maturity models in healthcare. J Med Syst 2018 Oct 16;42(12):235. [doi: 10.1007/s10916-018-1097-0] [Medline: 30327955]
18. Mettler T, Pinto R. Evolutionary paths and influencing factors towards digital maturity: an analysis of the status quo in Swiss hospitals. Technol Forecast Soc Change 2018 Aug;133:104-117. [doi: 10.1016/j.techfore.2018.03.009]
19. van Poelgeest R, van Groningen JT, Daniels JH, Roes KC, Wiggers T, Wouters MW, et al. Level of digitization in Dutch hospitals and the lengths of stay of patients with colorectal cancer. J Med Syst 2017 May;41(5):84 [FREE Full text] [doi: 10.1007/s10916-017-0734-3] [Medline: 28391455]
20. Neunaber T, Meister S. Digital maturity and its measurement of general practitioners: a scoping review. Int J Environ Res Public Health 2023 Feb 28;20(5):4377 [FREE Full text] [doi: 10.3390/ijerph20054377] [Medline: 36901387]
21. Teixeira F, Li E, Laranjo L, Collins C, Irving G, Fernandez MJ, et al. Digital maturity and its determinants in general practice: a cross-sectional study in 20 countries. Front Public Health 2022 Jan 13;10:962924 [FREE Full text] [doi: 10.3389/fpubh.2022.962924] [Medline: 36711349]
22. Kringos DS, Boerma W, van der Zee J, Groenewegen P. Europe's strong primary care systems are linked to better population health but also to higher health spending. Health Aff (Millwood) 2013 Apr;32(4):686-694. [doi: 10.1377/hlthaff.2012.1242] [Medline: 23569048]
23. The European definition of general practice / family medicine. Wonca Europe. URL: https://www.woncaeurope.org/file/61a77842-76c2-45dd-a435-e0a8b875f30a/Definition%20EURACTshort%20version%20revised%202011.pdf [accessed 2023-08-21]
24. Zaresani A, Scott A. Does digital health technology improve physicians' job satisfaction and work-life balance? A cross-sectional national survey and regression analysis using an instrumental variable. BMJ Open 2020 Dec 12;10(12):e041690 [FREE Full text] [doi: 10.1136/bmjopen-2020-041690] [Medline: 33310807]
25. Devaraj S, Easley RF, Crant JM. Research note—how does personality matter? Relating the five-factor model to technology acceptance and use. Inf Syst Res 2008 Mar;19(1):93-105. [doi: 10.1287/isre.1070.0153]
26. Su J, Dugas M, Guo X, Gao GG. Influence of personality on mHealth use in patients with diabetes: prospective pilot study. JMIR Mhealth Uhealth 2020 Aug 10;8(8):e17709 [FREE Full text] [doi: 10.2196/17709] [Medline: 32773382]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
30. Carvalho JV, Rocha Á, Abreu A. Maturity models of healthcare information systems and technologies: a literature review. J Med Syst 2016 Jun;40(6):131. [doi: 10.1007/s10916-016-0486-5] [Medline: 27083575]
31. VERBI software. MAXQDA. URL: https://www.maxqda.com/about [accessed 2023-04-05]
32. Kuckartz U, Rädiker S. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. Weinheim, Germany: Beltz Juventa; 2022.
33. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the Affinity for Technology Interaction (ATI) scale. Int J Hum Comput Interact 2018 Mar 30;35(6):456-467. [doi: 10.1080/10447318.2018.1456150]
34. Rammstedt B, John OP. Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. Diagnostica 2005 Oct;51(4):195-206. [doi: 10.1026/0012-1924.51.4.195]
35. Leiner DJ. Too fast, too straight, too weird: non-reactive indicators for meaningless data in internet surveys. Surv Res Method 2019;13(3):229-248 [FREE Full text] [doi: 10.18148/srm/2019.v13i3.7403]

XSL•FO
RenderX

36. Bais F, Schouten B, Toepoel V. Investigating response patterns across surveys: do respondents show consistency in undesirable answer behaviour over multiple surveys? Bull Sociol Methodol 2020 Aug 26;147-148(1-2):150-168. [doi: 10.1177/0759106320939891]
37. IBM Corp. IBM SPSS statistics for windows, version 29.0. IBM Corp. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-29 [accessed 2023-12-28]
38. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011 Jun 27;2:53-55 [FREE Full text] [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
39. Field A. Discovering Statistics Using IBM SPSS Statistics. Thousand Oaks, CA: SAGE Publications, Inc; 2017.
40. Welch BL. On the comparison of several mean values: an alternative approach. Biometrika 1951 Dec;38(3/4):330-336. [doi: 10.2307/2332579]
41. Agarwal A, Pritchard D, Gullett L, Amanti KG, Gustavsen G. A quantitative framework for measuring personalized medicine integration into US healthcare delivery organizations. J Pers Med 2021 Mar 12;11(3):196 [FREE Full text] [doi: 10.3390/jpm11030196] [Medline: 33809012]
42. Al-Kahtani N, Alruwaie S, Al-Zahrani BM, Abumadini RA, Aljaafary A, Hariri B, et al. Digital health transformation in Saudi Arabia: a cross-sectional analysis using healthcare information and management systems society' digital health indicators. Digit Health 2022 Aug 04;8:20552076221117742 [FREE Full text] [doi: 10.1177/20552076221117742] [Medline: 35959196]
43. Brice S, Almond H. Health professional digital capabilities frameworks: a scoping review. J Multidiscip Healthc 2020 Nov 2;13:1375-1390 [FREE Full text] [doi: 10.2147/JMDH.S269412] [Medline: 33173300]
44. Carvalho JV, Rocha Á, van de Wetering R, Abreu A. A maturity model for hospital information systems. J Bus Res 2019 Jan;94:388-399. [doi: 10.1016/j.jbusres.2017.12.012]
45. Chong J, Jason T, Jones M, Larsen D. A model to measure self-assessed proficiency in electronic medical records: validation using maturity survey data from Canadian community-based physicians. Int J Med Inform 2020 Sep;141:104218. [doi: 10.1016/j.ijmedinf.2020.104218] [Medline: 32574925]
46. Grooten L, Borgermans L, Vrijhoef HJ. An instrument to measure maturity of integrated care: a first validation study. Int J Integr Care 2018 Jan 25;18(1):10 [FREE Full text] [doi: 10.5334/ijic.3063] [Medline: 29588644]
47. Khanbhai M, Flott K, Manton D, Harrison-White S, Klaber R, Darzi A, et al. Identifying factors that promote and limit the effective use of real-time patient experience feedback: a mixed-methods study in secondary care. BMJ Open 2021 Dec 08;11(12):e047239 [FREE Full text] [doi: 10.1136/bmjopen-2020-047239] [Medline: 34880009]
48. Kose I, Rayner J, Birinci S, Ulgu MM, Yilmaz I, Guner S. Adoption rates of electronic health records in Turkish Hospitals and the relation with hospital sizes. BMC Health Serv Res 2020 Oct 21;20(1):967 [FREE Full text] [doi: 10.1186/s12913-020-05767-5] [Medline: 33087106]
49. Kouroubali A, Papastilianou A, Katehakis DG. Preliminary assessment of the interoperability maturity of healthcare digital services vs public services of other sectors. Stud Health Technol Inform 2019 Aug 21;264:654-658. [doi: 10.3233/SHTI190304] [Medline: 31438005]
50. Krasuska M, Williams R, Sheikh A, Franklin BD, Heeney C, Lane W, et al. Technological capabilities to assess digital excellence in hospitals in high performing health care systems: international eDelphi exercise. J Med Internet Res 2020 Aug 18;22(8):e17022 [FREE Full text] [doi: 10.2196/17022] [Medline: 32808938]
51. Martin G, Arora S, Shah N, King D, Darzi A. A regulatory perspective on the influence of health information technology on organisational quality and safety in England. Health Informatics J 2020 Jun;26(2):897-910 [FREE Full text] [doi: 10.1177/1460458219854602] [Medline: 31203707]
52. Orenstein EW, Muthu N, Weitkamp AO, Ferro DF, Zeidlhack MD, Slagle J, et al. Towards a maturity model for clinical decision support operations. Appl Clin Inform 2019 Oct;10(5):810-819 [FREE Full text] [doi: 10.1055/s-0039-1697905] [Medline: 31667818]
53. Pumplun L, Fecho M, Wahl N, Peters F, Buxmann P. Adoption of machine learning systems for medical diagnostics in clinics: qualitative interview study. J Med Internet Res 2021 Oct 15;23(10):e29301 [FREE Full text] [doi: 10.2196/29301] [Medline: 34652275]
54. Wherton J, Greenhalgh T, Shaw SE. Expanding video consultation services at pace and scale in Scotland during the COVID-19 pandemic: national mixed methods case study. J Med Internet Res 2021 Oct 07;23(10):e31374 [FREE Full text] [doi: 10.2196/31374] [Medline: 34516389]
55. Williams PA, Lovelock B, Cabarrus T, Harvey M. Improving digital hospital transformation: development of an outcomes-based infrastructure maturity assessment framework. JMIR Med Inform 2019 Jan 11;7(1):e12465 [FREE Full text] [doi: 10.2196/12465] [Medline: 30632973]
56. Woods L, Eden R, Pearce A, Wong YC, Jayan L, Green D, et al. Evaluating digital health capability at scale using the digital health indicator. Appl Clin Inform 2022 Oct;13(5):991-1001 [FREE Full text] [doi: 10.1055/s-0042-1757554] [Medline: 36261114]
57. Gesundheitsdaten. Kassenärztliche Bundesvereinigung. URL: https://gesundheitsdaten.kbv.de/cms/html/16393.php [accessed 2023-11-12]

XSL•FO
RenderX

58. Hammerton M, Benson T, Sibley A. Readiness for five digital technologies in general practice: perceptions of staff in one part of southern England. BMJ Open Qual 2022 Jun;11(2):e001865 [FREE Full text] [doi: 10.1136/bmjoq-2022-001865] [Medline: 35768171]
59. Hanna L, May C, Fairhurst K. Non-face-to-face consultations and communications in primary care: the role and perspective of general practice managers in Scotland. Inform Prim Care 2011;19(1):17-24 [FREE Full text] [doi: 10.14236/jhi.v19i1.789] [Medline: 22118332]
60. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
61. Gagnon MP, Ngangue P, Payne-Gagnon J, Desmartis M. m-Health adoption by healthcare professionals: a systematic review. J Am Med Inform Assoc 2016 Jan;23(1):212-220 [FREE Full text] [doi: 10.1093/jamia/ocv052] [Medline: 26078410]
62. Choi WS, Park J, Choi JY, Yang JS. Stakeholders' resistance to telemedicine with focus on physicians: utilizing the Delphi technique. J Telemed Telecare 2019 Jul;25(6):378-385. [doi: 10.1177/1357633X18775853] [Medline: 29792080]
63. Gagnon MP, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
64. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022 Mar 18;8:20552076221085810 [FREE Full text] [doi: 10.1177/20552076221085810] [Medline: 35340904]
65. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022 Dec 22;24(12):e43086 [FREE Full text] [doi: 10.2196/43086] [Medline: 36548034]
66. Hägglund M, Cajander Å, Rexhepi H, Kane B. Editorial: personalized digital health and patient-centric services. Front Comput Sci 2022 Mar 10;4. [doi: 10.3389/fcomp.2022.862358]
67. Leonardsen AC, Hardeland C, Helgesen AK, Grøndahl VA. Patient experiences with technology enabled care across healthcare settings- a systematic review. BMC Health Serv Res 2020 Aug 24;20(1):779 [FREE Full text] [doi: 10.1186/s12913-020-05633-4] [Medline: 32838784]
68. Holanda AA, do Carmo E Sá HL, Vieira AP, Catrib AM. Use and satisfaction with electronic health record by primary care physicians in a health district in Brazil. J Med Syst 2012 Oct;36(5):3141-3149. [doi: 10.1007/s10916-011-9801-3] [Medline: 22072279]
69. O'Donnell A, Kaner E, Shaw C, Haighton C. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using the clinical adoption framework. BMC Med Inform Decis Mak 2018 Nov 13;18(1):101 [FREE Full text] [doi: 10.1186/s12911-018-0703-x] [Medline: 30424758]
70. Disparity in broadband uptake between urban and rural households. Organisation for Economic Co-operation and Development Going Digital Toolkit. URL: https://goingdigital.oecd.org/en/indicator/17 [accessed 2023-08-21]
71. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857 [FREE Full text] [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
72. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians' attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021 Nov 23;9(11):e33012 [FREE Full text] [doi: 10.2196/33012] [Medline: 34817385]
73. McCrae RR, Costa PTJ. Validation of the five-factor model of personality across instruments and observers. J Pers Soc Psychol 1987 Jan;52(1):81-90. [doi: 10.1037//0022-3514.52.1.81] [Medline: 3820081]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
OECD: Organisation for Economic Co-operation and Development
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

XSL•FO
RenderX

Edited by K Williams; submitted 22.08.23; peer-reviewed by M Marques da Cruz, M Jones, C Strumann; comments to author 02.11.23; revised version received 18.11.23; accepted 16.12.23; published 22.01.24.

Please cite as:
Weik L, Fehring L, Mortsiefer A, Meister S
Big 5 Personality Traits and Individual- and Practice-Related Characteristics as Influencing Factors of Digital Maturity in General Practices: Quantitative Web-Based Survey Study
J Med Internet Res 2024;26:e52085
URL: https://www.jmir.org/2024/1/e52085
doi:10.2196/52085
PMID:38252468

©Lisa Weik, Leonard Fehring, Achim Mortsiefer, Sven Meister. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Attitudes Toward and Use of eHealth Technologies Among German Dermatologists: Repeated Cross-Sectional Survey in 2019 and 2021

Matthias Augustin[1*], MD; Patrick Reinders[1*], MSc; Toni Maria Janke[1], MSc; Klaus Strömer[2], MD; Ralph von Kiedrowski[3], MD; Natalia Kirsten[1], MD; Alexander Zink[4], PD; Marina Otten[1], PhD

[1]Institute for Health Services Research in Dermatology and Nursing (IVDP), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany
[2]Private Dermatological Practice, Ahaus, Germany
[3]Dermatology Specialty Practice, Selters, Germany
[4]Department of Dermatology and Allergy, School of Medicine Technical University of Munich, Munich, Germany
[*]these authors contributed equally

Corresponding Author:
Patrick Reinders, MSc
Institute for Health Services Research in Dermatology and Nursing (IVDP)
University Medical Center Hamburg-Eppendorf (UKE)
Martinistraße 52
Hamburg, 20251
Germany
Phone: 49 40741024721
Email: p.reinders@uke.de

Abstract

Background: In recent years, legal and infrastructural conditions have been set to improve the adoption of digital applications in health care in Germany. The impact of these actions was amplified by the COVID-19 pandemic. So far, no studies have confirmed this progress in dermatology.

Objective: The aim of this study was to measure changes in knowledge, interest, expectation, and use of digital applications in health care among dermatologists in Germany in 2019 and 2021.

Methods: We administered a repeated cross-sectional survey among dermatologists in medical practices and clinics in Germany at 2 time points: t_1 (2019; before the COVID-19 pandemic) and t_2 (2021; during the COVID-19 pandemic). We used a standardized questionnaire, including items on respondents' knowledge, interest, expectation, and use of digital applications, as well as their demographics. The survey was distributed by post and email. The data were analyzed descriptively as well as with multiple logistic regressions.

Results: At t_1, 585 (272/571, 47.6% female; mean age 52.4, SD 8.9 years) dermatologists and at t_2, 792 (360/736, 48.9% female; mean age 54.3, SD 8.6 years) dermatologists participated in this survey. Interest in digital medicine was higher at t_1 than at t_2 (381/585, 65.1% vs 458/792, 57.8%; $P \leq .001$). Nevertheless, 38.6% (306/792) had used digital applications more often since the beginning of the pandemic. For example, real-time telemedicine with patients (12/585, 2.1% vs 160/792, 7.6%; $P \leq .001$) and other specialists did increase (33/385, 5.7% vs 181/792, 22.8%; $P \leq .001$). Almost one-third expressed great concerns about digitalization (272/792, 34.3% vs 294/792, 37.1%; $P = .21$). Spatial analysis revealed higher interest in, more positive expectations toward, and higher use of digital applications in urban areas in comparison to rural areas. For instance, dermatologists from urban areas assessed future applications as having less risk (adjusted odds ratio [aOR] 0.51, 95% CI 0.35-0.76) than did dermatologists from rural areas. The situation was similar with the age groups, as, for example, dermatologists aged <50 years also expected lower risks (aOR 0.51, 95% CI 0.34-0.77) than those aged ≥50 years. There were no differences between sexes in use, but there were differences in knowledge and expectation; for example, male participants assessed their confidence in using digital applications as higher (aOR 1.44, 95% CI 1.01-2.04) than did female participants.

Conclusions: During the pandemic, the use of digital applications in dermatology increased but still remained at a moderate level. The regional and age-related disparities identified indicate the need for further action to ensure equal access to digital care.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e45817) doi:10.2196/45817

KEYWORDS

acceptance; adoption; attitude; COVID-19; dermatology; digital medicine; digitalization; pandemic; perception; teledermatology; telehealth; telemedicine

Introduction

Digital health, or eHealth, can improve the effectiveness, efficiency, and equity in the delivery of health care services. The term was defined by the World Health Organization (WHO) and encompasses traditional information and communication technologies, mobile phone–based apps, artificial intelligence (AI), and computer science-driven genomics in health care [1]. Many digital health applications are nowadays established in health care environments in Europe, ranging from administrative solutions, such as practice or hospital information systems, to telemedical applications in some medical areas (eg, radiology) [2,3].

In terms of health care digitalization, Germany, along with France and Poland, is lagging behind other European countries [4]. Therefore, many legal and reimbursement conditions have been set to foster a nationwide implementation of digital health, for example, the eHealth Act in 2016 [5-8]. These initiatives enabled the reimbursement of video consultations in 2019, following the eased ban on remote treatments in 2018 [9]. Another initiative from 2019 enabled the introduction of digital health applications (Digitale Gesundheitsanwendungen; DiGAs). DiGAs are medical device-classified smartphone apps or web applications that practitioners can prescribe for patients. In terms of digital infrastructure, the nationwide telematics infrastructure (TI) is another important key element in health care digitalization in Germany. The TI aims to network all health care providers (eg, clinics, pharmacies, and outpatient practices) and thereby foster efficiency in the system by making patient data more easily available for all stakeholders [10].

Dermatology is a visual field enabling the adoption of applications such as AI, teledermatology, or patient monitoring for chronic dermatological diseases [9,11-13]. The implications of the COVID-19 pandemic rapidly amplified the adoption and use of digital applications in dermatology worldwide, but also in Germany. For example, every third German dermatologist used some form of teledermatology in May 2020, and of those, 75% had introduced it during the pandemic [14]. Yet, limited data are available on the actual uptake and acceptance of digitalization in dermatology in Germany.

Therefore, the primary aim of our survey was to measure developments in knowledge, interest, expectation, and use of digital applications among dermatologists in practices and clinics from 2019 to 2021, that is, before and after the onset of the pandemic. Our secondary aim was to analyze age, sex, and regional differences on these factors in 2021.

Methods

Study Design and Questionnaire

The Consensus-Based Checklist for Reporting of Survey Studies (CROSS) [15] was used for reporting this study. In June 2019 (t_1), an anonymous, quantitative cross-sectional survey both on the web [16] and offline through a paper-based version was conducted among all German dermatologists that provided care in medical practices and clinics. The questionnaire contained items on knowledge, interest, expectation, and use of digital applications. The following demographic data were collected: the dermatologist's age, sex, and postal code of the office or clinic. The survey was also conducted in June 2021 (t_2), including a few additional items; the t_1 survey had 23 items, while the t_2 survey had 29 items (23 items were identical to t_1 and 6 items were related to developments due to the COVID-19 pandemic, medical device–classified smartphone apps, ie, DiGAs, AI, the electronic health care professional card [elektronischer Heilberufeausweis; eHBA], and the physician's place of work (outpatient medical practice or hospital clinic).

The majority of items were on a 5-point Likert scale. Further items with 2-4 response categories were used (eg, great importance, moderate importance, and no importance). The full questionnaires are provided in Multimedia Appendix 1. Participants had the opportunity to provide written reasons for their responses.

Study Population, Recruitment, and Data Entry

At both time points, all members of the Federal Association of German Dermatologists (BVDD) and the German Dermatological Society (DDG) providing care in medical practices and clinics were invited to anonymously answer one of the questionnaires (on the web or offline). During the recruitment period of 4 weeks, the dermatologists received multiple reminders for participation. The total pool of addresses in both years was approximately 5900 dermatologists, but this also included members who did not practice dermatologic care anymore or had an outdated address. After considering 350 undeliverable mails or emails and an additional 30% of retirees or nonactive dermatologists in the pool, we assumed a sample of 3500 active dermatologists. No sample size calculation was conducted. To reduce errors in data entry regarding the paper-based questionnaires, 25% of all data entries were randomly checked. No errors were identified. During data cleaning, we also checked demographics for duplicates in the web-based survey–generated data.

Statistical Analysis

Statistical analysis was performed with SPSS (version 27; IBM Corporation) for Windows, starting with descriptive statistics. For items with more than 2 answer categories, responses were binary-coded: for items with 5 answer categories, "very often"

XSL•FO
RenderX

and "often" were coded as 1, while "rarely," "never, but planned," and "never, not planned" were coded as 2; for items with 4 answer categories, "yes and "no, but ordered" were coded as 1, while "no, not ordered" and "no, I reject a connection" were coded as 2; for items with 3 answer categories, "great importance" was coded as 1, while "moderate importance" and "no importance" were coded as 2.

Based on the postal code, participants were assigned to urban and rural regions using data from the Federal Institute for Research on Building, Urban Affairs, and Spatial Development (Bundesinstitut für Bau-, Stadt- und Raumforschung; BBSR) [17]. Age was categorized as 25-39 years, 40-49 years, 50-59 years, 60-66 years, and ≥67 years. For subgroup analysis, age was binary coded: ≥50 years and <50 years.

To test for differences between t_1 and t_2 (the primary aim of the study), *t* tests (2-tailed) for interval-scaled variables and chi-square tests for ordinal and nominal-scaled variables were performed (significance level of P=.05). Due to the exploratory nature of this research, no correction of the P value for multiple testing (eg, Bonferroni adjustment) was applied to prevent an increased likelihood of type II errors [18,19]. Additional adjustment for demographic and geographic differences between t_1 and t_2 was done through multivariate logistic regression models. Each model included age, sex, urban or rural area, and type of survey (paper or internet-based). Derived from the models, adjusted odds ratios (aORs) and the associated 95% CI are presented. Items that were introduced at t_2 were analyzed descriptively.

To reach the secondary aim of the study, unadjusted differences between subgroups (age (≥50 vs <50 years, sex, and urban or rural) were calculated at t_2. Additionally, aORs and 95% CI were obtained using logistic regressions.

Missing Data

Missing data are reported for items with a rate >5%. We identified high missing values for some variables (age, sex, and region; Table 1). The Little test of missingness [20] was performed and was significant (χ^2_{3072}=3,704,038; P<.001). Hence, it was assumed that data were not missing completely at random but were missing at random [20,21]. As recommended, 20 data sets were imputed with fully conditional specification through the multiple imputation algorithm within SPSS (version 22) [22]. Thereby, linear regression was used for age, and logistic regressions were used for binary-coded variables (eg, urban or rural and sex), using all variables. After imputation, all analyses were run on each data set, and the results were pooled in accordance with the Rubin rules [23]. As there were minimal differences in the distribution of values before and after imputation, the imputed values are reported.

XSL•FO
RenderX

Table 1. Demographic and geographic distribution of participating dermatologists at both time points.

	t_1 (June 2019; n=585)	t_2 (June 2021; n=792)	*P* value
Age (years), mean (SD)	52.4 (8.9)	54.3 (8.6)	<.001[a]
Age groups (years), n (%)			<.001[b]
25-39	51 (9.3)	40 (5.6)	
40-49	139 (25.5)	162 (22.6)	
50-59	241 (44.1)	294 (41.2)	
60-66	95 (17.4)	181 (25.4)	
≥67	20 (3.7)	37 (5.2)	
Missing values	40 (6.8)	78 (9.9)	
Sex, n (%)			
Female participants	272 (47.6)	360 (48.9)	.64[b]
Missing values	14 (2.4)	56 (7.1)	
Activity, n (%)			
Activity in outpatient medical practice	N/A[c]	640 (87.9)	
Missing values	N/A[c]	60 (7.6)	
Region, n (%)			
Participants from urban areas	398 (76)	471 (76.1)	.97[b]
Participants from eastern federal states	112 (21.4)	130 (21)	.83[b]
Missing values	62 (10.6)	173 (21.8)	
Web-based participation, n (%)	284 (48.5)	261 (33)	<.001[b]

[a]2-tailed *t* tests were performed.

[b]Chi-square tests were performed.

[c]N/A: not applicable; item was first introduced in 2021.

Ethical Considerations

For this study, no ethical review was necessary as the study was anonymous and noninterventional. The Ethics Committee of the Medical Association in Hamburg states that no ethical approval is necessary for studies in which data have already been collected anonymously (Ethik-Kommission; Sonstige Studien: Ärztekammer Hamburg [24]). The study was conducted in accordance with the principles of the Declaration of Helsinki and the Guidelines for Good Clinical Practice.

Results

At t_1, a total of 585 dermatologists participated, and at t_2, a total of 792 dermatologists participated in the survey (Table 1). Considering approximately 3500 eligible dermatologists in both years, the participation rate is thereby estimated to be 16.7% (585/3500) and 22.6% (792/3500), respectively.

Sex and geographic distribution were comparable (no significant differences) between the 2 time points (Table 1), except for a mean age difference of 1.9 years between t_1 and t_2 ($P \leq .001$). There was also a significant difference in the mode of participation: while in 2019, about 48.5% (284/585) dermatologists participated through the web-based survey, only 32.9% (261/792) did so in 2021 ($P \leq .001$).

Most dermatologists expressed interest in digital medicine in both years, and half of them also felt confident in using it, but interest (381/585, 65.1% vs 458/792, 57.8%; $P<.001$) and confidence (297/585, 50.7% vs 333/792, 42.1%; $P=.002$) significantly decreased from t_1 to t_2 (Figure 1). The self-assessed knowledge of digital medicine was rather low in both years (228/585, 38.9% vs 272/792, 34.4%; $P=.08$)

The importance of the digital transformation within the health care system was rated highly by the participants. Yet a slight negative trend was noticed (t_1: 400/585, 68.3% vs t_2: 492/792, 62.1%; $P=.03$) (Figure 1). For around one-third of participants, digital medicine harbors a high risk (t_1: 201/585, 34.3% vs t_2: 294/792, 37.1%; $P=.21$) and supports daily activities (223/585, 38.2% vs 304/792, 38.4%; $P=.93$). Only a small number of dermatologists had read the guideline on teledermatology (t_1: 95/585, 16.2% vs t_2: 148/792, 18.7%; $P=.40$). DiGAs were seen as important in the future by 38.2% (303/792) of participants at t_2.

The most often used digital applications in dermatology at both survey time points were those ensuring asynchronous

XSL•FO
RenderX

communication (store-and-forward) with patients (t_1: 253/585, 43.2% and t_2: 348/792, 44%; *P*=.54) and other specialists (t_1: 273/585, 46.7% and t_2: 359/792, 45.3%; *P*=.58). Real-time video consultation with a patient was used 4 times more frequently in 2021 compared with 2019 but was still used rarely (t_1: 12/585, 2.1%, and t_2: 60/792, 7.6%; aOR 4.00, 95% CI 2.12-7.52) (Figure 1). Real-time communication with other specialists was used nearly 6 times more often (t_1: 33/585, 5.7%, and t_2: 181/792, 22.8%; *P*<.001). Remote patient monitoring (22/585, 3.8% vs 55/792, 7%; *P*=.02) and electronic appointment reminders (t_1: 121/585, 20.6% vs t_2: 212/792, 26.8%; *P*=.02) were also used more regularly by dermatologists in 2021. In contrast, electronic physician letters were sent significantly less often in 2021 compared with 2019 (t_1: 122/585, 20.8% vs t_2: 105/792, 13.3%; *P*<.001). Incorporation of web-based patient data into patient care was rarely used at both time points (t_1: 25/585, 4.2% vs 38/792, 4.8%; *P*=.57) (Figure 1).

Connection to the nationwide TI has progressed further, with 79% (626/792) of participants being connected or having requested a connection in comparison to 66.5% (389/792) in 2019 (*P*<.001). The eHBA was available or at least requested by 80.3% (636/792), and AI methods for diagnosing were used by 21.4% (169/792) of participating dermatologists at t_2. For the last 2 items, no data were available for t_1.

According to respondents, 38.6% (306/792) had used digital medicine procedures more frequently since the onset of the COVID-19 pandemic. Of these, 92.4% (283/306) estimated to make at least partial continued use of the newly introduced digital applications.

Younger dermatologists indicated higher interest, knowledge, and confidence in using digital applications (aORs in Table 2). They saw more of a benefit in using applications for their daily activities than older dermatologists. Younger participants also used asynchronous communication methods with patients and physicians and electronic appointment reminders for patients more often, whereas older dermatologists more often expected high risks regarding the implementation of digital medicine.

Participants practicing in urban areas had a higher likelihood of reporting a good level of knowledge, interest, and confidence in using digital applications (aORs in Table 3). Furthermore, they used many of the digital applications more often, including asynchronous communication with patients and colleagues and electronic patient reminders, and rated the risks associated with digital medicine as lower. In contrast, they were less likely connected to the TI.

Only a few significant differences in sexes were identified (aORs in Table 4). Male dermatologists were more likely than female dermatologists to state a good level of knowledge and confidence in using digital applications. Female dermatologists more often expected the great importance of DiGAs and the digital transformation in the future.

Figure 1. Forrest plot depicting the development of knowledge, use, and expectations of digital medicine in dermatology. Odds ratios were adjusted using logistic regressions for age groups (25-39, 40-49, 50-59, 60-66, and ≥67 years), sex (male or female), and region (urban or rural). The values in bold are significant. A chi-square test was performed to obtain *P* values. Percentages present crude or unadjusted values.

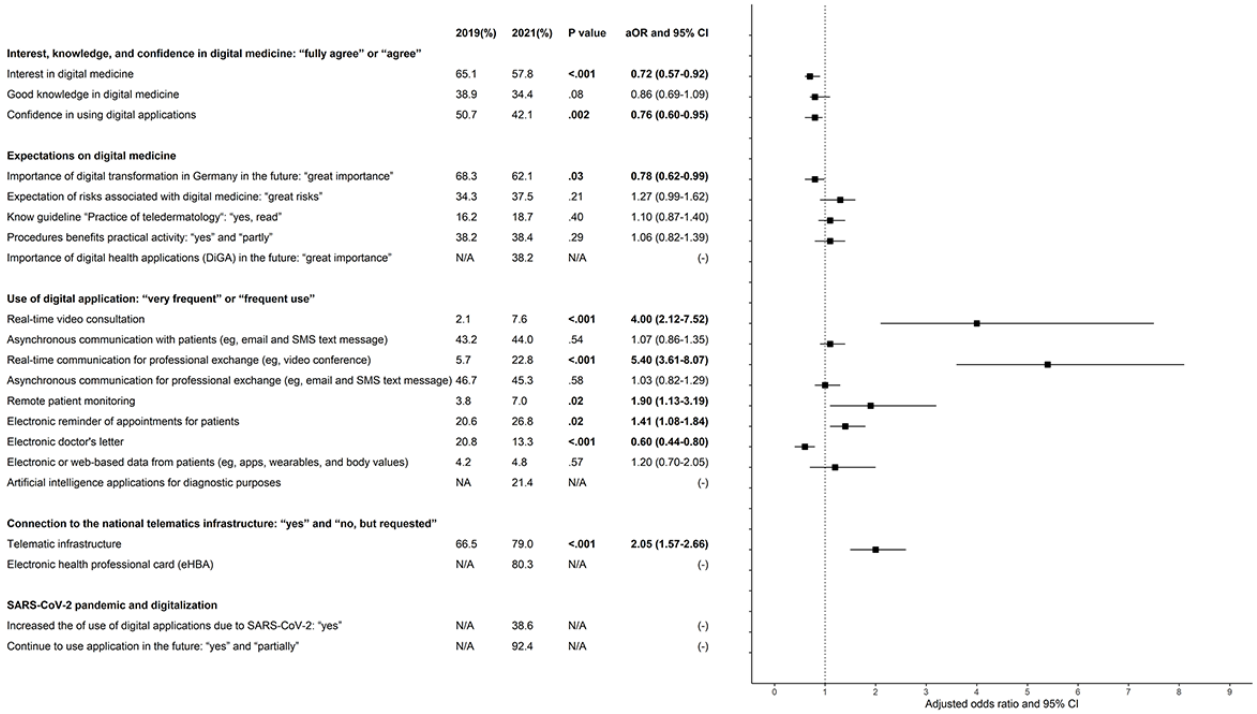

XSL•FO
RenderX

Table 2. Adjusted odds ratios (aORs) for associations between age groups (≥50 years vs <50 years) and knowledge, interest, expectation, and use of digital applications in dermatology (survey date: 2021). The values in italics are significant.

	Age ≥50 years (n=566[a]), n (%)[b]	Age <50 years (n=223[a]), n (%)[b]	*P* value[c,b]	aOR (95% CI) for age ≥50/< 50 years; reference ≥50 years[d]
Interest, knowledge, and confidence in digital medicine: "fully agree" or "agree"				
Interest in digital medicine	*304 (53.7)*	*158 (70.9)*	*<.001*	*2.15 (1.47-3.15)*
Good knowledge of digital medicine	*180 (31.8)*	*96 (43.0)*	*<.001*	*1.78 (1.25-2.54)*
Confidence in using digital applications	*220 (38.9)*	*118 (52.9)*	*<.001*	*1.91 (1.33-2.72)*
Use of digital application: "very frequent" or "frequent"				
Real-time video consultation	40 (7.1)	22 (9.9)	.24	1.29 (0.68-2.46)
Asynchronous communication with patients (eg, email and SMS text message)	*235 (41.5)*	*121 (54.3)*	*<.001*	*1.65 (1.16-2.34)*
Real-time communication for professional exchange (eg, video conference)	*118 (20.8)*	*66 (29.6)*	*<.001*	1.45 (0.98-2.47)
Asynchronous communication for professional exchange (eg, email and SMS text message)	*237 (41.9)*	*127 (57.0)*	*<.001*	*1.73 (1.23-2.44)*
Remote patient monitoring	36 (6.4)	20 (9.0)	.22	1.47 (0.72-2.99)
Electronic reminder of appointments for patients	*140 (24.7)*	*77 (34.5)*	*.005*	*1.63 (1.11-2.38)*
Electronic doctor's letter	77 (13.6)	32 (14.3)	.62	0.97 (0.56-1.67)
Electronic or web-based data from patients (eg, apps, wearables, and body values)	26 (4.6)	13 (5.8)	.28	1.63 (0.70-3.83)
Artificial intelligence applications for diagnostic purposes	115 (20.3)	52 (23.3)	.35	1.07 (0.72-1.60)
Expectations of digital medicine				
Importance of digital transformation in Germany in the future: "great importance"	*341 (60.2)*	*159 (71.3)*	*.007*	1.43 (0.98-2.10)
Expectation of risks associated with digital medicine: "great risks"	*247 (42.6)*	*58 (26.0)*	*<.001*	*0.51 (0.34-0.77)*
Know guideline "Practice of teledermatology:" "yes, read"	117 (20.7)	34 (15.2)	.10	0.85 (0.51-1.40)
Procedures benefits practical activity: "yes" and "partly"	*222 (39.2)*	*118 (52.9)*	*<.001*	*1.68 (1.19-2.39)*
Importance of digital health applications (DiGA[e]) in the future: "great importance"	211 (37.3)	90 (40.4)	.43	0.97 (0.68-1.38)
Connection to the national telematics infrastructure: "yes" and "no, but requested"				
Telematic infrastructure	473 (83.6)	173 (77.6)	.08	0.75 (0.47-1.18)
Electronic health professional card (eHBA[f])	461 (81.4)	171 (76.7)	.09	0.82 (0.52-1.31)
SARS-CoV-2 pandemic and digitalization				
Increased the use of digital applications due to SARS-CoV-2: "yes"	*184 (32.5)*	*111 (49.8)*	*<.001*	*1.85 (1.26-2.71)*
Continue to use application in the future: "yes" and "partially"	173 (94.0)	104 (93.7)	.78	0.91 (0.28-2.93)

[a]Does not add up to the overall total of 792, as data for an average of 3 participants were not imputed.

[b]Unadjusted or crude values.

[c]Chi-square tests were performed.

[d]Logistic regressions were performed. Following items were included as independent variables: sex (male or female) and regional allocation (urban or rural).

[e]DiGAs: Digitale Gesundheitsanwendungen.

[f]eHBA: elektronischer Heilberufeausweis.

XSL•FO
RenderX

Table 3. Adjusted odds ratios (aORs) for associations between region (urban or rural) and knowledge, interest, expectation, and use of digital applications in dermatology (survey date: 2021). The values in italics are significant.

	Urban (n=600), n (%)[a]	Rural (n=192), n (%)[a]	*P* value[b]	aOR (95% CI) for urban/rural; reference: rural[c]
Interest, knowledge, and confidence in digital medicine: "fully agree" or "agree"				
Interest in digital medicine	*372 (62)*	*91 (47.4)*	*<.001*	*1.70 (1.17-2.47)*
Good knowledge of digital medicine	*228 (38)*	*50 (26)*	*.01*	*1.64 (1.06-2.49)*
Confidence in using digital applications	*279 (46.5)*	*61 (31.8)*	*.002*	*1.73 (1.17-2.58)*
Use of digital application: "very frequent" or "frequent"				
Real-time video consultation	55 (9.2)	7 (3.6)	.05	2.68 (0.95-7.51)
Asynchronous communication with patients (eg, email and SMS text message)	*295 (49.2)*	*63 (32.8)*	*.01*	*1.79 (1.20-2.68)*
Real-time communication for professional exchange (eg, video conference)	*154 (25.7)*	*30 (15.6)*	*.02*	*1.90 (1.13-3.17)*
Asynchronous communication for professional exchange (eg, email and SMS text message)	*297 (49.5)*	*70 (36.5)*	*<.001*	*1.62 (1.10-2.39)*
Remote patient monitoring	45 (7.5)	12 (6.3)	.58	1.19 (0.56-2.52)
Electronic reminder of appointments for patients	*185 (30.8)*	*33 (17.2)*	*<.001*	*2.01 (1.24-3.27)*
Electronic doctor's letter	89 (14.8)	20 (10.4)	.22	1.54 (0.83-2.86)
Electronic or web-based data from patients (eg, apps, wearables, and body values)	32 (5.3)	8 (4.2)	.51	1.31 (0.49-3.56)
Artificial intelligence applications for diagnostic purposes	132 (22.0)	38 (19.8)	.52	1.12 (0.70-1.82)
Expectations of digital medicine				
Importance of digital transformation in Germany in the future: "great importance"	*400 (66.7)*	*104 (54.2)*	*<.001*	*1.64 (1.08-2.48)*
Expectation of risks associated with digital medicine: "great risks"	*205 (34.2)*	*101 (52.6)*	*<.001*	*0.51 (0.35-0.76)*
Know guideline "Practice of teledermatology:" "yes, read"	84 (14.0)	25 (13.0)	.79	1.23 (0.74-2.08)
Procedures facilitate practical activity: "yes" and "partly"	270 (45.0)	72 (37.5)	.10	1.28 (0.86-1.93)
Importance of digital health applications (DiGA[d]) in the future: "great importance"	185 (39.8)	48 (33.3)	.14	1.30 (0.87-1.95)
Connection to the national telematics infrastructure: "yes" and "no, but requested"				
Telematic infrastructure	*473 (78.8)*	*175 (91.1)*	*<.001*	*0.37 (0.21-0.68)*
Electronic health professional card (eHBA[e])	471 (78.5)	163 (84.9)	.06	0.61 (0.35-1.31)
SARS-CoV-2 pandemic and digitalization				
Increased the use of digital applications due to SARS-CoV-2: "yes"	*237 (39.5)*	*59 (30.7)*	*.05*	1.37 (0.91-2.06)
Continue to use application in the future: "yes" and "partially"	221 (93.2)	57 (96.6)	.50	0.63 (0.14-2.95)

[a]Unadjusted or crude values.

[b]Chi-square tests were performed.

[c]Logistic regressions were performed. Following items were included as independent variables: age (≥50 years or < 50 years) and sex (male or female).

[d]DiGAs: Digitale Gesundheitsanwendungen.

[e]eHBA: elektronischer Heilberufeausweis.

XSL•FO
RenderX

Table 4. Adjusted odds ratios (aORs) for associations between sex and knowledge, interest, expectation, and use of digital applications in dermatology (survey date: 2021). The values in italics are significant.

	Female (n=388), n (%)[a]	Male (n=404), n (%)[a]	*P* value[b]	aOR (95% CI) for male/female; reference: female[c]
Interest, knowledge, and confidence in digital medicine: "fully agree" or "agree"				
Interest in digital medicine	224 (57.7)	240 (59.4)	.94	1.22 (0.87-1.74)
Good knowledge of digital medicine	*119 (30.7)*	*158 (39.1)*	*.02*	*1.71 (1.18-2.46)*
Confidence in using digital applications	154 (39.7)	186 (46.0)	.09	*1.44 (1.01-2.04)*
Use of digital application: "very frequent" or "frequent"				
Real-time video consultation	31 (8.0)	31 (7.7)	.86	1.01 (0.52-1.97)
Asynchronous communication with patients (eg, email and SMS text message)	174 (44.8)	184 (45.5)	.86	1.19 (0.84-1.67)
Real-time communication for professional exchange (eg, video conference)	100 (25.8)	85 (21.0)	.10	0.79 (0.53-1.19)
Asynchronous communication for professional exchange (eg, email and SMS text message)	187 (48.2)	179 (44.3)	.24	0.92 (0.65-1.29)
Remote patient monitoring	30 (7.7)	27 (5.7)	.45	0.94 (0.48-1.86)
Electronic reminder of appointments for patients	98 (25.3)	119 (29.5)	.34	1.35 (0.92-1.98)
Electronic doctor's letter	53 (13.7)	56 (13.7)	.89	0.79 (0.48-1.31)
Electronic or web-based data from patients (eg, apps, wearables, and body values)	16 (4.1)	23 (5.7)	.30	1.57 (0.69-3.57)
Artificial intelligence applications for diagnostic purposes	93 (24.0)	77 (19.1)	.09	0.87 (0.57-1.31)
Expectations of digital medicine				
Importance of digital transformation in Germany in the future: "great importance"	*267 (68.8)*	*237 (58.7)*	*.04*	*0.67 (0.49-0.93)*
Expectation of risks associated with digital medicine: "great risks"	*132 (34.0)*	*173 (42.8)*	*.02*	1.39 (0.97-1.97)
Know guideline "Practice of teledermatology:" "yes, read"	58 (14.9)	79 (19.6)	.09	1.32 (0.85-2.05)
Procedures eases practical activity: "yes" and "partly"	175 (45.1)	167 (41.3)	.34	1.13 (0.80-1.60)
Importance of digital health applications (DiGA[d]) in the future: "great importance"	*167 (43.0)*	*134 (33.2)*	*.05*	*0.57 (0.40-0.80)*
Connection to the national telematics infrastructure: "yes" and "no, but requested"				
Telematic infrastructure	322 (83.0)	326 (80.7)	.43	0.96 (0.63-1.47)
Electronic health professional card (eHBA[e])	314 (80.9)	320 (79.2)	.41	0.97 (0.64-1.49)
SARS-CoV-2 pandemic and digitalization				
Increased the use of digital applications due to SARS-CoV-2: "yes"	*162 (41.8)*	*135 (33.4)*	*.02*	0.82 (0.32-2.10)
Continue to use applications in the future: "yes" and "partially"	151 (93.2)	127 (94.1)	.50	0.40 (0.13-1.28)

[a]Unadusted or crude values.

[b]Chi-square tests were performed.

[c]Logistic regressions were performed. The following items were included as independent variables: age (≥50 years or <50 years) and regional allocation (urban or rural).

[d]DiGAs: Digitale Gesundheitsanwendungen.

[e]eHBA: elektronischer Heilberufeausweis.

Discussion

Overview

This trend analysis is the first to measure developments in knowledge, interest, expectation, and use of digital applications in dermatology during the events of the COVID-19 pandemic. First, we identified a moderate increase in the use rate of digital applications in dermatological care, while interest in the topic decreased slightly; one-third of practitioners expect great risks with the introduction of digital medicine. Second, we noticed regional and age-related disparities in self-assessed knowledge,

XSL•FO
RenderX

confidence, expectations of digital medicine, and the use rate of digital applications. The survey also revealed differences between sexes in self-assessed knowledge and confidence in digital medicine among dermatologists in Germany.

The decline in interest in using digital applications might be explained by the more widespread implementation of digital applications, as verified by this survey. We hypothesize that digital applications may have moved from extraordinary to ordinary daily routines, resulting in a reduction of interest. Confidence and knowledge may have declined as practitioners were more often confronted with the actual use of applications in their daily routine, potentially disclosing the limits of their knowledge in digital medicine.

Interestingly, only one-third of physicians stated that digital applications support their daily activities, which is one of the main promises of health care digitalization. The outcome could be caused by the long, costly, and not yet fully implemented nationwide eHealth initiatives in Germany [25]. The overall perception will, however, differ depending on the type of digital application. The benefit of AI for diagnostic purposes, for example, might be high, as the technology could potentially improve the efficiency of screening and the diagnostic confidence of practitioners [26].

We observed an increase in real-time communication technologies to communicate with patients and other physicians, which could also be shown by other studies [14]. Compared to the United States, where nearly 42.6% of offices stated using video consultations, this service is rarely offered by dermatologists in Germany, as identified within our results (160/792, 7.6%) [27]. Reasons for the low use may involve insufficient reimbursement for providers and a 2019-loosened ban on remote treatments, which also explains the low use rate for this year [14]. In the future, it can be assumed that video consultation with patients may decline after the pandemic fades into the background again, as the main reason for its use was protection from infections [2]. Lower use of video consultations in times of low COVID-19 infection rates could already be shown in Germany [28]. Nevertheless, it is not assumed that the use rates of video consultation will go back to the use rates of the prepandemic era.

On the contrary, the use of asynchronous communication technologies was not affected by the pandemic or other changes during the observed period, which contradicts other results from Elsner [14] at first glance, wherein 75% of dermatologists in 2020 did introduce the technology with the onset of the pandemic. Within our survey, all purposes—administrative (eg, appointments and invoices), medical-related (eg, diagnosis), and technologies (eg, email, SMS text messages, or teledermatological platforms)—were covered by the item, whereas the item in the survey by Elsner [14] was specific to store-and-forward teledermatology, so a medical-related purpose. As the number of dermatologists adopting store-and-forward teledermatology substantially increased, we assume that the purpose of asynchronous communications has partly shifted from administrative purposes to medical-related purposes with the onset of the pandemic.

In terms of the TI, dermatologists are a bit behind other specialist groups: 14% of all practitioners that responded to a nationwide survey [2] and 21% (166/792) of dermatologists within our survey are not yet connected to the TI. Regarding DiGAs, they see a similar importance to other physicians: 40% nationwide of all physicians versus 38.2% (303/792) in our survey. We found a reduction in use of the electronic physician's letter in dermatology, although most outpatient physicians stated that it eases the administrative workload in clinical practice [2].

Within the survey, age, sex, and regional differences were identified. The lower self-assessed knowledge and confidence in digital medicine by female dermatologists could be confounded by social expectations. In other areas of life, no or only marginal differences in digital literacy between sexes could be measured when accounting for education and employment status [29]. Nevertheless, female dermatologists could thus have similar knowledge of digital medicine, resulting in similar use rates, as shown.

The phenomenon of regional and age disparities, noted in our survey, has generally been noted with the advent of digitalization in all areas of life and is known as the digital divide. Thus, not only access to the internet, internet speed, and availability of appropriate hardware, but also digital competencies among practitioners and patients are related to inequalities in the use of digital tools [30]. In Germany, the average use rates of the internet, as well as digital competence and openness to use digital applications, are more pronounced in urban regions and younger age groups in comparison to rural regions and older groups [31], partially explaining the differences identified within our survey. Consequently, digital applications have the potential to increase unequal access to care, at least in the short term. To ensure equity in the delivery of care in dermatology, all disparities related to sex, age, or region should be addressed by legal, infrastructure, and reimbursement initiatives. In addition, implementation strategies for evidence-based digital applications should be customized to specific target groups [32].

The response rates of our surveys were slightly lower than in previous surveys among German dermatologists [33]. Hence, a certain selection bias among participants cannot be excluded. Considering the official physician statistics, dermatologists aged between 50 and 59 years were slightly overrepresented, whereas dermatologists aged 35 years or younger were slightly underrepresented. In addition, female dermatologists were underrepresented in the survey in comparison to the federal statistic (360/736, 48% in our survey vs 60% at the national level) [34]. Although our surveys were not representative, the sample sizes were large enough to detect trends and variations among demographic and geographic parameters. To improve comparability between survey waves, an adjustment of demographic and geographic differences was performed through multiple logistic regressions. Groups at t_1 and t_2 may differ for unmeasured characteristics (eg, professional expertise, number of consultations per week, socioeconomic patient groups in the area, and the number of private and statutory insured patients), thus resulting in underestimated or overestimated differences. Nevertheless, the same population was contacted at both time

XSL•FO
RenderX

points, and important baseline characteristics did not present any larger differences.

Conclusions

The repeated survey revealed an increase in the adoption of the majority of digital applications in dermatology, amplified by the COVID-19 pandemic. However, regional and age-related disparities in the use of digital applications exist, reducing equal access to innovative health care solutions for patients. In the future, policy strategies must be developed to counteract these disparities and improve nationwide implementation. Developers of digital innovations should incorporate physicians' perspectives more commonly to ensure feasibility and use. Future research should support policymakers and developers by identifying barriers to implementation, monitoring disparities, and ensuring an evidence base for digital applications. Widespread implementation and acceptance among patients and physicians can then improve the efficiency, equity, and effectiveness of care.

Acknowledgments

The authors would like to thank the scientific communication team of the Institute for Health Services Research in Dermatology and Nursing, especially Merle Twesten and Mario Gehoff, for editing this manuscript.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MA was responsible for the conceptualization, methodology, software, supervision, writing, review, and editing of this study. PR handled data curation, formal analysis, investigation, project administration, visualization, and writing the original draft. TMJ supported the formal analysis. KS, AZ, and RVK supported the conceptualization of the survey and reviewed the manuscript. NK was responsible for the methodology, writing, review, and editing of the manuscript. MO initially developed the survey and supported data curation, formal analysis, investigation, methodology, project administration, validation, visualization, writing, review, and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey: Digital medicine in German dermatology practices and clinics 2019 and 2021.
[DOCX File , 19 KB - jmir_v26i1e45817_app1.docx]

References

1. Fifty-eighth world health assembly. World Health Organization. 2005. URL: https://apps.who.int/gb/ebwha/pdf_files/WHA58-REC1/A58_2005_REC1-en.pdf [accessed 2024-01-03]
2. Albrecht M, Otten M, Sander M, Temizdemir E. PraxisBarometer Digitalisierung 2020: Stand und Perspektiven der Digitalisierung in der vertragsärztlichen und -psychotherapeutischen Versorgung. Berlin, Germany: IGES Institut; 2020.
3. The Topol review. Preparing the healthcare workforce to deliver the digital future: an independent report on behalf of the Secretary of State for Health and Social Care. Health Education England. 2019. URL: https://topol.hee.nhs.uk/ [accessed 2021-07-28]
4. #SmartHealthSystems international comparison of digital strategies: digitisation requires political leadership, a strategy that is driven by health policy goals, and a coordinating institution. Bertelsmann Stiftung. 2019. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/Projekte/Der_digitale_Patient/VV_SHS_Europe_eng.pdf [accessed 2022-10-26]
5. E-Health-Gesetz verabschiedet. Bundesministerium für Gesundheit. 2015. URL: https://www.bundesgesundheitsministerium.de/ministerium/meldungen/2015/dezember-2015/e-health [accessed 2021-07-07]
6. Ärzte sollen Apps verschreiben können: Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation (Digitale-Versorgung-Gesetz—DVG). Bundesministerium für Gesundheit. 2020. URL: https://www.bundesgesundheitsministerium.de/digitale-versorgung-gesetz.html [accessed 2021-07-28]
7. Schnellere Termine, mehr Sprechstunden, bessere Angebote für gesetzlich Versicherte: Terminservice- und Versorgungsgesetz (TSVG). Bundesministerium für Gesundheit. 2019. URL: https://www.bundesgesundheitsministerium.de/terminservice-und-versorgungsgesetz.html [accessed 2021-07-28]
8. Bundesärztekammer. Hinweise und Erläuterungen zu § 7 Abs. 4 MBO-Ä—Behandlung im persönlichen Kontakt und Fernbehandlung. Deutsches Ärzteblatt 2019;19:116 [FREE Full text]

XSL•FO
RenderX

9. Augustin M, Wimmer J, Biedermann T, Blaga R, Dierks C, Djamei V, et al. Praxis der teledermatologie. J Dtsch Dermatol Ges 2018;16(Suppl 5):6-57. [doi: 10.1111/ddg.13512] [Medline: 29998512]
10. Thiel R, Deimel L, Schmidtmann D, Piesche K, Hüsing T, Rennoch J, et al. #SmartHealthSystems: Digitalisierungsstrategien im internationalen Vergleich. Bertelsmann Stiftung. 2018. URL: https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/smarthealthsystems [accessed 2021-10-05]
11. Augustin M, Strömer K. S2k-Leitlinie Teledermatologie. AWMF Online. 2020. URL: https://www.awmf.org/leitlinien/detail/ll/013-097.html [accessed 2021-07-28]
12. Armstrong AW, Chambers CJ, Maverakis E, Cheng MY, Dunnick CA, Chren MM, et al. Effectiveness of online vs in-person care for adults with psoriasis: a randomized clinical trial. JAMA Netw Open 2018;1(6):e183062 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.3062] [Medline: 30646223]
13. Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020;31(1):137-143 [FREE Full text] [doi: 10.1016/j.annonc.2019.10.013] [Medline: 31912788]
14. Elsner P. Teledermatology in the times of COVID-19—a systematic review. J Dtsch Dermatol Ges 2020;18(8):841-845 [FREE Full text] [doi: 10.1111/ddg.14180] [Medline: 33448667]
15. Sharma A, Duc NTM, Thang TLL, Nam NH, Ng SJ, Abbas KS, et al. A consensus-based Checklist for Reporting of Survey Studies (CROSS). J Gen Intern Med 2021;36(10):3179-3187 [FREE Full text] [doi: 10.1007/s11606-021-06737-1] [Medline: 33886027]
16. Unipark. URL: https://www.unipark.com/ [accessed 2024-01-08]
17. INKAR—Indikatoren und Karten zur Raum- und Stadtentwicklung. Bundesinstitut für Bau-, Stadt- und Raumforschung. URL: https://www.inkar.de/Default [accessed 2021-07-27]
18. Bender R, Lange S. Adjusting for multiple testing--when and how? J Clin Epidemiol 2001;54(4):343-349. [doi: 10.1016/s0895-4356(00)00314-0] [Medline: 11297884]
19. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502-508 [FREE Full text] [doi: 10.1111/opo.12131] [Medline: 24697967]
20. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
21. Mirzaei A, Carter SR, Patanwala AE, Schneider CR. Missing data in surveys: key concepts, approaches, and applications. Res Social Adm Pharm 2022;18(2):2308-2316. [doi: 10.1016/j.sapharm.2021.03.009] [Medline: 33775556]
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
23. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley; 1987.
24. Aerztekammer-Hamburg. URL: https://www.aerztekammer-hamburg.org/home.html [accessed 2024-01-08]
25. Stafford N. Germany is set to introduce e-health cards by 2018. BMJ 2015;350:h2991. [doi: 10.1136/bmj.h2991] [Medline: 26032607]
26. Nelson CA, Pachauri S, Balk R, Miller J, Theunis R, Ko JM, et al. Dermatologists' perspectives on artificial intelligence and augmented intelligence—a cross-sectional survey. JAMA Dermatol 2021;157(7):871-874 [FREE Full text] [doi: 10.1001/jamadermatol.2021.1685] [Medline: 34037674]
27. Gorrepati PL, Smith GP. Analysis of availability, types, and implementation of teledermatology services during COVID-19. J Am Acad Dermatol 2020;83(3):958-959 [FREE Full text] [doi: 10.1016/j.jaad.2020.06.022] [Medline: 32531304]
28. Mangiapane S, Kretschmann J, Czihal T, von Stillfried D. Veränderung der vertragsärztlichen Leistungsinanspruchnahme während der COVID-Krise: Tabellarischer Trendreport bis zum Ende des Jahres 2021. Zentralinstitut für die kassenärztliche Versorgung in Deutschland. 2021. URL: https://www.zi.de/fileadmin/Downloads/Service/Publikationen/Zi-TrendReport_2021-Q4_2022-06-10.pdf [accessed 2022-10-26]
29. Ertl B, Csanadi A, Tarnai C. Getting closer to the digital divide: an analysis of impacts on digital competencies based on the German PIAAC sample. Int J Educ Dev 2020;78:102259 [FREE Full text] [doi: 10.1016/j.ijedudev.2020.102259] [Medline: 32834463]
30. Robinson L, Cotten SR, Ono H, Quan-Haase A, Mesch G, Chen W, et al. Digital inequalities and why they matter. Inf Commun Soc 2015;18(5):569-582. [doi: 10.1080/1369118X.2015.1012532]
31. Initiative D21 e.V.. D21 Digital Index 2018/2019 Jährliches Lagebild zur Digitalen Gesellschaft. Berlin, Germany: Hochschule der Medien; 2019.
32. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. Lancet 2003;362(9391):1225-1230. [doi: 10.1016/S0140-6736(03)14546-1] [Medline: 14568747]
33. Zander N, Krensel M, Schäfer I, Weyergraf A, Mohr P, Strömer K, et al. Social rights claims for malignant melanoma patients: providing patient information in German dermatological practices. Hautarzt 2019;70(10):804-810 Sozialrechtliche Ansprüche beim malignen Melanom : Aufklärung von Patienten in deutschen dermatologischen Praxen. [doi: 10.1007/s00105-019-4448-y] [Medline: 31267179]

XSL•FO
RenderX

34. Bei den Ärztekammern registrierte Ärztinnen und Ärzte mit Gebiets- und Facharztbezeichnung (absolut, je 100.000 Einwohner und Einwohner je Arzt). Gliederungsmerkmale: Jahre, Region, Alter, Geschlecht, Gebiets-/Facharztbezeichnung, Tätigkeitsbereich: Anzahl Ärzte mit Facharztbezeichnung Haut- und Geschlechtskrankheiten. Gesundheitsberichterstattung des Bundes. 2021. URL: https://www.gbe-bund.de/gbe/ [accessed 2021-08-12]

Abbreviations

AI: artificial intelligence
BBSR: Bundesinstitut für Bau-, Stadt- und Raumforschung
BVDD: Federal Association of German Dermatologists
CROSS: Consensus-Based Checklist for Reporting of Survey Studies
DDG: German Dermatological Society
DiGAs: Digitale Gesundheitsanwendungen
eHBA: elektronischer Heilberufeausweis
TI: telematics infrastructure
WHO: World Health Organization

Edited by T de Azevedo Cardoso; submitted 20.01.23; peer-reviewed by S Xu, G Tobias; comments to author 16.06.23; revised version received 29.06.23; accepted 17.10.23; published 12.02.24.

Please cite as:
Augustin M, Reinders P, Janke TM, Strömer K, von Kiedrowski R, Kirsten N, Zink A, Otten M
Attitudes Toward and Use of eHealth Technologies Among German Dermatologists: Repeated Cross-Sectional Survey in 2019 and 2021
J Med Internet Res 2024;26:e45817
URL: https://www.jmir.org/2024/1/e45817
doi:10.2196/45817
PMID:38345855

©Matthias Augustin, Patrick Reinders, Toni Maria Janke, Klaus Strömer, Ralph von Kiedrowski, Natalia Kirsten, Alexander Zink, Marina Otten. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Acceptance and Use of Digital Technologies for Self-Reporting Medication Safety Events After Care Transitions to Home in Patients With Cancer: Survey Study

Yun Jiang[1,2], MSc, PhD; Misun Hwang[1], MSN; Youmin Cho[1,3], PhD, AGPCNP-BC; Christopher R Friese[1,2,4], PhD, AOCN; Sarah T Hawley[2,4,5,6], MPH, PhD; Milisa Manojlovich[1], PhD; John C Krauss[2,5], MD; Yang Gong[3], MD, PhD

[1]School of Nursing, University of Michigan, Ann Arbor, MI, United States
[2]Rogel Cancer Center, University of Michigan, Ann Arbor, MI, United States
[3]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States
[4]School of Public Health, University of Michigan, Ann Arbor, MI, United States
[5]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States
[6]VA Ann Arbor Center for Clinical Management Research, Ann Arbor, MI, United States

Corresponding Author:
Yun Jiang, MSc, PhD
School of Nursing
University of Michigan
400 North Ingalls Street
Ann Arbor, MI, 48109
United States
Phone: 1 734 763 3705
Email: jiangyu@umich.edu

Abstract

Background: Actively engaging patients with cancer and their families in monitoring and reporting medication safety events during care transitions is indispensable for achieving optimal patient safety outcomes. However, existing patient self-reporting systems often cannot address patients' various experiences and concerns regarding medication safety over time. In addition, these systems are usually not designed for patients' just-in-time reporting. There is a significant knowledge gap in understanding the nature, scope, and causes of medication safety events after patients' transition back home because of a lack of patient engagement in self-monitoring and reporting of safety events. The challenges for patients with cancer in adopting digital technologies and engaging in self-reporting medication safety events during transitions of care have not been fully understood.

Objective: We aim to assess oncology patients' perceptions of medication and communication safety during care transitions and their willingness to use digital technologies for self-reporting medication safety events and to identify factors associated with their technology acceptance.

Methods: A cross-sectional survey study was conducted with adult patients with breast, prostate, lung, or colorectal cancer (N=204) who had experienced care transitions from hospitals or clinics to home in the past 1 year. Surveys were conducted via phone, the internet, or email between December 2021 and August 2022. Participants' perceptions of medication and communication safety and perceived usefulness, ease of use, attitude toward use, and intention to use a technology system to report their medication safety events from home were assessed as outcomes. Potential personal, clinical, and psychosocial factors were analyzed for their associations with participants' technology acceptance through bivariate correlation analyses and multiple logistic regressions.

Results: Participants reported strong perceptions of medication and communication safety, positively correlated with medication self-management ability and patient activation. Although most participants perceived a medication safety self-reporting system as useful (158/204, 77.5%) and easy to use (157/204, 77%), had a positive attitude toward use (162/204, 79.4%), and were willing to use such a system (129/204, 63.2%), their technology acceptance was associated with their activation levels (odds ratio [OR] 1.83, 95% CI 1.12-2.98), their perceptions of communication safety (OR 1.64, 95% CI 1.08-2.47), and whether they could receive feedback after self-reporting (OR 3.27, 95% CI 1.37-7.78).

Conclusions: In general, oncology patients were willing to use digital technologies to report their medication events after care transitions back home because of their high concerns regarding medication safety. As informed and activated patients are more

XSL•FO
RenderX

likely to have the knowledge and capability to initiate and engage in self-reporting, developing a patient-centered reporting system to empower patients and their families and facilitate safety health communications will help oncology patients in addressing their medication safety concerns, meeting their care needs, and holding promise to improve the quality of cancer care.

(J Med Internet Res 2024;26:e47685) doi:10.2196/47685

KEYWORDS

digital technology; patient safety; patient participation; patient-reported outcomes; drug-related side effects and adverse reactions

Introduction

Background

The rapid growth in cancer treatment options has contributed to improved survival but has increased the complexity of care [1]. Most adults with cancer now receive their treatments in outpatient settings, and an increasing number of patients take cancer medications orally in their homes [2]. This shift in cancer care increases the likelihood of transitions across diverse settings, including primary care facilities, cancer centers, community infusion clinics, and homes [3]. Frequent care transitions may lead to medication safety events owing to inaccurate medication information sharing or poor communication [4]. Patients often face challenges in managing complex dosage schedules of their cancer medications, potentially life-threatening toxicities, and highly incident drug-drug and drug-food interactions at home [5]. Furthermore, many patients with cancer take concomitant medications for other chronic conditions [6]. During care transitions, these medications may be stopped, started, or changed, and unintentional changes at these interfaces can lead to discrepancies, which may, in turn, lead to adverse medication events. Currently, there is a lack of complete understanding of the nature, scope, and causes of medication safety events that patients with cancer experience at home because of a lack of patient engagement in self-reporting safety events from home [7].

Active engagement of patients with cancer and their families in self-monitoring and reporting adverse medication events from home is indispensable to achieving safe and effective care transitions and optimal patient outcomes [8]. However, patient and family engagement has been limited, especially in oncology settings [9-11]. There are numerous barriers to such engagement, including knowledge, attitudes and beliefs, health literacy, cultural differences, sex, age, education, economic status, and disease and symptom burdens [8,12]. Moreover, patients may be unable to engage when they receive conflicting recommendations or are excluded from the care-planning process [8]. As a result, patients may withhold their concerns regarding medication safety, be unwilling or unable to report to clinicians, or even fear reprisals from clinicians [7,13].

A patient-centered medication safety self-reporting system can guide and engage patients with cancer in self-management and reporting their experiences and concerns regarding medication events. The Chronic Care Model (CCM) highlights digital technologies' support for productive communications between informed, activated patients and a well-prepared, proactive practice team to improve outcomes [14]. Furthermore, emerging evidence demonstrates the effectiveness of patient-facing technology solutions to empower the patient's well-being and help strengthen the relationship and communication between patients with cancer and their health care providers [15,16]. However, there are challenges in initiating and engaging people with cancer using technology systems for health self-monitoring and health communications [17-19]. A 2017 pilot study of a web- and telephone-based safety reporting system received only 37 reports in 17 months [20]. Lessons from this pilot project include increased patient engagement and the focus on high-risk and high-reward populations at risk for notable adverse events [20,21], which are particularly applicable to patients with cancer during care transitions. However, the challenges faced by patients with cancer in adopting digital technologies and engaging in medication safety event self-reporting during transitions of care have not been fully understood.

Existing electronic patient-reported outcome systems for patients with cancer are often limited to reporting preselected common symptoms using survey questionnaires, which are not able to address patients' various experiences and concerns regarding medication safety during care transitions. In addition, these systems are not designed for patients to initiate timely self-reporting [21-23]. The willingness of oncology patients to use digital technologies to report medication safety events from home settings has been less studied [23,24]. The literature from other clinical settings suggests 3 prerequisites regarding the psychosocial aspects for patient willingness to share safety concerns: cognitive-cultural conditions (eg, patients' understanding and prioritization of patient safety); structural-procedural conditions (eg, the opportunity, means, and ease of providing feedback); and learning and change conditions (eg, feeling that their feedback would be acted upon and make a difference to patient safety) [7]. These prerequisite conditions can be further explored among patients with cancer to identify factors associated with their willingness to engage in medication safety event self-reporting after their care transitions to home.

Objectives

In this survey study, we aimed to assess oncology patients' perceptions of medication and communication safety during care transitions and their willingness to use digital technologies to self-report medication safety events after care transitions to home and to identify factors associated with their technology acceptance. This study's findings support subsequent development and testing of personalized technology systems for patients with cancer to self-report medication safety events to improve patient-centered cancer care, especially during care transitions.

XSL•FO
RenderX

Methods

Study Design and Participants

A cross-sectional survey study was conducted with patients with cancer who had received care at the University of Michigan Rogel Cancer Center, a National Cancer Institute–designated Comprehensive Cancer Center in the Midwest, from December 10, 2021, to August 30, 2022. This survey study was conducted following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [25] (Multimedia Appendix 1). Eligible participants were screened from electronic health records based on the following inclusion criteria: (1) receiving a diagnosis of invasive colorectal, lung, breast, or prostate cancer; (2) being discharged from the hospital or clinic to a home setting in the past 1 year; and (3) being aged ≥18 years. The 4 cancer types highlight the diversity and representation of participants, and the initial eligibility screening was determined based on the *International Classification of Diseases, Tenth Revision* codes. Multimedia Appendix 2 lists the corresponding *International Classification of Disease, Tenth Revision* codes for the 4 types of cancer.

Recruitment and Survey Administration

Figure 1 shows the survey participant recruitment and enrollment process. We recruited participants through convenience sampling. A total of 696 patients were initially identified from the medical record review, of whom 11 (1.6%) were excluded for the following reasons: poor health conditions (4/696, 0.6%), non–English speaker (4/696, 0.6%), or enrolled in an ongoing clinical trial (3/696, 0.4%). An additional 26 (3.7%) patients were excluded at their health care provider's discretion, resulting in 659 (94.7%) patients who were contacted via either a phone call or an email invitation. Of these 659 patients, 251 (38.1%) expressed their willingness to participate in the survey. Of the 251 participants, 17 (6.8%) withdrew from the study owing to their busy schedules and 30 (11.9%) were lost to follow-up. A total of 204 (81.3%) participants completed the surveys, including 181 (88.7%) participants who submitted using the web-based Qualtrics survey platform, 55 (26.9%) via email, and 9 (4.4%) via phone call.

Figure 1. Flowchart of survey participant recruitment and enrollment process.

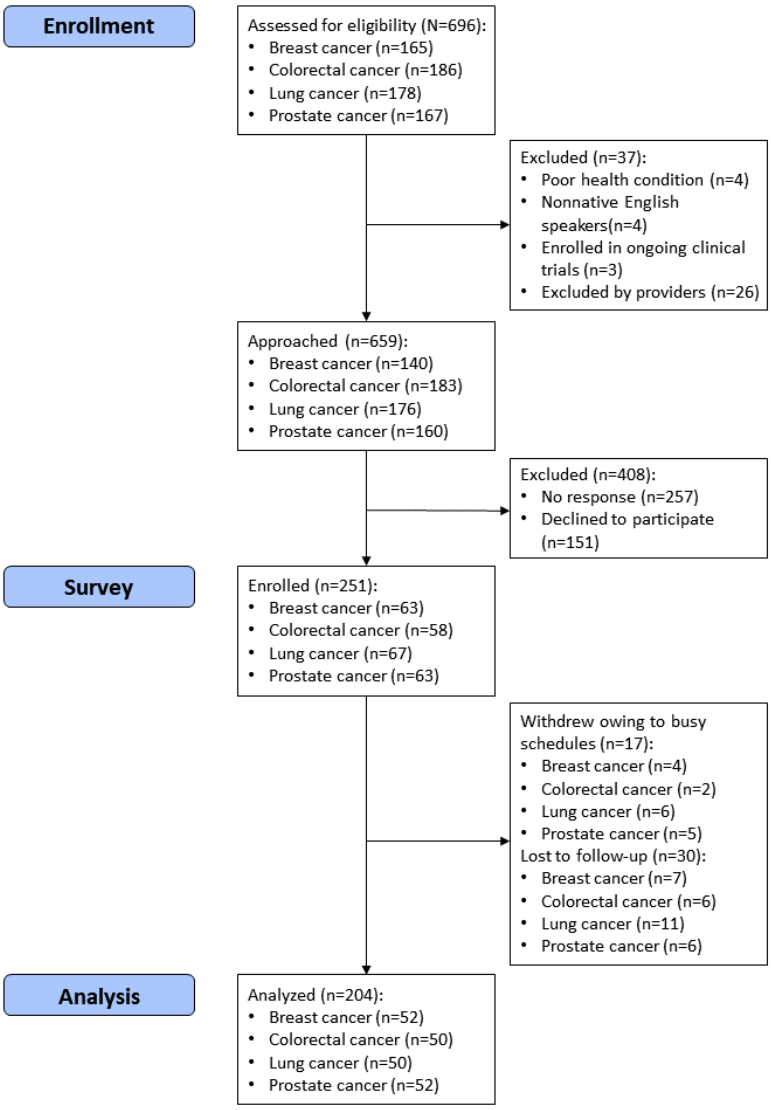

XSL•FO
RenderX

Ethical Considerations

This study was determined to be an exempt study by the University of Michigan Institutional Review Board (HUM00203239). Informed consent was obtained from all participants before data collection via phone, email, or the internet. Anonymized survey data were used for analysis. The respondents who completed the survey were given a US $25 gift card to appreciate their time and effort.

Measures

The survey included 66 items that were generated by the research team or adapted from the literature. Participants' acceptance of using digital technologies for self-reporting medication safety events from home as the primary outcome was assessed after a short scenario that described the functionalities of a web-based medication safety event self-reporting system and how it worked. In total, 4 questions were asked about participants' (1) perceived usefulness (ie, "How likely would you consider this online reporting tool is useful?"), (2) perceived ease of use ("How likely would you consider it is easy to use such an online reporting tool?"), (3) attitudes toward use ("How would you think about reporting your safety experiences or concerns through such an online reporting tool?"), and (4) intention to use this self-reporting system ("Do you intend to use an online reporting tool to report your safety concerns?"), using a 5-point Likert scale, ranging from "very unlikely" to "very likely." A total of 3 prerequisites, including cognitive-cultural conditions (eg, medication safety perception, perceived safety of communication with physicians, self-rated health, patient activation, and medication self-management ability); structural-procedural conditions (eg, experience with web-based self-reporting systems); and learning and change conditions (eg, perceived importance of feedback and influence of others' self-reporting responses), were assessed as potential psychosocial factors. Specifically, medication safety perceptions (4 items), perceived safety of communication with a physician (2 items) [7], self-rated health status (4 items), beliefs about medications (10 items) [26], patient activation (13 items) [27], and medication self-management ability (10 items) [28] were assessed as cognitive-cultural conditions. Prior technology use experiences (6 items), including prior experience with web-based health-related information self-reporting systems (1 item), were assessed as the structural-procedural conditions. Learning and change condition measures included the perceived importance of feedback on their reports (3 items) and the perceived influence of seeing others' self-reporting responses (1 item). Furthermore, we administered a brief sociodemographic questionnaire (9 items) to collect personal information, such as age, sex, race, ethnicity, income, education, marital status, and employment status. We extracted clinical factors from patients' electronic medical records, including cancer types and whether they were taking oral anticancer agents (OAAs) and other outpatient medications. The participants were informed that they could decline to answer any question that they preferred not to respond to, and they also had the choice to stop the survey at any point. Refer to Multimedia Appendix 3 for the survey questionnaires.

Statistical Analysis

The characteristics of participants were summarized using descriptive statistics (ie, mean, SD, frequency, and percentage). The associations between each potential factor and participants' safety perceptions and acceptance of the medication safety self-reporting system were assessed using bivariate correlation analyses. Multiple logistic regression was conducted to identify the adjusted associations between factors and each technology acceptance variable (ie, perceived usefulness, perceived ease of use, attitude toward use, and intention to use, recoded as "likely" vs "unlikely or uncertain") after controlling for all personal and clinical characteristics and selected psychosocial factors that presented $P<.20$ in bivariate analyses. Two-sided P values of $\leq .05$ were considered statistically significant. All statistical analyses were conducted using Stata SE software (version 17.0; StataCorp), and the correlation matrix figure was generated by R software (version 4.2.2; R Foundation for Statistical Computing). A power analysis was conducted to justify a sample size of 204 participants, which was deemed sufficient to identify the estimated squared multiple correlation coefficient of 0.35 through multiple logistic regression using the method introduced by Hsieh et al [29].

Results

Summary of Sample Characteristics and Covariate Factors

Personal and Clinical Factors

Table 1 provides a summary of participants' personal and clinical characteristics. The mean age of participants was 65.2 (SD 11) years. The sample had a slightly higher number of female participants (108/204, 52.9%), a majority of whom were White participants (179/204, 87.7%), college educated or above (108/204, 52.9%), currently married or living as married (153/204, 75%), and without a full-time or part-time job (150/204, 73.5%). The diagnosis of lung, breast, prostate, and colorectal cancer was approximately equally distributed among the participants. Most participants were taking OAAs currently or previously (140/204, 68.6%) and taking other outpatient medications (146/204, 71.6%) currently.

XSL•FO
RenderX

Table 1. Summary of personal and clinical factors of the study participants (N=204).

Characteristics	Participants
Age (y), mean (SD)	65.2 (11.0)
Sex, n (%)	
Male	96 (47.1)
Female	108 (52.9)
Race, n (%)	
Racial and ethnic minority individuals[a]	25 (12.3)
White[a]	179 (87.7)
Marital status, n (%)	
Currently married or living as married	153 (75)
Currently unmarried	51 (25)
Educational background, n (%)	
College graduate or postgraduate	108 (52.9)
Grade school, high school, or some college	96 (47.1)
Occupation, n (%)	
Unemployed, retired, disabled, or others	150 (73.5)
Full-time or part-time employee	54 (26.5)
Cancer type, n (%)	
Colorectal cancer	50 (24.5)
Prostate cancer	52 (25.5)
Lung cancer	50 (24.5)
Breast cancer	52 (25.5)
Taking OAAs[b] currently or previously, n (%)	
Yes	140 (68.6)
No	64 (31.4)
Taking other outpatient medications currently, n (%)	
Yes	146 (71.6)
No	58 (28.4)

[a]Racial and ethnic minority individuals include Black or African American, Asian, or those who identified with >1 race.

[b]OAA: oral anticancer agent.

Psychosocial Factors

Table 2 provides a summary of participants' psychosocial factors represented as 3 conditions: cognitive-cultural conditions (eg, medication safety perception, perceived safety of communication with physicians, self-rated health, patient activation, and medication self-management ability); structural-procedural conditions (eg, experience of web-based self-reporting systems); and learning and change conditions (eg, perceived importance of feedback and influence of others' self-reporting responses).

XSL•FO
RenderX

Table 2. Summary of psychosocial factors of the study participants (N=204).

Categories	Participants
Medication safety perception, mean (SD)[a]	4.4 (0.48)
Perceived safety of communication with a physician, mean (SD)[a]	4.5 (0.59)
Self-rated health, n (%)	
Good, very good, or excellent	131 (64.2)
Poor or fair	73 (35.8)
Patient activation level, n (%)	
1 (disengaged and overwhelmed)	15 (7.4)
2 (becoming aware but still struggling)	50 (24.5)
3 (taking action and gaining control)	92 (45.1)
4 (maintaining behaviors and pushing further)	47 (23)
Medication self-management, mean (SD)[b]	9.1 (2.0)
Medication self-management ability, n (%)	
≥10 (adequate)	95 (46.6)
<10 (inadequate)	109 (53.4)
Experience of web-based self-reporting systems, n (%)	
Yes	76 (37.3)
No	128 (62.7)
Perceived importance of feedback, n (%)	
Very important	161 (78.9)
Others	43 (21.1)
Influence of others' self-reporting responses, n (%)	
A lot or little	103 (50.5)
Not at all or do not know	101 (49.5)

[a]Scores range from 1 to 5, with higher scores indicating higher safety perception.

[b]Scores range from 0 to 12, with higher scores indicating a higher level of medication self-management.

Regarding cognitive-cultural conditions, participants reported strong perceptions of medication safety (mean 4.4, SD 0.48) and perceived safety of communication with a physician (mean 4.5, SD 0.59). More than half of the participants (131/204, 64.2%) reported good or better self-rated health, at least 68.1% (139/204) reported level 3 patient activation (ie, being able to take action to maintain and improve health), and approximately 46.6% (95/204) had adequate medication self-management ability. Regarding structural-procedural conditions, more than one-third of participants (76/204, 37.3%) had prior experience using web-based self-reporting systems. As for learning and change conditions, most participants (161/204, 78.9%) perceived receiving feedback on their self-reporting as very important, and approximately half of the participants (103/204, 50.5%) considered that seeing other people's self-reporting responses would influence their self-reporting.

Correlations Among Factors

Figure 2 shows the results of bivariate correlation analyses between variables. Specifically, perceived medication safety and perceived safety of communication with a physician were moderately correlated with each other (r=0.5). Patient activation level (r=0.43) and medication self-management ability (r=0.41) were positively correlated with perceived medication safety and perceived safety of communication with a physician (r=0.33 and r=0.2, respectively).

XSL•FO
RenderX

Figure 2. Correlation coefficient between variables. *Perceived safety of communication with a physician; **experience of web-based reporting system; ***perceived importance of feedback; ****influence of other's reporting responses.

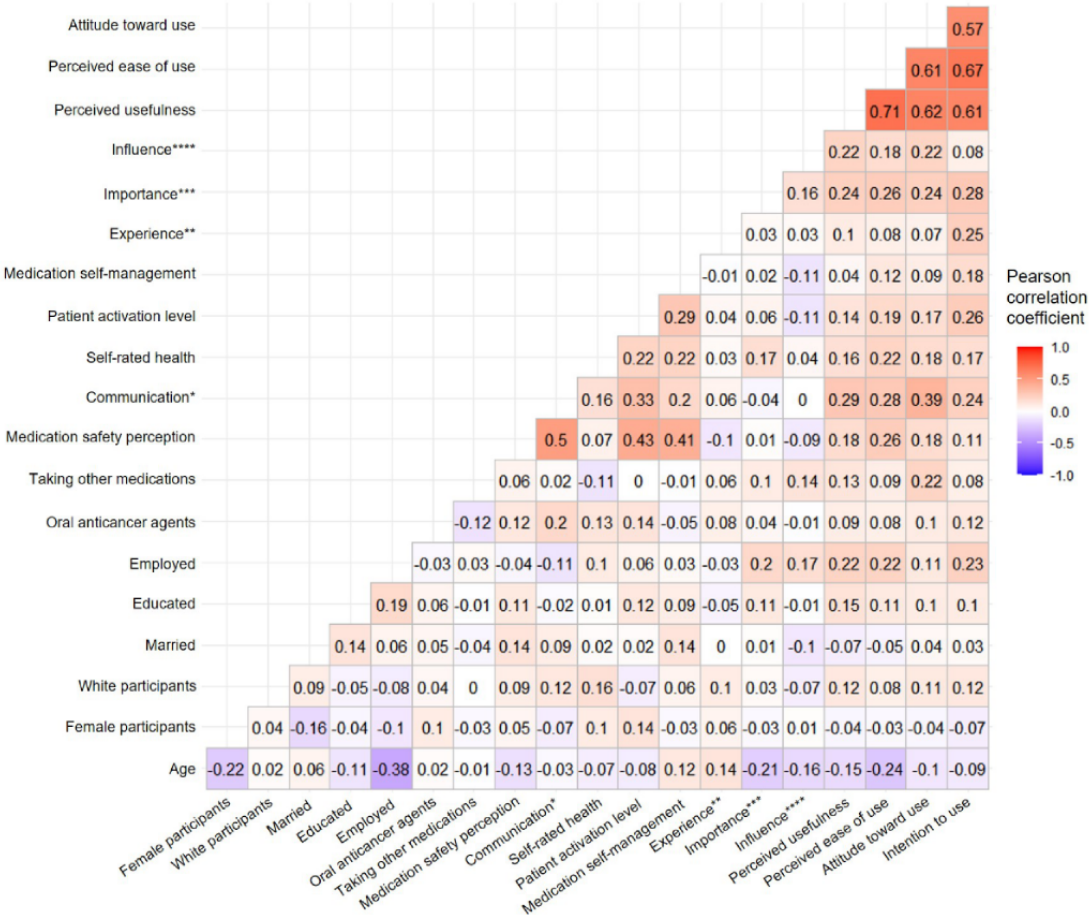

Summary of Technology Acceptance

As presented in Table 3, most participants perceived a medication safety event self-reporting system as useful (158/204, 77.5%) and easy to use (157/204, 77%). The majority (162/204, 79.4%) reported positive attitudes toward use. More than half of the participants (129/204, 63.2%) were willing to use the self-reporting system to report medication safety events or concerns.

Table 3. Summary of technology acceptance (N=204).

Categories	Participants, n (%)
Perceived usefulness	
Very likely or likely	158 (77.5)
Very unlikely, unlikely, or uncertain	46 (22.5)
Perceived ease of use	
Very likely or likely	157 (77)
Very unlikely, unlikely, or uncertain	47 (23)
Attitude toward using	
Positive	162 (79.4)
Negative or neither	42 (20.6)
Intention to use	
Very likely or likely	129 (63.2)
Very unlikely, unlikely, or uncertain	75 (36.8)

XSL•FO
RenderX

Factors Associated With Technology Acceptance

Overview

Bivariate analyses indicated that several factors, including perceived safety of communication with a physician, self-rated health, patient activation, and perceived importance of receiving feedback, were significantly correlated with all technology acceptance variables, including perceived usefulness, perceived ease of use, attitude toward use, and intention to use (correlation coefficients ranged from r=0.14 to r=0.39). Medication self-management ability (r=0.18) and prior experience with web-based self-reporting systems (r=0.25) were correlated with intention to use only. Medication safety perception and the influence of seeing others' self-reporting responses were significantly correlated with perceived usefulness, perceived ease of use, and attitude toward use (correlation coefficients ranged from r=0.18 to r=0.26) but not with the intention to use. Among other personal or clinical factors, age was negatively associated with perceived usefulness and ease of use (r=−0.15 and r=−0.24, respectively), and taking other outpatient medications was associated with a positive attitude toward use (r=0.22). All technology acceptance and intention to use variables were strongly correlated with each other, with coefficients ranging from r=0.57 to r=0.71 (Figure 2).

Perceived Usefulness

Logistic regression modeling indicated that the perceived safety of communication with a physician (odds ratio [OR] 3.59, 95% CI 1.52-8.48), the importance of receiving their self-reporting feedback (OR 2.59, 95% CI 1.03-6.51), and the influence of viewing others' self-reporting responses (OR 2.53, 95% CI 1.08-5.91) were independent predictors of the perceived usefulness of the web-based self-reporting system, after controlling for other variables in the model. In addition, the odds of perceived usefulness were 73% lower among participants who were married or those who were living as married (OR 0.27, 95% CI 0.09-0.81), higher among those with a college graduate or postgraduate educational background (OR 2.55, 95% CI 1.06-6.12), and higher among those with a full-time or part-time occupation (OR 6.64, 95% CI 1.65-26.79). Participants' clinical factors were not associated with the perceived usefulness of the web-based self-reporting system (Table 3).

Perceived Ease of Use

Logistic regression modeling indicated that the perceived medication safety (OR 3.37, 95% CI 1.07-10.58) and safety of communication with a physician (OR 2.44, 95% CI 1.05-5.68) were independent predictors of perceived ease of use of the web-based self-reporting system, after controlling for other variables in the model. In addition, the odds of perceived ease of use were 4 times higher among patients with a full-time or part-time occupation than among those without an occupation (OR 4, 95% CI 1.05-15.17). The participants' clinical factors were not associated with the perceived ease of use of the web-based self-reporting system (Table 3).

Attitude Toward Use

Logistic regression modeling indicated that the perceived safety of communication with a physician (OR 9.06, 95% CI 2.95-27.84), the importance of receiving their self-reporting feedback (OR 2.99, 95% CI 1.11-8.05), and the influence of viewing others' self-reporting responses (OR 3.36, 95% CI 1.3-8.68) were independent predictors of attitude toward use, after controlling for other variables in the model. In addition, the odds of having a positive attitude toward use were 4.23 times higher among patients taking outpatient medications (OR 4.23, 95% CI 1.63-10.99). Participants' personal factors were not associated with attitude toward use (Table 3).

Intention to Use

Logistic regression modeling indicated that the perceived safety of communication with a physician (OR 2.68; 95% CI 1.17-6.11), patient activation (OR 1.83, 95% CI 1.12-2.98), previous experience with the web-based self-reporting system (OR 3.8, 95% CI 1.73-8.36), and perceived importance of receiving their self-reporting feedback (OR 3.27, 95% CI 1.37-7.78) were independent predictors of intention to use the web-based self-reporting system, after controlling for other variables in the model. In addition, the odds of intention to use were 3.4 times higher among patients with a full-time or part-time occupation than among those without an occupation (OR 3.4, 95% CI 1.26-9.19). Participants' clinical factors were not associated with intention to use the web-based self-reporting system (Table 4).

XSL•FO
RenderX

Table 4. Factors associated with technology acceptance.

Factors	Perceived usefulness, OR[a] (95% CI)	Perceived ease of use, OR (95% CI)	Attitude toward use, OR (95% CI)	Intention to use, OR (95% CI)
Age (y)	1.01 (0.96-1.06)	0.97 (0.92-1.02)	1.01 (0.97-1.07)	0.99 (0.95-1.02)
Sex				
Male	Reference	Reference	Reference	Reference
Female	0.8 (0.32-1.96)	0.58 (0.23-1.43)	1 (0.38-2.63)	0.62 (0.28-1.37)
Race				
Racial and ethnic minority individuals	Reference	Reference	Reference	Reference
White	3.29 (0.96-11.28)	1.9 (0.53-6.8)	2.25 (0.61-8.25)	2.32 (0.79-6.77)
Marital status				
Currently unmarried	Reference	Reference	Reference	Reference
Currently married or living as married	*0.27 (0.09-0.81)* [b]	0.34 (0.12-1)	0.86 (0.29-2.54)	0.81 (0.34-1.88)
Educational background				
Grade school, high school, or some college	Reference	Reference	Reference	Reference
College graduate or postgraduate	*2.55 (1.06-6.12)*	1.56 (0.67-3.63)	1.89 (0.75-4.8)	1.22 (0.59-2.51)
Occupation				
Unemployed, retired, disabled, or others	Reference	Reference	Reference	Reference
Full-time or part-time employee	*6.64 (1.65-26.79)*	*4 (1.05-15.17)*	2.24 (0.64-7.84)	*3.4 (1.26-9.19)*
Taking OAAs[c] currently or previously				
No	Reference	Reference	Reference	Reference
Yes	1.05 (0.43-2.56)	1.07 (0.44-2.61)	1.37 (0.53-3.53)	1.55 (0.72-3.37)
Taking other medications currently				
No	Reference	Reference	Reference	Reference
Yes	1.85 (0.77-4.45)	1.53 (0.63-3.71)	*4.23 (1.63-10.99)*	1.43 (0.65-3.17)
Medication safety perception	2.01 (0.64-6.36)	*3.37 (1.07-10.58)*	0.74 (0.2-2.68)	0.51 (0.17-1.47)
Perceived safety of communication with a physician	*3.59 (1.52-8.48)*	*2.44 (1.05-5.68)*	*9.06 (2.95-27.84)*	*2.68 (1.17-6.11)*
Self-rated health				
Poor or fair	Reference	Reference	Reference	Reference
Good, very good, or excellent	1.86 (0.76-4.58)	2.4 (0.99-5.78)	1.85 (0.7-4.86)	1.23 (0.58-2.62)
Patient activation	1.02 (0.58-1.81)	1.08 (0.61-1.92)	1.09 (0.6-1.99)	*1.83 (1.12-2.98)*
Medication self-management	0.91 (0.71-1.16)	1 (0.79-1.27)	1.02 (0.79-1.33)	1.19 (0.97-1.46)
Experience of web-based self-reporting system				
No	Reference	Reference	Reference	Reference
Yes	1.9 (0.76-4.75)	2.2 (0.88-5.48)	1.14 (0.44-2.93)	*3.8 (1.73-8.36)*
Perceived importance of feedback				
Others	Reference	Reference	Reference	Reference
Very important	*2.59 (1.03-6.51)*	2.5 (1-6.29)	*2.99 (1.11-8.05)*	*3.27 (1.37-7.78)*
Influence of others' self-reporting responses				
Not at all or do not know	Reference	Reference	Reference	Reference
A lot or little	*2.53 (1.08-5.91)*	2.13 (0.92-4.92)	*3.36 (1.3-8.68)*	1.12 (0.55-2.3)

[a]OR: odds ratio.

[b]Italicized values denote the statistical significance of the *P* value (*P*<.05).

[c]OAA: oral anticancer agent.

XSL•FO
RenderX

Discussion

Principal Findings

This study demonstrated that patients with cancer had strong perceptions of medication and communication safety during transitions of care and a relatively high acceptance of digital technologies for self-reporting medication safety events after transitions back home. Furthermore, we identified significant factors associated with their technology acceptance, including patient activation, medication self-management ability, perceived medication and communication safety, perceived influence of seeing others' self-reporting responses, and perceived importance of receiving feedback. Patients with cancer often experience transitions between different care settings, which place them at risk for adverse medication events [3,4]. A patient-oriented medication safety self-reporting system has the potential to engage patients and their families in self-reporting safety events from home settings, which can consequently enhance the understanding of the nature, scope, and causes of medication safety events occurring after patients transition back home and improve patient-centered cancer care [7,30-32]. The findings of this study increased the understanding of oncology patients' willingness to adopt such a self-reporting system. This will contribute to the development of patient-facing technology systems tailored for self-reporting medication safety events. The associations between the factors, such as the perceived importance of feedback and the influence of others' self-reporting responses, and technology acceptance (perceived usefulness, perceived ease of use, attitude toward use, and intention to use) could be further translated into system functionality and data representation, which are essential tasks for engaging patients.

As indicated in this study, oncology patients had a strong sense of medication safety and communication safety with their providers. These factors were positively associated with their activation level, ability to self-manage their medications, and acceptance of technology for self-reporting medication safety events, particularly their intention to use the self-reporting system. Patients' concerns regarding their medication and communication safety were demonstrated to motivate their initiation and engagement in self-reporting medication safety events and their adoption of digital technology to improve medication safety. These findings are also perfectly aligned with the CCM, which emphasizes the effective interactions between patients who have been informed and activated and health care teams that are well-prepared and proactive, thus contributing to high-quality care outcomes [14,33]. Patient activation is not a new concept in oncology care settings. It has been reported to correlate with patients' confidence in managing their OAA side effects and subsequent adherence [34]. Interpreted by cognitive-cultural conditions for patient willingness to share safety concerns, patients and families who are well equipped with adequate patient activation and abilities for medication self-management can understand and prioritize medication safety events and actively engage in self-reporting of medication safety events after they transition back home [7,35,36]. As an independent predictor of patients' intention to use the self-reporting system, patient activation reflects their readiness and ability to be involved in their medication self-management. Therefore, to facilitate patient initiation and long-term engagement in using the medication safety event self-reporting system, the assessment and promotion of patient activation can be a fundamental design in developing the self-reporting system. It has been demonstrated that improving patient activation is feasible through continuous patient education, increased understanding of patient needs and expectations, and provision of personalized feedback and self-management recommendations [37,38]. In addition, digital technology as a tool, if designed and used appropriately, can significantly improve patients' knowledge, skills, and confidence in self-management [36]. Therefore, a well-designed and developed patient-centered medication safety self-reporting system should be able to positively affect patient activation by providing accessible and useful medication information to support patient needs [39].

It is not surprising that prior technology use experience was a predictor of patients' technology acceptance for self-reporting medication safety events, as theoretically, this factor is indicated in the Unified Technology Acceptance and Use Theory (UTAUT) [40]. Furthermore, it is understandable that patients' perceptions of communication safety with health care providers can facilitate their intention to use the safety self-reporting system, which aligns with the interpersonal process of care and communication [41,42]. One notable finding was the association between patients' perceptions of the importance of receiving their self-reporting feedback and their acceptance of the medication safety self-reporting system. Patients who preferred to receive feedback on their self-reporting were more likely to perceive the usefulness of the self-reporting system, have a positive attitude toward it, and have the intention to use it. This finding may confirm the prerequisite learning and change conditions from the literature, suggesting that a closed feedback loop between patients and clinicians is needed to improve patient safety [7,35,43]. Moreover, when patients and families serve as *vigilant partners* in medication safety self-monitoring and report their experiences and concerns after their care transitions back home, they can make a significant contribution to the understanding of the nature, scope, and causes of patient safety events outside health care systems that have been underreported historically [7,44,45].

To facilitate informed decision-making in cancer care, the medication safety self-reporting system should consider patient needs to enhance patient safety, promote self-management, and improve patient-centered care [46]. Patient-facing technologies, such as mobile health (mHealth) and wearable devices, are able to satisfy these expectations to increase the patient's access to health information, support engagement in self-management, and improve communication with health care providers [47]. From the perspective of structural-procedural conditions for patient willingness to share safety concerns, patient-facing technologies are capable of generating personalized feedback based on the patient's self-reporting and support interactive information exchanges to make the step-by-step process of self-reporting easy to follow, with downstream opportunities to increase patient engagement in the use of the system, improve patient outcomes, and reduce the cost of care in the long run

XSL•FO
RenderX

[7,35,48,49]. Certainly, it is important to design these technologies to be user-friendly, secure, and accessible to all patients, especially to those who are at high risk of adverse medication events and in high need of self-management support, that is, patients who take OAAs and experience care transitions back home, being expected to manage their cancer treatments by themselves at home [50-53]. As reported by Beauchemin et al [53], the OAA adherence rates among these patients are suboptimal, ranging from 14% to 100%, depending on the agents and measures. Suboptimal OAA adherence can lead to poor survival, severe toxicities, and increased use of health care resources [54-56]. This study solicited perceptions, willingness, and acceptance of technologies to support the subsequent development of effective programs for patient-centered and evidence-based cancer care.

Comparison With Prior Work

Constructs from the CCM, Technology Acceptance Model (TAM), and UTAUT were adopted in this study to help understand oncology patients' acceptance and use of digital technologies for self-reporting medication safety events after their care transitions back home [14,40]. Previous studies have explored the acceptance and use of mHealth or eHealth apps for self-management among the survivors of cancer and revealed whether personal factors, such as age, education, marital status, and employment status, and clinical factors, such as cancer diagnosis timeframe and that survivors are undergoing active treatment, mattered or not [22]. Consistently, this study suggested similar personal factors, including education, marital status, and employment status, for technology acceptance. However, since the previous study by de Brun et al [13] focused on general patient self-management instead of the self-reporting of medication safety events specifically, it was not able to address one of the main constructs of the CCM, which is a facilitator of involving the informed and activated persons in the process of productive interactions with health care teams, as indicated in this study [14]. Another study by Jiang et al [22] was a systematic review of the acceptance and use of home-based electronic symptom self-reporting systems by patients with cancer. This review is also guided by the CCM and UTAUT and has suggested that the interactive system features can improve patients' engagement in self-reporting, which is congruent with this study [23]. Furthermore, this review criticizes that existing home-based symptom self-reporting systems lack personalization features and only use questionnaires to collect patients' self-reporting data, which could not meet patients' various needs for self-reporting their safety concerns more conveniently and flexibly at any time [23]. This study had a similar finding, highlighting the patients' strong concerns regarding medication safety. It indicated that patients with cancer want to receive feedback on what they have reported at a personalized level. Although the TAM and UTAUT has been widely used to identify factors associated with technology system acceptance and use, its implications for the medical domain have been frequently criticized for lacking consideration of the complex health care context and need to be reassessed for additional and external factors that match with the targeted health context [57,58]. The implication of TAM and UTAUT in the context of self-reporting medication safety events in patients with cancer has not been reported before. Self-reporting of medication safety events includes but is not limited to self-reporting of symptoms (or adverse effects), medication nonadherence, medication self-administration errors, drug-drug or drug-food interactions, and safe handling or storage of OAAs. Overall, no previous study has targeted technology acceptance for self-reporting of medication safety events after oncology patients' care transitions back home [59].

Limitations

This study had several limitations. First, with a cross-sectional design, the study could not follow up with patients to understand the changes in their perception over time. It also could not demonstrate the causal relationships between the identified factors and technology acceptance (perceived usefulness, perceived ease of use, attitude toward use, and intention to use). However, the sample size of 204 participants had adequate power to reveal the potential associations among them. Furthermore, theoretical frameworks, such as the CCM, TAM, and UTAUT, were used to guide the identification and interpretation of potential factors. Second, the study was conducted at a single site, which had the potential limitation of reaching a homogenous group. Therefore, the findings may not be generalizable to other settings or regions. Third, no existing medication safety self-reporting system can be provided to assess patients' actual technology use experience. Although a scenario that describes the possible system was demonstrated in the survey, some participants may have had difficulty envisioning a pseudosystem unless they have actually used it. Therefore, longitudinal studies are needed to explore patients' acceptance and actual use of digital technologies for self-reporting medication safety events to completely understand patients' long-term engagement behaviors.

Conclusions

This study provides new insights into oncology patients' perceptions of safety during care transitions and their willingness to use digital technologies to self-report medication safety events after care transitions back home. Specifically, it highlights the importance of improving patient activation and medication self-management abilities to potentially increase their understanding and capabilities to prioritize medication safety events and engage them in using digital technology for self-reporting medication safety events. In addition, improving patients' technology use experience through appropriate training programs; assessing patients' perceptions of safety communication with health care providers; and integrating personalized features in the system design, such as providing individualized feedback on patient self-reporting, should be able to facilitate technology acceptance for self-reporting medication safety events in patients with cancer. As the informed and activated patients are more likely to have the knowledge and capability to initiate and engage in self-reporting, developing a patient-centered reporting system to empower patients and their families and facilitate safety health communications will help oncology patients in addressing their medication safety concerns, meeting their care needs, and holding promise to improve the quality of cancer care.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Agency for Healthcare Research and Quality (YJ and YG: R01HS027846) and supported by the National Cancer Institute (CRF: P30-CA-046592; T32-CA236621) and the UTHealth Houston, Cancer Prevention and Research Institute of Texas Innovation for Cancer Prevention Research Postdoctoral Fellowship (YC: RP210042). The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Agency for Healthcare Research and Quality, the National Cancer Institute, or the Cancer Prevention and Research Institute of Texas.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YJ and YG proposed the research questions and design. YJ and YC were responsible for participant selection and data acquisition. MH and YC conducted data cleaning and statistical analyses. YJ, YG, MH, and YC were involved in data interpretation and manuscript writing. CRF, STH, MM, and JCK were involved in manuscript editing and proofreading. All authors approved the final submitted version and agreed to be accountable for all aspects of this work.

Conflicts of Interest

CRF is on the Board of Governors of the Patient-Centered Outcomes Research Institute, is a member of the National Cancer Advisory Board, and receives institutional research funding from the National Cancer Institute and the NCCN Foundation or Pfizer (none of these roles are related to this manuscript). YJ is the co-editor-in-chief of JMIR Aging. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines.
[DOCX File , 33 KB - jmir_v26i1e47685_app1.docx]

Multimedia Appendix 2
International Classification of Diseases, Tenth Revision codes for the 4 eligible cancer types.
[DOCX File , 21 KB - jmir_v26i1e47685_app2.docx]

Multimedia Appendix 3
Survey questionnaires.
[PDF File (Adobe PDF File), 305 KB - jmir_v26i1e47685_app3.pdf]

References

1. Falzone L, Salomone S, Libra M. Evolution of cancer pharmacological treatments at the turn of the third millennium. Front Pharmacol 2018 Nov 13;9:1300 [FREE Full text] [doi: 10.3389/fphar.2018.01300] [Medline: 30483135]
2. Buffery D. Innovation tops current trends in the 2016 oncology drug pipeline. Am Health Drug Benefits 2016 Jun;9(4):233-238 [FREE Full text] [Medline: 27688835]
3. Rubin G, Berendsen A, Crawford SM, Dommett R, Earle C, Emery J, et al. The expanding role of primary care in cancer control. Lancet Oncol 2015 Sep;16(12):1231-1272. [doi: 10.1016/S1470-2045(15)00205-3] [Medline: 26431866]
4. Transitions of care: the need for a more effective approach to continuing patient care. Joint Commission Resources. 2012. URL: https://docplayer.net/9853408-Transitions-of-care-the-need-for-a-more-effective-approach-to-continuing-patient-care.html [accessed 2023-08-01]
5. Rogala BG, Charpentier MM, Nguyen MK, Landolf KM, Hamad L, Gaertner KM. Oral anticancer therapy: management of drug interactions. J Oncol Pract 2019 Feb;15(2):81-90. [doi: 10.1200/JOP.18.00483] [Medline: 30763198]
6. Fowler H, Belot A, Ellis L, Maringe C, Luque-Fernandez MA, Njagi EN, et al. Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. BMC Cancer 2020 Jan 28;20(1):2 [FREE Full text] [doi: 10.1186/s12885-019-6472-9] [Medline: 31987032]
7. Scott J, Heavey E, Waring J, de Brún A, Dawson P. Implementing a survey for patients to provide safety experience feedback following a care transition: a feasibility study. BMC Health Serv Res 2019 Aug 30;19(1):613 [FREE Full text] [doi: 10.1186/s12913-019-4447-9] [Medline: 31470853]

XSL•FO
RenderX

8. Backman C, Cho-Young D. Engaging patients and informal caregivers to improve safety and facilitate person- and family-centered care during transitions from hospital to home - a qualitative descriptive study. Patient Prefer Adherence 2019;13:617-626 [FREE Full text] [doi: 10.2147/PPA.S201054] [Medline: 31114175]
9. Shank B, Nguyen P, Pherson E. Transitions of care in patients with cancer. Am J Manag Care 2017;23(7 Spec No):SP280-SP284 [FREE Full text]
10. Hall J, Peat M, Birks Y, Golder S, PIPS Group, Entwistle V, et al. Effectiveness of interventions designed to promote patient involvement to enhance safety: a systematic review. Qual Saf Health Care 2010 Oct 27;19(5):e10. [doi: 10.1136/qshc.2009.032748] [Medline: 20427301]
11. Sharma AE, Rivadeneira NA, Barr-Walker J, Stern RJ, Johnson AK, Sarkar U. Patient engagement in health care safety: an overview of mixed-quality evidence. Health Aff (Millwood) 2018 Nov;37(11):1813-1820 [FREE Full text] [doi: 10.1377/hlthaff.2018.0716] [Medline: 30395509]
12. Bonetti L, Tolotti A, Anderson G, Nania T, Vignaduzzo C, Sari D, et al. Nursing interventions to promote patient engagement in cancer care: a systematic review. Int J Nurs Stud 2022 Sep;133:104289. [doi: 10.1016/j.ijnurstu.2022.104289] [Medline: 35751947]
13. de Brún A, Heavey E, Waring J, Dawson P, Scott J. PReSaFe: a model of barriers and facilitators to patients providing feedback on experiences of safety. Health Expect 2017 Aug 16;20(4):771-778 [FREE Full text] [doi: 10.1111/hex.12516] [Medline: 27860200]
14. Wagner EH. Chronic disease management: what will it take to improve care for chronic illness? Eff Clin Pract 1998;1(1):2-4. [Medline: 10345255]
15. ElKefi S, Asan O. How technology impacts communication between cancer patients and their health care providers: a systematic literature review. Int J Med Inform 2021 May;149:104430 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104430] [Medline: 33684711]
16. Penedo FJ, Oswald LB, Kronenfeld JP, Garcia SF, Cella D, Yanez B. The increasing value of eHealth in the delivery of patient-centred cancer care. Lancet Oncol 2020 May;21(5):e240-e251 [FREE Full text] [doi: 10.1016/S1470-2045(20)30021-8] [Medline: 32359500]
17. Viswanath K, Nagler RH, Bigman-Galimore CA, McCauley MP, Jung M, Ramanadhan S. The communications revolution and health inequalities in the 21st century: implications for cancer control. Cancer Epidemiol Biomarkers Prev 2012 Oct;21(10):1701-1708 [FREE Full text] [doi: 10.1158/1055-9965.EPI-12-0852] [Medline: 23045545]
18. Cho Y, Yang R, Gong Y, Jiang Y. Use of electronic communication with clinicians among cancer survivors: health information national trend survey in 2019 and 2020. Telemed J E Health 2023 Jun 10;29(6):866-874. [doi: 10.1089/tmj.2022.0203] [Medline: 36355055]
19. Howell D, Mayer DK, Fielding R, Eicher M, Verdonck-de Leeuw IM, Johansen C, Global Partners for Self-Management in Cancer. Management of cancer and health after the clinic visit: a call to action for self-management in cancer care. J Natl Cancer Inst 2021 May 04;113(5):523-531 [FREE Full text] [doi: 10.1093/jnci/djaa083] [Medline: 32525530]
20. Weingart SN, Weissman JS, Zimmer KP, Giannini RC, Quigley DD, Hunter LE, et al. Implementation and evaluation of a prototype consumer reporting system for patient safety events. Int J Qual Health Care 2017 Aug 01;29(4):521-526. [doi: 10.1093/intqhc/mzx060] [Medline: 28541430]
21. Jiang Y, Gong Y. Enhancing safety via medication self-management and timely reporting. Stud Health Technol Inform 2021 Dec 15;284:463-468 [FREE Full text] [doi: 10.3233/SHTI210773] [Medline: 34920572]
22. Jiang Y, West BT, Barton DL, Harris MR. Acceptance and use of eHealth/mHealth applications for self-management among cancer survivors. Stud Health Technol Inform 2017;245:131-135 [FREE Full text] [Medline: 29295067]
23. Cho Y, Zhang H, Harris MR, Gong Y, Smith EL, Jiang Y. Acceptance and use of home-based electronic symptom self-reporting systems in patients with cancer: systematic review. J Med Internet Res 2021 Mar 12;23(3):e24638 [FREE Full text] [doi: 10.2196/24638] [Medline: 33709929]
24. Pang NQ, Lau J, Fong S, Wong CY, Tan KK. Telemedicine acceptance among older adult patients with cancer: scoping review. J Med Internet Res 2022 Mar 29;24(3):e28724 [FREE Full text] [doi: 10.2196/28724] [Medline: 35348462]
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
26. Neame R, Hammond A. Beliefs about medications: a questionnaire survey of people with rheumatoid arthritis. Rheumatology (Oxford) 2005 Jun;44(6):762-767. [doi: 10.1093/rheumatology/keh587] [Medline: 15741193]
27. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
28. Bailey SC, Annis IE, Reuland DS, Locklear AD, Sleath BL, Wolf MS. Development and evaluation of the measure of drug self-management. Patient Prefer Adherence 2015;9:1101-1108 [FREE Full text] [doi: 10.2147/PPA.S85411] [Medline: 26257515]

XSL•FO
RenderX

29. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998 Jul 30;17(14):1623-1634. [doi: 10.1002/(sici)1097-0258(19980730)17:14<1623::aid-sim871>3.0.co;2-s] [Medline: 9699234]
30. Lee JL, Dy SM, Gurses AP, Kim JM, Suarez-Cuervo C, Berger ZD, et al. Towards a more patient-centered approach to medication safety. J Patient Exp 2018 Jun 19;5(2):83-87 [FREE Full text] [doi: 10.1177/2374373517727532] [Medline: 29978022]
31. Hahn-Goldberg S, Chaput A, Rosenberg-Yunger Z, Lunsky Y, Okrainec K, Guilcher S, et al. Tool development to improve medication information transfer to patients during transitions of care: a participatory action research and design thinking methodology approach. Res Social Adm Pharm 2022 Jan;18(1):2170-2177 [FREE Full text] [doi: 10.1016/j.sapharm.2021.04.002] [Medline: 33926826]
32. Leung K, Lu-McLean D, Kuziemsky C, Booth RG, Collins Rossetti S, Borycki E, et al. Using patient and family engagement strategies to improve outcomes of health information technology initiatives: scoping review. J Med Internet Res 2019 Oct 08;21(10):e14683 [FREE Full text] [doi: 10.2196/14683] [Medline: 31596241]
33. Flieger SP, Thomas CP, Prottas J. Improving interorganizational coordination between primary care and oncology: adapting a chronic care management model for patients with cancer. Med Care Res Rev 2021 Jun 28;78(3):229-239. [doi: 10.1177/1077558719870699] [Medline: 31462141]
34. Salgado TM, Mackler E, Severson JA, Lindsay J, Batra P, Petersen L, et al. The relationship between patient activation, confidence to self-manage side effects, and adherence to oral oncolytics: a pilot study with Michigan oncology practices. Support Care Cancer 2017 Jun 20;25(6):1797-1807. [doi: 10.1007/s00520-017-3584-0] [Medline: 28108821]
35. Bourgeois F, Fossa A, Gerard M, Davis ME, Taylor YJ, Connor CD, et al. A patient and family reporting system for perceived ambulatory note mistakes: experience at 3 U.S. healthcare centers. J Am Med Inform Assoc 2019 Dec 01;26(12):1566-1573 [FREE Full text] [doi: 10.1093/jamia/ocz142] [Medline: 31504576]
36. van der Hout A, van Uden-Kraan CF, Holtmaat K, Jansen F, Lissenberg-Witte BI, Nieuwenhuijzen GA, et al. Role of eHealth application Oncokompas in supporting self-management of symptoms and health-related quality of life in cancer survivors: a randomised, controlled trial. Lancet Oncol 2020 Jan;21(1):80-94. [doi: 10.1016/S1470-2045(19)30675-8] [Medline: 31838009]
37. Bates JS, Auten J, Sketch MR, Patel T, Clark SM, Morgan KP, et al. Patient engagement in first cycle comprehensive chemotherapy consultation pharmacist services and impact on patient activation. J Oncol Pharm Pract 2019 Jun 26;25(4):896-902. [doi: 10.1177/1078155219832644] [Medline: 30808278]
38. Melissant HC, Verdonck-de Leeuw IM, Lissenberg-Witte BI, Konings IR, Cuijpers P, Van Uden-Kraan CF. 'Oncokompas', a web-based self-management application to support patient activation and optimal supportive care: a feasibility study among breast cancer survivors. Acta Oncologica 2018 Feb 16;57(7):924-934. [doi: 10.1080/0284186x.2018.1438654]
39. Imeri H, Desselle S, Hetemi D, Hoti K. Mobile electronic devices as means of facilitating patient activation and health professional empowerment related to information seeking on chronic conditions and medications: qualitative study. JMIR Mhealth Uhealth 2021 Aug 11;9(8):e26300 [FREE Full text] [doi: 10.2196/26300] [Medline: 34383686]
40. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
41. Poon B, Kroenke C, Hillyer G. Patient experience of interpersonal processes of care and subsequent utilization of hormone therapy for non-metastatic breast cancer. Cancer Epidemiol Biomarkers Prev 2021;30(4):806-807. [doi: 10.1158/1055-9965.epi-21-0214]
42. Gallups SF, Demirci J, Nilsen M, Burke J, Bender C, Rosenzweig MQ. Using concept mapping to explore interpersonal communication components of patient navigation in breast cancer care. Cancer Nurs 2023;46(4):321-330. [doi: 10.1097/NCC.0000000000001118] [Medline: 35439221]
43. Strauss AT, Morgan C, El Khuri C, Slogeris B, Smith AG, Klein E, et al. A patient outcomes-driven feedback platform for emergency medicine clinicians: human-centered design and usability evaluation of linking outcomes of patients (LOOP). JMIR Hum Factors 2022 Mar 23;9(1):e30130 [FREE Full text] [doi: 10.2196/30130] [Medline: 35319469]
44. Schwappach DL. Review: engaging patients as vigilant partners in safety: a systematic review. Med Care Res Rev 2010 Apr 11;67(2):119-148. [doi: 10.1177/1077558709342254] [Medline: 19671916]
45. Marzban S, Najafi M, Agolli A, Ashrafi E. Impact of patient engagement on healthcare quality: a scoping review. J Patient Exp 2022 Sep 16;9:23743735221125439 [FREE Full text] [doi: 10.1177/23743735221125439] [Medline: 36134145]
46. Aapro M, Bossi P, Dasari A, Fallowfield L, Gascón P, Geller M, et al. Digital health for optimal supportive care in oncology: benefits, limits, and future perspectives. Support Care Cancer 2020 Oct 12;28(10):4589-4612 [FREE Full text] [doi: 10.1007/s00520-020-05539-1] [Medline: 32533435]
47. Lyles CR, Adler-Milstein J, Thao C, Lisker S, Nouri S, Sarkar U. Alignment of key stakeholders' priorities for patient-facing tools in digital health: mixed methods study. J Med Internet Res 2021 Aug 26;23(8):e24890 [FREE Full text] [doi: 10.2196/24890] [Medline: 34435966]
48. Ahern DK, Woods SS, Lightowler MC, Finley SW, Houston TK. Promise of and potential for patient-facing technologies to enable meaningful use. Am J Prev Med 2011 May;40(5 Suppl 2):S162-S172. [doi: 10.1016/j.amepre.2011.01.005] [Medline: 21521591]

XSL•FO
RenderX

49. Yang Y, Asan O. Designing patient-facing health information technologies for the outpatient settings: a literature review. J Innov Health Inform 2016 Apr 06;23(1):185 [FREE Full text] [doi: 10.14236/jhi.v23i1.185] [Medline: 27348487]
50. Wathne H, Morken IM, Storm M, Husebø AM. Designing a future eHealth service for posthospitalization self-management support in long-term illness: qualitative interview study. JMIR Hum Factors 2023 Feb 06;10:e39391 [FREE Full text] [doi: 10.2196/39391] [Medline: 36745492]
51. Husebø AM. Stakeholders' perspectives on eHealth support in colorectal cancer survivorship: qualitative interview study. JMIR Cancer 2021 Sep 07;7(3):e28279 [FREE Full text] [doi: 10.2196/28279] [Medline: 34491210]
52. Heffner C, Dillaman M, Hill J. Pharmacist-driven medication reconciliation reduces oral oncolytic medication errors during transitions of care. Am J Health Syst Pharm 2020 Nov 16;77(Supplement_4):S100-S104. [doi: 10.1093/ajhp/zxaa168] [Medline: 32725142]
53. Beauchemin MP, Lichtenstein MR, Raghunathan R, Doshi SD, Lee S, Law C, et al. Impact of a hospital-based specialty pharmacy in partnership with a care coordination organization on time to delivery and receipt of oral anticancer drugs. JCO Oncol Pract 2023 Mar;19(3):e326-e335. [doi: 10.1200/OP.22.00451] [Medline: 36473132]
54. Greer JA, Amoyal N, Nisotel L, Fishbein JN, MacDonald J, Stagl J, et al. A systematic review of adherence to oral antineoplastic therapies. Oncologist 2016 Mar 26;21(3):354-376 [FREE Full text] [doi: 10.1634/theoncologist.2015-0405] [Medline: 26921292]
55. McCue DA, Lohr LK, Pick AM. Improving adherence to oral cancer therapy in clinical practice. Pharmacotherapy 2014 May 18;34(5):481-494. [doi: 10.1002/phar.1399] [Medline: 24877187]
56. Jacobs JM, Pensak NA, Sporn NJ, MacDonald JJ, Lennes IT, Safren SA, et al. Treatment satisfaction and adherence to oral chemotherapy in patients with cancer. J Oncol Pract 2017 May;13(5):e474-e485. [doi: 10.1200/JOP.2016.019729] [Medline: 28398843]
57. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
58. Shachak A, Kuziemsky C, Petersen C. Beyond TAM and UTAUT: future directions for HIT implementation research. J Biomed Inform 2019 Dec;100:103315 [FREE Full text] [doi: 10.1016/j.jbi.2019.103315] [Medline: 31629923]
59. Mason M, Harris MR, Greer JA, Jiang Y. A concept analysis of oral anticancer agent self-management. Cancer Nurs 2022;45(2):E374-E387 [FREE Full text] [doi: 10.1097/NCC.0000000000000934] [Medline: 33654013]

Abbreviations

CCM: Chronic Care Model
mHealth: mobile health
OAA: oral anticancer agent
OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TAM: Technology Acceptance Model
UTAUT: Unified Technology Acceptance and Use Theory

Edited by A Mavragani; submitted 29.03.23; peer-reviewed by J Kabukye, Y Chu; comments to author 14.08.23; revised version received 18.09.23; accepted 09.02.24; published 08.03.24.

Please cite as:
Jiang Y, Hwang M, Cho Y, Friese CR, Hawley ST, Manojlovich M, Krauss JC, Gong Y
The Acceptance and Use of Digital Technologies for Self-Reporting Medication Safety Events After Care Transitions to Home in Patients With Cancer: Survey Study
J Med Internet Res 2024;26:e47685
URL: https://www.jmir.org/2024/1/e47685
doi:10.2196/47685
PMID:38457204

©Yun Jiang, Misun Hwang, Youmin Cho, Christopher R Friese, Sarah T Hawley, Milisa Manojlovich, John C Krauss, Yang Gong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Impact of In Basket Metrics on the Adoption of a New Electronic Health Record System Among Specialists in a Tertiary Hospital in Alberta: Descriptive Study

Melita Avdagovska[1], PhD; Craig Kuziemsky[2], PhD; Helia Koosha[1], MSc; Maliheh Hadizadeh[1], PhD; Robert P Pauly[3], MSc, MD; Timothy Graham[4], MSc, MD, CCFP, CHE; Tania Stafinski[1], PhD; David Bigam[5], MD; Narmin Kassam[3], MD; Devidas Menon[1], PhD

[1]School of Public Health, University of Alberta, Edmonton, AB, Canada

[2]Office of Research Services and School of Business, MacEwan University, Edmonton, AB, Canada

[3]Medicine Department, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada

[4]Department of Emergency Medicine, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada

[5]Surgery Department, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada

Corresponding Author:
Melita Avdagovska, PhD
School of Public Health
University of Alberta
11405 87 Avenue
Edmonton Clinic Health Academy
Edmonton, AB, T6G 1C9
Canada
Phone: 1 780 908 3334
Email: avdagovs@ualberta.ca

Abstract

Background: Health care organizations implement electronic health record (EHR) systems with the expectation of improved patient care and enhanced provider performance. However, while these technologies hold the potential to create improved care and system efficiencies, they can also lead to unintended negative consequences, such as patient safety issues, communication problems, and provider burnout.

Objective: This study aims to document metrics related to the In Basket communication hub (*time in In Basket per day, time in In Basket per appointment, In Basket messages received per day,* and *turnaround time*) of the EHR system implemented by Alberta Health Services, the province-wide health delivery system called Connect Care (Epic Systems). The objective was to identify how a newly implemented EHR system was used, the timing of its use, and the duration of use specifically related to In Basket activities.

Methods: A descriptive study was conducted. Due to the diversity of specialties, the providers were grouped into *medical* and *surgical* based on previous similar studies. The participants were further subgrouped based on their self-reported clinical full-time equivalent (FTE) measure. This resulted in 3 subgroups for analysis: *medical FTE <0.5*, *medical FTE >0.5*, and *surgical* (all of whom reported *FTE >0.5*). The analysis was limited to outpatient clinical interactions and explicitly excluded inpatient activities.

Results: A total of 72 participants from 19 different specialties enrolled in this study. The providers had, on average, 8.31 appointments per day during the reporting periods. The providers received, on average, 21.93 messages per day, and they spent 7.61 minutes on average in the *time in In Basket per day* metric and 1.84 minutes on average in the *time in In Basket per appointment* metric. The time for the providers to mark messages as *done* (*turnaround time*) was on average 11.45 days during the reporting period. Although the surgical group had, on average, approximately twice as many appointments per scheduled day, they spent considerably less *connected time* (based on almost all time metrics) than the medical group. However, the surgical group took much longer than the medical group to mark messages as *done* (*turnaround time).*

Conclusions: We observed a range of patterns with no consistent direction. There does not seem to be evidence of a “learning curve,” which would have shown a consistent reduction in time spent on the system over time due to familiarity and experience.

XSL•FO
RenderX

While this study does not show how the included metrics could be used as predictors of providers' satisfaction or feelings of burnout, the use trends could be used to start discussions about future Canadian studies needed in this area.

(J Med Internet Res 2024;26:e53122) doi:10.2196/53122

KEYWORDS

electronic health records; In Basket; metrics; descriptive study; inpatients

Introduction

Background

Electronic health record (EHR) systems have been implemented with many goals including streamlining information sharing among providers, empowering patients to be active partners in their care, supporting evidence-based individualized care, and monitoring population health. Health care organizations implement EHR systems with the expectation of improved patient care and enhanced provider performance [1,2].

EHR systems are not new in Canada [3]; however, their implementation has been faced with delays, changes in vendors, and reluctant adoption by users [4]. Canada continues to see activity in EHR implementation including in British Columbia [5], Saskatchewan [6], Ontario [7], Alberta (Connect Care; Epic Systems) [8], and Nova Scotia [9].

While EHR systems hold the potential to improve care delivery, they can also contribute to unintended negative consequences, such as patient safety issues, communication problems, and provider burnout [10-13]. Rather than implementing EHR systems and waiting to identify unintended consequences, we should proactively identify metrics to measure the impact of these EHR systems on the work of health care providers and enable ways to improve the diffusion and the subsequent adoption of EHR systems in Canada [9-12].

Advantages of the EHR System

EHRs are systems designed "to collect patient data in real time to enhance care by providing data at the provider's fingertips and enabling decision-making where it needs to occur" [14]. These systems provide functions such as viewing (eg, laboratory or test results); documenting (eg, entering data and notes); ordering (eg, referrals, prescriptions, and tests); web-based messaging (eg, notifying patients of test results); care management (eg, disease-specific tools and allergy alerts); analysis and reporting; and patient-directed engagement capabilities (eg, access to own laboratory values and web-based messaging care providers) [15-20]. EHRs can provide benefits such as easy access to accurate and timely point-of-care data, easy navigation to enhance workflow, automation of mundane tasks, evidence-based management pathways to individualize care, convenient sharing of data across organizations, and population health monitoring [13,19,21]. Furthermore, when fully implemented, in some instances, EHRs have resulted in the comprehensive replacement of traditional paper charts [13].

Burden of EHR System

Although EHR systems are designed to deliver positive outcomes, unintended and technology-specific negative outcomes have also been described related to the workflow, patient-provider interactions (technology seen as impersonal), and the challenges of implementing these technologies within current health care systems [10]. There is increasing awareness that physician well-being has an important impact on the health system, and concerns exist over increasing rates of burnout [11], job dissatisfaction, intention to leave practice, and job turnover [22]. Among the consistently reported drivers of burnout and job dissatisfaction are adverse clinician interactions with EHRs. While EHRs are intended to streamline workflows, they are cited as increasing the inefficiency of clinical work, adding to user frustration [23].

Studies suggest that in clinical environments that include an integrated EHR system, physicians spend an additional 1.5 hours [16] of time using the EHR system for every 1 hour of direct patient interaction, with an additional 6 to 30 hours [17,24] per month of cumulative time spent on EHR documentation and inbox management outside routine working hours [25]. Furthermore, studies have demonstrated a negative impact on providers' time due to managing test results and communications within In Basket, the EHR system's communication hub [24,26] for administrative, clerical, and documentation functions and after-hour activities related to accomplishing required tasks [23]. In Basket is where health care providers receive and manage various tasks; web-based messages; and notifications such as appointment requests, medication refill requests, laboratory or imaging results, consultation requests, and web-based messages from other health care team members [27]. Many of these tasks, previously performed by administrative staff, have been increasingly offloaded to providers.

In a primary care setting, a health care provider may receive anywhere from a few to several dozen In Basket messages per day [28]. In specialized settings, such as a hospital or specialized clinic, the volume of In Basket messages can be higher, especially for providers who are involved in complex cases or have a larger patient population [29,30]. Specialists may receive additional types of web-based messages, such as interdepartmental consult requests or referrals from primary care providers. Due to these challenges, the time spent in In Basket activities can vary depending on factors such as the volume of web-based messages, complexity of tasks, and individual work practices [26]. The number of In Basket messages received per day can vary significantly depending on several factors, such as the size of the health care organization, the specialty or department, and the individual provider's practice [31]. Managing In Basket time requires a balance between efficiency, prioritization, and effective communication to ensure timely and appropriate handling of tasks while delivering quality patient care [26]. Health care organizations have attempted to manage these challenges with one-on-one provider training, optimization and upgrade of processes,

XSL•FO
RenderX

increased availability of technical support, added or expanded use of scribes, voice recognition, and improved EHR governance [32,33].

While physician burnout and distress, more broadly, are prevalent issues in Canada [34], to our knowledge, research on EHR use and physician well-being in a Canadian context has been limited, so the relationship of EHRs to surrogate outcomes of well-being (eg, burnout) in the Canadian context is unknown. This knowledge gap is significant, given the substantial disparities between the Canadian and American health care systems, particularly concerning documentation requirements for billing, insurance, and medicolegal purposes, which EHR systems are designed to streamline [32]. Notably, while EHR system implementation is underway in Canadian hospitals, most US hospitals have already adopted these EHR systems [35,36].

Canada operates under a federated system where health governance is federal, but health care delivery and associated tasks such as EHR implementation are managed provincially. Moreover, unlike in the United States, where much of the charting is focused on billing, the Canadian health care landscape differs substantially, as billing is not part of the need for implementing EHR systems [35,36]. These contextual distinctions between the 2 health care systems mean that research findings from one setting cannot simply be extrapolated to the other [13,21,23,24,26,37].

Objectives

The extent to which EHRs contribute to physician dissatisfaction in Canada, akin to their presumed impact in the United States, remains uncertain. This study lays the groundwork for addressing this gap in knowledge by studying the use metrics of an EHR system implementation in Alberta. This study provides essential insights that not only pave the way for future investigations into the correlation between clinician well-being and EHRs in Canadian contexts but also inform interventional studies aimed at enhancing the user experience. In addition, our findings contribute to the development of best practices for EHR system implementation and use.

This study documented In Basket metrics of an EHR system implemented by Alberta Health Services (AHS), the province-wide health delivery system in Alberta branded as Connect Care. Understanding and documenting granular use metrics of Connect Care in Alberta is foundational for future studies examining the relationship between clinician wellness and EHRs in a Canadian setting, for example, understanding how the EHR-clinician interface contributes to adverse unintended consequences such as burnout. Through gaining an in-depth understanding of how the EHR system in Alberta captures In Basket metrics, this study was designed as a precursor to forthcoming studies examining the association between provider wellness and EHR systems in Canadian settings and to studies focused on improving the EHR system's user experiences and developing best practices for the EHR system rollout and subsequent use.

Methods

Study Design

This was a descriptive study of a volunteer cohort of multidisciplinary specialists working at the University of Alberta Hospital (UAH) [38]. As the goal of the study was to measure the trends in In Basket use in the EHR system from the launch of Connect Care by specialists at the UAH over 33 months of use, this study design allowed us to commence the identification of how the EHR was used, when it was used, and for how long it was used related to the In Basket activities. To the best of our knowledge, this study is the first to explore the In Basket use of Connect Care by specialists in Alberta.

Explored In Basket Domain Measures (Metrics)

In Basket metrics refer to performance indicators that assess the efficiency and effectiveness of managing tasks, web-based messages, or alerts within the EHR system. These metrics assess various aspects of workflow management and communication within the EHR environment. In this study, we used the following In Basket metrics to capture the use of the In Basket toolbar by study participants: *time spent in the In Basket per day*, *time spent in the In Basket per appointment*, number of *In Basket messages received per day*, and *turnaround Time*. The *time in In Basket per day* metric is defined as the average number of minutes a provider spends in In Basket per day. The *time in In Basket per appointment* metric is the average number of minutes a provider spends in In Basket per scheduled appointment. The *In Basket messages received per day* metric is the average number of In Basket messages a provider receives per day. The *turnaround time* metric is the average number of days a provider takes to mark a message of a specific type as done. Furthermore, In Basket metrics included *appointments per day*, which is the average number of appointments per day within the reporting period for comparison purposes (workload vs use) between the participating specialists.

Ethics Approval

Ethics approval was received from the University of Alberta Health Research Ethics Board (study ID Pro00119194), and operational approval was received from AHS (OA60778, OA60779, and OA60780).

EHR System

Connect Care is a comprehensive EHR system that allows users to access, generate, and manage documents, laboratory results, text reports, radiology images, notes, prescriptions, referrals, and web-based messages. Furthermore, Connect Care contains advanced auditing capabilities that record the actions of users when accessing the EHR system.

Study Setting

AHS is Canada's largest integrated provincial health system and is responsible for delivering health services to >4.3 million people. Health care programs and services are offered at >900 facilities throughout the province (eg, hospitals, clinics, continuing care facilities, cancer centers, mental health facilities, and community health sites) [39]. The UAH is a quaternary care research and teaching hospital in Edmonton, Alberta. This

XSL•FO
RenderX

hospital provides a wide range of inpatient and outpatient diagnostic and treatment services [40]. Study sites within the UAH were selected based on the length of time that they had been using the EHR system. The departments of medicine and surgery at the UAH were part of the first wave of the AHS Connect Care implementation. The specialists in these departments were considered to have used Connect Care for a time period that would provide sufficient use data required for this study.

Study Sample Recruitment

We decided on the following inclusion criteria for potential study participants: (1) any specialist located at the UAH and (2) ≥7 months of Connect Care use.

We used a purposive sampling method to recruit specialists. The clinical coinvestigators (RPP, DB, and NK) introduced and explained the project at departmental meetings. RPP developed a PowerPoint (Microsoft Corp) presentation, which was adapted by DB and NK to fit the context of their respective departments. During these presentations, the coinvestigators started by describing the potential impact of EHRs on provider well-being, the lack of Canadian use data, the need to understand the user experience, and the opportunity for EHR improvement driven by users. Furthermore, potential participants were informed that their individual results from the study would be shared with them. The clinical leads emailed all attending specialists asking them to complete the consent form (using REDCap [Research Electronic Data Capture]; Vanderbilt University) and provide the required information (eg, department, EHR login ID, clinical workload defined by the self-reported fraction of a full-time equivalent [FTE] measure, and work position) for data access.

Data Source

The raw In Basket data source was from Signal (an analytical platform developed by Epic Systems Corporation) using EHR user action log data (Epic Systems Corporation, unpublished data, April 2023). The user action log measures the time that the user interacted with the EHR system. The metrics captured in Signal are defined, and quantifiable measurements are used in reports to summarize information about processes or outcomes (Epic Systems Corporation, unpublished data, September 2020). Information about time spent in particular ambulatory (outpatient) In Basket activities (user action logs) was obtained for each participant from their first login to the EHR system. The analysis was limited to outpatient clinical interactions and explicitly excluded inpatient activities.

Once a specialist agreed to participate in the study, their name, login ID, and study ID were stored in a zipped and encrypted file and sent to the AHS Connect Care and Epic data team through REDCap to retrieve the required event logs data. REDCap is a secure web-based platform hosted by the Women and Children's Health Research Institute in collaboration with the Northern Alberta Clinical Trials and Research Centre at the University of Alberta. Once the Epic data team reviewed the requested information, data were pulled and transferred to the AHS Connect Care team. The anonymized data were zipped and encrypted before being transferred to the principal investigator for analysis.

Data Description

Participants

A total of 72 participants from 19 different specialties enrolled in this study. Of the 72 providers, 1 (1%) provider was excluded due to an absence of In Basket outpatient ambulatory Signal data. Due to the diversity of specialties, the providers were grouped into a *medical group* and a *surgical group* based on previous similar studies and the fact that these categories have similar EHR workflows [41,42].

The participants were further subgrouped based on their self-reported clinical FTE measure. Clinical FTE is a measure used in health care to quantify the work hours of health care providers or clinical staff in relation to a full-time position. This resulted in 3 subgroups for analysis: *medical FTE <0.5*, *medical FTE >0.5*, and *surgical* (all of whom reported FTE >0.5) groups.

In this study, providers in each group are independent of each other (ie, each provider contributes to the weighted means of only 1 group). However, for each In Basket metric, various subsets of providers in the group (ie, medical FTE <0.5, medical FTE >0.5, or surgical group) contribute to the weighted mean of various reporting periods.

Missing Values

Once the EHR data for each of the 72 providers was received, we identified missing values. As this study is one of the first to explore the provider's use of Connect Care in the Alberta context, we wanted to gain an in-depth understanding of the missing In Basket outpatient ambulatory provider-related Signal data.

On the basis of discussions with the Epic team, the study team identified 3 reasons for missing values in the data. The first reason was that a participant must be "registered" with Connect Care (AHS), be active, and must have logged in to the EHR system and seen at least 1 patient in the reporting period [38]. Second, for the *time in In Basket per appointment* metric, there was an additional inclusion criterion where the provider needed at least 5 appointments scheduled per week within the reporting period for Signal to capture user interactions in the EHR system [2]. We identified this as an issue as many part-time specialists might have ≤4 appointments per week; for example, if they were on ward duties, they would be managing only inpatients during that time. Although they interacted with the EHR system, no data would be recorded for these metrics. Since inpatient data were not studied, the true impact of EHR system use might be underestimated. The third reason for missing data is that the EHR system did not capture any data for certain metrics for all participating providers during certain months such as the *In Basket messagesreceived per day* (missing data for all providers during April 2021, May 2021, July 2021, August 2021, and September 2021) and *time in In Basket per appointment* (April 2021) metrics. Neither we nor the analysts from Epic Systems could determine the root cause of the missing data.

On the basis of these findings, we used a complete case analysis to address missing values [29]. The observations with denominator=0 were excluded. The weighted averages did not capture the missing values data. As each In Basket metric was

XSL•FO
RenderX

considered individually, a provider had to have at least 1 month of data for a particular In Basket metric to be included in the metric analysis.

Data Ranges

The start date was November 1, 2019 (the date of launch of Connect Care), and the end date was July 30, 2022, for the In Basket metrics.

Data Types

Depending on the available data for the metrics, the monthly reporting periods included in the analysis ranged between 14 and 33 months. The overall amount of data varied between 1528 (15.92%) observations for the *time in In Basket per appointment* metric and 2203 (22.95%) observations for the *time in In Basket per day* metric. The total number of observations for all included metrics was 9598.

Statistical Analysis

Data aggregation, analysis, and visualization were performed using SAS (version 9.4; SAS Institute) and Tableau (version 2021.4.3; Tableau Software, LLC) [43,44]. The numerator and denominator from each metric were used to calculate the weighted daily means of all participants and each group.

A 2-sample *t* test (2-tailed) was used to compare the weighted daily mean of every metric for the medical FTE >0.5 group with the medical FTE ≤0.5 group and compare those of the medical FTE >0.5 group with the surgical FTE >0.5 group. A weighted average calculates the mean of a data set while considering the varying importance or significance of each number within the set. This approach is commonly used in statistical analysis. It is a critical tool for addressing fluctuations, managing uneven or distorted data, and ensuring fair representation of similar data points based on their respective weights.

In time-series analysis, such as the one we have conducted, time-weighted averages were used because the time series was not evenly sampled. Ideally, data points in a time series are evenly spaced, such as hourly, daily, or monthly intervals, where each point carries equal weight. However, in our data set, reporting periods were irregular, with varying lengths ranging from 27 to 35 days. Consequently, these reporting periods had different weights. To address this, we converted the reporting periods to a daily scale, ensuring each data point carried equal weight. In summary, a time-weighted average assigns weight to each value based on its duration relative to surrounding points, leading to significantly improved accuracy in the final calculation.

Trend analysis was used to evaluate the use trends over time to determine changes in Connect Care use by the participating providers. A simple moving average (SMA) curve was used to explore the learning curves (changes over time) for each metric [45,46]. A linear trend line was fitted to the SMA curve for each group (ie, medical FTE <0.5, medical FTE >0.5, and surgical groups) based on each included metric to determine the changes in trends (ie, whether the slope increased, decreased, or remained unchanged).

In all these analyses, a *P* value of <.05 was considered statistically significant.

Results

Participant Characteristics

In total, 71 providers were included in the analysis. Of the 71 providers, 29 (40%) were women providers and 43 (60%) were men providers. The analysis did not compare results by age or gender because the numbers were small. The largest specialty group was internal medicine (n=14, 20%), followed by nephrology (n=10, 14%) and general surgery (n=9, 13%). The least represented specialties were dermatology, intensive care, neurosurgery, and cardiac surgery, at about 1% (n=1) each. Due to the diversity of specialties, the providers were grouped into a medical group (n=53, 75%) and a surgical group (n=18, 25%) based on previous similar studies (Multimedia Appendix 1) [41,42].

Furthermore, the self-reported FTE was used to further subgroup participants. Of the 53 participants in the medical group, 27 (51%) participants reported FTE <0.5 and 26 (49%) participants reported FTE >0.5. All 18 (100%) surgical specialists reported FTE >0.5. This resulted in 3 subgroups: medical FTE <0.5, medical FTE >0.5, and surgical (all FTE >0.5) groups.

Overall Results

Table 1 shows the weighted daily means for all participating providers (including weighted daily means for the medical and surgery groups) for each metric in this study. The use of weighted daily means indicates a more precise method for determining the average appointments per day compared to a simple average based solely on the number of providers and reporting periods. In this study, because the reporting periods varied in duration, they were assigned different weights based on the number of days within each period. This adjustment ensured a more accurate representation of daily appointment averages.

Table 1. Weighted daily means per metric for the medical and surgical groups.

Metric	Weighted daily mean (95% CI)		
	All specialties	Medical group	Surgical group
Appointments per day	8.31 (8.27-8.35)	6.41 (6.39-6.44)	14.01 (13.93-14.10)
In Basket messages received per day	21.93 (21.64-22.22)	22.02 (21.67-22.38)	21.70 (21.45-21.94)
Time spent in In Basket per day	7.61 (7.59-7.64)	8.86 (8.84-8.89)	3.95 (3.92-3.97)
Time spent in In Basket per appointment	1.84 (1.83-1.85)	2.78 (2.77-2.79)	0.60 (0.59-0.61)
Turnaround time	11.45 (10.83-12.08)	9.72 (9.21-10.23)	16.22 (14.69-17.76)

XSL•FO
RenderX

On the basis of the weighted daily means adjustment, each provider had, on average, 8.31 appointments per day during the entire reporting period. The providers received, on average, 21.93 web-based messages per day and spent 7.61 minutes on average in the *time in In Basket per day* metric and 1.84 minutes on average in the *time in In Basket per appointment* metric. The time for the providers to mark messages as "done" (meaning that they had completed tasks associated with them; Turnaround Time) was, on average, 11.45 days during the reporting period. Although the surgical group had, on average, approximately twice as many appointments per scheduled day, they spent considerably less "connected time" (based on almost all time metrics) than the medical group. However, the surgical group took much longer than the medical group to mark messages as done (Turnaround Time; Table 1).

Table 2 shows the weighted daily means per provider group (ie, medical FTE <0.5, medical FTE >0.5, and surgical groups) for each metric in this study. According to the raw data, the medical FTE <0.5 and the surgical groups had, on average, more appointments per day during the reporting period than the medical FTE >0.5 group. In addition, all the time metrics indicate that the medical FTE <0.5 group had less time on Connect Care than the medical group FTE >0.5. The same was observed between the medical FTE >0.5 and the surgical groups, except for the *turnaround time* metric (Table 2).

Table 2. Full-time equivalent (FTE) comparison between medical and surgical groups.

Metric	Weighted daily mean (95% CI)		
	Medical group FTE ≤0.5	Medical group FTE >0.5	Surgical group
Appointments per day	6.47 (6.44-6.49)	6.36 (6.33-6.39)	14.01 (13.93-14.10)
In Basket messages received per day	20.59 (20.39-20.78)	22.07 (21.75-22.38)	21.70[a] (21.45-21.94)
Time spent in In Basket per day	7.98 (7.95-8.02)	9.73 (9.68-9.77)	3.95 (3.92-3.97)
Time spent in In Basket per appointment	2.69 (2.68-2.71)	2.88 (2.86-2.90)	0.60 (0.59-0.61)
Turnaround time	6.23 (5.98-6.48)	12.24 (11.50-12.98)	16.22 (14.69-17.76)

[a]Surgical group versus medical FTE >0.5 group comparison: *P* value=.07.

Trend Analysis

Table 3 presents the results of the trend analysis. All 3 groups had a statistically significant increase in the appointments per day and turnaround time metrics over the study period.

As presented in Table 3, for the medical FTE ≤0.5 group, the *appointments per day*, *In Basket messages received per day*, *time in In Basket per appointment*, and *turnaround time* metrics showed statistically significant changing slopes (increasing trends over time), while the *time in In Basket per day* metric remained unchanged. The largest slope for this group was observed for the *turnaround time* metric with a value of 0.0055.

For the medical FTE >0.5 group, all metrics showed statistically significant changes (Table 3). This group showed the largest number of statistically significant trend changes among the 3 studied groups. A total of 3 metrics (ie, *appointments per day*, *In Basket messages received per day*, and *time in In Basket per day*) that showed statistically significant changes had increasing trends, while the *time in In Basket per appointment* metric showed statistically significant changes with a negative slope (decreasing trend).

Table 3. Trend analysis over time for the 3 groups: medical full-time equivalent (FTE) <0.5, medical FTE >0.5, and surgical groups.

Metric	Medical FTE <0.5 group			Medical FTE >0.5 group			Surgical group		
	Slope	*P* value	Trend	Slope	*P* value	Trend	Slope	*P* value	Trend
Appointments per day	0.0008	<.001	*Increasing*[a]	0.0014	<.001	*Increasing*	0.0015	<.001	*Increasing*
In Basket messages received per day	0.0047	<.001	*Increasing*	0.0194	<.001	*Increasing*	−0.0020	.17	Unchanged[b]
Time spent in In Basket per day	0.0002	.16	Unchanged	0.0009	<.001	*Increasing*	−0.0004	.01	*Decreasing*[c]
Time spent in In Basket per appointment	0.0003	<.001	*Increasing*	−0.0006	<.001	*Decreasing*	−0.0001	<.001	*Decreasing*
Turnaround time	0.0055	<.001	*Increasing*	0.0147	<.001	*Increasing*	0.0175	.001	*Increasing*

[a]Increasing: positive slope and *P* value is statistically significant.

[b]Unchanged: *P* value is not statistically significant.

[c]Decreasing: negative slope and *P* value is statistically significant.

For the surgical group, the *appointments per day* and *turnaroundtime* metrics showed a statistically significant

XSL•FO
RenderX

increasing trend, while the *time in In Basket per day* and *time in In Basket per appointment* metrics showed a statistically significant decreasing trend.

Although there were increasing and decreasing patterns among the included metrics, there were no obvious patterns across metrics and among groups. Therefore, there does not seem to be evidence of a "learning curve," which would have shown a consistent reduction in time spent in the EHR system over time due to familiarity and experience.

Findings by Metric

The following sections describe the findings for each metric.

Appointments Per Day

During the reporting period, the weighted daily average number of appointments per day was 8.31 (95% CI 8.27-8.35) for all providers. For the medical group, the daily weighted average was 6.41 (95% CI 6.39-6.44), while for the surgical group, this number was 14.01 (95% CI 13.93-14.10) appointments per day. The weighted daily mean for the medical FTE ≤0.5 group (mean 6.47, 95% CI 6.44-6.49), compared to the mean for the medical FTE >0.5 group (mean 6.36, 95% CI 6.33-6.39), was significantly different (Multimedia Appendix 2).

Although the slope changes were subtle, the SMA trends for the *appointments per day* metric for all 3 groups were statistically increasing over time (Multimedia Appendix 2).

In Basket Messages Received Per Day

The weighted daily mean of web-based messages received was 21.93 (95% CI 21.64-22.22) messages for all 71 providers. The weighted daily mean for the medical FTE >0.5 group was significantly larger than that for the medical FTE ≤0.5 group. Furthermore, the difference between the weighted daily mean values of the medical FTE >0.5 group (mean 23.29, 95% CI 22.70-23.57) and the surgical group (mean 21.70, 95% CI 21.45, 21.94; $P<.001$) was statistically significant (Multimedia Appendix 2).

In June 2021, Signal data recorded that 1 particular specialist received an unusually large number of In Basket messages. After an examination, it was determined that this was due to the EHR system sending a batch of all laboratory results from many patients to this particular medical specialist, who was probably on call. The spike from this individual's data is reflected in the 2 graphs related to the medical group (FTE >0.5) and the graph for all providers (Multimedia Appendix 2). While this particular case may be seen as an outlier, it serves as an illustration of what can potentially happen within an EHR system. Instances like this one may not be uncommon.

According to the SMA trend analysis, both medical groups experienced statistically significant increasing trends in this metric, while the surgery group's trend remained statistically unchanged. The trend change was much more pronounced for the medical FTE >0.5 group (slope=0.0194) than that of the medical FTE ≤0.5 group (slope=0.0047). Notably, for the medical FTE >0.5 group, this metric had the largest slope and was the fastest changing over time (Multimedia Appendix 2). This might have been because of the "anomaly" of a single physician in the medical FTE >0.5 group receiving a very large number of emails, as described in the previous paragraph. Furthermore, these results show the situations that are possible within the EHR system and need to be recognized.

Time in In Basket Per Day

The weighted daily mean for all providers was 7.61 (95% CI 7.59-7.64) minutes in In Basket per day. The weighted daily mean for the medical group was 8.86 (95% CI 8.84-8.89) minutes in In Basket per day, while that for the surgical group was 3.95 (95% CI 3.92-3.97) minutes per day. The medical FTE ≤0.5 group's weighted daily mean was 7.98 (95% CI 7.95-8.02) minutes in In Basket per day, and the weighted daily mean for the medical FTE >0.5 group ($P<.001$) was 9.73 (95% CI 9.68-9.77) minutes per day. The surgical group spent less time in In Basket per day than the medical FTE >0.5 group (mean 3.95, 95% CI 3.92-3.97, vs mean 9.73, 95% CI 9.68-9.77; $P<.001$; Multimedia Appendix 2).

On the basis of the trend analysis, the medical FTE >0.5 group showed a statistically significant increasing trend for this metric, while the surgery group showed a statistically significant decreasing trend and the medical FTE ≤0.5 group's trend stayed statistically unchanged. While the result of trend analysis for this metric is different for each group, it is important to note that the slopes for each group were very small and clinically insignificant (Multimedia Appendix 2).

Time in In Basket Per Appointment

After analyzing 1528 observations related to the time that providers spent in the In Basket per appointment, the total average time for both surgical and medical groups was 1.84 (95% CI 1.83-1.85) minutes. The surgical groups spent 0.60 (95% CI 0.59-0.61) minutes, while the medical group spent 2.78 (95% CI 2.77-2.79) minutes. The weighted daily mean for the medical FTE ≤0.5 group was significantly different compared to the mean for the medical FTE >0.5 group (mean 2.69, 95% CI 2.68-2.71, vs mean 2.88, 95% CI 2.86-2.90; $P<.001$). Furthermore, a significant difference was observed when comparing the medical FTE >0.5 group (mean 2.88, 95% CI 2.86-2.90) and the surgical group (mean 0.60, 95% CI 0.59, 0.61; $P<.001$; Multimedia Appendix 2).

For the *time in In Basket per appointment* metric, the medical FTE ≤0.5 group was the only group that saw a statistical increase in their use over time. The other 2 groups showed a statistical decrease in their use over time for this metric (Multimedia Appendix 2).

Turnaround Time

Turnaround time is a metric group under the In Basket category within Signal. It reports the average number of days a provider takes to mark a message of a specific type as "done." According to the data, the surgical group spent 16.22 (95% CI 14.69-17.76) days on average to mark messages as done. The medical group spent, on average, 9.72 (95% CI 9.21-10.23) days to mark messages as done (Multimedia Appendix 2). For this metric, a significant difference was observed when comparing the 2 medical groups and between the medical FTE >0.5 and the

XSL•FO
RenderX

surgical group. The study team was unable to identify the reasons for the delays.

For this metric, spikes in recorded data were observed (Multimedia Appendix 2) for 2 study participants (1 medical and 1 surgical specialist) over several months, indicating extremely long delays in marking received messages as "done." An explanation for these anomalies in data capture within the *turnaround time* metric remains elusive. Once more, we encounter an outlier; nonetheless, it serves as an example of potential EHR system use scenarios.

On the basis of the SMA trend analysis, all 3 groups experienced statistically increasing trends over time for their *turnaround time* metric (Multimedia Appendix 2). The largest slope (0.0175) belonged to the surgical group and the smallest slope (0.0055) belonged to the medical FTE ≤0.5 group for this metric.

Discussion

Principal Findings

Implementing Connect Care by AHS has transformed how providers capture and share information by establishing changes to workflows, processes, and charting approaches [47]. While the overall objective is to establish uniformity in the EHR system's use, this study has revealed disparities in the timing of task completion within the EHR system. Furthermore, in certain cases, outliers have emerged whose use patterns are not easily explained with the existing data. This study revealed significant gaps in our understanding of EHRs and In Basket management, highlighting the need for further exploration and comprehension in these areas.

Khairat et al [48] evaluated the time spent by general and specialist pediatricians performing clinical documentation and In Basket tasks outside work hours. Specialists spent more time in the EHR system, and "this may be because specialists see more complex patients and, therefore, need more time to review the patient chart and to respond to In Basket messages" [48]. Although in our study, we cannot say what percentage of workload the providers spent on In Basket activities, we identified that they spent 7.61 minutes in the *time in In Basket per day* metric and 1.84 minutes in the *time in In Basket per appointment* metric. According to the raw data, the medical FTE <0.5 group and the surgical group had, on average, more appointments per day during the reporting period than the medical FTE >0.5 group. It would be valuable to explore the main workflow drivers of In Basket time and try to optimize efficiency in this area for all specialties.

The proportion of time spent in the EHR system based on the included metrics between the providers within the medical groups (FTE ≤0.5 and FTE >0.5) was similar; however, little can be concluded about the similarities or differences in use due to the high variability within the specialties. Although data analysis showed statistical significance for all metrics, it is apparent that FTE made no difference to the workload between providers working (FTE ≤0.5 or FTE >0.5). Before comparing part-time medical providers with full-time ones, we could not definitively attribute the observed differences between the medical and surgical groups to the fact that some medical providers worked part time or to the fact that all surgical providers worked full time. Our study did not reveal important differences in In Basket metrics among medical specialists regardless of the clinical FTE . Significant differences were observed between medical and surgical colleagues. Presumably, these differences relate to broad differences in medical versus surgical consultation and their associated workflows.

When comparing the 3 groups, the medical FTE >0.5 group was more "connected" than the medical FTE <0.5 group and the surgical group when considering the *time in In Basket per day* and *time in In Basket per appointment* metrics. Although the surgical group treated more patients (on average, 14.01 appointments during the reporting period), they spent less time in In Basket per day and per appointment, so they were "less connected" than the 2 medical groups. Nonetheless, while the data do not provide a direct explanation for these differences, they do provide insight into the structuring of workloads. This insight is crucial for comprehending how various professionals use the EHR system and identifying areas where workflow enhancements could prove beneficial.

We identified several providers' data that were outliers in terms of their EHR use. For example, 2 providers took inordinately long times to mark received web-based messages as *done* (*turnaround time*), which impacted the data on between-group differences. Furthermore, there was a medical provider who received an extremely high number of web-based messages in June 2021. Such outliers demonstrate that certain scenarios can significantly influence the averages of various metrics, leading to skewed results. It underscores the possibility that data generated by the EHR system may not always be accurate, emphasizing the need for discussions and considerations with EHR system vendors regarding EHR functionality and measures to reduce outlier occurrences. Future research with a more robust statistical approach should be conducted to delve deeper into addressing and mitigating anomalies in the data.

One factor that we identified in our study is that Connect Care did not capture all interactions due to various vendor-imposed rules (eg, 5 appointments per reporting period). Similarly, Cohen et al [49] identified issues with vendor-derived metrics and how different vendors calculated the same activities in different ways and identified that not all EHRs (vendors) drew information from audit log data, which led to the inability to provide the whole picture of provider's interaction with the EHR system [49]. Therefore, using only vendor-derived metrics may miss important aspects of the true impact of the EHR system on users. In the study by Cohen et al [49], 1 participant stated that "if different EHR (vendors) are attacking the issue differently, you will get variations not related to burden but just how the math is done." Documentation time for In Basket use must be captured completely with the intent to understand how In Basket contributes to the overall workload of providers. If EHR systems are being associated with burnout, In Basket messages could be a starting point for common ground around the discussion of how web-based messages should be delivered and managed [50].

XSL•FO
RenderX

Future Directions

On the basis of the results from this study, we identified several future studies that can build upon this study. This study was descriptive and did not explore the correlation between the included metrics and provider satisfaction and burnout due to EHR system use. The next step would be to conduct a study exploring the circumstances around the individual EHR data. It would be valuable to explore the main workflow drivers of In Basket use time and try to optimize efficiency in this area for all specialties. A qualitative study should be conducted to explore the variances between actual and perceived EHR system use. While data from this study do not provide a direct explanation for these differences, they do provide insight into the structuring of workloads.

Furthermore, future studies should focus on the difference between providers with part-time and full-time clinical schedules and how that translates into EHR use. This insight is key for understanding how various professions use the EHR system in order to identify areas where workflow enhancements could prove beneficial. Moreover, future research should explore EHR use between different specialties and whether these specialties impact EHR use habits. In addition, studies should explore the association between other metrics and quality outcomes. Finally, future studies need to work on developing strategies for EHR data quality appraisal. In our study, we identified that the data generated by the EHR system may not always be accurate, emphasizing the need for discussions and considerations with EHR vendors regarding EHR functionality. In future studies with a more robust statistical approach, there may be an opportunity to delve deeper into addressing and mitigating anomalies in the data.

Limitations

This study has several limitations. The analyzed data were only the participating providers' ambulatory (outpatient) data. Inpatient data were not included, which might have provided additional information on some of the metrics (inpatient data were unavailable for all included metrics in this study). Another limitation is the underestimation of some metrics based on how Epic defines and captures activity (eg, a provider needs at least 5 appointments scheduled per week within the reporting period and inbox activities related to phone calls or chart review). Furthermore, to address some In Basket issues, a person may need to access other parts of the EHR system to gather more information or complete some other task (eg, write a prescription) and only then go back into the In Basket to sign off on it. Therefore, the actual time in the In Basket is a systematic underestimation of the actual time it took to complete a task.

Due to the high variability of specialties (19 in total) and the low number of recruited providers for each specialty (ranging between 1 and 14 providers), we were unable to explore and compare the differences in EHR use between the specialties. The small number of participants might have created a bias regarding the reasons for participation. Another study limitation was that the FTE was self-reported, which might have led to providers over- or underestimating their clinical schedules. The final limitation is that we did not evaluate the types of web-based messages that the providers received in the In Basket. As this is one of the first studies evaluating Connect Care, we deemed that the focus should be on the overall metrics rather than the submetrics or categories.

Conclusions

This study demonstrated the enormous promise of the ability to harvest data from an EHR that describes system use and the potential impact that it has on the workflow of physicians. To take complete advantage of this, there must be an appropriate understanding of how EHR systems capture and measure the use by providers. This would be foundational to forthcoming studies examining the association between provider wellness and EHR systems in Canadian settings and studies focused on improving the EHR system's user experiences, developing best practices for EHR systems rollout and subsequent use, and understanding how the interface of the user and the EHR systems interrelate. Although this study does not show how the included metrics could be used as predictors of providers' satisfaction or feeling of burnout, the use trends could be used to start discussions about future Canadian studies needed in this area.

Acknowledgments

The authors would like to acknowledge and thank the providers who participated in this study. Without their help, this research would not have been possible. The authors would also like to thank the staff from Alberta Health Services and Epic Corporation for their willingness to provide the required data. Furthermore, the authors would like to thank the Canadian Institutes of Health Research (grant 180993) for providing funding for this work. Without this funding support, this study would have not been possible.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Grouping of included providers according to specialties.
[DOCX File , 29 KB - jmir_v26i1e53122_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Findings by metric.
[DOCX File , 957 KB - jmir_v26i1e53122_app2.docx]

References

1. Donnelly C, Janssen A, Vinod S, Stone E, Harnett P, Shaw T. A systematic review of electronic medical record driven quality measurement and feedback systems. Int J Environ Res Public Health 2022 Dec 23;20(1):200 [FREE Full text] [doi: 10.3390/ijerph20010200] [Medline: 36612522]
2. Negro-Calduch E, Azzopardi-Muscat N, Krishnamurthy RS, Novillo-Ortiz D. Technological progress in electronic health record system optimization: systematic review of systematic literature reviews. Int J Med Inform 2021 Aug;152:104507 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104507] [Medline: 34049051]
3. Boothe C, Bhullar J, Chahal N, Chai A, Hayre K, Park M, et al. The history of technology in nursing: the implementation of electronic health records in Canadian healthcare settings. Can J Nurs Inform 2023 Sep 23;18(3) [FREE Full text]
4. Chang F, Gupta N. Progress in electronic medical record adoption in Canada. Can Fam Physician 2015 Dec;61(12):1076-1084 [FREE Full text] [Medline: 27035020]
5. Electronic medical and health records. HealthLink BC. URL: https://tinyurl.com/mrxyuky6 [accessed 2021-03-27]
6. MySaskHealthRecord. eHealth Saskatchewan. URL: https://www.ehealthsask.ca/MySaskHealthRecord/MySaskHealthRecord [accessed 2021-03-27]
7. What's an EHR? eHealth Ontario. URL: https://ehealthontario.on.ca/en/patients-and-families/ehrs-explained [accessed 2021-10-08]
8. Connect care. Alberta Health Services. URL: https://www.albertahealthservices.ca/cis/cis.aspx [accessed 2021-10-08]
9. Primary health information management and electronic medical records. Government of Nova Scotia. URL: https://novascotia.ca/dhw/primaryhealthcare/PHIM-EMR.asp [accessed 2021-10-08]
10. Buntin MB, Burke MF, Hoaglin MC, Blumenthal D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Aff (Millwood) 2011 Mar;30(3):464-471 [FREE Full text] [doi: 10.1377/hlthaff.2011.0178] [Medline: 21383365]
11. Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S, et al. Prevalence of burnout among physicians: a systematic review. JAMA 2018 Sep 18;320(11):1131-1150 [FREE Full text] [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
12. Physician burnout. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/prevention/clinician/ahrq-works/burnout/index.html [accessed 2022-01-13]
13. National Academies of Sciences, Engineering, and Medicine, Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being, National Academy of Medicine. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-being. Washington, DC: National Academies Press; 2019.
14. McGonigle D, Mastrian K. Nursing Informatics and the Foundation of Knowledge. Burlington, MA: Jones & Bartlett Learning; 2017.
15. Ouyang D, Chen JH, Hom J, Chi J. Internal medicine resident computer usage: an electronic audit of an inpatient service. JAMA Intern Med 2016 Feb;176(2):252-254 [FREE Full text] [doi: 10.1001/jamainternmed.2015.6831] [Medline: 26642261]
16. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Dec 06;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
17. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
18. Islam MM, Poly TN, Li YC. Recent advancement of clinical information systems: opportunities and challenges. Yearb Med Inform 2018 Aug;27(1):83-90 [FREE Full text] [doi: 10.1055/s-0038-1667075] [Medline: 30157510]
19. Piscotty RJ, Tzeng HM. Exploring the clinical information system implementation readiness activities to support nursing in hospital settings. Comput Inform Nurs 2011 Nov;29(11):648-656. [doi: 10.1097/NCN.0b013e31821a153f] [Medline: 21532469]
20. Sheikh A, Sood HS, Bates DW. Leveraging health information technology to achieve the "triple aim" of healthcare reform. J Am Med Inform Assoc 2015 Jul;22(4):849-856 [FREE Full text] [doi: 10.1093/jamia/ocv022] [Medline: 25882032]
21. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
22. Degen C, Li J, Angerer P. Physicians' intention to leave direct patient care: an integrative review. Hum Resour Health 2015 Sep 08;13:74 [FREE Full text] [doi: 10.1186/s12960-015-0068-5] [Medline: 26350545]
23. Eschenroeder HC, Manzione LC, Adler-Milstein J, Bice C, Cash R, Duda C, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. J Am Med Inform Assoc 2021 Apr 23;28(5):960-966 [FREE Full text] [doi: 10.1093/jamia/ocab053] [Medline: 33880534]

XSL•FO
RenderX

24. Akbar F, Mark G, Warton EM, Reed ME, Prausnitz S, East JA, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. J Am Med Inform Assoc 2021 Apr 23;28(5):923-930 [FREE Full text] [doi: 10.1093/jamia/ocaa229] [Medline: 33063087]
25. Berg S. Family doctors spend 86 minutes of “pajama time” with EHRs nightly. American Medical Association. 2017 Sep 11. URL: https://www.ama-assn.org/practice-management/digital/family-doctors-spend-86-minutes-pajama-time-ehrs-nightly [accessed 2021-10-12]
26. Tai-Seale M, Dillon E, Yang Y, Nordgren R, Steinberg R, Nauenberg T, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019 Jul;38(7):1073-1078 [FREE Full text] [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
27. Baxter SL, Saseendrakumar BR, Cheung M, Savides TJ, Longhurst CA, Sinsky CA, et al. Association of electronic health record Inbasket message characteristics with physician burnout. JAMA Netw Open 2022 Nov 01;5(11):e2244363 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.44363] [Medline: 36449288]
28. Lieu TA, Altschuler A, Weiner JZ, East JA, Moeller MF, Prausnitz S, et al. Primary care physicians' experiences with and strategies for managing electronic messages. JAMA Netw Open 2019 Dec 02;2(12):e1918287 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18287] [Medline: 31880798]
29. Nguyen OT, Turner K, Apathy NC, Magoc T, Hanna K, Merlo LJ, et al. Primary care physicians' electronic health record proficiency and efficiency behaviors and time interacting with electronic health records: a quantile regression analysis. J Am Med Inform Assoc 2022 Jan 29;29(3):461-471 [FREE Full text] [doi: 10.1093/jamia/ocab272] [Medline: 34897493]
30. Cutrona SL, Fouayzi H, Burns L, Sadasivam RS, Mazor KM, Gurwitz JH, et al. Primary care providers' opening of time-sensitive alerts sent to commercial electronic health record InBaskets. J Gen Intern Med 2017 Nov 14;32(11):1210-1219 [FREE Full text] [doi: 10.1007/s11606-017-4146-3] [Medline: 28808942]
31. Akbar F, Mark G, Prausnitz S, Warton EM, East JA, Moeller MF, et al. Physician stress during electronic health record inbox work: in situ measurement with wearable sensors. JMIR Med Inform 2021 Apr 28;9(4):e24014 [FREE Full text] [doi: 10.2196/24014] [Medline: 33908888]
32. Brunken M, Bice C. Achieving EHR satisfaction in any specialty: impact report. KLAS Research. 2019. URL: https://klasresearch.com/archcollaborative/report/achieving-ehr-satisfaction-in-any-specialty/310 [accessed 2024-04-08]
33. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 02;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
34. CMA national physician health survey: a national snapshot. Canadian Medical Association. 2018. URL: https://www.cma.ca/cma-national-physician-health-survey-national-snapshot [accessed 2024-04-08]
35. Holmgren AJ, Downing NL, Bates DW, Shanafelt TD, Milstein A, Sharp CD, et al. Assessment of electronic health record use between US and non-US health systems. JAMA Intern Med 2021 Feb 01;181(2):251-259 [FREE Full text] [doi: 10.1001/jamainternmed.2020.7071] [Medline: 33315048]
36. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):31 [FREE Full text] [doi: 10.1038/s41746-022-00573-1] [Medline: 35304561]
37. Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. J Am Med Inform Assoc 2021 Apr 23;28(5):998-1008 [FREE Full text] [doi: 10.1093/jamia/ocaa325] [Medline: 33434273]
38. Aggarwal R, Ranganathan P. Study designs: part 2 – descriptive studies. Perspect Clin Res 2019;10(1):34-36. [doi: 10.4103/picr.picr_154_18]
39. About AHS. Alberta Health Services. URL: https://www.albertahealthservices.ca/about/about.aspx [accessed 2020-08-19]
40. University of Alberta Hospital. University Hospital Foundation. URL: https://givetouhf.ca/university-of-alberta-hospital/ [accessed 2023-01-16]
41. Nath B, Williams B, Jeffery MM, O'Connell R, Goldstein R, Sinsky CA, et al. Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in New England. JAMA Netw Open 2021 Oct 01;4(10):e2131490 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.31490] [Medline: 34636917]
42. Saag HS, Shah K, Jones SA, Testa PA, Horwitz LI. Pajama time: working after work in the electronic health record. J Gen Intern Med 2019 Sep 9;34(9):1695-1696 [FREE Full text] [doi: 10.1007/s11606-019-05055-x] [Medline: 31073856]
43. SAS/STAT® 14.3 user’s guide high-performance procedures. SAS Institute Inc. 2017. URL: https://support.sas.com/documentation/onlinedoc/stat/143/stathpug.pdf [accessed 2024-04-08]
44. What is tableau? Tableau. URL: https://www.tableau.com/why-tableau/what-is-tableau [accessed 2023-01-16]
45. Dulku K, Toll E, Kwun J, van der Meer G. The learning curve of BiZact™ tonsillectomy. Int J Pediatr Otorhinolaryngol 2022 Jul;158:111155. [doi: 10.1016/j.ijporl.2022.111155] [Medline: 35584567]
46. Zou YM, Ma Y, Liu JH, Shi J, Fan T, Shan YY, et al. Trends and correlation of antibacterial usage and bacterial resistance: time series analysis for antibacterial stewardship in a Chinese teaching hospital (2009-2013). Eur J Clin Microbiol Infect Dis 2015 Apr 10;34(4):795-803. [doi: 10.1007/s10096-014-2293-6] [Medline: 25487131]

XSL•FO
RenderX

47. Frequently asked questions. Alberta Health Services. URL: https://www.albertahealthservices.ca/info/Page15938.aspx [accessed 2024-04-08]
48. Khairat S, Zalla L, Gartland A, Seashore C. Association between proficiency and efficiency in electronic health records among pediatricians at a major academic health system. Front Digit Health 2021 Sep 6;3:689646 [FREE Full text] [doi: 10.3389/fdgth.2021.689646] [Medline: 34713161]
49. Cohen GR, Boi J, Johnson C, Brown L, Patel V. Measuring time clinicians spend using EHRs in the inpatient setting: a national, mixed-methods study. J Am Med Inform Assoc 2021 Jul 30;28(8):1676-1682 [FREE Full text] [doi: 10.1093/jamia/ocab042] [Medline: 33899105]
50. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff (Millwood) 2017 Apr 01;36(4):655-662 [FREE Full text] [doi: 10.1377/hlthaff.2016.0811] [Medline: 28373331]

Abbreviations

AHS: Alberta Health Services
EHR: electronic health record
FTE: full-time equivalent
REDCap: Research Electronic Data Capture
SMA: simple moving average
UAH: University of Alberta Hospital

Edited by K Williams; submitted 10.10.23; peer-reviewed by M Zaidi, N Shaw; comments to author 26.01.24; revised version received 12.03.24; accepted 19.03.24; published 29.04.24.

Please cite as:
Avdagovska M, Kuziemsky C, Koosha H, Hadizadeh M, Pauly RP, Graham T, Stafinski T, Bigam D, Kassam N, Menon D
Exploring the Impact of In Basket Metrics on the Adoption of a New Electronic Health Record System Among Specialists in a Tertiary Hospital in Alberta: Descriptive Study
J Med Internet Res 2024;26:e53122
URL: https://www.jmir.org/2024/1/e53122
doi:10.2196/53122
PMID:38684079

©Melita Avdagovska, Craig Kuziemsky, Helia Koosha, Maliheh Hadizadeh, Robert P Pauly, Timothy Graham, Tania Stafinski, David Bigam, Narmin Kassam, Devidas Menon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Influence of Physical Attractiveness and Gender on Patient Preferences in Digital Doctor Consultations: Experimental Study

Xia Wei[1], BA, MBA, PhD; Shubin Yu[2], BA, MA, MSc, PhD; Changxu (Victor) Li[3], BA, MA, MSc

[1]College of Management, Shenzhen University, Shenzhen, China

[2]Department of Communication and Culture, BI Norwegian Business School, Oslo, Norway

[3]Department of Marketing, KU Leuven, Leuven, Belgium

Corresponding Author:
Shubin Yu, BA, MA, MSc, PhD
Department of Communication and Culture
BI Norwegian Business School
Nydalsveien 37
Oslo, 0484
Norway
Phone: 47 41228055
Email: shubin.yu@bi.no

Abstract

Background: The rise of digital health services, particularly digital doctor consultations, has created a new paradigm in health care choice. While patients traditionally rely on digital reviews or referrals to select health care providers, the digital context often lacks such information, leading to reliance on visual cues such as profile pictures. Previous research has explored the impact of physical attractiveness in general service settings but is scant in the context of digital health care.

Objective: This study aims to fill the research gap by investigating how a health care provider's physical attractiveness influences patient preferences in a digital consultation setting. We also examine the moderating effects of disease severity and the availability of information on health care providers' qualifications. The study uses signal theory and the sexual attribution bias framework to understand these dynamics.

Methods: Three experimental studies were conducted to examine the influence of health care providers' physical attractiveness and gender on patient preferences in digital consultations. Study 1 (n=282) used a 2×2 between-subjects factorial design, manipulating doctor attractiveness and gender. Study 2 (n=158) focused on women doctors and manipulated disease severity and participant gender. Study 3 (n=150) replicated study 2 but added information about the providers' abilities.

Results: This research found that patients tend to choose attractive doctors of the opposite gender but are less likely to choose attractive doctors of the same gender. In addition, our studies revealed that such an effect is more prominent when the disease severity is high. Furthermore, the influence of gender stereotypes is mitigated in both the high and low disease severity conditions when service providers' qualification information is present.

Conclusions: This research contributes to the literature on medical information systems research and sheds light on what information should be displayed on digital doctor consultation platforms. To counteract stereotype-based attractiveness biases, health care platforms should consider providing comprehensive qualification information alongside profile pictures.

(J Med Internet Res 2024;26:e46551) doi:10.2196/46551

KEYWORDS

digital doctor consultations; health care providers; gender stereotype; physical attractiveness; qualification information; experimental; telemedicine; digital consultation; disease severity; sex; gender; gender stereotypes; digital health

Introduction

Digital health services have become increasingly important in health care, particularly since the COVID-19 pandemic necessitated the development of new digital consultation systems [1-4]. Previous research indicates that patients often use digital reviews to select health care providers [5]. However, there is limited understanding of how decisions are made when such reviews are not available. In traditional health care settings (ie, offline health care), patients typically choose or are assigned a

XSL•FO
RenderX

general practitioner based on their geographical location. In contrast, digital health care settings remove these geographical restrictions, presenting patients with a vast array of choices and different information. This abundance of choices and inconsistent information can make it challenging for patients to compare and select health care providers [6]. For instance, the availability of digital information can lead patients to select doctors based on inherent biases, which can arise from their personal preferences or the way the information is presented on the platform. Given these limitations, patients may resort to evaluating health care providers based on available photographs. This situation raises questions about the influence of a health care provider's appearance on patient choices and whether gender differences affect these decisions. Previous research indicates that attractive service providers can positively influence consumer attitudes and purchase intentions [7]. High physical attractiveness is often associated with greater credibility and professional ability, affecting consumer attitudes accordingly [8-10]. However, high attractiveness can also have negative effects, such as difficulty in forming same-gender friendships due to competition or eliciting envy and perceived threats among members of the same gender [11-14].

Much of the existing research focuses on general services such as fitness clubs and educational consulting [15]. These findings may not be directly transferable to specialized digital services, such as digital doctor consultations. Such services are categorized as credence services, which are difficult for patients to evaluate due to their complex and specialized nature [16,17]. The challenge of evaluation is further exacerbated in a digital setting where information about the provider's abilities may be limited.

There is a research gap in understanding how physical appearance affects patient preferences in digital health care settings. This study aims to fill this gap by examining preferences for doctors of varying levels of physical attractiveness in digital consultations. We also consider the moderating effects of disease severity and the availability of information on health care providers' qualifications to offer both theoretical and practical insights.

Research on doctor selection uses signal theory to explore how patients navigate information asymmetry [18,19]. Text-mining research identified key service features, such as overall service experience and personality traits, that affects patients' trust and, consequently, consultation volumes [20]. Digital doctor consultation platforms typically display limited information about doctors, such as their names, profile pictures, and titles. The appearance of the doctor may influence patients' preferences. Prior studies have shown that the attire of physicians influenced patients' perceptions [21-23]. For example, physicians in white coats were viewed as more experienced and professional than those in casual jackets [24]. Despite this, there is still scant research on how such visual and textual information affect patients' decision-making.

In non–health care service settings, the impact of physical attractiveness on performance evaluation has shown mixed results. While some studies indicate that higher levels of attractiveness positively influence performance evaluations [25,26], others suggest a negative effect [27]. These mixed outcomes may be attributed to various contextual factors such as gender [28] and service quality conditions [29]. In digital settings, service providers' physical appearance has been shown to influence customer choice, preference, and purchase intent [30]. For example, in the sharing economy, such as Airbnb, facial characteristics contribute to reputation mechanisms [31]. Studies have found that the perceived trustworthiness and attractiveness of a host's profile photograph significantly affect Airbnb prices [32,33]. Physical attractiveness also influences digital consumer shopping, with more attractive avatars correlating with higher sales when product involvement is moderate [34].

Previous research indicates that gender differences significantly influence how patients perceive doctors [24,35,36]. The sexual attribution bias (SAB) offers an explanatory framework for these gender effects [37]. SAB leads individuals to attribute the success of same-gender individuals with high attractiveness to luck rather than ability, whereas for opposite-gender individuals, high attractiveness is attributed to ability [38]. This bias manifests in 2 ways: demeaning attractive individuals of the same gender and praising attractive individuals of the opposite gender. Studies have shown that negative vigilance against attractive same-gender individuals is strong and automatic due to intragender competition [39-42].

Unlike general services where attractiveness universally enhances provider popularity [43,44], we posit that in digital doctor platforms, gender plays a significant role in shaping preferences for providers with varying levels of attractiveness. Influenced by same-gender competition, consumers may perceive same-gender providers with high attractiveness as less qualified. Conversely, influenced by mating motivation, consumers may prefer highly attractive providers of the opposite gender. Given the importance of competence in selecting credence service providers [45], SAB suggests that individuals may make derogatory attributions about the competence of same-gender providers with high physical attractiveness.

- *Hypothesis 1*: In digital doctor consultations, people are more likely to perceive a more attractive doctor of the opposite (vs same) gender as more (less) competent, thereby influencing their likelihood of selecting that doctor.

We anticipate that disease severity will modulate the effects of gender and attractiveness on provider selection. Previous research has established a relationship between disease severity and behavior in various health care contexts [46-49]. The Elaboration Likelihood Model posits 2 routes of information processing: central and peripheral, determined by the individual's level of involvement [50]. In low-involvement situations, attitudes are influenced by simple cues, whereas high involvement leads to a deeper consideration of complex information. Applying this model and stereotype theory, we suggest that in high-involvement scenarios, consumers will scrutinize providers' abilities more closely, potentially leading to greater influence of SAB on their choices. For example, women may question the competence of attractive women providers, suspecting that their success is due to their appearance rather than merit [51]. This aligns with research showing that

XSL•FO
RenderX

physical attractiveness elicits more jealousy among women than men [14]. Conversely, in low-involvement scenarios, the impact of gender on preferences for providers' physical attractiveness is expected to be less pronounced.

- *Hypothesis 2*: The effect of gender on individual preferences for the doctor's physical attractiveness is moderated by disease severity. Such an effect is more prominent when disease severity is high (vs low).

Furthermore, a foundational assumption for the effects discussed earlier is that consumers lack additional information about the service providers' abilities. This is often the case in credence services, where consumers typically have less expertise and access to information compared with noncredence services, making them more reliant on extra information supplied by the provider [52]. Research has shown that the provision of such information can mitigate consumer-perceived risk or uncertainty [53].

- *Hypothesis 3*: The effect of gender on individual preferences for the service provider's physical attractiveness disappears when information about the doctor's abilities is present.

Methods

Ethical Considerations

This study has received ethical approval (20180420) from the College of Management, Shenzhen University. Informed consent was obtained from all participants. The consent form provided detailed information about the research's purpose, the involved institutions, the nature of their participation, and the use of their data. Participants were informed of their right to withdraw at any time and the procedures for data removal. The consent process was designed to comply with data protection legislation, ensuring that participants were aware of their rights and the protections in place. Privacy and confidentiality were protected by anonymizing study data. The data were stored securely and processed. For study 1, we paid CN ¥3 (a currency exchange rate of CN ¥7.21=US $1 is applicable.) for each response we collected. For study 2 and study 3, participants received a gift of approximately CN ¥5 for their participation.

Study 1

Overview

Study 1 used a 2 (attractiveness: high vs moderate)×2 (doctor gender: man vs woman) between-subjects factorial experiment. We created a digital doctor consultation scenario where participants were asked to imagine that they had abdominal pain while traveling. As most clinics are closed at night, they decided to consult a digital doctor via a platform called "*Doctor Online*."

Stimuli

An interface was designed for the platform. In the interface, the doctor's picture was shown on the left, and their name, title, and department were displayed on the right. There was also a button for digital consultation. To manipulate the physical attractiveness of the doctor, we used an artificial intelligence face generator to generate different faces for male and female doctors (Figure 1). We slightly adapted the physical features of the faces to make them less or more attractive.

Figure 1. Stimulus used in study 1. These profile pictures were generated by artificial intelligence.

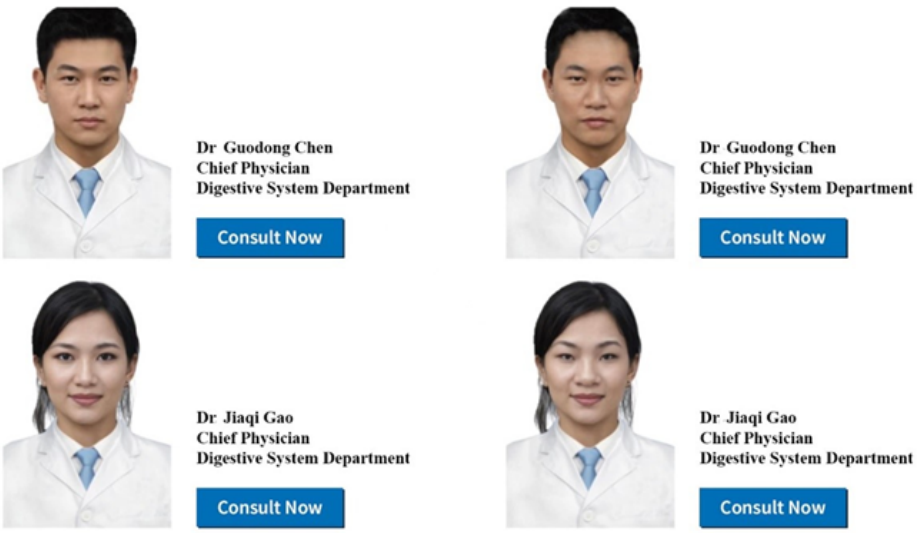

Pretest

We conducted a pretest to check whether the manipulation was successful. We recruited 111 participants ($mean_{age}$ 30.0, SD 7.86 years; women: n=70, 63.1%) to test the male doctor and 105 participants ($mean_{age}$ 30.5, SD 9.61 years; women: n=74, 70.5%) to test the female doctor via a digital panel service WJX. The results showed that for the male doctor, participants perceived the face in the high attractiveness condition to be more attractive than that in the moderate attractiveness condition ($mean_{high}$ 5.53, SD 0.98 vs $mean_{moderate}$ 4.32, SD 1.09; $F_{1,109}$=38.3; P<.001; ^{2}=0.26). For the female doctor, the face in the high attractiveness condition was perceived as more attractive than that in the moderate attractiveness condition ($mean_{high}$ 5.98, SD 0.85 vs $mean_{moderate}$ 4.94, SD 0.95; $F_{1,103}$=34.7; P<.001; ^{2}=0.25).

XSL•FO
RenderX

Subjects

We recruited 282 respondents via a digital panel service WJX in China. The participants were not necessarily actual patients in this experiment. The mean age of the respondents was 27.8 (SD 9.76) years. In all, 156 (55.3%) of the respondents were women. Based on the respondents' gender, we created a new variable "gender match." If the gender of the participant is the same as that of the doctor, we assigned the value 1 (same gender). Otherwise, we coded the variable as 0 (opposite gender). In the end, there were 147 participants in the same gender condition, and 135 participants in the opposite gender condition. The majority of respondents had a bachelor's degree (n=104, 36.9%), followed closely by those with a graduate degree or higher (n=89, 31.6%). Less than a quarter (n=61, 21.6%) of the respondents had an education level of high school or below. The majority of respondents were students (n=132, 46.8%), followed by those working in industrial or manufacturing enterprises (n=42, 14.9%). Less than 5% (n=17) were either retired or unemployed. The data revealed that a slight majority of respondents (n=153, 54.3%) had experience with digital consultations, while 45.7% (n=129) had not engaged in such services. Tables 1 and 2 show detailed information on educational level and occupational types.

Table 1. Distribution of respondents by educational level (study 1; n=282).

Educational level	Frequency, n (%)
Junior high or below	17 (6)
High school or vocational	44 (15.6)
Associate's degree	28 (9.9)
Bachelor's degree	104 (36.9)
Graduate or above	89 (31.6)

Table 2. Occupational types of respondents (study 1; n=282).

Occupation type	Frequency, n (%)
Commercial or service industry	35 (12.4)
Industrial or manufacturing	42 (14.9)
Self-employed	31 (11)
Retired	8 (2.8)
Unemployed	9 (3.2)
Students	132 (46.8)
Other	25 (8.9)

Manipulation Checks

For the manipulation check, we asked the participants to rate the physical attractiveness of the doctor using a single item: *What do you think of the doctor's physical appearance?* (1—not attractive at all; 7—very attractive). A single item was used to measure the doctor's perceived competence: *Do you think that this doctor is competent in this job*? (1—no; 7—yes). To assess the participants' intention to select the doctor, we asked the participants how likely he or she would be to select this doctor for a digital consultation (1—very unlikely; 7—very likely). To rule out alternative explanations, we also measured a number of the doctor's attributes: perceived seniority (*What do you think about the doctor's work experience?* 1—very limited; 7—very rich), anticipated embarrassment (*How likely are you to feel embarrassed about consulting this doctor?* 1—not embarrassed at all; 7—very embarrassed), perceived friendliness (*How friendly is this doctor?* 1—not friendly at all; 7—very friendly), and perceived willingness to help (*Do you think this doctor is willing to meet patients' needs?* 1—not at all; 7—very much). Finally, respondents were asked to provide some relevant demographic information about their age, gender, educational background, and occupation.

Study 2

Overview

In study 2, to reduce the complexity of the experimental design, we focused on women doctors. A between-subjects experiment was designed with 2 independent variables, each with 2 levels. The manipulated factors were disease severity (2 levels: high vs low) and gender of the participant (2 levels: man vs woman).

Stimuli

In the experiment, respondents were randomly assigned to 1 of 2 groups of scenarios with different degrees of involvement. They were first asked to read material about a scenario of abdominal pain during travel. The low-involvement group was described as having "slight abdominal pain, the same old symptom, not too worried, decided to consult a doctor." The high-involvement group heard it described as "severe abdominal pain, lumps when pressing with hands, never encountered such symptoms before, and felt both worried and afraid, decided to consult a doctor immediately."

To manipulate physical attractiveness in the dependent variable, we referred to the studies by Heilman and Stopeck [54] and Försterling et al [37], as well as previous studies on the use of

XSL•FO
RenderX

photographs as stimuli of attractiveness in job-seeking scenarios [55]. In total, 30 respondents were invited to participate in rating different women service providers' physical attractiveness. Respondents were asked to use a 5-point Likert scale to rate the attractiveness of providers in ten 2.54-cm photographs. All photographs were selected from several medical websites and their photograph styles were unified. Finally, 3 photographs, representing the mean value, higher than 1 SD and lower than 1 SD, were selected as photographs to represent the 3 different levels of physical attractiveness, as shown in Figure 2.

Figure 2. Stimulus used in study 2. The pictures were blurred because of privacy issues. PA: physical attractiveness.

Pretest

We also conducted a second pretest with 28 respondents to check whether confounding factors such as perceived friendliness, perceived patience, perceived seniority, perceived willingness to help, and anticipated embarrassment were controlled among 3 service providers. The results suggested that all 3 service providers were perceived as equally friendly ($mean_{high\ PA}$ 2.94, SD 1.692; $mean_{middle\ PA}$ 2.94, SD 2.016; $mean_{low\ PA}$ 3.75, SD 2.053; $F_{2,25}$=0.586; P=.56; ^{2}=0.031), patient ($mean_{high\ PA}$ 3.19, SD 1.682; $mean_{middle\ PA}$ 2.88, SD 1.857; $mean_{low\ PA}$ 3.88, SD 2.475; $F_{2,25}$=0.720; P=.50; ^{2}=0.037), experienced ($mean_{high\ PA}$ 2.44, SD 1.711; $mean_{middle\ PA}$ 2.81, SD 1.905; $mean_{low\ PA}$ 3.50, SD 2.390; $F_{2,25}$=0.806; P=.46; ^{2}=0.042), willing to help ($mean_{high\ PA}$ 3.44, SD 1.931; $mean_{middle\ PA}$ 2.56, SD 1.750; $mean_{low\ PA}$ 3.50, SD 2.268; $F_{2,25}$=1.037; P=.37; ^{2}=0.053), and embarrassing to consult with ($mean_{high\ PA}$ 3.38, SD 1.996; $mean_{middle\ PA}$ 3.44, SD 2.007; $mean_{low\ PA}$ 3.25, SD 1.488; $F_{2,25}$=0.025; P=.98; ^{2}=0.001).

Subjects

To observe how gender and degree of involvement influence individual preferences, this experiment recruited 158 citizens in the city center of Shenzhen, China. The participants were not necessarily actual patients during the experiment. Of the 158 respondents, 80 were men and 78 were women. We recruited only those participants who were older than 18 years. The age distribution was weighted toward adults, with 55.1% (n=87) aged 18-25 years, 22.8% (n=36) aged 26-30 years, and 16.5% (n=26) aged 31-40 years for both men and women. The respondents were assigned to either high-involvement (n=75, 48%) or low-involvement (n=83, 53%) groups. The data showed that the majority of respondents had a bachelor's degree (n=91, 57.6%), and less than 5% (n=7) had an educational level of junior high school or below. The data also indicated that the largest group of respondents were students (n=58, 36.7%), followed by those working in the commercial or service industry (n=31, 19.6%). Less than 1% (n=1) were retired. Tables 3 and 4 show detailed information on educational level and occupational types.

Respondents were asked to examine the names and photographs of 3 doctors and asked to select 1 from whom they would like to receive treatment. Then, they completed a questionnaire. A 7-point Likert scale was used in this experiment to measure the perceived severity of the disease. We used a 7-point Likert scale, adapted from previous studies on positive emotions caused by physical attractiveness [56,57] and research about the evaluation of attraction and ability [58], to evaluate the physical attractiveness of the physicians. Finally, respondents were asked to provide some relevant demographic information about their age, gender, educational background, and occupation.

Table 3. Distribution of respondents by educational level (study 2; n=158).

Educational level	Frequency, n (%)
Junior high or below	7 (4.4)
High school or vocational	22 (13.9)
Associate's degree	29 (18.4)
Bachelor's degree	91 (57.6)
Graduate or above	9 (5.7)

XSL•FO
RenderX

Table 4. Occupational types of respondents (study 2; n=158).

Occupation type	Frequency, n (%)
Commercial or service industry	31 (19.6)
Industrial or manufacturing	26 (16.5)
Self-employed	10 (6.3)
Retired	1 (0.6)
Students	58 (36.7)
Other	32 (20.3)

Study 3

Overview

All basic conditions and requirements were the same as in study 2. Study 3 examined whether the participants would make a different choice after being provided with extra information. We also sought to further verify that the presence of information about providers' abilities would reduce the influence of gender on respondents' preferences.

Stimuli

The only difference from the experimental procedure used in study 2 was the adjustment of stimuli. Study 3 added similar ability information about the provider's professional background and clinical experience below the photograph provided in study 2. The description read: "...graduated from...medical college, ...has participated in many research projects, ...has been working for 5 years, ...is an expert in diagnosis and treatment of common and frequently-occurring diseases..." To minimize confounding factors, the written description of the academic background, clinical experience, scientific research achievements, and areas of expertise of the service providers in the stimuli were very similar.

Subjects

Study 3 recruited 150 citizens in the city center of Shenzhen, with equal numbers of respondents by gender (n=75, 50% men and n=75, 50% women). The participants were not necessarily actual patients during the experiment. The age distribution was similar to that of study 2, with a high concentration of young adults: 54% (n=81) aged 18-25 years, 26.7% (n=40) aged 26-30 years, and only 12% (n=18) aged 31-40 years. The high- and low-involvement groups included 74 (49.3%) and 76 (50.7%) respondents, respectively. The data indicated that among the 150 respondents, the majority had a bachelor's degree (n=82, 54.7%). This was followed by those who had completed high school, vocational school, or technical school (n=26, 17.3%), and then by those with an associate degree (n=25, 16.7%). The data also indicated that among the 150 respondents, the largest group was students, comprising 44% (n=66) of the sample. This was followed by those working in the commercial or service industry, who made up 26.7% (n=40) of the respondents. Those in industrial or manufacturing roles accounted for 7.3% (n=11), and self-employed individuals made up 4.7% (n=7). Tables 5 and 6 show detailed information on educational level and occupational types.

Table 5. Distribution of respondents by educational level (study 3; n=150).

Educational level	Frequency, n (%)
Junior high or below	11 (7.3)
High school or vocational	26 (17.3)
Associate's degree	25 (16.7)
Bachelor's degree	82 (54.7)
Graduate or above	6 (4)

Table 6. Occupational types of respondents (study 3; n=150).

Occupation type	Frequency, n (%)
Commercial or service industry	40 (26.7)
Industrial or manufacturing	11 (7.3)
Self-employed	7 (4.7)
Retired	4 (2.7)
Unemployed	1 (0.7)
Students	66 (44)
Other	21 (14)

XSL•FO
RenderX

Results

Study 1

The individual item scores within a single scale were computed to yield an average score, serving as a composite measure for that particular scale. The results revealed that participants perceived doctors in the high attractiveness condition to be more attractive than those in the moderate attractiveness condition ($mean_{high}$ 5.62, SD 1.03 vs $mean_{moderate}$ 4.72, SD 1.36; $F_{1,280}$=39.1; P<.001; ^{2}=0.12). This suggested that our manipulation was successful.

We conducted an ANOVA to evaluate our hypotheses. The results showed that the physical attractiveness did not have a significant main effect on perceived competence ($mean_{high}$ 5.28, SD 0.98 vs $mean_{moderate}$ 5.10, SD 1.12; $F_{1,278}$=2.32; P=.13; ^{2}=0.008). In consistence with our expectation, we observed a significant interaction effect of physical attractiveness and gender match on perceived competence ($F_{1,278}$=5.95; P=.015; ^{2}=0.021). In particular, when the doctor and the participant were of different genders, attractive doctors were perceived to be more competent in the job than less attractive doctors ($mean_{high}$ 5.45, SD 0.93 vs $mean_{moderate}$ 4.95, SD 1.10; $F_{1,133}$=8.08; P=.005; ^{2}=0.057). However, when the doctor and the participant were of the same gender, attractiveness did not produce any positive effect on perceived competence ($mean_{high}$ 5.09, SD 1.01 vs $mean_{moderate}$ 5.21, SD 1.13; $F_{1,145}$=0.41; P=.52; ^{2}=0.003; see Tables 7 and 8 for detailed information).

Physical attractiveness had a positive effect on intention to select the doctor ($mean_{high}$ 5.49, SD 1.13 vs $mean_{moderate}$ 5.10, SD 1.32; $F_{1,278}$=5.10; P=.025; ^{2}=0.018). There was also a significant interaction effect of physical attractiveness and gender match on intention to select the doctor ($F_{1,278}$=3.96; P=.048; ^{2}=0.014). More specifically, participants were more likely to select attractive doctors than less attractive doctors of the opposite gender ($mean_{high}$ 5.49, SD 1.17 vs $mean_{moderate}$ 4.86, SD 1.36; $F_{1,133}$=8.15; P=.005; ^{2}=0.058). In contrast, when the participants and the doctor were of the same gender, there was no difference in intention to select between more attractive and less attractive doctors ($mean_{high}$ 5.31, SD 1.07 vs $mean_{moderate}$ 5.27, SD 1.27; $F_{1,145}$=0.40; P=.84; ^{2}<0.001).

Next, we performed a moderated mediation analysis using Hayes PROCESS (model 7) with 10,000 bootstrap samples. To rule out alternative explanations, we also examined the mediating effect of perceived seniority, perceived embarrassment, perceived friendliness, and perceived willingness to help. The results showed that the moderation effect of gender match on physical attractiveness on intention to select was mediated only by perceived competence (point estimation=0.34, 95% CI 0.0622-0.6593) but not by perceived seniority (point estimation=0.09, 95% CI –0.0601 to 0.2763), anticipated embarrassment (point estimation=0.01, 95% CI –0.0196 to 0.0663), perceived friendliness (point estimation=–0.01, 95% CI –0.0912 to 0.0477), and perceived willingness to help (point estimation=–0.001, 95% CI –0.1395 to 0.1423). In particular, when participants and doctors were of different genders, physical attractiveness led to higher perceived competence of the doctor, thereby increasing the intention to select the doctor (point estimation=0.27, 95% CI 0.0776-0.4994). However, when there was no difference in gender, such an indirect effect was insignificant (point estimation=–0.06, 95% CI –0.2716 to 0.1233).

We conducted a supplementary analysis in which we controlled for age. The results of this age-adjusted analysis remained consistent with our initial findings. Specifically, after accounting for age differences, the interaction effect of gender match and physical attractiveness on perceived competence ($F_{1,277}$=4.81; P=.03; ^{2}=0.017) and user intention ($F_{1,277}$=2.32; P=.006; ^{2}=0.027) remained stable. Also, the moderated mediation effect was significant (point estimation=0.55, 95% CI 0.1598-0.9762). Therefore, hypothesis 1 was supported.

Table 7. Descriptive statistics for perceived competence scores based on physical attractiveness and gender match.

Physical attractiveness and gender match	Score, mean (SD)	Participants, n
Moderate		
Concordance	5.21 (1.13)	82
Discordance	4.95 (1.10)	59
Total	5.10 (1.12)	141
High		
Concordance	5.09 (1.01)	65
Discordance	5.45 (0.93)	76
Total	5.28 (0.98)	141
Total		
Concordance	5.16 (1.08)	147
Discordance	5.23 (1.04)	135
Total	5.19 (1.06)	282

XSL•FO
RenderX

Table 8. Descriptive statistics for intention scores based on physical attractiveness and gender match.

Physical attractiveness and gender match	Score, mean (SD)	Participants, n
Moderate		
Concordance	5.27 (1.26)	82
Discordance	4.86 (1.35)	59
Total	5.10 (1.31)	141
High		
Concordance	5.31 (1.07)	65
Discordance	5.49 (1.17)	76
Total	5.40 (1.12)	141
Total		
Concordance	5.29 (1.18)	147
Discordance	5.21 (1.28)	135
Total	5.25 (1.23)	282

Study 2

A manipulation check on the perceived severity of the disease showed that the scores of the low-involvement group (mean 3, SD 1.47) were significantly lower than those of the high-involvement group (mean 4.45, SD 1.60; t_{158}=–5.95; *P*<.001; ^{2}=0.185). This indicates that the different degree (high vs low) of involvement was successfully manipulated. A 1-way ANOVA was used for a manipulation check on physical attractiveness. The results showed that physical attractiveness of doctors was successfully manipulated ($mean_{high\ PA}$ 5.61, SD 1.13; $mean_{middle\ PA}$ 4.77, SD 0.94; $mean_{low\ PA}$ 3.88, SD 1.28; $F_{1,58}$=98.24; *P*<.001; ^{2}=0.499).

A chi-square test was used to examine hypothesis 1. The result showed that men and women had a significant difference in selecting the doctors with 3 different physical attractiveness levels ($^2_2$=14.165; *P*<.001; =0.299). Furthermore, this research tried to identify the level of physical attractiveness upon which this kind of difference exists. Therefore, a post hoc test was conducted. The analysis finds that adjusted standardized residual of gender preference has a significant difference on high-level ($N_{men\ choose\ high\ physical\ attractiveness}$=48, $N_{women\ choose\ high\ physical\ attractiveness}$=29; *z* $score_{abs}$=2.9, *z* $score_{abs}$>1.96; *P*=.004) and middle-level ($N_{men\ choose\ middle\ physical\ attractiveness}$=25, $N_{women\ choose\ middle\ physical\ attractiveness}$=47; *z* $score_{abs}$=3.7, *z* $score_{abs}$>1.96; *P*<.001) service providers, while gender has an insignificant effect on those with low physical attractiveness ($N_{men\ choose\ middle\ physical\ attractiveness}$=7, $N_{women\ choose\ middle\ physical\ attractiveness}$=2; *z* $score_{abs}$=1.7, *z* $score_{abs}$<1.96; *P*=.09). This indicated that in terms of different physical attractiveness in service providers, men (vs women) preferred highly attractive women service providers, whereas women (vs men) preferred moderately attractive providers; however, there was no difference between men and women's choice for the low-attractiveness provider. Thus, hypothesis 1 is supported.

To further investigate how involvement in a credence service scenario influences men's (vs women's) preferences for different physical attractiveness in service providers, a post hoc chi-square test and an analysis of logistic regression were used. The test found that involvement moderated the effect of gender on preferences ($\beta_{gender\times involvement}$=–18.39; Wald=662.36; *P*=.003). The results illustrated that in the low-involvement condition, the effect of gender on preference was reduced. In particular, the effect of gender on individual preferences existed only in the high-involvement scenario ($^2_2$=9.78; *P*=.008). For the low-involvement scenario, there was no gender effect ($^2_2$=4.82; *P*=.09). Therefore, hypothesis 3 was supported. In the high-involvement condition, compared with women, men preferred the highly attractive service provider ($N_{men\ choose\ high\ physical\ attractiveness}$=19, $N_{women\ choose\ high\ physical\ attractiveness}$=12; *z* $score_{abs}$=2.1, *z* $score_{abs}$>1.96; *P*=.04); compared with men, women were more likely to select a service provider with moderate physical attractiveness ($N_{men\ choose\ middle\ physical\ attractiveness}$=13, $N_{women\ choose\ middle\ physical\ attractiveness}$=28; *z* $score_{abs}$=2.9; *z* $score_{abs}$>1.96; *P*=.004; Table 9).

Table 9. Influence of gender on consumers' preferences for service providers with different levels of physical attractiveness.

Involvement	Gender	High physical attractiveness	Middle physical attractiveness	Low physical attractiveness
High involvement	Men	19	13	3
High involvement	Women	12	28	0
Low involvement	Men	29	12	4
Low involvement	Women	17	19	2

XSL•FO
RenderX

Study 3

A manipulation check on the perceived severity of the disease showed that the scores of the low-severity group (mean 5.17, SD 1.14) were significantly lower than those of the high-severity group (mean 3.24, SD 1.13; t_{148}=–10.40; P<.001; ^{2}=0.422). This indicated that the degree (high vs low) of severity was successfully manipulated. A 1-way ANOVA was used for a manipulation check on physical attractiveness. The data showed that there was a significant difference between the 3 levels of physical attractiveness ($mean_{high\ PA}$ 5.81, SD 0.90; $mean_{middle\ PA}$=4.73, SD=0.94; $mean_{low\ PA}$ 3.82, SD 1.18; $F_{1,50}$=144.50; P<.001; ^{2}=0.661). Thus, the physical attractiveness of the 3 doctors was successfully manipulated.

We used a post hoc chi-square test and a logistic regression model to test our hypothesis. The chi-square test results showed that there was no significant difference in the physical attractiveness preferences of men and women (χ^2_2=1.147; P=.56; P>.05; =0.087) when they were provided with extra information about the service provider's qualification. A further post hoc chi-square test showed that in both the high-involvement (χ^2_2=1.730; P=.421; P>.05; =0.151) and low-involvement (χ^2_2=0.046; P=.98; P>.05; =0.025) scenarios, the effect of gender on preferences disappeared after respondents were given information on the service provider's abilities; Table 10).

Table 10. Influence of gender on consumers' preferences for service providers (including ability information).

Involvement	Gender	High physical attractiveness	Middle physical attractiveness	Low physical attractiveness
High involvement	Men	10	22	7
High involvement	Women	12	22	3
Low involvement	Men	19	13	4
Low involvement	Women	21	13	4

Discussion

Study 1 provided evidence to support hypothesis 1. In the context of digital doctor consultation, gender influences people's preferences for attractive health care providers. People are more likely to select a more (vs less) attractive doctor of the opposite (vs same) gender. This can be explained by the perceived competence of the doctor. In study 1, we adopted a separate evaluation method. On one hand, this method helps to rule out confounding factors and provides strong support for the hypothesis. However, in reality, when patients search for a digital doctor consultation, they may see a list of all available doctors and select the one they prefer. Therefore, in study 2, we adopted a joint evaluation method. Furthermore, we also manipulated the level of disease severity. The results of the second study show that when a service provider's profile picture was provided to consumers, the consumers' preferences were greatly influenced by the provider's gender. The impact of this gender stereotype existed only in the high-involvement condition. If involvement was low, consumers did not have higher requirements for the abilities of medical service providers and thus relied less on their physical attractiveness to infer the ability of the service provider. The results of study 3 illustrated that when extra qualification information was provided, the influence of gender on individual preferences for the service provider's attractiveness disappeared in both the high- and low-involvement conditions. It is important to note that the context of the research was based in China, a leading country in providing digital consultation services. According to data from the China Internet Network Information Center, by June 2023, the number of users accessing digital medical services in China had surged to 364 million, a rise of 1.62 million since December 2022. The data further reveal that the number of digital hospitals in China has surpassed 3000. These hospitals have offered digital diagnostic and treatment services to more than 25.9 million patients [59]. Thus, the participants in this study were familiar with digital doctor consultations. With the COVID-19 pandemic, an increasing number of European countries have started developing digital consultation services in health care. For instance, Estonia has already begun offering digital services to patients in remote areas. Therefore, the findings from this study can offer significant insights to other countries.

This research contributes to the literature on medical information systems research. First, this study further enhances our understanding of patients' behavior in digital doctor consultations. As a result of the information asymmetry, patients are always looking for other signals to aid in their decision-making [18]. The lack of trust is one of the biggest barriers to digital services [60]. Besides professional status and service feedback, which are known as important signals for patients' doctor selection [47], this study has shown that physical attractiveness is also a crucial signal for decision-making in the digital doctor consultation context. This study also helps to deepen the understanding of SAB in the context of digital health care services. Previous studies on SAB focus on contexts such as job interviews [61]. More recently, the literature has increasingly focused on general services, such as education services [15]. In addition, this study sheds light on the boundary condition of SAB. Our findings indicate that the degree of disease severity influences the effect of gender on preferences for attractive service providers. Based on the elaboration likelihood model, consumers with low disease involvement tend to rely on peripheral cues to make choices, which means that they may automatically choose a more attractive service provider. Nevertheless, our research points out that this assumption only holds for both men and women when qualification information is present. When qualification information is absent and information asymmetry exists, men

XSL•FO
RenderX

and women react differently to attractive service providers when they are highly involved.

These research findings have some crucial implications for digital health care service management. Providing qualification information of service providers on digital health consultation platforms can mitigate the effect of gender on the individual preference for the attractiveness of a service provider, regardless of the scenario's level of involvement. In the health care service sector, providing sufficient information to eliminate bias is a useful strategy. In the context of information asymmetry, patients may be influenced by the doctor's physical attractiveness and make irrational decisions. When marketing telehealth services, hospitals should focus on highlighting doctor expertise, patient satisfaction, and health outcomes instead of relying on doctor attractiveness or demographic characteristics to promote their services. Furthermore, hospitals should develop patient education materials that explain the importance of choosing a doctor based on their qualifications and expertise, rather than their appearance. Also, digital health care platforms could consider implementing features that allow patients to filter and search for doctors based on their qualifications, expertise, and patient satisfaction ratings. This can help patients to make more rational decisions.

This study has some clear limitations, which point, in turn, to avenues for future research. One notable limitation of this study is the restricted age range of both the health care providers depicted in the images and the study participants, who were predominantly younger than 25 years. This lack of age diversity could potentially limit the generalizability of the findings, as perceptions of physical attractiveness and professionalism may vary across different age groups. Although supplementary analyses controlling for age did not significantly alter the results, it remains an open question whether these findings would hold in a more age-diverse sample. Future research should aim to include a broader age range of both health care providers and participants to explore the potential moderating effect of age on the relationships examined in this study.

Another limitation of this study pertains to the clinical scenario presented, which focuses on an acute condition that could potentially be resolved in a single consultation. This may not fully capture the complexities involved when patients are seeking long-term care for chronic conditions. In such cases, the level of disease involvement is likely to be higher, and patients may seek more comprehensive information to assess a doctor's competence. The dynamics of how gender match and physical attractiveness influence patient choices could thus differ in a chronic care setting. Future research could benefit from exploring these nuances to provide a more holistic understanding.

Another limitation of this study is its focus on China. To enhance the generalizability of our research and generate a more worldwide impact, replication in various countries and across different cultures is necessary. Future research should, for instance, involve recruiting subjects from diverse nations. This is because health care systems vary significantly among countries. Systematic differences in health care structures, policies, and practices can substantially influence how patients evaluate doctors digitally. Consequently, these contextual differences should be carefully controlled for and considered in future studies to ensure the validity and applicability of research findings across different health care systems and cultural backgrounds. Furthermore, the cultural orientation of the patient can also play a role in this situation. Future studies should investigate how varying cultural dimensions, such as collectivism, power distance, and uncertainty avoidance, influence patient preferences and decision-making processes in digital health care settings. For instance, individuals from countries with high uncertainty avoidance may require more concrete information in the doctor's description to feel comfortable with their choice.

Another important research direction involves examining the impact of other forms of information asymmetry in digital health care settings. For instance, the availability and quality of digital reviews can play a significant role in shaping patient choices. By analyzing the content and credibility of web-based reviews, researchers can better understand how they contribute to patients' decision-making processes and their perceptions of service providers. In addition, assessing the role of digital reputation management and its influence on patient choices can offer valuable insights into digital health care.

Conflicts of Interest

None declared.

References

1. Lee SM, Lee D. Opportunities and challenges for contactless healthcare services in the post-COVID-19 era. Technol Forecast Soc Change 2021;167:120712 [FREE Full text] [doi: 10.1016/j.techfore.2021.120712] [Medline: 33654330]
2. Zhou F, Wang Z, Mai X, Liu X, Reid C, Sandover S, et al. Online clinical consultation as a utility tool for managing medical crisis during a pandemic: retrospective analysis on the characteristics of online clinical consultations during the COVID-19 pandemic. J Prim Care Community Health 2020;11:2150132720975517 [FREE Full text] [doi: 10.1177/2150132720975517] [Medline: 33213264]
3. Flodgren G, Rachas A, Farmer AJ, Inzitari M, Shepperd S. Interactive telemedicine: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2015;2015(9):CD002098 [FREE Full text] [doi: 10.1002/14651858.CD002098.pub2] [Medline: 26343551]
4. Yan Z, Bernardi R, Huang N, Chang Y. Guest editorial: the bright side and the dark side of digital health. Internet Res 2021;31(6):1993-1999 [FREE Full text] [doi: 10.1108/intr-12-2021-682]

XSL•FO
RenderX

5. Zhang S, Wang JN, Chiu YL, Hsu YT. Exploring types of information sources used when choosing doctors: observational study in an online health care community. J Med Internet Res 2020;22(9):e20910 [FREE Full text] [doi: 10.2196/20910] [Medline: 32936080]
6. Chen S, Guo X, Wu T, Ju X. Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. Inf Process Manag 2020;57(5):102253 [FREE Full text] [doi: 10.1016/j.ipm.2020.102253]
7. Caballero MJ, Pride WM. Selected effects of salesperson sex and attractiveness in direct mail advertisements. J Mark 2018;48(1):94-100. [doi: 10.1177/002224298404800110]
8. Bekk M, Spörrle M, Völckner F, Spieß E, Woschée R. What is not beautiful should match: how attractiveness similarity affects consumer responses to advertising. Mark Lett 2017;28(4):509-522. [doi: 10.1007/s11002-017-9428-3]
9. Chaker NN, Walker D, Nowlin EL, Anaza NA. When and how does sales manager physical attractiveness impact credibility: a test of two competing hypotheses. J Bus Res 2019;105:98-108. [doi: 10.1016/j.jbusres.2019.08.004]
10. Rule NO, Ambady N. She's got the look: inferences from female chief executive officers' faces predict their success. Sex Roles 2009;61(9-10):644-652. [doi: 10.1007/s11199-009-9658-9]
11. Fink B, Klappauf D, Brewer G, Shackelford TK. Female physical characteristics and intra-sexual competition in women. Pers Individ Differ 2014;58:138-141. [doi: 10.1016/j.paid.2013.10.015]
12. Kenrick DT, Montello DR, Gutierres SE, Trost MR. Effects of physical attractiveness on affect and perceptual judgments: when social comparison overrides social reinforcement. Pers Soc Psychol Bull 2016;19(2):195-199. [doi: 10.1177/0146167293192008]
13. Bleske-Rechek A, Lighthall M. Attractiveness and rivalry in women's friendships with women. Hum Nat 2010;21(1):82-97. [doi: 10.1007/s12110-010-9081-5]
14. Buunk AP, Aan 't Goor J, Solano AC. Intrasexual competition at work: sex differences in the jealousy-evoking effect of rival characteristics in work settings. J Soc Pers Relatsh 2010;27(5):671-684. [doi: 10.1177/0265407510368964]
15. Li Y, Xie L, Gao T, Guan X. Does being beautiful always help? contingency effects of physical attractiveness of the service providers on customer response. J Serv Mark 2019;33(3):356-368. [doi: 10.1108/jsm-05-2018-0140]
16. Keh HT, Pang J. Customer reactions to service separation. J Mark 2010;74(2):55-70. [doi: 10.1509/jmkg.74.2.55]
17. Zeithaml VA. How consumer evaluation processes differ between goods and services. Mark Serv 1981;9(1):25-32.
18. Connelly BL, Certo ST, Ireland RD, Reutzel CR. Signaling theory: a review and assessment. J Manag 2010;37(1):39-67. [doi: 10.1177/0149206310388419]
19. Dranove D. Demand inducement and the physician/patient relationship. Econ Inq 1988;26(2):281-298. [doi: 10.1111/j.1465-7295.1988.tb01494.x] [Medline: 10287407]
20. Wan Y, Peng Z, Wang Y, Zhang Y, Gao J, Ma B. Influencing factors and mechanism of doctor consultation volume on online medical consultation platforms based on physician review analysis. Internet Res 2021;31(6):2055-2075. [doi: 10.1108/intr-10-2020-0589]
21. Petrilli CM, Mack M, Petrilli JJ, Hickner A, Saint S, Chopra V. Understanding the role of physician attire on patient perceptions: a systematic review of the literature—Targeting Attire to Improve Likelihood of Rapport (TAILOR) investigators. BMJ Open 2015;5(1):e006578 [FREE Full text] [doi: 10.1136/bmjopen-2014-006578] [Medline: 25600254]
22. Rehman SU, Nietert PJ, Cope DW, Kilpatrick AO. What to wear today? effect of doctor's attire on the trust and confidence of patients. Am J Med 2005;118(11):1279-1286 [FREE Full text] [doi: 10.1016/j.amjmed.2005.04.026] [Medline: 16271913]
23. Au S, Khandwala F, Stelfox HT. Physician attire in the intensive care unit and patient family perceptions of physician professional characteristics. JAMA Intern Med 2013;173(6):465-467 [FREE Full text] [doi: 10.1001/jamainternmed.2013.2732] [Medline: 23420343]
24. Xun H, Chen J, Sun AH, Jenny HE, Liang F, Steinberg JP. Public perceptions of physician attire and professionalism in the US. JAMA Netw Open 2021;4(7):e2117779 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17779] [Medline: 34328503]
25. Ahearne M, Gruen TW, Jarvis CB. If looks could sell: moderation and mediation of the attractiveness effect on salesperson performance. Int J Res Mark 1999;16(4):269-284. [doi: 10.1016/s0167-8116(99)00014-2]
26. Jackson LA, Hunter JE, Hodge CN. Physical attractiveness and intellectual competence: a meta-analytic review. Soc Psychol Q 1995;58(2):108-122. [doi: 10.2307/2787149]
27. Eagly AH, Ashmore RD, Makhijani MG, Longo LC. What is beautiful is good, but…: a meta-analytic review of research on the physical attractiveness stereotype. Psychol Bull 1991;110(1):109-128. [doi: 10.1037/0033-2909.110.1.109]
28. Tsai WC, Huang TC, Yu HH. Investigating the unique predictability and boundary conditions of applicant physical attractiveness and non‐verbal behaviours on interviewer evaluations in job interviews. J Occup Organ Psychol 2010;85(1):60-79. [doi: 10.1348/2044-8325.002003]
29. Luoh HF, Tsaur SH. Physical attractiveness stereotypes and service quality in customer–server encounters. Serv Ind J 2009;29(8):1093-1104. [doi: 10.1080/02642060902764517]
30. Huneke T, Benoit S, Shams P, Gustafsson A. Does service employees' appearance affect the healthiness of food choice? Psychol Mark 2014;32(1):94-106 [FREE Full text] [doi: 10.1002/mar.20765]
31. Luca M. Designing online marketplaces: trust and reputation mechanisms. Innov Policy Econ 2017;17:77-93 [FREE Full text] [doi: 10.1086/688845]

XSL•FO
RenderX

32. Barnes SJ, Kirshner SN. Understanding the impact of host facial characteristics on Airbnb pricing: integrating facial image analytics into tourism research. Tour Manag 2021;83:104235. [doi: 10.1016/j.tourman.2020.104235]
33. Ert E, Fleischer A, Magen N. Trust and reputation in the sharing economy: the role of personal photos in Airbnb. Tour Manag 2016;55:62-73. [doi: 10.1016/j.tourman.2016.01.013]
34. Holzwarth M, Janiszewski C, Neumann MM. The influence of avatars on online consumer shopping behavior. J Mark 2006;70(4):19-36. [doi: 10.1509/jmkg.70.4.19]
35. Dunivin Z, Zadunayski L, Baskota U, Siek K, Mankoff J. Gender, soft skills, and patient experience in online physician reviews: a large-scale text analysis. J Med Internet Res 2020;22(7):e14455 [FREE Full text] [doi: 10.2196/14455] [Medline: 32729844]
36. Colletti LM, Mulholland MW, Sonnad SS. Perceived obstacles to career success for women in academic surgery. Arch Surg 2000;135(8):972-977 [FREE Full text] [doi: 10.1001/archsurg.135.8.972] [Medline: 10922261]
37. Försterling F, Preikschas S, Agthe M. Ability, luck, and looks: an evolutionary look at achievement ascriptions and the sexual attribution bias. J Pers Soc Psychol 2007;92(5):775-788. [doi: 10.1037/0022-3514.92.5.775] [Medline: 17484604]
38. Castro MR, Holvino E. Applying intersectionality in organizations: inequality markers, cultural scripts and advancement practices in a professional service firm. Gend Work Organ 2016;23(3):328-347. [doi: 10.1111/gwao.12129]
39. Agthe M, Spoerrle M. On the context sensitivity of the sexual attribution bias: a replication and extension to situations of failure. Open Psychol J 2009;2(1):19-24. [doi: 10.2174/1874350100902010019]
40. Maner JK, Gailliot MT, Rouby DA, Miller SL. Can't take my eyes off you: attentional adhesion to mates and rivals. J Pers Soc Psychol 2007;93(3):389-401. [doi: 10.1037/0022-3514.93.3.389] [Medline: 17723055]
41. Maner JK, Miller SL, Rouby DA, Gailliot MT. Intrasexual vigilance: the implicit cognition of romantic rivalry. J Pers Soc Psychol 2009;97(1):74-87. [doi: 10.1037/a0014055] [Medline: 19586241]
42. Rhodes G. The evolutionary psychology of facial beauty. Annu Rev Psychol 2006;57:199-226. [doi: 10.1146/annurev.psych.57.102904.190208] [Medline: 16318594]
43. Koernig SK, Page AL. What if your dentist looked like Tom Cruise? applying the match-up hypothesis to a service encounter. Psychol Mark 2001;19(1):91-110. [doi: 10.1002/mar.1003]
44. Lynch J, Schuler D. The matchup effect of spokesperson and product congruency: a schema theory interpretation. Psychol Mark 1994;11(5):417-445. [doi: 10.1002/mar.4220110502]
45. Shafei I, Walburg JA, Taher AF. Healthcare service quality: what really matters to the female patient? Int J Pharm Healthc Mark 2015;9(4):369-391. [doi: 10.1108/ijphm-05-2014-0028]
46. Weinstein ND. Perceived probability, perceived severity, and health-protective behavior. Health Psychol 2000;19(1):65-74. [doi: 10.1037//0278-6133.19.1.65] [Medline: 10711589]
47. Yang H, Du HS, Shang W. Understanding the influence of professional status and service feedback on patients' doctor choice in online healthcare markets. Internet Res 2020;31(4):1236-1261. [doi: 10.1108/intr-10-2019-0429]
48. Kasmaei P, Shokravi FAC, Hidarnia A, Hajizadeh E, Atrkar-Roushan Z, Shirazi KK, et al. Brushing behavior among young adolescents: does perceived severity matter. BMC Public Health 2014;14:8 [FREE Full text] [doi: 10.1186/1471-2458-14-8] [Medline: 24397588]
49. Geving AM. Identifying the types of student and teacher behaviours associated with teacher stress. Teach Teach Educ 2007;23(5):624-640. [doi: 10.1016/j.tate.2007.02.006]
50. Petty RE, Cacioppo JT, Schumann D. Central and peripheral routes to advertising effectiveness: the moderating role of involvement. J Consum Res 1983;10(2):135-146. [doi: 10.1086/208954]
51. DelPriore DJ, Bradshaw HK, Hill SE. Appearance enhancement produces a strategic beautification penalty among women. Evol Behav Sci 2018;12(4):348-366. [doi: 10.1037/ebs0000118]
52. Mortimer K, Pressey A. Consumer information search and credence services: implications for service providers. J Serv Mark 2013;27(1):49-58. [doi: 10.1108/08876041311296374]
53. Krishnamurthy S. A comprehensive analysis of permission marketing. J Comput-Mediat Commun 2001;6(2):JCMC623 [FREE Full text] [doi: 10.1111/j.1083-6101.2001.tb00119.x]
54. Heilman ME, Stopeck MH. Being attractive, advantage or disadvantage? performance-based evaluations and recommended personnel actions as a function of appearance, sex, and job type. Organ Behav Hum Decis Process 1985;35(2):202-215. [doi: 10.1016/0749-5978(85)90035-4]
55. Ruffle BJ, Shtudiner Z. Are good-looking people more employable? Manag Sci 2015;61(8):1760-1776. [doi: 10.1287/mnsc.2014.1927]
56. Fink B, Neave N, Manning JT, Grammer K. Facial symmetry and judgements of attractiveness, health and personality. Personal Individ Differ 2006;41(3):491-499. [doi: 10.1016/j.paid.2006.01.017]
57. Hahn AC, Perrett DI. Neural and behavioral responses to attractiveness in adult and infant faces. Neurosci Biobehav Rev 2014;46(Pt 4):591-603. [doi: 10.1016/j.neubiorev.2014.08.015] [Medline: 25199981]
58. Morrow PC, McElroy JC. The impact of physical attractiveness in evaluative contexts. Basic Appl Soc Psychol 2010;5(3):171-182. [doi: 10.1207/s15324834basp0503_1]
59. The 52nd statistical report on China's internet development. China Internet Network Information Center (CNNIC). 2023. URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202311/P020231121355042476714.pdf [accessed 2024-04-18]

XSL•FO
RenderX

60. Kim HW, Xu Y, Koh J. A comparison of online trust building factors between potential customers and repeat customers. J Assoc Inf Syst 2004;5(10):392-420. [doi: 10.17705/1jais.00056]
61. Pedulla DS. The positive consequences of negative stereotypes: race, sexual orientation, and the job application process. Soc Psychol Q 2014;77(1):75-94. [doi: 10.1177/0190272513506229]

Abbreviations

SAB: sexual attribution bias

Edited by A Mavragani; submitted 20.02.23; peer-reviewed by N Farre, D Verran; comments to author 20.07.23; revised version received 25.10.23; accepted 08.04.24; published 30.05.24.

Please cite as:
Wei X, Yu S, Li C(
Influence of Physical Attractiveness and Gender on Patient Preferences in Digital Doctor Consultations: Experimental Study
J Med Internet Res 2024;26:e46551
URL: https://www.jmir.org/2024/1/e46551
doi:10.2196/46551
PMID:38814690

©Xia Wei, Shubin Yu, Changxu (Victor) Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

How Informal Carers Support Video Consulting in Physiotherapy, Heart Failure, and Cancer: Qualitative Study Using Linguistic Ethnography

Lucas Martinus Seuren[1,2], PhD; Sara Shaw[1], PhD

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
[2]Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada

Corresponding Author:
Lucas Martinus Seuren, PhD
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford, OX2 6GG
United Kingdom
Phone: 1 4372607372
Email: lucas.seuren@phc.ox.ac.uk

Abstract

Background: Informal carers play an important role in the everyday care of patients and the delivery of health care services. They aid patients in transportation to and from appointments, and they provide assistance during the appointments (eg, answering questions on the patient's behalf). Video consultations are often seen as a way of providing patients with easier access to care. However, few studies have considered how this affects the role of informal carers and how they are needed to make video consultations safe and feasible.

Objective: This study aims to identify how informal carers, usually friends or family who provide unpaid assistance, support patients and clinicians during video consultations.

Methods: We conducted an in-depth analysis of the communication in a sample of video consultations drawn from 7 clinical settings across 4 National Health Service Trusts in the United Kingdom. The data set consisted of 52 video consultation recordings (of patients with diabetes, gestational diabetes, cancer, heart failure, orthopedic problems, long-term pain, and neuromuscular rehabilitation) and interviews with all participants involved in these consultations. Using Linguistic Ethnography, which embeds detailed analysis of verbal and nonverbal communication in the context of the interaction, we examined the interactional, technological, and clinical work carers did to facilitate video consultations and help patients and clinicians overcome challenges of the remote and video-mediated context.

Results: Most patients (40/52, 77%) participated in the video consultation without support from an informal carer. Only 23% (12/52) of the consultations involved an informal carer. In addition to facilitating the clinical interaction (eg, answering questions on behalf of the patient), we identified 3 types of *work* that informal carers did: facilitating the use of technology; addressing problems when the patient could not hear or understand the clinician; and assisting with physical examinations, acting as the eyes, ears, and hands of the clinician. Carers often stayed in the background, monitoring the consultation to identify situations where they might be needed. In doing so, copresent carers reassured patients and helped them conduct the activities that make up a consultation. However, carers did not necessarily help patients solve all the challenges of a video consultation (eg, aiming the camera while laying hands on the patient during an examination). We compared cases where an informal carer was copresent with cases where the patient was alone, which showed that carers provided an important safety net, particularly for patients who were frail and experienced mobility difficulties.

Conclusions: Informal carers play a critical role in making video consultations safe and feasible, particularly for patients with limited technological experience or complex needs. Guidance and research on video consulting need to consider the availability and work done by informal carers and how they can be supported in providing patients access to digital health care services.

(J Med Internet Res 2024;26:e51695) doi:10.2196/51695

XSL•FO
RenderX

KEYWORDS

delivery of health care; remote consultation; carer; telemedicine; videoconferencing; language; linguistics; gestures; physiotherapy; heart failure; care; patient care; feasibility; safety; communication; mobile phone

Introduction

Background

Video consulting has become an established health care service model since the outbreak of the COVID-19 pandemic [1]. Video consultations have been shown to be safe and effective in a range of clinical settings [2-6]. Patients and clinicians have largely reported positive experiences, particularly in secondary and tertiary care [7-10], with some patients even preferring video consultations over face-to-face consultations, especially for follow-up appointments and where a trusted relationship with the provider is already in place [11]. Given the policy push for remote health care services to continue beyond the COVID-19 pandemic [12-14], it is clear that video consulting is here to stay. However, significant concerns remain around when video consulting is feasible and appropriate (eg, for which patients and in which clinical settings) [15-17]. Some patients still do not have access to the necessary technology (ie, smartphone, tablet, or computer and high-quality internet) [18,19] and they may also lack the experience, confidence, or ability to use it for a medical consultation [20-22]. In these situations, carers, either professional or informal (eg, family and friends), can provide assistance [23,24].

There is extensive literature on video consulting, documenting the benefits and challenges for patients and clinicians [5]. However, very few studies address how informal carers participate in video consultations [25,26]. This is surprising, given that informal carers play an important role in patient care. Informal carers, usually family or friends, “[provide] unpaid care and support to a family member, friend or neighbour who is disabled, has an illness or long-term condition, or who needs extra help as they grow older” [27]. In the United Kingdom, approximately 6 million people provide unpaid care, many of whom play a vital role in coordinating and supporting care received by the person they care for [28,29]. Therefore, it is important to understand the role they play in supporting and delivering video-consulting services.

Contemporary health care systems and policy makers have been pushing a transition to patient-centered or person-centered care, that is, care that is “respectful and responsive to individual preferences, needs and values” [30]. However, person-centered care has often been taken to only mean patient-centered care. Guidelines do not always address family or carers, and where they do, they merely highlight that practitioners must involve carers in patient care, for example, by asking them to clarify the patient’s wishes [31]. In other words, the focus in person-centered care is still on the patient. Nevertheless, informal carers play a central role in the delivery of care, supporting patients (to varying degrees and in varying situations) with their needs and care. Carers may deliver up to 90% of the care and support for patients in the community [32]. Therefore, it is potentially important for guidance on video consultations to take carers, and the support of carers, into account. Given that the work done by carers can cause a significant mental and physical strain [33], practitioners and policy makers need to consider the preferences, needs, and values of patients and carers.

Health communication research has shown that carers sometimes play an active role in making in-person consultations work: carers may speak on behalf of the patient (eg, to provide additional medical or other information for children or patients who lack capacity), alongside the patient (eg, when planning a next appointment), or with the patient (eg, to help them answer questions about their medication) [34-36]. However, having a carer copresent, that is, physically with the patient in the consultation, can be challenging as patients, clinicians, and carers report that they have trouble managing turn-taking [37]. This raises questions regarding when the carer is able to talk, what they can talk about, and how they can determine this.

Participation problems may be more pronounced in video consultations. From research outside the health care setting, we know that it can be difficult for carers to facilitate a conversation over video [38]. The camera restricts the field of view, and generally, only 1 person is visible at a time on each end [39]. The clinical context adds additional challenges, with participants having to manage new interactional skills (eg, how to begin a video consultation) and accomplish activities that are constrained by the lack of physical copresence (eg, conducting a physical examination) [5,40,41].

To date, only 1 study has investigated how the constraints of technology affect communication in health care where informal carers are copresent, focusing on postoperative cancer consultations in the Netherlands and showing that carers often remain offscreen and do not actively participate, and when they do, they mostly talk to the patient [42]. Several other studies have investigated how professional carers (eg, copresent nurses or primary care physicians) participate in video consultations, with a focus on how these professional Despite their crucial role in health care delivery, informal carers have not yet benefited from the advancements made in this field [29].

Objectives

Overall, there is a need to understand how informal carers support video consultations when they are copresent with the patient. This study focuses on how informal carers support patients and clinicians during video consultations. Our focus is on the *work* (either interactional, clinical, or technological) that informal carers do to make video consultations work to provide key insights into how they affect the feasibility of video consulting. To support our analysis, we compared the consultations where informal carers provided support and the reflections of participants in subsequent interviews with consultations where patients were alone and the reflections of those participants.

XSL•FO
RenderX

Methods

Study Design

We conducted a qualitative, multimethods study using Linguistic Ethnography, which combines ethnographic approaches (ie, observation and interviews) with the close inspection of interactional data [43]. We used *ethnography of communication* [44] to guide our understanding of how the context of video consultations (eg, restricted visual field) may shape the ways in which patients, carers, and clinicians communicate over video. We combined this with conversation analysis, an inductive method that investigates the turn-by-turn construction of a conversation, to understand the communication practices (verbal and nonverbal) that make up a video consultation [45]. Combining these methods enabled us to show how the interactions in video consultations shape, and are shaped by, the wider sociocultural and clinical contexts (eg, established clinical routines, policy, and technology in use) [46].

For this study, we conducted secondary analysis of qualitative data that were previously collected for 3 separate studies on video consultations in different clinical settings across 4 National Health Service clinics in the United Kingdom (1 in Oxford and 3 in London):

- Supporting Consultations in Remote Physiotherapy (SCiP; 2021-2022) was funded by the National Institute for Health Research to investigate the feasibility and practicalities of video consultations for physiotherapy.
- Virtual Online Consultations: Advantages and Limitations (VOCAL; 2015-2017) was funded by the National Institute for Health Research and investigated (gestational) diabetes and cancer.
- Oxford Telehealth Qualitative Study (OTQS; 2015-2017) was funded by the Wellcome Trust to investigate the feasibility of video consulting in a specialist nurse service for patients with heart failure.

Data were chosen for convenience, having been collected as part of research studies that had already been conducted by members of the larger research team and available for secondary analysis [47,48].

Data Collection

We analyzed all 52 video recordings of video consultations that were recorded across the 3 studies. Data for VOCAL and OTQS were collected from 2015 to 2017 (refer to the study by Shaw et al [5] for an overview), and data for SCiP were collected from 2021 to 2022 (refer to the study by Seuren et al [47] for an overview).

In all 3 studies, recruitment was done based on convenience. For VOCAL and OTQS, which took place before the COVID-19 pandemic when video consulting was still relatively unfamiliar, patient participants were recruited in collaboration with clinical staff to ensure that patients were suitable to have a video consultation. The aim was to create a sample with a range of experiences with video consultations, "seeking maximum variety in clinical, ethnic and personal circumstances." Patients were initially contacted by their clinician, after which the research team sent out an invitation letter [5]. For SCiP, data collection took place between August 2021 and March 2022, during the COVID-19 pandemic. Initially, clinicians reached out to any patient who had an upcoming appointment by video. Those who showed an interest in the study were subsequently contacted by a member of the research team to explain the details of the study [47]. For all studies, exclusion criteria were the inability to give informed consent and comorbidity preventing participation. For VOCAL and OTQS, additional exclusion criteria were no 3G internet access at home and lack of familiarity with technology [5].

Video consultations for VOCAL and OTQS were recorded using small digital camcorders (Sony Handycam DCR-SR72; Sony Corporation) and a handheld iPad (Apple Inc), combined with a commercially available screen-capture software tool (ACA Systems), which was run directly from an encrypted USB memory stick. Whenever feasible, both the patient's and the clinician's end of the consultation had been recorded, capturing the consultations and the context in which they took place. The total data set from VOCAL and OTQS consisted of 37 video recordings and transcripts of the video consultations (cancer: 12/37, 32%; diabetes: 12/37, 32%; heart failure: 7/37, 19%; and gestational diabetes: 6/37, 16%), 35 transcripts of semistructured interviews conducted with staff and 26 transcripts of semistructured interviews conducted with patients involved in these consultations (Table 1) [5].

Video consultations for SCiP were recorded by the clinical team in the 2 National Health Service Trusts using the built-in recording option in Microsoft Teams (Microsoft Corp). The total data set consisted of 15 video recordings and transcripts of video consultations (neuromuscular rehabilitation: 5/15, 33%; long-term pain: 1/15, 7%; and orthopedics: 9/15, 60%), 15 transcripts of semistructured interviews with patients and 7 transcripts of semistructured interviews with clinicians involved in these consultations (Table 2) [47].

XSL•FO
RenderX

Table 1. Participant data from the Virtual Online Consultations: Advantages and Limitations study and Oxford Telehealth Qualitative Study (n=37).

Clinic	Total recordings, n (%)	Gender, n (%)	Age (years), median (IQR)	Ethnicity, n (%)
Diabetes	12 (32)	Male: 5 (42) and female: 7 (58)	23 (22-35)	Asian Bangladeshi: 1 (8), Asian Indian: 3 (25), Black Caribbean: 1 (8), White British: 5 (42), and White Other: 2 (17)
Gestational diabetes	6 (16)	Female: 6 (100)	34 (30-35)	Asian Bangladeshi: 1 (17), Asian Other: 3 (50), Black Caribbean: 1 (17), and White British: 1 (17)
Cancer	12 (32)	Male: 4 (33) and female: 8 (67)	74 (60-77)	Asian Indian: 1 (8), Black Caribbean: 1 (8), White British: 9 (75), and White Other: 1 (8)
Heart failure	7 (19)	Male: 3 (43) and female: 4 (57)	67 (56-76)	White British: 7 (100)

Table 2. Participant data from the Supporting Consultations in Remote Physiotherapy study (n=15).

Clinic	Total recorded, n (%)	Gender, n (%)	Age (years), median (IQR)	Ethnicity, n (%)
Orthopedics	9 (60)	Female: 8 (89) and male: 1 (11)	35 (29-57)	White British: 7 (78), White Irish: 1 (11), and White Mauritian: 1 (22)
Neurology (neuromuscular)	5 (33)	Female: 1 (20) and male: 4 (80)	41 (40-47)	White British: 5 (100)
Long-term pain	1 (7)	Female: 1 (100)	20[a]	White Scottish 1 (100)

[a]There was only one participant; hence, there is no IQR.

Analysis

An initial exploration of the 52 recorded video consultations across all 3 studies showed that informal carers performed a range of seemingly vital tasks in some (but not all) video consultations (12/52, 23%; eg, holding the tablet and laying hands on the patient). This raised questions about the role of carers in video consultations. We collected all instances in our video data where carers were involved at any point during a video consultation and corresponding interview data in which participants in these video consultations reflected on the work carers do. As a routine practice in conversation analysis [49,50], we then built "collections" of similar cases [51], organizing the data based on the type of work done by carers. To further refine our analysis, we compared our findings with consultations where no carer was present (40/52, 77%), combining researcher observations of potentially risky situations (eg, an older patient nearly fell) with clinician reflections on these consultations to identify cases where the lack of a copresent carer might have negatively affected the quality of care. On the basis of these collections, we then analyzed the qualitative interviews deductively using thematic analysis [52]. Themes were identified based on our analysis of the consultations and used deductively to analyze the interviews. We examined how participants talked about the 3 key themes, aiming to discern whether participants' reflections were in line with our findings of the consultations (eg, when and why do patients require assistance with technology) or whether they offered complementary (eg, additional work done by informal carers outside of the consultation) or even contradicting viewpoints (eg, informal carers not being able to offer support). Our analysis focused on the conversation analysis of the consultations, with supporting reflections from the participants.

As all data were selected for convenience, the consultations that involved a carer and those that did not involve a carer were not matched regarding, for example, clinical context, patient demographics, or type of technology used.

We transcribed all video consultations orthographically and subsequently used established conventions for verbal and nonverbal communication [53,54] for the data in our collections. This is a routine practice in conversation analysis and, for this paper, enabled us to track how and why carers assist in video consultations. In the Results section, we present simplified extracts from transcripts, providing orthographic transcripts complemented with notations for silence and overlapping talk to maintain legibility. We added screengrabs to allow readers to appreciate the context of consultations and how participants use their bodies and material objects (eg, how they move and hold a tablet). All interviews were transcribed orthographically. We extracted screengrabs using Adobe Premiere Pro 2023 (Adobe Inc), adding a video filter and facial blur to deidentify participants. Subsequently, we combined these screengrabs with the transcript in Adobe InDesign 2023 (Adobe Inc) and exported these at 600 dots per inch to generate the figures.

Ethical Considerations

All studies received ethics approval for detailed analysis of video recordings of video consultations and audio recordings of interviews. VOCAL was approved by the National Research Ethics Committee London-City Road and Hampstead in December 2014 (14/LO/1883), OTQS by the South Central-Berkshire Research Ethics Committee in September 2015 (15/SC/0553), and SCiP by the East Midland-Nottingham 1 Research Ethics Committee in April 2021 (21/EM/0082). All participating staff and patients provided their informed consent to be audio and video recorded during consultations and

XSL•FO
RenderX

interviews and for the data to be used for research purposes, including secondary analysis.

Patients were initially approached by a member of their clinical care team. After signaling an interest in the study, the patient's contact information was forwarded to a member of the research team. The author provided the patient with an information sheet to review. After providing an opportunity to ask questions, patients were asked if they wanted to participate, and if they agreed, they were asked to sign the consent form. For VOCAL and OTQS, patients provided consent during an in-person conversation with a member of the research team. For SCiP, to comply with infection control procedures during the COVID-19 pandemic, patients provided verbal consent during a video call. Participants did not receive compensation for participation in any of the 3 studies.

All transcriptions were anonymized by removing identifying data and replacing names with descriptions (eg, NAME, where someone's name is used). Participants consented to the analysis of the raw (ie, recognizable) video data. For publication, video data were anonymized using a visual filter and blur effect in Adobe Premiere Pro 2023.

Results

Main Findings

Of the 52 video consultations in our data, 12 (23%) involved a copresent carer: 8 (67%) with patients with cancer, 3 (25%) with patients with heart failure, and 1 (8%) with a patient consulting for physiotherapy. None of the patients with gestational diabetes had a copresent carer. In these 12 consultations, we identified three main types of work that carers performed: (1) facilitating the use of the technology, (2) helping the patient hear or understand what the clinician said, and (3) assisting the patient with and performing parts of the physical examination. Carers performed these tasks through the use of verbal and nonverbal communication strategies, as seen in the data extracts, screengrabs, and participants' reported experiences in the following sections. Furthermore, we found that in 10% (5/52) of the consultations the patient did not have a carer copresent, but either the patient or clinician expressed concerns regarding safety during the consultation (eg, a patient saying, "I'm not sure if I'll be able to get back up again") or the clinician, during the interview afterward, commented that they felt they might have put the patient in an unsafe situation.

Facilitating the Use of Technology in Video Consultations

Informal carers facilitated the use of technology for video consultations in 2 ways: they provided patients access to the technology, and their presence and perceived expertise provided patients with confidence and reassurance for using the technology.

In our data, some patients (5/52, 10%) either did not have the technology or had never used it for video chat. Therefore, they relied on carers to set up, and sometimes provide, the technology. This facilitation involved activities such as the carer bringing a tablet for the patient to use, registering a Skype (Skype Technologies) account, adding the clinician as a contact on Skype or FaceTime (Apple Inc), talking to the clinician beforehand regarding any practicalities, and explaining to the patient what to expect from the video consultation. For the patients who lacked experience with video-mediated communication, carers provided a sense of reassurance if something went wrong or if there were difficulties. This was evident both in how the informal carers acted in the consultations and how they discussed their experiences during the interviews. An older patient explained before her oncology consultation that she only agreed to a video consultation if her husband would be there:

> *First uh, I was a bit uh, I said uh, if he's here it's fine. I haven't got any problem.*

Another older patient stated after her heart failure consultation that, while she could learn to use the technology, she relied on her daughter being there and would not have been able to do it on her own:

> *Patient: that's what I really think, that for me,... it's easy. Because I don't have to sit here and think, what if I do something wrong?*
>
> *Carer: no*
>
> *Patient: for people, old people on their own, entirely different.*
>
> *Carer: yeah. it is entirely different.*
>
> *Patient: And I would not be able to do it on my own. ... I wouldn't have the confidence.*

During consultations, we found that carers often facilitated the use of technology while being silent (ie, nonspeaking) and offscreen. This involved carers performing 3 types of background activities that allowed the patient to consult with the clinician via video: they handled the "preopening," the work people do before they start a video consultation [55]; they handled the camera allowing the patient and clinician to adequately see each other; and they made sure that the patient and clinician could hear each other.

In 8 (67%) of the 12 consultations, carers took care of the "preopening" [55]: they set up the technology, logged in, and answered the call from the clinician when using a program such as Skype or FaceTime. Then, the patient took over when the consultation started.

In the example in Figure 1, the patient had never used FaceTime before and did not own a video communication technology (eg, a smartphone or tablet). The carer brought a tablet with her, signed into FaceTime, and held it ready for use. When the clinician called via video, the carer explained to the patient that they would accept the call (line 1). Then, she swiped to answer, pointed out to the patient when the connection was established (line 7) and answered the video call with a "hello, conveying to the clinician that the connection had been established and they were ready [56]. The carer stayed out of the frame (refer to screengrab 2 in Figure 1) and hence out of the interaction [42,57], allowing the patient to conduct her consultation while still remaining available in the background.

XSL•FO
RenderX

Figure 1. Carer helping the patient at the start of a video consultation.

```
01. Carer:      alright. so we're gonna accept that, alright,
02.             (.)
03. Patient:    yeah
04.             (0.5)
05. Carer:      and then you'll see,
06.             (0.5)
07.             there we go_=
08. Nurse:      =oh.
09.             (0.3)
10. Carer:      hello?
11. Nurse:      good morn- (0.3) good morning ((name patient))?
12. Patient:    uh good morning ((name nurse)). how are you.
```

Screengrab 1: Patient's view

Screengrab 2: Nurse's view

In 7 (58%) of the 12 consultations, carers did additional background work that enabled patients to talk to their clinician. This included making sure that the patient and clinician could adequately see and hear each other, for example, by acting as a cameraperson: positioning the technology and framing the patient throughout the consultation [38] to maintain a "talking heads" configuration for the patient and clinician [39], a setup in which both participants are visible from the shoulders up. In 71% (5/7) of these consultations, the carer held a tablet or smartphone, moving this to frame the patient while remaining outside the frame themselves. In 29% (2/7) of these consultations, the patient used a desktop PC, so the carer moved the patient instead of the technology.

In the example in Figure 2, from the start of an oncology consultation, the patient was at the left edge of the field of view of the camera and only half of her face was visible to the clinician. As soon as the physician told the patient to "move slightly" (line 3), the carer turned toward the patient and began to pull their chair. At the point where the physician completed his request (line 8), the patient was visible in the center of the screen. Our recording of the clinician's end does not capture the screen. However, on the screen on the patient's side, we can see that initially only the right half of her face is visible, and the carer then adjusts the chair so that the patient is centered and fully visible.

Figure 2. Carer helping to make the patient visible.

```
01. Pt: we're off.
02.     (0.6)
03. Dr: you need to (.) move slightly, or
04.     turn the camera slightly, so that
05.     you are in the centre of the
06.     picture. we can only see a part of
07.     your face.
08.     (0.9)
09.     yeah that's better,
```

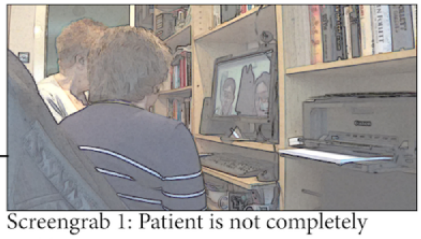
Screengrab 1: Patient is not completely visible

Screengrab 2: Carer pulls the patient's chair

In 4 (33%) of the 12 consultations, the carers acted as a technological facilitator to ensure the audio and video were working. Carers did most of this work at the start of the consultation. This was the first point where participants could determine whether the sound and video were working. In the example in Figure 3, the carer answered the clinician's call when he appeared on screen by saying "hello" (line 1), but the clinician did not respond. The carer treated this silence as indicating a problem: she said "hello" again but this time with a more *questioning* intonation (a strong rising pitch on the "o"), a typical communication strategy for testing if someone can still hear [58].

XSL•FO
RenderX

Figure 3. Carer checking the connection at the opening of a video consultation.

```
01. Cr:  hello?
02.      (3.6)
03.      hel[lo?
04. Dr:     [hi.
05.      (0.4)
06. Pt:  hello.
07.      (0.9)
08. Dr:  can you hear me?
09. Pt:  I [can hear you,
10. Cr:    [yeah
11.      can you hear us_
12. Dr:  (   )
13.      (0.5)
14.      yes. yes yes,
15.      and see you very well.
```

Screengrab 1: Carer leaning forward

Screengrab 2: Carer leaning back

After the clinician had said "hi" (line 4), he asked whether the patient (and carer) could hear him (line 8). The patient and carer confirmed (lines 9-10), and the carer checked whether the clinician could hear them. In other words, before the consultation began, the carer and clinician ensured that the technology was working and that the patient and clinician could see and hear each other. It was only when the physician had confirmed (line 14), that the consultation proceeded in a usual manner. At this point, the carer faded into the background.

Staying largely in the background (and so invisible to the clinician), carers typically maintained an active role, helping to address any problems (eg, lost connection or microphone on mute) that arose during the consultation. In these instances, carers temporarily became active participants while fixing the problem. In the example in Figure 4, the physician asked the patient "how are you." However, a technical disruption occurred and his turn was cut off after "ho." After a few seconds of silence, the physician said, "what happened" (line 3), taking the lack of response by the patient as indicative of a problem. It was the carer who then switched to become an active coparticipant, asking if the physician could hear them (line 5). Once all parties had established that they could see and hear each other, the physician acknowledged (line 13) [59] and repeated the question.

Figure 4. Carer troubleshooting the connection problem.

```
01. Doctor:   so (0.9) tell me, ho-
02.           (4.2)
03.           what happened,
04.           (2.4)
05. Carer:    can you- can you- can you hear us?
06. Doctor:   I can hear you. And I can see you.
07.           (0.2)
08.           can you hear me?
09.           (0.7)
10. Carer:    yes. We can hear.
11.           we [can see you and [hear you.
12. Patient:     [(we can hear)
13. Doctor:                       [okay.
14.           how are you?
15.           (0.6)
16. Patient:  I'm fine thank you.
```

Overall, carers in our data made video consultations feasible by facilitating the use of technology. Much of this work involved carers moving from being coparticipants to listening in the background, often unobservable to the clinician. They did this either by self-selecting to respond to a clinician's question (Figures 3 and 4) or by being selected by the clinician to answer a question. After responding, they would visually move out of the screen, or at least, no longer respond or take turns.

Making the Interaction Work in Video Consultations

Patients in our study occasionally had problems with hearing or understanding the clinician (eg, due to soft or distorted sound). Such problems happen routinely in any form of conversation [60], and people have a large array of *repair* strategies to fix them [61,62]. Normally, when trouble arises, recipients ask the speaker to repeat or clarify (part of) their turn (eg, by repeating the part of the turn they did hear or using exclamations such as "sorry," "what," or "huh" [63-65]). During in-person consultations, if patients have problems, they can ask the clinician to clarify [66].

In our video consultation data, we found that 25% (3/12) of the patients relied on their carer to help them hear or understand the clinician's talk (in all 3 consultations, the quality of the call

XSL•FO
RenderX

was problematic, eg, low volume and distortions). The example in Figure 5 illustrates how carers perform this type of interactional repair. In lines 1 to 4, the physician checked that the patient had seen one of his registrars the week before at an in-person consultation. At this point, the volume was low, making it hard to hear. Moving into the physician's turn, the patient started squinting (refer to screengrab 1 in Figure 5), indicating she had a problem. When the physician finished his question, the patient remained silent for 700 milliseconds (a substantial amount of time, given the usual response time for face-to-face interaction being 0-200 milliseconds [67]), indicating difficulty [68]. Instead of answering, the patient turned to the carer (refer to screengrab 2 in Figure 5), softly asking "what?" (indicated with the degree symbols) and expecting the carer to perform an interactional repair on the physician's question. The carer (offscreen) repeated the physician's verification question in line 8. Once the patient could answer, she started to nod, turned her gaze toward the physician (line 9), and answered (line 11) loudly, thereby making clear her response was now directed to the physician.

While this was a brief interaction, the carer in this example played a crucial role in the successful communication between the physician and the patient. The patient mobilized the carer to help her hear what the physician said. Akin to an interpreter, the carer "animated" the physician's talk [57]. Similar examples using indirect communication (eg, physically turning to the carer when something was unclear) were evident across our data set, where patients sought help from carers to enable repair and continuation of the interaction.

Figure 5. Carer repeating the clinician's question.

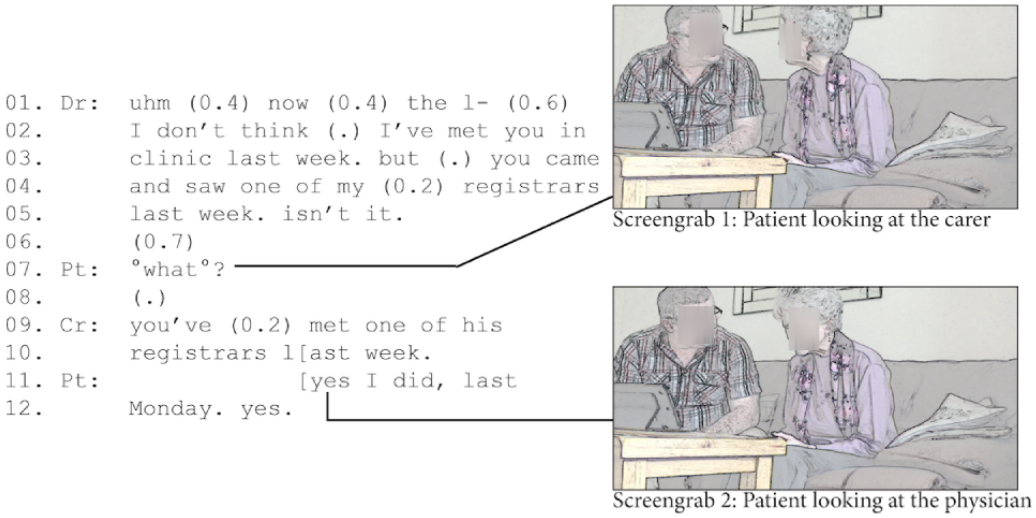

Making Physical Examination Possible in Video Consultations

The final area where carers made a vital contribution to video consultations was during physical examinations. The inability of clinicians to lay hands on the patient is one of the main concerns among clinicians and patients about video consultation [69-71]. Instead, patients have to describe and show their body and, where available, use their own devices such as oximeters (a device that people clip onto their finger to measure their blood oxygen saturation and heart rate) [40,72].

Carers supported remote physical examinations in 8 (67%) of the 12 video consultations in our data. This included helping to make the relevant parts of a patient's body visible, acting as the clinician's hands to perform tactile examinations and providing visual assessments, and assisting the patient with operating equipment such as blood pressure meters. Support was typically for patients who were frail, in cases where they were either unable to bend over (eg, due to the nature of their condition) or unable to move their tablet or laptop at the same time as moving their body (a complex sociotechnical task that was particularly challenging for those experiencing chronic illness) [72].

Figure 6 illustrates how carers can play a vital role in the feasibility of a physical examination. The patient had recently undergone surgery to remove a tumor and had complained to the physician about pain in her abdomen around the scar. The physician asked to examine the scar, requesting her to stand up (lines 1-3). The patient did not respond to this request. Instead, she waited for the carer to help out. After 1.3 seconds of silence, the physician made his request again, but at the same time, the carer said "hold on." Then, the carer helped the patient lift her sweater and aimed the camera toward the scar, allowing the physician to perform a visual assessment (lines 16-19).

In the example in Figure 6, the role of the carer was crucial for making the physical examination work. With limited physical capacity (and technological literacy), the patient was unable to hold the tablet and show the clinician her abdomen. It was only with the help of her carer that she could provide a sufficiently clear view for the physician to perform a visual assessment remotely.

XSL•FO
RenderX

Figure 6. Carer assisting in making the patient's scar visible.

```
01. Dr:  okay. let- let us have a look at
02.      your scar, can you stand up and
03.      show me your tummy,
04.      (1.3)
05.      show [me the scar.
06. Cr:      [hold on,
07.      (1.0)
08. Pt:  (that's really)(      ),
09.      (1.3)
10. Cr:  (   )(mum)
11.      (1.9)
12. Pt:  (              ) okay, sl::
13.      (2.6)
14. Pt:  can you see?
15.      (0.8)
16. Dr:  okay, yes, I can see it, the scar
17.      looks alright?
18.      (0.4)
19.      it looks nice and (.) clean,
```

Screengrab 1: Carer lifting the sweater

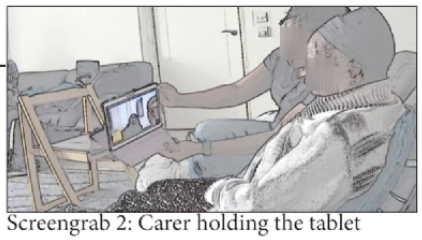
Screengrab 2: Carer holding the tablet

At times, clinicians relied on a carer to lay hands on the patient on their behalf during video consultations. In the heart failure consultations—both routine follow-up consultations—the specialist nurse wanted to check whether the patients had fluid build-up (edema) in their legs and ankles by pressing their thumb on the patient's leg and check whether this leaves an indentation. Carers played a vital role in performing these remote assessments, which involved patients who were frail, with restricted mobility and breathlessness, and for whom moving could cause severe discomfort [72]. Figure 7 illustrates an example in which the patient had just measured her blood oxygen saturation with an oximeter. Then, the nurse addressed the carer directly, announcing that she wanted to check the patient's legs (lines 1-4). Depicting how the carer should hold her hand (lines 11-16) [72,73], she explained how to press (lines 18-19). The carer followed these instructions and pressed the patient's legs several times. Using the camera on the back of the tablet, she not only performed the examination but also did so while simultaneously monitoring what the nurse could see (refer to screengrabs 2 to 4 in Figure 7). The carer's presence meant that the nurse was able to make a good assessment of the patient's legs, telling the carer that "you're doing a good job, and I can see it really clearly on the screen."

Figure 7. Carer assessing for edema.

```
01. Nurse:   darling, (0.3) ((carer name))?
02. Carer:   ye[s?
03. Nurse:     [I wanna che- I wanna check (.) your mum's legs.
04.          for swelling.
05. Nurse:   so (.) you know (.) in clinic? when I (.) check
06.          for any swelling in her le[gs?
07. Carer:                             [you feel, yes,
08.          (0.2)
09. Nurse:   yeah.
10.          so what you need to do, is just t-
11.          if you see my hand, you just use y'r thumb
12.          and your finger?
13.          (0.2)
14. Carer:   right,
15. Nurse:   (   ) you just need to press_=you know the bone,
16.          on her shin,
17.          (0.6)
18.          you just need to hold that. sort of quite firmly,
19.          for a couple of seconds,
```

Screengrab 1: Carer copying the grip

Screengrab 2: Carer pointing towards where to press

Screengrab 3: Carer pressing into the shin

Screengrab 4: Carer aiming camera at the shin

XSL•FO
RenderX

In total, 2 (4%) of the 52 cases in our wider data set of video consultations flagged questions regarding the safety of physical assessment where carers were not present. Figure 8 illustrates the example of a neuromuscular physiotherapy consultation with a patient with Charcot-Marie-Tooth disease (a neurological disorder that causes damage to the peripheral nerves leading to muscle weakness and atrophy), who struggled with walking and balance. At one point, the clinician asked the patient to stand up so that she could see her walk while holding onto a wall. The patient had to push herself from the bed, had difficulty standing up without losing her balance, and had to use both hands to help herself. In hindsight, the clinician acknowledged that this may have been too difficult.

Figure 8. Patient struggling to stand.

```
01. Physio:      yea::h. so standing still 02. is har:d, isn't
02.              it, I can [see that.
03. Patient:               [(   )
04.              (.)
05. Patient:     °yeah°.
```

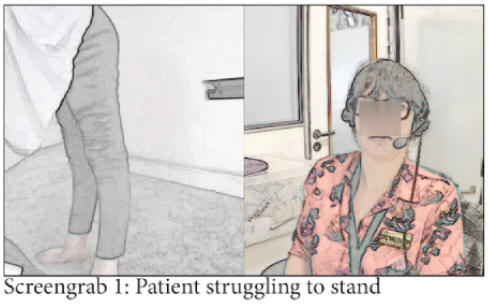

Screengrab 1: Patient struggling to stand

We identified a similar case in our heart failure data, where an older patient raised her leg to the camera, allowing her nurse to assess whether there was any swelling (Figure 9). The patient needed to stand and had to hold on to the chair in front of her to maintain her balance, but the uncomfortable position caused abdominal cramps and led her to drop her leg. This raised questions regarding safety while also placing limits on what was only a brief visual assessment.

Figure 9. Older patient holding her leg up to be seen in the camera.

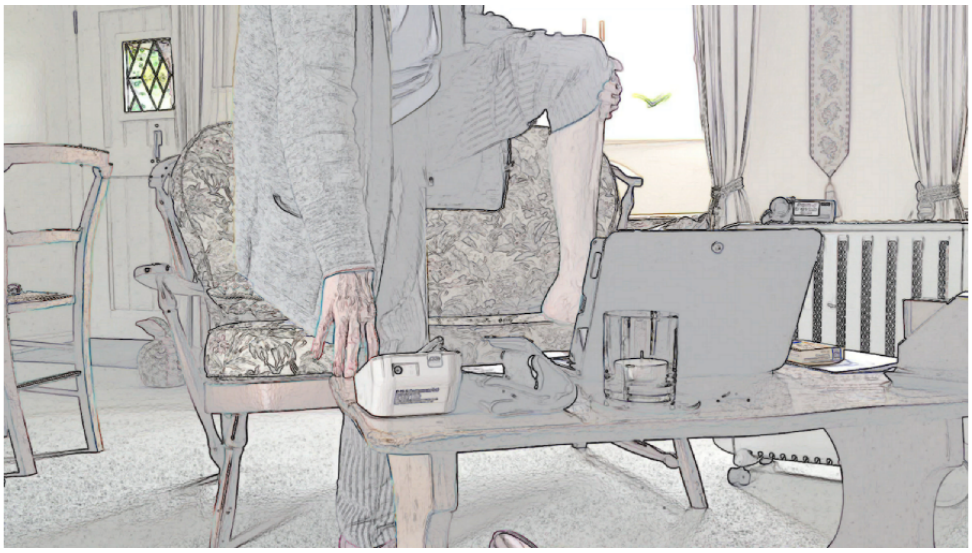

Furthermore, carers helped some patients (2/12, 10%) operate measuring devices during examination. This was particularly relevant in remote heart failure consultations, in which all 7 patients needed to measure their oxygen saturation, blood pressure, and heart rate. All 7 patients were able to use the oximeter; however, operating a blood pressure meter proved challenging for 2 patients, both experiencing frailty. In both cases, the patient's carer placed the cuff on their arm, held the monitor up to the screen to display the results, and adjusted positions so that the patient's blood pressure measurements could be obtained from both sitting and standing positions (Figure 10).

XSL•FO
RenderX

Figure 10. Carer operating the blood pressure monitor.

Remote physical examinations are complex sociotechnical tasks, involving (in our data) at least 3 people, multiple devices at both ends of the call, and a series of instructions and interactions conducted over a video consultation [72]. Hence, while carers were often needed to make physical examinations work, the assistance of a carer did not make them straightforward. In 1 consultation, video largely restricted the examination to a visual inspection. As the carer from Figure 6 reported after the consultation, with an in-person consultation, "[the patient] could probably explain more where it hurts and [the physician] could, you know, feel why it's, you know, still tender." Furthermore, doing the examination while making it visible to the clinician can be challenging for the carer [72]. This was summarized by the carer from Figure 1 as follows:

> *Well, look at me, fannying about just trying to get a picture of your leg. I mean it's not a matchstick. I just could not picture. But it's partly, because I'm holding it and I can't see what I'm looking at.*

Discussion

Principal Findings

Our findings demonstrate that, for some video consultations with some patients, informal carers play an important role in supporting setting up and running a video consultation. While most patients (40/52, 77%) in our data completed the consultation on their own, informal carers were the linchpin that made the video consultation safe and feasible, especially when the patients lacked technological literacy or experienced high frailty. We demonstrated this using recordings and observations to show 3 types of *work* that carers perform. First, they help patients use video technology by setting up everything beforehand and acting as technological support, providing patients with the confidence to even commit to using video. Second, where patients struggle to hear or understand the clinician, carers perform the interactional repair work, repeating or clarifying the clinician's words. Third, where physical assessments are needed, carers can lay hands on the patient's or the clinician's behalf or assist the patient with using the technology (either video technology or examination equipment). Even where patients seemed to manage on their own, patients performed maneuvers that put them at risk of falling, and this was not always clear to the remote clinician. Copresent carers provide an important safety net, making video consultations safe and feasible.

Comparison With Previous Research

There is an extensive body of research on the feasibility and acceptability of video consultations [6-10], which indicates that some patients may need assistance from carers [74]. However, to date, no study has investigated the work informal carers do to support video consultations. A total of 3 health communication studies have used robust methods for analyzing interaction to demonstrate how carers, whether professional or informal, can be involved in a consultation. Two (67%) of these 3 studies documented that nurses and general practitioners play an essential role in making physical examinations work when patients talk to a remote consultant [74,75]. One study on follow-up consultations after surgery showed that informal carers mostly act as bystanders: they remain invisible to the clinician and only occasionally facilitate the consultation [42].

Our study adds to this growing body of literature, demonstrating that informal carers may take a more active role than that of a bystander: in our data, they are attentive to the interaction, moving into and out of the field of view of the camera as needed; performing a range of technical, interactional, and clinical tasks; and taking a more active role depending on the needs of both the patient and clinician. Discrepancies between our findings and previous studies can be accounted for in many ways. First, as both the 3 previous studies and our study are qualitative in nature, they prioritize analytical depth, which mandates small patient samples that are not necessarily representative, and these are prone to bias in recruitment processes. Second, all studies have taken place in different geographical locations, at different points in time (eg, before or during the COVID-19 pandemic), and in different clinical settings, with patients with different sociodemographic backgrounds. While the methods may be transferable, more research is needed to appreciate to what extent the findings transfer.

The important role of carers is not limited to video consulting. For in-person consultations, research has shown that carers can be actively involved, talking about, alongside, or with the patients, to provide clinicians with relevant information [34-36]. Findings from our study extend this, demonstrating not only the other types of work that carers do to support video consultations but also how the technology shapes this work.

Videoconferencing technologies and the visual angle of webcams are designed for one-to-one conversations [39]. These aspects of technology add to the complexity of the interactional dynamics that already exist for triadic consultations (ie, involving a patient, clinician, and informal carer), where

XSL•FO
RenderX

participants may struggle with turn-taking [34,35]. Because of this added complexity, video consultations have a continuously shifting *participation framework* (ie, the roles of patient, clinician, and carer as, for example, an active coparticipant of overhearer) [57], where carers move in and out of a variety of interactional and technical support roles. Depending on the situation (eg, the patient's capacity and willingness to talk on their own behalf), carers may be expected to be more or less active coparticipants during consultations. Being offscreen makes carers less available for the clinician. They are more likely to act as overhearers [42], which can be beneficial in cases where patients wish to interact with the clinician themselves, but it may also be detrimental when patients need more continuous support. Therefore, our findings contribute to not only our appreciation of the important role of carers in the delivery of health care services but also the interactional organization of video consultations. Future research should investigate systematically how the affordances of the technology, particularly the camera's field of view, affect the norms regarding participation, quality of care, and participant satisfaction.

Meaning of the Study

Our findings suggest that when considering the feasibility of video consultations, some important considerations need to be taken into account. Video consulting has often only been considered a suitable service model for patients with technological competence and experience, where the goal of the consultation is expected to be relatively straightforward (eg, sharing test results and routine follow-up). However, our study shows that this unnecessarily limits to application of video for 2 reasons. First, where patients have a lack of experience with or have anxiety around technology, informal carers can help overcome technological or interactional difficulties. Furthermore, they offer reassurance, making patients comfortable with doing a video consultation. Help may not be needed, but where it would be needed, it would be available [76]. Second, where the goal of the consultation is more complex (eg, involves a physical assessment), video can still be an appropriate option if the patient has adequate support. Assessments in a video consultation often require the patient to move the camera around to frame themselves in a way that they are adequately visible to the clinician while performing movements that may be difficult for them to do safely or using devices that they are not familiar with (eg, oximeters). Copresent carers can overcome some of these challenges, for example, by taking care of the camera or laying hands on the patient, where patients are comfortable with that.

Since the outbreak of the COVID-19 pandemic, video consultations have become a more routinely used service model. While many patients and providers are moving back to in-person delivery of (health) care, hybrid service models that involve remote options, including video consultation, are likely to constitute the new normal. However, despite the routinization of video-consulting services, clinicians still have limited evidence on when they are a feasible and safe option. While the literature is growing quickly and many organizations have proposed guidelines, these often ignore the role of informal carers. Further rollout of this new service model needs to consider not only what patients themselves can do but also what informal carers can do. Given the important role that informal carers have in health care management, particularly for certain groups of patients (eg, young children, patients with high frailty, or patients who lack capacity), it is logical to assume that their role can be transferred to video-consulting models. The additional work for carers will have to be weighed against the potential benefits for each specific clinical context and each individual patient.

The importance of carers for making some video consultations work raises important questions for those providing and supporting services. Not all patients will have access to an informal carer, and those who have may not always want a carer to be present during the consultation. A systematic review found that patients are not necessarily as involved during consultations where they are accompanied by a carer, and while most patients say they appreciate having someone with them, they want to be able to decide whether a carer will be present during the consultation [77]. Patients should feel comfortable asking for their carer to leave the room at any point during a consultation. However, this might put an unnecessary burden on the patient. It may be necessary for clinicians to create opportunities to talk to the patient privately.

Strengths and Limitations

Physiotherapy consultations in our data set were conducted during the COVID-19 pandemic, with heart failure and diabetes consultations conducted before the pandemic when video consulting was not a routine service model and few patients, carers, or clinicians had experience with it. Given the uptake and learning around video consultations during the COVID-19 pandemic, it is possible that patients involved in heart failure and cancer consultations needed more support with the technology than they would now. The prepandemic data were also likely to involve early-adopter clinicians who were supportive of video consultations as a new service model. Furthermore, participants in our data set used mainly Skype (Microsoft Corporation) and FaceTime (Apple Inc), whereas video consultations now often take place on dedicated platforms such as Teams (Microsoft Corporation), Attend Anywhere, or AccuRx. Some of these platforms affect the opening of video consultations, with patients expected to join a virtual waiting room before joining the consultation with their clinician. In addition, we focused on the positive experiences of patients and carers, without actively considering whether and when clinical staff are receptive to carer involvement. Despite these limitations, we anticipate that many of our findings are transferable to current video-consulting services. Our use of methods focused on interaction and communication has enabled us to demonstrate in detail the active role that carers played in some video consultations. While the exact role of carers may differ during and after the COVID-19 pandemic, it is highly likely that some patients (eg, older patients, those experiencing frailty, or those with multimorbidity) will continue to need assistance.

To our knowledge, this study is the first to use robust methods for analyzing communication in triadic video consultations (ie, among clinician, patient, and informal carer) across multiple

XSL•FO
RenderX

clinical settings. Doing so has allowed us to show in detail not only that carers play a vital role in making video consulting work but also how they go about doing this. Our work adds to the existing literature by highlighting the interactional complexity of these consultations, demonstrating the sociotechnical nature of the work undertaken by informal carers, and underscoring the importance of focusing on the microlevel organization of consultations where carers are (and are not) involved [46,50]. Our work was exploratory in nature, relying on secondary analysis; future studies could investigate how the presence of carers affects the overall experience of patients and clinical staff with health care services, the patient-carer relationship, and the health outcomes for patients.

Conclusions

Video consulting remains a viable service option but depends on patient access to technology and their ability to use it. While many patients can manage a video consultation on their own, some (continue to) require assistance. In these circumstances, informal carers can play a unique, and often invisible, role in making video consultations work. To date, research and guidelines have not adequately considered the work of informal carers. This urgently needs addressing, not only to support the policy vision of the spread of video-consulting services but also to make visible and enable the informal carers (and the patients and clinicians they support) in this often vital role.

Acknowledgments

The authors are grateful to Marissa Bird and Joe Wherton for commenting on an earlier draft of this paper. The authors would like to thank Joe Wherton, Chrysanthi Papoutsi, Trish Greenhalgh, Christine A'Court, Gita Ramdharry, Anthony Gilbert, and Jackie Walumbe for their support in collecting the data used in this paper. This project was funded by UK Research and Innovation via the Economic and Social Research Council (ES/V010069/1), Wellcome Trust (WT104830MA), National Institute for Health Research (NIHR) Biomedical Research Centre (BRC-1215-20008), and the NIHR under its Research for Patient Benefit (PB-PG-1216-20012) and Policy Research Programme (grant NIHR202067). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to confidentiality and sensitivity of the material but are available from the corresponding author on reasonable request.

Authors' Contributions

LMS was the principal investigator for this study and led data collection and formal analysis and wrote the first draft of the manuscript. SS provided supervision as a co–principal investigator and supported the review and editing of the manuscript. Both authors were involved in all aspects of the study design and funding acquisition and have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Greenhalgh T, Wherton J, Shaw S, Morrison C. Video consultations for covid-19. BMJ 2020 Mar 12;368:m998. [doi: 10.1136/bmj.m998] [Medline: 32165352]
2. Haider Z, Aweid B, Subramanian P, Iranpour F. Telemedicine in orthopaedics during COVID-19 and beyond: a systematic review. J Telemed Telecare 2022 Jul;28(6):391-403 [FREE Full text] [doi: 10.1177/1357633X20938241] [Medline: 32762270]
3. Car J, Koh GC, Foong PS, Wang CJ. Video consultations in primary and specialist care during the covid-19 pandemic and beyond. BMJ 2020 Oct 20;371:m3945. [doi: 10.1136/bmj.m3945] [Medline: 33082127]
4. Gunasekeran DV, Liu Z, Tan WJ, Koh J, Cheong CP, Tan LH, et al. Evaluating safety and efficacy of follow-up for patients with abdominal pain using video consultation (SAVED study): randomized controlled trial. J Med Internet Res 2020 Jun 15;22(6):e17417 [FREE Full text] [doi: 10.2196/17417] [Medline: 32459637]
5. Shaw SE, Seuren LM, Wherton J, Cameron D, A'Court C, Vijayaraghavan S, et al. Video consultations between patients and clinicians in diabetes, cancer, and heart failure services: linguistic ethnographic study of video-mediated interaction. J Med Internet Res 2020 May 11;22(5):e18378 [FREE Full text] [doi: 10.2196/18378] [Medline: 32391799]
6. Donaghy E, Atherton H, Hammersley V, McNeilly H, Bikker A, Robbins L, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Aug 29;69(686):e586-e594 [FREE Full text] [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
7. Chaudhry H, Nadeem S, Mundi R. How satisfied are patients and surgeons with telemedicine in orthopaedic care during the COVID-19 pandemic? A systematic review and meta-analysis. Clin Orthop Relat Res 2021 Jan 01;479(1):47-56 [FREE Full text] [doi: 10.1097/CORR.0000000000001494] [Medline: 33009231]

XSL•FO
RenderX

8. Eichberg DG, Basil GW, Di L, Shah AH, Luther EM, Lu VM, et al. Telemedicine in neurosurgery: lessons learned from a systematic review of the literature for the COVID-19 era and beyond. Neurosurgery 2020 Dec 15;88(1):E1-12 [FREE Full text] [doi: 10.1093/neuros/nyaa306] [Medline: 32687191]
9. Fahey E, Elsheikh MF, Davey MS, Rowan F, Cassidy JT, Cleary MS. Telemedicine in orthopedic surgery: a systematic review of current evidence. Telemed J E Health 2022 May;28(5):613-635. [doi: 10.1089/tmj.2021.0221] [Medline: 34375150]
10. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020 May 1;4(1):bjgpopen20X101020 [FREE Full text] [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
11. Oates EV, Lim GH, Nevins EJ, Kanakala V. Are surgical patients satisfied with remote consultations? A comparison of remote versus conventional outpatient clinic follow-up for surgical patients: a systematic review and meta-analysis of randomized controlled trials. J Patient Exp 2021 Jul 31;8:23743735211035916 [FREE Full text] [doi: 10.1177/23743735211035916] [Medline: 34377778]
12. NHS long term plan. National Health Service. URL: https://www.longtermplan.nhs.uk/ [accessed 2024-05-17]
13. A plan for digital health and social care. Department of Health and Social Care United Kingdom Government. 2022 Jun 29. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2024-05-17]
14. Data saves lives: reshaping health and social care with data. Department of Health and Social Care United Kingdom Government. 2022 Jun 15. URL: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data [accessed 2024-05-17]
15. James HM, Papoutsi C, Wherton J, Greenhalgh T, Shaw SE. Spread, scale-up, and sustainability of video consulting in health care: systematic review and synthesis guided by the NASSS framework. J Med Internet Res 2021 Jan 26;23(1):e23775 [FREE Full text] [doi: 10.2196/23775] [Medline: 33434141]
16. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182 [FREE Full text] [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
17. Armfield NR, Bradford M, Bradford NK. The clinical use of Skype--for which patients, with which problems and in which settings? A snapshot review of the literature. Int J Med Inform 2015 Oct;84(10):737-742. [doi: 10.1016/j.ijmedinf.2015.06.006] [Medline: 26183642]
18. Moisan P, Barimani B, Antoniou J. Orthopedic surgery and telemedicine in times of COVID-19 and beyond: a review. Curr Rev Musculoskelet Med 2021 Apr;14(2):155-159 [FREE Full text] [doi: 10.1007/s12178-021-09693-9] [Medline: 33460020]
19. Dekker A, Bandell DL, Kortlever JT, Schipper IB, Ring D. Factors associated with patient willingness to conduct a remote video musculoskeletal consultation. Arch Bone Jt Surg 2020 Nov;8(6):656-660 [FREE Full text] [Medline: 33313344]
20. Wherton J, Shaw S, Papoutsi C, Seuren L, Greenhalgh T. Guidance on the introduction and use of video consultations during COVID-19: important lessons from qualitative research. BMJ Lead 2020;4(3):120-123. [doi: 10.1136/leader-2020-000262]
21. Bhatta R, Aryal K, Ellingsen G. Opportunities and challenges of a rural-telemedicine program in Nepal. J Nepal Health Res Counc 2015;13(30):149-153. [Medline: 26744201]
22. Moyle W, Jones C, Murfield J, Liu F. 'For me at 90, it's going to be difficult': feasibility of using iPad video-conferencing with older adults in long-term aged care. Aging Ment Health 2020 Feb;24(2):349-352. [doi: 10.1080/13607863.2018.1525605] [Medline: 30621436]
23. Grandizio LC, Foster BK, Klena JC. Telemedicine in hand and upper-extremity surgery. J Hand Surg Am 2020 Mar;45(3):239-242. [doi: 10.1016/j.jhsa.2019.09.007] [Medline: 31718846]
24. Tattersall R, Carty S, Meldrum D, Hardiman O, Murray D. The patient's perspective of remote respiratory assessments during the COVID-19 pandemic. Amyotroph Lateral Scler Frontotemporal Degener 2022 Feb;23(1-2):76-80. [doi: 10.1080/21678421.2021.1920982] [Medline: 34092157]
25. Wittenberg-Lyles E, Oliver DP, Demiris G, Washington KT, Regehr K, Wilder HM. Question asking by family caregivers in hospice interdisciplinary team meetings. Res Gerontol Nurs 2010 Apr;3(2):82-88 [FREE Full text] [doi: 10.3928/19404921-20090731-05] [Medline: 20415357]
26. Wittenberg-Lyles E, Oliver DP, Kruse RL, Demiris G, Gage LA, Wagner K. Family caregiver participation in hospice interdisciplinary team meetings: how does it affect the nature and content of communication? Health Commun 2013;28(2):110-118 [FREE Full text] [doi: 10.1080/10410236.2011.652935] [Medline: 22435889]
27. Why we're here. Carers UK. URL: https://www.carersuk.org/about-us/why-were-here/ [accessed 2023-07-03]
28. State of Caring 2022: a snapshot of unpaid care in the UK. Carers UK. 2022 Nov. URL: https://www.carersuk.org/media/ew5e4swg/cuk_state_of_caring_2022_report.pdf [accessed 2024-05-17]
29. Fenney D, Thorstensen-Woll C, Bottery S. Caring in a complex world: perspectives from unpaid carers and the organisations that support them. The King's Fund. 2023 May 26. URL: https://www.kingsfund.org.uk/insight-and-analysis/reports/unpaid-carers-caring-complex-world [accessed 2024-05-17]

XSL•FO
RenderX

30. National Academies of Sciences, Engineering, and Medicine, National Academies of Sciences, Engineering, and Medicine (U.S.). Committee on Improving the Quality of Health Care Globally, Board on Health Care Services. Crossing the Global Quality Chasm: Improving Health Care Worldwide. Washington, DC: National Academies Press; 2018.
31. Jardien-Baboo S, van Rooyen DR, Ricks EJ, Jordan PJ, Ten Ham-Baloyi W. Integrative literature review of evidence-based patient-centred care guidelines. J Adv Nurs 2021 May;77(5):2155-2165. [doi: 10.1111/jan.14716] [Medline: 33314226]
32. Parmar J, Anderson S, Duggleby W, Holroyd-Leduc J, Pollard C, Brémault-Phillips S. Developing person-centred care competencies for the healthcare workforce to support family caregivers: caregiver centred care. Health Soc Care Community 2021 Sep;29(5):1327-1338. [doi: 10.1111/hsc.13173] [Medline: 32949440]
33. Caregiving in the United States 2020. AARP and National Alliance for Caregiving. 2020 May 14. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving-in-the-united-states.html [accessed 2024-05-17]
34. Stivers T. Negotiating who presents the problem: next speaker selection in pediatric encounters. J Commun 2001 Jun;51(2):252-282 [FREE Full text] [doi: 10.1111/j.1460-2466.2001.tb02880.x]
35. Fioramonte A, Vásquez C. Multi-party talk in the medical encounter: socio-pragmatic functions of family members' contributions in the treatment advice phase. J Pragmat 2019 Jan;139:132-145. [doi: 10.1016/j.pragma.2018.11.001]
36. Halkowski T. Medical discourse. In: Hyland K, Paltridge P, editors. Continuum Companion to Discourse Analysis. London, UK: Continuum Publishers, Ltd; 2011:321-331.
37. Chinn D, Rudall D. Who is asked and who gets to answer the health-care practitioner's questions when patients with intellectual disabilities attend UK general practice health checks with their companions? Health Commun 2021 Apr;36(4):487-496. [doi: 10.1080/10410236.2019.1700440] [Medline: 31842630]
38. Gan Y, Greiffenhagen C, Reeves S. Connecting distributed families: camera work for three-party mobile video calls. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376704]
39. Licoppe C, Morel J. Video-in-interaction: “talking heads” and the multimodal organization of mobile and skype video calls. Res Lang Soc Interact 2012;45(4):399-429. [doi: 10.1080/08351813.2012.724996]
40. Stommel WJ, van Goor H, Stommel MW. The impact of video-mediated communication on closed wound assessments in postoperative consultations: conversation analytical study. J Med Internet Res 2020 May 05;22(5):e17791 [FREE Full text] [doi: 10.2196/17791] [Medline: 32310816]
41. Ilomäki S, Ruusuvuori J. Preserving client autonomy when guiding medicine taking in telehomecare: a conversation analytic case study. Nurs Ethics 2022 May;29(3):719-732 [FREE Full text] [doi: 10.1177/09697330211051004] [Medline: 35119321]
42. Stommel WJ, Stommel MW. Participation of companions in video-mediated medical consultations: a microanalysis. In: Meredith J, Giles D, Stommel W, editors. Analysing Digital Interaction. Cham, Switzerland: Palgrave Macmillan; May 04, 2021:177-203.
43. Tusting K. General introduction. In: The Routledge Handbook of Linguistic Ethnography. Milton Park, UK: Routledge; 2019:1-9.
44. Gumperz JJ, Hymes D. Directions in Sociolinguistics: The Ethnography of Communication. Hoboken, NJ: Wiley; 1986.
45. Maynard DW, Heritage J. Conversation analysis, doctor-patient interaction and medical communication. Med Educ 2005 Apr;39(4):428-435. [doi: 10.1111/j.1365-2929.2005.02111.x] [Medline: 15813766]
46. Seuren LM, Shaw SE. Using linguistic ethnography to study video consultations: a call to action and future research agenda. Qual Health Res 2022 Apr;32(5):800-813 [FREE Full text] [doi: 10.1177/10497323221077297] [Medline: 35245150]
47. Seuren LM, Gilbert A, Ramdharry G, Walumbe J, Shaw SE. Video analysis of communication by physiotherapists and patients in video consultations: a qualitative study using conversation analysis. Physiotherapy 2024 Jun;123:30-37 [FREE Full text] [doi: 10.1016/j.physio.2023.10.002] [Medline: 38262264]
48. Greenhalgh T, Shaw S, Wherton J, Vijayaraghavan S, Morris J, Bhattacharya S, et al. Real-world implementation of video outpatient consultations at macro, meso, and micro levels: mixed-method study. J Med Internet Res 2018 Apr 17;20(4):e150 [FREE Full text] [doi: 10.2196/jmir.9897] [Medline: 29625956]
49. Leydon GM, Barnes RK. Conversation analysis. In: Pope C, Mays N, editors. Qualitative Research in Health Care. Hoboken, NJ: Wiley; 2019:135-150.
50. Halkowski T, Gill VT. Conversation analysis and ethnomethodology: the centrality of interaction. In: The SAGE Handbook of Qualitative Methods in Health Research. Thousand Oaks, CA: SAGE Publications; 2010:212-228.
51. Stivers T, Sidnell J. Introduction. In: The Handbook of Conversation Analysis. Chichester, UK: Blackwell Publishing Ltd; 2012:77-99.
52. Miles MB, Huberman AM, Saldana J. Qualitative Data Analysis: A Methods Sourcebook. Thousand Oaks, CA: SAGE Publications; 2018.
53. Jefferson G. Glossary of transcript symbols with an introduction. In: Lerner GH, editor. Conversation Analysis: Studies from the First Generation. Amsterdam, The Netherlands: John Benjamins Publishing Company; Aug 31, 2004:13-31.
54. Mondada L. Challenges of multimodality: language and the body in social interaction. J Socioling 2016 Jul 11;20(3):336-366. [doi: 10.1111/josl.1_12177]
55. Schegloff EA. The routine as achievement. Hum Stud 1986 Jun;9:111-151. [doi: 10.1007/bf00148124]

XSL•FO
RenderX

56. Schegloff EA. Sequencing in conversational openings. Am Anthropol 1968 Dec;70(6):1075-1095. [doi: 10.1525/aa.1968.70.6.02a00030]
57. Goffman E. Forms of Talk. Philadelphia, PA: University of Pennsylvania Press; 1981.
58. Schegloff EA. Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge, UK: Cambridge University Press; 2007.
59. Beach WA. Transitional regularities for 'casual' "Okay" usages. J Pragmat 1993 Apr;19(4):325-352. [doi: 10.1016/0378-2166(93)90092-4]
60. Dingemanse M, Roberts SG, Baranova J, Blythe J, Drew P, Floyd S, et al. Universal principles in the repair of communication problems. PLoS One 2015 Sep 16;10(9):e0136100 [FREE Full text] [doi: 10.1371/journal.pone.0136100] [Medline: 26375483]
61. Schegloff EA, Jefferson G, Sacks H. The preference for self-correction in the organization of repair in conversation. Language 1977 Jun;53(2):361-382. [doi: 10.2307/413107]
62. Hayashi M, Raymond G, Sidnell J. Conversational repair and human understanding: an introduction. In: Conversational Repair and Human Understanding. Cambridge, UK: Cambridge University Press; 2013:1-40.
63. Robinson JD. Epistemics, action formation, and other-initiation of repair: the case of partial questioning repeats. In: Hayashi M, Raymond G, Sidnell J, editors. Conversational Repair and Human Understanding. Cambridge, UK: Cambridge University Press; 2013:261-292.
64. Drew P. 'Open' class repair initiators in response to sequential sources of troubles in conversation. J Pragmat 1997 Jul;28(1):69-101. [doi: 10.1016/s0378-2166(97)89759-7]
65. Enfield NJ, Dingemanse M, Baranova J, Blythe J, Brown P, Dirksmeyer T, et al. Huh? What? – a first survey in twenty-one languages. In: Hayashi M, Raymond G, Sidnell J, editors. Conversational Repair and Human Understanding. Cambridge, UK: Cambridge University Press; 2013:343-380.
66. McCabe R, Healey PG. Miscommunication in doctor-patient communication. Top Cogn Sci 2018 Apr;10(2):409-424 [FREE Full text] [doi: 10.1111/tops.12337] [Medline: 29749042]
67. Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, et al. Universals and cultural variation in turn-taking in conversation. Proc Natl Acad Sci U S A 2009 Jun 30;106(26):10587-10592 [FREE Full text] [doi: 10.1073/pnas.0903616106] [Medline: 19553212]
68. Seuren LM, Wherton J, Greenhalgh T, Shaw SE. Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. J Pragmat 2021 Jan;172:63-78 [FREE Full text] [doi: 10.1016/j.pragma.2020.11.005] [Medline: 33519050]
69. Fatehi F, Martin-Khan M, Smith AC, Russell AW, Gray LC. Patient satisfaction with video teleconsultation in a virtual diabetes outreach clinic. Diabetes Technol Ther 2015 Jan;17(1):43-48. [doi: 10.1089/dia.2014.0159] [Medline: 25296189]
70. Ignatowicz A, Atherton H, Bernstein CJ, Bryce C, Court R, Sturt J, et al. Internet videoconferencing for patient-clinician consultations in long-term conditions: a review of reviews and applications in line with guidelines and recommendations. Digit Health 2019 Apr 23;5:2055207619845831 [FREE Full text] [doi: 10.1177/2055207619845831] [Medline: 31069105]
71. Gilbert AW, May CR, Brown H, Stokes M, Jones J. A qualitative investigation into the results of a discrete choice experiment and the impact of COVID-19 on patient preferences for virtual consultations. Arch Physiother 2021 Sep 06;11(1):20 [FREE Full text] [doi: 10.1186/s40945-021-00115-0] [Medline: 34488898]
72. Seuren LM, Wherton J, Greenhalgh T, Cameron D, A'Court C, Shaw SE. Physical examinations via video for patients with heart failure: qualitative study using conversation analysis. J Med Internet Res 2020 Feb 20;22(2):e16694 [FREE Full text] [doi: 10.2196/16694] [Medline: 32130133]
73. Clark HH. Depicting as a method of communication. Psychol Rev 2016 Apr;123(3):324-347. [doi: 10.1037/rev0000026] [Medline: 26855255]
74. Cherba M, Grosjean S, Bonneville L, Nahon-Serfaty I, Boileau J, Waldolf R. The essential role of nurses in supporting physical examination in telemedicine: insights from an interaction analysis of postsurgical consultations in orthopedics. Nurs Inq 2022 Apr;29(2):e12452. [doi: 10.1111/nin.12452] [Medline: 34428319]
75. Pappas Y, Seale C. The physical examination in telecardiology and televascular consultations: a study using conversation analysis. Patient Educ Couns 2010 Oct;81(1):113-118. [doi: 10.1016/j.pec.2010.01.005] [Medline: 20144523]
76. Lynch J, Hughes G, Papoutsi C, Wherton J, A'Court C. "It's no good but at least I've always got it round my neck": a postphenomenological analysis of reassurance in assistive technology use by older people. Soc Sci Med 2022 Jan;292:114553 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114553] [Medline: 34799181]
77. Laidsaar-Powell RC, Butow PN, Bu S, Charles C, Gafni A, Lam WW, et al. Physician-patient-companion communication and decision-making: a systematic review of triadic medical consultations. Patient Educ Couns 2013 Apr;91(1):3-13. [doi: 10.1016/j.pec.2012.11.007] [Medline: 23332193]

Abbreviations

OTQS: Oxford Telehealth Qualitative Study
SCiP: Supporting Consultations in Remote Physiotherapy

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 08.08.23; peer-reviewed by P Traulsen, G Gauhe, T Halkowski, S White; comments to author 06.10.23; revised version received 24.11.23; accepted 19.04.24; published 31.05.24.

Please cite as:
Seuren LM, Shaw S
How Informal Carers Support Video Consulting in Physiotherapy, Heart Failure, and Cancer: Qualitative Study Using Linguistic Ethnography
J Med Internet Res 2024;26:e51695
URL: https://www.jmir.org/2024/1/e51695
doi:10.2196/51695
PMID:38819900

©Lucas Martinus Seuren, Sara Shaw. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

eHealth Literacy and the Use of NHS 111 Online Urgent Care Service in England: Cross-Sectional Survey

Joanne Turnbull[1], PhD; Jane Prichard[1], PhD; Jennifer MacLellan[2], PhD; Catherine Pope[2], PhD

[1]School of Health Sciences, University of Southampton, Southampton, United Kingdom

[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Joanne Turnbull, PhD
School of Health Sciences
University of Southampton
Highfield Campus
Southampton, SO17 1BJ
United Kingdom
Phone: 44 2380597940
Email: j.c.turnbull@soton.ac.uk

Abstract

Background: Many health care systems have used digital technologies to support care delivery, a trend amplified by the COVID-19 pandemic. “Digital first” may exacerbate health inequalities due to variations in eHealth literacy. The relationship between eHealth literacy and web-based urgent care service use is unknown.

Objective: This study aims to measure the association between eHealth literacy and the use of NHS (National Health Service) 111 online urgent care service.

Methods: A cross-sectional sequential convenience sample survey was conducted with 2754 adults (October 2020-July 2021) **from** primary, urgent, or emergency care; third sector organizations; and the NHS 111 online website. The survey included the eHealth Literacy Questionnaire (eHLQ), questions about use, preferences for using NHS 111 online, and sociodemographic characteristics.

Results: Across almost all dimensions of the eHLQ, NHS 111 online users had higher mean digital literacy scores than nonusers (P<.001). Four eHLQ dimensions were significant predictors of use, and the most highly significant dimensions were eHLQ1 (using technology to process health information) and eHLQ3 (ability to actively engage with digital services), with odds ratios (ORs) of 1.86 (95% CI 1.46-2.38) and 1.51 (95% CI 1.22-1.88), respectively. Respondents reporting a long-term health condition had lower eHLQ scores. People younger than 25 years (OR 3.24, 95% CI 1.87-5.62) and those with formal qualifications (OR 0.74, 95% CI 0.55-0.99) were more likely to use NHS 111 online. Users and nonusers were likely to use NHS 111 online for a range of symptoms, including chest pain symptoms (n=1743, 70.4%) or for illness in children (n=1117, 79%). The users of NHS 111 online were more likely to have also used other health services, particularly the 111 telephone service (χ_1^2=138.57; P<.001).

Conclusions: These differences in eHealth literacy scores amplify perennial concerns about digital exclusion and access to care for those impacted by intersecting forms of disadvantage, including long-term illness. Although many appear willing to use NHS 111 online for a range of health scenarios, indicating broad acceptability, not all are able or likely to do this. Despite a policy ambition for NHS 111 online to substitute for other services, it appears to be used alongside other urgent care services and thus may not reduce demand.

(J Med Internet Res 2024;26:e50376) doi:10.2196/50376

KEYWORDS

urgent care; digital health; access to health care; eHealth; health care system; COVID-19; urgent; emergency; health literacy; eHealth literacy; digital literacy; access; cross-sectional

XSL•FO
RenderX

Introduction

Background

"Digital first" as the central point of contact is increasingly being pursued in the delivery of a wide range of services, including health care [1]. The COVID-19 pandemic rapidly accelerated the use of apps, web-based digital technologies, and web-based triage in general practice and urgent care internationally [2] and in the United Kingdom [3]. Digital and telephone access are now core to primary [4,5], urgent, and emergency care provision in the UK NHS (National Health Service) [6,7], with a range of telephone and web-based services that triage and manage demand via e-consultation systems [6,8]. These systems typically offer urgent care call handling, web-based triage, and signposting to suitable services (eg, general practice, urgent care centers [UCCs], and emergency departments [EDs]) [9].

Digital health care offers the potential to improve the quality of patient care and provide timely and more convenient access to services [6]. They may also empower people to manage and maintain their own health [10]. Evidence suggests that participants often express high levels of satisfaction with web-based symptom checkers and assessment services [7]. However, important longstanding concerns that remain are socioeconomic and cultural factors [11], language difficulties, disability, and wider structural and technical infrastructure obstacles, which act as barriers to using, and benefitting from, digital access to health care [12]. Studies have shown that people from lower socioeconomic groups are typically less likely to use web-based information seeking [13] and symptom checkers [14]. Black or African American and Hispanic adults in the United States have been shown to be less likely to use technology for health-related purposes [15]. Conversely, younger and more highly educated people were more likely to use web-based triage and symptom checkers [7].

Accessing health care services via digital technologies predicates that people have sufficient knowledge, skills, resources, and motivation to access and use digital technologies to make decisions about a health problem [16]. The concept of eHealth literacy, which combines ideas about "health" and "health service" literacy (appreciation of symptoms and signs of illness and awareness of service provision) with digital literacy (ability to use digital technologies such as the web or smartphones) [16,17], has proved useful for examining this "digital divide" [11,18-20]. Studies using the eHealth literacy scale (eHEALS) [17] have demonstrated that lower eHealth literacy is associated with increased age [21,22], lower levels of education [23], lower socioeconomic status [22], and the presence of a long-term health condition (LTHC) [23]. Much of this literature has focused on eHealth literacy in relation to internet use for health information seeking rather than using symptom checkers or web-based triage. A survey using the eHealth Literacy Questionnaire (eHLQ) [24] reported that users who have digital access to health care services (eg, communicating with health professionals and accessing health-related information) scored higher on most dimensions of the scale [25]. However, little research has focused on eHealth literacy in the context of web-based urgent care triage and assessment.

NHS 111 Online

The NHS 111 online urgent care service was launched in 2017 across the 4 nations of the United Kingdom and is an exemplar of a policy push to "digital first" that is not unique to the NHS. NHS 111 online was designed to augment the NHS 111 telephone triage and assessment service, which was launched in England in 2011 [3]. NHS 111 online is freely available 24 hours a day, giving access to web-based assessment and triage (via a smartphone, tablet, or computer) for people with urgent (nonemergency) care needs aged older than 5 years. The NHS 111 telephone and online services are both underpinned by a computer decision support software system. NHS 111 online users follow a tailored algorithm, answering questions about symptoms or health concerns. This results in an outcome that directs users to appropriate services (eg, emergency ambulance, ED, and general practice) or provides self-care advice. Where indicated, a call back from a health care professional may be offered. Facilities for booking arrival at an ED were more recently added [26]. In a single month (April 2024), 661,987 NHS 111 online sessions were completed: 64,754 (10%) resulted in an ambulance outcome; 73,366 (11%) emergency treatment; 283,808 (43%) primary care; 102,182 (15%), a prescription; 39,622 (6%) dental care; and 46,572 (7%) another service. Only 51,683 (8%) of calls resulted in self-management [27]. There is some expectation that NHS 111 online may help reduce or ameliorate demand for face-to-face urgent and emergency care services [3], but there is some evidence to suggest that NHS 111 online had little impact on the number of calls to the NHS 111 telephone service [28].

There is little research to date about eHealth literacy and the use of web-based triage and assessment urgent care services. Since NHS 111 online is used directly by patients and the public—without a call handler or clinical intermediary—this raises additional concerns about eHealth literacy and equity of access via digital services. It is unclear whether the potential benefits of urgent web-based health services, such as improving access to services and greater empowerment or self-management of own health [6,10], may be hindered by eHealth literacy. This study provides the first large-scale survey that aims to measure eHealth literacy and the help-seeking preferences of users and nonusers of NHS 111 online in the context of web-based urgent care use.

Methods

Study Design

A cross-sectional survey was conducted in England between October 2020 and July 2021, including periods when COVID-19 restrictions were in place. The survey included eHLQ, a validated 35-item 7-dimension questionnaire [28] used to explore individuals' reported competencies, experiences, and interactions with technologies and services. The eHLQ consists of 7 dimensions: eHLQ1, using technology to process health information (5 items); eHLQ2, understanding of health concepts and language 5 items); eHLQ3, ability to actively engage with digital services (5 items); eHLQ4, feel safe and in control (5

XSL•FO
RenderX

items); eHLQ5, motivated to engage with digital services (5 items); eHLQ6, access to digital services that work (6 items); and eHLQ7, digital services that suit individual needs (4 items). The eHLQ was developed simultaneously in Danish and English using classical and modern test theory [28]. The instrument has been used in a range of countries and health care settings. Since its development, there have been several translations and cultural adaptations, and research indicates that the instrument is robust across a range of health care contexts [29-31].

The eHLQ is scored using a 4-point ordinal scale, from strongly disagree (1) to strongly agree (4). Each dimension contains between 4 and 6 items, with scores averaged to calculate each dimension. A higher mean score indicates a higher self-reported eHealth literacy score (a scale of 1 to 4). The highest score of 4 indicates individuals' self-reported positive experiences and self-reported strengths and comfort with using digitized health services. The eHLQ does not include cut-off points or a benchmarking score for high or low eHealth literacy levels.

The survey also included questions about age, gender, educational attainment, employment status, and the presence of an LTHC. Respondents were asked if they had "a long-term condition or chronic disease" (eg, diabetes, chronic obstructive pulmonary disease, arthritis, and hypertension). As such, having an LTHC is defined by the respondents themselves. Educational attainment was aggregated to four levels: (1) no formal qualifications; (2) comprehensive or secondary school education equivalent—International Standard Classification of Education 2011 (ISCED-2011) levels 1 and 2; (3) further (short) education equivalent to ISCED levels 3, 4, and 5; and (4) higher education (medium and long equivalent to ISCED levels 6, 7, and 8) [32].

Additionally, 10 scenarios describing common presenting conditions or urgent care needs were used to explore preferences for using NHS 111 online. These scenarios were informed by data from our previous research [33] and developed in consultation with NHS Digital and patient and public representatives. Scenario preferences for using NHS 111 online were rated on a 5-point Likert scale from "very likely" to "very unlikely." Respondents were also asked if they had ever previously used an urgent and emergency service (NHS 111 online, NHS 111 telephone service, UCC, general practice out-of-hours service, 999 emergency ambulance, and ED).

Survey Sampling and Participants

Nonprobability sequential convenience sampling was a pragmatic choice to access people who had and had not previously used NHS 111 online. The sampling and recruitment strategy meant that it was not possible to calculate a response rate. Respondents were recruited via 24 primary care organizations, 7 urgent or emergency care settings, the NHS 111 online website, and 2 non-NHS third sector (charity) organizations. The small number of respondents from the charity sites (n=5 respondents) have been combined with primary care data in the analysis presented here. Potential respondents (aged 18 years or older) were identified sequentially by administrative or clinical staff at participating sites or organizations (eg, by reception staff at EDs, or general practice surgeries). General practices used an SMS text message mail out of the survey to eligible patients registered at their practice who had agreed to receive practice information via text message. Practices were asked to select a minimum of 100 random patients on their practice list that had consented to SMS mail outs. Some practices chose to sample more patients than 100 to increase recruitment numbers (practices sampled between an additional 1 and 135 patients per practice). EDs and UCCs invited attendees to their services to take part either by providing them with a web-based link to the survey or by offering the opportunity to complete the survey on a computer tablet in the waiting room (assisted, if necessary, by a research nurse). Sequential patients were offered a survey until a minimum of 50 participants had been recruited at each site. Patients in England who completed the NHS 111 online triage were offered a tailored hyperlink to complete the survey. Of 2754 valid responses, 1621 (58.9%) were recruited via primary care and charity settings, 626 (22.7%) through ED and UCC, and 507 (18.4%) via NHS 111 online.

Patient and Public Involvement

A patient and public involvement (PPI) representative was on the project team and the study steering group and contributed to the design of the study and interpretation of the results. Additional PPI representatives (homeless health peer advocates of the charity Groundswell and members of the public from the Deep End Sheffield cluster PPI Group) took part in PPI events throughout the project, contributing to decisions about survey recruitment, helping to develop the scenario questions, and discussing the interpretation of results and how best to present information from the study for public audiences.

Data Analysis

The analysis compared those who had previously used NHS 111 online at least once (users) and those who had not (nonusers). Descriptive categorical data are summarized and presented as frequency counts and percentages. Chi-square analysis was used to compare users and nonusers and whether they had ever used other urgent and emergency services and the likelihood of using NHS 111 online for the 10 health scenarios with previous use or nonuse of NHS 111 online. We created a binary variable of "likely" or "not likely" by removing the small number of neutral responses. Neutral responses accounted for 8%-16% of the data depending on the scenario. While this grouping loses some of the details of responses, it facilitates comparison. Analyses of the difference between users and nonusers were conducted using Bonferroni adjusted α levels of .007 per test (.05/7). Effect sizes are reported due to the large sample size (Φ correlation coefficient).

A secondary analysis was performed to assess the effects of age, gender, education, and use of NHS 111 online on eHealth literacy scores. Continuous data are presented as means (SDs). When comparing a continuous variable between 2 groups, 2-tailed *t* tests were applied. The mean eHLQ score for each dimension was compared for users and nonusers of NHS 111 online. Analysis of the difference in eHLQ scores was conducted using Bonferroni adjusted α levels of .007 per test (0.05/7). Due to the large sample size, effect sizes are reported (Hedges *g*).

Logistic regression was used to extend the univariate analysis outlined above to explore use versus nonuse of NHS 111 online. Logistic regression reports odds ratios (ORs) associated with

XSL•FO
RenderX

each predictor value. The "enter" method (where all variables are entered into the model) was chosen so that all the chosen variables were entered into the model in a single step. Education was aggregated into a binary variable in the regression analysis since there was no strong association between eHLQ and education level (except that people with no formal qualifications had lower eHLQ scores compared to people with any level of qualification). The logistic regression model was examined for multicollinearity by examining tolerance, variance inflation factor, and variance of proportions.

We included respondents with incomplete data. Data for each analysis included all available values (case-by-case). In calculating the eHLQ dimensions, where more than 50% of the data were missing, a score was not calculated for that dimension and was excluded from the analysis.

Ethical Considerations

This study involves human participants and ethical approval was granted for the study by the London Stanmore Research Ethics Authority (20/ LO/0294). Participants gave informed consent to participate in the study before taking part.

Results

Characteristics of Users and Nonusers of NHS 111 Online

Of 2754 valid respondents, 1617 (58.7%) had previously used NHS 111 online ("users") and 1137 (41.3%) had not used NHS 111 online ("nonusers"). In total, 1745 (63.5%) of respondents were female, 1195 (43.6%) were aged between 45 and 64 years, and 1197 (44.2%) reported an LTHC (Table 1). More female participants reported using NHS 111 online, and the proportion of NHS 111 online users declined consistently with each increasing age and increased with the reported level of education. In total, there is a small difference in the proportion of people with a long-term or chronic condition who had used NHS 111 online compared to those who had not, 523 (46.7%) and 674 (42.4%), respectively.

Table 1. Characteristics of respondents by previous use of NHS 111 online (n=2754).

Characteristics	Nonuser, n (%)	User, n (%)	Total, n (%)
Sex[a]			
Female	1000 (62)	745 (65.7)	1745 (63.5)
Male	606 (37.6)	373 (32.9)	979 (35.7)
Nonbinary or prefer not to say	6 (0.4)	16 (1.4)	22 (0.8)
Age group (years)			
18-24	55 (3.4)	104 (9.2)	159 (5.8)
25-34	149 (9.3)	182 (16)	331 (12.1)
35-44	245 (15.2)	203 (17.9)	448 (16.3)
45-54	328 (20.4)	240 (21.2)	568 (20.7)
55-64	405 (25.2)	222 (19.6)	627 (22.9)
65-74	321 (19)	139 (12.2)	460 (16.8)
≥75	106 (6.6)	44 (3.9)	150 (5.5)
Educational level			
No formal qualifications	233 (14.7)	83 (7.4)	316 (11.7)
Comprehensive school or General Certificate of Secondary Education or equivalent	327 (20.7)	213 (19)	540 (20)
Further education	420 (26.6)	334 (29.7)	754 (27.9)
Higher education (degree or higher)	601 (38)	493 (43.9)	1094 (40.5)
Long-term health condition[b]			
Yes	674 (42.4)	523 (46.7)	1197 (44.2)
No	916 (57.6)	597 (53.3)	1513 (55.8)

[a]Sex: significant difference between male and female (χ_1^2=5.46; *P*=.02; Φ=0.05).

[b]Long-term health condition: significant difference between yes and no (χ_1^2=4.94; *P*=.03; Φ=0.04).

eHealth Literacy

Across almost all dimensions of the eHLQ, as might be expected, NHS 111 online users had higher eHealth literacy (Table 2). Significant differences were observed for all dimensions except eHLQ4 (feel safe and in control) and eHLQ6 (access to digital services that work). Effect size calculations revealed that differences between users and nonusers were largest for the dimensions of eHLQ1 (using technology to process health information), eHLQ3 (ability to actively engage with digital services), and eHLQ5 (motivated to engage with digital services).

XSL•FO
RenderX

Respondents who reported having an LTHC tended to have lower eHLQ scores on some dimensions and yet were also more likely to have used NHS 111 online (Table 1). Further analysis identified that the subset of people with an LTHC who were nonusers of NHS 111 had the lowest eHLQ mean score for each dimension (Table 3). This difference was statistically significant for 5 dimensions when compared to users both with and without an LTHC but was not significant for eHLQ4 (feel safe and in control) and eHLQ6 (access to digital services that work).

Table 2. eHLQ[a] dimensions by previous use of NHS 111 online.

Dimensions	Mean score (SD)	Mean difference (95% CI)	*t* value (*df*)	*P* value	Hedges *g*
eHLQ1: using technology to process health information		–0.22 (–0.26 to –0.17)	–9.00 (2677)	<.001	–0.35
Nonuser (n=1565)	2.69 (0.62)				
User (n=1114)	2.91 (0.60)				
eHLQ2: understanding of health concepts and language		–0.10 (–0.14 to –0.06)	–5.00 (2703)	<.001	–0.20
Nonuser (n=1584)	2.96 (0.50)				
User (n=1121)	3.06 (0.53)				
eHLQ3: ability to actively engage with digital services		–0.24 (–0.29 to –0.19)	–9.70 (2712)	<.001	–0.38
Nonuser (n=1590)	2.91 (0.67)				
User (n=1124)	3.15 (0.62)				
eHLQ4: feel safe and in control		0.003 (–0.04 to 0.05)	0.14 (2686)	.89	0.01
Nonuser (n=1574)	2.96 (0.58)				
User (n=1114)	2.95 (0.66)				
eHLQ5: motivated to engage with digital services		–0.17 (0.03 to –0.22)	–6.80 (2660)	<.001	–0.27
Nonuser (n=1552)	2.51 (0.62)				
User (n=1110)	2.68 (0.63)				
eHLQ6: access to digital services that work		–0.04 (0.02 to –0.09)	–1.82 (2728)	.068	–0.07
Nonuser (n=1598)	2.71 (0.57)				
User (n=1132)	2.75 (0.62)				
eHLQ7: digital services that suit individual needs		–0.09 (–0.15 to –0.04)	–3.50 (2668)	<.001	–0.09
Nonuser (n=1559)	2.48 (0.67)				
User (n=1111)	2.57 (0.70)				

[a]eHLQ: eHealth Literacy Questionnaire.

XSL•FO
RenderX

Table 3. eHLQ[a] mean, SD, and mean difference for each dimension grouped by self-reported long-term health condition (yes or no) and NHS 111 online (nonuser or user).

Dimension	Mean score (SD)	Mean difference[b]		
		Nonuser (no)	User (yes)	User (no)
eHLQ1				
Nonuser (yes)	2.63 (0.65)	*–0.11*[c]	*–0.24*	*–0.31*
Nonuser (no)	2.74 (0.60)		*–0.14*	*–0.21*
User (yes)	2.87 (0.62)			–0.07
User (no)	2.94 (0.57)			
eHLQ2				
Nonuser (yes)	2.94 (0.51)	–0.03	*–0.11*	*–0.12*
Nonuser (no)	2.97 (0.48)		*–0.79*	*–0.09*
User (yes)	3.05 (0.54)			–0.01
User (no)	3.07 (0.52)			
eHLQ3				
Nonuser (yes)	2.81 (0.72)	*–0.17*	*–0.29*	*–0.39*
Nonuser (no)	2.98 (0.61)		*–0.12*	*–0.22*
User (yes)	3.10 (0.66)			*–0.11*
User (no)	3.21 (0.58)			
eHLQ4				
Nonuser (yes)	2.96 (0.59)	0.01	0.02	–0.01
Nonuser (no)	2.95 (0.57)		0.02	–0.02
User (yes)	2.94 (0.67)			–0.03
User (no)	2.97 (0.65)			
eHLQ5				
Nonuser (yes)	2.47 (0.64)	–0.07	*–0.18*	*–0.24*
Nonuser (no)	2.54 (0.61)		*–0.10*	*–0.17*
User (yes)	2.64 (0.64)			–0.07
User (No)	2.71 (0.62)			
eHLQ6				
Nonuser (yes)	2.71 (0.58)	–0.01	–0.02	–0.07
Nonuser (no)	2.72 (0.56)		–0.01	–0.06
User (yes)	2.73 (0.63)			–0.05
User (no)	2.78 (0.61)			
eHLQ7				
Nonuser (yes)	2.41 (0.68)	*–0.12*	–0.08	*–0.24*
Nonuser (no)	2.53 (0.65		0.04	*–0.11*
User (yes)	2.49 (0.72)			*–0.15*
User (no)	2.65 (0.67)			

[a]eHLQ: eHealth Literacy Questionnaire.

[b]Grouped by self-reported long-term health condition (yes or no).

[c]Italic formatting indicates significant differences between groups (P<.001).

XSL•FO
RenderX

Use of Other Services In Addition to NHS 111 Online

The use of NHS 111 online is associated with increased previous use of other urgent and emergency services (Table 4; ie, if the respondent had ever used other urgent and emergency services). Notably, NHS 111 online users were likely to have also used the 111 telephone service.

Table 4. Previous use of urgent and emergency services by previous use of NHS 111 online.

Use of a service	Nonusers, n (%)	Users, n (%)	Chi-square (*df*=1)	*P* value	Φ
NHS 111 telephone			138.57	<.001	0.22
No (n=1211)	862 (53.3)	349 (30.7)			
Yes (n=1543)	755 (46.7)	788 (69.3)			
Urgent care center			90.63	<.001	0.18
No (n=1768)	1156 (71.5)	612 (53.8)			
Yes (n=986)	461 (28.5)	525 (46.2)			
General practice out-of-hours services			86.08	<.001	0.18
No (n=1941)	1249 (77.2)	692 (60.9)			
Yes (n=813)	368 (22.8)	445 (39.1)			
999 ambulance service			23.12	<.001	0.09
No (n=1835)	1136 (70.3)	699 (61.5)			
Yes (n=919)	481 (29.7)	438 (38.5)			
Emergency department			33.35	<.001	0.11
No (n=1428)	913 (56.6)	515 (45.3)			
Yes (n=1326)	704 (43.5)	622 (54.7)			

Scenarios Where NHS 111 Online Would Be Considered

There were 2 scenarios for which both users and nonusers reported they were especially likely to use NHS 111 online (Table 5); “young child with a temperature and crying” and “severe chest pain that goes away after a few minutes.” A sizeable proportion of nonusers reported that they might use NHS 111 online for seeking advice about young children (n=1117, 76.2%) or severe chest pain (n=1008, 69.3%). Nearly half of the nonusers also reported that they would be likely to use it for an itchy bite or sting (n=591, 42.5%), pain when urinating (n=696, 50.9%), and a headache for several hours (n=577, 43.3%).

XSL•FO
RenderX

Table 5. Likelihood of using NHS 111 online for different health scenarios.

Health scenarios and responses	Nonusers, n (%)	Users, n (%)	Chi-square (*df*=1)	*P* value	Φ
Itchy bite or sting			19.48	<.001	–0.09
Likely (n=1097)	591 (42.5)	506 (51.6)			
Unlikely (n=1275)	801 (57.5)	474 (48.4)			
Young child's temperature or crying			16.51	<.001	–0.08
Likely (n=1970)	1117 (76.2)	853 (83)			
Unlikely (n=523)	348 (23.8)	175 (17)			
Cough, cold, and sore throat			1.25	.26	–0.02
Likely (n=810)	462 (34.2)	348 (36.5)			
Unlikely (n=1494)	888 (65.8)	606 (63.5)			
Diarrhea or vomiting			3.87	.05	–0.04
Likely (n=972)	561 (40.8)	411 (44.9)			
Unlikely (n=1319)	815 (59.2)	504 (55.1)			
Scalded hand			12.54	<.001	–0.07
Likely (n=738)	399 (28.)	339 (34.8)			
Unlikely (n=1658)	1024 (72)	634 (65.2)			
Painful urinating			31.42	<.001	–0.12
Likely (n=1294)	696 (50.9)	598 (62.6)			
Unlikely (n=1029)	672 (49.1)	357 (37.4)			
Toothache >24 hours			4.71	.03	–0.05
Likely (n=894)	504 (36.3)	390 (40.7)			
Unlikely (n=1454)	886 (63.7)	568 (59.3)			
Headache for several hours			3.35	.06	–0.04
Likely (n=1014)	577 (43.1)	437 (46.9)			
Unlikely (n=1257)	763 (56.9)	494 (53.1)			
Tearful, not sleeping			4.56	.03	–0.04
Likely (n=620)	343 (24.9)	277 (28.9)			
Unlikely (n=1713)	1032 (75.1)	681 (71.1)			
Severe chest pain that subsides			2.15	.14	–0.03
Likely (n=1743)	1008 (69.3)	735 (72.1)			
Unlikely (n=731)	446 (30.7)	285 (27.9)			

Predicting Who Will Use NHS 111 Online

Logistic regression was used to predict the use (vs nonuse) of NHS 111 online for categorical variables such as age, gender, education, and LTHC and the mean scores for the 7 eHLQ continuous variables (Table 6). In the regression model, the reference group for age is the oldest group (≥75 years). For other variables, being female, no LTHC, and any qualification were the reference groups. Multicollinearity was tested in the model examining tolerance, the inverse of the tolerance, collinearity diagnostics, and the variance of proportions. Multicollinearity of greater than 0.5 occurred between dimensions 1 and 5 (0.63 for dimension 1 and 0.49 for dimension 5). Removing dimension 5 from the model improved the model fit slightly. Dimension 4 did not behave like the other dimension (there was little difference in this dimension between age, education, and LTHC), and so it was also removed from the model, providing a very small improvement in model fit. A total of 2534 respondents were included in the regression analysis, with 220 (8%) missing data either on at least 1 sociodemographic variable or eHLQ mean score. The model included 1055 respondents who had used NHS 111 online.

Age was a predictor of using NHS 111 online; people younger than 25 years (OR 3.24, 95% CI 1.87-5.62) and aged between 25 and 44 years (OR 2.35, 95% CI 1.47-3.75) were most likely to have used NHS 111 online. Although more women reported use of NHS 111 online, gender was not a significant predictor in the regression model. Education level was not a strong predictor of use, although those with formal qualifications were, perhaps unsurprisingly, more likely to report using NHS 111

XSL•FO
RenderX

online (95% CI 0.55-0.99). Respondents reporting LTHC had lower eHLQ scores and a subset of nonusers with an LTHC had the lowest eHLQ scores. Four eHLQ dimensions (eHLQ1, eHLQ2, eHLQ3, and eHLQ6) were significant predictors of NHS 111 online use, and most highly significant were dimensions eHLQ1 (using technology to process health information) and eHLQ3 (the ability to actively engage with digital services), with ORs of 1.86 (95% CI 1.46-2.38) and 1.51 (95% CI 1.22-1.88), respectively.

Table 6. Odds ratios for the likelihood of previous NHS 111 online use.

Characteristics	NHS 111 online users (n=1055)	Odds ratio (95% CI)	*P* value
Age (years)			
18-24 (n=132)	84	3.24 (1.87-5.62)	<.001
25-34 (n=304)	174	2.35 (1.47-3.75)	<.001
35-44 (n=418)	190	1.60 (1.02-2.49)	.04
45-54 (n=533)	229	1.44 (0.93-2.21)	.10
55-64 (n=583)	208	1.09 (0.71-1.67)	.69
65-74 (n=428)	130	0.90 (0.57-1.39)	.62
≥75 (n=136)	40	1.0	
Sex			.24
Male (n=909)	348	0.90 (0.76-1.07)	
Female (n=1625)	707	1.0	
Education			.04
No qualifications (n=291)	77	0.74 (0.55-0.99)	
Any qualifications (n=2243)	978	1.0	
LTHC[a]			<.001
Yes (n=1124)	495	1.61 (1.35-1.93)	
No (n=1410)	560	1.00	
eHLQ[b]			
1 (n=2534)	1055	1.86 (1.46-2.38)	<.001
2 (n=2534)	1055	0.77 (0.60-0.99)	.04
3 (n=2534)	1055	1.51 (1.22-1.88)	<.001
6 (n=2534)	1055	0.78 (0.61-1.00)	.05
7 (n=2534)	1055	0.80 (0.64-1.00)	.06

[a]LTHC: long-term health condition.

[b]eHLQ: eHealth Literacy Questionnaire.

Discussion

Principal Results and Comparison With Prior Work

Our findings are consistent with previous research, which shows that women [12,14,34] and younger people are more likely to use digital health services and that people with no formal qualifications are less likely to use NHS 111 online [11,12,35,36]. To our knowledge, this is the first time the eHLQ has been used to examine eHealth literacy in relation to the use of an urgent web-based health service (NHS 111 online). Despite relying on web-based data collection methods in some of our settings (due to COVID-19), we found clear differences in reported eHealth literacy between users and nonusers of NHS 111 online. This finding suggests that the digital divide may be even greater than our data indicate. Similar significant differences have been reported in other studies of users and nonusers of technologies, for example, in medical outpatients using the eHLQ [25] and the eHEALS instrument in a population of baby boomers and older adults seeking health information on the internet [34]. These eHealth literacy differences highlight the potential for digital exclusion and widening of health inequalities and warrant further investigation.

The survey showed that respondents who had an LTHC appeared more likely to use NHS 111 online compared to those without an LTHC. This is consistent with previous research [36] and might be used to argue that NHS 111 online is meeting a need for this group. However, our findings are more nuanced; respondents who reported having an LTHC tended to have lower eHLQ scores and the subset of respondents with very low eHLQ scores who reported having an LTHC had not used NHS 111 online. This apparent digital exclusion may be a cause for concern and a source of inequitable service provision.

XSL•FO
RenderX

People who currently use NHS 111 online appear to concurrently use a range of other urgent and emergency services. This may suggest that the web-based service is not a substitute for other services and does not seem to offer an alternative but an addition to the 111 telephone service in help seeking for urgent care. It is important to note that the survey question asked if respondents had *ever* used other urgent and emergency services, so we do not know if multiple services are used within a single episode of care (eg, using NHS 111 online in addition to other services such as NHS 111 telephone services), or whether different services are used at different time points for different reasons. The value and health benefits of NHS 111 online as an additional service are unclear but, given that one of the key functions of the service is to refer to and signpost to other services, it seems unlikely that NHS 111 online will reduce demand for other urgent and emergency care services.

Our survey showed that people would consider using NHS 111 online for a range of symptom presentations. It was worrying that significant numbers reported they might use NHS 111 online for potentially more serious chest pain symptoms. We asked 2 PPI groups to reflect on this finding and they suggested that the now ubiquitous use of internet searching might underlie this, ie, people experiencing a symptom for the first time would "Google it." The use of NHS 111 online for help seeking about illness in children may be similarly problematic as this service is not intended use children younger than 5 years. The use of NHS 111 online for potentially more serious symptom scenarios or younger children may introduce unnecessary delays in getting help. More targeted information to clarify the scope of NHS 111 online and encourage greater awareness of appropriate use is necessary.

Strengths and Limitations

This large cross-sectional survey is the first to report on the eHealth literacy of people using and not using urgent care triage and assess technology (NHS 111 online). We acknowledge the limitations of eHealth health literacy measures [37,38] and the problem of using self-reports to assess eHealth literacy, but the eHLQ has shown high construct validity, discriminant validity, and scale reliability [24,29,30]. The requirement to report the 7 dimensions separately adds analytical complexity compared to other measures, which offer a single digital literacy score (such as the eHEALS instrument) [16]. Our pragmatic recruitment strategy (designed to capture users and nonusers of NHS 111 online) meant that we were unable to calculate or estimate a response rate. Survey data collection was conducted primarily via the internet resulting inevitably in some bias toward digital literacy in our sample. Some population groups (such as older adults and people with very low educational attainment) may be underrepresented. Recruitment via general practices via text mail excluded those without access to text and those who had not consented to receiving text messages; again, this may disproportionally reduce the responses from some groups (eg, older people). Nonetheless, we have demonstrated differences in reported eHealth literacy and we contend that these are likely to underreport the digital divide, given that people with the lowest literacy and greatest barriers to access to digital technologies were less well represented in the survey.

Our data were collected from across England, including areas of deprivation and high health need. The survey took place during the COVID-19 pandemic 2020-2021 and health services will adjust coming out of the pandemic; however, NHS 111 online remains a core component of urgent care provision and demand management.

Conclusions and Future Research

Our findings about eHealth literacy and use of NHS 111 online may not be surprising; younger and more educated people are more digitally literate and may be expected to be better able to use this urgent care service. However, we have identified important differences in reported eHealth literacy between users and nonusers of NHS 111 online, notably for those with LTHCs. Going forward, the NHS must ensure that "digital first" policies do not entrench or exacerbate health inequalities.

One of the hopes for NHS 111 online was that it would substitute for other services, such as telephone or face-to-face urgent and emergency care [3]. Our survey shows that NHS 111 online users were more likely to have used other NHS urgent and emergency care services in addition to using NHS 111 online, and they had higher cumulative use across these services compared to nonusers. The implications of this, both in terms of health outcomes and service costs, warrant further investigation.

Our survey also suggests that people who have not previously used NHS 111 online appear likely to consider using it for a wide range of health scenarios. Understanding this reservoir of demand and their eHealth literacy will be important as web-based services continue to develop.

Acknowledgments

This study was funded by the National Institute for Health Research (NIHR) Health and Social Care Delivery Research (HS&DR) Programme (127590) and will be published in full in HS&DR. This paper reports independent research commissioned by the NIHR. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicts of Interest

CP is a National Institute for Health Research board member but was not on the board that commissioned this project. None declared for all other authors.

XSL•FO
RenderX

References

1. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2023-11-18]
2. Gunasekeran DV, Tham YC, Ting DSW, Tan GSW, Wong TY. Digital health during COVID-19: lessons from operationalising new models of care in ophthalmology. Lancet Digit Health 2021 Feb;3(2):e124-e134 [FREE Full text] [doi: 10.1016/S2589-7500(20)30287-9] [Medline: 33509383]
3. NHS 111 online. NHS Digital. URL: https://digital.nhs.uk/services/nhs-111-online [accessed 2023-05-30]
4. Paddison CAM, McGill I. Digital primary care: improving access for all? Nuffield Trust 2022:1-30 [FREE Full text]
5. Bryce C, O'Connell MD, Dale J, Underwood M, Atherton H. Online and telephone access to general practice: a cross-sectional patient survey. BJGP Open 2021 Aug;5(4):BJGPO.2020.0179 [FREE Full text] [doi: 10.3399/BJGPO.2020.0179] [Medline: 33910917]
6. The long term plan. NHS England. 2019. URL: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf [accessed 2023-11-18]
7. Cantrell A, Johnson M, Preston L, Baxter SK, Booth A, Turner J. Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review. Health Serv Deliv Res 2019;7(29):1-116. [doi: 10.3310/hsdr07290] [Medline: 31433612]
8. Next steps on the NHS five year forward review. NHS England. 2017. URL: https://www.england.nhs.uk/wp-content/uploads/2017/03/NEXT-STEPS-ON-THE-NHS-FIVE-YEAR-FORWARD-VIEW.pdf [accessed 2023-11-18]
9. Baird B, Maguire D. Understanding factors that enabled digital service change in general practice during the COVID-19 pandemic. The King's Fund. 2021. URL: https://www.kingsfund.org.uk/publications/digital-service-change-general-practice-during-covid-19 [accessed 2023-05-30]
10. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: How to proceed? Front Med (Lausanne) 2022;9:1040926 [FREE Full text] [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
11. Estacio EV, Whittle R, Protheroe J. The digital divide: examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019 Oct;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]
12. Alvarez-Galvez J, Salinas-Perez JA, Montagni I, Salvador-Carulla L. The persistence of digital divides in the use of health information: a comparative study in 28 European countries. Int J Public Health 2020 Apr;65(3):325-333. [doi: 10.1007/s00038-020-01363-w] [Medline: 32277246]
13. Guo Z, Zhao SZ, Guo N, Wu Y, Weng X, Wong JYH, et al. Socioeconomic disparities in eHealth Literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021 Apr 14;23(4):e24577 [FREE Full text] [doi: 10.2196/24577] [Medline: 33784240]
14. Kopka M, Scatturin L, Napierala H, Fürstenau D, Feufel MA, Balzer F, et al. Characteristics of users and nonusers of symptom checkers in Germany: cross-sectional survey study. J Med Internet Res 2023 Jun 20;25:e46231 [FREE Full text] [doi: 10.2196/46231] [Medline: 37338970]
15. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 09;59(1):6-14 [FREE Full text] [doi: 10.1093/geront/gny138] [Medline: 30452660]
16. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
17. Osborne R, Kayser L. Skills and characteristics of the e-health literate patient. BMJ 2018 Apr 16;361:k1656. [doi: 10.1136/bmj.k1656] [Medline: 29661943]
18. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
19. Gray DM, Joseph JJ, Olayiwola JN. Strategies for digital care of vulnerable patients in a COVID-19 world-keeping in touch. JAMA Health Forum 2020 Jun 01;1(6):e200734 [FREE Full text] [doi: 10.1001/jamahealthforum.2020.0734] [Medline: 36218529]
20. Kaihlanen AM, Virtanen L, Buchert U, Safarov N, Valkonen P, Hietapakka L, et al. Towards digital health equity - a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188 [FREE Full text] [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
21. Wong DKK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a Cross-sectional survey. J Med Internet Res 2019 Mar 27;21(3):e10831 [FREE Full text] [doi: 10.2196/10831] [Medline: 30916666]
22. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 02;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
23. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]

XSL•FO
RenderX

24. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
25. Holt KA, Karnoe A, Overgaard D, Nielsen SE, Kayser L, Røder ME, et al. Differences in the level of electronic health literacy between users and nonusers of digital health services: an exploratory survey of a group of medical outpatients. Interact J Med Res 2019 Apr 05;8(2):e8423 [FREE Full text] [doi: 10.2196/ijmr.8423] [Medline: 30950809]
26. MacLellan J, Turnbull J, Prichard J, Pope C. Emergency department staff views of NHS 111 First: qualitative interview study in England. Emerg Med J 2023 Sep;40(9):636-640 [FREE Full text] [doi: 10.1136/emermed-2022-212947] [Medline: 37414462]
27. NHS 111 online statistics. NHS England. URL: https://www.england.nhs.uk/statistics/statistical-work-areas/iucadc-new-from-april-2021/111-online-statistics/ [accessed 2023-11-18]
28. Simpson RM, Jacques RM, Nicholl J, Stone T, Turner J. Measuring the impact introducing NHS 111 online had on the NHS 111 telephone service and the wider NHS urgent care system: an observational study. BMJ Open 2022 Jul 12;12(7):e058964 [FREE Full text] [doi: 10.1136/bmjopen-2021-058964] [Medline: 35820752]
29. Martins S, Augusto C, Martins MRO, José Silva M, Okan O, Dadaczynski K, et al. Adaptation and validation of the digital health literacy instrument for Portuguese university students. Health Promot J Austr 2022 Oct;33 Suppl 1(Suppl 1):390-398. [doi: 10.1002/hpja.580] [Medline: 35124876]
30. Cheng C, Elsworth GR, Osborne RH. Validity evidence of the eHealth Literacy Questionnaire (eHLQ) part 2: mixed methods approach to evaluate test content, response process, and internal structure in the Australian community health setting. J Med Internet Res 2022 Mar 08;24(3):e32777 [FREE Full text] [doi: 10.2196/32777] [Medline: 35258475]
31. Poot CC, Meijer E, Fokkema M, Chavannes NH, Osborne RH, Kayser L. Translation, cultural adaptation and validity assessment of the Dutch version of the eHealth Literacy Questionnaire: a mixed-method approach. BMC Public Health 2023 May 30;23(1):1006 [FREE Full text] [doi: 10.1186/s12889-023-15869-4] [Medline: 37254148]
32. International Standard Classification of Education (ISCED). European Commission. 2019. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php/International_Standard_Classification_of_Education_(ISCED) [accessed 2023-11-21]
33. Turnbull J, Pope C, Prichard J, McKenna G, Rogers A. A conceptual model of urgent care sense-making and help-seeking: a qualitative interview study of urgent care users in England. BMC Health Serv Res 2019 Jul 12;19(1):481 [FREE Full text] [doi: 10.1186/s12913-019-4332-6] [Medline: 31299969]
34. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
35. Reiners F, Sturm J, Bouw LJW, Wouters EJM. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019 Feb 21;16(4):645 [FREE Full text] [doi: 10.3390/ijerph16040645] [Medline: 30795623]
36. Kontos E, Blake KD, Chou WYS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the health information national trends survey 2012. J Med Internet Res 2014;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
37. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021 Nov 15;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
38. Faux-Nightingale A, Philp F, Chadwick D, Singh B, Pandyan A. Available tools to evaluate digital health literacy and engagement with eHealth resources: A scoping review. Heliyon 2022 Aug;8(8):e10380 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10380] [Medline: 36090207]

Abbreviations

ED: emergency department
eHLQ: eHealth Literacy Questionnaire
eHEALS: eHealth literacy scale
LTHC: long-term health condition
NHS: National Health Service
OR: odds ratio
PPI: patient and public involvement
UCC: urgent care center

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.06.23; peer-reviewed by D Furstrand, M Bardus, S Kujala, L Kayser; comments to author 02.11.23; revised version received 21.11.23; accepted 11.04.24; published 04.06.24.

Please cite as:
Turnbull J, Prichard J, MacLellan J, Pope C
eHealth Literacy and the Use of NHS 111 Online Urgent Care Service in England: Cross-Sectional Survey
J Med Internet Res 2024;26:e50376
URL: https://www.jmir.org/2024/1/e50376
doi:10.2196/50376
PMID:38833297

©Joanne Turnbull, Jane Prichard, Jennifer MacLellan, Catherine Pope. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of Digital Early Warning Scores on Hospital Vital Sign Observation Protocol Adherence: Stepped-Wedge Evaluation

David Chi-Wai Wong[1], MEng, DPhil; Timothy Bonnici[2], BSc, MBBS, PhD; Stephen Gerry[3], BSc, MSc; Jacqueline Birks[3], MA, MSc; Peter J Watkinson[4,5,6], MD

[1]Leeds Institute of Health Sciences, School of Medicine, University of Leeds, Leeds, United Kingdom

[2]Critical Care Division, University College Hospital London NHS Foundation Trust, London, United Kingdom

[3]Centre for Statistics in Medicine, University of Oxford, Oxford, United Kingdom

[4]Oxford University Hospitals NHS Trust, Oxford, United Kingdom

[5]NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

[6]Nuffield Department of Clinical Neurosciences, Kadoorie Centre for Critical Care Research and Education, University of Oxford, Oxford, United Kingdom

Corresponding Author:
David Chi-Wai Wong, MEng, DPhil
Leeds Institute of Health Sciences
School of Medicine
University of Leeds
Level 10
Worsley Building
Leeds, LS2 9JT
United Kingdom
Phone: 44 113 343 0806
Email: d.c.wong@leeds.ac.uk

Abstract

Background: Early warning scores (EWS) are routinely used in hospitals to assess a patient's risk of deterioration. EWS are traditionally recorded on paper observation charts but are increasingly recorded digitally. In either case, evidence for the clinical effectiveness of such scores is mixed, and previous studies have not considered whether EWS leads to changes in how deteriorating patients are managed.

Objective: This study aims to examine whether the introduction of a digital EWS system was associated with more frequent observation of patients with abnormal vital signs, a precursor to earlier clinical intervention.

Methods: We conducted a 2-armed stepped-wedge study from February 2015 to December 2016, over 4 hospitals in 1 UK hospital trust. In the control arm, vital signs were recorded using paper observation charts. In the intervention arm, a digital EWS system was used. The primary outcome measure was time to next observation (TTNO), defined as the time between a patient's first elevated EWS (EWS ≥3) and subsequent observations set. Secondary outcomes were time to death in the hospital, length of stay, and time to unplanned intensive care unit admission. Differences between the 2 arms were analyzed using a mixed-effects Cox model. The usability of the system was assessed using the system usability score survey.

Results: We included 12,802 admissions, 1084 in the paper (control) arm and 11,718 in the digital EWS (intervention) arm. The system usability score was 77.6, indicating good usability. The median TTNO in the control and intervention arms were 128 (IQR 73-218) minutes and 131 (IQR 73-223) minutes, respectively. The corresponding hazard ratio for TTNO was 0.99 (95% CI 0.91-1.07; P=.73).

Conclusions: We demonstrated strong clinical engagement with the system. We found no difference in any of the predefined patient outcomes, suggesting that the introduction of a highly usable electronic system can be achieved without impacting clinical care. Our findings contrast with previous claims that digital EWS systems are associated with improvement in clinical outcomes. Future research should investigate how digital EWS systems can be integrated with new clinical pathways adjusting staff behaviors to improve patient outcomes.

(J Med Internet Res 2024;26:e46691) doi:10.2196/46691

XSL•FO
RenderX

KEYWORDS

vital signs; early warning score; track and trigger; electronic charting; stepped-wedge; vital; charting; documentation; deterioration; hospital management; clinical intervention; decision-making; patient risk; hospital; ICU; intensive care unit; UK; United Kingdom; intervention

Introduction

Avoidable mortality from unrecognized clinical deterioration is an internationally recognized problem [1]. Such deterioration often corresponds with deviations in patient vital signs early warning score (EWS) algorithms have been introduced to improve the recognition of abnormal vital signs [2]. They assign a score to each vital sign value according to the degree of abnormality. The total score is a measure of patient risk. Many EWS algorithms have been published and their use is mandated by the National Institute of Health and Care Excellence in the United Kingdom [3,4]. Since 2018, 1 standard EWS, the National Early Warning Score 2, has been mandated in acute hospital trusts [5].

EWS algorithms are accompanied by an escalation protocol, which dictates how frequently the patient should be monitored and what other actions staff should take for each value of the total score. If the EWS score exceeds the "trigger threshold" defined in the escalation protocol, the nursing staff must call a doctor to review the patient.

Despite the widespread adoption of EWS algorithms and associated escalation protocols, patient outcomes have not improved significantly [6,7]. It is possible that errors in the calculation of the EWS are partially to blame. Studies have shown that errors in the calculation of EWS are common and failure to calculate the correct EWS may result in failure to take the correct action [8,9]. Other barriers to escalation include delays in documentation, lack of familiarity with the escalation protocol, failure to follow the protocol, and poor communication [10,11].

Digital EWS systems have been proposed as a solution. These systems automatically calculate the EWS based on data input by staff and display relevant information from the escalation protocol. These data may be displayed to the staff at the bedside, on mobile devices, or at nursing station dashboards, enabling senior clinicians to rapidly survey patient acuity across an area.

At present there is no robust evidence of changes in clinical outcomes to support or refute the case for the introduction of electronic EWS systems. Most recent studies focus on improving the predictive ability of the scoring system itself [12,13], ignoring the complex interaction with health care staff and infrastructure required to affect clinical decision-making. The limited number of studies of digital EWS systems in clinical practice have shown inconsistent results [14-16]. Some have used uncontrolled "before and after" design methodologies, comparing data from periods several years apart, which are limited by their inability to control for temporal confounding such as changes in case mix [17]. Furthermore, very few existing studies have not provided insight into the mechanisms by which any reported improvements were achieved [18].

This study aimed to examine whether the introduction of a digital EWS charting system leads to improvements in patient care. Our causal hypothesis is that, compared with paper charting, the use of a digital EWS system leads to better recognition of patient deterioration and closer adherence to the hospital escalation protocol. These behavior changes would lead to the more frequent observation of patients with abnormal vital signs and therefore earlier escalation. Earlier escalation would lead to improvements in both process metrics and patient outcomes.

Methods

The staged replacement of paper EWS charting with a digital EWS charting system at the Oxford University Hospitals Foundation NHS Trust (OUHFT) provided us the opportunity to conduct a natural experiment using a nonrandomized stepped wedge trial design.

Ethical Considerations

The study protocol was reviewed by the OUHFT's Research and Development department, and based upon Health Care Quality Improvement Partnership guidelines and was deemed to be a service evaluation (ID: 3196), not requiring review by the National Research Ethics Service. All methods were carried out in accordance with the Declaration of Helsinki. As patient data were collected without their consent, permission for informed consent waiver was obtained from the Trust's Caldicott Guardian and Medical Director in accordance with the Health Research Authority Confidentiality Advisory Group guidelines. All study data were deidentified and patients were not compensated. The full study protocol has previously been published and is summarized below [19].

Study Setting

The OUHFT is comprised of 4 hospitals: 1 large teaching hospital, a small district general hospital, and 2 specialist hospitals that do not have emergency departments. Two of the hospitals have intensive care units (ICU) that also act as high-dependency units, and 2 have high-dependency units only.

The digital EWS implemented at OUHFT was the system for electronic notification and documentation (SEND) system [20], a system in which clinical users manually enter vital sign observation data onto a tablet PC. The system then automatically calculates an EWS and displays relevant advice from hospital escalation protocols.

The tablet is physically mounted to a roll-stand with a blood pressure monitor, as shown in Figure 1. The system displays historical vital sign observations of a patient (Figure 1), and ward-level and hospital-level overviews are available via desktop computers.

XSL•FO
RenderX

Figure 1. Overview of the system for electronic notification and documentation (SEND) digital early warning score (EWS) system. The left image shows how SEND is run on a tablet PC, housed alongside the equipment required to take a set of vital sign observations and a barcode scanner to positively identify the patient and clinician. The right image shows a screenshot from SEND, once a patient has been positively identified. Historical vital signs are presented in a familiar format, and an option to document a new set of observations is available via the top-right "Take Obs" button. Further details of this system are provided by Wong et al [20].

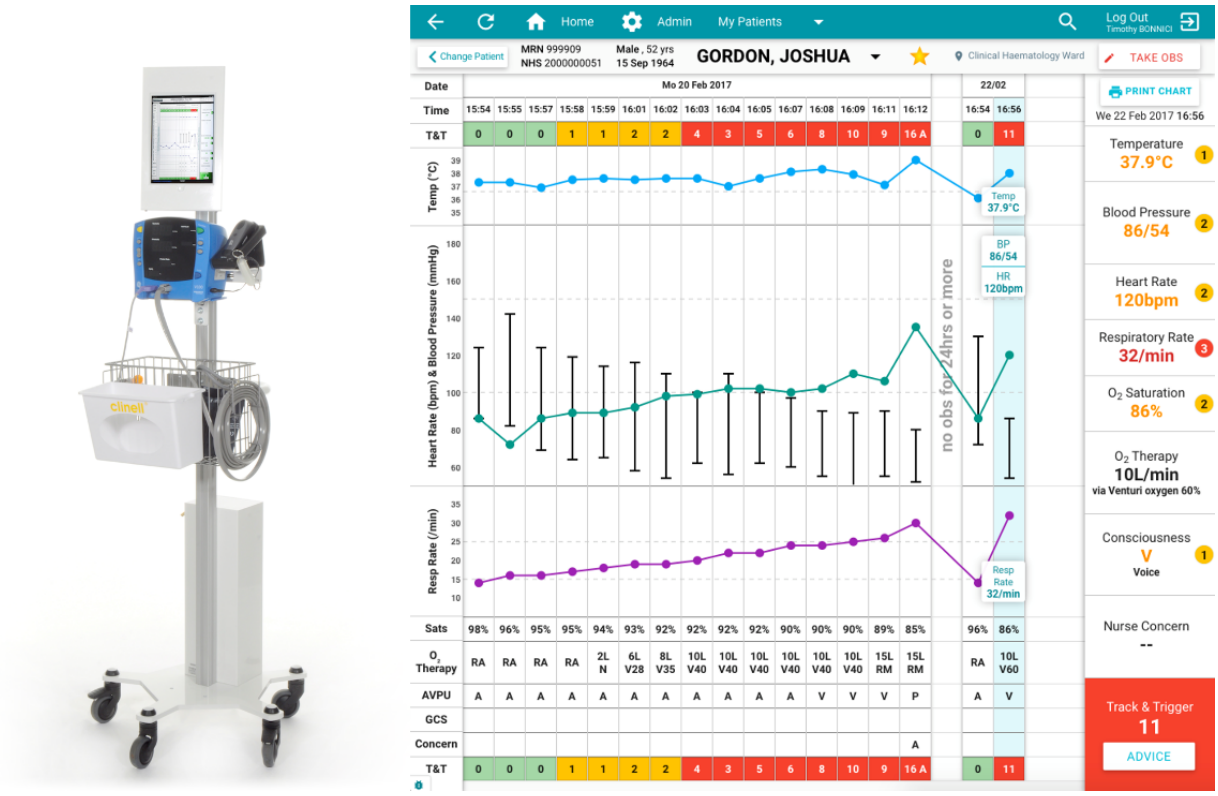

The EWS used at OUHFT was the centile early warning score (CEWS) [21]. CEWS uses 6 vital signs as input parameters, which are each scored from 0 to 3 (Table 1). The trigger threshold is set at 3. For any CEWS greater than or equal to the trigger threshold, the escalation protocol mandates hourly observations and review by a senior doctor. CEWS also allows a nurse to indicate clinical concern. When a nurse is concerned, hourly observations and escalation to a doctor are mandated, irrespective of the CEWS score. A copy of the paper EWS chart and a full description of the escalation protocol are provided in Multimedia Appendix 1.

Table 1. The centile early warning score (CEWS) algorithm used by clinicians at the Oxford University Hospitals Trust to identify patients at risk of deterioration. The subscores for each vital sign are tallied to generate the total EWS. Hospital protocol dictates that a score of 3 or greater warrants senior clinical review, and a new set of observations within 1 hour.

Vital signs	Subscores						
	3	2	1	0	1	2	3
Temperature (°C)	≤35.4	__[a]	35.5-35.9	36.0-37.3	37.4-38.3	—	≥38.4
Heart rate (/min)	≤42	43-49	50-53	54-104	105-112	113-127	≥128
Systolic blood pressure (mm Hg)	≤85	86-96	97-101	102-154	155-164	165-184	≥185
Respiratory rate (/min)	≤7	8-10	11-13	14-19	20-21	22-24	≥25
SpO_2[b] (%)	≤84	85-90	91-93	≥94	—	—	—
Level of consciousness (AVPU[c] or GCS[d])	P and U GCS ≤13	—	V GCS:14	A GCS: 15	—	—	—

[a]Not available.

[b]SpO_2: peripheral arterial oxygen saturation.

[c]AVPU: Alert, Voice, Pain, Unresponsive.

[d]GCS: Glasgow coma scale.

XSL•FO
RenderX

Trial Design

The stepped-wedge study comprised 2 arms, a control arm in which vital signs were recorded using paper observation charts and an intervention arm where the digital EWS system, SEND, was used. The EWS and escalation protocol were identical in both arms.

The study consisted of 20 clusters (and 21 steps). We defined a cluster as a group of between 1 and 5 eligible wards that implemented SEND simultaneously. All wards that were due to switch to the SEND system were eligible for inclusion in a cluster; we defined these as "study wards." Study wards included all adult wards across the Trust, except for the obstetric wards, emergency departments, day units, high dependency units, ICUs, and investigation suites, which were excluded as they did not use standard hospital observation recording and escalation policies. We also excluded the 3 wards where the SEND system was initially developed and piloted, as the control condition, paper charting, was no longer used at the commencement of the study.

Clusters of wards were determined by pragmatic considerations related to the safe conduct of the rollout. For example, each cluster only contained wards from an individual hospital. The sequence of study clusters was predetermined by the system rollout strategy and was therefore not randomized.

The rollout schedule is depicted in Figure 2. The time period between the start of each step was typically 2 weeks. The period was occasionally lengthened to account for project management issues such as reduced staffing over the Christmas holidays (exact dates are provided in Multimedia Appendix 2). The final period, which occurred after SEND was fully deployed to all wards, lasted 3 months. The extended period was designed to capture any delayed effects caused by wards adapting to the new system.

XSL•FO
RenderX

Figure 2. Stepped-wedge data collection with respect to the roll-out schedule. Gray represents periods where no study data were collected, yellow represents periods where data were collected for the control arm, green represents periods where study data were collected for the intervention arm, and brown represents the final 3-month post–roll-out during which study data were collected for all clusters and wards. Ward denotes that the system for electronic notification and documentation (SEND) was deployed to an individual ward. Cluster denotes that SEND was deployed on multiple wards simultaneously.

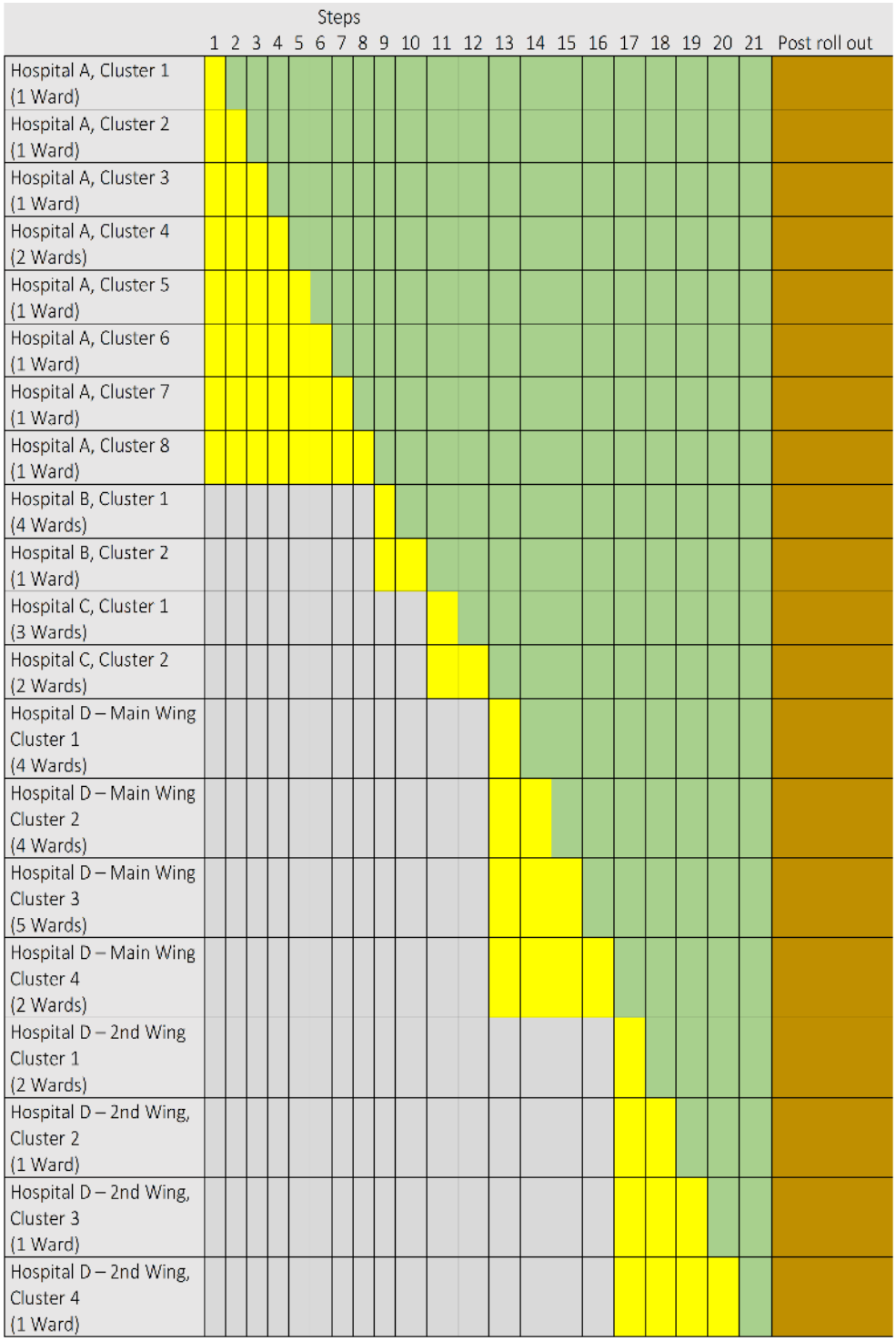

Each study ward admitted multiple patients during each step. Data for this study was obtained at an individual patient level. A patient's data belonged to only 1 step, that is, each cluster and period contained data pertaining to different people. We included all patient admissions to the study wards during the study period rather than censoring data from repeated admissions. Therefore, some patients could potentially contribute data to multiple steps on different admissions. We treated multiple episodes within the same patient as independent, reasoning that the primary outcome was unlikely to be causally related to patient characteristics. We excluded data from

XSL•FO
RenderX

admissions where patients crossed study arms (ie, the ward moved from paper to digital EWS) during their admission.

Data Collection

Data from the control arm were collected by 7 research assistants transcribing data from paper charts located on each study ward into a bespoke electronic form. This was a resource-intensive process, making it unfeasible to collect data from all clusters simultaneously for the duration of the study. Therefore, we commenced data for the control arm at the start of the roll-out to each hospital site and limited it to the site where SEND was actively rolled out (illustrated in Figure 2). To make this tractable, we further split the largest hospital (Hospital D), into 2 sites (Main Wing, second Wing). Data from the intervention arm was continued even once the roll-out of the intervention at a given hospital was complete such that patients from the hospital contributed more data to the intervention arm than data in subsequent hospitals. In summary, data collection may be considered as separate stepped wedges associated with each of the 5 sites, with varying lengths of data from after the intervention.

For each patient admission within each study cluster, we collected patient characteristics (age, gender, Charlson score, admission type, and admitting specialty), the date and times of admission to the ward; first observation with CEWS ≥3 and the immediate subsequent observation; hospital discharge; hospital mortality; transfer to ICU; cardiac arrest call; and theatre admission.

Outcome Measures

The primary outcome measure was the time to next observation (TTNO), defined as the time between a patient's first triggering observations set (CEWS score ≥3) and the subsequent observations set. To address potential confounding by length of ward stay, analysis of the primary outcome measure was restricted to triggering observation sets that occurred within 48 hours of transfer to the first study ward of an admission.

Secondary outcome measures were time to death in the hospital, time to unplanned ICU admission, time to cardiac arrest call, and hospital length of stay (LOS). In each case, the start time was the time of the initial triggering set of observations.

We reported these outcomes for the subgroup included in the analysis of the primary outcome measure (ie, those patients who had a CEWS score ≥3), in line with our causal hypothesis. We also reported the secondary outcomes for all eligible admissions. In these analyses, we used the time of admission to the study ward as the start time.

Finally, we reported system usability to provide further context. System usability was measured using the system usability scale, a validated 10-item questionnaire that is used to generate a score between 0 and 100 [22]. We delivered the questionnaire electronically to all users of the digital system. The questionnaire is included in Multimedia Appendix 3.

Sample Size

The upper bound on the number of patient admissions included in the study was determined by the pragmatic roll-out schedule of the intervention. To determine whether this would be sufficient, we initially undertook a power calculation for steps 1-8, using unpublished pilot data from the Computer Alerting Monitoring System 2 study [23]. We assumed that the proportion of patients who have a further observation within 3 hours of recording an EWS ≥3 would be 0.5 in the paper arm and 0.6 in the electronic arm, that there would be an average of 11 patients with an initial CEWS ≥3 per cluster, and conservatively that the intracluster correlation will be 0.15. The power was then estimated to be 79.3% for a 5% α level. While the calculation depended on statistics estimated from limited pilot data, it indicated that the inclusion of all steps would be sufficiently powerful to detect a difference of 10% in the primary outcome between groups. Full details of this calculation are provided in Multimedia Appendix 4.

Statistical Methods

The primary outcome, the difference in TTNO between arms, was analyzed using a mixed-effects Cox model with a random intercept for cluster and a fixed effect for time as described by Hussey and Hughes [24]. The model included in-hospital death, ICU admission, theatre admission, and cardiac arrest calls as competing events.

We conducted a sensitivity analysis using 5 variants of the basic Hussey and Hughes model, as originally proposed by Hemming et al [25]. The five variants were: (1) time by strata interaction (fixed effects), (2) time by cluster interaction (random effects), (3) treatment by strata interaction (fixed effects), (4) treatment by cluster interaction (random effects), and (5) treatment by time interaction (fixed effects). Secondary outcomes were analyzed using the same method.

To aid interpretation, we calculated the average TTNO in each arm as the mean of the median (IQR) TTNO within each unit of the stepped wedge cluster.

We reported baseline descriptive statistics on patient characteristics, including age and sex, by study arm. We also reported these data for each time period to help understand whether trends in baseline characteristics differed between the control and intervention arms.

Results

Overview

We conducted the study between January 2015 and September 2016, after the conclusion of the rollout of SEND. During this time, there were 90,262 admissions to the study wards. For 2927 (3%) of admissions, vital signs were recorded on both paper and SEND systems and thus excluded. Of the remaining 87,335 admissions, 40,885 (47%) had vital signs recorded exclusively on paper (control arm) and 46,450 (53%) admissions involved patients who had vital signs recorded exclusively using SEND (intervention arm). Of the admissions in the control arm, 11,597 occurred during the implementation period and were available for data capture. In total, 12,802 admissions were entered into the analysis, consisting of 1084 admissions in the control arm and 11,718 admissions in the intervention arm that had a triggering observation within 48 hours of arrival on their first study ward (Figure 3).

XSL•FO
RenderX

Figure 3. Recruitment flow diagram showing how patient admissions were recruited to this stepped-wedge study. Assignment to the Control or Intervention arm was determined by whether the admitting ward was using the SEND system, which followed the rollout schedule in Figure 2. The imbalance between arms is due to the limited data collection period (shown in Figure 2) prior to switching to SEND. CEWS: centile early warning score; SEND: system for electronic notification and documentation.

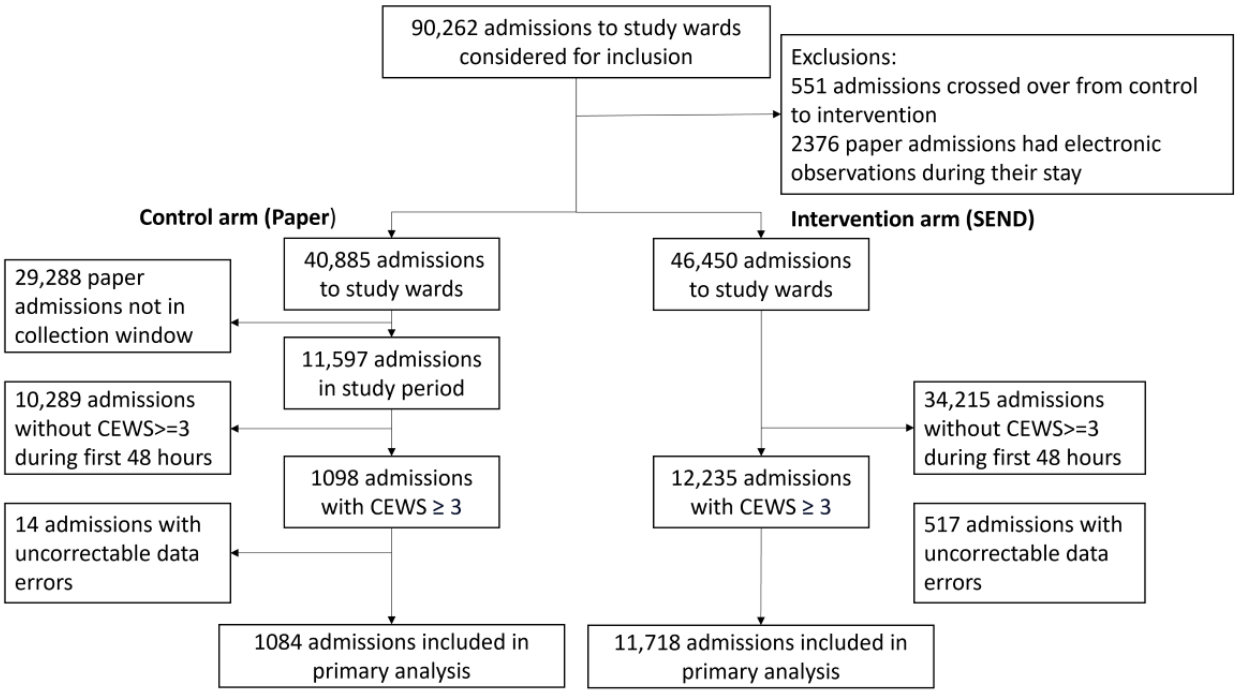

Admission characteristics for the control and intervention are presented in Table 2. Admissions in the intervention arm tended to be slightly older (median age 65 vs 70 years), more likely to be male (49.3% vs 45.6%), and have a higher number of comorbidities (median Charlson score 3 vs 4).

Table 2. Baseline characteristics for the study population, which included all hospital admissions in "study wards" to the Oxford University Hospitals Foundation Trust between January 2015 and September 2016. Study wards included all adult wards across the Trust, except for the obstetric wards, emergency departments, day units, high dependency units, intensive care units, and investigation suites.

Characteristics	Control (paper)	Intervention (SEND[a])
Admissions	1084	11,718
Patients	1048	10,708
Age (years), median (IQR)	65 (49-79)	70 (54-81)
Sex (male), n (%)	494 (45.6)	5777 (49.3)
Charlson score, median (IQR)	3 (0-10)	4 (0-12)
Admission type, n (%)		
Elective	392 (36.2)	4281 (36.5)
Emergency	692 (63.9)	7427 (63.4)
Other	0 (0)	10 (0.1)
Admitting specialty, n (%)		
Medical	430 (40)	5618 (47.9)
Surgical	645 (59.5)	5894 (50.3)
Other	9 (0.8)	206 (1.76)

[a]SEND: system for electronic notification and documentation.

The proportion of male to female sex in both study arms was similar across all steps apart from cluster 1, in which there were a small number of admissions on paper (n=10). There were no males in cluster 20, a cluster that contained only obstetrics and gynecology wards. Proportions of elective and emergency admissions, and medical and surgical admissions, were similar for each study arm across all clusters.

XSL•FO
RenderX

Primary Outcome

There was no significant difference in the TTNO between the 2 arms after adjustment for competing events (Table 3). The median TTNO in the control arm was 128 (IQR 73-218) minutes. The median TTNO in the observation arm was 131 (IQR 73-223) minutes. The hazard ratio of the TTNO using paper charting and the TTNO using SEND was 0.99 (95% CI 0.91-1.07, *P*=.73). All model variants in the sensitivity analysis gave results consistent with the Hussey and Hughes model primary analysis. The numbers of each type of competing events in each arm are shown in Table 4.

Table 3. Hazard ratio for time to next observation (TTNO) after an initial early warning score of 3 or greater. A hazard ratio <1 implies that the TTNO was shorter in the control (paper) arm and >1 implies that the TTNO was shorter in the intervention (system for electronic notification and documentation) arm. There was no significant difference in TTNO using the Hussey and Hughes model, or any other variants.

Model	Hazard ratio (95% CI)	*P* value
Hussey and Hughes model	0.99 (0.91-1.07)	.73
Time by strata interaction (FE[a])	Does not fit	—[b]
Time by cluster interaction (RE[c])	0.98 (0.91-1.07)	.72
Treatment by strata interaction (FE)	0.96 (0.83-1.12)	.63
Treatment by cluster interaction (RE)	0.99 (0.90-1.07)	.73
Treatment by time interaction (FE)	Does not fit	—

[a]FE: Fixed Effects.

[b]Not available.

[c]RE: Random Effects.

Table 4. Number of competing events in the control and intervention arms. The hospital admissions in both study arms had similar percentages of deaths, intensive care unit admissions, theatre admissions, and arrest calls.

Competing events	Control (paper), n (%)	Intervention (SEND[a]), n (%)
Death	50 (5)	826 (7)
ICU[b] admission	22 (2)	237 (2)
Theatre admission	181 (14)	1508 (12)
Arrest call	4 (<1%)	44 (<1%)

[a]SEND: system for electronic notification and documentation.

[b]ICU: intensive care unit.

Figure 4 shows the TTNO for each step during the study. Confidence intervals for the electronic arm were much narrower than the electronic arm because there was more electronic data (collected after the initial intervention rollout period). There was a marked variation in the TTNO according to cluster (Figure 4); the introduction of the digital system did not reduce this variance. There was insufficient power to determine if the intervention had an impact at a cluster level. However, we note that there appeared to be a large reduction in TTNO for cluster 12, which were acute general medicine wards.

XSL•FO
RenderX

Figure 4. Time to next observations after an initial early warning score of 3 or greater by cluster for the control arm (left) and intervention arm (right). There was insufficient power to determine a statistically significant effect at the cluster level, as evidenced by overlapping confidence intervals between the left and right columns.

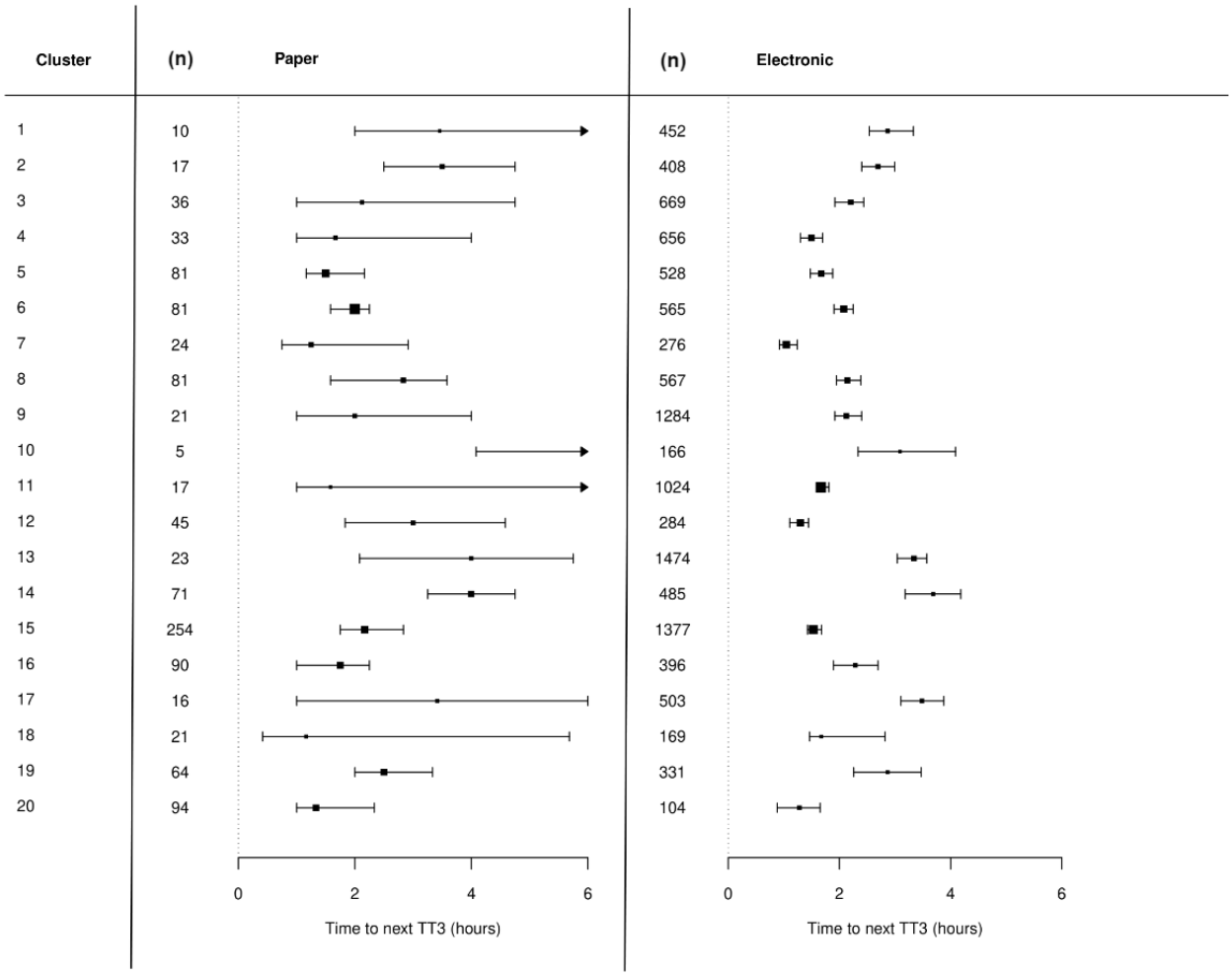

Secondary Outcomes

The introduction of SEND had no significant effect on time to death in hospital, LOS, or time to unplanned ICU admission for the cohort included in the primary analysis (Table 5). There were only 48 cardiac arrest calls across the 2 arms of the study, therefore, there were insufficient events to model this outcome. The findings were consistent irrespective of modeling assumptions. Sensitivity analyses are reported in Multimedia Appendix 5.

Table 5. Hazard ratio for secondary outcomes. A hazard ratio of <1 implies that the outcome was shorter in the control (paper) arm and >1 implies that the outcome was shorter in the intervention (system for electronic notification and documentation) arm. None of the secondary outcomes were statistically significant at the *P*=.05 level.

Outcome	Hazard ratio (95% CI)	*P* value
Time to death in hospital	0.96 (0.68-1.36)	.84
Time to ICU[a] admission	1.85 (0.98-3.49)	.06
Hospital length of stay	0.99 (0.65-1.51)	.97

[a]ICU: intensive care unit.

We also calculated the same secondary measures for the entire patient population (11,597 control and 46,450 intervention), including all those who did not score a CEWS ≥3 within the first 48 hours of admission (Multimedia Appendix 6). For this population, the start time was taken to be the time of admission to the study ward. In this group, there were no significant differences in time to death or LOS. However, there was a borderline reduction in time to ICU admission from the initial triggering set of observations in the intervention arm (hazard ratio 1.25, 95% CI 1.02-1.54).

Usability

System usability scores were only available from Hospital A. The feedback questionnaire was sent to 1891 users, of which 208 (11%) responded. The system usability score was 77.6.

XSL•FO
RenderX

Discussion

Principal Findings

In this large, stepped wedge trial conducted across 4 hospital sites of the same National Health Service trust, the introduction of a digital charting system did not affect the frequency of vital signs recording, nor was it associated with changes in hospital mortality, cardiac arrest rates, or hospital LOS within the subgroup of patients who had a triggering EWS.

Our findings contrast with previous studies of digital vital signs charting. Jones et al [26] reported a reduction in the mean LOS from 9.7 to 6.9 days following the introduction of Patientrack (Alcidion Group Ltd). Schmidt et al [15] reported a reduction in hospital mortality following the introduction of VitalPAC (System C Healthcare Ltd).

The differences between our findings and those of previous researchers may be related to trial design and statistical analysis. A significant strength of our work is the use of a stepped-wedge trial design and a large data set, in line with international recommendations regarding digital health evaluation [27]. Furthermore, we did not institute any new clinical workflows when implementing SEND, which would have confounded the results.

Beyond issues related to design and analysis, 4 other hypotheses could explain our findings. First, it might be that the design or usability of SEND meant that nurses did not engage with the system. However, the system has previously been shown to be more efficient than the charting on paper and the score of 77.6 on the system usability scale is representative of good usability [28,29].

A second possibility is that, although the system was well-liked by staff, advice was not presented at the right time or in the right context and was therefore ineffective in reminding nurses to recheck vital signs [30]. Advice from the hospital protocol was presented at the time of observation recording but there was no mechanism for automatically notifying staff that the next set of observations was due and our implementation did not include the display of the time to the next observations on a dedicated screen at the nursing station. The understanding of how digital systems influence behavior is poorly understood.

A third possibility is that the system was effective in reminding nurses to recheck observations more frequently, but that the reminder alone was insufficient to trigger behavior change. Behavior change requires a combination of capability, opportunity, and motivation [31]. Even if a digital charting system positively alters motivation (through user prompts) and capability (through increased efficiency), these influences may be nullified by competing demands.

Finally, there is the possibility that, even with an effective reminder and supportive context, nurses were exercising clinical judgment and deliberately choosing to deviate from the hospital protocol. The gap between hospital protocols ("work as imagined") and routine clinical practice ("work as done") is well recognized and is often an essential adaptation to ensuring that hospitals continue to function [32]. While the hospital protocol recommended the same frequency of monitoring for all patients with an EWS greater than or equal to 3, our results showed that nurses increased the frequency of vital signs monitoring with the EWS score. It is possible that increasing the frequency of vital signs recording would not improve patient outcomes and rather than the nurses changing practice to match the hospital protocol, the protocol should be changed to match nursing practice more closely.

An unexpected finding was that when including all patients, irrespective of whether they had a triggering observation, the time to ICU admission in the intervention arm was less than in the control arm. Similar reductions in time to ICU transfer have recently been observed in a pre and postintervention study of a digital EWS system that used the electronic Cardiac Arrest Risk Triage EWS [33]. The difference was observed without any difference in the primary outcome measure, which might be explained in 2 ways. Either the result may not correspond to a true effect (which is consistent with the associated wide confidence intervals), or else SEND may be exerting effects via a mechanism other than increased frequency of patient observations.

Limitations

The primary limitation of the study design was that clusters were not randomized but were instead determined by the predetermined phased rollout plan for SEND. Lack of randomization may be a problem since the estimate of the treatment may be unbiased if secular trends exist. To mitigate against this, we included a large number of clusters and explored a variety of analysis methods to examine the possibility of a secular trend. The stepped approach retains advantages over a simple before-after design. The presence of a control group available throughout the study period means that system-level changes may be detected.

A further limitation was the relatively small number of secondary end points. This led to instances in which some clusters had zero secondary end point events. Therefore, conclusions from the secondary outcome analysis ought to be interpreted with caution.

Caution is also required in interpreting the usability survey results. In our original study protocol, we had intended to obtain system usability score data from all new users of the system at the end of roll-out to each hospital site. However, flaws in our survey administration procedures inhibited us from identifying new users versus clinical users who worked in multiple hospitals. Therefore, we only surveyed users of the first site. It is possible that they were not representative of all users. Furthermore, there may be responder bias associated with the low response rate. However, the results obtained in this study are consistent with the findings of questionnaires from staff on pilot wards during the SEND development process [28].

Although data in this study were collected in 2016, we emphasize that the findings remain highly relevant to both the United Kingdom and international health care providers. In the United Kingdom, digital EWS systems are not yet ubiquitous and have been implemented at multiple hospital Trusts in the last year [34,35]. Internationally, the use of both EWS and an

XSL•FO
RenderX

accompanying digital system is an emerging practice [36]. More pertinently, the effectiveness of EWS and the mechanism by which any potential benefits are obtained is still an open question. Indeed, a recent pre- and postevaluation of a digital sepsis score system highlighted the ongoing need for understanding how the use of alert systems evolves over time and impacts clinical workflow [37].

Finally, the findings presented here likely underestimate the true overall benefit of the system. We only examined the effects of SEND using a single measure of observation recording practice, the time between observations, is primarily a reflection of the impact of the system on nursing processes. We did not examine the impact of SEND on other clinical processes or the benefits of secondary use of the data for clinical governance and research.

Conclusion

The introduction of a digital vital signs charting system had no effect on the frequency of vital signs observation or the time to ICU admission, hospital LOS, and hospital mortality in patients with a high EWS. Our findings stand in contrast to previous claims that the introduction of a digital vital signs charting system is associated with significant improvement in clinical outcomes. Future research should continue to investigate the mechanisms by which digital vital signs charting systems alter staff behaviors and improve patient outcomes.

Acknowledgments

The authors thank Soubera Yousefi, Samuel Wilson, Alan Dodge, David Vallance, Simon Kerr, Deolyn Makoni, and Giovanni Rizzo for transcribing information from paper observation charts during the study. This study was supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre, Oxford. System for electronic notification and documentation (SEND) was developed and implemented with funding from the National Health Service England Safer Wards Safer Hospitals Fund. PJW is employed by the OUH Foundation National Health Service Trust. PJW, TB, and DC-W were supported by the National Institute for Health and Care Research Biomedical Research Centre, Oxford.

Data Availability

The data sets generated and analyzed during this study are not publicly available as they are recorded at the patient level, such that it might be possible to reidentify individuals. They are available from the corresponding author on reasonable request.

Authors' Contributions

TB, SG, DC-W, JB, and PJW have substantially contributed to the design of the study and the writing of this manuscript. Statistical analysis was undertaken by SG and JB. All authors read and approved the final manuscript. The funders have not been involved in the study design or reporting.

Conflicts of Interest

TB, DC-W, and PJW were part of the team that developed the system for electronic notification and documentation (SEND). Sensyne Health has since purchased the sole license for SEND. DC-W has previously undertaken consultancy for Sensyne Health. PJW was previously employed part-time and held shares in Sensyne Health. SG and JB declare that they have no competing interests.

Multimedia Appendix 1
EWS chart and escalation protocol.
[DOCX File , 182 KB - jmir_v26i1e46691_app1.docx]

Multimedia Appendix 2
Dates of steps.
[DOCX File , 15 KB - jmir_v26i1e46691_app2.docx]

Multimedia Appendix 3
System Usability Scale questionnaire adapted for SEND.
[DOCX File , 15 KB - jmir_v26i1e46691_app3.docx]

Multimedia Appendix 4
Power calculation.
[DOCX File , 435 KB - jmir_v26i1e46691_app4.docx]

Multimedia Appendix 5

XSL•FO
RenderX

Sensitivity analysis.
[DOCX File , 15 KB - jmir_v26i1e46691_app5.docx]

Multimedia Appendix 6
Secondary outcomes for the entire patient population (11,597 control and 46,450 intervention), including all those who did not score a CEWS≥3 within the first 48 h of admission.
[DOCX File , 14 KB - jmir_v26i1e46691_app6.docx]

References

1. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ Qual Saf 2012;21(9):737-745 [FREE Full text] [doi: 10.1136/bmjqs-2011-001159] [Medline: 22927487]
2. Jones D, Mitchell I, Hillman K, Story D. Defining clinical deterioration. Resuscitation 2013;84(8):1029-1034. [doi: 10.1016/j.resuscitation.2013.01.013] [Medline: 23376502]
3. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. BMJ 2020;369:m1501 [FREE Full text] [doi: 10.1136/bmj.m1501] [Medline: 32434791]
4. Centre for Clinical Practice at NICE (UK). Acutely ill patients in hospital: recognition of and response to acute illness in adults in hospital. In: National Institute for Health and Care Excellence: Guidelines. London: National Institute for Health and Clinical Excellence (NICE); 2007.
5. Royal College of Physicians. National early warning score (NEWS) 2. In: Standardising the assessment of acute-illness severity in the NHS. UK: NHS; 2017.
6. Gao H, McDonnell A, Harrison DA, Moore T, Adam S, Daly K, et al. Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. Intensive Care Med 2007;33(4):667-679. [doi: 10.1007/s00134-007-0532-3] [Medline: 17318499]
7. Findlay GP, Shotton H, Kelly K, Mason M. Time to intervene? a review of patients who went cardiopulmonary resuscitation as a result of an in-hospital cardiorespiratory arrest. In: National Confidential Enquiry into Perioperative Deaths (NCEPOD). UK: National Confidential Enquiry into Perioperative Deaths (NCEPOD); 2012.
8. Prytherch DR, Smith GB, Schmidt P, Featherstone PI, Stewart K, Knight D, et al. Calculating early warning scores--a classroom comparison of pen and paper and hand-held computer methods. Resuscitation 2006;70(2):173-178. [doi: 10.1016/j.resuscitation.2005.12.002] [Medline: 16806641]
9. Clifton DA, Clifton L, Sandu D, Smith GB, Tarassenko L, Vollam SA, et al. 'Errors' and omissions in paper-based early warning scores: the association with changes in vital signs--a database analysis. BMJ Open 2015;5(7):e007376 [FREE Full text] [doi: 10.1136/bmjopen-2014-007376] [Medline: 26141302]
10. Downey CL, Tahir W, Randell R, Brown JM, Jayne DG. Strengths and limitations of early warning scores: a systematic review and narrative synthesis. Int J Nurs Stud 2017;76:106-119. [doi: 10.1016/j.ijnurstu.2017.09.003] [Medline: 28950188]
11. Watson A, Skipper C, Steury R, Walsh H, Levin A. Inpatient nursing care and early warning scores: a workflow mismatch. J Nurs Care Qual 2014;29(3):215-222. [doi: 10.1097/NCQ.0000000000000058] [Medline: 24569518]
12. Bedoya AD, Futoma J, Clement ME, Corey K, Brajer N, Lin A, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. JAMIA Open 2020;3(2):252-260 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa006] [Medline: 32734166]
13. Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE - an interpretable and recurrent neural survival model for early prediction of sepsis. Artif Intell Med 2021;113:102036 [FREE Full text] [doi: 10.1016/j.artmed.2021.102036] [Medline: 33685592]
14. Wong DC, Knight J, Birks J, Tarassenko L, Watkinson PJ. Impact of electronic versus paper vital sign observations on length of stay in trauma patients: stepped-wedge, cluster randomized controlled trial. JMIR Med Inform 2018;6(4):e10221 [FREE Full text] [doi: 10.2196/10221] [Medline: 30381284]
15. Schmidt PE, Meredith P, Prytherch DR, Watson D, Watson V, Killen RM, et al. Impact of introducing an electronic physiological surveillance system on hospital mortality. BMJ Qual Saf 2015;24(1):10-20. [doi: 10.1136/bmjqs-2014-003073] [Medline: 25249636]
16. Dawes TR, Cheek E, Bewick V, Dennis M, Duckitt RW, Walker J, et al. Introduction of an electronic physiological early warning system: effects on mortality and length of stay. Br J Anaesth 2014 Oct;113(4):603-609 [FREE Full text] [doi: 10.1093/bja/aeu107] [Medline: 24878563]
17. Goodacre S. Uncontrolled before-after studies: discouraged by Cochrane and the EMJ. Emerg Med J 2015;32(7):507-508 [FREE Full text] [doi: 10.1136/emermed-2015-204761] [Medline: 25820301]
18. Lang A, Simmonds M, Pinchin J, Sharples S, Dunn L, Clarke S, et al. The impact of an electronic patient bedside observation and handover system on clinical practice: mixed-methods evaluation. JMIR Med Inform 2019;7(1):e11678 [FREE Full text] [doi: 10.2196/11678] [Medline: 30839278]

XSL•FO
RenderX

19. Bonnici T, Gerry S, Wong D, Knight J, Watkinson P. Evaluation of the effects of implementing an electronic early warning score system: protocol for a stepped wedge study. BMC Med Inform Decis Mak 2016;16:19 [FREE Full text] [doi: 10.1186/s12911-016-0257-8] [Medline: 26860362]
20. Wong D, Bonnici T, Knight J, Morgan L, Coombes P, Watkinson P. SEND: a system for electronic notification and documentation of vital sign observations. BMC Med Inform Decis Mak 2015;15:68 [FREE Full text] [doi: 10.1186/s12911-015-0186-y] [Medline: 26268349]
21. Tarassenko L, Clifton DA, Pinsky MR, Hravnak MT, Woods JR, Watkinson PJ. Centile-based early warning scores derived from statistical distributions of vital signs. Resuscitation 2011;82(8):1013-1018. [doi: 10.1016/j.resuscitation.2011.03.006] [Medline: 21482011]
22. Brooke J. SUS: a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, editors. Usability Evaluation in Industry. United Kingdom: Taylor & Francis; 1996:4-7.
23. Watkinson P, Pimentel MCL. Early detection of physiological deterioration in post-surgical patients using wearable technology combined with an integrated monitoring system: a pre- and post-interventional study. 2020 Presented at: medRxiv; December 02 2020; UK URL: https://doi.org/10.1101/2020.12.01.20240770 [doi: 10.1101/2020.12.01.20240770]
24. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 2007;28(2):182-191. [doi: 10.1016/j.cct.2006.05.007] [Medline: 16829207]
25. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials 2017;18(1):101 [FREE Full text] [doi: 10.1186/s13063-017-1833-7] [Medline: 28259174]
26. Jones S, Mullally M, Ingleby S, Buist M, Bailey M, Eddleston JM. Bedside electronic capture of clinical observations and automated clinical alerts to improve compliance with an early warning score protocol. Crit Care Resusc 2011;13(2):83-88. [Medline: 21627575]
27. Michie S, van SMM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
28. Wong D, Bonnici T, Knight J, Gerry S, Turton J, Watkinson P. A ward-based time study of paper and electronic documentation for recording vital sign observations. J Am Med Inform Assoc 2017;24(4):717-721. [doi: 10.1093/jamia/ocw186] [Medline: 28339626]
29. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum-Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
30. Medlock S, Wyatt JC, Patel VL, Shortliffe EH, Abu-Hanna A. Modeling information flows in clinical decision support: key insights for enhancing system effectiveness. J Am Med Inform Assoc 2016 Sep;23(5):1001-1006. [doi: 10.1093/jamia/ocv177] [Medline: 26911809]
31. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
32. Johnson A, Clay-Williams R, Lane P. Framework for better care: reconciling approaches to patient safety and quality. Aust Health Rev 2019;43(6):653-655. [doi: 10.1071/AH18050] [Medline: 30111484]
33. Winslow CJ, Edelson DP, Churpek MM, Taneja M, Shah NS, Datta A, et al. The impact of a machine learning early warning score on hospital mortality: a multicenter clinical intervention trial. Crit Care Med 2022;50(9):1339-1347. [doi: 10.1097/CCM.0000000000005492] [Medline: 35452010]
34. NHS England launches tech trials to boost health and care connectivity. NHS Digital. 2023. URL: https://digital.nhs.uk/news/2023/nhs-england-launches-tech-trials-to-boost-health-and-care-connectivity [accessed 2023-11-16]
35. 100% of NEWS scores are not accurate as Wye Valley NHS Trust goes live with electronic observations. IMS Maxims. URL: https://www.imsmaxims.com/case-studies/wye-valley-nhs-trust-imsmaxims-eobs [accessed 2023-11-16]
36. Mills D, Schmid A, Najajreh M, Al Nasser A, Awwad Y, Qattush K, et al. Implementation of a pediatric early warning score tool in a pediatric oncology ward in Palestine. BMC Health Serv Res 2021;21(1):1159 [FREE Full text] [doi: 10.1186/s12913-021-07157-x] [Medline: 34702268]
37. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]

Abbreviations

CEWS: centile early warning score
EWS: early warning score
ICU: intensive care unit
LOS: length of stay
OUHFT: Oxford University Hospitals Foundation NHS Trust
SEND: system for electronic notification and documentation
TTNO: time to next observation

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.02.23; peer-reviewed by SB Ho, D Barra, C Subbe; comments to author 30.10.23; revised version received 17.11.23; accepted 08.04.24; published 20.06.24.

Please cite as:
Wong DCW, Bonnici T, Gerry S, Birks J, Watkinson PJ
Effect of Digital Early Warning Scores on Hospital Vital Sign Observation Protocol Adherence: Stepped-Wedge Evaluation
J Med Internet Res 2024;26:e46691
URL: https://www.jmir.org/2024/1/e46691
doi:10.2196/46691
PMID:38900529

©David Chi-Wai Wong, Timothy Bonnici, Stephen Gerry, Jacqueline Birks, Peter J Watkinson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Insights From a Clinically Orientated Workshop on Health Care Cybersecurity and Medical Technology: Observational Study and Thematic Analysis

Isabel Straw[1], BMedSci, BMBS, MPH, MSc; Irina Brass[2], BA, MSc, PhD; Andrew Mkwashi[3], BSc, MBA, PhD; Inika Charles[4], BA, LLB, MPA; Amelie Soares[4], BA, MPA; Caroline Steer[4], BA, MPA

[1]Institute of Health Informatics, University College London, London, United Kingdom
[2]Department of Science, Technology, Engineering and Public Policy, University College London, London, United Kingdom
[3]Newcastle University, Newcastle, United Kingdom
[4]University College London, London, United Kingdom

Corresponding Author:
Isabel Straw, BMedSci, BMBS, MPH, MSc
Institute of Health Informatics
University College London
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 020 7679 2000
Email: isabelstraw@doctors.org.uk

Abstract

Background: Health care professionals receive little training on the digital technologies that their patients rely on. Consequently, practitioners may face significant barriers when providing care to patients experiencing digitally mediated harms (eg, medical device failures and cybersecurity exploits). Here, we explore the impact of technological failures in clinical terms.

Objective: Our study explored the key challenges faced by frontline health care workers during digital events, identified gaps in clinical training and guidance, and proposes a set of recommendations for improving digital clinical practice.

Methods: A qualitative study involving a 1-day workshop of 52 participants, internationally attended, with multistakeholder participation. Participants engaged in table-top exercises and group discussions focused on medical scenarios complicated by technology (eg, malfunctioning ventilators and malicious hacks on health care apps). Extensive notes from 5 scribes were retrospectively analyzed and a thematic analysis was performed to extract and synthesize data.

Results: Clinicians reported novel forms of harm related to technology (eg, geofencing in domestic violence and errors related to interconnected fetal monitoring systems) and barriers impeding adverse event reporting (eg, time constraints and postmortem device disposal). Challenges to providing effective patient care included a lack of clinical suspicion of device failures, unfamiliarity with equipment, and an absence of digitally tailored clinical protocols. Participants agreed that cyberattacks should be classified as major incidents, with the repurposing of existing crisis resources. Treatment of patients was determined by the role technology played in clinical management, such that those reliant on potentially compromised laboratory or radiological facilities were prioritized.

Conclusions: Here, we have framed digital events through a clinical lens, described in terms of their end-point impact on the patient. In doing so, we have developed a series of recommendations for ensuring responses to digital events are tailored to clinical needs and center patient care.

(J Med Internet Res 2024;26:e50505) doi:10.2196/50505

KEYWORDS

digital health; clinical medicine; biotechnology; medical device; device regulation; medical education; eHealth; digital medicine; health care; health care cybersecurity; internet of medical things

XSL•FO
RenderX

Introduction

The patient who seeks medical care due to a medical device fault, cybersecurity exploit, or failure in digital health infrastructure may encounter a clinical team that lacks an understanding of the nature of their condition [1-7]. These digital events are often framed as computing issues, yet in practice they manifest as patient symptoms and signs and pose significant challenges to the treating clinicians at the point-of-care [1-9]. In our digitized society, where health care provision is increasingly reliant on technological infrastructure, the "Internet of Medical Things," and connected and intelligent medical devices, computing issues are increasingly translating into clinical complaints [2-8,10-16].

The Landscape of Digital Health Technologies

The proliferation of digital technologies in the health care sector has accelerated over the past decade, with new medical devices entering the market, the growth of consumer health technologies, and the introduction of novel digital tools to hospital workflows (eg, cloud-connected care platforms, digital assistants, and remote monitoring) [10-12,17-22]. These devices are connected to communication networks and the Internet to send, store, and process data in the cloud, forming integral components of the evolving "Internet of Medical Things" [10-12,21,23]. A subset of medical devices comprise stand-alone software, known as Software as a Medical Device, which may incorporate varying degrees of artificial intelligence (AI) approaches and locked or adaptive machine learning [10,17,24]. These novel digital tools present many opportunities for improving patient care, yet they have also introduced new vulnerabilities into the health care system that may impact patient safety [10,16,21,24-28].

New Risks in the Digital Health Landscape

Our increased reliance on digital infrastructure opens us up to new digital risks, exemplified by the increasing number of cyberattacks affecting the health care sector [2,25-28]. Borycki et al [29] have provided an overview of the new types of technology-induced errors that have arisen in the health care system with the introduction of health information technologies. The authors detailed the dangers of an overreliance on technology and explored the specific challenges faced by clinicians who are "digital natives," including unrealistic expectations towards fault tolerance and the availability of digital systems [29,30]. Sax et al [30] consider the potential patient harm associated with a range of IT failures, including loss of system availability, loss of data, and loss of data integrity. Alemzadeh et al [16] place technological events in their clinical context, linking adverse computer incidents to end-point clinical symptoms, uncovering a range of safety-critical computer failures that have resulted in significant patient harm and death.

In the United Kingdom, the retrospective analyses of the National Health Service (NHS) WannaCry attack highlighted the importance of cybersecurity for patient safety; yet, many NHS hospitals still lack guidance regarding the incident response to a cyberattack [28]. Furthermore, security researchers have sounded the alarm regarding the potential for individual-level health care attacks termed "MedJacking," referring to the remote manipulation of patients' medical devices such as insulin pumps and deep brain stimulators [6,31]. When such adverse events occur, research has described the challenges encountered by clinicians for whom the clinical manifestations of technological failures may be unfamiliar [1,5,15]. A recent review of patient illnesses stemming from digital technologies—termed "Biotechnological syndromes"—detailed a range of clinical presentations relating to implanted medical devices (eg, complications of neurostimulators) and technologies within the wider health care system (eg, harms from failures in drug delivery systems) [5].

Clinical complications may also arise from digital technologies not traditionally thought of as "digital health" technologies. In particular, the rise in biohacking technologies has presented unexpected challenges to clinicians, as described by Fram et al [7] and Gangadharbatla [32] in their review of the clinical considerations of consumer implants and microchips. Clinical presentations may also be affected by technology through unexpected mechanisms, as highlighted in cases of technology-facilitated abuse in health care settings [33,34]. Research from the domestic violence field has described the harms faced by patients encountering technology-facilitated abuse (eg, harm inflicted through the manipulation of smart devices) and the need for clinicians to update safeguarding protocols to encompass these risks [33,34].

The introduction of health care AI has brought forward novel ethical questions regarding accountability and liability, with researchers raising concerns about the removal of the "human-in-the-loop" in clinical systems that embed autonomous functions [10,26]. Farhud and Zokaei [35] and Habli et al [36] discuss the challenges of determining moral accountability in complex sociotechnical systems involving AI software, and Fahrud and Zokaei frame these challenges through the traditional pillars of medical ethics (autonomy, beneficence, nonmaleficence, and justice). Dufour et al [37] highlight the potential clinical dangers of autonomous systems in their report, which describes several cardiac arrests stemming from algorithmic errors in a series of ventilators. In addition to concerns surrounding AI autonomy, ethicists have illuminated further issues associated with these technologies, including the risks of bias and discrimination in AI-supported clinical decision tools [38-41].

Clinical Medicine and Digital Health Risks

Currently, clinicians receive little training on the emerging digital technologies that their patients rely on for care and that professionals depend on to perform their work [1-8,15]. As detailed by Sally Adee in her comprehensive history of the life sciences and physical sciences, the separation of these 2 scientific domains over the past 200 years has resulted in distinct professional languages and expertise, which often struggle to understand one another [42]. However, with the advance of biodigital convergence and the growing prevalence of digital health technologies, clinicians are increasingly likely to encounter patients that are no longer purely biological but rely on varied digital devices requiring an understanding of the physical sciences [5,42,43]. Furthermore, the government bodies tasked with overseeing digital health technologies have faced the challenge of understanding biomedical innovations arriving

XSL•FO
RenderX

from overseas, requiring an alignment of domestic standards with the global market [10,24,44]. As a result, the technologies that play a key role in shaping a patient's journey may be poorly understood by the health care practitioners and the governmental agencies responsible for public oversight.

Research Aim

The disciplines of clinical medicine, cybersecurity, and engineering have long operated in silos, and as a result, we lack effective frameworks for patients whose health complaints emerge from the intersection of these domains. In this paper, we present the findings of a confidential multistakeholder workshop that bridged the gap between these professions, facilitating an interdisciplinary discussion on the key challenges affecting digitally dependent patients. We provide insights from frontline staff on the difficulties faced during digital events (eg, cyberattacks and device failures), detail differing perspectives from varied stakeholder groups, and present a series of recommendations for ensuring best clinical practice in the evolving digital health care infrastructure.

Methods

Overview

The workshop titled "Emerging Digital Technologies in Patient Care: Dealing with connected, intelligent medical device vulnerabilities and failures in the health care sector" took place at Goodenough College, London, United Kingdom in February 2023. The workshop was part of an EPSRC-funded project investigating how health care systems, regulations, and standards are responding to the cybersecurity and algorithmic integrity challenges posed by the growing use of connected and intelligent medical devices.

Participant Recruitment

The workshop of 52 participants had representation from the European Union, the United Kingdom, and the United States, involving a wide range of stakeholder groups (Table 1).

All workshop participants were recruited based on their practical experience and expertise in digital health care. Limited snowball sampling was used to identify practitioners and experts in the field. Participants were recruited via email by the research team, leveraging the academic, clinical, and professional networks of IS, IB, and AM and their affiliated institutions. Thus, participants were aware of the researchers' academic profiles prior to the workshop and were provided with detailed information in advance describing the goals of the research.

No patient or health care data were used for the workshop, and all scenarios (Textbox 1 [43]) were fictitious and informed by published research. Workshop discussions were held under the Chatham House Rule and neither the identity nor the organizational affiliation of participants will be disclosed in research outputs derived from the workshop. Participants were identified solely by broad stakeholder categories (Table 1) using different badge colors for note-taking purposes. No compensation was provided to workshop participants.

Table 1. Details of workshop participants and their disciplinary backgrounds (N=52).

Stakeholder category	Participants
Health care professionals or clinicians	20
Public body representatives	3
Device manufacturers and developers	6
Standards bodies representatives	4
Regulatory consultants and advisers	5
Academic professionals	14

XSL•FO
RenderX

Textbox 1. Details of the clinical scenarios designed for table-top exercises specific to different specialties, all of which are based on published case reports.

Acute medicine

Caring for medical patients during a weekend cyberattack: In scenario 1, a hospital cyberattack compromised a (1) cloud-based platform that detailed chemotherapy regimens for oncology patients, (2) smart drug-delivery systems within the hospital, and (3) patient electronic health records (EHRs). The patients described within the scenario all required specific dosing of medications and careful fluid management; the group had to prioritize patients for care and design a wider hospital response.

Acute medicine

Managing unwell patients during a cyberattack on the acute medical unit (AMU): scenario 2 focused on clinical cases requiring careful management of acid-base conditions that necessitated effective blood gas analysis (diabetic ketoacidosis and renal failure) and attentive fluid management (decompensated heart failure); however, the laboratory, blood gas machine, and smart pumps were all compromised.

Acute medicine

Treating blind—patient care during a radiological cyberattack: scenario 3 focused on patient conditions where clinical decision-making relied on radiological information, including (1) identification of the pneumothorax for chest drain insertion, (2) use of a chest x-ray to confirm the position of the nasogastric tube, (3) magnetic resonance imaging to diagnose cauda equina syndrome, and (4) computed tomographic imaging to identify an intercranial bleed. In the scenario, the EHR system was unavailable, and the radiological imaging system was known to be compromised (although the impact on the integrity of scans was unclear).

Surgery and obstetrics

Mother, baby, and spinal cord stimulator: In scenario 4, the team needed to decide on the best management for a pregnant patient presenting with signs of labor, who reported having a closed-loop spinal cord stimulator in situ. The patient is likely to need a caesarean section due to the breech position of the baby but has not had a preanesthetic evaluation, and there is no available information regarding the spinal cord stimulator. The model of a spinal stimulator has artificial intelligence–integrated functionalities, including the self-adjustment of settings based on patient posture. The team must design a safe clinical care plan that accounts for the device and any complications.

Emergency and intensive care

Patient care and ventilator autonomy: scenario 5 was based on the intensive care unit, where a ventilator malfunction causes patients to go into cardiac arrest. The ventilators have integrated automated functions that allow them to update their own settings; however, the attending clinicians are unable to interpret the settings of the machine. The team was tasked with determining the immediate clinical response and discussing the wider implications of closed-loop life support systems in critical medical settings.

General practice

Seizure outbreaks in epilepsy management apps: In scenario 6, a teenager presents to her general practitioner with a relapse in seizure symptoms, having a known diagnosis of epilepsy that had been previously well controlled. It is suggested that the epilepsy management app on her phone has been compromised. The team was tasked with planning an appropriate response for this patient and the wider population's health implications.

Ethical Considerations

Ethical approval was obtained from the UCL Research Ethics Committee (no 222137/001-0023A). A participant information sheet detailing the research project and a consent form were circulated in advance, and participants were asked to return the consent form via email. All participants provided written informed consent for their contributions to be used as anonymized research data.

Workshop Structure

The workshop consisted of two parts: (1) a series of expert talks followed by a Question and Answer (Q&A) session and group discussion, and (2) breakout table-top exercises in which participants discussed and designed a response to a clinical scenario complicated by technology (Textbox 1). Clinical scenarios were written to account for the diversity of health care technologies and the disciplinary backgrounds represented in the room. All scenarios were based on published case reports and reported issues related to digital health technologies. The full case scenarios are provided in Multimedia Appendix 1 and detailed in the web-based workshop description [45]. Notetakers were present throughout the workshop to record details of Q&A sessions, group discussions, and table-top exercises (audio and visual recording was not used). Notetakers received in-person training in advance of the workshop to review the research materials, discuss the scenarios, address any questions, and agree on a framework for the data collection process.

Research Data Analysis

Extensive notes taken during the workshop were retrospectively analyzed and the data from the 5 note-takers was cross-referenced to ensure consistency in the reported results. Four of the researchers coded the qualitative data and undertook an inductive thematic analysis to extract major and minor themes present within the text. Anonymized examples of clinical cases involving technology provided by clinicians were collected as use cases and described in the results.

Results

Our results are divided into themes extracted from panel Q&As and group discussions, and those identified from the breakout exercises on different clinical scenarios.

XSL•FO
RenderX

Group Discussions: Identified Themes

Digital infrastructure: "The Hospital Still Has Windows 10"

The outdated IT infrastructure of the NHS in the United Kingdom is a known risk for cyberattacks, as these systems can more easily be exploited by malicious actors [25-28,45,46]. Participants described the compounding effect that poor IT infrastructure has on their digital behavior, which may further exacerbate the risk of cyber-exploitation. For example, the fact that IT systems are often slow, broken, or unreliable results in clinicians sharing computers, logging in and engaging in risky cyber-hygiene practices. Additionally, the current pressure on NHS staff due to underfunding and staff shortages results in a lack of capacity for additional training on cybersecurity.

Clinicians expressed that it was hard to get excited about the introduction of digital systems when the basic health care needs of their patients were not being met. Further, practitioners shared concerns regarding the impact of digital innovation on health care inequalities, as access to newer technologies is often mediated by wealth therefore deepening socioeconomic disparities in health outcomes. We heard examples of tech-poverty affecting patients, including accessibility issues related to the rise of telehealth, which is inaccessible to disadvantaged patient groups. Clinicians also raised concerns regarding technology-facilitated abuse, with one participant describing the malicious use of GPS tracking technologies to "Geofence" young women and girls within specific city boroughs, such that abusive parties would receive notifications if they left defined geographic areas (Textbox 2).

Textbox 2. Anecdotes and examples of patient presentations related to technology that were disclosed by participants at the workshop.

Hospital medicine

In a case of a malfunctioning piece of radiological equipment, we had a case where 3 dozen patients were exposed with higher doses, above the diagnostic reference level.

We know we have issues with pacemaker batteries, but we don't know how many are affected. We need to understand how to manage that risk.

Women who was a 35-year-old diabetic patient who died and it was a surprise to the clinical team. The patient had been on a pump and the clinician discussed with the coroner if the pump could have contributed? The husband had thrown the insulin pump in the bin, no one had looked at it so there was no clear cause. We need to collect devices after a death, maybe give them to the police.

Surgery

In Obstetrics and Gynecology, there was a case of the wrong woman being given an Emergency C-Section because the communication lines from 2 fetal CTGs overlapped, and one baby's readings had been assigned to the wrong mother.

Community care

In general practice surgery, Apple watches causing patients to think their heart rhythms are abnormal.

Seen issues in general practice with tech-abuse and GPS tracking. Young girls in some communities have to stay inside their borough, otherwise an alarm goes off to their family or partner.

Trust and Medical Devices: "The Device Should Be Considered Guilty Until Proven Otherwise"

Clinicians expressed concerns regarding the often-unchallenged assumption that exists in the medical community regarding device functionality, with one practitioner sharing the view that "99.9% of clinicians wouldn't expect a device to fail," reinforced by peers stating that they'd be highly unlikely to suspect embedded technology as a source of pathology. In cases where medical devices do fail, manufacturers shared that it is not always easy to identify the cause. Companies may be able to take a piece of technology back and attempt to replicate the failure mode; however, often these faults are attributed to unknown internal or external factors.

Manufacturers highlighted the risk of eroding patient trust when device malfunctions are poorly communicated, detailing the historic challenges that they have faced when communicating issues to patients, stating, "It is easy to spend millions on recalls when it is not necessary." Workshop participants discussed that manufacturers are often cognizant that technologies fail but may feel that publicizing all of these events could cause unnecessary alarm, especially in cases where a fault (eg, a software bug) is likely inconsequential but causes fear due to its presence in a consequential technology (eg, a ventilator). Lastly, representatives from the policy domain highlighted the potential for fake news and disinformation that could result from misinterpreted reports of technological failures. Participants also raised issues of public versus private interests, exploring the role that financial incentives play when making decisions regarding device fault disclosures.

Responsibility and Liability: "Clinicians Don't Have Time to Report This"

Views differed between clinicians and manufacturers on the topics of device failure reporting, post-market surveillance, and professional liability in cases of patient harm. Clinicians shared frustrations that they "don't have time to report this" (referring to device malfunctions), stating that responsibility for consistent surveillance of deployed devices should be with the manufacturer. The issue of transparency was evident when discussing the communication between clinicians, manufacturers, and regulatory agencies, as manufacturers often found themselves limited in the information they could share due to commercial confidentiality. Furthermore, there was confusion as to whether reporting device errors was mandatory, with clinicians stating that this was a voluntary (and unfortunately often underperformed) action, and other participants stating this was an obligation.

XSL•FO
RenderX

Workshop participants heard from representatives from the consumer implant industry, who described the increasing uptake of implanted radio frequency identification and near-field communication chips in younger generations. Such body modification technologies that are sited under the skin can cause medical complications (eg, infections, soft tissue injury); however, these technologies do not fall under medical device regulation. It manifested that members of the public occasionally approach these companies, expressing fears of being chipped by "the government" or "aliens"; yet, in the case of potentially serious mental health concerns, these individuals are not redirected to appropriate health care support. Additional ethically contentious cases were discussed, including requests from family members to "chip" older relatives with dementia, or younger children. There was a clearly identified need for improving safeguarding referral pathways.

Clinical Scenarios

The focused disciplinary table-top exercises facilitated a deeper dive into the specific clinical issues that may arise within each clinical domain.

Medical Scenarios—"With Most Tech, we Don't Know How it Works, so we Don't Know How to Trust it"

The 3 hospital-based medical scenarios described failures in cloud-based treatment platforms, laboratory equipment, radiological systems, and electronic health records. Clinicians initially drew parallels to previous experiences where IT had been compromised, referencing the WannaCry 2017 attack and climate events (eg, heat waves) affecting computer systems. Practitioners described the chaos of shifting to paper-based prescribing and note-taking during on-call and overnight hours, during which time they noted a lack of leadership and defined protocols for responding to the event clinically.

In designing their response, all groups reached consensus that the cyberattack should be classified as a "major incident," activating a chain of responses including the recruitment of additional staff, awareness at the national level, communication across sites, an effective PR response, and the allocation of roles to those with sufficient skills and seniority. Senior team members suggested the development of battle bags and action cards that are commonly used in other major incident events, citing Grenfell Tower and the Manchester Bombing as examples. Clinicians demonstrated some awareness of available resources, including NHS Digital and Chief Information Officers; however, these resources were relatively unheard of within the groups.

The scenario suggested there were cybersecurity vulnerabilities in the drug delivery and laboratory systems present on the wards, to which health care staff reported a preference for shutting down technology entirely as a safety measure (while acknowledging that this could be unnecessary and cause delays and even more harm). When discussing plans to turn off digital systems, the groups shared their concerns about what disabling devices would do and whether there was a default safe mode that could protect patients. In shutting down digital equipment, participants decided to return to rudimentary clinical techniques, including dripping (the act of delivering intravenous medications based on drips of liquid). The teams raised concerns for younger generations of clinicians who may not have this nondigital foundation to fall back on.

Lastly, practitioners identified the patients most at risk of harm in the context of IT manipulation in the hospital and developed a clinical hierarchy specific to digital threats. The patients with diabetic ketoacidosis and cauda equina were identified as likely to experience the worst outcomes due to their reliance on laboratory and radiological resources for treatment. By prioritizing the patients according to the role that technology played in their clinical management, participants framed the cyberattacks in clinical terms.

Surgical Scenario—"Mother, Baby, and Spinal Cord Stimulator"—"There Is no Way of Knowing How the Body Is Communicating With the Device"

The surgical group was tasked with managing a patient in active labor who had a potentially compromised implanted spinal cord stimulator. Through the discussion, it became apparent that there was a lack of clinical knowledge regarding the implications of the technology, and the health care staff opted to focus on the medical management that they did understand while putting the technical components to one side. In discussing the follow-up to the case, standard body representatives raised concerns that the incident would not be flagged as an adverse outcome, supported by clinicians who indicated that this would be unlikely to be reported.

Emergency and Intensive Care Scenario—"Patient Care and Ventilator Autonomy"

The clinicians first drew comparisons to historic crises events, including the 1952 Copenhagen crisis, in which medical students were recruited to manually ventilate patients [47]. When discussing the initial response to the scenario, the team discussed disconnecting all patients from the ventilators while acknowledging the challenges of doing this in an intensive care unit where patients are dependent on life support systems. The group discussed the specific implications of AI and closed-loop systems within medical equipment, identifying the central issue of trust and a lack of understanding of the technological mechanisms. The lack of training in AI systems was felt to be compounded by the absence of training regarding cyberattacks, with clinicians sharing that they had "never had a training day on what would happen in this scenario."

Community Care and General Practice Scenario—"Seizure Outbreaks in Epilepsy Management Apps"

The scenario based on community care described a teenage patient who experienced a relapse in seizure symptoms, suspected to be related to a malicious hack on an epilepsy app. The group began by identifying possible adverse health effects, including seizures, headaches, distress, loss of vision, loss of focus, visual effects, and airway compromise. At-risk patient groups were noted, including dementia patients and those with neurodiversity. For immediate clinical management, practitioners felt that staff were likely to tell the patient to avoid the app or their phone entirely, reiterating a theme heard in the other groups of completely disconnecting from the technology.

XSL•FO
RenderX

In contrast to the other groups, the general practitioners discussed reporting the case to the police and ministry of defense due to the concern of a malicious attack and the implications this could have for a large number of people using the app. Participants also described the need to involve parents as this was a pediatric case and the role of the app company in protecting their users.

Uncovered Examples

During the workshop, practitioners provided anonymized anecdotes of patient harm where technology played a role. Textbox 2 provides these example cases. In reflecting on a patient death, one participant discussed the lack of postmortem guidance, highlighting that the disposal of devices as medical waste precludes the evaluation of their involvement in a death.

Discussion

Principal Findings

In a patient journey, the individual is likely to encounter various forms of technology, from their electronic health record, to advanced, interconnected, automated, and intelligent health care technologies [15,45,48,49]. In these journeys, health care staff are the immediate point of contact when clinical care goes wrong, and patient health deteriorates [1-5]. In this paper, we have explored the points of digital vulnerability that may contribute to patient illness along these trajectories and discussed the issues of cybersecurity, device failures, and faulty AI systems from the perspective of treating clinicians.

Clinicians from a diverse range of specialties responded in a similar manner when confronted with failures in digital devices, opting to immediately shut down the technology. Whether this was disconnecting patients from ventilators, disabling all medication smart pumps, or advising patients to turn off phones that were vulnerable to malicious hacks, the safest measure was often considered to be preventing any ongoing interaction between the technology and the patient's physicality. The response is understandable given that clinicians receive little education on these tools and do not trust the systems or have the confidence to appropriately evaluate them.

Yet when clinicians are not informed of a potential technological failure, the default position appears to be the opposite and to trust devices entirely, such that "99.9% of clinicians wouldn't expect a device to fail." Hence, an interesting dichotomy exists—when a clinician has not been given reason to doubt a device, they will often trust the device over the patient (eg, believing the patient's blood glucose data, as opposed to their subjective symptoms); however, once doubt is introduced, the clinicians opt to disregard the technology completely. The polarity of these reactions mirrors the black-and-white nature of the black-box technology that remains opaque to clinicians using the digital tools. This delicate relationship reinforces the importance of effective communication around medical device failures with clinicians as well as with the public.

The lack of reporting regarding digital adverse events is a regulatory and public policy concern. Textbox 2 provides a list of events we collected in this workshop, of which several remained unreported and describe significant patient harm. Within these stories, we heard new examples of biotechnological syndromes and forms of technology-facilitated abuse that add to the existing literature reporting issues of technology in domestic violence and the risks posed to vulnerable patients [5,33,34].

We heard differing opinions on where responsibility and professional liability should lie when patient harm occurs because of a technology malfunction, with clinicians sharing the view that manufacturers are responsible for ongoing follow-up and manufacturers stating that clinicians are responsible for the outcomes of patients with embedded devices. Health care professionals advocated for a higher standard to be placed on manufacturers with regard to patient trust and transparency, citing parallels to the Hippocratic oath and fundamental medical ethics taught in medical school. These suggestions have previously been made by several health care cybersecurity researchers who developed a "Hippocratic Oath for Connected Medical Devices" [39-41].

Conclusions

Our research has taken the unique approach of positioning digital health care technology failures in their medical context, viewed through the lens of the clinician at the point of care. In doing this, we demonstrate how health care staff can form tailored clinical hierarchies when faced with health care cyberattacks, such as prioritizing patients' dependence on digitally vulnerable systems (eg, spinal injuries requiring radiological imaging) or identifying at-risk groups of mobile screen-based hacks (eg, patients with epilepsy, those who are neurodiverse, and those with dementia). Understanding cyberattacks as clinical attacks in this manner provides an opportunity to form the guidelines and major incident response protocols that our participants identified as urgent and lacking resources in hospital settings.

Our findings illustrate gaps in clinical knowledge regarding medical technology and a lack of confidence in managing these scenarios, which can only be addressed with improved clinical education and training. To ensure effective patient care in our environment of evolving digital infrastructure, historic IT responses to cyberattacks and device failures must be married with the clinical needs and perspectives provided in this report. Our research shines a light on a critical and understudied area at the intersection of clinical medicine and digital health that requires greater research and professional guidance. We provide a series of recommendations based on our findings in Textbox 3.

XSL•FO
RenderX

Textbox 3. Recommendations.

- **Hospital protocols and incident plans:**

  Cyberattacks should be treated as clinical attacks, framed similarly to other major incidents such as terrorist and extreme weather events [50]. To achieve this, cyberattack threat models should be developed with end-point clinical symptoms and signs in mind, necessitating input from clinicians and engineers. Hospitals and health care practices should develop major incident protocols that specify the clinical steps to be taken in a cyberattack, hierarchically prioritizing patient groups, and repurposing existing resources such as action cards, dedicated response teams, battle bags, and communication pathways for escalation.

- **Medical education and clinical training:**

  Health care practitioners require a fundamental understanding of the novel digital technologies that their patients rely on, in order to treat them effectively when digital complications arise. Software-based medical devices, especially those connected to communication networks and with AI-integrated functionalities, require continuous monitoring of their performance by hospital staff and their medical device inventory teams. Digitally themed professional courses through the UK Royal Colleges would incentivize an uptake of this training, in addition to integrating educational content into National Health Service Trust mandatory training modules, clinical orientation and induction weeks in hospitals, and medical school education curricula.

- **Academia and research:**

  Research focused on the health complications of digital technologies needs to advance at a parallel rate to the development and deployment of digital health care tools and devices. Future research focused on the symptoms and signs of digital failures and technological pathology would improve the ability of clinicians to diagnose these cases, and consequently report them to the appropriate bodies.

- **Manufacturer training and support:**

  An easily accessible interface between manufacturers and clinicians is required to ensure health care staff can find appropriate information about the performance of connected and intelligent medical devices in a timely manner. Manufacturers could develop “how-to” cards and clinically tailored resources about their digital medical devices, which would be more useful to health care staff than traditional user manuals.

- **Regulation and reporting:**

  Increased interaction is needed between regulatory agencies, such as the Medicines and Healthcare products Regulatory Agency, and clinical teams. Through events in medical schools, hospitals, and community health care practices, representatives of public and regulatory bodies may provide additional support to clinicians on the reporting processes and existing guidance regarding software-based and connected medical devices.

Limitations

While we engaged a range of critical stakeholders in our workshop, we were limited by sample size and the representativeness of our participants. In future research, it would be beneficial to engage a wider range of clinical specialties, such as dermatologists, oncologists, radiologists, and neurologists, where digital health technologies are expanding rapidly. Furthermore, we have limited our focus to emerging digital technologies, with functionalities including telemetry, internet connectivity, and AI. As a result, we have not examined issues associated with static medical devices; for example, adverse reactions to materials used for hip implants.

Acknowledgments

The authors would like to express their appreciation to the PETRAS team and all stakeholders who took part in our primary research and provided us with a wealth of information on this topic. Special thanks go to our project partners at BSI (the United Kingdom’s National Standards Body) for their guidance and support throughout this project, especially Rob Turpin, Paul Sim, Emma Glass, and Matthew Chiles. This research is funded by the PETRAS National Centre of Excellence in IoT Systems Cybersecurity (EPSRC grant number EP/S035362/1).

Data Availability

The data and clinical scenarios generated during and analyzed during this study are available in Multimedia Appendix 1.

Authors' Contributions

The authorship contributions are listed as per the CRediT authorship criteria. The study was conceptualized by IS, IB, and AM; data were curated by IS, IB, IC, AS, and CS; formal analysis was conducted by IS, IB, IC, AS, and CS; the investigation was carried out by IS, IB, AM, IC, AS, and CS; the methodology was designed by IS, IB, and AM; resources were provided by IB and AM; IB supervised the study; validation was carried out by IB; writing—original draft preparation was carried out by IS; writing—review and editing was carried out by IS and IB. All authors reviewed and approved the final version. The details of the researchers, as per the COREQ (consolidated criteria for reporting qualitative research) checklist, are as follows. At the time of the research study and qualitative data collection, IS (female) was a medical doctor and PhD candidate in artificial intelligence–enabled health care at University College London (UCL); IB (female) was an associate professor of regulation,

XSL•FO
RenderX

innovation and public policy at the Department of Science, Technology, Engineering and Public Policy (STEaPP) at UCL; AM (male) was a research fellow on the Regulation and Standardization of Connected, Intelligent Medical Devices (REG-MEDTECH) project at the STEaPP at UCL; and IC (female), AS (female), and CS (female) were candidates on the masters in public administration in digital technologies and policy at UCL and employed as research assistants in the REG-MEDTECH project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Material: Workshop Scenarios for Dealing with Connected, Intelligent Medical Device Vulnerabilities and Failures. [DOCX File , 42 KB - jmir_v26i1e50505_app1.docx]

References

1. Dameff CJ, Selzer JA, Fisher J, Killeen JP, Tully JL. Clinical cybersecurity training through novel high-fidelity simulations. J Emerg Med 2019;56(2):233-238. [doi: 10.1016/j.jemermed.2018.10.029] [Medline: 30553562]
2. Dameff C, Farah J, Killeen J, Chan T. Cyber disaster medicine: a new frontier for emergency medicine. Ann Emerg Med 2020;75(5):642-647. [doi: 10.1016/j.annemergmed.2019.11.011] [Medline: 31959537]
3. Sommer M, Stiksrud EM, von Eckardstein K, Rohde V, Paulus W. When battery exhaustion lets the lame walk: a case report on the importance of long-term stimulator monitoring in deep brain stimulation. BMC Neurol 2015;15:113 [FREE Full text] [doi: 10.1186/s12883-015-0365-6] [Medline: 26187628]
4. Shao W. Cochlear implant electrode failure secondary to silicone touch-up during device manufacturing. Otol Neurotol 2013;34(7):e72-e75. [doi: 10.1097/MAO.0b013e318298aaaf] [Medline: 23921929]
5. Straw I, Rees G, Nachev P. 21st century medicine and emerging biotechnological syndromes: a cross-disciplinary systematic review of novel patient presentations in the age of technology. BMC Digit Health 2023(1):41 [FREE Full text] [doi: 10.1186/s44247-023-00044-x]
6. Adashi EY, Thomasian NM. Medical devices in harm's way: medjacking. JAMA Health Forum 2020;1(1):e200007 [FREE Full text] [doi: 10.1001/jamahealthforum.2020.0007] [Medline: 36218531]
7. Fram BR, Rivlin M, Beredjiklian PK. On emerging technology: what to know when your patient has a microchip in his hand. J Hand Surg Am 2020;45(7):645-649. [doi: 10.1016/j.jhsa.2020.01.008] [Medline: 32164995]
8. Straw I, Ashworth C, Radford N. When brain devices go wrong: a patient with a malfunctioning deep brain stimulator (DBS) presents to the emergency department. BMJ Case Rep 2022;15(12):e252305 [FREE Full text] [doi: 10.1136/bcr-2022-252305] [Medline: 36572446]
9. Guidance for manufacturers on reporting adverse incidents involving Software as a Medical Device under the vigilance system. UK Government. 2023. URL: https://tinyurl.com/5n85fj7k [accessed 2023-06-30]
10. Mkwashi A, Brass I. The future of medical device regulation and standards: dealing with critical challenges for connected, intelligent medical devices (September 7, 2022). 2022 Presented at: PETRAS National Centre of Excellent in IoT Systems Cybersecurity: London; 07 September 2027; London URL: https://ssrn.com/abstract=4226057 [doi: 10.2139/ssrn.4226057]
11. Pani S, Patra P, Ferrari G, Kraleva R, Wang X. The internet of medical thingsnabling technologies and emerging applications. In Institution of Engineering and Technology (The IET). URL: https://air.unipr.it/handle/11381/2938191 [accessed 2024-05-30]
12. Amato C. Internet of bodies: digital content directive, and beyond. J Intell Property Inf Technol Electron Commerce Law 2021;12(2) [FREE Full text]
13. South L, Borkin M. Ethical considerations of photosensitive epilepsy in mixed reality. OSF Preprints Preprint posted online October 23, 2020 [FREE Full text] [doi: 10.31219/osf.io/y32td]
14. Nishihama K, Eguchi K, Maki K, Okano Y, Tanaka S, Inoue C, et al. Sudden death associated with severe hypoglycemia in a diabetic patient during sensor-augmented pump therapy with the predictive low glucose management system. Am J Case Rep 2021;22:e928090 [FREE Full text] [doi: 10.12659/ajcr.928090]
15. Straw I, Dobbin J, Reaver DL, Tanczer L. Medical cyber crises and biotechnological syndromes: a multisite clinical simulation study focused on digital health complaints. Lancet 2023;402(Suppl 1):S88. [doi: 10.1016/S0140-6736(23)02082-2] [Medline: 37997134]
16. Alemzadeh H, Iyer RK, Kalbarczyk Z, Raman J. Analysis of safety-critical computer failures in medical devices. IEEE Secur Privacy 2013;11(4):14-26. [doi: 10.1109/msp.2013.49]
17. Ronte H. Medtech and the internet of medical things: how connected medical devices are transforming health care. Deloitte. URL: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-medtech-iomt-brochure.pdf [accessed 2024-05-30]
18. Afolabi BA. Remote monitoring of patients with implanted cardiac devices—a review. 2012. URL: https://www.ecrjournal.com/articles/remote-monitoring-patients-implanted-cardiac-devices-review [accessed 2012-04-08]

XSL•FO
RenderX

19. Nurshod A, Khujamatov K, Lazarev A. Remote monitoring system architectures in healthcare. 2021 Presented at: International Conference on Information Science and Communications Technologies (ICISCT). 2021; 2023 Nov 03; Tashkent, Uzbekistan. [doi: 10.1109/icisct52966.2021.9670405]
20. Tarakji KG, Zaidi AM, Zweibel SL, Varma N, Sears SF, Allred J, et al. Performance of first pacemaker to use smart device app for remote monitoring. Heart Rhythm O2 2021;2(5):463-471 [FREE Full text] [doi: 10.1016/j.hroo.2021.07.008] [Medline: 34667961]
21. Turpin R, Hoefer E, Lewelling J, Baird P. Adapting regulatory frameworks and standards to ensure safety and performance. Association for the Advancement of Medical Instrumentation (AAMI), British Standards Institute (BSI). Machine Learning AI in Medical Devices. 2020. URL: https://www.bsigroup.com/en-US/medical-devices/resources/Whitepapers-and-articles/machine-learning-ai-in-medical-devices/ [accessed 2024-05-30]
22. Garg S, Williams NL, Ip A, Dicker AP. Clinical integration of digital solutions in health care: an overview of the current landscape of digital technologies in cancer care. JCO Clin Cancer Inform 2018;2:1-9 [FREE Full text] [doi: 10.1200/CCI.17.00159] [Medline: 30652580]
23. Artificial Intelligence and Machine Learning (AI/ML)—Software as a Medical Device: Action Plan. USA: Food and Drug Administration (FDA); 2021.
24. The Regulation of Artificial Intelligence as a Medical Device. Regulatory Horizons Council, UK Government. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1120503/RHC_regulation_of_AI_as_a_Medical_Device_report.pdf [accessed 2023-11-27]
25. Ghafur S, Grass E, Jennings N, Darzi A. The challenges of cybersecurity in health care: the UK national health service as a case study. Lancet Digit Health 2019;1(1):e10-e12 [FREE Full text] [doi: 10.1016/S2589-7500(19)30005-6] [Medline: 33323235]
26. Ghafur S, Fontana G, Martin G, Grass E, Goodman J, Darzi A. Improving Cyber Security in the NHS. London: Imperial College London URL: https://www.imperial.ac.uk/media/imperial-college/institute-of-global-health-innovation/Cyber-report-2020.pdf [accessed 2024-05-30]
27. O'Brien N, Ghafur S, Durkin M. Cybersecurity in health is an urgent patient safety concern: we can learn from existing patient safety improvement strategies to address it. J Patient Saf Risk Manag 2021;26(1):5-10. [doi: 10.1177/2516043520975926]
28. Ghafur S, Kristensen S, Honeyford K, Martin G, Darzi A, Aylin P. A retrospective impact analysis of the wannaCry cyberattack on the NHS. NPJ Digit Med 2019;2:98 [FREE Full text] [doi: 10.1038/s41746-019-0161-6] [Medline: 31602404]
29. Borycki E, Dexheimer JW, Hullin Lucay Cossio C, Gong Y, Jensen S, Kaipio J, et al. Methods for addressing technology-induced errors: the current state. Yearb Med Inform 2016(1):30-40 [FREE Full text] [doi: 10.15265/IY-2016-029] [Medline: 27830228]
30. Sax U, Lipprandt M, Röhrig R. The rising frequency of IT blackouts indicates the increasing relevance of IT emergency concepts to ensure patient safety. Yearb Med Inform 2016(1):130-137 [FREE Full text] [doi: 10.15265/IY-2016-038] [Medline: 27830241]
31. Pycroft L, Boccard SG, Owen SLF, Stein JF, Fitzgerald JJ, Green AL, et al. Brainjacking: implant security issues in invasive neuromodulation. World Neurosurg 2016;92:454-462. [doi: 10.1016/j.wneu.2016.05.010] [Medline: 27184896]
32. Gangadharbatla H. Biohacking: an exploratory study to understand the factors influencing the adoption of embedded technologies within the human body. Heliyon 2020;6(5):e03931 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03931] [Medline: 32426543]
33. Straw I, Tanczer L. Safeguarding patients from technology-facilitated abuse in clinical settings: a narrative review. PLoS Digit Health 2023;2(1):e0000089 [FREE Full text] [doi: 10.1371/journal.pdig.0000089] [Medline: 36812593]
34. Tanczer L, López-Neira I, Parkin S. 'I feel like we're really behind the game': perspectives of the United Kingdom's intimate partner violence support sector on the rise of technology-facilitated abuse. J Gender-Based Violence 2021;5(3):431.
35. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health 2021;50(11):i-v [FREE Full text] [doi: 10.18502/ijph.v50i11.7600] [Medline: 35223619]
36. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020;98(4):251-256 [FREE Full text] [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
37. Dufour N, Fadel F, Gelée B, Dubost JL, Ardiot S, Di Donato P, et al. When a ventilator takes autonomous decisions without seeking approbation nor warning clinicians: a case series. Int Med Case Rep J 2020;13:521-529 [FREE Full text] [doi: 10.2147/IMCRJ.S266969] [Medline: 33116944]
38. Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Health Care Inform 2022;29(1):e100457 [FREE Full text] [doi: 10.1136/bmjhci-2021-100457] [Medline: 35470133]
39. Wu H, Sylolypavan A, Wang M, Wild S. Quantifying health inequalities induced by data and AI models. 2022. URL: https://www.ijcai.org/proceedings/2022/721 [accessed 2024-05-30]
40. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

XSL•FO
RenderX

41. Straw I. The automation of bias in medical artificial intelligence (AI): decoding the past to create a better future. Artif Intell Med 2020;110:101965. [doi: 10.1016/j.artmed.2020.101965] [Medline: 33250145]
42. We are electric: the new science of our body's electrome. UK: Canongate Books; 2023.
43. Exploring biodigital convergence : what happens when biology and digital technology merge?. 2020. URL: https://horizons.gc.ca/wp- [accessed 2023-10-01]
44. Medical technology strategy. UK Government. URL: https://www.gov.uk/government/publications/medical-technology-strategy/medical-technology-strategy [accessed 2024-05-30]
45. Brass I, Straw I, Mkwashi A, Charles I, Soares MA, Steer C. Emerging digital technologies in patient care: dealing with connected, intelligent medical device vulnerabilities and failures in the healthcare sector. In: Workshop Report. 2023 Presented at: PETRAS National Centre of Excellence in IoT Systems Cybersecurity; June 6 2023; London URL: https://zenodo.org/record/8011139
46. Sharma S. Outdated IoT healthcare devices pose major security threats. CSO Online. URL: https://www.csoonline.com/article/3648592/outdated-iot-healthcare-devices-pose-major-security-threats.html [accessed 2024-05-30]
47. Awad A, Trenfield SJ, Pollard TD, Ong JJ, Elbadawi M, McCoubrey LE, et al. Connected healthcare: improving patient care using digital health technologies. Adv Drug Deliv Rev 2021;178:113958. [doi: 10.1016/j.addr.2021.113958] [Medline: 34478781]
48. Mitchell M, Kan L. Digital technology and the future of health systems. Health Syst Reform 2019;5(2):113-120. [doi: 10.1080/23288604.2019.1583040] [Medline: 30908111]
49. Woods B, Coravos A, Corman JD. The case for a hippocratic oath for connected medical devices: viewpoint. J Med Internet Res 2019;21(3):e12568 [FREE Full text] [doi: 10.2196/12568] [Medline: 30888323]
50. NHS England. Clinical guidelines for major incidents and mass casualty events. URL: https://www.england.nhs.uk/publication/clinical-guidelines-for-major-incidents-and-mass-casualty-events/ [accessed 2023-05-01]

Abbreviations

AI: artificial intelligence
NHS: National Health Service
Q&A: Question and Answer

Edited by T de Azevedo Cardoso; submitted 03.07.23; peer-reviewed by R Röhrig, E van der Velde; comments to author 03.10.23; revised version received 28.11.23; accepted 17.05.24; published 11.07.24.

Please cite as:
Straw I, Brass I, Mkwashi A, Charles I, Soares A, Steer C
Insights From a Clinically Orientated Workshop on Health Care Cybersecurity and Medical Technology: Observational Study and Thematic Analysis
J Med Internet Res 2024;26:e50505
URL: https://www.jmir.org/2024/1/e50505
doi:10.2196/50505
PMID:38990611

©Isabel Straw, Irina Brass, Andrew Mkwashi, Inika Charles, Amelie Soares, Caroline Steer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Patient-Driven Mobile Health Innovation in Cystic Fibrosis Care: Comparative Cross-Case Study

Pamela Mazzocato[1,2], PhD; Jamie Linnea Luckhaus[1,3], MPH; Moa Malmqvist Castillo[1], MPhil; Johan Burnett[4], RN; Andreas Hager[4], LLM; Gabriela Oates[5], PhD; Carolina Wannheden[1], PhD; Carl Savage[1], PhD

[1]Department of Learning, Informatics, Management and Ethics, Medical Management Centre, Karolinska Institutet, Stockholm, Sweden

[2]Södertälje Hospital, Södertälje, Sweden

[3]Participatory e-Health and Health Data, Department of Women's and Child's Health, Uppsala University, Uppsala, Sweden

[4]Upstream Dream, Bromma, Sweden

[5]Pulmonary, Allergy and Critical Care Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Pamela Mazzocato, PhD
Department of Learning, Informatics, Management and Ethics
Medical Management Centre
Karolinska Institutet
Tomtebodavägen 18 A
Stockholm
Sweden
Phone: 46 8 524 83 696
Email: pamela.mazzocato@ki.se

Abstract

Background: Patient-driven innovation in health care is an emerging phenomenon with benefits for patients with chronic conditions, such as cystic fibrosis (CF). However, previous research has not examined what may facilitate or hinder the implementation of such innovations from the provider perspective.

Objective: The aim of this study was to explain variations in the adoption of a patient-driven innovation among CF clinics.

Methods: A comparative multiple-case study was conducted on the adoption of a patient-controlled app to support self-management and collaboration with health care professionals (HCPs). Data collection and analysis were guided by the nonadoption, abandonment, spread, scale-up, and sustainability and complexity assessment tool (NASSS-CAT) framework. Data included user activity levels of patients and qualitative interviews with staff at 9 clinics (n=8, 88.9%, in Sweden; n=1, 11.1%, in the United States). We calculated the maximum and mean percentage of active users at each clinic and performed statistical process control (SPC) analysis to explore how the user activity level changed over time. Qualitative data were subjected to content analysis and complexity analysis and used to generate process maps. All data were then triangulated in a cross-case analysis.

Results: We found no evidence of nonadoption or clear abandonment of the app. Distinct patterns of innovation adoption were discernable based on the maximum end-user activity for each clinic, which we labeled as low (16%-23%), middle (25%-47%), or high (58%-95%) adoption. SPC charts illustrated that the introduction of new app features and research-related activity had a positive influence on user activity levels. Variation in adoption was associated with providers' perceptions of care process complexity. A higher perceived complexity of the value proposition, adopter system, and organization was associated with lower adoption. In clinics that adopted the innovation early or those that relied on champions, user activity tended to plateau or decline, suggesting a negative impact on sustainability.

Conclusions: For patient-driven innovations to be adopted and sustained in health care, understanding patient-provider interdependency and providers' perspectives on what generates value is essential.

(J Med Internet Res 2024;26:e50527) doi:10.2196/50527

KEYWORDS

chronic illness; implementation; adoption; spread; patient-driven innovation; mHealth; mobile health; innovation; health care provider; motivation; interdependency; adaptability

XSL•FO
RenderX

Introduction

Patient-driven innovation in health care is an emerging phenomenon. An example of coproduction and *prosumerism* (where consumers produce what they consume), patient-driven innovations can be seen as the next logical step in health care evolution [1,2] and could fundamentally challenge the essence of what it means to be a professional health care provider. The number of publications about patient-driven or informal caregiver-driven innovations (ie, innovations that are both initiated and driven by patients or informal caregivers or both) mostly concern chronic conditions and have increased substantially in recent years [3]. However, the current literature does not examine the factors that influence the adoption, spread, and scale-up of patient-driven innovations in health care organizations [3]. The paucity of research studies evaluating and reporting the outcomes of patient-driven innovations has been suggested as a potential obstacle to their adoption in health care [4].

Cystic fibrosis (CF) is a complex chronic and genetic condition that affects respiratory and other organ systems [5]. Disease activity varies over time, and treatment requires a high degree of discipline and self-care outside of the clinical microsystem [6,7]. Patient-driven innovations in CF care have resulted in the development and dissemination of mobile health (mHealth) apps that support patients with CF and their caregivers in self-care and information sharing with health care providers [8-11]. Sharing of patient-generated health data has been associated with improved symptom control and quality of life and reduced health care use [12]. The COVID-19 pandemic has further demonstrated the value of and opportunity for patient-generated health data [13], as well as the importance of actively involving patients with CF and caregivers in critical conversations about care and self-care management [14].

However, implementing innovations in health care can be challenging [9,15-18]. The issue is even more pronounced if innovations are created or driven by patients [19], as that can challenge traditional hierarchical values and structures and professional identities. When implementation is not appropriately managed, mHealth apps fail to be adopted, are abandoned, or falter when they are scaled up or spread [20,21]. To increase the ability of hospitals, staff, and patients to adopt technological innovations, implementation approaches need to be anchored in the needs of patients and adapted to the organizational context and the wider system in which the new technology is implemented [21,22]. The level of organization and system complexity will influence the level of adoption by patients and providers [20,21,23-25].

Building on the field of complexity, we see complexity as a characteristic and property that emerges from 3 variables: the number of elements or components, often referred to as nodes (eg, actors and stakeholders); the number of interactions and interdependences between these nodes; and the variation within these nodes and interactions [26-28]. Successful implementation often requires a higher level of alignment between the purpose of the organization and its users [20,21,24]. The risk for failure increases with the level of complexity as health care systems respond to changes in unpredictable and nonlinear ways due to fuzzy organizational boundaries and interconnected actions with other actors that are often difficult to predict or even be aware of [21,29].

With patients and informal caregivers playing an increased role in health care, we need to know more about how their potential contributions can be realized through the adoption of patient-driven innovations in health care. Therefore, the aim of this study was to explain variations in the adoption of a patient-driven innovation among CF clinics. We posed the following research questions:

- How does the adoption of a patient-driven innovation, based on patterns of patient use, vary among clinics?
- What factors influence the level of adoption of the patient-driven innovation over time?

Methods

Study Design

This comparative multiple-case study used mixed (quantitative and qualitative) methods to explain differences in the adoption of a patient-driven innovation among patients with CF at 9 clinics. The Consolidated Criteria for Reporting Qualitative Research (COREQ) guideline [30] was followed in reporting this study (Multimedia Appendix 1). The study is part of the *Patient in the Driver's Seat* research program conducted at the Medical Management Centre, Karolinska Institutet, Sweden. It is a 6-year program that studies how 5 patient-driven innovations are implemented in clinical practice and the daily lives of patients and their networks, 1 of which is the subject of this study.

Theoretical Framework

This study was guided by the framework for theorizing and evaluating nonadoption, abandonment, and challenges in the scale-up (ie, building infrastructure to support full-scale implementation across an organization, locality, or health systems), spread (ie, replicating an intervention somewhere else), and sustainability of health and care technologies (nonadoption, abandonment, spread, scale-up, and sustainability [NASSS]) and the NASSS complexity assessment tool (NASSS-CAT) [29]. We chose the NASSS framework because it was designed to both analyze and prepare for the implementation of technology in health care. Its development process has been well described, lending surface validity to the framework, and it has garnered attention among researchers [20,31,32]. According to the framework, the dynamic interactions that influence the nonadoption, abandonment, scale-up, spread, and sustainability of technological innovations are inherent to the complexity within and between 7 domains:

- Condition: nature of the condition or illness, comorbidities, sociocultural factors
- Technology: material features, type of data generated, knowledge needed to use, supply model
- Value proposition: supply-side value for the developer and for the patient
- Adopters: staff, patients, caregivers

XSL•FO
RenderX

- Organization: capacity to innovate, readiness for technology, nature of adoption/funding decision, extent of change in team routines, work needed
- Wider context: political/policy, regulatory/legal, professional and sociocultural
- Embedding and adaptation over time: scope, organizational resilience [29]

Simple systems consist of a few components that interact in straightforward and predicable ways. Complicated systems have multiple components that interact in a predictable fashion. Complex systems have multiple and intricately related interactions that are constantly changing, unpredictable, nonlinear, and difficult to deconstruct [29]. NASSS and, practically, CAT can be used to distinguish between simple, complicated, and complex elements in the 7 domains [29]. The intention is to identify the multiple influences that are at play; to determine how complexity once identified can be reduced, addressed, or navigated; and to provide information and guidance to the involved actors on how to do so.

The Patient-Driven Innovation

The technology (innovation) studied was a patient-controlled app (named *Genia*) that was originally designed as a patient-facing app to foster self-management. With the addition of a function to generate reports for providers, the app expanded its scope to become a patient-controlled information app for the coproduction of care that places the patient at the center of the decision-making process [10,11]. Founded in 2012 by a father of children with CF, Upstream Dream, which developed the patient-driven innovation, employs individuals with lived experience as patients or informal caregivers of a family member with a rare disease. The innovation was developed in collaboration with the Swedish CF community and upon research conducted at Karolinska Institutet and Dartmouth University. Launched in Sweden in 2015, it was introduced to all CF clinics in the country within a period of a few months. Subsequently, the innovation was piloted (2020-2021) and adopted (2021) as part of routine care in 1 pediatric CF program in the United States and is now the focus of a multicenter study. At launch, the patient-driven innovation was only compatible with the iOS platform. Android support was added later.

Upstream Dream is working to spread the innovation to other clinics in the United States and South America. The innovation has also been tested for use with other chronic medical conditions and has demonstrated improved patient engagement, patient-centered care, and practice-based learning, with the conclusion that the innovation can be recommended for other chronic conditions [31].

The main features of the patient-driven innovation are related to the tracking of symptoms and medications. What differentiates it from a regular personal health record for patients to record disease progression is that information can be shared with care providers, which is why the app was introduced through the clinics to reach patients with CF. Information that patients wish to share with their multidisciplinary clinical care team is submitted via previsit reports in the form of portable document format (PDF) files. Data are integrated into the Swedish National Cystic Fibrosis Quality Registry. In the United States, data are incorporated into the local electronic health record (EHR) [9]. Over time, additional features were added based on input from the clinics. These included an antibiotic use–reporting tool, “Antibiotic Check-in,” to support care and research on the use of new antibiotics and therapies; a medication-monitoring tool for the Orkambi medical treatment; and a “Health Check-in” feature to simplify communication of the information used by clinics for planning meetings and patient visits.

In Sweden, clinics were not charged to use the innovation, whereas in the United States, clinics paid an annual fee. The innovation was free for patients to use in both countries. Patients’ user activity can be seen as an outcome that reflects how well clinics integrate the app in their own care processes.

When the study was conducted, 2 generic health information and communication apps were provided by the regions and in use in the Swedish clinics, “1177” and “Always Open.” Neither app provides a comparable service (patient-controlled communication of disease activity) to the innovation; instead, all 3 apps provide complementary functions that did not interfere with one another and could therefore be used simultaneously. The 1177 app is a patient portal that provides information about illnesses and clinics, booking, and electronic medical records. The Always Open app (Swedish *Alltid Öppet*) was designed as a secure platform for providing remote care services, such as appointment reminders, prescription renewals, and video visits. Its availability was limited to 1 region, so only 2 clinics had access to it.

Study Setting

The study was conducted at 1 CF center in the United States (referred to as clinic A) and all 8 pediatric and adult CF centers in Sweden (referred to as clinics B-I). All but 1 clinic had used the innovation for 7 years (2 years for the most recent clinic). The clinics were small, with 7-10 staff members (typically including pulmonologists, registered nurses (RNs), physiotherapists or respiratory therapists, psychologists or social workers, and dietitians), and focused specifically on CF. All were located at academic medical centers. One clinic in Sweden provided both pediatric and adult care, and the same staff served both patient populations. However, because adoption levels were different between these 2 patient groups, the data were presented separately.

Based on the inclusion of many different implementation settings, we expected different patterns of adoption. This would allow us to develop an empirical basis for both literal replication (cases that predict similar results) and theoretical replication (cases that predict different results for predictable reasons) [32]. The US site was selected because it was the first use case outside Sweden. We expected its inclusion to provide further insights into relevant contextual aspects.

Data Collection

Quantitative data were collected to assess the level of adoption based on the number of active users (ie, patient activity). Since the innovation was designed to improve patient-provider communication, patient activity levels can be seen as an indicator of adoption. Anonymous user activity data for each

XSL•FO
RenderX

clinic were extracted from Upstream Dream's monthly reports on user activity for the period from March 2015 (when the innovation was implemented at the first clinic) to March 2022. At clinic A, where the innovation was implemented last, the user activity period was from February 2020 to March 2022.

Qualitative data were collected to understand how the innovation was perceived and used based on the NASSS domains and to identify specific interventions or factors that may have influenced user activity. A semistructured interview guide was developed in both Swedish and English (Multimedia Appendix 2). Questions were designed to elicit information about NASSS domains and their degrees of complexity [21], as well as to capture information about app integration into care processes. A precursor to the interview guide was first tested in a separate study [33] and revised based on that experience. Further input was sought from Upstream Dream to ensure that we would capture a holistic understanding of how the innovation was used in daily practice with respect to the different organizational contexts and potential differences in complexity. The interview guide was then piloted twice, with minor changes made to enhance the clarity of the questions in only the Swedish version.

Purposive and snowball sampling strategies were used to identify participants with knowledge of and personal experience with the innovation and who represented the professions involved in CF care. Upstream Dream's clinical coordinators contacted the staff coordinators at each CF center on our behalf, who then connected us with the staff who expressed an interest to participate in the study. We contacted these individuals via email and followed up by telephone. In total, 21 participants were interviewed, including 2 key members of Upstream Dream, one of whom was interviewed twice (first, to develop a contextual understanding of the innovation and the supplier perspective and then to provide additional information and clarity after the adopter interviews), RNs, physicians, respiratory therapists or physiotherapists, a psychologist, a dietician, and a researcher involved in the development, implementation, and evaluation of the innovation in 1 of the clinics, without being a health care provider. This role distribution reflected typical staff distribution at the clinics. The CF centers involved had a small number of employees, which limited the number of potential participants. The number of participants was also limited by the high workloads brought on by the COVID-19 pandemic. Thus, the participants involved represented between 28% and 40% of working staff. Participants' experiences with the innovation ranged from 3 months to 7 years.

Interviews were conducted online via the Zoom videoconferencing system due to the ongoing COVID-19 pandemic, in Swedish or English, and lasted 30-60 minutes. Interviews were conducted by 2 authors with training in qualitative research, with support from a senior researcher. The interviewers had no prior relationship with the participants. Audio was digitally recorded and transcribed verbatim.

Data Analysis

Quantitative data on user activity were analyzed using descriptive statistics. We calculated the percentage of active end users per month among the total number of patients and the percentage of active users on average over the entire period. Active users were defined as end users who logged on and used some of the basic app features within a 6-month period. As clinic A first introduced the patient-driven innovation as a pilot study, the percentage of active users for that period was calculated based on the pilot sample size and after the end of the pilot was calculated based on the clinic size. The descriptive statistics (ie, maximum end-user activity and average over the entire period) provided insights into the level of adoption. We defined low (<25%), middle (25-50%), and high (>50%) activity levels based on the maximum percentage of end-user level achieved.

To put user activity data into an organizational context, we complemented the descriptive data with statistical process control (SPC) charts. SPC charts were created for each clinic to identify whether and when statistically significant changes in app adoption levels occurred and whether these were sustained. SPC makes it possible to determine whether a change is a matter of chance (ie, common-cause variation) or due to a specific happening or intervention (ie, special-cause variation) [34-36]. We used P charts with the following rules to identify special-cause variation: rule 1/3 sigma violation (1 point +/– the upper control limit/lower control limit [UCL/LCL], with the control limits set to +/–3 sigma), rule 2/shift (8 successive consecutive points above or below the centerline), rule 3/trend (6 or more consecutive points steadily increasing or decreasing), and rule 5/hugging the centerline (15 or more consecutive points within +/–1 sigma on either side of the centerline) [37].

Qualitative data were analyzed using content analysis (directed and inductive) [38], process mapping, and complexity analysis. Interview transcripts were read through repeatedly to develop familiarity and then subjected to directed qualitative content analysis [39]. Two authors together identified meaning units, which they abstracted to condensed meaning units and added as "sticky notes" to the MIRO online whiteboard for visual collaboration, where they were directed to 1 of the NASSS domains (ie, categories). The condensed meaning units were labeled with descriptive codes. Where the codes did not fit the framework, additional subcategories were created through traditional inductive content analysis [39]. All codes and the categorization were reviewed and corroborated by 2 other authors. During the analysis, interview data from suppliers and HCPs were kept separate, and the former were used solely to provide contextual information about specific interventions and factors that were integrated in the SPC charts (eg, timeline of interventions that could have influenced user activity).

To support the cross-case comparison, the traditional approach to reporting qualitative analyses was then transformed into tabulated form based on the original NASSS framework and expanded to include the additional subcategories.

Process maps were created for each clinic based on interview data to illustrate how the innovation was integrated into clinical work processes. The process maps and the qualitative analysis were shared with participants in an informant validation process. Five clinics made small adjustments. The suppliers provided additional corroborating feedback, where needed, for the clinics that did not respond.

XSL•FO
RenderX

We diagnosed complexity levels (see the definitions of simple, complicated, complex in the theoretical framework) of the NASSS domains for each clinic by analyzing interview data using the NASSS complexity table (Multimedia Appendix 3) and the NASSS-CAT Short survey, which was specifically designed to assess and differentiate between complexity levels [29]. Both analyses were combined to generate the complexity assessment. Although the survey was originally intended to spark a reflective discussion, we used the reflective discussion created in the interview setting to diagnose the level of complexity. For the qualitative analyses and the complexity assessment, conflicting interpretations were discussed until consensus was reached.

Qualitative data were first analyzed case by case but then triangulated with quantitative data to develop initial explanations of variation in adoption. These were then tested against the data and the analyses to identify those domains or interacting domains that could explain the observed patterns. In a process akin to modified analytic induction [38], when falsifying evidence was found, the explanatory model was dropped. This process involved vigorous discussion and iterative cycles to narrow and refine the explanatory models that are presented in the cross-case comparison.

Ethical Considerations

Ethical approval was obtained from the Swedish Ethical Review Authority (approval number 2019-03849). The study followed the Swedish Research Council's ethical principles for humanities and social science research. Participants were informed orally and in writing about their rights and what study participation would entail. Written and verbal consent was obtained from all participants prior to commencement of interviews. Quantitative data were completely anonymous. Qualitative data were pseudonymized and deidentified prior to coding and analysis. In presenting the findings, we made efforts to maintain participants' privacy and confidentiality, referencing HCPs only with their pseudonym identifiers and clinic letters (eg, "HCP01, clinic A"). No participant received any compensation for participating in the study.

Results

In this section, we present the triangulation of the qualitative and quantitative data first with user activity levels, and then a cross-case comparison based on the complexity assessment.

Participant Details

In total, 21 participants (n=16, 76.2%, women and n=5, 23.8%, men) were interviewed. Of the 21 participants, 9 (42.9%) were RNs, 5 (23.8%) physicians, 4 (19%) respiratory therapists or physiotherapists, 1 (4.8%) psychologist, 1 (4.8%) dietician, and 1 (4.8%) researcher.

In 2 of the clinics (D and H), we were able to interview only 1 (4.8%) participant each; the remaining 7 clinics were represented by 2 (9.5%) or more participants: clinic A, n=3 (14.3%); clinic B, n=4 (19%); clinic C, n=4 (19%); clinic E, n=2 (9.5%); clinic F, n=2 (9.5%); clinic G, n=2 (9.5%); and clinic I, n=2 (9.5%).

User Activity Level

The end-user activity levels showed that the innovation was adopted by patients at all clinics; there was no evidence of nonadoption. Based on the maximum end-user level achieved, we found 3 clusters: high-adoption clinics A (maximum n=40, 65%), B (maximum n=60, 58%), and C (maximum n=66, 96%); medium-adoption clinics D (maximum n=125, 35%), E (maximum n=30, 47%), and F (maximum n=60, 25%); and low-adoption clinics G (maximum n=120, 19%), H (maximum n=116, 23%), and I (maximum n=155, 16%).

To further explore user activity levels over time, SPC charts were created (Figure 1) for each clinic to plot the user activity level over 81 months. In the SPC charts, specific interventions or events retrieved through the interviews, which may have influenced user activity levels over time, are indicated as vertical dotted lines and labeled as a-g. These included the introduction of new features, research-related activity, and the COVID-19 pandemic.

XSL•FO
RenderX

Figure 1. SPC analysis of end-user activity at the different clinics organized as high (top row), middle (middle row), and low (lower row) end-user adoption. The UCL and LCL were defined as +/– 3 SDs from the centerline [37]. The y-axis uses different scales to be able to better discern the trends over time. The blue circles represent common cause variation, whereas the red squares represent special cause variation, with numbers indicated according to SPC rules. LCL: lower control limit; SPC: statistical process control; UCL: upper control limit.

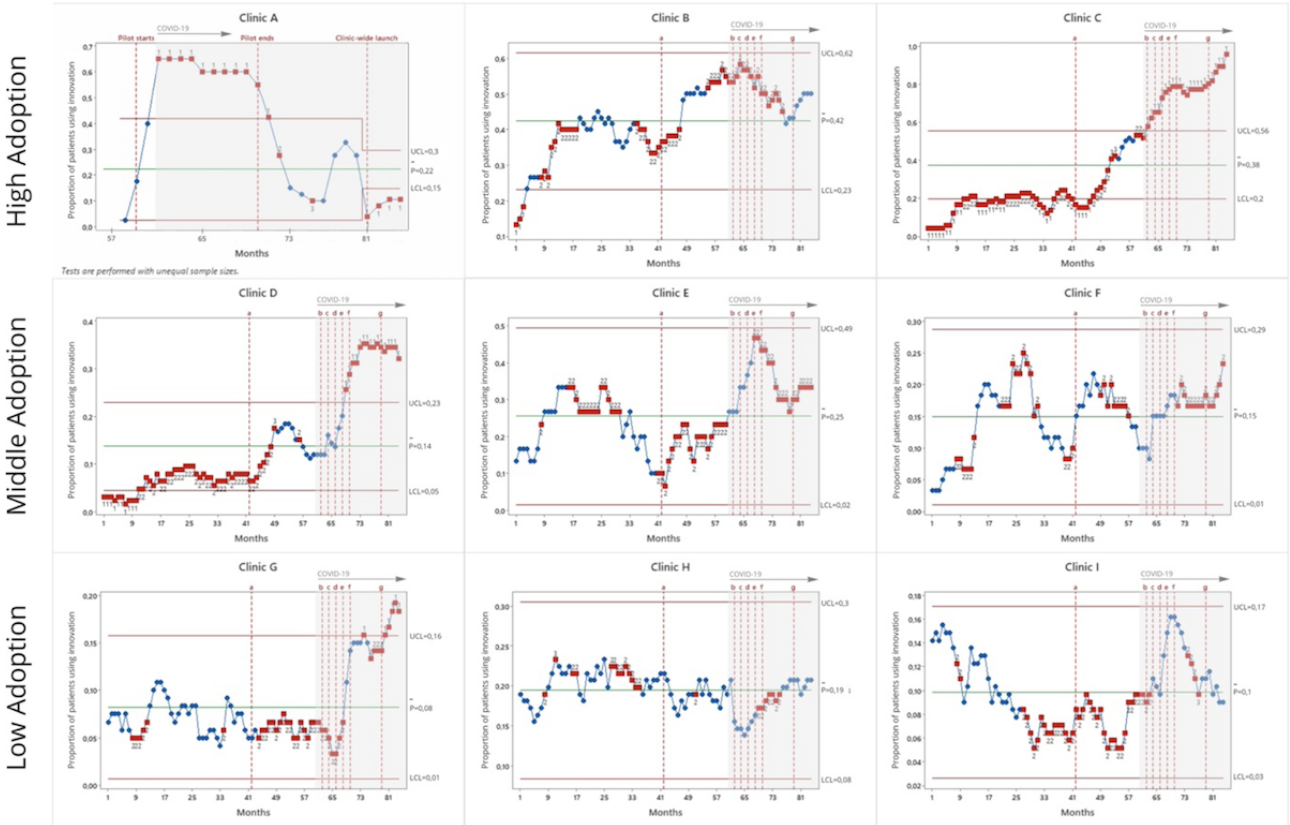

New features included the launch of Orkambi medication monitoring (intervention a) in month 42, which likely explains the increased use among 5 clinics (B-F), particularly among pediatric clinics. Antibiotic Check-in was launched in Swedish clinics in month 62 (intervention b), which was followed by a campaign (intervention d), which could explain the increased activity in 7 (87.5%) of the 8 Swedish clinics (C-I). The introduction of an Android-compatible version in month 68 (intervention e) opened the innovation up to all patients and caregivers with a smartphone or tablet. Health Check-in (intervention f) may have contributed to the observable increase around month 70 (clinics C, D, F, and H).

The influence of research-related activities was mainly identified in clinic A, in which there was a rapid increase in activity starting month 60, which plateaued. This corresponded to when the innovation was first introduced in clinic A as a year-long single-group pilot study (n=40 participants, pre-post design) [40]. When the patient quota for the pilot was met (first vertical dotted line, Figure 1), 40% of the invited patients were active users. Activity declined after the pilot ended but began to rise again (month 81) when the decision was made to launch the innovation clinic-wide. The consistency of this increase continued after the last point recorded in Figure 1 until activity was halted again in preparation for a new study (data not included in the SPC). Clinic C launched a digitization research project (intervention c), which required participants to use the innovation, and later launched a second digitization project (intervention g) in month 79.

The COVID-19 pandemic also appeared to influence user activity levels. Clinic A partially transitioned to telehealth visits during the pandemic, which made the Health Check-in feature desirable as patients could upload photos and other information. This may have contributed to increased user activity. In contrast, for clinics E, F, H, and I, a deterioration (months 70-78) was concomitant with an active choice not to focus on the innovation due to strained resources.

Cross-Case Comparison: Complexity Assessment Linked to Adoption Level

Specific characteristics of the clinics related to each NASSS domain gleaned from the interviews and the process maps are presented in Multimedia Appendix 4. In terms of complexity, all clinics viewed the nature of the *condition*, *technology*, and *wider context* domains as complicated (Figure 2). Differences were found in the *value proposition*, *adopter system*, *organization*, and *embedding* domains. The cross-case comparison presented next was organized around levels of adoption, integrating data from the SPC analysis and exploring differences and similarities in the NASSS domains.

XSL•FO
RenderX

Figure 2. Complexity assessment for NASSS domains grouped by adoption level based on maximum end-user activity. Green circles represent simple, blue labyrinths complicated, and red network symbols complex. *Minimum activity levels were not reported because data collection started when there were 0 active users for all clinics. **For clinic A, n was calculated based on active users over the pilot study. NASSS: nonadoption, abandonment, spread, scale-up, and sustainability.

NASSS domain	High adoption			Middle adoption			Low adoption		
	Clinic A	Clinic B	Clinic C	Clinic D	Clinic E	Clinic F	Clinic G	Clinic H	Clinic I
1. Condition	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated
2. Technology	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated
3. Value proposition	Simple	Simple	Simple	Complicated	Complicated	Complicated	Complex	Complex	Complex
4. Adopter system	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complex	Complex	Complex
5. Organization	Complicated	Complicated	Complicated	Complex	Complex	Complex	Complex	Complex	Complex
6. Wider context	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated	Complicated
7. Embedding	Simple	Complicated	Complicated	Complicated	Complex	Complex	Complex	Complex	Complex
Maximum active end users*, n/N (%)	26/40 (65)	35/60 (58)	63/66 (95)	44/125 (35)	14/30 (47)	15/60 (25)	23/120 (19)	27/116 (23)	25/155 (16)
Mean active end users over the study period, n (%)	21 (22)**	25 (42)	25 (38)	17 (14)	8 (25)	9 (15)	10 (8)	23 (19)	15 (10)

High-Adoption Cluster (Clinics A-C)

The first group perceived the value proposition as simple, the adopter system and organization as complicated, and embedding as either simple or complicated. The staff had a shared perception that there is a clear benefit to using the innovation (ie, the innovation-facilitated meeting of patient and provider creates value in health care for this chronic condition). This was aligned with their view that value in health care for this chronic condition is created in the patient-provider meeting. Although going through reports lengthened the previsit planning process, the innovation was perceived to save time through more concise patient visits.

> *Just doing that, [opening reports in weekly team meetings] will change it a little bit, it is going to lengthen our [team] meeting a bit (…) So it helps everybody to be prepared ahead of time and hopefully make the visit more concise.* [HCP14, clinic A]

For clinics B and C, where Antibiotic Check-in was used, these data were highly valued for research purposes and research was seen as an integral component of high quality care. For clinic C, this was corroborated by SPC analysis (intervention c).

The adopter system was viewed by clinics A-C as complicated. Patient-controlled data were viewed as integral to previsit planning, suggesting a view of patients as active contributors to the cocreation of care and of the staff as having a responsibility to ensure that patients contributed with these data.

> *Way more attention is being paid now to this patient-generated data rather than the core clinical data. I mean, we still look at lung function and microbiology and what not, but patient-reported symptoms and outcomes are a bigger part of the discussion now, as well as patient preferences and…and goals and what people want to focus on. So that is 1 change that has happened.* [HCP15, clinic A]

This person-centered culture was reflected in the organization domain, which the staff perceived as complicated, as organizational routines and care pathways needed to be flexible to reinforce the importance of patient input and the use of the innovation as an integral source of information for previsit planning. For example, if a patient had either not downloaded the innovation or submitted a report, clinics A and C had routines to complete those tasks together with the patient (Multimedia Appendix 4). Even though these clinics saw a clear value in the innovation, they still expressed the difficulties of cultural change and that it requires key drivers.

> *We often do not have the personnel for this. It needs to be carried out, in addition to the usual work. And there has to be someone driven to be able to push these questions forward.* [HCP01, clinic C]

Embedding was seen as simple or complicated based on how mature the routines for incorporating user information (the reports) into the patient pathway were. In clinic B, the continual increase in adoption observed in Figure 1 could be explained by the clinic's routines of using the innovation with the whole clinical team, as well as its regular communication with the developer. For example, clinic A had clear plans and used collective reflection among the staff for embedding the technology in the short and the long term.

> *The good thing is that the way we have set up this pragmatic trial is that they are [clinicians] not required to do anything extra than they normally do, so they are not required to log in on a separate platform or a dashboard or anything extra. Everything is embedded and integrated, which makes it an easier sell.* [HCP15, clinic A]

XSL•FO
RenderX

Clinics A-C saw a match between what they valued (the patient-provider interaction) and the value proposition of the innovation because it improved the quality and efficiency of the patient-provider encounter. These clinics had existing and further developed their routines to ensure and reinforce patient use of the innovation. Moreover, they behaved as if they "co-owned" the innovation either by conducting research studies or by taking responsibility for patient training. In the adopter system, most of the staff had clearly defined roles in relation to the use of the innovation, as well as established routines for group reflection. Participants valued patients' role in the cocreation of care.

Middle-Adoption Cluster (Clinics D-F)

The second group perceived the value proposition and adopter system as complicated and organization and embedding as complex. The value proposition was seen as complicated because the technology's desirability was contested, and the business case for adoption was deemed unclear. The participants experienced a mismatch between the condition and technology that reflected itself in their view of the value proposition: clinics D and F (adult clinics) believed that the technology better serves pediatric patients, who typically receive outpatient care and need to report symptoms, whereas adult patients with CF in Sweden are often hospitalized and their symptomatology easier to track. The staff felt that the graphical user interface is less appropriate for adults.

Regarding the adopter system, clinics E and F questioned the appropriateness of care providers in a public institution to convince their patients to use a product from a private company.

> *If a patient is completely on top of it with their treatment and medications and everything, then it feels a little like a car salesman if I am to try and sell something that is not a directive of the hospital. And that role we all feel is a bit annoying. But if you can show a direct benefit to using [the innovation], then it feels good.* [HCP12, clinic F]

Organization was seen as complex because of difficulties integrating the innovation in the workflow. For example, patient-generated reports were scheduled on the weekly agenda at clinics E and F but seldom discussed. Work routines also included contact with the supplier, who felt that clinic D demonstrated heightened commitment. This was mirrored in the SPC data, which showed a clear increase in adoption toward the end of the data collection period. However, patient workflow processes were not established. Adoption for all 3 clinics was largely the responsibility of individual clinicians and, due to individual levels of enthusiasm or work practices, gave rise to variation (Figure 1). This impacted patients.

> *We are a center where not everyone is on the team in the same way. I think that is a factor for whether, depending on which doctor one meets, there will be a question about [the innovation] or not (…) It becomes person dependent in a crazy way.* [HCP19, clinic E]

It impacted staff as well.

> *If everyone did it, there would not be any extra work, but since we cannot manage to get everyone here doing it, I end up trying to take the main responsibility (…) Sometimes I sit and go through [the reports] (…) So, I have a bit of extra work, but it is too difficult to create a routine for such a thing if not all patients use it, then it can become forgotten.* [HCP18, clinic H]

In clinics E and F, adoption relied heavily on champions, which could explain the decrease observed in months 26-43 (Figure 1), which was also compounded by a severe staff shortage. Embedding was deemed complex at clinics E and F as they were forced to prioritize resources due to the COVID-19 pandemic, which shifted focus away from the innovation and was reflected in the user activity decrease. Moreover, they felt the questions prior to the introduction of Antibiotic Check-in and Health Check-in were too generic.

Low-Adoption Cluster (Clinics G-I)

The third group perceived the value proposition, adopter system, organization, and embedding as complex. Like clinics D-E, these providers had a negative experience of advocating for a privately owned app. They were among the first to adopt the innovation, at a point where app features were minimal and the app was limited to 1 platform. They felt that this effort to get patients to use an underdeveloped app negatively impacted the patient-provider relationship. The SPC analysis for clinics H and I (Figure 1) confirmed that early adopters struggled to maintain adoption as the user activity level either remained stable (clinic H) or decreased (clinic I). The addition of new features, although potentially increasing the value of the innovation, was offset by previous experiences, which had worn them out.

> *It is now actually that you should start trying to get patients to use [the innovation]. But in [this region], we have sort of worn ourselves out because we already did it 4-5 years ago.* [HCP18, clinic H]

The innovation was seen as undesirable by most of the staff at these clinics, and they perceived their patients were equally uninterested. Several providers mentioned that patients do not want to "have their illness in an app," as symptom tracking can become a reminder of how sick one has been and add yet another task, when patients with CF are already "drowning in health care" (HCP06). Some also felt adults have a hard time changing their ways. These opinions demonstrated a mismatch between the value of the technology and the needs and challenges related to the condition.

Adopter systems were deemed complex because, although at all clinics the staff had to learn new skills, the staff at clinics G-I felt the innovation poses a threat to their professional identity and scope of practice and felt patients find the innovation challenging. Clinicians preferred digital technologies available from the regions, rather than from a private company. Here, they referred to 1177 and Always Open as examples of such tools, which they also perceived served a clearer purpose. Clinicians were skeptical of how the app was introduced and that it did not come from within the clinic.

XSL•FO
RenderX

> *It was not we as providers who went to an app developer and said, "We want a tool." Rather, it was they who came from the outside and said, "You need a tool, and here it is."* [HCP03, clinic G]

Clinicians at pediatric clinics believed there is a particular need for medication tracking among adult patients, whereas the staff at adult clinics expressed the opposite. Past experiences with other technological interventions and how well they were received also influenced how hesitant or open clinics were to innovation.

Organizations also demonstrated complexity; none of clinics G-I had integrated the innovation into their clinical workflow, and the innovation was used on an individual rather than on a team basis, which put pressure on the single user (Multimedia Appendix 3). This was reinforced by the special cause variation observable in clinic G (consecutive points below/above the center line, months 69-83), where the activity dipped down before shooting up and could be linked to a champion staff member who led the adoption but left temporarily before returning. The clinics saw funding as a barrier to implementing new technologies, including the innovation. Clinic I mentioned a severe resource pressure, including hiring stops (frozen posts), which halted the use of the innovation, especially under the stress of the pandemic. The embedding system was complex due to the clinic's inability to adapt the innovation use to critical and unforeseen events (eg, the COVID-19 pandemic).

Discussion

Principal Findings

In this study, the adoption of a patient-driven innovation was studied using a complexity-based framework and tools for the introduction of technology in health care, NASSS-CAT. The innovation was developed as a patient-controlled information app to support the self-management of CF and communicate disease related–activity with health care providers. Although we found no evidence of nonadoption or clear abandonment of the app, distinct patterns of innovation adoption were discernable based on user activity data (ie, low, medium, and high adoption). The perceived value proposition of the technology and the experienced complexity were associated with different levels of adoption. Research activity and the introduction of new app features positively impacted adoption. In clinics that adopted the innovation early or those that relied on champions, user activity tended to plateau or decline, suggesting a negative impact on sustainability.

Perceived Complexity Influencing Adoption

There was little variation between clinics regarding perceptions of the condition and technology domains. Differences in complexity were seen within the value proposition, adopter system, organization, and embedding and adaptation domains. The more complex these domains were perceived to be, the lower the level of adoption was.

The perceptions of value identified in this study demonstrate a patient-provider interdependency (ie, both the provider and the patient must value and use the patient-driven innovation, or else it will lead to a downward spiral of abandonment). As Floch et al [8] found, "Self-management enfolds a collaboration between patients and [health care providers]." This suggests that studying providers' experiences of using patient-driven innovations can be an important perspective as patients' and providers' behaviors are 2 sides of the proverbial same coin. One could expect that a patient-driven innovation, as an example of prosumerism, would entail a higher level of acceptance or be more highly valued than an innovation external to the clinical context.

All clinics seemed to agree that there is a clear need to focus clinic visits on what patients value. Where the value proposition was perceived as simple, HCPs saw the innovation as a solution, or at least worth testing as a solution. Although going through reports lengthened weekly previsit planning, the staff saw that the innovation enables shorter and more concise patient visits [9]. This mirrors the findings for another patient-centered care app for patients with CF [8]. Clinics with lower adoption described the innovation as interfering with the patient-provider interaction, since they believed they already knew their patients well due to the chronic nature of CF and perceived the app as an affront to their professionalism. The time providers spent with their patients to elicit this information was seen as a demonstration of how they valued their patients. The difference between the 2 interpretations of value could be paraphrased, from the perspective of the professional, as "we value our patients' time" versus "we value our time with patients." In the former, providers focus on what matters to the patient with the help of the innovation; in the latter, providers try to find out what is the matter with the patient through a person-to-person conversation without the innovation and the information it provides.

In terms of working with the adopter system and organization, our findings suggest a need to work with the context to integrate a new patient-driven innovation in health care. In this respect, despite prosumerism, this patient-driven innovation is not different from other innovations in health care [33]. Working with a broader group of adopters, not just champions, and integrating the patient-driven innovation with care processes appeared to facilitate adoption. If context is not addressed by suppliers, there is a risk that the perceived value of the innovation will be influenced by the perceived complexity of the setting. This lived experience of complexity could not be explained in terms of differences in the medical condition or technology. Instead, it appeared to be more dependent on how care processes had evolved and the human (in)ability to *deal* with variation, uncertainty, complexity, and ambiguity in everyday work life: complexity was in the eye of the beholder.

Co-ownership and Trust

When the app was first launched, the company took responsibility for staff and patient education, with the intention of having as little disturbance as possible in the clinics. However, clinics that took shared responsibility for the patient-driven rollout had higher and sustained adoption levels. As studies in behavioral economics have demonstrated [41], a higher level of perceived co-ownership leads to a higher evaluation of the object in question. Co-ownership invites the staff to learn and understand more about the patient-driven innovation, which could explain why merely relying on

XSL•FO
RenderX

champions can be associated with lower adoption and responsibility for patient training and research with higher adoption.

Most clinics raised questions about the financial motives of the company behind the patient-driven innovation. Two clinics questioned whether it is appropriate for them as medical providers to "sell" a product from a private company. Particularly, during the early rollout, the staff felt that pushing a premature version of the innovation on patients is a violation of their professional integrity. The distrust of a private business overshadowed the patient-driven prosumer nature of the innovation that should have engendered trust. This distrust could be related to a commonly held negative view of privatization in Sweden or may indicate that the company's patient origins had not been communicated clearly.

Utility of NASSS-CAT

This study used NASSS-CAT as a framework and tool to analyze the implementation of a patient-driven innovation. We found it helpful to characterize single domains but less suitable to explore the interaction between domains. The framework is innovation centric, which risks generating a bias that values innovation per se regardless of its suitability for addressing the challenge at hand or the purpose of the hosting clinic or organization.

Another issue relates to the essence of the complexity captured with CAT. Our findings support the basic tenet of NASSS-CAT that adoption is inversely related to the level of complexity. However, a closer look into the data suggests that what was captured may have been individuals' perceptions of complexity related to lived experiences rather than the actual contextual complexity related to the level of interdependency. This corroborates preliminary observations of the original CAT [29]. Thus, results may be more indicative of the maturity of the complexity mindset of individuals [42] rather than the contextual complexity itself. To develop adoption strategies based on such data would be tantamount to developing treatment strategies based on an incorrect diagnosis.

A more accurate assessment of contextual complexity could be achieved by exploring the level of agreement between understanding the challenge and the proposed response [25,26,43]. These 2 questions are simpler to ask and evaluate to generate actionable data: less agreement would suggest a higher degree of complexity and therefore a need for strategies that facilitate learning [43,44]. More direct implementation works when things are simple (ie, greater agreement).

Strengths and Limitations

Directed content analysis inherently has some biases due to the use of a preselected theory [39]. However, several measures were taken to mitigate this limitation. For example, we used open-ended questions in the interview guide, and multiple authors were involved in all the steps of the qualitative analysis.

There was variation in the number of participants per clinic, which could have influenced the analysis. Clinics where the innovation was perceived more positively and used to a greater degree also tended to yield more interviews. This difference may reflect both resource availability and the perceived value of the innovation. Overall, the number of interview participants was limited by the number of employees at each clinic and the COVID-19 pandemic. The literature suggests that theoretical saturation is usually attained at around 12 interviews [45]. This exceeds the number of staff members working with the innovation in most of the clinics we studied. Despite our small sample size, the participants' specific knowledge about the innovation and the care processes in each clinic contributed to strengthening information power [46]. Moreover, the triangulation of qualitative and quantitative data strengthened the trustworthiness of the findings.

Not all clinics provided feedback on the process maps, although all were given the opportunity. To further improve trustworthiness, process maps were checked again against the transcripts and reviewed by the developer's clinic coordinator, who had insight into the clinics and staff.

The total number of potential users was limited because the innovation was first released in an iOS version only. As we did not have data on the proportion of potential users who had an iOS smartphone, we may have overestimated the number of potential users, in particular prior to the release of the Android version.

This study did not examine patient outcomes related to the innovation or satisfaction with the innovation. Interviewing patients and informal caregivers would add valuable perspectives to that of providers.

Conclusion

Patient-driven innovations could be highly relevant for health care, but their adoption has seldom been explored from the perspective of health care providers. We found that providers play a significant role in the adoption of patient-driven innovations in health care: patients cannot do it alone. Health care providers who make an effort to reduce the perceived complexity in the adoption process, simplify their processes, take co-ownership of the innovation, and work on its adoption and improvement as a team, rather than relying on change champions, improve their capability to support the adoption and sustainability of innovative ideas developed by patients. For patient-driven innovations to be adopted and sustained in health care, understanding patient-provider interdependency and providers' perspectives on what generates value is essential.

Acknowledgments

The authors acknowledge the contributions of Rafiq Muhammad to the statistical process control analyses and the valuable input of Henna Hasson and other members of the Patient in the Driver's Seat research program to the manuscript draft. A hallmark of the research program is that the team, which includes patients and patient-innovators, has involved co-creative efforts throughout

XSL•FO
RenderX

all aspects of the research process, including meta-level discussions on how our different perspectives influence the research process.

This work was financially supported by the Swedish Research Council for Health, Working Life and Welfare (grant number 2018-01472). The funders had no involvement in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; and in the decision to submit the paper for publication.

Data Availability

No additional data are available. The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

CS, PM, and AH were responsible for conceptualization, with input from JLL, MMC, CW, JB, GO, and members of the Patient in the Driver's Seat research program. The interview guide was developed by JLL, MMC, JB, CS, and PM. Data collection was conducted by JLL and MMC, with support from CS. Qualitative data analysis was conducted by JLL, MMC, and CS. Statistical process control (SPC) analysis was conducted and interpreted by JLL, MMC, PM, CS, and Rafiq Muhammad. JB and CW provided input on the interpretation of the results. SPC graphs were compiled by JLL. The manuscript was drafted by CS, JLL, MMC, JB, and PM. The manuscript was revised by CS, JLL, MMC, and PM, with input from JB, AH, CW, and GO. AH and JB were involved in the revision of the manuscript and in the decision to submit the paper for publication. CW, CS, PM, and AH were coapplicants of the funding grant, with Professor Henna Hasson as program lead. CS and PM served as principal investigators.

Conflicts of Interest

AH and JB are employees of Upstream Dream, the company that developed the innovation. Upstream Dream was not involved in the study design but did facilitate data collection, and JB helped interpret the data based on knowledge of the participating CF clinics. To mitigate bias, interviews and preliminary analysis were conducted without participation of the supplier. AH and JB were interviewed, and their transcripts were subjected to the same data analysis methods, kept separate from the analyses of the other participants, and then used for clarification purposes for the statistical process control timeline analysis and complexity analysis.

Multimedia Appendix 1
Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist.
[DOCX File , 1243 KB - jmir_v26i1e50527_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File , 37 KB - jmir_v26i1e50527_app2.docx]

Multimedia Appendix 3
Nonadoption, abandonment, spread, scale-up, and sustainability (NASSS) domain areas and case-specific descriptions.
[PDF File (Adobe PDF File), 133 KB - jmir_v26i1e50527_app3.pdf]

Multimedia Appendix 4
Process maps.
[PDF File (Adobe PDF File), 707 KB - jmir_v26i1e50527_app4.pdf]

References

1. Batalden M, Batalden P, Margolis P, Seid M, Armstrong G, Opipari-Arrigan L, et al. Coproduction of healthcare service. BMJ Qual Saf 2016 Jul;25(7):509-517 [FREE Full text] [doi: 10.1136/bmjqs-2015-004315] [Medline: 26376674]
2. Tapscott D, Williams A. Wikinomics: How Mass Collaboration Changes Everything. New York: Portfolio; 2006.
3. Reinius M, Mazzocato P, Riggare S, Bylund A, Jansson H, Øvretveit J, et al. Patient-driven innovations reported in peer-reviewed journals: a scoping review. BMJ Open 2022 Jan 24;12(1):e053735 [FREE Full text] [doi: 10.1136/bmjopen-2021-053735] [Medline: 35074818]
4. Dahlberg M, Lek M, Malmqvist Castillo M, Bylund A, Hasson H, Riggare S, et al. Objectives and outcomes of patient-driven innovations published in peer-reviewed journals: a qualitative analysis of publications included in a scoping review. BMJ Open 2023 Jun 01;13(6):e071363 [FREE Full text] [doi: 10.1136/bmjopen-2022-071363] [Medline: 37263703]
5. Elborn JS. Cystic fibrosis. Lancet 2016 Nov;388(10059):2519-2531. [doi: 10.1016/s0140-6736(16)00576-6]

XSL•FO
RenderX

6. Sawicki GS, Sellers DE, Robinson WM. High treatment burden in adults with cystic fibrosis: challenges to disease self-management. J Cyst Fibros 2009 Mar;8(2):91-96 [FREE Full text] [doi: 10.1016/j.jcf.2008.09.007] [Medline: 18952504]
7. Ziaian T, Sawyer MG, Reynolds KE, Carbone JA, Clark JJ, Baghurst PA, et al. Treatment burden and health-related quality of life of children with diabetes, cystic fibrosis and asthma. J Paediatr Child Health 2006 Oct 08;42(10):596-600. [doi: 10.1111/j.1440-1754.2006.00943.x] [Medline: 16972965]
8. Floch J, Vilarinho T, Zettl A, Ibanez-Sanchez G, Calvo-Lerma J, Stav E, et al. Users' experiences of a mobile health self-management approach for the treatment of cystic fibrosis: mixed methods study. JMIR Mhealth Uhealth 2020 Jul 08;8(7):e15896 [FREE Full text] [doi: 10.2196/15896] [Medline: 32673237]
9. Longacre M, Grande S, Hager A, Montan M, Bergquist RP, Martensson M, et al. Clinical adoption of mHealth technology to support pediatric cystic fibrosis care in Sweden: qualitative case study. JMIR Pediatr Parent 2018 Dec 10;1(2):e11080 [FREE Full text] [doi: 10.2196/11080] [Medline: 31518297]
10. Rutland SB, Bergquist RP, Hager A, Geurs R, Mims C, Gutierrez HH, et al. A mobile health platform for self-management of pediatric cystic fibrosis: qualitative study of adaptation to stakeholder needs and integration in clinical settings. JMIR Form Res 2021 Jan 26;5(1):e19413 [FREE Full text] [doi: 10.2196/19413] [Medline: 33496667]
11. Hager A, Lindblad S, Brommels M, Salomonsson S, Wannheden C. Sharing patient-controlled real-world data through the application of the theory of commons: action research case study. J Med Internet Res 2021 Jan 19;23(1):e16842 [FREE Full text] [doi: 10.2196/16842] [Medline: 33464212]
12. Tiase VL, Hull W, McFarland MM, Sward KA, Del Fiol G, Staes C, et al. Patient-generated health data and electronic health record integration: protocol for a scoping review. BMJ Open 2019 Dec 17;9(12):e033073 [FREE Full text] [doi: 10.1136/bmjopen-2019-033073] [Medline: 31852707]
13. Shaw RJ, Boazak M, Tiase V, Porter G, Wosik J, Bumatay S, et al. Integrating Patient-Generated Digital Health Data into Electronic Health Records in Ambulatory Care Settings: An Environmental Scan. Report No.: AHRQ Publication No. 21-0031. Rockville, MD: Agency for Healthcare Research and Quality; 2021.
14. Wannheden C, Hasson H, Hager A, Karlgren K, Pukk Härenstam K. Now what? Collective sensemaking and sensegiving in the cystic fibrosis community in Sweden during the initial phase of the COVID-19 pandemic. Health Commun 2023 Dec 17;38(14):3102-3112. [doi: 10.1080/10410236.2022.2134705] [Medline: 36250348]
15. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
16. Standing C, Standing S, McDermott M, Gururajan R, Kiani Mavi R. The paradoxes of telehealth: a review of the literature 2000–2015. Syst Res 2016 Dec 01;35(1):90-101. [doi: 10.1002/sres.2442]
17. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]
18. Greenhalgh T, Papoutsi C. Spreading and scaling up innovation and improvement. BMJ 2019 May 10;365:l2068 [FREE Full text] [doi: 10.1136/bmj.l2068] [Medline: 31076440]
19. Riggare S. Patient researchers — the missing link? Nat Med 2020 Oct 07;26(10):1507-1507. [doi: 10.1038/s41591-020-1080-4] [Medline: 33029015]
20. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Analysing the role of complexity in explaining the fortunes of technology programmes: empirical application of the NASSS framework. BMC Med 2018 May 14;16(1):66 [FREE Full text] [doi: 10.1186/s12916-018-1050-6] [Medline: 29754584]
21. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
22. Pettigrew A, Whipp R. Managing Change for Competitive Success. Oxford, UK: Blackwell Publishing; 1991.
23. Mazzocato P, Thor J, Bäckman U, Brommels M, Carlsson J, Jonsson F, et al. Complexity complicates lean: lessons from seven emergency services. J of Health Org and Mgt 2014 May 13;28(2):266-288. [doi: 10.1108/jhom-03-2013-0060]
24. Chambers DA, Glasgow RE, Stange KC. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. Implement Sci 2013 Oct 02;8(1):117 [FREE Full text] [doi: 10.1186/1748-5908-8-117] [Medline: 24088228]
25. Storkholm M, Mazzocato P, Savage C. Make it complicated: a qualitative study utilizing a complexity framework to explain improvement in health care. BMC Health Serv Res 2019 Nov 14;19(1):842 [FREE Full text] [doi: 10.1186/s12913-019-4705-x] [Medline: 31727069]
26. Stacey R. Strategic Management and Organisational Dynamics: The Challenge of Complexity to Ways of Thinking about Organisations (6th Edition). London, UK: Pearson; 2011.
27. Holland J. Hidden Order: How Adaptation Builds Complexity. Reading, MA: Addison-Wesley; 1995.
28. Waldrop M. Complexity: The Emerging Science at the Edge of Order and Chaos. New York: Simon & Schuster; 1992.
29. Greenhalgh T, Maylor H, Shaw S, Wherton J, Papoutsi C, Betton V, et al. The NASSS-CAT tools for understanding, guiding, monitoring, and researching technology implementation projects in health and social care: protocol for an evaluation

XSL•FO
RenderX

study in real-world settings. JMIR Res Protoc 2020 May 13;9(5):e16861 [FREE Full text] [doi: 10.2196/16861] [Medline: 32401224]
30. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
31. Grande SW, Longacre MR, Palmblad K, Montan MV, Berquist RP, Hager A, et al. Empowering young people living with juvenile idiopathic arthritis to better communicate with families and care teams: content analysis of semistructured interviews. JMIR Mhealth Uhealth 2019 Feb 22;7(2):e10401 [FREE Full text] [doi: 10.2196/10401] [Medline: 30794202]
32. Yin RK. Case Study Research: Design and Methods. Third Edition. Thousand Oaks, CA: SAGE Publications; 2003.
33. Sacchi C, Andersson K, Roczniewska M, Luckhaus JL, Malmqvist M, Rodmalm LP, et al. Mind the gap: analysis of two pilot projects of a home telehealth service for persons with complex conditions in a Swedish hospital. BMC Health Serv Res 2023 May 09;23(1):463 [FREE Full text] [doi: 10.1186/s12913-023-09409-4] [Medline: 37161458]
34. Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. Qual Saf Health Care 2003 Dec 01;12(6):458-464 [FREE Full text] [doi: 10.1136/qhc.12.6.458] [Medline: 14645763]
35. Montgomery D. Introduction to Statistical Process Control. Hoboken, NJ: John Wiley & Sons; 2005.
36. Thor J, Lundberg J, Ask J, Olsson J, Carli C, Härenstam KP, et al. Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care 2007 Oct 01;16(5):387-399 [FREE Full text] [doi: 10.1136/qshc.2006.022194] [Medline: 17913782]
37. A guide to creating and interpreting run and control charts: turning data into information for improvement. NHS Institute for Innovation and Improvement. 2009. URL: https://www.england.nhs.uk/improvement-hub/wp-content/uploads/sites/44/2017/11/A-guide-to-creating-and-interpreting-run-and-control-charts.pdf [accessed 2024-07-04]
38. Storkholm MH, Mazzocato P, Savage M, Savage C. Money's (not) on my mind: a qualitative study of how staff and managers understand health care's triple Aim. BMC Health Serv Res 2017 Jan 31;17(1):98 [FREE Full text] [doi: 10.1186/s12913-017-2052-3] [Medline: 28143539]
39. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
40. Oates G, Mims C, Geurs R, Bergquist R, Hager A, Guimbellot J, et al. Mobile health platform for self-management of pediatric cystic fibrosis: impact on patient-centered care outcomes. J Cyst Fibros 2023 Sep;22(5):823-829 [FREE Full text] [doi: 10.1016/j.jcf.2023.04.009] [Medline: 37085386]
41. Norton MI, Mochon D, Ariely D. The IKEA effect: when labor leads to love. J Consum Psychol 2011 Sep 09;22(3):453-460. [doi: 10.1016/j.jcps.2011.08.002]
42. Kegan R, Laskow LL, Miller M, Fleming A, Helsing D. An Everyone Culture: Becoming a Deliberately Developmental Organization. Boston, MA: Harvard Business Review Press; 2016.
43. Heifetz RA. Leadership Without Easy Answers. Cambridge, MA: The Belknap Press of Harvard University Press; 1994.
44. Snowden D, Boone M. A Leader’s Framework for Decision Making. Boston, MA: Harvard Business Review Press; 2007.
45. Guest G, Bunce A, Johnson L. How many interviews are enough?: an experiment with data saturation and variability. Field Methods 2016 Jul 21;18(1):59-82. [doi: 10.1177/1525822X05279903]
46. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov 27;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]

Abbreviations

CAT: complexity assessment tool
CF: cystic fibrosis
HCP: health care professional
LCL: lower control limit
mHealth: mobile health
NASSS: nonadoption, abandonment, spread, scale-up, and sustainability
RN: registered nurse
SPC: statistical process control
UCL: upper control limit

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 04.07.23; peer-reviewed by D Rose, G Akrong; comments to author 06.03.24; revised version received 08.04.24; accepted 20.06.24; published 31.07.24.

Please cite as:
Mazzocato P, Luckhaus JL, Malmqvist Castillo M, Burnett J, Hager A, Oates G, Wannheden C, Savage C
A Patient-Driven Mobile Health Innovation in Cystic Fibrosis Care: Comparative Cross-Case Study
J Med Internet Res 2024;26:e50527
URL: https://www.jmir.org/2024/1/e50527
doi:10.2196/50527
PMID:39083342

©Pamela Mazzocato, Jamie Linnea Luckhaus, Moa Malmqvist Castillo, Johan Burnett, Andreas Hager, Gabriela Oates, Carolina Wannheden, Carl Savage. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring How Patients Are Supported to Use Online Services in Primary Care in England Through "Digital Facilitation": Survey Study

Rachel Winder[1], BSc; John L Campbell[1], MD; Nurunnahar Akter[1,2], PhD; Abodunrin Q Aminu[1], PhD; Jeffrey Lambert[3], PhD; Emma Cockcroft[1], PhD; Chloe Thomas[1], MSc; Christopher E Clark[1], PhD; Carol Bryce[4], PhD; Jon Sussex[5], MSc; Helen Atherton[6], PhD; Christine Marriott[1]; Gary Abel[1], PhD

[1]University of Exeter Medical School, University of Exeter, Exeter, United Kingdom
[2]University of Liverpool, Department of Health Data Science, Liverpool, United Kingdom
[3]Department for Health, University of Bath, Bath, United Kingdom
[4]Unit of Academic Primary Care, University of Warwick, Warwick, United Kingdom
[5]RAND Europe, Cambridge, United Kingdom
[6]Primary Care Research Centre, University of Southampton, Southampton, United Kingdom

Corresponding Author:
John L Campbell, MD
University of Exeter Medical School
University of Exeter
Smeall Building
St Luke's Campus
Exeter, EX1 2LU
United Kingdom
Phone: 44 44 1392 722821
Email: john.campbell@exeter.ac.uk

Abstract

Background: Health service policy in many jurisdictions is driving greater investment into digital primary care services. While some patients and practices may benefit, there are concerns that not all are able or wish to access primary care services online. "Digital facilitation" is the "range of processes, procedures, and personnel seeking to support patients in their uptake and use of online services" and may address such concerns.

Objective: As part of a multimethod research program, we undertook surveys of practice staff and patients to gain insight into the support being offered by practices and explore patients' experiences of this support.

Methods: General practices from 4 regions of England were sent a questionnaire exploring the modes of digital facilitation offered, the personnel involved in its delivery, and views on the motivations and drivers for providing support. Moreover, 12,822 patients registered with 62 general practices (predominantly those providing practice survey responses) were sent a questionnaire exploring their experiences of any support offered by their practice to use online services.

Results: Almost one-third of practices (156/500, 31.2%) responded to the practice survey, with most reporting using passive modes of digital facilitation (eg, display, leaflets, and SMS text messages) and few using active modes (eg, offering tablets or computers or using practice champions). However, 90.9% (130/143) reported providing ad hoc support. Practices agreed that it was the responsibility of both the practice (105/144, 72.9%) and the wider National Health Service (118/143, 82.5%) to support patients in using online services and that providing such support benefited the practice (126/144, 87.5%) and their patients (132/144, 91.7%). Nearly a quarter of the patients (3051/12,822, 23.8%) responded to the patient survey, with few (522/3051, 17.11% or less) reporting awareness of any modes of digital facilitation apart from text messages and emails (1205/3051, 39.5%) and only 13.36% (392/2935) reporting receiving support to use online services. Adjusted logistic regression analyses showed that older patients had a lower likelihood of 4 outcomes: being aware of, or of using, digital facilitation efforts, or being told about or being helped to use online services (all $P<.05$), particularly with regard to being helped to use online services (adjusted odds ratio for patients aged 85 years versus those aged 55-64 years: 0.08, 95% CI 0.02-0.36). However, ethnic minority participants or those for whom their first language was not English had positive associations with these outcomes.

XSL•FO
RenderX

Conclusions: General practices recognize that patients would benefit from support to access online services. However, the support provided is often passive or ad hoc, and patients were seldom aware of digital facilitation efforts that their practice provided. There is potential to increase engagement with online primary care services by providing more support for all patients, particularly to provide targeted support for older patients.

(J Med Internet Res 2024;26:e56528) doi:10.2196/56528

KEYWORDS

primary care; online services; access to online health care services; general practice; survey; digital support; inequalities; remote consultation; health services research; digital technology

Introduction

Background

The UK National Health Service (NHS) is experiencing increasing demand for primary care provision as a result of demographic change, technological progress, and changing service configuration [1,2]. In a drive by NHS policy makers for greater investment in technology to help address the increase in demand, the use of online services in primary care in England is being promoted [3]. Primary care services that are provided online include ordering repeat prescriptions, booking appointments, and checking test results, as well as alternatives to face-to-face consultations between practitioners and patients such as video consults or e-consults [4,5]. There is an assumption among policy makers that online service delivery offers benefits for both patients (eg, wider choice of access, flexibility of use, and convenience) and primary care practices (eg, helping to organize work via triage to a specific person or service), thus suggesting the need for promotion or advocacy in respect of the use of such services [6]. The *NHS Long Term Plan*, published in 2019, aimed to ensure that every patient was offered access to digital primary care services and that all general practitioner (GP) practices had a website [7]. The COVID-19 pandemic created the need for urgent action to increase access to, and the use of, online services to reduce infection risk [8,9]. More recently, the 2023 model of general practice care aimed to enable practices to optimize patients' access to primary care services by all routes, including by online access [2].

Despite benefits for some patients [10,11], there are concerns that not all are able or wish to access their general practice (primary medical care) services online [12,13]. Gaining such access depends on having the technology, the knowhow, and a willingness to use such services [12,14,15]. Some patients may need support to set up their access and then to continue using an online service [16].

"Digital facilitation" refers to "the range of processes, procedures and personnel seeking to support patients in their uptake and use of online services" [16-18]. While there are programs aimed at widening digital participation and providing support for digitally excluded people [19,20], there is little formal evidence to suggest that patients who need support and help to access and use digital services in health care settings are receiving it; furthermore, where it is available, it is unclear what form such support might be taking.

Surveys of Practice Staff and Patients

The wider mixed methods study [18] aimed to describe the range of digital facilitation activities in general practice in England and included a scoping review [16]; practice and patient surveys; an analysis of national patient survey data; an ethnographic case study, including observations of general practice staff; patient and stakeholder interviews; and a synthesis of the evidence from the study. The surveys of practice staff and patients, as detailed in this paper, aimed to gain insight into what support (if any) to use online primary care services was being offered by practices and to explore patients' experiences of this support.

Methods

Sampling and Distribution

Our practice sampling frame was identified using publicly available information covering 8 NHS commissioning group areas (610 practices) in England. Between December 2020 and May 2021, we approached a random sample of 500 of these practices, initially by the practices' generic email address or via practice website forms, and included a link in the email to complete the survey online (specific to that practice), with up to 3 reminders to nonresponders at intervals of 2 to 3 weeks. Approximately 8 weeks after initial contact, paper questionnaires with reply envelopes were posted to nonresponding practices for the attention of the practice manager and up to 4 named GPs listed as providing services at the practice. Potential respondents could respond via a survey link to the survey or complete the paper copy and post it back in a reply-paid envelope. Only the first complete response per practice was used in the analysis. We anticipated responses from 300 (60%) of the 500 practices by using this approach [21].

All practices from which we received a response in the practice survey were invited to participate in the patient survey. We sought to invite 12,000 patients aged ≥16 years across 60 practices, aiming to secure 4200 (35%) responses [22]. Practices were given detailed instructions on how to select a random sample of up to 320 eligible patients per practice. Practices were requested to check the list and exclude patients known to be experiencing severe mental illness or recent bereavement, as well as those unable to provide informed consent. The remainder (up to 285 patients per practice) were mailed a paper questionnaire by their practice (along with an invitation letter, information sheet, and a reply-paid envelope). Subsequently, 2 reminders were sent from the practice to all invitees over the next month. The research team did not have access to

XSL•FO
RenderX

identifiable information on participating patients, and practices were not aware of details of who had responded. All materials were presented in English (although, for the patient survey, the invitation advised that another person could complete the survey on behalf of the respondent, and a telephone number was provided for help with completion). In line with recognized survey processes [23,24], practices located in areas with populations experiencing relatively high levels of deprivation (according to their Index of Multiple Deprivation score) were requested to invite larger numbers of patients to participate.

The invitation letter and information sheets for each survey explained the purpose of the study; the expected time to complete it (10 minutes for the practice survey and 10-15 minutes for the patient survey); which data would be stored how, where, and for how long; and the names of the investigators.

Any data received via responses on paper were entered by the study team using the same online survey platform. To ensure accurate data entry, double data entry was performed for 4 surveys per practice by a separate member of the study team and checked for consistency.

Survey Instruments

Overview

The questionnaires were each developed by subgroups of the research team (and later by the wider research team), with input and feedback from our patient and public involvement and engagement (PPIE) advisory group, in addition to convenience samples of patients, GPs, and volunteers who provided feedback on usability and functionality. The questions were developed using an iterative process and drew on the findings of the recent scoping review of digital facilitation [16] that formed part of the wider study [18].

Practice Survey

The practice survey (Multimedia Appendix 1) consisted of 11 main items presented across 4 sections (“Promotion, help and support for patients to use online services,” “Changes in access to offline services since national lockdown,” “Your views,” and “Your role”). No scales were developed. The items were formulated to address the range of online services offered to patients at the time of the survey and before the COVID-19 pandemic (ie, before 2020), the activities used to promote online services or to support patients in using them, the staff involved in supporting patients, whether specific groups were targeted with this support, respondents’ views on responsibility for support, the influence of various factors relating to online service provision, and the rationale behind providing online services. For the purposes of the questionnaire, promotion activities were defined as “activities to either inform patients about online services or encourage their use without necessarily providing any help or ‘support’ to assist patients in using them.” Response options were generally *tick box* in nature with a 5-point Likert scale for items in the “Your views” section. One question provided the opportunity for an open, free-text response on any issues that the respondent saw as relevant (which will be reported elsewhere). Practice staff could respond using the paper questionnaire they were sent (and return it in the reply-paid envelope) or complete online via an individual link.

Patient Survey

The content of the patient survey included items based on existing questionnaires on computer competence and confidence [25-28] and the national General Practice Patient Survey of NHS patients in England [29]. Patient survey questions (Multimedia Appendix 2) reflected the content of the practice survey, with wording tailored for patients. In addition to sections exploring patients’ digital confidence, the survey included sections on patients’ awareness and uptake of online services, as well as their experience of any support provided by the practice to use online services. Two further questions asked respondents about what help the practice could provide to access and use their practice’s online services, the results of which are reported elsewhere [30]. The questionnaire was piloted with a convenience sample of 6 volunteers (members of the public: n=3, 50%; members of the PPIE group: n=3, 50%).

PPIE Group Input

This project was conducted in collaboration with a study-specific PPIE group, including both patients and caregivers. Eight PPIE group members participated in the development, methods, and interpretation of findings of the practice and patient surveys. The group attended an initial brainstorming session for the patient survey. At further meetings, the PPIE contributors generated discussion on whether or how to include points in the surveys. As a result of the input of the PPIE contributors, the research team added and amended some response options (eg, use of emails to and from practices as a further response choice); reworded some digital facilitation options (eg, included television displays in GP surgeries); incorporated further suggested inclusions (eg those with carer responsibilities as a potential group that practices might target with digital facilitation); explored further which patient groups may have needed targeted support when using online services; and identified words that might cause a barrier to respondents understanding the questionnaire (eg, “activities,” “facilitation,” and “engage with”), which were removed. The contributors also provided feedback on aspects of completing the survey and its length and requested further explanation of acronyms and abbreviations used in the patient invitation letter. The contributors attended meetings where the survey results were presented and discussed and contributed to the interpretation of initial findings from a patient perspective.

Ethical Considerations

Ethics approval was granted for the patient survey by the North East Newcastle and North Tyneside 2 Research Ethics Committee on April 27, 2021, and by the Health Research Authority on July 1, 2021 (Integrated Research Authority System 289425). Ethics approval was not required for the practice survey element (as advised by the Health Research Authority) because the survey did not intend to change practice or patient care. Patients were deemed to have consented to participate in the patient survey if they returned a questionnaire either by post or online (implied consent). The research team did not ask for any personal data from survey participants,

XSL•FO
RenderX

although participants could provide their contact details (which were kept separate from other survey data) if they wished to take part in the prize draw. Information on processing of personal data on the participant information sheet provided an explanation of our approach to handling personal data. Analysis of General Practice Patient Survey data was deemed service evaluation not requiring ethics approval.

Practices responding to the practice survey were entered into a prize draw for 1 of 10 £250 (US $316) vouchers. A voluntary prize draw for 1 of 10 £25 (US $32) vouchers was offered as an incentive for patients participating in the patients survey. Potential patient survey respondents were informed that consent would be assumed upon return of a questionnaire either by post or online.

Statistical Analysis

Examining Variability in Digital Facilitation Outcomes

Simple descriptives were used for the practice and patient surveys (after excluding missing responses to individual questions). The patient survey data were further analyzed using mixed effects logistic regression models, with a random effect for practice, to examine variability in 4 outcomes in relation to patient characteristics on a complete case basis. The four binary outcomes were developed from patient survey responses, considering (1) awareness of, and (2) use of, practice efforts aimed at providing digital facilitation (ie, by endorsing 1 of the following options in the relevant questions [Q10 and Q11]: displays in the practice, leaflets, email or text messages, practice website, social media, workshops or events, and making IT equipment available to access online services); (3) being told by someone from the practice about online services; and (4) being helped by someone from the practice to use online services. For each of the binary outcomes, univariable associations were calculated along with a multivariable model adjusting for respondent self-reported age, gender, deafness or hearing impairment, parental status, ethnicity, physical or mental long-term health condition, working status, first language, and whether they had repeat prescriptions (adjusted model 1). A second model (adjusted model 2) augmented adjusted model 1 with a composite measure of digital confidence constructed from patient responses to Q2 to Q6 (Multimedia Appendix 3). A further regression was run in this model using a binary version of the digital confidence scale (confident vs quite confident or not confident) as the outcome and using the same covariates as in adjusted model 1. Sensitivity analyses were performed, excluding respondents who specified having received help to complete the survey.

Comparison of Practice and Patient Survey Response Data: Combined Analyses

Finally, we combined data from the practice and patient surveys (restricted to practices that participated in both parts of the study) to examine whether practice responses were associated with patient awareness and use of digital facilitation. Three sets of comparisons were made using chi-square tests:

- The percentage of respondents reporting being aware of, and the percentage of respondents reporting using, particular modes of facilitation were compared between those registered at practices that reported offering that mode of facilitation and those registered at practices that did not make such a provision (eg, we compared the percentage of respondents who were aware of their practice providing leaflets about online services between practices that said they used leaflets to promote and support the use of online services and those that did not).
- The percentage of respondents reporting being aware of, and the percentage of respondents reporting using, any mode of facilitation were compared: (1) between respondents registered at practices that reported undertaking ad hoc promotion or support of online services and those that did not and (2) between respondents registered at practices that reported using a practice champion and those that did not (eg, we compared the percentage of respondents who were aware of any mode of facilitation between practices that said they used a practice champion and those that did not).
- We investigated respondent groups who described themselves as either an older adult (aged ≥65 years), having a physical health condition, having a mental health condition, having limited or no internet access, being a non-English speaker or speaking English as a second language, being a member of an ethnic minority, or as being a carer. We compared the percentage, separately, in each of these groups, who reported awareness or use of any modes of facilitation between those registered at practices that reported offering digital facilitation support targeted at that group and those that did not (eg, we compared the percentage of respondents aged ≥65 years who were aware of, or used, any mode of facilitation between practices that said they targeted older patients and those that did not).

Study Registration

The study was registered with the Research Registry (researchregistry6523).

Results

Practice Survey

Of the 499 practices invited (of the intended 500 practices, n=2, 0.4% were found to have merged), 156 (31.3%) sent back at least 1 questionnaire. Participating practices were more likely than other practices in England to serve less deprived populations (21/155, 13.6% in the least deprived quintile vs 1359/6745, 20.15%), to be in rural areas (46/156, 29.5% vs 1004/6842, 14.67%), and to have >12,000 registered patients (54/155, 34.8% vs 1477/6461, 22.86% [31-33]; Multimedia Appendix 4). The age profile and ethnicity of the registered populations showed smaller differences compared to other practices. Full summaries of survey responses are presented in Multimedia Appendices 4-12, but here we focus on those questions pertaining to digital facilitation.

There was a clear division in the practice survey between the endorsement of modes of digital facilitation that could be described as passive, where information is provided with no scope for 2-way interaction (displays, leaflets, text messages, emails, social media, and material on practice websites), and those that could be described as active and would involve 2-way

XSL•FO
RenderX

interaction between patients and staff members (ad hoc support, using a practice champion, holding workshops or events, and offering tablets or computers; Multimedia Appendix 6). A majority of practices reported using passive modes of facilitation for either promotion of online services or supporting patients to use them; for example, leaflets were used in 70.9% (100/141) of the practices (Figure 1). By contrast, with the exception of ad hoc support (which was reported in 130/143, 90.9% of the practices), active modes of facilitation were reported by only a minority of practices (eg, workshops or events were used by 18/145, 12.4% of the practices). Most of the practices reported using digital facilitation across a wide range of online services, with 96.5% (139/144) reporting using it to promote or support online repeat prescription ordering (Multimedia Appendix 7).

Figure 1. Digital facilitation activities used to promote online services and support patients to use them by percentage of practices responding to the practice survey.

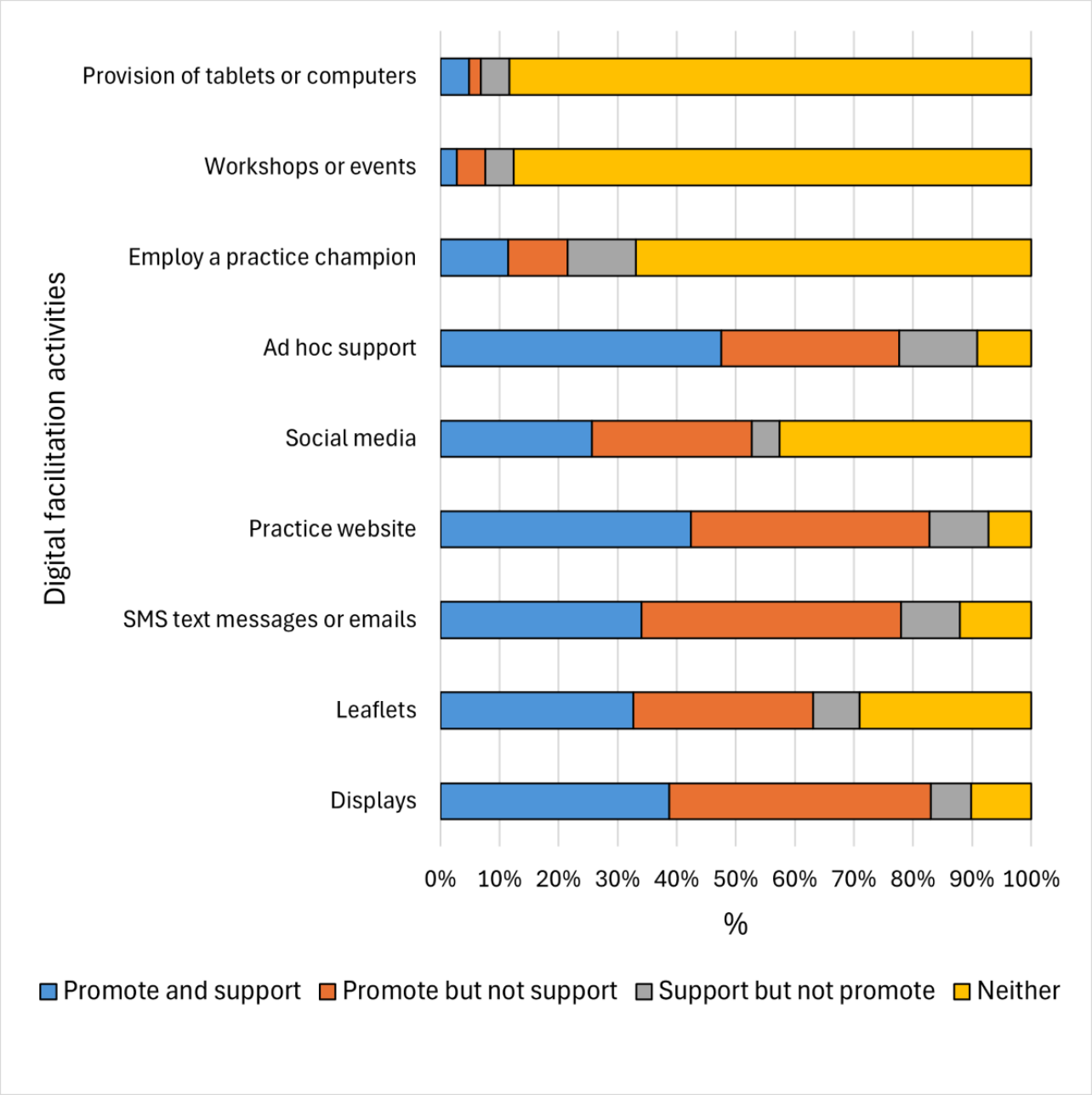

Of the 156 participating practices, the vast majority reported that promotion and supporting activities involved administrative (n=134, 85.9%) and reception (n=134, 85.9%) staff, with doctors (n=96, 61.5%), nurses (n=83, 53.2%), and other health care professionals (n=65, 41.7%) also frequently reported as being involved. The involvement of IT staff (44/156, 28.2%), volunteers (31/156, 19.9%), external contractors (6/156, 3.9%), and practice staff with specific responsibilities around digital facilitation (16/156, 10.3%) was reported less often (Multimedia Appendix 8). Of the 156 practices, 18 (11.5%) reported that they targeted their facilitation activities at every patient group listed in the questionnaire (Multimedia Appendix 9), with most of the practices (85/156, 54.5%) targeting >1 group. The only specific group reported to be targeted by most of the practices (87/156, 55.8%) were older adults aged ≥65 years (Multimedia Appendix 9).

XSL•FO
RenderX

Most of the practices (129/144, 89.6%) responding to a list of statements (Multimedia Appendix 11) agreed that the COVID-19 pandemic had been a key driver in the uptake of online services by patients, and many (109/144, 75.7%) reported that the pandemic had led to an increase in the support they provided to patients to use online services. Responsibility to alert patients about online services was seen to lie with the practice (123/144, 85.4%) or the wider NHS (122/144, 84.7%), with many also seeing responsibility of providing support to use online services lying with the practice (105/144, 72.9%) and the NHS (118/143, 82.5%). Current support provided to patients by practices was agreed to be in response to patient demand (119/142, 83.8%), rather than the Clinical Commissioning Group (47/142, 33.1%), practice staff (61/142, 43%), or monetary incentives (20/135, 14.8%).

Nevertheless, most of the respondents agreed that supporting patients to use online services benefited the practice (126/144, 87.5%) and patients (132/144, 91.7%); however, 60.3% (85/141) of the practices agreed that the practice lacked the capacity to provide adequate support for patients. Indeed, 78.4% (109/139) agreed that their practice had increased the uptake of online services by supporting patients to use them, and 72.7% (104/143) hoped to further increase uptake through support. It was acknowledged by nearly all respondents (135/143, 94.4%) that some patients were unlikely to use online services regardless of the support provided and that some patient groups required more support than others (131/143, 91.6%). Two-thirds of practice respondents (95/142, 66.9%) agreed that increasing patient uptake of online primary care services led to operational efficiencies for the practice. There was strong agreement that online access to primary care services was complementary to traditional forms of access (121/143, 84.6%), whereas 36.6% (52/142) agreed that online primary care services would ultimately replace traditional forms of access.

Patient Survey

In all, 62 practices sent invitations to 12,822 patients between August 2021 and May 2022, of whom 3051 (23.8%) responded (Multimedia Appendix 13). Analysis of double data entry of paper responses identified a difference in 0.4% of data points entered, which was considered acceptable. Despite lower patient response rates from practices serving populations considered deprived, due to our adopted approach to sampling, responses were largely representative in terms of deprivation with, for example, 21.4% (653/3051) of the responders registered at practices falling in the lowest quintile of deprivation nationally and a broadly even spread of responders (ranging from 438/3051, 14.36% to 698/3051, 22.88%) across all quintiles (Multimedia Appendix 14).

Of those responding to the patient survey, 56.59% (1710/3022) were female, 45.39% (1373/3025) were aged ≥65 years, most (2741/2958, 92.66%) were White, 8.8% (261/2966) reported that English was not their first language, 43.7% (1269/2904) were working either full time or part time, and 42.84% (1244/2904) were retired (Table 1). Of the respondents, 9% (280/3051) said they did not have access to the internet at home, with 98.6% (276/280) of this group completing the paper version of the survey. In terms of overall digital confidence (Multimedia Appendix 3) with respect to a range of digital tasks (eg, using search engines, completing online forms, sending personal messages, and installing apps), a little more than half (1589/2961, 53.66%) were categorized as being very confident, whereas 22.56% (668/2961) were categorized as not confident (Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Self-reported demographics of responders to the main patient survey (n=3051).

Demographic variables	Respondents, n (%)
Gender (n=3022)	
Male	1312 (43.41)
Female	1710 (56.59)
Age group (y; n=3025)	
16-24	140 (4.63)
25-34	195 (6.45)
35-44	262 (8.66)
45-54	436 (14.41)
55-64	619 (20.46)
65-74	727 (24.03)
75-84	515 (17.02)
≥85	131 (4.33)
Deaf or hearing impediment (n=2989)	
Yes	358 (11.98)
No	2631 (88.02)
Blind or partially sighted (n=2963)	
Yes	54 (1.82)
No	2909 (98.18)
Ethnicity (n=2958)	
Asian or Asian British	107 (3.62)
Black or African or Caribbean or Black British	57 (1.93)
White	2741 (92.66)
Mixed	31 (1.05)
Other	22 (0.74)
Mental health condition or disability (n=2876)	
Yes	472 (16.41)
No	2281 (79.31)
Don’t know	72 (2.50)
Prefer not to say	51 (1.77)
Working status (n=2904)	
Full-time paid work	906 (31.2)
Part-time paid work	363 (12.5)
Full-time student	75 (2.58)
Unemployed	59 (2.03)
Permanently sick	72 (2.48)
Fully retired	1244 (42.84)
Looking after family	104 (3.58)
Doing something else	75 (2.58)
Furloughed[a]	6 (0.21)
English as their first language (n=2966)	
Yes	2705 (91.2)
No	261 (8.80)

XSL•FO
RenderX

Demographic variables	Respondents, n (%)
Carer (n=2937)	
Yes	700 (23.83)
No	2237 (76.17)
Parent (n=2956)	
Yes	420 (14.21)
No	2536 (85.79)
Repeat prescriptions (n=2978)	
Yes	1957 (65.72)
No	1021 (34.28)
Help to complete the survey (n=2963)	
Yes	183 (6.18)
No	2780 (93.82)
Physical health condition or disabilities (n=2900)	
Yes	1046 (36.07)
No	1737 (59.9)
Don't know	74 (2.55)
Prefer not to say	43 (1.48)

[a]UK government scheme offering payment to employers to retain and pay staff while businesses were closed due to the COVID-19 pandemic.

Table 2 presents a summary of responses to questions concerning awareness, use, and provision of support for the use of general practices' online services. More than a third of those responding (1106/2998, 36.89%) had not attempted to use their practice's website, although of those who had tried, 77.27% (1462/1892) reported finding it *very easy* or *fairly easy* to use. For all services considered, ≤37% (≤1126/3051) of the respondents were aware of the various online services provided by their practice, with the exception of appointment booking and ordering repeat prescriptions online (1675/3051, 54.9% and 1944/3051, 63.72%, respectively). In terms of the use of online services, ≤16% (≤501/3051) of the respondents had used a range of online services, with the exception of ordering repeat prescriptions online, where a third (1003/3051, 32.87%) had used the service. Apart from the use of SMS text messages or emails (where 1205/3051, 39.5% respondents reported awareness of facilitation efforts), ≤17% (≤522/3051) of the respondents were alert to any forms of digital facilitation opportunities at their practice. The reported use of these facilitation efforts was even lower than patients' awareness, with <10% (≤291/3051) of the patients making use of any mode, apart from email or text messages (860/3051, 28.19%).

XSL•FO
RenderX

Table 2. Summary of responses to the core items of the patient survey (n=3051).

Questions	Respondents, n (%)
Ease of practice website use (n=2998)	
Very easy	541 (18.05)
Fairly easy	920 (30.69)
Not very easy	286 (9.54)
Not at all easy	145 (4.84)
Haven't tried	1106 (36.89)
Awareness of online services	
Appointment booking online	1675 (54.9)
Ordering repeat prescriptions	1944 (63.72)
Accessing medical records online	945 (30.97)
Accessing test results online	663 (21.73)
Email enquiries to the practice	1126 (36.91)
Help or advice using an online form (online consultations or e-consults)	945 (30.97)
Video consultations	458 (15.01)
Use of online services	
Appointment booking online	469 (15.37)
Ordering repeat prescriptions	1003 (32.87)
Accessing medical records online	393 (12.88)
Accessing test results online	260 (8.52)
Email enquiries to the practice	457 (14.98)
Help or advice using an online form (online consultations or e-consults)	501 (16.42)
Video consultations	118 (3.87)
Awareness of facilitation	
Displays in the practice (eg, posters or television displays)	522 (17.11)
Leaflets about online services	178 (5.83)
SMS text messages or emails	1205 (39.5)
Practice website content (eg, *how to* guide or video or prominent pop up)	229 (7.51)
Use of social media	118 (3.87)
Scheduled workshop or events (in person or online)	17 (0.56)
Making tablets or computers available to access online services	16 (0.52)
Use of facilitation	
Displays in the practice (eg, posters or television displays)	291 (9.54)
Leaflets about online services	93 (3.05)
SMS text messages or emails	860 (28.19)
Practice website content (eg, *how to* guide or video or prominent pop-up)	152 (4.98)
Use of social media	79 (2.59)
Scheduled workshop or events (in person or online)	15 (0.49)
Making tablets or computers available to access online services	15 (0.49)
Usefulness of help	
Booking appointments online (n=196)	
Not helpful	40 (20.41)
Quite helpful	77 (39.29)

XSL•FO
RenderX

Questions	Respondents, n (%)
Very helpful	79 (40.31)
Ordering repeat prescriptions online (n=211)	
Not helpful	22 (10.43)
Quite helpful	66 (31.28)
Very helpful	123 (58.29)
Accessing medical records online (n=121)	
Not helpful	24 (19.83)
Quite helpful	42 (34.71)
Very helpful	55 (45.45)
Accessing test results online (n=111)	
Not helpful	23 (20.72)
Quite helpful	34 (30.63)
Very helpful	54 (48.65)
Email enquiries to the practice (n=154)	
Not helpful	27 (17.53)
Quite helpful	53 (34.42)
Very helpful	74 (48.05)
Online video consultations with GP [a] or other health care professional (n=81)	
Not helpful	28 (34.57)
Quite helpful	19 (23.46)
Very helpful	34 (41.98)
Other (n=33)	
Not helpful	10 (30.3)
Quite helpful	8 (24.24)
Very helpful	15 (45.45)
Help from GP using online form (n=171)	
Not helpful	28 (16.37)
Quite helpful	59 (34.5)
Very helpful	84 (49.12)
Told about online services (n=2964)	
Yes	883 (29.79)
No	2081 (70.21)
Helped to use online services (n=2935)	
Yes	392 (13.36)
No	2543 (86.64)
Reasons why not using online services	
No internet access	247 (8.1)
Security concern	228 (7.47)
Confidentiality concerns	145 (4.75)
Not knowing how to register	440 (14.42)
Registration too difficult	175 (5.74)
Not knowing how to get support	259 (8.49)

XSL•FO
RenderX

Questions	Respondents, n (%)
Practice too busy to help	262 (8.59)
Prefer to speak in person	1432 (46.94)

[a]GP: general practitioner.

Only 883 (29.79%) of 2964 patients agreed that they had been told about online services by someone at their practice, and only 392 (13.36%) of 2935 patients reported that they had been supported to use such services. Those who did receive help largely rated this support as beneficial: 89.6% (189/211) selected *quite helpful* or *very helpful* for online ordering of repeat prescriptions, and 65% (53/81) reported the help to be *quite helpful* or *very helpful* for video consultations. A variety of reasons were reported for not using online services, but, most commonly, respondents preferred to speak in person (1432/3051, 46.94%), with 14.42% (440/3051) reporting that they did not know how to register and 5.74% (175/3051) reporting that they found registration too difficult.

Table 3 shows the results of the adjusted model 1 logistic regression analyses, with the results of the unadjusted model analyses shown in Multimedia Appendix 16, for 4 facilitation outcomes, that is, awareness (Q8) and use (Q9) of facilitation efforts, being told about online services (Q12), and being helped to use online services (Q13; Multimedia Appendix 2). Older patients were less likely than younger patients to report awareness and use of facilitation efforts, as well as to report having been told about or helped in the use of such services (all *P* values <.05). This was particularly noticeable in respect of being helped to use online services (adjusted odds ratio [OR] for patients aged ≥85 years vs those aged 55-64 years: 0.08, 95% CI 0.02-0.36; *P*=.006). Respondents in receipt of repeat prescriptions were more likely than those not receiving this provision to have experienced all 4 facilitation outcomes. Patients describing themselves as being of minority ethnicity were more likely than those describing themselves as being of White ethnicity to be aware of digital facilitation (adjusted OR 1.48, 95% CI 1.00-2.20; *P*=.05), to have made use of digital facilitation (adjusted OR 1.48, 95% CI 1.03-2.15; *P*=.04), and to have been helped to use online services (adjusted OR 1.80, 95% CI 1.14-2.86; *P*=.01). However, no difference was seen between respondents of White and minority ethnicity for being told about online services (adjusted OR 0.91, 95% CI 0.61-1.36; *P*=.65). There was weak evidence that patients for whom English was not their first language were more likely to be told about online services (adjusted OR 1.61, 95% CI 1.12-2.32; *P*=.01) or to use digital facilitation (adjusted OR 1.79, 95% CI 1.28-2.52; *P*=.001). The respondent's gender, being deaf or having a hearing impairment, and parental and employment status were not associated with any of the 4 outcomes (*P*>.10 for all).

XSL•FO
RenderX

Table 3. Adjusted logistic regression (model 1) considering participants' awareness of digital facilitation, use of digital facilitation, being told about online services, and being helped to use online services (all models; n=2587).

Characteristics	Awareness of any digital facilitation efforts[a]		Use of any digital facilitation efforts[b]		Being told about online services		Being helped to use online services	
	Adjusted OR[c] (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
Gender								
Male	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Female	0.92 (0.78-0.90)	.34	0.90 (0.76-1.07)	.23	1.14 (0.95-1.36)	.16	0.86 (0.68-1.10)	.24
Age group (y)								
16-24	0.84 (0.50-1.42)	<.001	0.64 (0.37-1.09)	<.001	1.58 (0.93-2.70)	.04	1.00 (0.47-2.17)	.006
25-34	1.08 (0.73-1.58)	<.001	1.07 (0.74-1.56)	<.001	1.14 (0.77-1.69)	.04	1.11 (0.65-1.91)	.006
35-44	0.98 (0.67-1.43)	<.001	1.01 (0.70-1.46)	<.001	1.08 (0.73-1.60)	.04	1.42 (0.85-2.37)	.006
45-54	0.94 (0.70-1.27)	<.001	0.94 (0.70-1.26)	<.001	1.01 (0.74-1.38)	.04	1.31 (0.87-1.98)	.006
55-64	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
65-74	0.71 (0.53-0.95)	<.001	0.88 (0.66-1.17)	<.001	0.77 (0.56-1.05)	.04	0.81 (0.52-1.24)	.006
75-84	0.43 (0.30-0.60)	<.001	0.62 (0.44-0.87)	<.001	0.66 (0.45-0.95)	.04	0.50 (0.30-0.84)	.006
≥85	0.32 (0.19-0.54)	<.001	0.27 (0.15-0.49)	<.001	0.38 (0.20-0.71)	.04	0.08 (0.02-0.36)	.006
Deafness or hearing impairment								
No	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Yes	0.93 (0.70-1.22)	.60	1.16 (0.87-1.53)	.31	0.82 (0.60-1.13)	.23	1.32 (0.88-2.00)	.18
Parent								
No	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Yes	1.28 (0.96-1.72)	.01	1.03 (0.78-1.37)	.83	1.28 (0.95-1.73)	.11	1.03 (0.70-1.52)	.88
Ethnicity								
White	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Other	1.48 (1.00-2.20)	.05	1.48 (1.03-2.15)	.04	0.91 (0.61-1.36)	.65	1.80 (1.14-2.86)	.01
Long-term physical or mental health condition								
No	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Yes	1.25 (1.03-1.50)	.02	1.17 (0.97-1.41)	.09	1.07 (0.88-1.31)	.48	1.29 (0.98-1.70)	.07
Work status								
Working	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Studying	0.73 (0.38-1.41)	.12	1.05 (0.54-2.04)	.19	1.09 (0.56-2.12)	.98	1.23 (0.49-3.10)	.62
Other	0.72 (0.54-0.95)	.12	0.78 (0.59-1.04)	.19	0.95 (0.70-1.28)	.98	1.23 (0.83-1.82)	.62
Retired	0.96 (0.72-1.26)	.12	0.77 (0.59-1.02)	.19	1.00 (0.74-1.35)	.98	1.24 (0.82-1.89)	.62

XSL•FO
RenderX

Characteristics	Awareness of any digital facilitation efforts[a]		Use of any digital facilitation efforts[b]		Being told about online services		Being helped to use online services	
	Adjusted OR[c] (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
First language								
English	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Other	1.38 (0.96-1.96)	.08	1.79 (1.28-2.52)	.001	1.61 (1.12-2.32)	.01	1.52 (0.97-2.37)	.07
Repeat prescription								
No	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Yes	1.28 (1.04-1.56)	.02	1.46 (1.19-1.79)	<.001	1.70 (1.36-2.13)	<.001	1.45 (1.06-1.97)	.02

[a]Awareness of any facilitation efforts includes respondents who ticked any of the first 7 options of Q10.

[b]Use of any facilitation efforts includes respondents who ticked any of the first 7 options of Q11.

[c]OR: odds ratio.

In adjusted model 2 (Multimedia Appendix 16), lower digital confidence was associated with both lower awareness and lower use of facilitation, as well as with less likelihood of being told about, or being helped to use, online services. Furthermore, the association between age and each of the outcomes was weaker in this model than in adjusted model 1, indicating that some of the differences ascribed to older age could be explained by older patients having, on average, lower digital confidence than younger patients. However, the changes in ORs were small (eg, OR for the awareness of facilitation for patients aged ≥85 years changed from 0.35 to 0.59), with a substantial age effect remaining after adjustment for digital confidence. By contrast, the association between whether a respondent speaks English as a first language and 3 of the 4 outcomes was stronger when adjusting for digital confidence (eg, OR for being helped to use online services increased from 1.52 to 1.71).

Sensitivity analyses excluding those who reported having had help to complete the questionnaire did not lead to materially different findings (Multimedia Appendix 17).

Comparison of Practice and Patient Survey Response Data: Combined Analyses

Patients often reported awareness and even use of various modes of facilitation in practices that did not report using that mode. This finding implies that either practice or patient responses were in error or that the practice's use of particular modes had changed over time. Moreover, when considering specific modes of facilitation, there was generally no evidence that patients registered at practices using that mode were any more likely to report being aware of it or using it; for example, 5.39% (92/1707) of the patients were aware of leaflets being used for digital facilitation in practices that reported using leaflets compared to 6.27% (67/1069) of the patients in practices that did not report using leaflets (Multimedia Appendix 18). However, exceptions were evident for in-practice displays, where 17.55% (466/2655) of the respondents reported awareness of such displays in practices that reported using them compared to 10.7% (30/280) in those that did not (*P*=.004), social media (105/1733, 6.06% vs 12/1242, 0.97% for awareness and 63/1733, 3.64% vs 15/1242, 1.21% for use; *P*<.001 for both), and workshops or events (4/267, 1.5% vs 12/2607, 0.46% for awareness; *P*=.03; and 4/267, 1.5% vs 10/2607, 0.38% for use; *P*=.01).

No difference was evident in respect of patient awareness or use of any form of digital facilitation when comparing those practices that reported using ad hoc support for digital facilitation to those that did not do so (Multimedia Appendix 19). However, while using a practice champion was associated with higher patient awareness of facilitation efforts (149/410, 36.3% vs 669/2304, 29.04%; *P*=.003), no greater use of facilitation efforts was seen (Multimedia Appendix 19). When considering practices that reported targeting specific patient groups with digital facilitation efforts, we found no evidence of differences in the patient awareness or use of any mode of digital facilitation within such groups (Multimedia Appendix 20).

Discussion

Principal Findings

We explored whether and how practices supported patients to use online primary care health services and patients' experiences of this support.

Against a backdrop of practices offering an increased number and variety of online services, in part as a response to the COVID-19 pandemic, our findings showed that practices commonly reported using passive modes (eg, displays, leaflets, and text messages) of digital facilitation to promote or support patients' use of online services, and reported that practice staff provided ad hoc support for patients. Active modes of digital facilitation (eg, using a practice champion) were rarely in place. We found that administrative (134/156, 85.9%) and reception (134/156, 85.9%) staff provided digital facilitation efforts in the vast majority of practices, with clinical staff (doctors: 96/156, 61.5% and nurses: 83/156, 53.2%) contributing in more than half of the practices. Although most practices viewed the potential for digital facilitation positively, benefitting both patients (132/144, 91.7%) and practices (126/144, 87.5%), many (85/141, 60.3%) felt that they lacked the resources to deliver it.

XSL•FO
RenderX

More than three-quarters of those patients (1461/1892, 77.22%) responding that they had accessed the practice website reported finding it *very easy* or *fairly easy* to use, which is surprising considering that many websites have historically pitched their usability levels too high for many potential users [34]. Nevertheless, despite the relatively positive reports from practices that they have provided some forms of digital facilitation (albeit mainly ad hoc), patients' reported awareness and use of online services was still generally low. Awareness and provision of any support for the use of these services was lower still, with the exception of text messages or emails, for which 39.5% (1205/3051) of the patients reported awareness of their practice providing support through this mode. Despite older patients tending to have higher health care use and practices reporting that they targeted the group of older adults aged >65 years more often than other patient groups, older patients were less likely to be aware of digital facilitation efforts reported to be provided by their practice or to use them or even to report having been informed about online services or being supported in their potential use of such services.

Comparison to Prior Work

Recent reports have focused on inequalities in accessing health services within the NHS from ethnic minority and other groups [35,36]. Perhaps surprisingly, in our study, ethnic minority individuals and those for whom English was not their first language were more likely to report engagement with, or being engaged in, digital facilitation. Being in receipt of repeat prescriptions was also associated with such engagement. Patients gave a variety of reasons for not engaging with online services, often expressing their preference to speak to someone and their difficulties in registering to use online services. Such difficulties suggest unmet need for effective digital facilitation to increase the use of online services.

It is not surprising that when people are not aware of, and do not use, online services, their awareness of any support to use them would be lower still. However, it is important to consider whether a lack of support leads to a reduced use of online services. Low patient awareness and use of online services has been shown in other surveys of patients. In a survey of general practice patients in the West Midlands of England in 2019, half (1362/2715, 50.17%) were aware they could order repeat prescriptions online, and 21.84% (584/2674) had used the service; for accessing records, 23.17% (629/2715) were aware, and 7.79% (172/2207) had used them [37]. Our study showed some increase on these figures but with levels remaining low overall. The same survey found that patients who used the internet daily were more likely to be aware of, and use, online services, suggesting that familiarity with the internet leads to the use of online services, and this is aligned with our findings that the level of digital confidence is associated with awareness and use of online services [37]. Ordering repeat prescriptions would, for many, mean frequent use of the online repeat prescription service. This higher engagement, evident in patients in our study, may have helped them embed the online process, or necessitated the need for them to seek out support.

Research applying secondary analysis of the English General Practice Patient Survey to examine awareness and use of online appointment booking reported that 45% of patients were aware of it, and 16% had used it; these values are comparable with our findings [38]. Qualitative research in the same study identified that older patients preferred other ways of booking appointments and that difficulties registering for online appointment booking impacted their decision on whether to use it [38]. In our study, 5.74% (175/3051) of the respondents reported registration being too difficult, and 14.42% (440/3051) reported not knowing how to register, as the reasons why they did not use online services.

A 2016 survey reporting on how UK general practices were providing (or planning) digital consultations (email and video) [21] found that GPs were concerned about the impact that digital consultation could have on older people. Our survey of practices found that digital facilitation efforts were mainly targeted at older adults, which fits with the perception that it is older adults who struggle most with digital health services. For the older population, Leach et al [16] found that despite the concern that older people need additional support, there were some studies that showed that older people were more likely to use technology to access online services once provided with digital facilitation efforts [39-41]. In considering ethnicity and language, despite reports on the implications of language barriers and their effect on health [42,43], there is a need to explore further whether online services are offering improved access via online over face-to-face services. It is possible that those of minority ethnicity and those whose first language was not English were more used to finding technological workarounds or have family and friends who can support them to use the services.

A Norwegian study examining older patients' use of e-consults found that they had neither been encouraged to use the service nor helped to use it, and patients highlighted the importance of having digital literacy to use the service [44]. Previous research examining the unintended consequences of introducing digital services in general practice found that patients expressed uncertainty about how to use online consultation systems and their online patient records, and that this could be countered by providing patients with clear information and instructions on how to use them [14,45]. In 1 study, there was some evidence that the more proactive, *hands-on* facilitation approaches were likely to be of more use and support for patients from older age groups [16].

Strengths and Limitations

One of the strengths of the study was that by oversampling patients in deprived areas, we gained a broadly similar number of responses from patients across all deprivation quintiles. The survey procedures were refined by conducting a pilot survey and was supported by extensive patient and public input in the design and implementation of the research. This work was undertaken as part of a large-scale research project that got underway as the COVID-19 pandemic broke out, which presented many challenges. However, we were able to deliver the surveys and provided added insights into a system of access to practices that was changing rapidly.

Second, our patient survey response rate (3051/12,822, 23.8%) is comparable to that of other primary care–based surveys [37,46,47] and should be considered credible when bearing in

XSL•FO
RenderX

mind the difficulties also presented to patients by the COVID-19 pandemic. We also managed to capture many responses from those respondents reporting no access to the internet by using paper invites as well as offering an online mode. Of the 280 respondents reporting no internet access, 276 (98.6%) took advantage of the paper version of the survey. At the start of the pandemic, practices were grappling with new and rapidly changing procedures, increasing staff absenteeism, and increased service use [48,49], and thus participation in research for some practices could have been too challenging. The moderate response rate for the patient survey could also have been a product of our wish to target patients in more deprived areas.

Third, by focusing the patient survey on practices that had responded to the practice survey, we were able to link the responses from the 2 surveys. However, often, the types of facilitation reported by practices had no statistical association with patient responses. This may be a limitation of using self-report, with responses being unreliable, or an indication that the intensity of digital facilitation efforts was too low to impact many patients. This uncertainty could not be investigated in detail because we were not able to examine the intensity of facilitation efforts.

Considering limitations, first, all materials for the patient survey were presented in English. Nevertheless, 8.8% (261/2966) of the respondents reported that their first language was not English, which is similar to the percentage reported in the UK Census 2021 for England and Wales (8.9%) [50]. Our advice that another person could complete the survey from the respondent's viewpoint on their behalf may have encouraged some who felt that they could not complete it without help.

A further limitation to the study was that we did not have information on the socioeconomic status (eg, deprivation) of patients or the rurality classification of their home location. We were not privy to patient addresses, which would allow linkage to data sets containing this information.

Future Directions

Despite the high levels of ad hoc support reported to be provided by practice staff in our survey, only 13.36% (392/2935) of the respondents to the patient survey reported having been supported to use an online service. It is possible that the lack of any formality in the process of receiving support was not recognized as bona fide assistance in their online endeavors, or it may be that most of the ad hoc support is provided to a small proportion of a practice's patients. Establishing whether this is truly the case should be a focus for future research.

Our findings raise questions regarding the success of NHS policy to increase patients' reliance on online services [51]. They draw further attention to the need for policy makers to provide resources in the form of finances for staff training, time, and infrastructure for practices to ensure that all patients have equitable access to online primary care services, as well as to allow for the ongoing increasing use of online services by patients [18,40,52-54].

Conclusions

There is a range of digital facilitation activities offered at general practices. This is limited to passive or ad hoc forms in most general practices, with few providing more proactive models of support. There is a recognition by general practices that patients would benefit from support in accessing and using online services. Despite most practices reporting provision of digital facilitation efforts, most patients are not aware of them. However, where patients do experience these facilitation efforts, they often find them helpful. Some patient groups are at risk of digital exclusion [15,20], and it is potentially concerning that older patients are less likely to be aware of, or make use of, digital facilitation efforts. More reassuringly, ethnic minority patients and those for whom English is not their first language seem to be better targeted. Given the number of patients who do not know how to access online services or do not know how to get support in using them, there is unmet need for increasing engagement with online primary care services by providing more digital facilitation to support the rollout of digital services and particularly to provide targeted support for older patients.

Acknowledgments

This study was funded by the National Institute for Health and Care Research Health and Social Care Delivery Research Programme (128268). The views expressed in this paper are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. The authors would like to acknowledge and thank the following people for their contributions to the research project from which this paper was developed: research team colleagues Wiktoria Bak, Hamish Evans, Evangelos Gkousis, Georgia Jenkins, Caroline Jenkinson, Nada Khan, Brandi Leach, Jennifer Newbould, Sarah Parkinson, Joanne Parsons, Emma Pitchforth, Laura Sheard, Stephanie Stockwell, and Bethan Treadgold; patient advisory group members Laiya Bailey, Julie Harvey, Diana Lewis-Frost, Ben Marshall, Andrea Shelly, Lynn Tatnell, Penny Thompson, and Malcolm Turner; steering group committee members Jeremy Horwood, David Parkes, and Amrit Takhar; sponsors Pam Baxter and Vivienne Shaw; sponsor team members Sarah Hider and Alexander Philips; administrator Ellie Kingsland; survey support staff members Jodie Button, Mayam Gomez-Cano, Sine MacDonald, Isabelle Mayne, Charlotte Reburn, and Claire Wright; initial study advice from Richard Moore; clinical research networks East of England, North Thames, North West Coast, South West Peninsula, and West Midlands; and all patients and practice staff who supported and participated in the study.

XSL•FO
RenderX

Authors' Contributions

RW was responsible for the overall project management as well as oversight of the conduct of the surveys and jointly led the drafting of the manuscript. JLC led and contributed to the overall study design and contributed to the design and conduct of the surveys as well as to the drafting of the manuscript. NA conducted data analysis, modeling, and synthesis for the surveys and contributed to the drafting of the manuscript. AA conducted data analysis, modeling, and synthesis for the surveys and contributed to the drafting of the manuscript. JL contributed to the design, conduct, and analysis of the surveys as well as to the drafting of the manuscript. EC coordinated and led patient and public involvement and engagement and contributed to the drafting of the manuscript. CT provided administrative and data processing support for the surveys and contributed to the drafting of the manuscript. CEC contributed to the design and conduct of the surveys as well as to the drafting of the manuscript. CB contributed to the design and conduct of the surveys as well as to the drafting of the manuscript. JS conceived and contributed to the overall study design as well as to the drafting of the manuscript. HA contributed to the overall study design and to the drafting of the manuscript. CM acted as patient and public involvement and engagement representative and contributed to the design of the surveys as well as to the drafting of the manuscript. GA was responsible for overall study design as well as for the coordination of the surveys' design, conduct, and analyses and jointly led the drafting of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Practice survey: questionnaire.
[DOCX File , 932 KB - jmir_v26i1e56528_app1.docx]

Multimedia Appendix 2
Patient survey: questionnaire.
[DOCX File , 1078 KB - jmir_v26i1e56528_app2.docx]

Multimedia Appendix 3
Digital confidence score development.
[DOCX File , 13 KB - jmir_v26i1e56528_app3.docx]

Multimedia Appendix 4
Patients in practices in England: comparisons between survey practices and other practices.
[DOCX File , 16 KB - jmir_v26i1e56528_app4.docx]

Multimedia Appendix 5
Online services offered by practices: current and before the COVID-19 pandemic.
[DOCX File , 16 KB - jmir_v26i1e56528_app5.docx]

Multimedia Appendix 6
Practice activities to promote, help or support patients use online services.
[DOCX File , 16 KB - jmir_v26i1e56528_app6.docx]

Multimedia Appendix 7
Services that have been promoted and supported by practices.
[DOCX File , 16 KB - jmir_v26i1e56528_app7.docx]

Multimedia Appendix 8
Staff roles involved.
[DOCX File , 14 KB - jmir_v26i1e56528_app8.docx]

Multimedia Appendix 9
Practice responses to questions about patient groups.
[DOCX File , 15 KB - jmir_v26i1e56528_app9.docx]

Multimedia Appendix 10
Removal or reduction of offline services since the start of the COVID-19 pandemic.

XSL•FO
RenderX

[DOCX File , 17 KB - jmir_v26i1e56528_app10.docx]

Multimedia Appendix 11
Practice survey: agreement or disagreement with statements.
[DOCX File , 19 KB - jmir_v26i1e56528_app11.docx]

Multimedia Appendix 12
Awareness of online toolkits.
[DOCX File , 15 KB - jmir_v26i1e56528_app12.docx]

Multimedia Appendix 13
Practice recruitment, patient survey invitation, and response.
[DOCX File , 14 KB - jmir_v26i1e56528_app13.docx]

Multimedia Appendix 14
Patient survey: invitations and response by deprivation quintile.
[PNG File , 246 KB - jmir_v26i1e56528_app14.png]

Multimedia Appendix 15
Patient survey: digital confidence score.
[DOCX File , 13 KB - jmir_v26i1e56528_app15.docx]

Multimedia Appendix 16
Logistic regression analyses: unadjusted model 1 and adjusted model 2.
[DOCX File , 41 KB - jmir_v26i1e56528_app16.docx]

Multimedia Appendix 17
Sensitivity analyses: adjusted models 1 and 2 (excludes those responding yes to the question “Did you need help to complete this survey?”).
[DOCX File , 31 KB - jmir_v26i1e56528_app17.docx]

Multimedia Appendix 18
Comparison of patient-reported awareness and the use of modes of digital facilitation with practice-reported use of facilitation mode.
[DOCX File , 15 KB - jmir_v26i1e56528_app18.docx]

Multimedia Appendix 19
Combined analyses of practice and patient surveys: patient awareness.
[DOCX File , 14 KB - jmir_v26i1e56528_app19.docx]

Multimedia Appendix 20
Combined analyses of practice and patient surveys: targeted patient awareness.
[DOCX File , 16 KB - jmir_v26i1e56528_app20.docx]

References

1. The future of human resources and organisational development. NHS England. 2021. URL: https://tinyurl.com/mr3mdjdc [accessed 2023-08-29]
2. Delivery plan for recovering access to primary care. NHS England. 2023. URL: https://www.england.nhs.uk/long-read/delivery-plan-for-recovering-access-to-primary-care-2/ [accessed 2024-03-20]
3. A plan for digital health and social care. Appendix A: our action plan for delivering a digital future. UK Department of Health and Social Care. 2022. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2023-01-17]
4. About GP online services. NHS England. URL: https://www.england.nhs.uk/gp-online-services/about-the-prog/ [accessed 2023-08-29]

XSL•FO
RenderX

5. Vimalananda VG, Gupte G, Seraj SM, Orlander J, Berlowitz D, Fincke BG, et al. Electronic consultations (e-consults) to improve access to specialty care: a systematic review and narrative synthesis. J Telemed Telecare 2015 Sep;21(6):323-330 [FREE Full text] [doi: 10.1177/1357633X15582108] [Medline: 25995331]
6. Clarke G, Dias A, Wolters A. Access to and delivery of general practice services: a study of patients at practices using digital and online tools. The Health Foundation, London. 2022. URL: https://www.health.org.uk/publications/access-to-and-delivery-of-general-practice-services [accessed 2023-01-17]
7. NHS England, BMA. Investment and evolution: a five-year framework for GP contract reform to implement the NHS Long Term Plan. BMA. 2019 Jan 31. URL: https://www.bma.org.uk/media/1461/bma-investment-and-evolution-five-year-framework-gp-contract-jan-2019.pdf [accessed 2023-02-28]
8. Solans O, Vidal-Alaball J, Roig Cabo P, Mora N, Coma E, Bonet Simó JM, et al. Characteristics of citizens and their use of teleconsultations in primary care in the Catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629 [FREE Full text] [doi: 10.2196/28629] [Medline: 33970867]
9. Clarke G, Pariza P, Wolters A. How has COVID-19 affected service delivery in GP practices that offered remote consultations before the pandemic? The Health Foundation. 2020. URL: https://www.health.org.uk/news-and-comment/charts-and-infographics/how-has-covid-19-affected-service-delivery-in-gp-practices [accessed 2023-01-10]
10. Johnsen TM, Norberg BL, Kristiansen E, Zanaboni P, Austad B, Krogh FH, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 08;23(2):e26433 [FREE Full text] [doi: 10.2196/26433] [Medline: 33465037]
11. Zanaboni P, Kummervold PE, Sørensen T, Johansen MA. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020 Feb 07;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
12. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440 [FREE Full text] [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
13. Honeyman M, Maguire D, Evans H, Davies A. Digital technology and health inequalities: a scoping review. Cardiff: Public Health Wales NHS Trust. 2020. URL: https://phw.nhs.wales/publications/publications1/digital-technology-and-health-inequalities-a-scoping-review/ [accessed 2024-01-17]
14. Turner A, Morris R, Rakhra D, Stevenson F, McDonagh L, Hamilton F, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2021 Oct 20;72(715):e128-e137. [doi: 10.3399/bjgp.2021.0426]
15. Heponiemi T, Kaihlanen A, Kouvonen A, Leemann L, Taipale S, Gluschkoff K. The role of age and digital competence on the use of online health and social care services: a cross-sectional population-based survey. Digit Health 2022;8:1-10 [FREE Full text] [doi: 10.1177/20552076221074485] [Medline: 35111333]
16. Leach B, Parkinson S, Gkousis E, Abel G, Atherton H, Campbell J, et al. Digital facilitation to support patient access to web-based primary care services: scoping literature review. J Med Internet Res 2022 Jul 14;24(7):e33911 [FREE Full text] [doi: 10.2196/33911] [Medline: 35834301]
17. Leach B, Gkousis E, Parkinson S, Newbould J, Sussex J. Digital facilitation to support patient access to online primary care services. National Institute for Health Research. 2022. URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=189019 [accessed 2023-02-07]
18. Abel G, Atherton H, Sussex J, Atker N, Aminu AQ, Bak W, et al. Current experience and future potential of facilitating access to digital NHS primary care services in England: the Di-Facto mixed-methods study. In: NIHR Health and Social Care Delivery Research Programme (forthcoming).
19. Widening digital participation programme. Good Things Foundation. 2022. URL: https://www.goodthingsfoundation.org/what-we-do/our-partnerships/health-and-wellbeing/widening-digital-participation/ [accessed 2023-01-23]
20. Locked out: digitally excluded people's experiences of remote GP appointments. Healthwatch England. 2021. URL: https://www.healthwatch.co.uk/report/2021-06-16/locked-out-digitally-excluded-peoples-experiences-remote-gp-appointments [accessed 2023-11-09]
21. Brant H, Atherton H, Ziebland S, McKinstry B, Campbell JL, Salisbury C. Using alternatives to face-to-face consultations: a survey of prevalence and attitudes in general practice. Br J Gen Pract 2016 Jul;66(648):e460-e466 [FREE Full text] [doi: 10.3399/bjgp16X685597] [Medline: 27215571]
22. GP patient survey 2021: technical annex. NHS England & Ipsos Mori. URL: https://gp-patient.co.uk/downloads/2021/GPPS_2021_Technical_Annex_PUBLIC.pdf [accessed 2024-05-21]
23. GP patient survey 2022: technical annex. NHS England & Ipsos Mori. URL: https://gp-patient.co.uk/downloads/2022/GPPS_2022_Technical_Annex_PUBLIC.pdf [accessed 2023-09-07]
24. Improving inclusion of under-served groups in clinical research: guidance from INCLUDE project. National Institute for Health and Care Research (NIHR). 2020. URL: https://www.nihr.ac.uk/documents/improving-inclusion-of-under-served-groups-in-clinical-research-guidance-from-include-project/25435 [accessed 2023-12-04]

XSL•FO
RenderX

25. Lin X, Wang M, Zuo Y, Li M, Lin X, Zhu S, et al. Health literacy, computer skills and quality of patient-physician communication in Chinese patients with cataract. PLoS One 2014 Sep 16;9(9):e107615 [FREE Full text] [doi: 10.1371/journal.pone.0107615] [Medline: 25226299]
26. Taylor AH, Taylor RS, Ingram WM, Anokye N, Dean S, Jolly K, et al. Adding web-based behavioural support to exercise referral schemes for inactive adults with chronic health conditions: the e-coachER RCT. Health Technol Assess 2020 Nov;24(63):1-106 [FREE Full text] [doi: 10.3310/hta24630] [Medline: 33243368]
27. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
28. Karnoe A, Furstrand D, Christensen KB, Norgaard O, Kayser L. Assessing competencies needed to engage with digital health services: development of the eHealth literacy assessment toolkit. J Med Internet Res 2018 May 10;20(5):e178 [FREE Full text] [doi: 10.2196/jmir.8347] [Medline: 29748163]
29. GP patient survey: see how your GP practice is doing or compare practices. GP Patient Survey. 2020. URL: https://www.gp-patient.co.uk/ [accessed 2023-12-03]
30. Khan N, Pitchforth E, Winder R, Abel G, Clark CE, Cockcroft E, et al. What helps patients access web-based services in primary care? free-text analysis of patient responses to the Di-Facto questionnaire. BMC Prim Care 2024 Jan 10;25(1):20 [FREE Full text] [doi: 10.1186/s12875-023-02257-5] [Medline: 38200431]
31. Patients registered at a GP practice December 2020. NHS England. 2020. URL: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/december-2020 [accessed 2023-06-08]
32. GP patient survey 2019: technical annex. NHS England & Ipsos Mori. 2019. URL: https://gp-patient.co.uk/downloads/archive/2019/GPPS_2019_Technical_Annex_PUBLIC.pdf [accessed 2023-01-18]
33. Fingertips: public health profiles. Public Health England. 2022. URL: https://fingertips.phe.org.uk/ [accessed 2023-01-18]
34. Rughani G, Hanlon P, Corcoran N, Mair FS. The readability of general practice websites: a cross-sectional analysis of all general practice websites in Scotland. Br J Gen Pract 2021 Jan 05;71(706):e391-e398. [doi: 10.3399/bjgp.2020.0820]
35. Ethnic health inequalities and the NHS: driving progress in a changing system. The King's Fund. 2022. URL: https://www.nhsrho.org/research/ethnic-inequalities-in-healthcare-a-rapid-evidence-review-2/ [accessed 2024-01-17]
36. The NHS Long Term Plan. NHS England. 2019. URL: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/01/nhs-long-term-plan-june-2019.pdf [accessed 2024-04-24]
37. Bryce C, O'Connell MD, Dale J, Underwood M, Atherton H. Online and telephone access to general practice: a cross-sectional patient survey. BJGP Open 2021 Jun 09;5(4):1-9 [FREE Full text] [doi: 10.3399/bjgpo.2020.0179]
38. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating Patient Use and Experience of Online Appointment Booking in Primary Care: Mixed Methods Study. J Med Internet Res 2024 Jul 08;26:e51931 (forthcoming) [FREE Full text] [doi: 10.2196/51931] [Medline: 38976870]
39. Nijland N, van Gemert-Pijnen JE, Kelders SM, Brandenburg BJ, Seydel ER. Factors influencing the use of a web-based application for supporting the self-care of patients with type 2 diabetes: a longitudinal study. J Med Internet Res 2011 Sep 30;13(3):e71 [FREE Full text] [doi: 10.2196/jmir.1603] [Medline: 21959968]
40. Rodgers M, Raine G, Thomas S, Harden M, Eastwood A. Informing NHS policy in 'digital-first primary care': a rapid evidence synthesis. Health Serv Deliv Res 2019 Dec;7(41):1-124 [FREE Full text] [doi: 10.3310/hsdr07410]
41. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
42. People who cannot speak English well are more likely to be in poor health. Office of National Statistics. 2015. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/articles/peoplewhocannotspeakenglishwellaremorelikelytobeinpoorhealth/2015-07-09 [accessed 2023-11-30]
43. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman Med J 2020 Mar;35(2):e122 [FREE Full text] [doi: 10.5001/omj.2020.40] [Medline: 32411417]
44. Kristiansen E, Atherton H, Austad B, Bergmo T, Norberg BL, Zanaboni P. Older patients' experiences of access to and use of e-consultations with the general practitioner in Norway: an interview study. Scand J Prim Health Care 2023 Mar;41(1):33-42 [FREE Full text] [doi: 10.1080/02813432.2022.2161307] [Medline: 36592342]
45. Turner A, Morris R, McDonagh L, Hamilton F, Blake S, Farr M, et al. Unintended consequences of patient online access to health records: a qualitative study in UK primary care. Br J Gen Pract 2023 Jan;73(726):e67-e74 [FREE Full text] [doi: 10.3399/BJGP.2021.0720] [Medline: 36316163]
46. Peeters JM, Krijgsman JW, Brabers AE, Jong JD, Friele RD. Use and uptake of eHealth in general practice: a cross-sectional survey and focus group study among health care users and general practitioners. JMIR Med Inform 2016 Apr 06;4(2):e11 [FREE Full text] [doi: 10.2196/medinform.4515] [Medline: 27052805]
47. Newbould J, Abel G, Ball S, Corbett J, Elliott M, Exley J, et al. Evaluation of telephone first approach to demand management in English general practice: observational study. BMJ 2017 Sep 27;358:j4197 [FREE Full text] [doi: 10.1136/bmj.j4197] [Medline: 28954741]
48. Next steps for integrating primary care: fuller Stocktake report. NHS England. 2022. URL: https://www.england.nhs.uk/wp-content/uploads/2022/05/next-steps-for-integrating-primary-care-fuller-stocktake-report.pdf [accessed 2023-01-18]

XSL•FO
RenderX

49. The Hewitt review: an independent review of integrated care systems. UK Government. 2023. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1148568/the-hewitt-review.pdf [accessed 2923-05-21]
50. Ethnic group, England and Wales: census 2021. Office of National Statistics. 2022. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021 [accessed 2023-01-12]
51. Delivery plan for recovering access to primary care: update and actions for 2024/25. NHS England. 2024. URL: https://www.england.nhs.uk/long-read/delivery-plan-for-recovering-access-to-primary-care-update-and-actions-for-2024-25/ [accessed 2024-05-16]
52. Castle-Clarke S. What will new technology mean for the NHS and its patients?: four big technological trends. The Health Foundation, The Institute for Fiscal Studies, The King's Fund and the Nuffield Trust. 2018. URL: https://www.nuffieldtrust.org.uk/files/2018-06/1530028974_the-nhs-at-70-what-will-new-technology-mean-for-the-nhs-and-its-patients.pdf [accessed 2023-02-07]
53. Campbell J, Atherton H, Abel G, Sussex J. Digital facilitation: supporting the use of NHS digital and online services. Br J Gen Pract 2024 Mar 27;74(741):150-151. [doi: 10.3399/bjgp24x736749]
54. Verity A, Tzortziou Brown V. Inclusion health patient perspectives on remote access to general practice: a qualitative study. BJGP Open 2023 Feb 22;7(2):1-12. [doi: 10.3399/bjgpo.2023.0023]

Abbreviations

GP: general practitioner
NHS: National Health Service
OR: odds ratio
PPIE: patient and public involvement and engagement

Edited by N Cahill; submitted 01.02.24; peer-reviewed by SQ Yew; comments to author 20.03.24; revised version received 11.04.24; accepted 26.06.24; published 07.08.24.

Please cite as:
Winder R, Campbell JL, Akter N, Aminu AQ, Lambert J, Cockcroft E, Thomas C, Clark CE, Bryce C, Sussex J, Atherton H, Marriott C, Abel G
Exploring How Patients Are Supported to Use Online Services in Primary Care in England Through "Digital Facilitation": Survey Study
J Med Internet Res 2024;26:e56528
URL: https://www.jmir.org/2024/1/e56528
doi:10.2196/56528
PMID:

©Rachel Winder, John L Campbell, Nurunnahar Akter, Abodunrin Q Aminu, Jeffrey Lambert, Emma Cockcroft, Chloe Thomas, Christopher E Clark, Carol Bryce, Jon Sussex, Helen Atherton, Christine Marriott, Gary Abel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Media Usage Behavior and Its Impact on the Physician-Patient Relationship: Cross-Sectional Study Among Individuals Affected by Psoriasis in Germany

Mert Ege Erbas[1], MD; Stefanie Ziehfreund[1], MSc; Hannah Wecker[1], MSc; Tilo Biedermann[1], MD; Alexander Zink[1,2], MPH, MD, PhD

[1]Department of Dermatology and Allergy, School of Medicine, Technical University of Munich, Munich, Germany

[2]Division of Dermatology and Venereology, Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden

Corresponding Author:
Alexander Zink, MPH, MD, PhD
Department of Dermatology and Allergy
School of Medicine
Technical University of Munich
Biedersteiner Str 29
Munich, 80802
Germany
Phone: 49 8941403176
Email: alexander.zink@tum.de

Abstract

Background: Psoriasis is a chronic skin disorder with a high burden of disease. People affected with psoriasis increasingly use the internet for health-related reasons, especially those with younger age, higher education, and higher disease severity. Despite advantages such as enhancing the individuals' knowledge with the use of digital media for health-related issues, disadvantages were also present such as quality control, and variability in the individuals' health information literacy. While patients with psoriasis within medical settings generally trust physicians over digital media, they commonly withhold their web-based research findings from health care providers.

Objective: The study aims to (1) identify further factors associated with regular psoriasis-related internet use, (2) rank specific digital media platforms used, and (3) examine digital media within the physician-patient relationship among individuals with and without dermatological treatment.

Methods: A cross-sectional, questionnaire-based study was conducted among individuals with self-reported psoriasis in Germany between September 2021 and February 2022. Participants were recruited via digital media platforms and in person at a University Hospital Department of Dermatology in southern Germany. The questionnaire asked about demographic and medical information, individual psoriasis-related digital media use, and the impact of digital media on the physician-patient relationship. Data were analyzed descriptively, and logistic regression models were performed to assess the factors associated with regular psoriasis-related internet use.

Results: Among 321 individuals with a median age of 53 (IQR 41-61) years (nonnormally distributed; females: 195/321), female sex, shorter disease duration, moderate mental burden of disease, and good self-assessed psoriasis-related knowledge were associated with regular psoriasis-related internet use. Of the 188 participants with a mean age of 51.2 (SD 13.9) years (normally distributed) who used digital media 106 (56.4%) usually searched for information on psoriasis-based websites and 98 (52.1%) on search engines, primarily for obtaining information about the disease and therapy options, while social media were less frequently used (49/188, 26.1%). Nearly two-thirds of internet users (125/188) claimed that their physicians did not recommend digital media platforms. About 44% (82/188) of the individuals reported to seek for additional information due to the insufficient information provided by their physician.

Conclusions: This study revealed the importance of digital media in the context of psoriasis, especially among women, individuals with shorter disease duration, and moderate mental disease severity. The lack of physicians' digital media recommendations despite their patients' desire to receive such and being more involved in health-related decisions seems to be a shortcoming within the physician-patient relationships. Physicians should guide their patients on digital media by recommending platforms with evidence-based information, thereby potentially creating an adequate framework for shared decision-making. Future research

XSL•FO
RenderX

should focus on strategies to prevent the spread of false information on digital media and address the needs of patients and physicians to enhance health-related digital media offerings.

(J Med Internet Res 2024;26:e57823) doi:10.2196/57823

KEYWORDS

psoriasis; dermatology; digital health; digital media; internet use; questionnaire; physician-patient relationship

Introduction

Background

Psoriasis is a chronic inflammatory skin disease with a global prevalence ranging from 0.51% to 11.43% among adults [1]. Its etiology is multifactorial and influenced by genetic, environmental, and immunological factors [2]. Potential triggers include stress, infections, smoking, alcohol consumption, and specific medications like lithium [3]. The choice of psoriasis treatment depends on the disease severity, presence of psoriatic arthritis, comorbidities, and patient preferences. Established therapeutic options include topical therapy, phototherapy, oral-systemic therapy (eg, methotrexate), and biologics [3]. Individuals with psoriasis experience an increased mental burden due to stigmatization [4], a higher risk of addictive behavior [5], depression, and reduced overall well-being [6].

There has been a notable surge in the health-related use of digital media by individuals with dermatological conditions, particularly young and highly educated individuals facing socially burdensome skin conditions [7-9]. Digital media involves the transmission of information as digital data, encompassing different types of content such as audio, video, graphics, and text via digital devices such as computers, tablets, and smartphones. Examples range from social media platforms (eg, Facebook) and video streaming services (eg, YouTube) to digital radio stations and literary websites (eg, Wikipedia) [10]. In Germany, the internet has been increasingly used for accessing psoriasis-related information [11]. A prior study indicated that the psoriasis-related use of social media and the internet was more prevalent among individuals, who were young, had higher education, had higher impairment in skin disease-specific quality of life and severity, and those experiencing symptoms such as arthritis or involvement of facial or genital areas [12]. Additionally, Google and Facebook were found to be the most frequently used platforms among individuals affected by psoriasis [12].

Engaging with digital media platforms for health-related information provides several advantages. Notably, it empowers individuals by enhancing their knowledge, competence, and engagement in health decision-making processes [13]. Furthermore, it strengthens the physician-patient relationship and enables affected individuals to find answers, explore sensitive topics, and ask additional questions, from the comfort of their homes [14,15]. Blogs and digital support communities provide personalized insights and reflections and reduce feelings of loneliness and isolation [13].

However, there are also challenges regarding the use of health-related digital media. Regulating digital health information proves to be challenging due to the difficulties in maintaining quality control and the variability in the affected individual's health information literacy [16]. Misleading or sensationalized content can result in harmful health decision-making, while the endorsement of unscientific practices further exacerbates the complexity [13,16]. Vulnerable individuals desiring empowerment may accept information without critical evaluation or may misinterpret it, thereby fostering erroneous knowledge and noncompliance [13,14]. Although the traditional role of physicians in providing personalized and evidence-based information may have been perceived as paternalistic, it ensured appropriate guidance based on individual comprehension [13].

Regarding the impact of digital media on the physician-patient relationship, a quantitative study indicated that physicians could recommend reputable digital media platforms; thereby, enhancing the quality of patient care [17]. Moreover, individuals with psoriasis tend to place greater faith in their physicians' knowledge when disparities arise between information obtained from social media and the advice provided by physicians [12]. However, patients typically choose not to disclose their web-based research findings to their physicians [12].

Existing research predominantly centers on social media and the broader internet, primarily surveying patients within medical settings [7,12,18]. However comprehensive investigations that offer a more nuanced examination of digital media, encompassing individuals beyond the medical setting, remain conspicuously scarce [19].

Objectives

The aims of this study were to (1) examine sociodemographic and medical factors influencing the health-related use of the internet among individuals affected by psoriasis, (2) rank specific digital media platforms and assess reasons for their use in the context of psoriasis, and (3) examine digital media within the physician-patient relationship from the perspective of individuals affected by psoriasis with and without dermatological treatment.

Methods

Study Design

For this cross-sectional study, an anonymous questionnaire was conducted among individuals affected by psoriasis in Germany between September 2021 and February 2022.

Data Collection Methods

The questionnaire was designed by an interdisciplinary team, including 1 dermatologist, 2 epidemiologists, and 2 medical students. The questions are based on literature research [11,19,20] and qualitative interviews about digital media

XSL•FO
RenderX

conducted by the same interdisciplinary team among psoriasis-affected individuals. The survey contained 20 items with multiple-choice and free-text answers and was split into 2 parts.

The first part contained the initial 12 items and could be filled out by every participant. This section covered demographic and medical information as well as information about psoriasis-related internet use: age (years), sex, educational level, place of residence, disease duration (years), general disease severity (1 item question), mental burden of disease (1 item question), presence of psoriasis arthritis, dermatological treatment (yes or no), psoriasis knowledge, type of recruitment (digital media or clinical setting) and frequency of internet use regarding psoriasis. The items assessing the mental burden of disease, general disease severity, and psoriasis knowledge relied on self-reported data rather than validated instruments.

Participants who used the internet at least "less than once a month" for psoriasis-related information were considered internet users and eligible for the second part of the questionnaire. The second part contained the remaining 8 items of the questionnaire and asked about the following digital media platforms in the context of psoriasis: search engines, Wikipedia, general health websites, psoriasis-related websites, Facebook, other social media (Instagram, YouTube, and Twitter), and other forums or blogs. Additionally, the physician-patient relationship in the context of digital media was analyzed. For the second part, the following components and variables were used: frequency of each portal used, the best portal (single choice), search for treatment centers (yes or no), the communication of digital media use to the physician (yes or no), and if applicable physician response (positive, neutral, or negative), various statements regarding the physician-patient relationship, and an indication of reasons for using individual portals (multiple answer options). For the reasons for using individual portals, Facebook was considered separately since it is the most used social media platform in Germany and worldwide [21,22].

The survey was pilot-tested by 2 patients, resulting in minor changes in the wording of some questions. The final version of the questionnaire was incorporated into the platform REDCap (Research Electronic Data Capture; Vanderbilt University) to provide the digital questionnaire and to digitize the paper-based questionnaires.

Sample Characteristics

Eligible for inclusion were individuals who were at least 18 years old, had psoriasis diagnosed by a physician (self-reported), and were able to answer a German questionnaire. Recruitment via digital media was initiated by sharing a link to the digital questionnaire in several psoriasis-related Facebook groups. These groups were general psoriasis support groups specifically aimed at German-speaking individuals. Collaboration with 2 prominent patient organizations ("Psoriasis-Netz" and "Farbenhaut") was established, and these organizations shared the questionnaire link on their websites and in newsletters. The largest national psoriasis group in Germany ("Deutscher Psoriasis Bund eV") also supported the recruitment efforts by disseminating the questionnaire link through their newsletter. Anyone who accessed the questionnaire via this link could participate in the survey. Moreover, patients were recruited personally at the Department for Dermatology and Allergy at the University Hospital of the Technical University of Munich. The clinical staff approached eligible inpatients consecutively, identified through their medical records as having psoriasis, and invited them to participate in the study. Similarly, clinical staff approached outpatients visiting the dermatology department for appointments consecutively. Both inpatients and outpatients were provided with detailed information about the study and, if they agreed to participate, were given a paper-based questionnaire to complete during their hospital stay or appointment. All participants, whether recruited via digital media or in person, received the same introductory information about the study to ensure consistency in understanding the study's purpose and procedures.

Ethical Considerations

The study was reviewed and approved by the local ethics committee of the Medical Faculty of the Technical University of Munich (reference 129/20 S). Informed consent was obtained from all participants. For the digital questionnaire, confirmation was necessary to complete the questionnaire. For the paper-based questionnaire, only those participants who confirmed the consent form were included in the statistical analysis. To preserve anonymity, the questionnaires did not ask for any identifying information such as name or address. For the paper-based questionnaires administered in the clinical setting, patients received the questionnaire from clinic staff who were not involved in this study. The patients then returned the filled-out questionnaires to the clinic staff who subsequently forwarded them to the study team. This process ensured that the questionnaires could not be traced back to individual participants. The participants in this study did not receive any form of compensation.

Statistical Analysis

Data were descriptively analyzed. Absolute and relative frequencies were calculated for all categorical variables. For the reasons for using the individual portals, multiple answers were allowed so that the cumulative frequency or percentage could exceed 100%. For continuous variables, descriptive parameters (ie, mean and SD in case of normal distribution, median and IQR if not normally distributed, range) were determined, and histograms were constructed to evaluate distributions. The frequency at each portal was used and the variable frequency of internet use regarding psoriasis referred to the last 6 months and was categorized into "irregular use" (containing the answer options "not at all" and "less than once a month") and "regular use" ("at least once a month," "at least once a week," and "daily"). The chi-square tests were used to analyze the association between the dichotomized variable frequency of internet use regarding psoriasis and the other nominal variables. In addition, Mann-Whitney and Kruskal-Wallis tests were used to assess the differences between the dichotomized variable frequency of internet use regarding psoriasis and the ordinal variables, as well as the nonnormally distributed metric variables disease duration and age. For the subsample of internet users, unpaired *t* tests were performed with the normally distributed metric variable age and the

XSL•FO
RenderX

variable frequency of use of each portal to discover a possible difference in the age regarding the frequency of use of individual digital media platforms.

Univariate and multivariate binary logistic regressions were used to determine the associations between the dependent variable frequency of internet use regarding psoriasis and the following independent variables: age, disease duration, sex, education level, place of residence, general disease severity, mental burden of disease, presence of psoriasis arthritis, dermatological treatment, and psoriasis knowledge. To assess multicollinearity, Spearman correlation (ρ) and variance inflation factor (VIF) were used, with correlation ρ>0.7 or VIF≥10 being used cutoff values for multicollinearity [23,24]. If at least 1 cutoff value was exceeded, 1 of the affected variables had to be removed. Results of the regression analyses were reported with odds ratios (ORs), 95% CI, and *P* values.

For the frequency of each portal used, the variable best portal, and the indication of reasons for using individual portals, free-text responses were possible in addition to the answer options given. The free-text responses were recoded whenever they described a response option already listed.

Only the complete cases were used for all statistical analyses. The level of significance was set at α=.05 for all analyses. The SPSS statistics software (version 29; IBM Corp) was used for data analyses.

Results

Distribution of Total Data

In total, 462 individuals participated between September 2021 and February 2022. Initially, 44 participants who did not agree to the consent form (n=13) or whose psoriasis was not diagnosed by a physician (n=31) were excluded. Thus, based on these 418 questionnaires (digital: n=297 and paper-based: n=121), the missing data were excluded according to the research question (Figure 1).

Figure 1. Flowchart of the study population and distribution.

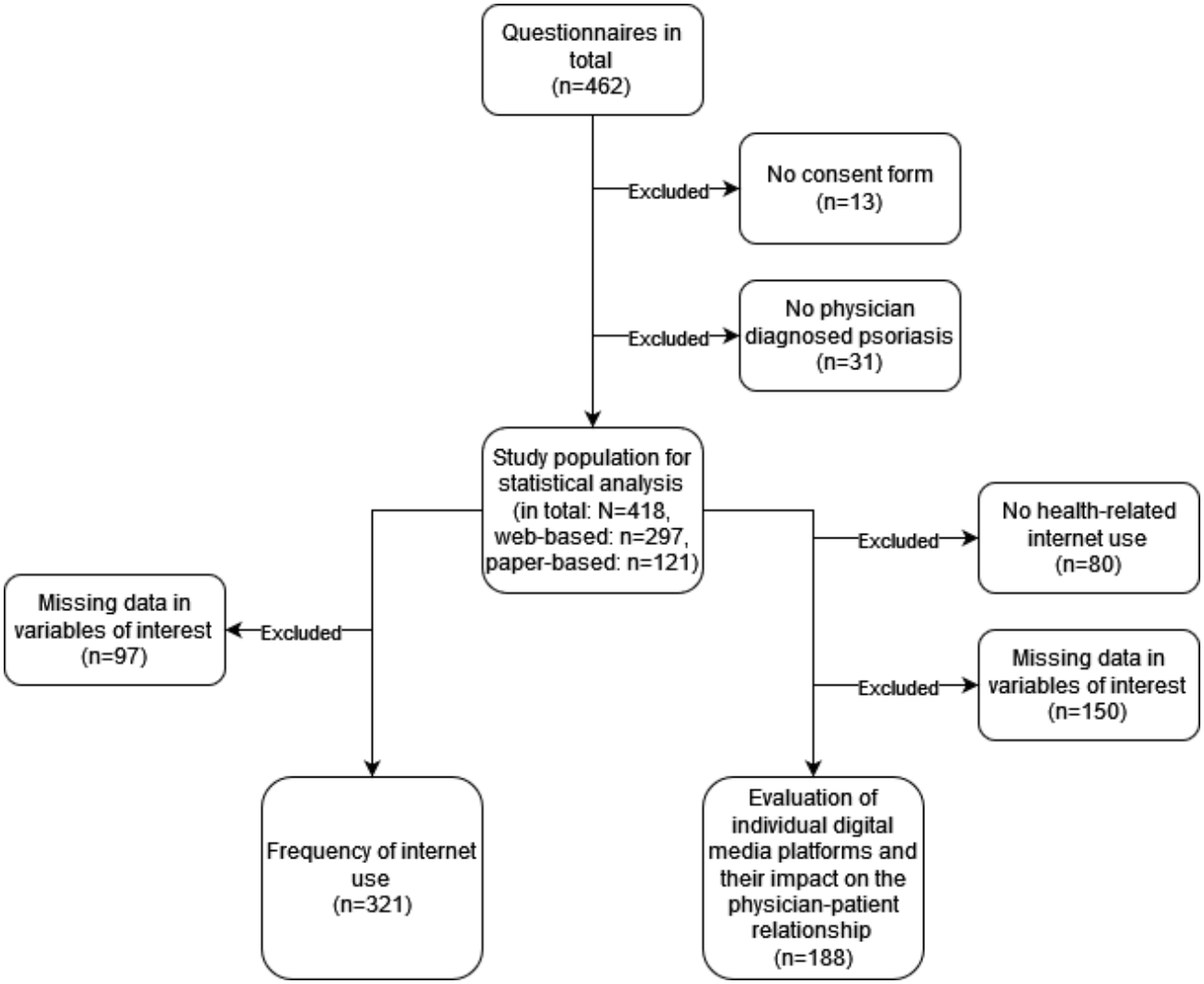

Frequency of Internet Use Regarding Psoriasis

In total, 321 participants were included with 195 (60.7%) being female (Table 1). The median age was 53.0 (IQR 41.0-61.0; range 18.0-82.0) years and the median disease duration was 22.0 (IQR 10.0-37.5) years with most of the participants reporting moderate or severe psoriasis (moderate: 147/321, 45.8%; severe: 142/321, 44.2%) and moderate to severe mental burden (moderate: 146/321, 45.5%; severe: 81/321, 29.3%). Overall, 71% of the participants (228/321) were treated by

XSL•FO
RenderX

dermatologists. In the last 6 months, 54.2% (174/321) of participants used the internet regarding psoriasis regularly. Participants with a shorter disease duration were more likely to use the internet regularly for psoriasis-related information in the last 6 months (regular use: median 19.5, IQR 8-36 years vs irregular use: median 26, IQR 15-40 years; *P*=.007). There was no age-related difference in terms of the frequency of internet use regarding psoriasis (regular use: median 52.5, IQR 43-60.3 years vs irregular use: median 53, IQR 40-62 years; *P*=.81). In addition, the female sex and good or very good psoriasis knowledge were associated with a more regular internet use regarding psoriasis compared to male sex (*P*=.001) or less than good knowledge (*P*=.08). Individuals recruited from the clinical setting were less likely to use digital media regularly compared to individuals recruited through digital media (*P*<.001).

No violations of multicollinearity could be detected by checking the Spearman correlations and the VIF. The binary logistic regression analyses (Table 2) showed that participants with shorter disease duration (univariate: OR 0.99, 95% CI 0.97-1.00; multivariate: OR 0.98, 95% CI 0.96-1.00) and who were female (univariate: OR 2.13, 95% CI 1.35-3.36; multivariate: OR 1.70, 95% CI 1.03-2.81) compared to males were more likely to use the internet regarding psoriasis regularly. However, for age (univariate: OR 1.00, 95% CI 0.98-1.01; multivariate: OR 1.01, 95% CI 0.99-1.03), education (univariate: 0.78, 95% CI 0.50-1.21; multivariate: OR 0.88, 95% CI 0.54-1.44), and place of residence (univariate: OR 0.78, 95% CI 0.50-1.24; multivariate: OR 0.74, 95% CI 0.45-1.21), an association with the frequency of psoriasis-related internet use could not be found. Participants considering their psoriasis knowledge as good or very good had higher odds of regularly using the internet for psoriasis than people with less knowledge about their disease (multivariate: OR 1.97, 95% CI 1.01-3.83). In addition, the odds for regular psoriasis-related internet use were 80% higher in participants with a moderate mental burden of disease (multivariate: OR 1.80, 95% CI 1.01-3.20) compared to participants with no or mild mental distress.

XSL•FO
RenderX

Table 1. Characteristics of the study population overall (n=321) and stratified by the frequency of internet use regarding psoriasis.

Patients' characteristics	Overall (n=321)	Internet use		*P* value
		Regular (n=174, 54.2 %)	Irregular (n=147, 45.8 %)	
Age (years)				.81[a]
Median (IQR)	53.0 (41.0-61.0)	52.5 (43.0-60.3)	53.0 (40.0-62.0)	
Range	18.0-82.0	21.0-82.0	18.0-81.0	
Sex, n (%)				.001[b]
Male	126 (39.3)	54 (31)	72 (49)	
Female	195 (60.7)	120 (69)	75 (51)	
Diverse	0 (0)	0 (0)	0 (0)	
Education level, n (%)				.27[a]
No access to university	144 (44.9)	83 (47.7)	61 (41.5)	
Access to university	177 (55.1)	91 (52.3)	86 (58.5)	
Place of residence, n (%)				.30[a]
Cities with up to 100,000 inhabitants	202 (62.9)	114 (65.5)	88 (59.9)	
Cities with over 100,000 inhabitants	119 (37.1)	60 (35.5)	59 (40.1)	
Disease duration (years)				.007[a]
Median (IQR)	22.0 (10.0-37.5)	19.5 (8.0-36.0)	26.0 (15.0-40.0)	
Range	0.5-77.0	0.5-70.0	1.5-77.0	
General disease severity, n (%)				.09[c]
Mild	32 (10)	14 (8)	18 (12.2)	
Moderate	147 (45.8)	89 (51.1)	58 (39.5)	
Severe	142 (44.2)	71 (40.8)	71 (48.3)	
Mental burden of disease, n (%)				.15[c]
None or mild	94 (29.3)	43 (24.7)	51 (34.7)	
Moderate	146 (45.5)	85 (48.9)	61 (41.5)	
Severe	81 (25.2)	46 (26.4)	35 (23.8)	
Presence of psoriasis arthritis, n (%)				.26[b]
Yes	166 (51.7)	95 (54.6)	71 (48.3)	
No	155 (48.3)	79 (45.4)	76 (51.7)	
Dermatological treatment by, n (%)				.36[b]
Yes	228 (71.0)	120 (69.0)	108 (73.5)	
No	93 (29.0)	54 (31.0)	39 (26.5)	
Psoriasis knowledge, n (%)				.08[a]
Less than good	55 (17.1)	24 (13.8)	31 (21.1)	
Good or very good	266 (82.9)	150 (86.2)	116 (78.9)	
Type of recruitment, n (%)				<.001[b]
Digital media	225 (70.1)	147 (84.5)	78 (53.1)	
Clinical setting	96 (29.9)	27 (15.5)	69 (46.9)	

[a]Mann-Whitney test.
[b]Chi-square test.
[c]Kruskal-Wallis test.

XSL•FO
RenderX

Table 2. Results of the univariate and multivariate binary logistic regression with the frequency of internet use regarding psoriasis as a dependent variable and the several independent variables (n=321).

Variable	Univariate		Multivariate	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
Age	1.00 (0.98-1.01)	.86	1.01 (0.99-1.03)	.50
Duration of disease	0.99 (0.97-1.00)	.03	0.98 (0.96-1.00)	.008
Sex (reference: male)	2.13 (1.35-3.36)	.001	1.70 (1.03-2.81)	.04
Education (reference: No access to university)	0.78 (0.50-1.21)	.27	0.88 (0.54-1.44)	.62
Place of residence (reference: cities up to 100,000 inhabitants)	0.78 (0.50-1.24)	.30	0.74 (0.45-1.21)	.23
General psoriasis severity (reference: mild)				
Mild to moderate	1.97 (0.91-4.27)	.09	1.62 (0.71-3.67)	.25
Mild to severe	1.29 (0.59-2.78)	.52	0.96 (0.37-2.31)	.92
Mental burden of disease (reference: none or mild)				
None or mild to moderate	1.65 (0.98-2.79)	.06	1.80 (1.01-3.20)	.046
None or mild to severe	1.56 (0.86-2.84)	.15	1.96 (0.96-4.00)	.06
Presence of psoriasis arthritis (reference: no)	1.29 (0.83-2.00)	.26	1.11 (0.69-1.79)	.67
Dermatological treatment (reference: no)	0.80 (0.49-1.31)	.38	0.84 (0.50-1.44)	.53
Psoriasis knowledge (reference: less than good)	1.67 (0.93-3.00)	.09	1.97 (1.01-3.83)	.046

[a]OR: odds ratio.

Evaluation of Individual Digital Media Platforms and Their Impact on the Physician-Patient Relationship

In total, 188 participants who were considered internet users were included to evaluate digital media platforms and their impact on the physician-patient relationship. The mean age was 51.2 (SD 13.9, range 18.0-88.0) years. The most regularly used portals were psoriasis-related websites (106/188, 56.4%) and search engines (98/188, 52.1%; Figure 2).

The participants' main reasons (Table 3) for using information websites and search engines were to seek information about psoriasis (132/188, 70.2%) and therapy options (133/188, 70.7%). Among the participants using social media and Facebook, the main reasons were to read experience reports from other patients (social media: 34/188, 18.1%; Facebook: 55/188, 29.3%), to seek information about the disease (social media: 32/188, 17%; Facebook: 36/188, 19.1%), and to engage with other affected people (social media: 27/188, 14.4%; Facebook: 58/188, 30.9%).

Regarding age, only Facebook showed a significant difference with people using Facebook regularly being generally younger (regular use: mean 47.4, SD 11.8 years vs irregular use: mean 53.4, SD 14.6 years; *P*=.004). By the majority of the participants (140/188, 74.5%) psoriasis-related websites were rated as the best digital portal regarding information about psoriasis.

Overall, 72.3% (n=136) did web-based research on psoriasis treatment facilities with 43.1% (n=81) actually seeking consultation. In total, 75% (n=141) informed their physician about their web-based research and 77 (54.6%) received a neutral, 44 (31.2%) a positive, and 20 (14.2%) a negative reaction by their physician. Nearly 67.6% (n=127) wanted more involvement regarding health-related decisions. Approximately 44% (n=82) of the participants stated to search for additional information since the information provided by the physician was not sufficient. In addition, 66.5% (n=125) reported that their physicians have not recommended health-related websites, but 45.2% (n=85) wanted their physicians to do so. About 54.3% (n=102) stated that digital health-related information helped them to communicate with their physician on an equal footing (Figure 3).

XSL•FO
RenderX

Figure 2. Several digital media platforms stratified by the frequency of use (irregular use: less than once a month, regular use: at least once a month; n=188).

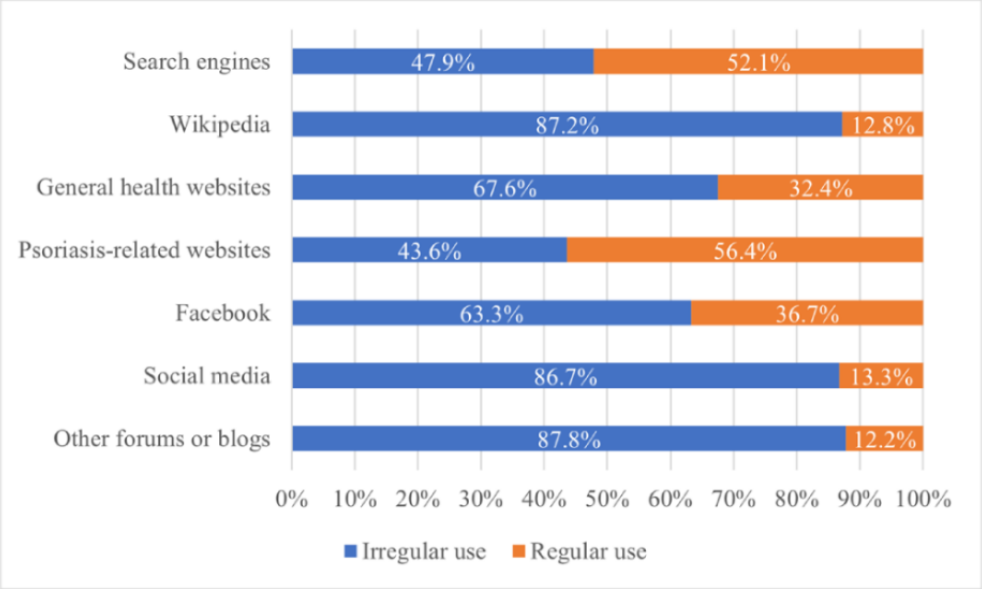

Table 3. Reasons for using digital media from the participants' perspective (n=188).

	Information websites and search engines	Social media	Facebook
I don't use it, n (%)	10 (5.3)	139 (73.9)	112 (59.6)
For information about my disease, n (%)	132 (70.2)	32 (17)	36 (19.1)
For information about therapy options, n (%)	133 (70.7)	31 (16.5)	39 (20.7)
For information about alternative medicine, n (%)	83 (44.1)	21 (11.2)	30 (16)
To be able to speak more at the next physician´s appointment, n (%)	45 (23.9)	12 (6.4)	10 (5.3)
To read/check what my physician told me, n (%)	61 (32.4)	10 (5.3)	10 (5.3)
For information about the causes of my disease, n (%)	97 (51.6)	—[a]	—
To be able to have an exchange with other affected individuals, n (%)	—	27 (14.4)	58 (30.9)
To read experience reports from other affected individuals, n (%)	—	34 (18.1)	55 (29.3)

[a]Not available.

Figure 3. Statements about the physician-patient relationship (n=188) with the percentages of participants' answers shown in the respective bars.

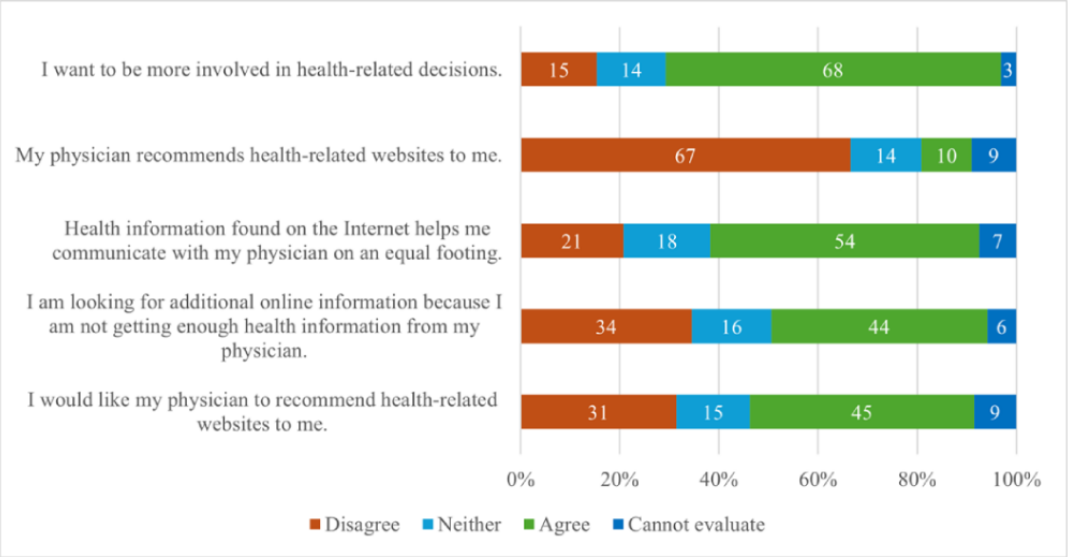

XSL•FO
RenderX

Discussion

Principal Results and Comparison With Prior Work

To gain a more differentiated understanding of the use of digital media and its influence on the physician-patient relationship to contribute to the appropriate care and support of those affected in the age of digital media, this study examined digital media use regarding psoriasis among people with and without dermatological treatment. Female sex, shorter disease duration, good self-assessed knowledge of psoriasis, and moderate mental burden were observed to be associated with regular psoriasis-related internet use. In addition, search engines and psoriasis-related websites were identified as the most frequently used platforms, especially for obtaining information about the disease and its therapy options. The fact that most physicians did not recommend health-related digital media to their patients despite their desire to receive such and being more involved in health-related decisions revealed shortcomings in the way physicians deal with digital media.

Our findings are in line with another study suggesting that women use the internet more for health-related information than men [25] although men generally used the internet more than women from 1997 until 2022 in Germany [26]. A prior study suggested that women perceived the internet as a more useful and entertaining medium for obtaining health information. Situational influences and a need to be well-informed contributed to women's higher involvement in health-related internet searches [25].

Although no age differences between irregular and regular internet users were found in this work, it was reported that the elderly generally use the internet less than younger people [26]. The elderly may lack experience and comfort in seeking health information resulting in a lack of trust in the internet. Consequently, older adults not using the internet for health-related reasons but those having the desire to do so might profit from receiving training on identifying trustworthy digital media platforms that consistently provide accurate health information [27]. Furthermore, it is possible that a longer disease duration and therefore dealing with continuous struggles for a long time might lead to frustration among the affected individuals [28]. To avoid this emotion, emotion-oriented coping could lead to an avoidant behavior [29] and, therefore, probably to an avoidance of digital information seeking as well. In addition, individuals with a shorter disease duration may want to acquire more health-related information as they know less about their disease whereas individuals with a longer disease duration and therefore more experience and knowledge may use digital media irregularly, for example, to search for new findings.

The factors influencing psoriasis-based internet use in another study (eg, younger age and higher education) [12] do not or only slightly align with the results of our study. This could be explained by the study's smaller sample size compared to the previous one (n=1520). Additionally, the previous study only included psoriasis-affected people from clinical settings.

One of the reasons for the limited use of social media platforms compared to, for example, psoriasis- or health-related websites, as identified in our study, may pertain to the quality of content present on certain platforms. For Facebook, it was priory revealed that the quality of psoriasis-related content was found to be inadequate, warranting improvements [19]. Another study found that approximately two-thirds of YouTube videos examined contained misleading or even dangerous content [30]. A further reason might be the age distribution in our study as a study from Germany found that the percentage of people using social media generally decreased as they got older [31]. In addition, some people may not want to discuss their disease on social media platforms because of the highly personal nature of the information. Concerns regarding the potential risks of sharing personal health data revolve around fears of discrimination by employers, insurance companies, and friends or family. These fears are intensified by stigmatized illnesses [32].

Search engines were one of the most consulted digital media when affected people sought health-related information [12,33]. Especially for these platforms a neutral controlling authority might be important to minimize false information. This could be achieved by ranking qualitative digital media platforms higher on search engines and thus making them more visible so that they are more likely to be accessed in health-related research on search engines. The quality of health-related digital media platforms could be ensured by including physicians in the development and supervision of these platforms [15]. Furthermore, there should be a regular exchange of opinions between physicians and patients to meet each other's needs. In this study, search engines and information websites are mainly used when affected individuals want to inform themselves about the disease and therapy options. In comparison, a popular reason for the use of social media was to read experience reports from other affected people. However, especially when reading subjective reports from other affected individuals, it is crucial to control the spread of false information [19,30,34]. This could be achieved by a review process supervised by medical experts before these reports are visible to everyone.

To address the increasing desire to be involved in health-related decisions, it is necessary to reevaluate the conventional physician-patient relationship. It is crucial to adopt a medical approach that emphasizes the relationship and incorporates the patient's view [35]. For patients seeking digital information, it was suggested to adopt a deliberative and participatory approach, which differs from the physician-centered and paternalistic models [36]. In this approach, physicians take on the role of educators and supporters, involving patients in a conversation to comprehend and choose from the various options presented during the decision-making process [37]. As identified in our study, the lack of digital media recommendations by physicians despite the patients' desire to receive such should be a reminder for the physicians to promote and guide the digital information seeking of their patients [33,38]. Physicians should ensure that shared decision-making is based on reliable and evidence-based digital information. Nevertheless, the proportion of positive responses from physicians when patients inform them about their digital media use should increase to empower and guide

XSL•FO
RenderX

patients who are interested in gaining more disease-related information. With this approach, patients can manage their disease more effectively and the physician-patient relationship may be improved by obtaining reliable digital health-related information [38-40]. However, it is important to acknowledge that physicians face a considerable burden with endorsing digital media, given the dynamic nature of content therein. Moreover, considering the breadth of diseases physicians treat, expecting them to be familiar with all available resources for each disease is demanding. This challenge could be mitigated by establishing a single high-quality, evidence-based platform that is supervised by medical professionals, thereby enabling physicians to recommend it.

Strengths and Limitations

A strength of this study was the mixed study sample, achieved through recruitment inside the clinical setting but also from digital media platforms, thereby encompassing individuals who have limited contact with health care professionals or infrequent physician appointments. Furthermore, the questionnaire was tailored based on recently conducted qualitative interviews, ensuring the inclusion of target group-specific and current aspects.

Recall bias is likely when asking questions about digital media use in the past. Moreover, a selection bias must be assumed. The recruitment from a dermatology-specialized clinic likely resulted in the inclusion of participants with higher overall disease severity. Meanwhile, individuals recruited via digital media were more inclined to engage with digital media regarding their condition. The average age of the study population was rather high (median 53, IQR 41-61 years and mean 51.2, SD 13.9 years for the respective research questions) compared to the average age in Germany which was 44.7 years in 2021 [41]. However, the average age in Germany also included people under the age of 18 years. According to a German study, the older generation does not use the internet in general compared to the younger ones [26] whereby that study considered individuals aged 14 years and over. In addition, different categorizations of the variables may lead to different results, especially concerning regular and irregular digital media use. In addition, the sample size is small due to the exclusion of questionnaires with missing data, resulting in a considerable loss of information. Therefore, there may have been a higher variance in the estimators and a loss of statistical power. Moreover, the process of missing data itself may not be random, but rather systematic, leading to potential non-response bias in the results and therefore limiting the generalizability of the results. The participants' self-reported diagnosis of psoriasis was not confirmed through clinical evaluation which could potentially affect the validity of the results.

Conclusions

The study underlines the significance of digital media in the health-related context, particularly with regard to women, individuals with shorter disease duration, and moderate mental disease severity. It also reveals the shortcomings within the physician-patient relationship that necessitate adaptation to the increasing digital media use. Physicians should guide their patients accordingly to ensure that they consume evidence-based information, thereby potentially creating an adequate framework for shared decision-making. Future studies should focus on analyzing and differentiating the individual digital media platforms in the context of psoriasis in more detail by assessing the quality of the information on the individual digital media platforms and additionally their potential for improving the physician-patient relationship. Additionally, to prevent the spread of false information on digital media, strategies against this prevalent problem should be developed in future studies. Furthermore, there should be a focus on the needs of psoriasis-affected people and physicians when discussing the future of health-related digital media. This will facilitate enhancements in, and, if necessary, expansion of the existing digital media offerings.

Acknowledgments

The authors are thankful to all participants for sharing their opinions and to Fabian Wallnöfer for helping with the creation of the questionnaire and recruitment of participants. The authors further thank Barbara Schuster and Linda Tizek for their valuable support during the development of the study questionnaire and the setup of the study. This study was financially supported by an unrestricted research grant from Novartis Pharma GmbH.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization and project administration were handled by AZ and SZ. Data curation was managed by MEE. Data analysis was performed by MEE, SZ, and HW. The original draft was written by MEE. Supervision, review, and editing were provided by AZ, SZ, HW, and TB.

Conflicts of Interest

MEE, SZ, and HW have no potential conflicts to declare. TB was president of the German Society of Dermatology. In addition, he has either received grants, consulting fees, support for attending meetings or travel, and speaker's honoraria or participated in a Data Safety Monitoring Board or Advisory Board of the following companies outside this work: AbbVie, Alk-Abello, Almirall,

XSL•FO
RenderX

Boehringer-Ingelheim, Celgene-BMS, GSK, Leo Pharma, Lilly, Novartis, Sanofi-Genzyme, Regeneron, and Viatris. AZ is a member of the German Society of Dermatology and leader of the Digital Dermatology group within this society. He has either been an advisor, received speaker's honoraria and grants, or participated in clinical trials of the following companies outside this work: AbbVie, Almirall, Amgen, Beiersdorf Dermo Medical, Bencard Allergie, BMS, Celgene, Eli Lilly, GSK, Incyte, Janssen Cilag, Leo Pharma, Novartis, Pfizer, Regeneron, Sanofi-Aventis, and UCB Pharma.

References

1. Michalek IM, Loring B, John SM. A systematic review of worldwide epidemiology of psoriasis. J Eur Acad Dermatol Venereol 2017;31(2):205-212. [doi: 10.1111/jdv.13854] [Medline: 27573025]
2. Yamanaka K, Yamamoto O, Honda T. Pathophysiology of psoriasis: a review. J Dermatol 2021;48(6):722-731. [doi: 10.1111/1346-8138.15913] [Medline: 33886133]
3. Griffiths CEM, Armstrong AW, Gudjonsson JE, Barker JNWN. Psoriasis. Lancet 2021;397(10281):1301-1315. [doi: 10.1016/S0140-6736(20)32549-6] [Medline: 33812489]
4. van Beugen S, van Middendorp H, Ferwerda M, Smit J, Zeeuwen-Franssen M, Kroft E, et al. Predictors of perceived stigmatization in patients with psoriasis. Br J Dermatol 2017;176(3):687-694. [doi: 10.1111/bjd.14875] [Medline: 27436615]
5. Zink A, Herrmann M, Fischer T, Lauffer F, Garzorz-Stark N, Böhner A, et al. Addiction: an underestimated problem in psoriasis health care. J Eur Acad Dermatol Venereol 2017;31(8):1308-1315. [doi: 10.1111/jdv.14204] [Medline: 28281329]
6. Schuster B, Peifer C, Ziehfreund S, Tizek L, Biedermann T, Zink A, et al. Happiness and depression in psoriasis: a cross-sectional study in Germany. Qual Life Res 2022;31(6):1761-1773 [FREE Full text] [doi: 10.1007/s11136-021-02991-2] [Medline: 34532835]
7. Wolf JA, Moreau JF, Patton TJ, Winger DG, Ferris LK. Prevalence and impact of health-related internet and smartphone use among dermatology patients. Cutis 2015;95(6):323-328. [Medline: 26125208]
8. DeBord LC, Patel V, Braun TL, Dao H. Social media in dermatology: clinical relevance, academic value, and trends across platforms. J Dermatolog Treat 2019;30(5):511-518. [doi: 10.1080/09546634.2018.1530444] [Medline: 30265614]
9. Corazza M, Forconi R, Patrizi A, Veneziano L, Offidani A, Paolinelli M, et al. The use of the internet by outpatients in dermatology: a multicentric, observational and cross-sectional study on frequency, motivations and feedback. J Eur Acad Dermatol Venereol 2021;35(9):e578-e582. [doi: 10.1111/jdv.17284] [Medline: 33871062]
10. What is digital media? all you need to know about this booming industry. Maryville University. 2020. URL: https://online.maryville.edu/blog/what-is-digital-media/ [accessed 2024-05-13]
11. Wallnöfer F, Erbas ME, Tizek L, Schuster B, Wecker H, Biedermann T, et al. Leveraging web search data to explore public interest in psoriasis in Germany. JEADV Clin Pract 2022;1(3):254-263. [doi: 10.1002/jvc2.37]
12. Aslan Kayıran M, Karadağ AS, Oğuz Topal I, Adışen E, Kılıç S, Keskin N, et al. Habits of using social media and the internet in psoriasis patients. Dermatol Pract Concept 2022;12(3):e2022143 [FREE Full text] [doi: 10.5826/dpc.1203a143] [Medline: 36159107]
13. Tonsaker T, Bartlett G, Trpkov C. Health information on the internet: gold mine or minefield? Can Fam Physician 2014;60(5):407-408 [FREE Full text] [Medline: 24828994]
14. Forgie EME, Lai H, Cao B, Stroulia E, Greenshaw AJ, Goez H. Social media and the transformation of the physician-patient relationship: viewpoint. J Med Internet Res 2021;23(12):e25230 [FREE Full text] [doi: 10.2196/25230] [Medline: 34951596]
15. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare 2021;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
16. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
17. Orgaz-Molina J, Cotugno M, Girón-Prieto MS, Arrabal-Polo M, Ruiz-Carrascosa J, Buendía-Eisman A, et al. A study of internet searches for medical information in dermatology patients: the patient-physician relationship. Actas Dermosifiliogr 2015;106(6):493-499 [FREE Full text] [doi: 10.1016/j.ad.2015.01.011] [Medline: 25805250]
18. Gantenbein L, Navarini AA, Maul LV, Brandt O, Mueller SM. Internet and social media use in dermatology patients: search behavior and impact on patient-physician relationship. Dermatol Ther 2020;33(6):e14098. [doi: 10.1111/dth.14098] [Medline: 32725746]
19. Schuster B, Ziehfreund S, Biedermann T, Zink A. Psoriasis 2.0: facebook als quelle krankheitsbezogener informationen für patienten mit psoriasis [Psoriasis 2.0: facebook as a source of disease-related information for patients with psoriasis]. J Dtsch Dermatol Ges [J German Soc Dermatol] 2020;18(6):571-581. [doi: 10.1111/ddg.14070_g] [Medline: 32519484]
20. Scherr S, Toma CL, Schuster B. Depression as a predictor of Facebook surveillance and envy: longitudinal evidence from a cross-lagged panel study in Germany. J Media Psychol 2019;31(4):196-202. [doi: 10.1027/1864-1105/a000247]
21. Koch W, Frees B. ARD/ZDF-onlinestudie 2017: neun von zehn deutschen online [ARD/ZDF online study 2017: nine out of ten Germans online]. ARD/ZDF-Onlinestudie [ARD/ZDF online study]. URL: https://www.ard-zdf-onlinestudie.de/files/2017/Artikel/917_Koch_Frees.pdf [accessed 2024-02-09]
22. Kemp S. Digital in 2018: world's internet users pass the 4 billion mark. We Are Social. 2018. URL: https://wearesocial.com/uk/blog/2018/01/global-digital-report-2018/ [accessed 2024-02-09]

XSL•FO
RenderX

23. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 2013;36(1):27-46. [doi: 10.1111/j.1600-0587.2012.07348.x]
24. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. J Interdiscip Math 2010;13(3):253-267. [doi: 10.1080/09720502.2010.10700699]
25. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015;17(6):e156 [FREE Full text] [doi: 10.2196/jmir.4127] [Medline: 26099325]
26. Entwicklung der onlinenutzung in Deutschland 1997 bis 2022 [Development of online usage in Germany from 1997 to 2022]. ARD/ZDF-Onlinestudie [ARD/ZDF online study]. URL: https://www.ard-zdf-onlinestudie.de/tabellen-onlinenutzung/entwicklung-der-onlinenutzung/ [accessed 2024-02-09]
27. Miller LMS, Bell RA. Online health information seeking: the influence of age, information trustworthiness, and search challenges. J Aging Health 2012;24(3):525-541. [doi: 10.1177/0898264311428167] [Medline: 22187092]
28. Masoudi Alavi N. Frustration: a challenge in chronic conditions. Nurs Midwifery Stud 2013;2(3):1-2 [FREE Full text] [doi: 10.5812/nms.14166] [Medline: 25414868]
29. Kristofferzon M, Engström M, Nilsson A. Coping mediates the relationship between sense of coherence and mental quality of life in patients with chronic illness: a cross-sectional study. Qual Life Res 2018;27(7):1855-1863 [FREE Full text] [doi: 10.1007/s11136-018-1845-0] [Medline: 29623597]
30. Mueller SM, Jungo P, Cajacob L, Schwegler S, Itin P, Brandt O. The absence of evidence is evidence of non-sense: cross-sectional study on the quality of psoriasis-related videos on youtube and their reception by health seekers. J Med Internet Res 2019;21(1):e11935 [FREE Full text] [doi: 10.2196/11935] [Medline: 30664460]
31. Social-media-nutzung 2019 bis 2022 [Social media usage 2019 to 2022]. ARD/ZDF-Onlinestudie [ARD/ZDF online study]. URL: https://www.ard-zdf-onlinestudie.de/tabellen-onlinenutzung/social-media-und-messenger/social-media/ [accessed 2024-02-09]
32. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010;12(2):e19. [doi: 10.2196/jmir.1549]
33. Wecker H, Höllerl L, Schick TS, Biedermann T, Zink A, Ziehfreund S. Patient journey and disease-related digital media usage: a cross-sectional study among dermatology patients across Germany. J Dtsch Dermatol Ges 2024;22(3):377-387. [doi: 10.1111/ddg.15311] [Medline: 38308182]
34. Gorrepati PL, Smith GP. Evaluating social media as a source of patient information regarding psoriasis treatments using the DISCERN instrument. J Dermatolog Treat 2022;33(5):2685-2686. [doi: 10.1080/09546634.2022.2037497] [Medline: 35107049]
35. Roter D. The enduring and evolving nature of the patient-physician relationship. Patient Educ Couns 2000;39(1):5-15 [FREE Full text] [doi: 10.1016/s0738-3991(99)00086-5] [Medline: 11013543]
36. Gerber BS, Eiser AR. The patient physician relationship in the internet age: future prospects and the research agenda. J Med Internet Res 2001;3(2):E15 [FREE Full text] [doi: 10.2196/jmir.3.2.e15] [Medline: 11720957]
37. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. JAMA 1992;267(16):2221-2226. [Medline: 1556799]
38. Schick TS, Höllerl L, Biedermann T, Zink A, Ziehfreund S. Impact of digital media on the patient journey and patient-physician relationship among dermatologists and adult patients with skin diseases: qualitative interview study. J Med Internet Res 2023;25:e44129 [FREE Full text] [doi: 10.2196/44129] [Medline: 37738078]
39. Stevenson FA, Kerr C, Murray E, Nazareth I. Information from the internet and the doctor-patient relationship: the patient perspective—a qualitative study. BMC Fam Pract 2007;8:47 [FREE Full text] [doi: 10.1186/1471-2296-8-47] [Medline: 17705836]
40. AlGhamdi KM, Moussa NA. Internet use by the public to search for health-related information. Int J Med Inform 2012;81(6):363-373. [doi: 10.1016/j.ijmedinf.2011.12.004] [Medline: 22217800]
41. Durchschnittsalter der bevölkerung in Deutschland (1871-2021) [Average age of the population in Germany (1871-2021)]. Bundesinstitut für Bevölkerungsforschung [Federal Institute for Population Research]. URL: https://www.bib.bund.de/Permalink.html?cms_permaid=1217910 [accessed 2024-02-14]

Abbreviations

OR: odds ratio
REDCap: Research Electronic Data Capture
VIF: variance inflation factor

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.02.24; peer-reviewed by P Reinders, S Gulliver; comments to author 13.04.24; revised version received 21.05.24; accepted 17.06.24; published 07.08.24.

Please cite as:
Erbas ME, Ziehfreund S, Wecker H, Biedermann T, Zink A
Digital Media Usage Behavior and Its Impact on the Physician-Patient Relationship: Cross-Sectional Study Among Individuals Affected by Psoriasis in Germany
J Med Internet Res 2024;26:e57823
URL: https://www.jmir.org/2024/1/e57823
doi:10.2196/57823
PMID:

©Mert Ege Erbas, Stefanie Ziehfreund, Hannah Wecker, Tilo Biedermann, Alexander Zink. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Long-Term Glycemic Control Improvement After the Home and Self-Care Program for Patients With Type 1 Diabetes: Real-World–Based Cohort Study

Dae-Jeong Koo[1*], MD; Sun-Joon Moon[2*], MD; Suhyeon Moon[3], MS; Se Eun Park[2], MD, PhD; Eun-Jung Rhee[2], MD, PhD; Won-Young Lee[2], MD, PhD; Cheol-Young Park[2], MD, PhD

[1]Division of Endocrinology and Metabolism, Department of Internal Medicine, Changwon Fatima Hospital, Changwon, Republic of Korea

[2]Division of Endocrinology and Metabolism, Department of Internal Medicine, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3]Division of Biostatistics, Department of Academic Research, Kangbuk Samsung Hospital, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Cheol-Young Park, MD, PhD
Division of Endocrinology and Metabolism
Department of Internal Medicine, Kangbuk Samsung Hospital
Sungkyunkwan University School of Medicine
29 Saemunan-ro, Jongno-gu
Seoul, 03181
Republic of Korea
Phone: 82 02 2001 1869
Fax: 82 02 2001 1588
Email: cydoctor@skku.edu

Abstract

Background: The prevalence of type 1 diabetes (T1D) is increasing worldwide, with a much higher proportion of adult patients. However, achieving stable glycemic control is difficult in these patients.

Objective: After periodic implementation of structured education for patients with T1D through the Home and Self-Care Program, a pilot home health care project promoted by the Korean government, we evaluated the program's effects on glycemic control.

Methods: This study was conducted from April 2020 to March 2023. We analyzed 119 participants with T1D aged >15 years. Nursing and nutrition education were provided separately up to 4 times per year, with physician consultation up to 6 times per year. A distinguishing feature of this study compared with previous ones was the provision of remote support using a general-purpose smartphone communication app offered up to 12 times annually on an as-needed basis to enhance the continuity of in-person education effects. Patients were followed up on at average intervals of 3 months for up to 24 months. The primary end point was the mean difference in glycated hemoglobin (HbA_{1c}) at each follow-up visit from baseline. For continuous glucose monitoring (CGM) users, CGM metrics were also evaluated.

Results: The mean HbA_{1c} level of study participants was 8.6% at baseline (mean duration of T1D 10.02, SD 16.10 y). The HbA_{1c} level reduction in participants who received at least 1 structured educational session went from 1.63% (SD 2.03%; P<.001; adjustment model=1.69%, 95% CI 1.24%-2.13% at the first follow-up visit) to 1.23% (SD 1.31%; P=.01; adjustment model=1.28%, 95% CI 0.78%-1.79% at the eighth follow-up visit). In the adjustment model, the actual mean HbA_{1c} values were maintained between a minimum of 7.33% (95% CI 7.20%-7.46% at the first follow-up visit) and a maximum of 7.62% (95% CI 7.41%-7.82% at the sixth follow-up visit). Among CGM users, after at least 1 session, the mean time in the target range was maintained between 61.59% (adjusted model, 95% CI 58.14%-65.03% at the second follow-up visit) and 54.7% (95% CI 50.92%-58.48% at the eighth follow-up visit), consistently staying above 54.7% (corresponding to an HbA_{1c} level of <7.6%). The mean time below the target range (TBR) also gradually improved to the recommended range (≤4% for TBR of <70 mg/dL and ≤1% for TBR of <54 mg/dL).

Conclusions: The Home and Self-Care Program protocol for glycemic control in patients with T1D is effective, producing significant improvement immediately and long-term maintenance effects, including on CGM indexes.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e60023) doi:10.2196/60023

KEYWORDS

type 1 diabetes; structured education; home health care; glycated hemoglobin; continuous glucose monitoring; mobile phone

Introduction

Background

As of 2021, approximately 8.4 million people worldwide are living with type 1 diabetes (T1D). Among them, only approximately 18% are aged <20 years [1]. Because the incidence of T1D is increasing worldwide, there is a need for comprehensive, tailored interventions that consider various needs for disease management, along with policy support from governments and the establishment of clinical evidence to support these interventions [2].

Research from the T1D Exchange Registry shows that only 17% of children and 21% of adults currently succeed in reducing their glucose to target levels [3]. Despite such a low glycemic control rate, level-2 hypoglycemia (<54 mg/dL) occurs quite frequently, ranging from once every 6 days to once every 2 days, in patients receiving multiple daily insulin injection treatment [4,5]. Due to the complexity and difficulty of managing T1D, it is well established that proper structured education is very important for controlling glucose levels and improving quality of life (QOL) [6-9]. A variety of programs, including education and support programs, technological interventions, psychosocial interventions, and transitional care programs, have been developed and applied to improve glycemic control or QOL in young adult patients with T1D [10]. Those interventions can be effective, but their effectiveness can vary depending on the specific design of the intervention and how it is implemented [10]. Many studies show promise, but further research is needed to identify the most effective strategies. The need for longitudinal studies to assess the long-term effects of these interventions is particularly great, along with studies based on personalized and integrated interventions that use a multifaceted approach [10]. However, clinic hours and platforms for structured education, which are needed to provide individualized and comprehensive treatment for patients with T1D, are lacking.

Objectives

Therefore, the Korean government promoted a pilot home health care project for patients with T1D in January 2020. The purpose of this project was to develop a program that continuously manages patients and provides feedback even when they are at home, thereby reducing unnecessary medical costs due to worsening disease and simultaneously improving the QOL of patients. Specifically, the goal was to develop periodic medical interventions that enable home health care services, including education and counseling for self-management and regular monitoring of conditions. Our team at Kangbuk Samsung Hospital participated in this project and developed a comprehensive and structured education program to meet those needs (tentative name *Home and Self-Care Program* [HELP]) that includes remote support using an existing universal smartphone app. For this paper, we evaluated the short- and long-term effects of HELP on glycemic control in patients with T1D from adolescence to middle age.

Methods

Study Setting and Context

Kangbuk Samsung Hospital is a prominent health care institution located in Seoul, South Korea. It is affiliated with Sungkyunkwan University and serves as a tertiary teaching hospital. The hospital is equipped with 700 beds and caters to a wide range of medical needs, providing advanced health care services to its patients. The hospital mainly serves patients referred from primary or secondary hospitals, ensuring that they receive specialized care. Before the implementation of HELP, the team included 5 diabetologists among 7 endocrinologists, 2 education nurses, and 1 nutritionist. During the HELP period, the team was expanded to include 3 education nurses and 2 nutritionists.

Kangbuk Samsung Hospital had limited structured education programs for patients with T1D before the implementation of HELP. Several factors contributed to this limitation. The standard practice in South Korea of having approximately 5-minute clinic visits per patient due to low fees for service and the need to see many patients resulted in insufficient time to provide comprehensive education and support during appointments. In addition, there were scarce structured education programs available for patients with T1D, which hindered the ability to deliver consistent and thorough education on diabetes management. The lack of a fee-for-service model or specific education fees for T1D care made it challenging to allocate resources and time for comprehensive patient education. In the clinic, insulin dosage adjustments were provided by physicians primarily based on laboratory test results and daily log data from patients. Specific education for T1D was scarce at the time.

Despite these challenges, our hospital's team of endocrinologists, education nurses, and nutritionists strived to manage patients with T1D to the best of their abilities within these constraints. The introduction of the HELP program aimed to address these gaps and provide a more structured and effective approach to education and management for patients with T1D.

Study Design

This study is a real-world–based cohort study conducted from April 2020 to March 2023 in South Korea. The last data collection for this study occurred until March 3, 2023. The definition of T1D applied in this study was established according to the guidelines of the Health Insurance Review and Assessment Service under the Ministry of Health and Welfare of South Korea [11]. T1D was defined as 1 or more of the following: basal C-peptide of ≤0.6 ng/mL, stimulated C-peptide of ≤1.8 ng/mL, or 24-hour urine C-peptide of <3 µg per 24 hours; history of diabetic ketoacidosis at initial diagnosis;

XSL•FO
RenderX

positive autoantibodies to pancreatic islets or insulin, such as the antiglutamic acid decarboxylase antibody; insulin administration required for proper glucose control; and disease code E10.x. During regular follow-up for each patient, the attending physician explained the purpose of HELP, and when the patient's understanding and verbal consent were ascertained at the clinic, an experienced nurse or attending physician directly explained the consent form in additional detail. All participants were included if they met the age limit and were able to understand the program and receive educational sessions. Eligible participants were those who completed the required number of educational session or sessions, visited the outpatient clinic on the scheduled date, and provided complete data for statistical analysis (Figure 1).

Figure 1. Diagram for study participants. The Home and Self-Care Program (HELP) means receiving at least 1 in-person structured education session (with remote digital support), which includes nursing and nutrition education as well as a physician consultation. f/u: follow-up.

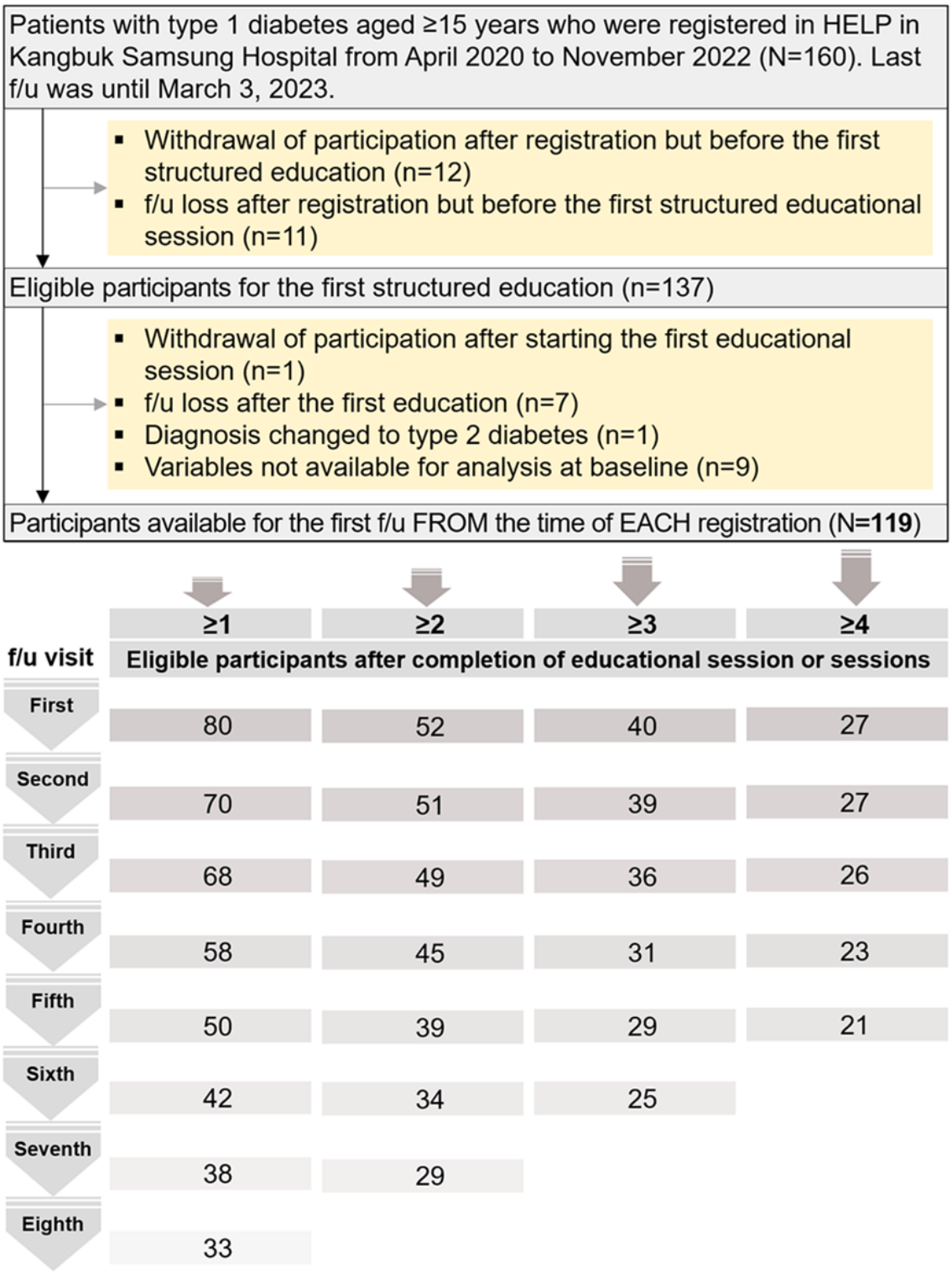

XSL•FO
RenderX

Structured Education

The definition of receiving structured education (HELP) was that the participant received physician consultation, nursing education, and nutrition education at least once each. Participants consulted with a physician at every follow-up visit from the beginning up to 6 times per year. Nursing and nutrition education was provided separately up to 4 times per year at or between visits by experienced board-certified nurses and nutritionists. Each educational area included basic interventions that were applicable across all areas, along with specialized, individualized, and advanced intervention components specific to each area. Remote support encompassed all educational areas (Figure 2 and Multimedia Appendices 1 and 2).

Figure 2. Schematic diagram for implementation of structured education sessions in the Home and Self-Care Program (HELP). HELP means receiving at least 1 structured education session (with remote digital support), which includes nursing and nutrition education as well as a physician consultation. AGP: ambulatory glucose profile; CGM: continuous glucose monitoring; f/u: follow-up; SES: socioeconomic status; T1D: type 1 diabetes.

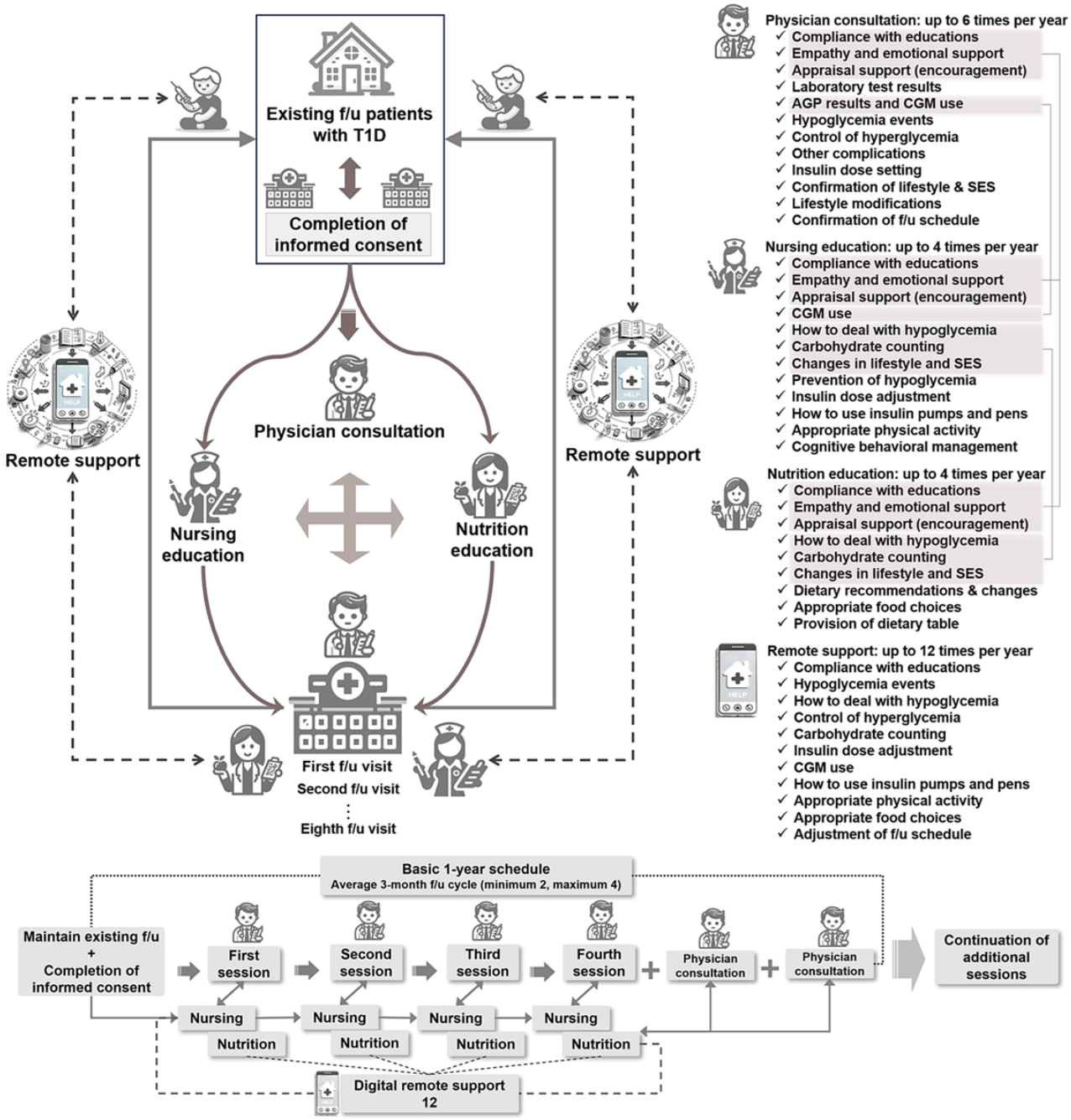

The structured education described how to prevent and respond to hypoglycemia, control hyperglycemia, determine proper insulin dose control with carbohydrate counting, and optimally use and understand continuous glucose monitoring (CGM) for CGM users only. To induce overall lifestyle modification, dietary recommendations (including appropriate food choices and provision of a dietary table), exercise therapy, and cognitive behavioral management were also discussed. We also provided emotional support with empathy and appraisal support with encouragement. These educational contents were created based on the guidelines of the Korean Diabetes Association [12] and the Korean Society for the Study of Obesity [13] and were

XSL•FO
RenderX

implemented in parallel with Kangbuk Samsung Hospital's own nursing and food nutrition education programs already in progress. The individualized educational sessions were tailored to the characteristics and conditions of each patient.

In addition, this program is characterized by actively using remote support through smartphones to strengthen and sustain the educational effect after in-person education, which had not been attempted in previous structured education programs. All participants took part in this study using their own personal smartphones, and no separate equipment was provided. When the patients were outside the hospital, remote support with feedback on an existing general-purpose smartphone app (KakaoTalk Channel; Kakao Corp) was provided up to 12 times a year between education sessions. Through the channel established by Kangbuk Samsung Hospital, patients communicated based on SMS text messages, and when necessary, photos provided by patients were used as an element of communication. In most cases, patients initiated the request for help, and the medical staff responded with real-time or delayed answers depending on the situation. Medical staff may initiate communication first to confirm that the feedback has been correctly implemented by the patient. In addition, communication can also take place through video or phone calls. Remote support was not just for cases in which participants did not receive in-person education. This support was provided to ensure that participants were effectively applying what they had learned from the education, even if they had completed more than 1 session. In addition, in situations in which in-person education was delayed (eg, due to COVID-19 infection), remote support was provided to ensure continuity of education until the next education session. While in-person education is the basis of HELP, this constitutes additional support to enhance the effectiveness of the program.

Primary Outcome

Participants were followed up on at average intervals of 3 months for up to 24 months. The timing of participation and the start of education varied for each patient, resulting in a different number of education sessions completed by participants during each follow-up period. We aimed to evaluate glycemic control status by follow-up visit according to the number of education sessions completed (Figure 1). We would like to emphasize that our study was based on a real-world setting. In general, most participants were observed based on the 3-month follow-up intervals, but some patients visited earlier than these intervals, whereas others visited at delayed intervals (Figure 2). Thus, we did not want to look at changes in glycated hemoglobin (HbA_{1c}) levels over a specific period but rather wanted to check the degree of glycemic control (HbA_{1c}) at the time of the follow-up visit after the education session or sessions. The key was to see whether glycemic control had improved at the next visit after receiving education session or sessions. The primary end point was the difference in mean HbA_{1c} values at each follow-up visit.

Secondary Outcome

Before the start of the study, only a small number of participants (6/119, 5%) were using CGM consistently and properly, and we recommended CGM use after the start of the study, resulting in 83.2% (99/119) of the participants using CGM throughout the study period and being used for analysis. For study participants who were using CGM during the study period, we included CGM metrics as additional end points in our analysis using cloud-based ambulatory glucose profile (AGP), which was a time in the target range (TIR) of 70 to 180 mg/dL, time below the target range (TBR; <70 mg/dL and <54 mg/dL), and coefficient of variation (CV). It is emphasized that appropriate education in CGM use was added only for existing CGM users. CGM users were encouraged to actively use CGM (>70% of time CGM active) during the follow-up period rather than for a specific period. However, the percentage of time CGM is active varies significantly among CGM users. Most CGM users used the FreeStyle Libre 1 system, but some also used the Dexcom G5 or G6 and Guardian Connect systems.

Data Collection

Information about each patient's medical history and socioeconomic status (SES) was obtained from official medical charts and forms filled out by the patients. Blood pressure was measured at least twice using a standard sphygmomanometer in a sitting position with sufficient rest. Data were collected through a combination of medical chart reviews and patient-completed forms. Trained research assistants abstracted the data from the medical charts, ensuring consistency and accuracy in the process. Patients completed the forms independently, providing self-reported information on their health status and treatment experiences. The abstracted data were entered into a structured database designed for the study. To ensure data quality, we implemented several data checking and cleaning steps. This included verifying data entries against the original medical charts and patient forms, performing range and consistency checks, and resolving any discrepancies through a secondary review by another researcher. All biochemical tests were performed in laboratories using standardized equipment, and anthropometric measurements were taken by well-trained staff. Patients' medical data were obtained at each visit.

CGM-related data were also obtained through cloud-based AGP at each follow-up visit after the first education session. As mentioned previously, HELP is defined as receiving in-person education, so patients who were supported only remotely due to delays in the visit cycle and unable to receive in-person education were excluded from the analysis of that follow-up cycle. CGM indexes (TIR, TBR, and CV) were confirmed from cloud-based AGP data. In some cases, AGP data could not be obtained intermittently during follow-up. In that case, they were excluded from the analysis related to that follow-up period. Blood glucose levels were assessed through the hexokinase method (Modular D2400; Hitachi). HbA_{1c} levels were measured using the immunoturbidimetric assay method (Cobra Integra 800 automatic analyzer; Roche Diagnostics). The urinary albumin-to-creatinine ratio was determined by measuring the spot urine albumin base in a radioimmunological competition assay (Immunotech) and the spot urine creatinine concentration using a modified Jaffe method.

XSL•FO
RenderX

Statistical Analysis

The analysis was performed according to the number of structured educational sessions provided to participants. For continuous variables, appropriate representative values (mean or median) and spreads (SDs) are presented, and for categorical variables, frequencies and ratios are presented. A 2-tailed *t* test was conducted to compare continuous variables to assess differences in both directions, and the chi-square test was used to compare categorical variables. Paired *t* tests were used to compare HbA_{1c} levels and other values between baseline and each follow-up time point, with *P* values adjusted using Bonferroni correction for multiple testing. Regression analyses were performed to test the association between the number of education sessions and changes in HbA_{1c} levels and CGM metrics. The last observation carried forward imputation method was used for missing variables. Confounders were adjusted for, and the results are presented with 95% CIs for the primary end point—model 1 was adjusted for age, sex, baseline BMI, and baseline HbA_{1c}, and model 2 was additionally adjusted for treatment type and diabetes mellitus duration. The supplemental analysis included variables such as educational level, social status, and employment status (Multimedia Appendix 3). CGM-related outcomes were adjusted for age, sex, BMI, HbA_{1c} level, treatment type, and duration of diabetes mellitus. All 2-tailed values with $P<.05$ were considered statistically significant. All statistical analyses were performed using R (version 4.3.2; R Foundation for Statistical Computing).

Ethical Considerations

This study was reviewed and approved by the institutional review board of Kangbuk Samsung Hospital (KBSMC 202004028), Seoul, South Korea. Informed consent was obtained from all participants before their inclusion in the study for the primary data collection, including the provision that the collected data may be used for secondary analyses without obtaining additional consent from the participants. Although the HELP program was conducted prospectively, with all participants providing informed consent at the outset, research consent was not fully obtained from some participants. For these cases, we retrospectively analyzed the data with approval from the institutional review board. All procedures performed in this study were in accordance with the ethical standards of the institution. All study data were collected and stored in an anonymous format, ensuring that no personally identifiable information was included in the data set. No financial compensation was offered to participants. The study was conducted on a voluntary basis, and participants were informed of this before consenting to take part. Participants were informed of the significant impact that their participation would have on advancing our understanding of T1D.

Results

Study Participants

Participants in this study were selected from patients with T1D (aged ≥15 years). A total of 160 patients participated sequentially after completing the consent form. Patients who withdrew their participation (12/160, 7.5%) or were lost to follow-up (11/160, 6.9%) before the first structured education session were excluded. After the program started, a participant who withdrew their participation after the first physician consultation and before completing 1 education session (1/160, 0.6%), participants with complete loss to follow-up (7/160, 4.4%), a participant whose diagnosis changed to type 2 diabetes (1/160, 0.6%), and participants missing baseline variables (9/160, 5.6%) were further excluded from the analysis. Therefore, ultimately, 119 patients were available for study analysis at baseline (Figure 1).

Characteristics of Participants

The median HbA_{1c} level of the study participants was 8.3% (IQR 7.3%-9.15%; mean 8.62%, SD 2.27%) at baseline, and the median duration of diabetes was 7.00 (IQR 2.00-13.00; mean 10.02, SD 16.10) years. Most of the study participants (110/119, 92.4%) controlled their glucose using multiple daily insulin injections (Table 1; Table S1 in Multimedia Appendix 3). After enrollment, most participants (99/119, 83.2%) checked their glucose using CGM. Although there was a slight difference, the BMI was statistically significantly lower in the group with relatively poor glycemic control (HbA_{1c}≥8%). Among those who provided their SES information (63/119, 52.9%), participants with relatively better glycemic control (HbA_{1c}<8%) had a relatively higher educational level (proportion of people with a bachelor's degree or higher: 20/29, 69% with HbA_{1c}<8% vs 10/34, 29% with HbA_{1c}≥8%; $P=.007$; Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics of participants by glycated hemoglobin (HbA_c) level (N=119)[a].

Characteristic	Total participants	HbA_{1c} level		*P* value
		<8% (n=50)	≥8% (n=69)	
Male participants, n (%)	61 (51.3)	20 (40)	41 (59.4)	.06
Age (y), mean (SD)	45.84 (16.96)	43.98 (16.74)	47.19 (17.11)	.31
Age (y), n (%)				.64
<35	35 (29.4)	17 (34)	18 (26.1)	
35-60	55 (46.2)	22 (44)	33 (47.8)	
≥60	29 (24.4)	11 (22)	18 (26.1)	
Duration of diabetes (y)				
Values, mean (SD)	10.02 (16.10)	12.08 (23.36)	8.52 (7.13)	.24
Values, median (IQR)	7.00 (2.00-13.00)	7.50 (2.00-13.00)	7.00 (2.00-12.00)	.94
Subcategories, n (%)				.80
<3.0	33 (27.7)	14 (28)	19 (27.5)	
3.0-10.0	37 (31.1)	17 (34)	20 (29)	
≥10.0	49 (41.2)	19 (38)	30 (43.5)	
HbA_{1c} level (%)				
Values, mean (SD)	8.62 (2.27)	6.95 (0.77)	9.84 (2.23)	<.001
Values, median (IQR)	8.30 (7.30-9.15)	7.20 (6.40-7.57)	9.00 (8.50-10.40)	<.001
FBG[b] (mg/dL), mean (SD)	146.97 (73.58)	132.02 (53.11)	158.11 (84.49)	.09
PP2[c] (mg/dL), mean (SD)	189.61 (90.86)	171.93 (93.77)	202.63 (88.91)	.35
Body weight (kg), mean (SD)	61.86 (12.95)	64.27 (16.86)	60.15 (8.99)	.09
BMI (kg/m^2), mean (SD)	22.88 (3.60)	23.74 (4.78)	22.27 (2.31)	.03
SBP[d] (mm Hg), mean (SD)	119.98 (14.62)	122.27 (15.55)	118.39 (13.82)	.16
DBP[e] (mm Hg), mean (SD)	73.57 (10.67)	75.46 (9.77)	72.26 (11.14)	.11
Creatinine (mg/dL), mean (SD)	0.78 (0.36)	0.77 (0.40)	0.79 (0.33)	.74
eGFR[f] (mL/min/1.73 m^2), mean (SD)	99.69 (26.53)	98.13 (27.92)	100.81 (25.69)	.63
uACR[g] (mg/g), mean (SD)	80.70 (184.74)	84.44 (241.40)	78.25 (139.61)	.90
Type of insulin treatment, n (%)				.88
MDI[h]	110 (92.4)	46 (92)	64 (92.8)	
Premixed insulin	3 (2.5)	1 (2)	2 (2.9)	
Pump	6 (5)	3 (6)	3 (4.3)	
TDD[i] of insulin (units), mean (SD)	45.90 (25.93)	46.80 (30.93)	45.25 (21.83)	.75
Socioeconomic status				
Educational level, n (%)				.007
Primary school	3 (4.8)[j]	1 (3.4)[k]	2 (5.9)[l]	
Middle or high school	30 (47.6)[j]	8 (27.6)[k]	22 (64.7)[l]	
Bachelor's degree or higher	30 (47.6)[j]	20 (69)[k]	10 (29.4)[l]	
Employment status, n (%)				.10
Active in work	28 (44.4)[j]	15 (51.7)[k]	13 (38.2)[l]	
Retired or unemployed	9 (14.3)[j]	6 (20.7)[k]	3 (8.8)[l]	

XSL•FO
RenderX

Characteristic	Total participants	HbA_{1c} level		*P* value
		<8% (n=50)	≥8% (n=69)	
Other[m]	26 (41.3)[j]	8 (27.6)[k]	18 (52.9)[l]	
Social status, n (%)				.58
Married	27 (42.9)[j]	14 (48.3)[k]	13 (38.2)[l]	
Other[n]	36 (57.1)[j]	15 (51.7)[k]	21 (61.8)[l]	
Caregiver, n (%)				.47
Self	18 (28.6)[j]	7 (24.1)[k]	11 (32.4)[l]	
Domestic partner[o]	44 (69.8)[j]	22 (75.9)[k]	22 (64.7)[l]	
Social support system	1 (1.6)[j]	0 (0)[k]	1 (2.9)[l]	
Income, n (%)				.41
Relatively low[p]	34 (54)[j]	15 (51.7)[k]	19 (55.9)[l]	
Relatively high[q]	13 (20.6)[j]	8 (27.6)[k]	5 (14.7)[l]	
No income or unknown	16 (25.4)[j]	6 (20.7)[k]	10 (29.4)[l]	

[a]Only 63 participants out of the total study population consented to provide their socioeconomic status information.
[b]FBG: fasting blood glucose.
[c]PP2: 2-hour postprandial glucose.
[d]SBP: systolic blood pressure.
[e]DBP: diastolic blood pressure.
[f]eGFR: estimated glomerular filtration rate.
[g]uACR: urine albumin-to-creatinine ratio.
[h]MDI: multiple daily injections.
[i]TDD: total daily dose.
[j]N=63.
[k]N=29.
[l]N=34.
[m]Housewives, students, and people with disabilities, among others.
[n]Unmarried, living alone, and widowed, among others.
[o]Spouse, parents, or descendants, among others.
[p]<5 million South Korean won (KRW) per month; assuming an average exchange rate of approximately 1200 KRW per US $ during the study period, the estimated amount is approximately US $4100 per month.
[q]≥5 million KRW per month.

Primary Outcome

The HbA_{1c} reduction in participants who received structured educational session or sessions went from 1.63% (SD 2.03%; *P*<.001; after model 2 adjustment: 1.69%, 95% CI 1.24%-2.13%) at the first follow-up visit to 1.23% (SD 1.31%; *P*=.01; after model 2 adjustment: 1.28%, 95% CI 0.78%-1.79%) at the eighth follow-up visit after ≥1 session (Table 2) and from 1.74% (SD 2.45%; *P*=.01; after model 2 adjustment: 1.75%, 95% CI 0.72%-2.78%) at the first follow-up visit to 1.18% (SD 1.3%; *P*=.009; after model 2 adjustment: 1.17%, 95% CI 0.55%-1.85%) at the fourth follow-up visit after ≥4 sessions (Table S2 in Multimedia Appendix 3).

XSL•FO
RenderX

Table 2. Effects of the Home and Self-Care Program (HELP)[a] on glycemic control in patients with type 1 diabetes (N=119).

Number of education sessions and follow-up visit[b]	Participants, n (%)	HbA_{1c}[c] level (%), mean (SD)		HbA_{1c} difference (%), mean (SD)	*P* value	Adjusted[d] HbA_{1c} difference (%; 95% CI)
		At baseline	At follow-up visit			
≥1						
First	80 (67.2)	8.54 (2.15)	7.33 (1.12)	1.63 (2.03)	<.001	1.69 (1.24-2.13)
Second	70 (58.8)	8.53 (2.05)	7.37 (1.18)	1.67 (1.94)	.001	1.74 (1.27-2.20)
Third	68 (57.1)	8.59 (2.10)	7.49 (1.38)	1.63 (2.01)	.002	1.69 (1.20-2.19)
Fourth	58 (48.7)	8.58 (2.06)	7.41 (1.21)	1.65 (2.07)	.004	1.70 (1.13-2.27)
Fifth	50 (42)	8.61 (2.11)	7.55 (1.09)	1.55 (1.98)	.02	1.66 (1.05-2.27)
Sixth	42 (35.3)	8.70 (2.12)	7.65 (1.12)	1.51 (1.88)	.03	1.54 (0.92-2.17)
Seventh	38 (31.9)	8.60 (1.99)	7.56 (1.17)	1.41 (1.66)	.02	1.44 (0.81-2.06)
Eighth	33 (27.7)	8.48 (1.35)	7.54 (1.18)	1.23 (1.31)	.01	1.28 (0.78-1.79)

[a]HELP means receiving at least 1 structured education session (with remote digital support) comprising nursing and nutrition education and a physician consultation.

[b]Follow-up visit after the corresponding number of structured education session or sessions.

[c]HbA_{1c}: glycated hemoglobin.

[d]Adjusted for age, sex, BMI, HbA_{1c} level, treatment type, and diabetes duration.

After full adjustment, the actual mean HbA_{1c} values remained between 7.33% (95% CI 7.2%-7.46%, at the first follow-up visit) and 7.62% (95% CI 7.41%-7.82%, at the sixth follow-up visit) after completing at least 1 session, and similarly remained between 7.35% (95% CI 7.09%-7.6%, at the first follow-up visit) and 7.62% (95% CI 7.34%-7.9%, at the second follow-up visit) after completing ≥4 sessions (Figure 3).

Figure 3. Improvement in glycated hemoglobin (HbA_{1c}) levels after the Home and Self-Care Program (HELP) in patients with type 1 diabetes and its long-term maintenance. HELP means receiving at least one structured education session (with remote digital support), which includes nursing and nutrition education as well as a physician consultation. *Mean values at baseline. **Follow-up visit values adjusted for age, sex, BMI, HbA_{1c} level, treatment type, and diabetes duration (with 95% CIs). ***HbA_{1c} difference after adjustment. f/u: follow-up.

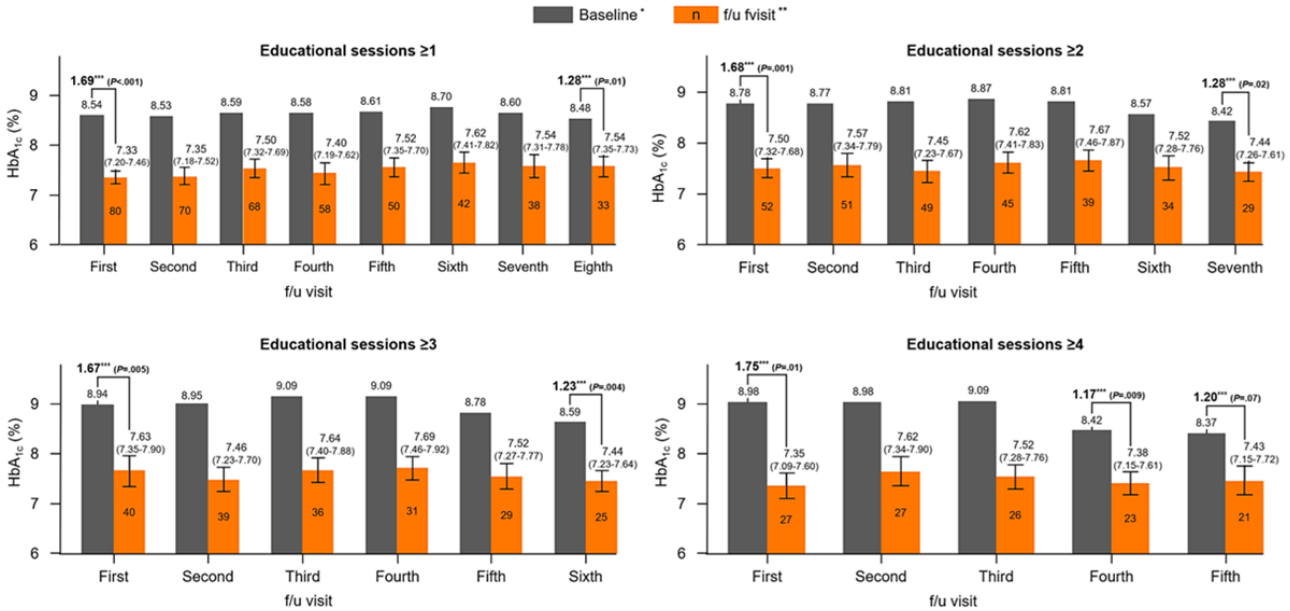

Even if no further education sessions were provided after the cessation of the program, the positive effect on glycemic control tended to be maintained for approximately 1 year, but the statistical significance was weak, and the number of participants in that analysis was small (Figure S1 and Table S3 in Multimedia Appendix 3).

In the analysis of participants who provided SES information, models adjusting for SES (ie, in the effort to exclude the effect of SES) showed a tendency for the degree of HbA_{1c} reduction to be more significant. This was particularly evident in patients who completed the educational sessions ≥2 times. However, when the number of sessions exceeded 4, the effect of SES

XSL•FO
RenderX

decreased again (Table S4 in Multimedia Appendix 3). These effects were particularly pronounced in CGM users, especially in active users with a CGM activation time of 70% or more (Tables S5-S7 in Multimedia Appendix 3).

Secondary Outcome

CGM metrics were officially recorded and analyzed after the first education session. The mean TIR values were maintained at >54.7% (corresponding to an HbA_{1c} level of approximately <7.6% [14])—from 61.77% (SD 16.79%; after adjustment: 61.59%, 95% CI 58.14%-65.03%) at the second follow-up visit to 55.85% (SD 20.61%; after adjustment: 54.7%, 95% CI 50.92%-58.48%) at the eighth follow-up visit after ≥1 session and from 59% (SD 13.72%; after adjustment: 59%, 95% CI 53.78%-64.22%) at the first follow-up visit to 55.86% (SD 14.91%; after adjustment: 55.92%, 95% CI 51.13%-60.72%) at the third follow-up visit after ≥4 sessions (Figure 4 and Table S8 and Figure S2 in Multimedia Appendix 3). The analysis of participants who provided SES information showed no significant differences according to SES (Table S9 in Multimedia Appendix 3).

Figure 4. Long-term effects of the Home and Self-Care Program (HELP) on continuous glucose monitoring (CGM) metrics in patients with type 1 diabetes using CGM after at least one educational session (n=99 at baseline). All values at the follow-up visits were adjusted for age, sex, BMI, glycated hemoglobin level, treatment type, and diabetes duration and expressed with 95% CIs. HELP means receiving at least one structured education session (with remote digital support), which includes nursing and nutrition education as well as a physician consultation. **P* value (in t test) for maintenance above the reference (time in the target range=55%) and below the reference (coefficient of variation [CV]=40%); TBR: time below the target range; TIR: time in the target range.

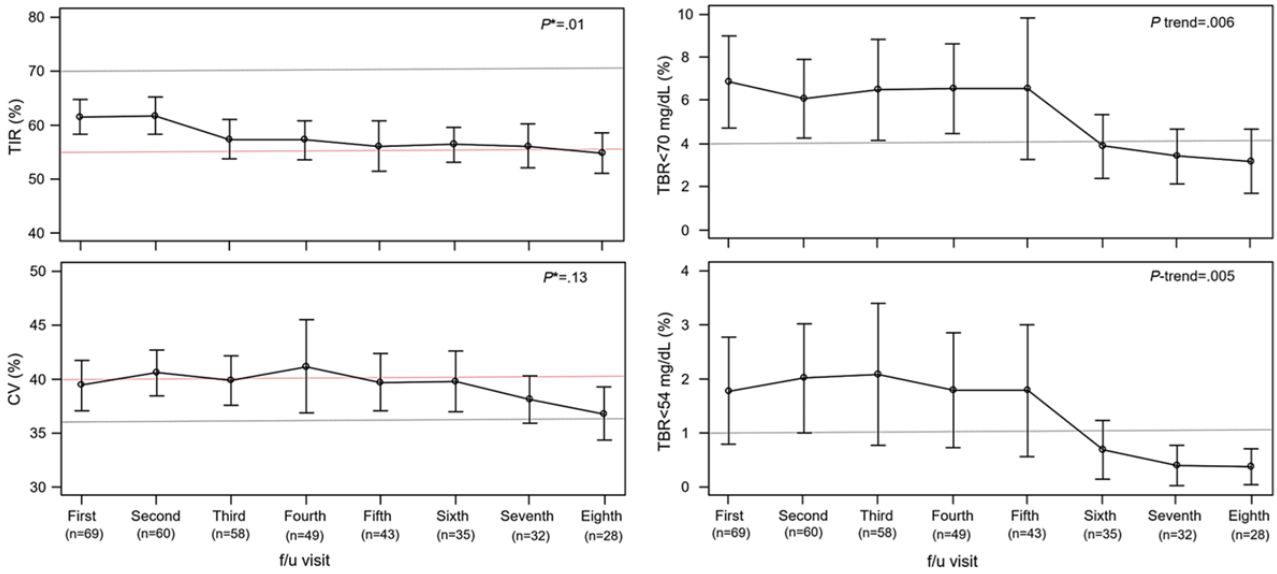

The mean TBR gradually converged to within the recommended ranges (≤4% for TBR of <70 mg/dL and ≤1% for TBR of <54 mg/dL) [14] as follow-up progressed (Figure 4 and Table S8 and Figure S2 in Multimedia Appendix 3)—from 6.82% (SD 8.11%; after adjustment: 6.85%, 95% CI 4.72%-8.98%) at the first visit to 3.43% (SD 4.37%; after adjustment: 3.16%, 95% CI 1.68%-4.64%) at the eighth visit for a TBR of <70 mg/dL and from 1.76% (SD 3.48%; after adjustment: 1.77%, 95% CI 0.78%-2.77%) at the first visit to 0.44% (SD 0.89%; after adjustment: 0.36%, 95% CI 0.44%-0.89%) at the eighth visit after at least 1 session for a TBR of <54 mg/dL.

The mean CV value was maintained at <41.12% (41% for participants with SES information) after adjustment throughout the follow-up period (Figure 4 and Tables S8 and S9 and Figure S2 in Multimedia Appendix 3). These effects were also pronounced in active CGM users (>70% of "% time CGM active"; Table S10 in Multimedia Appendix 3).

Discussion

Principal Findings

The aim of this study was to develop a structured education program for patients with T1D that was feasible in the real world and imposed no additional cost burden on patients. We tried to demonstrate that, by individualizing and expanding existing in-person education and support programs, their effects can be significantly amplified and prolonged with simple additional remote support. This study provides evidence that HELP is clinically effective in improving glycemic control in patients with T1D (see Multimedia Appendix 4 for a graphical summary). Mean HbA_{1c} values decreased by 1.63% (after adjustment: 1.69%) at the first follow-up visit after completing only 1 education session. This effect was higher among CGM users (1.69% [after adjustment: 1.78%] for all CGM users and 2.02% [after adjustment: 2.05%] for active CGM users with >70% of "% time CGM active"). As the follow-up period progressed, the degree of reduction tended to gradually decline, but an overall decrease in HbA_{1c} levels of at least 1.2% was maintained for the entire follow-up period (approximately 24 months). In addition, although statistical significance was weak,

XSL•FO
RenderX

the reduction in mean HbA_{1c} level tended to be sustained for approximately 12 months even with only simple follow-up after discontinuing the program. These results imply that, even after the educational program ends, patients keep using the knowledge and skills provided to them.

It is interesting to note that the reduction in HbA_{1c} levels did not continue to increase significantly when the number of structured educational sessions increased. When confounding factors such as SES, including educational background, were adjusted for and the effect of the structured education itself was clarified, the degree of HbA_{1c} level reduction increased slightly as the number of educational sessions increased. However, this pattern disappeared in patients who received >4 educational sessions. These findings require some careful consideration. Those who received fewer educational sessions had lower mean HbA_{1c} levels at baseline, suggesting that their glycemic control potential and educational compliance were likely to be relatively good. Therefore, we presume that even after only 1 educational session, the glucose-lowering effect can be maintained with simple follow-up without the need for additional sessions. In contrast, those who received more educational sessions had higher mean baseline HbA_{1c} levels, suggesting that they were likely having difficulty controlling their glucose. Those patients might need more educational sessions and longer follow-up periods to achieve effects similar to those in the former group.

Another important clinical outcome in this study was that hypoglycemia was not exacerbated despite the overall glycemic reduction; instead, the hypoglycemia indexes also improved gradually. The assessment of hypoglycemia events was based on TBR data from the AGPs due to the scarcity and irregularity of self-monitoring of blood glucose results. In a TBR of both <70 mg/dL and <54 mg/dL, the pattern of converging within the recommended level (4% and 1%, respectively) [14] was clear as the education sessions and follow-up visits progressed. It is interesting that both indexes fell below the recommended levels at similar times, likely due to the gradual appearance of preventive effects against hypoglycemia after education. CV (recommended level: <36% [14]) is an indicator of glycemic variability and presents many complexities because it is influenced not only by TIR and TBR but also by the time above the range [14]. In this study, mean CV values remained below the maximum of 42.1% (41.1% after adjustment) throughout the follow-up period and did not change significantly. However, unlike TIR, CV appeared to correlate somewhat with the improvement patterns in TBR as the number of education sessions and follow-up visits increased. In this study, we suspect that the gradually and significantly improved TBR had a greater impact on the change in CV values than TIR (which remained relatively stable despite a slight gradual decrease).

Comparison With Prior Work

The key point we wish to emphasize in our study is that, by adapting and integrating existing in-person educational support programs into a digital communication app, it is possible to achieve long-term glycemic control outcomes in patients with T1D that have not been realized in previous studies. This approach is not only immediately applicable in clinical settings but also highly effective, using widely available smartphone communication apps with minimal cost and resources. This approach could serve as a robust model for digital remote support programs for patients with diabetes that are being promoted by the Korean government.

We aimed to develop comprehensive content and protocols that would maximize the benefits reported in previous studies. The unique advantage of our study compared with previous studies on similar educational programs is that it not only consisted of 1-time in-person education but was also supplemented with remote support by using an existing general-purpose smartphone app to strengthen and sustain the educational effect. It is also true that efforts were made to personalize the usability of the app. Because the ability to understand education and use smartphone apps varies depending on conditions such as age, gender, occupation, and academic ability, our medical staff adjusted the level of education according to the patient's level of understanding and used the method preferred by the patient (text, photo, video, or call).

Our research stands out for its practical integration of structured education and remote support using a common smartphone communication app, which is both cost-effective and scalable. Studies such as the "bant" app trial [15] show promise but are often limited by user engagement challenges, whereas app cloud systems [16] demonstrate modest gains in self-management behaviors without substantial HbA_{1c} level reductions. Unlike other studies [15-17] that may show limited long-term engagement or modest clinical outcomes, our study demonstrated sustained glycemic control improvements over time. This long-term effect and the seamless integration into real-world clinical settings make our approach highly applicable and potentially more impactful in broader health care systems, especially as an economical prototype for government-supported telemedicine programs.

It has already been established by many previous studies that promoting the actual or potential self-management capabilities of patients with diabetes is essential for improving short- and long-term prognosis regardless of the type of education used [18]. The implementation of a structured education program and follow-up for 6 months to 1 year simultaneously improved mean HbA_{1c} values and QOL without worsening hypoglycemia [19,20]. The effects of structured education for intensive insulin treatment in reducing mean HbA_{1c} values and improving hypoglycemia or QOL are already well known [21-23]. Moreover, nutritional education, including carbohydrate counting, has been shown to be beneficial in improving HbA_{1c} levels and increasing the consumption of recommended foods, although no robust evidence indicates that it improves glycemic control in patients with T1D yet [24-26]. Although the content of the education sessions offered and the characteristics of the study participants differ, those results are consistent with our research results.

Many previous studies have investigated the benefits of CGM and found that it provides significant improvement in the hypoglycemic index (TBR) and mean glucose values (eg, HbA_{1c} or TIR) of patients with T1D only when appropriately structured education is provided [27-29]. Most of our participants in this

XSL•FO
RenderX

study were using CGM (99/119, 83.2%), and those CGM users showed better results than other patients, especially active CGM users. It is likely that the effects of structured education on them influenced the overall positive outcomes of this study. Because the TIR value could not be obtained from baseline, we attempted to confirm the patterns of change in TIR according to the number of education sessions beginning at the first follow-up visit. The recommended TIR value for patients with T1D using CGM is >70% [14]. Although it varies by study, HbA_{1c} values of 7.4% (95% CI 6.1%-8.8%) and 7.9% (95% CI 6.6%-9.2%) generally correspond to a TIR of 60% and 50%, respectively [14]. The participants in this study who completed at least 1 educational session maintained a mean TIR of ≥55% (HbA_{1c} of approximately ≤7.6%) during the entire follow-up period. It is noteworthy that these outcomes are slightly better than the results of previous studies that compared glycemic control between patients using CGM and capillary blood measurement [30,31]. We confirmed that appropriate consistency was maintained between the mean HbA_{1c} values measured in this study and the mean TIR values in the AGP data of the study participants. The pattern of change in TIR values was similar to that of HbA_{1c} values, being relatively constant throughout the follow-up period without significant changes.

One more interesting finding in this study was the association between educational level and glycemic control and CGM use. Of the total study participants, 52.9% (63/119) provided their educational background. Among them, the group with relatively good glycemic control (HbA_{1c}<8%) at baseline had a higher proportion of individuals with a bachelor's degree or higher compared to the other group (n=30: 20/29, 69% vs 10/34, 29%; *P*=.007; Table 1). It has been suggested that the level of education may be related to complication awareness and treatment compliance in patients with type 2 diabetes [32] and that overall SES is correlated with glycemic control and complications in patients with T1D using flash glucose monitoring [33]. Although that was not the main concern of our study, it might have had some influence on the results. Because schooling information was available for only approximately half of our participants, caution is needed in interpretation. The data in Multimedia Appendix 3 adjust the results for SES-related information.

Limitations

This study has some limitations. Even though it is prospective, cohort studies have inherent limitations. Because it was not a randomized controlled trial, patients with high compliance might have been selectively recruited, and many confounding factors might have influenced the results despite our adjustment efforts. A high level of attrition (loss to follow-up) could lead to biased results. We have no objective verification of patient-provided information. Not all participants were able to properly provide their SES, including their educational level and economic status. Hypoglycemic events could not be confirmed through serum or capillary blood test, so cloud-based AGP data had to be used. CGM activity varied widely among patients using CGM, so the consistency of the AGP data could be limited. Different manufacturers and brands of CGM used by patients may also have some limitations. HbA_{1c} is an indicator of a 2- to 3-month average of blood glucose levels. Most participants were observed based on the 3-month follow-up intervals, but some patients visited earlier than these intervals, whereas others visited at delayed intervals. We tried to reconfirm the consistency of the presented HbA_{1c} values by observing the TIR values (reflecting the average HbA_{1c} in CGM users) and continuously tracking the patients' sequential HbA_{1c} values over a long period. Finally, despite the unfavorable conditions for glycemic control posed by the COVID-19 pandemic, paradoxically, presenting good results in this study can be considered a strength. However, the pandemic may have influenced some of the intervention processes applied in this study, which could be a limitation in interpreting the results.

Conclusions

Despite some limitations, this study has clinical and policy significance because it is the first study to implement and clinically confirm the positive results of a home health care project for patients with T1D supported by the Korean government. The periodic implementation of HELP has been demonstrated to be very effective in improving both the short- and long-term glycemic control of patients with T1D. Our task now is to further simplify the program to enable 1-stop educational support and restructure the content so that even patients with lower educational levels can easily adhere to it. To increase the use of remote support, it will be necessary to secure dedicated medical staff and use simulations at the beginning of the education sessions in a hospital. By strengthening the reliability of these results through more objective studies, such as randomized controlled trials, we expect to facilitate the general application of HELP. In addition, future research could explore the interesting comparison between cases in which only in-person education is provided and cases in which remote support is provided alongside in-person education.

Acknowledgments

The authors acknowledge the efforts of the Department of Academic Research at Kangbuk Samsung Hospital. Special appreciation is extended to the nursing (Suji Lee, Yunhee Seo, and Jiyeon Lee) and nutrition (Kyunghee Oh and Jieun Lee) education team members who contributed to this project.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

DJK, S-JM, and CYP contributed to conception or design. SM performed formal analysis. DJK, SEP, EJR, WYL, S-JM, and CYP contributed to acquisition, analysis, or interpretation of data. DJK drafted the manuscript. DJK, S-JM, and CYP contributed to discussion and revision. S-JM and CYP provided final approval for the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Home patient review checklist.
[PDF File (Adobe PDF File), 195 KB - jmir_v26i1e60023_app1.pdf]

Multimedia Appendix 2
Basic nutritional education course.
[PDF File (Adobe PDF File), 143 KB - jmir_v26i1e60023_app2.pdf]

Multimedia Appendix 3
Supplementary tables and figures.
[PDF File (Adobe PDF File), 1292 KB - jmir_v26i1e60023_app3.pdf]

Multimedia Appendix 4
Graphical abstract.
[PDF File (Adobe PDF File), 1503 KB - jmir_v26i1e60023_app4.pdf]

References

1. Gregory GA, Robinson TI, Linklater SE, Wang F, Colagiuri S, de Beaufort C, International Diabetes Federation Diabetes Atlas Type 1 Diabetes in Adults Special Interest Group, et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. Lancet Diabetes Endocrinol 2022 Oct;10(10):741-760. [doi: 10.1016/S2213-8587(22)00218-2] [Medline: 36113507]
2. Guo SJ, Shao H. Growing global burden of type 1 diabetes needs multitiered precision public health interventions. Lancet Diabetes Endocrinol 2022 Oct;10(10):688-689. [doi: 10.1016/S2213-8587(22)00257-1] [Medline: 36113506]
3. Foster NC, Beck RW, Miller KM, Clements MA, Rickels MR, DiMeglio LA, et al. State of type 1 diabetes management and outcomes from the T1D exchange in 2016-2018. Diabetes Technol Ther 2019 Feb;21(2):66-72. [doi: 10.1089/dia.2018.0384] [Medline: 30657336]
4. Riddlesworth T, Price D, Cohen N, Beck RW. Hypoglycemic event frequency and the effect of continuous glucose monitoring in adults with type 1 diabetes using multiple daily insulin injections. Diabetes Ther 2017 Aug;8(4):947-951 [FREE Full text] [doi: 10.1007/s13300-017-0281-4] [Medline: 28616804]
5. Heinemann L, Freckmann G, Ehrmann D, Faber-Heinemann G, Guerra S, Waldenmaier D, et al. Real-time continuous glucose monitoring in adults with type 1 diabetes and impaired hypoglycaemia awareness or severe hypoglycaemia treated with multiple daily insulin injections (HypoDE): a multicentre, randomised controlled trial. Lancet 2018 Apr;391(10128):1367-1377. [doi: 10.1016/s0140-6736(18)30297-6]
6. Hopkins D, Lawrence I, Mansell P, Thompson G, Amiel S, Campbell M, et al. Improved biomedical and psychological outcomes 1 year after structured education in flexible insulin therapy for people with type 1 diabetes: the U.K. DAFNE experience. Diabetes Care 2012 Aug;35(8):1638-1642 [FREE Full text] [doi: 10.2337/dc11-1579] [Medline: 22619082]
7. Abolfotouh MA, Kamal MM, El-Bourgy MD, Mohamed SG. Quality of life and glycemic control in adolescents with type 1 diabetes and the impact of an education intervention. Int J Gen Med 2011 Mar 20;4:141-152 [FREE Full text] [doi: 10.2147/IJGM.S16951] [Medline: 21475630]
8. Hermanns N, Kulzer B, Ehrmann D, Bergis-Jurgan N, Haak T. The effect of a diabetes education programme (PRIMAS) for people with type 1 diabetes: results of a randomized trial. Diabetes Res Clin Pract 2013 Dec;102(3):149-157. [doi: 10.1016/j.diabres.2013.10.009] [Medline: 24210673]
9. Mouslech Z, Somali M, Sarantis L, Christos D, Alexandra C, Maria P, et al. Significant effect of group education in patients with diabetes type 1. Hormones (Athens) 2018 Sep;17(3):397-403. [doi: 10.1007/s42000-018-0054-0] [Medline: 30112705]
10. O'Hara MC, Hynes L, O'Donnell M, Nery N, Byrne M, Heller SR, et al. A systematic review of interventions to improve outcomes for young adults with Type 1 diabetes. Diabet Med 2017 Jun;34(6):753-769 [FREE Full text] [doi: 10.1111/dme.13276] [Medline: 27761951]

XSL•FO
RenderX

11. Announcement. Customer Center Consultation. URL: https://www.hira.or.kr/bbsDummy.do?pgmid=HIRAA020002000100&brdScnBltNo=4&brdBltNo=7873&pageIndex=1&pageIndex2=1#none [accessed 2024-04-29]
12. Hur KY, Moon MK, Park JS, Kim S, Lee S, Yun J, et al. 2021 clinical practice guidelines for diabetes mellitus of the Korean Diabetes Association. Diabetes Metab J 2021 Jul;45(4):461-481 [FREE Full text] [doi: 10.4093/dmj.2021.0156] [Medline: 34352984]
13. Kim B, Kang SM, Kang JH, Kang SY, Kim KK, Kim K, et al. 2020 Korean Society for the Study of Obesity guidelines for the management of obesity in Korea. J Obes Metab Syndr 2021 Jun 30;30(2):81-92 [FREE Full text] [doi: 10.7570/jomes21022] [Medline: 34045368]
14. Battelino T, Danne T, Bergenstal RM, Amiel SA, Beck R, Biester T, et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. Diabetes Care 2019 Aug;42(8):1593-1603. [doi: 10.2337/dci19-0028] [Medline: 31177185]
15. Goyal S, Nunn CA, Rotondi M, Couperthwaite AB, Reiser S, Simone A, et al. A mobile app for the self-management of type 1 diabetes among adolescents: a randomized controlled trial. JMIR Mhealth Uhealth 2017 Jun 19;5(6):e82 [FREE Full text] [doi: 10.2196/mhealth.7336] [Medline: 28630037]
16. Handa T, Onoue T, Kobayashi T, Maeda R, Mizutani K, Yamagami A, et al. Effects of digitization of self-monitoring of blood glucose records using a mobile app and the cloud system on outpatient management of diabetes: single-armed prospective study. JMIR Diabetes 2024 Jan 19;9:e48019 [FREE Full text] [doi: 10.2196/48019] [Medline: 38241065]
17. Zimmermann G, Venkatesan A, Rawlings K, Scahill MD. Improved glycemic control with a digital health intervention in adults with type 2 diabetes: retrospective study. JMIR Diabetes 2021 Jun 02;6(2):e28033 [FREE Full text] [doi: 10.2196/28033] [Medline: 34075880]
18. Funnell MM, Brown TL, Childs BP, Haas LB, Hosey GM, Jensen B, et al. National standards for diabetes self-management education. Diabetes Care 2008 Jan;31 Suppl 1(Suppl 1):S97-104 [FREE Full text] [doi: 10.2337/dc08-S097] [Medline: 18165344]
19. DAFNE Study Group. Training in flexible, intensive insulin management to enable dietary freedom in people with type 1 diabetes: dose adjustment for normal eating (DAFNE) randomised controlled trial. BMJ 2002 Oct 05;325(7367):746 [FREE Full text] [doi: 10.1136/bmj.325.7367.746] [Medline: 12364302]
20. Cooke D, Bond R, Lawton J, Rankin D, Heller S, Clark M, et al. Structured type 1 diabetes education delivered within routine care: impact on glycemic control and diabetes-specific quality of life. Diabetes Care 2013 Feb;36(2):270-272 [FREE Full text] [doi: 10.2337/dc12-0080] [Medline: 23139374]
21. Plank J, Köhler G, Rakovac I, Semlitsch BM, Horvath K, Bock G, et al. Long-term evaluation of a structured outpatient education programme for intensified insulin therapy in patients with Type 1 diabetes: a 12-year follow-up. Diabetologia 2004 Aug;47(8):1370-1375. [doi: 10.1007/s00125-004-1456-x] [Medline: 15258736]
22. Müller N, Kloos C, Sämann A, Wolf G, Müller UA. Evaluation of a treatment and teaching refresher programme for the optimization of intensified insulin therapy in type 1 diabetes. Patient Educ Couns 2013 Oct;93(1):108-113. [doi: 10.1016/j.pec.2013.05.008] [Medline: 23747089]
23. Burghardt K, Müller UA, Müller N, Kloos C, Kramer G, Jörgens V, et al. Adequate structured inpatient diabetes intervention in people with type 1 diabetes improves metabolic control and frequency of hypoglycaemia. Exp Clin Endocrinol Diabetes 2020 Dec;128(12):796-803. [doi: 10.1055/a-0873-1465] [Medline: 31091546]
24. Schmidt S, Schelde B, Nørgaard K. Effects of advanced carbohydrate counting in patients with type 1 diabetes: a systematic review. Diabet Med 2014 Aug;31(8):886-896. [doi: 10.1111/dme.12446] [Medline: 24654856]
25. Fu S, Li L, Deng S, Zan L, Liu Z. Effectiveness of advanced carbohydrate counting in type 1 diabetes mellitus: a systematic review and meta-analysis. Sci Rep 2016 Nov 14;6(1):37067 [FREE Full text] [doi: 10.1038/srep37067] [Medline: 27841330]
26. Sterner Isaksson S, Bensow Bacos M, Eliasson B, Thors Adolfsson E, Rawshani A, Lindblad U, et al. Effects of nutrition education using a food-based approach, carbohydrate counting or routine care in type 1 diabetes: 12 months prospective randomized trial. BMJ Open Diabetes Res Care 2021 Mar;9(1):e001971 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001971] [Medline: 33789909]
27. ISCHIA Study Group. Prevention of hypoglycemia by intermittent-scanning continuous glucose monitoring device combined with structured education in patients with type 1 diabetes mellitus: a randomized, crossover trial. Diabetes Res Clin Pract 2023 Jan;195:110147 [FREE Full text] [doi: 10.1016/j.diabres.2022.110147] [Medline: 36396114]
28. Schlüter S, Freckmann G, Heinemann L, Wintergerst P, Lange K. Evaluation of the SPECTRUM training programme for real-time continuous glucose monitoring: a real-world multicentre prospective study in 120 adults with type 1 diabetes. Diabet Med 2021 Feb;38(2):e14467. [doi: 10.1111/dme.14467] [Medline: 33230860]
29. Yoo JH, Kim G, Lee HJ, Sim KH, Jin S, Kim JH. Effect of structured individualized education on continuous glucose monitoring use in poorly controlled patients with type 1 diabetes: a randomized controlled trial. Diabetes Res Clin Pract 2022 Mar;184:109209. [doi: 10.1016/j.diabres.2022.109209] [Medline: 35065101]
30. Beck RW, Riddlesworth T, Ruedy K, Ahmann A, Bergenstal R, Haller S, et al. Effect of continuous glucose monitoring on glycemic control in adults with type 1 diabetes using insulin injections: the DIAMOND randomized clinical trial. JAMA 2017 Dec 24;317(4):371-378. [doi: 10.1001/jama.2016.19975] [Medline: 28118453]

XSL•FO
RenderX

31. Leelarathna L, Evans ML, Neupane S, Rayman G, Lumley S, Cranston I, et al. Intermittently scanned continuous glucose monitoring for type 1 diabetes. N Engl J Med 2022 Oct 20;387(16):1477-1487. [doi: 10.1056/nejmoa2205650]
32. Al-Rasheedi AA. The role of educational level in glycemic control among patients with type II diabetes mellitus. Int J Health Sci (Qassim) 2014 Apr;8(2):177-187 [FREE Full text] [doi: 10.12816/0006084] [Medline: 25246885]
33. Sebastian-Valles F, Martínez-Alfonso J, Arranz Martin JA, Jiménez-Díaz J, Hernando Alday I, Navas-Moreno V, et al. Impact of socioeconomic status on chronic control and complications of type 1 diabetes mellitus in users of glucose flash systems: a follow-up study. BMC Med 2024 Jan 25;22(1):37 [FREE Full text] [doi: 10.1186/s12916-024-03254-w] [Medline: 38273326]

Abbreviations

AGP: ambulatory glucose profile
CGM: continuous glucose monitoring
CV: coefficient of variation
HbA_{1c}: glycated hemoglobin
HELP: Home and Self-Care Program
QOL: quality of life
SES: socioeconomic status
T1D: type 1 diabetes
TBR: time below the target range
TIR: time in the target range

Edited by G Eysenbach, N Cahill; submitted 29.04.24; peer-reviewed by S Schembre, AN Ali; comments to author 24.05.24; revised version received 09.06.24; accepted 23.08.24; published 11.09.24.

Please cite as:
Koo DJ, Moon SJ, Moon S, Park SE, Rhee EJ, Lee WY, Park CY
Long-Term Glycemic Control Improvement After the Home and Self-Care Program for Patients With Type 1 Diabetes: Real-World–Based Cohort Study
J Med Internet Res 2024;26:e60023
URL: https://www.jmir.org/2024/1/e60023
doi:10.2196/60023
PMID:

©Dae-Jeong Koo, Sun-Joon Moon, Suhyeon Moon, Se Eun Park, Eun-Jung Rhee, Won-Young Lee, Cheol-Young Park. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the National Rollout of a Type 2 Diabetes Self-Management Intervention: Qualitative Interview Study With Local National Health Service Leads Responsible for Implementation

Lisa Brunton[1], PhD; Sarah Cotterill[1], PhD; Paul Wilson[1], BA (Hons)

Division of Population Health, Health Services Research & Primary Care, School of Health Sciences, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Lisa Brunton, PhD
Division of Population Health, Health Services Research & Primary Care
School of Health Sciences
University of Manchester
Suite 3, 6th Floor Williamson Building
Oxford Road
Manchester, M13 9PL
United Kingdom
Phone: 44 161 275 7603
Email: lisa.brunton@manchester.ac.uk

Abstract

Background: Approximately 4.5 million people live with type 2 diabetes mellitus (T2DM) in the United Kingdom. Evidence shows that structured education programs can improve glycemic control and reduce the risk of complications from T2DM, but they have low attendance rates. To widen access to T2DM structured education, National Health Service England commissioned a national rollout of Healthy Living, a digital self-management program.

Objective: The objectives were to understand the barriers and enablers to adopting, implementing, and integrating Healthy Living into existing T2DM care pathways across England.

Methods: We undertook a cross-sectional, qualitative telephone semistructured interview study to address the objectives. In total, 17 local National Health Service leads responsible for implementing Healthy Living across their locality were recruited. We conducted 16 one-time interviews across 16 case sites (1 of the interviews was conducted with 2 local leads from the same case site). Interview data were analyzed using thematic analysis.

Results: Three overarching themes were generated: (1) implementation activities, (2) where Healthy Living fits within existing pathways, and (3) contextual factors affecting implementation. Of the 16 sites, 14 (88%) were implementing Healthy Living; the barrier to not implementing it in 2 case sites was not wanting Healthy Living to compete with their current education provision for T2DM. We identified 6 categories of implementation activities across sites: communication strategies to raise awareness of Healthy Living, developing bespoke local resources to support general practices with referrals, providing financial reimbursement or incentives to general practices, promoting Healthy Living via public events, monitoring implementation across their footprint, and widening access across high-need groups. However, outside early engagement sites, most implementation activities were "light touch," consisting mainly of one-way communications to raise awareness. Local leads were generally positive about Healthy Living as an additional part of their T2DM structured education programs, but some felt it was more suited to specific patient groups. Barriers to undertaking more prolonged, targeted implementation campaigns included implementation not being mandated, sites not receiving data on uptake across their footprint, and confusion in understanding where Healthy Living fit within existing care pathways.

Conclusions: A passive process of disseminating information about Healthy Living to general practices rather than an active process of implementation occurred across most sites sampled. This study identified that there is a need for clearer communications regarding the type of patients that may benefit from the Healthy Living program, including when it should be offered and whether it should be offered instead of or in addition to other education programs. No sites other than early engagement sites received

XSL•FO
RenderX

data to monitor uptake across their footprint. Understanding variability in uptake across practices may have enabled sites to plan targeted referral campaigns in practices that were not using the service.

(J Med Internet Res 2024;26:e55546) doi:10.2196/55546

KEYWORDS

type 2 diabetes; structured education; self-management; digital interventions; implementation; qualitative methods; evaluation

Introduction

Background

Recent figures from Diabetes UK show that approximately 4.5 million people now live with type 2 diabetes mellitus (T2DM) in the United Kingdom [1], a major public health concern. People diagnosed with T2DM are at risk of developing complications that include cardiovascular disease, renal disease, sight problems, and limb amputations [2,3]. The cost of diabetes care accounts for approximately 10% (£8.8 billion [US $11.6 billion] per year and rising) of the total National Health Service (NHS) budget [4]. Most costs are related to managing complications of diabetes [5]; however, when the condition is managed well through taking medications as prescribed and modifying lifestyle factors (such as eating a healthy diet, increasing physical activity, and maintaining a healthy weight), complications can be minimized [1]. A systematic review and meta-analysis identified that group-based education programs were more effective than usual care in improving clinical, lifestyle, and psychosocial outcomes for people with T2DM [6].

In England, the National Institute for Health and Care Excellence recommends offering evidence-based structured education programs to all adults with T2DM (and their family members or carers) at the time of diagnosis as appropriate, with yearly reinforcement and review [7]. T2DM structured education programs such as X-PERT and Diabetes Education and Self-Management for Ongoing and Newly Diagnosed (DESMOND) are usually offered to newly diagnosed patients with T2DM as face-to-face group sessions [8,9]. However, uptake of structured education programs in England is low, with the latest National Diabetes Audit 2021 to 2022 data showing that, of the 65% of newly diagnosed patients with T2DM offered structured education within 12 months of diagnosis, only 7.5% of people attended a session within the first year following diagnosis [10]. Some people choose not to take up diabetes education for reasons such as lack of time due to family or work commitments, disability, or existing multimorbidity or due to associated costs such as traveling to the venue [11]. To widen access to T2DM support in England, the NHS Long Term Plan committed to expanding access to structured education and digital self-management tools, including widening access to Healthy Living for People With Type 2 Diabetes (HeLP Diabetes) [12]. HeLP Diabetes is a self-management, web-based program originally developed and tested in a randomized controlled trial by researchers at University College London (UCL) [13,14]. HeLP Diabetes was shown to significantly (but modestly) reduce blood glucose levels and be cost-effective compared to usual care and considered feasible to deliver at scale [15,16]. The UCL research team iteratively developed an implementation plan to support the implementation of HeLP Diabetes into routine care [17] using normalization process theory (NPT), an implementation theory that is used to explore how new processes become or do not become routine practice [18], as a guide to development. The recommended implementation strategies included engaging with relevant stakeholders in the early stages of implementation, providing educational support materials, conducting general practice site visits and meetings, and conducting audit and feedback to identify how each general practice was performing and providing regular emails and newsletters to services to promote continued engagement [17]. The UCL research team found that the main barrier to implementation centered on the requirement of staff to provide onboarding support, with some nurses feeling that it was not a legitimate part of their role. A main facilitator of successful implementation was providing staff with patient feedback on using HeLP Diabetes, which was considered to motivate staff to continue to engage with the program [17].

Previous literature has identified factors that influence implementation of eHealth technologies into health services [19,20]. A systematic review of reviews highlighted several factors, including the importance of relevant policies and incentives to drive implementation at the organizational level; implementation climate (ie, the fit between the intervention and the organization) and organizational readiness to implement the intervention; individuals' knowledge and beliefs; and the importance of planning for implementation, including early engagement of key stakeholders, and evaluation to identify system benefits to increase health professional acceptance [19]. A recent narrative review of technology-delivered diabetes self-care interventions identified that the main barriers to implementation included clinicians' time constraints and concern that interventions may increase workload and patients' and clinicians' lack of familiarity with technology or the intervention [20]. The main facilitators to successful implementation included patients' and clinicians' perceived value of the intervention and interventions being of low cost to use, deliver, and maintain [20].

National Rollout of Healthy Living

NHS England commissioned the national rollout of HeLP Diabetes under the name Healthy Living from 2019 onward. An external digital service provider was appointed to develop and deliver the national program. Unlike HeLP Diabetes, the national scale-up of the Healthy Living program is a self-contained and self-directed web-based program with no health professional support beyond initial referral and signposting [21]. NHS England originally planned a phased rollout with 10 early engagement areas expected to adopt Healthy Living first, working with NHS England to develop and test implementation strategies that could be used for national

XSL•FO
RenderX

rollout. Early engagement areas were given £30,000 (US $39,375.60) to support implementation efforts. NHS England's original intention was that the primary referral route would be through general practices. However, in March 2020, due to the COVID-19 pandemic, NHS England amended their implementation plans to (1) enable early engagement areas to delay implementation to address local pressures and (2) enable a self-referral route for people to self-register without invitation from general practices. Implementation across most early engagement areas was delayed until late 2020, and NHS England worked with early engagement areas to amend their original implementation plans as the prioritization of the development of referral routes changed.

National rollout via general practices started in May 2022. National rollout via general practices was to be organized through the 42 integrated care systems (ICSs) across England—ICSs are partnerships of organizations that join to plan and deliver health and care services across a geographic area [22]. NHS England considered 4 implementation options for national rollout (Textbox 1). In response to the multiple pressures placed on primary care following the COVID-19 pandemic, NHS England chose to roll out option 3, whereby Healthy Living was provided to ICSs as an offer of support but with no formal requirements placed upon ICSs to implement the program. NHS England communicated to general practices that Healthy Living was an addition to their existing locally commissioned T2DM structured education programs and should not be considered a replacement for them.

Textbox 1. Implementation options considered for the national rollout of Healthy Living (information received by personal correspondence with National Health Service England).

Implementation options

- Scenario 1: formal, national rollout of Healthy Living that requires that all integrated care systems (ICSs) implement the program
- Scenario 2: formal, partial rollout of Healthy Living that requires only those ICSs that have either none or limited Digital Structured Education provision; other ICSs with Digital Structured Education provision would be able to opt in to implement
- Scenario 3: Healthy Living program as an offer of support—offered to those who want it, but no formal requirement for any ICS to implement it
- Scenario 4: a mix of scenarios 1 to 3

Aim of Research

Healthy Living is a considerable investment of NHS funds and, although it is based on evidence from the HeLP Diabetes randomized controlled trial [15,16], evidence is required to show its effectiveness now that it has been rolled out nationally. Therefore, researchers from the University of Manchester undertook an independent evaluation of the real-world implementation of Healthy Living, including uptake and engagement, effectiveness and cost-effectiveness, fidelity, and how it is being implemented across sites in England [23]. This part of the research aimed to understand the barriers and enablers to adopting, implementing, and integrating Healthy Living into T2DM care across England.

Methods

Design

We undertook a cross-sectional, qualitative semistructured interview design.

Respondents

Respondents were local leads responsible for implementation of Healthy Living across their locality or ICS.

Sampling and Recruitment

We aimed to interview at least one participant from each ICS across England (N=42) whether they were currently implementing Healthy Living or not. We received names and work email addresses of potential respondents from the Healthy Living national delivery team at NHS England. Email interview invites, with a participant information sheet attached, were sent to potential respondents, and we sent a reminder email after 2 weeks to nonresponders. We used snowball recruitment, in which additional names were provided by local leads (often in cases in which people had changed job roles). We recruited participants from 16 case sites, and saturation of data was reached as later interviews produced little or no new information or themes.

Data Collection

We planned to conduct interviews in 2 stages, with local leads in the early engagement areas being interviewed first so that findings could inform subsequent rollout. However, implementation activities in the early engagement areas were delayed due to the COVID-19 pandemic, and so data collection occurred in all sites at 1 time point between August 2022 and December 2022.

Interviews were conducted by the first author, LB, who is an experienced qualitative researcher. Most interviews were conducted over the telephone, but a small number were conducted over Microsoft Teams at the request of the respondents. A topic guide was used to manage the interviews (Multimedia Appendix 1). All interviews were audio recorded on an encrypted digital audio recorder and transcribed verbatim. Transcriptions were subsequently checked for accuracy by LB and anonymized to be ready for analysis.

Data Analysis

Interview transcripts were uploaded to the NVivo qualitative software (version 12; QSR International) for data management and coding. The data were analyzed thematically [24] using the constant comparative method to systematically interrogate the

XSL•FO
RenderX

data [25]. LB read and reread transcripts and created a summary (memo) of each interview to familiarize herself with the data and develop an inductive coding framework. LB coded the interviews, and the coding framework was refined as more interviews were coded. Data were summarized into a table of categories; this was created in Microsoft Word and included selected verbatim quotes to support the categories. We continued to work between the table of categories and the full transcripts and codes in NVivo; this enabled us to generate themes and compare the data systematically to identify the similarities and differences within and across the data. Themes were refined after presenting early findings at 2 wider research team meetings and a patient and public involvement meeting that was attended by 7 members comprising people at risk of developing T2DM, those with a diagnosis of T2DM, and some who cared or had cared for older relatives with T2DM. Throughout the analysis process, LB and PW held regular research meetings to further scrutinize the data and generate the interpretive themes presented. The final set of themes presented was agreed upon by all authors.

Ethical Considerations

The full research program (of which this study is part) was reviewed and approved by the Yorkshire and the Humber–Leeds West NHS Research Ethics Committee (reference 20/YH/0250). Before the interview started, verbal informed consent was digitally audio recorded using a prepared form that was emailed to respondents ahead of the interview. Data were anonymized during the transcription process to remove any traceable information and maintain participants' confidentiality. Participants were not paid to take part in the interviews.

Results

Respondents

A total of 17 local leads were recruited into the study, and 16 interviews were conducted across 16 sites; 1 (6%) interview was carried out with 2 local leads at the same site. A total of 24% (4/17) of the local leads were from early engagement areas, and 76% (13/17) were from 12 sites that had started to implement Healthy Living following national rollout. The interviews lasted between 23 and 61 minutes (mean 35 minutes).

Most respondents (14/17, 82%) were project or program managers or support officers, but 18% (3/17) of the respondents held more senior strategic, commissioning roles. Most (15/17, 88%) had a remit to work across the whole footprint of their ICS; however, 12% (2/17) of the respondents only worked in a specific locality within their ICS. In total, 41% (7/17) of the local leads had a portfolio of work that involved diabetes alone, whereas most had a broader responsibility for other long-term conditions (8/17, 47%), prevention (1/17, 6%), or commissioning and transformation (1/17, 6%).

We generated three overarching themes from the analysis: (1) implementation activities, (2) where Healthy Living fits within existing pathways, and (3) contextual factors affecting implementation.

Theme 1: Implementation Activities

Overview

A total of 88% (14/16) of the sites were implementing Healthy Living at the time of the interview. Across the 14 sites that were implementing Healthy Living, we identified six categories of implementation activities that occurred: (1) organizing communication strategies to raise awareness of Healthy Living, (2) developing bespoke local resources to support general practices with referrals (in addition to the national resources provided), (3) providing financial reimbursement or incentives to general practices to encourage them to search their registers and send out SMS text messages, (4) promoting Healthy Living alongside other diabetes structured education programs to the local general population via public events, (5) monitoring the implementation of Healthy Living across their footprint, and (6) widening access to Healthy Living across high-need groups. Table 1 outlines each implementation activity reported and which sites used it.

Reasons for not implementing Healthy Living in 12% (2/16) of the sites focused on not wanting to compete with their current structured education provision. In site 1 (early engagement area), patients with T2DM were referred to structured education via a patient referral hub; they did not want Healthy Living to be offered as a first-line program, preferring instead for people to be offered face-to-face services in the first instance. Therefore, they were still trying to work out where to offer Healthy Living within their T2DM care pathway:

> *And at present in [site 1] [Healthy Living] is yet to be offered to anybody, because we didn't want to stray away from the actual pathway of [the patient referral hub]...cause the fear would be [if] we offer [Healthy Living] to everyone, and then they would never ever go on to the [patient referral hub]...[that] then stops everybody from having a physical choice of all the other things they could have.* [Local lead 03; site 1 (early engagement area)]

Site 10 chose to delay implementation of Healthy Living because they had recently commissioned a new digital program and intended to evaluate its effectiveness before they introduced another digital offer into their T2DM care pathway:

> *[T]he work with [the newly commissioned digital programme] was seen as...let's not make it more complicated...and then see how we could maybe bring [Healthy Living] into what we're doing [in future]...there are matters in terms of what are we paying for? And can we see the benefit of that? And this is where doing too much at once would muddy the water.* [Local lead 06; site 10]

XSL•FO
RenderX

Table 1. Implementation activities across case sites.

Implementation activity	Total number of sites undertaking the activity, n (%)	Early engagement areas undertaking the activity, n (%)
Communications to general practices	10 (63)	3 (75)
Include as item in newsletter	7 (44)	2 (50)
Direct contact with general practices via email or telephone	4 (25)	2 (50)
Disseminate national implementation support resources	4 (25)	2 (50)
Webinars undertaken with social prescribing teams and local pharmacy teams	1 (6)	1 (25)
Attend local general practice meetings	1 (6)	1 (25)
Promote with relevant staff at strategic meetings	6 (38)	2 (50)
Promote at local staff educational events	2 (13)	1 (25)
Add to general practice email correspondence for T2DM[a] options	1 (6)	0 (0)
Provide specialist teams with email banner to promote Healthy Living	1 (6)	0 (0)
Developing bespoke local resources	6 (38)	2 (50)
Create search template	1 (6)	1 (25)
Create SMS text message template	1 (6)	1 (25)
Create flowchart of T2DM structured education offers available (including Healthy Living)	4 (25)	1 (25)
Use of financial reimbursement or incentives	3 (19)	2 (50)
Provide general practices with direct financial incentive to carry out search of register, send SMS text messages and complete tracker	2 (13)	2 (50)
Add Healthy Living to local enhanced service[b] for T2DM care pathway	2 (13)	1 (25)
Raising awareness among the public	5 (31)	2 (50)
Promote at local community events	2 (13)	2 (50)
Promote via social media channels or request to promote on general practice website	4 (25)	1 (25)
Monitoring implementation	3 (19)	3 (75)
Request general practices to complete tracker	3 (19)	3 (75)
Use data on uptake from NHS[c] England to monitor implementation performance	3 (19)	3 (75)
Widening access	4 (25)	2 (50)
Plan to conduct focus groups with people from ethnic minority groups	1 (6)	1 (25)
Promote to homeless health care team	1 (6)	1 (25)
Promote to local traveler community	1 (6)	1 (25)
Translate Healthy Living poster into different languages	1 (6)	1 (25)
Plan to directly contact general practices in deprived areas that have not engaged with Healthy Living	2 (13)	2 (50)
Request that clinicians signpost patients who decline locally commissioned T2DM structured education to Healthy Living	3 (19)	1 (25)

[a]T2DM: type 2 diabetes mellitus.

[b]General practices receive financial reimbursement for meeting targets related to T2DM care.

[c]NHS: National Health Service.

XSL•FO
RenderX

Subtheme 1.1: Light-Touch Implementation

At the time of data collection, outside the early engagement areas, sites undertook "light touch" implementation activities, including communication strategies to raise awareness of Healthy Living among general practice staff responsible for referring patients to T2DM structured education, discussing the program at local board meetings with those responsible for commissioning T2DM structured education, and raising patient awareness via social media channels:

> *We haven't asked people to engage with us...we asked them to circulate to everyone that does the diabetes reviews...offering education is part of the review...so we just want to make sure that everyone that's carrying out a review is aware of it. And then we also asked them to put it on their websites.* [Local lead 08; site 9]

> *I did communicate it out to primary care through our GP bulletin, saying if you would like to promote within your practices, you're welcome to promote it. So, I haven't given them all the information the NHS England have given us...so it's been quite left to primary care if they want to do anything about it.* [Local lead 17; site 6]

> *Yeah, we do ongoing comms. I mean, recently, we did a big piece related to the Diabetes Week, as well, so we did some ICS-wide, which is the Integrated Care System, right across.* [Local lead 07; site 11]

None of the sites from national rollout had been prescriptive regarding how Healthy Living was to be promoted by general practice and other community teams to their populations of people living with T2DM. Some of the early engagement areas reported encouraging general practices to conduct searches of their practice registers to identify eligible people and send out SMS text messages. This was regarded as the method of referral that placed the least burden on primary care given other pressures they faced during and after the COVID-19 pandemic:

> *I think just sticking to this text messaging method of roll out during the pandemic was the easiest thing for practices to implement. So it, kind of, like broke down some of those barriers of, we don't have the staff to be printing off letters, or we don't have the money for postage...it just took away a few of those barriers and helped.* [Local lead 1; site 4 (early engagement area)]

Subtheme 1.2: Monitoring Referrals to Healthy Living

Early engagement areas monitored referrals into Healthy Living across their footprint. This entailed asking general practices to complete a tracker to show how many SMS text messages or letters had been sent out to eligible people; this was cross-referenced with the data received from NHS England to identify how many people had registered or activated an account with Healthy Living to identify the level of uptake. This enabled local leads to consider widening access to Healthy Living to high-need populations. For example, some early engagement areas identified that general practices in deprived areas were less likely to promote Healthy Living to their patients and suggested that more work was needed to raise awareness in these localities, whereas site 4 (early engagement area) had also commissioned a local community group to carry out focus groups as their data showed that people from ethnic minority groups were less likely to register for the program:

> *So, just looking at the spreadsheet now, those that signed up [to send out text messages to eligible patients]...probably looking at three from that [deprived] area...we need to do more work...and I think it's across the board in terms of diabetes [care] in general...the uptake from practices and PCNs [primary care networks] [in deprived localities] isn't as high as other [localities].* [Local lead 4; site 5 (early engagement area)]

> *Then we get to...one of our most deprived areas and the highest prevalence of diabetes and only two practices [out of 14] sent text messages...so we need to work there.... In addition, it's showing low uptake among ethnic minority groups. So, we've commissioned a piece of work with a local community group...to do some focus groups and understand why they're not signing up to Healthy Living.* [Local lead 01; site 4 (early engagement area)]

None of the sites from national rollout monitored how many people were offered Healthy Living across their area, and at the time of the interviews, they had received no data from NHS England regarding uptake.

Subtheme 1.3: How Implementation Monies Were Spent in Early Engagement Areas

As outlined previously, early engagement areas received monies to support their implementation efforts, and these were used in different ways. In one site (site 4), they employed a part-time support officer for 6 months whose main role was to telephone general practices to raise awareness of Healthy Living. They also used the money to commission the focus groups discussed previously. In total, 50% (2/4) of the sites chose to use their implementation monies to provide financial incentives to support general practices to carry out specific clinical register searches for eligible patients and send out SMS text messages. Financial incentives given ranged from £200 (US $262.50) per general practice (site 5) to £1800 (US $2362.54) per general practice (site 7) and were considered a good use of the funds to ensure that general practices engaged with the program in the early adoption phase:

> *We've still got money left over from the pilot still. So, that's what that was used for...some of them still haven't applied for the £200. But no, I think it's a good incentive for them, you know, looking at some spreadsheets that we've received back in terms of the trackers, you know, it takes them an hour, and hour and a half at most, so for £200...it's a good incentive rate for an hour's work.* [Local lead 04; site 5 (early engagement area)]

XSL•FO
RenderX

Theme 2: Where Does Healthy Living Fit Within Existing Pathways?

Overview

Most respondents perceived Healthy Living to be an addition to the existing T2DM structured education that they had in place, which was in line with the message from NHS England. Some interpreted this to mean that Healthy Living should be offered to patients only after they had attended (or declined to attend) a locally commissioned face-to-face T2DM structured education program:

> *And Healthy Living, from a lot of people's perceptions, including mine, is something that should be offered after a different education has been sought. Because the role of peer support and community [support] has a greater impact.* [Local lead 03; site 1 (early engagement area)]

> *The problem we have with local programmes, you can't make face to face—[people] work...if they can't access anything locally or haven't had the opportunity to access anything locally, there is a national programme. So, it is around just best managing their condition and health and wellbeing, et cetera.* [Local lead 13; site 14]

Local leads expressed mixed views regarding the Healthy Living program. Healthy Living was mainly perceived to be a positive addition to the portfolio of T2DM structured education. Some respondents described how they welcomed a free, national offer that was available to all patients (and their family members or carers) regardless of time since diagnosis of T2DM. This contrasted with their locally commissioned T2DM structured education offers that were often restricted to newly diagnosed patients:

> *And that's one of the reasons why I really love, actually, having the universal offer that's provided by NHS England, because it provides a sort of consistent base...it will just mean that everybody will be able to receive the same information from the same source as a standardised approach...it can be based upon what is needed, rather than what the demand...is for a structured education programme.* [Local lead 15; site 16]

However, positive views were often tempered, with local leads also expressing reservations regarding where Healthy Living fit within their existing pathways due to the concerns outlined in the following sections.

Subtheme 2.1: Duplication and Confusion

Some perceived Healthy Living to duplicate existing locally commissioned digital structured T2DM education services, and despite communication from NHS England that Healthy Living was an addition to and not a replacement for locally commissioned programs, one case site (site 16) reported that they had decided to stop commissioning their local digital offer:

> *So, with the new offer [Healthy Living] coming out from NHS England, and it being uncapped and available for anybody to access regardless of whether they are newly diagnosed or not, we decided that it would be better to actually focus on that [Healthy Living], rather than advertising another locally commissioned one as well.* [Local lead 15; site 16]

Others expressed general concern regarding perceived duplication and inefficiency within the NHS:

> *And this is sometimes the thing with the NHS...you end up with all these offers and all these services and there's quite often overlap and there's quite often duplication and it seems like no one quite knows how to best manage it or how to avoid those duplications.* [Local lead 02; site 7 (early engagement area)]

Moreover, respondents described the numerous lifestyle offers that existed for people with T2DM in primary care and how this led to confusion in understanding what type of patients was suitable for what type of program. Local leads from a number of sites reported that they were still grappling with where Healthy Living fit within their existing T2DM care pathways, which contributed to the "light touch" implementation approach described previously:

> *I think what we hear a lot from practices, and in fact our colleagues across the system really, it's the confusion around how many lifestyle programmes are on offer and where to signpost people, where to refer people to. We get that constantly and it is almost no matter how many times you go through the offer, people are still very confused.* [Local lead 14; site 15]

> *Question: how do you feel that [Healthy Living] fits with the existing services in your area?*

> *Local lead 10: Well, we found it quite difficult, didn't we? The reason we took a while to decide to roll it out is we were concerned how does it fit into the structured education on offer? What is the additional requirement?* [Site 12]

> *So we've tried to not roll it out unless clinicians understood where it might go on their pathway already.* [Local lead 09; Site 12]

> *[W]e still get quite a lot of feedback that there is confusion in the system. So if a GP diagnoses someone with diabetes, do they send them to the local structured education or do they suggest they use Healthy Living or do they suggest the other national programme being pushed quite a lot, the digital weight management [programme]? Which is the most appropriate? The GP staff don't necessarily know. So there's still quite a lot of confusion around that.* [Local lead 11; site 13]

Subtheme 2.2: Healthy Living Perceived to Be More Suited to Specific Patient Cohorts

Several respondents perceived that the Healthy Living program as a self-directed web-based program without health professional support was geared more toward certain population types; for example, they considered that it may be more suitable for (and appeal more to) native English-language speakers and people who were well educated and motivated to learn, with some suggesting that it would be useful for younger, working-age

XSL•FO
RenderX

populations who may not find time to attend a face-to-face course:

> *As our demographic [with T2DM] are getting younger, they'll want the app...so I think it will be quite reasonable for a younger person...who are tech savvy to be offered a structured education through an app.* [Local lead 16; site 3]

> *I definitely think it's going to be a good resource for people who are...happy accessing things online and maybe people who have gone through a fair bit of education as well, because you're used to using this type of resource and reading articles...* [Local lead 02; site 7 (early engagement area)]

> *Probably [more suited] to those that are more affluent and digital ready, and people that are of working age and this fits in around their lifestyle, which is the main asset of this programme.* [Local lead 03; site 1 (early engagement area)]

Local leads perceived it to be less suitable for (and appeal less to) populations who were likely to be more predisposed to T2DM, such as people from specific ethnic minority groups (including those whose first language is not English), people from more deprived areas (including those that were digitally excluded), and people with learning disabilities. This led some to express concern that national programs had the potential to widen health inequalities rather than address them because they perceived such programs as not being tailored toward those with the highest needs. Healthy Living as a web-based, self-directed program was not considered to help ICSs in reaching out to their high-need patients, particularly those who experienced barriers to accessing health care:

> *But we don't just want to reach out to the white, the worried well, those that are naturally aware and concerned about their mental and physical wellbeing; I'm not worried about those. Those will naturally reach out...it's those that don't and need a bit more support from us to be able to do that. And that's where I don't know how suitable Healthy Living would be best placed for that. It's an open challenge really, but you can say that for any large scale, national programme.* [Local lead 12; site 2]

> *And I guess the other thing to consider is the health inequalities aspect. So...a couple of our [localities] have got some quite distinct populations...So it's actually, how accessible is [Healthy Living] for a lot of our cohorts who are the ones who are most predisposed to having type 2 diabetes? So, it feels as though, even if [Healthy Living]'s there, we're still going to have to offer something that meets everyone else's needs.* [Local lead 05; site 8]

> *Obviously, I think we're not solely the only [ICS] that's raised the concern around digital exclusion. Especially noting that...some of our areas are very deprived and obviously, the cost of living is affecting everybody, so there's also the potential of the impact from that and the difference in broadband access as well, which should be recognised.* [Local lead 09; site 12]

Theme 3: Contextual Factors Affecting Implementation

This theme highlights the contextual factors that impeded sites from conducting more intensive and targeted implementation activities.

Subtheme 3.1: Lack of Feedback on Uptake Data

The main contextual factor affecting implementation, and the main barrier to conducting more intensive or targeted implementation activities for local leads from national rollout sites, was not receiving any data from NHS England on uptake. Local leads suggested that it would be useful to receive data on how many patients within their ICSs (or from which general practices) had registered on Healthy Living or how many had activated their accounts or completed the program. Some local leads compared Healthy Living to other nationally commissioned programs such as the NHS Digital Weight Management Programme [26], for which they did receive such data, and questioned why it was not possible to receive similar uptake data for Healthy Living:

> *So, I think sometimes, having the stats, so when we get the reports from the NHS Digital Weight Management Programme, I think that just gives us a way of communicating [to general practices] how the programme is going. It also gives us an opportunity to put it on as an agenda item as part of our diabetes structured education working group...So I think it would be helpful to hear whether [Healthy Living] is being taken up by people living in [site 15]. Because I wouldn't really know whether it's a useful programme for people to access, whether people are accessing it.* [Local lead 14; site 15]

> *...there's only any point in promoting [Healthy Living] if we know that people are actually going to it, completing it, using it and seeing the benefits of it. And I guess the difference for us is, if it's our own [locally commissioned structured education programme] we get all that information. If we direct them to an NHS England tool, we've got no idea whether they turn up or not. But, we're still measured on it. So there's a bit of a disconnect for me around, well, if we have no control, how can we influence?* [Local lead 05; site 8]

While lack of data was not a barrier to initially raising awareness of Healthy Living across their general practices, lack of data *was* a barrier to conducting subsequent, targeted implementation campaigns:

> *We're not even going to have any data on how many people are using it for a very long time. So, in the absence of that, it's quite hard to know what to do...it's not a barrier to rolling it out, more of a barrier to continuing to push it...* [Local lead 08; site 9]

> *What we've done is pushed it out, told everybody, said this is what's available...but we've not done any more detailed work with that, and because, we don't*

XSL•FO
RenderX

> *know, do we. So, we don't know whose picked it up and who hasn't, so we can't do that until we've got that information.* [Local lead 10; site 12]

Some respondents also reported confusion in knowing how patients who had registered and attended Healthy Living would be reported back to general practices to enable the latter to code them as attending a structured education program in general practice clinical systems, something they were audited on. This led to concerns that patients who attended Healthy Living may not count in their figures for the number of patients attending T2DM structured education and appeared to be another barrier to implementation. In addition, not receiving data on uptake led to confusion over whether Healthy Living did indeed count as attending a structured education program:

> *If practices are...not mailing out to all their patients and relying on in-practice advertisements, we won't be able to map [those patients] to that particular practice because there's no requirement for an NHS number [to register for Healthy Living]. So, for some patients, there's no way of knowing who has actually gone on to attend from a particular area, other than just the postcode. So, we're not going to be able to compare.* [Local lead 15; site 16]

> *...an offer of [Healthy Living], would that then meet the requirement of being offered structured education because they've got to opt into that themselves and do that self-care themselves...if they get a referral into secondary care [diabetes structured education] it feels more of a must do...as opposed to here's [Healthy Living] it's up to you to access it.... I have asked NHS England for our metrics on who's got access to Healthy Living and haven't heard anything back, so it's a bit of a problem.* [Local lead 16; site 6]

Subtheme 3.2: Implementing Healthy Living in a Crowded Market

Another contextual factor affecting implementation related to the perception from local leads that primary care was "bombarded" with communications on the many offers available for patients with T2DM. Local leads were cognizant of not wanting to "overload" general practices and so tread cautiously in the frequency of communications:

> *Our senior team are having discussions with the primary care leads to find out a better way of doing this so we're not bombarding them or drip-feeding different things.* [Local lead 01; site 4 (early engagement area)]

> *I think it is trying to get it out in a crowded market. In terms of diabetes at the minute...prevention, footcare, low calorie, there's a lot of stuff going off at the minute anyway.* [Local lead 04; site 5 (early engagement area)]

Subtheme 3.3: Timing of National Rollout

Early engagement sites were faced with implementing the program just as the COVID-19 pandemic happened, and this severely delayed their implementation. For sites that implemented Healthy Living following national rollout, local leads described how primary care was still dealing with COVID-19 recovery efforts and continued to be overstretched due to winter pressures and organizing influenza and COVID-19 vaccination programs. Given that there was no formal mandate for ICSs to implement Healthy Living, local leads were reluctant to request additional activities such as searches of patient registers or sending out SMS text messages to eligible people:

> *We could definitely do more around the promotion of NHS Healthy Living programme [but] we always play that fine line...we don't want to overwhelm primary care any more because they are under a lot of pressure. They have been under pressure through the pandemic, they are under the winter pressures now with COVID and flu vaccination programmes.* [Local lead 14; site 15]

> *When you talk to PCN managers or the practices themselves, they've got a lot going on and when you try and talk to them about DESMOND...there's potentially still a drawback...because they've got other, you know, the winter pressures, the COVID...so then also trying to fit in or send them resources about Healthy Living...it's like "well, we haven't got capacity at the moment to do that."* [Local lead 04; site 5 (early engagement area)]

National rollout occurred in May 2022, which was also a time of organizational change for the NHS as ICSs were in the process of being formed. Some local leads described how their focus was on conducting scoping exercises to identify diabetes offers across localities within their ICSs and working toward creating a more consistent and equitable offer of T2DM services across their localities. This suggested a lack of capacity to concentrate efforts on implementing Healthy Living beyond conducting light-touch implementation activities:

> *And then we are also doing lots of work around diabetes recovery at the moment. We are doing lots around recovery of the 8 care processes and the three treatment targets. So, trying to recover back from pre-pandemic levels.* [Local lead 14; site 15]

There was a sense among local leads that the national rollout of Healthy Living happened with little fanfare in comparison to other national programs, and they felt that regional NHS England leads had not been proactive in providing information about Healthy Living:

> *I'm not swimming in knowledge around Healthy Living, nor has it felt like there's been a tap on the shoulder around why are you not [implementing it].* [Local lead 06; site 10]

Discussion

Principal Findings

Most sites were implementing Healthy Living across their ICS, but this was limited to raising awareness of Healthy Living via one-way communications to general practices. Thus, a more passive process of dissemination (as opposed to active

XSL•FO
RenderX

implementation) was used by most sites. Early engagement areas received funds from NHS England to support implementation, and these were used to conduct more intensive implementation activities, with 50% (2/4) of the early engagement areas choosing to provide general practices with financial incentives for undertaking referral tasks. Local leads were generally positive about Healthy Living as an additional offer to existing service provision, although concern was raised that national programs such as Healthy Living had the potential to increase health inequalities rather than widen access to structured education because they were felt not to be geared toward patient cohorts with the greatest need. While most early engagement areas were monitoring uptake of Healthy Living across their area and making attempts to widen access to Healthy Living, no sites implementing Healthy Living following national rollout monitored general practice referrals. Contextual factors such as the absence of data received on the uptake of the program, no formal requirements being placed on ICSs to implement the program, and the timing of national rollout impeded more prolonged, targeted implementation activities to support rollout of Healthy Living across most sites.

Limitations

We initially aimed to analyze the interviews through the lens of NPT [18] as a more mandated and active implementation process was anticipated. However, expectations placed upon sites to implement Healthy Living changed from what we initially envisaged when we designed the study. In addition, the national rollout of the Healthy Living program removed the onboarding support provided by general practice clinicians in the HeLP Diabetes service, thereby removing any direct patient engagement with health professionals in the Healthy Living program [21]. Therefore, general practice implementation efforts were reduced to signposting eligible patients to Healthy Living so that patients could self-register for the program. As NPT is focused on the concept of "collective action," or what people do to implement new practices and ways of working, its use was less relevant to a process of passive dissemination [17]. Instead, we deployed a thematic analysis approach using the constant comparative method to systematically interrogate the data.

We aimed to recruit at least one local lead from every ICS in England (N=42) but recruited a total of 17 local leads across 16 different sites (16/42, 38% response rate). Therefore, the findings may be limited by nonresponder or "sampling bias," that is, the underrepresentation of sections of the population under study [27] Gaining views from a convenience sample of those who agreed to the interviews may reflect different views from those of sites that did not take part. Nevertheless, sites varied in terms of geographical areas of England, including rural and urban locations, and included sites that had adopted Healthy Living in the early rollout phase, those that had implemented Healthy Living following national rollout in May 2022, and those yet to implement the program. In addition, our sample of 16 ICSs was spread evenly across the range of deprivation seen among all ICSs in England [28], with 6 (38%) in the most deprived areas, 4 (25%) in the middle, and 6 (38%) in the least deprived areas. The proportion of the population that was of people of color ethnicity in our sample of 16 ICSs ranged from 3% to 54%, with a mean of 17%. This is the same as the national mean for the proportion of non-White populations among all ICSs [29]. Hence, we are satisfied that the sample provided an adequate mix of sites to evaluate the implementation of Healthy Living. In addition, while the number of interviews was lower than anticipated, we are satisfied that, in rigorously collecting and analyzing interview data, we were able to represent the "full dimensionality of [participant]'s experiences" [30]. A recent systematic review confirmed that qualitative studies can reach data saturation with sample sizes between 9 and 17 interviews [31].

Findings are also limited by the cross-sectional design as we undertook interviews at 1 time point; therefore, interviews report a "snapshot" in time. It would be useful for future evaluations to consider conducting interviews across different time points to investigate changes in implementation over time.

Comparison With Prior Work

We identified that the main barrier to prolonged or targeted implementation campaigns was the absence of data on uptake, which meant that local leads often did not go beyond raising awareness of Healthy Living across their localities. An updated systematic review of systematic reviews on factors that influence the implementation of eHealth technologies recommended that implementation should not stop with "go-live" and identified the need for continued monitoring and evaluation of eHealth technologies to ensure that intended goals were being met, benefits were realized, and barriers to use were understood [19].

Another barrier to implementation was confusion in understanding where Healthy Living fit within existing T2DM care pathways, with some indicating that Healthy Living should be offered to people only after attending or declining to attend a locally commissioned face-to-face T2DM structured education program. In contrast to local leads' views, qualitative research from the Healthy Living Diabetes-Long-Term Independent National Evaluation study exploring patients' experiences of using Healthy Living [32] identified that participants felt Healthy Living to be most suitable for newly diagnosed patients, perceiving the program to provide a reliable source of introductory information about T2DM. Some suggested that Healthy Living could be offered alongside face-to-face programs, which provide additional health care professional and peer support. In addition, most participants in the study had taken part in structured education programs but still learned new information from Healthy Living. They especially valued the emotional management content in Healthy Living as they had not been made aware of the emotional impact of T2DM previously, including when attending face-to-face T2DM structured education courses [32].

Research from the Healthier You NHS Diabetes Prevention Programme (DPP) showed that patients registered with practices identified as providing lower-quality care were less likely to be referred to the NHS DPP, leaving such patients disadvantaged, concluding that using primary care for referral into the NHS DPP had the potential to further increase health inequalities [33]. This study highlights that sites struggled to engage general practices in the most deprived areas to undertake referrals to Healthy Living and that some of the early engagement areas

XSL•FO
RenderX

were planning future engagement work with those practices but there were no similar plans in other sites. This highlights the importance of monitoring where referrals are being initiated from and providing support to practices with lower referral rates [33].

We found evidence that money allocated to the implementation of Healthy Living in the early engagement areas was used, in whole or in part, to provide general practices with "process based" financial incentives (ie, payments received for undertaking referral tasks). Payment for performance, or the use of financial incentives, in health care systems worldwide is increasingly used to try to improve quality and efficiency [34]. Previous research suggests that outcome-based incentives (ie, payments directly linked to referral volumes) are effective in increasing referrals but process incentives, such as those used in early engagement areas, are not [35].

Implications for Future Practice and Policy

Healthy Living is a self-contained, self-management digital health intervention without support for onboarding, health coaches, or any health professional support. Local leads responsible for implementing Healthy Living across their area perceived Healthy Living to be more suitable for patients who were self-motivated to learn and those who were more highly educated. This perception contributed to them not carrying out targeted implementation campaigns for Healthy Living. Healthy Living was implemented into an already "crowded market," with a number of other local and national programs competing for attention (such as locally commissioned T2DM structured education or lifestyle programs, the Digital Weight Management Programme [26], and the Path to Remission program [36]). There is a need for clearer communications regarding the type of patients who may benefit from the Healthy Living program, when it should be offered, and whether it should be offered instead of or in addition to other programs.

As reported, none of the sites other than early engagement areas received data to monitor uptake across their ICSs. Understanding variability in uptake between practices may have allowed sites to plan targeted referral campaigns in those practices that were not using the service.

Conclusions

This study describes the impact of shifting from an active to passive strategy for the implementation of Healthy Living. Passive dissemination, whereby communications were sent to general practices to raise awareness of Healthy Living, were favored over more active implementation processes. This was due mainly to wider contextual factors but the strategy of awareness raising, while necessary, is unlikely to drive uptake of and engagement with the program in general practice, if used alone.

Acknowledgments

The authors thank National Health Service England for facilitating this research, including assisting with recruitment of local leads for interviews. The authors are grateful to the local leads who took time to participate in the interviews. The authors wish to thank the following researchers in the Healthy Living Diabetes-Long-Term Independent National Evaluation team for providing valuable feedback during manuscript development: Dr Rhiannon Hawkes, Dr Brian McMillan, Dr Martin Rutter, and Mr Mike Spence. The authors also thank Dr Luke Munford for his advice on accessing Index of Multiple Deprivation and ethnicity data. This study is funded by the National Institute for Health and Care Research (Policy Research Programme, Healthy Living Diabetes-Long-Term Independent National Evaluation; NIHR200933). The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. Generative artificial intelligence was not used in any portion of manuscript writing.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SC and PW conceived the study, and SC secured funding. PW supervised the research. LB carried out data collection. LB and PW undertook data analysis. LB wrote the first draft of the manuscript, and PW and SC made substantive revisions to the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topic guide.
[DOC File , 54 KB - jmir_v26i1e55546_app1.doc]

References

XSL•FO
RenderX

1. Diabetes is serious. Diabetes UK. 2023. URL: https://www.diabetes.org.uk/support-us/campaign/diabetes-is-serious [accessed 2024-04-29]
2. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
3. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412 [FREE Full text] [Medline: 10938048]
4. Hex N, Bartlett C, Wright D, Taylor M, Varley D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. Diabet Med 2012 Jul;29(7):855-862. [doi: 10.1111/j.1464-5491.2012.03698.x] [Medline: 22537247]
5. Hewings R. The cost of diabetes report. Diabetes UK. 2014 May. URL: https://www.researchgate.net/publication/315474818_The_cost_of_diabetes_-_report_from_Diabetes_UK [accessed 2024-09-04]
6. Odgers-Jewell K, Ball LE, Kelly JT, Isenring EA, Reidlinger DP, Thomas R. Effectiveness of group-based self-management education for individuals with type 2 diabetes: a systematic review with meta-analyses and meta-regression. Diabet Med 2017 Aug;34(8):1027-1039 [FREE Full text] [doi: 10.1111/dme.13340] [Medline: 28226200]
7. Type 2 diabetes in adults: management. National Institute for Health and Care Excellence. 2015. URL: https://www.nice.org.uk/guidance/ng28 [accessed 2023-08-06]
8. Deakin TA, Cade JE, Williams R, Greenwood DC. Structured patient education: the diabetes X-PERT Programme makes a difference. Diabet Med 2006 Sep;23(9):944-954. [doi: 10.1111/j.1464-5491.2006.01906.x] [Medline: 16922700]
9. Davies MJ, Heller S, Skinner TC, Campbell MJ, Carey ME, Cradock S, Diabetes Education and Self Management for Ongoing and Newly Diagnosed Collaborative. Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomised controlled trial. BMJ 2008 Mar 1;336(7642):491-495 [FREE Full text] [doi: 10.1136/bmj.39474.922025.BE] [Medline: 18276664]
10. National diabetes audit core report 1: care processes and treatment targets 2012-22, underlying data. National Health Service Digital. URL: https://digital.nhs.uk/data-and-information/publications/statistical/national-diabetes-audit/nda-core-21-22/nda-core-21-22-data [accessed 2023-06-29]
11. Horigan G, Davies M, Findlay-White F, Chaney D, Coates V. Reasons why patients referred to diabetes education programmes choose not to attend: a systematic review. Diabet Med 2017 Jan;34(1):14-26 [FREE Full text] [doi: 10.1111/dme.13120] [Medline: 26996982]
12. The NHS long term plan. National Health Service England. 2019. URL: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/ [accessed 2024-04-29]
13. Murray E, Dack C, Barnard M, Farmer A, Li J, Michie S, et al. HeLP-Diabetes: randomised controlled trial protocol. BMC Health Serv Res 2015 Dec 29;15:578 [FREE Full text] [doi: 10.1186/s12913-015-1246-9] [Medline: 26715038]
14. Healthy living for people with type 2 diabetes. National Health Service England. URL: https://www.england.nhs.uk/diabetes/digital-innovations-to-support-diabetes-outcomes/healthy-living/ [accessed 2023-06-29]
15. Murray E, Sweeting M, Dack C, Pal K, Modrow K, Hudda M, et al. Web-based self-management support for people with type 2 diabetes (HeLP-Diabetes): randomised controlled trial in English primary care. BMJ Open 2017 Sep 27;7(9):e016009 [FREE Full text] [doi: 10.1136/bmjopen-2017-016009] [Medline: 28954789]
16. Li J, Parrott S, Sweeting M, Farmer A, Ross J, Dack C, et al. Cost-effectiveness of facilitated access to a self-management website, compared to usual care, for patients with type 2 diabetes (HeLP-diabetes): randomized controlled trial. J Med Internet Res 2018 Jun 08;20(6):e201 [FREE Full text] [doi: 10.2196/jmir.9256] [Medline: 29884608]
17. Ross J, Stevenson F, Dack C, Pal K, May C, Michie S, et al. Developing an implementation strategy for a digital health intervention: an example in routine healthcare. BMC Health Serv Res 2018 Oct 19;18(1):794 [FREE Full text] [doi: 10.1186/s12913-018-3615-7] [Medline: 30340639]
18. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554. [doi: 10.1177/0038038509103208]
19. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
20. Nelson LA, Williamson SE, Nigg A, Martinez W. Implementation of technology-delivered diabetes self-care interventions in clinical care: a narrative review. Curr Diab Rep 2020 Nov 18;20(12):71 [FREE Full text] [doi: 10.1007/s11892-020-01356-2] [Medline: 33206241]
21. Benton JS, Cotterill S, Hawkes RE, Miles LM, French DP. Changes in a digital type 2 diabetes self-management intervention during national rollout: mixed methods study of fidelity. J Med Internet Res 2022 Dec 07;24(12):e39483 [FREE Full text] [doi: 10.2196/39483] [Medline: 36476723]
22. Integrated care systems explained: making sense of systems, places and neighbourhoods. The Kings Fund. 2022. URL: https://tinyurl.com/3mujb3ja [accessed 2024-04-29]

XSL•FO
RenderX

23. Healthy living diabetes - long-term independent national evaluation (HED-LINE). National Institute for Health and Care Research (NIHR). URL: https://arc-gm.nihr.ac.uk/projects/HEDLINE [accessed 2023-06-29]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. Green J, Thorogood N. Qualitative Methods for Health Research. London, UK: Sage Publications; 2009.
26. The NHS digital weight management programme. National Health Service England. URL: https://www.england.nhs.uk/digital-weight-management/ [accessed 2024-04-29]
27. Polit DF, Beck CT. Essentials of Nursing Research: Appraising Evidence for Nursing Practice. Philadelphia, PA: Wolters Kluwer; 2022.
28. Dunn P, Fraser C, Williamson S, Alderwick H. Integrated care systems: what do they look like? The Health Foundation. URL: https://www.health.org.uk/sites/default/files/pdf/2022-07/2022%20-%20ICS%20characteristics_v3.pdf [accessed 2024-04-29]
29. Ethnic group by sex and age. Office for National Statistics. URL: https://www.ons.gov.uk/datasets/RM032/editions/2021/versions/1 [accessed 2024-04-29]
30. Young DS, Casey E. An examination of the sufficiency of small qualitative samples. Soc Work Res 2018;43(1):53-58. [doi: 10.1093/swr/svy026]
31. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
32. Hawkes R, Benton J, Cotterill S, Sanders C, French D. Service users' experiences of a nationwide digital type 2 diabetes self-management intervention (Healthy Living): qualitative interview study. JMIR Diabetes 2024 Jul 18;9:e56276 [FREE Full text] [doi: 10.2196/56276] [Medline: 39024002]
33. Parkinson B, McManus E, Sutton M, Meacock R. Does recruiting patients to diabetes prevention programmes via primary care reinforce existing inequalities in care provision between general practices? A retrospective observational study. BMJ Qual Saf 2023 May 06;32(5):274-285 [FREE Full text] [doi: 10.1136/bmjqs-2022-014983] [Medline: 36597995]
34. Eijkenaar F, Emmert M, Scheppach M, Schöffski O. Effects of pay for performance in health care: a systematic review of systematic reviews. Health Policy 2013 May;110(2-3):115-130. [doi: 10.1016/j.healthpol.2013.01.008] [Medline: 23380190]
35. McManus E, Elliott J, Meacock R, Wilson P, Gellatly J, Sutton M. The effects of structure, process and outcome incentives on primary care referrals to a national prevention programme. Health Econ 2021 Jun;30(6):1393-1416. [doi: 10.1002/hec.4262] [Medline: 33786914]
36. NHS type 2 diabetes path to remission programme. National Health Service England. URL: https://www.england.nhs.uk/diabetes/treatment-care/diabetes-remission/ [accessed 2024-04-29]

Abbreviations

DPP: Diabetes Prevention Programme
DESMOND: Diabetes Education and Self-Management for Ongoing and Newly Diagnosed
HeLP Diabetes: Healthy Living for People With Type 2 Diabetes
ICS: integrated care system
NHS: National Health Service
NPT: normalization process theory
T2DM: type 2 diabetes mellitus
UCL: University College London

Edited by T de Azevedo Cardoso; submitted 15.12.23; peer-reviewed by L Aljerf, K Fitzner; comments to author 26.02.24; revised version received 12.04.24; accepted 29.07.24; published 25.09.24.

Please cite as:
Brunton L, Cotterill S, Wilson P
Evaluating the National Rollout of a Type 2 Diabetes Self-Management Intervention: Qualitative Interview Study With Local National Health Service Leads Responsible for Implementation
J Med Internet Res 2024;26:e55546
URL: https://www.jmir.org/2024/1/e55546
doi:10.2196/55546
PMID:

©Lisa Brunton, Sarah Cotterill, Paul Wilson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the "Zindagi Mehfooz" Electronic Immunization Registry and Suite of Digital Health Interventions to Improve the Coverage and Timeliness of Immunization Services in Sindh, Pakistan: Mixed Methods Study

Patricia Mechael[1,2], MHS, PhD; Sara Gilani[3], MSc; Ahsan Ahmad[4], MBBS, MSc, PhD; Amnesty LeFevre[5], PhD; Diwakar Mohan[2], MPH, DrPH; Asra Memon[3], MSc; Mubarak Taighoon Shah[6], MPhil, PharmD; Danya Arif Siddiqi[6], MSc; Subhash Chandir[2,6,7], MBBS, MPH, PhD; Riswana Soundardjee[8], BSN, MSc

[1]HealthEnabled, Washington, DC, United States
[2]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[3]Gallup Pakistan, Karachi, Pakistan
[4]Independent Consultant, Islamabad, Pakistan
[5]University of Cape Town, Cape Town, South Africa
[6]IRD Global, Karachi, Pakistan
[7]IRD Pakistan, Karachi, Pakistan
[8]Gavi, The Vaccine Alliance, Geneva, Switzerland

Corresponding Author:
Patricia Mechael, MHS, PhD
HealthEnabled
3147 Tennyson Street Northwest
Washington, DC, 20015
United States
Phone: 1 9732228252
Email: pmechael@gmail.com

Abstract

Background: The Zindagi Mehfooz (safe life; ZM) electronic immunization registry (EIR) is a comprehensive suite of digital health interventions that aims to improve equitable access, timeliness, and coverage of child immunizations through a smartphone-based app for vaccinators, web-based dashboards for supervisors and managers, text message alerts and reminders for caregivers, and a call center. It has been implemented at scale in Sindh Province, Pakistan.

Objective: This study aimed to present findings from an evaluation of the ZM-EIR suite of digital health interventions in order to improve data availability and use as a contribution, among other immunization program interventions, to enhanced immunization outcomes for children aged 12-23 months in Sindh Province.

Methods: The mixed methods study included (1) analysis of ZM-EIR system data to identify high-, moderate-, and low-adoption and compliance sites; (2) in-depth interviews with caregivers, vaccinators, supervisors, and managers in the Expanded Program for Immunization (EPI); and (3) pre-post outcome evaluation using vaccine coverage from the Multiple Indicator Cluster Surveys (MICS) 2014 and 2018-2019. Key outcomes of interest were improved data availability, use and contribution to immunization outcomes, including receipt of individual antigens (Bacillus Calmette-Guérin [BCG], pentavalent [Penta] 1-3, measles), full immunization (all antigens), and zero-dose children defined as children aged 6-23 months who have not received the first dosage of the diphtheria-pertussis-tetanus 1/Penta vaccine.

Results: By registering newborns, providing alerts and reminders, and tracking their immunization completion, the ZM-EIR improved data availability and use in the EPI. The ZM-EIR was well received by EPI administrators, supervisors, vaccinators, and caregivers. The key benefit highlighted by ZM-EIR users was a list of children who missed scheduled vaccines (defaulters). Through greater availability and use of data, the ZM-EIR implementation, as part of a broader package of immunization program–strengthening activities in Sindh Province, may have contributed to an increase in immunization coverage and timeliness for BCG vaccinations and a decrease in zero-dose children in 2018-2019 from 2014. Additional findings from the study included

XSL•FO
RenderX

the dual burden of reporting on paper and gender-related considerations of female caregivers not wanting to provide their phone numbers to male vaccinators, creating barriers to greater uptake of the ZM-EIR.

Conclusions: The ZM-EIR is a promising technology platform that has increased the availability and use of immunization data, which may have contributed, along with other intensive immunization program interventions, to improvements in immunization outcomes through systematic registration of children, alerts and reminders, and increased use of data for planning and monitoring by the EPI.

Trial Registration: ISRCTN Registry ISRCTN23078223; https://doi.org/10.1186/ISRCTN23078223

(J Med Internet Res 2024;26:e52792) doi:10.2196/52792

KEYWORDS

electronic immunization registry; registry; Zindagi Mehfooz; vaccination; alert; reminder; dashboard; survey; cost; economic; digital health; immunization; children; pediatrics; equity; accessibility; text messages; SMS; zero dose

Introduction

The *Zindagi Mehfooz* (*safe life*; ZM) electronic immunization registry (EIR) is a comprehensive suite of digital health interventions that aims to improve equitable access, timeliness, and coverage of child immunizations through a smartphone-based app for vaccinators, web-based dashboards for supervisors and managers, text message alerts and reminders for caregivers, and a call center. It has been implemented at scale in Sindh Province, Pakistan.

Electronic immunization registries (EIRs) are digitized case-based immunization record–keeping systems designed to improve the collection, analysis, and use of immunization program data to improve the efficiency, equity, and coverage of immunization programs [1]. A 2015 systematic review [2] of aggregate immunization information systems in high-income settings highlighted the main pathways through which such systems improve immunization outcomes, such as coverage, although only 1 study in Australia measured such an association. The benefits of aggregate systems align with more recent reviews and studies of EIRs that record and track data for individual children and generate aggregate data for immunization program monitoring and planning [3]. These benefits include client reminders, client vaccination status, missed vaccinations, and vaccination program management and accountability [2,3]. Improved access to better data can support evidence-based decisions, planning, and operations for the immunization program, which, in turn, will lead to better delivery of immunization services [4]. When EIR data are triangulated with other primary health care services, recorded births, and population denominators, EIRs can enable population-level analysis to identify areas where the number of children vaccinated is lower than the target population [1,5]. In this way, EIRs can improve the identification and reach of zero-dose children (ie, children who have not received a diphtheria-pertussis-tetanus [DPT] 1 vaccine) and provide targeted service improvements and extended outreach services to those areas and communities. EIRs can also address demand-side challenges by sending automated reminders to caregivers, providing access to educational messages and service-related updates and improving the quality of services at vaccination centers [5].

Because EIRs are intended to support immunization programs with improved data-driven decision-making, the evaluation of data use is also important [6]. Drivers of EIR use include organizational factors (supervision, support, resources, facility type, location, and volume of clients), technical factors (infrastructure, electricity, and internet connectivity), and behavioral factors (training, capacity, and motivation of system users) [6]. A recent study [7] using EIR data to evaluate the timeliness of vaccination systematically found mixed results, with the conclusion that the overall reliance of the study on system-level data was insufficient to evaluate its impact on immunization outcomes. A review of the evidence for EIRs reveals a moderate number of process evaluations, few pilot studies, and no large-scale effectiveness studies of EIRs [3].

The ZM-EIR comprises complementary digital health interventions that satisfy and go beyond the conventional definition of an EIR. The ZM-EIR aims to strengthen Expanded Program for Immunization (EPI) service delivery and the broader health system by addressing both supply- and demand-side barriers that contribute to suboptimal use of immunization services. On the supply side, it is designed to enable health workers to track the immunization history for individual children and assess catch-up immunization schedules, support motivation among health workers by reducing the excessive burden of paperwork and demonstrating the value that quality data can contribute to their workflow [4], and reduce the occurrence of poor reporting, management, supervision, and monitoring of vaccinators. On the demand side, the ZM-EIR is designed to overcome challenges associated with low uptake of immunization services due to a lack of awareness and motivation among parents, as well as the inability to remember vaccine appointments.

Pakistan represents 1 of the 6 countries in the world with the highest concentration of zero-dose children. In 2012-2013, only about half (58.3%) of children aged 12-23 months in Pakistan received all basic vaccinations [8], according to the Pakistan Demographic and Health Survey 2012-2013, with marked geographic variation across the country. Sindh Province recorded full immunization coverage of 35% [9] in 2014, according to the Multiple Indicator Cluster Survey (MICS) conducted in 2014. Sindh Province is home to 47.9 million people and comprised 28 districts during the time frame covered by the evaluation, with around 1500 fixed immunization centers that conduct outreach activities, as well as over 200 private

XSL•FO
RenderX

clinics [10]. The provincial capital, Karachi, is Pakistan's largest capital city and economic hub representing a large urban population. Most childhood vaccines in the district are delivered through public sector immunization centers or outreach activities managed by the national EPI [10].

As of August 2022, the ZM-EIR was scaled to cover all of Sindh Province, including 2 new districts that were not covered during the evaluation period. The ZM-EIR has also been introduced in other areas of Pakistan, including the Federal Capital (Islamabad), Khyber Pakhtunkhwa Province, and the Gilgit-Baltistan region, and has been used by over 5000 vaccinators at public and private immunization clinics to enroll over 7 million children.

This paper aimed to summarize findings from a mixed methods evaluation of ZM-EIR implementation in Sindh Province. Retrospective evaluations of program implementation are difficult to design and carry out due to the lack of counterfactuals and recall biases. Habicht et al [11] proposed "adequacy" as the minimum level of evidence in support of a program's impact part of an evaluation. We used a range of data sources to support the hypothesis that the ZM-EIR intervention may have contributed to the outcomes achieved by the immunization program. We showed that ZM intervention activities were implemented adequately, accepted (provider and caregiver perceptions of the ZM-EIR), and adopted (vaccinator use of the ZM-EIR) and may have contributed to improved coverage (changes in antigen coverage) and timeliness.

Methods

Intervention Description

The ZM-EIR includes the following functionalities:

- Web interface for data visualization
- Mobile-based data entry and access for vaccinators
- Geospatial data for vaccination events
- 2-way interactive SMS reminders for parents/caregivers
- Call center to answer queries of parents/caregivers
- Offline mode to work in areas of poor connectivity
- Child registry for enrolling newborns or never-vaccinated children
- Immunization decision support system (iDSS) to guide vaccinators on routine and catch-up immunizations
- Defaulter reports for vaccinators
- Gamified videos for training vaccinators
- Artificial intelligence (AI)-based predictive analytics to identify children highly likely to drop out
- AI-based chatbot to address caregiver queries and concerns in real time
- Global System for Mobile Communications (GSM)-based tracking of vaccinators during work hours
- Quick response (QR) code–based tracking for unique identification of children and women

These functionalities center around a mobile phone–based registry with a suite of complementary interventions. More details and underlying research that have contributed to the development of the intervention over the past decade have been published elsewhere [12-18]. See Figure 1 for a graphical overview of the ZM-EIR intervention. Although the intervention can capture the immunization of pregnant women, this research focused on its use for routine immunization of children.

Figure 1. Overview of the ZM system. EIR: electronic immunization registry; EPI: Expanded Program for Immunization; iDSS: immunization decision support system; QR: quick response; ZM: Zindagi Mehfooz (safe life).

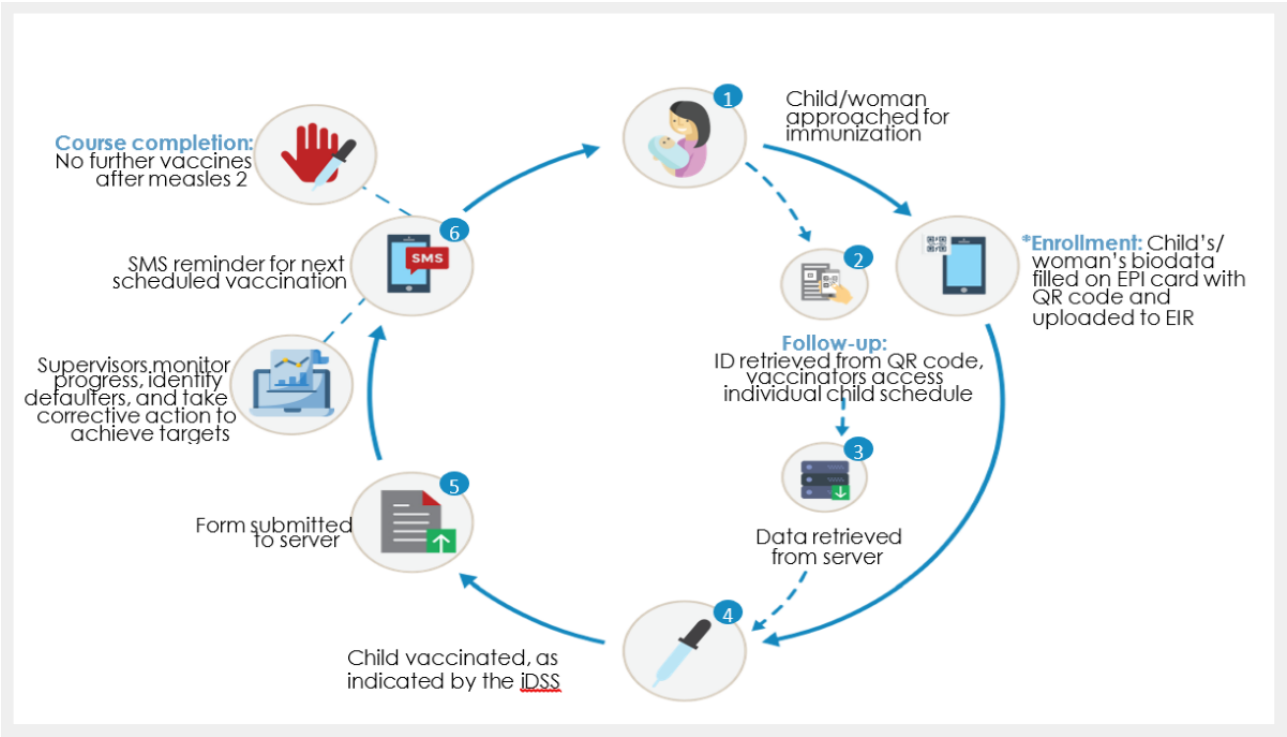

XSL•FO
RenderX

The ZM mobile Android app is used to capture data for each point of contact between the immunization service, caregivers, and their children. At the time of first enrollment, the child's data are entered on the paper EPI card and into the ZM-EIR. A unique QR code is printed and affixed to the home-based paper record and linked to the electronic system. During follow-up visits, the child's ID is retrieved by scanning the QR code, giving the vaccinator access to the child's vaccine schedule. For each vaccination received, the data are entered into the ZM app and linked to the server. All data entered into the ZM app are collated on a central web-based dashboard, which supervisors can use to monitor progress, identify defaulters, and take corrective action to achieve targets. The decision support system (DSS) uses a built-in algorithm to automatically calculate which vaccine is to be administered based on the child's date of birth and immunization history. This minimizes the chance of missed immunization and enables the vaccinators to provide age-appropriate and timely immunization. In addition to recording vaccination activities within the ZM intervention, vaccinators were required to document their work in paper-based registers that were then manually tabulated and submitted as part of monthly reporting activities.

To remind caregivers of when they need to next visit the clinic, up to 3 automatic personalized SMS reminders are sent 1 day before, on the day of, and 6 days after the scheduled immunization date (if the child did not make the scheduled appointment). A child is considered to have completed the full course of immunizations after the second dose of the measles vaccine.

In addition to the inclusion of standard EIR features (as illustrated in Table 1) [3], the ZM-EIR includes AI-based predictive analytics to identify children more likely to drop out; automated SMS reminders; a call center; and an AI chatbot for caregiver questions, concerns, and complaints. The suite of interventions also includes geospatial data for vaccination events, geospatial data to track each child's immunization progress, real-time tracking of vaccinators with Geographic Information System (GIS) tracking to improve monitoring and accountability, simultaneous supervision of multiple teams, monitoring of attendance and mobility of field staff, and gamified videos and training for vaccinators.

Table 1. Comparison of ZM[a] and ideal EIR[b] characteristics and functionalities.

Feature	ZM-EIR	Ideal requirements of an EIR [19]
Registration and search	The ZM-EIR is a registry to *enroll all newborns* or never-vaccinated children with a *unique ID/QR*[c]*code* assigned to each woman and child; a web interface *allows data visualization*, *mobile-based data entry*, and *access for vaccinators*.	Enrollment at birth, client management, unique and unequivocal ID
Patient records	Registration includes all relevant *demographic and location data of the patient* and facility with *geolocation*, *barcoded vaccine vials* linked with vaccine records, and an *offline mode* for areas with low mobile/data coverage (*not linked to other health areas*).	Individual demographic data, vaccine event data, nonroutine (campaign and outreach) vaccine events, and linkages to other health areas
Vaccination monitoring and follow-up	The *iDSS*[d] guides vaccinators on routine and catch-up immunizations and provides *defaulter reports* for vaccinators and interactive *SMS reminders* for parents/caregivers.	Clinical decision support, identification of undervaccinated children/missed vaccination, and reminder and recall messages
Registry	*Vaccinators are registered* within the system and assigned to a *health facility registered* to a district management unit.	*Health facility registration and management*
Stock	The *vaccine stock is tracked* through the number and type of vaccinations given by vaccinator, facility, and district.	*Stock management*
Data and reporting	*Geospatial analysis* maps pockets of undervaccinated children with exact locations to enable follow-up with households refusing vaccination. There is a call center for complaints and adverse events following immunization *(AEFI) reporting.*	Data aggregation at different geographic and administrative levels, adverse event reporting
Other system requirements	The ZM-EIR follows industry standards for the *data security and privacy* platform *interoperable withDistrict Health Information System 2 (DHIS2) and the government's EPI*[e]*management information system (MIS)* deployed at *full scale* in all districts of Sindh Province by over 3096 vaccinators at 1694 public and private sector immunization clinics to enroll 3.8 million children and 1.3 million women, with 31 million immunization visits recorded (as of 2021).	Data exchange and interoperability, offline capability, alignment with international standards, data privacy and security, scalability and capacity, usability

[a]ZM: *Zindagi Mehfooz* (*safe life*).

[b]EIR: electronic immunization registry.

[e]QR: quick response.

[d]iDSS: immunization decision support system.

[e]EPI: Expanded Program for Immunization.

XSL•FO
RenderX

Study Design

The pre-post mixed methods study design measured changes in the proportion of fully immunized children aged 12-23 months from 2014 to 2019. Fully immunized children are those who have received the complete recommended schedule of 5 antigens by 23 months of age. The pentavalent (Penta) vaccine refers to a DPT-Hib-HepB (where Hib-HepB refers to the *Haemophilus* b–hepatitis B conjugate) combination in this context. Additional aims and objectives of the study were assessed through qualitative research (observation and in-depth interviews), secondary data analysis (MICS from 2014 and 2018-2019), and comparison between districts and health facilities with high, moderate, and low ZM-EIR compliance (based on data from the platform or app of log-ins and entries of child encounter information). The results are presented later based on the study aim and objective. Experimental or quasi-experimental study designs were not feasible as the ZM-EIR was implemented throughout Sindh Province and a suitable comparator was not available.

Patient and Public Involvement

There was no patient or public involvement in the design or implementation of the study. Caregivers, health workers, and health administrators were included in the qualitative research component as key informants after they provided informed consent.

Ethical Considerations

The study received Institutional Review Board (IRB) approval from the IRD (approval number IRD_IRB_2021_11_001) and was registered with the ISRCTN registry (#23078223). Informed consent was obtained from all key stakeholder groups prior to interviews.

Data Sources and Analysis

Quantitative Data Collection

The 2 sources of quantitative data were the MICS conducted across the state of Sindh, Pakistan, by the United Nations Children's Fund (UNICEF) from 2014 and 2018-2019 and the ZM registry database. Details of the MICS are provided on the UNICEF website. The indicators across the study aims and objectives are listed in Table 2.

XSL•FO
RenderX

Table 2. Indicators included in quantitative analyses.

Indicator	Definition	Numerator	Denominator	Data source
Compliance	The vaccinator updates/uploads 1 record per vaccination day.	Number of days the vaccinator logged in to the ZM[a] app and uploaded/updated at least 1 child record	Number of vaccination days for that health center/vaccination center	ZM database
ZM registry coverage	N/A[b]	Total number of children receiving any immunization reported in the ZM	Estimated monthly target population of children in Sindh Province derived by the EPI[c] in close collaboration with the ZM team	ZM child records database
ZM registry use	The proportion of vaccinators trained to use ZM upload details for at least 1 client per day in Sindh Province from 2017 to 2019.	Number of days the vaccinator logged in to the ZM app and uploaded/updated at least 1 child record	Number of vaccination days for that health center/vaccination center	ZM vaccination staff database
ZM dashboard use	This is the number of unique user (manager/supervisor) logins to the dashboard.	Number of log-ins per district per year	N/A	ZM management staff usage of dashboard derived from the ZM database
Fully immunized	If the child has received the BCG[d] vaccine at birth and OPV[e] 1-3, Penta[f] 1-3 vaccines, PCV[g] 1-3, and the measles vaccine 1 at 12-23 months, the child will be considered fully immunized.	Number of children 12-23 months old who received the BCG vaccine at birth, OPV 1-3, Penta 1-3 vaccines, PCV 1-3, and the measles vaccine 1	Sample of children 12-23 months old as part of the survey in Sindh Province	MICS[h] 2014, 2018-2019
Zero doses	A child aged 6-23 months has not received the first dose of the DPT[i] 1/Penta 1 vaccine.	Number of children who did not receive the DPT 1/Penta 1 vaccine	Sample of children 12-23 months old as part of the survey in Sindh Province	MICS 2014, 2018-2019
Timely immunization	A child has received the dose of antigen at the recommended age specified in the EPI schedule.	Number of children 12-23 months old who received the BCG vaccine at birth, OPV 1-3, Penta 1-3 vaccines, PCV 1-3, and the measles 1 vaccine as per the timeliness criteria of the Pakistan/Sindh Province EPI	Sample of children 12-23 months old as part of the survey in Sindh Province	MICS 2014, 2018-2019

[a]ZM: *Zindagi Mehfooz* (*safe life*).

[b]N/A: not applicable.

[c]EPI: Expanded Program for Immunization.

[d]BCG: Bacillus Calmette-Guérin.

[e]OPV: oral polio vaccine.

[f]Penta: pentavalent.

[g]PCV: pneumococcal conjugate vaccine.

[h]MICS: Multiple Indicator Cluster Survey.

[i]DPT: diphtheria-pertussis-tetanus.

The ZM platform uses multiple sources to register children, including birth cohort registration, the recording of pregnancies and births, the EPI's estimated targets, and the date of enrollment. The analysis reflected the vaccine-wise coverage for the age group of 12-23 months. The registry-based coverage was calculated using the enrollment numbers for each calendar year birth cohort as the numerator and the annual targets provided by the EPI (ascertained population based on modeling using birth rates and recent census figures). Furthermore, the target population of children was derived based on the enrollment of children at the time of birth through a birth cohort registration and an EPI schedule–based continuum-of-care mechanism. The numerators were based on aggregate numbers for the different antigens and did not distinguish the age of the child for the receipt of vaccination, which is reflective of immunization data from administrative sources.

We analyzed MICS data using the design considerations for weighting, adjustment of clustering at the level of primary sample units, and strata for survey errors according to guidelines provided by the MICS. We presented survey design–adjusted estimates of vaccine coverage with 95% CIs.

XSL•FO
RenderX

Timeliness was calculated according to the following recommended age schedule. It is worth noting that timeliness for the Penta 3 vaccine does not correct for any minimal interval with previous doses, which makes being untimely almost normal if Penta 1 vaccination is already late (Table 3).

Table 3. Timeliness schedule of vaccination.

Antigen and dose	Minimum age	Recommended age	Timely schedule (days)
BCG[a]	Birth	0 weeks/0 days	0-28
Penta[b] 1/OPV[c] 1/PCV[d] 1	6 weeks	6 weeks/42 days	39-70
Penta 3/OPV 3/PCV 3	14 weeks	14 weeks/98 days	95-126
Measles 1	6 months	9 months/274 days	180-302

[a]BCG: Bacillus Calmette-Guérin

[b]Penta: pentavalent.

[c]OPV: oral polio vaccine.

[d]PCV: pneumococcal conjugate vaccine.

Qualitative Data Collection

Qualitative data collection was conducted in 3 districts in Sindh Province, selected for the high, medium, and low immunization coverage rates according to the analysis of ZM-EIR system data. Direct observations of immunization services were conducted for 1 day per facility in 6 health facilities and 3 district offices. Within each district, 2 health facilities were selected based on high and low ZM-EIR compliance. At each health facility, 2 vaccinators and 2 caregivers (1 male and 1 female) who were attending the facility for their child's immunization were interviewed. Selection was based on their availability and consent to participate during field visit days. For each district, the managing supervisor and the district field coordinators and EPI managers were also interviewed to capture views of the leadership, along with vaccinators and clients. A structured discussion guide was used to conduct the interviews, led by a researcher with experience in qualitative methods. Qualitative data were collected from caregivers of children <2 years old (n=12), vaccinators (n=12), supervisors (n=3), district-level EPI managers (n=7), provincial EPI managers (n=2), and ZM-EIR implementation and data teams (n=4). The interviews were audio-recorded and transcribed for qualitative analysis. Direct observations of immunization services were also conducted for 1 day per facility in the 6 health facilities and 3 district offices.

As guided by the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist, the research team comprised a Pakistani female principal investigator and health system researchers trained in qualitative methods, with master's degrees in public health and over 10 years of experience. There was no previous relationship with the study participants. District field supervisors facilitated contact with health facilities, and health facilities facilitated contact with caregivers. The sampling was purposive to ensure that all key stakeholder groups engaged in using or having data captured within the ZM-EIR: EPI managers, district health officers, EPI supervisors, vaccinators, and caregivers were included in the sample. Interviews were conducted face to face either in offices/places of work or in health facilities. There were no refusals, and no nonparticipants were present at the time of the interviews. Content analysis was conducted using a framework developed through the content captured within the in-depth interviews. Interview guides were developed and tested prior to data collection. There were no repeat interviews. Saturation was achieved through the structured sampling and interviews conducted. Interviews were approximately 30 minutes in length. They were audio-recorded and field notes captured. Microsoft Excel was used to organize data by theme for ease in the synthesis of findings and reporting. Thematic analysis was conducted using grounded theory through a 2-step process: (1) iterative debriefs conducted concurrently with data collection and (2) coding of transcripts and generation of subthemes and themes. A theoretical framework based on these themes and subthemes was then developed in line with the research study objectives and used to analyze and present findings.

Results

ZM-EIR Reach and Use by Providers

Vaccination staff training on the ZM-EIR started in quarter 3 of 2017 as part of a district-wide planning with the Sindh Provincial Department of Health. Data on the proportion of vaccination staff trained on the system from the total were not available. However, all the 1307 staff members nominated for training were trained across 52 batches between September and December 2017. Training batch sizes ranged from 12-40 participants per training session.

The Sindh Provincial EPI defines compliance with the ZM-EIR system by vaccinators as 1 upload or update for at least 1 client per vaccination day. Compliance estimates include instances in which a vaccinator may be present at the health facility but no child is brought in for vaccination. Findings suggested that across 27 districts that were active during the time frame covered by the study, compliance rates ranged from 46% to 85%. The average compliance increased from 40% in 2018 to 57% in 2019. Data are available in Multimedia Appendix 1.

Trends in compliance were assessed based on demographic characteristics of vaccinators. Compliance and usage during 2017-2019 were slightly higher among urban vaccination staff compared to their rural counterparts. Compliance among female vaccinators was higher compared to males, even though the proportion of females in the vaccination staff was only 15%.

XSL•FO
RenderX

There was a gradual increase in compliance in all age-groups during 2017 to 2019. Younger age groups (less than 40 years old) showed improvement from 40% in 2017 to 49% in 2019, with the highest proportional increase observed for those aged 26-30 years (49%) and those aged 18-25 years (47%). A more modest increase was found for vaccinators in higher age groups, with the 51-55-year group showing only a 28% increase.

Stakeholder Perceptions of the ZM-EIR

Stakeholder perceptions of the ZM-EIR were assessed for each set of functions within the ZM-EIR suite of digital health interventions, including the perceived benefits and challenges of use to improve immunization service delivery (Table 4). Overall, all stakeholder groups interviewed had a favorable impression of the ZM-EIR and the associated suite of interventions. Provincial managers felt that the real-time access to vaccination data facilitates longitudinal monitoring of vaccination coverage over time and across geographic areas. They further noted that expansion of the ZM-EIR would allow this to occur across the country as it would enable the provinces to see and evaluate each other's coverage. The ZM-EIR was thought to provide a blueprint of what a national repository of EPI records could look like and further be a mechanism for improving supply monitoring and fraud prevention. As the primary end users of the ZM-EIR, vaccinators expressed their satisfaction with ZM-EIR features, specifically the use of automatic lists generated of those who may have missed a scheduled vaccine dose or may not have received any vaccines (defaulter and zero-dose lists).

Table 4. Key informant general perceptions of the ZM-EIR[a].

Key informant	Example quotations
Caregiver	• "...It is very helpful. Otherwise, we are busy with our work and forget about it. When we receive a message, then we know we have to get the child vaccinated soon." • "The caregivers who used the call center, called it, 'good in every way'. According to them, these have been proved useful in gaining information about the availability of the vaccination teams, discussing any symptoms caused by the vaccination, and asking for a second dose."
Vaccinator supervision	• "...Wherever he [vaccinator] goes, we monitor it through Google Maps. When we visit a vaccinator, it is a surprise visit. We do not call them. We just follow them on mobile and visit them on field." • "These graphs depict which vaccinators have complied [based on required vaccinations for the day], and which are still noncompliant."
Manual (paper-based) data capture	• "Our people in the field write in the book, on the card, and all that writing is very difficult. It would be better if we just use ZM-EIR, and only 1 thing stays." • "If they had owned it properly, it would end. The vaccinators who work honestly in the field ask us to put an end to the manual data capture work so they can continue working with ZM only." • "All their registers will cease to be used, and all their data will be done on ZM. That will be their daily register and stock register. Work is being done on this, and I think in 1 or 2 months, they may become paper free."

[a]ZM-EIR: *Zindagi Mehfooz* (*safe life*) electronic immunization registry.

Registry Coverage

From 2017 to 2019, the ZM-EIR immunization registry enrolled nearly 5.4 million children and 1.9 million married women of reproductive age (15-49 years).

Gender-Related Barriers to Registry Uptake

Gender was identified to play a significant role in ZM-EIR uptake by caregivers owing to its embeddedness within the local, rural contexts. There is often a clear distinction between a caregiver's interaction with male vaccinators and that with female vaccinators. One key informant emphasized the need for female health workers who can bond with female caregivers (who are usually more responsible for taking care of their children than male caregivers are) to resolve any resistance or anxieties regarding vaccinations to further expand immunization efforts. A female caregiver shared:

> *The female health worker can come inside the house, and we [caregiver] know her, so it is good that a female health worker comes. If a man comes, then it is quite difficult.*

ZM-EIR Data Use

The provincial managers shared that the ZM-EIR intervention provides defaulter and zero-dose lists, which helps them meet their target and improve performance. It enables those who supervise to encourage the vaccinators to bridge the gaps in the coverage of immunization targets. The data are viewed as reliable due to the feature allowing the system to operate in both offline and online modalities, and the use of the platform is not dependent on the availability of internet services. These data are then communicated to other responsible representatives at the district and provincial levels. Among the stakeholders who can access the data reports and dashboards, the role of supervisors has been seen to be relatively limited compared to others. The supervisor's overall responsibilities include monitoring and providing support to vaccinators, checking the vaccination stock, and planning and conducting field visits. For their limited use of the ZM-EIR app to monitor the activities of vaccinators in their respective district areas, it was stated that newer and younger supervisors can use the app comfortably, while older supervisors find the app difficult and less engaging. Coordinators and supervisors also suggested that the latter be

XSL•FO
RenderX

given access to a separate app to monitor vaccinators so that they do not have to rely on district managers for the data.

Observation of Data Entry and Use

Among the health facilities observed, half were clean, organized, and relatively well maintained. The other half were seen as unclean, "cluttered," and "with broken furniture." The number of rooms in a health facility ranges from 2 to 7. All the health facilities provide vaccination and other health services. Most of the health facilities have 2 vaccinators present. Of these, 1 (50%) in each facility uses the ZM-EIR for registration and defaulter lists. In all 3 (50%) facilities, it was reported by vaccinators that when the caregivers would bring their children for vaccination, they did not use the ZM-EIR for registration in real time, but they recorded information in the paper-based register system and later entered it into the electronic app. Only 4 (67%) of the 6 facilities had electricity, 1 (25%) being backed by a solar power system. Network connectivity was limited in all facilities, which also hindered the use of the ZM-EIR in real time. All district offices observed had 1 supervisor on-site with access to a computer/laptop in the office, which enabled access to web-based ZM-EIR dashboards. The supervisors reported engaging in daily monitoring of the vaccinators, whereby they tracked the vaccinators' attendance, location, and registration. No other use of the ZM-EIR was reported or observed (Table 5).

Table 5. Key informant views of ZM-EIR[a] data use by key stakeholder group.

Key stakeholder group	Example quotations
District field coordinators and managers	• "When we have their attendance in the morning, the supervisors call right then to ask why this vaccinator isn't marked present, tell their timing, and all. So, we talk to them daily." • "I can also see everything about defaulter children and why we can't cover them."
Supervisors	• "[Supervisors] do monitoring by checking their vaccines, by checking their temperature in the monitoring. We check the graph and record to check whether they are taking records or not. We also check whether they [vaccinators] use ZM or not. This is our work." • "We can't tell if it is useful for the supervisors. Because we don't know where the vaccinators are. We should be able to check their work through it, even as a guest or monitors." • "All supervisors are old aged. If we make them use it, they won't be able to. They can't even see the dashboard, because those poor people aren't used to it since the start."
Vaccinators	• "The vaccinators are scared because when their defaulters increase, their UCa goes into categorization B or C. So, they are working on the coverage of defaulters, because with ZM, we are getting proper data now. Since the defaulters are being covered to keep their UCs up, that means vaccinations are happening timely."

[a]ZM-EIR: *Zindagi Mehfooz* (*safe life*) electronic immunization registry.

[b]UC: Union Council.

Vaccination Coverage and Timeliness

The vaccine-wise coverage across this evaluation's time frame of consideration (2017-2019) reflected that since its inception for scale-up in the province and phase-wise implementation in all districts, increases were observed for all vaccines (the EPI in Sindh Province had measles 1 as the last vaccine as per the child vaccination schedule till 2018. Subsequently, measles 2 was added as part of the EPI schedule; however, given the evaluation's considered time frame of 2017-2019, the quantitative analyses focused on measles 1). Findings suggested that improvements in coverage were achieved for BCG (+10.1%), Penta 1 (+8.8%), Penta 3 (+7.4%), and measles 1 (+5.5%) vaccines. Vaccination timeliness was assessed by antigen by comparing the proportion of children 12-23 months old immunized across MICSs in 2014 and 2018-2019. Findings suggested that improvements in timeliness were achieved for BCG (+10.0%), Penta 1 (+3.1%), Penta 3 (+2.0%), and measles 1 (+0.7%) vaccines. However, only the improvement in timeliness for the BCG vaccine was statistically significant (Table 6).

XSL•FO
RenderX

Table 6. Vaccination coverage and timeliness in Sindh Province between 2017-2018 and 2014 (MICS[a]).

Vaccination	Prevalence (95% CI)		Increase (%)
	2018-2019 (%)	2014 (%)	
BCG[b] received	80.2 (78.4-81.9)	70.1 (67.2-72.9)	10.1
Penta[c] 1 received	72.5 (70.3-74.6)	63.7 (60.7-66.7)	8.8
Penta 3 received	50.6 (48.2-53.1)	43.2 (40.0-46.4)	7.4
Measles 1 received	54.4 (52.1-56.8)	48.9 (45.9-51.9)	5.5
BCG timely	33.7 (31.6-35.7)	23.7 (21.0-26.4)	10.0
Penta 1 timely	22.7 (20.8-24.5)	19.5 (17.0-22.1)	3.1
Penta 3 timely	12.3 (10.9-13.8)	10.4 (8.4-12.3)	2.0
Measles 1 timely	11.7 (10.3-13.1)	11.0 (9.1-12.9)	0.7

[a]MICS: Multiple Indicator Cluster Survey.

[b]BCG: Bacillus Calmette-Guérin.

[c]Penta: pentavalent.

Additional analyses were conducted to compare increases in immunization coverage in Sindh Province, along with Punjab Province, which implemented a different EIR as part of its EPI. The analyses indicated increases in immunization coverage, making the increases in immunization coverage and timeliness difficult to attribute to the ZM intervention (Tables 7 and 8).

Table 7. Vaccination coverage in Sindh and Punjab Provinces between 2017-2018 and 2014 (MICS[a]).

Vaccination	Sindh Province			Punjab Province		
	2014 (%)	2018-2019 (%)	Increase (%)	2014 (%)	2017-2018 (%)	Increase (%)
BCG[c]	77.4	82.4	5.0	93.1	94.8	1.7
Penta[c] 1	70.7	75.1	4.4	85.7	92.7	7.0
Penta 3	55.3	57.6	2.3	73.3	87.3	14.0
Measles 1	58.6	62.7	4.1	77.8	82.7	4.9

[a]MICS: Multiple Indicator Cluster Survey.

[b]BCG: Bacillus Calmette-Guérin.

[c]Penta: pentavalent.

Table 8. Key informant views of the relationship of the ZM-EIR[a] to immunization coverage and equity.

Key informant	Example quotation
Immunization program performance	'…In what [performance] category are [2] UCs[b], that is, whether [they] are in red, green, blue, or yellow…every UC-wise, every vaccinator-wise, every town-wise, every district-wise, we are getting all the data."
Immunization coverage	"Because of ZM, the EPI[c] office has found out that we have 114 vaccinators. If only 67 vaccinators out of them are working, they ask your district where the rest of your vaccinators are. They ask us to enroll them, assign them a field, and send them to work. So now, I have over a hundred vaccinators enrolled, and they work themselves after returning. Coverage will thus ultimately increase."
Immunization equity	"In outreach, more female children are being vaccinated because they find it easy that they can get the immunization service at their doorstep…we can see [in defaulters list] that more females might be lagging behind; as a result of this, there is catch-up, and we do see better coverage among the female population."
Immunization timeliness	"…Three SMS reminders are sent so it is not that you only get one SMS reminder for the child's appointment day…there are these three different time points as a constant nudge to the caregiver to bring the child for immunization at the time that he is required to show up."

[a]ZM-EIR: *Zindagi Mehfooz* (*safe life*) electronic immunization registry.

[b]UC: Union Council.

[c]EPI: Expanded Program for Immunization.

XSL•FO
RenderX

From a ground-level perspective, vaccinators shared that the selection of local people for authority positions in rural governments has increased the trust of the people in the government, leading local caregivers to trust health authorities and their immunization efforts. In general, the vaccinators also believe that there has been an increase in awareness, coupled with an increase in diseases, which has had a concurrent effect on improved vaccination rates among rural communities.

Managers shared that the timeliness of immunization may be due to the DSS built within the app. Although SMS reminders as a ZM-EIR feature are more caregiver oriented, the DSS within the app integrated as an explicit function is more vaccinator oriented. In instances where vaccinators have not had substantive training, they are likely to get confused about which vaccine to give a child at a specific time.

Zero Doses

The prevalence of zero-dose children aged 6-23 months, defined as not receiving at least 1 dose of the Penta vaccine, went down from 36.3% to 28.5% (at 95% CI) between 2014 and 2019. Relative decreases in Punjab Province make attribution to the ZM intervention difficult to ascertain.

Discussion

Principal Findings

This external evaluation of the ZM-EIR and the associated suite of digital health interventions for childhood immunization in Sindh Province, Pakistan, highlights important insights and recommendations across stakeholder groups related to the benefits of an EIR, namely the availability and use of data through alerts and reminders to caregivers, identification of defaulters, supervision of vaccinators, and overall management of vaccination program activities [2].

Qualitative findings support this assertion; indicating that overall, the stakeholders interviewed had a favorable impression of the ZM-EIR suite of interventions and the perceived benefits of ZM defaulter lists on improved immunization coverage, as well as decreased percentage of zero-dose children. Several suggestions and areas for improvement were made by key informants to address some of the challenges experienced by users of the ZM-EIR.

First, gender dynamics, gender segregation (male vaccinators/female caregivers), education, literacy, and cultural norms impacted access and use by vaccinators and the ability to make direct contact with caregivers. Although their numbers are lower, female vaccinators have higher compliance with the ZM-EIR system than their male counterparts. Power, trust, and relationships were noted to present challenges between vaccinators and clients of the opposite gender. The implementation of the ZM-EIR brings to the fore the deep-rooted gender inequities in society that need to be addressed through relevant policy measures to leverage the full potential of the intervention. In addition, gender analyses at the start of future ZM-EIR intervention planning may also help identify strategies to help overcome this gap. These deeply embedded gendered preferences related to vaccination experiences reinforce the need to incorporate cultural and gender considerations when designing digital health interventions.

Second, we noted reliance on double record keeping, which respondents recommended phasing out to reduce the data entry burden (double requirement to complete both paper and mobile data entry) for field vaccinators and facility staff. Paper-based reporting is a requirement for health program audits. This has led to prioritization of paper reporting over the consistent use of the ZM-EIR for reporting. Our study found the perception that ZM-EIR uptake is hindered by physical record keeping, with a recommendation to transition to paperless reporting.

Although changes in immunization coverage across all antigens, timeliness for the BCG vaccine, and the reduction in the number of zero-dose children cannot be attributed wholly to the ZM-EIR program, the study findings are a promising indication that EIR data availability and use can support improvements in child health outcomes, such as vaccination, possibly through better tracking of clients and availability of data. Small pilot studies in countries such as Vietnam and Bangladesh with immunization programs rolling out EIRs report improvements in key immunization program milestones, such as timeliness of immunization (vaccinations delivered on time according to established schedule recommendations), a reduction in vaccination schedule dropouts, and improved ability for vaccinators to identify and follow-up with defaulters [3,19-22]. Similar to this evaluation, health workers using EIR systems report that they use the data to identify defaulters, coverage disparities, and vaccine stock levels and that they have more confidence to take action based on these data analyses [3]. However, this is the first known study of this degree of comprehensiveness, scale, and rigor to evaluate the availability and use of data of an EIR and the potential for associated effects on immunization outcomes.

Limitations of the Study

The study was a retrospective one, with a small sample of qualitative interviews, and information bias due to recall cannot be excluded. Stakeholders may also have expressed a positive view of the program due to reasons of social desirability. The MICS estimates show similar changes in Pakistan's Punjab Province, which was not a target of the ZM intervention. But since immunization has been at the forefront of public health issues in Pakistan, similar programs were likely in place in Punjab Province. These create challenges in the attribution of effects to programs such as ZM. ZM did not act in isolation but would have been greatly affected by the presence of other health system–strengthening activities, including improvements in supply chain logistics, training and supervision of providers, and demand generation activities targeting immunizations. The increase in immunization coverage may also be attributed to more structural factors (eg, public trust in the government).

Conclusion

This study contributes to the evidence base of the added value of increased data availability and use through the implementation of EIRs in low- and middle-income countries as part of broader EPI interventions. It aligns with key benefits identified in other studies of immunization information systems

XSL•FO
RenderX

and EIRs, namely alerts and reminders, missed vaccinations, supervision and oversight of vaccinators, and management and accountability. Study methods included a retrospective analysis of system-generated data, qualitative stakeholder interviews, and assessment of potential contributions to improvements in immunization coverage and timeliness and reduction in the number of zero-dose children. Findings suggest that the ZM-EIR is a promising technology platform that may have contributed to improvements in the coverage and timeliness of BCG vaccination, as well as the reduction in the number of zero-dose children as part of a broader package of immunization interventions in Sindh Province, Pakistan, from 2017 to 2019.

Acknowledgments

The external evaluation of the *Zindagi Mehfooz* (*safe life*) electronic immunization registry was led by HealthEnabled, in collaboration with a team of local and global researchers with the lead research group Gallup Pakistan, who supported the primary data collection for this study. We especially acknowledge inputs from Vijay Kumar Dharma, Sundus Iftikhar, and Bilal Gilani. This evaluation was funded by Gavi, the Vaccine Alliance.

Data Availability

A comprehensive report of all quantitative and qualitative data analyses conducted as part of the study has been included as Multimedia Appendix 1.

Authors' Contributions

MTS, DAS, and SC contributed to the research design and implementation and the review of study findings. RS provided a high-level review and inputs in the design, implementation, and review of study findings. The funders played an oversight role in the study design, data collection, analysis, interpretation, and writing of the paper. The authors/researchers in Pakistan had full access to all the data in the study, while those outside Pakistan had access to only postanalysis data in the study. All authors had final responsibility for the decision to submit for publication.

Conflicts of Interest

The core study team was independent of funders. MTS, DAS, and SC are employed by the Interactive Research and Development, the implementer of the Zindagi Mehfooz (safe life) electronic immunization registry (ZM-EIR). RS is from Gavi, which funded the implementation and evaluation of the ZM-EIR.

Multimedia Appendix 1
Zindagi Mehfooz Evaluation Comprehensive Report.
[PDF File (Adobe PDF File), 2431 KB - jmir_v26i1e52792_app1.pdf]

References

1. Pancholi J, Birdie R, Guerette J. Village Reach landscape analysis of electronic immunization registries: lessons learned from a landscape analysis of EIR implementations in low- and middle-income countries. Village Reach. 2020. URL: https://www.villagereach.org/wp-content/uploads/2020/07/Final-EIR-Landscape-Analysis.pdf [accessed 2023-06-10]
2. Groom H, Hopkins D, Pabst L. Immunization information systems to increase vaccination rates: a community guide systematic review. J Public Health Manag Pract 2015 May;21(3):227-248. [doi: 10.1097/phh.0000000000000100]
3. PATH, PAHO. A realist review of what works to improve data use for immunization: evidence from low- and middle-income countries. Project report. TechNet-21. URL: https://www.technet-21.org/en/resources/report/a-realist-review-of-what-works-to-improve-data-use-for-immunization [accessed 2024-08-22]
4. Werner L, Seymour D, Puta C, Gilbert S. Three waves of data use among health workers: the experience of the Better Immunization Data Initiative in Tanzania and Zambia. Glob Health Sci Pract 2019 Sep 26;7(3):447-456. [doi: 10.9745/ghsp-d-19-00024]
5. Landscape analysis of health information systems and data tools for identifying, reaching, and monitoring zero-dose and under-immunized children. USAID and MOMENTUM. 2022 Aug. URL: https://usaidmomentum.org/resource/landscape-analysis-of-health-information-systems-and-data-tools-for-identifying-reaching-and-monitoring-zero-dose-and-under-immunized-children/ [accessed 2024-08-22]
6. Carnahan E, Ferriss E, Beylerian E, Mwansa FD, Bulula N, Lyimo D, et al. Determinants of facility-level use of electronic immunization registries in Tanzania and Zambia: an observational analysis. Glob Health Sci Pract 2020 Aug 24;8(3):488-504. [doi: 10.9745/ghsp-d-20-00134]
7. Dolan SB, Burstein R, Shearer JC, Bulula N, Lyons H, Carnahan E, et al. Changes in on-time vaccination following the introduction of an electronic immunization registry, Tanzania 2016-2018: interrupted time-series analysis. BMC Health Serv Res 2022 Sep 20;22(1):1175 [FREE Full text] [doi: 10.1186/s12913-022-08504-2] [Medline: 36127683]

XSL•FO
RenderX

8. National Institute of Population Studies Islamabad, Pakistan, ICF International. Pakistan demographic and health survey 2012-13. DHS. 2013. URL: https://dhsprogram.com/pubs/pdf/fr290/fr290.pdf [accessed 2024-08-22]
9. Multiple indicator cluster survey – Pakistan (Sindh). UNICEF. 2014. URL: https://mics.unicef.org/news_entries/56/PAKISTAN-(SINDH)-2014-MICS-FINAL-REPORT-&-DATASETS-RELEASED [accessed 2024-07-04]
10. Mehmood M, Setayesh H, Siddiqi DA, Siddique M, Iftikhar S, Soundardjee R, et al. Prevalence, geographical distribution and factors associated with pentavalent vaccine zero dose status among children in Sindh, Pakistan: analysis of data from the 2017 and 2018 birth cohorts enrolled in the provincial electronic immunisation registry. BMJ Open 2022 May 18;12(5):e058985 [FREE Full text] [doi: 10.1136/bmjopen-2021-058985] [Medline: 35584879]
11. Habicht JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. Int J Epidemiol 1999 Feb;28(1):10-18. [doi: 10.1093/ije/28.1.10] [Medline: 10195658]
12. Siddiqi DA, Abdullah S, Dharma VK, Shah MT, Akhter MA, Habib A, et al. Using a low-cost, real-time electronic immunization registry in Pakistan to demonstrate utility of data for immunization programs and evidence-based decision making to achieve SDG-3: insights from analysis of big data on vaccines. Int J Med Inform 2021 May;149:104413. [doi: 10.1016/j.ijmedinf.2021.104413] [Medline: 33652259]
13. Chandir S, Dharma VK, Siddiqi DA, Khan AJ. Feasibility of using global system for mobile communication (GSM)-based tracking for vaccinators to improve oral poliomyelitis vaccine campaign coverage in rural Pakistan. Vaccine 2017 Sep 05;35(37):5037-5042. [doi: 10.1016/j.vaccine.2017.07.026] [Medline: 28802756]
14. Chandir S, Siddiqi DA, Hussain OA, Niazi T, Shah MT, Dharma VK, et al. Using predictive analytics to identify children at high risk of defaulting from a routine immunization program: feasibility study. JMIR Public Health Surveill 2018 Sep 04;4(3):e63 [FREE Full text] [doi: 10.2196/publichealth.9681] [Medline: 30181112]
15. Chandir S, Siddiqi DA, Dharma VK, Shah MT, Turab A, Khan MI, et al. Zindagi Mehfooz (Safe Life) digital immunization registry: Leveraging low-cost technology to improve immunization coverage and timeliness in Pakistan. Iproceedings 2018 Sep 17;4(2):e11770. [doi: 10.2196/11770]
16. Chandir S, Siddiqi DA, Setayesh H, Khan AJ. Impact of COVID-19 lockdown on routine immunisation in Karachi, Pakistan. Lancet Global Health 2020 Sep;8(9):e1118-e1120. [doi: 10.1016/s2214-109x(20)30290-4]
17. Khan AA, Siddique M, Abdo M, Pearman J, Prusik T, Chandir S. Scanning 2D barcodes on vaccine vials to link vials to immunized child, a pilot study. Eur J Public Health 2021;31(Supplement_3):165-173. [doi: 10.1093/eurpub/ckab165.173]
18. Siddique M, Iftikhar S, Dharma VK, Shah MT, Siddiqi DA, Malik AA, et al. Using geographic information system to track children and optimize immunization coverage and equity in Karachi, Pakistan. Vaccine 2023 May 02;41(18):2922-2931 [FREE Full text] [doi: 10.1016/j.vaccine.2023.03.051] [Medline: 37012115]
19. Moshtaq A. eGovernment for development eHealth case study no.6: electronic immunisation registry and tracking system in Bangladesh. eGovernment for Development. 2004. URL: https://www.egov4dev.org/health/case/banglaimmune.shtml [accessed 2024-08-20]
20. Akther M, Onishi T, Kidokoro T. E-government in a developing country: citizen-centric approach for success. IJEG 2007;1(1):38-51. [doi: 10.1504/ijeg.2007.014342]
21. Sharing lessons learned from Vietnam's national electronic immunization registry to improve global vaccination coverage. PATH. 2019. URL: https://www.path.org/our-impact/media-center/sharing-lessons-learned-from-vietnams-national-electronic-immunization-registry-to-improve-global-vaccination-coverage/ [accessed 2024-08-20]
22. Digital square electronic immunization registries in low- and middle-income countries. PATH. 2021. URL: https://static1.squarespace.com/static/59bc3457ccc5c5890fe7cacd/t/60aee1bfd163646306fb924c/1622073794356/Digital+Square+EIR+Landscape_Final.pdf [accessed 2022-11-15]

Abbreviations

AI: artificial intelligence
BCG: Bacillus Calmette-Guérin
DPT: diphtheria-pertussis-tetanus
DSS: decision support system
EIR: electronic immunization registry
EPI: Expanded Program for Immunization
iDSS: immunization decision support system
MICS: Multiple Indicator Cluster Survey
OPV: oral polio vaccine
Penta: pentavalent
QR: quick response
UC: Union Council
ZM: Zindagi Mehfooz (safe life)

XSL•FO
RenderX

Edited by A Mavragani; submitted 15.09.23; peer-reviewed by T Ryman, MC Danovaro-Holliday; comments to author 16.10.23; revised version received 24.11.23; accepted 20.08.24; published 11.10.24.

Please cite as:
Mechael P, Gilani S, Ahmad A, LeFevre A, Mohan D, Memon A, Shah MT, Siddiqi DA, Chandir S, Soundardjee R
Evaluating the "Zindagi Mehfooz" Electronic Immunization Registry and Suite of Digital Health Interventions to Improve the Coverage and Timeliness of Immunization Services in Sindh, Pakistan: Mixed Methods Study
J Med Internet Res 2024;26:e52792
URL: https://www.jmir.org/2024/1/e52792
doi:10.2196/52792
PMID:39162666

©Patricia Mechael, Sara Gilani, Ahsan Ahmad, Amnesty LeFevre, Diwakar Mohan, Asra Memon, Mubarak Taighoon Shah, Danya Arif Siddiqi, Subhash Chandir, Riswana Soundardjee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Accessibility, Cost, and Quality of an Online Regular Follow-Up Visit Service at an Internet Hospital in China: Mixed Methods Study

Kun Wang[1*], MMed; Wenxin Zou[2*], MPH; Yingsi Lai[1,3], PhD; Chun Hao[1,3], PhD; Ning Liu[4,5], PhD; Xiang Ling[6], BS; Xiaohan Liu[6], MMed; Ting Liu[6], MMgmt; Xin Yang[7], MPH; Chenxi Zu[6*], MPA; Shaolong Wu[2,3*], PhD

[1]Department of Medical Statistics, School of Public Health, Sun Yat-Sen University, Guangzhou, China
[2]School of Government, Sun Yat-sen University, Guangzhou, China
[3]Sun Yat-sen Global Health Institute, Institute of State Governance, Sun Yat-sen University, Guangzhou, China
[4]School of Management, Lanzhou University, Lanzhou, China
[5]China Research Center for Government Performance-Management, Lanzhou University, Lanzhou, China
[6]Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[7]Department of Health Policy and Management, School of Public Health, Sun Yat-sen University, Guangzhou, China
[*]these authors contributed equally

Corresponding Author:
Shaolong Wu, PhD
School of Government
Sun Yat-sen University
No. 135 Xingang Xi Road
Guangzhou, 510000
China
Phone: 86 02087335706
Email: wushlong@mail.sysu.edu.cn

Abstract

Background: Telemedicine provides remote health care services to overcome constraints of time and space in accessing medical care. Similarly, internet hospitals in China support and provide remote health care services. Due to the COVID-19 pandemic, there has been a proliferation of internet hospitals. Many new services, including online consultations and regular online follow-up visit services, can now be accessed via internet hospitals in China. However, the accessibility, cost, and quality advantages of regular online follow-up visit services remain unclear.

Objective: This study aimed to evaluate the accessibility, costs, and quality of an online regular follow-up visit service provided by an internet hospital in China. By analyzing the accessibility, costs, and quality of this service from the supply and demand sides, we can summarize the practical and theoretical experiences.

Methods: A mixed methods study was conducted using clinical records from 18,473 patients receiving 39,239 online regular follow-up visit services at an internet hospital in 2021, as well as interviews with 7 physicians, 2 head nurses, and 3 administrative staff members. The quantitative analysis examined patient demographics, diagnoses, prescriptions, geographic distribution, physician characteristics, accessibility (travel time and costs), and service hours. The qualitative analysis elucidated physician perspectives on ensuring the quality of online health care.

Results: Patients were predominantly middle-aged men with chronic diseases like viral hepatitis who were located near the hospital. The vast majority were from Guangdong province where the hospital is based, especially concentrated in Guangzhou city. The online regular follow-up visit service reduced travel time by 1 hour to 9 hours and costs by ¥6 to ¥991 (US $0.86-$141.32) depending on proximity, with greater savings for patients farther from the hospital. Consultation times were roughly equivalent between online and in-person visits. Physicians provided most online services during lunch breaks (12 PM to 2 PM) or after work hours (7 PM to 11 PM), indicating increased workload. The top departments providing online regular follow-up visit services were Infectious Diseases, Rheumatology, and Dermatology. The most commonly prescribed medications aligned with the prevalent chronic diagnoses. To ensure quality, physicians conducted initial in-person consultations to fully evaluate patients before allowing online regular follow-up visits, during which they communicated with patients to assess conditions and determine if in-person

XSL•FO
RenderX

care was warranted. They also periodically reminded patients to come in person for more comprehensive evaluations. However, they acknowledged online visits cannot fully replace face-to-face care.

Conclusions: Telemedicine services such as online regular follow-up visit services provided by internet hospitals must strictly adhere to fundamental medical principles of diagnosis, prescription, and treatment. For patients with chronic diseases, online regular follow-up visit services improve accessibility and reduce cost but cannot fully replace in-person evaluations. Physicians leverage various strategies to ensure the quality of online care.

(J Med Internet Res 2024;26:e54902) doi:10.2196/54902

KEYWORDS

internet hospital; medical service; accessibility; cost; quality; regular follow-up

Introduction

Significance of Online Health Care

Time and space are the basic environments for human survival and development. One of the goals of human development is to break through the shackles of time and space. In the health care sector, time and space have always constrained both the supply and demand sides in accessing and providing health services, known as spatial accessibility and temporal accessibility. Therefore, how can we improve the spatial and temporal accessibility of health care services? Telemedicine actually provides one viable approach. Telemedicine refers to the use of electronic and telecommunication technologies to deliver health care services remotely [1]. It has been applied in health care since the invention of the telegraph and telephone. Passing through successive evolution phases involving radio, television, the internet, mobile health, and artificial intelligence, telemedicine has now become an integral component of health care delivery [2].

Initially, telemedicine only enabled the transmission of text and voice using telegraphy and telephony. Later, television technology allowed the transmission of medical images. With the popularization of digital technology and the internet, multimedia medical information could be inexpensively transmitted to the masses. Especially with the proliferation of smartphones, almost everyone can now access health information services anytime, anywhere [2]. With the addition of artificial intelligence, 5G networks, virtual reality, and wearable technology, telemedicine appears almost omnipotent [3-6]. As telemedicine has advanced, more traditional health care services have shifted from offline to online. Its applications have expanded from basic consultations to even complex remote surgeries [7].

Instant online health care services offer a significant advantage in terms of spatial accessibility as they eliminate the need for face-to-face interactions between health care service providers and recipients. Generally, online health care services are also more cost-effective than offline services. However, in comparison with offline services, the quality of online health care services tends to be relatively lower [8]. Due to concerns related to quality and risks, health care professionals and institutions are often hesitant to provide core health care services online. Additionally, government regulations may restrict the provision of certain services, such as initial diagnosis, through online platforms [9]. Therefore, what types of services can be offered online? How substantial are their advantages in terms of accessibility and cost? How can the quality of health care services be ensured?

Existing research on health service delivery shows that telemedicine delivered through the internet can provide appointment scheduling, consultation, payment, examination, report checking, symptom monitoring, prescriptions, medical conferences, follow-up visits, and even simple surgeries online [7,10,11]. For the most essential aspects of health care services—diagnosis, prescription, and treatment—these online services are only applicable to specific situations, such as mild symptoms, certain diseases (eg, skin and mental diseases), follow-up visits, remote areas lacking vessels and aircraft, and other medical conditions suitable for online health care. In terms of large-scale application, telemedicine is mostly used for regular follow-up visit services. A “regular online follow-up visit” refers to a specific service provided by internet hospitals for patients who require regular follow-up visit services. Doctors prescribe medications online for stable follow-up patients and have the medications delivered to them. How much time and distance can online regular follow-up visit services save in terms of accessibility? How much can patients in different regions save in terms of expenses? How can the quality of online regular follow-up visit services be ensured? Through a mixed methods study with China’s internet hospital, this paper attempted to answer these questions. On this basis, the paper further discusses the impact of telemedicine on health care service delivery.

Online Health Care in China

In China, medical resource distribution is unbalanced, with large urban-rural and regional differences [12,13]. Most high-quality medical resources are concentrated in the city's top tertiary public hospital [11,14]. In addition, a sound system of tiered diagnosis and treatment has not yet been established. Patients can choose medical institutions freely. In order to obtain high-quality medical services, patients flock to the tertiary hospitals [15]. As a result, tertiary public hospitals are crowded, but primary health care institutions are deserted. Through internet-based online consultation services, patients in remote and rural areas can conveniently access the services of physicians from tertiary hospitals through an internet hospital. Internet hospitals are a platform to provide remote health care services [16]. Internet hospitals have greatly improved the accessibility of high-quality medical services from tertiary hospitals, especially those in Beijing, Shanghai, and Guangzhou.

XSL•FO
RenderX

After the internet entered China in the 1990s, a variety of health-related websites also emerged. In the beginning, most websites only provided services such as health counseling or drug searches. With the advancement of technology, some internet platforms began to provide non-core medical services such as registration, payment, the purchase of medicines, and chronic disease management [8]. Some internet health care companies have begun to penetrate into the core areas of health care by offering medical consultations called *light consultations*. Apart from not providing prescriptions or treatment plans in compliance with regulations, there is no clear boundary between a light consultation and medical diagnosis. Eventually, the internet hospital established by the Guangdong Second Provincial General Hospital made a breakthrough. It carried out online video-based diagnosis and treatment through an internet platform. After issuing electronic prescriptions, patients could obtain medications at partner pharmacies or primary care facilities [17]. According to the different initiators, China's internet hospitals can be divided into 3 categories: government-led internet hospitals, public hospital–led internet hospitals, and enterprise-led internet hospitals [16]. Our study mainly focused on an internet hospital led by a public hospital.

Prior to 2018, internet health care companies and hospitals were actively exploring internet hospitals, but internet hospitals were still in the initial stages of development. Policies for internet hospitals were also not yet mature. In 2018, the General Office of the State Council issued an opinion on Promoting the Development of Internet plus health care in order to encourage the development of internet hospitals [18]. In order to implement the policy, the National Health Commission and the State Administration of Traditional Chinese Medicine organized and formulated the Internet Diagnosis and Treatment Management Measures (trial implementation), Internet Hospital Management Measures (trial implementation), and Telemedicine Service Management Standards (trial implementation) [9]. The issue of these 3 documents was the first time that China put forward detailed regulations on internet hospitals. This was a symbol that internet hospitals had entered the stage of standardized development in China [17].

After the issue of these 3 documents and the COVID-19 pandemic, internet hospitals were rapidly established in China. In 2017, 68 internet hospitals were identified in research [19]. In 2019, there were 158 internet hospitals nationwide [20]. However, in 2021, there were more than 1600 internet hospitals in China [21]. During the initial period of the COVID-19 pandemic, the Chinese government took strict quarantine measures to prevent the spread of the virus. It also prevented citizens from getting face-to-face medical services. Simultaneously, most of the hospitals were used to treat patients who had been infected, and they reduced offline outpatient services [22]. With the control of COVID-19 spread, it was a challenge to provide constant medical services to the large number of patients with chronic diseases [23]. At this point in time, internet hospitals were favored due to their noncontact nature. During the COVID-19 epidemic, internet hospitals not only provided people with health information on COVID-19 but also ensured constant medical services for patients with chronic diseases. Patients with chronic diseases can register for online consultations on internet hospitals, and physicians can confirm the patients' conditions based on their past medical records and online consultations. Medicines are delivered to the patient by a specialized courier company after the prescription is issued [24]. During the pandemic, the government leveraged internet hospitals to ensure patients without COVID-19 could still access medical services while alleviating the burden on health care facilities [21,25,26].

Internet hospitals continued to play a role after China lifted a series of blocking policies against the COVID-19 pandemic. Internet hospitals break down the barriers of time and space and improve the accessibility of health services [15]. During the COVID-19 pandemic, people used internet hospitals due to the lack of direct contact between physicians and patients (ie, high accessibility). However, in the later stages of the pandemic, the quality of online health care services became an issue of greater concern. The advantage of the accessibility of online health care services is undisputed, but the services' quality has been under discussion [27-29]. In order to ensure the medical quality of online health care services, the Internet Diagnosis and Treatment Management Measures (trial implementation) clearly states that physicians can issue online prescriptions for some common diseases and chronic diseases only if they clearly understand the patient's medical record information [9]. Online regular follow-up visit services in public hospitals are conducted under this law. This study aimed to analyze the advantages and disadvantages of a particular service delivered by online hospitals—regular follow-up visit services—from the perspective of accessibility, cost, and quality. By analyzing the accessibility, cost, and quality of online regular follow-up visit services from the supply and demand sides, we can summarize the practical and theoretical experiences.

Methods

Conceptual Framework

There are 2 very famous models of health care service delivery: the World Health Organization (WHO) health system model focusing on the supply side and the model by Andersen and Davidson [30,31] concentrating on the demand side. In 2000, the WHO published a report that established a milestone health system framework [32]. The health system includes 4 functions—stewardship, creating resources, financing, and delivering services—that determine 3 goals: health, responsiveness, and fairness. Using this model, the WHO evaluated the health system performance of its 192 member states. In 2010, 4 intermediate goals—accessibility, coverage, quality, and safety—were added to the model in order to link functions and objectives [30]. The model by Andersen and Davidson, since its establishment in the 1960s, has evolved several times into a series of health service frameworks. The latest model by Andersen and Davidson consists of 4 components—environment, population characteristics, health behavior, and outcomes—elaborating the interrelationships between health care service accessibility, utilization, and satisfaction [31]. Specifically, satisfaction includes quality, financing, convenience, and availability. The model by Andersen and Davidson has been widely used in studies on the equity of

XSL•FO
RenderX

health care service accessibility and utilization. In both models, health services as the most essential element connect the supply and demand sides, aiming to promote, restore, and maintain health. The accessibility, quality, and cost of health services play critical roles in both models as key indispensable components. Accordingly, we established a health service triangle model covering both supply and demand sides, as shown in Figure 1, to analyze the impacts of telemedicine on the accessibility, quality, and cost of online regular follow-up visit services in China.

Figure 1. The health service triangle model.

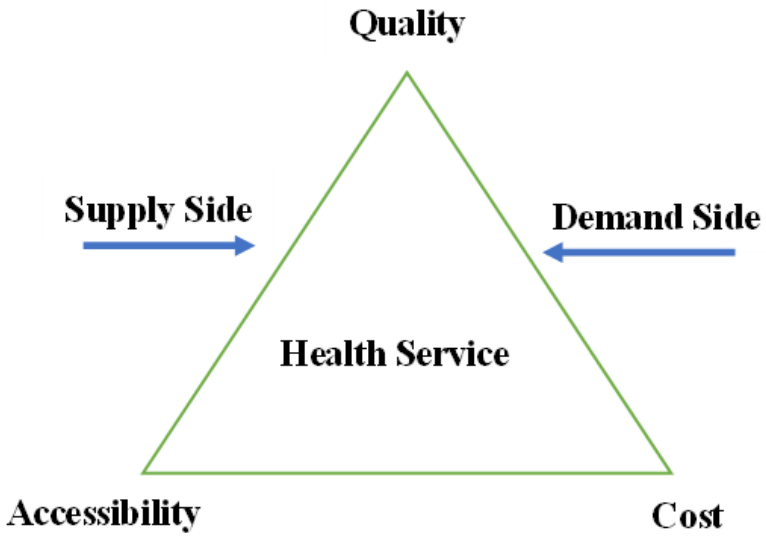

Data

Patient Care Access

Prior to receiving health care services, patients need to register either onsite at the hospital or online. Patients see a physician face to face within a week after they register online. However, via the internet hospital established by the Third Affiliated Hospital of Sun Yat-sen University (TAHSYSU), patients only need to log on to the TAHSYSU app on their cell phone and input their personal information, medical examination report, and previous prescriptions on the regular follow-up visit service page. Within 24 hours, the attending physician at TAHSYSU confirms their identity information and issues the appropriate prescription according to the patient's condition.

Demand Side

Data were collected for patients who received online regular follow-up visit services from TAHSYSU in 2021. The data included 64,202 drug records and 39,239 medical prescriptions for 18,473 patients who received the online regular follow-up visit service. The data provided 3 kinds of information. The first was general information about patients, including age and gender. The second was the address information where the patient will receive the medication delivery. The third was the department from which patients received online medical services, their disease, when a medical prescription was created, drug names, and the number of drugs.

Supply Side

Data were collected for physicians who provided online medical consultation and online regular follow-up visit services in the TAHSYSU from the WeChat mini program called TAHSYSU Internet Hospital. On the register page, the internet hospital departments are shown. If a patient selected a physician based on their condition, the physician's information, including their technical title and total number of service occurrences that the physician has provided, were shown. We also collected data about health care service accessibility, quality, and cost from physicians, nurses, and administrators through interviews.

Quantitative Methods

The features of the patients were described first. The patient address information was parsed into 3 levels of administrative regions, including district, city, and province (or municipality), as well as specific latitudes and longitudes, by calling the Amap web service application programming interface (API) using Python programs. The geographical distributions of patients at the district, city, province, and region levels were described. The characteristics of the patients were then described, including age, gender, departments in which patients received online medical services, diseases, and kinds of drugs.

Quantitative methods were mainly used to measure the advantages and disadvantages of online health care services in terms of cost and accessibility. The data to measure costs and accessibility came from map software, rather than patient surveys. A Python program was written to access the Amap web service API for driving route planning and public transportation route planning to estimate the time and cost for patients to travel to face-to-face visits. We used 9 AM to 11 AM and 2 PM to 4 PM as the simulation periods for patients' visits to the hospital. The time and cost for patients to travel to the hospital by public transportation or by car were simulated. The average of the simulation outcomes was used as the estimation result. The time and cost of drug delivery were obtained by checking the official website of SF Express. By comparing the time and costs, the advantages and disadvantages of offline and online regular follow-up visit services can be observed.

XSL•FO
RenderX

Finally, the features of physicians were described. The frequency of services provided by physicians in different departments and with different professional titles was described. By using the timing of medical prescription creation from the patient data, we could determine the number of prescriptions prescribed by physicians during different time periods within a day in order to analyze the accessibility of prescriptions.

Qualitative Methods

Qualitative methods were mainly used to discuss the advantages and disadvantages of online health care services in terms of medical quality. Semistructured interviews were conducted at TAHSYSU based on the descriptive statistics of the physicians. The interviews were designed to mainly answer 5 broad questions posed to physicians: (1) How can medical service quality be ensured during the online regular follow-up visit service process? (2) Which diseases and conditions would you recommend for patients to use the online regular follow-up visit service? (3) Is there any difference between the types of drugs prescribed online and those prescribed offline? (4) What other perspectives do you have on the quality of health care services provided by online hospitals? (5) What other thoughts do you have about the online regular follow-up visit service? For other interviewees, we asked them about their thoughts about internet hospitals. Interviews lasted approximately 30 minutes per interviewee. We stopped inviting new interviewees when the interviews stopped generating new information. Ultimately, 12 interviewees were selected from different departments in the hospitals. Of the 12 interviewees, there were 7 physicians, 2 head nurses, and 3 administrative staff members. Of the interviewees, 7 were physicians who regularly used such services from departments with high internet hospital utilization. After the interview, the audio data were catalogued and transcribed verbatim by a third-party company specializing in transcriptions in the Chinese language. Subsequently, using a grounded theory approach, we coded the text to identify motivations, risks, benefits, disease characteristics, quality control strategies, and medication traits associated with the online provision of health care. Ultimately, these data were used to elucidate how physicians determine patient suitability for online health care services and how they ensure quality of the online health care services provided.

Ethics Approval

The Human Studies Committee of Third Affiliated Hospital of Sun Yat-sen University (TAHSYSU) approved the study's protocol in compliance with the Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects (number 2022 02-279-01).

Results

Demand Side

Demographic Characteristics and Visits

The results are shown by age group and gender in Table 1. In total, 18,473 patients received 39,239 online regular follow-up visit services in 2021. Patients were predominantly 21 years to 40 years old (10,036/18,473, 54.33%) and male (10,538/18,473, 57.05%). The average number of visits per patient was 2.12. Patients aged 41 years to 60 years had the highest average number of visits (2.33), while those older than 80 years had the lowest number of visits (1.75). Male patients had a higher average number of visits (2.27) than female patients (1.93). Overall, middle-aged and male patients contributed to more visits on average.

Table 1. Demographic characteristics of patients (N=18,473) and the average number of visits (N=39,239; overall mean: 2.12 visits per person).

Characteristic	Visits, n (%)	Patients, n (%)[a]	Visits per person, mean
Age (years)			
≤20	1078 (2.75)	702 (3.8)	1.54
21-40	20,566 (52.41)	10,036 (54.33)	2.05
41-60	14,168 (36.11)	6093 (32.98)	2.33
61-80	3315 (8.45)	1578 (8.54)	2.10
≥80	112 (0.29)	64 (0.35)	1.75
Gender			
Female	15,298 (38.99)	7935 (42.95)	1.93
Male	23,941 (61.01)	10,538 (57.05)	2.27

[a]If 2 patients had the same district and ID information, they were considered the same person.

Geographical Distribution

Table 2 shows the geographical distribution of patients and average number of visits by region in China. The vast majority of patients were from south China (17,376/18,473, 94.06%), which also accounted for 94.71% (37,163/39,239) of total visits. Patients from east China comprised 2.37% (437/18,473) of the sample and 2.14% (840/39,239) of visits. The average number of visits per patient was highest in south China (2.14) and lowest in northeast China (1.52). The results show that the patient population was highly concentrated in south China, which drove a higher number of total visits and visits per patient compared with other regions. As shown in Table S1 in Multimedia Appendix 1 and Figure 2, the majority of patients from south China were located in Guangdong province, with 16,979 of the 18,473 patients (91.91%) from Guangdong. As shown in Tables S2 and S3 in Multimedia Appendix 1, Figure 3, and Figure 4, most patients were located in Guangzhou city, concentrated in

XSL•FO
RenderX

the Tianhe and Huangpu districts where the hospital is located. Tianhe and Huangpu had the highest number of patients due to proximity to the hospital. Yuexiu district attracts patients from nearby areas—Baiyun, Haizhu, and Liwan districts—due to its concentrated medical resources. In summary, the distribution of patients across provinces nationally and cities in Guangdong appears to follow a pattern of more patients in closer proximity to the hospital; however, the distribution across districts in Guangzhou city does not follow this pattern.

Table 2. Region-level geographical distribution of patients (N=18,473) and average number of visits (N=39,239; overall mean: 2.12 visits per person).

Regions of China	Visits, n (%)	Patients, n (%)	Visits per person, mean
South China	37,163 (94.71)	17,376 (94.06)	2.14
East China	840 (2.14)	437 (2.37)	1.92
Central China	795 (2.02)	426 (2.31)	1.87
Southwest China	260 (0.66)	128 (0.69)	2.03
North China	96 (0.24)	55 (0.3)	1.75
Northwest China	47 (0.12)	26 (0.13)	1.81
Northeast China	38 (0.1)	25 (0.13)	1.52

[a]If 2 patients had the same district and ID information, they were considered the same person.

Figure 2. Province-level geographical distribution of patients. TAHSYSU： Third Affiliated Hospital of Sun Yat-sen University.

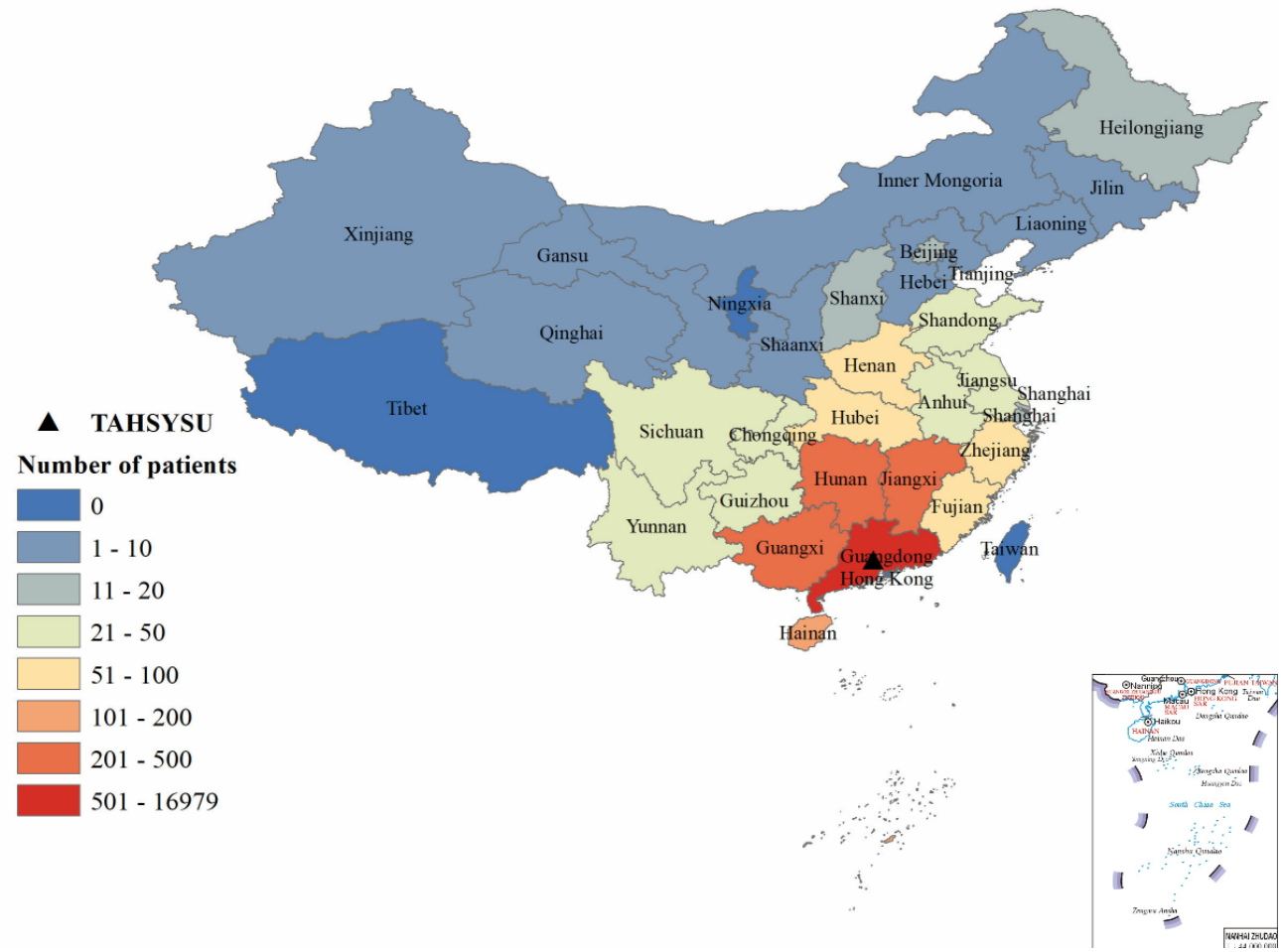

XSL•FO
RenderX

Figure 3. City-level geographical distribution of patients. TAHSYSU: Third Affiliated Hospital of Sun Yat-sen University.

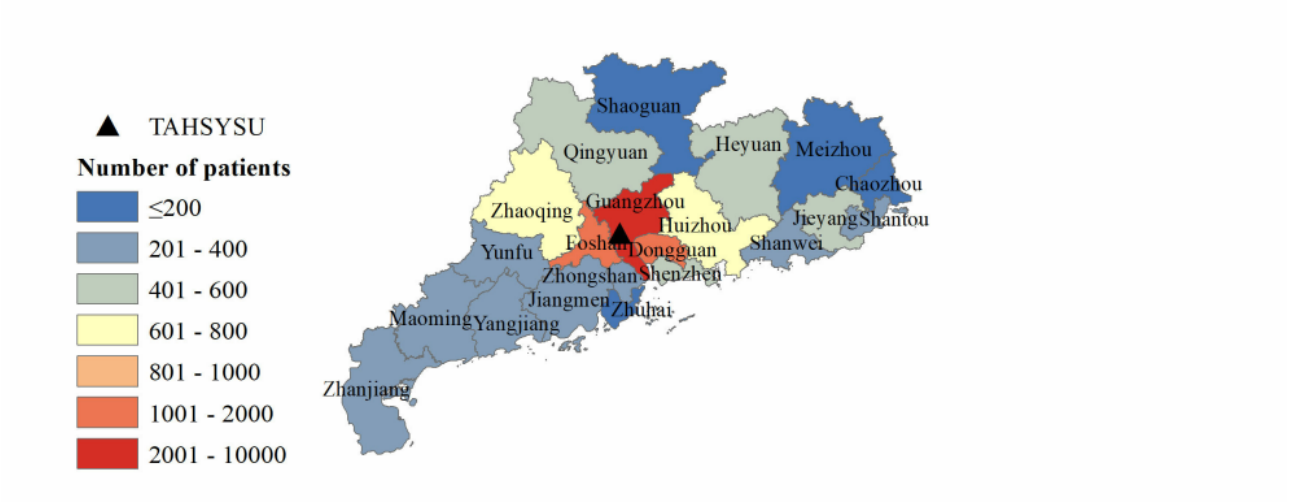

Figure 4. District-level geographical distribution of patients. TAHSYSU: Third Affiliated Hospital of Sun Yat-sen University.

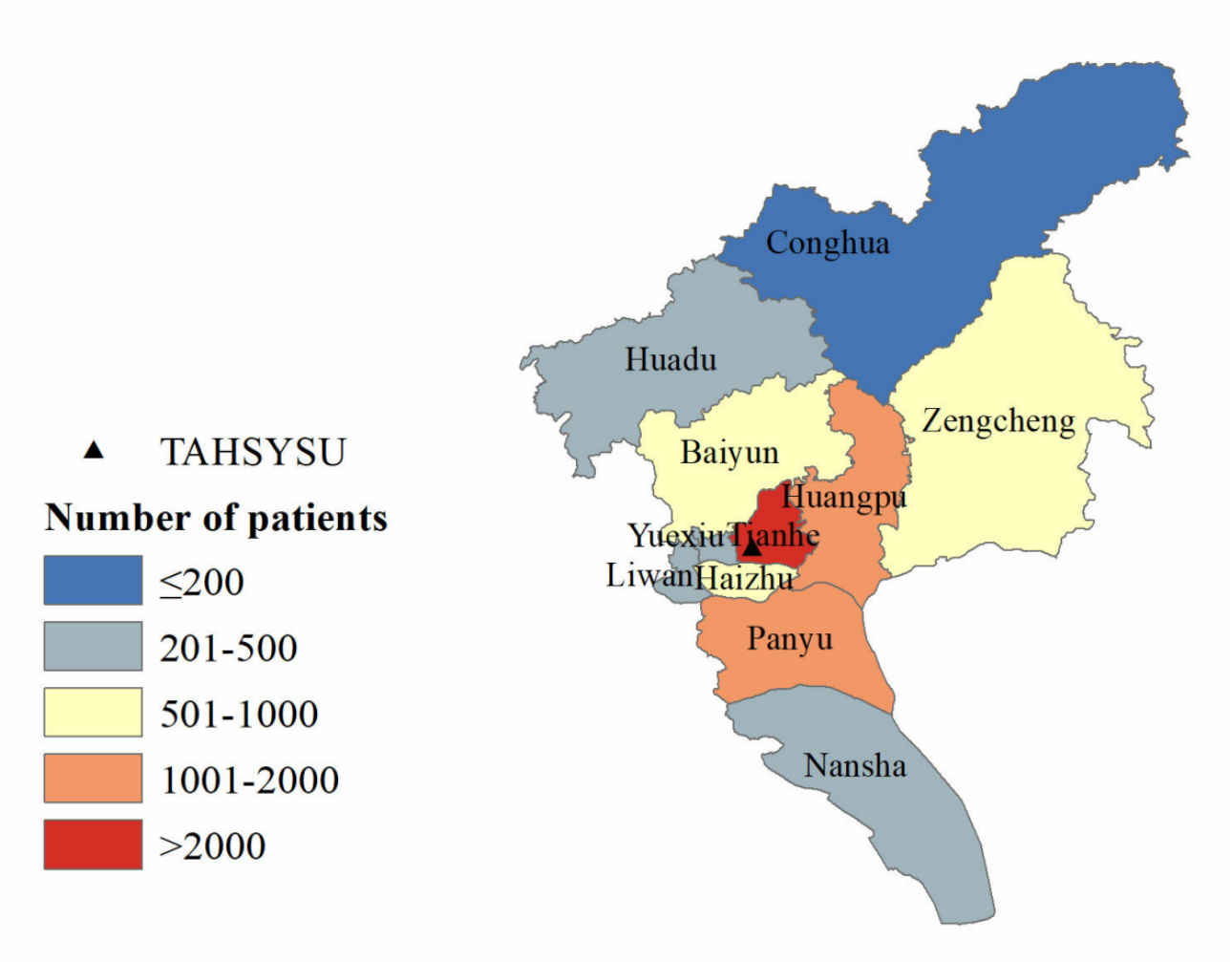

Description of Patients' Characteristics by Frequency of Regular Follow-Up Visits

According to the descriptive results described in the previous section, patients received online regular follow-up visit services on average 2.12 times a year. Therefore, the patients were divided into 3 groups, namely, patients who visited once a year, patients who visited twice a year, and patients who visited ≥3 times a year. As shown in Table 3, nearly one-half (9216/18,473, 49.89%) of the patients received online regular follow-up visit services once a year: the older the patient, the more often he or she received online regular follow-up visit services. There was also a gender difference, with slight differences between men and women in the patients who visited once, and men

XSL•FO
RenderX

representing significantly larger proportions of patients who visited twice (2343/4034, 58.08%) or ≥3 times (3417/5223, 65.42%). The proportion of patients who visited once a year was highest in Guangzhou (4427/9216, 48.04%), and as the number of visits increased, the proportion of patients from Guangzhou and other provinces decreased, while the proportion of patients from other cities in Guangdong Province increased.

Table 3. Characteristics of patients by frequency of regular follow-up visits.

Characteristics	Visits per year		
	Once (n=9216)	Twice (n=4034)	≥3 times (n=5223)
Age (years), mean (SD)	38.57 (14.51)	40.19 (13.73)	41.61 (12.54)
Gender, n (%)			
Female	4437 (48.14)	1691 (41.92)	1806 (34.58)
Male	4779 (51.86)	2343 (58.08)	3417 (65.42)
District, n (%)			
Guangzhou city	4427 (48.04)	1635 (40.53)	1942 (37.18)
Other cities in Guangdong province	3979 (43.17)	2063 (51.14)	2933 (56.16)
Other provinces	810 (8.79)	336 (8.33)	348 (6.66)
Departments (top 10), n (%)			
Infectious Diseases	4487 (48.69)	2537 (62.89)	3904 (74.75)
Dermatology	1292 (14.02)	363 (9)	179 (3.43)
Rheumatology	785 (8.52)	342 (8.48)	329 (6.30)
Gynecology	508 (5.51)	133 (3.3)	86 (1.65)
Endocrinology	324 (3.52)	87 (2.16)	93 (1.78)
Hepatobiliary Surgery	306 (3.32)	133 (3.3)	215 (4.12)
Cardiovascular Internal Medicine	235 (2.55)	59 (1.46)	70 (1.34)
Neurology	159 (1.73)	71 (1.76)	89 (1.70)
Nephrology	141 (1.53)	67 (1.66)	72 (1.38)
Gastroenterology	132 (1.43)	47 (1.17)	43 (0.82)

In the Infectious Diseases department, the proportions of patients who visited twice or ≥3 times were higher than the proportion of patients who visited once a year. However, the situation was the opposite in the Dermatology department. Therefore, the frequency of online regular follow-up services that patients receive varies among patients with different types of diseases.

Supply Side

Physician Departments

As shown in Table S4 in Multimedia Appendix 1, infectious diseases accounted for the largest percentage of visits (26,163/39,239, 66.68%) and patients (10,935/18,473, 59.2%), with an average of 2.39 visits per patient. Rheumatology and Immunology was the second highest ranked department by visits (2804/39,239, 7.15%) and patients (1442/18,473, 7.81%), with an average of 1.94 visits. Dermatology had the third highest proportion of patients (1828/18,473, 9.9%) but a lower visit rate of 1.45 visits per patient. The departments with the highest average visits per patient were Infectious Diseases (2.39 visits), Hematology (2 visits), and Interventional Radiology (2.22 visits). Anorectal Surgery and Vascular Surgery had the fewest visits and patients overall. It is worth noting that the departments with the highest volume of regular follow-up visit services were the hospital's signature specialty departments. The Infectious Diseases department at TAHSYSU primarily specializes in the diagnosis and treatment of various forms of viral hepatitis. In summary, the results show a concentration of online regular follow-up visit services in infectious diseases and rheumatology, with a significant variation in visit rates across departments.

Online Service Volume

The data in Table 4 reflects activity at the TAHSYSU internet hospital since the start of the online medical services, including online consultation services and online regular follow-up visit services, until February 25, 2023. A total of 315,172 online medical services were provided by 375 physicians. Associate chief physicians provided the most services (139,530/315,173, 44.27%) but accounted for only 41.07% (154/375) of physicians. Chief physicians provided 38.76% (122,146/315,173) of the services but accounted for only 22.40% (84/375) of the physicians, with the highest average number of services provided per physician (1454.12). Attending physicians accounted for 32% (120/375) of physicians but only 16.79% (52,913/315,172) of the services, with an average of 440.94 services per physician. Physicians with other titles accounted for only 0.18% (583/315,172) of the services. In brief, higher-ranked physicians including associate chief and chief physicians provided the

XSL•FO
RenderX

majority of services and saw higher service volumes per physician.

Table 4. Online service volume (N=315,172) provided by physicians with different technical titles (N=375) from the start of the online medical service until February 25, 2023 (average number of services per physician: 840.46).

Technical titles	Services provided, n (%)	Physicians, n (%)	Services provided per physician, mean
Associate chief physician	139,530 (44.27)	154 (41.07)	906.04
Chief physician	122,146 (38.76)	84 (22.4)	1454.12
Attending physician	52,913 (16.79)	120 (32)	440.94
Other technical titles	583 (0.18)	17 (4.53)	32.29

Online Service Hours

As shown in Figure 5, physicians had 2 peak times when they responded to patients via the internet hospital, from 12 PM to 1:59 PM and from 7 PM to 10:59 PM. These 2 time periods are usually freer times for physicians. For instance, 12 PM to 1:59 PM is their lunch break, and 7 PM to 10 PM is their off-duty time. Overall, after the physician had met with the patients who were seen face to face or during his or her own free time, they met with the patients via online regular follow-up visits.

Figure 5. Times of day physicians performed visits with patients online; each time on the x axis indicates a full hour (eg, 1 PM=1 PM to 1:59PM).

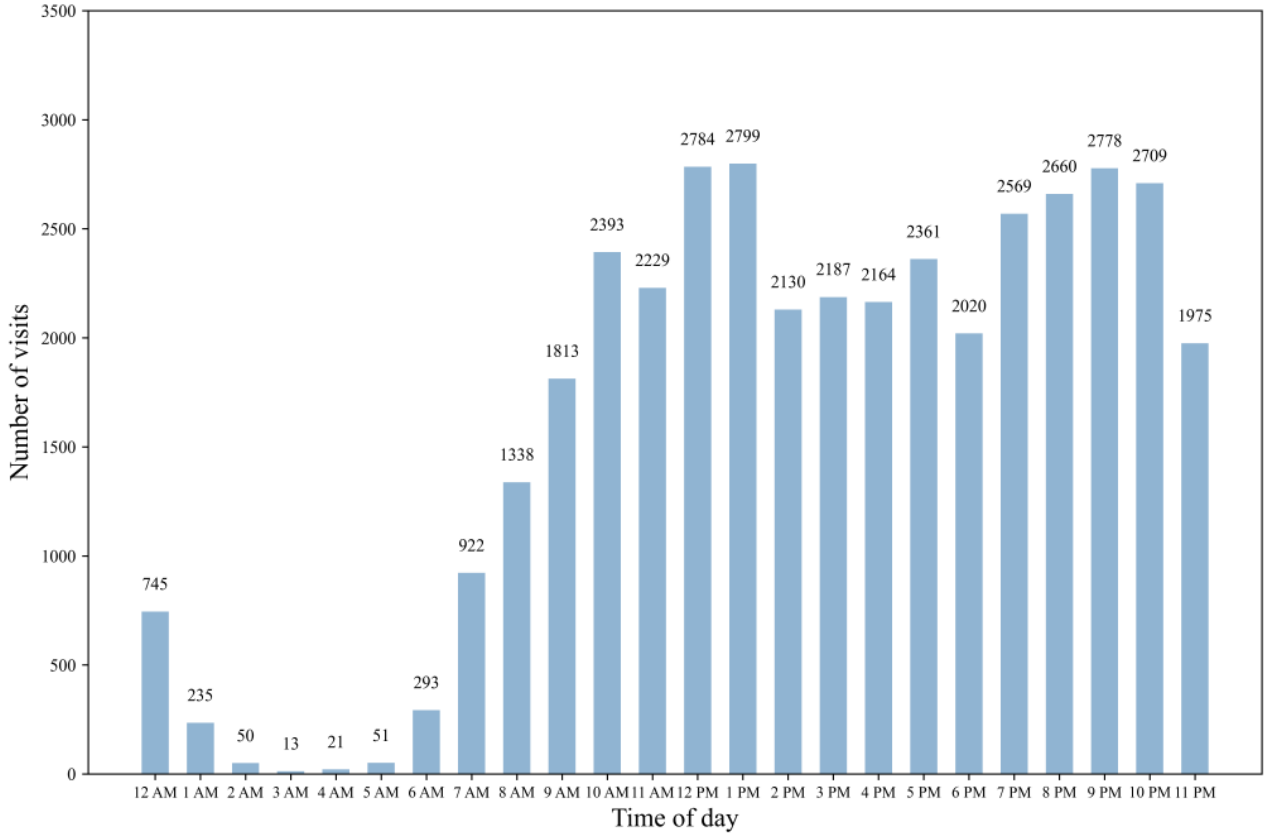

Internet hospital services have subtly increased physicians' working hours:

> *Internet hospitals are acting as a triage. Continuity of medical care for patients can be enhanced through online regular follow-up visit services at internet hospitals. But with the increase in the number of patients, we are also struggling. Many patients with stabilized conditions ask a lot of questions during online regular follow-up visit services out of concern for their condition, which increases our workload.* [P4, associate chief physician, Endocrinology department]

Health Services

Disease Diagnoses

After excluding patients with a diagnosis of counseling and missing records for a diagnosis, 5233 patients had a definite diagnosis. As shown in Table 5, chronic viral hepatitis B was the most common diagnosis, accounting for 49.3% (2580/5233) of the diagnoses. Chronic active viral hepatitis B (491/5233, 9.38%), hepatopathy (305/5233, 5.83%), and liver transplantation status (155/5233, 2.96%) were also among the top diagnoses, indicating a technology advantage for liver-related conditions at TAHSYSU.

XSL•FO
RenderX

Table 5. The diseases of patients using online health care services (N=5233; top 10 diagnoses, after removing counseling).

Diagnosis	Patients, n (%)
Chronic viral hepatitis B	2580 (49.30)
Chronic active viral hepatitis B	491 (9.38)
Hepatopathy	305 (5.83)
Liver transplantation status	155 (2.96)
Decompensated liver cirrhosis after viral hepatitis B	118 (2.25)
Acne	118 (2.25)
Systemic lupus erythematosus	80 (1.53)
Malignant tumor of liver	54 (1.03)
Rheumatoid arthritis	54 (1.03)
Chronic nephritic syndrome	50 (0.96)

The top 10 diagnosis were all chronic diseases. This means that patients have a clear diagnosis and a clear treatment plan, which makes the diseases more suitable for online treatment. One of the physicians interviewed said the following:

> *The most important characteristics of chronic diseases are that they require long-term medication and are not so dangerous. For patients who need to take medication for a long period of time, it is burdensome for them to travel to the pharmacy every time they need to refill their medication. Therefore, the internet hospital's chronic disease follow-up service is developing rapidly.* [P1, associate chief physician, Infectious Diseases department]

Prescribed Drugs

As shown in Table 6, of the 64,202 medications, the most commonly prescribed drugs were tenofovir disoproxil fumarate tablets (10,209/64,202, 15.9%) and entecavir dispersible tablets (9702/64,202, 15.11%), both used to treat chronic hepatitis B. In summary, the main drugs prescribed align with the common diagnoses of liver diseases and their complications.

Table 6. Drugs prescribed online by physicians (top 10; N=64,202).

Name of drug	Prescriptions, n (%)
Tenofovir disoproxil fumarate tablets	10,209 (15.9)
Entecavir dispersible tablets	9702 (15.11)
Tenofovir alafenamide fumarate tablets	3724 (5.8)
Isotretinoin soft capsules	1599 (2.49)
Danshnewuweizi tablet	1298 (2.02)
Calcium carbonate and vitamin D3 tablets	828 (1.29)
Methylprednisolone tablets	772 (1.2)
Compound glycyrrhizin tablets	746 (1.16)
Vitamin E soft capsules	734 (1.14)
Hydroxychloroquine sulfate tablets	721 (1.12)

Drugs represent a specialized and sensitive issue. For patients who are followed up at the TAHSYSU internet hospital, the physician prescribes medication based on the patient's past medical history and current condition. However, physicians are very cautious when prescribing medications for first-time patients who visit the internet hospital. One of the physicians interviewed said the following:

> *I have never prescribed medication to a patient at an internet hospital on a third-party platform such as Haodaifu. But on the TAHSYSU internet platform, sometimes I can't stand the patient's request to prescribe drugs to the first-time patients. Usually, the drugs I prescribe are safer and more conservative, and I will not prescribe drugs with stronger side effects to my patients.* [P2, associate chief physician, Cardiovascular department]

Accessibility of Online Services

As shown in Table 7, online medical services save the time patients spend on appointment registration, waiting in the hospital, and traveling. Patients from Guangzhou, other cities in Guangdong, or other provinces in China spend 1.11 hours, 3.66 hours, and 9.35 hours, respectively, traveling by public transportation. If they choose to drive, they spend 0.74 hours, 2.51 hours, and 9.62 hours, respectively. However, if patients see a physician online or face to face, it takes several minutes of the physician’s time to understand the patient’s condition

XSL•FO
RenderX

and prescribe medications. Online medical services have no time advantages related to the consultation and prescriptions. After physicians complete the diagnosis and prescriptions, patients using the online medical service receive their drugs by express shipments, and patients using the face-to-face medical service go to hospital pharmacy to get drugs. Getting to the hospital pharmacy costs less time than express delivery. Overall, the online medical service has advantages in appointment registration times, waiting times, and travel times but has no advantages in the time it takes for the consultation, prescriptions, and receiving the drugs. The outstanding advantages in travel time make the disadvantages negligible.

Table 7. Comparative estimation of accessibility associated with offline versus online regular follow-up visit services, by location.4

Accessibility features	Guangzhou City	Other cities in Guangdong province	Other provinces
Appointment registration and waiting times			
Face to face (weeks)	≤1	≤1	≤1
Online (hours)	≤24	≤24	≤24
Time to hospital (hours), mean (SD)			
One-way public transportation	1.11 (0.52)	3.66 (1.38)	9.35 (7.36)
Drive one way	0.74 (0.29)	2.51 (1.10)	9.62 (5.98)
Online	0	0	0
Consultation and prescribing time			
Face to face	Several minutes	Several minutes	Several minutes
Online	Several minutes	Several minutes	Several minutes
Time for patient to receive drugs (hours)			
Hospital pharmacy	≤1	≤1	≤1
Express delivery, mean (SD)	27.00 (0.62)	28.78 (5.69)	48.22 (17.00)

Cost of Online Services

Table 8 shows the characteristics of online health care services, such as the savings in indirect medical costs associated with the fact that patients do not need to travel to a hospital. No matter how far away the patient lives, the online health care service costs ¥0 for travel. Patients from Guangzhou, other cities in Guangdong, or other provinces in China spend ¥5.83, ¥98.38, and ¥359.41, respectively, to travel by public transportation to the hospital. If they choose to drive, they spend ¥21.14, ¥208.34, and ¥990.77, respectively. There is a standardized consultation fee of ¥10 for the online regular follow-up visit service. If you choose to have a face-to-face consultation, it costs at least ¥10 for the consultation, depending on the technical title of the physician you choose. Patients pay express delivery fees for online medical services but not for in-person hospital visits. Overall, online medical services save a lot of money for patients, especially patients located far from the hospital.

Table 8. Comparative estimation of costs associated with offline versus online regular follow-up visit services, by location.

Costs	Guangzhou City	Other cities in Guangdong province	Other provinces
Fees to get to the hospital (¥), mean (SD)			
One-way public transportation	5.83 (3.24)	98.38 (84.38)	359.41 (244.25)
Drive one way	21.14 (18.00)	208.34 (147.03)	990.77 (650.53)
Online	0	0	0
Consulting and prescribing fees (¥)			
Face to face	≥10	≥10	≥10
Online	10	10	10
Fees for drug prescriptions (¥), mean (SD)			
Hospital pharmacy	0	0	0
Express delivery (≤1 kg)	12.00 (0)	13.00 (0)	18.04 (0.34)

XSL•FO
RenderX

Quality of Online Services

According to the aforementioned examination of diseases and pharmaceuticals, the cohort using online regular follow-up visit services seems to be largely comprised of patients with chronic pathologies or those in the chronic stage of their diseases. Chronic pathologies are typified by discrete diagnostic and treatment paradigms. For patients with chronic diseases receiving standardized therapeutic interventions, their morbidity seems unlikely to exhibit precipitous, perilous perturbations in the short term, and their pharmaceutical regimen also does not have considerable variability. In accordance with guidelines, such characteristics appear conducive to return online visit services provided by the internet hospital.

> *I will decide whether the patient will have an online follow-up or a face-to-face follow-up based on the disease diagnosis and treatment guidelines. For some diseases, the guideline usually requires a three-month follow-up, so I will ask the patient to go to the internet hospital for a follow-up visit and prescription. At the third month, I will notify the patient to come to face-to-face. But for IgA nephropathy, the guideline requires patients to have a follow-up once a week, so I will not let IgA nephropathy patients go to the internet hospital for follow-up.* [P3, associate chief physician, Nephrology department]

Patients use online regular follow-up visit services, with some doing so voluntarily and others being recommended by their doctors. At such times that clinical evaluation or diagnostic testing is warranted, the physician may additionally counsel patients to pursue in-person medical attention. For patients exhibiting superior adherence, the paradigm of online pharmaceutical renewal consultations coupled with periodic in-person evaluations may ensure requisite medical care quality.

> *For patients in our department, our physicians generally recommend that patients follow up face-to-face once every 3 months. I will regularly post an announcement on the internet hospital platform for patients to follow up face to face once every 3 months. For patients with hepatitis B in our department, 3-monthly follow-up is good for controlling the progression of the disease. If the patient is not followed up at 6 months or even a year, then the situation will be very bad.* [P1, associate chief physician, Infectious Diseases department]

This study focused on the online regular follow-up visit service. In practice, all patients had previously visited the hospital for an initial face-to-face visit, during which the physician obtained a comprehensive understanding of their medical condition. After this visit, the physician recommended that the patient continue their care with the internet hospital for regular follow-up visits and online prescriptions. This model is likely to improve the quality of the online regular follow-up visit service by ensuring that patients are under the care of a physician who is familiar with their medical history.

> *The online regular follow-up visit service provided by the THYSYSU internet hospital requires a preliminary offline, face-to-face clinical evaluation. During the initial face-to-face consultation, I will determine the patient's condition through physical examination and some supporting examinations. I will let the patient whose condition is more stable and who is far away have a follow-up consultation online.* [P2, associate chief physician, Cardiovascular department]

There is communication between the physician and patient during the online regular follow-up visit service. As part of the registration process for internet hospital services, patients are required to provide a comprehensive description of their medical situation. During the online consultation, patients can ask the physician up to 3 questions. The physician will then assess the patient's condition based on the description of the condition and the questions asked. If the patient's condition is stable, the physician may prescribe a prescription online. However, if the patient's condition has changed significantly, the physician will advise the patient to seek in-person care at the hospital. This online communication between physician and patient allows the physician to assess the patient's condition and make treatment recommendations accordingly.

> *I make my decision based on the patient's description of their condition. For example, if a patient reports chest tightness and chest pain, I would advise them to go to the hospital immediately. If a patient reports poor blood pressure control during their usual blood pressure monitoring, I would adjust their medication online. The problem is analyzed on a case-by-case basis.* [P2, associate chief physician, Cardiovascular department]

These practices contribute to the quality of the online regular follow-up visit service. However, according to the interviewees, the effectiveness of the online regular follow-up visit service is still limited compared with face-to-face medical services. They cannot fully replace face-to-face medical services because patients may overlook symptoms that indicate a changing health condition due to their lack of medical expertise.

Discussion

Principal Findings

This article synthesizes the supply-side and demand-side dynamics of the TAHSYSU internet hospital together with our quantitative analysis and qualitative interview findings, in order to elucidate the benefits and limitations of online medical facilities with respect to the biological determinant of morbidity patterns as well as the 3 dimensions of health care access, expenditures, and quality. Empirical investigations have shown that the cohort that uses internet hospital services is predominantly composed of patients with chronic diseases with a spatially heterogeneous distribution. Online regular follow-up visit services reduce the time related to registration, waiting for the appointment, and travel, as well as transportation-related costs, but they also increase physician workloads. Physicians leverage patient selection, condition dialogue, and the offline regular follow-up visit service to ensure the quality of online medical care, as virtual services cannot fully replace in-person health care.

XSL•FO
RenderX

The ability of internet hospitals to overcome the constraints of time and space is a widely held view among internet hospital researchers [15,26,33]. However, in our study, the online regular follow-up visit service requires the patients to have a face-to-face consultation first to determine their condition. Therefore, the initial face-to-face consultation does not overcome the time and space constraints, but the subsequent online regular follow-up visit service does. This feature determines that most of the patients using the online regular follow-up visit service come from Guangdong province or provinces closer to Guangdong province. Patients can readily consult clinicians regarding their conditions at any time and in any locale, yet the inability to perform physical examinations or diagnostic tests severely limits clinical determinations based solely on patient narratives. Physicians are not inclined to make a definitive diagnosis and prescription when evaluating patients who have not been seen face to face. It is important to dialectically think about the features that internet hospitals can use to overcome the limitations of time and space.

The type of disease, as a biological factor, influences both the patient's and physician's use of internet hospitals. Different diseases require different diagnostic and therapeutic methods, and the characteristics of the disease also determine whether the disease is suitable for online regular follow-up visit services. For clinicians, the decision to provide online regular follow-up visit services depends on the chronicity or acuity of the conditions under their care, as well as the number of patients. Additionally, the income of physicians in different departments affects their willingness to provide online services. These points were confirmed in our interviews with physicians. After excluding patients with a diagnosis of counseling or who had missing diagnosis records, more than one-half of the patients were patients with liver disease. Nearly one-half of the medications on the medication record were antivirals for viral hepatitis. The Infectious Diseases department at TAHSUSY is a national key clinical specialty, and there are a lot of patients going to the Infectious Diseases department. The vast majority of patients who received online regular follow-up visit services were patients with chronic hepatitis. The chronic nature of these diseases allows physicians to reassure patients that they can be followed up online and to serve more patients at the same time. Patients with chronic diseases can also use the online regular follow-up visit service with confidence.

This analysis uses multiple dimensions of accessibility, including appointment, geographic, prescription, and medication availability, to evaluate the online regular follow-up visit service. Specifically, we measured appointment and waiting times, travel time via various transport modes, consultation and prescription times, and medication delivery times. The results of this study show that online regular follow-up visit services have superior appointment and geographic accessibility than in-person consultations. Prescription accessibility is equivalent between online and in-person services, while medication accessibility is inferior with the online regular follow-up visit service. For in-person consultations, patients must book appointments 1 week in advance, without immediate face-to-face access. Although online patients await physicians' replies postregistration, responses are furnished within 24 hours. Despite proximity, in-person visits entail travel time, whereas online consultations transcend spatiotemporal constraints without travel requirements. Regarding prescription accessibility, physicians expend minutes verifying conditions before prescribing. Consequently, the online regular follow-up visit service confers no advantages over in-person follow-ups in this dimension. However, the online regular follow-up visit service demonstrates inferior medication accessibility, as in-person patients can directly obtain their medications after the consultation and prescribing of medicines, while online patients rely on delivery. In summary, despite deficient medication accessibility, the online regular follow-up visit service demonstrates superior availability overall. Moreover, with proper planning, patients with chronic diseases accessing the online regular follow-up visit service need not urgently require medications.

The cost advantages of the online regular follow-up visit service are more pronounced than the advantages for service accessibility. In addition to the requisite medication and registration expenses, the online regular follow-up visit service saved transportation costs that constitute indirect medical fees. Greater distances from the hospital amplify the economic benefits conferred by internet hospitals. Research undertaken at Huashan Hospital, which is affiliated with Fudan University, used questionnaire instruments to quantify cost and time savings conferred by internet hospitals [34]. We used a new method to estimate the cost and time savings of internet hospitals, which was superior to the questionnaire method. Traditional questionnaire methodology necessitates completion by respondents. Conversely, our novel approach merely requires approximate patient address information. Upon converting addresses into geolocation coordinates and accessing the Amap web service API for vehicular and public transport routing, the time expenditure and travel expenses incurred by patients undergoing in-person clinical visits can be approximated. The predominant virtues of this methodology encompass heightened accuracy compared with traditional questionnaire protocols and mitigation of retrospective bias among the study population. Furthermore, this strategy facilitates analysis of indirect medical costs.

Comparison With Prior Work

According to Donabedian [35], quality can be defined as "structure-process-outcome" quality. Qi et al [34] and Chen et al [36] partially assessed the medical quality of internet hospitals by administering patient satisfaction surveys regarding their use of online services. Patient satisfaction can be an indicator of outcome quality. In our research, the quantitative research mainly focused on the process quality of online health care. In our opinion, the control of the quality of medical care in the medical process is more important than the satisfaction with the demand side. The difference in medical knowledge between physicians and patients is enormous. Patients' perceptions of the quality of medical care are relatively subjective, and the provider is the controller of the quality of care in the health service process. According to our quantitative research, patients visit the doctor an average of more than 2 times, with some patients even visiting 6 times. Patients' repeated use of the online regular follow-up visit service indicates their recognition

XSL•FO
RenderX

of the quality of care provided by the internet hospital. According to our qualitative research, physicians mainly understand the patient's condition during the initial consultation. During the online regular follow-up visit service, they understand the patient's disease status based on the patient's description and communication between the 2 parties. By comparing the patient's condition during the initial consultation and communication during the regular follow-up visit, physicians can judge whether the effect of the patient's online follow-up consultation is similar to that of the face-to-face consultation. These measures help physicians ensure the quality of care during the online follow-up.

Limitations

This article has several limitations. First, as a single-center study in 1 sample hospital, it may have biases in the research results. Although a focused hospital study can provide detailed descriptions, there are significant regional, urban-rural, and typological differences among China's public hospitals. Thus, the findings from this comprehensive hospital in a southeastern coastal region may not reflect the situation of public hospitals in underdeveloped areas in China. Wider collaboration with multiple hospitals in a representative, nationwide, multicenter study is needed in the future. Second, perspectives from the demand side—patients receiving online health care—regarding its quality are lacking. In the next phase, more detailed analysis on online health care quality from both demand and supply sides will be conducted.

Conclusions

Telemedicine services such as online regular follow-up visit services provided by internet hospitals must also strictly adhere to fundamental medical principles of diagnosis, prescription, and treatment. Therefore, the ability of internet hospitals to overcome the constraints of time and space depend upon the principles of disease diagnosis. The initial face-to-face consultation does not overcome the time and space constraints because of testing, but the subsequent regular online follow-up visit service does. Different diseases necessitate different diagnostic and therapeutic approaches. For chronic illnesses or chronic phases of disease, the patient's condition often changes relatively little over time. These stable cases with minimal fluctuation are well-suited for regular online follow-up visit services. Overall, for patients with chronic diseases, internet hospitals demonstrate significant advantages, both in accessibility and cost. Although physicians utilize various methods to ensure the quality of online health care, the online regular follow-up visit service cannot fully replace in-person medical services.

Acknowledgments

We would like to thank the 12 interviewees in TAHSYSU who participated in the study. In addition, this study was not supported by any funds.

Authors' Contributions

KW and WZ are joint first authors and contributed equally to this work. They conceived the study and drafted the manuscript. YL, CH, and NL conducted the data acquisition, provided statistical advice, and reviewed and revised the manuscript. XL, XL, TL, and XY were responsible for data collection and statistical analysis. SW and CZ are the corresponding authors of this article; they conceptualized the research; participated in its design and coordination; and assisted in drafting, reviewing, and revising the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of patients across geographies and hospital departments.
[DOCX File , 31 KB - jmir_v26i1e54902_app1.docx]

Multimedia Appendix 2
Interview transcripts.
[DOCX File , 23 KB - jmir_v26i1e54902_app2.docx]

References

1. Sood S, Mbarika V, Jugoo S, Dookhy R, Doarn CR, Prakash N, et al. What is telemedicine? A collection of 104 peer-reviewed perspectives and theoretical underpinnings. Telemed J E Health 2007 Oct;13(5):573-590. [doi: 10.1089/tmj.2006.0073] [Medline: 17999619]
2. Jagarapu J, Savani RC. A brief history of telemedicine and the evolution of teleneonatology. Semin Perinatol 2021 Aug;45(5):151416. [doi: 10.1016/j.semperi.2021.151416] [Medline: 34006382]
3. Riva G, Gamberini L. Virtual reality in telemedicine. Telemed J E Health 2000 Sep;6(3):327-340. [doi: 10.1089/153056200750040183] [Medline: 11110636]

XSL•FO
RenderX

4. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
5. Tian W. [Exploration and prospect of 5G application in telemedicine]. Zhonghua Wai Ke Za Zhi 2020 Jan 01;58(1):1-4. [doi: 10.3760/cma.j.issn.0529-5815.2020.01.001] [Medline: 31902160]
6. Rowan C, Hirten R. The future of telemedicine and wearable technology in IBD. Curr Opin Gastroenterol 2022 Jul 01;38(4):373-381. [doi: 10.1097/MOG.0000000000000845] [Medline: 35762696]
7. Zheng J, Wang Y, Zhang J, Guo W, Yang X, Luo L, et al. 5G ultra-remote robot-assisted laparoscopic surgery in China. Surg Endosc 2020 Nov 22;34(11):5172-5180. [doi: 10.1007/s00464-020-07823-x] [Medline: 32700149]
8. Jiang X. Reshaping the healthcare landscape: a brief history of China's mobile healthcare development. Internet Economy 2017:92-99. [doi: 10.19609/j.cnki.cn10-1255/f.2017.z2.019]
9. Circular of the Chinese Medicine Bureau on the Issuance of Three Documents, Including the Measures for the Administration of Internet Diagnosis and Treatment (for Trial Implementation). National Health and Medical Administration. 2018. URL: https://www.gov.cn/gongbao/content/2019/content_5358684.htm [accessed 2023-08-22]
10. Kitamura C, Zurawel-Balaura L, Wong RKS. How effective is video consultation in clinical oncology? A systematic review. Curr Oncol 2010 Jun;17(3):17-27 [FREE Full text] [doi: 10.3747/co.v17i3.513] [Medline: 20567623]
11. Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine during the COVID-19 pandemic: experiences from western China. J Med Internet Res 2020 May 08;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]
12. Anand S, Fan VY, Zhang J, Zhang L, Ke Y, Dong Z, et al. China's human resources for health: quantity, quality, and distribution. The Lancet 2008 Nov;372(9651):1774-1781. [doi: 10.1016/s0140-6736(08)61363-x]
13. Wan S, Chen Y, Xiao Y, Zhao Q, Li M, Wu S. Spatial analysis and evaluation of medical resource allocation in China based on geographic big data. BMC Health Serv Res 2021 Oct 12;21(1):1084 [FREE Full text] [doi: 10.1186/s12913-021-07119-3] [Medline: 34641850]
14. Wang T, Li J, Zhu C, Hong Z, An D, Yang H, et al. Assessment of utilization and cost-effectiveness of telemedicine program in western regions of China: a 12-year study of 249 hospitals across 112 cities. Telemed J E Health 2016 Nov;22(11):909-920. [doi: 10.1089/tmj.2015.0213] [Medline: 27314300]
15. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015 Aug;3(8):e445-e446 [FREE Full text] [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
16. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020 Jul 29;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
17. The Second People's Hospital of Guangdong Province. Sending Experts to the Village Entrance: Internet Hospitals. 2016 Sep 2. URL: https://www.gd2h.com/site/detail/38083.html
18. Opinions of the General Office of the State Council on Promoting the Development of "Internet+Medical Health". General Office of the State Council. 2018 Apr 25. URL: https://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm [accessed 2023-08-22]
19. Xie X, Zhou W, Lin L, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017 Jul 04;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]
20. National Health Commission: the country now has 158 Internet hospitals. Xinhua News Agency. 2019 May 08. URL: https://www.gov.cn/xinwen/2019-05/08/content_5389676.htm [accessed 2023-08-22]
21. Xiaohang T. Rapid development! China's Internet hospitals have reached more than 1,600. Xinhua News Agency. 2021 Aug 23. URL: https://www.gov.cn/xinwen/2021-08/23/content_5632844.htm [accessed 2023-08-22]
22. Wang W, Sun L, Liu T, Lai T. The use of E-health during the COVID-19 pandemic: a case study in China's Hubei province. Health Sociol Rev 2022 Nov;31(3):215-231 [FREE Full text] [doi: 10.1080/14461242.2021.1941184] [Medline: 34161186]
23. Sun S, Yu K, Xie Z, Pan X. China empowers Internet hospital to fight against COVID-19. J Infect 2020 Jul;81(1):e67-e68 [FREE Full text] [doi: 10.1016/j.jinf.2020.03.061] [Medline: 32251688]
24. There are 7,700 hospitals above the second level in the country providing online services Internet healthcare, making patients easier to see a doctor. People's Daily. 2021 Mar 26. URL: https://www.gov.cn/xinwen/2021-03/26/content_5595808.htm [accessed 2023-08-22]
25. Ding L, She Q, Chen F, Chen Z, Jiang M, Huang H, et al. The internet hospital plus drug delivery platform for health management during the COVID-19 pandemic: observational study. J Med Internet Res 2020 Aug 06;22(8):e19678 [FREE Full text] [doi: 10.2196/19678] [Medline: 32716892]
26. Gong K, Xu Z, Cai Z, Chen Y, Wang Z. Internet hospitals help prevent and control the epidemic of COVID-19 in China: multicenter user profiling study. J Med Internet Res 2020 Apr 14;22(4):e18908 [FREE Full text] [doi: 10.2196/18908] [Medline: 32250962]
27. Xu L. Internet Hospital: Challenges and Opportunities in China. In: Yin X, Geller J, Li Y, Zhou R, Wang H, Zhang Y, editors. Health Information Science. HIS 2016. Lecture Notes in Computer Science(), vol 10038. Cham, Switzerland: Springer; 2016:85-90.
28. Ge F, Qian H, Lei J, Ni Y, Li Q, Wang S, et al. Experiences and challenges of emerging online health services combating COVID-19 in China: retrospective, cross-sectional study of internet hospitals. JMIR Med Inform 2022 Jun 01;10(6):e37042 [FREE Full text] [doi: 10.2196/37042] [Medline: 35500013]

XSL•FO
RenderX

29. Li Y, Hu H, Rozanova L, Fabre G. COVID-19 and internet hospital development in China. Epidemiologia (Basel) 2022 Jun 02;3(2):269-284 [FREE Full text] [doi: 10.3390/epidemiologia3020021] [Medline: 36417257]
30. Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. World Health Organization. 2010. URL: https://iris.who.int/handle/10665/258734 [accessed 2023-08-22]
31. Andersen RM, Davidson PL. Improving Access to Care in America: Individual and Contextual Indicators. In: Andersen RM, Rice TH, Kominski GF, editors. Changing the U.S. health care system: Key issues in health services policy and management. San Francisco, CA: Jossey-Bass; 2007:3-31.
32. The world health report 2000. World Health Organization. 2000 Jun 14. URL: https://www.who.int/publications/i/item/924156198X [accessed 2024-09-30]
33. Xu X, Cai Y, Wu S, Guo J, Yang L, Lan J, et al. Assessment of internet hospitals in China during the COVID-19 pandemic: national cross-sectional data analysis study. J Med Internet Res 2021 Jan 20;23(1):e21825 [FREE Full text] [doi: 10.2196/21825] [Medline: 33417586]
34. Qi H, Ying Y, Zhu L, Li Q, Wang T, Chen B, et al. Exploration on the development of public hospital-sponsored telemedicine platform: A case study in China. J Telemed Telecare 2023 Jun 13:1. [doi: 10.1177/1357633x231176871]
35. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q 1966 Jul;44(3):Suppl:166-Suppl:206. [Medline: 5338568]
36. Chen Z, Liang W, Liu Q, He R, Chen Q, Li Q, et al. Use of a remote oncology pharmacy service platform for patients with cancer during the COVID-19 pandemic: implementation and user acceptance evaluation. J Med Internet Res 2021 Jan 21;23(1):e24619. [doi: 10.2196/24619] [Medline: 33395398]

Abbreviations

API: application programming interface
TAHSYSU: Third Affiliated Hospital of Sun Yat-sen University
WHO: World Health Organization

Edited by G Eysenbach; submitted 05.12.23; peer-reviewed by L Wreyford; comments to author 15.04.24; revised version received 10.07.24; accepted 05.09.24; published 21.10.24.

Please cite as:
Wang K, Zou W, Lai Y, Hao C, Liu N, Ling X, Liu X, Liu T, Yang X, Zu C, Wu S
Accessibility, Cost, and Quality of an Online Regular Follow-Up Visit Service at an Internet Hospital in China: Mixed Methods Study
J Med Internet Res 2024;26:e54902
URL: https://www.jmir.org/2024/1/e54902
doi:10.2196/54902
PMID:39432365

©Kun Wang, Wenxin Zou, Yingsi Lai, Chun Hao, Ning Liu, Xiang Ling, Xiaohan Liu, Ting Liu, Xin Yang, Chenxi Zu, Shaolong Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Implementation and Clinical Effectiveness of an Innovative Digital First Care Model for Behavioral Health Using the RE-AIM Framework: Quantitative Evaluation

Samuel S Nordberg[1*], PhD; Brittany A Jaso-Yim[1*], PhD; Pratha Sah[1], PhD; Keke Schuler[1], PhD; Mara Eyllon[1], PhD; Mariesa Pennine[1], BA; Georgia H Hoyler[2], BS; J Ben Barnes[1,3], PhD; Lily Hong Murillo[1], BA; Heather O'Dea[1]; Laura Orth[1], MSW; Elizabeth Rogers[1], MA; George Welch[1], MM; Gabrielle Peloquin[1], BS; Soo Jeong Youn[1,4], PhD

[1]Reliant Medical Group, OptumCare, Worcester, MA, United States

[2]United Health Group, Minnetonka, MN, United States

[3]University of Massachusetts Chan School of Medicine, Worcester, MA, United States

[4]Harvard Medical School, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Samuel S Nordberg, PhD
Reliant Medical Group, OptumCare
5 Neponset St.
Worcester, MA, 01606
United States
Phone: 1 5088560732
Email: Samuel.Nordberg@reliantmedicalgroup.org

Abstract

Background: In the United States, innovation is needed to address the increasing need for mental health care services and widen the patient-to-provider ratio. Despite the benefits of digital mental health interventions (DMHIs), they have not been effective in addressing patients' behavioral health challenges as stand-alone treatments.

Objective: This study evaluates the implementation and effectiveness of precision behavioral health (PBH), a digital-first behavioral health care model embedded within routine primary care that refers patients to an ecosystem of evidence-based DMHIs with strategically placed human support.

Methods: Patient demographic information, triage visit outcomes, multidimensional patient-reported outcome measure, enrollment, and engagement with the DMHIs were analyzed using data from the electronic health record and vendor-reported data files. The RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) framework was used to evaluate the implementation and clinical effectiveness outcomes of PBH.

Results: PBH had a 47.58% reach rate, defined as patients accepting the PBH referral from their behavioral health integrated clinician. PBH patients had high DMHI registration rates (79.62%), high activation rates (76.54%), and high retention rates at 15 days (57.69%) and 30 days (44.58%) compared to literature benchmarks. In total, 74.01% (n=168) of patients showed clinical improvement, 22.47% (n=51) showed no clinical change, and 3.52% (n=8) showed clinical deterioration in symptoms. PBH had high adoption rates, with behavioral health integrated clinicians referring on average 4.35 (SD 0.46) patients to PBH per month and 90%-100% of clinicians (n=12) consistently referring at least 1 patient to PBH each month. A third (32%, n=1114) of patients were offered PBH as a treatment option during their triage visit.

Conclusions: PBH as a care model with evidence-based DMHIs, human support for patients, and integration within routine settings offers a credible service to support patients with mild to moderate mental health challenges. This type of model has the potential to address real-life access to care problems faced by health care settings.

(J Med Internet Res 2024;26:e54528) doi:10.2196/54528

KEYWORDS

digital mental health interventions; implementation; clinical effectiveness; practice-oriented research; access to care

XSL•FO
RenderX

Introduction

Background

There is an increasing need for behavioral health services in the United States. It is estimated that 57.8 million, or 1 in 5 adults, in the United States experience mental health challenges [1], with some reports indicating challenges for 1 in 3 adults [2]. Despite this growing need for behavioral health services, it is well established that the United States will never be able to train enough behavioral health specialists to address the needs of the population adequately [3]. By 2024, the United States is expected to have a shortage of over 10,000 mental health care professionals [4] and will only have one provider available for every estimated 350 individuals with behavioral health needs [3]. Untreated behavioral health challenges have multi-level consequences. Mental health conditions have serious consequences for individuals' health and are considered important risk factors for excess morbidity, including cardiovascular disease and diabetes [5,6], as well as premature mortality [7]. Further, individuals with untreated depression and anxiety are more likely to experience lost productivity in the workplace and produce lower-quality work [8]. Health care providers report increased burden and burnout as they address patients' behavioral health needs without sufficient training, fewer opportunities for professional development, and facing competing demands between meeting productivity metrics and providing appropriate clinical care [9,10]. This gap in meeting behavioral health needs is also estimated to cost the United States over US $300 billion annually in overall lost productivity [11]. Thus, given the magnitude of consequences related to the access to care problem in behavioral health, it is imperative to find innovative and effective solutions.

Digital Mental Health Interventions

Digital mental health interventions (DMHIs) have been proposed as a solution to mitigate the supply and demand gap. DMHIs are defined as "discrete digital functionalities of technology used to achieve health sector objectives, […] designed to achieve a specific outcome" [12], and can be delivered via mobile apps, web-based programs, wearables, and virtual reality headsets [13]. DMHIs are well poised to equip patients with needed behavioral health interventions and have the potential to reach more people for minimal additional cost per user relative to traditional modalities [14]. These tools are highly flexible and customizable, serving a wide variety of indications, from depression and anxiety to more serious mental illnesses such as bipolar disorder [14-16]. Further, they are more convenient for patients than traditional mental health care, removing scheduling and transportation barriers, thus allowing patients to use them where and when they want [17]. Lastly, DMHIs' content is highly standardized and can, therefore, be delivered with high treatment fidelity [14].

Despite the significant benefits, stand-alone DMHIs have been shown to be insufficient in adequately addressing the demand for mental health services. The implementation of DMHIs has included significant end user adoption challenges, such as low registration, activation, and engagement or completion rates. For example, despite initial interest, only 42% of individuals visiting the website for a popular DMHI subsequently registered [15,18]. Even for those who do register, 70% of these users do not activate (ie, use the DMHI at least once post registration) [19], and low sustained engagement has been deemed the "Achilles heel" of DMHIs [20]. Baumel et al [21] reported that fewer than 4% of users who downloaded mental health apps engaged with the apps for at least 15-30 days. A systematic review of the uptake and use of digital self-help interventions identified that as few as 7% of those who registered with a DMHI had at least moderate engagement, and less than 1% completed the entire program [22].

A variety of barriers impede the adoption of DMHIs. Given there are between 10,000 and 20,000 DMHIs currently available [14,23,24], sorting through the options to select an appropriate DMHI may be challenging for users without input from a mental health clinician. Additionally, limited digital literacy has been shown to inhibit uptake and prolonged use [25,26]. Further, users seeking mental health treatment have expressed concerns about the efficacy of DMHIs [14,27] and indicated significantly greater interest in trying DMHIs that are explicitly recommended by clinicians and mental health care providers compared to those that are not [27]. Thus, there is an opportunity to address many of the current barriers to adopting DMHIs and maximize their benefits by embedding DMHIs within a clinical practice care model. Innovations that include the implementation of DMHIs in the clinic can efficiently increase access to timely and clinically appropriate care without the need for additional providers [28], as well as provide the resources patients need to successfully adopt them.

Innovation: Precision Behavioral Health

Precision behavioral health (PBH) is a digital-first care model in which clinicians and staff support patients in the use of digital interventions drawn from an ecosystem of evidence-based DMHIs as a frontline treatment option. PBH and its implementation have been extensively described elsewhere [29]. Briefly, PBH aims to increase access to quality treatment for patients experiencing mild to moderate anxiety and depression symptoms. To do so, PBH was designed to maximize the benefits of DMHIs while simultaneously addressing the barriers to adoption, all within the context of a primary care-integrated behavioral health setting. The clinical work is wrapped in a practice-research network [30] that allows for its systematic evaluation.

There are several key components to the clinical deployment of PBH. First, PBH includes an ecosystem of DMHIs that have been pre-vetted to be evidence-based and clinically appropriate for the needs of the patients being served [29]. The DMHIs selected for the ecosystem address particular evidence-based mechanisms of change using well-established techniques such as mindfulness, biofeedback, diaphragmatic breathing, cognitive reappraisal, and others. Techniques are delivered through various mediums, including but not limited to virtual reality headsets, phone apps, and inhaler-style devices. Multiple mediums were purposely selected to have a variety of methods that could be matched with patient preferences and needs. The ecosystem of DMHIs has evolved as some DMHIs have been rejected and new ones have been added. At the time of this evaluation, 6

XSL•FO
RenderX

DMHIs were active in the ecosystem. The complete programming length for the DMHIs ranged from 4 to 12 weeks.

Second, PBH is embedded within a broader primary care-integrated behavioral health workflow. Here, patients identified by their primary care providers as needing behavioral health support are handed off to licensed behavioral health clinicians for an assessment and triage visit. During the triage visit, where clinically appropriate, the clinician can make a personalized referral to specific DMHIs that would be best suited for the patient's needs. The triaging clinician also schedules a 4- or 6-week follow-up visit to check the patient's progress and make new recommendations if necessary. Throughout the PBH program, patients' symptoms are monitored at 2-week intervals via a computer adaptive outcomes questionnaire.

Lastly, PBH includes a strategically placed human connection explicitly designed to increase patients' likelihood of registration and engagement with the DMHIs. After triage with a licensed clinician, patients meet with a non-clinical digital care navigator for technical support. This step has been shown to increase the patients' registration rates from 20% to, on average, above 70% [31].

This paper aims to assess the clinical effectiveness and implementation outcomes of PBH. To evaluate whether an innovative care model such as PBH is a viable solution for providing and engaging patients with evidence-based digital mental health care, this paper uses the RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) framework [32] to examine the reach, effectiveness, adoption, implementation, and maintenance of the PBH program.

Methods

Setting

Reliant Medical Group (Reliant) is a large multispecialty group medical practice in central Massachusetts, with an integrated behavioral health department available to provide consultation, triage, and treatment to patients in primary care. The behavioral health department includes a team of behavioral health integrated clinicians who serve as the first point of contact for patients identified by providers as needing integrated care. They receive warm handoffs from primary care providers, conduct crisis consultations, and conduct assessments to triage patients to appropriate behavioral health services. The behavioral health department at Reliant also includes a practice research network established in 2020 to improve patients' access to high-quality care and efficiently generate effective and actionable outcomes to address research, clinical, and business needs [33]. The practice research network is embedded within the department and leverages practice-oriented data generated within the real-world practice setting to address clinically relevant problems through scientifically rigorous methods [30].

Procedure

For a detailed description of the procedure for PBH, see [29]. Briefly, as part of routine clinical care at Reliant, patients referred from primary care for behavioral services meet with a behavioral health integrated clinician for an initial triage assessment. Before this meeting, patients are asked to complete a multidimensional patient-reported outcome measure (Norse Feedback [NF]) through the electronic health record patient portal. Behavioral health integrated clinicians are licensed mental health providers trained to conduct clinical evaluations to determine the course of treatment recommendation for the patients. During the initial triage assessment, the behavioral health integrated clinician meets with the patient to understand their behavioral health challenges and needs and uses their discussion, clinical judgment, the NF results, and other relevant clinical information to determine and subsequently refer patients to appropriate behavioral health treatment options, such as short-term goal-oriented therapy with Reliant providers, long-term therapy in the community, or PBH.

Patients who accept PBH are referred to a specific DMHI within the PBH ecosystem by the behavioral health integrated clinician based on the clinical information and other relevant patient-level characteristics, such as access to Wi-Fi, available privacy level, and patient preference. Patients who accept the DMHI referral are then sent registration information via the electronic health record and are immediately scheduled for a visit with the digital care navigator. The digital care navigator helps the patients register for the DMHI and address any other technical challenges experienced. Patients are outreached a total of 4 times post-referral as needed by the digital care navigator to aid their registration process. If a patient does not follow through with the registration after these outreaches, they receive a final message through the electronic health record portal encouraging them to reach out to their integrated clinician for follow-up.

Patients then continue to complete the NF 2, 4, 6, 9, and 12 weeks post triage as part of the outcome monitoring component of PBH. The NF responses are closely monitored, and behavioral health integrated clinicians outreach patients who show clinically significant deterioration to evaluate further needs and discuss possible treatment changes. At 6 weeks, patients are also scheduled to meet with the behavioral health integrated clinician for a follow-up visit to inform any treatment decisions at that time. Figure 1 shows the NF assessment schedule.

XSL•FO
RenderX

Figure 1. PBH measurement timeline for enrolled patients. Data were extracted from the EHR as well as from DMHI vendors from December 1, 2022 through July 31, 2023. PBH: precision behavioral health; EHR: electronic health data; DMHI: digital mental health intervention.

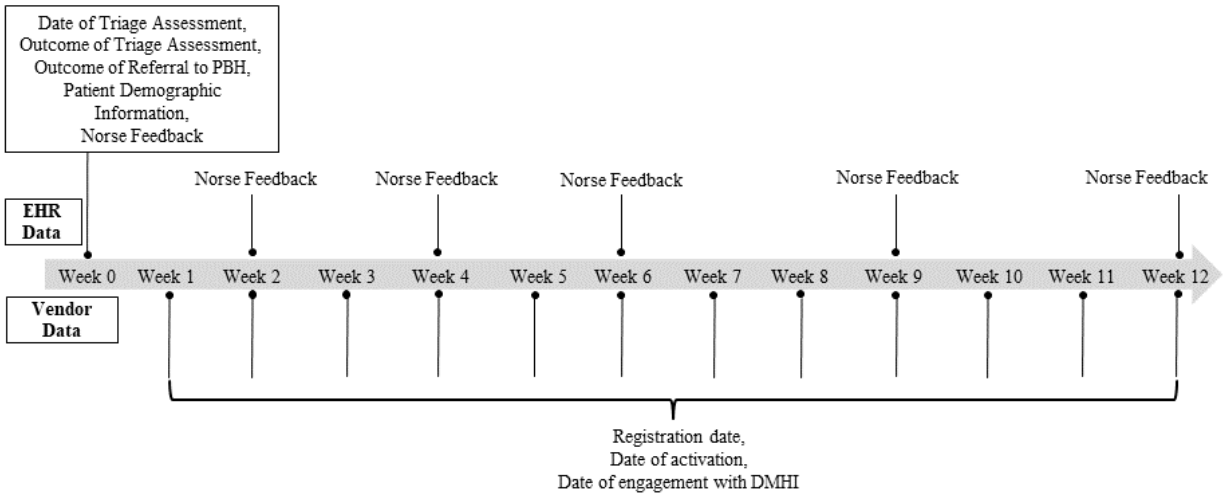

Ethical Considerations

The implementation and evaluation of PBH was determined to be an exempt quality improvement project by the Institutional Review Board at the Office of Human Research Affairs in United Health Group prior to data collection (Exemption action ID: 2023-0036-01). As part of routine care, referring providers collect verbal consent from patients referred to the PBH program to have the DMHI vendor share deidentified data using a randomly generated 9-digit ID regarding the patients' enrollment, engagement, and clinical outcomes with Reliant. Given that PBH was implemented as part of routine care at Reliant, enrolled patients were not compensated for their participation. Patients were not financially responsible for access to the PBH program components or the DMHI program due to grant funding provided by United Health Group's Strategy and Innovation Office to Reliant [29]. All results are shared at the group level to prevent any identification of individual patients.

Data Collection

Electronic Medical Record System

PBH-related documentation was captured within the electronic medical record system as it was collected as part of standard documentation practices for the behavioral health integrated clinicians. The system recorded the date the behavioral health integrated clinicians met with patients to conduct the triage assessment, the outcome of the assessment (eg, PBH, psychiatry evaluation, goal-oriented therapy, group therapy, etc), and the outcome of referral to PBH (ie, patient accepted PBH referral or patient declined PBH referral). Patient demographic information and the NF responses were also collected and stored in the patient electronic medical record system.

NF

The NF measure [34,35] is a multidimensional patient-adapted self-report questionnaire. The NF was developed by conducting qualitative interviews with patients and clinicians to identify areas of clinical interest [34]. Following item development, the early version of the NF underwent evaluation in both clinical and nonclinical samples using item response theory, and items were removed or edited as necessary [34,35]. Each administration begins with "trigger items" or a single item representing an entire subscale; if patients score above a threshold on that trigger item, the remaining subscale questions are offered later during the assessment for patient completion. Once the subscale has been activated in an assessment, it remains open, in that all the subscale items are shown to the patient for completion (regardless of trigger score item score) until the average of that subscale for the subsequent 3 NF administrations is below a subclinical closing threshold. Due to the dynamic nature of the questionnaire, the items in the NF can range from 27 to 96 questions. Total scores are not calculated for the NF; instead, each subscale has its own score.

The NF measure includes 22 different subscales covering various domains, such as psychopathology symptoms, chronic maintaining factors, consequences of mental illness, and personal resources. These 22 subscales are anger, cognitive problems, general functioning, hopelessness, internal avoidance, pain, physical anxiety, readiness for change, eating concerns, sad affect, social support, substance use, self-compassion, self-contempt, social avoidance, traumatic memories, urges, worry, and a few single-item questions that address sexuality/sex functioning, physical health, sleep, and self-harm. NF question responses are recorded on a 7-point Likert scale ranging from 1, "*This is not at all true for me*," to 7, "*This is completely true for me*," for all questions included in the analysis. Any subscale whose subscale total score is higher than the trigger score is identified as "clinically elevated" and highlighted in the patient's medical chart for clinical review and appropriate response as needed.

DMHI Vendor Data

The business relationship established between PBH and each DMHI vendor included systematic data collection and sharing from the vendors with PBH. The variables sent by each vendor differed as the data sent were responsive to each DMHI's unique programming. However, PBH asked each DMHI vendor to send standardized variables to evaluate the effectiveness of the PBH

XSL•FO
RenderX

care model as a whole. Specifically, all DMHI vendors sent PBH data related to the patient's registration date and date of activation, which was defined as the date when the patient engaged with the DMHI content for the first time post registration (this could be the same day, or this could be days after registration, or this variable would be missing for patients that registered but never engaged with the DMHI further), as well as the date of engagement with DMHI. We calculated engagement with the DMHI as the number of interactions the patient had with the DMHI per week. Initially, vendors shared data with PBH on weekly cadences, which changed into daily data transfers. Data were shared using a secure web-based file transfer application.

Data Selection

PBH was first integrated within routine clinical care at Reliant on April 25, 2022. The preimplementation phase [36] lasted until December 1, 2022. It focused on customizing the delivery components of PBH, including the gradual rollout of the various DMHIs in the ecosystem; the inclusion of the NF as the main outcome monitoring measure within the behavioral health department in October 2022; the development and refinement of the operational workflows related to the integration of PBH into routine clinical care; the implementation of extensive DMHIs and PBH-related training for the behavioral health–integrated clinicians; and the establishment of the infrastructure needed for data capture, ensuring data quality, and workflow support within the electronic health record system. Thus, the data pertaining to the evaluation of PBH's implementation outcomes and effectiveness include patients with a baseline and posttreatment assessment completed from December 1, 2022, through July 31, 2023 (date data pulled for this evaluation), during the implementation phase [36].

Demographic information for patients enrolled until November 30, 2022, is included in Table 1, and for patients enrolled after December 1, 2022, their demographic characteristics are summarized in Table 2. To evaluate whether there were meaningful differences in the patients enrolled during the preimplementation and implementation phases, we conducted chi-square tests, Fisher exact test, and the Welch 2-sample *t*-test. The results showed that there were no significant differences in the 2 samples' legal sex (χ^2_2[N=1024]=2.4, P=.30) race (P=.20), and ethnicity (χ^2_2[N=1024]=1.5, P=.50). Significant differences were observed for age (t_{1019}=2.28, P=.02), which implies that among the groups included in our analyses, patients who were offered and accepted the PBH referral after December 1, 2022, were older (mean 41.59, SD 16.04 years) than those who were offered and accepted the referral before November 30, 2022 (mean 37.88, SD 14.42 years).

Table 1. Demographics of patients (N=494) who accepted the PBH[a] referral from April 25, 2022, through November 30, 2022.

Variable	Values
Age (years), mean (SD)	37.88 (14.42)
Legal sex, n (%)	
Female	364 (73.68)
Male	128 (25.91)
N/A[b]	2 (0.40)
Race, n (%)	
White	332 (67.21)
Black or African American	17 (3.44)
Asian	3 (0.61)
Biracial	14 (2.83)
Native American	2 (0.40)
Multiracial	1 (0.20)
Pacific Islander	0 (0)
N/A	125 (25.30)
Ethnicity, n (%)	
Not Hispanic or Latino	283 (57.29)
Hispanic or Latino	53 (10.73)
Unknown	158 (31.98)

[a]PBH: precision behavioral health.

[b]N/A: not available.

XSL•FO
RenderX

Table 2. Demographics of patients who attended a triage visit through the different phases of PBH enrollment from December 1, 2022, through July 31, 2023.

	Triage visit	PBH offered[a]	PBH declined[b]	PBH accepted[b]	DMHI registered[c]	DMHI activated[d]
Total, n	3457	1114	584	530	422	323
Age (years), mean (SD)	41.59 (16.04)	40.23 (15.10)	40.45 (15.11)	39.99 (15.11)	39.88 (15.36)	40.33 (15.49)
Legal sex, n (%)						
Female	2310 (66.82)	798 (71.63)	399 (68.32)	399 (75.28)	322 (76.30)	252 (78.02)
Male	1144 (33.09)	316 (28.37)	185 (31.68)	131 (24.72)	100 (23.70)	71 (22.18)
Nonbinary	3 (0.09)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Race, n (%)						
White	2470 (71.45)	799 (71.72)	427 (73.12)	372 (70.19)	303 (71.80)	234 (72.45)
Black or African American	131 (3.79)	52 (4.67)	24 (4.11)	28 (5.28)	24 (5.69)	17 (5.26)
Asian	60 (1.74)	17 (1.53)	10 (1.71)	7 (1.32)	4 (0.95)	3 (0.93)
Biracial	54 (1.56)	17 (1.53)	9 (1.54)	8 (1.51)	5 (1.18)	3 (0.93)
Native American	16 (0.46)	5 (0.45)	2 (0.34)	3 (0.57)	3 (0.71)	2 (0.62)
Multiracial	7 (0.20)	2 (0.18)	1 (0.17)	1 (0.19)	1 (0.24)	1 (0.31)
Pacific Islander	4 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
N/A[e]	715 (20.69)	222 (19.93)	111 (19.01)	111 (20.94)	82 (19.43)	63 (19.51)
Ethnicity, n (%)						
Non-Hispanic or Latino	2090 (60.46)	686 (61.58)	367 (62.84)	319 (60.19)	259 (61.37)	194 (60.06)
Hispanic or Latino	340 (9.84)	118 (10.59)	58 (9.93)	60 (11.32)	42 (9.95)	34 (10.53)
N/A	1027 (29.71)	310 (27.83)	159 (27.23)	151 (28.49)	121 (28.68)	95 (29.41)

[a]PBH (precision behavioral health) offered: integrated clinicians believed the patients to be eligible for PBH and offered to refer the patients to the program.

[b]PBH declined/accepted: count of the patients' decision whether to accept or decline the PBH referral from the integrated clinician.

[c]DMHI (digital mental health intervention) registered: patients who were deemed eligible, offered PBH, accepted the referral and signed up using the DMHI's specific registration procedures.

[d]DMHI activated: patients who were deemed eligible, offered PBH, accepted the referral, signed up using the DMHI's specific registration procedures, and engaged with their DMHI at least once.

[e]N/A: not available.

Data Analysis

To evaluate the outcomes of PBH, we used the RE-AIM framework [32] to assess the relevant implementation and effectiveness outcomes [29].

Reach was determined as the proportion of patients that accepted the referral made by the behavioral health integrated clinician. To further characterize the reach patients for PBH in the context of all patients being referred for behavioral health services, we summarized and compared the demographic characteristics of patients who attended a behavioral health triage assessment visit, patients who were referred to PBH, patients who accepted the PBH referral, and patients who were identified as eligible for PBH by their behavioral health integrated clinician but the patient decided against participation.

The *effectiveness* of the PBH program was assessed in two ways. First, we analyzed the patients' behaviors as they related to the DMHIs. We summarized the patients' DMHI registration rates post referral, their activation rates (defined as patients who engaged at least once with the DMHI post-registration [19]), and retention rates at 15 and 30 days (defined as the percentage of patients who engaged with the DMHI at least once weekly out of the number of activated patients). Second, we evaluated patient-level clinical outcomes of interest. PBH's clinical response rate was analyzed as the number of patients classified as having clinically improved, clinically worsened, or shown no clinical improvement in the NF measure [34]. To assess

XSL•FO
RenderX

clinical change, we calculated the amount of change observed between the baseline and 6-week follow-up NF assessments for each of the NF subscales that were clinically elevated during baseline. The 6-week follow-up NF assessment, or the closest NF assessment to this time frame after 6 weeks, was selected as it was reflective of the time for most patients to have had access to half of the program for most of the digital interventions in the ecosystem. Following the definition by Hiller and colleagues [37], for a patient to be categorized as having clinically improved, they had to have achieved more than 50% improvement from baseline to 6-week follow-up NF scores in the pathological range, the follow-up NF score had to be at least 25% lower than the baseline score, and these criteria had to be met on at least 25% of the baseline elevated NF subscales. For a patient to be categorized as having clinically deteriorated, they had to show more than 50% symptom deterioration on the NF scores from baseline to 6-week follow-up in the pathological range, the follow-up NF score had to be at least 25% higher than the baseline score, and these criteria had to be met on at least 25% of the baseline elevated NF subscales. If a patient was not categorized as either clinically improved or clinically worsened, they were categorized as showing no clinical change. Detailed information on the categorization development can be found in Multimedia Appendix 1.

Given that PBH was implemented as part of routine care within the behavioral health department at Reliant, all behavioral health integrated clinicians *adopted* PBH [29]. Behavioral health integrated clinicians met biweekly to discuss PBH, and the DMHIs within the ecosystem, and share patient cases that were referred versus not referred to PBH to support the continued adoption of PBH. To further characterize the adoption behaviors of the providers, we tracked the average number of referrals made by the behavioral health integrated clinicians to PBH each month, as well as the percent of clinicians who made 1, 2, and 3 patient referrals to PBH per month among the 10-12 active clinicians.

Implementation was evaluated at the behavioral health integrated clinician level. We assessed the total number of patients referred to PBH out of those who attended a triage assessment visit.

Maintenance was defined as the extent to which PBH became a part of routine practice at Reliant, in addition to any long-term effects of PBH after it concluded [32]. Since PBH is still in its implementation phase [36] as part of clinical care at Reliant, we are unable to evaluate the maintenance of PBH at this time.

Results

Reach

The proportion of patients reached by the PBH program was 47.58% (530 patients who accepted the PBH referral/1114 patients referred to PBH by an integrated clinician). Table 2 shows the demographic characteristics of patients who attended a behavioral health triage assessment, patients who were referred to PBH, patients who accepted the PBH referral, patients who declined the PBH referral, patients who successfully registered, and patients who activated their DMHI. Chi-square tests and a 1-way ANOVA were used to assess significant demographic differences between the patient groups in Table 2. Our results indicate significant differences in legal sex (χ^2_5[N=6427]=42.6, P<.01). It should be noted that nonbinary sex was excluded from these results due to its absence in all groups except for patients who attended a triage visit. Logistic regression with estimated marginal means was used to further investigate the differences between the groups for legal sex. Results indicate that the female-to-male ratio of patients offered PBH was significantly higher (odds ratio 1.25, 95% CI 1.01-1.55) than the female-to-male ratio at triage, indicating that females were offered PBH more frequently than males. No additional significant differences between groups were found. Results were nonsignificant for ethnicity (χ^2_{10}[N=6430]=3.9, P=.90) and race (χ^2_{35}[N=6430]=16.4, P=.99). One-way ANOVA was significant for age ($F_{5,6424}$=3206, P<.05). Post hoc Bonferroni corrections did not yield any significant differences between the groups, likely due to a lack of sensitivity compared to the 1-way ANOVA.

Effectiveness

In terms of assessing PBH patients' behaviors with their referred DMHIs, among patients referred to PBH who accepted the referral (n=530), 79.62% (n=422) of patients who accepted the PBH referral continued on to register with their DMHI. Of those registered patients, 76.54% (n=323) activated their DMHI. Patient retention rates at 2 and 4 weeks were 57.69% and 44.58%, respectively.

For the clinical change analysis, 9 (2.79%) DMHI-activated patients did not have any clinical elevations in any of the NF subscales at baseline, and, thus, were not included. Additionally, 87 (26.93%) patients were excluded from the analysis because they lacked posttreatment assessment data. Demographic descriptions and the Welch t test and chi-square comparisons between these excluded patients and the included patients are found in Multimedia Appendix 2. The results show no significant differences in age (t_{130}=0.44, P=.66); legal sex (χ^2_1[N=410]=0.1, P=.78); race (P=.11); or ethnicity (χ^2_2[N=410]=0.4, P=.83) between the 2 groups. More than two-thirds of the remaining 227 patients with elevated NF baseline scores showed clinical improvement (n=168, 74.01%), 22.47% (n=51) of the patients showed no change in clinical symptoms, and 3.52% (n=8) of the patients showed deterioration in clinical symptoms.

Adoption

The average number of referrals made by behavioral health integrated clinicians to PBH each month over the course of the analytic period is illustrated in Figure 2. On average, the integrated clinicians made 5.33 (SD 1.17) PBH referrals per month across the study period. There was an average of 4.35 (SD 0.46) PBH referrals per month per clinician from December 2022 to March 2023. The average number of referrals per month increased to 6.78 (SD 0.21) per clinician between April 2023 and June 2023. In July 2023, there was a slight decrease in the average number of referrals per month to 4.92 (SD 3.60). Because the lowest mean number of referrals was three throughout the study period, Figure 3 shows the percentage of

XSL•FO
RenderX

clinicians (n=12) who made 1, 2, and 3 referrals each month. The percentage of clinicians who made at least 1 referral for any given month ranged from 90% to 100%; 72% to 100% for 2 referrals each month; and 58%-91% for 3 referrals.

Figure 2. Average number of PBH referrals made by behavioral health integrated clinicians per month from December 1, 2022, through July 31, 2023a. PBH: precision behavioral health; aError bars: standard error.

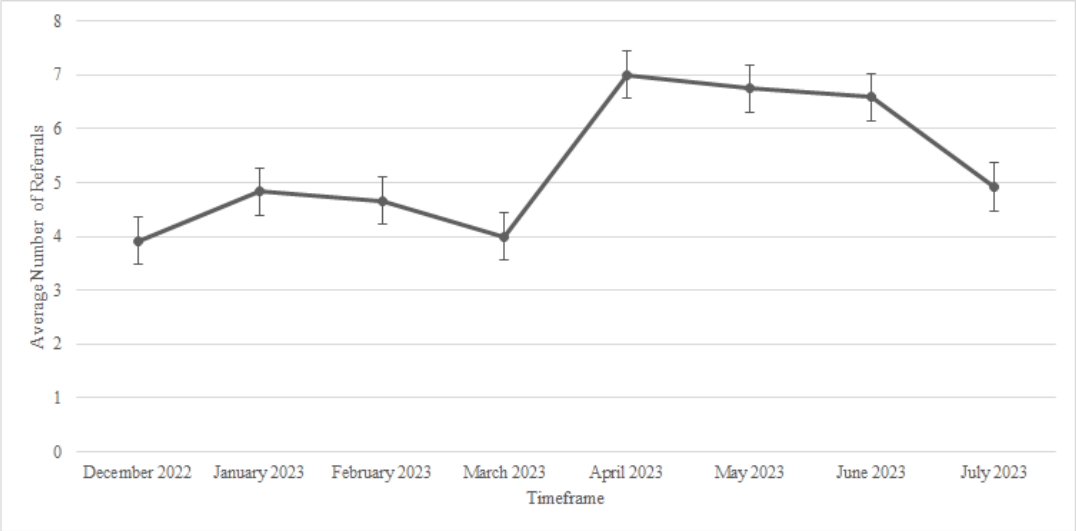

Figure 3. Bar graph representing the percent of integrated clinicians who made 1, 2, and 3 referrals to PBH each month from December 1, 2022, through July 31, 2023. PBH: precision behavioral health.

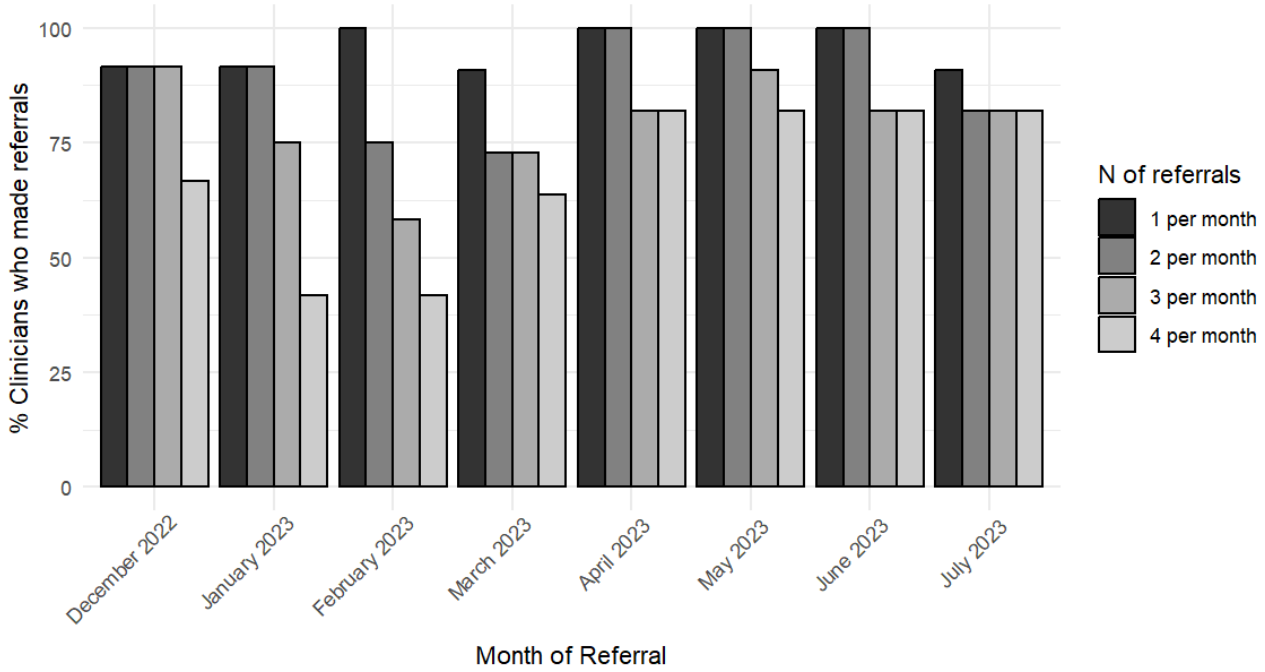

Implementation

During the analytic period, a total of 3457 patients attended a behavioral health triage visit, and 1114 (32%) were offered PBH (Table 2).

Discussion

To adequately address the access to care problem in behavioral health, we need innovative and scalable solutions that will have a meaningful impact on patients and health care organizations. Thus, this paper aimed to evaluate the implementation and effectiveness outcomes of PBH, a digital-first behavioral health care model that was integrated as part of routine clinical care, with an ecosystem of evidence-based DMHIs as treatment options and strategically placed human support for patients. Given the significant differences between naturalistic studies and randomized controlled trials in both engagement and effectiveness of DMHIs [38,39], the focus of our study was to assess the effectiveness of PBH as delivered in routine care to adequately understand its real-world impact and applications.

The implementation outcomes results show that almost 50% of patients were reached, defined as patients accepting the referral to PBH as a behavioral health treatment option. Despite some research showing that patients may be hesitant to accept digital

XSL•FO
RenderX

tools as valid mental health options [26], our results provide support for embedding DMHIs within a care model, with provider referral possibly addressing patient-level barriers and increasing reach rates [27]. There were significantly more females referred to PBH compared to males. This could reflect the fact that females are more represented in behavioral health patient populations [40] or that there was a provider bias present in the patients that were referred to PBH. Provider adoption rates were high. The results show that the behavioral health integrated clinicians, on average, referred more than 4 patients per month and that a high percentage of clinicians referred 2-3 patients in a given month. The current model of having providers adopt innovation in the context of routine clinical care and a practice research network shows that practice-oriented research [30] is a complementary paradigm to implementation efforts in increasing provider adoption rates [41]. The implementation outcomes further support the adoption results. The behavioral health integrated clinicians conducting the triage visit offered PBH to 1 out of 3 patients attending these visits as an alternative treatment option. These results show that innovative clinical care models such as PBH can bridge the gap in mental health care by expanding treatment pathways to include evidence-based DMHIs. Patients getting diverted from traditional psychotherapy referrals and into these types of digital interventions means that they will access treatment right away while freeing up therapy resources for patients that may be better suited for traditional treatment modalities.

The results of the clinical effectiveness evaluation further support PBH as an effective treatment option for patients. More than 70% of PBH patients showed clinical-level improvement rates, as compared to the 38% clinical improvement rates found in psychotherapy delivered in routine clinical care settings [42]. It should be noted that the 3.5% of PBH patients that showed clinical-level deterioration of symptoms is lower than the 5% deterioration rates found in routine psychotherapy [40]. However, given the small sample size of deteriorating patients in the program (n=8), these results should be interpreted with caution. PBH patients' usage of the DMHIs was also high; almost 80% of PBH patients registered with their DMHI, compared to 42% reported in previous studies [15,18], 76% activated their DMHI post registration, which is higher than the 30% found in the literature [19], and patient retention rates were almost 58% and 45% at 15 days and 30 days, compared to 4% and 3%, respectively, as assessed by the systematic review by Baumel et al [21]. The PBH program's inclusion of a digital care navigator, deployed to assist patients with their DMHIs post-referral from a clinician, has shown to increase registration success rates significantly [31]. Additionally, we hypothesize that the follow-up check-in visit with the behavioral health integrated clinician may aid PBH's high retention rates. Patients may feel a sense of accountability as well as support throughout their DMHI journey since they have a dedicated, scheduled visit with their referring provider to review treatment progress or discuss alternative treatment options.

The results of the study should also be considered in the context of a few limitations. About 27% of patients who had activated their DMHIs were not included in the effectiveness analysis because of a lack of post-outcome measure completion. This missing rate is comparable to other data collected in naturalistic settings [43], and there were no differences between patients who registered compared to those that activated. Nonetheless, the effectiveness results reported should be interpreted with this limitation. Additionally, this study focused on overall clinical improvement, regardless of DMHI usage. There is evidence to suggest that differential usage patterns of DMHIs lead to different outcomes in depression and anxiety, with some studies showing that more engagement is related to improved outcomes [44], whereas others not finding a relationship between engagement and effectiveness [45]. Thus, future studies should examine whether there are patterns of DMHI usage within the PBH ecosystem and how they relate to clinical improvement. Additionally, how DMHI usage is defined has varied in the field; some focused on daily usage, whereas others on weekly [46]. This variability is expected as DMHIs have different program components and suggested frequency of engagement. This study standardized these definitions within the DMHIs in the PBH ecosystem to define engagement as at least once a week engagement, which is consistent with how it has been done in the field. Future studies should explore the impact that differing ways of defining engagement have on overall metrics as well as outcomes. Lastly, this study did not evaluate patient-level characteristics that may have impacted them agreeing or declining PBH or registering with the DMHI post-referral. It would be beneficial for future investigations to focus on understanding what are patient-level factors that impact their decision to accept and start engaging with DMHIs.

Overall, the implementation and effectiveness results for PBH provide evidence for implementing these kinds of solutions in actual care delivery organizations. Chief medical officers or other leadership staff of health care settings would be equipped with the information they need to evaluate whether PBH and related programs could address the behavioral health access to care problems they are challenged with day in and day out. PBH's reach, effectiveness, adoption, and implementation outcome results show that PBH can and should be considered as a referral option for patients in addition to traditional psychotherapy options; the high uptake, retention rates, and effectiveness equivalence to psychotherapy as delivered in routine care provide leadership the metrics that they need to reconsider their existing treatment services and confidently refer patients to innovative and impactful referral options, such as PBH.

Acknowledgments

Funding for the PBH program was provided by United Health Group's Office of Strategy & Innovation. Georgia Hoyler is employed by the study sponsor. Her role in the study included developing and managing contractual partnerships with the digital mental health intervention vendors and providing feedback on the report.

XSL•FO
RenderX

Data Availability

The data analyzed during this study are not publicly available as data were collected as part of routine care provision at Reliant.

Conflicts of Interest

SSN has a financial relationship with Mental Health Informatics, which owns the Norse Feedback measure, a measurement-based care tool that has been integrated within routine care at Reliant Medical Group as part of the precision behavioral health initiative described in this paper. SSN declares a potential conflict of interest. SSN has a plan in place with OptumCare and Reliant Medical Group to monitor that the potential conflict of interest does not impact methods, results, and publications related to the Norse Feedback measure or precision behavioral health. No other authors have a conflict of interest to disclose.

Multimedia Appendix 1
Additional materials.
[DOCX File , 48 KB - jmir_v26i1e54528_app1.docx]

Multimedia Appendix 2
Additional materials 2.
[DOCX File , 40 KB - jmir_v26i1e54528_app2.docx]

References

1. Mental illness. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/statistics/mental-illness#part_2539 [accessed 2023-09-23]
2. Panchal N, Saunders H, Rudoqitz R, Cox C. Implications of COVID-19 for mental health and substance use. KFF. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/ [accessed 2023-03-20]
3. Reinert M, Fritze D, Nguyen T. The State of Mental Health in America 2022. Alexandria, VA: Mental Health America; 2022.
4. U.S. Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Workforce, National Center for Health Workforce Analysis. National Projections of Supply and Demand for Behavioral Health Practitioners: 2013-2025. Rockville, MD: HRSA Health Workforce; 2016.
5. Rotella F, Mannucci E. Depression as a risk factor for diabetes: a meta-analysis of longitudinal studies. J Clin Psychiatry 2013;74(1):31-37. [doi: 10.4088/JCP.12r07922] [Medline: 23419223]
6. Sowden GL, Huffman JC. The impact of mental illness on cardiac outcomes: a review for the cardiologist. Int J Cardiol 2009;132(1):30-37. [doi: 10.1016/j.ijcard.2008.10.002] [Medline: 19004512]
7. Druss BG, Zhao L, Von Esenwein S, Morrato EH, Marcus SC. Understanding excess mortality in persons with mental illness: 17-year follow up of a nationally representative US survey. Med Care 2011;49(6):599-604. [doi: 10.1097/MLR.0b013e31820bf86e] [Medline: 21577183]
8. Sanderson K, Andrews G. Common mental disorders in the workforce: recent findings from descriptive and social epidemiology. Can J Psychiatry 2006;51(2):63-75. [doi: 10.1177/070674370605100202] [Medline: 16989105]
9. Substance Abuse and Mental Health Services Administration (SAMHSA). Substance Abuse and Mental Health Services Administration. Rockville, MD: National Mental Health and Substance Use Policy Laboratory; 2022.
10. Rollins AL, Eliacin J, Russ-Jara AL, Monroe-Devita M, Wasmuth S, Flanagan ME, et al. Organizational conditions that influence work engagement and burnout: a qualitative study of mental health workers. Psychiatr Rehabil J 2021;44(3):229-237 [FREE Full text] [doi: 10.1037/prj0000472] [Medline: 33793289]
11. FY 2018 funding for mental health. National Alliance on Mental Illness. URL: https://www.nami.org/getattachment/Get-Involved/NAMI-National-Convention/Convention-Program-Schedule/Hill-Day-2017/FINAL-Hill-Day-17-Leave-Behind-all-(1).pdf [accessed 2024-09-24]
12. Substance Abuse and Mental Health Services Administration. Digital Therapeutics for Management and Treatment in Behavioral Health. Publication No. PEP23- 06-00-001. Rockville, MD: National Mental Health and Substance Use Policy Laboratory; 2023.
13. Khanna M, Rose R. Digital interventions in mental health. Cogn Behav Pract 2022;29(1):1-3.
14. Schueller SM, Torous J. Scaling evidence-based treatments through digital mental health. Am Psychol 2020;75(8):1093-1104 [FREE Full text] [doi: 10.1037/amp0000654] [Medline: 33252947]
15. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
16. Nicholas J, Larsen ME, Proudfoot J, Christensen H. Mobile apps for bipolar disorder: a systematic review of features and content quality. J Med Internet Res 2015;17(8):e198 [FREE Full text] [doi: 10.2196/jmir.4581] [Medline: 26283290]

XSL•FO
RenderX

17. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
18. Batterham PJ, Neil AL, Bennett K, Griffiths KM, Christensen H. Predictors of adherence among community users of a cognitive behavior therapy website. Patient Prefer Adherence 2008;2:97-105 [FREE Full text] [Medline: 19920949]
19. Agachi E, Bijmolt THA, van Ittersum K, Mierau JO. The effect of periodic email prompts on participant engagement with a behavior change mHealth app: longitudinal study. JMIR Mhealth Uhealth 2023;11:e43033 [FREE Full text] [doi: 10.2196/43033] [Medline: 37166974]
20. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the achilles heel? Front Psychiatry 2022;13:900615 [FREE Full text] [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
21. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
22. Christensen H, Griffiths KM, Korten AE, Brittliffe K, Groves C. A comparison of changes in anxiety and depression symptoms of spontaneous users and trial participants of a cognitive behavior therapy website. J Med Internet Res 2004;6(4):e46 [FREE Full text] [doi: 10.2196/jmir.6.4.e46] [Medline: 15631970]
23. Clay RA. Mental health apps are gaining traction: American Psychological Association 2021 annual trends report.: American Psychological Association URL: https://www.apa.org/monitor/2021/01/trends-mental-health-apps [accessed 2024-09-24]
24. Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. JAMA Psychiatry 2017;74(5):437-438. [doi: 10.1001/jamapsychiatry.2017.0262] [Medline: 28384700]
25. Nouri SS, Avila-Garcia P, Cemballi AG, Sarkar U, Aguilera A, Lyles CR. Assessing mobile phone digital literacy and engagement in user-centered design in a diverse, safety-net population: mixed methods study. JMIR Mhealth Uhealth 2019;7(8):e14250 [FREE Full text] [doi: 10.2196/14250] [Medline: 31469083]
26. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
27. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
28. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
29. Youn S, Jaso B, Eyllon M, Sah P, Hoyler G, Barnes JB, et al. Leveraging implementation science to integrate digital mental health interventions as part of routine care in a practice research network. Adm Policy Ment Health 2024;51(3):348-357. [doi: 10.1007/s10488-023-01292-9] [Medline: 37615809]
30. Castonguay LG, Barkham M, Youn S, Page AC. Practice-based evidence—findings from routine clinical settings. In: Bergin and Garfield's Handbook of Psychotherapy and Behavior Change: 50th Anniversary Edition, 7th ed. Hoboken, NJ: John Wiley & Sons, Inc; 2021:191-222.
31. Jaso-Yim B, Eyllon M, Sah P, Pennine M, Welch G, Schuler K, et al. Evaluation of the impact of a digital care navigator on increasing patient registration with digital mental health interventions in routine care. Internet Interv 2024 Dec;38:100777 [FREE Full text] [doi: 10.1016/j.invent.2024.100777] [Medline: 39410952]
32. What is RE-AIM?. URL: https://re-aim.org/learn/what-is-re-aim/ [accessed 2024-09-24]
33. Barnes BJ. Building a practice research network within an integrated BH department of a large medical practice. 2023 Presented at: Society for Psychotherapy Research 54th International Annual Meeting; 2023 June 21-24; Dublin, Ireland.
34. McAleavey AA, Nordberg SS, Moltu C. Initial quantitative development of the norse feedback system: a novel clinical feedback system for routine mental healthcare. Qual Life Res 2021;30(11):3097-3115 [FREE Full text] [doi: 10.1007/s11136-021-02825-1] [Medline: 33851326]
35. Nordberg SS, McAleavey AA, Moltu C. Continuous quality improvement in measure development: lessons from building a novel clinical feedback system. Qual Life Res 2021;30(11):3085-3096 [FREE Full text] [doi: 10.1007/s11136-021-02768-7] [Medline: 33591432]
36. Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R. Implementing evidence-based interventions in health care: application of the replicating effective programs framework. Implement Sci 2007;2:42 [FREE Full text] [doi: 10.1186/1748-5908-2-42] [Medline: 18067681]
37. Hiller W, Schindler AC, Lambert MJ. Defining response and remission in psychotherapy research: a comparison of the RCI and the method of percent improvement. Psychother Res 2012;22(1):1-11. [doi: 10.1080/10503307.2011.616237] [Medline: 21943215]
38. Granek B, Evans A, Petit J, James MC, Ma Y, Loper M, et al. Feasibility of implementing a behavioral economics mobile health platform for individuals with behavioral health conditions. Health Technol 2021;11(3):505-510. [doi: 10.1007/s12553-020-00492-9]

XSL•FO
RenderX

39. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
40. Terlizzi EP, Zablotsky B. Mental Health Treatment Among Adults: United States. Hyattsville, MD: National Center for Health Statistics; 2019.
41. Youn S, Boswell JF, Douglas S, Harris BA, Aajmain S, Arnold KT, et al. Implementation science and practice-oriented research: convergence and complementarity. Adm Policy Ment Health 2024;51(3):336-347. [doi: 10.1007/s10488-023-01296-5] [Medline: 37646966]
42. Cuijpers P, Karyotaki E, Ciharova M, Miguel C, Noma H, Furukawa TA. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. Acta Psychiatr Scand 2021;144(3):288-299 [FREE Full text] [doi: 10.1111/acps.13335] [Medline: 34107050]
43. Asnaani A, Benhamou K, Kaczkurkin AN, Turk-Karan E, Foa EB. Beyond the constraints of an RCT: naturalistic treatment outcomes for anxiety-related disorders. Behav Ther 2020;51(3):434-446 [FREE Full text] [doi: 10.1016/j.beth.2019.07.007] [Medline: 32402259]
44. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
45. Chien I, Enrique A, Palacios J, Regan T, Keegan D, Carter D, et al. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. JAMA Netw Open 2020;3(7):e2010791 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10791] [Medline: 32678450]
46. Lipschitz JM, Van Boxtel R, Torous J, Firth J, Lebovitz JG, Burdick KE, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022;24(10):e39204 [FREE Full text] [doi: 10.2196/39204] [Medline: 36240001]

Abbreviations

DMHI: digital mental health intervention
NF: Norse Feedback
PBH: precision behavioral health
RE-AIM: Reach, Effectiveness, Adoption, Implementation, Maintenance

Edited by A Mavragani; submitted 13.11.23; peer-reviewed by S Markham, P Baldwin, J Hensel, E Agachi; comments to author 10.03.24; revised version received 12.04.24; accepted 11.09.24; published 30.10.24.

Please cite as:
Nordberg SS, Jaso-Yim BA, Sah P, Schuler K, Eyllon M, Pennine M, Hoyler GH, Barnes JB, Murillo LH, O'Dea H, Orth L, Rogers E, Welch G, Peloquin G, Youn SJ
Evaluating the Implementation and Clinical Effectiveness of an Innovative Digital First Care Model for Behavioral Health Using the RE-AIM Framework: Quantitative Evaluation
J Med Internet Res 2024;26:e54528
URL: https://www.jmir.org/2024/1/e54528
doi:10.2196/54528
PMID:

©Samuel S Nordberg, Brittany A Jaso-Yim, Pratha Sah, Keke Schuler, Mara Eyllon, Mariesa Pennine, Georgia H Hoyler, J Ben Barnes, Lily Hong Murillo, Heather O'Dea, Laura Orth, Elizabeth Rogers, George Welch, Gabrielle Peloquin, Soo Jeong Youn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Clarifying the Concepts of Personalization and Tailoring of eHealth Technologies: Multimethod Qualitative Study

Iris ten Klooster[1], MSc; Hanneke Kip[1,2], PhD; Sina L Beyer[1], BSc; Lisette J E W C van Gemert-Pijnen[1], Prof Dr; Saskia M Kelders[1,3], PhD

[1]Centre for eHealth and Wellbeing Research, Department of Psychology, Health & Technology, University of Twente, Enschede, Netherlands

[2]Department of Research, Stichting Transfore, Deventer, Netherlands

[3]Optentia Research Unit, North-West University, Vanderbijlpark, South Africa

Corresponding Author:
Iris ten Klooster, MSc
Centre for eHealth and Wellbeing Research
Department of Psychology, Health & Technology
University of Twente
Drienerlolaan 5
Enschede, 7522 NB
Netherlands
Phone: 31 620730557
Email: i.tenklooster@utwente.nl

Abstract

Background: Although personalization and tailoring have been identified as alternatives to a “one-size-fits-all” approach for eHealth technologies, there is no common understanding of these two concepts and how they should be applied.

Objective: This study aims to describe (1) how tailoring and personalization are defined in the literature and by eHealth experts, and what the differences and similarities are; (2) what type of variables can be used to segment eHealth users into more homogeneous groups or at the individual level; (3) what elements of eHealth technologies are adapted to these segments; and (4) how the segments are matched with eHealth adaptations.

Methods: We used a multimethod qualitative study design. To gain insights into the definitions of personalization and tailoring, definitions were collected from the literature and through interviews with eHealth experts. In addition, the interviews included questions about how users can be segmented and how eHealth can be adapted accordingly, and responses to 3 vignettes of examples of eHealth technologies, varying in personalization and tailoring strategies to elicit responses about views from stakeholders on how the two components were applied and matched in different contexts.

Results: A total of 28 unique definitions of tailoring and 16 unique definitions of personalization were collected from the literature and interviews. The definitions of tailoring and personalization varied in their components, namely adaptation, individuals, user groups, preferences, symptoms, characteristics, context, behavior, content, identification, feedback, channel, design, computerization, and outcomes. During the interviews, participants mentioned 9 types of variables that can be used to segment eHealth users, namely demographics, preferences, health variables, psychological variables, behavioral variables, individual determinants, environmental information, intervention interaction, and technology variables. In total, 5 elements were mentioned that can be adapted to those segments, namely channeling, content, graphical, functionalities, and behavior change strategy. Participants mentioned substantiation methods and variable levels as two components for matching the segmentations with adaptations.

Conclusions: Tailoring and personalization are multidimensional concepts, and variability and technology affordances seem to determine whether and how personalization and tailoring should be applied to eHealth technologies. On the basis of our findings, tailoring and personalization can be differentiated by the way that segmentations and adaptations are matched. Tailoring matches segmentations and adaptations based on general group characteristics using if-then algorithms, whereas personalization involves the direct insertion of user information (such as name) or adaptations based on individual-level inferences. We argue that future research should focus on how inferences can be made at the individual level to further develop the field of personalized eHealth.

(J Med Internet Res 2024;26:e50497) doi:10.2196/50497

XSL•FO
RenderX

KEYWORDS

eHealth; personalization; tailoring; segmentation; adaptation; interviews; definition

Introduction

Background

eHealth technologies can provide opportunities to overcome the increased burden on health care [1]. For example, they can provide a more cost-efficient approach and improve the quality of care by exploiting the additional capabilities of technology such as continuous monitoring and allowing patients to move in a virtual world [2]. eHealth can be defined as "the use of technology to improve health, well-being, and health care" [2], and it is an umbrella term that encompasses a variety of digital health care technologies, such as mobile health apps, web-based monitoring systems, and web-based interventions [3]. Although eHealth technologies show significant improvements in health and well-being, these improvements are often small [4-7], adherence and engagement are regularly low [8], and the effectiveness seems to decline in the long term. This suggests that the full potential of eHealth technologies has not been realized. Because eHealth technologies that adapt to individuals are associated with more effective interventions [9-11], a possible explanation for the suboptimal effectiveness is that a "one-size-fits-all" approach is not sufficient. This means that the design of an eHealth technology needs to consider the variation in patients' needs, life experiences, and other factors [12]. For example, eHealth technologies focused on stress management appeared to be effective for those reporting increased stress, whereas this type of intervention is not effective when used by all workers [5]. While the adaptations of health care are largely based on information about the medical characteristics of the patient (eg, genes) [13], advances in eHealth technologies are creating new opportunities to continuously collect more holistic data about a patient. For example, information on stress levels might be obtained using ecological momentary assessment or wearables connected with the eHealth technology. Changing the eHealth technology according to the individual user is thought to increase effectiveness and implementation. This is referred to as tailoring and personalization in the context of eHealth technologies. Personalization and tailoring seem to be logical ways to overcome suboptimal effectiveness, but there is no clear agreement on how to define personalization and tailoring and what the differences and similarities are. In addition, it remains unclear how personalization and tailoring are being applied and can be applied to eHealth technologies. We need to understand more about personalization and tailoring to maximize their impact and move a step closer to realizing the full potential of eHealth technologies.

Personalization and Tailoring

What tailoring and personalization have in common is that certain adaptations are made for the individual user [14-16], although definitions vary in their operationalization. In the field of health communication, the focus is on presenting information that is perceived as relevant to the user, thereby increasing their engagement and, in turn, leading to improved behavior change outcomes according to the Elaboration Likelihood Model [17]. An example of personalization in this area is the inclusion of the user's name (identification) [18], while a broader definition includes other strategies, such as contextualization (eg, "This information is relevant to mothers.") and raising expectations (eg. "This information is especially designed for you.") [15]. These strategies merely focus on communicating that the information is relevant to the user, whereas the persuasive system design principles operationalize personalization as the actual provision of relevant information [16]. This does not necessarily involve explicitly stating that the information is relevant to the user, but the actual delivery of relevant information may indirectly be perceived as more relevant and reduce the burden on the user by excluding irrelevant information. An example of this operationalization is providing arguments for why eating fruit is important only to the users who have indicated that they do not eat fruits. A more recent example is adapting the mode of delivery of health communication [19].

In addition to health communication, personalization and tailoring are also used in other areas of research. An example is a study on customization and personalization of a dating website [20], where tailoring is operationalized as providing dating recommendations based on one's own qualities (personalization) or explicit preferences (customization). Another example is a study that investigates the effectiveness of web-based advertising when the advertisements are adapted to the preferences compared to location-based adaptations (showing advertisements for nearby businesses). Another example is in e-commerce, where cluster analysis is used to build user profiles that are used for real-time personalization [21].

Tailoring is sometimes described as an umbrella term for personalization, feedback, and content matching [15]. According to this definition, personalization is a form of tailoring. However, tailoring is also defined as a term that is distinct from personalization. Oinas-Kukkonen and Harjumma [16] distinguish personalization and tailoring by considering adaptations to groups of users as tailoring and to individual users as personalization. This contrasts with the definition by Revere and Dunbar [18], which sees tailoring as adaptations to individual users and adaptations to groups of users as a form of targeting eHealth technologies. Overall, we find that several attempts have been made to describe personalization and tailoring unambiguously, but these definitions are sometimes contradictory and the understanding of the differences and similarities between these 2 concepts vary.

Several distinctions are made within the concepts of tailoring and personalization, such as first, second, and third generation; dynamic versus static; deep versus surface; self- versus computer-based; and manual versus algorithm-based. First, distinctions are made that describe how tailoring has evolved over time, referred to as first generation (print), second generation (multimedia), and third generation (mobile and

XSL•FO
RenderX

handheld devices) [22]. Static and dynamic tailoring refers to whether data are collected once for segmentation and adapted accordingly (static) or whether data are collected multiple times over time (dynamic) [10]. The data may be behavioral data or observable factors (surface tailoring), determinants of health behavior (deep tailoring), or a combination of these [23]. A further distinction is whether the user determines how the technology is customized or whether it is determined for the user. The terms personalized, automated, and authoritative refer to technologies that are adapted *for* the user, whereas customization, autonomous, and self-tailoring refer to technologies that are adapted *by* the user [24,25].

Segmentation, Customization, and Matching

As mentioned earlier, it remains unclear how personalization and tailoring are applied to eHealth technologies. To this end, Hawkins et al [15] developed a framework in which the application of personalization and tailoring is described in terms of segmentation and customization. Segmentation is "the degree to which the audience is divided into increasingly more defined, homogenous groups," a concept that originated in marketing [26]. Segmentation takes place based on user characteristics such as lifestyle and the ability to understand health information or health status, creating groups ranging from very small (one person) to large (all women). Furthermore, based on these segmentations, customization of the content or design of an eHealth technology takes place [15]. Customization is described as "the degree to which the messages (ie, a combination of content, source, graphics, channel, etc) that the audience receives reflect relevant individual characteristics" [15]. In this study, we refer to what Hawkins called "customization" as "adaptation" to avoid confusion about how customization is defined in the field of design research and IT [24,27]. A meaningful link between segmentation and adaptation results in a personalized or tailored eHealth technology, which can be described on a continuum of segmentation and adaptation. A limitation of this model is that it only focuses on messages, whereas advances in eHealth technologies expand how they can be adapted. This means that the focus of message adaptation can, and should, be extended to, for example, changing virtual environments and intensity levels of exercise in a gamified sports game. Another limitation of this model is that it does not describe how segmentation and adaptation can be linked in a meaningful way. In this study, we refer to this aspect of creating a meaningful link as "matching."

Aims of This Study

The goals of this study are to describe in detail (1) how tailoring and personalization are defined according to eHealth experts and in literature and what the differences and similarities are, (2) on the basis of what type of variables users are segmented to allow for adapting of eHealth technologies, (3) what elements of eHealth technologies are adapted to those segments, and (4) how the segments are matched with eHealth adaptations.

Methods

Study Design

This study used a multimethod qualitative study design to explore the definitions, segmentation, and adaptation strategies used in personalized and tailored eHealth technologies. The study used semistructured interviews incorporating vignettes with individuals who work in eHealth, alongside an analysis of definitions of personalization and tailoring from references cited in eHealth evaluation studies.

Interviews

Participants

Participants were researchers who were fluent in English, who work or have been working with eHealth technologies, and who published at least 1 scientific article in the field of personalized or tailored eHealth technologies. They were included via purposive and snowball sampling by asking participants to identify researchers with expertise in personalizing and tailoring eHealth technologies. The participants were approached via email and received an information sheet about the interviewer's background and the purpose of the study, including the framework by Hawkins et al [15] to provide context. In total, 10 participants consented to participate in the interviews, of whom there were 4 (40%) men, 2 (20%) women, and 1 (10%) nonbinary participant (n=3, 30% participants did not disclose their gender). Their mean age was 39 years (SD 5.93 y; 4 missing values), and the majority (n=4, 40%) worked as a professor. Other professions were associate professor, professor emeritus, lecturer, and physiologist.

Materials and Procedure

The interviews were developed based on the framework proposed by Hawkins et al [15] and piloted among 2 individuals to ensure clarity of the questions. Feedback from the pilot interviews was used to refine the interview protocol. The interviews took place on the web to adapt to the COVID-19 restrictions, had a duration of around 60 minutes, and were video recorded after receiving permission from the participants. After a short introduction, questions were asked in two parts, namely (1) general questions about personalization and tailoring and (2) questions related to vignettes with 3 examples of eHealth technologies, varying in personalization and tailoring strategies (Multimedia Appendix 1 provides the full interview scheme). General questions were related to definitions of personalization and tailoring (eg, "Do you think there are differences between personalization and tailoring or any other similar terms, such as targeted, individualized, adapted?"), questions about how participants segmented users in their projects (eg, "How would you segment the users of eHealth technologies to provide them with personalized content?"), their experience with adapting eHealth technologies (eg, "How can eHealth technologies be customized to the end user?"), and how these 2 were matched (eg, "How do you think customization and segmentation are related?").

The second part of the interviews consisted of 3 vignettes to stimulate a reaction or opinion of the participant toward the depicted situation on the vignette. The vignette depicted 3

XSL•FO
RenderX

eHealth technologies that differ in their level of segmentation and the way in which the technologies are adapted. The first vignette is from the Brain Aging Monitor study [28], in which users are segmented into lifestyle profiles. Adaptations to those lifestyle profiles include feedback about lifestyle and the inclusion or exclusion of relevant content. The second vignette consists of a virtual reality (VR) behavior therapy for tobacco cessation [29], in which users are segmented into personas based on emotional response data. Adaptations entail the use of either positive (ie, cessation coach) or negative scenarios (ie, receiving a diagnosis of emphysema). The third vignette is a cardiovascular risk calculator [30], in which users are segmented based on several cardiovascular risk factors. The eHealth technology is adapted by indicating their cardiovascular risk and whether they are normal weight, overweight, or obese. The participants were introduced to a vignette and were asked to identify the aspects of the eHealth technology that are personalized to the user as well as how they would further segment and adapt the eHealth technology themselves.

The interviewer (SLB) was a female BSc student who received education in conducting qualitative research. There was no prior relationship established with the participants apart from the recruitment and planning required. The interviews took place with only the interviewer and the interviewee present. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [31] can be found in Multimedia Appendix 2.

Ethical Considerations

Ethics approval for this study was obtained from the ethics committee of the Faculty of Behavioral, Management, and Social Sciences of the University of Twente (request number 210375). All participants signed informed consent before the interviews, and they could opt out at any time. In such cases, no further data would be collected, and any data already collected would be deleted. Before analysis, the data were anonymized, and identifiable information was removed from the transcriptions (eg, when the participant mentioned names). The participants did not receive any compensation for participating in the study.

Literature

Search Strategy

For this study, we used definitions extracted from our earlier comprehensive review of personalization approaches in eHealth evaluation studies. The full review is reported elsewhere [32]. In the systematic review, we conducted an electronic search through the databases Scopus, PubMed, Embase, PsycINFO, and IEEE Xplore. We included peer-reviewed journal articles and conference papers describing personalized eHealth evaluation studies in which participants were randomly assigned to their condition, and outcomes were related to health or well-being. In addition, we only included studies that described how segmentations were matched with adaptations in a computerized way and with full text in German, English, or Dutch. In this study, we specifically used the definitions of personalization and tailoring cited in studies included in our review, along with the original references to these definitions as provided by those included studies. These definitions served as a foundation to explore expert perspectives through qualitative interviews, ensuring a distinct focus that builds on but does not substantially overlap with the previous work.

Data Extraction

Data were extracted from the definitions of tailoring and personalization used in the eHealth evaluation studies that were included. References to their definitions were collected. As a subsequent step, we extracted the definitions from the literature they referred to. By extracting these references, we aimed to identify key elements of definitions that are used in eHealth. References that did not provide a definition were excluded, and as a final step, only the unique definitions were selected.

Data Analysis

The interview data were transcribed manually and anonymized for analysis using Microsoft Excel. Reflexive thematic analysis, as described by Braun and Clarke [33] and further elaborated in their 2023 publication [34], was used to analyze the interview data, incorporating the reactions to the vignettes and the published definitions. This approach emphasizes the active role of the researcher and the interpretative nature of thematic analysis. Our analysis combined deductive and inductive coding strategies. Initially, deductive coding was used to identify fragments related to the research questions—definition, segmentation, adaptation, and matching—based on existing literature and theoretical frameworks. Subsequently, inductive coding allowed the identification of themes that emerged directly from the data without being constrained by predefined categories. Following this, these initial codes were merged into themes, and definitions and names for the themes were generated [35]. This dual approach ensured a comprehensive analysis that incorporated both preexisting knowledge and new insights from the data. To ensure reliability, 1 interview was coded by 2 researchers (HK and ITK), and the interrater reliability was 0.93 for the themes. This step was not only about measuring interrater reliability but also about engaging in a reflexive dialogue. Differences in coding perspectives were discussed, and changes to the thematic scheme were made accordingly, leading to a deeper understanding of the data. The resulting themes were reported with quotes from the interviews for illustration so the reader can confirm our interpretations.

Results

Definitions

Of the 412 eHealth evaluation studies included, 104 (25.2%) studies defined tailoring or personalization by referencing definitions from previous research. Of the 71 (68.3%) studies that referred to tailoring definitions, there were 20 (28.2%) unique references. In addition, of the 33 (31.7%) studies that referred to personalization definitions, there were 6 (18.2%) unique references.

eHealth evaluation studies referring to tailoring definitions mainly referenced to de Vries and Brug [36] (n=15, 21%), Kreuter and Wray [37] (n=11, 16%), Kreuter et al [38] (n=8, 11%), and Hawkins et al [15] (n=8, 11%). The unique definitions of tailoring with the number of references to these definitions

XSL•FO
RenderX

in eHealth evaluation studies can be found in Multimedia Appendix 3 [10,15,36-52]. In Table 1 the themes that we identified in the published definitions and definitions according to eHealth experts are summarized to identify key elements.

A total of 10 (36%) of the 28 definitions include the term *adaptation*, referring to adapting health education materials [36]; the provision of support [39]; adjusting message frames [40] and information [41]; customizing the source, message, and channel of a given communication [37]; and mimicking to some extent a classic tailoring technique, that of "person-to-person" counseling [42-44].

In 16 definitions, the term *individual* appeared, ranging from adaptations for the individual [15,36-38,43] to collecting data from individual users [10,39,42,44-48]. *User groups* appeared in 5 (17.9%) definitions of tailoring, such as "needs, interests, personality, usage context, or other factors relevant to a user group" [16], and 1 definition included the way that the tailoring strategy is developed:

> *...tailored medicine, then you're still looking for those outcomes of randomized control trials and try to adjust.* [Participant 6]

Several definitions included the segmentation variables for tailoring, namely *preferences* (7/28, 25% definitions), *symptoms* (2/28, 7% definitions), *characteristics* (9/28, 32% definitions), *context* (4/28, 14% definitions), and *behavior* (2/28, 7% definitions). In total, 20 (71%) of the 28 definitions included customizations (content, identification, feedback, channel, or design). Content and channeling were mainly included in the definitions of tailoring. Definitions that included content-related customizations were related to changes in health education materials [36], communications [10,15,37], information [16,38,42], health messages [49], health information [50], advice [43], individual treatment plans, behavior change plans, or dietary advice [44], information on health behavior change [51], content [45,52], treatment [46], messages [47,48], persuasive text [48], and to "what kind of messages you give what kind of information" (participant 2). Definitions that included channeling were related to adapting source and channel [37]; amount, type, and through channels of delivery [50]; the way that the content is presented [45]; the presentation modality [41]; or highlighting some bits of information while omitting others [40]. One definition did not concern adapting the channel but mentioned that tailored content can be presented through different channels [48].

In total, 8 (29%) of the 28 definitions included terms related to computerization, such as "computerized" [36,42], "decision rules" [10,49] or "automatically" [10], "expertise programmed into a computer" [44], and "algorithm" [47]. Outcomes were included in 8 (29%) definitions of tailoring, and these ranged from increasing relevance [15,37,42], persuasiveness [16,49], enhancing motivation [50], attention [50], reducing barriers to exposure [50], widely distributing at low cost [43], and reaching the target group (participant 10).

The unique definitions of personalization can be found in Multimedia Appendix 4 [15,36,53-55]. Table 2 summarizes the themes that we identified in the definitions of personalization in the literature and according to eHealth experts to identify key elements.

XSL•FO
RenderX

Table 1. Themes identified in definitions of tailoring.

Study and participant	Adaptation	Individuals	User groups	Preferences	Symptoms	Characteristics	Context	Behavior	Content	Identification	Feedback	Channel	Design	Computerized	Outcomes
Study															
de Vries and Brug [36], 1999	✓	✓							✓					✓	
Kreuter and Wray [37], 2003	✓	✓							✓			✓			✓
Kreuter et al [38], 2000		✓				✓			✓						
Hawkins et al [15], 2008		✓							✓						✓
Krebs et al [10], 2010		✓							✓		✓			✓	
Lustria et al [49], 2009									✓					✓	✓
Brug et al [42], 2003	✓	✓				✓			✓					✓	✓
Oinas-Kukkonen and Harjumma [16], 2009			✓	✓		✓	✓		✓						✓
Rimer and Kreuter [50], 2006					✓				✓			✓	✓		✓
Kroeze et al [43], 2006	✓	✓				✓	✓	✓	✓						✓
Brug et al [44], 1999	✓	✓							✓		✓			✓	
Smeets et al [51], 2007		✓							✓						
Bental et al [45], 1999		✓		✓					✓			✓			

XSL•FO
RenderX

Study and participant	Adaptation	Individuals	User groups	Preferences	Symptoms	Characteristics	Context	Behavior	Content	Identification	Feedback	Channel	Design	Computerized	Outcomes
Nahum-Shani et al [39], 2016	✓	✓				✓	✓								
Păsărelu et al [46], 2017		✓		✓	✓	✓			✓						
Strecher [47], 1999		✓		✓		✓			✓					✓	
Smit et al [40], 2015	✓			✓								✓			
Nguyen et al [41], 2016	✓			✓								✓			
Dijkstra [48], 2008		✓							✓	✓		✓			
Heron and Smyth [52], 2010		✓					✓	✓	✓		✓	✓			
Participant															
Participant 1				✓		✓									
Participant 2			✓						✓						
Participant 3														✓	
Participant 5	✓	✓				✓				✓			✓		
Participant 6	✓		✓												
Participant 7			✓												
Participant 8			✓												
Participant 10															✓

XSL•FO
RenderX

Table 2. Themes in the definitions of personalization.

Study and participant	Adaptation	Individuals	User groups	Preferences	Symptoms	Characteristics	Context	Behavior	Content	Identification	Feedback	Channel	Design	Computerized	Outcomes
Study															
de Vries and Brug [36], 1999										✓					
Hawkins et al [15], 2008										✓					✓
Dijkstra [53], 2005		✓								✓					
Oinas-Kukkonen [16], 2009									✓						✓
Di-Clemente et al [54], 2001		✓				✓					✓				
Evers et al [55], 2014									✓						
Participant															
Participant 1	✓														✓
Participant 2	✓	✓													
Participant 3							✓	✓	✓			✓			
Participant 4							✓								
Participant 5	✓	✓	✓						✓				✓		
Participant 6	✓	✓													
Participant 7		✓	✓						✓						
Participant 8		✓								✓	✓				✓
Participant 9			✓												
Participant 10				✓											

Interestingly, the definitions from the literature did not include *adaptation*, whereas 4 participants included adaptations, such as “trying to come close to the efficacy of having only really a one-on-one chat” (participant 1), adapting a digital health intervention (participant 2), “As long as there is some adaptation going on that changes the product from generic to specific to an individual or group of individuals. It’s personalization to me.” (participant 5), “adjust or adapt your, maybe also your assessment and the treatment of patients in a way that it’s really fits that person” (participant 6).

XSL•FO
RenderX

The theme *individuals* appeared in 7 (44%) of the 16 definitions of personalization, such as "recognizable feature or set of features refers undeniably to the person, e.g. the person's first name or the combination of objective behavioral features such as the number of cigarettes smoked, the number of years the person smoked, and the brand smoked" [53] or in another definition "personalized feedback represents the most individualized type of feedback" [54], and "helping people at an individual level to make healthy choices" (participant 8).

User groups did not appear in the definitions in the literature, while 3 (30%) of the 10 participants included this in their definitions. In 2 (12%) of the 16 definitions, participants mentioned that personalization entails both adaptations to individuals and user groups, while another definition described personalization as an adaptation to different stakeholders:

> *It is how to shape and design, intelligent design, technology to support the different stakeholders in health care.* [Participant 9]

Adaptations that were included in the definitions mainly referred to content and identification. An example of content is "offering specific physical, psychological, or social treatment modules instead of a standard treatment package" [55], "provide specific health information for women" (participant 5). Examples of identification included in the definitions are "listing a person's name" [36], "conveying, explicitly or implicitly, that the communication is designed specifically for 'you'" [15], "incorporating recognizable aspects of a person in a general text" [53], "the feedback that is being offered should be experienced as personal so that people think or experience or perceive that they are being spoken to in a personal way. It accounts to them and not per se to their fellow peers or something" (participant 8).

In total, 4 (25%) of the 16 definitions of personalization focused on the outcomes of personalization. These include "increase attention or motivation to process messages" [15], "greater capability for persuasion" [16], and "trying to come close to the efficacy of having only really a one-on-one chat, but at the same time making it available for a lot of people. So, the kind of this tradeoff is there" (participant 1), and "So, it's a lot about the subjective experience, I would say. And it should lead to actionable insights that can be applied to their personal life" (participant 8).

Segmentation

During the interviews, participants described 9 types of variables that can be used for segmenting eHealth users, namely demographic variables, preferences, health information, psychological variables, behavioral variables, individual determinants, environmental variables, intervention interaction, and technology variables (Table 3).

Table 3. Variable types for segmenting eHealth users (N=227).

Variables	Themes, n (%)	Participants, n (%)
Demographic	46 (20.3)	8 (80)
Preferences	43 (18.9)	8 (80)
Health	32 (14.1)	8 (80)
Psychological	25 (11)	8 (80)
Behavioral	25 (11)	5 (50)
Individual determinants	23 (10.1)	7 (70)
Environmental	20 (8.8)	4 (40)
Intervention interaction	9 (4.0)	4 (40)
Technology	6 (2.6)	4 (40)

Demographic variables were the most-mentioned variables (46/227, 20.3%). Examples of demographic variables that were mentioned are gender, age, education level, ethnicity, nationality, occupation, and place of residency. During the interviews, participants mentioned that demographic information largely consists of variables that are unchangeable, and therefore, other types of variables are preferably used for segmenting eHealth users.

Next to demographic variables, preferences were regularly mentioned as a variable for segmenting eHealth users (43/227, 18.9%). Behavioral preferences were related to whether eHealth users prefer different behaviors related to the target behavior of the eHealth system (eg, whether users like exercise or food and sport preferences). Participants also mentioned preferences that were related to the content of the eHealth technology. An example is to use the needs of the user for segmenting eHealth users:

> *People may themselves perceive themselves as, you know, maybe in need of some information, but not so much therapeutic strategies or changing lifestyle while from a professional or clinical perspective, that may be exactly what you want to offer to this specific person. Again, this is something that relates to perceived need.* [Participant 10]

Channeling preferences that were mentioned by the participants were related to preferences on how technology delivers the behavior change intervention. Examples are preferences on when to receive messages, how the users of eHealth technologies want to communicate with their health care provider, and preferences for communication styles. Next among the preferences, graphical preferences were mentioned as a variable

XSL•FO
RenderX

that can be used for segmenting eHealth users (eg, preference for graphics within eHealth technology, color preferences, and layout and interface preferences).

Participants regularly mentioned health-related variables as a way to segment eHealth users (32/227, 14.1%). First, participants mentioned biomedical risk factors, such as BMI, cholesterol levels, blood pressure, and weight. Next to biomedical risk factors, participants mentioned medical conditions as a way to segment eHealth users (eg, type of diabetes). Medications taken by the eHealth user and the treatment process were also mentioned as variables for segmenting eHealth users. Psychological variables were mentioned by participants as a way to segment users of eHealth technologies (25/227, 11%). These variables were related to the eHealth user's personality, well-being, psychological distress (eg, depression and anxiety), and emotions experienced by the user.

Behavioral variables that were mentioned by the participants (25/227, 11%) were related to things that can be observed from the outside, such as physical activity, sleeping, and extended sitting. Behavioral variables ranged from more specific behaviors (eg, steps and extended sitting) to more overarching behaviors that include >1 behavior (eg, lifestyle). Moreover, an example of a more indirect variable was weight behavior, which was used as an indication of whether the user was still engaged in losing weight.

Individual determinants mentioned by the participants were related to the users' internal factors that determine behavior, such as attitudes, knowledge, and stage of change (23/227, 10.1%). Participants mainly mentioned determinants that were related to theories and models of behavior and behavior change. An example of a determinant was given by participant 1, who described that using determinants to segment eHealth users increases the potential to change someone's behavior:

> *And if you personalize them on that then it almost doesn't matter which gender you are, you know, it's much closer to what you think about this topic. And you might be the female exception on this topic, or I may be the male exception. But if the content is tailored to what I think, I think that has much more potential there for change.* [Participant 1]

Participants also mentioned environmental variables (20/227, 8.8%) that are related to the surroundings of the eHealth user. Examples include time, place, exposure, and day of the week. More distant examples are culture and climate. An example of the place was given by participant 3:

> *Context is everything from the obvious, the obvious would be if you are outside of a McDonald's restaurant, message on snack food might be beneficial. So that's the low hanging fruit.* [Participant 3]

The user's interaction with the eHealth technology was also mentioned as a way of segmenting eHealth users (9/227, 4%). This includes the more specific parts of how they interact with the system (eg, how often a particular feature is used) as well as the broader use characteristics (eg, adherence, adoption, and engagement). Technology-related factors were the least mentioned by participants (6/227, 2.6%) to segment users into smaller homogeneous groups. These variables include the extent to which the user is able or experienced in using different forms of technology. Examples of technology-related variables were related to how skilled or experienced users are with technology in a broad sense, that is, independent of the form of technology (eg, digital skills, experience with technology, computer literacy, and attitudes toward health technology), or related to a specific form of technology (eg, experience with VR).

Adaptation

The adaptation strategies mentioned by the participants were grouped along the elements of the eHealth technologies that can be adapted to the eHealth user (Table 4), namely channeling (46/100, 46%), content (24/100, 24%), graphical parts of the technology (16/100, 16%), the functionalities (10/100, 10%), and the behavior change strategy that is used by the eHealth technology (4/100, 4%).

Table 4. Elements that can be adapted (N=100).

Customization	Themes, n (%)	Participants, n (%)
Channeling	46 (46)	8 (80)
Content	24 (24)	8 (80)
Graphical	16 (16)	8 (80)
Functionalities	10 (10)	4 (40)
Behavior change strategy	4 (4)	2 (20)

Strategies for channeling the eHealth message were related to adapting the way the eHealth technology was delivered. Participants mentioned channeling as a strategy (46/100, 46%), with several ways of adaptation mentioned, namely including personal details in the messages, adapting the timing of the messages, adapting the way the messages were delivered, and adapting the technology used to deliver the eHealth intervention. To begin with, including personal details in messages was mentioned as a way to adapt an eHealth technology, ranging from including the first name, second name, or gender in the communication to the eHealth user. One participant associated this way of adapting the eHealth technology with the cocktail party effect:

> *There are some indications that personalized material increases someone's attention to it so, the cocktail party effect, when we hear our name, we process that information differently.* [Participant 5]

XSL•FO
RenderX

Second, the timing of the delivery of eHealth technology was mentioned as a way of tailoring the channel to the eHealth user. Examples range from adapting the frequency of sending messages to the user, whether to send a reminder, when to send a reminder, and sending messages after an alarm has been generated for a certain value. An added value of timing the eHealth technology was mentioned by participant 8:

> *And well, if you're able to shape that in a personal way, then I think it's helpful and something that the human coach cannot do, especially when it comes to the timing. Yeah, at the moment, you feel tempted to eat unhealthily or to sit on the couch and do nothing or something, while actually it's better to move at that moment, then the human coach is not there in general. But the technology can help you to, at that moment, make a healthier choice. [Participant 8]*

Third, adapting the way messages are delivered to the eHealth user is mentioned by participants as a way to adapt eHealth technology. Examples range from adapting the person delivering a message in a video, adapting the tone of voice to the eHealth user (eg, one that is more reflective and one that is more directive), the amount of information provided to the user, and the order in which different parts of the eHealth technology are provided:

> *...as professionals assume that a certain order is helpful or is logic or is this is the way we do it? This is how we've always done it. While a person using that application or that program may think otherwise, may think, well, I want to start with this. I want to start with relaxation. I don't want to start with cognitive restructuring or whatever. So, I find it a bit of a struggle between what we think we know this is the way to do it in terms of order and allowing persons to do whatever they like at any point in time, which would be the ultimate personalization like this.* [Participant 10]

Finally, the medium used to deliver the eHealth technology is also mentioned by participants as an adaptation strategy for adapting the technology. Examples include using either an app or web-based eHealth technology, sending messages either by email or push messages via phone, or informing the user by text or video. An example is given by participant 5:

> *But maybe by offering it in different channels and then allowing people to choose whether they want to read, because that seems to be the only thing that they can do here, or maybe they want to use the information in another way or via video or audio. [Participant 5]*

Content was also mentioned by participants (24/100, 24%) as a part that can be adapted to the eHealth user. Adaptation of content ranged from receiving different content (eg, offering different therapeutic approaches), providing feedback related to user input (eg, comparing provided data with the eHealth user's goal), comparing the eHealth user's data with data from peers, and giving advice to the eHealth user based on the data provided by the eHealth user.

Graphical aspects of the eHealth technology were mentioned by participants as a part that could be adapted to the user (16/100, 16%). The examples given by the participants ranged from very simple adaptations (eg, changing the colors) to changing the layout of the technology to more complex adaptations, such as changing an environment in VR or creating avatars adapted to the eHealth user:

> *And people can also develop their own avatar, which, of course, gives already a nice personalization aspect to it.* [Participant 9]

Functionality was also mentioned by participants as a part that can be adapted to the eHealth user (10/100, 10%). This is related to including or excluding functionalities (eg, in or excluding functionality that allows inviting others as social contact) of the eHealth technology based on the data provided by the eHealth user. Finally, 4% (4/100) of the participants mentioned examples of adapting the behavior change strategy to the user. This refers to adapting the way in which the eHealth technology aims to change behavior, for example, by providing different persuasion strategies according to the characteristics of the eHealth user. An example was given by participant 2 where, depending on the data provided by the user, gamification is only used as a way to change behavior if it is an appropriate strategy for that person:

> *And I think sometimes, and there's been lots of work on, for example, whether or not gamification has added value. And the general feeling is that, yes, it does have some added value. But it's pretty small still. And I think that, at least, partially, because it works well for some and not so for others.* [Participant 2]

Matching

Participants mentioned substantiation methods (24/56, 43%) and variable level (32/56, 57%) as 2 components for matching the segmentations with adaptations (Table 5).

Table 5. Parts that can be adapted (N=56).

Matching	Themes, n (%)	Participants, n (%)
Substantiation	24 (43)	6 (60)
Variable level	32 (57)	7 (70)

Participants identified several methods of substantiation for matching segments of eHealth users to adaptations of eHealth technology. These are pilot studies (12/56, 21%), using theory for matching (6/56, 11%), conducting interviews (4/56, 7%) and the use of AI to derive insights specifically for individual users (3/56, 5%). First, participants mentioned several forms of pilot studies as examples of how segmentation and customization can be matched. These are studies in which a

XSL•FO
RenderX

(prototype) version of the eHealth technology is available for use and is used as material during the study. Examples of pilot studies mentioned are matched study design and asking users to rate which messages they like within the eHealth technology:

> *So, I did a large surveys with, you know, 500 people, ask them to rate certain messages and how motivated they thought these messages were or not, could also be demotivating. And then also asked their personality and of, course, age and gender, and other demographic information. Based on that I found, you know that certain messages fit better with certain personality types, and certain genders.* [Participant 7]

Mixed methods pilot studies were also mentioned by participants. For example, one way to match segmentation and adaptation is to pilot an eHealth technology and interview users who are congruent with the intended use and people who are not congruent with the intended use. In this way, users' views on why a technology (does not) fit their characteristics can be translated into a personalization or adaptation strategy. Another example of a pilot study was given by participant 2:

> *So, what you can do is get people to try out different versions of an intervention and just measure their response and engagement and as engagement, is a predictor of effectiveness, we could also use, well, the version of the intervention that provokes the most engagement to an individual might also be the version that's most, that's best personalized to them. And that has the highest chances of being effective for this individual.* [Participant 2]

Using an existing theory or developing a theory for a personalization or tailoring strategy was also mentioned by participants (5/56, 9%) as a way to match the segmentation and adaptation strategy of an eHealth technology. An example of using existing theory is to include segmentation variables that have been found in previous research to be predictors of target behavior.

> *...you need to have a very strong mixed methods approach, you really need to, so maybe if I write a few things down here. But any intervention on health should be based on a logic model, a logic model, since you really need to have a clear theory of what you're doing, and you need really to understand the issue...* [Participant 3]

The use of interviews means that data are collected in an open format and then translated into a segmentation and adaptation strategy (4/56, 7%). Participants mentioned that one way to gain information about a strategy is to explore whether there are differences within the target group during the interviews or to ask explicitly about their preferences. Participants mentioned that because of the open nature of the interviews, it is possible to go to go in more depth while gaining information about the strategy than with other methods; therefore, the personalization or tailoring strategy can also be explored in more depth. In addition, another added value mentioned was that segmentation variables or tailoring strategies can be developed, which was an aspect the developer or designer of the eHealth technology had not previously thought of. In an example related to interviews, the participant described that part of the user group expresses different needs in the way they could navigate through the eHealth technology:

> *And from the interviews with them, we found out that it was more that they were looking for someone who took them by the hand and did everything step by step. So how we translated that in the intervention was that the only control they had in terms of navigating an intervention was clicking next and previous...* [Participant 5]

The use of AI to derive insights specifically for individual users was mentioned as a way to substantiate the matching of segmentation and adaptation (3/56, 5%). This way of creating an adaptation strategy means that at an individual level, it is decided what works for whom (adaptation strategy) using data science techniques. This is different from other substantiation strategies because other substantiation is more focused on what works for whom at a group level:

> *Well, for example, if I take an extreme example now, but just to make a case. So, the kind of research that I do is I can model individual behavior. So, we model your sleeping behavior for three weeks, and based on modeling your own behavior, we design interventions that fit your pattern...* [Participant 3]

The use of guidelines was least mentioned by participants (2/56, 4%) for matching segmentation and adaptation. One example mentioned by participant 6 was related to the customization of the treatment, which needs to be in line with health care criteria and guidelines. In addition, the same participant mentioned working with dietitians to find guidelines for adapted dietary recommendations for the user.

In addition to "substantiation," the second main code related to the theme "matching" was "variable level." Participants mentioned several variable levels to match the segmentation with the adaptation strategy, namely grouping variables (18/56, 32%), direct input (8/56, 14%), and per variable (6/56, 11%).

Participants mentioned several ways in which variables can be grouped to personalize eHealth and varied in how these methods can be applied (18/56, 32%). First, they mentioned that users can be grouped into smaller segments using multiple segmentation variables. Personas and profiles are 2 specific examples that were mentioned. Personas consist of groups of users with similar characteristics across a diverse range of variables (eg, similarities in demographics, eHealth literacy, and preferences), whereas profiles consist of segments that are described on multiple variables associated with a particular concept, such as lifestyle profiles or risk profiles.

Participants gave 2 examples of how these personas and profiles can be used. First, personas can be created at the beginning of the eHealth development process, and these different personas can then be translated to develop an adaptation strategy that matches the characteristics of the persona descriptions. Second, these profiles or personas can be used within the eHealth system to segment the eHealth users to whom an adaptation strategy is applied. Participant 5 described this use of personas and profiles:

XSL•FO
RenderX

> *We assume it's personalized because it's personalized to the persona. But the persona is a fictional representation of a group of individuals. So, it's not personalized to the individual. It's personalized to individuals like that person, as seen by the designer or researcher.* [Participant 5]

In addition to using representatives for groups of users, 1 participant mentioned that variables can also be combined to create profiles similar to "Facebook profiles," where each user is unique. This can be translated into a continuum where personas and profiles can be created from a variety of variables on the individual level to a group level. Participant 5 described how one can decide at what level these personas and profiles can be created to allow for tailoring or personalization:

> *To the degree that they show maximum diversity between the groups and maximum homogeneity within the groups. And I would do that and have done that in a data driven way.* [Participant 5]

Direct input from the user was also mentioned by the participants as a way of matching the segments with adaptation strategies (8/56, 14%). One participant mentioned that this way of matching segmentation and adaptation is not included in the model by Hawkins et al [15], which assumes that the developer of the eHealth technology collects information about the user and translates it into an adaptation strategy. The use of direct input means that the user adapts the eHealth technology themselves, and no information about the user is collected beforehand for segmentation. Participant 5 described how this can be applied:

> *So that cuts out the middleman, so you no longer have to measure something because you're both the person that is being measured and the one that is adapting. So, if someone is going to choose the color of their phone, they don't have to ask themselves what is your favorite color and then process or produce that fitting phone case, they can immediately make one themself or choose one themselves...* [Participant 5]

Finally, participants mentioned that segments and adaptation strategies can be matched per variable (6/56, 11%). This means that users are segmented on 1 variable and that adaptation strategies are developed on that single variable. Participants mentioned that this can either be done on an absolute measure of variables (eg, a segment on the variable name and matching this one variable with the inclusion of a name in messages) and that changes to this single variable can be used to match segmentation and adaptation:

> *So, I think you have to also adapt immediately to changes in the data. I think that's one way to personalize and get well; then you don't even need groups, I think, to form groups.* [Participant 6]

Moreover, participant 5 mentioned that matching per variable is mainly useful for adaptation strategies related to the content of the eHealth intervention:

> *It's a different extent in whether it's on the level of the individual. The thing is, I know that you are able to generate a tailored message using computer tailoring, for example, that you measure psychological constructs, and on the basis of that, you use information and look at the message database to construct a completely unique message for the individual. And I like this principle. I think it works for content, but I think it doesn't work for graphics, channel, and source.* [Participant 5]

Discussion

Principal Findings

In this multimethod qualitative study, we sought to gain insights into the definitions and distinguishing factors of personalization and tailoring, their two components (segmentation and adaptation), and how these two are matched.

We identified 16 unique definitions of personalization and 28 unique definitions of tailoring in the literature and through interviews with eHealth experts. Some definitions focused on the more general description of "adaptations," "individual users" or "user groups," while other definitions described personalization and tailoring in terms of how segmentation and customization can be operationalized. Similar outcomes, such as increased motivation, relevance and persuasiveness, were described in definitions of tailoring and personalization, reflecting an optimistic perspective. Interestingly, none of the definitions describe outcomes that are related to the adverse effects of applying these concepts.

We found that definitions of tailoring included the term "computerization" or related terms, such as "decision rules," whereas definitions of personalization did not. Although the definitions do not include a description of how these decision rules should be established, the inclusion of terms such as "prepared texts" [48] implies that decision rules for tailoring are developed in advance and that these decision rules are not established based on the data collected from the user. We also observed that definitions of personalization from the literature regularly included "identification," while definitions from eHealth experts did not include this in their definitions. Moreover, we observed that "identification" is less common in definitions of tailoring and that the definitions of tailoring that do include "identification" describe it as an umbrella term for personalization and other strategies (such as feedback).

Along with the definitions of tailoring and personalization, we sought to gain insights into how they can be applied to eHealth technologies using an existing framework [15] expanded with "matching" to describe how segmentations and adaptations are linked in a meaningful way. We found that eHealth users can be segmented based on their demographics (eg, age), preferences (eg, like or dislike an exercise), health (eg, symptoms), psychological variables (eg, distress), behavioral variables (eg, physical activity), determinants (eg, self-efficacy), environmental information (eg, whether the user is at work), intervention interaction (eg, number of visits), and technology (eg, eHealth literacy). These segmentation variables can be matched with adaptations, where the adaptation reflects 1 variable, a group of variables, or the direct user input.

XSL•FO
RenderX

We identified 5 elements of eHealth technologies to which adaptations relate, namely channeling (eg, including one's name in feedback messages), content (eg, providing normative feedback), graphical (eg, avatars that look similar to the user), functionalities (eg, including or excluding self-monitoring), and behavior change strategy (eg, including or excluding certain persuasion strategies). The match between the segments and customizations can be based on pilot studies, theory, interviews, AI for deriving individual-level insights, and guidelines. During the interviews, participants mentioned that these substantiations contribute differently to the matching of segmentation and customization. On the basis of the interviews, we have summarized the substantiation methods for matching segmentation and customization with a description of their objectives and examples of research questions in Table 6.

Overall, we observed that variability and technology affordances seem to determine whether and how personalization and tailoring should be applied to eHealth technologies, according to participants. In Figure 1, we illustrate how these variables may be related to each other according to the participants. Variability in segmentation variables (shown on the y-axis) may be interindividual (such as differences in country of birth), intraindividual (such as changes in emotions), or both intra and interindividual (eg, a combination of emotions and country of birth).

Table 6. Substantiation methods for matching segmentations and adaptations.

Method	Goal	Examples of research questions
Pilot studies	Test an existing (prototype version of) segmentation and customization strategy	Which version of the eHealth technology fits best with which user segment?
Theory	Form a link between segmentation and customization	The target group is very different in terms of technology skills, how can we adapt our technology?
Interviews	Form hypotheses about which customizations must be included in the eHealth technology	Does the target group express different needs or preferences during the interviews?
AI for deriving individual-level insights	Develop customizations on the individual level	What patterns of behavior do we observe at an individual level, and how can we adapt the eHealth technology to these individual patterns?
Guidelines	Gather information on segmentation and adaptation at the group level	What kind of treatment should the eHealth technology offer to users with high blood pressure?

Figure 1. Model summarizing how variability and affordances are related.

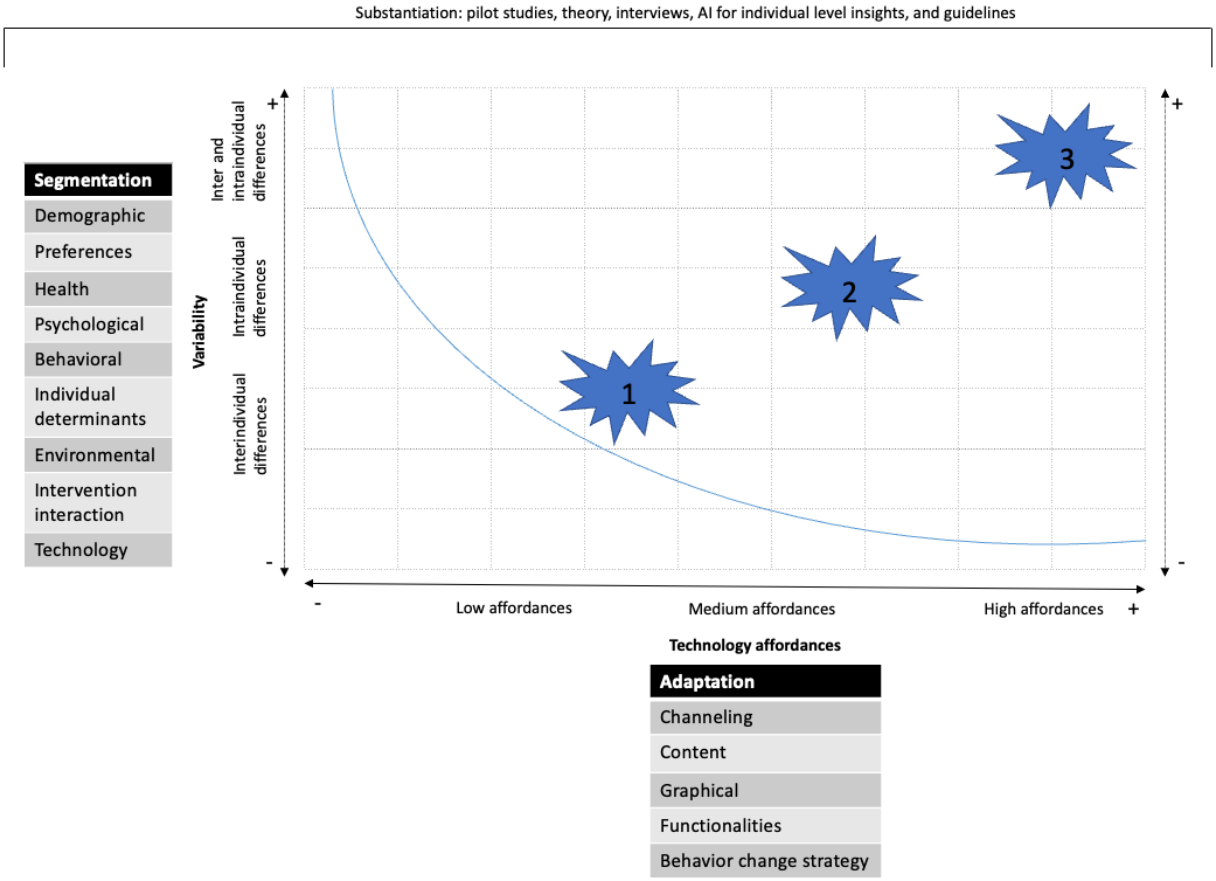

XSL•FO
RenderX

In addition, the affordances of the technology (x-axis) determine which adaptation strategy can be used. For example, a technology with high affordances such as VR allows adaptations in terms of channeling, textual content, graphical aspects, functionalities, and behavior change strategies, whereas an SMS text messaging–based eHealth technology can only be adapted in terms of textual content, channeling, and behavior change strategy. The area under the blue line depicts the area in which there is a mismatch between segmentation variability and technology affordances, resulting in ineffectively applied personalization and tailoring in eHealth technologies. The stars represent the desired combination of segmentation and adaptation levels, which, according to the participants, should ideally be at the same level. This means that low levels of segmentation should be combined with low levels of adaptations and the other way around. The illustration of each star is mentioned subsequently.

The first star represents the target group of the eHealth technology that shows interindividual differences (differences between users), and the technology with medium affordances allows for adaptations on the same level. An example is an eHealth technology that collects the name of the user (interindividual differences) at the first visit. On subsequent visits to the eHealth technology, a text is displayed saying "Welcome, John!" Another example is to segment users based on their visual appearance (interindividual differences) and to adapt characters in a virtual environment by making them look similar. As these examples illustrate, data can be collected once (or on a very long interval) because intraindividual differences are small, for example, through questionnaires and qualitative data collection.

The second star represents the target group of the eHealth technology that has intraindividual differences (differences within users), and the technology needs medium affordances to adapt to these differences over time. An example is an eHealth technology that segments users based on their location. If the eHealth user is in an environment where there is a place nearby for possible physical activity, the eHealth technology sends a message (channeling) to remind him or her to go for a run. As shown in this example, data must be collected multiple times to allow adaptations to these intraindividual differences, for example, through ecological momentary assessment, sensor data, log data, and activity trackers. The eHealth technology must have the affordances to translate these data into adaptations over time.

The third star represents the context in which eHealth users show both inter and intraindividual differences. An example is an eHealth technology that segments users by both their communication channel preference and the number of steps they take per day. Users indicate whether they want to use an eHealth technology on their computer or their mobile phone, through which it is then used. In addition, users receive motivational messages based on the number of steps, if this number is <2000. If the number of steps is >2000, the user will receive reinforcement messages. As shown in this example, data must be collected once for interindividual differences and several times related to intraindividual differences. The eHealth technology needs to have high affordances to allow adaptations that are stable and adaptations that occur over time.

To illustrate, one may use the theory about eHealth literacy to tailor or personalize the eHealth technology. However, if there are no intra or interindividual differences in eHealth literacy (low on the y-axis), there is no added value in segmenting and adapting based on this concept. Moreover, if an eHealth technology has very low affordances, such as an SMS text messaging–based eHealth technology, there is a low allowance for applying the different adaptation strategies (low on the x-axis); therefore, segmentation on multiple inter and intraindividual variables will often be obsolete.

Theoretical Implications

We argue that we can arrive at a clear definition of tailoring and personalization by merging the definitions from health communication research (mainly used in eHealth evaluation studies) with more recent definitions from other fields and eHealth experts. We observed a difference in the adaptations that are included in the definitions: while personalization shows a stronger focus on identification (including individual values) tailoring shows a focus on channeling and computerization (decision rules). We argue that the pattern underlying this difference cannot be captured in terms of segmentation, adaptation, and outcomes but in the way they are matched. While tailoring involves decision rules to match information collected about the individual user with general characteristics of groups, personalization does not include this in its definitions, as previous work in health communication research has focused on inserting identifiable aspects of the user. The affordances of the channels used in health communication research (brochures, leaflets, and later websites that largely mimic paper health information) were limited to adapting textual aspects and did not collect data from the user in real time, so definitions of personalization are limited to inserting someone's name or raising expectations. Nonetheless, advances in technology are broadening the area in which personalization can occur (more than just including one's name or raising expectations). The ability of technology to collect large amounts of data in real time and the potential of algorithms to extract meaningful insights from those data allow for individual-level matching at a deeper level than only "inserting" data from the user in communications. Therefore, based on published definitions and interviews with eHealth experts, we have formulated the definitions of eHealth personalization and eHealth tailoring as follows*:*

> *Segmentation of eHealth users based on (a combination of) demographics, preferences, health, psychological, behavioral, determinants, environment, intervention interaction, technology, matched with adaptations to content, channeling, graphics, functionalities, and/or behavior change strategies... by inserting this individual (segmentation) information or by translating this data into individual-level insights that are in turn translated into adaptations.* [eHealth personalization]

XSL•FO
RenderX

> *...by coupling segmentation variables with general characteristics of people with similar data using if-then algorithms.* [eHealth tailoring]

Both personalization and tailoring can be user-initiated, where this is referred to as customization in the case of personalization (technology has affordances for the user to directly change technology itself) and as self-tailoring in the case of tailored eHealth, where the user specifies their preferences or other information from which preferences are inferred using if-then algorithms.

Practical Implications

The literature shows mixed results on the effectiveness of tailoring and personalization. For example, the inclusion of a personalized meal planner [56] or tailoring for smokers with a low socioeconomic status [57] did not show better outcomes compared to a similar eHealth technology. In contrast, tailored advice on action planning [58] and daily tailored feedback on energy and fat intake [59] led to better outcomes. At this stage, it is difficult to draw conclusions about the effectiveness of personalization and tailoring because all the forms (all different segmentations, adaptations, and matching) are put together. It may be that one form works better than the other, which could explain the mixed results in effectiveness. We therefore suggest that the use of personalization and tailoring be reported in terms of segmentation, adaptation, and matching. By focusing reporting on segmentation and adaptation strategies (per "matched" segmentation and adaptation) with a description of how these were matched, we can compare and combine different forms, and subsequently, we may gain more insights into the working mechanisms of tailoring and personalization. For example, we can compare which theory gives the best results for building a segmentation and adaptation strategy, which segmentation variables are most relevant for personalizing and tailoring eHealth technologies, and which adaptation strategy increases the effectiveness of eHealth technologies. The items in Textbox 1 could be included as an extension of the CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH checklist [60] as a first step to gaining a better understanding of the working mechanisms of personalization and tailoring. This checklist is broader than the reporting standards described by Harrington and Noar [61], as it can also be used to report on channel adaptations, graphical properties, functionalities, and behavior change strategies.

Textbox 1. Suggested reporting standards for tailoring and personalization approaches in eHealth.

Segmentation and customization (description per "matched" link of segmentation and customization)

- Name the variables used to segment users into more homogeneous groups and which data collection method was used to collect data for segmenting eHealth users through the eHealth technology.
- Name which part of the eHealth technology was adapted (channeling, content, graphical, functionalities, and behavior change strategy) and how this part was adapted to represent the user segments.
- Describe how the match between segmentation and adaptation was substantiated (give details on the theory, pilot studies, artificial intelligence for deriving individual-level insights, guidelines, or interviews).
- Clarify the variable level through which segmentation and adaptation are linked (direct input, per variable, and grouping variables) and provide, where applicable, details about the algorithm or grouping that was used.

Limitations

An important limitation of this study is the lack of generalizability. Besides reviewing the literature for definitions, we have interviewed 10 eHealth experts through convenience sampling. As the context in which eHealth technologies use segmentation and adaptation is of great importance, it is not clear whether we have gained insights into these contextual factors and consequently, the different ways in which segmentation and adaptation are applied to eHealth technologies. However, we did include vignettes that are related to different contexts so that participants were able to provide input on eHealth technologies related to other contexts.

Conclusions

Overall, we found that personalization and tailoring are multidimensional concepts with multiple factors at play that determine how these concepts should be applied to eHealth and how effective they are. For example, continuously providing personalized feedback on segmentation variables that remain relatively stable over time may become more irritating than helpful, reducing the sense that the eHealth technology is truly personalized or tailored to the user. In short, some behaviors may require complex and idiosyncratic adaptations that personalization can provide. This requires a different study approach; while randomized controlled trials are appropriate for the development of if-then algorithms for tailoring, the development of personalization requires other study designs, such as an N-of-1 trial.

Acknowledgments

The authors would like to thank the eHealth experts for taking their time to participate in this study. Their valuable insights have significantly enriched this study, contributing to a better understanding of personalization and tailoring in the field of eHealth technologies. This work was supported by the Netherlands Organization for Scientific Research (NWO; grant 628.011.024), and

XSL•FO
RenderX

partners Podotherapy Center Wender, Reggeborgh, Stichting Diabetes en Vasculaire aandoeningen (DIAVASC), and Hospital Group Twente (ZGT).

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to the absence of explicit consent from participants but are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in the conceptualization and design of the study. ITK wrote the manuscript. ITK and HK carried out the analysis. SLB conducted the interviews. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview scheme.
[DOCX File , 14 KB - jmir_v26i1e50497_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 409 KB - jmir_v26i1e50497_app2.pdf]

Multimedia Appendix 3
Tailoring definitions.
[DOCX File , 31 KB - jmir_v26i1e50497_app3.docx]

Multimedia Appendix 4
Personalization definitions.
[DOCX File , 20 KB - jmir_v26i1e50497_app4.docx]

References

1. Durrani H. Healthcare and healthcare systems: inspiring progress and future prospects. Mhealth 2016;2:3 [FREE Full text] [doi: 10.3978/j.issn.2306-9740.2016.01.03] [Medline: 28293581]
2. van Gemert-Pijnen LJ, Kip H, Kelders SM, Sanderman R. Introducing eHealth. In: Kip H, Beerlage-de Jong N, van Gemert-Pijnen L, Sanderman R, Kelders SM, editors. eHealth Research, Theory and Development: A Multi-Disciplinary Approach. New York, NY: Routledge; 2018.
3. Car J, Tan WS, Huang Z, Sloot P, Franklin BD. eHealth in the future of medications management: personalisation, monitoring and adherence. BMC Med 2017 Apr 05;15(1):73 [FREE Full text] [doi: 10.1186/s12916-017-0838-0] [Medline: 28376771]
4. Deady M, Choi I, Calvo RA, Glozier N, Christensen H, Harvey SB. eHealth interventions for the prevention of depression and anxiety in the general population: a systematic review and meta-analysis. BMC Psychiatry 2017 Aug 29;17(1):310 [FREE Full text] [doi: 10.1186/s12888-017-1473-1] [Medline: 28851342]
5. Stratton E, Lampit A, Choi I, Calvo RA, Harvey SB, Glozier N. Effectiveness of eHealth interventions for reducing mental health conditions in employees: a systematic review and meta-analysis. PLoS One 2017;12(12):e0189904 [FREE Full text] [doi: 10.1371/journal.pone.0189904] [Medline: 29267334]
6. Capurro D, Ganzinger M, Perez-Lu J, Knaup P. Effectiveness of eHealth interventions and information needs in palliative care: a systematic literature review. J Med Internet Res 2014 Mar 07;16(3):e72 [FREE Full text] [doi: 10.2196/jmir.2812] [Medline: 24610324]
7. Champion KE, Parmenter B, McGowan C, Spring B, Wafford QE, Gardner LA, Health4Life team. Effectiveness of school-based eHealth interventions to prevent multiple lifestyle risk behaviours among adolescents: a systematic review and meta-analysis. Lancet Digit Health 2019 Sep;1(5):e206-e221 [FREE Full text] [doi: 10.1016/S2589-7500(19)30088-3] [Medline: 33323269]
8. Leslie E, Marshall AL, Owen N, Bauman A. Engagement and retention of participants in a physical activity website. Prev Med 2005 Jan;40(1):54-59. [doi: 10.1016/j.ypmed.2004.05.002] [Medline: 15530581]
9. Morrison LG, Yardley L, Powell J, Michie S. What design features are used in effective e-health interventions? A review using techniques from critical interpretive synthesis. Telemed J E Health 2012 Mar;18(2):137-144 [FREE Full text] [doi: 10.1089/tmj.2011.0062] [Medline: 22381060]

XSL•FO
RenderX

10. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010;51(3-4):214-221 [FREE Full text] [doi: 10.1016/j.ypmed.2010.06.004] [Medline: 20558196]
11. Lustria ML, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
12. Caserta MS, Lund DA, Utz RL, Tabler JL. "One size doesn't fit all" - partners in hospice care, an individualized approach to bereavement intervention. Omega (Westport) 2016 Jun;73(2):107-125 [FREE Full text] [doi: 10.1177/0030222815575895] [Medline: 27141124]
13. Schleidgen S, Klingler C, Bertram T, Rogowski WH, Marckmann G. What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC Med Ethics 2013 Dec 21;14(1):55 [FREE Full text] [doi: 10.1186/1472-6939-14-55] [Medline: 24359531]
14. Sebri V, Savioni L. An introduction to personalized eHealth. In: Pravettoni G, Triberti S, editors. P5 eHealth: An Agenda for the Health Technologies of the Future. Cham, Switzerland: Springer; 2020:53-70.
15. Hawkins RP, Kreuter M, Resnicow K, Fishbein M, Dijkstra A. Understanding tailoring in communicating about health. Health Educ Res 2008 Jun;23(3):454-466 [FREE Full text] [doi: 10.1093/her/cyn004] [Medline: 18349033]
16. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009;24(1):485-501. [doi: 10.17705/1CAIS.02428]
17. Petty R, Cacioppo J. The elaboration likelihood model of persuasion. In: Communication and Persuasion: Central and Peripheral Routes to Attitude Change. New York, NY: Springer; 1986.
18. Revere D, Dunbar PJ. Review of computer-generated outpatient health behavior interventions: clinical encounters "in absentia". J Am Med Inform Assoc 2001 Jan 01;8(1):62-79 [FREE Full text] [doi: 10.1136/jamia.2001.0080062] [Medline: 11141513]
19. Nguyen MH, Bol N, King AJ. Customization versus personalisation of digital health information. Eur J Health Commun 2020 Oct 01;1(1):30-54. [doi: 10.47368/ejhc.2020.003]
20. Cho E, Sundar SS. How do we like our online dates—customized or personalized? The differential effects of user vs. system tailoring on date preferences. Comput Human Behav 2022 Feb;127:107037. [doi: 10.1016/j.chb.2021.107037]
21. Kalyanaraman S, Sundar SS. The psychological appeal of personalized content in web portals: does customization affect attitudes and behavior? J Commun 2006 Mar;56(1):110-132. [doi: 10.1111/j.1460-2466.2006.00006.x]
22. Enwald HP, Huotari ML. Preventing the obesity epidemic by second generation tailored health communication: an interdisciplinary review. J Med Internet Res 2010;12(2):e24 [FREE Full text] [doi: 10.2196/jmir.1409] [Medline: 20584698]
23. Ryan K, Dockray S, Linehan C. A systematic review of tailored eHealth interventions for weight loss. Digit Health 2019;5:2055207619826685 [FREE Full text] [doi: 10.1177/2055207619826685] [Medline: 30783535]
24. Sundar SS, Marathe SS. Personalization versus customization: the importance of agency, privacy, and power usage. Hum Commun Res 2010;36(3):298-322 [FREE Full text] [doi: 10.1111/J.1468-2958.2010.01377.X] [Medline: 39436039]
25. Plow M, Mangal S, Geither K, Golding M. A scoping review of tailored self-management interventions among adults with mobility impairing neurological and musculoskeletal conditions. Front Public Health 2016;4:165 [FREE Full text] [doi: 10.3389/fpubh.2016.00165] [Medline: 27672633]
26. Moorthy KS. Market segmentation, self-selection, and product line design. Mark Sci 1984 Nov;3(4):288-307. [doi: 10.1287/mksc.3.4.288]
27. Serino CM, Furner CP, Smat C. Making it personal: how personalization affects trust over time. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences. 2005 Presented at: HICSS '05; January 6, 2005; Big Island, HI p. 170a URL: https://ieeexplore.ieee.org/document/1385576 [doi: 10.1109/hicss.2005.398]
28. Aalbers T, Baars MA, Qin L, de Lange A, Kessels RP, Olde Rikkert MG. Using an eHealth intervention to stimulate health behavior for the prevention of cognitive decline in Dutch adults: a study protocol for the brain aging monitor. JMIR Res Protoc 2015 Nov 10;4(4):e130 [FREE Full text] [doi: 10.2196/resprot.4468] [Medline: 26554496]
29. Kim B, Schwartz W, Catacora D, Vaughn-Cooke M. Virtual reality behavioral therapy. Proc Hum Factors Ergon Soc Annu Meet 2016 Sep 08;60(1):356-360. [doi: 10.1177/1541931213601081]
30. Ordúñez P, Tajer C. Disseminating cardiovascular disease risk assessment with a PAHO mobile app: a public eHealth intervention. Rev Panam Salud Publica 2015 Jul;38(1):82-85. [Medline: 26506325]
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Klooster IT, Kip H, van Gemert-Pijnen L, Crutzen R, Kelders S. A systematic review on eHealth technology personalization approaches. iScience 2024 Sep 20;27(9):110771 [FREE Full text] [doi: 10.1016/j.isci.2024.110771] [Medline: 39290843]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]

XSL•FO
RenderX

35. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
36. de Vries H, Brug J. Computer-tailored interventions motivating people to adopt health promoting behaviours: introduction to a new approach. Patient Educ Couns 1999 Feb;36(2):99-105. [doi: 10.1016/s0738-3991(98)00127-x] [Medline: 10223015]
37. Kreuter MW, Wray RJ. Tailored and targeted health communication: strategies for enhancing information relevance. Am J Health Behav 2003;27 Suppl 3:S227-S232. [doi: 10.5993/ajhb.27.1.s3.6] [Medline: 14672383]
38. Kreuter MW, Farrell DW, Brennan LK, Olevitch LR. Tailoring Health Messages: Customizing Communication With Computer Technology. New York, NY: Routledge; 2000.
39. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2016 Sep 23;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
40. Smit ES, Linn AJ, van Weert JC. Taking online computer-tailoring forward. Eur Health Psychol 2015;17(1):25-31 [FREE Full text]
41. Nguyen MH, van Weert JC, Bol N, Loos EF, Tytgat KM, van de Ven AW, et al. Tailoring the mode of information presentation: effects on younger and older adults' attention and recall of online information. Hum Commun Res 2016 Oct 21;43(1):102-126. [doi: 10.1111/hcre.12097]
42. Brug J, Oenema A, Campbell M. Past, present, and future of computer-tailored nutrition education. Am J Clin Nutr 2003 Apr;77(4 Suppl):1028S-1034S. [doi: 10.1093/ajcn/77.4.1028S] [Medline: 12663313]
43. Kroeze W, Werkman A, Brug J. A systematic review of randomized trials on the effectiveness of computer-tailored education on physical activity and dietary behaviors. Ann Behav Med 2006 Jun;31(3):205-223. [doi: 10.1207/s15324796abm3103_2] [Medline: 16700634]
44. Brug J, Steenhuis I, van Assema P, Glanz K, De Vries H. Computer-tailored nutrition education: differences between two interventions. Health Educ Res 1999 Apr;14(2):249-256. [doi: 10.1093/her/14.2.249] [Medline: 10387504]
45. Bental DS, Cawsey A, Jones R. Patient information systems that tailor to the individual. Patient Educ Couns 1999 Feb;36(2):171-180. [doi: 10.1016/s0738-3991(98)00133-5] [Medline: 10223021]
46. Păsărelu CR, Andersson G, Bergman Nordgren L, Dobrean A. Internet-delivered transdiagnostic and tailored cognitive behavioral therapy for anxiety and depression: a systematic review and meta-analysis of randomized controlled trials. Cogn Behav Ther 2017 Jan;46(1):1-28. [doi: 10.1080/16506073.2016.1231219] [Medline: 27712544]
47. Strecher VJ. Computer-tailored smoking cessation materials: a review and discussion. Patient Educ Couns 1999 Feb;36(2):107-117. [doi: 10.1016/s0738-3991(98)00128-1] [Medline: 10223016]
48. Dijkstra A. The psychology of tailoring‐ingredients in computer‐tailored persuasion. Soc Personal Psychol Compass 2008 Feb 21;2(2):765-784. [doi: 10.1111/j.1751-9004.2008.00081.x]
49. Lustria ML, Cortese J, Noar SM, Glueckauf RL. Computer-tailored health interventions delivered over the web: review and analysis of key components. Patient Educ Couns 2009 Feb;74(2):156-173. [doi: 10.1016/j.pec.2008.08.023] [Medline: 18947966]
50. Rimer BK, Kreuter MW. Advancing tailored health communication: a persuasion and message effects perspective. J Commun 2006 Aug;56(s1):S184-S201. [doi: 10.1111/j.1460-2466.2006.00289.x]
51. Smeets T, Kremers SP, Brug J, de Vries H. Effects of tailored feedback on multiple health behaviors. Ann Behav Med 2007 Apr;33(2):117-123. [doi: 10.1007/BF02879892] [Medline: 17447863]
52. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39 [FREE Full text] [doi: 10.1348/135910709X466063] [Medline: 19646331]
53. Dijkstra A. Working mechanisms of computer-tailored health education: evidence from smoking cessation. Health Educ Res 2005 Oct;20(5):527-539. [doi: 10.1093/her/cyh014] [Medline: 15701665]
54. DiClemente CC, Marinilli AS, Singh M, Bellino LE. The role of feedback in the process of health behavior change. Am J Health Behav 2001 May 01;25(3):217-227. [doi: 10.5993/ajhb.25.3.8] [Medline: 11322620]
55. Evers AW, Gieler U, Hasenbring MI, van Middendorp H. Incorporating biopsychosocial characteristics into personalized healthcare: a clinical approach. Psychother Psychosom 2014;83(3):148-157. [doi: 10.1159/000358309] [Medline: 24732828]
56. Brindal E, Freyne J, Saunders I, Berkovsky S, Smith G, Noakes M. Features predicting weight loss in overweight or obese participants in a web-based intervention: randomized trial. J Med Internet Res 2012 Dec 12;14(6):e173 [FREE Full text] [doi: 10.2196/jmir.2156] [Medline: 23234759]
57. Kock L, Brown J, Hiscock R, Tattan-Birch H, Smith C, Shahab L. Individual-level behavioural smoking cessation interventions tailored for disadvantaged socioeconomic position: a systematic review and meta-regression. Lancet Public Health 2019 Dec;4(12):e628-e644 [FREE Full text] [doi: 10.1016/S2468-2667(19)30220-8] [Medline: 31812239]
58. Bolman C, Eggers SM, van Osch L, Te Poel F, Candel M, de Vries H. Is action planning helpful for smoking cessation? Assessing the effects of action planning in a web-based computer-tailored intervention. Subst Use Misuse 2015;50(10):1249-1260. [doi: 10.3109/10826084.2014.977397] [Medline: 26440754]

XSL•FO
RenderX

59. Ambeba EJ, Ye L, Sereika SM, Styn MA, Acharya SD, Sevick MA, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. J Cardiovasc Nurs 2015;30(1):35-43 [FREE Full text] [doi: 10.1097/JCN.0000000000000120] [Medline: 24434827]
60. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
61. Harrington NG, Noar SM. Reporting standards for studies of tailored interventions. Health Educ Res 2012 Apr;27(2):331-342. [doi: 10.1093/her/cyr108] [Medline: 22156230]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
VR: virtual reality

Edited by T de Azevedo Cardoso; submitted 03.07.23; peer-reviewed by R Hawkins, O Korhonen; comments to author 02.11.23; revised version received 23.12.23; accepted 20.09.24; published 13.11.24.

Please cite as:
ten Klooster I, Kip H, Beyer SL, van Gemert-Pijnen LJEWC, Kelders SM
Clarifying the Concepts of Personalization and Tailoring of eHealth Technologies: Multimethod Qualitative Study
J Med Internet Res 2024;26:e50497
URL: https://www.jmir.org/2024/1/e50497
doi:10.2196/50497
PMID:

©Iris ten Klooster, Hanneke Kip, Sina L Beyer, Lisette J E W C van Gemert-Pijnen, Saskia M Kelders. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Health Platform for Improving the Effect of the Active Health Management of Chronic Diseases in the Community: Mixed Methods Exploratory Study

Zhiheng Zhou[1], MD, PhD; Danian Jin[1], PhD; Jinghua He[1], MD, PhD; Shengqing Zhou[2], BS; Jiang Wu[1], MD; Shuangxi Wang[1], MD; Yang Zhang[1], PhD; Tianyuan Feng[1], MD

[1]Pingshan Hospital of Southern Medical University, Shenzhen, China

[2]University of Washington, Seattle, WA, United States

Corresponding Author:
Zhiheng Zhou, MD, PhD
Pingshan Hospital of Southern Medical University
No. 16 Renmin Road, Pingshan District
Shenzhen, 518118
China
Phone: 86 13828493963
Email: zhihengz@163.com

Abstract

Background: China is vigorously promoting the health management of chronic diseases and exploring digital active health management. However, as most medical information systems in China have been built separately, there is poor sharing of medical information. It is difficult to achieve interconnectivity among community residents' self-testing information, community health care information, and hospital health information, and digital chronic disease management has not been widely applied in China.

Objective: This study aimed to build a digital health platform and improve the effectiveness of full-cycle management for community chronic diseases through digital active health management.

Methods: This was a single-arm pre-post intervention study involving the development and use of a digital health platform (2-year intervention; 2020 to 2022). The digital health platform included the "i Active Health" applet for residents and the active health information system (cardio-cerebrovascular disease risk management system) for medical teams. The digital active health management of chronic diseases involved creating health streets, providing internet-assisted full-cycle active health services for residents, implementing internet-based community management for hypertension and diabetes, and performing real-time quantitative assessment and hierarchical management of residents' risks of cardio-cerebrovascular disease. After the 2-year intervention, management effectiveness was evaluated.

Results: We constructed a digital health platform with interconnected health information and implemented a digital active health management model. After the intervention, the 2-way referral between community health care institutions and hospitals increased. Residents' health literacy rate increased from 30.6% (3062/10,000) in 2020 to 49.9% (4992/10,000) in 2022, with improvements in health knowledge, health behavior, and health skills. Moreover, the risk of cardio-cerebrovascular disease decreased after the intervention. The community hypertension and diabetes standardized management rates increased from 59.6% (2124/3566) and 55.8% (670/1200) in 2020 to 75.0% (3212/4285) and 69.4% (1686/2430) in 2022, respectively. The control rates of blood pressure in patients with hypertension and blood sugar in patients with diabetes increased from 51.7% (1081/2091) and 42.0% (373/888) in 2020 to 81.2% (1698/2091) and 73.0% (648/888) in 2022, respectively. The intervention improved patients' BMI, waist circumference, blood uric acid levels, and low-density lipoprotein cholesterol levels. The drug compliance rate of patients with hypertension and diabetes increased from 33.6% (703/2091) and 36.0% (320/888) in 2020 to 73.3% (1532/2091) and 75.8% (673/888) in 2022, respectively. The intervention greatly improved the diet behavior, exercise behavior, and drinking behavior of patients with hypertension and diabetes.

Conclusions: Our digital health platform can effectively achieve the interconnection and exchange of different health information. The digital active health management carried out with the assistance of this platform improved the effectiveness of community chronic disease management. Thus, the platform is worth promoting and applying in practice.

(J Med Internet Res 2024;26:e50959) doi:10.2196/50959

XSL•FO
RenderX

KEYWORDS

information platform; active health; chronic disease management; effectiveness; community; digital health; health literacy; cardio-cerebrovascular disease; China

Introduction

Since the promulgation of "Healthy China Action (2019-2030)" by the Chinese government, China has actively promoted a comprehensive and full life cycle health strategy, taking "popularizing knowledge and improving literacy, self-discipline, healthy living, early intervention, services, mass participation, joint construction, and sharing" as the implementation principle of Healthy China [1,2]. The concept of "active health" was first proposed by Chinese scholars in 2015 [3,4]. Active health is a novel medical model that involves actively applying controllable stimuli to the human body, increasing its micro complexity, promoting diverse adaptation, and achieving enhanced human function or reversal of chronic diseases. It emphasizes the long-term continuous dynamic tracking of the entire life cycle behavior system of individuals, and identifying and evaluating their own state, evolutionary direction, and degree, with a focus on selecting various elements of lifestyle, fully exerting their subjective initiative, and improving health behavior. By comprehensively using various medical methods, controllable and active interventions are carried out on human behavior, promoting self-organized adaptive changes in the human body and thereby achieving functional improvement in practical activities and knowledge systems to eliminate diseases and maintain a healthy state of the human body [5,6]. Internet technology and artificial intelligence technology are indispensable tools in achieving active detection and intervention among residents. Although international scholars did not directly put forward the word "active health," they carried out a lot of related research on active health, including digital health, internet-based medicine, artificial intelligence–assisted health care, and new digital management models, to promote patients' health knowledge, improve residents' health behaviors, and improve self-health management [7-9]. It could be predicted that internet-based active health management would be an important way to promote active health.

Internet-based health management is a new medical service model, which is based on the internet and is supported by cloud computing, big data, the Internet of Things, and other information technologies through cross-penetration, integration, and innovation with traditional medical services. This model has openness, interactivity, convenience, and cross-border characteristics, which could maximize the allocation and use of medical resources, diversify medical methods and pathways, facilitate medical processes and mechanisms, and personalize medical needs and services. The practice has proved that internet-based health management has changed the past model of the hospital serving patients, making it transform into a collaborative medical model including doctor-patient collaboration, general practitioner-specialist collaboration, and preventive medicine-clinical medicine integration. It could guide high-quality medical resources to primary care, realize medical resource sharing, and improve the accessibility of high-quality hospital and community health resources. The construction of an internet-based health service model will profoundly change the traditional way of community chronic disease health management, which would be a new trend and direction of health care development [10,11]. The internet and wearable detection devices automatically upload data, allowing residents to receive medical services at their homes. The application of these information technologies will play important guarantee and support roles in promoting the graded diagnosis and treatment of chronic diseases and improving the level of health management of chronic diseases [12,13].

With the development of China's social economy and the changes in medical conditions, the disease spectrum and death spectrum of residents have significantly changed. There is a continuous upward trend in both the prevalence and mortality of chronic diseases such as diabetes and hypertension. Chronic diseases have become a public health problem seriously threatening the health and life of Chinese residents. The main goal of China's health strategy is to carry out effective health management for chronic diseases and reduce the incidence of complications and disease burden [14,15]. At present, the overall level of community hypertension and diabetes health management in China is still unsatisfactory, internet-based health care has been applied to chronic disease management in China, and the initial results have been obtained. However, most medical information systems in China have been built separately, and there is limited information sharing available. Even in the same region, it is difficult to achieve interconnection between patients' information in community health service institutions and hospital medical information systems. This is a bottleneck that hinders the application and promotion of internet-assisted chronic disease management models in China [16,17]. Therefore, this study aimed to build a digital health platform and improve the effectiveness of chronic disease full-cycle management in the community through the implementation of a digital active health management model.

Methods

Overview

This was a single-arm pre-post intervention study that involved digital health platform building and use assessment conducted through a 2-year digital active health management intervention for residents from 2 streets. The intervention effectiveness was evaluated by qualitative and quantitative methods after 2 years. This mixed methods approach included quantitative evaluations of health literacy; changes in blood glucose, blood pressure, and blood lipids; and health behaviors of patients with chronic diseases, as well as qualitative evaluations of feedback from residents and health care staff.

XSL•FO
RenderX

Digital Health Platform Building

Objectives and Requirements of the Digital Health Platform

The objective of the digital health platform was to realize the interconnection of residents' wearable devices, applets for residents, community health service center information systems, hospital information systems, and government information systems, and promote residents to actively obtain health knowledge, monitor health and behaviors, and participate in health interventions through the platform. The platform would allow the medical team to comprehensively collect residents' health information, conduct intelligent health risk assessment and hierarchical management, and create personalized health interventions to finally achieve the goal of accurate health management for patients with hypertension and diabetes in the community. This platform included the "i Active Health" applet, risk of cardio-cerebrovascular disease management system, community health service information system, and hospital information system, which connected with the government information system and network maps. The architecture of the digital health platform is shown in Figure 1.

Figure 1. The architecture of the digital health platform. The C4 model and "classification of digital interventions, services, and applications in health" from the World Health Organization were referenced to draw the digital health platform architecture. (A) The system context of the architecture of the digital health platform. (B) The containers and components of the architecture of the digital health platform, which included users, systems (containers), contents (components), and their functions. The blue text boxes represent the systems within the platform, while the gray text boxes represent the external systems. Arrows represent data applications, and the solid line represents the content included. ASCVD: atherosclerotic cardiovascular disease.

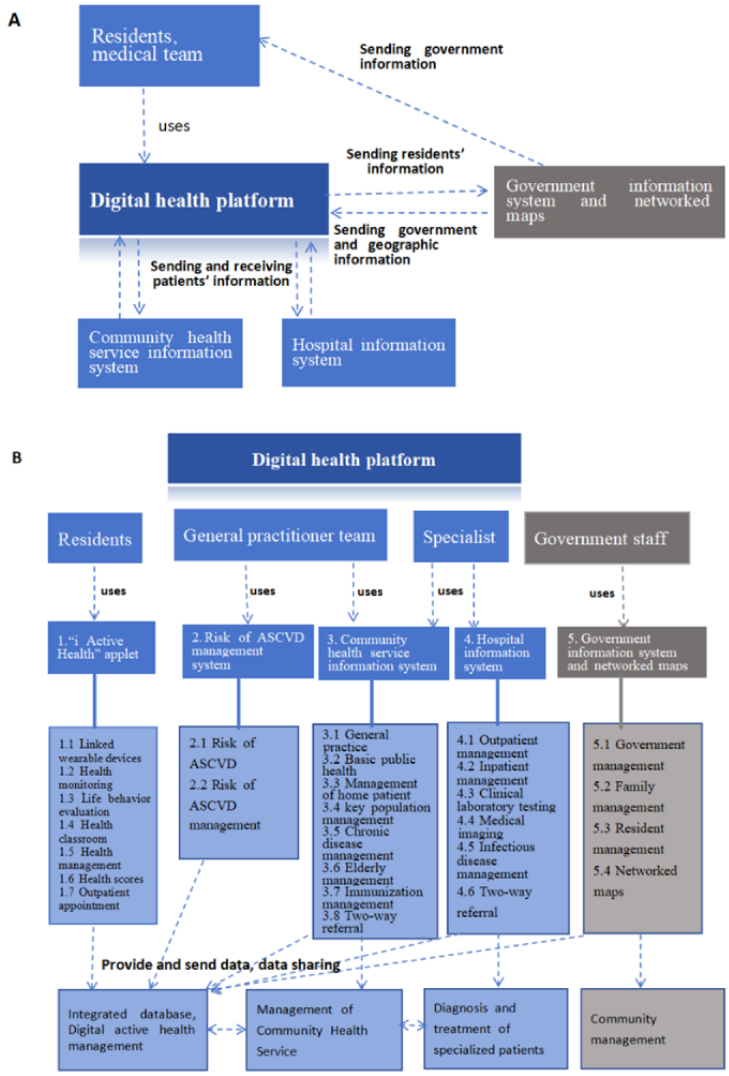

Establishment of the "i Active Health" Applet

We established China's first "i Active Health" applet for residents' active learning, active health monitoring, and active participation in interventions. The applet linked various wearable devices, such as a sphygmomanometer, glucose meter, health bracelet, sleep monitor, electrocardiogram device, and motion detection equipment. The results could be automatically transferred to the "i Active Health" applet after home testing by the residents, which could promote proactive health testing among residents. There were functions of life behavior evaluation and health prescription formulation. Residents' exercise and diet information could be evaluated, and targeted dietary prescriptions and exercise prescriptions could be formulated. The risk of cardio-cerebrovascular disease was automatically assessed by allowing residents to enter health

XSL•FO
RenderX

information or by capturing residents' health information from community health information systems and hospital information systems, which could promote residents to actively improve their living behaviors and participate in health management. There was a health classroom with health-related courses (graphics, text, video, and voice) available for residents, which could promote residents to actively obtain health knowledge and skills. There was a health management function. The general practitioner team developed health management plans for patients, provided online guidance, and scheduled follow-up appointments through the "i Active Health" applet, which could encourage patients to actively participate in health management (Figure 2). The "i Active Health" applet registration guidelines are provided in Multimedia Appendix 1. The training materials for using the "i Active Health" applet are presented in Multimedia Appendix 2. The health scores are presented in Multimedia Appendix 3.

Figure 2. "i Active Health" applet. (A) The homepage of the "i Active Health" applet, which displays health indicators, health scores, and the rank of local residents (top left of the homepage). (B) Health examination results. (C) Self-measured blood pressure results and health management recommendations. (D) Self-test blood glucose results and health management recommendations. (E) Record and analysis of 3 meal recipes. (F) Exercise record and analysis. (G) Smoking record. (H) Record of alcohol consumption.

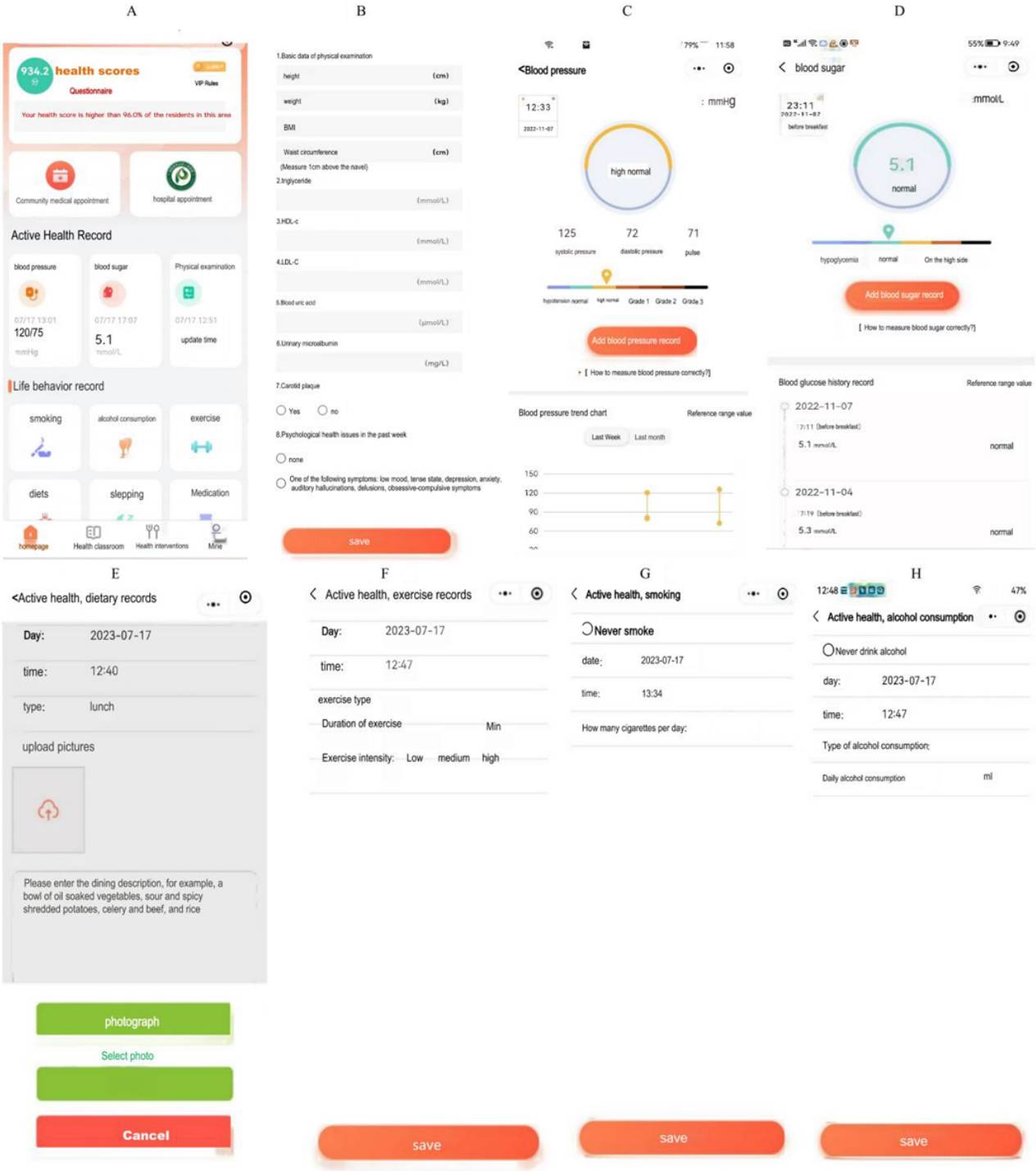

XSL•FO
RenderX

Establishment of a Risk Management System for Atherosclerotic Cardiovascular Disease

We screened 24 risk assessment indicators (including basic medical history risk, health behavior risk, and health effect risk) through a series of methods and established a risk assessment model for cardio-cerebrovascular disease (RACCVD). A risk management information system was established, and 4 risk assessment tools, including Stroke Risk Scorecard (SRSC) (China) [18], 10-year atherosclerotic cardiovascular disease (ASCVD) (China) [19], Systematic Coronary Risk Evaluation (SCORE) (Europe) [20], and Framingham scale (FRS) (United States) [21], were configured in the system. The system interconnected the "i Active Health" applet, community health information system, hospital information system, and network map. The risk assessment score of cardio-cerebrovascular disease could be calculated, and the system could automatically group hazard levels based on resident's risk scores, which could realize early identification of high-risk groups and accurate positioning, early warning, and early diagnosis and treatment (Figure 3).

Figure 3. Risk management system of cardio-cerebrovascular disease. (A) The homepage of the active health information platform. (B) The regional risk assessment page shows the risk status of cardio-cerebrovascular disease and the main data of health management of residents in all streets of Pingshan district. (C) The community risk assessment page shows the risk status of cardio-cerebrovascular disease and the main data of health management of residents in some communities. (D) The individual health risk page shows the health indicators and risk scores of cardio-cerebrovascular disease in individual residents.

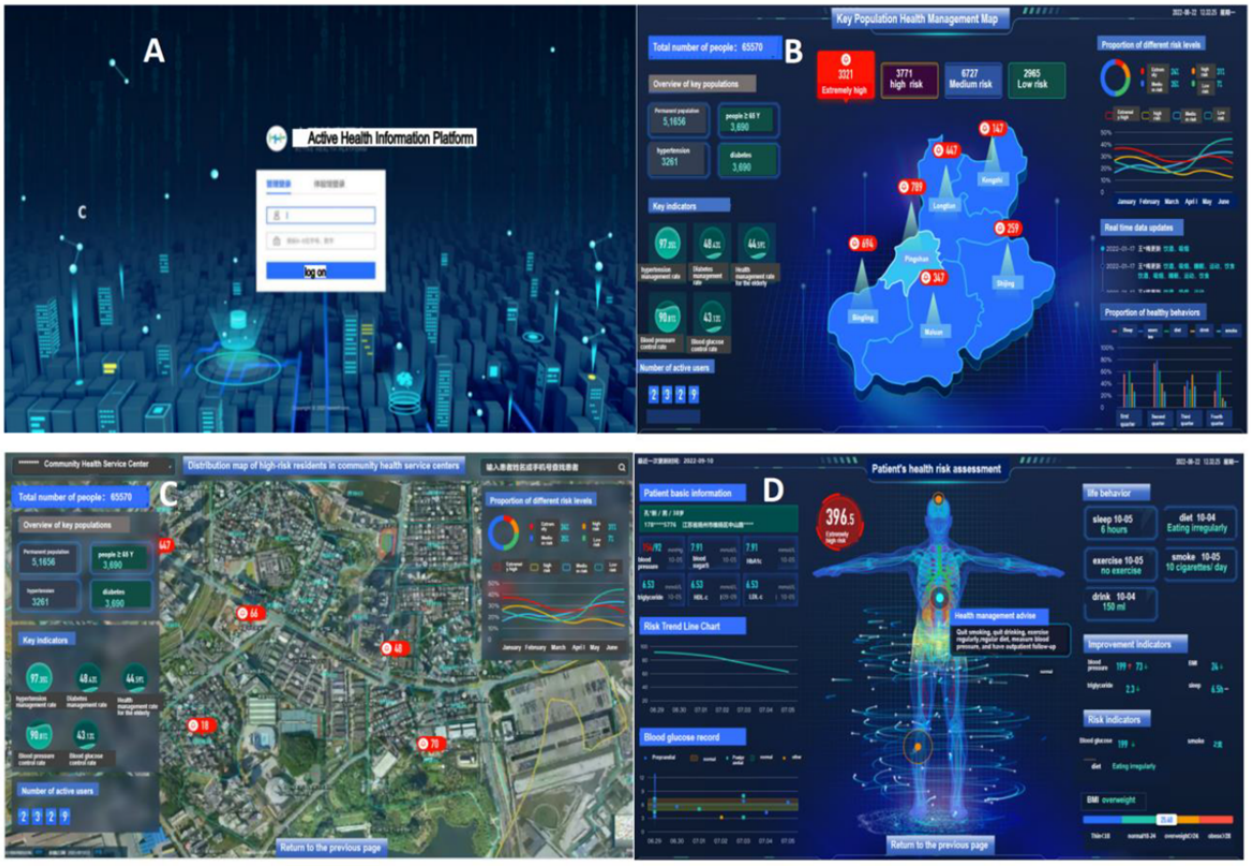

Modification of the Community Health Information System

The key population management module was added to the community health information system, which was used to interface with the data generated in the "i Active Health" applet and residents' self-test health data and health records online in a centralized manner, and the "risk score" and "risk level" were shown, based on which family doctors could quickly identify and pay attention to extremely high-risk and high-risk populations, and recommend transfer to higher-level hospitals for diagnosis and treatment (Figure 4).

XSL•FO
RenderX

Figure 4. The key population management module in the community health service information system. (A) The homepage of the community health service information system. (B) Residents' health risk scores and risk levels shown in the key population management module. (C) The risk indicators of cardio-cerebrovascular disease in the key population management module.

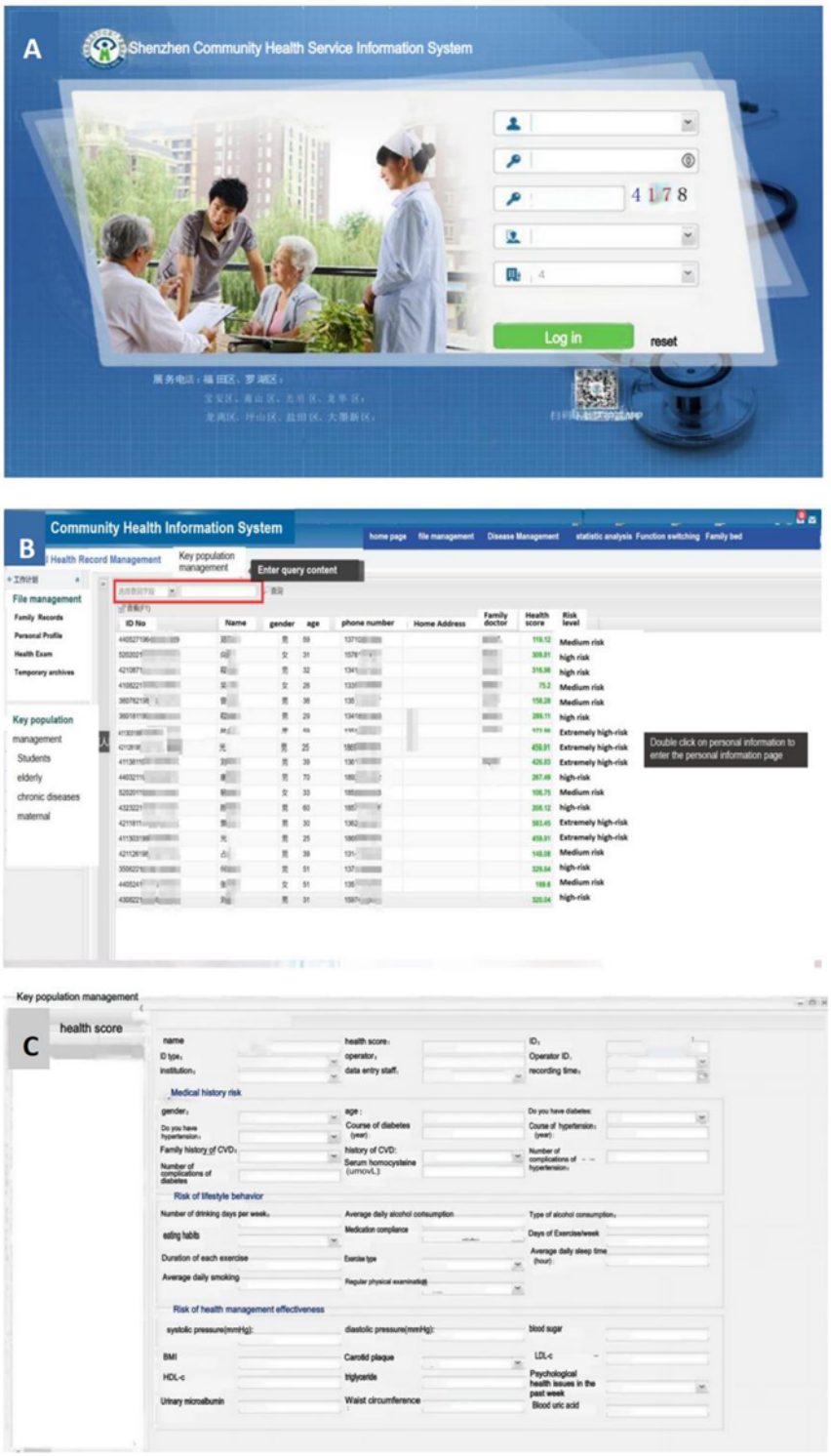

Modification of the Hospital Information System

A 2-way referral module was added to the hospital information system to improve the up and down transfer of patients between the hospital and the community health service center. A remote diagnosis and treatment information system module was added, which promoted hospital experts to participate in remote diagnosis and treatment in community consultations. Therefore, the health data collected by medical institutions and third-party systems (wearable devices) were interconnected, allowing unified collection and comprehensive management.

Connection of Government Information Platforms

Patients' demographic data, personal health behavior, and family status data were connected with the resident management information systems of the government department where they

XSL•FO
RenderX

were located to ensure the consistency of resident information in different departments and provide a basis for government departments to carry out health work and ensure the accuracy of patient medical information. At the same time, residents' information was integrated with the information of the data bureau for precise positioning of patient risk assessment and hierarchical management.

Platform Application: Digital Active Health Management

Selection of Experimental Streets

Biling street and Shijing street in Pingshan District, Shenzhen, China were selected as digital active health management intervention experimental streets.

Inclusion Criteria

The inclusion criteria were as follows: (1) diagnosis of hypertension or type 2 diabetes, age ≥65 years, and registration in the community health service center; (2) normal hearing and self-awareness; (3) presence of a smartphone in the family and ability to receive and read online information; (4) ability to complete self-health monitoring and online health classroom learning; and (5) willingness to accept health management and outpatient follow-up provided by general practitioner teams.

Exclusion Criteria

The exclusion criteria were as follows: (1) poor hearing and consciousness, and inability to communicate as usual; (2) inability to complete regular self-health tests; and (3) unwillingness to accept health managers or follow-up.

Digital Active Health Management Intervention Methods

Deep Cooperation Between the Government and Health Institutions

The active health service plans for the street have been formulated, which were under the government's annual key work for promotion. The street office released active health-related information through government information platforms and collaborated with street enterprises, schools, and other parties to build an active health system for residents.

Active Health Activities Provided by the Government

The services included establishing teams of health science popularization instructors, regularly providing health training to residents, building health volunteer service teams, cultivating health coaches and health clubs, and inviting community residents to participate in health activities. A total of 50 popular science videos were made. Moreover, a street health culture festival was held; healthy communities, health parks, and health canteens were introduced; and a good health culture atmosphere was created.

Active Health Services by Medical Institutions

Digital active health services included conducting health classrooms (video, voice, or text) to promote residents' active learning on health knowledge and skills, and screening hypertension, diabetes, dyslipidemia, and common tumors among residents. Moreover, the services included health monitoring, health warnings, diet monitoring, and sports health management for residents. Health activities were held on special days, such as world hypertension day, diabetes day, heart disease day, and chronic obstructive pulmonary disease day, and online and offline health consultations and interventions were provided for residents.

Optimization of the 2-Way Referral Process

The process of patient referral between community health institutions and hospitals was optimized by a 2-way referral information platform. With regard to transfer from the community health center to the hospital, the hospital prioritized the opening of specialized account sources and focused on priority reception, priority examination, and priority hospitalization for transferred patients. With regard to transfer from the hospital to the community health center, patients who completed diagnosis and treatment in the hospital and met the criteria for referral were referred to the community general practitioner team by specialist doctors through the information system, and joint management of general practitioners and specialists was carried out.

ASCVD Risk Assessment and Hierarchical Management

ASCVD risk assessment was conducted for adult residents aged 18 years or older who completed annual health examinations by the "i Active Health" applet and digital health platform. Hierarchical management was carried out according to the risk level of ASCVD. Residents with medium and low risk were managed by community general practitioner teams, while residents with high and extremely high risk were automatically sent to the cardiovascular and neurology departments of the hospital and were managed by specialists.

Internet-Based Management of Hypertension and Diabetes

Chronic Disease Management Team

Each practitioner team included a general practitioner (team leader), nurse, and health manager. At the same time, cardiovascular doctors, endocrinologists, clinical pharmacists, dietitians, and sports rehabilitation specialists were allocated according to the needs of residents to provide technical support for the chronic disease management team.

Intervention Process

The general practitioner team provided refined management through a combination of online and offline approaches. More than one face-to-face visit was simultaneously conducted per quarter, including condition consultation, physical examination, and clinical diagnosis and treatment. The management process is shown in Figure 5.

The patient self-measured blood pressure or blood sugar through an intelligent sphygmomanometer and glucose meter detection terminal, and the data were automatically transmitted. The active health applet and digital health platform obtained detection data and automatically sent it to the patient's family members and general practitioner team. The digital health platform grouped and managed patients. The active health applet provided health

XSL•FO
RenderX

education courses for patients and their families. The health managers and nurses followed up with patients and their families through WeChat and phone calls, provided health education, and promptly provided feedback on the patient's condition to the general practitioner. The patient's family members assisted in urging the patient to take medicine, follow the diet, and perform exercise at home. On-site health education lectures and exchange meetings were regularly held. General practitioners controlled the patient's condition in real time, provided health management for patients, and referred patients to specialized doctors. Medical experts were invited to conduct joint diagnosis and treatment at community health service centers. The operation manual is presented in Multimedia Appendix 4.

Figure 5. Internet-based management of hypertension and diabetes. (A) Internet-based management team including a general practitioner, a nurse, a health manager, a dietitian, sports rehabilitation specialists, clinical pharmacists, and specialists (cardiovascular doctors and endocrinologists). (B) The 5 steps of internet-based management. It includes patient training and enrollment, patient data collection and evaluation, online monitoring and health management, offline health intervention, and face-to-face outpatient diagnosis and treatment by physicians.

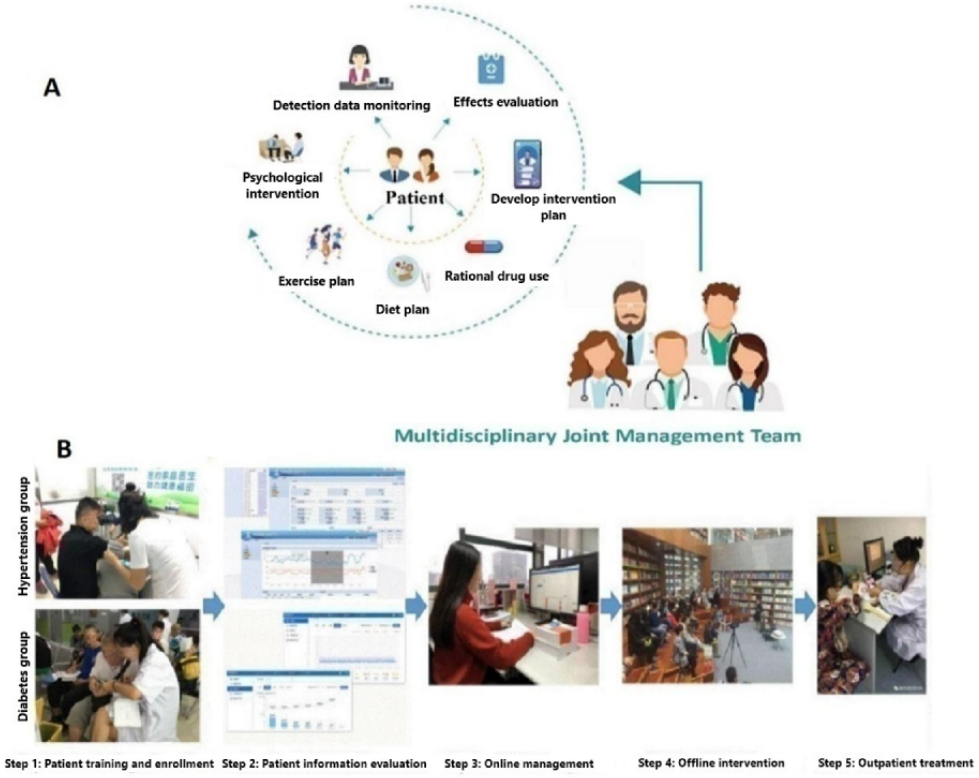

Effectiveness Evaluation

Qualitative Methods

The methods included conducting interviews and on-site supervisions to understand the construction of the digital health platform and the implementation of active health activities among residents and patients in the experimental area.

Quantitative Methods

For quantitative assessments, data were collected on residents' health, questionnaire survey findings, physical examination findings, laboratory testing results, etc, and evaluations were performed of the changes in residents' health literacy, the ASCVD risk, and the effectiveness of management for hypertension, diabetes, and old age.

Assessment Tools

Health Literacy Questionnaire

The Shenzhen Citizen Health Literacy Questionnaire was applied, which had a total of 32 questions with 5 parts: general condition (5 questions), physical health condition (5 questions), health knowledge (8 questions), daily living habits (8 questions), and health skills (6 questions). The Cronbach α of this questionnaire was .873, Kaiser-Meyer-Olkin (KMO) value was 0.893, and Bartlett test χ^2_{435} was 4267.20 (P<.001).

Risk Assessment Model of Cardio-Cerebrovascular Disease

The risk assessment model was developed through a series of methods and validated in the cohort population, which included 24 evaluation indicators for the dimensions of basic medical history risk (age, sex, history of hypertension, history of diabetes, history of cardiovascular and cerebrovascular events, family history of cardiovascular and cerebrovascular diseases, and hyperhomocysteinemia), health behavior risk (smoking, drinking, exercise, diet, sleep, medication, and physical examination), and health effect risk (current blood pressure, current blood sugar, BMI or waist circumference, triglyceride, high-density lipoprotein cholesterol [HDL-C], low-density lipoprotein cholesterol [LDL-C], hyperuricemia, microalbuminuria, carotid plaque, and psychological disorders). The total score was 0-1000, and a higher score indicated a higher

XSL•FO
RenderX

risk. According to the risk score, the risk level was divided into low risk, medium risk, high risk, and extremely high risk. The Cronbach α was .873, and the reliability coefficient was 0.891. The indicator system and its calculation method of the risk assessment of cardio-cerebrovascular disease are presented in Multimedia Appendix 5.

SRSC (China)

The SRSC was developed by the Committee of Stroke Screening, Prevention and Treatment Project of the National Health Commission of the People's Republic of China [18]. Overall, "8+2" stroke risk factors were used to assess the risk stratification of patients, among which "8" refers to hypertension, atrial fibrillation or valvular heart disease, diabetes, smoking history, dyslipidemia, lack of exercise, significant overweight or obesity, and a family history of stroke, and "2" refers to a history of stroke or transient ischemic attack. Individuals with 3 or more of the "8" risk factors for stroke or those with a history of transient ischemic attack or stroke were included in the high-risk group. Those who met 3 or fewer factors but had one of the chronic diseases (including hypertension, diabetes, atrial fibrillation, or valvular heart disease) were included in the medium-risk group. Those who met 3 or fewer factors and had no chronic diseases were included in the low-risk group. The scale was reasonably reliable, with a Cronbach α of .77 and a reliability coefficient of 0.803.

10-year ASCVD Risk Assessment Model (China)

The 10-year ASCVD risk assessment model was proposed by the Guidelines for the Prevention and Treatment of Dyslipidemia in Chinese Adults (2016 Revision) [19]. This evaluation model was based on the average incidence risk of diabetes and hypertension, the LDL-C level, the total cholesterol (TC) level, and the number of combined risk factors (men ≥45 years or women ≥55 years, smoking, and low HDL-C). An average incidence risk of <5% was considered low risk, 5%-9% was considered medium risk, and ≥10% was considered high risk. The Cronbach α was .813, and the reliability coefficient was 0.805.

SCORE (Europe)

The SCORE is a 10-year risk of fatal cardiovascular disease model in Europe by the SCORE project [20]. The model included 5 risk factors: age, gender, systolic blood pressure, TC, and smoking. An absolute risk of cardiovascular disease onset within 10 years of <1% was considered low risk, ≥1% but <5% was considered medium risk, ≥5% but <10% was considered high risk, and ≥10% was considered extremely high risk. The Cronbach α was .830, and the reliability coefficient was 0.871.

FRS (United States)

The FRS was developed and published by researchers from Framingham Heart Research in 1961 [21]. The model included 7 risk factors: gender, age, blood pressure, TC, HDL-C, smoking, and diabetes. An absolute risk of cardiovascular disease onset within 10 years of <5% was considered low risk, ≥5% but <10% was considered medium risk, and ≥10% was considered high risk. The Cronbach α was .701, the split-half reliability coefficient was 0.826, and the test-retest reliability coefficient was 0.94.

Medication Compliance Questionnaire

The Medication Compliance Questionnaire (MCQ) was developed by a team led by Professor Thomas Paraidathathu [22,23]. The MCQ included a total of 7 questions. Questions in the MCQ assessed patients' intentional and unintentional noncompliance to medication instructions, including reasons for noncompliance. A 4-point Likert scale was used for each question: "4," none of the time; "3," sometimes (1-4 times per month); "2," most of the time; and "1," all the time. A total score of 27 was considered to indicate compliance, which was the evaluation criterion of drug compliance for each patient in this study. The scores may range from 7 to 28. The Cronbach α was .782, and the interrater reliability of 2 interviewers demonstrated a Cohen kappa statistic of 0.796.

Assessment Methods of Life Behavior

Exercise Behavior

Regular exercise means exercising more than 3 times a week, with each continuous exercise session lasting more than 30 minutes.

Diet Behavior

Based on the patient's self-reported information, the dietary status in the past 2 weeks of the survey was determined to be the following: balanced diet, mainly meat and vegetables, mainly vegetarian, salt loving, oil loving, or sugar loving.

Smoking

Smoking was defined as smoking one or more cigarettes per day and smoking for 6 months or more.

Drinking

According to the "Dietary Guidelines for Chinese Residents (2022)," the definition of alcohol consumption was an average daily alcohol consumption of ≥15 g in the past 2 weeks.

Diagnosis Standard of Hypertension and Diabetes

Hypertension

Essential hypertension was in accordance with the 2023 "Guidelines for the Prevention and Treatment of Hypertension in China." Hypertension was defined as a systolic blood pressure ≥140 mmHg or a diastolic blood pressure ≥90 mmHg.

Diabetes

Diabetes was in accordance with the 2023 American Diabetes Association Diabetes Diagnosis and Treatment Standards. With regard to glucose testing, diabetes was considered as a fasting blood sugar level of ≥7.0 mmol/L, or an oral glucose tolerance test result or blood glucose level in 2 hours after a meal of ≥11.1 mmol/L, with or without clinical symptoms such as polydipsia, overeating, polyuria, and weight loss.

Ethical Considerations

All the human experiments were performed in accordance with the principles of the Declaration of Helsinki. All experimental protocols were approved by the Research Ethics Committee of

XSL•FO
RenderX

Pingshan Hospital of Southern Medical University (approval number: KY-2019-001-01). The personal information of samples involved in the study was not disclosed.

Statistical Methods

SPSS 20.0 software (IBM Corp) was used to establish a database and conduct statistical analysis. Measurement data have been expressed as mean (SD), and comparisons of these data before and after the intervention were performed using a paired *t* test. Count data have been expressed as rate or constituent ratio, and comparisons of these data before and after the intervention were performed using a paired chi-square test. Differences with $P \leq .05$ were considered statistically significant.

Results

Demographic Characteristics

The total population in the experimental streets before the intervention was 118,126, including 70,603 men and 46,980 women (men-to-women ratio of 1:0.665), and there were 5713 (4.8%) elderly people aged older than 60 years. Among the residents, 27,275 (23.1%) had a primary school or lower education level, 43,801 (37.1%) had a junior high school education level, 23,814 (20.2%) had a high school (technical secondary school) degree, and 23,235 (19.7%) had a university or higher education level. There were 3566 patients with hypertension and 1200 patients with diabetes registered in the community health service center.

Residents' Medical Treatment Situation

The number of chronic patients transferred from community health service centers to hospitals has increased from 79 in 2020 to 1634 in 2021 and 2381 in 2022 through digital health management. The number of patients transferred from hospitals to community health service centers has increased from 185 in 2020 to 1760 in 2021 and 3569 in 2022. The number of elderly people aged older than 65 years in the experimental streets who actively go to community health service centers for physical examination increased from 4876 in 2020 to 5579 in 2021 and 6809 in 2022.

Changes in Residents' Health Literacy

In order to monitor residents' health literacy, a health literacy questionnaire survey was conducted annually among 1000 adult residents aged 18 years or older in our region. The survey results showed that the health literacy of residents in the experimental streets rose slowly from 2017 to 2019, which was lower than the average level of residents in the city. The health literacy of residents in 2019 was 30.6% (3062/10,000). After the 2-year intervention, residents' health literacy scores significantly improved. In 2022 and 2023, residents' health literacy reached 48.0% (4802/10,000) and 49.9% (4992/10,000), respectively, with significant differences ($P<.001$). Residents' health knowledge, health behavior, and health technology literacy showed significant improvements (Figure 6).

Figure 6. Changes in the health literacy of residents before and after the intervention. (A) Health literacy of residents in Shenzhen city and the experimental streets from 2016 to 2022. (B) Health knowledge, behavior, and skills of residents in the experimental streets from 2016 to 2022.

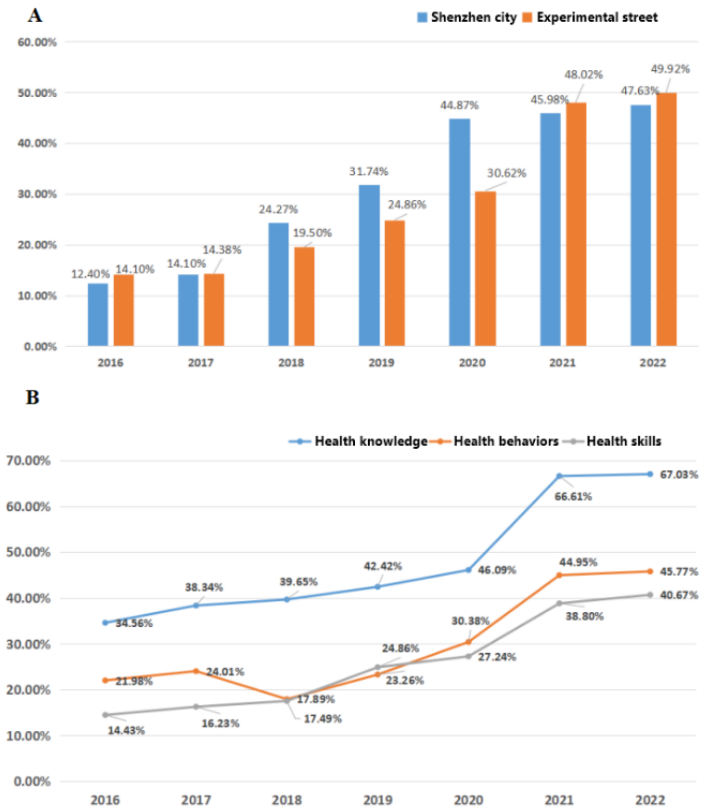

XSL•FO
RenderX

Changes in Residents' Risks of ASCVD

The residents' risks of ASCVD in the experimental streets were evaluated by the RACCVD, SRSC, ASCVD, FRS, and SCORE scales. We selected 6996 adult residents (those with hypertension or diabetes, or those aged older than 65 years) who completed the annual physical examination in 2020 and 2022 for ASCVD risk assessment. The results showed that after the 2-year intervention, the risk of ASCVD among the residents reduced, and the high-risk rate of residents in 2022 assessed by the 5 assessment models significantly reduced (*P*<.001) (Table 1).

Table 1. Risk of atherosclerotic cardiovascular disease in residents with hypertension or diabetes, or those aged older than 65 years before and after the intervention.

Assessment tool	Year 2020 (before; n=6996), n (%)			Year 2022 (after; n=6996), n (%)			*P* value
	Low risk	Medium risk	High risk or extremely high risk	Low risk	Medium risk	High risk and extremely high risk	
RACCVD[a]	2648 (37.9)	3067 (43.8)	1281 (18.3)	2977 (42.5)	3121 (44.6)	898 (12.8)	<.001
SRSC[b]	2534 (36.2)	1454 (20.8)	3011 (43.0)	2892 (41.4)	1226 (17.5)	2873 (41.1)	<.001
ASCVD[c]	3178 (45.4)	1652 (23.6)	2166 (31.0)	3572 (51.1)	1022 (14.6)	2402 (34.3)	<.001
FRS[d]	4260 (60.9)	2113 (30.2)	623 (8.9)	4431 (63.3)	2128 (30.4)	437 (6.2)	<.001
SCORE[e]	2708 (38.7)	3267 (46.7)	1021 (14.6)	3115 (44.5)	3088 (44.1)	793 (11.3)	<.001

[a]RACCVD: risk assessment model for cardio-cerebrovascular disease.

[b]SRSC: Stroke Risk Scorecard.

[c]ASCVD: 10-year atherosclerotic cardiovascular disease.

[d]FRS: Framingham scale.

[e]SCORE: Systematic Coronary Risk Evaluation.

Management Effect of Hypertension and Type 2 Diabetes Before and After the Intervention

Coverage of Patients With Hypertension and Type 2 Diabetes

The number of managed patients with hypertension was 3566 in 2020 and increased to 4285 in 2022. The standardized management rate of community hypertension increased from 59.6% (2124/3566) in 2020 to 75.0% (3212/4285) in 2022. The number of managed patients with diabetes increased from 1200 in 2019 to 2430 in 2022. The standardized management rate increased from 55.8% (670/1200) in 2020 to 69.4% (1686/2430) in 2022.

Effect of Blood Pressure and Blood Glucose Control

A total of 2091 patients with hypertension and 888 patients with diabetes who were enrolled in 2020 and who completed the 2-year internet-based management were statistically analyzed. The average blood pressure of patients with hypertension decreased from 148/96 mmHg before the intervention to 137/82 mmHg after the intervention, and the blood pressure control rate of patients with hypertension increased from 51.7% (1081/2091) to 81.2% (1698/2091). The average fasting blood glucose and glycated hemoglobin levels of patients with diabetes were 7.9 mmol/L and 8.2%, respectively, before the intervention and decreased to 6.4 mmol/L and 6.3%, respectively, after the intervention, and the blood sugar control compliance rate increased from 42.0% (373/888) to 73.0% (648/888).

Changes in Physical Examination Indicators

We evaluated the changes in patients' physical examination indicators before and after the intervention. The BMI, waist circumference, blood uric acid, TC, triglyceride, and serum LDL-C values of patients with hypertension decreased significantly after the intervention compared with the values before the intervention (*P*<.05). The BMI, waist circumference, aspartate aminotransferase, TC, triglyceride, and LDL-C values of patients with diabetes decreased significantly after the intervention compared with the values before the intervention (*P*<.05). The data are presented in Tables 2 and 3.

XSL•FO
RenderX

Table 2. Changes in physical examination indicators in patients with hypertension (n=2091).

Variable	Before the intervention, mean (SD)	After the intervention, mean (SD)	*P* value
BMI (kg/m^2)	25.67 (5.58)	23.52 (3.86)	<.001
Waist circumference (cm)	87.55 (12.28)	84.11 (13.12)	<.001
Aspartate transaminase (U/L)	25.50 (23.76)	25.07 (3.35)	.41
Alanine transaminase (U/L)	23.73 (11.27)	23.45 (10.22)	.40
Blood uric acid (µmol/L)	395.9 (99.52)	345.10 (64.23)	<.001
Total cholesterol (mmol/L)	5.15 (1.17)	4.61 (1.01)	<.001
Triglyceride (mmol/L)	2.23 (1.70)	2.12 (1.62)	.03
LDL-C[a] (mmol/L)	2.80 (1.03)	2.23 (1.01)	<.001
HDL-C[b] (mmol/L)	1.49 (0.54)	1.50 (0.42)	.50

[a]LDL-C: low-density lipoprotein cholesterol.
[b]HDL-C: high-density lipoprotein cholesterol.

Table 3. Changes in physical examination indicators in patients with type 2 diabetes (n=888).

Variable	Before the intervention, mean (SD)	After the intervention, mean (SD)	*P* value
BMI (kg/m^2)	24.67 (3.61)	23.14 (4.52)	<.001
Waist circumference (cm)	86.27 (13.24)	81.74 (11.85)	<.001
Aspartate transaminase (U/L)	26.22 (16.41)	37.81 (12.66)	<.001
Alanine transaminase (U/L)	21.35 (10.22)	21.01 (12.54)	.53
Blood uric acid (µmol/L)	363.10 (85.83)	341.20 (71.96)	<.001
Total cholesterol (mmol/L)	5.27 (1.18)	4.71 (1.25)	<.001
Triglyceride (mmol/L)	2.33 (2.41)	2.11 (1.42)	.02
LDL-C[a] (mmol/L)	2.86 (1.07)	2.41 (1.14)	<.001
HDL-C[b] (mmol/L)	1.29 (1.53)	1.33 (2.56)	.69

[a]LDL-C: low-density lipoprotein cholesterol.
[b]HDL-C: high-density lipoprotein cholesterol.

Changes in Medical Behavior

The self-measured blood pressure and blood glucose behaviors among the patients were investigated at the time of enrollment and after the 2-year internet-based management. At the time of enrollment, 25.4% (531/2091) and 26.0% (231/888) of patients with hypertension and type 2 diabetes, respectively, self-measured their blood pressure and glucose regularly (more than once a week). After the 2 years of management, 71.2% (1489/2091) and 69.4% (616/888) of patients with hypertension and type 2 diabetes, respectively, self-measured their blood pressure and glucose regularly, with a statistically significant difference (*P*<.001). The MCQ was used to evaluate patient compliance. After the 2-year intervention, the drug compliance of patients with hypertension increased from 33.6% (703/2091) to 73.30% (1532/2091), while that of patients with diabetes increased from 36.0% (320/888) to 75.8% (673/888) (*P*<.001). The data are presented in Table 4.

XSL•FO
RenderX

Table 4. Changes in medical behavior and life behavior among patients with hypertension and type 2 diabetes.

Variable	Before the intervention, n (%)	After the intervention, n (%)	*P* value
Patients with hypertension (n=2091)			
Self-measured blood pressure regularly	531 (25.4)	1489 (71.2)	<.001
Drug compliance	703 (33.6)	1532 (73.30)	<.001
Regular exercise	740 (35.4)	1296 (62.0)	<.001
Balanced dietary behavior	1641 (78.5)	1790 (85.6)	<.001
Smoking	309 (14.8)	217 (10.4)	<.001
Drinking	226 (10.8)	222 (10.6)	.84
Patients with type 2 diabetes (n=888)			
Self-measured glucose regularly	231 (26.0)	616 (69.4)	<.001
Drug compliance	320 (36.0)	673 (75.8)	<.001
Regular exercise	359 (40.4)	713 (80.3)	<.001
Balanced dietary behavior	687 (77.4)	786 (88.5)	<.001
Smoking	125 (14.1)	117 (13.2)	.58
Drinking	263 (29.6)	137 (15.4)	<.001

Changes in Life Behavior

Balanced dietary behavior, regular exercise, and smoking control among patients with hypertension significantly improved after the 2-year intervention than before the intervention (balanced diet rate: 1790/2091, 85.6% vs 1641/2091, 78.5%; regular exercise rate: 1296/2091, 62.0% vs 740/2091, 35.4%; smoking rate: 217/2091, 10.4% vs 309/2091, 14.8%; *P*<.001). Similarly, balanced diet behavior, regular exercise, and drinking behavior among patients with diabetes significantly improved after the 2-year intervention than before the intervention (balanced diet rate: 786/888, 88.5% vs 687/888, 77.4%; regular exercise rate: 713/888, 80.3% vs 359/888, 40.4%; drinking rate: 137/888, 15.4% vs 263/888, 29.6%; *P*<.001). On the other hand, drinking behavior among patients with hypertension and smoking behavior among patients with diabetes did not significantly improve. The data are presented in Table 4.

Discussion

Principal Findings

Active health is a new medical model initiated by Chinese scholars, which emphasizes the active health idea of “active discovery, scientific evaluation, active adjustment, and health promotion” [24,25]. From the perspective of the concept of active health, for individuals, active health is a medical model that promotes people’s subjective initiative for health through effective methods, focuses on actively improving health literacy, focuses on improving health behavior, and actively monitors and intervenes in the human body through various medical means to promote health and eliminate diseases. For the population, active health is a new health model that adheres to government leadership, fully mobilizes the enthusiasm of the entire society, emphasizes that individuals are the “first responsible person” for health, and promotes a healthy lifestyle with the support of new technologies such as informatics and biomics, achieving effective monitoring, accurate prediction, and active intervention, and promoting the health of the whole population. This model advocates the cooperation of governments, enterprises, medical institutions, individuals, and other parties, moving from passive medical care to active health care for everyone [9,26,27]. Active health has become one of the hot topics for scholars in China.

In recent years, many scholars in China have explored factors related to active health, such as the integration of medicine and sports, community chronic disease management, wearable active health devices, home care, smart communities, and biomechanics. However, most involved theoretical research, while empirical research was scarce [28]. There are currently no reports involving digital active health for an experimental street. Research has shown that digital medicine is an inevitable trend for the management of community chronic diseases. In developed countries, chronic disease management based on internet technology has been widely applied, and good results have been achieved. Our research results also showed that internet-based management had a positive effect on improving the compliance of chronic disease patients and improving their health behaviors.

A recent study [29] showed that the prevalence of cardiovascular risk factors has increased significantly in China from 2002 to 2018, with obesity increasing from 7.1% in 2002 to 16.4% in 2018, overweight/obesity increasing from 29.9% to 50.7%, hypertension increasing from 18.8% to 27.5%, diabetes increasing from 2.6% to 11.9%, and dyslipidemia increasing from 18.6% to 35.6%. The mortality rate of chronic noncommunicable diseases has increased from 533/100,000 in 2012 to 685/100,000 in 2018, an increase of 28.5% [29]. More than 50% of Chinese men currently smoke, and an increasing number of adults lack physical activity. According to research [29,30], the prevalence of chronic noncommunicable diseases in China is rising, while management is stagnant. It is estimated

XSL•FO
RenderX

that the prevalence and management of obesity, hypertension, and diabetes in 2030 will not be appropriate, and global or national goals will not be achieved. Therefore, there is an urgency to implement effective intervention models to improve the living behavior of Chinese residents and enhance the effectiveness of chronic disease control. Previous research [12,26] found that the majority of chronic disease patients in China are currently identified by social and family health records, hospital physical examinations, and treatment processes. However, the discovery and management of patient health issues have been relatively delayed, and patient participation in their own health management is not high. At present, China has established various methods to collect patient health information. However, community health service data are difficult to summarize and integrate with another platform, forming many information islands [30,31]. Therefore, carrying out internet-based active health management based on internet technology assistance, which integrates public health and clinical services, is an important approach to achieve intelligent health management of chronic diseases throughout the cycle.

Based on the current situation and existing problems of chronic diseases in China, our team built a digital health platform, and to our knowledge, we are the first to build an active health information platform and carry out digital active health management of the community in China. Our active health information platform could fully use information technology to enrich chronic disease prevention and treatment methods and work content, and promote network service applications such as appointment diagnosis and treatment, online follow-up, disease management, and health management. It effectively realized the interconnection of residents' wearable devices, residents' applets, community health service center information systems, hospital information systems, and government information systems. To our knowledge, it is rare in China to achieve the interconnection and sharing of this information. Through the 2-year application of the digital health platform, we effectively improved the efficiency of the 2-way referral of patients in the experimental streets between community health service centers and hospitals, and effectively improved the effectiveness of community chronic disease management.

We used the applet and network resources developed, and through deep cooperation between the government and health institutions, we took multiple measures to create an active and healthy environment for the residents on the experimental streets. We organized a large number of comprehensive active health activities for the residents. Medical institutions took the initiative to provide a series of active health services for community residents. After the 2-year continuous intervention, the levels of health knowledge, health behavior, and health technology literacy among residents significantly improved, and the total health literacy of residents increased from 30.6% before the intervention to 49.9% after the intervention. The practice has proven that the improvement of residents' health knowledge and health behavior is a long process that requires the assistance of multiple departments and the implementation of multichannel and sustained health education and promotion. An effective auxiliary tool is an important element in improving the effectiveness of health education. The results of this study suggest that with the use of convenient information platforms, residents can easily record their home testing results and smoking and drinking behaviors, and evaluate their health status. Moreover, health education and health services from the family doctor team can be easily obtained through an applet. These functions can encourage residents to actively obtain health knowledge, actively monitor health indicators, and actively participate in health interventions, which could play a positive role in improving residents' health literacy and improving health behaviors.

We realized that risk assessment and management are very important in health management [32,33]. Therefore, in the active health information platform, we established a risk management information system and configured various evaluation tools in the system, including our self-developed evaluation model and widely used stroke cards and scales (ASCVD, SCORE, and FRS) from domestic and international sources. The system could interconnect the "i Active Health" applet, community health information system, hospital information system, and network map. The patient's risk assessment score and risk level of cardio-cerebrovascular disease could be automatically calculated in real time, allowing early identification of high-risk groups, accurate positioning, early warning, and early diagnosis and treatment. In the past 2 years, we conducted ASCVD risk assessment and hierarchical management for adult residents who completed annual health examinations in community medical institutions and were registered in community health service centers. The practice has proven that we have effectively reduced the risk of ASCVD among residents in the experimental streets by virtue of the information platform, collected health information of residents in an all-round way through the medical team, and carried out health risk assessment and hierarchical management. This was a useful attempt to carry out the entire process management of ASCVD.

Focusing on the problems of a low disease control rate, low medication compliance, and unhealthy life behavior of chronic disease patients, we built a refined management model of hypertension and diabetes based on internet technology, and established a multidisciplinary management team consisting of general practitioners, community nurses, health managers, specialists, clinical pharmacists, dietitians, and sports rehabilitation specialists. Online and offline fine management for patients with hypertension and diabetes in the community were provided. We selected 2091 patients with hypertension and 888 patients with diabetes who were enrolled in 2020, and they completed a 2-year intervention for statistical analysis. The results showed that after the 2-year internet-based management for hypertension and diabetes, the patients' blood pressure and blood sugar decreased significantly, and their control rate of blood pressure and blood sugar increased significantly. Moreover, patients' BMI, waist circumference, blood uric acid, total cholesterol, and serum LDL-C decreased after the intervention compared to the values before the intervention. Furthermore, patients' self-measured blood pressure, blood sugar, medication adherence, and other medical behaviors significantly improved after the intervention, and their healthy behaviors, such as balanced dietary habits and regular exercise, significantly increased. However, the drinking

XSL•FO
RenderX

behavior of patients with hypertension did not significantly improve. To explore the reasons, we consulted relevant literature and conducted interviews with some patients. Some articles reported that internet-based hypertension health interventions did not significantly improve drinking behavior, which is consistent with the results of this study. Interviews with patients found that most patients had a significant increase in their awareness of the dangers of drinking, and their alcohol consumption was controlled after the intervention. However, drinking is an effective way of communication for Chinese people and might be a need for work [34], and they did not significantly reduce the frequency of drinking. Therefore, our further research will continue to attempt to enhance the effectiveness of an intervention for alcohol consumption behavior in patients with hypertension.

The results suggested that interdisciplinary team collaboration based on internet technology to carry out refined management of chronic diseases and create an online and offline precise health closed-loop service system might have a good effect on improving patient health behavior and improving the disease control rate [35-38]. Based on the above results, our digital health management model has a good effect on the health of both residents and patients and thus has high medical application value. However, digital medicine has not yet been widely promoted and applied worldwide though it could effectively improve the effectiveness of chronic disease management. According to the findings of this research and the literature, we have some strategic suggestions. First, further improve the convenience and equipment dependence of internet-assisted disease management technology with the help of artificial intelligence technology, and constantly improve the effectiveness and popularization of the internet-based chronic disease management model to allow more patients to participate and improve the management effect. Second, strengthen health education, change the old ideas on chronic disease management for both the chronic disease management team and patients, and encourage both doctors and patients to actively try new models of chronic disease management.

In conducting digital active health interventions for residents, we have realized 2 crucial factors. First, in-depth cooperation between the government and health institutions is important to achieve an active health system, which is essential for improving residents' active health literacy. Usually, the work performed by the health system is related to enhancing knowledge, guiding health behaviors, and intervening in health behaviors, but it is difficult to improve the health environment. With deep participation by the government, the difficulties in improving the health environment, integrating resources, and organizing residents to participate in active health activities can be overcome. Second, the internet and artificial intelligence play crucial roles in enhancing residents' learning and participation enthusiasm. At present, it is difficult to achieve good health intervention effects by simply using traditional health education methods, such as posters, flyers, and lectures. Therefore, digital health platforms and wearable devices, which are convenient to use, have a good sense of experience, are strongly interactive, and facilitate continuous intervention, play catalytic roles in improving the effectiveness of active health management [39,40]. We realize that the active health management model is new in China, and it is very difficult to build active health experimental streets. Moreover, we are aware that there are difficulties in promoting and applying this model. At present, there is a long process to change from "passive health" to "active health" for Chinese residents, health systems, and the government. However, the digital active health management model may provide direction for China to comprehensively carry out active health services in the future.

Limitations

Although this study pioneered the exploration of a digital active health management model in China and achieved some satisfactory results, there were some limitations. First, owing to the funding and manpower constraints of the research group, we only selected 2 streets as experimental streets for intervention research, which may not fully represent the current community health status in China. Second, we adopted a single-arm study design for comparison before and after the intervention but were unable to establish a control street for the general population in this region. Third, this study carried out a 2-year intervention and only analyzed the residents' health literacy, medical behavior, and life behavior, as well as the effect of hypertension and diabetes health management. It failed to compare the long-term outcomes and health economic indicators. Thus, the effectiveness of our digital active health management model needs to be further validated through studies involving larger sample sizes, longer assessment periods, and more in-depth analyses in the future.

Conclusions

This study demonstrates that our digital health platform can effectively achieve the interconnection and exchange of different health information. The digital active health management carried out with the assistance of this platform improved the effectiveness of community chronic disease management. Thus, the platform is worth promoting and applying in practice.

Acknowledgments

We are grateful to Professor Thomas Paraidathathu's research team for allowing the free use of the Medication Compliance Questionnaire (MCQ) to evaluate chronic disease management effects in this study. We thank all officers for their support and assistance in the coordination of this study, all community staff and residents who participated in this study, and the reviewers for reviewing and providing valuable feedback from academic perspectives to improve this manuscript. This work was supported by the National Key Research and Development Program (grant 2020YFC2006401), Shenzhen Science and Technology Innovation Project (grant JCYJ20210324135411031), Project of the National Administration of Traditional Chinese Medicine (YGZXKT2024327), and Pingshan District Health System Scientific Research Project (202289).

XSL•FO
RenderX

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Registration guidelines of the “i Active Health” applet.
[PDF File (Adobe PDF File), 124 KB - jmir_v26i1e50959_app1.pdf]

Multimedia Appendix 2
Training materials for using the “i Active Health” applet.
[PDF File (Adobe PDF File), 5389 KB - jmir_v26i1e50959_app2.pdf]

Multimedia Appendix 3
Health index indicators and algorithm.
[DOCX File , 18 KB - jmir_v26i1e50959_app3.docx]

Multimedia Appendix 4
Internet-based refined management of hypertension and diabetes operation manual.
[DOCX File , 1733 KB - jmir_v26i1e50959_app4.docx]

Multimedia Appendix 5
The indicator system and its calculation method of the risk assessment of cardio-cerebrovascular disease.
[DOCX File , 204 KB - jmir_v26i1e50959_app5.docx]

References

1. Healthy China Action Promotion Committee. Healthy China Action (2019-2030). The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/xinwen/2019-07/15/content_5409694.htm [accessed 2024-10-26]
2. Opinions of the state council on the implementation of the healthy China initiative. State Council of the People's Republic of China.: State Council of the People's Republic of China; 2019. URL: http://www.gov.cn/zhengce/content/2019-07/15/content_5409492.htm [accessed 2024-10-26]
3. Hu JP, Xu L, Feng W, Qu XH, Hu WS, Meng Q. Researches on the Theoretical Framework of Initiative Medical Health Service Model. Chinese Journal of Health Informatics and Management 2016 Jun 22;13(3):227-231. [doi: 10.3969/j.issn.1672-5166.2016.03.01]
4. Chen X, Giles J, Yao Y, Yip W, Meng Q, Berkman L, et al. The path to healthy ageing in China: a Peking University–Lancet Commission. The Lancet 2022 Dec;400(10367):1967-2006. [doi: 10.1016/s0140-6736(22)01546-x]
5. Li YH, Wang Y. Construction of modern public health system: considerations from an active health perspective. China Food Drug Admin Magazine 2022 Feb 15;19(2):114-121. [doi: 10.3969/j.issn.1673-5390.2022.02.016]
6. Chen X, Guo J. Construction of service framework of integration of sports and medicine in China under the background of proactive health. J Cap Univ Phys Educ Sports 2021;33(5):474-480. [doi: 10.14036/j.cnki.cn11-4513.2021.05.002]
7. Margariti A, Stathopoulou A, Koulouvari A, Sakellari E, Lagiou A. Digital health and health promotion: A scoping review of systematic reviews. Stud Health Technol Inform 2023 Oct 20;309(2):298-299. [doi: 10.3233/SHTI230802] [Medline: 37869863]
8. Lotman E, Viigimaa M. Digital health in cardiology: The Estonian perspective. Cardiology 2020 Dec 4;145(1):21-26 [FREE Full text] [doi: 10.1159/000504564] [Medline: 31801130]
9. Shilton T, Bauman A, Beger B, Chalkley A, Champagne B, Elings-Pers M, et al. More people, more active, more often for heart health - Taking action on physical activity. Glob Heart 2024;19(1):42 [FREE Full text] [doi: 10.5334/gh.1308] [Medline: 38708404]
10. Zhou H, Wang X, Yang Y, Chen Z, Zhang L, Zheng C, et al. Effect of a multicomponent intervention delivered on a web-based platform on hypertension control: A cluster randomized clinical trial. JAMA Netw Open 2022 Dec 01;5(12):e2245439 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.45439] [Medline: 36477479]
11. Vitija A, Amirthalingam A, Soltani A. The impact of digital interventions on medication adherence in paediatric populations with attention deficit hyperactivity disorder, depression, and/or anxiety: A rapid systematic review and meta-analysis. Res Social Adm Pharm 2022 Dec;18(12):4017-4027. [doi: 10.1016/j.sapharm.2022.07.042] [Medline: 35985977]

XSL•FO
RenderX

12. Kaihara T, Scherrenberg M, Intan-Goey V, Falter M, Kindermans H, Frederix I, et al. Efficacy of digital health interventions on depression and anxiety in patients with cardiac disease: a systematic review and meta-analysis. Eur Heart J Digit Health 2022 Sep;3(3):445-454 [FREE Full text] [doi: 10.1093/ehjdh/ztac037] [Medline: 36712158]
13. The Lancet Digital Health. Digital health equity for older populations. The Lancet Digital Health 2023 Jul;5(7):e395. [doi: 10.1016/s2589-7500(23)00114-0]
14. Su B, Li D, Xie J, Wang Y, Wu X, Li J, et al. Chronic disease in China: Geographic and socioeconomic determinants among persons aged 60 and older. J Am Med Dir Assoc 2023 Feb;24(2):206-212.e5 [FREE Full text] [doi: 10.1016/j.jamda.2022.10.002] [Medline: 36370750]
15. Sun X, Yan AF, Shi Z, Zhao B, Yan N, Li K, et al. Health consequences of obesity and projected future obesity health burden in China. Obesity (Silver Spring) 2022 Sep 24;30(9):1724-1751. [doi: 10.1002/oby.23472] [Medline: 36000246]
16. Liu L, Wu F, Tong H, Hao C, Xie T. The digital divide and active aging in China. Int J Environ Res Public Health 2021 Dec 01;18(23):12675 [FREE Full text] [doi: 10.3390/ijerph182312675] [Medline: 34886400]
17. Sun X, Yan W, Zhou H, Wang Z, Zhang X, Huang S, et al. Internet use and need for digital health technology among the elderly: a cross-sectional survey in China. BMC Public Health 2020 Sep 11;20(1):1386 [FREE Full text] [doi: 10.1186/s12889-020-09448-0] [Medline: 32917171]
18. Screening and comprehensive intervention technology plan for stroke population. Office of the Stroke Prevention and Treatment Engineering Committee of the National Health Commission. 2018. URL: http://www.nhc.gov.cn/jkj/s7930/201906/50f8c202839a4b1c9afd81441e1b88af/files/57ebc443c3b64515b20abb2c397dd29f.pdf [accessed 2024-10-20]
19. Goff D, Lloyd-Jones D, Bennett G, Coady S, D'Agostino R, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014 Jul 01;63(25 Pt B):2935-2959 [FREE Full text] [doi: 10.1016/j.jacc.2013.11.005] [Medline: 24239921]
20. Conroy R, Pyörälä K, Fitzgerald A, Sans S, Menotti A, De Backer G, SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003 Jun;24(11):987-1003. [doi: 10.1016/s0195-668x(03)00114-3] [Medline: 12788299]
21. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 1967 Jul;20(7):511-524. [doi: 10.1016/0021-9681(67)90082-3] [Medline: 6028270]
22. Paraidathathu T, Azuana, Islahudin F, Ahmad. Medication adherence in patients with type 2 diabetes mellitus treated at primary health clinics in Malaysia. Patient Preference and Adherence 2013 Jun:525. [doi: 10.2147/ppa.s44698]
23. Paraidathathu T, Azuana, Nur Sufiza A. Medication adherence among hypertensive patients of primary health clinics in Malaysia. Patient Preference and Adherence 2012 Aug:613. [doi: 10.2147/ppa.s34704]
24. Gong M, Liu L, Wang Y. Strategies for building an active health management model. Science and Technology Bulletin 2022 Mar 28;40(6):93-100. [doi: 10.3981/j.issn.1000-7857.2022.06.011]
25. Cao X, Wu L. Developing health management services for healthy China. Introduction to Chinese Medical Information 2020 Oct 27;35(20):4-5. [doi: 10.3760/cma.j.issn.1000-8039.2020.20.104]
26. Shen Q, Hao J, Qian J. Reflection on the transformation of disease control system construction from the perspective of active health. Administrative Management Reform 2020 Apr 20(4):29-36. [doi: 10.14150/j.cnki.1674-7453.2020.04.004]
27. Chen X, Guo J. Construction of a framework for integrated sports and medical services in China under the background of active health. Journal of Capital Sports University 2021 Oct 18;33(5):474-480. [doi: 10.14036/j.cnki.cn11-4513.2021.05.002]
28. Tan W, Bu Q, Liu J. New practice of "active health" promoting health in Suzhou construction. Health Education and Health Promotion 2019 Feb 20;14(1):10-13. [doi: 10.16117/j.cnki.31-1974/r.201901004]
29. Peng W, Chen S, Chen X, Ma Y, Wang T, Sun X, et al. Trends in major non-communicable diseases and related risk factors in China 2002-2019: an analysis of nationally representative survey data. Lancet Reg Health West Pac 2024 Feb;43:100809 [FREE Full text] [doi: 10.1016/j.lanwpc.2023.100809] [Medline: 38456095]
30. Yip W, Fu H, Chen AT, Zhai T, Jian W, Xu R, et al. 10 years of health-care reform in China: progress and gaps in Universal Health Coverage. The Lancet 2019 Sep;394(10204):1192-1204. [doi: 10.1016/s0140-6736(19)32136-1]
31. Dai T, Xie L, Wang Y, Wan Y, Hu H. Demand analysis and function design of health decision support system in China. Stud Health Technol Inform 2019 Aug 21;264:1383-1387. [doi: 10.3233/SHTI190454] [Medline: 31438153]
32. Guzik A, Bushnell C. Stroke epidemiology and risk factor management. Continuum (Minneap Minn) 2017 Feb;23(1, Cerebrovascular Disease):15-39. [doi: 10.1212/CON.0000000000000416] [Medline: 28157742]
33. Perry JJ, Yadav K, Syed S, Shamy M. Transient ischemic attack and minor stroke: diagnosis, risk stratification and management. CMAJ 2022 Oct 11;194(39):E1344-E1349 [FREE Full text] [doi: 10.1503/cmaj.220344] [Medline: 36220167]
34. Zhou Z, Huang Z, Zheng X, Chen B, Zhang X, Qiu K, et al. Research on the construction and application of "Internet plus" diabetes tripartite co management health management model. China General Practice 2020 Sep 14;18(10):1707-1710. [doi: 10.16766/j.cnki.issn.1674-4152.001599]
35. Kong D, Lu P, Lee Y, Wu B, Shelley M. Health behavior patterns and associated risk of memory-related disorders among middle-aged and older Chinese couples. Res Aging 2023 Feb 17;45(9-10):666-677. [doi: 10.1177/01640275231157784] [Medline: 36800501]

XSL•FO
RenderX

36. Nittas V, Zecca C, Kamm CP, Kuhle J, Chan A, von Wyl V. Digital health for chronic disease management: An exploratory method to investigating technology adoption potential. PLoS One 2023 Apr 13;18(4):e0284477 [FREE Full text] [doi: 10.1371/journal.pone.0284477] [Medline: 37053272]
37. Ownby R, Waldrop D, Davenport R, Simonson M, Caballero J, Thomas-Purcell K, et al. A mobile app for chronic disease self-management for individuals with low health literacy: A multisite randomized controlled clinical trial. medRxiv 2023 Apr 03:2023.04.01.23288020 [FREE Full text] [doi: 10.1101/2023.04.01.23288020] [Medline: 37066256]
38. Lei X, Zhou Y, Wu L. Is it "passive treatment" or "active health"? Behavioral choices of the disease-free population from the perspective of the Healthy China strategy. Medical Controversy 2019 Apr 18;10(4):75-78. [doi: 10.13276/j.issn.1674-8913.2019.04.018]
39. Zhai S, Chu F, Tan M, Chi N, Ward T, Yuwen W. Digital health interventions to support family caregivers: An updated systematic review. Digit Health 2023 May 19;9:20552076231171967 [FREE Full text] [doi: 10.1177/20552076231171967] [Medline: 37223775]
40. Sheng Y, Doyle J, Bond R, Jaiswal R, Gavin S, Dinsmore J. Home-based digital health technologies for older adults to self-manage multiple chronic conditions: A data-informed analysis of user engagement from a longitudinal trial. Digit Health 2022 Sep 22;8:20552076221125957 [FREE Full text] [doi: 10.1177/20552076221125957] [Medline: 36171962]

Abbreviations

ASCVD: atherosclerotic cardiovascular disease
FRS: Framingham scale
HDL-C: high-density lipoprotein cholesterol
LDL-C: low-density lipoprotein cholesterol
MCQ: Medication Compliance Questionnaire
RACCVD: risk assessment model of cardio-cerebrovascular disease
SCORE: Systematic Coronary Risk Evaluation
SRSC: Stroke Risk Scorecard

Edited by A Coristine; submitted 18.07.23; peer-reviewed by D Carvalho, W Zhang, N Deng; comments to author 15.05.24; revised version received 11.06.24; accepted 17.09.24; published 18.11.24.

Please cite as:
Zhou Z, Jin D, He J, Zhou S, Wu J, Wang S, Zhang Y, Feng T
Digital Health Platform for Improving the Effect of the Active Health Management of Chronic Diseases in the Community: Mixed Methods Exploratory Study
J Med Internet Res 2024;26:e50959
URL: https://www.jmir.org/2024/1/e50959
doi:10.2196/50959
PMID:

©Zhiheng Zhou, Danian Jin, Jinghua He, Shengqing Zhou, Jiang Wu, Shuangxi Wang, Yang Zhang, Tianyuan Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Virtual Care on Patient and Provider Experience of the Clinical Encounter: Qualitative Hermeneutic Study

Graham McCaffrey[1], BA, PhD; Erin Wilson[2], BSN, MSN, PhD; Lela V Zimmer[2], PhD; Anurag Singh[3], MBBS, PhD; Steinunn Jonatansdottir[2], MSc; Peter Zimmer[2], MA, MDiv; David Snadden[4], MBChB, MCISc, MD; Ian D Graham[5], PhD; Martha MacLeod[2], PhD

[1]Faculty of Nursing, University of Calgary, Calgary, AB, Canada
[2]School of Nursing, University of Northern British Columbia, Prince George, BC, Canada
[3]Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[4]Northern Medical Program, University of Northern British Columbia, Prince George, BC, Canada
[5]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:
Graham McCaffrey, BA, PhD
Faculty of Nursing
University of Calgary
2500 University Drive NW
Calgary, AB, T2N 1N4
Canada
Phone: 1 4032203516
Email: gpmccaff@ucalgary.ca

Abstract

Background: Virtual health care has transformed health care delivery, with its use dramatically increasing since the COVID-19 pandemic. While it has been quickly adopted for its convenience and efficiency, there has been a relative lack of in-depth exploration of its human impact, specifically how both patients and providers experience clinical encounters.

Objective: This analysis aims to identify and explore themes of change in how patients and providers in a geographically dispersed renal service described their experiences with virtual care, including those changes that occurred during the COVID-19 pandemic.

Methods: Hermeneutics is an interpretive research methodology that treats human experience as inherently interpretive, generating meaning through interactions with others in specific, historically conditioned, social contexts. A total of 17 patients and 10 providers from various disciplines were interviewed by phone as part of a study on health care implementation in the context of a kidney care service in northern British Columbia, Canada. The interview data were analyzed using a hermeneutic approach, which emphasizes careful attention to reported experiences in relation to the relationships and contexts of care.

Results: During analysis, the interdisciplinary team identified themes related to changes in the clinical encounter and how virtual care influenced perceptions of care among both providers and patients. We organized these themes into 2 categories: the structure and content of the encounter. The structure category included the convenience for patients, who no longer had to travel long distances for appointments, as well as changes in care networks. For example, communication between specialist services and local primary care providers became more crucial for ensuring continuity of care. The content category included issues related to trust-building and assessment. Providers expressed concerns about the difficulty in assessing and understanding their patients' physical and social well-being beyond laboratory results.

Conclusions: Patients in the study appreciated the convenience of not needing to travel for appointments, while still having the option for in-person contact with local providers or specialists if their condition changed. Providers were more concerned about the loss of visual cues and sensory data for assessments, as well as the reduced opportunity to build relationships through conversation with patients. Providers also described changes in the locus of control and boundaries, as patients could join phone encounters from anywhere, bypassing traditional privacy and confidentiality boundaries. The study offers a nuanced view of the effects of virtual care on clinical encounters in one setting, seen through the experiences of both patients and providers.

(J Med Internet Res 2024;26:e52552) doi:10.2196/52552

XSL•FO
RenderX

KEYWORDS

virtual medicine; telehealth; professional-patient relations; hermeneutics; kidney; health care facility; British Columbia; Canada; qualitative research; eHealth; health informatics; physician; COVID-19; pandemic; patient experience

Introduction

Virtual patient-provider encounters have increased dramatically in response to the exigencies of the COVID-19 pandemic. It is not known how virtual care will develop in the future, but it is clear that it has become familiar, even routine, to many providers and patients to a far greater degree than anyone might have expected at the end of 2019. We report on findings from a qualitative study of a virtual kidney care service in northern British Columbia, Canada. The focus of our study was to understand human processes in implementing health care innovation, in this case, the introduction of virtual care by a specialized clinic serving a large geographic area. Our study coincided with the COVID-19 pandemic which meant that work processes and use of technology for virtual care changed again in response. In addition to data about implementation processes, the interviews provided a great deal of information about experiences with virtual care from both provider and patient points of view. We brought these data together under the organizing theme of the clinical encounter, seeing it as a human interaction that is connected to institutional structures, clinical goals, professional roles, and of course, technological means of communication.

Recent literature has begun to explore the implications for health care practitioners and service users of virtual care. Virtual care has been defined as “any interaction between patients and/or members of their circle of care, occurring remotely, using any forms of communication or information technologies, with the aim of facilitating or maximizing the quality and effectiveness of patient care” [1]. A scoping review of 8 studies on remote renal care found overall satisfaction, with convenience, especially reduced travel time for appointments, identified as the primary advantage to patients [2]. Barriers were identified that affect communication during consultations, often making it more functional and hindering discussions of sensitive issues. One related concern was patients’ difficulty in finding a quiet, private space conducive to a confidential medical conversation [2]. Another scoping review of virtual renal care highlighted a shift from older audio-visual technology, which often required clinic visits, to the use of portable personal devices [3]. However, this shift was accompanied by concerns about disparities arising from social determinants of health, as well as access to technology and reliable internet connections—issues that are increasingly the responsibility of individuals rather than provider agencies. Positive outcomes included high levels of patient satisfaction and fewer emergency room visits or hospital admissions [3]. Several studies recommended hybrid approaches, where, over the course of a clinical relationship, there would be a combination of in-person and virtual contacts between providers and patients [2,4,5]. A pair of complementary guides issued by the Canadian Medical Association [6,7] made recommendations for conduct during virtual consultations. One suggestion for physicians was to wear a white coat as a visual reminder of the nature of the meeting, in the absence of office or clinic space [6]. For patients, the recommendation was to find a space where they could be alone and uninterrupted [7]. Overall, experience with virtual care continues to evolve, with changes in communication technology and new clinical applications. The resulting need to adapt to new modes of practice is reflected in the development of toolkits for clinicians and patients [8,9], designed to address issues of privacy, informed consent, equity, and appropriateness in virtual care encounters.

Not surprisingly, there has been a surge of literature addressing the rapid, and mostly unanticipated, adoption of virtual care in response to COVID-19 restrictions. Studies seem to reinforce existing trends, with perhaps greater emphasis on concerns about the challenges of ensuring clear communication during virtual appointments, especially over the phone [10-13]. Bajgain et al [10] reported in a systematic review of 102 studies from 20 countries that measures of patient experience showed greater convenience as a strongly positive factor, although communication was often perceived as more rushed or uncertain. Other studies focused on the implementation process itself [11] or on simple measures of patient satisfaction [12], but the effects on clinical relationships were less clear [13]. Our study, drawing on interview data from both providers and patients and exploring in detail their experiences with virtual care, adds to this emerging literature. Our objective was to understand virtual clinical encounters as a form of relational event, shaped by complex interactions with structures of care and the frames of meaning that participants bring to these encounters.

Methods

Study Design

The study used a qualitative, hermeneutic methodology to explore what it means for patients and providers to implement a virtual care service [14]. The subsequent analysis of interviews with patients and providers aimed to understand their experiences with virtual clinical encounters. The study was conducted in northern British Columbia, Canada, in partnership with knowledge users in Northern Health.

The virtual kidney care (Tele-Kidney Care) program [15] is provided through Northern Health, the health authority that delivers health care across the region to 32 communities and 55 First Nations communities. The sparsely populated region covers an area of 617,271 square kilometers—approximately the size of France—with a population of about 300,000 [16]. The kidney care clinic is located in the regional center, a city with a population of 80,000. Virtual care is offered to individuals with predialysis chronic kidney disease and those who have received kidney transplants. Patients attend their local health care facility for virtual appointments and communicate via video with the specialist kidney care staff located in the regional center. Some follow-up communication, particularly during the COVID-19 pandemic, occurred via telephone.

XSL•FO
RenderX

Study participants were 10 providers and 17 patients, all recruited through the kidney care program. Providers (2 men and 8 women) consisted of 8 clinical professionals (a nephrologist, nurses, a dietitian, a pharmacist, and a unit assistant) and 2 information technology specialists. The patients, 6 men and 11 women, were identified by the program's clinical leader as individuals who had provided feedback on virtual services to the kidney care team. They were contacted by the clinical leader in person or by telephone and provided with both oral and written information about the study. Those who chose to participate then contacted the research coordinator directly or consented for a care provider to pass their contact information to the coordinator, who then reached out to them. Potential participants were given further information, and if they chose to proceed, written consent was obtained. All providers and patients who were approached contacted the researchers and volunteered to be interviewed. None withdrew from the study.

Two senior authors (GM and MM) and a PhD research trainee (SJ) conducted 30- to 60-minute interviews by telephone between November 2020 and February 2021. Questions for providers focused on delivering kidney care virtually and how this changed with COVID-19. Questions for patients focused on their initial experience with virtual care for kidney treatment and how they had experienced virtual care since then. The interview questions for patients and providers are presented as Multimedia Appendices 1 and 2, respectively. All interviews were audio recorded and transcribed.

Hermeneutics is a qualitative research methodology rooted in the philosophical tradition of interpretation. It views human experience as emerging from the interrelations between individuals and their social environments, which are shaped historically over time [17]. Understanding of experience is developed through dialogue, where differing perspectives lead to new insights into a topic of interest. These principles have been operationalized in research using interviews [17] to address complex topics of human interrelationships, which are influenced and shaped by contextual factors. In the case of the renal service, geography and technology had significant impacts on how care was structured, delivered, and perceived. Data analysis in hermeneutic research goes beyond identifying themes from interview transcripts to develop interpretations that deepen the understanding of the topic in relation to transdisciplinary sources, revealing aspects that may be significant for practice. Hermeneutics is a suitable methodology for exploring a topic like virtual care encounters, which are relatively new and complex, and require investigation into their emerging implications for practice.

The research team analyzed the data through a systematic process, working iteratively through several stages of interpretation. These interpretations were compared, contrasted, and refined with insights from relevant literature and materials. Initial analysis was conducted by 6 academically based team members, with interpretations carried out by the full team. The team discussed the interpretations while remaining open to questions and alternative viewpoints. They included substantive quotes to preserve the depth and context of the data, ensuring rigor and allowing readers to assess the relevance of the findings in the context of virtual care encounters beyond the specialty focus of renal care.

Ethical Considerations

Harmonized ethical approval for this study was obtained from the University of Northern British Columbia and Northern Health (REB H20-02263). Informed consent was received from all participants, allowing for interpretive analysis of results beyond the original scope of the implementation study. All interview data were deidentified, with identifying details omitted from the reported data. Quotations were organized by theme, not by association with a particular participant. Each participant received a CAD $25 (US $17.86) gift certificate for a coffee shop chain.

Results

Exploring Patient and Provider Perspectives on In-Person and Virtual Care Encounters

When the research team met to analyze the data, we found that participants had a lot to say about their experiences with in-person encounters compared with virtual encounters, as well as video-link versus telephone encounters. What made the data so rich was that we interviewed patients who had experienced both in-person and virtual appointments at different stages of their kidney illness and treatment, and who saw these encounters change again during the COVID-19 pandemic. Combined with interviews with providers, we gained a comprehensive and varied perspective on clinical encounters related to chronic kidney disease. Through team discussions, we identified 2 broad themes.

The first theme focused on the structure of the encounter in a phenomenological sense—how 2 or more people connected around a specific occasion of health care needs, and the networks of human, technological, and material resources, required by both patients and providers, that were invoked in order for encounters to happen. The second theme focused on the content of the encounters—how relationships were formed, cultivated, or inhibited through different modes of interaction between patients and providers. Crucially, in the health care context, it also examined the role of trust in clinical encounters.

Structure of the Encounter: Health Care Systems, Peer Support, and Virtual Network

The rationale for adopting virtual care in regions like northern British Columbia is driven by factors such as distance, low population density, and limited availability of specialized health care resources. Patients clearly expressed how virtual care significantly reduced the time and planning required for appointments. One person compared it with having to make a 4-hour drive each way to the central clinic:

> *I don't have to take a full day off of work and even though my husband attends all my appointments he doesn't have to take full days off of work either and then we don't have to try to, cause we've got animals, we don't have to try to really worry about if they're going to be in the car all day, if we bring them with us. And then if we're staying overnight just not have*

XSL•FO
RenderX

> *to worry about trying to find care for them if we don't bring them with us.*

For this patient, as for most people, the impact goes beyond distance and travel time. There are ripple effects across family, work, and social networks when having to commit an entire day, or even an overnight stay, to attend a medical appointment. Climate and time of year were additional factors promoting the convenience of virtual visits compared with, "...a five hour drive through the [mountain] pass in the middle of winter often times." Another efficiency remarkable to patients was the structure of the visit, where they met with different clinical team members in sequence during a single encounter:

> *I find it very efficient to be able to conference with somebody and have your concerns addressed without you having to physically be in their office. I find that the timing and the appointment time is much, much better on the video things because they make a point of being there and so do you.*

Clinicians also recognized the importance of convenience for patients. One nurse observed, "I'm just guessing here, but I would say 95% appreciate the video conference and not having to travel. Like they choose it now."

Another form of convenience for providers was related to specific role requirements. For example, a pharmacist found that phone call encounters during COVID-19 restrictions were helpful for the discussions she needed to have with patients. As patients were now calling from home, they could easily access all their medications:

> *...on the phone, when I'm phoning them at home I say can you please go get your medications. They might be sitting in their chair or on the phone with me and their medications are in their room or somewhere else and I make them go get them because I need to confirm exactly what they're taking.*

Health Care Network

One immediately important network for patients, understandably, was their health care providers outside the renal team. Although our focus in the study was the kidney care team, participants frequently mentioned other clinicians in relation to their needs and preferences, often referring to primary care, usually their family physician. For some, there was a very positive link between the kidney care team and primary care. Even if they were unaware of the details, they felt reassured that information flowed easily between the 2 parts of the health care system.

> *Ya, a really important part of the team too. It's not just the kidney specialist because I really need someone here so that anything that has to actually be physically looked at for it should be done here instead of having to travel there anyway, so all of it has been through virtual because I have my own doctor here and they keep in touch with the clinic and they know everything that's going on and watch my blood tests every month and everything. They check up on all the stuff.*

Health care networks looked different for patients with comorbid conditions and more complex needs. One patient described an effective communication network involving the kidney care team, 2 other specialist teams managing a comorbid condition, and the primary care physician:

> *...they work in conjunction with each other. Dr. C is the doctor in Vancouver and anything that he does or reports that he gets or tests that he does, Dr. S [another specialist] is made aware of them as well. So they're hand in glove really. And then my GP here in [hometown] gets the same report and even from the [other specialists], all of that has been forwarded to all of them so they know what's going on.*

Another variation in health care networks was seen for peritoneal dialysis patients and providers, where material elements, such as dialysis equipment and the need for hands-on patient education and training, were central. Dialysis care required a network to facilitate the delivery or replacement of supplies, as well as training on when and how these actions were necessary. One patient mentioned that supplies were delivered during COVID-19 restrictions and left on the doorstep. A clinician described nurses in some local communities who "can change transfer sets for us." One of the nurses spoke about a coordinating role involving the nephrologist, nurses, local pharmacies, and families to ensure the timely delivery of dialysis supplies. She would also check that local nurses had the necessary support to assist dialysis patients. For those involved with dialysis, the introduction of virtual care required additional logistics and resources due to the essential material elements in their networks.

Peer Support

Not all the patients in the study mentioned peer support as part of their network, although one participant did bring it up:

> *I remember, well, this lady who became a friend of ours, she had found out about her kidneys going south and this mutual friend asked us, my wife and I to talk to her cause she was just going ballistic thinking life was over. So we went up and talked to her and now she's had a transplant and things are going good for her too.*

The same participant also mentioned the support he and his family received from the Kidney Foundation of Canada, which provided subsidized accommodation in Vancouver during an intensive treatment period that required him to stay in the city.

Content of the Encounter: Trust and Virtual Relationships

People requiring specialty care for kidney conditions maintain ongoing contact with their kidney care team over years or even decades, following a chronic care model. Patients and providers have regular interactions, and when a virtual option was introduced, most patient participants welcomed the change, finding little difference compared with face-to-face encounters. One participant said, "It's pretty much the same thing said and done as if you were actually sitting in a doctor's office" and another, "I mean it's good to meet them in person and talk about your problems...but no, the video conference filled that gap

XSL•FO
RenderX

really easily." While they acknowledged a difference, participants in our study did not place much value on it: "I guess it's a little bit nice seeing a person's face and everything but either way it doesn't bother me too much."

This contrasts with the trade-off of not having to travel, as mentioned earlier.

> *It's a lot of inconvenience to have a 15-20 min conference with a doctor and then turn around and come home again. You've absorbed a whole day, where if I have video conference I can get done in the morning and my day is my own.*

For some providers, by contrast, the loss of face-to-face encounters was seen as eliminating an essential aspect of care. One physician described the trade-off this way:

> *I think we're pretty satisfied with how we're able to provide these video conferencing visits, and I think seeing the patient satisfaction too is a good thing. But I think we all recognize that it doesn't make up for, or it's not as good as seeing someone face-to-face.*

Virtual visits were intended to replace in-person visits, and one provider involved with the technical side of implementation stated, "we pretty much follow the same template as if we're doing in-person care." However, a provider in a clinical management role identified a hierarchy of preference: "...The face-to-face visits are much more important and sometimes the video conference, it's almost like a touch-base in between."

The idea that video or phone encounters are substandard compared with face-to-face visits might be reinforced or complicated by other changes that came with switching to video encounters. One practitioner noted a change in the duration of visits: "...[N]ormally I would see a client for at least 30 minutes, if not maybe 45 minutes...virtually it's more like 15 minutes."

Both patient and clinician participants pointed out a correlation between the stability of illness and the need for in-person encounters. In other words, if a patient's condition became unstable with concerning symptoms or entered a new phase of treatment, such as dialysis, in-person communication became much more important. It should be noted that this flexibility was already built into the kidney program's practices.

A dietitian commented that video images of faces omitted a lot of visual information she would typically observe if seeing a patient in person:

> *...for me assessing someone, like looking at their stature or looking at their legs or their knees or their hands or how their watch fits on their wrist, like I can't see that on video conferencing.... It tells me about muscle wasting or that type of thing or just how they're, ya, muscle wasting, how nutritioned, how well-nourished they are.*

One nurse pointed out a less obvious distinction between video and "face-to-face" contact: "Video versus face to face doesn't change bad news and responses...[but] I just think they're going to hear it better [face to face]."

Sound and Vision

The innovation of the initial virtual kidney care project allowed patients to connect via video from a room in a health care facility in their home community. However, when public health orders related to COVID-19 restricted hospital visitations, all virtual appointments were conducted by phone for the first time. In this context, both patient and provider participants expressed a preference for video over phone visits. One patient commented, "...You get all the information but it's so much easier to actually see the person talking to you" and another said, "I think just as humans we like to see each other. So much of our communication is visual." Providers noted that the visual element helped them gather more information about patients. One nurse who followed transplant patients explained, "...When I see their face I remember their case better than if I'm just talking to them on the phone." A dialysis nurse added, "...I would say on the phone is even worse because you can't judge their reaction."

Virtual visits change many of the ways providers and patients are accustomed to interacting in an encounter, and the loss of sensory data during phone visits led to feelings of discomfort or awkwardness for team members, especially when trying to hit the right conversational tone or (re)establish a relational connection. One nurse noted the shift:

> *...It was so different when it was just over the phone. You don't tend to ask those questions. It's not as easy to just ask someone about, like you can't just work it into the conversation in the same way...The relationship is different, I can't explain it very well.*

As we explored with participants the differences in experiencing virtual encounters, we found that patient participants did not report feeling disconnected or that the encounters were impersonal. Instead, patients consistently described how much they valued the team's availability, their willingness to accommodate patient needs, and how approachable they were. Patients reported still feeling heard by the team and being seen as individuals. One said, "They give me their full attention every time we're on the video screen and they seem genuinely interested in what's going on with me. I haven't seen it go the wrong way as far as the quality of care" and another, "They're interested in what you're doing, and they ask about more than just technical information."

Viewing the Team

The patient participants described a sense of trust in the clinical team, where multiple individuals showed interest in their health, symptoms, and personhood, both in person and virtually. The patient participants described comprehensive care experiences that extended beyond the virtual encounter and seemed to view the team as a cohesive unit. By contrast, providers described a more significant shift in their experience of the encounter and how it impacted their one-on-one interactions with patients when conducting visits virtually rather than in person. One physician put it that,

> *...it's more focused on how is your blood pressure, are you taking your medication, do you have fatigue, shortness of breath, any problems with passing urine?*

XSL•FO
RenderX

> *I don't say how are you doing, you look really good today, looks like you got your hair done, oh wow, okay, it's your son getting married, that's lovely.*

In this study, however, patients described trusting the team and knowing they could rely more or less on the team's resources depending on the trajectory and severity of their condition. One patient, for example, said "...a quarterly touch base or anybody who will respond to an email in between is fine with me. I've become less dependent on them; I don't feel quite so terrified anymore. I've adapted to my life." Another participant described how reassuring it was to know that the team would advocate for their care if needed.

> *My general practitioner is great, but I'd have to tell you that he's far too busy. If I had to go to Emergency, I don't fancy my chances. I sat in Emergency for 5 hours after being told by him that my potassium was super high and I sat there and waited for 5 hours because I looked healthy and I don't fancy my chances. I feel that if I were to send a phone call or a message to them that they would phone ahead and pave my passage.*

This team already had an existing interdisciplinary approach to supporting patients before implementing virtual encounters. For example, there is a "reach-in" mechanism where patients can phone a team member or email a provider with a question without needing to make an appointment. Patients expressed how much they valued knowing they could access the team when needed: "...if I ever have any questions that I need to ask them directly, I can phone the clinic and they will either find out themselves or talk to the doctor and get back to me within a day at the latest." This enabled them to take initiative in their own care: "...it's good to have the opportunity to...instead of having to make an appointment and waiting to speak to somebody, you can actually just call."

A second critical factor in patients' successful acceptance of virtual visits was the preexisting relationship they had with the kidney care team. Patients acknowledged that they felt comfortable during virtual visits because they already knew the team through face-to-face encounters.

> *Because I have had a relationship with them before this, everything moving virtual that it hasn't had an impact on me. I certainly might be wary if I was new to this.*

> *I know these people so talking to them on the phone is okay but if I was new I don't know what my experience would've been.*

Another participant reflected on the value of in-person encounters, saying, "It was a great opportunity to actually talk to them face to face so that now when I talk to them on the phone I kind of know who I'm talking to."

However, a physician expressed that from the team's perspective, new patients who had not experienced in-person visits might not feel the same sense of connection:

> *Now the patients that we've never seen before we all just blend in for them. Because it's just three or four different people that step in and out of in front of their screen. I don't want to say that maybe our message is not so believable anymore but we're just a face now.*

Dialysis: In the Real World

Knowing patients is vital for providers, as it influences how they plan care, including follow-up or monitoring of patient conditions. It was notable that clinicians working with peritoneal dialysis patients were particularly concerned with enabling patients to use equipment to manage their dialysis at home. This raised broader implications about understanding each patient's unique circumstances and needs:

> *It's just me and them and you do get a good idea about how they learn, for one thing, and also you learn some things about their home situation and their life and their support system, that sort of thing. So those people who...I follow up and I'm talking to them frequently. So you have that established relationship. For those people then to do virtual assessment it's easier because you have a good idea of what their baseline is, what they would normally look like, sound like, what their numbers normally would be. But for people who you don't have that, for instance, a patient that I haven't trained, that I've only seen virtually, then I'm lacking all that information and that makes it much more challenging.*

A nurse working with dialysis patients also highlighted the importance of bodily presence and materials in providing care, especially when patients needed to learn hands-on skills for the correct use of equipment at home. One nurse described teaching patients about peritoneal dialysis using a dummy:

> *...so we can show them what it looks like, what it feels like, you know, where it sits in the belly. So I mean we can show them on video but they just don't get the same hands-on and they just don't get a feel for it so just not the same. It's impersonal, right.*

This nurse believed that the time spent teaching skills also provided valuable insights into understanding a patient's circumstances:

> *So I guess we have the good idea because of the face to face contact we've had in the past so we start with [peritoneal dialysis] PD training and I've spent four days with the patient. Usually with their family member as well as teaching them. So it's just me and them and you do get a good idea about how they learn, for one thing, and also you learn some things about their home situation and their life and their support system, that sort of thing.*

Providers in this study recognized that achieving lasting satisfaction with virtual encounters may require learning new skills beyond those typically addressed in standard guidelines for virtual care. Provider job satisfaction is an important factor that contributes to the quality of care [18]. One physician noted, "All these little cues of reassurance and compassion, it's harder. Because we see a face, we don't see a person...I don't have the same level of gratification with virtual appointments." Providers

XSL•FO
RenderX

identified the need to find ways to ensure their assessments were both complete and genuine. One nurse stated:

> *I like to see a face and I recognize faces, so I'm having a hard time if I haven't seen a person for a year, oh man, what did they look like, you know. And then if they were to come in and say oh hi and I'm like who are you?...In virtual care you do see a face but it's just not, it's incomplete I guess is what I would say. You do not have the same assessment, an incomplete assessment.*

A physician also expressed concern about the sense of incompleteness compared with in-person care, framing this as an ongoing challenge in adapting to virtual care:

> *...even though in various jurisdictions there are various levels of experience of virtual care, we're still learning from it. It's like an element of artificialness to it...So many other things that are going in people's lives, they want to share that with us and that's part of their connection with us and it can be better virtually, I'm not saying it's not. It can be, we just haven't developed that. We haven't learned to make the virtual experience be closer to the human experience.*

Discussion

For the purposes of analysis, we divided our findings into 2 themes: structure and content. The introduction of virtual care, along with the shift from video appointments to phone consultations due to COVID-19, created ripple effects in the context of clinical encounters. While patients highlighted the convenience of virtual appointments, they also adapted their networks of care, managing hybrid combinations involving local primary care providers and informal support systems. Patients reported overall satisfaction with the content of virtual encounters and felt that team collaboration continued, even though providers perceived this collaboration as more fragmented. However, providers expressed concerns about constrained communication and the loss of the more holistic assessments typically made during in-person encounters. Another issue arose regarding privacy and confidentiality norms, which are usually controlled by professionals in a traditional clinic setting but were disrupted when patients could join appointments from virtually any location. In the discussion, we interpret various aspects of the encounter, keeping in mind that it is a multidimensional experience. We used broad themes of structure and content as a heuristic to give basic shape to our data. However, in practice, encounters are shaped by a dialectical interchange between both sets of factors. For example, convenience can be seen not only as a matter of time management but also as a sense of control over one's time and resources. From the patients' perspective, there was a high degree of satisfaction with virtual care, though some expressed a preference for in-person visits when managing increases in the complexity or severity of illness. When participants indicated preferences between modes of communication, the video was generally favored over the telephone. On the provider side, concerns were raised about the impact of virtual care on assessment and the relational tone of encounters with patients.

Virtual care is inherently hybrid, consisting of a mix of physical presence, unshared spaces, and communication technology that bridges the distance between 2 bodies in time. While this is often taken for granted, it highlights, from a phenomenological perspective, the various ways in which physical presence and separate spaces are interconnected by clinicians and patients. Paying close attention to the phenomenon allows for the recognition of subtle effects on behavior and power dynamics in clinical encounters—whether positive or negative, depending on the expectations and values of the participants. Recent phenomenological research on health and health care, such as Havi Carel's [19] account of experiencing progressive respiratory disease, has shown the value of carefully examining the structures of experience.

Patient and provider perspectives revealed a marked contrast in the level of concern about the complications of remote communication. Patients felt that virtual visits met their needs, with few concerns raised. While the differences between in-person and virtual visits seemed intangible to patients, those in our study had been part of the kidney care program long enough to compare their experiences: attending the clinic in person, switching on the video in a local health care facility, or using the phone. Past experiences shape expectations in a health care encounter [20], and patients with an established relationship with a provider or team may have more success with virtual visits [13,21,22]. This may also reflect the foundational level of trust patients had in the kidney care team, which was further strengthened by the various "reach-in" features available. These allowed patients to call or email team members to have questions answered or seek advice outside of their scheduled appointments.

While patients still highly valued the ability to have in-person visits, whether scheduled in advance or made available if their condition changed, the need for an in-person encounter was weighed against the convenience of a virtual one. Other recent studies on the increase in virtual care usage have also highlighted the importance of balancing the complexity of illness or severity of symptoms with the patient's desire to have in-person care remain an option [23-25].

Both patients and providers described a hierarchy of preferences for encounters, with face-to-face interactions (either in-person or via video) being preferred over telephone consultations. This preference has been noted in other studies as well [26], underscoring what may be lost without visual cues and sensory data. However, this preference for video consultations contrasts with findings in other research [11], where family practitioners favored shifting to telephone consultations. Participants' language when describing their virtual care experiences reflects an awareness of the distinctions between in-person, video, and telephone encounters. "In-person" encounters clearly referred to times when both patient and provider were physically present in the same room. "Face-to-face" encounters, by contrast, were more ambiguous and could refer to either in-person or video appointments, but not telephone visits. Additionally, a distinction was made based on whether a visit required a

XSL•FO
RenderX

"hands-on" component, highlighting moments when touch or manual dexterity was involved, such as when using a mannequin for learning peritoneal dialysis.

Provider participants in the study acknowledged the convenience virtual visits offered to patients but expressed greater ambivalence about virtual encounters. Zhang et al [27], in a literature review on the effects of virtual care on clinical reasoning, described an increase in ambiguity for clinicians. The ways in which clinicians synthesize a range of subjective and objective data were seen as becoming less integrated, with the style of interviewing often reverting to a more rudimentary and impersonal approach, resembling a checklist of questions about symptoms. One participant gave an example of observing a patient walking into the consulting room, immediately gaining an impression of their difficulty walking—an observation that he did not perceive as significant when it was only reported verbally by the patient.

The scope of attention narrows in on-screen encounters, and even more so in phone encounters. Clinician participants in our study described this as a reduction in the clinical horizon of attention. While pertinent information was not necessarily missed, the range of available information—drawn from sensory perception, interaction, and responses to questions—was notably diminished. This, in turn, related to a sense of trust in the relationship as being continuous and reliable, a quality conveyed more fully by in-person encounters or carried over into virtual encounters based on recollections of previous in-person meetings. Findings from this study support Prasad's [4] argument that virtual care reduces the amount of information available to clinicians, making them more reliant on their own judgment to form a complete picture.

In virtual visits, providers described attempts to fill gaps in information or assessments that would have been easily addressed in in-person visits. Some providers noted that their interviewing style became more rudimentary, impersonal, and focused solely on medical symptoms. Other recent studies on the accelerated adoption of virtual care due to the COVID-19 pandemic have reported similar findings. In a survey of 696 patients and caregivers in Manitoba, Canada, Halas et al [24] noted that virtual encounters were experienced as more transactional. While patients did not report noticing or being concerned by these changes, our data revealed a tendency among patient participants to refer to their "numbers"—a shorthand for their kidney functioning—adopting a medical perspective that views measurable data as an objective indicator of their health status. This phenomenon is not solely a result of remote communication; it may also reflect how patients experience illness as both subjective and objective [28]. Patients might prioritize certain aspects of their health over others if clinicians, adjusting their focus based on a blend of subjective and objective data, have a narrowed horizon of attention in virtual compared with in-person encounters.

"The medium is the message," as McLuhan [29] famously proposed, and virtual visits alter the meaning and significance of a health care encounter. For providers, the difference lay not only in the absence of the full physical presence of the patient but also in the varied settings where encounters now take place. Patients could participate in virtual care from home, their cars pulled over in a school parking lot, airport lounges, or even their workplaces. Virtual care transforms a medical office visit into a more routine, everyday encounter, removing the "ceremony" of the clinic [30]. Some of the unease expressed by provider participants in describing changes with virtual care may stem from the need to renegotiate their sense of legitimacy in a medium that does not readily convey authority or, necessarily, professionalism.

The shift in the locus of control during virtual encounters is a subtle aspect not yet fully explored in the literature. Virtual care toolkits emphasize provider attire, demeanor, privacy, and control over what is visible within the camera's view. While this advice is typically placed under headings related to etiquette or privacy, it may also implicitly reinforce elements of power or legitimacy. Taken further, this may also influence how providers approach an encounter. Amid the many cognitive biases they try to identify in themselves, they may unconsciously reinforce the biomedical aspects of the encounter—not only to compensate for the objective data typically gathered during in-person visits but also to reassure the patient that this is a "real" medical interaction.

This study provides a nuanced understanding of virtual care encounters as experienced by kidney care providers and patients with chronic renal failure in a large, northern, sparsely populated region. One limitation is the specific setting, which may limit generalizability. Additionally, as with most qualitative research, the small sample size focuses on capturing detailed accounts of experiences rather than achieving statistical power. One implication of the small sample size is that we did not collect detailed information about participants' indigeneity, socioeconomic status, or other characteristics, as these could potentially identify individuals in this rural setting. Instead of being directly transferable to similar populations, this study is intended to encourage those engaged in virtual care to consider new perspectives on the nature of the encounter.

By exploring both provider and patient experiences with virtual care, this study adds nuance and complexity to the often simplistic narratives of virtual visits' convenience. It is important to remember that, at its core, virtual care is simply a form of communication between individuals, all of whom ultimately seek to have their bodily needs met and to find relief from suffering. Virtual care is inherently hybrid, and changes in how care is delivered create ripple effects across relationships and networks of care. Against this backdrop, it is well-established that patients and providers often have differing preferences regarding health care delivery [31]. As virtual care plays an increasingly significant role in improving access for patients facing geographical and financial barriers, further exploration of the unintended consequences of altering the medium of care delivery is crucial. Virtual care opens a long-closed gate, allowing patients to access care with minimal disruption to their everyday lives. However, it also raises new questions about legitimacy, power, and trust in health care encounters. This study contributes to the growing body of literature on virtual care post–COVID-19 by encouraging researchers and clinicians to consider the network of physical, social, and emotional

XSL•FO
RenderX

aspects of people's lives that are altered by a shift in the communication medium.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide used in the study for patient interviews.
[DOCX File , 17 KB - jmir_v26i1e52552_app1.docx]

Multimedia Appendix 2
Interview guide used with provider participants.
[DOCX File , 22 KB - jmir_v26i1e52552_app2.docx]

References

1. Canadian Institute for Health Information (CIHI). Virtual care in Canada. CIHI. 2023 Apr 13. URL: https://www.cihi.ca/en/virtual-care-in-canada [accessed 2024-10-10]
2. Ewart C, Baharani J, Wilkie M, Thomas N. Patient perspectives and experiences of remote consultations in people receiving kidney care: a scoping review. J Ren Care 2022 Sep;48(3):143-153 [FREE Full text] [doi: 10.1111/jorc.12419] [Medline: 35338610]
3. Young A, Orchanian-Cheff A, Chan C, Wald R, Ong S. Video-based telemedicine for kidney disease care. CJASN 2021 Dec;16(12):1813-1823 [FREE Full text] [doi: 10.2215/cjn.06660521]
4. Prasad GVR. Enhancing clinical judgement in virtual care for complex chronic disease. J Eval Clin Pract 2021 Jun;27(3):677-683 [FREE Full text] [doi: 10.1111/jep.13544] [Medline: 33559390]
5. Zhao B, Zhang L, Ji P, Lin J, Han J, Li J, et al. Bridging "office-based care" with the "virtual practice care model": evolving care for chronic kidney disease patients in the COVID-19 pandemic-and beyond. Front Med (Lausanne) 2020;7:568201 [FREE Full text] [doi: 10.3389/fmed.2020.568201] [Medline: 33240903]
6. Canadian Medical Association. Virtual care playbook. Canadian Medical Association. URL: https://www.cma.ca/virtual-care-playbook-canadian-physicians [accessed 2024-10-10]
7. Canadian Medical Association. How to navigate a virtual care visit: patient guide. Canadian Medical Association. URL: https://www.cma.ca/resources/how-navigate-virtual-care-visit-patient-guide [accessed 2024-10-10]
8. Canada Health Infoway and Healthcare Excellence Canada. Providing safe and high-quality virtual care: a guide for new and experienced users. Clinician change virtual care toolkit. Version 1.0. Canada Health Infoway and Healthcare Excellence Canada. 2022 May. URL: https://tinyurl.com/5n6f8hm5 [accessed 2024-10-10]
9. Accreditation Canada. COVID-19 toolkit. Virtual care V: 2.0. Accreditation Canada. URL: http://www.phsa.ca/health-professionals-site/Documents/Office%20of%20Virtual%20Health/VirtualCareToolkitV2.pdf [accessed 2024-10-10]
10. Bajgain B, Rabi S, Ahmed S, Kiryanova V, Fairie P, Santana MJ. Patient-reported experiences and outcomes of virtual care during COVID-19: a systematic review. J Patient Rep Outcomes 2023 Dec 01;7(1):126 [FREE Full text] [doi: 10.1186/s41687-023-00659-8] [Medline: 38038800]
11. Hedden L, Spencer S, Mathews M, Marshall EG, Lukewich J, Asghari S, et al. "There's nothing like a good crisis for innovation": a qualitative study of family physicians' experiences with virtual care during the COVID-19 pandemic. BMC Health Serv Res 2023 Apr 04;23(1):338 [FREE Full text] [doi: 10.1186/s12913-023-09256-3] [Medline: 37016330]
12. Buchanan S, Peixoto C, Belanger C, Archibald D, Bjerre L, Fitzsimon J. Investigating patient experience, satisfaction, and trust in an integrated virtual care (IVC) model: a cross-sectional survey. Ann Fam Med 2023 Mar 21;21(4):338-340 [FREE Full text] [doi: 10.1370/afm.2978] [Medline: 36944507]
13. Wu K, Dang Nguyen M, Rouleau G, Azavedo R, Srinivasan D, Desveaux L. Understanding how virtual care has shifted primary care interactions and patient experience: a qualitative analysis. J Telemed Telecare 2023 Apr 18:A [FREE Full text] [doi: 10.1177/1357633x231167905]
14. MacLeod MLP, McCaffrey G, Wilson E, Zimmer L, Snadden D, Zimmer P, et al. Exploring the intersection of hermeneutics and implementation: a scoping review. Syst Rev 2023 Mar 02;12(1):30 [FREE Full text] [doi: 10.1186/s13643-023-02176-7] [Medline: 36864488]
15. Tele-kidney care. Northern Health. URL: https://www.northernhealth.ca/health-topics/kidney-care [accessed 2024-10-10]
16. Moules NM, McCaffrey G, Field J, Laing C. Conducting Hermeneutic Research. New York City, NY: Peter Lang; 2015.
17. Quick facts. Northern Health. URL: https://www.northernhealth.ca/about-us/quick-facts [accessed 2024-10-10]

XSL•FO
RenderX

18. Nundy S, Cooper L, Mate K. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 08;327(6):521-522 [FREE Full text] [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
19. Carel H. Phenomenology of Illness. New York, NY: Oxford University Press; 2016.
20. El-Haddad C, Hegazi I, Hu W. Understanding patient expectations of health care: a qualitative study. J Patient Exp 2020 Dec;7(6):1724-1731 [FREE Full text] [doi: 10.1177/2374373520921692] [Medline: 33457636]
21. Javanparast S, Roeger L, Kwok Y, Reed R. The experience of Australian general practice patients at high risk f poor health outcomes with telehealth during the COVID-19 pandemic: a qualitative study. BMC Fam Pract 2021 Apr 08;22(1):69 [FREE Full text] [doi: 10.1186/s12875-021-01408-w] [Medline: 33832422]
22. Imlach F, McKinlay E, Middleton L, Kennedy J, Pledger M, Russell L, et al. Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. BMC Fam Pract 2020 Dec 13;21(1):269 [FREE Full text] [doi: 10.1186/s12875-020-01336-1] [Medline: 33308161]
23. Bevilacqua M, Melnyk Y, Chiu H, Williams J, Watson P, Lee B, et al. Patient and clinician experiences with the combination of virtual and in-person chronic kidney disease care since the COVID-19 pandemic. Can J Kidney Health Dis 2023;10:20543581231217833 [FREE Full text] [doi: 10.1177/20543581231217833] [Medline: 38107157]
24. Halas G, Baldwin A, Mackay K, Cardenas E, LaBine L, Cherrett P, et al. Patients' and caregivers' experiences of virtual care in a primary care setting during the COVID-19 pandemic: a patient-oriented research study. Digit Health 2024;10:20552076241232949 [FREE Full text] [doi: 10.1177/20552076241232949] [Medline: 38410792]
25. Nayyar D, Pendrith C, Kishimoto V, Chu C, Fujioka J, Rios P, et al. Quality of virtual care for ambulatory care sensitive conditions: patient and provider experiences. Int J Med Inform 2022 Sep;165:104812 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104812] [Medline: 35691260]
26. Mühlbacher AC, Bethge S, Reed S, Schulman K. Patient preferences for features of health care delivery systems: a discrete choice experiment. Health Serv Res 2016 Apr;51(2):704-727 [FREE Full text] [doi: 10.1111/1475-6773.12345] [Medline: 26255998]
27. Zhang Y, Stayt L, Sutherland S, Greenway K. How clinicians make decisions for patient management plans in telehealth. J Adv Nurs 2024 Sep;80(9):3516-3532 [FREE Full text] [doi: 10.1111/jan.16104] [Medline: 38380577]
28. Bhatti S, Dahrouge S, Muldoon L, Rayner J. Using the quadruple aim to understand the impact of virtual delivery of care within Ontario community health centres: a qualitative study. BJGP Open 2022 Sep 15;6(4):BJGPO.2022.0031 [FREE Full text] [doi: 10.3399/bjgpo.2022.0031]
29. McLuhan M. Understanding Media: The Extensions of Man. Cambridge, MA: MIT Press; 1964:7.
30. Strong P. The Ceremonial Order of the Clinic: Parents, Doctors and Medical Bureaucracies. London, UK: Routledge; 1979.
31. Harrison M, Milbers K, Hudson M, Bansback N. Do patients and health care providers have discordant preferences about which aspects of treatments matter most? Evidence from a systematic review of discrete choice experiments. BMJ Open 2017 May 17;7(5):e014719 [FREE Full text] [doi: 10.1136/bmjopen-2016-014719] [Medline: 28515194]

Edited by A Mavragani; submitted 11.09.23; peer-reviewed by S Kreindler, F Medina; comments to author 02.02.24; revised version received 22.03.24; accepted 19.10.24; published 26.11.24.

Please cite as:
McCaffrey G, Wilson E, Zimmer LV, Singh A, Jonatansdottir S, Zimmer P, Snadden D, Graham ID, MacLeod M
Effects of Virtual Care on Patient and Provider Experience of the Clinical Encounter: Qualitative Hermeneutic Study
J Med Internet Res 2024;26:e52552
URL: https://www.jmir.org/2024/1/e52552
doi:10.2196/52552
PMID:

©Graham McCaffrey, Erin Wilson, Lela V Zimmer, Anurag Singh, Steinunn Jonatansdottir, Peter Zimmer, David Snadden, Ian D Graham, Martha MacLeod. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Views and Experiences of Integrated Care System Commissioners About the Adoption and Implementation of Virtual Wards in England: Qualitative Exploration Study

Laura J McGowan[1*], PhD; Fiona Graham[1*], PhD; Jan Lecouturier[1], PhD; Louis Goffe[2], PhD; Carlos Echevarria[3], PhD; Michael P Kelly[4], PhD; Falko F Sniehotta[1,5], PhD

[1]NIHR Policy Research Unit in Behavioural and Social Sciences, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom

[2]Health Determinants Research Collaboration, Gateshead Council, Gateshead, United Kingdom

[3]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

[4]Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

[5]Centre of Preventive Medicine and Digital Health, Division of Public Health, Social and Preventive Medicine, Heidelberg University, Mannheim, Germany

[*]these authors contributed equally

Corresponding Author:
Laura J McGowan, PhD
NIHR Policy Research Unit in Behavioural and Social Sciences
Population Health Sciences Institute
Newcastle University
Baddiley-Clark Building
Richardson Road
Newcastle upon Tyne, NE2 4BN
United Kingdom
Phone: 44 191 2083463
Email: laura.mcgowan@newcastle.ac.uk

Abstract

Background: Virtual wards (VWs) are being introduced within the National Health Service (NHS) in England as a new way of delivering care to patients who would otherwise be hospitalized. Using digital technologies, patients can receive acute care, remote monitoring, and treatment in their homes. Integrated care system commissioners are employees involved in the planning of, agreeing to, and monitoring of services within NHS England and have an important role in the adoption and implementation of VWs in clinical practice.

Objective: This study aims to develop an understanding of the acceptability and feasibility of adopting and implementing VWs in England from integrated care system commissioners' perspectives, including the identification of barriers and facilitators to implementation.

Methods: Qualitative semistructured interviews were conducted with 20 commissioners employed by NHS England (NHSE) in various geographic regions of England. Thematic analysis was conducted, structured using the framework approach, and informed by the Consolidated Framework for Implementation Research. The COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines were followed.

Results: Four overarching themes were identified reflecting the acceptability and feasibility of key adoption and implementation processes: (1) assessing the need for VWs, (2) coordinating a system approach, (3) agreeing to Program Outcomes: NHSE Versus Organizational Goals, and (4) planning and adapting services. Commissioners expressed the need for system-level change in care provision within the NHS, with VWs perceived as a promising model that could reform patient-centered care. However, there was uncertainty over the financial sustainability of VWs, with questions raised as to whether they would be funded by the closure of hospital beds. There was also uncertainty over the extent to which VWs should be technology-enabled, and the specific ways technology may enhance condition-specific pathways. Differing interpretations of the NHSE instructions between different health care sectors and a lack of clarity in definitions, as well as use of hospital-centric language within national guidance, were considered hindrances to convening a system approach. Furthermore, narrow parameters of success measures in terms of goals and outcomes

XSL•FO
RenderX

of VWs, unrealistic timescales for planning and delivery, lack of interoperability of technology and time-consuming procurement procedures, liability concerns, and patient suitability for technology-enabled home-based care were identified as barriers to implementation. Motivated and passionate clinical leads were considered key to successful implementation.

Conclusions: VWs have the potential to reform patient-centered care in England and were considered a promising approach by commissioners in this study. However, there should be greater clarity over definitions and specifications for technology enablement and evidence provided about how technology can enhance patient care. The use of less hospital-centric language, a greater focus on patient-centered measures of success, and more time allowance to ensure the development of technology-enabled VW services that meet the needs of patients and staff could enhance adoption and implementation.

(J Med Internet Res 2024;26:e56494) doi:10.2196/56494

KEYWORDS

virtual wards; remote monitoring; whole systems; qualitative; implementation science; integrated care system; England; digital technology; acute care; clinical practice; semistructured interviews; thematic analysis; patient-centered care; hospital-centric language; eHealth; health services

Introduction

Background

Health and social care services across England are currently facing increasing waiting times and unmet demand [1], creating high levels of system pressures that negatively impact both patients and health care staff. What have been defined as avoidable hospital admissions along with prolonged in-patient stays cause patients distress, limit hospital bed capacity, and are costly to the National Health Service (NHS; the publicly funded health care system in the United Kingdom). Older people and those from socioeconomically disadvantaged areas are at a greater risk of emergency admissions [2,3]. The emergence of COVID-19 has brought new urgency to the adoption of strategies that increase hospital bed capacity and reduce the risk of hospital infections.

Virtual wards (VWs) are models of care delivery whereby patients receive the care they need at home rather than in a hospital. This includes either preventing unplanned hospital admissions or supporting people to safely leave the hospital sooner. There is considerable overlap between VW and hospital at home (HaH) models of care, and these terms are sometimes confused or used interchangeably [4]. Frequently used definitions state that HaH services provide face-to-face care at home through community-based multidisciplinary teams [5], whereas VWs are hospital-led and managed to enable the delivery of at-home acute care, monitoring, and treatment using a variable combination of face-to-face care and remote monitoring (eg, through apps, technology platforms, and wearable devices) [6,7]. The scope of VWs as a model of care varies from models using remote monitoring technology to predominantly operate without in-person intervention (eg, telehealth hubs with remote monitoring) to models more heavily reliant on multidisciplinary teams delivering in-person care in patient homes [8]. Therefore, the latter end of this continuum of care overlaps largely with HaH models. VW or HaH models of care have been used across various countries, including the United States [9,10], Australia [11], the Netherlands [12], and Saudi Arabia [13].

During the COVID-19 pandemic, VWs were implemented in the NHS England (NHSE) to manage some patients with COVID-19 infection through monitoring their oxygen levels at home using a portable pulse oximeter, a small device that attaches to a patient's finger to rapidly measure blood oxygen levels. Patients recorded readings from the pulse oximeter in a diary and were instructed to call the hospital should readings fall below a given level [14]. In some hospitals, oximetry readings were recorded in an app and displayed on a dashboard that the clinical teams could access continuously via a tablet or computer [15]. Since then, VWs have been expanded to different care pathways, primarily for acute respiratory infection and people living with frailty, as per NHSE's (a public body that oversees budget planning, delivery, and operation of commissioning of the entire NHS in England) instruction [16]. A key criterion of this instruction was that VW models should be enabled by technology, that is, the management of patients via a digital platform managed remotely by a clinical team. VWs have been adopted and introduced in different ways in both primary and secondary care settings across England. As a key part of NHSE's plans to improve the responsiveness of emergency and urgent care and increase capacity, the national ambition was to have 40 to 50 virtual beds per 100,000 population by December 2023, the equivalent of up to 24,000 beds in total [16]. As of September 2023, the NHS had delivered 10,000 VW beds, with >240,000 patients treated in VWs [17].

Evaluations of VWs are underway. As of July 2022, evidence from reviews of randomized trials indicated that there is consistently low to moderate certainty evidence that clinical outcomes, including mortality and readmission, for patients treated in VWs are as good or better than for those treated as inpatients [4,18]. The evidence on cost-effectiveness is less clear. Although there have been many studies investigating the costs of VWs to health care providers, nearly all of these studies have methodological problems, which may mean they overestimate cost savings. For example, few studies conducted cost-effectiveness analyses, with most having cost-minimization and cost-saving designs, and many studies disregarded informal care costs to patients and carers [4]. There is low-certainty evidence from reviews of randomized trials that patient satisfaction may be improved by VWs compared to inpatient care [4].

On July 1, 2022, 42 integrated care systems (ICSs) were established across England on a statutory basis [19]. These are

XSL•FO
RenderX

partnerships of organizations (eg, secondary, community, primary, social care, and mental health services) that come together to plan and deliver joined-up health and care services and to improve the lives of people who live and work in their area [19]. Each ICS has an integrated care board (ICB). The ICBs have strategic and financial responsibility for the planning and delivery of health services. Within the NHS, "commissioning" refers to the process of assessing needs, planning and prioritizing, and purchasing and monitoring health services and is undertaken by those in senior leadership roles [20]. The mandate for VWs was developed centrally by NHSE and detailed the national ambitions and instructions for ICSs to then adopt and implement in their own way according to local needs. ICS commissioners and ICB members have been tasked with developing detailed plans on how they intend to move from current practice to using technology-enabled VWs. These plans are to be based on partnerships between secondary, community, primary, and mental health services. ICS commissioners have an important role in the development of implementation plans and the effective adoption of this new model of care in clinical practice.

Effective adoption and rollout of VWs will depend upon the perceived acceptability and feasibility of this care delivery model to stakeholders. The Medical Research Council framework for developing and evaluating complex interventions [21,22] advocates the use of qualitative research to ascertain what is viewed as acceptable or unacceptable to key stakeholders involved in complex health care interventions. The use of qualitative inquiry in intervention evaluation can provide a contextualized understanding of important experiences, enablers, and constraints regarding the implementation of an intervention across a range of settings. Findings from such research can therefore provide useful insights for decision makers working across different contexts.

This qualitative study drew upon the Consolidated Framework for Implementation Research (CFIR) [23]. The use of theory is critical to understanding the likely processes through which an intervention may work [22]. However, there is an array of often-overlapping implementation theories described in the literature, making the selection of a particular theory for implementation evaluation problematic. The CFIR addresses this issue by providing an overarching, comprehensive framework, comprising common constructs from published implementation theories; the CFIR is therefore well-suited to guide the evaluation of the implementation of health care delivery interventions [24]. The updated version of the CFIR [25] was used in this study. This consists of 5 overarching domains (innovation, outer setting, inner setting, individuals, and implementation process), each with multiple constructs (eg, for the innovation domain, constructs include innovation relative advantage, innovation complexity, and innovation cost).

Aims and Objectives

This study aimed to develop an understanding of the acceptability and feasibility of adopting and implementing VWs across ICSs in England from commissioners' perspectives, following the issue of the VW mandate by NHSE in the April 2022. Our objectives were to identify commissioners' views on the potential benefit of VWs within the current health care system, identify the barriers and facilitators to adoption and implementation, and ascertain the ways in which the VW model of care could be optimized in future practice.

Methods

This study is reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 1) [26].

Study Design

We conducted a qualitative study using one-to-one semistructured interviews via an online platform (Microsoft Teams).

Sample

The sample comprised individuals working in commissioning roles in ICSs in England. In total, there are 42 ICSs that each provide health care to around 500,000 to 3 million people. Commissioners are staff in senior leadership roles who are involved in the planning of various health care services, including the purchasing of technologies and monitoring implementation of the VWs program. Purposive sampling was used to select staff from ICSs in different geographical regions and those at different stages of the adoption and implementation process to enable comparisons across sites and to identify important contextual factors affecting implementation.

Recruitment

Potential participants were initially identified by our NHSE collaborators as being eligible to participate in the study based on their job role in NHSE. They were approached about the study, and permission to share their contact details with the research team was obtained. The research team invited them by email to participate. Snowball sampling, whereby participants were asked to identify other potentially eligible participants to invite to the study, was subsequently used. Participants were provided with an information sheet that outlined the requirements for participation and data protection measures.

Data Collection

In-depth one-to-one semistructured interviews were conducted via Microsoft Teams by 2 members of the research team (LJM and FG). Only the researcher and participant were present during the interviews. The topic guide (Multimedia Appendix 2) used broad open-ended questions relating to the study aims rather than mapping questions directly to the CFIR. This enabled respondents to discuss issues of importance to them and encouraged a logical conversational flow to the interview, in line with published methodological guidance [13]. The topic guide was developed by the researchers with input from patient and public involvement and engagement (PPIE) representatives and NHSE collaborators. Members of the NHSE team responsible for the VWs program nationally reviewed the topic guide, and modifications were made according to their feedback. Questions were reviewed and revised during the course of data collection to explore emerging issues in more depth. Interviews were video-recorded via Teams and converted to audio-only (MP3) files for storage. Audio files were transcribed verbatim;

XSL•FO
RenderX

transcripts were anonymized, ensuring all identifiable names, places, and events were removed. Data collection stopped once the research team agreed data adequacy had been achieved [27,28], which was based on the quality and sufficiency of data in terms of the richness of and variations in the data in relation to the research aims and objectives. To aid this judgment, we applied a color-code rating system to our preliminary themes from the first stage of analysis (refer to Data Analysis section): "red" signified an abundance of data with no new insights gained from subsequent interviews; "yellow" signified data were voluminous but new insights or alternative perspectives were still emerging; and "green" signified that data were minimal and more insight was needed through further data collection. Color codes were updated with each subsequent interview until all preliminary themes were coded red.

Data Analysis

The analysis involved 2 separate stages. First, in line with rapid qualitative inquiry methods [29,30], we conducted the preliminary analysis using rapid appraisal procedure sheets. These summarized key findings after each interview based on interview field notes and were then discussed with our policy partners at NHSE. This facilitated the timely dissemination of key findings in a fast-moving policy research area.

The in-depth analysis took place in the second analysis stage. Interview transcripts were analyzed thematically and structured using the framework approach [31,32]. NVivo (version 14; Lumivero) was used to facilitate analysis. The analysis involved both inductive (ie, data driven) and deductive (ie, theory driven) processes and comprised 4 main steps. The first step was reading and rereading transcripts and listening to audio recordings, noting key ideas (to aid familiarization). The second step involved line-by-line coding of transcripts to the updated CFIR [25] (reflecting the deductive element of analysis). CFIR domains were adapted to this study, and a codebook with definitions of domains and subdomains was created. Where it was felt that data did not fit within the framework, new codes were created. In the third step, each CFIR domain was plotted onto a separate thematic matrix, with subdomains ("codes") presented in separate columns and participants ("cases") on individual rows. Data were summarized into cells in each matrix according to the respective case and code. In the fourth step, matrices were reviewed and thematically analyzed (reflecting the inductive element of analysis), drawing upon connections made between codes and cases and reflecting on the overall narrative in relation to the research objectives. This facilitated the determination of the final overarching themes and the key issues and meanings within those themes.

The use of a combined inductive and deductive approach in the analysis helped to ensure comprehensive coverage and theoretical consideration while retaining the narrative and nuance of traditional qualitative approaches [33]. The findings presented in this study's results reflected the final overarching inductive themes derived from the final stage of analysis.

Research Team Reflexivity

Interviews were conducted by 2 female researchers (both holding PhDs and employed as research associates at the time of data collection) with previous training and experience in qualitative interviewing but little prior knowledge of the VW program and no previous relationship to participants. Participants had no personal knowledge of the researchers. This reduced the likelihood of participants providing ambiguous responses relying upon existing shared understandings. Transcripts were coded and analyzed by multiple members of the research team with a variety of disciplinary backgrounds, including health psychology, sociology, behavioral science, and public health. We also included the perspectives of an NHS clinician as well as a PPIE representative within our analysis discussions to ensure that interpretations were considered from multiple perspectives and that findings were relevant to both practice and public interest.

Ethical Considerations

This study was approved by Newcastle University's ethics committee on July 28, 2022 (23846/2022). A consent form was emailed to potential participants before the interview, and informed consent was recorded verbally at the start of the interview. Transcripts were deidentified, with any information that could potentially identify an individual removed or replaced with vaguer descriptions to ensure participant anonymity. Participants were given a unique identification number on recruitment to the study, which has been used when reporting participant quotations in the paper. No compensation, financial or otherwise, was provided.

Results

Participant Characteristics and Context

A total of 20 semistructured interviews were conducted between September 2022 and January 2023 (initial implementation plans for VWs were submitted around April of 2022). No participants dropped out once recruited to the study. Interviews lasted between 30 and 80 minutes. Participants' characteristics are shown in Table 1. There are different models of VW, and their emphasis on technology varies in line with patient needs and preferences. Descriptive information regarding the types of VWs being implemented, as described by commissioners in this study, is available in Multimedia Appendix 3.

Four overarching themes were identified reflecting the acceptability and feasibility of key adoption and implementation processes: (1) assessing the need for VWs, (2) coordinating a system approach, (3) agreement on Program Outcomes: NHSE Versus Organizational Goals, and (4) planning and adapting services.

Overarching themes and subthemes are presented in Textbox 1.

XSL•FO
RenderX

Table 1. Participant characteristics (N=20).

Characteristics	Participants
Age (y), median (IQR)	48 (34-60)
Sex, n (%)	
Female	16 (64)
Male	4 (36)
Professional role, n (%)	
Executive or director	6 (30)
Program lead or manager	6 (30)
Senior manager	6 (30)
Clinical lead	2 (10)
Employer, n (%)	
NHS[a] ICS[b]	5 (25)
NHS Trust	12 (60)
AHSN[c]	5 (25)
Geographic location, n (%)	
North East and Yorkshire	12 (60)
North West	4 (20)
Midlands	1 (5)
South West	3 (15)

[a]NHS: National Health Service.
[b]ICS: integrated care system.
[c]AHSN: Academic Health Sciences Network.

Textbox 1. Overarching themes and subthemes.

Assessing the need for virtual wards

- Tension for change
- Replacing hospital beds versus additionality of service
- Need for technology enablement

Coordinating a system approach

- The translation of instruction: definitions, language, and terminology
- Partnerships and connections

Program outcomes: National Health Service England versus organizational goals

- Targets and evaluation
- Timelines

Planning and adapting services

- Standardization or flexibility
- Identifying interoperable technology “solution”
- Receptivity and need assessments of patients and staff

XSL•FO
RenderX

Assessing the Need for VWs

Tension for Change

Many participants expressed a need for change to the ways in which care is provided within the NHS to counter the high levels of unmet demand for hospital beds and clinician time:

> *The pressure in secondary care is unbearable...the whole system feels just about broken. And if we carry on doing things the way we've done them then we're going to get what we've always got and we've got to do something different.* [P05]

Increasing capability and capacity in community care services to mitigate pressures in hospitals was a key strategic system-level aim with which VWs aligned. VWs also aligned with commissioners' aspirations for the NHS to provide quality patient-centered care and consequently were consistently reported by commissioners to be the "right thing to do:"

> *I'm a passionate believer in community-based care. We hospitalise people far, far too often when we can wrap services around people in their own home and keep them in a safe place. So, it's absolutely the right thing to do.* [P01]

However, there was some concern around the timing of the NHSE VWs instruction to ICSs and whether implementing a service requiring high levels of system-level transformation was appropriate given the infancy of ICSs and the recent pressures of the COVID-19 pandemic. A perceived weariness within the secondary care workforce appeared to lead to a level of despondency among some staff members and negatively impacted the timeliness of delivery of VW implementation plans:

> *Throwing something new in at the time at which an already busy corporate powerhouse is churning to deliver the existing asks. To then get support with something new that they don't understand, that has come completely leftfield, that's got money attached to it, but only if you respond within five minutes, and you've got to plan your workforce, it's a big ask.* [P08]

> *Some of the implementation plans for last year weren't delivering. Now that probably is a consequence of COVID, so we had to support and put in mitigation for supporting them to actually achieve what they could achieve....All really struggled last year with still coming out of COVID.* [P03]

Replacing Hospital Beds Versus Additionality of Service

There was variation in the extent to which participants believed that VWs would reduce the number of patients receiving care in hospitals. It appeared that many participants interpreted the NHSE instruction to mean that referral of patients to VWs would reduce the number of hospital beds required and that costs saved through this reduction in hospital beds (ie, reduced staff resources) could help to provide longer-term funding for VWs. Though some were optimistic that having more patients on VW beds would mean fewer patients cared for in hospitals, many felt that the existing unmet demand for hospital care inevitably meant that the hospital beds "released" by VWs would be taken by someone else:

> *There's this simplistic thought that this person would be in that hospital bed, and now they're in their own bed, so we've then saved that bed. And again, we also know that's not true, because that bed will be used for someone else.* [P18]

In this respect, some participants reflected that it was likely that VWs would always be an addition to the health care system:

> *In real terms, it is about additionality...it's not like we're saying we're closing a hospital ward to fund this. We're saying that we're still going to have a hospital ward open...I think what it will do is provide additional [care delivery] capacity.* [P19]

Some felt that having this additionality of service could improve the flow of patients through the system:

> *For the system, it will hopefully improve patient flow out of hospital. It'll hopefully impact on people being presented at emergency departments or being admitted to hospital. If you can contain somebody at home, manage their acute episode, put the right care in for a certain period of time, and then reduce that.* [P07]

The perception that VWs would unlikely lead to the closing of hospital beds inevitably meant that participants questioned the potential for long-term cost savings:

> *I think it could only be cost saving...if you could close hospital beds...But I can't see that that's possible.* [P06]

However, the lack of cost-saving potential was not equated to the lack of success of the VW program, with clinical outcomes typically considered a priority:

> *Clinical success does not equal financial saving...This will be an additional cost to the system.* [P02]

Need for Technology Enablement

The need to improve and expand community-driven care was widely accepted among commissioners, but there was less consensus about the need to incorporate technology as part of the VW service offer. While NHSE recognizes that there are "different models of VWs, and their emphasis on technology varies in line with patient needs and preferences," the NHSE guidance with regard to the use of technology, published in April 2022, states: "Technology enablement means the management of patients via a digital platform managed remotely by a clinical team" [34]. This may include patients measuring and recording vital signs in an app or continuous monitoring of vital signs via a worn device, which reports to a clinician dashboard, with alerts sent when readings fall outside agreed parameters. Participants reflected on the limited evidence for the advantages of using technology, particularly to remotely monitor patients, and some felt this led to uncertainty and nervousness among clinical staff members and resistance to change, which hindered engagement with the instruction. As such, the technology element of the VW implementation plans was typically less developed than other aspects of the program:

XSL•FO
RenderX

> *You haven't got a machine in front of you, you're going on what the remote monitoring devices are telling you. You can't sit and play with a device. You're not sure that that's giving you the right recording.* [P03]

> *I think there's a bit of a nervousness and a bit of a reluctance to maybe see the benefit of tech...I think it's sometimes easier to revert to what you know, and what you feel comfortable with.* [P13]

Several participants expressed reservations about the extent to which technology could improve or enhance care provision in the community. This was particularly common among participants in areas where existing services do not currently involve digital technology. One participant who was working to make existing HaH services technology enabled to meet the NHSE definition reported:

> *[NHSE] have extrapolated that model into a broader virtual ward model, and I don't think the digital monitoring is the be-all and end-all...frightened breathless patients want somebody at the end of the phone...somebody who can come and see them and...sort them out if they need urgent intervention.* [P05]

Furthermore, the mandate for VWs to be enabled by technology appeared to instill an illogical approach of identifying technology solutions first and attempting to develop a pathway around this, rather than assessing the pathway needs first, and then identifying how technology may facilitate and enhance these needs:

> *[VWs should be based] on a clear clinical need and solving a problem that is already identified as existing, as opposed to creating a problem to solve it.* [P16]

> *I care about the patients, the care we deliver. And if I believe that adding tech will add value then I'll do it. But I won't do it just to get my numbers up, because it's flawed.* [P05]

This was exemplified in an area where various technology solutions had been trialed and implemented in previous projects with the support of an Academic Health Science Network:

> *Those were definitely the most successful, where there was already a need and we were able to facilitate that...The less successful ones were the ones where services were approached and offered it...Where they hadn't already identified that that might be a thing that they want to do...it was trying to shoehorn technology into a pathway where either they weren't ready [or] they didn't really see a need.* [P16]

This highlighted the importance of enabling local places to assess and identify the specific needs of their respective care pathways.

Coordinating a System Approach

Translation of Instruction: Definitions, Language, and Terminology

VW service provision involves both "top-down" instruction from NHSE and "bottom-up" innovation by clinicians in practice. Commissioners were typically responsible for translating the NHSE instruction to care providers and other organizations in their system locally.

There was apprehension among commissioners about whether organizations could successfully come together to deliver the service, given that different sectors had contrasting interpretations of the VW definition and what patients receiving care at home meant:

> *I know there will be different views of what a patient receiving care at home means to people...traditionally that would mean that that patient is under a GP's care, perhaps, and it is more of an outpatient environment.* [P04]

The difficulty in the translation of the instruction appeared to be compounded by the use of hospital-centric language in the NHSE mandate. Several participants reflected that the terminology used to describe VWs was confusing, stemming from an attempt to apply "acute care language" to community services:

> *If we're wanting to manage people in their own home...we need to stop calling them "beds," we need to be talking about caseload instead...the notion that you would have an empty virtual ward bed doesn't make sense.* [P18]

Indeed, some felt the term "VWs" was problematic in itself and equated the model too closely to hospital care rather than placing emphasis on community care delivery:

> *"Virtual wards" itself is quite a problematic term...we are very much a community-led model. "Virtual ward" can make it sound...that you're going to receive exactly the same as what you received when you were at hospital.* [P14]

These reflections highlighted a "hospital-centric" view of the VW service, which may be problematic given that VW models involve the broader health care system, so they could, therefore, impede coordinating a system-level approach to the service. Alternative terms used by a clinician, such as *extra bed days at home* rather than *early discharge* and *alternative to admission* rather than *admission avoidance* [P18], were reportedly preferable.

Participants agreed that each locality would need clear communications to patients and staff about what their VW model offer is and is not and how the service differs from other care pathways available to them, such as being an outpatient. This was challenging for ICS staff to develop centrally as the offer varies across and between trusts. One participant reported that their ICS had reached their own consensus on the scope of their VWs but asserted that there should be a consensus on a national level:

XSL•FO
RenderX

> *I think the national definitions are confusing...If the care at home is replacing an acute admission and there's a treatment plan and the right level of oversight, and we can be assured that if that virtual ward package of support wasn't available, the person would have to be in hospital, then that should count as a virtual ward patient. That's the consensus for you in our ICS. So, I think that needs further debate nationally.* [P06]

This perceived lack of consensus was alluded to by other participants:

> *At the moment we're building a new model of care with no clear understanding.* [P17]

In this respect, it seems more clarity is needed from NHSE to ensure a shared consensus on what constitutes a VW for all stakeholders to facilitate a system approach to the VW program.

Partnerships and Connections

Most partnerships and connections that supported the implementation of VWs were local to place, either across or within ICSs. These connections ranged from forming a network of foundation trusts (NHS organizations demonstrating the highest clinical standards, quality leadership, and record of patient responsiveness and safety) within the ICS to an individual clinician peer-to-peer model of sharing. Primarily, their function was to share learning, best practices, and achievements:

> *What are we doing across our two acute trusts is, learning and sharing across the two. So [trust name] sharing a lot of their ideas around their multi-specialty pathways, because they have actually got a well-established pathway. Then actually, the other trust sharing across with [trust name] as well. It's very much, a sort of, sharing and building kind of model.* [P13]

Some commissioners discussed the usefulness of having a local clinical lead for the VW program who shared knowledge across other ICSs in their region. In particular, having empathy for the challenges associated with the development and implementation of VWs was viewed as important. This suggested the importance of having local VW "champions" who can positively drive the enthusiasm and development of VWs at a regional level:

> *We have a clinical lead at region who is also working on implementing it in her own system, and she has been fantastic. So she has come and talked to our teams about what they're already doing where they are. But she's also, because she's living and breathing it, when I tell her about the challenges, she recognises that that is a genuine challenge, and not saying, "Yes, but why can't you just do it?"* [P18]

Again, at a local level, 1 ICS had help from a commissioning support unit in their region, which facilitated VW implementation. They provided project management support in developing the VW model, with plans and tasks to complete to ensure they met the delivery timescales and national deadlines:

> *That's been really quite helpful because they've been almost like a go between, between us and NHSE if you like, and helping us to develop the plans and things that we need to do in terms of meeting timescales for delivering. Keeping us abreast with everything that's been required nationally so that we don't miss a deadline, or we make sure everything's been delivered that we're meant to.* [P13]

Participants described having convened "multiorganizational" steering groups comprising relevant representatives from across the system to support the development of VW service provision. It was reported that the design of the VW care delivery model was largely led by clinicians in acute trusts. This may be due to clinicians having ultimate clinical responsibility for patients in VWs, leading to resistance to delegate to community staff they do not know or trust. However, participants acknowledged that having the service designed largely by acute staff was problematic, as acute clinicians do not necessarily understand the community care "world," particularly with respect to how technology may serve to support patients in their own homes. This was discussed in terms of risk aversion, with community staff perceived as being less risk averse than staff in acute environments.

> *Community staff will have a different view. Because they're [services are] run with a bigger risk...there's lots of people who are not well who live in the community, who never hit hospital, who are well cared for. So, they'll probably have perhaps more confidence around using remote digital technology.* [P07]

This highlighted the importance of developing and establishing ways of connecting health care staff from traditionally distinct care delivery settings and supporting them to share their knowledge and experience.

Program Outcomes: NHSE Versus Organizational Goals

Targets and Evaluation

Participants reported that, as commissioners, they had a key role in ensuring the targets set by NHSE were met and reporting and monitoring procedures were followed. However, there were tensions between the reporting requirements of NHSE and what commissioners believed were important indicators of success. Many felt that there was too much focus on the number of beds saved at the expense of other measures of success, for example, reduced waiting times:

> *Let's not get too caught up in, how many beds have we saved, let's think about what all of the other measures are that we're trying to achieve,...length of stay,...patient experience,...clinician experience."* [P18]

Concerns were expressed about the "short-termism" of the monitoring and evaluation measures in place that were perceived as "setting ourselves up to fail" (P15):

> *I think we're so politically driven to, on a few metrics...we don't necessarily think about this in a*

XSL•FO
RenderX

> *long-term perspective, i.e. we don't think about some of the big quality markers that really represent value in the long-term for people and for systems.* [P15]

It was clear that the most important markers of success for health care providers were patient experience and outcomes, as opposed to the hard metrics put forward by the national team. This was due to participants being driven by patient-centered care, often cited as a key motivator and driver of behavior:

> *That willingness to put the patient at the centre of what we're doing, and not the service that you work for, and not the organisation that you work for, or not the criteria that's in front of you, but the person in front of you.* [P08]

Other measures of success proposed by participants included patient empowerment and autonomy:

> *I think you empower people, you measure some quality if people feel supported, people feel care's coordinated. And they're some of the big markers of people not wanting or ending up in hospital or dying where they want to die.* [P15]

Timelines

The timescale for implementation was considered too short, with participants perceiving the 2-year funding period to be inadequate to evidence outcomes and evaluate the efficacy of the VW program:

> *We won't have enough evidence to say, "Look, it's working, let's carry on with this," because you won't have had enough time to do that.* [P14]

Two years to transform a system was not considered viable, regardless of the funding available. Many held the view that NHSE should commit funding to enable sites to run VWs for longer and obtain proof of concept, even if this meant less funding in the first instance:

> *It's unrealistic to think that this level of change can be switched on really, really quickly. And I'm yet to see any system, unless they were already on that path, who said, "Oh yes, because we've suddenly got all this money, then we find it really easy to transform services."* [P18]

> *The way we do these things needs to be much more sustainably thought through...starting small and incrementally building up with funding over a longer term would be the way I would always say we should try these things rather than full-scale, two-year implementation, no more money in year three.* [P15]

However, there was acknowledgment that uncertainty over funding across health and social care on a national level would make a prolonged funding model unfeasible:

> *I think the problem is the political element to this and how do you have the ongoing funding over five years when there's so much political uncertainty around funding in the NHS and social care.* [P15]

Without evidence of the impact of VWs, some believed it would be difficult to secure continued funding:

> *So in terms of sustainability, systems are expected to then find that [funding] within their own financial envelope. So it's really quite crucial that we can demonstrate an impact to both patients and cost.* [P19]

There was concern that the lack of time investment given to the VW initiative could lead to the national team overlooking something that is promising but is not being given the required time to provide evidence of real success:

> *I think clinicians do think that this is a way forward, and this is where we should be heading. But by trying to do it too quickly, with ridiculous punitive measures in around success, I think we could throw the baby out with the bathwater.* [P18]

Planning and Adapting Services

Standardization or Flexibility

The various VW models described by participants were all highly complex, requiring multiple members of staff across different organizations working together to deliver care. Where existing community or HaH services were in place, commissioners reported being involved in shaping service provision less frequently, appearing to give trusts more autonomy to adapt or develop their own pathways to meet the VW definition. However, commissioners recognized that this was leading to variability across the system.

Participants in places where services were less developed reported a key part of their role as commissioners was to support delivery planning teams to develop standard operating procedures (SOPs) and processes for VW models:

> *We have to do a lot of influencing and coaching around what a design could look like. Working with business analysts to map out requirements of the workstreams—to understand the current situation and "future state."* [P17]

While many participants reported that local delivery teams were best placed to develop the VW care pathway, some systems had made a conscious effort to create a standardized delivery model and adapt SOPs from other trusts. They were mindful that the templates would be adapted by the trusts to meet their specific needs but thought that this would bring some level of consistency in terms of service provision and equity of access:

> *We want to standardise delivery as much as possible so that the patient access is equitable across the system.* [P06]

Participants identified several important planning activities that were needed to provide care to patients in the community successfully and monitor them remotely, including identifying roles and responsibilities of clinical staff for the duration of patient care, defining patient eligibility criteria and how that is applied, clarifying the need for remote monitoring (frequency and modality), and creating a shared patient record accessible across the partnership organizations.

Several participants expressed the notion that once these key processes had been conducted, the same design approach could

XSL•FO
RenderX

be applied by colleagues to manage patients with other conditions as most governance, IT requirements, and processes could be replicated:

> *Risks will need to be considered and tailored to the new conditions that they're managing but the principles and the process, a lot of it will be the same.* [P09]

Providing general guidance about what implementation processes and considerations are needed may be helpful in ensuring that the VW services offered are of a similar care quality standard.

Identifying Interoperable Technology "Solution"

Incorporating the "technology-enabled" element of the VW mandate was challenging for those with existing VW services that were not technology-enabled. They reported having to develop a digital strategy and adhere to the strict information governance arrangements to ensure that the technology met requirements. One participant reported that they already had such processes completed for a prior service before the VW mandate was announced but commented on the length of time it had taken and noted this as a reason some areas were considerably further behind others in their VW implementation:

> *Things such as the right IG protocols in place, not having e-prescribing programmes in place, not having after care testing in place, etc., not having the shared care record in place. We had to do all of that, and that took us at least 18 months anyway, so we're ahead of the game.* [P01]

Participants described difficulties that teams had faced when attempting to integrate existing remote monitoring devices, procured through previous projects or initiatives, into the VW services they were developing. For example, several participants reported interoperability difficulties, requiring a need for the dashboard to integrate with the patient medical records. They reflected this was an issue that was broader than the VW program:

> *At the moment...we've got 19 different sign-ins for a patient...the same problems that we're hearing with acutes [team]...if we want to really mitigate risk around digital disconnect...we need to make a better online experience.* [P17]

Interoperability of the technology in terms of having shared electronic patient records (EPRs) that all care teams can access in real time was considered critical to ensuring safe services. One participant described a "complete patchwork quilt of tech" (P06) across ICSs, reporting that EPRs are not in place in some organizations. This was considered a wider issue across the NHS but was deemed to be a particular obstruction to the delivery of VWs:

> *The current technology profile across our system does not make the delivery of virtual care as easy as it could be. In fact, for some staff it makes it harder...It's an issue for the NHS full stop, but I think it's a real issue for virtual wards.* [P06]

Interoperability was also a key issue in terms of EPRs for medication prescriptions. One participant noted a near miss with regard to relative contraindication due to general practitioners not having access to VW patient prescribing data:

> *We did have a near miss where...the patient had been off to the doctor for something else and now was changing their medication. There would have been a contraindication with their virtual ward medication. Luckily, it was picked up. Nothing happened, but obviously, it alerted us to the fact that we needed to make sure we got the EPR so that the GPs could know.* [P20]

There was also added complexity as it was deemed important that appropriate information about a patient is shared, where relevant, with the local authority regarding social care and housing arrangements to allow tracing of patients being admitted and discharged from VWs. These data could help to monitor the effectiveness of VWs along with any implications for health inequalities.

Receptivity and Need Assessments of Patients, Carers, and Staff

Although patient and carer need assessments and prospective receptivity to VWs were not typically conducted, there was consensus among commissioners on the patient-level benefits of VWs (eg, being more comfortable, having social support networks around them, less risk of hospital-acquired infection, and being able to sleep and eat better). However, it was acknowledged that VW care models would not be appropriate for all patients due to inadequate home environments, rising costs of living, and digital exclusion and participation issues:

> *There's a very small niche of this absolutely tech enabled person, who's got a smartphone, who's onboarded to tech...got a warm house...got enough food to eat to keep them well...where they can afford to plug in the [implement] and afford the electricity.* [P01]

> *There is a risk that this type of care might only be accessible if you are already digitally enabled...and you're not poor.* [P06]

Participants reported clinical teams were still in the process of planning assessment criteria for patient suitability for VWs. Nevertheless, participants reflected upon the importance of patients being able to make an informed choice about whether to be referred to a VW:

> *Is the patient comfortable with being on a virtual ward as opposed to being on a hospital ward? And that they should have that choice. They might, kind of, fit the definition of the patient cohort that would be appropriate for being on a virtual ward, but are they okay with that?* [P16]

Some VW models required patients to self-test and input their data into an app or a paper diary. This was regarded positively, affording a more "personal approach" to care and giving patients greater autonomy and responsibility for their own care:

XSL•FO
RenderX

> *[Patients] can see then whether their condition is improving or whether it's deteriorating, and they can see what action is going to be taken as part of that.* [P10]

However, there was also concern that self-monitoring adherence may decline over time ("people that started to use the app but, for some reason, have tailed off and stopped;" P11). Thus, further exploration, evidence, and guidance may be required as to what remote monitoring technologies are appropriate for which patient groups and conditions.

Participants also acknowledged the impact of VWs on patient carers, with increased burden identified as a risk in terms of increased responsibility, lack of respite, and potential financial issues:

> *Because otherwise, their loved one would be in hospital having care, and now they're at home, an increased carer stress, and a feeling of additional responsibility...There may be an economic burden for carers that we don't fully understand, if they have the take more time off work.* [P06]

> *Carers who are actually caring for that individual when somebody goes and hospital for a few days, actually gives them a little bit of a break.* [P13]

By contrast, some positive feedback was reportedly received from carers, relating to having a greater understanding and input into the care the patient receives:

> *If their loved one was in hospital...they felt a bit out of the loop and they weren't really sure what was happening or what to expect when the person they cared for came home. Whereas now, they feel part of the conversation.* [P19]

Therefore, it seems imperative that carer perspectives are also built into the design of VW services:

> *Something that we definitely want to do is reach out and have conversations with carers as well as well as people who would access this service to get that understanding.* [P13]

With respect to staff, VW delivery teams (eg, community, acute, telemedicine, urgent care, and out-of-hours teams) needed to have a clear understanding of the service offer and areas of responsibility across pathways and partnerships. One participant reflected on their experience of working with care home staff and emphasized the importance of conducting a skills mapping exercise before introducing technology to understand the training needs of those delivering the program:

If we were to do work again in care homes or the care sector, we'd spend a lot more time upfront doing skills analysis, digital literacy analysis,like whether they're ready for different digital pathways. [P16]

In terms of staff receptivity, participants believed the VWs were more successful in areas where they were led by clinical enthusiasts who had an overwhelming conviction in the VW model:

> *It's having people...who are willing to try things, to really step up and put their head above the parapet and say, "Do you know what? I know we've been doing it this way for 30 years, but this is better." It's that investment in your clinical leadership that you need, to make this change.* [P02]

Indeed, 1 participant noted that their approach was driven through the deliberate identification of clinical enthusiasts in the first instance:

> *So that was our approach really, is go where the enthusiasm is, that those who are keen to do it and, actually, let the rest of the system see the benefits.* [P19]

However, participants reported that enthusiasm for VWs varied among clinicians. Resistance to change among clinical staff appeared to be a barrier to the willingness to adopt VWs, with lower receptivity to VWs perceived to be more common in staff who had worked in the system for longer and were used to caring for patients in more traditional ways:

> *One of the consultants who's really quite dead set against virtual wards is a consultant that's been in the system for quite some time...is [a] strictly traditional consultant [in] the way that he has traditionally managed that work...whereas one of our most enthusiastic consultants is a younger person that hasn't been around for quite so long.* [P19]

Participants also described some clinicians as being nervous about patient safety. Clinicians were apprehensive that VWs diminished opportunities to observe patient symptoms and expressions that may provide additional information to determine clinical decisions:

> *In someone's home, there is no observation. So often you might need to work more with the patient's carer and family to make sure that they are looking out for some of the idiosyncrasies that you may, or may not, think are important.* [P03]

The technology and digital literacy of staff were also highlighted as factors influencing the adoption and implementation of VW services. For some, the VW program was the first time they had used digital devices, and the use of technology was considered a steep learning curve:

> *This was their first attempt at doing remote working, remote monitoring, remote consultations...in some cases they were still quite new to digital at all. So...to then go, "Okay, well we're going to completely change the way you work and bring in this new technology." It was quite a steep learning curve.* [P16]

However, it was felt that adequate training of staff on new technologies mitigated initial apprehension and concern over using technology in VW services:

> *So I think the clinicians [are] finding it [technology] really straightforward and easy...I think there's that little bit of concern and worry prior, but once the*

XSL•FO
RenderX

> *training has all happened, they feel then confident with it.* [P20]

There were also reports of fear and uncertainty among clinicians regarding liability in the case of adverse events and where the ultimate responsibility for VW patients lay, "There is a fear of medico-legal litigation, if things go wrong" (P03). Clarification of these issues was considered an important prerequisite to addressing these concerns and improving engagement.

Furthermore, the limited capacity of clinical staff to plan through the care pathway was highlighted as a hindrance to the development of VW delivery plans:

> *The workforce, within the acute trusts at a consultant level, is minimal...they just don't have the capacity...taking medical oversight and ownership of [virtual wards], is an additionality to them. It's easier for them to continue in the way that things always have been, without seeing the bigger benefits.* [P12]

Discussion

Principal Findings

This qualitative study analyzed commissioners' perspectives on the adoption and implementation of VWs in ICSs in England. Findings demonstrated a strong desire for system-level change in care provision. Though commissioners had an enthusiasm for the VW program as a means to improve patient-centered care, there were concerns about the timing of the mandate and uncertainty over the extent to which hospital beds could be closed to release funds for VWs or whether their purpose is to free bed capacity for those more in need of in-person care. There was also uncertainty over the need to incorporate technology into the VW offer. The VW mandate involved convening a system approach to implementation; however, this was hindered by differing interpretations of the instruction due to a lack of clarity in definitions within the national mandate and the use of hospital-centric language. Other key perceived challenges to implementation included narrow parameters of success, unrealistic timescales to provide evidence of effectiveness the lack of interoperability of technology and time-consuming procurement procedures, and liability concerns. Commissioners felt VWs could offer numerous patient-level benefits but emphasized the need to assess the suitability of patients for the service. Successful implementation was considered contingent upon having motivated and passionate clinical leads.

To our knowledge, no other study has analyzed perspectives on the adoption and implementation of VWs specific to the national 2022 NHSE mandate. However, some of our findings can be understood in the context of research on previous VW and HaH care delivery models. Similar to this study, reviews [4,35] identified patient-level characteristics (eg, technology access, technology literacy, and appropriate home environments) as important issues and emphasized the need for careful consideration of patients' needs and characteristics before enrollment in VWs. Other studies [36] have highlighted the importance of convening multidisciplinary teams and ensuring integrated involvement of both primary care and hospital staff in VW models. This is in line with our findings on the importance of having multiorganizational steering groups, with particular emphasis on connecting acute care and community care sectors. Furthermore, a recent review on the use of technology in telemedicine [37] identified common barriers to implementation, including resistance to change, technology-challenged staff, interoperability, digital exclusion of patients, clinician perceptions of impersonal care, and issues relating to legal liability. These barriers link to our findings regarding the receptivity and needs of patients and staff (resistance to adapt to new ways of providing care, digital literacy of patients and staff, clinician concern over lack of in-person observation of patients, and liability concerns) and identification of an interoperable technology solution. Therefore, it appears these are common barriers across technology-enabled virtual care that need addressing in future models and iterations of VWs.

Strengths and Limitations

Strengths of this study include the involvement of multiple stakeholders in the design and analysis of the research. In addition to academic researchers, we had input from policy, clinical, and PPIE representatives at all stages of the research, including consultation on the topic guide and involvement from the early stages of analysis, to assist with the interpretation of findings and identification of key issues to further probe during data collection. Such involvement helped to ensure the rigor and trustworthiness of the data collected and the diversity in interpretation and relevance of our findings. More broadly, our collaboration with policy makers ensured that this work is policy-relevant, with a direct route to impact decision-making. This was facilitated by our use of rapid appraisal procedure sheets as per the rapid qualitative inquiry approach, ensuring key findings were disseminated quickly in such a fast-moving policy research area.

A further strength of this study is the use of an implementation framework (the CFIR). This helped to guide the analysis and facilitated the comprehensive identification of issues related to implementation. However, our use of inductive analytical techniques following initial coding to the framework ensured that the overall narrative, context, and nuance were retained in our findings [33].

There are also various limitations of this study. First, there was an overrepresentation of ICSs from northern parts of England, particularly the North East and Yorkshire. We attempted to mitigate this by asking our NHSE collaborators to identify commissioners involved in the VW program from southern geographic areas who would be willing to participate; however, our response rate for this was low. This may have been due to the limited capacity of staff in such areas. The overrepresentation of participants from northern ICSs in England may have influenced our findings in terms of the infrastructure and financial circumstances of respective ICSs and the wider social and economic circumstances of areas in northern compared to southern regions. Indeed, there is a worsening divide in terms of health care and health inequality between northern and southern England, with northern areas demonstrating lower life expectancy, higher infant mortality, and worse health and well-being [38]. In light of this, there may

XSL•FO
RenderX

well be different challenges to the implementation of VWs in northern England compared to southern regions, such as the impact of the cost of living and patient suitability in terms of access to technology, digital literacy, and having appropriate home environments. A second limitation is that we were unable to characterize systematically the implementation stages of each ICS represented in the study. These data could have helped to identify more nuance in implementation issues across ICSs.

Implications and Future Research

There are several implications of this research. First, if VWs are to be implemented effectively, there is a need to develop clearer communication strategies between NHSE and ICS leads about the purposes and intended goals of the program. In particular, uncertainty about whether VWs are intended as an additional service to increase care delivery capacity or intended solely to improve the quality of patient-centered care needs to be addressed. This may be achieved through the development of a shared logic model providing clear definitions of what does and does not constitute a VW, how VW models could reduce hospital beds, as well as detailing the potential sustainability of such models in terms of continued funding. Such communications should avoid using "hospital-centric" language and terminology to prevent misinterpretation of VWs as an acute care service, with more focus needed on the community-care aspect. Importantly, a shared consensus needs to be reached on the intended outcomes and success measures of the VW program, including patient-level criteria. Clarification and consensus around these issues could help garner support among operational staff and clinicians—strategic leadership and implementers need to push the same agenda to foster a climate where stakeholders are highly motivated and open to changing their way of working.

In terms of the technology enablement of VWs, there should be further clarification over what types of technology would be suitable for particular pathways, with examples given of how such technologies may enhance care within the pathway. This should mitigate the concern over developing pathways around technology solutions rather than assessing the pathway needs first and then identifying how technology may facilitate and enhance these needs. Providing existing and emerging evidence on the benefits and possibilities of using technology in care provision, particularly with respect to digitally enabled remote monitoring, via a national platform would also help alleviate clinician concern and trust issues regarding technologies, and this increased confidence in technology enablement could help to ease clinicians' resistance to change.

Furthermore, need assessments of both patients and staff should be conducted. Behavioral and skills analysis of different patient populations (eg, respiratory and frailty) should be conducted to develop an understanding of training and support requirements, particularly concerning the use of technology. This could inform the creation of a standardized pro forma to identify suitable patients for a VW. A behavioral and skill analysis of those involved in the implementation of VWs would also be useful to help understand the key barriers and facilitators to implementation tasks and procedures and could inform any required intervention to enable successful implementation. Skill analysis relating to the use of technology and digital literacy would be particularly pertinent and could enable the identification of appropriate training and resources, along with the most suitable type of technology enablement for VW programs. Such behavioral and skill analysis should be grounded in a behavioral framework to identify the most relevant behavior change constructs for each stakeholder group.

In addition, co-design approaches in the design of future iterations of VW models could ensure representation from both acute and community care teams, as well as patient and carer representatives. This could improve the acceptability of VWs to all stakeholders and may help to establish a tolerable level of clinical risk with clear escalation procedures in the case of deterioration in the health of patients.

Further research is also needed to identify and describe the various models of VW currently being implemented across ICSs within the NHS. Such work could enable the development of SOPs that could be distributed nationally and adapted at the local level. Appraisal of the various models could help to identify the most effective model components and could also help to establish the costs involved and predict any potential for future cost savings compared to usual care. Specifically, evidence regarding the effectiveness and cost-effectiveness of the technology components could help to determine the appropriateness and necessity of technology enablement within VW services. Finally, more qualitative research is needed to evaluate the implementation of VWs from staff perspectives as well as patient experiences of being treated in a VW. Such research can feed into future iterations of VW service design and development processes.

Conclusions

This study showed that commissioners involved in the adoption and implementation of VWs in ICSs in England perceived VWs to have the potential to reform some aspects of patient-centered care for some patients. However, for the model to work, there needs to be more clarity over exactly what constitutes a VW, and there should be less use of hospital-centric language in communication strategies. Furthermore, there is a need to include more focus on patient-centered measures of success, for example, clinical safety measures, measures of clinical effectiveness (patient health outcomes), and care quality (patient satisfaction), and sufficient time should be allowed to evidence these. Evidence of patient satisfaction and clinical effectiveness could help to secure future funding, enabling the sustainability of the VW model of care.

Acknowledgments

The authors would like to thank Dan Berry and Henry Dowell (National Health Service England) for their collaboration and contribution to this research. They also thank Irene Soulsby for her contributions from a Patient and Public Involvement and

XSL•FO
RenderX

Engagement perspective. This paper is independent research commissioned and funded by the National Institute for Health and Care Research Policy Research Unit in Behavioural Science (PR-PRU-1217-20501).

Authors' Contributions

LJM, FG, and JL were responsible for conceptualization, formal analysis, methodology, writing the original draft, and reviewing and editing the manuscript. LJM and FG were responsible for investigation and project administration. LG, MPK, and FFS were responsible for conceptualization, methodology, and reviewing and editing the manuscript. CE was responsible for methodology and reviewing and editing the manuscript. MPK and FFS were responsible for funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 27 KB - jmir_v26i1e56494_app1.docx]

Multimedia Appendix 2
Topic guide.
[DOCX File , 36 KB - jmir_v26i1e56494_app2.docx]

Multimedia Appendix 3
Description of virtual wards.
[DOCX File , 24 KB - jmir_v26i1e56494_app3.docx]

References

1. Baker C. NHS pressures in England: waiting times, demand, and capacity. UK Parliament. 2019. URL: https://commonslibrary.parliament.uk/nhs-pressures-in-england-waiting-times-demand-and-capacity/ [accessed 2024-09-26]
2. Sheringham J, Asaria M, Barratt H, Raine R, Cookson R. Are some areas more equal than others? Socioeconomic inequality in potentially avoidable emergency hospital admissions within English local authority areas. J Health Serv Res Policy 2017 Apr;22(2):83-90 [FREE Full text] [doi: 10.1177/1355819616679198] [Medline: 28429977]
3. van den Broek S, Heiwegen N, Verhofstad M, Akkermans R, van Westerop L, Schoon Y, et al. Preventable emergency admissions of older adults: an observational mixed-method study of rates, associative factors and underlying causes in two Dutch hospitals. BMJ Open 2020 Nov 20;10(11):e040431 [FREE Full text] [doi: 10.1136/bmjopen-2020-040431] [Medline: 33444202]
4. Norman G, Bennett P, Vardy ER. Virtual wards: a rapid evidence synthesis and implications for the care of older people. Age Ageing 2023 Jan 08;52(1):afac319 [FREE Full text] [doi: 10.1093/ageing/afac319] [Medline: 36633298]
5. Shepperd S, Iliffe S, Doll HA, Clarke MJ, Kalra L, Wilson AD, et al. Admission avoidance hospital at home. Cochrane Database Syst Rev 2016 Sep 01;9(9):CD007491 [FREE Full text] [doi: 10.1002/14651858.CD007491.pub2] [Medline: 27583824]
6. Milne-Ives M, Shankar R, McLean B, Duun-Henriksen J, Blaabjerg L, Meinert E. Remote electroencephalography monitoring of epilepsy in adults: protocol for a scoping review. JMIR Res Protoc 2022 Feb 25;11(2):e33812 [FREE Full text] [doi: 10.2196/33812] [Medline: 35212630]
7. Tech-enabled virtual wards: relieving pressure on the NHS while caring for patients at home. National Health Service (NHS) England. URL: https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/improving-individual-care-and-patient-safety/virtual-wards-relieving-pressure-on-the-nhs-while-caring-for-patients-at-home/ [accessed 2024-06-05]
8. Lewis C, Moore Z, Doyle F, Martin A, Patton D, Nugent LE. A community virtual ward model to support older persons with complex health care and social care needs. Clin Interv Aging 2017;12:985-993 [FREE Full text] [doi: 10.2147/CIA.S130876] [Medline: 28721026]
9. Lloyd P. Virtual wards: why are they not widely adopted across the US healthcare industry? 2023. Healthcare Technology. URL: https://www.healthcare.digital/single-post/virtual-wards-why-are-they-not-widely-adopted-across-the-us-healthcare-industry [accessed 2024-06-05]
10. Pandit JA, Pawelek JB, Leff B, Topol EJ. The hospital at home in the USA: current status and future prospects. NPJ Digit Med 2024 Feb 27;7(1):48 [FREE Full text] [doi: 10.1038/s41746-024-01040-9] [Medline: 38413704]
11. Nicholson C, Bowler S, Jackson C, Schollay D, Tweeddale M, O'Rourke P. Cost comparison of hospital- and home-based treatment models for acute chronic obstructive pulmonary disease. Aust Health Rev 2001;24(4):181-187. [doi: 10.1071/ah010181] [Medline: 11842709]

XSL•FO
RenderX

12. Utens CM, Goossens LM, Smeenk FW, Rutten-van Mölken MP, van Vliet M, Braken MW, et al. Early assisted discharge with generic community nursing for chronic obstructive pulmonary disease exacerbations: results of a randomised controlled trial. BMJ Open 2012;2(5):e001684 [FREE Full text] [doi: 10.1136/bmjopen-2012-001684] [Medline: 23075570]
13. A look at the Kingdom of Saudi Arabia's Seha virtual hospital – a vision for the future? Health Tech Newspaper. 2023. URL: https://htn.co.uk/2023/07/07/a-look-at-the-kingdom-of-saudi-arabias-seha-virtual-hospital-a-vision-for-the-future/ [accessed 2024-06-05]
14. COVID virtual wards. National Health Service (NHS) England. 2021. URL: https://www.england.nhs.uk/nhs-at-home/covid-virtual-wards/ [accessed 2024-12-15]
15. Vindrola-Padros C, Sidhu MS, Georghiou T, Sherlaw-Johnson C, Singh KE, Tomini SM, et al. The implementation of remote home monitoring models during the COVID-19 pandemic in England. EClinicalMedicine 2021 Apr;34:100799 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100799] [Medline: 33817610]
16. Enablers for success: virtual wards including hospital at home. National Health Service (NHS) England. 2022. URL: https://www.england.nhs.uk/long-read/virtual-wards-operational-framework/ [accessed 2024-04-29]
17. NHS delivers 10,000 virtual ward beds target with hundreds of thousands of patients treated at home. National Health Service (NHS) England. URL: https://www.england.nhs.uk/2023/10/nhs-delivers-10000-virtual-ward-beds-target-with-hundreds-of-thousands-of-patients-treated-at-home/ [accessed 2024-04-29]
18. Echevarria C, Brewin K, Horobin H, Bryant A, Corbett S, Steer J, et al. Early supported discharge/hospital at home for acute exacerbation of chronic obstructive pulmonary disease: a review and meta-analysis. COPD 2016 Aug;13(4):523-533. [doi: 10.3109/15412555.2015.1067885] [Medline: 26854816]
19. Integrated care systems. National Health Service (NHS) England. URL: https://www.england.nhs.uk/integratedcare/ [accessed 2024-06-05]
20. What is commissioning? National Health Service (NHS) England. 2022. URL: https://www.england.nhs.uk/commissioning/what-is-commissioning/ [accessed 2022-07-26]
21. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015 Mar 19;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
22. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
23. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
24. Keith R, Crosson JC, O'Malley AS, Cromp D, Taylor EF. Using the Consolidated Framework for Implementation Research (CFIR) to produce actionable findings: a rapid-cycle evaluation approach to improving implementation. Implement Sci 2017 Feb 10;12(1):15 [FREE Full text] [doi: 10.1186/s13012-017-0550-7] [Medline: 28187747]
25. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148 [FREE Full text] [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]
28. Levitt HM, Motulsky SL, Wertz FJ, Morrow SL, Ponterotto JG. Recommendations for designing and reviewing qualitative research in psychology: promoting methodological integrity. Qua Psych 2017 Feb;4(1):2-22. [doi: 10.1037/qup0000082]
29. Beebe J. Rapid Qualitative Inquiry: A Field Guide to Team-Based Assessment. 2nd edition. Lanham, MD: Rowman & Littlefield Publishers; 2014.
30. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
31. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess B, editors. Analyzing Qualitative Data. London, UK: Routledge; 1994:173-194.
32. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
33. McGowan LJ, Powell R, French DP. How can use of the theoretical domains framework be optimized in qualitative research? A rapid systematic review. Br J Health Psychol 2020 Sep;25(3):677-694. [doi: 10.1111/bjhp.12437] [Medline: 32558289]
34. A guide to setting up technology-enabled virtual wards. National Health Service (NHS) England. 2021. URL: https://transform.england.nhs.uk/key-tools-and-info/a-guide-to-setting-up-technology-enabled-virtual-wards/ [accessed 2024-12-14]

XSL•FO
RenderX

35. Chua CM, Ko SQ, Lai YF, Lim YW, Shorey S. Perceptions of hospital-at-home among stakeholders: a meta-synthesis. J Gen Intern Med 2022 Feb;37(3):637-650 [FREE Full text] [doi: 10.1007/s11606-021-07065-0] [Medline: 34363185]
36. Thomas EE, Taylor ML, Banbury A, Snoswell CL, Haydon HM, Gallegos Rejas VM, et al. Factors influencing the effectiveness of remote patient monitoring interventions: a realist review. BMJ Open 2021 Aug 25;11(8):e051844 [FREE Full text] [doi: 10.1136/bmjopen-2021-051844] [Medline: 34433611]
37. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
38. Munford L, Bambra C, Davies H, Pickett K, Taylor-Robinson D. Health Equity North, Newcastle. 2023. URL: https://www.healthequitynorth.co.uk/app/uploads/2023/04/HEN-REPORT.pdf [accessed 2024-04-29]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
COREQ: Consolidated Criteria for Reporting Qualitative Research
EPR: electronic patient record
HaH: hospital at home
ICB: integrated care board
ICS: integrated care system
NHS: National Health Service
NHSE: National Health Service England
PPIE: patient and public involvement and engagement
SOP: standard operating procedure
VW: virtual ward

Edited by N Cahill, T Leung; submitted 17.01.24; peer-reviewed by L Donelle, N Patel, T Vagg; comments to author 30.04.24; revised version received 24.06.24; accepted 30.09.24; published 27.11.24.

Please cite as:
McGowan LJ, Graham F, Lecouturier J, Goffe L, Echevarria C, Kelly MP, Sniehotta FF
The Views and Experiences of Integrated Care System Commissioners About the Adoption and Implementation of Virtual Wards in England: Qualitative Exploration Study
J Med Internet Res 2024;26:e56494
URL: https://www.jmir.org/2024/1/e56494
doi:10.2196/56494
PMID:

©Laura J McGowan, Fiona Graham, Jan Lecouturier, Louis Goffe, Carlos Echevarria, Michael P Kelly, Falko F Sniehotta. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Dimensions and Subcategories of Digital Maturity in General Practice: Qualitative Study

Timo Neunaber[1], MA; Achim Mortsiefer[2], Prof Dr; Sven Meister[1,3], Prof Dr

[1]Health Care Informatics, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[2]General Practice II and Patient-Centeredness in Primary Care, Institute of General Practice and Primary Care, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[3]Department Healthcare, Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Corresponding Author:
Sven Meister, Prof Dr
Health Care Informatics, Faculty of Health, School of Medicine
Witten/Herdecke University
Pferdebachstr. 11
Witten, 58448
Germany
Phone: 49 230292678629
Email: sven.meister@uni-wh.de

Abstract

Background: The status of the digitalization of companies and institutions is usually measured using maturity models. However, the concept of maturity in general practice is currently unclear, and herewith we examine the question of how maturity can be measured. There is a lack of empirical work on the dimensions and subcategories of digital maturity that provide information on the assessment framework.

Objective: The aim of the study was to answer the question of how many and which dimensions and subcategories describe digital maturity in general practice.

Methods: An explorative, qualitative research design based on semistructured expert interviews was used to investigate the dimensions of digital maturity. Twenty experts from various areas of the health care sector (care providers, interest groups, health care industry, and patient organizations) were interviewed. The interviews were analyzed based on a content-structuring analysis according to Kuckartz and Rädiker using MAXQDA software (VERBI GmbH).

Results: In total, 6 dimensions with a total of 26 subcategories were identified. Of these, 4 dimensions with a total of 16 subcategories (1) digitally supported processes, (2) practice staff, (3) organizational structures and rules, and (4) technical infrastructure and were deductively linked to digital maturity. In addition to the use of digital solutions, digital maturity included, for example, individual, organizational, and technical capabilities and resources of the medical practice. The 2 further dimensions, (5) benefits and outcomes and (6) external framework conditions of the medical practice, were identified inductively with a total of 10 subcategories. Digital maturity was associated with the beneficial use of digitalization, for example, with efficiency benefits for the practice, and external framework conditions were associated with influencing factors such as the local patient situation in the medical practice.

Conclusions: The results indicate that digital maturity is a multidimensional construct that is associated with many dimensions and variables. It is a holistic approach with human, organizational, and technical factors and concerns the way digitalization is used to shape patient care and processes. Furthermore, it is related to the maturity of the organizational environment as well as the benefits of a digitalized medical practice; however, this still needs to be confirmed. To measure the level of digital maturity in outpatient care as accurately as possible, maturity models should therefore be multilayered and take external influencing factors into account. Future research should statistically validate the identified dimensions. At the same time, correlations and dependencies between the measurement dimensions and their subcategories should be analyzed.

(J Med Internet Res 2024;26:e57786) doi:10.2196/57786

XSL•FO
RenderX

KEYWORDS

digital health; eHealth; digital maturity; maturity assessment; primary care; general practice; general practitioner; qualitative research; expert interviews; interview; explorative; dimension; subcategory; expert; care provider; content analysis

Introduction

Overview

"Digitalisation of health systems can significantly improve performance and outcomes" [1]. This is the conclusion reached by the Organisation for Economic Co-operation and Development in its latest Health at Glance report [2]. In its "Global Strategy on Digital Health 2020-2025," the World Health Organization also outlines the goal of strengthening health care through the use of digital technologies [3]. Outpatient care is an essential pillar of health care systems, as it is usually the first contact patients have with health care [4]. It is not uncommon for general practitioners (GPs) to take on the role of gatekeepers who manage the patient's subsequent care steps [5]. Due to the close networking with other service providers and institutions in the health care system, information often must be exchanged along complex care pathways. In addition to exchanging information [6], GPs can use digital technologies for a range of other purposes, such as online appointments [7] or patient treatment [8]. With the opportunities of digitalization for health care systems, calls for a measurement of digital maturity, a concept for determining the progress of digitalization, are becoming louder [2,3]. An assessment of digitalization supports policy makers in identifying gaps for investment in digital health [3]. However, this requires a deeper understanding of digital maturity, for example, for outpatient care.

Background of Maturity Measurements

Maturity measurements are based on the use of maturity models. They are a description of a state with the aim of achieving maturity, that is, "being complete, perfect or ready" [9]. Maturity models are often based on levels that characterize the path to maturity [10,11]. They can be used for self-assessment, benchmarking, or change management, for example, by evaluating and classifying one's own initial state based on a sequence of maturity levels [12,13]. However, this requires measurement dimensions [14]. The dimensions in turn depend on subcategories that represent second-level variables. They are determined by evaluation questions and characterize the dimensions to be measured [14]. Maturity models are used in a wide variety of areas. With the Capability Maturity Model, they were originally used for software development [15]. Application in other fields followed, such as business process management [16] or knowledge management [17]. Maturity models have also become established for measuring digital maturity in the health care sector. The Electronic Medical Record Adoption Model from the Healthcare Information and Management Systems Society is the best-known maturity model for inpatient care [18]. The World Health Organization sees potential in assessing the digital maturity of health care systems to support decisions for national funding measures [3]. In Germany, for example, the Federal Ministry of Health is promoting the measurement of the digital maturity of hospitals [19] and health authorities [20]. Despite the wide range of maturity models, their development and therefore their quality varies [14]. For example, maturity models derived from earlier models dominate [21]. Qualitative and quantitative approaches for determining the dimensions of maturity models, on the other hand, are often lacking. Therefore, empirical validation is required [22].

Maturity Models in Outpatient Care

Maturity models and their measurement dimensions have so far been little researched for outpatient care, in contrast to inpatient care, for which Duncan et al [23], for example, have summarized measurement dimensions of maturity models. For outpatient care, there is literature about digital maturity in aged and community care [24] or commercially available options such as the Healthcare Information and Management Systems Society Continuity of Care Maturity Model [25]. For GPs or medical practices, there is currently only a diffuse understanding of digital maturity, which has only been researched in rudimentary form at best. We came to this conclusion in a recently published scoping review in which we reported an overview of the state of the research literature on the digital maturity of GPs [26]. The research primarily identified gray literature from national initiatives [27-30]. To the best of our knowledge, there are only 2 scientifically published papers dedicated to digital maturity as a whole in outpatient care and with a focus on GPs [31,32]. Teixeira et al [31] used 6 dimensions to measure digital maturity, including the use of digital systems in medical practices, as well as organizational and individual resources and abilities. The work by Haverinen et al [32] attempts to examine the level of maturity in primary and specialized care in Finland. In addition to the use of digital technologies, the regional exchange with other service providers, data security, and individual competencies of users in medical practices are also taken into account [32]. Both studies attempt to measure digital maturity but do not use validated assessment frameworks. There is no guarantee that the variables are truly measuring digital maturity. We were unable to identify any source explicitly dedicated to the validation of maturity models and their measurement dimensions for digital maturity in general practice.

Objective

The absence of literature and the lack of empirical work on maturity models in outpatient care underscore the need for research. Although the benefits of maturity assessment for general practice as a pillar of health care systems are recognized, there is still no answer to the question of which dimensions and subcategories can be used to measure digital maturity. There is a lack of validated assessment frameworks. This is accompanied by a lack of research on dimensions of digital maturity and its subcategories. However, these are necessary to be able to make an adequate measurement. Our research question is, therefore, as follows: Which dimensions and subcategories describe the level of digital maturity in outpatient care using the example of general practice?

XSL•FO
RenderX

The overarching aim is to create the basis for further research into the level of digital maturity in outpatient care, for example, to develop validated maturity models. The specific objectives of this study are (1) to examine how many and which dimensions are associated with digital maturity in general practice and (2) to examine how many and which subcategories describe the measurement dimensions.

Methods

Design and Setting

An explorative qualitative research design based on semistructured, guided expert interviews was used to answer the research question. The decision to use a qualitative research design was made against the background of a lack of empirical approaches for determining the dimensions of maturity models. The aim was to examine the understanding of digital maturity independently of existing maturity models. In this respect, we opted for an inductive research approach to explicitly dedicate ourselves to researching measurement dimensions. This is intended to lay the foundation for possible later quantitative validation. The Consolidated Criteria for Reporting Qualitative Research (COREQ-32) were used to describe the results [33]. The checklist can be found in Multimedia Appendix 1.

Participant Selection and Recruiting

The target group for the interviews was experts on digitalization in the (outpatient) health care sector in Germany. Experts are defined as people with specialized and detailed knowledge in a specific field [34]. The authors discussed whether persons meet the requirements for being experts. We selected, especially, participants who hold leading positions in the field of digitalization in health care. In our study, experts were care providers (GPs, practice managers, and medical assistants), people from interest groups (for care provider, health care industry, and health insurance companies), health care industry (eg, IT service provider), and patient organizations. We set 2 requirements for our sample. First, it should be as heterogeneous as possible in terms of age, gender, and profession. Second, most of the experts interviewed should be care providers from medical practices, as they are directly affected by digitalization. The participants were recruited in various ways. The majority were contacted by email or through LinkedIn (Microsoft). We continue to use specialist congresses and associations to recruit experts. Most recently, we were also able to recruit participants using a snowball method, whereby people who had already been interviewed recommended other experts. When selecting the experts, a parallel documentation of the sample took place, so that the authors controlled the selection of the other experts to the extent that they fulfilled requirements, for example, with regard to heterogeneity, as far as possible. In case of an expression of interest, the interview partners received a declaration of consent as well as specific information on participation. We contacted 32 people during the recruitment process. We did not receive a response from 7 people after contacting them, and we did not get a response from a further 5 people after the initial acceptance, with 1 person cancelling due to time constraints.

Data Collection

For the expert interviews, we used an interview guide that was piloted in advance within the research team and reviewed externally by members of the target group (n=2). Based on the piloting, we refined the wording of some questions. The guide can be found in translated form in Multimedia Appendix 2. An excerpt was sent to the participants together with the information letter and the declaration of consent. The interview guide consisted of 3 sections. The first section comprised demographic questions on age, gender, occupation, and professional experience. To ensure an easy start to the interview, the second part of the guideline contained low-threshold introductory questions. These included, for example, the open question about associations with a digitalized medical practice. The third part contained specific questions on the dimensions of digital maturity in outpatient care. First, general questions were asked about possible criteria for determining digital maturity in medical practices. For the visualization and better understanding of dimensions, reference was made to the “DigitalRadar Hospital,” a current German maturity model for measuring digital maturity in inpatient care, where necessary [19]. The general question on digital maturity was further specified in the areas of “human,” “organization,” and “technology.” We derived this from Cresswell et al [35], among others, according to which digital maturity reflects organizational and human capabilities as well as technical functionalities. Although we referenced the DigitalRadar for the third part in case of comprehension problems, we deliberately decided against a stronger reference to maturity models from inpatient care in order to be able to examine the digital maturity level in outpatient care more freely. If further potentially relevant content outside of the guidelines emerged from the interview, further questions were asked. The participants were also given the opportunity to name further aspects at the end of the interview. The interviews were conducted in the period April-August 2023 exclusively by a male scientist (TN). Due to limited personal scheduling capacities of the experts, TN used Zoom (Zoom Video Communications) to conduct and record the interviews. TN also decided to use Zoom because the software met current data protection and security requirements due to its end-to-end encryption and was accepted by the participants. The software recorded both audio and video; the latter was deleted by the participant immediately after the interview was completed. Field notes were taken by TN during the interviews, for example, to record notes from participants on other experts or topics that seemed relevant. The interviews were scheduled to last from 20 to 30 minutes. On average, the interviews lasted 34 (SD 7.5) minutes. The transcription was based on the transcription rules according to Kuckartz and Rädiker [36]. Furthermore, 1 person exercised their right to comment on the transcribed interview afterward. The participant accepted the transcript on condition that a few linguistic changes were made.

Data Analysis

The interviews were analyzed using a content-structuring qualitative content analysis according to Kuckartz and Rädiker [36]. The MAXQDA 2022 software (VERBI GmbH) was used for the evaluation [37]. The data was initially coded exclusively by the study leader. In accordance with the recommendations

XSL•FO
RenderX

of Kuckartz and Rädiker [36], the main categories were first determined deductively on the basis of the topics in the interview guide. After reviewing the material, further main and subcategories were inductively expanded. To ensure the quality of the coded material, we incorporated aspects of intra- and intercoder reliability. For intracoder reliability, selected aspects of the data material were recoded after a period and compared with the original coding. For intercoder reliability, the code book was discussed together with members of the research team (n=3). All of them were members of the Chair of Health Informatics at the University of Witten/Herdecke and had several years of experience in qualitative research work. A discussion on theoretical saturation also took place in this context. We concluded that we had achieved a high level of theoretical saturation with the interviews of 20 participants. This resulted from the participants' statements, which could be repeatedly assigned to the same main and subcategories. We therefore ruled out increasing the number of participants in view of the low added value [38].

Reflexivity

TN, who was at the time of the study a PhD candidate, conducted the study consequently to the first study [26]. TN therefore had previous knowledge of the research gap in digital maturity in outpatient care and of maturity models in general. In this respect, there was also previous knowledge of possible dimensions of digital maturity. However, as the dimensions were researched independently of any existing maturity models, there was no influence. Instead, the previous knowledge could be used to discuss the results. For data collecting, TN had knowledge of qualitative research, which was acquired, among other things, through participation in a qualitative research workshop at the Interdisciplinary Center for Health Services Research of Witten/Herdecke University [39]. Apart from the level of awareness of people in the field of digital health in outpatient care, there were no previous relationships between the participants and TN. TN placed great emphasis on protecting the rights, privacy, and confidentiality of participants throughout the research process. This concerned, for example, statements by participants from interest groups who pointed out to the author the need for anonymity of the statements. Efforts to establish a good rapport with the participants were made throughout the study. The interviews were individually adapted to the flow of discussion made by each participant.

Ethical Considerations

Ethics approval was obtained from the Ethics Committee of Witten/Herdecke University on March 1, 2023; no ethical or legal concerns were raised (application number S-47/2023). All participants provided complete informed consent before the interview.

Results

Demographics of Participants

We included a total of 20 people in our sample. Care providers such as GPs and medical assistants made up the majority (n=8), followed by people from interest groups and health care industry (n=5). In some cases, experts also took on a dual role. For example, interviewees were active in interest groups and at the same time as care providers. In these cases, we counted the main activity for the demographic data. The analysis can be seen in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of experts interviewed (N=20).

Category	Participants, n (%)
Participant group	
Individual care providers (GPs[a], practice manager, and medical assistants)	8 (40)
Interest groups (for care providers, health care industry, and health insurance companies)	5 (25)
Health care industry (eg, IT service provider)	5 (25)
Patient organizations	2 (10)
Sex[b]	
Male	12 (60)
Female	7 (35)
Age (years)	
30-40	10 (50)
41-50	6 (30)
51-60	3 (15)
<60	1 (5)
Total work experience (years)	
1-10	9 (45)
11-20	5 (25)
21-30	3 (15)
<30	3 (15)

[a]GP: general practitioner.

[b]Not all respondents answered the question.

General Results

The experts' statements on possible dimensions of digital maturity could be assigned to 6 dimensions and 26 subcategories (Figure 1). Deductively, 4 dimensions could be derived that were associated with digital maturity. These included (1) digitally supported processes, (2) practice staff, (3) organizational structures and rules, and (4) the technical infrastructure of the medical practice. Two more dimensions, (5) benefits and outcome of digitalization and (6) external framework conditions of the medical practice, were identified inductively. The fewest subcategories were identified for dimension 1, digitally supported processes, and dimension 5, benefits and outcome (n=3) and the most for dimension 6, external framework conditions (n=7). The individual dimensions are examined in the following sections. Quotes from the participants are used and participant groups are assigned. A table with descriptions of every dimension and subcategory can be found in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 1. Dimensions and its subcategories of digital maturity in general practice.

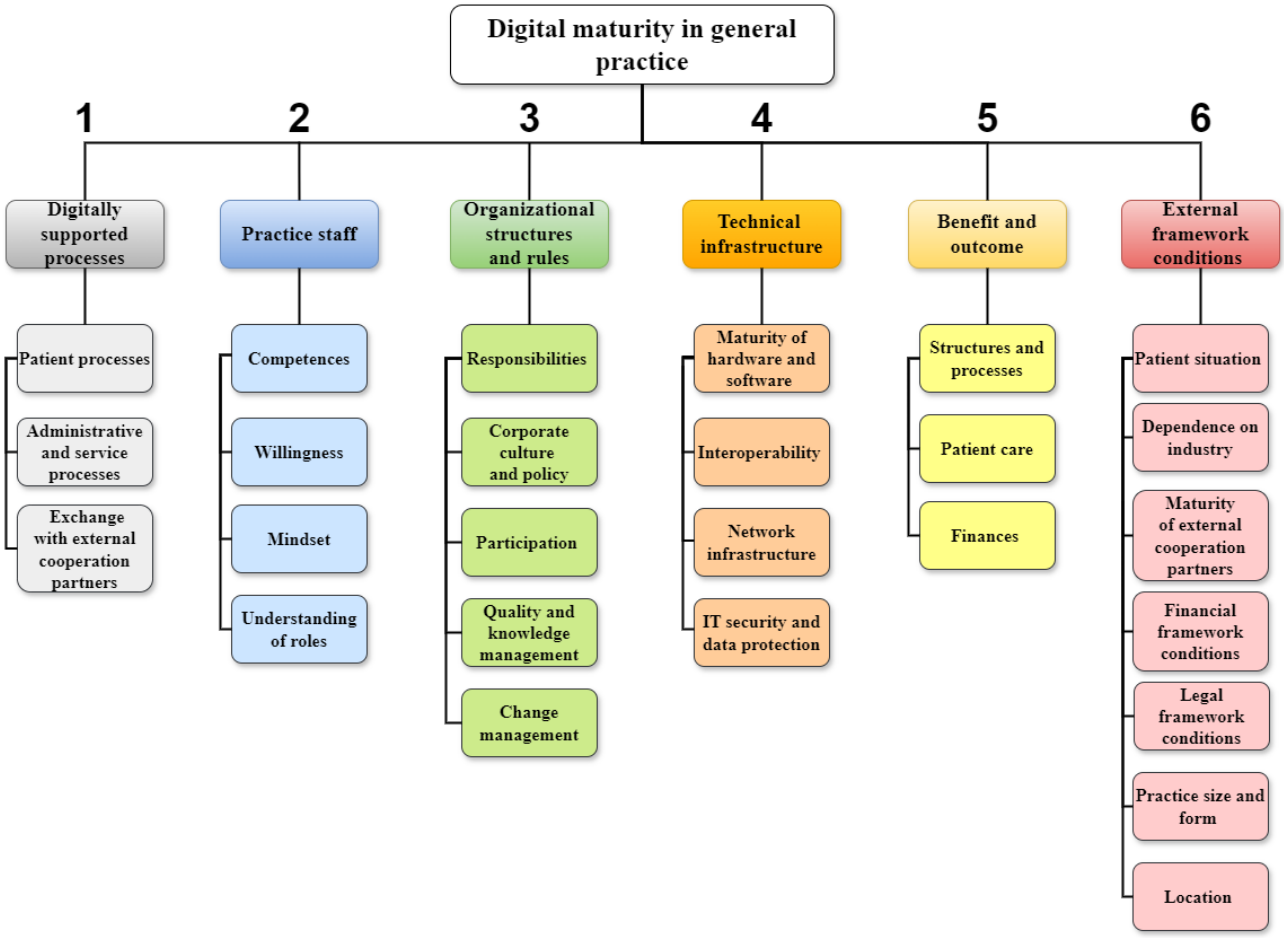

Digitally Supported Processes

The experts interviewed most frequently associated digital maturity with digitally supported processes (dimension 1). The experts' statements could be assigned to 3 subcategories, each representing process areas in the medical practice. Statements on patient processes concerned the use of digital solutions in processes directly related to patients and thus essentially covered the entire patient journey from the search for a medical practice to making an appointment and treating a patient.

> *And then there are the classics, such as online appointment management, video consultations, in other words telemedical services that can be offered to patients, and perhaps telemonitoring services in the future, which I would perhaps also include digital health applications that patients use and through which data should also flow back to the medical practice in the future.* [Participant 1, interest group]

In addition to the administrative processes within a practice, the respondents also addressed communication with external organizations and made the degree of maturity dependent on the communication channel.

> *...i.e. the communication channels to the outside world. Are they digitalized? That would also be another dimension for me, i.e. is the exchange with other physicians, specialists, clinics, health insurance companies by post or fax or is it digitalized?* [Participant 8, patient organization]

Practice Staff

The experts' statements on digital maturity in medical practices could also be assigned to the practice staff (dimension 2). For example, digital maturity depends on the competencies of the practice staff. This included statements on the necessary knowledge of digital solutions and the ability to use them.

> *In my view, however, the question of whether the practice team has digital expertise is very important. If I take a medical practice with a practice owner and several medical assistants and the willingness, but also the level of knowledge about digitalization, telematics, infrastructure, IT equipment is not available, then a very big prerequisite, a very big condition for digitalization to succeed and for processes to be digitalized, is not given.* [Participant 1, interest group]

The individual willingness of practice staff was also identified as a factor. This is primarily related to the willingness to use digital solutions and apply them (correctly).

> *…but of course, there also must be a willingness among people to embrace digital processes and if there are people who block this, then the entire level of digitalization can be jeopardized or counteracted*

XSL•FO
RenderX

> *because it is simply not being lived.* [Participant 7, care provider]

Closely related to this were statements on the mindset of the practice staff and thus on their attitude toward digitalization.

> *Also, the basic attitude, of course, i.e. what do you think about digital applications, digital therapies, diagnostic applications? What do you think about the use of artificial intelligence in diagnostics? Such things could certainly also be asked.* [Participant 8, patient organization]

Finally, the respondents also referred to the professional role perception of the practice staff, which has an influence on the level of digitalization of a practice. Among them, 1 participant, for example, reported on physicians who did not feel obliged to deal with digitalization in accordance with their professional role.

> *...we simply don't have practice managers in these medical practices, we really have physicians who are primarily there for the good of the patient and who prefer to treat patients accordingly.* [Participant 19, health care industry]

The understanding of professional roles was also expressed in the statement of one of the medical assistants interviewed, who on one hand was in favor of digitalization, but on the other hand would also like to maintain a personal exchange:

> *Of course, we now do a lot of things via the PC [personal computer], but I think it's nice when I still know the patients and know what's going on with them if I don't have to look at the PC first.* [Participant 18, care provider]

Organizational Structures and Rules

According to the respondents, the organizational setting with structures and rules of a medical practice also determined the level of digitalization (dimension 3). In structural terms, the organizational structure has a major influence on digital maturity. There is a need for clear responsibilities for the topic of digitalization in the team.

> *In our day-to-day work, we experience that everyone and no one is responsible for [digitalization] and that often causes problems because there is simply a lack of responsibility.* [Participant 20, health care industry]

With responsible persons, such as digitalization coordinators, change management, that is, dealing with changes in practice, could also be made more structured.

> *...if there is a digitalization coordinator, then the practice can respond to the new requirements of digitalization and meet them, whereas in traditional medical practices, where this is not the case, it is often an ad hoc reaction to changes, to new requirements, often with a lot of excessive demands and ignorance of the processes and procedures.* [Participant 1, interest group]

Last but not least, a vision for digitalization and the anchoring of digitalization in the corporate culture and policy of the medical practice were also considered important components of digital maturity.

> *And then there must be a commitment in practice: "Yes, we want digitalization under the maxim that we gain time."* [Participant 13, care provider]

Technical Infrastructure

When describing digital maturity, the experts interviewed referred several times to the area of technical infrastructure within a medical practice (dimension 4). The basic prerequisite for a digitalized medical practice is above all a well-developed network infrastructure.

> *Well, first the complete infrastructure, of course. I think without a well-developed and reliable broadband connection, it will be difficult to digitalize a medical practice, otherwise of course I won't be able to communicate with the outside world and something like video consultations wouldn't even be conceivable.* [Participant 8, patient organization]

The fulfillment of at least legal requirements in IT security and data protection is also important according to the interviewees.

> *..so, this degree of fulfillment of legal requirements in terms of IT security. This is also an important point because it correlates very strongly with digitalization because I would say that if a medical practice uses five systems for digitalization or for digital patient communication, for example, but is as open as a barn door, to put it bluntly, because it has no security, then in my eyes it is not a digitalized medical practice. The overall picture is part of it.* [Participant 20, health care industry]

Interoperable systems were also associated with a high level of maturity within medical practices. The different systems within the medical practice should be able to communicate with each other and enable data exchange.

> *...if examinations are carried out, also with the help of technical aids, it would of course be ideal if the sonography device, if the electrocardiogram device and the like were connected directly to the practice management system, i.e. if data is also transferred accordingly, so that there are no media breaks between the different systems.* [Participant 3, interest group]

However, digital maturity in medical practices was most often associated with the maturity of hardware and software in the practice and thus with factors such as usability, performance, stability, and whether the systems are up-to-date. This is primarily related to the practice management system of a medical practice.

> *...so, if I work with old software systems in the IT equipment, with practice management systems that are more in the category of old billing systems and not modern workflow management systems, then I would say you have a lower level of maturity.* [Participant 1, interest group]

XSL•FO
RenderX

Participants therefore wanted systems to be sufficiently tested before they are used in medical practices.

> *…it has to be applications that have been extensively tested, that simply run, to put it bluntly, that I don't have to deal with in the practice, either as a [medical assistant] or as a physician, or that anything slows down. So, I have nothing to gain from digital processes if I have to call the service desk of my [practice management system] provider five times a day, to put it bluntly.* [Participant 14, interest group]

Benefit and Outcome

In the benefits and outcomes category, the experts emphasized the meaningful and beneficial use of digitalization within the medical practice (dimension 5). A digitally mature medical practice is therefore not only characterized by the extent of digitalization but rather by the added value created by digitalization. The added value mentioned primarily related to structures and processes within the medical practice, which are associated with efficiency benefits and easier work for practice staff. One care provider interviewed emphasized this as follows:

> *I think we need to think about things like process optimization. Where can we really organize things in the process, in the medical practice, so that they don't cost us so much time and so that we only see the patients we need to see? In other words, if you measure now, does the medical practice has a messenger, does the medical practice has a homepage, does the medical practice has an online appointment calendar, I say, then you can survey who uses which digital helpers, but not the degree of digitalization.* [Participant 13, care provider]

Closely related to this, the participants also associated a digitally mature medical practice with benefits for patient care. Digitalization must create real added value for patients.

> *Digitalization must not be a sure-fire success. You don't digitalize something to digitalize it, but so that it makes work easier and relieves physicians so that they have as much time as possible for the patient. That's what I would focus on when you examine something when you measure something. Does this save time, which in the best case is then available for the patient?* [Participant 14, interest group]

External Framework Conditions

The experts interviewed mentioned external framework conditions of medical practices that could influence a possible maturity-level measurement (dimension 6). In addition to examples of dependencies on the industry or the digital maturity of external cooperation partners, the participants cited the local patient situation as an external framework condition. This should not lead to the exclusion of patients who either do not have the necessary skills or are not willing to use digital solutions. Instead, the use of digitalization must above all be adapted to the respective situation of the patient. There are situations in which a personalized approach is still necessary.

> *..but there is a moment when the [patient] wants to be held by another person, in the truest sense of the word, and that is the moment when I need it.* [Participant 11, care provider]

Another participant expressed the following concerns:

> *But when it comes to not being able to speak to the physician at all because you can no longer ask them any questions, then digitalization has overshot the mark again and it still must be possible.* [Participant 9, patient organization]

In addition to limitations due to the patient situation, respondents also referred to other local restrictions, such as the location of the medical practice.

> *..but it also depends on whether the medical practice is in a rural area or in a metropolitan area. This is a very relevant external factor, as is the question of the IT infrastructure in the region.* [Participant 1, interest group]

Highlighted Dimensions

Although this was not explicitly part of the interview guidelines and our objective, the participants also commented on possible connections between individual dimensions and emphasized individual dimensions. Dimension 2, the role of the practice staff, was an example focus. One participant described the willingness of the practice staff as a basic prerequisite for a digitally mature medical practice.

> *I can have the best-equipped medical practice, but if they live well with the paper appointment calendar and don't want to change it, they won't change their processes, not even in the direction of their patients. Not even if the technology allows it. So, it always takes both.* [Participant 1, interest group]

Competences were also seen as a key factor in positive results in other measurement dimensions.

> *If people don't know how to switch a computer on and off, then I probably won't have such great results in the [...] other structures or parameters, dimensions. So, for me, the structural area, the education, the training and so on is the basis for that.* [Participant 3, interest group]

Discussion

Principal Findings

In our study, we investigated the research question of which dimensions and subcategories describe the digital maturity level in outpatient care using the example of general practice. To this end, we conducted expert interviews to understand the concept of digital maturity. Our aim was to contribute to further empirical research into maturity models in outpatient care. Our results can be interpreted as follows. First, we found that digital maturity is associated with a variety of dimensions. From the statements of the experts interviewed, we were able to derive a heterogeneous portfolio of factors ranging from the use of digital applications in processes to individual, organizational, and technical capabilities and the benefit-oriented use of

XSL•FO
RenderX

digitalization. Digital transformation is therefore a multidimensional process [40]. According to Ebert and Duarte [41], digital transformation is "about adopting disruptive technologies to increase productivity, value creation, and the social welfare." It is a departure from the original understanding of digitization, which is understood as the transformation of analog into digital processes [42]. In this respect, the results support the frequently voiced criticism of existing maturity models in inpatient care with too strong a focus on technical factors [35]. Our results indicate that the digital transformation in outpatient care is also a profound change with effects on people and organizations; practice staff and organizational structures and rules of medical practices should therefore also be explicitly considered. For example, a recent study emphasizes the influence of the personality of GPs on the digital maturity of general practices [43].

Second, a comparison between our results and those of previous studies for outpatient care allows us to derive key points for the measurement dimensions. For example, the availability and use of digital solutions in medical practices appear to be highly correlated with digital maturity. This can be seen both in our identified dimension 1 and in other maturity models or measurement approaches [31,32,44]. This result is hardly surprising, as digital processes are the result of digitalization projects in medical practices. Another focus is on the topic of capabilities, that is, what the practice staff, organization, and technical infrastructure in the medical practice can achieve [27-29,44]. For example, the competencies of physicians and medical assistants were mentioned several times in connection with digital maturity. Although previous maturity models have been criticized for focusing too heavily on technical factors, the technical infrastructure (dimension 4) nevertheless appears to play a major role in digital maturity. This is underpinned by the statements of the experts interviewed on factors such as the maturity of hardware and software or interoperability in medical practices, which can also be found in other maturity measurements [27-31].

Third, our results show similarities and differences between existing digital maturity models in other areas of health care, for example, inpatient care. While organizational, human, and technical capabilities are taken into account, management activities, such as strategy and governance, are largely missing from our results [23,45]. At most, they can be found to some extent in the area of organization in the subcategories of corporate culture and change management but were mentioned very little by the experts compared with other subcategories. In our opinion, this is because the organizational form and size of a medical practice is not comparable with that of larger organizations, such as hospitals, where management plays a greater role. In Germany, the individual practice is still the most common form of establishment in outpatient care at 58%, for which digitalization strategies are irrelevant [46]. Furthermore, digital skills and knowledge of practice management are not an integral part of the curriculum for medical students in Germany, nor are they included in the training regulations for other medical professions [47]. The job description of a practice manager is still relatively new. Last but not least, our results also underline that physicians have a major impact as role models in the digitalization of a medical practice. We conclude from the results that the digital maturity of organizational forms of outpatient care should be assessed differently than in other care areas and that management activities are still given less importance compared with actual patient care.

Fourth, our results support the sensible use of digitalization in medical practices. Although dimension 1 alone emphasizes the availability and use of digital solutions in processes, it is supplemented by dimension 5 (ie, benefits and outcomes). It is not only the frequency of use that is decisive but also the benefits for patient care or for practice staff associated with its use. This result is similar to the assessment framework of Flott et al [44] for evaluating digital maturity of health services, which was also used by Teixeira et al [31] to assess the maturity level in outpatient care. For example, the effects of digital systems on patients, structures, processes, and finances were surveyed. A decisive factor appears to be not only the frequency of use but also the associated benefits for patient care or for practice staff. However, to measure the impact of the use of digital solutions, more suitable evaluation frameworks are required. To date, there are a number of evaluation frameworks, especially for digital solutions, such as NICE (National Institute for Health and Care Excellence) in the United Kingdom, which are already attempts to bring more evidence into the digital health world [48]. With patient-reported outcome measures and patient-reported experience measures, there are more ways to measure the benefits of digital technologies that explicitly take the patient's perspective into account [49]. It remains to be seen whether assessment approaches for the benefits of individual digital solutions are suitable for assessing the impact of digitalization in medical practices. Assessment approaches that deal more comprehensively with the circumstances of the organizational form exist, for example, the workflow composite score for hospitals [50]. With reference to information logistics, hospitals are examined according to the extent to which their IT systems describe the availability of data and information at the point of care. To the best of our knowledge, there are no explicit approaches for assessing digitalization in medical practices. If these are available, in our opinion the question remains open as to whether the effects of digitalization should be considered as causes or consequences of the digital maturity level of medical practices. In this respect, follow-up studies on the correlations between digital maturity and effects need to be carried out.

Fifth, our identified dimension 6 indicates that digital maturity is highly dependent on local circumstances such as the patient situation and other external factors that cannot be directly influenced. This poses a challenge for maturity models, as they allow little flexibility due to predefined maturity levels and a predefined maturity target. However, the statements of the experts interviewed confirm that the requirements for digital maturity in medical practices can vary. They therefore support calls for a differentiated maturity level measurement that takes local circumstances into account [35,51]. Cresswell et al [35], for example, are in favor of maturity models in which the institutions themselves define the goal of digital maturity and align it with local needs. Otherwise, maturity measurements would lead to frustration if prescribed targets are unrealistic and

XSL•FO
RenderX

therefore cannot be achieved [35]. Furthermore, it can be concluded from the results for dimension 6 that dimensions and subcategories of the digital maturity of medical practices are dependent on the maturity of other players in the health care system. An example is the exchange with service providers and institutions, such as public authorities, whose digital maturity for networking is often not fully developed. A joint digitalization of institutions in the health care sector is therefore also required. This is the only way to achieve integrated care in the health care sector [35,44,51]. It is therefore also necessary to assess the digital maturity of other institutions, which in turn must consider networking with institutions such as medical practices. Other examples of the dependence of digital maturity on external framework conditions are digital health literacy and the care needs of patients. The participants interviewed expressed that the pursuit of digital maturity should not lead to patients being excluded from health care. These are important statements for the use of maturity models, as there is a risk of neglecting disadvantaged population groups with increasing digitalization [52-54]. However, high-quality health care also means equitable care that is accessible to everyone [55]. The findings on the various external framework conditions support the diverse effects of digital transformation on society. Digital maturity is therefore not just a concern of individual care areas such as general practice but must be considered in the context of a task for society.

Finally, our brief digression makes it clear that there can also be more significant dimensions within the many dimensions and subcategories of digital maturity. For example, a positive change in 1 dimension could also have an impact on other factors. The results also point to a possible weighting of factors within the construct of digital maturity. They motivate further research to identify the dimensions that have the greatest impact on digital health when promoted by national institutions.

Future Implications

Our study builds on previous research and can in turn be used for a range of other research and applications. First, future quantitative studies should statistically test our identified dimensions. Our study did not aim to investigate statistical representativeness. However, it is necessary to contribute to a validation of the dimensions. In the context of quantitative research, correlations between individual dimensions should also be investigated in more detail. Second, future research should investigate the measurement dimensions we have identified individually. The depth of content for the individual dimensions was limited within the scope of this study. It may therefore make sense to investigate individual dimensions in more detail in studies to gain further insights into the measurement of digital maturity. Third, further studies should be carried out to examine the transferability of the dimensions to outpatient care. Digital maturity should also be analyzed for other care providers such as specialists. Finally, our results can be used in conjunction with quantitative studies for the construction of maturity models. Maturity models can be used by decision makers in the health care sector to promote digitalization in outpatient care.

Limitations

The decision in favor of a qualitative research design enabled the largely unbiased investigation of dimensions detached from any reference models of digital maturity. Nevertheless, our conclusions must be seen in the context of the limitations of our chosen research design. In principle, the results of qualitative research are difficult to verify [56]. Control over the content essentially lies with the participants. The statements made are also subjective and are based on the personal experience and knowledge of the interviewees. Due to participants with dual professional roles, limitations in the allocation to groups of people should also be mentioned. Interpretations are therefore only possible to a limited extent. We tried to meet the challenges by deciding in favor of piloted guided interviews. In this way, we wanted to give the statements of the experts interviewed more structure. However, in favor of more structure, we cannot rule out the possibility that we have somewhat restricted the scope for freer answers as a result. Nevertheless, we were able to inductively derive 2 further dimensions despite the guidelines. Furthermore, we opted for a heterogeneous sample with actors from different areas of the health care sector. In this way, we wanted to counteract the dependence on activity-related interest formation. Our results support the literature, according to which digital transformation is a multidimensional process, so that we feel that our approach of interviewing different professional groups has been confirmed. Limitations can also be seen in our approach to data collection. Although we discussed the coding book with the members of the research team, there may be limitations in the reliability of the coding. In addition, the interview participants were all from Germany. This means that no international maturity models could have been considered by the experts. However, the aim of the study was to examine the level of digital maturity in outpatient care largely independently of existing maturity models. Instead, the understanding of digital maturity was to be explained by the experts on the basis of their knowledge and experience, for example from everyday clinical practice. This study does not claim to have researched validated measurement dimensions. This requires objective results, which are achieved by means of quantitative, statistical analyses. However, mathematical analyses were not the aim of this study and can build on our research findings. Finally, we deliberately chose to focus on general practice in this study, as it plays a key role in outpatient care. An extension to all service providers in outpatient care would have gone beyond the scope of this study. The results of our study should therefore be examined for an extension to all outpatient care.

Conclusions

This study provides empirically analyzed dimensions and subcategories for measuring the level of digital maturity in outpatient care using the example of general practice. It is the first scientific study explicitly dedicated to the digital maturity level of general practice in a qualitative research design. The results indicate that digital maturity is a multidimensional construct and is related to human, organizational, and technical factors. Furthermore, they show that digital maturity influences the entire medical practice and thus the way digitalization is used to shape patient care. Digital maturity is also associated

XSL•FO
RenderX

with external conditions of the medical practice as well as benefits of a digitalized medical practice, which, however, still need to be confirmed. In order to measure the level of digital maturity in outpatient care in medical practices as accurately as possible, maturity models should therefore be multilayered and also take external influencing factors into account.

Acknowledgments

This research received no external funding.

Authors' Contributions

TN contributed to conceptualization, methodology, formal analysis, investigation, data curation, writing (review and editing), and visualization. AM provided a critical review. SM contributed to conceptualization, methodology, validation, and writing (review and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 41 KB - jmir_v26i1e57786_app1.docx]

Multimedia Appendix 2
Translated semistructured interview guide.
[DOC File , 66 KB - jmir_v26i1e57786_app2.doc]

Multimedia Appendix 3
Descriptions of dimensions and subcategories of digital maturity in general practice.
[DOCX File , 20 KB - jmir_v26i1e57786_app3.docx]

References

1. Digitalisation of health systems can significantly improve performance and outcomes. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/newsroom/digitalisation-of-health-systems-can-significantly-improve-performance-and-outcomes.htm [accessed 2024-02-02]
2. Health at a glance. Organisation for Economic Co-operation and Development. 2023. URL: https://www.oecd.org/en/publications/health-at-a-glance-2023_7a7afb35-en/full-report.html [accessed 2024-12-13]
3. Global strategy on digital health. World Health Organization. 2021. URL: https://apps.who.int/iris/handle/10665/344249 [accessed 2022-10-11]
4. Primary health care: report of the International Conference on Primary Health Care Alma-Ata, USSR. World Health Organization. 1978. URL: https://brill.com/display/book/9789004635753/back-4.xml [accessed 2024-12-13]
5. Health at a glance: Europe 2016: state of health in the EU Cycle. Organisation for Economic Co-operation and Development. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2016_9789264265592-en [accessed 2024-12-13]
6. Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices. J Am Board Fam Med 2010;23(5):655-670 [FREE Full text] [doi: 10.3122/jabfm.2010.05.090192] [Medline: 20823361]
7. Paré G, Raymond L, Castonguay A, Grenier Ouimet A, Trudel MC. Assimilation of medical appointment scheduling systems and their impact on the accessibility of primary care: mixed methods study. JMIR Med Inform 2021;9(11):e30485 [FREE Full text] [doi: 10.2196/30485] [Medline: 34783670]
8. Bashshur RL, Howell JD, Krupinski EA, Harms KM, Bashshur N, Doarn CR. The empirical foundations of telemedicine interventions in primary care. Telemed J E Health 2016;22(5):342-375 [FREE Full text] [doi: 10.1089/tmj.2016.0045] [Medline: 27128779]
9. Mettler T. Maturity assessment models: a design science research approach. Int J Soc Syst Sci 2011;3(1/2):81. [doi: 10.1504/ijsss.2011.038934]
10. Fraser P, Moultrie J, Gregory M. The use of maturity models/grids as a tool in assessing product development capability. 2002 Presented at: IEEE International Engineering Management Conference; August 18, 2002; Cambridge, UK p. 244. [doi: 10.1109/iemc.2002.1038431]
11. Becker J, Knackstedt R, Pöppelbuß J. Developing maturity models for IT management. Bus Inf Syst Eng 2009;1(3):213-222. [doi: 10.1007/s12599-009-0044-5]

XSL•FO
RenderX

12. Carvalho JV, Rocha, Abreu A. Maturity models of healthcare information systems and technologies: a literature review. J Med Syst 2016 Jun;40(6):131. [doi: 10.1007/s10916-016-0486-5] [Medline: 27083575]
13. Gomes J, Romão M. Information system maturity models in healthcare. J Med Syst 2018;42(12):235. [doi: 10.1007/s10916-018-1097-0] [Medline: 30327955]
14. Lasrado LA, Vatrapu R, Andersen KN. Maturity models development in IS research: a literature review. Association for Information Systems. 2015. URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1005&context=iris2015 [accessed 2024-01-16]
15. Humphrey WS. Characterizing the software process: a maturity framework. IEEE Softw 1988;5(2):73-79. [doi: 10.1109/52.2014]
16. de Bruin T, Rosemann M. Towards a business process management maturity model. In: Proceedings of the 13th European Conference on Information Systems. Reading, UK: Academic Conferences Limited; 2005.
17. Freeze R, Kulkarni U. Knowledge management capability assessment: validating a knowledge assets measurement instrument. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences. New York, NY: IEEE; 2005:251a.
18. Electronic medical record adoption model (EMRAM). Healthcare Information and Management Systems Society. URL: https://www.himss.org/what-we-do-solutions/digital-health-transformation/maturity-models/electronic-medical-record-adoption-model-emram [accessed 2022-10-23]
19. DigitalRadar Krankenhaus Konsortium. Instrument zur Evaluierung des Reifegrades der Krankenhäuser hinsichtlich der Digitalisierung. St. EconStor. URL: http://hdl.handle.net/10419/246794 [accessed 2023-04-24]
20. Doctor E, Eymann T, Fürstenau D, Gersch M, Hall K, Kauffmann AL, et al. A maturity model for assessing the digitalization of public health agencies. Bus Inf Syst Eng 2023;65(5):539-554. [doi: 10.1007/s12599-023-00813-y]
21. Wendler R. The maturity of maturity model research: a systematic mapping study. Inf Softw Technol 2012;54(12):1317-1339. [doi: 10.1016/j.infsof.2012.07.007]
22. Lahrmann G, Marx F, Mettler T, Winter R, Wortmann F. Inductive Design of Maturity Models: Applying the Rasch Algorithm for Design Science Research. In: Jain H, Sinha AP, Vitharana P, editors. Service-oriented perspectives in design science research. Berlin, Germany: Springer; 2011.
23. Duncan R, Eden R, Woods L, Wong I, Sullivan C. Synthesizing dimensions of digital maturity in hospitals: systematic review. J Med Internet Res 2022;24(3):e32994 [FREE Full text] [doi: 10.2196/32994] [Medline: 35353050]
24. Pearce F, Livingstone A, Gould G, Alexander G. Digital maturity in aged and community care: the current state and resources required. Aged Care Industry Information Technology Council. URL: https://www.aciitc.com.au/wp-content/uploads/2024/01/ACIITC_Digital_Maturity_Report_Final-31012024.pdf [accessed 2024-08-06]
25. Continuity of care maturity model (CCMM). Healthcare Information and Management Systems Society. URL: https://www.himss.org/what-we-do-solutions/digital-health-transformation/maturity-models/continuity-care-maturity-model-ccmm [accessed 2024-08-06]
26. Neunaber T, Meister S. Digital maturity and its measurement of general practitioners: a scoping review. Int J Environ Res Public Health 2023;20(5):4377 [FREE Full text] [doi: 10.3390/ijerph20054377] [Medline: 36901387]
27. Reddy S, Nguyen L, Cooper P, Huggins K, Ugalde A, Peeters A, et al. Digital maturity models for primary health care. Deeble Institute for Health Policy Research. URL: https://ahha.asn.au/wp-content/uploads/2023/11/perspectives_brief_no_26_digital_maturity_models_2.pdf [accessed 2023-12-12]
28. Azar D, Cuman A, Blake T, Proposch A. A digital health maturity assessment for general practice. Analysis & Policy Observatory. 2020. URL: https://apo.org.au/node/317828 [accessed 2022-10-06]
29. Victoria's digital health maturity model. Ernst & Young. URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en_au/noindex/ey-digital-health-maturity-model-whitepaper.pdf [accessed 2022-08-15]
30. Digital maturity assurance. NHS England. URL: https://digital.nhs.uk/services/cloud-centre-of-excellence/ccoe-services/capability-maturity-assessment [accessed 2022-09-25]
31. Teixeira F, Li E, Laranjo L, Collins C, Irving G, Fernandez MJ, et al. Digital maturity and its determinants in general practice: a cross-sectional study in 20 countries. Front Public Health 2022;10:962924 [FREE Full text] [doi: 10.3389/fpubh.2022.962924] [Medline: 36711349]
32. Haverinen J, Keränen N, Tuovinen T, Ruotanen R, Reponen J. National development and regional differences in eHealth maturity in finnish public health care: survey study. JMIR Med Inform 2022;10(8):e35612 [FREE Full text] [doi: 10.2196/35612] [Medline: 35969462]
33. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. Wassermann S. Das qualitative Experteninterview. In: Niederberger M, Wassermann S, editors. Methoden der Experten- und Stakeholdereinbindung in der sozialwissenschaftlichen Forschung. Wiesbaden, Germany: Springer Fachmedien Wiesbaden; 2015:51-67.
35. Cresswell K, Sheikh A, Krasuska M, Heeney C, Franklin BD, Lane W, et al. Reconceptualising the digital maturity of health systems. Lancet Digit Health 2019 Sep;1(5):e200-e201 [FREE Full text] [doi: 10.1016/S2589-7500(19)30083-4] [Medline: 33323267]

XSL•FO
RenderX

36. Kuckartz U, Rädiker S. Auflage. In: Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung: Grundlagentexte Methoden. Weinheim, Germany: Beltz Juventa; 2022.
37. MAXQDA. 2022. URL: https://www.maxqda.com/about [accessed 2024-11-16]
38. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
39. Interdisziplinäres Zentrum für Versorgungsforschung (IZVF). Universität Witten/Herdecke. URL: https://www.uni-wh.de/gesundheit/forschung-an-der-fakultaet-fuer-gesundheit/izvf/ [accessed 2024-01-22]
40. Fernández A, Beratarrechea A, Rojo M, Ridao M, Celi L. Starting the path of digital transformation in health innovation in digital health: conference proceeding. Cienc Innov Salud 2020;e74:68-75 [FREE Full text] [doi: 10.17081/innosa.74] [Medline: 32656302]
41. Ebert C, Duarte CHC. Digital transformation. IEEE Softw 2018;35(4):16-21. [doi: 10.1109/ms.2018.2801537]
42. Definition of digitization - Gartner IT Glossary. Gartner. URL: https://www.gartner.com/en/information-technology/glossary/digitization [accessed 2023-02-25]
43. Weik L, Fehring L, Mortsiefer A, Meister S. Big 5 personality traits and individual- and practice-related characteristics as influencing factors of digital maturity in general practices: quantitative web-based survey study. J Med Internet Res 2024;26:e52085 [FREE Full text] [doi: 10.2196/52085] [Medline: 38252468]
44. Flott K, Callahan R, Darzi A, Mayer E. A patient-centered framework for evaluating digital maturity of health services: a systematic review. J Med Internet Res 2016;18(4):e75 [FREE Full text] [doi: 10.2196/jmir.5047] [Medline: 27080852]
45. Woods L, Eden R, Duncan R, Kodiyattu Z, Macklin S, Sullivan C. Which one? A suggested approach for evaluating digital health maturity models. Front Digit Health 2022;4:1045685 [FREE Full text] [doi: 10.3389/fdgth.2022.1045685] [Medline: 36506845]
46. KBV - Einzelpraxis. Kassenärztliche Bundesvereinigung. URL: https://www.kbv.de/html/14365.php [accessed 2024-08-08]
47. Approbationsordnung für Ärzte. Bundesministerium für Gesundheit. 2002. URL: https://www.gesetze-im-internet.de/_appro_2002/%C3%84ApprO_2002.pdf [accessed 2024-08-13]
48. Unsworth H, Dillon B, Collinson L, Powell H, Salmon M, Oladapo T, et al. The NICE evidence standards framework for digital health and care technologies - developing and maintaining an innovative evidence framework with global impact. Digit Health 2021;7:20552076211018617 [FREE Full text] [doi: 10.1177/20552076211018617] [Medline: 34249371]
49. Knapp A, Harst L, Hager S, Schmitt J, Scheibe M. Use of patient-reported outcome measures and patient-reported experience measures within evaluation studies of telemedicine applications: systematic review. J Med Internet Res 2021;23(11):e30042 [FREE Full text] [doi: 10.2196/30042] [Medline: 34523604]
50. Liebe JD, Hübner U, Straede MC, Thye J. Developing a workflow composite score to measure clinical information logistics. Methods Inf Med 2018;54(05):424-433. [doi: 10.3414/me14-02-0025]
51. Phiri P, Cavalini H, Shetty S, Delanerolle G. Digital Maturity Consulting and Strategizing to Optimise Services: An Overview. JMIR Preprints. Posted online on Nov 16, 2022 2024 [FREE Full text] [doi: 10.2196/preprints.37545]
52. Greenhalgh T, Shaw SE, Alvarez Nishio A, Byng R, Clarke A, Dakin F, et al. Remote care in UK general practice: baseline data on 11 case studies. NIHR Open Res 2022;2:47 [FREE Full text] [doi: 10.3310/nihropenres.13290.2] [Medline: 36814638]
53. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. Inquiry 2022;59:469580221096272 [FREE Full text] [doi: 10.1177/00469580221096272] [Medline: 35471138]
54. Greer B, Robotham D, Simblett S, Curtis H, Griffiths H, Wykes T. Digital exclusion among mental health service users: qualitative investigation. J Med Internet Res 2019;21(1):e11696 [FREE Full text] [doi: 10.2196/11696] [Medline: 30626564]
55. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
56. Taherdoost H. What are different research approaches? Comprehensive review of qualitative, quantitative, and mixed method research, their applications, types, and limitations. J Manag Sci Eng Res 2022 Apr;5(1):53-63. [doi: 10.30564/jmser.v5i1.4538]

Abbreviations

COREQ-32: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioners
NICE: National Institute for Health and Care Excellence

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 26.02.24; peer-reviewed by L Woods, T Schmahl, M Esdar; comments to author 31.07.24; revised version received 20.08.24; accepted 04.11.24; published 19.12.24.

Please cite as:
Neunaber T, Mortsiefer A, Meister S
Dimensions and Subcategories of Digital Maturity in General Practice: Qualitative Study
J Med Internet Res 2024;26:e57786
URL: https://www.jmir.org/2024/1/e57786
doi:10.2196/57786
PMID:39699948

©Timo Neunaber, Achim Mortsiefer, Sven Meister. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Reporting Guidelines for the Early-Phase Clinical Evaluation of Applications Using Extended Reality: RATE-XR Qualitative Study Guideline

Johan H Vlake[1,2], BSc; Denzel L Q Drop[1], MD; Jasper Van Bommel[1], MD, PhD; Giuseppe Riva[3,4], PhD; Brenda K Wiederhold[5], BCN, BCB, MBA, PhD; Pietro Cipresso[3,6], PhD; Albert S Rizzo[7], PhD; Barbara O Rothbaum[8], PhD; Cristina Botella[9,10], PhD; Lotty Hooft[11], PhD; Oscar J Bienvenu[12], MD, PhD; Christian Jung[13,14], MD, PhD; Bart Geerts[15], MSc, MBA, MD, PhD; Evert-Jan Wils[2], MD, PhD; Diederik Gommers[1], MD, PhD; Michel E van Genderen[1], MD, PhD; RATE-XR Expert Group[16]

[1]Department of Intensive Care, Erasmus Medical Center, Rotterdam, Netherlands
[2]Department of Intensive Care, Franciscus Gasthuis & Vlietland, Rotterdam, Netherlands
[3]Applied Technology for Neuro-Psychology Lab, IRCCS Istituto Auxologico Italiano, Milan, Italy
[4]Department of Psychology, Catholic University of the Sacred Heart, Milan, Italy
[5]Virtual Reality Medical Center, San Diego, CA, United States
[6]Department of Psychology, University of Turin, Turin, Italy
[7]Medical Virtual Reality Lab, University of Southern California Institute for Creative Technologies, Los Angeles, CA, United States
[8]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, United States
[9]Department of Basic Psychology, Clinic, and Psychobiology, University Jaume I, Castellón, Spain
[10]CIBER de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Instituto Salud Carlos III, Madrid, Spain
[11]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands
[12]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[13]Department of Cardiology, Pulmonology, and Vascular Medicine, Medical Faculty, University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany
[14]Cardiovascular Research Institute Düsseldorf (CARID), Medical Faculty, University Hospital of Düsseldorf, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
[15]R&D BV, Healthplus.ai, Amsterdam, Netherlands
[16]See Acknowledgments

Corresponding Author:
Michel E van Genderen, MD, PhD
Department of Intensive Care
Erasmus Medical Center
Doctor Molewaterplein
Rotterdam, 3015 GD
Netherlands
Phone: 31 107040704
Email: m.vangenderen@erasmusmc.nl

Abstract

Background: Extended reality (XR), encompassing technologies such as virtual reality, augmented reality, and mixed reality, has rapidly gained prominence in health care. However, existing XR research often lacks rigor, proper controls, and standardization.

Objective: To address this and to enhance the transparency and quality of reporting in early-phase clinical evaluations of XR applications, we present the "Reporting for the early-phase clinical evaluation of applications using extended reality" (RATE-XR) guideline.

Methods: We conducted a 2-round modified Delphi process involving experts from diverse stakeholder categories, and the RATE-XR is therefore the result of a consensus-based, multistakeholder effort.

XSL•FO
RenderX

Results: The guideline comprises 17 XR-specific (composed of 18 subitems) and 14 generic reporting items, each with a complementary Explanation & Elaboration section.

Conclusions: The items encompass critical aspects of XR research, from clinical utility and safety to human factors and ethics. By offering a comprehensive checklist for reporting, the RATE-XR guideline facilitates robust assessment and replication of early-stage clinical XR studies. It underscores the need for transparency, patient-centeredness, and balanced evaluation of the applications of XR in health care. By providing an actionable checklist of minimal reporting items, this guideline will facilitate the responsible development and integration of XR technologies into health care and related fields.

(J Med Internet Res 2024;26:e56790) doi:10.2196/56790

KEYWORDS

extended reality; XR; virtual reality; augmented reality; mixed reality; reporting guideline; Delphi process; consensus; computer-generated simulation; simulation; virtual world; simulation experience; clinical evaluation

Introduction

Extended reality (XR) encompasses various forms of computer-generated reality, including augmented reality (AR), mixed reality (MR), and virtual reality (VR). XR, mainly in the form of VR, has rapidly emerged in health care, particularly in fields such as mental health, intensive care medicine, surgery, pain management, and rehabilitation [1-4]. Much like other transformative technologies as artificial intelligence (AI) algorithms, the field of XR has witnessed an exponential surge in research and applications: from 1992 to 2005, merely up to 100 publications were recorded yearly, but this number has steadily increased, with over 1000 publications annually since 2018 [2]. Notably, the US Food and Drug Administration has been approving a growing number of XR-based devices, underscoring its escalating clinical significance [5].

Despite this expanding landscape of XR research in health care, most studies primarily focus on treatment effects, tend to be small and heterogeneous, and often lack proper control conditions [6]. Consequently, comparing XR studies is challenging, as scientific rigor is often lacking and they pose several unique implementation and technological challenges in health care [7-9]. To overcome these challenges and to promote optimal reporting of XR-based interventions' clinical utility, a more structured approach is essential. This approach should encompass technological, methodological, and safety aspects to support a more objective understanding of the validity and generalizability of findings [10]. The challenges of early-stage clinical evaluation of applications using XR (Textbox 1) share similarities to those of other innovative technologies and interventions such as developing and implementing surgical innovations or AI models [11-14].

Early-stage clinical evaluation of XR systems plays a pivotal role in bridging the gap between preclinical technological development and large-scale effectiveness trials. Existing guidelines such as the VR core model stage 2/3 recommendations, SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) and CONSORT (Consolidated Standards of Reporting Trials) statements, including their AI extensions, and the IDEAL guidelines offer valuable insights into the design and reporting of clinical trials (Figure 1) [12,13,15,16]. However, they generally focus on later stages of clinical research or do not adequately address the specific challenges of developing and evaluating XR technologies in early clinical settings. These include the rapid evolution of XR hardware and software and the specific safety, usability, and ethical considerations that arise in these contexts. To address these gaps, the reporting for the early-phase clinical evaluation of applications using extended reality (RATE-XR) guideline specifically tailors its recommendations to support transparent reporting and effective evaluation of XR applications from the developmental phase through to early clinical trials, ensuring that these innovations can be safely and effectively integrated into health care practice.

Textbox 1. The challenges of early-phase clinical evaluation of applications using extended reality (XR). the clinical evaluation of applications using XR presents several challenges, all of which will likely be encountered at early stages. This textbox represents several of these challenges.

- Allow for a continuous changing nature of the software applications using XR and its hardware (due to early prototyping and version updates)
- Account for technical errors negatively impacting the reliability and consistency of clinical evaluations
- Evaluate the generalizability of findings across sites and populations
- Deal with ethical considerations in this early stage of research
- Deal with a variety of clinical trial endpoints for applications due to the wide range of intended uses
- Account for user variability and consequently the arising bias as users may not be trained adequately or may not be familiar with XR technology
- Incorporate XR applications into the standard workflow
- Absence of established methodologies and frameworks
- Having sufficient image quality of XR devices for users, due to their diversity and constantly evolving technological characteristics
- Create a usability profile applicable to various working environments

XSL•FO
RenderX

Figure 1. Comparison of development pathways for drug therapies, surgical innovation, artificial intelligence, and extended reality in health care. The colored lines represent reporting guidelines, some of which are study design–specific (SPIRIT or CONSORT and SPIRIT or CONSORT-AI); others are stage-specific (IDEAL and RATE-XR). Depending on the context, more than one study design can be appropriate for each stage. CONSORT: Consolidated Standards of Reporting Trials; RATE-XR: reporting for the early-phase clinical evaluation of applications using extended reality; SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials.

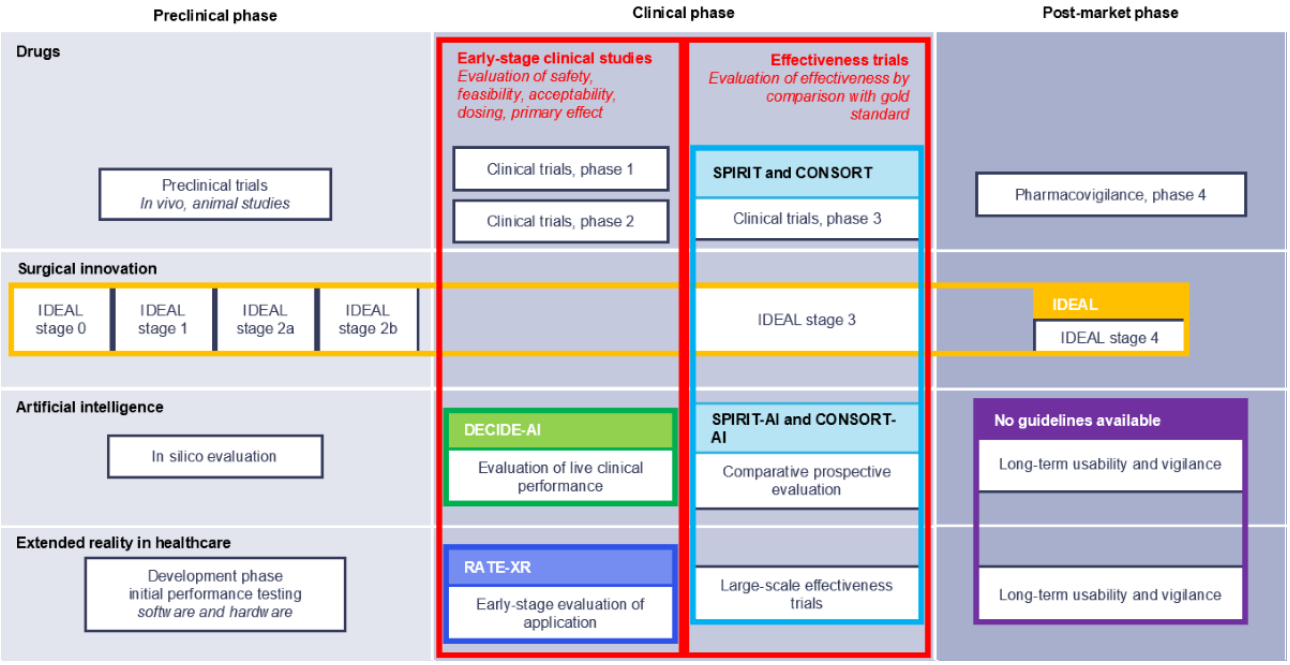

Early-stage clinical evaluation of XR interventions must prioritize clinical utility, safety, and human factors challenges in real-life clinical settings. Factors such as cybersickness, customizability and duration of the XR content, treatment frequency, and immersiveness contribute to safety and feasibility considerations and must be addressed transparently. A one-size-fits-all model is often not feasible, and neglecting safety profiles and rushing into large-scale trials can jeopardize patient well-being, which is ethically unacceptable. In terms of ethics, patient-centeredness and commercial interests must be addressed in the early stages of XR development. Currently, studies too often prioritize assessing commercial products but fail to address essential clinical research elements or tailoring content for medical applications, while understanding the interaction between XR technology and human factors is essential [17-19]. The clinical context should be the starting point in the development of medical XR systems, involving patients and health care providers as primary stakeholders early in the process to design systems that optimally address their needs, beyond placing an initial focus on commercial product evaluation. Moreover, variations in hardware, software, and content selection are complex to assess and often underreported [20].

To address these challenges and with the aim of improving the consistency, safety, knowledge generation, and applicability of XR research in the health care domain, we undertook a robust 2-phase modified Delphi process. This collaborative effort engaged strong and diverse stakeholder engagement and resulted in the development of the RATE-XR guideline. Here, we present the development process, key recommendations, and their implications for the XR health care field.

Methods

RATE-XR Guideline Development

The RATE-XR guideline was developed through an international expert consensus process adhering to the EQUATOR Network's recommendations for guideline development [21].

Establishment of the Steering Committee

To guide the development of the RATE-XR guideline, a Steering Committee was assembled, detailed in Table S1 in Multimedia Appendix 1. The committee was selected by the project initiators—JHV, JvB, and MEvG—to ensure a diverse representation of expertise within the XR and research domains. Members of our study group, including CJ, DLQD, DG, EJW, and OJB, were included for their direct contributions to this project. Additionally, we engaged 5 of the most influential authors in XR research as cited in a recent JMIR article—GR, BKW, PC, ASR, and CB—to incorporate a broad range of perspectives and expertise [2]. We also invited BOR, a pioneer in clinical VR, and BJB, known for his work in VR trial design and implementation. Finally, to incorporate expertise in methodology and guideline development, LH and BG, both experienced in developing previous reporting guidelines, were included [13]. We conducted a modified Delphi process based on previous guideline development consisting of 2 rounds of feedback from participating experts followed by virtual consensus meetings and qualitative evaluation by an independent evaluation committee [13,22,23].

Ethical Considerations

The project was approved by the Medical Ethics Committee of the Erasmus Medical Centre, Rotterdam (approval 2022-0623) and registered with the EQUATOR (Enhancing the Quality and

XSL•FO
RenderX

Transparency of Health Research) network. Informed consent was obtained from all members of the Steering Committee, all participants of the Delphi rounds, and all members of the evaluation committee.

Generation of the Initial Item List

An initial list of 61 candidate items (with subitems) was composed by 2 authors (JV and MEvG) and was based on (1) scientific reports on trials examining XR-based studies in health care [24-27], (2) recently published innovative technology guidelines [13,28], (3) methodological and evaluative challenges concerning the application of XR in health care [14,16], (4) a Cochrane Systematic Review on the clinical use of XR [29], and (5) institutional documents [30-32]. Hereafter, the candidate item list was commented on by the Steering Group members (Steering Group Round).

Expert Recruitment

Experts were recruited using five distinct approaches: (1) invitations to experts that were endorsed by the Steering Group members, (2) invitations to authors of publications identified through the preliminary literature search, (3) a call for contributions published within a medical journal [33], (4) consideration of professionals proactively reaching out to the Steering Group, and (5) invitations to experts recommended by the Delphi participants (snowballing). Prior to the initiation of recruitment, 17 target stakeholder groups were defined: clinicians, engineers or computer scientists, methodologists, statisticians, implementation specialists, entrepreneurs, epidemiologists, journal editors, allied health professionals, policy makers or official institutional staff, administrators or hospital management, researchers, ethicists, private sector representatives, patient representatives, funders, and psychologists or psychiatrists.

The Delphi Process

The Delphi process consisted of 2 rounds, and the Delphi surveys were designed and distributed using the Castor Electronic Data Capture web application (Castor EDC). The first round encompassed 2 parts. In the first part, participants answered 6 open-ended inquiries that address facets considered essential to be reported on during early-phase clinical evaluation. In the second part, Delphi participants were tasked with rating, from 1 to 9, the significance of items in the initial list. Ratings of 1 to 3 indicated insignificance, 4 to 6 denoted importance without being pivotal, and 7 to 9 implied that items were both important and critical. In addition to rating the items, participants were prompted to offer commentary on items and propose new additions. Thematic analysis of the open-ended questions was independently conducted by 2 Steering Group members (JV and MEvG), with any disagreements being resolved through consensus. Identified themes were used to determine whether any important themes were missing in the item list, along with newly proposed items to complement the item list. Hereafter, a summary score, including the median, 25th percentile, 75th percentile, mean, SD, proportion of participants scoring the item above 7 or below 3, and stakeholder groups with a median of ≤2 or ≥2 points from the overall median were calculated for each item. Prespecified inclusion cutoffs were determined as an item scoring a mean ≥7 and exclusion as an item scoring a mean ≤3. Based on these results, a revised item list for the second Delphi round was generated.

In the second Delphi round, participants were presented with the outcomes of the first round, along with the revised item list. Participants were tasked with reevaluating all items in a manner akin to the first Delphi round and were invited to comment on content and wording. Both Delphi round surveys and outcomes are accessible through the Open Science Framework (OSF) [34]. All analyses were performed using R for Statistics (R Foundation for Statistical Computing).

Consensus Meeting

Virtual consensus meetings were held on 3 separate occasions between June 12 and 15, 2023, with the aim of finalizing content and refining the phrasing of items within the RATE-XR reporting guideline. To ensure a balanced representation of key stakeholders throughout the XR field and geographic diversity, we engaged 18 experts with diverse expertise and backgrounds (Tables S2-S4 in Multimedia Appendix 1). Throughout the consensus meetings, all items from the second Delphi round were subject to discussion and anonymous voting, facilitated by the Mentimeter platform [35]. The voting process was overseen by a chairman and observed by a designated observer. For an item to be ultimately included in the definitive guideline, a predefined threshold of 80% agreement among Consensus Group members was necessary, excluding abstentions and blank votes.

Qualitative Evaluation

After finalizing both the guideline and the Explanation & Elaboration note, a qualitative evaluation was conducted by a panel of 14 experts (Note 1 in Multimedia Appendix 1) possessing significant experience in implementing or peer-reviewing literature relevant to applications using XR. None of the experts involved in the qualitative evaluation were affiliated with the Consensus Group. Their input focused on evaluating the clarity and usability of each XR-specific item using a custom form, which is available on the OSF platform [34]. In total, 3 reviewers (JV, DD, and MEvG) independently reviewed the provided comments to assess the necessity for revisions in the wording of items or their corresponding Explanation & Elaboration sections. The review process was structured to ensure comprehensive coverage and unbiased analysis of the feedback. Disagreements between reviewers were resolved through consensus, ensuring a balanced interpretation of the qualitative data. This methodological rigor enhances the reliability of the modifications made to the RATE-XR guidelines based on stakeholder feedback. To enhance comprehension of key concepts within the guideline, a glossary of terms (Table 1) was composed. All Consensus Group members approved the modifications, the final guideline, and the complementary Explanation & Elaboration note.

XSL•FO
RenderX

Table 1. Glossary of terms[a].

Terms	Explanations
Application	The software, program, intervention, or modality using an XR device or hardware.
Artificial intelligence algorithm	”Science of developing computer systems which can perform tasks normally requiring human intelligence” based on a mathematical model responsible for learning from data and producing an output [36].
Augmented reality	A technology that overlays digital information onto the real-world environment, viewed through an augmented reality headset or glasses in order to enhance the user’s perception of reality. Augmented reality is part of the extended reality (XR) technologies.
Bias	Systematic difference in treatment of certain objects, people, or groups in comparison to others [37].
Clinical utility	The practical value and usefulness of the application using XR.
Commercial name of application	Trademarked or branded name under which the specific software application or hardware device is sold.
Commercial product	An item originally manufactured for sale, lease, or license to the general public.
Cybersickness	A form of motion sickness or discomfort experienced by individuals while using XR devices.
Early-phase studies	Studies in the initial stages of investigation where applications using XR devices are tested and evaluated, focusing on safety, dosage, feasibility, and potential efficacy of the intervention involving a relatively small number of participants.
XR	An umbrella term to encompass the spectrum of immersive technologies consisting of virtual reality, augmented reality, and mixed reality.
Flow diagram	A visual representation using symbols and arrows to illustrate the sequential steps, processes, or decisions within the workflow or procedure.
Hardware	The physical components and equipment that make up the XR system.
Human factors	Also called ergonomics. “The scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance” (International Ergonomics Association).
Immersiveness	The degree to which the experience captivates and engages the user, involving a deep sense of presence and absorption within a simulated environment.
Immersive virtual reality	A computer-generated, 3D artificial environment using a head-mounted display and therefore completely surrounding the user’s senses. This way the user is brought from the real world into the artificial, virtual world.
Mixed reality	A technology combining elements of both virtual reality and augmented reality. It integrates digital content and virtual objects into the real-world environment, allowing users to interact with and manipulate these virtual elements as if they were part of their physical surroundings.
Performance	How well the application, device, system, or technology functions and executes its intended task or functionalities.
Preclinical	Pertaining to the phase of research prior to clinical trials targeting actual patients.
Prespecified outcomes	Specific defined results, goals, or expectations that are determined and established in advance, prior to performing the study.
Real clinical setting	Pertaining to the observation and treatment of actual patients, instead of preclinical users or simulated scenarios.
Software	The applications, programs, and digital content are specifically designed to interact with the XR system.
Usability	“Extend to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” [38].

[a]The definitions provided apply to the specific context of RATE-XR and the use of the terms in the guideline. They are not necessarily generally accepted definitions and may not always be entirely suitable for other research domains.

Results

Initial Item List

Based on the yielded 97 comments and 22 proposals for new items, 18 items were appended and 29 were subjected to reorganization (through merging or splitting, resulting in 16 items). Additionally, wording was amended, and items were categorized into XR-specific and general items. The final initial item list resulted in 71 Delphi items, subdivided into 41 XR-specific and 30 general reporting items, 6 categories, and 22 subcategories, and was approved on by all Steering Group members (see digital file in the OSF platform) [34].

Delphi Rounds

A total of 124 individuals expressed their interest and completed the participation form for the first Delphi round, of whom 22 were unqualified due to lack of XR-related experience. Among the 102 experts who received the first Delphi questionnaire, 93

XSL•FO
RenderX

(91%) completed the questionnaire. The participants included 13 Steering Group members, 38 identified from Steering Group recommendations, 13 from proactive contacts or correspondence, and 29 through snowballing. In total, 112 experts were invited to participate in the second Delphi round, of which 96 (86%) responded. In total, 82 of these experts also participated in the first Delphi round (continuity rate: 88%). Collectively, the participating experts represented 14 countries, and all stakeholders were represented (Supplementary Note 1 and Tables S5-S8 in Multimedia Appendix 1).

The first Delphi round yielded over 17,300 words of unstructured text to the open-ended inquiries, along with 6603 item scores, 256 item comments, and 97 newly proposed items. Thematic analysis identified 146 themes, of which 88 were covered in existing items, 22 were integrated into or added to the provisory Explanation & Elaboration note, 28 were used to amend existing items, 2 were selected as new items, and 6 were dropped as they were determined to be outside of the reporting guideline scope. Eventually, 5 items remained unchanged, 27 items were amended or rephrased, 36 items were merged or split into 14 items, 3 items were dropped, and 5 items were added (Figures S1 and S2 in Multimedia Appendix 1). The 3 items that were dropped were related to production costs of the XR module and were dropped due to low consensus in the scoring exercise and congruent comments that these items were out of scope. The revised item list eventually comprised 51 Delphi items in 45 reporting items, subdivided into 22 XR-specific and 23 general reporting items. The second Delphi round yielded 4896 item scores and 372 comments.

Consensus Meeting

In total, 32 items received endorsement for integration into the RATE-XR guideline during the consensus meetings—17 items specific to XR and 14 encompassing general reporting. A summary of the Consensus Meetings votes is presented in Table S9 in Multimedia Appendix 1.

Qualitative Evaluation

A total of 95 comments were provided. Subsequently, wording of 7 items was refined in the checklist, and of 9 items, there were modifications in their corresponding Explanation & Elaboration section in Multimedia Appendix 1. The evolutionary trajectory of the item list is presented in Figures S1 and S2 in Multimedia Appendix 1.

Final Reporting Item Checklist

Table 2 presents the RATE-XR checklist and consists of 17 XR-specific reporting items (composed of 18 subitems) and 14 generic reporting items, selected by the Consensus Group.

XSL•FO
RenderX

Table 2. RATE-XR (reporting for the early-phase clinical evaluation of applications using extended reality) checklist.

Theme	Item number[a]	Recommendation
Title and abstract		
Title	1	• Identify the study as an early clinical evaluation, or a similar term, of an application using XR[b], or a more specific term, in the title, including its intended aim.

Abstract	I	• Provide a (structured) summary of the study.
		• Consider including the following:
		  • A concise description of the clinical problem or knowledge gap and the rationale for using an application using XR
		  • A concise description of the study methods, including a short description of the application including its name, study population, study setting, main outcomes, and assessment methods.
		  • A concise description of the results, including safety and harm outcomes
		  • A short conclusion
		  • If applicable, details about the registration of the study in a publicly available database.

Introduction		
Clinical problem and existing evidence	2	• Introduce the clinical problem for which the application using XR was used, including its relevance and a description of (the efficacy of) evidence-based or commonly used interventions or the treatment as usual, which is intended to be replaced by the application using XR.

Introduction of the application	3	• Introduce the application using XR, including the following:
		  • Hypotheses for the potential effect; how the application is expected to contribute to the clinical problem.
		  • If available, a concise description of, or a reference to, previous research on the same (or a similar) application.

Objectives	II	• Specify the study objectives or hypotheses.
Methods and analysis		
Trial design and reporting	III	• Provide a reference to ethical approval and, if available, to any (published) study protocol and registration of the study in a publicly available repository.
Trial design and reporting	IV	• Describe, and mention the rationale for, the study design. For clarification, it is recommended to use a flow diagram.
Participants and setting	4	• Describe the setting and locations, including country, where data were collected and processed, and where the application using XR was applied and evaluated.
Participants and setting	5a	• Describe how participants were selected and recruited and provide eligibility criteria.
Participants nad setting	5b	• Describe who will be applying the application and whether they were trained.
Intervention and procedures	6	• Provide a description of the application, including its content, hardware, protocol, and set-up, or provide a reference to previous publications where this information is described. Consider supplementing the description with an image, figure, or film.
Intervention and procedures	7	• Describe, or provide a reference to, the development process of the application.
Intervention and procedures	8	• Describe the participant timeline in sufficient detail to allow replication, including all procedures, co-interventions (if applicable), and (follow-up) assessments.
Intervention and procedures	V	• Describe and give a rationale for the control conditions or provide a rationale for not using one.
Outcomes	VI	• Describe all prespecified primary and secondary outcomes, including how and when assessed.
Outcomes	9	• Describe how safety and harm outcomes were assessed. Describe which, and how, other XR-specific outcomes were assessed, such as performance, usability, presence, perspectives, and acceptability.
Sample size	VII	• Provide a justification for the sample size.

XSL•FO
RenderX

Theme	Item number[a]	Recommendation
Analysis	VIII	• Provide a detailed description of how primary and secondary outcomes were analyzed, including any prespecified comparisons or stratifications.
Protocol alterations	IX	• Describe changes to the methods or protocol, including procedures, study outcomes, eligibility criteria, and analysis plan, after study commencement, with reasons, and, if applicable, report whether the study registration was updated.
Results		
Participant flow and recruitment	X	• Describe the time frame of recruitment and follow-up and the participant flow, including the number of patients screened and included, receiving the intervention, and being included in each analysis. Report if, and why, the study was prematurely terminated. The use of a flow diagram is highly recommended.
Baseline data	XI	• Describe, or add a table depicting, baseline and treatment-related characteristics. If applicable, describe and specify any concurrent measures.
Main results	XII	• Report on all prespecified outcomes that are available. Consider using tables, figures, or graphs to illustrate results.
XR and human factors	10	• Include information about the usage of the application, such as duration, frequency, number of sessions, error rates, and number of sessions requiring interruption or discontinuation, including reasons.
XR and human factors	11	• If assessed, report on XR-specific outcomes, such as performance, usability, presence, perspectives, and acceptability.
Safety and harms	12	• Report on safety and harms, including unintended effects, both during and after using the application.
Discussion and conclusion		
Generalizability and impact	13	• Discuss (potential) impact of study findings and generalizability, including barriers for the use and implementation of the application.
Safety and harms	14	• Discuss safety and instances of harm, including their possible effects on study findings, implications for future use of the applications, and whether they can be prevented or mitigated.
Ethics	15	• Describe ethical considerations, including benefits and risks, for the current and future use of the application.
Strengths and limitations	XIII	• Discuss study strengths and limitations, including sources of potential bias.
Conclusion	16	• Provide a conclusion that accurately interprets study findings, including future perspectives.
Statements		
Funding and conflicts of interest	XIV	• Disclose any potential conflict of interest, real or apparent, including the funding sources and their roles in the design, conduct, analysis, and report of the study, potential roles of commercial companies, and personal conflicts of interest for each author.
Application	17	• Indicate whether the application is a commercial product, it is publicly available, it can be accessed, it complies with the medical device regulations, and whether the application was approved for its intended use by a formal regulatory body or if the study is part of the clinical evaluation for future certification.

[a]AI-specific items are numbered in Arabic numerals; generic items are numbered in Roman numerals.

[b]XR: extended reality.

XSL•FO
RenderX

Discussion

Reporting Item Checklist

The RATE-XR guideline serves as a checklist for reporting studies that focus on the early-phase evaluation of clinical applications using immersive technologies, regardless of the chosen study design (Figure 1). Depending on the specific study design selected, authors may also find it contributing to complement their reporting with guidelines tailored to that study type, such as the CONSORT guideline for randomized trials or the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for observational studies [39,40]. This provides a helpful source to support researchers and reviewers assess manuscript compliance with the guideline. For a more elaborate understanding of each item's relevance and recommendations on how to report, we added an in-depth Explanation & Elaboration section for every item in Note 2 in Multimedia Appendix 1.

It is important to recognize that reporting guidelines, including the RATE-XR guideline, offer a framework of reporting recommendations, yet they may not comprehensively cover every aspect relevant to a particular study or guide the conduct of research. While not exhaustive, familiarity with RATE-XR can help researchers in the design and execution of studies within the guideline's scope. Given the challenge of reporting all required information into a single manuscript, authors may need to refer to other documents, such as study protocols, previous publications, and supplementary materials from digital repositories.

Lessons Learned

The RATE-XR guideline represents the outcome of an extensive international consensus effort from a diverse and representative group of experts with a broad range of professional expertise and backgrounds. The high response rate and the remarkable level of engagement from stakeholders, along with the fact that 5 of the 7 most productive authors of XR research in psychology or medicine were willing to be included in the Steering Group, underscore the necessity for comprehensive reporting guidance in the early-phase clinical evaluation of XR applications [2]. This growing recognition highlights the increasing importance attributed to thorough clinical evaluation as a cornerstone for the effective implementation of XR technologies. The development of this reporting guideline was shaped by the Steering Group's belief that the use of XR-related health care applications will continue to expand, with an increasing requirement for high-quality, comprehensive, consistent, and generalizable reporting of early-phase evaluations.

The RATE-XR guideline is a pioneering endeavor, being the world's first reporting guideline specifically tailored to medical or psychology research involving XR-related applications. We focused on the early-phase clinical evaluation of XR-related applications, as existing guidelines insufficiently represented the essential reporting items for this type of research. Studies on late-phase evaluation often have the option to adhere to more general reporting guidelines, such as CONSORT for randomized controlled trials. However, we acknowledge that beyond this initial guideline, there is merit in further developing XR-specific extensions for existing guidelines. Thus, our efforts mark the crucial first step in harmonizing XR research in the health care sector. Beyond its primary aim, the RATE-XR guideline may also serve as a compass for authors, guiding them in study design, protocol development, and the registration of early-phase studies involving XR applications.

To attract experts with diverse backgrounds across the health care field, we published a call for contributions in the form of a correspondence paper. In this publication, we mainly focused on the terminology “virtual reality,” although we already had the more inclusive terminology “XR” in scope. During extensive deliberations during the Delphi rounds and after discussion within the Steering Group and Consensus Group meetings, we decided to use the more inclusive term XR instead of VR. This change allows the guideline to cover a broader range of applications, including augmented and MR, as well as any future immersive technologies. We concluded that XR terminology better represents all these applications, and all eventually will have to undertake equal and similar steps during early-stage evaluation and research. The decision to change the project's name from RATE-VR to RATE-XR reflects the guideline's adaptability and commitment to serving a diverse spectrum of applications beyond just focusing on VR alone.

The Delphi process, while invaluable for achieving consensus on guideline development, presented challenges such as maintaining a high follow-up rate among participants and reconciling diverse expert opinions. Our approach to addressing these challenges involved adaptive communication strategies and fostering an environment conducive to open dialogue, which were instrumental in enhancing participant engagement and consensus quality. These experiences provide key learnings that could inform similar guideline development efforts in emerging research fields.

Throughout the guideline's development, several topics generated more dynamic discussions than others, leading to a number of critical decisions. First, a discussion concerned the depth of information required about those administering XR applications. Some participants advocated for gathering baseline researcher or provider characteristics and offering detailed accounts of their training and qualifications. Given the novelty of the technology, the consensus was that mentioning that application providers had sufficient training was sufficient with no specific detailed requirement or baseline demographics unless deemed pivotal for study outcomes. In concordance, a recent review concluded that XR providers need training to improve adoption, which can be achieved using a variety of training programs, strategies, or educational resources, and that there is no minimum amount or golden standard of XR training [41].

Second, the deliberation on XR-specific outcomes, including factors such as acceptability, usability, user experiences, immersiveness, and cybersickness or other negative side-effects, sparked lengthy discussions among participants. While the importance of reporting cybersickness and safety-related aspects in early-phase evaluations was unanimously acknowledged for building trust within the research community, participants recognized the challenges of encompassing all XR-specific outcomes comprehensively in every article. Consequently, it

XSL•FO
RenderX

was agreed that cybersickness and safety should be mandatory reporting items, while other XR-specific outcomes should be considered optional, acknowledging the difficulty in fully addressing all these outcomes in every manuscript. This approach allows researchers the flexibility to focus on the most relevant outcomes for their specific studies.

Third, intensive debates centered around the level of detail necessary when describing XR application hardware, software, and development processes in the RATE-XR guideline. While some argued for comprehensive and mandatory disclosure, others championed flexibility. Ultimately, agreement was reached to consolidate these aspects into a single section in the guideline, with the Explanation & Elaboration note providing guidance on what to include and how to effectively report them.

Fourth, the need for items on randomization in the current guideline was discussed. Participants felt that most early-phase evaluations are seldom randomized and acknowledged that if a study has a randomized design, adhering to established guidelines such as CONSORT would be more appropriate than duplicating information in the RATE-XR guideline. It was agreed that the guideline should not delve into specific items related to randomization, as it would be more beneficial for researchers to consult CONSORT and ensure consistency in reporting across various study designs. A similar consideration and strategy was recently adopted in the DECIDE-AI guideline to prevent a too exhaustive reporting checklist in the early stage of innovative technology development [13].

Fifth, the topic of blinding and the utilization of control groups triggered significant debate during the consensus process. Participants recognized the inherent challenges of blinding in XR studies due to the immersive nature of the technology. While some argued that specific items on blinding-related items should be included, others emphasized the challenges in doing so effectively. Furthermore, in alignment with the guideline's aim to avoid duplicating items covered by study design–specific guidelines, the decision was made to omit the item on blinding. Nevertheless, control groups were deemed valuable for comparison purposes, resulting in the inclusion of items that address the presence or absence of control groups and outline their characteristics as essential components within the guideline.

Sixth, discussions occurred regarding the necessity of justifying the sample sizes in the early-phase clinical assessment of XR applications. Participants held differing perspectives as to whether a formal sample size calculation should be mandated for all research types within the RATE-XR guideline. Ultimately, it was determined that while it is essential to provide some form of justification for the chosen sample size, not all research designs necessitate a formal sample size calculation. This decision recognizes the diversity of study designs and acknowledges that certain types of early-phase evaluations may have inherent limitations that preclude the use of traditional sample size calculations. Nevertheless, authors are strongly encouraged to provide a rationale and justification for their chosen sample size. This proactive step enhances transparency within the research process, enabling readers to assess the study's reliability of its findings.

Lastly, discussions concerning the inclusion of standardized statements such as data protection, ethics approval, and data and code availability were extensive. While their significance was acknowledged, the consensus was that most journals already require authors to address these elements in their manuscripts. Therefore, specific items related to these aspects were omitted from the RATE-XR guideline, as they are comprehensively covered by existing publication requirements and publication guidelines.

In conclusion, the RATE-XR guideline is a pioneering effort in facilitating comprehensive and standardized reporting in the early-phase clinical evaluation of XR applications. Its development involved extensive expert-informed debates and critical decisions that aim to ensure its purpose as a valuable resource for researchers while maintaining its adaptability to a dynamic and evolving field. We anticipate that this guideline will foster transparency, enhance the quality of reporting, and ultimately contribute to the responsible and effective integration of XR technologies in health care and related fields. Furthermore, we encourage the development of XR-specific extensions for existing guidelines to further advance the harmonization of XR research practices.

Acknowledgments

There was no funding in any form for this study. We have not received payments from any agency or pharmaceutical company to establish this manuscript.

The members of RATE-XR Expert Group are Aisling Flynn, Alice Chirico, Amir H Sadeghi, Andrea Gaggioli, Andrea S Won, Annelotte P van Haaps, Azucena Garcia-Palacios, Beate Dejaco, Bram Dierckx, Carsten Finke, Catheleine van Driel, Chris NW Geraets, Christopher Eccleston, Christopher R Madan, Clarine van Oel, Constantinos Panayi, Darcy Ummels, Dave Thomas, Debra L Safer, Elsbeth Zandee, Emil R Høeg, Felix J Hüttner, Floris van der Breggen, Geert-Jan van Geffen, Gido A Hakvoort, Giulia Corno, Hadi Hosseini, Hafize Demirci, Hanne Konradsen, Henry Xiang, Ivan Phelan, Jan D Rölfing, Jan Gödeke, Jennifer N Stinson, Jeremy Bailenson, Jeroen Legerstee, Jiabin Shen, Joost Huiskens, Jordan Tsigarides, Jose F Costa, Juana M Bretón-López, Karamveer Narang, Karin Valkenet, Kim D Bullock, Kimmy Rosielle, Krista Hoek, Line K Pedersen, Loes Bulle-Smid, Lonneke M Staals, Kemal Kuscu, Mareine GT Koornneef, Margaux Sageot, Margot D Paul, Maria Bajwa, Maria Matsangidou, Marie-Madlen Jeitziner, Mariju F Baluyot, Marlies P Schijven, Martine J van Bennekom, Matthew Browning, Melissa L Morris, Merel A Oskam, Merlijn Smits, Michael Gaebler, Mienke Rijsdijk, Njin-Zu Chen, Omar Aly, Pablo Campo-Prieto, Panagiotis Kourtesis, Philipp Kellmeyer, Les Posen, Rachel Reeves, Rami A Ahmed, Raphael R Bruno, Rob JEM

XSL•FO
RenderX

Smeets, Robbert Brouwer, Robert J Fine, Robert M Lundin, Roderick F van Beek, Rosa M Baños, Roselinde MCA Pot-Kolder, Sarah E MacPherson, Silvia Serino, Sophia Rekers, Srinivasan S Pillay, Stephan Krohn, Stéphane Bouchard, Sulayman el Mathari, Susan Persky, Syl Slatman, Synthia Guimond, Thomas J Caruso, Thomas Sauter, Thomas Wolbers, Tjitske D Groenveld, Tobias Loetscher, Todd Chang, Tonnie Staring,Vishnunarayan G Prabhu, Wim Veling, and Winnie WS Mak.

Data Availability

The datasets and necessary documents generated and analyzed during this study are available in Multimedia Appendix 1 and the digital Open Science Framework RATE-XR platform [34]. Additionally, all study materials are available from the corresponding author upon reasonable request.

Authors' Contributions

MEvG initiated the study. JHV, JvB, and MEvG designed the study. Members of the RATE-XR Steering Group (DD, JvB, GR, ASR, LH, BG, EJW, and DG) provided methodological input and oversaw the conduct of the study. JHV and DD conducted the thematic analysis and Delphi rounds analysis and produced the Delphi round summaries. All members of the steering group (JHV, DD, JvB, GR, BKW, PC, ASR, BOR, CB, LH, OJB, CJ, BG, EJW, BJB, DG, and MEvG) selected the final content and wording of the guidelines. JHV, DD, and MEvG chaired the consensus meeting. JHV, DD, and MEvG drafted the final manuscript and Explanation & Elaboration note. All authors reviewed and commented on the final manuscript and Explanation & Elaboration note. All members of the steering group collaborated in the development of the guidelines by participating in the Delphi process, the qualitative evaluation of the guidelines, or both.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The RATE-XR (reporting for the early-phase clinical evaluation of applications using extended reality) qualitative study guideline. [DOCX File , 335 KB - jmir_v26i1e56790_app1.docx]

References

1. Topol E. The Topol Review: Preparing the Healthcare Workforce to Deliver the Digital Future. England: Health Education England; 2019:104 URL: https://topol.hee.nhs.uk/
2. Yeung AWK, Tosevska A, Klager E, Eibensteiner F, Laxar D, Stoyanov J, et al. Virtual and augmented reality applications in medicine: analysis of the scientific literature. J Med Internet Res 2021;23(2):e25499 [FREE Full text] [doi: 10.2196/25499] [Medline: 33565986]
3. Bruno RR, Bruining N, Jung C, VR-ICU Study group. Virtual reality in intensive care. Intensive Care Med 2022;48(9):1227-1229 [FREE Full text] [doi: 10.1007/s00134-022-06792-0] [Medline: 35816236]
4. Kanschik D, Bruno RR, Wolff G, Kelm M, Jung C. Virtual and augmented reality in intensive care medicine: a systematic review. Ann Intensive Care 2023;13(1):81 [FREE Full text] [doi: 10.1186/s13613-023-01176-z] [Medline: 37695464]
5. Cipresso P, Giglioli IAC, Raya MA, Riva G. The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. Front Psychol 2018;9:2086 [FREE Full text] [doi: 10.3389/fpsyg.2018.02086] [Medline: 30459681]
6. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2017;11(11):CD008349 [FREE Full text] [doi: 10.1002/14651858.CD008349.pub4] [Medline: 29156493]
7. Jung C, Wolff G, Wernly B, Bruno RR, Franz M, Schulze PC, et al. Virtual and augmented reality in cardiovascular care: state-of-the-art and future perspectives. JACC Cardiovasc Imaging 2022;15(3):519-532 [FREE Full text] [doi: 10.1016/j.jcmg.2021.08.017] [Medline: 34656478]
8. Bruno RR, Wolff G, Wernly B, Masyuk M, Piayda K, Leaver S, et al. Virtual and augmented reality in critical care medicine: the patient's, clinician's, and researcher's perspective. Crit Care 2022;26(1):326 [FREE Full text] [doi: 10.1186/s13054-022-04202-x] [Medline: 36284350]
9. Tsai TY, Onuma Y, Złahoda-Huzior A, Kageyama S, Dudek D, Wang Q, et al. Merging virtual and physical experiences: extended realities in cardiovascular medicine. Eur Heart J 2023;44(35):3311-3322 [FREE Full text] [doi: 10.1093/eurheartj/ehad352] [Medline: 37350487]
10. Rizzo AS, Koenig ST. Is clinical virtual reality ready for primetime? Neuropsychology 2017;31(8):877-899. [doi: 10.1037/neu0000405] [Medline: 29376669]
11. Hirst A, Philippou Y, Blazeby J, Campbell B, Campbell M, Feinberg J, et al. No surgical innovation without evaluation: evolution and further development of the IDEAL framework and recommendations. Ann Surg 2019;269(2):211-220. [doi: 10.1097/SLA.0000000000002794] [Medline: 29697448]

XSL•FO
RenderX

12. McCulloch P, Altman DG, Campbell WB, Flum DR, Glasziou P, Marshall JC, Balliol Collaboration, et al. No surgical innovation without evaluation: the IDEAL recommendations. Lancet 2009;374(9695):1105-1112. [doi: 10.1016/S0140-6736(09)61116-8] [Medline: 19782876]
13. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
14. Beams R, Brown E, Cheng W, Joyner JS, Kim AS, Kontson K, et al. Evaluation challenges for the application of extended reality devices in medicine. J Digit Imaging 2022;35(5):1409-1418 [FREE Full text] [doi: 10.1007/s10278-022-00622-x] [Medline: 35469355]
15. Fleetcroft C, McCulloch P, Campbell B. IDEAL as a guide to designing clinical device studies consistent with the new European medical device regulation. BMJ Surg Interv Health Technol 2021;3(1):e000066 [FREE Full text] [doi: 10.1136/bmjsit-2020-000066] [Medline: 35047800]
16. Birckhead B, Khalil C, Liu X, Conovitz S, Rizzo A, Danovitch I, et al. Recommendations for methodology of virtual reality clinical trials in health care by an international working group: iterative study. JMIR Ment Health 2019;6(1):e11973 [FREE Full text] [doi: 10.2196/11973] [Medline: 30702436]
17. Kellmeyer P, Biller-Andorno N, Meynen G. Ethical tensions of virtual reality treatment in vulnerable patients. Nat Med 2019;25(8):1185-1188. [doi: 10.1038/s41591-019-0543-y] [Medline: 31359003]
18. Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE, Louvet-de Verchère F, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol 2021;31(6):3786-3796 [FREE Full text] [doi: 10.1007/s00330-020-07684-x] [Medline: 33666696]
19. Lewis CH, Griffin MJ. Human factors consideration in clinical applications of virtual reality. Stud Health Technol Inform 1997;44:35-56. [Medline: 10175342]
20. Riva G, Wiederhold BK. What the metaverse Is (really) and why we need to know about It. Cyberpsychol Behav Soc Netw 2022;25(6):355-359. [doi: 10.1089/cyber.2022.0124] [Medline: 35696299]
21. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7(2):e1000217 [FREE Full text] [doi: 10.1371/journal.pmed.1000217] [Medline: 20169112]
22. Humphrey-Murto S, Wood TJ, Gonsalves C, Mascioli K, Varpio L. The Delphi method. Acad Med 2020 Jan;95(1):168. [doi: 10.1097/ACM.0000000000002887] [Medline: 31335812]
23. Dalkey N, Helmer O. An experimental application of the DELPHI method to the use of experts. Manage Sci 1963;9(3):458-467. [doi: 10.1287/mnsc.9.3.458]
24. Vlake JH, Wils E, van Bommel J, Korevaar TIM, Gommers D, van Genderen ME. Virtual reality tailored to the needs of post-ICU patients: a safety and immersiveness study in healthy volunteers. Crit Care Explor 2021;3(5):e0388 [FREE Full text] [doi: 10.1097/CCE.0000000000000388] [Medline: 34079940]
25. Vlake JH, Van Bommel J, Wils E, Korevaar TIM, Bienvenu OJ, Klijn E, et al. Virtual reality to improve sequelae of the postintensive care syndrome: a multicenter, randomized controlled feasibility study. Crit Care Explor 2021;3(9):e0538 [FREE Full text] [doi: 10.1097/CCE.0000000000000538] [Medline: 34549192]
26. Vlake JH, Wils EJ, van Bommel J, Gommers D, van Genderen ME, HORIZON-ICU study group. Familiarity with the post-intensive care syndrome among general practitioners and opportunities to improve their involvement in ICU follow-up care. Intensive Care Med 2022;48(8):1090-1092. [doi: 10.1007/s00134-022-06782-2] [Medline: 35776162]
27. Vlake JH, van Bommel J, Wils E, Bienvenu J, Hellemons ME, Korevaar TI, et al. Intensive care unit-specific virtual reality for critically ill patients with COVID-19: multicenter randomized controlled trial. J Med Internet Res 2022;24(1):e32368 [FREE Full text] [doi: 10.2196/32368] [Medline: 34978530]
28. Bilbro NA, Hirst A, Paez A, Vasey B, Pufulete M, Sedrakyan A, IDEAL Collaboration Reporting Guidelines Working Group. The IDEAL reporting guidelines: a Delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. Ann Surg 2021;273(1):82-85. [doi: 10.1097/SLA.0000000000004180] [Medline: 32649459]
29. Dockx K, Bekkers EM, Van den Bergh V, Ginis P, Rochester L, Hausdorff JM, et al. Virtual reality for rehabilitation in Parkinson's disease. Cochrane Database Syst Rev 2016;12(12):CD010760 [FREE Full text] [doi: 10.1002/14651858.CD010760.pub2] [Medline: 28000926]
30. Clinical Evaluation. IMDRF Medical Device Clinical Evaluation Form. 2019. URL: https://tinyurl.com/5n6jxst9 [accessed 2024-09-27]
31. Executive Summary for the Patient Engagement Advisory Meeting: Augmented Rality and Virtual Reality Medical Device. 2022. URL: https://www.fda.gov/media/159709/download: [accessed 2024-09-27]
32. Public Workshop - Medical Extended Reality: Towards Best Evaluation Practices for Virtual And Augmented Reality in Medicine. 2020. URL: https://www.fda.gov/media/159709/download: [accessed 2024-09-27]
33. Vlake JH, van Bommel J, Riva G, Wiederhold BK, Cipresso P, Rizzo AS, et al. Reporting the early stage clinical evaluation of virtual-reality-based intervention trials: RATE-VR. Nat Med 2023;29(1):12-13. [doi: 10.1038/s41591-022-02085-7] [Medline: 36526724]
34. Vlake JH, Drop DLQ, van Genderen ME. Open Science Framework RATE-XR. 2023. URL: https://osf.io/cw6yq/ [accessed 2024-09-27]

XSL•FO
RenderX

35. Warström J. Mentimeter. URL: https://www.mentimeter.com/ [accessed 2023-06-15]
36. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020;2(10):e537-e548 [FREE Full text] [doi: 10.1016/S2589-7500(20)30218-1] [Medline: 33328048]
37. International Organization for Standardization. Information technology - artificial intelligence (AI) - bias in AI systems and AI aided decision making. URL: https://www.iso.org/standard/77607.html [accessed 2024-09-27]
38. International Organization for Standardization. Ergonomics of human-system interaction—part 11: usability: defnitions and concepts. 2018. URL: https://www.iso.org/ [accessed 2024-09-27]
39. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Br Med J 2010;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
40. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Br Med J 2007;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
41. Kouijzer MMTE, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]

Abbreviations

AI: artificial intelligence
AR: augmented reality
CONSORT: Consolidated Standards of Reporting Trials
EQUATOR: Enhancing the Quality and Transparency of Health Research
MR: mixed reality
OSF: Open Science Framework
RATE-XR: reporting for the early-phase clinical evaluation of applications using extended reality
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VR: virtual reality
XR: extended reality

Edited by N Cahill, T Leung; submitted 26.01.24; peer-reviewed by J Peek, X Liu; comments to author 25.04.24; revised version received 03.06.24; accepted 11.09.24; published 29.11.24.

Please cite as:
Vlake JH, Drop DLQ, Van Bommel J, Riva G, Wiederhold BK, Cipresso P, Rizzo AS, Rothbaum BO, Botella C, Hooft L, Bienvenu OJ, Jung C, Geerts B, Wils EJ, Gommers D, van Genderen ME, RATE-XR Expert Group
Reporting Guidelines for the Early-Phase Clinical Evaluation of Applications Using Extended Reality: RATE-XR Qualitative Study Guideline
J Med Internet Res 2024;26:e56790
URL: https://www.jmir.org/2024/1/e56790
doi:10.2196/56790
PMID:

©Johan H Vlake, Denzel LQ Drop, Jasper Van Bommel, Giuseppe Riva, Brenda K Wiederhold, Pietro Cipresso, Albert S Rizzo, Barbara O Rothbaum, Cristina Botella, Lotty Hooft, Oscar J Bienvenu, Christian Jung, Bart Geerts, Evert-Jan Wils, Diederik Gommers, Michel E van Genderen, RATE-XR Expert Group. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Needs Expressed in Peer-to-Peer Web-Based Interactions Among People With Depression and Anxiety Disorders Hospitalized in a Mental Health Facility: Mixed Methods Study

Dawid Storman[1], MD; Paweł Jemioło[2], MA, MSc; Zuzanna Sawiec[3]; Mateusz Jan Swierz[1], MD; Ewa Antonowicz[3]; Malgorzata M Bala[1], Prof Dr; Anna Prokop-Dorner[4], PhD

[1]Chair of Epidemiology and Preventive Medicine, Department of Hygiene and Dietetics, Jagiellonian University Medical College, Kraków, Poland
[2]AGH University of Krakow, Kraków, Poland
[3]Students' Scientific Research Group of Systematic Reviews, Jagiellonian University Medical College, Kraków, Poland
[4]Chair of Epidemiology and Preventive Medicine, Department of Medical Sociology, Jagiellonian University Medical College, Kraków, Poland

Corresponding Author:
Dawid Storman, MD
Chair of Epidemiology and Preventive Medicine
Department of Hygiene and Dietetics
Jagiellonian University Medical College
Kopernika 7
Kraków, 31-034
Poland
Phone: 48 124223720
Email: dawid.storman@uj.edu.pl

Abstract

Background: Hospitalization in psychiatric wards is a necessary step for many individuals experiencing severe mental health issues. However, being hospitalized can also be a stressful and unsettling experience. It is crucial to understand and address the various needs of hospitalized individuals with psychiatric disorders to promote their overall well-being and support their recovery.

Objective: Our objectives were to identify and describe individual needs related to mental hospitals through peer-to-peer interactions on Polish web-based forums among individuals with depression and anxiety disorders and to assess whether these needs were addressed by peers.

Methods: We conducted a search of web-based forums focused on depression and anxiety and selected samples of 160 and 176 posts, respectively, until we reached saturation. A mixed methods analysis that included an in-depth content analysis, the Pearson χ^2 test, and ϕ coefficient was used to evaluate the posts.

Results: The most frequently identified needs were the same for depression and anxiety forums and involved informational (105/160, 65.6% and 169/393, 43%, respectively), social life (17/160, 10.6% and 90/393, 22.9%, respectively), and emotional (9/160, 5.6% and 66/393, 16.8%, respectively) needs. The results show that there is no difference in the expression of needs between the analyzed forums. The needs were directly (42/47, 89% vs 98/110, 89.1% of times for depression and anxiety, respectively) and not fully (27/47, 57% vs 86/110, 78.2% of times for depression and anxiety, respectively) addressed by forum users. In quantitative analysis, we found that depression-related forums had more posts about the need for informational support and rectification, the expression of anger, and seeking professional support. By contrast, anxiety-related forums had more posts about the need for emotional support; social life; and information concerning medications, hope, and motivation. The most common co-occurrence of expressed needs was between sharing own experience and the need for professional support, with a strong positive association. The qualitative analysis showed that users join web-based communities to discuss their fears and questions about psychiatric hospitals. The posts revealed 4 mental and emotional representations of psychiatric hospitals: the hospital as an unknown place, the ambivalence of presumptions and needs, the negative representation of psychiatric hospitals, and the people associated with psychiatric hospitals. The tone of the posts was mostly negative, with discussions revolving around negative stereotypes; traumatic experiences; and beliefs that increased anxiety, shock, and fright and deterred users from hospitalization.

XSL•FO
RenderX

Conclusions: Our study demonstrates that web-based forums can provide a platform for individuals with depression and anxiety disorders to express a wide range of needs. Most needs were addressed by peers but not sufficiently. Mental health professionals can benefit from these findings by gaining insights into the unique needs and concerns of their patients, thus allowing for more effective treatment and support.

(J Med Internet Res 2024;26:e51506) doi:10.2196/51506

KEYWORDS

anxiety disorders; depression; peer-to-peer web-based interactions; needs; psychiatric hospitalization

Introduction

Background

Over recent decades, the prevalence of depression has risen globally, affecting approximately 280 million people in 2019 [1]. Major depressive disorder is often characterized by recurrent episodes and imposes substantial challenges, including social difficulties, increased health care use, diminished work productivity, and heightened mortality risk [2,3]. The COVID-19 pandemic exacerbated mental health issues, leading to a 25% surge in depression and anxiety rates worldwide, including among various populations in Poland [4-9]. During mental health crises, individuals may seek help independently or from others, encountering barriers in accessing professional psychiatric care, especially pronounced in countries such as Poland with a shortage of psychiatrists and other specialists [10-15]. Web-based forums provide a less stigmatizing platform for seeking mental health information and support, enabling users to interact with peers and access resources anonymously. Other key advantages of web-based mental health forums include anonymity, easy access, continuous availability, the option for passive content consumption, peer-to-peer type of interaction, informality compared with a physician-patient relationship, and free accessibility [16].

However, misinformation on social media presents risks, with 40% of shared health-related links being categorized as fake news [17]. Using web-based platforms for mental health support has become increasingly common, facilitating knowledge exchange, emotional support, and community engagement [18]. These platforms also regard the topic of hospitalization in mental health facilities.

There are not many strict indications regarding who should be hospitalized in a psychiatric hospital, and these indications may vary by country [19-21]. However, it is generally agreed that hospitalization is needed when a patient experiences an exacerbation of the disease and presents a risk to themselves or others [19]. The World Health Organization Division of Mental Health and Prevention of Substance Abuse published 10 basic principles of mental health care law [22], such as the promotion of mental health, access to basic mental health care, guidelines on assessing necessary treatment options, and patient autonomy, along with suggestions on selected actions to promote the implementation of these principles. However, not all countries follow or implement them.

Other indications for hospitalization include a rapid or severe deterioration of the patient's mental state, a skewed perception of reality, an inability to handle the patient's condition in the ambulatory setting, or the need for the patient to leave the home environment [19,23]. However, hospitalization can cause shock and trauma in patients, especially if it is forced. Therefore, such decisions should always be thoroughly considered [24]. In the context of psychiatric hospital admissions, there are generally 2 primary ways through which individuals can be admitted: voluntarily or involuntarily. Voluntary admission occurs when an individual willingly seeks treatment and agrees to be admitted to a psychiatric hospital. They may recognize the need for assistance with their mental health and voluntarily choose to seek help. By contrast, involuntary admission takes place when an individual is admitted to a psychiatric hospital against their will. Involuntary admission typically happens when someone's mental health condition poses a significant risk to themselves or others and they are unable or unwilling to seek help voluntarily. Involuntary admission usually involves legal procedures and may require the involvement of mental health professionals, law enforcement, or the judicial system. Across the European Union, the frequency of involuntary admissions to psychiatric hospitals varies considerably, ranging from approximately 3% to 30% of all inpatient episodes [25].

The rules and regulations for involuntary admission and treatment of patients with mental disorders differ markedly across countries [26]. Essential legal criterion for detention, in addition to mental disorder in most European countries (eg, Poland, Austria, Belgium, France, Germany, and the Netherlands), is a threat or an actual danger to one's own life or others' safety, the need for treatment (eg, Italy, Spain, and the United Kingdom), or both (eg, Denmark, Finland, Greece, and Ireland) [27]. In case of the maximum length of initial hospitalization, Italy and Poland are characterized by the shortest period (7 days and 10 days, respectively), while Belgium and Germany allow for the longest hospitalization (up to 2 years). In Poland, involuntary admission to a mental health facility is regulated by the Mental Health Act of 1994 (in Polish, Ustawa o ochronie zdrowia psychicznego). It is allowed in the following situations: (1) when the patient experiences an exacerbation of the mental state and poses a risk to themselves (article 23); (2) when the patient develops symptoms of a new disease and is admitted for diagnostic workup, just for observation, without pharmacotherapy, for a maximum stay of 10 days (article 24); and (3) when the court orders compulsory hospitalization (article 29) [28]. Admission to a psychiatric hospital is regulated by the Convention on the Rights of Persons with Disabilities drafted in 2006 [29]. Up to February 21, 2023, a total of 186 countries, including Poland, gave their explicit consent to be bound by the treaty [30]. The convention has a strong position on hospitalization without consent, and Poland is repeatedly reprimanded for failing to comply with its provisions [31].

XSL•FO
RenderX

The length of stay in a psychiatric hospital can vary depending on the severity of the mental condition and response to treatment, and again, this differs significantly between countries [32,33]. The mean time of hospital stay for any mental disorder ranges from 6.8 (SD 3) to 55.1 (SD 62.4) days [34-37]. On average, hospitalization is the longest in patients with schizophrenia, ranging from 21 to 290 days [38,39]. For patients with dual diagnosis, the mean reported hospitalization duration was 12 days [38]. The mean length of hospital stay for manic, depressive, and mixed episodes was 29.2, 29.9, and 43.3 days, respectively [40]. Finally, according to the literature, factors associated with longer hospitalization include relationship status (being single) and unemployment [41].

After every hospitalization, patients need to return to their regular environment. To facilitate this transition, the time of hospital treatment should be reduced, and dedicated programs should be available in the outpatient setting. The process of change that led to the deinstitutionalization of psychiatric care began in Western Europe and the United States in the 1950s. The revolution in psychopharmacology had a significant impact on these transformations, as it improved the effectiveness of treatment, allowing many patients to function outside the hospital setting. Pioneering attempts at new therapeutic methods began in Anglo-Saxon countries; France promoted psychiatry without psychiatrists, and the United States started preparations for the deinstitutionalization process. New legislation was introduced to foster a noninstitutional approach to psychiatric care (1946-1959 in the United Kingdom and 1955-1963 in the United States). In France, the sectorization system was introduced in the 1960s [42]. In Italy, reforms were started by the representatives of a new European generation that was critically oriented toward the revolution of psychotropic drugs and psychoanalysis and was fascinated by phenomenology, poststructuralism, and Marxism. The most radical current emerging from these attitudes was called *antipsychiatry*. The 1960s were the times of increasing left-wing and antiauthoritarian social moods. Controversies around psychiatry resulted in a tendency to question its traditional prerogatives. A trend that saw psychiatry as an oppressive element of the capitalist system emerged. The developing ideas of antipsychiatry corresponded well with this atmosphere, and around 1968, they became an important counterculture and political current. While British, French, and American antipsychiatrists contested the psychiatric care system, radical Italian psychiatrists, led by Basaglia, attempted to dismantle it from within and effectively created social support for the deinstitutionalization process [43]. In Germany, many hospital wards were closed as a result of a psychiatric reform, and a net of health care professionals providing integrated care was created. This form of treatment offers the possibility of treating more patients by applying different treatment intensities and contents [44].

Depending on their mental condition and the ability to function in day-to-day activities, patients can be admitted to an inpatient facility (eg, a hospital ward, psychiatric hospital, residential house, or, in specific cases, a high-security forensic unit), or treatment can be provided in the outpatient setting, with daily visits to the hospital without an overnight stay. However, hospitalization is the best way to control a patient's behavior and adjust treatment [45,46]. During psychiatric hospitalization, a person typically receives a range of mental health treatments, such as medication, psychotherapy, and support from a multidisciplinary team of mental health professionals. However, hospital wards are not always perceived as a safe space. Patients face challenges as they adapt to new conditions; cope with restrictions regarding visits and leaving the ward; and deal with other patients or personnel at the facility, where mistreatment, aggressiveness, theft, or substance abuse take place [47,48]. There have been numerous cases of mistreatment by staff members or other patients. Examples of such behaviors include physical and verbal abuse, neglect, and isolation. In some cases, staff members may go too far when using restraints or other forms of physical intervention, causing harm to patients [49-51].

Objectives

The role of the media in reinforcing mental health stereotypes and stigmas is an important issue. For example, some news outlets tend to sensationalize incidents of violence or other disruptive behaviors, thus portraying patients as dangerous or unstable. This can create a climate of fear and mistrust toward patients with psychiatric disorders, making it even harder for them to receive the care and support they need [52-55]. Alzahrani [56] suggested that there have not been enough studies that identify and examine the effect of hospitalization in a mental health facility on the emotional status and well-being of patients. Unmet needs are particularly problematic in patients with psychiatric disorders because their mental condition might be an obstacle to expressing these needs [57,58]. One such condition is dementia, as studied by Røsvik and Rokstad [59] in a systematic review aimed at identifying the needs of patients, caregivers, and staff in the context of acute hospitalization. There have been studies conducted with the objective of deepening our understanding of patients' needs. They concern identifying needs by professionals [60], a sense of safety during psychiatric care [61], or a feeling of being cared for by nurses [62]. However, it is not common for investigators to examine the needs of people in the hospital setting via social media. Therefore, the aim of this study was to fill in this gap by identifying and exploring the needs of people regarding hospitalization in mental health facilities via internet forums.

Methods

Overview

For this study, we used a mixed methods design (qualitative and quantitative methods). To ensure the scientific rigor and transparency of our research, we adhered to standards for reporting qualitative research and qualitative meta-analysis, which provide systematic guidelines for describing research questions, study design, data collection and analysis, findings, and study implications [63]. Methods used in this study are briefly described in the subsequent sections. Detailed methodology was reported previously (Storman, D, unpublished data, May 2024).

XSL•FO
RenderX

Data Collection

We conducted a search for Polish-language web-based forums dedicated to individuals with depression and anxiety disorders. We used the following search terms in Polish: “forum” (in English, forum), “depresja” (depression), and “zaburzenia lękowe” (anxiety disorders) and scanned the first 10 pages of Google (Google LLC) search results. Of the 10 identified forums, we selected 2 (20%) forums with the highest number of registered users and posts: one forum related specifically to depression (depresja forum) and the other related to anxiety disorders (nerwica forum). Posts were obtained using a web scraping technique with Python (Python Software Foundation) code on July 18, 2021.

The selected forums were free of charge, open to the public, noncommercial, and nongovernmental. In 2005, the depresja forum [64] was established for individuals experiencing depression, while nerwica forum [65], also established in 2005, was intended for people with anxiety disorders. The statute of the forums indicates that the key objective is to facilitate anonymous interactions and mutual support among individuals experiencing depression or anxiety disorders. The content is monitored by volunteer peers who ensure adherence to forum rules. The users were free to open threads and answer them. The 2 selected forums did not require users to register to create a discussion thread, post comments, or reply to comments from other users. Forum discussions were asynchronous, with posts appearing in sequence, which means that the simultaneous presence of all participants in the thread was not necessary. The moderators of the forums did not provide any therapeutic interventions. While individuals without personal experience of depression or anxiety disorders, such as family or friends, could participate in the forums, only posts indicating that their authors had personal experience with mental disorders were considered for the analysis.

Sampling

We sampled the posts (160 from depression forums and 176 from anxiety forums) until achieving saturation of the data, that is, the content provided rich, detailed information relevant to the research question, and no new data were obtained in the following posts. We did not impose any limits regarding time, that is, when the post was published or which stage of hospitalization (before, during, or after) its content considered.

Coding

We used a deductive coding approach to assess the selected material, based on a codebook developed in the previous study (Storman, D, unpublished data, May 2024). Our unit of coding was a single interaction, consisting of a presentation of a need (which could be found in the opening post or within the discussion) and a response (posts within the same forum thread where the authors quoted a specific post, mentioned a specific user, or referred to a specific content). The unit of analysis was a forum thread. We considered only threads with at least 2 posts, as this was the minimum number of posts required to observe interactions. We coded the interactions between forum users that revealed their needs related to psychiatric hospitalization. Every post was coded based on how the need was presented: explicitly (a direct question) or implicitly (a problem disclosed indirectly). We also coded how peers responded to the needs, categorizing responses as fully or not fully addressed (eg, off topic) and as direct or indirect (eg, referring to the asker in the third person). We evaluated the adequacy of responses only when it was indicated by the asker, such as by thanking the responder or by expressing dissatisfaction with the response.

Reliability

To ensure coding quality, 4 coders underwent coding training with an experienced qualitative researcher involved in the study. Independent coders achieved an agreement rate of >70%. Any disagreements were resolved through discussion.

We began by immersing ourselves in the data to gain a comprehensive understanding of the content of the posts. Each coder kept a logbook to record their observations and insights as they worked through the data. After coding the data by 2 independent coders and comparing the coded fragments across different users, threads, and forums, we aggregated segments of posts that were coded with the same code and compared them within and between cases. This enabled us to identify patterns and variations in the data, identify the main themes, and gain a deeper understanding of the main themes and trends in the posts. Throughout the analysis, we wrote summaries and memos to document our observations, hypotheses, and insights, which helped us better understand the emerging patterns.

Our team included interdisciplinary investigators with different professional backgrounds, such as medicine and health sciences (including psychiatry, epidemiology, and public health), psychology, and sociology, which enriched our discussions and reflexivity throughout the study. The coding team convened on multiple occasions to ensure a shared understanding of the material. Upon the completion of the coding process, the first author continued with subsequent analysis steps. To foster reflexivity, routine group discussions were conducted throughout the study, focusing on scrutinizing premature observations and challenging underlying assumptions.

Analysis

In this study, we used a mixed methods approach to content analysis. Qualitative content analysis was used to understand experiences presented in the posts, while quantitative analysis was used to provide numerical descriptions of the data and explore relationships between needs and trends. We coded the interactions between forum users that revealed their needs related to hospitalization in psychiatric units using a codebook developed in our previous study [66]. Coding quality was ensured by training all the coders. Any discrepancies were resolved through discussion. The qualitative analysis was conducted by following the approach suggested by Krippendorff [67] and using MAXQDA 2022 (VERBI Software) [68,69], a qualitative data analysis software, for coding, annotating, retrieving, and reviewing textual data.

Using the SPSS (IBM Corp) software [70] to determine whether there is a significant association between the groups (forum members experiencing depression and anxiety disorders), the Pearson χ^2 test was applied, with the significance level set at $P<.05$. The co-occurrence of codes was presented using a

XSL•FO
RenderX

circular chart (co-occurrence frequencies) to investigate possible associations between interactions. The graph was prepared using NetworkX (version 2.5.1; Los Alamos National Laboratory) for Python. We used the ϕ coefficient [71] to examine the associations between the types of different interactions.

To differentiate the authors of the posts, we used 2 criteria: self-reported gender (woman, man, or unknown) and the level of activity on the forum based on the number of posts published. The categories of activity levels were defined as follows: 1 to 500 posts: occasional user; 501 to 1000 posts: regular user; 1001 to 5000 posts: super user; and >5000 posts: power user. For those who were not registered on the forum and remained anonymous (guests), it was impossible to determine the level of activity. Therefore, their level of activity was not reported.

Ethical Considerations

On the basis of the guidance and ethical recommendations of the Secretary's Advisory Committee on Human Research Protections, our study did not require ethics approval and informed consent because it involved the collection and analysis of publicly available information from internet forums. All data collected in this study were anonymized and deidentified. The study was strictly archival and cross-sectional in nature, without any intervention or interaction with the forum users, and only used preexisting information [72-74].

Results

Quantitative Description of the Material

Overview

The posts included in this analysis were published between 2013 and 2021. The mean length of coded posts from the anxiety-related forum was 485 (SD 477; range 24-2255) signs and that from the depression-related forum was 518 (SD 691; range 41-5773) signs. A detailed description of the forums was compiled previously (Storman, D, unpublished data, May 2024). In this study, we focused exclusively on posts discussing psychiatric hospitalization.

The frequencies of codes and their descriptions are presented in Table 1 (main codes) and Table S1 in Multimedia Appendix 1 (all codes).

XSL•FO
RenderX

Table 1. The frequencies of the main codes and subcodes[a].

Codes	Code description	Depression (n=160), n (%)	Anxiety disorders, (n=393), n (%)	*P* value	χ^2 (*df*)
Need for emotional self-expression					
All posts	—[b]	13 (8)	22 (6)	.28	1.2 (1)
Negative emotions	• We categorized posts containing verbal or nonverbal (emoticons) indicators of negative emotions, such as hostility, frustration, and the author's anger, which were directed toward previous statements, such as "You're stupid. It's impossible to read" and ">_<." • However, we did not categorize posts that expressed sadness in the context of empathy or when reflecting the emotions of another forum's user, such as "I feel sad when I read this."	8 (5)	10 (3)	.15	2.1 (1)
Positive emotions	• We categorized posts that contained verbal or nonverbal (emoticons) expressions of joy, gratitude, and pride in response to the content of forum users, such as "(ˆ v ˆ)" and "I'm happy when I read that you succeeded." • However, we did not categorize posts that conveyed emotions that hinted at sarcasm or spite, such as "I'm very happy about your misfortune."	5 (3)	12 (3)	.98	0.0 (1)
Need for support (receiving and offering)					
All posts	—	116 (72)	246 (63)	.10	2.8 (1)
Informational	• We categorized posts that discussed symptoms, treatments, and medication side effects based on either book knowledge or personal experience, for example, "Duloxetine can be used to relieve pain," "I was in treatment for 15 years, no one could help me until I finally tried...," and "Article 53 of the Mental Health Protection Act states that giving false information about the symptoms of mental disorders is a crime."	105 (66)	169 (43)	<.001	17.2 (1)
Emotional	• We categorized posts that showed normalization, compassion, self-esteem support, tension release, hope, motivation, and well wishes, for example, "Don't give up hope; everyone makes mistakes, but you have many strengths, and you are resilient."	9 (6)	66 (17)	<.001	11.7 (1)
Instrumental	• We categorized posts that offered or sought advice on how, when, or where to do something regarding mental health, such as "If you crush the pill first and then wash it down with plenty of water, it will be easier to swallow."	2 (1)	11 (3)	.27	1.2 (1)
Need for social interactions					
All posts	—	31 (19)	126 (32)	.004	8.2 (1)
Rectification	• We categorized posts that involved admitting to a mistake and attempting to resolve a communication misunderstanding with another user, such as "I misunderstood and thought the discussion was about venlafaxine."	4 (2)	2 (1)	.04	4.1 (1)
Disagreement	• We categorized posts where users expressed disagreement with others and presented a different opinion, for example, "I disagree; herbal medicine is more effective than drugs."	2 (1)	7 (2)	.65	0.2 (1)

XSL•FO
RenderX

Codes	Code description	Depression (n=160), n (%)	Anxiety disorders, (n=393), n (%)	*P* value	χ^2 (*df*)
Admitting that someone is right	• We categorized posts that demonstrated agreement with another user's statements and expressed approval, such as "I agree with you" or "B wrote the truth."	4 (2)	11 (3)	.84	0.0 (1)
Community values	• We coded posts that promoted attitudes of openness and tolerance and provided information about the community's values and prohealth behaviors, such as "It's great to see you back on medication; it will make a difference." • However, we did not categorize posts that advertised products or services, promoted unhealthy behaviors, or expressed sympathy toward someone else, such as "All differences are accepted here."	3 (2)	13 (3)	.36	0.8 (1)
Social life	• We categorized posts that encouraged others to share their own experiences, tell jokes, ask questions, and post comments in general, such as "What medications were you taking at that time?" • However, we did not categorize posts that focused only on revealing one's own experiences and exhibited signs of aggression, hostile sarcasm, or irony.	17 (11)	90 (23)	.001	10.4 (1)
Private messages	• We coded posts that indicated an interaction outside the forum, such as "I sent you a private message."	0 (0)	1 (0.3)	.52	0.4 (1)
Referring to the rules	• We coded posts related to forum regulations and what is allowed and not allowed, such as "According to the rules, we cannot post information about drug sales here."	1 (1)	2 (1)	.87	0.0 (1)

[a]Differences between forums for people experiencing depression versus anxiety disorders. A *P* value of <.05 was considered significant.
[b]Not applicable.

Depression Versus Anxiety Disorders

There was a difference in the dynamics of depression and anxiety forums, as evidenced by the higher number of codes for the identified needs on the anxiety forum than for those on the depression forum (394 vs 160). In general, the most frequently identified needs were the same for depression and anxiety forums and involved informational (105/160, 65.6% and 169/393, 43%, respectively), social life (17/160, 10.6% and 90/393, 22.9%, respectively), and emotional (9/160, 5.6% and 66/393, 16.8%, respectively) needs.

Looking into subcategories of codes, there was a significant difference between depression and anxiety disorders in terms of the need for social interactions (*P*=.004). Forum users expressed their need for informational support and rectification significantly more often on the depression-related forum than on the anxiety-related forum (*P*<.001 and *P*=.04, respectively). By contrast, needs for emotional support and social life were expressed more frequently on the forum for people with anxiety disorders (*P*<.001 and *P*=.001, respectively).

The following needs were expressed more frequently on the forum for depression than on the forum for anxiety disorders: need for expressing anger (*P*=.03), need for information concerning general knowledge or disorders (*P*<.001), need for professional support (*P*=.007), need for information based on own experience (*P*<.001), need for sharing own experience (*P*=.004), and need for the assessment of the situation (*P*=.004).

The posts describing needs for information concerning medications (*P*=.003), hope and motivation (*P*=.02), supporting self-esteem or admiration (*P*=.03), querying or refining (*P*=.005), and joking (*P*=.01) occurred more frequently on the anxiety-related forum than on the depression-related forum.

The co-occurrence of all expressed needs (codes) is shown in Figure 1. The co-occurrence was observed most often for the need for sharing own experience (node F3 in Figure 1) and the need for information regarding professional support (node C1 in Figure 1; 11 times). There was a positive association between these 2 needs (ϕ=0.966; *P*<.001), which was the strongest positive association noted between 2 interactions.

XSL•FO
RenderX

Figure 1. Circular chart of the co-occurrence of needs (component of specific interactions) on both forums together. Node size corresponds to the number of posts where a specific need was expressed. The edges of the graphs indicate co-occurrence with other codes, and their size is proportional to the number of co-occurrences. A1: anger; A2: reluctance; A3: irony and sarcasm; A4: rejection; B1: joy; B2: gratitude; C1: regarding professional support; C2: concerning general knowledge or disorders; C3: concerning medications; C4: regarding adverse effects; D1: I have that too and normalization; D2: wishes; D3: hope and motivation; D4: being with someone; D5: tension release; D6: supporting self-esteem and admiration; D7: compassion; E1: other sources and see more; F1: assessment of the situation; F2: unprofessional information; F3: sharing own experience; G1: instrumental; G2: warnings; H1: rectification; H2: disagreement; H3: admitting that someone is right; I1: acceptance; I2: promotion of expected behaviors; I3: the importance of relationships; I4: mutual protection; J1: encouraging sharing; J2: querying and refining; J3: small talk; J4: joking; K1: referring to the rules.

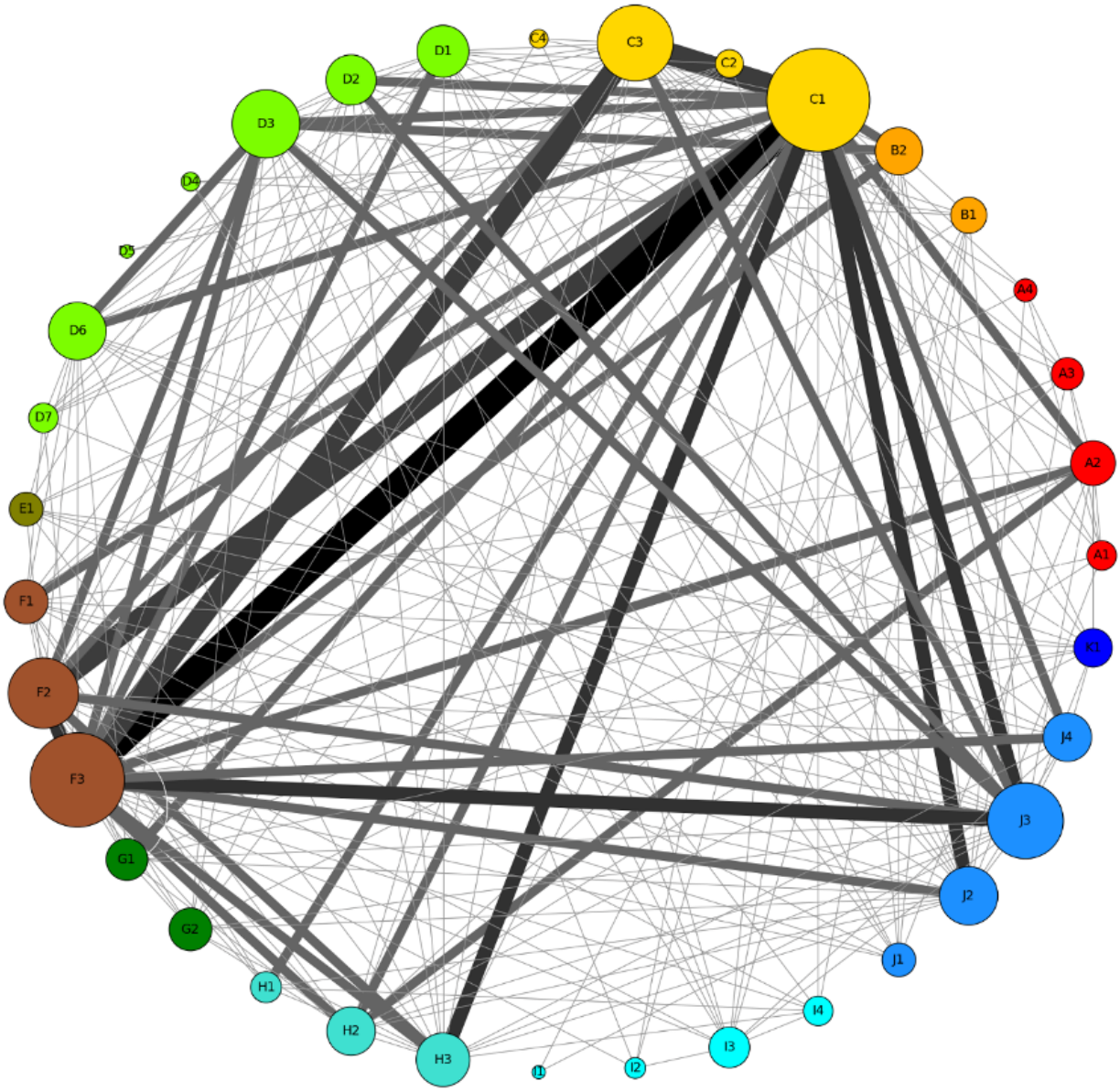

Implicit and Explicit Needs

Differences in the frequencies of expressed needs (explicitly and implicitly) and addressed needs (fully, not fully, directly, and indirectly) between the 2 forums are summarized in Table 2. There were no significant differences between the forums regarding the explicit and implicit ways of expressing the needs (Table 2). The needs were fully addressed more frequently on the depression-related forum than on the anxiety-related forum (Table 2). No other significant differences were observed between the forums in the way needs were addressed (Table 2).

XSL•FO
RenderX

Table 2. Frequencies of expressed and addressed needs on the forum for people with depression versus on the forum for people with anxiety disorders[a].

	Depression forum posts, n (%)	Anxiety disorder forum posts, n (%)	*P* value	χ^2 (*df*)
Expressing the need[b]				
Explicitly	14 (50)	30 (43)	.71	0.1 (1)
Implicitly	14 (50)	39 (57)	.58	0.3 (1)
Addressing the need				
Fully	20 (43)	24 (22)	.01	6.5 (1)
Not fully	27 (57)	86 (78)	.08	3.1 (1)
Directly	42 (89)	98 (89)	.86	0.0 (1)
Indirectly	5 (11)	11 (11)	.88	0.0 (1)

[a]Differences between forums for people experiencing depression versus anxiety disorders. A *P* value of <.05 was considered significant. Depression: for expressing the need, n=28, and for addressing it, n=47; anxiety: expression, n=69, addressing, n=110.

[b]Every post was coded based on how the need was presented: explicitly (a direct question) or implicitly (a problem disclosed indirectly).

In-Depth Content Analysis

The analysis suggests that forum users turn to web-based communities to discuss their questions and fears regarding psychiatric hospitalization. The analysis of interactions between forum users revealed 4 mental and emotional representations of a psychiatric hospital: a sense of the unknown, a fear of hospitalization, utterly negative images, and a longing for feeling at home when at the hospital.

The Hospital as an Unknown Place

From the users' posts, an image of a psychiatric hospital as a place shrouded in mystery emerged, with information about it derived mostly from other people, including professionals (Table S2 in Multimedia Appendix 1, quotation 1). This stimulates speculations, fantasies, and various beliefs about the psychiatric hospital. Confusion may generate anxiety among users about what to expect in psychiatric hospitals (quotation 2). Even for people who have been hospitalized many times and have previously collided their ideas about the hospital with reality, hospitalization can be surprising or even scary (quotation 3).

One of the users suggested that he would report his hospitalization on the forum; for him, this was a kind of mission, or a sacrifice, that would provide insights for others into the psychiatric hospital (quotation 4). A decision to go to the hospital was associated with fear, even for one's own life (quotation 5). The degree of emotional tension before a planned hospitalization is evidenced by the farewell to family, as described by a forum member (quotation 6).

The Ambivalence of Presumptions and Needs

Many forum users wrote about the dilemma of whether and when to be hospitalized. The lack of knowledge about what it is like to stay at the psychiatric hospital generated fears fueled by common negative stereotypes. The questions asked by forum members revealed a conflict between these fears and the need to feel safe and cared for. According to the posts, hospitalization was considered necessary when the mental health condition became more severe due to the discontinuation of medication (quotation 7), when the patient posed a threat to themselves or others (quotation 8), when there was a need to verify the diagnosis (quotations 9 to 11), or when the current treatment was ineffective (quotations 12 and 13). The decision of whether to go to the hospital was difficult because users were faced with the need to choose between 2 unappealing options: either take the risk of experiencing the unknown in the hospital or continue to experience emotional problems without any specialist treatment. Doubts were resolved on the forums, where users looked for information about hospitals or asked others to assess whether their condition required hospitalization (quotation 14). A user described hospitalization as a last resort (quotation 15).

"Hospitals Are a Total Mess"

Many posts reflected a negative approach to psychiatric hospitals and their functioning or organization. There were 4 areas that stirred particular interest among forum users: accommodation, a community of people (staff and other patients), drugs, and infrastructure. The tone of posts regarding these areas was mostly negative based on the user's own experience or the experience of family and friends as well as beliefs and emotions. This increased anxiety, shock, and fright and generally made users less inclined to go to the hospital.

Some users described their stay in the hospital, and many responders expressed interest in the admission procedure. Numerous posts indicated that forum users do not know the indications for involuntary hospitalization (quotation 16). Some forum members described scenarios in which the freedom to decide about oneself could be taken away, but others would not accept them. Those who did not agree with suggestions of forced admission maintained that individuals had an unconditional right of self-determination in any situation (quotations 17 and 18). A total of 2 users described situations where they were transported to the ward by the police (quotation 19). They shared traumatic memories and compared the situation to the apprehension of a criminal by the services or a deportation to a concentration camp. They felt as if they were treated unlawfully and in an unprofessional manner:

> *[The doctor] twisted my words and made up psychiatric epithets on the basis of these distortions.*

XSL•FO
RenderX

The reported experiences of involuntary admission revolved around the feeling of objectification, the deprivation of basic rights, slavery, and even dehumanization:

> *They kept [me]...in confinement like some animal.* [Quotation 20]

In addition, for some users, the work schedule at the ward resembled rigorous training as in the army, with patients being disciplined by orders (quotation 21).

Some users discussed how hospitalization might affect their mental health. Some forum users had expectations of "miraculous healing" that collided with the actual hospital experience of other users in whom symptoms persisted after hospitalization. Failure to get rid of symptoms during hospitalization inclined users to negatively assess the effectiveness of hospitalization or even consider therapeutic effects to be harmful (quotations 22 and 23).

People associated with the psychiatric hospital, both the staff and other patients, were an important topic among users when sharing observations from the hospital stay or beliefs not supported by their own experience. When writing about hospitalized individuals, users focused mainly on how different patients in closed wards were. These patients were described using vivid imagery and stereotypical representations of a psychiatric hospital popular in the media, which bring to mind a dehumanized image of a patient who is constantly in a straitjacket (quotation 24), is "dependent" (quotation 25), or gets into fights (quotations 26 and 27). Patients may fear that they will be seen as "criminals" who are "constrained for months" or that they will be harmed by "dangerous patients" (quotations 28 and 29). Such beliefs were shared by forum users themselves. They devalued other patients by calling them "freaks" or "plants that are soiling themselves" (quotations 30 and 31).

Medical care in the hospital was considered insufficient. In the opinion of forum users, clinicians do not work hard enough because they focus on more profitable activities in the private sector (quotation 21). According to some forum users, the staff lacks basic empathy and understanding regarding how difficult it is for their patients. One of the forum members suggested that to overcome these gaps, patients and specialists should switch roles (quotation 32).

Forum posts also revealed the anticipation of the stigma associated with psychiatric hospitalization. Users experienced it on many levels, for example, from staff, colleagues, or even themselves. Some users expressed concerns that they will experience prejudice due to common stereotypes around psychiatric hospitalization. One such stereotype is that people with mental disorders require special care and even control because of their unpredictable behavior. Patients actually encountered this kind of treatment in hospitals—they described how their right to privacy was restricted or how they were treated like a child who is under control all the time (quotation 33). Their posts revealed feelings of frustration, resentment, and anger. Forum users also described the experience of being stigmatized by hospital staff or hearing insults in the ward (quotations 34 and 35).

Pharmacological treatment was the third domain where the need for information support was revealed on the forum. Some users wrote about unclear restrictions on their access to medications in a place where, in their opinion, these medications should be provided. They described helplessness and that they had to humiliate themselves and sometimes even beg for medicines that they believed would help them (quotation 36). By contrast, some forum users pointed out that there may be a tendency to solve all the problems of the patient with drugs, without digging into the cause of the condition. When one set of drugs does not help, it is replaced with another and then with another and yet another, which builds an impression that physicians, due to their powerlessness or incompetence, have to experiment on the patient by trial and error (quotations 37 and 38). In addition, there was a feeling of being misunderstood by the staff, for whom the only method to deal with the patient's difficult emotions was to suppress them (quotation 37).

One of the posts indicated that the patient did not trust the staff. When symptoms became unbearable, the patient could not count on the help of specialists, for example, to be given a medication as needed, but they had to manage on their own and use their hidden medications (quotation 39).

Finally, another aspect of the psychiatric hospital that was negatively represented on the forum was poor infrastructure. The posts revealed that hospitals are short not only of amenities but also of basic equipment for everyday functioning, such as electricity (quotation 40). From what users shared on the forum, hospitalization appeared to be a dehumanizing experience due to "shitty food" and the fact that patients were placed in dark, dingy dormitory rooms "without handles" (quotations 41-43). Moreover, the hospital was viewed as a place that does not provide a sense of security, either regarding one's health and life (quotation 44) or regarding personal belongings (quotation 45). The extent of negative experiences and the lack of hope that hospital conditions would improve caused anger among forum users. One of them said that only the demolition of the hospital and building it from scratch might improve the current situation (quotation 46).

"Almost Like Home"

In contrast to the negative images of a psychiatric hospital that were predominant on the forum, there were also statements of people with positive experiences with institutional care. A user said that he felt "almost at home" in the hospital (quotation 47). Some people voiced their disagreement with negative statements. Users described a hospital that had a busy schedule in terms of classes and therapy and provided round-the-clock specialist care (quotation 48), and patients could feel that their private belongings were safe (quotation 49). Other users pointed out that although patients have to wait for a place in the ward, hospitalization offers various interesting activities (quotation 50) and helps reduce symptoms (quotation 51), establish a diagnosis and adjust medication (quotation 8), or provide a sense of control over medication (quotation 52). Users underlined the unique experience of staying at open wards and described them as pretty and a place of fun where one can feel as in a sanatorium or at a summer camp where discos take place (quotations 53 and 54).

XSL•FO
RenderX

However, another advantage of hospitalization that users mentioned in their posts was the opportunity to make new friends and share experiences with patients who are no different from other people and might be cool (quotation 55). According to some other people, the hospital has an advantage over outpatient treatment, for which there is a long waiting time. In the hospital, medications can be administered as needed and modified very quickly depending on the patient's current mental state. This can accelerate recovery and is in line with the principles of patient-centered care (quotation 56).

Building Bonds

Our analysis showed that the forum enhances the mechanisms that facilitate building a community based on giving hope, motivating others, and showing gratitude and admiration (quotations 4 and 57-61), for example, in response to posts expressing a fear of the first hospitalization in a psychiatric hospital (quotations 57 and 62) or doubts about the treatment received (quotation 63). Some users used irony to tighten bonds on the forum or reduce tension caused by the descriptions of hospital admission (quotation 64) or facing an unacceptable diagnosis during hospitalization (quotation 65). Moreover, some forum users were irritated when others described a division into "better and worse hospitals," which might be associated with a sense of injustice related to "better and worse treatment" (quotation 66).

Discussion

Principal Findings and Context of Other Studies

This study assessed the needs and emotions expressed in web-based forums for people with depression and anxiety disorders, with a focus on the representation of psychiatric hospitals. The most frequently identified needs were the same for depression and anxiety forums and involved informational, social life, and emotional needs. The needs were directly and not fully addressed by forum users. The needs expressed differed between the forums for depression and anxiety disorders, with a more frequent search for emotional support and social life on the anxiety forum and for informational support and rectification on the depression forum. Because anxiety can be accompanied by feelings of restlessness or agitation, individuals who are struggling with anxiety may be more likely to seek out emotional support and social interaction as a way of managing their symptoms [75,76]. They may also be more likely to experience social anxiety or difficulty forming and maintaining relationships, which could further drive their need for emotional support and social connection [77,78]. As for depression, it is often characterized by a persistent negative view of oneself, the world, and the future, which can lead to distorted thinking patterns and difficulty making decisions. Therefore, individuals who are struggling with depression may be more likely to seek out information and guidance that could help them make sense of their experiences and understand their symptoms [79,80]. By contrast, anxiety is often characterized by excessive worry or fear, which can lead to rumination and overthinking. As a result, individuals who are struggling with anxiety may be less likely to seek out information or rectification, as they may already be overwhelmed by the amount of information and may be struggling to make sense of it [81,82]. Another possible explanation is that depression may be more closely associated with feelings of helplessness or hopelessness, which can drive individuals to seek out information or resources that may help them regain a sense of control over their lives [83,84]. Moreover, individuals with depression may be more likely to experience feelings of lethargy or apathy, which could make it more difficult for them to seek out social interaction or emotional support. They may also feel shame or stigma surrounding their condition, which could further discourage them from asking others for help or support [85-87].

We noticed that posts about sharing experiences after staying in the psychiatric hospitals most often occurred together with the need for information regarding professional support. One possible explanation is that individuals who have had similar experiences with mental health issues may be more likely to seek out information and support from mental health professionals. They may also be more willing to share their experiences with other people who are going through similar struggles and may be seeking advice or guidance [88-90]. In addition, it is possible that individuals who have shared their own experiences may have identified gaps in their knowledge or understanding of mental health issues and, therefore, may be more likely to seek out information and support from mental health professionals [91-94].

Psychiatric hospitalization is an important step in the treatment of some patients. Our qualitative analysis revealed that forum users turn to web-based communities to discuss their questions and fears regarding psychiatric hospitals. Moreover, it revealed 4 mental and emotional representations of psychiatric hospitals: a sense of the unknown, a fear of hospitalization, utterly negative images, and a longing for feeling at home.

Many authors of the posts analyzed in this study had a dilemma of whether and when to go to a psychiatric hospital. The dilemma was caused by the lack of knowledge on what hospitalization looked like and what to expect. This increased the level of fear and anxiety.

Our analysis revealed that there were many negative stereotypes about psychiatric hospitals, and most posts about hospitalization were written in a negative tone. This caused greater anxiety, shock, and fear and made users less willing to decide on hospitalization. The perception of hospitalization is influenced by numerous factors, and despite obvious benefits, patients are often dissatisfied with their stay at the hospital [95]. The common reasons for negative experiences include safety concerns, unpleasant communication with hospital staff, and the use of coercive measures [96-98].

Each hospitalized patient should feel safe during their stay at the hospital; however, this is often not the case. It was reported that hospitalization in a closed ward was associated with poor patient satisfaction, which may be linked to restrictions on freedom and the use of coercive measures [99,100]. If coercive measures are used, the patient may feel a sense of powerlessness and humiliation and develop posttraumatic stress disorder [100]. Importantly, this reflects on the relationship with the hospital staff, which is critical for the healing process [99].

XSL•FO
RenderX

The negative perception of hospitalization in a psychiatric hospital may result from stereotypes perpetuated in the media and society that have been consistently harmful and imprecise [101]. This refers, among others, to movies and stories in which psychiatric hospitals are presented as scary and haunted, while people with mental illness are presented as violent and unstable [102]. The way mental illness is described in the media is inaccurate and exaggerated, which only reinforces the stigma [103]. Moreover, a history of psychiatry documented cases of abuse and exploitation, such as locking political prisoners in psychiatric hospitals in the Soviet Union regime [104].

The negative perception of psychiatric hospitalization may also stem from the fact that many psychiatric hospitals are housed in old and unwelcoming buildings, as is the case in many Central European countries, such as Poland, where psychiatric care has been underfinanced for decades [105,106]. Funds for the public sector of psychiatry are extremely low, resulting in understaffing, limited resources, and the lack of modern amenities [107].

The low effectiveness of treatment in psychiatric hospitals can be linked to multiple rehospitalizations [108]. This can be caused by various factors, such as inadequate follow-up care, lack of community resources, and insufficient support for patients after discharge [109]. The issue of psychiatric hospital readmissions is strictly related to the health care system and other contextual factors, such as the financing system and governance structure [110].

One of the important topics that emerged from our analysis was the relationship between patients and staff members. It was reported that maintaining high-quality relationships facilitates the inpatient care pathway and contributes to a reduction in the use of coercive measures [111]. The lack of meaningful communication, absence of regular ward staff, poor staff attitude, and inconsistent and abusive behavior of staff members are some of the potential barriers for patients. Forcible restraint prevents patients from seeking help earlier and is perceived negatively by most patients [111,112]. Different modes of admission to psychiatric units, including involuntary admission, can raise concerns about personal autonomy and boundaries. Patients may feel devalued and uncertain about their rights and decision-making capacity, which can contribute to negative perceptions [99,113].

Our analysis showed that hospitalization in a psychiatric ward is often associated with stigma. This is in line with other studies by Volpe et al [114] and Mutschler et al [115]. People with a history of hospitalization for mental health issues may feel labeled as “crazy” or “unstable” by others. This can lead to feelings of shame, embarrassment, and isolation and may prevent individuals from seeking help in the future [116].

To combat stigma and promote acceptance, it is important to raise awareness about mental health and the experience of people who have been hospitalized in a mental health facility. This can be achieved via education campaigns, support groups, and other initiatives that seek to reduce discrimination and promote understanding. Successful reduction of stigma has the potential to promote a healthy workplace environment, increase the quality of care, and improve treatment outcomes in individuals with stigmatized health conditions [117].

While there were many negative reflections among forum users, we also identified posts describing positive experiences with psychiatric hospitalization. This was linked, among other factors, to the concept of peer support, in which individuals with similar experiences help other patients. This observation was in line with previous research by Katsakou and Priebe [113]. The sense of community can be fostered by psychoeducation programs run by patient-led organizations and psychiatric rehabilitation organizations [118,119]. Peer support can help patients feel less lonely in their struggles and give hope for recovery. In addition, as these programs are often run by individuals who have firsthand experience with hospitalization, they can offer valuable insights to patients and their families. By incorporating peer support and community building into psychiatric care, patients can feel more empowered on their journey to recovery and less stigmatized by their hospitalization experience [120].

Limitations

This study has several limitations. We identified needs only based on the experiences described by forum users, which may not represent the full spectrum of needs. Other needs may not have been expressed because users either found them difficult to articulate or projected them onto others. There may also be groups of people who have unmet needs but do not have access to web-based forums. The anonymity of the forums made it challenging to verify the identity of users or collect demographic data.

Future Studies

Nevertheless, our content analysis revealed the diversity of ways in which individuals present their experiences from various events of the treatment trajectory. The findings of this study can serve as a starting point for more in-depth research that considers narrative development (sequence of posts) and collects more detailed data on users’ hospitalization experiences. As our results indicate, there are similarities and unique needs in the 2 groups of the analyzed population. Because there is a lack of studies considering people experiencing psychiatric diagnoses other than depression and anxiety via social media and forums, we believe it would be beneficial to explore the needs of these patients. It may contribute to a more comprehensive understanding of mental health experiences and facilitate more inclusive and effective mental health support systems.

Conclusions

Web-based forums offer a platform for people experiencing depression and anxiety disorders to express their needs and engage with other individuals who seek information on psychiatric hospitalization. Our study demonstrates that web-based forums can be used by individuals with depression and anxiety disorders to express a wide range of needs. Most needs were addressed by peers but not sufficiently. Our study revealed that forum users share fears and stereotypes regarding hospital stay. Our findings can help mental health professionals understand their patients’ needs and fears, thus guiding the choice of the most appropriate treatment. Using a patient-centered approach, mental health professionals can tailor

XSL•FO
RenderX

treatment to address specific concerns and underlying issues, resulting in better patient satisfaction and outcomes. Shared decision-making with patients can ensure the most effective and personalized approach, which can help build trust and establish a therapeutic relationship. This is particularly important for individuals who may be hesitant to seek treatment or have had negative experiences with mental health care. Understanding patients' needs can help clinicians provide information in a clear and concise language, offer visual aids, refer to reliable sources, and encourage questions. It is also important for clinicians to be cautious of the resources they rely on to avoid misinformation. By adopting these strategies, clinicians can improve the health literacy of patients, thus encouraging them to make informed decisions and take better care of themselves. We believe our findings could hold significant implications for mental health policy makers as well. By gaining insights into the needs of patients, policy makers can shape health care systems and policies to better support patient-centered care. Allocating resources toward addressing barriers, such as financial constraints or logistical difficulties, can enhance access to mental health services, ensuring that individuals receive timely and appropriate treatment. Furthermore, policies aimed at promoting education and awareness about mental health conditions can help reduce stigma and encourage individuals to seek treatment without hesitation.

Acknowledgments

This research was not supported by external funding.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Frequencies of all codes and quotations from users' posts on the forums for people with depression and anxiety disorders.
[DOCX File , 96 KB - jmir_v26i1e51506_app1.docx]

References

1. Global burden of disease study 2019 (GBD 2019) data resources. Institute for Health Metrics and Evaluation. URL: https://ghdx.healthdata.org/gbd-2019 [accessed 2024-06-27]
2. Bains N, Abdijadid S. Major depressive disorder. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2023.
3. Wulsin LR, Vaillant GE, Wells VE. A systematic review of the mortality of depression. Psychosom Med 1999;61(1):6-17. [doi: 10.1097/00006842-199901000-00003] [Medline: 10024062]
4. Markiewicz-Gospodarek A, Górska A, Markiewicz R, Chilimoniuk Z, Czeczelewski M, Baj J, et al. The relationship between mental disorders and the COVID-19 pandemic-course, risk factors, and potential consequences. Int J Environ Res Public Health 2022 Aug 04;19(15):9573 [FREE Full text] [doi: 10.3390/ijerph19159573] [Medline: 35954930]
5. Alon-Tirosh M, Hadar-Shoval D, Asraf K, Tannous-Haddad L, Tzischinsky O. The association between lifestyle changes and psychological distress during COVID-19 lockdown: the moderating role of COVID-related stressors. Int J Environ Res Public Health 2021 Sep 15;18(18):9695 [FREE Full text] [doi: 10.3390/ijerph18189695] [Medline: 34574620]
6. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
7. COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. World Health Organization. 2022 Mar 02. URL: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide [accessed 2024-06-27]
8. Demand for mental health treatment continues to increase, say psychologists. American Psychological Association. 2021 Oct 19. URL: https://www.apa.org/news/press/releases/2021/10/mental-health-treatment-demand [accessed 2024-06-27]
9. 40 proc. Polaków twierdzi, że w czasie pandemii pogorszyło się ich zdrowie psychiczne. UCE Group Ltd. 2022 Apr 5. URL: https://uce-pl.com/news/40-proc-polakow-twierdzi-ze-w-czasie-pandemii-pogorszylo-sie-ich-zdrowie-psychiczne [accessed 2024-06-27]
10. Barić H, Đorđević V, Cerovečki I, Trkulja V. Complementary and alternative medicine treatments for generalized anxiety disorder: systematic review and meta-analysis of randomized controlled trials. Adv Ther 2018 Mar;35(3):261-288. [doi: 10.1007/s12325-018-0680-6] [Medline: 29508154]

XSL•FO
RenderX

11. Berenzon-Gorn S, Saavedra-Solano N, Alanís-Navarro S. Estrategias utilizadas por un grupo de mujeres mexicanas para cuidar su salud emocional: autoatención y apoyo social. Salud Pública Méx 2009;51(6):465-473. [doi: 10.1590/s0036-36342009000600005]
12. Menéndez EL. Autoatención de los padecimientos y algunos imaginarios antropológicos. Desacatos 2018 Sep 05(58):104-113. [doi: 10.29340/58.1999]
13. MacDonald K, Fainman-Adelman N, Anderson KK, Iyer SN. Pathways to mental health services for young people: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2018 Oct;53(10):1005-1038 [FREE Full text] [doi: 10.1007/s00127-018-1578-y] [Medline: 30136192]
14. Lake J, Turner MS. Urgent need for improved mental health care and a more collaborative model of care. Perm J 2017;21:17-024 [FREE Full text] [doi: 10.7812/TPP/17-024] [Medline: 28898197]
15. Number of psychiatrists: how do countries compare? Eurostat. 2020 May 6. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20200506-1 [accessed 2024-06-22]
16. Liu C, Liu X, Wu F, Xie M, Feng Y, Hu C. Using artificial intelligence (Watson for oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. J Med Internet Res 2018 Sep 25;20(9):e11087 [FREE Full text] [doi: 10.2196/11087] [Medline: 30257820]
17. Waszak PM, Kasprzycka-Waszak W, Kubanek A. The spread of medical fake news in social media – the pilot quantitative study. Health Policy Technol 2018 Jun;7(2):115-118. [doi: 10.1016/j.hlpt.2018.03.002]
18. Communication in the Real World: An Introduction to Communication Studies. Minneapolis, MN: University of Minnesota Libraries Publishing; 2013.
19. Saya A, Brugnoli C, Piazzi G, Liberato D, Di Ciaccia G, Niolu C, et al. Criteria, procedures, and future prospects of involuntary treatment in psychiatry around the world: a narrative review. Front Psychiatry 2019 Apr 29;10:271 [FREE Full text] [doi: 10.3389/fpsyt.2019.00271] [Medline: 31110481]
20. Kallert TW, Glöckner M, Schützwohl M. Involuntary vs. voluntary hospital admission. A systematic literature review on outcome diversity. Eur Arch Psychiatry Clin Neurosci 2008 Jun;258(4):195-209. [doi: 10.1007/s00406-007-0777-4] [Medline: 18080170]
21. Yang Y, Li W, An FR, Wang YY, Ungvari GS, Balbuena L, et al. Voluntary and involuntary admissions for severe mental illness in China: a systematic review and meta-analysis. Psychiatr Serv 2020 Jan 01;71(1):83-86. [doi: 10.1176/appi.ps.201900106] [Medline: 31575350]
22. World Health Organization. Division of Mental Health and Prevention of Substance Abuse. Mental health care law : ten basic principles : with annotations suggesting selected actions to promote their implementation. World Health Organization. 1996. URL: https://iris.who.int/handle/10665/63624 [accessed 2023-04-15]
23. Perkes IE, Birmaher B, Hazell PL. Indications for psychiatric hospitalization of children and adolescents. Aust N Z J Psychiatry 2019 Aug;53(8):729-731. [doi: 10.1177/0004867419835930] [Medline: 30873847]
24. Morris NP, Kleinman RA. Taking an evidence-based approach to involuntary psychiatric hospitalization. Psychiatr Serv 2023 Apr 01;74(4):431-433. [doi: 10.1176/appi.ps.20220296] [Medline: 36128697]
25. Maina G, Rosso G, Carezana C, Mehanović E, Risso F, Villari V, et al. Factors associated with involuntary admissions: a register-based cross-sectional multicenter study. Ann Gen Psychiatry 2021 Jan 07;20(1):3 [FREE Full text] [doi: 10.1186/s12991-020-00323-1] [Medline: 33413499]
26. Zhang S, Mellsop G, Brink J, Wang X. Involuntary admission and treatment of patients with mental disorder. Neurosci Bull 2015 Feb;31(1):99-112 [FREE Full text] [doi: 10.1007/s12264-014-1493-5] [Medline: 25595369]
27. Carballedo A, Doyle M. Criteria for compulsory admission in some European countries. Int Psychiatry 2018 Jan 02;8(3):68-71. [doi: 10.1192/s1749367600002617]
28. Markiewicz I, Heitzman J, Gardyńska-Ziemba E. Involuntary psychiatric holds - the structure of admissions on the example of Institute of Psychiatry and Neurology in Warsaw. Psychiatr Pol 2016;50(1):7-18 [FREE Full text] [doi: 10.12740/PP/33336] [Medline: 27086324]
29. United Nations. Convention on the rights of persons with disabilities. Eur J Health Law 2007 Nov;14(3):281-298. [doi: 10.1515/9783110208856.203] [Medline: 18348362]
30. Convention on the rights of persons with disabilities. The Office of the High Commissioner for Human Rights, United Nations. URL: https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities [accessed 2024-04-02]
31. UN Committee is asking Poland how persons with disabilities are supported in their independent life. Speech by the Deputy Commissioner for Human Rights. Rzecznik Praw Obywatelskich. 2018 Sep 4. URL: https://bip.brpo.gov.pl/pl/content/un-committee-asking-poland-how-persons-disabilities-are-supported-their-independent-life-speech [accessed 2023-04-11]
32. Chebar D, Alba P, Alarcán H, Fassi G, Losanovsky Perel V, Mazaira S, et al. [Length of psychiatric hospitalizations: a clinical and socio-demographic characterization]. Vertex 2020 Apr;XXX(148):1-6. [doi: 10.53680/vertex.v30i148.124] [Medline: 33890930]
33. Masters GA, Baldessarini RJ, Öngür D, Centorrino F. Factors associated with length of psychiatric hospitalization. Compr Psychiatry 2014 Apr;55(3):681-687 [FREE Full text] [doi: 10.1016/j.comppsych.2013.11.004] [Medline: 24387922]

XSL•FO
RenderX

34. Lee S, Rothbard AB, Noll EL. Length of inpatient stay of persons with serious mental illness: effects of hospital and regional characteristics. Psychiatr Serv 2012 Sep 01;63(9):889-895. [doi: 10.1176/appi.ps.201100412] [Medline: 22751995]
35. Dimitri G, Giacco D, Bauer M, Bird VJ, Greenberg L, Lasalvia A, et al. Predictors of length of stay in psychiatric inpatient units: does their effect vary across countries? Eur Psychiatry 2018 Feb;48:6-12 [FREE Full text] [doi: 10.1016/j.eurpsy.2017.11.001] [Medline: 29331601]
36. Tulloch AD, Fearon P, David AS. Length of stay of general psychiatric inpatients in the United States: systematic review. Adm Policy Ment Health 2011 May;38(3):155-168. [doi: 10.1007/s10488-010-0310-3] [Medline: 20924662]
37. Length of hospital stay. Organization for Economic Co-operation and Development. URL: https://doi.org/10.1787/8dda6b7a-en [accessed 2023-04-15]
38. Schmidt LM, Hesse M, Lykke J. The impact of substance use disorders on the course of schizophrenia--a 15-year follow-up study: dual diagnosis over 15 years. Schizophr Res 2011 Aug;130(1-3):228-233. [doi: 10.1016/j.schres.2011.04.011] [Medline: 21592731]
39. Cheng P, Wang L, Xu L, Zhou Y, Zhang L, Li W. Factors related to the length of stay for patients with schizophrenia: a retrospective study. Front Psychiatry 2021 Jan 24;12:818254 [FREE Full text] [doi: 10.3389/fpsyt.2021.818254] [Medline: 35140640]
40. Osby U, Tiainen A, Backlund L, Edman G, Adler M, Hällgren J, et al. Psychiatric admissions and hospitalization costs in bipolar disorder in Sweden. J Affect Disord 2009 Jun;115(3):315-322 [FREE Full text] [doi: 10.1016/j.jad.2008.09.011] [Medline: 18930322]
41. Walker S, Mackay E, Barnett P, Sheridan Rains L, Leverton M, Dalton-Locke C, et al. Clinical and social factors associated with increased risk for involuntary psychiatric hospitalisation: a systematic review, meta-analysis, and narrative synthesis. Lancet Psychiatry 2019 Dec;6(12):1039-1053 [FREE Full text] [doi: 10.1016/S2215-0366(19)30406-7] [Medline: 31777340]
42. Fakhoury W, Priebe S. Deinstitutionalization and reinstitutionalization: major changes in the provision of mental healthcare. Psychiatry 2007 Aug;6(8):313-316. [doi: 10.1016/j.mppsy.2007.05.008]
43. Morzycka-Markowska M, Drozdowicz E, Nasierowski T. [Deinstitutionalization in Italian psychiatry - the course and consequences Part II. The consequences of deinstitutionalization]. Psychiatr Pol 2015 Apr 30;49(2):403-412 [FREE Full text] [doi: 10.12740/PP/28614] [Medline: 26093602]
44. Köhler S, Roth-Sackenheim C, Navarro Ureña A, Mönter N. [Innovative outpatient models]. Nervenarzt 2020 Nov;91(11):1003-1016. [doi: 10.1007/s00115-020-01010-y] [Medline: 33044578]
45. Deng M, Zhai S, Ouyang X, Liu Z, Ross B. Factors influencing medication adherence among patients with severe mental disorders from the perspective of mental health professionals. BMC Psychiatry 2022 Jan 07;22(1):22 [FREE Full text] [doi: 10.1186/s12888-021-03681-6] [Medline: 34996394]
46. Phan SV. Medication adherence in patients with schizophrenia. Int J Psychiatry Med 2016;51(2):211-219. [doi: 10.1177/0091217416636601] [Medline: 27079779]
47. Jones J, Nolan P, Bowers L, Simpson A, Whittington R, Hackney D, et al. Psychiatric wards: places of safety? J Psychiatr Ment Health Nurs 2010 Mar;17(2):124-130 [FREE Full text] [doi: 10.1111/j.1365-2850.2009.01482.x] [Medline: 20465757]
48. Agoritsas T, Bovier PA, Perneger TV. Patient reports of undesirable events during hospitalization. J Gen Intern Med 2005 Oct;20(10):922-928 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.0225.x] [Medline: 16191139]
49. The dark side of mental hospitals: abuse neglect and mistreatment. Excel Medical. 2022 Oct 6. URL: https://web.archive.org/web/20221130140534/www.excel-medical.com/the-dark-side-of-mental-hospitals-abuse-neglect-and-mistreatment/ [accessed 2024-07-01]
50. Panorama Team, Lee J. 'Toxic culture' of abuse at mental health hospital revealed by BBC secret filming. BBC News. 2022 Sep 28. URL: https://www.bbc.com/news/uk-63045298 [accessed 2023-04-11]
51. Top 10 forms of psychiatric institution abuse. Citizens Commission on Human Rights. 2014 Nov 12. URL: https://www.cchrint.org/2014/11/13/top-10-forms-of-psychiatric-institution-abuse/ [accessed 2023-04-11]
52. Wahl OF. Media Madness: Public Images of Mental Illness. New Brunswick, NJ: Rutgers University Press; 1995.
53. Goepfert NC, Conrad von Heydendorff S, Dreßing H, Bailer J. Effects of stigmatizing media coverage on stigma measures, self-esteem, and affectivity in persons with depression - an experimental controlled trial. BMC Psychiatry 2019 May 07;19(1):138 [FREE Full text] [doi: 10.1186/s12888-019-2123-6] [Medline: 31064374]
54. McCrae N, Sharif L, Norman I. Media portrayals of mental disorder in Saudi Arabia: a review of popular newspapers. Transcult Psychiatry 2019 Apr;56(2):428-442 [FREE Full text] [doi: 10.1177/1363461518819117] [Medline: 30632948]
55. Parrott S. Media stereotypes of mental illness: nurturing and mitigating stigma. In: Billings AC, Parrott S, editors. Media Stereotypes: From Ageism to Xenophobia. New York, NY: Peter Lang Publishing, Inc; 2020.
56. Alzahrani N. The effect of hospitalization on patients' emotional and psychological well-being among adult patients: an integrative review. Appl Nurs Res 2021 Oct;61:151488. [doi: 10.1016/j.apnr.2021.151488] [Medline: 34544571]
57. Hermann DM, Muck S, Nehen HG. Supporting dementia patients in hospital environments: health-related risks, needs and dedicated structures for patient care. Eur J Neurol 2015 Feb;22(2):239-45, e17. [doi: 10.1111/ene.12530] [Medline: 25103994]

XSL•FO
RenderX

58. Afilalo M, Soucy N, Xue X, Colacone A, Jourdenais E, Boivin JF. Characteristics and needs of psychiatric patients with prolonged hospital stay. Can J Psychiatry 2015 Apr;60(4):181-188 [FREE Full text] [doi: 10.1177/070674371506000405] [Medline: 26174218]
59. Røsvik J, Rokstad AM. What are the needs of people with dementia in acute hospital settings, and what interventions are made to meet these needs? A systematic integrative review of the literature. BMC Health Serv Res 2020 Aug 07;20(1):723 [FREE Full text] [doi: 10.1186/s12913-020-05618-3] [Medline: 32767987]
60. Lauri S, Lepistö M, Käppeli S. Patients' needs in hospital: nurses' and patients' views. J Adv Nurs 1997 Feb;25(2):339-346. [doi: 10.1046/j.1365-2648.1997.1997025339.x] [Medline: 9044009]
61. Berg SH, Rørtveit K, Aase K. Suicidal patients' experiences regarding their safety during psychiatric in-patient care: a systematic review of qualitative studies. BMC Health Serv Res 2017 Jan 23;17(1):73 [FREE Full text] [doi: 10.1186/s12913-017-2023-8] [Medline: 28114936]
62. van Lankeren JE, Testerink AE, Daggenvoorde TH, Poslawsky IE, Goossens PJ. Patient experiences with nursing care during hospitalization on a closed ward due to a manic episode: a qualitative study. Perspect Psychiatr Care 2020 Jan;56(1):37-45. [doi: 10.1111/ppc.12370] [Medline: 30820964]
63. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
64. Witamy w naszej społeczności!. Depresja.ws. URL: https://www.depresja.ws/ [accessed 2024-05-28]
65. Forum. Nerwica. URL: https://www.nerwica.com/ [accessed 2024-05-28]
66. Storman D, Jemioło P, Swierz MJ, Sawiec Z, Antonowicz E, Prokop-Dorner A, et al. Meeting the unmet needs of individuals with mental disorders: scoping review on peer-to-peer web-based interactions. JMIR Ment Health 2022 Dec 05;9(12):e36056 [FREE Full text] [doi: 10.2196/36056] [Medline: 36469366]
67. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: SAGE Publications; Oct 1980.
68. Kuckartz U, Rädiker S. Analyzing Qualitative Data with MAXQDA: Text, Audio, and Video. Berlin, Heidelberg: Springer International Publishing; 2019.
69. MAXQDA 2022 manual. VERBI Software. URL: https://www.maxqda.com/help-mx22/welcome [accessed 2024-06-27]
70. IBM SPSS statistics for windows, version 22.0. IBM Corporation. 2013. URL: https://www.ibm.com/support/pages/spss-statistics-220-available-download [accessed 2024-06-27]
71. Krysinska K, Lester D. Post-traumatic stress disorder and suicide risk: a systematic review. Arch Suicide Res 2010;14(1):1-23. [doi: 10.1080/13811110903478997] [Medline: 20112140]
72. Hewson C, Buchanan T. Ethics guidelines for internet-mediated research. The British Psychological Society. 2013. URL: https://oro.open.ac.uk/42132/ [accessed 2024-06-27]
73. Williams AJ, Nielsen E, Coulson NS. "They aren't all like that": perceptions of clinical services, as told by self-harm online communities. J Health Psychol 2020;25(13-14):2164-2177 [FREE Full text] [doi: 10.1177/1359105318788403] [Medline: 30024273]
74. Attachment B: considerations and recommendations concerning internet research and human subjects research regulations, with revisions. U.S. Department of Health and Human Services. 2013. URL: https://www.hhs.gov/ohrp/sachrp-committee/recommendations/considerations-for-research-with-the-internet/index.html [accessed 2024-06-23]
75. Hutchins N, Allen A, Curran M, Kannis-Dymand L. Social anxiety and online social interaction. Aust Psychol 2021 Mar 11;56(2):142-153. [doi: 10.1080/00050067.2021.1890977]
76. van der Velden PG, Contino C, Das M, van Loon P, Bosmans MW. Anxiety and depression symptoms, and lack of emotional support among the general population before and during the COVID-19 pandemic. A prospective national study on prevalence and risk factors. J Affect Disord 2020 Dec 01;277:540-548 [FREE Full text] [doi: 10.1016/j.jad.2020.08.026] [Medline: 32889378]
77. Simning A, Seplaki CL, Conwell Y. The association of an inability to form and maintain close relationships due to a medical condition with anxiety and depressive disorders. J Affect Disord 2016 Mar 15;193:130-136 [FREE Full text] [doi: 10.1016/j.jad.2015.12.079] [Medline: 26773918]
78. Porter E, Chambless DL. Social anxiety and social support in romantic relationships. Behav Ther 2017 May;48(3):335-348. [doi: 10.1016/j.beth.2016.12.002] [Medline: 28390497]
79. Hofheinz C, Germar M, Schultze T, Michalak J, Mojzisch A. Are depressed people more or less susceptible to informational social influence? Cogn Ther Res 2017 Apr 8;41(5):699-711. [doi: 10.1007/s10608-017-9848-7]
80. Depression. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/depression [accessed 2023-04-01]
81. Dar KA, Iqbal N. Worry and rumination in generalized anxiety disorder and obsessive compulsive disorder. J Psychol 2015;149(8):866-880. [doi: 10.1080/00223980.2014.986430] [Medline: 25495066]
82. Johnson EM, Coles ME. Failure and delay in treatment-seeking across anxiety disorders. Community Ment Health J 2013 Dec;49(6):668-674. [doi: 10.1007/s10597-012-9543-9] [Medline: 23054147]

XSL•FO
RenderX

83. Pryce CR, Azzinnari D, Spinelli S, Seifritz E, Tegethoff M, Meinlschmidt G. Helplessness: a systematic translational review of theory and evidence for its relevance to understanding and treating depression. Pharmacol Ther 2011 Dec;132(3):242-267. [doi: 10.1016/j.pharmthera.2011.06.006] [Medline: 21835197]
84. Liu RT, Kleiman EM, Nestor BA, Cheek SM. The hopelessness theory of depression: a quarter‐century in review. Clin Psychol Sci Pract 2015;22(4):345-365. [doi: 10.1037/h0101732]
85. Gariépy G, Honkaniemi H, Quesnel-Vallée A. Social support and protection from depression: systematic review of current findings in Western countries. Br J Psychiatry 2016 Oct;209(4):284-293. [doi: 10.1192/bjp.bp.115.169094] [Medline: 27445355]
86. Wolpert L. Stigma of depression--a personal view. Br Med Bull 2001;57:221-224. [doi: 10.1093/bmb/57.1.221] [Medline: 11719920]
87. Marroquín B. Interpersonal emotion regulation as a mechanism of social support in depression. Clin Psychol Rev 2011 Dec;31(8):1276-1290. [doi: 10.1016/j.cpr.2011.09.005] [Medline: 21983267]
88. Søvold LE, Naslund JA, Kousoulis AA, Saxena S, Qoronfleh MW, Grobler C, et al. Prioritizing the mental health and well-being of healthcare workers: an urgent global public health priority. Front Public Health 2021 May 7;9:679397 [FREE Full text] [doi: 10.3389/fpubh.2021.679397] [Medline: 34026720]
89. Billings J, Biggs C, Ching BC, Gkofa V, Singleton D, Bloomfield M, et al. Experiences of mental health professionals supporting front-line health and social care workers during COVID-19: qualitative study. BJPsych Open 2021 Mar 23;7(2):e70 [FREE Full text] [doi: 10.1192/bjo.2021.29] [Medline: 33752774]
90. Seera G, Arya S, Sethi S, Nimmawitt N, Ratta-Apha W. Help-seeking behaviors for mental health problems in medical students: studies in Thailand and India. Asian J Psychiatr 2020 Dec;54:102453. [doi: 10.1016/j.ajp.2020.102453] [Medline: 33271732]
91. Akhther N, Sopory P. Seeking and sharing mental health information on social media during COVID-19: role of depression and anxiety, peer support, and health benefits. J Technol Behav Sci 2022;7(2):211-226 [FREE Full text] [doi: 10.1007/s41347-021-00239-x] [Medline: 35036521]
92. Rodrigues NC, Ham E, Kirsh B, Seto MC, Hilton NZ. Mental health workers' experiences of support and help-seeking following workplace violence: a qualitative study. Nurs Health Sci 2021 Jun;23(2):381-388. [doi: 10.1111/nhs.12816] [Medline: 33496379]
93. Naslund JA, Bondre A, Torous J, Aschbrenner KA. Social media and mental health: benefits, risks, and opportunities for research and practice. J Technol Behav Sci 2020 Sep;5(3):245-257. [doi: 10.1007/s41347-020-00134-x] [Medline: 33415185]
94. Wall A, Lovheden T, Landgren K, Stjernswärd S. Experiences and challenges in the role as peer support workers in a Swedish mental health context - an interview study. Issues Ment Health Nurs 2022 Apr;43(4):344-355. [doi: 10.1080/01612840.2021.1978596] [Medline: 34569894]
95. Bird V, Miglietta E, Giacco D, Bauer M, Greenberg L, Lorant V, et al. Factors associated with satisfaction of inpatient psychiatric care: a cross country comparison. Psychol Med 2019 Jan 30;50(2):284-292. [doi: 10.1017/s0033291719000011]
96. Chieze M, Clavien C, Kaiser S, Hurst S. Coercive measures in psychiatry: a review of ethical arguments. Front Psychiatry 2021 Dec 14;12:790886 [FREE Full text] [doi: 10.3389/fpsyt.2021.790886] [Medline: 34970171]
97. Pelto-Piri V, Wallsten T, Hylén U, Nikban I, Kjellin L. Feeling safe or unsafe in psychiatric inpatient care, a hospital-based qualitative interview study with inpatients in Sweden. Int J Ment Health Syst 2019 Apr 08;13:23 [FREE Full text] [doi: 10.1186/s13033-019-0282-y] [Medline: 30996733]
98. Gilburt H, Rose D, Slade M. The importance of relationships in mental health care: a qualitative study of service users' experiences of psychiatric hospital admission in the UK. BMC Health Serv Res 2008 Apr 25;8:92 [FREE Full text] [doi: 10.1186/1472-6963-8-92] [Medline: 18439254]
99. Woodward S, Berry K, Bucci S. A systematic review of factors associated with service user satisfaction with psychiatric inpatient services. J Psychiatr Res 2017 Sep;92:81-93. [doi: 10.1016/j.jpsychires.2017.03.020] [Medline: 28412601]
100. Aguilera-Serrano C, Guzman-Parra J, Garcia-Sanchez JA, Moreno-Küstner B, Mayoral-Cleries F. Variables associated with the subjective experience of coercive measures in psychiatric inpatients: a systematic review. Can J Psychiatry 2018 Feb;63(2):129-144 [FREE Full text] [doi: 10.1177/0706743717738491] [Medline: 29069981]
101. Rössler W. The stigma of mental disorders: a millennia-long history of social exclusion and prejudices. EMBO Rep 2016 Sep;17(9):1250-1253 [FREE Full text] [doi: 10.15252/embr.201643041] [Medline: 27470237]
102. Stuart H. Media portrayal of mental illness and its treatments: what effect does it have on people with mental illness? CNS Drugs 2006;20(2):99-106. [doi: 10.2165/00023210-200620020-00002] [Medline: 16478286]
103. Klin A, Lemish D. Mental disorders stigma in the media: review of studies on production, content, and influences. J Health Commun 2008;13(5):434-449. [doi: 10.1080/10810730802198813] [Medline: 18661386]
104. van Voren R. Political abuse of psychiatry--an historical overview. Schizophr Bull 2010 Jan;36(1):33-35 [FREE Full text] [doi: 10.1093/schbul/sbp119] [Medline: 19892821]
105. Is private investment the answer to the crisis in psychiatric care in Poland? Healthcare Market Experts. URL: https://healthcaremarketexperts.com/en/news/is-private-investment-the-answer-to-the-crisis-in-psychiatric-care-in-poland/ [accessed 2023-04-15]

XSL•FO
RenderX

106. Mahomed F. Addressing the problem of severe underinvestment in mental health and well-being from a human rights perspective. Health Hum Rights 2020 Jun;22(1):35-49 [FREE Full text] [Medline: 32669787]
107. Sagan A, Kowalska-Bobko I, Biechowska D, Rogala M, Gałązka-Sobotka M. Implementation of mental health centres pilots in Poland since 2018: a chance to move towards community-based mental health services. Int J Environ Res Public Health 2022 May 09;19(9):5774 [FREE Full text] [doi: 10.3390/ijerph19095774] [Medline: 35565173]
108. Montgomery P, Kirkpatrick H. Understanding those who seek frequent psychiatric hospitalizations. Arch Psychiatr Nurs 2002 Feb;16(1):16-24. [doi: 10.1053/apnu.2002.30494] [Medline: 11877602]
109. Owusu E, Oluwasina F, Nkire N, Lawal MA, Agyapong VI. Readmission of patients to acute psychiatric hospitals: influential factors and interventions to reduce psychiatric readmission rates. Healthcare (Basel) 2022 Sep 19;10(9):1808 [FREE Full text] [doi: 10.3390/healthcare10091808] [Medline: 36141418]
110. Kalseth J, Lassemo E, Wahlbeck K, Haaramo P, Magnussen J. Psychiatric readmissions and their association with environmental and health system characteristics: a systematic review of the literature. BMC Psychiatry 2016 Nov 07;16(1):376 [FREE Full text] [doi: 10.1186/s12888-016-1099-8] [Medline: 27821155]
111. Staniszewska S, Mockford C, Chadburn G, Fenton SJ, Bhui K, Larkin M, et al. Experiences of in-patient mental health services: systematic review. Br J Psychiatry 2019 Jun;214(6):329-338 [FREE Full text] [doi: 10.1192/bjp.2019.22] [Medline: 30894243]
112. Modini M, Burton A, Abbott MJ. Factors influencing inpatients perception of psychiatric hospitals: a meta-review of the literature. J Psychiatr Res 2021 Apr;136:492-500. [doi: 10.1016/j.jpsychires.2020.10.020] [Medline: 33127069]
113. Katsakou C, Priebe S. Outcomes of involuntary hospital admission--a review. Acta Psychiatr Scand 2006 Oct;114(4):232-241. [doi: 10.1111/j.1600-0447.2006.00823.x] [Medline: 16968360]
114. Volpe U, Mihai A, Jordanova V, Sartorius N. The pathways to mental healthcare worldwide: a systematic review. Curr Opin Psychiatry 2015 Jul;28(4):299-306 [FREE Full text] [doi: 10.1097/YCO.0000000000000164] [Medline: 26001921]
115. Mutschler C, Lichtenstein S, Kidd SA, Davidson L. Transition experiences following psychiatric hospitalization: a systematic review of the literature. Community Ment Health J 2019 Nov;55(8):1255-1274. [doi: 10.1007/s10597-019-00413-9] [Medline: 31104176]
116. Teshager S, Kerebih H, Hailesilassie H, Abera M. Pathways to psychiatric care and factors associated with delayed help-seeking among patients with mental illness in Northern Ethiopia: a cross-sectional study. BMJ Open 2020 Jul 26;10(7):e033928 [FREE Full text] [doi: 10.1136/bmjopen-2019-033928] [Medline: 32713844]
117. Nyblade L, Stockton MA, Giger K, Bond V, Ekstrand ML, Lean RM, et al. Stigma in health facilities: why it matters and how we can change it. BMC Med 2019 Feb 15;17(1):25 [FREE Full text] [doi: 10.1186/s12916-019-1256-2] [Medline: 30764806]
118. Mueser KT, Corrigan PW, Hilton DW, Tanzman B, Schaub A, Gingerich S, et al. Illness management and recovery: a review of the research. Psychiatr Serv 2002 Oct;53(10):1272-1284. [doi: 10.1176/appi.ps.53.10.1272] [Medline: 12364675]
119. Mueser KT, Meyer PS, Penn DL, Clancy R, Clancy DM, Salyers MP. The illness management and recovery program: rationale, development, and preliminary findings. Schizophr Bull 2006 Oct;32 Suppl 1(Suppl 1):S32-S43 [FREE Full text] [doi: 10.1093/schbul/sbl022] [Medline: 16899534]
120. Åkerblom KB, Ness O. Peer workers in co-production and co-creation in mental health and substance use services: a scoping review. Adm Policy Ment Health 2023 Mar;50(2):296-316 [FREE Full text] [doi: 10.1007/s10488-022-01242-x] [Medline: 36396756]

Edited by K Williams; submitted 02.08.23; peer-reviewed by G Mayer, K Lee; comments to author 09.03.24; revised version received 03.04.24; accepted 06.05.24; published 12.07.24.

Please cite as:
Storman D, Jemioło P, Sawiec Z, Swierz MJ, Antonowicz E, Bala MM, Prokop-Dorner A
Needs Expressed in Peer-to-Peer Web-Based Interactions Among People With Depression and Anxiety Disorders Hospitalized in a Mental Health Facility: Mixed Methods Study
J Med Internet Res 2024;26:e51506
URL: https://www.jmir.org/2024/1/e51506
doi:10.2196/51506
PMID:

©Dawid Storman, Paweł Jemioło, Zuzanna Sawiec, Mateusz Jan Swierz, Ewa Antonowicz, Malgorzata M Bala, Anna Prokop-Dorner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Success Factors of Growth-Stage Digital Health Companies: Systematic Literature Review

Estelle Pfitzer[1]; Laura Bitomsky[1]; Marcia Nißen[1]; Christoph Kausch[2]; Tobias Kowatsch[1,3,4]

[1]Centre for Digital Health Interventions, School of Medicine, University of St. Gallen, St. Gallen, Switzerland

[2]MTIP AG, Basel, Switzerland

[3]Centre for Digital Health Interventions, Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[4]Centre for Digital Health Interventions, Department of Management, Technology, and Economics, ETH Zürich, Zurich, Switzerland

Corresponding Author:
Tobias Kowatsch
Centre for Digital Health Interventions
Institute for Implementation Science in Health Care
University of Zurich
Universitätstr. 84
Zurich, 8006
Switzerland
Phone: 41 71 224 72 44
Email: tobias.kowatsch@uzh.ch

Abstract

Background: Over the past decade, digital health technologies (DHTs) have grown rapidly, driven by innovations such as electronic health records and accelerated by the COVID-19 pandemic. Increased funding and regulatory support have further pushed the sector's expansion. Despite early success, many DHT companies struggle to scale, with notable examples like Pear Therapeutics and Proteus Digital Health, which both declared bankruptcy after initial breakthroughs. These cases highlight the challenges of sustaining growth in a highly regulated health care environment. While there is research on success factors across industries, a gap remains in understanding the specific challenges faced by growth-stage DHT companies.

Objective: This study aims to identify and discuss key factors that make growth-stage DHT companies successful. Specifically, we address three questions: (1) What are the success factors of growth-stage digital companies in general and (2) digital health companies in particular? (3) How do these success factors vary across DHTs?

Methods: Following established PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a systematic literature review was conducted to answer the questions. A comprehensive literature search was conducted using management and medical literature databases: EBSCO, ProQuest, PubMed, Scopus, and Web of Science. The review spanned scientific articles published from 2000 to 2023, using a rigorous screening process and quality assessment using the Critical Appraisal Skills Programme (CASP) checklist.

Results: From the 2972 studies initially screened, 36 were selected, revealing 52 success factors. We categorized them into internal factor categories (Product and Services, Operations, Business Models, and Team Composition) and external factor categories (Customers, Health Care System, Government and Regulators, Investors and Shareholders, Suppliers and Partners, and Competitors). Of the 52 factors, 19 were specific to DHT companies. The most frequently cited internal success factors included financial viability (n=18) and market demand and relevance of the product and service (n=13). External success factors emphasized the regulatory environment and policy framework (n=15). Key differences were observed between DHTs and broader digital companies in areas such as data security (P=.03), system interoperability (P=.01), and regulatory alignment (P=.02), with DHTs showing a higher frequency of these factors. In addition, success factors varied across different DHT categories. Health System Operational Software companies emphasized affordability and system integration, while Digital Therapeutics prioritized factors related to government regulations and regulatory approval.

Conclusions: Essential characteristics contributing to the success of growth-stage digital health companies have been identified. This work, therefore, fills a knowledge gap and provides relevant stakeholders, including investors and entrepreneurs, with a valuable resource that can support informed decision-making in investment decisions and, in turn, enhance the success of fast-growing digital health companies. In addition, it provides the research community with a direction for future studies, enhancing the understanding, implementation, and growth of DHTs.

XSL•FO
RenderX

International Registered Report Identifier (IRRID): RR2-10.1101/2024.05.06.24306674

(J Med Internet Res 2024;26:e60473) doi:10.2196/60473

KEYWORDS

digital health; health information technology; success factors; systematic literature review; growth-stage companies; clinical softwares; EHR; electronic health records; health companies; funding; digital therapeutics; substance use disorder; stakeholders

Introduction

Digital health technologies (DHTs), including health system clinical software like electronic health records, have seen remarkable growth in the past decade. Their adoption in US hospitals surged from 7% in 2009 to over 81% by 2019 [1]. The COVID-19 pandemic further accelerated the adoption of DHTs, pushing for considerable innovation and prompting supportive regulatory changes [2]. At the same time, there was a surge in investor interest, with funding for digital health companies increasing more than tenfold [3,4], driving the development of solutions promising to revolutionize modern medicine by making health care more accessible, efficient, and cost-effective [5,6].

Yet, within the dynamic realm of digital health, numerous companies struggle to maintain their initial momentum, failing to progress beyond early-stage adopters despite being initially hailed as industry unicorns. One example includes Pear Therapeutics Inc., a once pioneering firm that gained US Food and Drug Administration (FDA) approval for its standalone digital therapeutics (DTx) solution, reSET, in September 2017. This prescription DTx, which aimed at addressing substance use disorder in tandem with outpatient therapy [7], helped Pear Therapeutics secure substantial funding, pushing its valuation to over US $1.5 billion [8] and earning a position as a leader among DTx companies [9]. Nevertheless, in 2023, Pear Therapeutics Inc declared bankruptcy, providing a cautionary tale for the sector at large. Similarly, Proteus Digital Health Inc.'s bankruptcy in 2020 sent ripples through the industry, particularly after achieving the landmark development of the first FDA-approved "smart pill" that could monitor medication intake and assess its effectiveness. Despite its groundbreaking technology earning a valuation of US $1.5 billion and the proof of concept supported by studies and clinical trials [10], Proteus struggled to scale up and secure a significant market presence. The lesson here is clear: the most revolutionary DHTs cannot significantly impact people if they fail to align with market dynamics and stakeholders' priorities.

Previous research has examined success factors via literature reviews across a range of industries, including agile transformations [11], product innovation [12,13], and IT companies [14]. However, a notable gap remains concerning growth-stage digital companies, particularly within the digital health sector, with a particular absence of comprehensive literature reviews.

The "fail fast, fail often" philosophy common among technology companies is often impeded by the complex regulatory environment of the health care industry. This cultural clash is further intensified by the deliberate, step-by-step, and lengthy process of health care innovation, which is rooted in the risk-averse clinical principle of "first, not harm" [6,15]. This highlights the need for a study specific to digital health. A thorough analysis of success factors would not only guide investors and entrepreneurs toward informed decision-making in investments and strategy but also enable the digital health research community to develop a theoretical framework of these factors. This unique contribution, in turn, could pave the way for DHTs to have a higher likelihood of success and a broader implementation within the health care system [16,17].

Our study, therefore, seeks to address this gap by examining success factors of growth-stage digital companies in general, with a particular focus on companies offering DHTs. Growth-stage refers to businesses beyond the startup or existence phase and now focusing on expansion and dealing with the complexities of scaling and establishing a competitive market presence [18]. In addition, digital refers to companies that use technologies as their main product or service. Instead of confining ourselves to a single definition of success, we explored a range of interpretations across various studies, highlighting the diversity in understanding success. Building on prior research [14,19-24], we will identify factors that are directly related to companies' success. To this end, we formulate the following research questions (RQs):

RQ1: What are the general success factors of growth-stage digital companies?

RQ2: What are the specific success factors of growth-stage digital health companies?

RQ3: How do digital health companies' success factors differ across different DHT categories?

Methods

Overview

To answer our RQs, we undertook a systematic literature review. The study was preregistered on the Open Science Framework Registry on August 21, 2023, and the protocol was published on medRxiv [25]. We followed Snyder and Tranfield's established methodological guidelines [26,27] to create evidence-informed management knowledge, incorporating appropriate elements from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Multimedia Appendix 1) [28].

Eligibility Criteria

Table 1 specifies our inclusion and exclusion criteria, focusing on studies from 2000 to 2023 to capture the significant shifts post-Dot-Com Bubble—a time marking the entry of numerous digital firms into the market [29].

XSL•FO
RenderX

Table 1. Inclusion and exclusion criteria for study selection.

Criteria	Inclusion criteria	Exclusion criteria
Research focus	Studies that identify the success factors in digital growth-stage companies	Studies that do not show research methodology, analysis, or discussion. Research on companies without a digital or tech-enabled component.
Publication type	Peer-reviewed academic literature	Gray literature (eg, news articles, company publications, annual reports, studies by nongovernmental organizations, presentations, catalogs)
Year	Only studies from 2000 and onward are considered	Studies published before 2000
Language	Only English studies are considered	Studies published in languages other than English

Information Sources and Search Strategy

To conduct a thorough search, we used the following databases: (1) EBSCO and ProQuest: renowned and widely used in business research, these databases provide a rich collection of business-related journals, articles, and reports; (2) PubMed: specializing in biomedical and life sciences research, PubMed offers valuable insights at the intersection of health care and business; And (3) Scopus and Web of Science: these multidisciplinary databases cover a broad array of subject areas, ensuring extensive coverage of our topic.

By exploring these databases, we will collect a diverse range of sources covering business-specific research, health care-business intersections, and interdisciplinary perspectives, ensuring a comprehensive search for our study.

To identify studies that focused on success factors in growth-stage digital companies, we used specific search terms centered around the keywords “success factor” and its synonyms indicating high performance, “growth-stage” to ensure we targeted studies on companies with an established product or service, “companies,” and “digital.” For RQ2 and RQ3, we added “digital health” to uncover success factors specific to the DHT sector. We used these search terms to navigate the databases mentioned above, focusing on titles, abstracts, and keywords. By including synonyms, abbreviations, alternative spellings, and related topics, we ensured a thorough search (Multimedia Appendix 2). This targeted approach helped us to identify relevant literature [27], ensuring a comprehensive and systematic collection of studies for our review.

Study Records

Data Management

Following comprehensive database searches, we imported all citations into Mendeley (version 2.91.0) for efficient management. Mendeley served as our secure repository for all data and documents, preserving data integrity and supporting collaborative efforts.

Selection Process

Co-authors EP and LB independently reviewed titles and abstracts using Rayyan to assess eligibility, with EP retrieving full texts for closer evaluation. To mitigate selection bias, paired authors examined these texts, resolving any differences through discussion or consultation with the co-author TK. The selection was meticulously recorded, including a PRISMA flow diagram to detail exclusions. Publications selected for RQ1, which pertain to digital health, informed RQ2 and RQ3. As described above, this secondary selection process was conducted systematically and involved 2 independent reviewers.

Data Items and Synthesis

To ensure uniform and precise data extraction, we designed a detailed template to record key characteristics of each study, such as the first author, publication year, methodological details, and definitions of success.

For data analysis, the studies were saved into PDFs to facilitate line-by-line coding. This process enabled us to identify and systematically categorize key data points and recurring themes. Following parts of the active categorization framework of Grodal et al [30], we refined our initially identified categories by dropping, merging, and splitting categories. This iterative process led to the creation of a dynamic list of success factors, which was continuously updated and refined through repeated coding processes. These refined categories were then synthesized into higher-order constructs, forming second-order categories.

To enhance our analysis of success factors, we segmented them into internal and external factors, also referred to as third-order categories, thereby differentiating between elements within a company’s control and those outside its direct sphere of influence but are nonetheless essential for strategic planning. Our rationale for distinguishing between internal and external factors comes from foundational management theory [31,32], which suggests that an organization’s success and sustained competitive advantage are influenced by both its internal capabilities and its environment (ie, external factors).

For RQ3, which aims to identify DHT category–specific success factors, we used the DTx Alliance DHT ecosystem. This ecosystem contains the non–patient-facing Health System Operational Software—the software supporting nonclinical systems like operations and finance—and the Health System Clinical Software, which provides clinical support for managing patient care. In addition, patient-facing solutions include Health and Wellness, Patient Monitoring, Care Support, Digital Diagnostics, and Digital Therapeutics. The categorization of DHT-related papers into these specific categories was carried out by two independent authors who resolved any discrepancies through discussion.

Risk of Bias and Quality Assessment

The Critical Appraisal Skills Programme (CASP) checklist for qualitative research was used to assess the quality of our selected qualitative studies, aiming to discern and evaluate any methodological limitations impacting the research outcomes

XSL•FO
RenderX

[33]. Endorsed by reputable organizations like Cochrane and the World Health Organization, CASP's effectiveness in qualitative evidence synthesis is well-recognized [33-35]. This tool methodically probes a study's strengths and weaknesses across 10 specific questions, addressing everything from the clarity of objectives to the study's overall impact and contribution. Using the CASP framework, responses were categorized as "Yes," "Can't tell," or "No," with studies requiring at least a 70% "Yes" rate to be included in our analysis.

Results

Study Selection

Our comprehensive review identified 36 relevant studies from an initial pool of 2972 unduplicated studies. During the initial screening phase, 2721 articles were excluded based on our predefined inclusion and exclusion criteria. We then conducted a detailed screening of the abstracts for the remaining 251 publications, resulting in the further exclusion of 197 articles that did not meet our inclusion and exclusion criteria (Figure 1). This selection followed the assessment of 54 full-text papers for their eligibility, out of which 15 were excluded due to misaligned focus: 6 studies examined non–digital-related companies, while five concentrated on the internal scaling of software projects rather than the broader scope of growing and operating digital enterprises. Predominantly, the studies used qualitative research methods (25/36, 69%), including interviews (8/36, 22.2%) and case studies (10/36, 27.7%). Among the quantitative studies included (11/36, 30.6%), surveys were the primary method (8/36, 22.2%) used to investigate the success factors of growth-stage digital companies.

Of the 36 studies included, 15 are related to DHT (15/36, 41.7%), while the remaining studies describe success factors of companies in other digital sectors such as e-commerce [36-38], e-media [39], and digital companies without a specific sector focus [40-44]. We extracted between 2 and 20 success factors per paper, with a median count of 7 (IQR 5-9.25). Table 2 presents a summary of selected studies, including the publication year, title, research sector, methods used to identify success factors, and how success is defined in each study.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the review search, selection, and inclusion process.

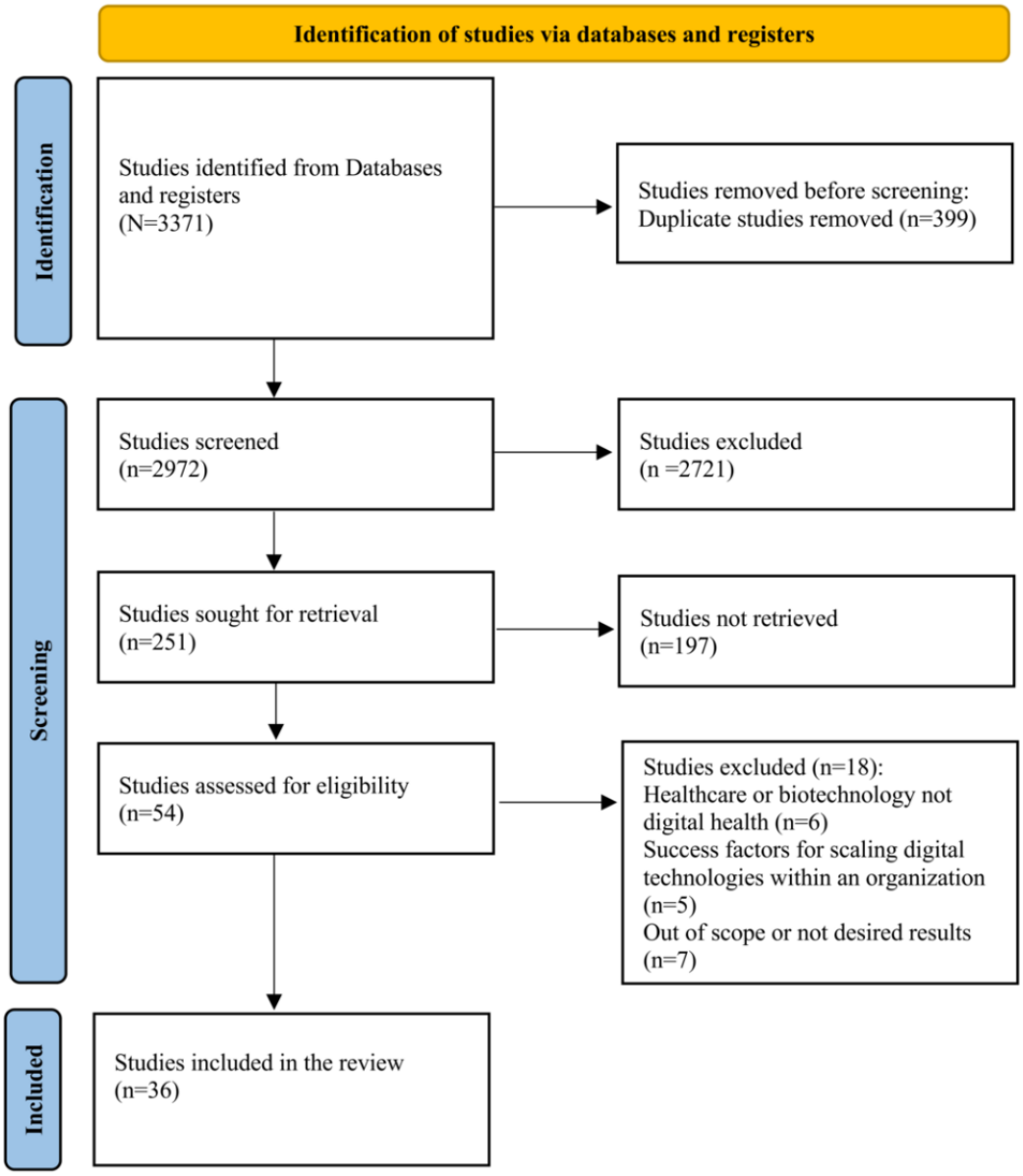

XSL•FO
RenderX

Table 2. Summary of selected studies on success factors of digital companies, including the first author, year of publication, title, sector, methods used, and definition of success

Study	Year of publication	Title	Sector	Method characteristics	Success definition
Abel et al [45]	2008	From technology imitation to market dominance: the case of iPod	Other digital	Single case study	Establish industry leadership
Azevedo et al [46]	2022	Scaling-up digital follow-up care services: collaborative development and implementation of remote patient monitoring pilot initiatives to increase access to follow-up care	DHT[a]	Multiple case studies	The scalability and implementation of remote patient monitoring services
Chakrabarti et al [47]	2022	Scaling an internet of things start-up: can alliance strategy help?	Other digital	Single case study	Reach US $10 million revenue in the next 4 years
Chakraborty et al [48]	2023	Critical success factors of startups in the e-health domain	DHT	Interviews	Reach the under-served market by reducing health care costs and increasing service delivery speed
Chanakira [49]	2009	Determining critical success factors of international expansion in African telephony industry	Other digital	Surveys	International expansion in the mobile telephony industry
Cresswell et al [50]	2013	Ten key considerations for the successful implementation and adoption of large-scale health information technology	DHT	Qualitative analysis and literature review	Effective integration and adoption of health IT in large-scale settings
Dong et al [51]	2022	Performance prediction of listed companies in smart health care industry: based on machine learning algorithms	DHT	Machine learning algorithms	Business performance of listed smart health care companies
Feindt et al [36]	2002	Identifying success factors for rapid growth in SME[b] e-commerce	Other digital	Multiple case studies	Successful competitive performance for the organization
Gengatharen and Standing [37]	2005	A framework to assess the factors affecting success or failure of the implementation of government-supported regional e-marketplaces for SMEs	Other digital	Literature review and multiple case studies	Best practices, theoretical frameworks, and models in the area of e-marketplaces
Greenhalgh et al [52]	2017	Beyond adoption: a new framework for theorizing and evaluating non-adoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies	DHT	Hermeneutic systematic review and empirical case studies	Achieving large-scale, sustained adoption
Halminen et al [53]	2021	Demonstrating the value of digital health: guidance on contextual evidence gathering for companies in different stages of maturity	DHT	Implementation of WHO[c]'s digital health evaluation guide and CIMO[d] framework	The appropriate level of evidence required to assess what digital health intervention companies develop
Hughes et al [54]	2021	Scaling digital health innovation: developing a new "service readiness level" framework of evidence	DHT	Interviews	Scaling up DHIs[e], expanding pilot projects for broader implementation and scalability
Kassner et al [55]	2023	The PropTech investors' dilemma – what are the key success factors that secure survival?	Other digital	Cox proportional hazards model with nonlinear splines	Survival over a certain period of time
König et al [40]	2023	Evaluation of tech ventures' evolving business models: rules for performance-related classification	Other digital	Principal component analysis and rough set theory mixed method	Turnover > €300,000 (US $316,000) and tripled the turnover when initially between €100,000 (US $106,000) and €500,00 (US $527,000) and doubled when initially above €500,000 (US $527,000)

XSL•FO
RenderX

Study	Year of publication	Title	Sector	Method characteristics	Success definition
Lennon et al [56]	2017	Readiness for delivering digital health at scale: lessons from a longitudinal qualitative evaluation of a National Digital Health Innovation Program in the United Kingdom	DHT	Longitudinal qualitative evaluation using interviews, focus groups, and surveys	Actions required in order to prepare the market and accelerate uptake
Li et al [41]	2022	Founders' creativity, business model innovation, and growth	Other digital	Surveys and empirical analysis	New venture growth and sustainable competitive advantage
Liu et al [57]	2021	The critical factors affecting the deployment and scaling of healthcare AI: viewpoint from an experienced medical center	DHT	Literature review and case observation	Enhancing medical centers income, reducing costs, and improving care quality
Mishra and Sharma [58]	2022	Digital transformation evaluation of telehealth using convergence, maturity, and adoption	DHT	Multiple case studies	Analyzing telehealth providers, identifying areas for improvement, and contributing to theory development
Ng et al [42]	2017	Exploring performance drivers for technology-based ventures from early stage to expansion: perspectives of venture capitalists	Other digital	Survey and interviews	Technology-based ventures' development and growth
Nirjar [59]	2008	Innovations and evolution of software SMEs: exploring the trajectories for sustainable growth	Other digital	Survey and model development	The sustained performance of entrepreneurial enterprises [...] in the software sector
Prodan et al [60]	2022	Success factors for scaling up the adoption of digital therapeutics towards the realization of P5 medicine	DHT	Literature review, interviews, and workshop	Scale-up of P5 medicine (personalized, participatory, predictve, precision medicine)
Raravi et al [61]	2013	Critical success factors: service industries	Other digital	Surveys, statistical analysis with IBM SPSS Statistics 19	Successful competitive performance for the individual, department, or organization
Richter and Harst [62]	2021	Tackling the scaling-up problem of digital health applications	DHT	Editorial	Strategies for successfully and sustainably implementing digital health solutions while considering barriers
Rohn et al [63]	2021	Digital platform-based business models–an exploration of critical success factors	Other digital	Semistructured interviews and qualitative content analysis	The configurations under which these business models (digital platform–based business models) are successful and ought to outperform traditional business models
Santisteban and Mauricio [14]	2017	Systematic literature review of critical success factors of IT startups	Other digital	Systematic literature review	The growth of the company and the number of jobs generated
Schlieter et al [17]	2022	Scale-up of digital innovations in healthcare: expert commentary on enablers and barriers	DHT	Workshop	Sustainable digital health diffusion
Sista and De [64]	2021	Scaling up smart city logistics projects: the case of the smooth project	Other digital	Qualitative case study, semistructured interviews, and thematic analysis	The ability of a system to improve its scale by aiming to meet the increasing volume demand
Smagalla [65]	2014	The truth about software startups	Other digital	Analysis of financial results and interviews	Rule of 126: US $100 million in revenue and 20% earnings before interest and taxes margins within 6 years of formation
Srinivasan and Venkatraman [66]	2018	Entrepreneurship in digital platforms: a network-centric view	Other digital	Network-centric theory	Positioning products and services within dynamic digital networks
Sweeney et al [67]	2023	Scaling AI-based industry 4.0 projects in the medical device industry: An exploratory analysis	DHT	Literature review and survey	Sustainable and scalable AI[f]-based industry 4.0 projects in the medical device manufacturing industries
TenBrink et al [43]	2017	Turnaround success in high technology growth stage firms	Other digital	Application of life cycle theory to technology-based firms	Actions and firm circumstances that are associated with successful growth stage turnarounds

XSL•FO
RenderX

Study	Year of publication	Title	Sector	Method characteristics	Success definition
Tseng et al [68]	2018	Catalyzing healthcare transformation with digital health: Performance indicators and lessons learned from a Digital Health Innovation Group	DHT	Case study	The "quadruple aim" of enhanced patient experience, reduced cost, improved population, health and improved clinician efficiency
Tsironis et al [69]	2017	e-Business critical success factors: toward the development of an integrated success model	Other digital	Quantitative survey, factor analysis, and structural equation modeling	The successful implementation of e-Business
Tsourvakas and Riskos [39]	2018	Emergent success factors for entrepreneurial e-media companies	Other digital	Case study	Unique characteristics of successful e-media business models
Yang et al [38]	2016	Scale, congestion, efficiency, and effectiveness in e-commerce firms	Other digital	Data envelopment analysis	Operational efficiency and economic effectiveness in e-commerce firms, with considerations for scale, technology, and resource allocation
Yin et al [44]	2019	The success factors of Korean global start-ups in the digital sectors through internationalization	Other digital	Single case study	Scale-up their businesses and gain success through global strategy

[a]DHT: digital health technologies.

[b]SME: small and medium scale enterprises.

[c]WHO: World Health Organization.

[d]CIMO: Context, Interventions Mechanisms, Outcomes Framework.

[e]DHI: digital health intervention.

[f]AI: artificial intelligence.

Success Definition and Success Metrics

In investigating definitions for success within the digital company landscape, our review revealed various approaches. These are categorized into 5 principal themes: industry position and market dominance [36,37,45,63]; financial performance, operational capabilities, and scalability [17,46,53,54,58,60,64]; market expansion and internationalization [44,49]; technological adoption and innovation [50,51,56,57,62,66,68]; and health impact and service delivery [38,39,41,42,48,52,59,61,67,69,70]. These themes represent a comprehensive view of success, from securing market leadership and achieving financial goals to executing scalable operations, broadening market presence, advancing technological innovation, and effecting social change. This variety in success narratives highlighted multiple interpretations of success across different studies and illustrated the importance of assigning relevant and context-specific success factors to different industry sectors.

RQ1: General Success Factors of Growth-Stage Digital Companies

In addressing RQ1, we identified 52 distinct success factors across 36 papers. These factors are listed in Table 3. An extended version, with quote examples that illustrate the categories, can be found in Multimedia Appendix 3. All factors contribute to the growth and sustainability of digital companies. We divided the success factors into two main categories: "internal," those within the company's control that can be optimized, and "external," those outside the company's control but still crucial for strategic planning.

XSL•FO
RenderX

Table 3. Breakdown of 52 success factors for growth-stage digital companies, categorized by Internal (Product and Services, Business Models, Operations, and Team Composition) and External (Customers, Healthcare System, Government and Regulators, Investors and Shareholders, Suppliers and Partners, and Competitors) factors

Third-order, second-order, and first-order categories (in order)	Rate	Occurrence (references)
I. Internal		
Product and Services		
Market demand and relevance of the product or service	13	[17,37,40,47,50,52,54,55,57,58,62,63,68]
User-centered design	10	[42,46,48,50,53,54,58,60,62,63]
Product or service innovation	10	[14,36,39,41-43,45,57,59,67]
Quality, performance, and brand trust	9	[39,48,54,56,58,59,63,67,68]
Data security, intellectual property protection, and ethical considerations	8	[17,46,48,54,56,60,64,67]
Convenience and standardization	7	[17,36,37,44,57,58,67]
Training and support programs	7	[37,48,50,54,56,61,67]
Price sensitivity	4	[36,48,61,64]
Number of patents	2	[40,59]
Operations		
Integration or technology change management	11	[17,36-38,42,45,50,52,59,67,69]
Optimized internal processes	10	[36,37,41,42,45,47,65,67-69]
Performance monitoring	9	[42,47,50,53,58,65,67-69]
Interoperability	6	[17,50,54,56,60,62]
Modern infrastructure	6	[42,48,50,56,64,67]
Risk management	5	[14,47,49,53,64]
Sales and marketing effectiveness	3	[36,37,65]
Workforce management	2	[17,47]
Business models		
Financially viable	18	[14,17,37,38,47-49,51-54,57,59,60,63-65,67]
Adaptability and flexibility to market changes and disruptions	11	[17,38,41-44,46,48,52,58,61]
Use of existing and emerging technologies	9	[17,42,44,47,51,55,58,59,61]
Value proposition and differentiation in the market	5	[17,41,47,48,52]
Business model consideration at an early stage	5	[17,38,50,59,64]
Competitive awareness and strategic positioning	4	[45,48,66,67]
Team composition		
Leadership experience and qualities of team members	11	[14,17,37,41-43,48,50,54,63,69]
Diversity of skills and expertise within the team	8	[17,38,42,46,48,57,59,68]
Innovativeness and technical competence	5	[14,17,42,52,59]
Organizational size	4	[40,43,55,70]
Experience in the sector	3	[14,41,42]
High motivation and focus	3	[14,37,44]
Training and development	2	[42,70]
Stock options	1	[42]
E. External		
Customers		
Customer feedback and satisfaction with the product or service	12	[36,39,41,42,47,48,54,61,62,65,67,69]
Customers awareness raising	6	[17,37,56,57,60,63]

XSL•FO
RenderX

Third-order, second-order, and first-order categories (in order)	Rate	Occurrence (references)
Regional market size, consumer needs, and behavior	5	[14,37,42,49,58]
Brand image and community building	2	[36,39]
Health care system		
Long-term integration with existing systems	7	[17,46,53,54,56,57,62]
Increase in affordability	5	[50,56,60,67,68]
Improved diagnosis or care	4	[54,58,60,68]
Enhanced coordination	4	[57,62,67,68]
Resource allocation and utilization optimization	3	[46,54,60]
Enhanced patient experience	1	[68]
Government and regulators		
Regulatory environment and policy framework	15	[14,17,37,48,52-54,56-58,60,62,64,67,68]
Government endorsement and direct support	3	[14,60,64]
Investors and shareholders		
Synergy and alignment of shareholders' goals and objectives	4	[37,64,66,69]
Investor backing and fit	4	[14,40,55,66]
Access to resources, networks, and expertise through partnerships	4	[43-45,66]
Suppliers and partners		
Alliance strategy in place	7	[14,17,47,49,54,64,66]
Collaboration with larger organizations	4	[36,45,66,68]
End-to-end value chain involvement	2	[17,46]
Competitors		
Market entry timing and organizational maturity	3	[14,45,55]
Intensity of competitors in the market	1	[14]
Collaborative opportunities with competitors for mutual growth	1	[14]

Internal Factors

Internal factors play an important role in shaping the operational and strategic framework of companies. In our comprehensive analysis, we identified 31 internal success factors, which are organized into 4 distinct second-order categories.

Factors within the I.1. Product and Services category emerged as the most prevalent, with a total of 70 occurrences (22.7% of all occurrences) in the reviewed literature. The leading factor in this category, I.1.a. Market Demand and Relevance of the product or service, was identified 13 times (13/70, 18.6%), underscoring the critical need to align products with consumer preferences and market trends. I.1.b. User-Centered Design and I.1.c. Product or Service Innovation were each highlighted 10 times (10/70, 14.3%), emphasizing the importance of prioritizing customer experience in product or service design and of continuously innovating such offerings. Other significant factors identified were I.1.d. Quality and Performance and Brand Trust (n=9, 2.9%), I.1.e. Data Security and Intellectual Property Protection (8/70, 2.6%), and I.1.f. Convenience and Standardization of Offerings (n=7, 2.3%). Together, these highlight the central role of quality, security, and customer-centered design as key drivers of a company's performance.

The I.2. Business Models category was highlighted 52 times (16.9% of all occurrences), with I.2.a. Financial Viability (18/52, 34.6%) as the most mentioned success factor. This category emphasizes the significance of the I.2.c. Use of Existing and Emerging Technologies (9/52, 17.3%) and a I.2.d. Strong Value Proposition and Differentiation in the Market (5/52, 9.6%), associated with I.2.f. Competitive Awareness (4/52). In addition, I.2.e. Early Business Model Planning (5/52, 9.6%) emerged as another important management practice vital for securing a firm's long-term viability.

I.3. Operations-related success factors drew 52 statements (16.9% of all occurrences), emphasizing the importance of seamless I.3.a. Technology Integration and Change Management (11/52, 21.2%) and I.3.b. Optimized Internal Processes (10/52, 19.2%) in fostering operational efficiency. I.3.c. Performance Monitoring (9/52, 17.3%) and I.3.d. Interoperability (6/52, 11.5%) highlighted important management practices, as well as again a key technology requirement. I.3.f. Risk Management (5/52, 9.6%) and I.3.g. Effective Sales and Marketing (n=3/52, 5.8%) also help support the success of digital companies.

With 37 (12.0% of all occurrences) mentions, the I.4. Composition and Dynamics of the Team emerged as another key domain impacting a company's success trajectory. I.4.a.

XSL•FO
RenderX

Leadership Qualities (11/37, 29.7%) and I.4.b. Diversity of Skills Within the Team (8/37, 21.6%) rank high in enabling the success of (digital) enterprises. In addition, I.4.d. Organizational Size (4/37, 10.8%), I.4.e. Sector-Specific Experience (3/37, 8.1%), and I.4.f. High Motivation Levels (3/37, 8.1%) pointed to important internal people-related factors to companies' success.

External Factors

The external factors emphasize the interactions of companies with various stakeholders. Overall, we distinguished 21 factors within 6 second-order categories. The most frequently mentioned external category concerns the engagement with E.1. Customers (n=25, 8.1% of all occurrences). E.1.a. Customer Feedback and Satisfaction (12/25, 48.0%) emerged as critical for refining product or service offerings and ensuring alignment with market needs. E.1.c. Understanding Regional Consumer Behavior (5/25, 20%) and E.1.d. Efforts in Brand and Community Building (2/25, 8%) are essential for cultivating customer loyalty and expanding market presence.

Impact on the E.2. Healthcare System was also highlighted as a crucial success factor in DHT-related papers, receiving 24 mentions (7.8% of all occurrences). This category highlights the importance of E.2.a. Long-Term Integration with Existing Systems (7/24, 29.2%), E.2.a. Affordability (5/24, 20.8%), and E.2.c. Improved Care Outcomes (4/24, 16.7%) for health care–related digital solutions.

Alignment with E.3. Government and Regulators is crucial, as raised 18 times (5.8% of all occurrences) in the literature review. The E.3.a. Regulatory Environment and Policy Framework (15/18, 83.3%) plays a fundamental role in shaping the market landscape for digital companies. E.3.b. Government Endorsement and Direct Support (3/18, 16.7%) are essential for navigating key legal and financial hurdles.

The alignment and interaction with other stakeholders, including investors, suppliers, and competitors, were mentioned less frequently. However, E.5.a. Alliance Strategy in Place (7/13, 53.8%) and E.5.b. Collaboration with Larger Organizations (4/13, 30.8%) within E.5. Suppliers and Partners (n=13, 4.3% of all occurrences) surfaced as essential to strengthening the value chain and enhancing product or service offerings, for example, through robust partnerships.

RQ2: Specific Success Factors of Growth-Stage Digital Health Companies

Our analysis examined success factors pertinent to the digital health sector relative to those in broader digital industries (Figure 2). Within the first category I.1. Product and Services, I.1.e. Data Security, Intellectual Property Protection, and Ethical Considerations ($n_{OtherDigital}$=1, $n_{DigitalHealth}$=7; $P_{Fisher\ exact\ test}$=.03) occurred with more than twice the frequency in digital health literature than in combined mentions within other digital industries underscoring the sector's unique priorities and challenges associated with handling patient data in the Digital Health sector. In reviewing the I.2. Business Models category, our analysis revealed that the representation of most factors was comparable between health care–focused companies and those in broader digital sectors. Fisher exact test showed no significant differences, with all factors yielding P>.05. This suggests similar strategic approaches across these industry segments. The I.3. Operations analysis showed I.3.d. Interoperability considerations ($n_{OtherDigital}$=0, $n_{DigitalHealth}$=6; $P_{Fisher\ exact\ test}$=.01) as having greater significance to digital health companies. This element is crucial for the seamless functioning within and between established health care infrastructures. In addition, within I.4. Team Composition, the I.4.b. Diversity of Skills and Expertise within the team was notably more prevalent among DHT companies, indicating the necessity for a wide range of competencies to address the multifaceted nature of health care and health technology challenges, though the difference did not reach statistical significance ($n_{OtherDigital}$=3, $n_{DigitalHealth}$=5; $P_{Fisher\ exact\ test}$=.49).

Regarding the influence of external stakeholders on digital companies, beyond just health care stakeholders, which appear only in DHT-related papers, E.3. Government Bodies and Regulatory Agencies are especially pertinent to digital health companies. All factors within this category show a higher representation in the health care domain, with the E.3.a. Regulatory Environment and Policy Framework as the most frequently mentioned ($n_{OtherDigital}$=3 vs $n_{DigitalHealth}$=12; $P_{Fisher\ exact\ test}$=.02).

XSL•FO
RenderX

Figure 2. Comparative analysis of success factor occurrences in growth-stage digital health companies versus other digital sectors, categorized by Internal (I) and External (E) factors. Light blue represents success factors in Digital Health companies, while dark blue represents success factors in broader digital industries. The numbers indicate the frequency of occurrences within the studies included in our literature review.

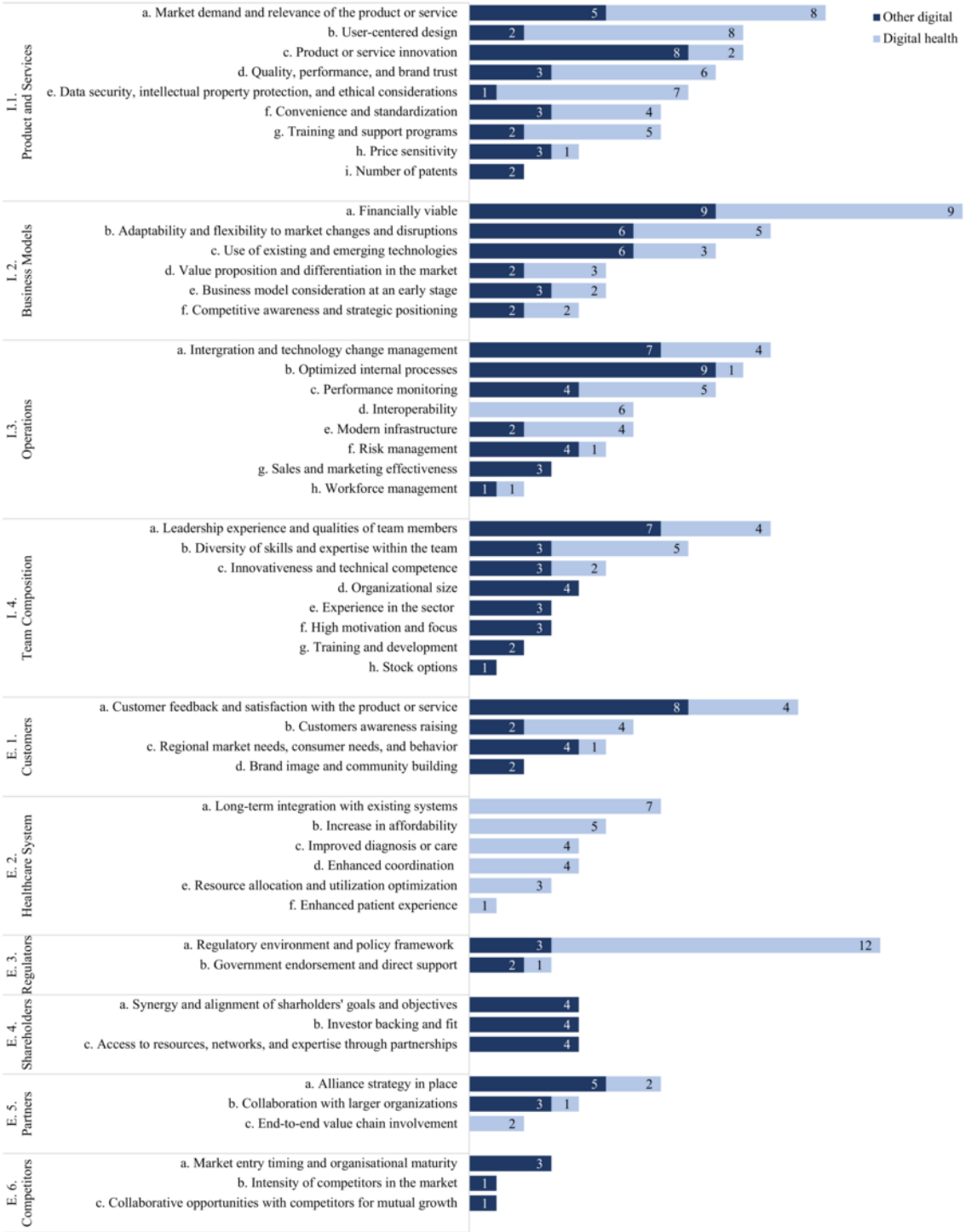

RQ3: Success Factors Differences Across the DTx Alliance Digital Health Technology Categories

We categorized the included papers focusing on digital health companies (15/36, 41.7%) based on the DHT categories defined by the DTx Alliance [71]: Health System Operational Software (2/14, 14.3%) [50,67], Health System Clinical Software (3/14, 21.4%) [48,57,62], Care Support (3/14, 14.3%) [52,54,58], Patient Monitoring (4/14, 28.6%) [46,52,54,58], and Digital Therapeutics (1/14, 7.1%) [60]. Not all papers were classifiable within these subcategories, as 5 focused on the broader scope of digital health without fitting into distinct segments [17,51,53,56,68]. None of the papers reviewed fell into the categories of Health and Wellness or Digital Diagnostics (Figure 3).

XSL•FO
RenderX

Figure 3. Distribution of success factors across DTx Alliance categories. A "-" symbol indicates that the success factor is not referenced. The shade of blue represents the frequency of occurrence, with the lightest shade signifying one occurrence and the darkest shade indicating 4 occurrences. The categories are as follows: Health System Operational Software, Health System Clinical Software, Health and Wellness, Patient Monitoring, Care Support, Digital Diagnosis, and Digital Therapeutics.

	HSOS	HSCS	HW	PM	CS	DD	DT
I.1. Product and Services							
a. Market demand and relevance of the product or service	1	3	-	3	3	-	-
b. User-centered design	1	2	-	2	3	-	1
c. Product and service innovation	1	1	-	-	-	-	-
d. Quality and performance/brand trust	1	1	-	2	2	-	-
e. Data security, IP protection and ethics	1	1	-	1	2	-	1
f. Convenience and standardization	1	1	-	1	1	-	-
g. Training and support programs	2	1	-	1	1	-	-
h. Price sensitivity	-	1	-			-	-
I.2. Operations	-	-	-	-	-	-	-
a. Intergration and technology change management	2	1	-	1	1	-	-
c. Performance monitoring	2	-	-	1	1	-	-
d. Interoperability	1	1	-	1	1	-	1
e. Modern infrastructure	2	1	-	-	-	-	
I.3. Business Models	-	-	-	-	-	-	
a. Financially viable	1	3	-	2	2	-	1
b. Adaptability and flexibility to market changes and disruptions	-	2	-	2	3	-	-
c. Use of existing and emerging technologies	-	-	-	1	1	-	-
d. Value proposition and differentiation in the market	-	2	-	1	1	-	-
e. Business model consideration at an early stage	1	-	-	-	-	-	-
f. Competitive awareness and strategic positioning	1	1	-	-	-	-	-
I.4. Team Composition	-	-	-	-	-	-	-
a. Leadership experience and qualities of team members	1	1	-	1	1	-	-
b. Diversity of skills and expertise within the team	-	2	-	-	1	-	-
c. Innovativeness and technical competence	-	1	-	1	1	-	-
E.1. Customers	-	-	-	-	-	-	-
a. Customer feedback and satisfaction with the product/service	1	2	-	1	1	-	-
b. Regional market needs, consumer needs, and behavior	-	-	-	1	1	-	-
c. Customers awareness raising	-	1	-	-	-	-	1
E.2. Healthcare Sytem	-	-	-	-	-	-	-
a. Long-term integration with existing systems	-	2	-	1	2	-	-
b. Increase in affordability	2	-	-	-	-	-	1
c. Improved diagnosis or care	-	-	-	2	2	-	1
d. Enhanced coordination	1	2	-	-	-	-	-
e. Resource allocation and utilization optimization	-	-	-	1	2	-	1
E.3. Government and Regulators	-	-	-	-	-	-	-
a. Regulatory environment and policy framework	1	4	-	3	3	-	1
b. Government endorsement and direct support	-	-	-	-	-	-	1
E.5. Suppliers and Partners	-	-	-	-	-	-	-
a. Alliance strategy in place	-	-	-	1	1	-	-
b. End-to-end value chain involvement	-	-	-	-	1	-	-

By focusing on factors more commonly mentioned within each category relative to the other 6, we observe that in Health System Operational Software, I.1.g. Training and Support Programs for Products or Services is prominent. In addition, I.2.a. Interoperability and Technology Change Management and I.2.e. Modern Infrastructure were found more often (2/5, 40%) in this category, highlighting their role in maintaining the integrity of operational software in health care systems. Moreover, E.2.b. Increase in Affordability of Health Care (2/3, 66.7%) positively impacts the success of Health System Operational Software.

In the category of Clinical Software, the emphasis on navigating the E.3.a. Regulatory Environment of the Location (4/12, 33.3%) emphasizes the critical nature of regulatory compliance. It is vital for such software not only to serve as a functional tool but to deliver tangible value, underscored by mentions of being I.3.a Financially Viable (3/9, 33.3%) and offering an attractive I.3.d. Value Proposition and Differentiation in the market (2/4, 50%). Effective Health System Clinical Software also depends on E.2.d. Enhanced Workforce Coordination (2/3, 66.6%) within health care environments and a team's I.4.b. Diversity of Skills and Expertise (2/3, 66.6%), whereby both clinical knowledge and software development proficiency are needed.

For Care Support and Patient Monitoring solutions to be broadly embraced, key aspects like the service's I.1.d. Quality and Performance and Brand trust (2/6, 33.3%) and E.2.a. Improved Diagnosis or Care (2/5, 40.0%) must be prioritized. Addressing the hesitation to switch from traditional to Telehealth and other innovative technologies involves not only focusing on these critical factors but also integrating I.1.b. User-Centered Design. In addition, the adoption of Digital Care Support by health care providers necessitates that these technologies significantly

XSL•FO
RenderX

improve E.2.e. Resource Allocation and Utilization Optimization (2/4, 50%) and be I.3.b. Flexibility to Market Changes and Disruptions (3/7, 42.9%) thereby enhancing the efficacy of the health care system. Only one paper fell into the DTx category, so our insights are limited. However, all the factors related to Government and Regulation were mentioned, underscoring their significance in DTx adoption.

Discussion

Theoretical Contribution

Our initial results indicate that, despite numerous advancements in DHT, the body of research specifically focused on the success factors of growth-stage digital health companies remains surprisingly limited. From a pool of 2972 studies, only 36 were relevant, with just 15 targeting DHTs directly. This scarcity of focused research suggests that one contributing factor could be a misalignment between current research priorities and the critical needs of the evolving digital health sector.

Internal success factors for digital companies, including those in the DHT sector, have been identified more frequently than external factors, which may be attributed to the limited control companies have over external influences. Among the internal factors, financial viability emerges as particularly critical. Indeed, the future of these companies is more closely tied to the demand for and proper payment for their products or services than to their contributions to improved health outcomes and more efficient health care provision. While extensive research focuses on the health impact of DHT, understanding their financial implications and assessing their economic value is equally crucial.

Our findings highlight that in external factors for DHT companies, ongoing support from regulatory bodies and policymakers is particularly essential for the continued success of these companies. Such support should include consistent guidance, assistance, and the development of policy frameworks that promote business growth. We urge stakeholders in DHT to clearly articulate the benefits their solutions offer to the health care system and to actively engage with regulatory authorities. This interaction not only ensures compliance but also encourages regulators to recognize and address any existing regulatory deficiencies within the DHT sector. Germany's adoption of the Digital Healthcare Act and the Digital Health Application Regulation highlights its forward-thinking approach to enhancing the digital health sector. This legislative framework not only accelerates the integration of digital health solutions, like prescription health apps but also sets a precedent for digital health advancement, fostering innovation and ensuring equitable access to digital health services [72]. For DHT companies, maintaining dialogs with public payers and insurance providers is important to ensure appropriate reimbursement. Moreover, adapting existing laws to accommodate new business models and technologies and efforts to ensure market fairness are vital steps in supporting the digital health ecosystem [54,60,73].

Integrating insights from the literature on success factors in other domains, such as product innovation, our study identifies both commonalities and unique aspects pertinent to DHT companies. Like general product innovation, the success factors we identified for DHT emphasize the importance of market demand, robust internal processes, and early customer involvement [12,74,75]. However, unique challenges, such as stringent regulatory compliance and the need for interoperability with existing health systems, set digital health apart from other sectors. While environmentally sustainable product innovation research underscores the importance of interfunctional collaboration and Research and Development investments [13], DHT also demands a heightened focus on data security and patient privacy. In addition, our findings highlight that adaptability and flexibility to market changes and disruptions are particularly crucial for DHT, especially given the regulatory uncertainty [76].

In examining the critical success factors across the various DTx Alliance DHT categories, as highlighted by RQ3, it becomes clear that the digital health sector is diverse, with the different categories having common as well as distinct success factors. For instance, Care Support solutions underscore the vital role of adaptability in the face of external disruptions, such as those brought about by global health crises like the COVID-19 pandemic [77]. This ability to adapt ensures the continuity of care and showcases the agility of DHTs to pivot in response to evolving health needs. To avoid encountering unforeseen and avoidable obstacles, digital health companies must understand the unique challenges and success factors specific to their DHT categories and the overarching factors driving their business forward.

Furthermore, RQ3 points out a notable research gap in the areas of Health and Wellness digital solutions and Digital Diagnostics. There's a clear need for more studies on best practices within these fields to provide a fuller picture of the critical focus areas for companies based on their DHT applications. Encouraging further research in these under-explored areas could greatly enrich our understanding of success factors in the evolving landscape of DHT.

Our analysis paints a picture of success in the digital health realm as a complex interplay of technological innovation, user-centric design, regulatory engagement, and systemic flexibility, among others. Given this complexity, stakeholders within the digital health ecosystem must adopt a comprehensive approach when developing and deploying DHTs. This means that technologies should push for innovation and be crafted to align with regulatory standards and operational exigencies, optimize business models, and meet market demands and stakeholders' expectations. By embracing this holistic strategy, digital health initiatives can ensure disruption of our health care landscape and sustained efficacy and impact in enhancing health care delivery.

Managerial Implications

The findings of this study hold significant implications for the digital health research community, highlighting the need for a deeper and more thorough exploration of success factors in the digital health sector. It advocates for a collaborative effort to narrow the divide between theoretical research and practical application, encouraging researchers to validate and refine identified success factors and to discover new ones, particularly

XSL•FO
RenderX

in underresearched areas such as digital health technologies for Health and Wellness (prevention), DTx or Digital Diagnostics. This effort will not only enrich the theoretical framework guiding the development and assessment of DHTs but also improve practical guidelines for their effective implementation and integration into the healthcare system.

This meticulously curated list of success factors is valuable for entrepreneurs and stakeholders active in the digital health industry. This resource can empower them to prioritize and focus on the most influential factors for success in this dynamic and competitive market. Founders in the digital health sector, particularly those with an established product or service and initial paying customers, often face challenges in scaling significantly after gaining initial traction. This research will be invaluable for them, offering targeted guidelines on emphasizing specific strategies, collaborative efforts, and organizational practices to facilitate their transition. They can go through the list one by one, establishing their current position regarding each factor and self-diagnosing areas where they are lagging and need more attention. For example, a company might notice they lack an alliance strategy and can brainstorm potential larger partners who could benefit from or deploy their product, ultimately boosting their growth. It provides a roadmap for leveraging initial successes into broader, sustainable growth within the digital health industry.

Investors in growth-stage digital health companies will find this research helpful for making informed decisions about which businesses have established the right priorities for scaling to success. It offers detailed insights into identifying companies poised for effective expansion. Furthermore, the research provides structured recommendations on scaling strategies, which are especially valuable for investors who serve on the boards of these companies, guiding them in advising digital health businesses toward sustainable growth.

As insurance companies begin implementing DHTs and offering them to their patients [78], this research provides essential guidance for selecting which DHT companies to collaborate with. Insurance providers can leverage the identified success factors to assess the viability of various digital health solutions. By choosing partners that align with these success factors, insurance companies can ensure they are integrating effective and reliable technologies into their offerings. The insights from this study can help insurance companies develop strategic partnerships and implement DHTs that support their goals of delivering high-quality, cost-effective care [79].

Limitations and Future Research Recommendations

The main limitation of the current work is that the success of DHT can significantly change depending on the health care environment, the demographics of the patient population, and the specifics of local regulations. Such diversity presents challenges universally applying identified success factors to all situations. We recommend that future research on the success of companies should focus on specific geographic regions to provide more tailored guidance.

Another limitation arises when comparing success across different digital health categories. Categories such as Digital Diagnostics or Health and Wellbeing have not been sufficiently analyzed regarding success factors. This lack of research limits our understanding of success factors in these evolving areas of DHTs. Ideally, we would have extensive research for each DTx Alliance category to better understand the key aspects that contribute to success in each sector. This would allow companies to receive tailored recommendations specific to their DHT category. Therefore, it is essential for researchers specialized in specific DHTs to conduct further studies to identify and elaborate on the critical and category-specific success factors.

Despite thorough attempts to search relevant literature extensively, the selection process, informed by precise inclusion and exclusion criteria, could have led to bias in the outcomes. Our search was limited to peer-reviewed literature to maintain high academic rigor and reliability, which excluded industry reports and gray literature. While we acknowledge that incorporating gray literature could yield additional insights, our aim for this systematic review was to focus on peer-reviewed sources to ensure the highest quality of included papers. We systematically screened papers, diligently identifying and coding quotes that revealed success factors. We categorized these factors based on our best judgment, acknowledging the nuanced interrelations among them. It is important to note that another group might have selected slightly different success factors; however, our use of broader categories ensures a comprehensive overview.

In addition, the rapid evolution of DHT might outstrip existing research, risking the obsolescence of identified success factors as innovations and updates arise. This dynamic could also reveal new factors not identified in this study. Therefore, we suggest future researchers in this field use our comprehensive list as a baseline to identify which success factors have become outdated or have emerged as new.

Conclusion

This systematic review explored success factors critical for growth-stage DHT companies, highlighting the interplay between internal and external factors such as market demand, regulatory compliance, and financial viability. Significant gaps were identified, notably in the under-represented categories like Digital Diagnostics, pointing to the need for focused research. The review emphasizes the necessity of adaptive strategies and innovation to meet changing market and regulatory demands. This study not only enhances academic understanding but also offers practical insights for stakeholders in the digital health industry, providing recommendations for DHT companies.

Data Availability

The data supporting this study's findings are available from the corresponding author upon request. The research protocol and other relevant materials can be provided.

XSL•FO
RenderX

Authors' Contributions

TK and EP developed the research questions and concept of this research. EP formulated the review approach and design with inputs from TK, MN, and CK. EP and LB performed the coding of the scientific articles. EP initiated the manuscript drafting, incorporating valuable feedback from TK throughout the iterations. The final version received full approval from all authors.

Conflicts of Interest

EP, LB, MN, and TK are affiliated with the Centre for Digital Health Interventions (CDHI), a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich, the Department of Management, Technology, and Economics at ETH Zurich, and the Institute of Technology Management and School of Medicine at the University of St. Gallen. CDHI is funded in part by CSS, a Swiss health insurer, the Swiss growth-stage investor MTIP, and the Austrian health provider Mavie Next. TK is also a co-founder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, this research did not involve CSS, Mavie Next, or Pathmate Technologies. EP and CK work at MTIP, a Swiss health tech growth equity firm that helps founders scale up successful and sustainable digital health businesses.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 34 KB - jmir_v26i1e60473_app1.docx]

Multimedia Appendix 2
Detailed search terms based on alternative keywords.
[DOCX File , 18 KB - jmir_v26i1e60473_app2.docx]

Multimedia Appendix 3
Detailed examples for each 1st-order category, illustrating thematic synthesis with quotes from relevant studies.
[DOCX File , 549 KB - jmir_v26i1e60473_app3.docx]

References

1. Jiang J, Qi K, Bai G, Schulman K. Pre-pandemic assessment: a decade of progress in electronic health record adoption among U.S. hospitals. Health Aff Sch 2023;1(5):qxad056 [FREE Full text] [doi: 10.1093/haschl/qxad056] [Medline: 38756982]
2. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
3. 2021 digital health funding - Rock Health. URL: https://rockhealth.com/insights/2021-year-end-digital-health-funding-seismic-shifts-beneath-the-surface/ [accessed 2024-11-01]
4. 2022 year-end digital health funding: Lessons at the end of a funding cycle Rock Health. 2023. URL: https://rockhealth.com/insights/2022-year-end-digital-health-funding-lessons-at-the-end-of-a-funding-cycle/ [accessed 2024-11-01]
5. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
6. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2(1):1-9 [FREE Full text] [doi: 10.1038/s41746-019-0111-3]
7. FDA permits marketing of mobile medical application for substance use disorder. US Food and Drug Administration. URL: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-mobile-medical-application-substance-use-disorder [accessed 2023-07-13]
8. SoftBank-backed Pear Therapeutics agrees $1.6 billion SPAC deal. Reuters. URL: https://www.reuters.com/article/us-pear-therapeutics-spac-idUSKCN2DY158 [accessed 2024-11-01]
9. Hong JS, Wasden C, Han DH. Introduction of digital therapeutics. Comput Methods Programs Biomed 2021;209:106319 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106319]
10. Litvinova O, Klager E, Tzvetkov N, Kimberger O, Kletecka-Pulker M, Willschke H, et al. Digital pills with ingestible sensors: patent landscape analysis. Pharmaceuticals (Basel) 2022;15(8):1025 [FREE Full text] [doi: 10.3390/ph15081025] [Medline: 36015173]
11. Dikert K, Paasivaara M, Lassenius C. Challenges and success factors for large-scale agile transformations: a systematic literature review. J Syst Softw 2016;119:87-108. [doi: 10.1016/j.jss.2016.06.013]
12. Johne FA, Snelson PA. Success factors in product innovation: a selective review of the literature. J Prod Innov Manag 2003;5(2):114-128. [doi: 10.1111/1540-5885.520114]
13. De Medeiros JF, Ribeiro JLD, Cortimiglia MN. Success factors for environmentally sustainable product innovation: a systematic literature review. J Clean Prod 2014;65:76-86. [doi: 10.1016/j.jclepro.2013.08.035]

XSL•FO
RenderX

14. Santisteban J, Mauricio D. Systematic literature review of critical success factors of information technology startups. Acad Entrep J 2017;23:1-23.
15. Smith CM. Origin and uses of primum non nocere--above all, do no harm!. J Clin Pharmacol 2005;45(4):371-377. [doi: 10.1177/0091270004273680] [Medline: 15778417]
16. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. it - Information Technology 2019;61(5-6):253-263 [FREE Full text] [doi: 10.1515/itit-2019-0019]
17. Schlieter H, Marsch L, Whitehouse D, Otto L, Londral A, Teepe GW, et al. Scale-up of digital innovations in health care: expert commentary on enablers and barriers. J Med Internet Res 2022;24(3):e24582 [FREE Full text] [doi: 10.2196/24582] [Medline: 35275065]
18. Lewis V, Churchill NC. The five stages of small business growth. 1983. URL: https://papers.ssrn.com/abstract=1504517 [accessed 2024-03-22]
19. Fitria SE, Hakim FR. Identification of critical success factor startup in business incubators (Case Study: Bandung Techno Park). Int J Soc Sci Res 2022;2(10):881-895 [FREE Full text] [doi: 10.46799/ijssr.v2i10.162]
20. Iyer AK. Building digital health and therapeutic solutions for the future: what's required for success? In: Digital Therapeutics: Strategic, Scientific, Developmental, and Regulatory Aspects. New York, NY: Chapman and Hall/CRC Press; 2022:245-260.
21. Werth O, Cardona DR, Torno A, Breitner M, Muntermann J. What determines FinTech success? A taxonomy-based analysis of FinTech success factors. Electron Mark 2023;33(1):21 [FREE Full text] [doi: 10.1007/s12525-023-00626-7] [Medline: 37252675]
22. Skawi?ska E, Zalewski R. Success factors of startups in the EU—a comparative study. Sustainability 2020;12(19):8200 [FREE Full text] [doi: 10.3390/su12198200]
23. Santisteban J, Mauricio D, Cachay O. Critical success factors for technology-based startups. Int J Entrep Small Bus 2021;42(4):397-421.
24. Żbikowski K, Antosiuk P. A machine learning, bias-free approach for predicting business success using crunchbase data. Inf Process Manag 2021;58(4):102555. [doi: 10.1016/j.ipm.2021.102555]
25. Pfitzer E, Bitomsky L, Nissen M, Kausch C, Kowatsch T. Success factors of growth-stage digital health companies: protocol for a systematic literature review. medRxiv 2024:24306674 Preprint posted online May 06, 2024 [FREE Full text] [doi: 10.2196/preprints.60473]
26. Tranfield D, Denyer D, Smart P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Br J Manag 2003;14(3):207-222 [FREE Full text] [doi: 10.1111/1467-8551.00375]
27. Snyder H. Literature review as a research methodology: an overview and guidelines. J Bus Res 2019;104:333-339. [doi: 10.1016/j.jbusres.2019.07.039]
28. Page MJ, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10(1):1-11 [FREE Full text] [doi: 10.31222/osf.io/jb4dx]
29. Umesh UN, Huynh MQ, Jessup L. Creating successful entrepreneurial ventures in IT. Commun ACM 2005;48(6):82-87. [doi: 10.1145/1064830.1064831]
30. Grodal S, Anteby M, Holm AL. Achieving rigor in qualitative analysis: the role of active categorization in theory building. Acad Manag Rev 2021;46(3):591-612 [FREE Full text] [doi: 10.5465/amr.2018.0482]
31. Barney J. Firm resources and sustained competitive advantage. J Manage 1991;17(1):99-120 [FREE Full text] [doi: 10.1177/014920639101700108]
32. Teece DJ, Pisano G, Shuen A. Dynamic capabilities and strategic management. Strateg Manag J 1997;18:509-533 [FREE Full text] [doi: 10.1002/(sici)1097-0266(199708)18:7<509::aid-smj882>3.0.co;2-z]
33. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences 2020;1(1):31-42 [FREE Full text] [doi: 10.1177/2632084320947559]
34. McCliskey A, O'Connor D. Implementing evidence: closing research-practice gaps. In: Hoffman T, Bennett S, Mar CD, editors. Evidence-Based Practice Across Health Professions. Sydney, NSW, Australia: Elsevier; 2017:384-408.
35. Hannes K, Macaitis K. A move to more systematic and transparent approaches in qualitative evidence synthesis: update on a review of published papers. Qual Res J 2012;12(4):402-442 [FREE Full text] [doi: 10.1177/1468794111432992]
36. Feindt S, Jeffcoate J, Chappell C. Identifying success factors for rapid growth in SME E-commerce. Small Bus Econ 2002;19(1):51-62 [FREE Full text]
37. Gengatharen DE, Standing C. A framework to assess the factors affecting success or failure of the implementation of government-supported regional e-marketplaces for SMEs. Eur J Inf Syst 2005;14(4):417-433 [FREE Full text] [doi: 10.1057/palgrave.ejis.3000551]
38. Yang Z, Shi Y, Yan H. Scale, congestion, efficiency and effectiveness in e-commerce firms. Electron Commer Res Appl 2016;20:171-182. [doi: 10.1016/j.elerap.2016.07.003]
39. Tsourvakas G, Riskos K. Emergent success factors for entrepreneurial E-media companies. J Entrep Innov Emerg Econ 2018;4(2):101-120 [FREE Full text] [doi: 10.1177/2393957518782040]
40. König M, Enjolras M, Ungerer C, Camargo M, Baltes G. Evaluation of tech ventures' evolving business models: rules for performance-related classification. Int J Entrep Ventur 2022;14(2):145-167 [FREE Full text] [doi: 10.1504/ijev.2022.122639]

XSL•FO
RenderX

41. Li Y, Li B, Lu T. Founders' creativity, business model innovation, and business growth. Front Psychol 2022;13:892716 [FREE Full text] [doi: 10.3389/fpsyg.2022.892716] [Medline: 35756233]
42. Ng A, Macbeth D, Yip L. Exploring performance drivers for technology-based ventures from early stage to expansion: perspectives of venture capitalists. Venture Cap 2017;19(4):335-359 [FREE Full text] [doi: 10.1080/13691066.2017.1334302]
43. TenBrink C, Keller RT, Gelb B. Turnaround success in high technology growth stage firms. J Policy Pract 2017;18(3):1707 [FREE Full text]
44. Yin W, Moon HC, Lee Y. The success factors of Korean global start-ups in the digital sectors through internationalization. International Journal of Global Business and Competitiveness 2019;14(1):42-53 [FREE Full text] [doi: 10.1007/s42943-019-00003-2]
45. Abel I. From technology imitation to market dominance: the case of iPod. Compet Rev 2008;18(3):257-274. [doi: 10.1108/10595420810906028]
46. Azevedo S, Guede-Fernández F, von Hafe F, Dias P, Lopes I, Cardoso N, et al. Scaling-up digital follow-up care services: collaborative development and implementation of remote patient monitoring pilot initiatives to increase access to follow-up care. Front Digit Health 2022;4:1006447 [FREE Full text] [doi: 10.3389/fdgth.2022.1006447] [Medline: 36569802]
47. Chakrabarti D, Sarkar S, Mukherjee A. Scaling an Internet of things start-up: can alliance strategy help? J Inf Technol Teach Cases 2022;12(1):43-49 [FREE Full text] [doi: 10.1177/2043886920986165]
48. Chakraborty I, Ilavarasan P, Edirippulige S. Critical success factors of startups in the e-health domain. Health Policy Technol 2023;12(3):100773. [doi: 10.1016/j.hlpt.2023.100773]
49. Chanakira M. Determining critical success factors of international expansion in African telephony industry. Journal of Global Business and Management Research 2009;1(1):50.
50. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful implementation and adoption of large-scale health information technology. J Am Med Inform Assoc 2013;20(e1):e9-e13 [FREE Full text] [doi: 10.1136/amiajnl-2013-001684] [Medline: 23599226]
51. Dong B, Wang X, Cao Q. Performance prediction of listed companies in smart healthcare industry: based on machine learning algorithms. J Healthc Eng 2022;2022:8091383 [FREE Full text] [doi: 10.1155/2022/8091383] [Medline: 35035859]
52. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
53. Halminen O, Chen A, Tenhunen H, Lillrank P. Demonstrating the value of digital health: guidance on contextual evidence gathering for companies in different stages of maturity. Health Serv Manage Res 2021;34(1):13-20. [doi: 10.1177/0951484820971447] [Medline: 33291977]
54. Hughes J, Lennon M, Rogerson R, Crooks G. Scaling digital health innovation: developing a new 'Service Readiness Level' framework of evidence. Int J Environ Res Public Health 2021;18(23):12575 [FREE Full text] [doi: 10.3390/ijerph182312575] [Medline: 34886302]
55. Kassner AJ, Cajias M, Zhu B. The PropTech investors' dilemma – what are the key success factors that secure survival? J Prop Invest Finance 2022;41(1):76-91. [doi: 10.1108/jpif-01-2022-0007]
56. Lennon M, Bouamrane M, Devlin A, O'Connor S, O'Donnell C, Chetty U, et al. Readiness for delivering digital health at scale: lessons from a longitudinal qualitative evaluation of a national digital health innovation program in the United Kingdom. J Med Internet Res 2017;19(2):e42 [FREE Full text] [doi: 10.2196/jmir.6900] [Medline: 28209558]
57. Liu CF, Huang CC, Wang JJ, Kuo KM, Chen C. The critical factors affecting the deployment and scaling of healthcare AI: viewpoint from an experienced medical center. Healthcare (Basel) 2021;9(6):685 [FREE Full text] [doi: 10.3390/healthcare9060685] [Medline: 34200189]
58. Mishra V, Sharma MG. Digital transformation evaluation of telehealth using convergence, maturity, and adoption. Health Policy and Technology 2022;11(4):100684. [doi: 10.1016/j.hlpt.2022.100684]
59. Nirjar A. Innovations and evolution of software SMEs: exploring the trajectories for sustainable growth. Vision J Bus Perspect 2008;12(2):47-59 [FREE Full text] [doi: 10.1177/097226290801200205]
60. Prodan A, Deimel L, Ahlqvist J, Birov S, Thiel R, Toivanen M, et al. Success factors for scaling up the adoption of digital therapeutics towards the realization of P5 medicine. Front Med (Lausanne) 2022;9:854665 [FREE Full text] [doi: 10.3389/fmed.2022.854665] [Medline: 35492346]
61. Raravi P, Bagodi V, Mench R. Critical success factors: service industries. SCMS Journal of Indian Management 2013;100:7.
62. Richter P, Harst L. Tackling the scaling-up problem of digital health applications. Z Gesundh Wiss 2022;30(1):1-3 [FREE Full text] [doi: 10.1007/s10389-021-01599-7] [Medline: 34109100]
63. Rohn D, Bican PM, Brem A, Kraus S, Clauss T. Digital platform-based business models – an exploration of critical success factors. J Eng Technol Manag 2021;60:101625. [doi: 10.1016/j.jengtecman.2021.101625]
64. Sista E, De GP. Scaling up smart city logistics projects: the case of the smooth project. Smart Cities 2021;4:1337-1365 [FREE Full text] [doi: 10.3390/smartcities4040071]
65. Smagalla D. The truth about software startups: it's not the size of the budget but how it is used that determines success or failure of the enterprise. MIT Sloan Manag Rev 2004;45(2):7.

XSL•FO
RenderX

66. Srinivasan A, Venkatraman N. Entrepreneurship in digital platforms: a network-centric view. Strateg Entrep J 2018:54-71 [FREE Full text] [doi: 10.1002/sej.1272]
67. Sweeney D, Nair S, Cormican K. Scaling AI-based industry 4.0 projects in the medical device industry. Procedia Comput Sci 2023;219:719-766 [FREE Full text] [doi: 10.1016/j.procs.2023.01.349]
68. Tseng J, Samagh S, Fraser D, Landman AB. Catalyzing healthcare transformation with digital health: performance indicators and lessons learned from a digital health innovation group. Healthc (Amst) 2018;6(2):150-155 [FREE Full text] [doi: 10.1016/j.hjdsi.2017.09.003]
69. Tsironis L, Gotzamani K, Mastos T. e-Business critical success factors: toward the development of an integrated success model. Bus Process Manag J 2017;2309(5):874-896 [FREE Full text] [doi: 10.1108/bpmj-02-2016-0030]
70. Santisteban J, Mauricio D. Systematic literature review of critical success factors of information technology startups. Acad Entrep J 2017;2309(2):1-23 [FREE Full text]
71. Digital Therapeutics Alliance, Health Advances. Digital health technology ecosystem categorization (Factsheet). URL: https://dtxalliance.org/wp-content/uploads/2023/06/DTA_FS_DHT-Ecosystem-Categorization.pdf [accessed 2023-08-11]
72. Sauermann S, Herzberg J, Burkert S, Habetha S. DiGA - a chance for the german healthcare system. J Eur CME 2022;11(1):2014047 [FREE Full text] [doi: 10.1080/21614083.2021.2014047] [Medline: 34992948]
73. Vayena E, Haeusermann T, Adjekum A, Blasimme A. Digital health: meeting the ethical and policy challenges. Swiss Med Wkly 2018;148:w14571 [FREE Full text] [doi: 10.4414/smw.2018.14571] [Medline: 29376547]
74. Evanschitzky H, Eisend M, Calantone R, Jiang Y. Success factors of product innovation: an updated meta-analysis. J Prod Innov Manag 2012;29:21-37 [FREE Full text] [doi: 10.1111/j.1540-5885.2012.00964.x]
75. Ernst H. Success factors of new product development: a review of the empirical literature. Int J Manag Rev 2003;4(1):1-40. [doi: 10.1111/1468-2370.00075]
76. Landers C, Vayena E, Amann J, Blasimme A. Stuck in translation: Stakeholder perspectives on impediments to responsible digital health. Front Digit Health 2023;5:1069410 [FREE Full text] [doi: 10.3389/fdgth.2023.1069410] [Medline: 36815171]
77. Filip R, Gheorghita Puscaselu R, Anchidin-Norocel L, Dimian M, Savage WK. Global challenges to public health care systems during the COVID-19 pandemic: a review of pandemic measures and problems. J Pers Med 2022;12(8):1295 [FREE Full text] [doi: 10.3390/jpm12081295] [Medline: 36013244]
78. Pfitzer E, Jörn B, Kowatsch T. Harmonizing innovation and regulation: swiss health insurance navigates digital health technologies for enhanced patient engagement. 2024 Presented at: 17th International Joint Conference on Biomedical Engineering Systems and Technologies; 2024 February 21-23; Rome, Italy p. 833-837. [doi: 10.5220/0012368300003657]
79. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GLD, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]

Abbreviations

CASP: Critical Appraisal Skills Programme
DHT: digital health technologies
DTx: digital therapeutics
FDA: Food and Drug Administration
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research questions

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by U Sinha, V Astha; comments to author 29.06.24; revised version received 08.08.24; accepted 29.10.24; published 11.12.24.

Please cite as:
Pfitzer E, Bitomsky L, Nißen M, Kausch C, Kowatsch T
Success Factors of Growth-Stage Digital Health Companies: Systematic Literature Review
J Med Internet Res 2024;26:e60473
URL: https://www.jmir.org/2024/1/e60473
doi:10.2196/60473
PMID:

©Estelle Pfitzer, Laura Bitomsky, Marcia Nißen, Christoph Kausch, Tobias Kowatsch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation and Comparison of Academic Impact and Disruptive Innovation Level of Medical Journals: Bibliometric Analysis and Disruptive Evaluation

Yuyan Jiang[1], MPhil; Xue-li Liu[1,2], BM; Zixuan Zhang[1], MPhil; Xinru Yang[1], MPhil

[1]Henan Research Center for Science Journals, Xinxiang Medical University, Xinxiang, China

[2]Faculty of Humanities & Social Sciences, Xinxiang Medical University, Xinxiang, China

Corresponding Author:
Xue-li Liu, BM
Faculty of Humanities & Social Sciences
Xinxiang Medical University
Library and Information Building, 2nd Fl.
No. 601, Jinsui Avenue, Hongqi District
Xinxiang, 450003
China
Phone: 86 1 383 736 0965
Email: liueditor03@163.com

Abstract

Background: As an important platform for researchers to present their academic findings, medical journals have a close relationship between their evaluation orientation and the value orientation of their published research results. However, the differences between the academic impact and level of disruptive innovation of medical journals have not been examined by any study yet.

Objective: This study aims to compare the relationships and differences between the academic impact, disruptive innovation levels, and peer review results of medical journals and published research papers. We also analyzed the similarities and differences in the impact evaluations, disruptive innovations, and peer reviews for different types of medical research papers and the underlying reasons.

Methods: The general and internal medicine Science Citation Index Expanded (SCIE) journals in 2018 were chosen as the study object to explore the differences in the academic impact and level of disruptive innovation of medical journals based on the OpenCitations Index of PubMed open PMID-to-PMID citations (POCI) and H1Connect databases, respectively, and we compared them with the results of peer review.

Results: First, the correlation coefficients of the Journal Disruption Index (JDI) with the Journal Cumulative Citation for 5 years (JCC_5), Journal Impact Factor (JIF), and Journal Citation Indicator (JCI) were 0.677, 0.585, and 0.621, respectively. The correlation coefficient of the absolute disruption index (Dz) with the Cumulative Citation for 5 years (CC_5) was 0.635. However, the average difference in the disruptive innovation and academic influence rankings of journals reached 20 places (about 17.5%). The average difference in the disruptive innovation and influence rankings of research papers reached about 2700 places (about 17.7%). The differences reflect the essential difference between the two evaluation systems. Second, the top 7 journals selected based on JDI, JCC_5, JIF, and JCI were the same, and all of them were H-journals. Although 8 (8/15, 53%), 96 (96/150, 64%), and 880 (880/1500, 58.67%) of the top 0.1%, top 1%, and top 10% papers selected based on Dz and CC_5, respectively, were the same. Third, research papers with the “changes clinical practice” tag showed only moderate innovation (4.96) and impact (241.67) levels but had high levels of peer-reviewed recognition (6.00) and attention (2.83).

Conclusions: The results of the study show that research evaluation based on innovative indicators is detached from the traditional impact evaluation system. The 3 evaluation systems (impact evaluation, disruptive innovation evaluation, and peer review) only have high consistency for authoritative journals and top papers. Neither a single impact indicator nor an innovative indicator can directly reflect the impact of medical research for clinical practice. How to establish an integrated, comprehensive, scientific, and reasonable journal evaluation system to improve the existing evaluation system of medical journals still needs further research.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e55121) doi:10.2196/55121

KEYWORDS

medical journals; journal evaluation; innovative evaluation; journal disruption index; disruptive innovation; academic impact; peer review

Introduction

Scientific and technical journals play a crucial role in showcasing research findings, and the value orientation of their published results is closely intertwined with their evaluation orientation. However, since Garfield [1] put forward the idea that "citation analysis can be used as an evaluation tool for journals" in 1972, the evaluation system of journals based on academic impact has become mainstream. However, relying too much on impact indicators for evaluation may hurt academic research and discipline progress.

On the one hand, some scholars have long pointed out that ranking journals according to their impact factors is noncomprehensive and may lead to misleading conclusions [2,3]. Meanwhile, many academic journals and publishers have engaged in strategic self-citation, leading to an overinflated journal impact factor (JIF) [4]. Some editorial behaviors to enhance the JIF clearly violate academic norms [5]. Some scholars are overciting each other's work to enhance their academic impact [6]. External contingencies can have a devastating effect on citation indicators [7]. Scientists themselves also present a mixed attitude toward impact factors [8].

On the other hand, despite all the benefits of increased academic impact to journals, there is a nonnegligible problem in the evaluation of journals that citation indicators essentially characterize the impact of journals rather than their disruptive innovation [9]. Relevant studies have confirmed that the level of disruptive innovation of scientific research is getting increasingly lower [10] and the progress in various disciplines is slowing [11]. This is often overlooked against the background of impact-only evaluations. Therefore, despite the urgent need for disruptive innovations in science [12], impact-based journal rankings have made it more difficult to accept novel results [13], replacing the "taste of science" with the "taste of publication" [14] in the actual environment.

The evaluation of academic journals is about not only the journals themselves [15] but also the wide use of the evaluation results in academic review, promotion, and tenure decisions [16]. Meanwhile, the quality and results of medical research are directly related to human health and life and have a direct impact on human health and well-being. Therefore, the general and internal medicine journals indexed in Science Citation Index Expanded (SCIE) in 2018 were chosen as the study object. The OpenCitations Index of PubMed open PMID-to-PMID citations (POCI) and H1 Connect databases were selected as the sources for citation relationship data and peer review data. We investigated the connections and contrasts between the academic impact, disruptive innovation level, and results of peer review for medical journals and published research papers. We also analyzed the similarities and differences as well as the fundamental causes of the varying evaluation results in terms of impact evaluation, disruptive innovation, and peer review for various types of medical research papers. We aimed to provide a reference for the correct understanding of the innovation level of the results published by journals; the scientific and reasonable evaluation of medical journals; and the construction of a scientific, objective, and fair academic evaluation system.

Methods

Research Object

Because there is basically no disruptive innovation in the review literature, this paper only focuses on the general and internal medicine journals indexed in SCIE in 2018 and the research papers involved. In addition, considering the computational efficiency, accuracy, and difficulty of data acquisition, research papers for which citation relationships in the aforementioned journals were not included in POCI and the journals with too few references (less than 10) were excluded. Finally, 114 journals were retained at the journal level, and 15,206 research papers were retained at the paper level.

Data Resource

The data acquired in this study included journal information, literature information, citation relationship data, and peer review data. The data were obtained through the Journal Citation Reports (JCR), Web of Science (WoS), POCI, and H1 Connect databases.

Of these databases, POCI is a Resource Description Framework (RDF) data set containing details of all the citations from publications bearing PMIDs to other PMID-identified publications harvested from the National Institutes of Health Open Citations Collection. POCI covers more than 29 million bibliographic resources and more than 717 million citation links (as of the last update in January 2023). Citations in POCI are not considered simple links but as data entities. This means that it is permissible to assign descriptive attributes to each citation, such as the date of creation of the citation, its time span, and its type [17].

H1 Connect (formerly Faculty Opinions), the world's most authoritative peer review database in the biomedical field, incorporates the combined efforts of more than 8000 international authoritative experts from around the globe and is a knowledge discovery tool used to evaluate published research. H1 Connect's reviewers are authoritative experts in the life sciences and medical fields. They provide commentary, opinions, and validation of key papers in their own field. The quality and rigor of the reviewers mean that researchers can be assured of the quality of the papers they recommend, and H1 Connect brings these recommendations together to recommend high-quality research to a wider audience. H1 Connect's experts

XSL•FO
RenderX

typically evaluate the "high level" of research literature in the field within 2 months of publication, with over 90% of recommendations made within 6 months of publication.

Data Acquisition and Processing

The data acquired for this study consisted of 3 parts. The specific steps for data acquisition and processing were (1) data acquisition and (2) data processing.

Data Acquisition

The steps taken to acquire the data included the following: log in to JCR; select Medicine, General & Internal in the "Browse categories" page; select JCR Year=2018 and Citation Indexes=SCIE in the filters; export the result to XLS format; according to the acquired journal title, select the publication year as 2018 and the literature type as Article to search the WoS core database; and export the full record of the journal literature in XLS format. Finally, we downloaded the related H1 Connect literature data and POCI data according to the list of journals.

Data Processing

To process the data, we undertook the following steps: use Navicat to import the full record of research papers into the local SQLite database, process the downloaded POCI data, extract the PMID numbers of all the focus papers from the full record, retrieve the references and citations of the focus papers as well as the citations of the references of the focus papers in the local database transformed based on the POCI data, and establish the relevant data tables for the subsequent calculations.

Evaluation Indicators

Innovation Indicators

Some researchers have observed, at an early stage, that some technological innovations complete and improve current technologies without replacing them while others outright eliminate technologies that were used in the past. However, for a long time, scholars did not analyze and explain the essence of this phenomenon. It was not until 1986 that Tushman and Anderson [18] summarized the phenomenon as follows: There are 2 types of major technological shifts that disrupt or enhance the capabilities of existing firms in an industry. However, Christensen [19], a professor at Harvard Business School in the United States, argued that disruptive innovations are new technologies that replace existing mainstream technologies in unexpected ways. Building on these views, Funk and Owen-Smith [20] provided deeper and more insightful insights. They argued that the dichotomy between disruptive and augmentative technologies lacks nuance and that the impact of new technologies on the status quo is a matter of degree rather than absolute impact.

In this regard, Govindarajan and Kopalle [21] also pointed out that disruptive innovation lacks reliable and valid measurement standards. Therefore, Funk and Owen-Smith [20] created the consolidation-disruption (CD) index, which aims to quantify the degree of technological change brought about by new patents. The index drew the attention of Wu et al [22], who analogized the basic principle of the CD index to measure disruption by calculating the citation substitution of the focus paper in the citation network and who were the first to apply the evaluation of disruptive innovation to the world of bibliometrics.

As an important carrier of academic results, it is important to evaluate papers quantitatively, rationally, and efficiently in terms of their innovation [23]. The disruption (D) index has received widespread attention after it was proposed by Wu et al [22]. A subset of scholars then explored disruptive papers in specific subject areas, including scientometrics [24], craniofacial surgery [25], pediatric surgery [26], synthetic biology [27], energy security [28], colorectal surgery [29], otolaryngology [30], military trauma [31], breast cancer [32], radiology [33], ophthalmology [34], plastic surgery [35], urology [36], and general surgery [37], based on the D index. Park et al [10] also analyzed the annual dynamics of the disruption level of papers and patents across the subject areas.

Another group of scholars conducted in-depth research on the index itself: Bornmann et al [38] explored the convergent validity of the index and the variants that may enhance the effectiveness of the measure and tested the validity of the D index based on the literature in the field of physics [39]. Ruan et al [40] provided an in-depth reflection on the limitations of the application of the D index as a measure of scientific and technological progress. Liu et al [41,42] empirically investigated the stabilization time window of the D index in different subject areas and addressed the mathematical inconsistency of the traditional D index and proposed an absolute disruption index (Dz; as in Equation 2) [43].

This series of studies has made it possible to evaluate the disruption of research papers based on the D index, which has gradually matured. On this basis, Jiang and Liu [44] proposed the Journal Disruption Index (JDI) to evaluate the disruptive innovation level of journals (as in Equation 3) and validated the evaluation effect of this indicator based on Chinese SCIE journals [45].

In Equations 1, 2, and 3, N_F refers to the literature that only cites the focus paper (FP), N_B refers to the literature that cites both the focus paper and at least one reference (R) of the focus paper, and N_R refers to the literature that only cites at least one reference (R) of the focus paper but not the focus paper. n is the number of "Article" type pieces of literature contained in the journal, and Dz_i is the Dz of the ith article in the journal.

In this study, Dz and JDI were chosen to evaluate the disruption of the selected studies at the literature and journal levels, respectively. Bornmann and Tekles [46] argued that 3 years is necessary regardless of the measurement for that discipline. Considering that, in the determination of the stabilization time window of the D index for each discipline conducted by Liu et al [42], the stabilization time window for clinical medicine is 4 years after publication. Therefore, in the process of calculating Dz in this paper, the citation time window of the focus papers was set to 2018 to 2022 to ensure the validity of the results. In addition, because the JDI value was too small, we multiplied the JDI by 1000 in all subsequent presentations.

XSL•FO
RenderX

Peer Review Indicators

The peer review indicators selected for this study included the peer review score (PScore), weighted peer review stars (PStar_w), and weighted peer review evaluation times (PTime_w). In this case, the weighted indicators refer to the weighting of ratings and number of evaluations by the number of evaluators when an evaluation was completed by more than one reviewer.

The advantage of using peer review indicators is that it can make up for the lag and one-sidedness of relying solely on the citations among the literature to assess the quality of the literature. It also corrects the shortcomings of the traditional JIF to judge the quality of the literature. Compared with a single impact indicator, it is more scientific.

Impact Indicators

The impact indicators selected for this study included the Cumulative Citation for 5 years (CC_5), JIF, Journal Citation Indicator (JCI), and Journal Cumulative Citation for 5 years (JCC_5). Among these, JIF is the total number of citations for scholarly articles published in the journal in the past 2 years divided by the total number of citable articles. JCI is the average Category Normalized Citation Impact (CNCI) of the citable literature published in the specific journal in the previous 3 years. The JCC_5 is obtained by dividing the sum of the citation frequencies recorded in the POCI repository of the corresponding research papers (focus papers) of the selected journals for the years 2018 to 2022 by the number of research papers published by the journal (as in Equation 4).

In Equation 4, a_iC_t represents the number of citations of the ith research paper of the journal in year t, and n is the number of research papers published by the journal.

Results

Evaluation and Correlation Analyses of Journals Under Different Evaluation Perspectives

Analysis of Differences in Academic Influence and Level of Disruptive Innovation of Journals

The academic impact and level of disruptive innovation of the selected 114 journals, the results of the correlation analyses of the indicators, and the differences in the rankings are shown in Table 1, Table 2, and Figure 1, respectively. From these, we can see that (1) journals that are at the top of the ranking of impact indicators are usually also ranked at the top in the ranking of disruptive innovation, (2) journals at the bottom of the impact rankings are usually at the bottom of the innovation rankings, and (3) there is little difference in the ranking results of journals under different impact indicators, but there is a big difference in the ranking results of influence and disruptive innovation. Although there are moderate correlations between the JDI of journals and the 3 influence indicators of JIF, JCI, and JCC_5, the average difference between the disruptive innovation ranking and academic influence ranking of the journals included in the study reached 20 places.

XSL•FO
RenderX

Table 1. Comparison of academic influence and disruptive innovation level of journals.

Journal	JCC_5[a]	JDI[b]	JIF[c]	JCI[d]
New England Journal of Medicine	337.90	1167.64	70.67	24.83
Lancet	259.17	883.11	59.10	18.05
Journal of the American Medical Association (JAMA)	110.67	467.92	51.27	13.08
Nature Reviews Disease Primers	245.25	466.91	32.27	9.48
Annals of Internal Medicine	144.84	293.23	19.32	6.37
British Medical Journal (BMJ)	85.27	278.50	27.60	5.57
JAMA Internal Medicine	54.80	235.91	20.77	6.03
International Journal of Evidence-Based Healthcare	24.94	165.18	1.16	0.95
Medical Journal of Australia	24.67	162.99	5.44	1.18
PLoS Medicine	49.55	158.76	11.05	3.91
Canadian Medical Association Journal	23.64	138.67	6.94	2.06
Journal of Travel Medicine	12.63	136.54	4.16	0.98
Journal of the Royal Society of Medicine	6.00	121.29	3.54	0.83
Amyloid	24.28	105.39	4.92	1.13
Medical Clinics of North America	20.90	104.36	2.72	1.22
Clinical Medicine	17.59	101.48	2.05	0.52
American Journal of Preventive Medicine	19.70	95.50	4.44	1.74
BMC Medicine	28.06	82.00	8.29	2.74
Annals of Family Medicine	17.24	78.87	4.19	2.31
Mayo Clinic Proceedings	20.34	71.01	7.09	2.47
American Journal of Chinese Medicine	14.53	65.77	3.51	1.42
British Journal of General Practice	15.76	64.69	4.43	1.49
Archives of Medical Science	11.89	62.18	2.38	0.88
Postgraduate Medical Journal	7.54	61.38	1.98	0.61
Journal of General Internal Medicine	18.38	61.18	4.61	1.69
Palliative Medicine	17.67	60.71	4.96	1.56
American Journal of Medicine	16.70	55.79	4.76	1.56
Deutsches Ärzteblatt International	20.13	55.74	4.47	1.59
Journal of the Chinese Medical Association	8.10	53.26	1.89	0.62
Journal of the National Medical Association	6.91	48.41	0.83	0.24
African Health Sciences	5.25	44.92	0.78	0.44
Journal of Urban Health	11.17	42.83	2.15	0.83
Canadian Family Physician	11.27	42.82	2.19	0.77
Family Practice	8.13	41.90	1.99	0.88
Journal of Pain and Symptom Management	13.62	39.72	3.38	1.22
Patient Preference and Adherence	8.88	38.48	2.10	0.69
Preventive Medicine	13.94	37.36	3.45	1.35
Saudi Medical Journal	5.90	34.58	1.06	0.35
Military Medicine	6.43	34.01	0.85	0.28
Journal of Korean Medical Science	7.79	33.23	1.72	0.62
Indian Journal of Medical Research	5.59	32.76	1.25	0.39
Sexual Medicine	7.66	31.84	1.44	0.53

XSL•FO
RenderX

Journal	JCC_5[a]	JDI[b]	JIF[c]	JCI[d]
Journal of the Royal Army Medical Corps	5.90	31.27	0.99	0.28
Journal of Evaluation in Clinical Practice	10.02	30.32	1.54	0.64
European Journal of General Practice	6.68	29.86	1.62	0.76
Journal of Women's Health	10.20	29.34	2.01	1.18
International Journal of Medical Sciences	10.61	29.30	2.33	0.81
European Journal of Internal Medicine	11.88	28.40	3.66	1.01
BMJ Open	11.00	28.20	2.38	0.83
Journal of the Formosan Medical Association	10.55	27.39	2.84	0.77
Pakistan Journal of Medical Sciences	3.64	26.87	0.83	0.27
Medicine	6.58	26.36	1.87	0.60
Chinese Medical Journal, Peking	9.08	25.93	2.10	0.85
Scandinavian Journal of Primary Health	6.49	25.93	1.56	0.49
International Journal of Clinical Practice	9.61	25.87	2.61	0.73
Singapore Medical Journal	5.55	25.68	1.14	0.37
BMC Family Practice	10.09	25.57	2.43	1.08
Journal of Internal Medicine	28.23	24.96	6.05	2.09
Journal of the American Academy of Physician Associates	3.91	24.76	0.46	0.14
Journal of Postgraduate Medicine	4.96	24.49	1.32	0.42
Journal of Hospital Medicine	9.25	24.46	2.28	0.79
Internal Medicine Journal	6.76	24.36	1.77	0.61
Pain Medicine	12.50	24.22	2.76	0.98
Frontiers in Medicine (Lausanne)	11.41	24.19	3.11	0.89
Irish Journal of Medical Science	5.26	23.52	1.03	0.37
Acta Clinica Croatica	4.28	23.25	0.40	0.14
Clinics	6.12	22.49	1.13	0.41
Internal and Emergency Medicine	10.29	21.32	2.34	0.80
Medicina (Lithuania)	5.25	21.31	1.47	0.39
Medical Principles and Practice	5.85	20.72	1.10	0.50
Open Medicine (Warsaw, Poland)	5.53	20.12	1.22	0.25
Korean Journal of Internal Medicine	8.34	19.75	2.71	0.70
Colombia Médica	4.33	17.23	0.98	0.24
Journal of Clinical Medicine	11.34	17.19	5.69	1.30
Journal of Nippon Medical School	2.39	17.11	0.62	0.19
Postgraduate Medicine	7.78	16.98	2.24	0.60
British Journal of Hospital Medicine	3.45	16.87	0.53	0.12
Yonsei Medical Journal	7.58	16.86	1.76	0.60
São Paulo Medical Journal	4.00	16.71	1.09	0.30
Upsala Journal of Medical Sciences	11.03	16.17	2.75	0.65
Current Medical Research and Opinion	8.00	15.73	2.35	0.83
American Journal of the Medical Sciences	7.71	15.52	1.96	0.63
Revista Clínica Española	5.34	15.19	1.04	0.40
QJM: An International Journal of Medicine	6.80	13.92	2.65	0.77

XSL•FO
RenderX

Journal	JCC_5[a]	JDI[b]	JIF[c]	JCI[d]
Revista da Associacao Medica Brasileira	3.78	13.73	0.80	0.23
Libyan Journal of Medicine	6.89	13.49	1.41	0.49
Translational Research	18.57	13.21	4.92	1.72
Atención Primaria	3.29	12.97	1.35	0.43
La Presse Médicale	4.16	12.95	1.17	0.29
Balkan Medical Journal	5.33	12.73	1.20	0.37
Internal Medicine	4.65	12.73	0.96	0.32
Journal of Investigative Medicine	8.00	12.54	1.99	0.71
Southern Medical Journal	4.12	12.42	0.87	0.33
Scottish Medical Journal	2.00	12.18	0.68	0.19
World Journal of Clinical Cases	4.74	12.01	1.15	0.45
Acta Clinica Belgica	4.98	11.76	0.96	0.27
Diagnostics	9.34	11.71	2.49	0.71
Annals of Saudi Medicine	5.22	11.08	0.81	0.29
Annals of Medicine	11.85	11.06	3.05	1.00
Medicina Clinica (Barcelona)	4.25	11.01	1.28	0.31
Journal of Research in Medical Sciences	5.63	10.95	1.47	0.40
European Journal of Clinical Investigation	12.57	10.76	2.78	0.92
Journal of Nepal Medical Association	1.27	10.23	0.21	0.07
Wiener Klinische Wochenschrift	6.66	10.10	1.17	0.39
Journal of Medical Economics	9.02	9.45	1.89	0.84
Tohoku Journal of Experimental Medicine	7.07	8.93	1.58	0.54
Journal of Cachexia, Sarcopenia and Muscle	41.24	6.90	10.75	3.08
National Medical Journal of India	1.68	5.92	0.64	0.18
Croatian Medical Journal	3.13	5.68	1.62	0.51
Disease-a-Month	8.44	5.13	0.94	0.34
Revue de Médecine Interne	2.97	5.02	0.81	0.28
Medizinische Klinik, Intensivmedizin und Notfallmedizin	3.34	3.52	0.85	0.31
Internist	1.58	2.67	0.43	0.18
Hippokratia	1.52	2.53	0.52	0.14

[a]JCC_5: Journal Cumulative Citation for 5 years.

[b]JDI: Journal Disruption Index.

[c]JIF: Journal Impact Factor.

[d]JCI: Journal Citation Indicator.

XSL•FO
RenderX

Table 2. Correlation analysis of journal evaluation indicators.

Journal evaluation indicators	JDI[a]	JCC_5[b]	JIF[c]	JCI[d]
JDI				
r	1	0.677	0.585	0.621
P value	__[e]	<.001	<.001	<.001
JCC_5				
r	0.677	1	0.909	0.935
P value	<.001	—	<.001	<.001
JIF				
r	0.585	0.909	1	0.955
P value	<.001	<.001	—	<.001
JCI				
r	0.621	0.935	0.955	1
P value	<.001	<.001	<.001	—

[a]JDI: Journal Disruption Index.

[b]JCC_5: Journal Cumulative Citation for 5 years.

[c]JIF: Journal Impact Factor.

[d]JCI: Journal Citation Indicator.

[e]Not applicable.

Figure 1. Comparison of ranking differences based on different evaluation indicators. JCC5: Journal Cumulative Citation for 5 years; JCI: Journal Citation Indicator; JDI: Journal Disruption Index; JIF: Journal Impact Factor.

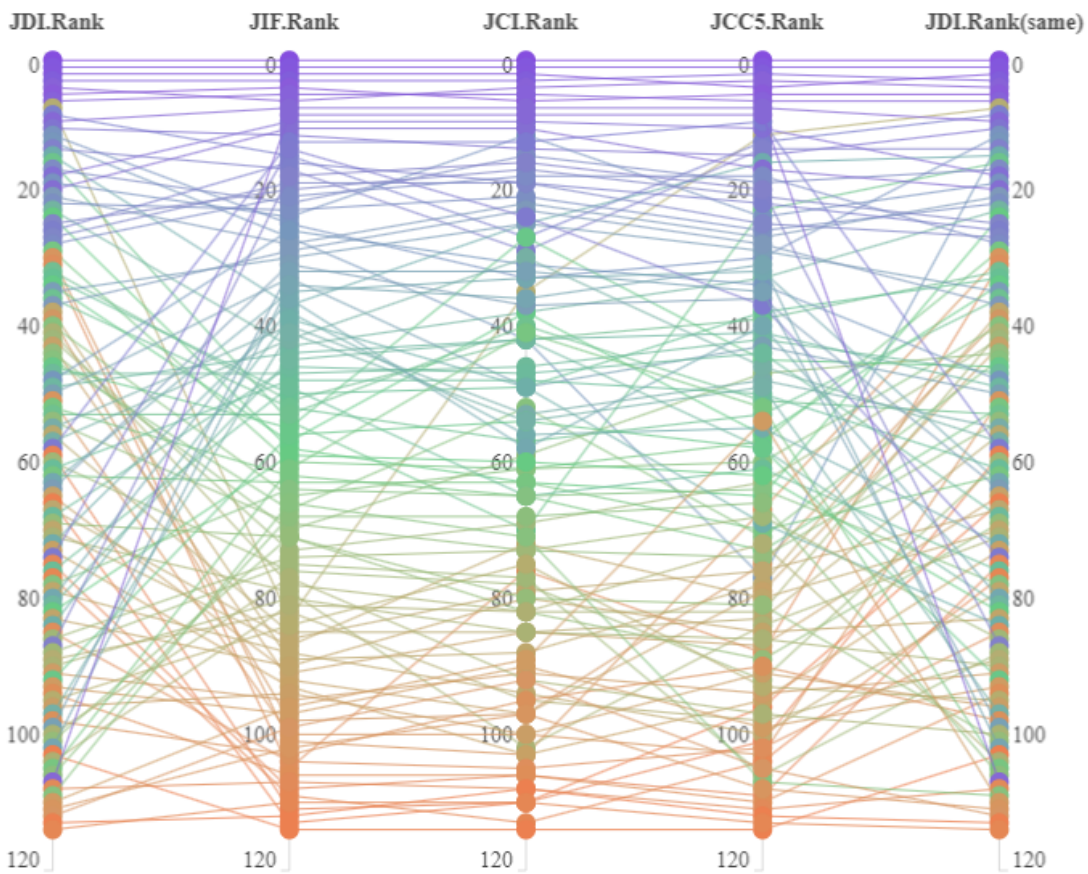

XSL•FO
RenderX

Analysis of the Differences Between the Journals' Impact and Disruption and the Results of Peer Review

In order to better analyze the differences between the academic impact and level of disruptive innovation with peer-reviewed results of journals in the field of general and internal medicine, papers indexed in H1 Connect were referred to as "H-papers," and the source journals of "H-papers" are referred to as "H-journals." The evaluation indicators, ranking of H-journals, and the percentage of H-papers are shown in Table 3 and Table 4. We can see that (1) the top 7 journals in terms of both academic impact and disruptive innovation are all H-journals, (2) the average impact ranking of H-journals is higher than the average innovation ranking, and (3) some journals with low impact and innovation also became H-journals.

Table 3. Evaluation indicators of H-journals (sources of papers indexed in the H1 Connect database).

Journal	Count, n	JPScore[a]	JPTime_w[b]	JFPStar_w[c]
New England Journal of Medicine	112	2192.2	234	528
Lancet	43	574.1	65	142
Journal of the American Medical Association	27	313.7	36	70
Annals of Internal Medicine	11	118.8	14	25
Nature Reviews Disease Primers	4	50.3	6	11
British Medical Journal	11	115.8	18	30
Sexual Medicine	3	14.1	7	7
PLoS Medicine	11	103.5	15	25
JAMA Internal Medicine	6	48.2	6	10
International Journal of Clinical Practice	3	23.2	3	5
Journal of Cachexia, Sarcopenia and Muscle	2	14	2	3
American Journal of Medicine	3	18.6	4	5
BMC Medicine	4	24.6	6	7
Journal of Investigative Medicine	1	4.7	1	4
Medical Clinics of North America	1	4.9	1	1
Canadian Medical Association Journal	1	9.4	1	2
Medicina Clinica	1	4.6	1	1
Journal of Women's Health	2	9.5	2	2
Wiener Klinische Wochenschrift	1	4.6	1	1
Mayo Clinic Proceedings	1	4.6	1	1
Frontiers in Medicine	1	4.7	1	1
Current Medical Research and Opinion	1	4.7	1	1
Pain Medicine	1	4.7	1	1
Preventive Medicine	1	9.4	1	2
BMJ Open	5	28.3	5	6
Journal of Clinical Medicine	1	4.6	1	1
Medicine	5	28	5	6

[a]JPScore: journal peer review score.

[b]JPTime_w: weighted journal peer review evaluation time.

[c]JFPStar_w: weighted journal peer review star.

XSL•FO
RenderX

Table 4. Ranking and percentage of H-papers (papers indexed in the H1 Connect database) in H-journals (sources of H-papers).

Journal	JDI[a] ranking	JIF[b] ranking	JCI[c] ranking	H-papers, n (%)
New England Journal of Medicine	1	1	1	112 (40.58)
Lancet	2	2	2	43 (22.51)
Journal of the American Medical Association	3	3	3	27 (18.24)
Annals of Internal Medicine	5	7	5	11 (11.46)
Nature reviews. Disease primers	4	4	4	4 (9.30)
British Medical Journal	6	5	7	11 (7.97)
Sexual Medicine	42	73	69	3 (7.14)
PLoS Medicine	10	8	8	11 (6.11)
JAMA Internal Medicine	7	6	6	6 (5.50)
International Journal of Clinical Practice	55	40	53	3 (4.62)
Journal of Cachexia, Sarcopenia and Muscle	107	9	9	2 (2.56)
American Journal of Medicine	27	19	19	3 (2.22)
BMC Medicine	18	10	10	4 (1.99)
Journal of Investigative Medicine	92	56	54	1 (1.52)
Medical Clinics of North America	15	37	25	1 (1.47)
Canadian Medical Association Journal	11	12	14	1 (1.28)
Medicina Clinica	100	77	92	1 (1.20)
Journal of Women's Health	46	55	27	2 (1.19)
Wiener Klinische Wochenschrift	104	81	82	1 (1.10)
Mayo Clinic Proceedings	20	11	11	1 (0.94)
Frontiers in Medicine	64	31	37	1 (0.53)
Current Medical Research and Opinion	81	45	42	1 (0.50)
Pain Medicine	63	35	33	1 (0.48)
Preventive Medicine	37	29	23	1 (0.39)
BMJ Open	49	44	42	5 (0.24)
Journal of Clinical Medicine	74	14	24	1 (0.23)
Medicine	52	62	65	5 (0.16)

[a]JDI: Journal Disruption Index.

[b]JIF: Journal Impact Factor.

[c]JCI: Journal Citation Indicator.

Evaluation and Correlation Analyses of Papers Under Different Evaluation Perspectives

Analysis of Differences in Academic Impact and Level of Disruptive Innovation of Papers

Ideally, if an article is accepted by a specific journal, it is because its overall quality is similar to other papers previously published in that journal [47]. However, journal-level indicators are, at best, only moderately suggestive of the quality of an article [48], which makes indicators that measure specific articles more popular [49]. Therefore, in this study, research papers in the field of general and internal medicine were also evaluated in terms of their academic impact and level of disruptive innovation, and the results are shown in Figure 2. From the results, we can see that research papers that rank high in the impact ranking usually also rank high in the disruptive innovation ranking. There were 8 (8/15, 53%), 96 (96/150, 64%), and 880 (880/1500, 58.67%) of the Top 0.1%, Top 1%, and Top 10% papers, respectively, selected based on the Dz and CC_5 that were the same. Second, the level of disruptive innovation of research papers with the same level of impact varied greatly, and the impact level of research papers with the same level of disruptive innovation also varied greatly. Third, despite the high correlation (r=0.635, P<.001) between the Dz and CC_5 of the selected research papers, the average difference between their innovation and impact rankings reached about 2700. Fourth, the actual analysis results showed no correlation between the innovation of the selected research paper and the number of references, which indicates that the Dz index is

XSL•FO
RenderX

basically unaffected by the difference in the number of references in the actual evaluation process (r=0.006, P=.43).

Figure 2. Comparison of the ranking of the absolute disruption index (Dz) and Cumulative Citation for 5 years (CC5) of selected research papers.

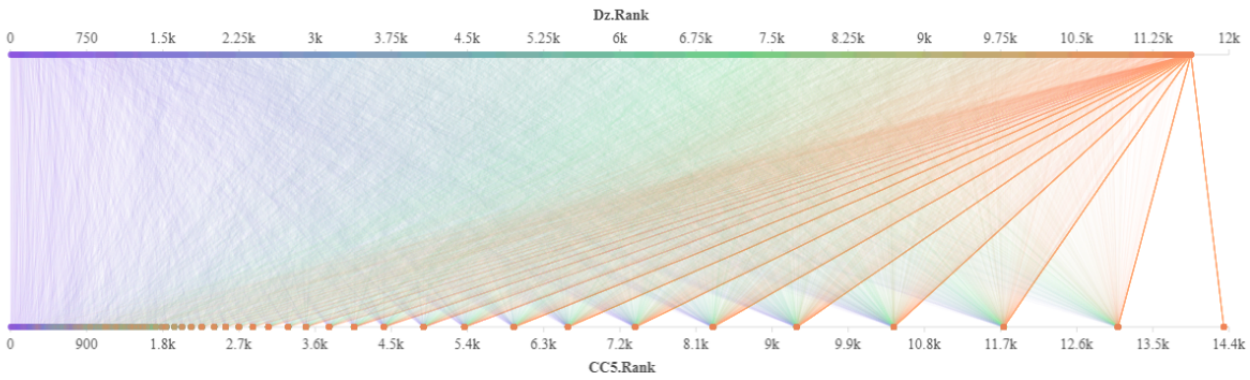

Analysis of Differences Between Papers' Impact and Disruption and the Results of Peer Review

In order to better analyze the differences between the academic impact, level of disruptive innovation, and peer review results of journal research papers in the field of general and internal medicine, the differences between the academic impact and disruptive innovation level and the peer review results of H-papers were analyzed. The specific results are shown in Table 5 and Figure 3. From the results, we see that there were 8 (8/15, 53%), 65 (65/150, 43.3%), and 187 (187/1500, 12.47%) H-papers among the top 0.1%, top 1%, and top 10% papers, respectively, selected based on Dz. There were 5 (5/15, 33%), 74 (74/150, 49.3%), and 220 (220/1500, 14.67%) H-papers among the top 0.1%, top 1%, and top 10% papers, respectively, selected based on CC_5. Second, there was a significant positive correlation between the peer review indicators, disruptive innovation indicators, and academic impact indicators of H-papers, reflecting the consistency between the quantitative evaluation and peer review at the overall level. Third, the average impact ranking of H-papers was 865 (top 5.68%), and the average disruptive innovation ranking was 1726 (top 11.35%), which means that the average impact ranking of H-papers was higher than the average disruptive innovation ranking. Fourth, some papers with low academic impact and disruptive innovation level also became H-papers. Fifth, compared with the CC_5, Dz has a minor correlation advantage with PTime_w and PStar_w.

XSL•FO
RenderX

Table 5. Validation of the correlation between 3 kinds of indicators of H-papers (papers indexed in the H1 Connect database).

Indicators	CC_5[a]	Dz[b]	PTime_w[c]	PStar_w[d]	PScore[e]
CC_5					
r	1	0.749	0.292	0.385	0.613
P value	__[f]	<.001	<.001	<.001	<.001
Dz					
r		1	0.322	0.393	0.544
P value	<.001	—	<.001	<.001	<.001
PTime_w					
r	0.292	0.322	1	0.751	0.534
P value	<.001	<.001	—	<.001	<.001
PStar_w					
r	0.385	0.393	0.751	1	0.886
P value	<.001	<.001	<.001	—	<.001
PScore					
r	0.613	0.544	0.534	0.886	1
P value	<.001	<.001	<.001	<.001	—

[a]CC_5: Cumulative Citation for 5 years.

[b]Dz: absolute disruption index.

[c]PTime_w: weighted peer review evaluation times.

[d]PStar_w: weighted peer review stars.

[e]PScore: peer review score.

[f]Not applicable.

Figure 3. Comparison of ranking differences between H-papers (papers indexed in the H1 Connect database) in general and internal medicine based on different evaluation indicators. CC5: Cumulative Citation for 5 years; Dz: absolute disruption index; PStar: peer review stars.

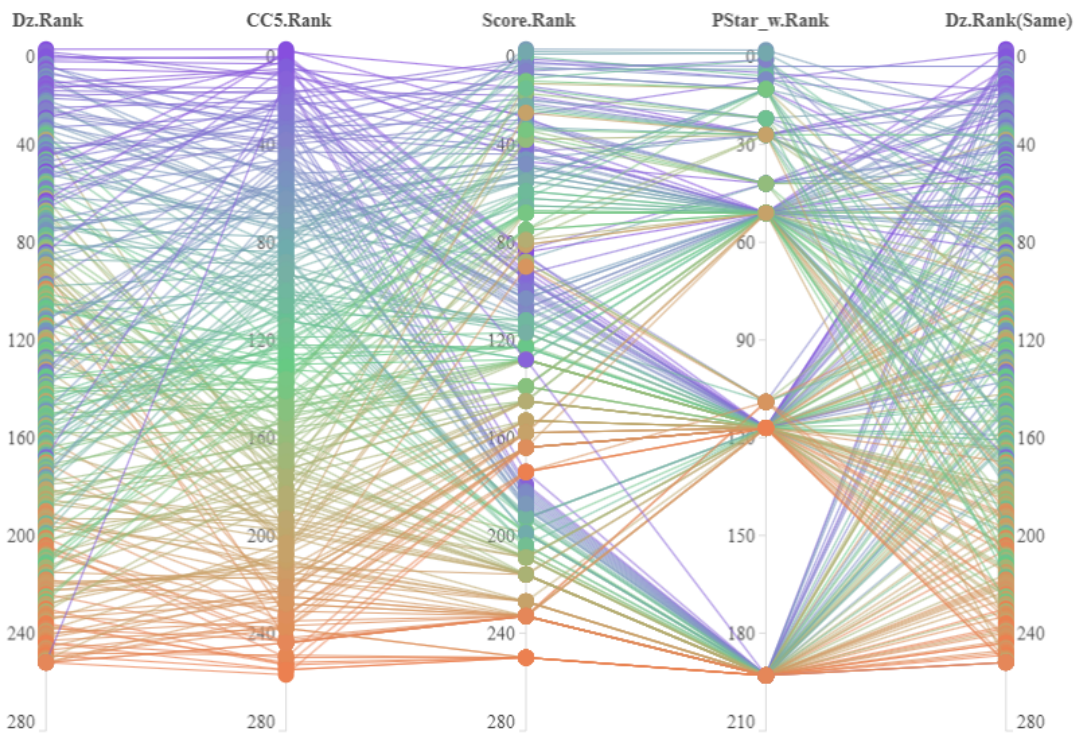

XSL•FO
RenderX

In addition to rating and commenting on the included research papers, different labels are added by the reviewers of articles in H1 Connect according to their different internal characteristics. The relevant definitions are shown in Table 6 (refer to Du et al [50]). We categorized and counted the 257 H-papers (in this study, if a paper had more than 1 tag, it was counted separately in the calculation of each tag), and the results are shown in Table 7. From this, we can see that (1) research papers with the tags in the "Novel Drug Target," "Technical Advance," and "New Finding" categories had a high academic impact and a high disruptive innovation level; (2) research papers with the "Changes Clinical Practice" tag showed only a moderate academic impact and disruptive innovation level but had high levels of peer-reviewed recognition and attention; (3) experts showed higher recognition and concern for research papers with tags of "Negative/Null Result," "Controversial," "Refutation," and "Interesting Hypothesis," but their academic impact and disruptive innovation level were lower than others.

Table 6. Categorization and definition of peer review tags.

Tag name	Definition
Confirmation	The article confirms previously published data or hypotheses.
Controversial	The article challenges established dogma.
Changes clinical practice	The article recommends that clinicians make complete, concrete and immediate changes in their practice.
Good for teaching	The article is a key article of significance in the field.
Interesting hypothesis	The article presents a meaningful model or hypothesis.
Negative/null result	The article has invalid or negative results
New finding	The article proposes a new data model or hypothesis.
Novel drug target	The article proposes a new drug target.
Refutation	The article refutes previously published data or hypotheses
Technical advance	The article introduces a new practical/theoretical technique, or a new use of an existing technique.

Table 7. Details of research papers with different peer-reviewed tags.

Type	Count, n	CC_5[a]	Dz[b]	PScore[c]	PStar_w[d]	PTime_w[e]
Novel drug target	7	511.57	24.50	16.84	4.43	1.57
Technical advance	16	577.44	19.69	19.98	5.25	2.38
New finding	185	336.48	9.80	15.31	3.60	1.74
Confirmation	60	315.47	8.49	19.72	4.63	2.02
Good for teaching	82	248.77	7.47	18.12	4.63	2.22
Changes clinical practice	6	241.67	4.96	22.50	6.00	2.83
Negative/null result	27	281.81	3.84	17.81	4.52	2.30
Controversial	33	227.06	3.83	25.41	6.48	2.73
Refutation	4	167.75	2.94	22.53	6.00	3.25
Interesting hypothesis	37	207.86	1.77	20.68	5.38	2.24

[a]CC_5: Cumulative Citation for 5 years.

[b]Dz: absolute disruption index.

[c]PScore: peer review score.

[d]PStar_w: weighted peer review stars.

[e]PTime_w: weighted peer review evaluation times.

Discussion

Principal Findings

Evaluation of Disruptive Innovation Is Detached From the Traditional Evaluation System

From these research results, large differences were seen between the innovation and impact rankings of the journals and research papers included in the study. This is also consistent with the findings of Guo and Zhou [51]. This phenomenon reflects the essential difference between the 2 different evaluation systems. It also proves that the evaluation of the disruptive innovation of research papers and journals based on Dz and JDI is not consistent with the traditional evaluation system of impact.

The essence of disruptive evaluation is to measure the innovation from the substitution level of knowledge structure. This evaluation method brings new ideas to the field of scientific

XSL•FO
RenderX

research evaluation, helps relevant institutions and scholars remove the constraints of the traditional evaluation system, and helps to establish a value orientation of encouraging innovation for scientific research and scientific and technological journals, so as to promote the benign development of the academic ecology.

The 3 Evaluation Systems Only Have High Consensus on the Top Object

Although, given the consistency of scientific evaluation, there will be some uniformity in the level of disruptive innovation, level of academic impact, and peer review results of journals as well as research papers.

However, from the research results presented in this paper, we can see that (1) the top 7 journals in terms of both academic impact and disruptive innovation were all H-journals, (2) more than one-half of the top papers selected based on Dz and CC_5 were the same, (3) the average H-papers were ranked at the top in terms of impact and innovation, and (4) the results of the different evaluation systems only had high consensus on the authoritative journals and top papers in the field.

These findings are also consistent with those of Goman [52] and Chi et al [53]. A fundamental reason for this phenomenon is that the purposes of the 3 evaluation systems are inherently different. Therefore, in the actual evaluation process, the 3 kinds of indicators are not interchangeable, and the combination of the 3 evaluation systems may be a feasible solution to establish a comprehensive, scientific, and reasonable journal evaluation system.

A Single Indicator Cannot Accurately Reflect the Impact of Medical Research on Clinical Practice Alone

From the aforementioned findings, we can see that research papers of the "Novel Drug Target," "Technical Advance," and "New Finding" types have high academic impact and high levels of innovation, which is in line with their classification definitions. This is also consistent with the findings of Du et al [50], Thelwall et al [54], and Jiang and Liu [55]. Second, research papers of the "Changes Clinical Practice" type showed only moderate levels of innovation and impact but had high levels of peer-reviewed recognition and attention. This reflects the difficulty of evaluating the impact of a particular academic paper on clinical practice, whether the evaluation system is based on academic impact or level of disruptive innovation. Third, peer-reviewed experts show higher recognition and concern for research papers of the types "Negative/Null Result," "Controversial," "Refutation," and "Interesting Hypothesis," but the level of impact and disruptive innovation of these papers are lower. This partly reflects the current academic community's excessive focus on positive results; deliberate avoidance of negative results; and overall lack of support for debatable, falsified, and other types of research.

Several scholars have recently suggested that the evaluation system of medical journals should be redesigned for contemporary clinical impact and utility [56]. In this regard, Thelwall and Maflahi [57] advocated that the references of a guideline are an important basis for studying clinical value. However, Traylor and Herrmann-Lingen [58] found only a weak correlation between the number of citations to individual journals in the guidelines and their respective JIF. Therefore, the JIF is not a suitable tool for assessing the clinical relevance of medical research. The results of this study similarly found that research papers of the "Changes Clinical Practice" type showed only moderate levels of disruptive innovation and academic impact, but such research papers received higher recognition and attention from peer review experts. Therefore, combining quantitative evaluation with peer review may be a feasible way to measure the impact of medical research on clinical practice.

Limitations

However, this study also has the following limitations. First, papers in the medical field have a preference for citing review articles [59], which has a certain impact on the evaluation of the disruptive innovation of research papers. Second, the scoring mechanism provided to its reviewers by H1 Connect has a low differentiation degree and cannot perfectly distinguish the differences in quality between papers yet. In addition, H1 Connect has too few evaluators. Brezis and Birukou [60] illustrated that, if the number of reviewers is increased to about 10, the correlation between the results and the quality of the paper will be significantly improved. However, it is difficult to seek so many high-quality experts who are willing to accept open peer review in the high pressure environment of "Publish or Perish" [61]. Third, since the citation data sources used in this study are all based on PubMed data, this study also suffers from the problem of missing references and citations that are not labeled with a PMID, which affects the accuracy of the evaluation results to a certain extent. In future studies, we will obtain more accurate measurement results by jointly using multiple sources of citation data.

Conclusions

In this study, the general and internal medicine journals indexed in SCIE in 2018 were chosen as the study object. The POCI and H1 Connect databases were selected as sources of citation relationship data and peer review data. We investigated the connections and contrasts between the academic impact, level of disruptive innovation, and results of peer review for medical journals and published research papers. We also analyzed the similarities and differences as well as the fundamental causes of the varying evaluation results in terms of impact evaluation, disruptive innovation, and peer review for various types of medical research papers.

The results of this study show that the evaluation of scientific research based on the innovation index is detached from the traditional impact evaluation system, the 3 evaluation systems only have high consistency for authoritative journals and top papers, and neither the single impact index nor the innovation index can directly reflect the impact of medical research on clinical practice.

In addition, with the increasing importance of replicative science, the accuracy of statistical reports, evidential value of reported data, and replicability of given experimental results [62] should also be included in the examination of journal

XSL•FO
RenderX

quality. How to establish a comprehensive, all-encompassing, scientific, and reasonable journal evaluation system needs to be further investigated.

Acknowledgments

XL was supported by the National Social Science Foundation of China (23BTQ085), Major Project of Basic Research on Philosophy and Social Sciences in Henan Universities (2024-JCZD-23).

Data Availability

All citation data can be obtained from [63].

Authors' Contributions

XL conceptualized the study, reviewed and edited the manuscript, and supervised the study. YJ was responsible for the methodology, provided the software and other resources, performed the formal analysis, and wrote the original manuscript draft. YJ and XL acquired funding. YJ, XY, and ZZ curated the data. XY and ZZ also contributed to the study investigation.

Conflicts of Interest

None declared.

References

1. Garfield E. Citation analysis as a tool in journal evaluation. Science 1972 Nov 03;178(4060):471-479. [doi: 10.1126/science.178.4060.471] [Medline: 5079701]
2. Hansson S. Impact factor as a misleading tool in evaluation of medical journals. Lancet 1995 Sep 30;346(8979):906. [doi: 10.1016/s0140-6736(95)92749-2] [Medline: 7564699]
3. Metze K, Borges da Silva FA. Ranking of journals by journal impact factors is not exact and may provoke misleading conclusions. J Clin Pathol 2022 May 11;75(10):649-650. [doi: 10.1136/jclinpath-2021-208051] [Medline: 35545355]
4. Siler K, Larivière V. Who games metrics and rankings? Institutional niches and journal impact factor inflation. Research Policy 2022 Dec;51(10):104608. [doi: 10.1016/j.respol.2022.104608]
5. Martin BR. Editors' JIF-boosting stratagems – Which are appropriate and which not? Research Policy 2016 Feb;45(1):1-7. [doi: 10.1016/j.respol.2015.09.001]
6. Zaidi SJA, Taqi M. Citation cartels in medical and dental journals. J Coll Physicians Surg Pak 2023 Jun 01;33(6):700-701. [doi: 10.29271/jcpsp.2023.06.700] [Medline: 37300268]
7. Fassin Y. Research on Covid-19: a disruptive phenomenon for bibliometrics. Scientometrics 2021 May 07;126(6):5305-5319 [FREE Full text] [doi: 10.1007/s11192-021-03989-w] [Medline: 33994601]
8. Buela-Casal G, Zych I. What do the scientists think about the impact factor? Scientometrics 2012 Feb 22;92(2):281-292. [doi: 10.1007/s11192-012-0676-y]
9. Thelwall M, Kousha K, Stuart E, Makita M, Abdoli M, Wilson P, et al. In which fields are citations indicators of research quality? Asso for Info Science & Tech 2023 May 04;74(8):941-953. [doi: 10.1002/asi.24767]
10. Park M, Leahey E, Funk R. Papers and patents are becoming less disruptive over time. Nature 2023 Jan;613(7942):138-144. [doi: 10.1038/s41586-022-05543-x] [Medline: 36600070]
11. Chu JSG, Evans JA. Slowed canonical progress in large fields of science. Proc Natl Acad Sci U S A 2021 Oct 12;118(41):1 [FREE Full text] [doi: 10.1073/pnas.2021636118] [Medline: 34607941]
12. Schattner U, Shatner M. Science desperately needs disruptive innovation. Qeios 2023 Jun 07:1. [doi: 10.32388/KY5Y9L.2]
13. Liang Z, Mao J, Li G. Bias against scientific novelty: A prepublication perspective. Asso for Info Science & Tech 2022 Nov 19;74(1):99-114. [doi: 10.1002/asi.24725]
14. Osterloh M, Frey BS. Ranking games. Eval Rev 2014 Aug 04;39(1):102-129. [doi: 10.1177/0193841x14524957]
15. Kulczycki E, Huang Y, Zuccala AA, Engels TCE, Ferrara A, Guns R, et al. Uses of the Journal Impact Factor in national journal rankings in China and Europe. Asso for Info Science & Tech 2022 Aug 13;73(12):1741-1754. [doi: 10.1002/asi.24706]
16. McKiernan E, Schimanski L, Muñoz Nieves C, Matthias L, Niles M, Alperin J. Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. Elife 2019 Jul 31;8:8 [FREE Full text] [doi: 10.7554/eLife.47338] [Medline: 31364991]
17. Heibi I, Peroni S, Shotton D. Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. Scientometrics 2019 Sep 14;121(2):1213-1228. [doi: 10.1007/s11192-019-03217-6]
18. Tushman ML, Anderson P. Technological discontinuities and organizational environments. Administrative Science Quarterly 1986 Sep;31(3):439. [doi: 10.2307/2392832]
19. Christensen CM. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail (Management of Innovation and Change). Boston, MA: Harvard Business Review Press; 1997.

XSL•FO
RenderX

20. Funk RJ, Owen-Smith J. A dynamic network measure of technological change. Management Science 2017 Mar;63(3):791-817. [doi: 10.1287/mnsc.2015.2366]
21. Govindarajan V, Kopalle PK. Disruptiveness of innovations: measurement and an assessment of reliability and validity. Strategic Management Journal 2005 Dec 22;27(2):189-199. [doi: 10.1002/smj.511]
22. Wu L, Wang D, Evans JA. Large teams develop and small teams disrupt science and technology. Nature 2019 Feb 13;566(7744):378-382. [doi: 10.1038/s41586-019-0941-9] [Medline: 30760923]
23. Kostoff RN, Boylan R, Simons GR. Disruptive technology roadmaps. Technological Forecasting and Social Change 2004 Jan;71(1-2):141-159. [doi: 10.1016/s0040-1625(03)00048-9]
24. Bornmann L, Devarakonda S, Tekles A, Chacko G. Disruptive papers published in Scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). Scientometrics 2020 Mar 14;123(2):1149-1155. [doi: 10.1007/s11192-020-03406-8]
25. Horen S, Hansdorfer M, Kronshtal R, Dorafshar A, Becerra A. The most disruptive publications in craniofacial surgery (1954-2014). J Craniofac Surg 2021 Oct 01;32(7):2426-2430. [doi: 10.1097/SCS.0000000000007804] [Medline: 34054087]
26. Sullivan GA, Skertich NJ, Gulack BC, Becerra AZ, Shah AN. Shifting paradigms: The top 100 most disruptive papers in core pediatric surgery journals. J Pediatr Surg 2021 Aug;56(8):1263-1274. [doi: 10.1016/j.jpedsurg.2021.02.002] [Medline: 33678406]
27. Meyer C, Nakamura Y, Rasor BJ, Karim AS, Jewett MC, Tan C. Analysis of the innovation trend in cell-free synthetic biology. Life (Basel) 2021 Jun 11;11(6):551 [FREE Full text] [doi: 10.3390/life11060551] [Medline: 34208358]
28. Jiang Y, Liu X. A bibliometric analysis and disruptive innovation evaluation for the field of energy security. Sustainability 2023 Jan 05;15(2):969. [doi: 10.3390/su15020969]
29. Becerra A, Grimes C, Grunvald M, Underhill J, Bhama A, Govekar H, et al. A new bibliometric index: the top 100 most disruptive and developmental publications in colorectal surgery journals. Dis Colon Rectum 2022 Mar 01;65(3):429-443. [doi: 10.1097/DCR.0000000000002118] [Medline: 34108364]
30. Sheth AH, Hengartner A, Abdou H, Salehi PP, Becerra AZ, Lerner MZ. Disruption index in otolaryngology: uncovering a bibliometric history of a rapidly evolving field. Laryngoscope 2024 Feb 19;134(2):629-636. [doi: 10.1002/lary.30883] [Medline: 37466290]
31. Dilday J, Gallagher S, Bram R, Williams E, Grigorian A, Matsushima K, et al. Citation versus disruption in the military: Analysis of the top disruptive military trauma research publications. J Trauma Acute Care Surg 2023 May 15;95(2S):S157-S169. [doi: 10.1097/ta.0000000000004009]
32. Grunvald M, Williams M, Rao R, O'Donoghue C, Becerra A. 100 disruptive publications in breast cancer research. Asian Pac J Cancer Prev 2021 Aug 01;22(8):2385-2389. [doi: 10.31557/apjcp.2021.22.8.2385]
33. Abu-Omar A, Kennedy P, Yakub M, Robbins J, Yassin A, Verma N, et al. Extra credit for disruption: trend of disruption in radiology academic journals. Clin Radiol 2022 Dec;77(12):893-901. [doi: 10.1016/j.crad.2022.07.003] [Medline: 36150935]
34. Patel PA, Javed Ali M. Characterizing innovation in science through the disruption index. Semin Ophthalmol 2022 Aug 18;37(6):790-791. [doi: 10.1080/08820538.2022.2112851] [Medline: 35979757]
35. Neubauer DC, Blum JD, Labou SG, Heskett KM, Calvo RY, Reid CM, et al. Using the disruptive score to identify publications that changed plastic surgery practice. Ann Plast Surg 2022;88(4):S385-S390. [doi: 10.1097/sap.0000000000003144]
36. Khusid JA, Gupta M, Sadiq AS, Atallah WM, Becerra AZ. Changing the status quo: the 100 most-disruptive papers in urology? Urology 2021 Jul;153:56-68. [doi: 10.1016/j.urology.2020.10.073] [Medline: 33412220]
37. Williams MD, Grunvald MW, Skertich NJ, Hayden DM, O'Donoghue C, Torquati A, et al. Disruption in general surgery: Randomized controlled trials and changing paradigms. Surgery 2021 Dec;170(6):1862-1866. [doi: 10.1016/j.surg.2021.05.011] [Medline: 34340818]
38. Bornmann L, Devarakonda S, Tekles A, Chacko G. Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. Quantitative Science Studies 2020 Aug;1(3):1242-1259. [doi: 10.1162/qss_a_00068]
39. Bu Y, Waltman L, Huang Y. A multidimensional framework for characterizing the citation impact of scientific publications. Quantitative Science Studies 2021;2(1):155-183. [doi: 10.1162/qss_a_00109]
40. Ruan X, Lyu D, Gong K, Cheng Y, Li J. Rethinking the disruption index as a measure of scientific and technological advances. Technological Forecasting and Social Change 2021 Nov;172:121071. [doi: 10.1016/j.techfore.2021.121071]
41. Liu X, Shen Z, Liao Y, Yang L. The research about the improved disruption index and its influencing factors. Library and Information Service 2020;64(24):84-91. [doi: 10.13266/j.issn.0252-3116.2020.24.010]
42. Liu X, Shen Z, Liao Y, Zhu M, Yang L. Research on the stable time window of disruption index. Library and Information Service 2021;65(18):49-57. [doi: 10.13266/j.issn.0252-3116.2021.18.006]
43. Liu X, Liao Y, Zhu M. A preliminary study on the application of disruption index to scientific research evaluation. Information Studies: Theory & Application 2021;44(12):34-40. [doi: 10.16353/j.cnki.1000-7490.2021.12.005]
44. Jiang Y, Liu X. A construction and empirical research of the journal disruption index based on open citation data. Scientometrics 2023 May 18;128(7):3935-3958 [FREE Full text] [doi: 10.1007/s11192-023-04737-y] [Medline: 37287879]

XSL•FO
RenderX

45. Jiang Y, Wang L, Liu X. Innovative evaluation of Chinese SCIE-indexed scientific journals: An empirical study based on the disruption index. Chinese Journal of Scientific and Technical Periodicals 2023;34(8):1060-1068. [doi: 10.11946/cjstp.202302190096]
46. Bornmann L, Tekles A. Disruption index depends on length of citation window. EPI 2019 Mar 25;28(2):1. [doi: 10.3145/epi.2019.mar.07]
47. Migheli M, Ramello GB. The unbearable lightness of scientometric indices. Manage Decis Econ 2021 Nov 04;42(8):1933-1944. [doi: 10.1002/mde.3486]
48. Thelwall M, Kousha K, Makita M, Abdoli M, Stuart E, Wilson P, et al. In which fields do higher impact journals publish higher quality articles? Scientometrics 2023 May 18;128(7):3915-3933. [doi: 10.1007/s11192-023-04735-0]
49. Iyengar KP, Vaishya R. Article-level metrics: A new approach to quantify reach and impact of published research. J Orthop 2023 Jun;40:83-86. [doi: 10.1016/j.jor.2023.05.001] [Medline: 37215294]
50. Du J, Tang X, Wu Y. The effects of research level and article type on the differences between citation metrics and 1000 recommendations. Asso for Info Science & Tech 2015 Jun;67(12):3008-3021. [doi: 10.1002/asi.23548]
51. Guo L, Zhou Q. Research on characteristics of disruptive papers and their correlation with traditional bibliometric indicators. Journal of Intelligence 2021;40(1):183-207. [doi: 10.3969/j.issn.1002-1965.2021.01.030]
52. Gorman DM, Huber C. Ranking of addiction journals in eight widely used impact metrics. J Behav Addict 2022 Jul 13;11(2):348-360 [FREE Full text] [doi: 10.1556/2006.2022.00020] [Medline: 35895608]
53. Chi P, Ding J, Leng F. Characteristics of the technology impact of breakthrough papers in biology and medicine. Journal of the China Society for Scientific and Technical Information 2022;41(7):675. [doi: 10.3772/j.issn.1000-0135.2022.07.001]
54. Thelwall M, Kousha K, Abdoli M. Is medical research informing professional practice more highly cited? Evidence from AHFS DI Essentials in drugs.com. Scientometrics 2017 Feb 21;112(1):509-527. [doi: 10.1007/s11192-017-2292-3]
55. Jiang Y, Liu X. The relationship between absolute disruption index, peer review index and CNCI: a study based on virology papers. Library and Information Service 2023;67(3):96-105. [doi: 10.13266/j.issn.0252-3116.2023.03.009]
56. Klein-Fedyshin M, Ketchum A. PubMed's core clinical journals filter: redesigned for contemporary clinical impact and utility. J Med Libr Assoc 2023 Jul 10;111(3):665-676 [FREE Full text] [doi: 10.5195/jmla.2023.1631] [Medline: 37483370]
57. Thelwall M, Maflahi N. Guideline references and academic citations as evidence of the clinical value of health research. Asso for Info Science & Tech 2015 Mar 17;67(4):960-966. [doi: 10.1002/asi.23432]
58. Traylor C, Herrmann-Lingen C. Does the journal impact factor reflect the impact of German medical guideline contributions? Scientometrics 2023 Feb 03;128(3):1951-1962. [doi: 10.1007/s11192-023-04647-z]
59. Marks MS, Marsh MC, Schroer TA, Stevens TH. An alarming trend within the biological/biomedical research literature toward the citation of review articles rather than the primary research papers. Traffic 2013 Jan 19;14(1):1-1 [FREE Full text] [doi: 10.1111/tra.12023] [Medline: 23163598]
60. Brezis ES, Birukou A. Arbitrariness in the peer review process. Scientometrics 2020 Feb 03;123(1):393-411. [doi: 10.1007/s11192-020-03348-1]
61. Jiang Y, Liu X. Open peer review: Mode, technology, problems, and countermeasures. Chinese Journal of Scientific and Technical Periodicals 2022;33(9):1196-1205. [doi: 10.11946/cjstp.202202280120]
62. Dougherty MR, Horne Z. Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences. R Soc Open Sci 2022 Aug 17;9(8):220334 [FREE Full text] [doi: 10.1098/rsos.220334] [Medline: 35991336]
63. COCI CSV dataset of all the citation data. Figshare. 2023 Feb 07. URL: https://figshare.com/articles/dataset/Crossref_Open_Citation_Index_CSV_dataset_of_all_the_citation_data/6741422?file=37483791 [accessed 2024-05-24]

Abbreviations

CC5: Cumulative Citation for 5 years
CD: consolidation-disruption
CNCI: Category Normalized Citation Impact
D: disruption
Dz: absolute disruption index
JCC5: Journal Cumulative Citation for 5 years
JCI: Journal Citation Indicator
JCR: Journal Citation Reports
JDI: Journal Disruption Index
JIF: Journal Impact Factor
POCI: the OpenCitations Index of PubMed open PMID-to-PMID citations
PScore: peer review score
PStar_w: weighted peer review star
PTime_w: weighted peer review time
RDF: Resource Description Framework

XSL•FO
RenderX

SCIE: Science Citation Index Expanded
WoS: Web of Science

Edited by S Ma; submitted 04.12.23; peer-reviewed by O Beiki, C Oo; comments to author 28.02.24; revised version received 16.04.24; accepted 29.04.24; published 31.05.24.

Please cite as:
Jiang Y, Liu XL, Zhang Z, Yang X
Evaluation and Comparison of Academic Impact and Disruptive Innovation Level of Medical Journals: Bibliometric Analysis and Disruptive Evaluation
J Med Internet Res 2024;26:e55121
URL: https://www.jmir.org/2024/1/e55121
doi:10.2196/55121
PMID:38820583

©Yuyan Jiang, Xue-li Liu, Zixuan Zhang, Xinru Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mapping Knowledge Landscapes and Emerging Trends in AI for Dementia Biomarkers: Bibliometric and Visualization Analysis

Wenhao Qi[1], BSc; Xiaohong Zhu[1], BSc; Danni He[1,2], MSc; Bin Wang[1], BSc; Shihua Cao[1], PhD; Chaoqun Dong[1], BSc; Yunhua Li[3], MSc; Yanfei Chen[1,4], MSc; Bingsheng Wang[1], BSc; Yankai Shi[1], BSc; Guowei Jiang[5], MSc; Fang Liu[6], PhD; Lizzy M M Boots[5], PhD; Jiaqi Li[1], BSc; Xiajing Lou[1], BSc; Jiani Yao[1], BSc; Xiaodong Lu[7], PhD; Junling Kang[8], PhD

[1]School of Nursing, Hangzhou Normal University, Hangzhou, China

[2]Nursing Department, Zhejiang Provincial People's Hospital, Hangzhou, China

[3]College of Education, Chengdu College of Arts and Sciences, Sichuan, China

[4]Nursing Department, Affiliated Hospital of Hangzhou Normal University, Hangzhou, China

[5]Department of Psychiatry and Neuropsychology and Alzheimer Center Limburg, School for Mental Health and Neuroscience (MHeNS), Maastricht University, Maastricht, Netherlands

[6]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China

[7]Department of Neurology, Affiliated Hospital of Hangzhou Normal University, Hangzhou, China

[8]Department of Neurology, The Third Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China

Corresponding Author:
Shihua Cao, PhD
School of Nursing
Hangzhou Normal University
No. 2318, Yuhangtang Road, Yuhang District
Hangzhou, 310021
China
Phone: 86 13777861361
Email: csh@hznu.edu.cn

Abstract

Background: With the rise of artificial intelligence (AI) in the field of dementia biomarker research, exploring its current developmental trends and research focuses has become increasingly important. This study, using literature data mining, analyzes and assesses the key contributions and development scale of AI in dementia biomarker research.

Objective: The aim of this study was to comprehensively evaluate the current state, hot topics, and future trends of AI in dementia biomarker research globally.

Methods: This study thoroughly analyzed the literature in the application of AI to dementia biomarkers across various dimensions, such as publication volume, authors, institutions, journals, and countries, based on the Web of Science Core Collection. In addition, scales, trends, and potential connections between AI and biomarkers were extracted and deeply analyzed through multiple expert panels.

Results: To date, the field includes 1070 publications across 362 journals, involving 74 countries and 1793 major research institutions, with a total of 6455 researchers. Notably, 69.41% (994/1432) of the researchers ceased their studies before 2019. The most prevalent algorithms used are support vector machines, random forests, and neural networks. Current research frequently focuses on biomarkers such as imaging biomarkers, cerebrospinal fluid biomarkers, genetic biomarkers, and blood biomarkers. Recent advances have highlighted significant discoveries in biomarkers related to imaging, genetics, and blood, with growth in studies on digital and ophthalmic biomarkers.

Conclusions: The field is currently in a phase of stable development, receiving widespread attention from numerous countries, institutions, and researchers worldwide. Despite this, stable clusters of collaborative research have yet to be established, and there is a pressing need to enhance interdisciplinary collaboration. Algorithm development has shown prominence, especially the application of support vector machines and neural networks in imaging studies. Looking forward, newly discovered biomarkers are expected to undergo further validation, and new types, such as digital biomarkers, will garner increased research interest and attention.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e57830) doi:10.2196/57830

KEYWORDS

artificial intelligence; AI; biomarker; dementia; machine learning; bibliometric analysis

Introduction

Background

As the global population ages and life expectancy increases, the number of individuals with dementia is rising at an alarming rate. It is estimated that >55 million people are currently affected by dementia, and this number is expected to continue to grow [1]. The 4 most common subtypes of dementia are Alzheimer disease (AD), vascular dementia (VaD), dementia with Lewy bodies (DLB), and frontotemporal dementia (FTD). Their typical symptoms include cognitive dysfunction, memory loss, and mood fluctuations [2], significantly impacting patients' quality of life and social function. Currently, there is no complete cure for these diseases, posing a substantial burden on patients and their families [3]. Therefore, early diagnosis is crucial for the intervention and management of these diseases [4]. At present, the diagnosis of these conditions largely relies on manual assessments by neurologists or other medical experts, which can be challenging to access in economically disadvantaged areas, leading to many cases of dementia going undiagnosed or misdiagnosed [5]. In addition, neurologists may take a considerable amount of time to make a final diagnosis for a single patient [6].

Biomarkers, as measurable biological indicators that can reflect normal physiological processes, disease progression, or responses to treatment [7], are crucial for the clinical diagnosis, management, and treatment of dementia. The National Institute on Aging and Alzheimer's Association in the United States have recognized the use of biomarkers for diagnosing AD and monitoring its progression [8]. These markers aid clinicians in identifying high-risk groups, making early diagnoses [9], determining subtypes [10], predicting prognosis [11], and assessing drug responses or adverse events. However, with the exponential growth of multiomics and multimodal data, traditional statistical methods are no longer sufficient to meet the needs of discovering new biomarkers [12]. Artificial intelligence (AI), a widely used tool in the health care sector, offers a new perspective for accelerating the discovery of more reliable and clinically applicable biomarkers for dementia [13].

AI, an interdisciplinary field merging computer and data sciences, aims to simulate and extend human intelligence through machines [14]. Core technologies in AI, such as machine learning (ML), natural language processing, and computer vision [15,16], allow researchers to analyze and mine vast amounts of clinical and biomarker data. Through techniques such as ML and deep learning, more accurate and personalized predictions and diagnoses for dementia are made possible [12]; for instance, deep learning and ML as well as using diverse biomarker data types such as imaging, genetic information, and proteomics have been highly accurate in early diagnosis and classification of dementia [17-19]. Genetic and neurobiological data reveal the neuroglial activation and inflammatory states in dementia, identifying pathological stages of the disease [20,21], thereby deepening the understanding of its onset and progression. Similarly, AI identifies patterns and features in these data sets, analyzing potential disease biomarkers. This helps researchers save significant time and resources as well as identify more diagnostic biomarkers for earlier interventions and treatments, ultimately leading to better therapeutic outcomes.

To assess effective diagnostic biomarkers, the Alzheimer's Disease Neuroimaging Initiative (ADNI) has used a multifaceted approach, including imaging and cerebrospinal fluid (CSF) tests, aimed at identifying the most predictive biomarkers for dementia [22]. Yang and Qu [23] analyzed AD biomarker research from 2000 to 2023, using network analysis to highlight CSF and beta amyloid (Aβ) protein as research hot spots and cutting-edge areas. Noda et al [24] identified the research dynamics involving the emerging biomarker neurofilament light (NFL) through keyword trend analysis. Similarly, Wu et al [25] emphasized the significance of AI in dementia research using bibliometrics. In review studies, Aberathne et al [26] highlighted the effectiveness of AI and ML in processing magnetic resonance imaging (MRI) and positron emission tomography (PET) imaging data. Blanco et al [27] and Falahati et al [28] demonstrated the application of algorithms in fluid biomarker research and imaging biomarker performance, respectively, while Chang et al [13] emphasized that ML combined with novel biomarkers and multivariate data could enhance the sensitivity and specificity of AD diagnosis. In addition, Li et al [29] reviewed the use of AI in digital biomarkers. Tzimourta et al [30] reviewed the application of various AI algorithms in 49 experimental studies analyzing electroencephalography (EEG) recordings, summarizing EEG features associated with AD.

However, the existing reviews summarizing the latest findings on AI algorithms and biomarkers often focus solely on 1 type of biomarker, failing to conduct multicategory induction and identify specific patterns. Current bibliometric studies have not yet explored the specific applications of AI in the field of dementia biomarkers. Therefore, this study combines bibliometric and content mining analysis to provide a comprehensive overview of research hot spots and developmental trends, offering valuable insights for future research directions.

Research Problem and Aim

Bibliometrics, as a method for analyzing quantitative information in scholarly literature [31], plays a crucial role in the evaluation of scientific advances within research areas [32]. Through bibliometric analysis as well as content mining and analysis, our study aims to achieve the following objectives:

1. Thoroughly analyze the current status and various stage applications of AI in dementia biomarkers
2. Highlight the research hot spots and future trends in this field

XSL•FO
RenderX

3. Identify and emphasize the contributions of prolific authors, leading countries, and the most productive academic institutions in this field
4. Explore potential future collaborative opportunities
5. Examine the connections and application scale between biomarkers and AI methods

Through this research, we aim to comprehensively understand and evaluate the application of AI in the field of dementia biomarkers and make substantive contributions to the future research development in this area.

Methods

Leveraging the Web of Science Core Collection database and various bibliometric tools, we conducted a detailed collaborative analysis of annual publication volume and trends, author publication dynamics and collaboration networks, institutional publications and collaboration networks, national publications, collaboration networks, distribution of disciplines and interdisciplinary activities, and keyword clustering. By using literature mining and content analysis, we captured the prevalence, trends, connections, newly discovered biomarkers associated with AI algorithms, and various types of dementia biomarkers, distinguishing and analyzing them according to the classification of dementia subtypes.

Data Sources and Search Strategy

Following the suggestion by Donthu et al [33] to minimize potential human errors during format conversion among different databases (manual calibration is required to standardize different database formats, including manually establishing and entering profiles for funds, authors, etc; in addition, discrepancies in citation statistics from different databases and the untraceability of local citations have been noted), we decided to collect bibliometric data from only 1 database. This study selected the Web of Science Core Collection as the platform for the literature search. To ensure comprehensive coverage, all editions of the citation index database were chosen to avoid any omission of relevant literature. This database is widely recognized as a core resource for interdisciplinary academic research and has received high acclaim in numerous bibliometric studies [25,34,35]. Before conducting the search, all team members underwent professional training based on the *Medical Literature Information Retrieval* textbook [36], and a web-based search of the Web of Science Core Collection was conducted on November 2, 2023. The search used keywords such as “artificial intelligence,” “dementia,” and “biomarker,” along with their derivatives, synonyms, and Boolean operators, to construct the search formula (Multimedia Appendix 1). The scope of the search extended from the database’s inception to the date of the search. A total of 2315 relevant documents were retrieved, exported with full records and complete citations, and saved in plain-text format. To avoid bias due to daily updates of the Web of Science Core Collection database, all searching and exporting tasks were completed within the same day.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: (1) document types restricted to “articles” or “reviews,” (2) papers written in “English,” and (3) research topics related to “artificial intelligence” and “dementia biomarkers.” The exclusion criteria were as follows: (1) duplicate publications; (2) nonjournal literature such as conference papers, books, and comments; (3) documents with missing abstract, keywords, or main text; and (4) studies unrelated to “artificial intelligence” and “dementia biomarkers.”

Screening Strategy

After establishing the inclusion and exclusion criteria, to ensure the reliability of the material selection process, 2 evaluators (WQ and XZ) conducted a preliminary screening trial of 50 papers based on the titles, abstracts, and keywords [37]. The Cohen κ coefficient was calculated to be approximately 0.88, indicating a high level of agreement between the evaluators (the Cohen κ coefficient ranges from –1 to 1, with higher values denoting better consistency [38,39]; the specific formulas and methods are provided in Multimedia Appendix 2).

Therefore, we decided not to make any changes to the inclusion and exclusion criteria or to the evaluators. In case of any disagreements during the official selection process, 3 authors (WQ, XZ, and SC) would discuss the matter until a consensus was reached in a team meeting. The literature screening and verification were successfully completed on November 25, 2023. Of the 2315 papers identified, 1070 (46.22%) were included, while 1245 (53.78%) were excluded (type mismatch: n=60, 4.82%; irrelevant to the topic: n=1184, 95.1%; missing abstract: n=1, 0.08%). The detailed search and selection process is recorded in Figure 1.

XSL•FO
RenderX

Figure 1. Search and filter process diagram.

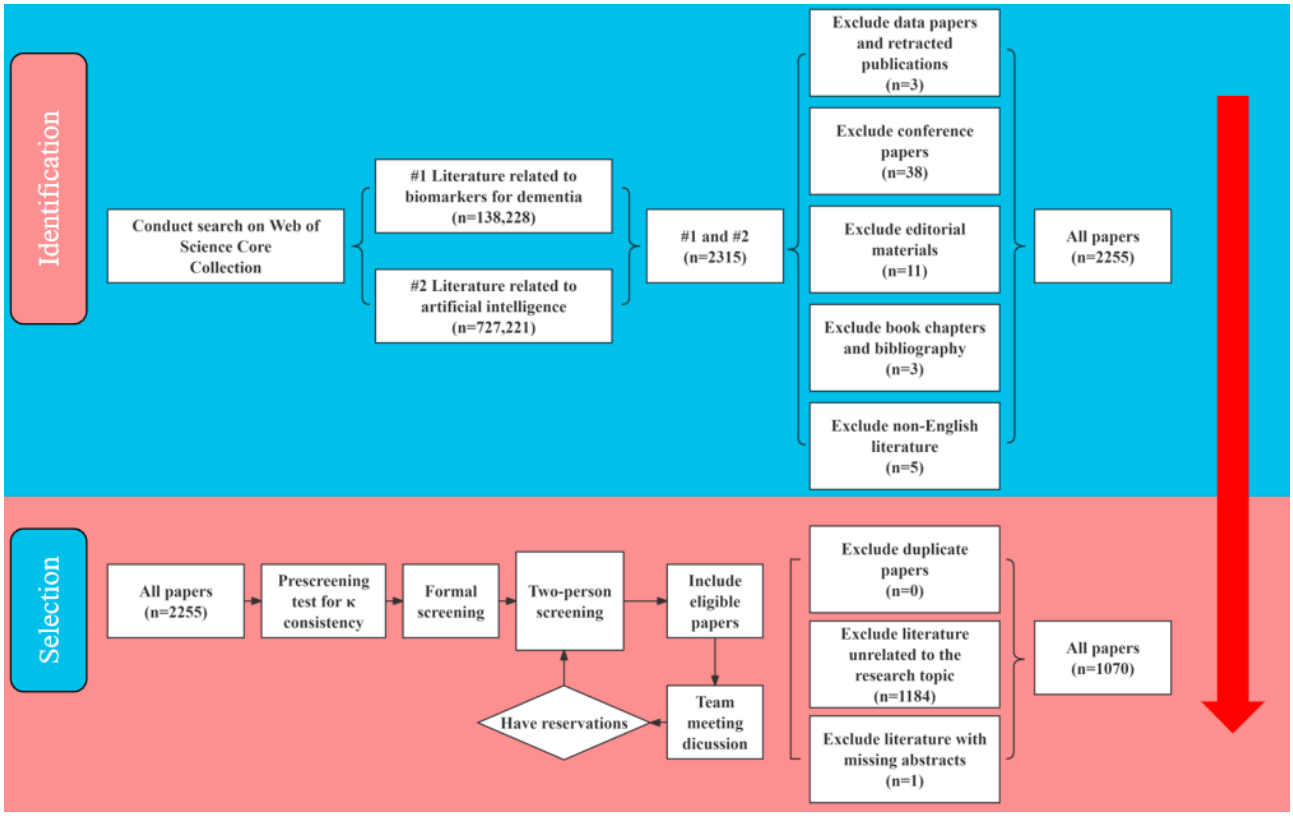

Data Cleaning

In the author analysis process, we conducted further reviews for authors with similar names to determine whether they were the same individual and decide whether further merging was necessary. The review was facilitated by examining the consistency of their Open Researcher and Contributor ID records, publication history, affiliation with the same institution, and information on professional sites such as ResearchGate. During the institutional analysis, we adopted the *institutional affiliation standardizationmodel* developed by Nam et al [40], selecting the first-listed institution, usually the primary affiliation, for authors associated with multiple institutions. In addition, we consolidated various institutions' full names and abbreviations. In analyzing international collaborations, we acknowledged authors affiliated with multiple international institutions because this could indicate potential transnational visiting scholarships or other forms of international cooperation. For funding analysis, we reviewed and appropriately merged various forms of sponsor names, including full names and abbreviations. Before the keyword analysis, to ensure the uniformity and accuracy of author keywords, we used the *Bibliometrix* package in R to merge synonyms; for instance, "Alzheimer disease" and "AD" were unified under "Alzheimer's disease" (specific merged keywords are detailed in Multimedia Appendix 3).

Data Analysis

Currently, single bibliometric tools still have limitations in information extraction and content analysis [41]. To avoid bias and ensure the completeness and detail of information, we adopted a joint analysis strategy based on the strengths of various tools, as detailed in Multimedia Appendix 4. Brief introductions to the tools used are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Brief introductions to the tools used.

Tools and brief introductions

- CiteSpace (version 5.7.R5; Drexel University): a Java scientometric tool developed by Chen [42], used for visualizing trends and patterns in scientific literature as well as revealing hot spots and the evolution of knowledge structures
- VOSviewer (version 1.6.19; Leiden University): free Java document-mapping software developed by the Centre for Science and Technology Studies at Leiden University, Leiden, Netherlands, assists in building various visualization networks [43]
- *Bibliometrix*: an R-based tool for extracting, processing, and analyzing literature data from the Web of Science database [44]
- gCLUTO (version 1.0; Kerapis Lab): focuses on data clustering, offering various clustering methods and visualization options [45]
- Publish or Perish (Harzing.com): used for assessing the publication and citation records of scholars, providing multiple metrics for comprehensive and fair academic research evaluation
- Gephi (version 0.10.1; Gephi.org): software for visualizing social and citation networks, providing significant flexibility in graph rendering
- Joinpoint (version 5.0.2; National Cancer Institute, United States): software designed for identifying and analyzing trend change points in time series data, allowing for the detection of points where there is a significant shift in the slope of the trend [46]
- Scimago Graphica (version 1.0.16; Scimago Lab) [35] and Pajek (64-bit version) Portable (version 5.18; University of Ljubljana) [47]: for enhanced readability of knowledge maps, Scimago Graphica and Pajek (64-bit version) Portable were incorporated for layout purposes

The analysis for each section adopted the bibliometrics analysis scheme proposed by Cobo et al [48].

Statistical Analysis

Extraction and Classification of Biomarkers and AI Algorithms

We specifically established an interdisciplinary professional team responsible for reading the full texts of research papers to extract and classify specific biomarkers and AI algorithms and to handle discussions and disputes that arose. The team consisted of 2 neurology experts, 2 AI domain experts, and 1 medical informatics expert. The classification process for biomarkers and algorithms was conducted independently by the neurology experts and the AI domain experts, without interference from each other. In addition, each expert conducted evaluations independently, and in cases of dispute, the medical informatics expert intervened to discuss the issue and take a decision. We referred to the classification of ML algorithms by Gutierrez [15] and Silva-Spínola et al [49], classified the biomarkers based on their nature and acquisition methods, and ultimately used Gephi (version 0.10.1) to construct a co-occurrence network between them. The specific classification process and network construction are shown in Figure 2. The detailed classification methods of biomarkers are presented in Multimedia Appendix 5.

XSL•FO
RenderX

Figure 2. Classification and network construction diagram of dementia biomarkers and artificial intelligence algorithms. Different colored squares represent various types of algorithms or biomarkers. The upright arrow on the far left represents the workflow. After organizing and compiling these, connections were established by constructing a matrix.

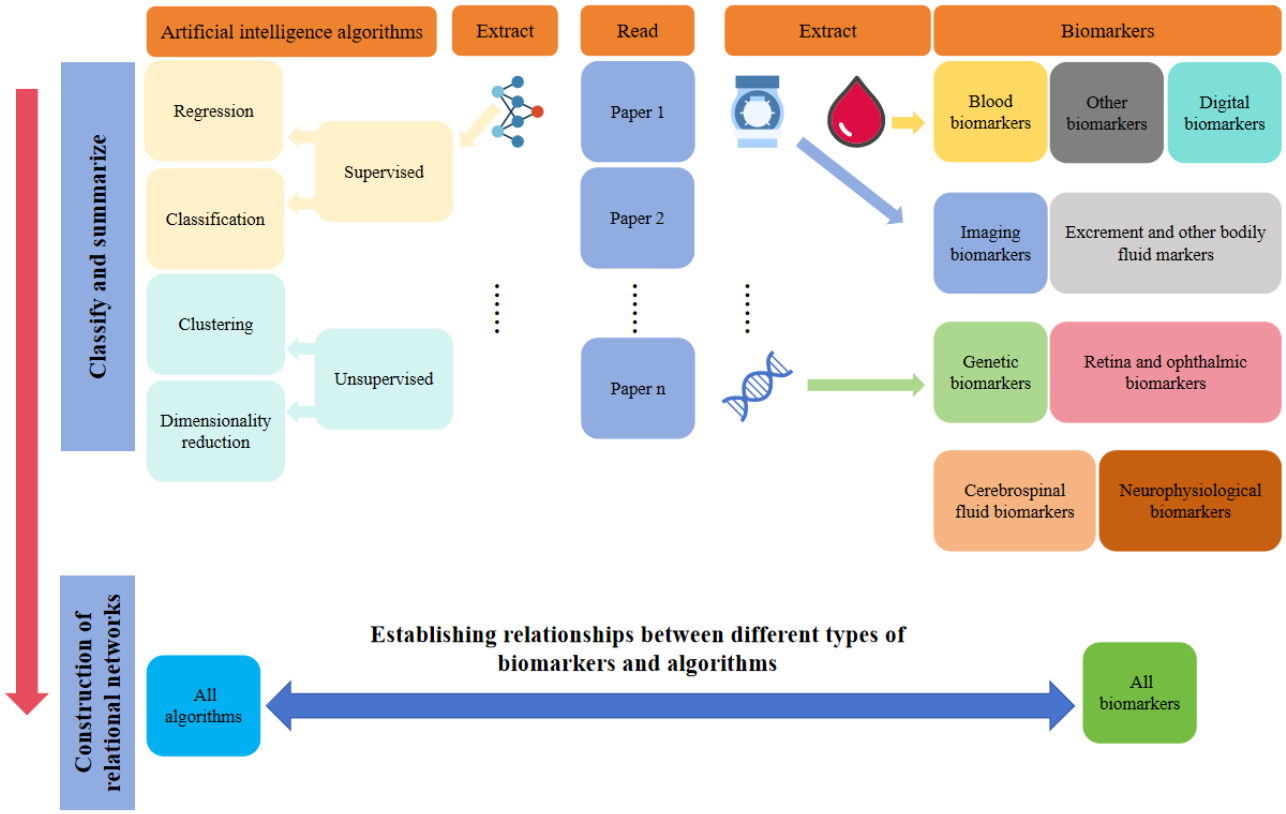

Publication Output and Growth of Research Interest

We used CiteSpace to analyze the annual publication trends of the literature and applied polynomial fitting using the least squares method in OriginPro 2021 (OriginLab Corporation) [50]. The R^2 value is an indicator of the fit of a trend line, reflecting the degree of fit between the estimated values of the trend line and the corresponding actual data. The closer the R^2 value is to 1, the higher the degree of fit and the greater the reliability of the trend line [51]. The annual growth rate of publications was calculated using the following compound formula [52,53]:

> *Growth rate = ([number of publications in the last year / number of publications in the first year]1 / (last year − first year) − 1) × 100*

Joinpoint software was used to evaluate time trends in a structured manner and to test which trends between junction points were statistically significant [54]. The software applies recommended schemes for the number of turning points in the model. To indicate the direction and magnitude of trends, this study calculated the changes in the trend slope. The slope represents the rate of change of the dependent variable over a specific period. When the difference in slopes between 2 line segments is significantly different from 0, it indicates a significant change in the trend at the corresponding time point (ie, the node). $P<.05$ was considered statistically significant.

Author Analysis

We used VOSviewer and *Bibliometrix* to analyze key information of the top 10 authors with the highest publication volume. Considering the differences in interdisciplinary citation habits, we used Publish or Perish software to calculate the h-index [55], g-index [56], and e-index [56,57] scores, thus avoiding assessment biases that might arise from relying on a single metric [58]. A higher e-index score indicates that an author has produced a series of high-quality, high-impact research works in their field, rather than just a few widely cited papers. Detailed methods and formulas for calculating the e-index score are provided in Multimedia Appendix 2.

We used Microsoft Excel 2019 to compile the annual output of all authors, analyzing their publication dynamics to identify new researchers and *terminators* [59]. New researchers are defined as those who started publishing in a specific year without any prior related publications, while *terminators* are those who published articles before a specific year but did not publish any article after that year [59]. The Price law formula [60] was applied to identify the core group of authors and calculate their productivity. The specific formulas and methods are provided in Multimedia Appendix 2.

Journal Analysis

To identify core journals in the field, we applied the Bradford law [61,62]. We conducted a fair and comprehensive evaluation of the journals' academic impact, integrating metrics such as CiteScore 2022 [63,64], Scimago Journal Rank [63,65], Journal Citation Reports Quartile rankings [66,67], and Impact Factor

XSL•FO
RenderX

[67,68]. These measures help in assessing the journals' influence and relevance in the field accurately [58,63].

Country Analysis

A detailed analysis of the countries leading in global publication volume was performed using VOSviewer. The Scimago Graphica tool was used to create a world map illustrating publication volumes and regional densities. The gross domestic product of these countries was estimated and analyzed, taking into account data from the International Monetary Fund's *World Economic Outlook* report [69]. In addition, the prevalence and mortality rates of dementia in these countries were examined by consulting reports from the World Health Organization's Global Dementia Observatory [70] and age-standardized dementia mortality rates [71].

Analysis of Highly Cited Papers

On the basis of the local citation index, the top 10 highly cited papers were identified, and their standardized citation indices were calculated. The normalized citation score is derived by dividing the number of citations of a key paper by the average number of citations for comparable papers in the same field or subfield and publication year. A final impact score (normalized citation score) of >1 indicates that the paper's citation rate is above the average for that field or subfield, while a score of <1 indicates that it is below average [72].

Author Keywords

High-frequency keywords were then clustered using gCLUTO based on their proximity, using hierarchical clustering with repeated bisection, and using the cosine function to calculate similarity. The clustering criterion function was set to I^2, and the results were selected for display based on high internal similarity and low external similarity, with the results displayed using matrix and mound visualization techniques [73]. The selection of high-frequency keywords for clustering is based on the method described by Bai et al [74], which involves extracting keywords that cumulatively account for >30% of the total frequency. If the number of included keywords is <30, the threshold is adjusted to include high-frequency words that cumulatively account for >40% until the number exceeds 30. Building on this approach, we observed the importance of subsequent keywords and incorporated them appropriately.

Disciplinary Analysis

Through disciplinary analysis, we can gain a comprehensive understanding of the research content within a field and interdisciplinary collaborations. The fields of study form the subject classification scheme shared across all Web of Science product databases. Each document indexed in the Web of Science Core Collection is assigned to at least 1 subject category, which maps to a research field. Using VOSviewer, we constructed a disciplinary collaboration network to understand the distribution of disciplines within the field and the nature of interdisciplinary collaborations, where each node represents a discipline, and the connections between nodes represent collaborations among disciplines [75].

Ethical Considerations

Ethics committee approval was not required because this study was a retrospective bibliometric analysis of existing published studies.

Results

The Annual Trends of Publications

Our study incorporated 1070 research papers, of which 993 (92.8%) were articles and 77 (7.2%) were reviews, indicating that the research in the field of dementia biomarkers using AI is primarily driven by original articles.

The change in publication volume reflects the dynamic development of this field. The earliest study on this topic dates back to 2007. In 2020, of the 1070 included papers, 131 (12.24%) were published (the 100-paper mark was crossed for the first time), and publication peaked at 229 (21.4%) papers in 2022. To visually represent the change in publication volume, we used a cubic trendline model. As shown in Figure 3A, the red dashed line represents the fitted trendline, with an R^2 value of 0.95760 and an adjusted R^2 value of 0.94783, indicating a good model fit and accurately reflecting the growth trend in publication volume. On the basis of the trend analysis using Joinpoint software, 2 potential turning points were identified in the years 2018 and 2021. The slopes calculated for these periods are as follows: slope 1 (from 2007 to 2018)=4.02, slope 2 (from 2018 to 2021)=48.58, and slope 3 (from 2021 to 2023)=16.33. The differences in slopes between slope 1 and slope 2 as well as those between slope 2 and slope 3 have P values <.05, indicating significant changes in the growth trends, as illustrated in Figure 3B.

XSL•FO
RenderX

Figure 3. (A) Annual and total publication outputs and the model-fitting curve of the time trend of artificial intelligence in dementia biomarkers. (B) The distribution across 3 phases and the respective slopes. *Indicates that the Slope is significantly different from zero at the alpha = 0.05 level. Final selected model: 2 Joinpoints".

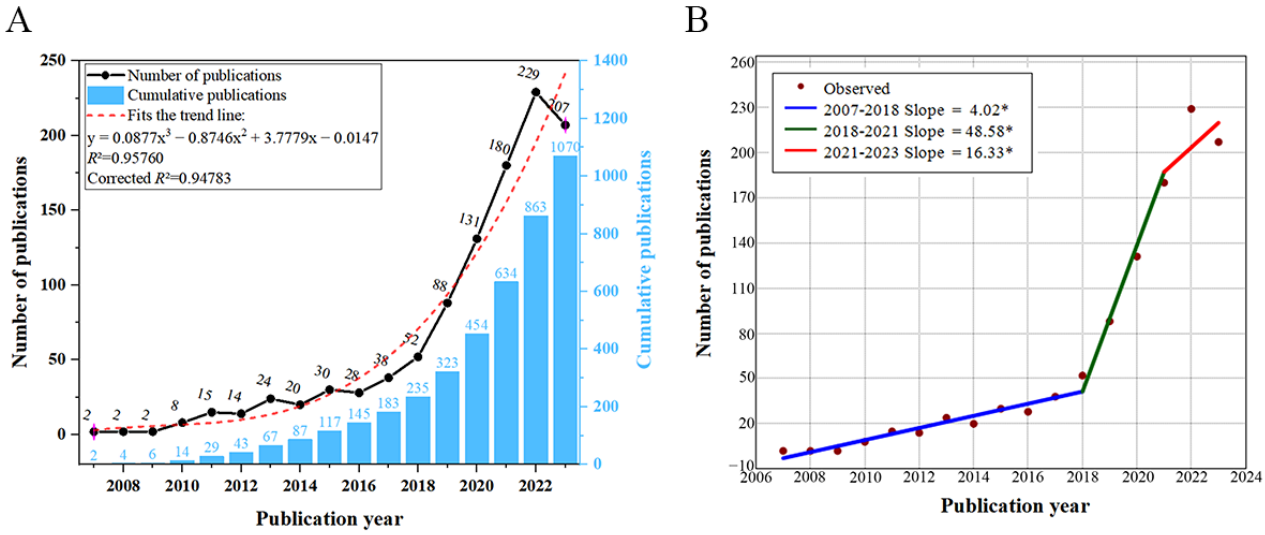

On the basis of the changes in publication volume and slope, the development of this field can be preliminarily divided into 3 stages. The first stage (2007-2017) is the initiation stage, with 183 (17.1%) of the 1070 papers published during this period, and an annual publication volume not exceeding 50 papers (growth rate of 34.2%). The second stage (2018-2020) is marked by rapid growth, with 271 (25.33%) of the 1070 papers published during this period, and an annual publication volume not exceeding 100 papers (growth rate of 58.7%). The third stage (2021-2023) is characterized as a stable development phase, influenced by a larger publication base, with 616 (57.57%) of the 1070 papers published during this period (growth rate of 7.2%).

Author Analysis

The participation of researchers in the field reflects the level of interest in it. A total of 6455 authors have been involved in publishing papers. The top 10 authors have collectively contributed 125 (1.35%) of the 9246 studies. Among them, Morris, JC, is the most prolific author (16/9246, 0.17%). Shen, DG, has the highest h-index and e-index scores among these prolific authors. The majority of these prolific authors (8/10, 80%) published their works between 2018 and 2023, while the publications of Shen, DG, and Zhang, DQ, are mainly concentrated between 2007 and 2017, as shown in Table 1.

Adhering to the Price law, the minimum publication threshold for core authors is approximately 3 papers. Using VOSviewer for analysis, 663 (10.27%) of the 6455 core authors were identified, contributing a total of 2635 (28.5%) of the 9246 papers, which does not meet the standard of the Price law (>50%) [60]. In the collaboration network diagram, the co-occurrence network among core authors is relatively independent with fewer connections, indicating a pattern of high cohesion and low coupling. Networks centered around the top 10 most prolific authors are more developed compared to those of others, as illustrated in Figure 4.

Figure 5 illustrates the annual influx of researchers into the field of AI in dementia biomarkers. Of the 6455 authors involved in publishing papers in the field, there were only 14 (0.22%) in 2007, while in 2023, the number of new researchers entering the field soared to 1208 (18.71%). The trend line indicates that there will be an increasing number of new researchers joining this field in the future. On the basis of the influx of new authors, the year 2019 was selected as a specific point in time [59] to identify new researchers and those who ceased their research in this area at the current stage. Among them, 5023 (77.81%) of the 6455 researchers are new to this field since 2019, and of the 1432 researchers who were active before 2019, a total of 994 (69.41%) ceased publishing after 2019. In addition, in exploring the demographics of new researchers, it was found that 372 (56.1%) of the 663 core authors identified by the Price law are newcomers to the field.

XSL•FO
RenderX

Table 1. Top 10 authors' production distribution and academic impact evaluation.

Rank	Author	Output (n=9246), n (%)	h-index[a] score	e-index[b] score	g-index[c] score	Period 1: 2007-2017, n (%)[d]	Period 2: 2018-2020, n (%)[d]	Period 3: 2021-2023, n (%)[d]
1	Morris, JC	16 (0.2)	10	19.4	16	2 (12.5)	3 (18.7)	11 (68.7)
2	Jack, CR	13 (0.1)	9	20.1	13	3 (23.1)	3 (23.1)	7 (53.8)
3	Liu, Y	13 (0.1)	9	11.6	13	0 (0)	5 (38.5)	8 (61.5)
4	Saykin, AJ	13 (0.1)	9	30.4	13	4 (30.8)	7 (53.8)	2 (15.4)
5	Shen, DG	13 (0.1)	13	42.5	13	7 (53.8)	6 (46.1)	0 (0)
6	O'Bryant, SE	12 (0.1)	9	22.2	12	4 (33.3)	3 (25)	5 (41.7)
7	Zetterberg, H	12 (0.1)	9	17.6	12	0 (0)	6 (50)	6 (50)
8	Han, Y	11 (0.1)	8	14.3	11	1 (9.09)	7 (63.6)	3 (27.3)
9	Wang, L	11 (0.1)	6	13.4	11	2 (18.2)	3 (27.3)	6 (54.5)
10	Zhang, DQ	11 (0.1)	9	40.3	11	7 (63.6)	2 (18.2)	2 (18.2)

[a]At least *h* papers have been cited *h* times each.

[b]The supplementary measure of the h-index score.

[c]The total citation count of the first *g* papers is $\geq g^2$.

[d]The denominator is the n value in "Output" column.

Figure 4. Graph of core authors' collaboration network. Color coding is used to display clusters, with authors within the same cluster sharing the same color. The size of the circles increases with the number of publications.

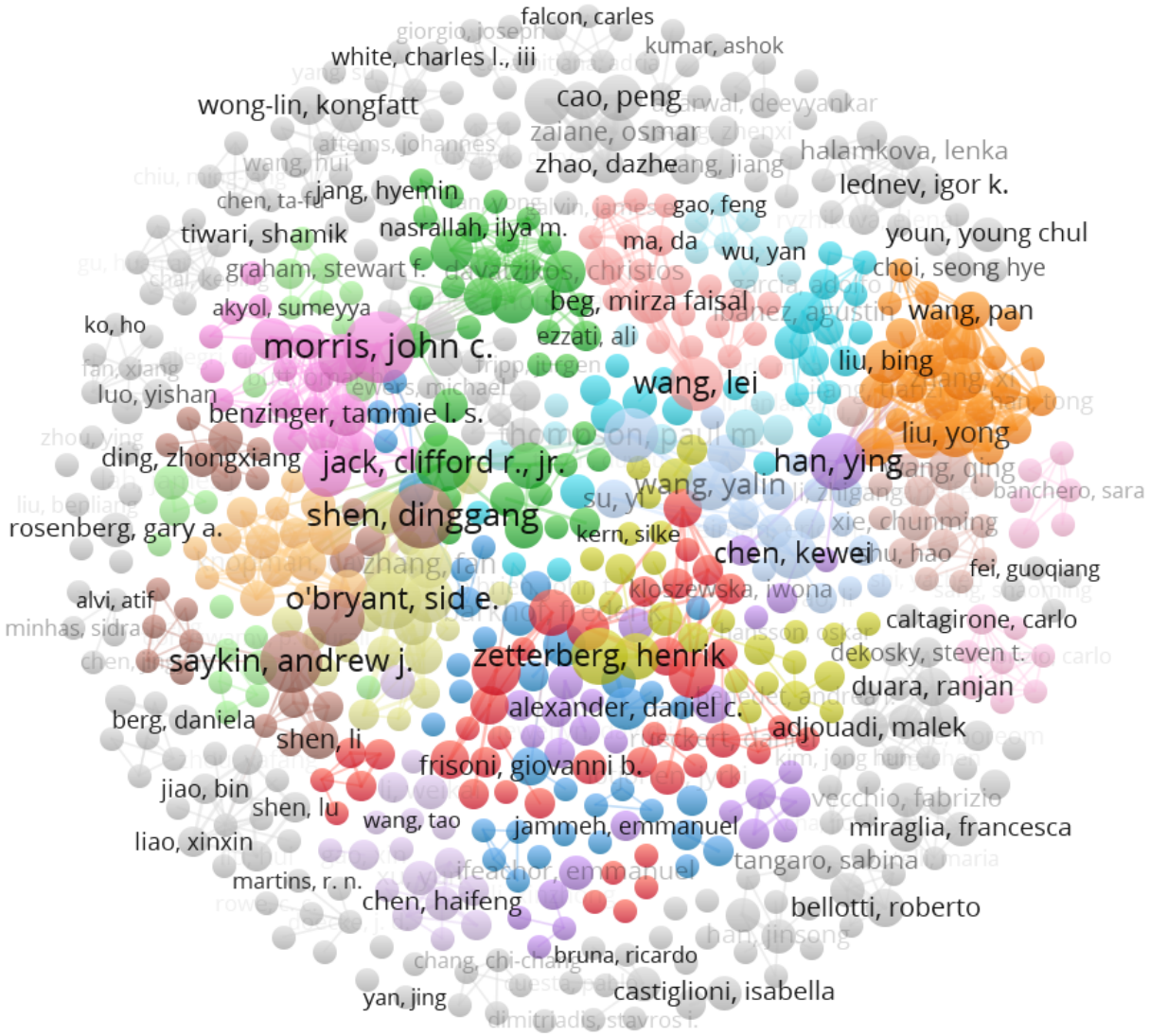

XSL•FO
RenderX

Figure 5. Time distribution chart of new research personnel entering the field of artificial intelligence in dementia biomarkers.

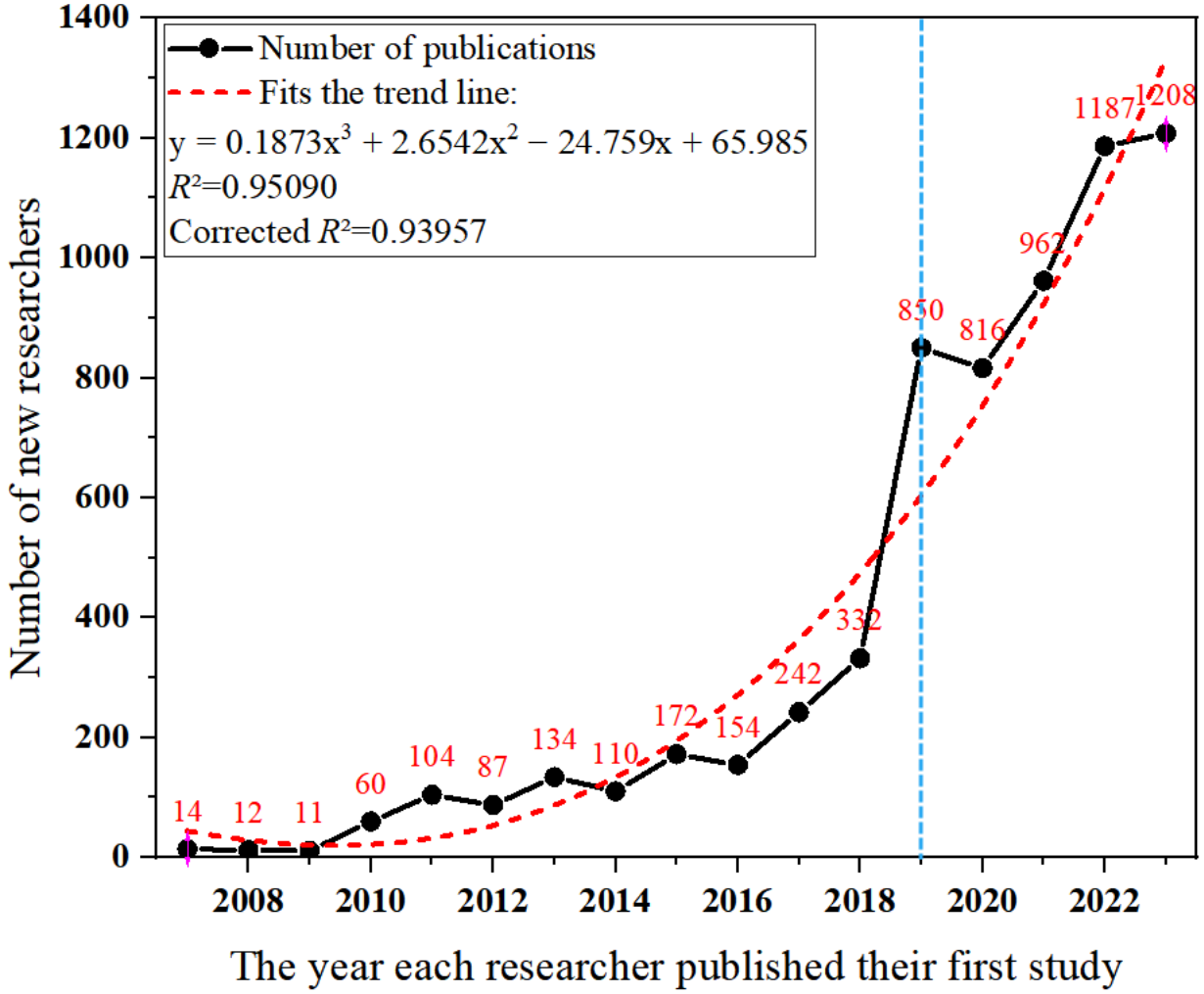

Journal Analysis

The journal analysis showcases the structure and characteristics of the field. A total of 362 journals have published relevant articles. Following the Bradford law [61,62], we identified 12 core journals in this field that collectively contributed 360 (33.6%) of the 1070 studies. Of these, the *Journal of Alzheimer's Disease* (Netherlands) had the highest output with 22.7% (78/344) of the published papers. In terms of citation frequency, *NeuroImage* (United States) leads, with a citation percentage of 13.8% (3293/23,842), averaging 122 citations per paper. The journal with the highest impact factor is *Alzheimer's and Dementia* (United States). These journals are all ranked in the top 2 quartiles of the Journal Citation Reports Quartile rankings and have achieved notable CiteScore 2022 and Scimago Journal Rank rankings, as shown in Table 2.

The dual map overlay of the journals reveals the thematic distribution across academic journals (Figure 6). Figure 6A shows the citing journals, while Figure 6B shows the cited journals; the colored paths indicate citation relationships. There are 5 cited paths: 2 yellow, 2 pink, and 1 green. The analysis indicates that papers in psychology, education, or sociology journals are often cited by journals from fields such as molecular biology, immunology, medicine, clinical studies, ophthalmology, kinesiology, and neurology. Similarly, papers from molecular biology, genetics, or genomics journals are often cited by journals from fields such as medicine, clinical studies, and neurology, highlighting the importance of interdisciplinary research.

XSL•FO
RenderX

Table 2. Top 12 journals with the highest publication volumes on the application of artificial intelligence in dementia biomarkers.

Rank	Journal	Output (n=1070), n (%)	Citations (n=23,842), n (%)	CiteScore 2022	Impact Factor 2022[a]	JCR[b]	SJR[c]	Country
1	*Journal of Alzheimer's Disease*	78 (7.3)	1445 (6.1)	6.4	4.0	Q[d]2	1.146	Netherlands
2	*Frontiers in Aging Neuroscience*	58 (5.4)	887 (3.7)	5.2	4.8	Q2	1.211	Switzerland
3	*Scientific Reports*	34 (3.2)	741 (3.1)	7.5	4.6	Q2	0.973	United Kingdom
4	*NeuroImage*	27 (2.5)	3293 (13.8)	11.6	5.7	Q1	2.512	United States
5	*Alzheimer's & Dementia*	26 (2.4)	826 (3.5)	14.7	14.0	Q1	3.288	United States
6	*PLOS ONE*	26 (2.4)	1140 (4.8)	6.0	3.7	Q2	0.885	United States
7	*Alzheimer's Research & Therapy*	21 (2.0)	374 (1.6)	12.0	9.0	Q1	2.650	United Kingdom
8	*Frontiers in Neuroscience*	20 (1.9)	369 (1.5)	6.8	5.2	Q2	1.161	Switzerland
9	*NeuroImage: Clinical*	19 (1.8)	652 (2.7)	8.1	4.2	Q2	1.395	Netherlands
10	*Human Brain Mapping*	18 (1.7)	670 (2.8)	9.1	4.8	Q1	1.688	United States
11	*IEEE Access*	17 (1.6)	166 (0.7)	9.0	3.9	Q2	0.926	United States
12	*Frontiers in Neurology*	16 (1.5)	192 (0.8)	4.8	3.4	Q2	0.978	Switzerland

[a]Impact factor based on Clarivate Analytics Journal Citation Report (2022).

[b]JCR: Journal Citation Reports.

[c]SJR: Scimago Journal Rank.

[d]Q: quartile ranking position.

Figure 6. The dual-map overlay of journals that have published papers on artificial intelligence in dementia biomarkers. The lines indicate the pathways between the Web of Science Core Collection categories of (A) the citing journals and (B) the cited journals. Thicker lines signify a stronger citation relationship. The colors represent the origins of the Web of Science categories of the citing journals.

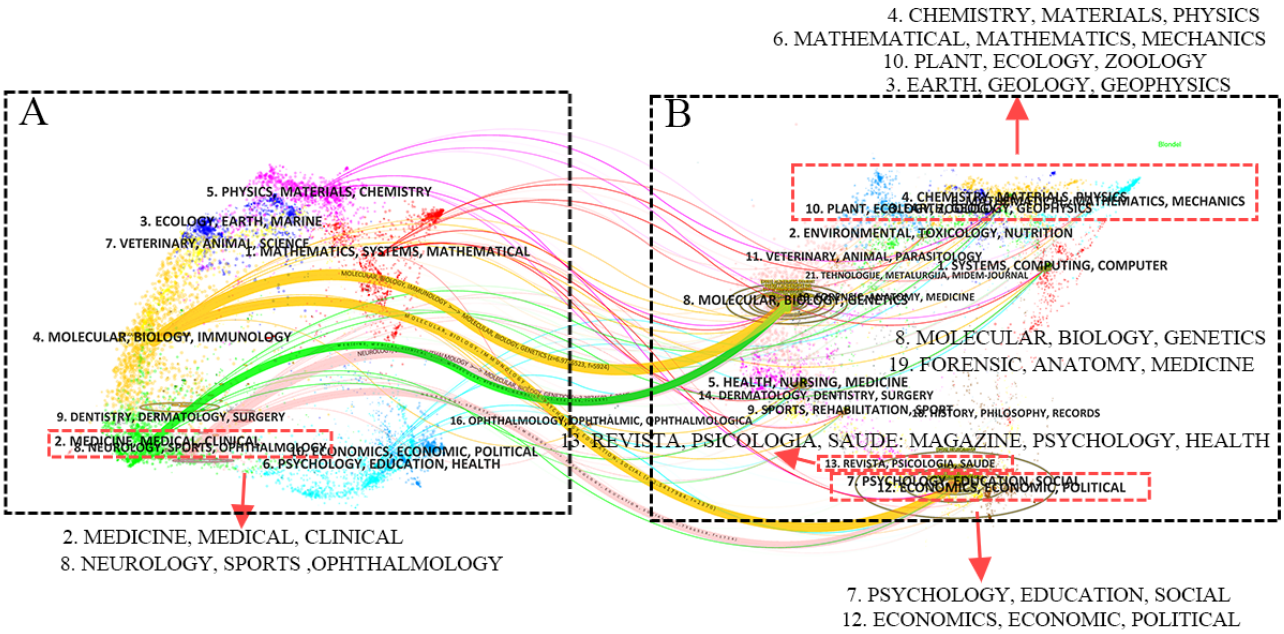

Institutional Analysis

The institutional analysis reveals the organizational structure characteristics of research in the field of dementia biomarkers. A total of 1793 institutions have conducted research on AI in dementia biomarkers and published papers. The highest publishing volumes come from the University of Pennsylvania in Philadelphia, United States, which contributed 31 (0.9%) of the 3442 papers. The University of North Carolina Chapel Hill in North Carolina, United States, has the highest citation index, with 2235 (2.73%) of the 81,952 citations, averaging 111.8 citations per paper. Among the top 10 institutions in terms of publication volume, 5 (50%) are located in the United States, 3 (30%) in the United Kingdom, 2 (20%) in China, and 1 (10%) is the globally renowned Mayo Clinic in the United States, as shown in Table 3.

To further explore the collaboration patterns among these institutions, we selected the top 100 institutions by publication

XSL•FO
RenderX

volume (the list includes 102 institutions due to institutional ties, collectively publishing 992/3442, 28.82% of the papers, with a minimum publication count of 6) to construct a collaboration network map. The map reveals that most of these institutions (94/102, 92.2%) are research-intensive universities. Notably, institutions from China, the United States, and the United Kingdom form 3 major collaborative networks, with specific network relationships detailed in Figure 7.

Table 3. Top 10 organizations in the field of artificial intelligence in dementia biomarkers.

Rank	Organization	Output (n=3442), n (%)	Citations (n=81,952), n (%)	PPC[a]	Country
1	University of Pennsylvania	31 (0.9)	790 (0.96)	25.5	United States
2	University College London	26 (0.76)	682 (0.83)	26.2	United Kingdom
3	King's College London	25 (0.73)	1931 (2.36)	77.2	United Kingdom
4	Mayo Clinic	22 (0.64)	640 (0.78)	29.1	United States
5	Capital Medical University	21 (0.61)	338 (0.41)	16.1	China
6	University of Cambridge	21 (0.61)	582 (0.71)	27.7	United Kingdom
7	University of California San Francisco	20 (0.58)	814 (0.99)	40.7	United States
8	University of North Carolina Chapel Hill	20 (0.58)	2235 (2.72)	111.8	United States
9	Chinese Academy of Sciences	19 (0.55)	467 (0.57)	24.6	China
10	Washington University	17 (0.49)	623 (0.76)	36.6	United States

[a]PPC: per-paper citations.

Figure 7. Graph of top 100 organizations in the artificial intelligence in dementia biomarker collaboration network. Colors represent clusters, with institutions within the same cluster sharing the same color. The size of the circles increases with the number of publications.

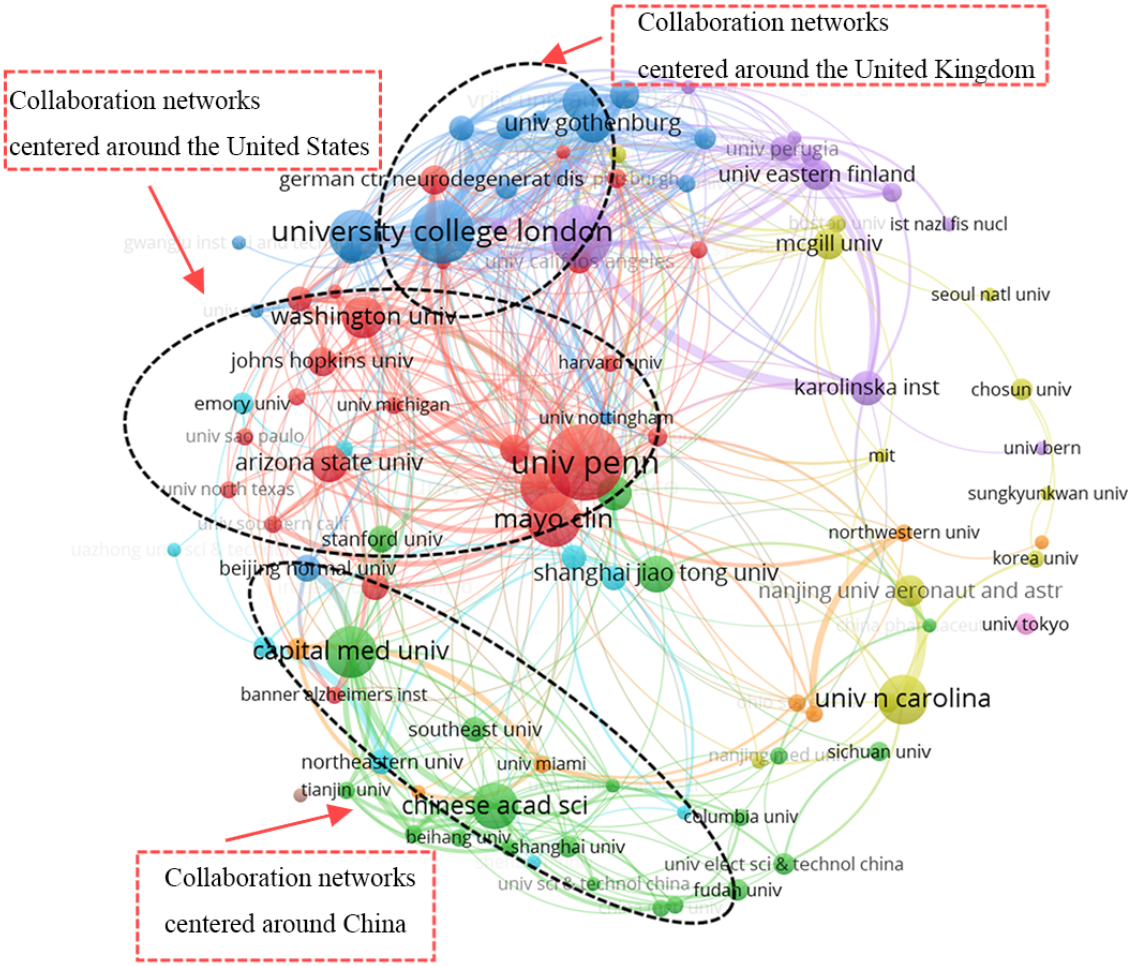

XSL•FO
RenderX

Country Analysis

The participation of 74 countries in dementia biomarker research highlights the global interest in the topic. The 10 most productive countries contributed 1216 (69.64%) of the 1746 papers and 1110 (61.91%) of the 1793 research institutions. The United States led in publication and citation counts with 346 (19.82%) of the 1746 papers and 10,745 (25.28%) of the 42,496 citations. China had the most research institutions (281/1793, 15.67%). South Korea had a dementia prevalence rate of 7%, China 4.5%, and India 3.7%. The standardized dementia mortality rates in the United States and the United Kingdom were higher than in other countries, as detailed in Table 4.

A visualization map was created using Scimago Graphica software to display the level of attention different regions pay to the field. In the map, the size of the circles and the color of the circles represent the publication volume of each country. The European region shows a higher interest in this field than other continents, with 30 countries participating in publishing research, as seen in Figure 8.

A chord diagram of international collaboration based on the number of joint papers was produced. The lines represent collaborative relationships between countries, with the width indicating the strength of collaboration. Each country's end point on its own axis represents its total number of collaborations with other countries. Among the top 10 productive countries, the United States is at the core of a network covering 69% (51/74) of the countries, with 377 collaborations; the United Kingdom covers 66% (49/74) of the countries, with 366 collaborations; and China covers 42% (31/74) of the countries, with 191 collaborations, as illustrated in Figure 9.

Table 4. Countries with the top 10 publications on artificial intelligence in dementia biomarkers.

Rank	Country	Output (n=1746), n (%)	Citations (n=42,496), n (%)	Organizations (n=1793), n (%)	2023 GDP[a] rank	Partner countries (n=74), n (%)	Prevalence rate[b] (%)	Mortality rate[c], n ( )
1	United States	346 (19.82)	10,745 (25.28)	241 (13.44)	1	51 (68.9)	6.4	3.33
2	China	282 (16.15)	3782 (8.9)	281 (15.67)	2	31 (41.9)	4.5	1.74
3	United Kingdom	143 (8.19)	5079 (11.95)	82 (4.57)	6	49 (66.2)	__[d]	4.27
4	Italy	79 (4.52)	2687 (6.32)	103 (5.74)	8	36 (48.6)	6.9	1.49
5	South Korea	70 (4.01)	1355 (3.19)	52 (2.9)	13	31 (41.9)	7	1.63
6	India	70 (4.01)	920 (2.16)	95 (5.3)	5	15 (20.3)	3.7	1.46
7	Germany	65 (3.72)	2376 (5.59)	79 (4.41)	3	34 (45.9)	6.9	1.55
8	Spain	58 (3.32)	912 (2.15)	87 (4.85)	15	32 (43.2)	—	2.15
9	Canada	57 (3.26)	1170 (2.75)	43 (2.4)	10	28 (37.8)	6.4	2.79
10	Australia	46 (2.63)	1786 (4.2)	47 (2.62)	14	40 (54.1)	6.7	2.26

[a]GDP: gross domestic product.

[b]The World Health Organization's Global Dementia Observatory's estimate of the unstandardized prevalence rate of dementia in the Global Burden of Disease region report for the year 2017.

[c]The World Health Organization's age-standardized dementia mortality rates per 100,000 population in 2019 by country.

[d]Not available.

XSL•FO
RenderX

Figure 8. World map of production distribution by country.

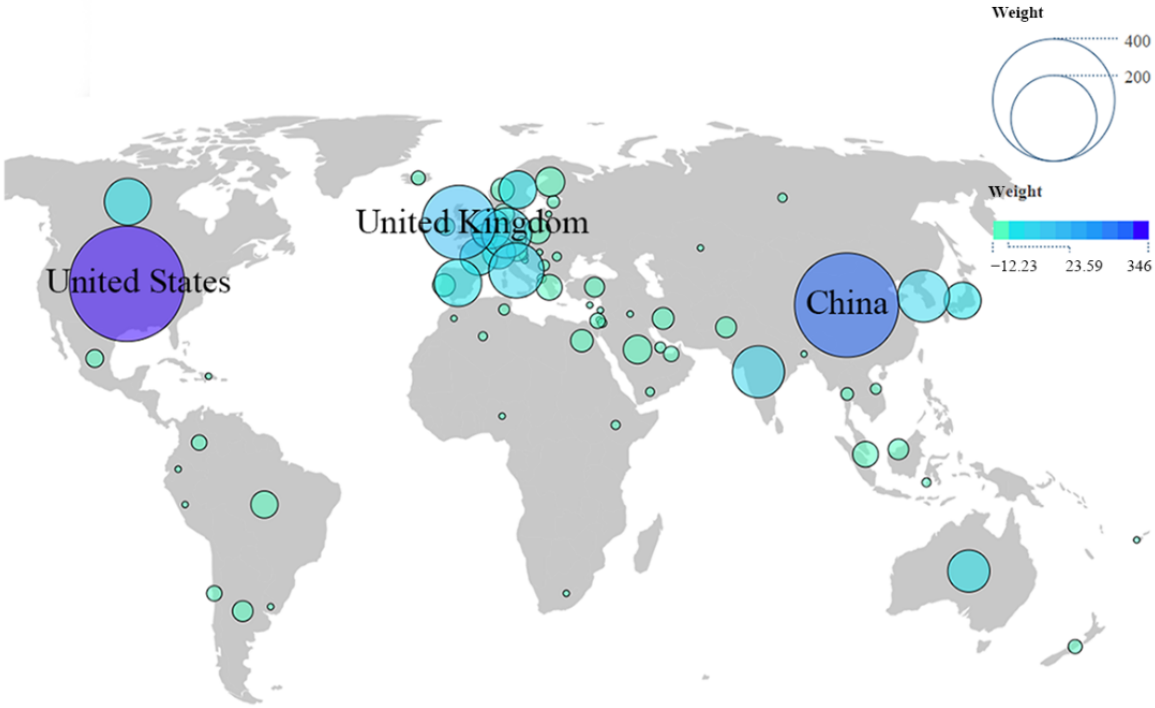

Figure 9. The world cooperation research network map. The colors representing the countries have no specific meaning; only the thickness of the lines between them is significant, indicating the frequency of collaborations between different countries. The thickness of the lines corresponds to the numerical values on their respective axes. The radial axes end points for each country represent the total number of collaborations with other countries.

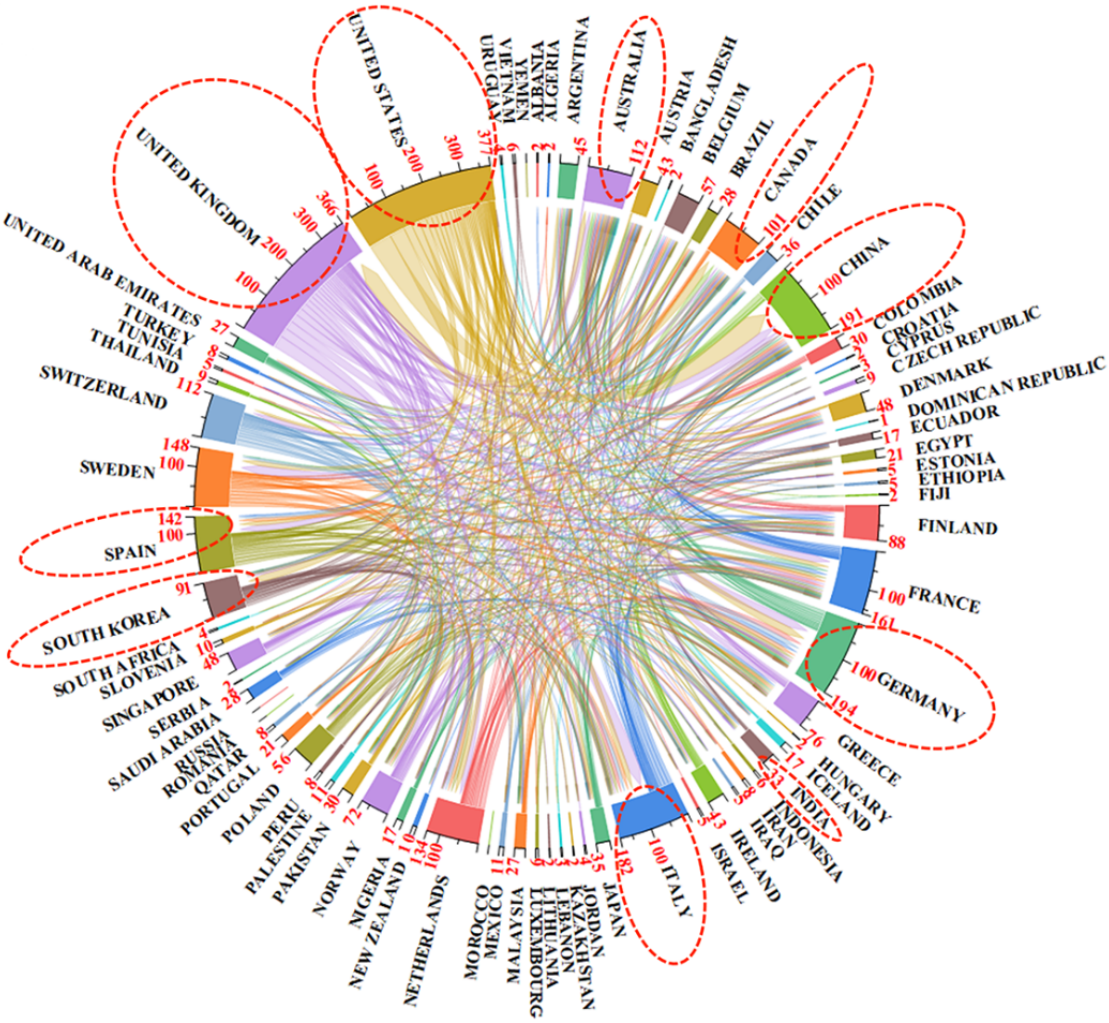

XSL•FO
RenderX

Fund Analysis

The funding situation for projects in this field is a key indicator of the level of investment and government emphasis in each country. The study identified 450 funding projects providing 1604 instances of support for such research. Upon reviewing the top 10 funding projects with the most contributions, it was found that 5 (50%) are from the United States, 2 (20%) from China, 1 (10%) each from South Korea and the United Kingdom, and 2 (20%) from international organizations. Notably, of the 1604 studies in this area, the National Institutes of Health in the United States provided funding for 161 (10.04%), and the ADNI funded 135 (8.42%). Detailed information can be found in Table 5.

Table 5. Top 10 funders for studies on artificial intelligence in dementia biomarkers (N=1604).

Rank	Funders	Studies, n (%)	Country
1	National Institutes of Health	161 (10.03)	United States
2	Alzheimer's Disease Neuroimaging Initiative	135 (8.42)	United States
3	National Natural Science Foundation of China	119 (7.41)	China
4	Department of Defense–Alzheimer's Disease Neuroimaging Initiative	118 (7.36)	United States
5	National Institute on Aging	61 (3.8)	United States
6	National Research Foundation of Korea	28 (1.75)	South Korea
7	Medical Research Council	26 (1.62)	United Kingdom
8	National Key Research and Development Program of China	23 (1.43)	China
9	Alzheimer's Association	21 (1.31)	United States
10	European Union	16 (1)	__[a]

[a]Not applicable.

Analysis of Highly Cited Papers

Compared to global citations, local citations, or peer citations, more accurately reflect the academic community's recognition and importance of specific articles locally, as well as the influence, quality, and collaboration status of the literature in local academic research [75,76]. The top 10 high-value publications, based on local citations, accumulated a total of 392 (18.2%) of the 2157 local peer citations, averaging 39.2 citations per year for each publication. The local and global normalized citation indices for these studies are both >1, indicating that their citation rates exceed the average level for research published in the same year. Of the 10 highly cited papers, 8 (80%) were published between 2007 and 2017 (for detailed information, refer to Table 6).

A deeper analysis of these 10 highly cited papers revealed valuable information regarding their specific tasks and research outcomes. Of the 10 papers, 1 (10%) is a review paper [11], and 9 (90%) are research papers [77-85]. These studies predominantly conducted binary classification analyses using the ADNI data set, with 9 (90%) of the 10 papers using multimodal biomarkers. Of the 10 papers, 8 (80%) applied ML methods, and 2 (20%) used deep learning techniques. These studies detailed their methods for classifying specific diseases; the types of biomarkers used; and the accuracy, sensitivity, specificity, and fitting of their classification tasks. However, not all studies reported these specific values in detail. More details about these studies can be found in Table 7 and Multimedia Appendix 6. The top 10 locally normalized cited documents can be found in Multimedia Appendix 7.

XSL•FO
RenderX

Table 6. The top 10 locally cited articles on the application of artificial intelligence in dementia biomarkers.

Rank	Article title	Study	LCS[a] (n=2157), n (%)	GCS[b] (n=23,842), n (%)	NLCS[c]	NGCS[d]	PY[e]
1	Multimodal classification of Alzheimer's disease and mild cognitive impairment	Zhang et al [77]	109 (5.05)	883 (3.7)	7.6	6.0	2011
2	Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects	Moradi et al [78]	52 (2.41)	421 (1.77)	7.7	5.2	2015
3	Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease	Zhang and Shen [79]	40 (1.85)	453 (1.9)	6.2	4.0	2012
4	Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data	Jo et al [11]	35 (1.62)	229 (0.96)	9.3	6.9	2019
5	Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment	Young et al [80]	31 (1.44)	183 (0.77)	6.4	3.2	2013
6	Predicting Alzheimer's disease progression using multimodal deep learning approach	Lee et al [81]	29 (1.34)	158 (0.66)	7.7	4.8	2019
7	Random forest–based similarity measures for multi-modal classification of Alzheimer's disease	Gray et al [82]	25 (1.16)	306 (1.28)	5.1	5.4	2013
8	Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease	Liu et al [83]	25 (1.16)	323 (1.35)	3.7	4.0	2015
9	Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset	Hinrichs et al [84]	23 (1.07)	157 (0.66)	2.0	1.7	2009
10	Early detection of Alzheimer's disease using MRI hippocampal texture	Sorensen et al [85]	23 (1.07)	120 (0.5)	6.4	2.9	2016

[a]LCS: local citation score.

[b]GCS: global citation score.

[c]NLCS: normalized local citation score.

[d]NGCS: normalized global citation score.

[e]PY: publication year.

XSL•FO
RenderX

Table 7. Artificial intelligence classifiers and biomarker input features for highly cited local literature.

Study	PY[a]	Database	Classifier	Input features
Hinrichs et al [84]	2009	ADNI[b]	Spatially augmented LPboosting[c]	MRI[d]+FDG-PET[e]
Zhang et al [77]	2011	ADNI	Multiple-kernel SVM[f]	MRI+PET+CSF[g]
Zhang and Shen [79]	2012	ADNI	M3T[h]	MRI+PET+CSF
Young et al [80]	2013	ADNI	SVM+GP[i]	MRI+FDG-PET+CSF+APOE[j]
Gray et al [82]	2013	ADNI	Random forest	MRI+FDG-PET+CSF+APOE
Moradi et al [78]	2015	ADNI	LDS[k]+random forest	MRI+aggregate biomarker
Liu et al [83]	2015	ADNI	SAE[l]+softmax regression+SVM	MRI+FDG-PET
Sorensen et al [85]	2016	ADNI+AIBL[m]+Metropolit	SVM+logistic regression	MRI+CSF
Lee et al [81]	2019	ADNI	CNN[n]	MRI+CSF+APOE

[a]PY: publication year.
[b]ADNI: Alzheimer's Disease Neuroimaging Initiative.
[c]LPboosting: linear programming boosting.
[d]MRI: magnetic resonance imaging.
[e]FDG-PET: fluorodeoxyglucose positron emission tomography.
[f]SVM: support vector machine.
[g]CSF: cerebrospinal fluid.
[h]M3T: multimodal multitask.
[i]GP: Gaussian process.
[j]APOE: apolipoprotein E.
[k]LDS: low density separation.
[l]SAE: stacked autoencoder.
[m]AIBL: Australian Imaging, Biomarker & Lifestyle.
[n]CNN: convolutional neural network.

Analysis of Author Keywords

By analyzing keywords in a specific field, we can gain insights into its research directions and trends. In this study, the most frequent keywords identified were "Alzheimer's disease" (603/5467, 11.03%), "machine learning" (302/5467, 5.52%), "mild cognitive impairment" (166/5467, 3.04%), "biomarker" (152/5467, 2.78%), and "deep learning" (127/5467, 2.32%). Notably, "Alzheimer's disease," "mild cognitive impairment," "biomarker," and "magnetic resonance imaging" were high-frequency keywords used consistently throughout all 3 stages (2007-2023), while "deep learning" emerged in the first stage (2007-2017) and increased in the third stage (2021-2023), as shown in Table 8. A detailed time-segmented analysis of the 20 high-frequency keywords was conducted, resulting in a heat map where lighter blue indicates lower frequency in a given year and deep red indicates higher frequency; for instance, "artificial neural networks" appeared as early as 2007, decreased in frequency, and then consistently appeared at a high frequency in recent years. The keyword "Alzheimer's disease" shows a progressive increase in occurrences each year. Nearly all keywords shifted toward orange and red in 2021 and through the third phase (2021-2023). However, the keyword "support vector machine" changed from orange-red to light blue in 2023. In addition, as classification is one of the primary tasks of AI, its frequency of appearance has remained stable annually, as seen in Figure 10.

XSL•FO
RenderX

Table 8. Top 20 most frequent keywords related to the application of artificial intelligence in the dementia biomarker field.

Rank	Keyword	Occurrences (N=5467), n (%)	Period 1 (2007-2017), n (%)[a]	Period 2 (2018-2020), n (%)[a]	Period 3 (2021-2023), n (%)[a]
1	Alzheimer's disease	603 (11.03)	102 (16.92)	143 (23.71)	358 (59.37)
2	Machine learning	302 (5.52)	24 (7.95)	77 (25.5)	201 (66.56)
3	Mild cognitive impairment	166 (3.04)	43 (25.9)	40 (24.1)	83 (50)
4	Biomarker	153 (2.8)	34 (22.22)	33 (21.57)	86 (56.21)
5	Deep learning	128 (2.34)	3 (2.34)	25 (19.53)	100 (78.13)
6	Magnetic resonance imaging	126 (2.3)	31 (24.6)	29 (23.02)	66 (52.38)
7	Dementia	83 (1.52)	10 (12.05)	22 (26.51)	51 (61.45)
8	Support vector machine	78 (1.43)	23 (29.49)	22 (28.21)	33 (42.31)
9	Classification	57 (1.04)	22 (38.6)	17 (29.82)	18 (31.58)
10	Artificial Intelligence	48 (0.88)	1 (2.08)	9 (18.75)	38 (79.17)
11	Convolutional neural network	43 (0.79)	0 (0)	10 (23.26)	33 (76.74)
12	Neuroimaging	39 (0.71)	8 (20.51)	12 (30.77)	19 (48.72)
13	Random forest	39 (0.71)	4 (10.26)	10 (25.64)	25 (64.1)
14	Diagnosis	31 (0.57)	7 (22.58)	5 (16.13)	19 (61.29)
15	Feature selection	31 (0.57)	10 (32.26)	4 (12.9)	17 (54.84)
16	Amyloid-Beta	28 (0.51)	7 (25)	3 (10.71)	18 (64.29)
17	Cerebrospinal fluid biomarker	25 (0.46)	7 (28)	5 (20)	13 (52)
18	Amyloid	24 (0.44)	5 (20.83)	5 (20.83)	14 (58.33)
19	Artificial neural network	24 (0.44)	4 (16.67)	8 (33.33)	12 (50)
20	Structural magnetic resonance imaging	24 (0.44)	8 (33.33)	9 (37.5)	7 (29.17)

[a]The denominator is the n value in "Occurrences" column.

Figure 10. Heat map of top 20 high-frequency keywords related to the application of artificial intelligence in the dementia biomarker field.

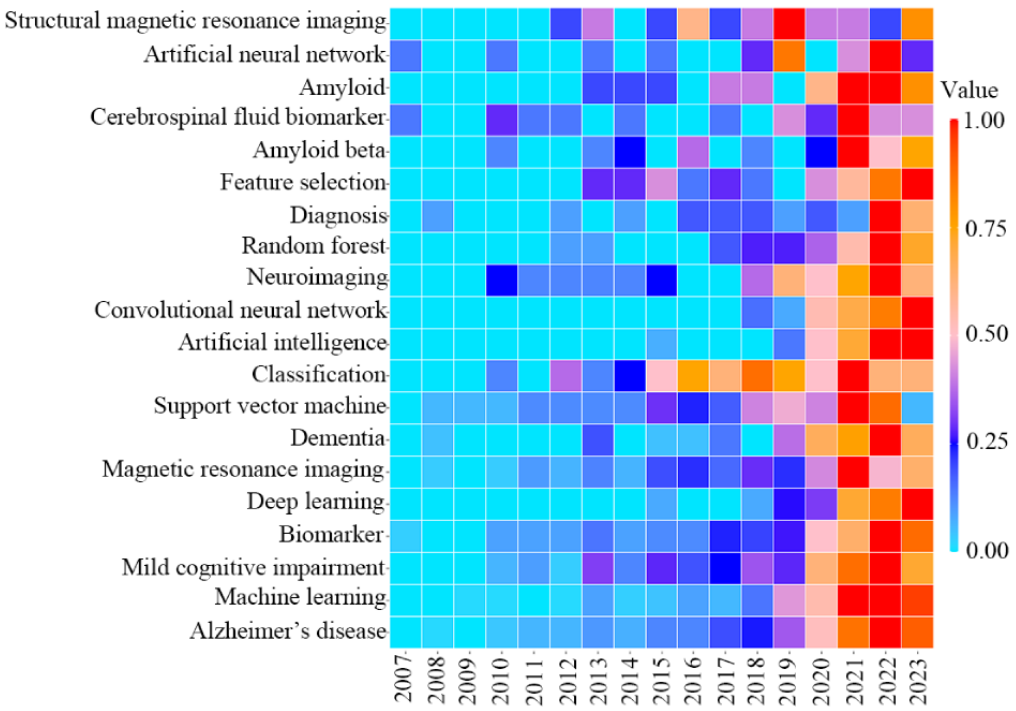

XSL•FO
RenderX

Analysis of Keyword Clusters

Identifying keyword clusters allows for an intuitive understanding of subfields within specific research areas. A total of 36 high-frequency keywords were included for clustering. These keywords accounted for 41.92% (2292/5467) of the occurrences, meeting the requirements for clustering. High-frequency keywords were analyzed using gCLUTO software to generate dendrograms and mound maps, revealing 6 distinct clusters. Each mound represents a unique cluster, with its height and volume indicating the similarity and number of documents, respectively. The colors on the mound tops signify different levels of internal SDs, with red indicating low internal SD and blue high internal SD [73]. The tops of these 6 mounds are not blue, indicating no high internal SD, especially in clusters 0 and 4, where the peaks are red and the internal SDs are lower, as shown in Figure 11.

Figure 11. Keyword clustering mound map of publications related to artificial intelligence in dementia biomarkers.

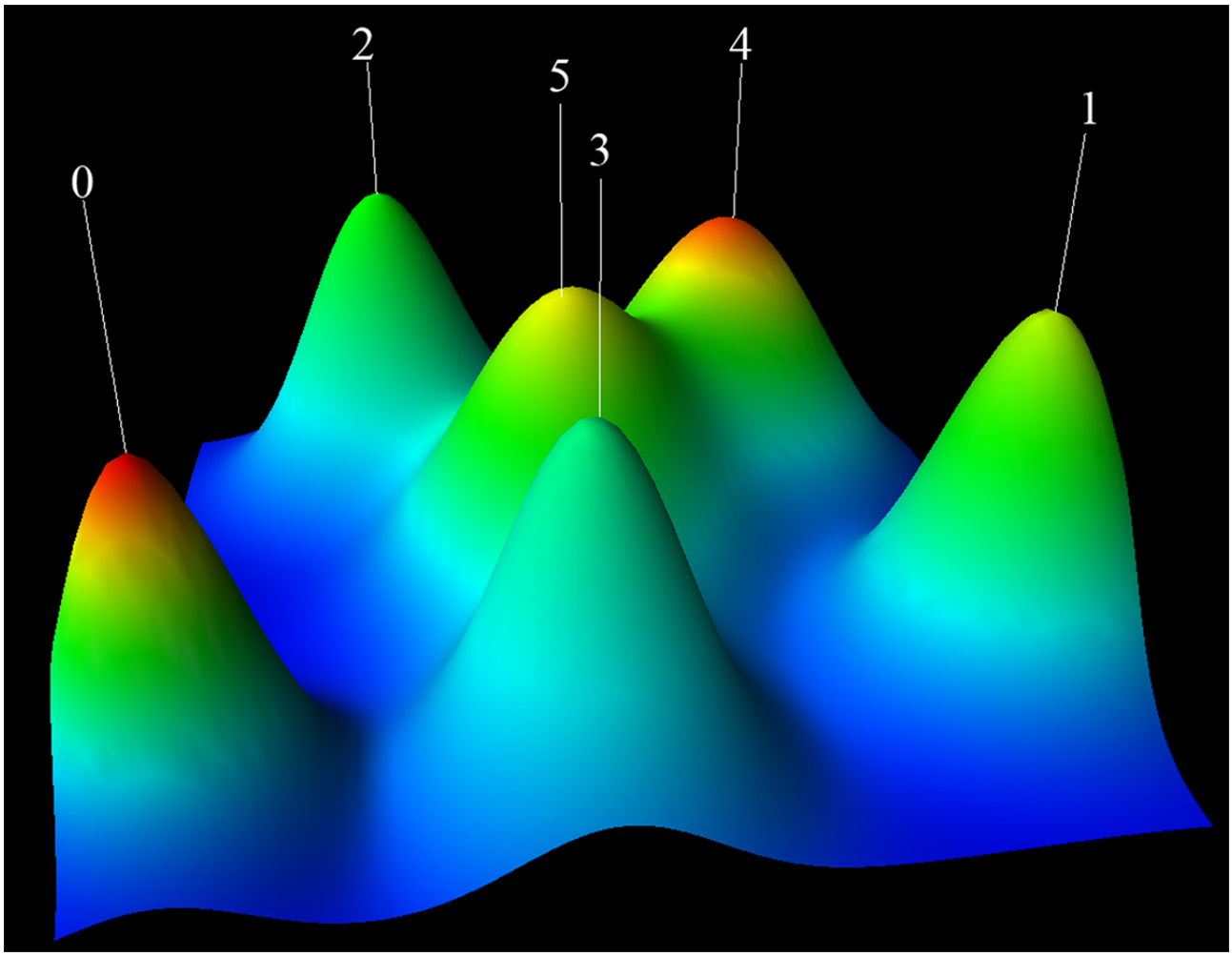

In the dendrogram, the depth of the color blocks indicates the strength of the association between the keywords on the vertical axis and those on the horizontal axis. Deep red signifies a high association strength, while white indicates a lower association strength. The dendrogram shows that AI research hot spots in dementia biomarkers primarily focus on diseases such as “Alzheimer’s disease,” “Dementia with Lewy bodies,” “mild cognitive impairment,” and “frontotemporal dementia.” Cluster 4 is the largest cluster, containing 10 keywords that can be categorized into 3 aspects: AI (“artificial neural network,” “machine learning,” “diagnosis,” and “feature extraction”), diseases (“Alzheimer’s disease,” “Parkinson’s disease,” “disease,” and “Dementia with Lewy bodies”), and biomarkers (“Electroencephalogram” and “Electroencephalography”). The theme reflected here is the application of neural networks in neurodegenerative diseases, with EEG features used for diagnosing such diseases. Cluster 5 includes 8 keywords, divided into 2 aspects: algorithms (“random forests,” “support vector machines,” “classification,” and “feature selection”) and biomarkers (“structural magnetic resonance imaging” “ADNI,” “mild cognitive impairment,” and “radiomics”). This cluster reflects the theme of traditional ML algorithms classifying biomarkers in neuroimaging. Cluster 0, the smallest cluster, contains just 3 keywords, succinctly summarizing the application of AI in FTD. Cluster 2 consists of 6 keywords mainly related to CSF biomarkers: “tau,” “beta-amyloid,” and “proteomics.” This cluster highlights the primary protein markers in CSF. Cluster 1 contains 5 keywords, divided into deep learning and imaging biomarkers. Deep learning (“deep learning,” “transfer learning” and “Convolutional Neural Network”) and imaging markers (“magnetic resonance imaging” and “hippocampus”) reflect the application of nontraditional ML methods in imaging biomarkers. Cluster 3 contains 4 keywords related to imaging markers, as shown in Figure 12.

XSL•FO
RenderX

Figure 12. Artificial intelligence keyword dendrogram in the field of dementia biomarkers. ADNI: Alzheimer's Disease Neuroimaging Initiative; PET: positron emission tomography.

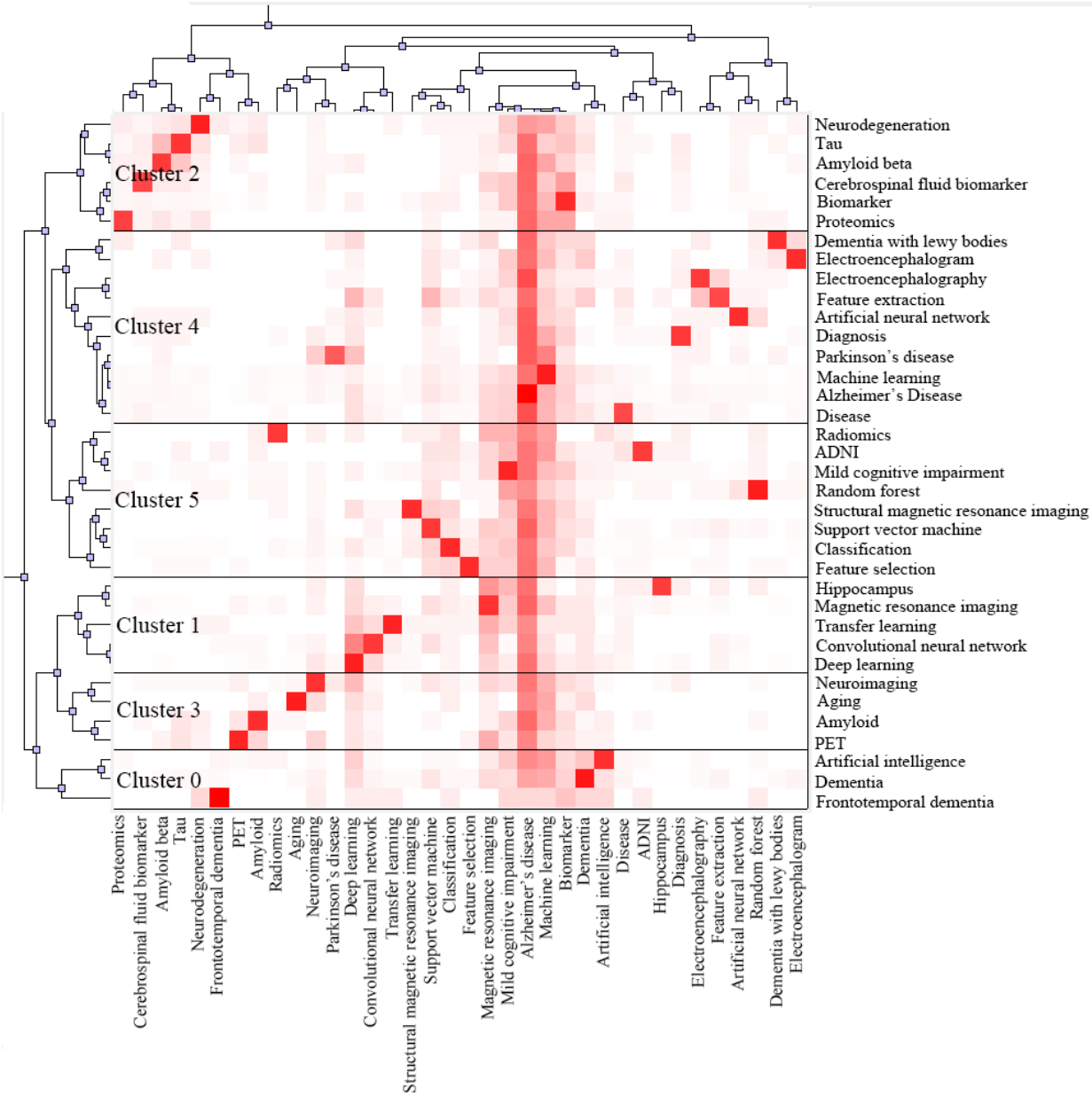

Disciplinary Analysis

We identified cross-disciplinary connections among 46 subjects, finding that each paper involved an average of 1.55 disciplines. Neuroscience and neurology (524/1661, 31.55%) were the most frequently involved disciplines, significantly more than other subjects. Engineering (128/1661, 7.71%) and computer science (126/1661, 7.59%) followed, highlighting the central role of neuroscience in this research area. Network analysis revealed 117 interdisciplinary connections, most of which were weak, indicating that direct collaboration between different disciplines is relatively limited. By contrast, collaborations within the same disciplinary group were more frequent. Specifically, the connections between neurology and geriatric medicine were the closest, followed by radiology, nuclear medicine, and medical imaging. Computer science was most closely connected to engineering. However, the connection strength between the neurosciences representing AD and the engineering and computer sciences representing AD appeared to be weak, suggesting that interdisciplinary research between these 2 fields has potential for growth, as shown in Figure 13.

XSL•FO
RenderX

Figure 13. Interdisciplinary collaboration network diagram.

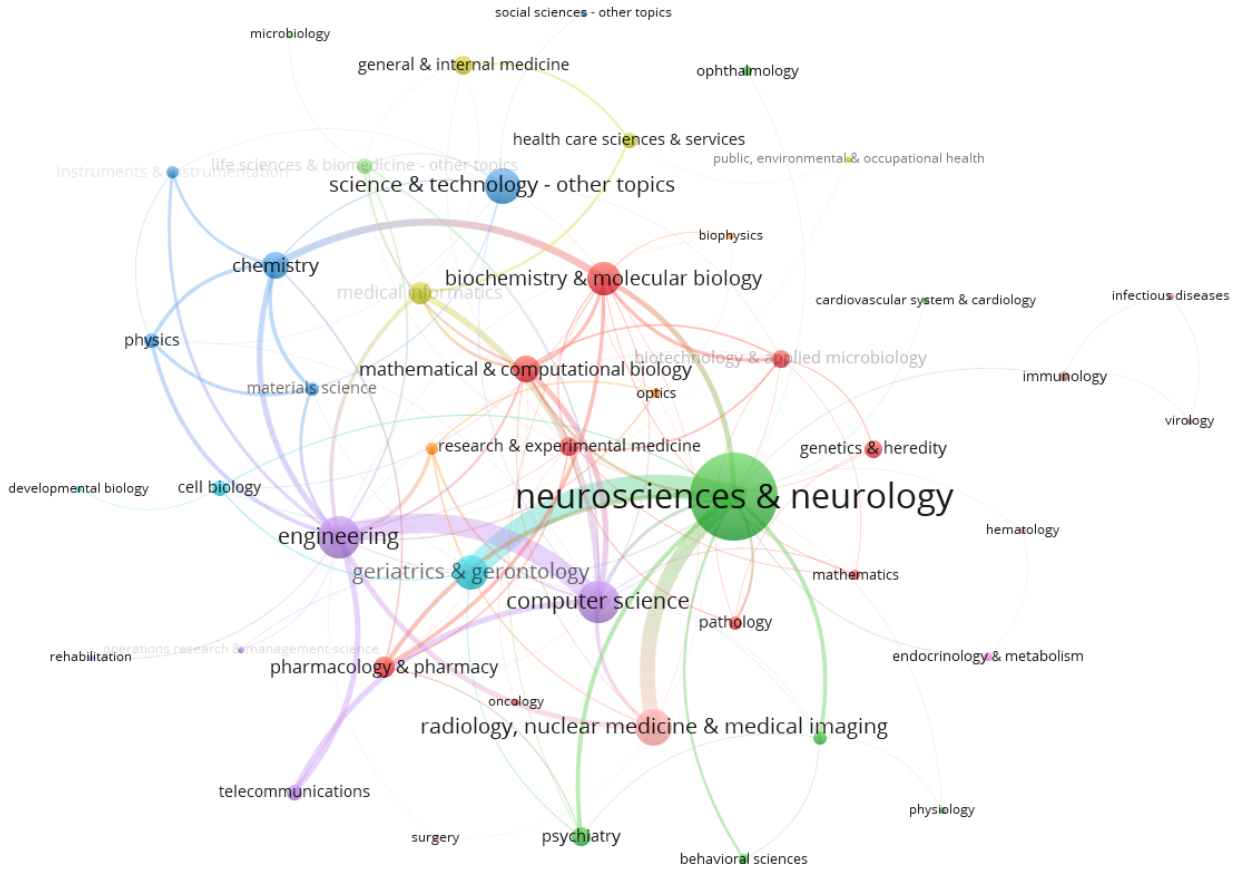

Biomarker and AI Method Analysis

Given that review articles often cover algorithms and biomarkers that overlap with those discussed in research literature, we focused on the content of 993 articles to classify biomarkers into 9 major categories based on their sources and characteristics: imaging biomarkers, CSF biomarkers, genetic markers, blood biomarkers, digital biomarkers, ophthalmic and retinal markers, neurophysiological markers, fecal and other bodily fluid markers, and other types of markers. Among the 993 studies, 973 (98%) addressed AD, 32 (3.2%) discussed FTD, 17 (1.7%) referenced DLB, and 10 (1%) focused on VaD. Overall, the main biomarkers across these subtypes were imaging, genetic, CSF, and blood biomarkers, each mentioned >100 times. Specifically, of the 1060 citations, imaging biomarkers were cited 473 (44.62%) times, genetic biomarkers 187 (17.64%) times, CSF biomarkers 148 (13.96%) times, and blood biomarkers 111 (10.47%) times.

In terms of trends, the use of AD biomarkers has been notably increasing year by year, with imaging biomarkers consistently being the most used annually. The use of genetic biomarkers surged in 2021, surpassing both CSF and blood biomarkers. CSF biomarkers have shown a fluctuating upward trend, while the use of blood biomarkers has gradually increased, recently approaching the use levels of CSF biomarkers. In addition, after 2018, various types of biomarkers have shown some intermittent growth trends. Among the other 3 subtypes, only the imaging biomarkers for FTD and the CSF biomarkers for DLB exhibited brief spikes in growth in 2022 and 2020, respectively. The trends for the other subtypes are not as pronounced, as shown in Figure 14.

The AI methods extracted from the literature were categorized into 2 main classes: supervised learning and unsupervised learning, further subdivided according to the tasks performed. In this field, classification tasks predominate. Among the algorithms used for the 4 subtypes of dementia, support vector machines (SVMs; 302/1581, 19.1%) were the most frequently applied. Various neural network algorithms (229/1581, 14.48%) ranked second overall, followed by random forests (221/1581, 13.98%). However, it is noteworthy that in 2023, SVMs were used 52 times, a stark contrast to their mere 2 mentions in keyword heat map analyses.

Regarding trends in algorithm use for AD, there has been a noticeable increase over time. Neural networks started to become popular after 2018 and surpassed SVMs by 2022. Since 2016, random forests have been used nearly as frequently as SVMs. In addition, after 2018, various types of algorithms have demonstrated a clear growth trend. In the other 3 subtypes, although there is a slight growth trend in algorithm use for FTD, the use of algorithms in DLB and VaD has not shown a significant growth trend, as depicted in Figure 15.

In the co-occurrence network of biomarkers and the 20 most commonly used AI methods, the thickness of the lines and the depth of their colors intuitively reflect the frequency and strength of their associations: thicker lines and darker colors indicate higher co-occurrence frequency and tighter connections (Figure 16). Overall, clustering, regression, and dimension reduction

XSL•FO
RenderX

algorithms are significantly less used in various types of biomarkers than classification algorithms. In AD, only 2 clustering algorithms appear among the top 20 most frequently used, with no use in other subtypes.

Figure 14. Annual use of various dementia biomarkers. (A) Dynamics of biomarkers for Alzheimer disease (AD). (B) Dynamics of biomarkers for frontotemporal dementia (FTD). (C) Dynamics of biomarkers for dementia with Lewy bodies (DLB). (D) Dynamics of biomarkers for vascular dementia (VaD).

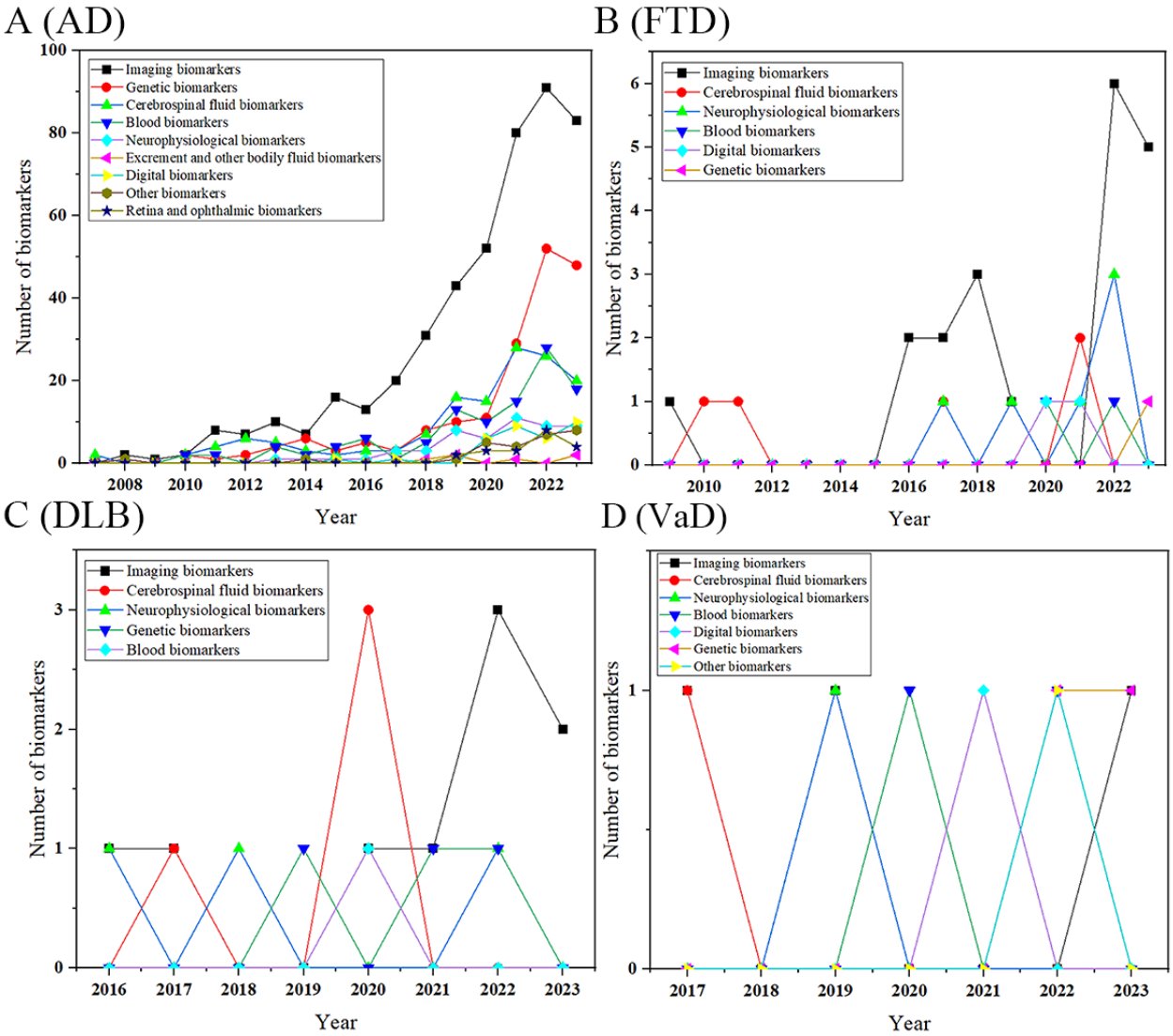

XSL•FO
RenderX

Figure 15. Annual use of various algorithms in the field of dementia biomarkers. (A) Algorithm use for Alzheimer disease (AD). (B) Algorithm use for frontotemporal dementia (FTD). (C) Algorithm use for dementia with Lewy bodies (DLB). (D) Algorithm use for vascular dementia (VaD). AI: artificial intelligence.

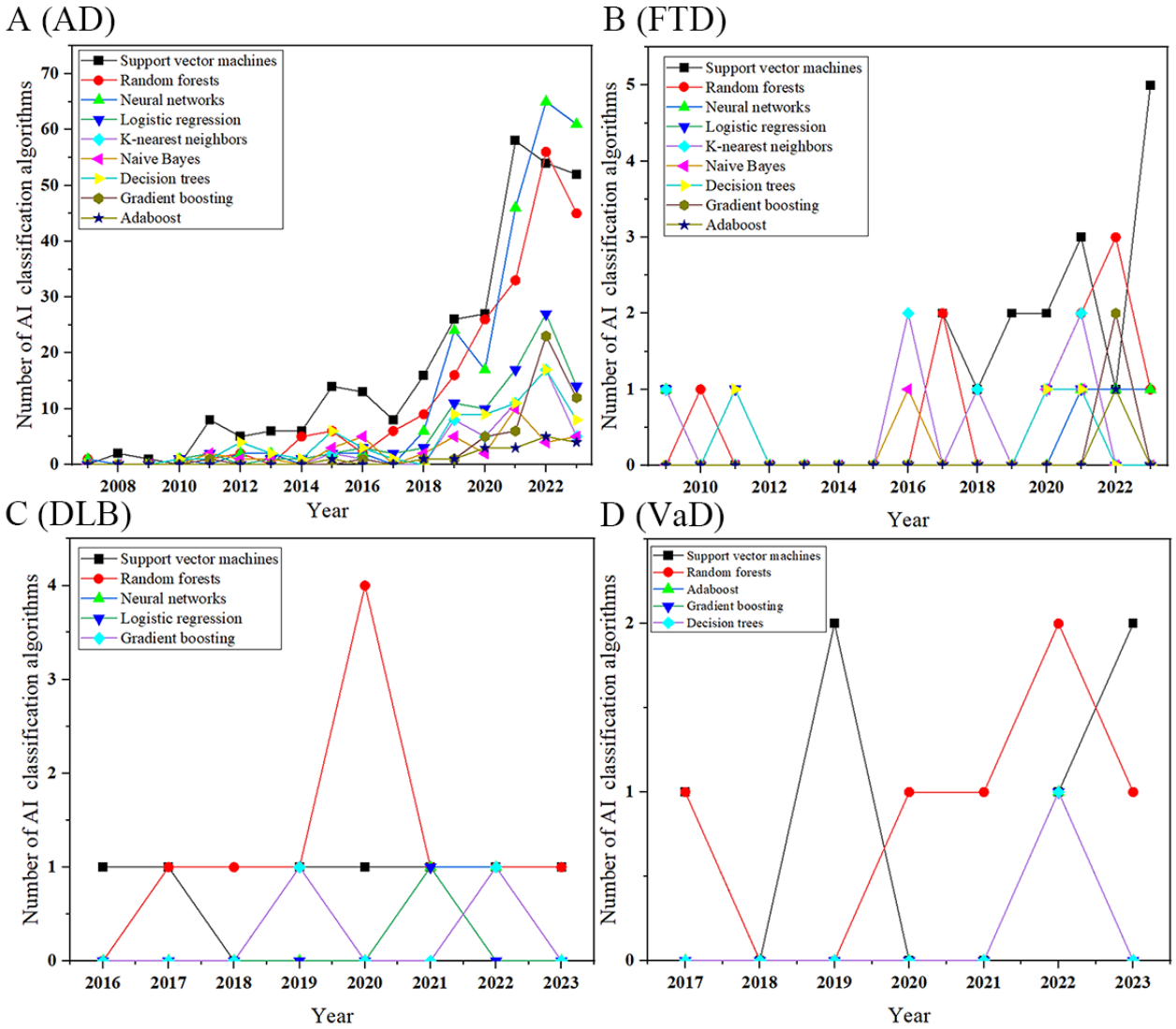

XSL•FO
RenderX

Figure 16. Graph of the correspondence between artificial intelligence algorithms and dementia biomarkers. (A) Connection between biomarkers and algorithms for Alzheimer disease (AD). (B) Connection between biomarkers and algorithms for frontotemporal dementia (FTD). (C) Connection between biomarkers and algorithms for dementia with Lewy bodies (DLB). (D) Connection between biomarkers and algorithms for vascular dementia (VaD). t-SNE: t-distributed stochastic neighbor embedding.

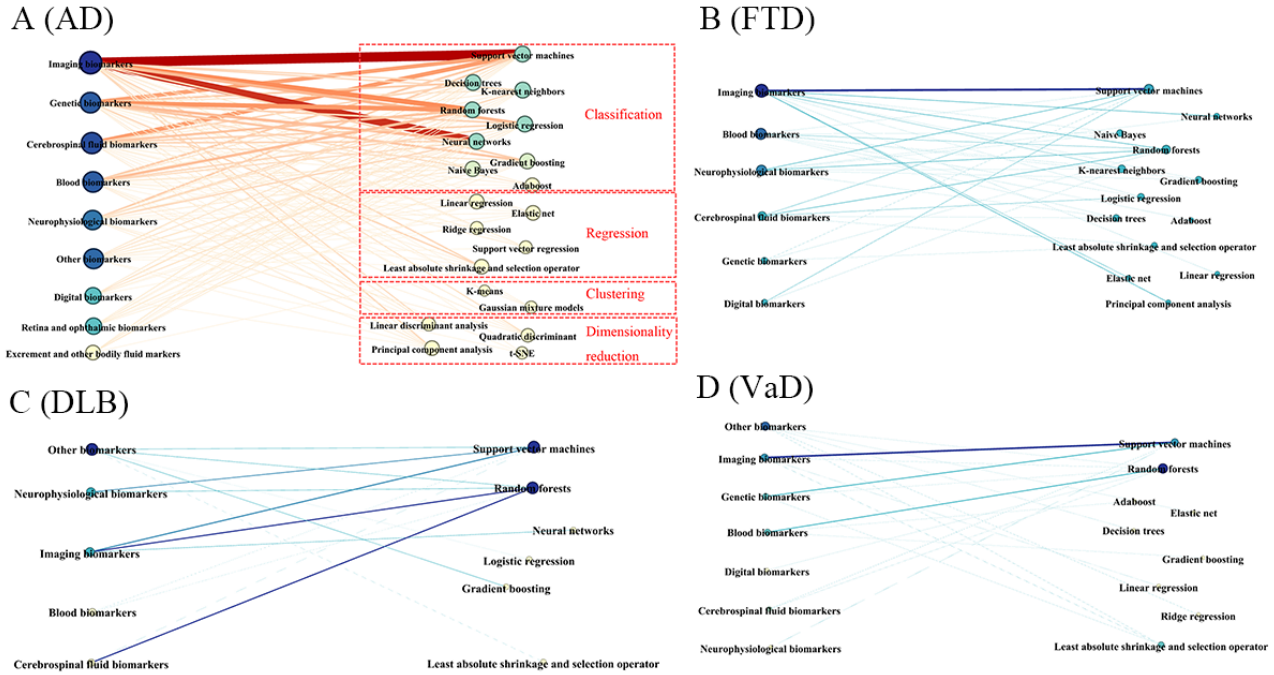

In each dementia subtype, the connections between classification algorithms and biomarkers are generally thicker and darker, especially the link between SVMs and imaging biomarkers in AD, followed by the connection between neural networks and imaging biomarkers. The thickest line in blood biomarkers is associated with random forests. In the other 3 subtypes, the connections between algorithms and biomarkers are weaker, particularly in VaD. The variety of algorithms used in FTD is second only to those used in AD, with the most notable associations being between imaging biomarkers and SVMs, which is also observed in VaD. In DLB, random forests appear to be more frequently used with imaging and CSF biomarkers, as illustrated in Figure 16.

Discoveries of New Biomarkers

Overall, there have been significant new findings in dementia biomarkers. A total of 244 research reports have identified new biomarkers: 231 (94.7%) for AD, 3 (1.2%) for FTD, 5 (2%) for DLB, and 5 (2%) for VaD. Of these, 211 (86.5%) new biomarkers were discovered after 2018. Among these 211 biomarkers, imaging biomarkers and genetic biomarkers have been found most frequently, with 68 (32.2%) and 70 (33.2%) new findings, respectively, followed by blood biomarkers with 34 (16.1%) new findings. New biomarkers in emerging areas such as ophthalmology and retinal studies as well as digital biomarkers have also been identified in recent years, as shown in Figure 17.

XSL•FO
RenderX

Figure 17. The number of studies on new biomarkers discovered for various subtypes of dementia using artificial intelligence.

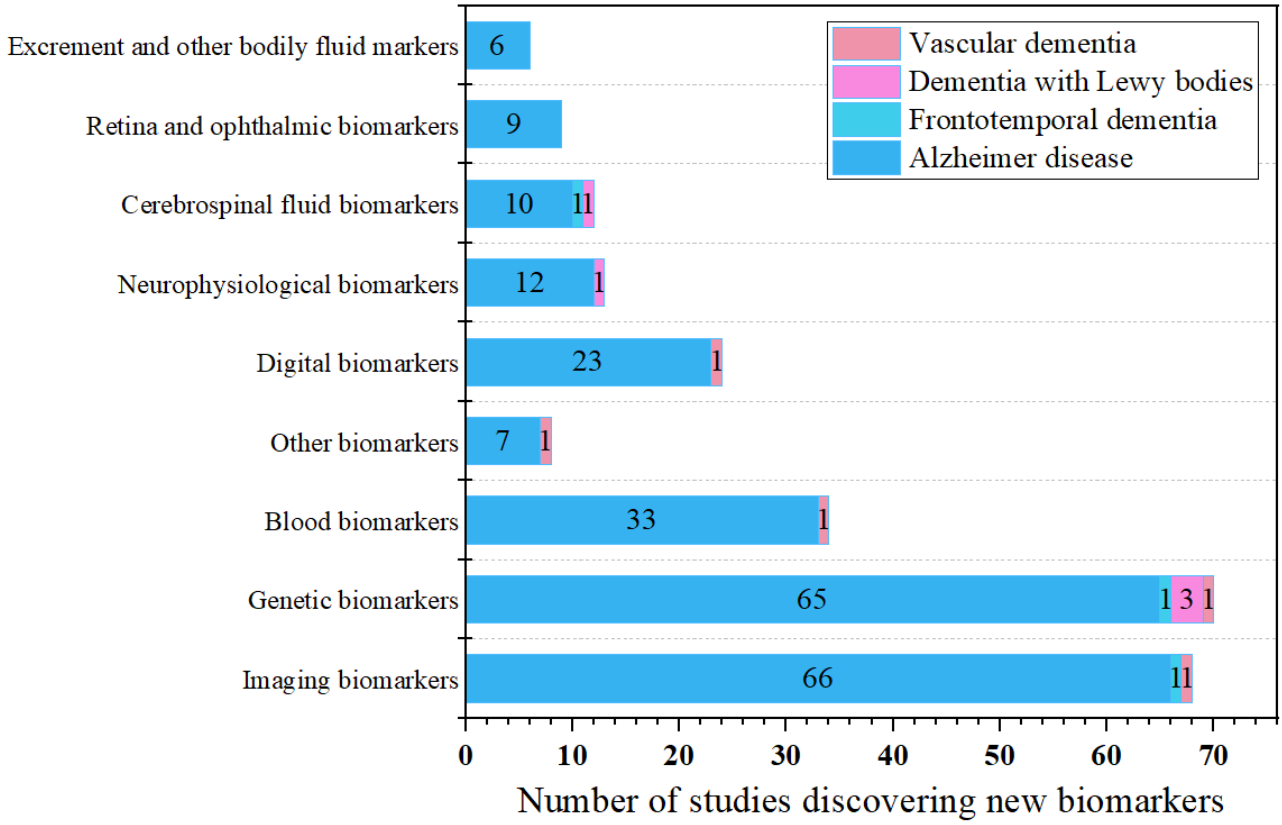

Discussion

Summary

Compared to other bibliometric studies on dementia biomarkers [23,24], our research not only reveals basic data, such as publication volumes, institutions, and national trends, but also delves deeply into the phenomena of author turnover and collaboration network flaws and more specifically highlights the contributions of prolific authors and key national efforts. In addition, we have successfully captured and quantified the developmental trends and dynamics of various biomarkers. In contrast to another study [12], we have detailed the contributions of various algorithms in this domain and followed the latest advances in biomarkers. Our analysis supports earlier research [31,33] regarding the prevalence of SVMs in imaging biomarkers and further augments the significance of other algorithms in biomarker research. Specifically, through mining analyses of high-frequency author keywords, keyword clustering, and literature content, we identified research hot spots, including the diagnosis and classification of dementia subtypes and neurodegenerative diseases, an exploration of CSF proteomic markers, and the application of traditional algorithms and neural networks in imaging biomarkers. SVMs, neural networks, and random forests are widely used as popular algorithms. Random forests are most frequently used in blood and genetic biomarkers. Newly discovered biomarkers primarily focus on imaging, genetics, and blood domains. We discuss these key findings in detail in the following subsections.

Publication Output and Growth of Research Interest

Overview

In dividing the development stages of research on AI in dementia biomarkers, the analysis went beyond just publication volume and annual growth rates. It also considered key factors such as changes in publication numbers of prolific authors, fluctuations in high-frequency keywords each year, and the evolution of algorithms observed in 973 research papers. This comprehensive analysis supported the definition of 3 development stages, outlined in the following subsections.

Initial Exploration and Methodological Advances (2007-2017)

This stage is characterized by limited publications and growing interest in AI in dementia biomarkers. Key reasons included nascent AI technology in the field, limited availability of data sets [12,49,86-88], and immature development of biomarkers; for example, early PET radioactive tracers were not yet capable of specifically measuring the burden of neurofibrillary tangles and other tau protein abnormalities [89].

Rapid Development Period (2018-2020)

This stage marked a turning point with a surge in high-quality research methods. This was driven by the rise of deep learning [90], multimodal biomarker use [91], and expansion of public data sets (eg, ADNI) [92,93].

Stable Development Period (2021-2023)

This stage is characterized by a substantial increase in research volume, indicating a period of fast growth. Advances in image

XSL•FO
RenderX

segmentation [94], deep learning algorithms [95-97], large public data sets [12], and digital biomarkers [98] contributed to this growth.

Enhance Collaboration Among Authors and Maintain Their Interest in Research

The field in question has attracted considerable attention from researchers, with the majority being newcomers who entered after 2019 in particular. This influx of new researchers indicates a strong interest within the scientific community toward this field. According to the Price law, the current output from core authors has not yet reached 50% of the total output, suggesting that a core group of authors has not been fully established. More than half of the current core authors (372/663, 56.1%) are new researchers from recent years, an indication perhaps that more researchers will emerge as leading figures in this domain. However, an important observation is that 69.41% (994/1432) of the researchers active before 2019 have not continued to produce related research, potentially indicating a decline in interest or a shift in research focus. While the contributions of most authors may be transient, a small number of researchers, such as the 10 highly productive researchers identified, have maintained consistent research output. Sustained knowledge accumulation in a research field greatly relies on ongoing studies and the establishment of a core group of authors [59].

Furthermore, establishing collaboration networks is a critical issue. Although most researchers (451/663, 68%) have formed collaborative groups, the majority of these networks (39/57, 68%) are still underdeveloped. Given the potential of AI in processing and analyzing large-scale biomedical data, as well as the need for the validation and correct use of new biomarkers, close collaboration among computer scientists, neuroscientists, and biostatisticians becomes particularly important [99]. The Brookings Institution in the United States also highlights the critical role of interdisciplinary collaboration in research innovation and standard setting within the AI field [100]. Therefore, both core authors and new researchers need to strengthen collaborations, especially interdisciplinary ones. New researchers, in particular, face challenges such as geography and costs in the process of interdisciplinary collaboration [101,102], and they often lack a deep understanding of other disciplines, which hinders the smooth progress of collaboration.

Interdisciplinary Collaborative Innovation

In the construction of cross-industry innovation systems between AI and medicine, AI often plays the role of outbound innovation, introducing AI technologies into the medical field. Conversely, the medical sector tends to embrace inbound innovation, adopting AI to address medical issues. This division primarily stems from the medical sector's needs for diagnosis and treatment [102]. However, the ultimate goal is to achieve a close integration of both domains, advancing the integration of science and technology by developing new knowledge through collaboration with partners from various industries [103].

In the medical field, leadership teams proactively seek external knowledge based on their experience and standards to build interdisciplinary collaborations; for example, radiomics research teams can seek collaboration with partners skilled in image segmentation techniques. In addition, the shift from a closed to an open team model is crucial and involves adopting analogical thinking. This approach can draw from successful interdisciplinary collaborations already established in the medical field; for instance, the field of cardiology has set a commendable example with its multi-institutional interdisciplinary collaborations on AI [104]. For the AI sector, the main challenges lie in technological support and innovation, necessitating enhancements to algorithms and the development of new technological frameworks in response to medical needs. This not only requires medical knowledge but also entails the acquisition, assimilation, transformation, and development of knowledge within interdisciplinary teams. These learning processes demand active participation from team members and standardized sharing of information and knowledge, thereby facilitating advances in AI and its commercialization. Establishing connections between different disciplinary teams and building bridges for communication across fields are essential starting points. Cross-disciplinary academic conferences and web-based public courses serve as effective means to construct initial cooperative bridges. In addition, the establishment of cross-departmental digital platforms enables researchers to access and collaboratively analyze existing research data, exemplified by several searchable professional websites related to AI medical devices [105], fostering the development of tacit cooperation. Furthermore, several forward-thinking higher education institutions have already begun to informally incorporate the principles of AI into undergraduate courses through lectures. A new graduate module on radiology AI has also been established [106]. At Stanford University in Stanford, California, United States, leaders across various disciplines have formed interdisciplinary teams dedicated to teaching and researching AI to address health care issues [107].

Despite these measures aiding in the establishment of initial collaborative networks, the involvement of government and social enterprises as intermediaries is necessary to overcome informational disparities and promote deeper exchanges. Forming multidisciplinary societies, such as dedicated biomarker research associations, and enhancing interdisciplinary integration through research funding and incentive mechanisms are crucial measures to foster cooperation. The participation of diverse organizations, including universities, medical institutions, and corporations, will provide a broader scope and vision for the development of these associations. Finally, we also advocate for interdisciplinary information exchange within the respective fields of medicine and AI. Although this may provoke some potential internal competition, the convenience of this communication method and the potential for innovative benefits significantly outweigh the challenges it presents.

Regional Proximity Collaboration

Regional proximity has long been recognized as a crucial objective factor influencing innovation activities. Participants concentrated in a specific area benefit from the knowledge externalities produced by short distances, facilitating the exchange of knowledge between proximate entities and thereby fostering the development of innovation and the flow of tacit

XSL•FO
RenderX

knowledge [108]. The convenience of such networks, coupled with cultural and institutional similarities, helps to keep cooperative networks vibrant [109]. For newcomers to the field, considering the advantages brought by regional proximity is key to building a stable foundational cooperative network. As the importance of complementary capabilities in partners continues to increase [110], seeking technological complementarity has become essential for maintaining active and robust cooperative networks. Particularly in the field of dementia research, the high heterogeneity of the disease requires us to construct knowledge networks from a global perspective, making full use of the differences in AI technologies across different countries. Relying solely on cooperation networks within a single country may overlook the value of global and nonlocalized knowledge networks, hindering the further integration of technology; for example, constructing diversified data sets will benefit from the inclusion of different regions and ethnicities. For transnational collaboration, the successful cases across multiple European countries serve as instructive examples. These nations have demonstrated the advantages of collaboration facilitated by regional proximity. Moreover, collaborating with high-output countries in the field is also a wise choice because these countries typically possess advanced technology and extensive resources. These nations are distributed across various continents, playing a significant radiative role, thus providing a more diversified array of options for establishing cooperative networks. Therefore, we recommend building foundational cooperative networks based on the principle of regional proximity and actively seeking partnerships with technologically leading countries to stimulate sustained network activity. In addition, governments and research institutions should support the construction of these transnational cooperative networks by increasing research funding and establishing incentive mechanisms to ensure the continuity and development of research.

Preferred Journals

In the field of dementia biomarkers, AI-related research has identified 12 core journals. These journals rank well across multiple platforms, reflecting the favorability of AI research in dementia biomarkers among numerous prestigious publications, including well-known journals such as *Alzheimer's & Dementia* and *NeuroImage*. Dual-map overlays of the journals indicate extensive coverage of topics such as psychology, education, molecular biology, medicine, genetics, and immunology in this field. Therefore, scholars eager to delve into AI in dementia biomarkers should follow these high-output, influential journals. Simultaneously, they should explore interdisciplinary reports aligned with their research interests and content. This approach will help them comprehensively understand the latest developments and trends in the field.

Leading Countries and International Collaboration

Currently, dementia biomarker research involving AI has seen participation from 74 countries worldwide, demonstrating widespread international interest. In particular, the European region exhibits a higher level of attention toward this type of research, which correlates with its dementia incidence rates exceeding the global average at 1123 cases per 10,000 individuals [111], underscoring the urgent need to address this challenging issue. Similarly, the higher rates of dementia incidence and mortality in the majority of high-producing countries reflect how research is influenced by the dementia situation in each country. However, the concentration of research activities is closely related to the scientific capabilities and resource allocation of specific countries. The leading positions of the United States, China, and the United Kingdom in this field not only reflect these countries' strong capabilities in research infrastructure, funding support, and technological innovation but also highlight their proactive roles in addressing global health challenges. This situation also suggests a potential issue of uneven resource distribution globally and the challenges other countries and regions may face in enhancing their research capabilities.

Therefore, to further enhance the contribution and impact of global research on dementia biomarkers, it is necessary to take measures to strengthen international cooperation, promote resource sharing, and encourage countries to increase research investment, especially in countries and regions with fewer resources. Fortunately, in terms of international collaboration, most high-producing countries have >100 instances of cross-border cooperation, indicating a strong willingness for international collaboration, particularly the United States and the United Kingdom, which lead not only in the number of countries they collaborate with but also in the frequency of such collaborations. Their implementation of AI in health care provides guidance for development and regulation for other countries; for example, the National Institutes of Health in the United States, in collaboration with multiple countries, has established one of the largest public AD data set in the world (ADNI) [112], offering data support for numerous studies. The United Kingdom's *Code of Conduct for Data-Driven Health and Care Technology* provides funding, research opportunities, and tools for researchers in low- and middle-income countries, encouraging their participation in AI research and fostering connections [113]. By contrast, although China is the second largest producer of research outputs globally, it has fewer instances of international collaboration. This is related to China's later start in AI compared to the United States and the United Kingdom, with its current AI strategy focusing more on the localization and training of AI talents [114], and international cooperation has not yet fully taken off. However, it cannot be denied that China possesses many research institutions and leading funding support, harboring significant potential for international collaboration that will play a substantial role in future international efforts.

For researchers, this information is valuable for considering international collaborations, applying for visiting scholar positions, or participating in educational projects. For nations, actively engaging with leading countries in this field and establishing collaborations can foster development in this area, particularly for low- and middle-income countries that have high dementia rates but lack AI technology.

Highly Cited Papers

A substantial body of ML research has focused on integrating brain imaging with both structured and unstructured clinical

XSL•FO
RenderX

data to predict disease progression. The integration of multimodal data, such as imaging, CSF biomarkers, and demographic data, is considered one of the best approaches to address data heterogeneity [115]. This study identified 10 highly cited papers that provide significant insights into the analysis and application of various multimodal biomarkers, especially in terms of feature selection and the construction of new multimodal data sets. Among them, Zhang et al [77] adopted a multimodal classification strategy and a multikernel SVM in 2011 to enhance the classification performance for AD and mild cognitive impairment, demonstrating higher accuracy and sensitivity. This approach, based on the construction and kernel combination of multimodal heterogeneous biological data, overcomes the limitations of traditional studies that rely on a single biomarker, offering a more comprehensive and precise analytical framework. Subsequently, Zhang and Shen [79] introduced a method that combines multimodal data and multitask learning to jointly predict multivariate regression and classification variables from baseline multimodal data, providing new perspectives and tools for subsequent research. Gray et al [82] used the random forest algorithm to extract paired similarity measures, constructing a manifold-based representation that integrates information from multiple modalities. This approach enhances the accuracy and efficiency of classification. For newcomers to this field, a thorough examination of these high-value publications will facilitate a deeper understanding and inspiration.

Research Trends in Clusters of Highly Productive Authors

Since 2018, research on various biomarkers has been gradually increasing. Among them, 10 prolific authors are noteworthy. The research themes of Shen, DG, and Zhang, DQ, are highly similar. Their research teams mainly focus on imaging biomarkers, particularly the multimodal fusion of brain structural data, and they lean toward algorithmic improvements. Their research, initially centered on multikernel SVM studies, has progressively shifted toward studies using deep learning architectures [116]. Similarly, the research team led by Zetterberg [117] also shows interest in imaging biomarkers but tends toward newer biomarkers. In recent research, a deep learning–based model developed by them used computed tomography imaging biomarkers to distinguish people with dementia from healthy individuals with performance similar to MRI [117]. In addition, in brain age difference studies, the use of algorithms such as extreme gradient boosting has revealed a positive correlation between brain age difference and NFL [118]. Jack, CR, also exhibits a certain interest in imaging biomarkers; he used SVMs to achieve multimodal fusion of imaging data as early as 2010 [119]. However, in recent years, his research interests have diversified: he is not only using deep learning for predicting brain age [120] but also exploring blood biomarkers [121] and genetic biomarkers [122], which have become increasingly popular in recent years.

Furthermore, the collaborative group led by Morris, JC, is the most prolific in recent years, with their research primarily focused after 2021. Moreover, they seem to have a broader interest in emerging biomarkers such as imaging-based brain age difference [123], digital biomarkers based on driving behavior [124], gut microbiota [125], and MTBR-tau243 in CSF [126]. The research team led by Saykin, AJ, has been more focused on genetic biomarkers in recent years. They have used deep learning methods to identify potential AD-risk single nucleotide polymorphisms, discovering that rs561311966 (located in the *APOC1* gene) and rs2229918 (located in the *ERCC1*/*CD3EAP* genes) are significant factors influencing AD risk [127]. Similarly, the research team led by O'Bryant, SE, tends to focus on blood biomarkers, initiating the search for dementia-related blood biomarkers using random forests in 2011 [128]. Their recent research has found that a combination of serum and plasma biomarkers has higher predictive performance than serum or plasma biomarkers alone, providing a new approach to diagnosis using blood biomarkers [129]. Different prolific research groups seem to have certain differences in research interests and trends, but they generally converge on the study of imaging, blood, genetic, and some emerging biomarkers. Regarding algorithm use, besides traditional algorithms, there is also a growing trend toward the use of deep learning algorithms. Keeping tabs on the latest research trends concerning these highly prolific authors will aid in grasping the cutting-edge developments in various types of biomarkers.

Research Hot Spots

Research on Dementia Subtypes

In the context of AI, the research focus on different dementia subtypes varies significantly, with AD dominating the field. This predominance is primarily due to the high prevalence of AD and its significant societal impact, which have attracted more resources and efforts. By contrast, research on other dementia subtypes started later, and most studies are either based on AD or aimed at differentiating from AD. Independent research on other subtypes such as FTD has shown some modest increases in the number of studies and algorithm use recently, but no similar trend is evident for VaD and DLB. Therefore, for VaD and DLB, we only discuss their latest biomarker findings based on AI.

New Biomarkers for FTD

Distinct from AD, FTD represents the second most common subtype of dementia. Since 2015, research on FTD has shown a growing trend. However, the diagnosis of FTD remains challenging due to the high symptom overlap with AD. Studies have shown that imaging biomarkers can distinctly differentiate AD from FTD [130,131]. This success is partly due to FTD subtypes affecting different brain regions; for example, behavioral variant FTD is typically associated with atrophy in the frontal and anterior temporal lobes; progressive nonfluent aphasia mainly impacts the left inferior frontal gyrus, leading to motor speech disorders; and semantic dementia primarily affects the left anterior temporal region [132]. This also explains why the use of imaging biomarkers is more widespread in FTD than in other biomarkers. Recently, significant white matter (WM) damage revealed by ML has been validated as an effective imaging biomarker for FTD, with WM degeneration in behavioral variant FTD being more pronounced than in AD, supporting the hypothesis that neurodegenerative changes in FTD start in the WM [133,134].

XSL•FO
RenderX

In addition, several CSF cobiomarkers have been proposed for FTD, such as NFL chain and TDP-43 [135]. In recent research, Bergström et al [136] used the least absolute shrinkage and selection operator (LASSO) and random forest methods to analyze protein data obtained from CSF samples, identifying NFM, aquaporin-4, neuronal pentraxin 2, and the neurosecretory protein VGF as potential diagnostic tools for FTD. In the genetic domain, Magen et al [137] developed a nonlinear predictive model based on gradient boosting trees, successfully identifying 13 microRNA (miRNA) features, offering new possibilities for early diagnosis and treatment of FTD. In other biomarker research, EEG features have achieved an accuracy rate of 73% in distinguishing FTD from AD using decision tree algorithms [138]. However, the current number of biomarkers available for FTD is still limited, necessitating more research to develop novel biomarkers to aid in distinguishing FTD from other types of dementia, particularly AD [139]. In this process, the application of AI will undoubtedly play an increasingly vital role.

New Biomarkers for DLB

The pathological hallmark of DLB is the presence of Lewy bodies containing alpha-synuclein in the neocortex and limbic areas [140]. Research has identified several potential biomarkers for the diagnosis of DLB, such as alpha-synuclein, Aβ42, and phosphorylated tau (p-tau) [141]. However, autopsy results indicate that 50% to 80% of DLB cases show cortical Aβ deposits similar to those in patients with AD [142]. In addition, the early cognitive symptoms of DLB highly overlap with those of AD, posing a challenge for clinical diagnosis. van Steenoven et al [143] used a random forest algorithm to identify 6 proteins in CSF—VGF, SCG2, neuronal pentraxin 2, NPTXR, PDYN, and PCSK1N—as candidate biomarkers for DLB. Moreover, EEG has become a research focus due to its accuracy in reflecting brain electrophysiological activity, surpassing neuroimaging and CSF biomarkers [144]. EEG has revealed specific electrophysiological patterns associated with DLB, particularly a dominant frequency of <8 Hz, which helps differentiate DLB from AD in 85% to 100% of patients [145]. Suzuki et al [146] used an EEG-based ML algorithm, MC-004, to distinguish DLB from AD with an accuracy rate of 79.5%. Recently, changes in miRNA expression have also been linked to various neurodegenerative diseases, providing new hope for diagnosing and differentiating DLB; for example, the pathological link between the genes *BCL2L1* and *PIK3R2* has been further supported [147]. The latest research by Soto et al [148] using ML has revealed 12 miRNAs with continuous expression dysregulation throughout the development of DLB. Zhou et al [141] used logistic regression and SVMs to build a predictive model and identified 5 potential DLB hub genes—*SRF*, *MAPK1*, *YWHAE*, *RPS6KA3*, and *KDM7A*—that may provide new biomarkers for the diagnosis and treatment of DLB.

New Biomarkers for VaD

Research on VaD is the least extensive, primarily because its pathological mechanisms involve complex issues related to cerebral vascular health, unlike specific intracellular pathogenic protein accumulation seen in other dementias, such as AD. VaD is mainly associated with cerebrovascular disease, and its onset and progression are often abrupt. This makes the development of biomarkers for VaD more challenging than for other types of dementia. Some researchers believe that VaD may be linked to systemic autoimmune diseases, and through bioinformatics and ML methods, genes such as *C1QA*, *CD163*, *LY96*, and *MS4A4A* have been identified as potential biomarkers for the link between VaD and systemic lupus erythematosus [149]. In addition, other studies have identified potential biomarkers for VaD, including digital clock drawing tests [150], lipids [151], the *REPS1* gene [152], and brain tissue volume [153]. While these findings have opened new research avenues, no class of biomarkers has been widely applied in clinical settings to date. Future research needs to further validate these potential biomarkers and explore more from a multiomics perspective. This could help establish reliable biomarkers, thereby enhancing the diagnostic accuracy and treatment efficacy for VaD.

New Biomarkers for AD

Hippocampal atrophy, cortical thinning, and ventricular enlargement are classic manifestations of AD in MRI scans. The use of brain PET scans to detect tau and Aβ proteins has been extensively applied in ML models, with their effectiveness continually validated. With advances in imaging technology and AI, we can now process high-dimensional data, identify relevant patterns in complex data sets, and decipher the brain's intricate network structures. In the hippocampal region, multivariate morphometry statistics [154], feature sets [155], and principal curvature ratios [24] provide new perspectives for analyzing structural changes in the AD brain. Compared to studies on physical structural changes, those on brain functional connectivity have revealed insights into the brain's functional organization and operational mechanisms, becoming a vital resource for exploring new biomarkers; for instance, dynamic functional connectivity obtained from functional MRI [156] and correlated transfer function connectivity strength [157] have demonstrated potential as biomarkers. Zhao et al [158] have confirmed the excellent feature selection performance of dynamic functional connectivity by analyzing the functional connections between gray matter and WM and using SVMs for feature evaluation. Recent studies, such as that by Zhu et al [159], have combined SVMs with the *apolipoprotein E* (*APOE*) genotype, CSF biomarkers (Aβ, tau, and p-tau), and neuroimaging markers, finding that connections between the left insula and the left posterior middle temporal gyrus, the left medial superior frontal gyrus, and the right lingual gyrus are significant for cognitive functions. Sadiq et al [160] demonstrated the potential value of these signals in diagnosing AD by using SVMs to process nonfractal connectivity features extracted from resting-state functional MRI data through wavelet-based fractal analysis. In addition, dynamic connectivity anomalies between the hippocampus and the default mode network [156], as well as functional connectivity abnormalities in the posterior brain regions [28] and corticosubcortical circuits [161], have been identified as newly discovered key biomarkers.

Brain age discrepancies, evaluated by comparing the deviation of predicted brain functional connectivity age from actual age, have been shown to correlate with genetic markers such as *APOE ε4* alleles across multiple study cohorts [118]. Lee et al

XSL•FO
RenderX

[120] used a deep learning model based on structural MRI and fluorodeoxyglucose-PET to predict brain age, demonstrating that brain age differences can effectively predict the transition from no cognitive impairment to mild cognitive impairment or AD. Zhang et al [162] used SVMs and arterial spin labeling to reveal significant declines in blood flow in the posterior cingulate cortex and precuneus, providing evidence for regional cerebral blood flow as a new biomarker. Moreover, changes in the microstructure and integrity of brain WM fiber tracts captured via diffusion tensor imaging, such as changes in the parietal WM, limbic and high-order association areas WM, medial temporal WM, posterior cingulate and precuneal WM [163], and whole-brain WM fiber connectivity [164], also show great potential for predicting AD precursors. These imaging biomarkers discovered through AI offer significant research prospects, and their application could aid in the early diagnosis and development of treatment strategies for AD.

In genetic biomarker research, recent years have seen the identification of multiple genes associated with AD using large training data sets and complex analyses of genetic relationships. Zhuang et al [165] used methods such as random forest and LASSO to identify, for the first time in AD research, 10 biomarkers related to immune infiltration. Similarly, Zhou et al [166] successfully identified 5 potential AD predictive biomarkers—*FAM71E1*, *DDB2*, *AP4M1*, *GPR4*, and *DOC2A*—using transcriptome-wide association studies and weighted gene coexpression network analysis, combined with random forest and SVM algorithms. In addition, recent research has shown that genes such as *BAG2*, *HSC70*, *STUB1*, and *MAPT* are closely related to the occurrence and progression of AD [167]. Small noncoding RNA molecules, or miRNAs, have also garnered significant attention in recent years for their multifaceted roles in AD development, including regulating the formation of Aβ plaques, phosphorylation of the tau protein, and involvement in inflammatory processes [168]. Tan et al [169] used an integrated framework of statistical and ML methods to perform differential expression analysis of miRNA, identifying 3 highly significant and relevant miRNA candidates: has-miR-6501-5p, has-miR-4433b-5p, and has-miR-143-3p. Likewise, Alamro et al [170] identified 6 AD-related miRNAs using ML and deep learning models. In addition, ferroptosis has been implicated in the pathogenesis of AD [171], and Deng et al [172] used various ML methods to build models and identify 5 genes related to ferroptosis (*RAF1*, *NFKBIA*, *MOV10L1*, *IQGAP1*, and *FOXO1*). Wang et al [173] used a random forest classifier to screen 12 differentially expressed genes associated with ferroptosis.

Although these discoveries are significant for understanding the genetic foundation of AD, the new gene biomarkers identified are often limited to specific gene data sets and lack validation across broader data sets. Furthermore, additional comprehensive studies are needed to elucidate the specific mechanisms of these genes and their impact on the pathological progression of AD.

In AD diagnostic research, biomarkers such as the ratio of Aβ42/Aβ40 in plasma and p-tau proteins (p-tau181, p-tau231, and p-tau217) have demonstrated high diagnostic accuracy, further supporting their potential as noninvasive diagnostic tools. In addition to these biomarkers, which are also present in CSF, changes in the expression of the RTN1 protein in the blood, related to the production of Aβ and BACE1 enzyme activity, may affect the pathological process of AD [174]. Yu et al [175] achieved a diagnostic accuracy rate of 99% using a random forest model constructed with 8 different serum proteins, providing potential new biomarkers for a noninvasive serum diagnostic platform for AD. Moreover, discoveries of more related blood biomarkers, such as tumor necrosis factor-alpha and monocyte chemoattractant protein-1 [176], plasma levels of D-glutamate [177], changes in platelet proteins [178], and expression changes in immune cells [179,180], are continually increasing, but the specific mechanisms behind them still require further investigation.

As the association between metabolic abnormalities and the onset of AD is increasingly confirmed, blood-based metabolic biomarkers are receiving more attention [181]. Recent studies have shown that cystatin C and carboxypeptidase B2 have potential as blood biomarkers, with a diagnostic model based on logistic regression algorithms showing a high accuracy rate of 93.8% [182]. In addition, lipids [183-185], arginine, and pentanoylcarnitine [84] as blood metabolic biomarkers also show diagnostic potential.

The development of new biomarkers in CSF has been slower than anticipated due to challenges in sample collection, high costs, and analytical complexities. Besides traditional biomarkers such as Aβ and p-tau, new CSF proteins such as NFL, soluble triggering receptor expressed on myeloid cells 2, and YKL-40 have been identified as indicators of neuronal damage. In recent research, Gaetani et al [186] performed a quantitative analysis of multiple biomarkers in CSF and used ML models, including penalized logistic regression, to identify biomarkers indicative of neuroinflammation's role in AD, such as SIRT2, HGF, MMP-10, and CXCL5. In addition, the study by Horie et al [126] on MTBR-tau243 in CSF demonstrated that its association with tau tangles and cognitive impairment in AD exceeds that of traditional p-tau biomarkers [126]. This discovery provides new insights for updating the amyloid, tau, and neurodegeneration diagnostic framework for AD.

In recent years, a series of emerging biomarkers for dementia have been continually identified and validated. These biomarkers are at the initial stages of research. In the realm of digital biomarkers, Bayat et al [124] achieved an accuracy rate of 89% in predicting preclinical AD by analyzing natural driving GPS data and building a random forest model. Thompson et al [187] used ML to analyze the graphics and features during the digital clock drawing test, finding a potential correlation between lower scores and a higher presence of *APOE ε4* alleles. In ophthalmology, Cheung et al [188] discovered new biomarkers related to dementia risk through the diameters of retinal blood vessels using a deep learning model. Recent confirmations also show that macular thickness and volume obtained from optical coherence tomography measurements [189] and the thickness of the retinal nerve fiber layer [190] have potential as AD biomarkers. In addition, metabolites in urine [191] and EEG features [192] have also demonstrated new research outcomes with the aid of AI.

XSL•FO
RenderX

Overall, many of the newly discovered biomarkers are still in the initial stages of discovery and validation. For these biomarkers to be translated into clinical applications, they must undergo thorough validation in broader data sets and larger population cohorts. In addition, assessing their diagnostic efficacy and reliability through longitudinal studies is a prerequisite for their future integration into clinical practice, a process that may take considerable time. However, it is encouraging that with the assistance of AI, researchers have discovered more biomarkers, significantly aiding in the refinement of dementia's pathological mechanisms and the exploration of potential therapeutic avenues.

Application of Popular Algorithms in Imaging Biomarkers

On the basis of our research, biomarkers obtained through various imaging techniques, such as MRI, offer detailed information about brain structure and are among the most widely used biomarkers today. Incorporating imaging biomarkers in multimodal data fusion strategies often significantly enhances classification accuracy [115]. SVMs are not only the most commonly used classification algorithms to date but also the most prevalent method for processing imaging data. SVMs excel in handling high-dimensional neuroimaging data, leveraging their kernel trick to handle nonlinear data in high-dimensional spaces, which is crucial for capturing complex biomarkers [193].

However, we have observed an interesting phenomenon: despite the low frequency of the keyword "SVM" in 2023, the actual use of SVMs in research as classifiers and as part of ensemble learning architectures has not shown a significant downward trend, with only a slight decrease of 2 instances compared to 2022.

This discrepancy is linked to a reduction in studies focusing solely on SVMs as independent algorithms, shifting toward comparative studies of various ML models, multivariate classification research, and an increase in ensemble learning approaches; for instance, Zubrikhina et al [99] found that SVMs showed the best performance among various ML models when classifying MRI data. Similarly, Tan et al [194] demonstrated that an ensemble model comprising gradient boosting, logistic regression, and SVMs outperformed single classifiers in multiple performance aspects [194]. Shukla et al [195] achieved an accuracy rate of 96% in a ternary classification of individuals with AD versus individuals with MCI versus cognitively normal individuals using multimodal imaging data combined with gradient boosting and SVMs. These studies indicate that SVMs remain a key component in many research projects. However, the use of "SVM" as a keyword may have declined due to a tendency to highlight emerging or innovative methods, leading to a reduced frequency of "SVM" in keyword use.

In recent years, neural networks have gained unprecedented popularity in the imaging biomarker domain, surpassing SVMs, primarily due to their reduced dependency on manual feature engineering and their proficiency in automatically identifying and learning the most significant features within data. Specifically, convolutional neural networks (CNNs) have demonstrated exceptional performance in the realm of image processing. A notable instance is the work of Lee et al [120], who used a deep learning model (3D-DenseNet) to process fluorodeoxyglucose-PET and MRI images. This model's architecture includes multiple dense blocks and convolutional layers capable of autonomously extracting complex features from imaging data. The feedforward connections within each dense block aid in acquiring a rich feature representation. Using the discrepancy between actual age and estimated brain age (brain age gap), they conducted classification diagnostics for dementia. Furthermore, a CNN model introduced by Ahmed et al [196] achieved an accuracy rate of 94% in distinguishing between patients with AD and healthy individuals through the analysis of imaging data from the ADNI data set.

Moreover, the application of transfer learning has significantly reduced the need for extensive data and computational resources required for training new models, thereby allowing researchers to forgo the necessity of developing CNN models from scratch. Hence, the synergy of imaging biomarkers and neural networks holds considerable potential and prospects for future research, particularly in terms of processing complex imaging data with greater precision and efficiency.

CSF Proteomics Biomarkers

CSF biomarkers have been among the earliest studied markers in dementia research due to their direct link with the brain and spinal cord, serving as a vital source of biochemical information. Specifically, tau proteins and Aβ in the CSF are core markers for dementia diagnosis. The pursuit of new proteomic biomarkers in CSF has been a continual area of interest; for instance, Gogishvili et al [197] used a random forest classification model to analyze proteomics data from CSF, successfully identifying new biomarkers such as CLEC1B, TNFRSF4, and TGF-β-1. However, obtaining CSF samples requires an invasive procedure known as lumbar puncture, and the analysis is costly, which somewhat limits the feasibility of large-scale data collection.

In more recent studies, increasing evidence has shown that tau proteins and Aβ, as well as their derivative forms, such as p-tau and total tau [198,199], can be obtained through multiple pathways [200]. This not only adds dimensions and richness to the data but also allows more research institutions access to these biomarkers. For regions with limited resources or underdeveloped technology, this accessibility helps reduce the costs of diagnosis and monitoring and provides more strategies to construct diverse cohorts and data sets for a more comprehensive understanding of neuropathological diversity. This shift in accessibility might explain why CSF biomarkers were more prevalently used in early research than other types of biomarkers but are now gradually being surpassed by other types of markers.

The Use of Random Forests in Blood and Genetic Markers

Compared to CSF and imaging biomarkers, genetic and blood markers have garnered considerable attention from researchers in recent years due to their minimally invasive collection process and ease of acquisition. Our research indicates that random forests have become more popular than neural networks and

XSL•FO
RenderX

SVMs in the application of blood and genetic markers. By integrating multiple decision trees, random forests can effectively capture the complex nonlinear relationships in data and handle various types of data. They not only possess robust predictive capabilities but also prevent overfitting [201], offering an intuitive understanding of the most critical features (such as specific biomarkers) in model predictions [202].

In genome-wide association studies, random forests can capture complex epistatic interactions and select key genetic variations [203], which is invaluable for identifying potential biomarkers in genes and blood; for instance, Kelly et al [204] used various ML models and gene expression profiles in their study of blood-based biomarkers, finding that random forests performed best in AD diagnostic models with an accuracy rate of 81%, identifying 159 gene markers. Beltrán et al [205] compared 4 ML methods, noting that random forests could achieve competitive results with costly medical imaging techniques when applied to readily available measurements (such as cognitive scores, genetic risk, and plasma biomarkers), identifying *APOE* and plasma C-reactive protein as the most significant features. However, each of these prevalent ML methods has its shortcomings; for example, neural networks have issues with interpretability and training costs; SVMs are highly sensitive to parameter selection, where inappropriate use of the kernel function or regularization parameters can lead to poor model performance; and, by contrast, random forests require extensive experimentation to adjust the number of trees, depth, and other parameters.

The Relationship Between Other Algorithms and Biomarkers

The connection between other algorithms and biomarkers is not as prominent or popular as that between the aforementioned algorithms and biomarkers. However, LASSO is observed to be frequently used in genetic biomarkers due to its efficiency in selecting disease-related feature genes from high-dimensional data [206]. By contrast, linear discriminant analysis and principal component analysis are more often applied in imaging biomarkers for feature reduction in MRI and PET modalities and fusion analysis of multimodal data [207]. Gradient boosting seems to be more inclined toward imaging and genetic markers, the k-nearest neighbors algorithm leans more toward imaging and neurophysiological markers, and logistic regression is more favored for imaging markers. Currently, many ML models lack standard settings and guidelines, making a robust comparison of these experiments difficult. Moreover, the specific combinations of ML methods and biomarkers may be influenced by various factors, such as the accessibility of variables, cost-effectiveness, and the adaptability of the model to the application context (eg, clinical and research environments) [208]. The diversity and complexity of these factors mean that the same algorithm might show different effectiveness and applicability in different studies. Nonetheless, by conducting an in-depth analysis of numerous studies to explore the relationship between different ML models and biomarker research, valuable insights and references can be provided for the field.

The Progress of AI

From the transition of traditional ML algorithms to the widespread application of deep learning and neural networks, significant progress has been marked in the field of medical AI [209]. Notably, since 2018, neural networks have increasingly dominated the research of dementia biomarkers, showcasing the potential to become the leading algorithms. A similar trend has been observed in other medical disciplines, such as cardiology, which has broadly implemented neural networks and deep learning technologies since 2015 [210]. In gastric cancer research, Shichijo et al [211] first used CNNs in 2017 to evaluate their effectiveness in diagnosing *Helicobacter pylori* infection. By 2020, deep learning technologies were extensively applied in the study of biomarkers for gastric cancer [212].

In addition, oncology is at the forefront of using multiomics data for patient stratification and personalized treatment [213]. In the imaging of brain tumors, neural networks have significantly enhanced the accuracy of detection and classification; for instance, Özkaraca et al [214] successfully applied a dense CNN architecture, using MRI images to precisely classify different brain tumors, thus supporting the development of accurate treatment plans. In the research of genetic and hematologic biomarkers, neural networks have opened new pathways for the early detection and classification of various cancer types. The studies by Liu et al [215] and Almarzouki [216] have demonstrated the potential of neural networks, with their capability to identify biomarkers with high specificity and sensitivity, in processing complex biological data. Advanced algorithms are also extensively used in specific tumor subtyping, grading, and staging [209], as well as predicting treatment outcomes [217]. These are directions that dementia research needs to learn from and emulate. Currently, dementia research mainly revolves around diagnosing and classifying AD, and there is a need to strengthen the study of other subtypes and expand the scope and objectives of the research.

Commercialization of AI in Dementia

Although the potential for AI technology in the medical field is immense, the use of commercial AI products for dementia in clinical settings remains relatively limited. This is partly due to significant unresolved limitations associated with ML applications. Furthermore, obtaining regulatory approval for AI products in the tightly regulated health care sector is a major challenge and a prerequisite for their practical application.

However, in recent years, AI-based methods have made significant strides. Particularly following the release of the US Food and Drug Administration *Action Plan*, which classifies AI- and ML-based software as a medical device [218], the market has begun to see approvals for such products. While no AI devices specifically targeted at dementia have been approved yet, in the field of radiology, AI software such as SubtlePET and SubtleMR, which process imaging data, have been approved [105,219], indirectly advancing AI in the clinical diagnosis of dementia. In addition, Cheung et al [188] recently developed the Singapore I Vessel Analyzer deep learning system for automatic measurement of retinal vessel calibers in dementia, and the commercially available Idx software for diagnosing

XSL•FO
RenderX

retinal diseases through retinal examination [105] may further promote the application of ophthalmic biomarkers in the clinical diagnosis of dementia. The use of computer-aided diagnosis systems [220], which provide radiologists with areas of interest or risk assessments, also aids in better guiding clinicians in diagnosing dementia.

Development of Emerging Technologies

The application of emerging technologies has provided more opportunities for the use of AI and the discovery of new biomarkers. Specifically, advanced imaging techniques such as structural MRI, functional MRI, Pittsburgh compound B PET, and diffusion tensor imaging have significantly enhanced our ability to capture detailed information about the brain. Single molecule array technology, genome-wide association studies, and high-throughput sequencing techniques also play a crucial role in identifying blood and genetic biomarkers.

The introduction of a series of emerging digital and wearable devices has created new opportunities for the diagnosis and assessment of dementia. Zhang et al [221] recorded participants' trail making test hand-drawn strokes using an electromagnetic tablet and used random forest analysis to examine the drawing features, discovering that models combining paper-based and electronic trail making tests improved the accuracy of assessing cognitive impairments. Ghosh et al [222] used GPS tracking to measure ecological outdoor behavior and differentiated individuals with AD from healthy individuals using data-driven ML methods. In addition, gait data obtained from accelerometers and inertial measurement units, eye movement variations captured by eye trackers, voice data recorded by microphones, and a range of digital biomarkers captured by other devices show promising applications [208].

Furthermore, digital biomarkers are significant due to their close connection with daily life. The deployment of Internet of Things devices based on environmental sensors and monitoring software in homes can enable long-term monitoring and assessment of the behaviors of patients with dementia. Khodabandehloo and Riboni [223] used environmental sensors to monitor real-life activities to detect wandering behavior, combined with ML methods to detect cognitive decline. Lotfi et al [224] conducted studies using various standard home automation sensors to monitor activities and movements at home, using neural networks for data analysis to detect abnormal behaviors in dementia. All these studies provide new insights into the exploration of dementia biomarkers. As the Internet of Things, particularly wearable devices, becomes more prevalent, it will further drive the development and commercialization of software as a medical device.

Potential Biases in AI

Currently, the application of AI in dementia biomarkers faces multiple challenges. First, many studies rely heavily on specific data sets, particularly the ADNI, which, although they provide high-quality data, may limit the universality of the research due to overreliance. These data sets may not adequately represent all races, cultures, or geographic locations, potentially leading to algorithmic bias and affecting the broad applicability and clinical translation of the research findings.

Second, although many AI studies show promise in the preliminary stages, they often lack external validation on independent data sets during the validation phase. External validation is a crucial step to assess the model's generalizability, ensuring the effectiveness of research outcomes across different populations and clinical settings.

Moreover, although AI technologies such as deep learning excel in identifying and predicting dementia biomarkers, the *black box* nature of these models poses challenges in enhancing transparency and gaining trust from medical professionals. The limited interpretability of deep learning models restricts their practical application in clinical decision-making. Therefore, using techniques such as feature importance analysis and model visualization tools to help medical professionals and patients understand the logic behind AI decisions—explainable AI—is becoming an important research area, aiming to make the ML process more transparent and comprehensible [27].

Finally, the imbalance in the number of samples for each category label within training data sets also imposes additional constraints on the model's robustness and clinical applicability. Addressing these challenges requires broader sample collection; more rigorous model design and testing, such as using synthetic minority oversampling technique, adjusting class weights, or using specific loss functions to minimize the impact of minority categories; and other new methods to enhance model interpretability.

Ethical and Privacy Challenges

Data collection involves handling a significant amount of sensitive personal information. Without appropriate data protection measures, this could infringe on the participants' privacy rights. In addition, during the data storage process, it is essential to ensure data encryption and anonymization. It is also necessary to clearly define the ownership and use rights of the data, ensuring that only authorized personnel can access this information to prevent misuse. When sharing these data as data sets or in other forms publicly, patient consent is also required.

From an ethical standpoint, researchers have the responsibility to ensure that participants fully understand the significance of their involvement in the research, which should be based on voluntary principles and clear consent, especially for the collection of novel biomarkers such as digital biomarkers. This is particularly important for patients with AD who may not fully comprehend the research content. Ensuring the reasonableness and fairness of the consent process is essential. When errors occur in the diagnostic process using AI systems, a clear accountability mechanism should be in place. This involves how to handle medical errors caused by AI decisions and how to correct these errors.

Moreover, it is necessary to establish relevant policies and regulations to regulate the use of AI in the medical field, ensuring that it complies with medical ethical standards. In facing these challenges, researchers, technology developers, and policy makers need to work together to ensure that the development of AI technology proceeds under the premise of respecting patient rights and ensuring data security. By

XSL•FO
RenderX

establishing strict industry standards and ethical guidelines, the responsible use of AI in the medical field can be facilitated.

Application and Development Trend of Research

Strengthening Interdisciplinary Collaboration

As more new researchers join the field, the demand for external disciplinary knowledge continues to expand, making it especially important to establish a stable and continually active collaboration network, particularly an interdisciplinary one. In the future, exploring this model of interdisciplinary collaboration will become a focus of research.

Exploration of New Biomarkers

AI is widely used in the research of various dementia biomarkers, including imaging, CSF, and genetic markers, which remain the primary subjects of current research. However, there is a growing demand for economically efficient and noninvasive biomarkers [89,225]. Digital biomarkers and ophthalmic biomarkers hold significant research prospects for the future. Currently, Alzheimer's Research UK is studying combinations of various digital biomarkers and exploring the application of ML algorithms [208]. Once these new biomarkers are validated through neuroimaging and CSF tests [226], they may become a more cost-effective tool for the early detection of dementia, especially in resource-limited areas [227].

Validation of Newly Discovered Biomarkers

Researchers have used AI to successfully identify patterns and correlations that may have gone unnoticed previously, uncovering a cohort of new candidate biomarkers, particularly in imaging and genetic biomarkers. However, these novel biomarkers generally lack external validation. Hence, future research trends will focus on further validation and comparison of these biomarkers in larger data sets or cohorts to confirm their effectiveness.

Enhancing Interpretability

In the medical domain, the interpretability and transparency of algorithms are paramount. With the increasing popularity of neural networks, researchers must carefully select or design algorithms, focusing on interpretable ML and AI use. This emphasis will drive further innovation and development in the fields of neuroscience and medical research.

Increasing Research on Dementia Subtypes

AI has demonstrated the potential to differentiate between subtypes of dementia and identify new biomarkers for these conditions. However, research into subtypes other than AD remains scarce. Increasing the number of studies on these specific subtypes and expanding the diversity of research will help to enhance our comprehensive understanding of dementia.

Strengths and Limitations

To our knowledge, this study is the first to conduct a comprehensive analysis of AI in the field of dementia biomarkers using bibliometric methods. By integrating a strategy of multiple tools, we not only improved the accuracy of the analysis but also expanded the dimensions of the comprehensive analysis. We presented the current status and research hot spots of the field from multiple aspects (eg, keywords, countries or regions, authors, and funding) and, for the first time, used text mining methods to specifically quantify the scale and relationship of biomarker and algorithm use.

Nevertheless, this study has limitations. To mitigate potential human errors associated with manual database management, only 1 database was included, which may have resulted in the omission of a small number of relevant studies. Furthermore, our investigation was restricted to studies published in English, potentially overlooking high-quality research in other languages, and did not account for potential self-citation bias, although its impact on the trends displayed is likely minimal. Therefore, future studies should leverage programming languages such as Python or R to expand database inclusion and analyze research across multiple languages. In addition, the quality and potential biases of the included studies were not assessed, which might affect the depicted trends due to the influence of low-quality and biased research. Future efforts should include a detailed quality evaluation of the studies.

Conclusions

In this study, we conducted a comprehensive analysis of research on AI in dementia biomarkers, using the Web of Science Core Collection. The objective was to summarize the latest advances and trends in this field. Our findings reveal significant progress since 2018, with numerous biomarkers identified in the areas of imaging and genetics. The United States, China, and the United Kingdom have been instrumental in driving progress in this domain. In addition, we noted a trend of author turnover and a need for stronger collaboration, suggesting that governments and researchers should develop strategies to facilitate the involvement and initiation of research by new scholars. Furthermore, through content mining and analysis, this study explored the popularity trends of various algorithms and biomarkers and delved into the pivotal applications of AI technologies across different types of dementia biomarkers. It also summarized newly discovered biomarkers identified through AI. In conclusion, as new biomarkers continue to be developed, and new algorithmic architectures are constructed, the application of AI in the field of dementia biomarkers is emerging as a promising area of research.

Acknowledgments

This study was supported by the Key Project of University Laboratory Research Work of Zhejiang Province (ZD202202); the Zhejiang Province Traditional Chinese Medicine, Science, and Technology Project (2023ZF134); the Medical and Health Technology Plan of Zhejiang Province (2022507615); the Zhejiang Provincial Medical and Health Science and Technology Program Project (2022KY1052); and the First-Class Course of Zhejiang Province (2022-1133).

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request. The code required for software operation is also available from the corresponding author on reasonable request.

Authors' Contributions

WQ conceptualized the study, set the research methodology, performed data visualization, and edited the manuscript. XZ was responsible for the methodology and wrote the original draft. DH provided software and other resources and wrote and reviewed the original draft. BW, CD, YL, YC, and BW organized the data and set the research methodology. SC conceptualized the study, reviewed and edited the manuscript, acquired funding, managed the project, and performed formal analysis. YS provided software and other resources and set the research methodology. FL, LMMB, XL, JK, and GJ conducted formal analysis and supervision. XL, JY, and JL were responsible for data visualization. All authors reviewed the draft, provided comments and revisions, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search query.
[DOCX File , 12 KB - jmir_v26i1e57830_app1.docx]

Multimedia Appendix 2
Calculation formulas used for data statistics and analysis.
[DOCX File , 15 KB - jmir_v26i1e57830_app2.docx]

Multimedia Appendix 3
Keyword merging.
[DOCX File , 14 KB - jmir_v26i1e57830_app3.docx]

Multimedia Appendix 4
Use strategies for each tool.
[DOCX File , 16 KB - jmir_v26i1e57830_app4.docx]

Multimedia Appendix 5
Specific classification methods for various types of biomarkers.
[DOCX File , 12 KB - jmir_v26i1e57830_app5.docx]

Multimedia Appendix 6
Artificial intelligence classification of diseases and performance of highly cited local literature.
[DOCX File , 24 KB - jmir_v26i1e57830_app6.docx]

Multimedia Appendix 7
Top 10 research studies of artificial intelligence in dementia biomarkers, ranked by a standardized local citation index.
[DOCX File , 20 KB - jmir_v26i1e57830_app7.docx]

References

1. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020 Aug 08;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
2. -. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023 Apr;19(4):1598-1695. [doi: 10.1002/alz.13016] [Medline: 36918389]
3. -. 2021 Alzheimer's disease facts and figures. Alzheimers Dement 2021 Mar;17(3):327-406. [doi: 10.1002/alz.12328] [Medline: 33756057]
4. Mirkin S, Albensi BC. Should artificial intelligence be used in conjunction with neuroimaging in the diagnosis of Alzheimer's disease? Front Aging Neurosci 2023 Apr 18;15:1094233 [FREE Full text] [doi: 10.3389/fnagi.2023.1094233] [Medline: 37187577]

XSL•FO
RenderX

5. Qiu S, Miller MI, Joshi PS, Lee JC, Xue C, Ni Y, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun 2022 Jun 20;13(1):3404 [FREE Full text] [doi: 10.1038/s41467-022-31037-5] [Medline: 35725739]
6. AlSharabi K, Salamah YB, Aljalal M, Abdurraqeeb AM, Alturki FA. EEG-based clinical decision support system for Alzheimer's disorders diagnosis using EMD and deep learning techniques. Front Hum Neurosci 2023 Aug 31;17:1190203 [FREE Full text] [doi: 10.3389/fnhum.2023.1190203] [Medline: 37719771]
7. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001 Mar;69(3):89-95. [doi: 10.1067/mcp.2001.113989] [Medline: 11240971]
8. Jack CRJ, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimers Dement 2018 Apr;14(4):535-562 [FREE Full text] [doi: 10.1016/j.jalz.2018.02.018] [Medline: 29653606]
9. Simrén J, Leuzy A, Karikari TK, Hye A, Benedet AL, Lantero-Rodriguez J, AddNeuroMed consortium, et al. The diagnostic and prognostic capabilities of plasma biomarkers in Alzheimer's disease. Alzheimers Dement 2021 Jul;17(7):1145-1156 [FREE Full text] [doi: 10.1002/alz.12283] [Medline: 33491853]
10. Kapasi A, DeCarli C, Schneider JA. Impact of multiple pathologies on the threshold for clinically overt dementia. Acta Neuropathol 2017 Aug;134(2):171-186 [FREE Full text] [doi: 10.1007/s00401-017-1717-7] [Medline: 28488154]
11. Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Front Aging Neurosci 2019 Aug 20;11:220 [FREE Full text] [doi: 10.3389/fnagi.2019.00220] [Medline: 31481890]
12. Winchester LM, Harshfield EL, Shi L, Badhwar A, Khleifat AA, Clarke N, et al. Artificial intelligence for biomarker discovery in Alzheimer's disease and dementia. Alzheimers Dement 2023 Dec;19(12):5860-5871. [doi: 10.1002/alz.13390] [Medline: 37654029]
13. Chang CH, Lin CH, Lane HY. Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease. Int J Mol Sci 2021 Mar 09;22(5):2761 [FREE Full text] [doi: 10.3390/ijms22052761] [Medline: 33803217]
14. Zhang C, Lu Y. Study on artificial intelligence: the state of the art and future prospects. J Ind Inf Integr 2021 Sep;23:100224. [doi: 10.1016/j.jii.2021.100224]
15. Gutierrez G. Artificial intelligence in the intensive care unit. Crit Care 2020 Mar 24;24(1):101 [FREE Full text] [doi: 10.1186/s13054-020-2785-y] [Medline: 32204716]
16. Kameyama M, Umeda-Kameyama Y. Applications of artificial intelligence in dementia. Geriatr Gerontol Int 2024 Mar;24 Suppl 1:25-30. [doi: 10.1111/ggi.14709] [Medline: 37916614]
17. Aschenbrenner AJ, Gordon BA, Benzinger TL, Morris JC, Hassenstab JJ. Influence of tau PET, amyloid PET, and hippocampal volume on cognition in Alzheimer disease. Neurology 2018 Aug 28;91(9):e859-e866 [FREE Full text] [doi: 10.1212/WNL.0000000000006075] [Medline: 30068637]
18. Koychev I, Gunn RN, Firouzian A, Lawson J, Zamboni G, Ridha B, et al. PET Tau and Amyloid-β burden in mild Alzheimer's disease: divergent relationship with age, cognition, and cerebrospinal fluid biomarkers. J Alzheimers Dis 2017;60(1):283-293 [FREE Full text] [doi: 10.3233/JAD-170129] [Medline: 28800330]
19. Li CC, Wang ZY, Wang LJ, Zhang CY. Biosensors for epigenetic biomarkers detection: a review. Biosens Bioelectron 2019 Nov 01;144:111695. [doi: 10.1016/j.bios.2019.111695] [Medline: 31526982]
20. Nordengen K, Kirsebom BE, Henjum K, Selnes P, Gísladóttir B, Wettergreen M, et al. Glial activation and inflammation along the Alzheimer's disease continuum. J Neuroinflammation 2019 Feb 21;16(1):46 [FREE Full text] [doi: 10.1186/s12974-019-1399-2] [Medline: 30791945]
21. Gordon BA, Blazey TM, Su Y, Hari-Raj A, Dincer A, Flores S, et al. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. Lancet Neurol 2018 Mar;17(3):241-250 [FREE Full text] [doi: 10.1016/S1474-4422(18)30028-0] [Medline: 29397305]
22. Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, et al. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. Neurology 2010 Jan 19;74(3):201-209 [FREE Full text] [doi: 10.1212/WNL.0b013e3181cb3e25] [Medline: 20042704]
23. Yang X, Qu H. Bibliometric review on biomarkers for Alzheimer's disease between 2000 and 2023. Medicine (Baltimore) 2023 Sep 08;102(36):e34982 [FREE Full text] [doi: 10.1097/MD.0000000000034982] [Medline: 37682187]
24. Noda K, Lim Y, Sengoku S, Kodama K. Global biomarker trends in Alzheimer's research: a bibliometric analysis. Drug Discov Today 2023 Aug;28(8):103677. [doi: 10.1016/j.drudis.2023.103677] [Medline: 37390962]
25. Wu CC, Su CH, Islam MM, Liao MH. Artificial intelligence in dementia: a bibliometric study. Diagnostics (Basel) 2023 Jun 19;13(12):2109 [FREE Full text] [doi: 10.3390/diagnostics13122109] [Medline: 37371004]
26. Aberathne I, Kulasiri D, Samarasinghe S. Detection of Alzheimer's disease onset using MRI and PET neuroimaging: longitudinal data analysis and machine learning. Neural Regen Res 2023 Oct;18(10):2134-2140 [FREE Full text] [doi: 10.4103/1673-5374.367840] [Medline: 37056120]
27. Blanco K, Salcidua S, Orellana P, Sauma-Pérez T, León T, Steinmetz LC, et al. Systematic review: fluid biomarkers and machine learning methods to improve the diagnosis from mild cognitive impairment to Alzheimer's disease. Alzheimers Res Ther 2023 Oct 14;15(1):176 [FREE Full text] [doi: 10.1186/s13195-023-01304-8] [Medline: 37838690]

XSL•FO
RenderX

28. Falahati F, Westman E, Simmons A. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimers Dis 2014;41(3):685-708. [doi: 10.3233/JAD-131928] [Medline: 24718104]
29. Li R, Wang X, Lawler K, Garg S, Bai Q, Alty J. Applications of artificial intelligence to aid early detection of dementia: a scoping review on current capabilities and future directions. J Biomed Inform 2022 Mar;127:104030 [FREE Full text] [doi: 10.1016/j.jbi.2022.104030] [Medline: 35183766]
30. Tzimourta KD, Christou V, Tzallas AT, Giannakeas N, Astrakas LG, Angelidis P, et al. Machine learning algorithms and statistical approaches for Alzheimer's disease analysis based on resting-state EEG recordings: a systematic review. Int J Neural Syst 2021 May;31(5):2130002. [doi: 10.1142/S0129065721300023] [Medline: 33588710]
31. Ninkov A, Frank JR, Maggio LA. Bibliometrics: methods for studying academic publishing. Perspect Med Educ 2022 Jun;11(3):173-176 [FREE Full text] [doi: 10.1007/s40037-021-00695-4] [Medline: 34914027]
32. Bordons M, Zulueta MA. [Evaluation of the scientific activity through bibliometric indices]. Rev Esp Cardiol 1999 Oct;52(10):790-800 [FREE Full text] [doi: 10.1016/s0300-8932(99)75008-6] [Medline: 10563155]
33. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res 2021 Sep;133:285-296. [doi: 10.1016/j.jbusres.2021.04.070]
34. Wu CC, Huang CW, Wang YC, Islam MM, Kung WM, Weng YC, et al. mHealth research for weight loss, physical activity, and sedentary behavior: bibliometric analysis. J Med Internet Res 2022 Jun 08;24(6):e35747 [FREE Full text] [doi: 10.2196/35747] [Medline: 35675126]
35. He D, Cao S, Le Y, Wang M, Chen Y, Qian B. Virtual reality technology in cognitive rehabilitation application: bibliometric analysis. JMIR Serious Games 2022 Oct 19;10(4):e38315 [FREE Full text] [doi: 10.2196/38315] [Medline: 36260388]
36. Aijing L, Shuangcheng Y, Lu M. Medical Literature Information Retrieval. Third Edition. Beijing, China: People's Medical Publishing House; 2005.
37. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. J Med Internet Res 2020 Jul 29;22(7):e18228 [FREE Full text] [doi: 10.2196/18228] [Medline: 32723713]
38. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
39. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37-46 [FREE Full text] [doi: 10.1177/001316446002000104]
40. Nam S, Kim D, Jung W, Zhu Y. Understanding the research landscape of deep learning in biomedical science: scientometric analysis. J Med Internet Res 2022 Apr 22;24(4):e28114 [FREE Full text] [doi: 10.2196/28114] [Medline: 35451980]
41. Osinska V, Klimas R. Mapping science: tools for bibliometric and altmetric studies. Inf Res 2021 Dec;26(4). [doi: 10.47989/irpaper909]
42. Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci Technol 2005 Dec 14;57(3):359-377. [doi: 10.1002/asi.20317]
43. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
44. Aria M, Cuccurullo C. bibliometrix : an R-tool for comprehensive science mapping analysis. J Informetr 2017 Nov;11(4):959-975. [doi: 10.1016/j.joi.2017.08.007]
45. Yin S, Zhu F, Li Z, Che D, Li L, Zhang L, et al. Research hotspots and trends in music therapy intervention for patients with dementia: a bibliometrics and visual analysis of papers published from 2010 to 2021. Front Psychiatry 2022 Apr 28;13:860758 [FREE Full text] [doi: 10.3389/fpsyt.2022.860758] [Medline: 35573325]
46. Joinpoint trend analysis software. National Cancer Institute Division of Cancer Control and Population Sciences. URL: https://surveillance.cancer.gov/joinpoint/ [accessed 2024-04-05]
47. Niazi MA, Vasilakos A, Temkin A. Review of “Exploratory social network analysis with Pajek” by Wouter de Nooy, Andrej Mrvar and Vladimir Batageli. Complex Adapt Syst Model 2019 Jun 17;7:1. [doi: 10.1186/s40294-019-0062-1]
48. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. Science mapping software tools: review, analysis, and cooperative study among tools. J Am Soc Inf Sci Technol 2011 May 02;62(7):1382-1402. [doi: 10.1002/asi.21525]
49. Silva-Spínola A, Baldeiras I, Arrais JP, Santana I. The road to personalized medicine in Alzheimer's disease: the use of artificial intelligence. Biomedicines 2022 Jan 29;10(2):315 [FREE Full text] [doi: 10.3390/biomedicines10020315] [Medline: 35203524]
50. Golub GH, van Loan CF. An analysis of the total least squares problem. SIAM J Numer Anal 1980;17(6):883-893. [doi: 10.1137/0717073]
51. Yao J, Cao S, Le Y, He D, Chen Y, Huang C, et al. Mapping knowledge landscapes and emerging trends of non-contact vital signs monitoring: a bibliometric and visualization analysis from 2002 to 2023. Alex Eng J 2024 Feb;88:197-209. [doi: 10.1016/j.aej.2024.01.015]
52. Tang R, Zhang S, Ding C, Zhu M, Gao Y. Artificial intelligence in intensive care medicine: bibliometric analysis. J Med Internet Res 2022 Nov 30;24(11):e42185 [FREE Full text] [doi: 10.2196/42185] [Medline: 36449345]
53. Garrett S. An Introduction to the Mathematics of Finance: A Deterministic Approach. Amsterdam, The Netherlands: Elsevier Science; 2013.

XSL•FO
RenderX

54. Qiu H, Cao S, Xu R. Cancer incidence, mortality, and burden in China: a time-trend analysis and comparison with the United States and United Kingdom based on the global epidemiological data released in 2020. Cancer Commun (Lond) 2021 Oct;41(10):1037-1048 [FREE Full text] [doi: 10.1002/cac2.12197] [Medline: 34288593]
55. Shah FA, Jawaid SA. The h-index: an indicator of research and publication output. Pak J Med Sci 2023;39(2):315-316 [FREE Full text] [doi: 10.12669/pjms.39.2.7398] [Medline: 36950422]
56. Ali M. Understanding the 'g-index' and the 'e-index'. Semin Ophthalmol 2021;36(4):139. [doi: 10.1002/9781119597797.index]
57. Zhang CT. The e-index, complementing the h-index for excess citations. PLoS One 2009 May 5;4(5):e5429 [FREE Full text] [doi: 10.1371/journal.pone.0005429] [Medline: 19415119]
58. Choudhri AF, Siddiqui A, Khan NR, Cohen HL. Understanding bibliometric parameters and analysis. Radiographics 2015;35(3):736-746. [doi: 10.1148/rg.2015140036] [Medline: 25969932]
59. Gordon A. Transient and continuant authors in a research field: the case of terrorism. Scientometrics 2007 Jun 17;72(2):213-224. [doi: 10.1007/s11192-007-1714-z]
60. Price DJ. Networks of scientific papers. Science 1965 Jul 30;149(3683):510-515. [doi: 10.1126/science.149.3683.510] [Medline: 14325149]
61. Weinstock M. Bradford's Law. Nature 1971 Oct 08;233(5319):434. [doi: 10.1038/233434a0] [Medline: 16063422]
62. Venable GT, Shepherd BA, Loftis CM, McClatchy SG, Roberts ML, Fillinger ME, et al. Bradford’s law: identification of the core journals for neurosurgery and its subspecialties. J Neurosurg 2016 Feb;124(2):569-579. [doi: 10.3171/2015.3.jns15149]
63. Roldan-Valadez E, Salazar-Ruiz SY, Ibarra-Contreras R, Rios C. Current concepts on bibliometrics: a brief review about impact factor, Eigenfactor score, CiteScore, SCImago Journal Rank, Source-Normalised Impact per Paper, H-index, and alternative metrics. Ir J Med Sci 2019 Aug 3;188(3):939-951. [doi: 10.1007/s11845-018-1936-5] [Medline: 30511320]
64. Sources. Scopus Preview. URL: https://www.scopus.com/sources.uri?RN_AG_Sourced_400000654 [accessed 2023-12-10]
65. Scimago journal and country rank. Scimago Lab. URL: https://www.scimagojr.com/ [accessed 2023-12-12]
66. Garfield E. How can impact factors be improved? BMJ 1996 Aug 17;313(7054):411-413 [FREE Full text] [doi: 10.1136/bmj.313.7054.411] [Medline: 8761234]
67. Sign in to continue with Journal Citation Reports. Clarivate. URL: https://jcr.clarivate.com/jcr/browse-journals[accessed [accessed 2024-07-17]
68. Garfield E. The history and meaning of the journal impact factor. JAMA 2006 Jan 04;295(1):90-93. [doi: 10.1001/jama.295.1.90] [Medline: 16391221]
69. List of countries by GDP (nominal). Wikipedia: The Free Encyclopedia. URL: https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal) [accessed 2023-12-12]
70. Global Dementia Observatory (GDO). World Health Organization. URL: https://who.int/data/gho/data/themes/global-dementia-observatory-gdo [accessed 2024-01-20]
71. Global health estimates: leading causes of death. World Health Organization. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death [accessed 2024-01-20]
72. Bornmann L, Wohlrabe K. Normalisation of citation impact in economics. Scientometrics 2019 Jun 14;120(2):841-884. [doi: 10.1007/s11192-019-03140-w]
73. Rasmussen M, Karypis G. gCLUTO -- an interactive clustering, visualization, and analysis system. University of Minnesota. 2004 May 25. URL: https://conservancy.umn.edu/items/5deba1b8-6c14-4dee-9f82-da3b2210e7ed [accessed 2024-04-12]
74. Bai J, Li W, Huang YM, Guo Y. Bibliometric study of research and development for neglected diseases in the BRICS. Infect Dis Poverty 2016 Sep 06;5(1):89 [FREE Full text] [doi: 10.1186/s40249-016-0182-1] [Medline: 27595987]
75. Shen L, Wang S, Dai W, Zhang Z. Detecting the interdisciplinary nature and topic hotspots of robotics in surgery: social network analysis and bibliometric study. J Med Internet Res 2019 Mar 26;21(3):e12625 [FREE Full text] [doi: 10.2196/12625] [Medline: 30912752]
76. Wang J, Liang Y, Cao S, Cai P, Fan Y. Application of artificial intelligence in geriatric care: bibliometric analysis. J Med Internet Res 2023 Jun 23;25:e46014 [FREE Full text] [doi: 10.2196/46014] [Medline: 37351923]
77. Zhang D, Wang Y, Zhou L, Yuan H, Shen D, Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 2011 Apr 01;55(3):856-867 [FREE Full text] [doi: 10.1016/j.neuroimage.2011.01.008] [Medline: 21236349]
78. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Alzheimer's Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 2015 Jan 01;104:398-412 [FREE Full text] [doi: 10.1016/j.neuroimage.2014.10.002] [Medline: 25312773]
79. Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 2012 Jan 16;59(2):895-907 [FREE Full text] [doi: 10.1016/j.neuroimage.2011.09.069] [Medline: 21992749]
80. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. Neuroimage Clin 2013;2:735-745 [FREE Full text] [doi: 10.1016/j.nicl.2013.05.004] [Medline: 24179825]

XSL•FO
RenderX

81. Lee G, Nho K, Kang B, Sohn KA, Kim D, for Alzheimer's Disease Neuroimaging Initiative. Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep 2019 Feb 13;9(1):1952 [FREE Full text] [doi: 10.1038/s41598-018-37769-z] [Medline: 30760848]
82. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Alzheimer's Disease Neuroimaging Initiative. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage 2013 Jan 15;65:167-175 [FREE Full text] [doi: 10.1016/j.neuroimage.2012.09.065] [Medline: 23041336]
83. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng 2015 Apr;62(4):1132-1140 [FREE Full text] [doi: 10.1109/TBME.2014.2372011] [Medline: 25423647]
84. Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. Neuroimage 2009 Oct 15;48(1):138-149 [FREE Full text] [doi: 10.1016/j.neuroimage.2009.05.056] [Medline: 19481161]
85. Sørensen L, Igel C, Liv Hansen N, Osler M, Lauritzen M, Rostrup E, et al. Early detection of Alzheimer's disease using MRI hippocampal texture. Hum Brain Mapp 2016 Mar 21;37(3):1148-1161 [FREE Full text] [doi: 10.1002/hbm.23091] [Medline: 26686837]
86. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. Recent publications from the Alzheimer's disease neuroimaging initiative: reviewing progress toward improved AD clinical trials. Alzheimers Dement 2017 Apr;13(4):e1-85 [FREE Full text] [doi: 10.1016/j.jalz.2016.11.007] [Medline: 28342697]
87. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 2007 Sep;19(9):1498-1507. [doi: 10.1162/jocn.2007.19.9.1498] [Medline: 17714011]
88. Ellis KA, Bush AI, Darby D, de Fazio D, Foster J, Hudson P, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 2009 Aug;21(4):672-687. [doi: 10.1017/S1041610209009405] [Medline: 19470201]
89. Jack CRJ, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 2010 Jan;9(1):119-128 [FREE Full text] [doi: 10.1016/S1474-4422(09)70299-6] [Medline: 20083042]
90. Cavedoni S, Chirico A, Pedroli E, Cipresso P, Riva G. Digital biomarkers for the early detection of mild cognitive impairment: artificial intelligence meets virtual reality. Front Hum Neurosci 2020 Jul 24;14:245 [FREE Full text] [doi: 10.3389/fnhum.2020.00245] [Medline: 32848660]
91. Teipel SJ, Sabri O, Grothe M, Barthel H, Prvulovic D, Buerger K, et al. Perspectives for multimodal neurochemical and imaging biomarkers in Alzheimer's disease. J Alzheimers Dis 2013;33 Suppl 1:S329-S347. [doi: 10.3233/JAD-2012-129030] [Medline: 22735677]
92. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Cedarbaum J, et al. Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. Alzheimers Dement 2015 Jul;11(7):865-884 [FREE Full text] [doi: 10.1016/j.jalz.2015.04.005] [Medline: 26194320]
93. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. The Alzheimer's Disease Neuroimaging Initiative 3: continued innovation for clinical trial improvement. Alzheimers Dement 2017 May;13(5):561-571 [FREE Full text] [doi: 10.1016/j.jalz.2016.10.006] [Medline: 27931796]
94. Xu M, Zhang T, Li Z, Liu M, Zhang D. Towards evaluating the robustness of deep diagnostic models by adversarial attack. Med Image Anal 2021 Apr;69:101977. [doi: 10.1016/j.media.2021.101977] [Medline: 33550005]
95. Khojaste-Sarakhsi M, Haghighi SS, Ghomi SM, Marchiori E. Deep learning for Alzheimer's disease diagnosis: a survey. Artif Intell Med 2022 Aug;130:102332. [doi: 10.1016/j.artmed.2022.102332] [Medline: 35809971]
96. Gautam R, Sharma M. Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis. J Med Syst 2020 Jan 04;44(2):49. [doi: 10.1007/s10916-019-1519-7] [Medline: 31902041]
97. Segato A, Marzullo A, Calimeri F, de Momi E. Artificial intelligence for brain diseases: a systematic review. APL Bioeng 2020 Dec;4(4):041503 [FREE Full text] [doi: 10.1063/5.0011697] [Medline: 33094213]
98. Ding Z, Lee TL, Chan AS. Digital cognitive biomarker for mild cognitive impairments and dementia: a systematic review. J Clin Med 2022 Jul 19;11(14):4191 [FREE Full text] [doi: 10.3390/jcm11144191] [Medline: 35887956]
99. Zubrikhina MO, Abramova OV, Yarkin VE, Ushakov VL, Ochneva AG, Bernstein AV, et al. Machine learning approaches to mild cognitive impairment detection based on structural MRI data and morphometric features. Cognit Syst Res 2023 Mar;78:87-95. [doi: 10.1016/j.cogsys.2022.12.005]
100. Meltzer JP, Kerry CF. Strengthening international cooperation on artificial intelligence. Policy Commons. 2021 Feb 17. URL: https://policycommons.net/artifacts/4143996/strengthening-international-cooperation-on-artificial-intelligence/4951885/ [accessed 2023-12-20]
101. Gaebel W, Jessen F, Kanba S. Neurocognitive disorders in ICD-11: the debate and its outcome. World Psychiatry 2018 Jun;17(2):229-230 [FREE Full text] [doi: 10.1002/wps.20534] [Medline: 29856541]

XSL•FO
RenderX

102. Kusters R, Misevic D, Berry H, Cully A, Le Cunff Y, Dandoy L, et al. Interdisciplinary research in artificial intelligence: challenges and opportunities. Front Big Data 2020 Nov 23;3:577974 [FREE Full text] [doi: 10.3389/fdata.2020.577974] [Medline: 33693418]
103. Carmona-Lavado A, Gimenez-Fernandez EM, Vlaisavljevic V, Cabello-Medina C. Cross-industry innovation: a systematic literature review. Technovation 2023 Jun;124:102743. [doi: 10.1016/j.technovation.2023.102743]
104. Brown SA, Sparapani R, Osinski K, Zhang J, Blessing J, Cheng F, et al. Establishing an interdisciplinary research team for cardio-oncology artificial intelligence informatics precision and health equity. Am Heart J Plus 2022 Jan;13:100094 [FREE Full text] [doi: 10.1016/j.ahjo.2022.100094] [Medline: 35434676]
105. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020 Sep 11;3(1):118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
106. Malamateniou C, Knapp KM, Pergola M, Woznitza N, Hardy M. Artificial intelligence in radiography: where are we now and what does the future hold? Radiography (Lond) 2021 Oct;27 Suppl 1:S58-S62. [doi: 10.1016/j.radi.2021.07.015] [Medline: 34380589]
107. Liu DS, Abu-Shaban K, Halabi SS, Cook TS. Changes in radiology due to artificial intelligence that can attract medical students to the specialty. JMIR Med Educ 2023 Mar 20;9:e43415 [FREE Full text] [doi: 10.2196/43415] [Medline: 36939823]
108. Lazzeretti L, Capone F. How proximity matters in innovation networks dynamics along the cluster evolution. A study of the high technology applied to cultural goods. J Bus Res 2016 Dec;69(12):5855-5865. [doi: 10.1016/j.jbusres.2016.04.068]
109. Geldes C, Felzensztein C, Turkina E, Durand A. How does proximity affect interfirm marketing cooperation? a study of an agribusiness cluster. J Bus Res 2015 Feb;68(2):263-272. [doi: 10.1016/j.jbusres.2014.09.034]
110. Xu G, Dong C, Meng L. Research on the collaborative innovation relationship of artificial intelligence technology in Yangtze river delta of China: a complex network perspective. Sustainability 2022 Oct 27;14(21):14002. [doi: 10.3390/su142114002]
111. Cao Q, Tan CC, Xu W, Hu H, Cao XP, Dong Q, et al. The prevalence of dementia: a systematic review and meta-analysis. J Alzheimers Dis 2020;73(3):1157-1166. [doi: 10.3233/JAD-191092] [Medline: 31884487]
112. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am 2005 Nov;15(4):869-77, xi [FREE Full text] [doi: 10.1016/j.nic.2005.09.008] [Medline: 16443497]
113. Morley J, Floridi L. NHS AI lab: why we need to be ethically mindful about AI for healthcare. SSRN. 2019 Sep 6. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3445421 [accessed 2023-12-21]
114. Wu F, Lu C, Zhu M, Chen H, Zhu J, Yu K, et al. Towards a new generation of artificial intelligence in China. Nat Mach Intell 2020 Jun 16;2(6):312-316. [doi: 10.1038/s42256-020-0183-4]
115. Li Z, Jiang X, Wang Y, Kim Y. Applied machine learning in Alzheimer's disease research: omics, imaging, and clinical data. Emerg Top Life Sci 2021 Dec 21;5(6):765-777 [FREE Full text] [doi: 10.1042/ETLS20210249] [Medline: 34881778]
116. Zhu Q, Xu B, Huang J, Wang H, Xu R, Shao W, et al. Deep multi-modal discriminative and interpretability network for Alzheimer's disease diagnosis. IEEE Trans Med Imaging 2023 May;42(5):1472-1483. [doi: 10.1109/TMI.2022.3230750] [Medline: 37015464]
117. Srikrishna M, Ashton NJ, Moscoso A, Pereira JB, Heckemann RA, van Westen D, et al. CT-based volumetric measures obtained through deep learning: association with biomarkers of neurodegeneration. Alzheimers Dement 2024 Jan;20(1):629-640 [FREE Full text] [doi: 10.1002/alz.13445] [Medline: 37767905]
118. Cumplido-Mayoral I, García-Prat M, Operto G, Falcon C, Shekari M, Cacciaglia R, et al. Biological brain age prediction using machine learning on structural neuroimaging data: multi-cohort validation against biomarkers of Alzheimer's disease and neurodegeneration stratified by sex. Elife 2023 Apr 17;12:e81067 [FREE Full text] [doi: 10.7554/eLife.81067] [Medline: 37067031]
119. Kohannim O, Hua X, Hibar DP, Lee S, Chou YY, Toga AW, et al. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol Aging 2010 Aug;31(8):1429-1442 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2010.04.022] [Medline: 20541286]
120. Lee J, Burkett BJ, Min HK, Senjem ML, Lundt ES, Botha H, et al. Deep learning-based brain age prediction in normal aging and dementia. Nat Aging 2022 May;2(5):412-424 [FREE Full text] [doi: 10.1038/s43587-022-00219-7] [Medline: 37118071]
121. Ramanan VK, Gebre RK, Graff-Radford J, Hofrenning E, Algeciras-Schimnich A, Figdore DJ, et al. Genetic risk scores enhance the diagnostic value of plasma biomarkers of brain amyloidosis. Brain 2023 Nov 02;146(11):4508-4519 [FREE Full text] [doi: 10.1093/brain/awad196] [Medline: 37279785]
122. Luckett PH, Chen C, Gordon BA, Wisch J, Berman SB, Chhatwal JP, et al. Biomarker clustering in autosomal dominant Alzheimer's disease. Alzheimers Dement 2023 Jan;19(1):274-284 [FREE Full text] [doi: 10.1002/alz.12661] [Medline: 35362200]
123. Millar PR, Gordon BA, Luckett PH, Benzinger TL, Cruchaga C, Fagan AM, Dominantly Inherited Alzheimer Network, et al. Multimodal brain age estimates relate to Alzheimer disease biomarkers and cognition in early stages: a cross-sectional observational study. Elife 2023 Jan 06;12:e81869 [FREE Full text] [doi: 10.7554/eLife.81869] [Medline: 36607335]

XSL•FO
RenderX

124. Bayat S, Babulal GM, Schindler SE, Fagan AM, Morris JC, Mihailidis A, et al. GPS driving: a digital biomarker for preclinical Alzheimer disease. Alzheimers Res Ther 2021 Jun 14;13(1):115 [FREE Full text] [doi: 10.1186/s13195-021-00852-1] [Medline: 34127064]
125. Ferreiro AL, Choi J, Ryou J, Newcomer EP, Thompson R, Bollinger RM, et al. Gut microbiome composition may be an indicator of preclinical Alzheimer's disease. Sci Transl Med 2023 Jun 14;15(700):eabo2984 [FREE Full text] [doi: 10.1126/scitranslmed.abo2984] [Medline: 37315112]
126. Horie K, Salvadó G, Barthélemy NR, Janelidze S, Li Y, He Y, et al. CSF MTBR-tau243 is a specific biomarker of tau tangle pathology in Alzheimer's disease. Nat Med 2023 Aug;29(8):1954-1963 [FREE Full text] [doi: 10.1038/s41591-023-02443-z] [Medline: 37443334]
127. Bae J, Logan PE, Acri DJ, Bharthur A, Nho K, Saykin AJ, et al. A simulative deep learning model of SNP interactions on chromosome 19 for predicting Alzheimer's disease risk and rates of disease progression. Alzheimers Dement 2023 Dec;19(12):5690-5699. [doi: 10.1002/alz.13319] [Medline: 37409680]
128. O'Bryant SE, Xiao G, Barber R, Huebinger R, Wilhelmsen K, Edwards M, et al. A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI. PLoS One 2011;6(12):e28092 [FREE Full text] [doi: 10.1371/journal.pone.0028092] [Medline: 22163278]
129. Zhang F, Petersen M, Johnson L, Hall J, O'Bryant SE. Combination of serum and plasma biomarkers could improve prediction performance for Alzheimer's disease. Genes (Basel) 2022 Sep 27;13(10):1738 [FREE Full text] [doi: 10.3390/genes13101738] [Medline: 36292623]
130. Nguyen HD, Clément M, Planche V, Mansencal B, Coupé P. Deep grading for MRI-based differential diagnosis of Alzheimer's disease and frontotemporal dementia. Artif Intell Med 2023 Oct;144:102636 [FREE Full text] [doi: 10.1016/j.artmed.2023.102636] [Medline: 37783553]
131. Pérez-Millan A, Contador J, Juncà-Parella J, Bosch B, Borrell L, Tort-Merino A, et al. Classifying Alzheimer's disease and frontotemporal dementia using machine learning with cross-sectional and longitudinal magnetic resonance imaging data. Hum Brain Mapp 2023 Apr 15;44(6):2234-2244 [FREE Full text] [doi: 10.1002/hbm.26205] [Medline: 36661219]
132. Planche V, Mansencal B, Manjon JV, Tourdias T, Catheline G, Coupé P. Anatomical MRI staging of frontotemporal dementia variants. Alzheimers Dement 2023 Aug;19(8):3283-3294. [doi: 10.1002/alz.12975] [Medline: 36749884]
133. Canu E, Agosta F, Mandic-Stojmenovic G, Stojković T, Stefanova E, Inuggi A, et al. Multiparametric MRI to distinguish early onset Alzheimer's disease and behavioural variant of frontotemporal dementia. Neuroimage Clin 2017;15:428-438 [FREE Full text] [doi: 10.1016/j.nicl.2017.05.018] [Medline: 28616383]
134. Feis RA, Bouts MJ, Panman JL, Jiskoot LC, Dopper EG, Schouten TM, et al. Single-subject classification of presymptomatic frontotemporal dementia mutation carriers using multimodal MRI. Neuroimage Clin 2019;22:101718 [FREE Full text] [doi: 10.1016/j.nicl.2019.101718] [Medline: 30827922]
135. Teunissen CE, Elias N, Koel-Simmelink MJ, Durieux-Lu S, Malekzadeh A, Pham TV, et al. Novel diagnostic cerebrospinal fluid biomarkers for pathologic subtypes of frontotemporal dementia identified by proteomics. Alzheimers Dement (Amst) 2016;2:86-94 [FREE Full text] [doi: 10.1016/j.dadm.2015.12.004] [Medline: 27239539]
136. Bergström S, Öijerstedt L, Remnestål J, Olofsson J, Ullgren A, Seelaar H, et al. A panel of CSF proteins separates genetic frontotemporal dementia from presymptomatic mutation carriers: a GENFI study. Mol Neurodegener 2021 Nov 27;16(1):79 [FREE Full text] [doi: 10.1186/s13024-021-00499-4] [Medline: 34838088]
137. Magen I, Yacovzada NS, Warren JD, Heller C, Swift I, Bobeva Y, et al. microRNA-based predictor for diagnosis of frontotemporal dementia. Neuropathol Appl Neurobiol 2023 Aug;49(4):e12916. [doi: 10.1111/nan.12916] [Medline: 37317649]
138. Miltiadous A, Tzimourta KD, Giannakeas N, Tsipouras MG, Afrantou T, Ioannidis P, et al. Alzheimer's disease and frontotemporal dementia: a robust classification method of EEG signals and a comparison of validation methods. Diagnostics (Basel) 2021 Aug 09;11(8):1437 [FREE Full text] [doi: 10.3390/diagnostics11081437] [Medline: 34441371]
139. Swift IJ, Sogorb-Esteve A, Heller C, Synofzik M, Otto M, Graff C, et al. Fluid biomarkers in frontotemporal dementia: past, present and future. J Neurol Neurosurg Psychiatry 2021 Feb 13;92(2):204-215 [FREE Full text] [doi: 10.1136/jnnp-2020-323520] [Medline: 33188134]
140. Bonanni L, Thomas A, Onofrj M. Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. Neurology 2006 May 09;66(9):1455; author reply 1455. [doi: 10.1212/01.wnl.0000224698.67660.45] [Medline: 16682691]
141. Zhou S, Meng Q, Li L, Hai L, Wang Z, Li Z, et al. Identification of a qualitative signature for the diagnosis of dementia with Lewy bodies. Front Genet 2021;12:758103 [FREE Full text] [doi: 10.3389/fgene.2021.758103] [Medline: 34868234]
142. Villemagne VL, Chételat G. Neuroimaging biomarkers in Alzheimer's disease and other dementias. Ageing Res Rev 2016 Sep;30:4-16. [doi: 10.1016/j.arr.2016.01.004] [Medline: 26827785]
143. van Steenoven I, Koel-Simmelink MJ, Vergouw LJ, Tijms BM, Piersma SR, Pham TV, et al. Identification of novel cerebrospinal fluid biomarker candidates for dementia with Lewy bodies: a proteomic approach. Mol Neurodegener 2020 Jun 18;15(1):36 [FREE Full text] [doi: 10.1186/s13024-020-00388-2] [Medline: 32552841]

XSL•FO
RenderX

144. Ferreira D, Jelic V, Cavallin L, Oeksengaard AR, Snaedal J, Høgh P, et al. Electroencephalography is a good complement to currently established dementia biomarkers. Dement Geriatr Cogn Disord 2016;42(1-2):80-92. [doi: 10.1159/000448394] [Medline: 27595479]
145. Law ZK, Todd C, Mehraram R, Schumacher J, Baker MR, LeBeau FE, et al. The role of EEG in the diagnosis, prognosis and clinical correlations of dementia with Lewy bodies-a systematic review. Diagnostics (Basel) 2020 Aug 20;10(9):616 [FREE Full text] [doi: 10.3390/diagnostics10090616] [Medline: 32825520]
146. Suzuki Y, Suzuki M, Shigenobu K, Shinosaki K, Aoki Y, Kikuchi H, et al. A prospective multicenter validation study of a machine learning algorithm classifier on quantitative electroencephalogram for differentiating between dementia with Lewy bodies and Alzheimer's dementia. PLoS One 2022;17(3):e0265484 [FREE Full text] [doi: 10.1371/journal.pone.0265484] [Medline: 35358240]
147. Shigemizu D, Akiyama S, Asanomi Y, Boroevich KA, Sharma A, Tsunoda T, et al. A comparison of machine learning classifiers for dementia with Lewy bodies using miRNA expression data. BMC Med Genomics 2019 Oct 30;12(1):150 [FREE Full text] [doi: 10.1186/s12920-019-0607-3] [Medline: 31666070]
148. Soto M, Iranzo A, Lahoz S, Fernández M, Serradell M, Gaig C, et al. Serum MicroRNAs predict isolated rapid eye movement sleep behavior disorder and Lewy body diseases. Mov Disord 2022 Oct;37(10):2086-2098 [FREE Full text] [doi: 10.1002/mds.29171] [Medline: 35962561]
149. Chen J, Zhao X, Huang C, Lin J. Novel insights into molecular signatures and pathogenic cell populations shared by systemic lupus erythematosus and vascular dementia. Funct Integr Genomics 2023 Nov 16;23(4):337. [doi: 10.1007/s10142-023-01270-2] [Medline: 37971684]
150. Davoudi A, Dion C, Amini S, Tighe PJ, Price CC, Libon DJ, et al. Classifying non-dementia and Alzheimer's disease/vascular dementia patients using kinematic, time-based, and visuospatial parameters: the digital clock drawing test. J Alzheimers Dis 2021 Jun 29;82(1):47-57. [doi: 10.3233/jad-201129]
151. Liu Y, Chan DK, Thalamuthu A, Wen W, Jiang J, Paradise M, et al. Plasma lipidomic biomarker analysis reveals distinct lipid changes in vascular dementia. Comput Struct Biotechnol J 2020 Jun 9;18:1613-1624 [FREE Full text] [doi: 10.1016/j.csbj.2020.06.001] [Medline: 32670502]
152. Luo J, Chen L, Huang X, Xie J, Zou C, Pan M, et al. REPS1 as a potential biomarker in Alzheimer's disease and vascular dementia. Front Aging Neurosci 2022;14:894824 [FREE Full text] [doi: 10.3389/fnagi.2022.894824] [Medline: 35813961]
153. Zheng Y, Guo H, Zhang L, Wu J, Li Q, Lv F. Machine learning-based framework for differential diagnosis between vascular dementia and Alzheimer's disease using structural MRI features. Front Neurol 2019 Oct 25;10:1097 [FREE Full text] [doi: 10.3389/fneur.2019.01097] [Medline: 31708854]
154. Zheng W, Liu H, Li Z, Li K, Wang Y, Hu B, et al. Classification of Alzheimer's disease based on hippocampal multivariate morphometry statistics. CNS Neurosci Ther 2023 Sep;29(9):2457-2468 [FREE Full text] [doi: 10.1111/cns.14189] [Medline: 37002795]
155. Zhang Y, Li H, Zheng Q. A comprehensive characterization of hippocampal feature ensemble serves as individualized brain signature for Alzheimer's disease: deep learning analysis in 3238 participants worldwide. Eur Radiol 2023 Aug;33(8):5385-5397. [doi: 10.1007/s00330-023-09519-x] [Medline: 36892643]
156. Qiao J, Wang R, Liu H, Xu G, Wang Z. Brain disorder prediction with dynamic multivariate spatio-temporal features: application to Alzheimer's disease and autism spectrum disorder. Front Aging Neurosci 2022;14:912895 [FREE Full text] [doi: 10.3389/fnagi.2022.912895] [Medline: 36110425]
157. Mousa D, Zayed N, Yassine IA. Alzheimer disease stages identification based on correlation transfer function system using resting-state functional magnetic resonance imaging. PLoS One 2022 Apr 12;17(4):e0264710 [FREE Full text] [doi: 10.1371/journal.pone.0264710] [Medline: 35413053]
158. Zhao J, Ding X, Du Y, Wang X, Men G. Functional connectivity between white matter and gray matter based on fMRI for Alzheimer's disease classification. Brain Behav 2019 Oct;9(10):e01407 [FREE Full text] [doi: 10.1002/brb3.1407] [Medline: 31512413]
159. Zhu Y, Wu Y, Lv X, Wu J, Shen C, Tang Q, et al. The relationship between APOE genotype, CSF Tau and cognition across the Alzheimer's disease spectrum, moderation and mediation role of insula network connectivity. CNS Neurosci Ther 2024 Jan;30(1):e14401 [FREE Full text] [doi: 10.1111/cns.14401] [Medline: 37577852]
160. Sadiq A, Yahya N, Tang TB, Hashim H, Naseem I. Wavelet-based fractal analysis of rs-fMRI for classification of Alzheimer's disease. Sensors (Basel) 2022 Apr 19;22(9):3102 [FREE Full text] [doi: 10.3390/s22093102] [Medline: 35590793]
161. Xu X, Wang T, Li W, Li H, Xu B, Zhang M, et al. Morphological, structural, and functional networks highlight the role of the cortical-subcortical circuit in individuals with subjective cognitive decline. Front Aging Neurosci 2021 Jul 9;13:688113 [FREE Full text] [doi: 10.3389/fnagi.2021.688113] [Medline: 34305568]
162. Zhang Q, Wang Q, He C, Fan D, Zhu Y, Zang F, et al. Altered regional cerebral blood flow and brain function across the Alzheimer's disease spectrum: a potential biomarker. Front Aging Neurosci 2021 Feb 22;13:630382. [doi: 10.3389/fnagi.2021.630382] [Medline: 33692680]
163. Zhou Y, Song Z, Han X, Li H, Tang X. Prediction of Alzheimer's disease progression based on magnetic resonance imaging. ACS Chem Neurosci 2021 Nov 17;12(22):4209-4223. [doi: 10.1021/acschemneuro.1c00472] [Medline: 34723463]

XSL•FO
RenderX

164. He F, Li Y, Li C, Zhao J, Liu T, Fan L, et al. Changes in the connection network of whole-brain fiber tracts in patients with Alzheimer's disease have a tendency of lateralization. Neuroreport 2021 Oct 06;32(14):1175-1182. [doi: 10.1097/WNR.0000000000001708] [Medline: 34334777]
165. Zhuang X, Zhang G, Bao M, Jiang G, Wang H, Li S, et al. Development of a novel immune infiltration-related diagnostic model for Alzheimer's disease using bioinformatic strategies. Front Immunol 2023 Jul 20;14:1147501 [FREE Full text] [doi: 10.3389/fimmu.2023.1147501] [Medline: 37545529]
166. Zhou S, Ma G, Luo H, Shan S, Xiong J, Cheng G. Identification of 5 potential predictive biomarkers for Alzheimer's disease by integrating the unified test for molecular signatures and weighted gene coexpression network analysis. J Gerontol A Biol Sci Med Sci 2023 Mar 30;78(4):653-658. [doi: 10.1093/gerona/glac179] [Medline: 36048437]
167. Yang X, Guo W, Yang L, Li X, Zhang Z, Pang X, et al. The relationship between protein modified folding molecular network and Alzheimer's disease pathogenesis based on BAG2-HSC70-STUB1-MAPT expression patterns analysis. Front Aging Neurosci 2023;15:1090400 [FREE Full text] [doi: 10.3389/fnagi.2023.1090400] [Medline: 37251806]
168. Wei W, Wang ZY, Ma LN, Zhang TT, Cao Y, Li H. MicroRNAs in Alzheimer's disease: function and potential applications as diagnostic biomarkers. Front Mol Neurosci 2020;13:160 [FREE Full text] [doi: 10.3389/fnmol.2020.00160] [Medline: 32973449]
169. Tan MS, Cheah PL, Chin AV, Looi LM, Chang SW. Differential expression analysis of blood MicroRNA in identifying potential genes relevant to Alzheimer’s disease pathogenesis, using an integrated bioinformatics and machine learning approach. Appl Sci 2023 Feb 27;13(5):3071. [doi: 10.3390/app13053071]
170. Alamro H, Thafar MA, Albaradei S, Gojobori T, Essack M, Gao X. Exploiting machine learning models to identify novel Alzheimer's disease biomarkers and potential targets. Sci Rep 2023 Mar 27;13(1):4979 [FREE Full text] [doi: 10.1038/s41598-023-30904-5] [Medline: 36973386]
171. Zhang C, Rodriguez C, Spaulding J, Aw TY, Feng J. Age-dependent and tissue-related glutathione redox status in a mouse model of Alzheimer's disease. J Alzheimers Dis 2012;28(3):655-666 [FREE Full text] [doi: 10.3233/JAD-2011-111244] [Medline: 22045490]
172. Deng Y, Feng Y, Lv Z, He J, Chen X, Wang C, et al. Machine learning models identify ferroptosis-related genes as potential diagnostic biomarkers for Alzheimer's disease. Front Aging Neurosci 2022 Sep 28;14:994130 [FREE Full text] [doi: 10.3389/fnagi.2022.994130] [Medline: 36262887]
173. Wang X, Tian Y, Li C, Chen M. Exploring the key ferroptosis-related gene in the peripheral blood of patients with Alzheimer's disease and its clinical significance. Front Aging Neurosci 2022;14:970796 [FREE Full text] [doi: 10.3389/fnagi.2022.970796] [Medline: 36118694]
174. Yu H, Li M, Pan Q, Liu Y, Zhang Y, He T, et al. Integrated analyses of brain and platelet omics reveal their common altered and driven molecules in Alzheimer's disease. MedComm (2020) 2022 Dec;3(4):e180 [FREE Full text] [doi: 10.1002/mco2.180] [Medline: 36254251]
175. Yu S, Liu YP, Liu HL, Li J, Xiang Y, Liu YH, et al. Serum protein-based profiles as novel biomarkers for the diagnosis of Alzheimer's disease. Mol Neurobiol 2018 May;55(5):3999-4008. [doi: 10.1007/s12035-017-0609-0] [Medline: 28567666]
176. Lin H, Himali JJ, Satizabal CL, Beiser AS, Levy D, Benjamin EJ, et al. Identifying blood biomarkers for dementia using machine learning methods in the Framingham Heart Study. Cells 2022 Apr 30;11(9):1506 [FREE Full text] [doi: 10.3390/cells11091506] [Medline: 35563811]
177. Chang CH, Lin CH, Liu CY, Huang CS, Chen SJ, Lin WC, et al. Plasma d-glutamate levels for detecting mild cognitive impairment and Alzheimer's disease: machine learning approaches. J Psychopharmacol 2021 Mar;35(3):265-272. [doi: 10.1177/0269881120972331] [Medline: 33586518]
178. Yu H, Liu Y, He B, He T, Chen C, He J, et al. Platelet biomarkers for a descending cognitive function: a proteomic approach. Aging Cell 2021 May;20(5):e13358 [FREE Full text] [doi: 10.1111/acel.13358] [Medline: 33942972]
179. Gericke C, Kirabali T, Flury R, Mallone A, Rickenbach C, Kulic L, et al. Early β-amyloid accumulation in the brain is associated with peripheral T cell alterations. Alzheimers Dement 2023 Dec;19(12):5642-5662. [doi: 10.1002/alz.13136] [Medline: 37314431]
180. Wu CT, Chu CI, Wang FY, Yang HY, Tseng WS, Chang CR, et al. A change of PD-1/PD-L1 expression on peripheral T cell subsets correlates with the different stages of Alzheimer's disease. Cell Biosci 2022 Sep 30;12(1):162 [FREE Full text] [doi: 10.1186/s13578-022-00897-1] [Medline: 36180897]
181. Feng Y, Chen X, Zhang XD, Huang C. Metabolic pathway pairwise-based signature as a potential non-invasive diagnostic marker in Alzheimer's disease patients. Genes (Basel) 2023 Jun 17;14(6):1285 [FREE Full text] [doi: 10.3390/genes14061285] [Medline: 37372465]
182. Yang L, Xuan C, Yu C, Zheng P, Yan J. Diagnostic model of Alzheimer’s disease in the elderly based on protein and metabolic biomarkers. J Alzheimers Dis 2022 Feb 01;85(3):1163-1174. [doi: 10.3233/jad-215119]
183. Wang Q, Zang F, He C, Zhang Z, Xie C, Alzheimer’s Disease Neuroimaging Initiative. Dyslipidemia induced large-scale network connectivity abnormality facilitates cognitive decline in the Alzheimer's disease. J Transl Med 2022 Dec 06;20(1):567 [FREE Full text] [doi: 10.1186/s12967-022-03786-w] [Medline: 36474263]

XSL•FO
RenderX

184. Gurke R, Etyemez S, Prvulovic D, Thomas D, Fleck SC, Reif A, et al. A data science-based analysis points at distinct patterns of lipid mediator plasma concentrations in patients with dementia. Front Psychiatry 2019 Feb 11;10:41 [FREE Full text] [doi: 10.3389/fpsyt.2019.00041] [Medline: 30804821]
185. Varma VR, Oommen AM, Varma S, Casanova R, An Y, Andrews RM, et al. Brain and blood metabolite signatures of pathology and progression in Alzheimer disease: a targeted metabolomics study. PLoS Med 2018 Jan;15(1):e1002482 [FREE Full text] [doi: 10.1371/journal.pmed.1002482] [Medline: 29370177]
186. Gaetani L, Bellomo G, Parnetti L, Blennow K, Zetterberg H, Di Filippo M. Neuroinflammation and Alzheimer's disease: a machine learning approach to CSF proteomics. Cells 2021 Jul 29;10(8):1930 [FREE Full text] [doi: 10.3390/cells10081930] [Medline: 34440700]
187. Thompson LI, Cummings M, Emrani S, Libon DJ, Ang A, Karjadi C, et al. Digital clock drawing as an Alzheimer's disease susceptibility biomarker: associations with genetic risk score and APOE in older adults. J Prev Alzheimers Dis 2024;11(1):79-87 [FREE Full text] [doi: 10.14283/jpad.2023.48] [Medline: 38230720]
188. Cheung CY, Wong WL, Hilal S, Kan CN, Gyanwali B, Tham YC, et al. Deep-learning retinal vessel calibre measurements and risk of cognitive decline and dementia. Brain Commun 2022 Aug 17;4(4):fcac212 [FREE Full text] [doi: 10.1093/braincomms/fcac212] [Medline: 36043139]
189. Wang X, Jiao B, Liu H, Wang Y, Hao X, Zhu Y, et al. Machine learning based on optical coherence tomography images as a diagnostic tool for Alzheimer's disease. CNS Neurosci Ther 2022 Dec;28(12):2206-2217 [FREE Full text] [doi: 10.1111/cns.13963] [Medline: 36089740]
190. Lemmens S, van Craenendonck T, van Eijgen J, de Groef L, Bruffaerts R, de Jesus DA, et al. Combination of snapshot hyperspectral retinal imaging and optical coherence tomography to identify Alzheimer's disease patients. Alzheimers Res Ther 2020 Nov 10;12(1):144 [FREE Full text] [doi: 10.1186/s13195-020-00715-1] [Medline: 33172499]
191. Yilmaz A, Ugur Z, Bisgin H, Akyol S, Bahado-Singh R, Wilson G, et al. Targeted metabolic profiling of urine highlights a potential biomarker panel for the diagnosis of Alzheimer's disease and mild cognitive impairment: a pilot study. Metabolites 2020 Aug 31;10(9):357 [FREE Full text] [doi: 10.3390/metabo10090357] [Medline: 32878308]
192. Simfukwe C, Youn YC, Kim MJ, Paik J, Han SH. CNN for a regression machine learning algorithm for predicting cognitive impairment using qEEG. Neuropsychiatr Dis Treat 2023 Apr 12;19:851-863 [FREE Full text] [doi: 10.2147/NDT.S404528] [Medline: 37077704]
193. Gomes TA, Prudêncio RB, Soares C, Rossi AL, Carvalho A. Combining meta-learning and search techniques to select parameters for support vector machines. Neurocomputing 2012 Jan;75(1):3-13. [doi: 10.1016/j.neucom.2011.07.005]
194. Tan WY, Hargreaves C, Chen C, Hilal S. A machine learning approach for early diagnosis of cognitive impairment using population-based data. J Alzheimers Dis 2023;91(1):449-461 [FREE Full text] [doi: 10.3233/JAD-220776] [Medline: 36442196]
195. Shukla A, Tiwari R, Tiwari S. Alzheimer’s disease detection from fused PET and MRI modalities using an ensemble classifier. MAKE 2023 May 18;5(2):512-538. [doi: 10.3390/make5020031]
196. Ahmed S, Kim BC, Lee KH, Jung HY, Alzheimer’s Disease Neuroimaging Initiative. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. PLoS One 2020;15(12):e0242712 [FREE Full text] [doi: 10.1371/journal.pone.0242712] [Medline: 33290403]
197. Gogishvili D, Vromen EM, Koppes-den Hertog S, Lemstra AW, Pijnenburg YA, Visser PJ, et al. Discovery of novel CSF biomarkers to predict progression in dementia using machine learning. Sci Rep 2023 Apr 21;13(1):6531 [FREE Full text] [doi: 10.1038/s41598-023-33045-x] [Medline: 37085545]
198. Lee WJ, Brown JA, Kim HR, La Joie R, Cho H, Lyoo CH, et al. Regional Aβ-tau interactions promote onset and acceleration of Alzheimer's disease tau spreading. Neuron 2022 Jun 15;110(12):1932-43.e5 [FREE Full text] [doi: 10.1016/j.neuron.2022.03.034] [Medline: 35443153]
199. Mahaman YA, Embaye KS, Huang F, Li L, Zhu F, Wang JZ, et al. Biomarkers used in Alzheimer's disease diagnosis, treatment, and prevention. Ageing Res Rev 2022 Feb;74:101544. [doi: 10.1016/j.arr.2021.101544] [Medline: 34933129]
200. Chandra A, Dervenoulas G, Politis M, Alzheimer’s Disease Neuroimaging Initiative. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. J Neurol 2019 Jun;266(6):1293-1302 [FREE Full text] [doi: 10.1007/s00415-018-9016-3] [Medline: 30120563]
201. Byeon H. Can the random forests model improve the power to predict the intention of the elderly in a community to participate in a cognitive health promotion program? Iran J Public Health 2021 Feb;50(2):315-324 [FREE Full text] [doi: 10.18502/ijph.v50i2.5346] [Medline: 33747995]
202. Wu X, Peng C, Nelson PT, Cheng Q. Random forest-integrated analysis in AD and LATE brain transcriptome-wide data to identify disease-specific gene expression. PLoS One 2021 Sep 7;16(9):e0256648 [FREE Full text] [doi: 10.1371/journal.pone.0256648] [Medline: 34492068]
203. Bettencourt C, Skene N, Bandres-Ciga S, Anderson E, Winchester LM, Foote IF, Deep Dementia Phenotyping (DEMON) Network, et al. Artificial intelligence for dementia genetics and omics. Alzheimers Dement 2023 Dec;19(12):5905-5921. [doi: 10.1002/alz.13427] [Medline: 37606627]
204. Kelly J, Moyeed R, Carroll C, Luo S, Li X. Blood biomarker-based classification study for neurodegenerative diseases. Sci Rep 2023 Oct 11;13(1):17191 [FREE Full text] [doi: 10.1038/s41598-023-43956-4] [Medline: 37821485]

XSL•FO
RenderX

205. Beltrán JF, Wahba BM, Hose N, Shasha D, Kline RP, Alzheimer's Disease Neuroimaging Initiative. Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's Disease Neuroimaging (ADNI) database. PLoS One 2020 Jun 27;15(7):e0235663 [FREE Full text] [doi: 10.1371/journal.pone.0235663] [Medline: 32716914]
206. Zou C, Su L, Pan M, Chen L, Li H, Zou C, et al. Exploration of novel biomarkers in Alzheimer's disease based on four diagnostic models. Front Aging Neurosci 2023 Feb 16;15:1079433 [FREE Full text] [doi: 10.3389/fnagi.2023.1079433] [Medline: 36875704]
207. Lin W, Gao Q, Du M, Chen W, Tong T. Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. Comput Biol Med 2021 Jul;134:104478. [doi: 10.1016/j.compbiomed.2021.104478] [Medline: 34000523]
208. Lyall DM, Kormilitzin A, Lancaster C, Sousa J, Petermann-Rocha F, Buckley C, Deep Dementia Phenotyping (DEMON) Network, et al. Artificial intelligence for dementia-applied models and digital health. Alzheimers Dement 2023 Dec;19(12):5872-5884 [FREE Full text] [doi: 10.1002/alz.13391] [Medline: 37496259]
209. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun (Lond) 2020 Apr;40(4):154-166 [FREE Full text] [doi: 10.1002/cac2.12012] [Medline: 32277744]
210. Suero-Abreu GA, Hamid A, Akbilgic O, Brown SA. Trends in cardiology and oncology artificial intelligence publications. Am Heart J Plus 2022 May;17:100162 [FREE Full text] [doi: 10.1016/j.ahjo.2022.100162] [Medline: 38559882]
211. Shichijo S, Nomura S, Aoyama K, Nishikawa Y, Miura M, Shinagawa T, et al. Application of convolutional neural networks in the diagnosis of helicobacter pylori infection based on endoscopic images. EBioMedicine 2017 Nov;25:106-111 [FREE Full text] [doi: 10.1016/j.ebiom.2017.10.014] [Medline: 29056541]
212. Zhang G, Song J, Feng Z, Zhao W, Huang P, Liu L, et al. Artificial intelligence applicated in gastric cancer: a bibliometric and visual analysis via CiteSpace. Front Oncol 2022;12:1075974 [FREE Full text] [doi: 10.3389/fonc.2022.1075974] [Medline: 36686778]
213. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17 [FREE Full text] [doi: 10.1016/j.csbj.2014.11.005] [Medline: 25750696]
214. Özkaraca O, Bağrıaçık OI, Gürüler H, Khan F, Hussain J, Khan J, et al. Multiple brain tumor classification with dense CNN architecture using brain MRI images. Life (Basel) 2023 Jan 28;13(2):349 [FREE Full text] [doi: 10.3390/life13020349] [Medline: 36836705]
215. Liu B, Liu Y, Pan X, Li M, Yang S, Li SC. DNA methylation markers for pan-cancer prediction by deep learning. Genes (Basel) 2019 Oct 04;10(10):778 [FREE Full text] [doi: 10.3390/genes10100778] [Medline: 31590287]
216. Almarzouki HZ. Deep-learning-based cancer profiles classification using gene expression data profile. J Healthc Eng 2022 Jan 7;2022:4715998 [FREE Full text] [doi: 10.1155/2022/4715998] [Medline: 35035840]
217. Jiang Y, Wang H, Wu J, Chen C, Yuan Q, Huang W, et al. Corrigendum to "noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer": annals of oncology 2020; 31: 760-768. Ann Oncol 2021 Apr;32(4):578 [FREE Full text] [doi: 10.1016/j.annonc.2021.01.064] [Medline: 33546918]
218. Merkin A, Krishnamurthi R, Medvedev ON. Machine learning, artificial intelligence and the prediction of dementia. Curr Opin Psychiatry 2022 Mar 01;35(2):123-129. [doi: 10.1097/YCO.0000000000000768] [Medline: 34861656]
219. Pemberton HG, Zaki LA, Goodkin O, Das RK, Steketee RM, Barkhof F, et al. Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis-a systematic review. Neuroradiology 2021 Nov;63(11):1773-1789 [FREE Full text] [doi: 10.1007/s00234-021-02746-3] [Medline: 34476511]
220. Carnevale L, Lembo G. Innovative MRI techniques in neuroimaging approaches for cerebrovascular diseases and vascular cognitive impairment. Int J Mol Sci 2019 May 30;20(11):2656 [FREE Full text] [doi: 10.3390/ijms20112656] [Medline: 31151154]
221. Zhang W, Zheng X, Tang Z, Wang H, Li R, Xie Z, et al. Combination of paper and electronic trail making tests for automatic analysis of cognitive impairment: development and validation study. J Med Internet Res 2023 Jun 09;25:e42637 [FREE Full text] [doi: 10.2196/42637] [Medline: 37294606]
222. Ghosh A, Puthusseryppady V, Chan D, Mascolo C, Hornberger M. Machine learning detects altered spatial navigation features in outdoor behaviour of Alzheimer's disease patients. Sci Rep 2022 Feb 24;12(1):3160 [FREE Full text] [doi: 10.1038/s41598-022-06899-w] [Medline: 35210486]
223. Khodabandehloo E, Riboni D. Collaborative trajectory mining in smart-homes to support early diagnosis of cognitive decline. IEEE Trans Emerg Topics Comput 2021 Jul 1;9(3):1194-1205. [doi: 10.1109/tetc.2020.2975071]
224. Lotfi A, Langensiepen C, Mahmoud SM, Akhlaghinia MJ. Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour. J Ambient Intell Human Comput 2011 Jan 7;3(3):205-218. [doi: 10.1007/s12652-010-0043-x]
225. Petermann-Rocha F, Lyall DM, Quinn TJ, Ho FK, Pell JP, Celis-Morales C. Associations between physical frailty and dementia incidence: a prospective study from UK Biobank - authors' reply. Lancet Healthy Longev 2021 Feb;2(2):e68 [FREE Full text] [doi: 10.1016/S2666-7568(21)00008-8] [Medline: 36098159]

XSL•FO
RenderX

226. Korolev IO, Symonds LL, Bozoki AC, Alzheimer's Disease Neuroimaging Initiative. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. PLoS One 2016;11(2):e0138866 [FREE Full text] [doi: 10.1371/journal.pone.0138866] [Medline: 26901338]
227. Godfrey A, Brodie M, van Schooten KS, Nouredanesh M, Stuart S, Robinson L. Inertial wearables as pragmatic tools in dementia. Maturitas 2019 Sep;127:12-17. [doi: 10.1016/j.maturitas.2019.05.010] [Medline: 31351515]

Abbreviations

AD: Alzheimer disease
ADNI: Alzheimer's Disease Neuroimaging Initiative
AI: artificial intelligence
APOE: apolipoprotein E
Aβ: beta amyloid
CNN: convolutional neural network
CSF: cerebrospinal fluid
DLB: dementia with Lewy bodies
EEG: electroencephalography
FTD: frontotemporal dementia
LASSO: least absolute shrinkage and selection operator
miRNA: microRNA
ML: machine learning
MRI: magnetic resonance imaging
NFL: neurofilament light
PET: positron emission tomography
p-tau: phosphorylated tau
SVM: support vector machine
VaD: vascular dementia
WM: white matter

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 27.02.24; peer-reviewed by X Ouyang, Y Zhang, J Liang, K Kodama, A Azizan, S Sengoku; comments to author 22.03.24; revised version received 04.05.24; accepted 25.06.24; published 08.08.24.

Please cite as:
Qi W, Zhu X, He D, Wang B, Cao S, Dong C, Li Y, Chen Y, Wang B, Shi Y, Jiang G, Liu F, Boots LMM, Li J, Lou X, Yao J, Lu X, Kang J
Mapping Knowledge Landscapes and Emerging Trends in AI for Dementia Biomarkers: Bibliometric and Visualization Analysis
J Med Internet Res 2024;26:e57830
URL: https://www.jmir.org/2024/1/e57830
doi:10.2196/57830
PMID:

©Wenhao Qi, Xiaohong Zhu, Danni He, Bin Wang, Shihua Cao, Chaoqun Dong, Yunhua Li, Yanfei Chen, Bingsheng Wang, Yankai Shi, Guowei Jiang, Fang Liu, Lizzy M M Boots, Jiaqi Li, Xiajing Lou, Jiani Yao, Xiaodong Lu, Junling Kang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating Algorithmic Bias in 30-Day Hospital Readmission Models: Retrospective Analysis

H Echo Wang[1], DrPH; Jonathan P Weiner[1,2], DrPH; Suchi Saria[3], PhD; Hadi Kharrazi[1,2], MD, PhD

[1]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[2]Johns Hopkins Center for Population Health Information Technology, Baltimore, MD, United States

[3]Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Hadi Kharrazi, MD, PhD
Bloomberg School of Public Health
Johns Hopkins University
624 N Broadway, Hampton House
Baltimore, MD
United States
Phone: 1 443 287 8264
Email: kharrazi@jhu.edu

Abstract

Background: The adoption of predictive algorithms in health care comes with the potential for algorithmic bias, which could exacerbate existing disparities. Fairness metrics have been proposed to measure algorithmic bias, but their application to real-world tasks is limited.

Objective: This study aims to evaluate the algorithmic bias associated with the application of common 30-day hospital readmission models and assess the usefulness and interpretability of selected fairness metrics.

Methods: We used 10.6 million adult inpatient discharges from Maryland and Florida from 2016 to 2019 in this retrospective study. Models predicting 30-day hospital readmissions were evaluated: LACE Index, modified HOSPITAL score, and modified Centers for Medicare & Medicaid Services (CMS) readmission measure, which were applied *as-is* (using existing coefficients) and *retrained* (recalibrated with 50% of the data). Predictive performances and bias measures were evaluated for all, between Black and White populations, and between low- and other-income groups. Bias measures included the parity of false negative rate (FNR), false positive rate (FPR), 0-1 loss, and generalized entropy index. Racial bias represented by FNR and FPR differences was stratified to explore shifts in algorithmic bias in different populations.

Results: The retrained CMS model demonstrated the best predictive performance (area under the curve: 0.74 in Maryland and 0.68-0.70 in Florida), and the modified HOSPITAL score demonstrated the best calibration (Brier score: 0.16-0.19 in Maryland and 0.19-0.21 in Florida). Calibration was better in White (compared to Black) populations and other-income (compared to low-income) groups, and the area under the curve was higher or similar in the Black (compared to White) populations. The retrained CMS and modified HOSPITAL score had the lowest racial and income bias in Maryland. In Florida, both of these models overall had the lowest income bias and the modified HOSPITAL score showed the lowest racial bias. In both states, the White and higher-income populations showed a higher FNR, while the Black and low-income populations resulted in a higher FPR and a higher 0-1 loss. When stratified by hospital and population composition, these models demonstrated heterogeneous algorithmic bias in different contexts and populations.

Conclusions: Caution must be taken when interpreting fairness measures' face value. A higher FNR or FPR could potentially reflect missed opportunities or wasted resources, but these measures could also reflect health care use patterns and gaps in care. Simply relying on the statistical notions of bias could obscure or underplay the causes of health disparity. The imperfect health data, analytic frameworks, and the underlying health systems must be carefully considered. Fairness measures can serve as a useful routine assessment to detect disparate model performances but are insufficient to inform mechanisms or policy changes. However, such an assessment is an important first step toward data-driven improvement to address existing health disparities.

(J Med Internet Res 2024;26:e47125) doi:10.2196/47125

XSL•FO
RenderX

KEYWORDS

algorithmic bias; model bias; predictive models; model fairness; health disparity; hospital readmission; retrospective analysis

Introduction

Background of Algorithmic Bias

Predictive algorithms and machine learning tools are increasingly integrated into clinical decision-making and population health management. However, with the increasing reliance on predictive algorithms comes a growing concern of exacerbating health disparities [1-3]. Evidence has shown that widely used algorithms that use past health care expenditures to predict high-risk patients have systematically underestimated the health care needs of Black patients [4]. In addition, studies have shown that predictive performances of models predicting intensive care unit mortality, 30-day psychiatric readmission, and asthma exacerbation were worse in populations with lower socioeconomic status [5,6].

With algorithmic bias as a potentially pervasive issue, a few checklists have been published to qualitatively identify and understand the potential biases derived from predictive models [7,8]. However, no agreed-upon quantitative method exists to routinely assess whether deployed models will lead to biased results and exacerbate health disparities faced by marginalized groups [2,9]. In this study, we define algorithmic bias as the differential results or performance of predictive models that may lead to differential allocation or outcomes between subgroups [10-12]. In addition, we define disparity as the difference in the quality of health care (the degree to which health services increase the likelihood of desired health outcomes) received by a marginalized population that is not due to access-related factors, clinical needs, preferences, and appropriateness of intervention [10,13]. Fairness metrics, which are a set of mathematical expressions that formalize certain equality between groups (eg, equal false negative rates [FNRs]), were proposed to measure and detect biases in machine learning models [12,14]. Although the machine learning community has shown that fairness metrics are a promising way to identify algorithmic bias, these metrics are criticized for being insufficient to reflect the heterogeneous and dynamic nature of health care [15,16]. Fairness metrics can also be misleading or conflicting due to their narrow focus on equal rates between groups [12,15]. Furthermore, these metrics could be interpreted without context-specific judgment or domain knowledge, thus failing to connect predictions to interventions and the downstream health care disparity [15,17]. Most importantly, these measures are often not fully tested in real-world predictive tasks and lack evidence on how well these measures' interpretation could guide intervention planning.

Background of Disparity in 30-Day Hospital Readmission

Predicting hospital readmissions is widely studied in health care management and delivery [18-21]. Hospital readmissions, especially unplanned or avoidable readmissions, are not only associated with a high risk of in-hospital mortality but also costly and burdensome to the health care system [19,22]. Since 2012, the Hospital Readmission Reduction Program by the Centers for Medicare & Medicaid Services (CMS) has imposed financial penalties for hospitals with excessive readmission rates [22]. CMS has consequently incentivized hospitals to segment patients by risk so that hospitals can target the delivery of these resource-intensive interventions to the patients at greatest risk, such as transitional care intervention and better discharge planning [19,23,24]. Many hospital readmission predictive models have been published, with >350 models predicting 30-day readmission identified in prior systematic reviews and our prior work [7,18,19,21,25]. The disparity in hospital readmission rates is well studied. For example, past studies have shown that Black patients have higher readmission rates after adjusting for demographic and clinical characteristics [26-29]. In addition to racial disparity, patients receiving care at racial and ethnic minority-serving hospitals [29,30] or living in disadvantaged neighborhoods have higher rates of readmission [31-33]. Research has also shown that disparity in health care use, including hospital readmission, is related to not only individuals' racial and ethnic identity but also their communities [34]. Other research has also suggested that social environments, either the place of residence or the hospital where one receives care, may explain a meaningful portion of health disparity [35,36].

Objectives

Despite model abundance and known disparity in hospital readmissions, research has been limited in evaluating how algorithmic bias or the disparate performances of these predictive models may impact patient outcomes and downstream health disparities once deployed. Lack of evidence is more prominent in how the model-guided intervention allocation may reduce or aggravate existing health disparities between different populations. To address this gap in evidence, in this study, we aimed to (1) implement a selection of fairness metrics to evaluate whether the application of common 30-day readmission predictive models may lead to bias between racial and income groups and (2) interpret the selected fairness metrics and assess their usefulness in the context of facilitating equitable allocation of interventions. In this paper, we represent the perspective of a health system or payer who uses an established, validated algorithm to identify patients at high risk of unplanned readmission so that targeted intervention can be planned for these patients. Thus, our main concern for algorithmic bias is the unequal allocation of intervention resources and the unequal health outcome as a result. Specifically, we are concerned about risk scores systematically underestimating or overestimating needs for a certain group, assuming the model we deploy is validated and has acceptable overall predictive performance.

Methods

Study Population and Data

This retrospective study included 1.9 million adult inpatient discharges in Maryland and 8.7 million inpatient discharges in Florida from 2016 to 2019. The State Inpatient Databases (SIDs) are maintained by the United States Agency for Healthcare

XSL•FO
RenderX

Research and Quality, as part of the Healthcare Cost and Utilization Project (HCUP), were used for this analysis. The SIDs include longitudinal hospital care data in the United States, inclusive of all insurance payers (eg, Medicare, Medicaid, private insurance, and the uninsured) and all patient ages [37]. The SIDs capture >97% of all eligible hospital discharges in each state [38]. Maryland and Florida were selected due to their different population sizes, compositions (eg, racial and ethnic distribution and urban to rural ratio), and health care environment (Maryland's all-payer model vs Medicaid expansion not adopted in Florida) [39,40]. In addition, Maryland and Florida are among a small subset of states in which the SIDs contain a "VisitLink" variable that tracks unique patients within the state and across years from 2016 to 2019, allowing for the longitudinal analysis of readmissions across hospitals and different calendar years [41]. The SIDs were also linked to the American Hospital Association's Annual Survey Database to obtain hospital-level information. The study population excluded admissions where patients were aged <18 years, died in hospitals, were discharged against medical advice, or had insufficient information to calculate readmission (eg, missing the VisitLink variable or length of stay).

Study Outcome

The calculation of 30-day readmission followed the definition used by the HCUP [42]. Any inpatient admission was counted as an index admission. The all-cause 30-day readmission rate was defined as the number of admissions with at least 1 subsequent hospital admission within 30 days, divided by the total number of admissions during the study period. Unplanned, all-cause 30-day hospital readmissions were identified using the methodology developed by CMS [43,44]. The study cohort selection process and determination of unplanned readmission are outlined in Figure 1.

Figure 1. Determination of the study cohort and unplanned all-cause 30 - day readmission.

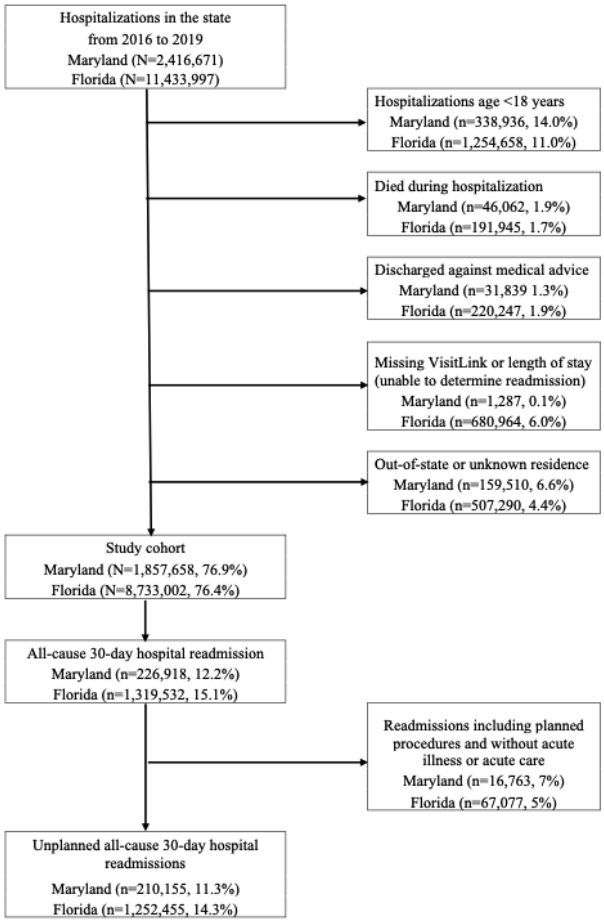

Predictive Models

The LACE index [45], the HOSPITAL score [46], and the CMS hospital-wide all-cause readmission measure [43] were included in the analysis as they were validated externally and commonly used in practice based on our prior review [7]. The LACE index and the HOSPITAL score were designed for hospital staff to identify patients at high risk of readmission for targeted intervention efforts and have been converted to a scoring system and extensively validated. Thus, the 2 models were applied to obtain the predicted risk scores without retraining, to mimic how the models were used in practice. In total, 2 of the HOSPITAL score predictors—low hemoglobin and low sodium levels at discharge—were not available in the SIDs, and thus were excluded. The total risk scores were adjusted as a result. Details of model variables and how the 2 models were implemented are reported in Multimedia Appendices 1 and 2. The CMS measure was evaluated using 2 approaches: applied *as-is* with existing coefficients and *retrained* to generate new coefficients using 50% of the sample. To ensure comparability between the CMS measure and other models, the predicted

XSL•FO
RenderX

patient-level risk was used without the hospital-level effect from the original measure, and the CMS measure was limited to the "medicine cohort" [43]. On the basis of the CMS measure's specification report, the patient population was divided into 5 mutually exclusive cohorts: surgery or gynecology, cardiorespiratory, cardiovascular, neurology, and medicine. The cohorts were determined using the Agency for Healthcare Research and Quality Clinical Classifications Software categories [43]. The medicine cohort was randomly split 50-50 into a retraining and testing data set. The CMS measure includes age and >100 variables, representing a wide range of condition categories. The measure was trained on the retraining data set with 5 cross-validations and then run on the testing data set using the new coefficients to obtain the performance and bias metrics for the CMS retrained model. Separately, the CMS measure with the published coefficients was run on the full medicine cohort data set to obtain performance and bias metrics for the CMS as-is model. The existing model thresholds were used to classify a positive, or high-risk, class: 10 points for LACE, and high-risk (5 in the adjusted scoring) for modified HOSPITAL. The optimal threshold identified using the Youden Index [47] on the receiver operating characteristic curve was used for the 2 CMS measures.

Measures

We measured predictive performances and biases between Black and White subpopulations and between low-income and other-income subpopulations. Race is a normalized variable in the HCUP that indicates race and ethnicity. The low-income group was defined as the fourth quartile of the median state household income, whereas the remaining 3 quartiles were grouped as other income. The median state income quartiles were provided in HCUP SIDs and were calculated based on the median income of the patient's zip code. Predictive performances of each model were derived for all population and each subpopulation using area under the curve (AUC), Brier statistic, and Hosmer-Lemeshow goodness of fit. Bias was represented by the group difference of the mathematical measures: false positive rate (FPR) difference (eg, FPR between Black and White patients), FNR difference, 0-1 loss difference, and generalized entropy index (GEI). FNR was calculated as the ratio between false negatives (those predicted as low risk while having an unplanned 30-day readmission) and the total number of positives. Similarly, the FPR was calculated as the ratio of false positives out of the total number of negative cases. Normalized total error rates is 0-1 loss, and it is calculated as the percentage of incorrect predictions. Bias measured by FPR, FNR, and 0-1 loss differences focus on unequal error rates. The GEI is a measure of income inequality and proposed to measure algorithm fairness between groups with a range between 0 and infinity, in which lower scores represent more equity [48].

Ethical Considerations

This study was not human subjects research, as determined by the Johns Hopkins School of Public Health Institutional Review Board. No compensation was provided.

Statistical Analysis

Primary analyses were conducted using R (version 4.0.2; R Foundation for Statistical Computing). The aggregate condition categories required to calculate unplanned readmission and CMS measures were calculated in SAS software (version 9.4; SAS Institute) using the programs provided by the agencies [49,50]. GEI measures were calculated using the AI Fairness 360 package published by IBM Corp [51]. The unit of analysis was admission. FNR and FPR results were first stratified by individual hospital and visualized in a scatter plot. The racial bias results were then stratified by hospital population composition (eg, percentage of Black patients), which was shown to associate with the overall outcome of a hospital [35]. Hospitals were binned by the percentage of Black patients served in a hospital (eg, >10% and >20%), and the racial bias measures with their 95% CIs were calculated for each bin. For FNR difference, FPR difference, and 0-1 loss difference, the distribution across 2 groups was calculated, and the significance of the measure difference was assessed using the Student *t* test (2-tailed) under the null hypothesis that the group difference was equal to 0. For all statistical tests, an α of .05 was used.

Results

Demographic and Clinical Characteristics

As presented in Table 1, among the 1,857,658 Maryland inpatient discharges from 2016 to 2019, a total of 55.41% (n=1,029,292) were White patients and 33.71% (n=626,280) were Black patients, whereas in Florida, 64.49% (5,632,318/8,733,002) of the inpatient discharges were White patients and 16.59% (1,448,620/8,733,002) were Black patients.

White patients in both states were older, more likely to be on private insurance, and less likely to reside in large metropolitan areas or be treated in major teaching or large hospitals in urban areas. Compared to White patients, Black patients in Maryland had a longer length of inpatient stay, more inpatient procedures, fewer inpatient diagnoses, higher inpatient charges, and more comorbidities and were more likely to be discharged to home or self-care. However, Black patients in Florida had fewer inpatient diagnoses, fewer procedures, and fewer total charges. These patients also had longer lengths of inpatient stays, more comorbidities, and were more likely to be discharged to home or self-care. In both Maryland and Florida, those in the lowest income quartile were younger, had a longer length of inpatient stay, had higher inpatient charges, had more comorbidities, and had fewer procedures than other-income groups. The low-income group was less likely to reside in metropolitan areas but was more likely to be treated in major teaching hospitals. Except for those noted in footnote c of Table 1, all characteristics showed statistically significant differences between racial and income groups (all *P* values <.001).

XSL•FO
RenderX

Table 1. Demographic characteristics by race and by income in Maryland (n=1,857,658) and Florida (n=8,733,002).

Characteristics and state	Race			Income	
	White	Black	Other	Low income	Other income
Discharges, n (%)					
MD[a]	1,029,292 (55.41)	626,280 (33.71)	187,935 (10.12)	627,013 (33.75)	1,225,820 (65.99)
FL[b]	5,632,318 (64.49)	1,448,620 (16.59)	1,598,392 (18.3)	2,600,326 (29.78)	6,028,720 (69.03)
Age (y), mean (SD)					
MD	61.4 (19.9)	54.0 (19.4)	46.8 (20.7)	56.3 (19.8)	58.0 (20.8)
FL	63.1 (19.4)	51.1 (19.8)	56.1 (21.6)	58.0 (20.4)	60.6 (20.5)
Sex (female; yes), n (%)					
MD	586,641 (56.99)	377,063 (60.21)	128,138 (68.18)	364,718 (58.17)	733,135 (59.81)
FL	3,050,611 (54.16)	856,343 (59.11)	942,996 (59)	1,454,239 (55.93)	3,372,155 (55.93)[c]
Payer, n (%)					
Medicare					
MD	550,364 (53.47)	262,512 (41.92)	43,461 (23.13)	293,009 (46.73)	566,503 (46.21)
FL	3,386,956 (60.13)	589,574 (40.7)	715,483 (44.76)	1,358,285 (52.24)	3,301,163 (54.76)
Medicaid					
MD	142,138 (13.81)	192,443 (30.73)	71,450 (38.02)	192,624 (30.72)	215,530 (17.58)
FL	504,531 (8.96)	373,896 (25.81)	315,306 (19.73)	505,805 (19.45)	682,879 (11.33)
Private					
MD	306,929 (29.82)	148,781 (23.76)	59,381 (31.6)	120,346 (19.19)	398,459 (32.51)
FL	1,183,304 (21.01)	287,528 (19.85)	398,712 (24.94)	422,946 (16.27)	1,439,482 (23.88)
Residence (large metropolitan), n (%)					
MD	839,688 (81.58)	586,868 (93.72)	178,015 (94.71)	460,587 (73.46)	1,154,341 (94.17)
FL	2,939,039 (52.18)	1,027,033 (70.9)	1,319,976 (82.58)	1,530,136 (58.84)	3,738,731 (62.02)
Length of stay, mean (SD)					
MD	4.73 (6.10)	5.22 (7.20)	4.19 (6.22)	5.17 (6.89)	4.68 (6.32)
FL	4.96 (6.37)	5.24 (7.98)	4.76 (6.76)	5.16 (7.26)	4.89 (6.51)
Total charges, mean (SD)					
MD	17,000 (22,700)	17,800 (25,800)	14,600 (23,000)	18,000 (25,100)	16,500 (23,200)
FL	68,500 (88,800)	62,800 (95,900)	68,900 (100,000)	67,300 (93,600)	67,800 (92,000)
Discharge type (home or self-care), n (%)					
MD	780,238 (75.8)	490,051 (78.25)	165,538 (88.08)	486,882 (77.65)	956,510 (78.03)
FL	4,408,065 (78.26)	1,240,314 (85.62)	1,367,677 (85.57)	2,102,035 (80.84)	4,873,457 (80.84)[c]
CCI[d] score, mean (SD)					
MD	0.498 (1.04)	0.594 (1.18)	0.359 (0.962)	0.573 (1.13)	0.486 (1.06)
FL	0.516 (1.04)	0.616 (1.23)	0.529 (1.13)	0.575 (1.13)	0.516 (1.07)
Number of procedures, mean (SD)					
MD	1.68 (2.44)	1.71 (2.60)	2.00 (2.38)	1.69 (2.59)	1.74 (2.44)
FL	1.57 (2.34)	1.50 (2.34)	1.57 (2.31)	1.49 (2.33)	1.59 (2.33)
Number of diagnoses, mean (SD)					
MD	15.5 (8.20)	14.6 (7.84)	11.1 (7.07)	15.3 (8.04)	14.4 (8.08)
FL	13.3 (7.33)	11.7 (7.19)	10.8 (6.91)	12.6 (7.33)	12.5 (7.29)

XSL•FO
RenderX

Characteristics and state	Race			Income	
	White	Black	Other	Low income	Other income
Hospital type (major teaching; yes), n (%)					
MD	181,493 (17.63)	120,649 (19.26)	27,808 (14.80)	172,092 (27.45)	158,693 (12.95)
FL	576,819 (10.24)	231,379 (15.97)	242,754 (15.19)	354,901 (13.65)	691,635 (11.47)
Hospital beds (≥200 beds), n (%)					
MD	741,956 (72.08)	460,667 (73.56)	150,745 (80.21)	482,604 (76.97)	878,817 (71.69)
FL	3,546,967 (62.98)	981,735 (67.77)	967,036 (60.5)	1,735,704 (66.75)	3,730,066 (61.87)
Urban hospital (yes), n (%)					
MD	943,928 (91.71)	602,779 (96.25)	179,091 (95.29)	548,230 (87.44)	1,187,298 (96.86)
FL	2,859,038 (50.76)	858,993 (59.3)	903,705 (56.54)	1,340,735 (51.56)	3,256,784 (54.02)

[a]MD: Maryland.

[b]FL: Florida.

[c]*P* values were computed between racial groups and between income groups, respectively. All *P* values are <.001 except for the ones in this footnote: *P* value for female between income groups=.80 and for discharge type between income groups=.99.

[d]CCI: Charlson Comorbidity Index.

Predictive Performance

The observed 30-day unplanned readmission rates in Maryland were higher in the Black and low-income patient groups (ie, 11.13% for White patients, 12.77% for Black patients, 10.59% for other-income patients, and 12.73% for low-income patients; Table 2).

Table 2. Observed and predicted 30-day unplanned readmission rates by model and state.

	Observed (%)		Predicted[a] (%)							
	MD (n=1,857,658)	FL (n=8,733,002)	LACE[b]		HOSPITAL[c]		CMS[d] as-is		CMS retrained	
			MD[e] (n=1,857,658)	FL[f] (n=8,733,002)	MD (n=1,857,658)	FL (n=8,733,002)	MD (n=714,917)	FL (n=2,636,671)	MD (n=357,458)	FL (n=1,318,335)
Total	11.31	14.34	14.48	15.88	12	14.61	10.48	10.26	14.71	16.5
Race										
White	11.13	13.94	12.92	14.81	11.62	14.07	10.27	10.11	14.14	15.97
Black	12.77	17.14	18.62	21.33	14.12	18.26	10.82	10.85	15.84	18.93
Income										
Other	10.59	13.6	12.88	14.62	10.84	13.66	10.37	10.19	14.3	16.17
Low	12.73	16.03	17.6	18.72	14.29	16.77	10.67	10.43	15.46	17.25

[a]Predicted: the predicted readmission rates for LACE and HOSPITAL were calculated as the percentage of patients at high risk of unplanned readmission based on the model output for the group; and the predicted readmission rates for the two CMS models were the predicted probability of being at high risk of unplanned readmission for the group.

[b]LACE: The LACE Index for readmission risk.

[c]HOSPITAL: The modified HOSPITAL score for readmission risk.

[d]CMS: Centers for Medicare & Medicaid Services (readmission measure).

[e]MD: Maryland.

[f]FL: Florida.

A fair and well-calibrated predictive model would be assumed to overpredict or underpredict readmission rates to a similar degree across racial or income groups. Compared to the observed readmission rates, the LACE index overestimated readmission rates in all subpopulations and was more pronounced in Black and low-income populations. The readmission rates estimated by the modified HOSPITAL score were closest to the observed rates. The CMS as-is model underestimated across subpopulations, and the estimated rates of readmission were similar between subpopulations, while the retrained CMS model overestimated in all subpopulations to a similar degree. In Florida, the observed 30-day unplanned readmission rates were higher than those in Maryland in all populations. Similar to Maryland, Florida's observed readmission rates were also higher in the Black and low-income groups (ie, 13.94% for White populations, 17.14% for Black

XSL•FO
RenderX

populations, 13.6% for other-income populations, and 16.03% for low-income populations) and had similar overestimation and underestimation patterns (Table 2).

As presented in Table 3, in Maryland, the retrained CMS model had better predictive performance (AUC 0.74 in all subpopulations) than the other 3 models, which only achieved moderate predictive performance (AUC between 0.65 and 0.68). The modified HOSPITAL score had the best calibration (Brier score=0.16−0.19 in all subpopulations), whereas the CMS as-is model performed poorly on the Brier score. Calibration was better in the White (compared to the Black) population and other-income (compared to low-income) populations in both states, and the AUC was higher or similar in the Black (compared to the White) population. In Florida, the CMS retrained model also performed better than the other models in all subpopulations (AUC 0.68-0.72), and the modified HOSPITAL score had the best calibration (Brier score 0.19-0.21). All models demonstrated excellent goodness of fit across subpopulations (Table 3).

Table 3. Predictive performances of each 30-day readmission model in Maryland and Florida.

All vs group and performance measure	LACE[a]		HOSPITAL[b]		CMS[c] as-is		CMS retrained	
	MD[d]	FL[e]	MD	FL	MD	FL	MD	FL
All								
AUC[f]	0.68	0.68	0.65	0.66	0.66	0.65	0.74	0.69
Brier statistic	0.19	0.21	0.17	0.20	0.44	0.37	0.32	0.32
Hosmer-Lemeshow, *P* value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Group								
Race								
White								
AUC	0.68	0.67	0.64	0.65	0.65	0.63	0.74	0.68
Brier statistic	0.18	0.21	0.17	0.20	0.41	0.36	0.31	0.32
Hosmer-Lemeshow, *P* value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Black								
AUC	0.68	0.68	0.67	0.69	0.67	0.68	0.74	0.72
Brier statistic	0.22	0.25	0.19	0.21	0.49	0.43	0.33	0.36
Hosmer-Lemeshow, *P* value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Income								
Other								
AUC	0.69	0.68	0.65	0.66	0.65	0.64	0.74	0.69
Brier statistic	0.17	0.20	0.16	0.19	0.43	0.36	0.31	0.32
Hosmer-Lemeshow, *P* value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Low								
AUC	0.68	0.67	0.66	0.67	0.67	0.66	0.74	0.70
Brier statistic	0.21	0.23	0.19	0.21	0.47	0.39	0.33	0.34
Hosmer-Lemeshow, *P* value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

[a]LACE: The LACE Index for readmission risk.

[b]HOSPITAL: The modified HOSPITAL score for readmission risk.

[c]CMS: Centers for Medicare & Medicaid Services (readmission measure).

[d]MD: Maryland.

[e]FL: Florida.

[f]AUC: area under the curve.

XSL•FO
RenderX

Bias Measures

Misclassification rates (ie, FPR difference and FNR difference) indicate relative between-group bias, whereas 0-1 loss differences indicate the overall error rates between groups. The between-group GEI indicates how unequally an outcome is distributed between groups [48]. In Maryland, the retrained CMS model and the modified HOSPITAL score had the lowest racial and income bias (Table 4).

Specifically, the modified HOSPITAL score demonstrated the lowest racial bias based on 0-1 loss, FPR difference, and GEI, and the lowest income bias based on FPR and GEI. Retrained CMS demonstrated the lowest racial bias based on 0-1 loss and FNR difference, and the lowest income bias on all 4 measures. In Florida, racial biases based on FPR and FNR differences was generally greater than that in Maryland, especially for FNR differences. In Florida, the modified HOSPITAL score showed the lowest racial bias based on 0-1 loss, FPR difference, and GEI; the LACE index showed the lowest racial bias in FNR difference. Each model scored the best in at least one measure of income bias, but the overall HOSPITAL score and retrained CMS showed the lowest income bias in Florida. In both states, the White and other-income patient groups had a higher FNR, indicating that they were more likely to be predicted as low risk while having a 30-day unplanned readmission. The Black and low-income patient groups had a higher FPR, indicating that they were more likely to be predicted to be high-risk and not have a 30-day unplanned readmission. The overall error rates were higher in the Black and low-income patient groups compared to the White and other-income patient groups, respectively. Except for GEI and the values noted with a footnote in Table 4, all other measures showed statistically significant differences (all P values <.001) between racial and income groups, respectively.

XSL•FO
RenderX

Table 4. Bias measures of evaluated 30-day readmission models in Maryland and Florida.

Measures and state	All	White	Black	Difference (B-W)[a]	Other income	Low income	Difference (L-O)[a]
LACE[b]							
Maryland							
0-1 loss	0.19	0.18	0.22	0.04	0.17	0.21	0.04
FNR[c]	0.69	0.72	0.63	−0.10[d]	0.71	0.65	−0.06
FPR[e]	0.12	0.11	0.16	0.05	0.11	0.15	0.04
GEI[f] (between-group)	N/A[g]	N/A	N/A	0.03	N/A	N/A	0.02
Florida							
0-1 loss	0.21	0.21	0.25	0.04	0.20	0.23	0.03
FNR	0.68	0.71	0.60	−0.11	0.70	0.64	−0.06
FPR	0.13	0.12	0.18	0.05	0.12	0.16	0.04
GEI (between-group)	N/A	N/A	N/A	0.02	N/A	N/A	0.01
HOSPITAL[h]							
Maryland							
0-1 loss	0.17	0.17	0.19	0.02	0.16	0.19	0.03
FNR	0.73	0.75	0.69	−0.07	0.75	0.69	−0.06
FPR	0.10	0.10	0.12	0.02	0.10	0.12	0.03
GEI (between-group)	N/A	N/A	N/A	0.01	N/A	N/A	0.01
Florida							
0-1 loss	0.20	0.20	0.21	0.01	0.19	0.21	0.02
FNR	0.68	0.71	0.59	−0.12	0.71	0.64	−0.07
FPR	0.12	0.12	0.14	0.02	0.11	0.14	0.02
GEI (between-group)	N/A	N/A	N/A	0.01	N/A	N/A	0.01
CMS[i] (as-is)							
Maryland							
0-1 loss	0.44	0.41	0.49	0.07	0.43	0.47	0.04
FNR	0.30	0.34	0.24	−0.10	0.32	0.26	−0.06
FPR	0.47	0.43	0.54	0.11	0.45	0.51	0.06
GEI (between-group)	N/A	N/A	N/A	0.05	N/A	N/A	0.02
Florida							
0-1 loss	0.37	0.36	0.43	0.08	0.36	0.39	0.02
FNR	0.42	0.47	0.28	−0.18	0.44	0.38	−0.06
FPR	0.36	0.34	0.47	0.13	0.35	0.40	0.04
GEI (between-group)	N/A	N/A	N/A	0.05	N/A	N/A	0.02
CMS (retrained)							
Maryland							
0-1 loss	0.32	0.31	0.33	0.02	0.31	0.33	0.02
FNR	0.29	0.31	0.26	−0.05	0.30	0.27	−0.03[d]
FPR	0.32	0.31	0.35	0.04	0.31	0.35	0.03
GEI (between-group)	N/A	N/A	N/A	0.02	N/A	N/A	0.01
Florida							

XSL•FO
RenderX

Measures and state	All	White	Black	Difference (B-W)[a]	Other income	Low income	Difference (L-O)[a]
0-1 loss	0.32	0.32	0.36	0.04	0.32	0.34	0.02
FNR	0.38	0.41	0.28	−0.13	0.40	0.35	−0.05[d]
FPR	0.31	0.30	0.38	0.08	0.31	0.34	0.04
GEI (between-group)	N/A	N/A	N/A	0.03	N/A	N/A	0.01

[a]The columns Difference (B-W) and Difference (L-O) indicate algorithmic bias measured as the difference in the bias measure (eg, FNR and FPR) between Black and White patients and between low-income and other-income groups.

[b]LACE: The LACE Index for readmission risk.

[c]FNR: false negative rate.

[d]All *P* values of the bias measures are <.001 except for the ones in this footnote: the *P* value for FNR difference of LACE in MD is .41, and the FNR difference of CMS retrained in MD is .45, and FNR difference of CMS retrained in FL is .005. Statistical tests were not conducted for the GEI as this measure produces one value for the population.

[e]FPR: false positive rate.

[f]GEI: generalized entropy index.

[g]N/A: not applicable.

[h]HOSPITAL: The modified HOSPITAL score for readmission risk.

[i]CMS: Centers for Medicare & Medicaid Services (readmission measure).

Stratification Analyses

The results were first stratified by hospital and then by patient population composition (percentage of Black patients). As shown in Figure 2, the models' FNR differences and FPR differences between the Black and White patients varied by hospital within the state, indicating hospital shifts when applying the same model. The modified HOSPITAL score was more likely to cluster near the "equality lines" (ie, when the FNR or FPR difference is 0) than other models in both states. Colors representing LACE and CMS as-is were mostly distributed in the first quadrant in Maryland, indicating that the majority of hospitals had a positive FPR difference (ie, Black patients with higher FPR) and a negative FNR difference (ie, White patients with higher FNR) when applying these 2 models (Figure 2). Despite most hospitals falling in the first quadrant, the variance between hospitals appeared to be greater in Florida (Figure 3). In addition, more hospitals in Florida fell in the far corners of the first and fourth quadrants than those in Maryland, indicating more hospitals with severe bias (eg, large racial differences in FPR or FNR). Refer to Multimedia Appendix 3 for the measures of income bias and hospital distribution for Maryland and Florida.

Figure 2. Measures of racial bias in predicting readmission across individual hospitals in Maryland. CMS: Centers for Medicare & Medicaid Services; FNR: false negative rate; FPR: false positive rate.

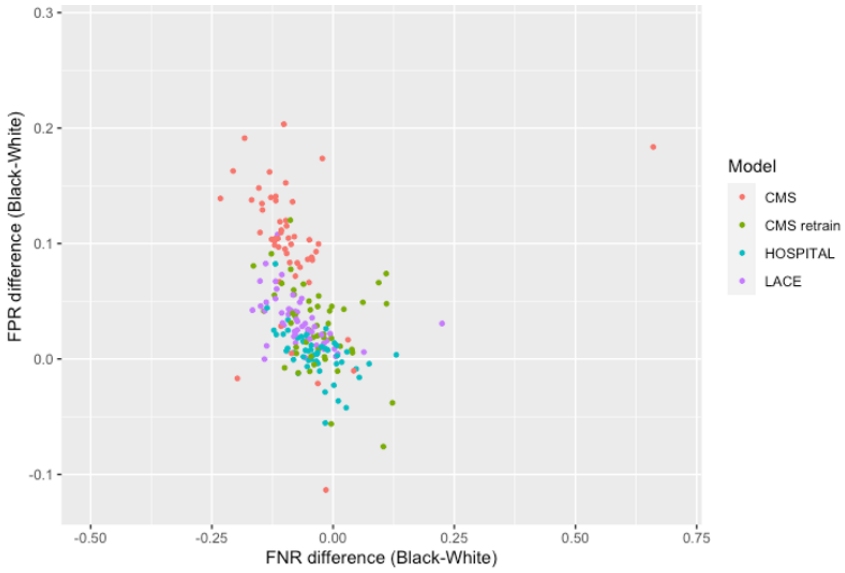

XSL•FO
RenderX

Figure 3. Measures of racial bias in predicting readmission across individual hospitals in Florida. CMS: Centers for Medicare & Medicaid Services; FNR: false negative rate; FPR: false positive rate.

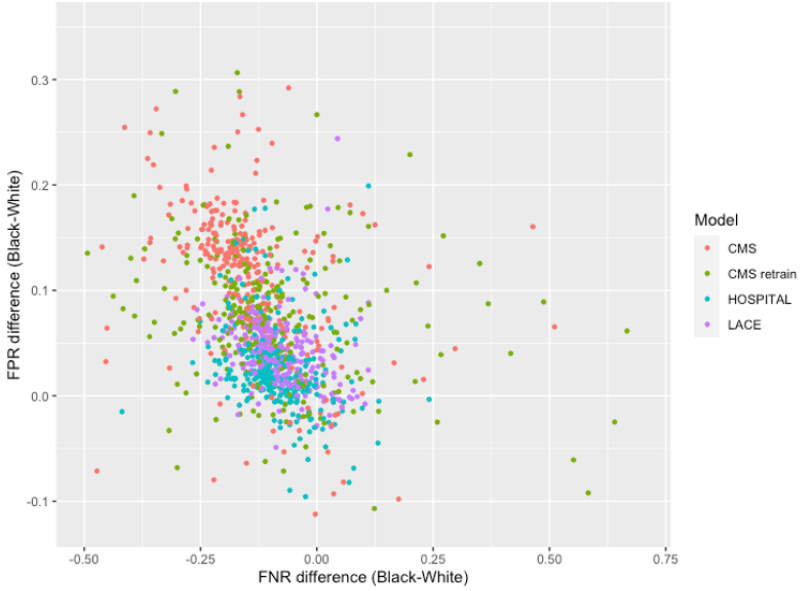

Hospitals with a higher percentage of Black patients have been shown to be associated with low resources and poorer outcomes for their patients [35]; thus, the results were stratified by the proportion of Black patients served in a hospital. In Figures 4 and 5, each data point represents the racial bias (FNR difference or FPR difference) in a stratum of hospitals with a certain percentage of Black patients (eg, hospitals with at least 20% of Black patients). The error bars show the 95% CI of the bias measure in the strata. In both figures, the racial biases of all models, represented as FNR and FPR differences, decreased and approached zero as the hospital population became more diverse. In Maryland, the diminishing racial bias was particularly notable in hospitals where >50% of patients were Black (Figure 4). The diminishing racial bias was also observed in Florida's hospitals (Figure 5). The direction of bias flipped for the LACE index and the modified HOSPITAL score in Florida hospitals with >50% of Black patients. In hospitals with a lower percentage of Black patients, Black patients had a lower FNR compared to White patients, while in hospitals with a higher percentage of Black patients, White patients had a higher FNR (Figure 4). In Florida, the widening gap shown in the 2 CMS models for hospitals serving >60% of Black patients was likely attributed to the small number of hospitals and small sample size in the strata (Figure 5). Refer to Multimedia Appendix 4 for the details on the bias measures stratified by payers for both Maryland and Florida.

Figure 4. Bias measures by hospital patient population composition (percentage of Black patients) in Maryland. CMS: Centers for Medicare & Medicaid Services; FNR: false negative rate; FPR: false positive rate.

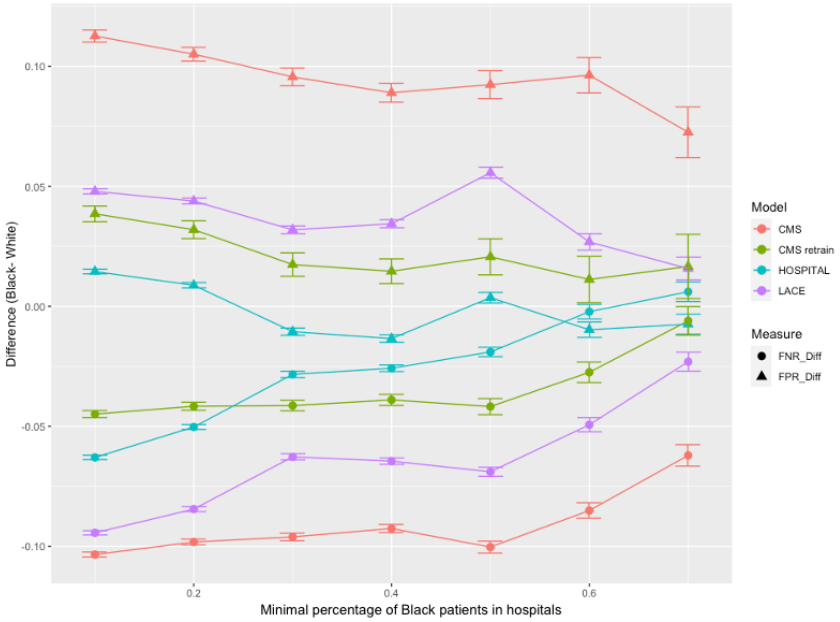

XSL•FO
RenderX

Figure 5. Bias measures by hospital patient population composition (a percentage of Black patients) in Florida. CMS: Centers for Medicare & Medicaid Services; FNR: false negative rate; FPR: false positive rate.

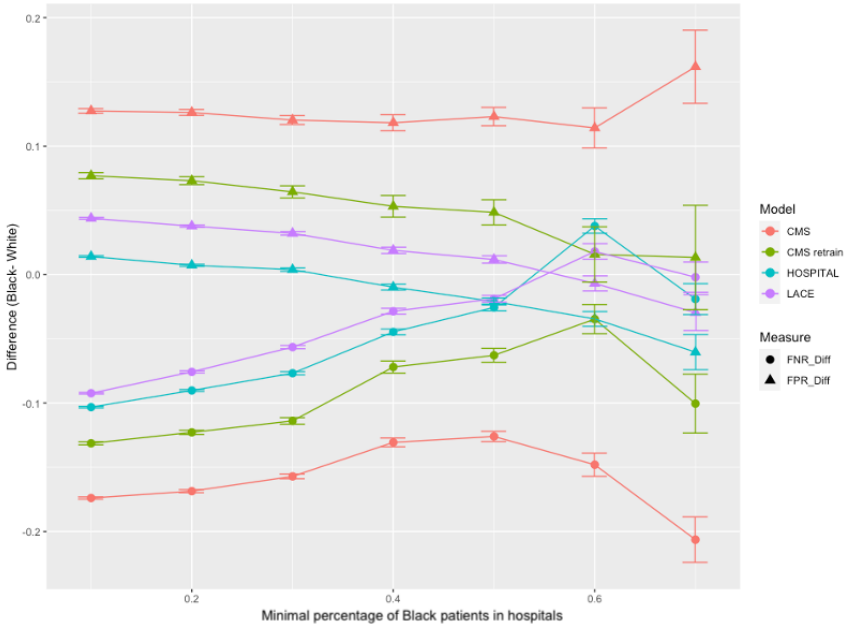

Discussion

Overall Findings

The abundance of research on fairness and bias has provided potential means to quantify bias, but there has been a gap to operationalize these metrics, interpret them in specific contexts, and understand their impact on downstream health disparity [7]. Our analysis demonstrated a practical use case for measuring algorithmic bias when applying or deploying previously validated 30-day hospital readmission predictive models in a new setting. Our approach to testing the fairness measures could serve as a framework for routine assessment of algorithmic bias for health care predictive models, and our results also revealed the complexity and limitations of using mathematical bias measures. According to these bias measures, the retrained CMS model and the modified HOSPITAL score showed the best predictive performance and the lowest bias in Maryland and Florida. However, the CMS as-is model showed subpar performance in both states, indicating that retraining on the local data not only improved predictive performance but also reduced group bias. In addition, large variations were detected between hospitals, and system- or hospital-level factors needed to be considered when interpreting algorithmic bias.

Measure Interpretation

Caution must be taken when using algorithmic bias to guide equitable intervention allocation, as the bias measures may not include key context. When designing a risk-based intervention based on model output, we would be naturally more concerned about FNR, as a higher FNR means a groups that is more likely to be predicted low in risk of readmission will indeed be readmitted, indicating missed opportunities for intervention [52]. Looking at bias measures alone, our results suggest that the risk to White and higher-income patients has a systematically higher proportion of false negatives estimated by common readmission models, suggesting more missed opportunities to intervene and prevent unplanned readmissions. This observation is contrary to our assumption, and other parts of the results show that White and higher-income patients were less sick with lower readmission rates. An explanation would suggest that the higher FNR observed in the White and higher-income patient groups might be attributed to health care use patterns. For example, research has shown that White individuals and higher socioeconomic patient groups were more likely to overuse health care resources, while Black patients and disadvantaged groups tended to underuse them [53-55]. The overutilizers could have more unplanned visits to the hospital when the risk was not high, while the underusing group may be more likely to defer or skip care and only use costly hospital resources when they must. Similarly, a higher FPR in Black and low-income patient groups would indicate more wasted resources on “false positives.” However, such a conclusion did not align with the rest of the study findings. These subpopulations, on average, had more chronic comorbidities and longer inpatient stays, indicating that Black and low-income patient groups were more likely to have conditions that warrant an unplanned readmission but did not show up in the observed data, potentially alluding to a health care access gap in these groups. In this case, drawing a conclusion simply based on the face value of higher FPR would lead to a reduction in the resources allocated to the sicker, more vulnerable populations. It is also important to note that, despite the racial difference in health behaviors and outcomes, race merely represents a social classification rather than the driver of the observed differences [56]. Although the performance of the evaluated readmission models differed by race, we do not recommend including race as a variable in a predictive model unless race is a biological or clinical risk factor for the predictive outcome.

The interpretation of measurable bias requires considering models’ predictive performance, the nature of health data, analytic frameworks, and the underlying health care delivery system. In our analysis, all models had modest performance,

XSL•FO
RenderX

and the high FNRs may deter their application in a real setting, especially in the score-based models of LACE and HOSPITAL (ie, FNR ranges from 0.63 to 0.75). When calculating these measures, we assumed the observed outcome (ie, 30-day unplanned readmission) as the ground truth; however, it was important to recognize the key limitations of this truth and the measured bias. First, despite the HCUP state inpatient data being one of the most comprehensive and high-quality data for studying readmission, no guarantee existed that all readmissions and their causes were captured. It is possible that a patient had conditions that warranted an unplanned revisit to the hospital but either did not occur due to the patient's unwillingness to seek treatment in time [57,58] or did not get documented (eg, out-of-state admissions were not captured in HCUP's state-wide inpatient data by design). Such underdocumentation was more likely to impact disadvantaged populations and those with fragmented care, thus introducing embedded bias into the underlying data. Second, a higher percentage of Black patients sought care in academic teaching institutions (eg, 120,649/626,280, 19.26% of Black patients in Maryland and 231,379/1,448,620, 15.97% of Black patients in Florida, compared to 181,493/1,029,292, 17.63% of White patients in Maryland and 576,819/5,632,318, 10.24% of White patients in Florida), which were generally considered to deliver high-quality care [35,59,60]. These hospitals may have a more effective readmission prevention program while serving sicker patients, contributing to a higher FPR among Black and low-income patients. Third, as shown in Figure 2, we observed that hospitals that served a high proportion of Black patients had a lower algorithmic bias. For example, in Maryland, the majority Black hospitals (>70% of patients served are Black) were in resource-poor neighborhoods, and both White and Black patients had similar higher-than-average readmission rates in these hospitals (data not shown). The fairer model performance in these hospitals was not necessarily a reflection of a higher quality of care, as all patients served in those hospitals had higher unplanned readmission rates. Finally, whether a readmission was unplanned or planned was determined using a well-established algorithm developed by CMS [43,44], which categorized readmissions based on the nature of the diagnoses and procedures (eg, acute vs routine). Research demonstrated that different diagnosis intensities existed between regions and hospitals, and a higher intensity of services was associated with a higher prevalence of common chronic diseases [61]. If diagnosis was not just a patient attribute but indeed reflected the systematic characteristics of the health care environment [62], the quality of unplanned readmission classification and other predictors in our models would be subject to encoded bias in the health care system. In fact, in our population, the average number of diagnoses was higher in White patients than in Black patients and higher in Maryland than in Florida, indicating the presence of such systematic variation (Table 1). Of course, this is not a unique issue with our data set; electronic health records and other health data sets also reflect histories of unequal access to health care and carry racial, ethnic, socioeconomic, and other societal biases due to how the data are collected [2,3,63].

Utility of Bias Measures

Once the limitations of real-world health data are acknowledged, the expectation of equity and interpretation of the measurable bias should adjust accordingly. First, it will be too restrictive to expect mathematical equality for measurable bias; rather, it is best viewed as a relative value to aid in the selection of a less biased model. Most real-world problems are based on imperfect data, and pushing the model to perform equally on these measures will inevitably create unintended results (eg, sacrificing accuracy and potentially increasing bias for other subpopulations) [15]. Second, a validated and accurate model may reveal the gap between the "supposed-to-be" state and the reality in the underlying data, showing areas of unmet needs [16,64], as we observed in our Black and low-income populations. Finally, the bias measures alone provide limited evidence about which group is being biased against and in which way. A conclusion based solely on the face value of a few bias measures can be misleading and may exacerbate the disparity already faced by marginalized groups. These quantitative bias measures are useful to evaluate a model's disparate group performance on a given data set, but they are insufficient to inform the intervention allocation or mechanisms of potential bias, which are key to the mitigation strategies [15]. In addition, our study did not evaluate other definitions of bias, such as calibration or predictive parity, which do not focus on error rates and may require unique interpretation considerations.

This analysis addressed a fundamental gap in operationalizing fairness techniques. The selection of a bias definition and appropriate bias measures is as important as detecting bias itself, yet it has remained a blind spot in practice [2]. In addition to the fact that these mathematical notions cannot be satisfied simultaneously, using the appropriate measures is also highly contextual and data dependent [65,66]. For example, having a model with equal positive predictions across groups (known as demographic or statistical parity) would not be a meaningful measure for inherently unbalanced outcomes such as 30-day readmissions; however, based on the fairness concept, satisfying any of the bias measures would mean a fair model. In this study, the 4 evaluated bias measures showed consistent results, despite each measuring a different definition of bias. All selected measures were able to demonstrate the magnitude of bias, but FNR and FPR differences were the most informative, as they indicated the direction of bias and were more interpretable in the context of mitigation actions. In our attempt to translate the algorithmic bias findings to intervention planning, we found that the bias measures could serve as a quick and routine assessment to compare algorithms, subpopulations, or localities (eg, hospitals) to help target further investigation of drivers of potential disparity. However, simply relying on these statistical notions to make decisions could obscure or underplay the causes of health care disparities, and a more comprehensive approach is necessary. In real-world applications, the practical goal of predictive modeling must incorporate predictive accuracy and algorithmic bias, among other operational considerations. As there is usually a trade-off between these 2 model performance goals, the best model is likely the one that balances the 2 goals rather than the one achieving the highest possible accuracy or fairness alone.

XSL•FO
RenderX

Limitations

Our analysis has several limitations and caveats. First, none of the models evaluated in this analysis had high accuracy, which may affect the measurement of misclassifications. For simplicity and the focus on interpreting the bias measures, we did not evaluate machine learning models that usually improve local accuracy [20]. While the LACE index and the HOSPITAL score were used by hospitals to manage readmissions, the CMS measure was mostly used in payer operations or population health management in addition to CMS purposes (eg, budget allocation and hospital penalties); thus, it was not used as a typical predictive model. Although we believe the models evaluated in this study represented practical scenarios, we were unable to assess if a particular type of models, variables, weights, or modeling structures were more likely to be algorithmically biased. Second, we did not evaluate the scenario in which models can be optimized to minimize and constraint bias during training or retraining. Model optimization has been a popular approach to developing fair models but, it was considered out of scope as this analysis focused on model application and bias identification. Third, we only included bias measures that are algorithm-agnostic and can be routinely calculated; thus, they were not comprehensive or exclusive. Fourth, the conclusion was based on Maryland and Florida data, which would not represent all states nor the national average. For example, Maryland is a small state with an all-payer model payment system [39] and a high percentage of patients seeking care in neighboring states, whereas Florida is a large state with a large Hispanic population and has not adopted Medicaid expansion [40]. In addition, the data set we used was administrative in nature and did not have the detailed medical information (eg, medications, laboratory results, and clinical notes) to fully evaluate the potential drivers of our results, such as selection bias [67], data quality factors [68], and more accurate ascertainment of the outcome (ie, unplanned readmissions).

Conclusions

In conclusion, our analysis found that fairness metrics were useful to serve as a routine assessment to detect disparate model performance in subpopulations and to compare predictive models. However, these metrics have limited interpretability and are insufficient to inform mechanisms of bias or guide intervention planning. Further testing and demonstration will be required before using mathematical fairness measures to guide key decision-making or policy changes. Despite these limitations, demonstrating the differential model performances (eg, misclassification rates) is often the first step in recognizing potential algorithmic bias, which will be necessary as health care organizations move toward data-driven improvement in response to existing health care disparities. The potential subtle—and not so subtle—imperfections of underlying health data, analytic frameworks, and the underlying health care delivery system must be carefully considered when evaluating the potential bias that exists within predictive models. Finally, future research is required to improve the methodology of measuring algorithmic bias and to test more fairness definitions and measures (eg, calibration parity) through an operational lens. Future studies should also explore how modeling factors influence algorithmic bias (eg, how variable inclusions, weights, or scoring schemes affect the model's differential performance). We hope that algorithmic bias assessment can be incorporated into routine model evaluation and ultimately inform meaningful actions to reduce health care disparity.

Acknowledgments

The authors acknowledge the contributions of Dr Darrell Gaskin and Dr Daniel Naiman of Johns Hopkins University for their input into the study conceptualization and results interpretation.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The data sets analyzed during this study are available for a fee from the Agency for Healthcare Research and Quality Healthcare Cost and Utilization Project [37].

Authors' Contributions

HEW and HK conceived the study concept, and all authors contributed to the study design. HEW analyzed the data and wrote the manuscript. HK and HEW interpreted the results, and all authors provided input for the interpretations. All authors reviewed and contributed to the final manuscript.

Conflicts of Interest

HEW is an employee of Merck & Co, and the employer had no role in the development or funding of this work. SS has received funding from NIH, NSF, CDC, FDA, DARPA, AHA, and Gordon Betty Moore Foundation. She is an equity holder in Bayesian Health, a clincial AI platform company; Duality Tech, a privacy preserving technology comlany and sits on the scientific advisory board of large life sciences (Eg Sanofi) and digital health startups (eg Century Health). She has received honoraria for talks from a number of biotechnology, research and health-tech companies. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict-of-interest policies.

XSL•FO
RenderX

Multimedia Appendix 1
The LACE index.
[DOCX File , 14 KB - jmir_v26i1e47125_app1.docx]

Multimedia Appendix 2
The modified HOSPITAL score.
[DOCX File , 15 KB - jmir_v26i1e47125_app2.docx]

Multimedia Appendix 3
Income bias and hospital distribution in Maryland and Florida.
[DOCX File , 142 KB - jmir_v26i1e47125_app3.docx]

Multimedia Appendix 4
Racial and income bias measures by payer in Maryland and Florida.
[DOCX File , 21 KB - jmir_v26i1e47125_app4.docx]

References

1. Rojas JC, Fahrenbach J, Makhni S, Cook SC, Williams JS, Umscheid CA, et al. Framework for integrating equity into machine learning models: a case study. Chest 2022 Jun;161(6):1621-1627 [FREE Full text] [doi: 10.1016/j.chest.2022.02.001] [Medline: 35143823]
2. Ferryman K. Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. J Am Med Inform Assoc 2020 Dec 09;27(12):2016-2019 [FREE Full text] [doi: 10.1093/jamia/ocaa133] [Medline: 32951036]
3. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021 Jul;4:123-144 [FREE Full text] [doi: 10.1146/annurev-biodatasci-092820-114757] [Medline: 34396058]
4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
5. Juhn YJ, Ryu E, Wi C, King KS, Malik M, Romero-Brufau S, et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. J Am Med Inform Assoc 2022 Jun 14;29(7):1142-1151 [FREE Full text] [doi: 10.1093/jamia/ocac052] [Medline: 35396996]
6. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 2019 Feb 01;21(2):E167-E179 [FREE Full text] [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]
7. Wang HE, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc 2022 Jul 12;29(8):1323-1333 [FREE Full text] [doi: 10.1093/jamia/ocac065] [Medline: 35579328]
8. Obermeyer Z, Nissan R, Stern M. Algorithmic bias playbook. Center for Applied AI. 2021. URL: https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias/playbook [accessed 2024-03-05]
9. Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, et al. Algorithmic fairness in computational medicine. EBioMedicine 2022 Oct;84:104250 [FREE Full text] [doi: 10.1016/j.ebiom.2022.104250] [Medline: 36084616]
10. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care, Smedley BD, Stith AY, Nelson AR. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: National Academies Press; 2003.
11. Rathore SS, Krumholz HM. Differences, disparities, and biases: clarifying racial variations in health care use. Ann Intern Med 2004 Oct 19;141(8):635-638. [doi: 10.7326/0003-4819-141-8-200410190-00011] [Medline: 15492343]
12. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the 40th International Workshop on Software Fairness. 2018 Presented at: FairWare '18; May 29, 2018; Gothenburg, Sweden p. 1-7 URL: https://dl.acm.org/doi/10.1145/3194770.3194776
13. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
14. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021 Jul 13;54(6):1-35. [doi: 10.1145/3457607]
15. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. J Biomed Inform 2021 Jan;113:103621 [FREE Full text] [doi: 10.1016/j.jbi.2020.103621] [Medline: 33220494]
16. Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR Public Health Surveill 2020 Oct 22;6(4):e22400 [FREE Full text] [doi: 10.2196/22400] [Medline: 33090117]
17. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc 2020 Dec 09;27(12):2020-2023 [FREE Full text] [doi: 10.1093/jamia/ocaa094] [Medline: 32574353]

XSL•FO
RenderX

18. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. Comput Methods Programs Biomed 2018 Oct;164:49-64. [doi: 10.1016/j.cmpb.2018.06.006] [Medline: 30195431]
19. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011 Oct 19;306(15):1688-1698 [FREE Full text] [doi: 10.1001/jama.2011.1515] [Medline: 22009101]
20. Huang Y, Talwar A, Chatterjee S, Aparasu RR. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. BMC Med Res Methodol 2021 May 06;21(1):96 [FREE Full text] [doi: 10.1186/s12874-021-01284-z] [Medline: 33952192]
21. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ 2020 Apr 08;369:m958 [FREE Full text] [doi: 10.1136/bmj.m958] [Medline: 32269037]
22. Hospital readmissions reduction program (HRRP). Centers for Medicare & Medicaid Services. 2023 Sep. URL: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program [accessed 2023-12-05]
23. Teo K, Yong CW, Muhamad F, Mohafez H, Hasikin K, Xia K, et al. The promise for reducing healthcare cost with predictive model: an analysis with quantized evaluation metric on readmission. J Healthc Eng 2021;2021:9208138 [FREE Full text] [doi: 10.1155/2021/9208138] [Medline: 34765104]
24. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. Implementation of artificial intelligence-based clinical decision support to reduce hospital readmissions at a regional hospital. Appl Clin Inform 2020 Aug 02;11(4):570-577 [FREE Full text] [doi: 10.1055/s-0040-1715827] [Medline: 32877943]
25. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open 2016 Jun 27;6(6):e011060 [FREE Full text] [doi: 10.1136/bmjopen-2016-011060] [Medline: 27354072]
26. Pandey A, Keshvani N, Khera R, Lu D, Vaduganathan M, Joynt Maddox KE, et al. Temporal trends in racial differences in 30-day readmission and mortality rates after acute myocardial infarction among medicare beneficiaries. JAMA Cardiol 2020 Feb 01;5(2):136-145 [FREE Full text] [doi: 10.1001/jamacardio.2019.4845] [Medline: 31913411]
27. Rodriguez-Gutierrez R, Herrin J, Lipska KJ, Montori VM, Shah ND, McCoy RG. Racial and ethnic differences in 30-day hospital readmissions among US adults with diabetes. JAMA Netw Open 2019 Oct 02;2(10):e1913249 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.13249] [Medline: 31603490]
28. Jiang HJ, Andrews R, Stryer D, Friedman B. Racial/ethnic disparities in potentially preventable readmissions: the case of diabetes. Am J Public Health 2005 Sep;95(9):1561-1567. [doi: 10.2105/AJPH.2004.044222] [Medline: 16118367]
29. Tsai TC, Orav EJ, Joynt KE. Disparities in surgical 30-day readmission rates for medicare beneficiaries by race and site of care. Ann Surg 2014 Jun;259(6):1086-1090 [FREE Full text] [doi: 10.1097/SLA.0000000000000326] [Medline: 24441810]
30. Joynt KE, Orav EJ, Jha AK. Thirty-day readmission rates for medicare beneficiaries by race and site of care. JAMA 2011 Feb 16;305(7):675-681 [FREE Full text] [doi: 10.1001/jama.2011.123] [Medline: 21325183]
31. Kind AJ, Jencks S, Brock J, Yu M, Bartels C, Ehlenbach W, et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Ann Intern Med 2014 Dec 02;161(11):765-774 [FREE Full text] [doi: 10.7326/M13-2946] [Medline: 25437404]
32. Hu J, Kind AJ, Nerenz D. Area deprivation index predicts readmission risk at an urban teaching hospital. Am J Med Qual 2018;33(5):493-501 [FREE Full text] [doi: 10.1177/1062860617753063] [Medline: 29357679]
33. Gershon AS, Thiruchelvam D, Aaron S, Stanbrook M, Vozoris N, Tan WC, et al. Socioeconomic status (SES) and 30-day hospital readmissions for chronic obstructive pulmonary (COPD) disease: a population-based cohort study. PLoS One 2019;14(5):e0216741 [FREE Full text] [doi: 10.1371/journal.pone.0216741] [Medline: 31112573]
34. Gaskin DJ, Dinwiddie GY, Chan KS, McCleary R. Residential segregation and disparities in health care services utilization. Med Care Res Rev 2012 Apr;69(2):158-175 [FREE Full text] [doi: 10.1177/1077558711420263] [Medline: 21976416]
35. López L, Jha AK. Outcomes for whites and blacks at hospitals that disproportionately care for black medicare beneficiaries. Health Serv Res 2013 Feb;48(1):114-128 [FREE Full text] [doi: 10.1111/j.1475-6773.2012.01445.x] [Medline: 22816447]
36. LaVeist T, Pollack K, Thorpe Jr R, Fesahazion R, Gaskin D. Place, not race: disparities dissipate in southwest Baltimore when blacks and whites live under similar conditions. Health Aff (Millwood) 2011 Oct;30(10):1880-1887 [FREE Full text] [doi: 10.1377/hlthaff.2011.0640] [Medline: 21976330]
37. Healthcare cost and utilization project (HCUP). Agency for Healthcare Research and Quality. 2023. URL: https://www.ahrq.gov/data/hcup/index.html [accessed 2024-03-05]
38. Metcalfe D, Zogg CK, Haut ER, Pawlik TM, Haider AH, Perry DC. Data resource profile: state inpatient databases. Int J Epidemiol 2019 Dec 01;48(6):1742-172h [FREE Full text] [doi: 10.1093/ije/dyz117] [Medline: 31280297]
39. Maryland all-payer model. Center for Medicare & Medicaid Services. 2022 Aug. URL: https://www.cms.gov/priorities/innovation/innovation-models/maryland-all-payer-model [accessed 2023-03-05]
40. Status of state action on the Medicaid expansion decision. Kaiser Family Foundation. 2022 Sep. URL: https://www.cms.gov/priorities/innovation/innovation-models/maryland-all-payer-model [accessed 2023-03-05]

XSL•FO
RenderX

41. User guide: HCUP supplemental variables for revisit analyses. Agency for Healthcare Research and Quality. 2022. URL: https://www.hcup-us.ahrq.gov/toolssoftware/revisit/UserGuide-SuppRevisitFilesCD.pdf [accessed 2024-03-06]
42. Statistical brief no. 248. characteristics of 30-day all-cause hospital readmissions, 2010-2016. Healthcare Cost and Utilization Project. 2019 Feb. URL: https://hcup-us.ahrq.gov/reports/statbriefs/sb248-Hospital-Readmissions-2010-2016.jsp [accessed 2023-03-05]
43. Hospital-wide all-cause risk-standardized readmission measure: measure methodology report. Center for Medicare & Medicaid Services. 2012 Jul. URL: https://www.cms.gov/priorities/innovation/files/fact-sheet/bpciadvanced-fs-nqf1789.pdf [accessed 2023-03-04]
44. Horwitz LI, Grady JN, Cohen DB, Lin Z, Volpe M, Ngo CK, et al. Development and validation of an algorithm to identify planned readmissions from claims data. J Hosp Med 2015 Oct;10(10):670-677 [FREE Full text] [doi: 10.1002/jhm.2416] [Medline: 26149225]
45. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010 Apr 06;182(6):551-557 [FREE Full text] [doi: 10.1503/cmaj.091117] [Medline: 20194559]
46. Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Intern Med 2013 Apr 22;173(8):632-638. [doi: 10.1001/jamainternmed.2013.3023] [Medline: 23529115]
47. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J 2005 Aug;47(4):458-472. [doi: 10.1002/bimj.200410135] [Medline: 16161804]
48. Speicher T, Heidari H, Grgic-Hlaca N. A unified approach to quantifying algorithmic unfairness: measuring individual and group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018 Presented at: KDD '18; August 19-23, 2018; London, UK p. 2239-2248 URL: https://dl.acm.org/doi/10.1145/3219819.3220046 [doi: 10.1145/3219819.3220046]
49. Clinical classifications software (CCS) for ICD-10-PCS (beta version). Healthcare cost and utilization project (HCUP). Agency for Healthcare Research and Quality. Rockville, MD; 2019 Nov. URL: https://hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp [accessed 2021-11-10]
50. Risk adjustment 2020 model software/ICD-10 mappings 2020. Center for Medicare & Medicaid Services. 2020 Sep. URL: https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors [accessed 2023-12-05]
51. AI fairness 360. IBM Corp. 2020. URL: https://aif360.res.ibm.com/ [accessed 2021-04-05]
52. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872 [FREE Full text] [doi: 10.7326/M18-1990] [Medline: 30508424]
53. Chandler D. The underutilization of health services in the black community: an examination of causes and effects. J Black Stud 2008 Aug 06;40(5):915-931. [doi: 10.1177/0021934708320723]
54. Best MJ, McFarland EG, Thakkar SC, Srikumaran U. Racial disparities in the use of surgical procedures in the US. JAMA Surg 2021 Mar 01;156(3):274-281 [FREE Full text] [doi: 10.1001/jamasurg.2020.6257] [Medline: 33439237]
55. Kressin NR, Groeneveld PW. Race/Ethnicity and overuse of care: a systematic review. Milbank Q 2015 Mar;93(1):112-138 [FREE Full text] [doi: 10.1111/1468-0009.12107] [Medline: 25752352]
56. Jones CP. Levels of racism: a theoretic framework and a gardener's tale. Am J Public Health 2000 Aug;90(8):1212-1215. [doi: 10.2105/ajph.90.8.1212] [Medline: 10936998]
57. Hewins-Maroney B, Schumaker A, Williams E. Health seeking behaviors of African Americans: implications for health administration. J Health Hum Serv Adm 2005;28(1):68-95. [Medline: 16521616]
58. Moser DK, Kimble LP, Alberts MJ, Alonzo A, Croft JB, Dracup K, et al. Reducing delay in seeking treatment by patients with acute coronary syndrome and stroke: a scientific statement from the American Heart Association Council on cardiovascular nursing and stroke council. Circulation 2006 Jul 11;114(2):168-182. [doi: 10.1161/CIRCULATIONAHA.106.176040] [Medline: 16801458]
59. Allison JJ, Kiefe CI, Weissman NW, Person SD, Rousculp M, Canto JG, et al. Relationship of hospital teaching status with quality of care and mortality for medicare patients with acute MI. JAMA 2000 Sep 13;284(10):1256-1262. [doi: 10.1001/jama.284.10.1256] [Medline: 10979112]
60. Popescu I, Nallamothu BK, Vaughan-Sarrazin MS, Cram P. Racial differences in admissions to high-quality hospitals for coronary heart disease. Arch Intern Med 2010 Jul 26;170(14):1209-1215. [doi: 10.1001/archinternmed.2010.227] [Medline: 20660838]
61. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES. Regional variations in diagnostic practices. N Engl J Med 2010 Jul 01;363(1):45-53 [FREE Full text] [doi: 10.1056/NEJMsa0910881] [Medline: 20463332]
62. Welch HG, Sharp SM, Gottlieb DJ, Skinner JS, Wennberg JE. Geographic variation in diagnosis frequency and risk of death among medicare beneficiaries. JAMA 2011 Mar 16;305(11):1113-1118 [FREE Full text] [doi: 10.1001/jama.2011.307] [Medline: 21406648]
63. Ghassemi M, Nsoesie EO. In medicine, how do we machine learn anything real? Patterns (N Y) 2022 Jan 14;3(1):100392 [FREE Full text] [doi: 10.1016/j.patter.2021.100392] [Medline: 35079713]

XSL•FO
RenderX

64. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med 2020 Jan;26(1):16-17. [doi: 10.1038/s41591-019-0649-2] [Medline: 31932779]
65. Srivastava M, Heidari H, Krause A. Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019 Presented at: KDD '19; August 4-8, 2019; Anchorage, AK p. 2459-2468 URL: https://dl.acm.org/doi/10.1145/3292500.3330664 [doi: 10.1145/3292500.3330664]
66. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform 2021 Apr;28(1):e100289 [FREE Full text] [doi: 10.1136/bmjhci-2020-100289] [Medline: 33910923]
67. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak 2014 Jun 11;14:51 [FREE Full text] [doi: 10.1186/1472-6947-14-51] [Medline: 24916006]
68. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013 Oct;46(5):830-836 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.010] [Medline: 23820016]

Abbreviations

AUC: area under the curve
CMS: Centers for Medicare & Medicaid Services
FNR: false negative rate
FPR: false positive rate
GEI: generalized entropy index
HCUP: Healthcare Cost and Utilization Project
SIDs: State Inpatient Databases

Edited by A Mavragani; submitted 11.03.23; peer-reviewed by D Nerenz, J Herington; comments to author 07.12.23; revised version received 28.12.23; accepted 27.02.24; published 18.04.24.

Please cite as:
Wang HE, Weiner JP, Saria S, Kharrazi H
Evaluating Algorithmic Bias in 30-Day Hospital Readmission Models: Retrospective Analysis
J Med Internet Res 2024;26:e47125
URL: https://www.jmir.org/2024/1/e47125
doi:10.2196/47125
PMID:38422347

©H Echo Wang, Jonathan P Weiner, Suchi Saria, Hadi Kharrazi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research

Joseph D Romano[1,2,3], MA, MPhil, PhD; Van Truong[1,4,5], MS; Rachit Kumar[1,4,5,6], BS; Mythreye Venkatesan[7], BE, MS; Britney E Graham[7], PhD; Yun Hao[1,4], PhD; Nick Matsumoto[7], BA; Xi Li[7], MS; Zhiping Wang[7], MS, PhD; Marylyn D Ritchie[1,3,5], PhD; Li Shen[1,3], PhD; Jason H Moore[7], PhD

[1]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]Center of Excellence in Environmental Toxicology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[3]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[4]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[5]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[6]Medical Scientist Training Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[7]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States

Corresponding Author:
Joseph D Romano, MA, MPhil, PhD
Institute for Biomedical Informatics
Perelman School of Medicine
University of Pennsylvania
403 Blockley Hall
423 Guardian Drive
Philadelphia, PA, 19104
United States
Phone: 1 2155735571
Email: joseph.romano@pennmedicine.upenn.edu

Abstract

Background: As global populations age and become susceptible to neurodegenerative illnesses, new therapies for Alzheimer disease (AD) are urgently needed. Existing data resources for drug discovery and repurposing fail to capture relationships central to the disease's etiology and response to drugs.

Objective: We designed the Alzheimer's Knowledge Base (AlzKB) to alleviate this need by providing a comprehensive knowledge representation of AD etiology and candidate therapeutics.

Methods: We designed the AlzKB as a large, heterogeneous graph knowledge base assembled using 22 diverse external data sources describing biological and pharmaceutical entities at different levels of organization (eg, chemicals, genes, anatomy, and diseases). AlzKB uses a Web Ontology Language 2 ontology to enforce semantic consistency and allow for ontological inference. We provide a public version of AlzKB and allow users to run and modify local versions of the knowledge base.

Results: AlzKB is freely available on the web and currently contains 118,902 entities with 1,309,527 relationships between those entities. To demonstrate its value, we used graph data science and machine learning to (1) propose new therapeutic targets based on similarities of AD to Parkinson disease and (2) repurpose existing drugs that may treat AD. For each use case, AlzKB recovers known therapeutic associations while proposing biologically plausible new ones.

Conclusions: AlzKB is a new, publicly available knowledge resource that enables researchers to discover complex translational associations for AD drug discovery. Through 2 use cases, we show that it is a valuable tool for proposing novel therapeutic hypotheses based on public biomedical knowledge.

(J Med Internet Res 2024;26:e46777) doi:10.2196/46777

KEYWORDS

Alzheimer disease; knowledge graph; knowledge base; artificial intelligence; drug repurposing; drug discovery; open source; Alzheimer; etiology; heterogeneous graph; therapeutic targets; machine learning; therapeutic discovery

XSL•FO
RenderX

Introduction

Background

Alzheimer disease (AD) is a progressive, neurodegenerative disease affecting an estimated 6.5 million Americans aged ≥65 years and represents a significant clinical, economic, and emotional burden worldwide [1]. AD is often cited as one of the greatest health care problems of the 21st century, particularly in high-income nations with an increasing proportion of older adults. Despite its societal impact, effective pharmaceutical treatments for AD remain notoriously elusive. The US Food and Drug Administration has approved 5 drugs for the treatment of AD, 4 of which (donepezil, rivastigmine, galantamine, and memantine) only temporarily treat symptoms but do not alter the overall progression of the disease [2], whereas the fifth (aducanumab) is highly controversial in terms of evidence of effectiveness and its safety profile [3]. AD researchers have prioritized the discovery and approval of new therapies for the disease both in terms of newly discovered compounds and by repurposing drugs that are already approved to treat other (non-AD) human diseases.

AD is associated with substantial changes in pathology, including the presence of neuritic plaques associated with the amyloid-β protein, extracellular deposition of amyloid-β, and neurofibrillary tangles. Previous research has shown that these neuropathological changes begin to occur years before clinical symptoms are apparent [4,5]. Despite decades of research, why this pathology begins to develop remains largely unknown [6]. Current consensus is that AD risk is multifactorial. The most well-established risk factors include age; family history; and certain genetic factors, especially the presence of the σ4 allele of the *apolipoprotein E* gene, which is involved in fat metabolism and cholesterol transport. However, the exact mechanism through which these factors—including *APOE*-σ4 presence—cause or contribute to AD risk is unknown [7].

Of the many techniques used in AD therapeutics research, there is a wealth of computer-aided approaches that leverage recent advances in bioinformatics, epidemiology, artificial intelligence (AI), and machine learning (ML). For example, Rodriguez et al [8] developed an ML framework to assess gene lists constructed by differential gene expression data in response to drug treatment to determine whether those drugs would be candidates for repurposing in AD. Tsuji et al [9] used an autoencoder neural network to perform dimensionality reduction of a high-density protein interaction network to identify new possible drug targets and then found drugs associated with those targets. Genome-wide association studies have long been used for the identification of genes that confer AD risk, particularly for rare genes or genes with small (but statistically significant) contributions to disease risk [10].

In this paper, we describe the design and deployment of a major new knowledge resource for computational AD research—named The Alzheimer's Knowledge Base (AlzKB) [11]—with a particular focus on drug discovery and drug repurposing. The overall structure and contents of AlzKB are summarized in Figure 1. At its core, AlzKB consists of a large, heterogeneous graph database describing entities related to AD at multiple levels of biological organization, with rich semantic relationships describing how those entities are linked to one another. To demonstrate its value, we present two data-driven analyses involving ML on AlzKB's knowledge graph: (1) predicting Parkinson disease (PD) genes that may also be associated with AD and (2) generating and explaining drug repurposing hypotheses for treating AD, both of which replicate existing knowledge while proposing entirely novel directions for future experimental validation. AlzKB is free, open source, and publicly available [11] and consists entirely of publicly sourced knowledge integrated from 22 diverse web-based biomedical databases. We hypothesized that the relationships and entities in AlzKB contain valuable knowledge that cannot be effectively captured in existing data resources, with the additional advantage of improving the explainability of new predictions.

XSL•FO
RenderX

Figure 1. Schematic overview of the Alzheimer's Knowledge Base (AlzKB).

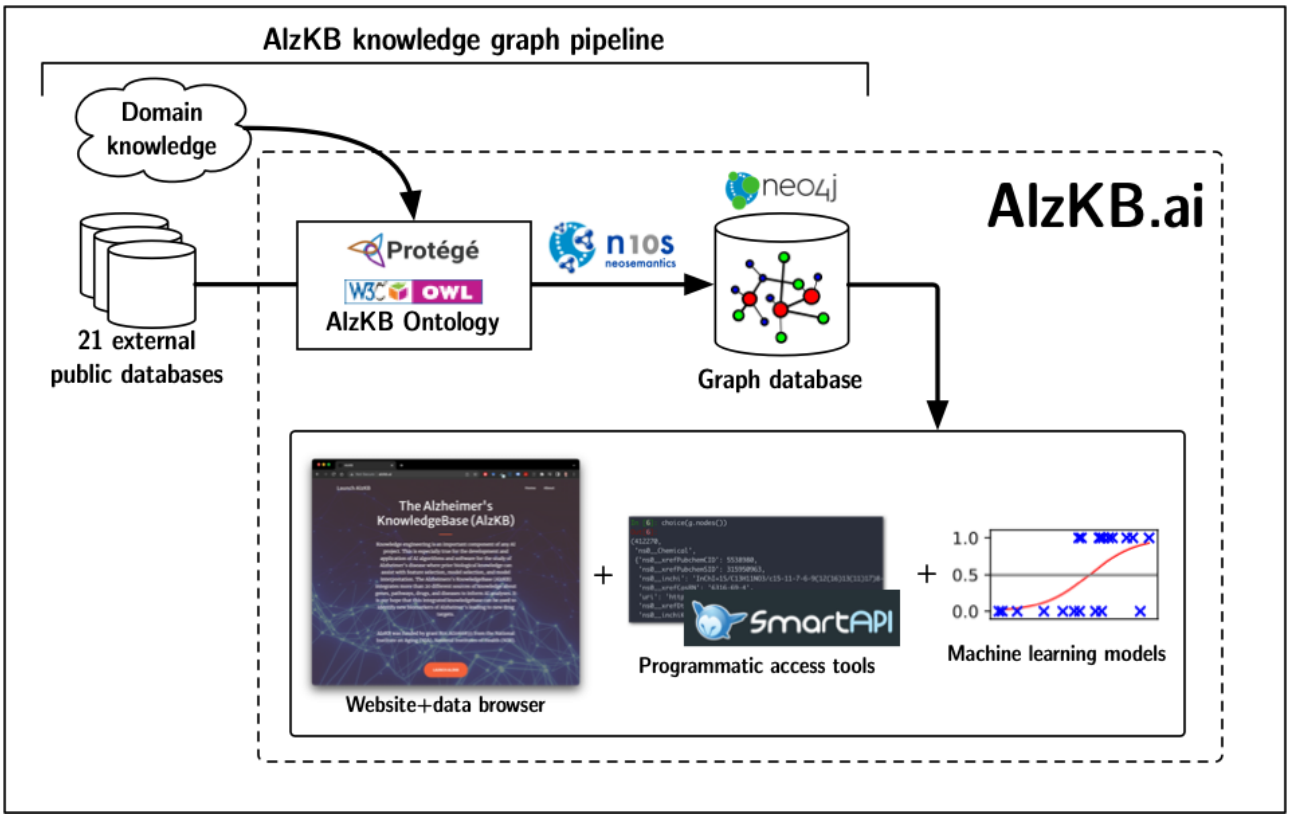

Existing Graph-Based Approaches to AD Research

Due to the increased popularity and success of analyses using integrated knowledge, previous efforts have used knowledge graphs in AD research for a variety of purposes, including drug repurposing [12-14] and gene identification [15] and as general informational resources [16]. Similar to AlzKB, these bodies of work draw from a variety of sources to construct the underlying knowledge graphs, including scientific literature and formally structured biomedical databases. Some, including the Alzheimer Disease Knowledge Graph [14] and the Heterogeneous network-based data set for AD [16], have been released as publicly accessible resources similar to AlzKB. Other studies have used existing resources not specifically intended for AD research (such as the Semantic MEDLINE Database [13]) to answer questions related to AD. To our knowledge, AlzKB is the largest graph-based knowledge representation that focuses solely on AD and draws from the greatest number of source databases. For comparison, the next largest AD-specific knowledge graph that we are aware of is AD-KG, which contains 30,729 nodes and 398,544 edges (compared to AlzKB's 118,902 nodes and 1,309,527 edges). Our emphasis on merging similar nodes or edges and cleaning the graph structure using an underlying biomedical ontology reduces the amount of noise that tends to be associated with many different node or edge types in a single graph, enabling more robust inference about relationships in AD, especially when used with emerging graph ML algorithms. Furthermore, AlzKB offers a public, web interface that allows for easy access and application to new research questions, whereas existing resources have either restricted access or are entirely unavailable for reuse. Given the challenge of identifying new or repurposed drugs for etiologically complex diseases such as AD, AlzKB represents a major step forward by improving both quantitatively and structurally on existing resources.

Methods

AlzKB Ontology

Graph databases are renowned for their flexibility in representing data that do not conform to a rigid, tabular structure, but this comes at the expense of implicitly enforcing consistency and semantic standardization [17]. To mitigate this issue, we designed a Web Ontology Language (OWL) 2 ontology—describing the types of entities relevant to AD and treatment of AD, as well as the types of relationships that link those entities—that serves as a *template* for nodes and edges in the knowledge graph. Ontologies (including OWL 2 ontologies) are formal representations of knowledge that are frequently used in biomedicine to computationally structure, retrieve, and make inferences about knowledge within a domain of interest [18]. Briefly, as many of the components of a graph database have a 1-to-1 correspondence with components of an OWL 2 ontology (eg, OWL 2 classes are equivalent to graph database node labels, and OWL 2 object properties are equivalent to edge types in a graph database), it is possible to populate the ontology using biomedical knowledge and translate the contents of the populated ontology into an equivalent graph database. Therefore, enforcing consistency in the ontology becomes equivalent to enforcing consistency in the graph database.

We constructed the ontology manually using the Protégé ontology editor (version 5.5.0; Stanford Center for Biomedical Informatics Research) [19] following an iterative process guided

XSL•FO
RenderX

by expert domain knowledge. First, we prototyped a class hierarchy containing the types of nodes (eg, gene, disease, pathway, and drug) desired in the knowledge base. We then annotated these classes with data properties (eg, drugs can be assigned a property value corresponding to molecular weight) and object properties (relationship types that link 2 entities, such as "drug treats disease"). A thorough description of the components of OWL 2 ontologies is provided by Hitzler et al [20]. Finally, we placed restrictions on the ontology to reflect biology and clinical practice. For example, we specified restrictions stating that all pathways must contain one or more genes or that all drugs in the knowledge base must have a valid DrugBank ID. We repeated these steps several times, making revisions on previous iterations until several domain experts agreed that the semantic contents of the ontology were consistent with current AD knowledge and systems biology processes involved in AD etiology. After collecting the data sources used to populate the ontology (see the following section), we included additional data properties corresponding to identifiers in those source databases, enabling data provenance and facilitating both interoperability and validation. The final ontology structure consists of entity types involved in AD etiology (modeled as OWL 2 classes), types of semantic relationships that can link those entity types (modeled as OWL 2 object properties), and properties that can be annotated onto entities of specific types (modeled as OWL 2 data properties). Both before and after populating the ontology with individuals (see the *Implementing AlzKB* section), we validated its contents and structure by running FaCT++—an ontology inference engine that identifies errors by evaluating all assertions in the ontology against the ontology's class or property hierarchy and other restrictions [21].

Collecting and Assembling Third-Party Data Sources

Using the AlzKB ontology's class hierarchy as a starting point, we determined a set of the most important entity types to include in the first release of the knowledge base. For example, we prioritized inclusion of entities representing diseases (specifically AD and its various subtypes), genes, and drugs, among others. Similarly, we identified important relationship types (eg, "DRUG_BINDS_GENE" or "GENE_ASSOCIATED_WITH_DISEASE") to include in the knowledge base. For each of these entity and relationship types, we identified a third-party, public data source that would serve as a collection of "ground truth knowledge" for that entity or relationship type. In the assembled knowledge base, there is roughly a 1-to-1 correspondence between a data record in the original "ground truth" data source and its corresponding entity or relationship in AlzKB, with some important exceptions. For example, we made the decision to only include neurological diseases in AlzKB rather than all diseases described in the "ground truth" data source (in this case, the Disease Ontology). We also identified instances in which properties from additional data sources could be used to augment the "ground truth" entities. For example, while DrugBank is used to specify the drugs described in AlzKB, we also used fields from Distributed Structure-Searchable Toxicity and PubChem to augment the properties annotated onto drugs (such as molecular weight, chemical fingerprint, and synonyms).

Implementing AlzKB

We populated the ontology by sequentially carrying out the following steps:

1. Import distinct entities from each data source corresponding to the corresponding ontology class and define those entities as ontology individuals (ie, instances of that class). For example, the drug memantine is defined as an instance of the ontology class Drug.
2. Populate data properties for all instances of each ontology class using data from relevant sources. For example, memantine is annotated with the Chemical Abstracts Service Registry number 19982-08-2.
3. Populate object properties as the semantic relationships linking pairs of entities using the appropriate data source. For example, an object property of type "DRUG_TREATS_DISEASE" links memantine to the instance of Disease named Alzheimer's Disease.

After populating the AlzKB ontology with entities, relationships, and data properties, we serialized the ontology into the Resource Description Framework (RDF) or XML graph data format, which is compatible with modern graph database software as an input format. A complete list of the data sources used in AlzKB at the time of writing is provided in Table 1. We then populated a Neo4j graph database (version 4.4.5; Neo4j, Inc) [22] with the contents of the RDF or XML file using the neosemantics library [23], which parses the RDF data, inserting semantic triples into the graph database corresponding to each entity or relationship. Finally, we stripped the newly populated graph database of unnecessary artifacts that are components of the OWL 2 standard, leaving only nodes, relationships, and properties defined within the hierarchy. For the publicly hosted version of AlzKB, we created a web server that hosts both the static AlzKB website (containing information, documentation, and use details) and the Neo4j graph database, which is available by navigating to a subdomain [24] of the main website [11]. For reproducibility, this entire pipeline (including mappings to source databases) is provided as a single Python script available on GitHub (the most recent version) [25] or Zenodo (an archived version of the code at the time of publication) [26].

XSL•FO
RenderX

Table 1. Third-party public data sources used in the Alzheimer's Knowledge Base (AlzKB), which data elements were used from them, and total size of the data source (counts of entities of relevant data types only)[a].

Data source	Use in AlzKB	Size (number of entities)
AOP-DB[b] [27]	Adverse outcome pathways and chemical-gene associations	1,207,456
Bgee[c] [28]	Tissue-specific gene expression data; only human gene expression data were used in AlzKB	9,093,494
Disease Ontology[c] [29]	Human diseases—only AD[d], subtypes of AD, and related neurodegenerative diseases were included in AlzKB	8043
DisGeNET [30]	Diseases, genes, and disease-gene associations with scores representing levels of evidence; only AD and related neurodegenerative terms were used for diseases	51,841
DrugBank [31]	On-market and experimental pharmaceutical drugs	15,550
EPA[e] DSSTox[f] [32]	Chemical toxicity data—filtered in AlzKB to drugs contained in DrugBank	1,200,059
EPA ACToR[g] [33]	Chemical toxicity data—filtered in AlzKB to drugs contained in DrugBank	504,871
Gene Ontology[c] [34,35]	Biological processes, molecular functions, and cellular components	42,950
GWAS[h] Catalog[c] [36]	Gene-disease associations	60,071
Hetionet [37]	Graph modeling and entity resolution (for data sources marked with footnote indicator "c")	47,031
Human Reference Protein Interactome Mapping Project[c] [38]	Human protein-protein interactions (modeled as gene-gene interactions)	9094
LINCS[i] L1000[c] [39]	Human differential gene expression data	7467 genes
NCBI[j] MeSH[k] [40]	Clinical and biomedical concepts (annotated to various node types)	Approximately 27,000
NCBI Entrez Gene [41]	Human genes and gene synonyms	62,407
Pathway Interaction Database[c] [42]	Pathways and gene-pathway membership	223
PharmacotherapyDB[c] [43]	Drug indications for human diseases	698
PubChem [44]	Chemical structures and identifiers—only chemicals in DrugBank were included in AlzKB	115,067,800
Reactome pathway database[c] [45]	Pathways and gene-pathway membership	1341
SIDER[c,l] [46]	Drug side effects (modeled as diseases)	5868
TISSUES[c] [47]	Tissue-specific gene expression data	__[m]
Uberon[c] [48]	Human anatomical structures	402
WikiPathways[c] [49]	Pathways and gene-pathway membership	298

[a]As source data elements do not correspond in a 1-to-1 manner with entities in the graph (eg, entities may be merged, filtered, or used as edges rather than nodes), actual counts for entities in AlzKB stratified by source are not available. The sizes are the best available estimates at the time of publication. Table 2 and Table S1 in Multimedia Appendix 1 [50-56] provide actual node and edge type counts in AlzKB.

[b]AOP-DB: Adverse Outcome Pathway Database.

[c]The derived data are structured in part using Hetionet.

[d]AD: Alzheimer disease.

[e]EPA: Environmental Protection Agency.

[f]DSSTox: Distributed Structure-Searchable Toxicity.

[g]ACToR: Aggregated Computational Toxicology Resource.

[h]GWAS: genome-wide association studies.

[i]LINCS: Library of Integrated Network-Based Cellular Signatures.

[j]NCBI: National Center for Biotechnology Information.

[k]MeSH: Medical Subject Headings.

[l]SIDER: Side Effect Resource.

XSL•FO
RenderX

[m]Counts not applicable (TISSUES associations map to edges rather than nodes in the graph).

Table 2. Node types and counts in the Alzheimer's Knowledge Base listed in descending order by prevalence. Additional node types will be added over time, and counts will increase as new data sources are incorporated or existing sources are updated to newer versions.

Node label	Total nodes, N
Gene	62,407
Drug	35,063
BiologicalProcess	11,381
Pathway	4570
MolecularFunction	2884
CellularComponent	1391
Symptom	438
BodyPart	402
DrugClass	345
Disease	20

Validating AlzKB Using Real-World Use Cases

After building AlzKB's knowledge graph, we designed two ML-based use cases that resemble real-world tasks for which AlzKB was originally designed: (1) proposing genetic targets for new drugs based on disease similarity and topological graph features and (2) predicting new edges in the knowledge graph linking AD to repurposed drugs via a graph completion model. These 2 use cases are intended to assess the external validity of AlzKB—for the ML models to perform well on tasks defined using real-world evaluation end points (eg, effective drugs or etiologically important genes), the informative patterns and phenomena underlying those end points need to be adequately captured in the knowledge graph.

In the first use case (identifying genetic targets via graph topology measures), we trained a random forest (RF) classifier (implemented in the scikit-learn library [Python Software Foundation] for the Python programming language) using the following topological graph features, which are computed for every node pair in the graph (regardless of whether an edge does or does not exist between them): common neighbors, total neighbors, preferential attachment, Adamic-Adar, and resource allocation [57-60]. Each feature gives a different measure of network "relatedness" for a pair of nodes, which are then used as predictive features in the RF model. For a given node pair (n_1, n_2), these metrics are defined as follows:

where $N(n_1)$ is the set of neighbor (adjacent) nodes of node *i*. Our training procedure for the RF model included 3-fold grid search cross-validation to optimize hyperparameters, an 80%/20% train/test split, and repeating the procedure 10 times with random sampling.

To accomplish the second use case (drug repurposing via graph completion models), we implemented and compared the performance of 5 graph completion algorithms applied to the entire AlzKB knowledge graph. These models learn low-dimensional representations of graph nodes as vector embeddings. The embeddings are then combined to propose all possible triples in the graph (source node, edge, and target node), and scores are generated to indicate the plausibility of the triple. The 5 models we evaluated are TransE, RotatE, DistMult, ComplEx, and ConvE [60].

We implemented the 5 models using PyKEEN—a Python library for knowledge graph embeddings [50]. We randomly split the data set of all triples into 80/10/10 training/validation/testing sets and used grid search to empirically set embedding dimensions to 256 and the number of epochs to 100 with early stopping allowed. All remaining hyperparameters were set to the PyKEEN defaults. We trained the models on Google Colab using a single Tesla T4 graphics processing unit and evaluated the results using the rank-based evaluation metrics hits@k (k=1, 3, and 10) and mean reciprocal rank (MRR) [61]. Ranking-based evaluation sorts the scores of triples in descending order and sets their rank as the index in the sorted list. In the case of multiple "true" triples having an equal score, we used the average of the most optimistic (best) and pessimistic (worst) ranks across the metrics. Briefly, hits@k is the ratio of true triples in the test set that have been ranked within the top k predictions of the model. Higher values indicate better performance. The MRR, also known as inverse harmonic mean rank, is the arithmetic mean of the inverse rank of the true triples. We performed evaluation on both left- and right-side predictions (ie, how well they can predict missing entities in partial triples without either the head [source] or tail [target] entities).

XSL•FO
RenderX

Ethical Considerations

No human participants were involved in this research. All data used to build and evaluate AlzKB were derived from publicly available biomedical knowledge retrieved from open access databases. None of the data included were derived from individual human participants. Similarly, AlzKB is entirely open source and publicly available and complies with the licensing terms of all 22 source databases used to build the knowledge base.

Results

Knowledge Base Description

The first release of AlzKB (version 1.0) [26] contains 118,902 distinct nodes (representing biomedical entities) and 1,309,527 relationships linking those nodes. A full summary of node and relationship types with counts, respectively, is provided in Table 2 and Table S1 in Multimedia Appendix 1. Users can interact with AlzKB in their web browser using the built-in Neo4j interface or programmatically by connecting to the graph database over the internet. We also provide instructions for installing a local copy of the graph database as well as how to build the database from its original data sources.

Proposing New Therapeutic Targets for AD

As a proof of concept, we performed an analysis to predict whether known PD genes are also linked to AD etiology. PD is a chronic, progressive neurological disorder characterized by uncontrollable movements and possible mental and behavioral changes. Similar to AD, the precise etiology of PD is not fully understood, but the disease is characterized by the death or dysfunction of basal ganglia neurons. A growing body of work has established physiological and genetic similarities between PD and AD [62], and it has been proposed that drugs targeting PD genes could potentially treat AD as well. To approach this hypothesis computationally, we defined a binary classification task to predict whether gene nodes in the AlzKB knowledge graph are or are not AD genes [63]. To assemble the data set, we considered all gene nodes adjacent to AD as positive (n=101) and all gene nodes not adjacent to AD as negative (n=62,306). The negative samples are assumed to contain a mixture of true negatives and false negatives; in link prediction tasks, the goal is to recover the false negatives. We further filtered the negative nodes to omit PD genes (n=73) and orphan gene nodes (n=43,032) and down sampled the remaining genes to 303 (ie, 3 times the number of positive samples). To evaluate the performance, we used accuracy, balanced accuracy, precision, recall, F_1-score, area under the receiver operating characteristic curve, and area under the precision-recall curve, as shown in Figure 2.

The RF model predicted gene-disease relationships with an average balanced accuracy of 96.2% (precision=0.88; recall=0.98). We applied the trained models to predict PD genes that are likely to also be AD genes. Of the 73 PD genes in AlzKB, 8 (11%; *FYN*, *DCTN1*, *SNCA*, *SYNJ1*, *RSP12*, *ATXN2*, *KCNIP3*, and *CHRNB1*; described in Table 3) were predicted to be AD genes. A total of 10% (7/73) of the genes were predicted to be AD genes in all 10 models, whereas *CHRNB1* was predicted in 7 of the 10 models.

XSL•FO
RenderX

Figure 2. Random forest classifier performance (over 10 independent training runs) on the task of predicting whether Parkinson disease genes are also Alzheimer disease genes based on patterns of graph connectivity in the Alzheimer's Knowledge Base's heterogeneous knowledge graph. Across all metrics, a score of 1.00 represents the best possible performance. AUCPR: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

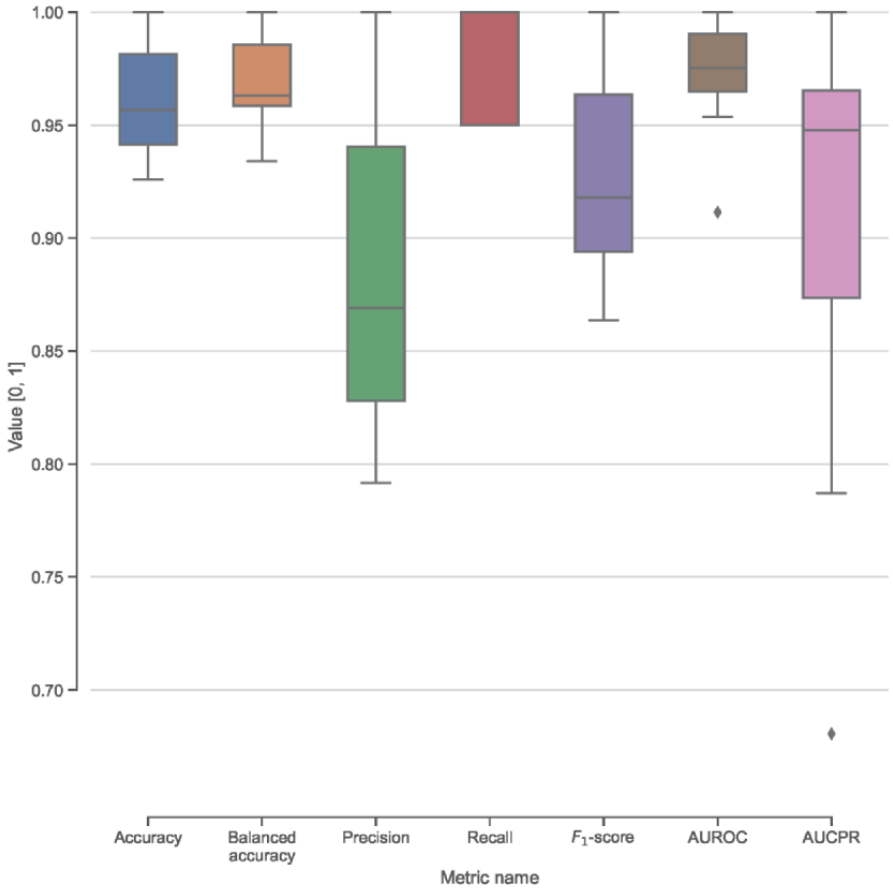

Table 3. Parkinson disease genes predicted by a graph-augmented random forest model to also be associated with Alzheimer disease.

Gene symbol	Gene name	Notes (from the Entrez Gene summary)
ATXN2	Ataxin-2	Modulates endocytosis, ribosomal translation, and mitochondrial function; aberrations are linked to diverse neurodegenerative diseases, diabetes, and obesity
CHRNB1	Cholinergic receptor nicotinic β1 subunit	Beta subunit of muscle acetylcholine receptor involved in transmitting signals at neuromuscular junction
DCTN1	Dynactin subunit 1	Dynactin is a macromolecular complex involved in many cellular functions, including the formation of neuronal pathways
FYN	FYN proto-oncogene, Src tyrosine kinase family	Membrane-associated tyrosine kinase involved in control of cell growth; highly expressed in brain tissue
KCNIP3	Potassium voltage-gated channel interacting protein 3	Voltage-gated potassium channel–interacting protein that is critical to neuronal excitability

Drug Repurposing via Graph Data Science

As a second use case, we considered the task of repurposing existing drugs—currently used to treat other diseases—based on patterns in the knowledge graph that suggest that they may also treat AD. To do this, we trained 5 state-of-the-art knowledge graph completion methods (TransE, RotatE, DistMult, ComplEx, and ConvE) [51] on AlzKB and selected the highest-performing one to predict links between drugs and AD. Additional details about the differences between these methods are provided in Multimedia Appendix 1.

The performance of the 5 different knowledge graph completion models is shown in Table 4. Among them, RotatE performed best, with the highest MRR and hits@k values. Therefore, we used RotateE to make predictions on the test set to obtain missing head entities with the template ([*drug*], DRUG_TREATS_DISEASE, AD). The top 10 predicted drugs are listed in Table 5 along with their current approved use and relevant clinical trial status pertaining to AD efficacy. Of the top 10 predictions, 3 (30%) have been investigated in clinical trials to treat symptoms of AD. To further explore these

XSL•FO
RenderX

predictions, we generated visualizations of a minimum spanning tree linking the 10 drugs to AD in AlzKB's knowledge graph, as shown in Figure 3. The visualization shows that the shortest paths between the drugs and AD are mediated by a small set of AD-associated genes, each of which is associated with one or more of the proposed drugs. The visualization is suggestive of interpretable biological mechanisms through which the drugs could act on AD etiology and provides hypotheses to further explore their validity.

Table 4. Ranking-based evaluation metrics of 5 embedding-based link prediction models on the Alzheimer's Knowledge Base knowledge graph. Metrics are derived from the likelihood of existing (known) links being predicted by the models. Higher scores indicate better performance.

Model name	Hits@1	Hits@3	Hits@10	MRR[a]
RotatE	*0.126* [b]	*0.220*	*0.358*	*0.202*
TransE	0.046	0.097	0.198	0.097
DistMult	0.027	0.056	0.126	0.061
ComplEx	0.074	0.142	0.263	0.136
ConvE	0.002	0.005	0.013	0.006

[a]MRR: mean reciprocal rank.

[b]Italicized values indicate maximum scores within a given column.

Table 5. Drug repurposing predictions made by the best-performing topological link prediction model (RotatE). Also shown are current approved indications and (if available) clinical trials investigating the efficacy of the drug for treating Alzheimer disease (AD).

Drug name	Approved indications	AD-related clinical trials
Sumatriptan	Migraines and cluster headaches	__[a]
Nicotine	Nicotine withdrawal symptoms	NCT00018278 (completed)
Pimozide	Tourette disorder	—
Risperidone	Schizophrenia, bipolar mania, and psychosis	NCT00034762 (completed)
Flurbiprofen	Osteoarthritis and rheumatoid arthritis	—
Sertraline	Depressive disorder and social anxiety disorder	NCT00086138 (completed); NCT00009191 (completed)
Clozapine	Schizophrenia	—
Tamoxifen	ER+[b] breast cancer	—
Amiodarone	Recurrent hemodynamically unstable ventricular tachycardia and recurrent ventricular fibrillation	—
Chlorpromazine	Nausea, vomiting, preoperative anxiety, schizophrenia, bipolar disorder, and severe behavioral problems in children	—

[a]No known AD-related clinical trials for the given drug.

[b]ER+: estrogen-receptor positive.

XSL•FO
RenderX

Figure 3. Spanning tree linking the 10 highest-scoring Alzheimer disease (AD) drug predictions (listed in Table 5) to AD. Blue nodes are drugs, pink nodes are genes, and the orange node is AD. Genes on the shortest path between a drug and AD can be considered putative mechanistic explanations for how the drug may act on AD etiology.

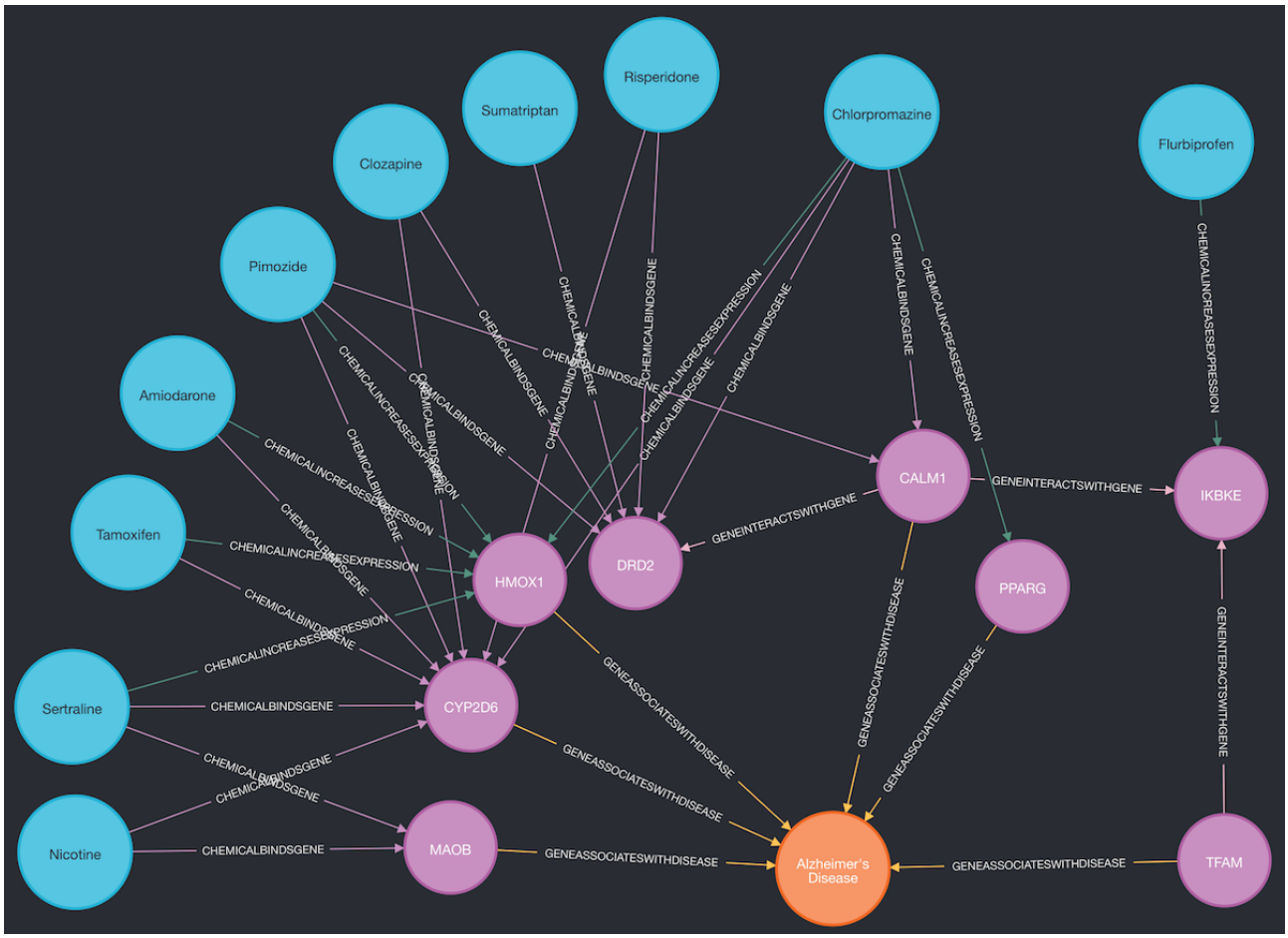

Discussion

Principal Findings

AlzKB is a freely available resource for the biomedical research community, with the primary goal of expanding the repertoire of therapies for AD via drug repurposing. In the previous sections, we described the current contents of AlzKB, the process of constructing it, and 2 specific data-driven use cases that illustrate how it can be applied to drug repurposing tasks. These use cases consisted of predicting the shared genetic architecture of AD and PD (potentially allowing for PD therapies to be repurposed for AD) and directly proposing drugs to repurpose for treating AD by predicting new links between drug and disease nodes in the knowledge graph. In both cases, the results are both biologically plausible and supported by quantitative metrics, yielding new hypotheses that merit experimental validation. AlzKB is a flexible resource that is not limited to these analyses, and we encourage other research teams to use it for different and complementary knowledge discovery tasks.

The Role of AlzKB in Biomedical Knowledge Discovery

AD and other neurodegenerative diseases present one of the greatest challenges in modern biomedicine. AD is by and large a disease of old age, and as improvements to health care continue to increase the overall global life expectancy, we can expect the number of people with various forms of dementia to also increase. As the etiology and pathophysiology of AD are highly multifactorial, there is likely no single "cure" for the disease. Instead, researchers and public health officials have shifted much of their focus toward finding therapies that reduce risk, slow the progression of the disease, or reverse neuronal damage. In addition, as there are various subtypes of AD with underlying mechanisms, any therapy might be effective for only some patients with AD. Therefore, an essential step for reducing global disease burden is to propose many new therapeutic agents that target various aspects of AD pathology. This is precisely the motivating use case for AlzKB. As we have demonstrated, AlzKB provides a rich representation of existing knowledge about AD and the biological context in which it acts. The 2 ML-based use cases we presented previously use real-world end points to demonstrate that the knowledge captured in AlzKB is meaningful and representative of the biological processes underlying the disease. AlzKB stands to become a major resource in the AD research community, where pattern analysis and integration with observational data can be used to propose a diverse array of new therapeutic hypotheses along with interpretable mechanistic explanations of how those therapies may act in the human body.

Building the initial release of AlzKB was a highly interdisciplinary effort involving contributions from experts in translational bioinformatics, data science, and clinical

XSL•FO
RenderX

informatics as well as medical scientists. Although each of these domains was essential in delivering a knowledge base that reflects important biomedical patterns describing AD etiology and treatment, a key need during the design and implementation phases was data literacy. To support future work in this and related areas, we encourage the inclusion of informatics and data analysis techniques in all types of biomedical curricula. Beyond AlzKB, our approach for building the knowledge graph is generalizable to practically any domain and depends on (1) defining an ontology using expert knowledge that formally describes the domain of interest and (2) identifying source databases that provide the entities and relationships described in the ontology. We are directly involved in the ongoing development of other knowledge bases using this same approach, including ComptoxAI—a knowledge base that supports AI research in toxicology [64]. As both knowledge bases share many of the same "core" entities (genes, diseases, pathways, and anatomical structures), the knowledge graphs are already semantically harmonized and ready for integration in larger, cross-disciplinary biomedical knowledge applications.

Discovering Putative Therapies Through Graph Data Science

Of the PD genes predicted to also be AD genes (see the *Proposing New Therapeutic Targets for AD* section; Table 3), some are involved in neuronal signaling and structure, and some are known to be involved in a wide range of neurological disorders. *FYN* has seen recent attention and investigation into its possible link to AD due to its broad expression in brain tissue and known interactions with tau proteins [65,66]. Among the other identified genes, one (*CHRNB1*) is known to be involved in acetylcholine signaling [67,68], and another (*KCNIP3*) codes a protein that interacts with presenilin, and mutations in presenilin are causal for hereditary AD [69,70]. Some of these gene hits (*ATXN2* and *DCTN1*) have limited or no current research directly linking them to AD but are biologically plausible. As such, they may represent novel therapeutic targets or targets for further research and investigation [71]. For example, *DCTN1* encodes the dynactin-1 protein, and deficits in dynactin are connected to several neurodegenerative diseases; however, there is limited research linking this gene to AD [72,73].

Among the drug repurposing predictions (see the *Drug Repurposing via Graph Data Science* section; Table 5) are some agents that have previously been proposed for the treatment of AD (risperidone and sertraline) or for symptoms associated with AD (nicotine). Sumatriptan has been the subject of several studies focused on AD [74] and is connected to a strong comorbidity of migraine headaches and dementia in women [75]. Pimozide has been shown to reduce the aggregation of tau protein in mice [76] and is linked to AD in a number of unrelated *in silico* models [77]. The inclusion of nicotine is also noteworthy as it has seen recent interest among AD researchers and is the subject of an ongoing clinical trial to improve memory [78]. Other drugs listed in Table 5 have not yet been identified as AD treatments and represent novel repurposing candidates. Each can be considered a testable hypothesis meriting further investigation, giving credence to the increased detective power of AlzKB's knowledge graph approach over existing AD data resources. It should be noted that this approach can only propose new indications for existing drugs and is based on existing knowledge and derived from known biological associations with those drugs. Other approaches (including emerging techniques in graph ML) could be used to propose entirely new drugs that could treat AD.

Future Directions With AlzKB

AlzKB is a growing resource, and we have plans for adding new features and data types that are in various stages of implementation. As a central hypothesis of AD pathogenesis revolves around the atypical accumulation of proteins within and around brain cells, an important step will be to adequately distinguish and differentiate genes from the proteins that those genes code for. Existing data resources available for inclusion in AlzKB largely fail to make this distinction in a way that is accepted by the scientific community, so we are currently evaluating options to use either postprocessing of existing knowledge sources or synthesis of new knowledge to achieve a good representation of genes, proteins, and functional or structural variants that are key to understanding AD.

Current ML models often do not generalize well to heterogeneous graphs such as the one that constitutes AlzKB's knowledge graph. This is largely because traditional models cannot use the network structure and heterogeneous nature of different entity types. Several promising algorithms can be used for prediction on heterogeneous graphs—including GraphSAGE [79] and metapath2vec [80]—but most fail to scale effectively when the number of node or edge types increases. As any effective therapy must be accompanied by a mechanistic understanding of how it functions, we also need to ensure that new heterogeneous graph ML models are *explainable*. With this in mind, we are using AlzKB as a motivating resource for designing new, cutting-edge algorithms that produce interpretable predictions from highly heterogeneous knowledge graphs. Furthermore, the increasing popularity of large language models (LLMs; such as GPT-4) presents a wealth of opportunities for incorporating knowledge graphs such as AlzKB into diverse AI applications [81]. One application we are considering is using AlzKB to provide LLMs with formalized knowledge about AD that allows them to more effectively produce informative outputs about AD etiology. Currently, LLMs can perform poorly on technically complex or poorly understood domains due to a scarcity of relevant content in their training corpora, and augmenting their performance using domain-specific knowledge graphs is an emerging strategy for fixing that issue. As we do so, these will be released alongside AlzKB with educational resources that facilitate ease of use and adoptability by various stakeholders.

Knowledge graphs—including AlzKB—come with several important limitations that will be crucial to address in coming years. One of these is the subjective nature of determining what does and does not constitute "knowledge," implying broad acceptance by the scientific community (as opposed to "data," which consist of individual observations). Currently, we use expert domain knowledge and careful screening of source databases to accomplish this, but with the advent of broadly accessible generative AI tools, there may be emerging strategies

XSL•FO
RenderX

that minimize sources of human bias [82]. Furthermore, new predictions made using knowledge graphs still necessitate costly and time-consuming experimental or observational follow-up studies to validate those predictions. This is due in part to the absence of negative samples for training predictive models. While the presence of an edge between 2 nodes in a knowledge graph is interpreted as a "positive sample" for model training, the absence of an edge simply means that we do not know whether a relationship does or does not exist, and therefore, it may not in fact be a negative sample. New methods, including self-supervised contrastive learning, show promise in alleviating this issue [83], but further work is needed to determine whether these generalize well to AlzKB and similar highly heterogeneous biomedical knowledge graphs. Nonetheless, these are active areas of research in the AI, informatics, and computer science communities, and in spite of them, our results are still robust enough to provide compelling evidence demonstrating AlzKB's scientific value.

Ultimately, we aim to provide AlzKB as a robust resource that helps unravel the etiology of AD. It is already a large, high-quality knowledge base from which graph-based AI or ML approaches can be developed for drug repurposing and drug discovery. As we and the rest of the biomedical research community make these discoveries in the coming years, they will be included and publicized on the AlzKB website as a public resource to drive innovation and scientific progress.

Obtaining AlzKB for Local Use and Extending the Knowledge Graph

As it is a public and open-source resource for scientific discovery, we provide AlzKB through a variety of interfaces with distinct advantages for different use cases and user types. Casual users who wish to browse the knowledge base or perform simple analyses can do so directly through the Neo4j browser interface [24]. However, for more advanced use cases (or when computational needs exceed those available on the public version of the knowledge base), AlzKB can be either downloaded and populated locally into a Neo4j installation or built from the original source data files via the tools included on the AlzKB GitHub repository [25]. The latter of these options also allows users to extend the knowledge base to include additional data sources, entity types, or relationships beyond those provided in the official knowledge base distribution. We also encourage users who make modifications to the knowledge base to submit their changes for review to be included in the main code distribution. Instructions for how to contribute to AlzKB are also available on the GitHub repository.

As the data sources included in AlzKB are all, themselves, from open-source databases, we urge users to ensure that any new data sources they merge into AlzKB similarly comply with open-source standards. In brief, AlzKB can only be maintained under the most restrictive license terms of its included third-party sources, so restrictive license terms in a database being considered decrease that database's suitability for inclusion. We hope for AlzKB to be recognized as a community effort for aggregating and democratizing the discovery of new AD therapeutics and, therefore, encourage public discussion of new methods and data sources to be included.

Conclusions

In this work, we introduced AlzKB as a free, publicly available toolkit and data resource for novel discoveries in AD research, with a particular focus on therapeutic approaches to treating AD. AlzKB is both new and continually growing, and we aim to cultivate a community of researchers to collaboratively increase the impact, speed, and throughput of AD research, along with rapid dissemination to health care, academia, and the pharmaceutical industry. In the future, we will develop new AI and data science methods to continually extract knowledge from AlzKB, but in this study, we already demonstrate through graph data science that AlzKB can both replicate existing AD knowledge and generate entirely new, testable hypotheses to drive the future of drug repurposing and drug discovery.

Acknowledgments

The Alzheimer's Knowledge Base is supported by US National Institutes of Health grants U01-AG066833, R01-LM010098, R01-LM013463 (principal investigator [PI]: JHM), and R00-LM013646 (PI: JDR).

Data Availability

The data sets generated during and analyzed during this study are available in the GitHub and Zenodo repositories [25,26].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental information providing expanded details on the knowledge graph completion methods used to validate Alzheimer's Knowledge Base, as well as counts for relationship types in the knowledge graph.
[DOCX File , 23 KB - jmir_v26i1e46777_app1.docx]

References

XSL•FO
RenderX

1. 2022 Alzheimer's disease facts and figures. Alzheimers Dement 2022 Apr;18(4):700-789. [doi: 10.1002/alz.12638] [Medline: 35289055]
2. Yiannopoulou KG, Papageorgiou SG. Current and future treatments in Alzheimer disease: an update. J Cent Nerv Syst Dis 2020 Feb 29;12:1179573520907397 [FREE Full text] [doi: 10.1177/1179573520907397] [Medline: 32165850]
3. Rabinovici GD. Controversy and progress in Alzheimer's disease - FDA approval of aducanumab. N Engl J Med 2021 Aug 26;385(9):771-774. [doi: 10.1056/NEJMp2111320] [Medline: 34320284]
4. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. Mol Neurodegener 2019 Aug 02;14(1):32 [FREE Full text] [doi: 10.1186/s13024-019-0333-5] [Medline: 31375134]
5. Aisen PS, Cummings J, Jack CRJ, Morris JC, Sperling R, Frölich L, et al. On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimers Res Ther 2017 Aug 09;9(1):60 [FREE Full text] [doi: 10.1186/s13195-017-0283-5] [Medline: 28793924]
6. Fan L, Mao C, Hu X, Zhang S, Yang Z, Hu Z, et al. New insights into the pathogenesis of Alzheimer's disease. Front Neurol 2019;10:1312 [FREE Full text] [doi: 10.3389/fneur.2019.01312] [Medline: 31998208]
7. Silva MV, de Mello Gomide Loures C, Alves LC, de Souza LC, Borges KB, Carvalho MD. Alzheimer's disease: risk factors and potentially protective measures. J Biomed Sci 2019 May 09;26(1):33 [FREE Full text] [doi: 10.1186/s12929-019-0524-y] [Medline: 31072403]
8. Rodriguez S, Hug C, Todorov P, Moret N, Boswell SA, Evans K, et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nat Commun 2021 Feb 15;12(1):1033 [FREE Full text] [doi: 10.1038/s41467-021-21330-0] [Medline: 33589615]
9. Tsuji S, Hase T, Yachie-Kinoshita A, Nishino T, Ghosh S, Kikuchi M, et al. Artificial intelligence-based computational framework for drug-target prioritization and inference of novel repositionable drugs for Alzheimer's disease. Alzheimers Res Ther 2021 May 03;13(1):92 [FREE Full text] [doi: 10.1186/s13195-021-00826-3] [Medline: 33941241]
10. Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet 2007 Apr 15;16(8):865-873. [doi: 10.1093/hmg/ddm031] [Medline: 17317784]
11. The Alzheimer's KnowledgeBase (AlzKB). AlzKB. URL: https://alzkb.ai/ [accessed 2023-02-24]
12. Daluwatumulle G, Wijesinghe R, Weerasinghe R. In silico drug repurposing using knowledge graph embeddings for Alzheimer's disease. In: Proceedings of the 9th International Conference on Bioinformatics Research and Applications. 2022 Presented at: ICBRA '22; September 18-20, 2022; Berlin, Germany. [doi: 10.1145/3569192.3569203]
13. Nian Y, Hu X, Zhang R, Feng J, Du J, Li F, et al. Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing. BMC Bioinformatics 2022 Sep 30;23(Suppl 6):407 [FREE Full text] [doi: 10.1186/s12859-022-04934-1] [Medline: 36180861]
14. Hsieh KL, Plascencia-Villa G, Lin KH, Perry G, Jiang X, Kim Y. Synthesize heterogeneous biological knowledge via representation learning for Alzheimer's disease drug repurposing. iScience 2022 Nov 26;26(1):105678 [FREE Full text] [doi: 10.1016/j.isci.2022.105678] [Medline: 36594024]
15. Binder J, Ursu O, Bologa C, Jiang S, Maphis N, Dadras S, et al. Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Commun Biol 2022 Feb 11;5(1):125 [FREE Full text] [doi: 10.1038/s42003-022-03068-7] [Medline: 35149761]
16. Sügis E, Dauvillier J, Leontjeva A, Adler P, Hindie V, Moncion T, et al. HENA, heterogeneous network-based data set for Alzheimer's disease. Sci Data 2019 Aug 14;6(1):151. [doi: 10.1038/s41597-019-0152-0] [Medline: 31413325]
17. Robinson I, Webber J, Eifrem E. Graph Databases: New Opportunities for Connected Data. Sebastopol, CA: O'Reilly Media; 2015.
18. Davis R, Shrobe H, Szolovits P. What is a knowledge representation? AI Mag 1993;14(1):17. [doi: 10.1609/aimag.v14i1.1029]
19. Musen MA, Protégé Team. The Protégé project: a look back and a look forward. AI Matters 2015 Jun;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
20. Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S. OWL 2 Web ontology language primer. World Wide Web Consortium. 2009 Apr 21. URL: https://www.w3.org/TR/2009/WD-owl2-primer-20090421/ [accessed 2024-03-25]
21. Tsarkov D, Horrocks I. FaCT++ description logic reasoner: system description. In: Proceedings of the International Joint Conference on Automated Reasoning. 2006 Presented at: IJCAR 2006; August 17-20, 2006; Seattle, WA. [doi: 10.1007/11814771_26]
22. Neo4j. URL: https://neo4j.com/ [accessed 2022-10-25]
23. Barrasa J, Cowley A. neosemantics (n10s): Neo4j RDF and semantics toolkit. Neo4j. URL: https://neo4j.com/labs/neosemantics/ [accessed 2022-10-25]
24. Neo4j browser. Neo4j. URL: http://neo4j.alzkb.ai/browser/ [accessed 2023-02-24]
25. EpistasisLab/AlzKB. GitHub. URL: https://github.com/EpistasisLab/AlzKB [accessed 2023-02-24]
26. Romano J, Wang P. EpistasisLab/AlzKB: AlzKB first DOI release. Zenodo. 2022 Aug 22. URL: https://zenodo.org/records/7015728 [accessed 2024-03-27]
27. Mortensen HM, Senn J, Levey T, Langley P, Williams AJ. The 2021 update of the EPA's adverse outcome pathway database. Sci Data 2021 Jul 12;8(1):169 [FREE Full text] [doi: 10.1038/s41597-021-00962-3] [Medline: 34253739]

XSL•FO
RenderX

28. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: integrating and comparing heterogeneous transcriptome data among species. In: Proceedings of the Data Integration in the Life Sciences. 2008 Presented at: DILS 2008; June 25-27, 2008; Evry, France. [doi: 10.1007/978-3-540-69828-9_12]
29. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 2019 Jan 08;47(D1):D955-D962 [FREE Full text] [doi: 10.1093/nar/gky1032] [Medline: 30407550]
30. Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford) 2015;2015:bav028 [FREE Full text] [doi: 10.1093/database/bav028] [Medline: 25877637]
31. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008 Jan;36(Database issue):D901-D906 [FREE Full text] [doi: 10.1093/nar/gkm958] [Medline: 18048412]
32. Grulke CM, Williams AJ, Thillanadarajah I, Richard AM. EPA's DSSTox database: history of development of a curated chemistry resource supporting computational toxicology research. Comput Toxicol 2019 Nov 01;12:100096 [FREE Full text] [doi: 10.1016/j.comtox.2019.100096] [Medline: 33426407]
33. Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, et al. ACToR--Aggregated computational toxicology resource. Toxicol Appl Pharmacol 2008 Nov 15;233(1):7-13. [doi: 10.1016/j.taap.2007.12.037] [Medline: 18671997]
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000 May;25(1):25-29 [FREE Full text] [doi: 10.1038/75556] [Medline: 10802651]
35. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021 Jan 08;49(D1):D325-D334 [FREE Full text] [doi: 10.1093/nar/gkaa1113] [Medline: 33290552]
36. Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019 Jan 08;47(D1):D1005-D1012 [FREE Full text] [doi: 10.1093/nar/gky1120] [Medline: 30445434]
37. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 2017 Sep 22;6:e26726 [FREE Full text] [doi: 10.7554/eLife.26726] [Medline: 28936969]
38. The human reference protein interactome mapping project. The Human Reference Interactome. URL: http://www.interactome-atlas.org/ [accessed 2023-02-24]
39. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl 2016 Aug 04;2(1):16015 [FREE Full text] [doi: 10.1038/npjsba.2016.15] [Medline: 28413689]
40. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265-266 [FREE Full text] [Medline: 10928714]
41. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2011 Jan;39(Database issue):D52-D57 [FREE Full text] [doi: 10.1093/nar/gkq1237] [Medline: 21115458]
42. Schaefer C, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nat Prec 2008 Aug 29. [doi: 10.1038/npre.2008.2243.1]
43. Himmelstein D, Pouya K, Hessler CS, Green AJ, Baranzini S. PharmacotherapyDB 1.0: the open catalog of drug therapies for disease. Figshare. 2016. URL: https://tinyurl.com/mv8k46em [accessed 2024-03-25]
44. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021 Jan 08;49(D1):D1388-D1395 [FREE Full text] [doi: 10.1093/nar/gkaa971] [Medline: 33151290]
45. Wu G, Haw R. Functional interaction network construction and analysis for disease discovery. Methods Mol Biol 2017;1558:235-253. [doi: 10.1007/978-1-4939-6783-4_11] [Medline: 28150241]
46. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016 Jan 04;44(D1):D1075-D1079 [FREE Full text] [doi: 10.1093/nar/gkv1075] [Medline: 26481350]
47. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ. TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database (Oxford) 2018 Jan 01;2018:2 [FREE Full text] [doi: 10.1093/database/bay028] [Medline: 30403794]
48. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol 2012 Jan 31;13(1):R5 [FREE Full text] [doi: 10.1186/gb-2012-13-1-r5] [Medline: 22293552]
49. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers KA, et al. WikiPathways: connecting communities. Nucleic Acids Res 2021 Jan 08;49(D1):D613-D621 [FREE Full text] [doi: 10.1093/nar/gkaa1024] [Medline: 33211851]
50. Ali M, Berrendorf M, Hoyt CT, Vermue L, Sharifzadeh S, Tresp V, et al. PyKEEN 1.0: a python library for training and evaluating knowledge graph embeddings. J Mach Learn Res 2021;22(82):1-6 [FREE Full text]
51. Zamini M, Reza H, Rabiei M. A review of knowledge graph completion. Information 2022 Aug 21;13(8):396. [doi: 10.3390/info13080396]

XSL•FO
RenderX

52. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, Nevada.
53. Sun Z, Deng ZH, Nie JY, Tang J. RotatE: knowledge graph embedding by relational rotation in complex space. arXiv Preprint posted online February 26, 2019 [FREE Full text]
54. Yang B, Yih WT, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. arXiv Preprint posted online December 20, 2014 [FREE Full text]
55. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. arXiv Preprint posted online June 20, 2016 [FREE Full text]
56. Dettmers T, Minervini P, Stenetorp P, Riedel S. Convolutional 2D knowledge graph embeddings. arXiv Preprint posted online July 5, 2017 [FREE Full text] [doi: 10.1609/aaai.v32i1.11573]
57. Newman ME. Clustering and preferential attachment in growing networks. Phys Rev E 2001 Jul 26;64(2):025102. [doi: 10.1103/physreve.64.025102]
58. Barabasi AL, Albert R. Emergence of scaling in random networks. Science 1999 Oct 15;286(5439):509-512. [doi: 10.1126/science.286.5439.509] [Medline: 10521342]
59. Adamic LA, Adar E. Friends and neighbors on the web. Soc Netw 2003 Jul;25(3):211-230. [doi: 10.1016/s0378-8733(03)00009-1]
60. Zhou T, Lü L, Zhang YC. Predicting missing links via local information. Eur Phys J B 2009 Oct 10;71(4):623-630. [doi: 10.1140/epjb/e2009-00335-8]
61. Gao Z, Ding P, Xu R. KG-Predict: a knowledge graph computational framework for drug repurposing. J Biomed Inform 2022 Aug;132:104133 [FREE Full text] [doi: 10.1016/j.jbi.2022.104133] [Medline: 35840060]
62. Nussbaum RL, Ellis CE. Alzheimer's disease and Parkinson's disease. N Engl J Med 2003 Apr 03;348(14):1356-1364. [doi: 10.1056/NEJM2003ra020003] [Medline: 12672864]
63. Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, et al. Application of network link prediction in drug discovery. BMC Bioinformatics 2021 Apr 12;22(1):187 [FREE Full text] [doi: 10.1186/s12859-021-04082-y] [Medline: 33845763]
64. Romano JD, Hao Y, Moore JH, Penning TM. Automating predictive toxicology using ComptoxAI. Chem Res Toxicol 2022 Aug 15;35(8):1370-1382 [FREE Full text] [doi: 10.1021/acs.chemrestox.2c00074] [Medline: 35819939]
65. Iannuzzi F, Sirabella R, Canu N, Maier TJ, Annunziato L, Matrone C. Fyn tyrosine kinase elicits amyloid precursor protein Tyr682 phosphorylation in neurons from Alzheimer's disease patients. Cells 2020 Jul 30;9(8):1807 [FREE Full text] [doi: 10.3390/cells9081807] [Medline: 32751526]
66. Nygaard HB, van Dyck CH, Strittmatter SM. Fyn kinase inhibition as a novel therapy for Alzheimer's disease. Alzheimers Res Ther 2014 Feb 5;6(1):8 [FREE Full text] [doi: 10.1186/alzrt238] [Medline: 24495408]
67. Lardenoije R, Roubroeks JA, Pishva E, Leber M, Wagner H, Iatrou A, et al. Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood. Clin Epigenetics 2019 Nov 27;11(1):164 [FREE Full text] [doi: 10.1186/s13148-019-0755-5] [Medline: 31775875]
68. Lombardo S, Maskos U. Role of the nicotinic acetylcholine receptor in Alzheimer's disease pathology and treatment. Neuropharmacology 2015 Sep;96(Pt B):255-262 [FREE Full text] [doi: 10.1016/j.neuropharm.2014.11.018] [Medline: 25514383]
69. Jo DG, Lee JY, Hong YM, Song S, Mook-Jung I, Koh JY, et al. Induction of pro-apoptotic calsenilin/DREAM/KChIP3 in Alzheimer's disease and cultured neurons after amyloid-beta exposure. J Neurochem 2004 Feb;88(3):604-611 [FREE Full text] [doi: 10.1111/j.1471-4159.2004.02159.x] [Medline: 14720210]
70. Jin JK, Choi JK, Wasco W, Buxbaum JD, Kozlowski PB, Carp RI, et al. Expression of calsenilin in neurons and astrocytes in the Alzheimer's disease brain. Neuroreport 2005 Apr 04;16(5):451-455. [doi: 10.1097/00001756-200504040-00007] [Medline: 15770150]
71. Rosas I, Martínez C, Clarimón J, Lleó A, Illán-Gala I, Dols-Icardo O, et al. Role for ATXN1, ATXN2, and HTT intermediate repeats in frontotemporal dementia and Alzheimer's disease. Neurobiol Aging 2020 Mar;87:139.e1-139.e7. [doi: 10.1016/j.neurobiolaging.2019.10.017] [Medline: 31810584]
72. Aboud O, Parcon PA, DeWall KM, Liu L, Mrak RE, Griffin WS. Aging, Alzheimer's, and APOE genotype influence the expression and neuronal distribution patterns of microtubule motor protein dynactin-P50. Front Cell Neurosci 2015 Mar 25;9:103 [FREE Full text] [doi: 10.3389/fncel.2015.00103] [Medline: 25859183]
73. Caroppo P, Le Ber I, Clot F, Rivaud-Péchoux S, Camuzat A, De Septenville A, et al. DCTN1 mutation analysis in families with progressive supranuclear palsy-like phenotypes. JAMA Neurol 2014 Feb;71(2):208-215 [FREE Full text] [doi: 10.1001/jamaneurol.2013.5100] [Medline: 24343258]
74. Zochodne DW, Ho LT. Sumatriptan blocks neurogenic inflammation in the peripheral nerve trunk. Neurology 1994 Jan;44(1):161-163. [doi: 10.1212/wnl.44.1.161] [Medline: 8290056]
75. Liu CT, Wu BY, Hung YC, Wang LY, Lee YY, Lin TK, et al. Decreased risk of dementia in migraine patients with traditional Chinese medicine use: a population-based cohort study. Oncotarget 2017 Oct 03;8(45):79680-79692 [FREE Full text] [doi: 10.18632/oncotarget.19094] [Medline: 29108348]

XSL•FO
RenderX

76. Kim YD, Jeong EI, Nah J, Yoo SM, Lee WJ, Kim Y, et al. Pimozide reduces toxic forms of tau in TauC3 mice via 5' adenosine monophosphate-activated protein kinase-mediated autophagy. J Neurochem 2017 Sep 11;142(5):734-746 [FREE Full text] [doi: 10.1111/jnc.14109] [Medline: 28632947]
77. Kumar S, Chowdhury S, Kumar S. In silico repurposing of antipsychotic drugs for Alzheimer's disease. BMC Neurosci 2017 Oct 27;18(1):76 [FREE Full text] [doi: 10.1186/s12868-017-0394-8] [Medline: 29078760]
78. van Duijn CM, Hofman A. Relation between nicotine intake and Alzheimer's disease. BMJ 1991 Jun 22;302(6791):1491-1494 [FREE Full text] [doi: 10.1136/bmj.302.6791.1491] [Medline: 1855016]
79. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. arXiv Preprint posted online June 7, 2017 [FREE Full text]
80. Dong Y, Chawla NV, Swami A. metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 Presented at: KDD '17; August 13-17, 2017; Halifax, NS URL: https://dl.acm.org/doi/10.1145/3097983.3098036
81. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. arXiv Preprint posted online June 14, 2023 [FREE Full text] [doi: 10.48550/arXiv.2306.08302]
82. Zhu Y, Wang X, Chen J, Qiao S, Ou Y, Yao Y, et al. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. arXiv Preprint posted online May 22, 2023 [FREE Full text]
83. Kefato ZT, Girdzijauskas S. Self-supervised Graph Neural Networks without explicit negative sampling. arXiv Preprint posted online March 27, 2021 [FREE Full text]

Abbreviations

AD: Alzheimer disease
AI: artificial intelligence
AlzKB: Alzheimer's Knowledge Base
LLM: large language model
ML: machine learning
MRR: mean reciprocal rank
OWL: Web Ontology Language
PD: Parkinson disease
RDF: Resource Description Framework
RF: random forest

Edited by T de Azevedo Cardoso; submitted 24.02.23; peer-reviewed by P Dabas, N Mungoli, B Xie, C Sun; comments to author 21.04.23; revised version received 23.06.23; accepted 07.11.23; published 18.04.24.

Please cite as:
Romano JD, Truong V, Kumar R, Venkatesan M, Graham BE, Hao Y, Matsumoto N, Li X, Wang Z, Ritchie MD, Shen L, Moore JH
The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research
J Med Internet Res 2024;26:e46777
URL: https://www.jmir.org/2024/1/e46777
doi:10.2196/46777
PMID:38635981

©Joseph D Romano, Van Truong, Rachit Kumar, Mythreye Venkatesan, Britney E Graham, Yun Hao, Nick Matsumoto, Xi Li, Zhiping Wang, Marylyn D Ritchie, Li Shen, Jason H Moore. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Searching COVID-19 Clinical Research Using Graph Queries: Algorithm Development and Validation

Francesco Invernici[1], MSc; Anna Bernasconi[1], PhD; Stefano Ceri[1], PhD

Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy

Corresponding Author:
Anna Bernasconi, PhD
Department of Electronics, Information, and Bioengineering
Politecnico di Milano
Via Ponzio 34/5
Milan, 20133
Italy
Phone: 39 23993494
Fax: 39 23993411
Email: anna.bernasconi@polimi.it

Abstract

Background: Since the beginning of the COVID-19 pandemic, >1 million studies have been collected within the COVID-19 Open Research Dataset, a corpus of manuscripts created to accelerate research against the disease. Their related abstracts hold a wealth of information that remains largely unexplored and difficult to search due to its unstructured nature. Keyword-based search is the standard approach, which allows users to retrieve the documents of a corpus that contain (all or some of) the words in a target list. This type of search, however, does not provide visual support to the task and is not suited to expressing complex queries or compensating for missing specifications.

Objective: This study aims to consider small graphs of concepts and exploit them for expressing graph searches over existing COVID-19–related literature, leveraging the increasing use of graphs to represent and query scientific knowledge and providing a user-friendly search and exploration experience.

Methods: We considered the COVID-19 Open Research Dataset corpus and summarized its content by annotating the publications' abstracts using terms selected from the Unified Medical Language System and the Ontology of Coronavirus Infectious Disease. Then, we built a co-occurrence network that includes all relevant concepts mentioned in the corpus, establishing connections when their mutual information is relevant. A sophisticated graph query engine was built to allow the identification of the best matches of graph queries on the network. It also supports partial matches and suggests potential query completions using shortest paths.

Results: We built a large co-occurrence network, consisting of 128,249 entities and 47,198,965 relationships; the GRAPH-SEARCH interface allows users to explore the network by formulating or adapting graph queries; it produces a bibliography of publications, which are globally ranked; and each publication is further associated with the specific parts of the query that it explains, thereby allowing the user to understand each aspect of the matching.

Conclusions: Our approach supports the process of query formulation and evidence search upon a large text corpus; it can be reapplied to any scientific domain where documents corpora and curated ontologies are made available.

(J Med Internet Res 2024;26:e52655) doi:10.2196/52655

KEYWORDS

big data corpus; clinical research; co-occurrence network; COVID-19 Open Research Dataset; CORD-19; graph search; Named Entity Recognition; Neo4j; text mining

Introduction

Since the COVID-19 pandemic outbreak in early 2020, important clinical research efforts have been targeted at understanding the COVID-19 disease. More than 1 million studies have been collected within the COVID-19 Open Research Dataset (CORD-19), a corpus of manuscripts created to accelerate the research against the disease. Their related

XSL•FO
RenderX

abstracts hold a wealth of information that remains largely unexplored and difficult to search due to its unstructured nature.

Searching over the literature is a nontrivial task, as it strongly relies on the quality of the data corpus, the characteristics of the search portal, and the language used to express the search.

Keyword-based search is the standard search approach, which allows users to retrieve the documents of a corpus that contain some of the words in a specified target list [1,2]. However, this type of search lacks visual support for the task and is not suitable for expressing complex research queries or compensating for missing specifications.

The development of frontend tools and visualizations for COVID-19 knowledge graphs has been motivated by several works [3,4]. We then explored the use of small graph-based queries that can be built visually [5] to empower a literature exploration tool; the GRAPH-SEARCH system stems from this motivation, providing both a visual language to express search queries and a friendly tool to explore relevant publications, which highlights the relationships between the original graph queries and an underlying corpus of scientific evidence, in the spirit of literature-based discovery [6].

To support this idea, the underlying textual corpus must first be analyzed and enriched; in our approach, the CORD-19 was expressed in the form of a co-occurrence network. First, we annotated all the abstracts with terms from the Unified Medical Language System (UMLS) [7] and the Coronavirus Infectious Disease Ontology (CIDO) [8]. This step closely aligns with classical work on ontology-based annotation (refer to Semantic MEDLINE [9] and our previous study on genomic metadata annotation [10,11]). Second, we built a comprehensive co-occurrence network that includes all relevant clinical and biological concepts mentioned in the corpus, linking them based on their co-occurrence in given abstracts.

The visual language used to express a query over the network describes concepts as nodes and their copresence within research abstracts as undirected edges; some concepts are associated with medical conditions, whereas others are associated with treatments or biological entities. We also allow modifiers. Queries run on the network may correspond to the expressed graph pattern or to a selected subpart.

The query semantics corresponds to extracting scientific evidence (ie, publications) from the corpus, in support of the existence of the relationships linking the expressed concepts; each search process extracts the references that best explain the relationships occurring within the query. When a specified path is not present in the co-occurrence network, alternative scored and ranked shortest paths connecting the nodes expressed in the query are proposed to the user (refer to the *Methods* section). The search output provides a ranking of references because of their weight, summing up the support that they provide to several relationships in the query.

Our GRAPH-SEARCH implementation is supported by a graphical interface (refer to the *Data Availability* section) that allows the user to express the queries and to interpret the results in terms of concepts explained by each discovered reference, thus enabling the users to better qualify the query during the interaction; in addition, users can read the textual abstracts of the retrieved references. Such interactive exploration of the search space allows for exploring assumptions and for progressively adapting them as a result of existing evidence.

The manuscript is organized as follows: we first describe the CORD-19, the characteristics of the co-occurrence network representing CORD-19 abstracts, the technological process of building the network, the graph search operation, and the web user interface that allows us to express graph queries and explore the retrieved results. We then present a series of example use case (UC) queries relevant to COVID-19 research and review the current state of the art. We then evaluate the benefits of using our GRAPH-SEARCH as opposed to full-text indexed databases and keyword search. Finally, we draw our conclusions.

Methods

The CORD-19 Corpus

CORD-19 [12] is a corpus of academic publications about COVID-19 and related coronavirus research; it was released and maintained by the Allen Institute for AI in collaboration with The White House Office of Science and Technology Policy and other partners. Published articles and preprints were collected from several archives, including PubMed, PubMedCentral, bioRxiv, and arXiv; since its release, it has served as the basis of many COVID-19 text mining and discovery systems [12]. The final release of June 2, 2022, indexes >1 million publications. As summarized in Figure 1, approximately 79% of the documents in CORD-19 have an abstract. Out of them, around 41% have a full-text JSON file available, while <11% of available full-text publications have no abstract in the metadata table. Thus, we decided to focus on data set records with an abstract. The file containing the metadata of publications in the data set is a comma-separated table (CORD-19 metadata.csv) including the following:

XSL•FO
RenderX

Figure 1. Euler-Venn diagram of the overlap of publications with abstract and publications with full-text JSON from PDF or from PubMedCentral (PMC) in the COVID-19 Open Research Dataset (CORD-19).

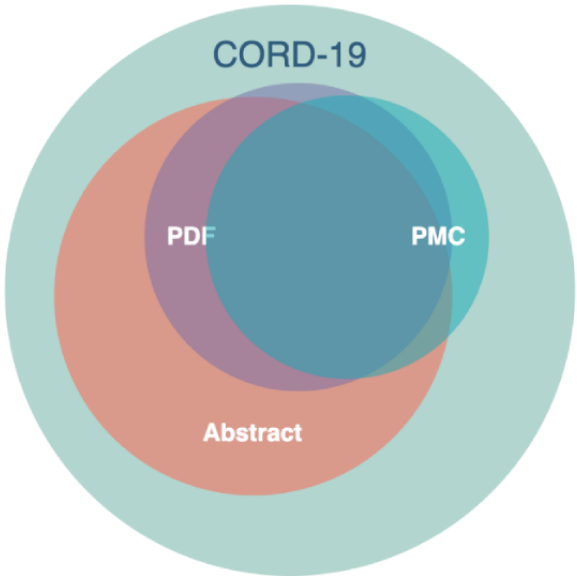

- A unique identifier cord_uid for a cluster of different records of the same publication—upon it, we performed deduplication and subsequent reconciliation of the other metadata of the cluster into a single record.
- Title of the publication—we detected the language and filtered out those not in English.
- Abstract of the publication—only records with an actual abstract were retained.
- Publish_time—the distribution of publication times, shown in Figure 2, shows that COVID-19 publications increased in the first half of 2020. Spikes at the beginning of each year correspond to publications whose publish time is incomplete (ie, only the year field was filled). Publications before 2020 that are concerned with Middle East Respiratory Syndrome, Severe Acute Respiratory Syndrome, and the coronavirus were removed.
- Journal's abbreviated name—fuzzy matching of the abbreviated names was performed with a list of full names obtained from Scopus [13].
- Authors and DOI of the publication
- Number of citations received (ie, through numCitedBy), obtained by SemanticScholar application programming interfaces (APIs) [14]

Records from CORD-19 are already harmonized (refer to the study by Wang et al [12]), resulting in distinct cord_uid keys. However, several records of the same publication are included, with different metadata. We deduplicated them and retained just 1 record (the one published in a peer-reviewed journal, if available, else the richest one in metadata).

Figure 2. Line plot showing the 10-base logarithm of the number of publications (y-axis) per publish time and date (x-axis).

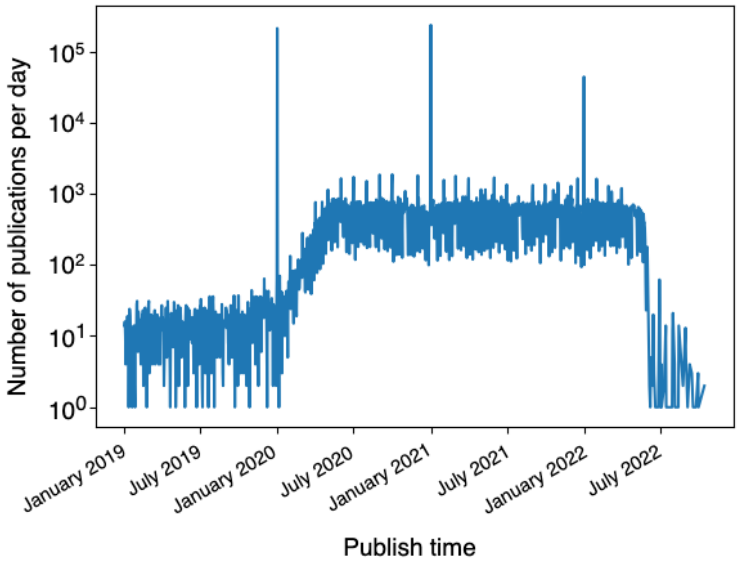

XSL•FO
RenderX

Co-Occurrence Network

The co-occurrence network was built to support graph search; it consists of entities and relationships mined from the title and abstract fields of the metadata table. For building it, we considered 2 sources: UMLS and CIDO. UMLS [7] is a generic source that includes many vocabularies and covers the entire spectrum of medicine; CIDO [8] is a community-driven open-source biomedical ontology in the area of COVID-19.

While CIDO has a simple concept structure, UMLS concepts have a taxonomy that includes macrocategories at a coarse level; each macrocategory is further characterized by a type. Currently, we consider the following UMLS macrocategories: ACTIVITIES_AND_BEHAVIORS, ANATOMY, CHEMICALS_AND_DRUGS, CONCEPTS_AND_IDEAS, DEVICES, DISORDERS, ENTITY, GENES_AND_MOLECOLAR_SEQUENCES, GEOGRAPHIC_AREAS, LIVING_BEINGS, OBJECTS, OCCUPATIONS, ORGANIZATIONS, PHENOMENA, PHYSIOLOGY, and PROCEDURES.

As attributes, entities of the co-occurrence network include the name, an Umls_id when the entity is extracted from UMLS, and the frequency associated with the entity (ie, the number of documents in CORD-19 capturing that concept). Relationships in the co-occurrence network express the co-occurrence of 2 entities in ≥1 documents of CORD-19. Each relationship has the following attributes: a name (ie, built as concatenation in alphabetic order of the names of the entities that co-occur); a frequency (ie, the number of abstracts that mention such co-occurring entities); and several statistical indicators of the relationship's strength within the corpus, such as the pointwise mutual information value (comparing the relative frequency of 2 concepts occurring together in the text to the probability of either concept occurring independently [15]), the normalized pointwise mutual information (NPMI) value (normalized by the Shannon self-information, ranging from −1 to 1 [16]), and the Cramer V value (measuring the statistical significance of the co-occurrence between 2 entities [17]).

Figure 3 illustrates the process of ontology creation at a conceptual level. The process applies to textual abstracts (refer to Figure 3 where we consider an excerpt of the textual abstract of the study by Logette et al [18]) and consists of an entity recognition task aiming to extract the known ontological terms (ie, either from UMLS or from CIDO), followed by an entity linking task; eventually, we produce a co-occurrence network, whose entities are extracted terms and whose relationships connect entities that co-occur, weighted by the strength of the co-occurrence. Next, we detail the data extraction and transformation process.

Figure 3. Rationale of co-occurrence network construction. Ontological terms are recognized in textual abstracts using entity recognition; then, this process is reiterated with approximately 660,000 publications' abstracts. Terms are connected to each other using entity linking; each relationship between entities is associated with several properties representing the co-occurrence weight, using different statistical methods. The generated connected co-occurrence network has approximately 128,000 concepts and approximately 47 million relationships. ACE2: angiotensin-converting enzyme 2; CIDO: Coronavirus Infectious Disease Ontology; NPMI: normalized pointwise mutual information; UMLS: Unified Medical Language System.

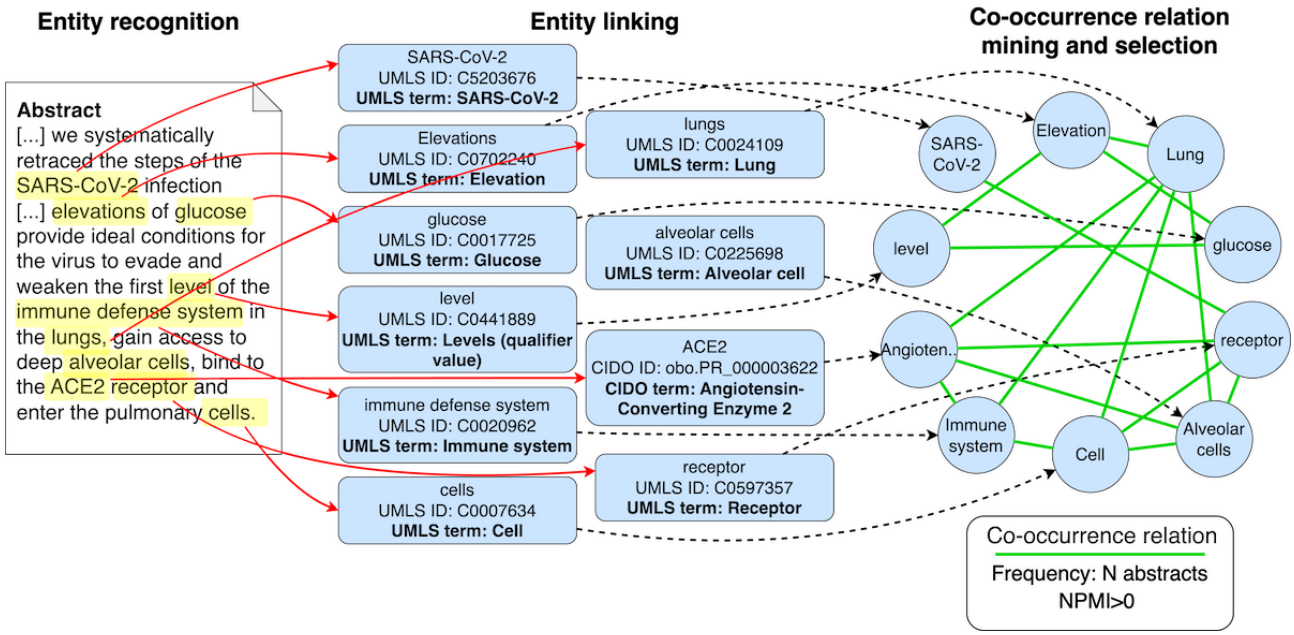

Data Provisioning and Co-Occurrence Network Construction

Overview

The data provision workflow is represented in Figure 4 [18]; it follows the extract-load-transform paradigm. Data were extracted from CORD-19 and loaded into the data storage system. The pipeline produces 3 data objects: the co-occurrence network; the metadata table; and the *inverted index*, that is, a simple postings list whose keys are the relationships of the co-occurrence network and whose elements are links to the relevant publications where such relationships co-occur. Other data tables contain intermediate results of the extraction and curation of the entities, that is, the nodes of the co-occurrence network and the computation of the co-occurrence measures used for the relationships. For storing data tables, we selected

XSL•FO
RenderX

the MariaDB relational engine [19]; for storing the co-occurrence network, we selected the Neo4j graph data engine [20].

Figure 4. Workflow diagram of the GRAPH-SEARCH data provision pipeline. Tasks are performed sequentially; each task uses data objects and produces data objects, starting from the raw COVID-19 Open Research Dataset (CORD-19) metadata.csv file present in CORD-19, which is translated into metadata.csv once cleaned. The final outcome of the pipeline is a Neo4j database containing the network.

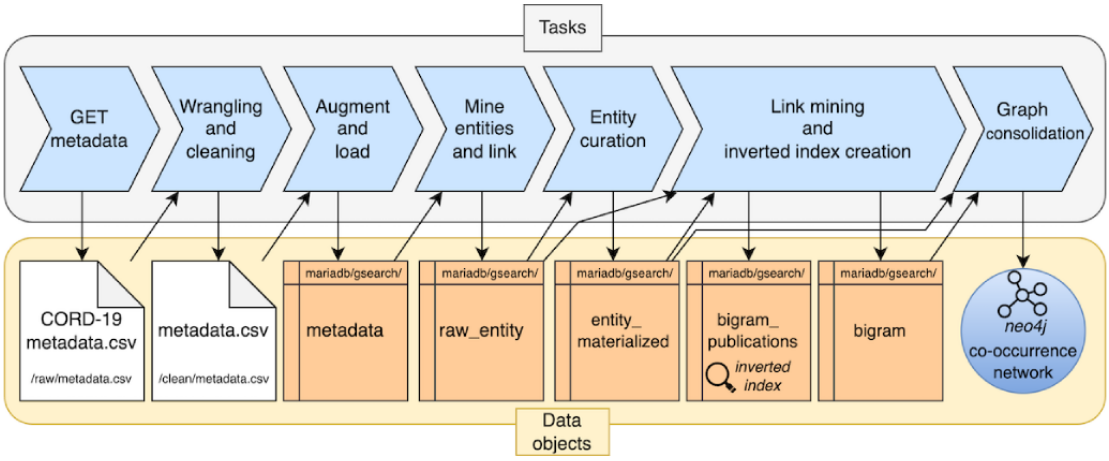

Data Loading

Three tasks apply to raw CORD-19 data and produce a metadata table. Metadata were obtained by using the "GET metadata" from the S3 bucket of Allen Institute of AI; then, we performed a "Wrangling and cleaning" step and the "Augment and load" step on the cleaned metadata table with information from the external APIs.

Entities Mining and Linking

The "Mine entities and link" task takes as input the curated and augmented metadata table and produces the raw_entity table. With a single pass over the title and abstract, we performed typical information retrieval steps such as lexical analysis, removal of stopwords, stemming, and lemmatization. Then, we performed named entity recognition (NER), consisting of the identification and extraction of entities from unstructured text and linking to UMLS and CIDO; specifically, we used the en_core_sci_lg model of the *scispaCy* Python package. The selected model is particularly suited for processing English-based scientific literature, providing an approximately 785,000 word vocabulary with 600k word vectors, with a declared F_1-score for mentions of 68.67 (refer to the study by Nuemann et al [21] for details on the achieved performances). Entities are linked to UMLS and CIDO by associating each concept with the UMLS identifier (with its related type and macroclass) and the CIDO identifier (if available).

Entity Curation

The "Entity curation" task aggregates the occurrences in the raw_entity table and outputs the entity_materialized table, collecting all the entities to be used as nodes of the co-occurrence network. In this pass, we excluded the occurrences of the entities that score a low similarity with UMLS or CIDO concepts; we used a normalized string similarity measure based on the Levenshtein distance and a threshold value of 0.7. We also included within entities some *utility* terms that indicate level modifiers (eg, "high" and "increased") or causative connectors (ie, "induces"). Eventually, we added the entity type and macrocategory using their names in UMLS.

Link Mining

The "Link mining and inverted index creation" task uses the raw_entity table and the entity_materialized table to generate the bigram table (ie, information on the links of the co-occurrence network) and the bigram_publications table that we use as an *inverted index* in the information retrieval process.

A co-occurrence is a relationship between 2 concepts, and it exists when those 2 concepts occur in the same document. Each relationship is named using the convention "X.name—Y.name," where X and Y are the 2 concepts expressed as nodes, which it connects, and X.name precedes Y.name alphabetically.

We designed a greedy algorithm—optimized for big data contexts—to extract the relationships in a single pass over the publications. This algorithm requires 2 *read-only* lookup tables, built before the execution: publication_entities (ie, for each publication a list of mentioned entities) and entity_publications (ie, for each entity, a list of mentioning publications). The complexity of the algorithm is $o(N^2)$, where N is the number of entities in the entity_materialized list; in practice, the number of required comparisons is low, as the number of entities in each publication is much lower than the total number of entities selected in the "Entity Curation" step.

Graph Consolidation

The "Graph consolidation" task selects data from the entity_materialized and bigram tables and migrates them to the Neo4j instance to create the co-occurrence network.

The nodes are curated in the previous "Entity Curation" step. The relationships of co-occurrence are chosen at this stage, based on their NPMI, which is the point estimate of the Mutual Information, normalized by the Shannon self-information (taking a value between −1 and +1); this compares the probability that the 2 entities occur together. We exclude the relationships with NPMI≤0, as a nonpositive NPMI indicates that the relationship is not significant.

XSL•FO
RenderX

The resulting co-occurrence network has 128,249 entities and 47,198,965 relationships, extracted from 662,105 initial publications. Using the Neo4j Graph Data Science library [22], we verified that the graph is a unique connected component; such a condition is essential to ensure that every possible formulated graph query can be matched on the co-occurrence network.

Graph Query Search

A graph query *Q* is a connected graph formed by nodes and undirected relationships, where nodes are the set of entities appearing in *Q* and *rels(Q)* is a set of arbitrary relationships connecting some pairs of entities in *Q*. A subgraph *Q'* is simply a connected subset of the nodes and relationships of *Q*. The search strategy is composed of 2 steps: matching of graph query against the co-occurrence network and extracting the relevant publications.

Graph query matching is the operation of comparing the graph query *Q* with the co-occurrence network *N* created along the procedure described in the *Data Provisioning and Co-Occurrence Network Construction* section. By construction, each entity in *Q* is contained in *N*, whereas relationships in *rels(Q)*, arbitrarily created in *Q*, may not be present in *N*. Both *Q* and *N* are connected graphs with undirected relationships; then, matching *Q* within *N* can be seen as an instance of inexact graph matching [23].

Figure 5 guides the intuition of the matching operation. A graph query A-B (in blue) is searched over a co-occurrence network (in white). No direct relationship exists between A and B on the network. However, several alternative finite paths exist (ie, A-X-B, A-Y-Z-B, or A-V-Y-Z-B). Among these, A-X-B is found to be the "shortest path" between A and B, as its length or distance (in green) equals 1.

Figure 5. Graph query matching operation.

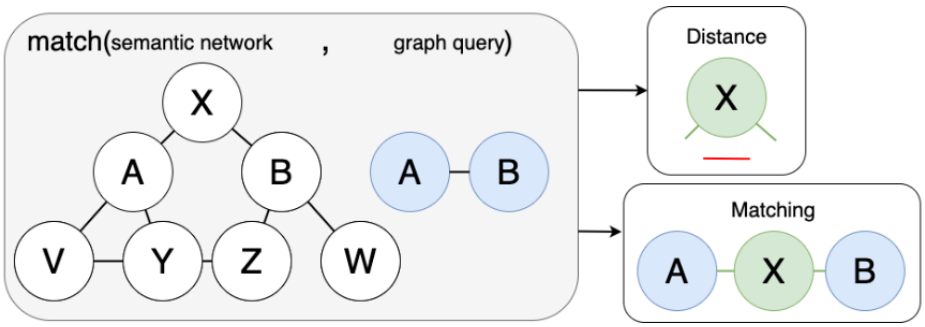

All entities in *Q* are matched in *N*; then, for each relationship *r* in *rels(Q)*, connecting nodes α and β, we retrieve the "shortest paths" within *N* that connect α and β, that is, a chain of relationships r_1', r_2', ..., r_n', where *r'* is in *rels(N)*, r_1' starts from node α, and r_n' ends in node β.

Shortest paths are computed using the All Pairs Shortest Path function allShortestPaths available in Cypher, Neo4j v4.4 [20]. Candidate shortest paths are ranked by the average of the NPMI property associated with each relationship along the path; we retain the top 10 paths in the ranking.

We refer to the set of candidate shortest paths as *expansion*; the selection of exactly 1 preferred path among the candidates of the expansion is performed interactively by the user of the search system, as it is strictly domain or context specific.

Relevant publications extraction corresponds to the retrieval of the publications that mention concepts of the matched graph, using the inverted index. We access the inverted index by relationship name, using either *r* when it appears in the relationships *rels(N)* of the co-occurrence network or all the relationships r_1', r_2', ..., r_n' appearing in the specified *path(r)*. The score of a publication *P* relative to a query *Q* (ie, the number of explained relationships) is computed as follows:

The addends of the external summation represent a score assigned to each relationship *r* in *Q*. Each addend captures how well *P* represents *r*; it is equal to 1 if *P* directly mentions the relationship of *Q* (eg, when *path(r)=r'*, with length 1) or if P mentions all the relationships of *path(r)*. Otherwise, it equals a fraction of 1, counting the number of relationships r_1', r_2', …, r_n' of *path(r)* mentioned in *P*, divided by the length of *path(r)*.

Extracted publications are ordered by their score; they are further described by other properties, such as the sum of the NPMI of all the mentioned relationships and the date of publication.

Running Example

Consider Figure 6 as an example of the four steps performed during the search:

1. *Create graph query* (Figure 6A): Nodes are chosen among the concepts existing in the co-occurrence network; node names can be found through a dedicated browser working either by autocompletion of user-typed content (ie, matching terminologies concepts) or by selection of category and type and the contained concepts; search on multiple terminologies at the same time is allowed. For each concept, we provide a description and ID from the original source. Relationships can be drawn between any pair of nodes.
2. *Find paths* (Figure 6B): For each pair of entities connected by a relationship in the graph query, the Neo4j graph is queried to find the shortest paths (at most 10) with top average NPMI scores.
3. *Select paths* (Figure 6C): The user selects the most relevant path for each original relationship that has been expanded.
4. *Retrieve publications and return ranking to the user* (Figure 6D): The system collects the names of all the relationships from the expanded graph query (computed in step B and

XSL•FO
RenderX

selected in step C) and exploits them to retrieve the posting lists of publications (from the inverted index). It computes the relationships explained by each publication. Then, it ranks the publications by (1) the number of explained relationships of the original graph query (refer to equation 1), (2) the sum of NPMI scores of the relationships, and (3) the publication date. Finally, it shows the complete list with the publications' metadata.

Figure 6. (A) Example of a graph query with 6 concepts and 5 relationships. (B) Match of graph query on the co-occurrence network, with the search of shortest paths (in the dashed spaces called expansions). Considering the relationship between SARS-CoV-2 and angiotensin-converting enzyme 2, its expansion includes 3 paths of length 3, each characterized by 2 intermediate nodes. Light green paths have the highest average normalized pointwise mutual information (NPMI) of each expansion. (C) Regardless of the suggested paths with the highest average NPMI, users can select any path (dark green). (D) A list of publications, ranked by their score, is extracted; the score is computed using equation 1 and considers all the relationships in the selected paths that are mentioned in the publication. ACE2: angiotensin-converting enzyme 2; AngII: angiotensin II.

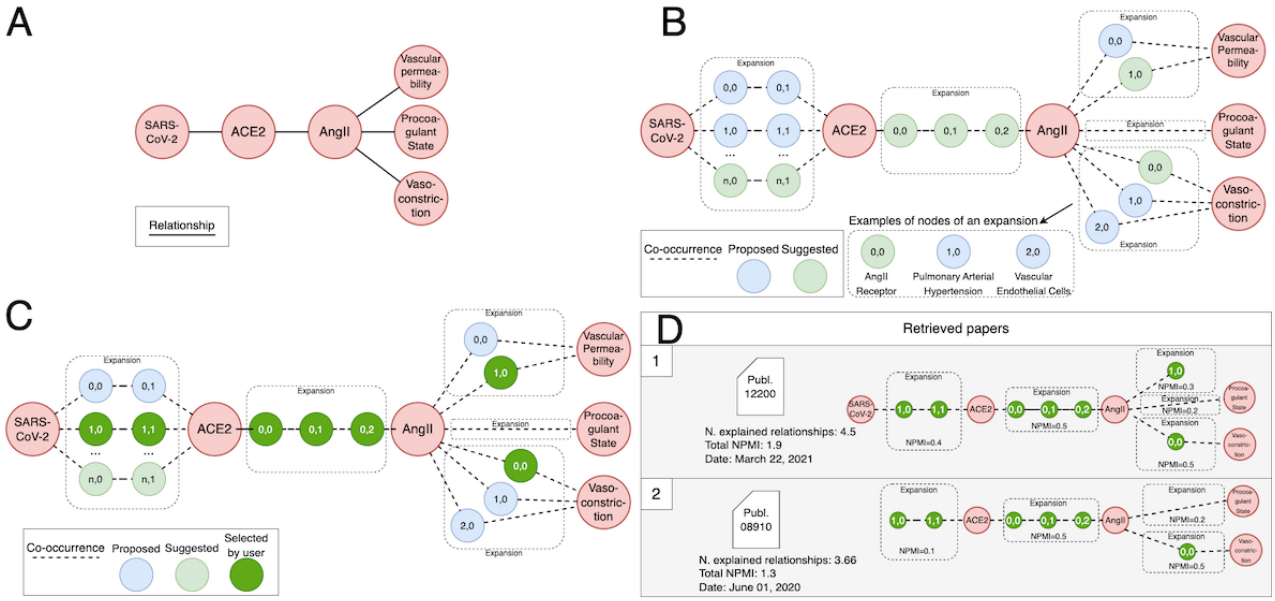

In Figure 6D, we observe that 5 expansions are produced: the first publication scores 1 in 4 expansions and 1/2 in the expansion at the top-right end of the graph query. Indeed, publication 1 only includes the relationship (AngII)-(1,0), which is half of the selected shortest path that connects (AngII) and (Vascular Permeability).

The second publication scores 0 in 1 expansion, as there is no path between (AngII) and (Vascular Permeability); 1 in 3 expansions; and 2/3 in the expansion at the left end of the graph query; the relationship (SARS-CoV-2)-(1,0) is not mentioned.

Ethical Considerations

Ethics approval was not applicable for this study.

Results

Web Interface

With GRAPH-SEARCH, the researcher can express a query in the form of a graph query on a web interface and retrieve a list of CORD-19 publications that best correspond to the query. During the search process, each link in the original graph query is expanded and matched with the co-occurrence network. When a relationship in the query is not available in the co-occurrence network, an expansion may suggest that several sets of concepts can explain a relationship in the original graph query; therefore, 10 ranked paths are proposed to the user, who may express a preference according to their interest. After selecting 1 path for each expanded relationship, GRAPH-SEARCH provides a list of publications ranked by the number of explained relationships of the original graph query.

The GRAPH-SEARCH application service exposes a web user interface to query the co-occurrence network and exploit the graph-driven search methodology described in the *Graph Query Search* section; it contains a backend (ie, web server that exposes a Representational State Transfer Application Programming Interface for high-level retrieval operations) and a frontend (ie, visual interface that exploits the Representational State Transfer Application Programming Interfaces to use the backend).

The web interface has been designed and implemented following the major steps of the algorithm described in the *Running Example* subsection above. The user experience has been modeled as a multipage application; for each step of the retrieval strategy, different API services and a different page were implemented.

The frontend is built with the Vue.js framework and the D3.js library for graph illustrations; instead, the backend is written in Python and includes two components:

1. Swagger_server, which implements the web service logic, interfaces, and the models necessary to handle the persistence and asynchronicity behaviors of a multiuser system. We used the connexion framework, a flask-based web framework, and SQLAlchemy as the database abstraction layer.
2. Core, which implements the retrieval strategy and provides high-level programming interfaces for it. This package has

XSL•FO
RenderX

been designed as an independent library that can be embedded in other applications, as it has been done with the backend service. Its implementation relies on several Python libraries, such as Neo4j, networkx, and SQLAlchemy.

UC Queries

UC1 emphasizes the strength of exploratory search over graphs by supporting users in selecting graph portions, considering eventually accepting proposed expansions, and browsing results in terms of NPMI and explained relationships. UCs of increasing complexity are provided next, offering examples of searches upon graph queries with different shapes: UC2 and UC3 introduce very simple linear graph queries (ie, 1 chain of nodes), UC4 shows the use of a Y-shaped graph query, and UC5 and UC6 represent more complex shapes with nodes forming triangles.

UC1: Genetic Mechanisms of Critical Illness in COVID-19

Pairo-Castineira et al [23] revealed previously undescribed molecular mechanisms of critical illness in patients with COVID-19 with genome-wide studies. The results of such studies may provide therapeutic targets to modulate the host immune response to promote survival. Inspired by this publication, we create a graph query including relevant human genes that are related to higher or lower severity of COVID-19 (eg, *IFNAR2, CCR2, and TYK2* genes), and we link them to the change in the severity of the disease (Figure 7A). Since the research idea is broad, we start the exploratory process focusing on a subgraph of the graph query (refer to the nodes in red selected in Figure 7A). Here, we only consider the effect of the increase of expression in the *CCR2* gene. Figure 7B shows how GRAPH-SEARCH expands the path between the concepts "High" and "Gene Expression" (not otherwise connected in the co-occurrence network). According to NPMI values, the most relevant concept connecting them is "Up-Regulation (Physiology)." Figure 7C shows that the path going through this concept has been selected by the user among the other proposed. The Results page (Figure 7D) shows a publication (Teixeira et al [24]) that covers 4 (80%) out of 5 explained relationships of the original graph query. This means that out of the 5 original relationships of the selected portion of the graph query, only 4 (80%) are explained by the publication (all except for the one between "Gene Expression" and "High"). At this point, the user can consider other portions of the graph query or the entire query.

Figure 7. GRAPH-SEARCH screens dedicated to use case 1 (UC1): (A) graph query, (B) find paths, (C) select paths, and (D) first publication on the results page.

UC2: COVID-19 and Cystic Fibrosis

Cystic fibrosis is a disorder that affects mostly the lungs, the digestive system, and other organs in the body. It is widely known that COVID-19 also affects the respiratory system. *How has their connection been investigated in CORD-19?* The simplest possible graph query in GRAPH-SEARCH holds 2 nodes (ie, cystic fibrosis and COVID-19) connected by 1 relationship of co-occurrence. "Cystic fibrosis" is represented by UMLS concept ID C0010674, and "COVID-19" is represented by the UMLS concept ID C5203670. The 2 concepts are not directly connected within the network; among the proposed paths in the expansion, we choose the one through the concept "Respiratory secretion viscosity alteration" (UMLS ID 3537094). Only 1 publication in CORD-19 explains this path, covering it completely, with an NPMI sum of 0.5668. Kratochvil et al [25] characterized the composition of respiratory secretions of intubated patients with COVID-19 infection, finding that they closely resemble those of cystic fibrosis, a minor observation unrelated to clinical severity. In general, the lack of relevant clinical references confirmed our expectation that cystic fibrosis did not impact COVID-19 severity.

XSL•FO
RenderX

UC3. COVID-19 and Nonsteroidal Anti-Inflammatory Drugs

During the second year of the pandemic, interest arose in the possibility of intervening at the onset of mild to moderate COVID-19 symptoms in outpatients (instead of hospitalized patients); it was suggested that this could prevent the progression to a more severe illness and long-term complications. More specifically, Perico et al [26] investigated the use of anti-inflammatory drugs, especially nonsteroidal anti-inflammatory drugs (NSAIDs) as a therapeutic strategy. In our graph query, we include the following as main concepts: "COVID-19" (C5203670), "Outpatients" (C0029921), "Anti-Inflammatory Agents, Non Steroidal" (C0003211), and "Cyclooxygenase 2 Inhibitors" (C1257954), with the last being a specific class of NSAIDs. In this case, no expansion of the original graph query is performed, as all the relationships are present in the co-occurrence network. The Results page contains a list of 440 publications, whose abstracts discuss the concepts in the graph query from different perspectives and approaches. The top 3 results include work from Consolaro et al [27], a home-treatment algorithm based on anti-inflammatory drugs; Popovych et al [28], discussing the therapeutic efficacy of the BNO 1030 extract, which is a phytotherapeutic anti-inflammatory agent; and Sava et al [29], exposing the results of a 90-day treatment of patients with severe COVID-19 with a specific NSAID drug, tocilizumab.

UC4: Elevated Blood Glucose Levels and COVID-19 Severity

Elevated blood glucose levels are considered a risk factor for the severity of the disease. With GRAPH-SEARCH, we compose a Y-shaped graph query (Figure 8), expressing that high levels of blood glucose or increasing blood glucose can induce a severe illness. This example makes sophisticated use of *utility* terms; these are provided in a specific list of the concepts' browser of GRAPH-SEARCH.

Figure 8. Graph query of use case 4 (UC4), with Unified Medical Language System concepts IDs in red.

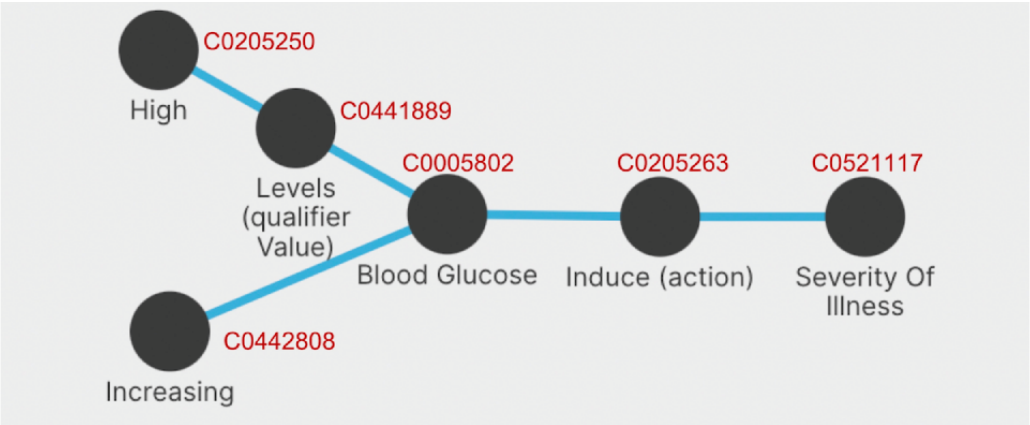

Consequently, we obtain a list of 395 results, where the top-ranked publication explains 5 out of 5 relationships. Logette et al [18] reported on the relationship between blood glucose levels and the severity of COVID-19. All following publications, ranked in descending order by the number of explained relationships of the original graph query, explain at most 3 out of 5 relations.

UC5: COVID-19, Angiotensin-Converting Enzyme 2, and Cardiovascular Diseases

Patel et al [30] hypothesized that the infection caused by SARS-CoV-2 could be associated with the shedding of angiotensin-converting enzyme 2 (ACE2). In their study, it is suggested that in patients with cardiovascular diseases, there is increased shedding of ACE2; consequently, higher levels of ACE2 in blood circulation are associated with the downregulation of membrane-bound ACE2. The graph query in Figure 9A expresses this query by connecting patients with COVID-19 infection with cardiovascular diseases; as they have more circulating ACE2, there is a downregulation of membrane-bound ACE2. When running this query, 2 relationships are not found in the co-occurrence network; the first paths suggested by the system as possible explanations are not meaningful with regard to the context; thus, we select alternative concepts, that is, "Subacute Endocarditis" and "Intensive Care Unit" (Figure 9B). Results can be ranked by the number of citations; we found the studies by Yamaguchi et al [31] and Gupta et al [32] particularly interesting, as they propose solutions for the prevention and treatment of the side effects of COVID-19 for patients with cardiovascular diseases.

XSL•FO
RenderX

Figure 9. (A) Graph query of use case 5 (UC5) and (B) found paths.

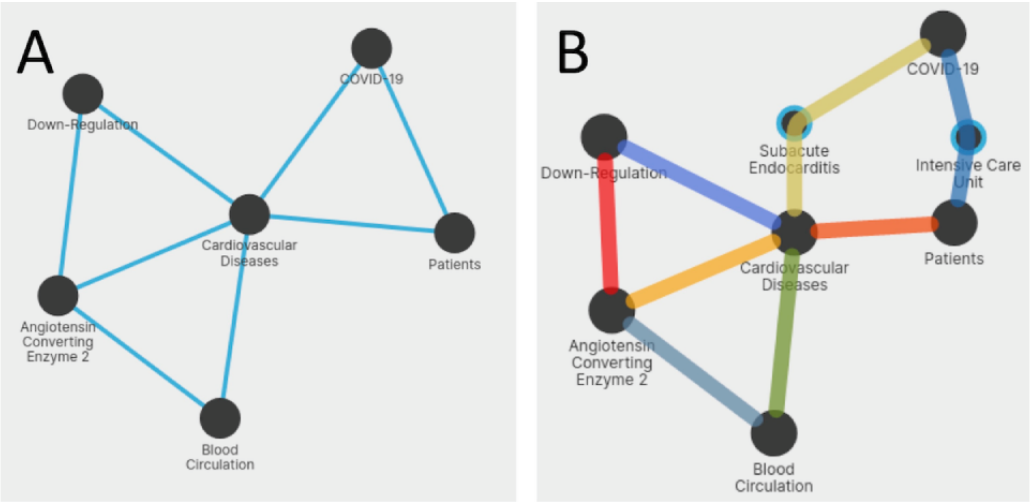

UC6: COVID-19 Vaccines and Myocarditis

The side effects of vaccines are a topic of relevance. Here, we investigate the connection between events of heart inflammation (eg, myocarditis) among adolescents and the COVID-19 Moderna vaccine. We compose a graph query in GRAPH-SEARCH with 4 nodes (Figure 10A); a triangle is formed by "Adolescent (age group)" (C0205653), "Myocarditis" (C0027059), and the "Moderna COVID-19 Vaccine" (CIDO ID obo.VO 0005157); the vaccine entity is connected to the "COVID-19" (C5203670) node. COVID-19 and Moderna COVID-19 vaccine are not directly connected; among the possible paths suggested by GRAPH-SEARCH, the 2 scoring the highest sum of mutual information are through "Vaccination" and "Myopericarditis." The latter refers to both myocarditis and pericarditis (ie, the inflammation of the pericardium, which is the sac that surrounds the heart). The latter concept allows us to expand the initial query to complete the match with the co-occurrence network (Figure 10B). On the Results page, 190 bibliographic resources are provided. The top-ranked one, which explains all 4 relationships of the graph query, is a report by Gargano et al [33] that highlights the implications of the use of messenger RNA vaccines with a higher risk for myocarditis in male individuals aged 12 to 29 years. The following results do not explain the relationship between the COVID-19 Moderna vaccine and COVID-19 through myopericarditis, as they explain only 3 relations. These results, for instance, report adverse events of myocarditis after vaccination in the United States [34] and Korea [35].

Figure 10. (A) Graph query of use case 6 (UC6) and (B) found paths.

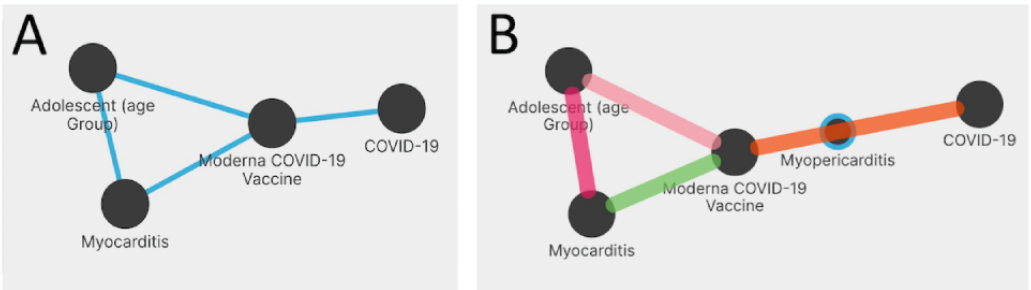

Query Performances

GRAPH-SEARCH queries are composed of two computationally intensive steps: (1) the graph query matching over the co-occurrence network and (2) the retrieval and ranking of publications related to the query. For each such step, we run a performance analysis.

Specifically, we simulated random queries with 2, 4, 6, 8, or 10 nodes from the existing co-occurrence network; we assume that these are the typical UC scenarios, as queries represent small queries of researchers created through the graphical interface. We separately measure computation times of the first and second steps (shown in Figures 11A and 11B, respectively); each experiment has been repeated on 10 queries, generated randomly using the "Random walk with restarts sampling" method of Neo4j. We observe that the computational times for graph matching in all cases is <2.2 seconds, and its growth is less-than-linear with the number of the nodes, whereas the retrieval operation typically takes up to 3 seconds, with a small number of outliers due to cache misses; the resulting user delay in these scenarios seems quite acceptable.

XSL•FO
RenderX

Figure 11. Box plots measuring the time for (A) the graph matching operation and (B) the publication retrieval operation performed using complete graph queries of 2, 4, 6, 8, and 10 nodes (each repeated 10 times).

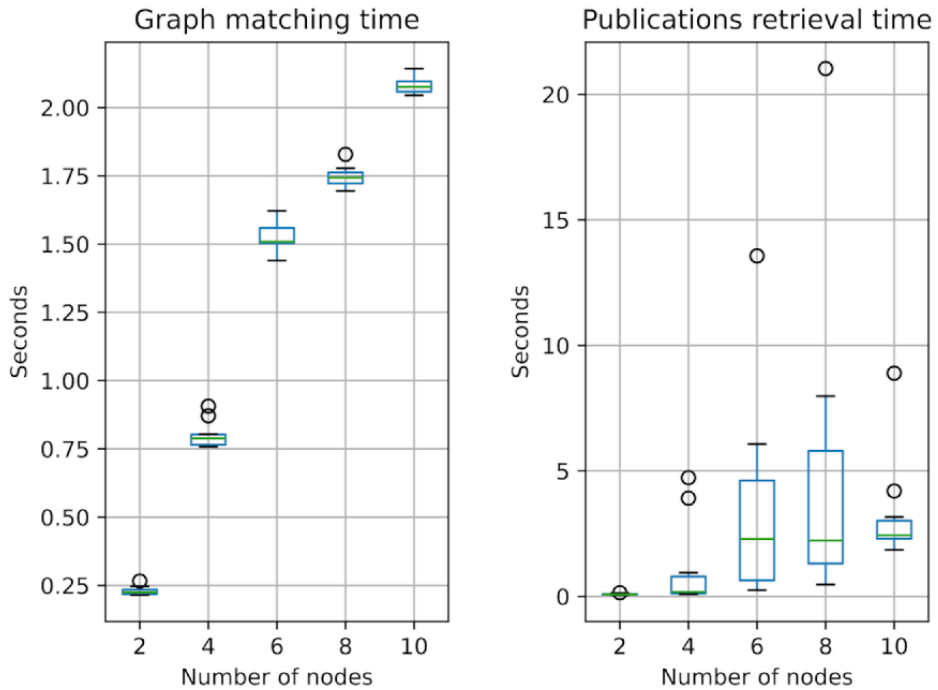

We also created random graph queries by removing increasing percentages of their relationships to simulate the difference between exact and inexact graph search (thereby triggering the search for alternative shortest paths); computational times (not shown for brevity) are not significantly affected.

Related Work

In this section, we review classic approaches to search over co-occurrence networks. Then, we focus on the specific use of bio-ontologies in information extraction systems, and finally, we propose a close comparison with COVID-19–specific search systems.

Semantic-Network Search

The task of searching and extracting literature documents over co-occurrence networks with graph-based queries can be considered through the subproblems that compose it. To query a co-occurrence network with a graph-like query, a similarity measure between graphs must be defined. Existing methods in the context of graph databases include definitions of graph edit distances and maximum common subgraphs [36], but a later approach introduced a similarity measure based on a graph kernel between pairs of documents, which exploits the shortest paths between nodes as units to compare graphs [37]. Considering the construction of the co-occurrence networks from data sets of literature documents, different approaches are available to extract concepts to represent nodes in the network and connections between them. The survey by Han et al [38] and the study by Shi et al [39] present all the main methodologies and text mining pipeline architectures, which are applied in this study to engineering and design (ie, subsets of scientific literature). G-Bean [40] is also a relevant related work, that is, a graph-based tool that exploits ontologies for graph-based query expansion to support the user search intention discovery.

Literature Annotation and Bio-Ontologies

The incorporation of bio-ontologies in information extraction and information retrieval has demonstrated its efficacy through diverse applications, such as patent information retrieval [41] and identification of concept domains [42]. Bio-ontologies are also applied in natural language processing tasks, such as NER [43]. Moreover, Wang et al [44] illustrated the application of bio-ontologies in retrieving biomedical data sets, while Maraver et al [45] emphasized their role in literature search facilitation and metadata organization. The potential for refining search queries through ontology-guided expansion is also a recurring theme in the biomedical literature for information retrieval. Diaz-Galiano et al [46] and Dong et al [47] show query expansion methodologies using different medical vocabularies.

A fundamental aspect of research in this domain pertains to the availability and use of suitable corpora and data sets; previous studies [48,49] have provided foundational annotated and curated resources that underpin the experimental frameworks addressing these tasks. Lately, the integration of bio-ontologies with language models has also gained traction within the context of bioinformation extraction [50,51].

COVID-19–Specific Literature Discovery

With the outbreak of the COVID-19 pandemic, several open-access data sets have been collected, including the National Institute of Health's COVID-19 [52], the Human Coronaviruses Data Initiative [53], and COVIDScholar [54].

The CORD-19 data set received the widest attention. Several knowledge graphs that exploit this data set were proposed at the beginning of the pandemic for representing biomedical entities (eg, CORD-NER [55] and COVID-19 KG [56]) or publications metadata (eg, COVID-19-Literature [57]). More recently, CovidPubGraph [58] has provided a comprehensive

XSL•FO
RenderX

and updated knowledge graph, which integrates information from multiple sources, making results available through a SPARQL end point. Finally, CovidGraph [59] exposed a knowledge graph in the Neo4j browser; several external ontologies are used to tag entities. The focus of these resources is more on organization and semantic enrichment than on exploration.

The goal of the TREC-COVID initiative [60] was to establish targeted retrieval tasks in response to the pandemic, to be shared and collectively addressed by the community. Instead, GRAPH-SEARCH aims to make the literature about COVID-19 searchable and explorable. This objective is common to other 2 systems, LitCovid and Outbreak.info; these support enhanced keyword-based search, but they do not offer any graph-based search support.

LitCovid [1] was developed within the US National Institutes of Health as a comprehensive resource of literature on COVID-19 (372,221 publications at the time of writing), updated regularly starting from PubMed. Publications are manually screened to assess their relevance to COVID-19. They are then categorized (eg, overview, disease mechanism, transmission dynamics, treatment, case report, and epidemic forecasting); assigned geographical locations; and annotated with any drug or chemical-related information found in their title and abstract, if applicable. The updated version [61] introduced the long-covid category, added annotations on variants and vaccines, and supported with machine learning algorithms the topic categorization (with a more updated model) and entity recognition (with NER). The interface allows us to apply filters on country, journal, drug, variant, and vaccine and compose search strings combining AND, OR, and NOT operators (ie, not documented); results are ranked by relevance, based on the widely used BM25 ranking function of Lucene. LitCovid positively compares its performances to the classical keyword search of PubMed (where annotations or tags are not used).

Outbreak.info Research Library [2] is a project of the Hughes, Su, Wu, and Andersen laboratories at Scripps Research. It offers a searchable interface of COVID-19 publications (complementing the content of LitCovid integrating preprint servers), together with clinical trials, data sets, protocols, and other resources. The data structure upon which the search is performed is supported by a schema; entities are connected by links with various semantics. The visual interface allows the use of some filters and keyword search; results are ranked by relevance based on the Lucene Practical Scoring Function on Elasticsearch (prioritizing the query normalization factor, coordination factor, term frequency, and inverse document frequency).

Discussion

In this section, we discuss how the proposed graph query search could be compared to other information extraction setups. For this purpose, we focus on 2 UC queries, that is, the linear query presented in UC3 (4 nodes in a linear pattern) and the red subgraph shown in UC1 (a nonlinear 6 nodes query, expanded with an additional node in GRAPH-SEARCH).

Comparison With COVID-19 Literature Search Systems

First, we considered running the UCs on the COVID-19 literature–dedicated search systems LitCovid and Outbreak.info. Both systems were queried using concept names corresponding to UMLS terms in the nodes; unfortunately, they both suffer from the limitations of Boolean search. Specifically, if we search with conjunctive clauses and exact search (eg, using “Outpatients” AND “Anti-Inflammatory Agents, Non Steroidal” AND “Cyclooxygenase 2 Inhibitors” AND “COVID-19” for UC3), no system returns any result. Dealing with exact search is hard. For instance, with LitCovid, the query “Cyclooxygenase Inhibitors” produces 3 results, whereas the query “Cyclooxygenase 2 Inhibitors” produces 5 results, although apparently more restrictive; instead, the query Cyclooxygenase Inhibitors (no quotes), without exact search, produces 12,287 results (including all references referring to generic inhibitors). Table 1 reports the results of LitCovid with conjunctive queries but no exact matching, while a similar search is not supported by Outbreak.info. In comparison, GRAPH-SEARCH reports 327 results for UC1 and 440 results for UC3. These outputs are hardly comparable, mainly because with LitCovid it is not possible to build a unique graph-shaped query; therefore, results of single conjunctive queries need to be evaluated one after the other, whereas GRAPH-SEARCH aggregates together the results of several conjunctive chains; it also expands given concepts with their acronyms (eg, “anti-inflammatory agents, non steroidal” is also searched as “NSAIDs”). In addition, GRAPH-SEARCH allows for the expansion of specific links by adding new concepts (eg, “Up-Regulation [Physiology]” in UC1). No domain-specific system for COVID-19 supports graph-based search, allowing a more insightful comparison.

XSL•FO
RenderX

Table 1. Results of the evaluation of use case (UC) 1 (Figure 7) and UC3 queries when performed on the LitCovid search interface, on the full-text indexed MariaDB database, and on GRAPH-SEARCH.

Query	Retrieved pubs LitCovid, n	Retrieved pubs MariaDB, n
UC1: GRAPH-SEARCH retrieves 327 publications overall		
(Severe (severity modifier)) AND (Disease) AND (Associated With) AND (Gene Expression) AND (High) AND (CCR2 gene)	316	11
(Severe (severity modifier)) AND (Disease) AND (Associated With) AND (Gene Expression) AND (High) AND (CCR2 gene) AND Up-Regulation (Physiology)	52	12
(Outpatients) AND (Anti-Inflammatory Agents, Non Steroidal) AND (Cyclooxygenase 2 Inhibitors)	972	4
UC 3: GRAPH-SEARCH retrieves 440 publications overall		
(Outpatients) AND (Anti-Inflammatory Agents, Non Steroidal) AND (COVID-19)	1714	3
(Outpatients) AND (Cyclooxygenase 2 Inhibitors) AND (COVID-19)	3018	1
(Anti-Inflammatory Agents, Non Steroidal) AND (Cyclooxygenase 2 Inhibitors) AND (COVID-19)	37322	5
(Outpatients) AND (Anti-Inflammatory Agents, Non Steroidal) AND (Cyclooxygenase 2 Inhibitors) AND (COVID-19)	902	5
“Outpatients” AND “Anti-Inflammatory Agents, Non Steroidal” AND “Cyclooxygenase 2 Inhibitors” AND “COVID-19”	0	5

Comparison With the Search on Full-Text Indexed Corpora

We also attempted a comparison with search operations performed on a baseline created by full-text indexing the CORD-19 titles and abstracts. Specifically, we used the full-text indexing option of MariaDB, an open-source fork of MySQL [19]. Typically, full-text indexes work well for regular text; they build an index over specific words rather than the whole text, and consequently, they show good performances for searches of specific words. The same queries used on LitCovid and Outbreak.info were used on this setup: on MariaDB, we used the “Natural language mode” documented on MariaDB [62] and, thus, we removed the “AND” Boolean operators and parentheses. To be part of the index, words must appear in <50% of the documents to be considered potentially relevant and to be used in searches (consequently, “COVID-19” and “SARS-CoV-2” are not considered relevant). Results are returned in descending order of relevance; limitations include the exclusion of partial (or very short or long) words.

Notwithstanding our attempts, we note that the comparison of the GRAPH-SEARCH approach with the full-text indexing setup is very difficult for many reasons:

1. The databases upon which search is performed are built on different assumptions (eg, to be part of the index, words must appear in <50% of the documents; the co-occurrence network only includes entities that score high similarity with ontology concepts and exclude relationships with a negative NPMI).
2. In 1 case, we perform separate keyword search sessions with separate results (with associated precision and recall measures); in the other case, we retrieve aggregated results (with summarized measures).
3. On one side, the ranking produced is only on single query result sets; on the other side, it is a global ranking.

The results are reported in Table 1; they must be read considering all these aspects. Note that results achieved with keyword search are restricted to manipulating Boolean expressions, adding, keywords and dropping keywords. On the contrary, the results on GRAPH-SEARCH (327 and 440, respectively for UC1 and UC3) are inspectable, with ranking, ordering, filtering, and visualization options dedicated to the explained chains of entities; using our search paradigm, users can compose graph queries; more complex topologies also allow a stronger explainability of results.

Conclusions

GRAPH-SEARCH is the first search engine to propose the exploration of COVID-19 scientific literature using visual graph queries. GRAPH-SEARCH provides several unique features such as the possibility to describe concepts using well-known ontologies, to establish co-occurrence relationships between any 2 concepts of choice, to support search queries with concepts proposed and ranked by the system, and to browse resulting publications exploiting several visual and analytic measures.

The completeness and accuracy of the information captured in the co-occurrence network strictly depend on the advances of the NER methods used during the steps of entity mining and linking. Other systems have used expert curation (eg, LitCovid) or community-driven curation (eg, Outbreak.info). Although expert curation can improve the search experience, it does not properly scale; we opted for the exploitation of well-known biomedical ontologies such as UMLS and CIDO and state-of-the-art natural language processing models used for Entity Recognition in our data provision pipeline.

The ability of our system to extract results was evaluated, attempting a comparison with existing published systems (eg,

XSL•FO
RenderX

LitCovid and Outbreak.info) and with full-text indexing search. We recognize that comparisons between the results retrieved from these systems are not ideal, as it is very critical to compare single search runs with a system where the result is built progressively on the graph—considering a set of aspects altogether (ie, how the network was built and pruned, shortest path computation, completion with additional nodes, and global ranking of results).

Co-occurrence networks are conventionally used for analyzing extensive text and big data. Common applications have involved sentiment analysis [63] and detection of prevailing topics [64]. Here, each node is a word occurring in a set of user-generated social media posts. Moreover, word-co-occurrence networks are present in clinical applications, for example, Millington and Luz [65] proposed to encode recordings of speech data used for recognizing patients with Alzheimer and controls. In all such cases, GRAPH-SEARCH may be used to find specific subgraphs and propose completions of missing links.

In this study, we have demonstrated the capability of domain-specific (even inexact) graph query matching when semantics is considered only for nodes; we are aware of the limitations of this approach, which, at this stage, is considered a modeling choice. In future work, we plan to extend our search system to semantically rich knowledge graphs with both entities and relationships, thereby enriching the expressivity of graph queries (also including the possibility to capture the semantics of relationships, with state-of-the-art methods [66] or as we already experimented in a previous study [67]). Then, we aim to formalize the use of graph queries in the context of graph databases by studying the complexity of graph search and connecting it to classical theories of subgraph matching, shortest path search, and conjunctive query processing.

We also aim to conduct extensive empirical studies to measure user satisfaction with systems such as GRAPH-SEARCH analyzed along the 3 dimensions of usability, usefulness in deepening their knowledge of certain connected topics, and support of user's intentions in knowledge exploration.

Acknowledgments

The authors would like to thank Luca Minotti for implementing the frontend of the GRAPH-SEARCH web application. This research is supported by the PNRR-PE-AI FAIR project funded by the NextGenerationEU program.

Data Availability

The data processing pipeline is available as a Docker image [68]. The GRAPH-SEARCH application is available [69] and documented in the wiki [70].

Authors' Contributions

AB and SC conceived the work; FI, AB, and SC jointly conceptualized and designed the framework; FI implemented the pipelines of data ingestion, network construction, and graph query engine; FI and AB curated the user experience; AB drafted the manuscript; FI and SC improved it; all authors revised the final version of the manuscript; and SC supervised the project.

Conflicts of Interest

None declared.

References

1. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res 2021 Jan 08;49(D1):D1534-D1540 [FREE Full text] [doi: 10.1093/nar/gkaa952] [Medline: 33166392]
2. Tsueng G, Mullen JL, Alkuzweny M, Cano M, Rush B, Haag E, et al. Outbreak.info Research Library: a standardized, searchable platform to discover and explore COVID-19 resources. Nat Methods 2023 Apr;20(4):536-540 [FREE Full text] [doi: 10.1038/s41592-023-01770-w] [Medline: 36823331]
3. Kejriwal M. Knowledge graphs and COVID-19: opportunities, challenges, and implementation. Harvard Data Science Review. 2020 Dec 1. URL: https://hdsr.mitpress.mit.edu/pub/xl0yk6ux/release/4 [accessed 2024-05-08]
4. Peng J, Xu D, Lee R, Xu S, Zhou Y, Wang K. Expediting knowledge acquisition by a web framework for Knowledge Graph Exploration and Visualization (KGEV): case studies on COVID-19 and Human Phenotype Ontology. BMC Med Inform Decis Mak 2022 Jun 02;22(Suppl 2):147 [FREE Full text] [doi: 10.1186/s12911-022-01848-z] [Medline: 35655307]
5. Ware C. Visual queries: the foundation of visual thinking. In: Tergan SO, Keller T, editors. Knowledge and Information Visualization. Berlin, Heidelberg: Springer; 2005.
6. Cheerkoot-Jalim S, Khedo KK. Literature-based discovery approaches for evidence-based healthcare: a systematic review. Health Technol (Berl) 2021;11(6):1205-1217 [FREE Full text] [doi: 10.1007/s12553-021-00605-y] [Medline: 34722102]
7. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]

XSL•FO
RenderX

8. He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Sci Data 2020 Jun 12;7:181 [FREE Full text] [doi: 10.1038/s41597-020-0523-6] [Medline: 32533075]
9. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: an advanced information management application for biomedicine. Inf Serv Use 2011 Sep 06;31(1-2):15-21. [doi: 10.3233/ISU-2011-0627]
10. Bernasconi A, Canakoglu A, Masseroli M, Ceri S. META-BASE: a novel architecture for large-scale genomic metadata integration. IEEE/ACM Trans Comput Biol Bioinf 2022;19(1):543-557. [doi: 10.1109/tcbb.2020.2998954]
11. Fernandez JD, Lenzerini M, Masseroli M, Venco F, Ceri S. Ontology-based search of genomic metadata. IEEE/ACM Trans Comput Biol Bioinf 2016 Mar 1;13(2):233-247. [doi: 10.1109/tcbb.2015.2495179]
12. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. CORD-19: the COVID-19 Open Research Dataset. arXiv Preprint posted online on April 22, 2020 [FREE Full text] [doi: 10.48550/arXiv.2004.10706]
13. Rose ME, Kitchin JR. pybliometrics: scriptable bibliometrics using a Python interface to Scopus. SoftwareX 2019 Jul;10:100263. [doi: 10.1016/j.softx.2019.100263]
14. danielnsilva / semanticscholar. GitHub. URL: https://github.com/danielnsilva/semanticscholar [accessed 2023-11-15]
15. Church KW, Hanks P. Word association norms, mutual information, and lexicography. In: Proceedings of the 27th Annual Meeting on Association for Computational Linguistics. 1989 Presented at: ACL '89; June 26-29, 1989; Vancouver, BC. [doi: 10.3115/981623.981633]
16. Bouma G. Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of Conferences of the German Society for Computational Linguistics and Language Technology. 2009 Presented at: GSCL 2009; September 30, 2009; Potsdam, Germany.
17. Cramér H. Mathematical Methods of Statistics. Princeton, NJ: Princeton University Press; 1999.
18. Logette E, Lorin C, Favreau C, Oshurko E, Coggan JS, Casalegno F, et al. A machine-generated view of the role of blood glucose levels in the severity of COVID-19. Front Public Health 2021 Jul 28;9:695139 [FREE Full text] [doi: 10.3389/fpubh.2021.695139] [Medline: 34395368]
19. MariaDB foundation homepage. MariaDB Foundation. URL: https://mariadb.org/ [accessed 2023-11-15]
20. Neo4j graph database and analytics. Neo4j. URL: https://neo4j.com/ [accessed 2023-11-15]
21. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP 2019; August 1, 2019; Florence, Italy. [doi: 10.18653/v1/w19-5034]
22. neo4j / graph-data-science. GitHub. URL: https://github.com/neo4j/graph-data-science [accessed 2023-11-15]
23. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, GenOMICC Investigators, ISARIC4C Investigators, COVID-19 Human Genetics Initiative, 23andMe Investigators, BRACOVID Investigators, Gen-COVID Investigators, et al. Genetic mechanisms of critical illness in COVID-19. Nature 2021 Mar;591(7848):92-98. [doi: 10.1038/s41586-020-03065-y] [Medline: 33307546]
24. Teixeira PC, Dorneles GP, Santana Filho PC, da Silva IM, Schipper LL, Postiga IA, et al. Increased LPS levels coexist with systemic inflammation and result in monocyte activation in severe COVID-19 patients. Int Immunopharmacol 2021 Nov;100:108125 [FREE Full text] [doi: 10.1016/j.intimp.2021.108125] [Medline: 34543980]
25. Kratochvil MJ, Kaber G, Demirdjian S, Cai PC, Burgener EB, Nagy N, Stanford COVID-19 Biobank Study Group, et al. Biochemical, biophysical, and immunological characterization of respiratory secretions in severe SARS-CoV-2 infections. JCI Insight 2022 Jun 22;7(12):e152629 [FREE Full text] [doi: 10.1172/jci.insight.152629] [Medline: 35730564]
26. Perico N, Cortinovis M, Suter F, Remuzzi G. Home as the new frontier for the treatment of COVID-19: the case for anti-inflammatory agents. Lancet Infect Dis 2023 Jan;23(1):e22-e33 [FREE Full text] [doi: 10.1016/S1473-3099(22)00433-9] [Medline: 36030796]
27. Consolaro E, Suter F, Rubis N, Pedroni S, Moroni C, Pastò E, et al. A home-treatment algorithm based on anti-inflammatory drugs to prevent hospitalization of patients with early COVID-19: a matched-cohort study (COVER 2). Front Med (Lausanne) 2022 Apr 22;9:785785 [FREE Full text] [doi: 10.3389/fmed.2022.785785] [Medline: 35530041]
28. Popovych V, Koshel I, Malofiichuk A, Pyletska L, Semeniuk A, Filippova O, et al. A randomized, open-label, multicenter, comparative study of therapeutic efficacy, safety and tolerability of BNO 1030 extract, containing marshmallow root, chamomile flowers, horsetail herb, walnut leaves, yarrow herb, oak bark, dandelion herb in the treatment of acute non-bacterial tonsillitis in children aged 6 to 18 years. Am J Otolaryngol 2019;40(2):265-273 [FREE Full text] [doi: 10.1016/j.amjoto.2018.10.012] [Medline: 30554882]
29. Sava M, Sommer G, Daikeler T, Woischnig AK, Martinez AE, Leuzinger K, et al. Ninety-day outcome of patients with severe COVID-19 treated with tocilizumab - a single centre cohort study. Swiss Med Wkly 2021 Aug 10;151:w20550 [FREE Full text] [doi: 10.4414/smw.2021.20550] [Medline: 34375986]
30. Patel SK, Juno JA, Lee WS, Wragg KM, Hogarth PM, Kent SJ, et al. Plasma ACE2 activity is persistently elevated following SARS-CoV-2 infection: implications for COVID-19 pathogenesis and consequences. Eur Respir J 2021 May 13;57(5):2003730 [FREE Full text] [doi: 10.1183/13993003.03730-2020] [Medline: 33479113]

XSL•FO
RenderX

31. Yamaguchi T, Hoshizaki M, Minato T, Nirasawa S, Asaka M, Niiyama M, et al. ACE2-like carboxypeptidase B38-CAP protects from SARS-CoV-2-induced lung injury. Nat Commun 2021 Nov 23;12(1):6791 [FREE Full text] [doi: 10.1038/s41467-021-27097-8] [Medline: 34815389]
32. Gupta S, Mitra A. Challenge of post-COVID era: management of cardiovascular complications in asymptomatic carriers of SARS-CoV-2. Heart Fail Rev 2022 Jan;27(1):239-249 [FREE Full text] [doi: 10.1007/s10741-021-10076-y] [Medline: 33426593]
33. Gargano JW, Wallace M, Hadler SC, Langley G, Su JR, Oster ME, et al. Use of mRNA COVID-19 vaccine after reports of myocarditis among vaccine recipients: update from the advisory committee on immunization practices - United States, June 2021. MMWR Morb Mortal Wkly Rep 2021 Jul 09;70(27):977-982 [FREE Full text] [doi: 10.15585/mmwr.mm7027e2] [Medline: 34237049]
34. Oster ME, Shay DK, Su JR, Gee J, Creech CB, Broder KR, et al. Myocarditis cases reported after mRNA-based COVID-19 vaccination in the US from December 2020 to August 2021. JAMA 2022 Jan 25;327(4):331-340 [FREE Full text] [doi: 10.1001/jama.2021.24110] [Medline: 35076665]
35. Lee E, Kim K, Kim M, Yang H, Yum HY, Lee M, et al. Adverse reactions to coronavirus disease 2019 vaccines in children and adolescents. Allergy Asthma Respir Dis 2022 Jan;10(1):9-14. [doi: 10.4168/aard.2022.10.1.9]
36. Zhu Y, Qin L, Yu JX, Cheng H. Finding top-k similar graphs in graph databases. In: Proceedings of the 15th International Conference on Extending Database Technology. 2012 Presented at: EDBT '12; March 27-30, 2012; Berlin, Germany. [doi: 10.1145/2247596.2247650]
37. Nikolentzos G, Meladianos P, Rousseau F, Stavrakas Y, Vazirgiannis M. Shortest-path graph kernels for document similarity. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: EMNLP 2017; September 9-11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/d17-1202]
38. Han J, Sarica S, Shi F, Luo J. Semantic networks for engineering design: state of the art and future directions. J Mech Des 2022 Feb;144(2):020802. [doi: 10.1115/1.4052148]
39. Shi F, Chen L, Han J, Childs P. A data-driven text mining and semantic network analysis for design information retrieval. J Mech Des 2017 Nov;139(11):111402. [doi: 10.1115/1.4037649]
40. Wang JZ, Zhang Y, Dong L, Li L, Srimani PK, Yu PS. G-Bean: an ontology-graph based web tool for biomedical literature retrieval. BMC Bioinform 2014 Nov 6;15(Suppl 12):S1. [doi: 10.1186/1471-2105-15-s12-s1]
41. Taduri S, Law KH, Kesan JP, Sriram RD. Utilization of bio-ontologies for enhancing patent information retrieval. In: Proceedings of the IEEE 43rd Annual Computer Software and Applications Conference. 2019 Presented at: COMPSAC 2019; July 15-19, 2019; Milwaukee, WI. [doi: 10.1109/compsac.2019.10189]
42. Dinh D, Tamine L. Identification of concept domains and its application in biomedical information retrieval. Inf Syst E Bus Manage 2014 Sep 20;13:647-672. [doi: 10.1007/s10257-014-0259-y]
43. Gao S, Kotevska O, Sorokine A, Christian JB. A pre-training and self-training approach for biomedical named entity recognition. PLoS One 2021 Feb 9;16(2):e0246310 [FREE Full text] [doi: 10.1371/journal.pone.0246310] [Medline: 33561139]
44. Wang X, Huang Z, van Harmelen F. Ontology-based semantic similarity approach for biomedical dataset retrieval. In: Proceedings of the Health Information Science. 2020 Presented at: HIS 2020; October 20-23, 2020; Amsterdam, The Netherlands. [doi: 10.1177/0002764289032005003]
45. Maraver P, Armañanzas R, Gillette TA, Ascoli GA. PaperBot: open-source web-based search and metadata organization of scientific literature. BMC Bioinformatics 2019 Jan 24;20(1):50 [FREE Full text] [doi: 10.1186/s12859-019-2613-z] [Medline: 30678631]
46. Díaz-Galiano MC, Martín-Valdivia MT, Ureña-López LA. Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput Biol Med 2009 Apr;39(4):396-403. [doi: 10.1016/j.compbiomed.2009.01.012] [Medline: 19268924]
47. Dong L, Srimani PK, Wang JZ. Ontology graph based query expansion for biomedical information retrieval. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2011 Presented at: BIBM 2011; November 12-15, 2011; Atlanta, GA. [doi: 10.1109/bibm.2011.15]
48. Mohan S, Li D. MedMentions: a large biomedical corpus annotated with UMLS concepts. arXiv Preprint posted online on February 25, 2019 [FREE Full text] [doi: 10.24432/C5G59C]
49. Basaldella M, Liu F, Shareghi E, Collier N. COMETA: a corpus for medical entity linking in the social media. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP 2020; November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-main.253]
50. Fei H, Ren Y, Zhang Y, Ji D, Liang X. Enriching contextualized language model from knowledge graph for biomedical information extraction. Brief Bioinform 2021 May 20;22(3):bbaa110. [doi: 10.1093/bib/bbaa110] [Medline: 32591802]
51. Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. J Biomed Inform 2022 Feb;126:103982 [FREE Full text] [doi: 10.1016/j.jbi.2021.103982] [Medline: 34974190]
52. Welcome to the COVID-19 portfolio. National Institute for Health Office of Portfolio Analysis. URL: https://icite.od.nih.gov/covid19/search/ [accessed 2023-11-15]
53. Human coronaviruses data initiative. Lens. URL: https://about.lens.org/covid-19/ [accessed 2023-11-15]

XSL•FO
RenderX

54. Dagdelen J, Trewartha A, Huo H, Fei Y, He T, Cruse K, et al. COVIDScholar: an automated COVID-19 research aggregation and analysis platform. PLoS One 2023 Feb 1;18(2):e0281147 [FREE Full text] [doi: 10.1371/journal.pone.0281147] [Medline: 36724184]
55. Wang X, Song X, Li B, Guan Y, Han J. Comprehensive named entity recognition on CORD-19 with distant or weak supervision. arXiv Preprint posted online on March 27, 2020 [FREE Full text] [doi: 10.48550/arXiv.2003.12218]
56. Chen C, Ross KE, Gavali S, Cowart JE, Wu CH. COVID-19 Knowledge Graph from semantic integration of biomedical literature and databases. Bioinformatics 2021 Dec 07;37(23):4597-4598 [FREE Full text] [doi: 10.1093/bioinformatics/btab694] [Medline: 34613368]
57. Steenwinckel B, Vandewiele G, Rausch I, Heyvaert P, Taelman R, Colpaert P, et al. Facilitating the analysis of COVID-19 literature through a knowledge graph. In: Proceedings of the 19th International Semantic Web Conference. 2020 Presented at: ISWC 2020; November 2-6, 2020; Athens, Greece. [doi: 10.1007/978-3-030-62466-8_22]
58. Pestryakova S, Vollmers D, Sherif MA, Heindorf S, Saleem M, Moussallem D, et al. COVIDPUBGRAPH: a FAIR knowledge graph of COVID-19 publications. Sci Data 2022 Jul 08;9(1):389 [FREE Full text] [doi: 10.1038/s41597-022-01298-2] [Medline: 35803947]
59. Gütebier L, Bleimehl T, Henkel R, Munro J, Müller S, Morgner A, et al. CovidGraph: a graph to fight COVID-19. Bioinformatics 2022 Oct 14;38(20):4843-4845 [FREE Full text] [doi: 10.1093/bioinformatics/btac592] [Medline: 36040169]
60. Roberts K, Alam T, Bedrick S, Demner-Fushman D, Lo K, Soboroff I, et al. Searching for scientific evidence in a pandemic: an overview of TREC-COVID. J Biomed Inform 2021 Sep;121:103865 [FREE Full text] [doi: 10.1016/j.jbi.2021.103865] [Medline: 34245913]
61. Chen Q, Allot A, Leaman R, Wei CH, Aghaarabi E, Guerrerio JJ, et al. LitCovid in 2022: an information resource for the COVID-19 literature. Nucleic Acids Res 2023 Jan 06;51(D1):D1512-D1518 [FREE Full text] [doi: 10.1093/nar/gkac1005] [Medline: 36350613]
62. Full-text index overview. MariaDB. URL: https://mariadb.com/kb/en/full-text-index-overview/ [accessed 2023-11-15]
63. Fudolig MI, Alshaabi T, Arnold MV, Danforth CM, Dodds PS. Sentiment and structure in word co-occurrence networks on Twitter. Appl Netw Sci 2022 Feb 14;7(1):1-27. [doi: 10.1007/s1109-022-00446-2]
64. Segev E. Textual network analysis: detecting prevailing themes and biases in international news and social media. Sociol Compass 2020 Feb 14;14(4):e12779. [doi: 10.1111/soc4.12779]
65. Millington T, Luz S. Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. Front Comput Sci 2021 Apr 29;3:1-12. [doi: 10.3389/fcomp.2021.649508]
66. Singhal A, Simmons M, Lu Z. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. PLoS Comput Biol 2016 Nov 30;12(11):e1005017 [FREE Full text] [doi: 10.1371/journal.pcbi.1005017] [Medline: 27902695]
67. Serna García G, Al Khalaf R, Invernici F, Ceri S, Bernasconi A. CoVEffect: interactive system for mining the effects of SARS-CoV-2 mutations and variants based on deep learning. Gigascience 2022 Dec 28;12:giad036 [FREE Full text] [doi: 10.1093/gigascience/giad036] [Medline: 37222749]
68. frinve/graph-search. Docker. URL: https://hub.docker.com/r/frinve/graph-search [accessed 2024-05-14]
69. GRAPH-SEARCH homepage. GRAPH-SEARCH. URL: http://geco.deib.polimi.it/graph-search/ [accessed 2024-05-14]
70. FrInve / graph-search. GitHub. URL: https://github.com/FrInve/graph-search/wiki/ [accessed 2024-05-14]

Abbreviations

ACE2: angiotensin-converting enzyme 2
API: application programming interface
CIDO: Coronavirus Infectious Disease Ontology
CORD-19: COVID-19 Open Research Dataset
NER: named entity recognition
NPMI: normalized pointwise mutual information
NSAID: nonsteroidal anti-inflammatory drugs
UC: use case
UMLS: Unified Medical Language System

XSL•FO
RenderX

Edited by G Tsafnat; submitted 26.11.23; peer-reviewed by S Hume, A Rasool; comments to author 28.02.24; revised version received 06.03.24; accepted 30.03.24; published 30.05.24.

Please cite as:
Invernici F, Bernasconi A, Ceri S
Searching COVID-19 Clinical Research Using Graph Queries: Algorithm Development and Validation
J Med Internet Res 2024;26:e52655
URL: https://www.jmir.org/2024/1/e52655
doi:10.2196/52655
PMID:38814687

©Francesco Invernici, Anna Bernasconi, Stefano Ceri. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Federal Communications Commission and Microsoft Estimates of Broadband Access for Mental Health Video Telemedicine Among Veterans: Retrospective Cohort Study

Amy MJ O'Shea[1,2,3], PhD; Kailey Mulligan[1,4], MS; Knute D Carter[4], PhD; Bjarni Haraldsson[1], MS; Charlie M Wray[5,6], MS, DO; Ariana Shahnazi[1], PhD; Peter J Kaboli[1,2,3], MD, MS

[1]Center for Access and Delivery Research and Evaluation (CADRE), Iowa City VA Healthcare System, Iowa City, IA, United States
[2]Veterans Rural Health Resource Center-Iowa City, VA Office of Rural Health, Iowa City, IA, United States
[3]Department of Internal Medicine, University of Iowa Carver College of Medicine, Iowa City, IA, United States
[4]Department of Biostatistics, University of Iowa College of Public Health, Iowa City, IA, United States
[5]Department of Medicine, University of California, San Francisco, CA, United States
[6]Section of Hospital Medicine, San Francisco Veterans Affairs Medical Center, San Francisco, CA, United States

Corresponding Author:
Amy MJ O'Shea, PhD
Center for Access and Delivery Research and Evaluation (CADRE)
Iowa City VA Healthcare System
601 US Hwy 6
Iowa City, IA, 52246
United States
Phone: 1 3193380581
Email: amy.oshea@va.gov

Abstract

Background: The COVID-19 pandemic highlighted the importance of telemedicine in health care. However, video telemedicine requires adequate broadband internet speeds. As video-based telemedicine grows, variations in broadband access must be accurately measured and characterized.

Objective: This study aims to compare the Federal Communications Commission (FCC) and Microsoft US broadband use data sources to measure county-level broadband access among veterans receiving mental health care from the Veterans Health Administration (VHA).

Methods: Retrospective observational cohort study using administrative data to identify mental health visits from January 1, 2019, to December 31, 2020, among 1161 VHA mental health clinics. The exposure is county-level broadband percentages calculated as the percentage of the county population with access to adequate broadband speeds (ie, download >25 megabits per second) as measured by the FCC and Microsoft. All veterans receiving VHA mental health services during the study period were included and categorized based on their use of video mental health visits. Broadband access was compared between and within data sources, stratified by video versus no video telemedicine use.

Results: Over the 2-year study period, 1,474,024 veterans with VHA mental health visits were identified. Average broadband percentages varied by source (FCC mean 91.3%, SD 12.5% vs Microsoft mean 48.2%, SD 18.1%; $P<.001$). Within each data source, broadband percentages generally increased from 2019 to 2020. Adjusted regression analyses estimated the change after pandemic onset versus before the pandemic in quarterly county-based mental health visit counts at prespecified broadband percentages. Using FCC model estimates, given all other covariates are constant and assuming an FCC percentage set at 70%, the incidence rate ratio (IRR) of county-level quarterly mental video visits during the COVID-19 pandemic was 6.81 times (95% CI 6.49-7.13) the rate before the pandemic. In comparison, the model using Microsoft data exhibited a stronger association (IRR 7.28; 95% CI 6.78-7.81). This relationship held across all broadband access levels assessed.

Conclusions: This study found FCC broadband data estimated higher and less variable county-level broadband percentages compared to those estimated using Microsoft data. Regardless of the data source, veterans without mental health video visits lived in counties with lower broadband access, highlighting the need for accurate broadband speeds to prioritize infrastructure

XSL•FO
RenderX

and intervention development based on the greatest community-level impacts. Future work should link broadband access to differences in clinical outcomes.

(J Med Internet Res 2024;26:e47100) doi:10.2196/47100

KEYWORDS

broadband; telemedicine; Federal Communications Commission; veterans; United States Department of Veterans Affairs; internet; mental health care; veteran health; broadband access; web-based; digital

Introduction

During the COVID-19 pandemic, in-person visits were heavily restricted, allowing for the rapid expansion of video-based telemedicine. With substantial prepandemic investments in training and technology, the Veterans Health Administration (VHA) was well positioned to quickly expand the use of video visits during the pandemic [1]. Within VHA, video-based visits are delivered primarily through Veterans Affairs (VA) Video Connect, a secure telemedicine app that connects VHA providers and patients via live video at a location of the veteran's choosing (eg, home and work).

However, video telemedicine requires adequate broadband speeds, defined by the Federal Communications Commission (FCC) and for this study as ≥25 megabits per second (Mbps) download and ≥3 Mbps upload speeds [2]. As broadband services expand, some have argued higher broadband speeds are essential for daily life, including health care services [3,4]. Beyond infrastructural availability, access to adequate broadband internet is influenced by socioeconomic factors, age, digital literacy, and rurality [5-9]. Homes in rural areas are less likely to have access to terrestrial broadband infrastructure [9,10]. This is amplified among veterans, of whom 30% reside in rural areas and 15% lack home internet [11,12]. Previous research has shown that rural, homeless, and older populations typically experience fewer telehealth visits [13,14]. Altogether, these individual- and area-level characteristics and their relationship with broadband access may exacerbate disparities for already underserved or medically vulnerable populations [15,16].

One of the most challenging and critical tasks in telemedicine research is to define broadband access. In the United States, the most common publicly available data comes from the FCC and is reported directly by fixed broadband internet providers on a biannual basis [17-21]. These data include the number of providers, speeds, and technologies offered within a census block. Although this detail can make the data prohibitively large, they are longitudinally available nationally and by state. Additionally, in the timeframe of this research, it is well known that the FCC data likely overrepresent broadband coverage as internet service providers report speed and technology combinations aggregated at the census block [22]. Alternatively, the Connect2Health platform links December 2015 broadband data with health and demographic data from a variety of sources in a preprocessed, ready-to-use structure [10,23], although these data have become somewhat dated.

Beyond the FCC, the publicly available Microsoft US broadband use data measure the percentage of a county's population using the internet at a minimum download speed of 25 Mbps [19]. The FCC data originate from internet providers (eg, Comcast, Mediacom) and can be considered a measure of availability, whereas the Microsoft data aggregate broadband speeds people are actively using, illustrating broadband accessibility. This can be thought of as a difference between an ideal internet landscape without individual-level influences (FCC) versus actual use (Microsoft), which is inherently influenced by unmeasurable outside factors. This is especially true as reports to the FCC during our study period required broadband providers to report census blocks where a given technology and speed was delivered to at least 1 location, not necessarily the entire census block. This means that FCC-defined broadband availability may not be universally or equally available to all residents of a census block [9]. Although Microsoft data improve upon the measurement of "real" access, there are also several limitations. Most importantly, these data require Microsoft products to be in use, which may bias results toward those able to afford a Microsoft license. In the same regard, this "downstream" measurement may not reflect the highest speed one could purchase due to perceived speed needs, desirability, or affordability issues. Finally, other issues such as internet traffic, user download bandwidth limits, and technology type (eg, digital subscriber line, cable modem, and fiber) may lead to discrepancies in broadband speed measurement.

To our knowledge, only 1 published research study has compared the divergence between the FCC and Microsoft broadband measures to assess the relationship between county-level social vulnerability and broadband access [24]. The goal of this study adds to this effort by comparing (1) FCC and (2) Microsoft estimates of broadband access and further by applying both broadband measures to the evaluation of VHA mental health video visit use as a case study of how the definition of broadband access may impact health services research. We hypothesized that (1) higher broadband access would be estimated using FCC versus Microsoft data and (2) veterans living in counties with lower broadband access would have fewer video mental health visits. Identifying how the definition of broadband speeds may influence research conclusions will inform the development of more robust research methodologies, aid in developing targeted health system initiatives to reduce the digital divide, and ultimately improve patient access to care. Accurate measurement at sufficiently granular and temporal levels is critical for telemedicine research, as it has implications for broadband improvements, government policy, health care, work, and day-to-day life in the digital age.

XSL•FO
RenderX

Methods

Study Design and Data Sources

This is a retrospective, cross-sectional, observational, cohort study of VHA mental health outpatient visits between January 1, 2019, and December 31, 2020. We chose mental health care as a case study of how 2 different national broadband measures may impact the assessment of the association between broadband access and telemedicine use because studies have shown that 58% of outpatient mental health visits in VHA are provided through video-based means [15]. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline [25].

Patient-level data, including patient demographics and outpatient mental health visits, were obtained from the Corporate Data Warehouse within the VA Informatics and Computing Infrastructure, an integrated electronic health records system including VHA electronic health records and administrative data. The 2010 Census Bureau TIGER/Line shapefile contains geographic entity codes (ie, census block, census block group, and county) [26]. Broadband access was obtained at the census block, the smallest geography of the US Census Bureau, for FCC fixed broadband data and at the county level for Microsoft US broadband use data [27]. The area deprivation index (ADI), a measure of neighborhood socioeconomic disadvantage, is reported in the census block group [28]. Both broadband and ADI data were spatially merged with the census shapefiles to overlay the latitude and longitude of each veteran's home address within a year to either the census block polygon for FCC broadband, the county polygon for Microsoft use data, or the census block group for the ADI. Population estimates are from the 2019 and 2020 census block-level estimates created by FCC staff based on public data from the US Census Bureau [29], while the estimate of veteran enrollees residing in each county is from the VA National Center for Veterans Analysis and Statistics [30].

Patient Population

The study cohort includes veterans aged 18 and older with a US home address having at least 1 VHA mental health visit between January 1, 2019, and December 31, 2020. Visits were classified as occurring in person, by telephone, or by video. All veterans receiving VHA mental health services during the study period were categorized based on their history of ever using video-based mental health care during the study period. The video cohort is thus all veterans who experienced 1 or more video-based mental health visits. We excluded care received at residential rehabilitation centers, nursing homes, or domiciliary visits, as well as care received in US territories and freely associated states.

Broadband Measures

Our primary exposure was nationwide county-level broadband access identified by (1) FCC or (2) Microsoft. FCC fixed broadband data originate from internet service providers, who report available broadband speeds by census block and technology type to the FCC. These data are available biannually in June and December. A total of 2 reporting timeframes were selected: December 2019 [31] and June 2020 [32]. To match the format of the Microsoft data, FCC data were aggregated at the county level (Multimedia Appendix 1), excluding satellite technology. Satellite technology was excluded because although it is widely available at the time of this study, it suffers from interference and delays, has relatively low subscription rates, and could be prohibitively expensive [22,33,34]. To be considered adequate, FCC data were required to demonstrate ≥25 Mbps download and ≥3 Mbps upload speeds, in line with speed recommendations for videoconferencing [2]. This FCC availability percentage can be interpreted as the percent of the county population (ie, 0%-100%) with access to adequate fixed terrestrial broadband download and upload speeds.

The Microsoft data originate from the Microsoft Airband Initiative and report the percent of each county's population with download speeds >25 Mbps as measured using the time required to download Microsoft products and updates [27]. Due to the nature of software updates, upload speeds could not be assessed. We note this is slightly different than the calculation using FCC data, which uses both download and upload speeds to determine speed adequacy. The Microsoft data are available at 2 time points: November 2019 and October 2020. This Microsoft use percentage represents the percent of the county population (ie, 0%-100%) using adequate download broadband speeds (ie, >25 Mbps).

For both FCC and Microsoft data, broadband access percentages were matched to patient-level data by county using the latitude and longitude of the veteran's residential address, matched to the timeframe of the broadband data (eg, 2019 or 2020).

Video Visit Outcome

Mental health clinic visits were assessed by clinic stop codes with visits categorized as occurring either in-person, by telephone, or via video-based care. Each patient's mental health visits per quarter were identified, by modality and overall, and then aggregated by county of residence. Visits were also categorized as occurring prepandemic (from January 2019 to March 2020) or postpandemic onset (from April 2020 to December 2020).

Covariates

Patient demographics included age (as of March 2020), sex, race, ethnicity, and patient rurality. Patient rurality was identified using the geocoded location of the patient's home and dichotomized into urban and rural (ie, rural and highly rural) according to rural-urban commuting area classifications [35]. Race was categorized as Black or African American, other (including American Indian or Alaska Native, Asian, Multiracial, or Pacific Islander), and White. Ethnicity was reported as being Hispanic or not Hispanic. The patient census block was updated quarterly, based on the veteran residential location. Patient demographics and rurality were also summarized at the county level as a percentage.

Statistical Analysis

Initially, we describe the cohort of patients receiving VHA mental health care stratified by their use of mental health video visits during the study period (ie, video and no video cohorts).

XSL•FO
RenderX

To better understand the differences between FCC-based broadband availability and Microsoft-based broadband use, we compare differences over time within and between data sources using empirical cumulative distribution functions. These functions plot the broadband percentage on the x-axis and the cumulative percent of veterans living in a county with the corresponding broadband percentage on the y-axis. Similar plots illustrate FCC and Microsoft broadband percentages stratified by veteran rurality (ie, rural or urban) and video cohort status (ie, at least 1 video visit in the study period or not).

Mental health visit rates in 2019 and 2020 were visualized within the COVID-19 pandemic period by video cohort status and broadband access categorized into quartiles (ie, cutoffs at 25%, 50%, and 75%). Finally, to better understand the association between quarterly mental health visit rates and broadband access leading into and during the COVID-19 pandemic, we quantify this association using a multivariate generalized estimating equation with a negative binomial distribution, log link function, and an offset for the county population. A total of 2 models are generated: 1 using each broadband measure. Broadband access was aligned by study year (eg, 2019 or 2020). These models were adjusted for patient characteristics as listed previously; included a binary indicator variable for time before and after the onset of the COVID-19 pandemic, broadband percentage, and the interaction of these 2 variables; and implemented random effects to account for repeated measurement within the county.

The authors had full access to and take full responsibility for the integrity of the data. All analyses were conducted using the SAS Enterprise Guide (version 8.2, 64-bit; SAS Institute).

Ethical Considerations

This work was completed as a quality improvement evaluation and was designated as nonresearch by the VHA Office of Connected Care (COR 20-199-05), the University of Iowa institutional review board (#202010441), and the VA Research Administration. This VHA documentation ensures work was not carried out under a human participants protocol and is part of institutionally sanctioned quality improvement activities. This study was conducted without direct patient contact using data routinely collected in the electronic health record and of minimal risk; therefore, a waiver of informed consent was obtained.

Results

Patient Population

Over the 2-year study period, 1,474,024 veterans were seen across 1161 VHA mental health clinics. Of these, 418,898 (28.4%) participated in at least 1 video visit in the study period (ie, video cohort). The video cohort compared to those without a video visit was less likely to be rural (115,421/418,898, 27.6% vs 346,945/1,055,126, 32.9%; $P<.001$) and more likely to be female (90,677/418,898, 21.7% vs 130,350/1,055,126, 12.4%; $P<.001$; Table 1). Further, veterans in the video cohort, on average, displayed higher broadband availability via FCC data (92.3% vs 91%; $P<.001$) and higher broadband use via Microsoft data (50.8% vs 47.1%; $P<.001$) in 2019. Similar differences were detected in 2020.

XSL•FO
RenderX

Table 1. Veteran demographics and broadband percentage by data source overall and by mental health video visit status.

Category	Overall cohort (n=1,474,024)	Video cohort[a] (n=418,898)	No video cohort (n=1,055,126)
Ever rural residence[b], n (%)	462,366 (31.4)	115,421 (27.6)	346,945 (32.9)
Age (years), mean (SD)	55.4 (16.0)	50.4 (14.9)	57.4 (16.0)
Sex, n(%)			
Female	221,027 (15)	90,677 (21.6)	130,350 (12.4)
Male	1,252,602 (85)	328,158 (78.4)	924,454 (87.6)
Race[c], n(%)			
Black or African American	329,879 (22.4)	94,616 (22.6)	235,263 (22.3)
White	989,465 (67.1)	274,971 (65.6)	714,494 (67.7)
Other	65,239 (4.4)	21,728 (5.2)	43,511 (4.1)
Unknown or missing	89,441 (6.1)	27,583 (6.6)	61,858 (5.9)
Ethnicity[d], n (%)			
Hispanic or Latino	115,063 (7.8)	40,897 (9.8)	74,166 (7)
Not Hispanic or Latino	1,305,467 (88.6)	361,968 (86.4)	943,499 (89.4)
Other	53,494 (3.6)	16,033 (3.8)	37,461 (3.6)
ADI[e], mean (SD)			
National rank	53.9 (25.3)	50.1 (24.8)	55.4 (25.4)
State rank	5.6 (2.7)	5.3 (2.7)	5.7 (2.7)
FCC[f] broadband availability (%)			
Mean (SD)			
2019	91.3 (12.5)	92.3 (11.2)	91.0 (12.9)
2020	92.3 (11.2)	93.1 (10.1)	92.0 (11.6)
Median (IQR)			
2019	95.7 (8.1)	96.0 (6.5)	95.7 (8.8)
2020	96.2 (6.9)	96.4 (6.1)	96.2 (7.6)
Microsoft broadband use (%)			
Mean (SD)			
2019	48.2 (18.1)	50.8 (17.4)	47.1 (18.2)
2020	61.4 (18.6)	64.2 (17.5)	60.2 (18.9)
Median (IQR)			
2019	51.0 (26.0)	54.0 (24.0)	49.0 (26.0)
2020	64.7 (23.4)	68.5 (22.4)	64.1 (23.4)

[a]Differences between the video and no video cohorts are significant at $P<.001$ for all variables, where continuous variables were assessed using a 2-tailed *t* test and categorical variable using a two-sided chi-square test.

[b]Includes patients who lived in a rural area at any time during the study period.

[c]Other races include American Indian or Alaskan Native, Asian, Multiracial, or Pacific Islander categories.

[d]Other ethnicity includes missing, multiple ethnicity, and unknown categories.

[e]ADI: area deprivation index; ranks neighborhoods by socioeconomic disadvantage on a scale of 0-100 with lower rankings indicating less social disadvantage.

[f]FCC: Federal Communications Commission.

Broadband Measures

Next, we plotted the empirical cumulative distribution functions, which represent the cumulative percent of veterans living in a county with a specified broadband percentage, split by data set and year (Figure 1). For example, comparing the 2020 data from FCC (blue) and Microsoft (green), 1.4% (21,338/1,474,024) versus 22.4% (n=327,584/1,474024) of veterans live in a county

XSL•FO
RenderX

with a broadband percentage of 50% or less. We observed an overall increase in both broadband availability and use, although increases in use were more pronounced.

Figure 1. Empirical cumulative distribution functions of broadband penetration rates among veterans using mental health services by data set and year. FCC: Federal Communications Commission.

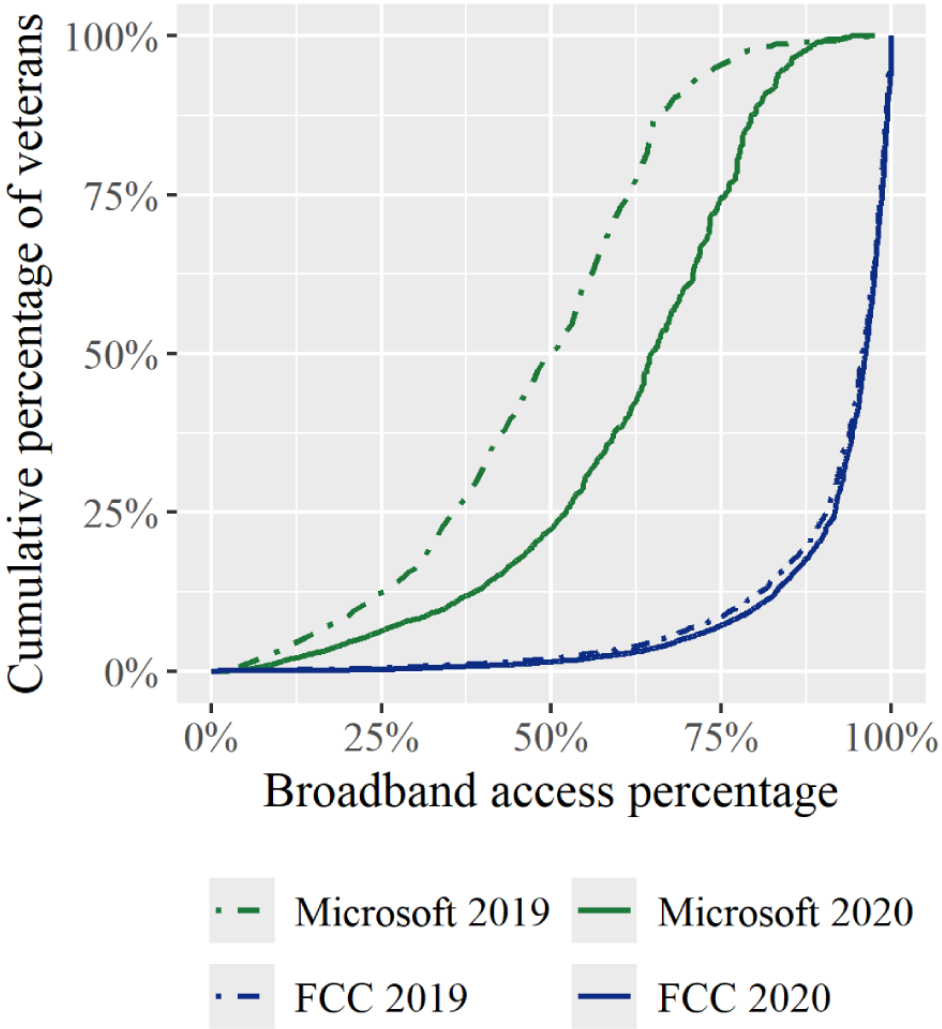

Video Visit Outcome

Stratifying the 2020 veteran broadband percentages by video cohort and rurality (Figure 2) we note that, compared to veterans without video visits, veterans who had at least 1 video visit had greater broadband access. Regardless of broadband data source, veterans in urban areas had higher broadband percentages than those living in rural areas.

XSL•FO
RenderX

Figure 2. Empirical cumulative distribution functions of veteran penetration rates of 2020 data set stratified by video cohort. FCC: Federal Communications Commission.

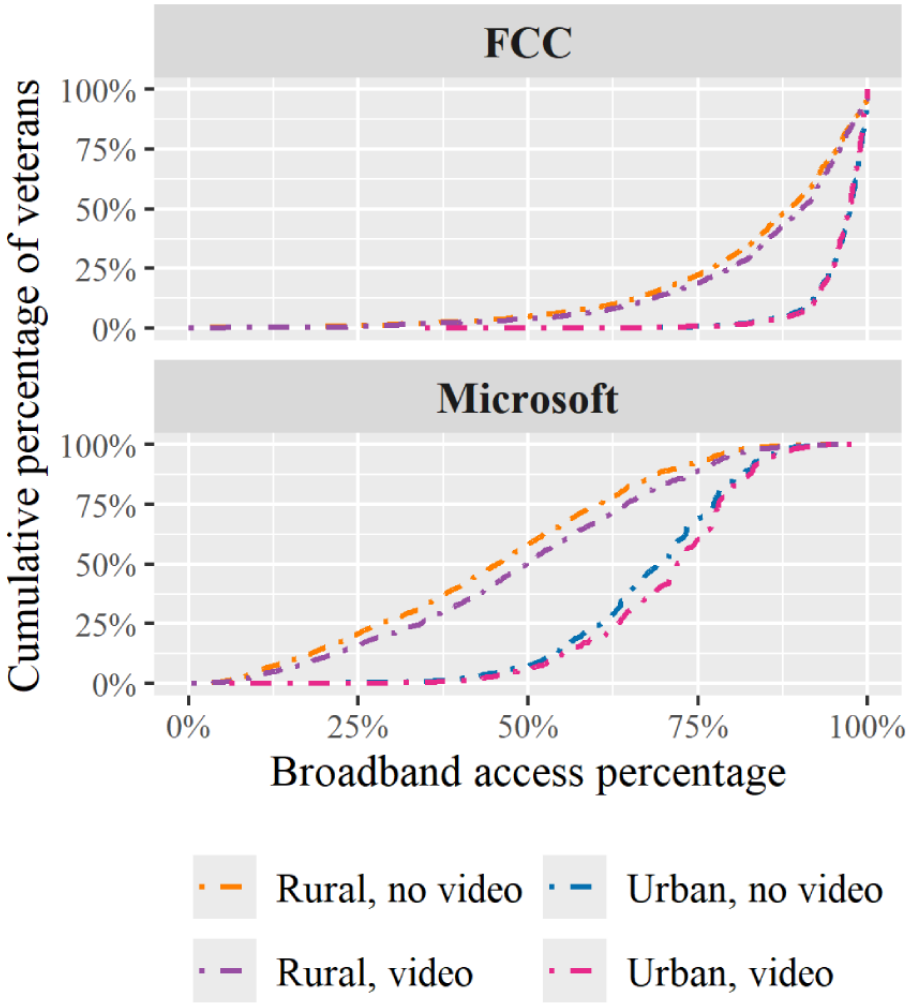

Similarly, as broadband access increased, a higher proportion of veterans used video visits (Multimedia Appendix 2). Based on our findings, we estimate that of the 19.2 million living veterans in 2020, between 1.4 million (7.3%) and 7.2 million (37.4%) veterans do not have access to sufficient broadband to participate in video telemedicine (Multimedia Appendix 3).

After the onset of the COVID-19 pandemic, there has been a clear association between group and mental health visit use, where those living in a county with a higher broadband percentage have a higher mental health visit rate (Figure 3). Both before and during the COVID-19 pandemic, the video cohort had a higher quarterly rate of mental health visits compared to the nonvideo cohort. Furthermore, among the video cohort, the mental health visit rate doubled compared to before the COVID-19 pandemic.

Finally, 2 longitudinal multivariate generalized estimating equation models were developed identically but used either FCC or Microsoft broadband percentages to evaluate the association between broadband access and quarterly mental health video visit use (Table 2).

XSL•FO
RenderX

Figure 3. Overall quarterly mental health visit rate by data set, penetration rate, video cohort, and pandemic time period. FCC: Federal Communications Commission.

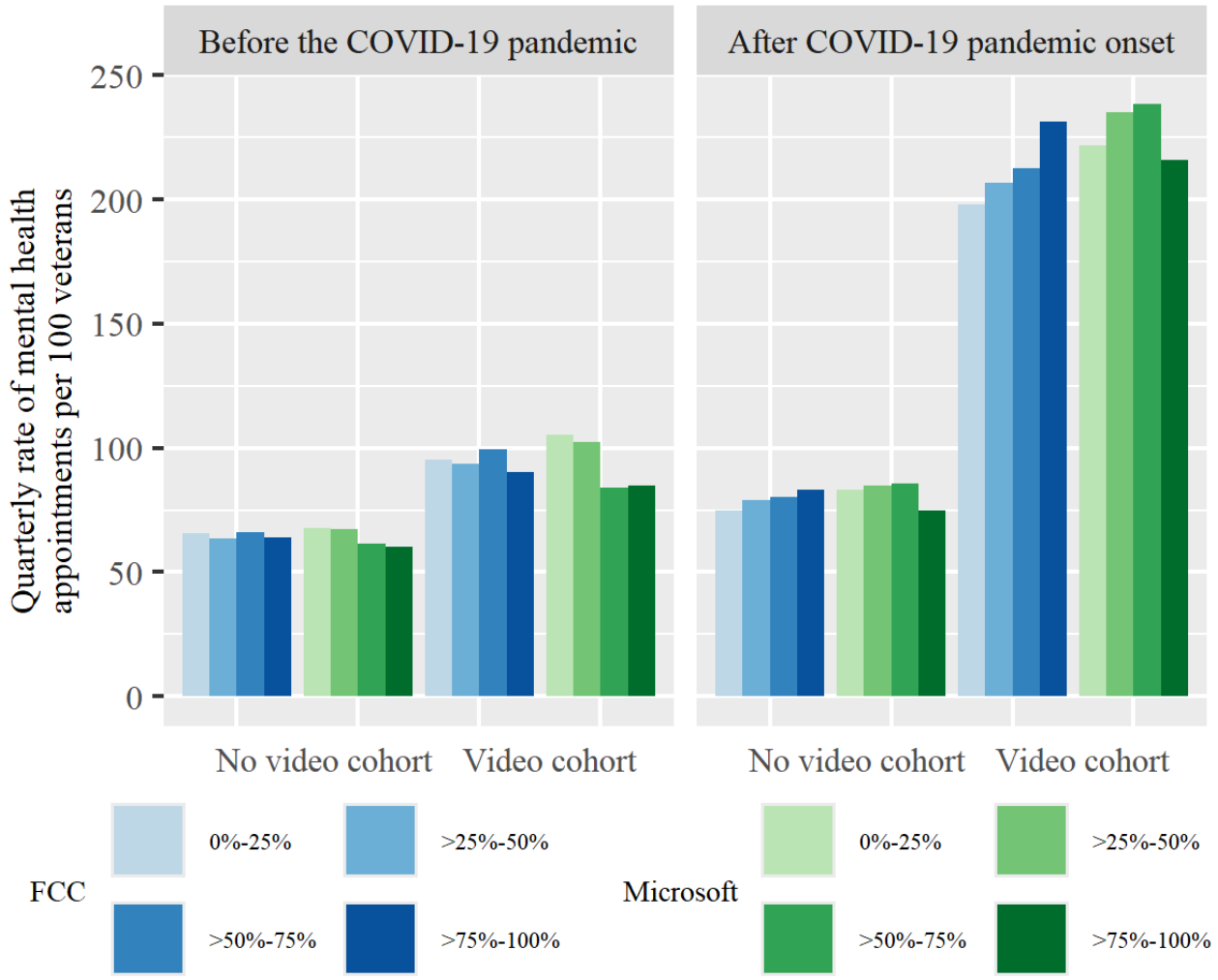

Table 2. Multivariate negative binomial regression models using generalized estimating equations to predict quarterly mental health video visit rate by county for each broadband data source.

Pandemic indicator (after vs before)[a]	Federal Communications Commission model; IRR[b] (95% CI)	Microsoft broadband use model; IRR (95% CI)
At broadband percentage (%)		
10	5.74 (4.85-6.80)	6.05 (5.67-6.45)
20	5.91 (5.10-6.84)	6.24 (5.93-6.57)
30	6.08 (5.36-6.89)	6.43 (6.16-6.72)
40	6.25 (5.64-6.93)	6.64 (6.36-6.92)
50	6.43 (5.92-6.98)	6.84 (6.53-7.17)
60	6.62 (6.21-7.04)	7.06 (6.66-7.48)
70	6.81 (6.49-7.13)	7.28 (6.78-7.81)
80	7.00 (6.74-7.28)	7.51 (6.89-8.18)
90	7.20 (6.90-7.51)	7.74 (6.99-8.57)

[a]The models were adjusted for rural residency, race, and Hispanic ethnicity summarized at the county level as percentages, as well as average county-level area deprivation index and age.

[b]IRR: incidence rate ratio.

Using FCC model estimates, given all other covariates are constant and assuming the FCC broadband percentage was set at 70%, the incident rate ratio (IRR) of county-level quarterly mental health video visits during the pandemic was 6.81 times

XSL•FO
RenderX

(95% CI 6.49-7.13) the rate before the pandemic. In comparison, the Microsoft model estimates a higher rate of mental health video visits during versus before the pandemic at the same broadband percentage (IRR 7.28; 95% CI 6.78-7.81). This was the case for all broadband percentages assessed.

Discussion

Principal Findings

In this study, we found that (1) the measurement of county-level broadband access varied by broadband data source; (2) regardless of data source, overall, adequate broadband access improved over time, although these gains were most pronounced in the measurement of Microsoft use; and (3) there was a positive association between county-level broadband access percentage and mental health video visits during compared to before the pandemic, but the strength of this association varied based on which broadband access measure was used. As health care systems continue to expand their use of video-based care, our findings highlight the importance of accurately defining and measuring broadband access from both research and clinical standpoints.

Comparison With Prior Work

While telemedicine is known to increase access to care, studies have suggested that a patient's ability to sufficiently access the internet may be a limiting factor [7,9,10,36,37]. Importantly, access to the internet is no longer the only defining feature of taking part in the digital age, but instead depends on broadband technology, speeds, and—at the individual level—economic cost, digital literacy, the number of simultaneous users, and other issues. For researchers, the measurement of broadband availability using FCC data is flexible regarding questions about varying speed levels and technology types (eg, digital subscriber line, cable modem, and fiber) and can be aggregated to a variety of census-based geographies. The largest hurdle to processing the FCC data is the sheer size of the files, although our study team has created a publicly available R (version 4.4.0; R Foundation for Statistical Computing) package to facilitate download and processing [38]. In comparison, Microsoft developed and released its broadband use data in a ready-to-use file, over which researchers have no control. Choosing which of these data sources is best suited to a particular research question will likely depend on the question of interest, the required control over additional broadband-specific features, the data management skills of the study team, and the desire to control for unmeasurable social determinants that influence internet access.

These findings also have clinical implications. After the onset of the COVID-19 pandemic, the VHA and other health care systems experienced a major shift to web-based mental health care. However, the most cited barriers to adoption were unreliable broadband, insufficient internet speeds, and technological barriers at the user level [39-42]. The VHA has attempted to address these barriers by providing internet-connected devices to patients meeting specific criteria [43], negotiating with telecommunication companies to provide free unlimited data while using VHA video telemedicine [44], and initiating test calls prior to a video visit to ensure the veteran knows how to use the VA Video Connect platform [45]. As health care systems and providers shift more of their care to video telemedicine, accurate measurement of broadband availability and access is crucial to improving health care system interventions adopted to close the digital divide, as well as the allocation of federal funding for the development of broadband infrastructure across the United States [46]. Targeting areas where infrastructure investment has been impacted by low population density (eg, fully rural counties) or geographic challenges (ie, mountains or valleys) may be a critical factor in improving equitable access to telemedicine [18]. Importantly, although increasing the availability of telemedicine opens additional avenues to care, limited access to adequate broadband may favor certain groups in unintended ways and worsen existing disparities among rural or already underserved populations [47].

Limitations

This research has limitations. First, we did not assess causal relationships between video visit use and broadband percentages at the county level, nor did we account for patient comorbidities, familiarity with web-based care, or other factors related to telemedicine use. However, we highlight that we did adjust for age and other common sociodemographic factors that prior work has shown to be highly associated with telemedicine readiness and use [48]. Second, we were unable to distinguish when a video visit could not be completed, leading to a possible overcount of video visits. Third, this work focused only on mental health-based care. However, we contend that limitations in broadband access are likely generalizable to other types of care (eg, primary and subspecialty care), and thus we would expect similar findings if other types of care were assessed. Finally, data inadequacies, including our inability to account for cellular coverage areas may not accurately reflect those veterans using mobile versus terrestrial broadband connections.

Conclusions

Despite limitations, this work is the first we are aware of to consider the use of FCC and Microsoft broadband data to triangulate the uneven impact of broadband access and the use of video mental health services. As health care systems expand video telemedicine, it is important to accurately measure where broadband infrastructure and related interventions can make the most impact and diminish access barriers to using this care modality. Future research should use the improved and more granular broadband data that are currently being collected by the FCC, investigate how differential broadband access may impact health care use, and whether increased access to care through video telemedicine has a beneficial impact on health care outcomes.

XSL•FO
RenderX

Acknowledgments

This material is based upon work supported (or supported in part) by the Department of Veterans Affairs (VA), Veterans Health Administration (VHA), VA Office of Connected Care (COR 20-199-05), and the Office of Research and Development, Health Services Research and Development (HSR&D) Service through the Comprehensive Access and Delivery Research and Evaluation (CADRE) Center (CIN 13-412). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The data sets generated and analyzed are not publicly available due to VHA privacy and confidentiality requirements but are available to VHA researchers from the corresponding author on reasonable request. This work was designated nonresearch since it was carried out as a quality improvement evaluation under the terms of a signed attestation of nonresearch from the VHA Office of Connected Care. This VHA documentation ensures that work not carried out under a human subjects protocol is part of institutionally sanctioned quality improvement activities. The views expressed in this paper are those of the authors and do not necessarily represent the views of the Department of VA or the US government.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The protocol for calculating county-level penetration rates using Federal Communications Commission broadband data reported at the census block.
[PDF File (Adobe PDF File), 83 KB - jmir_v26i1e47100_app1.pdf]

Multimedia Appendix 2
Percentage of veterans in the video use cohort and non–video use cohort within quintiles of 2019 penetration rate stratified by data set.
[PDF File (Adobe PDF File), 101 KB - jmir_v26i1e47100_app2.pdf]

Multimedia Appendix 3
The estimation of enrolled veterans lacking adequate broadband for a video telemedicine visit by year and data source.
[PDF File (Adobe PDF File), 72 KB - jmir_v26i1e47100_app3.pdf]

References

1. Baum A, Kaboli PJ, Schwartz MD. Reduced in-person and increased telehealth outpatient visits during the COVID-19 pandemic. Ann Intern Med 2021 Jan;174(1):129-131 [FREE Full text] [doi: 10.7326/M20-3026] [Medline: 32776780]
2. Broadband speed guide. Federal Communications Commission. 2022. URL: https://www.fcc.gov/sites/default/files/broadband_speed_guide.pdf [accessed 2020-02-05]
3. Household broadband guide. Federal Communications Commission. 2020. URL: https://www.fcc.gov/consumers/guides/household-broadband-guide [accessed 2021-10-15]
4. Bennet MF, King AS, Portman R, Manchin J. Bipartisan senator broadband speed letter. United States Senate. 2021 Mar 04. URL: https://www.bennet.senate.gov/public/_cache/files/c/7/c76028fb-488d-498e-8506-7d8a2dce3172/05DDC9148CC7F12A9F09235F77BB7A0D.bipartisan-broadband-speed-letter.pdf [accessed 2023-06-12]
5. Brake D, Bruer A. Broadband myths: are high broadband prices holding back adoption? Information Technology & Innovation Foundation. 2021. URL: https://itif.org/publications/2021/02/08/broadband-myths-are-high-broadband-prices-holding-back-adoption/ [accessed 2021-02-08]
6. Mishori R, Antono B. Telehealth, Rural America, and the digital divide. J Ambul Care Manage 2020;43(4):319-322. [doi: 10.1097/JAC.0000000000000348] [Medline: 32858734]
7. Lai J, Widmar NO. Revisiting the digital divide in the COVID-19 era. Appl Econ Perspect Policy 2021 Mar;43(1):458-464 [FREE Full text] [doi: 10.1002/aepp.13104] [Medline: 33230409]
8. Serrano‐Cinca C, Muñoz‐Soro JF, Brusca I. A multivariate study of internet use and the digital divide*. Soc Sci Q 2018 Feb 27;99(4):1409-1425. [doi: 10.1111/ssqu.12504]
9. Hollman AK, Obermier TR, Burger PR. Rural measures: a quantitative study of the rural digital divide. J Inf Policy 2021;11:176-201. [doi: 10.5325/jinfopoli.11.2021.0176]
10. Drake C, Zhang Y, Chaiyachati KH, Polsky D. The limitations of poor broadband internet access for telemedicine use in Rural America: an observational study. Ann Intern Med 2019 Sep 03;171(5):382-384. [doi: 10.7326/M19-0283] [Medline: 31108509]
11. Rural veterans. US Department of Veterans Affairs. 2021. URL: https://www.ruralhealth.va.gov/aboutus/ruralvets.asp [accessed 2022-08-16]

XSL•FO
RenderX

12. Report on promoting internet access service for veterans. Federal Communications Commission. 2019. URL: https://www.fcc.gov/document/report-promoting-internet-access-service-veterans [accessed 2022-08-16]
13. Connolly SL, Stolzmann KL, Heyworth L, Sullivan JL, Shimada SL, Weaver KR, et al. Patient and provider predictors of telemental health use prior to and during the COVID-19 pandemic within the Department of Veterans Affairs. Am Psychol 2022;77(2):249-261 [FREE Full text] [doi: 10.1037/amp0000895] [Medline: 34941310]
14. Hsiao V, Chandereng T, Lankton RL, Huebner JA, Baltus JJ, Flood GE, et al. Disparities in telemedicine access: a cross-sectional study of a newly established infrastructure during the COVID-19 pandemic. Appl Clin Inform 2021 May;12(3):445-458 [FREE Full text] [doi: 10.1055/s-0041-1730026] [Medline: 34107542]
15. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the Veterans Health Administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc 2021 Mar 01;28(3):453-462 [FREE Full text] [doi: 10.1093/jamia/ocaa284] [Medline: 33125032]
16. Adams SV, Mader MJ, Bollinger MJ, Wong ES, Hudson TJ, Littman AJ. Utilization of interactive clinical video telemedicine by rural and urban veterans in the veterans health administration health care system. J Rural Health 2019 Jun;35(3):308-318. [doi: 10.1111/jrh.12343] [Medline: 30600557]
17. Cortelyou-Ward K, Atkins DN, Noblin A, Rotarius T, White P, Carey C. Navigating the digital divide: barriers to telehealth in rural areas. J Health Care Poor Underserved 2020;31(4):1546-1556. [doi: 10.1353/hpu.2020.0116] [Medline: 33416736]
18. Wilcock AD, Rose S, Busch AB, Huskamp HA, Uscher-Pines L, Landon B, et al. Association between broadband internet availability and telemedicine use. JAMA Intern Med 2019 Nov 01;179(11):1580-1582 [FREE Full text] [doi: 10.1001/jamainternmed.2019.2234] [Medline: 31355849]
19. Gallardo R, Whitacre B, Kumar I, Upendram S. Broadband metrics and job productivity: a look at county-level data. Ann Reg Sci 2020 Jul 30;66(1):161-184. [doi: 10.1007/s00168-020-01015-0]
20. Jacobs J, Ferguson JM, van Campen J, Yefimova M, Greene L, Heyworth L, et al. Organizational and external factors associated with video telehealth use in the veterans health administration before and during the COVID-19 pandemic. Telemed J E Health 2022 Feb;28(2):199-211. [doi: 10.1089/tmj.2020.0530] [Medline: 33887166]
21. Mack EA, Helderop E, Ma K, Grubesic TH, Mann J, Loveridge S, et al. A broadband integrated time series (BITS) for longitudinal analyses of the digital divide. PLoS One 2021;16(5):e0250732 [FREE Full text] [doi: 10.1371/journal.pone.0250732] [Medline: 34038407]
22. Fries J, Doan T, Gupta R, Bajpai V, editors. An eight years perspective on the internet broadband infrastructure in the USA. 2022 Presented at: 2022 IFIP Networking Conference (IFIP Networking); June 13-16, 2022; Catania, Italy. [doi: 10.23919/ifipnetworking55013.2022.9829788]
23. Jella TK, Cwalina TB, Schmidt JE, Wu VS, Haglin JM, Kamath AF. Do Americans lacking a local orthopaedic surgeon have adequate broadband for telehealth? Clin Orthop Relat Res 2023 Feb 01;481(2):347-355 [FREE Full text] [doi: 10.1097/CORR.0000000000002374] [Medline: 36040749]
24. Acharya M, Shoults CC, Hayes CJ, Brown CC. Association between broadband capacity and social vulnerability factors in the United States: a county-level spatial analysis. Popul Health Manag 2022 Dec;25(6):798-806. [doi: 10.1089/pop.2022.0189] [Medline: 36450124]
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
26. TIGER/Line shapefiles. US Census Bureau. 2021. URL: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html [accessed 2021-02-28]
27. Kahan J, Ferres JL. United States Broadband Usage Percentages Dataset. URL: https://github.com/microsoft/USBroadbandUsagePercentages [accessed 2021-10-13]
28. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood atlas. N Engl J Med 2018 Jun 28;378(26):2456-2458 [FREE Full text] [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
29. Staff block estimates. Federal Communications Commission. 2019. URL: https://www.fcc.gov/staff-block-estimates [accessed 2022-02-03]
30. Table 9L: VetPop2020 county-level veteran population by state, age group, gender, 2020-2050. U.S. Department of Veterans Affairs. 2023. URL: https://www.va.gov/vetdata/Veteran_Population.asp [accessed 2023-01-13]
31. Form 477 broadband deployment data—December 2019 (version 1). Federal Communications Commission. 2019. URL: https://www.fcc.gov/form-477-broadband-deployment-data-december-2019-version-1 [accessed 2021-05-14]
32. Form 477 broadband deployment data—June 2020 (version 1). Federal Communications Commission. 2020. URL: https://www.fcc.gov/form-477-broadband-deployment-data-june-2020-version-1 [accessed 2021-05-14]
33. Fourteenth broadband deployment report. Federal Communications Commission. 2021. URL: https://www.fcc.gov/reports-research/reports/broadband-progress-reports/fourteenth-broadband-deployment-report [accessed 2022-02-03]
34. Mack EA, Helderop E, Keene T, Loveridge S, Mann J, Grubesic TH, et al. A longitudinal analysis of broadband provision in tribal areas. Telecommun Policy 2022 Jun;46(5):102333. [doi: 10.1016/j.telpol.2022.102333]
35. Rural-urban commuting area codes. University of Washington Rural Health Research Center and USDA Economic Research Service (ERS). 2018. URL: http://depts.washington.edu/uwruca/ruca-uses.php [accessed 2018-12-13]

XSL•FO
RenderX

36. Ali MM, Ghertner R. Broadband access and telemedicine adoption for opioid use disorder treatment in the United States. J Rural Health 2023 Jan;39(1):233-239. [doi: 10.1111/jrh.12699] [Medline: 35838414]
37. Okoye SM, Mulcahy JF, Fabius CD, Burgdorf JG, Wolff JL. Neighborhood broadband and use of telehealth among older adults: cross-sectional study of national survey data linked with census data. J Med Internet Res 2021 Jun 14;23(6):e26242 [FREE Full text] [doi: 10.2196/26242] [Medline: 34125071]
38. Mulligan K. Package to process FCC fixed broadband datasets. processFCC. R: A language and environment for statistical computing: R Core Team; 2022. URL: https://github.com/kdmulligan/processFCC [accessed 2022-11-11]
39. Hermes EDA, Burrone L, Heapy A, Martino S, Perez E, Rosenheck R, et al. Beliefs and attitudes about the dissemination and implementation of internet-based self-care programs in a large integrated healthcare system. Adm Policy Ment Health 2019 May;46(3):311-320. [doi: 10.1007/s10488-018-0913-7] [Medline: 30600402]
40. Story KM, Flanagan ME, Brown HA, Robb SL, Damush TM, Otto D. Veterans hospital administration telehealth utilization for recreation and creative arts therapies: a brief report. Telemed e-Health 2022 May;28(5):752-757 [FREE Full text] [doi: 10.1089/tmj.2021.0363] [Medline: 34432540]
41. Ganapathy A, Clough BA, Casey LM. Organizational and policy barriers to the use of digital mental health by mental health professionals. Telemed e-Health 2021 Dec;27(12):1332-1343. [doi: 10.1089/tmj.2020.0455] [Medline: 33646057]
42. Moreau JL, Cordasco KM, Young AS, Oishi SM, Rose DE, Canelo I, et al. The use of telemental health to meet the mental health needs of women using department of veterans affairs services. Womens Health Issues 2018;28(2):181-187. [doi: 10.1016/j.whi.2017.12.005] [Medline: 29339013]
43. Bridging the digital divide. US Department of Veteran Affairs. URL: https://telehealth.va.gov/digital-divide [accessed 2021-09-15]
44. VA's telehealth system grows as veterans have access to unlimited data while using VA video connect. US Department of Veteran Affairs. 2020. URL: https://www.va.gov/opa/pressrel/pressrelease.cfm?id=5426 [accessed 2021-03-01]
45. Wray CM, Sridhar A, Young A, Noyes C, Smith WB, Keyhani S. Assessing the impact of a pre-visit readiness telephone call on video visit success rates. J Gen Intern Med 2023 Jan;38(1):252-253 [FREE Full text] [doi: 10.1007/s11606-022-07627-w] [Medline: 35501627]
46. Varn J, Gong L. What is the FCC's new broadband map and why does it matter?: the effort to highlight where high-speed internet is and is not available will affect state funding. The Pew Charitable Trusts. 2022. URL: https://www.pewtrusts.org/en/research-and-analysis/articles/2022/11/18/what-is-the-fccs-new-broadband-map-and-why-does-it-matter [accessed 2022-11-18]
47. Summers-Gabr NM. Rural-urban mental health disparities in the United States during COVID-19. Psychol Trauma 2020 Aug;12(S1):S222-S224. [doi: 10.1037/tra0000871] [Medline: 32478541]
48. Wray CM, Tang J, Shah S, Nguyen OK, Keyhani S. Sociodemographics, social vulnerabilities, and health factors associated with telemedicine unreadiness among US adults. J Gen Intern Med 2022 May;37(7):1811-1813 [FREE Full text] [doi: 10.1007/s11606-021-07051-6] [Medline: 34331214]

Abbreviations

ADI: area deprivation index
FCC: Federal Communications Commission
IRR: incident rate ratio
Mbps: megabits per second
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VA: Veterans Affairs
VHA: Veterans Health Administration

Edited by T Leung, K Williams; submitted 08.03.23; peer-reviewed by A Wilcock, A Hollman, J Prieger; comments to author 04.03.24; revised version received 27.04.24; accepted 11.06.24; published 08.08.24.

Please cite as:
O'Shea AMJ, Mulligan K, Carter KD, Haraldsson B, Wray CM, Shahnazi A, Kaboli PJ
Comparing Federal Communications Commission and Microsoft Estimates of Broadband Access for Mental Health Video Telemedicine Among Veterans: Retrospective Cohort Study
J Med Internet Res 2024;26:e47100
URL: https://www.jmir.org/2024/1/e47100
doi:10.2196/47100
PMID:

XSL•FO
RenderX

©Amy MJ O'Shea, Kailey Mulligan, Knute D Carter, Bjarni Haraldsson, Charlie M Wray, Ariana Shahnazi, Peter J Kaboli. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Maternal Health Dashboards in the United States: Descriptive Assessment

Jennifer A Callaghan-Koru[1,2], MHS, PhD; Paige Newman Chargois[3], MPH; Tanvangi Tiwari[1], MPP; Clare C Brown[3], MPH, PhD; William Greenfield[2,4], MBA, MD; Güneş Koru[1], PhD

[1]Fay W Boozman College of Public Health, University of Arkansas for Medical Sciences, Springdale, AR, United States
[2]Department of Obstetrics & Gynecology, University of Arkansas for Medical Sciences, Little Rock, AR, United States
[3]Fay W Boozman College of Public Health, University of Arkansas for Medical Sciences, Little Rock, AR, United States
[4]Family Health Branch, Arkansas Department of Health, Little Rock, AR, United States

Corresponding Author:
Jennifer A Callaghan-Koru, MHS, PhD
Fay W Boozman College of Public Health
University of Arkansas for Medical Sciences
2708 S. 48th St.
Springdale, AR, 72762
United States
Phone: 1 479 713 8102
Email: jck@uams.edu

Abstract

Background: Data dashboards have become more widely used for the public communication of health-related data, including in maternal health.

Objective: We aimed to evaluate the content and features of existing publicly available maternal health dashboards in the United States.

Methods: Through systematic searches, we identified 80 publicly available, interactive dashboards presenting US maternal health data. We abstracted and descriptively analyzed the technical features and content of identified dashboards across four areas: (1) scope and origins, (2) technical capabilities, (3) data sources and indicators, and (4) disaggregation capabilities. Where present, we abstracted and qualitatively analyzed dashboard text describing the purpose and intended audience.

Results: Most reviewed dashboards reported state-level data (58/80, 72%) and were hosted on a state health department website (48/80, 60%). Most dashboards reported data from only 1 (33/80, 41%) or 2 (23/80, 29%) data sources. Key indicators, such as the maternal mortality rate (10/80, 12%) and severe maternal morbidity rate (12/80, 15%), were absent from most dashboards. Included dashboards used a range of data visualizations, and most allowed some disaggregation by time (65/80, 81%), geography (65/80, 81%), and race or ethnicity (55/80, 69%). Among dashboards that identified their audience (30/80, 38%), legislators or policy makers and public health agencies or organizations were the most common audiences.

Conclusions: While maternal health dashboards have proliferated, their designs and features are not standard. This assessment of maternal health dashboards in the United States found substantial variation among dashboards, including inconsistent data sources, health indicators, and disaggregation capabilities. Opportunities to strengthen dashboards include integrating a greater number of data sources, increasing disaggregation capabilities, and considering end-user needs in dashboard design.

(J Med Internet Res 2024;26:e56804) doi:10.2196/56804

KEYWORDS

dashboard; maternal health; data visualization; data communication; perinatal health

XSL•FO
RenderX

Introduction

Background

Data dashboards are "visual displays that feature the most important information needed to achieve specific goals captured on a single screen" [1]. The term dashboard is borrowed from the vehicle dashboard, highlighting the expectation that a data dashboard will display key indicators used to understand performance and make decisions. Data dashboards can be designed to meet different types of goals, including informing strategic decisions, monitoring organizational performance, and communicating information to raise awareness and motivation [1-3]. Similar to businesses, health care organizations and public health agencies have widely adopted dashboards to support internal operations [4,5]. More recently, dashboards are being developed for the public communication of health-related data, a practice that accelerated during the COVID-19 pandemic [6,7]. Examples of public health dashboards include the City Health Dashboard [8], national and state substance use and overdose dashboards [9-11], and various COVID-19 dashboards [12,13].

The interest in public health dashboards has also expanded to the area of maternal health. The status of maternal health in the United States has been characterized as a crisis [14], with outcomes much worse than those in peer nations as well as large racial and ethnic disparities across perinatal outcomes [15,16]. National plans for improving maternal health have highlighted the need for better, more timely maternal health data [14,15]. Public reporting of maternal health data has also been proposed as a key mechanism for enabling quality improvement and empowering patient decision-making [17]. Consistent with these priorities, recent federal investments in maternal health have supported improving data systems. Several federal agencies are partnering on maternal data infrastructure projects to standardize measurement and expand surveillance to include underreported conditions and outcomes across settings [18]. At the state level, the Centers for Disease Control and Prevention (CDC) granted 39 awards to support maternal mortality review committees (MMRCs) in the collection and reporting of data on the causes of maternal deaths [19]. In the same year, the Health Resources and Services Administration awarded 9 new state maternal health innovation awards, with an emphasis on maternal health data. One of the primary suggested data-related innovations is the development of a state-focused maternal health dashboard "to easily access and report on maternal health outcomes" [20].

Maternal health data in the United States currently comes from multiple sources that are typically disconnected outside of limited data linkage efforts [18,21]. These include maternal mortality reviews conducted by state committees [22], national and state vital statistics systems [15], hospital discharge data [23], claims data [24], and surveys such as the Pregnancy Risk Assessment Monitoring System (PRAMS) [25]. Additional sources of health system, demographic, and social data may also identify risks and opportunities for improving maternal health. Maternal health dashboards have the potential to reduce access barriers to these data, particularly when integrating multiple sources of information in 1 location. Dashboards may also support efforts to address maternal health disparities through disaggregation and comparison features [26].

Alongside the growth of public health dashboards, questions have been raised regarding the extent to which dashboards present the most valuable data in a way that meets users' information needs [4,7,13]. Published assessments of dashboards in other health areas have identified substantial heterogeneity and gaps in their actionability [7,27]. Segmenting data presentations for different audiences is considered essential for the utility of data dashboards, as audiences may differ in their information needs as well as their health and data literacy levels [3,4]. However, dashboard developers are often disconnected from public users and may have a limited understanding of users' needs and abilities [4]. Given that there are few published reports on the development process for maternal health dashboards [28,29], little is known about the factors considered in maternal health dashboard design.

This Study

To inform the development of a maternal health dashboard for the state of Arkansas, we undertook a rigorous descriptive assessment of existing publicly available maternal health dashboards in the United States. This study, the first step in a user-centered design process, allowed the team to identify the range of data sources and technical features in existing dashboards of similar scope. This paper provides the results of our evaluation of the content and features of existing publicly available maternal health dashboards in the United States. Given the strong interest in improving access to maternal health data [17] and the importance of maintaining and continually improving current dashboards [6], the results of this review of public dashboards are relevant to all stakeholders concerned with the dissemination of maternal health data.

Methods

Overview

This descriptive assessment was conducted by a team of researchers with expertise in information systems, software engineering, maternal health surveillance, and quantitative and qualitative methods. Similar to scoping reviews of published literature [30], we sought to identify the scope and key attributes of public dashboards presenting maternal health data for the United States. The methods used for this assessment were also informed by prior studies of hospital patient safety dashboards [27] and public COVID-19 dashboards [7,31].

Identification of Dashboards

Between January and March 2023, the research team performed multiple searches of publicly available maternal health dashboards using the Google (Google LLC) search engine. The topical keywords included in each search were "dashboard," "maternal," "perinatal," and "birth." Given the geographic focus of the study, each search contained a geographic keyword, either "United States" or the name of 1 of the 50 states (eg, Alabama, Alaska, and Arizona). A total of 51 separate searches were performed, and each search was completed independently by 2 team members. Team members accessed and reviewed the first

XSL•FO
RenderX

10 results for each search they performed to determine whether the websites met the eligibility criteria.

We adopted the functional definition of a dashboard from the study by Sarikaya et al [3], "an interactive display that enables...monitoring of dynamically updating data," as the basis for our inclusion criteria. A dashboard was eligible for inclusion if it (1) reported data related to maternal health, (2) included data from the United States, (3) was accessible by the public, and (4) included at least 1 interactive feature for data presentation. A dashboard was considered ineligible if it (1) was presented on a password-protected, nonpublic website; (2) provided data only as downloadable files (eg, PDF reports and Excel spreadsheets); (3) did not contain any interactive features; (4) did not include any US data. All dashboards determined to be eligible by the team were included in a list for data extraction. Through the searches, 76 websites were identified. An additional 4 dashboards were identified during the review of the content of included dashboards and state health department websites and were added to the abstraction list, resulting in a total of 80 dashboards (Multimedia Appendix 1).

Abstraction of Dashboard Characteristics

The features and content of identified dashboards were abstracted using a standardized form in REDCap (Research Electronic Data Capture; Vanderbilt University). After a preliminary review of the first quarter of identified dashboards, the team developed an abstraction form to capture the variability in features and data observed across this subset of dashboards. Following the initial abstraction of all dashboards, the form underwent a revision to explicitly include dashboard features and content that were most commonly included in "other" categories, as described subsequently. The final form included >250 fields across four areas: (1) scope and origins, (2) technical capabilities, (3) data sources and indicators, and (4) disaggregation capabilities (Multimedia Appendix 2).

"Scope and origins" fields addressed the geographic areas and health topics included in the dashboard, the dashboard's stated purpose and audience, and the hosting organization or website. "Technical capabilities" fields included types of visualizations, export and download features, responsiveness and mobile friendliness, and available comparisons (eg, longitudinal trends and benchmarks). Both the "scope and origins" and "technical capabilities" fields were assessed for the website as a whole, except for visualizations, which were abstracted only for maternal health data. "Data sources and indicators" fields addressed the maternal health data sources explicitly reported on the dashboard as well as the presence, period, and format of specific maternal or perinatal indicators. A total of 40 indicators were abstracted into 4 categories: health status or outcomes (n=11, 28% indicators), health behaviors and use of health services (n=11, 28% indicators), individual characteristics and risk factors (n=12, 30% indicators), and health system characteristics (n=5, 12% indicators). "Disaggregation capabilities," also assessed for maternal health data, included fields for abstracting disaggregation by race, geography, and multiple perinatal characteristics. A variable that could be used for disaggregation was considered to additionally be an indicator on the dashboard only if the variable was also presented independently of disaggregation functions. Most fields on the abstraction form were assessed as binary variables, indicating the presence or absence of a data point or feature. "Other" options were included for each section on the abstraction form, with the text specification of the variable or feature observed. A small set of fields, related to the stated purpose and audience of each dashboard, also required the abstraction of textual information. Operational definitions were developed by the team for select abstraction fields that required interpretation by abstractors.

The original abstraction process was completed by 5 team members between February and May 2023. A total of 2 (40%) master's-trained team members were responsible for the primary abstraction of all dashboards. A total of 3 (60%) PhD-trained team members, each with at least 5 years of experience working with US maternal health data, completed a thorough secondary review for each dashboard. All corrections made during the secondary review were logged by REDCap, with a note of explanation from the reviewer. Given that website features may change or websites may become unavailable at any time, 1 team member also recorded brief videos demonstrating the primary features of each website at the time of abstraction. A second round of abstraction for an additional 9 variables was completed in May 2024 to respond to reviewer comments. Of the 80 included dashboards, 5 (6%) were no longer accessible, and additional abstraction for these websites was completed to the extent possible using the recorded videos.

Analysis

Following the completion of reviews, the abstracted data were exported from REDCap into R (version 4.3.1; The R Foundation) [32] for analysis. The number and proportion of dashboards with each characteristic were calculated. To characterize the stated purpose and audience, the team followed a qualitative content analysis approach, led by a PhD-trained team member with 15 years of qualitative and mixed methods research experience. The abstracted text that described purpose and audience was exported into a matrix and read for familiarization and the inductive identification of common themes. A total of 8 thematic categories were identified in the dashboard purpose descriptions, and 6 categories were identified in the dashboard audience descriptions. Operational definitions were developed for each of these 14 categories. After the REDCap form was modified to add these categories, 1 team member reviewed all definitions again to apply the relevant categories, and a second team member checked all purpose and abstract characterizations with reference to the abstracted text and original websites, as needed. Illustrative quotes for each purpose category were selected for presentation in the Results section.

Ethical Considerations

This study collected data from publicly available websites. It did not include individual-level data, and did not involve any interaction with human subjects. Therefore, the study did not meet the definition of human subjects research according to the University of Arkansas for Medical Science's "Human Subjects Protection Program Plan" (version 6/24/2021).

XSL•FO
RenderX

Results

Scope and Origins

Among the 80 publicly available maternal health dashboards included in this study, 24 (30%) were solely focused on maternal health; 28 (35%) included both maternal and child health data; and 28 (35%) included data for health areas outside of maternal and child health, such as environmental health and chronic diseases (Table 1). A total of 46 (58%) dashboards included the term "dashboard" in the website title or description, while 32 (40%) did not. Most dashboards (58/80, 72%) had a focus on state-level data, 8 (10%) presented national or multinational data, and 14 (18%) presented data for a substate region (eg, a city, multiple counties, or a tribal area). The most common hosting website for included dashboards was a health department website (48/80, 60%), followed by a nongovernmental organization or program website (14/80, 18%). Other types of hosting websites included those of federal agencies and universities (Multimedia Appendix 3). The software platforms used to create reviewed dashboards included Tableau (Tableau Software, LLC; 30/80, 38%); unspecified or custom systems (26/80, 32%); and off-the-shelf solutions (16/80, 20%), such as Conduent (Conduent Inc), Clear Impact (Clear Impact, LLC), MySidewalk (MySidewalk, Inc), and IBM Cognos (IBM Corp). A small number of dashboards were created with Power BI (Microsoft Corp; 4/80, 5%) or ArcGIS (Environmental Systems Research Institute, Inc; 4/80, 5%).

XSL•FO
RenderX

Table 1. Design and features of reviewed dashboards (N=80).

Characteristic[a]	Dashboards, n (%)
Topical scope	
Maternal and child health	28 (35)
Broader than maternal and child health	28 (35)
Maternal health only	24 (30)
Geographic scope	
National or multinational	8 (10)
State focus	58 (72)
Region or other area	14 (18)
Use of the term "dashboard" in the title or description[b]	
Includes the term "dashboard"	46 (58)
Does not include the term "dashboard"	32 (40)
Hosting organization or website	
Health department web page	48 (60)
NGO[c] or program web page	14 (18)
Other or unspecified	18 (22)
Visualizations of maternal health data	
Table	63 (79)
Bar chart	58 (72)
Line graph	52 (65)
Map	46 (58)
Pie chart	14 (18)
Large numbers or card visualizations	9 (11)
Arrows representing the direction of change	7 (9)
Gauge chart	5 (6)
Other visualization type	9 (11)
CIs in any visualizations	24 (30)
Software platform	
Tableau	30 (38)
Unspecified or custom system	26 (32)
Off-the-shelf solutions[d]	16 (20)
Power BI	4 (5)
ArcGIS	4 (5)
Technical capabilities	
User can download data	46 (58)
User can download complete dashboard view	43 (54)
All indicators visible on one page	38 (48)
Responsive interface	37 (46)
Instructions for interactive features	35 (44)
Mobile-friendly interface	31 (39)
User can download individual visualizations	26 (32)
User can select visualization type	16 (20)

XSL•FO
RenderX

Characteristic[a]	Dashboards, n (%)
Adaptive visualization (select to filter)	8 (10)
Full screen	6 (8)
Other	21 (26)
Statements of purpose and audience	
Statement of purpose	55 (69)
Statement of intended audience	30 (38)
Stated audience categories[e]	
Legislators or policy makers	17 (57)
Public health agencies or organizations	16 (53)
General audience descriptor	15 (50)
Health care organizations or providers	11 (37)
Community members or health care consumers	10 (33)
Researchers	7 (23)
Other	8 (27)

[a]All characteristics were assessed for the dashboard as a whole, except for visualizations, which was restricted to maternal health data.

[b]Could not be assessed for 2 dashboards that were no longer accessible during the second round of abstraction.

[c]NGO: nongovernmental organization.

[d]Off-the-shelf systems include Conduent and MySidewalk.

[e]Percentages calculated among the subset of 30 dashboards with a stated audience.

Technical Capabilities and Visualizations

Included dashboards used a range of data visualizations and technical features (Table 1). A total of 4 visualization types were used in most dashboards: tables (63/80, 79%); bar charts (58/80, 72%); line graphs (52/80, 65%); and maps (46/80, 58%), predominantly choropleth maps. Less common visualizations included pie charts (14/80, 18%), large numbers or card visualizations (9/80, 11%), directional arrows (7/80, 9%), and gauge charts (5/80, 6%). The most common technical features observed for included dashboards were the ability to download data (46/80, 58%) and the ability to download the dashboard view as an image or document file (eg, in PDF) (43/80, 54%). The classic dashboard style of presenting all indicators on 1 page was observed in roughly half of the reviewed dashboards (38/80, 48%). Some dashboards also accommodated different screen sizes with responsive interfaces (37/80, 46%) and mobile-friendly interfaces (31/80, 39%). Less common technical features included the option allowing the user to select the type of visualization (16/80, 20%), adaptive visualizations (8/80, 10%), and full-screen mode (6/80, 8%).

Audience and Purpose

While most included dashboards provided an explicit statement regarding the purpose of the dashboard (55/80, 69%), descriptions of the intended audience were less common (30/80, 38%), with close to two-thirds (50/80, 62%) of dashboards not specifying their audience (Table 1). A total of 8 categories of audiences were identified from the dashboards' textual statements regarding audience. Audience categories specified in at least half of the audience statements included legislators or policy makers (17/30, 57%); public health agencies or organizations (16/30, 53%); and general audience descriptors (15/30, 50%), such as "stakeholders," "partners," or general "public" audiences. Health care organizations or providers (11/30, 37%) and community members or health care consumers (10/30, 33%) were identified in around one-third of the statements, while the least common category was researchers (7/30, 23%).

The thematic analysis of dashboards' purpose statements identified 8 distinct categories of purpose: data accessibility, program planning or needs assessment, accountability for government organizations, program evaluation and monitoring, research, quality improvement, policy making and goal setting, and individual decision-making (Table 2). Among the 55 dashboards with a statement of purpose, general data accessibility was the most common purpose and was noted in 35 (64%) dashboards. Often, the dashboard was described as a tool to achieve a public agency's goals or mandates for data accessibility. The second most common purpose was program planning or needs assessment (n=24, 44%), which included identifying health needs, developing plans to address these needs, and applying for funding to support program plans. Accountability for government-supported programs (eg, Medicaid) and health care organizations was included in some purpose statements (n=14, 25%), as was program evaluation and monitoring (n=14, 25%). The least common purpose categories were research (n=8, 15%), quality improvement for health care organizations (n=7, 13%), and policy making and goal setting (n=4, 7%). Supporting individual decision-making by health care consumers or community members, such as selecting a health care provider or identifying a healthy behavior, was observed in a small proportion of purpose statements (n=4,

XSL•FO
RenderX

7%). Other types of purposes included journalism and promoting residency or tourism in the area (Multimedia Appendix 3).

Table 2. Purpose categories and example statements among dashboards that provided a statement of purpose (n=55).

Purpose category	Values, n (%)	Example statements
Data accessibility	35 (64)	Idaho Vital Statistics Natality Dashboard [33][a]: "The Division of Public Health's Strategic Plan has identified priority areas which include leveraging and using data more effectively across the Division as well as improving data accessibility for use by stakeholders (i.e., public health agencies, health systems, decision-makers, and the public). The data found on this website are intended to help the Division of Public Health achieve those priority area strategic goals."
Program planning and needs assessment	24 (44)	Georgia Online Analytical Statistical Information System (OASIS[b]) [34]: "OASIS plays an integral role in program planning, which includes determining target population areas, formulating financial plans, monitoring program effectiveness, program evaluation and reporting program outcomes. Use OASIS data querying tools to: develop profiles and report cards for counties or districts; assess community health needs, prioritize health problems, and evaluate programs; assemble data for grant writing, health analysis, special projects or state legislative reporting."
Accountability for government and organizations	14 (25)	Missouri HealthNet Managed Care Quality Dashboard [35]: "The purpose of the Medicaid Managed Care Quality Dashboard is to provide transparency and accountability in the health care provided to Missouri's Medicaid participants."
Program evaluation and monitoring	14 (25)	Montana Children's Health Data Dashboard [36]: "The Montana Children's Health Data Dashboard highlights ten shared measures identified by early childhood health stakeholders. Stakeholders and early childhood coalitions can use these measures to track outcomes and demonstrate the impact of their work."
Research	8 (15)	Hawai'i Health Data Warehouse [37]: "The Hawai'i Health Data Warehouse is dedicated to providing useful data to support public health professionals, researchers, the community and health agencies to become more effective in the application of health data."
Quality improvement	7 (13)	Agency for Healthcare Research and Quality Perinatal Dashboard [38]: "This rich data source makes it possible to identify and track patient safety concerns for the purpose of learning how to mitigate patient safety risks and reduce harm across healthcare settings nationally."
Policy making and goal setting	4 (7)	Kansas Health Matters [39]: "Kansas Health Matters is intended to help hospitals, health departments, community members and policy makers learn about the health of the community and ways to help improve it.... Master planners and government representatives can use this data to establish community goals on a variety of platforms."
Individual decision-making	4 (7)	New Jersey Maternal Health Hospital Report Card [40]: "This is an informational resource tool that provides important data on maternal health care provided in New Jersey licensed birthing general acute care hospitals. In determining the best hospital for you, you may review the information provided in this report."
Other	10 (18)	March of Dimes Peristats [41]: "Data are updated throughout the year, and useful for multiple tasks, including fact-finding,...lectures and presentations."

[a]Reference numbers refer to the order of appearance in Multimedia Appendix 1.

[b]OASIS: Online Analytical Statistical Information System.

Data Sources

Most dashboards reported data from only 1 (33/80, 41%) or 2 (23/80, 29%) sources (Table 3). Over half (47/80, 59%) of the dashboards reported state vital statistics as a data source. The second most common data source was the state PRAMS (22/80, 28%), followed by state health departments (12/80, 15%), the US Census (9/80, 11%), the CDC's Wide-ranging Online Data for Epidemiologic Research system (9/80, 11%), and the National Center for Health Statistics (8/80, 10%). Data sources reported in <10% (8/80) of dashboards include hospital discharge data (6/80, 8%), the Health Resources and Services Administration (5/80, 6%), and insurance claims data (4/80, 5%). Of the 80 dashboards, 2 (2%) reported the state MMRC as a source of data. Narrative interpretations of the meaning of maternal health data were more commonly included on reviewed dashboards than narrative interpretations of data quality.

XSL•FO
RenderX

Table 3. Sources, comparisons, and disaggregation capabilities for maternal health data within reviewed dashboards (N=80).

Characteristic	Dashboards, n (%)
Number of data sources reported	
1	33 (41)
2	23 (29)
3	5 (6)
4	6 (8)
≥5	10 (12)
None stated	3 (4)
Specified data sources	
State vital statistics	47 (59)
PRAMS[a]	22 (28)
State health department	12 (15)
US Census	9 (11)
CDC WONDER[b]	9 (11)
National Center for Health Statistics	8 (10)
Hospital discharge data	6 (8)
Health Resources and Services Administration	5 (6)
Insurance claims data	4 (5)
Maternal mortality reviews	2 (2)
Other[c]	31 (39)
Narrative interpretations of data quality	
Completeness or validity	31 (39)
Privacy protections	16 (20)
Biases	2 (2)
Narrative interpretations of meaning	
Definition of variable	51 (64)
Normative interpretation of value	32 (40)
Contextualization with factors influencing value[d]	52 (65)
Comparisons	
Temporal	63 (79)
Other geographic areas	42 (52)
National data	25 (31)
Benchmarks or targets	18 (22)

[a]PRAMS: Pregnancy Risk Assessment Monitoring System.

[b]CDC WONDER: Centers for Disease Control and Prevention Wide-ranging Online Data for Epidemiological Research.

[c]See Multimedia Appendix 3 for a description of other responses.

[d]Includes discussion of the influence of sample size on the reliability of estimates as well as policy or environmental influences.

Comparisons and Disaggregation Capabilities

The comparison and data disaggregation capabilities also varied considerably between dashboards (Table 3). While a large majority of dashboards (63/80, 79%) provided temporal comparisons for included indicators, roughly half (42/80, 52%) allowed for comparisons between geographic areas. Some dashboards also provided comparisons against national data (25/80, 31%) and benchmark or target values (18/80, 22%), such as the Healthy People goals. The predominant period for reporting data was annual, followed by multiyear reporting periods (Multimedia Appendix 4).

Data disaggregation by race or ethnicity was possible for any data on 55 (69%) dashboards and for all data on 26 (32%)

XSL•FO
RenderX

dashboards (Table 4). The next most common disaggregation variables for any data were maternal age (42/80, 52%) and insurance status or type (24/80, 30%). Few dashboards presented data by health care facility (7/80, 9%) or health care provider (5/80, 6%).

Table 4. Disaggregation capabilities of included dashboards (N=80).

Disaggregating variable	Dashboards, n (%)	
	Any data	All data
Time interval	65 (81)	54 (68)
Geography	65 (81)	53 (66)
Race or ethnicity	55 (69)	26 (32)
Maternal age	42 (52)	14 (18)
Insurance status or type	24 (30)	13 (16)
Education	18 (22)	5 (6)
Marital status	16 (20)	5 (6)
Receipt of social support	10 (12)	3 (4)
Birth weight	9 (11)	0 (0)
Level of poverty	7 (9)	3 (4)
Sex or gender of the mother or baby	7 (9)	2 (2)
Household income	7 (9)	2 (2)
Plurality	7 (9)	1 (1)
Health care facility	7 (9)	1 (1)
Nativity or citizenship	7 (9)	1 (1)
Gestational age	6 (8)	1 (1)
Health care provider	5 (6)	1 (1)
Other characteristics[a]	21 (26)	7 (9)

[a]See Multimedia Appendix 3 for a description of other responses.

Indicators

Table 5 displays the presence of 40 specific perinatal health indicators in 4 groups, health status or outcomes, health behaviors and health care use, individual characteristics and risk factors, and health system characteristics, along with the format in which and the geographic level at which each is reported. Rates or percentages were more commonly used as the reporting format than counts for almost all health indicators. The most common geographic level for reporting health indicators was the state, followed by county and regions within states. Among health status indicators, birth weight (54/80, 68%), preterm birth (45/80, 56%), and infant mortality (42/80, 52%) were reported in at least half of the reviewed dashboards. Birth statistics were reported in 48% (38/80) of dashboards, and mode of delivery (eg, vaginal or cesarean delivery) was included in 26% (21/80) of dashboards. Indicators of any pregnancy-related problems, such as gestational diabetes or preeclampsia, were reported in 20% (16/80) of dashboards, while maternal mental health conditions were reported in 18% (14/80) of dashboards. Severe maternal morbidity and maternal mortality were each reported in <1 (20%) in 5 of the reviewed dashboards.

Among the 23 health indicators grouped under the “health behaviors and health care use” and “individual characteristics and risk factors” groups, only 3 (13%) were reported on half or more of the reviewed dashboards: receipt of prenatal care (57/80, 71%), maternal smoking (42/80, 52%), and maternal age (43/80, 54%). Indicators reported in 20% to 30% of dashboards include breastfeeding (22/80, 28%), preconception health (18/80, 22%), maternal BMI (17/80, 21%), health insurance status (17/80, 21%), and maternal education (17/80, 21%). The remaining 15 (65%) health indicators in the “health behaviors and health care use” and “individual characteristics and risk factors” groups were reported in <20% (16/80) of dashboards. Finally, “health system characteristics” was the least commonly included group of indicators. Among the 80 dashboards, indicators of population characteristics (eg, sex ratios, fertility rates, and pregnancy rates) and access to the Special Supplemental Nutrition Program for Women, Infants, and Children were reported in 18 (22%) and 16 (20%) dashboards, respectively. The availability of maternity care was reported in 12 (15%) dashboards, while policy measures and expenditures related to maternal health care were reported only in <5 (6%) dashboards.

XSL•FO
RenderX

Table 5. Maternal health indicators and reporting format among reviewed dashboards (N=80).

Category	Any indicator, n (%)	Format[a], n (%)		Geographic level[a], n (%)				
		Count	Percentage or rate	National	State	Substate region	County	Smaller than county
Health status and outcomes								
Birth weight	54 (68)	27 (34)	50 (62)	15 (19)	48 (60)	20 (25)	37 (46)	7 (9)
Preterm birth	45 (56)	20 (25)	41 (51)	11 (14)	42 (52)	17 (21)	30 (38)	3 (4)
Infant mortality	42 (52)	19 (24)	39 (49)	16 (20)	35 (44)	13 (16)	31 (39)	3 (4)
Births	38 (48)	32 (40)	23 (29)	7 (9)	30 (38)	13 (16)	30 (38)	7 (9)
Mode of delivery	21 (26)	11 (14)	19 (24)	8 (10)	20 (25)	5 (6)	10 (12)	4 (5)
Pregnancy-related problems	16 (20)	8 (10)	15 (19)	2 (2)	15 (19)	6 (8)	8 (10)	3 (4)
Maternal depression or anxiety	14 (18)	5 (6)	14 (18)	2 (2)	14 (18)	6 (8)	4 (5)	1 (1)
Severe maternal morbidity	12 (15)	5 (6)	9 (11)	4 (5)	12 (15)	3 (4)	2 (2)	1 (1)
Maternal mortality	10 (12)	4 (5)	9 (11)	4 (5)	9 (11)	1 (1)	6 (8)	0 (0)
Birth defects	9 (11)	5 (6)	9 (11)	2 (2)	6 (8)	1 (1)	4 (5)	0 (0)
Neonatal abstinence syndrome	8 (10)	3 (4)	8 (10)	4 (5)	8 (10)	3 (4)	3 (4)	0 (0)
Other	14 (18)	9 (11)	13 (16)	2 (2)	12 (15)	4 (5)	9 (11)	2 (2)
Health behaviors and health care use								
Receipt of prenatal care	57 (71)	21 (26)	54 (68)	13 (16)	53 (66)	20 (25)	35 (44)	3 (4)
Maternal smoking	42 (52)	16 (20)	41 (51)	10 (12)	39 (49)	16 (20)	22 (28)	0 (0)
Breastfeeding	22 (28)	8 (10)	21 (26)	6 (8)	21 (26)	7 (9)	7 (9)	2 (2)
Substance use	14 (18)	5 (6)	13 (16)	2 (2)	14 (18)	6 (8)	5 (6)	0 (0)
Oral health	13 (16)	4 (5)	13 (16)	1 (1)	12 (15)	3 (4)	2 (2)	0 (0)
Postpartum care	12 (15)	4 (5)	10 (12)	2 (2)	11 (14)	3 (4)	2 (2)	0 (0)
Maternal nutrition	11 (14)	4 (5)	11 (14)	1 (1)	11 (14)	4 (5)	4 (5)	0 (0)
Immunization	10 (12)	5 (6)	10 (12)	1 (1)	10 (12)	3 (4)	3 (4)	1 (1)
Contraception	9 (11)	3 (4)	9 (11)	2 (2)	9 (11)	5 (6)	4 (5)	1 (1)
Preventive health care visit	5 (6)	1 (1)	5 (6)	1 (1)	5 (6)	1 (1)	2 (2)	1 (1)
Smoking household	5 (6)	2 (2)	5 (6)	1 (1)	5 (6)	3 (4)	3 (4)	2 (2)
Other	19 (24)	9 (11)	17 (21)	3 (4)	18 (22)	7 (9)	8 (10)	3 (4)
Individual characteristics and risk factors								
Maternal age	43 (54)	19 (24)	37 (46)	14 (18)	36 (45)	14 (18)	26 (32)	5 (6)
Preconception health	18 (22)	8 (10)	17 (21)	1 (1)	16 (20)	5 (6)	6 (8)	3 (4)
Maternal BMI	17 (21)	6 (8)	17 (21)	2 (2)	16 (20)	4 (5)	10 (12)	2 (2)
Health insurance status	17 (21)	9 (11)	17 (21)	3 (4)	15 (19)	3 (4)	8 (10)	3 (4)
Maternal education	17 (21)	11 (14)	14 (18)	4 (5)	16 (20)	5 (6)	11 (14)	3 (4)
Pregnancy intention	13 (16)	5 (6)	13 (16)	2 (23)	13 (16)	5 (6)	3 (4)	0 (0)
Stress or abuse	11 (14)	5 (6)	11 (14)	2 (2)	10 (12)	2 (2)	3 (4)	1 (1)
Maternal marital status	10 (12)	9 (11)	8 (10)	1 (1)	10 (12)	4 (5)	9 (11)	1 (1)
Plurality	9 (11)	7 (9)	8 (10)	1 (1)	9 (11)	3 (4)	8 (10)	2 (2)
Birth spacing	7 (9)	6 (8)	6 (8)	1 (1)	6 (8)	3 (4)	6 (8)	2 (2)
Number of prior births	7 (9)	5 (6)	7 (9)	0 (0)	7 (9)	3 (4)	4 (5)	1 (1)
Maternal race or ethnicity	5 (6)	2 (2)	4 (5)	0 (0)	5 (6)	2 (2)	1 (1)	1 (1)
Other	2 (2)	1 (1)	2 (2)	1 (1)	1 (1)	0 (0)	1 (1)	0 (0)

XSL•FO
RenderX

Category	Any indicator, n (%)	Format[a], n (%)		Geographic level[a], n (%)				
		Count	Percentage or rate	National	State	Substate region	County	Smaller than county
Health system characteristics								
Population characteristic	18 (22)	7 (9)	17 (21)	3 (4)	16 (20)	5 (6)	13 (16)	4 (5)
WIC access[b]	16 (20)	4 (5)	16 (20)	3 (4)	11 (14)	3 (4)	9 (11)	0 (0)
Availability of maternity care	12 (15)	6 (8)	6 (8)	2 (2)	9 (11)	3 (4)	8 (10)	3 (4)
Policy measures	4 (5)	3 (4)	1 (1)	3 (4)	2 (2)	2 (2)	1 (1)	0 (0)
Health care expenditure	2 (2)	2 (2)	0 (0)	0 (0)	2 (2)	1 (1)	2 (2)	1 (1)
Other	3 (4)	1 (1)	2 (2)	0 (0)	3 (4)	3 (4)	2 (2)	1 (1)

[a]Dashboards may use >1 reporting format and >1 geographic level.

[b]WIC access: Special Supplemental Nutrition Program for Women, Infants, and Children access.

Discussion

Principal Findings

This study demonstrates that public dashboards presenting US maternal health data are widespread. While most reviewed dashboards are hosted by state or local health departments, the content and features across maternal health dashboards are not uniform. Reviewed dashboards were developed with various software platforms and offered diverse technical capabilities. Only half (38/80, 48%) of the maternal health dashboard designs reflected the traditional, single-view presentation of key metrics, demonstrating the expanding conceptualization of dashboards [3]. The included dashboards also presented a range of maternal health indicators, generally reflective of their reported scope and included data sources. The heterogeneity in the design and content of maternal health dashboards is similar to that observed in public COVID-19 dashboards [7,31].

In addition to mapping the breadth of options for US maternal health dashboards, this study also suggests opportunities for improving the utility of current and future dashboards. Greater integration of multiple data sources is the first opportunity for improvement suggested by these results. Overall, 70% (56/80) of included dashboards in this study provide data from only 1 or 2 sources, yet individual data sources provide an incomplete picture of maternal health in the United States. Birth records, the primary vital statistics data reported on dashboards, exclude key outcomes, notably maternal mortality rates. It is particularly striking that only 2 (2%) of the 80 dashboards included data from MMRCs, the gold standard assessment of state-level maternal mortality [42]. Similarly, hospital discharge data files, the primary data source for calculating severe maternal mortality rates using the CDC's algorithm [23], were a reported data source in only 6 (8%) of the 80 dashboards. State-focused dashboards, in particular, could enable more efficient access to maternal health data by including key maternal indicators from all state data sources—vital statistics (including death certificates), PRAMS, MMRCs, and hospital discharge data files—in 1 comprehensive dashboard.

A second area where public maternal health dashboards can be strengthened is their ability to identify disparities through data disaggregation. Maternal health disparities in the United States are of grave concern, with Black women and American Indian or Alaska Native women having 2 to 3 times greater risk of maternal death compared to White non-Hispanic women [14] and increased likelihood of receiving poor quality care [43]. Women residing in rural areas [44] and those with public insurance [43] also experience worse maternal outcomes. The ability of data systems to detect these disparities is considered critical for efforts aimed at reducing adverse perinatal outcomes overall as well as disparities in these outcomes [43,45,46]. While most of the reviewed dashboards permitted the disaggregation of all indicators by geography, only one-third (26/80, 32%) had the capability of disaggregation by race or ethnicity for all indicators. Disaggregation by other social determinants of maternal health, such as insurance type, education, and language spoken, was even less common. Although privacy protections may prohibit the stratified presentation of infrequent outcomes [47], such as maternal mortality, most individual-level indicators reported on reviewed dashboards would not require such data suppression.

Third, maternal dashboards may increase attention to the information needs of their audiences and segmentation of information presented for different audiences. Fewer than half (30/80, 38%) of the dashboards in this study provided statements about their intended audience; available audience statements typically listed multiple, diverse stakeholders. Critical assessments of public health dashboards have noted that "not all data are necessary for all users at all times, and often the information that citizens need to make informed decisions is absent" [4]. The interest and needs of policy makers and providers for publicly reportable maternal health data are not well studied. For maternity patients, documented information needs include identifying geographically close health care providers who provide the services and quality that the patient prefers [48-50]. These patient needs appear to be particularly underserved by current maternal health dashboards. Among reviewed dashboards, <1 (10%) in 10 presented any data by health care facility. The technical features and visual displays of dashboards should also be differentiated for different audiences [51]; while program planners and researchers may prioritize access to a wide range of variables available through sophisticated queries, maternity patient decision-making has been shown to benefit from explanations about the relevance

XSL•FO
RenderX

of included data [52]. Engaging intended audience members during dashboard design and development is a best practice that was followed in 1 (50%) [29] of the 2 reports of public maternal dashboard development [28,29]. Future qualitative research with maternal health dashboard development teams, similar to COVID-19 dashboard studies [53], could lead to a better understanding of the factors guiding their decisions and opportunities to improve processes and dashboard actionability for intended users.

Limitations

This study could not evaluate the impact of maternal health dashboards and was limited in the extent to which it could assess aspects of the user experience, such as usability. Although comprehensive usability standards have recently been proposed for public health data dashboards [54], further operationalization is needed before they can be consistently applied. The features of some dashboards suggest limited usability for many users. For example, fewer than half (31/80, 39%) of dashboards had a responsive interface that would reformat for viewing on a mobile device. This study was also unable to address the usefulness of the data and features of maternal health dashboards from the perspective of intended users. We could not identify any published articles evaluating the user experience, interaction effectiveness, system efficacy [55], or actionability [13] of a deployed maternal health dashboard, although these evaluations are critical to understanding and optimizing the contribution of public health dashboards to patient- and system-level decision-making [6].

Further Considerations

For maternal health dashboards to successfully improve in sophistication and meet the expectations of multiple stakeholders, prior research suggests that it is important to involve collaborative interdisciplinary teams in dashboard design and development [53]. The expertise of systems analysts, software engineers specializing in backend and front-end web development, and user-interface designers is likely to result in dashboards with improved usefulness and usability. These fields have established best practices for designing usable systems that are often neglected in health information systems [56]. When resources are limited, general-purpose data analytics tools such as Tableau or Microsoft Power BI are promising low-cost options for prototyping. Ongoing data improvement efforts by national funders and organizations may also contribute resources and standards to support the efficient development of maternal data dashboards. For example, data standards could facilitate sharing, reuse, and integration of maternal health data [57], and user-interface standards could lead to more effective and efficient development by offering specific widgets and visualization solutions for certain maternal health data or for certain purposes (eg, benchmarking).

Conclusions

This national descriptive assessment of maternal health dashboards in the United States found substantial variation among dashboards, including inconsistent data sources, health indicators, and disaggregation capabilities. Few dashboards included information regarding maternity patients as an intended dashboard audience. Careful consideration of the design of publicly available health dashboards, with specific intent to develop dashboards that cater to end-user needs and to include data from multiple sources, is critical for ensuring that dashboards achieve their intended goals.

Acknowledgments

This work was made possible by the grant U7A46847 from the Maternal and Child Health Bureau of the Health Resources and Services Administration (HRSA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Department of Health and Human Services. This work was also supported by the University of Arkansas for Medical Sciences Translational Research Institute funding awarded through the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH; UL1 TR003107). CCB was supported by the National Institute on Minority Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH; 1K01MD018072). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, who had no role in the design, analysis, or writing of this article. The authors wish to thank Erin Gloster for assistance with manuscript submission.

Authors' Contributions

JAC-K contributed to conceptualization, methodology, validation, resources, writing the original draft, and funding acquisition. PNC contributed to investigation, data curation, and reviewing and editing the manuscript. TT contributed to software, investigation, data curation, formal analysis, and reviewing and editing the manuscript. CCB contributed to methodology, validation, and reviewing and editing the manuscript. WG contributed to methodology and reviewing and editing the manuscript. GK contributed to conceptualization, methodology, investigation, validation, data curation, project administration, supervision, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

List of reviewed maternal health dashboards.
[DOCX File , 34 KB - jmir_v26i1e56804_app1.docx]

Multimedia Appendix 2
REDCap (Research Electronic Data Capture) data abstraction instrument.
[DOCX File , 105 KB - jmir_v26i1e56804_app2.docx]

Multimedia Appendix 3
Description of other categories for Tables 1, 3, 4, and 5.
[DOCX File , 18 KB - jmir_v26i1e56804_app3.docx]

Multimedia Appendix 4
Indicator reporting periods.
[DOCX File , 17 KB - jmir_v26i1e56804_app4.docx]

References

1. Smith VS. Data dashboard as evaluation and research communication tool. New Dir Eval 2013 Dec 10;2013(140):21-45. [doi: 10.1002/ev.20072]
2. Pappas L, Whitman L. Riding the technology wave: effective dashboard data visualization. In: Proceedings of the Symposium on Human Interface 2011, Held as Part of HCI International 2011. 2011 Presented at: HCI International 2011; July 9-14, 2011; Orlando, FL. [doi: 10.1007/978-3-642-21793-7_29]
3. Sarikaya A, Correll M, Bartram L, Tory M, Fisher D. What do we talk about when we talk about dashboards? IEEE Trans Vis Comput Graph 2019 Jan;25(1):682-692. [doi: 10.1109/tvcg.2018.2864903]
4. Dasgupta N, Kapadia F. The future of the public health data dashboard. Am J Public Health 2022 Jun;112(6):886-888. [doi: 10.2105/ajph.2022.306871]
5. Rabiei R, Almasi S. Requirements and challenges of hospital dashboards: a systematic literature review. BMC Med Inform Decis Mak 2022 Nov 08;22(1):287 [FREE Full text] [doi: 10.1186/s12911-022-02037-8] [Medline: 36348339]
6. Thorpe LE, Gourevitch MN. Data dashboards for advancing health and equity: proving their promise? Am J Public Health 2022 Jun;112(6):889-892 [FREE Full text] [doi: 10.2105/AJPH.2022.306847] [Medline: 35446603]
7. Ivanković D, Barbazza E, Bos V, Brito Fernandes Ó, Jamieson Gilmore K, Jansen T, et al. Features constituting actionable COVID-19 dashboards: descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. J Med Internet Res 2021 Mar 24;23(2):e25682 [FREE Full text] [doi: 10.2196/25682] [Medline: 33577467]
8. Gourevitch MN, Athens JK, Levine SE, Kleiman N, Thorpe LE. City-level measures of health, health determinants, and equity to foster population health improvement: the city health dashboard. Am J Public Health 2019 Apr;109(4):585-592. [doi: 10.2105/AJPH.2018.304903] [Medline: 30789770]
9. About the drug overdose surveillance and epidemiology (DOSE) system. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/overdose-prevention/data-research/facts-stats/about-dose-system.html [accessed 2023-10-01]
10. Hughes P, Denslow S, Ostrach B, Fusco C, Tak C. Exploration of the STOP Act and opioid deaths in North Carolina, 2010-2018. Am J Public Health 2020 Oct;110(10):1573-1577. [doi: 10.2105/AJPH.2020.305748] [Medline: 32816537]
11. Anderson J, Demeter N, Pasquires M, Wirtz S. Using the CA opioid overdose surveillance dashboard to track opioid overdose deaths. Online J Public Health Inform 2019;11(1):e62570. [doi: 10.5210/ojphi.v11i1.9938]
12. Gardner L. The COVID-19 dashboard for real-time tracking of the pandemic: the Lasker-Bloomberg public service award. JAMA 2022 Oct 04;328(13):1295-1296. [doi: 10.1001/jama.2022.15590] [Medline: 36170054]
13. Barbazza E, Ivanković D, Wang S, Gilmore KJ, Poldrugovac M, Willmington C, et al. Exploring changes to the actionability of COVID-19 dashboards over the course of 2020 in the Canadian context: descriptive assessment and expert appraisal study. J Med Internet Res 2021 Aug 06;23(8):e30200 [FREE Full text] [doi: 10.2196/30200] [Medline: 34280120]
14. The White House blueprint for addressing the maternal health crisis: two years of progress. The White House. 2024 Jul 10. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2024/07/10/the-white-house-blueprint-for-addressing-the-maternal-health-crisis-two-years-of-progress/ [accessed 2024-08-22]
15. Chinn JJ, Eisenberg E, Artis Dickerson S, King RB, Chakhtoura N, Lim IA, et al. Maternal mortality in the United States: research gaps, opportunities, and priorities. Am J Obstet Gynecol 2020 Oct;223(4):486-92.e6 [FREE Full text] [doi: 10.1016/j.ajog.2020.07.021] [Medline: 32682858]
16. MacDorman MF, Thoma M, Declcerq E, Howell EA. Racial and ethnic disparities in maternal mortality in the United States using enhanced vital records, 2016 2017. Am J Public Health 2021 Oct;111(9):1673-1681 [FREE Full text] [doi: 10.2105/AJPH.2021.306375] [Medline: 34383557]
17. Mayer R, Dingwall A, Simon-Thomas J, Sheikhnureldin A, Lewis K. The United States maternal mortality rate will continue to increase without access to data. Health Affairs. 2019 Feb 4. URL: https://www.healthaffairs.org/content/forefront/united-states-maternal-mortality-rate-continue-increase-without-access-data [accessed 2023-10-01]

XSL•FO
RenderX

18. Enhancing the data infrastructure for women's health research to improve women's and maternal health outcomes. U.S. Department of Health and Human Services, Assistant Secretary for Planning and Evaluation. 2021. URL: https://aspe.hhs.gov/sites/default/files/2021-07/womens-maternal-health.pdf [accessed 2023-10-01]
19. Enhancing reviews and surveillance to eliminate maternal mortality. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/maternal-mortality/php/erase-mm/index.html [accessed 2023-03-03]
20. State maternal health innovation and data capacity program. Health Resources and Services Administration. URL: https://www.hrsa.gov/grants/find-funding/HRSA-22-149 [accessed 2023-10-01]
21. Strahle S. Linking datasets to address racial equity in maternal and child health outcomes. Association of State and Territorial Health Officials. 2023 Aug 01. URL: https://www.astho.org/topic/brief/linking-datasets-to-address-racial-equity-in-maternal-and-child-health-outcomes/ [accessed 2023-08-31]
22. Zaharatos J, St Pierre A, Cornell A, Pasalic E, Goodman D. Building U.S. capacity to review and prevent maternal deaths. J Womens Health (Larchmt) 2018 Jan;27(1):1-5. [doi: 10.1089/jwh.2017.6800] [Medline: 29240525]
23. Snowden JM, Lyndon A, Kan P, El Ayadi A, Main E, Carmichael SL. Severe maternal morbidity: a comparison of definitions and data sources. Am J Epidemiol 2021 Sep 01;190(9):1890-1897 [FREE Full text] [doi: 10.1093/aje/kwab077] [Medline: 33755046]
24. Daw JR, Auty SG, Admon LK, Gordon SH. Using modernized medicaid data to advance evidence-based improvements in maternal health. Am J Public Health 2023 Jul;113(7):805-810. [doi: 10.2105/AJPH.2023.307287] [Medline: 37141557]
25. Shulman HB, D'Angelo DV, Harrison L, Smith RA, Warner L. The pregnancy risk assessment monitoring system (PRAMS): overview of design and methodology. Am J Public Health 2018 Oct;108(10):1305-1313. [doi: 10.2105/AJPH.2018.304563] [Medline: 30138070]
26. Kauh TJ, Read JG, Scheitler AJ. The critical role of racial/ethnic data disaggregation for health equity. Popul Res Policy Rev 2021;40(1):1-7 [FREE Full text] [doi: 10.1007/s11113-020-09631-6] [Medline: 33437108]
27. Kuznetsova M, Frits ML, Dulgarian S, Iannaccone C, Mort E, Bates DW, et al. An analysis of the structure and content of dashboards used to monitor patient safety in the inpatient setting. JAMIA Open 2021 Nov 17;4(4):ooab096 [FREE Full text] [doi: 10.1093/jamiaopen/ooab096] [Medline: 34805777]
28. Hutchinson-Colas JA, Balica A, Chervenak FA, Friedman D, Locke LS, Bachmann G, et al. New Jersey maternal mortality dashboard: an interactive social-determinants-of-health tool. J Perinat Med 2022 Feb 28;51(2):188-196 [FREE Full text] [doi: 10.1515/jpm-2021-0673] [Medline: 35224952]
29. Fostering livable communities for birthing people: a platform for data-driven investment in the wellbeing of mothers. Ariadne Labs. 2021. URL: https://www.ariadnelabs.org/wp-content/uploads/2021/07/Cities-White-Paper-Fostering-Livable-Communities.pdf [accessed 2023-10-01]
30. Peters M, Godfrey CM, McInerney P, Munn Z, Trico A, Khalil H. Scoping reviews. In: JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; Jan 2020.
31. Fareed N, Swoboda CM, Chen S, Potter E, Wu DT, Sieck CJ. U.S. COVID-19 state government public dashboards: an expert review. Appl Clin Inform 2021 Mar;12(2):208-221 [FREE Full text] [doi: 10.1055/s-0041-1723989] [Medline: 33853140]
32. R: a language and environment for statistical computing. R Core Team. 2015 Feb 10. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2023-10-01]
33. Idaho births (vital statistics). Get Health Idaho. URL: https://www.gethealthy.dhw.idaho.gov/idaho-births-vital-statistics [accessed 2024-08-27]
34. OASIS GA DPH. Georgia Department of Public Health. URL: https://oasis.state.ga.us/ [accessed 2024-08-27]
35. Managed Care Health Plans. Missouri Department of Social Services. URL: https://mydss.mo.gov/mhd/managed-care-health-plans [accessed 2024-08-27]
36. Montana KIDS COUNT. Montana Budget and Policy Center. URL: https://montanabudget.org/kids-count [accessed 2024-08-27]
37. Hawai i Health Data Warehouse. URL: https://hhdw.org/ [accessed 2024-08-27]
38. Perinatal Dashboard. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/npsd/data/dashboard/perinatal.html [accessed 2024-08-27]
39. Kansas health indicators. Kansas Health Matters. URL: https://www.kansashealthmatters.org/indicators [accessed 2024-08-27]
40. New Jersey maternal data center. State of New Jerseyv Department of Health. URL: https://www.nj.gov/health/maternal/morbidity/mhh_reportcard/2018-2019/index.shtml [accessed 2024-08-27]
41. PeriStats. March of Dimes. URL: https://www.marchofdimes.org/peristats/ [accessed 2024-08-27]
42. Collier AR, Molina RL. Maternal mortality in the United States: updates on trends, causes, and solutions. Neoreviews 2019 Oct;20(10):e561-e574 [FREE Full text] [doi: 10.1542/neo.20-10-e561] [Medline: 31575778]
43. Howell EA, Egorova NN, Janevic T, Brodman M, Balbierz A, Zeitlin J, et al. Race and ethnicity, medical insurance, and within-hospital severe maternal morbidity disparities. Obstet Gynecol 2020 Feb;135(2):285-293 [FREE Full text] [doi: 10.1097/AOG.0000000000003667] [Medline: 31923076]

XSL•FO
RenderX

44. Kozhimannil KB, Interrante JD, Henning-Smith C, Admon LK. Rural-urban differences in severe maternal morbidity and mortality in the US, 2007-15. Health Aff (Millwood) 2019 Dec;38(12):2077-2085. [doi: 10.1377/hlthaff.2019.00805] [Medline: 31794322]
45. Taylor KJ, Benatar S. The pandemic has increased demand for data and accountability to decrease maternal health inequity. Urban Institute. 2020 Dec 04. URL: https://www.urban.org/research/publication/pandemic-has-increased-demand-data-and-accountability-decrease-maternal-health-inequity [accessed 2023-10-01]
46. Howell EA, Ahmed ZN. Eight steps for narrowing the maternal health disparity gap: step-by-step plan to reduce racial and ethnic disparities in care. Contemp Ob Gyn 2019 Jan;64(1):30-36 [FREE Full text] [Medline: 31673195]
47. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [accessed 2023-09-04]
48. Lagan BM, Sinclair M, Kernohan WG. What is the impact of the internet on decision-making in pregnancy? A global study. Birth 2011 Dec;38(4):336-345. [doi: 10.1111/j.1523-536X.2011.00488.x] [Medline: 22112334]
49. Gourevitch RA, Mehrotra A, Galvin G, Karp M, Plough A, Shah NT. How do pregnant women use quality measures when choosing their obstetric provider? Birth 2017 Jun;44(2):120-127 [FREE Full text] [doi: 10.1111/birt.12273] [Medline: 28124390]
50. Hebert LE, Freedman L, Stulberg DB. Choosing a hospital for obstetric, gynecologic, or reproductive healthcare: what matters most to patients? Am J Obstet Gynecol MFM 2020 Feb;2(1):100067. [doi: 10.1016/j.ajogmf.2019.100067] [Medline: 33345982]
51. Vazquez-Ingelmo A, Garcia-Penalvo FJ, Theron R. Information dashboards and tailoring capabilities - a systematic literature review. IEEE Access 2019;7:109673-109688. [doi: 10.1109/access.2019.2933472]
52. Maurer M, Carman KL, Yang M, Firminger K, Hibbard J. Increasing the use of comparative quality information in maternity care: results from a randomized controlled trial. Med Care Res Rev 2019 Apr;76(2):208-228. [doi: 10.1177/1077558717712290] [Medline: 29148346]
53. Barbazza E, Ivanković D, Davtyan K, Poldrugovac M, Yelgezekova Z, Willmington C, et al. The experiences of 33 national COVID-19 dashboard teams during the first year of the pandemic in the World Health Organization European Region: a qualitative study. Digit Health 2022 Aug 29;8:20552076221121154 [FREE Full text] [doi: 10.1177/20552076221121154] [Medline: 36060614]
54. Ansari B, Martin EG. Development of a usability checklist for public health dashboards to identify violations of usability principles. J Am Med Inform Assoc 2022 Oct 07;29(11):1847-1858 [FREE Full text] [doi: 10.1093/jamia/ocac140] [Medline: 35976140]
55. Zhuang M, Concannon D, Manley E. A framework for evaluating dashboards in healthcare. IEEE Trans Vis Comput Graph 2022 Apr 1;28(4):1715-1731. [doi: 10.1109/tvcg.2022.3147154]
56. Carayon P, Hoonakker P. Human factors and usability for health information technology: old and new challenges. Yearb Med Inform 2019 Aug;28(1):71-77 [FREE Full text] [doi: 10.1055/s-0039-1677907] [Medline: 31419818]
57. Mayer R, Kinling S, Dingwall A. Saving women's lives: a data-driven approach to reduce maternal mortality. MITRE Corporation. 2021. URL: https://www.mitre.org/sites/default/files/2021-11/prs-21-1148-saving-womens-lives-a-data-driven-approach-to-reduce-maternal-mortality.pdf [accessed 2023-10-01]

Abbreviations

CDC: Centers for Disease Control and Prevention
MMRC: maternal mortality review committee
PRAMS: Pregnancy Risk Assessment Monitoring System
REDCap: Research Electronic Data Capture

Edited by N Cahill; submitted 26.01.24; peer-reviewed by D Ivanković, E Barbazza, B Ansari; comments to author 21.03.24; revised version received 14.05.24; accepted 05.06.24; published 17.09.24.

Please cite as:
Callaghan-Koru JA, Newman Chargois P, Tiwari T, Brown CC, Greenfield W, Koru G
Public Maternal Health Dashboards in the United States: Descriptive Assessment
J Med Internet Res 2024;26:e56804
URL: https://www.jmir.org/2024/1/e56804
doi:10.2196/56804
PMID:

XSL•FO
RenderX

©Jennifer A Callaghan-Koru, Paige Newman Chargois, Tanvangi Tiwari, Clare C Brown, William Greenfield, Güneş Koru. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Data Visualization Preferences in Remote Measurement Technology for Individuals Living With Depression, Epilepsy, and Multiple Sclerosis: Qualitative Study

Sara Simblett[1], BSc, MSc, DClinPsy, PhD; Erin Dawe-Lane[1], BSc, MSc; Gina Gilpin[1], BSc; Daniel Morris[1], BSc; Katie White[1], PhD; Sinan Erturk[1], BSc; Julie Devonshire[2], BSc, BA; Simon Lees[2]; Spyridon Zormpas[2]; Ashley Polhemus[3], BSc, MSc, PhD; Gergely Temesi[3], PhD; Nicholas Cummins[1], PhD; Matthew Hotopf[1,4], PhD; Til Wykes[1,4], PhD; RADAR-CNS Consortium[1]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Remote Assessment of Disease And Relapse - Central Nervous System Patient Advisory Board, King's College London, London, United Kingdom

[3]Medical Science Division IT Services, Prague, Czech Republic

[4]South London and Maudsley NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Sara Simblett, BSc, MSc, DClinPsy, PhD
Institute of Psychiatry, Psychology and Neuroscience
King's College London
16 De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 02078480762
Email: sara.simblett@kcl.ac.uk

Abstract

Background: Remote measurement technology (RMT) involves the use of wearable devices and smartphone apps to measure health outcomes in everyday life. RMT with feedback in the form of data visual representations can facilitate self-management of chronic health conditions, promote health care engagement, and present opportunities for intervention. Studies to date focus broadly on multiple dimensions of service users' design preferences and RMT user experiences (eg, health variables of perceived importance and perceived quality of medical advice provided) as opposed to data visualization preferences.

Objective: This study aims to explore data visualization preferences and priorities in RMT, with individuals living with depression, those with epilepsy, and those with multiple sclerosis (MS).

Methods: A triangulated qualitative study comparing and thematically synthesizing focus group discussions with user reviews of existing self-management apps and a systematic review of RMT data visualization preferences. A total of 45 people participated in 6 focus groups across the 3 health conditions (depression, n=17; epilepsy, n=11; and MS, n=17).

Results: Thematic analysis validated a major theme around design preferences and recommendations and identified a further four minor themes: (1) data reporting, (2) impact of visualization, (3) moderators of visualization preferences, and (4) system-related factors and features.

Conclusions: When used effectively, data visualizations are valuable, engaging components of RMT. Easy to use and intuitive data visualization design was lauded by individuals with neurological and psychiatric conditions. Apps design needs to consider the unique requirements of service users. Overall, this study offers RMT developers a comprehensive outline of the data visualization preferences of individuals living with depression, epilepsy, and MS.

(J Med Internet Res 2024;26:e43954) doi:10.2196/43954

KEYWORDS

mHealth; qualitative; technology; depression; epilepsy; multiple sclerosis; wearables; devices; smartphone apps; application; feedback; users; data; data visualization; mobile phone

XSL•FO
RenderX

Introduction

Remote measurement technology (RMT) involves the use of wearable devices and smartphone apps to measure health outcomes in everyday life. Real-time, continuous symptom tracking has the potential to revolutionize the self-management of chronic conditions [1]. Feedback of data, in the form of data visual representations, can facilitate patient-driven health care and present opportunities for intervention by raising awareness of symptom patterns or prompting health appointments [2-5]. There is an emerging literature on "data visualization," which is defined as "the visual representation of data, encoded using position, length, size, and or color, among others, to reduce complexity and effectively communicate information to support discovery and understanding of patterns within data, decision-making, and memory" [6].

RMT users living with neurological and psychiatric conditions consistently express a desire to visualize their recorded data [7-9]. This includes both condition-specific metrics, as well as measures of broader health such as physical activity and sleep patterns. This could help to increase RMT users' engagement with the recording technology [10,11], improve communication with others involved in their care [12,13], validate their experiences [11], and enhance data-driven self-management [7,8,12], but only if the feedback is meaningful, and therefore, useful [14].

There is heterogeneity in the data visualization preferences of RMT users living with neurological and psychiatric conditions [15,16], but also some common overarching themes. Multiple studies have found that graphical data depictions [15,17], images, and color [16,17] enhance the RMT experience. A key principle is the generation of visualizations that are intuitive and clear [9,18], with a sensitivity to how users' current health status affects their interests and ability to engage with RMT [19], making it clinically beneficial. Users have expressed a preference for data visualization that incentivizes and rewards positive health outcomes and behaviors, as opposed to providing negative feedback [16,20] and the ability to customize their visualization experience, for example, altering how often data is presented, and in what format and quantity it is visualized [12,21,22]. While user experience research has been conducted for pre-existing apps tailored toward the management of our 3 chosen relapsing conditions: multiple sclerosis (MS) [9], epilepsy [18], and depression [10,23], and has identified themes to consider in device selection and RMT design for individuals living with these conditions [19], none of these previous studies have focused specifically on data visualization. Existing studies focus broadly on multiple dimensions of service users' design preferences and RMT user experiences (eg, health variables of perceived importance and perceived quality of medical advice provided) as opposed to data visualization preferences. This study was designed to address this gap. We chose to separate the 3 groups in terms of the participants' diagnoses so that we could compare and contrast needs, and find common themes that were relevant transdiagnostically or different themes that were specific to each group.

A triangulation approach was adopted to integrate multiple methods of data collection and multiple sources of data [24]. This multimethod qualitative design facilitated the identification and validation of a thematic coding frame to improve our understanding of service users' preferences and experiences of data visualization in RMT. Triangulation has been found to be an effective method for enhancing the credibility and validity of research findings [24,25] as synthesizing multiple data sources can help reduce bias and enrich data sets by broadening the scope and providing a more comprehensive picture [26,27]. This study builds on two sources of data (already published): (1) a systematic review [28], and (2) analyses of user reviews from existing self-management apps [29,30].

Methods

Overview

We ran several focus group sessions that allowed for the identification of new insights and an exploration of themes that had not been discovered in the previous data sources. This approach enabled a well-rounded exploration of visualization design preferences and permitted cross-validation of the thematic coding frame.

Ethical Considerations

We sought and were granted ethical approval via the UK Health Research Authority's research ethics committee (19/LO/1759).

Design of the Focus Groups

We recruited people with one of two neurological conditions: epilepsy or MS, or a diagnosis of major depressive disorder (MDD). Participants were eligible to participate if they met the following inclusion criteria; were at least 18 years old, fluent in English, able to provide informed consent, and had a clinical diagnosis of epilepsy, MS, or MDD (with an episode within the last 2 years). Participants with a confirmed diagnostic history of depression were recruited through the RADAR-CNS (Remote Assessment of Disease And Relapse - Central Nervous System) study [31], those with epilepsy and MS self-identified as having a diagnosis of one of these conditions and were recruited through nonprofit organizations (Epilepsy Action and MS Society) and were invited to join condition-specific focus groups.

Contextual Measures

The following measures were completed to help characterize our samples of people with epilepsy, MS, or MDD: (1) demographic information (age, gender, ethnicity) was completed by all participants; (2) Patient Health Questionnaire [32] was completed by the MDD group. This is a validated 9-item measure of the severity of symptoms associated with MDD, with a threshold of ≥10 indicating current clinical problems; (3) Liverpool Seizure Severity Scale [33] was completed by the epilepsy group. It is a validated 20-item measure of patient-perceived seizure severity, and (4) Patient Determined Disease Steps scale [34] was completed by the MS group. It is a validated patient-reported outcome of disability in MS.

XSL•FO
RenderX

Procedure

We held 12 focus groups, 4 for each health condition. Participants discussed data visualization with prompts from a topic guide (Multimedia Appendix 1). An initial analysis of the topics discussed in that focus group was sent to the participants who considered them and provided feedback in one of a further 6 "theme-checking" focus groups, which were held 2 weeks later (Figure 1). All questionnaires and consent forms were gathered by email and focus groups were conducted digitally, using videoconferencing software. The focus groups were recorded and transcribed prior to data analysis.

Figure 1. Flowchart of participants included in the focus groups and member checking sessions. MDD: major depressive disorder; MS: multiple sclerosis.

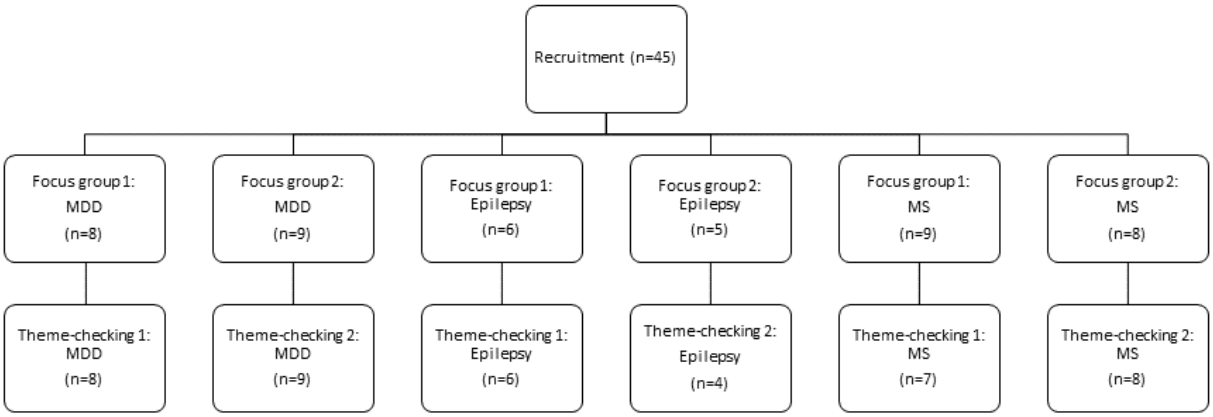

Data Analysis

The focus group discussions were shaped by a coding frame, which was developed through an iterative process, incorporating emerging themes from (1) the systematic review and (2) the analyses of user reviews for self-management apps (Figure 2). Focus group discussions were analyzed using a thematic approach. Content was analyzed deductively using the coding frame and refined inductively to reflect emerging themes. A total of 2 reviewers (ED-L and GG) coded the focus group transcripts and following coding, each reviewer suggested additions and revisions to the original coding frame. All disagreements were resolved through discussion. Reviews were then reread, and codes were compiled into descriptive themes. Novel themes emerging from the analysis were incorporated into the final version of the coding frame. The software NVivo (Lumivero) was used to help manage the data.

Figure 2. Coding frame development process.

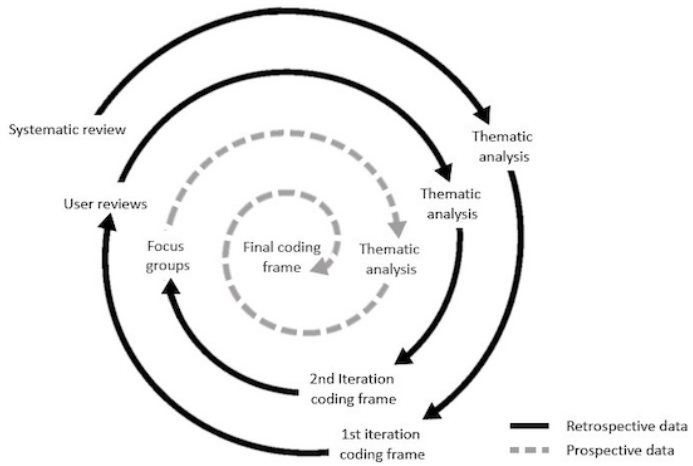

XSL•FO
RenderX

Patient and Public Involvement

This research was codeveloped with an advisory board of people with lived experience of either depression, epilepsy, or MS who were gathered together to advise on the following program of study. The advisory board had oversight of the themes in the coding frame and helped with the interpretation of results. Representatives of this board are coauthors who have critically reviewed this paper.

Results

Focus Group Participant Characteristics

In total, 45 people were recruited to take part in the focus groups across the 3 health conditions (Figure 1). Participant characteristics are summarized in Table 1. People with epilepsy were mostly symptomatic at the time of participation (and rated themselves on average relatively high on severity), people in the depression group scored in the mild range but had a history of at least one episode of MDD in the last 2 years, and most people in the MS group reported a moderate level of ongoing disability.

Table 1. Participant characteristics (N=45).

Characteristic	Depression	Epilepsy	MS[a]	Overall
Sex assigned at birth, n/n (%)				
Female	15/18 (83)	10/11 (91)	11/16 (69)	36/45 (80)
Male	3/18 (17)	1/11 (9)	5/16 (31)	9/45 (20)
Ethnicity, n/n (%)				
White or White British	15/18 (83)	11/11 (100)	16/16 (100)	42/45 (93)
Mixed	1/18 (6)	0/11 (0)	0/16 (0)	1/45 (2)
Other	2/18 (11)	0/11 (0)	0/16 (0)	2/45 (5)
Highest education completed, n/n (%)				
Degree level	14/18 (78)	6/10 (60)	10/16 (63)	30/44 (68)
A level	2/18 (11)	3/10 (30)	5/16 (31)	10/44 (23)
GSCE[b] level	2/18 (11)	1/10 (10)	1/16 (6)	4/44 (9)
Age (in years), n (mean, SD; range)	17 (55, 13; 27-69)	10 (41, 15; 23-62)	14 (50, 14; 33-73)	41 (50, 15; 23-73)
Age first diagnosed (in years), n (mean, SD; range)	17 (25, 12; 13-49)	10 (24, 16; 3-60)	14 (36, 11; 22-54)	41 (29, 13; 3-60)
MS diagnosis, n/n (%)				
Relapsing remitting	0/0 (0)	0/0 (0)	13/16 (81)	__[c]
Primary progressive	0/0 (0)	0/0 (0)	1/16 (6)	—
Secondary progressive	0/0 (0)	0/0 (0)	2/16 (13)	—
Symptom severity, n (mean, range)	17 (10.24, 2-17)	17 (57.35, 44.71)	10 (2.38, 0-6)	—

[a]MS: multiple sclerosis.

[b]GSCE: general certificate of secondary education.

[c]Not applicable.

Thematic Coding Frame

In general, focus group discussions aligned with the first iteration of the coding frame for app-based health data visualization developed by Polhemus et al [28,29,30] (Table S1 in Multimedia Appendix 2). However, the focus groups yielded additional insights that had not been discussed in academic studies or app reviews, resulting in 14 novel subthemes and 4 expanded core themes (Figure 3). Despite the additions, the thematic structure of the coding frame remained unchanged, comprising five central themes: (1) design preferences and recommendations, (2) data reporting, (3) impact of visualization, (4) moderators of visualization preferences, and (5) system-related factors and features. Table S2 in Multimedia Appendix 2 provides a list of quotes that are illustrative of these themes.

XSL•FO
RenderX

Figure 3. Visualization of the 14 novel subthemes and 4 expanded core themes.

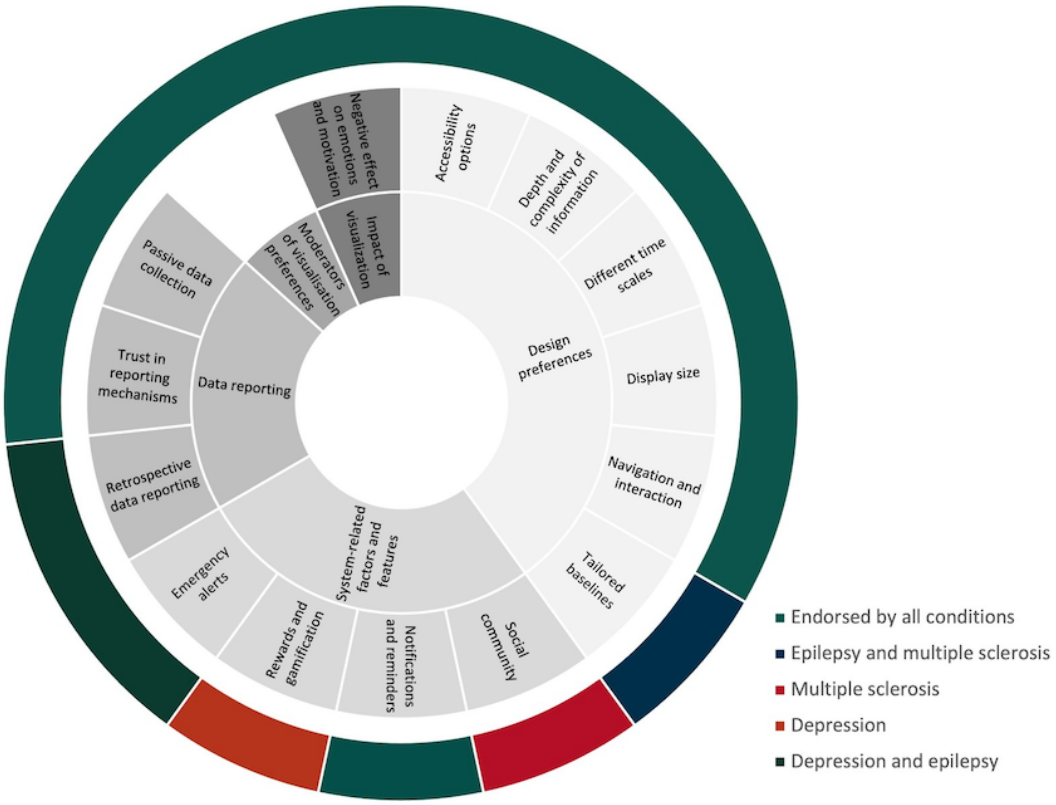

Major Theme: Design Preferences and Recommendations

As expected, this was the largest of all the themes emerging from the data. Bar and line graphs were the preferred format, regardless of diagnostic group, as they are easy to understand and can effectively visualize trends. All groups appreciated how numbers can help quantify experiences and that text can provide supplementary information (eg, contextual annotations). There was agreement within the groups that color influenced clarity and supported participants' understanding of the data. Across conditions, the ability to manipulate the data display size was considered beneficial but was particularly critical for individuals with MS who often have visual impairments. For some, images, and the ability to upload photographs were thought useful. Members of the depression group thought animation might also be an engaging way of presenting data. The overarching sentiment across groups was that it was crucial for data to be simple and clear, with the option to view a greater depth and complexity of information, when required. All suggested that accessibility options, such as the ability to select color schemes and the availability of voice-activated software, were key to accommodating any visual and motor impairments. More generally, participants wanted a good user experience with ease of use, emphasizing straightforward navigation and interaction in the app.

All 3 groups expressed a desire for customization of data visualizations such as choosing app displays to suit their individual needs or interests. Participants also described the advantages of adaptive interfaces to select and present data of relevance to the user. For instance, when a user is experiencing an episode of depression, a seizure, or MS relapse, the app might respond by reducing visualizations of a decline in health or by increasing positive feedback through responses to behaviors captured in the app. People with MS were interested in the idea of having tailored baselines, which would ensure goals were realistic and that progress visualizations were motivating. All groups saw the value in being able to select, manipulate, and compare data streams, spotlighting the use of filtering data by different timescales (eg, daily, weekly, and monthly) to visualize immediate information and long-term trends.

Minor Theme 1: Data Reported

Participants from all 3 diagnostic groups highlighted additional metrics of importance that they agreed that they wished to track (Figure 4). Although there were condition-specific metrics, there were several intersecting measures that were important across conditions. The epilepsy and depression groups highlighted the need for a retrospective reporting capability, which allowed for flexibility around data input. The benefits of continuous, passive data collection were proposed by all groups. However, to maintain trust in reporting mechanisms, participants emphasized that data accuracy must be a top priority for RMT developers. These concerns were evident in instances where participants described experiences of apps that had collected data that was perceived to be inaccurate. Despite these concerns, participants still agreed on the benefits of translating subjective feelings (eg, mood) into "objective" or numeric data.

XSL•FO
RenderX

Figure 4. Independent and overlapping metrics of importance for the 3 diagnostic groups.

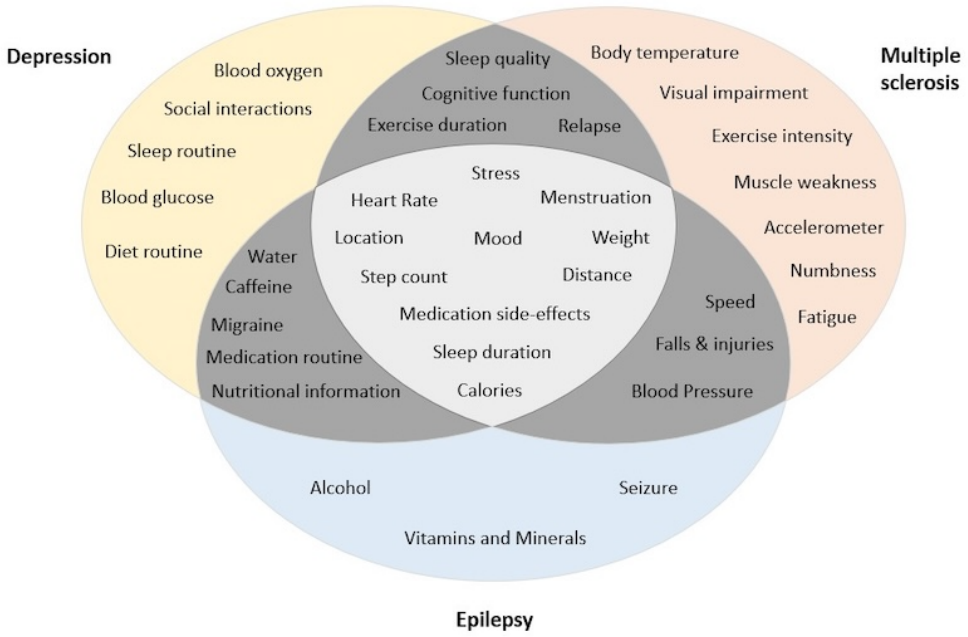

Minor Theme 2: Impact of Visualization

Participants from all 3 diagnostic groups believed that data visualization would improve RMT engagement, participants felt that seeing their data would help to validate experiences, as well as structure and organize their symptoms, enabling proactive self-management. By identifying patterns and triggers, participants could become more self-aware of their health condition, effectively monitor progress, and possibly alter people's experiences of their health condition. Another key impact was how data visualizations would affect an individual's emotions and self-image. They could improve self-esteem by providing a sense of empowerment over their condition, on the other hand, negative data could lead to anxiety. All groups highlighted the potential of data visualizations to facilitate sharing and communication about their condition, particularly with health care professionals, but also government agencies and employers to validate and help accommodate their disability. This information could help family, friends, and carers be better informed, as well as members of the public (eg, in emergency situations such as a seizure).

Minor Theme 3: Moderators of Visualization Preferences

This was the only theme in the coding frame not expanded by the focus group discussions, and all subthemes were mentioned across the 3 diagnostic groups. Participants described how their health status and intended use of the app affected their needs and requirements. Preferences were further moderated by experience with health monitoring, as it allowed participants to gain a better understanding of RMT and how best to use it.

Minor Theme 4: System-Related Factors and Features

Participants had clear ideas about system-related factors and features. People in the depression and epilepsy groups wanted an app to provide visual emergency alerts if data indicated risk. More general "pop-up" notifications and reminders about upcoming appointments, medication, and positive behaviors (eg, exercise) were seen as beneficial by all diagnostic groups. Apps that used reward systems or gamification were considered a useful way of encouraging long-term engagement, especially in periods of low mood for people with depression. A social community feature would allow them to link with people with the same health condition and was seen as a benefit. Having the ability to export data was also seen as crucial for facilitating communication with health care practitioners.

Discussion

Principal Findings

This study provides the first cross-diagnostic analysis of data visualization preferences for RMT apps. Preferences only varied slightly between the conditions and visualization preferences coalesced under five broad themes (Table S1 in Multimedia Appendix 2). These themes corroborate and expand on our previous research looking at RMT user experience for individuals with epilepsy [18], MS [9,19], mental health conditions [10,12,15,17,21,23], and other neurological conditions [16,20,22], using a triangulation approach. We found that when asked directly, people had preferences for a combination of passive and active data collection that had not emerged in the previous data sources (ie, in our systematic review and analysis of app user reviews). A clearer emphasis on the importance of accuracy also arose. Some design preferences were unique, including discussions on the need for a mobile app to be accessible, with the size and complexity of the information displayed seen as key considerations to enable ease of navigation and interaction. While an interest in customization has been raised in the previous literature [35], the idea of using tailored symptom baselines, as well as flexible

XSL•FO
RenderX

time scales, were novel insights from the focus group discussions. Participants expressed a desire for adaptive interfaces, which could use machine learning techniques to provide a more personalized user experience and feedback. Incentivization was considered important to support engagement with a mobile app, with new ideas around the use of gamification, social comparison, and the use of notifications emerging.

The impact of visualizations is such that health apps can enhance engagement and awareness, provide structure, and empower individuals to manage their conditions. Participants appreciate apps that are simple and intuitive but are simultaneously able to provide flexibility to accommodate individual health management needs. All diagnostic groups expressed a preference for pictures or graphics to display data and emphasized the need for clear formats with sufficient contextual information to allow users to navigate and understand their data. Overall, findings suggest that data visualizations can be an effective tool for health self-management when they align with users' individual needs, interests, and goals. However, this is heavily reliant on the type of data that can be reported and the wider system that the app is placed within. These factors need to be considered in relation to data visualization preferences in future research.

Limitations and Future Directions

The study design was a notable strength, as the triangulated framework enhanced validity and credibility, and the cross-diagnostic sample enhanced applicability and served as a point of comparison. The authors acknowledge that it was difficult to focus participants on the topic of data visualization, particularly as RMT design and data reporting are so intrinsically linked to data visualization. We recommend that future research studies adopt a more active approach to gathering individuals' views on data visualization, for example, by asking participants to design their app display and then discuss their ideas. It should also be noted that where the topic guides in this study focused on the use of RMT for reviewing raw symptom data, an emerging subsection of the field is further exploring the potential for RMT data to predict deterioration or relapse [31]. Indeed, some participants mentioned the perceived value of emergency "risk" alerts within apps. Further work should look to understand how best to implement relapse prediction within data visualization components. Some individuals with depression had greater experience of using wearable devices, nonetheless, this transdiagnostic study detected few differences between the groups. However, examining data visualization preferences in a broader sample and recruiting from clinical services will ensure that these findings generalize to the wider population.

Practical Implications and Conclusions

This study offers an opportunity for app designers to understand the data visualization preferences and priorities of individuals living with depression, epilepsy, and MS. The practical implications are considerable. RMT limits visualizations to mobile devices and introduces challenges related to scalability including reduced display sizes and resolution, lower computational power, and storage and interaction challenges associated with the smaller device size [14,36]. The need for simple visualizations matches with the current state-of-the-art data visualization style guide [36] and could be realized through a myriad of "out-of-the-box" solutions, for example, Google Charts. However, building in qualities, such as flexibility, intuitiveness, customization, and enhanced accessibility, requires increased development and resources. Similarly, considerable research efforts are still required to improve the generalizability and robustness of machine learning approaches within mobile health to provide a level of personalization that is reliable and supports clinical practice [14,37]. Together, these challenges point to the need for truly multidisciplinary co-design processes to meet patient preferences and priorities.

Acknowledgments

This paper was written in support of the RADAR-CNS (Remote Assessment of Disease and Relapse - Central Nervous System) program [38], a collaborative research effort focusing on the development of RMT for monitoring central nervous system disease progression. We acknowledge all partners in the RADAR-CNS consortium for their input and support of this work including our patient advisory board and the support we received from Epilepsy Action and the MS Society with recruitment. The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (grant 115902). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). This communication reflects the views of the RADAR-CNS consortium and neither the Innovative Medicines Initiative nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. This paper represents an independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre in South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. TW and MH are National Institute of Health Research Senior Investigators.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group topic guide.
[DOCX File , 14 KB - jmir_v26i1e43954_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Additional tables.
[DOCX File , 23 KB - jmir_v26i1e43954_app2.docx]

References

1. Chiauzzi E, Rodarte C, DasMahapatra P. Patient-centered activity monitoring in the self-management of chronic health conditions. BMC Med 2015;13(1):77 [FREE Full text] [doi: 10.1186/s12916-015-0319-2] [Medline: 25889598]
2. Simblett S, Greer B, Matcham F, Curtis H, Polhemus A, Ferrão J, et al. Barriers to and facilitators of engagement with remote measurement technology for managing health: systematic review and content analysis of findings. J Med Internet Res 2018;20(7):e10480 [FREE Full text] [doi: 10.2196/10480] [Medline: 30001997]
3. Simblett S, Matcham F, Siddi S, Bulgari V, di San Pietro CB, López JH, et al. Barriers to and facilitators of engagement with mHealth technology for remote measurement and management of depression: qualitative analysis. JMIR Mhealth Uhealth 2019;7(1):e11325 [FREE Full text] [doi: 10.2196/11325] [Medline: 30698535]
4. Simblett SK, Bruno E, Siddi S, Matcham F, Giuliano L, López JH, et al. Patient perspectives on the acceptability of mHealth technology for remote measurement and management of epilepsy: a qualitative analysis. Epilepsy Behav 2019;97:123-129. [doi: 10.1016/j.yebeh.2019.05.035] [Medline: 31247523]
5. Simblett SK, Evans J, Greer B, Curtis H, Matcham F, Radaelli M, et al. Engaging across dimensions of diversity: a cross-national perspective on mHealth tools for managing relapsing remitting and progressive multiple sclerosis. Mult Scler Relat Disord 2019;32:123-132. [doi: 10.1016/j.msard.2019.04.020] [Medline: 31125754]
6. Backonja U, Chi N, Choi Y, Hall AK, Le T, Kang Y, et al. Visualization approaches to support healthy aging: a systematic review. J Innov Health Inform 2016;23(3):860 [FREE Full text] [doi: 10.14236/jhi.v23i3.860] [Medline: 28059694]
7. Bardram JE, Frost M, Szántó K, Marcu G. The MONARCA self-assessment system: a persuasive personal monitoring system for bipolar patients. 2012 Presented at: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; January, 2012; Florida, Miami, USA p. 21-30. [doi: 10.1145/2110363.211037]
8. Eisner E, Drake RJ, Berry N, Barrowclough C, Emsley R, Machin M, et al. Development and long-term acceptability of express, a mobile phone app to monitor basic symptoms and early signs of psychosis relapse. JMIR Mhealth Uhealth 2019;7(3):e11568 [FREE Full text] [doi: 10.2196/11568] [Medline: 30924789]
9. Schleimer E, Pearce J, Barnecut A, Rowles W, Lizee A, Klein A, et al. A precision medicine tool for patients with multiple sclerosis (the open MS bioscreen): human-centered design and development. J Med Internet Res 2020;22(7):e15605 [FREE Full text] [doi: 10.2196/15605] [Medline: 32628124]
10. Rohani DA, Tuxen N, Lopategui AQ, Faurholt-Jepsen M, Kessing LV, Bardram JE. Personalizing mental health: a feasibility study of a mobile behavioral activation tool for depressed patients. 2019 Presented at: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 20, 2019; New York, NY, United States URL: https://www.bardram.net/wp-content/papercite-data/pdf/rohani2019personalizing.pdf
11. Matthews M, Voida S, Abdullah S, Doherty G, Choudhury T, Im S, et al. In situ design for mental illness: considering the pathology of bipolar disorder in mHealth design. 2015 Presented at: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services; August, 2015; New York, NY, United States p. 86-97. [doi: 10.1145/2785830.278586]
12. Murnane EL, Cosley D, Chang P, Guha S, Frank E, Gay G, et al. Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: implications for the design of technologies to manage mental health. J Am Med Inform Assoc 2016;23(3):477-484. [doi: 10.1093/jamia/ocv165] [Medline: 26911822]
13. Wannheden C, Revenäs Å. How people with Parkinson's disease and health care professionals wish to partner in care using eHealth: co-design study. J Med Internet Res 2020;22(9):e19195 [FREE Full text] [doi: 10.2196/19195] [Medline: 32955448]
14. Marzano L, Bardill A, Fields B, Herd K, Veale D, Grey N, et al. The application of mHealth to mental health: opportunities and challenges. Lancet Psychiatry 2015;2(10):942-948. [doi: 10.1016/S2215-0366(15)00268-0] [Medline: 26462228]
15. Forchuk C, Reiss JP, O'Regan T, Ethridge P, Donelle L, Rudnick A. Client perceptions of the mental health engagement network: a qualitative analysis of an electronic personal health record. BMC Psychiatry 2015;15:250 [FREE Full text] [doi: 10.1186/s12888-015-0614-7] [Medline: 26467210]
16. Garzo A, Silva PA, Garay-Vitoria N, Hernandez E, Cullen S, De Cock VC, et al. Design and development of a gait training system for Parkinson's disease. PLoS One 2018;13(11):e0207136 [FREE Full text] [doi: 10.1371/journal.pone.0207136] [Medline: 30418993]
17. McClelland GT, Fitzgerald M. A participatory mobile application (app) development project with mental health service users and clinicians. Health Educ J 2018;77(7):815-827. [doi: 10.1177/0017896918773790]
18. Yoo S, Lim K, Baek H, Jang S, Hwang G, Kim H, et al. Developing a mobile epilepsy management application integrated with an electronic health record for effective seizure management. Int J Med Inform 2020;134:104051. [doi: 10.1016/j.ijmedinf.2019.104051] [Medline: 31837500]
19. Giunti G, Kool J, Romero OR, Zubiete ED. Exploring the specific needs of persons with multiple sclerosis for mHealth solutions for physical activity: mixed-methods study. JMIR Mhealth Uhealth 2018;6(2):e37 [FREE Full text] [doi: 10.2196/mhealth.8996] [Medline: 29426814]

XSL•FO
RenderX

20. Erten-Uyumaz B, Ahuja M, Vacaretu T, Rama MD, Overeem S, Visser T, et al. Design and evaluation of a negotiation-based sleep scheduler app for insomnia treatment. 2019 Presented at: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 20, 2019; New York, NY, United States p. 225-233. [doi: 10.1145/3329189.3329218]
21. Proudfoot J, Parker G, Pavlovic DH, Manicavasagar V, Adler E, Whitton A. Community attitudes to the appropriation of mobile phones for monitoring and managing depression, anxiety, and stress. J Med Internet Res 2010;12(5):e64 [FREE Full text] [doi: 10.2196/jmir.1475] [Medline: 21169174]
22. Quaedackers L, De Wit J, Pillen S, Van Gilst M, Batalas N, Lammers GJ, et al. A mobile app for longterm monitoring of narcolepsy symptoms: design, development, and evaluation. JMIR Mhealth Uhealth 2020;8(1):e14939 [FREE Full text] [doi: 10.2196/14939] [Medline: 31909723]
23. Fuller-Tyszkiewicz M, Richardson B, Klein B, Skouteris H, Christensen H, Austin D, et al. A mobile app-based intervention for depression: end-user and expert usability testing study. JMIR Ment Health 2018;5(3):e54 [FREE Full text] [doi: 10.2196/mental.9445] [Medline: 30139722]
24. Denzin NK. The Research Act: A Theoretical Introduction to Sociological Methods. Oxfordshire, England, UK: Routledge; 2017.
25. Noble H, Heale R. Triangulation in research, with examples. Evidence-Based Nurs 2019;22(3):67-68. [doi: 10.1136/ebnurs-2019-103145] [Medline: 31201209]
26. Johnson M, O'Hara R, Hirst E, Weyman A, Turner J, Mason S, et al. Multiple triangulation and collaborative research using qualitative methods to explore decision making in pre-hospital emergency care. BMC Med Res Methodol 2017;17(1):11 [FREE Full text] [doi: 10.1186/s12874-017-0290-z] [Medline: 28118817]
27. Thurmond VA. The point of triangulation. J Nurs Scholarship 2001;33(3):253-258. [doi: 10.1111/j.1547-5069.2001.00253.x] [Medline: 11552552]
28. Polhemus A, Novak J, Majid S, Simblett S, Morris D, Bruce S, et al. Data visualization for chronic neurological and mental health condition self-management: systematic review of user perspectives. JMIR Ment Health 2022;9(4):e25249 [FREE Full text] [doi: 10.2196/25249] [Medline: 35482368]
29. Polhemus A, Simblett S, Dawe Lane E, Elliott B, Jilka S, Negbenose E, et al. Experiences of health tracking in mobile apps for multiple sclerosis: a qualitative content analysis of user reviews. Mult Scler Relat Disord 2023;69:104435. [doi: 10.1016/j.msard.2022.104435] [Medline: 36493561]
30. Polhemus A, Simblett S, Dawe-Lane E, Gilpin G, Elliott B, Jilka S, et al. Health tracking via mobile apps for depression self-management: qualitative content analysis of user reviews. JMIR Hum Factors 2022;9(4):e40133 [FREE Full text] [doi: 10.2196/40133] [Medline: 36416875]
31. Matcham F, di San Pietro CB, Bulgari V, de Girolamo G, Dobson R, Eriksson H, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019;19(1):72 [FREE Full text] [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
32. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
33. Baker GA, Smith DF, Jacoby A, Hayes JA, Chadwick DW. Liverpool seizure severity scale revisited. Seizure 1998;7(3):201-205 [FREE Full text] [doi: 10.1016/s1059-1311(98)80036-8] [Medline: 9700832]
34. Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D. Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. BMC Neurol 2013;13:37 [FREE Full text] [doi: 10.1186/1471-2377-13-37] [Medline: 23617555]
35. Bauer AM, Iles-Shih M, Ghomi RH, Rue T, Grover T, Kincler N, et al. Acceptability of mHealth augmentation of collaborative care: a mixed methods pilot study. Gen Hosp Psychiatry 2018;51:22-29 [FREE Full text] [doi: 10.1016/j.genhosppsych.2017.11.010] [Medline: 29272712]
36. Choe EK, Dachselt R, Isenberg P, Lee B. Mobile data visualization (Dagstuhl seminar 19292). Dagstuhl Reports 2019;9(7):78-93. [doi: 10.4230/DagRep.9.7.78]
37. Cummins N, Schuller BW. Five crucial challenges in digital health. Front Digit Health 2020;2:536203 [FREE Full text] [doi: 10.3389/fdgth.2020.536203] [Medline: 34713029]
38. RADAR-CNS (Remote Assessment of Disease and Relapse – Central Nervous System). King's College London. URL: www.radar-cns.org

Abbreviations

MDD: major depressive disorder
MS: multiple sclerosis
RADAR-CNS: Remote Assessment of Disease And Relapse - Central Nervous System
RMT: remote measurement technology

XSL•FO
RenderX

Edited by T Leung; submitted 21.11.22; peer-reviewed by H Veldandi, E Andrikopoulou; comments to author 26.12.23; revised version received 18.03.24; accepted 31.03.24; published 18.10.24.

Please cite as:
Simblett S, Dawe-Lane E, Gilpin G, Morris D, White K, Erturk S, Devonshire J, Lees S, Zormpas S, Polhemus A, Temesi G, Cummins N, Hotopf M, Wykes T, RADAR-CNS Consortium
Data Visualization Preferences in Remote Measurement Technology for Individuals Living With Depression, Epilepsy, and Multiple Sclerosis: Qualitative Study
J Med Internet Res 2024;26:e43954
URL: https://www.jmir.org/2024/1/e43954
doi:10.2196/43954
PMID:

©Sara Simblett, Erin Dawe-Lane, Gina Gilpin, Daniel Morris, Katie White, Sinan Erturk, Julie Devonshire, Simon Lees, Spyridon Zormpas, Ashley Polhemus, Gergely Temesi, Nicholas Cummins, Matthew Hotopf, Til Wykes, RADAR-CNS Consortium. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Internet-Based Social Connections of Black American College Students in Pre–COVID-19 and Peri–COVID-19 Pandemic Periods: Network Analysis

Eun Lee[1], PhD; Heejun Kim[2], PhD; Yildiz Esener[2], MEd; Terika McCall[3,4,5], MBA, MPH, PhD

[1]Department of Scientific Computing, Pukyong National University, Busan, Republic of Korea
[2]Department of Information Science, University of North Texas, Denton, TX, United States
[3]Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, United States
[4]Division of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT, United States
[5]Center for Interdisciplinary Research on AIDS, Yale School of Public Health, New Haven, CT, United States

Corresponding Author:
Eun Lee, PhD
Department of Scientific Computing
Pukyong National University
45, Yongso-ro
Nam-gu
Busan, 48513
Republic of Korea
Phone: 82 10 7356 7890
Email: eunlee@pknu.ac.kr

Abstract

Background: A global-scale pandemic, such as the COVID-19 pandemic, greatly impacted communities of color. Moreover, physical distancing recommendations during the height of the COVID-19 pandemic negatively affected people's sense of social connection, especially among young individuals. More research is needed on the use of social media and communication about depression, with a specific focus on young Black Americans.

Objective: This paper aims to examine whether there are any differences in social-networking characteristics before and during the pandemic periods (ie, pre–COVID-19 pandemic vs peri–COVID-19 pandemic) among the students of historically black colleges and universities (HBCUs). For the study, the researchers focus on the students who have posted a depression-related tweet or have retweeted such posts on their timeline and also those who have not made such tweets. This is done to understand the collective patterns of both groups.

Methods: This paper analyzed the social networks on Twitter (currently known as X; X Corp) of HBCU students through comparing pre–COVID-19 and peri–COVID-19 pandemic data. The researchers quantified the structural properties, such as reciprocity, homophily, and communities, to test the differences in internet-based socializing patterns between the depression-related and non–depression related groups for the 2 periods.

Results: During the COVID-19 pandemic period, the group with depression-related tweets saw an increase in internet-based friendships, with the average number of friends rising from 1194 (SD 528.14) to 1371 (SD 824.61; $P<.001$). Their mutual relationships strengthened (reciprocity: 0.78-0.8; $P=.01$), and they showed higher assortativity with other depression-related group members (0.6-0.7; $P<.001$). In a network with only HBCU students, internet-based and physical affiliation memberships aligned closely during the peri–COVID-19 pandemic period, with membership entropy decreasing from 1.0 to 0.5. While users without depression-related tweets engaged more on the internet with other users who shared physical affiliations, those who posted depression-related tweets maintained consistent entropy levels (modularity: 0.75-0.76). Compared with randomized networks before and during the COVID-19 pandemic ($P<.001$), the users also exhibited high homophily with other members who posted depression-related tweets.

Conclusions: The findings of this study provided insight into the social media activities of HBCU students' social networks and communication about depression on social media. Future social media interventions focused on the mental health of Black college students may focus on providing resources to students who communicate about depression. Efforts aimed at providing relevant resources and information to internet-based communities that share institutional affiliation may enhance access to social

XSL•FO
RenderX

support, particularly for those who may not proactively seek assistance. This approach may contribute to increased social support for individuals within these communities, especially those with a limited social capacity.

(J Med Internet Res 2024;26:e55531) doi:10.2196/55531

KEYWORDS

COVID-19 pandemic; college students; Black American; African American; social network analysis; social media; mental health; depression

Introduction

Overview

The COVID-19 pandemic significantly impacted the mental health of college students [1]. Reports suggest that symptoms of depression, anxiety, and stress increased during the COVID-19 pandemic period [2]. Many factors that were exacerbated by the COVID-19 pandemic, such as housing, food insecurity, and physical isolation, may have negatively affected the mental health of individuals [1,2]. In addition, racial and ethnic minoritized communities were disproportionately impacted by the higher rates of COVID-19 infection and mortality in the United States [3-6]. Owing to structural health inequities, Black American college students are more likely to experience difficulties in accessing resources for basic needs, which may lead to deleterious mental health symptoms [7].

Expanding upon our previous study [8], which mainly focused on the differences in the internet-based conversation topics of the students of historically black colleges and universities (HBCUs; defined as "...any historically Black college or university established before 1964, with the principal mission of educating Black Americans..." [9]) depending on their depression expression using topic analysis and network analysis, this study shifts its focus to examine changes in internet-based social-connection patterns among 2 groups of HBCU students before and during the COVID-19 pandemic. The study includes (1) a depression-related group (d-r group), who have posted a depression-related tweet or a retweet in the pre–COVID-19 pandemic period, and (2) a non–d-r group, who have not posted depression-related tweets. We examined the changes in the 2 groups' internet-based social networks and investigated the social connectedness of this minoritized group based on whether each Twitter (currently known as X; X Corp) user had a history of mentioning depression. The pre–COVID-19 characteristics of their social networks, such as their mutuality, the extent of the clustering, and other structural properties, have been compared with the same characteristics during the peri–COVID-19 period. This was done to understand whether the aforementioned social network measures have changed during the COVID-19 pandemic or not and the implications of these findings.

Background

A global-scale pandemic, such as COVID-19, can greatly impact society and result in changes in personal social network structures. The physical restrictions imposed during the COVID-19 pandemic limited face-to-face interactions, which facilitated the expansion of their internet-based social networks [10]. For example, local area–based and internet-based networks emerged in the United Kingdom due to geographic restrictions (eg, stay-at-home orders) for regional volunteers and neighboring supports [11]. Limited social activities and communication, which may lower one's sense of social connection, can negatively affect mental health. People have used internet-based social platforms to fulfil their need for connectedness, resulting in a global increase in the average time spent on social media during the COVID-19 pandemic. The average time spent has increased from 54 minutes in 2018 to 65 minutes in 2022 [12,13]. Researchers found that those who engage in these internet-based social activities can experience relief from negative mental-health symptoms of depression, loneliness, and anxiety during the COVID-19 pandemic [14-16]. In particular, this internet-based connectedness can be crucial to minoritized groups during the COVID-19 pandemic. This is crucial not only to attenuate the negative mental effects but also to prompt people to share accurate and up-to-date health-related information [17].

The pandemic also had negative implications for mental health and well-being among the young communities of color. A study found that Generation Z adults (aged 18 to 23 years) were the most stressed generation during the COVID-19 pandemic, reporting increased mood symptoms, stress, and alcohol use during the COVID-19 pandemic [18].

In particular, Black communities experienced significantly higher rates of COVID-19 hospitalization and mortality [3-6,19], which can be attributed to systemic health inequalities, including disenfranchisement from the health care system and escalating mental health risks [20,21]. Acknowledging this unequal impact may correlate with distinct coping strategies for social connections in these communities [22]. Black college students, as younger members of their communities, likely encountered unique challenges due to the swift changes in their academic, occupational, housing, and social environments, exacerbated by the disproportionate effects of the COVID-19 pandemic [19]. The survey study by Molock and Parchem [23] revealed substantial disruptions in various aspects of life of college students of color. These aspects included finances, living situations, academics, and career goals, demonstrating increased stress, anxiety, and depression, along with challenges related to racial injustice among the study participants.

Even though age and racial factors could influence minoritized communities (eg, college students of color) significantly during the COVID-19 pandemic, there is a paucity of research on the impact of the COVID-19 pandemic on their mental health and coping behaviors during this time [23-26]. Some studies investigating the emotional well-being of HBCU students found that it strongly correlates with their social connectedness to others, which is mediated by internet-based social platforms

XSL•FO
RenderX

[25]. Researchers have observed that strong social networks are a crucial resource for younger generations to cope with depressive symptoms [27]. This is particularly important for college students in Black communities [25]. While these studies offer valuable insights, they are constrained to conducting individual (ie, egocentric) survey analyses and do not focus on the collective behavioral patterns exhibited by HBCU groups.

Objectives

To fill this gap, this research focuses on the structural changes in the internet-based social networks of HBCU students to understand the collective internet-based behaviors of the group during the COVID-19 pandemic. By categorizing the students on the basis of their creation and engagement with depression-related content, this study aims to investigate how the aforementioned groups (d-r group and non–d-r group) have shown distinctive behavioral patterns in making friends on the internet and to describe the collective behavioral characteristics among the groups before and during the COVID-19 pandemic.

Methods

Overview

Firstly, the researchers selected 20 HBCUs and identified active users who are students of the HBCUs. We separated them into the d-r group and non–d-r group depending on depression-related mentions in their tweets (refer to detailed methods in our previously published paper [8]). Social network data (ie, IDs of followers and friends) of these Twitter users were collected for the pre–COVID-19 and peri–COVID-19 pandemic periods. We compared the changes in the 2 groups' social networks with a focus on their friends networks during the pre–COVID-19 and peri–COVID-19 pandemic periods. The variations of basic network characteristics such as (1) the number of friends or followers, (2) reciprocity, and (3) assortativity of the d-r group before and during the COVID-19 pandemic were assessed; their statistical differences were tested with the paired 2-tailed *t* test. In addition, we separated the HBCU students' social networks from other users to focus on the structural changes in the HBCU groups and the differences in structural patterns between d-r and non–d-r groups during the COVID-19 pandemic.

Data Source and Processing

Initially, we compiled a list of 20 HBCUs with active Twitter followers. To ensure representative sampling, HBCUs were chosen on the basis of their geographical distribution, considering their historical concentration in the Southeastern region owing to past racial segregation practices. Therefore, HBCUs in the Southeast (14/20, 70%), Northeast (3/20, 15%), Southwest (2/20, 10%), and Midwest (1/20, 5%) of the United States were selected.

Subsequently, we identified Twitter users associated with the HBCUs. The examination of tweets pertaining to these educational institutions was conducted by one of the authors (TM), who compiled the list of mentions and hashtags relevant to each HBCU (eg, #school_name, @school_name_abbreviation). Consequently, we compiled a comprehensive list of users who authored tweets featuring these mentions and hashtags using Twitter's search application programming interface. We categorized these users into 2 groups: one group (d-r group) mentioning depression-related terms (eg, *depressed*, *#depressed*, *depression*, *#depression*, *dark place*, *depressing*, *#depressing*, *sad sad #SadSad*) on Twitter and another group (non–d-r group) without such indications.

The first dataset of these users for the pre–COVID-19 pandemic period was collected between May 13, 2019, and May 15, 2019. These data yielded 682 users for the non–d-r group and the same number of users for the d-r group. The second dataset of the same users for the peri–COVID-19 pandemic period was collected between July 19, 2022, and July 21, 2022. This was done to gain insight into the changes in social networks during the COVID-19 pandemic. The peri–COVID-19 pandemic period dataset yielded 508 users for the d-r group (508/682, 74.5% of the d-r group in the pre–COVID-19 pandemic period) and 512 users for the non–d-r group (512/682, 75.1% of the non–d-r group in the pre–COVID-19 pandemic period). There has been a notable decline in user subscriptions to the Twitter platform throughout the COVID-19 pandemic. However, we could track >70% of the previous users in the peri–COVID-19 pandemic period. The datasets encompassed various user attributes such as user ID, account description, and IDs of their followers and friends. In addition, tweet-related attributes, such as tweet ID, were included. Comprehensive details about compiling 20 HBCUs, selecting Twitter users associated with these HBCUs, and separating users into 2 groups are noted in our prior work [8].

Friendship Network Construction

On Twitter, there are 2 types of relationships: followers and friends. Followers denote people who decided to follow a person's account; friends are people who an individual account follows, as shown in Figure 1A. We analyzed only "friends" on Twitter as a friendship network because making friends reflects one's intentional social activities on Twitter, but following relationships are passively decided by others. When V follows W (ie, V→W), W is V's friend, as shown in Figure 1A. However, V is merely a follower and not a friend of W from W's perspective. Because we had followers and friends of the sampled HBCU students' in May 2019 and July 2022, these relationship data were the internet-based social network of HBCU students, including followers and friends, as shown in Figure 1B. By focusing on the friend relationships on Twitter, the network in panel Figure 1B can be transformed to Figure 1C.

From the 2 data collection periods (pre–COVID-19 and peri–COVID-19 pandemic), there were 2 friendship networks for each period; we compared the structural changes of those networks in the context of the COVID-19 pandemic to understand how these 2 groups' collective internet-based friendships looked different. For instance, Figure 1D displays an exemplary structural change in that W stopped following V and made a new friend, Z. To make a fair comparison, we analyzed the networks of 2019 users solely with those have remained in the 2022 dataset. As the first step, we incorporated all users along with their friends from the collected data to grasp macrolevel patterns. Subsequently, as the second step, we

XSL•FO
RenderX

included users only affiliated with the selected HBCUs to explore internal dynamics and microlevel patterns in greater detail.

Figure 1. An exemplary friendship network construction from a Twitter account relationship. (A) Two possible interpretations for a relationship between V and W: when V follows W, one can say W is a friend of V or V is a follower of W. (B) A sample social network of person V: dashed arrows represent followers of V, and plain arrows represent the friends of V. Each color follows the originating node's color. (C) Focusing on V's friendship network, the internet-based social network can be represented by 3 individuals (V, W, and Y nodes) and their 2 relationships (2 gray arrows). (D) An exemplary network change from pre–COVID-19 to peri–COVID-19 pandemic periods. The pre–COVID-19 network had 3 individuals or nodes V, W, and Y. Then, an individual Z is added to the peri–COVID-19 network, and transparent links between individuals W and V and Y and V disappear in the peri–COVID-19 network.

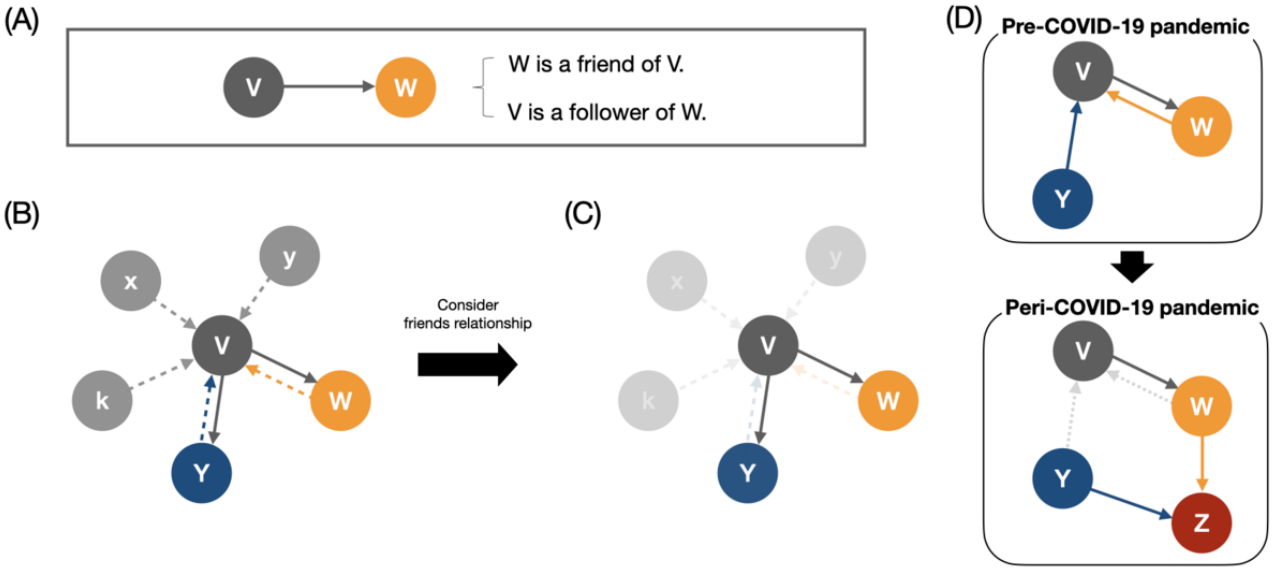

The Macroview: Network Analysis With All Users

Reciprocity

We defined reciprocity r_i for a node, which bears conceptual resemblance to "mutuality" in sociometry [28] and a network-level "reciprocity" in network science [29]. This was done to quantify the proportion of node i's friends who are in a mutual friendship with node *i*. For instance, let us consider nodes V and W in the upper panel of Figure 1D. Node V has 1 outgoing link to node W (eg, W is a friend of V), and node W has the same number of outgoing links to node V (eg, V is also a friend of W). Because V and W are friends of each other, one can say that there is 1 reciprocal link between nodes V and W. Then, the reciprocity r_V of node V is 1.0 based on equation 1,

(1)

In equation 1, Γ_i denotes a set of i's neighboring friends, and A represents an adjacency matrix of a network. When $A_{ij} = 1$, it means there is a link from node i to j, and $A_{ij} = 0$ represents a disconnection between node *i* and *j*. Therefore, r_i is in the range of (0-1) by its definition. The larger the reciprocity, the higher the mutual friendship between nodes.

Assortativity With the D-R Group of Users

The proportion of the d-r group of users among a node's friends (eg, a node's number of outgoing links) could inform us of an individual's tendency to make friends with the d-r group of users. This type of tendency connecting with similar ones is proposed in many studies [29,30], and in this study, we defined an individual-level "assortativity with the d-r group of users" of node i (ρ_i) as follows:

(2)

In equation 2, d_j denotes a binary value 1 (or 0) when a node is classified as a d-r user in the 2019 dataset (or a non–d-r user in the 2019 dataset). For these 2 network metrics and the number of friends, we used the friendship network for the 2019 and 2022 datasets (as shown in Figure 2).

XSL•FO
RenderX

Figure 2. Distributions of the number of friends, reciprocity, and assortativity of the students of historically black colleges and universities with depression-related users (d-r group) and non–depression-related users (non–d-r group). (A-C) The number of friends, reciprocity, and assortativity of the d-r group. (D-F) The number of friends, reciprocity, and assortativity of the non–d-r group. (A and D) Distributions of the number of friends of the d-r group and non–d-r group for pre–COVID-19 (2019; red) and peri–COVID-19 (2022; blue) pandemic periods. The inset shows the cumulative distribution function (CDF) of the number of friends. (B and E) The CDF of the reciprocity (*r*) of the d-r group and non–d-r group for 2019 and 2022. (C and F) The CDF of the assortativity with d-r group (ρ) for each year.

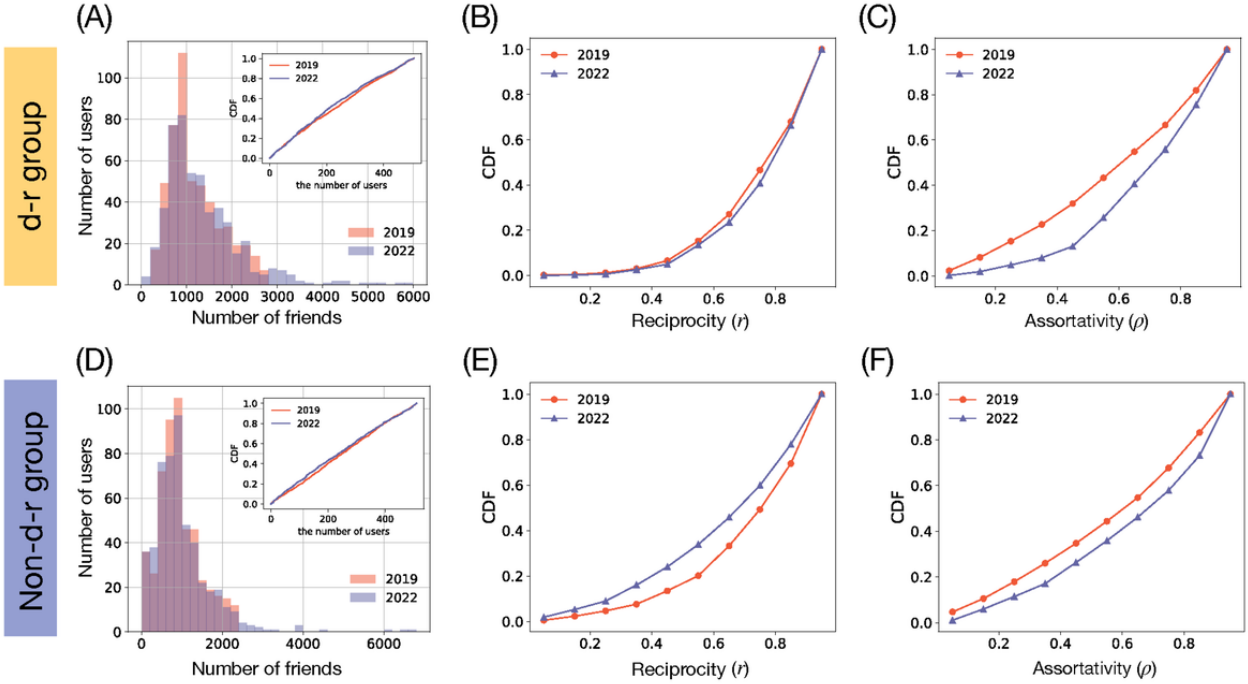

Analysis of Networks Among HBCU Students

Community Detection

From the friendship network (ie, focusing on the friends' relationships on Twitter), the largest connected component (LCC) was extracted for each period (pre– or peri–COVID-19 pandemic periods), and the LCC was used for community detection, excluding non-HBCU students. Thus, the size of the LCC for each period was distinct because of different connectivity structures. The LCC of the pre–COVID-19 pandemic period network contains 939 nodes, and the LCC of the peri–COVID-19 pandemic period network includes 827 nodes. We applied the *louvain_communities* function of the *NetworkX* module (Python Software Foundation). This function finds the best partition of a graph based on the Louvain community detection algorithm [31]. Therefore, a node in a network can have the best community membership with an integer number.

Membership Entropy

We compared the community membership with the physical affiliations of HBCU students' accounts. The similarity between a user's physical affiliation (eg, a university) and their community membership in internet-based network is evaluated as an information entropy, where the fraction of users with a school affiliation in a community is applied as a probability (p) in the information entropy equation: $S = -p\log(p)$. This entropy quantifies a community's affiliation complexity in LCC, and we averaged the entropy across all communities as follows:

(3)

In equation 3, C represents each community, |C| is the total number of communities in a network, and *Sc* denotes the entropy of a community C. The smaller the entropy for a network, the more identical affiliation membership exists in the network's communities. Putting it differently, the communities in a network are closely matched with the physical affiliation of a college.

Coleman Homophily Index

The Coleman homophily index typically ranges from −1 to 1 [32]. A value of 0 indicates a neutral mix of the attributes, representing random connections. A value of 1 signifies complete homophily, where all connections occur within the group; a value of −1 denotes complete heterophily, indicating that all connections with individuals outside the group. When there are 2 groups called *V* and *W* with connections between them, we can assess the homophily of *V* (or *W*) as follows:

(4)

In equation 4, L_{VV} represents the number of connections between members in group *V*, M_V denotes the total number of connections made by group *V*, and f_V signifies the fraction of group *V*'s size. In this study, the groups are defined as the d-r and non–d-r groups, and h_c illustrates the homogeneity of each group in their

XSL•FO
RenderX

internet-based connections during both the pre– and peri–COVID-19 pandemic periods.

Attribute Randomized Network Construction

To compare the characteristics of the HBCU students' friendship network, this research randomized depression-related attributes, which were empirically observed in the LCCs of the friendship network from pre–COVID-19 and peri–COVID-19 pandemic periods. The study generated 100 randomized network samples for each period. The sample networks were structurally identical to the original HBCU students' friendship networks; however, only depression-related attributes were randomly mixed. To assess the statistical significance of the Coleman homophily index and membership entropy of the randomized networks with those of empirical networks, one can apply the permutation test to those metrics to confirm the statistical significance of their difference.

Ethical Considerations

To protect the rights and privacy of Twitter users, we followed strict ethical principles throughout the study. All personally identifiable information that could reveal users' identities was anonymized, and no sample tweet was presented in this paper. Only aggregated information was reported. Data collection, management, and analysis were conducted on a secure research cluster computer. No data were made available as open access.

The University of North Texas (IRB-21-623) and Yale University (2000032441) institutional review boards reviewed and approved the study protocol, determining that it meets the criteria for exemption for research involving human participants. IRB exemptions covered secondary analysis without additional consent for the secondary analyses of this study.

Results

The Macroview: Friendship Networks of HBCU Students on Twitter

To understand the HBCU students' collective internet-based behaviors during the COVID-19 pandemic, first, we compared their friendship networks on Twitter between pre–COVID-19 (May 2019) and peri–COVID-19 (July 2022) pandemic periods, including all users regardless of their affiliation to the sampled HBCUs (Figure 2). In terms of the number of friends, d-r and non–d-r groups' internet-based friendships have grown on average with time. In particular, the d-r group (with depression-related tweets) expanded the average number of friends from 1194 before the COVID-19 pandemic to 1371 during the COVID-19 pandemic. When we statistically tested the average number of friends for 2 groups using the paired *t* test, the depression-related tweet users (d-r group) had statistically different numbers of friends during the pre–COVID-19 and peri–COVID-19 pandemic periods ($P<.001$), with an increased average number of friends (1194-1371). Conversely, the users without depression-related tweets (non–d-r group) showed a less significant increment ($P=.06$) than the d-r group, even though the former group's average number of friends slightly increased from 926 to 1007 before and during the COVID-19 pandemic.

The average number of friends and the distribution of the number of friends have changed in the context of the COVID-19 pandemic. The distributions of the number of friends for the d-r group of users in data published in 2019 and 2022 are displayed in Figure 2A. It is noticeable that the tail part stretched out more in 2022 than in 2019, showing the increased number of d-r group of users, with a large number of friends during the COVID-19 pandemic. The heavy-tailed number of friends (approximately 6000) confirms active engagement in making internet-based friends in the d-r group during the COVID-19 pandemic period. The non–d-r group of users' number of friends also increased in the tail part during the COVID-19 pandemic period (refer to Figure 2D); however, most non–d-r group of users maintained the same number of friends as that before the COVID-19 pandemic, resulting in insignificant difference in the average number of friends ($P>.05$).

Regarding the extent of mutual interactions, the reciprocal relationship between users in both groups displayed statistically significant change during the periods as shown in Table 1 (paired *t* test; d-r group $P=.01$ and non–d-r group $P<.001$) and Figure 2 (for the d-r group, refer to Figure 2B and for the non–d-r group, refer to Figure 2E). However, the pattern of changes in average reciprocity differed between groups. On an average, the d-r group has built up their mutual relationships from 0.78 to 0.8, whereas the non–d-r group has lowered their reciprocity from 0.73 to 0.67, as shown in Figures 2B and 2E). The growth in the d-r group's reciprocity suggests that users who mention depression on internet-based platforms may be more likely to form mutual relationships on the platform, unlike the non–d-r group.

In addition, the d-r group of users' assortativity with other d-r group members has been amplified from 0.6 to 0.7 ($P<.001$). Considering that the measure represents the proportion of the d-r group of users when compared with all their friends, it means the d-r group has been connected more to other d-r group of users during the COVID-19 pandemic period (Figure 2C). The non–d-r group of users kept a similar level of assortative connections with the d-r group of users during the period on average; it was not statistically significant (from 0.483 to 0.475, $P=.62$; Table 1 and Figure 2F). The statistically significant increments in the reciprocity (r) and the assortativity of the d-r users (ρ) show that the d-r group members have made mutual connections with other internet-based users and tend to connect themselves during the COVID-19 pandemic (as seen in Table 1).

XSL•FO
RenderX

Table 1. Paired *t* test to measure changes in the number of friends, reciprocity, and assortativity between the pre–COVID-19 and peri–COVID-19 pandemic periods for the depression-related (d-r) and non–depression-related (non–d-r) groups (with their *P* value regarding the paired *t* test).

Measures	Pre–COVID-19 pandemic period (May 2019)	Peri–COVID-19 pandemic period (July 2022)	*P* value
Non–d-r group			
Number of friends	926	1007	.06
Reciprocity	0.73	0.67	<.001
Assortativity	0.483	0.475	.62
D-r group			
Number of friends	1194	1371	<.001
Reciprocity	0.78	0.8	.01
Assortativity	0.6	0.7	<.001

The Microview 1: Connection Patterns of HBCU Students' Friendship Networks

The friendship networks used in the previous analysis included Twitter users who were not identified as HBCU students. Therefore, focusing on HBCU students' internet-based behaviors could be difficult. Hence, to observe specific characteristics of the HBCU communities, we separated their network from the whole network. After removing users who not identified as HBCU students from the previous LCC, we extracted the largest connected cluster that can represent most of the HBCU students' network and its connection properties. In the largest connected cluster that connects solely HBCU students, the remaining pre–COVID-19 and peri–COVID-19 pandemic networks included 939 and 827 HBCU students, respectively, including the d-r and non–d-r groups of users (Figures 3A and 3B). These numbers correspond to about 68.8% (939/1364) and 81.1% (827/1020) of the HBCU student sample in the pre–COVID-19 and peri–COVID-19 pandemic periods.

The communities are shown in Figures 3A and 3B, and those communities are determined by the Louvain algorithm [31] in Python's NetworkX package. This algorithm found 11 and 15 communities for pre–COVID-19 (Figure 3A) and peri–COVID-19 (Figure 3B) pandemic networks, respectively. It is noticeable that the peri–COVID-19 pandemic network is more segregated than the pre–COVID-19 pandemic network even though we displayed the network with the same parameters and layout algorithm (Cytoscape's Prefuse Force Directed Layout with default options), and the large numbers of communities for peri–COVID-19 pandemic confirm the different segregation (ie, 11 for pre–COVID-19 and 15 for peri–COVID-19 pandemic periods; Figure 3B). The increase of the clustering in the peri–COVID-19 pandemic period is also confirmed by the modularity, a metric of the segregation quality [33]. The pre–COVID-19 pandemic network has a modularity of 0.7; the peri–COVID-19 pandemic has a higher modularity of 0.75, confirming a more distinct and internally cohesive structure of the peri–COVID-19 pandemic network.

The pre–COVID-19 pandemic network has a shorter diameter of 11 (the longest distance between 2 nodes in an entire network) than the peri–COVID-19 pandemic one (diameter of 13). This is because the pre–COVID-19 pandemic network is more densely connected among the communities. The longer diameter in the peri–COVID-19 pandemic friendship network implies that the network requires a greater effort to disseminate information across the entire network. This change suggests that during the COVID-19 pandemic, students tended to form smaller, tightly-knit communities, potentially concentrating information sharing and communications within these groups.

When we compared the alignment with the community members' affiliation with the university (ie, their university name), we found that the communities identified by the friendship network were well aligned with the members' actual university (or college) affiliations. The community members' affiliation composition heterogeneity is calculated with the membership entropy to quantify the affiliation complexity (as referred in the Methods section). The peri–COVID-19 pandemic network demonstrates a smaller variance in *S* compared to the variance of *S* in the pre–COVID-19 pandemic network, as shown in the kernel density estimate plot in Figure 3C. The mean of the membership entropy during the pre–COVID-19 pandemic period is approximately 1.0; this value during the peri–COVID-19 pandemic period is decreased to approximately 0.5. Because larger entropy means a large discrepancy between physical affiliation and community membership in internet-based social networks, the relatively small membership entropy during the peri–COVID-19 pandemic period demonstrates that a structurally identified community contains more homogeneous users in terms of their affiliated universities when compared with the pre–COVID-19 pandemic period.

XSL•FO
RenderX

Figure 3. Detected communities in the internet-based social networks of the students from historically black colleges and universities during pre–COVID-19 and peri–COVID-19 pandemic periods. (A) This panel shows communities in the pre–COVID-19 pandemic period (data from May 2019). The colors represent 11 different communities detected by the Louvain community detection algorithm (Python Software Foundation). (B) The panel shows communities in the peri–COVID-19 pandemic period (data from July 2022), which are distinguished by 15 communities. (C) The kernel density estimation plot of the detected communities' group membership entropy is shown. The vertical lines display the average membership entropy for each period (pre–COVID-19: yellow; peri–COVID-19: blue), and the networks' layouts use Cytoscape program's Prefuse Force Directed Layout with default parameter options.

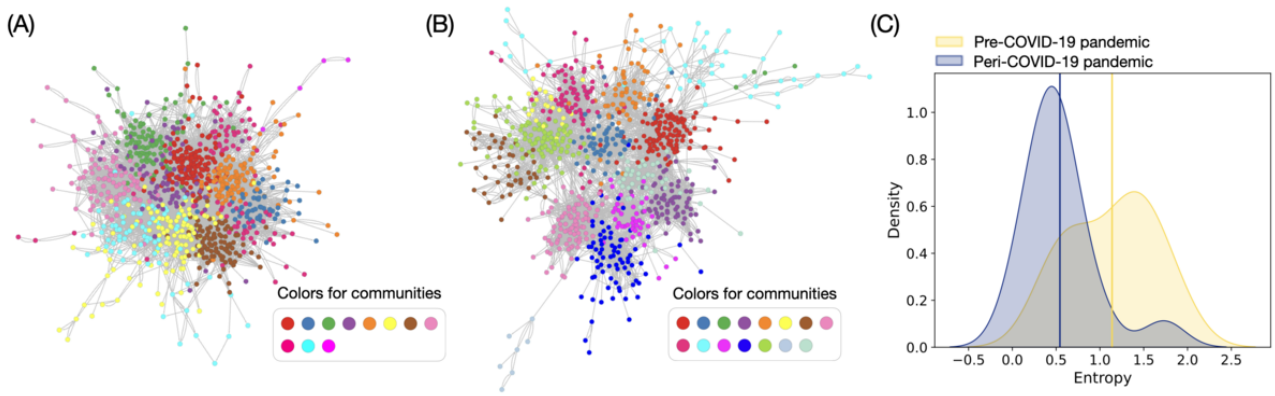

The Microview 2: Connection Patterns of the D-R and Non–D-R Groups Focusing on Filtered Friendship Networks of HBCU Students

When we extracted the friendship networks of a group of HBCU students (ie, d-r or non–d-r groups), the distinct structural patterns and homophilic connections could be clearly observed (as seen in Figures 4A and 4B). The more segregated community patterns of the non–d-r group during the COVID-19 pandemic, which is observed in Figure 4B, are correlated with the increased modularity of the non–d-r group of users (from 0.67 during pre–COVID-19 pandemic period to 0.73 during peri–COVID-19 pandemic period), while d-r group's modularity has not largely changed, as shown in Figure 4A (from 0.75 during pre–COVID-19 pandemic period to 0.76 during peri–COVID-19 pandemic period).

The mixing patterns of each group also exhibited different arrangements. The d-r group's average Coleman homophily index h_c = 0.41 in pre-COVID-19 pandemic has remained approximately the same with a slight decrease during the COVID-19 pandemic. To assess the significance of these changes in h_c, we conducted the permutation test by randomizing users' depression-related attributes to generate 100 sample networks that represent each group's friendship networks during pre–COVID-19 and peri–COVID-19 pandemic periods (seen in the Methods section). Figure 4C displays the distribution of h_c for the randomized friendship networks of the d-r users before COVID-19 pandemic (2019 dataset); Figure 4D shows these data for the d-r users during the COVID-19 pandemic (2022 dataset). As one can see, the homogeneity of the d-r group regarding the depression-related attribute is significantly different from the randomized results (P<.001). In short, this result emphasizes the d-r group's collective connectivity patterns with other d-r group members.

By contrast, the non–d-r group's homophily has increased from h_c= –0.16 before the COVID-19 pandemic to h_c= –0.04 during the COVID-19 pandemic. This means that the non–d-r group members connected more with other non–d-r group members during the pandemic, as shown in Figures 4E and 4F. In the pre–COVID-19 pandemic period, non–d-r users connected with different group members (negative h_c); however, they were linked more with other members of the same group in the peri–COVID-19 pandemic period (as seen in Figures 4B, 4E, and 4F).

The d-r group's average membership entropy has been preserved during the COVID-19 pandemic from 0.51 to 0.52, as shown in Figures 4G and 4H, even though for the peri–COVID-19 pandemic period is statistically insignificant. In contrast, the non–d-r group demonstrated a statistically significant decrease in from 1.22 to 0.73 (as seen in Figures 4I and 4J). It made us infer that the non–d-r group's communities are aligned more with users' physical affiliations in the peri–COVID-19 pandemic period as compared with that in the pre–COVID-19 pandemic period (indicating low membership entropy). These results highlight the distinct internet-based social behaviors between the d-r and non–d-r groups in the course of the COVID-19 pandemic period. In other words, the d-r members' communities maintained connections through shared physical affiliations during the COVID-19 pandemic period. These connections were different from the connection patterns of the non–d-r HBCU student group.

XSL•FO
RenderX

Figure 4. The depression-related (d-r) and non–depression related (non–d-r) groups' friendship networks, composed of students from historically black colleges and universities before and during the COVID-19 pandemic. (A) The panel demonstrates the structural change of the d-r group's friendship network in pre–COVID-19 (2019 dataset) and in peri–COVID-19 (2022 dataset) pandemic periods. (B) The panel displays the same structural changes of the non–d-r group's friendship network in pre–COVID-19 (2019 dataset) and peri–COVID-19 (2022 dataset) pandemic periods. (C to F). The distribution of Coleman homophily index (h_c) of randomized networks for the d-r group (C: 2019 and D: 2022) and for the non–d-r group (E: 2019 and F: 2022) are compared with the empirically observed h_c of the d-r and non–d-r groups (dashed black line). *P* values are evaluated from the h_c of random networks after the permutation tests. Highlighted bars in red represent the dataset having less the empirically observed h_c. (G to J) The distribution of <inline-graphic xlink:href="jmir_v26i1e55531_fig9.png" mimetype="image" xlink:type="simple"/> of randomized networks for d-r group (G:2019, H:2022) and for non-d-r group (I:2019, J:2022). Dashed black lines represent the empirically observed <inline-graphic xlink:href="jmir_v26i1e55531_fig9.png" mimetype="image" xlink:type="simple"/>, and red bars denote the dataset having less <inline-graphic xlink:href="jmir_v26i1e55531_fig9.png" mimetype="image" xlink:type="simple"/> than the empirically observed one.

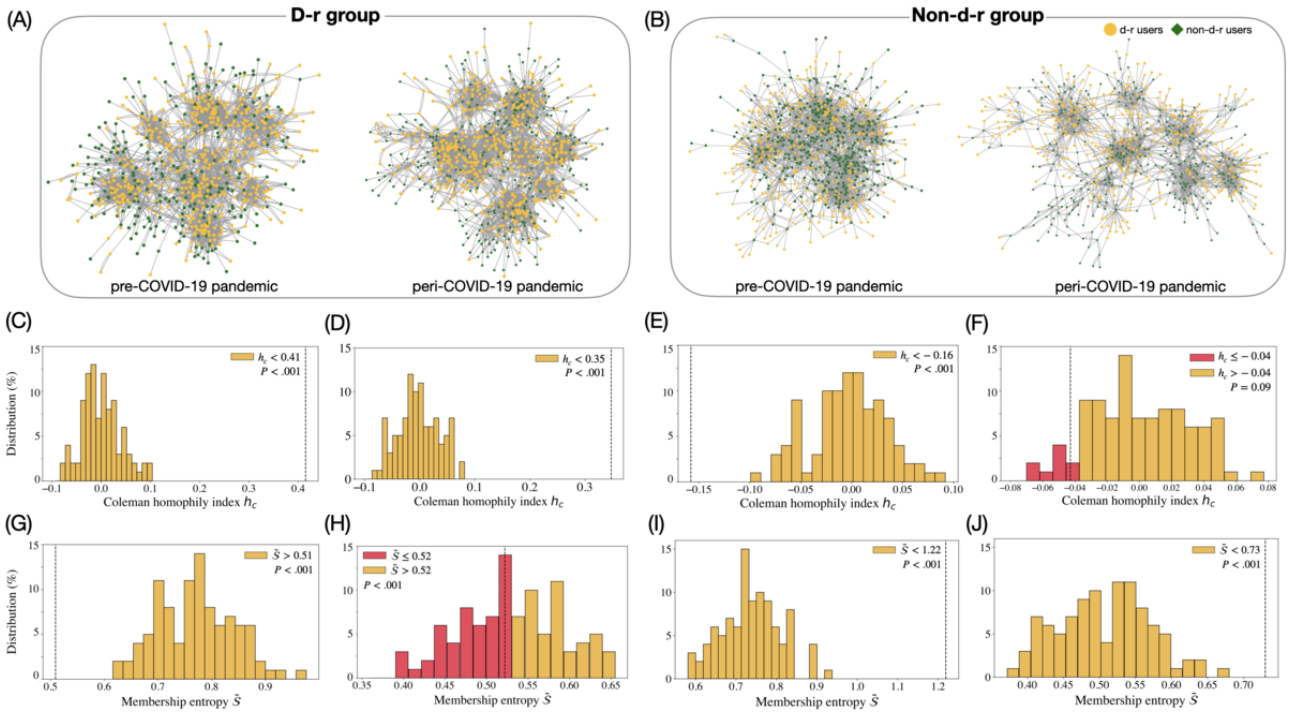

Discussion

Principal Findings

The COVID-19 pandemic profoundly impacted individuals, communities, economies, health care systems, and daily life worldwide [34-36]. The way individuals constructed social networks was no exception. Owing to the requirements of physical distancing, people use social media more than in the past to connect with others [12,13]. The increase in social media use on Twitter during the COVID-19 pandemic period was observed in both the d-r and non–d-r groups. However, the increase was statistically significant only in the d-r group. Moreover, users in the d-r group significantly increased mutual relationships (ie, reciprocity) and assortativity with other d-r users, whereas the non–d-r group's reciprocity and assortativity decreased. The community detection and membership entropy analyses revealed an interesting pattern of internet-based socialization among HBCU students during the COVID-19 pandemic. The results show that the students strengthened their internet-based relationships on the basis of their physical affiliations in the course of the COVID-19 pandemic period, indicating localization of such socialization. The increase in agreement between internet-based relationships and offline affiliation is mostly observed in the non–d-r group. However, the d-r group has kept similar levels of connectivity during the COVID-19 pandemic, connecting with more d-r group of users in their internet-based community than the expectation from randomized networks. This study offers unique insights into the collective social networking patterns of young Black communities in HBCUs, with focus on depression-related expressions. This approach differs from traditional social relationship analysis conducted on the general population without providing information on the population group's initial mental conditions [10,37].

Distinct Internet-Based Socializing Patterns of the D-R Group

In this study, the increase in the number of friends per user during the COVID-19 pandemic for both groups may indicate a coping strategy of people, resulting in more activities on social media [12,13,37]. The internet has been a major source of health information related to the etiologies, interventions, and treatments for disease and illness [38]. Internet-based forms of social support also gained prominence and experienced substantial growth during the COVID-19 pandemic. This is because they allowed interactions with a broader network of people beyond geographic limitations, including the use and expansion of formal internet-based support spaces [10]. The statistically significant expansion of friendship shown only in the d-r group indicates a desire for greater connectedness in this group. This indirectly reflects a disproportional impact of the

XSL•FO
RenderX

COVID-19 pandemic on people who experienced depression before the COVID-19 pandemic. This finding may inform higher education administrators and health organizations of the need for tailored mental wellness interventions for Black college students who mentioned depression on internet-based platforms.

In particular, the d-r group showed a statistically significant increase in reciprocity (mutual dependence) and assortativity (tendency to make friends in the same group) during the COVID-19 pandemic. In contrast, the non–d-r group experienced a decrease in both reciprocity (significant) and assortativity (not significant). The increased reciprocity and assortativity of the d-r group in an internet-based society might indicate that the tendency to associate with similar individuals became more prominent [10]; it might reflect that the d-r group pursued mutual exchange for information and social support, while the non–d-r group pursued more diverse information and social-support sources.

This distinct internet-based socializing pattern may bring several merits and benefits in different contexts, such as strong social bonds [27], fostered empathy [39], and efficient communication [40]. However, internet-based media can be a source of misinformation about health-related topics [40,41]. Considering the possibilities, we need to carefully understand the collective internet-based behaviors of a group of interest, especially of a minoritized group, which requires effort to understand their information consumption patterns [42]. Given the significant assortative tendency of the d-r group of users, identifying the d-r groups and designing a strategy to deliver accurate health information tailored for this group might be crucial to providing support during outbreaks such as the COVID-19 pandemic.

Strengthening Physical Relationships via an Internet-Based Platform

On an average, the social networks of 2 HBCU groups showed that user interactions became more localized during the COVID-19 pandemic compared with that before the pandemic. When we focus solely on HBCU students, the results show more modular interactions that are densely connected between users with the same affiliations during the COVID-19 pandemic. The size of the communities becomes small, leading to increased modularity. The decreased membership entropy in the course of the COVID-19 pandemic indicates that people tend to make internet-based connections based on physical affiliations. Another study found evidence of network homophily, revealing that most teachers interacted primarily with colleagues from similar disciplines and the same country [43].

Even though the densely connected internet-based communities aligned with one's physical affiliations can improve one's connectedness and emergency support, information among the communities could often revolve around campus updates (eg, remote learning), local COVID-19 pandemic status updates (eg, vaccine availability), and connections within their immediate community [11,40]. These localized and assortative connections might influence the likelihood of exposure to reputable sources of information [44]. Strategies that are culturally appropriate, community competent, and mindful of the nuances of different populations and communities play a crucial role in promoting health equity [45]. The dissemination of health-related information, especially of minoritized groups on internet-based platforms, should be further explored.

However, the increase in Coleman homophily and modularity in the filtered HBCU students' friendship network is mostly observed in non–d-r group of users than in the d-r group of users. This implies that the tendency to connect with local communities on the internet can be different by an individual's engagement in depression-related content on the internet. The d-r group of users remained in a similar level of relational alignment with physical affiliations in their internet-based communities during the COVID-19 pandemic; the non–d-r group of users reached a comparable level of Coleman homophily and modularity to their counterparts in the course of the COVID-19 pandemic period. This result suggests that HBCU students with depression-related expressions before the pandemic primarily sought social connections from their preexisting interactions, unlike non-d-r users, who tended to seek more localized connections. Therefore, a different strategy may be needed to deliver appropriate health information to those with depression-related expressions.

As these results are shown, understanding a group's social networks, risk experiences, and historical coping tendencies can help shape tailored strategies for health communication and treatment, as this information influences both internet-based and offline behavioral choices. In this study, insights from the internet-based socializing patterns of HBCU students can inform health care workers on therapeutic strategies tailored to patients within these communities. This might entail diversifying health-related information within their communities and nurturing sustainable localized communities with accurate health information and direct links to health care systems and support groups [42,43].

Limitations

A limitation of this study is that Twitter does not allow easy filtering of users on the basis of demographics; therefore, verification of whether users meet the inclusion criteria was subjective and based on the available content in the user's Twitter profile (eg, profile picture, account description, and tweets). This may exclude users who may have been eligible for inclusion. Because we could not reverify affiliation in the dataset published in 2022, the number of HBCU students in this study may be overestimated. In addition, because this study focuses on the social media activities and connections of the d-r group of users, the categorization of depression-related tweets cannot be used to confirm that the user has a clinical diagnosis of a depressive disorder. Therefore, the d-r group of users (or non–d-r group of users) could be overrepresented (or underrepresented) in the categorization.

Conclusions

The results demonstrate that the social media activity of HBCU students with depression-related tweets can be significantly different from those without depression-related tweets. The d-r group of users tend to have more mutual relationships and try to connect with other d-r group of users. The HBCU students also strengthened their connections among themselves in the course of the pandemic. It is unclear whether having more

XSL•FO
RenderX

friends in the same group can be positive (eg, social support) or negative (eg, harmful action) and how the information flow could differ from other demographic groups; thus, understanding these structural patterns related to the health-related information spreading dynamics could be an interesting future direction.

In addition, understanding minoritized groups' internet-based social behaviors, such as that of Black American college students during the COVID-19 pandemic, can help to understand their social support–seeking behaviors during stressful times and inform the development of interventions to support their mental health. Future studies should further explore the relationship between social media and mental health to determine whether there is a pattern of increased social media activity, mental health information seeking, and social-support seeking in different periods of time during the year for Black college students.

Acknowledgments

EL was supported by the National Research Foundation (NRF) of Korea grant funded by the Korean Ministry of Science and Information and Communication Technology (MSIT; grant number RS-2022-00165916617820643400 02), Learning & Academic research institution for Master's PhD students, Postdocs (LAMP) Program of the NRF of Korea grant funded by the Ministry of Education (grant number RS-2023-00301702), and the Pukyong National University Industry-university Cooperation Research Fund in 2023 (202312230001). TM was supported by funding from the National Library of Medicine (NLM) under award number R01LM013477 at the time the study was conducted. The authors would like to especially thank the Consumer Health Informatics Lab at Yale University for their support in publishing the results of this study.

Data Availability

The datasets collected during and analyzed during this study are not publicly available due to ethical considerations, since sharing the data may raise the possibility of revealing the identities of the Twitter users included in this study, which could put individuals at risk for harassment and discrimination.

Authors' Contributions

EL, HK, and TM conceptualized the study and conducted project administration. EL conducted formal analysis and investigation. TM and HK conducted data curation. All authors contributed to writing the original draft.

Conflicts of Interest

TM is a member of the Digital Wellbeing Advisory Board at Peer Health Exchange, the advisory board at RACE Space Inc, and a consultant for SecondMuse LLC. Peer Health Exchange, RACE Space, and SecondMuse were not involved in the writing of this paper or the decision to submit it for publication. TM is also an associate editor of the Journal of Medical Internet Research. All other authors declare no other conflicts of interest.

References

1. Huckins JF, daSilva AW, Wang W, Hedlund E, Rogers C, Nepal SK, et al. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 2020 Jun 17;22(6):e20185 [FREE Full text] [doi: 10.2196/20185] [Medline: 32519963]
2. Charles NE, Strong SJ, Burns LC, Bullerjahn MR, Serafine KM. Increased mood disorder symptoms, perceived stress, and alcohol use among college students during the COVID-19 pandemic. Psychiatry Res 2021 Feb;296:113706 [FREE Full text] [doi: 10.1016/j.psychres.2021.113706] [Medline: 33482422]
3. Alcendor DJ. Racial disparities-associated COVID-19 mortality among minority populations in the US. J Clin Med 2020 Jul 30;9(8):2442 [FREE Full text] [doi: 10.3390/jcm9082442] [Medline: 32751633]
4. Magesh S, John D, Li WT, Li Y, Mattingly-App A, Jain S, et al. Disparities in COVID-19 outcomes by race, ethnicity, and socioeconomic status: a systematic-review and meta-analysis. JAMA Netw Open 2021 Nov 01;4(11):e2134147 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.34147] [Medline: 34762110]
5. Cunningham TJ, Croft JB, Liu Y, Lu H, Eke PI, Giles WH. Vital signs: racial disparities in age-specific mortality among Blacks or African Americans - United States, 1999-2015. MMWR Morb Mortal Wkly Rep 2017 May 05;66(17):444-456 [FREE Full text] [doi: 10.15585/mmwr.mm6617e1] [Medline: 28472021]
6. Millett GA, Jones AT, Benkeser D, Baral S, Mercer L, Beyrer C, et al. Assessing differential impacts of COVID-19 on Black communities. Ann Epidemiol 2020 Jul;47:37-44 [FREE Full text] [doi: 10.1016/j.annepidem.2020.05.003] [Medline: 32419766]
7. Olaniyan M, Magnelia S, Coca V, Abeyta M, Vasquez MC, Harris FIII, et al. Two pandemics: racial disparities in basic needs insecurity among college students during the COVID-19 pandemic. The Hope Center. 2023 Oct 03. URL: https://hope.temple.edu/racial-disparities-basic-needs-insecurity-college-students-covid-19-pandemic#:~:text=These%20disparities%20were%20 further%20magnified,Black%20and%20Native%20American%20men [accessed 2024-09-25]

XSL•FO
RenderX

8. Mccall T, Kim H, Lee E, Lakdawala A, Bolton Iii CS. Content and social network analyses of depression-related tweets of African American college students. In: Proceedings of the 54th Hawaii International Conference on System Sciences. 2021 Presented at: HICSS 2021; January 5-8, 2021; Maui, HI. [doi: 10.24251/hicss.2021.318]
9. Higher Education Act of 1965. U.S. Government Publishing Office. URL: https://www.govinfo.gov/content/pkg/COMPS-765/pdf/COMPS-765.pdf [accessed 2024-09-25]
10. Long E, Patterson S, Maxwell K, Blake C, Bosó Pérez R, Lewis R, et al. COVID-19 pandemic and its impact on social relationships and health. J Epidemiol Community Health 2022 Feb;76(2):128-132 [FREE Full text] [doi: 10.1136/jech-2021-216690] [Medline: 34413184]
11. A look at volunteering during the response to COVID-19. Department for Digital, Culture, Media & Sport, United Kingdom Government. 2021 Jan 11. URL: https://www.gov.uk/government/publications/a-look-at-volunteering-during-the-response-to-covid-19/a-look-at-volunteering-during-the-response-to-covid-19 [accessed 2024-09-25]
12. Pandya A, Lodha P. Social connectedness, excessive screen time during COVID-19 and mental health: a review of current evidence. Front Hum Dyn 2021 Jul 22;3:684137. [doi: 10.3389/fhumd.2021.684137]
13. Juvonen J, Schacter HL, Lessard LM. Connecting electronically with friends to cope with isolation during COVID-19 pandemic. J Soc Pers Relatsh 2021 Mar 17;38(6):1782-1799. [doi: 10.1177/0265407521998459]
14. Stuart J, O'Donnell K, O'Donnell A, Scott R, Barber B. Online social connection as a buffer of health anxiety and isolation during COVID-19. Cyberpsychol Behav Soc Netw 2021 Aug;24(8):521-525. [doi: 10.1089/cyber.2020.0645] [Medline: 33601944]
15. Rauschenberg C, Schick A, Goetzl C, Roehr S, Riedel-Heller SG, Koppe G, et al. Social isolation, mental health, and use of digital interventions in youth during the COVID-19 pandemic: a nationally representative survey. Eur Psychiatry 2021 Mar 09;64(1):e20 [FREE Full text] [doi: 10.1192/j.eurpsy.2021.17] [Medline: 33686930]
16. Gabbiadini A, Baldissarri C, Durante F, Valtorta RR, De Rosa M, Gallucci M. Together apart: the mitigating role of digital communication technologies on negative affect during the COVID-19 outbreak in Italy. Front Psychol 2020 Oct 21;11:554678 [FREE Full text] [doi: 10.3389/fpsyg.2020.554678] [Medline: 33192807]
17. Li J. Digital technologies for mental health improvements in the COVID-19 pandemic: a scoping review. BMC Public Health 2023 Mar 01;23(1):413 [FREE Full text] [doi: 10.1186/s12889-023-15302-w] [Medline: 36859184]
18. Stress in America™ 2020: a national mental health crisis. American Psychological Association. 2020. URL: https://www.apa.org/news/press/releases/stress/2020/report-october [accessed 2024-09-25]
19. Provisional death counts for COVID-19. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm [accessed 2024-09-25]
20. Kemei J, Tulli M, Olanlesi-Aliu A, Tunde-Byass M, Salami B. Impact of the COVID-19 pandemic on Black communities in Canada. Int J Environ Res Public Health 2023 Jan 15;20(2):1580 [FREE Full text] [doi: 10.3390/ijerph20021580] [Medline: 36674335]
21. Yancy CW. COVID-19 and African Americans. JAMA 2020 May 19;323(19):1891-1892. [doi: 10.1001/jama.2020.6548] [Medline: 32293639]
22. Mercier CM, Abbott DM, Ternes MS. Coping matters: an examination of coping among Black Americans during COVID-19. Couns Psychol 2022 Mar 12;50(3):384-414. [doi: 10.1177/00110000211069598]
23. Molock SD, Parchem B. The impact of COVID-19 on college students from communities of color. J Am Coll Health 2022;70(8):2399-2405. [doi: 10.1080/07448481.2020.1865380] [Medline: 33502970]
24. Ramos MS, Corona R, Dempster KW, Morton SC, Everhart RS. The COVID-19 pandemic: asthma control, tobacco use, and mental health among African American and Latinx college students. J Asthma 2023 Mar;60(3):496-507. [doi: 10.1080/02770903.2022.2062673] [Medline: 35385676]
25. Huang HY, Li H, Hsu YC. Coping, COVID knowledge, communication, and HBCU student's emotional well-being: mediating role of perceived control and social connectedness. J Community Psychol 2022 Aug;50(6):2703-2725 [FREE Full text] [doi: 10.1002/jcop.22824] [Medline: 35187691]
26. Fetter A, Thompson M. Understanding the impacts of COVID-19 pandemic for undergraduate students attending an HBCU: insights from student voices. Center for Research on College-Workforce Transitions, Wisconsin Center for Education Research. 2020 Nov. URL: https://ccwt.wisc.edu/wp-content/uploads/2022/04/ccwt_report_Understanding-the-Impacts-of-COVID-19-Pandemic-for-Undergraduate-Students-attending-an-HBCU.pdf [accessed 2024-09-25]
27. Perry BL, Smith NC, Coleman ME, Pescosolido BA. Social networks, the COVID-19 pandemic, and emerging adults' mental health: resiliency through social bonding and cohesion. Am J Public Health 2024 Mar;114(S3):S258-S267. [doi: 10.2105/AJPH.2023.307426] [Medline: 37948054]
28. Moreno JL, Jennings HH. Statistics of social configurations. Sociometry 1938;1(3/4):342-374. [doi: 10.2307/2785588]
29. Newman M. Networks. Oxford, UK: Oxford University Press; Jul 26, 2018.
30. Moody J. Race, school integration, and friendship segregation in America. Am J Sociol 2001 Nov;107(3):679-716. [doi: 10.1086/338954]
31. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theor Exp 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/p10008]

XSL•FO
RenderX

32. Coleman J. Relational analysis: the study of social organizations with survey methods. Hum Org 1958;17(4):28-36. [doi: 10.17730/humo.17.4.q5604m676260q8n7]
33. Newman ME. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006 Jun 06;103(23):8577-8582 [FREE Full text] [doi: 10.1073/pnas.0601602103] [Medline: 16723398]
34. Global state of small business report. Facebook, OECD, World Bank. 2020. URL: https://about.fb.com/wp-content/uploads/2020/07/Global_State_of_Small_Business_Report.pdf [accessed 2024-09-25]
35. Reinhart CM. From health crisis to financial distress. Policy research working paper. World Bank. 2021 Apr. URL: https://openknowledge.worldbank.org/handle/10986/35411 [accessed 2024-09-25]
36. Haleem A, Javaid M, Vaishya R. Effects of COVID-19 pandemic in daily life. Curr Med Res Pract 2020 Mar;10(2):78-79 [FREE Full text] [doi: 10.1016/j.cmrp.2020.03.011] [Medline: 32292804]
37. Kovacs B, Caplan N, Grob S, King M. Social networks and loneliness during the COVID-19 pandemic. Socius 2021 Jan 19;7. [doi: 10.1177/2378023120985254]
38. National Research Council, Commission on Physical Sciences, Mathematics, and Applications, Computer Science and Telecommunications Board, Committee on Enhancing the Internet for Health Applications: Technical Requirements and Implementation Strategies. Networking Health: Prescriptions for the Internet. Washington, DC: National Academies Press; Jun 12, 2000.
39. Chen Y, Xu Y. Exploring the effect of social support and empathy on user engagement in online mental health communities. Int J Environ Res Public Health 2021 Jul 26;18(13):6855 [FREE Full text] [doi: 10.3390/ijerph18136855] [Medline: 34206719]
40. Baker H, Concannon S, So E. Information sharing practices during the COVID-19 pandemic: a case study about face masks. PLoS One 2022 May 5;17(5):e0268043 [FREE Full text] [doi: 10.1371/journal.pone.0268043] [Medline: 35511962]
41. Cuan-Baltazar JY, Muñoz-Perez MJ, Robledo-Vega C, Pérez-Zepeda MF, Soto-Vega E. Misinformation of COVID-19 on the internet: infodemiology study. JMIR Public Health Surveill 2020 May 09;6(2):e18444 [FREE Full text] [doi: 10.2196/18444] [Medline: 32250960]
42. Reddy B, Gupta A. Importance of effective communication during COVID-19 infodemic. J Family Med Prim Care 2020 Aug;9(8):3793-3796 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_719_20] [Medline: 33110769]
43. Alwafi E. Tracing changes in teachers' professional learning network on Twitter: comparison of teachers' social network structure and content of interaction before and during the COVID-19 pandemic. J Comput Assist Learn 2021 Dec;37(6):1653-1665 [FREE Full text] [doi: 10.1111/jcal.12607] [Medline: 34903905]
44. Hâncean MG, Lerner J, Perc M, Molina JL, Geantă M. Assortative mixing of opinions about COVID-19 vaccination in personal networks. Sci Rep 2024 Mar 09;14(1):3385 [FREE Full text] [doi: 10.1038/s41598-024-53825-3] [Medline: 38336858]
45. Webb Hooper M, Nápoles AM, Pérez-Stable EJ. COVID-19 and racial/ethnic disparities. JAMA 2020 Jul 23;323(24):2466-2467 [FREE Full text] [doi: 10.1001/jama.2020.8598] [Medline: 32391864]

Abbreviations

d-r: depression-related
HBCU: historically black colleges and universities
LCC: largest connected component
non–d-r: non–depression-related

Edited by T de Azevedo Cardoso; submitted 20.12.23; peer-reviewed by MDP Marques León, M Martínez-Fernández, R Leirós-Rodríguez; comments to author 11.04.24; revised version received 06.06.24; accepted 12.08.24; published 28.10.24.

Please cite as:
Lee E, Kim H, Esener Y, McCall T
Internet-Based Social Connections of Black American College Students in Pre–COVID-19 and Peri–COVID-19 Pandemic Periods: Network Analysis
J Med Internet Res 2024;26:e55531
URL: https://www.jmir.org/2024/1/e55531
doi:10.2196/55531
PMID:

©Eun Lee, Heejun Kim, Yildiz Esener, Terika McCall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Electronic Medical Record–Based Prognostic Model for Inpatient Falls: Development and Internal-External Cross-Validation

Rex Parsons[1], PhD; Robin Blythe[1], PhD; Susanna Cramb[1,2], PhD; Ahmad Abdel-Hafez[3], PhD; Steven McPhail[1,4], PhD

[1]Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Queensland University of Technology, Kelvin Grove, Australia

[2]Jamieson Trauma Institute, Royal Brisbane and Women's Hospital, Metro North Health, Herston, Australia

[3]College of Computing and Information Technology, University of Doha for Science and Technology, Doha, Qatar

[4]Clinical Informatics Directorate, Metro South Health, Woolloongabba, Australia

Corresponding Author:
Rex Parsons, PhD
Australian Centre for Health Services Innovation and Centre for Healthcare Transformation
School of Public Health and Social Work
Queensland University of Technology
60 Musk Ave
Kelvin Grove, 4059
Australia
Phone: 61 31380905
Email: rex.parsons@hdr.qut.edu.au

Abstract

Background: Effective fall prevention interventions in hospitals require appropriate allocation of resources early in admission. To address this, fall risk prediction tools and models have been developed with the aim to provide fall prevention strategies to patients at high risk. However, fall risk assessment tools have typically been inaccurate for prediction, ineffective in prevention, and time-consuming to complete. Accurate, dynamic, individualized estimates of fall risk for admitted patients using routinely recorded data may assist in prioritizing fall prevention efforts.

Objective: The objective of this study was to develop and validate an accurate and dynamic prognostic model for inpatient falls among a cohort of patients using routinely recorded electronic medical record data.

Methods: We used routinely recorded data from 5 Australian hospitals to develop and internally-externally validate a prediction model for inpatient falls using a Cox proportional hazards model with time-varying covariates. The study cohort included patients admitted during 2018-2021 to any ward, with no age restriction. Predictors used in the model included admission-related administrative data, length of stay, and number of previous falls during the admission (updated every 12 hours up to 14 days after admission). Model calibration was assessed using Poisson regression and discrimination using the area under the time-dependent receiver operating characteristic curve.

Results: There were 1,107,556 inpatient admissions, 6004 falls, and 5341 unique fallers. The area under the time-dependent receiver operating characteristic curve was 0.899 (95% CI 0.88-0.91) at 24 hours after admission and declined throughout admission (eg, 0.765, 95% CI 0.75-0.78 on the seventh day after admission). Site-dependent overestimation and underestimation of risk was observed on the calibration plots.

Conclusions: Using a large dataset from multiple hospitals and robust methods to model development and validation, we developed a prognostic model for inpatient falls. It had high discrimination, suggesting the model has the potential for operationalization in clinical decision support for prioritizing inpatients for fall prevention. Performance was site dependent, and model recalibration may lead to improved performance.

(J Med Internet Res 2024;26:e59634) doi:10.2196/59634

XSL•FO
RenderX

KEYWORDS

clinical prediction model; falls; patient safety; prognostic; electronic medical record; EMR; intervention; hospital; risk assessment; clinical decision; support system; in-hospital fall; survival model; inpatient falls

Introduction

Falls in hospitals cause serious injuries and deaths [1]. Inpatient falls are difficult to predict [2] or prevent [3], although some inpatient fall prevention strategies that require efforts early in patients' admissions have been effective [4-6]. The use of health professional–completed fall risk assessment and screening tools to identify patients at high risk of falls has been a popular approach to stratify the risk of falling for inpatients, as has the development of clinical prediction models incorporating these assessments [7,8]. A clinical prediction model is a model that estimates an individual's probability of a current health condition (diagnostic) or one that may occur in the future (prognostic). These models are usually estimating a probability, or risk, for the given patient and health outcome. As a result, a prognostic model is a form of a clinical prediction model (or risk prediction model) that would be relevant to estimating the risk of falling for an inpatient in a manner that could guide the use of a fall prevention strategy by clinical teams. Recent efforts include work to use prediction tools to guide the selection of specific prevention strategies to prevent inpatient falls [9]. However, the discriminatory performance of inpatient falls risk assessment tools and prediction models is often poor, which may have contributed to why many risk model–guided treatments have failed to demonstrate effectiveness [9]. This study used the Northern Hospital Modified St Thomas' Risk Assessment Tool, which was developed for acute inpatients and had a sensitivity and specificity of 0.65 and 0.79, respectively [10]. A recent multihospital trial demonstrated the noninferiority and potential superiority of divestment away from conventional in-hospital fall risk assessment toward a simple clinical decision support that prompted consideration of potential patient interventions [11]. The tools used in both of these studies required clinical teams to complete questionnaires and manually score patient risk based on tallying up the identified risk factors. Conventional fall risk assessments may therefore be introducing additional data collection burdens to health professionals without leading to better patient outcomes.

Our previous review that investigated the methods and data used to develop prognostic models for inpatient falls found that reporting quality was typically poor, with a largely homogeneous approach to model development [8]. No studies applied time-to-event methods, also known as survival modeling approaches, to predict falls. Instead, studies typically used data available at admission to predict a binary end point of fall at any time during the patient admission over unknown and heterogenous lengths of time [8]. This approach has the potential to be misleading if used in a clinical decision support system, as the estimated risk (probability of a fall) may be primarily driven by exposure time through long lengths of stay rather than the proximal risk of experiencing a fall event. If exposure time is not appropriately accounted for during model development, fall prevention strategies may be disproportionately assigned to those who are likely to stay longer at the expense of those who are at an elevated relative risk despite shorter anticipated lengths of stay.

Many existing fall prediction models rely upon a wide variety of data sources, often requiring manual collection and data entry, which limits the potential for integration into real-time risk prediction and clinical decision support [12]. Development of a robust, interpretable prediction model that can estimate the risk of inpatient falls using solely routinely recorded electronic medical record (EMR) data to guide the assignment of effective inpatient fall prevention strategies remains a priority. An important anticipated advantage of relying on routinely recorded EMR data is that risk estimates can be generated without the requirement for additional fall risk screening tasks to be completed by clinical staff, likely being resource-saving when compared with widely used questionnaire-based falls risk screening. It is also particularly important for the first days of an admission period when in-hospital fall prevention strategies are typically initiated and when most falls occur, as risk estimates can be generated immediately as data are available [13].

Primarily, our previous review highlighted that relying on data collected as part of routine clinical care and stored within EMR systems and using time-to-event modeling approaches is underexplored and may be particularly useful and relevant for the given clinical context of falls risk screening, assessment, and prevention [8]. Our primary interest was to develop a prognostic model for inpatient falls using a survival modeling approach and EMR data from 5 hospitals that performed well early in admission, with the potential for implementation in a clinical decision support tool to help augment clinician decision-making regarding the initiation of effective inpatient fall prevention interventions.

This study is reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement and checklist [14].

Methods

Overview

Model development choices were informed by a previous review [8] and best-practice recommendations [15,16]. The model was developed using Cox proportional hazards regression and included time-varying covariates and predictor variables available within the EMR system. Performance was assessed in terms of discrimination and calibration using internal-external cross-validation, and the final model was fit using the full dataset.

Setting and Data Cleaning

After consultation with fall prevention experts and hospital geriatricians, we censored observations after 14 days of admission. After 14 days, hospital falls occurred at a lower rate and may be associated with a different clinical context for

XSL•FO
RenderX

long-stay patients that may benefit from a more targeted modeling strategy. To ensure the data focused on the period in which fall interventions are initiated and most falls occur, we used EMR data for the first 14 days of admission from patients admitted to 5 Australian hospitals between 2018 and 2021 to develop our model. Our intention was for study findings to be representative of inpatient hospital wards broadly. There were no exclusion criteria applied to eliminate particular subgroups of hospital patients; patients from any ward and of any age were included for model development and validation. In each of our included hospitals, inpatient falls were recorded in the EMR using RiskMan (RiskMan International Pty Ltd), a proprietary patient safety system for tracking and managing adverse events [17]. RiskMan reports were linked to each inpatient admission record using a unique identifier, including date and time stamps. When the time of fall was not recorded, it was assumed to be at midnight on the morning of the date specified to ensure no predictions used data recorded after the fall.

To maximize the potential for generalizability and ease of potential transferability to external hospital systems while eliminating the need for clinical staff time for additional fall risk assessment and data entry, predictor variables used in the model were primarily derived from administrative information routinely recorded for all hospital admissions. This included the source of the admission (eg, whether they came from residential aged care or through the emergency department), type of admitting medical service (eg, neurosurgery or orthopedics), sex (coded as the binary variable “Female”), age in years, as well as years since 2018 to incorporate potential adjustment for longer-term trends in fall risk over the study period. We also included 2 time-varying covariates in the model—the number of previous inpatient falls during the current admission up until the time of prediction and the time since admission, in hours. The time since admission variable was updated every 12 hours throughout the admission.

Because coding for admission source and medical service categories with similar or overlapping meanings varied across hospital facilities, we aggregated these into overarching categories that could be readily applied to data from all hospitals based on clinical groupings, shown in Table 1. Using these overarching clinical groupings also mitigated the risk of categorical variables for admission source or medical service being exclusive to the training or validation datasets. The “Other” categories for these 2 variables were used to include missing data as well. There were no missing data for other covariates included in the model. Restricted cubic splines with 4 knots were used for patient age and the years since 2018, and a restricted cubic spline with 3 knots was used for the hours since admission variable. A spline function with n knots is made up of n+1 piecewise polynomial functions where the knot locations describe the place (value of the variable to which spline is being applied) where these 2 sequential polynomial functions meet. For example, in the full model, the restricted cubic spline applied to patient age has 4 knots, meaning that there are 5 polynomial functions that meet at the 4 locations (values of patient age; described in Multimedia Appendix 1). Further information on use and interpretation of spline terms are illustrated in Regression Modeling Strategies by Harrell [18]. The restricted cubic splines used in this study have an advantage for use in prediction models in that they are constrained to be linear in their 2 tails (above the last knot and before the first knot). This is particularly important considering the use of “years since 2018” as a predictor and the possible use of the model in years that exceed the range of the data used to fit it.

XSL•FO
RenderX

Table 1. Demographics and summary of key predictor variables used in the model.

Variable and measure	Hospital					All
	1	2	3	4	5	
Inpatient admissions, n						
Total	462,891	359,075	150,458	124,149	10,983	1,107,556
Days	1,198,736	612,658	256,887	273,179	31,882	2,373,343
Days (truncated at 14 days)	794,398	485,189	197,951	228,770	23,825	1,730,133
Female, n (%)[a]	191,732 (41.42)	205,809 (57.32)	89,113 (59.23)	65,440 (53.71)	7364 (67.05)	559,458 (50.51)
Falls (truncated at 14 days)						
Total, n	3027	1377	638	828	134	6004
Fallers, n (%)[a]	2678 (0.58)	1248 (0.35)	562 (0.37)	731 (0.59)	122 (1.11)	5341 (0.48)
Age (year), median (IQR)	63 (48-74)	42 (26-63)	52 (31-72)	61 (43-77)	38 (26-71)	56 (35-72)
Length of stay (days), median (IQR)	0.3 (0.1-1.3)	0.3 (0.2-1.2)	0.3 (0.2-1.2)	0.4 (0.2-2.1)	1.0 (0.2-2.3)	0.3 (0.2-1.4)
Medical service, n						
General medicine	22,837	26,694	22,678	22,869	5042	100,120
Orthopedics	15,833	10,694	<100[b]	8138	<100	34,665
Psychiatry	16,932	12,209	2729	<100	<100	31,870
Rehabilitation	7840	905	<100	8912	<100	17,657
Neurosurgery	5449	<100	<100	<100	<100	5449
Palliative medicine	<100[a]	1553	613	1573	<100	3739
Geriatrics	2145	<100	<100	<100	<100	2145
Psychogeriatric	628	<100	<100	<100	<100	628
Spinal	378	<100	<100	<100	<100	378
Other	390,849	307,020	124,438	82,657	5941	910,905
Admission source, n						
Emergency department—this hospital	128,782	182,224	77,243	84,570	3990	476,809
Outpatient department—this hospital	208,239	91,621	38,291	29,378	4160	371,689
Routine readmission no referral required	84,698	63,475	25,576	7002	327	181,078
Admitted patient transferred from other hospital	20,187	3906	1186	2084	744	28,107
Private medical practitioner (not psychiatry)	11,119	136	340	616	<100	12,221
Residential aged care service	2140	<100	<100	<100	<100	2166
Community service	257	1237	237	363	<100	2130
Other	7469	16,471	7582	118	1716	33,356

[a]Percentages were calculated with the total number of inpatient admissions as the denominator.

[b]Cells with small values have been replaced with "<100" to address privacy concerns.

Evaluation

Following advice from Steyerberg and Harrell [16], we developed the model using the full dataset but evaluated model performance using internal-external cross-validation. Internal-external cross-validation is a process in which the folds used for cross-validation correspond to individual study sites, in this case, hospitals. This method is useful for reporting prediction performance, as allocation to validation folds is nonrandom, is reproducible, and provides better estimates of model generalizability [19]. It also means that every validation fold is an entirely separate hospital from those used for fitting the model, meaning that it provides a better estimate of generalizable performance to new settings than internal validation (where data from a single site are divided into a development dataset and an evaluation dataset). Table 2 shows a summary of how the internal-external cross-validation folds are constructed from the dataset comprising 5 hospitals. In this

XSL•FO
RenderX

process, 5 models are fit—one for each hospital being used as a validation dataset, with the remaining 4 hospitals being used to fit the model. Table 2 shows a summary of the data used to fit each of the 5 models (fold 1 to 5) and the final model (fit using the data from all hospitals). Using the entire dataset for model development but internal-external cross-validation for reporting ensures that the model does not suffer from avoidable losses in performance by using a subset of the data and potentially omitting useful observations [16].

Model performance was evaluated in terms of discrimination and calibration. The area under the time-dependent receiver operating characteristic (ROC) curve was calculated for each fold and a "combined" curve was derived to show the moving average across folds. The time-dependent ROC curve was evaluated using prediction times of every 24 hours from 1 day post admission to 1 week after admission. The *survivalROC* R package [20] was used to estimate the time-dependent ROC curves. To estimate the parameter uncertainty of our performance metrics, we applied the approach described by LeDell and colleagues [21] by modifying the *cvAUC* R package [22] to combine area under the curve (AUC) values from multiple cross-validation folds and construct 95% CIs. For visualization purposes, the curves labeled "combined" on the ROC plots in this study were constructed using vertical averaging [23].

To evaluate model calibration, we followed advice from McLernon and colleagues [24] to obtain moderate calibration curves over the full-time range. As described in the supplement of their study, we used a Poisson model with the outcome regressed against the linear predictors, fit with a restricted cubic spline, and the log of the predicted cumulative hazard as an offset. Smoothed curves were applied to investigate whether observed fall event rates were equal to the predicted risk among patients with the same predicted risk. We created curves separately for each validation fold.

Table 2. Sample size calculations and summary.

Model[a]	Inpatient admissions, n	Patient days (truncated at 14 days), n	Falls, n	Model parameters, n	Events per parameter, n
Fold: 1	644,665	935,735	2977	22	135
Fold: 2	748,481	1,244,944	4627	26	178
Fold: 3	957,098	1,532,182	5366	26	206
Fold: 4	983,407	1,501,363	5176	26	199
Fold: 5	1,096,573	1,706,308	5870	26	226
Final	1,107,556	1,730,133	6004	26	231

[a]Model represents the cross-validation fold models and the final model fit with all patient data. The fold models are those that fit during internal-external cross-validation and incorporate all patient data except for the associated hospital of the same number. For example, the "Fold: 1" model was fit using patient data from hospitals 2 to 5, with hospital 1 being the validation set.

Sample Size

We considered a model with a possible 30 parameters, a fall rate of 0.0054, a conservative estimate of 0.143 Cox-Snell R^2 for sample size calculation, a selected prediction time of 24 hours, and mean follow-up of 37.5 hours (the mean length of stay within the truncated dataset). The Cox-Snell R^2 was obtained by using an estimated C-index of 0.7, the approximate performance of recently published inpatient fall prediction models [8], and transforming it into an estimate for the Cox-Snell R^2 using methods described in related works by Jinks et al [25] and Royston and Sauerbrei [26]. Using the *pmsampsize* R package [27], we estimated that the minimum sample size for fitting this Cox model would be 1619 patients, with 330 fall events, and 11.75 events per parameter. During our internal-external cross-validation, the least powered model that was fit included 2977 falls and 135 events per parameter (22 parameters; a summary of model parameters, patients, and events used for each model fit during cross-validation is illustrated in Multimedia Appendix 2). The final model was fit with 6004 falls and 231 events per parameter (26 parameters).

Ethical Considerations

Approval was granted by the Metro South Human Research Ethics Committee (HREC/2020/QMS/64807). The requirement for informed consent was waived.

To protect the privacy and confidentiality of candidates, the data used in the study cannot be publicly shared. However, analysis code, including software and package versions, is available on GitHub [28].

Results

Patient characteristics and frequencies for admission source and medical service are described for each included hospital in Table 1. The admitted patient records over 4 years (2018 to 2021) used in this study included 1,107,556 inpatient admissions. There was approximately even representation of male and female patients (559,458/1,107,556, 50.51% female) and the median (IQR) age was 56 (35-72) years. There were 6004 falls and 5241 individual fallers.

There were expected and notable differences in patient age and sex characteristics between hospitals, consistent with the clinical services offered at each facility and populations served. In this context, hospitals differed by size and level of intensive care

XSL•FO
RenderX

unit services, specialized geriatric care units, rural or inner-city location, and number of beds (Table 1) [29].

The final model fit to the entire dataset is provided in Table 3 with associated locations for spline terms in Multimedia Appendix 1. This study was conducted with the primary aim of developing a prognostic model for inpatient falls, not to identify prognostic factors. Interpreting estimates from Table 3 allows us to identify which factors were associated with higher risk estimates but should not be interpreted as causally related to falls. Male patients, patients receiving geriatric or rehabilitation care, and those with previous falls were associated with higher risk estimates.

The AUC was 0.899 (95% CI 0.88-0.91) when a prediction time of 24 hours after admission was used. Discriminatory performance reduced as the admission progressed (Figure 1). There were substantial differences in performance between hospitals, particularly as time since admission increased. Model calibration plots indicated that overestimation and underestimation of risk occurred, with the degree of deviation from optimal calibration dependent on the cross-validation fold (Figure 2 [24]).

Table 3. Summary of the final model.

Parameter	Estimate (95% CI)	Hazard ratio (95% CI)
Female	–0.2189 (–0.2705 to –0.1673)	0.8034 (0.7630-0.8460)
Age (years)[a]	0.0221 (0.0153 to 0.0290)	1.0224 (1.0154-1.0294)
Age (years)[b]	–0.0051 (–0.0169 to 0.0068)	0.9949 (0.9832-1.0068)
Age (years)[c]	–0.0297 (–0.0802 to 0.0209)	0.9708 (0.9229-1.0211)
Time since 2018 (years)[a]	0.0228 (–0.0868 to 0.1324)	1.0231 (0.9169-1.1416)
Time since 2018 (years)[b]	–0.0370 (–0.3597 to 0.2857)	0.9637 (0.6979-1.3307)
Time since 2018 (years)[c]	0.0451 (–0.9285 to 1.0186)	1.0461 (0.3952-2.7692)
Admission source (community service)	0.0054 (–0.2627 to 0.2736)	1.0055 (0.7690-1.3147)
Admission source (emergency department - this hospital)	–0.0003 (–0.0828 to 0.0822)	0.9997 (0.9205-1.0857)
Admission source (other)	–0.5548 (–0.8137 to –0.2959)	0.5742 (0.4432-0.7439)
Admission source (outpatient department - this hospital)	–0.3924 (–0.5040 to –0.2807)	0.6755 (0.6041-0.7552)
Admission source (private medical practitioner [not psychiatry])	–0.7988 (–1.2337 to –0.3639)	0.4499 (0.2912-0.6950)
Admission source (residential aged care service)	–0.0465 (–0.3199 to 0.2270)	0.9546 (0.7262-1.2549)
Admission source (routine readmission no referral required)	–0.8621 (–1.1727 to –0.5516)	0.4223 (0.3095-0.5760)
Medical service (geriatrics)	0.2673 (0.1288 to 0.4059)	1.3065 (1.1375-1.5006)
Medical service (neurosurgery)	–0.0202 (–0.1929 to 0.1525)	0.9800 (0.8245-1.1648)
Medical service (orthopedics)	–0.3137 (–0.4298 to –0.1976)	0.7307 (0.6506-0.8207)
Medical service (other)	–0.5710 (–0.6393 to –0.5027)	0.5650 (0.5277-0.6049)
Medical service (palliative medicine)	0.2389 (0.0888 to 0.3889)	1.2698 (1.0928-1.4754)
Medical service (psychiatry)	–0.3366 (–0.4531 to –0.2201)	0.7142 (0.6356-0.8025)
Medical service (psychogeriatric)	–0.0213 (–0.3123 to 0.2697)	0.9789 (0.7317-1.3095)
Medical service (rehabilitation)	0.0417 (–0.0969 to 0.1802)	1.0425 (0.9076-1.1975)
Medical service (spinal)	–0.5627 (–1.0919 to –0.0335)	0.5697 (0.3356-0.9671)
Time since admission (hours)[a]	0.3008 (0.2454 to 0.3562)	1.3510 (1.2781-1.4279)
Time since admission (hours)[b]	–0.3621 (–0.5374 to –0.1867)	0.6962 (0.5842-0.8297)
Previous falls (n)	0.9053 (0.8640 to 0.9467)	2.4728 (2.3727-2.5771)

[a]Level 1 for the spline terms (refer to Multimedia Appendix 1 for knot locations for each term).

[b]Level 2 for the spline terms.

[c]Level 3 for the spline terms.

XSL•FO
RenderX

Figure 1. Discrimination performance. (A) Time series of discrimination performance for prediction times between 1 day and 1 week. The colored lines represent performance at difference centers when used as the holdout set during cross-validation, and the dashed black line represents the aggregate measure of discrimination with the gray region as the 95% CI for this estimate. (B-H) Time-dependent ROC curves for predicted times from 1 day to 1 week, each including a curve for each of the 5 models fit during internal-external cross-validation as well as a "combined" curve derived using vertical-averaging. AUC: area under the curve; ROC: receiver operating characteristic.

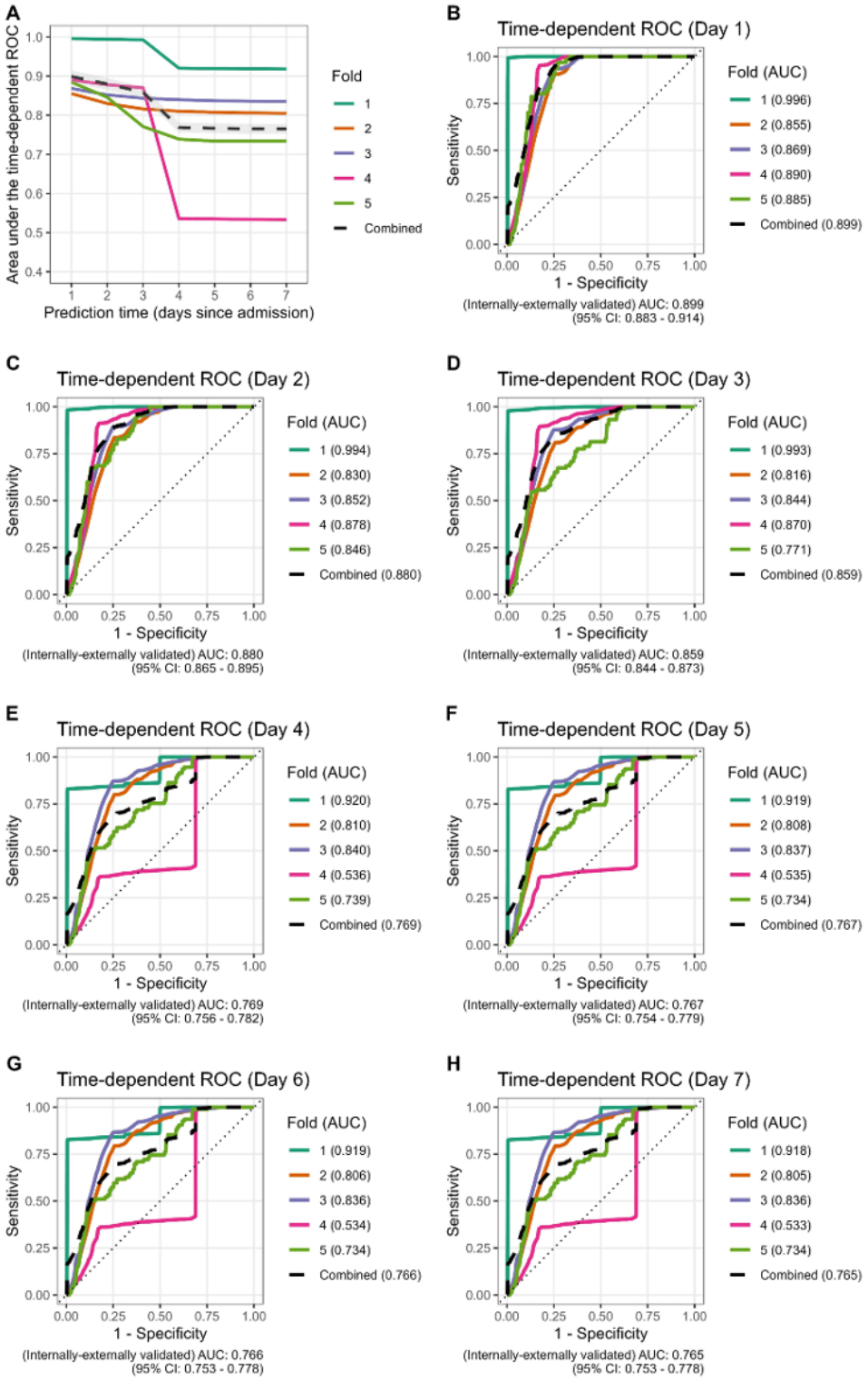

XSL•FO
RenderX

Figure 2. Calibration performance for each cross-validation fold. The top row (A and B) shows the full plots, and the bottom row (C and D) shows the same curves but with a focused view to better visualize the largest portion of predicted values (predicted cumulative hazards between 0 and 0.015). (A and C) Empirical cumulated density functions of the cumulative hazard from the Cox regression model. (B and D) Moderate calibration assessment as described by McLernon et al.

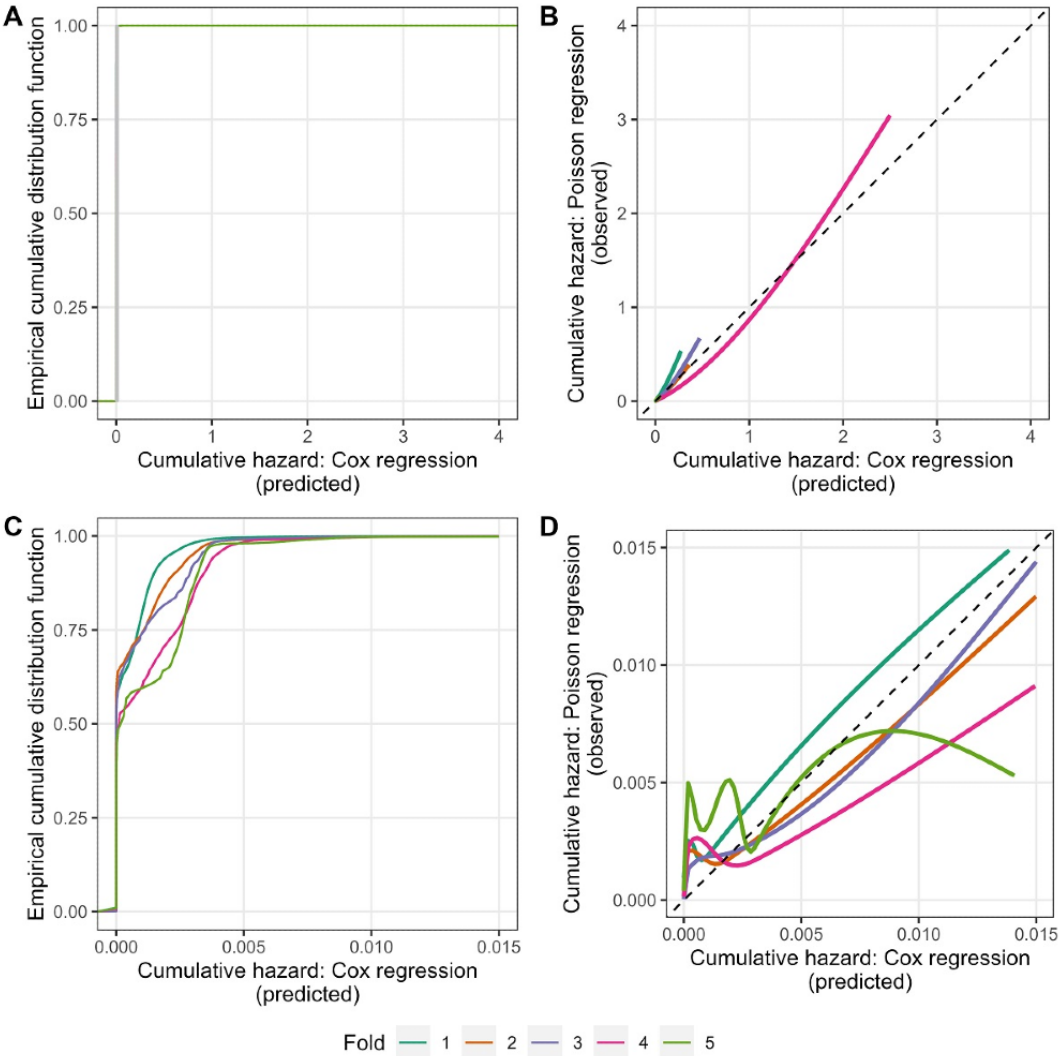

Discussion

Principal Findings

Using only routinely recorded information in EMR systems from 5 hospitals of varying capacities and capabilities, we successfully developed a novel prognostic model for inpatient falls using a time-to-event modeling approach. Our approach has the potential to add value to inpatient falls prevention efforts by updating prognostic risk predictions with new information throughout a patient's admission as it becomes available without requiring clinical staff to collect additional fall-risk screening or assessment data and manually estimate risks. It was encouraging to observe useful predictive performance that was not inferior to more labor- and data-intensive fall risk prediction approaches that featured in our previous review [8]. Studies reporting AUC within our previous review that included external validation and used data from adult inpatients typically reported AUC in the range of 0.7 to 0.85 [30-32]. By developing our model using only data from EMR, this labor-efficient approach to estimation of inpatient fall risk has the potential to be implemented in clinical decision support systems that aid in appropriate allocation of effective inpatient fall prevention interventions without dependency on physical functional performance assessment [33,34].

Overall, our model demonstrates similar or greater discriminatory performance than many other published models for inpatient falls, particularly those that have been externally validated [2,8]. Calibration was site dependent but generally poor. The variable model performance, in terms of both discrimination and calibration, and its association with hospitals may be explained by differences in patient cohorts, fall rates, and the nature of care provided in each hospital.

The poor calibration performance may suggest that the absolute model predicted risk should be interpreted with caution. However, predictions are likely to remain useful for if the models were implemented in a manner that prioritized higher-risk patients (based on rank of estimated risk) toward a fall prevention intervention rather than applying a selected cutoff point above which the patient receives the intervention.

This approach would better take advantage of the high discriminatory performance without relying heavily on the need

XSL•FO
RenderX

for the model to be well calibrated and for the selected cutoff point to be used on potentially biased estimates of risk. The use of a cutoff point may not work effectively due to the observed miscalibration based on local clinical contexts, but model recalibration by hospital could be beneficial [35].

We observed that there was a high level of between-hospital variability in ROC curves, particularly for prediction times at day 4 and later. We suspect that this is due to there being relatively few patients and fallers who are still admitted at that time. For example, the particularly unstable ROC curve for hospital 4 observed in Figure 1 had a median length of stay of 0.4 (IQR 0.2-2.1) days, with likely relatively few admitted patients, and very few admitted patients with falls, remaining after day 3. We suspect that the ROC curve for this hospital for these later prediction times is likely influenced by relatively few data.

While some aspects of our model are directly comparable to existing inpatient fall prediction models, several are unique. We have not identified any time-to-event prediction models in the literature that evaluate performance in terms of discrimination and calibration [8]. As binary prediction models typically identify fallers as patients who fell over the remaining time in the admission without time-varying covariates, perhaps the closest comparison from our study would be our model's discriminatory performance at day 1 before time-varying covariates had influenced risk predictions. For this scenario, the discrimination of our model demonstrates superior performance during internal-external cross-validation. In addition, the advantages of a time-to-event approach include the ability to update predictions as the admission progresses and new data become available, and better incorporation of current length of stay into estimated risk.

We used internal-external cross-validation because this is the most appropriate way to evaluate a clinical prediction model when (1) data are available from multiple sites and (2) there are enough data at each site to ensure that models developed from each cross-validation fold are adequately powered [16]. Many clinical prediction model development studies may only rely on data from 1 to 3 sites. In our previous review, we found that only 2 of 51 studies used data from more than 3 hospitals to develop their model (43 of which only used data from a single hospital) [8]. When only a few sites, it may be impractical to conduct internal-external cross-validation as there would only be 2- or 3-folds, and potentially have too small a portion of the entire dataset for those models to be adequately powered. Fortunately, in our study, we had data from 5 sites and, of the 5 models fit during cross-validation, the lowest number of events per parameter was 135, which was much higher than the minimum number estimated by our sample size calculations (11.75).

For evaluation of the discrimination, we used the area under the time-dependent ROC curve. To create a single summary measure, we used vertical averaging of the 5-folds [23]. A potential limitation of this is that hospitals are equally influential on the "combined" curve and the final estimate of discrimination, there is no additional weight given to hospitals with more data than others. However, in this study, the models with the lowest performance were typically those evaluated on hospitals with fewer admitted patients and events, rather than the largest hospital ("Fold 1" curve in Figure 1), which was consistently the best performing.

Most existing approaches to predicting inpatient falls use data collected upon admission to estimate risk for the duration of the entire admission. This may lead to confusing interpretations; if clinicians interpret predicted risks as a function of admission time, this can be misleading. For example, a user may interpret the patient's daily fall risk as the total estimated risk divided by the expected length of stay in days, which inappropriately assumes falls risk is evenly distributed over the entire admission. A patient may have a relatively high risk of falling on any given day, but if they are admitted for a condition that is associated with shorter lengths of stay, then a binary prediction model may predict a deceptively low risk. Time-to-event models avoid this potentially erroneous conclusion by accounting for length of stay inherently. We anticipate that a model of this level of discriminatory performance may be useful when allocating patients for fall prevention intervention, but this would need to be evaluated prospectively or possibly within a simulation study [36].

The use of machine learning methods for clinical prediction models has been increasing in popularity, including neural networks or tree-based ensemble models such as random forests or extreme gradient boosting [37,38]. There are variants of these approaches capable of handling time-to-event models but none, to our knowledge, have been applied to develop a prognostic model for falls. For example, random survival forests have been used to predict colorectal cancer prognosis [39], and "pycox" [40], an implementation of a Cox model for a neural network using *PyTorch*, has been used to predict in-hospital events including mortality and discharge [41]. There has also been a recent increase in popularity of sequence models, including attention models and recurrent neural networks, which can incorporate a sequence of input data of variable lengths and of irregular intervals [42]. We expect that combining these approaches with a Cox layer, as is used in "pycox," may lead to the benefits of both being able to use time-varying covariates and being able to appropriately predict falls as a time-to-event outcome. A recent study has implemented a similar approach to fit a time-dependent Cox survival neural network to generate dynamic predictions, with example datasets including age-related eye disease and time-to-liver transplant [43]. The machine learning methods included in our recent review [8] were limited to those predicting falls as either occurring any time in the admission or within a fixed time horizon [44,45], but none used a time-to-event outcome. Consequently, it is difficult to know whether falls may be better estimated by the additional complexity and ability to estimate interaction effects within these methods.

We expect that by incorporating more time-varying predictors in a Cox model to predict falls may improve performance but may require a larger dataset with more falls to avoid overfitting. Another limitation of using this model is that it requires computation to estimate risk and cannot be easily implemented without an EMR, as other commonly used analog falls risk assessment tools can [33,46]. Although a reduction in data

XSL•FO
RenderX

collection burden for health professionals is a strength of our model, it is a potential barrier to implementation in settings without the required levels of digital maturity.

Conclusions

In this study, we have presented a prognostic model using time-to-event modeling. We used internal-external cross-validation with data from 5 hospitals with notable heterogeneity to estimate its generalizable performance and present the final model that incorporated data from all 1,107,556 inpatient admissions. We anticipate that the reported model estimates may be adjusted to improve calibration performance at each hospital, and that this model has potential use for improving clinical care in the context of prompting the initiation of fall prevention strategies. However, evaluating our model in a simulated clinical environment or prospective clinical trial is a requisite step in determining whether the model is clinically useful. These studies may guide the implementation of the presented model within a decision support system for falls prevention strategies in hospital inpatient environments. Given the use of time-to-event modeling approaches, we also anticipate that machine learning methods that account for time-varying predictors may lead to improved discrimination and calibration.

Acknowledgments

Special thanks for Mr Daniel Cross for assistance with data acquisition. This work was supported by the Digital Health Cooperative Research Centre (DHCRC). DHCRC is funded under the Commonwealth's Cooperative Research Centres (CRC) Program. SM and SC are supported by National Health and Medical Research Council–administered fellowships (#1181138 and #2008313, respectively).

The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

The data used for this study cannot be shared publicly due to ethical and privacy considerations; however, the data analysis scripts are available on a public GitHub repository[28].

Authors' Contributions

RP had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design were contributed by RP, SC, and SMP. Acquisition, analysis, or interpretation of data were contributed by RP, SC, AA-H, and SMP. Drafting of the manuscript was contributed by RP, RB, SC, and SMP. Critical revision of the manuscript for important intellectual content was done by RP, RB, SC, AA-H, and SMP. Statistical analysis was managed by RP. Funding was obtained by SMP. Supervision and administrative, technical, or material support were handled by SC and SMP.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Knot locations for restricted cubic spline terms in the final model.
[DOCX File , 16 KB - jmir_v26i1e59634_app1.docx]

Multimedia Appendix 2
Summary of model parameters, patients, and events used for each model fit during cross-validation.
[DOCX File , 14 KB - jmir_v26i1e59634_app2.docx]

References

1. LeLaurin JH, Shorr RI. Preventing falls in hospitalized patients: state of the science. Clin Geriatr Med 2019;35(2):273-283 [FREE Full text] [doi: 10.1016/j.cger.2019.01.007] [Medline: 30929888]
2. Park S. Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. Aging Clin Exp Res 2018;30(1):1-16. [doi: 10.1007/s40520-017-0749-0] [Medline: 28374345]
3. Cameron ID, Dyer SM, Panagoda CE, Murray GR, Hill KD, Cumming RG, et al. Interventions for preventing falls in older people in care facilities and hospitals. Cochrane Database Syst Rev 2018;9(9):CD005465 [FREE Full text] [doi: 10.1002/14651858.CD005465.pub4] [Medline: 30191554]
4. Morris ME, Webster K, Jones C, Hill AM, Haines T, McPhail S, et al. Interventions to reduce falls in hospitals: a systematic review and meta-analysis. Age Ageing 2022;51(5):afac077 [FREE Full text] [doi: 10.1093/ageing/afac077] [Medline: 35524748]

XSL•FO
RenderX

5. Hill AM, McPhail SM, Waldron N, Etherton-Beer C, Ingram K, Flicker L, et al. Fall rates in hospital rehabilitation units after individualised patient and staff education programmes: a pragmatic, stepped-wedge, cluster-randomised controlled trial. Lancet 2015;385(9987):2592-2599. [doi: 10.1016/S0140-6736(14)61945-0] [Medline: 25865864]
6. Haines TP, Bennell KL, Osborne RH, Hill KD. Effectiveness of targeted falls prevention programme in subacute hospital setting: randomised controlled trial. BMJ 2004;328(7441):676 [FREE Full text] [doi: 10.1136/bmj.328.7441.676] [Medline: 15031238]
7. Matarese M, Ivziku D, Bartolozzi F, Piredda M, de Marinis MG. Systematic review of fall risk screening tools for older patients in acute hospitals. J Adv Nurs 2015;71(6):1198-1209. [doi: 10.1111/jan.12542] [Medline: 25287867]
8. Parsons R, Blythe RD, Cramb SM, McPhail SM. Inpatient fall prediction models: a scoping review. Gerontology 2023;69(1):14-29 [FREE Full text] [doi: 10.1159/000525727] [Medline: 35977533]
9. Barker AL, Morello RT, Wolfe R, Brand CA, Haines TP, Hill KD, et al. 6-PACK programme to decrease fall injuries in acute hospitals: cluster randomised controlled trial. BMJ 2016;352:h6781 [FREE Full text] [doi: 10.1136/bmj.h6781] [Medline: 26813674]
10. Barker A, Kamar J, Graco M, Lawlor V, Hill K. Adding value to the STRATIFY falls risk assessment in acute hospitals. J Adv Nurs 2011;67(2):450-457. [doi: 10.1111/j.1365-2648.2010.05503.x] [Medline: 21105896]
11. Morris ME, Haines T, Hill AM, Cameron ID, Jones C, Jazayeri D, et al. Divesting from a Scored Hospital Fall Risk Assessment Tool (FRAT): a cluster randomized non-inferiority trial. J Am Geriatr Soc 2021;69(9):2598-2604 [FREE Full text] [doi: 10.1111/jgs.17125] [Medline: 33834490]
12. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
13. Rapp K, Ravindren J, Becker C, Lindemann U, Jaensch A, Klenk J. Fall risk as a function of time after admission to sub-acute geriatric hospital units. BMC Geriatr 2016;16(1):173 [FREE Full text] [doi: 10.1186/s12877-016-0346-7] [Medline: 27717326]
14. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211-219 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.014508] [Medline: 25561516]
15. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-387. [doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4] [Medline: 8668867]
16. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245-247 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
17. Practical innovative risk management. RiskMan. 2015. URL: http://www.riskman.net.au/ [accessed 2024-10-05]
18. Harrell FE. Package 'rms': regression modeling strategies. R Project. 2012. URL: https://tinyurl.com/mphz4ezb [accessed 2024-11-06]
19. Takada T, Nijman S, Denaxas S, Snell KIE, Uijl A, Nguyen TL, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. J Clin Epidemiol 2021;137:83-91 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.03.025] [Medline: 33836256]
20. Saha-Chaudhuri P, Heagerty P. Package 'survivalROC': time-dependent ROC curve estimation from censored survival data. R Project. 2022. URL: https://cran.r-project.org/web/packages/survivalROC/survivalROC.pdf [accessed 2024-10-15]
21. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electron J Stat 2015;9(1):1583-1607 [FREE Full text] [doi: 10.1214/15-EJS1035] [Medline: 26279737]
22. LeDell E, Petersen M, van der Laan M. Package 'cvAUC': cross-validated area under the ROC curve confidence intervals. R Project. 2022. URL: https://cran.r-project.org/web/packages/cvAUC/cvAUC.pdf [accessed 2024-10-15]
23. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006;27(8):861-874 [FREE Full text] [doi: 10.1016/j.patrec.2005.10.010]
24. McLernon DJ, Giardiello D, van Calster B, Wynants L, van Geloven N, van Smeden M, et al. Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for Cox proportional hazards models. Ann Intern Med 2023;176(1):105-114. [doi: 10.7326/M22-0844] [Medline: 36571841]
25. Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. BMC Med Res Methodol 2015;15:82 [FREE Full text] [doi: 10.1186/s12874-015-0078-y] [Medline: 26459415]
26. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23(5):723-748. [doi: 10.1002/sim.1621] [Medline: 14981672]
27. Ensor J, Martin E, Riley R. Package 'pmsampsize': sample size for development of a prediction model. R Project. 2023. URL: https://cran.r-project.org/web/packages/pmsampsize/pmsampsize.pdf [accessed 2024-10-15]
28. RWParsons/inpatient-falls-cox-cpm. GitHub. URL: https://github.com/RWParsons/inpatient-falls-cox-cpm [accessed 2024-11-06]

XSL•FO
RenderX

29. Metro South Health hospitals and centres. Queensland Government. 2023. URL: https://tinyurl.com/5fzpnxuk [accessed 2024-06-01]
30. Jacobi L, Petzold T, Hanel A, Albrecht M, Eberlein-Gonska M, Schmitt J. [Epidemiology and prediction of the risk of falling in patients in acute care settings: analysis of routine data from a university hospital]. Z Evid Fortbild Qual Gesundhwes 2017;120:9-15. [doi: 10.1016/j.zefq.2016.12.006] [Medline: 28284369]
31. Hou WH, Kang CM, Ho MH, Kuo JMC, Chen HL, Chang WY. Evaluation of an inpatient fall risk screening tool to identify the most critical fall risk factors in inpatients. J Clin Nurs 2017;26(5-6):698-706. [doi: 10.1111/jocn.13510] [Medline: 27533486]
32. Higaonna M. The predictive validity of a modified Japanese Nursing Association fall risk assessment tool: a retrospective cohort study. Int J Nurs Stud 2015;52(9):1484-1494. [doi: 10.1016/j.ijnurstu.2015.05.009] [Medline: 26138481]
33. Podsiadlo D, Richardson S. J Am Geriatr Soc 1991;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]
34. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
35. Parsons R, Blythe R, Cramb S, McPhail SM. Integrating economic considerations into cutpoint selection may help align clinical decision support toward value-based healthcare. J Am Med Inform Assoc 2023;30(6):1103-1113 [FREE Full text] [doi: 10.1093/jamia/ocad042] [Medline: 36970849]
36. Parsons R, Blythe RD, Barnett AG, Cramb SM, McPhail SM. predictNMB: An R package to estimate if or when a clinical prediction model is worthwhile. JOSS 2023;8(84):5328. [doi: 10.21105/joss.05328]
37. Hsu Y, Weng H, Kuo C, Chu T, Tsai Y. Prediction of fall events during admission using eXtreme gradient boosting: a comparative validation study. Sci Rep 2020;10(1):16777 [FREE Full text] [doi: 10.1038/s41598-020-73776-9] [Medline: 33033326]
38. Lindberg DS, Prosperi M, Bjarnadottir RI, Thomas J, Crane M, Chen Z, et al. Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: a machine-learning approach. Int J Med Inform 2020;143:104272 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104272] [Medline: 32980667]
39. Li C, Zhao K, Zhang D, Pang X, Pu H, Lei M, et al. Prediction models of colorectal cancer prognosis incorporating perioperative longitudinal serum tumor markers: a retrospective longitudinal cohort study. BMC Med 2023;21(1):63 [FREE Full text] [doi: 10.1186/s12916-023-02773-2] [Medline: 36803500]
40. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. arXiv preprint arXiv.00825 2019;20(129):1-30.
41. Kumar N, Qi S, Kuan L, Sun W, Zhang J, Greiner R. Learning accurate personalized survival models for predicting hospital discharge and mortality of COVID-19 patients. Sci Rep 2022;12(1):4472 [FREE Full text] [doi: 10.1038/s41598-022-08601-6] [Medline: 35296767]
42. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med (Lausanne) 2021;8:607952 [FREE Full text] [doi: 10.3389/fmed.2021.607952] [Medline: 34124082]
43. Zeng L, Zhang J, Chen W. Dynamic prediction using time-dependent cox survival neural network. Statistics Machine Learning 2023.
44. Yokota S, Ohe K. Construction and evaluation of FiND, a fall risk prediction model of inpatients from nursing data. Jpn J Nurs Sci 2016;13(2):247-255. [doi: 10.1111/jjns.12103] [Medline: 27040735]
45. Yokota S, Endo M, Ohe K. Establishing a classification system for high fall-risk among inpatients using support vector machines. Comput Inform Nurs 2017;35(8):408-416. [doi: 10.1097/CIN.0000000000000332] [Medline: 28800580]
46. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 1997;315(7115):1049-1053 [FREE Full text] [doi: 10.1136/bmj.315.7115.1049] [Medline: 9366729]

Abbreviations

AUC: area under the curve
EMR: electronic medical record
ROC: receiver operating characteristic
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, A Coristine; submitted 17.04.24; peer-reviewed by C El-Hayek, E Sükei; comments to author 21.05.24; revised version received 03.06.24; accepted 16.09.24; published 13.11.24.

Please cite as:
Parsons R, Blythe R, Cramb S, Abdel-Hafez A, McPhail S
An Electronic Medical Record–Based Prognostic Model for Inpatient Falls: Development and Internal-External Cross-Validation
J Med Internet Res 2024;26:e59634
URL: https://www.jmir.org/2024/1/e59634
doi:10.2196/59634
PMID:

©Rex Parsons, Robin Blythe, Susanna Cramb, Ahmad Abdel-Hafez, Steven McPhail. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factor Analysis of Patients Who Find Tablets or Capsules Difficult to Swallow Due to Their Large Size: Using the Personal Health Record Infrastructure of Electronic Medication Notebooks

Masaki Asano[1], MS; Shungo Imai[1], PhD; Yuri Shimizu[1], HSG; Hayato Kizaki[1], MS; Yukiko Ito[2], BS; Makoto Tsuchiya[2], AA; Ryoko Kuriyama[2], BA; Nao Yoshida[2], BS; Masanori Shimada[2], BS; Takanori Sando[2], BS; Tomo Ishijima[2], MS; Satoko Hori[1], PhD

[1]Division of Drug Informatics, Faculty of Pharmacy and Graduate School of Pharmaceutical Sciences, Keio University, Tokyo, Japan
[2]harmo Co, Ltd, Tokyo, Japan

Corresponding Author:
Satoko Hori, PhD
Division of Drug Informatics
Faculty of Pharmacy and Graduate School of Pharmaceutical Sciences
Keio University
1-5-30 Shibakoen
Minato-ku
Tokyo, 105-8512
Japan
Phone: 81 354002650
Email: satokoh@keio.jp

Abstract

Background: Understanding patient preference regarding taking tablet or capsule formulations plays a pivotal role in treatment efficacy and adherence. Therefore, these preferences should be taken into account when designing formulations and prescriptions.

Objective: This study investigates the factors affecting patient preference in patients who have difficulties swallowing large tablets or capsules and aims to identify appropriate sizes for tablets and capsules.

Methods: A robust data set was developed based on a questionnaire survey conducted from December 1, 2022, to December 7, 2022, using the harmo smartphone app operated by harmo Co, Ltd. The data set included patient input regarding their tablet and capsule preferences, personal health records (including dispensing history), and drug formulation information (available from package inserts). Based on the medication formulation information, 6 indices were set for each of the tablets or capsules that were considered difficult to swallow owing to their large size and concomitant tablets or capsules (used as controls). Receiver operating characteristic (ROC) analysis was used to evaluate the performance of each index. The index demonstrating the highest area under the curve of the ROC was selected as the best index to determine the tablet or capsule size that leads to swallowing difficulties. From the generated ROCs, the point with the highest discriminative performance that maximized the Youden index was identified, and the optimal threshold for each index was calculated. Multivariate logistic regression analysis was performed to identify the risk factors contributing to difficulty in swallowing oversized tablets or capsules. Additionally, decision tree analysis was performed to estimate the combined risk from several factors, using risk factors that were significant in the multivariate logistic regression analysis.

Results: This study analyzed 147 large tablets or capsules and 624 control tablets or capsules. The “long diameter + short diameter + thickness” index (with a 21.5 mm threshold) was identified as the best indicator for causing swallowing difficulties in patients. The multivariate logistic regression analysis (including 132 patients with swallowing difficulties and 1283 patients without) results identified the following contributory risk factors: aged <50 years (odds ratio [OR] 1.59, 95% CI 1.03-2.44), female (OR 2.54, 95% CI 1.70-3.78), dysphagia (OR 3.54, 95% CI 2.22-5.65), and taking large tablets or capsules (OR 9.74, 95% CI 5.19-18.29). The decision tree analysis results suggested an elevated risk of swallowing difficulties for patients with taking large tablets or capsules.

Conclusions: This study identified the most appropriate index and threshold for indicating that a given tablet or capsule size will cause swallowing difficulties, as well as the contributory risk factors. Although some sampling biases (eg, only including

XSL•FO
RenderX

smartphone users) may exist, our results can guide the design of patient-friendly formulations and prescriptions, promoting better medication adherence.

(J Med Internet Res 2024;26:e54645) doi:10.2196/54645

KEYWORDS

tablet; tablets; capsules; capsule; size; personal health record; electronic medication notebook; patient preference; drug; drugs; pharmacy; pharmacies; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; preference; preferences; pill; pills; machine learning; decision tree; swallow; swallowing; throat; pharynx; risk; risks; dysphagia; speech; mobile phone

Introduction

Tablets and capsules are widely used portable drugs. However, the size and shape of these preparations can pose challenges for some patients, leading to difficulties in swallowing and, consequently, suboptimal medication adherence [1]. To improve this situation, it is necessary to share patient preferences regarding tablets or capsules related to the difficulty of swallowing tablets or capsules, such as what size of tablet or capsule product is best for patients and which patients are likely to find tablets or capsules difficult to swallow because of their size. Therefore, several studies have been conducted on the difficulty of swallowing tablets and capsules owing to their size [1-5]. For instance, Kabeya et al [4], in a survey based on 40 patients, identified that “long diameter (longer oval diameter) + short diameter (shorter oval diameter) + thickness” as the best index and a size of less than 22 mm and thickness of 2 to 6 mm was easy for patients to swallow. Sugiyama et al [2], using endoscopic evaluation in healthy subjects, showed increased dysphagia in patients prescribed tablets exceeding 7 mm in diameter. However, these previous studies have notable limitations, including small sample size [1-4], limitations in reflecting real-world patient experiences [1-4], difficulties in accurately identifying patient backgrounds and medications taken because the information is collected primarily through questionnaires [2-4], and reporting bias [5].

To overcome these limitations, we focused on “electronic medication notebooks” to directly collect information on patient preferences regarding tablets and capsules. In Japan, the introduction of the “medication notebook” system has revolutionized the way patients and health care providers document medication use, precautions, and side effects [6]. This prevents side effects and duplicate administration of multiple medications. In recent years, “electronic medication notebook,” which are “medication notebook” in electronic form using functions such as smartphones, have become widespread [7]. These electronic medication notebooks effectively serve as personal health records (PHR) and enable centralized management of the contents of dispensed medications. The electronic medication notebook also enables secondary uses, such as analysis of the accumulated PHR as big data. Furthermore, questionnaires can be sent directly to patients using an electronic medication notebook as a medium, and PHR, such as dispensing histories, can be collected simultaneously as the answers to questionnaires are sent out. However, the use of PHR data based on electronic medication notebooks has not been explored yet.

Owing to the advantages of the electronic medication notebook, we hypothesized that using the electronic medication notebook would reflect the patient assessment in the real world and accurately identify patient backgrounds and medications taken.

In this study, we collected patient preferences regarding tablets or capsules related to “tablets or capsules difficult to swallow due to their large size” from patients themselves via questionnaires sent through the electronic medication notebook app. The purpose of this study was to clarify the appropriate size of these pharmaceutical forms and to identify factors contributing to swallowing difficulties in patients who “find tablets or capsules difficult to swallow due to their large size.”

Methods

Data Sets and Collection Items

A data set encompassing the needs of the patients regarding tablets or capsules was created through a questionnaire survey conducted from December 1, 2022, to December 7, 2022, via harmo, an electronic medication notebook smartphone app operated by harmo Co, Ltd (part 1 in Figures 1 and 2). The data set also included the PHR, such as the dispensing history of the respondent users and drug formulation information in the package insert.

XSL•FO
RenderX

Figure 1. Analysis flow: identification of indices and threshold of the size of tablets or capsules difficult to swallow. ROC: receiver operating characteristic.

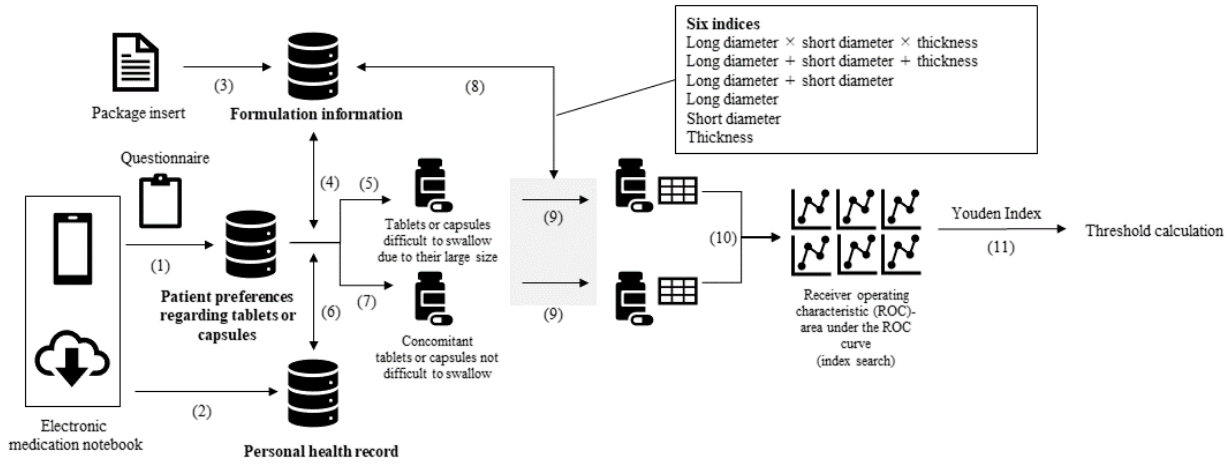

The target population comprised 49,505 men and women aged 20 years or older who had been prescribed either tablets or capsules within 90 days from 2 weeks (November 16) before the survey distribution date (December 1). We sent notifications about the questionnaire to the target population, including a link to the survey details, through the harmo smartphone app. Users were then asked to provide consent and respond through the linked platform. The questionnaire included the following items: whether they found the tablets or capsules too large to swallow, the names of the medications that corresponded to the above question, and whether they had swallowing difficulties. Since larger numbers of questions are associated with lower response rates, it was necessary to set the lowest number of questions possible. Thus, we set simple questions to evaluate whether dysphagia was present without using objective measures such as the 10-item Eating Assessment Tool in this study [8].

For PHR, we obtained the following information from the data stored in the electronic medication notebook (part 2 in Figures 1 and 2) while maintaining anonymity: dispensing history and basic information (gender and age).

Figure 2. Analysis flow: factor analysis in patients who tend to face swallowing difficulties because of tablet or capsule sizes.

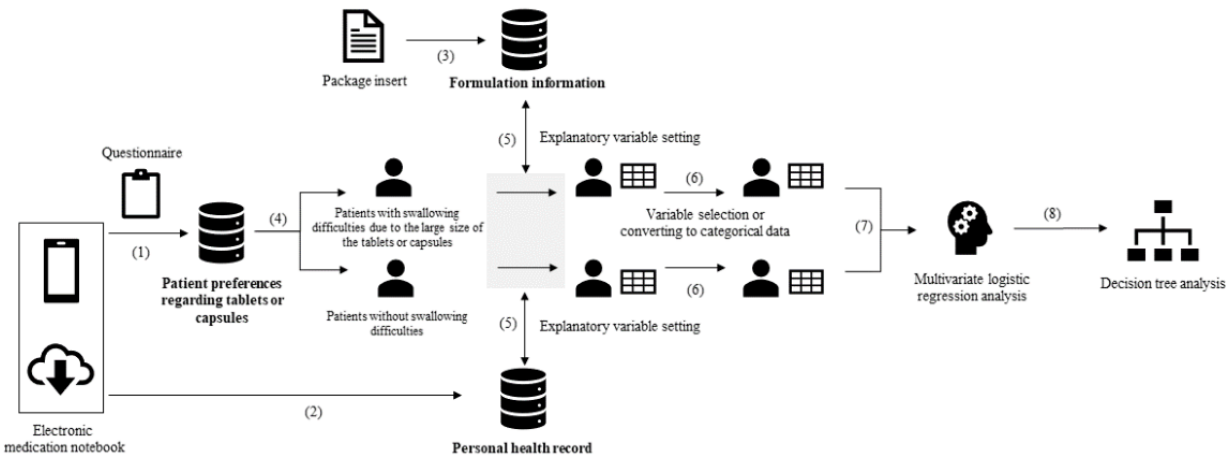

For medication formulation information, we obtained the following information from the package inserts (part 3 in Figures 1 and 2): long diameter, short diameter, and thickness (Figure 3). At this time, we distributed tablets and capsules of the same formulation. This was because a previous study conducted by Kabeya et al [4] demonstrated that the difficulty of swallowing tablets and capsules is almost identical. The diameters were calculated as follows: diameter = long diameter = short diameter for regular circular formulations; capsule diameter = short diameter = thickness for capsule formulations; and diameter = long diameter = short diameter = thickness for spherical formulations.

XSL•FO
RenderX

Figure 3. Schematic showing the measurement methods of long diameter, short diameter, and thickness (each unit is a millimeter).

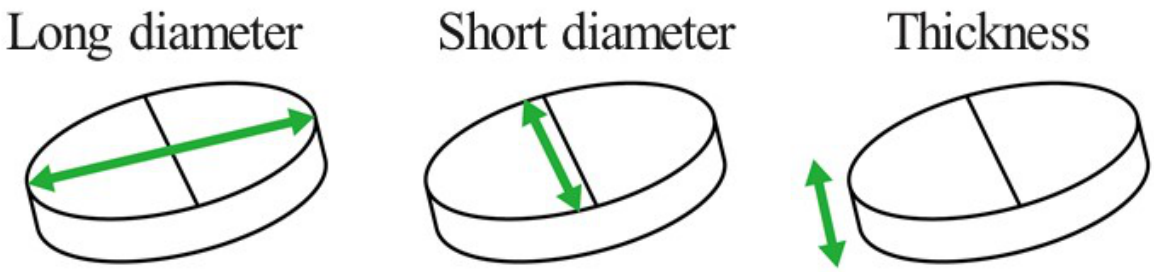

Outcomes

The primary outcome was the risk factors affecting the difficulty in swallowing tablets or capsules owing to their large size. The secondary outcomes were an appropriate index of the size of tablets or capsules difficult to swallow due to their large size, such as "long diameter + short diameter," and a threshold for the size of tablets or capsules difficult to swallow due to their large size.

Identification of Tablet or Capsule Size

We identified the names of medications in the dispensing history of those medications that were reported to be difficult to swallow owing to their large size in the questionnaire. Tablets or capsules with identifiable "long diameter," "short diameter," and "thickness" from the package insert were defined as "tablets or capsules difficult to swallow due to their large size" (parts 4 and 5 in Figure 1).

Tablets or capsules that could be identified from the dispensing history as being taken by the same patient or at the same period as tablets or capsules difficult to swallow due to their large size were defined as "concomitant tablets or capsules" and formed the control group (parts 6 and 7 in Figure 1). In Japan, tablets and capsules are sometimes powdered to make them easier to ingest, for example, by crushing the tablets or removing the powder from the capsule. In this study, we handled tablets and capsules as powders if instructions related to crushing could be obtained from the dispensing comments in the dispensing history.

Search for Indices and Thresholds for the Size of Tablets or Capsules Difficult to Swallow

Based on the medication formulation information, 6 indices were set for each of the tablets or capsules that were difficult to swallow owing to their large size and concomitant tablets or capsules (used as controls; parts 8 and 9 in Figure 1), and the receiver operating characteristic (ROC) was calculated for each index. The index with the largest area under the receiver operating characteristic curve (AUC) was used as the best index of the size of tablets or capsules difficult to swallow because of their large size (part 10 in Figure 1).

From the generated ROCs, the point with the highest discriminative performance that maximized the Youden index was identified, and the optimal threshold for each index was calculated (part 11 in Figure 1) [9].

Factor Analysis of Patients Who Tended to Find Swallowing Difficult Because of the Tablet or Capsule Size

Multivariate logistic regression analysis was performed to identify risk factors for patients who tended to find swallowing difficulties due to the size of the tablets or capsules. Based on information on patient preferences regarding tablets or capsules, the objective variable was defined as the presence or absence of tablets or capsules that were difficult to swallow because of their large size (part 4 in Figure 2). The following explanatory variables were set from patient preferences regarding tablets or capsules, PHR, and formulation information (part 5 in Figure 2): gender, age, dysphagia, number of oral medications per day (including powdered medicine and liquid), number of tablets or capsules per day, number of tablets or capsules per timing, total size of tablets or capsules per day, taking powdered medicine, taking liquid, single-packaging, and taking large tablets or capsules (Note S1 in Multimedia Appendix 1).

Among these explanatory variables, "number of oral medications per day (including powdered medicine and liquid)," "number of tablets or capsules per day," "tablets or capsules per timing," and "total size of tablets or capsules per day" were assumed to have correlations, and therefore were selected from the clinical perspective that the correlation ratio with "taking large tablets or capsules" was the lowest and could be easily calculated by health care workers. For variables that were not binary, cutoff values were determined based on the results of stratified group comparisons and were classified as binary (part 6 in Figure 2). In addition, sensitivity analysis was conducted to evaluate whether different factors were extracted when patients were stratified based on their age (by dividing the identified cutoff value). For this analysis, the same explanatory variables for formal analysis were used except for age. The significance level was set at 5%, and the odds ratios (ORs) and 95% CIs were calculated for each explanatory variable (part 7 in Figure 2).

To estimate the risk of a combination of multiple factors, a decision tree analysis was performed using the significant risk factors in the multivariate logistic regression analysis (part 8 in Figure 2). Chi-Squared Automatic Interaction Detection (CHAID), which divides features using a χ^2 test, was also used [10]. CHAID is an algorithm that creates multiple cross-tabulation tables between the objective variable and each explanatory variable, and then branches the tree by selecting the variable that is most significant from the χ^2 test. The analysis conditions were based on a previous study using the same method [11].

XSL•FO
RenderX

Python (version 3.8.8; Anaconda Inc) was used for data sampling, index search, threshold calculation, and multivariate logistic regression analysis, and SPSS Statistics (version 28.0.1.0; IBM Corp) was used for decision tree analysis.

Ethical Consideration

This study was a self-administered, anonymous questionnaire survey (web-based questionnaire) and all data obtained in this study were anonymized and deidentified. Only those who gave their consent to the questionnaire survey and the provision of PHR owned by the harmo Co, Ltd were asked to complete the questionnaire. This study's protocol was approved by the ethics committee of the Keio University Faculty of Pharmacy (221111-1). No incentives were distributed for survey respondents.

Results

Questionnaire Results on Tablet or Capsule Size

Out of 49,505 patients, 5528 (11.2%) patients opened the survey request notification, 1976 (4%) patients accessed the questionnaire page, and 1501 (3%) patients responded to the questionnaire. Among them, 212 (14.1%) patients answered, "There are tablets or capsules difficult to swallow due to their large size." Regarding dysphagia, 189 (12.6%) patients reported that they "usually found it difficult to swallow while eating" (Table 1).

Table 1. Questionnaire results (N=1501).

Questions and answers	Patients, n (%)
Are there any tablets or capsules difficult to swallow due to their large size?	
Yes	212 (14.1)
No	1271 (84.7)
Do not know	18 (1.2)
Do you usually find it difficult to swallow when eating?	
Yes	189 (12.6)
No	1283 (85.5)
Do not know	29 (1.9)

Search for Indices and Thresholds for the Size of Tablets or Capsules Difficult to Swallow

Figure 4 shows the ROC of 147 tablets or capsules that were difficult to swallow because of their large size and 624 control tablets or capsules and their AUCs. The highest AUC was obtained for "long diameter + short diameter + thickness" (0.9125). Table 2 presents the threshold values calculated from the illustrated ROCs. The threshold value was calculated to be 21.5 mm for "long diameter + short diameter + thickness," which had the highest AUC in the index search.

XSL•FO
RenderX

Figure 4. ROC and AUC of medical tablets or capsule size based on 6 indices. AUC: area under the receiver operating characteristic curve; FPR: false positive rate; ROC: receiver operating characteristic; TPR: true positive rate.

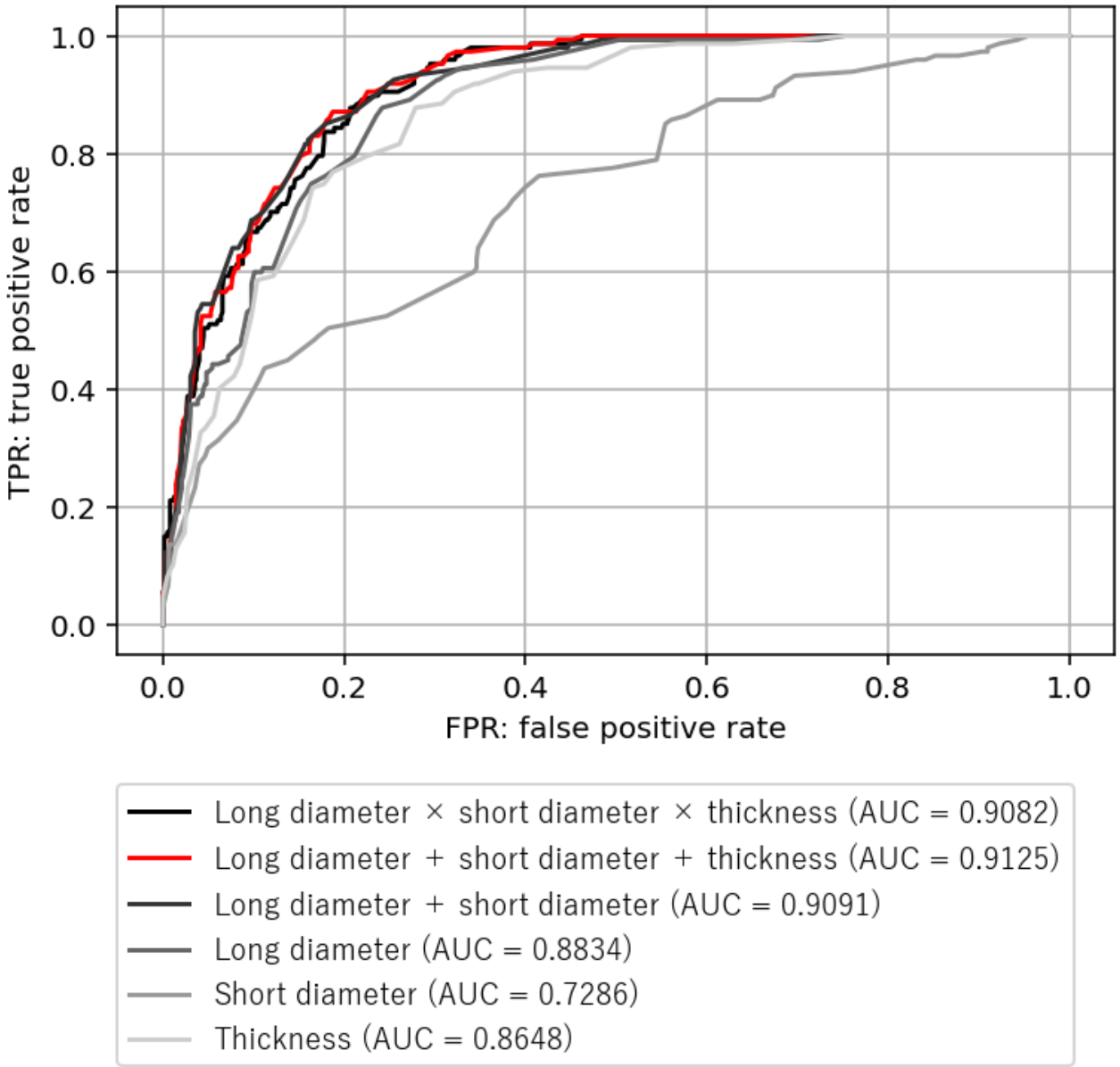

Table 2. Calculation of the size threshold of tablets or capsules that patients find difficult to swallow due to their large size.

Index	Average (SD)	Threshold	Sensitivity	Specificity
Long diameter × short diameter × thickness	278.2 (208.4)	279.5	0.89115	0.78045
Long diameter + short diameter + thickness	20.0 (4.7)	21.5	0.87048	0.81250
Long diameter + short diameter	16.2 (3.7)	17.4	0.85034	0.82051
Long diameter	8.9 (3.1)	8.7	0.87755	0.75801
Short diameter	7.3 (1.5)	7.2	0.76190	0.58494
Thickness	3.8 (1.2)	4.0	0.87755	0.72115

Factor Analysis of Patients Who Tended to Find Swallowing Difficult Because of the Tablet or Capsule Size

Of the 1501 patients, 86 patients with missing data were excluded. The analysis included 1415 patients (132 patients with tablets or capsules swallowing difficulties owing to the size of the formulations and 1283 patients without swallowing difficulties). As shown in Table S1 in Multimedia Appendix 1, "number of oral medications per day (including powdered medicine and liquid)," "number of tablets or capsules per day," "number of tablets or capsules per timing," and "total size of

XSL•FO
RenderX

tablets or capsules per day" showed strong correlations. We selected "number of tablets or capsules per timing" as a variable because of its lowest correlation ratio with "taking large tablets or capsules" (Table S2 in Multimedia Appendix 1) and the clinical perspective that it can be easily calculated by health care workers.

The nonbinary variables "age" and "number of tablets or capsules per timing" were classified as binary variables by setting cutoff values (aged 50 years and number of tablets or capsules per timing: 4) based on the results of stratified group comparison (Tables S3 and S4 in Multimedia Appendix 1).

An elevated percentage of patients who showed difficulties swallowing tablets or capsules because of their size were more likely to be female or aged younger than 50 years (Table S5 in Multimedia Appendix 1). The results of the multivariate logistic regression analysis are presented in Table 3. The variables identified as risk factors were aged <50 years (OR 1.59, 95% CI 1.03-2.44), female (OR 2.54, 95% CI 1.70-3.78), dysphagia (OR 3.54, 95% CI 2.22-5.65), and taking large tablets or capsules (OR 9.74, 95% CI 5.19-18.29). Similar trends were obtained from the sensitivity analysis when patients were divided based on whether they were "younger than 50 years" or "aged 50 years or older" (Tables S6 and S7 in Multimedia Appendix 1).

Table 3. Multivariate logistic regression analysis (N=1415).

Variable name	Odds ratio (95% CI)	P value
Aged <50 years	1.59 (1.03-2.44)	.04
Female	2.54 (1.70-3.78)	<.001
Dysphagia	3.54 (2.22-5.65)	<.001
Number of tablets or capsules per timing ≥4[a]	1.12 (0.74-1.68)	.59
Taking powdered medicine	0.98 (0.62-1.53)	.91
Taking liquid	0.83 (0.22-3.12)	.79
Single-packaging	0.34 (0.04-2.98)	.33
Taking large tablets or capsules[b]	9.74 (5.19-18.29)	<.001

[a]The number of tablets or capsules per time indicates the number of tablets or capsules with the same dosing time (time to reach the maximum number of pieces) as tablets or capsules that were difficult to swallow because of their large size in patients who answered that they had tablets or capsules that they found large and difficult to swallow. In the case of patients who answered that they had no tablets or capsules that were difficult to swallow because of their size, the number was defined as the number of tablets or capsules at the same dosing time (time to reach the maximum number of pieces) as the latest prescription tablet or capsule.

[b]Taking large tablets or capsules indicates the presence or absence of tablets or capsules in the medications taken that exceeded the threshold for the index of tablets or capsule size (long diameter + short diameter + thickness was 21.5 mm).

In the subsequent decision tree analysis based on these risk factors, the risk was higher in patients with "taking large tablets or capsules" and "dysphagia" (percentage of patients who reported difficulty in swallowing: 32 out of 94 patients, 34% patients). Additionally, even in the absence of dysphagia, the risk was higher in patients "taking large tablets or capsules" and who were "female" (percentage of patients who reported difficulty in swallowing: 62 out of 299 patients, 20.7% patients; Figure 5).

XSL•FO
RenderX

Figure 5. Decision tree analysis.

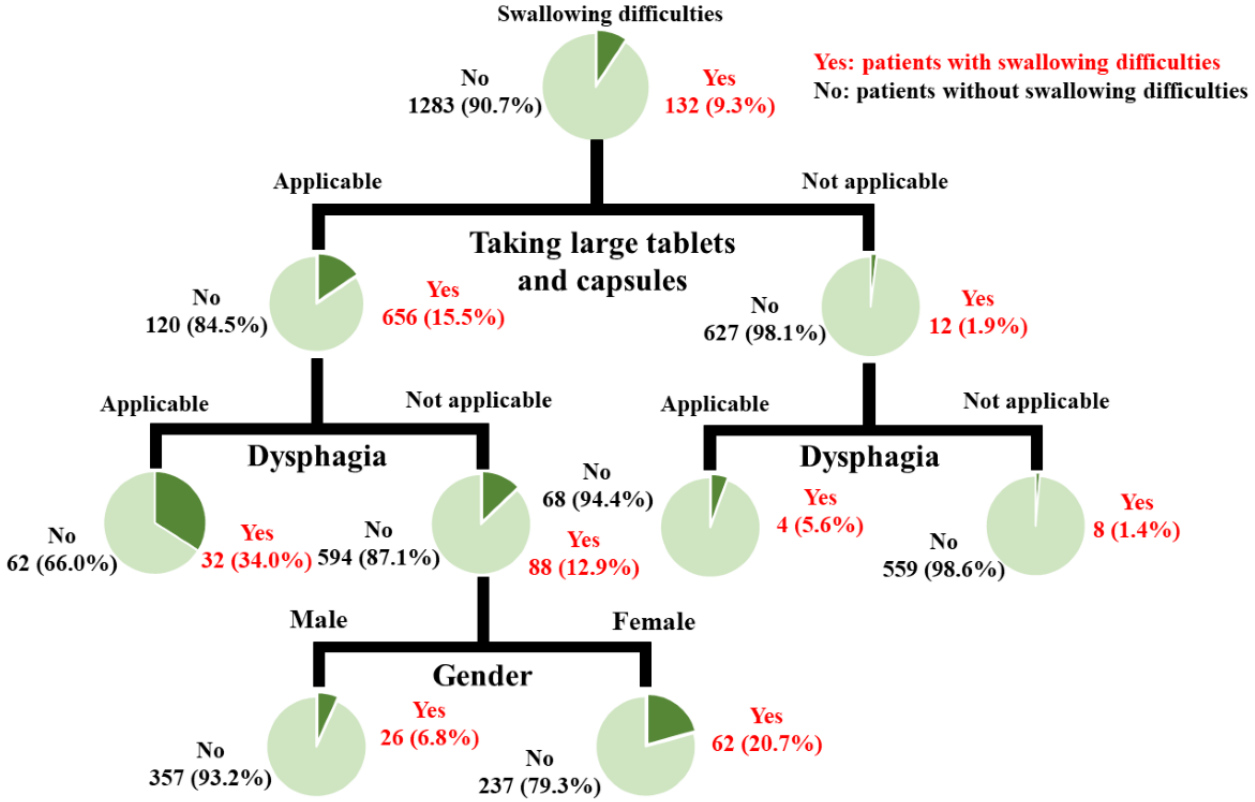

Discussion

Principal Results

In this study, we conducted an analysis using a questionnaire with an electronic medication notebook, PHR, and tablets or capsules formulation information.

We identified the size index and threshold of tablets or capsules that patients find difficult to swallow, as well as the risk factors for patients who find tablets or capsules "too large to swallow." The threshold for the best index was "long diameter + short diameter + thickness" = 21.5 mm, and the risk factors were aged <50 years, female sex, dysphagia, and the use of large tablets or capsules. In addition, patients with "taking large tablets or capsules" and "having dysphagia" and those with "taking large tablets or capsules" and "no dysphagia" and "female" were at relatively high risk.

For the index search and threshold calculation, the PHR-based study design defined "difficult-to-swallow" as well as "not difficult-to-swallow" tablets or capsules. Compared with a previous study based on a geometric analysis of histograms [5], this study enabled conducting a threshold or indicator search using the ROC-AUC and obtaining more generalized findings while strengthening the validity of the previous report.

Results of the Questionnaire About the Size of Tablets or Capsules

Although it is difficult to estimate the percentage of people who find tablets or capsules difficult to swallow because of the differences in methods and study populations, it has been shown that 10%-40% of adult patients in general practice find tablets or capsules difficult to swallow [12-15]. Among the patients evaluated in this study, 212 (14.1%) patients felt that some tablets or capsules were difficult to swallow because their sizes fell within the range of values obtained in these studies. A 2007 US survey on dysphagia in general practice patients aged 18 years showed a prevalence of dysphagia of 22.6% [16]. In this study, 189 (12.6%) patients reported that they usually found it difficult to swallow when eating showed a lower percentage than that in the previous report. This may be due to the lower response rate among older adults in this study (who generally have an elevated prevalence of dysphagia) and racial differences than those in previous studies.

Indicator Search and Threshold Calculation for Tablets or Capsules Difficult to Swallow Owing to Their Large Size

The AUC (0.9125) for "long diameter + short diameter + thickness" was extremely good in the index search and threshold calculation, suggesting that this index has high discrimination power as an index of medical tablets or capsule size. Furthermore, the threshold value of 21.5 mm for "long diameter + short diameter + thickness" was similar to that in a previous report [4] (22 mm: a small clinical study of 40 patients) and was considered to have a certain validity.

Multivariate Logistic Regression Analysis

In multivariate logistic regression analysis, aged <50 years, which was identified as a risk factor, was at odds with the general perception that dysphagia due to tablets or capsules was more frequent in the older age group. However, the result was consistent with those of previous studies, in which younger

XSL•FO
RenderX

patients were more likely to report difficulty in swallowing tablets or capsules than older patients [12,17].

These results suggest that the measures commonly taken to make medication easier for older patients may not be considered for younger patients. It is also possible that this was due to younger and middle-aged patients' unfamiliarity with taking tablets or capsules compared to older patients who take medications daily and are accustomed to taking tablets or capsules. However, we believe that further verification of whether age is a true risk factor for swallowing difficulties is needed. This is because age was not extracted in the DT model, that is, its association with swallowing difficulties may be relatively weaker than that of other variables. In addition, as described above, age was likely influenced by external factors such as the absence of measures to make it easier to swallow tablets and capsules.

Female gender has been identified as a risk factor in several studies [12,13,18]. Schiele et al [12] attributed this to physiological processes and anatomical differences in the dimensions and function of the mouth, pharynx, upper esophageal sphincter, and esophagus, as well as a higher incidence of depression and anxiety disorders. Dysphagia was identified as a risk factor in a previous study in which patients attending general practice were asked to complete a questionnaire, and the same trend was observed [12]. The same trend was observed for large tablets or capsules because the size of the solid dosage form has been cited as a cause of dysphagia in many studies [1-5]. Furthermore, the high OR (9.74) and 95% CI (5.19-18.29) suggest that the size of the medical tablet or capsule is the most important factor in avoiding risk among the risk factors identified in this study. The validity of the threshold value of 21.5 mm for "long diameter + short diameter + thickness," which was calculated only for patients who reported having a difficult-to-swallow medical tablet or capsule, was strengthened by conducting a factor analysis that included patients who reported having no difficult-to-swallow tablets or capsules.

Decision Tree Analysis

This study investigated the risk estimation of swallowing by combining factors in a decision tree analysis. In the CHAID model used in this study, the order of the branches reflects the importance of each feature. Since the explanatory variable in the highest branch was "taking large tablets or capsules," it was the explanatory variable that contributed the most to the objective variable. The risk was higher in patients with "taking large tablets or capsules" and "dysphagia" (32/94, 34% of patients). Additionally, even in the absence of dysphagia, the risk was higher in patients "taking large tablets or capsules" and who were "female" (62/299, 20.7%). In contrast, the risk was low in patients who did not take large tablets or capsules (12/639, 1.9%). These results identified a more detailed group of patients for whom health care workers should provide easier-taking measures and advice when prescribing medications compared with previous studies.

Clinical Application

Tablets or capsules that surpass the determined threshold should be subject to formulation adjustments, such as transformation into orally disintegrating tablets, divisible through a dividing line, or formulation of smaller dose units. It is important to design a medical tablet or capsule formulation that does not exceed the threshold size to reduce swallowing difficulty. Furthermore, it is important to design prescriptions and provide guidance to younger patients aged younger than 50 years to ease intake of such formulations, ultimately enhancing compliance with prescribed doses. Moreover, the implementation of straightforward monitoring practices, such as asking patients about their daily experiences with swallowing difficulties during meals and beverages, can serve as an effective means of identifying individuals at a higher risk of dysphagia. Early intervention, including a potential change in dosage form, can then be promptly initiated in response to these findings.

Limitations

This study acknowledges limitations stemming from the characteristics of the electronic medication notebook and the questionnaire design. Among the limitations associated with the electronic medication notebook, population bias is a notable concern. Furthermore, in this questionnaire survey conducted using harmo, there was a tendency for younger to middle-aged people rather than older people, which could be attributed to the penetration rate. The skewed distribution contrasts with the age profile of hospital outpatients in Japan [19], raising the possibility that the results may have been influenced by this sampling bias. Furthermore, our focus was on those individuals among older people who can complete the questionnaire using their smartphones. A survey conducted among residents of a nursing home revealed that 68% of older participants experienced difficulty in swallowing at mealtime [20]. Therefore, the older people who find it difficult to swallow tablets or capsules may be underestimated. In addition, the medication intake in patients cannot be ensured in that patients actually take the medication, as recorded in the dispensing history, and there is no guarantee that all tablets or capsules taken by patients are recorded in the electronic medication notebook. Furthermore, there is no certainty that all prescribed medications are recorded in the electronic medication book as part of the dispensing history.

The limitations stemming from the questionnaire-based design include the potential for response bias, as this study relies on active participation and interest in swallowing tablets or capsules. Another factor was the high exclusion rate due to missing data, resulting in a reduced sample size from 212 to 132 patients in the factor analysis. Furthermore, we could not use objective measures of dysphagia (such as the 10-item Eating Assessment Tool) due to the need to minimize the number of included questions.

Despite these limitations, this study presents an innovative approach aimed at generating tablets or capsules and prescription considerations using the data collected from electronic medication notebooks. The novelty lies in risk estimation through the combination of various factors and the validation of the usefulness of the PHR infrastructure. We anticipate that

XSL•FO
RenderX

the use of PHR will continue to grow in the future, and this study serves as a pioneering step in this promising direction.

Conclusions

Combining the information on patient preferences regarding tablets or capsules obtained from the questionnaire survey of the electronic medication notebook, the user's PHR and medication formulation information revealed that the most suitable indicator for determining the size of medical tablets or capsules causing swallowing difficulties is "long diameter + short diameter + thickness," with a threshold of 21.5 mm. In addition, factors such as "aged <50 years," "female," "dysphagia," and "taking large tablets or capsules" were identified as patient backgrounds and prescribing backgrounds that are likely to cause patients to feel difficulty in swallowing large tablets or capsules. Furthermore, patients who ingested "large tablets or capsules" and had "dysphagia" were shown to be at an elevated risk of difficulty in swallowing. Although some sampling biases could not be avoided due to the low response rate (1501/49,505, 3% of patients) and specific survey population (ie, only smartphone users), these findings have significant implications for optimizing the design of patient-friendly formulations and prescriptions.

Acknowledgments

We would like to thank all participants who responded to our survey.

Data Availability

The questionnaires and PHR data for this study are available upon reasonable request.

Authors' Contributions

MA, SI, HK, and SH designed this study. MA, YS, HK, SI, and SH created and evaluated the questionnaire. MA and YS processed the data and performed the statistical analysis. YI, MT, RK, NY, MS, and TS provided support for questionnaire development, distribution, collection of responses, and provision of PHRs. MA, SI, and SH drafted and completed this paper. TI and SH supervised this study. YS, YI, MT, RK, NY, MS, TS, and TI reviewed this paper. All authors have approved this paper.

Conflicts of Interest

TS and TI are co-CEOs of harmo Co, Ltd, which operates harmo. YI, MT, RK, NY, and MS are employees of harmo Co, Ltd. Other authors declare no conflicts of interest.

Multimedia Appendix 1
A note and tables.
[DOCX File , 34 KB - jmir_v26i1e54645_app1.docx]

References

1. Shariff Z, Kirby D, Missaghi S, Rajabi-Siahboomi A, Maidment I. Patient-centric medicine design: key characteristics of oral solid dosage forms that improve adherence and acceptance in older people. Pharmaceutics 2020;12(10):905 [FREE Full text] [doi: 10.3390/pharmaceutics12100905] [Medline: 32977445]
2. Sugiyama S, Iida T, Morimoto Y, Yamazaki Y, Mikuzuki L, Hayashi M. Effects of tablet size and head posture on drug swallowing: a preliminary examination using endoscopy in healthy subjects. Acta Med Okayama 2021;75(4):495-503 [FREE Full text] [doi: 10.18926/AMO/62402] [Medline: 34511617]
3. Hummler H, Stillhart C, Meilicke L, Grimm M, Krause E, Mannaa M, et al. Impact of tablet size and shape on the swallowability in older adults. Pharmaceutics 2023;15(4):1042 [FREE Full text] [doi: 10.3390/pharmaceutics15041042] [Medline: 37111528]
4. Kabeya K, Satoh H, Hori S, Sawada Y. Experimental study on patient preferences regarding the shape and size of medical tablets and capsules using three-dimensionally printed plastic model formulations. Patient Prefer Adherence 2021;15:863-870 [FREE Full text] [doi: 10.2147/PPA.S306582] [Medline: 33953546]
5. Kabeya K, Satoh H, Hori S, Miura Y, Sawada Y. Threshold size of medical tablets and capsules: based on information collected by Japanese medical wholesaler. Patient Prefer Adherence 2020;14:1251-1258 [FREE Full text] [doi: 10.2147/PPA.S253663] [Medline: 32801655]
6. Yamaura K, Masuda M, Motohashi M, Kuramochi K, Ishizuka T, Someya T, et al. [Investigation of the "medication notebook" about its utilization in patients and its evaluation by doctors in Iwai City. Yakugaku Zasshi 2003;123(3):151-155 [FREE Full text] [doi: 10.1248/yakushi.123.151] [Medline: 12693015]
7. Orii T. Current status and issues of electronic information provision for risk minimization: coordination between electronic medicine notebook and patient drug information. Yakugaku Zasshi 2018;138(3):325-329 [FREE Full text] [doi: 10.1248/yakushi.17-00185-4] [Medline: 29503423]

XSL•FO
RenderX

8. Belafsky PC, Mouadeb DA, Rees CJ, Pryor JC, Postma GN, Allen J, et al. Validity and reliability of the Eating Assessment Tool (EAT-10). Ann Otol Rhinol Laryngol 2008;117(12):919-924. [doi: 10.1177/000348940811701210] [Medline: 19140539]
9. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32-35 [FREE Full text] [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
10. Kass GV. An exploratory technique for investigating large quantities of categorical data. J R Stat Soc Ser C Appl Stat 1980;29(2):119-127. [doi: 10.2307/2986296]
11. Imai S, Yamada T, Kasashi K, Kobayashi M, Iseki K. Usefulness of a decision tree model for the analysis of adverse drug reactions: evaluation of a risk prediction model of vancomycin-associated nephrotoxicity constructed using a data mining procedure. J Eval Clin Pract 2017;23(6):1240-1246. [doi: 10.1111/jep.12767] [Medline: 28544476]
12. Schiele JT, Quinzler R, Klimm HD, Pruszydlo MG, Haefeli WE. Difficulties swallowing solid oral dosage forms in a general practice population: prevalence, causes, and relationship to dosage forms. Eur J Clin Pharmacol 2013;69(4):937-948. [doi: 10.1007/s00228-012-1417-0] [Medline: 23052416]
13. 40% of American adults report experiencing difficulty swallowing pills—national survey shows many failed to take medication as directed—because of difficulty swallowing pills. PR Newswire Association LLC. New York; 2013. URL: https://jacksonbondenterprises.wordpress.com/2013/09/18/40-of-american-adults-report-experiencing-difficulty-swallowing-pills/ [accessed 2023-10-25]
14. Marquis J, Schneider MP, Payot V, Cordonier AC, Bugnon O, Hersberger KE, et al. Swallowing difficulties with oral drugs among polypharmacy patients attending community pharmacies. Int J Clin Pharm 2013;35(6):1130-1136. [doi: 10.1007/s11096-013-9836-2] [Medline: 23963541]
15. Tahaineh L, Wazaify M. Difficulties in swallowing oral medications in Jordan. Int J Clin Pharm 2017;39(2):373-379. [doi: 10.1007/s11096-017-0449-z] [Medline: 28281225]
16. Wilkins T, Gillies RA, Thomas AM, Wagner PJ. The prevalence of dysphagia in primary care patients: a HamesNet Research Network study. J Am Board Fam Med 2007;20(2):144-150 [FREE Full text] [doi: 10.3122/jabfm.2007.02.060045] [Medline: 17341750]
17. Lau ETL, Steadman KJ, Cichero JAY, Nissen LM. Dosage form modification and oral drug delivery in older people. Adv Drug Deliv Rev 2018;135:75-84. [doi: 10.1016/j.addr.2018.04.012] [Medline: 29660383]
18. Andersen O, Zweidorff OK, Hjelde T, Rødland EA. Problems when swallowing tablets. A questionnaire study from general practice. Tidsskr Nor Laegeforen 1995;115(8):947-949. [Medline: 7709385]
19. Labour and Welfare. Kanja Chosa no Gaiyo. Ministry of Health. 2022. URL: https://www.mhlw.go.jp/toukei/saikin/hw/kanja/20/dl/gaiyou.pdf [accessed 2023-10-25]
20. Steele CM, Greenwood C, Ens I, Robertson C, Seidman-Carlson R. Mealtime difficulties in a home for the aged: not just dysphagia. Dysphagia 1997;12(1):43-50; discussion 51. [doi: 10.1007/pl00009517] [Medline: 8997832]

Abbreviations

AUC: area under the receiver operating characteristic curve
CHAID: Chi-Squared Automatic Interaction Detection
EAT-10: 10-Item Eating Assessment Tool
OR: odds ratio
PHR: personal health record
ROC: receiver operating characteristic

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 21.11.23; peer-reviewed by S Javorszky, N Okada; comments to author 05.12.23; revised version received 22.12.23; accepted 19.03.24; published 24.04.24.

Please cite as:
Asano M, Imai S, Shimizu Y, Kizaki H, Ito Y, Tsuchiya M, Kuriyama R, Yoshida N, Shimada M, Sando T, Ishijima T, Hori S
Factor Analysis of Patients Who Find Tablets or Capsules Difficult to Swallow Due to Their Large Size: Using the Personal Health Record Infrastructure of Electronic Medication Notebooks
J Med Internet Res 2024;26:e54645
URL: https://www.jmir.org/2024/1/e54645
doi:10.2196/54645
PMID:38657229

©Masaki Asano, Shungo Imai, Yuri Shimizu, Hayato Kizaki, Yukiko Ito, Makoto Tsuchiya, Ryoko Kuriyama, Nao Yoshida, Masanori Shimada, Takanori Sando, Tomo Ishijima, Satoko Hori. Originally published in the Journal of Medical Internet Research

XSL•FO
RenderX

(https://www.jmir.org), 24.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tanzania's and Germany's Digital Health Strategies and Their Consistency With the World Health Organization's Global Strategy on Digital Health 2020-2025: Comparative Policy Analysis

Felix Holl[1], MPH, MSc, PhD; Jennifer Kircher[1], BA; Attila J Hertelendy[2], BHS, MSc, PhD; Felix Sukums[3], MSc, PhD; Walter Swoboda[1], Dr med

[1]DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany

[2]Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, FL, United States

[3]MUHAS Digital Health and Innovation Research Group, Muhimbili University of Health & Allied Sciences, Dar es Salaam, United Republic of Tanzania

Corresponding Author:
Felix Holl, MPH, MSc, PhD
DigiHealth Institute
Neu-Ulm University of Applied Sciences
Wileystraße 1
Neu-Ulm, 89231
Germany
Phone: 49 7319762 ext 1613
Email: felix.holl@hnu.de

Abstract

Background: In recent years, the fast-paced adoption of digital health (DH) technologies has transformed health care delivery. However, this rapid evolution has also led to challenges such as uncoordinated development and information silos, impeding effective health care integration. Recognizing these challenges, nations have developed digital health strategies (DHSs), aligning with their national health priorities and guidance from global frameworks. The World Health Organization (WHO)'s Global Strategy on Digital Health 2020-2025 (GSDH) guides national DHSs.

Objective: This study analyzes the DHSs of Tanzania and Germany as case studies and assesses their alignment with the GSDH and identifies strengths, shortcomings, and areas for improvement.

Methods: A comparative policy analysis was conducted, focusing on the DHSs of Tanzania and Germany as case studies, selected for their contrasting health care systems and cooperative history. The analysis involved a three-step process: (1) assessing consistency with the GSDH, (2) comparing similarities and differences, and (3) evaluating the incorporation of emergent technologies. Primary data sources included national eHealth policy documents and related legislation.

Results: Both Germany's and Tanzania's DHSs align significantly with the WHO's GSDH, incorporating most of its 35 elements, but each missing 5 distinct elements. Specifically, Tanzania's DHS lacks in areas such as knowledge management and capacity building for leaders, while Germany's strategy falls short in engaging health care service providers and beneficiaries in development phases and promoting health equity. Both countries, however, excel in other aspects like collaboration, knowledge transfer, and advancing national DHSs, reflecting their commitment to enhancing DH infrastructures. The high ratings of both countries on the Global Digital Health Monitor underscore their substantial progress in DH, although challenges persist in adopting the rapidly advancing technologies and in the need for more inclusive and comprehensive strategies.

Conclusions: This study reveals that both Tanzania and Germany have made significant strides in aligning their DHSs with the WHO's GSDH. However, the rapid evolution of technologies like artificial intelligence and machine learning presents challenges in keeping strategies up-to-date. This study recommends the development of more comprehensive, inclusive strategies and regular revisions to align with emerging technologies and needs. The research underscores the importance of context-specific adaptations in DHSs and highlights the need for broader, strategic guidelines to direct the future development of the DH ecosystem. The WHO's GSDH serves as a crucial blueprint for national DHSs. This comparative analysis demonstrates the value and challenges of aligning national strategies with global guidelines. Both Tanzania and Germany offer valuable insights into developing and implementing effective DHSs, highlighting the importance of continuous adaptation and context-specific considerations. Future

XSL•FO
RenderX

policy assessments require in-depth knowledge of the country's health care needs and structure, supplemented by stakeholder input for a comprehensive evaluation.

(J Med Internet Res 2024;26:e52150) doi:10.2196/52150

KEYWORDS

digital health strategies; comparative policy analysis; DHS; eHealth; digital health; strategy; strategies; policy; policies; document analysis; document analyses; GSDH; Germany; Europe; Africa; Tanzania

Introduction

Technological advances have fostered the adoption of new digital health (DH) technologies [1]. Such adoption has been so rapid that negative side effects such as uncoordinated implementation have resulted [2]. Uncoordinated development can lead to siloed information systems with limited interpretability with other information systems [3]. To combat uncoordinated development, countries have developed digital health strategies (DHSs) to set strategic guidance for the future development of DH technologies, often aligned with their own national health priorities. The World Health Organization (WHO) developed the Global Strategy on Digital Health 2020-2025 (GSDH) to provide global guidance to member countries in defining their DHSs [4]. It is the first comprehensive strategy for DH globally [5]. The GSDH encourages the development of national DHSs based on expert knowledge and WHO member state consensus. National DHSs often fail to capture the most emergent technologies such as artificial intelligence (AI), machine learning, or drones because of their fast development cycles, which can lead to uncoordinated development and adoption [6]. While no country has reported that they used the GSDH specifically to develop their national DHS yet, the GSDH may have informed elements of national DHS and existing DHSs can be assessed for their consistency with the GSDH.

Tanzania's Ministry of Health recognizes the importance of DH technologies in health care delivery and has published 2 strategy documents within the last decade. The *Tanzania National eHealth Strategy 2013-2018* focused on the infrastructure and technologies needed to support the transformation of the health sector [7]. In contrast, the latest *Digital Health Strategy July 2019-June 2024* addresses, among other aspects, challenges that have not been addressed or solved and the best use of technology to improve patient care [8]. It also focuses on strategic priorities aligned with the WHO health system building blocks and advocating user-centric, interoperable, and data-driven DH interventions.

Germany has a federal system of government, with 16 state-level ministries of health and 1 Federal Ministry of Health (FMoH) plus other institutions responsible for different areas of DH. Between 2015 and 2021, each entity published corresponding strategies, strategic documents, recommendations, or legislation about the digitalization of health care in Germany, many of which were included in this study. Before the publication of its first National Digitalization Strategy in March 2023 [9], Germany lacked a uniform strategy with binding goals and guidelines.

Tanzania and Germany have a history of cooperation in socioeconomic development, including health sector improvement. The 2 countries were chosen as case studies of the development of DHS in the global South and North. We draw on our own extensive experience in developing and implementing DH systems in these 2 countries. The purpose of this study is to investigate the key elements and possible shortcomings of the DHSs of Tanzania and Germany, investigate their alignment with the WHO GSDH, and make applicable recommendations to improve the DHS for the 2 countries. We also aim to identify aspects of the WHO GSDH that are challenging for countries to implement. The comparison of the WHO GSDH with the DHSs of Tanzania and Germany is intended as a case study for this policy comparison approach to conduct additional analyses and share this method with other researchers to stimulate similar analyses in other countries.

Methods

We compared the DHSs of Tanzania and Germany and assessed their consistency with the WHO's GSDH, using a document analysis and comparative policy analysis approach [10,11]. We selected Tanzania and Germany as case studies for this comparison, as the researchers had in-depth knowledge about the health care systems and the state of DH in the 2 countries. This in-depth understanding is a requirement for case studies in comparative policy analyses [11]. We searched for the relevant primary data sources through searches in academic databases, through search engines, and based on the expert knowledge of the researchers and other subject matter experts. We identified primary data sources that were national eHealth policy documents and technology and health-related policy documents (see Textbox 1).

Textbox 1. Primary data sources from the World Health Organization, Tanzania, and Germany used for the analysis with references.

Global Strategy on Digital Health 2020-2025—World Health Organization [4]
Digital Health Strategy July 2019-June 2024—Tanzania [8]
Different Approaches of Digital Health Strategies—Germany (various sources, a list of all included laws and regulations can be found in the Results section)

XSL•FO
RenderX

Once the primary data sources were identified, data extraction and analysis were done in six steps: (1) development of a category system for data extraction and analysis; (2) data extraction; (3) assessment of the data from Germany and Tanzania for their compliance with the GSDH; (4) analysis of similarities and differences between of the strategies of Germany and Tanzania; (5) categorical summary and regrouping into policy, infrastructural, and human factors; and (6) assessment of how new technologies are incorporated in each strategy.

In the first step of this study, we created a deductive category system to assess consistency with the WHO's GSDH 2020-2025. This step was conducted by 1 researcher (JK) and validated by a second researcher (FH). The categories were derived from the 4 dimensions defined in the GSDH:

- collaboration and knowledge transfer
- advance the implementation of national DHSs
- strengthen governance for DH at global, regional, and national levels
- integrated people-centered health systems enabled by DH technologies

Data from the primary data sources were extracted and classified into 1 of the 4 categories by 1 researcher (JK). A second researcher (FH) validated both the extrication and the classification. After the initial classification, the results within each category were categorized into one of the three subsections: (1) policy options, (2) measures, or (3) outcomes. The initial categorization was done by 1 researcher (JK) and validated by a second researcher (FH).

After the extrication and categorization were completed, we assessed all extracted data for compliance with the GSDH (initial assessment by JK, validation for Germany by FH and for Tanzania by FS). In the second step of the analysis, we identified similarities and differences in the strategies (initial assessment by JK, validation for Germany by FH and for Tanzania by FS). Following an inductive content analysis (by JK), in a third step, the extracted data were summarized categorically and regrouped into *policy*, *infrastructural*, and *human factors* [12] (by JK, and validated by FH and FS). Finally, as the fourth and last step, the extent to which new technologies such as AI, machine learning, or drones were included in the DHSs was assessed (initial assessment by JK, validation for Germany by FH and for Tanzania by FS). A detailed breakdown of the results of the policy comparison by aspect can be found in the Multimedia Appendix 1.

Ethical Considerations

This study only analyzed policy documents. No human subjects were investigated or personal data analyzed. All data that were handled as part of this study were stored and analyzed on encrypted devices. Any personal identifiers from the primary data, such as author names or contact details were removed during data extraction of the primary data into our data set. The extracted data are available in a summarized form in the Multimedia Appendix 2 [4,7-9,13-27] and the full data set is available on request from the authors.

Results

Consistency With the WHO's GSDH 2020-2025

Both Germany's and Tanzania's DHS include most of the 35 elements from the WHO's GSDH. Both DHSs do not include 5 elements each. The items that are not included are shown in Table 1.

XSL•FO
RenderX

Table 1. Items of the World Health Organization's GSDH[a] not included in the DHSs[b] of Tanzania and Germany

Section	Item
Tanzania	
(1) Collaboration and knowledge transfer: policy options and actions	(2) Establish a knowledge management approach to identify and share good practices, knowledge about the implementation of new methods and techniques, evidence, and lessons learned on digital health across countries and international communities
(3) Strengthen governance for digital health at global, regional, and national levels: policy options and actions	(4) Promote capacity-building for leaders of public health authorities and affiliated agencies, and policy makers to make informed decisions to support digital health investments
(3) Strengthen governance for digital health at global, regional, and national levels: outputs	(3) Global guidance on planning, development, and use of digital hospitals, digital clinical trials, and digital therapeutics is developed
(3) Strengthen governance for digital health at global, regional, and national levels: outputs	(4) A set of recommendations is developed for pseudonymization and anonymization of health data
(4) Advocate people-centered health systems that are enabled by digital health: outputs	(4) Global guidance on personalized medicine is developed
Germany	
(2) Advance the implementation of national DHSs: policy options and actions	(2) Ensure that institutions, decision makers, and personnel involved in the provision of health care services and all end user communities and beneficiary populations are adequately engaged in the design and development phases
(2) Advance the implementation of national DHSs: policy options and actions	(7) Design, implement, and monitor a change management plan, to support conducive organizational behavior surrounding newly digitized health processes and practices
(3) Strengthen governance for digital health at global, regional, and national levels: policy options and actions	(4) Promote capacity-building for leaders of public health authorities, affiliated agencies, and policy makers to make informed decisions to support digital health investments
(4) Advocate people-centered health systems that are enabled by digital health: policy options and actions	(4) Strengthen gender equality and health equity approaches and accessibility for people with disabilities to promote an inclusive digital society with enhanced digital health skills. When planning and prioritizing digital health interventions, relevant factors of inequality should be assessed to ensure that the introduction of digital health technologies does not aggravate these ("do no harm") and that access for specific population groups is guaranteed. In addition, the specific potential of digital technologies to promote health equity should be leveraged. Designed properly, digital solutions can propel inclusiveness as digital connectivity can transcend physical barriers
(4) Advocate people-centered health systems that are enabled by digital health: outputs	(2) A framework allowing individual feedback in validating the performance of digital health tools and services, diffusion of increasing digital health demand is implemented and used

[a]GSDH: Global Strategy on Digital Health 2020-2025.

[b]DHS: digital health strategy.

(1) Collaboration and Knowledge Transfer

As a member of the European Union, Germany is involved in shaping a global DHS and published the strategy document *Strategy of the Federal Government on Global Health* in 2020 [13]. Germany participates in various programs such as Horizon Europe, aiming to promote a knowledge- and innovation-based society and a competitive, sustainable economy [14]. Multi-stakeholder meetings are convened to overcome the implementation hurdles of digitale gesundheitsanwendungen (DiGAs) and innovations through various initiatives. For example, the goal of the *German Alliance for Global HealthResearch* is to expand its research network [15]. To facilitate the exchange of results with partners and institutions across countries, the focus of the *German Alliance for Global HealthResearch* is also on a research-compatible data infrastructure with international standards [16]. Knowledge transfer, especially to the global south is an integral part of the German strategy.

Tanzania's strategy focuses primarily on national challenges and opportunities. The aim is to improve health services at all levels of the country's health system [8]. Key stakeholders are assigned to sectors in the strategy [8]. According to priority 3 of Tanzania's strategy, a knowledge management approach will be expanded and developed only at the population and health worker level, for example, through e-learning platforms [8]. More digital solutions are being developed, according to priority 10 [8], to improve surveillance of and reporting on notifiable

XSL•FO
RenderX

diseases, disease outbreaks, and disasters to prevent loss of life and socioeconomic impact [28].

(2) Advance the Implementation of National DHSs

Due to the prevailing federalism in Germany, each federal state has its own Ministry of Health at the state level and can therefore enact its own laws and rules. However, higher-level legislation such as the *E-Health Law* [17], which was introduced in 2015, can force the federal states to adapt to the national strategy. The E-Health Law has set the initial course for the development of the secure telematics infrastructure (TI) and the introduction of DiGAs. The law is driving forward digitalization in the health care sector for the benefit of patients. It contains a roadmap for the introduction of a digital infrastructure and allows insured persons to benefit from specific applications [17]. In addition, at annual conferences, the health ministers of all federal states represent and discuss the interests of the federal states and health policy issues [18]. Instead of a national digitization strategy, Germany has formal roadmaps based on legislation and recommendations, such as the *Roadmap Digital Health* [19]. Within the framework of the Innovation Forum *Digital Health 2025* of the FMoH, 5 fields of action for the future have been defined, which function as a target and implementation blueprint. The five fields are (1) building a sustainable basis, (2) digital care as the normality, (3) overcoming institutional and sectoral (digital) boundaries, (4) strengthening data literacy—making health data usable, and (5) using new technologies to enable individualized medicine [16]. However, a clear and structured prioritization to achieve the goals is not evident.

To lead Tanzania into the digital age and guide progress and development, the government developed the *Digital Health Strategy July 2019-June 2024*. This strategy includes a clear prioritization derived from the *National Health Policy 2019* and complemented by a rigorous consultation process with key stakeholders in the health sector [8]. It sets standards for data and technologies and strengthens interoperability between systems and sectors [8]. “Improving the legal and regulatory framework to ensure client safety, data security, confidentiality, and privacy” is identified in priority 1 [8]. It identifies the development of a change management plan as an essential factor for the successful implementation and adoption of DH solutions.

(3) Strengthen Governance for DH at Global, Regional, and National Levels

In Germany, several laws and regulations came into force in 2019 and the following years, leading to a stable legal framework for DH. Most recently, the *Health IT Interoperability Governance Regulation* was published in 2021, specifying the tasks of the interoperability coordination unit and the expert panel [20]. Other programs, such as the innovation initiative “Data for Health: Roadmap for Better Patient Care through Health Research and Digitalization” [21], and various working groups, deal with data context [22]. DiGAs are described as particularly innovative and are defined as prescription applications reimbursable by the public health insurance system [29]. As such, DiGAs must meet high requirements and demonstrate an evidence-based medical benefit. All these laws and regulations were related to DH. An overview of them can be found in Table 2.

Table 2. Laws and regulations regarding digital health in Germany with their respective year of publication.

Reference	Law and regulation	Year of publication
[9]	DIGITAL TOGETHER Germany’s Digitalisation Strategy for Health and Care	2023
[13]	Strategie der Bundesregierung zur globalen Gesundheit	2020
[14]	Horizon Europe Program	2021
[16]	Digitale Gesundheit 2025	2020
[17]	Gesetz für sichere digitale Kommunikation und Anwendungen im Gesundheitssektor	2015
[19]	Roadmap Digitale Gesundheit	2018
[20]	Gesundheits-IT Interoperabilität Governance Verordnung	2021
[21]	Daten helfen heilen	2020
[22]	Datenschutz und IT-Sicherheit im Gesundheitssektor	2022
[23]	Gesetz zur digitalen Modernisierung von Versorgung und Pflege (Digitale-Versorgung-und-Pflege-Modernisierungs-Gesetz-DVPMG)	2022
[24]	Hightech-Strategie 2025	2021
[25]	European Health Data Space	2022

For Tanzania, Digital Health Strategy July 2019-June 2024 identifies responsibilities and organizations for a sustainable governance structure and specified projects for developing guidelines for implementation and creating a legal and regulatory framework for DH under priority 1 [8]. According to priority 6, the aim is to develop analytical tools and indicators with the data and use them for evidence-based interventions and decision-making [8].

Tanzania explicitly mentions strengthening programs for continuous professional development of health workers in the use of data. Priority 6 is to include aspects of data use in health care worker educational and professional development curricula [8]. In addition, there is a focus on networking health professionals by introducing digital platforms such as e-learning (priority 3) [8].

XSL•FO
RenderX

(4) Integrated People-Centered Health Systems Enabled by DH Technologies

The German federal government wants to promote patient autonomy. Implementing a national electronic patient file enables patients to retain complete control over their data and retain decision-making authority over their medical records [30]. To promote accessibility to health tools, a national health portal was created [31]. In addition, the German government wants to promote personalized treatment approaches in all important disease categories and foster early cooperation between stakeholders from science, industry, regulatory authorities, and the medical and patient communities [32]. The *Digital Care and Nursing Modernisation Act* stipulates that, in the future, more DiGAs, digital nursing applications, and telemedical applications should support doctors and nurses and help them perform their tasks more efficiently [23]. To enable the government to monitor the current state of digitization, the gematik (gematik GmbH) has developed a dashboard of key performance indicators that tracks the TI in Germany. Further, gematik is the public entity tasked with developing and maintaining the TI in the country [33]. Germany has not adopted formal strategic approaches to strengthening gender equality and inclusion in the context of digitalization.

Tanzania's DHS focuses heavily on using digital technologies and improving DH competencies. The aim of priority 3 is to provide specialized care to underserved facilities using digital technologies and train health workers with the appropriate skills [8]. The objective is to improve health care facility processes through digital solutions, such as electronic referrals, which also helps to relieve the demand on staff resources (priority 2) [8]. Health information should be shared and disseminated using mobile health, short messaging services, mobile apps, and web apps, thus contributing to patient education (priority 4) [8]. For the government to assess and improve the quality of health services efficiently, priority 6 is to introduce digital solutions for monitoring facilities [8]. Priority 8 also describes measures using digital tools to monitor human resources [8]. However, approaches to strengthen gender equality and accessibility for people with disabilities to promote an inclusive digital society are not addressed.

Similarities and Differences in the Strategies

Our evaluation system is based on the degree to which the policy options, measures, and outputs of each country fulfill the respective dimensions of the WHO strategy from today's perspective. The 4 rating options are 0=not present, 1=partly fulfilled, 2=largely fulfilled, and 3=completely fulfilled. Multimedia Appendix 3 illustrates the results. The analysis highlights Tanzania's plans to strengthen national health policy through targeted implementation measures in the strategy. On the other hand, Germany fulfills all the points recommended by the WHO by expanding its cross-standard data infrastructure. While both countries consider a people-centered health system to be highly relevant, we nevertheless rate the current measures as just above average due to, for example, the lack of gender equality and inclusion concepts in the context of digitalization in the strategies. Strengthening governance for DH at global, regional, and national levels is only partially met in both strategies. In particular, the development of leadership for informed decision-making and the expansion of the strategy to a global perspective limit the fulfillment of the dimension. There is an overall similarity between both countries in terms of their overall progress on their respective DHS. Individual different focus points have therefore already been identified.

In the following, the extracted data were summarized categorically and regrouped into *policy*, *infrastructural*, and *human factors*.

Policy Factors

The WHO GSDH includes a vision, strategic goals, a framework for action, and principles for implementation to advance DH globally and at the state level. Until March 2023, Germany did not develop a national digitization strategy. Priority topics can be found in political documents such as Digital Health 2025. Although most elements of the WHO are mentioned, various aspects are weighted differently. Germany focuses on overcoming sectoral care boundaries and developing innovative technologies such as AI.

While Germany uses national legislation to create a comprehensive legal framework aligned with its national health system, Tanzania has already established governance structures and is now focusing on building the capacity of new members and stakeholders at lower levels. Tanzania's national strategy is based on a vision, associated goals, and clearly articulated priorities. Among these strategic priorities are all 4 main key points identified in the WHO GSDH. Tanzania also complements the WHO GSDH and strives, among other things, to improve supply chain management of health commodities and to improve human resource management at all levels of the health system. Compared to Germany, Tanzania's strategy details developing a change management plan.

Infrastructure Factors

The WHO aims to promote international cooperation by intensifying knowledge transfer among member states. As part of the European Union, Germany pools its resources for international projects and programs and supports DH globally. One of the goals is to create a research-compatible data infrastructure to strengthen the interoperability of systems following international data exchange standards. In Germany, the gematik dashboard is used as an assessment tool to make an initial assessment of the maturity of digital solutions. As part of the European Union, Germany has an overarching basic regulation in the form of the General Data Protection Regulation [33].

Tanzania has also developed the *Tanzania Health Enterprise Architecture*, an approach developed to simplify the complexity of health information systems, guide the development of DH solutions, and facilitate system interoperability [26]. The Tanzanian government highly prioritizes safeguarding the security of sensitive personal data, such as medical information, and is dedicated to improving the legal and regulatory framework to ensure data security, confidentiality, and privacy protection.

XSL•FO
RenderX

Human Factors

Both Tanzania and Germany aim to place patients at the center of their health care systems but take different approaches to achieve this end. Germany actively strives to integrate the patient into health care delivery processes. Germany also aims to intensify research into personalized medicine. Digital solutions such as DiGAs or digital nursing applications (digital applications to support nursing care) focus on providing relief to medical and nursing staff, for example, by supporting labor-intensive tasks such as medical documentation.

Tanzania plans to use client-centric technologies to respond to clients' needs through user-centered design to ensure a responsive, resilient, and inclusive health system. In terms of health care staff, Tanzania is investing in education and training to improve the digital skills of its health care system staff. Only Tanzania's DHS mentions the need to promote health care management staff briefly, and neither Tanzania's nor Germany's DHS defines such steps. Similarly, neither strategy addresses strengthening gender equality and inclusion in digitalization.

Incorporation of New Technologies

New technologies and developments in the health care system can be used to overcome care bottlenecks by increasing efficiency and effectiveness and thus reduce overuse, underuse, and misuse of health care [34].

The WHO recommends introducing sustainable financing models to benefit from the opportunities, realize the full potential of these innovations, and support the exchange of knowledge [4].

Tanzania

To strengthen public health services, the government of Tanzania has focused on developing and expanding information and communication technologies (ICTs). However, further investment is needed to harness the ICT infrastructure for effective data systems such as AI. The goal is to use the huge amounts of data already being collected today effectively and productively. In collaboration with development partners, the Tanzanian government is looking for ways to increase investment in and use of health data systems [27].

In a scoping review conducted by Sukums et al [35] in 2021, a total of 16 publications were identified to have explored the use of AI-driven solutions in Tanzania's health sector. The review called for Tanzania to establish a national AI policy and a regulatory framework for adopting reliable AI solutions in the health sector in line with the WHO guidance on ethics and governance of AI for health.

Germany

The German federal government's *Hightech-Strategy 2025*, creates a basis for investing in innovation by providing incentives for investment [24]. The federal government thus supports local health structures with direct funding and promotes synergies with the private sector and health promotion and disease prevention approaches.

The Federal Ministry of Education and Research has allocated around 250 million euros from 2018 to 2025 for therapy and care concepts implementing AI [36]. The Innovation Committee of the Joint Federal Committee also continuously promotes new technologies used in innovative forms of health care provision [37]. There are many different models for funding new technologies in Germany, many of which involve limited and temporary funding. A sustainable strategy or roadmap for developing and implementing innovative solutions is lacking.

Global Digital Health Monitor

We also retrieved the assessments for Tanzania and Germany from the Global Digital Health Monitor [38]. Tanzania received a rating in terms of an overall DH phase of 4 (out of 5, with 1 being the lowest phase and 5 the highest) and Germany received a 5. For Germany, 4 out of 7 categories were not assessable. Scores for all categories are displayed in Table 3.

Table 3. Ratings on the Global Digital Health Monitor for Tanzania and Germany.

Characteristics	Tanzania	Germany
Leadership and governance	5	5
Strategy and investment	4	N/A[a]
Legislation, policy, and compliance	4	5
Workforce	2	N/A
Standards and interoperability	5	N/A
Infrastructure	4	5
Service and applications	4	N/A
Overall digital health phase	4	5

[a]N/A: not applicable.

Discussion

Principal Findings

Both Germany's and Tanzania's DHS include most of the 35 elements from the WHO's GSDH. While both DHSs fail to include 5 of the 35 elements, they each include additional aspects not included in the GSDH that relate to their own country-specific health care system challenges. The Tanzanian DHS emphasizes (1) quality aspects and data use, (2) digital solutions for supply chains and resource optimization, and (3) human resources. These aspects are oriented toward national

XSL•FO
RenderX

health priorities at all health system levels. In contrast, the German DHS focuses mainly on overcoming sectoral boundaries. Overall, both countries consider context in formulating goals and priorities. Both countries also receive high ratings on the Global Digital Health Monitor, with an overall DH phase rating of 4 out of 5 for Tanzania and 5 out of 5 for Germany. Tanzania has already developed its digitalization strategy for the health sector and identified tangible goals, priorities, and measures. In contrast, Germany does not have a unified strategy and is trying to cover a large distance by taking small steps. All the policies from Germany included in this analysis are from the federal level, meaning that even though Germany is a federal republic, the investigated strategies apply to the whole country.

Tanzania mainly uses new technologies to strengthen its health care system and improve the health and quality of life of its population through intensive health education of the population. Further, one requirement to achieving these goals is to move toward universal access to fiber optic cable and mobile. In recent years, the Tanzanian government has intensified the expansion of ICTs and focused intently on expanding the use of mobile health and telemedicine. It also recognizes the enormous potential of processing large data and has entered public-private funding partnerships to accelerate data infrastructure development and drive innovation. However, as evidenced by the not entirely successful use of machine learning to predict and stem cholera outbreaks, such practices still have weaknesses, and there is a lack of experts skilled in securely collecting, processing, and interpreting large data [35]. In its latest DHS, Tanzania commits to deepening the competencies of health care workers by intensifying e-learning, training, and new curricula. The Tanzanian government is actively trying to improve the health care situation in its country and is looking into sustainable financing models. Addressing the cost implications of DH is a key factor for future DHS development. Germany, meanwhile, is involved in helping to shape global health priority-setting and policy making. Its national ministries are promoting financing models for new technologies and innovations from the public sector and initiatives from the private sector. The field of AI has enormous potential for optimizing processes and improving medicines and other treatments. Developing countries like Tanzania are critical partners for research, innovation, and development, and their progress is crucial for strengthening global health.

A main recommendation for Germany based on the presented analysis would have been to develop a unified DHS. Given that the FMoH published its first digitization strategy in March 2023 [9], it is advisable to review the previous results and compare them with the goals and measures actually achieved. Developing, implementing, and assessing a DHS requires human capacity, expert knowledge of the current health care system challenges and deployments as well as the legal, social, and ethical framework are needed.

In addition, we recommend the development of step-by-step guidelines and a digital tool for the assessment of national DHSs and their consistency with the WHO's GSDH, including a recommended benchmarking with a digital maturity model. An important aspect of the guidelines needs to be the ability to capture context-specific elements and adaptations of national DHSs. Having such an assessment tool will enable the comparison of DHSs between countries and provide an overview of the usage of the WHO GSDH as it is a very valuable blueprint for national DHSs.

Limitations

This study has several limitations. First of all, there is a temporal limitation. DHSs evolve and are updated frequently. The Tanzanian DHS was published 1 year before the WHO GSDH, while different parts of the German strategies were published over several years. Therefore, this work is the state of DHSs in the 2 case study countries as of February 2023, when the analysis was conducted. For instance, the new German Digitalisation Strategy for Health and Care, which was released in March 2023, is not included yet. In addition, there are methodical limitations associated with the policy analysis approach that solely rely on the analysis of policy documents. This may have led to certain aspects not being included in the analysis or being underreported. This applies especially to the case study of Germany with its fragmented strategy across a large number of disparate documents. The level of granularity of the current analysis could be another limitation. Certain aspects of the analysis may have been neglected due to this level of granularity. A more comprehensive evaluation with interviews or surveys with stakeholders could help to provide a deeper understanding.

Conclusion

The WHO's GSDH is a valuable blueprint for developing DHSs. Both Tanzania and Germany have developed strategic guidelines aligned with their own national health care priorities. A federal governmental structure, such as in Germany, makes implementing a national DHS more challenging, often leading to many different strategic approaches and priorities. The extremely rapid development and advancement of emerging technologies is a challenge when their development outpaces the speed at which strategies are adapted and implemented, potentially leading to uncoordinated development. Countries need to develop broad DHSs that guide the future development of the DH ecosystem. These strategies need to include frameworks to support the implementation of new technologies to ensure that these technologies are strategically aligned and revised regularly to ensure alignment with new developments and needs.

For the policy assessment of DHSs, in-depth knowledge of the respective country, its health care needs, and health care system structure is needed. Additional data collection, for example, through interviews and surveys with stakeholders, is needed in addition to document reviews to conduct a holistic assessment of a country's DHS.

XSL•FO
RenderX

Acknowledgments

The authors thank Prof Dr Zahra Meidanii for her input on the idea of this study. This study would not have been possible without the support of Neu-Ulm University of Applied Sciences.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed breakdown of the results of the policy comparison by aspect.
[DOCX File , 42 KB - jmir_v26i1e52150_app1.docx]

Multimedia Appendix 2
Overview of the policy documents mentioned in the methods section with web links.
[DOCX File , 18 KB - jmir_v26i1e52150_app2.docx]

Multimedia Appendix 3
Evaluation of the strategies of Tanzania and Germany in term of their fulfillment of the 4 dimensions of WHO strategy.
[PNG File , 89 KB - jmir_v26i1e52150_app3.png]

References

1. Frank SR. Digital health care--the convergence of health care and the Internet. J Ambul Care Manage 2000;23(2):8-17. [doi: 10.1097/00004479-200004000-00003] [Medline: 10848396]
2. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon MP. Methods to evaluate the effects of internet-based digital health interventions for citizens: systematic review of reviews. J Med Internet Res 2018;20(6):e10202 [FREE Full text] [doi: 10.2196/10202] [Medline: 29880470]
3. Stroetmann KA. From siloed applications to national digital health ecosystems: a strategic perspective for African countries. Stud Health Technol Inform 2019;257:404-412. [doi: 10.3233/978-1-61499-951-5-404] [Medline: 30741231]
4. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://tinyurl.com/58d488cu [accessed 2024-02-21]
5. Mariano B. Towards a global strategy on digital health. Bull World Health Organ 2020;98(4):231-231A [FREE Full text] [doi: 10.2471/BLT.20.253955] [Medline: 32284641]
6. Mbunge E, Muchemwa B, Jiyane S, Batani J. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. Glob Health J 2021;5(4):169-177 [FREE Full text] [doi: 10.1016/j.glohj.2021.11.008]
7. Tanzania national eHealth strategy 2013-2018. The United Republic of Tanzania, Ministry of Health and Social Welfare. 2013. URL: https://tinyurl.com/ykr823up [accessed 2024-02-07]
8. Digital health strategy July 2019-June 2024. Ministry of Health, Community Development, Gender, Elderly and Children. 2019. URL: https://tinyurl.com/hsnc4wdn [accessed 2024-02-06]
9. DIGITAL TOGETHER Germany's Digitalisation Strategy for Health and Care. German Federal Ministry of Health. 2023. URL: https://tinyurl.com/4se9vynp [accessed 2024-02-21]
10. Bowen GA. Document analysis as a qualitative research method. Qual Res J 2009;9(2):27-40 [FREE Full text] [doi: 10.3316/qrj0902027]
11. Peters BG, Fontaine G, editors. Handbook of Research Methods and Applications in Comparative Policy Analysis. UK: Edward Elgar Publishing; 2020.
12. Mayring P. Combination and integration of qualitative and quantitative analysis. Forum Qual Sozialforschung Forum Qual Soc Res 2001;2(1):6 [FREE Full text]
13. Strategie der bundesregierung zur globalen gesundheit. Bundesministerium für Gesundheit. 2020. URL: https://tinyurl.com/mspzuvup [accessed 2024-02-06]
14. Horizont Europa programm 2021. Bundesministerium für Bildung und Forschung. 2021. URL: https://www.horizont-europa.de/de/Programm-1710.html [accessed 2024-02-06]
15. German Alliance for Global Health Research. 2022. URL: https://globalhealth.de [accessed 2024-02-06]
16. Digitale Gesundheit 2025. Bundesministerium für Gesundheit. 2020. URL: https://tinyurl.com/3wcv7fzf [accessed 2024-02-06]

XSL•FO
RenderX

17. Gesetz für sichere digitale Kommunikation und Anwendungen im Gesundheitssektor. Bundesministerium für Gesundheit. 2015. URL: https://www.bundesgesundheitsministerium.de/ [accessed 2024-02-06]
18. Die GMK. Gesundheitsministerkonferenz. 2018. URL: https://www.gmkonline.de/Die-GMK.html [accessed 2024-02-06]
19. Roadmap digitale gesundheit. Bertelsmann Stiftung. 2018. URL: https://tinyurl.com/46znvkbd [accessed 2024-02-06]
20. Gesundheits-IT Interoperabilität Governance Verordnung. In: Bundesministerium für Gesundheit. Verlag: Bundesanzeiger; 2021.
21. Bundesregierung. Daten helfen heilen. Germany: Federal Ministry for Education and Research; 2020. URL: https://tinyurl.com/3pvzjzns [accessed 2024-02-07]
22. Arbeitsgruppe: Datenschutz und IT-Sicherheit im Gesundheitswesen (DIG). GMDS e.V. 2022. URL: https://www.gesundheitsdatenschutz.org/download/einwilligung_2021.pdf [accessed 2024-02-06]
23. Entwurf eines gesetzes zur digitalen Modernisierung von Versorgung und Pflege (Digitale-Versorgung-und-Pflege-Modernisierungs-Gesetz – DVPMG). Bundesministerium für Gesundheit. 2022. URL: https://tinyurl.com/mryefka6 [accessed 2024-02-06]
24. Hightech-Strategie 2025. Bundesministerium für Bildung und Forschung. 2021. URL: https://tinyurl.com/m5kcuua [accessed 2024-02-06]
25. Directorate-General for Health and Food Safety. Proposal for a regulation - The European Health Data Space. European Commission. 2022. URL: https://tinyurl.com/mfrst9j4 [accessed 2024-02-06]
26. Tanzania Health Enterprise Architecture (TZHEA). Ministry of Health, Community Development, Gender, Elderly and Children. 2019. URL: http://api-hidl.afya.go.tz/uploads/library-documents/1617092274-3I9GuToO.pdf [accessed 2024-02-06]
27. Tanzania digital health investment road map 2017-2023. Ministry of Health, Community Development, Gender, Elderly and Children. 2017. URL: http://api-hidl.afya.go.tz/uploads/library-documents/1573688147-xrkVuNtD.pdf [accessed 2024-02-06]
28. Mremi IR, Rumisha SF, Sindato C, Kimera SI, Mboera LEG. Comparative assessment of the human and animal health surveillance systems in Tanzania: opportunities for an integrated one health surveillance platform. Glob Public Health 2023;18(1):2110921 [FREE Full text] [doi: 10.1080/17441692.2022.2110921] [Medline: 35951768]
29. Das DiGA-verzeichnis. Antworten zur nutzung von DiGA. Bundesinstitut für Arzneimittel und Medizinprodukte. 2022. URL: https://diga.bfarm.de/de [accessed 2024-02-06]
30. ePA, Persönliche daten, persönliche entscheidungen. Gematik. 2022. URL: https://www.gematik.de/anwendungen/e-patientenakte [accessed 2024-02-06]
31. Gesundheit und digitalisierung. Gesund.bund.de. 2022. URL: https://tinyurl.com/4n9be32u [accessed 2024-02-06]
32. Personalisierte medizin. Bundesministerium für Bildung und Forschung. 2022. URL: https://www.gesundheitsforschung-bmbf.de/de/personalisierte-medizin-9459.php [accessed 2024-02-06]
33. Gematik. TI Dashboard, Digitalisierung im Überblick. URL: https://www.gematik.de/telematikinfrastruktur/ti-dashboard [accessed 2024-02-06]
34. Künstliche intelligenz und assistenzroboter: neue technologien im gesundheitswesen stärker nutzen. Zukunftsrat der Bayerischen Wirtschaft. 2022. URL: https://tinyurl.com/4htpu6xp [accessed 2024-02-06]
35. Sukums F, Mzurikwao D, Sabas D, Chaula R, Mbuke J, Kabika T, et al. The use of artificial intelligence-based innovations in the health sector in Tanzania: a scoping review. Health Policy Technol 2023;12(1):100728 [FREE Full text] [doi: 10.1016/j.hlpt.2023.100728]
36. Karliczek: mit künstlicher intelligenz das gesundheitssystem besser auf künftige krisen vorbereiten. Bundesministerium für Bildung und Forschung. 2021. URL: https://tinyurl.com/2f8d6jb8 [accessed 2024-02-06]
37. Innovationsfonds. Gemeinsamer Bundesausschuss. 2022. URL: https://innovationsfonds.g-ba.de [accessed 2024-02-06]
38. Global Digital Health Monitor. 2023. URL: https://digitalhealthmonitor.org/ [accessed 2024-02-06]

Abbreviations

AI: artificial intelligence
DH: digital health
DHS: digital health strategy
DiGA: digital health application (digitale gesundheitsanwendungen)
FMoH: Federal Ministry of Health
GSDH: Global Strategy on Digital Health 2020-2025
ICT: information and communication technology
TI: telematics infrastructure
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.08.23; peer-reviewed by N O'Brien, E Vashishtha, H Durrani; comments to author 02.11.23; revised version received 28.11.23; accepted 31.01.24; published 18.03.24.

Please cite as:
Holl F, Kircher J, Hertelendy AJ, Sukums F, Swoboda W
Tanzania's and Germany's Digital Health Strategies and Their Consistency With the World Health Organization's Global Strategy on Digital Health 2020-2025: Comparative Policy Analysis
J Med Internet Res 2024;26:e52150
URL: https://www.jmir.org/2024/1/e52150
doi:10.2196/52150
PMID:38498021

©Felix Holl, Jennifer Kircher, Attila J Hertelendy, Felix Sukums, Walter Swoboda. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Self-Management Using eHealth Technologies for Liver Transplant Recipients: Scoping Review

Soo Hyun Kim[1], MSN; Kyoung-A Kim[2], PhD; Sang Hui Chu[3], PhD; Hyunji Kim[4], MSN; Dong Jin Joo[5], MD; Jae Geun Lee[5], MD; JiYeon Choi[6], PhD

[1]School of Nursing, Johns Hopkins University, Baltimore, MD, United States
[2]College of Nursing, Suwon Women's University, Suwon, Republic of Korea
[3]College of Nursing, Mo-Im Kim Nursing Research Institute, Yonsei University, Seoul, Republic of Korea
[4]School of Nursing, Yale University, Orange, CT, United States
[5]Department of Surgery, College of Medicine, Yonsei University, Seoul, Republic of Korea
[6]College of Nursing, Mo-Im Kim Nursing Research Institute, Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea

Corresponding Author:
JiYeon Choi, PhD
College of Nursing, Mo-Im Kim Nursing Research Institute, Institute for Innovation in Digital Healthcare, Yonsei University
50-1 Yonsei-ro, Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 3301
Fax: 82 2 2227 8303
Email: jychoi610@yuhs.ac

Abstract

Background: Liver transplantation has become increasingly common as a last-resort treatment for end-stage liver diseases and liver cancer, with continually improving success rates and long-term survival rates. Nevertheless, liver transplant recipients face lifelong challenges in self-management, including immunosuppressant therapy, lifestyle adjustments, and navigating complex health care systems. eHealth technologies hold the potential to aid and optimize self-management outcomes, but their adoption has been slow in this population due to the complexity of post–liver transplant management.

Objective: This study aims to examine the use of eHealth technologies in supporting self-management for liver transplant recipients and identify their benefits and challenges to suggest areas for further research.

Methods: Following the Arksey and O'Malley methodology for scoping reviews, we conducted a systematic search of 5 electronic databases: PubMed, CINAHL, Embase, PsycINFO, and Web of Science. We included studies that (1) examined or implemented eHealth-based self-management, (2) included liver transplant recipients aged ≥18 years, and (3) were published in a peer-reviewed journal. We excluded studies that (1) were case reports, conference abstracts, editorials, or letters; (2) did not focus on the posttransplantation phase; (3) did not focus on self-management; and (4) did not incorporate the concept of eHealth or used technology solely for data collection. The quality of the selected eHealth interventions was evaluated using (1) the Template for Intervention Description and Replication guidelines and checklist and (2) the 5 core self-management skills identified by Lorig and Holman.

Results: Of 1461 articles, 15 (1.03%) studies were included in the final analysis. Our findings indicate that eHealth-based self-management strategies for adult liver transplant recipients primarily address lifestyle management, medication adherence, and remote monitoring, highlighting a notable gap in alcohol relapse interventions. The studies used diverse technologies, including mobile apps, videoconferencing, and telehealth platforms, but showed limited integration of decision-making or resource use skills essential for comprehensive self-management. The reviewed studies highlighted the potential of eHealth in enhancing individualized health care, but only a few included collaborative features such as 2-way communication or tailored goal setting. While adherence and feasibility were generally high in many interventions, their effectiveness varied due to diverse methodologies and outcome measures.

Conclusions: This scoping review maps the current literature on eHealth-based self-management support for liver transplant recipients, assessing its potential and challenges. Future studies should focus on developing predictive models and personalized eHealth interventions rooted in patient-generated data, incorporating digital human-to-human interactions to effectively address

XSL•FO
RenderX

the complex needs of liver transplant recipients. This review emphasizes the need for future eHealth self-management research to address the digital divide, especially with the aging liver transplant recipient population, and ensure more inclusive studies across diverse ethnicities and regions.

(J Med Internet Res 2024;26:e56664) doi:10.2196/56664

KEYWORDS

liver transplantation; self-management; transplant management; eHealth; eHealth technology; digital health; review; mobile phone; liver transplant; liver disease; scoping review

Introduction

Background

As the last treatment resort for individuals with end-stage liver diseases or liver cancer [1], liver transplantation (LT) has become one of the fastest-growing solid organ transplant procedures worldwide. Since its first case in 1963, LT has evolved into a more viable treatment option for those living with end-stage liver conditions. In the United States, >9000 individuals receive LTs annually [2]. In South Korea, the number of LT cases increased from 1066 in 2010 to 1515 in 2021 [3]. Over the past decades, there has been notable progress in the success of LT surgery and long-term survival rates [4]. In the United States, the 1-year survival rate reached 89%, and the 5-year survival rate is >74%, although variations exist depending on donor types, underlying diagnoses, recipient age, and region [5].

Despite the improving trend of posttransplant survival, optimizing the benefits of LT remains a complex and challenging issue for LT recipients. Research on post-LT outcomes to date has primarily focused on graft function and overall survival, corresponding to rapidly advancing surgical techniques and drug development [6]. Relatively little attention has been paid to promoting posttransplant self-management and its impact on long-term quality of life (QOL) [7]. After transplant, LT recipients must manage risks of complications, such as intestinal adhesion, bleeding, and bile leakages, and maintain a balance between graft failure risks and the side effects of immunosuppressant therapy, which often necessitate frequent dosage changes [8,9]. Additional lifelong challenges include management of common side effects of immunosuppressant therapy, such as hyperlipidemia, high blood pressure, chronic kidney failure, obesity, diabetes, and infection [10,11].

To address these challenges and maximize the benefits of LT, vigilant posttransplant self-management is crucial. Self-management has been defined in the literature as a comprehensive process that encompasses focusing on one's illness needs (eg, acquiring knowledge and skills, monitoring symptoms, problem-solving, and decision-making), using available resources, building partnerships with health care providers (HCPs), and integrating illness management into daily life [12-14]. For LT recipients, major self-management issues include symptom monitoring, medication management, and engaging in healthy lifestyles after transplant [15-17]. Furthermore, LT recipients must navigate complex health care systems, face changes in social roles, and cope with uncertainty and mental distress associated with the ever-present risk of graft rejection [18]. During the COVID-19 pandemic, with reduced human contact support and the strained health care systems reallocating resources, self-management became more challenging for immunocompromised individuals such as LT recipients [19]. However, the increased accessibility to the internet and digital devices, coupled with the challenges posed by the pandemic, has rapidly escalated interest in interventions using the internet to promote or manage health (eHealth). These technologies hold the potential to overcome challenges related to resource allocation, geographical accessibility, and health care cost.

Prior Work

eHealth, defined as the use of the internet and communication technologies to deliver and improve health care services [20], has been rapidly expanding to promote self-management in acute and chronic conditions [21-23]. However, the application of eHealth for LT recipients has been relatively slow due to the complexity of LT management, the need for close physical examinations, and the importance of building rapport with HCPs for lifelong posttransplant care [24]. Although there have been studies investigating the application of eHealth among solid organ transplant recipients [25,26], to date, no review has specifically focused on eHealth for self-management support in LT recipients. Therefore, this study aimed to map the current state of the literature on self-management using eHealth technologies for LT recipients and assess their benefits and challenges to suggest areas needing further investigation in the field.

Goal of This Study

This review aimed to examine the current literature on eHealth-based self-management among adult LT recipients and its associated factors by addressing the following questions: (1) what are the characteristics and associated factors of eHealth strategies in the adult LT recipient population? (2) how effective and feasible are eHealth-based self-management interventions after LT? and (3) what are the future potential and challenges of eHealth in facilitating self-management among this population? By mapping the existing literature through this scoping review, we aimed to identify gaps and propose future directions for the development and application of eHealth-based self-management interventions for adult LT recipients.

Methods

Study Design

We conducted a scoping review based on the 6-stage scoping review framework by Arksey and O'Malley [27]. This

XSL•FO
RenderX

methodology was chosen to examine the breadth and depth of knowledge in an emerging field of research. We used the population, concept, and context criteria to devise the research question for this review: adult LT recipients (population), eHealth (concept), and facilitating self-management (context) [28]. Given that previous studies on self-management among LT recipients have primarily concentrated on medication adherence, alcohol recidivism, and healthy lifestyle maintenance [15], our review specifically focused on these areas of self-management.

Databases and Search Terms

The search and screening procedure of this scoping review adhered to the guidelines provided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist. We systematically searched 5 electronic databases: PubMed, CINAHL, Embase, PsycINFO, and Web of Science. The search terms comprised a combination of the following: *adult LTR*, *e-Health*, and *self-management*. Our concept of eHealth encompasses a range of technology, including telehealth and internet-, computer-, or mobile-based health, and strategies using video, audio, SMS text messaging, wearables, and virtual reality. To conduct the search on self-management, we included and modified terms such as *alcohol*, *nutrition*, *exercise*, *physical activity*, *medication*, *medication adherence*, *self-care*, *self-management*, and *health behavior*. Our search spanned inception to June 19, 2023, without any restrictions on specific study design. We validated our search strategies through consultation with a university librarian and present the detailed search strategies in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Articles were included if they were studies that (1) examined or implemented eHealth-based self-management, (2) included LT recipients aged ≥18 years, and (3) were published in a peer-reviewed journal. We excluded studies that (1) were case reports, conference abstracts, editorials, or letters; (2) did not focus on the posttransplantation phase; (3) did not focus on self-management; and (4) did not incorporate the concept of eHealth or solely used technology for the purpose of data collection. We also considered studies that targeted various solid organ transplant recipients provided they included adult LT recipients. Our review specifically targeted adult patients in the post-LT phase as LT in children often involves dissimilar underlying conditions for transplant, such as biliary atresia [29], and post-LT self-management in pediatric patients may present additional challenges related to life stage development.

Article Selection

A total of 1460 articles were identified. After removing duplicates (51/1460, 3.49%), 1409 articles were imported to Microsoft Excel (Microsoft Corp) for screening using the inclusion and exclusion criteria. First, 2 authors (SHK and HK) independently screened the titles and excluded 95.95% (1352/1409) of the articles, which did not include LT recipients or were not about eHealth or self-management (eg, drug trials and surgery). This yielded 57 articles. In the second stage, 2 authors independently reviewed the abstracts of the 57 articles. Finally, a full-text review was conducted on 42% (24/57) of the articles. The references of the selected articles were manually examined to search for additional articles that could be eligible for this review. Consequently, 1 article was added through the manual search. A total of 15 articles were included in the final sample. The whole process was supervised by the corresponding author, and disagreements between authors were discussed in research team meetings until a consensus was reached. Figure 1 illustrates the PRISMA flow diagram.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

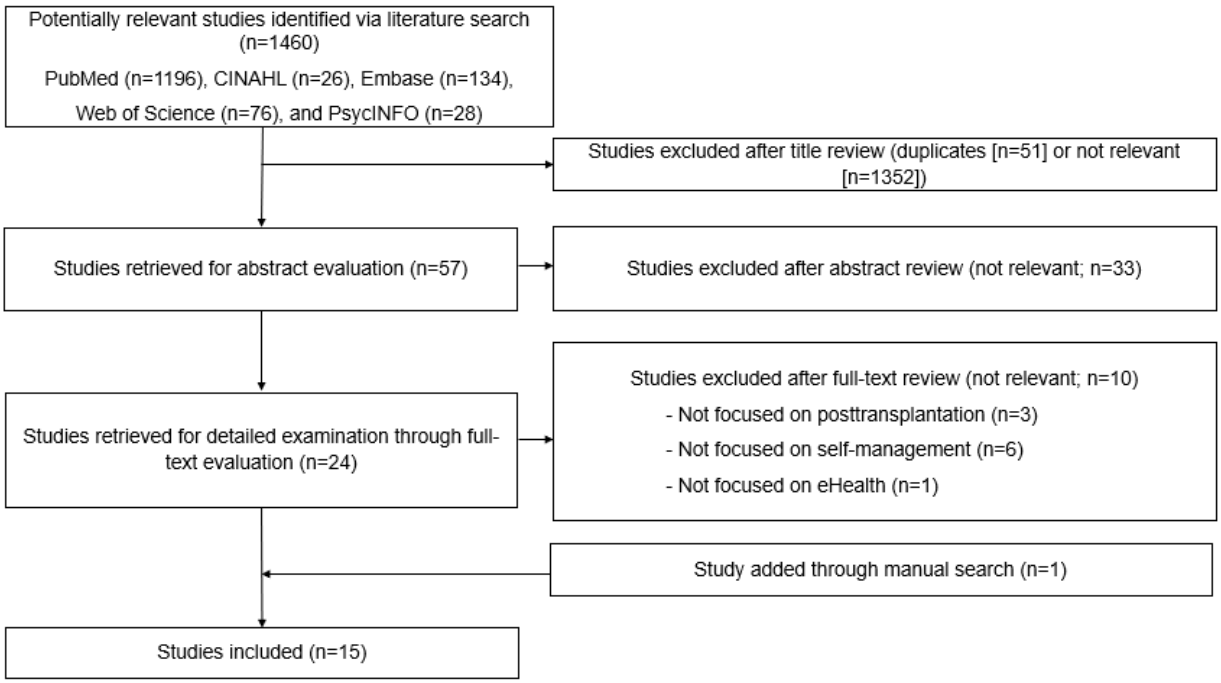

XSL•FO
RenderX

Data Analysis

Due to the heterogeneity of the interventions and measured outcomes, a meta-analysis or meta-synthesis of the results was not feasible. Instead, we presented the organized data in tables, which include the summaries of the descriptive and intervention studies and a detailed summarization of the interventions.

The following information was extracted using data-charting forms: first author's last name, publication year, country, study design, sample characteristics (sample size, average age, percentage of male individuals, and time since transplant), type of technology, variables with measures, and outcomes. For intervention studies, we analyzed the details of the intervention and control groups and extracted the following characteristics of the interventions: duration, providers, adherence, adverse events, and reasons for attrition.

We also evaluated the quality of the interventions reported in these studies using the Template for Intervention Description and Replication (TIDieR) checklist and guide [30]. The TIDieR checklist and guide address the items of interventions that need to be thoroughly described to enhance the replicability of trials and facilitate the appraisal of intervention reporting. To gauge the comprehensiveness of intervention content, we analyzed whether the interventions addressed the 5 core self-management skills identified by Lorig and Holman [14]. These core skills, which include problem-solving, decision-making, using resources, partnering with HCPs, and taking action, have been considered as fundamental elements of self-management in the literature that examines the definitions, components, and processes of self-management [12,13,31,32].

Ethical Considerations

No ethics approval was required for this scoping review as published articles rather than primary data were used in the analysis.

Results

Study Characteristics

The general characteristics of the 5 descriptive studies and 10 intervention studies are summarized in Tables 1-4. Of the 5 descriptive studies, 2 (40%) were qualitative and 3 (60%) were quantitative studies. Of the 10 intervention studies, 4 (40%) were randomized controlled trials (RCTs), 4 (40%) were prospective observational studies with cohorts, 1 (10%) was a study with a historical control group, and 1 (10%) was a single-group qualitative evaluation of a feasibility study. A total of 67% (10/15) of the studies were published in 2020 or later, with all studies (15/15, 100%) being published after 2016. Most studies were conducted in North America (7/15, 47% in the United States and 1/15, 7% in Canada) and Europe (1/15, 7% in Austria; 2/15, 13% in Belgium; and 1/15, 7% in Spain). The sample sizes of the selected studies varied, ranging from 19 to 710 participants, with the average age of the participants ranging from 46 to 63 years.

Table 1. Characteristics of the descriptive studies.

Study	Year	Country	Study design	Sample size	Age (y)	Sex (male; %)	Time since transplant
Lieber et al [33]	2021	United States	Qualitative study	20	Median 61	65	Not reported
Maroney et al [34]	2021	United States	Quantitative survey	178 KTRs[a] and 110 LTRs[b]	Mean 52.6	54.5	46.88% between 1 and 2 years after transplant
Mathur et al [35]	2021	Canada	Qualitative study	5 LTRs out of 21 SOTs[c]	Mean 47	48	2 years to 8 years
Vanhoof et al [36]	2018	Belgium	Cross-sectional descriptive study	30 LTRs out of 122 SOTs	Mean 55.9	57.4	Median 6 years
Wedd et al [37]	2019	United States	Cross-sectional descriptive study	455 KTRs and 255 LTRs	Median 49 (KTRs) and 53 (LTRs)	55.2 (KTRs) and 59.6 (LTRs)	6 months before transplant to 2 years after transplant (study period)

[a]KTR: kidney transplant recipient.

[b]LTR: liver transplant recipient.

[c]SOT: solid organ transplant.

XSL•FO
RenderX

Table 2. Results of the descriptive studies.

Study	Type of technology	Variable (measure)	Outcome
Lieber et al [33]	App	• Smartphone use • Acceptability and preference regarding smartphone app: thematic analysis of semistructured interviews	• 90% of participants used a smartphone. • The most frequently cited purpose of smartphone use for posttransplant recovery was setting alarm reminders to take medications. • 80% of participants were interested in using an app to connect with peer LT[a] recipients and HCPs[b], gain educational information and support medication taking, and log biometric data.
Maroney et al [34]	Patient portal (internet or smartphone)	• eHealth literacy: eHEALS[c] • Routine patient portal use (defined as routine if use frequency was every day, weekly, or monthly)	• eHealth literacy: KTRs[d]>LTRs[e] (P=.04) • eHealth literacy: college education or higher>high school education or lower (P<.001) • eHealth literacy was higher in those with mobile internet access than in those without (P=.04) • Routine users had higher eHealth literacy than nonroutine users and nonusers (eHEALS score 31.97 vs 29.97 vs 28.20; P<.001).
Mathur et al [35]	Digital health tools	• Barriers and motivators to PA[f] • Core features of a digital health tool to support PA: thematic analysis of semistructured interviews	• Barriers to PA included risk aversion, managing nonlinear health trajectories, physical limitations, and lack of access to appropriate PA resources for SOT[g] population. • Facilitators of PA included desire to live healthy lives and honor the donors, renewed physical abilities, and access to appropriate fitness guidelines and facilities. • Desired features included safe PA guidelines, a reward system, affordability, integration of multiple functions specifically designed for after SOT, and the ability to share information with HCPs and peer recipients.
Vanhoof et al [36]	IHT[h]	• Use of ICTs[i] • Willingness to use IHT as self-management support • Technology acceptance • Preferences for specific IHT features: 35-item newly designed interview questionnaire	• Only 27.9% of participants had a smartphone, whereas 72.1% owned a computer with internet access. • Participants gave a median score of 7 on a 10-point scale on whether they thought IHT was important to support their self-management. • Technology acceptance: Patients who were single, married, or living together>patients who were widowed or divorced; Patients with high school education or lower>patients with college education or higher; Patients with previous technology use>patients without previous technology use • Patients preferred automatic data transfer over sending data by themselves and feedback via visual aids over SMS text messages.
Wedd et al [37]	Patient portal (internet)	• Use of web portal (whether a patient had any recorded activity based on Cerner server logs at any point during the study period)	• 48.2% of LTR participants used the patient portal • LTR portal use: White>Black (P=.003); college education or higher>high school education or lower (P<.001) • The patient portal was most frequently used for viewing laboratory test results by KTRs and LTRs (43.9% and 37%, respectively).

[a]LT: liver transplant.

[b]HCP: health care provider.

[c]eHEALS: eHealth Literacy Scale.

[d]KTR: kidney transplant recipient.

[e]LTR: liver transplant recipient.

[f]PA: physical activity.

[g]SOT: solid organ transplant.

[h]IHT: interactive health technology.

[i]ICT: information and communications technology.

XSL•FO
RenderX

Table 3. Characteristics of the intervention studies.

Study	Year	Country	Study design	Sample size, n	Age (y)	Sex (male; %)	Time since transplant
Barnett et al [38]	2021	Australia	Qualitative evaluation of a feasibility study	19	Mean 52	63	Median 4.4 years
Hickman et al [39]	2021	Australia	RCT[a] with delayed intervention control	IG[b]: 23; CG[c]: 12	IG: mean 51; CG: mean 50	IG: 65; CG: 83	Median 4 years (IG) and 3 years (CG)
Ertel et al [40]	2016	United States	Prospective observational study	IG: 20; CG: 20	Mean 56	80	Not reported
Koc et al [41]	2022	Belgium	Prospective cohort study	Autonomous IG: 39; nonautonomous IG: 48; CG: 28	Median 59.1 (autonomous IG); 67.2 (nonautonomous IG); 66.0 (CG)	61.5 (autonomous IG); 60.4 (nonautonomous IG); 50 (CG)	Median 6.4 years (autonomous IG); 7.4 (nonautonomous IG); 7.0 (CG)
Lee et al [42]	2019	United States	RCT	IG: 50; CG: 50	IG: median 60; CG: median 58.5	IG: 52; CG: 60	Not reported
Tian et al [43]	2021	China	RCT	IG: 52; CG: 50	Mean 46.65	70.6	Not reported
Andrä et al [44]	2022	Austria	Prospective cohort study	124 (IG: 42)	Mean 63.2 (IG: mean 55.4)	74.1	Mean 6.5 years
Melilli et al [45]	2021	Spain	Prospective observational trial	84 KTRs[d] and 6 LTRs[e]	Mean 46	73	Mean 69 months
Serper et al [46]	2020	United States	RCT	61 KTRs and 66 LTRs	Mean 52	64	Median 9.5 months
Zanetti-Yabur et al [47]	2017	United States	Prospective cohort study	67 KTRs and 7 LTRs (IG: 21; CG: 53)	IG: mean 52.6; CG: mean 54.1	60.8	Not reported

[a]RCT: randomized controlled trial.

[b]IG: intervention group.

[c]CG: control group.

[d]KTR: kidney transplant recipient.

[e]LTR: liver transplant recipient.

XSL•FO
RenderX

Table 4. Results of the intervention studies.

Study	Intervention	Control	Measurement timeline	Variable (measure)	Outcome
Barnett et al [38]	Telehealth-delivered diet and exercise program	__[a]	After the intervention (at 12 weeks)	• End-user experiences • Perspectives • Program feasibility: thematic analysis of semistructured interviews	• Telehealth reduced the perceived pressure and burdens of face-to-face care. • Due to employment, participants desired flexible access to the program. • Mediterranean diet was well accepted through enhancing confidence, finding practical solutions, and integration into family meals. • Through tailored exercise prescriptions, participants experienced increase in exercise confidence, awareness of their exercise capabilities, and the ability to self-direct and prioritize exercise routine.
Hickman et al [39]	Telehealth-delivered lifestyle intervention	Delayed intervention control (12-24 weeks)	• At baseline • After the intervention (at 12 weeks in IG[b] and 24 weeks in CG[c])	• Feasibility: recruitment rate, attendance rate, attrition rate, self-reported adequacy, adverse events, and perception of safety • Dietary adherence: MEDAS[d] • QOL[e]: SF-12[f] • Metabolic syndrome: MetSSS[g]	• Recruitment rate: 26% • Attendance rate: 60% (diet: 71%; exercise: 52%) • Attrition rate: 22.9% • Staff-reported adequacy: 90% of sessions • Participant-reported confidence of ≥8/10: 96% of sessions • Participant-reported adequacy: 91% of sessions • Adverse events and perception of unsafeness: none • Increase in MEDAS: IG>CG (P=.004) • MCS-12[h]: IG>CG (P=.03); increase in IG after the intervention (P=.03) • Decrease in MetSSS after the intervention from before the intervention in the IG (P=.01)
Ertel et al [40]	Educational video program and telehealth monitoring of vital statistics using Bluetooth peripheral	Standard care (historic control)	• At 30 days • At 90 days	• 30- and 90-day readmission rate • Effectiveness: 9 binary questionnaires • Satisfaction: six 5-point Likert-scale questionnaires	• 30-day readmission rate: 20% (IG) and 40% (CG) • 90-day readmission rate: 30% (IG) and 45% (CG; statistical analysis not conducted due to small sample size) • Report of video program being helpful: 100% (response rate: 95%) • Mean composite satisfaction score: 29/30 (IG and CG)

XSL•FO
RenderX

Study	Intervention	Control	Measurement timeline	Variable (measure)	Outcome
Koc et al [41]	Telemedicine-based remote monitoring program	Standard follow-up	• At baseline • After the intervention (median follow-up 2.0 years in the autonomous IG, 2.1 years in the nonautonomous IG, and 2.4 years in the CG)	• Feasibility: initiation and continuation rate of autonomous CDSS[i], and satisfactory score • Safety: outpatient visits, emergency visits, unanticipated hospitalizations, tacrolimus level determinations, tacrolimus concentration, liver graft rejections, and retransplantations	• 75.7% of eligible patients initiated the program. • 44.8% in the IG continued to use the CDSS (autonomous IG group) • Median satisfactory score: 4/5 in IG • Number of outpatient visits at follow-up: IG<CG (*P*<.001) • Number of hospitalizations, rejections, or retransplantations did not differ between groups. • Tacrolimus level determinations: IG>CG (*P*<.001 and *P*=.003 for the autonomous and nonautonomous IG, respectively) • Tacrolimus blood level concentrations could be maintained lower after intervention than before in the IG (*P*=.04 and *P*=.002 for the autonomous and nonautonomous IG, respectively).
Lee et al [42]	THMP[j]	Standard care	• At discharge • At 30 days • At 90 days	• 90-day readmission rate • QOL: SF-36[k] • THMP participation: frequency	• 90-day readmission rate: IG<CG (28% vs 58%; *P*=.004) • SF-36—physical functioning increase in IG (*P*=.02); general health increase in IG (*P*=.05) • Frequency of vital sign monitoring and use of the devices to input the data was 86%.
Tian et al [43]	Telemedicine-based follow-up management	Usual care	• At baseline • At 30 days • At 12 months	• Hospital use • 30-day readmission rate • Survival rate, mortality, and morbidity	• Length and expense of initial hospitalization were lower in IG than CG (*P*=.03 and *P*=.049, respectively). • 30-day readmission rate: IG<CG (*P*=.02) • Cumulative survival rate: no substantial difference but a 2-year stable period after transplant in IG • Survival rates at 12 months: 94.2% (IG) and 90% (CG; *P*=.65) • Occurrence of complications: no significant difference
Andrä et al [44]	Medication tracking and healthy lifestyle management app	—	• At baseline • After the intervention (at 2 months)		

XSL•FO
RenderX

Study	Intervention	Control	Measurement timeline	Variable (measure)	Outcome
				• Basic requirements and knowledge about the app • Potential benefits and expectations regarding the app • Usability of and satisfaction with the app	• 57.3% owned a mobile device. • 14.5% had previously used a health care app and had a younger mean age than that of the overall cohort. • 31.8% of those requiring medication reminders relied on their spouses and were older than those using electronic alarms. • 86.3% expected to obtain helpful information from the app. • The individuals who tried the app (33.8%; IG) had a younger mean age than that of the overall cohort. • 47.6% of app users used all its functions. • 66.7% found the app to be adequate. • 57.1% used the app daily for 2 months. • The most frequently used functions were information access, reminders, and medication management.
Melilli et al [45]	App to monitor immunosuppressant adherence, classified into regular users, random users, and nonusers	—	• At baseline • At 3 months • At 6 months • At 1 year	• Engagement with the app: willingness to use and continuation rate • Medication nonadherence: number of missing drug intakes or doses taken out of time and nonadherent profiles according to distinct clinical characteristics	• 68% were regular users up to 6 months after the intervention. • Of the regular users, 59% remained active up to 1 year after the intervention. • Correct dose intakes (regular users): 69%-76% • Intakes out of time (regular users): 12%-19% • Missed doses (regular users): 9%-12% • 4% of patients with high out-of-time intakes were LTRs[l]. • Intakes out of time of >20% was the only variable independently predicting a tacrolimus CV[m] of >30%. • No differences in adherence levels were observed in relation to the type of immunosuppression formulations or combinations as well as the main clinical and demographic patient characteristics.

XSL•FO
RenderX

Study	Intervention	Control	Measurement timeline	Variable (measure)	Outcome
Serper et al [46]	Home-based exercise program using wearable devices, health engagement questions, and loss-framed financial incentives	CG arm 1: standard instructions regarding healthy diet and physical activity; CG arm 2: accelerometer without financial incentives	• At baseline • Daily (steps) • At 4 months	• Weight change • Proportion of participant days in which ≥7000 steps were achieved	• At 3 months, there were no significant differences in weight change across the 3 arms. • 62% of the IG achieved ≥7000 steps compared to 45% of the CG arm 2 (*P*<.001). • LTRs were associated with lower likelihood of achieving ≥7000 steps (*P*=.001).
Zanetti-Yabur et al [47]	Mobile app with medication-taking alarm system	Non–app users	• At 3 months	• Medication beliefs: BMQ[n] • Adherence: MMAS-8[o] • Patients' ability to recall immunosuppressive treatment: IAT[p] • Serum tacrolimus, creatinine, and rejection assessment	• Participants had adverse views about medication. • CG participants were more likely to have negative beliefs about medicine in general compared to IG participants (*P*=.006). • No difference in adherence patterns between IG and CG • IG generally scored higher than CG in IAT, but this was not statistically significant (*P*=.19). • No difference in laboratory test results or rejection

[a]Not applicable.
[b]IG: intervention group.
[c]CG: control group.
[d]MEDAS: Mediterranean Diet Adherence Screener.
[e]QOL: quality of life.
[f]SF-12: 12-item Short Form Health Survey.
[g]MetSSS: Metabolic Syndrome Severity Score.
[h]MCS-12: Mental Component Summary–12 (vitality, social functioning, role limitations due to emotional health, and mental health).
[i]CDSS: clinical decision support system.
[j]THMP: telemedicine-based home management program.
[k]SF-36: 36-item Short Form Health Survey.
[l]LTR: liver transplant recipient.
[m]CV: intrapatient variability.
[n]BMQ: Beliefs About Medicines Questionnaire.
[o]MMAS-8: Morisky Medication Adherence Scale-8.
[p]IAT: immunosuppression assessment test.

Intervention Characteristics

Tables 5 and 6 summarize the details of the interventions reported in the selected studies. The most frequently addressed aspect of self-management was lifestyle education (6/10, 60%). Other topics of the interventions, allowing for overlap, included telemedicine-based remote monitoring (4/10, 40%) and medication management (3/10, 30%). The technological methods to deliver the interventions varied, including mobile apps (4/10, 40%; one of which included a wearable app), a 2-way videoconference portal (3/10, 30%), a web-based platform (1/10, 10%), Bluetooth peripherals (2/10, 20%), and a robot (1/10, 10%). The duration of the interventions ranged from short term (2 weeks after transplantation) to long term (up to 2 years after transplantation), with most (6/10, 60%) lasting 12 weeks.

Among the 6 studies that involved lifestyle interventions, 3 (50%) used synchronous video streaming to deliver exercise or diet educational sessions [38,39,43]. A total of 33% (2/6) of the studies offered mobile apps, with 17% (1/6) of the studies encompassing both lifestyle education and medication management [44]. The latter study provided information on self-management issues and allowed patients to self-document in a patient diary [44]. Another study provided a wearable

XSL•FO
RenderX

accelerometer app, which was paired with financial incentives and questions regarding health engagement available on a patient portal [46]. A total of 40% (4/10) of the studies used telehealth to remotely monitor recipients' vital signs and blood glucose levels [40,42,43] and conduct postoperative management regarding medication, gastrointestinal function, wound care, and laboratory test results [41-43]. For remote monitoring via telehealth, daily vital signs were collected using Bluetooth devices [40,42] or a robot [43].

In terms of medication management, 30% (3/10) of the studies used mobile app interventions that emphasized scheduled immunosuppressant intake [44,45,47]. These interventions used methods such as QR codes, reminder systems, and access to various resources (eg, medication and dose converter, medication lists) to facilitate medication adherence. In addition, 10% (1/10) of the studies, which delivered a telemedicine-based remote program, also monitored laboratory test results using an alarm system of predefined thresholds [41].

XSL•FO
RenderX

Table 5. Key features of the interventions.

Type of technology and study	Focus of self-management	Collaboration	Personalization	Adherence
Videoconference portal				
Barnett et al [38]	Lifestyle management (diet and exercise education)	2-way videoconference portal	Exercise sessions were tailored to individual needs, capabilities, and preference for supervision.	• Median attendance: 10 sessions out of 14
Hickman et al [39]	Lifestyle management (diet and exercise education)	2-way videoconference portal	Participants received up to 3 SMS text messages between sessions based on preference. At the end of the exercise sessions, participants received advice with personal prescriptions	• Attendance rate: 60% (diet: 71%; exercise: 52%)
Telemedicine (website or tablet or mobile app or robot)-based remote monitoring				
Ertel et al [40]	Lifestyle education (posttransplant management) and remote monitoring (vital signs)	Not described	Not described	• 78% attended a listing class with educational information given at that time. • 95% watched all the videos. • 9 patients responded to ≥80% of the daily assessments. • 12 patients recorded ≥80% of the daily vital statistics.
Koc et al [41]	Remote monitoring (laboratory test result management)	Not described	Not described	• Patients in the IG[a] entered 1526 (90.9%) of the 1679 required data items. • 55.2% of the IG were not willing to report data in the CDSS[b] and switched to the nonautonomous group but still communicated with the nurse. • None switched to standard follow-up.
Lee et al [42]	Remote monitoring (vital sign management, posttransplantation education, and communication)	Video communication and phone calls were available.	Not described	• Vital sign monitoring and use of devices: 86% • Response to SMS text messages: 60% (0-30 days) and 25% (31-90 days) • Use of video messaging or FaceTime: 6% • Use of phone calls for issues: 70% • Educational video slides viewed: 75%
Tian et al [43]	Remote monitoring (vital sign and posttransplantation management and communication) and lifestyle management (exercise)	Synchronous and asynchronous communication was available via robot, which was controlled by specialists via computer, phone, or iPad.	Not described	• Not reported
Mobile app				
Andrä et al [44]	Lifestyle management and medication management	No direct connection with physicians	Not described	• 47.6% of app users used all its functions. • 11.9% did not use the app at all. • 57.1% used the app daily for 2 months.

XSL•FO
RenderX

Type of technology and study	Focus of self-management	Collaboration	Personalization	Adherence
Melilli et al [45]	Medication management	Not described	Not described	• 68% were regular users up to 6 months after the intervention. • Among regular users, 59% remained active up to 1 year after the intervention.
Serper et al [46]	Lifestyle management (exercise)	Questions and answers related to health engagement were exchanged through bidirectional SMS text messaging.	Biweekly walking goals were tailored to their baseline based on the mean steps.	• Adherence to target step goals was 74% in the IG. • 84% of health engagement questions were answered, and among those, 95% were answered correctly. • 92.1% (117/127) of participants were retained in the study.
Zanetti-Yabur et al [45]	Medication management	Not described	Not described	• Not reported

[a]IG: intervention group.

[b]CDSS: clinical decision support system.

XSL•FO
RenderX

Table 6. Summary of the interventions.

Type of technology and study	Adverse events	Reasons for attrition	Duration	Provider	TIDieR[a] checklist score (1-12)
Videoconference portal					
Barnett et al [38]	Not reported	—[b]	12 weeks	Dietitians and exercise physiologists	10
Hickman et al [39]	None	• Withdrew, no reason given (n=2) • Lost to follow-up (n=6) • Withdrew in delayed intervention control (n=4)	12 weeks	Dietitian and exercise physiologist	10
Telemedicine (website or tablet or mobile app or robot)-based remote monitoring					
Ertel et al [40]	Not reported	—	Perioperative period education until 90 days after discharge and telemonitoring after discharge	Unspecified health care providers	8
Koc et al [41]	Not reported	—	Median follow-up 2.0 years in the autonomous IG[c], 2.1 years in the nonautonomous IG, and 2.4 years in the CG[d]	Physicians and specialized nurses	9
Lee et al [42]	Not reported	• Withdrew due to recurrent readmissions or death within 90 days after transplant (IG: n=1; CG: n=1)	3 months (daily monitoring)	Nurse care coordinators and providers	10
Tian et al [43]	Not reported	• Death (IG: n=3; CG: n=5)	2 weeks	Transplant specialists	9
Mobile app					
Andrä et al [44]	Not reported	—	2 months	Unspecified health care providers	7
Melilli et al [45]	Not reported	• Lost to follow-up (n=2) • Withdrew, no reason given (n=1)	12 months	Physicians	9
Serper et al [46]	None	• Died before study completion for a reason unrelated to the study (n=1)	2-week run-in period and 16 weeks of intervention	Unspecified health care providers	9
Zanetti-Yabur et al [47]	Not reported	—	3 months	Unspecified health care providers	6

[a]TIDieR: Template for Intervention Description and Replication.

[b]Not applicable.

[c]IG: intervention group.

[d]CG: control group.

Intervention Reporting and Comprehensiveness

On the basis of the TIDieR guide, the included intervention studies' scores ranged from 6 to 10 (Tables 5 and 6). Detailed scores for each criterion can be found in Multimedia Appendix 2 [38-47]. All studies described the content and procedure of interventions, the type of technology used, and the locations where they were implemented. Of the 10 studies, 7 (70%) reported the duration and doses of the interventions, and 6 (60%) included the providers of the interventions. Personalized interventions were delivered to participants in 30% (3/10) of the studies. A total of 30% (3/10) of the studies included strategies or assessments to improve intervention fidelity, whereas 70% (7/10) of the studies monitored the adherence and fidelity of the delivered interventions. No intervention study reported any unforeseen changes in the interventions during the study process.

Table 7 shows how the selected interventions addressed the 5 core self-management skills identified by Lorig and Holman [14]. Of the 10 interventions, 4 (40%) included training in all 5 self-management skills. The number of core skills included in other studies ranged from 1 to 3. Taking action was a component in all interventions (10/10, 100%). Partnering with health care providers was featured in 80% (8/10) of the interventions, whereas Problem-solving was included in 70% (7/10) of the interventions. Decision-making was part of 60%

XSL•FO
RenderX

(6/10) of the interventions, and Using resources was the least included, addressed in 50% (5/10) of the interventions.

Table 7. Core self-management skills addressed in the selected intervention studies.

Study	Problem-solving	Decision-making	Using resources	Partnering with health care providers	Taking action
Barnett et al [38]	●[a]	●	●	●	●
Hickman et al [39]	●	●	●	●	●
Ertel et al [40]	●	●	●	●	●
Koc et al [41]	○[b]	○	○	●	●
Lee et al [42]	●	●	○	●	●
Tian et al [43]	●	●	○	●	●
Andrä et al [44]	◐[c]	◐	●	◐	●
Melilli et al [45]	●	○	○	●	●
Serper et al [46]	●	●	●	●	●
Zanetti-Yabur et al [47]	○	○	○	○	●

[a]Reported.

[b]Not reported.

[c]Unclear.

Study Outcomes

The use of and preferences regarding eHealth technologies were the most commonly examined outcomes in the descriptive studies. Study results concerning the use of technologies and its correlation with educational level and age varied. In the qualitative study by Lieber et al [33], 90% of the 20-person LT recipient sample reported using smartphones. In contrast, in the study by Vanhoof et al [36], only 27.9% of 122 solid organ transplant recipients owned a smartphone despite >70% having access to a computer with internet connection.

Concerning educational characteristics, 40% (2/5) of the studies found that individuals with college-level education or higher demonstrated greater eHealth literacy and more frequent use of the web patient portal than those with a high school education or lower [34,37]. Conversely, the study by Vanhoof et al [36] indicated that the group with a college education and higher had lower technology acceptance than those with a high school education or lower. Regarding age, younger patients had higher eHealth literacy in one study [34]. Another study found that those with previous experience using health apps, as well as those who tried the new app, had a younger average age than that of the entire cohort [44]. However, age was not a substantial factor in the willingness to use health technologies or patient portals in another 40% (2/5) of the studies [36,37]. Previous or routine use of health technologies was associated with higher eHealth literacy [34] and greater technology acceptance [36].

Moreover, 40% (2/5) of the studies highlighted that patients currently used smartphones or patient portals for reminders to take medications or to view laboratory test results. However, their preferences for future eHealth technologies extended beyond these uses. They expressed interest in connecting through web-based platforms with peer recipients and HCPs and gaining access to additional supportive features that included educational resources, medication management tools, and reward systems that consider affordability [33,35].

The most frequently measured outcomes in the intervention studies were adherence to the intervention and feasibility. A total of 80% (8/10) of the studies reported intervention adherence using various methods, including session attendance, device use frequency, response rate, and monitoring rate [38-42,44-46]. In 62% (5/8) of these studies, the overall adherence rate was >70% despite variability in measures [32,40-42,46]. Feasibility was measured using various methods, including qualitative interviews; recruitment rates; attendance rates; initiation and continuation rates; and reported levels of adequacy, confidence, effectiveness, or satisfaction [38-41]. Although rates of recruitment, attendance, initiation, and continuation varied between studies, the interventions were generally well received. Participant-reported satisfactory scores were >80% in 20% (2/10) of the studies [40,41], and participant-reported adequacy, confidence, and effectiveness levels were >90% in 30% (3/10) of the studies [39-41]. In a qualitative study, participants expressed that perceived burdens of face-to-face care were reduced regarding travel, time, or pressure and that confidence level in exercise increased due to tailored and self-directed telehealth education sessions [38].

A total of 4 studies involving telemedicine-based remote monitoring included 30- or 90-day readmission rates (n=3, 75%) and QOL (n=2, 50%) as outcomes. Regarding readmission rates, all studies reported a decrease in readmission rates after the interventions, 20% (2/10) of which were RCTs that reported a significant reduction in readmission rates [40,42,43]. A total of 20% (2/10) of the studies reported improvement in varying components of QOL, such as mental and physical function and general health [39,42].

Due to the heterogeneity of the interventions, various clinical outcomes were measured, such as dietary or medication

XSL•FO
RenderX

adherence, weight change, level of metabolic syndrome, and tacrolimus drug level. Among the studies that incorporated mobile app interventions for medication management, the study by Melilli et al [45] reported that correct dose intakes on schedule occurred 69% to 76% of the time among the regular users of the delivered mobile app, whereas the study by Zanetti-Yabur et al [47] reported no difference in medication adherence on a validated scale between the intervention and control groups. Positive results on dietary adherence and metabolic syndrome were reported on a telehealth-delivered diet and exercise education program [38,39], but there was no significant change in weight after a home-based exercise program using a wearable accelerometer [46]. In a study incorporating remote monitoring focused on laboratory test results, tacrolimus level determinations were higher and blood level concentrations remained lower in the intervention group [41]. Regarding the reporting of adverse events, most studies (8/10, 80%) did not provide specific descriptions. However, 20% (2/10) of the studies reported no adverse events during the intervention period [39,46].

Discussion

Principal Findings

In this scoping review, we investigated the breadth of application and associated factors of eHealth-based self-management strategies in adult LT recipients. While previous literature on self-management among LT recipients has primarily focused on medication adherence, alcohol abstinence, and health maintenance (eg, smoking cessation, vaccination, and health screening) [15], the eHealth-based self-management studies in our review predominantly covered lifestyle management, medication adherence, and remote monitoring. Given the heterogeneity of the interventions and study measures, we assessed outcomes by describing trends, consistencies, or discrepancies among studies reporting similar outcomes. Although significant effects on reducing readmission rates or improving QOL were observed, synthesizing quantitative outcomes was not feasible due to a small number of RCTs. However, based on outcomes measured in various ways and participants' qualitative reports, the interventions were well received, with generally high levels of feasibility, adherence, and satisfaction. The lifestyle management interventions used various modes of delivery, such as videoconferencing, web-based prerecorded videos, mobile apps, and patient portals. Remote monitoring was facilitated through a range of telehealth platforms, whereas interventions for medication adherence mainly used mobile apps. Notably, our review found that none of the eHealth-based self-management studies addressed the topic of prevention and management of alcohol relapse, a well-known concern among LT recipients [48].

Compared to intervention studies focusing on lifestyle management, the intervention studies focusing on medication adherence included relatively basic features, such as alarm reminders and logs for tracking medication intake. While these features target the action of medication taking, they fall short in promoting other essential skills such as decision-making or effective resource use to improve adherence to immunosuppressants. Given the complexity and variety of self-management behaviors required for LT recipients, eHealth technologies should be designed to support them in navigating multiple concurrent problematic situations to integrate disease management into their daily lives [15]. This objective extends beyond ensuring compliance with self-management behaviors and requires strategies to improve the self-efficacy of LT recipients and foster self-tailoring strategies [14]. This involves integrating their values, preferences, and readiness to increase motivation and confidence while also considering the various challenging scenarios that these LT recipients face, such as managing multiple medications due to comorbidities, coping with side effects, and handling varying schedules [9]. In interventions focused on lifestyle management and remote monitoring, features such as prerecorded or synchronous education sessions and platforms for asking questions and receiving answers were available. Such components could be beneficial in improving immunosuppressant adherence. In addition, incorporating elements such as role-play and quizzes that reflect challenging medication-taking scenarios along with web-based chatbots developed using frequently asked question algorithms and supplemented with emergency hotlines could enhance the decision-making and resource engagement skills of LT recipients.

While most of the reviewed studies (12/15, 80%) referred to the potential of eHealth in enhancing collaborative and individualized health care, only 33% (5/15) of them featured 2-way communication, and just 20% (3/15) incorporated personalized prescriptions or tailored goal setting. Videoconferencing emerged as the most common method for 2-way communication and building personalized strategies. Regular web-based meetings with HCPs or coaches and facilitators during these sessions can be effective, offering benefits such as reinforcing socially desirable behaviors; increasing accountability through clear, reciprocal goals and expectations; and enhancing interpersonal connectedness through support and feedback [49]. In addition, collaborative goal-setting strategies tailored to individual needs have been reported as effective in posttransplant recovery by acknowledging the variability in posttransplant health level, strength, and capacity among LT recipients [50]. The interest of LT recipients in connecting with peer recipients web-based also indicates a need for incorporating peer support groups to foster higher motivation and interpersonal connectedness [33].

In contrast, the studies that scored the lowest on the core self-management skills included minimal human collaborative elements such as communication or education sessions with HCPs [41,45,47]. These interventions primarily leveraged eHealth for its advantages in reducing labor-intensive tasks, enabling immediate evaluations, and improving accessibility while reducing exposure to infection sources [24,51,52] but overlooked the value of human support. Previous findings have suggested that digital person-to-person components can significantly improve effectiveness and adherence in eHealth interventions [49,53]. The selected interventions with lower levels of guidance and support included phone calls or SMS text messaging. However, assessing the relationship between the level of human support and postintervention outcomes was

XSL•FO
RenderX

not feasible due to the heterogeneity of intervention strategies and the lack of clear causality between specific strategies and outcomes. Although it has been demonstrated that eHealth interventions with feedback channels are generally more effective than those without [49], further research is warranted to understand how the directiveness, interactivity, and immediacy of feedback impact the effectiveness of these interventions for sustainable behavior change in LT recipients with complex self-management needs.

In exploring untapped benefits of eHealth among the reviewed studies, we suggest that future research focus on developing predictive models and tailored interventions based on patient-generated data. The potential for creating algorithms to identify behavior patterns could be promising for personalized management or decision support systems. For instance, algorithms analyzing medication-taking logs could proactively identify individuals at risk, enabling more intensive monitoring to prevent medication errors or nonadherence [9]. The capacity of eHealth to collect extensive patient information remotely and conveniently should be maximized to create personalized management plans. By leveraging advanced data analysis and machine learning techniques, coupled with the incorporation of the preferences, needs, and circumstances of LT recipients, there is potential for a more sophisticated, patient-centered design of self-management interventions.

Another critical consideration when developing and implementing eHealth interventions is the age of the LT recipient population. Notably, in 87% (13/15) of the reviewed studies, the average age among the LT recipients was >50 years. Older age has been identified as contributing to the digital divide [54,55], potentially affecting the ability and access of LT recipients to self-management support using eHealth technologies [56,57]. While the proportion of LT recipients aged ≥65 years has increased in the past decade [58], research on older LT recipients has primarily revolved around graft function and long-term survival [59,60]. This indicates a need for further research addressing self-management and QOL in this demographic group [59,61]. The digital literacy of older adults should be assessed as a potential influential factor in eHealth self-management intervention studies [62]. The study by Andrä et al [44] examined the variability in mobile device use and usability with age as a key factor in the discrepancy between patients interested in the mobile app and those who actually used it. The study recommended a longer trial period and repeated training for the older individuals [44]. Moreover, intervention designs should accommodate older users or those with limited digital literacy. This could involve simplifying interfaces, using intuitive features, and providing clear instructions or support to aid their understanding and use of technology.

Furthermore, there is a pressing need for more studies on eHealth use and self-management outcomes among diverse ethnic groups and regions worldwide. The LT recipients in our review were primarily from North America and Europe, which does not proportionately represent the increasing number of LTs in many Asian countries. While the reviewed descriptive studies did not cover a wide array of eHealth-related characteristics, future studies examining the relationship between the digital divide and social determinants—such as ethnicity, educational level, economic status, health care access, and community resources among LT recipients—should consider the variability across different countries and regions. Such research would more accurately reflect the current global situation of LT recipients and validate the effects of eHealth in this population.

Limitations

This review has several limitations. First, as we searched for articles explicitly including terms related to self-management based on previous literature on self-management among LT recipients, it is possible that articles including nuanced aspects of self-management were excluded. Second, it should be considered that the samples of the studies in this review comprised relatively old individuals as our review specifically focused on adult LT recipients. Thus, caution should be exercised when interpreting findings as age may influence adherence and self-management outcomes. As adherence and self-management needs significantly differ across developmental stages, future reviews focusing on younger populations are warranted. Third, this review included only studies published in English. Therefore, we may have missed relevant studies published in non-English languages, which should be considered with our finding related to the disproportional geographical distribution of the studies. Finally, it was inherently challenging to stratify and compare results for a detailed synthesis because this scoping review involved a small number of studies with heterogeneous designs, aims, and contents. Consistent with the purpose of the scoping review approach [63], our focus was on providing an overall mapping of the identified literature in this topic area rather than conducting an in-depth comparison of quantitative findings. Because the topic of eHealth interventions for self-management after LT is in its infancy but rapidly evolving, analyzing the replicability of the interventions to date using the TIDieR checklist may provide better insights for researchers and clinicians interested in further advancing this topic area. We suggest that future reviews prioritize analyzing the effectiveness of eHealth-based self-management interventions on various health outcomes and examine the interactions of social determinant factors as more evidence becomes available.

Conclusions

This scoping review has highlighted the significant potential and emerging challenges of eHealth-based self-management strategies for LT recipients. The reviewed studies predominantly focused on lifestyle management, medication adherence, and remote monitoring. However, there is a noticeable gap in eHealth research concerning alcohol recidivism and the psychosocial and cognitive dimensions of progressing and evaluating self-management (eg, self-efficacy and self-regulation). Future research should aim to develop tailored eHealth interventions that encompass multifaceted elements of self-management skills. These interventions should not only leverage the benefits of technology but also incorporate digital human-to-human interactions to adequately address the complex needs of LT recipients. In addition, ensuring inclusive and equitable self-management support requires addressing the

XSL•FO
RenderX

challenges of digital literacy, catering to the unique needs of older LT recipients, and considering the sociocultural contexts of LT recipients from diverse geographic regions.

Acknowledgments

This research was supported by the National Research Foundation of Korea funded by the Ministry of Education (2022R1I1A2053635; JC); the Institute for Innovation in Digital Healthcare, Yonsei University (JC); the Mo-Im Kim Nursing Research Institute (SHC and JC); and the Multidisciplinary Research Fund (6-2021-0199; SHC and JC) and Faculty Research Fund (6-2021-0177; SHC and JC) from Yonsei University College of Nursing. The authors used ChatGPT (OpenAI) [64] for grammar and language correction.

Authors' Contributions

JC and SHC conceptualized and supervised the study. JC, SHK, and SHC developed the methodology. SHK and HK screened the studies and performed the formal analysis. SHK and KK performed the validation. JC, SHK, KK, SHC, DJJ, and JGL wrote, reviewed, and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of database search strategies.
[DOCX File , 19 KB - jmir_v26i1e56664_app1.docx]

Multimedia Appendix 2
Detailed scores according to the Template for Intervention Description and Replication checklist for the selected intervention studies.
[DOCX File , 23 KB - jmir_v26i1e56664_app2.docx]

References

1. Maynard E. Liver transplantation: patient selection, perioperative surgical issues, and expected outcomes. Surg Clin North Am 2019 Feb;99(1):65-72. [doi: 10.1016/j.suc.2018.09.005] [Medline: 30471742]
2. National data. Organ Procurement & Transplantation Network. URL: https://optn.transplant.hrsa.gov/data/view-data-reports/national-data [accessed 2024-01-06]
3. 2020 Organ transplant cohort KOTRY annual report. Korean Organ Transplantation Registry. 2022 Jan 05. URL: http://www.kotryfoundation.org/rang_board/list.html?num=176&code=mem_ref [accessed 2024-01-06]
4. Black CK, Termanini KM, Aguirre O, Hawksworth JS, Sosin M. Solid organ transplantation in the 21st century. Ann Transl Med 2018 Oct;6(20):409 [FREE Full text] [doi: 10.21037/atm.2018.09.68] [Medline: 30498736]
5. Kim WR, Lake JR, Smith JM, Schladt DP, Skeans MA, Harper AM, et al. OPTN/SRTR 2016 annual data report: liver. Am J Transplant 2018 Jan;18 Suppl 1:172-253 [FREE Full text] [doi: 10.1111/ajt.14559] [Medline: 29292603]
6. Watt KD. Keys to long-term care of the liver transplant recipient. Nat Rev Gastroenterol Hepatol 2015 Nov;12(11):639-648. [doi: 10.1038/nrgastro.2015.172] [Medline: 26460351]
7. Yang LS, Shan LL, Saxena A, Morris DL. Liver transplantation: a systematic review of long-term quality of life. Liver Int 2014 Oct;34(9):1298-1313. [doi: 10.1111/liv.12553] [Medline: 24703371]
8. Ho MC, Wu YM, Hu RH, Ko WJ, Ni YH, Chang MH, et al. Surgical complications and outcome of living related liver transplantation. Transplant Proc 2004 Oct;36(8):2249-2251. [doi: 10.1016/j.transproceed.2004.08.099] [Medline: 15561208]
9. Neuberger J, Bechstein W, Kuypers D, Burra P, Citterio F, De Geest S, et al. Practical recommendations for long-term management of modifiable risks in kidney and liver transplant recipients: a guidance report and clinical checklist by the consensus on managing modifiable risk in transplantation (COMMIT) group. Transplantation 2017 Apr;101(4S Suppl 2):S1-56 [FREE Full text] [doi: 10.1097/TP.0000000000001651] [Medline: 28328734]
10. Dhanasekaran R. Management of immunosuppression in liver transplantation. Clin Liver Dis 2017 May;21(2):337-353. [doi: 10.1016/j.cld.2016.12.007] [Medline: 28364817]
11. Drent G, Moons P, De Geest S, Kleibeuker JH, Haagsma EB. Symptom experience associated with immunosuppressive drugs after liver transplantation in adults: possible relationship with medication non-compliance? Clin Transplant 2008;22(6):700-709. [doi: 10.1111/j.1399-0012.2008.00864.x] [Medline: 18673378]
12. Schulman-Green D, Jaser S, Martin F, Alonzo A, Grey M, McCorkle R, et al. Processes of self-management in chronic illness. J Nurs Scholarsh 2012 Jun;44(2):136-144 [FREE Full text] [doi: 10.1111/j.1547-5069.2012.01444.x] [Medline: 22551013]

XSL•FO
RenderX

13. Van de Velde D, De Zutter F, Satink T, Costa U, Janquart S, Senn D, et al. Delineating the concept of self-management in chronic conditions: a concept analysis. BMJ Open 2019 Jul 16;9(7):e027775 [FREE Full text] [doi: 10.1136/bmjopen-2018-027775] [Medline: 31315862]
14. Lorig KR, Holman H. Self-management education: history, definition, outcomes, and mechanisms. Ann Behav Med 2003 Aug;26(1):1-7. [doi: 10.1207/S15324796ABM2601_01] [Medline: 12867348]
15. Ko D, Muehrer RJ, Bratzke LC. Self-management in liver transplant recipients: a narrative review. Prog Transplant 2018 Jun;28(2):100-115. [doi: 10.1177/1526924818765814] [Medline: 29649931]
16. Vanhoof JM, Vandenberghe B, Geerts D, Philippaerts P, De Mazière P, DeVito Dabbs A, et al. Shedding light on an unknown reality in solid organ transplant patients' self-management: a contextual inquiry study. Clin Transplant 2018 Aug 02;32(8):e13314. [doi: 10.1111/ctr.13314] [Medline: 29897621]
17. Serper M, Patzer RE, Reese PP, Przytula K, Koval R, Ladner DP, et al. Medication misuse, nonadherence, and clinical outcomes among liver transplant recipients. Liver Transpl 2015 Jan;21(1):22-28 [FREE Full text] [doi: 10.1002/lt.24023] [Medline: 25312406]
18. Moayed MS, Ebadi A, Khodaveisi M, Nassiri Toosi M, Soltanian AR, Khatiban M. Factors influencing health self-management in adherence to care and treatment among the recipients of liver transplantation. Patient Prefer Adherence 2018 Nov 16;12:2425-2436 [FREE Full text] [doi: 10.2147/PPA.S180341] [Medline: 30510406]
19. Choudhury A, Varshney M, Sahoo B, Pamecha V, Sinha P, Patil NS, et al. Impact of covid-19 pandemic on quality of life and psychosocial difficulties among liver transplant recipients. J Family Med Prim Care 2022 Feb;11(2):744-750 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_1798_21] [Medline: 35360795]
20. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
21. Badawy SM, Cronin RM, Hankins J, Crosby L, DeBaun M, Thompson AA, et al. Patient-centered eHealth interventions for children, adolescents, and adults with sickle cell disease: systematic review. J Med Internet Res 2018 Jul 19;20(7):e10940 [FREE Full text] [doi: 10.2196/10940] [Medline: 30026178]
22. Liu S, Li J, Wan DY, Li R, Qu Z, Hu Y, et al. Effectiveness of eHealth self-management interventions in patients with heart failure: systematic review and meta-analysis. J Med Internet Res 2022 Sep 26;24(9):e38697 [FREE Full text] [doi: 10.2196/38697] [Medline: 36155484]
23. Tistad M, Lundell S, Wiklund M, Nyberg A, Holmner Å, Wadell K. Usefulness and relevance of an ehealth tool in supporting the self-management of chronic obstructive pulmonary disease: explorative qualitative study of a cocreative process. JMIR Hum Factors 2018 Oct 26;5(4):e10801 [FREE Full text] [doi: 10.2196/10801] [Medline: 30368440]
24. Fix OK, Serper M. Telemedicine and telehepatology during the COVID-19 pandemic. Clin Liver Dis (Hoboken) 2020 May 21;15(5):187-190 [FREE Full text] [doi: 10.1002/cld.971] [Medline: 32537134]
25. Duettmann W, Naik MG, Zukunft B, Osmonodja B, Bachmann F, Choi M, et al. eHealth in transplantation. Transpl Int 2021 Jan;34(1):16-26 [FREE Full text] [doi: 10.1111/tri.13778] [Medline: 33111393]
26. Tang J, James L, Howell M, Tong A, Wong G. eHealth interventions for solid organ transplant recipients: a systematic review and meta-analysis of randomized controlled trials. Transplantation 2020 Aug;104(8):e224-e235. [doi: 10.1097/TP.0000000000003294] [Medline: 32732828]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2007 Feb 23;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Lockwood C, Dos Santos KB, Pap R. Practical guidance for knowledge synthesis: scoping review methods. Asian Nurs Res (Korean Soc Nurs Sci) 2019 Dec;13(5):287-294 [FREE Full text] [doi: 10.1016/j.anr.2019.11.002] [Medline: 31756513]
29. Smith SK, Miloh T. Pediatric liver transplantation. Clin Liver Dis 2022 Aug;26(3):521-535. [doi: 10.1016/j.cld.2022.03.010] [Medline: 35868688]
30. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348(mar07 3):g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
31. Grady PA, Gough LL. Self-management: a comprehensive approach to management of chronic conditions. Am J Public Health 2014 Aug;104(8):e25-e31. [doi: 10.2105/AJPH.2014.302041] [Medline: 24922170]
32. McCorkle R, Ercolano E, Lazenby M, Schulman-Green D, Schilling LS, Lorig K, et al. Self-management: enabling and empowering patients living with cancer as a chronic illness. CA Cancer J Clin 2011;61(1):50-62 [FREE Full text] [doi: 10.3322/caac.20093] [Medline: 21205833]
33. Lieber SR, Kim HP, Baldelli L, Evon DM, Teal R, Nash R, et al. What liver transplant recipients want in a smartphone intervention to enhance recovery: prototype for the LiveRight transplant app. Liver Transpl 2021 Apr;27(4):584-589. [doi: 10.1002/lt.25967] [Medline: 33300288]
34. Maroney K, Curtis LM, Opsasnick L, Smith KD, Eifler MR, Moore A, et al. eHealth literacy and web-based patient portal usage among kidney and liver transplant recipients. Clin Transplant 2021 Feb;35(2):e14184. [doi: 10.1111/ctr.14184] [Medline: 33278846]

XSL•FO
RenderX

35. Mathur S, Janaudis-Ferreira T, Hemphill J, Cafazzo JA, Hart D, Holdsworth S, et al. User-centered design features for digital health applications to support physical activity behaviors in solid organ transplant recipients: a qualitative study. Clin Transplant 2021 Dec;35(12):e14472. [doi: 10.1111/ctr.14472] [Medline: 34510558]
36. Vanhoof JM, Vandenberghe B, Geerts D, Philippaerts P, De Mazière P, DeVito Dabbs A, et al. Technology experience of solid organ transplant patients and their overall willingness to use interactive health technology. J Nurs Scholarsh 2018 Mar;50(2):151-162. [doi: 10.1111/jnu.12362] [Medline: 29193654]
37. Wedd J, Basu M, Curtis LM, Smith K, Lo DJ, Serper M, et al. Racial, ethnic, and socioeconomic disparities in web-based patient portal usage among kidney and liver transplant recipients: cross-sectional study. J Med Internet Res 2019 Apr 22;21(4):e11864 [FREE Full text] [doi: 10.2196/11864] [Medline: 31008707]
38. Barnett A, Campbell KL, Mayr HL, Keating SE, Macdonald GA, Hickman IJ. Liver transplant recipients' experiences and perspectives of a telehealth-delivered lifestyle programme: a qualitative study. J Telemed Telecare 2021 Oct;27(9):590-598. [doi: 10.1177/1357633X19900459] [Medline: 31986966]
39. Hickman I, Hannigan A, Johnston H, Elvin-Walsh L, Mayr HL, Staudacher HM, et al. Telehealth-delivered, cardioprotective diet and exercise program for liver transplant recipients: a randomized feasibility study. Transplant Direct 2021 Feb 04;7(3):e667 [FREE Full text] [doi: 10.1097/TXD.0000000000001118] [Medline: 33564717]
40. Ertel AE, Kaiser TE, Abbott DE, Shah SA. Use of video-based education and tele-health home monitoring after liver transplantation: results of a novel pilot study. Surgery 2016 Oct;160(4):869-876. [doi: 10.1016/j.surg.2016.06.016] [Medline: 27499142]
41. Koc Ö, Pierco M, Remans K, Van den Hende T, Verbeek J, Van Malenstein H, et al. Telemedicine based remote monitoring after liver transplantation: feasible in a select group and a more stringent control of immunosuppression. Clin Transplant 2022 Jan;36(1):e14494 [FREE Full text] [doi: 10.1111/ctr.14494] [Medline: 34596918]
42. Lee TC, Kaiser TE, Alloway R, Woodle ES, Edwards MJ, Shah SA. Telemedicine based remote home monitoring after liver transplantation: results of a randomized prospective trial. Ann Surg 2019 Sep;270(3):564-572. [doi: 10.1097/SLA.0000000000003425] [Medline: 31356267]
43. Tian M, Wang B, Xue Z, Dong D, Liu X, Wu R, et al. Telemedicine for follow-up management of patients after liver transplantation: cohort study. JMIR Med Inform 2021 May 17;9(5):e27175 [FREE Full text] [doi: 10.2196/27175] [Medline: 33999008]
44. Andrä M, Sibinovic M, Pfeiffer K, Kniepeiss D. Implementation of a mobile application for outpatient care after liver transplantation. Digit Health 2022 Dec 25;8:20552076221145855 [FREE Full text] [doi: 10.1177/20552076221145855] [Medline: 36601283]
45. Melilli E, Cestone G, Revuelta I, Meneghini M, Lladó L, Montero N, et al. Adoption of a novel smart mobile-health application technology to track chronic immunosuppression adherence in solid organ transplantation: results of a prospective, observational, multicentre, pilot study. Clin Transplant 2021 May;35(5):e14278. [doi: 10.1111/ctr.14278] [Medline: 33682207]
46. Serper M, Barankay I, Chadha S, Shults J, Jones LS, Olthoff KM, et al. A randomized, controlled, behavioral intervention to promote walking after abdominal organ transplantation: results from the LIFT study. Transpl Int 2020 Jun;33(6):632-643 [FREE Full text] [doi: 10.1111/tri.13570] [Medline: 31925833]
47. Zanetti-Yabur A, Rizzo A, Hayde N, Watkins AC, Rocca JP, Graham JA. Exploring the usage of a mobile phone application in transplanted patients to encourage medication compliance and education. Am J Surg 2017 Oct;214(4):743-747. [doi: 10.1016/j.amjsurg.2017.01.026] [Medline: 28256241]
48. Shenoy A, Salajegheh A, Shen NT. Multimodal multidisciplinary management of alcohol use disorder in liver transplant candidates and recipients. Transl Gastroenterol Hepatol 2022 Jul 25;7:28 [FREE Full text] [doi: 10.21037/tgh.2020.02.22] [Medline: 35892051]
49. Santarossa S, Kane D, Senn CY, Woodruff SJ. Exploring the role of in-person components for online health behavior change interventions: can a digital person-to-person component suffice? J Med Internet Res 2018 Apr 11;20(4):e144 [FREE Full text] [doi: 10.2196/jmir.8480] [Medline: 29643048]
50. Dunn MA, Rogal SS, Duarte-Rojo A, Lai JC. Physical function, physical activity, and quality of life after liver transplantation. Liver Transpl 2020 May;26(5):702-708 [FREE Full text] [doi: 10.1002/lt.25742] [Medline: 32128971]
51. Faust A, Stine JG. Leveraging the coronavirus disease 2019 pandemic: is it time to consider incorporating mobile applications into standard clinical management of the liver transplantation patient? Liver Transpl 2021 Apr;27(4):479-481 [FREE Full text] [doi: 10.1002/lt.25991] [Medline: 33484229]
52. Lee H, Shin BC, Seo JM. Effectiveness of eHealth interventions for improving medication adherence of organ transplant patients: a systematic review and meta-analysis. PLoS One 2020 Nov 5;15(11):e0241857 [FREE Full text] [doi: 10.1371/journal.pone.0241857] [Medline: 33152010]
53. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
54. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]

XSL•FO
RenderX

55. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
56. Deshpande N, Arora VM, Vollbrecht H, Meltzer DO, Press V. eHealth literacy and patient portal use and attitudes: cross-sectional observational study. JMIR Hum Factors 2023 Jan 27;10:e40105 [FREE Full text] [doi: 10.2196/40105] [Medline: 36705947]
57. Cheng J, Arora VM, Kappel N, Vollbrecht H, Meltzer DO, Press V. Assessing disparities in video-telehealth use and ehealth literacy among hospitalized patients: cross-sectional observational study. JMIR Form Res 2023 May 12;7:e44501 [FREE Full text] [doi: 10.2196/44501] [Medline: 37171854]
58. Durand F, Levitsky J, Cauchy F, Gilgenkrantz H, Soubrane O, Francoz C. Age and liver transplantation. J Hepatol 2019 Apr;70(4):745-758. [doi: 10.1016/j.jhep.2018.12.009] [Medline: 30576701]
59. Dolnikov S, Adam R, Cherqui D, Allard MA. Liver transplantation in elderly patients: what do we know at the beginning of 2020? Surg Today 2020 Jun;50(6):533-539 [FREE Full text] [doi: 10.1007/s00595-020-01996-7] [Medline: 32279191]
60. Wilson GC, Quillin RC3, Wima K, Sutton JM, Hoehn RS, Hanseman DJ, et al. Is liver transplantation safe and effective in elderly (≥70 years) recipients? A case-controlled analysis. HPB (Oxford) 2014 Dec;16(12):1088-1094 [FREE Full text] [doi: 10.1111/hpb.12312] [Medline: 25099347]
61. Werkgartner G, Wagner D, Manhal S, Fahrleitner-Pammer A, Mischinger HJ, Wagner M, et al. Long-term quality of life of liver transplant recipients beyond 60 years of age. Age (Dordr) 2013 Dec;35(6):2485-2492 [FREE Full text] [doi: 10.1007/s11357-013-9527-x] [Medline: 23529506]
62. Wang Y, Song Y, Zhu Y, Ji H, Wang A. Association of eHealth literacy with health promotion behaviors of community-dwelling older people: the chain mediating role of self-efficacy and self-care ability. Int J Environ Res Public Health 2022 May 17;19(10):6092 [FREE Full text] [doi: 10.3390/ijerph19106092] [Medline: 35627627]
63. Daudt HM, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. BMC Med Res Methodol 2013 Mar 23;13:48 [FREE Full text] [doi: 10.1186/1471-2288-13-48] [Medline: 23522333]
64. ChatGPT | OpenAI. URL: https://openai.com/chatgpt/ [accessed 2024-06-24]

Abbreviations

HCP: health care provider
LT: liver transplantation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QOL: quality of life
RCT: randomized controlled trial
TIDieR: Template for Intervention Description and Replication

Edited by G Eysenbach, S Ma; submitted 23.01.24; peer-reviewed by K Bul; comments to author 20.02.24; revised version received 24.03.24; accepted 03.06.24; published 04.07.24.

Please cite as:
Kim SH, Kim KA, Chu SH, Kim H, Joo DJ, Lee JG, Choi J
Self-Management Using eHealth Technologies for Liver Transplant Recipients: Scoping Review
J Med Internet Res 2024;26:e56664
URL: https://www.jmir.org/2024/1/e56664
doi:10.2196/56664
PMID:

©Soo Hyun Kim, Kyoung-A Kim, Sang Hui Chu, Hyunji Kim, Dong Jin Joo, Jae Geun Lee, JiYeon Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Open-Access Database and Traditional Intensive Care Studies Using Machine Learning: Bibliometric Analysis Study

Yuhe Ke[1*], MBBS; Rui Yang[2*], MSc; Nan Liu[2], PhD

[1]Division of Anesthesiology and Perioperative Medicine, Singapore General Hospital, Singapore, Singapore

[2]Centre for Quantitative Medicine, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Nan Liu, PhD
Centre for Quantitative Medicine
Duke-NUS Medical School
National University of Singapore
8 College Road
Singapore, 169857
Singapore
Phone: 65 66016503
Email: liu.nan@duke-nus.edu.sg

Abstract

Background: Intensive care research has predominantly relied on conventional methods like randomized controlled trials. However, the increasing popularity of open-access, free databases in the past decade has opened new avenues for research, offering fresh insights. Leveraging machine learning (ML) techniques enables the analysis of trends in a vast number of studies.

Objective: This study aims to conduct a comprehensive bibliometric analysis using ML to compare trends and research topics in traditional intensive care unit (ICU) studies and those done with open-access databases (OADs).

Methods: We used ML for the analysis of publications in the Web of Science database in this study. Articles were categorized into “OAD” and “traditional intensive care” (TIC) studies. OAD studies were included in the Medical Information Mart for Intensive Care (MIMIC), eICU Collaborative Research Database (eICU-CRD), Amsterdam University Medical Centers Database (AmsterdamUMCdb), High Time Resolution ICU Dataset (HiRID), and Pediatric Intensive Care database. TIC studies included all other intensive care studies. Uniform manifold approximation and projection was used to visualize the corpus distribution. The BERTopic technique was used to generate 30 topic-unique identification numbers and to categorize topics into 22 topic families.

Results: A total of 227,893 records were extracted. After exclusions, 145,426 articles were identified as TIC and 1301 articles as OAD studies. TIC studies experienced exponential growth over the last 2 decades, culminating in a peak of 16,378 articles in 2021, while OAD studies demonstrated a consistent upsurge since 2018. Sepsis, ventilation-related research, and pediatric intensive care were the most frequently discussed topics. TIC studies exhibited broader coverage than OAD studies, suggesting a more extensive research scope.

Conclusions: This study analyzed ICU research, providing valuable insights from a large number of publications. OAD studies complement TIC studies, focusing on predictive modeling, while TIC studies capture essential qualitative information. Integrating both approaches in a complementary manner is the future direction for ICU research. Additionally, natural language processing techniques offer a transformative alternative for literature review and bibliometric analysis.

(J Med Internet Res 2024;26:e48330) doi:10.2196/48330

KEYWORDS

BERTopic; critical care; eICU; machine learning; MIMIC; Medical Information Mart for Intensive Care; natural language processing

XSL•FO
RenderX

Introduction

The start of critical care as a medical subspecialty can be traced back to a polio epidemic during which a substantial number of patients needed prolonged mechanical ventilation [1]. Over time, the field of critical care has experienced significant growth and continual evolution. Research in this field has played a pivotal role in unraveling the complexities of numerous diseases and treatment modalities, driving substantial advancements in clinical practice over the past decades [2]. Groundbreaking studies have investigated critical areas such as sepsis, mechanical ventilation, acute lung and kidney injuries, intensive care unit (ICU) delirium, and sedation in critically ill patients [3].

These research studies have often been conducted in traditional ways such as prospective and randomized controlled trials [4], cohort and observational studies, clinical trials [5], and clinical and translational research [6]. These traditional methods have revolutionized patient care and improved outcomes significantly. For instance, the implementation of protocol-driven, goal-directed management of sepsis and appropriate fluid therapy has led to remarkable reductions in mortality rates [7,8], and these findings have been integral in developing evidence-based practice guidelines that are now the gold standard [9,10].

Despite their undeniable merits, traditional research methods in intensive care also come with several limitations [11]. Clinical trials are known for their high costs [12], stringent standardization requirements, and ethical oversight [13]. Data collection can be laborious, prone to human errors, and constrained in terms of quantity and granularity [14]. Moreover, obtaining patient consent for most randomized controlled trials in the ICU poses challenges [15], necessitating alternative consent models. These limitations have become increasingly apparent as medical complexity continues to grow exponentially [16].

The advent of electronic health records (EHRs) has heralded a new era in clinical research by facilitating the digitization of health care systems [17]. In this era of data science, a more integrated approach can be adopted, using machine learning (ML) algorithms to tackle the complexity of critical illness [18]. Open-access databases (OADs), such as the Medical Information Mart for Intensive Care (MIMIC) database [19] and the Philips eICU Collaborative Research Database (eICU-CRD) [20], have played a transformative role by enabling free data sharing.

The concept of free and open databases plays a pivotal role in promoting data sharing and advancing medical knowledge in accordance with the findable, accessible, interoperable, and reusable (FAIR) guiding principle. The FAIR principles, which emphasize that data should be findable, accessible, interoperable, and reusable, are essential for fostering a collaborative and transparent scientific research environment [21,22]. By removing barriers to access, free, and open databases allow researchers, regardless of their affiliations or resources, to contribute to and benefit from the collective pool of information. Accessibility fosters inclusivity and diversity in research, promoting a broader range of perspectives and approaches to medical challenges. This democratization of knowledge leads to a more equitable distribution of information. Researchers can now leverage these vast repositories of information for ML and artificial intelligence studies, marking a departure from traditional intensive care (TIC) research approaches.

Conducting a literature review [23] to investigate the disparities between traditional ICU research and studies based on open-access data sets holds significant importance as it provides a comprehensive understanding of the strengths and limitations of the latter. However, conventional methods of literature reviews and bibliometric analysis have their limitations, especially when dealing with large-scale literature due to computational complexity and the labor-intensive nature of manual interpretations [24-26]. To address these challenges, natural language processing (NLP) offers a promising avenue, while topic modeling techniques can be used to extract various topic themes from extensive data sets [27,28].

Built on the foundations of bidirectional encoder representations from transformers (BERT), BERTopic introduces a novel approach to topic modeling [29,30]. Unlike traditional unsupervised models like latent Dirichlet allocation, which rely on “bag-of-words” model [31], BERTopic overcomes the problem of semantic information loss, significantly enhancing the accuracy of generated topics, and providing more interpretable compositions for each topic, which greatly facilitates the classification of topics.

With the aid of BERTopic, this study aims to shed light on the disparities and commonalities between studies conducted through OADs and TIC research. By analyzing the overall trends and patterns in these 2 groups, we seek to identify knowledge gaps and explore avenues for complementary contributions between these research approaches.

Methods

Data

Data Filtering

We performed an ML-based analysis of research abstracts in the Web of Science (WoS) database to automatically categorize the research papers to conduct this literature mapping analysis. There was no limit to the year of publication of the articles. The search query consisted of the following keyword to identify all the studies that were published under the umbrella of intensive care: (“ICU” OR “intensive care”). The search terms were deliberately left to be broad to cover broad spectrums of journals in the field.

The inclusion criteria were as follows: (1) written in English, (2) articles that had keywords related to intensive care, (3) articles that had the article type of “article” or “review.” We excluded articles with incomplete data fields (eg, title, abstract, publication year, and paper citation). The articles included were then further processed to identify if they were studies using OADs. These articles were labeled as “open-access database,” while the rest of the articles extracted were labeled as “traditional intensive care.”

XSL•FO
RenderX

The search used for this study was performed on January 18, 2023, from WoS. This generated 227,893 search results, which were subsequently reselected using Python. An advanced search from PubMed was done based on the broad search terms of ICU studies used from previous Cochrane ICU literature review [32] to ensure the accuracy of the results. The numbers corroborated with a discrepancy of 4.9% (227,893 WoS keyword search vs 239,748 PubMed ICU keyword search).

Selection Criteria for OADs

A title search using keywords from all currently existing OADs was conducted to identify OAD studies. These include (1) MIMIC [19], (2) eICU-CRD [20], (3) Amsterdam University Medical Centers Database (AmsterdamUMCdb) [33], (4) High Time Resolution ICU Dataset (HiRID) [34], and (5) Pediatric Intensive Care database [35]. We avoided including only keywords in the search and restricted the search years by the year that the OAD was made publicly available to reduce the inadvertent inclusion of incorrect articles due to keywords. For instance, the search term for OADs published with the MIMIC database included title keyword search with the terms (“MIMIC-IV” OR “MIMIC-III” OR “MIMIC-II” OR “MIMIC Dataset” OR “medical information mart for intensive care” OR “MIMIC IV” OR “MIMIC III” OR “MIMIC II”) in studies that were published after 2003. The title keyword search for the searches and the year of cutoff for each OAD are presented in Multimedia Appendix 1.

Furthermore, to ensure the accuracy of the supervised keyword classification, a manual review of the classification by 2 critical care physicians was done for 100 articles from each category that were randomly selected. The review was done independently with the physicians labeling the extract publications into OAD and TICs. An accuracy of 99% was achieved on independent reviews, and full agreement was achieved after discussion on the discrepancy. The final results were matched with the supervised keyword classification.

We performed a bibliometric analysis by directly extracting publication details from the WoS database using Python (Python Software Foundation). The analysis involved assessing the number of articles published per year, calculating total citation counts, and identifying the top journals that published intensive care-related articles. Comprehensive results are presented in Multimedia Appendix 2.

Data Analysis

Uniform Manifold Approximation and Projection

Uniform manifold approximation and projection (UMAP) is a manifold learning technique for dimension reduction, which can identify key structures in high-dimensional data space and map them to low-dimensional space to accomplish dimensionality reduction. Compared to other dimensionality reduction algorithms, such as principal component analysis [36], UMAP can retain more global features [37]. In this paper, we constructed a corpus consisting of abstract words from all studies. However, due to the massive size of the corpus, visualizing and analyzing the high-dimensional data to explore the differences in the vocabulary patterns between the OAD and TIC studies is a challenge. The UMAP package in Python, which implements the UMAP algorithm, was used to project the high-dimensional corpus to 4 dimensions. By cross plotting each dimension, we were able to investigate underlying differences in corpus distribution between OAD and TIC studies.

BERTopic

Topic modeling can help us explore the similarities and differences between research topics in OAD and TIC studies. Unlike conventional topic modeling models, BERTopic uses the BERT framework for embeddings, enabling a deeper understanding of semantic relationships [30]. The BERTopic model was implemented by the BERTopic package in Python and divided 146,727 studies into 30 topic IDs. We also performed latent Dirichlet allocation topic modeling through Python’s LdaModel package for comparison. Through the review of topic keywords by 2 critical care physicians, BERTopic exhibited superior accuracy and sophistication in topic identification, with enhanced interpretability and scientific rigor.

Consequently, the BERTopic model was used for the final analysis. Each of these topics was given a corresponding clinical research category. The overlapping categories were merged into topic families for easier comparisons. By using these advanced techniques, we were able to uncover hidden patterns and relationships within the literature and provide insights into the current state of intensive care research.

Results

A total of 227,893 records were identified from the WoS database on January 18, 2023, of which 195,463 full records were subsequently processed. Records were excluded if they are not “article” or “review” or if they do not contain keywords related to intensive care. After exclusions, 145,426 articles were identified as TIC studies and 1301 articles were categorized as OAD (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) diagram of the study. The final studies were divided into open-access database (OAD) and traditional intensive care (TIC) studies.

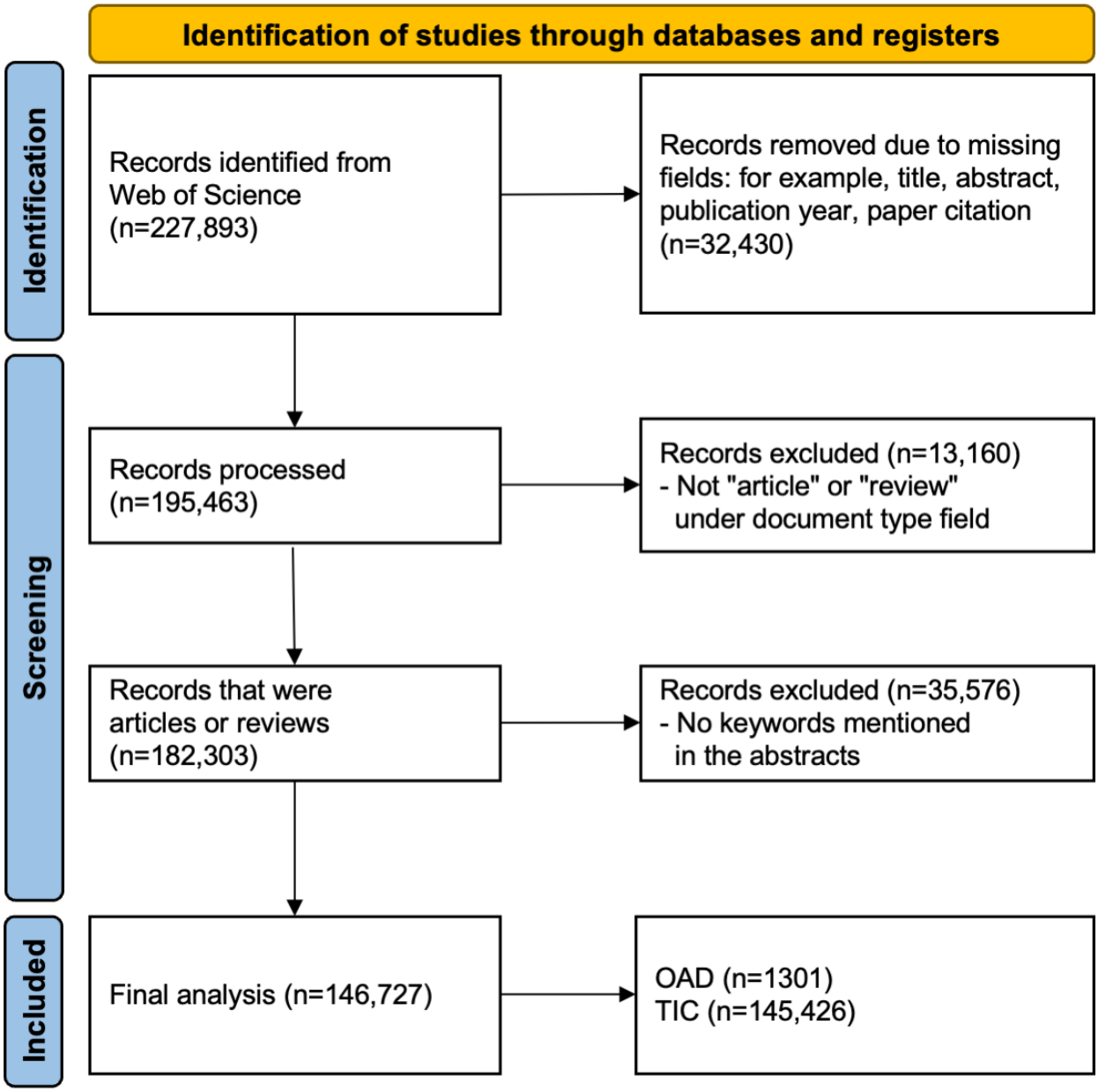

We examined the number of articles published per year to analyze the trends in TIC and OAD studies (Figure 2). Over the past 2 decades, TIC studies have experienced exponential growth, culminating in a peak of 16,378 articles in 2021. A subsequent decline in the number of publications occurred in 2022, likely attributable to delayed indexing within the WoS database and a reduction in COVID-19–related studies as the pandemic stabilized [38]. In contrast, the first OAD study emerged in 2003, with its popularity experiencing a consistent upsurge since 2018. Nonetheless, the number of OAD publications remains markedly lower in comparison to TIC publications.

XSL•FO
RenderX

Figure 2. Number of publications of open-access database (OAD) and traditional intensive care (TIC) studies by year. The first study in the OAD category started in the year 2003.

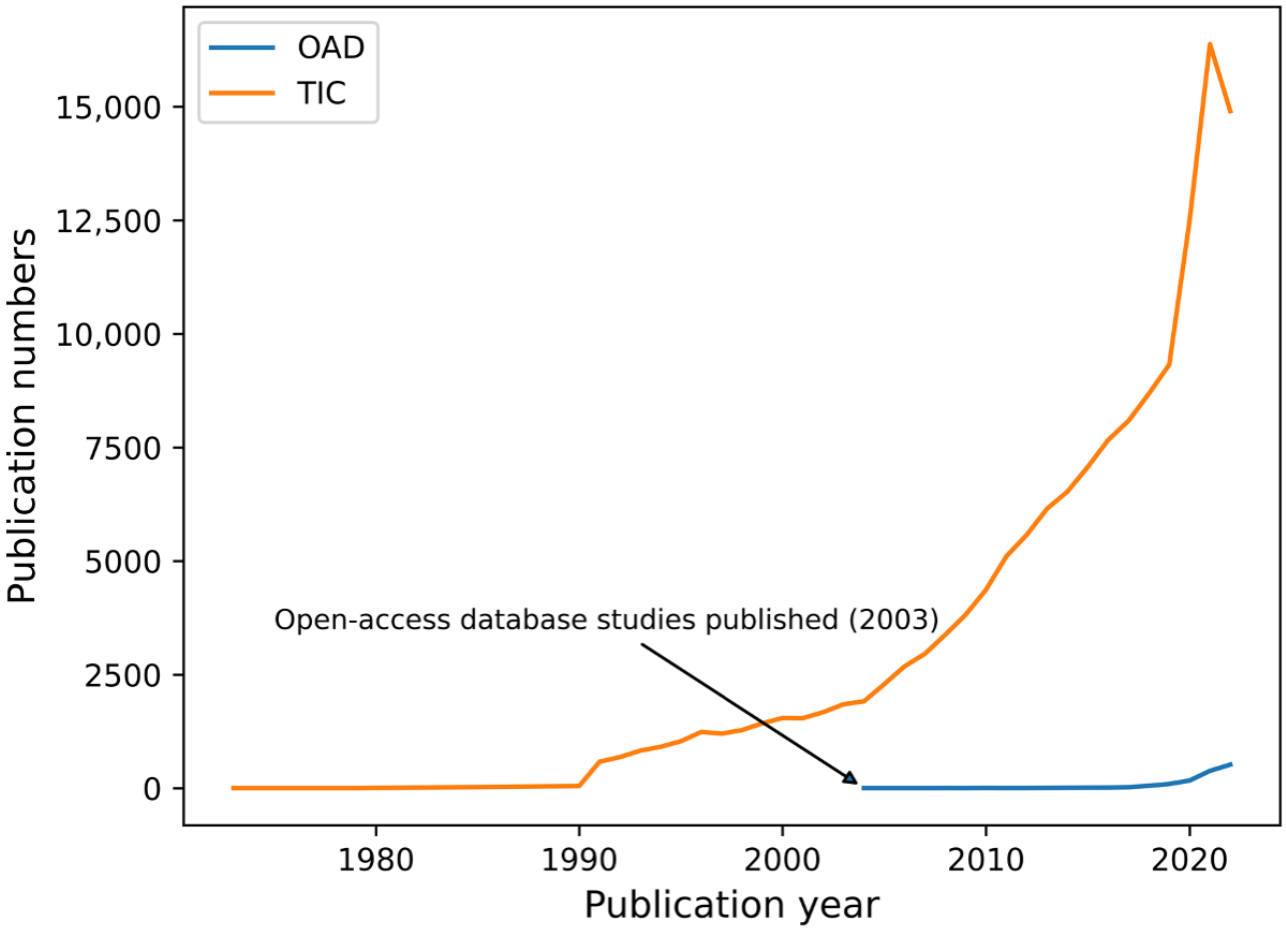

The OAD studies were published most frequently in new open-access journals such as *Frontiers in Medicine*, *Frontiers in Cardiovascular Medicine*, and *Scientific Reports* while the TIC studies were published most frequently in established journals like *Critical Care Medicine*, *Intensive Care Medicine*, and *Critical Care* (Multimedia Appendix 2). Further analysis of keywords from the abstracts showed 2.4% (3492/145,426) TIC studies were meta-analyses or systematic reviews, while only 0.08% (1/1301) OAD study was in this category. There were 5.61% (73/1301) OAD studies, and 7.43% (10,799/145,426) TIC studies that had the keyword of "cost." Examples of the data fields that are available within OADs such as MIMIC and eICU-CRD are listed in Textbox 1. Some information fields such as end-of-life goals and values and health care provider psychology are not available within the current EHRs extracted for OADs.

XSL•FO
RenderX

Textbox 1. Examples of information available in the current open-access databases (OADs) and examples of information not available in OADs.

Examples of information that is available in current OADs

- Patient information: demographics and social set-up
- Hospital context: admission time and discharge time, intensive care unit (ICU) and hospital admissions, and pre-ICU admission
- Diagnosis: physician-curated ICU diagnosis and data-driven phenotypes
- Intervention: medications, procedures, and organ support
- Diagnostics: blood test, microbiology, and scans
- Clinical texts: clinical notes and diagnostic reports
- Physiological monitoring: basic monitoring and waveforms

Examples of information that is not readily available in current OADs

- Patient information: family set up and visiting, financial information, and special populations
- Hospital context: post-ICU discharge details, delayed admission or discharge, and health personnel psychology
- Diagnosis: pre-ICU history and diagnosis requiring clinical symptoms
- Intervention: indications for interventions, complications, and intraoperative and postoperative
- Diagnostics: pathology photographs, imaging, and molecular or genetic studies
- Clinical texts: patient narratives, end-of-life goals and patient value, and health personnel behavior
- Physiological monitoring: advanced monitoring

The UMAP algorithm was used to project the high-dimension corpus to 4 dimensions and allowed exploration of the vocabulary patterns between the OAD and TIC studies (Figure 3). The projection values are represented by the x-axis, while the densities are represented by the y-axis. The presence of considerable overlap between TIC studies and OAD studies suggests that they share a substantial number of common terminologies, which may correspond to similar research topics. Nonetheless, TIC studies exhibit a more extensive coverage than OAD studies, which may stem from broader research scope and extended research duration.

XSL•FO
RenderX

Figure 3. Corpus distribution of open-access database and traditional intensive care studies along dimensions 1-4. OAD: open-access database; TIC: traditional intensive care.

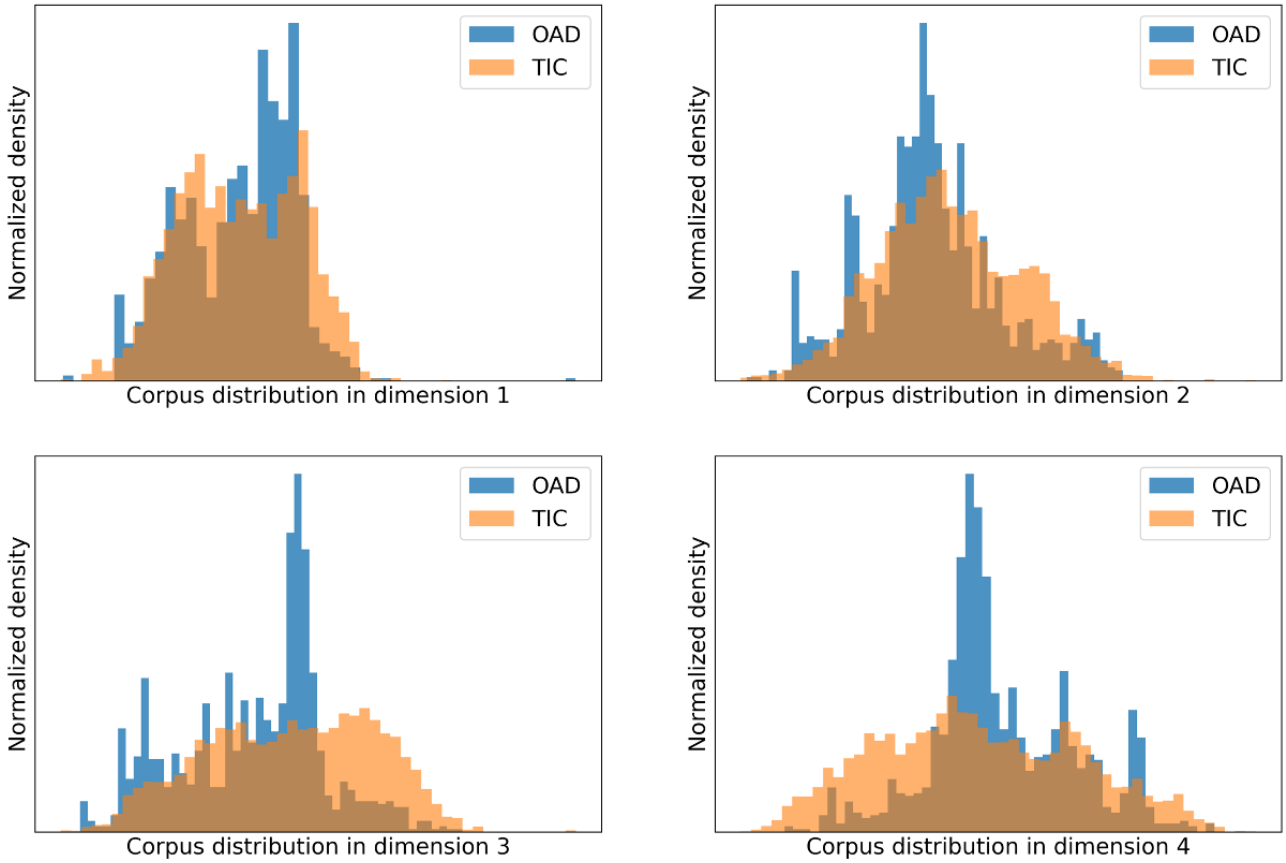

Subsequently, the BERTopic model was then used to generate 30 topic IDs (Figure 4). The internal commonalities of each topic ID were reviewed by critical care physicians and assigned a specific subtopic in intensive care research. The model was able to automatically classify the topics with high interpretability and the topic components were interpreted with ease. For instance, components in topic ID 5 consist of, in decreasing order of weightage: “learning,” “model,” “machine,” “machine learning,” “models,” “data,” “prediction,” and “performance.” This topic was consequently labeled “predictive model” (topic ID 5 in Multimedia Appendix 3).

Figure 4. The ratio of open-access database and traditional intensive care studies within each topic was identified by BERTopic.

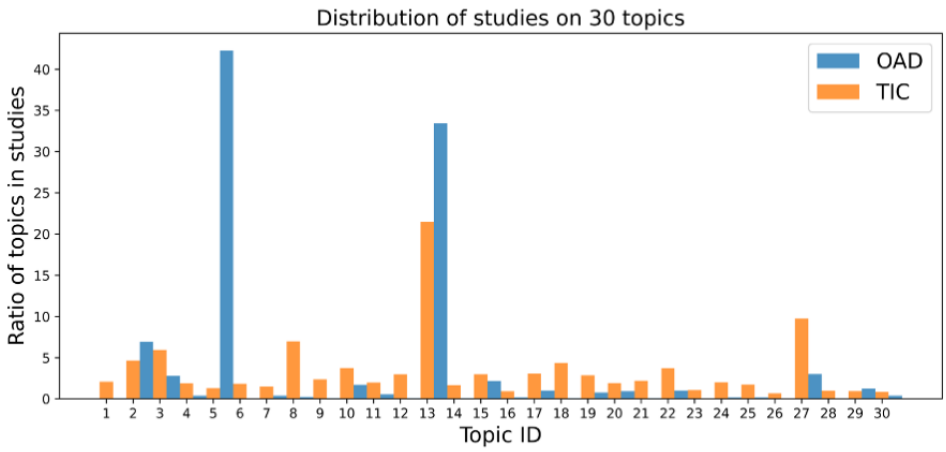

ID	Topic Family
1	Cellular / molecular studies
2	Kidney injury
3	Cardiac surgeries
4	Fungal infections
5	Predictive model
6	Pediatrics care
7	Antibiotics resistance
8	Pediatrics care
9	Healthcare associated infections
10	Delirium
11	Nutrition and rehabilitation
12	Pregnancy related
13	Sepsis
14	Antibiotics resistance
15	Glucose control
16	Cellular / molecular studies
17	Thromboembolism
18	Thoracic surgeries
19	Medication errors
20	Ventilation related
21	Viral infections
22	Traumatic brain injury
23	Health personnel and psychology
24	Antibiotics resistance
25	End of life and ethics
26	Pediatrics care
27	Ventilation related
28	Nutrition and rehabilitation
29	Obesity
30	Ventilation related

XSL•FO
RenderX

The overall topic distribution in TIC studies was more uniform, while the OAD studies tended to be concentrated on several topics including topic ID 2 (kidney injury), 5 (predictive model), and 13 (sepsis). Some topics that were missing in OAD studies included 6 (pediatrics care), 21 (viral infections), 23 (health personnel and psychology), and 28 (nutrition and rehabilitation).

The similarity matrix shows that there was little overlap between the topics (Multimedia Appendix 4). To facilitate the interpretability of the categories, the overlapping topic IDs were merged to form the final 22 topic families (Multimedia Appendix 3).

Topics such as "healthcare associated infection," "thoracic surgeries," and "pregnancy related" research were among the more frequently discussed 15 topics in TIC studies but have limited publications in OAD studies. The topics of "predictive model," "obesity," and "fungal infections" were popular in OAD studies but not the TIC studies. Overall, the topic distributions of the TIC studies were distributed more evenly with the topic family of sepsis accounting for a quarter of the studies, while publications in the OAD studies were heavily skewed toward the predictive model (>40%) and sepsis (>30%; Figure 5).

Figure 5. The top 15 topic families represented in open-access database and traditional intensive care studies.

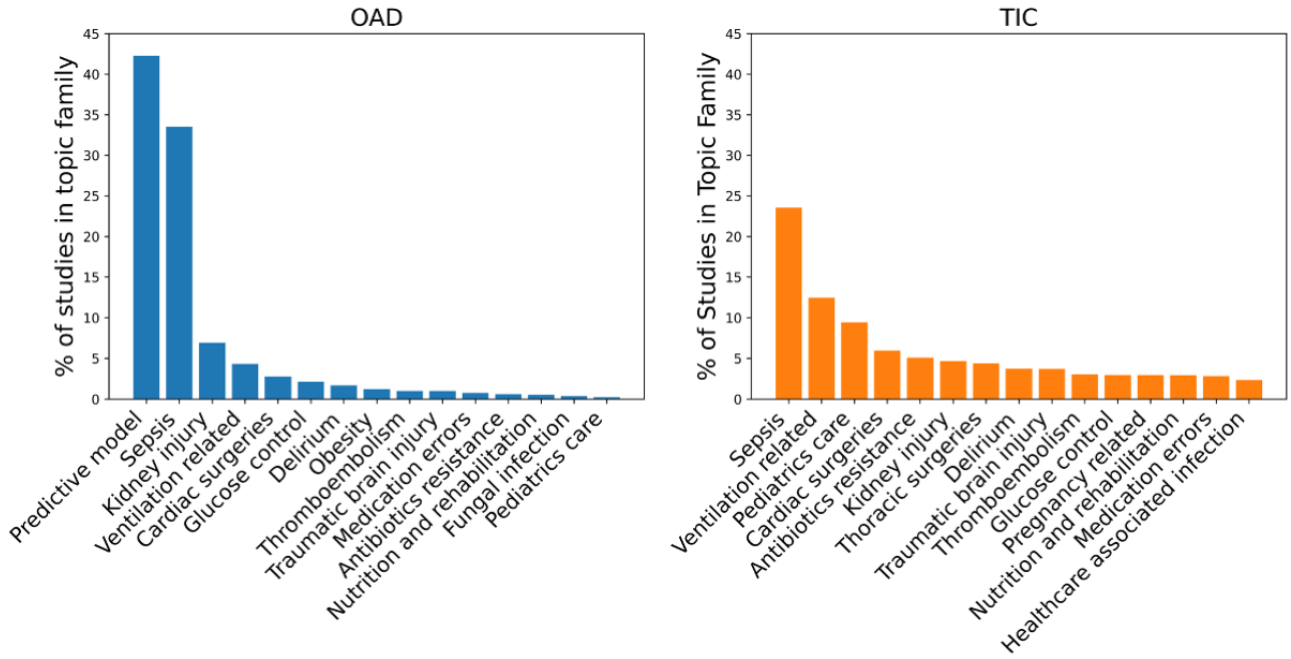

Discussion

Principal Results

This study conducted a comprehensive review and bibliometric analysis of OAD and TIC studies. NLP was used to facilitate this large-scale literature review. Studies using OADs mainly concentrated on a few topics, such as predictive modeling, while TIC studies covered a wider range of topics with a more balanced distribution.

Advantages of OAD Studies

OAD studies offer several advantages that have contributed to their increasing popularity in intensive care research. The granularity of data and easy access to large-cohort databases, such as MIMIC [39], has enabled researchers to perform predictive modeling and conduct various secondary analyses efficiently [40,41]. This accessibility has provided valuable opportunities for exploring specific aspects of patient care, evident in studies investigating phenomena like "weekend effects" and circadian rhythms in ICU patients before discharge [42-46]. The vast amount of longitudinal and time series data available in OADs has also facilitated the implementation of complex ML and deep learning methods [47].

Limitations of OAD Studies

However, it is crucial to acknowledge the retrospective nature of OAD data, which inherently limits the assessment of confounding factors and the ability to draw strong causal conclusions. The observational design of OAD studies may result in lower-quality evidence according to the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) framework [48,49], and thus, the research from OAD studies has yet to be fully integrated into existing evidence-based guidelines, as exemplified by the omission of OAD studies in the 2021 sepsis guidelines [50]. Nevertheless, OADs remain a valuable resource for supplementing and complementing TIC studies, providing unique insights and enhanced predictive scores for intensive care settings.

Furthermore, approximately 50% of the studies using OADs published focused on predictive modeling. The increased usage of ML methods in predictive modeling has not been without critique. Some medical prediction problems inherently possess linear characteristics, and the selection of features may predominantly focus on already known strong predictors, leading to limited improvements in prediction accuracy with ML [51]. Additionally, interstudy heterogeneity poses a challenge in comparing results obtained from different ML models applied to the same data sets [52]. The ethical implications of relying

XSL•FO
RenderX

solely on ML models to make high-risk health care decisions instead of involving clinical expertise are also relevant considerations [51,53].

While OADs provide comprehensive patient data, there are certain limitations in their ability to capture specific information essential for certain critical care research areas. Notably, data fields related to qualitative aspects such as ethics and end-of-life care [54,55], and health care personnel psychology [56] may be challenging, if not impossible, to obtain through OADs generated from EHRs. Consequently, TIC studies have played a pivotal role in addressing these limitations by capturing critical information that is integral to understanding ethical considerations, patient experiences, and health care provider psychology in intensive care [57,58].

Synergy Between OAD and TIC Studies

The synergy between OAD and TIC studies is a promising approach to enhance the comprehensiveness and robustness of intensive care research. OADs, with their large cohort sizes, can serve as external validation cohorts for ML models developed from TIC studies, potentially reducing the sample sizes required for prospective research. Furthermore, OAD studies can corroborate the results of TIC studies, benefiting from larger sample sizes and real-world data, thus providing more practical insights for implementation in intensive care settings [43]. The integration of OAD and TIC studies presents an opportunity to bridge the gaps in data availability and research methodologies, ultimately enriching the understanding and practice of critical care medicine.

Potential Impact of NLP

The usage of large language models such as BERTopic has proven to be a valuable tool for large-scale literature review and topic extraction [58]. This approach has enabled accurate, reliable, and granular topic generation, offering clinicians a more effective means of interpreting data compared to traditional bag-of-words models [59]. The potential of NLP to analyze scientific articles and identify trends and knowledge gaps holds promise for shaping the future of research in critical care medicine. As the volume of publications in critical care continues to grow and large language modeling continues to advance in health care [60], AI technology will be crucial in efficiently identifying and predicting emerging trends.

Future Directions

Future research in the field of critical care can explore novel applications of ML beyond predictive modeling. For instance, using ML to study patterns in how papers are cited, shared, and discussed on the web could help predict their potential impact on the scientific community. This analysis can aid in identifying highly influential papers and understanding the factors that contribute to their recognition. Additionally, investigations into methods for enhancing the interpretability and transparency of ML algorithms in critical care research would further facilitate the ethical and responsible use of AI technologies.

Strengths and Limitations

The study's application of NLP-driven in analyzing scientific articles and identifying trends highlights the potential impact of AI technologies in streamlining literature reviews and identifying emerging trends more efficiently.

Another notable strength of this study is the usage of the WoS database, the world's oldest and most extensively used repository of research publications and citations, encompassing approximately 34,000 journals [61]. The comprehensiveness of this database provides a robust representation of the literature in the field of intensive care research. Nevertheless, it is essential to acknowledge that some articles published in nonindexed journals might not have been captured, and future studies could benefit from considering additional databases to supplement our findings.

One other limitation lies in the classification of OAD and TIC studies, which may be subject to variations in the interpretation of keywords. However, we optimized the keyword combinations during the search process in the WoS database and implemented Python filtering techniques, resulting in a relatively high level of accuracy in our classifications. The number of studies was further corroborated with a manual search on PubMed and a review of the classifications of the studies was done by critical care physicians.

Although there were no specific language restrictions, the nature of the search term being in English inadvertently excluded valuable contributions from non-English research. This may potentially limit the generalizability of our findings to a broader international audience. In future investigations, the inclusion of articles from various languages could offer a more comprehensive and diverse perspective on intensive care research.

Conclusions

This study has provided valuable insights into the expanding landscape of intensive care research through a comprehensive bibliometric analysis of a large number of publications by leveraging NLP technologies. While OAD studies have demonstrated significant promise, it is essential to view them as a complementary approach rather than a replacement for TIC studies. The unique strength of TIC studies lies in their ability to capture crucial qualitative information, which is essential for comprehensive and ethical decision-making. The integration of both OAD and TIC studies offers a synergistic approach to enriching our understanding of critical care medicine and advancing patient care outcomes. As NLP technology continues to advance, it holds the potential to offer a feasible and transformative alternative for literature review and bibliometric analysis.

Acknowledgments

We thank Dr Nicholas Brian Shannon for assistance with the manual review of the supervised keyword classification. This work was supported by the Duke-NUS Signature Research Programme, funded by the Ministry of Health, Singapore.

XSL•FO
RenderX

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. The complete set of code used in this study is readily available for download on GitHub [62].

Authors' Contributions

YK and NL played key roles in the conceptualization of the project. RY was responsible for formalizing the methodology and conducting data curation with the advisory of YK. YK contributed to the validation of the data, ensuring its relevance to the research objectives. RY took the lead in visualizing the data. Both YK and RY drafted the original manuscript. NL served as the project supervisor, overseeing the implementation, and providing valuable input in the writing, review, and editing phases.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms for open-access database (OAD) studies with the cutoff by the years of publications.
[DOCX File , 15 KB - jmir_v26i1e48330_app1.docx]

Multimedia Appendix 2
Top 20 journals ranked by total citation in which the open-access database and traditional intensive care studies were published. The average citation per article was obtained with the total citation/total number of articles. The citation counts were obtained from Web of Science.
[DOCX File , 12 KB - jmir_v26i1e48330_app2.docx]

Multimedia Appendix 3
Topic ID and topic family and the components and weightage in each of the categories.
[DOCX File , 18 KB - jmir_v26i1e48330_app3.docx]

Multimedia Appendix 4
Similarity matrix of 30 topics.
[DOCX File , 107 KB - jmir_v26i1e48330_app4.docx]

References

1. Kelly FE, Fong K, Hirsch N, Nolan JP. Intensive care medicine is 60 years old: the history and future of the intensive care unit. Clin Med (Lond) 2014;14(4):376-379 [FREE Full text] [doi: 10.7861/clinmedicine.14-4-376] [Medline: 25099838]
2. Cook D, Brower R, Cooper J, Brochard L, Vincent JL. Multicenter clinical research in adult critical care. Crit Care Med 2002;30(7):1636-1643. [doi: 10.1097/00003246-200207000-00039] [Medline: 12130991]
3. Rosenberg AL, Tripathi RS, Blum J. The most influential articles in critical care medicine. J Crit Care 2010;25(1):157-170 [FREE Full text] [doi: 10.1016/j.jcrc.2008.12.010] [Medline: 19327953]
4. Granholm A, Alhazzani W, Derde LPG, Angus DC, Zampieri FG, Hammond NE, et al. Randomised clinical trials in critical care: past, present and future. Intensive Care Med 2022;48(2):164-178 [FREE Full text] [doi: 10.1007/s00134-021-06587-9] [Medline: 34853905]
5. Markey KA, Ottridge R, Mitchell JL, Rick C, Woolley R, Ives N, et al. Assessing the efficacy and safety of an 11β-hydroxysteroid dehydrogenase type 1 inhibitor (AZD4017) in the idiopathic intracranial hypertension drug trial, IIH:DT: clinical methods and design for a phase II randomized controlled trial. JMIR Res Protoc 2017;6(9):e181 [FREE Full text] [doi: 10.2196/resprot.7806] [Medline: 28923789]
6. Verdonk F, Feyaerts D, Badenes R, Bastarache JA, Bouglé A, Ely W, et al. Upcoming and urgent challenges in critical care research based on COVID-19 pandemic experience. Anaesth Crit Care Pain Med 2022;41(5):101121 [FREE Full text] [doi: 10.1016/j.accpm.2022.101121] [Medline: 35781076]
7. Gurnani PK, Patel GP, Crank CW, Vais D, Lateef O, Akimov S, et al. Impact of the implementation of a sepsis protocol for the management of fluid-refractory septic shock: a single-center, before-and-after study. Clin Ther 2010;32(7):1285-1293. [doi: 10.1016/j.clinthera.2010.07.003] [Medline: 20678676]
8. Wang JL, Chin CS, Chang MC, Yi CY, Shih SJ, Hsu JY, et al. Key process indicators of mortality in the implementation of protocol-driven therapy for severe sepsis. J Formos Med Assoc 2009;108(10):778-787 [FREE Full text] [doi: 10.1016/S0929-6646(09)60405-8] [Medline: 19864198]
9. Levy MM, Evans LE, Rhodes A. The surviving sepsis campaign bundle: 2018 update. Crit Care Med 2018;46(6):997-1000 [FREE Full text] [doi: 10.1097/CCM.0000000000003119] [Medline: 29767636]

XSL•FO
RenderX

10. Fan E, Del Sorbo L, Goligher EC, Hodgson CL, Munshi L, Walkey AJ, et al. An official American Thoracic Society/European Society of Intensive Care Medicine/Society of Critical Care Medicine clinical practice guideline: mechanical ventilation in adult patients with acute respiratory distress syndrome. Am J Respir Crit Care Med 2017;195(9):1253-1263 [FREE Full text] [doi: 10.1164/rccm.201703-0548ST] [Medline: 28459336]
11. Goldfrad C, Vella K, Bion JF, Rowan KM, Black NA. Research priorities in critical care medicine in the UK. Intensive Care Med 2000;26(10):1480-1488. [doi: 10.1007/s001340000628] [Medline: 11126260]
12. Moore TJ, Heyward J, Anderson G, Alexander GC. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015-2017: a cross-sectional study. BMJ Open 2020;10(6):e038863 [FREE Full text] [doi: 10.1136/bmjopen-2020-038863] [Medline: 32532786]
13. Umscheid CA, Margolis DJ, Grossman CE. Key concepts of clinical trials: a narrative review. Postgrad Med 2011;123(5):194-204 [FREE Full text] [doi: 10.3810/pgm.2011.09.2475] [Medline: 21904102]
14. Maré IA, Kramer B, Hazelhurst S, Nhlapho MD, Zent R, Harris PA, et al. Electronic data capture system (REDCap) for health care research and training in a resource-constrained environment: technology adoption case study. JMIR Med Inform 2022;10(8):e33402 [FREE Full text] [doi: 10.2196/33402] [Medline: 36040763]
15. O'Hearn K, Gibson J, Krewulak K, Porteous R, Saigle V, Sampson M, et al. Consent models in Canadian critical care randomized controlled trials: a scoping review. Can J Anaesth 2022;69(4):513-526 [FREE Full text] [doi: 10.1007/s12630-021-02176-y] [Medline: 34907503]
16. Ghassemi M, Celi LA, Stone DJ. State of the art review: the data revolution in critical care. Crit Care 2015;19(1):118 [FREE Full text] [doi: 10.1186/s13054-015-0801-4] [Medline: 25886756]
17. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33(7):1123-1131. [doi: 10.1377/hlthaff.2014.0041] [Medline: 25006137]
18. Mlodzinski E, Wardi G, Viglione C, Nemati S, Crotty Alexander L, Malhotra A. Assessing barriers to implementation of machine learning and artificial intelligence-based tools in critical care: web-based survey study. JMIR Perioper Med 2023;6:e41056 [FREE Full text] [doi: 10.2196/41056] [Medline: 36705960]
19. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
20. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
21. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of FAIR (findable, accessible, interoperable, and reusable) data principles in health data stewardship practice: protocol for a scoping review. JMIR Res Protoc 2021;10(2):e22505 [FREE Full text] [doi: 10.2196/22505] [Medline: 33528373]
22. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
23. Bahadoran Z, Mirmiran P, Kashfi K, Ghasemi A. Importance of systematic reviews and meta-analyses of animal studies: challenges for animal-to-human translation. J Am Assoc Lab Anim Sci 2020;59(5):469-477 [FREE Full text] [doi: 10.30802/AALAS-JAALAS-19-000139] [Medline: 32727637]
24. Haidich AB. Meta-analysis in medical research. Hippokratia 2010;14(Suppl 1):29-37 [FREE Full text] [Medline: 21487488]
25. Thompson DF, Walker CK. A descriptive and historical review of bibliometrics with applications to medical sciences. Pharmacotherapy 2015;35(6):551-559. [doi: 10.1002/phar.1586] [Medline: 25940769]
26. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res 2021;133:285-296 [FREE Full text] [doi: 10.1016/j.jbusres.2021.04.070]
27. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 2015;16(Suppl 13):S8 [FREE Full text] [doi: 10.1186/1471-2105-16-S13-S8] [Medline: 26424364]
28. Doanvo A, Qian X, Ramjee D, Piontkivska H, Desai A, Majumder M. Machine learning maps research needs in COVID-19 literature. Patterns (N Y) 2020;1(9):100123 [FREE Full text] [doi: 10.1016/j.patter.2020.100123] [Medline: 32959032]
29. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on May 24 2019 2018. [doi: 10.48550/arXiv.1810.04805]
30. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. ArXiv. Preprint posted online on March 11 2022 2022. [doi: 10.48550/arXiv.2203.05794]
31. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text]
32. Delaney A, Bagshaw SM, Ferland A, Manns B, Laupland KB, Doig CJ. A systematic evaluation of the quality of meta-analyses in the critical care literature. Crit Care 2005;9(5):R575-R582 [FREE Full text] [doi: 10.1186/cc3803] [Medline: 16277721]
33. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine joint data science collaboration:

XSL•FO
RenderX

the Amsterdam university medical centers database (AmsterdamUMCdb) example. Crit Care Med 2021;49(6):e563-e577 [FREE Full text] [doi: 10.1097/CCM.0000000000004916] [Medline: 33625129]
34. Faltys M, Zimmermann M, Lyu X, Hüser M, Hyland S, Rätsch G, et al. HiRID, a high time-resolution ICU dataset. PhysioNet. 2021. URL: https://physionet.org/content/hirid/1.1.1/ [accessed 2024-04-02]
35. Zeng X, Yu G, Lu Y, Tan L, Wu X, Shi S, et al. PIC, a paediatric-specific intensive care database. Sci Data 2020;7(1):14 [FREE Full text] [doi: 10.1038/s41597-020-0355-4] [Medline: 31932583]
36. Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Comput Geosci 1993;19(3):303-342. [doi: 10.1016/0098-3004(93)90090-R]
37. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv. Preprint posted online on September 18 2020 2018. [doi: 10.48550/arXiv.1802.03426]
38. Murray CJL. COVID-19 will continue but the end of the pandemic is near. Lancet 2022;399(10323):417-419 [FREE Full text] [doi: 10.1016/S0140-6736(22)00100-3] [Medline: 35065006]
39. Mark R. The story of MIMIC. In: Secondary Analysis of Electronic Health Records. Cham, Switzerland: Springer International Publishing; 2016:43-49.
40. Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. JMIR Med Inform 2021;9(5):e21347 [FREE Full text] [doi: 10.2196/21347] [Medline: 33949961]
41. Liu D, Zheng M, Sepulveda NA. Using artificial neural network condensation to facilitate adaptation of machine learning in medical settings by reducing computational burden: model design and evaluation study. JMIR Form Res 2021;5(12):e20767 [FREE Full text] [doi: 10.2196/20767] [Medline: 34889747]
42. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019;23(1):112 [FREE Full text] [doi: 10.1186/s13054-019-2411-z] [Medline: 30961662]
43. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018;6(12):905-914. [doi: 10.1016/S2213-2600(18)30300-X] [Medline: 30274956]
44. Chen H, Zhu Z, Zhao C, Guo Y, Chen D, Wei Y, et al. Central venous pressure measurement is associated with improved outcomes in septic patients: an analysis of the MIMIC-III database. Crit Care 2020;24(1):433 [FREE Full text] [doi: 10.1186/s13054-020-03109-9] [Medline: 32665010]
45. Faust L, Feldman K, Chawla NV. Examining the weekend effect across ICU performance metrics. Crit Care 2019;23(1):207 [FREE Full text] [doi: 10.1186/s13054-019-2479-5] [Medline: 31171026]
46. Davidson S, Villarroel M, Harford M, Finnegan E, Jorge J, Young D, et al. Vital-sign circadian rhythms in patients prior to discharge from an ICU: a retrospective observational analysis of routinely recorded physiological data. Crit Care 2020;24(1):181 [FREE Full text] [doi: 10.1186/s13054-020-02861-2] [Medline: 32345354]
47. Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inform 2022;126:103980 [FREE Full text] [doi: 10.1016/j.jbi.2021.103980] [Medline: 34974189]
48. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ, GRADE Working Group. What is "quality of evidence" and why is it important to clinicians? BMJ 2008;336(7651):995-998 [FREE Full text] [doi: 10.1136/bmj.39490.551019.BE] [Medline: 18456631]
49. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
50. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Crit Care Med 2021;49(11):e1063-e1143 [FREE Full text] [doi: 10.1097/CCM.0000000000005337] [Medline: 34605781]
51. Volovici V, Syn NL, Ercole A, Zhao JJ, Liu N. Steps to avoid overuse and misuse of machine learning in clinical research. Nat Med 2022;28(10):1996-1999 [FREE Full text] [doi: 10.1038/s41591-022-01961-6] [Medline: 36097217]
52. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
53. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? J Med Ethics 2022;48(9):581-585 [FREE Full text] [doi: 10.1136/medethics-2020-107102] [Medline: 34006600]
54. Gillett GR. Intensive care unit research ethics and trials on unconscious patients. Anaesth Intensive Care 2015;43(3):309-312. [doi: 10.1177/0310057X1504300304] [Medline: 25943602]
55. Aulisio MP, Chaitin E, Arnold RM. Ethics and palliative care consultation in the intensive care unit. Crit Care Clin 2004;20(3):505-523, x-xi. [doi: 10.1016/j.ccc.2004.03.006] [Medline: 15183216]

XSL•FO
RenderX

56. Raudenská J, Steinerová V, Javůrková A, Urits I, Kaye AD, Viswanath O, et al. Occupational burnout syndrome and post-traumatic stress among healthcare professionals during the novel coronavirus disease 2019 (COVID-19) pandemic. Best Pract Res Clin Anaesthesiol 2020;34(3):553-560 [FREE Full text] [doi: 10.1016/j.bpa.2020.07.008] [Medline: 33004166]
57. Davidson JE, Jones C, Bienvenu OJ. Family response to critical illness: postintensive care syndrome-family. Crit Care Med 2012;40(2):618-624. [doi: 10.1097/CCM.0b013e318236ebf9] [Medline: 22080636]
58. White DB, Angus DC, Shields AM, Buddadhumaruk P, Pidro C, Paner C, et al. A randomized trial of a family-support intervention in intensive care units. N Engl J Med 2018;378(25):2365-2375 [FREE Full text] [doi: 10.1056/NEJMoa1802637] [Medline: 29791247]
59. Popoff B, Occhiali É, Grangé S, Bergis A, Carpentier D, Tamion F, et al. Trends in major intensive care medicine journals: a machine learning approach. J Crit Care 2022;72:154163. [doi: 10.1016/j.jcrc.2022.154163] [Medline: 36209696]
60. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61]
61. Birkle C, Pendlebury DA, Schnell J, Adams J. Web of science as a data source for research on scientific and scholarly activity. Quant Sci Stud 2020;1(1):363-376 [FREE Full text] [doi: 10.1162/qss_a_00018]
62. GitHub. URL: https://github.com/YangRui525/Comparing-OAD-and-TIC-Studies

Abbreviations

AmsterdamUMCdb: Amsterdam University Medical Centers Database
BERT: bidirectional encoder representations from transformers
EHR: electronic health record
eICU-CRD: eICU Collaborative Research Database
FAIR: findable, accessible, interoperable, and reusable
GRADE: Grading of Recommendations, Assessment, Development, and Evaluations
HiRID: High Time Resolution ICU Dataset
ICU: intensive care unit
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
NLP: natural language processing
OAD: open-access database
TIC: traditional intensive care
UMAP: uniform manifold approximation and projection
WoS: Web of Science

Edited by A Mavragani; submitted 19.04.23; peer-reviewed by D Chrimes, S Pesälä; comments to author 14.07.23; revised version received 01.08.23; accepted 14.01.24; published 17.04.24.

Please cite as:
Ke Y, Yang R, Liu N
Comparing Open-Access Database and Traditional Intensive Care Studies Using Machine Learning: Bibliometric Analysis Study
J Med Internet Res 2024;26:e48330
URL: https://www.jmir.org/2024/1/e48330
doi:10.2196/48330
PMID:38630522

©Yuhe Ke, Rui Yang, Nan Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of a Digital Mental Health Intervention on Perceived Stress and Rumination in Adolescents Aged 13 to 17 Years: Randomized Controlled Trial

Eliane M Boucher[1], PhD; Haley Ward[1], BA; Cynthia J Miles[1], BS; Robert D Henry[1], PhD; Sarah Elizabeth Stoeckl[1], BS

Twill, New York, NY, United States

Corresponding Author:
Eliane M Boucher, PhD
Twill
114 Fifth Avenue
10th Floor
New York, NY, 10011
United States
Phone: 1 432 258 5233
Email: eliane@twill.health

Abstract

Background: Although adolescents report high levels of stress, they report engaging in few stress management techniques. Consequently, developing effective and targeted programs to help address this transdiagnostic risk factor in adolescence is particularly important. Most stress management programs for adolescents are delivered within schools, and the evidence for these programs is mixed, suggesting a need for alternative options for stress management among adolescents.

Objective: The aim of the study is to test the short-term effects of a self-guided digital mental health intervention (DMHI) designed for adolescents on perceived stress and rumination (ie, brooding).

Methods: This was a 12-week, 2-arm decentralized randomized controlled trial of adolescents aged 13 to 17 years who presented with elevated levels of perceived stress and brooding. Participants were randomly assigned to engage with a self-guided DMHI (Happify for Teens) or to a waitlist control. Participants assigned to the intervention group were given access to the program for 12 weeks. Happify for Teens consists of various evidence-based activities drawn from therapeutic modalities such as cognitive behavioral therapy, positive psychology, and mindfulness, which are then organized into several programs targeting specific areas of concern (eg, Stress Buster 101). Participants in the waitlist control received access to this product for 12 weeks upon completing the study. Participants in both groups completed measures of perceived stress, brooding, optimism, sleep disturbance, and loneliness at baseline, 4 weeks, 8 weeks, and 12 weeks. Changes in outcomes between the intervention and waitlist control groups were assessed using repeated-measures multilevel models.

Results: Of the 303 participants included in data analyses, 132 were assigned to the intervention and 171 to the waitlist. There were significantly greater improvements in the intervention condition for perceived stress (intervention: B=–1.50; 95% CI –1.82 to –1.19; P<.001 and control: B=–0.09; 95% CI –0.44 to 0.26; P=.61), brooding (intervention: B=–0.84; 95% CI –1.00 to –0.68; P<.001 and control: B=–0.30; 95% CI –0.47 to –0.12; P=.001), and loneliness (intervention: B=–0.96; 95% CI –1.2 to –0.73; P<.001 and control: B=–0.38; 95% CI: –0.64 to –0.12; P=.005) over the 12-week study period. Changes in optimism and sleep disturbance were not significantly different across groups (Ps≥.096).

Conclusions: Happify for Teens was effective at reducing perceived stress, rumination, and loneliness among adolescents over 12 weeks when compared to a waitlist control group. Our data reveal the potential benefits of DMHIs for adolescents, which may present a more scalable, destigmatized, and cost-effective alternative to school-based programs.

Trial Registration: ClinicalTrials.gov NCT04567888; https://clinicaltrials.gov/ct2/show/NCT04567888

International Registered Report Identifier (IRRID): RR2-10.2196/25545

(J Med Internet Res 2024;26:e54282) doi:10.2196/54282

XSL•FO
RenderX

KEYWORDS

digital intervention; adolescents; adolescent; stress management; mental health; mobile phone; mobile health; mHealth; teen; teens; teenager; teenagers; stress; mental health; rumination; brooding; randomized controlled trial; RCT; randomized; controlled trial; controlled trials; digital mental health intervention; DMHI; digital health

Introduction

Background

The onset of many mental health disorders occurs in childhood and adolescence, and these disorders often continue into adulthood [1,2]. Already, approximately 20% of adolescents worldwide have at least 1 mental health disorder [3]; however, rates of mental illness among adolescents are increasing [4] and at steeper rates than among adults [5]. Addressing mental health in adolescence is critical to reducing the prevalence and consequences of mental health disorders in adulthood.

Despite public health efforts to increase access to mental health services for youths, the use of these services by adolescents with mental health disorders remains low [6]. Research suggests that only 50% of US adolescents with mental illness seek treatment [7], and as few as 25% actually receive treatment [8]. Paradoxically, national trends indicate that, overall, the use of outpatient mental health services by US adolescents, including both psychotherapy and psychotropic medications, has increased over time [9]. However, this increase is primarily attributed to adolescents with less severe or no mental health impairment [9]. This is problematic, given the shortage of mental health service providers for children and adolescents, particularly in rural and low-income areas [10]. The increased use of these services by adolescents likely places a strain on the available services for youths and, in turn, adolescents, with more serious mental health disorders may have difficulties receiving necessary treatment. Interventions that are focused on promoting flourishing and preventing future dysfunction and impairment may help to reduce the burden on mental health services by offering alternative services to adolescents with little or no mental health impairment [11].

Stress as a Risk Factor in Adolescence

Adolescence is a developmental stage where individuals may be particularly sensitive to stress due to shifts in hypothalamic-pituitary-adrenal axis reactivity, leading to more intense hormonal responses to stressors, particularly in older adolescents [12]. Over the last decade, researchers have become increasingly interested in adolescent stress as an important transdiagnostic risk factor. Although animal models suggest that predictable, chronic mild stress levels in adolescence can actually increase resilience in adulthood [13], research suggests many adolescents are coping with more severe stress levels. In one study, 22% of adolescents aged 15 to 17 years reported moderate to severe levels of perceived stress [14]. In addition, national surveys conducted by the American Psychological Association [15,16] have found that adolescents report higher levels of stress than most adults, particularly during the school year.

Such high levels of stress have been linked to lower life satisfaction [17], poorer academic performance [18,19], cigarette smoking [20], emotional eating [21], poorer diet [22], more frequent subjective health complaints (eg, headaches, fatigue, and sleep difficulties) [23], as well as internalizing symptoms including depression [24,25] and anxiety [25]. Interpersonal stress, in particular, predicts the onset of a first major depressive episode among adolescents [26]. Thus, addressing stress in adolescents is critical to avoiding more long-term problems.

However, 25% of adolescents self-report not doing enough to manage their stress [16]. Approximately 55% of adolescents report setting aside time for stress management “a few times a month or less,” 13% report never setting aside time for stress management, and just 5% report having seen a mental health professional about their stress [15]. Based on these data, the American Psychological Association noted the need for opportunities to help adolescents address and cope with their stress in order “to break this unhealthy legacy of stress in America” [15].

Although there has been an increase in interventions targeting adolescent stress [27], evidence-based interventions for stress are much less common than those for anxiety and depression [28]. Furthermore, most interventions for stress are school-based, and the evidence for these interventions is mixed. In one systematic review of stress management interventions, only 58% of the reviewed studies found significant improvements in physiological indicators of stress (eg, blood pressure) or self-reported stress [29]. More recent meta-analyses of school-based programs further suggest that these programs may only be effective for school rather than social stress [27]. Given these findings, we need more research exploring the effectiveness of stress management programs with adolescents, particularly those that could be delivered outside schools.

Negative Cognitions and Stress

In addition, given that stress in adolescence stems, at least in part, from feelings of helplessness and negative affect [30], effective interventions for adolescent stress may need to target underlying negative cognitions as well as perceived stress. However, recommended approaches to addressing stress tend to focus on stress management training, relaxation training, and problem-solving skill or decision-making skill training [27,28,31], which may not adequately address the related negative cognitions and cognitive processes.

Indeed, there is evidence that improving the content of adolescents’ cognitions can help attenuate the consequences of stress. For example, interpersonal stress predicts depressive symptoms in adolescents only when coupled with negative cognitions [26] or persistent low positive affect [32], whereas self-compassion buffers against the negative consequences of stress on internalizing symptoms [25]. In other words, addressing both negative cognitions and stress together may be particularly important to mitigate the negative consequences of stress [26].

XSL•FO
RenderX

Rumination—a pattern of repeatedly thinking about the causes, consequences, and symptoms of one's negative affect [33]—is one cognitive process that may be particularly relevant to stress and future risk of mental health concerns. Because rumination can occur in response to stressful events, it also plays a role in the negative consequences of chronic stress. Longitudinal research has shown that more frequent stressful life events predict increased rumination among both adults and adolescents, and that rumination mediates the relationship between stressful life events and anxiety symptoms in adolescents [34]. Other research has shown that adolescents who ruminate in response to stress are at greater risk of depression and substance misuse [35].

Rumination may also increase perceived stress. In one study, rumination exacerbated the relationship between life hassles and depression, anxiety, and stress in adolescents [36]. Other research shows that rumination increases adolescents' likelihood of experiencing interpersonal stress which, in turn, predicts more internalizing symptoms [37]. Therefore, applying approaches that are more traditionally applied to addressing negative cognitive processes like rumination, such as cognitive restructuring, self-monitoring, acceptance strategies, and attention control training [31], may be helpful in reducing stress as well.

These findings suggest there may be a common pathway to reducing both rumination and perceived stress in adolescents. Therefore, identifying interventions that can address both would be beneficial as they would target 2, rather than 1, important transdiagnostic risk factors in the adolescent population.

Objectives

The aim of this study was to test the impact of a digital mental health intervention (DMHI) designed for adolescents called Happify for Teens (Twill Inc). This program has been adapted for adolescents between 13 and 17 years of age from another DMHI developed for adults (Happify; Twill Inc). Happify for Teens consists of gamified versions of evidence-based activities adapted from various therapeutic approaches including cognitive behavioral therapy (CBT) [38], mindfulness-based stress reduction [39], and positive psychology [40,41]. Consequently, these programs integrate a variety of recommended approaches to addressing negative cognitions [31] that also help manage stress [42-45].

Observational studies of real-world Happify users have shown significant improvements in subjective well-being and anxiety among adults after 6-8 weeks of use, with greater gains among users who completed more activities [46-48]. Randomized controlled trials (RCTs) with adult participants have similarly shown improvements in depressive symptoms, anxiety, and resilience among those who completed at least 2 activities per week relative to a control group [49,50]. However, Happify for Teens has yet to be tested empirically. Therefore, in this study, we recruited adolescents aged 13 to 17 years who reported elevated stress and rumination. Participants were randomly assigned to have access to Happify for Teens for 12 weeks or to a corresponding waitlist control group. Changes in perceived stress after 4, 8, and 12 weeks across the 2 groups were compared. We also examined changes in brooding (the maladaptive component of rumination) as a secondary outcome and in sleep disturbance, loneliness, and optimism as exploratory outcomes.

Methods

Study Design

This was a 12-week, 2-arm RCT of adolescents aged 13 to 17 years residing in the United States with elevated levels of perceived stress and brooding (Multimedia Appendix 1).

Ethical Considerations

A study protocol was previously published [51], and the trial was registered with ClinicalTrials.gov (NCT04567888). Study procedures were reviewed and approved by IntegReview (HLS-20), an independent institutional review board, which was acquired by Advarra after the study was launched. Parental consent was obtained electronically as part of the prescreening questionnaire, and participant assent was obtained electronically before participants began the baseline assessment, after qualifying for the study. The use of e-signatures was approved by IntegReview. Participants received US $20 for completing each of the 4-, 8-, and 12-week assessments and a US $20 bonus if they completed all 4 assessments. Participants were not compensated for engaging in the intervention. All data were deidentified for analysis, and responses to the baseline, 4-, 8-, and 12-week assessments were matched using a user-generated ID code, which was entered at the beginning of each assessment.

Participants

The study was advertised to parents or guardians via social media, emails to Happify users, within US schools, and using a snowballing recruitment method. Parents or guardians were directed to complete a prescreening questionnaire to report their child's age, country of residence, and previous Happify use and to provide electronic consent for each eligible child. Adolescents who were qualified based on the prescreening questionnaire were then invited to complete a screening questionnaire via email; adolescents were eligible to participate if they reported being 13 to 17 years of age, residing in the United States, having never used Happify, and reporting elevated stress (Perceived Stress Scale [52] scores >14) and brooding (Ruminative Response Scale—Short Form—Brooding Subscale scores [53] ≥10).

Procedure

Eligible participants were directed to complete the baseline assessment and provide electronic assent to participate in the study. To enroll in the study, participants had to provide electronic assent, complete the baseline assessment, and pass at least 1 attention check.

Once enrolled, participants were randomized to condition; randomization was completed automatically by Qualtrics (Qualtrics) using block randomization upon completing the baseline assessment and meeting final eligibility criteria. Those assigned to the intervention condition received instructions to download the Happify mobile app and were directed to the Happify for Teens platform after creating an account. Like the original Happify program, Happify for Teens consists of digital

XSL•FO
RenderX

versions of evidence-based activities, which are categorized into 6 different skills: *Savor* (activities focused on developing mindfulness), *thank* (activities focused on gratitude), *Aspire* (activities focused on optimism, goal-setting, and finding meaning and purpose), *Give* (activities focused on acts of kindness, forgiveness, or promoting prosocial behavior), *Empathize* (activities fostering self-compassion and perspective-taking), and *Revive* (activities focused on physical health). In Happify for Teens, these activities were modified and then reviewed by a panel of adolescents to ensure the language and content were both appropriate and relevant. As in the original Happify program, activities are organized into "tracks," 4-week programs intended to address a specific area of concern, like increasing one's confidence or reducing stress. Users are able to switch tracks at any time and can also access activities outside these tracks, on demand, through the Instant Play feature. More information on the Happify for Teens program, including screenshots, can be found elsewhere [51].

Participants were given access to all Happify for Teens tracks; however, the same track (*Stop the Worry Cycle*) was shown as the featured track for all participants and recommended upon signing up for the program. Participants received no instructions on how much to engage with the intervention but received push notifications and weekly engagement emails as part of the intervention. We also contacted participants who completed no activities over 7 days via email or SMS text message (based on participant preference). Participants had access to the intervention for the full study period but stopped receiving push notifications or emails regarding use after 8 weeks; this was done to provide a period of time where engagement with the program may be more naturalistic, given that RCT procedures tend to be associated with inflated engagement with DMHIs [54]. Participants in both conditions also received regular emails or SMS text messages (depending on participant preference) about their progress to boost retention. These emails were identical regardless of condition and reminded participants of an upcoming assessment (when appropriate) or asked if the participant had any concerns. In total, 5 such emails were sent throughout the 12-week study period. Waitlist participants received access to the intervention after 12 weeks.

Every 4 weeks, participants were prompted to complete outcome measures via Qualtrics. Participants were contacted via email or SMS text message (based on participant preference) if an assessment was not completed within 7 days of its scheduled date.

Measurements and Outcomes

Primary Outcome: Perceived Stress

The Perceived Stress Scale [52] consists of 10 items asking participants how often they felt each feeling or thought in the previous month (eg, "How often have you felt nervous and 'stressed'?"). Items are rated on a scale from 0 (never) to 4 (very often), and ratings are summed, so higher scores indicate greater perceived stress. Across the 4 time points, internal reliabilities ranged from acceptable to good (αs=.77 to .88).

Secondary Outcome: Brooding

The Brooding Subscale of the Ruminative Response Scale [53] asks respondents to indicate how often they engage in 5 behaviors (eg, "Think 'Why can't I handle things better'?") on a scale from 1 (almost never) to 4 (almost always). Participants in this study were asked to reflect on the previous month. Ratings are summed, so higher scores indicate more brooding. Internal reliability values were fair to acceptable (αs=.61 to .78).

Exploratory Outcomes

Sleep Disturbance

The PROMIS (Patient-Reported Outcomes Measurement Information System) Pediatric Sleep Disturbance Scale—Short Form 4a [55] measures the extent to which participants experienced sleep disturbances over the previous 7 days (eg, "In the past 7 days, I had trouble sleeping"). Items are rated on a scale from 1 (never) to 5 (always), and ratings are summed, so higher scores indicate more sleep disturbance. Internal consistency ranged from good to excellent (αs=.89 to .91).

Loneliness

The Roberts UCLA Loneliness Scale [56] consists of 8 items (eg, "I lack companionship"). Respondents rate each statement on a 4-point scale from 0 (never) to 3 (often), and ratings are summed, so higher scores indicate greater loneliness. Internal reliability values ranged were all good (αs=.82 to .85).

Optimism

The Life Orientation Test—Revised [57] consists of 10 items (eg, "I'm always optimistic about my future"). Respondents indicate their level of agreement with each statement on a 5-point scale from 0 (strongly disagree) to 4 (strongly agree). After dropping 4 filler items, ratings are summed, so higher scores indicate more optimism. Internal reliabilities were acceptable (αs=.72 to .77).

Statistical Analysis

To assess whether changes in outcomes differed across groups, we created repeated-measures multilevel models (MLMs) for each outcome. MLMs were computed in R software (R Foundation for Statistical Computing) using the *lme4* package [58] and restricted maximum likelihood estimation. All models included a random slope for time ($\Delta\chi^2_2$s≥10.3; *P*s≤.01) and cross-level interactions for condition (waitlist and intervention)×time (baseline, 4 weeks, 8 weeks, and 12 weeks). We examined regression assumptions (eg, residual normality and influential observations), and final models were run with the small number of influential outliers removed (2-4 observations, depending on the outcome). Note that these analyses differ from those in the published protocol [51] because data violated the sphericity assumption for repeated-measures ANOVAs. Additionally, MLMs can manage missing outcome data without imputation, allowing us to retain more participants. Due to power constraints, we focus only on analyses testing changes in primary, secondary, and exploratory outcomes across conditions.

XSL•FO
RenderX

In addition, because careless responding is problematic with web-based surveys and may artificially increase relationships between variables, we used two a priori mechanisms for identifying low-quality data: (1) failing 3 or more attention checks in any given assessment and (2) completing an assessment at a rate faster than 1 second per item [59]. Specific assessments that met either of these criteria were dropped rather than dropping the participant altogether, and analyses were rerun without these assessments as a sensitivity analysis.

Results

Sample

Study enrollment began on June 3, 2021, and continued until December 18, 2021. Overall, 2332 parents or guardians completed the screening questionnaire, of which 1604 met initial inclusion criteria and provided consent. In total, 631 adolescents proceeded to complete the screener, of which 353 met inclusion criteria, completed the baseline assessment, and were enrolled into the study and assigned to condition (Happify for Teens: n=178 and waitlist: n=175). However, 22 participants in the intervention condition did not create a Happify for Teens account and were withdrawn from the study.

Note that this sample is smaller than the target sample size reported in the published protocol [51]. As described in that protocol, however, power analyses indicated that 200 participants would be sufficient for 80% power to detect a small effect for our primary outcome. Due to the high levels of dropout observed with digital interventions [60], including with adolescents, we assumed a 75% attrition rate and set a target sample size of 800 participants. Because attrition rates were much lower than expected (assessment completion rates ranged from 78.95% to 87.13%), we decided to stop enrollment once 100 participants per condition completed the 12-week assessment to avoid oversampling, resulting in a smaller sample.

Data Exclusion

Of the 331 participants who completed the study, 28 were excluded from data analyses (Figure 1). Of note, 22 participants in the intervention condition failed to complete any activities throughout the study period (and thus were not exposed to the intervention). In total, 5 participants were excluded for providing odd or missing birth years on the baseline assessment, and 1 additional participant was excluded for having duplicate responses to assessments. This resulted in a total of 303 participants included in our analyses (Happify for Teens: n=132 and waitlist n=171). Because the failure to engage with the intervention is subject to self-selection bias, and thus excluding these participants from the analysis means the sample is not fully randomized, we also conducted sensitivity analyses including participants who signed up for the program but completed no activities (n=22) as well as those participants randomized to the Happify for Teens condition who never created an account on the platform (n=22).

XSL•FO
RenderX

Figure 1. Flow of the participants in the randomized controlled trial.

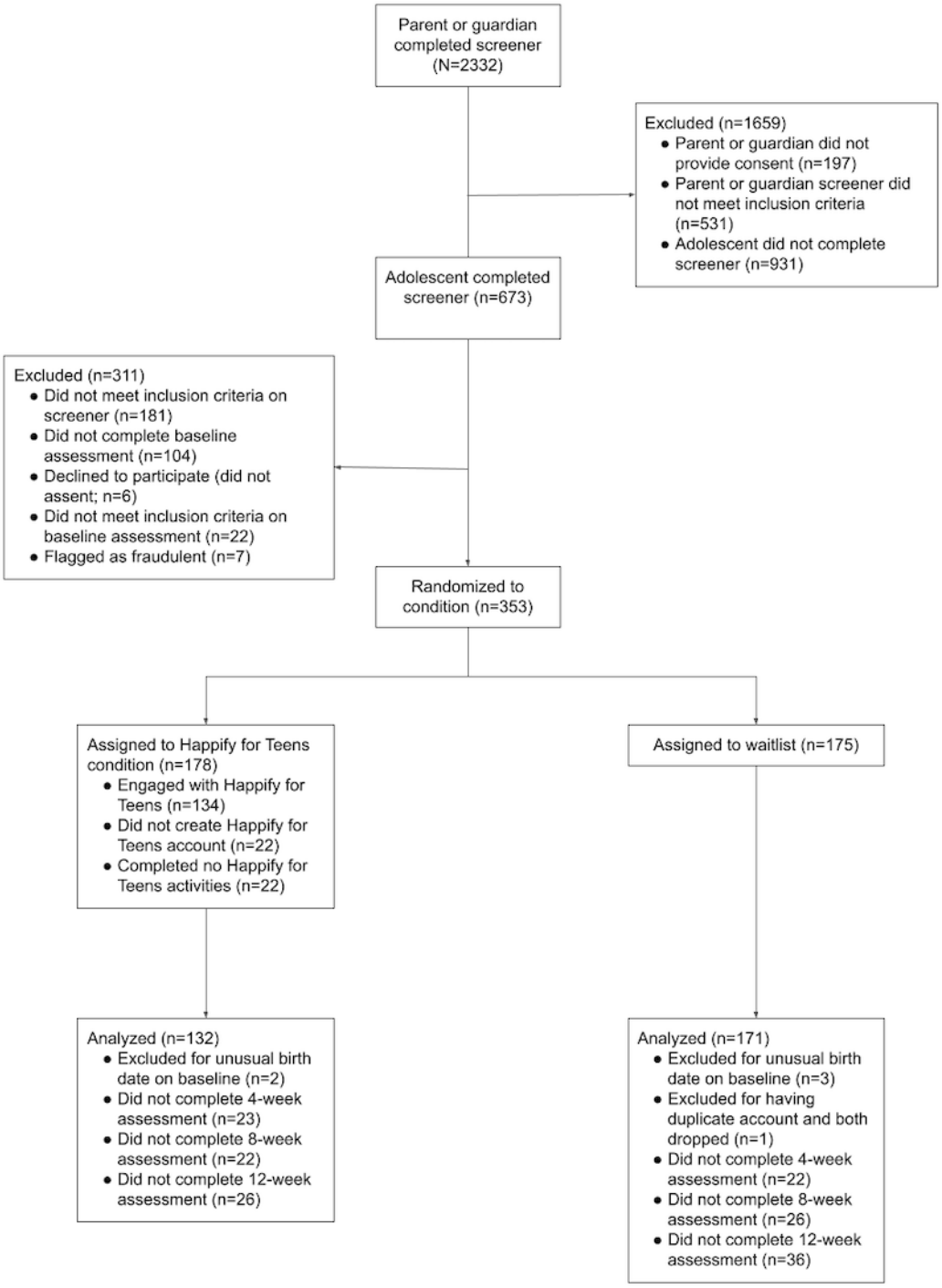

Sample Demographics

Sample demographics are presented in detail in Table 1. The majority of the sample identified as female (n=202, 66.7%) and as White (n=262, 86.5%). Fewer participants reported being 13 years of age relative to other age groups, but all qualifying age groups were represented in our sample. When starting the study, the majority of participants (n=226, 74.6%) were enrolled in school full-time, in person.

XSL•FO
RenderX

Table 1. Demographic characteristics for analytical sample.

	Total sample (N=303), n (%)	Happify for Teens (n=132), n (%)	Waitlist (n=171), n (%)
Gender[a]			
Female	202 (66.7)	94 (71.2)	108 (63.2)
Male	76 (25.1)	27 (20.4)	49 (28.6)
Neither	24 (7.9)	11 (8.3)	13 (7.6)
Not reported	1 (0.003)	0 (0)	1 (0.6)
Age[b] (years)			
13	38 (12.5)	14 (10.6)	24 (14)
14	75 (24.7)	33 (25)	42 (24.6)
15	53 (17.5)	22 (16.7)	31 (18.1)
16	69 (22.8)	27 (20.4)	42 (24.6)
17	68 (22.4)	36 (27.3)	32 (18.7)
Schooling situation			
Not attending school	6 (2)	2 (1.5)	4 (2.3)
Attending school, in person	226 (74.6)	97 (73.5)	129 (75.4)
Attending school, virtual	22 (7.3)	8 (6.1)	14 (8.2)
Attending school, hybrid	22 (7.3)	13 (9.8)	9 (5.3)
Homeschooled	27 (8.9)	12 (9.1)	15 (8.8)
Race[c]			
American Indian or Alaska Native	9 (3)	4 (3)	5 (2.9)
Asian or Asian American	14 (4.6)	10 (7.6)	4 (2.3)
Black or African American	32 (10.6)	10 (7.6)	22 (12.9)
Hispanic or Latinx	33 (10.9)	12 (9.1)	21 (12.3)
Native Hawaiian, other Pacific Islander	2 (0.7)	0 (0)	2 (1.2)
White	262 (86.5)	117 (88.6)	145 (84.8)
Other	9 (3)	3 (2.3)	6 (3.5)
Current use of wellness apps			
Yes	35 (11.6)	15 (11.4)	20 (11.7)
No	268 (88.4)	117 (88.6)	151 (88.3)

[a]Gender was assessed using an open-ended question (What is your gender?); responses clearly identifying male or female were coded as such, and other responses were combined into the “neither” category.

[b]Age was estimated using participants’ reported birth year at baseline. Adolescents reported their age during screening, and those reporting ages 13 years and younger or 17 years and older were disqualified. Nevertheless, some participants may have turned 18 years before the baseline assessment or during the study period.

[c]Race was a multioption question, consequently percentages across categories sum to more than 100%.

Most participants were not currently using any self-care or wellness apps (n=268, 88.4%). Among those currently using these programs, there was a great deal of variability in programs, but *I am Sober* (n=4) and *Calm* (n=3) were reported most frequently. Additionally, although we used current Happify members as one recruitment mechanism, few participants reported their parents used Happify (n=25, 8.3%).

Happify for Teens Use

Of the participants assigned to the Happify for Teens condition who signed up for the program, 85.7% (132/154) completed at least 1 activity over the 12-week study period. The mean number of activities completed among these participants was 46.24 (SD 53.02), ranging from 1 to 235 activities.

Consistent with other DMHI research, engagement dropped over the course of the 12-week study period. More specifically, among participants who completed at least 1 activity, 3 (2.3%) completed no activities during the first 4 weeks, whereas 25 (18.9%) completed no activities between week 5 and week 8, and 60 (45.5%) completed no activities between week 9 and week 12. Similarly, mean levels of engagement dropped over time, with an average of 22.49 (SD 22.29) activities completed

XSL•FO
RenderX

between week 1 and week 4 compared to an average of 15.25 (SD 21.11) activities completed between week 5 and week 8 and 8.5 (SD 14.49) activities completed between week 9 and week 12.

Although participants were not given explicit instructions on how often to engage with the Happify for Teens platform or how many activities to complete, previous research with adults suggests that completing an average of 2 or more activities is optimal [49,50,61,62]. Among participants who completed at least 1 activity, 75.8% (n=100) engaged at the optimal level during the first 4 weeks, 50% (n=66) engaged at the optimal level between week 5 and week 8, and 29.5% (n=39) engaged at the optimal level between week 9 and week 12. These levels are consistent or better than what has been observed in previous research using adults [49,61,62].

Changes in Outcomes

Overview

Descriptive statistics for each outcome by assessment period are presented in Table 2. All final MLMs reported below included a random slope for time ($\Delta\chi^2_2$s≥10.3; *P*s≤.01). Only 4 participants were flagged for attention checks (1 on 3 of 4 assessments, 1 on 2 assessments, and 2 on just 1 assessment). No participants were flagged for completion rate. Primary and secondary outcomes did not change meaningfully in these analyses when these assessments were excluded. In addition, outcomes did not change meaningfully in sensitivity analyses including participants who had been randomly assigned to the intervention condition but never engaged with the Happify for Teens program. More specifically, no main effects of time or condition changed for any of the primary, secondary, or exploratory outcomes nor did any of the time×condition interactions. However, some nonsignificant effects for control variables did become significant in the sensitivity analyses. The effect of age became significant for perceived stress (*P*=.04), such that stress increased with age. The effect of gender also became significant for both optimism (*P*=.03) and loneliness (*P*=.02), such that adolescents who identified as male reported higher levels of optimism but also higher levels of loneliness compared to other gender identities.

XSL•FO
RenderX

Table 2. Descriptive statistics for all outcomes by assessment period and condition.

	α	Total sample, mean (SD)	Happify for Teens, mean (SD)	Waitlist control, mean (SD)
Perceived stress[a]				
Baseline	.77	25.73 (4.77)	26.33 (4.43)	25.26 (4.97)
4 weeks	.84	23.55 (5.82)	22.01 (6.10)	24.67 (5.36)
8 weeks	.86	23.47 (6.27)	21.95 (6.20)	24.61 (6.10)
12 weeks	.88	22.91 (6.62)	20.96 (6.57)	24.44 (6.27)
Brooding[b]				
Baseline	.61	14.29 (2.47)	14.29 (2.36)	14.29 (2.56)
4 weeks	.75	13.32 (3.18)	12.53 (3.24)	13.89 (3.01)
8 weeks	.72	12.65 (2.93)	11.94 (2.89)	13.20 (2.85)
12 weeks	.78	12.68 (3.22)	11.68 (3.18)	13.46 (3.03)
Sleep disturbance [c]				
Baseline	.89	12.60 (3.87)	12.51 (4.01)	12.67 (5.04)
4 weeks	.91	11.88 (4.07)	11.32 (4.10)	12.29 (4.01)
8 weeks	.90	11.46 (3.92)	10.86 (3.63)	11.92 (4.09)
12 weeks	.90	11.31 (4.15)	10.88 (3.76)	11.65 (4.42)
Loneliness[d]				
Baseline	.83	21.15 (4.93)	21.25 (4.80)	21.06 (5.04)
4 weeks	.82	20.29 (4.82)	19.69 (5.15)	20.72 (4.52)
8 weeks	.85	19.62 (5.08)	19.14 (5.33)	19.98 (4.87)
12 weeks	.82	19.09 (4.68)	18.05 (4.98)	19.91 (4.28)
Optimism[e]				
Baseline	.72	10.31 (3.40)	9.99 (3.40)	10.56 (3.39)
4 weeks	.77	10.89 (3.60)	10.80 (3.64)	10.95 (3.57)
8 weeks	.77	11.05 (3.70)	11.21 (3.76)	10.93 (3.67)
12 weeks	.75	11.41 (3.70)	11.42 (3.76)	11.39 (3.66)

[a]Perceived stress was measured with the Perceived Stress Scale [52].

[b]Brooding was measured using the Ruminative Response Scale—Brooding Subscale [53].

[c]Sleep disturbance was measured using the PROMIS (Patient-Reported Outcomes Measurement Information System) Pediatric Sleep Disturbance Scale—Short Form 4a [54].

[d]Loneliness was measured using the Roberts UCLA Loneliness Scale [55].

[e]Optimism was measured using the Life Orientation Test—Revised [56].

Primary Outcome: Perceived Stress

We found a significant main effect of time ($P<.001$) but no significant effect of condition ($P=.97$). These were qualified by the predicted condition×time interaction ($P<.001$). To break down this interaction, we calculated simple slopes for each condition, which revealed a significant reduction in perceived stress among participants in the intervention condition ($B=-1.50$; 95% CI –1.82 to –1.19; $P<.001$) but not among those in the waitlist control ($B=-0.09$; 95% CI –0.44 to 0.26; $P=.61$; Figure 2). With the exception of a significant effect of gender ($P<.001$) suggesting that participants identifying as male had significantly lower PSS scores overall, no other effects were significant ($Ps\geq.06$).

XSL•FO
RenderX

Figure 2. Model-predicted changes in perceived stress scores across time and condition. PSS: Perceived Stress Scale.

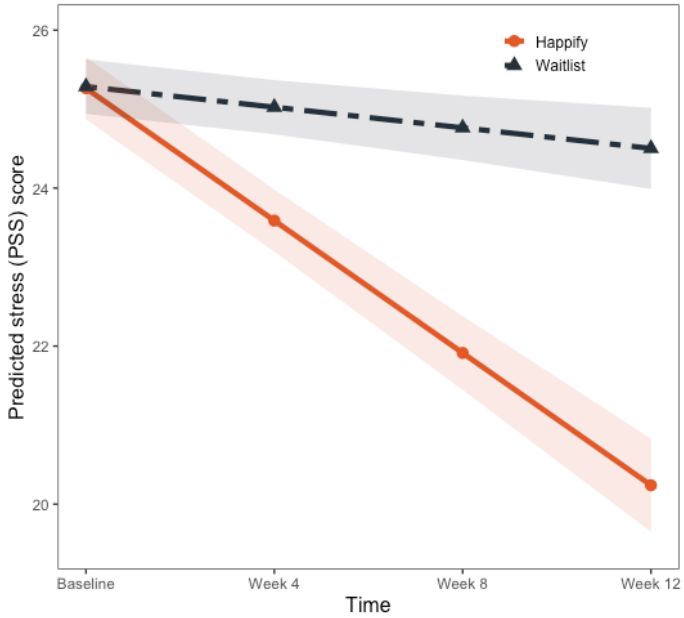

Secondary Outcome: Brooding

We also found a significant effect of time (*P*<.001) but not of condition (*P*=.15) for brooding. These effects were qualified by a significant condition×time interaction (*P*<.001). There were significant reductions in brooding among participants in the waitlist control (*B*=–0.30; 95% CI –0.47 to –0.12; *P*=.001) as well as those in the Happify for Teens condition (*B*=–0.84; 95% CI –1.00 to –0.68; *P*<.001); however, CIs for the 2 groups did not overlap, suggesting that improvements were relatively greater among those in the intervention condition (Figure 3). Again, with the exception of a significant effect of gender suggesting that participants identifying as male reported significantly lower levels of brooding (*P*<.001), no other effects were significant (*P*s≥.35).

Figure 3. Model-predicted changes in RRS—Brooding Subscale scores across time and condition. RRS: Ruminative Response Scale.

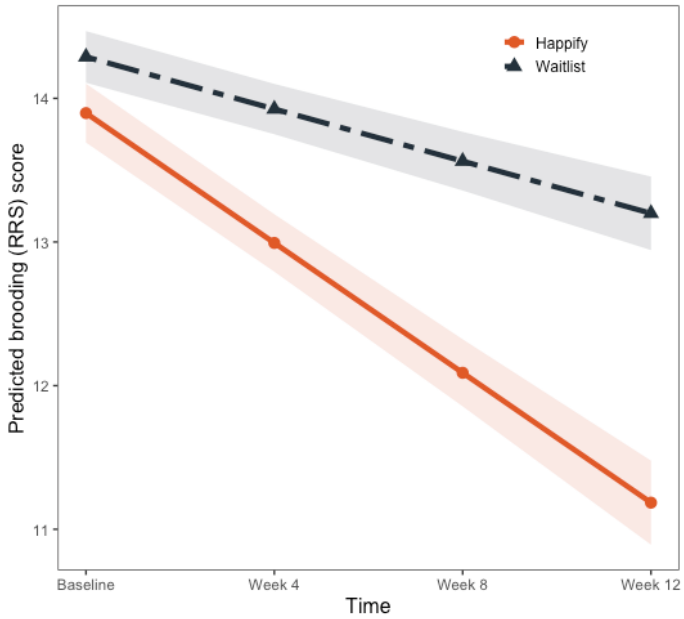

Exploratory Outcomes

For sleep disturbance and optimism, we found a significant effect of time (*P*s<.001) and no significant effect of condition (*P*s≥.28). However, the corresponding condition×time interactions were not significant (*P*s≥.10), suggesting that while there were improvements in sleep disturbance and optimism over time, improvements did not differ by condition. We did find a significant effect of age for sleep disturbance (*P*=.04), suggesting that sleep disturbance was greater as age increased, but no other effects were significant (*P*s≥.11). When assessments failing the attention check criterion were dropped, the condition×time interaction did become significant for sleep disturbance (*P*=.04). Because no other outcomes changed substantially in the sensitivity analyses, we did not interpret this interaction to avoid making a type I error.

XSL•FO
RenderX

For loneliness, we also found a significant effect of time (P<.001) and no significant effect of condition (P=.79), but the corresponding condition×time interaction was significant (P<.001). As with perceived stress and brooding, simple slopes for each condition were calculated, which revealed significant reductions in Roberts UCLA Loneliness Scale-8 scores among participants in the waitlist control (B=–0.38; 95% CI –0.64 to –0.12; P=.005) as well as those in the intervention condition (B=–0.96; 95% CI –1.2 to –0.73; P<.001). Similar to changes in brooding, however, the CIs for the 2 groups did not overlap, suggesting that improvements in loneliness were greater in the Happify for Teens group than in the waitlist control (Figure 4). No other effects were significant (Ps≥.16).

Figure 4. Model-predicted changes in UCLA loneliness scores across time and condition.

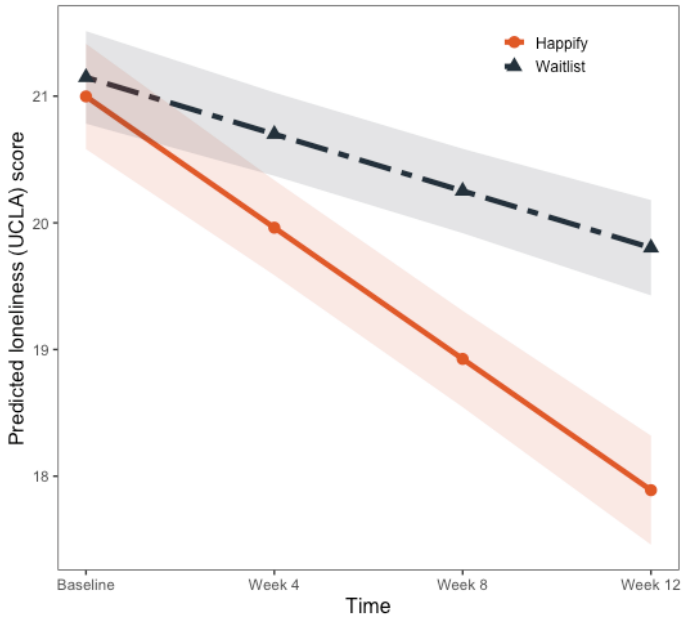

Discussion

Principal Findings

The purpose of this study was to test the efficacy of a DMHI designed for adolescents on perceived stress and brooding. Our findings suggest that adolescents who were engaged with the intervention had significant reductions in perceived stress, rumination, and loneliness over 12 weeks relative to those in a waitlist control but not in sleep disturbance or optimism.

Previous DMHI research with adolescents has focused primarily on depression and anxiety. For instance, in a recent review of 18 meta-analyses and systematic reviews of digital health interventions for adolescents [63], only 3 included stress as an outcome. Consistent with the present findings, these studies found significant improvements in stress among adolescents engaging with a digital intervention, which suggests DMHIs can be effective in improving stress management among adolescents. To our knowledge, however, this study was the first to explore the effects of a digital intervention on rumination among adolescents. Given that rumination predicts future dysfunction including anxiety, depression, and substance use [34,35], intervening early to reduce the frequency of brooding is another important mechanism for prevention.

We also found that the intervention significantly improved participants' loneliness but not optimism or sleep disturbance. Although these were exploratory outcomes, the effect on loneliness is not unexpected, given that Happify for Teens includes information targeting loneliness. Research conducted on the adult version of Happify also suggests that users perceive several activities that appear in the adolescent version as helpful for addressing their feelings of loneliness [64]. In contrast, sleep is not a focus within the program; and, consequently, to impact sleep outcomes, a more targeted intervention may be necessary.

The reasons why we did not find an effect on optimism are less clear. One potential explanation is that the Life Orientation Test—Revised is a measure of dispositional optimism and thus may be less sensitive to changes in optimism over time. Indeed, a recent study suggested that a state measure of optimism may be more suitable for assessing within-person changes [65]. Another possibility is that the study was conducted during the COVID-19 pandemic, which may have impacted participants' levels of optimism and the ability of the intervention to impact optimism. Consequently, additional research is needed to better understand whether this particular intervention can also improve optimism as well as perceived stress, brooding, and loneliness.

Leveraging DMHIs for Mental Health Support in Adolescence

Given that rates of mental illness among adolescents are increasing [4], and at a steeper rate than among adults [5], identifying opportunities to increase access to mental health care in this population is critical. Currently, most stress management programs for adolescents continue to be delivered within schools [27-29]. Although school-based prevention programs can be effective [28,29], there are numerous barriers to implementing evidence-based interventions within schools, including lack of time and resources and financial constraints [66]. Costs associated with implementing such interventions

XSL•FO
RenderX

may be particularly prohibitive in socially and economically disadvantaged areas [28], where students may need these interventions most [27]. Even when such programs can be implemented, student participation may be negatively impacted by fears of stigmatization [67]. Given that 95% of US adolescents own, or have access to, a cellular phone, and 88% have daily access to a computer [68], digital interventions may offer a better, and arguably more cost-effective, opportunity to reach more adolescents while potentially reducing concerns with stigma by increasing privacy and confidentiality.

DMHIs have become increasingly popular [69], and although interventions developed specifically for youths are still relatively scarce [70], interest in developing digital tools specifically for adolescents is also increasing. For example, Woebot Health has developed an intervention for adolescents with mild to moderate depression and anxiety that uses a relational agent to deliver CBT, interpersonal psychotherapy for adolescents, and elements of dialectical behavior therapy. Although results are not yet published, they recently completed an RCT testing this DMHI (ClinicalTrials.gov NCT05486611). Similarly, Limbix developed SparkRx, a DMHI drawing on behavioral activation and CBT designed for adolescents with depression, which has been shown to significantly improve symptoms of depression among adolescents who were engaged consistently with the intervention [71]. However, both interventions have been developed for adolescents with clinical levels of depression, which offer important support to adolescents struggling with mental illness but do not address the increased burden on mental health services coming from adolescents with less severe or no mental health impairment.

Limitations and Future Directions

One limitation of this design is the lack of an active control group. Although the inclusion of a waitlist control accounts for some threats to internal validity (eg, regression to the mean and maturation), we cannot rule out placebo effects. In particular, some researchers argue there are unique considerations for placebo effects in the context of digital interventions, including participants' beliefs about technology and how the mobile app is designed [72]. Research has shown that the impact of digital interventions is typically weaker when compared to active control, though these interventions still appear to outperform attentional controls in terms of depression and anxiety [73]. Previous research conducted on the adult version of Happify also showed superior effects on depression, anxiety, and resilience compared to an active control [49,50]. Nevertheless, additional research comparing the effects of Happify for Teens to a digital sham condition will be important to rule out any potential placebo effects.

Another limitation of this study was the relative homogeneity of participants, particularly in terms of race and schooling situation. Specifically, our sample was predominantly White and attended school full-time in person. Moreover, all participants in the study were US residents. Consequently, it is possible that our results may not generalize to adolescents from other countries, or effects may differ based on schooling situation or race. In addition, because we required participants to complete the baseline assessment before qualifying for the study, our sample may have included adolescents who were particularly motivated, conscientious, etc. Although this helped to reduce the risk of attrition, these participants may have engaged with the intervention differently than we would see in another, less motivated, sample of adolescents.

Another important consideration is that because we advertised to caregivers and obtained parental consent, our participants may have also differed from the general population in terms of attachment style, quality of relationship with their caregiver, parenting styles, or even the caregiver's own mental health; all of which could impact the extent to which the adolescent responds to and benefits from the program [74,75]. Research suggests that caregiver perceptions of an adolescent's mental health as well as their attitudes toward mental health care are more predictive of treatment-seeking among adolescents than the adolescents' own attitudes;, consequently, caregivers may act as the "gatekeepers" to mental health services [76]. Therefore, an important step to improving access to mental health support among adolescents may be to identify opportunities to reach those whose caregivers are less supportive. In future research, obtaining a waiver of parental consent may permit accessing those adolescents who are personally interested in DMHIs, but whose caregivers may be less interested or supportive.

Finally, previous research on DMHIs for adolescents is also criticized for a lack of follow-up assessments [77-79]. However, some interventions, including school-based stress management interventions, may have stronger effects over time than immediately after intervention [27], emphasizing the importance of tracking effects longitudinally. Although this study followed participants for 12 weeks, providing some insight into the longitudinal effects of the intervention, future research should include longer-term follow-up to better ascertain whether effects diminish or strengthen over time. In particular, because rumination can stem from both perceived stress [34] and loneliness [80], it is plausible that there could be continued improvements in rumination resulting from the reductions in perceived stress and loneliness. Similarly, additional research assessing the long-term implications of Happify for Teens on other mental or physical health outcomes that are predicted by stress, rumination, and loneliness is needed.

Conclusions

Research suggests that adolescents are experiencing increasing levels of stress, particularly while in school, but few engage in stress management activities [15,16]. Although several stress management interventions have been developed for adolescents, many of these are school-based interventions, and the evidence supporting their effectiveness is mixed [27-29]. Digital interventions may offer more cost-effective and scalable options for addressing important transdiagnostic risk factors in adolescence to promote positive mental health and prevent future dysfunction; however, DMHIs designed specifically for adolescents are rare, particularly those focused on wellness. This study provides promising evidence for the use of a DMHI for addressing perceived stress, brooding, and loneliness in adolescence.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Acacia Parks and Julia Stafford for their input in planning this study.

Authors' Contributions

EMB was the principal investigator for this study and led research planning as well as writing the paper. HW assisted in the design and implementation of the study. CJM was responsible for implementing and conducting the study. RDH was responsible for data analysis and supported paper writing. SES supported paper writing.

Conflicts of Interest

EMB is a salaried employee of Twill, Inc (the study sponsor) and holds stock options; all other authors (HW, RDH, CJM, and SES) were employees of Twill, Inc while working on this research. However, all authors received salaries that were not contingent on research outcomes.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1291 KB - jmir_v26i1e54282_app1.pdf]

References

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62(6):593-602 [FREE Full text] [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
2. Kessler RC, Angermeyer M, Anthony JC, De Graaf R, Demyttenaere K, Gasquet I, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. World Psychiatry 2007;6(3):168-176 [FREE Full text] [Medline: 18188442]
3. Mental health atlas 2014. World Health Organization. URL: https://www.who.int/publications/i/item/mental-health-atlas-2014 [accessed 2022-04-21]
4. Twenge JM, Cooper AB, Joiner TE, Duffy ME, Binau SG. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. J Abnorm Psychol 2019;128(3):185-199. [doi: 10.1037/abn0000410] [Medline: 30869927]
5. Olfson M, Blanco C, Wang S, Laje G, Correll CU. National trends in the mental health care of children, adolescents, and adults by office-based physicians. JAMA Psychiatry 2014;71(1):81-90 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.3074] [Medline: 24285382]
6. Merikangas KR, He JP, Burstein M, Swendsen J, Avenevoli S, Case B, et al. Service utilization for lifetime mental disorders in U.S. adolescents: results of the National Comorbidity Survey-Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry 2011;50(1):32-45 [FREE Full text] [doi: 10.1016/j.jaac.2010.10.006] [Medline: 21156268]
7. Whitney DG, Peterson MD. US national and state-level prevalence of mental health disorders and disparities of mental health care use in children. JAMA Pediatr 2019;173(4):389-391 [FREE Full text] [doi: 10.1001/jamapediatrics.2018.5399] [Medline: 30742204]
8. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
9. Olfson M, Druss BG, Marcus SC. Trends in mental health care among children and adolescents. N Engl J Med 2015;372(21):2029-2038 [FREE Full text] [doi: 10.1056/NEJMsa1413512] [Medline: 25992747]
10. Thomas CR, Holzer CE. The continuing shortage of child and adolescent psychiatrists. J Am Acad Child Adolesc Psychiatry 2006;45(9):1023-1031. [doi: 10.1097/01.chi.0000225353.16831.5d] [Medline: 16840879]
11. Kazdin AE. Adolescent mental health. Prevention and treatment programs. Am Psychol 1993;48(2):127-141. [doi: 10.1037//0003-066x.48.2.127] [Medline: 8442568]
12. Romeo RD. The teenage brain: the stress response and the adolescent brain. Curr Dir Psychol Sci 2013;22(2):140-145 [FREE Full text] [doi: 10.1177/0963721413475445] [Medline: 25541572]
13. Suo L, Zhao L, Si J, Liu J, Zhu W, Chai B, et al. Predictable chronic mild stress in adolescence increases resilience in adulthood. Neuropsychopharmacology 2013;38(8):1387-1400 [FREE Full text] [doi: 10.1038/npp.2013.67] [Medline: 23478858]
14. Østerås B, Sigmundsson H, Haga M. Perceived stress and musculoskeletal pain are prevalent and significantly associated in adolescents: an epidemiological cross-sectional study. BMC Public Health 2015;15:1081 [FREE Full text] [doi: 10.1186/s12889-015-2414-x] [Medline: 26498498]
15. Stress in America™: are teens adopting adults' stress habits? American Psychological Association. 2014. URL: https://www.apa.org/news/press/releases/stress/2013/stress-report.pdf [accessed 2020-10-29]

XSL•FO
RenderX

16. Stress in America 2020™: a national mental health crisis. American Psychological Association. 2020. URL: https://www.apa.org/news/press/releases/stress/2020/report-october [accessed 2022-10-29]
17. Burger K, Samuel R. The role of perceived stress and self-efficacy in young people's life satisfaction: a longitudinal study. J Youth Adolesc 2017;46(1):78-90. [doi: 10.1007/s10964-016-0608-x] [Medline: 27812840]
18. DuBois DL, Felner RD, Brand S, Adan AM, Evans EG. A prospective study of life stress, social support, and adaptation in early adolescence. Child Dev 1992;63(3):542-557. [doi: 10.1111/j.1467-8624.1992.tb01645.x] [Medline: 1600821]
19. Kaplan DS, Liu RX, Kaplan HB. School related stress in early adolescence and academic performance three years later: the conditional influence of self expectations. Soc Psychol Educ 2005;8(1):3-17. [doi: 10.1007/s11218-004-3129-5]
20. Wills TA, Sandy JM, Yaeger AM. Stress and smoking in adolescence: a test of directional hypotheses. Health Psychol 2002;21(2):122-130. [Medline: 11950102]
21. Nguyen-Rodriguez ST, Chou CP, Unger JB, Spruijt-Metz D. BMI as a moderator of perceived stress and emotional eating in adolescents. Eat Behav 2008;9(2):238-246 [FREE Full text] [doi: 10.1016/j.eatbeh.2007.09.001] [Medline: 18329603]
22. De Vriendt T, Clays E, Huybrechts I, De Bourdeaudhuij I, Moreno LA, Patterson E, et al. European adolescents' level of perceived stress is inversely related to their diet quality: the healthy lifestyle in Europe by nutrition in adolescence study. Br J Nutr 2012;108(2):371-380 [FREE Full text] [doi: 10.1017/S0007114511005708] [Medline: 22054044]
23. Wiklund M, Malmgren-Olsson EB, Ohman A, Bergström E, Fjellman-Wiklund A. Subjective health complaints in older adolescents are related to perceived stress, anxiety and gender—a cross-sectional school study in Northern Sweden. BMC Public Health 2012;12:993 [FREE Full text] [doi: 10.1186/1471-2458-12-993] [Medline: 23158724]
24. Wang S, Zhao Y, Zhang L, Wang X, Wang X, Cheng B, et al. Stress and the brain: perceived stress mediates the impact of the superior frontal gyrus spontaneous activity on depressive symptoms in late adolescence. Hum Brain Mapp 2019;40(17):4982-4993 [FREE Full text] [doi: 10.1002/hbm.24752] [Medline: 31397949]
25. Lathren C, Bluth K, Park J. Adolescent self-compassion moderates the relationship between perceived stress and internalizing symptoms. Pers Individ Dif 2019;143:36-41 [FREE Full text] [doi: 10.1016/j.paid.2019.02.008] [Medline: 32042216]
26. Carter JS, Garber J. Predictors of the first onset of a major depressive episode and changes in depressive symptoms across adolescence: stress and negative cognitions. J Abnorm Psychol 2011;120(4):779-796. [doi: 10.1037/a0025441] [Medline: 21928863]
27. van Loon AWG, Creemers HE, Beumer WY, Okorn A, Vogelaar S, Saab N, et al. Can schools reduce adolescent psychological stress? A multilevel meta-analysis of the effectiveness of school-based intervention programs. J Youth Adolesc 2020;49(6):1127-1145 [FREE Full text] [doi: 10.1007/s10964-020-01201-5] [Medline: 32034632]
28. Feiss R, Dolinger SB, Merritt M, Reiche E, Martin K, Yanes JA, et al. A systematic review and meta-analysis of school-based stress, anxiety, and depression prevention programs for adolescents. J Youth Adolesc 2019;48(9):1668-1685 [FREE Full text] [doi: 10.1007/s10964-019-01085-0] [Medline: 31346924]
29. Rew L, Johnson K, Young C. A systematic review of interventions to reduce stress in adolescence. Issues Ment Health Nurs 2014;35(11):851-863. [doi: 10.3109/01612840.2014.924044] [Medline: 25353298]
30. Lindahl M, Archer T. Depressive expression and anti-depressive protection in adolescence: stress, positive affect, motivation and self-efficacy. Psychology 2013;4(6):495-505 [FREE Full text] [doi: 10.4236/psych.2013.46070]
31. Dozois DJA, Seeds PM, Collins KA. Transdiagnostic approaches to the prevention of depression and anxiety. J Cogn Psychother 2009;23(1):44-59. [doi: 10.1891/0889-8391.23.1.44]
32. Kuhlman KR, Chiang JJ, Bower JE, Irwin MR, Cole SW, Dahl RE, et al. Persistent low positive affect and sleep disturbance across adolescence moderate link between stress and depressive symptoms in early adulthood. J Abnorm Child Psychol 2020;48(1):109-121 [FREE Full text] [doi: 10.1007/s10802-019-00581-y] [Medline: 31446530]
33. Smith JM, Alloy LB. A roadmap to rumination: a review of the definition, assessment, and conceptualization of this multifaceted construct. Clin Psychol Rev 2009;29(2):116-128 [FREE Full text] [doi: 10.1016/j.cpr.2008.10.003] [Medline: 19128864]
34. Michl LC, McLaughlin KA, Shepherd K, Nolen-Hoeksema S. Rumination as a mechanism linking stressful life events to symptoms of depression and anxiety: longitudinal evidence in early adolescents and adults. J Abnorm Psychol 2013;122(2):339-352 [FREE Full text] [doi: 10.1037/a0031994] [Medline: 23713497]
35. Skitch SA, Abela JRZ. Rumination in response to stress as a common vulnerability factor to depression and substance misuse in adolescence. J Abnorm Child Psychol 2008;36(7):1029-1045. [doi: 10.1007/s10802-008-9233-9] [Medline: 18461438]
36. Marks ADG, Sobanski DJ, Hine DW. Do dispositional rumination and/or mindfulness moderate the relationship between life hassles and psychological dysfunction in adolescents? Aust N Z J Psychiatry 2010;44(9):831-838. [doi: 10.3109/00048674.2010.487478] [Medline: 20815670]
37. McLaughlin KA, Nolen-Hoeksema S. Interpersonal stress generation as a mechanism linking rumination to internalizing symptoms in early adolescents. J Clin Child Adolesc Psychol 2012;41(5):584-597 [FREE Full text] [doi: 10.1080/15374416.2012.704840] [Medline: 22867280]
38. Beck AT. Cognitive Therapy of Depression. New York, NY: Guilford; 1979.
39. Praissman S. Mindfulness-based stress reduction: a literature review and clinician's guide. J Am Acad Nurse Pract 2008;20(4):212-216. [doi: 10.1111/j.1745-7599.2008.00306.x] [Medline: 18387018]

XSL•FO
RenderX

40. Duckworth AL, Steen TA, Seligman ME. Positive psychology in clinical practice. Annu Rev Clin Psychol 2005;1:629-651. [doi: 10.1146/annurev.clinpsy.1.102803.144154] [Medline: 17716102]
41. Seligman MEP, Steen TA, Park N, Peterson C. Positive psychology progress: empirical validation of interventions. Am Psychol 2005;60(5):410-421. [doi: 10.1037/0003-066X.60.5.410] [Medline: 16045394]
42. Edwards M, Adams EM, Waldo M, Hadfield OD, Biegel GM. Effects of a mindfulness group on Latino adolescent students: examining levels of perceived stress, mindfulness, self-compassion, and psychological symptoms. J Spec Group Work 2014;39(2):145-163. [doi: 10.1080/01933922.2014.891683]
43. Hamdan-Mansour AM, Puskar K, Bandak AG. Effectiveness of cognitive-behavioral therapy on depressive symptomatology, stress and coping strategies among Jordanian university students. Issues Ment Health Nurs 2009;30(3):188-196. [doi: 10.1080/01612840802694577] [Medline: 19291496]
44. Brinkborg H, Michanek J, Hesser H, Berglund G. Acceptance and commitment therapy for the treatment of stress among social workers: a randomized controlled trial. Behav Res Ther 2011;49(6-7):389-398. [doi: 10.1016/j.brat.2011.03.009] [Medline: 21513917]
45. Wersebe H, Lieb R, Meyer AH, Hofer P, Gloster AT. The link between stress, well-being, and psychological flexibility during an Acceptance and Commitment Therapy self-help intervention. Int J Clin Health Psychol 2018;18(1):60-68 [FREE Full text] [doi: 10.1016/j.ijchp.2017.09.002] [Medline: 30487911]
46. Carpenter J, Crutchley P, Zilca RD, Schwartz HA, Smith LK, Cobb AM, et al. Seeing the "big" picture: big data methods for exploring relationships between usage, language, and outcome in internet intervention data. J Med Internet Res 2016;18(8):e241 [FREE Full text] [doi: 10.2196/jmir.5725] [Medline: 27580524]
47. Carpenter J, Crutchley P, Zilca RD, Schwartz HA, Smith LK, Cobb AM, et al. Correction: seeing the "big" picture: big data methods for exploring relationships between usage, language, and outcome in internet intervention data. J Med Internet Res 2017;19(12):e347 [FREE Full text] [doi: 10.2196/jmir.8099] [Medline: 29258059]
48. Parks AC, Williams AL, Kackloudis GM, Stafford JL, Boucher EM, Honomichl RD. The effects of a digital well-being intervention on patients with chronic conditions: observational study. J Med Internet Res 2020;22(1):e16211 [FREE Full text] [doi: 10.2196/16211] [Medline: 31922491]
49. Parks AC, Williams AL, Tugade MM, Hokes KE, Honomichl RD, Zilca RD. Testing a scalable web and smartphone based intervention to improve depression, anxiety, and resilience: a randomized controlled trial. Int J Wellbeing 2018;8(2):22-67 [FREE Full text] [doi: 10.5502/ijw.v8i2.745]
50. Williams AL, Parks AC, Cormier G, Stafford J, Whillans A. Improving resilience among employees high in depression, anxiety, and workplace distress. Int J Manag Res 2018;9(1-2):4-22 [FREE Full text]
51. Boucher EM, Ward HE, Stafford JL, Parks AC. Effects of a digital mental health program on perceived stress in adolescents aged 13-17 years: protocol for a randomized controlled trial. JMIR Res Protoc 2021;10(4):e25545 [FREE Full text] [doi: 10.2196/25545] [Medline: 33871377]
52. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385-396. [Medline: 6668417]
53. Treynor W, Gonzalez R, Nolen-Hoeksema S. Rumination reconsidered: a psychometric analysis. Cogn Ther Res 2003;27:247-259. [doi: 10.1023/A:1023910315561]
54. Baumel A, Edan S, Kane JM. Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs. Transl Behav Med 2019;9(6):1020-1033. [doi: 10.1093/tbm/ibz147] [Medline: 31689344]
55. Forrest CB, Meltzer LJ, Marcus CL, de la Motte A, Kratchman A, Buysse DJ, et al. Development and validation of the PROMIS Pediatric Sleep Disturbance and Sleep-Related Impairment item banks. Sleep 2018;41(6):zsy054 [FREE Full text] [doi: 10.1093/sleep/zsy054] [Medline: 29546286]
56. Roberts RE, Lewinsohn PM, Seeley JR. A brief measure of loneliness suitable for use with adolescents. Psychol Rep 1993;72(3 Pt 2):1379-1391. [doi: 10.2466/pr0.1993.72.3c.1379] [Medline: 8337350]
57. Scheier MF, Carver CS, Bridges MW. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. J Pers Soc Psychol 1994;67(6):1063-1078. [doi: 10.1037//0022-3514.67.6.1063] [Medline: 7815302]
58. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48 [FREE Full text] [doi: 10.18637/jss.v067.i01]
59. Wood D, Harms PD, Lowman GH, DeSimone JA. Response speed and response consistency as mutually validating indicators of data quality in online samples. Soc Psychol Pers Sci 2017;8(4):454-464. [doi: 10.1177/1948550617703168]
60. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
61. Boucher E, Honomichl R, Ward H, Powell T, Stoeckl SE, Parks A. The effects of a digital well-being intervention on older adults: retrospective analysis of real-world user data. JMIR Aging 2022;5(3):e39851 [FREE Full text] [doi: 10.2196/39851] [Medline: 36053569]

XSL•FO
RenderX

62. Montgomery RM, Boucher EM, Honomichl RD, Powell TA, Guyton SL, Bernecker SL, et al. The effects of a digital mental health intervention in adults with cardiovascular disease risk factors: analysis of real-world user data. JMIR Cardio 2021;5(2):e32351 [FREE Full text] [doi: 10.2196/32351] [Medline: 34806986]
63. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
64. Boucher EM, McNaughton EC, Harake N, Stafford JL, Parks AC. The impact of a digital intervention (Happify) on loneliness during COVID-19: qualitative focus group. JMIR Ment Health 2021;8(2):e26617 [FREE Full text] [doi: 10.2196/26617] [Medline: 33498011]
65. Hoeppner SS, Millstein RA, Siegel KR, Carlon HA, Harnedy LE, Chung WJ, et al. A secondary analysis examining the performance of the State Optimism Measure (SOM) compared to the Life Orientation Test-Revised (LOT-R) in measuring optimism over time. Psychol Health 2022:1-16. [doi: 10.1080/08870446.2022.2126472] [Medline: 36154764]
66. Hicks TB, Shahidullah JD, Carlson JS, Palejwala MH. Nationally Certified School Psychologists' use and reported barriers to using evidence-based interventions in schools: the influence of graduate program training and education. Sch Psychol Q 2014;29(4):469-487. [doi: 10.1037/spq0000059] [Medline: 24708281]
67. Gronholm PC, Nye E, Michelson D. Stigma related to targeted school-based mental health interventions: a systematic review of qualitative evidence. J Affect Disord 2018;240:17-26. [doi: 10.1016/j.jad.2018.07.023] [Medline: 30041074]
68. Teens, social media and technology 2018. Pew Research Center. 2018. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2020-10-29]
69. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. J Med Internet Res 2013;15(6):e120 [FREE Full text] [doi: 10.2196/jmir.2600] [Medline: 23770578]
70. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
71. Kulikov V, Padmanabhan A, Peake E, Lake J. 1.29 An evaluation of the effectiveness of a CBT-based digital therapeutic in reducing adolescent depressive symptoms. J Am Acad Child Adolesc Psychiatry 2022;61(10):S150. [doi: 10.1016/j.jaac.2022.09.045]
72. Torous J, Firth J. The digital placebo effect: mobile mental health meets clinical psychiatry. Lancet Psychiatry 2016;3(2):100-102. [doi: 10.1016/S2215-0366(15)00565-9] [Medline: 26851322]
73. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002 [FREE Full text] [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
74. Melchior M, van der Waerden J. Parental influences on children's mental health: the bad and the good sides of it. Eur Child Adolesc Psychiatry 2016;25(8):805-807 [FREE Full text] [doi: 10.1007/s00787-016-0891-9] [Medline: 27456961]
75. DeWalt DA, Hink A. Health literacy and child health outcomes: a systematic review of the literature. Pediatrics 2009;124(Suppl 3):S265-S274. [doi: 10.1542/peds.2009-1162B] [Medline: 19861480]
76. Tsang YT, Franklin M, Sala-Hamrick K, Kohlberger B, Simon VA, Partridge T, et al. Caregivers as gatekeepers: professional mental health service use among urban minority adolescents. Am J Orthopsychiatry 2020;90(3):328-339. [doi: 10.1037/ort0000432] [Medline: 31971405]
77. Richardson T, Stallard P, Velleman S. Computerised cognitive behavioural therapy for the prevention and treatment of depression and anxiety in children and adolescents: a systematic review. Clin Child Fam Psychol Rev 2010;13(3):275-290. [doi: 10.1007/s10567-010-0069-9] [Medline: 20532980]
78. Li J, Theng YL, Foo S. Game-based digital interventions for depression therapy: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2014;17(8):519-527 [FREE Full text] [doi: 10.1089/cyber.2013.0481] [Medline: 24810933]
79. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual research review: digital health interventions for children and young people with mental health problems—a systematic and meta-review. J Child Psychol Psychiatry 2017;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
80. Vanhalst J, Luyckx K, Van Petegem S, Soenens B. The detrimental effects of adolescents' chronic loneliness on motivation and emotion regulation in social situations. J Youth Adolesc 2018;47(1):162-176. [doi: 10.1007/s10964-017-0686-4] [Medline: 28497208]

Abbreviations

CBT: cognitive behavioral therapy
DMHI: digital mental health intervention
MLM: multilevel model
PROMIS: Patient-Reported Outcomes Measurement Information System
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by G Eysenbach; submitted 03.11.23; peer-reviewed by E Bernstein; comments to author 30.11.23; revised version received 15.12.23; accepted 17.02.24; published 29.03.24.

Please cite as:
Boucher EM, Ward H, Miles CJ, Henry RD, Stoeckl SE
Effects of a Digital Mental Health Intervention on Perceived Stress and Rumination in Adolescents Aged 13 to 17 Years: Randomized Controlled Trial
J Med Internet Res 2024;26:e54282
URL: https://www.jmir.org/2024/1/e54282
doi:10.2196/54282
PMID:38551617

©Eliane M Boucher, Haley Ward, Cynthia J Miles, Robert D Henry, Sarah Elizabeth Stoeckl. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceptions of Illness Control, Coherence, and Self-Efficacy Following a Web-Based Lifestyle Program for Multiple Sclerosis: A Qualitative Analysis of Semistructured Interviews

Rebekah Davenport[1], MPsych(Clin); William Bevens[2,3], PhD; Sandra Neate[3], MD, PhD; Pia Jelinek[3], MD; Maggie Yu[3], MA, MPH; George Jelinek[3], MD, PhD; Jeanette Reece[3], PhD

[1]Melbourne School of Psychological Sciences, Faculty of Medicine, Dentistry & Health Sciences, The University of Melbourne, Melbourne, Australia
[2]Department of Psychiatry, IN STEP Children's Mental Health Research Center, University of California, San Diego, La Jolla, CA, United States
[3]Melbourne School of Theology, Centre for Epidemiology and Biostatistics, The University of Melbourne, Melbourne, Australia

Corresponding Author:
William Bevens, PhD
Department of Psychiatry
IN STEP Children's Mental Health Research Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA, 92093
United States
Phone: 1 858 534 7792
Email: wbevens@health.ucsd.edu

Abstract

Background: Evidence suggests that illness perceptions held by people living with multiple sclerosis (MS) impact affective distress and physical health outcomes. In a randomized controlled trial, we developed 2 MS Online Courses—the standard care course and the intervention course (IC). The IC was adapted from an evidence-based lifestyle program. Modifying lifestyle risk factors offers an opportunity to impact illness perceptions. Research on illness perceptions in people living with MS has focused predominately on quantitative methods.

Objective: This study aimed to explore the experiences and health outcomes of randomized controlled trial participants, including individuals' perceived changes in attitudes toward MS and their health (ie, illness perceptions).

Methods: Qualitative responses provided by 38 participants (22 in the IC and 16 in the standard care course) were derived from semistructured interviews 1 month after completing the MS Online Course. The impact of course engagement and lifestyle modification on illness perceptions was explored using inductive thematic analysis.

Results: Themes of changes in illness perceptions were most prominent in the IC arm. Qualitative responses from 22 people living with MS informed the development of three themes: "self-efficacy for disease management," "personal control," and "illness coherence."

Conclusions: Findings provide novel insights into the formation and modification of illness perceptions. Lifestyle modification may serve as a valuable adjunct to clinician-administered therapies, improving the sense of personal control over MS, illness coherence, and self-efficacy for disease management.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12621001605886; https://tinyurl.com/2vyve9p9

International Registered Report Identifier (IRRID): RR2-10.1186/s12883-023-03298-0

(J Med Internet Res 2024;26:e60240) doi:10.2196/60240

KEYWORDS

illness perceptions; personal control perceptions; illness coherence; self-efficacy; lifestyle modification; disease management; multiple sclerosis; MS; qualitative; health behavior; thematic analysis

XSL•FO
RenderX

Introduction

People living with multiple sclerosis (MS) are tasked with adjusting to substantial prognostic uncertainty, in the absence of clarity regarding the etiology of symptoms and without access to a cure. Critical events, such as receiving an MS diagnosis and ongoing illness stressors, can disrupt emotional equilibrium [1]. It is well established that depressive, anxiety, and stress symptoms are highly prevalent in the population with MS [2]. Approximately 30% of people living with MS report clinically significant depressive symptoms, while 34% experience symptoms of anxiety [2]. Seminal works by Lazarus and Folkman [3] and Lazarus and Opton [4] on cognitive appraisal propose that the psychological impact of an event (eg, diagnosis and relapse) is partially impacted by an individual's appraisal of the event and their self-efficacy to access and use effective coping responses. Leventhal [5] and Leventhal et al [6] developed the Common-Sense Model of Self-Regulation theory (CSM) to formulate how attitudes and beliefs are specifically formed in response to chronic illness and health events. Research has examined the impact of illness perceptions in people living with MS, highlighting significant contributions of perceptions relating to perceived personal control over MS, a strong illness identity (attribution of physical symptoms to disease), and illness coherence (perceived understanding of MS) on depressive symptoms [7]. Furthermore, self-efficacy perceptions have been shown to impact symptom severity and disability [8] and quality of life (QoL) [9].

Identifying factors predictive of health outcomes and amenable to intervention is critical for fostering hope in disease management, particularly in the context of a highly uncertain disease such as MS. Given their robust links to mental and physical health outcomes in people living with MS, illness perceptions hold substantial clinical significance. In addition, interventional data demonstrate the amenability of illness perceptions. Changes to illness perceptions, facilitated by cognitive behavioral therapy, have been shown to coincide with decreases in depressive and anxiety symptoms [10].

Involving health care practitioners, including mental health professionals, in MS care is crucial but often regarded as time-intensive, difficult to access, and potentially costly [11]. As such, there has been an emerging focus on lifestyle self-management as an intervention for people living with MS, supplemental to clinician-administered psychological treatments and pharmaceutical disease-modifying therapies. Emerging evidence supports the effectiveness of lifestyle self-management interventions, encompassing improved diet quality, increased exercise, mindfulness practices, and stress management, in enhancing QoL [12,13] and improving depressive and anxiety symptoms [14]. Furthermore, self-management interventions place the patient at the center of their own care, emphasizing their responsibility and control over identifying and addressing health challenges.

According to the CSM for chronic disease populations [5,6], exposure to new and alternative information from expert sources is important in influencing the nature and severity of illness perceptions. When presented within the context of a programmatic, educational intervention delivered by experts, lifestyle modification information may provide people living with MS with the required scaffolding to reevaluate their mental representations of illness and efficacy to manage their own MS. To date, research into illness perceptions in the population living with MS has almost exclusively been conducted within a quantitative research paradigm. While findings offer important insights into group-level relationships between illness perceptions and health outcomes in MS, they are unable to address questions relating to the highly-varied individual experience of MS or the impact of self-managed lifestyle modifications on illness perceptions. Few studies have explored the patient perspective on MS using qualitative methods [15]. This information holds considerable value for medical and allied health professionals in tailoring treatment plans for patients with MS [14]. Providing additional options for multimodal lifestyle interventions may enrich the overall approach to patient care.

This study is an additional qualitative analysis to the larger ancillary randomized controlled trial (RCT) that assessed the effectiveness of the MS Online Course (MSOC) [16]. The MSOC provided information tailored on lifestyle-related risk factors for people living with MS, compared with a standard care course (SCC) with general lifestyle information. This qualitative study aimed to assess individuals' perceived changes in attitudes toward MS and health (ie, illness perceptions) 1 month after completing the MSOC. Responses were obtained through semistructured interviews and analyzed using inductive thematic analysis.

Methods

Ethical Considerations

The study involving humans was conducted according to the guidelines of the Declaration of Helsinki and approved by The University of Melbourne Human Research Ethics subcommittee (ID: 22140). This study received approval from the University of Melbourne Human Research Ethics subcommittee on November 2, 2021 (ID: 1851781.2). This ancillary RCT is a CONSORT-R (Consolidated Standards of Reporting Trials–Routine)–compliant RCT (Multimedia Appendix 1). The RCT protocol was reviewed and approved by the Australian New Zealand Clinical Registry on November 25, 2021 (ACTRN12621001605886). Written informed consent was obtained from all participants before inclusion in the RCT with additional verbal consent obtained from all participants in this qualitative study to undertake semistructured qualitative interviews for the purposes of this research. This qualitative study adheres to the Consolidated Criteria for Reporting Qualitative Research (COREQ) [17] and the American Psychological Association (APA) Style Journal Article Reporting Standards [18]. Participants received no compensation for their involvement. Qualitative data are presented anonymously, with participants categorized by study arm and assigned random numbers (1-16 for the SCC; 1-22 for the intervention course [IC]).

XSL•FO
RenderX

Study Design and Procedure

The MSOC effectiveness RCT is a randomized, single-blinded trial of a 6-week online lifestyle modification intervention that follows participants who completed the MSOC over a 2.5-year period. Consequently, short-term primary RCT outcomes were not yet available at the execution of this study.

Details of the MSOC have been comprehensively described previously [16]. Participants were recruited internationally through MS society websites (eg, MS Australia), research newsletters (eg, Clinical Trials Australia), Instagram (Meta), and public MS Facebook (Meta) groups focused on the dissemination of health information. Eligible participants were aged 18 years or older, fluent in English, and self-reported a physician's diagnosis of MS. Exclusion criteria included significant comorbid neurological conditions (eg, stroke) or chronic disease and participation in another RCT. At baseline (0 months), participants completed an online survey within the first MSOC education module, covering sociodemographic and disease variables (eg, MS subtype), lifestyle habits, and physical and mental health outcomes.

The MSOC encompassed 2 courses (study arms)—the IC and the SCC—developed by researchers in collaboration with people living with MS to address lifestyle-related risk factors. The objective is to compare the effectiveness of each course in a larger RCT for improving QoL and health outcomes. Both the IC and SCC consist of seven modules: (1) Introduction, (2) Eat Well, (3) Sunlight and Vitamin D, (4) Exercise, (5) Meditation and Mind-body Connection, (6) Medication and Family Prevention, and (7) Conclusion. The SCC provided general lifestyle information from reputable MS websites, while the IC delivered information tailored to people living with MS, based on an evidence-based MS program [19] (Multimedia Appendix 2).

Qualitative Study

Participant Recruitment

Participants of the RCT who completed all course modules, the precourse baseline survey, and postcourse evaluation survey across both study arms (n=53) were invited by email to participate in a 45-60 minute semistructured interview 1 month after completing the course (Multimedia Appendix 3). In total, 38 participants accepted the invitation. Written informed consent was obtained from all participants in the baseline survey, within the first MSOC module of the RCT, and verbal consent was obtained before participating in semistructured interviews purposed for qualitative analyses.

Demographic and clinical quantitative data presented in the current study were obtained from the pre-course baseline survey. The Hospital Anxiety and Depression Scale (HADS) [20] was reported to characterize the severity of self-reported depressive and anxiety symptoms. The HADS, used extensively in the population living with MS (eg, Dahl et al [21] and Janssens et al [22]), has demonstrated accurate sensitivity and specificity for major depressive disorder (90% and 87%, respectively) and generalized anxiety disorder (88% and 81%, respectively) within the population living with MS [23]. Ambulation levels were evaluated using the Patient-Determined Disease Steps (PDDS) scale [24], recognized as a guide of overall disability in individuals with mild-to-moderate disabilities due to MS [25].

Interview Procedure

Participants were invited by email to participate in semistructured interviews 1 month after completing the MSOC. Interviews occurred at a time convenient for participants between October 14, 2022, and November 16, 2022, and were conducted by SN (lead interviewer), RD, JR, and PJ using Zoom (Zoom Video Communications, Inc) software licensed by The University of Melbourne or WhatsApp (Meta). Interviewers had a research background and either (1) had substantial experience conducting qualitative research or (2) were highly trained health care professionals routinely involved in the clinical management of MS and other chronic diseases. There were minor variations in the interview scripts for the intervention versus the standard care group. Interview questions explored perceived changes in attitudes toward MS and health, initial health changes, and broader aspects of the course experience (eg, engagement with the online forum and views of course content; Multimedia Appendix 4).

To ensure consistency in the delivery of the interview schedule between interviewers, authors (RD, JR, and PJ) observed 1 interview conducted by the lead interviewer (SJ). Minor changes to the structure of the interview schedule were made following the 3 initial interviews with the intention of ensuring the schedule followed a coherent, logical order (eg, all questions probing mental health variables were asked consecutively). With participant permission, interviews were audio recorded and transcribed verbatim, using the voice recognition software Temi (Rev) [26]. Transcripts were saved in a restricted folder and were accessed by interviewers. Transcript text was altered for purposes such as upholding participant confidentiality by redacting sensitive disclosures as requested by participants and ensuring that text were a verbatim account participants' verbal responses. Transcripts were imported into NVivo qualitative data analysis software (version 12, 2018; Lumivero [formerly QSR International]). In line with best-practice data management, document coding and theme development were maintained within NVivo outputs.

Thematic Analysis

Inductive, reflexive thematic analysis was used as a means of identifying, analyzing, and reporting patterns in the data [27]. Reflexive thematic analysis was selected as it allows for researcher autonomy in the processes of determining themes (latent or semantic) and establishing the most relevant analysis to use (inductive vs theoretical). This study applied the 6 distinct phases of thematic analysis: data familiarization, generating codes, constructing themes, reviewing potential themes, defining and naming themes, and producing the report [28]. The dataset was divided into 2, according to the IC and SCC arms, and both were analyzed and coded independently by RD and WB, respectively.

Following data familiarization and coding, the research team met to review potential themes to ensure alignment with the raw data and participants' perspectives. To enhance transparency, verbatim quotes have been included to illustrate

XSL•FO
RenderX

themes, allowing readers to assess the validity of researchers' interpretations.

Participants in the SCC arm reported some changes in mental health; however, these were varied and often attributed to external factors. As such, only data from participants in the IC arm were retained. In total, 4 superordinate themes were discerned from the data corpus (Figure 1). Due to their distinct nature and relevance to diverse fields, the research team partitioned the data based on these superordinate themes and conducted separate analyses. This study focuses on the superordinate theme "changes in mental health aspects," delving into psychological phenomena such as participants' experiences in modifying illness perceptions over the course. Another superordinate theme, "information seeking," which addresses online information-seeking behaviors among people living with MS, was reported separately [29]. The remaining 2 themes required quantitative examination to derive their meaning, in alignment with the larger RCT's goals of quantifying changes in MS symptoms and lifestyle after MSOC completion.

Figure 1. Schematic of the qualitative analysis process stages 1-3. MS: multiple sclerosis.

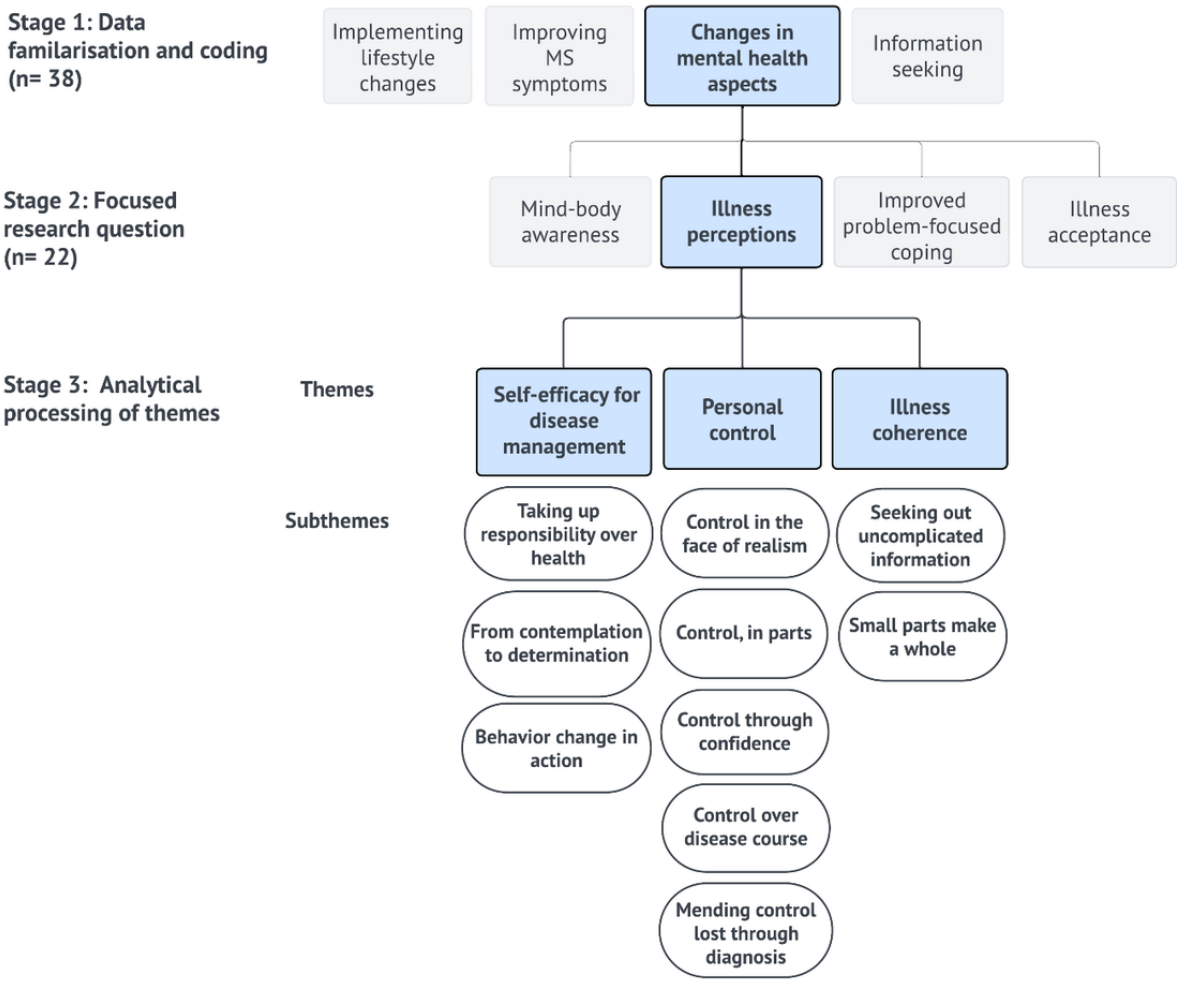

Results

Participant Characteristics

Table 1 provides demographic and clinical characteristics. Among the 22 participants who completed the IC (mean age 48, SD 12 years), most identified as female (19/22, 86%) and self-reported a relapsing-remitting MS phenotype (18/22, 82%). Most participants scored within the "normal" (<8 HADS score) ranges for depression (19/22, 86%), while over half reported "normal" anxiety levels (13/22, 59%). Clinically significant fatigue was reported by a large proportion (13/22, 59%), and 68% (15/22) of participants had no or mild disability.

XSL•FO
RenderX

Table 1. Self-reported demographic and clinical characteristics of the multiple sclerosis online course interventional arm interviewees (n=22) at baseline.

Demographics	Statistical value
Age (years), n (%)	48.14 (12.25)
Gender, n (%)	
Men	3 (14)
Women	19 (86)
Other	0 (0)
Country of residence, n (%)	
Australia or New Zealand	5 (23)
United States or Canada	5 (23)
United Kingdom	3 (14)
Other	9 (41)
Highest educational level, n (%)	
Postgraduate degree	7 (32)
University degree	7 (32)
Secondary education	2 (9)
Higher education qualification (eg, certificate and diploma)	6 (27)
Employment status, n (%)	
Currently employed	13 (59)
Unemployed	8 (36)
Missing	1 (5)
Currently in a relationship, n (%)	
Yes	14 (64)
No	8 (36)
Clinical characteristics	
MS[a] phenotype, n (%)	
RRMS[b]	18 (82)
PPMS[c]	2 (9)
SPMS[d]	1 (5)
Unsure or other	1 (5)
Years since MS diagnosis, mean (SD)	5.36 (4.30)
Disability level, mean (SD)	1.55 (1.82)
Normal or mild (PDDS[e]≤2), n (%)	15 (68)
Moderate (PDDS≥3 to 5), n (%)	6 (27)
Severe (PDDS≥6), n (%)	1 (5)
DMT[f] use, n (%)	
Yes	13 (59)
No	9 (41)
Comorbidities, n (%)	
Yes	16 (73)
No	4 (18)
Missing	2 (9)
Anemia or blood disease	1 (5)

XSL•FO
RenderX

Demographics	Statistical value
Anxiety	4 (18)
Back pain	4 (18)
Lung disease	2 (9)
Other (eg, high cholesterol)	11 (50)
Fatigue levels, mean (SD)	4.81 (1.59)
Normal or mild, n (%)	9 (41)
Clinically significant (FSS[g]>5), n (%)	13 (59)
Depressive symptoms, mean (SD)	4.91 (3.07)
Normal (HADS[h]<8), n (%)	19 (86)
Mild-moderate (HADS≥8 to 10), n (%)	1 (5)
Clinically significant (HADS≥11), n (%)	2 (9)
Anxiety symptoms, mean (SD)	7.14 (4.61)
Normal (HADS<8), n (%)	13 (59)
Mild-moderate (HADS≥8 to 10), n (%)	5 (23)
Clinically significant (HADS≥11), n (%)	4 (18)

[a]MS: multiple sclerosis.
[b]RRMS: relapsing-remitting multiple sclerosis.
[c]PPMS: primary progressive multiple sclerosis.
[d]SPMS: secondary progressive multiple sclerosis.
[e]PDDS: Patient-Determined Disease Steps.
[f]DMT: disease-modifying treatments.
[g]FSS: fatigue severity scale.
[h]HADS: Hospital Anxiety and Depression Scale.

Themes

Overview

Within "changes in mental-health aspects," 3 themes relating to illness perceptions were identified. The first theme, "self-efficacy for disease management," explored perceptions of increased confidence and motivation to manage MS, gained through engaging with and applying lifestyle modification information. Data relating to this theme reflected beliefs about one's capacity to execute behaviors necessary for disease management, resembling the concept of self-efficacy assessed in the Multiple Sclerosis Self-Efficacy (MSSE) Scale [30], the gold-standard assessment for self-efficacy in persons with MS. Thus, the designation of this theme title was deemed appropriate.

The second theme, "personal control," explored perceptions of improved sense of personal control over MS, developed through engaging with and applying lifestyle modification information from the IC. Finally, the third theme, "illness coherence," explored individuals' experiences of an increased and more coherent cognitive representation of their MS. These 2 themes were named due to the high similarity to the factors "personal control" and "illness coherence," as described by the Illness Perceptions Questionnaire (IPQ) [31].

Quotations are reported to illustrate themes, identified by the study arm (IC) and participant number.

Self-Efficacy for Disease Management

The theme "self-efficacy for disease management" explored individuals' perceptions of increased confidence and motivation gained through undertaking the IC, and by commencing the recommended lifestyle changes provided in the IC. Increased self-efficacy had 3 effects, referred to as subthemes, including "taking responsibility over health," "from contemplation to determination," and "behavior change in action."

Participants described improvements in their self-efficacy to manage MS through learning about lifestyle modifications for diet, exercise, and mindfulness practice. Course information appeared to foster motivation to adopt a future-oriented focus and prioritize what changes may be important for improving health, based on evidence presented in the course. Equipped with this lifestyle-modification information, participants described feeling an increased sense of confidence to take up responsibility over their own health and motivation to make recommended lifestyle changes.

> *I felt confident (from the course) that I now have the data behind making changes. I can believe it.* [IC18]
>
> *It's had a positive effect because I feel more confidence that I know what I need to do and that I'll be able to do it.* [IC6]

Lifestyle modification information increased participants' sense of responsibility and ownership over managing their health,

XSL•FO
RenderX

enhancing their confidence to adopt course recommendations and seek further lifestyle-related information.

> *The course gave me a launching pad of things I can look at myself.* [IC14]
>
> *Having a course like this has helped me know that I am on the right path. I'm doing the right things. I've got more information while I was doing the course that boosted me.* [IC9]

Increased confidence and motivation, stemming from engagement, enabled many participants to progress through stages of change in lifestyle management, as delineated by the Stages of Change Model [32]. Improvements in self-efficacy facilitated a shift away from merely contemplating change toward a more determined and committed approach, evident in the development of action plans and intentions to make life-overhauling changes.

> *Now I know what I need to do, to achieve what I want to achieve. I know how to get there. I've got a plan.* [IC6]

Some set specific targets for change, such as supplementing with omega-3 fatty acids, increasing vitamin-D exposure, exercising, and re-engaging with mindfulness techniques including mindful breathing and awareness. For some, this included introducing the first form of intentional exercise in years.

> *It helped. I already knew some of the changes I needed to make but it pushed me into making them. So, you know, in addition to just relying on vitamin D supplements, I need to actually get out in the sun.* [IC22]
>
> *I have a plan moving forward. I'm really gonna step that (exercise) up. I've got a stronger awareness of the boxes I need to tick. I've got a target.* [IC6]

For some, plans had begun to take effect; actions were underway and new patterns of behavior were beginning to form. Participants described improvements in mood and attributed this to the integration of lifestyle changes.

> *I'm not as depressed as I used to be.* [IC2]
>
> *I'm happier now.* [IC20]
>
> *I feel more in control. Because when you feel that something is stressful if I go and meditate, afterwards I am more relaxed and those bad feelings are better.* [IC20]

Participant's responses also raised the possibility of a bidirectional relationship between self-efficacy and change; self-efficacy prompting lifestyle modifications, which may in turn create positive mastery experiences that further enhance self-efficacy.

> *When I started to do some changing, I started to feel more confident, and overall, it's been better for my mental-health.* [IC18]
>
> *I definitely feel more confident. It helped me be free of confusion about how my immune system works on MS drugs and leave behind fear of the unknown. I'm remembering to breathe and reduce stress and that it's important to fuel your body, make good choices, one at a time.* [IC4]

Personal Control

The theme "personal control" explored individuals' awareness of their improved personal control over their MS through engaging with lifestyle-related modification information. This theme included 5 subthemes relating to different forms of control, including "control in the face of realism," "control, in parts," "control through confidence," "control over disease course," and "mending control lost through diagnosis."

Overwhelmingly, participants expressed an increased sense of control over their MS, although tempered by the realism of the disease. They found a balance between focusing on controllable aspects of MS and accepting that their actions alone may not cure their disease.

> *I believe it [lifestyle modification] won't necessarily get rid of the MS, but it will improve my day to day, symptom-wise.* [IC1]
>
> *I'm trying to cover all bases. Trying to take some control of a condition that I don't really have much control of.* [IC19]

Having accepted that MS is a chronic disease, some participants began to focus on the possibility that lifestyle modifications may help to control the progression of symptoms.

> *So it (the course), made me realise I can actually, through diet and lifestyle change, I potentially hold off anything getting worse.* [IC14]
>
> *I'm interested in all the factors I can change that are in my hands, to change the course of the disease.* [IC12]

Many expressed a conservative view of lifestyle modification, stating that even if modifications may not control the disease or symptoms directly, lifestyle modification provided the opportunity to experience a sense of control.

> *I cannot control it (the disease), but at least I have a feeling I'm controlling it. Which is also very important. It motivated me more to do it (make lifestyle modifications).* [IC17]
>
> *When I was diagnosed I asked my neurologist about the diet changes and she said, 'as a doctor, I cannot tell you that diet change will help you, but I think you should change your diet because it will make you feel like you're in control.' Even if it's just that this course gave me information on flax-seed oil. It's something new that gives me an idea that I am at least a bit more in control than I was.* [IC17]

Participants were prompted to evaluate the utility of all-or-nothing thinking in relation to personal control, that is, the polarity of being "in control" of their MS versus being "out of control" of their MS. Through learning about controllable aspects of health (eg, diet and exercise), participants were able to reconceptualize control as a feeling which varies along a continuum between these 2 possibilities. That is,

XSL•FO
RenderX

self-management lifestyle modification offers individuals the opportunity to gain more control over their MS.

> *I don't feel like I can control whether or not I'm going to get new lesions…but I'm doing as much as I can within my control, which is to exercise more, continue to eat a health diet, meditation, yoga.* [IC4]

> *It (the course) gave me the push to take more control over my life, with my knowledge about lifestyle modification. I think it helped more than I realised at the time.* [IC13]

The IC made participants feel they could extend their sense of control over their MS and explore other avenues to gain further control of their MS in the future.

> *It (the course) just made me think '...well, wait a minute, I know I can manage this. I know I can control this, or control elements of it.' The course gave me a bit more motivation to consider, what else can I do here?* [IC13]

For some, confidence was highly interrelated with control, such that increases in confidence coincided with increases in perceptions of personal control.

> *I think it helped me. I was sort of 80% there, but after the course, I'm at 100%. It comes hand in hand with more confidence, you feel like you do have more control. I know now what I need to do and it's only up to me to do it.* [IC6]

> *I am not at the stage where I can defend and justify following the diet, but I'm trying my best to get to that stage… confidence for me comes from having a better understanding of the oils and that. I'm not there yet, but I am more confident, and my confidence comes from feeling a little bit of control.* [IC19]

For 1 newly diagnosed participant, there was a gap between their current reality and a state of perceived control over their MS. However, the IC, and possibly other educational courses, appeared to help them feel more in control.

> *I completely panicked after I was diagnosed and I was planning things in advance, and I was like, no, I can't do that because maybe I won't be able to walk by then…I felt completely out of control the entire time…Once your realise you can't actually live like that, then that (personal control) slowly comes back and it's still coming back and this course has helped.* [IC15]

Illness Coherence

The theme of "illness coherence" encompassed individuals' beliefs in a more coherent understanding of MS and included 2 subthemes: "seeking out uncomplicated information" and "small parts make a whole."

Participants emphasized the challenges associated with complex health information, complicated by medical terminology, and the barrier this imposes for constructing a coherent representation of MS. In contrast, the IC was regarded as a helpful source of information, scaffolding participants' understanding of health information.

> *Medical information is difficult to understand, but the more I read about it, the easier it becomes.* [IC1]

One participant described the protective effects of uncomplicated communication within the acute adjustment period following diagnosis.

> *It (the course) has been helpful. Normalising MS, especially for people in their twenties or thirties who are unexpectedly catapulted into a chronic situation. It's a lot to take in because we're at a point in our lives where everything feels like it's starting. Then it feels like you're figuring out a whole new identity. So, simple communication on health information is helpful. This is a lifestyle change, and also the way we view it, or the way we can support others in understanding it is helpful.* [IC5]

The opportunity to understand and piece together complex health information enabled participants to develop a more coherent mental representation of their disease, which they viewed as being important for their general well-being.

> *Overall I think it (doing the IC) was good for my wellbeing…more information makes me more informed and then the changes that I'm implementing should be positive on my overall wellbeing.* [IC1]

This allowed some to begin translating health information into everyday well-being practices, such as mediation, with beneficial results.

> *I think things have gotten better since the course, I can say that. And the course helped me to put things together, especially for me, meditation, I'm doing a bit more meditation, which I find really does help.* [IC9]

Discussion

Principal Findings

Overview

This study assessed participants' responses 1 month after completing the IC of an online, tailored education lifestyle intervention for people living with MS. To the authors' knowledge, this study is the first to qualitatively explore the impact of lifestyle modification on illness perceptions in a population living with MS. Specifically, it aimed to understand the sequelae of the IC on individuals' cognitive representations of their MS. Within the superordinate theme of "changes in mental health aspects," 3 subthemes were identified—"self-efficacy for disease management," "personal control" perceptions, and "illness coherence" perceptions. Findings contribute to an evidence base on the role of illness perceptions in understanding the experiences of people living with MS. In addition, findings offer preliminary support for the CSM by demonstrating a pathway between exposure to new and valid information and changes in control perceptions, illness coherence, and self-efficacy beliefs [15].

XSL•FO
RenderX

The positive effects of self-managed lifestyle modification on mental health outcomes, such as depression and anxiety, in the population with MS are well established [13]. This study provides evidence that exposure to lifestyle-related information may yield benefits for individuals' perceptions of their disease. Illness perceptions have been proposed as cognitive mechanisms underlying changes in depression and anxiety [31] and may be helpful in linking lifestyle modification to positive mental health outcomes. Positioned as a valuable adjunct to clinician-administered therapies, lifestyle modification has the potential to empower individuals to assume greater responsibility over their disease management, thereby enhancing the comprehensiveness of the treatment approach.

Deciphering the Connection: Lifestyle Modification and Personal Control

Gaining a sense of control, which is commonly reduced following MS diagnosis [33], was a primary motivating factor for engaging with the MSOC, as identified in an earlier qualitative study nested within a pilot feasibility study of the MSOC [34]. Furthermore, evidence substantiates the importance of personal control perceptions for mental health outcomes in people living with MS (eg, Brown et al [35] and Jopson and Moss-Morris [36]). While the current analysis confirms the importance of personal control perceptions and the impact of the IC on this aspect, it also offers novel insights into the multidimensional nature of personal control perceptions in MS. Findings illustrate how personal control may be formed, the extent of control gained, and its relevance across MS-related experiences.

Participants' reports suggested a link between control perceptions and confidence, with each potentially influencing the other. Gaining more information about lifestyle modification, helped to bolster individuals' confidence and enabled them to feel more in control of MS. Importantly, this was tendered by a pragmatic, continuum-based view of control over MS. While some participants described retrospective feelings of helplessness and a lack of control at the time of diagnosis, they acknowledged the possibility of exerting partial control over MS, aided through their engagement with the IC. This is unexpected as most study participants were within the first 5 years of diagnosis, a period marked by increased distress and tendencies toward all-or-nothing cognitive patterns [37]. Notably, most participants in this analysis scored within the normal ranges for depression and anxiety, contrasting findings from previous observational research [2]. This suggests that the current sample may represent a subgroup of people living with MS who have more positive mental health and may face fewer barriers to lifestyle modification. Considering that motivation difficulties and fatigue—hallmark symptoms of depression [38,39]—are known barriers to lifestyle modification [40], future research should explore how mental health influences engagement with lifestyle modification information, the initiation of lifestyle changes, and the extent to which mental health may impact the effectiveness of self-management interventions on personal control perceptions.

Building Illness Coherence: When Information Becomes Clear

Participants' responses in the current study highlighted the changing nature of illness coherence perceptions, triggered by an MS diagnosis, and shaped through their interaction with the IC. One participant described their experience of diagnosis as a sudden shift, "catapulting" them into a situation that required them to reconstruct their identity in midlife. Notably, given that MS is frequently misunderstood and concealed, the period following diagnosis often requires individuals to assert their understanding of MS and how MS uniquely manifests within them [41]. Therefore, sourcing health information represents a high priority for people living with MS [41,42].

The concept of illness coherence represents a relatively new idea, distinct from the construct of "sense of coherence," which enables individuals to cope with adverse experiences and is regarded as a dispositional trait (ie, a characteristic embedded within the personality) [42]. Sense of coherence has demonstrated predictive power across a spectrum of physical and mental health outcomes in people living with MS [43,44]; however, given its supposed trait-like structure, its amenability to intervention may be limited. In contrast, illness coherence perceptions have been shown to exhibit temporal variability [45], and their connections to mental health outcomes, encompassing depression and anxiety, have been documented in a handful of MS studies (eg, Kroencke et al [46] and Lynch et al [47]) and are firmly established within the context of other chronic disease populations [48].

The CSM posits that exposure to new and expert information may alter illness perceptions, including psychological adjustment outcomes [15]. The present study extends this concept, suggesting a precondition for such change—health information must be presented in a way that can be easily understood by individuals without specialized medical knowledge. Participant responses suggested that their engagement with uncomplicated health information within the IC facilitated a process of reintegration, helping them to "put things together" again, toward a coherent mental representation of MS. Consequently, participants described improvements in well-being and the practical application of information provided in the IC into daily life, including mindfulness.

The Role of Self-Efficacy in Advancing Through Stages of Change

Individuals reported increased self-efficacy for disease management following engagement with the IC which manifested in diverse ways. The Stages of Change Model [32] is a highly influential framework for understanding health-related behavior, and its application can lend meaning to the present findings. Through this lens, participant responses can be allocated to 2 different stages of readiness. The first set of responses suggested that the IC helped them overcome ambivalence toward change, progressing from Prochaska and Velicer's [32] "contemplation stage to preparation or determination" [32]. At this stage, individuals had established the belief that lifestyle modification may lead to a healthier life and described planning steps toward change, with some already experiencing positive effects.

XSL•FO
RenderX

The second set of responses suggested a more advanced stage of change, narrating experiences consistent with a transition from "preparation/determination" into "action." By incorporating recommended adjustments, including increasing exercise, mindfulness practices, and making dietary modifications, participants noted improvements in their well-being, within a mere month, further enhancing their self-efficacy. These responses raise the possibility of a bidirectional relationship between self-efficacy and lifestyle modification changes, whereby changes in 1 factor feed into changes in the other. The pathway from self-efficacy to behavior change has been tested and substantiated in previous research [49], although operating under the assumption that changes in self-efficacy precede behavioral changes.

It should be noted that an individual's baseline stage of change (ie, where individuals are at in terms of contemplating or actioning change) may predict the likelihood of behavior change [50]. The act of enrolling in and completing a self-guided lifestyle modification course likely signifies contemplation of change. Future research may consider assessing individuals' baseline self-efficacy for disease management as one index of readiness for change. Tailoring lifestyle modification educational programs according to the individual's stage of readiness may help to optimize lifestyle modification change. This approach may also help cater to those who are less inclined to engage with lifestyle modification (eg, in the precontemplation stage), as well as those focused on maintaining changes and preventing regression to earlier stages.

Limitations and Future Research

A 1-month time lag between completion of the MSOC and the interviews yielded insights into the immediate impacts of engagement and implementation of lifestyle modifications. Nonetheless, the long-term effects of the MSOC, specifically the IC, could not be determined due to the scope of this study. Future research should examine changes in illness perceptions to determine if changes are maintained or bolstered through longer-term implementation of lifestyle changes, as individuals witness the benefits of modifications. Of equal importance, researchers should seek to assess cases where individuals initially implemented learned information but experienced a subsequent decline in their efforts over time.

Despite the strengths in the study's international scope and voluntary sampling, certain participant characteristics were underrepresented. In alignment with the study's objective to evaluate the course's impact, a deliberate selection process was necessary wherein only a small percentage of participants who had completed the larger RCT were eligible for qualitative interviews. However, this approach introduces constraints on the generalizability of findings, warranting caution in extending these results to the broader population with MS. Underrepresented characteristics encompass male individuals, individuals from non-WEIRD (Western, educated, industrialized, rich, and democratic) countries, those identifying outside the binary categories of male and female, individuals with progressive MS phenotypes, and those exhibiting moderate-to-higher levels of depressive and anxiety symptoms.

Conclusions

This study provides evidence for the therapeutic impact of engagement with educational information on lifestyle modification. Representing one of the first qualitative examinations in this area, this study provides insights into the formation and modification of illness perceptions which are commonly linked to mental and physical health outcomes. Findings suggest that the IC, a self-guided online lifestyle modification program, may serve as a valuable adjunct to clinician-administered therapies, improving sense of personal control over MS, disease coherence, and self-efficacy for disease management.

Acknowledgments

The authors gratefully acknowledge the generous contribution of those who participated in the Multiple Sclerosis Online Course (MSOC) feasibility trial, those who participated in qualitative interviews, the individuals currently participating in the MSOC effectiveness randomized controlled trial, and specifically the participants interviewed in this study.

We thank our industry partners, JMA Creative, who created the web platform and developed content. The technical development of the MSOC through JMA creative was funded by the Overcoming MS Charity (United Kingdom). This work was supported by the Neuroepidemiology Unit at The University of Melbourne which is funded by anonymous philanthropic donors.

Recognizing that our identities can shape our scientific perspectives [51], we aim to offer readers insight into our backgrounds as authors. When the manuscript was submitted, 5 authors self-identified as women (RD, SN, PJ, MY, and JR) and 2 authors as men (WB and GJ). Regarding race, 6 self-identified as White (RD, SN, PJ, JR, WB, and GJ), and 1 as East Asian or Chinese (MY).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RD contributed to conceptualization, project administration, validation, investigation, methodology, data analysis, visualization, and writing. WB contributed to conceptualization, validation, methodology, data analysis, and writing—review and editing. SN contributed to conceptualization, validation, investigation, methodology, visualization, writing—review and editing, and supervision.

XSL•FO
RenderX

PJ contributed to validation, investigation, methodology, writing—review and editing. MY contributed to conceptualization, investigation, methodology, data analysis, writing—review and editing. GJ contributed to conceptualization, methodology, writing—review and editing, funding acquisition. JR contributed to conceptualization, project administration, validation, investigation, methodology, writing—review and editing, and supervision.

Conflicts of Interest

GJ is the author of "Overcoming Multiple Sclerosis." GJ and SN are coeditors of "Overcoming Multiple Sclerosis Handbook. Roadmap to Good Health" and facilitators of past residential lifestyle modification workshops. PJ is a contributor to "Overcoming Multiple Sclerosis Handbook. Roadmap to Good Health." The other authors report there are no competing interests to declare.

Multimedia Appendix 1
CONSORT-eHEALTH Checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1151 KB - jmir_v26i1e60240_app1.pdf]

Multimedia Appendix 2
Outline of intervention and standard care course format.
[DOCX File , 14 KB - jmir_v26i1e60240_app2.docx]

Multimedia Appendix 3
Participant flow diagram. IC: intervention course; MSOC: Multiple Sclerosis Online Course; SCC: standard care course.
[PNG File , 84 KB - jmir_v26i1e60240_app3.png]

Multimedia Appendix 4
One-month postcourse completion interview schedule.
[DOCX File , 36 KB - jmir_v26i1e60240_app4.docx]

References

1. Ackroyd K, Fortune DG, Price S, Howell S, Sharrack B, Isaac CL. Adversarial growth in patients with multiple sclerosis and their partners: relationships with illness perceptions, disability and distress. J Clin Psychol Med Settings 2011;18(4):372-379. [doi: 10.1007/s10880-011-9265-0] [Medline: 21986996]
2. Boeschoten RE, Braamse AM, Beekman AT, Cuijpers P, van Oppen P, Dekker J, et al. Prevalence of depression and anxiety in multiple sclerosis: a systematic review and meta-analysis. J Neurol Sci 2017;372:331-341. [doi: 10.1016/j.jns.2016.11.067] [Medline: 28017241]
3. Lazarus RS, Folkmam S. Stress, Appraisal, and Coping. New York: Springer publishing company; 1984:0826141927.
4. Lazarus RS, Opton Jr EM. The study of psychological stress: A summary of theoretical formulations and experimental findings. Anxiety and Behavior 1966;76(34):1-35 [FREE Full text]
5. Leventhal H. Toward a comprehensive theory of emotion in advances in experimental social psychology. Academic Press 1980;13:139-207. [doi: 10.1016/s0065-2601(08)60132-x]
6. Leventhal H, Phillips LA, Burns E. The common-sense model of self-regulation (CSM): a dynamic framework for understanding illness self-management. J Behav Med 2016;39(6):935-946. [doi: 10.1007/s10865-016-9782-2] [Medline: 27515801]
7. Davenport RA, Krug I, Rickerby N, Dang PL, Forte E, Kiropoulos L. Personality and cognitive factors implicated in depression and anxiety in multiple sclerosis: a systematic review and meta-analysis. Journal of Affective Disorders Reports 2024;17:100832. [doi: 10.1016/j.jadr.2024.100832]
8. Tingey JL, Valentine TR, Kratz AL, Bentley JA, Ehde DM, Alschuler KN. Self-efficacy trajectories of individuals newly diagnosed with multiple sclerosis. Rehabil Psychol 2023;68(2):204-211. [doi: 10.1037/rep0000487] [Medline: 36951739]
9. Dehghan M, Sharifi P, Hasani J, Young CA, Langdon D. Healthier living with MS: The key role of self-efficacy and emotion regulation. Mult Scler Relat Disord 2023;73:104608. [doi: 10.1016/j.msard.2023.104608] [Medline: 36963171]
10. Hazhir F, Ahadi H, Pourshahbaz A, Reeiza M. The effectiveness of cognitive-behavior therapy on illness representations of multiple-sclerosis and improving their emotional states. Journal of Kermanshah University of Medical Sciences 2012;15(6):e78893.
11. Bebo B, Cintina I, LaRocca N, Ritter L, Talente B, Hartung D, et al. The economic burden of multiple sclerosis in the United States. Neurology 2022;98(18):e1810-e1817. [doi: 10.1212/wnl.0000000000200150]
12. Simpson-Yap N, Probst Y, Reece JC, Jelinek GA, Neate S. Prospective associations of better quality of the diet with improved quality of life over 7.5 years in people with multiple sclerosis. Multiple Sclerosis and Related Disorders 2022;60:103710. [doi: 10.1016/j.msard.2022.103710]

XSL•FO
RenderX

13. Wills OC, Probst YC. Understanding lifestyle self-management regimens that improve the life quality of people living with multiple sclerosis: a systematic review and meta-analysis. Health Qual Life Outcomes 2022;20(1):153 [FREE Full text] [doi: 10.1186/s12955-022-02046-1] [Medline: 36434609]
14. Kidd T, Carey N, Mold F, Westwood S, Miklaucich M, Konstantara E, et al. A systematic review of the effectiveness of self-management interventions in people with multiple sclerosis at improving depression, anxiety and quality of life. PLoS One 2017;12(10):e0185931 [FREE Full text] [doi: 10.1371/journal.pone.0185931] [Medline: 29020113]
15. Lynd LD, Henrich NJ, Hategeka C, Marra CA, Mittmann N, Evans C, et al. Perspectives of patients with multiple sclerosis on drug treatment: a qualitative study. Int J MS Care 2018;20(6):269-277 [FREE Full text] [doi: 10.7224/1537-2073.2017-109] [Medline: 30568564]
16. Reece JC, Yu M, Bevens W, Simpson-Yap S, Jelinek G, Jelinek P, et al. Study protocol for an online lifestyle modification education course for people living with multiple sclerosis: the Multiple Sclerosis Online Course (MSOC). BMC Neurol 2023;23(1):249 [FREE Full text] [doi: 10.1186/s12883-023-03298-0] [Medline: 37386385]
17. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
18. Kazak AE. Editorial: journal article reporting standards. Am Psychol 2018;73(1):1-2. [doi: 10.1037/amp0000263] [Medline: 29345483]
19. Jelinek G, Neate S, O' donoghue M. Overcoming Multiple Sclerosis Handbook: Roadmap to Good Health. Australia: Allen & Unwin; 2022.
20. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
21. Dahl OP, Stordal E, Lydersen S, Midgard R. Anxiety and depression in multiple sclerosis. a comparative population-based study in Nord-Trøndelag county, Norway. Mult Scler 2009;15(12):1495-1501. [doi: 10.1177/1352458509351542] [Medline: 19965515]
22. Janssens AC, Buljevac D, van Doorn PV, van der Meché FGA, Polman CH, Passchier J, et al. Prediction of anxiety and distress following diagnosis of multiple sclerosis: a two-year longitudinal study. Mult Scler 2006;12(6):794-801. [doi: 10.1177/1352458506070935] [Medline: 17263009]
23. Honarmand K, Feinstein A. Validation of the Hospital Anxiety and Depression Scale for use with multiple sclerosis patients. Mult Scler 2009;15(12):1518-1524. [doi: 10.1177/1352458509347150] [Medline: 19965520]
24. Hohol MJ, Orav EJ, Weiner HL. Disease steps in multiple sclerosis: a simple approach to evaluate disease progression. Neurology 1995;45(2):251-255. [doi: 10.1212/wnl.45.2.251] [Medline: 7854521]
25. Ann Marrie R, McFadyen C, Yaeger L, Salter A. A systematic review of the validity and reliability of the Patient-Determined Disease Steps scale. Int J MS Care 2023;25(1):20-25 [FREE Full text] [doi: 10.7224/1537-2073.2021-102] [Medline: 36711220]
26. Temi. Temi. Speech to text transcription in 5 minutes: Advanced speech recognition software. Voice recognition software. Temi URL: https://temi.com [accessed 2022-11-22]
27. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
28. Bevens W, Davenport RA, Neate S, Yu M, Jelinek P, Jelinek GA, et al. Web-based health information seeking by people living with multiple sclerosis: qualitative investigation of the multiple sclerosis online course. J Med Internet Res 2024;26:e53372 [FREE Full text] [doi: 10.2196/53372] [Medline: 38335016]
29. Rigby SA, Domenech C, Thornton EW, Tedman S, Young CA. Development and validation of a self-efficacy measure for people with multiple sclerosis: the Multiple Sclerosis Self-Efficacy Scale. Mult Scler 2003;9(1):73-81. [doi: 10.1191/1352458503ms870oa] [Medline: 12617272]
30. Weinman J, Petrie KJ, Moss-morris R, Horne R. The illness perception questionnaire: a new method for assessing the cognitive representation of illness. Psychology & Health 1996;11(3):431-445. [doi: 10.1080/08870449608400270]
31. Bassi M, Cilia S, Falautano M, Grobberio M, Niccolai C, Pattini M, et al. Illness perceptions and psychological adjustment among persons with multiple sclerosis: the mediating role of coping strategies and social support. Disability and Rehabilitation 2019;42(26):3780-3792. [doi: 10.1080/09638288.2019.1610511]
32. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
33. Garabedian M, Perrone E, Pileggi C, Zimmerman V. Support group participation: effect on perceptions of patients with newly diagnosed multiple sclerosis. Int J MS Care 2020;22(3):115-121. [doi: 10.7224/1537-2073.2018-099] [Medline: 32607073]
34. Neate SL, Bevens W, Jelinek PL, Gray KM, Weiland TJ, Nag N, et al. A multiple sclerosis lifestyle behavior online course: qualitative analysis of participants' motivations, expectations and experiences. Front Public Health 2022;10:1022185 [FREE Full text] [doi: 10.3389/fpubh.2022.1022185] [Medline: 36568793]
35. Brown RF, Valpiani EM, Tennant CC, Dunn SM, Sharrock M, Hodgkinson S, et al. Longitudinal assessment of anxiety, depression, and fatigue in people with multiple sclerosis. Psychol Psychother 2009;82(Pt 1):41-56. [doi: 10.1348/147608308X345614] [Medline: 18727845]

XSL•FO
RenderX

36. Jopson NM, Moss-Morris R. The role of illness severity and illness representations in adjusting to multiple sclerosis. J Psychosom Res 2003;54(6):503-513. [doi: 10.1016/s0022-3999(02)00455-5] [Medline: 12781303]
37. Dennison L, Moss-Morris R, Silber E, Galea I, Chalder T. Cognitive and behavioural correlates of different domains of psychological adjustment in early-stage multiple sclerosis. J Psychosom Res 2010;69(4):353-361. [doi: 10.1016/j.jpsychores.2010.04.009] [Medline: 20846536]
38. Schreurs KM, de Ridder DT, Bensing JM. Fatigue in multiple sclerosis: reciprocal relationships with physical disabilities and depression. J Psychosom Res 2002;53(3):775-781. [doi: 10.1016/s0022-3999(02)00326-4] [Medline: 12217451]
39. Solaro C, Gamberini G, Masuccio FG. Depression in multiple sclerosis: epidemiology, aetiology, diagnosis and treatment. CNS Drugs 2018;32(2):117-133. [doi: 10.1007/s40263-018-0489-5] [Medline: 29417493]
40. Neate SL, Donald A, Jelinek GA, Nag N. Experiences of and attitudes to lifestyle modification for the management of multiple sclerosis: a qualitative analysis of free-text survey data. Health Expect 2022;25(1):214-222 [FREE Full text] [doi: 10.1111/hex.13364] [Medline: 34599857]
41. Desborough J, Brunoro C, Parkinson A, Chisholm K, Elisha M, Drew J, et al. 'It struck at the heart of who I thought I was': a meta-synthesis of the qualitative literature examining the experiences of people with multiple sclerosis. Health Expect 2020;23(5):1007-1027. [doi: 10.1111/hex.13093] [Medline: 32578287]
42. Antonovsky A. Unraveling the Mystery of Health: How People Manage Stress and Stay Well. US San Francisco: Jossey-Bass Inc; 1987.
43. Broersma F, Oeseburg B, Dijkstra J, Wynia K. The impact of self-perceived limitations, stigma and sense of coherence on quality of life in multiple sclerosis patients: results of a cross-sectional study. Clin Rehabil 2018;32(4):536-545 [FREE Full text] [doi: 10.1177/0269215517730670] [Medline: 28895427]
44. Calandri E, Graziano F, Borghi M, Bonino S. Coping strategies and adjustment to multiple sclerosis among recently diagnosed patients: the mediating role of sense of coherence. Clin Rehabil 2017;31(10):1386-1395. [doi: 10.1177/0269215517695374] [Medline: 28933614]
45. Bonsaksen T, Lerdal A, Fagermoen MS. Trajectories of illness perceptions in persons with chronic illness: an explorative longitudinal study. J Health Psychol 2015;20(7):942-953. [doi: 10.1177/1359105313504235] [Medline: 24140616]
46. Kroencke DC, Denney DR, Lynch SG. Depression during exacerbations in multiple sclerosis: the importance of uncertainty. Mult Scler 2001;7(4):237-242. [doi: 10.1177/135245850100700405] [Medline: 11548983]
47. Lynch SG, Kroencke DC, Denney DR. The relationship between disability and depression in multiple sclerosis: the role of uncertainty, coping, and hope. Mult Scler 2001;7(6):411-416. [doi: 10.1177/135245850100700611] [Medline: 11795464]
48. Dempster M, Howell D, McCorry NK. Illness perceptions and coping in physical health conditions: a meta-analysis. J Psychosom Res 2015;79(6):506-513. [doi: 10.1016/j.jpsychores.2015.10.006] [Medline: 26541550]
49. Schwarzer R, Lippke S, Luszczynska A. Mechanisms of health behavior change in persons with chronic illness or disability: the health action process Approach (HAPA). Rehabil Psychol 2011;56(3):161-170. [doi: 10.1037/a0024509] [Medline: 21767036]
50. Engels ES, Nigg CR, Reimers AK. Predictors of physical activity behavior change based on the current stage of change-an analysis of young people from Hawai'i. J Behav Med 2022;45(1):38-49. [doi: 10.1007/s10865-021-00255-5] [Medline: 34471990]
51. Roberts SO, Bareket-Shavit C, Dollins FA, Goldie PD, Mortenson E. Racial inequality in psychological research: trends of the past and recommendations for the future. Perspect Psychol Sci 2020;15(6):1295-1309. [doi: 10.1177/1745691620927709] [Medline: 32578504]

Abbreviations

APA: American Psychological Association
CONSORT-R: Consolidated Standards of Reporting Trials–Routine
COREQ: Consolidated Criteria for Reporting Qualitative Research
CSM: Common-Sense Model of Self-Regulation Theory
HADS: Hospital Anxiety and Depression Scale
IC: intervention course
IPQ: Illness Perceptions Questionnaire
MS: multiple sclerosis
MSOC: Multiple Sclerosis Online Course
MSSE: Multiple Sclerosis Self-Efficacy
PDDS: Patient-Determined Disease Steps
QoL: quality of life
RCT: randomized controlled trial
SCC: standard care course
WEIRD: Western, educated, industrialized, rich, and democratic

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 06.05.24; peer-reviewed by J Desborough, E Unni; comments to author 30.08.24; revised version received 23.09.24; accepted 16.10.24; published 29.11.24.

Please cite as:
Davenport R, Bevens W, Neate S, Jelinek P, Yu M, Jelinek G, Reece J
Perceptions of Illness Control, Coherence, and Self-Efficacy Following a Web-Based Lifestyle Program for Multiple Sclerosis: A Qualitative Analysis of Semistructured Interviews
J Med Internet Res 2024;26:e60240
URL: https://www.jmir.org/2024/1/e60240
doi:10.2196/60240
PMID:39612211

©Rebekah Davenport, William Bevens, Sandra Neate, Pia Jelinek, Maggie Yu, George Jelinek, Jeanette Reece. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Pharmacogenetics Clinical Decision Support Systems for Primary Care in England: Co-Design Study

Videha Sharma[1,2], BMBS, BMedSci, PhD; John McDermott[3,4], BSc, MBChB; Jessica Keen[3], MSc, MPharm; Simon Foster[1], BA; Pauline Whelan[1], BA, MSc, PhD; William Newman[3], BSc, MBChB, MA, PhD

[1]Centre for Health Informatics, Division of Informatics, Imaging and Data Science, University of Manchester, Manchester, United Kingdom
[2]Pankhurst Institute for Health Technology Research and Innovation, University of Manchester, Manchester, United Kingdom
[3]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester, United Kingdom
[4]Division of Evolution, Infection and Genomics, School of Biological Sciences, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Videha Sharma, BMBS, BMedSci, PhD
Centre for Health Informatics
Division of Informatics, Imaging and Data Science
University of Manchester
Oxford Road
Manchester, M13 9PL
United Kingdom
Phone: 44 7735360958
Email: videha.sharma@manchester.ac.uk

Abstract

Background: Pharmacogenetics can impact patient care and outcomes through personalizing the selection of medicines, resulting in improved efficacy and a reduction in harmful side effects. Despite the existence of compelling clinical evidence and international guidelines highlighting the benefits of pharmacogenetics in clinical practice, implementation within the National Health Service in the United Kingdom is limited. An important barrier to overcome is the development of IT solutions that support the integration of pharmacogenetic data into health care systems. This necessitates a better understanding of the role of electronic health records (EHRs) and the design of clinical decision support systems that are acceptable to clinicians, particularly those in primary care.

Objective: Explore the needs and requirements of a pharmacogenetic service from the perspective of primary care clinicians with a view to co-design a prototype solution.

Methods: We used ethnographic and think-aloud observations, user research workshops, and prototyping. The participants for this study included general practitioners and pharmacists. In total, we undertook 5 sessions of ethnographic observation to understand current practices and workflows. This was followed by 3 user research workshops, each with its own topic guide starting with personas and early ideation, through to exploring the potential of clinical decision support systems and prototype design. We subsequently analyzed workshop data using affinity diagramming and refined the key requirements for the solution collaboratively as a multidisciplinary project team.

Results: User research results identified that pharmacogenetic data must be incorporated within existing EHRs rather than through a stand-alone portal. The information presented through clinical decision support systems must be clear, accessible, and user-friendly as the service will be used by a range of end users. Critically, the information should be displayed within the prescribing workflow, rather than discrete results stored statically in the EHR. Finally, the prescribing recommendations should be authoritative to provide confidence in the validity of the results. Based on these findings we co-designed an interactive prototype, demonstrating pharmacogenetic clinical decision support integrated within the prescribing workflow of an EHR.

Conclusions: This study marks a significant step forward in the design of systems that support pharmacogenetic-guided prescribing in primary care settings. Clinical decision support systems have the potential to enhance the personalization of medicines, provided they are effectively implemented within EHRs and present pharmacogenetic data in a user-friendly, actionable, and standardized format. Achieving this requires the development of a decoupled, standards-based architecture that allows for the separation of data from application, facilitating integration across various EHRs through the use of application programming interfaces (APIs). More globally, this study demonstrates the role of health informatics and user-centered design in realizing the

XSL•FO
RenderX

potential of personalized medicine at scale and ensuring that the benefits of genomic innovation reach patients and populations effectively.

(J Med Internet Res 2024;26:e49230) doi:10.2196/49230

KEYWORDS

personalized medicine; genomic medicine; pharmacogenetics; user-centred design; medical informatics; clinical decision support systems; side effect; information technology; data; primary care; health informatic

Introduction

The effectiveness, safety, and tolerability of medicines vary considerably across the population, a phenomenon that can be challenging for health care systems, clinicians, and patients. There are several factors that contribute to this including the chosen dosing strategy, the accuracy of the initial diagnosis, or individual factors such as medical comorbidities, age, polypharmacy, and adherence issues [1]. There is now strong evidence that common genetic variation also plays a significant role, a concept known as pharmacogenetics [2].

There are many laboratory technologies that can be used for pharmacogenetic testing. The challenge is returning these results to physicians in a clinically relevant format and clinically relevant timeframe at the point of prescribing. Integrating the results of a pharmacogenetic test within a model of service delivery to inform prescribing practice (hereafter "pharmacogenetics") could lead to more accurate prescribing, improving outcomes for patients and ensuring better use of health care resources [3]. At many centers globally, pharmacogenetics is part of routine clinical practice. Although there is notable heterogeneity in the design and implementation of these programs, most make use of pharmacogenetic gene panels as their modality for testing [4]. These panels can survey a range of genes simultaneously, meaning the data produced has utility beyond a single prescription at a single time point.

Consider a 45-year-old patient presenting to their primary care physician requiring an antidepressant medicine in March 2023. Providing the genetic data can be returned to the clinician in a readily interpretable format and in a relevant timeframe, the patient's therapy could be optimized [5]. However, the patient might go on to present in 2033 at 55 years of age, requiring statin therapy for high cholesterol. That pharmacogenetic test, undertaken a decade earlier, could be used to guide the selection and dosing of the patient's statin [6]. As such, for pharmacogenetics to deliver maximal benefit, the data must be stored and readily accessible throughout a patient's life (Figure 1).

Figure 1. Pharmacogenetic data collected through panel testing may be used throughout the lifetime of a patient at repeated clinical encounters. PGx: pharmacogenetics.

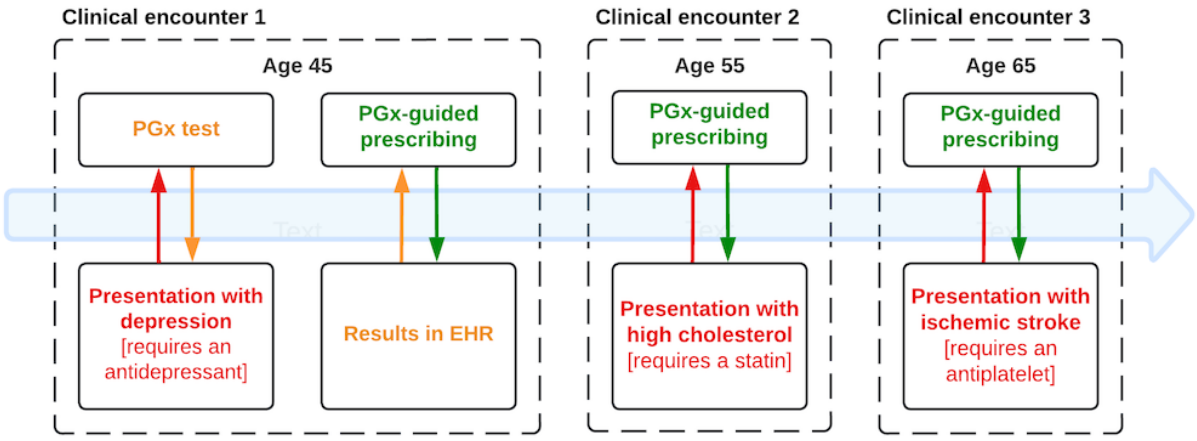

Many institutions have approached the challenge of persisting data across an individual's life by integrating pharmacogenetic data within their electronic health care records (EHRs) [7]. This approach makes use of EHR clinical decision support functionality, including alerts or pop-ups. These trigger when a medicine is prescribed where there are clinically actionable pharmacogenetic prescribing recommendations available in that patient's records. We recently published an international review of pharmacogenetics implementation programs and identified that most are undertaken at specialist centers, primarily based in the United States where health care institutions typically use a single EHR provider [4]. This consistency removes certain challenges related to the interoperability of any solution, something which will be a key issue in the National Health Service (NHS, United Kingdom) where there is a diverse EHR landscape with multiple vendors [8].

At the time of writing, there has been only limited implementation of panel-based pharmacogenetic testing in the NHS. The underlying reasons behind this have been discussed extensively in the past, but a lack of awareness of pharmacogenetics and an inability to return data to clinicians in a suitable format are 2 key barriers [9]. The majority of prescriptions are issued in primary care, and previous studies have suggested that clinicians in this setting are enthusiastic to embrace pharmacogenetic-guided prescribing [10,11]. Although there has been some research into the potential of

XSL•FO
RenderX

pharmacogenetics in primary care, there is no granular understanding of how primary care prescribers might want to interact with these data in their EHRs [12]. As an increasing volume of pharmacogenetic data is generated across the next decade, understanding stakeholder preferences, and designing clinical decision support systems for the presentation of these data will be essential to ensure uptake of pharmacogenetic-guided prescribing is maximized.

This study aims to apply a co-design approach to understand how pharmacogenetics might be best implemented in clinical practice in primary care. The objectives of this study were to (1) understand the requirements of a pharmacogenetic service from the perspective of primary care clinicians and (2) use these functional requirements to iteratively design interactive prototypes. The overall aim was to use these findings and prototypes to inform the design and technical architecture of a scalable service that can deliver pharmacogenetics across a range of diverse health care organizations in the NHS.

Methods

Overview

This was an iterative co-design study that gathered qualitative data to inform the user interface and user experience for prescribers using pharmacogenetics in primary care. Co-design of digital health interventions supports the active participation of the target end users throughout the development of the intervention aiming to ensure that the end product matches user needs. Co-design has been described as supporting "collective creativity" and is increasingly advocated across the digital health technologies landscape [13,14]. We used ethnographic observations using the "think aloud" method to capture current practices. We subsequently ran workshops with primary care clinicians to explore how pharmacogenetics may be introduced into the workflow and iteratively prototyped a solution that prioritized safety and effectiveness. The overall aim was to better understand perceived barriers to the real-world implementation of pharmacogenetics and identify strategies that would promote adoption.

Context

This study was the NHS North West Genomic Medicine Service Alliance based in Manchester, United Kingdom. The alliance hosts 1 of the 7 Genomic Laboratory Hubs and serves a wide geographical area covering a population of over 9 million citizens. This study was part of a wider project titled "Pharmacogenetics Roll Out: Gauging Response to Service" (PROGRESS), which is a pilot program to implement pharmacogenetic testing at 4 preliminary primary care centers in the North West England.

Participants

The target end users for delivering pharmacogenetics were expected to be general practitioners (GPs) and primary care pharmacists. We therefore advertised for participants within these communities through national interest groups and social media platforms. In addition to open advertising, we used mailing lists of clinical networks in primary care and pharmacies covering the whole of the United Kingdom. We received a total of 88 email responses from interested participants. Inclusion and exclusion criteria for selecting participants are summarized in Textbox 1.

For think-aloud ethnographic observations, we selected participants with whom the research team had a previous research relationship. This ensured we had active initial engagement and an even representation of professional and personal backgrounds and participation from end users across a wide geography within the United Kingdom. For workshops, we emailed the pool of interested participants with the date and time of the workshops. We were keen to include participants with little previous experience and or expertise in pharmacogenetics. We included all eligible participants that expressed interest and workshop attendance was determined by who could attend the dates or times set.

XSL•FO
RenderX

Textbox 1. Participant inclusion and exclusion criteria for the co-design process.

Inclusion criteria

- Clinical area
  - Primary care and community pharmacy
- Professional role
  - General practitioners and pharmacists
- Career stage
  - Qualified general practitioners, fully registered pharmacists, and general practitioner specialty trainees
- Geography
  - United Kingdom-wide

Exclusion criteria

- Clinical area
  - Hospital care and social care
- Professional role
  - Other health care professionals and patients
- Career stage
  - Foundation doctors on general practitioner rotation and preregistration pharmacists
- Geography
  - International

Co-Design

A co-design approach was adopted by the multidisciplinary project team to (1) understand the system end users by identifying their needs, goals, and the context of prescribing in primary care, (2) identify end-user requirements for using pharmacogenetics-based prescribing in their daily practice, (3) understand how pharmacogenetics can best be integrated into the clinical workflow, and (4) identify the barriers to embedding pharmacogenetics-based prescribing into routine practice.

Co-design allows expertise from multiple stakeholders to organically contribute to understanding a problem and developing novel solutions. Co-design aims to move from designing for users to designing with users. It relies on participatory creativity, lived experience, and open ideation throughout the design process [15]. With the increasing complexity of health care and the competing interests of stakeholders, there is a growing recognition of how user-centered design methods may be a powerful driver of change [16].

Ethnographic Observations

We arranged a Zoom (Zoom Video Communications) interview with selected participants to observe their on-screen prescribing workflows and asked them to verbalize their actions and thoughts (“think aloud”). During the interviews, the participants shared their screens, which we recorded for research purposes. We asked the participants to use the sandbox environment of the EHR and observed them logging in, selecting a patient, and prescribing medications, among other routine clinical tasks. We asked them to share any pain points or positive experiences.

Based on the observed workflows, we created an early prototype using Adobe XD, which represented the current EHRs in use in primary care in the NHS. We used the prototype as a starting point in the subsequent workshops and allowed participants to co-design how pharmacogenetics may be best integrated into the user interface and workflows.

User Workshops

We undertook 3 user workshops using Zoom. Each workshop had a different topic guide and built upon data captured in the preceding one. We purposefully invited new participants to each workshop to allow a range of perspectives to inform our final design. A bespoke slide deck was made for each workshop to facilitate the discussion and following consent from all participants, the meeting was recorded. The titles for the workshops were as follows: (1) workshop 1: personas and brainstorming; (2) workshop 2: workflows and potential for clinical decision support (alerts); and (3) workshop 3: feedback on prototype including alert timing and content.

Iterative Prototyping

Discussion in the workshops was informed by iterative prototyping developed using Adobe XD and presented to

XSL•FO
RenderX

participants using the Marvel software system. High-fidelity design prototypes were developed by the team user experience designer based on the ethnographic observations and iteratively refined based on participant feedback after each workshop. The prototypes visualized how a pharmacogenetics-based clinical decision support system could be integrated into an EHR. The prototypes visualized the planned changes to the EHR enabling participants to see how pharmacogenetics components could be displayed and actioned. Figure 2 is a flowchart summary of the iterative co-design process used in this study.

Figure 2. Iterative co-design process demonstrating how different data capture methods were used and co-ordinated to inform the design of an interactive prototype.

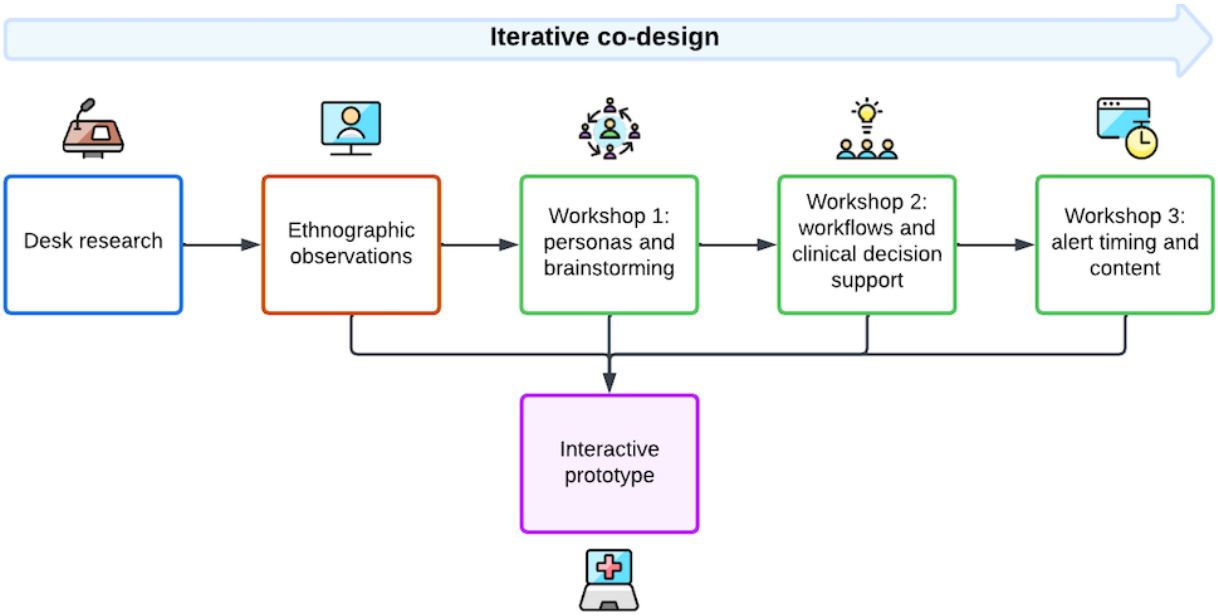

Agile Data Analysis

The team combined elements of affinity diagramming with a collaborative Agile approach to analyze the detailed notes taken during the workshops. Affinity diagramming is a user research method commonly used to group patterns and themes in qualitative data [17]. The user research team used affinity diagramming to group key themes from the notes after each workshop. Key themes were recorded in Microsoft Excel. After key themes had been identified from the workshop notes, the video recordings were used to revisit the workshops and expand the categories. Key themes and data summaries from the workshop were developed and circulated to the wider research team for review. The project team then met to review and discuss the key themes and workshop insights collectively. Key findings were updated during the group team meeting and were used to inform the preparation of the subsequent workshop (for workshops 1 and 2) and to develop the project report.

Ethical Considerations

This study voluntarily invited health care professionals to share experiences and perspectives directly related to their routine work. Such research does not require explicit ethical approval as per the NHS Research Ethics Committee decision tool (results of the system attached as Multimedia Appendix 1). All participants provided informed verbal consent before interviews or workshops. Secondary analysis of research data beyond the current study was not included in the consent. All data collected from participants was treated confidentially and stored on the secure research network at the University of Manchester. In addition, all data were deidentified apart from the professional role of participants in relation to the data contributed to this study. All participants were offered compensation for their contribution to the study in the form of a US $40 web-based shopping voucher. Payment rates were set to align with the National Institute for Health and Social Care Research guidance on reimbursement for patients and public contributors to research [18].

Sample Sizes and Theme Saturation

Research in usability and user research has recommended iterative testing with 5 users per testing round, noting that 85% of usability issues can be identified by 5 users [19]. Although our qualitatively informed approaches in this study did not require definitive sample sizes and we did not design our recruitment strategy to achieve a predefined number of participants, we used the overarching usability samples as guiding principles (5 participants in the ethnographic observations; at least 5 in the workshops). Our overarching aim was to achieve a diversity of engaged practitioners (GPs and pharmacists) at different career stages who could guide us on the design and implementation practicalities of our proposed designs. To ensure our data collection was sufficient to capture the diversity of experience, we did not define the number of user research workshops up front, but instead planned each workshop following preliminary analysis of the previous stage. This enabled the research team to review and discuss findings at each stage and to identify the unknowns that needed further discussion with practitioners in subsequent workshops. The team identified saturation as the point at which no new actionable themes or insights were emerging from the workshops or interviews. When this point was reached, no further data collection activities were conducted.

XSL•FO
RenderX

Results

Participants

At the start of each workshop, participants were asked to introduce themselves and provide some background about their experience with pharmacogenetics and digital health. Participants reported varied career stages (eg, from GP trainee to GP with >30 years of experience) and different levels of familiarity with pharmacogenetics and digital health technologies (Table 1). For each co-design element, we included the following participants (total N=24): ethnographic observations – GPs (3) and pharmacists (2); workshop 1 – GPs (3) and pharmacists (2); workshop 2 – GPs (2) and pharmacists (3); workshop 3 – GPs (5) and pharmacists (4).

Four participants used the EMIS Web HER and 1 participant used the SystmOne EHR. Through thinking aloud, users described their experience of logging in and selecting a patient as part of a clinical encounter and described initial frustrations with connectivity and start-up times. One user proactively configured their EHR system with shortcuts, templates, and filters to enhance their workflow. With regards to the prescribing workflow, all users demonstrated a similar on-screen workflow of a pop-up box appearing at the center of the EHR when clicking the “add drug” button. However, users did highlight that the preceding screen from which the prescribing workflow is launched may differ. For example, a user may launch the prescribing workflow from the “current medication” or the “current encounter” screen. From where the user decides to enter the prescribing workflow did not impact the pop-up box or the subsequent prescribing user experience.

We used the recordings of these observational sessions to mock up an EHR for exploring how pharmacogenetics may be introduced into the workflow. This was presented to subsequent workshop participants.

Table 1. Summary of participants by career level, pharmacogenetics, and digital health experience.

Workshop	Career level: junior	Career level: mid	Career level: senior	PGx[a] experience	Digital health experience
1	2	1	2	1	0
2	1	1	3	3	3
3	0	1	8	4	1

[a]PGx: pharmacogenetics.

Co-Design Workshop Results

A total of 19 people participated across the 3 co-design workshops. Each group co-design workshop lasted 1.5 hours and was facilitated by the lead author of this paper. They were further supported by a clinical geneticist and the user research team (1 senior user experience designer and 1 senior digital health researcher).

The key themes from the 3 workshops are summarized in Table 2. The summary column illustrates how each theme was discussed in the workshop at a high level. The summary is not a direct quote from participants; it is a simplified, anonymized summary of participants’ comments on the theme.

Participants in workshop 1 emphasized the importance of developing a pharmacogenetics-prescribing system that was easily usable by all key prescribers, including health care professionals in training. Integrating the system into the EHR was identified as a mandatory requirement. However, the complexity of clinical prescribing pathways was discussed at length in the workshop. Participants described how prescribing can happen in many ways and that the pharmacogenetics-prescribing system needed to be embedded at multiple points in the EHR to ensure that the recommendations were not missed. The importance and potential of pharmacogenetics to improve patient care was widely recognized by participants. National policy-level incentivization was discussed to ensure that the system would be used, particularly given the busy context of primary care prescribing.

Workshop 2 participants echoed the need expressed in workshop 1 to embed the pharmacogenetics system within the EHR at multiple points, both as alert pop-ups and by integrating pharmacogenetics recommendations into routine prescription reviews. Participants described how prescribers can have different prescribing workflows and how this creates complexity in the implementation of any singular pharmacogenetics-based prescribing approach. Like the comments in workshop 1, clear incentivization to use a pharmacogenetics-based prescribing system was identified as necessary for prescribing teams who are already very busy. Participants identified that the prescribing recommendations should provide direct links to the underlying evidence base to instill confidence in the credibility of the recommendations. Concerns about the clinical safety of prescribing recommendations were also raised; it was suggested that safety assurances should be embedded in the EHR.

While many of the comments in workshop 3 echoed the participants of workshops 1 and 2, particularly around the importance of good usability and linking back to the evidence base for recommendations, several safety concerns were also raised. Specific risks and concerns focused on protecting prescribers from legal challenges if recommendations were not correctly followed. To mitigate these risks, participants advised enabling a full audit trail of past decisions to be reviewed and suggested that it should be possible to record the reasons for overriding prescribing recommendations.

XSL•FO
RenderX

Table 2. Summary of key themes identified across 3 user workshops.

Workshop number	Theme	Summary
1	Usability	Pharmacogenetics must be easy to access and integrated into the EHR[a]. There are multiple EHRs
1	Usability	Messaging around pharmacogenetics must be clear and accessible; directions must be easy to follow even by trainees
1	Integration into clinical pathway	Multiple options should be available–flags in EHR and a thorough review of a patient list
1	Implementation complexity	Time is critical for prescribers; they are very busy
1	Incentivization	Incentivization for pharmacogenetics is important
1	Implementation complexity	Do not underestimate how complex this will be to implement in the real world
1	Importance of pharmacogenetics	There are compelling arguments why people would want to adopt this
2	Integration into clinical pathway	Pop-ups and review lists with pharmacogenetics information are useful
2	Incentivization	Incentivization for pharmacogenetics is important; ideally, incentivization would come from a national level
2	Integration into clinical pathway	Integrate with existing systems and mechanisms where possible. Where exactly to locate within the EHR was not universally agreed and noted that prescribers interact differently with the EHRs
2	Credibility of system	Ensuring trustworthiness and credibility of pharmacogenetics information is important
2	Safety concerns	Reassurances about patient safety are important for people engaging with pharmacogenetics
2	Implementation complexity	Sharing data between EHRs (interoperability) is very difficult–a data infrastructure that supports the movement of pharmacogenetic data between organizations is required
3	Integration into clinical pathway	Confirmation that (1) pop-ups are useful but (2) pop-up fatigue is a problem
3	Implementation complexity	Quality of EHR data can be poor; reliance on EHR data for clinical decision-making can be difficult
3	Credibility of system	Having more detailed information available with links to evidence is very important
3	Risk or safety concerns	Making details of previous pharmacogenetics decisions available is important so that this can be reviewed when prescribing
3	Risk or safety concerns	Overriding pharmacogenetics recommendations should require action that is recorded with clear reasons for overriding
3	Risk or safety concerns	What are the legal consequences of overriding a pharmacogenetics recommendation? How do we protect prescribers?

[a]EHR: electronic health records.

Iterative Prototyping Outputs

Across all workshops, participants reviewed high-fidelity design prototypes of how pharmacogenetic recommendations could be integrated into an EHR system. The prototype review of a realistic EHR enabled participants to try out a real-life prescribing scenario and to visually identify any barriers to the designed pharmacogenetics-prescribing process. A screenshot of the prototype is shown in Figure 3. The prototype was iterated and updated between each workshop based on feedback and gathered requirements. We have subsequently publicly published the prototype screens on figshare (Figshare LLC) [20]. Prototyping has subsequently informed the business case for the procurement of a technology partner to develop the technical solution to implement this system as part of the ongoing PROGRESS project.

XSL•FO
RenderX

Figure 3. Prototype screens showing where pharmacogenetic results could be surfaced to users. Screen (2) shows the results within the prescribing box and (3) shows specific prescribing guidance surfaced as a "high severity warning" once a medicine has been selected. Screen (4) is an alternative interruptive alert surfacing the guidance if a user ignores or misses the first warning (all data is fictitious).

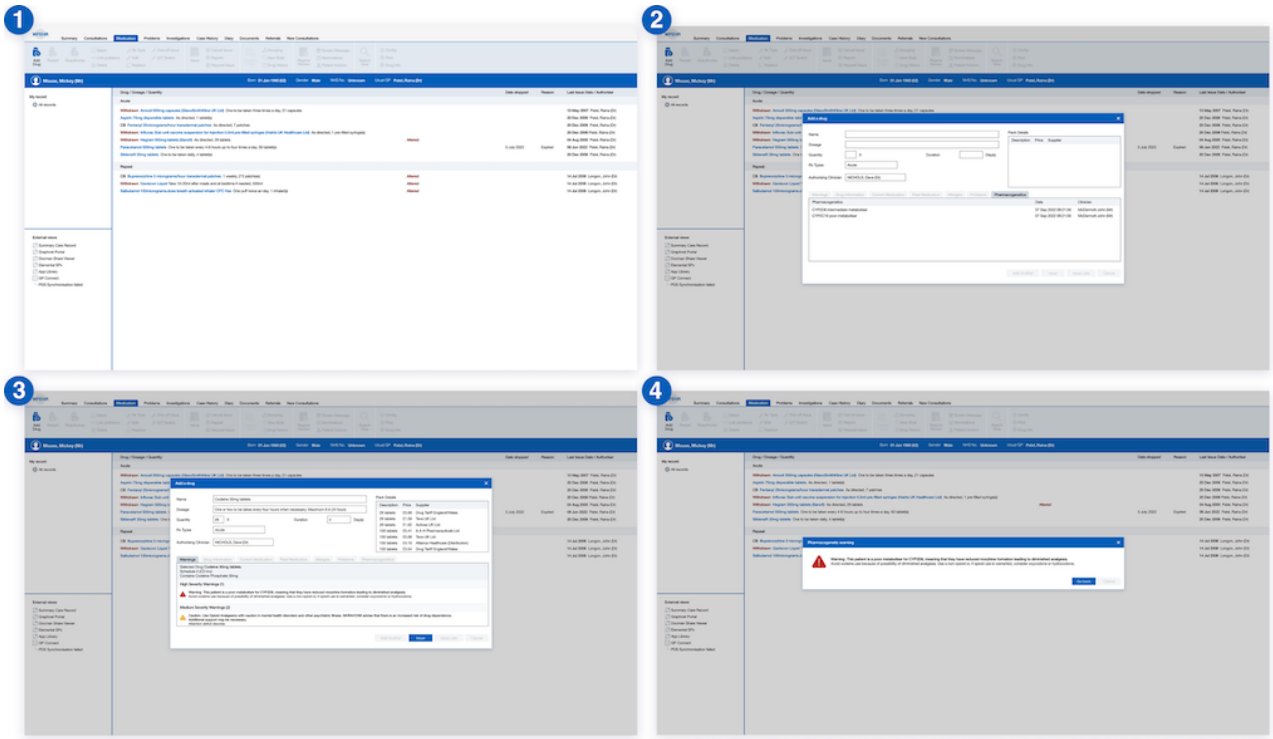

Discussion

Principal Findings

This study applied a co-design approach to better understand the needs and requirements of frontline clinicians for a pharmacogenetic clinical service. We identified that, in order to be used in practice, pharmacogenetic data must be incorporated within existing EHRs and prescribing workflows. The information presented to primary care clinicians must be clear and understandable as the service will be used by a range of end users with varying knowledge of pharmacogenetics. Finally, the prescribing recommendations should be authoritative, with direct links to the underlying evidence base to provide confidence in the validity of results. Based on these functional requirements, we designed an interactive prototype, which demonstrates pharmacogenetic clinical decision support integrated with a primary care EHR system. This study demonstrated a novel method to overcome the barriers to the implementation of pharmacogenetics in routine clinical practice.

Prescribing Clinical Decision Support Systems in Primary Care

Clinical decision support systems play a significant role in primary care with most EHRs providing a range of alerts to help clinicians make better, evidence-based decisions. They have been shown to improve the quality and safety of care with a recent review by Sutton et al [21] suggesting a potential reduction in patient harm. However, the authors also caution against the challenges of designing and developing systems that are adopted and have a sustained impact [21].

Several prescribing clinical decision support systems have been reported in primary care. For example, Rogero-Blanco et al [22] describes the implementation of a tool to identify patients at risk of adverse drug reactions due to polypharmacy. This tool was embedded within a single EHR and worked on an overall patient list, rather than individual patient records [22]. An example of active alerts was implemented by van Staa et al [23] to inform antibiotic prescribing in primary care with a view to improve adherence to best practice guidelines. Although such tools are useful, their impact on care and outcomes can be limited. A recent meta-analysis on clinical decision support systems by Kwan et al [24] showed only small to moderate improvements in clinical endpoints. In addition, due to the significant heterogeneity, it remains challenging to define the predictors of impactful clinical decision support systems [24]. NHS England recently published national guidance for the implementation of clinical decision support systems in the NHS, highlighting a design-led approach as 1 of the key recommendations for success [25].

This study is the first of its kind to apply an iterative codesign methodology to inform the needs and requirements of a pharmacogenetic clinical decision support system. This method has been shown to work in other clinical areas and by working with end users, we generated ideas that may reduce barriers to adoption [26]. By converting these into high-fidelity prototypes, we were able to drive visual engagement with workshop participants resulting in valuable feedback. This has established the user requirements of a pharmacogenetic clinical decision support system and also helps to inform the technical back-end infrastructure and integrations required to realize the desired user experience.

XSL•FO
RenderX

Implications for Clinical Practice

The evidence for the impact of pharmacogenetic-based prescribing is well-established and continues to expand. The recent Pre-emptive Pharmacogenomic Testing for Preventing Adverse Drug Reactions (PREPARE) study found, as part of an international multicenter randomized controlled trial in nearly 7000 participants, that pharmacogenetics reduced the risk of adverse drug events by at least 30% [27]. The authors make a compelling case for widespread adoption and development of infrastructure that supports clinical implementation. In the United Kingdom, the joint Royal College of Physicians and British Pharmaceutical Society report on pharmacogenetics calls for personalized testing for the safety and effectiveness of common medicines to be offered throughout the NHS [28]. Our study provides a critical perspective from frontline clinicians who would be expected to use pharmacogenetics as it becomes more mainstream and part of routine clinical practice. This aligns with the literature that highlights the clear requirement to integrate pharmacogenetic data within EHRs to drive clinical decision support systems [29]. We identified that interoperability and data sharing across health care organizations remains a key implementation challenge for personalized medicine.

For health IT solutions to effectively impact patient care, they must provide a view of clinical data that complements the clinical workflow [30]. It became clear through prototyping that primary care clinicians did not consider being able to view pharmacogenetic results on a routine basis as a priority. Pharmacogenetics only became relevant during the prescribing workflow. Furthermore, to provide meaningful clinical decision support, interruptive alerts were only relevant if the prescriber selected a drug for which the patient had a significant gene-medicine interaction, and a therapeutic implication was relevant. This complements previous work by Keeling et al [31], that merely storing pharmacogenetic data as static information in the EHR does not maximize the benefits of the intervention. Rather, querying a patient's pharmacogenetic data and returning relevant results at the time of prescribing enhances the potential for greater quality decision-making and complements the current workflow.

The findings from this study suggest that clinicians are eager to engage and use pharmacogenetic results to inform prescribing decisions. The general awareness of the roles genomics plays in the delivery of routine clinical services is increasing. A systematic review by Qureshi et al [32] on the barriers and facilitators of pharmacogenetics in primary care suggested that implementation may be facilitated by educating clinicians about the benefits of pharmacogenetic testing. A UK-based study on perceptions of primary care clinicians revealed interesting concerns about the cost-effectiveness of pharmacogenetics and the risks of direct-to-consumer consumer testing [33]. Reflecting on these findings as well as on the results from a structured review of implementation models by Hayward et al [10], it is clear that any pharmacogenetic program in primary care must have an educational component as one of its core pillars in order to be successful [10].

Implications for Policy and Strategy

Based on the functional needs to provide a query-response of pharmacogenetic data in a disparate landscape of EHRs, a decoupled architecture, which separates data from application may provide the appropriate infrastructure to implement pharmacogenetics at scale. A vendor-neutral data platform for pharmacogenetic data that uses standards-based application programming interfaces may allow a range of EHR and clinical decision support system providers to integrate pharmacogenetics within workflows. As highlighted by Blagec et al [34] as part of the Ubiquitous Pharmacogenomics consortium, this is dependent on the development of robust data standards that allow pharmacogenetic data to be stored in an open and structured manner [34]. Building on this, a systematic review on the inclusion of standards in the implementation of pharmacogenetics by Roosan et al [35] found that only 9 out of 32 articles mentioned standards, with only 4 providing solutions for the lack of interoperability. The authors discussed how interoperability is clearly essential for widespread implementation of pharmacogenetics, and how a lack of focus on standardized data remains a significant challenge [35]. Policymakers should thus promote the development standards including open EHR data models and Fast Healthcare Interoperability Resources through established standards development organizations such as the Global Alliance for Genomics and Health (GA4GH) [36].

The usability of EHRs is known to impact clinicians' experience of practice and the potential risk of burnout [37]. It is therefore imperative to design solutions that exert minimal additional cognitive load on end users while maximizing the added value technology can provide to clinical workflows. By involving clinicians early in our study, we identified potential barriers, which may be overcome by adopting a user-centered approach to the design process. This included presenting pharmacogenetic data within existing workflows and supporting clinical decision-making in real-time. Such so-called "active" clinical decision support systems have been reported extensively in the literature, including their application in pharmacogenetics [38]. However, implementers must be conscious of the risk of alert fatigue. Clinicians presented with inappropriate alerts, or an overwhelming number of alerts may distrust, disagree with, or ignore them [39]. Alerts should appear at the right time and provide context-relevant information that enhances clinical decision making, in a practical, standardized format. Importantly, prescriptive recommendations need to be reflective of local formulary and agreed upon by appropriate medicines optimization committees and medicines governance bodies to limit the burden of conflicting alerts. Future pharmacogenetic clinical decision support systems should publish data on alert responses (acceptance vs override) to provide a real-world evaluation of the effectiveness of the implementation.

Limitations

We sought input from health care professionals across the United Kingdom during this study. However, having included a limited number of participants, the generalizability of our findings may be limited. Results will particularly apply to a UK public health care setting, with a focus on primary care, rather than secondary

XSL•FO
RenderX

care, where requirements may slightly differ. Alternative methods, which capture more responses across a wider range of users and geographies, such as surveys, may be used in the future to complement and validate this study. We used remote think-aloud methods for our ethnographic observations through screen sharing, and, as a result, we may not have been able to fully capture additional details, such as the physical environment or distractions within the clinic room. Finally, the areas of genomics and digital health are fast evolving [40]. The way prescribing clinical decision support may be implemented may change with emerging technology or service models, such as patient portals or tools embedded within pharmacies rather than at the point of prescribing.

Conclusion

The implementation of pharmacogenetics in primary care has the potential to improve patient care and save health care costs. This study demonstrated how systems could be designed that are acceptable to primary care clinicians increasing the probability of achieving intended benefits. Active clinical decision support systems (alerts) are likely to personalize medication selection for patients if implemented within existing EHRs and designed to present information in a practical and standardized format. To achieve this, a decoupled architecture, which separates data storage from clinical application is preferable as it will allow the integration of pharmacogenetics across a range of EHRs using application programming interfaces. Alongside user experience and technical development, there remains a need to synchronously educate and upskill nongenomics health care professionals in pharmacogenetics and integrate within routine medicines governance pathways to achieve widespread adoption and impact on care outcomes.

Acknowledgments

We would like to express our sincere gratitude to all the participants who generously gave their time and shared their experiences for this study. This work was supported by the National Health Service England Genomics Medicine Service Transformation Program and the National Institute for Health and Care Research Manchester Biomedical Research Centre (NIHR203308).

Authors' Contributions

VS was responsible for conceptualization, methodology, investigation, writing– original draft, review, and editing. JMD contributed to conceptualization, investigation, writing–original draft, review, and editing. JK managed investigation, writing–review and editing. SF contributed to software, project administration, writing–review, and editing. PW managed investigation, methodology, formal analysis, writing–original draft, review, and editing. WN was responsible for writing–review and editing and supervision.

Conflicts of Interest

PW is a director and shareholder of CareLoop Health Ltd, a digital mental health company. WN is the cofounder of Fava Health, a health technology consultancy. The other authors declare no conflicts of interest.

Multimedia Appendix 1
National Health Service Research Ethics Committee decision tool result.
[PDF File (Adobe PDF File), 195 KB - jmir_v26i1e49230_app1.pdf]

References

1. Sweeney G. Variability in the human drug response. Thromb Res Suppl 1983;4:3-15. [doi: 10.1016/0049-3848(83)90353-5] [Medline: 6356464]
2. Pirmohamed M. Pharmacogenomics: current status and future perspectives. Nat Rev Genet 2023;24(6):350-362. [doi: 10.1038/s41576-022-00572-8] [Medline: 36707729]
3. Roden DM, Van Driest SL, Mosley JD, Wells QS, Robinson JR, Denny JC, et al. Benefit of preemptive pharmacogenetic information on clinical outcome. Clin Pharmacol Ther 2018;103(5):787-794 [FREE Full text] [doi: 10.1002/cpt.1035] [Medline: 29377064]
4. McDermott JH, Wright S, Sharma V, Newman WG, Payne K, Wilson P. Characterizing pharmacogenetic programs using the consolidated framework for implementation research: a structured scoping review. Front Med (Lausanne) 2022;9:945352 [FREE Full text] [doi: 10.3389/fmed.2022.945352] [Medline: 36059837]
5. Hicks JK, Bishop JR, Sangkuhl K, Müller DJ, Ji Y, Leckband SG, Clinical Pharmacogenetics Implementation Consortium. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. Clin Pharmacol Ther 2015;98(2):127-134 [FREE Full text] [doi: 10.1002/cpt.147] [Medline: 25974703]
6. Cooper-DeHoff RM, Niemi M, Ramsey LB, Luzum JA, Tarkiainen EK, Straka RJ, et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and statin-associated musculoskeletal symptoms. Clin Pharmacol Ther 2022;111(5):1007-1021 [FREE Full text] [doi: 10.1002/cpt.2557] [Medline: 35152405]

XSL•FO
RenderX

7. Hicks J, Dunnenberger H, Gumpper K, Haidar C, Hoffman J. Integrating pharmacogenomics into electronic health records with clinical decision support. Am J Health Syst Pharm 2016;73(23):1967-1976 [FREE Full text] [doi: 10.2146/ajhp160030] [Medline: 27864204]
8. Robertson A, Cresswell K, Takian A, Petrakaki D, Crowe S, Cornford T, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: qualitative analysis of interim results from a prospective national evaluation. BMJ 2010;341:c4564 [FREE Full text] [doi: 10.1136/bmj.c4564] [Medline: 20813822]
9. Turner RM, Newman WG, Bramon E, McNamee CJ, Wong WL, Misbah S, et al. Pharmacogenomics in the UK national health service: opportunities and challenges. Pharmacogenomics 2020;21(17):1237-1246. [doi: 10.2217/pgs-2020-0091] [Medline: 33118435]
10. Hayward J, McDermott J, Qureshi N, Newman W. Pharmacogenomic testing to support prescribing in primary care: a structured review of implementation models. Pharmacogenomics 2021;22(12):761-776 [FREE Full text] [doi: 10.2217/pgs-2021-0032] [Medline: 34467776]
11. Prescribing Costs in Hospitals and the Community - England.: NHS Business Services Authority; 2022. URL: https://tinyurl.com/5ymmd9j3 [accessed 2023-08-03]
12. Gammal RS, Berenbrok LA, Empey PE, Massart MB. Documenting pharmacogenomic test results in electronic health records: practical considerations for primary care teams. J Pers Med 2021;11(12):1296 [FREE Full text] [doi: 10.3390/jpm11121296] [Medline: 34945768]
13. Bevan Jones R, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
14. Sanders EB, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008;4(1):5-18. [doi: 10.1080/15710880701875068]
15. The Field Guide to Human-centered Design: Design Kit. San Francisco, California: IDEO; 2015.
16. Ku B, Lupton E. Health Design Thinking, second edition: Creating Products and Services for Better Health. Cambridge, MA: MIT Press; 2022:0262543605.
17. Marsh S. User Research: A Practical Guide to Designing Better Products and Services. New York, NY: Kogan Page Publishers; 2018.
18. Payment guidance for researchers and professionals. National Institute for Health and Care Research. URL: https://www.nihr.ac.uk/documents/payment-guidance-for-researchers-and-professionals/27392 [accessed 2023-02-03]
19. Nielsen J, Landauer T. A mathematical model of the finding of usability problems. 1993 Presented at: Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems; 1993; New York, NY.
20. Sharma V, Mcdermott J, Keen J, Foster S, Whelan P, Newman W. Software design prototypes for a pharmacogenetics service for primary care in England. Manchester, UK: The University of Manchester; 2023. URL: https://doi.org/10.48420/21842001.v1 [accessed 2024-07-03]
21. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
22. Rogero-Blanco E, Lopez-Rodriguez JA, Sanz-Cuesta T, Aza-Pascual-Salcedo M, Bujalance-Zafra MJ, Cura-Gonzalez I, MultiPAP Group. Use of an electronic clinical decision support system in primary care to assess inappropriate polypharmacy in young seniors with multimorbidity: observational, descriptive, cross-sectional study. JMIR Med Inform 2020;8(3):e14130 [FREE Full text] [doi: 10.2196/14130] [Medline: 32149715]
23. van Staa T, Sharma A, Palin V, Fahmi A, Cant H, Zhong X, et al. Knowledge support for optimising antibiotic prescribing for common infections in general practices: evaluation of the effectiveness of periodic feedback, decision support during consultations and peer comparisons in a cluster randomised trial (BRIT2) - study protocol. BMJ Open 2023;13(8):e076296. [doi: 10.1136/bmjopen-2023-076296] [Medline: 37607793]
24. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020;370:m3216 [FREE Full text] [doi: 10.1136/bmj.m3216] [Medline: 32943437]
25. National Health Service. Standard General Medical Services (GMS) Contract October 2020. England: NHS; 2023 Jul. URL: https://www.england.nhs.uk/wp-content/uploads/2020/12/20-21-GMS-Contract-October-2020.pdf [accessed 2024-07-03]
26. Sharma V, Ali I, van der Veer S, Martin G, Ainsworth J, Augustine T. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. BMJ Health Care Inform 2021;28(1):e100253 [FREE Full text] [doi: 10.1136/bmjhci-2020-100253] [Medline: 33608259]
27. Swen J, van der Wouden CH, Manson L, Abdullah-Koolmees H, Blagec K, Blagus T, Ubiquitous Pharmacogenomics Consortium. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. Lancet 2023;401(10374):347-356. [doi: 10.1016/S0140-6736(22)01841-4] [Medline: 36739136]

XSL•FO
RenderX

28. Pirmohammed M, O'Donoghue D, Turner R, Magavern E, Roebuck D, Ross P. Personalised prescribing. Using pharmacogenomics to improve patient outcomes. A report from the Royal College of Physicians and British Pharmacological Society joint working party. United Kingdom: British Pharmacological Society; 2022. URL: https://www.bps.ac.uk/getmedia/b43a3dca-1bbf-4bff-9379-20bef9349a8c/Personalised-prescribing-full-report.pdf.aspx [accessed 2024-06-27]
29. Freimuth R, Formea C, Hoffman J, Matey E, Peterson J, Boyce R. Implementing genomic clinical decision support for drug-based precision medicine. CPT Pharmacometrics Syst Pharmacol 2017;6(3):153-155 [FREE Full text] [doi: 10.1002/psp4.12173] [Medline: 28109071]
30. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AAT. Context-based electronic health record: toward patient specific health care. IEEE Trans Inf Technol Biomed 2012;16(2):228-234 [FREE Full text] [doi: 10.1109/TITB.2012.2186149] [Medline: 22395637]
31. Keeling NJ, Dunn TJ, Bentley JP, Ramachandran S, Hoffman JM, Rosenthal M. Approaches to assessing the provider experience with clinical pharmacogenomic information: a scoping review. Genet Med 2021;23(9):1589-1603 [FREE Full text] [doi: 10.1038/s41436-021-01186-x] [Medline: 33927377]
32. Qureshi S, Latif A, Condon L, Akyea RK, Kai J, Qureshi N. Understanding the barriers and enablers of pharmacogenomic testing in primary care: a qualitative systematic review with meta-aggregation synthesis. Pharmacogenomics 2022;23(2):135-154 [FREE Full text] [doi: 10.2217/pgs-2021-0131] [Medline: 34911350]
33. Rafi I, Crinson I, Dawes M, Rafi D, Pirmohamed M, Walter FM. The implementation of pharmacogenomics into UK general practice: a qualitative study exploring barriers, challenges and opportunities. J Community Genet 2020;11(3):269-277 [FREE Full text] [doi: 10.1007/s12687-020-00468-2] [Medline: 32468238]
34. Blagec K, Swen JJ, Koopmann R, Cheung K, Crommentuijn-van Rhenen M, Holsappel I, Ubiquitous Pharmacogenomics Consortium. Pharmacogenomics decision support in the U-PGx project: results and advice from clinical implementation across seven European countries. PLoS One 2022;17(6):e0268534 [FREE Full text] [doi: 10.1371/journal.pone.0268534] [Medline: 35675343]
35. Roosan D, Hwang A, Law AV, Chok J, Roosan MR. The inclusion of health data standards in the implementation of pharmacogenomics systems: a scoping review. Pharmacogenomics 2020;21(16):1191-1202 [FREE Full text] [doi: 10.2217/pgs-2020-0066] [Medline: 33124487]
36. Rehm H, Page A, Smith L, Adams J, Alterovitz G, Babb L, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genom 2021;1(2):100029 [FREE Full text] [doi: 10.1016/j.xgen.2021.100029] [Medline: 35072136]
37. Melnick ER, Dyrbye LN, Sinsky CA, Trockel M, West CP, Nedelec L, et al. The association between perceived electronic health record usability and professional burnout among us physicians. Mayo Clin Proc 2020;95(3):476-487 [FREE Full text] [doi: 10.1016/j.mayocp.2019.09.024] [Medline: 31735343]
38. Bell GC, Crews KR, Wilkinson MR, Haidar CE, Hicks JK, Baker DK, et al. Development and use of active clinical decision support for preemptive pharmacogenomics. J Am Med Inform Assoc 2014;21(e1):e93-e99 [FREE Full text] [doi: 10.1136/amiajnl-2013-001993] [Medline: 23978487]
39. Khalifa M, Zabani I. Improving utilization of clinical decision support systems by reducing alert fatigue: strategies and recommendations. Stud Health Technol Inform 2016;226:51-54. [Medline: 27350464]
40. Bombard Y, Ginsburg GS, Sturm AC, Zhou AY, Lemke AA. Digital health-enabled genomics: opportunities and challenges. Am J Hum Genet 2022;109(7):1190-1198 [FREE Full text] [doi: 10.1016/j.ajhg.2022.05.001] [Medline: 35803232]

Abbreviations

EHR: electronic health record
GA4GH: Global Alliance for Genomics and Health
GP: general practitioner
NHS: National Health Service
PREPARE: Pre-emptive Pharmacogenomic Testing for Preventing Adverse Drug Reactions
PROGRESS: Pharmacogenetics Roll Out: Gauging Response to Service

Edited by T de Azevedo Cardoso; submitted 23.05.23; peer-reviewed by P McEnhill, J Walsh, D Parry; comments to author 15.09.23; revised version received 22.12.23; accepted 13.05.24; published 23.07.24.

Please cite as:
Sharma V, McDermott J, Keen J, Foster S, Whelan P, Newman W
Pharmacogenetics Clinical Decision Support Systems for Primary Care in England: Co-Design Study
J Med Internet Res 2024;26:e49230
URL: https://www.jmir.org/2024/1/e49230
doi:10.2196/49230
PMID:

XSL•FO
RenderX

©Videha Sharma, John McDermott, Jessica Keen, Simon Foster, Pauline Whelan, William Newman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Definitions and Characteristics of Patient Digital Twins Being Developed for Clinical Use: Scoping Review

David Drummond[1,2,3,4], MD, PhD; Apolline Gonsard[3], MD

[1]Health Data- and Model-Driven Knowledge Acquisition Team, National Institute for Research in Digital Science and Technology, Paris, France

[2]Faculté de Médecine, Université Paris Cité, Paris, France

[3]Department of Pediatric Pulmonology and Allergology, University Hospital Necker-Enfants Malades, Assistance Publique – Hôpitaux de Paris, Paris, France

[4]Inserm UMR 1138, Centre de Recherche des Cordeliers, Paris, France

Corresponding Author:
David Drummond, MD, PhD
Health Data- and Model-Driven Knowledge Acquisition Team
National Institute for Research in Digital Science and Technology
2 Rue Simone IFF
Paris, 75012
France
Phone: 33 144494848
Email: david.drummond@aphp.fr

Abstract

Background: The concept of digital twins, widely adopted in industry, is entering health care. However, there is a lack of consensus on what constitutes the digital twin of a patient.

Objective: The objective of this scoping review was to analyze definitions and characteristics of patient digital twins being developed for clinical use, as reported in the scientific literature.

Methods: We searched PubMed, Scopus, Embase, IEEE, and Google Scholar for studies claiming digital twin development or evaluation until August 2023. Data on definitions, characteristics, and development phase were extracted. Unsupervised classification of claimed digital twins was performed.

Results: We identified 86 papers representing 80 unique claimed digital twins, with 98% (78/80) in preclinical phases. Among the 55 papers defining "digital twin," 76% (42/55) described a digital replica, 42% (23/55) mentioned real-time updates, 24% (13/55) emphasized patient specificity, and 15% (8/55) included 2-way communication. Among claimed digital twins, 60% (48/80) represented specific organs (primarily heart: 15/48, 31%; bones or joints: 10/48, 21%; lung: 6/48, 12%; and arteries: 5/48, 10%); 14% (11/80) embodied biological systems such as the immune system; and 26% (21/80) corresponded to other products (prediction models, etc). The patient data used to develop and run the claimed digital twins encompassed medical imaging examinations (35/80, 44% of publications), clinical notes (15/80, 19% of publications), laboratory test results (13/80, 16% of publications), wearable device data (12/80, 15% of publications), and other modalities (32/80, 40% of publications). Regarding data flow between patients and their virtual counterparts, 16% (13/80) claimed that digital twins involved no flow from patient to digital twin, 73% (58/80) used 1-way flow from patient to digital twin, and 11% (9/80) enabled 2-way data flow between patient and digital twin. Based on these characteristics, unsupervised classification revealed 3 clusters: simulation patient digital twins in 54% (43/80) of publications, monitoring patient digital twins in 28% (22/80) of publications, and research-oriented models unlinked to specific patients in 19% (15/80) of publications. Simulation patient digital twins used computational modeling for personalized predictions and therapy evaluations, mostly for one-time assessments, and monitoring digital twins harnessed aggregated patient data for continuous risk or outcome forecasting and care optimization.

Conclusions: We propose defining a patient digital twin as "a viewable digital replica of a patient, organ, or biological system that contains multidimensional, patient-specific information and informs decisions" and to distinguish simulation and monitoring digital twins. These proposed definitions and subtypes offer a framework to guide research into realizing the potential of these personalized, integrative technologies to advance clinical care.

(J Med Internet Res 2024;26:e58504) doi:10.2196/58504

XSL•FO
RenderX

KEYWORDS

patient simulation; cyber-physical systems; telemonitoring; personalized medicine; precision medicine; digital twin

Introduction

Each industrial revolution has transformed the practice of medicine. The first 2 led to the development of new techniques for the industrial collection of new data (biological, imaging, etc) on the human body. The third—or digital—revolution transformed this analogue data into digital data, accelerating the exchange of information, and allowing the emergence of computer models to propose a diagnosis, establish a prognosis, and recommend a treatment [1-3].

For some contemporaries, we are currently in the midst of the fourth industrial revolution, which is the merging of the physical and digital worlds, based on 3 pillars: the internet of things, increasing connectivity, and machine learning–based decisions [4,5].

Digital twins are emblematic of this new industrial revolution. Grieves [6] introduced the concept in 2002 as a system consisting of a physical product, its virtual counterpart, and 2-way data exchange between the 2 entities. However, the term "digital twin" was first used and defined in 2010 by NASA (National Aeronautics and Space Administration) engineers as "an integrated multi-physics, multi-scale, probabilistic simulation of a vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its flying twin" [7]. Following the popularization of the concept of the fourth industrial revolution in 2015, digital twins attracted interest from all industries [5]. With the objective of reducing production times through monitoring, coordination, and control of production systems, the manufacturing industry became the most active sector in terms of research and implementation of digital twins [8]. Digital twins also extended to construction, energy, transport, smart cities, agriculture, education, and health [8].

In the health sector, the concept of digital twins attracted interest from industry, scientists, clinicians, and patients [9-11]. Besides digital twins of hospitals, there is a strong rationale for the development of digital twins of patients, as these systems could offer personalized medicine through information gathered by the internet of things, real-time adaptation of treatments through efficient connectivity, and even automation of certain aspects of medical management through predictions based on machine learning [12]. An increasing number of scientific publications claim to be developing or to have developed "digital twins" of organs, physiological systems, or patients. However, when the term "digital twin" is used in this context, it remains unclear whether it corresponds to the definitions used in industry, or whether unique concepts and characteristics emerge for patient digital twins.

The 3 previous systematic literature reviews related to digital twins in health did not focus on patient digital twins [12-14]. To contribute to the understanding of patient digital twins, we conducted a scoping review to systematically map research in this area. The following research question was formulated: "What are the definitions and characteristics of digital twins provided by research teams claiming to develop patient digital twins for clinical applications, as reported in the scientific literature?"

Methods

Overview

We conducted this scoping review in accordance with the Joanna Briggs Institute guidelines for scoping reviews [15-17] and reported it following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Multimedia Appendix 1) [18]. The authors received training on the methodology of scoping reviews using the Joanna Briggs Institute Reviewers' Manual 2020 [19]. We registered the protocol for this review at the Open Science Framework [20].

Search Strategy

We created search strategies with a medical librarian to identify published papers on patient digital twins. Our initial search was conducted in MEDLINE via PubMed, using the terms "digital twin" or "digital twins," combined with a broad range of medical-related terms pertaining to patients and precision medicine. We then adapted this search strategy for each included database and information source, namely, PubMed, Scopus, Embase, IEEE, Web of Science, and Google Scholar. When searching Google Scholar, we used incognito mode to minimize effects of past search histories and screened the first 500 results. Full search strategy is reported in Multimedia Appendix 2.

Selection Process

We collated the identified citations from the comprehensive search and uploaded them into Rayyan, a web-based review tool [21]. Two independent reviewers (DD and AG) removed duplicates and screened titles and abstracts for inclusion. All screening was performed in a masked, duplicate fashion. Any disagreements between reviewers were resolved through discussion. Reasons for exclusion at the stage of full-text screening were recorded.

Inclusion and Exclusion Criteria

We included peer-reviewed research papers in which the authors claimed to be developing or to have developed or tested a digital twin in health care, subsequently referred to as a "claimed digital twin" (CDT). Only literature published in English, French, or German was included up to August 31, 2023. Finally, to match the context of this review, we restricted included research to papers focused on developing or testing digital twins specifically representing patients, or components of patients.

Exclusion criteria were (1) animal studies, abstracts only, conference papers, reviews, editorials or correspondence, and non–peer-reviewed papers; (2) studies in languages other than English, French, or German; (3) studies unrelated to health; (4) digital twins not representing patients or part of a patient; and (5) studies limited to proposing a framework for a digital twin with no case study.

XSL•FO
RenderX

Data Charting

Two reviewers (DD and AG) jointly developed a data charting form using Microsoft Excel to determine which variables to extract. For each included study, we extracted the following characteristics: study title, publication year, first author name, country of the corresponding author, and journal. We compiled the definitions of digital twins provided by the study authors and summarized their various dimensions (eg, digital replica, high-fidelity representation, and 2-way data exchange). Drawing from NASA's digital twin definition, we assessed whether the developed digital twins were multiscale, integrated multiphysics modeling, and multiple data sources. We incorporated additional descriptors specific to digital twins for patient care, including medical discipline, organs or systems represented, categorization of whether the models constituted anatomical representations (eg, 3D) and models of physiological systems, and the study's objective (simulation, prediction, monitoring, visualization, and generation of synthetic patients). Technical digital twin characteristics extracted included types of patient data used, approach used (mechanistic, data-driven, and hybrid), whether the digital twin included analytics and advanced visualization, constituted a model for simulations versus a simulation itself, capacity and frequency of updates, inclusion of none, 1-way or 2-way data exchange with the patient, and for 2-way exchange, the nature of feedback (recommendation or other type). Finally, we determined the clinical research phase of each digital twin by deriving the clinical research phases developed for artificial intelligence in medicine (phase 0: discovery and invention; phase 1: technical performance and safety; phase 2: efficacy and side effects; phase 3: therapeutic efficacy; and phase 4: safety and effectiveness) [22]. Additional information on each variable is given in Multimedia Appendix 2.

To ensure consistency, 2 reviewers (AG and DD) independently extracted data from the first 10 papers prior to full data extraction. Any disagreements were resolved through discussion. Following this pilot extraction and adaptation of the data charting form, we extracted data from the remaining studies in a masked, duplicate manner. Studies not meeting inclusion criteria were excluded. Any disagreements between reviewers were resolved through discussion.

Unsupervised Classification of CDTs

To perform unsupervised classification of CDT, we analyzed dissimilarities by using the Gower distance metric on a subpart of our data set (the variables included are marked with an asterisk in Multimedia Appendix 2) [23]. The optimal cluster count was identified using the Partitioning Around Medoids method, guided by silhouette width analysis across 2-10 clusters [24]. We then used t-Distributed Stochastic Neighbor Embedding for dimensionality reduction to facilitate visualization in a 2D space. These analyses were performed in R (version 4.2.3) with the packages cluster, Rtsne, and ggplot2. We finally compared the characteristics of the CDTs of each cluster using the package gtsummary.

Results

Characteristics of Included Studies

A total of 7224 citations were identified from the search strategies (Figure 1).

After removing duplicates and screening titles and abstracts, 154 papers were examined in full text, and 68 papers were excluded (see Multimedia Appendix 2 for a full list of excluded full-text reviews with exclusion reasons). Finally, 86 papers [25-110] representing 80 unique CDTs were included in this scoping review. The full results, with individual data for each study, can be viewed in tabular form [20].

The first publications of CDTs appeared in 2019. Since then, publication volume increased yearly from 3 papers in 2019 rising to 23 papers in 2022 and 32 papers in the partial 2023 year (Multimedia Appendix 2). Although incomplete, 2023 totals through August 31 still reflect rising publication volume.

The included publications originated from 22 countries (Multimedia Appendix 2). The United States was the most common country of origin, with 25% (22/80 papers) of included publications. At the continental level, Europe was leading with 49% (43/80) publications, followed by North America with 29% (25/80) and Asia with 18% (16/80). No papers from South America or Africa were found.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

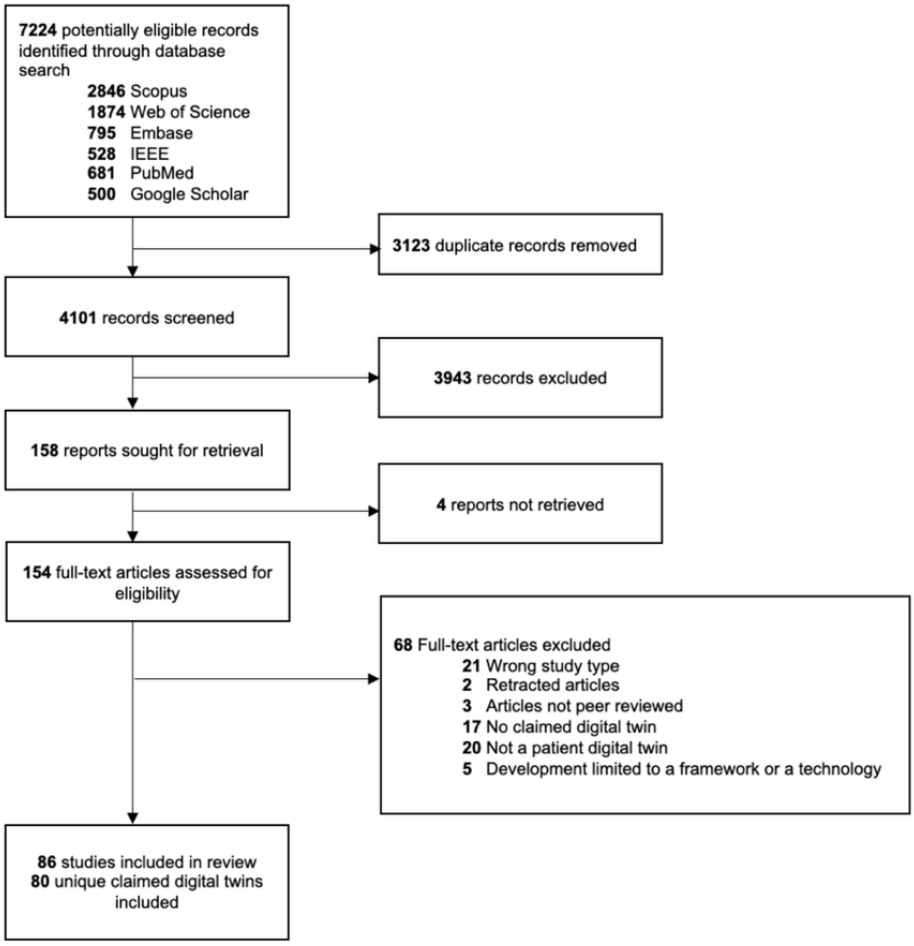

Definitions of Digital Twin Provided by Authors

Definitions of "digital twin" were provided in 63% (55/80 papers) of publications (Multimedia Appendix 2). Among papers with definitions, a digital twin was defined as a digital replica of a real object in 76% (42/55) of publications, with real-time update in 42% (23/55), a patient-specific approach in 24% (13/55), and a 2-way communication between the real and the digital object in 15% (8/55; Figure 2).

XSL•FO
RenderX

Figure 2. Dimensions of digital twins included in definitions provided by authors (gray) and included in the claimed digital twins (black). Created with BioRender [111] which is published under Creative Commons Attribution 4.0 International License [112].

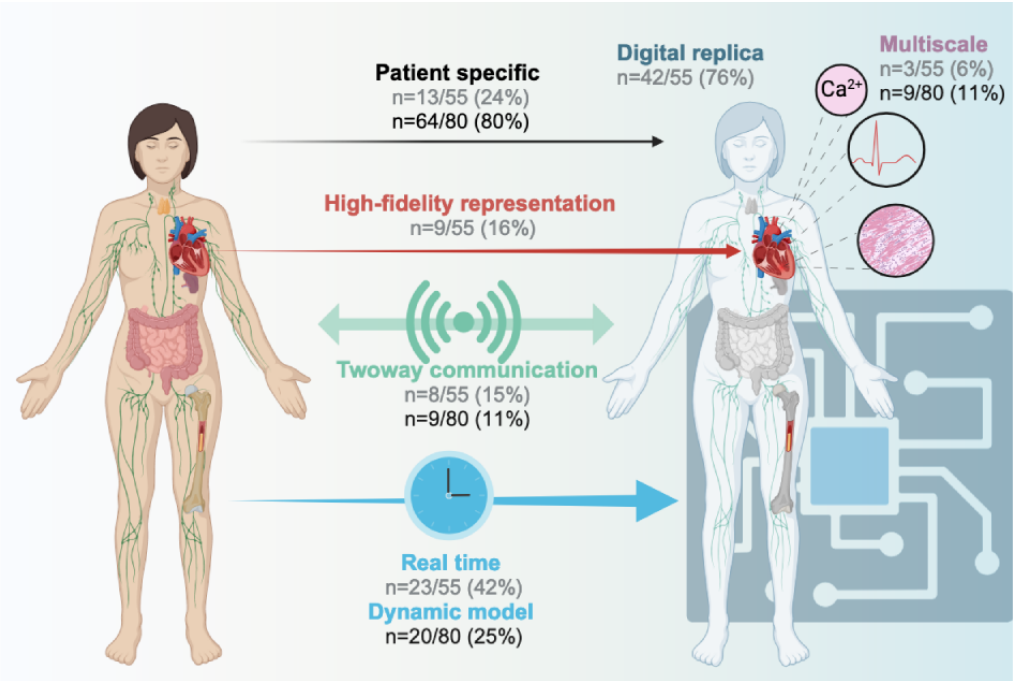

Medical Domains

Of the 80 CDTs identified, 60% (48/80) represented specific organs or anatomical regions, while 14% (11/80) embodied biological systems (eg, immune system). The remaining 26% (21/80) of publications described other types of CDTs (eg, systems for emotion recognition, fetal heart monitoring, etc). Among the 48 publications with CDT representing specific organs, the most widely modeled organs were the heart (15/48 CDTs, 31%), the bones and joints (10/48, 21%), the lung (6/48, 12%), and the arteries (5/48, 10%; Figure 3 and Multimedia Appendix 2).

Approximately 6% (3/48) of CDTs involved multiple organs. Among the 11 CDTs representing biological systems, the endocrine (4/11, 36%) and the immune system (3/11, 27%) were the most widely involved (Multimedia Appendix 2). The most highly represented medical disciplines were cardiology (16/80 CDTs, 20%), oncology (10/80, 13%), and orthopedics (9/80, 11%; Multimedia Appendix 2).

XSL•FO
RenderX

Figure 3. Distribution of organs or systems modeled in the claimed digital twins identified. N=48 claimed digital twin representing specific organs. Created with BioRender [111] which is published under Creative Commons Attribution 4.0 International License [112].

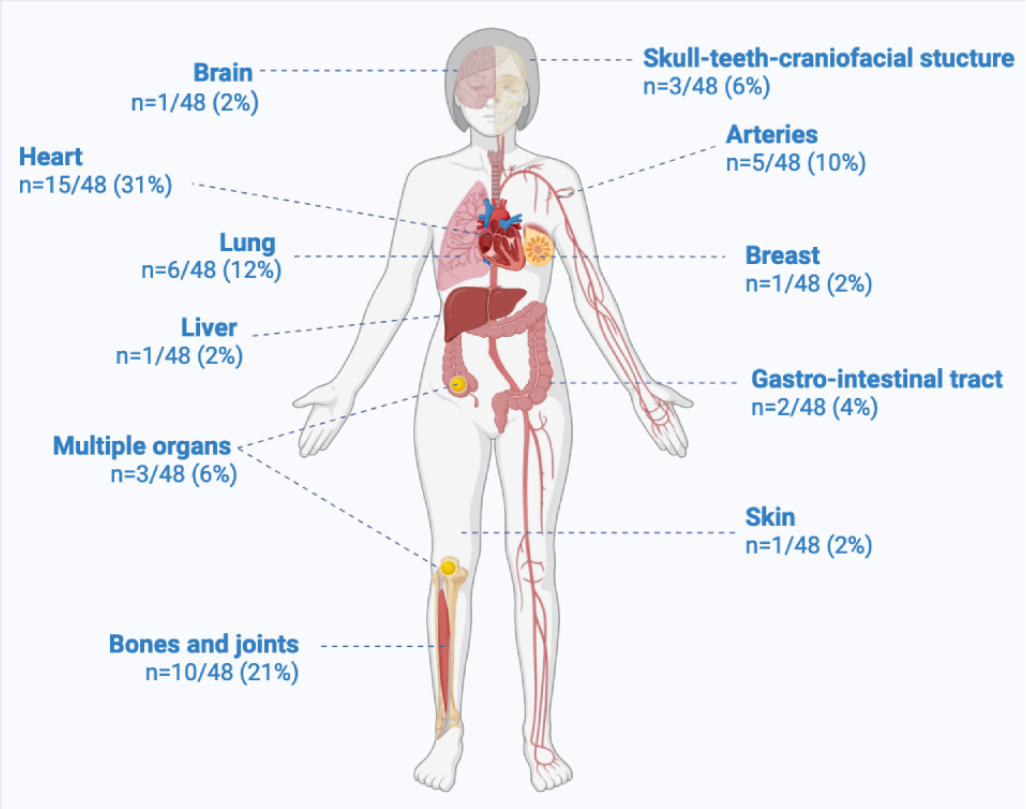

Characteristics of the Claimed "Digital Twins"

The patient data used to develop and run the 80 unique CDTs encompassed data from clinical notes (15/80 CDTs, 19% of publications), laboratory test results (13/80 CDTs, 16%), medical imaging examinations (35/80 CDTs, 44%), wearable device data (12/80 CDTs, 15%), and other data modalities such as electrocardiography, optical tracking, diet data, and intraoperative hemodynamic measurements (in 32/80 CDTs, 40%; Multimedia Appendix 2). Multimodal data incorporation, synthesizing various data types and sources, was present in 49% (39/80) of CDTs. Nearly all CDTs (78/80, 98%) involved data analytics and 36% (29/80) involved some form of advanced visualization. Moreover, 10% (8/80) of CDTs involved multiphysics and 11% (9/80) were multiscale. In terms of implementation, 50% (40/80) of CDTs followed a mechanistic approach, 28% (22/80) followed a data-driven approach, 21% (17/80) combined both approaches, and 1% (1/80) none of these approaches. Most of the CDTs developed (56/80, 70%) corresponded to models for simulation, that is, entities (organ and biological system) that can be used to simulate different states.

In terms of medical approach, 12% (10/80) of CDTs were categorized as anatomical, 28% (22/80) as physiological, and 35% (28/80) combined anatomical and physiological features. The objectives of CDTs were prediction (61/80, 76% CDTs), simulation (52/80, 65% CDTs), monitoring (11/80, 14% CDTs), visualization (10/80, 12% CDTs) and generation of synthetic patient data (8/80, 10% CDTs).

Among the 80% (64/80) of total patient-specific CDTs, 69% (44/80) were static models, whereas 31% (20/80) were designed as dynamic models with regular data inputs or outputs on a daily, hourly, or real-time basis (Multimedia Appendix 2).

Data flow topology characterizes the exchange of information between patient and digital twin. Moreover, 16% (13/80) of CDTs involved no flow from patient to their virtual counterpart, 73% (58/80) used 1-way flow from patient to their virtual counterpart, and 11% (9/80) enabled 2-way data flow between patient and their virtual counterpart (Figure 2). Within the nine 2-way digital twins, automated feedback occurred via recommendations to the patient or the physician in 6 cases and via direct feedback during surgical navigation in 3 cases.

Finally, regarding the clinical research phase of the CDTs, 98% (78/80) were categorized as phase 0 or 1 (preclinical phases) and only 2% (2/80) involved clinical assessments and thus reached phase 2. No CDT corresponded to phase 3 or 4.

Unsupervised Classification of CDTs

Three clusters were identified using the Partitioning Around Medoids method (Figure 4 and Multimedia Appendix 2). The highest average silhouette width was observed for a 3-cluster solution (average silhouette width=0.39), indicating that 3 clusters provided the best separation and cohesion for our data.

Cluster 1 included 43 unique CDTs, which corresponded to patient-specific organs or systems models for simulations and predictions, which we will refer to as "simulation digital twins." Indeed, 98% (42/43) were models to perform simulations, of an organ or system, and these were anatomical and physiological models in 100% (43/43) cases relying on a mechanistic approach

XSL•FO
RenderX

in 95% (41/43) cases. The output of these CDTs was patient-specific in 98% (42/43) cases, and their use was planned for onetime use in 84% (36/43) cases rather than for dynamic use.

Cluster 2 included 22 CDTs and mainly corresponded to prediction models and monitoring systems, which we will refer to as "monitoring digital twins." In this category, the CDT most often did not involve an anatomical or physiological model of the patient (18/22, 82% CDTs). The approach was mainly data-driven (20/22, 91% CDTs). These CDTs were often dynamic (13/22, 59% CDTs) and provided feedback in 23% (5/22) of cases.

Cluster 3 included 15 CDTs and usually corresponded to models not linked to patients but designed for research, which we will refer to as "research-oriented models." Indeed, in this group, only 7% (1/15) of CDTs were patient-specific [67]. Most CDTs were models for simulation (12/15, 80%) and involved a mechanistic approach (9/15, 60% CDTs). Examples of such CDTs are models of the immune response to vaccines [66] but also CDTs that corresponded to the generation of synthetic patients for in-silico trials [107] (5/15, 33% CDTs).

Figure 4. Clusters of claimed digital twins identified. Each point and its associated number correspond to 1 study and its study identification number. Cluster 1: simulation digital twins; cluster 2: monitoring digital twins; cluster 3: research-oriented models.

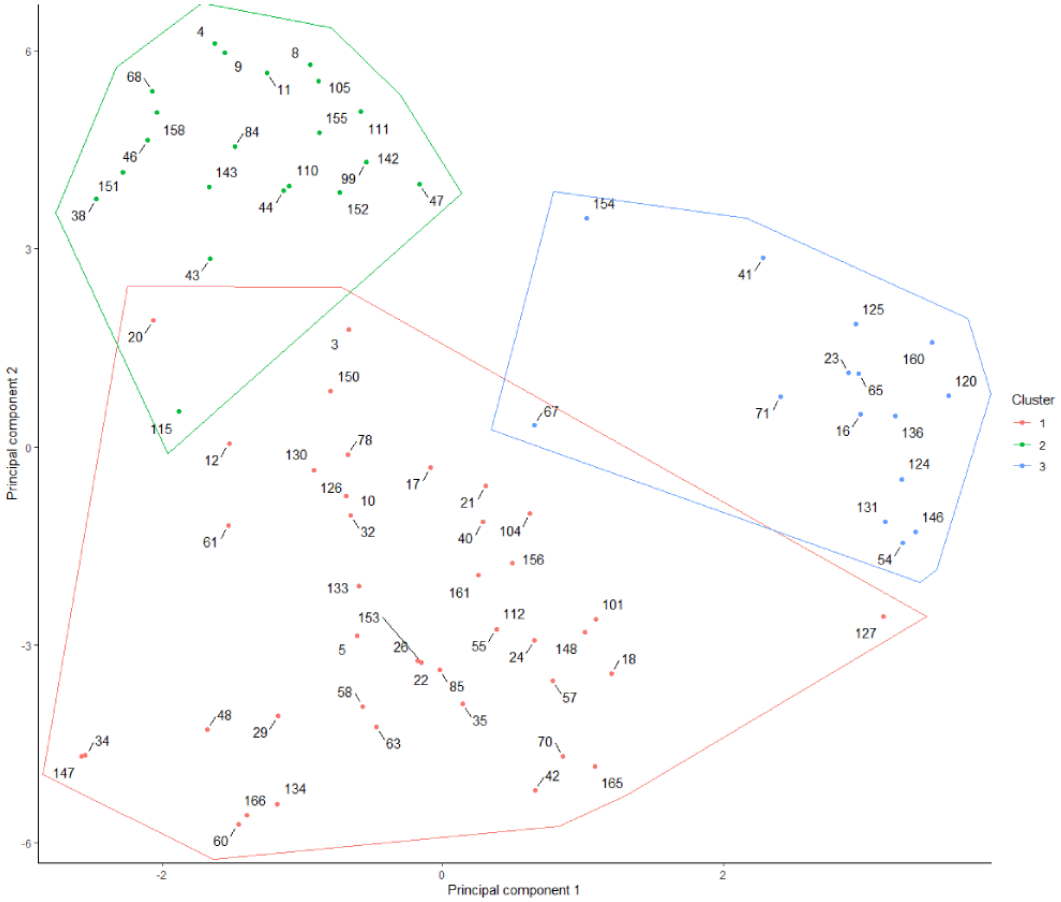

Discussion

Principal Findings

This scoping review highlights the growing interest in the concept of patient digital twins while at the same time revealing the lack of a uniform vision of this same concept among research teams.

The publication of scientific papers on digital twins across all sectors has experienced exponential growth, starting from 1 paper in 2014 to more than 1000 papers per year in 2022, predominantly in the manufacturing industry [113,114]. Studies focusing on the development and evaluation of patient digital twins emerged later, from 2019 onward, but have seen the same growth, particularly in Europe and North America.

Lack of Consensus in Definitions of Digital Twins and Patient Digital Twins

Despite growing interest, there is no consensus definition of the digital twin concept. Depending on the research teams and fields of application, the definitions and characteristics of digital twins differ [114], and this is also reflected in our study, with a diversity of definitions and types of CDTs.

If one were to strictly import the digital twin definition used by Grieve [6] and NASA [7] from the manufacturing industry and aerospace sector to patients, a digital twin of a patient would be an integrated multiscale, probabilistic simulation of a patient that uses the best available models, sensor updates, medical history, and so forth, to mirror the health status of the patient in real time and act on them. But this definition is not currently achievable since no patient digital twin is able to capture the

XSL•FO
RenderX

complexity of the human body in real time. Indeed, creating a patient's digital twin is different from creating a digital twin of an object such as an airplane. An airplane is designed and developed entirely by humans, who have mastery over the composition and physical properties of each part of an airplane, the assembly of all these parts, and the interaction of the whole. In contrast, we still have an imperfect understanding of the functioning of the human body and the interaction of its different organs and systems. It is therefore simpler to model an airplane than the human body with a high degree of fidelity and in a multiscale way. It is also easier to integrate additional sensors and actuators on an airplane's controls than to have patients continuously wear or have implanted sensors or actuators to obtain the 2-way, real-time data exchange between the plane or patients and their digital twin. Moreover, while the digital twin of the airplane relies essentially on the physical properties of the aircraft, the digital twin of a patient will have to go beyond anatomical and physiological models of the patient's various organs and systems to include models of the patient's cognitive and emotional functioning [115]. Finally, concerning the practical use of the term "patient digital twin," they must be able to echo patients' own perceptions of what a digital twin is. These terms are not neologisms, but each refers to concepts already used in everyday language by patients. Because twin refers to a multiple birth, it is likely that the patients picture a digital twin as a realistic avatar of themselves in the same health state. This implies that in the mind of a clinician or patient not well versed in computer science, the first representation of a patient digital twin will be that of a 3D representation of the patient's body and its functioning.

A Practical Definition for Patient Digital Twins

For all these reasons, the concept of a digital twin, which is appropriate in manufacturing, could be considered inappropriate in medicine. However, on the basis of the various papers in this review, we believe that this concept actually corresponds to current advances brought by the multiplication of data sources, data analysis, and the visualization of patients' state of health. It could also be a useful educational tool for the communication with patients when discussing models and predictions made from their data. Since it is currently not feasible to bring together all the characteristics of the digital twin in manufacturing, we propose the following definition of a patient digital twin: "A patient digital twin is a viewable digital replica of a patient, organ, or biological system that contains multidimensional, patient-specific information and informs decisions." This definition is based on the characteristics of the digital twins identified in this review (digital replica, multidimensional, and patient-specific) and aligns with the broader definition of the National Academies of Sciences, Engineering, and Medicine, which defines a digital twin as "a set of virtual information constructs that mimics the structure, context, and behaviour of a natural, engineered, or social system (or system-of-systems), is dynamically updated with data from its physical twin, has a predictive capability, and informs decisions that realize value" [116].

The advantage of our definition is that it can exclude what is not a patient digital twin: generic models of cells, tissues, organs, or biological systems not linked to a patient but used to study disease progression or drug development [33,58,84], which corresponded to our cluster 3 (research-oriented models); pure cyber-physical systems, that is, systems such as implantable cardioverter-defibrillators, or artificial pancreas, which do not use a representation of the patient and therefore not a "viewable" digital replica of the patient; digital patient data created from patient databases for in-silico trials [51,103,107]; often using generative adversarial networks, which we propose to call "synthetic patients" instead; data sets from another patient, similar to those of the index patient [81]; machine learning–based classifiers, trained on a population to predict a diagnosis [73]; and patient models built from a single data source, such as demographic characteristics or imaging [35,51,53].

Simulation and Monitoring Patient Digital Twins

This definition of patient digital twins also encompasses the 2 major trends revealed by this study. On one hand, digital twins offering a high degree of fidelity, combining advanced anatomical and physiological models, based on mechanistic approaches or combining mechanistic and data-driven approaches [47,70,109]. However, their limitation is their common reliance on hospital-acquired data such as imaging data (computed tomography and magnetic resonance imaging) or intraoperative hemodynamic data. They are therefore generally restricted to onetime purposes and not in a dynamic form. On the other hand, digital twins corresponding to real-time representations of patients through home-based data collection via wearable, using machine learning techniques for analysis and alert detection [29,31]. This type of digital twin is dynamic, can be 2-way by giving recommendations to patients or by modifying a biological parameter of the patient via medical devices, but often offers little or no information on the patient's anatomy or physiology. For it to be differentiable from a cyber-physical system or a telemonitoring system, it must integrate different data sources, data analysis systems, and a visualization of the patient organ, system, or body. We thus propose 2 major categories of patient digital twins based on our clustering analysis (Figure 5).

Simulation patient digital twins (derived from cluster 1): personalized, static, viewable digital replicas of patients' anatomy and physiology based on computational modeling to run simulations predicting outcomes in hypothetical scenarios or evaluating therapeutic approaches. These digital twins are mostly used for onetime assessments rather than continuous monitoring.

Monitoring patient digital twins (derived from cluster 2): personalized, dynamic, viewable digital replicas of patients leveraging aggregated health data and analytics to enable continuous predictions of risks and outcomes over time and provide feedback for optimizing care. These digital twins are mostly focused on continuous tracking rather than detailed mechanistic simulation.

An emblematic example of a simulation patient digital twin is a personalized cardiac model that integrates a patient's anatomical and physiological data to create a 3D virtual representation of his or her heart [61]. Computational simulations are then performed on this digital twin to predict

XSL•FO
RenderX

outcomes or optimize treatments, such as evaluating different ablation strategies for atrial fibrillation. These digital twins enable onetime, patient-specific assessments through detailed mechanistic modeling.

An emblematic example of a monitoring patient digital twin would be a personalized representation of the bronchi of a child with asthma based on real-time data from connected objects such as smart inhalers, spirometers, and portable air quality monitors [117]. This may allow prediction of the risk of an asthma attack and enable children to visualize the targets of the asthma treatments they are taking, namely, bronchial inflammation and bronchoconstriction.

Figure 5. Proposed definitions and characteristics of patient digital twins. PDT: patient digital twin. Created with BioRender [111] which is published under Creative Commons Attribution 4.0 International License [112].

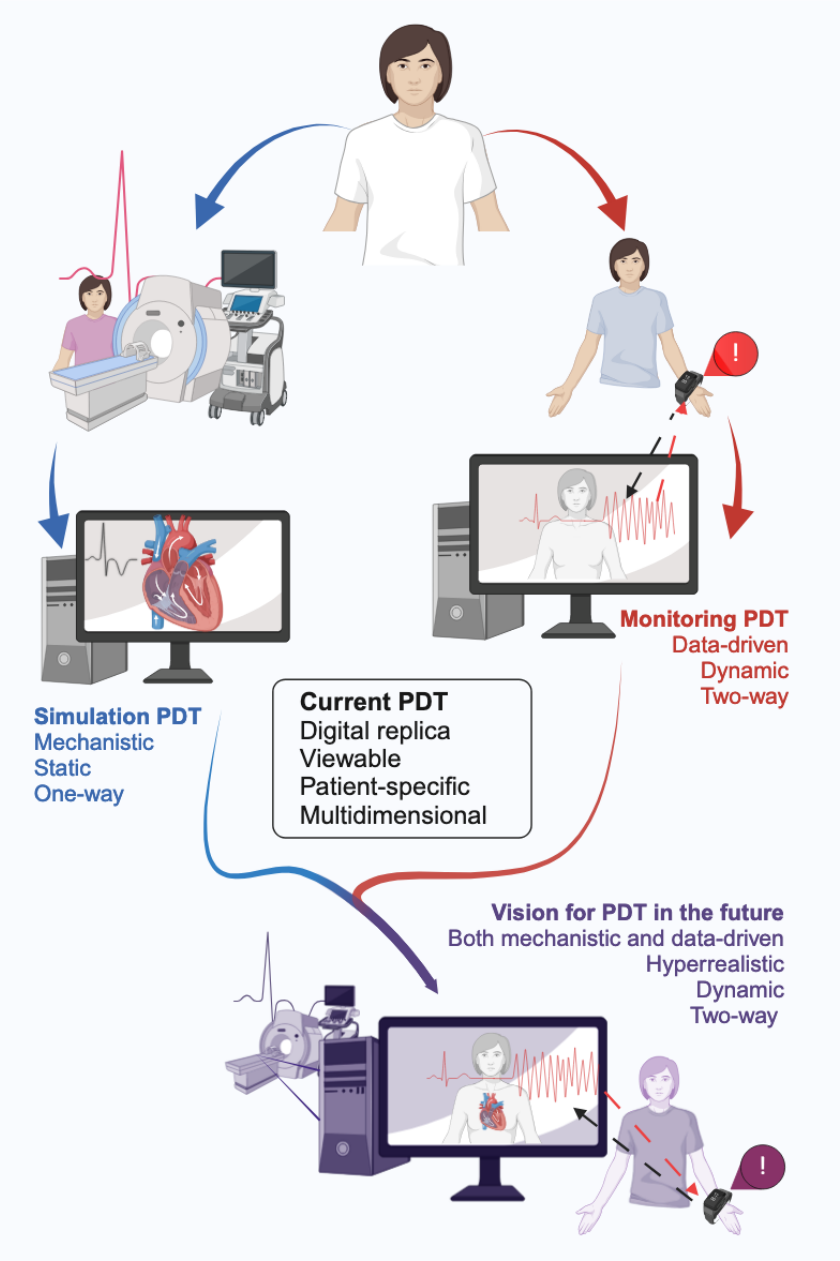

Future Directions

As technical progress is made, it is likely that these 2 types of patient digital twins will converge, combining both mechanistic models and increasingly efficient continuous data collection and data-driven approaches [118]. However, currently, only 14 CDTs from this review would meet the definition of patient digital twins combining multidimensional, patient-specific data, data analytics, and advanced visualization, all being simulation patient digital twins [27,38,41,47,52,61,62,65,70,74,92,102,109,110]. Digital twins for surgical navigation systems are the most advanced, with 3 teams presenting patient digital twins combining both a mechanistic approach and a dynamic real-time adaptation [62,92,110].

The main benefit expected from these patient digital twins is to offer increasingly personalized medicine, taking into account all the available data and patient-specific simulations and predictions. This would make it possible, for example, to choose the most appropriate drug treatment based on the patient's medical history, allergies, comorbidities, and genetic profile

XSL•FO
RenderX

[26]; the most appropriate stent for the configuration of the aortic dilatation [27]; and the least risky surgical approach thanks to preoperative simulations. But the benefits could also be for patient education and engagement: patients would interact with their digital replica to better understand their own body, health conditions, and influence of behaviors, and collaborate more with providers.

Strengths and Limitations

The strength of this review is to be the first to have carried out a systematic analysis of the literature concerning patient digital twins and to have been able to identify the major trends. We also acknowledge several limitations. First, in the absence of a consensual definition of the digital twin concept, we chose to include all papers in which the authors claimed that they had developed a digital twin or part of a digital twin. This approach undoubtedly led us to exclude some papers that embody the concept of a digital twin but did not use that specific term, such as works from the Physiome Project [119] or the Virtual Physiological Human initiative [120]. Conversely, it may have resulted in the inclusion of some papers that, in fact, did not develop a digital twin but only a simple prediction model. To mitigate this limitation, we proposed our own definition and typology of patient digital twins, derived from characteristics found in the reviewed health care literature and aligned with well-established definitions from other fields, such as NASA's definition. This approach ensures that our definition is grounded in current health care applications and fundamental digital twin principles, enhancing its robustness and adaptability.

Second, we realized that some characteristics of digital twins were not appropriate for consistent assessment, either because they were too industry-specific, such as the "multiphysic" characteristic, or because they did not allow for consistent categorization. For example, we often struggled to determine whether a presented digital twin was at phase 0 or phase 1 of development and chose to merge these 2 categories into a single "preclinical" category. This may have affected the granularity of our analysis. Future research could develop more precise and universally applicable criteria for classifying digital twin characteristics.

Third, we did not include papers written in languages other than English, French, or German. This language limitation may have led to underrepresentation of work from regions where other languages are predominant, potentially biasing our findings toward research conducted in these languages.

Fourth, we did not include conference papers. This exclusion might have omitted recent or preliminary findings that are often first presented at conferences, possibly affecting the comprehensiveness of our review.

Finally, our cluster analysis has limitations due to potential sensitivity to variable selection and the choice of clustering methods, which may influence the results. In addition, the relatively small sample size of 80 CDTs could affect the stability and generalizability of the clusters identified. Although we carefully selected variables and methods, these factors may still influence the findings. Therefore, our analysis provides valuable insights, but these should be interpreted with caution. Future research with larger sample sizes and exploration of different clustering techniques could enhance the reliability and generalizability of such analyses.

Conclusions

We propose that a patient digital twin be defined as "a viewable digital replica of a patient, organ, or biological system that contains multidimensional, patient-specific information and informs decision." We currently identify 2 categories, simulation patient digital twins and monitoring patient digital twins. In the future, we envisage a fusion of these 2 types of digital twins that will combine a high degree of fidelity based on anatomical and physiological models with real-time updating and feedback to the patient.

Acknowledgments

The authors wish to acknowledge the use of Claude 2.1, an artificial intelligence assistant created by Anthropic, for the translation of selected passages from French to English and the rewriting of these passages into academic English.

Data Availability

The data related to the studies included in this manuscript can be found at on Open Science Framework [20].

Authors' Contributions

DD led all aspects of the Review: conceptualization, data curation (title and abstract screening, full-text screening, and data extraction), formal analysis, project administration, supervision, validation, and writing (original draft, review, and editing); AG was involved in conceptualization, data curation (title and abstract screening, full-text screening, and data extraction), formal analysis, validation, and writing (review, and editing). All authors have had full access to the data and accept responsibility to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 154 KB - jmir_v26i1e58504_app1.pdf]

Multimedia Appendix 2
Additional methods, tables and figures.
[PDF File (Adobe PDF File), 938 KB - jmir_v26i1e58504_app2.pdf]

References

1. Drummond D. Between competence and warmth: the remaining place of the physician in the era of artificial intelligence. NPJ Digit Med 2021;4(1):85 [FREE Full text] [doi: 10.1038/s41746-021-00457-w] [Medline: 33990682]
2. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
3. Seyhan AA, Carini C. Are innovation and new technologies in precision medicine paving a new era in patients centric care? J Transl Med 2019;17(1):114 [FREE Full text] [doi: 10.1186/s12967-019-1864-9] [Medline: 30953518]
4. Xu LD, Xu EL, Li L. Industry 4.0: state of the art and future trends. Int J Prod Res 2018;56(8):2941-2962. [doi: 10.1080/00207543.2018.1444806]
5. Schwab K. The Fourth Industrial Revolution. Germany: Crown; 2017.
6. Grieves M. Digital Twin: Manufacturing Excellence through Virtual Factory Replication. Boston, MA: Digital Twin Institute; 2015.
7. Shafto M, Conroy M, Doyle R, Glaessgen E, Kemp C, LeMoigne J, et al. Modeling, simulation, information technology and processing roadmap. NASA 2010:TA11-7.
8. Singh M, Srivastava R, Fuenmayor E, Kuts V, Qiao Y, Murray N, et al. Applications of digital twin across industries: a review. Appl Sci 2022;12(11):5727. [doi: 10.3390/app12115727]
9. Gonsard A, AbouTaam R, Prévost B, Roy C, Hadchouel A, Nathan N, et al. Children's views on artificial intelligence and digital twins for the daily management of their asthma: a mixed-method study. Eur J Pediatr 2023;182(2):877-888 [FREE Full text] [doi: 10.1007/s00431-022-04754-8] [Medline: 36512148]
10. Abdoul C, Cros P, Coutier L, Hadchouel A, Neuraz A, Burgun A, et al. Parents' views on artificial intelligence for the daily management of childhood asthma: a survey. J Allergy Clin Immunol Pract 2021;9(4):1728-1730.e3. [doi: 10.1016/j.jaip.2020.11.048] [Medline: 33290917]
11. Casola L, editor. Opportunities and Challenges for Digital Twins in Biomedical Research: Proceedings of a Workshop-in Brief. Washington, DC: National Academies Press; 2023.
12. Coorey G, Figtree GA, Fletcher DF, Snelson VJ, Vernon ST, Winlaw D, et al. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. NPJ Digit Med 2022;5(1):126 [FREE Full text] [doi: 10.1038/s41746-022-00640-7] [Medline: 36028526]
13. Xames MD, Topcu TG. A Systematic literature review of digital twin research for healthcare systems: research trends, gaps, and realization challenges. IEEE Access 2024;12:4099-4126. [doi: 10.1109/access.2023.3349379]
14. Sun T, He X, Li Z. Digital twin in healthcare: recent updates and challenges. Digit Health 2023;9:20552076221149651 [FREE Full text] [doi: 10.1177/20552076221149651] [Medline: 36636729]
15. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
16. Peters MDJ, Godfrey C, McInerney P, Khalil H, Larsen P, Marnie C, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
17. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Aromataris E, Munn Z. JBI Manual for Evidence Synthesis. 2024. URL: https://jbi-global-wiki.refined.site/space/MANUAL [accessed 2024-02-08]
20. Scoping review of patient digital twins. OSF. 2024. URL: https://osf.io/7twn8/ [accessed 2024-05-30]
21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
22. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open 2020;3(3):326-331 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa033] [Medline: 33215066]
23. Gower J. Properties of euclidean and non-euclidean distance matrices. Linear Algebra Appl 1985;67:81-97. [doi: 10.1016/0024-3795(85)90187-9]

XSL•FO
RenderX

24. van der Laan M, Pollard K, Bryan J. A new partitioning around medoids algorithm. UC Berkeley Division of Biostatistics Working Paper. 2002. URL: https://biostats.bepress.com/ucbbiostat/paper105 [accessed 2024-10-11]
25. Chakshu NK, Carson J, Sazonov I, Nithiarasu P. A semi-active human digital twin model for detecting severity of carotid stenoses from head vibration-a coupled computational mechanics and computer vision method. Int J Numer Method Biomed Eng 2019;35(5):e3180 [FREE Full text] [doi: 10.1002/cnm.3180] [Medline: 30648344]
26. Liu Y, Zhang L, Yang Y, Zhou L, Ren L, Wang F, et al. A novel cloud-based framework for the elderly healthcare services using digital twin. IEEE Access 2019;7:49088-49101. [doi: 10.1109/access.2019.2909828]
27. Hemmler A, Reeps C, Lutz B, Gee MW. Der digitale zwilling in der endovaskulären versorgung. Gefässchirurgie 2019;24(7):548-556. [doi: 10.1007/s00772-019-00569-4]
28. Hirschvogel M, Jagschies L, Maier A, Wildhirt SM, Gee MW. An in silico twin for epicardial augmentation of the failing heart. Int J Numer Method Biomed Eng 2019;35(10):e3233. [doi: 10.1002/cnm.3233] [Medline: 31267697]
29. Barricelli BR, Casiraghi E, Gliozzo J, Petrini A, Valtolina S. Human digital twin for fitness management. IEEE Access 2020;8:26637-26664. [doi: 10.1109/access.2020.2971576]
30. Lal A, Li G, Cubro E, Chalmers S, Li H, Herasevich V, et al. Development and verification of a digital twin patient model to predict specific treatment response during the first 24 hours of sepsis. Crit Care Explor 2020;2(11):e0249 [FREE Full text] [doi: 10.1097/CCE.0000000000000249] [Medline: 33225302]
31. Shamanna P, Saboo B, Damodharan S, Mohammed J, Mohamed M, Poon T, et al. Reducing HbA1c in type 2 diabetes using digital twin technology-enabled precision nutrition: a retrospective analysis. Diabetes Ther 2020;11(11):2703-2714 [FREE Full text] [doi: 10.1007/s13300-020-00931-w] [Medline: 32975712]
32. Goodwin GC, Seron MM, Medioli AM, Smith T, King BR, Smart CE. A systematic stochastic design strategy achieving an optimal tradeoff between peak BGL and probability of hypoglycaemic events for individuals having type 1 diabetes mellitus. Biomed Signal Process Control 2020;57:101813. [doi: 10.1016/j.bspc.2019.101813]
33. Subramanian K. Digital twin for drug discovery and development—the virtual liver. J Indian Inst Sci 2020;100(4):653-662. [doi: 10.1007/s41745-020-00185-2]
34. Sinisi S, Alimguzhin V, Mancini T, Tronci E, Mari F, Leeners B. Optimal personalised treatment computation through in silico clinical trials on patient digital twins. Fundamenta Informaticae IOS Press 2020;174(3-4):283-310. [doi: 10.3233/fi-2020-1943]
35. Fu W, Sharma S, Abadi E, Iliopoulos A, Wang Q, Lo JY, et al. iPhantom: a framework for automated creation of individualized computational phantoms and its application to CT organ dosimetry. IEEE J Biomed Health Inform 2021;25(8):3061-3072 [FREE Full text] [doi: 10.1109/JBHI.2021.3063080] [Medline: 33651703]
36. Béthencourt L, Dabachine W, Dejouy V, Lalmiche Z, Neuberger K, Ibnouhsein I, et al. Guiding measurement protocols of connected medical devices using digital twins: a statistical methodology applied to detecting and monitoring lymphedema. IEEE Access 2021;9:39444-39465. [doi: 10.1109/access.2021.3063786]
37. Zhou C, Chase JG, Knopp J, Sun Q, Tawhai M, Möller K, et al. Virtual patients for mechanical ventilation in the intensive care unit. Comput Methods Programs Biomed 2021;199:105912. [doi: 10.1016/j.cmpb.2020.105912] [Medline: 33360683]
38. Camps J, Lawson B, Drovandi C, Minchole A, Wang ZJ, Grau V, et al. Inference of ventricular activation properties from non-invasive electrocardiography. Med Image Anal 2021;73:102143 [FREE Full text] [doi: 10.1016/j.media.2021.102143] [Medline: 34271532]
39. Barbiero P, Viñas Torné R, Lió P. Graph representation forecasting of patient's medical conditions: toward a digital twin. Front Genet 2021;12:652907 [FREE Full text] [doi: 10.3389/fgene.2021.652907] [Medline: 34603366]
40. Aubert K, Germaneau A, Rochette M, Ye W, Severyns M, Billot M, et al. Development of digital twins to optimize trauma surgery and postoperative management. a case study focusing on tibial plateau fracture. Front Bioeng Biotechnol 2021;9:722275 [FREE Full text] [doi: 10.3389/fbioe.2021.722275] [Medline: 34692655]
41. Förster KM, Roth CJ, Hilgendorff A, Ertl-Wagner B, Flemmer AW, Wall WA. In silico numerical simulation of ventilator settings during high-frequency ventilation in preterm infants. Pediatr Pulmonol 2021;56(12):3839-3846. [doi: 10.1002/ppul.25626] [Medline: 34432956]
42. Shamanna P, Joshi S, Shah L, Dharmalingam M, Saboo B, Mohammed J, et al. Type 2 diabetes reversal with digital twin technology-enabled precision nutrition and staging of reversal: a retrospective cohort study. Clin Diabetes Endocrinol 2021;7(1):21 [FREE Full text] [doi: 10.1186/s40842-021-00134-7] [Medline: 34776010]
43. Hernigou P, Olejnik R, Safar A, Martinov S, Hernigou J, Ferre B. Digital twins, artificial intelligence, and machine learning technology to identify a real personalized motion axis of the tibiotalar joint for robotics in total ankle arthroplasty. Int Orthop 2021;45(9):2209-2217. [doi: 10.1007/s00264-021-05175-2] [Medline: 34351462]
44. Shamanna P, Dharmalingam M, Sahay R, Mohammed J, Mohamed M, Poon T, et al. Retrospective study of glycemic variability, BMI, and blood pressure in diabetes patients in the digital twin precision treatment program. Sci Rep 2021;11(1):14892 [FREE Full text] [doi: 10.1038/s41598-021-94339-6] [Medline: 34290310]
45. Golse N, Joly F, Combari P, Lewin M, Nicolas Q, Audebert C, et al. Predicting the risk of post-hepatectomy portal hypertension using a digital twin: a clinical proof of concept. J Hepatol 2021;74(3):661-669. [doi: 10.1016/j.jhep.2020.10.036] [Medline: 33212089]

XSL•FO
RenderX

46. Cho SW, Byun SH, Yi S, Jang WS, Kim JC, Park IY, et al. Sagittal relationship between the maxillary central incisors and the forehead in digital twins of Korean adult females. J Pers Med 2021;11(3):203 [FREE Full text] [doi: 10.3390/jpm11030203] [Medline: 33805617]
47. Gillette K, Gsell MAF, Prassl AJ, Karabelas E, Reiter U, Reiter G, et al. A framework for the generation of digital twins of cardiac electrophysiology from clinical 12-leads ECGs. Med Image Anal 2021;71:102080 [FREE Full text] [doi: 10.1016/j.media.2021.102080] [Medline: 33975097]
48. Gillette K, Gsell MAF, Bouyssier J, Prassl AJ, Neic A, Vigmond EJ, et al. Automated framework for the inclusion of a his-purkinje system in cardiac digital twins of ventricular electrophysiology. Ann Biomed Eng 2021;49(12):3143-3153 [FREE Full text] [doi: 10.1007/s10439-021-02825-9] [Medline: 34431016]
49. Conte Alcaraz J, Moghaddamnia S, Peissig J. Efficiency of deep neural networks for joint angle modeling in digital gait assessment. EURASIP J Adv Signal Process 2021;2021(1):67-74. [doi: 10.1186/s13634-020-00715-1]
50. van Osta N, Kirkels FP, van Loon T, Koopsen T, Lyon A, Meiburg R, et al. Uncertainty quantification of regional cardiac tissue properties in arrhythmogenic cardiomyopathy using adaptive multiple importance sampling. Front Physiol 2021;12:738926 [FREE Full text] [doi: 10.3389/fphys.2021.738926] [Medline: 34658923]
51. Gonzalez-Abril L, Angulo C, Ortega JA, Lopez-Guerra JL. Generative adversarial networks for anonymized healthcare of lung cancer patients. Electronics 2021;10(18):2220. [doi: 10.3390/electronics10182220]
52. Gerach T, Schuler S, Fröhlich J, Lindner L, Kovacheva E, Moss R, et al. Electro-mechanical whole-heart digital twins: a fully coupled multi-physics approach. Mathematics 2021;9(11):1247. [doi: 10.3390/math9111247]
53. Geissler F, Heiß R, Kopp M, Wiesmüller M, Saake M, Wuest W, et al. Personalized computed tomography -automated estimation of height and weight of a simulated digital twin using a 3D camera and artificial intelligence. Rofo 2021;193(4):437-445 [FREE Full text] [doi: 10.1055/a-1253-8558] [Medline: 33142337]
54. Allen A, Siefkas A, Pellegrini E, Burdick H, Barnes G, Calvert J, et al. A digital twins machine learning model for forecasting disease progression in stroke patients. Appl Sci 2021;11(12):5576. [doi: 10.3390/app11125576]
55. Elayan H, Aloqaily M, Guizani M. Digital twin for intelligent context-aware iot healthcare systems. IEEE Internet Things J 2021;8(23):16749-16757. [doi: 10.1109/jiot.2021.3051158]
56. Tardini E, Zhang X, Canahuate G, Wentzel A, Mohamed ASR, Van Dijk L, et al. Optimal treatment selection in sequential systemic and locoregional therapy of oropharyngeal squamous carcinomas: deep Q-learning with a patient-physician digital twin dyad. J Med Internet Res 2022;24(4):e29455 [FREE Full text] [doi: 10.2196/29455] [Medline: 35442211]
57. Tai Y, Zhang L, Li Q, Zhu C, Chang V, Rodrigues JJPC, et al. Digital-twin-enabled IoMT system for surgical simulation using rAC-GAN. IEEE Internet Things J 2022;9(21):20918-20931. [doi: 10.1109/jiot.2022.3176300]
58. Schütt M, O'Farrell C, Stamatopoulos K, Hoad CL, Marciani L, Sulaiman S, et al. Simulating the hydrodynamic conditions of the human ascending colon: a digital twin of the dynamic colon model. Pharmaceutics 2022;14(1):184 [FREE Full text] [doi: 10.3390/pharmaceutics14010184] [Medline: 35057077]
59. Hernigou P, Safar A, Hernigou J, Ferre B. Subtalar axis determined by combining digital twins and artificial intelligence: influence of the orientation of this axis for hindfoot compensation of varus and valgus knees. Int Orthop 2022;46(5):999-1007. [doi: 10.1007/s00264-022-05311-6] [Medline: 35138455]
60. Ahmadian H, Mageswaran P, Walter BA, Blakaj DM, Bourekas EC, Mendel E, et al. Toward an artificial intelligence-assisted framework for reconstructing the digital twin of vertebra and predicting its fracture response. Int J Numer Method Biomed Eng 2022;38(6):e3601 [FREE Full text] [doi: 10.1002/cnm.3601] [Medline: 35403831]
61. Azzolin L, Eichenlaub M, Nagel C, Nairn D, Sanchez J, Unger L, et al. Personalized ablation vs. conventional ablation strategies to terminate atrial fibrillation and prevent recurrence. Europace 2023;25(1):211-222 [FREE Full text] [doi: 10.1093/europace/euac116] [Medline: 35943361]
62. Qin Y, Ma M, Shen L, Wang H, Han J. Virtual and real bidirectional driving system for the synchronization of manipulations in robotic joint surgeries. MDPI 2022;10(7):530. [doi: 10.3390/machines10070530]
63. Venkatapurapu SP, Iwakiri R, Udagawa E, Patidar N, Qi Z, Takayama R, et al. A computational platform integrating a mechanistic model of Crohn's disease for predicting temporal progression of mucosal damage and healing. Adv Ther 2022;39(7):3225-3247 [FREE Full text] [doi: 10.1007/s12325-022-02144-y] [Medline: 35581423]
64. Ahmadian H, Mageswaran P, Walter BA, Blakaj DM, Bourekas EC, Mendel E, et al. A digital twin for simulating the vertebroplasty procedure and its impact on mechanical stability of vertebra in cancer patients. Int J Numer Method Biomed Eng 2022;38(6):e3600 [FREE Full text] [doi: 10.1002/cnm.3600] [Medline: 35347880]
65. Poletti G, Antonini L, Mandelli L, Tsompou P, Karanasiou G, Papafaklis M, et al. Towards a digital twin of coronary stenting: a suitable and validated image-based approach for mimicking patient-specific coronary arteries. Electronics MDPI 2022;11(3):502. [doi: 10.3390/electronics11030502]
66. Zohdi TI. Machine-learning and digital-twins for rapid evaluation and design of injected vaccine immune-system responses. Comput Methods Appl Mech Eng 2022;401:115315. [doi: 10.1016/j.cma.2022.115315]
67. Bahrami F, Rossi RM, Defraeye T. Predicting transdermal fentanyl delivery using physics-based simulations for tailored therapy based on the age. Drug Deliv 2022;29(1):950-969 [FREE Full text] [doi: 10.1080/10717544.2022.2050846] [Medline: 35319323]

XSL•FO
RenderX

68. Sharotry A, Jimenez JA, Mediavilla FAM, Wierschem D, Koldenhoven RM, Valles D. Manufacturing operator ergonomics: a conceptual digital twin approach to detect biomechanical fatigue. IEEE Access 2022;10:12774-12791. [doi: 10.1109/access.2022.3145984]
69. Zhang J, Tai Y. Secure medical digital twin via human-centric interaction and cyber vulnerability resilience. Connect Sci 2021;34(1):895-910. [doi: 10.1080/09540091.2021.2013443]
70. Jung A, Gsell MAF, Augustin CM, Plank G. An integrated workflow for building digital twins of cardiac electromechanics—a multi-fidelity approach for personalising active mechanics. Mathematics (Basel) 2022;10(5):823 [FREE Full text] [doi: 10.3390/math10050823] [Medline: 35295404]
71. Li X, Lee EJ, Lilja S, Loscalzo J, Schäfer S, Smelik M, et al. A dynamic single cell-based framework for digital twins to prioritize disease genes and drug targets. Genome Med 2022;14(1):48 [FREE Full text] [doi: 10.1186/s13073-022-01048-4] [Medline: 35513850]
72. Silfvergren O, Simonsson C, Ekstedt M, Lundberg P, Gennemark P, Cedersund G. Digital twin predicting diet response before and after long-term fasting. PLoS Comput Biol 2022;18(9):e1010469 [FREE Full text] [doi: 10.1371/journal.pcbi.1010469] [Medline: 36094958]
73. Chakshu NK, Nithiarasu P. An AI based digital-twin for prioritising pneumonia patient treatment. Proc Inst Mech Eng H 2022;236(11):1662-1674 [FREE Full text] [doi: 10.1177/09544119221123431] [Medline: 36121054]
74. Wu C, Jarrett AM, Zhou Z, Elshafeey N, Adrada BE, Candelaria RP, et al. MRI-based digital models forecast patient-specific treatment responses to neoadjuvant chemotherapy in triple-negative breast cancer. Cancer Res 2022;82(18):3394-3404 [FREE Full text] [doi: 10.1158/0008-5472.CAN-22-1329] [Medline: 35914239]
75. Kim JK, Lee SJ, Hong SH, Choi IY. Machine-learning-based digital twin system for predicting the progression of prostate cancer. Appl Sci MDPI 2022;12(16):8156. [doi: 10.3390/app12168156]
76. Doste R, Lozano M, Jimenez-Perez G, Mont L, Berruezo A, Penela D, et al. Training machine learning models with synthetic data improves the prediction of ventricular origin in outflow tract ventricular arrhythmias. Front Physiol 2022;13:909372 [FREE Full text] [doi: 10.3389/fphys.2022.909372] [Medline: 36035489]
77. Ang CYS, Lee JWW, Chiew YS, Wang X, Tan CP, Cove ME, et al. Virtual patient framework for the testing of mechanical ventilation airway pressure and flow settings protocol. Comput Methods Programs Biomed 2022;226:107146 [FREE Full text] [doi: 10.1016/j.cmpb.2022.107146] [Medline: 36191352]
78. Al-Zyoud I, Laamarti F, Ma X, Tobón D, El Saddik A. Towards a machine learning-based digital twin for non-invasive human bio-signal fusion. Sensors (Basel) 2022;22(24):9747 [FREE Full text] [doi: 10.3390/s22249747] [Medline: 36560115]
79. Bahrami F, Rossi RM, De Nys K, Defraeye T. An individualized digital twin of a patient for transdermal fentanyl therapy for chronic pain management. Drug Deliv Transl Res 2023;13(9):2272-2285 [FREE Full text] [doi: 10.1007/s13346-023-01305-y] [Medline: 36897525]
80. Cen S, Gebregziabher M, Moazami S, Azevedo CJ, Pelletier D. Toward precision medicine using a "digital twin" approach: modeling the onset of disease-specific brain atrophy in individuals with multiple sclerosis. Sci Rep 2023;13(1):16279 [FREE Full text] [doi: 10.1038/s41598-023-43618-5] [Medline: 37770560]
81. Xu Q, Kowalski J. myCMIE: my cancer molecular information exchange. iScience 2023;26(8):107324 [FREE Full text] [doi: 10.1016/j.isci.2023.107324] [Medline: 37575188]
82. O'Hara RP, Meijborg VMF, Jelvehgaran P, van der Waal J, Boink GJJ, Trayanova NA, et al. Site-specific prolongation of repolarization prevents postmyocardial infarction tachycardia. Heart Rhythm O2 2023;4(7):466-468 [FREE Full text] [doi: 10.1016/j.hroo.2023.06.003] [Medline: 37520014]
83. Thamotharan P, Srinivasan S, Kesavadev J, Krishnan G, Mohan V, Seshadhri S, et al. Human digital twin for personalized elderly type 2 diabetes management. J Clin Med 2023;12(6):2094 [FREE Full text] [doi: 10.3390/jcm12062094] [Medline: 36983097]
84. Maksymenko K, Clarke AK, Mendez Guerra I, Deslauriers-Gauthier S, Farina D. A myoelectric digital twin for fast and realistic modelling in deep learning. Nat Commun 2023;14(1):1600 [FREE Full text] [doi: 10.1038/s41467-023-37238-w] [Medline: 36959193]
85. Yang PC, Rose A, DeMarco KR, Dawson JRD, Han Y, Jeng MT, et al. A multiscale predictive digital twin for neurocardiac modulation. J Physiol 2023;601(17):3789-3812 [FREE Full text] [doi: 10.1113/JP284391] [Medline: 37528537]
86. Chasseloup E, Hooker AC, Karlsson MO. Generation and application of avatars in pharmacometric modelling. J Pharmacokinet Pharmacodyn 2023;50(5):411-423 [FREE Full text] [doi: 10.1007/s10928-023-09873-9] [Medline: 37488327]
87. Reimann AM, Schalk E, Jost F, Mougiakakos D, Weber D, Döhner H, et al. AML consolidation therapy: timing matters. J Cancer Res Clin Oncol 2023;149(15):13811-13821 [FREE Full text] [doi: 10.1007/s00432-023-05115-0] [Medline: 37535164]
88. Zhu L, Lu W, Soleimani M, Li Z, Zhang M. Electrical impedance tomography guided by digital twins and deep learning for lung monitoring. IEEE Trans Instrum Meas 2023;72:1-9. [doi: 10.1109/tim.2023.3298389]
89. Cappon G, Vettoretti M, Sparacino G, Favero SD, Facchinetti A. ReplayBG: a digital twin-based methodology to identify a personalized model from type 1 diabetes data and simulate glucose concentrations to assess alternative therapies. IEEE Trans Biomed Eng 2023;70(11):3227-3238. [doi: 10.1109/TBME.2023.3286856] [Medline: 37368794]

XSL•FO
RenderX

90. Montgomery L, Willing R, Lanting B. Virtual joint motion simulator accurately predicts effects of femoral component malalignment during TKA. Bioengineering (Basel) 2023;10(5):503 [FREE Full text] [doi: 10.3390/bioengineering10050503] [Medline: 37237573]
91. Dubs L, Charitatos V, Buoso S, Wegener S, Winklhofer S, Alkadhi H, et al. Assessment of extracranial carotid artery disease using digital twins—a pilot study. Neuroimage Clin 2023;38:103435 [FREE Full text] [doi: 10.1016/j.nicl.2023.103435] [Medline: 37245493]
92. Shibuya S, Shido N, Shirai R, Sase K, Ebina K, Chen X, et al. Proposal of simulation-based surgical navigation and development of laparoscopic surgical simulator that reflects motion of surgical instruments in real-world. IJAT 2023;17(3):262-276. [doi: 10.20965/ijat.2023.p0262]
93. Maleki A, Crispino E, Italia SA, Di Salvatore V, Chiacchio MA, Sips FS, et al. Moving forward through the in silico modeling of multiple sclerosis: treatment layer implementation and validation. Comput Struct Biotechnol J 2023;21:3081-3090 [FREE Full text] [doi: 10.1016/j.csbj.2023.05.020] [Medline: 37266405]
94. Moztarzadeh O, Jamshidi MB, Sargolzaei S, Jamshidi A, Baghalipour N, Malekzadeh Moghani M, et al. Metaverse and healthcare: machine learning-enabled digital twins of cancer. Bioengineering (Basel) 2023;10(4):455 [FREE Full text] [doi: 10.3390/bioengineering10040455] [Medline: 37106642]
95. Moztarzadeh O, Jamshidi MB, Sargolzaei S, Keikhaee F, Jamshidi A, Shadroo S, et al. Metaverse and medical diagnosis: a blockchain-based digital twinning approach based on MobileNetV2 algorithm for cervical vertebral maturation. Diagnostics (Basel) 2023;13(8):1485 [FREE Full text] [doi: 10.3390/diagnostics13081485] [Medline: 37189587]
96. Viola F, Del Corso G, De Paulis R, Verzicco R. GPU accelerated digital twins of the human heart open new routes for cardiovascular research. Sci Rep 2023;13(1):8230 [FREE Full text] [doi: 10.1038/s41598-023-34098-8] [Medline: 37217483]
97. Lee JH, Lee HL, Park IY, On SW, Byun SH, Yang BE. Effectiveness of creating digital twins with different digital dentition models and cone-beam computed tomography. Sci Rep 2023;13(1):10603 [FREE Full text] [doi: 10.1038/s41598-023-37774-x] [Medline: 37391453]
98. Straughan R, Kadry K, Parikh SA, Edelman ER, Nezami FR. Fully automated construction of three-dimensional finite element simulations from optical coherence tomography. Comput Biol Med 2023;165:107341. [doi: 10.1016/j.compbiomed.2023.107341] [Medline: 37611423]
99. Batagov A, Dalan R, Wu A, Lai W, Tan CS, Eisenhaber F. Generalized metabolic flux analysis framework provides mechanism-based predictions of ophthalmic complications in type 2 diabetes patients. Health Inf Sci Syst 2023;11(1):18 [FREE Full text] [doi: 10.1007/s13755-023-00218-x] [Medline: 37008895]
100. Sarp S, Kuzlu M, Zhao Y, Gueler O. Digital twin in healthcare: a study for chronic wound management. IEEE J Biomed Health Inform 2023;27(11):5634-5643. [doi: 10.1109/JBHI.2023.3299028] [Medline: 37549083]
101. Zhang Y, Deng Y, Zhang X, Jiao P, Zhang X, Zhao Z. DT-CTNet: a clinically interpretable diagnosis model for fetal distress. Biomed Signal Process Control 2023;86:105190. [doi: 10.1016/j.bspc.2023.105190]
102. Azzolin L, Eichenlaub M, Nagel C, Nairn D, Sánchez J, Unger L, et al. AugmentA: patient-specific augmented atrial model generation tool. Comput Med Imaging Graph 2023;108:102265. [doi: 10.1016/j.compmedimag.2023.102265] [Medline: 37392493]
103. Lin TY, Chiu SYH, Liao LC, Chen SLS, Chiu HM, Chen THH. Assessing overdiagnosis of fecal immunological test screening for colorectal cancer with a digital twin approach. NPJ Digit Med 2023;6(1):24 [FREE Full text] [doi: 10.1038/s41746-023-00763-5] [Medline: 36765093]
104. Manocha A, Afaq Y, Bhatia M. Digital Twin-assisted blockchain-inspired irregular event analysis for eldercare. Knowledge Based Syst 2023;260:110138. [doi: 10.1016/j.knosys.2022.110138]
105. Kirkels FP, van Osta N, Rootwelt-Norberg C, Chivulescu M, van Loon T, Aabel EW, Lie, et al. Monitoring of myocardial involvement in early arrhythmogenic right ventricular cardiomyopathy across the age spectrum. J Am Coll Cardiol 2023;82(9):785-797 [FREE Full text] [doi: 10.1016/j.jacc.2023.05.065] [Medline: 37612010]
106. Avanzato R, Beritelli F, Lombardo A, Ricci C. Heart DT: monitoring and preventing cardiac pathologies using AI and IoT sensors. Future Internet 2023;15(7):223. [doi: 10.3390/fi15070223]
107. Susilo ME, Li CC, Gadkar K, Hernandez G, Huw LY, Jin JY, et al. Systems-based digital twins to help characterize clinical dose-response and propose predictive biomarkers in a phase I study of bispecific antibody, mosunetuzumab, in NHL. Clin Transl Sci 2023;16(7):1134-1148 [FREE Full text] [doi: 10.1111/cts.13501] [Medline: 36908269]
108. Comte N, Pujades S, Courvoisier A, Daniel O, Franco JS, Faure F, et al. Multi-modal data correspondence for the 4D analysis of the spine with adolescent idiopathic scoliosis. Bioengineering (Basel) 2023;10(7):874 [FREE Full text] [doi: 10.3390/bioengineering10070874] [Medline: 37508901]
109. Strocchi M, Longobardi S, Augustin CM, Gsell MAF, Petras A, Rinaldi CA, et al. Cell to whole organ global sensitivity analysis on a four-chamber heart electromechanics model using Gaussian processes emulators. PLoS Comput Biol 2023;19(6):e1011257 [FREE Full text] [doi: 10.1371/journal.pcbi.1011257] [Medline: 37363928]
110. Shu H, Liang R, Li Z, Goodridge A, Zhang X, Ding H, et al. Twin-S: a digital twin for skull base surgery. Int J Comput Assist Radiol Surg 2023;18(6):1077-1084 [FREE Full text] [doi: 10.1007/s11548-023-02863-9] [Medline: 37160583]
111. bioRender. URL: https://www.biorender.com/ [accessed 2024-11-12]

XSL•FO
RenderX

112. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-11-12]
113. Wang J, Li X, Wang P, Liu Q. Bibliometric analysis of digital twin literature: a review of influencing factors and conceptual structure. Technol Anal Strateg Manag 2022;36(1):166-180. [doi: 10.1080/09537325.2022.2026320]
114. Liu X, Jiang D, Tao B, Xiang F, Jiang G, Sun Y, et al. A systematic review of digital twin about physical entities, virtual models, twin data, and applications. Adv Eng Inform 2023;55:101876. [doi: 10.1016/j.aei.2023.101876]
115. Wang B, Zhou H, Li X, Yang G, Zheng P, Song C, et al. Human digital twin in the context of industry 5.0. Robotics Comput Integr Manufacturing 2024;85:102626. [doi: 10.1016/j.rcim.2023.102626]
116. National Academies of Sciences, Engineering, and Medicine. Foundational Research Gaps and Future Directions for Digital Twins. Washington, DC: National Academies Press; 2024.
117. Drummond D, Roukema J, Pijnenburg M. Home monitoring in asthma: towards digital twins. Curr Opin Pulm Med 2023;29(4):270-276. [doi: 10.1097/MCP.0000000000000963] [Medline: 37102597]
118. Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, et al. The "Digital Twin" to enable the vision of precision cardiology. Eur Heart J 2020;41(48):4556-4564 [FREE Full text] [doi: 10.1093/eurheartj/ehaa159] [Medline: 32128588]
119. Physiome Project. URL: https://physiomeproject.org/ [accessed 2024-11-12]
120. VPH Institute. URL: https://www.vph-institute.org/ [accessed 2024-11-12]

Abbreviations

CDT: claimed digital twin
NASA: National Aeronautics and Space Administration
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by G Eysenbach; submitted 17.03.24; peer-reviewed by C Delclaux, S Ricci; comments to author 21.05.24; revised version received 31.05.24; accepted 23.09.24; published 13.11.24.

Please cite as:
Drummond D, Gonsard A
Definitions and Characteristics of Patient Digital Twins Being Developed for Clinical Use: Scoping Review
J Med Internet Res 2024;26:e58504
URL: https://www.jmir.org/2024/1/e58504
doi:10.2196/58504
PMID:

©David Drummond, Apolline Gonsard. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association of Remote Patient-Reported Outcomes and Step Counts With Hospitalization or Death Among Patients With Advanced Cancer Undergoing Chemotherapy: Secondary Analysis of the PROStep Randomized Trial

Christopher R Manz[1], MSHP, MD; Emily Schriver[2,3], MS; William J Ferrell[4,5], MPH; Joelle Williamson[6], MPH; Jonathan Wakim[6], BS; Neda Khan[6,7], MHCI; Michael Kopinsky[7], BSE; Mohan Balachandran[7], MA, MS; Jinbo Chen[5,8], PhD; Mitesh S Patel[9], MBA, MD; Samuel U Takvorian[5,6,10], MSHP, MD; Lawrence N Shulman[5,6,10], MD; Justin E Bekelman[5,10], MD; Ian J Barnett[8*], PhD; Ravi B Parikh[2,4,5,10,11*], MPP, MD

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, United States
[2]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States
[3]Penn Medicine Predictive Healthcare, University of Pennsylvania Health System, Philadelphia, PA, United States
[4]Division of Health Policy, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[5]Penn Center for Cancer Care Innovation, Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, United States
[6]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[7]Center for Health Care Innovation, Penn Medicine, Philadelphia, PA, United States
[8]Department of Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[9]Ascension, St. Louis, MO, United States
[10]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, United States
[11]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, United States
[*]these authors contributed equally

Corresponding Author:
Ravi B Parikh, MPP, MD
Division of Health Policy
Perelman School of Medicine
University of Pennsylvania
423 Guardian Drive
Room 1102
Philadelphia, PA, 19104
United States
Phone: 1 3524224285
Email: Ravi.Parikh@pennmedicine.upenn.edu

Abstract

Background: Patients with advanced cancer undergoing chemotherapy experience significant symptoms and declines in functional status, which are associated with poor outcomes. Remote monitoring of patient-reported outcomes (PROs; symptoms) and step counts (functional status) may proactively identify patients at risk of hospitalization or death.

Objective: The aim of this study is to evaluate the association of (1) longitudinal PROs with step counts and (2) PROs and step counts with hospitalization or death.

Methods: The PROStep randomized trial enrolled 108 patients with advanced gastrointestinal or lung cancers undergoing cytotoxic chemotherapy at a large academic cancer center. Patients were randomized to weekly text-based monitoring of 8 PROs plus continuous step count monitoring via Fitbit (Google) versus usual care. This preplanned secondary analysis included 57 of 75 patients randomized to the intervention who had PRO and step count data. We analyzed the associations between PROs and mean daily step counts and the associations of PROs and step counts with the composite outcome of hospitalization or death using bootstrapped generalized linear models to account for longitudinal data.

XSL•FO
RenderX

Results: Among 57 patients, the mean age was 57 (SD 10.9) years, 24 (42%) were female, 43 (75%) had advanced gastrointestinal cancer, 14 (25%) had advanced lung cancer, and 25 (44%) were hospitalized or died during follow-up. A 1-point weekly increase (on a 32-point scale) in aggregate PRO score was associated with 247 fewer mean daily steps (95% CI –277 to –213; *P*<.001). PROs most strongly associated with step count decline were patient-reported activity (daily step change –892), nausea score (–677), and constipation score (524). A 1-point weekly increase in aggregate PRO score was associated with 20% greater odds of hospitalization or death (adjusted odds ratio [aOR] 1.2, 95% CI 1.1-1.4; *P*=.01). PROs most strongly associated with hospitalization or death were pain (aOR 3.2, 95% CI 1.6-6.5; *P*<.001), decreased activity (aOR 3.2, 95% CI 1.4-7.1; *P*=.01), dyspnea (aOR 2.6, 95% CI 1.2-5.5; *P*=.02), and sadness (aOR 2.1, 95% CI 1.1-4.3; *P*=.03). A decrease in 1000 steps was associated with 16% greater odds of hospitalization or death (aOR 1.2, 95% CI 1.0-1.3; *P*=.03). Compared with baseline, mean daily step count decreased 7% (n=274 steps), 9% (n=351 steps), and 16% (n=667 steps) in the 3, 2, and 1 weeks before hospitalization or death, respectively.

Conclusions: In this secondary analysis of a randomized trial among patients with advanced cancer, higher symptom burden and decreased step count were independently associated with and predictably worsened close to hospitalization or death. Future interventions should leverage longitudinal PRO and step count data to target interventions toward patients at risk for poor outcomes.

Trial Registration: ClinicalTrials.gov NCT04616768; https://clinicaltrials.gov/study/NCT04616768

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-054675

(J Med Internet Res 2024;26:e51059) doi:10.2196/51059

KEYWORDS

wearables; accelerometers; patient-reported outcomes; step counts; oncology; accelerometer; patient-generated health data; cancer; death; chemotherapy; symptoms; gastrointestinal cancer; lung cancer; monitoring; symptom burden; risk; hospitalization; mobile phone

Introduction

Patients with advanced cancer who receive chemotherapy often experience significant symptoms, declines in functional status, and hospitalization [1-3]. Patient-reported outcomes (PROs) measure symptom burden and well-being. In clinical trials, routine collection of PROs for patients with cancer undergoing treatment is associated with decreased acute care use and improved overall survival [4,5]. Similar to increased patient symptom burden, the decline in patient functional status often presages adverse events, hospitalizations, disease progression, and death [6,7]. While PROs are useful for monitoring changes in symptoms and reported activity levels, they do not provide objective measures of functional status. Step counts are a proxy measure of functional status and thus identify patients who are at a high risk of poor outcomes [6]. However, among people with advanced cancer undergoing treatment, the prognostic use of step count monitoring has never been explored [8].

Previously published studies in oncology demonstrate that lower step counts are associated with higher odds of adverse events, hospitalization, and death [6,9-11]. Yet, these studies have shortcomings that limit their generalizability to patients with advanced cancers. First, most studies were small (<50 patients), tracked step counts for less than 1 month, and focused only on patients receiving therapy with curative intent. Second, these studies measure associations between cross-sectional PROs or step counts and adverse outcomes. Few studies measured longitudinal PROs and step counts over several weeks, thus precluding evaluation of how patient-reported symptoms and objective measures of functional status interact and change over time.

Emerging value-based oncology models, including Medicare's Enhancing Oncology Model that began in 2023, required measurement of electronic PROs. There is an urgent need to identify whether step count data complement PRO monitoring and improve early identification of patients with advanced cancer who are at risk of future adverse outcomes. The PROStep trial was a pragmatic, randomized controlled trial of patients with advanced gastrointestinal (GI) and lung cancers treated with intravenous chemotherapy [12]. Intervention patients received remote, longitudinal PRO surveys and step count monitoring. The objectives of this preplanned secondary analysis of the PROStep trial were to (1) evaluate the association between longitudinal PROs and step counts and (2) assess the association of PROs and step counts with the composite outcome of hospitalization or death. Our overarching hypothesis was that lower step counts would be associated with greater symptom burden measured by PROs and that higher symptom burden and lower step counts would be independently associated with subsequent hospitalization or death.

Methods

Study Design and Cohort

This is a preplanned secondary analysis of the PROStep randomized trial (ClinicalTrials.gov NCT04616768). The trial's design, protocol, and main results have been previously published [13]. Briefly, PROStep tested the effect of clinician and patient-centered dashboards combining weekly PRO data, collected via text message, and step count monitoring, collected via a wearable accelerometer, on the primary outcomes of patient-reported clinician understanding of the patient's symptoms and functional status. The study population consisted of 108 patients with stage IV GI or lung cancers undergoing

XSL•FO
RenderX

intravenous cytotoxic chemotherapy at a tertiary academic cancer center between November 17, 2020, and June 17, 2021. Eligible patients were English-speaking, used a smartphone with SMS and Bluetooth capabilities, and received their primary oncology care at the study center. Patients undergoing monotherapy with checkpoint inhibitor therapy, targeted therapies (eg, cetuximab and trastuzumab), or oral chemotherapy without concurrent intravenous chemotherapy were excluded. Additionally, patients on active therapeutic interventional trials or confined to a wheelchair or bed were excluded.

Patients were electronically randomized in a 1:1:1 fashion, stratified by cancer type (GI or lung), to 1 of the 3 arms—(A) standard care (control), (B) PROStep intervention, or (C) PROStep intervention with active choice prompts (see Figure S1 in Multimedia Appendix 1 for CONSORT diagram). Patients randomized to the PROStep intervention received weekly 8-question text-based PRO surveys and passive step count monitoring via a wearable accelerometer (Fitbit Inspire HR; Google). PROs and step count level data were summarized in a dashboard and provided to the patient's medical oncologist or advanced practitioner before a clinic visit. Patients in arm C also received an automated active choice text on the morning of each oncology visit, which prompted patients to discuss concerning symptoms with their oncologist.

In this preplanned secondary analysis, we evaluated the association between step counts and PROs among 57 of the 75 patients originally randomized to arms B or C who had any PRO and step count data; 18 of the 75 randomized patients did not have step counts and PRO data and could not be analyzed. As the primary analysis showed no difference in any outcome for arms B and C, study arms B and C were combined for this analysis. We then evaluated the independent associations between step count levels and PROs with subsequent hospitalization or death. The University of Pennsylvania Institutional Review Board approved the study and participants provided written or electronic informed consent during the trial.

PRO and Step Count Assessment

Symptoms were assessed using an 8-question, text-based PRO survey drawn from the PRO version of the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE), scored on a 5-point Likert scale, from 0 (not present) to 4 (disabling; see Table S1 in Multimedia Appendix 1) [14,15]. PRO surveys were sent to participants weekly on Monday mornings at 10 AM via a text message on their mobile phones. For patients who did not respond to the PRO survey, automatic reminder alerts were sent on Tuesdays and Thursdays. To measure overall symptom burden, we aggregated the scores from the 8 PRO questions in an aggregate PRO score on a 0-32 scale.

Step counts were measured using the Fitbit Inspire HR. Patients enrolled in the trial using the Way2Health app and then were given Fitbits linked to the app so that they could submit step count data by opening the app. Fitbits had a 5-day memory, so the patients received a text reminder to synchronize their Fitbit 2 times per week as well as 2 days before a clinic visit unless the data were synchronized in the prior 24 hours. The Fitbit Inspire HR measured exact daily step counts; these were summarized on a weekly basis as average daily step counts each week. Only days in which the Fitbit was appropriately synced were used in the average step count calculation (ie, days with no Fitbit data were excluded from the step count calculation).

Outcomes

The primary outcome was a composite outcome of hospitalization or death, collected by the study research coordinator during the course of the trial for all enrolled patients. Associations between PROs and step counts were also assessed.

Statistical Analysis

We reported descriptive patient characteristics, mean adherence to weekly PRO surveys (number of completed weekly surveys divided by the number of weeks enrolled in the study), and mean weekly adherence to step count monitoring (percentage of weeks where step counts were available for >3 days in a given week). To determine the association between PROs and step counts in a given week, we modeled their concurrent association. To determine the association of change in PRO scores and the change in step counts in a given week, we calculated bootstrapped means with 1000 iterations using bootstrapped generalized linear models. To determine the association of the longitudinal measurement of composite PRO score and average weekly steps on the composite outcome of hospitalization or death, we used bootstrapped generalized linear models with multiple "outputation" to account for repeated measures from the longitudinal data [16]. To account for the decline in the degree of association between aggregate PRO score and average weekly steps as the time between activity data and the event increases, we assumed exponential decay between the time of the PRO or step count being recorded and the composite outcome. Specifically, we multiplied regression coefficients with an exponentially decaying term and used a grid search to determine the maximum likelihood exponential decay parameter estimate used for both PRO scores and Fitbit. Finally, to evaluate the trends in PRO scores and steps in the weeks leading up to the outcomes, we computed bootstrapped means. Analyses were performed using R (version 4.2.1; R Core Team) and Python (version 3.9.13; Python Software Foundation). Two-sided hypothesis testing with α=.05 was used to assess significance.

Ethical Considerations

This study underwent ethical review and was approved by the University of Pennsylvania Institutional Review Board (843616) and was registered on ClinicalTrials.gov (NCT04616768). Written informed consent was obtained from all participants in the trial. The study data were anonymized and deidentified. Participants in arms A, B, and C were compensated up to US $50 in gift cards upon completing their use surveys at 3 and 6 months after enrollment (US $25 each). Participants in arms B and C were permitted to keep their Fitbit as part of the trial (US $80 value).

XSL•FO
RenderX

Results

Baseline Characteristics

The 57 patients had a mean age of 57 (SD 10.9) years, 24 (42%) were female, 49 (86%) were White, and 3 (5%) were Black. A total of 43 (75%) patients had advanced GI cancers and 14 (25%) had advanced lung cancer (Table 1). A total of 79% (n=45) of patients completed 24 weeks of the study; the most common reasons for disenrollment were death (n=8) and voluntary drop out (n=4). Mean adherence to weekly PRO surveys was 77% (SD 29.7%), with 84% (n=48) of patients reporting PROs more than 50% of enrolled weeks. Mean weekly adherence to step count monitoring was 69% (SD 36.5%), with 70% (n=40) of patients recording step counts more than 50% of weeks enrolled in the study.

Table 1. Demographic characteristics of patients included in this secondary analysis of the PROStep trial. Sex and race were self-reported by patients.

Characteristics	Overall (N=57)
Age, mean (SD)	57.1 (10.9)
Cancer group, n (%)	
Gastrointestinal	43 (75)
Lung	14 (25)
Sex, n (%)	
Female	24 (42)
Male	33 (58)
Race, n (%)	
Asian	3 (5)
Black	3 (5)
White	49 (86)
Other	2 (4)

Associations Between PROs and Step Counts in a Given Week

In univariate analyses evaluating associations of PROs and step counts within a single week, a 1-point higher aggregate PRO score (out of 32) was associated with 150 fewer mean daily steps (95% CI –183 to –120 steps; $P<.001$) (Figure S2 in Multimedia Appendix 1). In a given week, 1-point higher scores (up to a score of 4) in shortness of breath ($P=.03$), sadness ($P<.001$), anxiety ($P<.001$), patient-reported activity ($P<.001$), nausea ($P<.001$), and diarrhea ($P<.01$) were associated with 112, 244, 125, 792, 564, and 248 fewer mean daily steps, respectively (Table S2 in Multimedia Appendix 1). Pain (–71 mean daily steps; $P=.23$), and constipation (–90; $P=.06$) were not independent factors. When examining between-PRO correlations, sadness and anxiety ($r=0.67$) and shortness of breath and patient-reported activity ($r=0.64$) were highly correlated. Patient-reported activity ($r=-0.41$) and shortness of breath ($r=-0.36$) were most correlated with decreased mean daily step count (Figure 1).

XSL•FO
RenderX

Figure 1. Correlations between patient-reported symptoms and step counts in a given week. Cell values represent correlation coefficients (r) between symptoms (expressed on a 0-4 scale) and step counts (expressed in thousands of steps) recorded in the same week. Blue values represent positive associations and red values represent negative associations.

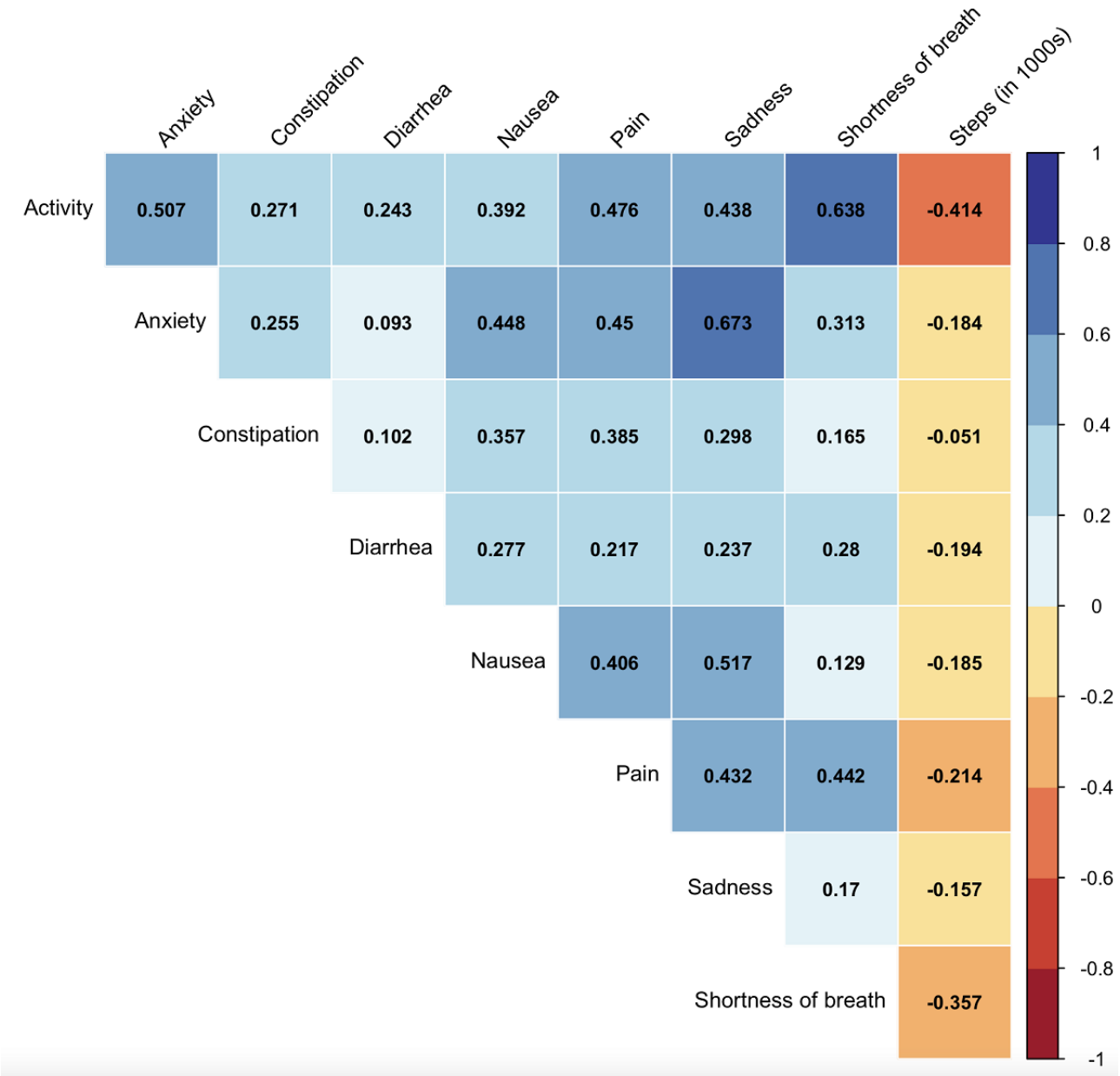

Associations Between Longitudinal Changes in PROs and Step Counts

In univariate analyses assessing week-to-week changes in PROs and changes in step counts, changes in aggregate PRO score from the prior week to the current week were associated with stepwise increases or decreases in the mean daily step counts (Figure 2). A 1-point increase in aggregate PRO score from the prior week was associated with a decrease of 247 mean daily steps (95% CI –277 to –213; *P*<.001). One-point increases in following PROs from the prior week had the strongest associations with decreased mean daily steps—patient-reported activity (mean daily step change –892, 95% CI –1050 to –758; *P*<.001), nausea (–677, 95% CI –770 to –588; *P*<.001), constipation (–524, 95% CI –614 to –431; *P*<.001), shortness of breath (–399, 95% CI –498 to –302; *P*<.01), pain (–304, 95% CI –409 to –204; *P*<.001), sadness (–382, 95% CI –462 to –302; *P*<.001), and anxiety (–125, 95% CI –213 to –42; *P*<.001) (Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Associations between longitudinal changes in patient-reported symptoms and step counts. Longitudinal changes are displayed in box and whisker plots, measured from week n–1 to week n. Open circles represent outliers. PRO: patient-reported outcome.

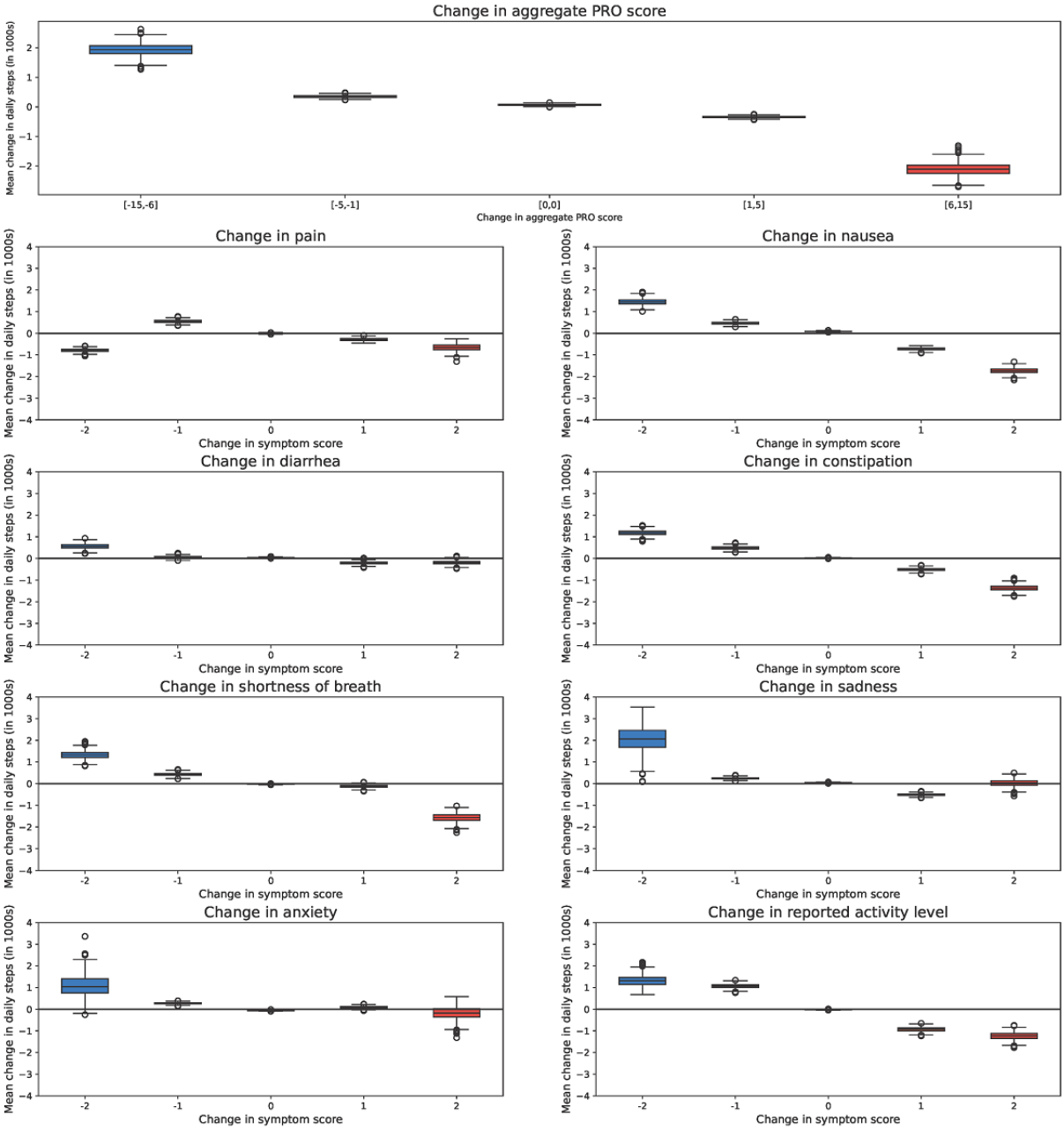

Associations Between PROs, Step Counts, and Hospitalization or Death

Among 57 patients, 21 (37%) patients were hospitalized and 8 (14%) died during the follow-up period. The rate of the composite outcome of hospitalization or death was 44% (n=25). On average, a 1-point increase in the aggregate PRO score was associated with a 20% increase in adjusted odds of hospitalization or death (adjusted odds ratio [aOR] 1.2, 95% CI 1.1-1.4; *P*=.01; Table 2). In a given week, a 1-point increase in patient-reported pain (aOR 3.2, 95% CI 1.6-6.5; *P*=.01), activity (aOR 3.2, 95% CI 1.4-7.1; *P*=.01), shortness of breath (aOR 2.6, 95% CI 1.2-5.5; *P*=.01), sadness (aOR 2.1, 95% CI 1.1-4.3; *P*=.03), and anxiety (aOR 2.7, 95% CI 1.3-5.6; *P*=.01) were most associated with increased odds of hospitalization or death. After adjusting for aggregate PRO score, a decrease in 1000 mean daily steps was associated with 16% increased odds of hospitalization or death (aOR 1.16, 95% CI 1.01-1.33; *P*=.03). Patients who were hospitalized or died had a progressive increase in aggregate PRO score and a decrease in mean daily steps in the 4 weeks leading up to the event, but patients who did not experience an event had no change in aggregate PRO score or mean daily steps in the prior 4 weeks (Figure 3). Compared with baseline, mean daily step count decreased 7% (n=274/4112 steps), 9% (n=351/4112 steps), and 16% (n=667/4112 steps) in the 3, 2, and 1 weeks before hospitalization or death, respectively. Mean aggregate PRO score increased by 11% (n=0.8), 25% (n=1.9), and 36% (n=2.8), in the 3, 2, and 1 weeks before hospitalization or death, respectively.

XSL•FO
RenderX

Table 2. Association of symptoms and step counts with the composite outcome of hospitalization or death[a].

Symptoms[a]	Symptom		Mean daily step count (per 1000 steps)	
	aOR[b] (95% CI)	*P* value	aOR (95% CI)	*P* value
Composite PRO[c] score	1.20 (1.1-1.4)	.01	0.86 (0.75-0.99)	.03
Pain	3.23 (1.6-6.5)	<.001	0.84 (0.72-0.98)	.02
Activity	3.16 (1.4-7.1)	.01	0.9 (0.79-1.03)	.14
Anxiety	2.69 (1.3-5.6)	.01	0.82 (0.70-0.95)	.01
Shortness of breath	2.58 (1.2-5.5)	.02	0.87 (0.76-1.01)	.07
Sadness	2.14 (1.1-4.3)	.03	0.82 (0.65-0.95)	.01
Constipation	1.92 (0.98-3.8)	.06	0.81 (0.69-0.57)	.01
Diarrhea	1.21 (0.78-1.9)	.40	0.81 (0.69-0.94)	.01
Nausea	0.91 (0.59-1.4)	.66	0.81 (0.70-0.95)	.01

[a]Each row represents a separate model composed of the symptom in the left column and step counts (in thousands) as the predictor variables and the composite outcome of hospitalization and death as the outcome.

[b]aOR: adjusted odds ratio.

[c]PRO: patient-reported outcome.

Figure 3. Longitudinal association between aggregate PRO symptom score, step counts, and hospitalization or death. The x-axis refers to the number of weeks before an event (hospitalization or death). Purple lines represent averages for patients without an event. Green lines represent averages for patients with an event. PRO: patient-reported outcome.

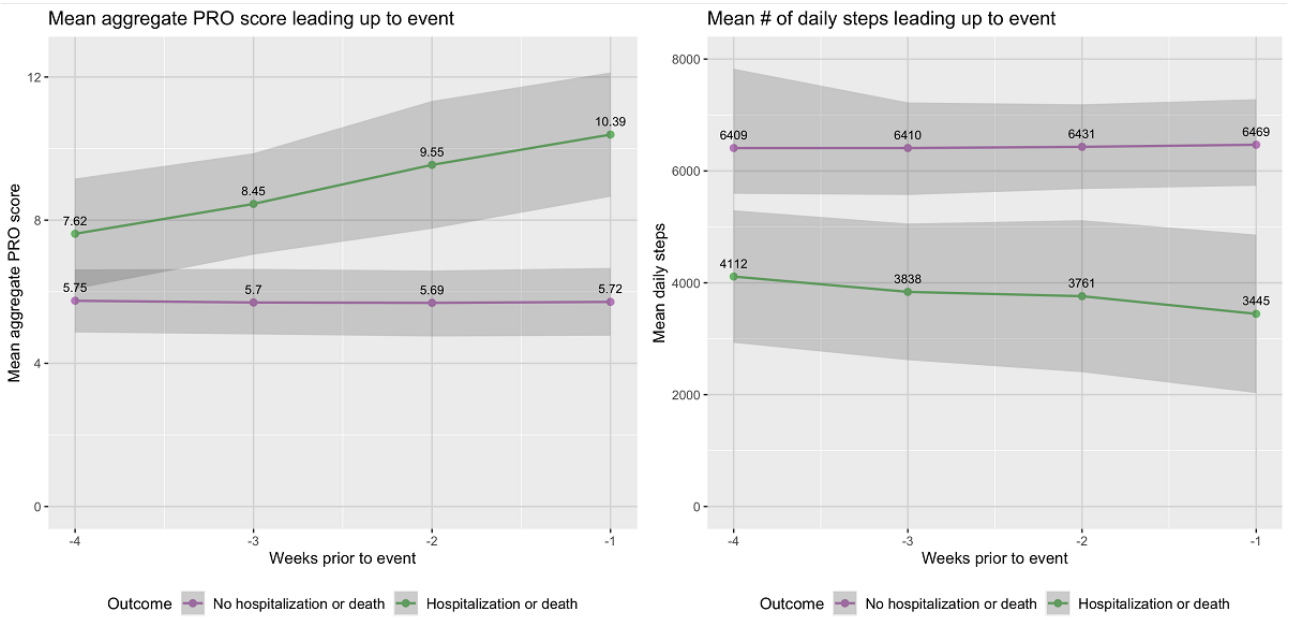

Discussion

Principal Findings

In this preplanned secondary analysis of a randomized clinical trial among patients with incurable GI and lung cancers receiving chemotherapy, higher symptom burden measured by remote PROs was associated with lower daily step count. Moreover, higher symptom burden and decreased daily step counts were independently associated with an increased risk of hospitalization or death. PROs and step counts predictably worsened in the 4 weeks before hospitalization or death, whereas the individuals who did not experience hospitalization or death had stable PROs and step counts.

Prior studies among patients with cancer undergoing chemoradiation demonstrated that lower step counts, measured cross-sectionally over brief periods, are associated with subsequent acute care use and worse prognosis [6,9-11]. We build upon these findings by demonstrating that longitudinally measured PROs and step counts are predictive of hospitalization and death. PROs and step count predictably worsened in the 4 weeks before hospitalization and death. This indicates that among patients with advanced cancer receiving chemotherapy, longitudinal remotely monitored patient-generated health data may identify at-risk patients weeks before hospitalization or death—a window in which proactive interventions, including goals-of-care conversations, may improve outcomes or goal-concordant care. Importantly, decreasing patient-reported

XSL•FO
RenderX

activity level was highly predictive of hospitalization or death (aOR 3.2, 95% CI 1.4-7.1) even when adjusting for step counts, suggesting that patient-reported activity may complement objective measures of step counts.

Our study is among the first to assess associations between longitudinal PROs and step counts and downstream use, among patients with cancer. We show that patients with higher symptom burden also experience fewer daily step counts in a given week. Furthermore, in longitudinal analyses, daily step counts decreased as symptom burden increased. Worsening nausea, constipation, shortness of breath, and sadness had the largest associations with decreased step count.

There is growing evidence to support routine collection of PROs, and Medicare's Enhancing Oncology Model incentivizes PRO measurement. Our results may facilitate efforts to use these PRO data in predictive models by identifying those PROs most likely to contribute to adverse events. This study provides novel provocative data that step count monitoring independently identifies at-risk patients and could be used synergistically with PROs as part of targeted care delivery interventions to prevent acute care use or target supportive care interventions. Such interventions may include additional symptom support, home services like physical therapy, treatment modifications, palliative care consultation, or goals-of-care conversations. For instance, worsening step counts may prompt physical therapy interventions to improve functional status and perhaps reduce the risk of hospitalization. Policymakers may consider strategically incorporating activity monitoring with electronic PRO monitoring mandates to enhance risk prediction and stratification of high-risk populations. Future studies should examine how longitudinal PRO and step count data could identify decreasing chemotherapy tolerance, enabling earlier dose reduction that might improve quality of life and extend treatment tolerability. Future studies should also investigate whether longitudinal PRO and step count data predict disease progression or symptomatic disease, potentially prompting earlier imaging and switches to new therapy.

Limitations

This study has several limitations. It includes patients with only 2 types of cancer enrolled in a clinical trial at a single tertiary cancer center and all undergoing chemotherapy, which may enrich for a sicker population that may have the most pronounced changes in symptoms and step counts. Adherence to weekly PROs and step counts among patients in this analysis was 77% (SD 29.7%) and 69% (SD 37%), respectively, which is higher than the adherence reported in real-world studies [17]. However, we excluded patients who did not have PRO or step count data, and adherence in the full cohort was similar to the levels reported in these real-world studies [12]. Moreover, step counts could only be tracked if patients wore the wearable monitor—if patients were less likely to wear the monitors in certain circumstances (eg, when symptoms worsen), data missingness may be informative. A subsequent analysis will explore how data missingness can also be used to improve predictive power. The study also used an abbreviated sample of PRO- CTCAE questions targeted to patients with GI and lung cancers; findings may differ using the full PRO-CTCAE question panel and in patients with other malignancies. Finally, our sample size and overall number of outcomes were too small to disaggregate predictors of hospitalizations or death individually. We hope future prospective work with larger sample sizes will facilitate the disaggregation of these outcomes. Nonetheless, prediction of both death and hospitalizations near the end of life can be used to trigger interventions such as goals-of-care conversations that may enable better goal-concordant care.

Conclusions

This study of remote monitoring of PROs and step counts shows that changes in these measures predict hospitalization and death for patients with advanced cancer undergoing treatment. Future work should validate these findings in larger, more diverse populations and translate these results into interventions that can avoid acute care use and improve supportive care.

Acknowledgments

This work was supported by funding from the University of Pennsylvania Institute of Translational Medicine and Therapeutics (grant or award 400-4133-4-578849-xxxx-2446–4732) and National Institutes of Health (NIH) (grants 5T32CA092203-18 to CM and K08CA263541 to RBP). The funders had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit this paper for publication.

Data Availability

The data supporting the findings of this study are available upon request from the corresponding author. The data are not publicly available because of privacy concerns.

Conflicts of Interest

The authors are solely responsible for the design and conduct of this study, study analyses, and the drafting of this paper. CRM reports funding from Genentech and the American Cancer Society, outside the submitted work. MSP reports employment at Ascension.org and being owner of Catalyst Health, a behavior change and technology consulting firm, outside the submitted work. RBP reports receiving grants from the National Institutes of Health, Prostate Cancer Foundation, National Palliative Care Research Center, National Comprehensive Cancer Network Foundation, Conquer Cancer Foundation, Humana, Emerson Collective, and Veterans Health Administration; receiving personal fees and equity from GNS Healthcare, Thyme Care, and Onc.AI; receiving

XSL•FO
RenderX

personal fees from the Cancer Study Group, Biofourmis, ConcertAI, CreditSuisse, Humana, Genetic Chemistry Therapeutics, and Nanology; filing a patent related to electronic health record–based predictive algorithms, receiving honoraria from Flatiron and Medscape; being an unpaid board member of the Coalition to Transform Advanced Care and American Cancer Society; and serving on a leadership consortium (unpaid) at the National Quality Forum, all outside the submitted work. JEB reports personal fees from Reimagine Care, Healthcare Foundry, and Astrazeneca and grants from Emerson Collective (co-investigator), Loxo@Lilly (co-investigator), and Gilead (co-investigator), outside the submitted work.

Multimedia Appendix 1
Consolidated Standards of Reporting Trials (CONSORT) diagram from PROStep trial; PRO survey: National Cancer Institute's common terminology criteria for adverse events and Patient Reported Functional Status (PRFS) tool; associations between patient-reported symptoms and step counts in a given week; and association between patient-reported symptoms and mean daily step count.
[DOCX File , 243 KB - jmir_v26i1e51059_app1.docx]

References

1. Nipp RD, El-Jawahri A, Moran SM, D'Arpino SM, Johnson PC, Lage DE, et al. The relationship between physical and psychological symptoms and health care utilization in hospitalized patients with advanced cancer. Cancer 2017;123(23):4720-4727 [FREE Full text] [doi: 10.1002/cncr.30912] [Medline: 29057450]
2. Goldberg SL, Paramanathan D, Khoury R, Patel S, Jagun D, Arunajadai S, et al. A patient-reported outcome instrument to assess symptom burden and predict survival in patients with advanced cancer: flipping the paradigm to improve timing of palliative and end-of-life discussions and reduce unwanted health care costs. Oncologist 2019;24(1):76-85 [FREE Full text] [doi: 10.1634/theoncologist.2018-0238] [Medline: 30266893]
3. Morrison EJ, Novotny PJ, Sloan JA, Yang P, Patten CA, Ruddy KJ, et al. Emotional problems, quality of life, and symptom burden in patients with lung cancer. Clin Lung Cancer 2017;18(5):497-503 [FREE Full text] [doi: 10.1016/j.cllc.2017.02.008] [Medline: 28412094]
4. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
5. Basch E, Schrag D, Henson S, Jansen J, Ginos B, Stover AM, et al. Effect of electronic symptom monitoring on patient-reported outcomes among patients with metastatic cancer: a randomized clinical trial. JAMA 2022;327(24):2413-2422 [FREE Full text] [doi: 10.1001/jama.2022.9265] [Medline: 35661856]
6. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, et al. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. NPJ Digit Med 2018;1:27 [FREE Full text] [doi: 10.1038/s41746-018-0032-6] [Medline: 31304309]
7. Kerrigan K, Patel SB, Haaland B, Ose D, Chalmers AW, Haydell T, et al. Prognostic significance of patient-reported outcomes in cancer. JCO Oncol Pract 2020;16(4):e313-e323 [FREE Full text] [doi: 10.1200/JOP.19.00329] [Medline: 32048943]
8. Gresham G, Schrack J, Gresham LM, Shinde AM, Hendifar AE, Tuli R, et al. Wearable activity monitors in oncology trials: current use of an emerging technology. Contemp Clin Trials 2018;64:13-21 [FREE Full text] [doi: 10.1016/j.cct.2017.11.002] [Medline: 29129704]
9. Izmailova ES, Cohen A, Kabarriti R, Eichler J, Huang C, Ohri N. Daily step counts to predict hospitalizations during chemoradiotherapy for head and neck cancer. J Clin Oncol 2021;39(15_suppl):1571-1571. [doi: 10.1200/jco.2021.39.15_suppl.1571]
10. France D, Nath P, Mahadevan S, Slagle J, Agarwal R, Kohutek Z, et al. Using Fitbit data to predict clinical deterioration and unplanned treatment events in cancer outpatients. J Clin Oncol 2021;39(15_suppl):e13560-e13560. [doi: 10.1200/jco.2021.39.15_suppl.e13560]
11. Ohri N, Halmos B, Bodner WR, Cheng H, Guha C, Kalnicki S, et al. Daily step counts: a new prognostic factor in locally advanced non-small cell lung cancer? Int J Radiat Oncol Biol Phys 2019;105(4):745-751. [doi: 10.1016/j.ijrobp.2019.07.055] [Medline: 31398385]
12. Manz C, Ferrell W, Williamson J, Schriver E, Wakim J, Khan N, et al. Remote patient-reported symptoms and passive activity monitoring to improve patient-clinician communication regarding symptoms and functional status: a randomized controlled trial (PROStep). J Clin Oncol 2022;40(16_suppl):1506-1506. [doi: 10.1200/jco.2022.40.16_suppl.1506]
13. Parikh RB, Ferrell W, Wakim J, Williamson J, Khan N, Kopinsky M, et al. Patient and clinician nudges to improve symptom management in advanced cancer using patient-generated health data: study protocol for the PROStep randomised controlled trial. BMJ Open 2022;12(5):e054675 [FREE Full text] [doi: 10.1136/bmjopen-2021-054675] [Medline: 35551088]
14. Common Terminology Criteria for Adverse Events (CTCAE). National Cancer Institute. 2023. URL: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm [accessed 2023-01-18]

XSL•FO
RenderX

15. Patient Reported Functional Status (PRFS) tool. Cancer Care Ontario. URL: https://www.cancercareontario.ca/sites/ccocancercare/files/assets/CCOPRFS-English.pdf [accessed 2023-01-18]
16. Follmann D, Proschan M, Leifer E. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. Biometrics 2003;59(2):420-429. [doi: 10.1111/1541-0420.00049] [Medline: 12926727]
17. Takvorian SU, Anderson RT, Gabriel PE, Poznyak D, Lee S, Simon S, et al. Real-world adherence to patient-reported outcome monitoring as a cancer care quality metric. JCO Oncol Pract 2022;18(9):e1454-e1465 [FREE Full text] [doi: 10.1200/OP.21.00855] [Medline: 35675586]

Abbreviations

aOR: adjusted odds ratio
CTCAE: Common Terminology Criteria for Adverse Events
GI: gastrointestinal
PRO: patient-reported outcome

Edited by A Mavragani; submitted 21.07.23; peer-reviewed by S Goldberg, P Innominato; comments to author 02.11.23; revised version received 09.11.23; accepted 29.01.24; published 17.05.24.

Please cite as:
Manz CR, Schriver E, Ferrell WJ, Williamson J, Wakim J, Khan N, Kopinsky M, Balachandran M, Chen J, Patel MS, Takvorian SU, Shulman LN, Bekelman JE, Barnett IJ, Parikh RB
Association of Remote Patient-Reported Outcomes and Step Counts With Hospitalization or Death Among Patients With Advanced Cancer Undergoing Chemotherapy: Secondary Analysis of the PROStep Randomized Trial
J Med Internet Res 2024;26:e51059
URL: https://www.jmir.org/2024/1/e51059
doi:10.2196/51059
PMID:38758583

©Christopher R Manz, Emily Schriver, William J Ferrell, Joelle Williamson, Jonathan Wakim, Neda Khan, Michael Kopinsky, Mohan Balachandran, Jinbo Chen, Mitesh S Patel, Samuel U Takvorian, Lawrence N Shulman, Justin E Bekelman, Ian J Barnett, Ravi B Parikh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of Save Medical Corporation (SMC)–01, a Smartphone App Designed to Support Type 2 Diabetes Self-Management Based on Established Guidelines: Randomized Controlled Trial

Nicholas Leung[1], BA; Kayo Waki[2,3,4], MD, MPH, PhD; Satoshi Nozoe[5], MEng; Shunpei Enomoto[2], MAS; Ryo Saito[2]; Sakurako Hamagami[6]; Toshimasa Yamauchi[4], MD, PhD; Masaomi Nangaku[7], MD, PhD; Kazuhiko Ohe[2,3], MD, PhD; Yukiko Onishi[8], MD, PhD

[1]Donald and Barbara Zucker School of Medicine, Hempstead, NY, United States
[2]Department of Planning, Information and Management, University of Tokyo Hospital, Tokyo, Japan
[3]Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[4]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[5]Save Medical Corporation, Tokyo, Japan
[6]Faculty of Biology, University of Cambridge, Cambridge, United Kingdom
[7]Division of Nephrology and Endocrinology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[8]The Institute of Medical Science, Asahi Life Foundation, Tokyo, Japan

Corresponding Author:
Kayo Waki, MD, MPH, PhD
Department of Planning, Information and Management
University of Tokyo Hospital
Hongo 7-3-1
Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 0358009129
Email: kwaki-tky@m.u-tokyo.ac.jp

Abstract

Background: Lifestyle modifications are a key part of type 2 diabetes mellitus treatment. Many patients find long-term self-management difficult, and mobile apps could be a solution. In 2010, in the United States, a mobile app was approved as an official medical device. Similar apps have entered the Japanese market but are yet to be classified as medical devices.

Objective: The objective of this study was to determine the efficacy of Save Medical Corporation (SMC)–01, a mobile app for the support of lifestyle modifications among Japanese patients with type 2 diabetes mellitus.

Methods: This was a 24-week multi-institutional, prospective randomized controlled trial. The intervention group received SMC-01, an app with functions allowing patients to record data and receive personalized feedback to encourage a healthier lifestyle. The control group used paper journals for diabetes self-management. The primary outcome was the between-group difference in change in hemoglobin A_{1c} from baseline to week 12.

Results: The change in hemoglobin A_{1c} from baseline to week 12 was –0.05% (95% CI –0.14% to 0.04%) in the intervention group and 0.06% (95% CI –0.04% to 0.15%) in the control group. The between-group difference in change was –0.11% (95% CI –0.24% to 0.03%; *P*=.11).

Conclusions: There was no statistically significant change in glycemic control. The lack of change could be due to SMC-01 insufficiently inducing behavior change, absence of screening for patients who have high intention to change their lifestyle, low effective usage of SMC-01 due to design issues, or problems with the SMC-01 intervention. Future efforts should focus on these issues in the early phase of developing interventions.

Trial Registration: Japan Registry of Clinical Trials jRCT2032200033; https://jrct.niph.go.jp/latest-detail/jRCT2032200033

(J Med Internet Res 2024;26:e53740) doi:10.2196/53740

XSL•FO
RenderX

KEYWORDS

behavioral change; HbA1c; hemoglobin A1c; mHealth; randomized controlled trial; smartphone application; T2DM; diabetes mellitus; mobile apps; mobile app; type 2 diabetes; diabetes; self-management; Japan; multi-institutional; RCT; RCTs; efficacy; app development; safety; mobile phone

Introduction

In recent years, the number of patients with type 2 diabetes mellitus (T2DM) has been on the rise due to unhealthy diets, sedentary lifestyles, and an aging population [1]. The epidemic is only expected to worsen, with the number of adults with diabetes predicted to reach 415 million by 2040, costing the health care system US $802 billion a year [2]. Improving diabetes treatment is essential.

Treatment approaches for diabetes currently consist of pharmacological interventions and lifestyle modifications of diet and physical exercise. Lifestyle modifications have been shown to be effective in treating diabetes in several studies [3-5], enhancing insulin sensitivity and improving glycemic control [4]. Lifestyle interventions have improved diabetes control and cardiovascular risk factors and have reduced the use of diabetic medications [6]. They have shown a substantial and parallel reduction in glucose-lowering medication [7]. However, many patients struggle with adhering to lifestyle treatments [4,8]. This may be due to several factors, including socioeconomic and temporal factors that make it difficult to fully implement behavioral change [9]. There may be a lack of actionable advice from physicians and other health care workers on how to maintain a healthy lifestyle [10]. The way the advice is communicated is also important—a holistic understanding of the patient's situation and motivational interviewing improve self-efficacy and adherence to lifestyle interventions [11-13].

Mobile health (mHealth; the use of mobile phones and other connected devices to improve health) could address some of these barriers. By using apps on mobile devices, mHealth enables remote monitoring of patients and delivery of personalized clinical advice [8]. mHealth has been shown to increase adherence to and efficacy of lifestyle interventions while reducing the financial burden on the health care system [14-17]. Mobile phone ownership rates reached 95% in developed regions in 2022 [18], and in Japan, the number of households owning smartphones reached 86.8% in 2020 [19]. Given this, and that nearly 0.5 billion people use health-related apps worldwide already, mHealth is likely to be a key addition to standard diabetes treatment [16].

Many diabetes-related health care apps have been introduced in Japan, but none are classified as medical devices, meaning they cannot be prescribed and can only be adopted by patients independently. In the United States, the app BlueStar (WellDoc Communications Inc) was approved by the Food and Drug Administration as a class II medical device in 2010 [20,21], becoming the first app in the world that could be prescribed by physicians. BlueStar provides personalized feedback in response to patient data per US diabetes treatment guidelines. Although the lifestyle of diabetes patients in Japan may differ from that of US patients, a similar app may be effective in combatting the disease in Japan, and trials in Japan have yielded promising outcomes [22,23]. The acceptance of mHealth among Japanese patients with lifestyle diseases is relatively high [24].

In this trial, we investigated the efficacy of Save Medical Corporation (SMC)–01, an app developed by Save Medical Inc designed to assist the self-management of T2DM by providing feedback according to the Japan Diabetes Society's 2019 diabetes care guidelines and 2018-2019 diabetes treatment guide [25]. The main objective of our trial was to assess whether glycemic control could be improved through the usage of SMC-01 among patients with T2DM. We also studied whether any safety issues arose while using the app. We hypothesized that using the app would be beneficial to patients by helping them improve their glycemic control and lifestyle modification adherence.

Methods

Participant Inclusion Criteria and Recruitment

Patients were recruited from 4 medical institutions across Japan with significant experience in conducting clinical trials (Textbox 1). Recruitment was conducted by attending physicians during the patients' regular consultations at outpatient clinics. Before patients enrolled in the study, the study's supervising physician presented the review board–approved consent form to each patient and explained the contents, which included information about diabetes, how the app functions, how to operate the app, data collection, risks and benefits of participating, and compensation. Once the study was explained and it was determined that patients understood the explanation, patients were requested to participate in the study, and consent given by the patient's own free will was collected in writing via the consent form.

Textbox 1. Participating institutions.

- The Institute of Medical Science, Asahi Life Foundation
- Tokyo-eki Center Building Clinic
- Fukuwa Clinic
- Tokyo Center Clinic

XSL•FO
RenderX

Patients who met the inclusion criteria (Textbox 2) and did not meet the exclusion criteria (Textbox 3) were eligible for the study. To focus the study on patients most likely to benefit and most suitable for participation, the criteria selected patients who have inadequate glycemic control but have not recently changed their blood glucose management. Minors and in-patient participants were not eligible due to concerns of ethics and safety. Participants were also limited to those who own a smartphone. Patients who might expect changes in their blood glucose levels during the study, such as women who are already pregnant or who are planning on becoming pregnant, were not eligible. The study focused on the population with T2DM, excluding patients with type 1 diabetes mellitus due to differences in the pathophysiology between the 2 diseases.

Following obtaining consent and a subsequent 2-week observational period, patients were screened against the criteria, and eligible patients were enrolled in the study. The 24-week study began with a 12-week treatment period that included consultations every 4 weeks, followed by a 12-week treatment sustainment period that included 1 consultation. Patients were randomized into either the intervention group or the control group in a 1:1 ratio by block randomization (block size of 4) with baseline hemoglobin A_{1c} (HbA_{1c}; less than 8% or 8% or greater) and hypoglycemic medication use (yes or no) as stratification factors. Patients were assigned an ID by the registration center for data tracking.

Textbox 2. Inclusion criteria of the study.

- Outpatients who were aged 20 years or older when consent was obtained, regardless of sex.
- Patients who give written consent during visit 1.
- Patients who have been diagnosed with type 2 diabetes mellitus for at least 12 weeks (84 days) at visit 1.
- Patients who were currently using diet therapy or exercise therapy only, or who used diet therapy or exercise therapy in addition to diabetes medication and have had no change in their treatment for 12 weeks (84 days) or longer at the time of visit 1.
- Patients whose hemoglobin A_{1c} was 7% or greater and 9% or lower at the time of visit 1.
- Patients who had a smartphone and had been using it for 12 weeks (84 days) or longer at the time of visit 1.
- Patients who could enter information into the device being studied without difficulty as determined by the study's supervising physician.

XSL•FO
RenderX

Textbox 3. Exclusion criteria of the study.

- Patients who were planning on conceiving, were currently pregnant, or who were nursing.
- Patients who are premenopausal and who tested positive on a urine pregnancy test.
- Patients who had been diagnosed with type 1 diabetes mellitus.
- Patients who had been diagnosed with secondary diabetes mellitus.
- Patients who received an insulin injection within 12 weeks (84 days) of visit 1.
- Patients who required the assistance of a 3rd party in treatment of hypoglycemia in the past.
- Patients who had been diagnosed with and were currently undergoing treatment for proliferative diabetic retinopathy.
- Patients who had developed cardiovascular disease within 12 weeks (84 days) of visit 1.
- Patients with decompensated heart failure.
- Patients with severe liver impairment (alanine transaminase, as measured at the central testing location during visit 1) greater than 3 times the standard maximum value.
- Patients with kidney disease (estimated glomerular filtration rate less than 45 mL/min/1.73m^2 or urine albumin of 300 mg/gCre or higher, both as measured at the central testing location during visit 1).
- Patients with a chronic illness that required treatment with continuous pharmacotherapy (oral, injected, or inhaled) such as corticosteroids, immunosuppressants, or loop diuretics.
- Patients with malignancy (less than 5 years without recurrence) or a communicable disease (sepsis).
- Patients who planned to be admitted to a hospital or undergo surgery during the study period.
- Patients with drug addiction, alcohol use disorder, or an unstable psychological illness.
- Patients who were limited in their physical activity due to a medical condition other than diabetes.
- Patients whose hemoglobin A_{1c} value differed by more than 1% between visit 1 (measured at the hospital) and when measured at some point 4 to 10 weeks before visit 1.
- Patients who planned to, were currently, or had (within the past 12 weeks; 84 days) been enrolled in a different clinical trial for a pharmacologic agent or medical device, or a postmarketing clinical trial.
- Patients who, within the past 12 weeks (84 days), had used a mobile app on their smartphone or have been self-monitoring their blood glucose for diabetes self-management, and whose evaluation of the efficacy of the device in this study might be influenced by said experience, as determined by the study's supervising physician.
- Patients who plan to switch their smartphone within 24 weeks of visit 2.
- Medical staff involved in this experiment and their close relatives.
- Save Medical employees or employees of companies that had been hired to complete tasks related to this experiment.
- Patients who could not comply with or refused to comply with conditions put forth by study staff, such as rules for visiting the hospital, or rules for taking medication.
- Patients determined by this study's supervising physician to be unfit for involvement in this study for any other reason.

Study Design

This study was a multi-institutional, prospective, and unblinded 2-armed randomized controlled trial carried out across 4 medical institutions in Tokyo, Japan. We chose to conduct an unblinded trial because we predicted that the use of a placebo app with minimal function would lead to the control group unblinding themselves.

We provided both groups with Bluetooth-enabled blood pressure cuffs (A&D, UC-352BLE) and scales (A&D, UA-651BLE). Patients in the intervention group were instructed to install the SMC-01 app onto their smartphones and use it to track diet, medication adherence, exercise, body weight, and blood pressure. Recordings of blood pressure and weight were collected automatically through the Bluetooth-enabled devices but could also be input manually. The rest of the data was input manually.

Patients in the control group recorded information about their diabetes self-management in a journal. They were asked to record which meals they ate (breakfast, lunch, dinner, or snack), whether they took their diabetes medication, whether they exercised and what type of exercise (aerobic or resistance training), their weight in kilograms, and their blood pressure along with when they measured it (morning, noon, or evening). Unlike the SMC-01 group, the control group did not record the content of meals, and only recorded whether they ate a meal or not. The format of a journal entry is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Entry in patient journal (English translation).

Date	Meals	Taking medication	Exercise	Weight and blood pressure
/ ( )	☐ Break-fast ☐ Lunch ☐ Dinner ☐ Snack	☐ Taken	☐ Aerobic exercise ☐ Resistance or weight training	☐ Weight (_____kg) ☐ Blood pressure AM (____/____mm Hg) Noon (____/____mm Hg) PM (____/____mm Hg)

SMC-01 Design

SMC-01 (Figure 2) was designed to assist the treatment and management of T2DM, complementing, not replacing, physicians. SMC-01 breaks down goals set by physicians into smaller and specific actions and sends them to patients as daily feedback and challenges. Through this, it aims to increase the likelihood of long-term behavioral change by making goals more achievable and by addressing the issue of patients not knowing how to implement their physician's advice.

SMC-01 consists of 4 main functions: facilitating data input, providing feedback in response to the data input, sending reminders, and setting daily challenges.

Patients receive feedback per Japanese diabetes treatment guidelines based on their daily behavior. This is implemented using Sketto, a chatbot within SMC-01 that sends messages to the patient based on patient-entered data. The content of these messages depends largely on whether the data the patient enters is within normal limits according to Japanese diabetes treatment guidelines. If the patient's data are within normal limits, Sketto praises the patient for their hard work and encourages them to continue their diabetes self-management as well as their data logging. On the other hand, if the patient's data are outside of normal limits, Sketto warns the patient about the possible implications of the measurement, such as poor health outcomes, and encourages the patient to seek the advice of a health professional. For example, when the patient records their weight, Sketto sends a message providing positive feedback on healthy lifestyle choices or caution on unhealthy lifestyle choices depending on the patient's BMI, which SMC-01 calculates from the patient's height registered at the initial set-up. When the patient records his or her meal data, Sketto sends a message depending on the patient's reported satiety level and the time of day. For medication adherence, Sketto sends reminders if patients forget to record that they took their medication, and praises patients for recording medication adherence several days in a row. Examples of some of these messages are shown in Figure 3.

Figure 2. Overview of SMC-01. SMC: Save Medical Corporation.

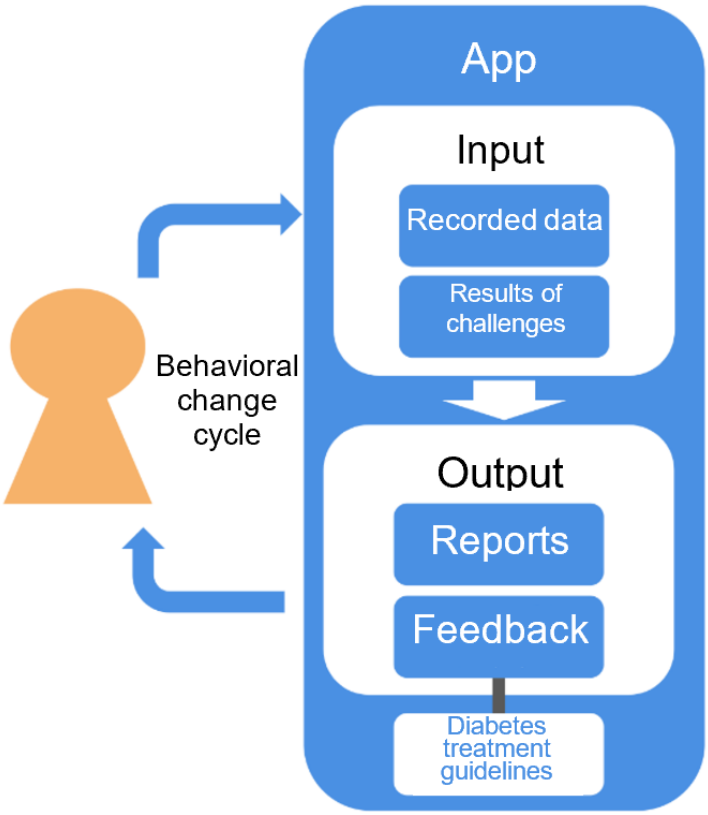

XSL•FO
RenderX

Figure 3. Example messages and challenges sent by Sketto (English translation).

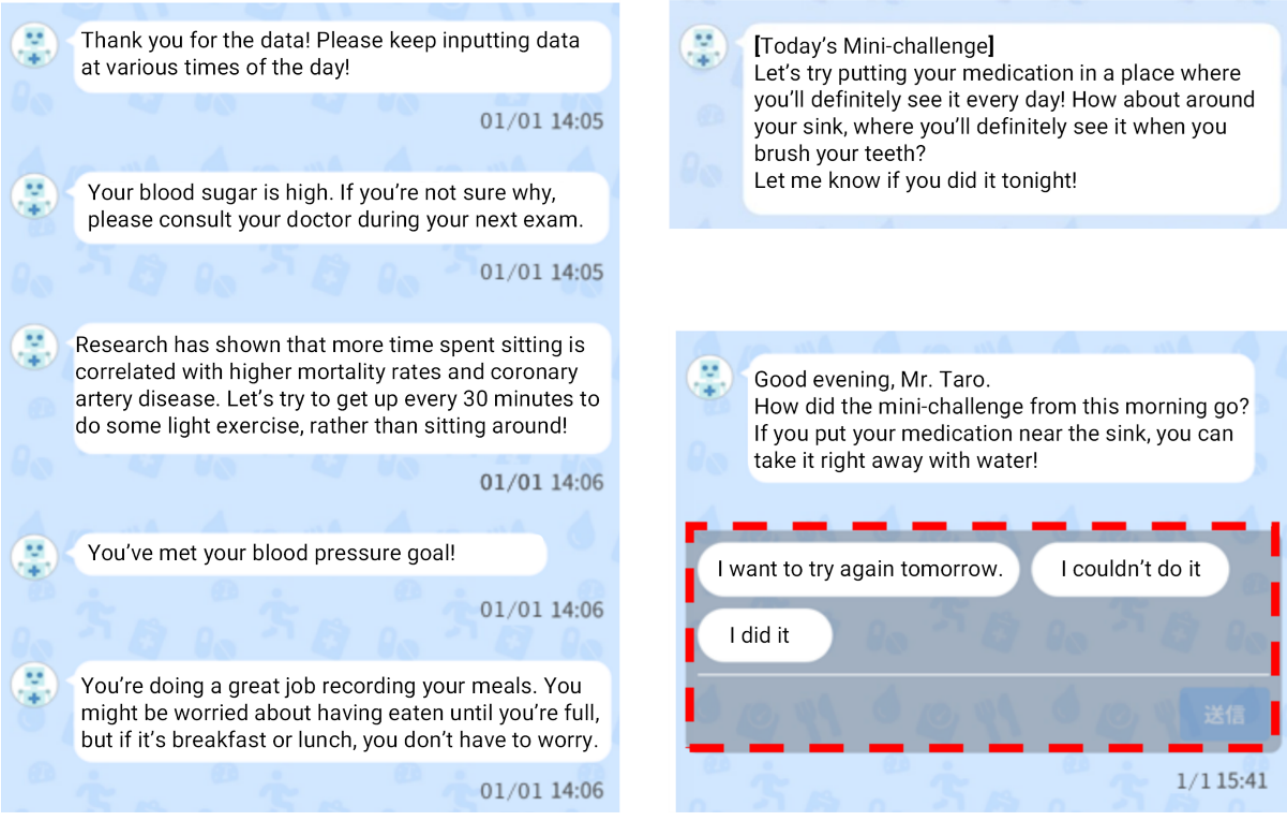

Sketto also sends patients health-related articles to increase patients' knowledge about diabetes and health in general. Unlike the feedback described above, these articles are not personalized based on patient data. Patients can also review their records in the form of reports. These reports show the patient's self-management over time, such as how many days the patient has exercised, their change in weight, and the days they took their medication. The supervising physicians of the study could review the patients' self-management by viewing the app's screen. Review of data input into the app by the study's supervising physicians could be accomplished at any time, but the review was mainly carried out during visits 3 through 6.

Every morning, Sketto sends patients mini-challenges within the app that are designed to augment the patients' assigned diet, medication adherence, and exercise-based behavior modifications. These challenges begin at the lowest level, and, depending on whether the patient succeeds, a similar challenge that is slightly harder or easier may be assigned the following day. Otherwise, a mini-challenge from another category will be assigned. The patient can also choose to repeat the same challenge on the following day. Examples of mini-challenges can be found in Table S1 in Multimedia Appendix 1.

Data Collection

Initial measurements were made at visit 1 at the beginning of the observation period. The study (Figure 4) included 6 in-person visits, a 2-week observation period, a 12-week intervention period, and a 12-week intervention sustainment period. If patients withdrew from the study at any point, observation and testing were carried out when the patient came to the health care facility to withdraw from the study. A detailed timeline of the study can be found in Table S2 in Multimedia Appendix 1.

From visit 1 until visit 6 or until measurement of efficacy outcomes had been completed, physicians in the study were prohibited from providing specific calorie intake and exercise amount goals to patients. Nonspecific advice, such as general directions to eat more slowly or eat less salt, was permitted. These limitations were set to ensure that changes in exercise amount and calorie intake could be attributed to the app and not instructions from physicians.

XSL•FO
RenderX

Figure 4. Overview of the study timeline. SMC: Save Medical Corporation.

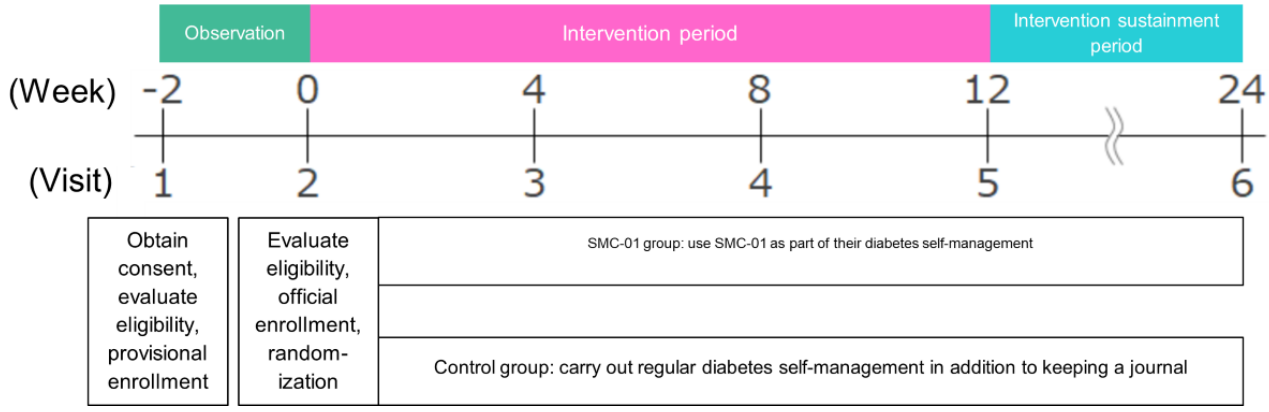

Outcome Measures

The primary outcome was the between-arm difference in change in HbA_{1c} values between the baseline and the end of the intervention period.

The study evaluated the following secondary measurements:

- HbA_{1c} value: between-arm difference in the change from baseline to the end of the intervention sustainment period
- Fasting blood glucose: between-arm difference in the change from baseline to the end of the intervention period and to the end of the intervention sustainment period.
- Fasting insulin: between-arm difference in the change from baseline to the end of the intervention period and to the end of the intervention sustainment period.
- Fasting intact-proinsulin or insulin ratio: between-arm difference in the change from baseline to the end of the intervention period and to the end of the intervention sustainment period.
- Self-maintenance adherence rate (diet, exercise, and medication adherence; limited to patients taking diabetes medication)

Self-maintenance adherence rates were calculated as follows:

Ethical Considerations

The study was registered in the Japan Registry of Clinical Trials (jRCT2032200033) on May 17, 2020. It was approved by the research ethics committee of The Institute for Adult Diseases, Asahi Life Foundation (179-8). We conducted the study per the Ministerial Ordinance on Standards for Conducting Clinical Trials of Medical Devices (The Ministerial Ordinance on Good Clinical Practice). The study has been reported as per the eHealth CONSORT (Consolidated Standards of Reporting Trials) statement. We obtained written informed consent from all participants before the study, and study data have been anonymized. Participants were compensated JP ¥10,000 (roughly US $64) for their time.

Statistical Analysis

We set the mean difference in the change in HbA_{1c} from the baseline that we were designing the study to detect to 0.5, as this is generally considered a clinically significant change in HbA_{1c} [26,27]. The SD common to both groups was set to 1 based on values from previous studies [23,28,29], and the 2-sided significance level was 5% with 90% power. The resulting required sample size is 86 for each group. Taking into account the likelihood of dropouts, the target sample size for each group was set at 100 cases (200 cases in total).

Per the intention-to-treat (ITT) analysis principle, all randomly assigned patients were included in the full analysis set (FAS). In the case that a patient was excluded from the analysis, justification for the exclusion was documented. Patients in the FAS who met all of the following criteria that did not violate the selection exclusion criteria and protocol were included in the per protocol set (PPS):

- Met the inclusion criteria and did not meet the exclusion criteria
- HbA_{1c} value both at baseline and at the end of the intervention period was not missing
- No use of any of the prohibited medications, used conditional medications according to the study guidelines, and did not carry out any other prohibited treatments
- Self-management adherence rate above 80%
- Did not commit any serious study protocol violations by the end of the intervention period

All patients that participated in the intervention period of the study were included in the safety analysis set. Demographic data and baseline values were compared between groups using the Fisher exact test for discrete values and a 2-tailed *t* test for continuous values.

As the primary analysis, an analysis of covariance was performed on the change in HbA_{1c} from baseline to the end of the intervention period (wk 12) in the FAS, with baseline values as the covariate, and the between-group differences and their 95% CIs were calculated. As a sensitivity analysis, the same analysis was performed for the PPS and for patients with HbA_{1c} measured at all time points (wk 0, 4, 8, 12, and 24). In addition, the FAS was analyzed with a linear mixed effects model that included all time points. The objective variable was the change in HbA_{1c} from baseline, the random effect was the patient, and the fixed effects were group, time point, group × time point interaction, and baseline HbA_{1c}. The amount of change for each group, the between-group difference in change, and their 95% CIs at each time point were calculated.

XSL•FO
RenderX

As a secondary analysis, an analysis of covariance was performed in the FAS on the change from baseline in HbA_{1c}, fasting blood glucose, fasting insulin, and fasting intact-proinsulin or insulin ratio at the end of the intervention period (wk 12) and at the end of the intervention sustainment period (wk 24) with baseline values as covariates. In addition, summary statistics of change and measurements were calculated for each time point (wk 0, 4, 8, 12, and 24) for each group. The percentage of HbA_{1c} less than 7% at each time point (wk 4, 8, 12, and 24) for each group was also calculated. However, the denominator was the FAS population excluding patients with HbA_{1c} less than 7% at baseline. Self-maintenance adherence rates for diet, exercise, and medication were assessed in the following time periods and compared between groups by a 2-tailed *t* test: week 0 to week 4, week 4 to week 8, week 8 to week 12, week 12 to week 24, week 0 to week 12, week 0 to week 24.

In addition, the same analyses on HbA_{1c} levels were performed for the following subgroups as in the primary and secondary analyses.

- Sex: female or male
- Age: younger than 65 years or 65 years or older
- Baseline HbA_{1c}: less than 8% or 8% or greater
- Baseline BMI: less than 25 kg/m^2 or 25 kg/m^2 or greater
- Hypoglycemic medication usage at week 2: yes or no
- Self-management adherence rate during the intervention period: less than 80% or 80% or greater

For analysis of safety for the safety analysis set, adverse events were classified for each group using system organ classes and preferred terms of the Japanese version of the Medical Dictionary for Regulatory Activities (version 23.0) [18], and the number of patients and occurrence rate for each system organ class or preferred term categories were tabulated. Additionally, the occurrence rates were compared between groups by Fisher exact test. For bugs in the SMC-01 group, the number of patients with bugs, occurrence rate, and the number of cases were tabulated. For each clinical test item and vital sign, levels and the difference from the beginning of the intervention period were calculated at each time point in each group as summary statistics.

All statistical tests were 2-tailed, with a significance level of 5%. CIs were calculated as 2-tailed 95% CIs. SAS (version 9.4; SAS Institute Inc) was used for statistical analysis.

Results

Overview

This study was conducted from May 11, 2020, to April 5, 2021. In total, 210 participants were randomly assigned between groups, yielding 107 (intervention) and 103 (control) at baseline, 106 (intervention) and 102 (control) at the end of 12 weeks, and 105 (intervention) and 101 (control) at the end of 24 weeks (Figure 5). Patients were well matched in baseline characteristics (Table 1), although urine albumin or creatinine ratio was significantly higher in the intervention group than in the control group (42.50 mg/g of Cr in the intervention group versus 20.82 mg/g of Cr in the control group, *P*=.01).

XSL•FO
RenderX

Figure 5. CONSORT flow diagram. CONSORT: Consolidated Standards of Reporting Trials; SMC: Save Medical Corporation.

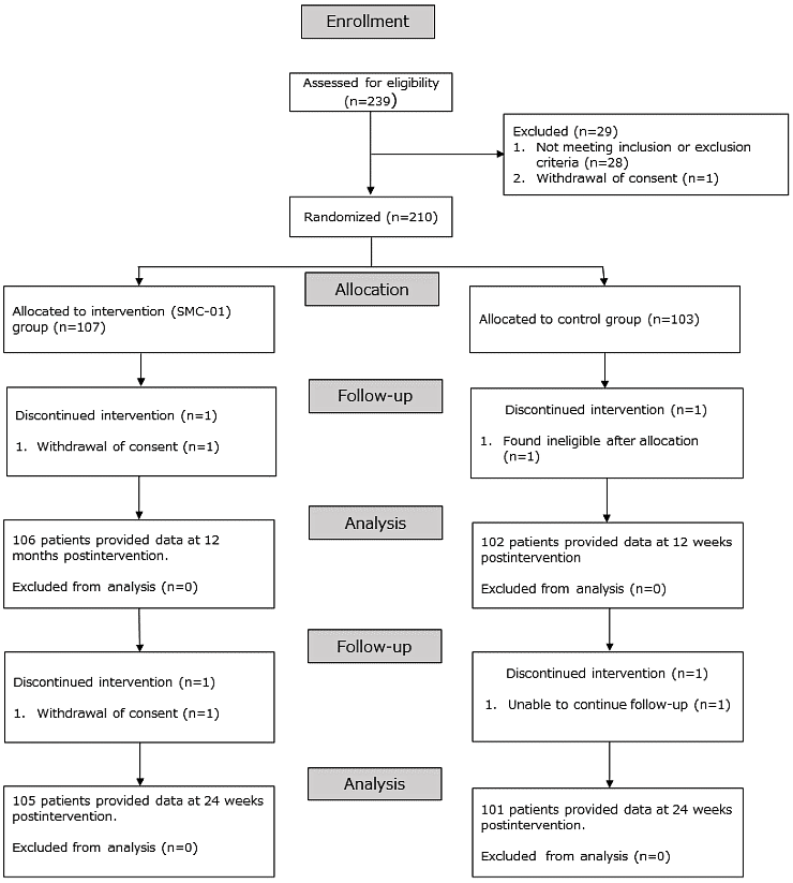

XSL•FO
RenderX

Table 1. Baseline characteristics[a].

Characteristics	SMC[b]-01 group (n=107)	Control group (n=103)	*P* value[c]
Age (years), mean (SD)	57.6 (8.35)	58 (8.79)	.75
<65, n (%)	83 (77.6)	78 (75.7)	.87
≥65, n (%)	24 (22.4)	25 (24.3)	
Sex, n (%)			.60
Female	22 (20.6)	18 (17.5)	
Male	85 (79.4)	85 (82.5)	
Disease duration (years), mean (SD)	10.93 (5.908)	10.10 (5.334)	.29
Height (cm), mean (SD)	167.04 (8.207)	167.62 (6.915)	.58
Weight (kg), mean (SD)	73.80 (14.668)	73.81 (14.163)	.99
BMI (kg/m^2), mean (SD)	26.34 (4.283)	26.18 (4.197)	.78
<25, n (%)	45 (42.1)	47 (45.6)	.68
≥25, n (%)	62 (57.9)	56 (54.4)	
Waist circumference (cm), mean (SD)	92.61 (11.255)	92.16 (10.207)	.77
Insulin dependence or independence, n (%)			
Insulin dependent	0 (0)	0 (0)	
Insulin independent	107 (100)	103 (100)	
C-peptide (ng/mL), mean (SD)	2 (0.799)	1.97 (0.829)	.78
Urine albumin or creatinine ratio (mg/g of Cr), mean (SD)	42.50 (84.277)	20.82 (31.552)	.01
Baseline HbA_{1c}, mean (SD)	7.74 (0.554)	7.67 (0.542)	.37
<8, n (%)	71 (66.4)	74 (71.8)	.46
≥8, n (%)	36 (33.6)	29 (28.2)	
Hypoglycemic medication use, n (%)			.72
Do not use	20 (18.7)	17 (16.5)	
Use	87 (81.3)	86 (83.5)	

[a]Baseline values were taken at the beginning of the intervention period (visit 2). If there was no measurement from the beginning of the intervention period, the latest value before the start of the intervention period was used.

[b]SMC: Save Medical Corporation.

[c]*P* value: Discrete variables were analyzed using the Fisher exact test. Continuous variables were analyzed using an unpaired 2-tailed *t* test.

Self-Management Adherence Rates

Overall adherence rates were high during the intervention period (Table 2), at 82.28% in the SMC-01 group and 79.35% in the control group (*P*=.20). Average diet adherence rates were significantly higher in the control group than in the SMC-01 group at 4 weeks (96.39%, SD 10.67% and 99.01%, SD 2.42%, respectively; *P*=.02) and week 24 (91.57%, SD 22.44% and 97.63%, SD 10.37%, respectively; *P*=.01). Medication adherence rates were similar in both groups, with no statistically significant differences found at any time point (all *P*>.05). Average exercise self-management adherence rates were significantly higher in the SMC-01 group than in the control group, with 60.79% (SD 37.22%) and 49.23% (SD 37.92%; *P*=.03) at 4 weeks, 60.32% (37.24%) and 49.33% (38.52%; *P*=.04) at 8 weeks, and 59.29% (35.95%) and 47.34% (37.01%; *P*=.03) at the end of the intervention period.

XSL•FO
RenderX

Table 2. Self-management adherence rate[a].

Adherence rate and period	SMC[b]-01 group	Control group	*P* value[c]
Overall			
Wk 0-12, mean (SD)	82.28 (16.56)	79.35 (16.27)	.20
≤80%, n (%)	40 (37.4)	50 (48.5)	.13
>80%, n (%)	67 (62.6)	53 (51.5)	
Meal, mean (SD)			
Wk 0-4	96.39 (10.67)	99.01 (2.42)	.02
Wk 4-8	96.63 (12.28)	98.87 (2.63)	.07
Wk 8-12	95.67 (12.96)	97.94 (10.03)	.16
Wk 12-24	91.57 (22.44)	97.63 (10.37)	.01
Wk 0-12	95.75 (11.18)	97.79 (9.70)	.16
Wk 0-24	93.16 (14.98)	97.27 (9.78)	.02
Medication, mean (SD)			
Wk 0-4	94.15 (15.23)	96.90 (11.05)	.18
Wk 4-8	95.22 (14.38)	96.43 (11.44)	.55
Wk 8-12	94.59 (12.35)	95.57 (13.65)	.62
Wk 12-24	94.13 (16.48)	96.27 (11.97)	.33
Wk 0-12	94.05 (14.67)	95.46 (14.30)	.52
Wk 0-24	93.68 (14.69)	95.53 (12.69)	.38
Exercise, mean (SD)			
Wk 0-4	60.79 (37.22)	49.23 (37.92)	.03
Wk 4-8	60.32 (37.24)	49.33 (38.52)	.04
Wk 8-12	57.88 (37.49)	47.97 (37.45)	.06
Wk 12-24	55.27 (38.55)	46.02 (38.87)	.09
Wk 0-12	59.29 (35.95)	48.34 (36.89)	.03
Wk 0-24	57.05 (35.80)	47.34 (37.01)	.06

[a]The overall self-management adherence rate was defined as the average of diet, exercise, and medication (only for those taking medication) adherence rates.

[b]SMC: Save Medical Corporation.

[c]*P* value: Discrete variables were analyzed using the Fisher exact test. Continuous variables were analyzed using an unpaired 2-tailed *t* test.

Efficacy of SMC-01

Based on the analysis of covariance (Table 3), the change in HbA_{1c} from baseline to the end of the intervention period (12 wk) was –0.05% (95% CI –0.14% to 0.04%) in the SMC-01 group and 0.06% (95% CI –0.04% to 0.15%) in the control group, with a between-group difference (SMC-01 group – control group) of –0.11% (95% CI –0.24% to 0.03%, *P*=.11). The difference is not statistically significant.

Secondary analysis of the change in HbA_{1c} from baseline to the end of the intervention sustainment period (24 wk) revealed a between-group difference (SMC-01 group – control group) of –0.16% (95% CI –0.36% to 0.03%; *P*=.11). The difference is not statistically significant.

Secondary analysis of the change in fasting blood glucose from baseline to the end of the intervention period (12 wk) yielded a between-group difference (SMC-01 group – control group) of –4 (95% CI –10.6 to 2.6; *P*=.23). From baseline to the end of the intervention sustainment period (24 wk) the between-group difference (SMC-01 group – control group) was –4.8 (95% CI –12.6 to 3; *P*=.23). The differences are not statistically significant.

Secondary analysis of the change in fasting insulin from baseline to the end of the intervention period (12 wk) yielded a between-group difference (SMC-01 group – control group) of 0.24 (95% CI –0.64 to 1.13; *P*=.59). From baseline to the end of the intervention sustainment period (24 wk) the between-group difference (SMC-01 group – control group) was 0.19 (95% CI –0.47 to 0.85; *P*=.56). The differences are not statistically significant.

Secondary analysis of the change in fasting intact-proinsulin or insulin ratio from baseline to the end of the intervention period

XSL•FO
RenderX

(12 wk) yielded a between-group difference (SMC-01 group – control group) of –0.01 (95% CI –0.04 to 0.02; *P*=.44). From baseline to the end of the intervention sustainment period (24 wk) the between-group difference (SMC-01 group – control group) was –0.01 (95% CI –0.03 to 0.01; *P*=.42). The differences are not statistically significant.

In the sensitivity analysis of the PPS and the complete set, ANCOVA of the between-group difference in the average change in HbA_{1c} from baseline to the end of the intervention period (wk 12) was not statistically significant (*P*=.14 and *P*=.11). Mixed model estimates of the change in HbA_{1c} from baseline in the FAS yielded a statistically significant difference at week 8 of –0.12 points (*P*=.04) and week 12 of –0.12 points (*P*=.04). The full sensitivity analysis results have been included in Multimedia Appendix 1.

In subgroup analysis of change in HbA_{1c} from baseline to the end of the intervention period (wk 12), statistically significant differences were found in the aged younger than 65 years subgroup (estimated values in the SMC-01 group and control group were –0.07% and 0.08%, respectively, *P*=.04, same order in subsequent listings) and the baseline HbA_{1c} less than 8% subgroup (–0.04% and 0.11%; *P*=.04). Subgroup analysis of change in HbA_{1c} from baseline to the end of the intervention sustainment period (wk 24) revealed no statistically significant difference between the 2 groups at any time point. The full subgroup analysis results have been included in Multimedia Appendix 1.

Table 3. Summary of efficacy analyses[a,b,c,d].

	SMC[e]-01 group (n=107)[f]	Control group (n=103)[e]	Difference[e]	*P* value
HbA_{1c}[g] (% points)				
Wk 12	–0.05 (–0.14 to 0.04)	0.06 (–0.04 to 0.15)	–0.11 (–0.24 to 0.03)	.11
Wk 24	0.11 (–0.03 to 0.25)	0.28 (0.13 to 0.42)	–0.16 (–0.36 to 0.03)	.11
Fasting blood glucose (mg/dL)				
Wk 12	–3.2 (–7.8 to 1.4)	0.9 (–3.9 to 5.6)	–4.0 (–10.6 to 2.6)	.23
Wk 24	1.7 (–3.8 to 7.1)	6.5 (0.9 to 12.0)	–4.8 (–12.6 to 3.0)	.23
Fasting insulin (μU/mL)				
Wk 12	0.17 (–0.45 to 0.79)	–0.08 (–0.71 to 0.56)	0.24 (–0.64 to 1.13)	.59
Wk 24	–0.11 (–0.57 to 0.35)	–0.30 (–0.78 to 0.17)	0.19 (–0.47 to 0.85)	.56
Fasting intact-proinsulin or insulin ratio				
Wk 12	–0.01 (–0.03 to 0.01)	0.00 (–0.02 to 0.02)	–0.01 (–0.04 to 0.02)	.44
Wk 24	–0.02 (–0.03 to 0.00)	–0.01 (–0.02 to 0.01)	–0.01 (–0.03 to 0.01)	.42

[a]The response variable is the absolute change from baseline, with the baseline value as the covariate in the analysis of covariance.

[b]Baseline measurement was measured at the start of the intervention period (visit 2). If there was no measurement at the start of the intervention period, the last measurement before the start of the intervention period was used.

[c]If there was no measurement at the 12-week visit, the last measurement taken between the second day and the 92nd day of the study was used.

[d]If there was no measurement at the 24-week visit, the last measurement taken between the second day and the 175th day of the study was used.

[e]SMC: Save Medical Corporation.

[f]Data are expressed as mean absolute change from baseline and its 95% CI.

[g]HbA_{1c}: hemoglobin A_{1c}.

Safety of SMC-01

The results of the safety analysis are included in Multimedia Appendix 1. There were no adverse events or deaths caused by the device under study.

Discussion

Principal Findings

We did not find any statistically significant difference in outcomes between the SMC-01 and control groups. Further, HbA_{1c} values essentially did not change in either group. Patients who participate in clinical trials normally experience a change in health outcomes due to the Hawthorne effect and the placebo effect, regardless of the intervention. These effects in this study may have been small because this trial was conducted during the COVID-19 pandemic, when people were advised to stay home, possibly reducing changes in behavior such as exercise.

We can assess possible reasons for the failure of the intervention to change outcomes using behavioral change theory, specifically, the Theory of Planned Behavior [30]. When designing a lifestyle intervention, it is critical to view it as requiring a behavioral change, a perspective that was not used in the SMC-01 intervention. The first step is to clearly define the desired change in behavior. In this study, the desired change was a loose collection of behaviors associated with following physician guidance on a healthy lifestyle. This may have been too vague to be actionable compared to past successful interventions in

XSL•FO
RenderX

diet and exercise that defined concrete actions [31-33]. The next consideration is the baseline lifestyle of the patients. It is possible that patients were already following physician lifestyle instructions, and that there was not much need for further lifestyle intervention via SMC-01.

Intention is key in the Theory of Planned Behavior framework. The intervention sought to have patients make a behavior change large enough to produce a significant change in glycemic control. Behavioral change is difficult for most people and requires strong motivation, and motivation was not evaluated as one of the inclusion criteria of this trial. If a patient does not intend to change or perform a behavior, they are very unlikely to do so [30]. Other studies have had success screening for patients who are in the contemplation, preparation, or action stages of the transtheoretical model of behavioral change, increasing the likelihood of strong motivation [34-36]. A 2015 study showed that 92% of Japanese patients with T2DM met this criterion [37]. A different intervention would be appropriate for patients who are not motivated to change [38]. Our study did not specify any screening test for high intention.

Another concern is the lack of clear information about patient behavior. The study examined elements of patient behavior such as app usage but was highly focused on the ultimate health changes rather than behavior. We can assume that the desired behavioral change would have led to improved glycemic control, as has been shown in many other studies [31,32,39], but we do not have enough data to confirm whether their behavior truly changed. Future studies of lifestyle intervention should carefully investigate how to measure behavioral changes. Additionally, SMC-01 may not have been used as intended due to problems with the protocol and its implementation. There may have been insufficient intervention design features. The study found high levels of usage, and yet the intervention did not change glycemic control. In this study, usage was measured as the percentage of days for which data was input at least once. User engagement with other features was not measured. These usage measurements may not have been the appropriate indicators of effective usage of SMC-01. Studies have shown that simply recording data does not improve health care outcomes [40,41]. A review of 32 studies found that adherence rates in mHealth interventions were inconsistent [42]. Screen-time measurements may be used to determine how often patients accessed certain features, providing insight into patient engagement with feedback and whether the patient truly received the intervention.

Another concern is possible insufficient data entry. Objective health data are crucial to mHealth interventions, as self-reported data are often unreliable [43,44]. Automatic data collection objectively measures behavior and was used for the measurement of weight and blood pressure in this study. These are intermediate health changes on the way to changed glycemic control, not behaviors. Future studies should apply automatic data collection of behavioral change. Finally, there may have been insufficient behavioral change information and directions suggested by SMC-01. Closed feedback loops form the backbone of mHealth interventions, but this feedback must be personalized and actionable [45-47].

We do not have sufficient data from this study to determine which, if any, of these possible issues contributed to the outcome. In future studies, addressing these possible issues during the study and app design phase is recommended. In future app design, a theoretical framework of behavioral change can be used to guide the implementation of features. It has been shown that basing interventions on behavior change theory improves health outcomes [47].

This study has some limitations. It was conducted in a Japanese population, and there are known differences between Japanese and other populations about lifestyle and the pathophysiology of T2DM [48,49]. This study was limited to patients who owned smartphones, which may have led to biases related to digital literacy and socioeconomic status. Social desirability bias also likely influenced results due to a lack of blinding.

To our knowledge, this study was the first trial in Japan that sought to obtain official approval for a smartphone app for T2DM as a medical device and will serve as a foundation for the design of future apps and clinical trials seeking medical device registration. mHealth devices for the management of chronic diseases will continue to improve, and future studies in Japan should continue to focus on gaining official approval. Information from this trial, along with the discussion of possible shortfalls, should help improve future efforts.

Conclusions

No statistically significant between-arm difference was found in the change in HbA_{1c} from baseline to the end of the 12-week intervention period. Future studies should improve clarity regarding the theory of the intervention, improve the methods used, and ensure the measurement of behavior and use throughout the study.

Acknowledgments

We would like to acknowledge the valuable contribution of the research consortium, our national and international collaborators, health care providers involved in the study, and the patients who participated in the study. The funders had no role in designing the study; collecting, managing, or analyzing the data; interpreting the results; writing the paper; or deciding to submit the paper for publication. Finally, the authors would like to thank Mr Daniel Lane, MAS, MBA for providing critical appraisals. The study was supported by Sumitomo Pharma Co, Ltd.

XSL•FO
RenderX

Conflicts of Interest

Save Medical Corporation (SMC), Inc planned and conducted the study, and SN is an employee of the company. KW is an advisor to the company. YO participated in this clinical trial as a principal investigator, received consulting fees from SMC, and received speaker fees from Sumitomo Pharma Co, Ltd, Eli Lilly Japan KK, and Novo Nordisk Pharma Ltd.

Multimedia Appendix 1
Tables depicting examples of mini-challenges, study schedule, summary of sensitivity analyses for HbA_{1c}, summary of subgroup analyses for HbA_{1c}, all adverse events, and breakdown of bugs and sections depicting PPS definition and safety analysis. Hb: hemoglobin; PPS: per protocol set.
[DOCX File , 45 KB - jmir_v26i1e53740_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 9349 KB - jmir_v26i1e53740_app2.pdf]

References

1. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of type 2 diabetes—global burden of disease and forecasted trends. JEGH 2020;10(1):107. [doi: 10.2991/jegh.k.191028.001] [Medline: 32175717]
2. Herman WH. The global burden of diabetes: an overview. In: Dagogo-Jack S, editor. Diabetes Mellitus in Developing Countries and Underserved Communities. US: Springer International Publishing; 2017:1-5.
3. Ministry of Health, Labour and Welfare. Diabetes Mellitus. URL: https://www.mhlw.go.jp/www1/topics/kenko21_11/b7.html [accessed 2023-08-13]
4. Wing RR, Goldstein MG, Acton KJ, Birch LL, Jakicic JM, Sallis JF, et al. Behavioral science research in diabetes: lifestyle changes related to obesity, eating behavior, and physical activity. Diabetes Care 2001;24(1):117-123. [doi: 10.2337/diacare.24.1.117] [Medline: 11194216]
5. Sami W, Ansari T, Butt NS, Hamid MRA. Effect of diet on type 2 diabetes mellitus: a review. Int J Health Sci (Qassim) 2017;11(2):65-71 [FREE Full text] [Medline: 28539866]
6. Chen L, Pei J, Kuang J, Chen H, Chen Z, Li Z, et al. Effect of lifestyle intervention in patients with type 2 diabetes: a meta-analysis. Metabolism 2015;64(2):338-347 [FREE Full text] [doi: 10.1016/j.metabol.2014.10.018] [Medline: 25467842]
7. Johansen MY, MacDonald CS, Hansen KB, Karstoft K, Christensen R, Pedersen M, et al. Effect of an intensive lifestyle intervention on glycemic control in patients with type 2 diabetes: a randomized clinical trial. JAMA 2017;318(7):637-646 [FREE Full text] [doi: 10.1001/jama.2017.10169] [Medline: 28810024]
8. Kitsiou S, Paré G, Jaana M, Gerber B. Effectiveness of mHealth interventions for patients with diabetes: an overview of systematic reviews. PLoS One 2017;12(3):e0173160 [FREE Full text] [doi: 10.1371/journal.pone.0173160] [Medline: 28249025]
9. Burgess E, Hassmén P, Pumpa KL. Determinants of adherence to lifestyle intervention in adults with obesity: a systematic review. Clin Obes 2017;7(3):123-135. [doi: 10.1111/cob.12183] [Medline: 28296261]
10. White ND, Bautista V, Lenz T, Cosimano A. Using the SMART-EST goals in lifestyle medicine prescription. Am J Lifestyle Med 2020;14(3):271-273 [FREE Full text] [doi: 10.1177/1559827620905775] [Medline: 32477026]
11. Noordman J, van der Weijden T, van Dulmen S. Communication-related behavior change techniques used in face-to-face lifestyle interventions in primary care: a systematic review of the literature. Patient Educ Couns 2012;89(2):227-244. [doi: 10.1016/j.pec.2012.07.006] [Medline: 22878028]
12. Selçuk-Tosun A, Zincir H. The effect of a transtheoretical model-based motivational interview on self-efficacy, metabolic control, and health behaviour in adults with type 2 diabetes mellitus: a randomized controlled trial. Int J Nurs Pract 2019;25(4):e12742. [doi: 10.1111/ijn.12742] [Medline: 31090161]
13. Naroee H, Rakhshkhorshid M, Shakiba M, Navidian A. The effect of motivational interviewing on self-efficacy and continuation of exclusive breastfeeding rates: a quasi-experimental study. Breastfeed Med 2020;15(8):522-527 [FREE Full text] [doi: 10.1089/bfm.2019.0252]
14. Nundy S, Dick JJ, Chou C, Nocon RS, Chin MH, Peek ME. Mobile phone diabetes project led to improved glycemic control and net savings for Chicago plan participants. Health Aff (Millwood) 2014;33(2):265-272 [FREE Full text] [doi: 10.1377/hlthaff.2013.0589] [Medline: 24493770]
15. Fortmann AL, Gallo LC, Garcia MI, Taleb M, Euyoque JA, Clark T, et al. Dulce digital: an mHealth SMS-based intervention improves glycemic control in hispanics with type 2 diabetes. Diabetes Care 2017;40(10):1349-1355 [FREE Full text] [doi: 10.2337/dc17-0230] [Medline: 28600309]
16. Fleming GA, Petrie JR, Bergenstal RM, Holl RW, Peters AL, Heinemann L. Diabetes digital app technology: benefits, challenges, and recommendations. A consensus report by the European association for the study of diabetes (EASD) and the American Diabetes Association (ADA) diabetes technology working group. Diabetes Care 2020;43(1):250-260 [FREE Full text] [doi: 10.2337/dci19-0062] [Medline: 31806649]

XSL•FO
RenderX

17. Orsama AL, Lähteenmäki J, Harno K, Kulju M, Wintergerst E, Schachner H, et al. Active assistance technology reduces glycosylated hemoglobin and weight in individuals with type 2 diabetes: results of a theory-based randomized trial. Diabetes Technol Ther 2013;15(8):662-669. [doi: 10.1089/dia.2013.0056] [Medline: 23844570]
18. Mobile phone ownership. Facts and Figures 2022. URL: https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-mobile-phone-ownership [accessed 2023-08-02]
19. Information and communication paper 2021 version. The current status of digital activities. Ministry of Internal Affairs and Communications. URL: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/html/nd111100.html# [accessed 2023-08-02]
20. Shan R, Sarkar S, Martin SS. Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. Diabetologia 2019;62(6):877-887. [doi: 10.1007/s00125-019-4864-7] [Medline: 30963188]
21. Neubeck L, Lowres N, Benjamin EJ, Freedman SB, Coorey G, Redfern J. The mobile revolution—using smartphone apps to prevent cardiovascular disease. Nat Rev Cardiol 2015;12(6):350-360. [doi: 10.1038/nrcardio.2015.34] [Medline: 25801714]
22. Treatment guidance using a self-management ICT system for type 2 diabetes mellitus patients. URL: https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000126725.pdf [accessed 2024-08-10]
23. Waki K, Fujita H, Uchimura Y, Omae K, Aramaki E, Kato S, et al. DialBetics: a novel smartphone-based self-management support system for type 2 diabetes patients. J Diabetes Sci Technol 2014;8(2):209-215 [FREE Full text] [doi: 10.1177/1932296814526495] [Medline: 24876569]
24. Yamaguchi S, Waki K, Tomizawa N, Waki H, Nannya Y, Nangaku M, et al. Previous dropout from diabetic care as a predictor of patients' willingness to use mobile applications for self-management: a cross-sectional study. J Diabetes Investig 2017;8(4):542-549 [FREE Full text] [doi: 10.1111/jdi.12613] [Medline: 28012247]
25. Diabetes diagnosis and treatment guidelines 2019. The Japan Diabetes Society. 2019. URL: http://www.jds.or.jp/modules/publication/index.php?content_id=4 [accessed 2023-06-23]
26. Ang SH, Thevarajah M, Alias Y, Khor SM. Current aspects in hemoglobin A1c detection: a review. Clin Chim Acta 2015;439:202-211 [FREE Full text] [doi: 10.1016/j.cca.2014.10.019] [Medline: 25451954]
27. Little RR, Rohlfing CL. The long and winding road to optimal HbA1c measurement. Clin Chim Acta 2013;418:63-71 [FREE Full text] [doi: 10.1016/j.cca.2012.12.026] [Medline: 23318564]
28. Sze WT, Waki K, Enomoto S, Nagata Y, Nangaku M, Yamauchi T, et al. StepAdd: a personalized mHealth intervention based on social cognitive theory to increase physical activity among type 2 diabetes patients. J Biomed Inform 2023;145:104481 [FREE Full text] [doi: 10.1016/j.jbi.2023.104481] [Medline: 37648101]
29. Waki K, Nara M, Enomoto S, Mieno M, Kanda E, Sankoda A, et al. Effectiveness of DialBetesPlus, a self-management support system for diabetic kidney disease: randomized controlled trial. NPJ Digit Med 2024;7(1):104 [FREE Full text] [doi: 10.1038/s41746-024-01114-8] [Medline: 38678094]
30. Ajzen I. The Theory of Planned Behavior: Habit, Perceived Control, and Reasoned Action. Germany: Mannheimer Zentrum für Europäische Sozialforschung; 2000.
31. Taheri S, Zaghloul H, Chagoury O, Elhadad S, Ahmed SH, El Khatib N, et al. Effect of intensive lifestyle intervention on bodyweight and glycaemia in early type 2 diabetes (DIADEM-I): an open-label, parallel-group, randomised controlled trial. Lancet Diabetes Endocrinol 2020;8(6):477-489. [doi: 10.1016/S2213-8587(20)30117-0] [Medline: 32445735]
32. Coppell KJ, Kataoka M, Williams SM, Chisholm AW, Vorgers SM, Mann JI. Nutritional intervention in patients with type 2 diabetes who are hyperglycaemic despite optimised drug treatment—lifestyle over and above drugs in diabetes (LOADD) study: randomised controlled trial. BMJ 2010;341:c3337 [FREE Full text] [doi: 10.1136/bmj.c3337] [Medline: 20647285]
33. Kirk A, Mutrie N, MacIntyre P, Fisher M. Effects of a 12-month physical activity counselling intervention on glycaemic control and on the status of cardiovascular risk factors in people with type 2 diabetes. Diabetologia 2004;47(5):821-832. [doi: 10.1007/s00125-004-1396-5] [Medline: 15138687]
34. Sze WT, Waki K, Enomoto S, Nagata Y, Nangaku M, Yamauchi T, et al. StepAdd: a personalized mHealth intervention based on social cognitive theory to increase physical activity among type 2 diabetes patients. J Biomed Inform 2023;145:104481 [FREE Full text] [doi: 10.1016/j.jbi.2023.104481] [Medline: 37648101]
35. Shibuta T, Waki K, Tomizawa N, Igarashi A, Yamamoto-Mitani N, Yamaguchi S, et al. Willingness of patients with diabetes to use an ICT-based self-management tool: a cross-sectional study. BMJ Open Diabetes Res Care 2017;5(1):e000322 [FREE Full text] [doi: 10.1136/bmjdrc-2016-000322] [Medline: 28243450]
36. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
37. Nakagawa Y, Moriya K, Ishikawa Y, Yoneshiro T, Musashi M. Influence of duration of diabetes and counseling frequency of dietitian-led medical nutrition therapy based on the transtheoretical model on its beneficial effects in patients with type 2 diabetes. J Jpn Health Med Assoc 2015;24(2):103-113 [FREE Full text]
38. Klonoff DC. Behavioral theory: the missing ingredient for digital health tools to change behavior and increase adherence. J Diabetes Sci Technol 2019;13(2):276-281 [FREE Full text] [doi: 10.1177/1932296818820303] [Medline: 30678472]
39. Salinero-Fort MA, Carrillo-de Santa Pau E, Arrieta-Blanco FJ, Abanades-Herranz JC, Martín-Madrazo C, Rodés-Soldevila B, et al. Effectiveness of PRECEDE model for health education on changes and level of control of HbA1c, blood pressure,

XSL•FO
RenderX

lipids, and body mass index in patients with type 2 diabetes mellitus. BMC Public Health 2011;11:267 [FREE Full text] [doi: 10.1186/1471-2458-11-267] [Medline: 21524316]
40. Fuji KT, Abbott AA, Galt KA. A mixed-methods evaluation of standalone personal health record use by patients with type 2 diabetes. Perspect Health Inf Manag 2021;18(4):1e [FREE Full text] [Medline: 34975354]
41. Cui M, Wu X, Mao J, Wang X, Nie M. T2DM self-management via smartphone applications: a systematic review and meta-analysis. PLoS One 2016;11(11):e0166718 [FREE Full text] [doi: 10.1371/journal.pone.0166718] [Medline: 27861583]
42. Madujibeya I, Lennie T, Aroh A, Chung M, Moser D. Measures of engagement with mHealth interventions in patients with heart failure: scoping review. JMIR mHealth uHealth 2022;10(8):e35657 [FREE Full text] [doi: 10.2196/35657] [Medline: 35994345]
43. McLean A, Osgood N, Newstead-Angel J, Stanley K, Knowles D, van der Kamp W, et al. Building research capacity: results of a feasibility study using a novel mHealth epidemiological data collection system within a gestational diabetes population. Stud Health Technol Inform 2017;234:228-232. [Medline: 28186046]
44. Reychav I, Beeri R, Balapour A, Raban D, Sabherwal R, Azuri J. How reliable are self-assessments using mobile technology in healthcare? The effects of technology identity and self-efficacy. Comput Hum Behav 2019;91:52-61 [FREE Full text] [doi: 10.1016/j.chb.2018.09.024]
45. Kawai Y, Sankoda A, Waki K, Miyake K, Hayashi A, Mieno M, et al. Efficacy of the self-management support system DialBetesPlus for diabetic kidney disease: protocol for a randomized controlled trial. JMIR Res Protoc 2021;10(8):e31061 [FREE Full text] [doi: 10.2196/31061] [Medline: 34402802]
46. Chomutare T, Fernandez-Luque L, Arsand E, Hartvigsen G. Features of mobile diabetes applications: review of the literature and analysis of current applications compared against evidence-based guidelines. J Med Internet Res 2011;13(3):e65 [FREE Full text] [doi: 10.2196/jmir.1874] [Medline: 21979293]
47. Glanz K, Bishop DB. The role of behavioral science theory in development and implementation of public health interventions. Annu Rev Public Health 2010;31:399-418. [doi: 10.1146/annurev.publhealth.012809.103604] [Medline: 20070207]
48. Møller JB, Pedersen M, Tanaka H, Ohsugi M, Overgaard RV, Lynge J, et al. Body composition is the main determinant for the difference in type 2 diabetes pathophysiology between Japanese and Caucasians. Diabetes Care 2014;37(3):796-804. [doi: 10.2337/dc13-0598] [Medline: 24130359]
49. Ma RCW, Chan JCN. Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. Ann N Y Acad Sci 2013;1281(1):64-91 [FREE Full text] [doi: 10.1111/nyas.12098] [Medline: 23551121]

Abbreviations

FAS: full analysis set
HbA_{1c}: hemoglobin A_{1c}
ITT: intention-to-treat analysis
mHealth: mobile health
PPS: per protocol set
SMC: Save Medical Corporation
T2DM: type 2 diabetes mellitus

Edited by G Eysenbach, N Cahill; submitted 17.10.23; peer-reviewed by M Savolainen, HY Chen; comments to author 07.06.24; revised version received 29.06.24; accepted 26.07.24; published 10.09.24.

Please cite as:
Leung N, Waki K, Nozoe S, Enomoto S, Saito R, Hamagami S, Yamauchi T, Nangaku M, Ohe K, Onishi Y
Efficacy of Save Medical Corporation (SMC)–01, a Smartphone App Designed to Support Type 2 Diabetes Self-Management Based on Established Guidelines: Randomized Controlled Trial
J Med Internet Res 2024;26:e53740
URL: https://www.jmir.org/2024/1/e53740
doi:10.2196/53740
PMID:

©Nicholas Leung, Kayo Waki, Satoshi Nozoe, Shunpei Enomoto, Ryo Saito, Sakurako Hamagami, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe, Yukiko Onishi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Concerns About the Generalizability Associated With a South African Randomized Controlled Trial on Prenatal Mothers

Yongjian Lin[1], MD

Department of Gastrointestinal and Gland Surgery, The First Affiliated Hospital of Guangxi Medical University, Nanning, China

Corresponding Author:
Yongjian Lin, MD
Department of Gastrointestinal and Gland Surgery
The First Affiliated Hospital of Guangxi Medical University
No. 22, Shuangyou Road
Qingxiu District
Nanning, 530021
China
Phone: 86 13878870525
Email: linyongjian@stu.gxmu.edu.cn

Related Articles:

Comment on: http://www.jmir.org/2023/1/e47266/

Comment in: http://www.jmir.org/2024/1/e55930/

(J Med Internet Res 2024;26:e53861) doi:10.2196/53861

KEYWORDS

letter; maternal child health; mHealth; mobile health; randomized controlled trial; short animated storytelling; South Africa; video health messaging

We have carefully reviewed the study conducted by Adam et al [1], which assessed the impact of short animated storytelling (SAS) on prenatal mothers' maternal knowledge and user satisfaction. The findings of this study indicated that while the SAS videos had high user satisfaction, the measured knowledge gains were limited, which could work as a broad-reaching perinatal health program [1]. In light of our thorough research for this paper, we would like to offer the following suggestions.

First, as the authors described in the Trial Design section, "participants were randomly assigned to either the SAS Intervention group or the Standard-of-Care (SOC) Control group," rather than using propensity score matching [2]. Most patients have individual polymorphisms in the real world, whereas no significant differences were seen between the SAS intervention group and the control group, which is not consistent with individual polymorphism. Furthermore, when describing data where population characteristics are unknown (eg, knowledge and maternal satisfaction outcomes), a nonparametric test using rank or approximate *t* test may be considered. Therefore, the results of this study may not truly reflect the situation of patients in the real world.

Second, we explored whether the SAS videos played a role in innovative approaches to perinatal health services, and it remains ambiguous whether the relationship between the variables examined in this study (eg, age, education, area, race, and working status) and maternal knowledge was linear or curvilinear. After adjusting for age, sex, education level, living area, and working status, it is advisable to use generalized estimating equations for estimating intervention effects instead of multivariate regression analysis [3]. This is due to the capability of the generalized linear model to account not only for the linear relationship between the two variables but also for the curvilinear relationship between them.

Finally, interaction is important in clinical studies because environmental factors and lifestyle habits may have an impact on clinical outcomes. For example, Yakupova and Liutsko [4] noted that physical health during pregnancy and marital satisfaction significantly predicted prenatal depression based on data from middle-income countries. Another study showed that the climate and physical environment of the care setting consistently influenced women's satisfaction, despite varying ethnic backgrounds [5]. In this study, many variables were described in Table 1, but subgroup analyses were not conducted based on factors such as age, race, education level, and income. Hence, this paper would be more meaningful if these factors were analyzed in layers.

In conclusion, we thank the authors for this excellent work that provides an approach to perinatal care. However, we believe

XSL•FO
RenderX

that the conclusions of this study would change if the previously mentioned issues were further addressed.

Acknowledgments

Our deepest gratitude goes to the authors for this excellent work that provides important evidence.

Authors' Contributions

YL conceptualized the study, prepared the original draft, was the project administrator, and reviewed and edited the manuscript.

Conflicts of Interest

None declared.

References

1. Adam M, Kwinda Z, Dronavalli M, Leonard E, Nguy n VK, Tshivhase V, et al. Effect of short, animated video storytelling on maternal knowledge and satisfaction in the perinatal period in South Africa: randomized controlled trial. J Med Internet Res 2023 Oct 13;25:e47266 [FREE Full text] [doi: 10.2196/47266] [Medline: 37831505]
2. Gillespie IA, Floege J, Gioni I, Drüeke TB, de Francisco AL, Anker SD, et al. Propensity score matching and persistence correction to reduce bias in comparative effectiveness: the effect of cinacalcet use on all-cause mortality. Pharmacoepidemiol Drug Saf 2015 Jul;24(7):738-747 [FREE Full text] [doi: 10.1002/pds.3789] [Medline: 26011775]
3. Pan L, Li Y, He K, Li Y, Li Y. Generalized linear mixed models with gaussian mixture random effects: inference and application. J Multivar Anal 2020 Jan;175:104555 [FREE Full text] [doi: 10.1016/j.jmva.2019.104555] [Medline: 32063658]
4. Yakupova V, Liutsko L. Perinatal depression, birth experience, marital satisfaction and childcare sharing: a study in Russian mothers. Int J Environ Res Public Health 2021 Jun 04;18(11):6086 [FREE Full text] [doi: 10.3390/ijerph18116086] [Medline: 34200046]
5. Handler A, Raube K, Kelley MA, Giachello A. Women's satisfaction with prenatal care settings: a focus group study. Birth 1996 Mar;23(1):31-37. [doi: 10.1111/j.1523-536x.1996.tb00458.x] [Medline: 8703255]

Abbreviations

SAS: short animated storytelling

Edited by T Leung; submitted 22.10.23; this is a non–peer-reviewed article;accepted 29.12.23; published 12.02.24.

Please cite as:
Lin Y
Concerns About the Generalizability Associated With a South African Randomized Controlled Trial on Prenatal Mothers
J Med Internet Res 2024;26:e53861
URL: https://www.jmir.org/2024/1/e53861
doi:10.2196/53861
PMID:38345847

©Yongjian Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Authors' Reply: Concerns About the Generalizability Associated With a South African Randomized Controlled Trial on Prenatal Mothers

Maya Adam[1,2*], MD; Vān Kính Nguy n[2*], PhD

[1]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, United States
[2]Institute of Global Health, Heidelberg University, Heidelberg, Germany
[*]all authors contributed equally

Corresponding Author:
Maya Adam, MD
Department of Pediatrics
Stanford University School of Medicine
291 Campus Drive
Stanford, CA, 94305
United States
Phone: 1 6508393600
Email: madam@stanford.edu

Related Articles:

Comment on: http://www.jmir.org/2024/1/e53861/

Comment on: http://www.jmir.org/2023/1/e47266/

(J Med Internet Res 2024;26:e55930) doi:10.2196/55930

KEYWORDS

maternal child health; mHealth; mobile health; randomized controlled trial; short animated storytelling; South Africa; video health messaging

We are responding to a letter to the editor [1] regarding our recent article, "Effect of Short, Animated Video Storytelling on Maternal Knowledge and Satisfaction in the Perinatal Period in South Africa: Randomized Controlled Trial" [2].

Unfortunately, we believe the letter's author (LA) may not have read the paper as carefully as they have claimed, and because of this, we find their criticisms to be somewhat misguided. Please see our detailed responses below.

The LA's comment that the "study may not truly reflect the situation of patients in the real world" because we did not choose to use propensity score matching is unfounded. Propensity score matching has many applications [3]; one of them is matching to conduct *quasi*–randomized controlled trials. Large-scale and *ad hoc* population trials could benefit from a propensity score matching approach, but this was not the aim of our study. Importantly, one applies this type of matching to ensure that the background variables are similarly distributed among the comparison groups. In our study, we conducted proper random assignment of the participants and found that, quoting the letter, "no significant differences were seen between the SAS Intervention group and the control group" (Table 1). Thus, while we did not control for all the potential covariates, our design provides the highest level of confidence to identify the intervention effects. We do not see a link between the statistical test used in this study and the LA's statement that the study "may not truly reflect the situation of patients." We respectfully refer the LA to the Statistical Analysis section of our paper [2] where we detail why we used the beta-binomial model to describe the score distribution and control for covariate effects instead of a nonparametric univariate analysis of the knowledge score.

The LA's comment advising us to use generalized estimating equations (GEEs) [4] has several unfounded statements: there are no longitudinal data so GEEs would not be advisable here, GEEs are a form of multivariate regression analysis, generalized linear model (GLM) was already used in our analysis, simply using GLM does not address whether a curvilinear relationship exists, and it is not necessary to use GLM to model curvilinear relationships.

Finally, while we agree that environmental factors might affect marital satisfaction, this is unrelated to our paper. The LA misread and did not realize that we measured *maternal*

XSL•FO
RenderX

satisfaction with the video intervention; however, our main outcome was not maternal satisfaction but the knowledge score. The LA also erroneously states that "subgroup analyses were not conducted based on factors such as age, race, education level, and income. Hence, this paper would be more meaningful if these factors were analyzed in layers." In Figure 3 of our paper [2], we show the results of our multivariable regression analysis, including the factors listed by the LA.

Conflicts of Interest

None declared.

References

1. Lin Y. Concerns about the generalizability associated with randomized controlled trials. J Med Internet Res 2024;6:e53861. [doi: 10.2196/53861]
2. Adam M, Kwinda Z, Dronavalli M, Leonard E, Nguy n VK, Tshivhase V, et al. Effect of short, animated video storytelling on maternal knowledge and satisfaction in the perinatal period in South Africa: randomized controlled trial. J Med Internet Res 2023 Oct 13;25:e47266 [FREE Full text] [doi: 10.2196/47266] [Medline: 37831505]
3. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. J Econ Surveys 2008 Jan 31;22(1):31-72. [doi: 10.1111/j.1467-6419.2007.00527.x]
4. Hanley J, Negassa A, Edwardes M, Forrester J. Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 2003 Feb 15;157(4):364-375. [doi: 10.1093/aje/kwf215] [Medline: 12578807]

Abbreviations

GEE: generalized estimating equation
GLM: generalized linear model
LA: letter's author

Edited by T Leung; submitted 29.12.23; this is a non–peer-reviewed article;accepted 24.01.24; published 12.02.24.

Please cite as:
Adam M, Nguy n VK
Authors' Reply: Concerns About the Generalizability Associated With a South African Randomized Controlled Trial on Prenatal Mothers
J Med Internet Res 2024;26:e55930
URL: https://www.jmir.org/2024/1/e55930
doi:10.2196/55930
PMID:38345840

©Maya Adam, Vān Kính Nguy n. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Ambiguity in Statistical Analysis Methods and Nonconformity With Prespecified Commitment to Data Sharing in a Cluster Randomized Controlled Trial

Yasaman Jamshidi-Naeini[1], PhD; Lilian Golzarri-Arroyo[1], MS; Deependra K Thapa[1], PhD; Andrew W Brown[2,3], PhD; Daniel E Kpormegbey[1], PhD; David B Allison[1], PhD

[1]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, Bloomington, IN, United States
[2]Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, AR, United States
[3]Arkansas Children's Research Institute, Little Rock, AR, United States

Corresponding Author:
David B Allison, PhD
Department of Epidemiology and Biostatistics
School of Public Health
Indiana University Bloomington
1025 E 7th St, PH 111
Bloomington, IN, 47405
United States
Phone: 1 8128551250
Email: allison@iu.edu

Related Articles:

Comment on: http://www.jmir.org/2023/1/e43018/

Comment in: http://www.jmir.org/2024/1/e57422/

(J Med Internet Res 2024;26:e54090) doi:10.2196/54090

KEYWORDS

cluster randomized trial; clustering; nesting; data availability; random allocation; data sharing; reproducibility

In a published cluster randomized controlled trial (cRCT) [1] on the effects of desks on physical behaviors, 66 individuals were randomized into 3 groups by their office space (ie, clusters): the seated desk control (n=21; 8 clusters), sit-to-stand desk (n=23; 9 clusters), or treadmill desk (n=22; 7 clusters) group.

In the article, there is ambiguity regarding whether clustering (potential nonindependence of observations within the same office space) and nesting (due to the hierarchical structure of the data; see definitions from Jamshidi-Naeini et al [2]) have been accounted for. Furthermore, it appears that the data underlying the published results are not available to other researchers, which is contrary to the journal's policy indicating "a submission to JMIR journals implies that…all relevant raw data, will be freely available to any researcher wishing to use them for non-commercial purposes…."

The description of methods indicates using random intercept mixed linear models accounting for repeated measures and clusters. However, it is stated elsewhere that "[t]he cluster effect did not significantly (all *P* values >.05) account for the variability in any of the outcome variables….Therefore, aim 1 and aim 2 outcome observations…were analyzed at the participant level instead of cluster…" [1].

The statement regarding accounting for the clustering effect is ambiguous. If the authors ignored the clustering effect based on the reasoning that participant outcomes within the same cluster are unrelated, as indicated by intraclass correlation coefficient (ICC) values that one chooses to describe as "small" or nonstatistical significance of the ICC values at some nominal α level (eg, 0.05), such reasoning is erroneous. A sample ICC or its associated *P* value is not an appropriate metric on which to determine whether one should account for clustering. Ignoring clustering, regardless of a sample ICC's magnitude or associated *P* value, potentially leads to miscalculation of variance components and type I error rates above the nominal significance level [3,4].

In a cRCT with such unequal cluster sizes (ranging from 1 to 11 participants), there is no exact size α test, and type I error inflation may occur. Therefore, to ensure control of type I error rate, it is essential to apply appropriate weighting for unequal

XSL•FO
RenderX

cluster sizes. In addition, the nesting effect that arises from the hierarchical structure of the data in cRCTs was not considered in the statistical analyses. Adjusting *df* (eg, by between-within determination [4]) could have accounted for this nesting effect.

We requested the raw data to reproduce the analyses (see definition of "reproducing" in *Reproducibility and Replicability in Science* [5]) and potentially use alternative corrected methods to reanalyze the data. Data were not shared with us. The authors stated that this decision was made "to ensure the integrity of ongoing research being conducted using the same dataset." Nevertheless, sharing data for the purpose of reproducing published results does not compromise the integrity of further analyses on the same data set. Withholding data, on the other hand, renders the study irreproducible and thus compromises the trustworthiness of the published results.

The concerns raised herein should be addressed to ensure the integrity, transparency, and reproducibility of the published findings.

Acknowledgments

The authors are supported in part by R25DK099080, R25HL124208, R25GM141507, and the Gordon and Betty Moore Foundation. The opinions expressed are those of the authors and do not necessarily represent those of the National Institutes of Health (NIH) or any other organization.

Conflicts of Interest

Collectively, the authors and their institutions have received payments for consultations, grants, contracts, in-kind donations, and contributions from multiple for-profit and not-for-profit entities interested in statistical design and analysis of experiments but not directly related to the research questions addressed in the paper in question.

References

1. Arguello D, Cloutier G, Thorndike AN, Castaneda Sceppa C, Griffith J, John D. Impact of sit-to-stand and treadmill desks on patterns of daily waking physical behaviors among overweight and obese seated office workers: cluster randomized controlled trial. J Med Internet Res 2023 May 16;25:e43018 [FREE Full text] [doi: 10.2196/43018] [Medline: 37191995]
2. Jamshidi-Naeini Y, Brown AW, Mehta T, Glueck DH, Golzarri-Arroyo L, Muller KE, et al. A practical decision tree to support editorial adjudication of submitted parallel cluster randomized controlled trials. Obesity (Silver Spring) 2022 Mar;30(3):565-570 [FREE Full text] [doi: 10.1002/oby.23373] [Medline: 35195364]
3. Brown AW, Li P, Brown MMB, Kaiser KA, Keith SW, Oakes JM, et al. Best (but oft-forgotten) practices: designing, analyzing, and reporting cluster randomized controlled trials. Am J Clin Nutr 2015 Aug;102(2):241-248 [FREE Full text] [doi: 10.3945/ajcn.114.105072] [Medline: 26016864]
4. Golzarri-Arroyo L, Dickinson SL, Jamshidi-Naeini Y, Zoh RS, Brown AW, Owora AH, et al. Evaluation of the type I error rate when using parametric bootstrap analysis of a cluster randomized controlled trial with binary outcomes and a small number of clusters. Comput Methods Programs Biomed 2022 Mar;215:106654 [FREE Full text] [doi: 10.1016/j.cmpb.2022.106654] [Medline: 35093646]
5. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press; 2019.

Abbreviations

cRCT: cluster randomized controlled trial
ICC: intraclass correlation coefficient

Edited by T Leung; submitted 29.10.23; this is a non–peer-reviewed article;accepted 22.02.24; published 03.04.24.

Please cite as:
Jamshidi-Naeini Y, Golzarri-Arroyo L, Thapa DK, Brown AW, Kpormegbey DE, Allison DB
Ambiguity in Statistical Analysis Methods and Nonconformity With Prespecified Commitment to Data Sharing in a Cluster Randomized Controlled Trial
J Med Internet Res 2024;26:e54090
URL: https://www.jmir.org/2024/1/e54090
doi:10.2196/54090
PMID:38568721

XSL•FO
RenderX

©Yasaman Jamshidi-Naeini, Lilian Golzarri-Arroyo, Deependra K Thapa, Andrew W Brown, Daniel E Kpormegbey, David B Allison. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Authors' Reply: Ambiguity in Statistical Analysis Methods and Nonconformity With Prespecified Commitment to Data Sharing in a Cluster Randomized Controlled Trial

Diego Arguello[1], BA, MSc, PhD

Human Performance and Exercise Science Lab, Department of Health Sciences, Northeastern University, Boston, MA, United States

Corresponding Author:
Diego Arguello, BA, MSc, PhD
Human Performance and Exercise Science Lab
Department of Health Sciences
Northeastern University
360 Huntington Avenue
Boston, MA, 02115
United States
Phone: 1 6173734427
Email: arguello.d@Northeastern.edu

Related Articles:

Comment on: https://mhealth.jmir.org/2024/1/e54090/

Comment on: https://mhealth.jmir.org/2023/1/e43018/

(J Med Internet Res 2024;26:e57422) doi:10.2196/57422

KEYWORDS

prolonged sedentary behavior; sedentary behavior; sit-to-stand desks; treadmill desks; physical activity promotion; workplace wellness; seated office workers; move more and sit less

We appreciate the thoughtful commentary [1] on our study [2] and address the raised concerns.

One pertains to perceived ambiguity in reporting statistical methodology, presumably prompting inquiry into whether analyses accounted for clustering and nesting. We clarify that our choice of a cluster randomized controlled trial design was driven by practical implications (eg, space modification and experimental contamination)—valid reasons for the study design. While this determined the study design, we were interested in the participant-average treatment effect. We feel this is affirmed in our aims and note the perceived ambiguity and level of detail requested by the letter writers were not raised during multiple reviews, and this response to the letter further clarifies our analytical approach.

We agree that type I error inflation is a concern if clustering is not addressed in the study design phase or during analyses. We also recognize that this is an active area of study, and work is ongoing to determine optimal approaches for small-sample study designs [3]. Importantly, the current recommendation is to develop the analytical plan based on goals of the analyses (ie, exploratory outcomes in this study) and various study-specific characteristics (selected examples for our study: random cluster size independent of other data and wide physical distribution of within-cluster participants) [3]. Due to such study-specific considerations, clusters were deemed to be noninformative [4], and our approach—"random-intercept mixed linear models that accounted for repeated measures and clustering effects" [2]—included a random effect for clusters to model any potential correlational structure and interparticipant dependency within clusters [5]. Additionally, we used the Kenward-Rogers method to preserve nominal type I error. The method adjusts for *df* to account for hierarchical complexity of data, including potential nesting and variable or small cluster sizes [3]. We also acknowledged uneven cluster size as a study limitation [2].

Second, while we clearly conveyed the exploratory nature of the analyses aimed at developing hypotheses, we were also conservative by avoiding confirmatory conclusions based on type I error rates [2]. While it is known that conservative analyses may be counterproductive for exploration [6], our careful approach—relying on sound, disparate methods sufficiently accounting for any potential within-cluster dependence and variable cluster size, along with a cautious interpretation strategy—was appropriate for the study objective.

XSL•FO
RenderX

Regarding data sharing, we have previously shared other data and received data from the community to advance multiple areas of inquiry. However, we primarily rejected this request and a competing industry request, given ongoing small business and financial interests leveraging this and the associated body of work; we communicated this to the journal (August 2023) after the request in question was made. Such considerations are necessary to avoid setting a precedent and in the context of any potential for these interests to be compromised. An example of such considerations includes federal funding agencies supporting small business innovation research, allowing awardees to withhold related data to protect endeavors similar to ours. Implying that findings are untrustworthy due to such considerations would incorrectly render a substantial body of such work as the same.

Conflicts of Interest

Since June 2023, DA has received financial compensation for consultations leveraging the published work.

References

1. Jamshidi-Naeini Y, Golzarri-Arroyo L, Thapa DK, Brown AW, Kpormegbey DE, Allison DB. Ambiguity in statistical analysis methods and nonconformity with prespecified commitment to data sharing in a cluster randomized controlled trial. J Med Internet Res 2024:e54090. [doi: 10.2196/54090]
2. Arguello D, Cloutier G, Thorndike AN, Castaneda Sceppa C, Griffith J, John D. Impact of sit-to-stand and treadmill desks on patterns of daily waking physical behaviors among overweight and obese seated office workers: cluster randomized controlled trial. J Med Internet Res 2023 May 16;25:e43018 [FREE Full text] [doi: 10.2196/43018] [Medline: 37191995]
3. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? Int J Epidemiol 2018 Feb 1;47(1):321-331. [doi: 10.1093/ije/dyx169] [Medline: 29025158]
4. Wang B, Harhay MO, Tong J, Small DS, Morris TP, Li F. On the mixed-model analysis of covariance in cluster-randomized trials. arXiv Preprint posted online on Oct 9, 2023 [FREE Full text] [doi: 10.48550/arXiv.2112.00832]
5. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London, England: Arnold Publishers; 2000.
6. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol 2014 May 20;12(5):e1001863 [FREE Full text] [doi: 10.1371/journal.pbio.1001863] [Medline: 24844265]

Edited by T Leung; submitted 16.02.24; this is a non–peer-reviewed article;accepted 22.02.24; published 03.04.24.

Please cite as:
Arguello D
Authors' Reply: Ambiguity in Statistical Analysis Methods and Nonconformity With Prespecified Commitment to Data Sharing in a Cluster Randomized Controlled Trial
J Med Internet Res 2024;26:e57422
URL: https://www.jmir.org/2024/1/e57422
doi:10.2196/57422
PMID:38568734

©Diego Arguello. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications

Kuan-Ju Huang[1], MS

Department of Obstetrics and Gynecology, National Taiwan University Hospital Yunlin Branch, Yunlin County, Taiwan

Corresponding Author:
Kuan-Ju Huang, MS
Department of Obstetrics and Gynecology
National Taiwan University Hospital Yunlin Branch
No 579, Sec 2, Yunlin Rd, Douliu City
Yunlin County, 640
Taiwan
Phone: 886 55323911 ext 563413
Fax: 886 55335325
Email: restroomer@icloud.com

Related Articles:

Comment on: http://www.jmir.org/2024/1/e52113/

Comment in: http://www.jmir.org/2024/1/e57778/

(J Med Internet Res 2024;26:e56997) doi:10.2196/56997

KEYWORDS

artificial intelligence; ChatGPT; Bloom taxonomy; AI; cognition

We are inspired by Herrmann-Werner et al's [1] article, which assesses GPT-4's cognitive functions based on the Bloom taxonomy. Adopting the Bloom taxonomy for evaluating GPT-4's understanding of specific knowledge, traditionally applied to humans, is a novel concept. The results could also offer insights into whether GPT-4 can think like a human. However, some points in this article need clarification.

First, in Figure 3, the difficulty of the questions might have been inversely reported in the abstract, with 0 representing a very difficult question and 1 representing a very easy question, according to the description in the Quantitative Data Analysis subsection of the Methods. Consequently, GPT-4 performed better on easy questions than on hard ones.

Second, since a large language model (LLM) like GPT-4 operates by predicting the next word from its memory-based archive [2], it seems unlikely that GPT-4 would perform worst in the "remember" domain of the Bloom taxonomy in this study (42.65%) and excel in higher cognitive domains such as analyze, evaluate, and create, with incorrect reasoning accounting for 0%, 0.15%, and 0%, respectively, as reported in Table 3 [1]. The Bloom taxonomy categorizes the aims of questions, not the answers, in evaluating a "student's" cognitive level within specific domains. Therefore, evaluating GPT-4's cognitive functions by analyzing its responses presupposes that GPT-4 can think like a human. However, given our current understanding of how LLMs generate answers—essentially predicting the next word based on probabilities within a database—it is doubtful that GPT-4's cognitive levels in responses can be accurately assessed using the Bloom taxonomy, especially with high scores in advanced cognitive domains [2].

For example, when evaluating "memory" (eg, definitions, guidelines, or facts), if the combination of elements exists in its database, GPT-4 can readily produce the most likely answers from its "memory." Conversely, when elements are incorrectly combined, it may produce "hallucinated" answers [2]. In complex questions that test higher cognitive domains (eg, analyzing a previously unpublished case report with findings from subjective and objective medical evaluations to deduce the most likely diagnosis), if a similar case or key elements exist in GPT-4's database, it might still produce a result from its "memory," seemingly "analyzing, evaluating, and creating" an answer as it has "learned" from human problem-solving in similar cases. This "memory" function, considered LLM's most potent capability compared to humans, can yield incorrect answers if the "memory" does not exist in the database (eg, news) or is not predicted as the next word. The apparent high cognitive function might result from the model's ability to extract multiple human thought processes about a specific question from its vast database, akin to a well-trained system mimicking human cognitive processes [3,4].

XSL•FO
RenderX

Since most medical qualifying exams consist mainly of “memory” tests, the actual count of incorrect reasoning in the “memory” domain could be lower when both correct and incorrect answers are combined. Until more evidence proving that LLMs can think like humans is available, evaluating LLM-generated answers through the Bloom taxonomy may yield misleading results.

Conflicts of Interest

None declared.

References

1. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024 Jan 23;26:e52113 [FREE Full text] [doi: 10.2196/52113] [Medline: 38261378]
2. GPT-4 system card. OpenAI. 2023 Mar 23. URL: https://cdn.openai.com/papers/gpt-4-system-card.pdf [accessed 2024-04-09]
3. Stanovich KE, West RF. Individual differences in reasoning: implications for the rationality debate? Behav Brain Sci 2000 Oct;23(5):645-65; discussion 665. [doi: 10.1017/s0140525x00003435] [Medline: 11301544]
4. Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux; 2011.

Abbreviations

LLM: large language model

Edited by T Leung; submitted 01.02.24; this is a non–peer-reviewed article;accepted 04.04.24; published 16.04.24.

Please cite as:
Huang KJ
Evaluating GPT-4’s Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications
J Med Internet Res 2024;26:e56997
URL: https://www.jmir.org/2024/1/e56997
doi:10.2196/56997
PMID:38625725

©Kuan-Ju Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Authors' Reply: "Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications"

Anne Herrmann-Werner[1,2], MME, Prof Dr Med; Teresa Festl-Wietek[1], MSc, Dr Rer Nat; Friederike Holderried[1,3], MME, Dr Med; Lea Herschbach[1], MSc; Jan Griewatz[1], MA; Ken Masters[4], Prof Dr; Stephan Zipfel[2], Prof Dr Med; Moritz Mahling[1,5], MHBA, Dr Med

[1]Tübingen Institute for Medical Education, Faculty of Medicine, University of Tübingen, Tübingen, Germany
[2]Department of Psychosomatic Medicine and Psychotherapy, University Hospital Tübingen, Tübingen, Germany
[3]University Department of Anesthesiology and Intensive Care Medicine, University Hospital Tübingen, Tübingen, Germany
[4]Medical Education and Informatics Department, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Oman
[5]Department of Diabetology, Endocrinology, Nephrology, Section of Nephrology and Hypertension, University Hospital Tübingen, Tübingen, Germany

Corresponding Author:
Teresa Festl-Wietek, MSc, Dr Rer Nat
Tübingen Institute for Medical Education
Faculty of Medicine
University of Tübingen
Elfriede-Aulhorn-Strasse 10
Tübingen, 72076
Germany
Phone: 49 7071 29 73715
Email: teresa.festl-wietek@med.uni-tuebingen.de

Related Articles:

Comment on: https://www.jmir.org/2024/1/e56997/

Comment on: https://www.jmir.org/2024/1/e52113

(J Med Internet Res 2024;26:e57778) doi:10.2196/57778

KEYWORDS

answer; artificial intelligence; assessment; Bloom's taxonomy; ChatGPT; classification; error; exam; examination; generative; GPT-4; Generative Pre-trained Transformer 4; language model; learning outcome; LLM; MCQ; medical education; medical exam; multiple-choice question; natural language processing; NLP; psychosomatic; question; response; taxonomy

We appreciate the thoughtful commentary titled "Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications" [1] and welcome the opportunity to clarify and expand upon our research findings [2] regarding GPT-4's cognitive evaluation using the Bloom taxonomy.

First, we acknowledge the confusion surrounding the use of the term "difficulty" in our manuscript. Traditionally in educational testing, "difficulty" is quantified by the ratio of correct responses against the number of students taking the test [3]; thus, a rating of 1 indicates an extremely simple question (100% correct responses), and a rating of 0 indicates a significantly challenging question (0% correct responses). Throughout the manuscript, we used "difficulty" as a measurement scale.

Consequently, "higher difficulty" means it is higher on the scale and thus easier. This also applies to Figure 3. Because "lower" means less easy (ie, closer to 0 on the scale from 0 to 1), it shows that the questions answered correctly were easier compared to those answered wrong. Although our use of the measurement "difficulty" is correct, on reflection, we agree that we could have been clearer, and we apologize for any confusion.

Second, the commentary on GPT-4's approach to "memory" tasks adds a valuable dimension to our discussion. We agree that GPT-4 "remembers" through technical and programmatic means, highlighting the critical difference between GPT-4's architecture and human cognitive processes, a distinction that was central to our study.

However, GPT-4's material selection is far more complex than a flat-file database with simple mapping (unless the exam questions had been in the testing data, but this is not applicable in our case). Generative tools like GPT-4 have other weaknesses and strengths. For example, they may perform relatively poorly on pure memory-recall problems but excel in topics requiring

XSL•FO
RenderX

subtlety and nuanced work. This is demonstrated by GPT-4's high performance on soft-skill questions from the USMLE (United States Medical Licensing Examination) and AMBOSS [4]. Part of our study went further by using the Bloom taxonomy as a framework for tracing the logical process of GPT-4's *explanations* (not *answers*) and determining the stages at which its errors occurred.

This discussion underscores a critical point: the complexity of assessing artificial intelligence and the processes underlying the output of models like GPT-4. This methodology allows us to critically examine where GPT-4's responses fall within a spectrum of cognitive tasks, from simple recall to more complex analytical and evaluative processes.

Third, while it is quite true that many questions in medical qualifying exams are simple memory-type questions, we see this as a weakness rather than an optimum aiming point. While our understanding is that medical schools are trying to move away from those types of questions, this is an area of further research.

Again, we thank the author for the thoughtful critique of our paper and the resultant continued discussion, which underscores the importance of ongoing dialogue and research into artificial intelligence's cognitive processes and how they parallel and diverge from human cognition.

Conflicts of Interest

None declared.

References

1. Huang KJ. Evaluating GPT-4's cognitive functions through the Bloom taxonomy: insights and clarifications. J Med Internet Res 2024 Apr 16;26:e56997 [FREE Full text] [doi: 10.2196/56997]
2. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024 Jan 23;26:e52113 [FREE Full text] [doi: 10.2196/52113] [Medline: 38261378]
3. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Zeitschrift Medizinische Ausbildung 2006;23(3):Doc53.
4. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 01;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]

Abbreviations

USMLE: United States Medical Licensing Examination

Edited by T Leung; submitted 26.02.24; this is a non–peer-reviewed article;accepted 04.04.24; published 16.04.24.

Please cite as:
Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, Zipfel S, Mahling M
Authors' Reply: "Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications"
J Med Internet Res 2024;26:e57778
URL: https://www.jmir.org/2024/1/e57778
doi:10.2196/57778
PMID:38625723

©Anne Herrmann-Werner, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, Moritz Mahling. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

How to Enhance Adherence in Patients with Nonalcoholic Fatty Liver Disease: A Thought-Provoking Issue

Chunlan Liu[1*], MSN; Run Zhou[1*], MSN; Junping Shi[2,3,4], MD, PhD

[1]School of Nursing, Hangzhou Normal University, Hangzhou, China

[2]Department of Infectious Disease and Hepatology, The Affiliated Hospital of Hangzhou Normal University, Hangzhou, China

[3]Zhejiang Key Laboratory of Medical Epigenetics, Hangzhou, China

[4]Institute of Hepatology and Metabolic Disease, Hangzhou Normal University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Junping Shi, MD, PhD
Department of Infectious Disease and Hepatology
The Affiliated Hospital of Hangzhou Normal University
Hangzhou Normal University Jinhua Road Campus
120 Jinhua Road
Hangzhou, 311121
China
Phone: 86 13957121199
Email: 20131004@hznu.edu.cn

Related Article:

Comment on: https://www.jmir.org/2024/1/e49839/

(J Med Internet Res 2024;26:e59208) doi:10.2196/59208

KEYWORDS

NAFLD; adherence; digital therapeutics; lifestyle modification; mobile health; nonalcoholic fatty liver disease; self-management; randomized controlled trial

We read with great interest the article by Kwon et al [1] in the *Journal of Medical Internet Research* that described an app for lifestyle coaching intervention for patients with nonalcoholic fatty liver disease (NAFLD). With the paradigm shift in medicine, we are delighted to find that the study took into account the importance of psychological and social factors for patients. Digital therapeutics have shown remarkable contributions in various fields [2]. Our team has done some research on the use of digital therapeutics in NAFLD [3]. We consider that the remote intervention developed by researchers based on the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) model can be regarded as a form of digital therapeutics. Having closely examined the research, we were inspired by several findings.

In this research, patients with adherence of 90% or more had improved BMI, liver fat score, and alanine transferase after 6 months of intervention compared to patients with adherence of less than 90%. In addition, participants in the higher adherence group also showed the highest performance in the areas of self-management and knowledge [1]. This shows that adherence plays a crucial role in the out-of-hospital treatment and management of NAFLD.

However, in terms of assessment, the researchers only evaluated participant adherence to the intervention through dietary recording rates (82.6% at 3 months and 79.8% at 6 months) [1]. We were concerned about whether this could give rise to a one-sided assessment. According to our knowledge, adherence assessments for mHealth interventions typically encompass a broader range of metrics, including completion of multifunctional modules, frequency of application visits, and the number of self-reported activities [4]. Therefore, in order to accurately assess patient adherence, researchers should update and refine subsequent iterations of the SMART-Liver app.

Furthermore, it is worth noting that while the average adherence was good, only 15 patients maintained adherence levels above 90% [1]. In addition, our team recently conducted a behavioral survey among 380 adults with NAFLD and found significant room for improvement in treatment adherence for the majority of patients [5]. With that in mind, here are some suggestions. On the one hand, researchers could personalize the management program by matching participants' characteristics such as age, gender, education, dietary preferences, and movement disorders before the intervention. This approach would tailor the program to individual needs and potentially enhance adherence. On the

XSL•FO
RenderX

other hand, fostering interest is essential for long-term adherence. Therefore, we suggest that the app be designed with attractive functional sections, such as artificial intelligence exercise coaches, virtual therapists, and interactive knowledge games. These features could engage users and motivate them to adhere to their treatment regimen more consistently [3].

This research represents a breakthrough in the out-of-hospital treatment and management of NAFLD. Having used the features of digital therapeutics to their fullest potential, researchers have to not only be free from the constraints of time and space but also bring substantial economic advantages to stakeholders. Undoubtedly, this development holds immense significance for NAFLD patients, health care professionals, and research and development companies.

Conflicts of Interest

None declared.

Editorial Notice

The corresponding author of "Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial" declined to respond to this letter.

References

1. Kwon OY, Lee MK, Lee HW, Kim H, Lee JS, Jang Y. Mobile app-based lifestyle coaching intervention for patients with nonalcoholic fatty liver disease: randomized controlled trial. J Med Internet Res 2024 Feb 15;26:e49839 [FREE Full text] [doi: 10.2196/49839] [Medline: 38358794]
2. Makin S. The emerging world of digital therapeutics. Nature 2019 Sep 25;573(7775):S106-S109. [doi: 10.1038/d41586-019-02873-1] [Medline: 31554990]
3. Zhou R, Gu Y, Zhang B, Kong T, Zhang W, Li J, et al. Digital therapeutics: emerging new therapy for nonalcoholic fatty liver disease. Clin Transl Gastroenterol 2023 Apr 01;14(4):e00575 [FREE Full text] [doi: 10.14309/ctg.0000000000000575] [Medline: 36854062]
4. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
5. Zhou R, Zhang B, Zhang W, Kong T, Fu J, Li J, et al. Self-management behaviours in adults with non-alcoholic fatty liver disease: a cross-sectional survey from China. BMJ Open 2024 Feb 22;14(2):e078333 [FREE Full text] [doi: 10.1136/bmjopen-2023-078333] [Medline: 38388505]

Abbreviations

ADDIE: Analysis, Design, Development, Implementation, Evaluation
NAFLD: nonalcoholic fatty liver disease

Edited by T Leung, L Beri; submitted 05.04.24; this is a non–peer-reviewed article;accepted 27.04.24; published 14.05.24.

Please cite as:
Liu C, Zhou R, Shi J
How to Enhance Adherence in Patients with Nonalcoholic Fatty Liver Disease: A Thought-Provoking Issue
J Med Internet Res 2024;26:e59208
URL: https://www.jmir.org/2024/1/e59208
doi:10.2196/59208
PMID:38743941

©Chunlan Liu, Run Zhou, Junping Shi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Mental Health Problems Among Children and Adolescents From a Sports Sociology Perspective

Yuan Li[1], MSc; Qun Zhai[1], PhD; Weihang Peng[2], MEng

[1]Faculty of Health Sciences and Sports, Macao Polytechnic University, Macao, Macao

[2]Faculty of Applied Sciences, Macao Polytechnic University, Macao, Macao

Corresponding Author:
Qun Zhai, PhD
Faculty of Health Sciences and Sports
Macao Polytechnic University
Comes Street
Macao, 999078
Macao
Phone: 853 68911106
Email: zqun@mpu.edu.mo

Related Articles:

Comment on: http://www.jmir.org/2024/1/e43994

Comment in: http://www.jmir.org/2024/1/e62775

(J Med Internet Res 2024;26:e60513) doi:10.2196/60513

KEYWORDS

sociology of sport; children; adolescents; mental health; systematic review; intervention; digital technology, parenting; technology; parenting program; engagement; support; adverse childhood experiences

We read the systematic review by Aldridge et al [1] recently published in the *Journal of Medical Internet Research* with interest. First, I would like to commend the author group for their efforts in publishing this article. The topic of the research is not only relevant for the current times but also showcases a high level of academic rigor in research methods and data analysis.

The study makes a particularly prominent contribution to mental health problems among adolescents and adverse experiences among children. The authors effectively provide a comprehensive evaluation of technology-assisted parenting programs through systematic review methods and address the challenges of adverse childhood experiences (ACEs) at the family level. The data analysis section deserves special mention, where the authors used the Stouffer method to combine *P* values, providing preliminary quantitative evidence for the association between specific engagement strategies and engagement outcomes.

Although the article is excellent in multiple aspects, there are some limitations. The heterogeneity in the definition and measurement of engagement and the lack of engagement outcome data were mentioned in the article, which may have limited the depth and breadth of the analysis.

Mental health problems among adolescents are closely related to ACEs. The field of sports sociology can play an important role in addressing this issue. From the perspective of sports sociology, parental involvement plays a crucial role in the development of children. Parents are not only their children's first coaches but also their life mentors, and their behaviors and attitudes have a profound impact on their children's mental health and participation in physical activity [2].

First, parents' support and encouragement can enhance their children's confidence and self-esteem, which is essential for their mental health. Parental involvement can also stimulate their children's interest in physical activity and help them establish positive exercise habits [3,4]. Regular participation in physical activity helps children maintain a healthy weight, prevent chronic diseases, and improve physical fitness. Similarly, positive family interactions and a supportive family environment can promote a child's overall development, including their physical health and social skills [5].

Future research can explore the impact of family physical activity on parenting behavior, how positive interactions among family members can be promoted through physical activity, and how these interactions positively affect the family environment from a sports sociology perspective. Moreover, we can further explore physical activity to reduce factors within the family that

XSL•FO
RenderX

may adversely affect child development, such as stress and conflict.

Overall, this article provides us with valuable knowledge and insights that make an important contribution to the development of the field of adolescent mental health problems and childhood adversity experiences. Despite some limitations, these are precisely the areas that future research can continue to explore and expand. By combining theories and methods from the sociology of sports, we can aim to find innovative solutions to current and future challenges.

Thank you for the opportunity to share our thoughts on the mental health problems of children and adolescents from a sports sociology viewpoint.

Conflicts of Interest

None declared.

References

1. Aldridge G, Tomaselli A, Nowell C, Reupert A, Jorm A, Yap MBH. Engaging parents in technology-assisted interventions for childhood adversity: systematic review. J Med Internet Res 2024 Jan 19;26:e43994 [FREE Full text] [doi: 10.2196/43994] [Medline: 38241066]
2. Wei C, Kendall PC. Parental involvement: contribution to childhood anxiety and its treatment. Clin Child Fam Psychol Rev 2014 Dec 15;17(4):319-339. [doi: 10.1007/s10567-014-0170-6] [Medline: 25022818]
3. Kohl HW3, Hobbs KE. Development of physical activity behaviors among children and adolescents. Pediatrics 1998 Mar;101(3 Pt 2):549-554. [Medline: 12224661]
4. Trost SG, Sallis JF, Pate RR, Freedson PS, Taylor WC, Dowda M. Evaluating a model of parental influence on youth physical activity. Am J Prev Med 2003 Nov;25(4):277-282. [doi: 10.1016/s0749-3797(03)00217-4] [Medline: 14580627]
5. Welk GJ, Wood K, Morss G. Parental influences on physical activity in children: an exploration of potential mechanisms. Pediatric Exercise Science 2003;15(1):19-33. [doi: 10.1123/pes.15.1.19]

Abbreviations

ACE: adverse childhood experience

Edited by T Leung, L Beri; submitted 14.05.24; this is a non–peer-reviewed article;accepted 07.06.24; published 03.07.24.

Please cite as:
Li Y, Zhai Q, Peng W
Mental Health Problems Among Children and Adolescents From a Sports Sociology Perspective
J Med Internet Res 2024;26:e60513
URL: https://www.jmir.org/2024/1/e60513
doi:10.2196/60513
PMID:

©Yuan Li, Qun Zhai, Weihang Peng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Author's Reply: Mental Health Problems Among Children and Adolescents From a Sports Sociology Perspective

Grace Aldridge[1], BPsych (Hons), PhD

Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton, Australia

Corresponding Author:
Grace Aldridge, BPsych (Hons), PhD
Turner Institute for Brain and Mental Health
School of Psychological Sciences
Monash University
18 Innovation Walk
Clayton Campus
Clayton, 3168
Australia
Phone: 61 39905 072
Email: grace.aldridge@monash.edu

Related Articles:

Comment on: https://www.jmir.org/2024/1/e60513

Comment on: https://www.jmir.org/2024/1/e43994

(J Med Internet Res 2024;26:e62775) doi:10.2196/62775

KEYWORDS

parenting interventions; technology; sports sociology; child mental health; adolescent mental health; adverse childhood experiences; systematic review; intervention; digital technology, parenting; parenting program; engagement; support

We thank the authors for their thoughtful response and comments on our systematic review [1]. As discussed in our paper, comprehensive evaluations (including quantitative syntheses) of engagement strategies and outcomes in technology-assisted parenting programs underline a significant gap in the parenting and child mental health literature. This gap is especially notable in the context of the COVID-19 pandemic, highlighting the clear utility technology has for maintaining or, at times, improving parents' engagement with services and programs to support their young person's mental health. While heterogeneous definitions, measurement of engagement, and a lack of engagement outcome data were indeed limitations to our review findings, we hope that our work provides a taxonomy of engagement strategies and measures that can help future work in the field to overcome these limitations.

The authors proposed that the field of sports sociology can play an important role in the prevention of mental health problems among adolescents. In particular, they outlined how family physical activity can promote a positive family environment, and how parenting behavior that supports and encourages family physical activity can promote young people's mental and physical health. Given the strong link between adverse childhood experiences (ACEs) and poorer mental and physical health outcomes, we agree that exploring theories and methods from a sports sociology perspective is a worthwhile avenue for innovative solutions to preventing and reducing the impact of ACEs on young people's mental health. Some technology-assisted parenting programs included in our study's analysis included a physical activity component, and we wish to offer the authors some insights that may inform future research.

Our review included findings from the "Grow" parenting program [2] that aims to teach parents about key factors for healthy child development (such as physical activity) and has demonstrated effectiveness in helping parents increase child outdoor playtime and meet health recommendations. Parents appeared to engage well with both face-to-face and online versions of the "Grow" program; however, greater parental satisfaction with the face-to-face version led the authors to propose hybrid program models for future research to consider [3]. Our review also included findings from the web-based "Parenting Resilient Kids" program [4] that includes a module on encouraging healthy habits such as promoting regular physical exercise. While parenting behaviors specific to physical activity were not independently evaluated, this program demonstrated effectiveness in improving target parenting behaviors covered across all modules selected by parents. Furthermore, parents' engagement with the program was shown

XSL•FO
RenderX

to predict child health-related quality of life outcomes at the 12-month follow-up [5].

We believe these findings highlight technology-assisted parenting programs as an innovative solution for integrating sports sociology in efforts to prevent young people's mental health problems. Given the link between parents' engagement with these programs and program outcomes, designing engaging programs is important for these solutions to be effective. We therefore propose that the authors and those contributing to future research in this space consider drawing on the taxonomy of engagement strategies and measures presented in our review when designing such programs, to conceptualize engagement that is evidence-informed and facilitates evaluation.

Conflicts of Interest

None declared.

References

1. Aldridge G, Tomaselli A, Nowell C, Reupert A, Jorm A, Yap MBH. Engaging parents in technology-assisted interventions for childhood adversity: systematic review. J Med Internet Res 2024 Jan 19;26:e43994 [FREE Full text] [doi: 10.2196/43994] [Medline: 38241066]
2. Chesnut R, DiNallo JM, Czymoniewicz-Klippel MT, Perkins DF. The Grow parenting program: demonstrating proof of concept. Health Education 2018 Jul 04;118(5):413-430. [doi: 10.1108/HE-01-2018-0005]
3. Czymoniewicz-Klippel M, Chesnut R, DiNallo J, Perkins D. Patterns of participation in the Grow parenting program. Journal of Children's Services 2019 Mar 18;14(1):27-41. [doi: 10.1108/jcs-06-2018-0014]
4. Sim WH, Fernando LMN, Jorm AF, Rapee RM, Lawrence KA, Mackinnon AJ, et al. A tailored online intervention to improve parenting risk and protective factors for child anxiety and depression: medium-term findings from a randomized controlled trial. J Affect Disord 2020 Dec 01;277:814-824. [doi: 10.1016/j.jad.2020.09.019] [Medline: 33065822]
5. Sim WH, Jorm AF, Yap MBH. The role of parent engagement in a web-based preventive parenting intervention for child mental health in predicting parenting, parent and child outcomes. Int J Environ Res Public Health 2022 Feb 15;19(4):2191 [FREE Full text] [doi: 10.3390/ijerph19042191] [Medline: 35206394]

Abbreviations

ACE: adverse childhood experience

Edited by T Leung, L Beri; submitted 31.05.24; this is a non–peer-reviewed article;accepted 07.06.24; published 03.07.24.

Please cite as:
Aldridge G
Author's Reply: Mental Health Problems Among Children and Adolescents From a Sports Sociology Perspective
J Med Internet Res 2024;26:e62775
URL: https://www.jmir.org/2024/1/e62775
doi:10.2196/62775
PMID:

©Grace Aldridge. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

The Power of Collaboration in Facilitating Mobile Technology Adoption in Health Care

Weijing Fan[1], MD; Guobin Liu[1], MD

Department of Peripheral Vascular Surgery, Institute of Surgery of Traditional Chinese Medicine, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China

Corresponding Author:
Guobin Liu, MD
Department of Peripheral Vascular Surgery, Institute of Surgery of Traditional Chinese Medicine
Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine
Zhangheng road 528
Shanghai, 201203
China
Phone: 86 15800885533
Email: 15800885533@163.com

Related Articles:

Comment on: http://www.jmir.org/2024/1/e53651/

Comment in: http://www.jmir.org/2024/1/e62891/

(J Med Internet Res 2024;26:e60350) doi:10.2196/60350

KEYWORDS

social workers; government; mobile technology; mobile apps; fitness trackers; SMS; SMS text messaging; physical activity; exercise; sedentary behavior; Middle East; Africa; movement; physical inactivity; smartphone; mobile phone; mobile health; mHealth; digital health; behavior change; intervention

We read the recently published article titled “The Use of Mobile Technologies to Promote Physical Activity and Reduce Sedentary Behaviors in the Middle East and North Africa Region: Systematic Review and Meta-Analysis” by Tong HL et al [1] with great interest. The authors explored the impact of mobile technologies on physical activity and sedentary habits in the Middle East and North Africa (MENA) region. Even though solid conclusions were not obtained due to insufficient number and quality of the included studies, the analysis yielded some promising insights. However, several aspects merit further exploration.

Primarily, the authors did not search the commonly used databases PubMed, Web of Science, and Cochrane in their systematic review [2]. An explanation for how to ensure completeness of the search is thus needed. Second, behavior change is usually a gradual process of change, so the frequency and duration of the intervention have an important impact on the outcome of behavior change; too much repetitive information can lead to user fatigue and ignored messages, and the authors did not analyze the relevant information. Third, the authors should conduct a more specific analysis of potential confounding factors that may affect the outcome, as well as variables such as the nature of work, family support, and education that significantly affect the actual effect of mobile technologies [3]. Without controlling for these variables, attributing their effects to mobile technologies will be very challenging.

Finally, the authors emphasize the crucial role that governments play in the context of mobile technologies—a point well-taken. Social workers also play a key part in this process. Many countries in the MENA region have vast differences in terms of economy, politics, and culture [4]; these countries can leverage their understanding of local economics, culture, and health determinants to advocate for policies that boost the uptake of mobile technologies uptake. Additionally, social workers can foster dialogue between medical professionals and government officials, offering social insights to enhance the deployment of mobile technologies. We look forward to seeing governments, social workers, doctors, and families collaborate to leverage mobile technologies for improved public health management.

XSL•FO
RenderX

Authors' Contributions

WF contributed to the methodology, formal analysis, and writing of the original draft. GL contributed to the conceptualization, methodology, supervision, writing, as well as review and editing of the draft.

Conflicts of Interest

None declared.

References

1. Tong HL, Alnasser A, Alshahrani NZ, Bawaked RA, AlAhmed R, Alsukait RF, et al. The use of mobile technologies to promote physical activity and reduce sedentary behaviors in the Middle East and North Africa region: systematic review and meta-analysis. J Med Internet Res 2024 Mar 19;26:e53651 [FREE Full text] [doi: 10.2196/53651] [Medline: 38502160]
2. Trivedi R, Elshafie S, Tackett R, Young H, Sattler ELP. Effectiveness and feasibility of telehealth-based dietary interventions targeting cardiovascular disease risk factors: systematic review and meta-analysis. J Med Internet Res 2024 Feb 16;26:e49178 [FREE Full text] [doi: 10.2196/49178] [Medline: 38363635]
3. Al Ansari FS, Alfayez A, Alsalman D, Alanezi F, Alhodaib H, Al-Rayes S, et al. Using mobile health applications to enhance physical activity in Saudi Arabia: a cross-sectional study on users' perceptions. Int Health 2023 Jan 03;15(1):47-55 [FREE Full text] [doi: 10.1093/inthealth/ihac008] [Medline: 35348719]
4. Middle East and North Africa. World Bank Group. URL: https://www.worldbank.org/en/region/mena [accessed 2023-04-24]

Abbreviations

MENA: Middle East and North Africa

Edited by T Leung, L Beri; submitted 08.05.24; this is a non–peer-reviewed article;accepted 07.06.24; published 08.07.24.

Please cite as:
Fan W, Liu G
The Power of Collaboration in Facilitating Mobile Technology Adoption in Health Care
J Med Internet Res 2024;26:e60350
URL: https://www.jmir.org/2024/1/e60350
doi:10.2196/60350
PMID:

©Weijing Fan, Guobin Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Authors' Reply: The Power of Collaboration in Facilitating Mobile Technology Adoption in Health Care

Huong Ly Tong[1], PhD; Severin Rakic[2], PhD; Hazzaa M Al-Hazzaa[3,4], PhD; Saleh A Alqahtani[5,6], PhD

[1]Cultural and Indigenous Research Centre Australia, Redfern, Australia
[2]The World Bank Group, Washington, DC, United States
[3]Lifestyle and Health Research Center, Health Sciences Research Center, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[4]School of Sport Sciences, University of Jordan, Amman, Jordan
[5]Organ Transplant Center of Excellence, King Faisal Specialist Hospital & Research Center, Riyadh, Saudi Arabia
[6]Division of Gastroenterology and Hepatology, Weill Cornell Medicine, New York, NY, United States

Corresponding Author:
Severin Rakic, PhD
The World Bank Group
J5-068, 1818 H Street NW
Washington, DC, 20433
United States
Phone: 1 2024585552
Email: srakic@worldbank.org

Related Articles:

Comment on: http://www.jmir.org/2024/1/e60350/

Comment on: http://www.jmir.org/2024/1/e53651/

(J Med Internet Res 2024;26:e62891) doi:10.2196/62891

KEYWORDS

mobile apps; fitness trackers; SMS; SMS text messaging; physical activity; exercise; sedentary behavior; Middle East; Africa, Northern; movement; physical inactivity; smartphone; mobile phone; mobile health; mHealth; digital health; behavior change; intervention

We are responding on behalf of our author group to a letter to the editor [1] regarding our recent article titled "The Use of Mobile Technologies to Promote Physical Activity and Reduce Sedentary Behaviors in the Middle East and North Africa Region: Systematic Review and Meta-Analysis" [2], published in the *Journal of Medical Internet Research*. We thank the letter's authors for their interest in our article and the journal editor for providing us the opportunity to respond. Our response is limited to the 4 aspects of critique in the letter.

First, regarding the database search, we agree with the letter's authors that the more databases searched, the more comprehensive the results would be. In our review, we conducted a search across 5 major databases, including MEDLINE, Embase, CINAHL, Scopus, and Global Index Medicus. This is in line with the standard practices outlined by the Cochrane Handbook for Systematic Reviews of Interventions, which recommends using at least 3 databases [3]. Given the focus of the review on the Middle East and North Africa, we used Global Index Medicus for our search to enhance comprehensiveness. Moreover, a gray literature search was also conducted, in Google Scholar, to ensure that we identified the relevant studies. Paradoxically, the letter's authors referred to a systematic review [4] that used fewer databases in their search than we did to argue the database search point.

Second, we agree that frequency and duration of the intervention are important aspects. Indeed, we analyzed this information and discussed it in Tables 1 and 2 of our article, as well as in the main text: "The average duration of experimental studies was 20 (SD 14.4; range 6-52) weeks." and "Half (11/22, 50%) of the interventions used SMS text messaging to deliver educational content…; the frequency of delivery varied from 2 messages per day to 1 message per week" [1].

Third, we agree with the letter's authors that potential confounding factors can influence the findings of the reported studies. However, we found that important confounders were not reported in the many studies reviewed. While we planned to explore the role of confounders via planned meta-regression [5], it would have been methodologically flawed to do so, given that the meta-analysis would have been based on 7 studies only. At least 10 studies should be available for each characteristic

XSL•FO
RenderX

modeled [3]. The influence of some reported confounders was discussed in the article, including socioeconomic status, context fit, and cultural appropriateness.

Finally, we agree with the authors that social workers can facilitate mobile technologies deployment. However, exploration of the social workers' role in this process was not an objective of the review nor a primary focus of the articles included in our review.

We would welcome future publications that discuss specifically the role of social workers in the deployment of mobile technologies.

Conflicts of Interest

None declared.

References

1. Fan W, Liu G. The power of collaboration in facilitating mobile technology adoption in health care. J Med Internet Res 2024. [doi: 10.2196/60350]
2. Tong HL, Alnasser A, Alshahrani NZ, Bawaked RA, AlAhmed R, Alsukait RF, et al. The use of mobile technologies to promote physical activity and reduce sedentary behaviors in the Middle East and North Africa region: systematic review and meta-analysis. J Med Internet Res 2024 Mar 19;26:e53651 [FREE Full text] [doi: 10.2196/53651] [Medline: 38502160]
3. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester, UK: John Wiley & Sons; 2019.
4. Trivedi R, Elshafie S, Tackett R, Young H, Sattler ELP. Effectiveness and feasibility of telehealth-based dietary interventions targeting cardiovascular disease risk factors: systematic review and meta-analysis. J Med Internet Res 2024 Feb 16;26:e49178 [FREE Full text] [doi: 10.2196/49178] [Medline: 38363635]
5. Al-Hazzaa H, Alnasser A, AlQahtani S, Alshahrani N, Alsukait R, Bawaked R, et al. The use of mobile technologies to promote physical activity and reduce sedentary behaviors in the Middle East and North Africa region: A systematic review and meta-analysis. PROSPERO 2023 CRD42023392699. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023392699 [accessed 2024-06-14]

Edited by L Beri, T Leung; submitted 04.06.24; this is a non–peer-reviewed article;accepted 07.06.24; published 08.07.24.

Please cite as:
Tong HL, Rakic S, Al-Hazzaa HM, Alqahtani SA
Authors' Reply: The Power of Collaboration in Facilitating Mobile Technology Adoption in Health Care
J Med Internet Res 2024;26:e62891
URL: https://www.jmir.org/2024/1/e62891
doi:10.2196/62891
PMID:

©Huong Ly Tong, Severin Rakic, Hazzaa M Al-Hazzaa, Saleh A Alqahtani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Considerations and Challenges in the Application of Large Language Models for Patient Complaint Resolution

Bin Wei[1], MD; Xin Hu[1], MD; XiaoRong Wu[1], MD

The 1st Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, China

Corresponding Author:
XiaoRong Wu, MD
The 1st Affiliated Hospital
Jiangxi Medical College
Nanchang University
No. 17 Yongwai Zheng Street
Donghu District
Nanchang, 330000
China
Phone: 86 13617093259
Email: wxr98021@126.com

Related Article:

Comment on: https://www.jmir.org/2024/1/e56413

(J Med Internet Res 2024;26:e65527) doi:10.2196/65527

KEYWORDS

ChatGPT; large language model; LLM; artificial intelligence; AI; patient complaint; empathy; efficiency; patient satisfaction; resource allocation

We are writing to express our appreciation for the recent publication of the study titled "Performance of Large Language Models in Patient Complaint Resolution: Web-Based Cross-Sectional Survey" in the *Journal of Medical Internet Research* [1]. The authors provide an innovative perspective on the application of large language models (LLMs) like ChatGPT in resolving patient complaints, demonstrating their potential to enhance patient satisfaction through thoughtful and well-constructed responses. Although we greatly appreciate this study's meticulous work and significant contributions, we would like to highlight additional considerations and potential limitations for future research.

First, the decision-making process of LLMs is often a "black box." While these models can generate reasonable and satisfactory responses, the underlying logic and reasoning remain unknown [2]. This lack of transparency and interpretability could pose challenges to the legal responsibilities of health care institutions and potentially undermine patient trust. Additionally, when artificial intelligence (AI) handles sensitive patient information, even with deidentification measures, ensuring data security and privacy remains a critical concern. Therefore, as the application of LLMs expands, establishing appropriate regulatory frameworks and accountability mechanisms may be necessary.

Second, this study may have been conducted within the English-speaking health care system of Singapore, but patient complaints and requirements in different regions heavily depend on local cultural, legal, and social contexts. LLMs may struggle to capture and differentiate these subtle cultural nuances, potentially leading to responses that are not fully appropriate or that even cause misunderstandings. Moreover, while AI models can generate responses that appear empathetic, such responses are merely based on the model's imitation of language patterns, rather than a true understanding of the patient's psychological state and emotional needs [3,4]. In the tense environment of doctor-patient relationships in China, if patients perceive a difference in responses, it could exacerbate their anxiety and dissatisfaction. In certain circumstances, using AI to handle patient complaints may lack the necessary human touch, thereby intensifying conflict between patients and health care providers.

Third, over time, LLMs need continuous updates to reflect the latest medical practices, policy changes, and societal expectations, which undoubtedly increases maintenance costs. If AI models generate responses based on incorrect or outdated medical information, they may unintentionally mislead patients, particularly in critical medical decision-making and health guidance, which could be especially dangerous. Furthermore, given the varying levels of health care across different regions,

XSL•FO
RenderX

and the differing approaches to treatment, this variability presents a significant challenge for AI.

In conclusion, while LLMs like ChatGPT show great potential in enhancing patient communication and handling complaints, their practical application requires careful consideration of their suitability and the challenges they may present. Therefore, establishing robust regulatory frameworks, ethical guidelines, and stringent oversight mechanisms is crucial. This approach not only helps build trust between patients and health care providers but also contributes to creating a more patient-centered and efficient health care system.

Conflicts of Interest

None declared.

Editorial Notice

The corresponding author of "Performance of Large Language Models in Patient Complaint Resolution: Web-Based Cross-Sectional Survey" declined to respond to this letter.

References

1. Yong LPX, Tung JYM, Lee ZY, Kuan WS, Chua MT. Performance of large language models in patient complaint resolution: web-based cross-sectional survey. J Med Internet Res 2024 Aug 09;26:e56413 [FREE Full text] [doi: 10.2196/56413] [Medline: 39121468]
2. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual Event p. 610-623. [doi: 10.1145/3442188.344592]
3. D'Alfonso S, Santesteban-Echarri O, Rice S, Wadley G, Lederman R, Miles C, et al. Artificial intelligence-assisted online social therapy for youth mental health. Front Psychol 2017;8:796-720 [FREE Full text] [doi: 10.3389/fpsyg.2017.00796] [Medline: 28626431]
4. Hoermann S, McCabe KL, Milne DN, Calvo RA. Application of synchronous text-based dialogue systems in mental health interventions: systematic review. J Med Internet Res 2017 Jul 21;19(8):e267 [FREE Full text] [doi: 10.2196/jmir.7023] [Medline: 28784594]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by T Leung, L Beri; submitted 18.08.24; this is a non–peer-reviewed article;accepted 23.08.24; published 17.09.24.

Please cite as:
Wei B, Hu X, Wu X
Considerations and Challenges in the Application of Large Language Models for Patient Complaint Resolution
J Med Internet Res 2024;26:e65527
URL: https://www.jmir.org/2024/1/e65527
doi:10.2196/65527
PMID:

©Bin Wei, Xin Hu, XiaoRong Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Expanding the Scope: Reflections on Digital Smoking Cessation Strategies for Diverse Age Groups

Bin Wei[1], MD; Xin Hu[1], MD; XiaoRong Wu[1], MD

The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, Jiangxi, China

Corresponding Author:
XiaoRong Wu, MD
The First Affiliated Hospital
Jiangxi Medical College
Nanchang University
No. 17, Yongwai Zhengjie
Donghu District
Nanchang, Jiangxi, 330000
China
Phone: 86 13617093259
Email: wxr98021@126.com

Related Articles:

Comment on: http://www.jmir.org/2024/1/e52919/

Comment in: http://jmir.org/2024/1/e67749/

(J Med Internet Res 2024;26:e65929) doi:10.2196/65929

KEYWORDS

digital smoking cessation; age group comparisons; behavioral health interventions; older adults; digital cessation treatment; cigarettes; tobacco; quit; telehealth; behavioral health; public health

We are writing to express our appreciation for the recent publication of the study entitled, “Expectations and Preferences for Digital Cessation Treatment: Multimethods Study Among Older Adults Who Smoke Cigarettes” in the *Journal of Medical Internet Research* [1]. I commend the authors for their efforts in exploring smoking cessation treatments within the older adult population and appreciate their focus on this often-overlooked demographic group. However, as I perused the article, I had a few concerns and suggestions that I would like to share with the authors and the editorial team to aid in the enhancement of future research.

First, the authors selected individuals aged 65 years and older as the focus of their study, specifically exploring their attitudes toward digital smoking cessation interventions. This choice is crucial, as older adults are at higher risk for tobacco-related health issues [2]. However, we believe that expanding the scope of the study to include younger populations, such as middle-aged adults, young adults, and adolescents, is equally important. Younger individuals also need effective smoking cessation tools, and their greater adaptability to digital technologies might allow them to benefit even more [3]. Future research could consider comparing the effectiveness of digital smoking cessation treatments across different age groups to gain a more comprehensive understanding of their applicability in various populations.

Moreover, for older adults with a long history of smoking, quitting can be an exceptionally challenging process. While digital smoking cessation tools offer convenience and accessibility, it remains to be seen whether these tools can effectively help older individuals overcome decades-long nicotine addiction. Future research might consider conducting long-term follow-ups to assess the sustained impact of digital cessation tools on older adults and explore whether they might benefit from additional support, such as in-person counseling or community-based programs, to enhance their chances of successfully quitting smoking.

Older adults in the process of quitting smoking require not only behavioral support but also emotional and social assistance. Whether digital smoking cessation tools can meet these more complex needs remains an open question. As a social worker, I believe that integrating more social and emotional support features into digital platforms, alongside traditional face-to-face counseling, could potentially offer a more comprehensive approach to smoking cessation support.

Overall, this article offers highly valuable insights into smoking cessation approaches for older adults, which can foster further

XSL•FO
RenderX

exploration in the field of digital smoking cessation research. We would like to extend our gratitude once again to the authors for their efforts and to the editorial team for providing a platform to publish this important work.

Conflicts of Interest

None declared.

References

1. Fahey MC, Carpenter MJ, O'Neal R, Pebley K, Schick MR, Ware E, et al. Expectations and preferences for digital cessation treatment: multimethods study among older adults who smoke cigarettes. J Med Internet Res 2024 Aug 28;26:e52919 [FREE Full text] [doi: 10.2196/52919] [Medline: 39196628]
2. Schnack H, Griessen R, Lensink J, Wen HH. Generalized inversion scheme for the determination of activation energies from flux-creep experiments in high-Tc superconductors. Phys Rev B Condens Matter 1993 Nov 01;48(17):13178-13181. [doi: 10.1103/physrevb.48.13178] [Medline: 10007706]
3. Rowley R, Phillips EN, Schroeder AL. The effects of ionizing radiation on DNA synthesis in eukaryotic cells. Int J Radiat Biol 1999 Mar;75(3):267-283. [doi: 10.1080/095530099140456] [Medline: 10203178]

Edited by T Leung, L Beri; submitted 29.08.24; this is a non–peer-reviewed article;accepted 28.10.24; published 18.11.24.

Please cite as:
Wei B, Hu X, Wu X
Expanding the Scope: Reflections on Digital Smoking Cessation Strategies for Diverse Age Groups
J Med Internet Res 2024;26:e65929
URL: https://www.jmir.org/2024/1/e65929
doi:10.2196/65929
PMID:

©Bin Wei, Xin Hu, XiaoRong Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Author's Reply: Expanding the Scope: Reflections on Digital Smoking Cessation Strategies for Diverse Age Groups

Margaret C Fahey[1], PhD

Department of Psychology, Middle Tennessee State University, Murfreesboro, TN, United States

Corresponding Author:
Margaret C Fahey, PhD
Department of Psychology
Middle Tennessee State University
Academic Classroom Building
Murfreesboro, TN, 37130
United States
Phone: 1 615 898 5949
Email: mfahey@mtsu.edu

Related Articles:

Comment on: http://www.jmir.org/2024/1/e65929/

Comment on: http://www.jmir.org/2024/1/e52919/

(J Med Internet Res 2024;26:e67749) doi:10.2196/67749

KEYWORDS

digital smoking cessation; age group comparisons; behavioral health intervention; older adult; cigarette; tobacco; quitting; telehealth; behavioral health; public health

We thank the authors for their positive response to our paper [1], which provided thoughtful comments and suggestions for future research [2]. First, the authors mention the importance of expanding the scope of our research to include younger age groups. We agree that smoking cessation is an important public health initiative across the entire lifespan. We focused on an older adult (≥65 years) population not to discount the importance of expanding digital cessation treatment among younger adults, but rather to focus on an underserved population in the field of tobacco cessation. For decades, older adults have been ignored by antitobacco public health initiatives, given a focus on tobacco prevention in adolescence and ageist misconceptions that cessation does not benefit health in later life [3]. Hence, older adults have historically been less likely to receive evidence-based cessation treatment [3]. As the field of tobacco cessation treatment is evolving to include digital platforms, older adults continue to be ignored. To our knowledge, no digital cessation treatment is tailored to the unique needs of this age group, despite numerous programs tailored specifically for adolescents or young adults [4]. Although younger age groups are generally more likely to interact with technology, older adults are using technology for their health at increasing rates [5] and should not be excluded from digital cessation treatment research. Our findings are hypothesis-generating and provide recommendations for being more inclusive of older age groups in the development of these treatments. Further, we hope that our study highlights an underrepresented voice that needs better inclusion in the field of tobacco cessation.

Second, the authors suggest that integrating social and emotional support in digital platforms, alongside traditional face-to-face counseling, could offer a comprehensive approach to cessation for this age group. We agree that incorporating social components, such as group chats within app-based programs or telehealth group counseling, might be appealing and beneficial approaches for older adults. When prompted about group-based components in digital platforms, our study population believed that the benefits of doing so would include learning new strategies for quitting and connecting with other older adults attempting to quit cigarette smoking. However, our sample discussed concerns about interpersonal challenges (eg, conflicts and negativity) among individuals, which many had witnessed on social media platforms. Perhaps, moderated or asynchronous interactions in group-based digital platforms might be more appealing for this age group. However, given ageist misconceptions that older adults are unable or unwilling to quit cigarettes [3], this population might uniquely benefit from interacting with others in their age group who are motivated to quit smoking. We thank the authors for this comment and believe that future research could consider evaluating the effectiveness of group-based digital cessation treatment for older age groups. In summary, our study challenges the bias that older adults are unwilling or uninterested in

XSL•FO
RenderX

engaging with digital platforms to aid with smoking cessation. We are encouraged by the commentary that our study is eliciting, and we hope that researchers and clinicians working with older adults might benefit from our findings.

Conflicts of Interest

None declared.

References

1. Fahey MC, Carpenter MJ, O'Neal R, Pebley K, Schick MR, Ware E, et al. Expectations and preferences for digital cessation treatment: multimethods study among older adults who smoke cigarettes. J Med Internet Res 2024 Aug 28;26:e52919 [FREE Full text] [doi: 10.2196/52919] [Medline: 39196628]
2. Wei B, Hu X, Wu X. Expanding the scope: Reflections on digital smoking cessation strategies for diverse age groups. J Med Internet Res 2024:e65929. [doi: 10.2196/65929]
3. McAfee T, Malone RE, Cataldo J. Ignoring our elders: tobacco control's forgotten health equity issue. Tob Control 2021 Sep 12;30(5):479-480. [doi: 10.1136/tobaccocontrol-2021-056945] [Medline: 34389689]
4. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. Cochrane Database Syst Rev 2019 Oct 22;10(10):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub5] [Medline: 31638271]
5. Fahey MC, Huang S, Dahne J. Technology use for health-related purposes among US older adults (65+ years). J Am Geriatr Soc 2024 Aug 19. [doi: 10.1111/jgs.19122] [Medline: 39158124]

Edited by L Beri, T Leung; submitted 20.10.24; this is a non–peer-reviewed article;accepted 28.10.24; published 18.11.24.

Please cite as:
Fahey MC
Author's Reply: Expanding the Scope: Reflections on Digital Smoking Cessation Strategies for Diverse Age Groups
J Med Internet Res 2024;26:e67749
URL: https://www.jmir.org/2024/1/e67749
doi:10.2196/67749
PMID:

©Margaret C Fahey. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

The Vast Potential of ChatGPT in Pediatric Surgery

Ran Tang[1], MD; Shi-qin Qi[1], MD

Department of Pediatric Surgery, Anhui Provincal Children's Hospital, Hefei, China

Corresponding Author:
Shi-qin Qi, MD
Department of Pediatric Surgery
Anhui Provincal Children's Hospital
39 Wangjiang Street
Hefei, 230051
China
Phone: 86 13637080508
Email: qishiqin@163.com

Related Article:

Comment on: https://www.jmir.org/2024/1/e54985

(J Med Internet Res 2024;26:e66453) doi:10.2196/66453

KEYWORDS

ChatGPT; pediatric; surgery; artificial intelligence; AI; diagnosis; surgeon

We came across the publication "The Diagnostic Ability of GPT-3.5 and GPT-4.0 in Surgery: Comparative Analysis" [1] in the *Journal of Medical Internet Research* by Liu and colleagues. As highlighted in this article, the diagnostic prowess of artificial intelligence (AI) systems like GPT-4 continues to advance rapidly, showcasing remarkable improvements over earlier models [2]. These findings emphasize the potential of AI in diagnosing colorectal cancer and provide insights into its possible applications in pediatric surgery.

Pediatric surgery, distinct from adult surgery, presents unique challenges: children's developmental and anatomical differences, a limited disease spectrum, and the delicate balance between surgical trauma and recovery. Moreover, communication barriers with young patients often place a greater burden on clinicians for accurate diagnosis and effective care planning. In areas like Anhui Province, China, which is home to over 60 million people but served by only one pediatric center, nonspecialist doctors must often perform surgeries, heightening the risk of complications.

The application of AI in pediatric surgery has begun to bridge these gaps. GPT-4's capabilities in multimodal data integration, such as processing medical records, images, and lab results, offer substantial potential in preoperative planning, providing tailored insights that pediatric surgeons can use to navigate complex cases. Furthermore, GPT-4's utility in enhancing surgical education is noteworthy. Through integration with 3D models and its conversational abilities, AI can offer virtual simulations and detailed anatomical explanations, greatly benefiting students and junior doctors who need hands-on experience in congenital anomalies and tumor resections.

Postoperative care is another critical area where AI can provide substantial support. As children are less able to articulate their needs, GPT-4's real-time data processing and monitoring abilities can aid in identifying early signs of complications and offer personalized recovery plans. In resource-limited regions, AI can facilitate remote collaboration, linking local hospitals with specialists in pediatric centers, ultimately improving care outcomes.

Liu et al [1] emphasized the potential of AI in refining diagnostic tools. However, it is important to note that AI lacks the clinical intuition required in many pediatric surgery cases. While AI systems like GPT-4 provide valuable assistance, they cannot replace the clinical judgment of experienced pediatric surgeons, especially when dealing with complex or nonstandard scenarios.

In conclusion, GPT-4 offers a promising future for pediatric surgery, from improving diagnostic accuracy and surgical education to enhancing postoperative care. However, its limitations in clinical decision-making highlight the need for careful integration of AI technologies into practice, with clear guidelines to ensure ethical and effective use. With proper regulation, GPT-4 can become an indispensable tool in pediatric surgery, particularly in regions with limited access to specialized care.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Editorial Notice

The corresponding author of "The Diagnostic Ability of GPT-3.5 and GPT-4.0 in Surgery: Comparative Analysis" declined to respond to this letter.

References

1. Liu J, Liang X, Fang D, Zheng J, Yin C, Xie H, et al. The diagnostic ability of GPT-3.5 and GPT-4.0 in surgery: comparative analysis. J Med Internet Res 2024 Sep 10;26:e54985 [FREE Full text] [doi: 10.2196/54985] [Medline: 39255016]
2. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg 2018 Jul;268(1):70-76 [FREE Full text] [doi: 10.1097/SLA.0000000000002693] [Medline: 29389679]

Abbreviations

AI: artificial intelligence

Edited by T Leung, L Beri; submitted 13.09.24; this is a non–peer-reviewed article;accepted 31.10.24; published 18.11.24.

Please cite as:
Tang R, Qi SQ
The Vast Potential of ChatGPT in Pediatric Surgery
J Med Internet Res 2024;26:e66453
URL: https://www.jmir.org/2024/1/e66453
doi:10.2196/66453
PMID:

©Ran Tang, Shi-qin Qi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Reassessing AI in Medicine: Exploring the Capabilities of AI in Academic Abstract Synthesis

Zijian Wang[1], MS; Chunyang Zhou[2], MS

[1]Shandong University of Traditional Chinese Medicine, College of Traditional Chinese Medicine, Jinan, China

[2]Department of Radiation Oncology, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, Qingdao, China

Corresponding Author:
Chunyang Zhou, MS
Department of Radiation Oncology, Qilu Hospital (Qingdao)
Cheeloo College of Medicine
Shandong University
758 Hefei Road, Qingdao
Qingdao, 266000
China
Phone: 86 18561813085
Email: chunyangzhou29@163.com

Related Articles:

Comment on: http://www.jmir.org/2023/1/e51229/

Comment in: http://www.jmir.org/2024/1/e65123/

Abstract

(J Med Internet Res 2024;26:e55920) doi:10.2196/55920

KEYWORDS

ChatGPT; AI-generated scientific content; plagiarism; AI; artificial intelligence; NLP; natural language processing; LLM; language model; text; textual; generation; generative; extract; extraction; scientific research; academic research; publication; abstract; comparative analysis; reviewer bias

In their study, Cheng and colleagues [1] delved into a comparison between academic abstracts generated by ChatGPT and those written by humans. Their aim was to evaluate the applicability of an artificial intelligence (AI) model, specifically ChatGPT 3.5, in generating abstracts for basic preclinical research. This investigation, involving 30 articles from renowned journals, revealed that the quality of ChatGPT-generated abstracts generally falls short compared to human-written counterparts, particularly in unstructured formats, occasionally leading to erroneous conclusions. The study also concluded that such AI-generated abstracts can be identified by humans. However, several aspects of this study invite further discussion.

First, the selection process for the experts involved in the study was not explicitly detailed, potentially leading to subjective biases. These biases could arise from the experts' personal preferences, academic or cultural backgrounds, or their familiarity with the subject matter. Such biases might distort the results, causing the quality of AI-generated abstracts to be overestimated or underestimated.

Second, the study used ChatGPT 3.5 as the primary research model. However, there are concerns that ChatGPT 3.5 might not be ideally suited for processing lengthy texts, potentially leading to misunderstandings or the provision of inaccurate information [2]. Contrastingly, its accuracy might surpass that of humans when dealing with short and concise texts [3,4]. We propose that employing ChatGPT in a "layered progressive" manner for text generation could address this issue. This method involves dividing an article into smaller sections, having ChatGPT summarize each section individually, and then compiling these summaries into a cohesive whole. Such an approach is likely to yield better results than generating a summary from the entire text.

Lastly, the study concluded that blinded reviewers could accurately identify ChatGPT-generated abstracts with an

XSL•FO
RenderX

accuracy of 93%. Does this conclusion actually hold practical significance? Does the realization that an abstract is AI generated influence the reviewers' judgment of its content? In scientific research, the primary goal of writing a paper is to present research results to the readers. If AI assistance in refining an article does not alter the accuracy of the research and aids in clearer communication of research content, such practice should be deemed acceptable. A paper generated by AI and reviewed by humans, conveying high-quality research findings, deserves acceptance without bias toward its AI-generated origin. Therefore, what needs to be avoided is the complete reliance on AI language models for labor-intensive tasks, as demonstrated in the study, where an entire article is entrusted to AI. This could lead to a lack of critical thinking and the production of low-quality or erroneous articles by human researchers. We must recognize the importance of AI language models as assistants in our writing endeavors. They can inspire our ideas and help refine our text (as researchers often lack the literary flair of writers). If the authors of the article consider the above suggestions, the conclusions of their study might change.

Conflicts of Interest

None declared.

References

1. Cheng S, Tsai S, Bai Y, Ko C, Hsu C, Yang F, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e51229 [FREE Full text] [doi: 10.2196/51229] [Medline: 38145486]
2. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023 Jul 13;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
3. Herbold S, Hautli-Janisz A, Heuer U, Kikteva Z, Trautsch A. A large-scale comparison of human-written versus ChatGPT-generated essays. Sci Rep 2023 Oct 30;13(1):18617 [FREE Full text] [doi: 10.1038/s41598-023-45644-9] [Medline: 37903836]
4. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]

Abbreviations

AI: artificial intelligence

Edited by T Leung, L Beri; submitted 29.12.23; this is a non–peer-reviewed article;accepted 27.09.24; published 23.12.24.

Please cite as:
Wang Z, Zhou C
Reassessing AI in Medicine: Exploring the Capabilities of AI in Academic Abstract Synthesis
J Med Internet Res 2024;26:e55920
URL: https://www.jmir.org/2024/1/e55920
doi:10.2196/55920
PMID:

©Zijian Wang, Chunyang Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Letter to the Editor

Authors' Reply: Reassessing AI in Medicine: Exploring the Capabilities of AI in Academic Abstract Synthesis

Tien-Wei Hsu[1,2], MD; Chih-Sung Liang[3,4], MD

[1]Department of Psychiatry, E-Da Dachang Hospital, I-Shou University, Kaohsiung, Taiwan
[2]Department of Psychiatry, E-Da Hospital, I-Shou University, Kaohsiung, Taiwan
[3]Department of Psychiatry, Tri-service Hospital, Beitou Branch, Taipei, Taiwan
[4]Department of Psychiatry, National Defense Medical Center, Taipei, Taiwan

Corresponding Author:
Chih-Sung Liang, MD
Department of Psychiatry
Tri-service Hospital, Beitou Branch
No. 60, Xinmin Road
Beitou District
Taipei, 112
Taiwan
Phone: 886 2 28959808
Email: lcsyfw@gmail.com

Related Articles:

Comment on: http://www.jmir.org/2024/1/e55920/

Comment on: http://www.jmir.org/2023/1/e51229/

(J Med Internet Res 2024;26:e65123) doi:10.2196/65123

KEYWORDS

ChatGPT; AI-generated scientific content; plagiarism; AI; artificial intelligence; NLP; natural language processing; LLM; language model; text; textual; generation; generative; extract; extraction; scientific research; academic research; publication; abstract; comparative analysis; reviewer bias

We appreciate Wang and Zhou's [1] keen interest in our publication [2]. Below are our responses to their comments.

First, Wang and Zhou [1] argued potential biases could arise from the 8 experts' personal preferences, academic or cultural backgrounds, or their familiarity with the subject matter. However, the 30 papers we selected were from 3 different journals and explored various topics (as outlined in Table S3 in Multimedia Appendix 1 of our paper [2]). Therefore, these biases are unlikely to occur, considering the significant variation among the articles.

Second, we did not use ChatGPT 3.5. We used ChatPDF, which is based on ChatGPT 3.5 but can efficiently analyze PDF content.

Third, Wang and Zhou argued that if artificial intelligence (AI) assistance in refining an article does not alter the accuracy of the research and aids in clearer communication of the content, such a practice should be deemed acceptable. However, in our study, we found that 3 of the 30 ChatGPT-generated abstracts contained incorrect conclusions.

Finally, Wang and Zhou are right in cautioning against complete reliance on AI language models for labor-intensive tasks, which may lead to a lack of critical thinking and the production of low-quality or erroneous articles. However, their perspective differs from the purpose of our study. We aimed to assess the applicability of an AI model in generating abstracts for basic preclinical research. We found that the quality of ChatGPT-generated abstracts for basic preclinical research was suboptimal and that accuracy was not 100%.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Wang Z, Zhou C. Reassessing AI in medicine: exploring the capabilities of AI in academic abstract synthesis. J Med Internet Res 2024 [FREE Full text] [doi: 10.2196/55920]
2. Cheng S, Tsai S, Bai Y, Ko C, Hsu C, Yang F, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e51229 [FREE Full text] [doi: 10.2196/51229] [Medline: 38145486]

Abbreviations

AI: artificial intelligence

Edited by L Beri, T Leung; submitted 05.08.24; this is a non–peer-reviewed article;accepted 27.09.24; published 23.12.24.

Please cite as:
Hsu TW, Liang CS
Authors' Reply: Reassessing AI in Medicine: Exploring the Capabilities of AI in Academic Abstract Synthesis
J Med Internet Res 2024;26:e65123
URL: https://www.jmir.org/2024/1/e65123
doi:10.2196/65123
PMID:

©Tien-Wei Hsu, Chih-Sung Liang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Social Media Use and Its Concurrent and Subsequent Relation to a Biological Marker of Inflammation: Short-Term Longitudinal Study

David Lee[1], PhD; Tao Jiang[2], PhD; Jennifer Crocker[3], PhD; Baldwin Way[3], PhD

[1]Department of Communication, University at Buffalo, The State University of New York, Buffalo, NY, United States
[2]Institute for Policy Research, Northwestern University, Evanston, IL, United States
[3]Department of Psychology, The Ohio State University, Columbus, OH, United States

Corresponding Author:
David Lee, PhD
Department of Communication
University at Buffalo, The State University of New York
327 Baldy Hall
Buffalo, NY, 14260
United States
Phone: 1 716 645 1167
Email: dslee9@buffalo.edu

Related Article:

Correction of: https://www.jmir.org/2023/1/e46309

(J Med Internet Res 2024;26:e56808) doi:10.2196/56808

In "Social Media Use and Its Concurrent and Subsequent Relation to a Biological Marker of Inflammation: Short-Term Longitudinal Study" (J Med Internet Res 2023;25:e46309) one error has been noted.

In the originally published manuscript, Reference 1 contained a broken URL.

This URL has been corrected to: https://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x [1]

The correction will appear in the online version of the paper on the JMIR Publications website February 1, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Boyd DM, Ellison NB. Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 2007;13(1):210-230. [doi: 10.1111/j.1083-6101.2007.00393.x]

Submitted 26.01.24; this is a non–peer-reviewed article;accepted 26.01.24; published 01.02.24.

Please cite as:
Lee D, Jiang T, Crocker J, Way B
Correction: Social Media Use and Its Concurrent and Subsequent Relation to a Biological Marker of Inflammation: Short-Term Longitudinal Study
J Med Internet Res 2024;26:e56808
URL: https://www.jmir.org/2024/1/e56808
doi:10.2196/56808
PMID:38301251

XSL•FO
RenderX

©David Lee, Tao Jiang, Jennifer Crocker, Baldwin Way. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media

Feipeng Guo[1,2], PhD; Zixiang Liu[1], MSc; Qibei Lu[3], PhD; Shaobo Ji[4], PhD, Prof Dr; Chen Zhang[5], MSE

[1]Modern Business Research Center, Zhejiang Gongshang University, Hangzhou, China
[2]School of Management and E-Business, Zhejiang Gongshang University, Hangzhou, China
[3]School of International Business, Zhejiang International Studies University, Hangzhou, China
[4]Sprott School of Business, Carleton University, Ottawa, ON, Canada
[5]General Manager's Office, Hangzhou Gaojin Technology Co, Ltd, Hangzhou, China

Corresponding Author:
Qibei Lu, PhD
School of International Business
Zhejiang International Studies University
299 Liuhe Road
Xihu District
Hangzhou, 310030
China
Phone: 86 13857659112
Email: luqibei@zisu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2024/1/e47508

(J Med Internet Res 2024;26:e57233) doi:10.2196/57233

In “Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media” (J Med Internet Res 2024;26:e47508) the authors noted one error.

In the “Methods” section of the Abstract, the following sentence:

> *Finally, a machine learning linear regression (ML-LR) model combined with a sparse matrix was proposed to explore the evolutionary trend in public opinion on social media and verify the high accuracy of the model.*

Has been changed to read as follows:

> *Finally, a machine learning logistic regression (ML-LR) model combined with a sparse matrix was proposed to explore the evolutionary trend in public opinion on social media and verify the high accuracy of the model.*

The correction will appear in the online version of the paper on the JMIR Publications website on February 14, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 08.02.24; this is a non–peer-reviewed article;accepted 08.02.24; published 14.02.24.

Please cite as:
Guo F, Liu Z, Lu Q, Ji S, Zhang C
Correction: Public Opinion About COVID-19 on a Microblog Platform in China: Topic Modeling and Multidimensional Sentiment Analysis of Social Media
J Med Internet Res 2024;26:e57233
URL: https://www.jmir.org/2024/1/e57233
doi:10.2196/57233
PMID:38354396

XSL•FO
RenderX

©Feipeng Guo, Zixiang Liu, Qibei Lu, Shaobo Ji, Chen Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis

Jian Huo[1*], MSc; Yan Yu[2*], MMS; Wei Lin[3], MMS; Anmin Hu[2,3,4], MMS; Chaoran Wu[2], MD, PhD

[1]Boston Intelligent Medical Research Center, Shenzhen United Scheme Technology Company Limited, Boston, MA, United States

[2]Department of Anesthesia, Shenzhen People's Hospital, The First Affiliated Hospital of Southern University of Science and Technology, Shenzhen Key Medical Discipline, Shenzhen, China

[3]Shenzhen United Scheme Technology Company Limited, Shenzhen, China

[4]The Second Clinical Medical College, Jinan University, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Chaoran Wu, MD, PhD
Department of Anesthesia
Shenzhen People's Hospital, The First Affiliated Hospital of Southern University of Science and Technology
Shenzhen Key Medical Discipline
No 1017, Dongmen North Road
Shenzhen, 518020
China
Phone: 86 18100282848
Email: wu.chaoran@szhospital.com

Related Article:

Correction of: https://www.jmir.org/2024/1/e51250

(J Med Internet Res 2024;26:e59628) doi:10.2196/59628

In "Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis" (J Med Internet Res 2024;26:e51250) a redundant repetition was found.

The title was originally published as:

> *Application of AI in in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis*

It will now be changed to:

> *Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis*

The correction will appear in the online version of the paper on the JMIR Publications website on April 17, 2024, together with the publication of this correction notice. Because this correction was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 17.04.24; this is a non–peer-reviewed article;accepted 17.04.24; published 17.04.24.

Please cite as:
Huo J, Yu Y, Lin W, Hu A, Wu C
Correction: Application of AI in Multilevel Pain Assessment Using Facial Images: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e59628
URL: https://www.jmir.org/2024/1/e59628
doi:10.2196/59628
PMID:38631027

©Jian Huo, Yan Yu, Wei Lin, Anmin Hu, Chaoran Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial

Hailey N Miller[1], RN, PhD; John A Gallis[2,3], ScM; Miriam B Berger[4], MPH; Sandy Askew[4], MPH; Joseph R Egger[3], PhD; Melissa C Kay[5], RD, MS, MPH, PhD; Eric Andrew Finkelstein[6], PhD; Mia de Leon[7], BSE; Abigail DeVries[8], MD; Ashley Brewer[9], RD; Marni Gwyther Holder[10], RN, MSN; Gary G Bennett[4,11], PhD

[1]School of Nursing, Johns Hopkins University, Baltimore, MD, United States
[2]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, United States
[3]Duke Global Health Institute, Duke University, Durham, NC, United States
[4]Duke Digital Health Science Center, Duke University, Durham, NC, United States
[5]Department of Pediatrics, Duke University, Durham, NC, United States
[6]Duke-NUS Medical School Singapore, Duke Global Health Institute, Duke University, Durham, NC, United States
[7]Caraway, New York, NY, United States
[8]Medical Home Network, Chicago, IL, United States
[9]Piedmont Health Services, Inc, Chapel Hill, NC, United States
[10]Department of Family Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[11]Trinity College of Arts & Sciences, Duke University, Durham, NC, United States

Corresponding Author:
Miriam B Berger, MPH
Duke Digital Health Science Center
Duke University
417 Chapel Drive, Room 048
Campus Box 90086
Durham, NC, 27708
United States
Phone: 1 919 613 8398
Email: miriam.berger@duke.edu

Related Article:

Correction of: https://www.jmir.org/2024/1/e50330

(J Med Internet Res 2024;26:e60137) doi:10.2196/60137

In "Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial" (J Med Internet Res 2024;26:e50330) an error was noted:

In Figure 3, the titles "Panel A" and "Panel B" were missing at the top of the two graphs.

Also in Figure 3, the y-axis title of the second graph:

Risk ratio of ≤3% weight gain

Has been revised to:

Risk difference of ≤3% weight gain

Hence, the revised Figure 3 will be the following:

XSL•FO
RenderX

Figure 3. Risk ratio (panel A) and risk difference (panel B) comparing intervention and control on ≤3% weight gain. In panel A, values above 1 indicate that the intervention group had a greater relative probability of ≤3% weight gain versus control. In panel A, values above 0 indicate that the intervention group had a greater absolute probability of ≤3% weight gain versus control.

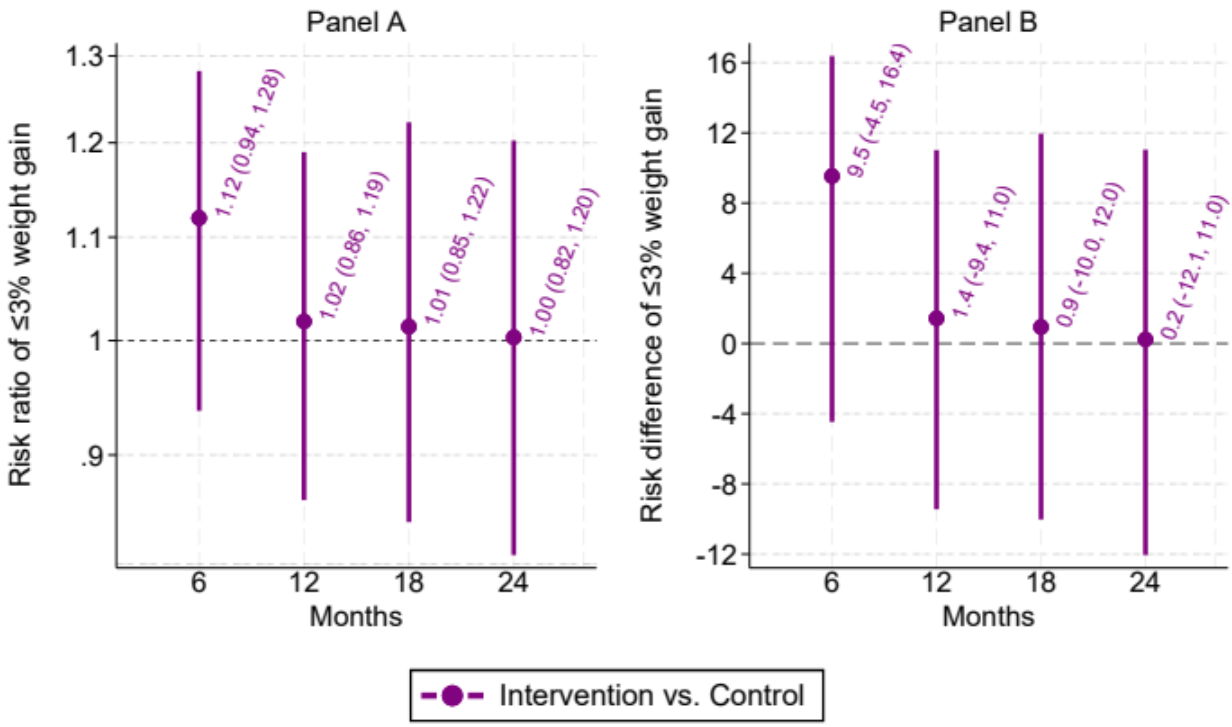

The correction will appear in the online version of the paper on the JMIR Publications website on May 13, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 02.05.24; this is a non–peer-reviewed article;accepted 02.05.24; published 13.05.24.

Please cite as:
Miller HN, Gallis JA, Berger MB, Askew S, Egger JR, Kay MC, Finkelstein EA, de Leon M, DeVries A, Brewer A, Holder MG, Bennett GG
Correction: Weight Gain Prevention Outcomes From a Pragmatic Digital Health Intervention With Community Health Center Patients: Randomized Controlled Trial
J Med Internet Res 2024;26:e60137
URL: https://www.jmir.org/2024/1/e60137
doi:10.2196/60137
PMID:38739917

©Hailey N Miller, John A Gallis, Miriam B Berger, Sandy Askew, Joseph R Egger, Melissa C Kay, Eric Andrew Finkelstein, Mia de Leon, Abigail DeVries, Ashley Brewer, Marni Gwyther Holder, Gary G Bennett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Habit Strength, Medication Adherence, and Habit-Based Mobile Health Interventions Across Chronic Medical Conditions: Systematic Review

Sherif M Badawy[1,2*], MS, MD, MBBCh; Richa Shah[3*]; Usman Beg[4], BS; Mallorie B Heneghan[1,2], MD

[1]Department of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[2]Division of Hematology, Oncology and Stem Cell Transplant, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, United States
[3]Weinberg College of Arts and Sciences, Northwestern University, Evanston, IL, United States
[4]Midwestern University Arizona College of Osteopathic Medicine, Glendale, AZ, United States
[*]these authors contributed equally

Corresponding Author:
Sherif M Badawy, MS, MD, MBBCh
Department of Pediatrics
Northwestern University Feinberg School of Medicine
225 E Chicago Avenue, Box #30
Chicago, IL, 60654
United States
Phone: 1 3122274836
Email: sbadawy@luriechildrens.org

Related Article:

Correction of: https://www.jmir.org/2020/4/e17883/

(J Med Internet Res 2024;26:e63980) doi:10.2196/63980

In "Habit Strength, Medication Adherence, and Habit-Based Mobile Health Interventions Across Chronic Medical Conditions: Systematic Review" (J Med Internet Res 2020;22(4):e17883) the authors made one addition.

In the "Acknowledgements" section, which previously appeared as:

> *This project was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (K23HL150232, principal investigator: Badawy). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.*

The authors have added an additional paragraph preceding the original, as follows:

> *We wish to acknowledge MMAS Research LLC for the use of ©Morisky Widget MMAS-8 software. Morisky Widget MMAS-8 software is protected by U.S. copyright and trademark laws. Permission for use of Morisky Widget MMAS-8 software copyright registration number TX 8-816-517 is required and was obtained for this research. A license agreement is available from MMAS Research LLC, 1501 Spring Hill Red, Petaluma CA, 94952 USA; strubow@morisky.org. We would like to express our gratitude to Steve Trubow for the generous training and the timely license provision.*

The correction will appear in the online version of the paper on the JMIR Publications website on July 8, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 05.07.24; this is a non–peer-reviewed article;accepted 05.07.24; published 08.07.24.

Please cite as:
Badawy SM, Shah R, Beg U, Heneghan MB
Correction: Habit Strength, Medication Adherence, and Habit-Based Mobile Health Interventions Across Chronic Medical Conditions: Systematic Review
J Med Internet Res 2024;26:e63980
URL: https://www.jmir.org/2024/1/e63980
doi:10.2196/63980
PMID:

©Sherif M Badawy, Richa Shah, Usman Beg, Mallorie B Heneghan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties

Felix Busch[1], MD; Tianyu Han[2], PhD; Marcus R Makowski[3], MD, PhD; Daniel Truhn[2], MSc, MD; Keno K Bressem[4*], MD; Lisa Adams[3*], MD

[1]Department of Neuroradiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[2]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

[3]Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

[4]Institute for Radiology and Nuclear Medicine, German Heart Center Munich, Technical University of Munich, Munich, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Busch, MD
Department of Neuroradiology
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin
Charitépl. 1
Berlin, 10117
Germany
Phone: 49 3045050
Email: felix.busch@charite.de

Related Article:

Correction of: http://mhealth.jmir.org/2024/1/e54948/

(J Med Internet Res 2024;26:e64411) doi:10.2196/64411

In "Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties" (J Med Internet Res 2024;26:e54948) one error was noted.

In the second paragraph of the "Methods" section, the phrase:

> *GPT-4 and GPT-4V were accessed between November 6, 2023, and November 17, 2024.*

Has been changed to read as follows:

> *GPT-4 and GPT-4V were accessed between November 6, 2023, and November 17, 2023.*

The correction will appear in the online version of the paper on the JMIR Publications website on July 19, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 17.07.24; this is a non–peer-reviewed article;accepted 17.07.24; published 19.07.24.

Please cite as:
Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L
Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties
J Med Internet Res 2024;26:e64411
URL: https://www.jmir.org/2024/1/e64411
doi:10.2196/64411
PMID:

XSL•FO
RenderX

©Felix Busch, Tianyu Han, Marcus R Makowski, Daniel Truhn, Keno K Bressem, Lisa Adams. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis

Nuša Farič[1,2], BSc, MSc, PhD; Henry WW Potts[3], BA, MSc, PhD; James M Heilman[4], BSc, MD

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[2]Advanced Care Research Center, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[3]University College London, Institute of Health Informatics, London, United Kingdom
[4]Department of Emergency Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Henry WW Potts, BA, MSc, PhD
University College London
Institute of Health Informatics
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 0207679200
Email: h.potts@ucl.ac.uk

Related Article:

Correction of: https://www.jmir.org/2024/1/e47562/

(J Med Internet Res 2024;26:e66590) doi:10.2196/66590

In "Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis" (J Med Internet Res 2024;26:e47562) an error has been noted.

In the Abstract, the following excerpt:

> *Of the 1000 Wikipedia health articles, 993 (93.3%) were not sex specific and 67 (6.7%) were sex specific...*

has been revised to:

> *Of the 1000 Wikipedia health articles, 933 (93.3%) were not sex specific and 67 (6.7%) were sex specific...*

The correction will appear in the online version of the paper on the JMIR Publications website on September 18, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 17.09.24; this is a non–peer-reviewed article;accepted 17.09.24; published 25.09.24.

Please cite as:
Farič N, Potts HWW, Heilman JM
Correction: Quality of Male and Female Medical Content on English-Language Wikipedia: Quantitative Content Analysis
J Med Internet Res 2024;26:e66590
URL: https://www.jmir.org/2024/1/e66590
doi:10.2196/66590
PMID:

©Nuša Farič, Henry WW Potts, James M Heilman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study

Angela Karellis[1], PhD; Duncan Webster[2], MD; Jean Boulanger[3], RN; Kate Harland[2], RN; Paige Feltmate[2], RN; Stefanie Materniak[2], MSc; Gabriel Daunais-Laurin[3], BA; Christine Mesa[4], MSc; Olivia Vaikla[1], BSc; John Kim[4], PhD; Nitika Pant Pai[1,5], MPH, MD, PhD

[1]Infectious Diseases and Immunity in Global Health, Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[2]Centre for Research, Education & Clinical Care of At-Risk Populations, Saint John, NB, Canada
[3]REZO, Montreal, QC, Canada
[4]National Laboratory for HIV Reference Services, Public Health Agency of Canada, Winnipeg, MB, Canada
[5]Faculty of Medicine, McGill University, Montreal, QC, Canada

Corresponding Author:
Nitika Pant Pai, MPH, MD, PhD
Infectious Diseases and Immunity in Global Health
Research Institute of the McGill University Health Centre
5252 boul de Maisonneuve Ouest
Montreal, QC, H4A3S5
Canada
Phone: 1 514 934 1934 ext 44729
Email: nitika.pai@mcgill.ca

Related Article:

Correction of: https://www.jmir.org/2024/1/e55075

(J Med Internet Res 2024;26:e67994) doi:10.2196/67994

In "Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study" (J Med Internet Res 2024;26:e55075), 2 errors were made during the publication process.

1) Affiliations 1 and 4:

> *1. Research Institute of the McGill University Health Centre, Montreal, QC, Canada*
>
> *4. National Laboratory for HIV Reference Services, Winnipeg, MB, Canada*

Have been revised to include subunits, as follows:

> *1. Infectious Diseases and Immunity in Global Health, Research Institute of the McGill University Health Centre, Montreal, QC, Canada*
>
> *4. National Laboratory for HIV Reference Services, Public Health Agency of Canada, Winnipeg, MB, Canada*

2) Tables 3 and 4 have been revised to adjust some numbers in "Sensitivity" and "Specificity" columns, as follows:

XSL•FO
RenderX

Table 3. Diagnostic accuracy of the Chembio and MedMira point-of-care (POC) tests in comparison to local laboratory tests in the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (2021-2022).

POC test	Sensitivity, % (95% CI)	Specificity, % (95% CI)
HIV		
Chembio	100.00 (79.41-100.00)	98.71 (96.72-99.65)
MedMira	100.00 (78.20-100.00)	100.00 (98.82-100.00)
Syphilis		
Chembio	86.84 (71.91-95.59)	99.65 (98.06-99.99)
MedMira	57.89 (40.82-73.69)	100.00 (98.72-100.00)
Hepatitis C		
Chembio	—[a]	—
MedMira	90.32 (80.12-96.37)	99.62 (97.92-99.99)

[a]Not applicable.

Table 4. Diagnostic accuracy of the Chembio and MedMira point-of-care (POC) tests in comparison to dried blood spot tests in the AideSmart! cross-sectional multiplexed rapid sexually transmitted blood-borne infection screening study in Canada (2021-2022).

POC test	Sensitivity, % (95% CI)	Specificity, % (95% CI)
HIV		
Chembio	100.00 (76.84-100.00)	98.94 (97.30-99.71)
MedMira	93.75 (69.77-99.84)	100.00 (99.02-100.00)
Syphilis		
Chembio	80.00 (64.35-90.95)	99.03 (97.19-99.80)
MedMira	55.00 (38.49-70.74)	100.00 (98.81-100.00)
Hepatitis C		
Chembio	—[a]	—
MedMira	91.15 (84.33-95.67)	99.29 (97.46-99.91)

[a]Not applicable.

The correction will appear in the online version of the paper on the JMIR Publications website on October 28, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 25.10.24; this is a non–peer-reviewed article;accepted 25.10.24; published 28.10.24.

Please cite as:
Karellis A, Webster D, Boulanger J, Harland K, Feltmate P, Materniak S, Daunais-Laurin G, Mesa C, Vaikla O, Kim J, Pant Pai N Correction: Feasibility, Preference, and Impact of a Rapid Multiplexed Point-of-Care Digital Innovation (AideSmart!) to Expedite Screening of Sexually Transmitted Blood-Borne Infections in At-Risk Populations in Canada: Cross-Sectional Study
J Med Internet Res 2024;26:e67994
URL: https://www.jmir.org/2024/1/e67994
doi:10.2196/67994
PMID:

©Angela Karellis, Duncan Webster, Jean Boulanger, Kate Harland, Paige Feltmate, Stefanie Materniak, Gabriel Daunais-Laurin, Christine Mesa, Olivia Vaikla, John Kim, Nitika Pant Pai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Quality Assessment of Digital Health Apps: Umbrella Review

Maciej Marek Zych[1], BSc; Raymond Bond[1], BSc, PhD; Maurice Mulvenna[1], BSc, MPhil, PhD; Jorge Martinez Carracedo[1], PhD, BSc, MSc; Lu Bai[2], BEng, PhD; Simon Leigh[3], BSc, MSc, PhD

[1]School of Computing, Ulster University, Belfast, United Kingdom
[2]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom
[3]ORCHA, Daresbury, United Kingdom

Corresponding Author:
Maciej Marek Zych, BSc
School of Computing
Ulster University
2-24 York St
Belfast, BT15 1AP
United Kingdom
Phone: 44 7526852505
Email: maciejmarekzych@gmail.com

Related Article:

Correction of: https://www.jmir.org/2024/1/e58616

(J Med Internet Res 2024;26:e67849) doi:10.2196/67849

In "Quality Assessment of Digital Health Apps: Umbrella Review" (J Med Internet Res 2024;26:e58616) one error has been noted.

> *Multimedia Appendix 1 presents the PRISMA checklist.*

has been revised to:

> *Multimedia Appendix 1 presents the PRIOR checklist.*

The correction will appear in the online version of the paper on the JMIR Publications website on October 28, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 22.10.24; this is a non–peer-reviewed article;accepted 22.10.24; published 28.10.24.

Please cite as:
Zych MM, Bond R, Mulvenna M, Martinez Carracedo J, Bai L, Leigh S
Correction: Quality Assessment of Digital Health Apps: Umbrella Review
J Med Internet Res 2024;26:e67849
URL: https://www.jmir.org/2024/1/e67849
doi:10.2196/67849
PMID:

©Maciej Marek Zych, Raymond Bond, Maurice Mulvenna, Jorge Martinez Carracedo, Lu Bai, Simon Leigh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review

Jie Gong[1], BD; Dandan Gu[1], MD; Suyun Dong[1], BD; Wangqin Shen[2], PhD; Haiou Yan[1], BD; Juan Xie[1], BD

[1]Department of Nursing, Affiliated Hospital of Nantong University, Nursing and Rehabilitation School of Nantong University, Nantong, China
[2]School of Nursing and Rehabilitation, Nantong University, Nantong, China

Corresponding Author:
Juan Xie, BD
Department of Nursing
Affiliated Hospital of Nantong University
Nursing and Rehabilitation School of Nantong University
No. 20 West Temple Road
Nantong, 226001
China
Phone: 86 13912288333
Email: xjjx@ntu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2024/1/e52738

(J Med Internet Res 2024;26:e68783) doi:10.2196/68783

In "Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review" (J Med Internet Res 2024;26:e52738) one error was noted.

In the Methods section of the text, the PROSPERO registration number was revised. Therefore, the following sentence:

> *The study protocol was registered in PROSPERO (number CRD42063451612).*

Has been revised to:

> *The study protocol was registered in PROSPERO (number CRD42023451612).*

The correction will appear in the online version of the paper on the JMIR Publications website on November 15, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 13.11.24; this is a non–peer-reviewed article;accepted 13.11.24; published 15.11.24.

Please cite as:
Gong J, Gu D, Dong S, Shen W, Yan H, Xie J
Correction: Effects of Message Framing on Human Papillomavirus Vaccination: Systematic Review
J Med Internet Res 2024;26:e68783
URL: https://www.jmir.org/2024/1/e68783
doi:10.2196/68783
PMID:

©Jie Gong, Dandan Gu, Suyun Dong, Wangqin Shen, Haiou Yan, Juan Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study

Minseok Hong[1,2], MD; Ri-Ra Kang[3], BS; Jeong Hun Yang[2,4], MD, PhD; Sang Jin Rhee[1], MD, PhD; Hyunju Lee[1], MD, PhD; Yong-gyom Kim[3], BS; KangYoon Lee[3,5], PhD; HongGi Kim[6], BS; Yu Sang Lee[7], MD, PhD; Tak Youn[8,9], MD, PhD; Se Hyun Kim[1,2], MD, PhD; Yong Min Ahn[1,2,10], MD, PhD

[1]Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea
[2]Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea
[3]Department of IT Convergence Engineering, Gachon University, Seongnam-si, Republic of Korea
[4]Department of Psychiatry, Chungnam National University Sejong Hospital, Sejong, Republic of Korea
[5]Department of Computer Engineering, Gachon University, Seongnam-si, Republic of Korea
[6]Healthconnect Co. Ltd., Seoul, Republic of Korea
[7]Department of Psychiatry, Yong-In Mental Hospital, Yongin-si, Republic of Korea
[8]Department of Psychiatry and Electroconvulsive Therapy Center, Dongguk University International Hospital, Goyang-si, Republic of Korea
[9]Institute of Buddhism and Medicine, Dongguk University, Seoul, Republic of Korea
[10]Institute of Human Behavioral Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea

Corresponding Author:
Yong Min Ahn, MD, PhD
Department of Neuropsychiatry
Seoul National University Hospital
101 Daehak-ro
Jongno-Gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 2450
Fax: 82 2 766 2450
Email: aym@snu.ac.kr

Related Article:

Correction of: https://www.jmir.org/2024/1/e65994/

(J Med Internet Res 2024;26:e69042) doi:10.2196/69042

In "Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study" (J Med Internet Res 2024;26:e65994), a few errors were noted.

1. The following sentence:

> *However, significant variations were observed across wards, which presents a key challenge…*

has been corrected to:

> *However, significant variations were observed across wards, which present a key challenge…*

2. The following sentence:

> *Despite lower computational costs, Multi models demonstrated equivalent or superior performance than Single models…*

has been corrected to:

> *Despite lower computational costs, Multi models demonstrated equivalent or superior performance to Single models…*

3. The following sentence:

> *…installed in each patient's room and common areas*

has been corrected to:

> *…installed in each patient's room and common areas.*

4. The following sentence:

> *The analysis code is publicly [77].*

has been corrected to:

> *The analysis code is publicly available [77].*

5. The following sentence:

XSL•FO
RenderX

> *…the YMRS performed the least accurately in in external validation…*

has been revised to:

> *…the YMRS performed the least accurately in external validation…*

The correction will appear in the online version of the paper on the JMIR Publications website on December 3, 2024, together with the publication of this correction notice. Because these corrections were made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 20.11.24; this is a non–peer-reviewed article;accepted 20.11.24; published 03.12.24.

Please cite as:
Hong M, Kang RR, Yang JH, Rhee SJ, Lee H, Kim YG, Lee K, Kim H, Lee YS, Youn T, Kim SH, Ahn YM
Correction: Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study
J Med Internet Res 2024;26:e69042
URL: https://www.jmir.org/2024/1/e69042
doi:10.2196/69042
PMID:

©Minseok Hong, Ri-Ra Kang, Jeong Hun Yang, Sang Jin Rhee, Hyunju Lee, Yong-gyom Kim, KangYoon Lee, HongGi Kim, Yu Sang Lee, Tak Youn, Se Hyun Kim, Yong Min Ahn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

JOURNAL OF MEDICAL INTERNET RESEARCH

Corrigenda and Addenda

Correction: Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line

JMIR Editorial Office[1]

JMIR Publications, Toronto, ON, Canada

Corresponding Author:
JMIR Publications
130 Queens Quay East
Suite 1100-1102
Toronto, ON, M5A 0P6
Canada
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Correction of: https://jmir.org/2019/1/e11507/

(J Med Internet Res 2024;26:e67880) doi:10.2196/67880

KEYWORDS

conflicts of interest; disclosure

In "Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line" (J Med Internet Res 2019;21(1):e11507) [1], a change has been made to the Conflicts of Interest section.

According to a commentary by Tim Reierson [2], the following relationships were originally not explicitly declared in the Conflicts of Interest section (author initials were added by the editor for clarity):

> *One other author was a Crisis Text Line staff member, the study's open data collaborations manager Nitya Kanuri. Her staff member affiliation is undeclared in the paper. The following affiliations to Crisis Text Line, Inc. were also undeclared: the corresponding author [AP] and one other author [MG] were serving on the clinical advisory board to Crisis Text Line. Seven other authors [BP, DR, JEM, LSL, MLR, RL, SY] were members of the Data Ethics Committee to the study. The Data Ethics Committee was listed on Crisis Text Line's official website as an advisory panel of experts to the organization. [2]*

The corresponding author Pisani, in their response [3], argues that "far from being 'undisclosed,' the participation and collaboration of members of the author group with Crisis Text Line in documenting the results of their pilot academic data sharing program was a key feature of the article" and further adds that "an appendix was provided listing members of the data ethics committee, which included several of the authors."

We agree with this characterization. We do not see any attempts to deceive reviewers or editors by concealing that this is a collaborative project that involved staff members of Crisis Text Line and academic experts and ethicists comprising the Data Ethics Committee, who advised Crisis Text Line on best practices on data sharing with researchers, which was the topic of the viewpoint article.

However, JMIR Publications' policy has stated since 2017 that conflicts of interest disclosures for authors include "non-financial competing interests," including "membership in a government or other advisory board" [4]. This is consistent with the guidance from the International Committee of Medical Journal Editors (ICMJE) on conflicts of interest disclosures [5].

Given the Loris.ai controversy, which unfolded after publication of the 2019 paper, and which involved sharing of anonymized Crisis Text Line data with the commercial subsidiary Loris.ai for machine learning purposes [6], these relationships as well as relationships to Loris.ai could have been disclosed more explicitly, which is why we are updating the Conflicts of Interest section for this publication accordingly.

Crisis Text Line emphasizes that Loris.ai did not exist during the time of the academic data sharing pilot [7]. From a publication ethics perspective, the most relevant question is whether any of the authors (including the three coauthors affiliated with Crisis Text Line) benefitted or had the potential to benefit financially from Loris.ai, at the time of submission or acceptance of their article. As Reierson points out:

XSL•FO
RenderX

> *None of the authors made any disclosures or acknowledged the existence of the for-profit or its licensing agreement with Crisis Text Line. Loris.ai, Inc. had been in operation for months by the time the paper was submitted for publication. [2]*

In response [3], the authors assured us that "no one on the CTL [Crisis Text Line] data advisory board had a role in or even knew about Loris.ai." In addition, we received confirmation from the three authors who were staff members of Crisis Text Line (Shairi Turner, Current Chief Health Officer at Crisis Text Line; Bob Filbin, former Chief Data Science officer at Crisis Text Line who left in 2021; and Nitya Kanuri, former Open Data Collaborations Manager at Crisis Text Line) that they did *not* have a financial stake in Loris.ai. (NK did not include Crisis Text Line as an affiliation because that author was affiliated with Crisis Text Line only in 2016-2017, and they had moved on by the time when the paper was published; past affiliations are often not listed, and there is a lack of a consensus in the scholarly publishing industry on what affiliation should actually be declared upon submission [8,9]).

Any financial stake in Loris.ai certainly should have been disclosed with the 2019 paper published in the *Journal of Medical Internet Research*, but would not necessarily invalidate the viewpoint article (which focuses on a *noncommercial* sharing scenario). While Crisis Text Line staff Bob Filbin and Shairi Turner conceded that they had knowledge of the license agreement between Loris.ai and Crisis Text Line at the time of submission, it is an open question whether such knowledge could have biased the work of the independent Data Ethics Committee. However, we are satisfied to hear that the Data Ethics Committee, which evidently spearheaded the guidelines and processes for noncommercial academic data sharing, did not know about Loris.ai and did not benefit from any commercial data sharing agreement personally, nor did any of the authors personally benefit from such an arrangement financially.

JMIR Publications' policy on conflicts of interest follows ICMJE recommendations [4,5] and includes procedures relating to post-publication handling of undisclosed conflicts of interest [10]. As a member of the Committee on Publication Ethics (COPE), the *Journal of Medical Internet Research* adheres to COPE guidance on these matters and hereby updates the Conflicts of Interest section of the published viewpoint article by Pisani et al [1], as follows:

> ***Conflicts of Interest***
>
> *(updated December 2024) NK was a salaried full-time employee with Crisis Text Line in 2016-2017, overseeing the Robert Wood Johnson Foundation–funded, open data pilot described in this article; they were no longer in that role at the time of publication of this article. BF was a salaried Chief Data Scientist at Crisis Text Line and chaired the Data Ethics Committee. ST served as Chief Medical Officer for Crisis Text Line and received a salary. AP and MG were uncompensated members of the clinical advisory board to Crisis Text Line. As outlined in Multimedia Appendix 1, several of the authors (BP, DR, JEM, LSL, MLR, RL, and SY) served as uncompensated members of the Data Ethics Committee for Crisis Text Line. None of the authors reported any financial benefits from a data sharing agreement between Crisis Text Line and its for-profit subsidy Loris.ai, including a financial stake in Loris.ai. None of the authors except Crisis Text Line leaders (BF and ST) had contemporaneous knowledge of a license agreement between Crisis Text Line and Loris.ai. The data sharing agreement between Crisis Text Line and Loris.ai was terminated in 2022.*

The *Journal of Medical Internet Research* editors do not think that this revised conflicts of interest statement affected the peer review process or the validity of this viewpoint article. Consequently, the noted issue of concern regarding undisclosed conflicts of interest, given the information currently available, did not appear to meet criteria for further action, such as a notice of editorial concern or retraction [11].

The correction will appear in the online version of the paper [1] on the JMIR Publications website on December 20, 2024, together with the publication of this correction notice. Because this correction was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

References

1. Pisani AR, Kanuri N, Filbin B, Gallo C, Gould M, Lehmann LS, et al. Protecting user privacy and rights in academic data-sharing partnerships: principles from a pilot program at Crisis Text Line. J Med Internet Res 2019 Jan 17;21(1):e11507 [FREE Full text] [doi: 10.2196/11507] [Medline: 30664452]
2. Reierson TD. Commentary to "Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles from a Pilot Program at Crisis Text Line". JMIR Preprints 2024 Jun 28 [FREE Full text]
3. Pisani AR, Gallo C, Gould M, Kanuri N, Marcotte JE, Pascal B, et al. Science and ethics advance through publication, critique, and refinement: author reply to commentary five years after publication. JMIR Preprints 2024 Jul 05 [FREE Full text]
4. How does JMIR Publications define a Conflict of Interest (COI)? JMIR Publications Knowledge Base and Help Center. URL: https://support.jmir.org/hc/en-us/articles/115001252671 [accessed 2024-07-05]
5. Disclosure of Financial and Non-Financial Relationships and Activities, and Conflicts of Interest. International Committee of Medical Journal Editors. URL: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/author-responsibilities--conflicts-of-interest.html [accessed 2023-03-06]

XSL•FO
RenderX

6. Eysenbach G. Crisis Text Line and Loris.ai controversy highlights the complexity of informed consent on the internet and data sharing ethics for big data. J Med Internet Res (forthcoming/in press) 2024 Oct 24:67878 (forthcoming) [FREE Full text] [doi: 10.2196/67878]
7. Trujillo D. How Crisis Text Line protects data privacy, digital rights and research ethics. J Med Internet Res (forthcoming/in press) 2024 Dec 20:64015 [FREE Full text] [doi: 10.2196/64015]
8. Claiming institutional affiliations. COPE: Committee on Publication Ethics. URL: https://publicationethics.org/topic-discussions/claiming-institutional-affiliations [accessed 2024-07-05]
9. Clancy L. Which author affiliation do I use? International Science Editing. 2022 Jan 13. URL: https://www.internationalscienceediting.com/which-author-affiliation-do-i-use/ [accessed 2024-07-05]
10. Undisclosed conflict of interest in a published article. COPE: Committee on Publication Ethics. URL: https://publicationethics.org/guidance/flowchart/undisclosed-conflict-interest-published-article [accessed 2024-07-05]
11. Retraction guidelines. COPE: Committee on Publication Ethics. URL: https://publicationethics.org/guidance/guideline/retraction-guidelines [accessed 2024-12-18]

Edited by T Leung; submitted 23.10.24; this is a non–peer-reviewed article;accepted 24.10.24; published 20.12.24.

Please cite as:
JMIR Editorial Office
Correction: Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line
J Med Internet Res 2024;26:e67880
URL: https://www.jmir.org/2024/1/e67880
doi:10.2196/67880
PMID:

© JMIR Editorial Office. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Electronic Health Record–Nested Reminders for Serum Lithium Level Monitoring in Patients With Mood Disorder: Randomized Controlled Trial

Tomotsugu Seki[1,2*], MD, MPH; Morio Aki[3,4*], MD; Toshi A Furukawa[5], MD, PhD; Hirotsugu Kawashima[3,4], MD; Tomotaka Miki[3,4], MD; Yujin Sawaki[3,4,6], MD; Takaaki Ando[3,4], MD; Kentaro Katsuragi[3,4], MD; Takahiko Kawashima[3,4], MD; Senkei Ueno[3,4], MD; Takashi Miyagi[3,4,7], MD; Shun'ichi Noma[3,4,8], MD; Shiro Tanaka[9], PhD; Koji Kawakami[1], MD, PhD

[1]Department of Pharmacoepidemiology, School of Public Health, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[2]Department of Cardiovascular Medicine, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan
[3]Department of Psychiatry, Toyooka Hospital, Toyooka, Japan
[4]Department of Psychiatry, Kyoto University Hospital, Kyoto, Japan
[5]Department of Health Promotion and Human Behavior, School of Public Health, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[6]National Epilepsy Center, National Hospital Organization Shizuoka Institute of Epilepsy and Neurological Disorders, Shizuoka, Japan
[7]Department of Psychiatry, Kyoto-Katsura Hospital, Kyoto, Japan
[8]Noma-Kokoro Clinic, Kyoto, Japan
[9]Department of Clinical Biostatistics, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[*]these authors contributed equally

Corresponding Author:
Toshi A Furukawa, MD, PhD
Department of Health Promotion and Human Behavior
School of Public Health, Graduate School of Medicine
Kyoto University
Yoshida Konoe-cho
Sakyo-ku
Kyoto, 606-8501
Japan
Phone: 81 75 753 9491
Email: furukawa@kuhp.kyoto-u.ac.jp

Related Article:

Correction of: https://www.jmir.org/2023/1/e40595

(J Med Internet Res 2024;26:e55805) doi:10.2196/55805

The authors of “Electronic Health Record–Nested Reminders for Serum Lithium Level Monitoring in Patients With Mood Disorder: Randomized Controlled Trial” (J Med Internet Res 2023;25:e40595) noted an error in the *Sample Size Calculation* subsection of the *Methods* section.

Originally, the power in the sample size calculation was reported as “90%,” but the correct power for the study was “80%.”

The sentence

> *In this trial, we estimated that a total of 108 patients would provide 90% power if 48 (80%) patients in the reminder group and 33 (55%) patients in the usual care group achieved the primary outcome using the 2-sided chi-square test with a 2-side alpha level of .05. Assuming a dropout rate of 10%, the final sample size was set at 120.*

was changed to

> *In this trial, we estimated that a total of 108 patients would provide 80% power if 48 (80%) patients in the reminder group and 33 (55%) patients in the usual care group achieved the primary outcome using the 2-sided chi-square test with a 2-side alpha level of .05. Assuming a dropout rate of 10%, the final sample size was set at 120.*

This error does not affect the study's methodology, results, or conclusions, but it is important to correct the record for accuracy and clarity.

XSL•FO
RenderX

The correction will appear in the online version of the paper on the JMIR Publications website on January 3, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 25.12.23; this is a non–peer-reviewed article;accepted 29.12.23; published 03.01.24.

Please cite as:
Seki T, Aki M, Furukawa TA, Kawashima H, Miki T, Sawaki Y, Ando T, Katsuragi K, Kawashima T, Ueno S, Miyagi T, Noma S, Tanaka S, Kawakami K
Correction: Electronic Health Record–Nested Reminders for Serum Lithium Level Monitoring in Patients With Mood Disorder: Randomized Controlled Trial
J Med Internet Res 2024;26:e55805
URL: https://www.jmir.org/2024/1/e55805
doi:10.2196/55805
PMID:38170973

©Tomotsugu Seki, Morio Aki, Toshi A Furukawa, Hirotsugu Kawashima, Tomotaka Miki, Yujin Sawaki, Takaaki Ando, Kentaro Katsuragi, Takahiko Kawashima, Senkei Ueno, Takashi Miyagi, Shun'ichi Noma, Shiro Tanaka, Koji Kawakami. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Creatively Adapting Touch-Based Practices to the Web Format During the COVID-19 Pandemic: Systematic Review

Greta Gauhe[1], BA, MA, MFA; Rosemary E Kostic Cisneros[1], BS, MA, PhD; Jade Ward[1], BEd; David J Hohenschurz-Schmidt[2,3], MOst, MSc, PhD

[1]Centre for Dance Research, Coventry University, Coventry, United Kingdom
[2]Pain Group, Department of Surgery & Cancer, Imperial College London, London, United Kingdom
[3]Research Department, University College of Osteopathy, London, United Kingdom

Corresponding Author:
David J Hohenschurz-Schmidt, MOst, MSc, PhD
Pain Group
Department of Surgery & Cancer
Imperial College London
Chelsea & Westminster Hospital
369 Fulham Road
London, SW10 9NH
United Kingdom
Phone: 44 0 20 3315 88
Email: d.hohenschurz@imperial.ac.uk

Related Article:

Correction of: https://www.jmir.org/2023/1/e46355

(J Med Internet Res 2024;26:e55713) doi:10.2196/55713

In "Creatively Adapting Touch-Based Practices to the Web Format During the COVID-19 Pandemic: Systematic Review" (J Med Internet Res 2023;25:e46355) the authors noted two errors:

1. In the originally published article, in Tables 2 and 3 the in-text citation for reference [34] appeared as:

Lord Ferguson et al [34]

This has been corrected to:

Lord [34]

2. In the Reference list of the originally published article, reference [34] appeared as follows:

Lord Ferguson S, Smith C, Kietmann J. Hands-off? Lessons from high-touch professionals about going virtual. Bus Horiz. 2022 May;65(3):303-13. Doi:10.1016/j.bushor.2021.03.002.

This has been corrected to:

Lord SA. COVID Couple Therapy: Telehealth and Somatic Action Techniques. Aust N Z J Fam Ther. 2022;43(2):197-209. doi:10.1002/anzf.1487

The correction will appear in the online version of the paper on the JMIR Publications website on January 05, 2023, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

34. Lord SA. COVID Couple Therapy: Telehealth and Somatic Action Techniques. Aust N Z J Fam Ther 2022 Jun;43(2):197-209 [FREE Full text] [doi: 10.1002/anzf.1487] [Medline: 35942399]

XSL•FO
RenderX

Submitted 21.12.23; this is a non–peer-reviewed article;accepted 21.12.23; published 05.01.24.

Please cite as:
Gauhe G, Kostic Cisneros RE, Ward J, Hohenschurz-Schmidt DJ
Correction: Creatively Adapting Touch-Based Practices to the Web Format During the COVID-19 Pandemic: Systematic Review
J Med Internet Res 2024;26:e55713
URL: https://www.jmir.org/2024/1/e55713
doi:10.2196/55713
PMID:38181492

©Greta Gauhe, Rosemary E Kostic Cisneros, Jade Ward, David J Hohenschurz-Schmidt. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Correlation Between Remote Symptom Reporting by Caregivers and Adverse Clinical Outcomes: Mixed Methods Study

Ingrid Oakley-Girvan[1*], MPH, PhD; Reem Yunis[1*], MS, PhD; Stephanie J Fonda[2], PhD; Michelle Longmire[1], MD; Tess L Veuthey[3], MD, PhD; Jennifer Shieh[3], MD; Sara Aghaee[4], MPH; Ai Kubo[4], PhD; Sharon W Davis[1], BS, MPA; Raymond Liu[4,5], MD; Elad Neeman[6], MD

[1]Medable Inc, Palo Alto, CA, United States
[2]Estenda Solutions Inc, Wayne, PA, United States
[3]Kaiser Permanente Northern California, San Francisco, CA, United States
[4]Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States
[5]Department of Hematology Oncology, Kaiser Permanente Northern California, San Francisco, CA, United States
[6]Kaiser Permanente Northern California, San Rafael, CA, United States
[*]these authors contributed equally

Corresponding Author:
Ingrid Oakley-Girvan, MPH, PhD
Medable Inc
525 University Ave
Suite A70
Palo Alto, CA, 94103
United States
Phone: 1 877 820 6259
Email: oakley@stanford.edu

Related Article:

Correction of: https://www.jmir.org/2023/1/e49100

(J Med Internet Res 2024;26:e56368) doi:10.2196/56368

In "Correlation Between Remote Symptom Reporting by Caregivers and Adverse Clinical Outcomes: Mixed Methods Study" (J Med Internet Res. 2023 Nov 21:25:e49100) the authors Sara Aghaee, Ai Kubo, and Raymond Liu were listed with the incorrect affiliations. The author Raymond Liu was also missing an affiliation.

In the original manuscript, the affiliations for the authors appeared as follows:

> *Sara Aghaee: Department of Hematology Oncology, The Kaiser Permanente Medical Group, San Francisco, CA, United States*
>
> *Ai Kubo: Department of Hematology Oncology, The Kaiser Permanente Medical Group, San Francisco, CA, United States*
>
> *Raymond Liu: Department of Hematology Oncology, The Kaiser Permanente Medical Group, San Francisco, CA, United States*

This has been corrected to:

> *Sara Aghaee: Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA*
>
> *Ai Kubo: Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA*
>
> *Raymond Liu: Department of Hematology Oncology, Kaiser Permanente Northern California, San Francisco, CA, United States*

Additionally, the following affiliation has been included for the author Raymond Liu:

> *Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States*

The correction will appear in the online version of the paper on the JMIR Publications website on January 30, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 14.01.24; this is a non–peer-reviewed article;accepted 15.01.24; published 30.01.24.

Please cite as:
Oakley-Girvan I, Yunis R, Fonda SJ, Longmire M, Veuthey TL, Shieh J, Aghaee S, Kubo A, Davis SW, Liu R, Neeman E
Correction: Correlation Between Remote Symptom Reporting by Caregivers and Adverse Clinical Outcomes: Mixed Methods Study
J Med Internet Res 2024;26:e56368
URL: https://www.jmir.org/2024/1/e56368
doi:10.2196/56368
PMID:38290116

©Ingrid Oakley-Girvan, Reem Yunis, Stephanie J Fonda, Michelle Longmire, Tess L Veuthey, Jennifer Shieh, Sara Aghaee, Ai Kubo, Sharon W Davis, Raymond Liu, Elad Neeman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Table Correction: Effectiveness of eHealth Smoking Cessation Interventions: Systematic Review and Meta-Analysis

Yichen E Fang[1], BA, BSc; Zhixian Zhang[1], BA; Ray Wang[1], MSc; Bolu Yang[1], MSc; Chen Chen[2], PhD; Claudia Nisa[1,3], PhD; Xin Tong[1,4], PhD; Lijing L Yan[1,2,5,6], PhD

[1]Global Health Research Center, Duke Kunshan University, Kunshan, China
[2]Department of Global Health, School of Public Health, Wuhan University, Wuhan, China
[3]Division of Social Sciences, Duke Kunshan University, Kunshan, China
[4]Data Science Research Center, Duke Kunshan University, Kunshan, China
[5]Duke Global Health Institute, Duke University, Durham, NC, United States
[6]Institute for Global Health and Development, Peking University, Beijing, China

Corresponding Author:
Lijing L Yan, PhD
Global Health Research Center
Duke Kunshan University
3rd floor, Academic Building
Kunshan, 215316
China
Phone: 86 0 512 5777 9
Email: lijing.yan@duke.edu

Related Article:

Correction of: https://www.jmir.org/2023/1/e45111

(J Med Internet Res 2024;26:e56438) doi:10.2196/56438

In “Effectiveness of eHealth Smoking Cessation Interventions: Systematic Review and Meta-Analysis” (J Med Internet Res 2023;25:e45111) the authors noted three errors associated with the cited paper by Bricker et al [1] in Table 5.

In Table 5 row “Bricker et al” of the originally published manuscript, there were several errors.

1) The number of quits and smoking participants at 12-month follow-up in the intervention group was entered incorrectly as "12 months: 256” in “Intervention quit, n” column and “12 months: 856” in “Intervention smoking, n” column. The numbers have been replaced by “12 months: 356” in “Intervention quit, n” column and “12 months: 858” in “Intervention smoking, n”.

This error then led to the following:

2) The incorrect number of “12 months: 0.92 (0.79-1.06)” in the “RR 95% CI” column

3) The incorrect statement of “No significant increase on cessation outcome at 12 months follow-ups; significant increase on cessation outcome at 3- and 6-month follow-ups” in the “Summary of outcome” column

The entry in column “RR 95% CI” has now been corrected to “12 months: RR=1.17 (1.02-1.33)” and the entry in column “Summary of outcome” has now been corrected to “significant increase in cessation outcome at 3-, 6- and 12-month follow-ups”.

The error did not affect any other part of the paper as the error happened during the final formatting stage on the authors’ end and the manuscript content was drafted with the correct numbers.

The correction will appear in the online version of the paper on the JMIR Publications website on February 7, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

XSL•FO
RenderX

1. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. JAMA Intern Med 2020 Nov 01;180(11):1472-1480 [FREE Full text] [doi: 10.1001/jamainternmed.2020.4055] [Medline: 32955554]

Submitted 16.01.24; this is a non–peer-reviewed article;accepted 18.01.24; published 07.02.24.

Please cite as:
Fang YE, Zhang Z, Wang R, Yang B, Chen C, Nisa C, Tong X, Yan LL
Table Correction: Effectiveness of eHealth Smoking Cessation Interventions: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e56438
URL: https://www.jmir.org/2024/1/e56438
doi:10.2196/56438
PMID:38324769

©Yichen E Fang, Zhixian Zhang, Ray Wang, Bolu Yang, Chen Chen, Claudia Nisa, Xin Tong, Lijing L Yan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Exploring the Effect of the Dynamics of Behavioral Phenotypes on Health Outcomes in an mHealth Intervention for Childhood Obesity: Longitudinal Observational Study

Sarah Woo[1*], PhD; Sunho Jung[2*], PhD; Hyunjung Lim[3], PhD; YoonMyung Kim[4], PhD; Kyung Hee Park[5], MPH, MD, PhD

[1]Department of Medical Sciences, College of Medicine, Hallym University, Chuncheon-si, Republic of Korea
[2]School of Management, Kyung Hee University, Seoul, Republic of Korea
[3]Department of Medical Nutrition, Kyung Hee University, Yongin-si, Republic of Korea
[4]University College, Yonsei University International Campus, Incheon, Republic of Korea
[5]Department of Family Medicine, Hallym University Sacred Heart Hospital, Hallym University, Anyang-si, Gyeonggi-do, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Kyung Hee Park, MPH, MD, PhD
Department of Family Medicine
Hallym University Sacred Heart Hospital
Hallym University
Gwanpyeong-ro 170beon-gil, Dongan-gu
Anyang-si, Gyeonggi-do, 14068
Republic of Korea
Phone: 82 31 380 3805
Fax: 82 31 380 1782
Email: beloved920@gmail.com

Related Article:

Correction of: https://www.jmir.org/2023/1/e45407

(J Med Internet Res 2024;26:e57013) doi:10.2196/57013

In “Exploring the Effect of the Dynamics of Behavioral Phenotypes on Health Outcomes in an mHealth Intervention for Childhood Obesity: Longitudinal Observational Study” (J Med Internet Res 2023;25:e45407) a figure number and some symbols were misnamed. The following corrections have been made:

In the second paragraph of the ‘FPCA Results’ subsection under the Results section, Figure 4 was misnamed as Figure 3. Changes were made in the following sentences.

The original passage:

> *In PC1, adherence to missions 1, 4, and 5 decreased at the end of the intervention (Figure 3A, 3G, and 3I), but compliance to missions 2 and 3 maintained high scores throughout the intervention (Figure 3C and 3E). Participants with –1 SD from the average had a low score at the beginning of the intervention and maintained a low level below average throughout the treatment.*

has been changed to:

> *In PC1, adherence to missions 1, 4, and 5 decreased at the end of the intervention (Figure 4A, 4G, and 4I), but compliance to missions 2 and 3 maintained high scores throughout the intervention (Figure 4C and 4E). Participants with –1 SD from the average had a low score at the beginning of the intervention and maintained a low level below average throughout the treatment.*

The following passage:

> *Participants with +1 SD from the average on PC 2 of behavioral phenotypes 1, 2, 3, and 5 showed a pattern of a low score at the beginning of the treatment, followed by a rapid increase at 8 to 10 weeks of the intervention (Figure 3B, 3D, 3F, and 3J).*

was changed to:

> *Participants with +1 SD from the average on PC 2 of behavioral phenotypes 1, 2, 3, and 5 showed a pattern of a low score at the beginning of the treatment, followed by a rapid increase at 8 to 10*

XSL•FO
RenderX

weeks of the intervention (Figure 4B, 4D, 4F, and 4J).

The following passage:

PC 2 of the fourth behavioral phenotype exhibited a slightly distinctive pattern, wherein participants with +1 SD showed a gradual increase throughout the intervention. In addition, the crossing point of participants with +1 SD and −1 SD was between 10 and 15 weeks of intervention, which is a later point compared with other phenotypes (Figure 3H).

has been changed to:

PC 2 of the fourth behavioral phenotype exhibited a slightly distinctive pattern, wherein participants with +1 SD showed a gradual increase throughout the intervention. In addition, the crossing point of participants with +1 SD and −1 SD was between 10 and 15 weeks of intervention, which is a later point compared with other phenotypes (Figure 4H).

In Figure 5, the seventh symbol from the left originally appeared as 'M3 (PC2)', and now reads 'M4 (PC1)'. The eighth symbol originally appeared as 'M4 (PC1)' and now reads 'M4 (PC2).

The correction will appear in the online version of the paper on the JMIR Publications website on February 23, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Figure 5. Effect of behavioral phenotypes on BMI z score change. PC: principal component. M1: screen time less than 2 hours; M2: eating more than 5 servings of fruits and vegetables; M3: exercising for more than 1 hour; M4: drinking water or plain milk; M5: sleeping for more than 8 hours; PC 1=high or low adherence level; PC 2=late or early behavioral change. Data were presented as standardized β coefficients and standard regression coefficients with 95% CIs. M: mission.

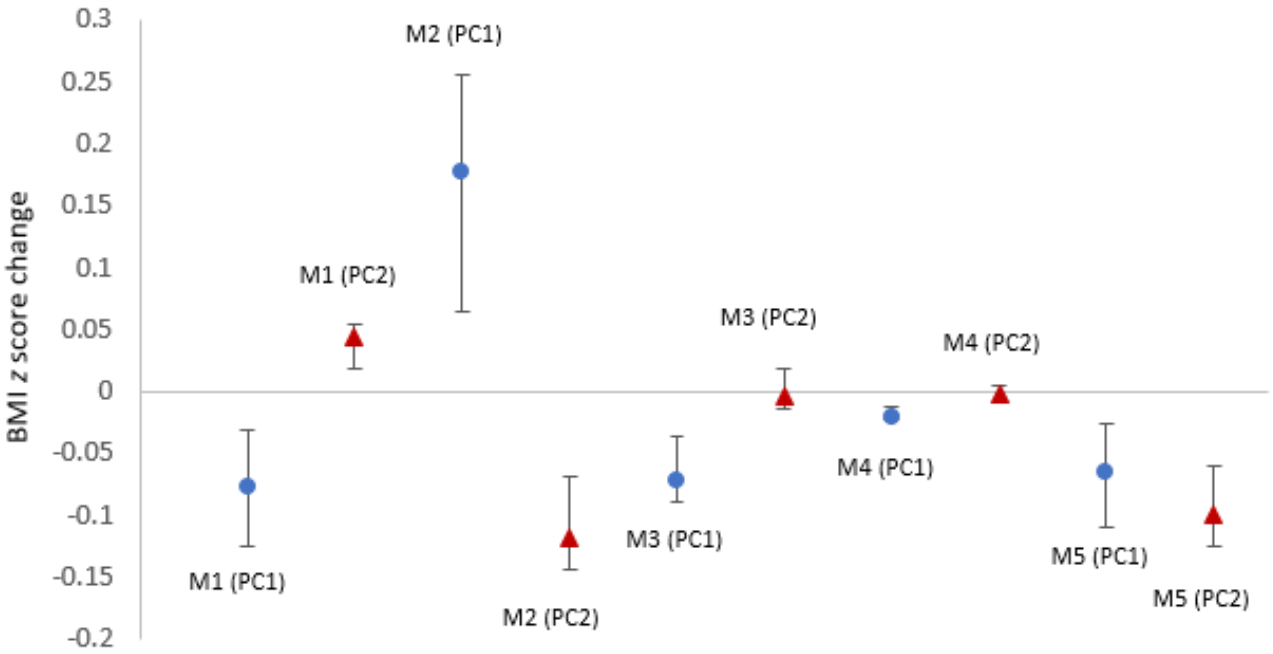

Submitted 01.02.24; this is a non–peer-reviewed article;accepted 05.02.24; published 23.02.24.

Please cite as:
Woo S, Jung S, Lim H, Kim Y, Park KH
Correction: Exploring the Effect of the Dynamics of Behavioral Phenotypes on Health Outcomes in an mHealth Intervention for Childhood Obesity: Longitudinal Observational Study
J Med Internet Res 2024;26:e57013
URL: https://www.jmir.org/2024/1/e57013
doi:10.2196/57013
PMID:38394656

XSL•FO
RenderX

©Sarah Woo, Sunho Jung, Hyunjung Lim, YoonMyung Kim, Kyung Hee Park. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial

Oh Young Kwon[1,2], PhD; Mi Kyung Lee[3], PhD; Hye Won Lee[4,5], MD, PhD; Hyerang Kim[6], PhD; Jae Seung Lee[4,5], MD, PhD; Yeonsoo Jang[2], PhD

[1]College of Nursing, Brain Korea 21 FOUR Project, Yonsei University, Seoul, Republic of Korea
[2]College of Nursing, Mo-Im Kim Nursing Research Institute, Yonsei University, Seoul, Republic of Korea
[3]Frontier Research Institute of Convergence Sports Science, Yonsei University, Seoul, Republic of Korea
[4]Department of Internal Medicine, College of Medicine, Yonsei University, Seoul, Republic of Korea
[5]Yonsei Liver Center, Severance Hospital, Seoul, Republic of Korea
[6]Department of Nursing Science, VISION College of Jeonju, Jeollabuk-Do, Republic of Korea

Corresponding Author:
Yeonsoo Jang, PhD
College of Nursing, Mo-Im Kim Nursing Research Institute
Yonsei University
50-1 Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 3343
Email: ysjang517@yuhs.ac

Related Article:

Correction of: https://www.jmir.org/2024/1/e49839

(J Med Internet Res 2024;26:e57499) doi:10.2196/57499

In "Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial" (J Med Internet Res 2024;26:e49839) the authors made two corrections.

In the first box of Figure 2, the following phrase:

> *Assessed for eligibility*

Has been replaced with the following, as visible in the attached figure:

> *Assessed for eligibility (n=138)*

Furthermore, the Acknowledgments section has been changed from:

> *This research was supported by the Brain Korea 21 FOUR project funded by the National Research Foundation of Korea, Yonsei University College of Nursing. We would like to thank Doing Lap for their cooperation and help throughout this research with their special technology. During the preparation of this, work the authors did not use generative artificial intelligence in any portion of the manuscript writing.*

And will now read as follows:

> *This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2017R1D1A1B04032264) and the Brain Korea 21 FOUR project funded by the National Research Foundation (NRF) of Korea, Yonsei University College of Nursing. We would like to thank Doing Lap for their cooperation and help throughout this research with their special technology. During the preparation of this, work the authors did not use generative artificial intelligence in any portion of the manuscript writing.*

The corrections will appear in the online version of the paper on the JMIR Publications website on February 27, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Figure 2. Summary of participation flow diagram.

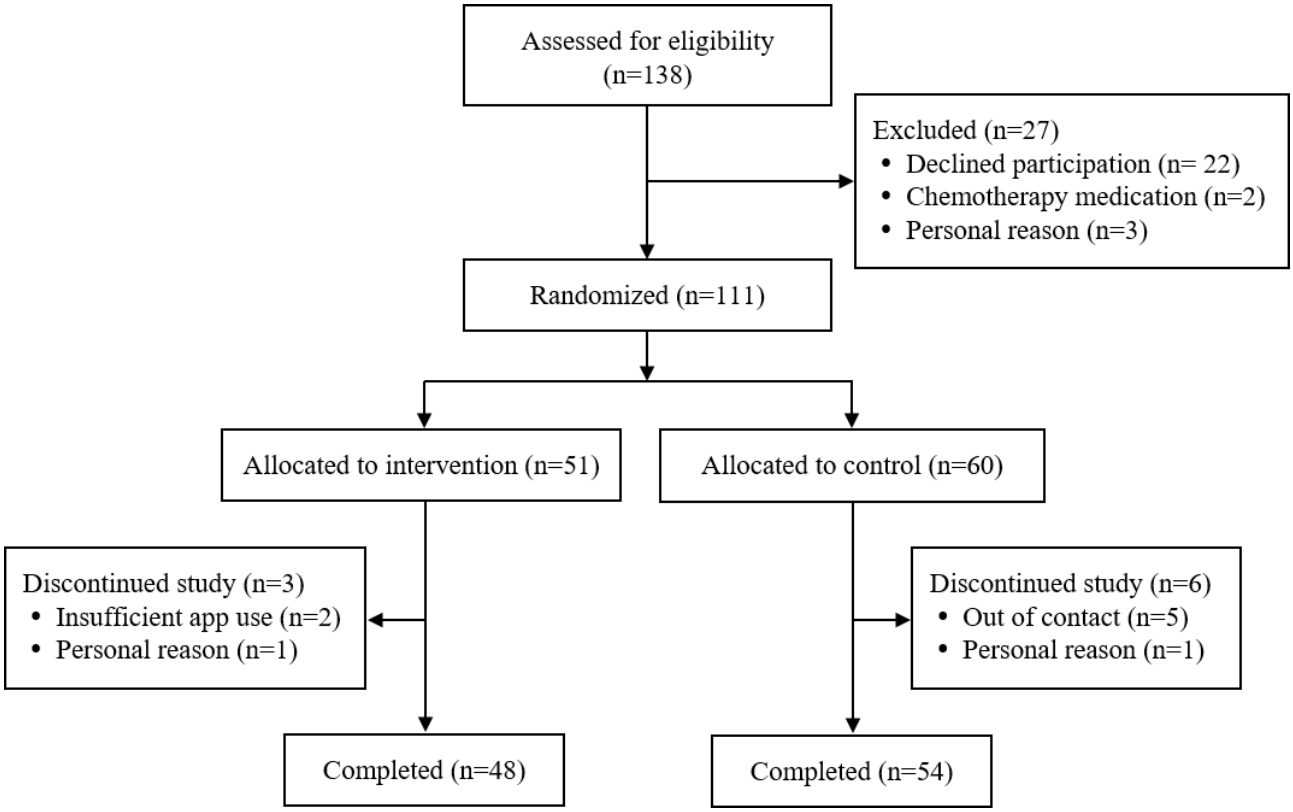

Submitted 19.02.24; this is a non–peer-reviewed article;accepted 22.02.24; published 27.02.24.

Please cite as:
Kwon OY, Lee MK, Lee HW, Kim H, Lee JS, Jang Y
Correction: Mobile App–Based Lifestyle Coaching Intervention for Patients With Nonalcoholic Fatty Liver Disease: Randomized Controlled Trial
J Med Internet Res 2024;26:e57499
URL: https://www.jmir.org/2024/1/e57499
doi:10.2196/57499
PMID:38412479

©Oh Young Kwon, Mi Kyung Lee, Hye Won Lee, Hyerang Kim, Jae Seung Lee, Yeonsoo Jang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Internet Tool to Support Self-Assessment and Self-Swabbing of Sore Throat: Development and Feasibility Study

Mark Lown[1], PhD; Kirsten Smith[2], PhD; Ingrid Muller[1], PhD; Catherine Woods[1], PhD; Emma Maund[3], PhD; Kirsty Rogers[4]; Taeko Becque[1], PhD; Gail Hayward[5], DPhil; Michael Moore[1], BMBS; Paul Little[1], MD, MRCP, MRCGP; Margaret Glogowska[5], PhD; Alastair Hay[6], MD; Beth Stuart[7], PhD; Efi Mantzourani[8], MPharm, MSc, PhD; Christopher R Wilcox[1], MRCGP; Natalie Thompson[1], MSc; Nick A Francis[1], BA, MD, PhD

[1]Primary Care Research Centre, University of Southampton, Southampton, United Kingdom
[2]School of Computing, University of Portsmouth, Portsmouth, United Kingdom
[3]School of Healthcare Enterprise and Innovation, University of Southampton, Southampton, United Kingdom
[4]Local Clinical Research Network Wessex, Southampton, United Kingdom
[5]Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford, United Kingdom
[6]Centre for Academic Primary Care, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[7]Pragmatic Clinical Trials Unit, Queen Mary University of London, London, United Kingdom
[8]Cardiff School of Pharmacy, Cardiff University, Cardiff, United Kingdom

Corresponding Author:
Margaret Glogowska, PhD
Nuffield Department of Primary Health Care Sciences, University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford, OX2 6GG
United Kingdom
Phone: 44 01865617872
Email: margaret.glogowska@phc.ox.ac.uk

Related Article:

Correction of: https://www.jmir.org/2023/1/e39791/

(J Med Internet Res 2024;26:e59420) doi:10.2196/59420

In "Internet Tool to Support Self-Assessment and Self-Swabbing of Sore Throat: Development and Feasibility Study" (J Med Internet Res 2023;25:e39791) the authors noted one error.

In the originally published article, the affiliations appeared as follows:

> *Mark Lown1 , PhD; Kirsten A Smith1 , PhD; Ingrid Muller1 , PhD; Catherine Woods1 , PhD; Emma Maund2 , PhD; Kirsty Rogers3 ; Taeko Becque1 , PhD; Gail Hayward4 , DPhil; Michael Moore1 , BMBS; Paul Little1 , MD, MRCP, MRCGP; Margaret Glogowska4 , PhD; Alastair Hay5 , MD; Beth Stuart6 , PhD; Efi Mantzourani7 , MPharm, MSc, PhD; Christopher R Wilcox1 , MRCGP; Natalie Thompson1 , MSc; Nick A Francis1 , BA, MD, PhD*

1. School of Computing, University of Portsmouth, Portsmouth, United Kingdom

2. School of Healthcare Enterprise and Innovation, University of Southampton, Southampton, United Kingdom

3. Local Clinical Research Network Wessex, Southampton, United Kingdom

4. Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford, United Kingdom

5. Centre for Academic Primary Care, Bristol Medical School, University of Bristol, Bristol, United Kingdom

6. Pragmatic Clinical Trials Unit, Queen Mary University of London, London, United Kingdom

7. Cardiff School of Pharmacy, Cardiff University, Cardiff, United Kingdom

The affiliations have been corrected as follows:

> *Mark Lown1 , PhD; Kirsten Smith2 , PhD; Ingrid Muller1 , PhD; Catherine Woods1 , PhD; Emma Maund3 , PhD; Kirsty Rogers4 ; Taeko Becque1 , PhD; Gail Hayward5 , DPhil; Michael Moore1 , BMBS; Paul Little1 , MSc, PhD; Margaret Glogowska5 , PhD; Alastair Hay6 , MD; Beth Stuart7 , PhD; Efi Mantzourani8 , MPharm, MSc, PhD;*

XSL•FO
RenderX

Christopher R Wilcox1 , MRCGP; Natalie Thompson1 , MSc; Natalie Thompson1 , MSc; Nick A Francis1 , BA, MD, PhD

1. Primary Care Research Centre, University of Southampton, Southampton, United Kingdom

2. School of Computing, University of Portsmouth, Portsmouth, United Kingdom

3. School of Healthcare Enterprise and Innovation, University of Southampton, Southampton, United Kingdom

4. Local Clinical Research Network Wessex, Southampton, United Kingdom

5. Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford, United Kingdom

6. Centre for Academic Primary Care, Bristol Medical School, University of Bristol, Bristol, United Kingdom

7. Pragmatic Clinical Trials Unit, Queen Mary University of London, London, United Kingdom

8. Cardiff School of Pharmacy, Cardiff University, Cardiff, United Kingdom

The correction will appear in the online version of the paper on the JMIR Publications website on April 26, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 11.04.24; this is a non–peer-reviewed article;accepted 16.04.24; published 26.04.24.

Please cite as:
Lown M, Smith K, Muller I, Woods C, Maund E, Rogers K, Becque T, Hayward G, Moore M, Little P, Glogowska M, Hay A, Stuart B, Mantzourani E, Wilcox CR, Thompson N, Francis NA
Correction: Internet Tool to Support Self-Assessment and Self-Swabbing of Sore Throat: Development and Feasibility Study
J Med Internet Res 2024;26:e59420
URL: https://www.jmir.org/2024/1/e59420
doi:10.2196/59420
PMID:38669681

©Mark Lown, Kirsten Smith, Ingrid Muller, Catherine Woods, Emma Maund, Kirsty Rogers, Taeko Becque, Gail Hayward, Michael Moore, Paul Little, Margaret Glogowska, Alastair Hay, Beth Stuart, Efi Mantzourani, Christopher R Wilcox, Natalie Thompson, Nick A Francis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation

Hyo-Jeong Ahn[1], MD; Eue-Keun Choi[1,2], MD, PhD; So-Ryoung Lee[1,2], MD, PhD; Soonil Kwon[1], MD; Hee-Seok Song[3], MSc; Young-Shin Lee[3], MSc; Seil Oh[1,2], MD, PhD

[1]Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[2]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[3]Seers Technology Co, Ltd, Seongnam-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Eue-Keun Choi, MD, PhD
Department of Internal Medicine
Seoul National University Hospital
101 Daehak-ro
Seoul, 03080
Republic of Korea
Phone: 82 220720688
Email: choiek417@gmail.com

Related Article:

Correction of: https://www.jmir.org/2024/1/e46098

(J Med Internet Res 2024;26:e59984) doi:10.2196/59984

In "Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation" (J Med Internet Res 2024;26:e46098) the authors noted one error.

In the "Ethics Approval" section, the institutional review board (IRB) number was originally published as:

E-2001-111-1096.

This has been changed to:

H-2103-010-1201.

The correction will appear in the online version of the paper on the JMIR Publications website on May 9, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 28.04.24; this is a non–peer-reviewed article;accepted 03.05.24; published 09.05.24.

Please cite as:
Ahn HJ, Choi EK, Lee SR, Kwon S, Song HS, Lee YS, Oh S
Correction: Three-Day Monitoring of Adhesive Single-Lead Electrocardiogram Patch for Premature Ventricular Complex: Prospective Study for Diagnosis Validation and Evaluation of Burden Fluctuation
J Med Internet Res 2024;26:e59984
URL: https://www.jmir.org/2024/1/e59984
doi:10.2196/59984
PMID:38723255

XSL•FO
RenderX

©Hyo-Jeong Ahn, Eue-Keun Choi, So-Ryoung Lee, Soonil Kwon, Hee-Seok Song, Young-Shin Lee, Seil Oh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study

Amy Thomas[1], MD; Andrea Asnes[1], MSW, MD; Kyle Libby[2], BA, MS; Allen Hsiao[1], MD, FAAP, FAMIA; Gunjan Tiyyagura[1], MHS, MD

[1]Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States
[2]3M | M*Modal, 3M Health Information Systems, 3M Company, Maplewood, MN, United States

Corresponding Author:
Gunjan Tiyyagura, MHS, MD
Department of Pediatrics
Yale University School of Medicine
Pediatric Emergency Medicine, PO Box 208064
New Haven, CT, 06520
United States
Phone: 1 203 464 6343
Fax: 1 203 737 7447
Email: gunjan.tiyyagura@yale.edu

Related Article:

Correction of: https://www.jmir.org/2024/1/e51058

(J Med Internet Res 2024;26:e60444) doi:10.2196/60444

In “Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study” (J Med Internet Res 2024;26:e51058), the authors noted one error.

The "Background" section of the abstract was originally published as:

> *Despite the impact of physical abuse on children, it is often underdiagnosed, especially among children evaluated in general emergency departments (EDs) and those belonging to racial or ethnic minority groups. Electronic clinical decision support (CDS) can improve the recognition of child physical abuse.*

It will now be changed to:

> *Despite the impact of physical abuse on children, it is often underdiagnosed, especially among children evaluated in emergency departments (EDs). Electronic clinical decision support (CDS) can improve the recognition of child physical abuse.*

The correction will appear in the online version of the paper on the JMIR Publications website on May 27, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 11.05.24; this is a non–peer-reviewed article;accepted 16.05.24; published 27.05.24.

Please cite as:
Thomas A, Asnes A, Libby K, Hsiao A, Tiyyagura G
Correction: Developing and Testing the Usability of a Novel Child Abuse Clinical Decision Support System: Mixed Methods Study
J Med Internet Res 2024;26:e60444
URL: https://www.jmir.org/2024/1/e60444
doi:10.2196/60444
PMID:38801770

XSL•FO
RenderX

©Amy Thomas, Andrea Asnes, Kyle Libby, Allen Hsiao, Gunjan Tiyyagura. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Experiences of Using Digital Mindfulness-Based Interventions: Rapid Scoping Review and Thematic Synthesis

Emma Louise Osborne[1], BA, MRES; Ben Ainsworth[1,2], PhD; Nic Hooper[3], PhD; Melissa Jayne Atkinson[1], PhD

[1]Department of Psychology, University of Bath, Bath, United Kingdom
[2]School of Psychology, University of Southampton, Southampton, United Kingdom
[3]School of Psychology, Cardiff University, Cardiff, United Kingdom

Corresponding Author:
Emma Louise Osborne, BA, MRES
Department of Psychology
University of Bath
10 West
Claverton Down
Bath, BA2 7AY
United Kingdom
Phone: 44 383843 ext 01225
Email: elo25@bath.ac.uk

Related Article:

Correction of: https://www.jmir.org/2023/1/e44220

(J Med Internet Res 2024;26:e60967) doi:10.2196/60967

In "Experiences of Using Digital Mindfulness-Based Interventions: Rapid Scoping Review and Thematic Synthesis" (J Med Internet Res 2023;25:e44220) the authors noted one error.

In the "Results" section of the Abstract, the following sentence:

> *The search identified 510 studies, 22 (4.3%) of which met the inclusion criteria.*

Has been corrected to:

The search identified 530 studies, 22 (4.2%) of which met the inclusion criteria.

The correction will appear in the online version of the paper on the JMIR Publications website on May 29, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 27.05.24; this is a non–peer-reviewed article;accepted 28.05.24; published 29.05.24.

Please cite as:
Osborne EL, Ainsworth B, Hooper N, Atkinson MJ
Correction: Experiences of Using Digital Mindfulness-Based Interventions: Rapid Scoping Review and Thematic Synthesis
J Med Internet Res 2024;26:e60967
URL: https://www.jmir.org/2024/1/e60967
doi:10.2196/60967
PMID:38809581

©Emma Louise Osborne, Ben Ainsworth, Nic Hooper, Melissa Jayne Atkinson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation

Han Shi Jocelyn Chew[1], PhD; Nicholas WS Chew[2], MBBS; Shaun Seh Ern Loong[3], MBBS; Su Lin Lim[4], PhD; Wai San Wilson Tam[1], PhD; Yip Han Chin[3], MBBS; Ariana M Chao[5], PhD; Georgios K Dimitriadis[6], MBBS, MSc; Yujia Gao[7]; Jimmy Bok Yan So[8], MB ChB, FRCS, MPH; Asim Shabbir[8], MBBS, MMed, FRCS; Kee Yuan Ngiam[9], MBBS, FRCS

[1]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2]Department of Cardiology, National University Hospital, Singapore, Singapore
[3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[4]Department of Dietetics, National University Hospital, Singapore, Singapore
[5]School of Nursing, Johns Hopkins University, Baltimore, MD, United States
[6]Department of Endocrinology ASO/EASO COM, King's College Hospital NHS Foundation Trust, London, United Kingdom
[7]Division of Hepatobiliary & Pancreatic Surgery, Department of Surgery, National University Hospital, Singapore, Singapore
[8]Division of General Surgery (Upper Gastrointestinal Surgery), Department of Surgery, National University Hospital, Singapore, Singapore
[9]Division of Thyroid & Endocrine Surgery, Department of Surgery, National University Hospital, Singapore, Singapore

Corresponding Author:
Han Shi Jocelyn Chew, PhD
Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore
Level 3, Clinical Research Centre
Block MD11, 10 Medical Drive
Singapore, 117597
Singapore
Phone: 65 65168687
Email: jocelyn.chew.hs@nus.edu.sg

Related Article:

Correction of: https://www.jmir.org/2024/1/e46036

(J Med Internet Res 2024;26:e62767) doi:10.2196/62767

In “Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation” (J Med Internet Res 2024;26:e46036) the authors noted one error.

The last name of author GKD:

Dimitriadish

Has been revised to:

Dimitriadis

The correction will appear in the online version of the paper on the JMIR Publications website on June 7, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 30.05.24; this is a non–peer-reviewed article;accepted 04.06.24; published 07.06.24.

Please cite as:
Chew HSJ, Chew NWS, Loong SSE, Lim SL, Tam WSW, Chin YH, Chao AM, Dimitriadis GK, Gao Y, So JBY, Shabbir A, Ngiam KY
Correction: Effectiveness of an Artificial Intelligence-Assisted App for Improving Eating Behaviors: Mixed Methods Evaluation
J Med Internet Res 2024;26:e62767
URL: https://www.jmir.org/2024/1/e62767
doi:10.2196/62767
PMID:38848558

XSL•FO
RenderX

©Han Shi Jocelyn Chew, Nicholas WS Chew, Shaun Seh Ern Loong, Su Lin Lim, Wai San Wilson Tam, Yip Han Chin, Ariana M Chao, Georgios K Dimitriadis, Yujia Gao, Jimmy Bok Yan So, Asim Shabbir, Kee Yuan Ngiam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Author Correction: Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study

Dianna Vidas[1], BPsycSc, MSc, PhD; Romina Carrasco[1,2], BA, MA, PhD; Ryan M Kelly[1,3], BSc, MSc, PhD; Jenny Waycott[1], BA, PhD; Jeanette Tamplin[4], BMus (Hons), MMus, PhD; Kate McMahon[4], BA, MMusThrpy, PhD; Libby M Flynn[4], BMus, MMusThrpy, PhD; Phoebe A Stretton-Smith[4], BMus, MMusThrpy; Tanara Vieira Sousa[4], BA, MSc, PhD; Felicity A Baker[4,5], BMus, GDip, MMus, RMT, PhD

[1]School of Computing and Information Systems, The University of Melbourne, Parkville, Australia
[2]Facultad de Comunicación y Artes Audiovisuales, University of Las Américas Ecuador, Quito, Ecuador
[3]School of Computing Technologies, RMIT University, Melbourne, Australia
[4]Faculty of Fine Arts and Music, The University of Melbourne, Melbourne, Australia
[5]Centre for Research in Music and Health, Norwegian Academy of Music, Oslo, Norway

Corresponding Author:
Dianna Vidas, BPsycSc, MSc, PhD
School of Computing and Information Systems
The University of Melbourne
Grattan Street
Parkville, 3010
Australia
Phone: 61 3 90355511
Email: dianna.vidas@unimelb.edu.au

Related Article:

Correction of: https://www.jmir.org/2024/1/e54186

(J Med Internet Res 2024;26:e65977) doi:10.2196/65977

In "Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study" (J Med Internet Res 2024;26:e54186) the authors noted one error.

The ninth author's name was listed as:

Tanara Viera Sousa

This has been changed to the following:

Tanara Vieira Sousa

The correction will appear in the online version of the paper on the JMIR Publications website on September 9, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 30.08.24; this is a non–peer-reviewed article;accepted 04.09.24; published 09.09.24.

Please cite as:
Vidas D, Carrasco R, Kelly RM, Waycott J, Tamplin J, McMahon K, Flynn LM, Stretton-Smith PA, Vieira Sousa T, Baker FA
Author Correction: Everyday Uses of Music Listening and Music Technologies by Caregivers and People With Dementia: Survey and Focus Group Study
J Med Internet Res 2024;26:e65977
URL: https://www.jmir.org/2024/1/e65977
doi:10.2196/65977
PMID:

XSL•FO
RenderX

©Dianna Vidas, Romina Carrasco, Ryan M Kelly, Jenny Waycott, Jeanette Tamplin, Kate McMahon, Libby M Flynn, Phoebe A Stretton-Smith, Tanara Vieira Sousa, Felicity A Baker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study

Yun Kwan Kim[1,2], MSc; Won-Doo Seo[1], MSc; Sun Jung Lee[1], PhD; Ja Hyung Koo[1], PhD; Gyung Chul Kim[1], MSE; Hee Seok Song[1], MSE; Minji Lee[3], PhD

[1]Technology Development, Seers Technology Co. Ltd., Pyeongtaek-si, Gyeonggi-do, Republic of Korea

[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

[3]Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Minji Lee, PhD
Department of Biomedical Software Engineering
The Catholic University of Korea
43, Jibong-ro
Bucheon-si, Gyeonggi-do, 14662
Republic of Korea
Phone: 82 2 2164 4364
Email: minjilee@catholic.ac.kr

Related Article:

Correction of: https://www.jmir.org/2024/1/e62890

(J Med Internet Res 2024;26:e67135) doi:10.2196/67135

In “Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study” (J Med Internet Res 2024;26:e62890) the authors noted one error:

In "Results", "Subgroup Analysis", Figure 4 has been replaced with Figure 5, as follows:

XSL•FO
RenderX

Figure 4. Model performance in difference patient cohorts from MIMIC-IV and eICU-CRD. (A) AUROC on ICU types of MIMIC-IV. (B) AUROC on ICU types of eICU-CRD. Boxes in the box plot show IQR and the cross marks are outliers with values that lie outside the minimum and maximum ranges of the whiskers, where minimum = Q1 - 1.5 × IQR and maximum = Q3 + 1.5 × IQR. * Statistically significant (*P*<.05). AUROC: area under the receiver operating characteristic curve; DEWS: Deep Learning–Based Early Warning Score; eICU-CRD: eICU-Collaborative Research Database; ICU: intensive care unit; KNN: k-nearest neighbors; LGBM: light gradient boosting method; LR: logistic regression; MIMIC: Medical Information Mart for Intensive Care; MLP: Multilayer perceptron; NEWS: National Early Warning Score; Q1: first quartile; Q3: third quartile; RETAIN: reverse time attention; SAPS: Simplified Acute Physiology Score.

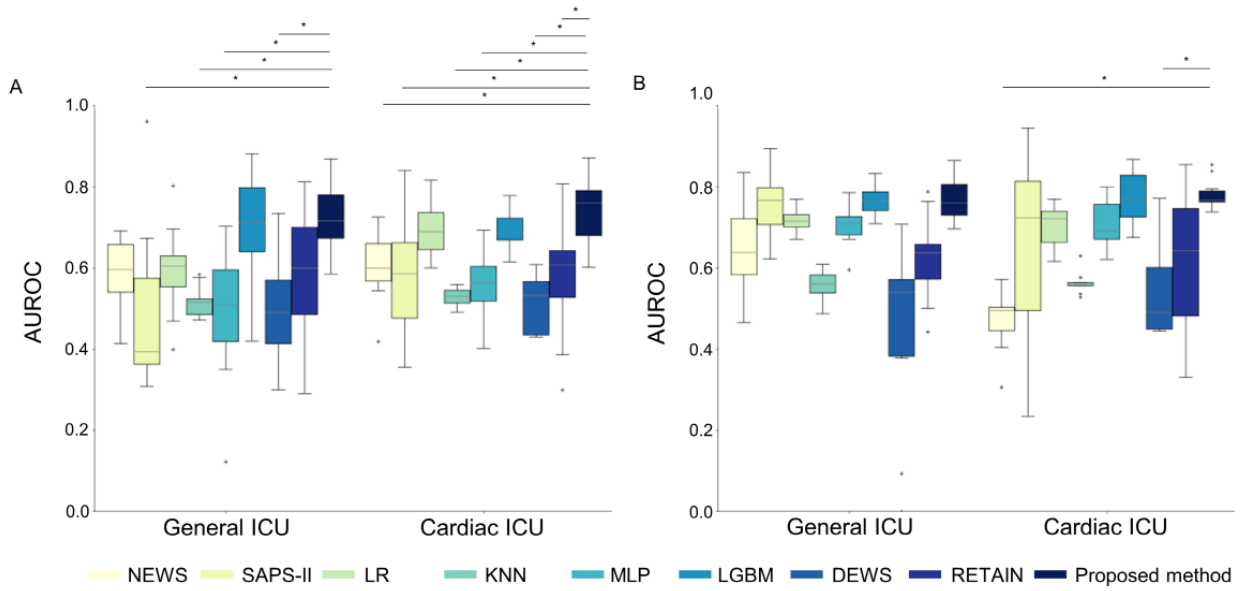

In "Results", "External Validation", Figure 5 has been replaced with Figure 4, as follows:

Figure 5. Cross–data set external validation AUROC performance. (A) eICU after training MIMIC-IV. (B) MIMIC-IV after training eICU-CRD. AUROC: area under the receiver operating characteristic curve; DEWS: Deep Learning–Based Early Warning Score; eICU-CRD: eICU-Collaborative Research Database; KNN: k-nearest neighbors; LGBM: light gradient boosting method; LR: logistic regression; MIMIC: Medical Information Mart for Intensive Care; MLP: multilayer perceptron; NEWS: National Early Warning Score; RETAIN: reverse time attention.

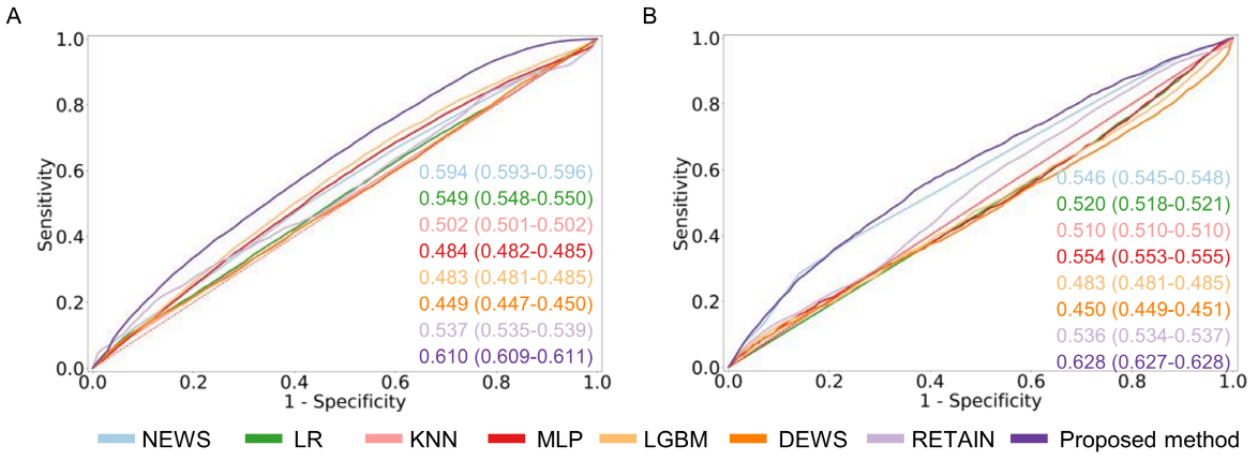

The correction will appear in the online version of the paper on the JMIR Publications website on October 28, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 09.10.24; this is a non–peer-reviewed article;accepted 09.10.24; published 28.10.24.

Please cite as:
Kim YK, Seo WD, Lee SJ, Koo JH, Kim GC, Song HS, Lee M
Correction: Early Prediction of Cardiac Arrest in the Intensive Care Unit Using Explainable Machine Learning: Retrospective Study
J Med Internet Res 2024;26:e67135
URL: https://www.jmir.org/2024/1/e67135
doi:10.2196/67135
PMID:

©Yun Kwan Kim, Won-Doo Seo, Sun Jung Lee, Ja Hyung Koo, Gyung Chul Kim, Hee Seok Song, Minji Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study

Kelly Quinn[1], MBA, MSLIS, PhD; Sarah Leiser Ransom[1], MA; Carrie O'Connell[1], MA; Naoko Muramatsu[2], PhD; David X Marquez[3], PhD; Jessie Chin[4], PhD

[1]Department of Communication, University of Illinois Chicago, Chicago, IL, United States

[2]Division of Community Health Sciences, School of Public Health, University of Illinois Chicago, Chicago, IL, United States

[3]Department of Kinesiology and Nutrition, University of Illinois Chicago, Chicago, IL, United States

[4]School of Information Sciences, University of Illinois Urbana-Champaign, Urbana, IL, United States

Corresponding Author:
Kelly Quinn, MBA, MSLIS, PhD
Department of Communication
University of Illinois Chicago
1007 W Harrison St
1140 BSB, MC 132
Chicago, IL, 60607
United States
Phone: 1 312 996 3187
Email: kquinn8@uic.edu

Related Article:

Correction of: https://jmir.org/2024/1/e54800/

(J Med Internet Res 2024;26:e66813) doi:10.2196/66813

In “Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study” (J Med Internet Res 2024;26:e54800) the authors would like to make an addendum to the Acknowledgements:

The corrected text should read:

> *This work was supported by the the National Center for Advancing Translational Sciences, National Institutes of Health [UL1TR002003] and the National Institute on Aging of the National Institutes of Health [R24AG064191, P30AG022849]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. The authors would like to thank Smit Desai for his contribution to an earlier version of the manuscript. We also would like to thank the following individuals for their assistance with this project: Smit Desai, Jamie Gradishar, Daniel Hlavacek, Faraz Hussain, Caniece Leggett, Fernando Mendez, and Melissa Pierce.*

The correction will appear in the online version of the paper on the JMIR Publications website on November 15, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 23.09.24; this is a non–peer-reviewed article;accepted 26.09.24; published 15.11.24.

Please cite as:
Quinn K, Leiser Ransom S, O'Connell C, Muramatsu N, Marquez DX, Chin J
Correction: Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study
J Med Internet Res 2024;26:e66813
URL: https://www.jmir.org/2024/1/e66813
doi:10.2196/66813
PMID:

©Kelly Quinn, Sarah Leiser Ransom, Carrie O'Connell, Naoko Muramatsu, David X Marquez, Jessie Chin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey

Yuka Sugawara[1], MD, PhD; Yosuke Hirakawa[1], MD, PhD; Masao Iwagami[2], MSc, MD, PhD; Haruo Kuroki[3], MD, PhD; Shuhei Mitani[1], MD; Ataru Inagaki[4], MD, PhD; Hiroki Ohashi[5], MD; Mitsuru Kubota[6], MD, PhD; Soichi Koike[7], MD, PhD; Rie Wakimizu[8], PhD; Masaomi Nangaku[1], MD, PhD

[1]Division of Nephrology and Endocrinology, The University of Tokyo, Tokyo, Japan
[2]Department of Health Services Research, Institute of Medicine, University of Tsukuba, Ibaraki, Japan
[3]Clinic Pauroom, Tokyo, Japan
[4]Department of Education, College of Education, Psychology and Human Studies, Aoyama Gakuin University, Tokyo, Japan
[5]TAMA Family Clinic, Tokyo, Japan
[6]Department of General Pediatrics and Interdisciplinary Medicine, National Center for Child Health and Development, Tokyo, Japan
[7]Division of Health Policy and Management, Center for Community Medicine, Jichi Medical University, Tochigi, Japan
[8]Department of Child Health and Development Nursing, Institute of Medicine, University of Tsukuba, Ibaraki, Japan

Corresponding Author:
Masaomi Nangaku, MD, PhD
Division of Nephrology and Endocrinology
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 3 3815 5411
Email: mnangaku@m.u-tokyo.ac.jp

Related Article:

Correction of: https://www.jmir.org/2024/1/e64159

(J Med Internet Res 2024;26:e68439) doi:10.2196/68439

In "Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey" ([J Med Internet Res 2024;1(26):e64159]) the authors made on correction.

In the originally published manuscript, the following sentence:

> *"Consequently, telepsychiatry has become widespread in many countries, and in 15 of 17 regions in Japan,…"*

Was corrected as follows:

> *"Consequently, telepsychiatry has become widespread in many countries, and in 15 of 17 regions in the world,…".*

The correction will appear in the online version of the paper on the JMIR Publications website on November 27, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 12.11.24; this is a non–peer-reviewed article;accepted 14.11.24; published 27.11.24.

Please cite as:
Sugawara Y, Hirakawa Y, Iwagami M, Kuroki H, Mitani S, Inagaki A, Ohashi H, Kubota M, Koike S, Wakimizu R, Nangaku M
Correction: Issues in the Adoption of Online Medical Care: Cross-Sectional Questionnaire Survey
J Med Internet Res 2024;26:e68439
URL: https://www.jmir.org/2024/1/e68439
doi:10.2196/68439
PMID:

XSL•FO
RenderX

©Yuka Sugawara, Yosuke Hirakawa, Masao Iwagami, Haruo Kuroki, Shuhei Mitani, Ataru Inagaki, Hiroki Ohashi, Mitsuru Kubota, Soichi Koike, Rie Wakimizu, Masaomi Nangaku. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers

Anna Jolliff[1], MS, LMHC; Richard J Holden[1], PhD; Rupa Valdez[2,3], PhD; Ryan J Coller[4], MPH, MD; Himalaya Patel[1,5], PhD; Matthew Zuraw[6], MBA; Anna Linden[7], PhD; Aaron Ganci[8], PhD; Christian Elliott[6]; Nicole E Werner[1], PhD

[1]Department of Health & Wellness Design, School of Public Health - Bloomington, Indiana University Bloomington, Bloomington, IN, United States

[2]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, United States

[3]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, United States

[4]Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin - Madison, Madison, WI, United States

[5]Health Systems Research Center for Health Information and Communication, Richard L Roudebush Veterans Affairs Medical Center, United States Department of Veterans Affairs, Indianapolis, IN, United States

[6]CareVirtue Technologies, San Diego, CA, United States

[7]Department of Industrial & Systems Engineering, University of Wisconsin - Madison, Madison, WI, United States

[8]Department of Visual Communication Design, Herron School of Art & Design, Indiana University Indianapolis, Indianapolis, IN, United States

Corresponding Author:
Anna Jolliff, MS, LMHC
Department of Health & Wellness Design
School of Public Health - Bloomington
Indiana University Bloomington
1025 E. 7th Street
Bloomington, IN, 47405
United States
Phone: 1 812 855 1561
Email: annjoll@iu.edu

Related Article:

Correction of: https://www.jmir.org/2024/1/e60353

(J Med Internet Res 2024;26:e69714) doi:10.2196/69714

In "Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers," (J Med Internet Res 2024;26:e60353), the authors noted one error.

The following author's affiliation was listed as:

> *Himalaya Patel[5]*
>
> *[5]Health Systems Research Center for Health Information and Communication, Richard L Roudebush Veterans Affairs Medical Center, United States Department of Veterans Affairs, Indianapolis, IN, United States*

The author's affiliation has been corrected to:

> *Himalaya Patel[1,5]*
>
> *[1]Department of Health & Wellness Design, School of Public Health - Bloomington, Indiana University Bloomington, Bloomington, IN, United States*
>
> *[5]Health Systems Research Center for Health Information and Communication, Richard L Roudebush Veterans Affairs Medical Center, United States Department of Veterans Affairs, Indianapolis, IN, United States*

Additionally, middle initial J has been added to author name Ryan J Coller. The name has been changed from:

> *Ryan Coller*

to

> *Ryan J Coller*

The correction will appear in the online version of the paper on the JMIR Publications website on December 31, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 06.12.24; this is a non–peer-reviewed article;accepted 11.12.24; published 31.12.24.

Please cite as:
Jolliff A, Holden RJ, Valdez R, Coller RJ, Patel H, Zuraw M, Linden A, Ganci A, Elliott C, Werner NE
Correction: Investigating the Best Practices for Engagement in Remote Participatory Design: Mixed Methods Analysis of 4 Remote Studies With Family Caregivers
J Med Internet Res 2024;26:e69714
URL: https://www.jmir.org/2024/1/e69714
doi:10.2196/69714
PMID:

©Anna Jolliff, Richard J Holden, Rupa Valdez, Ryan J Coller, Himalaya Patel, Matthew Zuraw, Anna Linden, Aaron Ganci, Christian Elliott, Nicole E Werner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Outpatient Video Visits During the COVID-19 Pandemic: Cross-Sectional Survey Study of Patients' Experiences and Characteristics

Stefanie C van den Bosch[1*], MD, DDS; Demi van Dalen[2*], MD; Marjan Meinders[3], PhD; Harry van Goor[2], MD, PhD; Stefaan Bergé[1], MD, PhD, DDS; Martijn Stommel[2*], MD, PhD; Sandra van Dulmen[4,5,6*], PhD

[1]Department Oral and Maxillofacial Surgery, Radboud University Medical Center, Nijmegen, Netherlands
[2]Department of Surgery, Radboud University Medical Center, Nijmegen, Netherlands
[3]IQ Healthcare, Radboud University Medical Center, Nijmegen, Netherlands
[4]Department of Primary and Community Care, Radboud University Medical Center, Nijmegen, Netherlands
[5]Faculty of Caring Science, Work Life and Social Welfare, University of Boras, Boras, Sweden
[6]Nivel, Netherlands Institute for Health Services Research, Utrecht, Netherlands
[*]these authors contributed equally

Corresponding Author:
Demi van Dalen, MD
Department of Surgery
Radboud University Medical Center
Geert Grooteplein Zuid 10
6525 GA
Nijmegen, 6525 GA
Netherlands
Phone: 31 243611111
Email: demi.vandalen@radboudumc.nl

Abstract

Background: During the first lockdown of the COVID-19 pandemic, an exponential increase in video consultations replacing in-person outpatient visits was observed in hospitals. Insight into patients' experiences with this type of consultation is helpful for a broad, sustainable, and patient-centered implementation of video consultation.

Objective: This study aims to examine patients' experiences with video consultation during the COVID-19 pandemic and identify discriminative patient and consultation characteristics to determine when video consultation is most feasible.

Methods: A cross-sectional survey study was conducted. Patients aged ≥18 years and scheduled for a video consultation at the outpatient clinic of a Dutch university medical center from August 2020 to December 2020 for all medical specialties were eligible. Patients' experiences were explored through a study-specific survey using descriptive quantitative statistics. Open-ended questions were qualitatively analyzed and thematically categorized into appreciated aspects and aspects for improvement. Discriminative patient and consultation characteristics were identified using 3 distinctive survey items. Characteristics of patients who scored and those who did not score all 3 items positively were analyzed using binary logistic regression.

Results: A total of 1054 patients were included in the analysis. Most patients (964/1054, 91.46%) were satisfied with their video consultation, with a mean overall grade of 8.6 (SD 1.3) of 10. In the qualitative analyses, 70.02% (738/1054) of the patients cited aspects they appreciated and 44.97% (474/1054) mentioned aspects for improvement during their consultation. Patients with better self-rated health reported a positive evaluation significantly more often (P=.001), which also held true for other medical specialties (vs surgical and nonsurgical specialties; P<.001).

Conclusions: Video consultation was perceived as highly satisfactory by patients during the COVID-19 pandemic, with the best experience reported by healthy participants and those undergoing their first consultation. Appreciated aspects are mainly at the individual professional level, organizational level, and innovation level itself. The aspects that were mentioned for improvement can be changed for the better.

(J Med Internet Res 2024;26:e49058) doi:10.2196/49058

XSL•FO
RenderX

KEYWORDS

telemedicine; video visit; remote consultation; eHealth; patient-centered care; COVID-19; mobile phone

Introduction

Background

In the Netherlands, the COVID-19 pandemic led to the first lockdown from March 2020 to June 2020 and the second lockdown from mid-December 2020 to April 2021. Throughout these periods, in-person visits were minimized to reduce the risk of potential virus transmission. As a substitute for in-person visits, the use of video visits significantly increased across many medical specialties.

Video visits were already in practice on a limited scale in a wide range of contexts: for speech evaluation in patients with cleft palate [1], genetic counseling [2], follow-up after facial plastic surgery [3], and postoperative wound assessment [4]. Video visits were found to be satisfactory for patients [5] and enabled empathetic patient-professional relationships remotely [1,6,7]. However, large-scale implementation in daily practice was found to be challenging owing to the multilevel complexity of implementation, where people, organizations, and technology continuously interconnect and develop [8-11]. For example, attitudes and beliefs of individual professionals have been shown to act as both facilitators and barriers in the implementation of eHealth applications [8]. Furthermore, video visits appeared to be particularly successful in follow-up appointments, when a preexisting relationship of trust is established between the patient and clinician [10]. For instance, video visits were more easily adopted in follow-up care after cancer surgery compared with a multidisciplinary context of antenatal diabetes care [10].

Owing to the pandemic-driven, accelerated application of video visits, many clinicians gained experience with this mode of health care delivery. This way of providing care offers several advantages, including saving travel time and costs for both patients and their companions and the efficient use of health care resources, such as outpatient clinic space and support [12-14]. As we move into the post–COVID-19 era, video visits are expected to persist as a routine practice, but large-scale use has seemed to stagnate, presumably owing to a lack of guidance, vision, and attention to patients' needs, as observed in the United States [15,16]. The use and implementation of video visits are expected to be most successful when tailored to the needs of patients, clinicians, and health care organizations [17,18]. Therefore, it is crucial to understand patients' perspectives and experiences with video visits and identify specific patient groups that show a greater or lesser degree of appreciation for and suitability to video visits [19]. Large studies with diverse patient populations covering all medical specialties need to be conducted to learn more about patients' perspectives [16,18]. However, we are concerned that only a limited number of studies have been published that evaluated video visits for patients within large, diverse populations. Consequently, this study contributes significantly to the existing body of knowledge in this area [20-22].

Aims

The primary aim of this study was to examine patients' evaluations of video visits in the context of the COVID-19 pandemic. The secondary aim was to identify patient groups for whom video visits are relatively more suitable, given their positive experiences. With these results, the first step toward patient-tailored choices for type of visit can be made.

Methods

Study Design and Population

A cross-sectional study was conducted from August 2020 to December 2020 at the Radboud University Medical Center (Radboudumc). Zaurus was used as the video visit app, which is compatible with all smartphones and tablets. Patients were invited via email to register and download the app.

Patients scheduled for a video visit were automatically selected based on the registered mode of visit. Links to the questionnaire were sent by an independent research firm (Expoints) on behalf of the Radboudumc. Selected patients received the survey within 8 days after their visit via email to evaluate the visit and collect their sociodemographic details. The survey had to be completed within 2 weeks, and a reminder was sent 1 week after the initial invitation. An incomplete survey could be saved to be completed later (within 2 weeks). No reminder for completion was sent.

All patients aged ≥18 years who received a video visit at an outpatient clinic at Radboudumc were eligible for inclusion. When a patient had multiple video visits in the selected period, the most recent video visit was selected.

Patients were excluded if they had cognitive problems; had difficulties with reading and understanding Dutch owing to a hindering language barrier; were deceased at the time of selection; completed a survey regarding video visits in the 180 days before the start of our study; completed a survey regarding their admission experience or experience with an in-person visit 30 days before the start of our study; or were admitted to the hospital, as priority was given to the patient experience survey regarding admission. In addition, when the video visit was a follow-up visit after giving birth or when the video visit was replaced with another visit modality, the patient was excluded.

Ethical Considerations

All patients participated voluntarily and anonymously in the survey and gave informed consent to use their data in accordance with the General Data Protection Regulation. Ethics approval was requested and waived by the local Medical Research Ethics Committee of Radboudumc (CMO [committee on research involving human subjects] Oost-Nederland; registration number 2021-8415).

Survey

A combined survey was used, which consisted of the Patient Experience Monitor (PEM) for adult outpatient experience [23], developed by the Dutch Federation of University Medical

XSL•FO
RenderX

Centers, and the patient satisfaction survey for video visits created by Hanna et al [7]. This combined survey was constructed after extensive deliberation by an expert panel. In this process, a literature review of surveys specifically about video visits was performed. The experts found the survey by Hanna et al [7] to be the most suitable for the aim of our study.

The PEM survey was constructed by adapting a validated Picker Institute survey following a comprehensive theory-driven approach of item selection by an expert panel, cognitive interviews with patients, analysis of psychometric properties, and member checking. This survey of 14 items represents eight key domains of person-centered care: (1) fast access to reliable health care advice; (2) effective treatment delivered by trusted professionals; (3) continuity of care and smooth transitions; (4) involvement and support for family and caregivers; (5) clear information, communication, and support for self-care; (6) involvement in making decisions and respect for preferences; (7) emotional support, empathy, and respect; and (8) attention to physical and environmental needs [24]. The Picker Institute surveys are measures for evaluating patients' experiences in outpatient and inpatient clinical care and have been validated and extensively used in university medical centers in the Netherlands since 2019 [25]. The PEM survey was adapted to the videoconferencing setting by rephrasing the questions. Overall, 2 items were open-ended questions (Multimedia Appendix 1). Quantitative analyses of the PEM survey were based on individual survey items, whereas open-ended questions were analyzed using qualitative methods.

The 13-item survey by Hanna et al [7] is designed according to the principles of survey development for telemedicine to evaluate patients' experiences with video visits in pain clinics [26]. This survey was translated into Dutch, and 1 item was removed, as it was already covered by the PEM survey (Multimedia Appendix 1, items 16-27). In total, 4 items were rephrased based on the advice of the patient communication experts. Analysis of the survey by Hanna et al [7] is based on an overall sum score, where a higher overall sum score represents greater satisfaction. For correct analysis and to calculate an overall sum score, the 3 negatively phrased questions (items 17, 21, and 23) were reversed (eg, "No, definitely not" was converted to "Yes, definitely"), according to protocol.

Finally, a question was added to assess the visit by assigning a score on a scale ranging from 1 to 10 (with 10 being most positive). The survey was conducted according to CHERRIES (Checklist for Reporting Results of Internet-Based e-Surveys; Multimedia Appendix 2) [27].

Statistical Analysis

Quantitative Analysis

Descriptive statistics were used for the closed-ended items (Multimedia Appendix 1, items 1-11 and 16-27). To identify patient and visit characteristics associated with positive evaluation of video visits, 4 authors found consensus upon the 3 key items from the survey by Hanna et al [7] that stood out the most (item 18: "the care I received by a video visit was just as good as with an in-person appointment"; item 22: "I was comfortable talking by video to the healthcare professional"; item 27: "I would recommend the video visit option to other patients"). Throughout the paper, these 3 items have been referred to as "crucial" components of the survey by Hanna et al [7], as they best displayed a positive experience.

Patient and visit characteristics were determined for the group answering the 3 crucial items positively. In this analysis, the following characteristics were included: sex, age category, level of education, self-rated health, type of visit (first vs follow-up), and medical specialty (surgical, nonsurgical, or other). High self-rated health was defined as a score that indicates "very well" or "excellent."

Statistical analysis was performed using SPSS Statistics (version 25; IBM Corp). Binary logistic regression analysis was used to calculate differences in patient and visit characteristics between the patient subgroup that scored positively on all crucial items and the patient subgroup that did not score positively, as the dependent variable was not normally distributed.

Qualitative Analysis

Qualitative analysis was performed on the open-ended questions (items 13 and 14) to identify appreciated aspects and aspects for improvement for video visits. Overall, 2 authors (SCvdB and DD) independently categorized all the answers into six categories of factors that influence the implementation of innovations: (1) the innovation itself, (2) the individual professional, (3) the patient, (4) social context, (5) organizational context, and (6) economic and political context [28]. In case of conflicts in the categorization, consensus was reached through discussion between the authors. Responses including multiple levels within a single response were counted as individual items. Examples of answers for both aspects in each category have been cited in the Results section.

Results

Quantitative Results

From August 2020 to December 2020, a total of 1244 surveys were completed, with a response rate of 28.32% (1244/4392). After excluding 15.27% (190/1244) of the patients who reported that the visit was either a telephone consultation or replaced by telephone after technical difficulties, 84.73% (1054/1244) of the surveys were used in the analysis.

Table 1 shows the patients' demographics. An equal distribution across age categories was observed. Clinical genetics, neurology, and medical oncology accounted for 65.84% (694/1054) of the total number of evaluated video visits, whereas the distribution across the other medical specialties varied widely. The numbers of first and follow-up visits were equal, with most follow-up visits (480/1054, 45.54%) performed by a known clinician. After a video visit, 36.91% (389/1054) of the patients had to make an appointment for an additional in-person visit or medical examination. A follow-up visit via video was planned in 40.32% (425/1054) of the evaluated video visits. Clinicians from medical oncology and neurology more frequently scheduled an in-person follow-up visit for their patients—18.6% (40/215) and 23.3% (50/215), respectively. Moreover, in 39.7% (69/174) and 14.4% (25/174) of cases, the visits provided by clinical genetics and

XSL•FO
RenderX

neurology respectively, were followed by a consecutive visit for additional (diagnostic) testing.

The overall grading for the video visit had a mean of 8.6 (SD 1.3; median 9) of 10. For 5 PEM items, >80% of the patients answered positively, that is, patients waited no longer than 5 minutes, clinicians had read their medical records well, patients received understandable answers, patients trusted the clinician, and patients had enough time to discuss their problems with the clinician. Refer to Table 2 for details about the responses of patients.

For the remaining 6 items, more than 21% stated that the item was either not applicable or answered positively. For instance, 80.25% (829/1033) indicated not receiving any new medication for the question about whether the professional explained the adverse effects of new medication. Analysis of the items in the survey by Hanna et al [7] showed that 91.46% (964/1054) of the patients was satisfied with their video visit, 66.98% (706/1054) found it to be just as good as an in-person visit, and 68.69% (724/1054) would recommend video visits to other patients, as shown in Table 3.

Of 1054 patients, 574 (54.46%) answered all 3 crucial items on the survey by Hanna et al [7] positively, 234 (22.2%) answered 2 of 3 positively, 138 (13.09%) answered only 1 item positively, 72 (6.83%) patients answered “not applicable” or responded negatively, and 36 (3.42%) responses were missing. Patient and visit characteristics of patients who positively answered all 3 crucial Hanna [7] items versus the group who did not are shown in Multimedia Appendix 3.

Results of the binary logistic regression are shown in Table 4. Negative association was found between the positive evaluation of a visit and the surgical and “other” medical specialties (B=−0.64; $P<.001$). Positive evaluation was also associated with the patient category who described their health as “very well” (B=1.12; $P=.01$). Sex, age, and educational status had no influence on whether a patient rated the visit positively. In addition, the reason for the visit was not found to have any influence on the positive evaluation of a video visit.

XSL•FO
RenderX

Table 1. Demographic characteristics.

Characteristics	Patients (n=1054), n (%)
Sex	
Male	448 (42.5)
Female	606 (57.49)
Age (y)	
18-34	131 (12.43)
35-54	333 (31.59)
55-64	272 (25.81)
65-79	304 (28.84)
80-99	14 (1.33)
Educational level	
None	5 (0.47)
Primary	105 (9.96)
Secondary	378 (35.86)
Higher and university	468 (44.4)
Different	29 (2.75)
Missing	69 (6.55)
Self-rated health	
Excellent	95 (9.01)
Very well	167 (15.84)
Good	466 (44.21)
Not so well	255 (24.19)
Bad	49 (4.65)
Missing	22 (2.08)
Specialty	
Nonsurgical medical specialisms	498 (47.24)
Cardiology	33 (3.13)
Dermatology	3 (0.28)
Gastroenterology	16 (1.52)
Hematology	40 (3.79)
Internal medicine	66 (6.26)
Medical oncology	121 (11.48)
Nephrology	14 (1.33)
Neurology	153 (14.52)
Pulmonary medicine	25 (2.37)
Reproductive medicine	25 (2.37)
Rheumatology	2 (0.18)
Surgical specialties	101 (9.58)
ENT[a]	6 (0.57)
General surgery[b]	36 (3.41)
Gynecology	9 (0.85)
Neurosurgery	13 (1.23)

XSL•FO
RenderX

Characteristics	Patients (n=1054), n (%)
Obstetrics	1 (0.09)
OMF[c] surgery	1 (0.09)
Orthopedic surgery	3 (0.28)
Plastic surgery	8 (0.76)
Urology	24 (2.27)
Other medical specialties	455 (43.17)
Clinical genetics	420 (39.85)
Psychiatry	29 (2.75)
Radiology	2 (0.18)
Radiotherapy	4 (0.38)
Type of visit	
First visit—referred by primary care	195 (18.5)
First visit—different health care professional	270 (25.62)
First visit—second opinion	15 (1.42)
Follow-up—known health care professional	480 (45.54)
Follow-up—new health care professional	71 (6.74)
Missing	23 (2.18)
Follow-up after video consultation	
Follow-up visit at the outpatient clinic	215 (20.39)
Follow-up video visit	425 (40.32)
Additional medical examination	174 (16.51)
Referred to the referring physician	41 (3.89)
There is no follow-up needed	126 (11.95)
I do not know	61 (5.79)
Missing	12 (1.14)

[a]ENT: ear, nose, and throat.

[b]Includes trauma surgery, visceral surgery, surgical oncology, and vascular surgery.

[c]OMF: oral and maxillofacial surgery.

XSL•FO
RenderX

Table 2. Responses of patients to the Patient Experience Monitor (PEM) items[a].

PEM item and response options	Responses (n=1054), n (%)
Did you have to wait for the video visit?	
The professional was on time	719 (68.22)
I waited <5 min	143 (13.57)
I waited >5 min	62 (5.88)
I waited >10 min	42 (3.98)
I waited >15 min	63 (5.97)
I don't know	22 (2.08)
Did the professional read your health record?	
Yes	991 (94.02)
Not thoroughly	25 (2.37)
No	2 (0.19)
I don't know	21 (1.99)
Did you get answers you could understand?	
Yes	985 (93.45)
Most answers were understandable	26 (2.46)
No	2 (0.19)
No opportunity to ask questions	0 (0)
I didn't have any questions	26 (2.47)
Did you trust the professional?	
Yes	1021 (96.87)
Less than I wanted	17 (1.61)
No	4 (0.38)
Was there enough time to talk to the professional about your disease or problem?	
Yes	1015 (96.29)
I needed more time	14 (1.33)
No	1 (0.09)
I don't know	3 (0.28)
Could you participate in the decision making process?	
Yes	732 (69.45)
Less possible than I wanted	18 (1.71)
No	10 (0.95)
No need	272 (25.81)
I don't know	7 (0.66)
Could a relative participate in the conversation?	
Yes	499 (47.34)
More possible than I wanted	8 (0.76)
Less possible than I wanted	4 (0.38)
No	8 (0.76)
There were no relatives present	241 (22.87)
No need	277 (26.28)
I don't know	4 (0.38)
Did the professional explain the advantages and disadvantages of treatment or diagnostic process?	

XSL•FO
RenderX

PEM item and response options	Responses (n=1054), n (%)
Yes	743 (70.49)
Some aspects weren't clearly explained	20 (1.89)
No	4 (0.38)
There was no explanation	2 (0.19)
No need	268 (25.43)
Did the professional explain the adverse effects of the new medication?	
Yes	170 (16.13)
There wasn't enough explanation	23 (2.18)
No	11 (1.04)
I didn't receive new medication	829 (78.65)
Did the professional tell you what to do after the visit?	
Yes	784 (74.38)
Not enough	14 (1.33)
No	7 (0.66)
No need	228 (21.63)
I don't know	9 (0.85)
Could you speak to anyone of the hospital about your worries and anxieties?	
Yes	235 (22.29)
Sometimes	24 (2.28)
No	25 (2.37)
No need	758 (71.92)

[a]The total in some sections is not 100% owing to missing responses.

XSL•FO
RenderX

Table 3. Responses of patients for the items on the survey by Hanna et al [7].

Items of the survey	Responses (n=1054), n (%)						
	Yes, definitely	Yes, I think so	Maybe yes, maybe no	I don't think so	No, definitely not	Not applicable	Missing
I am satisfied with my video visit.	671 (63.66)	293 (27.79)	46 (4.36)	17 (1.61)	10 (0.95)	0 (0)	17 (1.61)
I worried about my privacy.	13 (1.23)	18 (1.71)	97 (9.20)	366 (34.72)	549 (52.09)	0 (0)	11 (1.04)
The care I received by video visit was just as good as with an in-person appointment.	321 (30.45)	385 (36.53)	162 (15.37)	102 (9.68)	23 (2.18)	44 (4.17)	17 (1.61)
It is important to me that the video visit saved me travel time.	428 (40.61)	280 (26.56)	191 (18.12)	107 (10.15)	32 (3.04)	0 (0)	16 (1.52)
It is important to me that the video visit saved me money.	169 (16.03)	179 (16.98)	276 (26.19)	140 (13.28)	106 (10.06)	165 (15.65)	19 (1.8)
I experienced technical problems during the video visit which caused a disturbance of the visit.	33 (3.13)	74 (7.02)	92 (8.73)	265 (25.14)	575 (54.55)	0 (0)	15 (1.42)
I was comfortable talking by video to the health care professional.	580 (55.03)	359 (34.06)	71 (6.74)	25 (2.37)	8 (0.76)	0 (0)	11 (1.04)
I would rather travel to have my next visit in-person than use video visit.	100 (9.48)	142 (13.47)	272 (25.81)	257 (24.38)	211 (20.02)	56 (5.31)	16 (1.52)
I was able to develop a friendly relationship with my health care professional.	305 (28.94)	477 (45.25)	209 (19.83)	43 (4.08)	4 (0.38)	0 (0)	16 (1.52)
I was able to explain my problems clearly to my healthcare professional during the video visit.	482 (45.73)	472 (44.78)	65 (6.17)	11 (1.04)	3 (0.28)	0 (0)	21 (1.99)
The video visit was convenient.	389 (36.91)	474 (44.97)	141 (13.38)	32 (3.04)	8 (0.76)	0 (0)	10 (0.95)
I would recommend the video visit option to other patients.	329 (31.21)	395 (37.47)	254 (24.09)	42 (3.98)	12 (1.14)	0 (0)	22 (2.09)

XSL•FO
RenderX

Table 4. Binary logistic regression model results[a].

Variable	B (95% CI)	*P* value
Sex (male)	0.03 (0.78-1.37)	.83
Age category (y)		
18-34	Reference	N/A[b]
35-54	0.17 (0.76-1.86)	.45
55-64	0.14 (0.72-1.85)	.55
65-79	0.06 (0.65-1.73)	.81
80-99	−0.17 (0.20-3.54)	.82
Educational level		
None	Reference	N/A
Primary	0.05 (0.16-7.07)	.96
Secondary	0.05 (0.16-6.89)	.95
Higher and university	−0.04 (0.15-6.22)	.96
Different	−0.16 (0.11-6.43)	.88
Self-rated health		
Excellent	Reference	N/A
Very well	1.12 (1.39-6.72)	.01
Good	0.52 (0.85-3.35)	.13
Not so well	0.14 (0.61-2.15)	.66
Bad	0.27 (0.69-2.50)	.41
Specialty		
Nonsurgical	Reference	N/A
Surgical	−0.64 (0.37-0.75)	<.001
“Other”	−0.45 (0.38-1.06)	.08
Type of visit		
First visit	Reference	N/A
Follow-up visit	0.17 (0.86-1.65)	.30

[a] Nagelkerke R^2=0.08.

[b]N/A: not applicable.

Qualitative Results

Overview

Table 5 displays the frequencies of appreciated aspects and aspects for improvement. Most patients (738/1054, 70.02%) cited appreciated aspects of the use of video visits in the open-ended questions. The most frequently cited appreciated aspects were expressed at the individual professional level, followed by the organizational context level and the innovation level. Few aspects were mentioned at the patient level, economic and political context level, and social context level. Approximately half of the patients (474/1054, 44.97%) cited aspects for improvement. Most were cited at the innovation, organizational context, and patient levels. In contrast, no improvable aspects were reported at the economic and political context level.

XSL•FO
RenderX

Table 5. Aspects that were appreciated and aspects for improvement obtained from patients' experiences with video visits, extracted from answers to open-ended questions.

Level	Citations of aspects as appreciated[a] (n=926)[b], n (%)	Citations of aspects for improvement[c] (n=388)[b], n (%)
Innovation	185 (19.9)	138 (35.6)
Individual professional	399 (43.1)	46 (11.9)
Patient	82 (8.9)	91 (23.5)
Social context	11 (1.2)	17 (4.4)
Organizational context	204 (22)	96 (24.7)
Economic and political context	45 (4.9)	0 (0)

[a]Overall, 70.01% (738/1054) of the patients cited appreciated aspects.

[b]Responses including multiple levels within a single response were counted as individual items.

[c]Overall, 44.97% (474/1054) of the patients cited aspects for improvement.

Innovation

The perceived ease of use and audio-visual quality were frequently mentioned as appreciated aspects. The intuitive character of the app was seen as valuable, as not all patients were familiar with using web-based apps. Patients appreciated the audio-visual quality, allowing the video visit to be a good alternative for an in-person visit. However, not all the patients experienced the same ease of use, as the most reported aspect for improvement was poor audio and video quality, sometimes clearly caused by an unstable internet connection. Although Zaurus is compatible with all electronic devices, users have reported issues with video size specifically on smartphones (Textbox 1).

Textbox 1. Responses related to innovation.

Appreciated aspect

"The application is straightforward and easy to understand. Conversation went well, the doctor even asked me if I could hear her well." [Female; aged 18-34 y; clinical genetics]

Aspect for improvement

"The video connection was really bad. Almost immediately the app crashed, and the audio stuttered, so I could not understand what the doctor was saying. After two attempts, we continued the visit by telephone." [Female; aged 65-79 y; excluded for further analysis; clinical genetics]

Individual Professional

Patients often mentioned what they valued in the clinician's professional behavior, such as their attitude, and communicative style. In contrast, a lack of adequate or visible body language and lack of knowledge about someone's medical history were mentioned as aspects for improvement (Textbox 2).

Textbox 2. Responses related to the individual professional.

Appreciated aspect

"There is still a kind of personal touch in the contact, which is nice for the perception as well. The doctor radiated tranquility and was understanding, and she had read my personal record well. That gives me confidence." [Female; aged 55-64 y; clinical genetics]

Aspect for improvement

"The doctor did not look at us during the video visit. Both my daughter, who was also present, and I had noticed. That felt a little awkward. He was mainly looking down (I guess at a file or something like that, which was in front of him)." [Female; aged 80-99 y; clinical genetics]

Patient

The possibility to have face-to-face interactions remotely was often mentioned as valuable, as patients were able to watch the clinicians' nonverbal reactions. It made video visits a safe alternative for patients with a weak immune system, for example, during the pandemic. Personal lack of experience with video visits was a hindering factor, as not all patients were familiar with the use of videoconferencing apps. Some of them preferred an in-person visit, as they felt uncomfortable owing to inexperience (Textbox 3).

XSL•FO
RenderX

Textbox 3. Responses related to the patient.

Appreciated aspect

"It is nice to see the doctor, but for a first meeting, it is something I need to get used to. However, this feels safer regarding the coronavirus and a vulnerable immune status." [Female; aged 35-54 y; neurology]

Aspect for improvement

"I'd rather have the first visit in person. Maybe I'm old fashioned, but I prefer physical contact, even during this COVID pandemic. Feelings and emotions might be more difficult to pick up on screen." [Male; aged 55-64 y; neurology]

Social Context

The possibility of the involvement of others, such as next of kin or other family members, was one of the mentioned appreciated aspects. Creating a culture in which a patient can share their preference or opt for a certain visit modality could stimulate the use of video visits. Patients expressed that they would like to have a say in choosing which visit modality they like, especially when the nature of the visit is sensitive (Textbox 4).

Textbox 4. Responses related to the social context.

Appreciated aspect

"On time, pleasant conversation, space for questions, clear explanation. It was nice that my partner could join with his phone." [Female; aged 18-34 y; reproductive medicine]

Aspect for improvement

"It was a shame they communicated the results by a video visit. I was shocked and found they acted a bit indignant about my reaction. I was not capable anymore to follow the conversation." [Female; aged 18-34 y; clinical genetics]

Organizational Context

Internet-based assistance, clear instructions, and time management by the clinician during the visit were often mentioned as appreciated aspects at this level. Many patients were called in advance of the visit to check for technical problems. However, patients were not always informed correctly if the visit would start later than scheduled, and in some cases, patients received the link for the video visit just before the visit started, which was an aspect for improvement (Textbox 5).

Textbox 5. Responses related to the organizational context.

Appreciated aspect

"The support was really good, as I am not so technical and there was enough explanation. Great." [Female; aged 55-64 y; clinical genetics]

Aspect for improvement

"I would like to receive a notification when the doctor is held up, especially when it's a first visit. Also, I would like to receive a heads up when I get another doctor than the one the appointment was originally scheduled with." [Male; aged 55-64 y; medical oncology]

Economic and Political Context

Time and financial savings were identified as valuable aspects at the economic and political context level. Patients often cited less travel time and costs as beneficial. There were no improvable aspects reported at the economic and political context level (Textbox 6).

Textbox 6. Responses related to the economic and political context.

Appreciated aspect

"It is pleasant that there is no need for traveling to the hospital (regarding travel time and travel distance) and still have 'personal' contact with the doctor through a video connection." [Male; aged 35-54 y; neurology]

Discussion

Principal Findings

In this study, we comprehensively analyzed evaluations of visits via video to a tertiary clinic made by a large, diverse patient population, including appreciated aspects and aspects for improvement. Most patients (964/1054, 91.46%) evaluated the video visits positively, with significantly more positive evaluations when the visits were provided by a clinician from "other" medical specialties, as compared to surgical and nonsurgical specialties, or when the patient rated their health status as "very well." The appreciated aspects were mostly at the individual professional level, whereas aspects for improvement were reported at the innovation level itself.

Comparison With Previous Studies

Our finding that high self-rated health of patients is an influencing patient characteristic for suitability of video visits echoes the finding that patients with less complex, more straightforward clinical needs are more suitable candidates for video visits than those with complex, high-risk diseases [10,29].

XSL•FO
RenderX

Similar findings were identified in an oncological study wherein telemedicine was received favorably for low-acuity cancer care [16]. In contrast to findings that video visits appear to be more appropriate when the clinician knows the patient beforehand and when it is a follow-up visit [10], we found that the reason for the visit did not have any influence on whether patients rated the video visit positively. Remarkably, the medical specialty providing the visit was found to be a significant associated factor. Nonsurgical visits were found to be most suitable for telemedicine. One can imagine that these visits are less dependent on physical examination, such as internal medicine or dermatology, for instance, as these specialties can easily review laboratory abnormalities or skin disorders on screen [30]. In addition to specific aspects of the visit that may depend on the medical specialty, other dimensions or elements during the visit could affect the patient experience, such as the communication strategy used by the clinician [31]. Nonetheless, more studies are needed to get a clear overview about whether medical specialty is a truly discriminative characteristic or whether it is more dependent on the attitude of certain clinicians and patient groups.

Qualitative analysis of the open-ended questions revealed both facilitating and hindering factors for broad implementation and upscaling of video visits. Following Grol and Wensing [28], these factors were categorized into innovation, professional, patient, social context, organizational context, and economic and political context levels. Appreciation was mostly centered on the professionals' skill in adapting communication to the video setup; however, there was scope for improvement among some individuals, as they might benefit from investing additional effort in making visual contact. The way in which the video visits were organized was also highly valued, especially for the provision of technical support to patients as and when needed.

The attitude of the clinician during the video visit was one of the most frequently mentioned aspects for improvement. The bedside manner, which may be better described as the "webside" manner, of a clinician refers to how the clinician behaves, approaches the patient, and communicates during the visit. Clinicians sometimes seem to lack awareness of how their nonverbal behavior looks on screen, as was also shown by a study that analyzed a large data set of >5000 patients [32]. Patients prefer increased expression of nonverbal empathy from clinicians when they show signs of distress. Inadequate nonverbal communication and body language are often reported as barriers for telemedicine adoption [3,33,34]. Despite expert recommendations dating back several years to raise awareness for nonverbal and paraverbal communication, our study also indicates that there is still considerable scope for improvement and training at the clinician level [31,35].

In accordance with several survey studies conducted during the pandemic, our response rate was low. This might be explained by the questionnaire fatigue that was frequently observed during the COVID-19 pandemic, as patients received multiple questionnaires and messages from the outpatient clinic, apart from research [36].

Regarding future perspectives, clinicians should seek guidelines to assess the suitability of a video visit, and the following recommendations might be useful. The Dutch Center of Expertise on Health Disparities recommends checking the patients' digital skills beforehand, providing digital support, and evaluating whether the information is correctly understood through techniques such as "teach back" at the end of the visit [37]. Video consulting guidelines advise considering several factors while deciding whether video visits may be suitable, such as whether there is an established relationship with the patient, whether it entails nonurgent care, whether there is a need for physical examination, and whether there are factors in favor of the patient staying at home [38].

Limitations

The findings of this study must be considered in the light of some limitations. First, our study might have been exposed to selection and sampling biases for several reasons. It was an "open" survey, where patients could decide voluntarily whether they would participate in the survey, which might have led to a sample of patients that is not representative of the entire population of the hospital. In addition, owing to the exclusion criteria, not all video visits were evaluated.

Second, the validity of the combined survey was not tested. As the analysis of our data was reported at the item level, calculating the internal consistency using Cronbach α was not applicable. PEM is known to be a validated survey; however, the psychometric properties of the survey by Hanna et al [7] are not known and should be determined. As the PEM items were rephrased to the videoconferencing setting, reliability of this new PEM survey will have to be reassessed.

The educational status of patients attending a university medical center is, in general, often higher than the mean educational status of the general population. In our study population, 44.4% (468/1054) of the patients were highly educated, compared with 30% in the Dutch population in 2018 [39]. As teaching hospitals and referring hospitals might serve different populations, the generalizability of our results might be limited, and further studies including different types of hospitals are recommended.

A total of 139 patients reported a failed video visit and noted that the visit was replaced by telephone. However, the exact number of times this occurred is not known, as not all patients might have reported this failure, which also may have resulted in selection bias.

In this evaluation, we deliberately focused on the patient evaluation of video visits. As it is known that patients and clinicians have different views about quality of information and visits [40-42], the clinicians' point of view should also be explored for a comprehensive evaluation of the use of video visits. Health care providers, such as clinicians, might experience different barriers and facilitators compared with patients, thus influencing the successful implementation of video visits. Key barriers to successful implementation such as the lack of training and motivation to offer video visits need to be addressed [3,8,43,44].

Conclusions

Video visits were perceived as highly satisfactory by patients during the COVID-19 pandemic, with the best experiences

XSL•FO
RenderX

reported by healthy participants and participants who scheduled a visit with a clinician outside the realms of surgical and nonsurgical medical specialties, such as clinical genetics or radiotherapy. Appreciated aspects were mainly at the individual professional level, organizational level, and at the level of the innovation itself. The mentioned aspects for improvement can be changed for the better.

The findings cannot be directly generalized as they were collected in a university medical center with a specific patient population, but they provide additional results for understanding the suitability of video visits in a broad patient population. To be able to truly tailor the use of video visits to patients' needs, a patient-centered perspective involving both patients and health care professionals is needed.

Data Availability

The data used in this study will be made available by the authors upon reasonable request.

Authors' Contributions

SCvdB, MM, MS, and SvD contributed to the study's conception and design. SCvdB and DvD performed the statistical analyses. SCvdB managed the project administration and data curation. SCvdB, DvD, MS, and SvD drafted the manuscript. HvG, SB, MS, and SvD supervised the research project. SCvdB and DvD contributed equally to this work and share first authorship. MS and SvD contributed equally to this work and share senior authorship. All authors interpreted the data, critically revised the paper, and approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey on video visits during the COVID-19 outbreak (August 2020 to December 2020).
[DOCX File , 18 KB - jmir_v26i1e49058_app1.docx]

Multimedia Appendix 2
CHERRIES (Checklist for Reporting Results of Internet-Based e-Surveys) checklist.
[DOCX File , 17 KB - jmir_v26i1e49058_app2.docx]

Multimedia Appendix 3
Patient and visit characteristics of those who positively answered all 3 crucial items in the checklist by Hanna et al [7] versus the group that did not answer positively.
[DOCX File , 17 KB - jmir_v26i1e49058_app3.docx]

References

1. Whitehead E, Dorfman V, Tremper G, Kramer A, Sigler A, Gosman A. Telemedicine as a means of effective speech evaluation for patients with cleft palate. Ann Plast Surg 2012 Apr;68(4):415-417 [FREE Full text] [doi: 10.1097/SAP.0b013e31823b6897] [Medline: 22421491]
2. Hilgart JS, Hayward JA, Coles B, Iredale R. Telegenetics: a systematic review of telemedicine in genetics services. Genet Med 2012 Sep;14(9):765-776 [FREE Full text] [doi: 10.1038/gim.2012.40] [Medline: 22498847]
3. Westra I, Niessen FB. Implementing real-time video consultation in plastic surgery. Aesthetic Plast Surg 2015 Oct;39(5):783-790. [doi: 10.1007/s00266-015-0526-4] [Medline: 26169952]
4. Stommel WJ, van Goor H, Stommel MW. The impact of video-mediated communication on closed wound assessments in postoperative consultations: conversation analytical study. J Med Internet Res 2020 May 05;22(5):e17791 [FREE Full text] [doi: 10.2196/17791] [Medline: 32310816]
5. Aneja J, Goyal T, Verma M, Kaur G, Mirza M, Gupta S. Client satisfaction with telemedicine services during COVID-19 pandemic: a cross-sectional survey from a teaching institute of North India. J Family Med Prim Care 2022 Sep;11(9):5187-5193 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_2217_21] [Medline: 36505639]
6. van Gurp J, van Selm M, Vissers K, van Leeuwen E, Hasselaar J. How outpatient palliative care teleconsultation facilitates empathic patient-professional relationships: a qualitative study. PLoS One 2015;10(4):e0124387 [FREE Full text] [doi: 10.1371/journal.pone.0124387] [Medline: 25902263]
7. Hanna GM, Fishman I, Edwards DA, Shen S, Kram C, Liu X, et al. Development and patient satisfaction of a new telemedicine service for pain management at Massachusetts general hospital to the island of Martha's vineyard. Pain Med 2016 Sep;17(9):1658-1663 [FREE Full text] [doi: 10.1093/pm/pnw069] [Medline: 27121891]

XSL•FO
RenderX

8. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
9. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
10. Greenhalgh T, Shaw S, Wherton J, Vijayaraghavan S, Morris J, Bhattacharya S, et al. Real-world implementation of video outpatient consultations at macro, meso, and micro levels: mixed-method study. J Med Internet Res 2018 Apr 17;20(4):e150 [FREE Full text] [doi: 10.2196/jmir.9897] [Medline: 29625956]
11. Ossebaard HC, Van Gemert-Pijnen L. eHealth and quality in health care: implementation time. Int J Qual Health Care 2016 Jun;28(3):415-419. [doi: 10.1093/intqhc/mzw032] [Medline: 27029590]
12. Bradford NK, Armfield NR, Young J, Smith AC. Paediatric palliative care by video consultation at home: a cost minimisation analysis. BMC Health Serv Res 2014 Jul 28;14:328 [FREE Full text] [doi: 10.1186/1472-6963-14-328] [Medline: 25069399]
13. Versleijen M, Martin-Khan MG, Whitty JA, Smith AC, Gray LC. A telegeriatric service in a small rural hospital: a case study and cost analysis. J Telemed Telecare 2015 Dec;21(8):459-468. [doi: 10.1177/1357633X15611327] [Medline: 26556059]
14. Andino J, Zhu A, Chopra Z, Daignault-Newton S, Ellimoottil C, Dupree JM. Video visits are practical for the follow-up and management of established male infertility patients. Urology 2021 Aug;154:158-163 [FREE Full text] [doi: 10.1016/j.urology.2021.03.050] [Medline: 34022261]
15. Assing Hvidt E, Atherton H, Keuper J, Kristiansen E, Lüchau EC, Lønnebakke Norberg B, et al. Low adoption of video consultations in post-COVID-19 general practice in Northern Europe: barriers to use and potential action points. J Med Internet Res 2023 May 22;25:e47173. [doi: 10.2196/47173] [Medline: 37213196]
16. Mackwood M, Butcher R, Vaclavik D, Alford-Teaster JA, Curtis KM, Lowry M, et al. Adoption of telemedicine in a rural US cancer center amid the COVID-19 pandemic: qualitative study. JMIR Cancer 2022 Aug 16;8(3):e33768 [FREE Full text] [doi: 10.2196/33768] [Medline: 35895904]
17. Communication on enabling the digital transformation of health and care in the digital single market; empowering citizens and building a healthier society. European Commission. 2018. URL: https://digital-strategy.ec.europa.eu/en/library/communication-enabling-digital-transformation-health-and-care-digital-single-market-empowering [accessed 2024-01-29]
18. Wherton J, Greenhalgh T, Hughes G, Shaw SE. The role of information infrastructures in scaling up video consultations during COVID-19: mixed methods case study into opportunity, disruption, and exposure. J Med Internet Res 2022 Nov 10;24(11):e42431 [FREE Full text] [doi: 10.2196/42431] [Medline: 36282978]
19. Esber A, Teufel M, Jahre L, In der Schmitten J, Skoda EM, Bäuerle A. Predictors of patients' acceptance of video consultation in general practice during the coronavirus disease 2019 pandemic applying the unified theory of acceptance and use of technology model. Digit Health 2023;9:20552076221149317 [FREE Full text] [doi: 10.1177/20552076221149317] [Medline: 36815005]
20. Rasmussen B, Perry R, Hickey M, Hua X, Wong ZW, Guy L, et al. Patient preferences using telehealth during the COVID-19 pandemic in four Victorian tertiary hospital services. Intern Med J 2022 May;52(5):763-769 [FREE Full text] [doi: 10.1111/imj.15726] [Medline: 35194927]
21. Splinter MJ, Ikram MK, Helsper CW, Bindels PJ, de Schepper EI, Licher S. Patient perspectives on telemedicine during the COVID-19 pandemic: a mixed-methods community-based study. BMC Health Serv Res 2023 Jul 27;23(1):803 [FREE Full text] [doi: 10.1186/s12913-023-09794-w] [Medline: 37501087]
22. Farrer LM, Batterham PJ, Gulliver A, Morse A, Calear AL, McCallum S, et al. The factors associated with telehealth use and avoidance during the COVID-19 pandemic: longitudinal survey. J Med Internet Res 2023 Feb 08;25:e43798 [FREE Full text] [doi: 10.2196/43798] [Medline: 36649254]
23. Nieuwe patiëntervaringsmonitor helpt zorg te verbeteren. Netherlands Federation of University Medical Centres. URL: https://www.nfu.nl/themas/kwaliteit-van-zorg/patientervaringen-en-uitkomsten-van-zorg [accessed 2024-01-29]
24. Principles of person-centered care. Picker Institute Europe. URL: https://picker.org/who-we-are/the-picker-principles-of-person-centred-care/ [accessed 2024-01-29]
25. Bastemeijer CM, Boosman H, Zandbelt L, Timman R, de Boer D, Hazelzet JA. Patient experience monitor (PEM): the development of new short-form picker experience questionnaires for hospital patients with a wide range of literacy levels. Patient Relat Outcome Meas 2020;11:221-230 [FREE Full text] [doi: 10.2147/PROM.S274015] [Medline: 33312007]
26. Demiris G. Principles of survey development for telemedicine applications. J Telemed Telecare 2006;12(3):111-115. [doi: 10.1258/135763306776738549] [Medline: 16638231]
27. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet E-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
28. Grol R, Wensing M. What drives change? Barriers to and incentives for achieving evidence-based practice. Med J Aust 2004 Mar 15;180(S6):S57-S60. [doi: 10.5694/j.1326-5377.2004.tb05948.x] [Medline: 15012583]

XSL•FO
RenderX

29. James HM, Papoutsi C, Wherton J, Greenhalgh T, Shaw SE. Spread, scale-up, and sustainability of video consulting in health care: systematic review and synthesis guided by the NASSS framework. J Med Internet Res 2021 Jan 26;23(1):e23775 [FREE Full text] [doi: 10.2196/23775] [Medline: 33434141]
30. Jaklitsch E, Shah VK, Smith B, Agarwal A, Chen J, Sweeney A, et al. Melanoma detected through teledermatology versus in-person visits. Telemed J E Health (Forthcoming) 2023 Oct 17. [doi: 10.1089/tmj.2023.0364] [Medline: 37937942]
31. Mazouri-Karker S, Braillard O, Lüchinger R, Bajwa N, Achab S, Hudelson P, et al. Patients preferences for communication during video consultations. Patient Educ Couns 2023 Oct;115:107894 [FREE Full text] [doi: 10.1016/j.pec.2023.107894] [Medline: 37480793]
32. Chandrasekaran R, Bapat P, Venkata PJ, Moustakas E. Face time with physicians: how do patients assess providers in video-visits? Heliyon 2023 Jun;9(6):e16883 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e16883] [Medline: 37292342]
33. Demiris G, Speedie SM, Finkelstein S. Change of patients' perceptions of TeleHomeCare. Telemed J E Health 2001;7(3):241-248. [doi: 10.1089/153056201316970948] [Medline: 11564360]
34. Armfield N, Bradford N, White MM, Spitzer P, Smith AC. Humour sans frontieres: the feasibility of providing clown care at a distance. Telemed J E Health 2011 May;17(4):316-318 [FREE Full text] [doi: 10.1089/tmj.2010.0166] [Medline: 21457119]
35. Jiménez-Rodríguez D, Ruiz-Salvador D, Rodríguez Salvador MD, Pérez-Heredia M, Muñoz Ronda FJ, Arrogante O. Consensus on criteria for good practices in video consultation: a Delphi study. Int J Environ Res Public Health 2020 Jul 27;17(15):5396 [FREE Full text] [doi: 10.3390/ijerph17155396] [Medline: 32727042]
36. de Koning R, Egiz A, Kotecha J, Ciuculete AC, Ooi SZ, Bankole ND, et al. Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. Front Surg 2021 Aug 12;8:690680 [FREE Full text] [doi: 10.3389/fsurg.2021.690680] [Medline: 34458314]
37. Checklist beeldbellen. Pharos. URL: https://www.pharos.nl/kennisbank/stappenplan-laagdrempelig-toegankelijk-beeldbellen/ [accessed 2024-01-29]
38. Gilbert AW, Billany JC, Adam R, Martin L, Tobin R, Bagdai S, et al. Rapid implementation of virtual clinics due to COVID-19: report and early evaluation of a quality improvement initiative. BMJ Open Qual 2020 May;9(2):e000985 [FREE Full text] [doi: 10.1136/bmjoq-2020-000985] [Medline: 32439740]
39. Imlach F, McKinlay E, Middleton L, Kennedy J, Pledger M, Russell L, et al. Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. BMC Fam Pract 2020 Dec 13;21(1):269 [FREE Full text] [doi: 10.1186/s12875-020-01336-1] [Medline: 33308161]
40. Barsom EZ, Meijer HA, Blom J, Schuuring MJ, Bemelman WA, Schijven MP. Emergency upscaling of video consultation during the COVID-19 pandemic: contrasting user experience with data insights from the electronic health record in a large academic hospital. Int J Med Inform 2021 Jun;150:104463 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104463] [Medline: 33872824]
41. Kim J, Kim S. Physicians' perception of the effects of Internet health information on the doctor-patient relationship. Inform Health Soc Care 2009 Sep;34(3):136-148. [doi: 10.1080/17538150903102422] [Medline: 19670004]
42. Murray E, Lo B, Pollack L, Donelan K, Catania J, Lee K, et al. The impact of health information on the internet on health care and the physician-patient relationship: national U.S. survey among 1.050 U.S. physicians. J Med Internet Res 2003;5(3):e17 [FREE Full text] [doi: 10.2196/jmir.5.3.e17] [Medline: 14517108]
43. Smith WR, Atala AJ, Terlecki RP, Kelly EE, Matthews CA. Implementation guide for rapid integration of an outpatient telemedicine program during the COVID-19 pandemic. J Am Coll Surg 2020 Aug;231(2):216-2.e2 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2020.04.030] [Medline: 32360960]
44. Jiménez-Rodríguez D, Santillán García A, Montoro Robles J, Rodríguez Salvador MD, Muñoz Ronda FJ, Arrogante O. Increase in video consultations during the COVID-19 pandemic: healthcare professionals' perceptions about their implementation and adequate management. Int J Environ Res Public Health 2020 Jul 15;17(14):5112 [FREE Full text] [doi: 10.3390/ijerph17145112] [Medline: 32679848]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet-Based e-Surveys
PEM: Patient Experience Monitor

XSL•FO
RenderX

Edited by T Leung; submitted 16.05.23; peer-reviewed by AW Zahoor, J Hayden; comments to author 25.10.23; revised version received 08.12.23; accepted 31.01.24; published 27.03.24.

Please cite as:
van den Bosch SC, van Dalen D, Meinders M, van Goor H, Bergé S, Stommel M, van Dulmen S
Outpatient Video Visits During the COVID-19 Pandemic: Cross-Sectional Survey Study of Patients' Experiences and Characteristics
J Med Internet Res 2024;26:e49058
URL: https://www.jmir.org/2024/1/e49058
doi:10.2196/49058
PMID:38536236

©Stefanie C van den Bosch, Demi van Dalen, Marjan Meinders, Harry van Goor, Stefaan Bergé, Martijn Stommel, Sandra van Dulmen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

#ProtectOurElders: Analysis of Tweets About Older Asian Americans and Anti-Asian Sentiments During the COVID-19 Pandemic

Reuben Ng[1,2], PhD; Nicole Indran[1], BSocSci (Hons)

[1]Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore, Singapore

[2]Lloyd's Register Foundation Institute for the Public Understanding of Risk, National University of Singapore, Singapore, Singapore

Corresponding Author:
Reuben Ng, PhD
Lee Kuan Yew School of Public Policy
National University of Singapore
469C Bukit Timah Road
Singapore, 259772
Singapore
Phone: 65 66013967
Email: spprng@nus.edu.sg

Abstract

Background: A silver lining to the COVID-19 pandemic is that it cast a spotlight on a long-underserved group. The barrage of attacks against older Asian Americans during the crisis galvanized society into assisting them in various ways. On Twitter, now known as X, support for them coalesced around the hashtag #ProtectOurElders. To date, discourse surrounding older Asian Americans has escaped the attention of gerontologists—a gap we seek to fill. Our study serves as a reflection of the level of support that has been extended to older Asian Americans, even as it provides timely insights that will ultimately advance equity for them.

Objective: This study explores the kinds of discourse surrounding older Asian Americans during the COVID-19 crisis, specifically in relation to the surge in anti-Asian sentiments. The following questions guide this study: What types of discourse have emerged in relation to older adults in the Asian American community and the need to support them? How do age and race interact to shape these discourses? What are the implications of these discourses for older Asian Americans?

Methods: We retrieved tweets (N=6099) through 2 search queries. For the first query, we collated tweets with the hashtag #ProtectOurElders. For the second query, we collected tweets with an age-based term, for example, “elderly” or “old(er) adults(s)” and either the hashtag #StopAAPIHate or #StopAsianHate. Tweets were posted from January 1, 2020, to August 1, 2023. After applying the exclusion criteria, the final data set contained 994 tweets. Inductive and deductive approaches informed our qualitative content analysis.

Results: A total of 4 themes emerged, with 50.1% (498/994) of posts framing older Asian Americans as “vulnerable and in need of protection” (theme 1). Tweets in this theme either singled them out as a group in need of protection because of their vulnerable status or discussed initiatives aimed at safeguarding their well-being. Posts in theme 2 (309/994, 31%) positioned them as “heroic and resilient.” Relevant tweets celebrated older Asian Americans for displaying tremendous strength in the face of attack or described them as individuals not to be trifled with. Tweets in theme 3 (102/994, 10.2%) depicted them as “immigrants who have made selfless contributions and sacrifices.” Posts in this section referenced the immense sacrifices made by older Asian Americans as they migrated to the United States, as well as the systemic barriers they had to overcome. Posts in theme 4 (85/994, 8.5%) venerated older Asian Americans as “worthy of honor.”

Conclusions: The COVID-19 crisis had the unintended effect of garnering greater support for older Asian Americans. It is consequential that support be extended to this group not so much by virtue of their perceived vulnerability but more so in view of their boundless contributions and sacrifices.

(J Med Internet Res 2024;26:e45864) doi:10.2196/45864

XSL•FO
RenderX

KEYWORDS

AAPI; anti-Asian hate; anti-Asian; Asian Americans and Pacific Islanders; Asian-American; content analysis; coronavirus; COVID-19; discourse; discriminate; discrimination; discriminatory; Pacific Islander; racial; racism; racist; SARS-CoV-2; social media; tweet; Twitter

Introduction

Not unlike other public health crises, the COVID-19 pandemic brought with it a disconcerting onslaught of racism and xenophobia [1]. The number of anti-Asian hate crimes in the United States quadrupled in 2021, escalating from the already significant uptick it experienced in 2020, when the COVID-19 outbreak was declared a global pandemic [2]. In the Asian American and Pacific Islanders (AAPI) community, those aged 60 years or older accounted for 7.3% of the 2808 self-reported incidents in 2020 [3]. Though not a particularly large figure, underreporting in this community is fairly common [4]. Moreover, older adults have reported being physically assaulted and having to deal with civil rights violations more than the general AAPI community [3]. When the crisis first emerged, older Asian Americans were beleaguered by increased economic insecurity [5] and poorer health outcomes [6] due to a confluence of structural inequities [5].

A silver lining to the COVID-19 pandemic is that it cast a spotlight on a long-underserved group. The barrage of attacks against older Asian Americans galvanized both individuals and organizations into assisting them in various ways, such as by distributing safety whistles and meal vouchers [7]. On Twitter, now known as X, support for them coalesced around the hashtag #ProtectOurElders [4]. The objective of this study is to explore the kinds of discourse surrounding older Asian Americans during the COVID-19 crisis, specifically in relation to the surge in anti-Asian sentiments.

Dating back to the nineteenth century, one of the most pervasive stereotypes of Asian Americans is that they are a high-achieving demographic [8]. While seemingly innocuous, this myth of them as a “model minority” has been criticized as highly problematic. Not only does it run counter to their lived realities—plenty of evidence has exposed the widespread inequalities confronted by various subgroups within the community [8,9]—it also delegitimizes their struggles and feeds the misconception that they require no assistance whatsoever [5].

Racial discrimination is well known to be a key social determinant of health [6,10]. Among Asian Americans in the United States, experiences of discrimination are linked to poorer mental health outcomes, including anxiety, depression, hypertension, and elevated blood pressure [10]. Racism may exacerbate health issues brought about by the aging process, such as the onset of chronic diseases or functional impairment [11], rendering older Asian Americans more susceptible to detrimental health outcomes.

Studies have indicated that social support has a positive impact on both the mental and physical health of older adults [12]. Social support likewise serves as a protective buffer against the negative effects of racial discrimination on one’s health [13,14]. The role of social support may be especially critical for Asian Americans. Although the Asian American populace includes a diverse array of ethnicities, cultures, and languages, collectivism appears to be a cultural orientation shared among many Asian American groups [15]. Evidence revealed that social support improved health outcomes among Asian Americans during the start of the pandemic, when anti-Asian sentiments were rampant [14].

It is widely acknowledged that in Asian societies, attitudes toward older adults are typically informed by values of respect and filial piety [11,16]. Old age bespeaks knowledge and wisdom, and younger people are expected to honor and respect their older counterparts [11]. Despite concerns that such values have eroded, there is evidence that they continue to resonate with Asian Americans [17]. One study concluded that Asian Americans are twice as likely as the general population to care for their parents [18]. Even so, ageism has been discovered to be pan-cultural [19]. A meta-analysis comparing Western and Eastern attitudes toward older adults revealed that Easterners actually harbored more negative views of older adults than Westerners [20]. In this analysis, Western countries included anglophone countries in the West such as Australia, Canada, the United Kingdom, and the United States, as well as Western European countries like Switzerland and France. Eastern countries covered countries in different regions of Asia, such as East Asia, South Asia, and Southeast Asia [20].

First proposed in 2002, the stereotype content model maintains that people stereotype others on the basis of warmth and competence [21]. The dimension of warmth includes qualities such as friendliness and sincerity, while the dimension of competence includes traits such as intelligence and skillfulness [21]. According to the stereotype content model, perceptions of social groups can be categorized into four clusters: (1) warm and competent, (2) incompetent and cold, (3) competent and cold, and (4) warm and incompetent. These 4 combinations of stereotypes produce distinct emotional responses among those who hold them. Groups stereotyped as warm and competent elicit admiration. Those evaluated as incompetent and cold elicit contempt. Groups stereotyped as competent and cold evoke envy. Those evaluated as warm but incompetent evoke pity [21].

A large body of work has evinced that older adults are generally stereotyped as warm but incompetent [21]. Although they elicit feelings of admiration occasionally, they predominantly evoke pity. Evidence attests to the universality of these stereotypes in both individualistic and collectivistic societies [19]. The evaluation of older adults as warm but lacking in competence may lend itself to benevolent ageism—a paternalistic form of prejudice founded on the assumption that older adults are helpless or pitiful [22]. Benevolent ageism has intensified over the course of the pandemic owing to recurring depictions of older adults as an at-risk group [23].

Asian Americans—older or otherwise—are one of the most underresearched ethnic groups in peer-reviewed literature

XSL•FO
RenderX

[24,25]. In spite of the discomfiting rise in violence directed at them during the COVID-19 outbreak, discourse surrounding older adults from the Asian American community has escaped the attention of gerontologists. Most social media analyses conducted before and during the pandemic have focused on the discursive construction of the older population as a whole [26-28]. Other social media analyses have concentrated on the general Asian American population [29-31]. This study is therefore conceptually significant in that it is the first to dissect the content of tweets posted about older Asian Americans during the COVID-19 crisis.

At the heart of the concept of intersectionality is the notion that various social positions—such as race, age, gender, and socioeconomic status—interact to shape the types of biases one confronts [32]. From an intersectional standpoint, age and race may converge in ways that worsen the experience of discrimination for older Asian Americans [33]. In addition to being part of a racial group that faces more systemic challenges compared to White people, older Asian Americans also face age-related hurdles [34]. In terms of practical significance, this study serves as a reflection of the level of support being extended to older Asian Americans, even as it provides timely insights that will ultimately advance equity for them.

This study pivots around the following questions: What types of discourse have emerged in relation to older Asian Americans and the need to support them? How do age and race interact to shape these discourses? What are the implications of these discourses for older Asian Americans?

Methods

Data Set

We retrieved the data using version 2 of Twitter's application programming interface (API) [35], which was accessed through Twitter's academic research product track [36]. Compared to what was achievable with the standard version 1.1 API, the version 2 API grants users a higher monthly tweet cap and access to more precise filters [37].

To build an extensive data set, we collected the tweets using 2 search queries. For both queries, "retweets" were excluded, and only English tweets posted from January 1, 2020, to August 1, 2023, were collated. We excluded retweets to avoid including duplicate content in the data set, which could skew the significance of particular topics. Tweets collected through the first query (n=1549) contained the hashtag #ProtectOurElders. For the second query (n=4550), we gathered tweets that met the following inclusion criteria: (1) contained either the hashtag #StopAAPIHate or #StopAsianHate; (2) included "elder," "elderly," "old(er) adult(s)," "old(er) people," "old(er) person(s)," "senior(s)," "aged," "old folk(s)," "grandparent(s)," "grandfather(s)," "grandmother(s)," "grandpa," or "grandma." The 2 queries yielded a total of 6099 tweets.

We removed posts that were (1) contextually irrelevant, that is, discussed content not pertaining to anti-Asian attacks, such as tweets related to getting vaccinated to protect older people, or tweets related to protecting older adults from cybercrime (n=1384); (2) repeated in the 2 queries (n=20); (3) incorrectly retrieved by the API, that is, they did not fulfill the inclusion criteria of either search query (n=258); and (4) informative, factual, or descriptive (eg, tweets that were newspaper headlines) or that brought up the older person in a tangential fashion (eg, tweets that mentioned older Asian Americans alongside several other groups; n=3443). After applying the aforementioned exclusion criteria, the data set consisted of 994 tweets. Figure 1 provides a flowchart of the data collection process.

XSL•FO
RenderX

Figure 1. Process of collecting tweets about older Asian Americans in relation to anti-Asian sentiments during the COVID-19 pandemic.

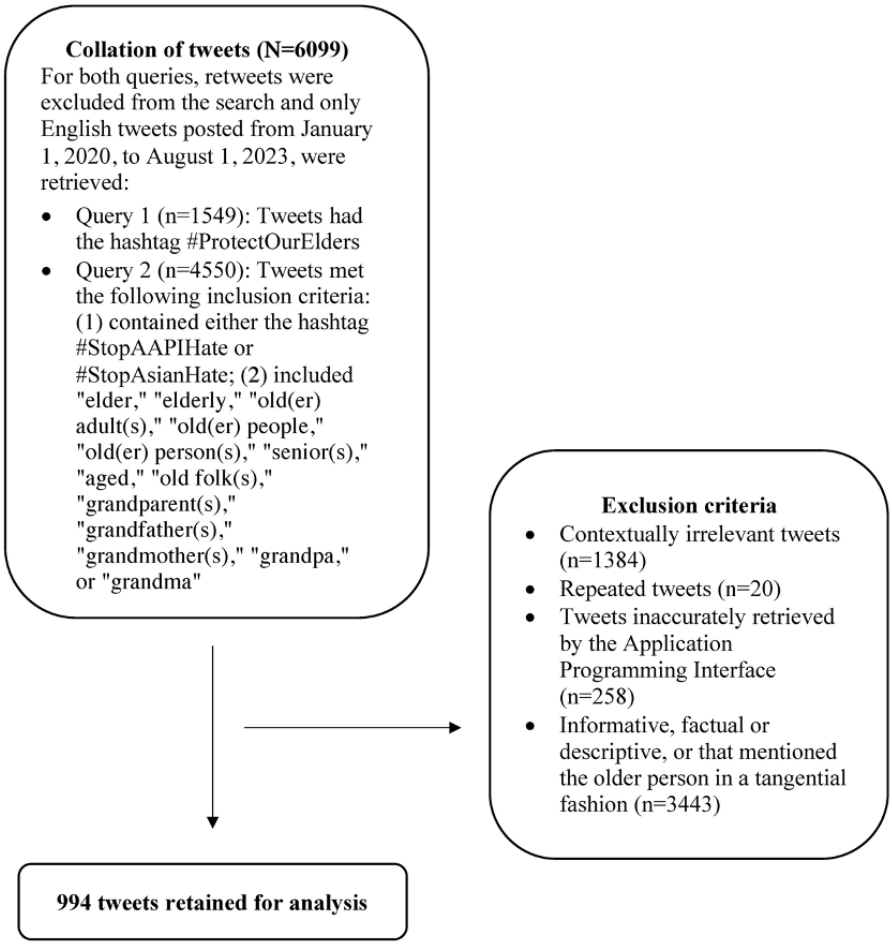

Tweet Content Coding

Consistent with past research [27,38-41], the codebook was designed through both deductive and inductive modes of reasoning [42]. Analyses led by a directed or deductive approach begin with the identification of an initial set of codes based on previous literature [43]. Conversely, in inductive content analyses, codes are derived directly from the data [43]. We used both deductive and inductive approaches to make sure certain pertinent assumptions guided the analysis while also being aware that new categories would surface inductively [42].

To create a preliminary codebook, we first identified a set of categories based on previous literature regarding the perceptions of older adults in Asia [44]. The content analysis was subsequently conducted in several stages, with each tweet read twice by 2 researchers trained in gerontology to ensure familiarity with and immersion in the data [43]. The goal of the first reading was to ascertain the validity of the initial set of categories as well as to generate codes systematically across the whole data set. Each researcher modified the codebook independently until all variables were refined and clearly defined. During this first reading, a new category was added whenever a post featured a particular trait that could not be suitably coded into any of the existing categories and which was recurrent in the data. During the second reading, the 2 coders had frequent discussions where any discrepancies were reviewed and adjudicated to ensure rigor in the analysis. At this point, both coders discussed what the codes meant, confirmed the relevance of the codes to the research question, and identified areas of significant overlap to finalize the coding rubric.

The percentage agreement between the 2 raters was 92.5% with a weighted Cohen κ of 0.89 (P<.001), indicating high interrater reliability. A total of 4 themes emerged from the whole process. The frequency of each theme was identified after the analysis. As mentioned in past scholarship, categories in a content analysis need not be mutually exclusive, although they should be internally homogeneous (ie, coherent within themes) and externally heterogeneous (ie, distinct from each other) as far as possible [27,45].

Ethical Considerations

Ethical approval was not deemed necessary for this study, as all the data used were publicly available and anonymized.

Results

Summary of Insights From Content Analysis of Tweets

A total of 4 themes emerged from our content analysis of 994 tweets. Half of the posts (498/994, 50.1%) were filed under the theme “vulnerable and in need of protection” (theme 1). Tweets in this theme either singled out older Asian Americans as a group in need of protection because of their vulnerable status or discussed initiatives aimed at safeguarding their well-being. The theme “heroic and resilient” (theme 2) was present in 31.1%

XSL•FO
RenderX

(309/994) of the posts. Relevant tweets celebrated older Asian Americans for displaying tremendous strength in the face of attack or described them as individuals not to be trifled with. The theme "immigrants who have made selfless contributions and sacrifices" (theme 3) appeared in 10.2% (102/994) of the posts. Posts in this section referenced the immense sacrifices made by older Asian Americans as they migrated to the United States, as well as the systemic barriers they had to overcome. Theme 4 "worthy of honor" (85/994, 8.5%) consisted of tweets that venerated older Asian Americans. Textbox 1 provides a summary of the themes.

Textbox 1. Description of themes present in tweets about older Asian Americans in relation to anti-Asian sentiments during the COVID-19 pandemic.

Vulnerable and in need of protection (498/994, 50.1%)

- "Isn't it so cowardly that they attack the elderly mostly? Not that violence is acceptable for any age, but to hurt the defenseless only means they got loose screws. #StopAsianHate"
- "Conducting walking patrols everyday to protect our elders and community #StopAAPIHate #HateisaVirus #StopAsianHate #SFChinatown #SafeNeighborhood #ProtectOurElders #TogetherWeCan"

Heroic and resilient (309/994, 31.1%)

- "Underestimating the terror wrought by old Chinese ladies with sticks was his first mistake #grannygoals #StopAsianHate"
- "Don't mess with Asian grandmas. But also sad this is happening. #StopAsianHate #StopAAPIHate"

Immigrants who have made selfless contributions and sacrifices (102/994, 10.2%)

- "Come to America they said..

It's the land of Opportunities they said...

Feeling so sad seeing this video 2 underage over privileged girls get to do this to a man ,a father ,a grandfather and not even have their identities revealed ...devastating

#MuhammadAnwar #StopAsianHate"

- "These are my grandparents. They came to America to build a new life. (That's my dad on the right wearing a tie.) My grandfather was a very well respected doctor in the Chinese community. America is built on the backbone of hard-working immigrants. #StopAsianHate"

Worthy of honor (85/994, 8.5%)

- "What's been shocking to me about these increased attacks on #AAPI is how often the elderly have been the focus. It's such a shock because one thing that has been common amongst #AAPI culture is the reverence/respect of elders. #StopAAPIHate #StopAsianHate"
- "It really makes me weak and cry seeing videos of those elderly being hit and hurt. We, Asians, value and esteem our elderly. We even live with them in the same house, take care of them. I can't imagine how someone can simply push them. Just like that. #StopAsianHate"

Theme 1: Vulnerable and in Need of Protection

The vulnerability of older adults was a throughline in this category (498/994, 50.1%). Although concern was directed at the entire Asian American population, older adults were singled out as deserving of more sympathy because of their advanced age. Adjectives commonly used to frame them include "infirm," "weak," "defenseless," and "powerless." A person described them as lacking "the strength to even unclasp a grip." Sympathy for older adults was magnified in view of other challenges they had been confronting since the outbreak of COVID-19. For instance, one poster expressed sorrow over how older Asian Americans had to grapple with the "fear of getting attacked" on top of "already [being] really afraid of COVID-19 because it disproportionately affects" them.

What made the act "especially egregious" in the eyes of many was the fact that assailants targeted older adults of all people. Users lambasted attackers for their "coward[ice]," asserting that they should have "picked on someone [their] own size" instead of attacking "people who can't even defend themselves." Several posters insisted that it was incumbent upon society to "be watchdogs" for older adults since they are more vulnerable.

A large number of tweets featured a call-to-action aimed at mobilizing members of the Twitter community to assist older Asian Americans. Fundraising campaigns were conducted to raise money for "alarms and pepper spray" for older Asian Americans. Others lobbied for donations to causes that deliver food to this group. The following tweet is one such example: "Wondering how you can support elderly Asians and show you will not tolerate #Asianhate? Join me in making a contribution to @heartofdinner, which brings food to elderly Asians in NYC so they can eat safely in their homes #StopAsianHateCrimes #StopAAPIHate." The Twitter audience was also invited to escort older persons who walk alone: "United Peace Collaborative protects the #SF Chinatown community with daily walking patrols, providing protection & assistance to the elderly & residents. Please join us & volunteer!"

There were many tweets concerning the suite of initiatives aimed at supporting older Asian Americans. The Yellow Whistle—a campaign involving the distribution of free whistles for Asian Americans to signal danger in the event of an assault—was held up as one such example to "keep older Asian Americans safe." Select community partners also received plaudits for their

XSL•FO
RenderX

"wonderful work in distributing and training use of the alarms to" older persons.

Theme 2: Heroic and Resilient

Tweets in this theme (309/994, 31.1%) mainly revolved around a high-profile incident in San Francisco in which Xiao Zhen Xie, an older woman of Asian descent, put her assailant on a stretcher in an unexpected turn of events. She earned kudos from the Twitter community for "hold[ing] her ground," "fighting back," and sending him "to the hospital with his face bloodied." Many saluted her for being "feisty," "resilient," and "[as] tough as nails," dubbing her a "hero" who made them feel "#HonoredToBeAsian." One user used the hashtag "#GrannyGoals," quipping that the attacker made a "mistake" "underestimating the terror" that "old Chinese ladies" could wreak. Xiao Zhen Xie was also applauded for "refusing to be a statistic" as well as for defying the image of older adults as a group most expect "not to fight back."

This episode involving Xiao Zhen Xie set in motion a series of tweets in which users warned others not to get on their grandparents' bad side. A user cautioned that the incident was a lesson to everyone not to "mess with ahjummas, lolas, and all the elderly Asian women." Another claimed that Asian grandmothers possess a special kind of "Asian grandma strength." Some took the opportunity to underline the importance of not belittling older adults, with one in particular commenting on how his or her grandparents embodied grit and "toughness" because they "lived through war."

Besides Xiao Zhen Xie, a few other older Asian Americans were celebrated for their resilience. A Filipina immigrant, Vilma Kari, was lauded for saying she "forgives" and "prays" for her attacker. A handful of tweets focused on a group of older Asian Americans who made the headlines for having filmed a music video in which they condemned the racially motivated acts of violence targeting their community.

Theme 3: Immigrants Who Have Made Selfless Contributions and Sacrifices

Members of the Twitter community frequently shared stories of their grandparents' immigration (102/994, 10.2%). A common thread running through these posts was that their forefathers made immense sacrifices, uprooting themselves to move to the United States in order that their children might receive "the best education they can get" and "enjoy a "better future." A user portrayed his or her grandmother as a "fighter" who "worked two to three jobs" while struggling to acculturate in a new society at a time when she knew "very little English."

Attention was drawn to how the string of attacks against Asian Americans was ironic given the national ethos of the country commonly touted as the "American dream." A few posters implied that labeling the United States as a "land of opportunity" was a misnomer: "Come to America,' they said... 'It's the land of opportunities,' they said..." A user said that the Asian "elderly did not escape communism" only to become a target of racism.

Tweets in this theme also discussed the burden of racism that older Asian Americans had endured before the COVID-19 pandemic. Users commented on their grandparents' day-to-day experiences of racial discrimination. A handful were dismayed by how their grandparents were survivors of "prejudice and xenophobia" during World War II when they were forcibly relocated to Japanese internment camps. Others bemoaned that their older family members were "imprisoned for being the wrong-colored Americans." One user deplored the fact that his or her grandfather "could not come to [the United States] because of his race" due to the Chinese Exclusion Act of 1882, a law that suspended Chinese immigration for 10 years and declared Chinese immigrants ineligible for naturalization. Another poster pinpointed how his or her grandfather felt compelled to dress in an "extremely patriotic" manner in order to camouflage his Asian identity and better assimilate into America.

Users considered older Asian Americans as foundational to the growth of America and foregrounded the need to acknowledge that "America is built on the backbone of hardworking immigrants," who "made 90%" of what society has. Examples of contributions made by those of Asian ancestry include how they "oversaw" the construction of the transcontinental railroad in the "Old West" as well as their "service in the #442RCT (442nd Infantry Regiment)"—a highly decorated infantry regiment that mainly comprised second-generation American soldiers of Japanese descent who served in World War II. One user mentioned Chien-Shiung Wu, a groundbreaking Chinese American physicist whose scientific accomplishments were a core part of "U.S. WW II efforts" and that "helped win Nobel Prizes for Americans," without which the "country would be so much worse off." Artworks inspired by "hustling, elderly Asian folks" were also broadcasted under a hashtag that deified them as "#ChinatownGods."

Several attempts were made to deconstruct the myth of the model minority. Individuals were aggrieved at how the looming specter of anti-Asian violence compounded the plight of older Asian Americans, who had already been dealt multiple blows during the COVID-19 crisis. These posters raised awareness of how many of them are in "precarious living situations" or "working in low-wage jobs." Some pleaded for the Asian American community to be seen and understood, as captured in the following tweet: "See what's happening to our elderly and community. Understand us. Understand why no matter how model of a minority we seem to be... we are just like you. #StopAsianHate #StopAAPIHate #StandWithAsians."

Theme 4: Worthy of Honor

Many users (85/994, 8.5%) were outraged at how older adults appeared to be prime targets of violence against the Asian American community, perceiving these acts as a flagrant transgression of Asian cultural mores that "revere" them as "the most important people" in society. Some tweets exalted them as wellsprings of "wisdom" and "thoughtful guidance"—one user even likened them to "gold"—to "value and esteem." Tweets in this theme also alluded to how deference to the older community was practically nonnegotiable in the Asian household. A poster tweeted, "No one should be assaulted, especially the elderly. I grew up respecting my elders. You never even argued with them ... They pass on wisdom."

XSL•FO
RenderX

Values of collectivism were prized by certain users. These posters made reference to the notion of intergenerational reciprocity by stressing that younger people had an obligation to "protect" the older generation in return. The idea of solidarity was also raised. For instance, some viewed the attack of an older adult—related or otherwise—as an affront to the entire Asian community: "Many are saying 'she could've been MY grandma.' To that I say, she is ALL OUR GRANDMAS. Fight hate, love justice, stand with our elders always. #ForTheLoveOfLolas #StopAsianHate #StopAAPIHate #StopAsianHateCrimes."

Discussion

Overview

This study serves as a substantive first step in understanding discourses surrounding older Asian Americans. In our content analysis of tweets posted about the rash of attacks targeting them during the COVID-19 crisis, 4 main discourses surfaced. The first positioned them as "vulnerable and in need of protection" (theme 1). The second characterized them as "heroic and resilient" (theme 2). The third portrayed them as "immigrants who have made selfless contributions and sacrifices" (theme 3), and the fourth extolled them as "worthy of honor" (theme 4).

Our findings demonstrate an outpouring of support for the older Asian American community, which manifested itself in various local initiatives such as the distribution of safety whistles and the delivery of food. Scholars have drawn attention to how social support is particularly crucial for those in their later years [12] as well as those who experience racial discrimination [13,14]. The fact that older Asian Americans are finally being given support and assistance is therefore a step in the right direction.

However, even well-intentioned acts may be met with negative repercussions. In the wake of the COVID-19 crisis, older adults were reduced to a uniform group of at-risk individuals [46]. Assumptions of their vulnerability led to paternalistic behaviors, which denied them their autonomy [23]. Our results indicate that the rise in violence toward older Asian Americans sparked much-needed dialogue regarding their everyday struggles. Nevertheless, an unfortunate corollary is that this may have predisposed them to being recipients of benevolent prejudice on the basis of both age and race. Older Asian Americans may have been viewed as especially defenseless or vulnerable, perhaps more so than the general older population. This was made amply clear in the findings, where half of the tweets branded older Asian Americans as "weak" and "powerless."

Notwithstanding concerns that Asian values of respect and filial piety have become irrelevant in the face of modernization [17], findings from themes 2-4 show emphatically that older adults retain their revered status, at least among some in the Asian American community. Tweets in theme 2 featured users enthusing over the way Xiao Zhen Xie held her ground when she was attacked in San Francisco, which led to deliberations on the strength and tenacity of older Asian women in general. Discourses of gratitude emerged in theme 3 as users ruminated over the sacrifices their forefathers had made in migrating to the United States, as well as the attendant systemic challenges they had to navigate. Posts in theme 4 indicate that users perceived the violence against older Asian Americans as a contravention of cultural norms, which emphasize the importance of honoring older adults. These provide a countervailing force to the various ageist tropes that came to the fore during the COVID-19 pandemic, such as the #BoomerRemover hashtag, which saw the lives of older people being discounted [27,28].

Theoretical Contribution and Implications

Findings from this study show that during the COVID-19 pandemic, age and race interfaced in complex ways to shape discourses on older Asian Americans. Specifically, our content analysis demonstrates that the stereotypes of warmth and incompetence, which are often thought to shape evaluations of older adults, cannot be applied indiscriminately to older Asian Americans as a subcategory of the older demographic. Theme 1, which positions older Asian Americans as vulnerable and in need of protection, does indeed align with traditional evaluations of older adults as warm and incompetent. However, the remaining themes celebrate older Asian Americans for their numerous contributions to society, the sacrifices they have made, and their unwavering resilience during the pandemic, all of which challenge the stereotype of incompetence under the stereotype content model. These findings add complexity to the commonly held notion of older adults as a pitiful social group by highlighting that older Asian Americans evoke not just pity but also admiration. The stereotype content model should therefore be expanded or modified in a way that accounts for attitudes toward older adults of different ethnicities.

Additionally, gerontological scholarship would benefit from a cross-cultural analysis of benevolent ageism. At present, little is known about how displays of benevolent ageism are affected by cultural norms of parental respect and filial piety and the extent to which these norms affect one's perception of an older adult's competence. Several studies have been conducted to make sense of ageism in different cultures [47,48], but there has been limited research on the cross-cultural differences in benevolent ageism specifically. The ways in which evaluations of older Asian Americans may be complicated by the deeply ingrained myth of the model minority as well as the pandemic-induced rise in anti-Asian hate are important avenues for future study.

This study has a number of implications for policy and practice. First, although care toward one's parents or grandparents is not the prerogative of Asians [49], Asia's adherence to collectivism nonetheless offers a useful learning point for the West. Many of the posters were Asian Americans, who held older adults in high regard, whether related or otherwise. Fostering a cultural emphasis on solidarity and interconnectedness in the West may promote respect not only for one's parents but also for older adults outside of one's family [44]. Second, ongoing efforts to reframe aging [50] could highlight the need to respect older adults, not in a way that advances their supremacy or absolves them from wrongdoing, but in a way that teaches society to view them as people whose experience may render them wise and worth learning from. Educators could also incorporate

XSL•FO
RenderX

lessons on age-related stereotypes in schools to guard against the formation of ageist beliefs [51].

Third, current moves to redress the longstanding omission of Asian American history from national curricula [52] should ensure that students in every state are taught about the sacrifices, struggles, and contributions of older Asian Americans. Public campaigns could be organized as well to raise awareness of the aforementioned. This will help counter the myth of the model minority and get more people to acknowledge older Asian Americans as a significant part of America's social fabric. Fourth, our findings underscore the need to reflect on the diversity of the older population in terms of socioeconomic status. Older adults—particularly those from the baby boomer generation—have been stereotyped as having made significant financial gains compared to their predecessors, at times even seen as stealing resources from the young [53]. However, as highlighted by some of the Twitter users as well as scholars, many older Asian Americans are in dire economic straits [5]. Rectifying the structural inequities that have contributed to their immiseration should hence be a key component of the agenda moving forward.

There are limitations inherent in this study. First, we acknowledge that Twitter users might not be representative of the wider population and that only publicly available tweets were included in the data set. Some of the users whose tweets were included in the study appeared to be Asian Americans, who are likely to be more passionate about supporting individuals in their community. Relatedly, as we did not collect information regarding users' demographics—not all users publish demographic information, and there are certain limitations to using publicly provided demographic information on social media [54]—we could not contextualize the motivations of those whose tweets were included in the analysis. Ultimately, social support for older Asian Americans—whether from the Asian American community or society as a whole—has important implications for their well-being [14]. Subsequent research could focus on conducting interviews among individuals from different ethnic groups to tease out any differences in the level of support extended to older Asian Americans.

Second, we queried the hashtag #StopAAPIHate as a way to understand sentiments toward Asian Americans, even though the term "AAPI" refers to 2 different racial groups: Asian Americans and Pacific Islanders. As the tweets analyzed paid more attention to older Asian Americans, we were not able to offer insight into the types of discourses that emerged in relation to older Pacific Islanders. Future studies are needed to expound on such discourses. Third, it is vital to highlight that both the Asian American community and the older population are heterogeneous. The Asian American community encompasses numerous ethnicities, all with distinct languages, cultures, immigration histories, values, and beliefs [34]. The older demographic, too, is a diverse group composed of people with vastly different health trajectories [55]. Given the brevity of the tweets uploaded, we were unable to assess how discourses on older Asian Americans vary across different ethnicities. Finally, we collected only textual data, although tweets often contain visual elements such as photos, videos, and GIFs. This is a drawback that can be overcome in the future when multimodal techniques are developed to analyze both textual and visual content on Twitter.

Another direction for future inquiry involves an assessment of how discourses surrounding older Asian Americans have changed over time. The level of support shown to this group is likely to fluctuate over time, depending on the frequency at which anti-Asian attacks are reported in the news as well as other types of news being covered. Sentiment and narrative analyses [56-58] could be performed to glean such insights.

Conclusion

Even as older Asian Americans contended with a rise in racism alongside other struggles during the COVID-19 pandemic, our findings reveal that the crisis had the unintended effect of garnering greater support for this group. In the future, it is important that support be extended to older Asian Americans not so much by virtue of their perceived vulnerability but more so in view of their boundless contributions and sacrifices.

Acknowledgments

The authors would like to thank L Liu for preprocessing the data. We gratefully acknowledge support from the Social Science Research Council SSHR Fellowship (MOE2018-SSHR-004). The funder had no role in study design, data collection, analysis, writing, or the decision to publish this study.

Data Accessibility

Data are publicly available on Twitter [59].

Authors' Contributions

RN designed the study, developed the methodology, analyzed the data, wrote the paper, acquired the funding. RN and NI analyzed the data and wrote the paper.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Elias A, Ben J, Mansouri F, Paradies Y. Racism and nationalism during and beyond the COVID-19 pandemic. Ethn Racial Stud 2021;44(5):783-793. [doi: 10.1080/01419870.2020.1851382]
2. Yamauchi N. Anti-Asian hate crimes quadrupled in U.S. last year. Nikkei Asia. 2022. URL: https://asia.nikkei.com/Spotlight/Society/Anti-Asian-hate-crimes-quadrupled-in-U.S.-last-year [accessed 2022-02-16]
3. Turton N. Stop AAPI Hate: new data on anti-Asian hate incidents against elderly and total national incidents in 2020. Stop AAPI Hate. 2021. URL: https://stopaapihate.org/wp-content/uploads/2021/04/Stop-AAPI-Hate-Press-Statement-Bay-Area-Elderly-Incidents-210209.pdf [accessed 2024-02-23]
4. Huang J. How the #ProtectOurElders movement helped create a wave of first-time Asian American Activists. LAist. 2021. URL: https://laist.com/news/how-protectourelders-helped-created-a-wave-of-first-time-asian-american-activists [accessed 2022-04-03]
5. Ma KPK, Bacong AM, Kwon SC, Yi SS, Đoàn LN. The impact of structural inequities on older Asian Americans during COVID-19. Front Public Health 2021;9:690014 [FREE Full text] [doi: 10.3389/fpubh.2021.690014] [Medline: 34490181]
6. Chen JA, Zhang E, Liu CH. Potential impact of COVID-19-related racial discrimination on the health of Asian Americans. Am J Public Health 2020;110(11):1624-1627. [doi: 10.2105/AJPH.2020.305858] [Medline: 32941063]
7. Lee E. Asian American community rallies to support elders. AARP. 2021. URL: https://www.aarp.org/home-family/friends-family/info-2021/asian-american-support-communities.html [accessed 2022-02-16]
8. Yi V, Museus S. Model minority myth. In: Smith AD, Hou X, Stone J, Dennis RM, Rizova P, editors. The Wiley Blackwell Encyclopedia of Race, Ethnicity, and Nationalism. Oxford, UK: John Wiley & Sons, Ltd; 2015.
9. Li G, Wang L. Model Minority Myth Revisited: An Interdisciplinary Approach to Demystifying Asian American Educational Experiences. Charlotte, NC: Information Age Publishing; 2008.
10. Paradies Y, Ben J, Denson N, Elias A, Priest N, Pieterse A, et al. Racism as a determinant of health: a systematic review and meta-analysis. PLoS One 2015;10(9):e0138511 [FREE Full text] [doi: 10.1371/journal.pone.0138511] [Medline: 26398658]
11. Min J, Moon A. Older Asian Americans. In: Berkman B, D'Ambruoso S, editors. Handbook of Social Work in Health and Aging. New York, NY: Oxford University Press; 2006.
12. Antonucci TC, Ajrouch KJ, Birditt KS. The convoy model: explaining social relations from a multidisciplinary perspective. Gerontologist 2014;54(1):82-92 [FREE Full text] [doi: 10.1093/geront/gnt118] [Medline: 24142914]
13. Ajrouch KJ, Reisine S, Lim S, Sohn W, Ismail A. Perceived everyday discrimination and psychological distress: does social support matter? Ethn Health 2010;15(4):417-434 [FREE Full text] [doi: 10.1080/13557858.2010.484050] [Medline: 20582775]
14. Lee S, Waters SF. Asians and Asian Americans' experiences of racial discrimination during the COVID-19 pandemic: impacts on health outcomes and the buffering role of social support. Stig Health 2021 Feb;6(1):70-78. [doi: 10.1037/sah0000275]
15. Markus HR, Kitayama S. The cultural construction of self and emotion: implications for social behavior. In: Emotion and Culture: Empirical Studies of Mutual Influence. Washington, DC: American Psychological Association; 1994:89-130.
16. Ingersoll-Dayton B, Saengtienchai C. Respect for the elderly in Asia: stability and change. Int J Aging Hum Dev 1999;48(2):113-130. [doi: 10.2190/G1XR-QDCV-JRNM-585P] [Medline: 10376957]
17. Harrington B. "It's more us helping them instead of them helping us": how class disadvantage motivates Asian American college students to help their parents. J Fam Issues 2022;44(7):1773-1795. [doi: 10.1177/0192513x211064867]
18. Montenegro X. Caregiving among Asian Americans and Pacific Islanders age 50+. AARP Research. 2014. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving-asian-americans-pacific-islanders/ [accessed 2024-02-23]
19. Cuddy AJC, Norton MI, Fiske ST. This old stereotype: the pervasiveness and persistence of the elderly stereotype. J Social Issues 2005;61(2):267-285. [doi: 10.1111/j.1540-4560.2005.00405.x]
20. North MS, Fiske ST. Modern attitudes toward older adults in the aging world: a cross-cultural meta-analysis. Psychol Bull 2015;141(5):993-1021. [doi: 10.1037/a0039469] [Medline: 26191955]
21. Fiske ST, Cuddy AJC, Glick P, Xu J. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. J Pers Soc Psychol 2002;82(6):878-902. [doi: 10.1037/0022-3514.82.6.878]
22. Cary LA, Chasteen AL, Remedios J. The ambivalent ageism scale: developing and validating a scale to measure benevolent and hostile ageism. Gerontologist 2017;57(2):e27-e36 [FREE Full text] [doi: 10.1093/geront/gnw118] [Medline: 27520730]
23. Vervaecke D, Meisner B. Caremongering and assumptions of need: the spread of compassionate ageism during COVID-19. Gerontologist 2021;61(2):159-165 [FREE Full text] [doi: 10.1093/geront/gnaa131] [Medline: 32920642]
24. Yi SS. Taking action to improve Asian American Health. Am J Public Health 2020;110(4):435-437. [doi: 10.2105/AJPH.2020.305596] [Medline: 32160000]
25. Đoàn LN, Takata Y, Sakuma KLK, Irvin VL. Trends in clinical research including Asian American, Native Hawaiian, and Pacific Islander participants funded by the US National Institutes of Health, 1992 to 2018. JAMA Netw Open 2019;2(7):e197432 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.7432] [Medline: 31339543]
26. Levy BR, Chung PH, Bedford T, Navrazhina K. Facebook as a site for negative age stereotypes. Gerontologist 2014;54(2):172-176 [FREE Full text] [doi: 10.1093/geront/gns194] [Medline: 23392643]

XSL•FO
RenderX

27. Sipocz D, Freeman JD, Elton J. "A toxic trend?": generational conflict and connectivity in Twitter discourse under the #BoomerRemover hashtag. Gerontologist 2021;61(2):166-175 [FREE Full text] [doi: 10.1093/geront/gnaa177] [Medline: 33159524]
28. Skipper AD, Rose DJ. #BoomerRemover: COVID-19, ageism, and the intergenerational twitter response. J Aging Stud 2021;57:100929 [FREE Full text] [doi: 10.1016/j.jaging.2021.100929] [Medline: 34082999]
29. Hswen Y, Xu X, Hing A, Hawkins JB, Brownstein JS, Gee GC. Association of "#covid19" versus "#chinesevirus" with Anti-Asian sentiments on Twitter: march 9-23, 2020. Am J Public Health 2021;111(5):956-964 [FREE Full text] [doi: 10.2105/AJPH.2021.306154] [Medline: 33734838]
30. Nguyen TT, Criss S, Dwivedi P, Huang D, Keralis J, Hsu E, et al. Exploring U.S. shifts in Anti-Asian sentiment with the emergence of COVID-19. Int J Environ Res Public Health 2020;17(19):7032 [FREE Full text] [doi: 10.3390/ijerph17197032] [Medline: 32993005]
31. Darling-Hammond S, Michaels EK, Allen AM, Chae DH, Thomas MD, Nguyen TT, et al. After "The China Virus" went viral: racially charged coronavirus coverage and trends in bias against Asian Americans. Health Educ Behav 2020;47(6):870-879 [FREE Full text] [doi: 10.1177/1090198120957949] [Medline: 32911985]
32. Krekula C, Nikander P, Wilińska M. Multiple marginalizations based on age: gendered ageism and beyond. In: Ayalon L, Tesch-Römer C, editors. Contemporary Perspectives on Ageism. Cham, Switzerland: Springer Open; 2018:33-50.
33. Gutterman AS. Ageism, race and ethnicity. SSRN Journal. Preprint posted online on December 8 2021 2021 [FREE Full text] [doi: 10.2139/ssrn.3972849]
34. Kim G, Wang SY, Park S, Yun SW. Mental health of Asian American older adults: contemporary issues and future directions. Innov Aging 2020;4(5):igaa037 [FREE Full text] [doi: 10.1093/geroni/igaa037] [Medline: 33274302]
35. Twitter API v2: early access. Twitter. 2021. URL: https://developer.twitter.com/en/docs/twitter-api/early-access [accessed 2021-10-13]
36. Tornes A, Trujillo L. Enabling the future of academic research with the Twitter API. Twitter Developer Platform Blog. 2021. URL: https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api [accessed 2021-10-13]
37. Barrie C, Ho JC. academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. JOSS 2021;6(62):3272 [FREE Full text] [doi: 10.21105/joss.03272]
38. Ng R, Indran N. Does age matter? Tweets about gerontocracy in the United States. J Gerontol B Psychol Sci Soc Sci 2023;78(11):1870-1878. [doi: 10.1093/geronb/gbad113] [Medline: 37584274]
39. Ng R, Indran N. Innovations for an aging society through the lens of patent data. Gerontologist 2024;64(2) [FREE Full text] [doi: 10.1093/geront/gnad015] [Medline: 37497634]
40. Ng R, Indran N. Not too old for TikTok: how older adults are reframing aging. Gerontologist 2022;62(8):1207-1216 [FREE Full text] [doi: 10.1093/geront/gnac055] [Medline: 35512401]
41. Ng R, Indran N. Age advocacy on Twitter over 12 years. Gerontologist 2024;64(1). [doi: 10.1093/geront/gnac183] [Medline: 36516463]
42. Armat MR, Assarroudi A, Rad M, Sharifi H, Heydari A. Inductive and deductive: ambiguous labels in qualitative content analysis. TQR 2018;23(1):219-221 [FREE Full text] [doi: 10.46743/2160-3715/2018.2872]
43. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
44. Hwang KK. Filial piety and loyalty: two types of social identification in Confucianism. Asian J of Social Psycho 2002;2(1):163-183. [doi: 10.1111/1467-839x.00031]
45. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14 [FREE Full text] [doi: 10.1016/j.npls.2016.01.001]
46. Ayalon L. There is nothing new under the sun: ageism and intergenerational tension in the age of the COVID-19 outbreak. Int Psychogeriatr 2020;32(10):1221-1224 [FREE Full text] [doi: 10.1017/S1041610220000575] [Medline: 32284078]
47. Ng R, Lim-Soh JW. Ageism linked to culture, not demographics: evidence from an 8-billion-word corpus across 20 countries. J Gerontol B Psychol Sci Soc Sci 2021;76(9):1791-1798 [FREE Full text] [doi: 10.1093/geronb/gbaa181] [Medline: 33099600]
48. Löckenhoff CE, De Fruyt F, Terracciano A, McCrae RR, De Bolle M, Costa PT, et al. Perceptions of aging across 26 cultures and their culture-level associates. Psychol Aging 2009;24(4):941-954 [FREE Full text] [doi: 10.1037/a0016901] [Medline: 20025408]
49. Lim AJ, Lau CYH, Cheng CY. Applying the Dual Filial Piety Model in the United States: a comparison of filial piety between Asian Americans and Caucasian Americans. Front Psychol 2021;12:786609 [FREE Full text] [doi: 10.3389/fpsyg.2021.786609] [Medline: 35185688]
50. Sweetland J, Volmert A, O'Neil M. Finding the frame: an empirical approach to reframing aging and ageism. FrameWorks Institute. Washington, DC; 2017. URL: https://www.frameworksinstitute.org/wp-content/uploads/2020/05/aging_research_report_final_2017.pdf [accessed 2024-02-23]
51. Russell ER, Thériault É, Colibaba A. Facilitating age-conscious student development through lecture-based courses on aging. Can J Aging 2022;41(2):283-293 [FREE Full text] [doi: 10.1017/S0714980821000246] [Medline: 35534789]

XSL•FO
RenderX

52. Chavez N. New Jersey becomes second state to require Asian American history to be taught in schools. CNN. 2022. URL: https://www.cnn.com/2022/01/18/us/new-jersey-schools-asian-american-history/index.html [accessed 2022-04-11]
53. Ng R, Indran N. Hostility toward baby Boomers on TikTok. Gerontologist 2022;62(8):1196-1206 [FREE Full text] [doi: 10.1093/geront/gnac020] [Medline: 35106587]
54. Sloan L. Who Tweets in the United Kingdom? Profiling the Twitter population using the British Social Attitudes Survey 2015. Soc Media Soc 2017;3(1):205630511769898 [FREE Full text] [doi: 10.1177/2056305117698981]
55. Diehl M, Smyer MA, Mehrotra CM. Optimizing aging: a call for a new narrative. Am Psychol 2020;75(4):577-589 [FREE Full text] [doi: 10.1037/amp0000598] [Medline: 32378951]
56. Ng R, Indran N, Suarez P. Communicating risk perceptions through Batik art. JAMA 2023;330(9):790-791. [doi: 10.1001/jama.2023.14278] [Medline: 37486662]
57. Ng R, Indran N. Reframing aging: foregrounding familial and occupational roles of older adults is linked to decreased ageism over two centuries. J Aging Soc Policy 2023:1-18. [doi: 10.1080/08959420.2023.2238538] [Medline: 37534573]
58. Ng R, Indran N. Impact of old age on an occupation's image over 210 years: an age premium for doctors, lawyers, and soldiers. J Appl Gerontol 2023;42(6):1345-1355. [doi: 10.1177/07334648231155025] [Medline: 37092180]
59. Twitter. 2021. URL: https://twitter.com/ [accessed 2021-10-13]

Abbreviations

AAPI: Asian American and Pacific Islanders
API: application programming interface

Edited by A Mavragani; submitted 19.01.23; peer-reviewed by A Atalay, A Bacong; comments to author 22.02.23; revised version received 12.03.23; accepted 14.09.23; published 29.03.24.

Please cite as:
Ng R, Indran N
#ProtectOurElders: Analysis of Tweets About Older Asian Americans and Anti-Asian Sentiments During the COVID-19 Pandemic
J Med Internet Res 2024;26:e45864
URL: https://www.jmir.org/2024/1/e45864
doi:10.2196/45864
PMID:38551624

©Reuben Ng, Nicole Indran. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identification of Predictors for Clinical Deterioration in Patients With COVID-19 via Electronic Nursing Records: Retrospective Observational Study

Sumi Sung[1], PhD; Youlim Kim[2], MPH; Su Hwan Kim[3], PhD; Hyesil Jung[4], PhD

[1]Department of Nursing Science, Research Institute of Nursing Science, Chungbuk National University, Cheongju, Chungcheongbuk-do, Republic of Korea

[2]Department of Radiation Oncology, College of Medicine, Seoul National University, Seoul, Republic of Korea

[3]Department of Information Statistics, Gyeongsang National University, Jinju, Gyeongsangnam-do, Republic of Korea

[4]Department of Nursing, College of Medicine, Inha University, Incheon, Republic of Korea

Corresponding Author:
Hyesil Jung, PhD
Department of Nursing
College of Medicine
Inha University
100 Inha-ro, Michuhol-gu
Incheon, 22212
Republic of Korea
Phone: 82 32 860 8206
Fax: 82 32 874 5880
Email: hsjung@inha.ac.kr

Abstract

Background: Few studies have used standardized nursing records with Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) to identify predictors of clinical deterioration.

Objective: This study aims to standardize the nursing documentation records of patients with COVID-19 using SNOMED CT and identify predictive factors of clinical deterioration in patients with COVID-19 via standardized nursing records.

Methods: In this study, 57,558 nursing statements from 226 patients with COVID-19 were analyzed. Among these, 45,852 statements were from 207 patients in the stable (control) group and 11,706 from 19 patients in the exacerbated (case) group who were transferred to the intensive care unit within 7 days. The data were collected between December 2019 and June 2022. These nursing statements were standardized using the SNOMED CT International Edition released on November 30, 2022. The 260 unique nursing statements that accounted for the top 90% of 57,558 statements were selected as the mapping source and mapped into SNOMED CT concepts based on their meaning by 2 experts with more than 5 years of SNOMED CT mapping experience. To identify the main features of nursing statements associated with the exacerbation of patient condition, random forest algorithms were used, and optimal hyperparameters were selected for nursing problems or outcomes and nursing procedure–related statements. Additionally, logistic regression analysis was conducted to identify features that determine clinical deterioration in patients with COVID-19.

Results: All nursing statements were semantically mapped to SNOMED CT concepts for “clinical finding,” “situation with explicit context,” and “procedure” hierarchies. The interrater reliability of the mapping results was 87.7%. The most important features calculated by random forest were “oxygen saturation below reference range,” “dyspnea,” “tachypnea,” and “cough” in “clinical finding,” and “oxygen therapy,” “pulse oximetry monitoring,” “temperature taking,” “notification of physician,” and “education about isolation for infection control” in “procedure.” Among these, “dyspnea” and “inadequate food diet” in “clinical finding” increased clinical deterioration risk (dyspnea: odds ratio [OR] 5.99, 95% CI 2.25-20.29; inadequate food diet: OR 10.0, 95% CI 2.71-40.84), and “oxygen therapy” and “notification of physician” in “procedure” also increased the risk of clinical deterioration in patients with COVID-19 (oxygen therapy: OR 1.89, 95% CI 1.25-3.05; notification of physician: OR 1.72, 95% CI 1.02-2.97).

Conclusions: The study used SNOMED CT to express and standardize nursing statements. Further, it revealed the importance of standardized nursing records as predictive variables for clinical deterioration in patients.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53343) doi:10.2196/53343

KEYWORDS

COVID-19; infectious; respiratory; SARS-CoV-2; nursing records; SNOMED CT; random forest; logistic regression; EHR; EHRs; machine learning; documentation; deterioration; health records; health record; patient record; patient records; nursing; standardization; standard; standards; standardized; standardize; nomenclature; term; terms; terminologies; terminology

Introduction

As of September 27, 2023, the World Health Organization reported more than 770 million confirmed cases of COVID-19, including approximately 6.9 million deaths [1]. In South Korea, among 34,436,586 confirmed cases between January 3, 2020, and August 31, 2023, there were 35,812 deaths attributed to COVID-19 [2]. Among them, older patients or those with underlying diseases or comorbidities died due to severe conditions during the beginning of COVID-19; however, other cases exhibited initially mild symptoms that gradually worsened, causing death. Therefore, early detection of aggravating factors and symptoms is crucial for timely intervention and treatment.

COVID-19 symptoms include fever, cough, sore throat, nasal congestion, malaise, headache or muscle pain, dehydration, pneumonia, sepsis, and shortness of breath [3]. COVID-19 mortality prediction relies on patient age, blood oxygen saturation, and body temperature [4]. Through a systematic evaluation and external validation of 22 prognostic models for COVID-19 [5], admission oxygen saturation on room air and age have been shown to be strong predictors of clinical deterioration and mortality, respectively, in adults hospitalized with COVID-19.

The electronic health records (EHR) system contains detailed information about symptoms, problems, and services or care provided to patients with COVID-19. Previous studies have used EHR data to predict the COVID-19 prognosis [6-8]. However, nursing documentation records, which account for a substantial portion of EHR data, are underused for research owing to their low quality of documentation [9,10]. Unlike other EHR data, nursing documentation lacks standardization than other data such as diagnoses, laboratory, and medication data. The common data model (CDM) of the Observational Medical Outcome Partnership (OMOP) is designed to standardize the structure and content of observational data from multiple sites for efficient analyses that can produce reliable evidence; it is rarely used to standardize nursing documentation. One study [11] standardized 6277 nursing statements data using Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and converted them into the OMOP CDM to develop a fall-prediction model along with other CDM data. SNOMED CT is the most comprehensive multilingual clinical terminology [12,13] used by more than 40 countries and 30,000 individuals or organizations. The SNOMED CT International Edition, published on April 30, 2023, includes 360,942 concepts, 1,595,980 descriptions (synonyms) of concepts, and 3,261,032 relationships between concepts. Despite these advancements, there is a lack of empirical studies analyzing patient outcomes using standardized nursing documentation or records with SNOMED CT.

The American Nurses Association in 2015 and 2018 position statements also recommended the use of SNOMED CT to code nursing assessments or outcomes to facilitate the interoperability of nursing data. Several studies have examined the content coverage of SNOMED CT in the nursing domain by mapping standardized nursing terminologies, such as the Nursing Intervention Classification, International Classification for Nursing Practice (ICNP), Clinical Care Classification System, North American Nursing Diagnosis Association-International, and Omaha system, into SNOMED CT [14-18]. Thoroddsen et al [19] described the nursing care of patients with COVID-19 using ICNP and SNOMED CT. SNOMED CT outperformed ICNP in representing COVID-19 diagnoses and interventions. In particular, SNOMED CT comprehensively covered nursing interventions for patients with COVID-19. However, prior research only evaluated its usability or content coverage in the nursing domain, with limited analysis of specific phenomena or problems. Therefore, analyzing specific phenomena or problems using standardized nursing documentation data with SNOMED CT is necessary.

This study aimed to (1) standardize the nursing documentation records of patients with COVID-19 using SNOMED CT and (2) identify predictive factors of clinical deterioration in patients with COVID-19 using standardized nursing records.

Methods

Study Design and Setting

This retrospective observational study used data extracted from a clinical data warehouse (CDW) at a tertiary hospital with 1782 beds in Seoul, South Korea. The hospital operated 2 general wards and 4 disaster intensive care units (DICU) to treat COVID-19.

Outcomes

In this study, the clinical outcome is the patient's clinical deterioration. We defined clinical deterioration as an event in which patients with COVID-19 were transferred to the ICU from general wards.

Study Participants

In total, 460 patients with COVID-19 were admitted to 2 general wards for infectious disease treatment at the hospital between December 2019 and June 2022. Patients transferred from or to other general wards or ICUs other than the DICU were excluded. Among the 460 patients, 19 patients were transferred to the DICU because of their deteriorating clinical condition. The mean length of stay (LOS) in the general ward of the 441 patients who were not transferred to the DICU during hospitalization was 8.71 days. The mean and maximum LOS of the 19 patients transferred to the DICU were 3.11 and 7 days, respectively.

XSL•FO
RenderX

Considering LOS differences between the 2 groups, the final analysis of the study included 226 admitted patients with COVID-19 at 2 general wards who were discharged or transferred to the DICU within 7 days between December 2019 and June 2022. Among these, 207 patients who were discharged in a stable clinical condition from the hospital within 7 days were assigned to the control group. A total of 19 patients who were transferred to the DICU within 7 days because of their deteriorating clinical conditions were assigned to the case group.

Data Sources

Using CDW at a study hospital, the general characteristics of patients including age, gender, and initial respiratory symptoms were extracted from initial nursing assessment records. The patient's acuity level and nursing statements were extracted from the nursing records. A total of 57,558 nursing statements from 226 patients were extracted. Among these, 45,852 nursing statements were obtained from the nursing records of 207 patients in the control group. Of the remaining records, 11,706 nursing statements were obtained from 19 patients in the case group.

Mapping to International Standard Terminology, SNOMED CT

Nursing statements were standardized by mapping them to SNOMED CT concepts. The 57,558 nursing statements consisted of 1776 unique nursing statements which were interface terminologies developed by Seoul National University Hospital. Among the 1776 unique nursing statements, 260 unique nursing statements which were recorded 50,867 in total (cumulative recorded rate of 90%) were selected as the mapping source. Among these 260 unique statements, the control group used 237 nursing statements, and the case group used 204 nursing statements.

According to the mapping guide in the previous study [20], 2 experts with more than 5 years of SNOMED CT mapping experience performed the mapping. The scope of the map was defined to restrict mapping to the precoordinated SNOMED CT concept only. Korean nursing statements were translated into English, and search terms were chosen considering the clinical context. SNOMED CT concepts are organized into 19 top-level concepts or hierarchies under the root concept. Among them, the Clinical finding (finding) hierarchy contains concepts related to symptoms and disorders. The procedure hierarchy contains concepts related to activities performed in the provision of health care and regime or therapy. The Situation with explicit context hierarchy includes concepts that specifically define the context information of a clinical finding or procedure [21]. Therefore, the nursing statements in nursing problems or outcomes and diagnosis domains were mapped to concepts of "clinical finding" and "situation with explicit context" hierarchies, and those in the nursing interventions domain were mapped to concepts of "procedure" hierarchy within the SNOMED CT International edition, released on November 30, 2022. If a SNOMED CT concept semantically consistent with the nursing statement could not be found, it was mapped to broader precoordinated concepts.

The mapping results were classified according to the level of correspondence as 3 types: "exact map," "narrow to broad map," and "no map." Mappings were categorized as follows: "exact map" when the meaning of the nursing statement matched an equivalent SNOMED CT concept, "narrow to broad map" when the meaning of the nursing statement matched a broader SNOMED CT concept, and "no map" when there was no broad match concept for the meaning of the nursing statement.

Internal validation of the final map was conducted by calculating interrater reliability. Two experts reviewed the mapping results. If the mapper and reviewer selected the same result, the map was deemed correct. If the maps differed, the results were evaluated via a group discussion, and one of them was selected.

Statistical Analyses

The general characteristics of patients (eg, gender, age, and patient acuity level) in both the case and control groups were explored. For patients in the case group transferred to the DICU (average ward LOS, 3.1 days), nursing statements documented up to the third day were extracted. The total number of nursing statements was divided by the LOS to calculate the mean number of nursing statements per day. Thereafter, the data set was separated according to the top-level hierarchies of SNOMED CT mapped to nursing statements, "clinical finding" or "situation with explicit context," and "procedure." Among the data set, we excluded SNOMED CT concepts that do not describe the patient's clinical problems, such as no breathlessness, free of symptoms, no sputum, and no cough. To identify key features of nursing statements associated with the exacerbation of patient condition, random forest algorithms were used, and optimal hyperparameters were selected with 2 data sets ("clinical finding" or "situation with explicit context" and "procedure").

To assess the effects of the features identified by random forest algorithms on the exacerbation of patient conditions, we performed logistic regression. Patient DICU transfer status was the dependent variable, and the top 5 features from each hierarchy identified by random forest algorithms were independent variables at a significance level of α=.05. Considering the small sample size for binary logistic regression analysis, the model was fitted using a modified estimation procedure known as Firth correction [22]. R (version 4.2.2; R Foundation for Statistical Computing) was used for all analyses.

Ethical Considerations

This study was approved by the Institutional Review Board of the Seoul National University Hospital (H-2207-097-1341). The requirement for informed consent was waived according to the relevant guidelines and regulations of the institutional review board. Identifiers, such as the patient's ID and name, were encrypted so that individuals could not be identified during the data analysis. Participant compensation was not offered since this study was a retrospective observational study.

XSL•FO
RenderX

Results

SNOMED CT Mapping Results

The results of mapping nursing statements to SNOMED CT are presented in Table 1. The interrater reliability of the mapping results was excellent (87.7%). Of the 260 unique nursing statements, 157 were on nursing problems or outcomes, 16 were on nursing diagnoses, and 87 were on nursing intervention statements. A total of 138 (87.8%) of nursing problems or outcomes statements were mapped to concepts within "clinical finding" and "situation with explicit context" hierarchies on SNOMED CT. Through a validation process, 19 nursing problems and outcome statements were mapped into concepts within the procedure hierarchy. All 16 nursing diagnosis statements were mapped to clinical finding concepts. Among the 87 nursing intervention statements, 85 (97.7%) nursing intervention statements were mapped to concepts within the procedure hierarchy. Only 2 statements were mapped to the clinical finding concepts, "267038008 |Edema|" and "29658002 |Oxygen supply absent|." Of the 260 nursing statements, 244 (93.8%) nursing statements were classified as "exact map" and 16 (6.2%) nursing statements were classified as "broad map." The mapped concepts with high frequency are presented in Multimedia Appendix 1.

Table 1. Results of mapping nursing statements to Systematized Nomenclature of Medicine–Clinical Terms.

Type of statements and top-level hierarchy	Concepts, n[a]
Nursing problems or outcomes (n=157 statements)	
Clinical finding	94
Situation with explicit context	22
Procedure	17
Nursing diagnoses (n=16 statements)	
Clinical finding	15
Nursing interventions (n=87 statements)	
Procedure	64
Clinical finding	2
Total (n=260 statements)	
Clinical finding	109
Procedure	73
Situation with explicit context	22

[a]The number of concepts was calculated after removing duplicated concepts.

Clinical Characteristics

The clinical characteristics of the study participants are presented in Table 2. The mean age was 55.9 (SD 17.3) years, and the mean LOS was 5.3 (SD 2.0) days. Patients in the control group were hospitalized in the ward for approximately 5.04 (SD 1.6) days. Patients in the case group were admitted to the ward for approximately 3.1 (SD 1.9) days. Overall, 119 (52.7%) patients were male. A total of 197 (86.7%) patients had a patient acuity level of 4, and 26 (11.5%) patients had a patient acuity level of 3. Additionally, 59 (26.1%) patients had a cough, 47 (20.8%) patients had sputum, and 26 (11.5%) patients had dyspnea as their initial symptoms.

XSL•FO
RenderX

Table 2. Clinical characteristics of study participants.

Variables	Control group (n=207)	Case group (n=19)	Total (n=226)	Chi-square (*df*) or *t* test (*df*)	*P* value
Age (years), mean (SD)	55.8 (17.3)	57.0 (18.6)	55.9 (17.3)	–0.28 (224)[a]	.78
Age (years), n (%)				3.7 (5)[b]	.60
<30	14 (6.8)	1 (5.3)	15 (6.6)		
30-39	30 (14.5)	4 (21.1)	34 (15)		
40-49	30 (14.5)	0 (0)	30 (13.3)		
50-59	38 (18.4)	4 (21.1)	42 (18.6)		
60-69	42 (20.3)	5 (26.3)	47 (20.8)		
>70	53 (25.6)	5 (26.3)	58 (25.7)		
Sex, n (%)				0.1 (1)[b]	.81
Male	110 (53.1)	9 (47.4)	119 (52.7)		
Female	97 (46.9)	10 (52.6)	107 (47.3)		
LOS[c], mean (SD)					
LOS ward	5.04 (1.6)	3.1 (1.9)	4.9 (1.7)	5.06 (224)[a]	<.001
LOS DICU[d]	0.0 (0)	5 (2.9)	0.4 (1.6)	–7.50 (18)[a]	<.001
LOS total	5.04 (1.6)	8.1 (3.2)	5.3 (2)	–4.11 (18.882)[a]	<.001
Patient acuity level, n (%)				8.4 (2)[b]	.01
Level 1	0 (0)	0 (0)	0 (0)		
Level 2	4 (1.9)	0 (0)	4 (1.8)		
Level 3	20 (9.7)	6 (31.6)	26 (11.5)		
Level 4	183 (88.4)	13 (68.4)	196 (86.7)		
Initial respiratory symptoms, n (%)					
Cough	53 (25.6)	6 (31.6)	59 (26.1)	0.1 (1)[b]	.76
Sputum	40 (19.3)	7 (36.8)	47 (20.8)	2.3 (1)[b]	.13
Dyspnea	19 (9.2)	7 (36.8)	26 (11.5)	10.5 (1)[b]	<.001
Other	6 (2.9)	0 (0)	6 (2.7)	0.0 (1)[b]	.99
None	144 (55.1)	11 (57.9)	71 (31.4)	0.6 (1)[b]	.43
Nursing statements					
Frequency in total, n	45,852	11,706	57,558	N/A[e]	N/A
Frequency per day, per patient; mean (SD)	42.73 (14.60)	55.48 (22.01)	43.80 (15.70)	–2.48 (19.48)[a]	.023

[a]*t* test.
[b]Chi-square.
[c]LOS: length of stay.
[d]DICU: disaster intensive care units.
[e]N/A: not applicable.

Feature Selection via Random Forest

The feature importance calculated using the random forest method is shown in Figure 1. The most important feature in "clinical finding" showed "449171008 |Oxygen saturation below reference range (finding)|." Prominent concepts that express respiratory issues included "267036007 |Dyspnea (finding)|," "271823003 |Tachypnea (finding)|," and "49727002 |Cough (finding)|."

XSL•FO
RenderX

Figure 1. Top 30 features extracted via random forest (left: clinical finding or situation with explicit context concepts; right: procedure concepts).

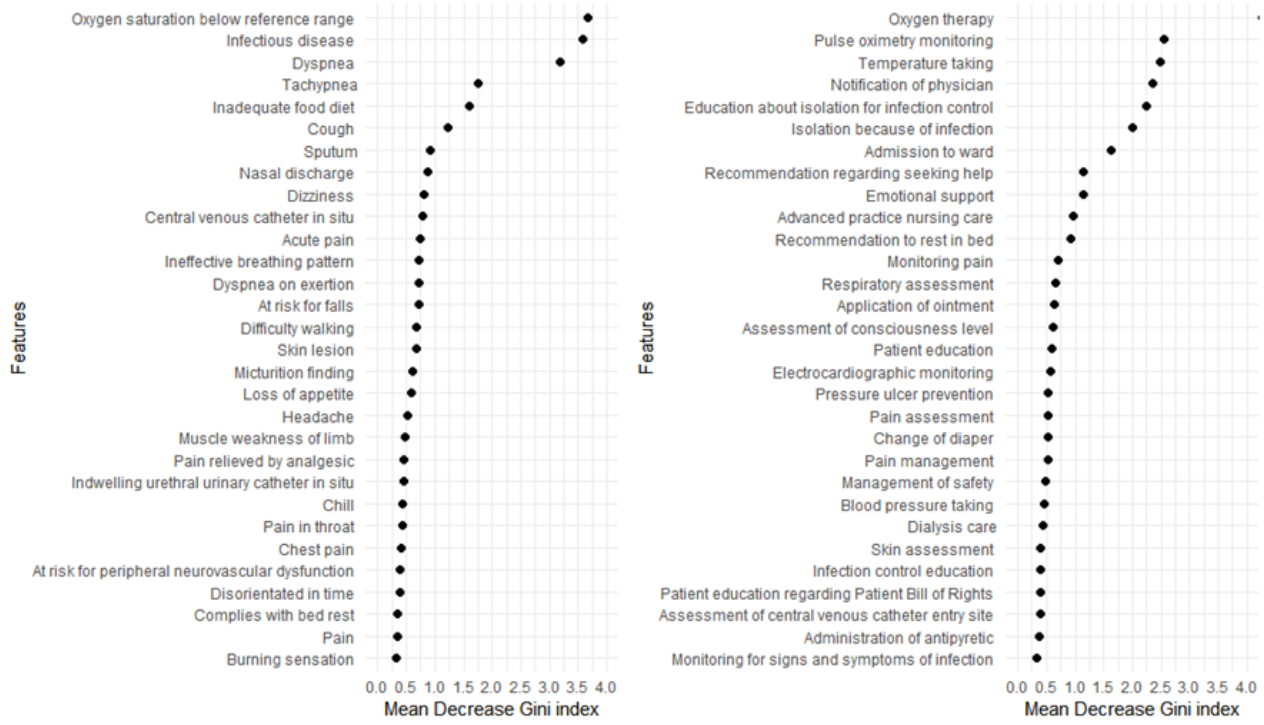

Among procedure concepts, “57485005 |Oxygen therapy|” was the most important feature, followed by “284034009 |Pulse oximetry monitoring|,” “56342008 |Temperature taking (procedure)|,” “428426009 |Notification of physician|,” and “737612005 |Education about isolation for infection control|.”

Association Between Patient’s DICU Transfer and Selected Features

Table 3 presents the results of logistic regression analysis, which explored the association between patient DICU transfer status and features in both “clinical finding” or “situation with explicit context” and “procedure” selected via random forest. In “clinical finding,” patients with dyspnea and inadequate food diet were more likely to experience DICU transfer (dyspnea: odds ratio [OR] 5.99, 95% CI 2.25-20.29; inadequate food diet: OR 10.0, 95% CI 2.71-40.84). Patients with infectious disease in nursing records were less likely to be associated with the DICU transfer in patients with COVID-19 (OR 0.32, 95% CI 0.16-0.66). In the procedure, patients with “oxygen therapy” and “notification of physician” were more likely to experience DICU transfer (oxygen therapy: OR 1.89, 95% CI 1.25-3.05; notification of physician: OR 1.72, 95% CI 1.02-2.97). Patients with “temperature taking” were less likely to experience DICU transfer (OR 0.36, 95% CI 0.18-0.70).

Table 3. Results of logistic regression.

SNOMED CT[a] top-level hierarchy and SNOMED CT code	SNOMED CT fully specified name	Odds ratio (95% CI)	*P* values
Clinical finding			
449171008	Oxygen saturation below reference range	4.21 (0.88-18.78)	.07
40733004	Infectious disease	0.32 (0.16-0.66)	.003
267036007	Dyspnea	5.99 (2.25-20.29)	<.001
271823003	Tachypnea	2.22 (0.34-44.75)	.44
102610002	Inadequate food diet	10.0 (2.71-40.84)	.001
Procedure			
57485005	Oxygen therapy	1.89 (1.25-3.05)	.002
284034009	Pulse oximetry monitoring	0.84 (0.46-1.51)	.55
56342008	Temperature taking	0.36 (0.18-0.70)	.002
428426009	Notification of physician	1.72 (1.02-2.97)	.04
737612005	Education about isolation for infection control	1.26 (0.76-2.02)	.35

[a]SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms.

XSL•FO
RenderX

Discussion

Principal Results

This study standardized the nursing statements of patients with COVID-19 in a hospital in South Korea by mapping them to SNOMED CT, an international standard terminology. Using these standardized statements, machine learning analysis with a random forest was conducted. Concepts that could predict patient deterioration in nursing problems or outcomes and intervention domains were identified. In the nursing problems or outcomes domain, concepts in nursing statements related to dyspnea, tachypnea, and oxygen saturation below the reference range were associated with patient deterioration. In nursing interventions, concepts in nursing statements related to respiratory intervention, including oxygen therapy, pulse oximetry monitoring, and physician notification, were associated with DICU transfer. These concepts were different from the findings of another study that analyzed nursing statements of patients with ovarian cancer. Kim et al [23] analyzed standardized nursing statements of patients who underwent curative surgery for epithelial ovarian cancer, and urination, food supply, bowel mobility, and pain were identified as the most common concepts.

To the best of our knowledge, this study is the first to identify predictive factors from nursing records standardized using SNOMED CT. Although previous studies [24-29] have used standardized nursing records such as the North American Nursing Diagnosis Association-International, Nursing Intervention Classification, Nursing Outcomes Classification, and ICNP to predict the patient's condition or explore the nursing care provided, no study has predicted patient clinical deterioration using SNOMED CT–standardized nursing records. To achieve interoperable health information exchange, the Centers for Medicare and Medicaid Services and Office of the National Coordinator for Health Information Technology recommended the use of SNOMED CT and Logical Observation Identifier Names and Codes (LOINC) as reference terminologies. In 2020, the Office of the National Coordinator for Health Information Technology introduced the United States Core Data for Interoperability, a standardized set of health data classes and constituent data elements that recommended the use of international standard terminologies, including SNOMED CT and LOINC [30]. Consequently, standardization of nursing records using reference terminology before conducting data analysis is required in the nursing field.

Our study comprehensively standardized all nursing statements in study participants using SNOMED CT concepts. Several previous studies have evaluated SNOMED CT coverage in various health care domains, such as frailty and wound care [31-34]. In nursing, efforts have been made to map ICNP 7-axis concepts, nursing problem lists, and ICU nursing flowsheets to SNOMED CT [16,32,35-37], which provides the most detailed semantic expression compared with other standard terminologies [19]. Notably, this study unexpectedly mapped 19 nursing statements on nursing assessments or outcomes to the "procedure" hierarchy, despite the lack of corresponding concepts in the "clinical finding" or "situation with explicit context" hierarchies. This was due to the fact that the problem or phenomenon was the result of specific nursing procedures or practices. For example, the statement "a patient is applying a mattress to prevent bedsores" in nursing assessment or outcome was mapped to concept "733920005 |Provision of pressure-relieving mattress|" in the procedure hierarchy.

The identified features affecting patient deterioration (DICU transfer) in the hierarchy of "clinical finding" included respiratory issues such as low oxygen saturation, dyspnea, and tachypnea. Izquierdo et al [7] reported patient signs and symptoms, especially tachypnea, to be reliable predictors of DICU admission, which is consistent with our results. In addition, inadequate diet was significantly associated with DICU transfer, potentially due to the inability of patients to eat independently owing to their deteriorated condition. These findings highlighted the importance of signs and symptoms in nursing records for predicting a patient's deteriorating clinical condition. Alternatively, nursing statements related to "infectious disease" within "clinical finding" and "temperature taking" within "procedure" decreased the risk of clinical deterioration in patients with COVID-19. Since patients with COVID-19 were included as study participants, it was valid that the concept of "infectious disease" appeared frequently in nursing statements, but repeated studies are needed to explain the finding that the more statements related to this concept, the lower the risk of clinical deterioration. Considering previous studies [4,38] reporting "fever" as a significant factor affecting deterioration in patients with COVID-19, our study revealed that increased nursing activity of "temperature taking" notably reduced patient deterioration risk. This suggested that nurses may have preemptively frequently monitored temperature to prevent fever; however, to accurately interpret this result, analyzing "body temperature" data is necessary. The features identified in this study highlight the potential of nursing records as a valuable real time predictor of a patient's clinical condition.

Limitations

This study had some limitations. First, the analyzed data were extracted from a single tertiary university hospital, which limits the generalizability of this study. Second, sample bias is possible because patients admitted for more than 7 days to the ward were excluded. Third, other clinical findings or nursing interventions may not have been documented as nursing statements in the nursing records. Considering that this study extracted only structured nursing statements stored in the CDW, records written in free text were not included in the analysis. Fourth, as SNOMED CT concepts were used in the analysis, the results may vary depending on the hierarchical structure or level of the mapped SNOMED CT concept. For example, concepts such as "chest pain," "headache," and "pain" derived through random forest are connected in a hierarchical structure where "chest pain" and "headache" are subcategories of the "pain" concept. If "chest pain" and "headache" were grouped together as "pain," the effect of "pain" on clinical deterioration might have been significant.

Conclusions

This study showed that standardized nursing records are an important source of data that can be used to predict clinical

XSL•FO
RenderX

deterioration in patients with COVID-19. In total, 260 nursing statements were mapped to the SNOMED CT, including 109 concepts in the clinical finding hierarchy, 73 concepts in the procedure hierarchy, and 22 concepts in "situation" with an explicit context hierarchy. Among the standardized nursing statements, key clinical findings were respiratory issues, including low oxygen saturation, dyspnea, and tachypnea. The primary procedure-related features included oxygen therapy, pulse oximetry, and temperature monitoring. In specific, low oxygen saturation, dyspnea, tachypnea, and oxygen therapy are associated with the risk of clinical deterioration in patients with COVID-19. This study validates the use of nursing records as variables for predicting the deterioration of patients with COVID-19. Future research should investigate the integration of standardized nursing records with diagnoses, laboratory, and medication data to develop a highly reliable predictive model.

Acknowledgments

This study was supported by an Institute of Information and Communications Technology Planning and Evaluation grant funded by the Korean government (MSIT; 2021-0-00312, "Development of Non-Face-to-face Patient Infection Activity Prediction and Protection Management SW Technology at Home and Community Treatment Centers for Effective Response to Infectious Disease") and the National Research Foundation of Korea funded by the Ministry of Science and Information Communication Technology (NRF-2021R1A2C1091261).

Data Availability

The data analyzed during this study are not readily available because of hospital regulation restrictions and patient privacy concerns. Requests to access the data sets should be directed to the corresponding author.

Authors' Contributions

SS and HJ conceptualized the study and reviewed and edited the manuscript. SS, SHK, and HJ developed the study methodology and drafted the manuscript. SS, SHK, YK, and HJ conducted formal analyses. HJ supervised the study and acquired funding. All the authors reviewed the results and approved the final version of the manuscript. Generative AI was not used in any part of the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The mapped concepts with high frequency.
[DOCX File , 18 KB - jmir_v26i1e53343_app1.docx]

References

1. WHO Coronavirus (COVID-19) dashboard. World Health Organization. 2023. URL: https://covid19.who.int/ [accessed 2023-09-27]
2. Coronavirus disease 19 (COVID-19) occurrence status. Korea Disease Control and Prevention Agency. URL: https://ncov.kdca.go.kr/bdBoardList_Real.do [accessed 2023-09-27]
3. Clinical management of severe acute respiratory infection when novel coronavirus (2019-nCoV) infection is suspected: interim guidance. World Health Organization. 2020. URL: https://iris.who.int/handle/10665/330893 [accessed 2023-09-27]
4. Yadaw AS, Li YC, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health 2020;2(10):e516-e525 [FREE Full text] [doi: 10.1016/S2589-7500(20)30217-X] [Medline: 32984797]
5. Gupta RK, Marks M, Samuels THA, Luintel A, Rampling T, Chowdhury H, et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. Eur Respir J 2020;56(6):2003498 [FREE Full text] [doi: 10.1183/13993003.03498-2020] [Medline: 32978307]
6. Campbell TW, Wilson MP, Roder H, MaWhinney S, Georgantas RW, Maguire LK, et al. Predicting prognosis in COVID-19 patients using machine learning and readily available clinical data. Int J Med Inform 2021;155:104594 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104594] [Medline: 34601240]
7. Izquierdo JL, Ancochea J, Savana COVID-19 Research Group, Soriano JB. Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. J Med Internet Res 2020;22(10):e21801 [FREE Full text] [doi: 10.2196/21801] [Medline: 33090964]
8. Gong K, Wu D, Arru CD, Homayounieh F, Neumark N, Guan J, et al. A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records. Eur J Radiol 2021;139:109583 [FREE Full text] [doi: 10.1016/j.ejrad.2021.109583] [Medline: 33846041]

XSL•FO
RenderX

9. Wang N, Hailey D, Yu P. Quality of nursing documentation and approaches to its evaluation: a mixed-method systematic review. J Adv Nurs 2011;67(9):1858-1875. [doi: 10.1111/j.1365-2648.2011.05634.x] [Medline: 21466578]
10. Fennelly O, Grogan L, Reed A, Hardiker NR. Use of standardized terminologies in clinical practice: a scoping review. Int J Med Inform 2021;149:104431 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104431] [Medline: 33713915]
11. Jung H, Yoo S, Kim S, Heo E, Kim B, Lee HY, et al. Patient-level fall risk prediction using the observational medical outcomes partnership's common data model: pilot feasibility study. JMIR Med Inform 2022;10(3):e35104 [FREE Full text] [doi: 10.2196/35104] [Medline: 35275076]
12. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
13. Park HA, Yu SJ, Jung H. Strategies for adopting and implementing SNOMED CT in Korea. Healthc Inform Res 2021;27(1):3-10 [FREE Full text] [doi: 10.4258/hir.2021.27.1.3] [Medline: 33611871]
14. Park HT, Lu DF, Konicek D, Delaney C. Nursing interventions classification in systematized nomenclature of medicine clinical terms: a cross-mapping validation. Comput Inform Nurs 2007;25(4):198-208; quiz 209. [doi: 10.1097/01.NCN.0000280590.35690.7d] [Medline: 17625400]
15. Park HA, Lundberg C, Coenen A, Konicek D. Evaluation of the content coverage of SNOMED CT representing ICNP seven-axis version 1 concepts. Methods Inf Med 2011;50(5):472-478. [doi: 10.3414/ME11-01-0004] [Medline: 21947166]
16. Kim TY, Coenen A, Hardiker N. Semantic mappings and locality of nursing diagnostic concepts in UMLS. J Biomed Inform 2012;45(1):93-100 [FREE Full text] [doi: 10.1016/j.jbi.2011.09.002] [Medline: 21951759]
17. Monsen KA, Finn RS, Fleming TE, Garner EJ, LaValla AJ, Riemer JG. Rigor in electronic health record knowledge representation: lessons learned from a SNOMED CT clinical content encoding exercise. Inform Health Soc Care 2016;41(2):97-111. [doi: 10.3109/17538157.2014.965302] [Medline: 25325887]
18. Kim J, Yao Y, Macieira TGR, Keenan G. An examination of the coverage of the SNOMED CT coded nursing problem list subset. JAMIA Open 2019;2(3):386-391 [FREE Full text] [doi: 10.1093/jamiaopen/ooz023] [Medline: 31984371]
19. Thoroddsen A, Rúnarsdóttir ER, Örlygsdóttir B. Description of COVID-19 patients and mapping nursing data to ICNP 2021 reference set in SNOMED CT. Int Nurs Rev 2023;70(1):28-33 [FREE Full text] [doi: 10.1111/inr.12824] [Medline: 36580398]
20. Sung S, Park HA, Jung H, Kang H. A SNOMED CT mapping guideline for the local terms used to document clinical findings and procedures in electronic medical records in South Korea: methodological study. JMIR Med Inform 2023;11:e46127 [FREE Full text] [doi: 10.2196/46127] [Medline: 37071456]
21. SNOMED CT editorial guide. SNOMED CT International. 2023. URL: https://confluence.ihtsdotools.org/display/DOCEG/SNOMED+CT+Editorial+Guide [accessed 2023-10-25]
22. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016;16(1):163 [FREE Full text] [doi: 10.1186/s12874-016-0267-3] [Medline: 27881078]
23. Kim K, Han Y, Jeong S, Doh K, Park HA, Lee K, et al. Prediction of postoperative length of hospital stay based on differences in nursing narratives in elderly patients with epithelial ovarian cancer. Methods Inf Med 2019;58(6):222-228. [doi: 10.1055/s-0040-1705122] [Medline: 32349156]
24. Azzolin K, Mussi CM, Ruschel KB, de Souza EN, de Fátima Lucena A, Rabelo-Silva ER. Effectiveness of nursing interventions in heart failure patients in home care using NANDA-I, NIC, and NOC. Appl Nurs Res 2013;26(4):239-244. [doi: 10.1016/j.apnr.2013.08.003] [Medline: 24050914]
25. Escalada-Hernández P, Muñoz-Hermoso P, González-Fraile E, Santos B, González-Vargas JA, Feria-Raposo I, et al. A retrospective study of nursing diagnoses, outcomes, and interventions for patients with mental disorders. Appl Nurs Res 2015;28(2):92-98. [doi: 10.1016/j.apnr.2014.05.006] [Medline: 25002344]
26. da Silva Thomé E, Centena RC, da Silva Behenck A, Marini M, Heldt E. Applicability of the NANDA-I and nursing interventions classification taxonomies to mental health nursing practice. Int J Nurs Knowl 2014;25(3):168-172. [doi: 10.1111/2047-3095.12033] [Medline: 24758497]
27. Yang MJ, Kim HY, Ko E, Kim HK. Identification of nursing diagnosis-outcome-intervention linkages for inpatients in the obstetrics department nursing unit in South Korea. Int J Nurs Knowl 2019;30(1):12-20. [doi: 10.1111/2047-3095.12187] [Medline: 28884935]
28. Palomar-Aumatell X, Subirana-Casacuberta M, Mila-Villarroel R. Critical care nursing interventions and the time required for their completion in intensive care units: a Delphi study. Intensive Crit Care Nurs 2017;43:87-93. [doi: 10.1016/j.iccn.2017.05.001] [Medline: 28781007]
29. Sung S, Jung H, Kim Y. Exploring nursing care for patients with COVID-19 using international classification for nursing practice-based nursing records. Comput Inform Nurs 2024;42(2):127-135 [FREE Full text] [doi: 10.1097/CIN.0000000000001048] [Medline: 37579774]
30. United States Core Data for Interoperability (USCDI). Office of the National Coordinator for Health Information Technology. 2023. URL: https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi#uscdi-v1 [accessed 2023-09-27]
31. Thandi M, Brown S, Wong ST. Mapping frailty concepts to SNOMED CT. Int J Med Inform 2021;149:104409. [doi: 10.1016/j.ijmedinf.2021.104409] [Medline: 33677397]

XSL•FO
RenderX

32. Block LJ, Wong ST, Handfield S, Hart R, Currie LM. Comparison of terminology mapping methods for nursing wound care knowledge representation. Int J Med Inform 2021;153:104539. [doi: 10.1016/j.ijmedinf.2021.104539] [Medline: 34358804]
33. Hüsers J, Przysucha M, Esdar M, John SM, Hübner UH. Expressiveness of an international semantic standard for wound care: mapping a standardized item set for leg ulcers to the systematized nomenclature of medicine-clinical terms. JMIR Med Inform 2021;9(10):e31980 [FREE Full text] [doi: 10.2196/31980] [Medline: 34428171]
34. Block L, Handfield S. Mapping wound assessment data elements in SNOMED CT. In: Sermeus W, Procter PM, Weber P, editors. Nursing Informatics 2016: eHealth for All: Every Level Collaboration—From Project to Realization. Amsterdam, Netherlands: IOS Press; 2016:1078-1079.
35. Matney SA, Warren JJ, Evans JL, Kim TY, Coenen A, Auld VA. Development of the nursing problem list subset of SNOMED CT®. J Biomed Inform 2012;45(4):683-688 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.003] [Medline: 22202620]
36. So EY, Park HA. Exploring the possibility of information sharing between the medical and nursing domains by mapping medical records to SNOMED CT and ICNP. Healthc Inform Res 2011;17(3):156-161 [FREE Full text] [doi: 10.4258/hir.2011.17.3.156] [Medline: 22084810]
37. Park HA, Lundberg CB, Coenen A, Konicek DJ. Evaluation of the content coverage of SNOMED-CT to represent ICNP version 1 catalogues. In: Saranto K, Brennan PF, Park HA, Tallberg M, Ensio A, editors. Connecting Health and Humans. Amsterdam, Netherland: IOS Press; 2009:303-307.
38. Chen SL, Feng HY, Xu H, Huang SS, Sun JF, Zhou L, et al. Patterns of deterioration in moderate patients with COVID-19 from jan 2020 to mar 2020: a multi-center, retrospective cohort study in China. Front Med (Lausanne) 2020;7:567296 [FREE Full text] [doi: 10.3389/fmed.2020.567296] [Medline: 33344469]

Abbreviations

CDM: common data model
CDW: clinical data warehouse
DICU: disaster intensive care unit
EHR: electronic health record
ICNP: International Classification for Nursing Practice
LOINC: Logical Observation Identifier Names and Codes
LOS: length of stay
NANDA-I: North American Nursing Diagnosis Association–International
NIC: Nursing Intervention Classification
OMOP: Observational Medical Outcome Partnership
ONC: Office of the National Coordinator for Health Information Technology
OR: odds ratio
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms

Edited by T de Azevedo Cardoso; submitted 03.10.23; peer-reviewed by J Lee, N Hardiker, D Liu; comments to author 19.10.23; revised version received 26.10.23; accepted 27.02.24; published 29.03.24.

Please cite as:
Sung S, Kim Y, Kim SH, Jung H
Identification of Predictors for Clinical Deterioration in Patients With COVID-19 via Electronic Nursing Records: Retrospective Observational Study
J Med Internet Res 2024;26:e53343
URL: https://www.jmir.org/2024/1/e53343
doi:10.2196/53343
PMID:38414056

©Sumi Sung, Youlim Kim, Su Hwan Kim, Hyesil Jung. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding the Impact of Communicating Uncertainty About COVID-19 in the News: Randomized Between-Subjects Factorial Experiment

Rui Zhao[1], PhD; Xuerong Lu[2], PhD; Jiayi Yang[3], PhD; Biao Li[4], PhD

[1]School of Journalism and Communication, Minzu University of China, Beijing, China
[2]School of Communication, Oregon State University, Corvallis, OR, United States
[3]School of Chinese Language and Literature, University of International Business and Economics, Beijing, China
[4]School of Journalism, Renmin University of China, Beijing, China

Corresponding Author:
Biao Li, PhD
School of Journalism
Renmin University of China
19 Zhong Guancun Street
Beijing, 100081
China
Phone: 86 18810386586
Email: libiao@ruc.edu.cn

Abstract

Background: Whether and how the uncertainty about a public health crisis should be communicated to the general public have been important and yet unanswered questions arising over the past few years. As the most threatening contemporary public health crisis, the COVID-19 pandemic has renewed interest in these unresolved issues by both academic scholars and public health practitioners.

Objective: The aim of this study was to investigate the impact of communicating uncertainty about COVID-19–related threats and solutions on individuals' risk perceptions and misinformation vulnerability, as well as the sequential impact of these effects on health information processing and preventative behavioral intentions.

Methods: A 2×2 (threat uncertainty [presence vs absence]×solution uncertainty [presence vs absence]) full-fractional between-subjects online experiment was conducted with 371 Chinese adults. Focusing on the discussion of whether the asymptomatic cases detected during the COVID-19 pandemic would further lead to an uncontrolled pandemic, news articles were manipulated in terms of whether the infectiousness of asymptomatic cases and the means to control the transmission are presented in terms of their certainty or uncertainty. Participants were randomly assigned to one of the four experimental conditions, being instructed to read one news article. After reading the news article assigned, participants were asked to respond to a series of questions to assess their cognitive and behavioral responses.

Results: Individuals were more susceptible to believing false COVID-19–related information when a certain threat and uncertain solution were framed in the news article. Moreover, individuals' perceptions of crisis severity increased when exposed to news information containing uncertain solutions. Both misinformation vulnerability and perceived severity were positively associated with information processing. Information seeking was positively associated with protective behavioral intention, whereas information avoidance was negatively associated with protective behavioral intention.

Conclusions: Our findings imply that uncertainty, depending on its aspect, can be effectively communicated to the public during an emerging public health crisis. These results have theoretical and practical implications for health communicators and journalists. Given its limited influence on individuals' cognitive and behavioral responses, uncertainty related to a health threat should be disseminated to meet the public's expectation of information transparency. However, caution is advised when communicating uncertainty related to potential solutions, as this factor exhibited a mixed impact on individual responses during a crisis.

(J Med Internet Res 2024;26:e51910) doi:10.2196/51910

XSL•FO
RenderX

KEYWORDS

information uncertainty; health communication; uncertainty management; COVID-19; public health perception; health information

Introduction

Background

The question of how to communicate uncertainty to the general public has been raked up during the COVID-19 pandemic, which is considered to be the most threatening public health crisis that emerged over the past 10 years, characterized by a high level of uncertainty. Since its outbreak, news coverage of COVID-19 has largely been emphasizing the "unknowns" about the source, infectivity, treatment, prevention, and control measures of the virus [1]. However, whether (or not) uncertainties should be communicated to the general public remains a controversial issue, given the general low tolerance of the public for uncertainty along with a high expectation for information transparency. On the one hand, uncertainty is an undesirable experience in which people fear losing control of their lives, leading to negative consequences [2,3]. On the other hand, uncertainty may also have positive effects, as some scholars suggest that when uncertainty is perceived, people tend to actively seek for information to ease this feeling, and in this process can gain more information and a deeper understanding of the event [4]. Thus, it is important to understand whether the uncertainty presented in news articles influences individuals' cognitive and behavioral intentions during public health emergencies.

In actual news framing, uncertainty does not appear as a monolithic entity, and each new challenge presented by COVID-19 involves different aspects (eg, threats and solutions) with varying degrees of uncertainty [5]. According to the Centers for Disease Control and Prevention, the uncertainties related to threats and solutions are the two greatest concerns among the public during public health emergencies [6]. Therefore, this study focused on the uncertainties related to threats and solutions associated with COVID-19 that the news media might (or might not) communicate to the public, with the goal of exploring how this communication of uncertainty might influence individuals' health behaviors.

Impacts of Uncertainty on Risk Perception and Misinformation Vulnerability

Risk perception is always associated with uncertainty in the public health context. It is assumed that individuals will only begin to manage uncertainty through information processing or preventive behaviors when they perceive a given issue to be associated with a certain level of risk [7]. The perceived risk by individuals includes the severity and susceptibility of a public health crisis. Severity refers to the magnitude of harm caused by the threat, whereas susceptibility refers to the probability of occurrence of a threat to a specific subject [8]. Empirically, Lalot et al [9] found that individuals' uncertainty about how the novel coronavirus would affect people significantly increased their perceived threat of the pandemic. Pine et al [10] found that the partial and changing information would cultivate the uncertainty surrounding COVID-19, which would further influence individuals' risk perception of the pandemic.

Moreover, the potential impact of uncertainty on exacerbating the misinformation effect has been raised as a concern in recent years. Lu et al [11] found that communicating uncertainty about preliminary evidence related to COVID-19 was positively associated with the number of likes and retweets of related misinformation on social media. Consequently, the communication of uncertainty during a pandemic might unexpectedly facilitate engagement with misinformation.

To better understand the impact of threat uncertainty and solution uncertainty that are communicated to the general public during a public health crisis, we established the following research question: How, if at all, does the uncertainty frame of a threat and solution exert main and interaction effects on individuals' risk perceptions and misinformation vulnerability?

Impacts of Uncertainty on Information Processing

Information seeking is regarded as a key communication outcome during a public health crisis, guiding individuals to understand public health issues and consequently adopt appropriate health behaviors (eg, [7,12]). Information seeking refers to individuals' active information-searching activity through human interaction [13], online inquiry [14], and passive observation [15]. Theoretical and empirical evidence suggests that individuals tend to engage in information-seeking behaviors when experiencing psychological discomfort such as confusion and anxiety resulting from exposure to uncertainty (eg, [16,17]). Although the motivations for information-seeking behaviors according to various demographic characteristics such as age, gender, and health status have been extensively examined in the public health context (eg, [12,18]), little is known about how uncertainty communicated by the media and experts influences individuals' information-seeking behavior.

Case et al [19] pointed out that individuals might also engage in information avoidance to reduce feelings of uncertainty. They found that the mental discomfort that arises due to uncertainty, especially in a health context, could facilitate information-avoidance behaviors. Information avoidance refers to an individual's absence from or ignorance of information and its source [4].

Thus, we further aimed to understand whether (or not) and how the uncertainty framed in the news would motivate individuals' information-seeking and information-avoiding behavior differently.

Additionally, the relationship between risk perception and information processing has been documented in the health risk literature. Goodall and Reed [20] suggested that individuals would seek more information when the perceived threat is high. Conversely, Jiang et al [21] found that a higher risk perception of the COVID-19 pandemic would reduce individuals' information-seeking behavior. They also found that individuals' propensity to believe COVID-19 misinformation would also influence their additional information-seeking behavior.

Therefore, we sought to examine how, if at all, uncertainty framed in severity and susceptibility influences information

XSL•FO
RenderX

processing (ie, information seeking and information avoidance) through risk perceptions and misinformation vulnerability.

Information Seeking and Avoidance Influence Preventive Behaviors

In the context of COVID-19, studies on how organizations and individuals perceive risks during crises have centered on the changes in preventive behaviors during infectious disease outbreaks and how these behavioral shifts can be facilitated by engagement in informational behaviors.

Previous studies suggest that information seeking through different channels and sources is positively associated with preventive behaviors during crises (eg, [22,23]). Individuals who engage in more effortful information seeking and processing are more likely to develop risk-related cognitions, attitudes, and behaviors [24]. However, the health information environment in a pandemic is often filled with uncertain information, false claims, or even conspiracy theories [25], which can bias people's pandemic-related beliefs and impede their adoption of effective actions [26].

Nevertheless, this situation does not imply that refraining from active information seeking is a wise choice. By contrast, the impact of information avoidance on preventive behaviors is not less significant than that of information seeking [27]. While information avoidance minimizes the chances of interaction with unnecessary information, it simultaneously diminishes the opportunities to receive relevant information. From a cognitive perspective, individuals have limited capacity to process information, and if not adequately addressed, the outcome can be information overload [28]. Avoiding information acquisition may lead individuals to make preventive decisions based on limited information [29]. Particularly when faced with uncertain information, information avoidance may lead to incorrect preventive behaviors. Therefore, we sought to determine how, if at all, preventative behavioral intentions might be associated with information seeking (1) and information avoidance (2) separately.

Methods

Research Design

This study adopted a controlled experiment approach. Based on the question "*Will asymptomatic cases lead to an uncontrolled epidemic?*" a 2×2 (threat certainty vs uncertainty×solution certainty vs uncertainty) online experiment was designed using asymptomatic cases, an emerging challenge in the COVID-19 epidemic, as a risk scenario. Data were collected in May 2020 through this anonymous online experiment. Participants were recruited from Sojump, which is the largest online survey platform in China. As the context of the experiment, at this time, China was gradually implementing measures to prevent and control the COVID-19 epidemic. However, at the same time, the detection of an increasing number of asymptomatic cases was raising concern. On March 31, 2020, China's National Health Commission announced that as of the following day (April 1), it would disclose the detection, transition, and management of asymptomatic cases in its daily briefings on the epidemic to respond to these societal concerns in a timely manner [30]. The risk threat posed by the emerging challenge and COVID-19 prevention measures were still being explored and discussed at that time and were fraught with uncertainties. Based on this context, we developed 4 simulated online news reports based on real news coverage and expert interview data; the details of the simulations are provided in Multimedia Appendix 1 with a brief summary in Table 1.

Although the content orientation of the simulated news coverage differed across the 4 scenarios, the format, word count, structural design, and information volume of the news coverage remained consistent. Participants were randomly assigned to one of the scenarios and were asked to read the simulated news coverage before completing the questionnaire. A total of 592 people participated in the study online and 371 valid questionnaires were obtained after postchecking, including 88 valid questionnaires for condition 1, 99 for condition 2, 88 for condition 3, and 96 for condition 4.

Figure 1 shows the general flow of participants in the study. Participants were from a wide range of age groups. The sample included a relatively equal sex ratio (with 41.80% of the sample identifying as female), and the majority of the participants had an education level of college degree or above (73%). Table 2 summarizes the main sociodemographic characteristics of the sample.

Table 1. News article exposure simulation conditions and core messages.

Condition	Core message in simulated news article
1: Threat certainty × Solution certainty	The infectiousness of asymptomatic cases is limited and there are adequate guarantees to control the transmission range of asymptomatic cases
2: Threat certainty × Solution uncertainty	The infectiousness of asymptomatic cases is limited, but the means to control the transmission range of asymptomatic cases are uncertain
3: Threat uncertainty × Solution certainty	The infectiousness of asymptomatic cases is uncertain, but there are adequate guarantees to control the transmission range of asymptomatic cases
4: Threat uncertainty × Solution uncertainty	The infectiousness of asymptomatic cases is uncertain and the means to control the transmission range of asymptomatic cases are uncertain

XSL•FO
RenderX

Figure 1. Study flow diagram.

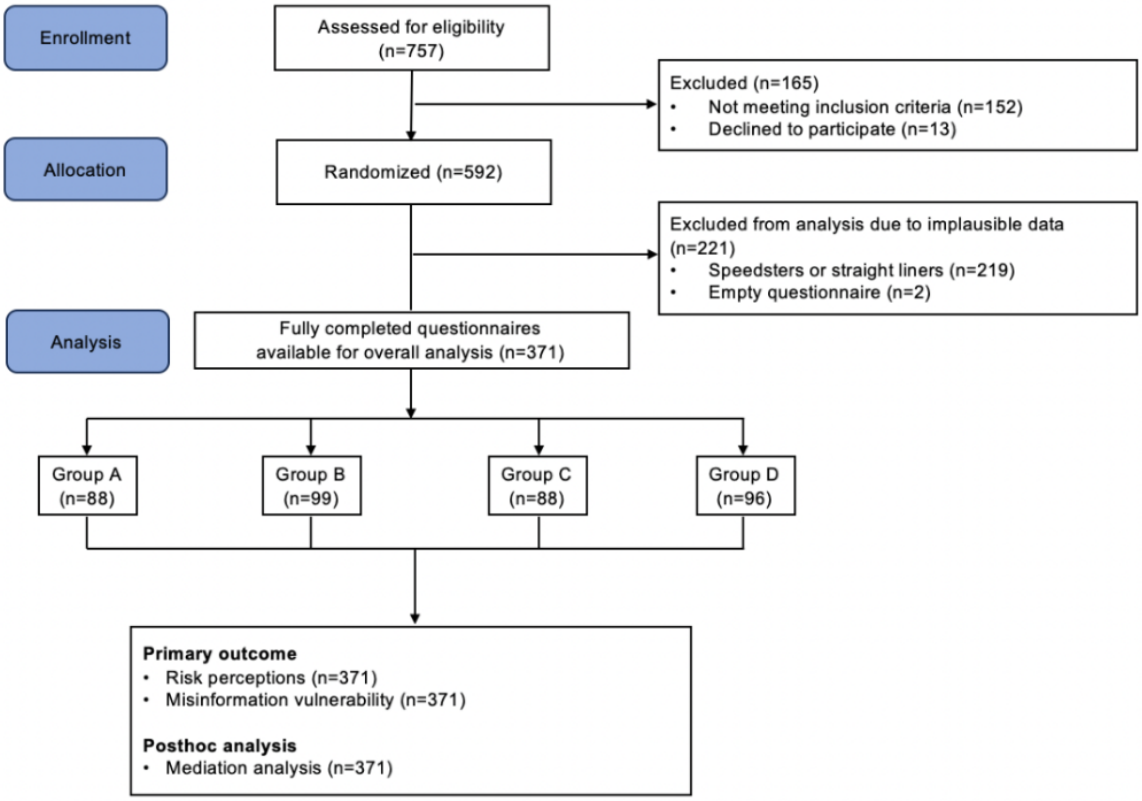

Table 2. Sociodemographic characteristics of the participants (N=371).

Characteristics	Participants, n (%)
Sex	
Male	216 (58.2)
Female	155 (41.8)
Education	
Completed PhD degree	25 (6.7)
Completed graduate degree	88 (23.7)
Completed college	158 (42.6)
Completed high school	49 (13.2)
Completed middle school	51 (13.7)
Employment status	
Full-time job	167 (45.0)
Part-time job	12 (3.2)
Unemployed or student	176 (47.4)
Retired	16 (4.3)
Monthly income (US $)	
<428	80 (21.6)
428-857	98 (26.4)
857-1428	77 (20.7)
>1428	60 (16.2)

XSL•FO
RenderX

Ethical Considerations

The Institutional Review Board office of Renmin University of China approved the research protocol, and responses were collected via Sojump. Before the experiment, participants were asked if they agree to participate in the research voluntarily for receiving compensation. Participants received a reward of RMB 7 (US $1). The final data set is anonymized, without any identifiable private information connected to participants.

Variables

Perceived Severity

Perceived severity was measured with five items adapted from a previous study [8]: (1) I think asymptomatic cases are a serious problem for us; (2) At some point in the future, we could all be threatened by asymptomatic cases; (3) I think asymptomatic cases have very serious consequences; (4) I think asymptomatic cases are a very serious problem; and (4) I think the presence of asymptomatic cases is a serious threat to my health (mean 5.26, SD 1.18; Cronbach α=0.89).

Perceived Susceptibility

Perceived susceptibility was measured with two items taken from an existing scale [8]: (1) I feel that I am also at risk of being infected by asymptomatic cases and (2) I feel that I may also be infected by asymptomatic cases (mean 4.86, SD 1.76; Pearson r=0.83, P<.001).

Information Seeking

Information seeking was measured based on the scale developed by Brashers et al [31] with the following eight items: (1) I would like to learn more information on asymptomatic cases than what is provided in this report; (2) I may discuss asymptomatic cases with people close to me (eg, friends, family); (3) I may ask my doctor about asymptomatic cases; (4) I may seek other news reports about asymptomatic cases; (5) I may pay close attention to news reports about asymptomatic cases that I encounter in the future; (6) I may check the internet for other information about asymptomatic cases; (7) I am likely to pay close attention to information on asymptomatic cases that I encounter on the internet in the future; and (8) I may read the scientific research literature on asymptomatic cases (mean 5.46, SD 1.30; Cronbach α=0.94).

Information Avoidance

Information avoidance was measured using the scales developed by Afifi and Weiner [13] and Evans et al [14] with the following six items: (1) I may try to change the subject if people close to me (eg, friends and family) discuss the issue of asymptomatic cases; (2) I may try to change the subject if my doctor discusses asymptomatic cases; (3) I may avoid exposure to other news reports about asymptomatic cases; (4) I may avoid exposure to information on the internet about asymptomatic cases; (5) I may try not to think too much about asymptomatic cases; and (6) I may try to forget about asymptomatic cases (mean 2.76, SD 1.73; Cronbach α=0.96).

Willingness to Adopt Preventive Behaviors

Willingness to adopt preventive behaviors was measured via 10 items using a 7-point Likert scale (1=strongly agree to 7=strongly disagree), which measured respondents' willingness to adopt a range of preventive behaviors such as strengthening protection, following health instructions, trying more preventive measures, and getting vaccinated. The specific items were: (1) I have decided to strengthen my preventive measures against COVID-19 immediately; (2) I intend to strengthen my protective measures against COVID-19 in the future; (3) I will pay more attention to asymptomatic cases; (4) I will persuade people around me to pay more attention to asymptomatic cases; (5) I will strictly follow professional health instructions to prevent catching the disease; (6) I will persuade people around me to strictly follow professional health instructions to prevent disease; (7) I will try as many ways as possible to prevent disease; (8) I will persuade those around me to try as many ways as possible to prevent disease; (9) I will get vaccinated as soon as a COVID-19 vaccine is developed; and (10) I will persuade those around me to get vaccinated as soon as a COVID-19 vaccine is developed (mean 5.54, SD 1.27; Cronbach α=0.89).

Vulnerability to False Information

Vulnerability to false information was measured by presenting respondents with 10 pieces of false news on COVID-19 (with five real news articles provided as distractors) to measure their trust in the false news. Sample items included the following:

> *Isatis root is suitable for treating conditions such as the common cold and viral influenza with heat symptoms. It has an antiviral effect and can help to prevent COVID-19.*
>
> *Tobacco particles are at the nanometer scale and can evenly cover the surface of lung cells, forming a barrier that keeps viruses out of the body. Therefore, smoking can reduce the infection of the novel coronavirus.*
>
> *The novel coronavirus is primarily a respiratory infection virus. Gargling with saline solution can eliminate the novel coronavirus bacteria that enter through the mouth.*

These items were measured on a scale of 1 (strongly disbelieve) to 10 (strongly believe) (mean 3.68, SD 1.46; Cronbach α=0.89).

Data Analysis

To answer the four research questions, ANOVA was performed to examine the main and interaction effects of the manipulated independent variables (ie, uncertainty) on dependent outcomes (ie, risk perceptions and misinformation vulnerability). The mediation model was used as a posthoc analysis to estimate statistically significant differences between experimental conditions on outcomes. SPSS (version 28.0) and PROCESS (version 4.2) were used for all statistical analyses.

Results

Effects of Uncertainty Framing in News Coverage

According to the ANOVA results (Figure 2), there was a significant interaction effect of threat uncertainty and solution uncertainty on individuals' vulnerability to misinformation ($F_{1,367}$=5.10, P=.02; partial η^2=0.01). Figure 2 provides the

XSL•FO
RenderX

complete data on group comparisons. Specifically, individuals who read news containing a certain threat and uncertain solution (mean 3.86, SD 0.15) were more likely to believe misinformation than those who read news containing a certain threat and certain solution (mean 3.47, SD 0.16). However, there was neither a significant main effect of threat uncertainty (partial η^2=0.00) nor a significant main effect of solution uncertainty (partial η^2=0.00) on individuals' vulnerability to false information (Table 3).

In terms of information processing (Table 4), the threat uncertainty of health information showed a significant effect in promoting information-seeking behavior (partial η^2=0.012). This implies that people who read news with the presence of an uncertain threat were more likely to search for additional relevant information than those who read news with absence of an uncertain threat. However, the uncertainty framing in the news, regardless of the type of uncertainty, did not affect individuals' information avoidance behaviors.

Figure 2. Interaction effect of an uncertainty frame of news coverage on people's misinformation vulnerability.

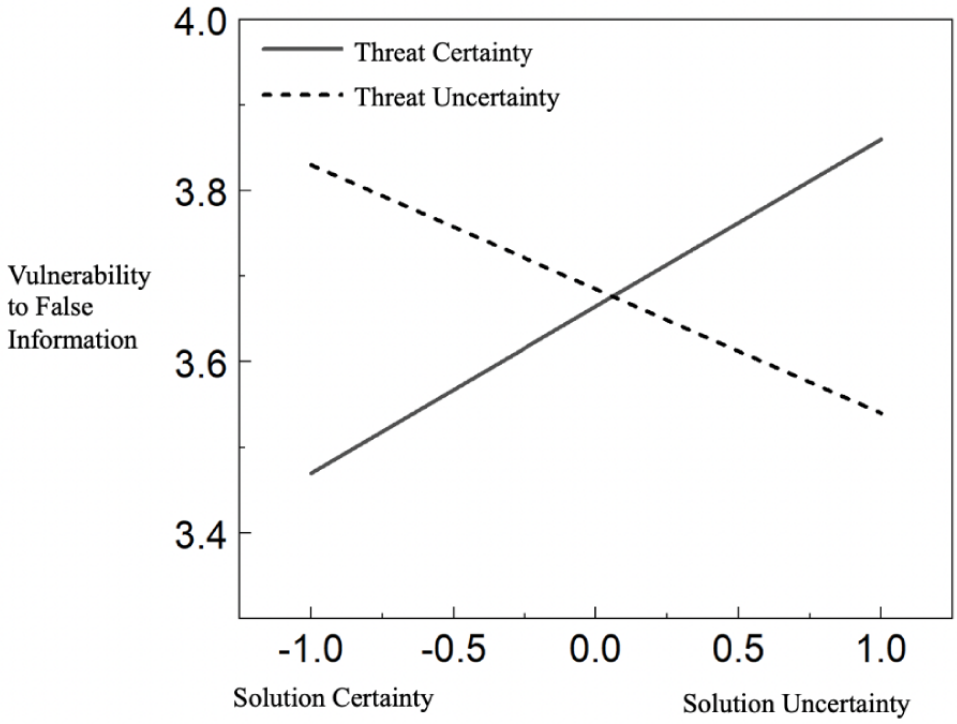

Table 3. ANOVA results for determining the main effects of an uncertainty frame of a threat and solution on individuals' risk perceptions and misinformation vulnerability.

Effect	Severity	Susceptibility	Misinformation vulnerability
Threat			
Certainty, mean (SD)	5.12 (0.10)	4.86 (0.13)	3.67 (0.11)
Uncertainty, mean (SD)	5.34 (0.10)	4.86 (0.13)	3.68 (0.11)
F (*df*=1, 367)	1.68	0.00	0.01
P value	.20	>.99	.91
Solution			
Certainty, mean (SD)	5.10 (0.10)	4.78 (0.13)	3.65 (0.11)
Uncertainty, mean (SD)	5.39 (0.10)	4.94 (0.13)	3.70 (0.10)
F (*df*=1, 367)	4.29	0.78	0.09
P value	.04	.38	.76

XSL•FO
RenderX

Table 4. ANOVA results for main effects of uncertainty framed in the news on motivating individuals' information-seeking and information-avoiding behaviors.

Effect	Information seeking	Information avoidance
Threat		
Certainty, mean (SD)	5.31 (0.10)	2.89 (0.13)
Uncertainty, mean (SD)	5.60 (0.10)	2.62 (0.13)
F (*df*=1, 367)	4.47	2.26
P value	.04	.13
Solution		
Certainty, mean (SD)	5.40 (0.10)	2.60 (0.13)
Uncertainty, mean (SD)	5.51 (0.10)	2.91 (0.12)
F (*df*=1, 367)	0.61	3.07
P value	.43	.08

Sequential Mediation Analysis

Serial linear regression with PROCESS macro model 81 was used to analyze how the dummy-coded variables (1=uncertain threat and certain solution, 2=certain threat and uncertain solution, 3=uncertain threat and uncertain solution, and reference (0)=certain threat and certain solution) influence preventive behavioral intention through the factors perceived severity, information seeking, and information avoidance. A significant serial mediation model was detected, in which the perceived severity and information seeking would sequentially mediate the relationship between the exposure to news containing a certain threat and an uncertain solution (point estimate 0.17, SE 0.08; 95% CI 0.02-0.33) or news containing an uncertain threat and an uncertain solution (point estimate 0.18, SE 0.08; 95% CI 0.03-0.34) and the protective behavioral intention (see Figure 3 for the path significance and coefficients).

The mediation models obtained with PROCESS macro model 81, including misinformation vulnerability, information seeking, and information avoidance as the three mediators, were established with the same three dummy-coded comparison variables described above. Although the serial mediating effect of misinformation vulnerability and information processing on the relationship was nonsignificant, significant associations between misinformation vulnerability on information processing and protective behavioral intention were detected (see Figure 4 for the path significance and coefficients).

Figure 3. Sequential mediators of perceived severity and information processing with exposure to uncertainty for preventive behavioral intention (N=371). Significant paths are presented with solid lines and nonsignificant paths are presented with dotted lines. Dummy code: 1=uncertain threat and certain solution; 2=certain threat and uncertain solution; 3=uncertain threat and uncertain solution; and reference (0)=certain threat and certain solution.

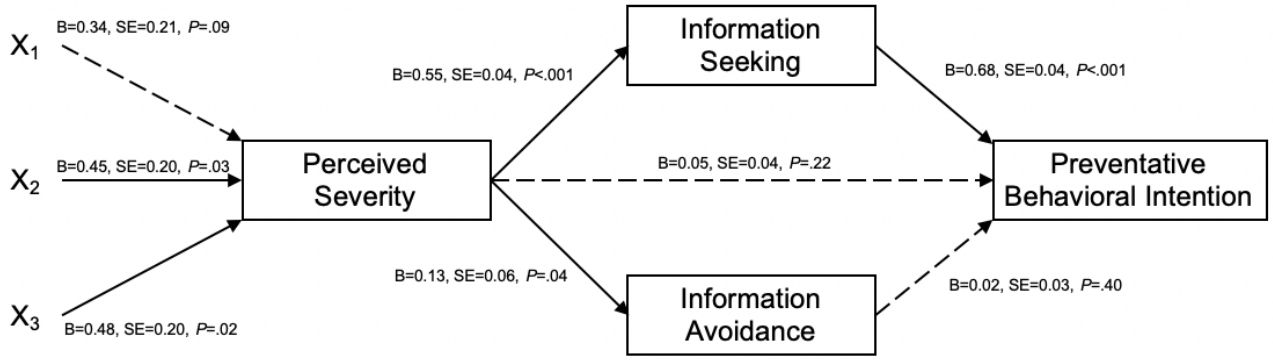

XSL•FO
RenderX

Figure 4. Sequential mediators of misinformation vulnerability and information processing with exposure to uncertainty for preventative behavioral intention (N=371). Significant paths are presented with solid lines and nonsignificant paths are presented with dotted lines. Dummy code: 1=uncertain threat and certain solution; 2=certain threat and uncertain solution; 3=uncertain threat and uncertain solution; and reference (0)=certain threat and certain solution.

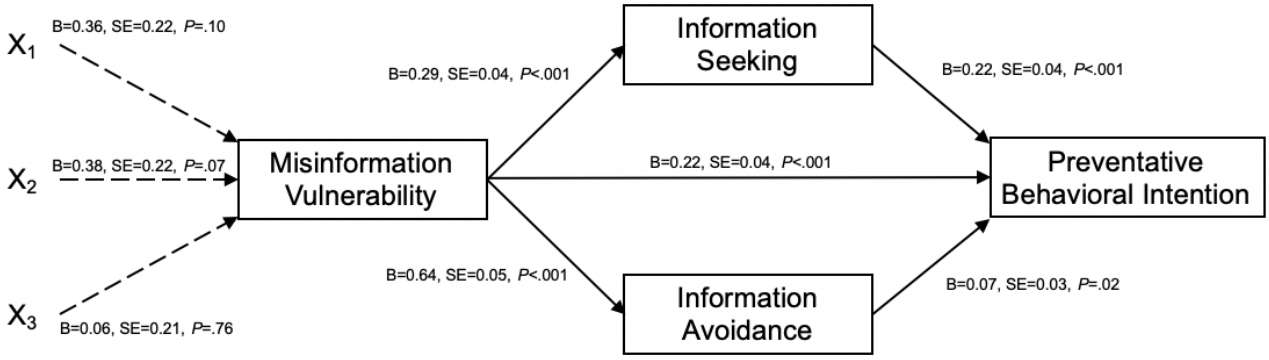

Discussion

Principal Findings

This study explored the impact of incorporating different levels of uncertainty in news articles on individuals' risk perception, susceptibility to misinformation, subsequent information processing, and intention toward preventive behaviors. These factors play a crucial role in guiding individuals to protect themselves during a public health crisis.

First, as expected, the findings showed that individuals would perceive the crisis to be of greater severity when reading news framed withing a context of solution uncertainty, regardless of whether or not the threat of the crisis was framed as certain or uncertain in the news article. Furthermore, the perceived severity would in turn motivate individuals' information-seeking and information-avoidance behaviors. As suggested by Gudykunst [32], the experience of fear and concern is also associated with uncertainty. Individuals' fear might be evoked when they fail to find a solution to solve the risky problem, which would consequently increase their perception of the severity of the health threat. Meanwhile, to cope with the fear, individuals would either seek more information or avoid more information [31].

Second, individuals tended to believe false information during a public health crisis, especially after reading a news article containing certain threats and uncertain solutions. Furthermore, posthoc regression analysis suggested that both information seeking and information avoidance were positively associated with misinformation vulnerability. As expected, information seeking was positively associated with protective behavioral intentions, while information avoidance was negatively associated with protective behavioral intention. That said, communicating crisis uncertainty in a news article would be risky in terms of increasing the public's vulnerability to misinformation. Moreover, misinformation vulnerability would further motivate information avoidance, which would consequently dissuade individuals' intention to adopt preventive behaviors. However, communicating crisis uncertainty could also be beneficial because the increased misinformation vulnerability that arises after reading news with related uncertainty could simultaneously motivate both expected health information–seeking and protective behaviors.

Third, our findings suggest that compared to an uncertain threat, uncertain solutions are more potentially problematic. An uncertain solution with either an uncertain or certain threat had a greater impact on individuals' perceived severity of the crisis. Moreover, an uncertain solution with an uncertain threat led to higher vulnerability to misinformation. Although both perceived severity and misinformation vulnerability could motivate expected health information–seeking or protective behavioral intentions, they were likely to trigger information avoidance, which would further impair the protective behavioral intention. Given these conflicting findings, more research is needed to understand the mechanism behind the effects of communicating solution uncertainty, especially during a public health crisis.

Implications and Limitations

A few limitations of this study should be noted to provide inspiring suggestions for further research. First, the generalizability of the findings needs to be addressed. This study is based on an exceptionally unique context of a public health event. In April 2020, the ambiguity surrounding the perceived risk of COVID-19 and the appropriate preventive measures was prevalent. Different regions in China implemented varying degrees of epidemic control measures, leading to divergent strategies. Media coverage of "asymptomatic carriers" and other aspects of the COVID-19 pandemic often exhibited inconsistent or contradictory information. Thus, the diverse experimental scenarios were based on natural contexts for presenting four distinct threat-and-solution scenarios and would not induce a perceptual conflict for the participants. However, when extrapolating to other research topics, it is imperative to consider the coherence between experimental scenarios and real-world settings.

Second, the impact of the research topic on participants with different characteristics requires further discussion. This study endeavored to achieve a balanced representation of participants with respect to sex, education level, and income bracket. Consequently, an exhaustive examination of the differential effects stemming from various sociodemographic factors on the outcomes was not performed. However, in distinct research inquiries, this divergence may prove consequential. Hence,

XSL•FO
RenderX

future research endeavors could delve deeper into understanding the perception of information uncertainty and management behaviors across populations with diverse backgrounds and characteristics.

Third, although this type of experimental design typically includes 80-90 participants per group to achieve sufficient statistical power, the unexpected smaller effect size in this study resulted in a reduced statistical power of approximately 60%. This diminutive effect size may compromise the study's sensitivity in detecting meaningful relationships, thereby affecting the reliability and generalizability of the findings. To enhance statistical power and ensure robust results in similar experimental studies in China, a larger sample size is recommended.

Regarding the implications, this study aimed to elucidate the impact of health information uncertainty on individuals' information-processing mechanisms. The results thus provide further evidence for the impacts of individuals' perceptions and behaviors underlying uncertainty management theory. Amid public health events inundated with uncertain information, individuals' perceptions and behaviors related to uncertainty management often determine their attitudes toward addressing health threats and the potential adoption of health measures. Therefore, comprehending this process contributes to facilitating more effective health communication between the public health system and the general public. Faced with uncertain public health events, participants such as public health institutions, media, and the general public should all take into account the implications of information uncertainty, ensuring the effective dissemination of information throughout all stages of the crisis.

Acknowledgments

This work was supported by the National Social Science Fund of China (project name: Construction and Impact Research of Trust Mechanisms in Health Communication during the "Information Epidemic"; project number 22CXW022).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Simulated news articles presented to the participants in the online experiment under four conditions varying in threat and solution uncertainty about the COVID-19 pandemic.
[DOCX File , 1829 KB - jmir_v26i1e51910_app1.docx]

References

1. Capurro G, Jardine CG, Tustin J, Driedger M. Communicating scientific uncertainty in a rapidly evolving situation: a framing analysis of Canadian coverage in early days of COVID-19. BMC Public Health 2021 Nov 29;21(1):2181 [FREE Full text] [doi: 10.1186/s12889-021-12246-x] [Medline: 34844582]
2. Büchter RB, Betsch C, Ehrlich M, Fechtelpeter D, Grouven U, Keller S, et al. Communicating uncertainty in written consumer health information to the public: parallel-group, web-based randomized controlled trial. J Med Internet Res 2020 Aug 10;22(8):e15899 [FREE Full text] [doi: 10.2196/15899] [Medline: 32773375]
3. van den Bos K, Ham J, Lind EA, Simonis M, van Essen WJ, Rijpkema M. Justice and the human alarm system: the impact of exclamation points and flashing lights on the justice judgment process. J Exp Soc Psychol 2008 Mar;44(2):201-219. [doi: 10.1016/j.jesp.2007.03.001]
4. Afifi WA, Weiner JL. Toward a theory of motivated information management. Commun Theory 2004 May;14(2):167-190. [doi: 10.1111/j.1468-2885.2004.tb00310.x]
5. Chu TH, Yeo TED, Su Y. Effects of exposure to COVID-19 news and information: a meta-analysis of media use and uncertainty-related responses during the pandemic. Journalism Mass Commun Quart 2022 Jan 04;99(1):89-112. [doi: 10.1177/10776990211068857]
6. Bed bugs FAQs. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/parasites/bedbugs/faqs.html [accessed 2020-09-16]
7. So J, Kuang K, Cho H. Information seeking upon exposure to risk messages: predictors, outcomes, and mediating roles of health information seeking. Commun Res 2016 Nov 24;46(5):663-687. [doi: 10.1177/0093650216679536]
8. Witte K, Cameron KA, McKeon JK, Berkowitz JM. Predicting risk behaviors: development and validation of a diagnostic scale. J Health Commun 1996 Oct 29;1(4):317-341. [doi: 10.1080/108107396127988] [Medline: 10947367]
9. Lalot F, Abrams D, Travaglino GA. Aversion amplification in the emerging COVID‐19 pandemic: the impact of political trust and subjective uncertainty on perceived threat. Commun Appl Soc Psy 2020 Dec 13;31(2):213-222. [doi: 10.1002/casp.2490]
10. Pine KH, Lee M, Whitman SA, Chen Y, Henne K. Making sense of risk information amidst uncertainty: individuals' perceived risks associated with the COVID-19 pandemic. 2021 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445051]

XSL•FO
RenderX

11. Lu J, Zhang M, Zheng Y, Li Q. Communication of uncertainty about preliminary evidence and the spread of its inferred misinformation during the COVID-19 pandemic-a Weibo case study. Int J Environ Res Public Health 2021 Nov 13;18(22):11933 [FREE Full text] [doi: 10.3390/ijerph182211933] [Medline: 34831688]
12. Zhao X, Fan J, Basnyat I, Hu B. Online health information seeking using "#COVID-19 Patient Seeking Help" on Weibo in Wuhan, China: descriptive study. J Med Internet Res 2020 Oct 15;22(10):e22910 [FREE Full text] [doi: 10.2196/22910] [Medline: 33001838]
13. Afifi WA, Weiner JL. Seeking information about sexual health: applying the theory of motivated information management. Human Comm Res 2006 Jan;32(1):35-57. [doi: 10.1111/j.1468-2958.2006.00002.x]
14. Evans BM, Kairam S, Pirolli P. Do your friends make you smarter?: an analysis of social strategies in online information seeking. Inf Process Manage 2010 Nov;46(6):679-692. [doi: 10.1016/j.ipm.2009.12.001]
15. Hogan TP, Brashers DE. The theory of communication and uncertainty management: implications from the wider realm of information behavior. In: Afifi T, Afifi W, editors. Uncertainty, Information Management, and Disclosure Decisions: Theories and Applications. New York: Routledge/Taylor & Francis Group; 2009:45-66.
16. Charpentier CJ, Cogliati Dezza I, Vellani V, Globig LK, Gädeke M, Sharot T. Anxiety increases information-seeking in response to large changes. Sci Rep 2022 May 05;12(1):7385. [doi: 10.1038/s41598-022-10813-9] [Medline: 35513397]
17. Keller AM, Taylor HA, Brunyé TT. Uncertainty promotes information-seeking actions, but what information? Cogn Res Princ Implic 2020 Sep 07;5(1):42 [FREE Full text] [doi: 10.1186/s41235-020-00245-2] [Medline: 32894402]
18. Brainin E, Neter E. Refined analysis of older eHealth users from an agency perspective: quantitative telephone interview study. JMIR Aging 2023 Apr 26;6:e40004 [FREE Full text] [doi: 10.2196/40004] [Medline: 37121572]
19. Case DO, Andrews JE, Johnson JD, Allard SL. Avoiding versus seeking: the relationship of information seeking to avoidance, blunting, coping, dissonance, and related concepts. J Med Libr Assoc 2005 Jul;93(3):353-362 [FREE Full text] [Medline: 16059425]
20. Goodall CE, Reed P. Threat and efficacy uncertainty in news coverage about bed bugs as unique predictors of information seeking and avoidance: an extension of the EPPM. Health Commun 2013 Jan;28(1):63-71. [doi: 10.1080/10410236.2012.689096] [Medline: 23330859]
21. Jiang S, Liu PL, Ngien A, Wu Q. The effects of worry, risk perception, information-seeking experience, and trust in misinformation on COVID-19 fact-checking: a survey study in China. Chinese J Commun 2022 Sep 12;16(2):132-149. [doi: 10.1080/17544750.2022.2121931]
22. Liu BF, Viens J. Crisis and risk communication scholarship of the future: reflections on research gaps. J Int Crisis Ris Commun Res 2020 Mar 9;3(1):7-14. [doi: 10.30658/jicrcr.3.1.1]
23. Liu BF, Xu S, Rhys Lim J, Egnoto M. How publics' active and passive communicative behaviors affect their tornado responses: an integration of STOPS and SMCC. Public Relat Rev 2019 Nov;45(4):101831. [doi: 10.1016/j.pubrev.2019.101831]
24. Griffin RJ, Dunwoody S, Neuwirth K. Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. Environ Res 1999 Feb;80(2 Pt 2):S230-S245. [doi: 10.1006/enrs.1998.3940] [Medline: 10092438]
25. Bode L, Vraga EK. See something, say something: correction of global health misinformation on social media. Health Commun 2018 Sep;33(9):1131-1140. [doi: 10.1080/10410236.2017.1331312] [Medline: 28622038]
26. Tan ASL, Lee C, Chae J. Exposure to health (mis)information: lagged effects on young adults' health behaviors and potential pathways. J Commun 2015 Jul 06;65(4):674-698. [doi: 10.1111/jcom.12163]
27. Soroya SH, Farooq A, Mahmood K, Isoaho J, Zara S. From information seeking to information avoidance: understanding the health information behavior during a global health crisis. Inf Process Manag 2021 Mar;58(2):102440 [FREE Full text] [doi: 10.1016/j.ipm.2020.102440] [Medline: 33281273]
28. Sharit J, Czaja SJ. Overcoming older adult barriers to learning through an understanding of perspectives on human information processing. J Appl Gerontol 2020 Mar 30;39(3):233-241 [FREE Full text] [doi: 10.1177/0733464818794574] [Medline: 30160573]
29. Dai B, Ali A, Wang H. Exploring information avoidance intention of social media users: a cognition–affect–conation perspective. Internet Res 2020 Jul 14;30(5):1455-1478. [doi: 10.1108/intr-06-2019-0225]
30. "Asymptomatic infections" enter daily bulletin: epidemic prevention and control must be realistic and pragmatic. Xinhua News Agency. URL: http://www.xinhuanet.com/politics/2020-04/02/c_1125803965.htm [accessed 2020-04-02]
31. Brashers DE, Goldsmith DJ, Hsieh E. Information seeking and avoiding in health contexts. Human Comm Res 2002 Apr;28(2):258-271. [doi: 10.1111/j.1468-2958.2002.tb00807.x]
32. Gudykunst W. Anxiety/uncertainty management (AUM) theory: current status. In: Wiseman R, editor. Intercultural Communication Theory. New York: Sage Publications; 1995:8-58.

XSL•FO
RenderX

Edited by G Eysenbach, L Beri; submitted 16.08.23; peer-reviewed by S Su, X Xu, A Chang; comments to author 06.12.23; revised version received 08.01.24; accepted 12.04.24; published 14.05.24.

Please cite as:
Zhao R, Lu X, Yang J, Li B
Understanding the Impact of Communicating Uncertainty About COVID-19 in the News: Randomized Between-Subjects Factorial Experiment
J Med Internet Res 2024;26:e51910
URL: https://www.jmir.org/2024/1/e51910
doi:10.2196/51910
PMID:38743940

©Rui Zhao, Xuerong Lu, Jiayi Yang, Biao Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Difference in Rumor Dissemination and Debunking Before and After the Relaxation of COVID-19 Prevention and Control Measures in China: Infodemiology Study

Xiaoqi Liu[1], BS; Qingyuan Hu[2], MS; Jie Wang[3], BS; Xusheng Wu[4], PhD; Dehua Hu[1], PhD

[1]Department of Biomedical Informatics, School of Life Sciences, Central South University, Changsha, Hunan, China
[2]Department of Research, The Third Xiangya Hospital of Central South University, Changsha, Hunan, China
[3]School of Life Sciences, Central South University, Changsha, Hunan, China
[4]Shenzhen Health Care Development Research and Data Management Center, Shenzhen, Guangdong, China

Corresponding Author:
Dehua Hu, PhD
Department of Biomedical Informatics
School of Life Sciences
Central South University
172 Tongzipo Road
Changsha, Hunan, 410013
China
Phone: 86 13975869106
Email: hudehua@csu.edu.cn

Abstract

Background: The information epidemic emerged along with the COVID-19 pandemic. While controlling the spread of COVID-19, the secondary harm of epidemic rumors to social order cannot be ignored.

Objective: The objective of this paper was to understand the characteristics of rumor dissemination before and after the pandemic and the corresponding rumor management and debunking mechanisms. This study aimed to provide a theoretical basis and effective methods for relevant departments to establish a sound mechanism for managing network rumors related to public health emergencies such as COVID-19.

Methods: This study collected data sets of epidemic rumors before and after the relaxation of the epidemic prevention and control measures, focusing on large-scale network rumors. Starting from 3 dimensions of rumor content construction, rumor propagation, and rumor-refuting response, the epidemic rumors were subdivided into 7 categories, namely, involved subjects, communication content, emotional expression, communication channels, communication forms, rumor-refuting subjects, and verification sources. Based on this framework, content coding and statistical analysis of epidemic rumors were carried out.

Results: The study found that the rumor information was primarily directed at a clear target audience. The main themes of rumor dissemination were related to the public's immediate interests in the COVID-19 field, with significant differences in emotional expression and mostly negative emotions. Rumors mostly spread through social media interactions, community dissemination, and circle dissemination, with text content as the main form, but they lack factual evidence. The preferences of debunking subjects showed differences, and the frequent occurrence of rumors reflected the unsmooth channels of debunking. The χ^2 test of data before and after the pandemic showed that the *P* value was less than .05, indicating that the difference in rumor content before and after the pandemic had statistical significance.

Conclusions: This study's results showed that the themes of rumors during the pandemic are closely related to the immediate interests of the public, and the emotions of the public accelerate the spread of these rumors, which are mostly disseminated through social networks. Therefore, to more effectively prevent and control the spread of rumors during the pandemic and to enhance the capability to respond to public health crises, relevant authorities should strengthen communication with the public, conduct emotional risk assessments, and establish a joint mechanism for debunking rumors.

(J Med Internet Res 2024;26:e48564) doi:10.2196/48564

XSL•FO
RenderX

KEYWORDS

new stage; public health emergency; information epidemic; propagation characteristic; debunking mechanism; China

Introduction

Background

Since the outbreak of COVID-19, with the development of information technology and the improvement of digital technology, a large number of epidemic and related information has been fermented and disseminated on various network platforms. The false information has been overwhelming, making it difficult to distinguish truth from falsehood. This phenomenon has also been referred to as an "infodemic" by the World Health Organization [1]. Under the infodemic, rumors such as "Delta and Omicron are co-circulating" and "the State Council Joint Prevention and Control Group is formally removed from its duties" spread widely, and the harm of rumors cannot be underestimated. First of all, such epidemic-related rumors not only lead to the adoption of excessive or incorrect epidemic prevention measures by the public but also easily undermine the public's confidence in epidemic prevention, causing secondary harm to social order. In addition, with rumors abound, scientific voices will be drowned out. As the infodemic spreads, false information is spread erroneously, whereas scientific, correct, and official epidemic and related information is not properly disseminated due to various rumors. This makes it difficult for some people to discern reliable sources of information from the massive amount of information available, thereby causing social panic and seriously threatening social stability.

The information environment is highly uncertain, and the characteristics of rumor dissemination will also change with different pandemic stages. Therefore, comparing and analyzing the characteristics of rumor dissemination before and after the new stage of epidemic prevention and control, as well as exploring the debunking mechanism under the new stage of epidemic prevention and control, is of great significance for the governance of rumors, for relevant departments to carry out debunking work, and to avoid public health crises. This study will comprehensively analyze large-scale network rumors related to the epidemic before and after the release of the epidemic prevention and control measures and explore them from multiple dimensions, such as the construction of rumor content, rumor dissemination, and debunking response. Based on this, the characteristics of the dissemination of COVID-19 network rumors before and after the release of the epidemic prevention and control measures and the debunking mechanism will be proposed, providing a reference for the governance of network rumors of sudden public health events by relevant departments and the development of debunking work under the information epidemic.

Conceptual Definition and Related Research

Definition and Characteristics of Rumors

The definition of "rumor" has evolved over time, and the "rumor" initially mentioned in "Eastern Zhou States" refers to false information intentionally fabricated without a factual basis [2]. Further research introduced various definitions across disciplines. In communication studies, rumors are information lacking a factual basis, potentially including defamatory elements [3]. Knapp [4] describes rumors as declarations that are meant to be believed, related to current events, and spread without official confirmation. In social psychology, rumors are defined as baseless content spread by individuals or groups with specific motives [5]. Sun [6] views rumors as propositions related to current events, spread mainly through oral communication, but without a solid basis for their accuracy. Rumors are fundamentally unverified and lack evidence. The rise of social media has brought concepts such as "misinformation," "disinformation," and "fake news" to the forefront. "Misinformation" refers to inadvertently shared false information, whereas "disinformation" is the deliberate spread of falsehoods [7]. "Fake news" is defined as falsehoods designed to deceive by mimicking real news [8], with Tandoc [9] categorizing fake news into 6 types, including news satire and parody. Network rumors emerge with internet technology, differing from traditional rumors in spread speed, range, and methods. Defined as rumors spread through the internet without deviating from the basic characteristics of traditional rumors [10], they play a crucial role in the generation, dissemination, and impact of rumors on the internet [11]. Compared to traditional rumors, network rumors feature rapid spread, wide reach, and high influence and are challenging to control effectively [12,13].

The Process and Patterns of Rumor Propagation

Studying how rumors spread can help curb the impact of destructive misinformation on the internet [14]. Research into rumor spreading dates back to Allport and Postman [15], who introduced the rumor formula as a function of importance and ambiguity in their book *The Psychology of Rumors*. Recent studies have explored rumor spreading patterns, with Zhang et al [16] identifying epidemic-related, web-based rumors as community, micro, and circle spreading, and Tong and Yu [17] summarizing web-based rumor spreading patterns as 1-way linear, multidirectional network, and umbrella-shaped propagation. Simulation models for analyzing rumor content have also advanced. For instance, Xi and Yong [18] proposed a model for public opinion dissemination on social networks, Han et al [19] analyzed network topology in rumor spreading, Wu et al [20] developed a rumor-spreading model based on recommendation systems, and Fang and Yang [21] studied rumor spreading and refutation effectiveness on Weibo under real-name systems. Moreover, epidemic models have been applied to social network rumor propagation studies, based on similarities between rumor spread and disease transmission mechanisms [22], including SI (susceptible and infectious) [23-25], SIS (susceptible, infectious, and susceptible) [26-28], SIR (susceptible, infectious, and recovered) [29-31], and SEIR (susceptible, exposed, infectious, recovered) [32-34] models, representing susceptible, infected, recovered, and exposed states, respectively [35]. Zhu et al [36] introduced a SIS model with a forced silence function, Sun et al [37] applied uncertain

XSL•FO
RenderX

differential equations to study an SIR model, Hosseini and Zandvakili [38] added rumor delay and countermeasure mechanisms to an SIR model, and Chen et al [39] proposed a new SEIOR (ignorant, hesitators, spreaders, rumor debunkers, and immunizers) model incorporating debunkers.

Governance and Refutation Mechanisms of Rumors

Rumor governance involves 5 main mechanisms: web-based rumor warning and monitoring, improving information disclosure, postevent evaluation, legal mechanisms for web-based rumor punishment, and long-term education [40]. Effective curbing of fake news spread can be achieved through enhancing individual abilities to recognize fake news and structural changes in platform policies [41]. Media literacy training is suggested to prevent misinformation uptake, with a recommendation for tools to identify harmful misinformation [42]. The refutation mechanism, addressing rumors that have already occurred, includes top-down and bottom-up approaches. The top-down approach involves proactive dissemination of official information, whereas the bottom-up approach refers to seeking expert verification of public feedback or rumors and providing feedback through refutation articles [43]. Governments are advised to release official rumor refutation information promptly and ensure its continuous dissemination [44]. Establishing rumor monitoring and refutation mechanisms on various platforms, making policy adjustments according to different stages, and timely releasing official information are recommended [45]. Active dissemination of official information and public engagement in verification can effectively eliminate false information spread.

Methods

Sampling and Sample

This study focused on web-based rumors and used rumors related to the COVID-19 epidemic as the research sample. Since rumors originate from various social media platforms such as WeChat, Weibo, Douyin, Kuaishou, and Toutiao, during the epidemic, various media outlets have also launched various epidemic information refutation zones, such as the Hubei Network Refutation WeChat public account, Toutiao Refutation, and Gansu Network Refutation Reporting Platform, etc. Refutation information is relatively scattered. In comparison, the China Internet Joint Rumor Refutation Platform is sponsored by the Central Cyberspace Affairs Commission's Illegal and Harmful Information Reporting Center. Hosted by Xinhua Net, it has the function of reporting and verifying rumors and can obtain authoritative refutation information from relevant departments and experts. Therefore, the rumor samples in this study are more representative. In addition, since the new stage of epidemic prevention and control has just begun, the number of rumor samples is relatively small. Therefore, we decided to conduct further searches on Toutiao to expand the sample size.

In this study, web crawling tools were initially used to crawl all entries to date (March 2, 2023) from the "Epidemic Prevention and Control Refutation Zone" of the China Internet Joint Rumor Refutation Platform. Subsequently, searches were conducted on Toutiao using the keyword "epidemic debunking" for each of the 31 provinces (direct-controlled municipalities and autonomous regions), and entries from the "News" section were collected. The samples from both platforms were then consolidated, using a fuzzy matching method based on title fields to eliminate duplicate rumor samples while excluding rumors already confirmed as facts. In the end, a total of 672 rumors were obtained as the final research sample. The timeframe of the rumor samples spanned from April 19, 2022, to March 2, 2023, covering both the period before the relaxation of epidemic measures (from April 19, 2022, to November 30, 2022) and the early stage of the new phase of epidemic prevention and control (from December 1, 2022, to March 2, 2023).

Coding Category Construction

Combining with the Laswell 5W communication model [46], the content to be studied in this paper can be roughly divided into "communicators," "content," "channels," "audiences," and "effects." However, based on the content of the selected web-based rumors and considering the complexity and operability of the analysis of "audiences" and "effects," the categories of this study were composed of "involved entity," "propagation content," "emotional expression," "propagation channel," "propagation form," "debunking entity," and "verification source" [47]. The specific category analysis framework is shown in the Table 1.

XSL•FO
RenderX

Table 1. Analysis framework and category construction.

Analytical and statistical categories	Category description	Observation pathway
Content construction		
Involved entity	Entities involved in the spread of rumors	1. Virus 2. Patients 3. Transportation department 4. Enterprises 5. Schools 6. Hospitals 7. Government agencies 8. Others
Propagation content	Division of the themes of rumor content	1. Epidemic situation 2. Epidemic anecdotes 3. Epidemic prevention knowledge 4. Epidemic prevention measures 5. Others
Emotional expression	Emotional tendencies conveyed in rumor content	1. Fear 2. Worry 3. Sadness 4. Anger 5. Doubt 6. Recognition 7. Opposition 8. Surprise 9. No clear emotion
Rumor dissemination		
Propagation channel	Platforms and methods of rumor dissemination	1. Douyin 2. Toutiao 3. WeChat groups or Moments 4. Weibo 5. Others or unknown
Propagation form	Modes of presenting rumor content	1. Images 2. Text 3. Video 4. Others or unknown
Refutation and response		
Debunking entity	Publisher of refutation information	1. Media 2. Government 3. Refutation platforms 4. Collaboration
Verification source	Sources used for verifying rumor information	1. Relevant government departments 2. Authoritative experts 3. Enterprises 4. Others or unknown

During the COVID-19 epidemic, the entities involved were complex and diverse. Rumors involving multiple entities were classified based on the most important entity. The specific categories of the content are shown in Table 2. The epidemic situation included information directly related to the occurrence and development of the epidemic, such as epidemic data, trends, regions, policies, and other information. Epidemic curiosities included all kinds of strange stories related to the epidemic, such as epidemic events, aid, heroes, and other information. Epidemic prevention knowledge included related knowledge on interpreting and coping with the epidemic, such as epidemic prevention and control, transmission, causes, science popularization, and other information. Epidemic prevention measures included the epidemic prevention and control measures taken by relevant departments, such as epidemic detection, isolation, treatment, and other information. Anything that did not fall into these 4 categories was classified as other.

XSL•FO
RenderX

Table 2. Subcategories of rumor content in epidemic spread.

Subcategory	Rumor content
Epidemic situation	• Epidemic data: including statistics such as the number of infections, number of recoveries, number of deaths, etc • Epidemic trend: including predictions on the development, changes, and direction of the epidemic • Epidemic region: including information on epidemic areas, severity, risk levels, etc • Epidemic policies: including epidemic-related policies and measures issued by various levels of government
Epidemic oddities	• Epidemic events: including various events and accidents related to the epidemic • Epidemic aid: including information on domestic and foreign aid, donations, etc • Epidemic heroes: including the deeds of frontline workers such as medical staff and volunteers
Epidemic prevention knowledge	• Epidemic prevention and control: including preventive measures, personal protective knowledge, etc • Epidemic transmission: including information on transmission routes, transmission methods, etc • Epidemic causes: including relevant knowledge on virus sources, infection mechanisms, etc • Epidemic science popularization: including various popular science knowledge, epidemic interpretation, scientific research results, etc
Epidemic prevention measures	• Epidemic detection: including virus detection, nucleic acid detection, and other related information • Epidemic isolation: including home isolation, centralized isolation, and other epidemic prevention measures • Epidemic treatment: including various treatment plans, drugs, vaccines, and other information • Epidemic resumption: including information on resumption plans and measures of enterprises, schools, and other units

Emotional expression is the emotional tendency of rumor information presentation. Due to the extremely rich types of emotions and the fact that the same rumor may express multiple emotions, to facilitate statistical classification, this paper defined a rumor's emotional expression category based on the most obvious emotional tendency in the rumor information. The specific subcategories of emotional expression are shown in Table 3. The emotional expression of rumors can be divided into negative emotions, positive emotions, and no obvious emotions. Among them, negative emotions included fear, doubt, worry, opposition, anger, and sadness, and positive emotions included approval [48]. As surprise can represent both positive and negative emotions in different contexts, it was not classified into emotional attributes here.

Table 3. Subcategories of emotions expressed in COVID-19 rumors.

Emotion subcategory	Description
Fear	Focuses on information related to the spread of the virus and the death rate, which can cause panic and fear among people
Worry	Focuses on the impact of the pandemic on individuals, families, and society, which can cause people to worry and feel anxious
Sadness	Focuses on information and reports related to victims of the pandemic, which can lead to feelings of sadness
Anger	Focuses on information related to the origin of the virus and the handling of the pandemic, which can cause people to feel dissatisfied and angry
Doubt	Focuses on information related to the authenticity of pandemic data and the credibility of government agencies, which can cause people to be suspicious and doubtful
Approval	Focuses on information related to known methods of prevention and scientific research findings, which can cause people to accept and approve of such information
Opposition	Focuses on voices of opposition to pandemic response measures, which can cause people to feel resentment and dissatisfaction
Surprise	Focuses on strange and unusual events related to the pandemic and heroic deeds of medical workers, which can cause people to feel surprised and moved
Not apparent	Focuses on other information related to the pandemic that may not have an obvious emotional tendency

Rumors often spread through multiple media platforms. Based on content analysis of rumor samples, this paper classified the channels of rumor dissemination into 5 main forms listed in Table 1. The propagation form referred to the way the rumor information was presented, mainly divided into 3 categories: images, text, and video. The debunking entity referred to the entity that clarified the rumor, mainly including media, government, and debunking platforms. The verification source

XSL•FO
RenderX

referred to authoritative sources of information used to verify the rumor. In the context of epidemic rumors, verification primarily came from relevant government departments, authoritative experts, and enterprises, which are the parties referred to in the rumor information or authoritative entities in the relevant field.

In addition, in the coding process, it was necessary to scientifically design and control the coding process to ensure the reliability of the coding. The coding reliability assurance measures in this study were as follows: (1) three coders with qualitative research and coding experience encoded the analysis text and conducted a Krippendorff α coefficient test [49]; (2) all personnel received coding training, selected some indicators at random for trial coding, discussed the coding results, improved coding standards, and reached a consensus for formal coding; and (3) during the formal coding, coders independently completed the coding work, communicated the coding results together after the completion of the coding work, and discussed the points of disagreement to ensure the coding reliability. By sorting out the data of coders and using SPSS statistical software (version 25.0; IBM Corp) to conduct a Krippendorff α reliability test on the data, the results showed that the reliability coefficients were all greater than 0.80, demonstrating a high level of reliability.

Statistical Analysis

Data were extracted from the "Epidemic Prevention and Control Rumor Refutation Special Area" of the China Internet Joint Rumor Refutation Platform and Toutiao. The extracted data were organized according to the analysis framework and category construction of this study, processed using Microsoft Excel, and analyzed using SPSS statistical software. The χ^2 test was used for categorical data, and the Pearson χ^2 test or likelihood ratio χ^2 test was selected based on the sample size and whether the theoretical frequency in each cell was greater than 5. A difference was considered statistically significant at $P<.05$.

Ethical Considerations

The data mentioned above are publicly accessible, and none of the data were used to trace personal identification details such as ID numbers, birth dates, or IP addresses. The authors sought advice from the institutional review board of the respective institution and were informed that no approvals were required.

Results

Content Construction: Features of Entities, Content, and Emotional Expression

Involved Entity: The Vast Majority of Rumor Information Dissemination Entities Were Relatively Clear

Regarding the entities involved in rumor content, the relevant entities that society was widely concerned about such as viruses, patients, transportation departments, enterprises, schools, hospitals, and government agencies were mentioned in the rumors during the epidemic period. Comparing the keyword clouds of rumor dissemination before and after the epidemic outbreak (Multimedia Appendix 1), rumors in both periods mainly targeted "nucleic acid testing," "hospitals," "patients," and other entities. However, after the epidemic was lifted, rumors related to "virus" and "vaccines" increased relatively, which may have been caused by the adjustment of epidemic prevention and control policies and the public's panic about the uncertain direction of virus mutation. Among them, "virus" and "patients" had a strong correlation, and the rumors related to "virus" mainly focused on its development status and prevention and treatment, such as "COVID-19 is not a severe flu, it is a minor AIDS," "Delta and other variants are circulating in Beijing's northern region," and "Delta and Omicron are co-circulating and may recombine." Rumors related to "patients" were closely related to the occurrence and development process of COVID-19, including prevention, treatment, rehabilitation, and various stages of patient symptoms and behavioral dynamics: for example, "A COVID-19 positive patient in Dadixiang Township died after taking a bath during the recovery period" and "Many people who relapsed were infected again because they forgot to change their toothbrush after their test results turned negative." Such rumors make it difficult for the public to distinguish between true and false information, causing great panic in society.

Content Dissemination: The Content Themes Mainly Focused on the Areas Closely Related to the Public's Vital Interests in the Context of the Epidemic

Analysis of the content themes of rumor samples showed that during the COVID-19 pandemic, rumors can be mainly categorized into 4 types: the pandemic situation, strange occurrences related to the pandemic, knowledge about preventing the spread of the virus, and measures to prevent the spread of the virus. Among them, the pandemic prevention and control measures were the most frequent topic, accounting for 78.3% (526/672) of all content, as shown in Table 4. The difference in content themes between the pre- and postlockdown periods of the pandemic was statistically significant ($\chi^2_4=29.374$, $P<.001$). Comparing the distribution of content themes before and after the lockdown period (Multimedia Appendix 2), it was found that rumors about measures to prevent the spread of the virus dominated in both periods. However, during the high-intensity outbreak period, there was a relative increase in rumors about knowledge related to prevention and a decrease in rumors about pandemic prevention and control measures. This indicated that in situations where the prevention and control mechanism changed and the uncertainty of the pandemic increased, people were more inclined to obtain information about preventing the virus rather than general information about the pandemic. In addition, it was observed that the specific content of rumors was concentrated in areas of public interest during the COVID-19 pandemic, such as silent management during the pandemic, city lockdown, traffic control, and delayed resumption of work and school. Such rumors were particularly prominent. COVID-19–related rumors also took various forms, including fabrication, tampering, and misinterpretation, and were generated for various reasons, such as to attract attention, hostility, or subjective conjecture.

XSL•FO
RenderX

Table 4. Overall distribution of epidemic rumors.

Content theme	Rumors (N=672), n (%)
Epidemic situation	133 (19.8)
Epidemic oddities	114 (17)
Epidemic prevention knowledge	56 (8.3)
Epidemic prevention measures	279 (41.5)
Others	90 (13.4)

Emotional Expression: Rumors About the Pandemic Showed Significant Differences in Emotional Expression, With Many Trending Toward Negative Emotions

Different themes of COVID-19 rumors had distinct emotional preferences (Figure 1). Rumors related to the pandemic situation tended to express fear and concern, whereas those related to bizarre incidents tended to express surprise. Rumors related to knowledge and measures of epidemic prevention tended to express approval, whereas those related to epidemic prevention measures not only showed a tendency toward fear and concern but also aimed to satisfy the public's need for information without any clear emotional expression. The difference in emotional expression between rumors before and after the pandemic was statistically significant (χ^2_8=28.924, P<.001). Comparing the emotional expressions of rumors before and after the pandemic (Figure 2), it was observed that during the initial stage of the new epidemic prevention and control period, the proportion of fear, concern, anger, and surprise expressed in pandemic rumors decreased, whereas the proportion of sadness and questioning increased substantially. At the same time, the proportion of rumors that expressed approval also increased substantially. Moreover, both before and after the pandemic, pandemic rumors tended to express negative emotions (291/577, 50.4% and 40/95, 42%, respectively). More than half of the negative emotions tended to arouse the fear and concern of the public (173/291, 59.4% and 21/40, 52%, respectively), such as "asymptomatic infections will cause a new outbreak of the pandemic." Such rumors often quickly triggered the fear of the public about the severity of the pandemic, and in an environment where everyone was at risk, people tended to trust them without much discernment. In contrast, rumors related to epidemic prevention knowledge mainly expressed positive emotions and idealized "wishes" for virus prevention and treatment, with the proportion of approval emotions accounting for nearly two-thirds of rumors (20/33, 61% and 13/23, 57%, respectively), such as "drinking garlic water can cure COVID-19, so we should drink more garlic water." Such rumors often contained positive action hints.

Figure 1. Heatmap of overall sentiment expression in COVID-19 rumor content.

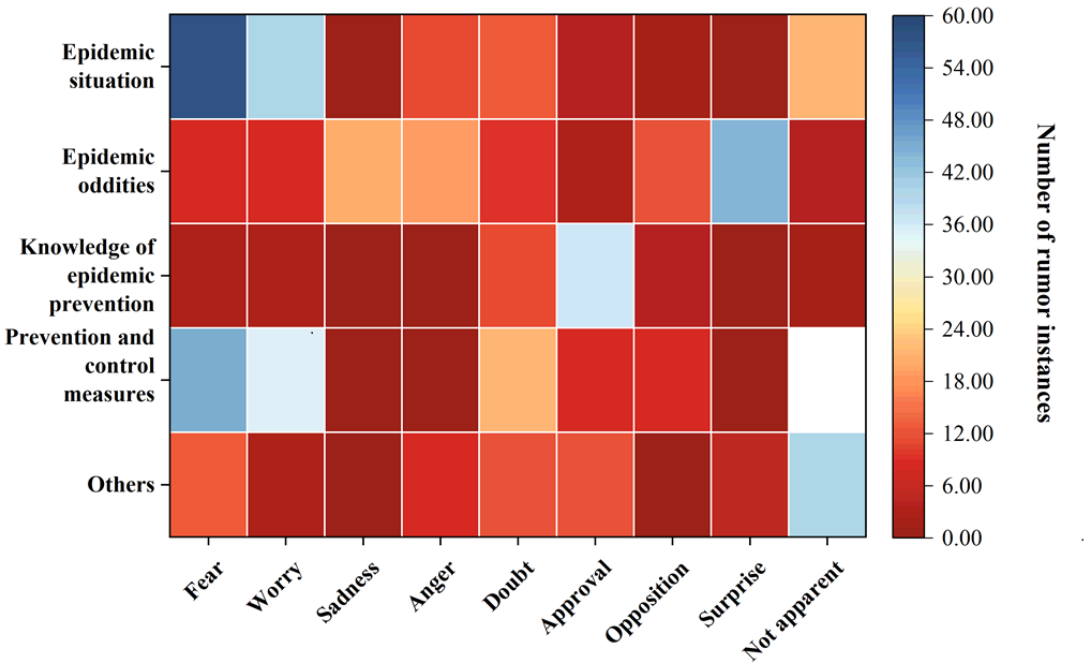

XSL•FO
RenderX

Figure 2. Comparison of emotional expressions in rumors before and after the relaxation of the epidemic prevention and control measures.

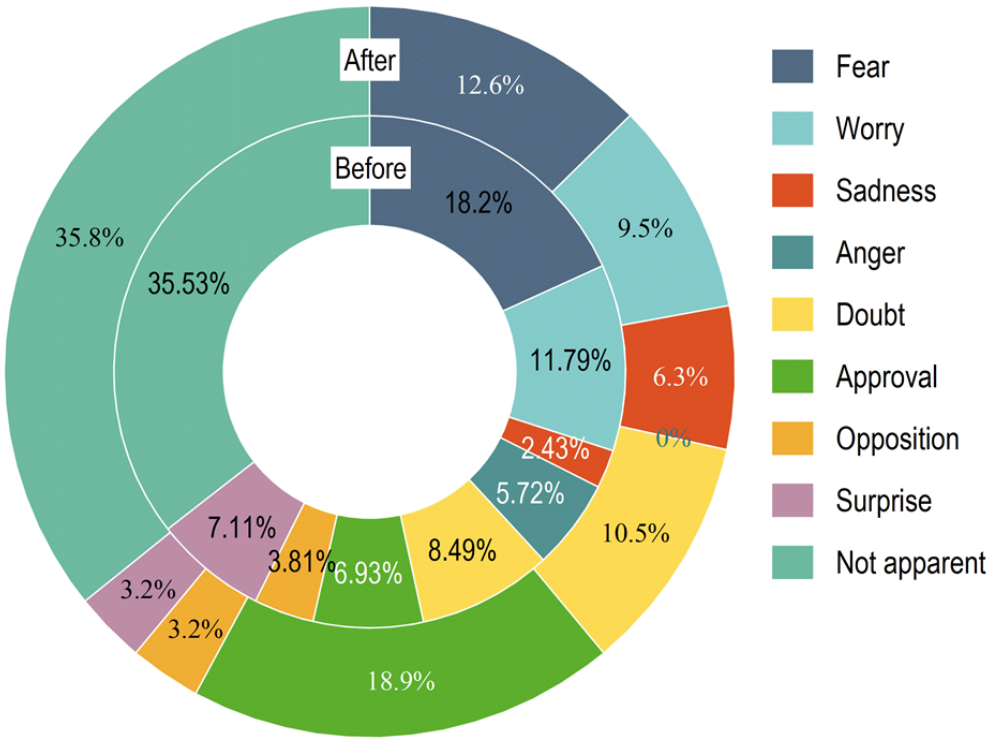

Process of Rumor Transmission: Rumors Were Mainly Spread Through Social Media Interactions and Community Transmission but Lacked a Factual Basis in the Content

The spread of COVID-19 rumors mainly relied on various social media platforms, and after excluding rumor samples with unknown dissemination channels, it was found that most COVID-19 rumors (306/672, 45.5%) before and after the lifting of restrictions were spread through WeChat groups and Moments (Figure 3). The differences in the dissemination channels and forms of rumors between the 2 periods were statistically significant (χ^2_5=15.498, P=.01; χ^2_3=33.510, P<.001). Among the known forms of rumor samples, text- and image-based rumors were more common before and after the lifting of restrictions. Images and texts are more easily manipulated and distorted compared to videos, making them easier to fabricate. In addition, there were secondary COVID-19 rumors based on images and texts, which were widely spread through social media platforms such as Douyin, Weibo, Toutiao, and Xiaohongshu. At the same time, it is worth noting the phenomenon of using authoritative experts to spread false information on a large scale, such as “The First Hospital of Ningbo warmly reminds that COVID-19 is not an ordinary cold, and each infected person should rest for a month like postpartum women...” There were also a series of rumors fabricated using people’s respect for authoritative experts, such as “The chairman of Changzhou Red Cross claims that drinking water boiled with brown sugar, ginger, scallion, and garlic can almost eliminate the chance of virus infection.” These types of rumors are often titled with eye-catching headlines such as “Emergency Notification from the Community” or “Important Reminder” or were presented in the tone of official announcements such as “Notice from the Joint Prevention and Control Office of COVID-19 in the Community” or “Morning Assessment by the CDC.” They also included phrases such as “Just informing your relatives and family members” or “Forward it to the people you care about” to facilitate circulation and expand the range of dissemination through social relationships.

XSL•FO
RenderX

Figure 3. Distribution of channels and forms of rumor dissemination (A) before and (B) after the release of epidemic control measures.

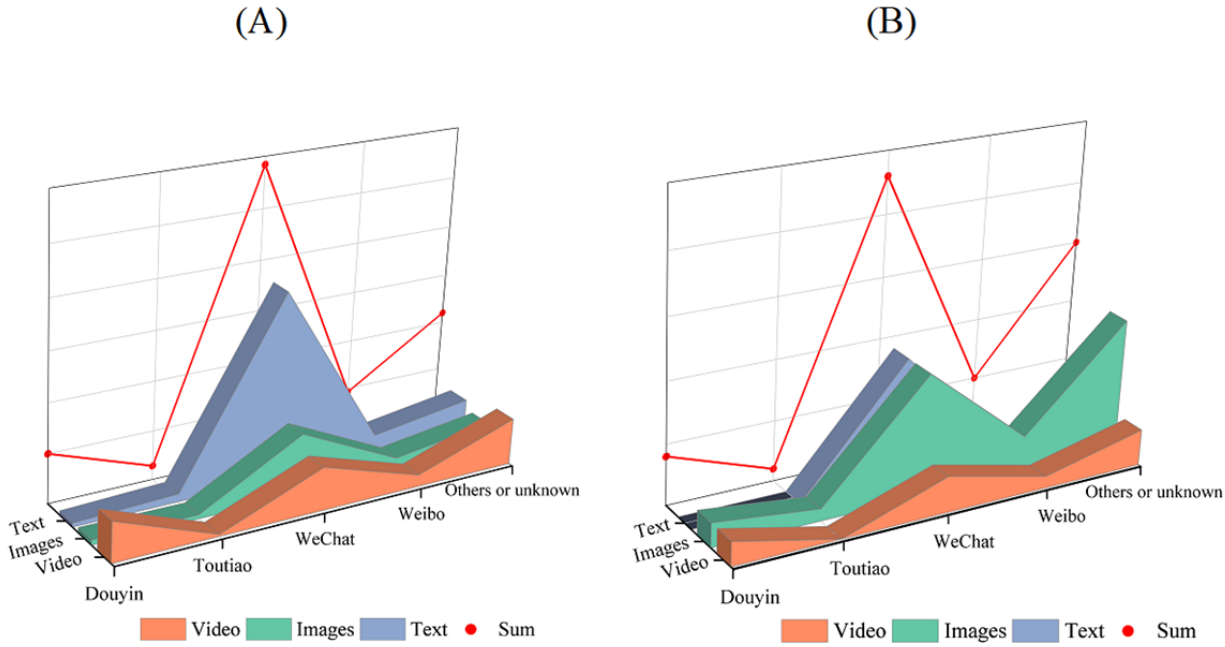

Response to Rumors: Different Rumor-Refuting Entities Had Content Preference Differences, and Frequent Rumors Reflected Poor Unobstructed Channels for Rumor Clarification

The main entities involved in debunking rumors were the media, government, and rumor-refuting platforms, among which the media played a leading role both before and after the release of pandemic control measures (Multimedia Appendix 3), followed by the government and rumor-refuting platforms. The difference in rumor debunking between the 2 periods (before and after the lifting of pandemic control measures) was statistically significant (χ^2_3=13.894, *P*=.003). By comparing the debunking entities and the content of rumor dissemination (Figure 4), it was found that all 3 entities tended to focus on rumors related to epidemic prevention measures. In addition, compared with the government and rumor-refuting platforms, the media was the main force in addressing and debunking rumors related to epidemic prevention measures, the epidemic situation, and epidemic prevention knowledge. The government, as the overall manager of epidemic prevention and control, can effectively intervene in epidemic-related rumors at the macro level and stabilize social emotions. The media can collaborate with official authorities to obtain firsthand information on epidemic developments, such as virus transmission, confirmed patients, and quarantine treatment, while also meeting the public's need for epidemic prevention information. In addition, tracing and verifying rumors by contacting relevant departments or individuals is currently the main method of rumor management and response. The National Health Commission, Centers for Disease Control and Prevention, and transportation departments, as well as authoritative experts, are the main channels for verification, along with enterprises, schools, hospitals, and individuals. Although the intervention of debunking entities and the use of debunking channels to verify rumors can effectively curb the spread of rumors, the emergence of old rumors in new forms also reflects the phenomenon of unsmooth debunking channels to a certain extent.

XSL•FO
RenderX

Figure 4. Radar chart of rumor-debunking entities and content categories. The 5 rumor content categories are S1: epidemic situation, S2: epidemic oddities, S3: epidemic oddities, S4: epidemic prevention measures, and S5: others.

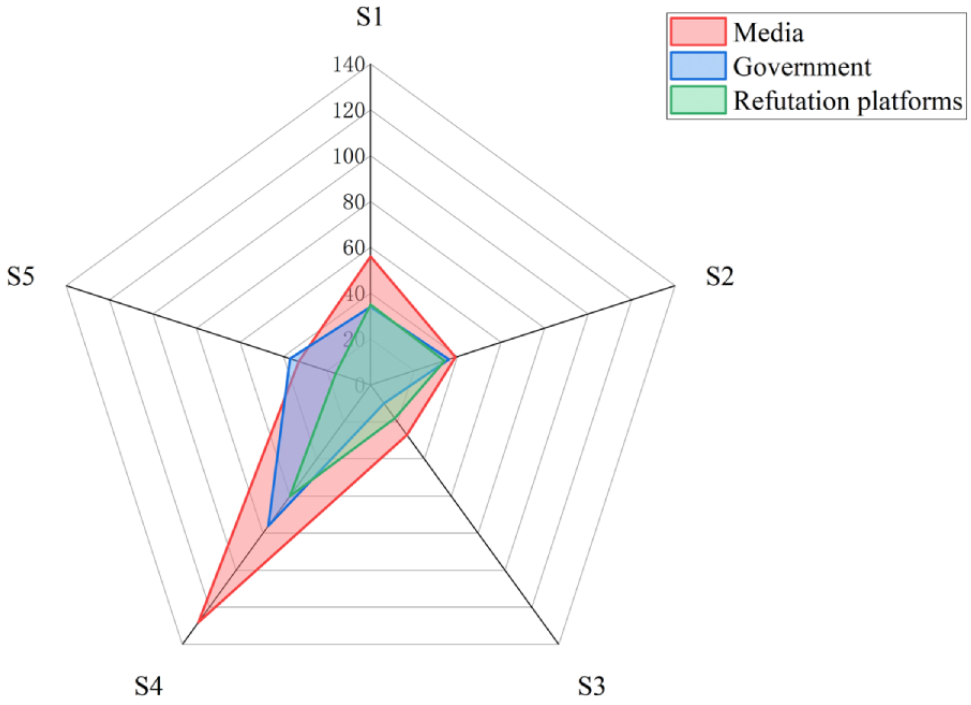

Discussion

Principal Findings

This paper discussed the characteristics and refutation mechanisms of COVID-19–related, web-based rumors before and after the lifting of pandemic restrictions, covering aspects such as rumor content construction, dissemination, and refutation responses. The study found that most rumors clearly targeted specific entities, with content themes closely related to the public's immediate interests in the pandemic domain. Rumor dissemination often showed significant emotional expression differences, which were predominantly negative. Rumors were mainly spread through social media interactions within communities and circles, with the text lacking a factual basis. Additionally, the entities refuting rumors showed content preference differences, indicating inefficiencies in rumor-refuting channels.

Interpretation of Findings

The content themes of epidemic rumors focused on closely related areas of public interest. Relevant departments should attach importance to the public's interests and adopt targeted measures to prevent rumors. First, due to the lack of understanding of the epidemic situation, some members of the public are prone to fear, which leads to the spread of false information about the epidemic situation, thus causing damage to social order and instability in the social environment [50,51]. Second, due to differences in personal information literacy and unsmooth information communication channels [52], some members of the public are prone to listen to and believe in certain "folk remedies" for COVID-19 and even regard some fake experts' false epidemic prevention knowledge as a guiding principle. Some unscrupulous merchants take advantage of the public's epidemic prevention demands and fears to falsely advertise the efficacy of drugs, leading to shortages of some drugs and making it difficult for those who truly need them to obtain them, thus resulting in market chaos [53]. Therefore, relevant departments should proactively understand and grasp the public's needs and develop differentiated rumor prevention and control plans. Authoritative departments and senior experts should strengthen communication with the public to reduce information asymmetry that leads to the proliferation of rumors [54]. Meanwhile, official releases and authoritative information exchange should be used to compress the space for the spread of rumors [44].

Public emotions are fuel for the spread of rumors during the COVID-19 outbreak, and it is important to assess the risks of public emotions and guide them through media cooperation [55]. During the outbreak, negative emotions such as panic, doubt, opposition, and concern among the public were widespread, and rumors could easily fuel negative emotions, leading to impulsive and irrational behavior [56]. Relevant departments should establish an effective system for assessing social emotional risks, grasp the psychological trends of the public, and provide psychological counseling to mitigate negative emotions and prevent their spread, thereby avoiding serious social harm. At the same time, it is important to also recognize the impact of rumors on positive emotions and take appropriate measures to prevent them based on different emotional responses. Additionally, media coverage plays an important role in guiding public emotions. On the one hand, media coverage can help the public clarify information, correct information biases, and meet their cognitive needs; on the other hand, it can also become a breeding ground for rumors and an ignition source for public opinion [57]. Therefore, relevant departments should work with the media to effectively guide

XSL•FO
RenderX

public opinion, create a positive social atmosphere, interact positively with the public, and promote the smooth implementation of COVID-19 prevention and rumor control work.

The spread of rumors often relies on strong ties in social networks [58]. Therefore, it is necessary to enhance the precision of rumor governance and encourage various social actors to actively participate in rumor control. Social media is a major channel for emotional communication and information sharing among the public, and emotional transmission within interpersonal networks can easily cause group resonance. Social networks with strong ties such as WeChat groups and friend circles have become easy platforms for public opinion fermentation and rumor proliferation. In the prevention and control of rumors, relevant departments should attach importance to community and circle transmissions, closely monitor the trend of public opinion, respond in a timely manner to topics that have aroused public discussion, and conduct proper tracing and verification of information, to timely cut off the spread of rumors and respond to them [59]. At the same time, a multiparty joint rumor control mechanism should be established, and relevant government departments should provide corresponding communication channels and upload and issue relevant information in a timely manner [60]. Authoritative experts should proactively communicate with the public in a timely manner to crush false rumors. Relevant media should also collaborate with authoritative officials, fulfill their social responsibilities, fully play the role of public opinion propaganda, and timely meet the information needs of the public.

Limitations

However, this study still has certain limitations. On the one hand, with the development of internet technology, the spread and governance of network rumors have become more complex. Future research needs to further consider the characteristics of communication subjects and network users, establish more detailed and comprehensive rumor propagation models, and identify more targeted debunking mechanisms. On the other hand, as the pandemic was still in its early stages, the sample size of postpandemic rumors included in this study was relatively small and static. Therefore, future research can explore ways to incorporate more relevant data to establish a complex and timely dynamic model. Furthermore, despite the call in this paper for relevant authorities to closely monitor the trends in public opinion and actively carry out fact-checking, there is currently a lack of clear standards to determine which information should be subject to verification and which may not require detailed scrutiny. In the future, there is a need for in-depth research to establish clear and practical debunking standards. This includes specifying the types of information more likely to capture public attention, identifying information that may pose threats to public safety, and determining which sources are more likely to disseminate accurate information. This approach aims to enhance the effectiveness of debunking efforts and optimize resource use, ensuring that debunking work is more targeted and efficient.

Conclusions

As a sudden and serious public health emergency of international concern, the COVID-19 pandemic has caused significant harm to social order. Through an analysis of rumors during the COVID-19 pandemic, it is evident that the themes of these rumors revolve around the immediate interests of the public, requiring targeted measures from relevant authorities. Additionally, recognizing the role of public emotions as an accelerant for rumor dissemination, there is a need for risk assessment and media-driven emotional guidance. Rumors often exploit strong social networks for propagation, emphasizing the necessity to enhance precise rumor governance and encourage diverse societal participation. In conclusion, it is emphasized that authorities should prioritize communication with the public, conduct emotional risk assessments, and establish collaborative rumor control mechanisms. This study provides a theoretical basis and effective approach to enhance debunking mechanisms during public health emergencies.

Acknowledgments

Funding was provided by the Hunan Province Philosophy and Social Science Foundation of China (grant 22YBQ014).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Word cloud maps of entities involved in rumor propagation (A) before and (B) after the release of epidemic control measures.
[PNG File , 604 KB - jmir_v26i1e48564_app1.png]

Multimedia Appendix 2
Comparison of the distribution of epidemic rumors before and after the relaxation of the epidemic prevention and control measures.
[PNG File , 33 KB - jmir_v26i1e48564_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3
Comparison of rumor-debunking entities before and after the relaxation of the epidemic prevention and control measures.
[PNG File , 40 KB - jmir_v26i1e48564_app3.png]

References

1. Zhang W, Yang L. Multidimensional dialogue: information dissemination and governance of public health emergencies [Article in Chinese]. Journal of Shandong University 2020 Sep 15;242(05):24-30. [doi: 10.19836/j.cnki.37-1100/c.2020.05.003]
2. Chen R. The hazards and prevention of rumor propagation in the era of online media [Article in Chinese]. Journal of Inner Mongolia Radio & TV University 2012 Jul;134(04):36-38. [doi: 10.16162/j.issn.1672-3473.2012.04.019]
3. Ran Y. Governance of network rumors in public health emergencies [Article in Chinese]. Party School of Chongqing Municipal Committee of CPC 2021 Oct(05):6. [doi: 10.27762/d.cnki.gzgcq.2021.000046]
4. Knapp RH. A psychology of rumor. Public Opin Q 1944;8(1):22-37. [doi: 10.1086/265665]
5. Ming J. Research on rumor spreading in emergency public events in social media environment [Article in Chinese]. Shanghai Normal University 2018 May(08):1-2 [accessed 2023-01-11] [FREE Full text]
6. Sun XT. Psychological mechanisms and governance strategies of health-related online rumors [Article in Chinese]. All-Media Explorations 2022 Jun;10(06):65-66 [accessed 2023-01-26] [FREE Full text]
7. Wardle C. Fake news. it's complicated. Medium. 2017 Feb 17. URL: https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79 [accessed 2023-01-11]
8. Tandoc EC. The facts of fake news: a research review. Sociol Compass 2019 Sep;13(9):e12724. [doi: 10.1111/soc4.12724]
9. Tandoc EC, Lim ZW, Ling R. Defining “fake news”. Digit Journal (Abingdon) 2017 Aug 30;6(2):137-153. [doi: 10.1080/21670811.2017.1360143]
10. Wu QQ, Shi SP, Lin W, Zhai HL. Research on the characteristics and strategies of online rumors during the epidemic period: based on the content analysis of rumors during the COVID-19 epidemic [Article in Chinese]. Public Communication of Science & Technology 2021 Aug;13(15):150-152. [doi: 10.16607/j.cnki.1674-6708.2021.15.049]
11. Ma XP. Rumor propagation and government governance in social media [Article in Chinese]. Journal of the Yinchuan Municipal Party College of CPC 2022 Jun;121(03):91-96 [accessed 2023-01-14] [FREE Full text]
12. Bordia P, Rosnow RL. Rumor rest stops on the information highway: transmission patterns in a computer-mediated rumor chain. Human Comm Res 1998 Dec;25(2):163-179. [doi: 10.1111/j.1468-2958.1998.tb00441.x]
13. Wang Y. Analysis of the characteristics, formation reasons, and governance strategies of "information epidemic" [Article in Chinese]. Business Culture 2021 Jul;509(20):142-144 [accessed 2023-01-14]. [doi: 10.27012/d.cnki.gdhuu.2021.001051]
14. DiFonzo N. Rumour research can douse digital wildfires. Nature 2013 Jan 10;493(7431):135. [doi: 10.1038/493135a] [Medline: 23302823]
15. Allport GW, Postman LJ. The basic psychology of rumor. Trans NY Acad Sci 1945 Dec;8:61-81. [doi: 10.1111/j.2164-0947.1945.tb00216.x] [Medline: 21012248]
16. Zhang HG, Bai G, Mao QL. The dissemination mode of online rumors during the COVID-19 epidemic [Article in Chinese]. Young Journalist 2020 Jun;674(18):44-45. [doi: 10.15997/j.cnki.qnjz.2020.18.019]
17. Tong L, Yu H. Characteristics of online rumor propagation, propagation modes, and dissolution mechanisms [Article in Chinese]. Hebei Enterprises 2022 Nov 10;400(11):96-98. [doi: 10.19885/j.cnki.hbqy.2022.11.017]
18. Xi X, Yong H. Research on the dynamics of opinion spread based on social network services [Article in Chinese]. Acta Phys Sin 2012;61(15):150509. [doi: 10.7498/aps.61.150509]
19. Han S, Zhuang FZ, He Q, Shi Z, Ao X. Energy model for rumor propagation on social networks. Physica A: Statistical Mechanics and its Applications 2014 Jan 15;394:99-109. [doi: 10.1016/j.physa.2013.10.003]
20. Wu RM, Han YL, Guo KY, Li YB. A rumor propagation model under recommendation mode [Article in Chinese]. Cyber Security and Data Governance 2022 Jul 15;41(07):56-62. [doi: 10.20044/j.csdg.2097-1788.2022.01.009]
21. Fang YQ, Yang LJ. Empirical analysis and model construction of Weibo rumor propagation under real-name system in network background [Article in Chinese]. Communication and Copyright 2017 Nov 15;54(11):164-166. [doi: 10.16852/j.cnki.45-1390/g2.2017.11.062]
22. Li CR, Ma ZJ, Wang YY. Dynamics of a delayed rumor spreading model with discontinuous threshold control. Heliyon 2022 Oct;8(10):e11231 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e11231] [Medline: 36311360]
23. Zou W, Xie J. An SI epidemic model with nonlinear infection rate and stage structure. Int J Biomath 2011 Nov 20;02(01):19-27. [doi: 10.1142/s1793524509000479]
24. Lu ZH, Gao SJ, Chen LS. Analysis of an SI epidemic model with nonlinear transmission and stage structure. Acta Math Sci 2003 Oct;23(4):440-446. [doi: 10.1016/s0252-9602(17)30486-1]
25. Zhao L, Huo HF. Spatial propagation for a reaction-diffusion SI epidemic model with vertical transmission. Math Biosci Eng 2021 Jul 05;18(5):6012-6033 [FREE Full text] [doi: 10.3934/mbe.2021301] [Medline: 34517521]
26. Qin Y, Zhong XX, Jiang H, Ye Y. An environment aware epidemic spreading model and immune strategy in complex networks. Appl Math Comput 2015 Jun 15;261:206-215. [doi: 10.1016/j.amc.2015.03.110]

XSL•FO
RenderX

27. Zhang FQ, Li JQ, Li J. Epidemic characteristics of two classic SIS models with disease-induced death. J Theor Biol 2017 Jul 07;424:73-83. [doi: 10.1016/j.jtbi.2017.04.029] [Medline: 28479003]
28. Neal P. Endemic behaviour of SIS epidemics with general infectious period distributions. Adv Appl Probab 2016 Feb 22;46(01):241-255. [doi: 10.1017/s0001867800007023]
29. Enatsu Y, Nakata Y, Muroya Y. Global stability for a class of discrete SIR epidemic models. Math Biosci Eng 2010 Apr;7(2):347-361 [FREE Full text] [doi: 10.3934/mbe.2010.7.347] [Medline: 20462293]
30. Cadoni M, Gaeta G. Size and timescale of epidemics in the SIR framework. Physica D 2020 Oct;411:132626 [FREE Full text] [doi: 10.1016/j.physd.2020.132626] [Medline: 32834247]
31. Acedo L, González-Parra G, Arenas AJ. An exact global solution for the classical SIRS epidemic model. Nonlinear Anal Real World Appl 2010 Jun;11(3):1819-1825. [doi: 10.1016/j.nonrwa.2009.04.007]
32. Qiu ZZ, Sun YY, He X, Wei J, Zhou R, Bai J, et al. Application of genetic algorithm combined with improved SEIR model in predicting the epidemic trend of COVID-19, China. Sci Rep 2022 May 26;12(1):8910 [FREE Full text] [doi: 10.1038/s41598-022-12958-z] [Medline: 35618751]
33. Pongsumpun P, Kongnuy R, López DG, Tang I, Dubois MA. Contact infection spread in an SEIR model: an analytical approach. ScienceAsia 2013 Jun 6;39(4):410-415. [doi: 10.2306/scienceasia1513-1874.2013.39.410]
34. Wang WB, Xia ZY. Study of COVID-19 epidemic control capability and emergency management strategy based on optimized SEIR model. Mathematics 2023 Jan 08;11(2):323. [doi: 10.3390/math11020323]
35. Wu Y. Research on rumor dissemination and tracing based on extended epidemic model [Article in Chinese]. Donghua University 2021 Jun 15(03):7-9. [doi: 10.27012/d.cnki.gdhuu.2021.001051]
36. Zhu LH, Yang F, Guan G, Zhang Z. Modeling the dynamics of rumor diffusion over complex networks. Information Sciences 2021 Jul;562:240-258. [doi: 10.1016/j.ins.2020.12.071]
37. Sun H, Sheng YH, Cui Q. An uncertain SIR rumor spreading model. Adv Differ Equ 2021 Jun 12;2021(1):286. [doi: 10.1186/s13662-021-03386-w]
38. Hosseini S, Zandvakili A. Information dissemination modeling based on rumor propagation in online social networks with fuzzy logic. Soc Netw Anal Min 2022 Feb 07;12(1):34. [doi: 10.1007/s13278-022-00859-y]
39. Chen JH, Ma HC, Yang S. SEIOR rumor propagation model considering hesitating mechanism and different rumor-refuting ways in complex networks. Mathematics 2023 Jan 05;11(2):283. [doi: 10.3390/math11020283]
40. Tao L. Innovative governance mechanisms for online rumors during public health emergencies: a comparative analysis of rumors during the SARS and COVID-19 outbreaks [Article in Chinese]. Future Communication 2021 Apr 20;28(02):59-68. [doi: 10.13628/j.cnki.zjcmxb.2021.02.008]
41. Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, et al. The science of fake news. Science 2018 Mar 09;359(6380):1094-1096. [doi: 10.1126/science.aao2998] [Medline: 29590025]
42. Thorson E, Sheble L, Southwell BG. Conclusion: an agenda for misinformation research. In: Southwell BG, Thorson EA, Sheble L, editors. Misinformation and Mass audiences. Austin, TX: University of Texas Press; 2018:289-293.
43. Wang X. Mitigating "knowledge risk": research on the mechanism of refuting rumors by "fact-checking" in the post-epidemic era [Article in Chinese]. Public Communication of Science & Technology 2022 Feb;14(04):48-51. [doi: 10.16607/j.cnki.1674-6708.2022.04.020]
44. Zhang Y, Xu JP, Wu Y. A rumor control competition model considering intervention of the official rumor-refuting information. Int J Mod Phys C 2020 Aug 19;31(09):2050123. [doi: 10.1142/s0129183120501235]
45. Chen B, Chen XY, Pan J, Liu K, Xie B, Wang W, et al. Dissemination and refutation of rumors during the COVID-19 outbreak in China: infodemiology study. J Med Internet Res 2021 Feb 15;23(2):e22427 [FREE Full text] [doi: 10.2196/22427] [Medline: 33493124]
46. Shu T. Analysis of rumor propagation in public emergencies based on the 5W model: a content analysis of 513 online rumors [Article in Chinese]. China Newspaper Industry 2021 Oct 25;525(20):62-64. [doi: 10.13854/j.cnki.cni.2021.20.029]
47. Zhou MJ, Du YX. Rumor propagation and collaborative governance mechanism during major public emergencies: a study of rumors during the COVID-19 epidemic [Article in Chinese]. Media Observer 2020 Sep 10;441(09):15-23. [doi: 10.19480/j.cnki.cmgc.2020.09.002]
48. Liu C, Xie YG, Wan XA. Relationship between Weibo emotions and Weibo dissemination: an empirical analysis based on 24 public events [Article in Chinese]. Journalism & Communication 2015 Sep 25;22(09):92-106 [accessed 2023-01-23] [FREE Full text]
49. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas 2007 Apr;1(1):77-89. [doi: 10.1080/19312450709336664]
50. Ho KKW, Chan JY, Chiu DKW. Fake news and misinformation during the pandemic: what we know and what we do not know. IT Prof 2022 Mar 1;24(2):19-24. [doi: 10.1109/mitp.2022.3142814]
51. Barrientos-Báez A, Martínez-Sala AM, Altamirano VP, Caldevilla Domínguez D. Fake news: la pandemia de la COVID-19 y su cronología en el sector turístico [Article in Spanish]. Hist Comun Soc 2021 Feb 10;26(Especial):135-148. [doi: 10.5209/hics.74248]
52. Gerosa T, Gui M, Hargittai E. (Mis)informed during COVID-19: how education level and information sources contribute to knowledge gaps. International Journal of Communication 2021;15:2196-2217. [doi: 10.5167/UZH-204498]

XSL•FO
RenderX

53. Nanfang Daily. Severely punish false advertising that plays the "epidemic card" [Article in Chinese]. Credit China. 2020 Apr 1. URL: https://www.creditchina.gov.cn/home/zhuantizhuanlan/fengxiantishi/xinyongfengxiandajiatan/202004/t20200401_190338.html [accessed 2023-01-23]
54. Li H. Communication for coproduction: increasing information credibility to fight the coronavirus. Am Rev Public Adm 2020 Jul 16;50(6-7):692-697. [doi: 10.1177/0275074020942104]
55. Dong W, Tao J, Xia X, Ye L, Xu H, Jiang P, et al. Public emotions and rumors spread during the COVID-19 epidemic in China: web-based correlation study. J Med Internet Res 2020 Nov 25;22(11):e21933 [FREE Full text] [doi: 10.2196/21933] [Medline: 33112757]
56. Rezapour M, Dehzangi A, Saadati F. Students' negative emotions and their rational and irrational behaviors during COVID-19 outbreak. PLoS One 2022;17(3):e0264985 [FREE Full text] [doi: 10.1371/journal.pone.0264985] [Medline: 35255106]
57. Simanjuntak S. Memahami manajemen media massa untuk publikasi artikel bagi penulis pemula [Article in Indonesian]. Tadbir 2021 Dec 30;3(2):229-250. [doi: 10.24952/tad.v3i2.4878]
58. Wang LL, Zhang LL. A study on the diffusion mechanism of strong relationship social network rumors in public emergencies: based on situational awareness theory [Article in Chinese]. Journal of Information Technology 2019;42(05):112-118 [FREE Full text]
59. Ahsan M, Kumari M, Sharma T. Rumors detection, verification and controlling mechanisms in online social networks: a survey. Online Soc Netw Media 2019 Nov;14:100050. [doi: 10.1016/j.osnem.2019.100050]
60. Zhao Y, Xu Y, Zhang H. Rumor data analysis and path research from the perspective of co-governance. 2021 Presented at: 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE); Mar 26-28, 2021; Nangchang, China p. 1030-1034. [doi: 10.1109/icbaie52039.2021.9389865]

Abbreviations

SEIOR: ignorant, hesitators, spreaders, rumor debunkers, and immunizers
SEIR: susceptible, exposed, infectious, and recovered
SI: susceptible and infectious
SIR: susceptible, infectious, and recovered
SIS: susceptible, infectious, and susceptible

Edited by A Mavragani; submitted 28.04.23; peer-reviewed by W Ceron, B Southwell; comments to author 22.12.23; revised version received 11.01.24; accepted 11.04.24; published 15.05.24.

Please cite as:
Liu X, Hu Q, Wang J, Wu X, Hu D
Difference in Rumor Dissemination and Debunking Before and After the Relaxation of COVID-19 Prevention and Control Measures in China: Infodemiology Study
J Med Internet Res 2024;26:e48564
URL: https://www.jmir.org/2024/1/e48564
doi:10.2196/48564
PMID:38748460

©Xiaoqi Liu, Qingyuan Hu, Jie Wang, Xusheng Wu, Dehua Hu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of Inpatient Electronic Consultations During the COVID-19 Crisis and Its Sustainability Beyond the Pandemic: Quality Improvement Study

Anna S Aledia[1], BS; Amish A Dangodara[1], MD; Aanya A Amin[1]; Alpesh N Amin[1], MD, MBA

Department of Medicine & Hospital Medicine, University of California, Irvine, Orange, CA, United States

Corresponding Author:
Alpesh N Amin, MD, MBA
Department of Medicine & Hospital Medicine
University of California, Irvine
333 City Boulevard West
Suite 500
Orange, CA, 92868
United States
Phone: 1 714 456 3785
Fax: 1 714 456 3871
Email: anamin@uci.edu

Abstract

Background: Limiting in-person contact was a key strategy for controlling the spread of the highly infectious novel coronavirus (COVID-19). To protect patients and staff from the risk of infection while providing continued access to necessary health care services, we implemented a new electronic consultation (e-consult) service that allowed referring providers to receive subspecialty consultations for patients who are hospitalized and do not require in-person evaluation by the specialist.

Objective: We aimed to assess the impact of implementing e-consults in the inpatient setting to reduce avoidable face-to-face referrals during the COVID-19 pandemic.

Methods: This quality improvement study evaluated all inpatient e-consults ordered from July 2020 to December 2022 at the University of California Irvine Medical Center. The impact of e-consults was assessed by evaluating use (eg, number of e-consults ordered), e-consult response times, and outcome of the e-consult requests (eg, resolved electronically or converted to the in-person evaluation of patient).

Results: There were 1543 inpatient e-consults ordered across 11 participating specialties. A total of 53.5% (n=826) of requests were addressed electronically, without the need for a formal in-person evaluation of the patient. The median time between ordering an e-consult and a specialist documenting recommendations in an e-consult note was 3.7 (IQR 1.3-8.2) hours across all specialties, contrasted with 7.3 (IQR 3.6-22.0) hours when converted to an in-person consult (P<.001). The monthly volume of e-consult requests increased, coinciding with surges of COVID-19 cases in California. After the peaks of the COVID-19 crisis subsided, the use of inpatient e-consults persisted at a rate well above the precrisis levels.

Conclusions: An inpatient e-consult service was successfully implemented, resulting in fewer unnecessary face-to-face consultations and significant reductions in the response times for consults requested on patients who are hospitalized and do not require an in-person evaluation. Thus, e-consults provided timely, efficient delivery of inpatient consultation services for appropriate problems while minimizing the risk of direct transmission of the COVID-19 virus between health care providers and patients. The service also demonstrated its value as a tool for effective inpatient care coordination beyond the peaks of the pandemic leading to the sustainability of service and value.

(J Med Internet Res 2024;26:e55623) doi:10.2196/55623

KEYWORDS

COVID-19; electronic consultation; eConsult; e-consult; inpatient; consultations; pandemic; infectious; novel coronavirus; coronavirus; patients; patient; staff; health care; health care providers; tool; tools; effectiveness

XSL•FO
RenderX

Introduction

When the novel coronavirus (COVID-19), the disease caused by SARS-CoV-2, began to quickly spread around the world, the high transmissibility of this disease urged health care systems to explore alternatives to face-to-face interactions that would reduce the risk of exposure for both the patient and the provider. Electronic consultations (e-consults) are asynchronous, non–face-to-face, provider-to-provider exchanges that have been shown to improve patient access to specialty care for appropriate referral problems that do not require an in-person evaluation of the patient by the specialist [1-3]. The rapid rise in COVID-19 cases induced a demand for the adoption of e-consult services and triggered an increase in the use of e-consults [4]. Although its use in the outpatient setting is well established [5,6], e-consults in the inpatient arena are relatively new.

As the only academic health system in the sixth largest county in the United States, University of California Irvine (UCI) Health has been a leader in the advancement of telehealth technologies that expand access to care and improve health care efficiency and resource use [7]. UCI already has a well-developed e-consults program in the ambulatory setting [8], and to complement this existing service, we expanded e-consults to patients who were hospitalized to further help reduce in-person contacts between consulting providers and patients, thereby minimizing disease transmission and conserving scarce personal protective equipment (PPE) during the COVID-19 crisis. Other health systems have implemented similar e-consult services for inpatients, but provider use of e-consults was temporary in response to the pandemic, favoring in-person consultative care instead [9]; inpatient e-consults were offered by only a single specialty consulting service [10,11]; and use cases involved early inpatient e-consult models [12,13]. In this study, we describe our rapid implementation of inpatient e-consults in multiple specialties and its sustained use beyond the peaks of the pandemic.

Methods

Setting

The inpatient e-consult service was implemented at the UCI Medical Center, a 478-bed acute care hospital providing tertiary and quaternary care, ambulatory and specialty medical clinics, behavioral health care, and rehabilitation services. Located in Orange County, California, it serves a diverse population of close to 4 million persons with broad health care needs. With more than 500 specialty and primary care physicians, UCI offers a full scope of acute and general care services. It is also the primary teaching location for UCI medical and nursing students, medical residents, and fellows, and it is home to Orange County's only adult level 1 and pediatric level 2 trauma centers, a National Cancer Institute–designated comprehensive cancer center, a regional burn center, the county's only hematopoietic stem cell and bone marrow transplant program, and the region's only high-risk perinatal and neonatal program and maternal-fetal transport system. In winter 2020, UCI Medical Center opened a temporary mobile field hospital that added up to 50 acute care beds in response to a surge of patients with COVID-19.

Implementation

The design and implementation of inpatient e-consults were guided by a steering committee, which included the Chair of the Department of Medicine and Executive Director of Hospital Medicine (who was the lead to design and develop e-consults at the UCI), a clinical informaticist, specialty physician leads, an IT build team, representatives from the Compliance and Privacy Office and Physicians Billing Group, and a project manager. Early on, members of this committee engaged UCI leadership to affirm support for the new service and obtain the IT resources needed to build the inpatient e-consults workflow. Regular steering committee meetings were established to discuss the design of the inpatient e-consults workflow and develop a process for provider reimbursement or credit. Prior to the go-live, the inpatient e-consult service was publicized by members of the steering committee through email communications with house staff. Steering committee members also hosted Zoom training (Zoom Technologies) and orientation sessions with participating consulting services, and they distributed tip sheets summarizing the steps to complete the requesting provider and responding consultant workflows.

Our IT team was able to efficiently implement our inpatient e-consult service by designing workflows similar to those for traditional in-person consults. Thus, the processes for requesting and responding to inpatient e-consults were not unfamiliar to providers (Figure 1). To request an e-consult, the inpatient service or team places a consult order in the electronic health record (EHR), indicating that the request is for an e-consult. The patient is then added to the physician e-consults system list of the appropriate specialty. A follow-up call or page is also sent to the specialty by the requesting team to alert the inpatient consulting team of the e-consult and, if necessary, provide them with any additional details. To respond to the e-consult request, the inpatient consulting team reviews the relevant clinical information available in the EHR and documents their assessment and recommendations in a consult note. If the case is deemed too complex to be addressed electronically, the consulting team converts the e-consult to a traditional in-person consultation and the patient is examined before documenting guidance in the EHR. The requesting provider and responding consultant are each credited with 0.7 work relative value units (a measure of the provider's time and effort required to perform the service) for every completed e-consult that does not result in an in-person evaluation of the patient by the consulting service, while usual billing or relative value unit credit applies for in-person consultations.

XSL•FO
RenderX

Figure 1. Workflows for requesting and responding to inpatient e-consults. e-consult: electronic consultation; EHR: electronic health record.

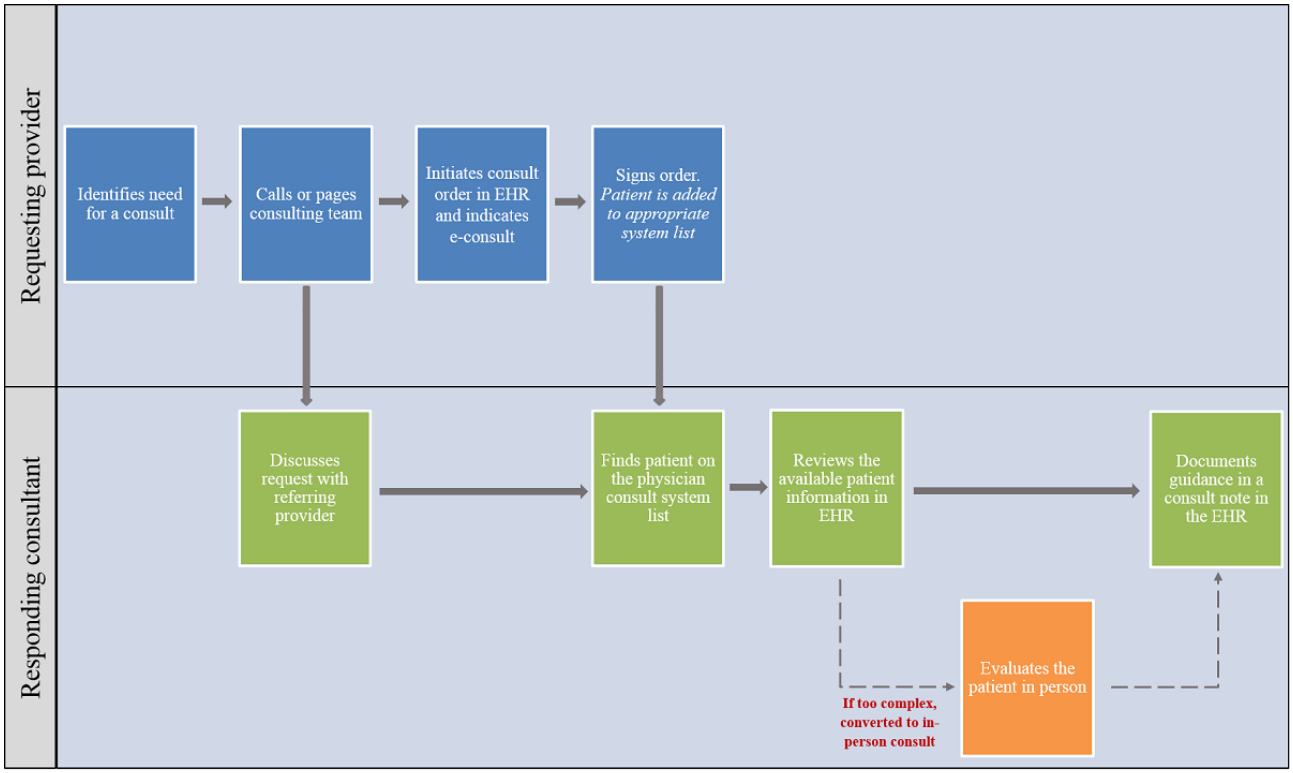

Measurement and Analysis

We conducted a retrospective evaluation of all inpatient e-consults ordered at the UCI Medical Center from implementation in July 2020 to December 2022 to assess use, outcomes, and response times. Use was tracked by examining the volume of e-consults ordered per specialty over the 2.5-year period and comparing it with the volume of traditional in-person consults ordered for the specialties offering inpatient e-consults. To assess outcomes, we categorized the result of each e-consult order as either "resolved electronically" if the consulting team addressed the request without a face-to-face evaluation of the patient or "converted to in-person" if the consulting team deemed the case too complex and the patient required a physical examination. The response time was defined as the interval between the documented consult order in the EHR and the consulting team filing recommendations in a consult note. We calculated the median response time and the IQR in hours for each specialty and performed nonparametric Mann-Whitney *U* tests in SPSS (version 28; IBM Corp) to compare the median response times of requests resolved electronically and converted to in person. All *P* values were 2-tailed, and $P<.05$ was considered statistically significant.

Ethical Considerations

Our implementation and retrospective analysis of the inpatient e-consults service constituted as quality improvement activities and not human subjects research. Thus, our study did not require institutional review board review. This study followed the Standards for Quality Improvement Reporting Excellence guidelines.

Results

UCI's e-consults service was launched in 11 total specialties (allergy and immunology, cardiology, dermatology, endocrinology, infectious diseases, nephrology, palliative care, pediatric endocrinology, pulmonary and critical care, radiation oncology, and rheumatology). Over a 2.5-year period, 1543 e-consults were requested out of 14,974 total consult orders (e-consult and traditional in-person consults) across the 11 participating specialties (Figure 2). Thus, the average proportion of consult orders requested as e-consults is 10.3%, although this proportion varied widely among participating specialties. The specialty with the lowest e-consult proportion was pulmonary and critical care, which had 1.5% (13/850) of total consult orders requested as e-consults, while the specialty with the highest e-consult proportion was pediatric endocrinology, which had 48% (12/25) of total consult orders requested as e-consults. However, with only 25 total consult orders, pediatric endocrinology had the fewest number of total consult orders of all participating specialties.

The most requested e-consult specialties were infectious diseases (which received 574/1543, 37.2% of the e-consult requests), cardiology (261/1543, 16.9% of the e-consult requests), endocrinology (229/1543, 14.8% of the e-consult requests), and dermatology (226/1543, 14.6% of the e-consult requests; Table 1). A total of 53.5% (826/1543) of e-consult requests across all participating specialties were addressed without the need for an in-person evaluation of the patient by the consulting team. The specialty with the fewest e-consult requests resolved electronically was pulmonary and critical care, which completed 0% (0/13) of requests electronically, while the specialty with the most e-consult requests resolved electronically was pediatric

XSL•FO
RenderX

endocrinology, which completed 100% (12/12) of e-consult requests, without needing to physically examine the patient. However, both specialties had the smallest volumes of e-consult requests of all participating specialties.

We found that the overall median response time of e-consult requests resolved electronically was significantly lower than requests converted to an in-person consultation (Figure 3). The median time between ordering an e-consult and a specialist documenting recommendations in a consult note was 3.7 (IQR 1.3-8.2) hours across all specialties when resolved electronically, contrasted with 7.3 (IQR 3.6-22.0) hours when converted to an in-person consult ($P<.001$). Over half (6/11, 55%) of the participating specialties had significantly faster median e-consult response times for requests resolved electronically compared to requests converted to an in-person consultation. The specialties with the fastest e-consult response times were dermatology and radiation oncology, which had median response times of 1.3 (IQR 0.4-3.0) hours and 0.9 (IQR 0.3-1.5) hours when resolved electronically, respectively. However, radiation oncology had one of the smallest volumes of e-consult requests among participating specialties.

The overall response times of e-consult requests were much faster than the turnaround goal mandated by our institutional guidelines, which require a same-day response by 8 PM if the consult is ordered before noon or a response by the following morning if ordered after noon. For reference, the overall median response time for completion of a traditional in-person consult by the same 11 specialties during the same 2.5-year period is 25.8 (IQR 10.8-65.7) hours (Multimedia Appendix 1). Thus, regardless of whether an e-consult request was resolved electronically or converted to an in-person consult, e-consults significantly improved the turnaround times for inpatient consultations.

The average volume of requests was 19 inpatient e-consults per month during the first 5 months that inpatient e-consults were live (Figure 4). Then, California experienced surges of COVID-19 cases throughout the pandemic and we saw corresponding increases in inpatient e-consults use. During the winter 2020 surge, the average volume of requests increased to 52 inpatient e-consults per month. Then, the Delta variant wave arrived in summer 2021, and the average volume of requests increased to 61 inpatient e-consults per month. When the Omicron variant wave emerged in winter 2021, the average volume of requests peaked at 75 inpatient e-consults per month. During a sustained wave in spring-summer 2022 driven by Omicron subvariants, the average volume of requests was 62 inpatient e-consults per month. After these surges subsided and COVID-19 cases declined, the use of inpatient e-consults remained at a high-level baseline with an average of 53 inpatient e-consults per month. Interestingly, similar patterns of increased e-consults use were observed in the ambulatory setting.

Figure 2. e-Consult and traditional in-person consult volume by specialty. Percentages of e-consults (blue) and traditional in-person consults (gray) requested by specialty are shown on the primary axis. Total consult orders (e-consult and traditional in-person consults) by specialty are shown on the secondary axis. e-consult: electronic consultation.

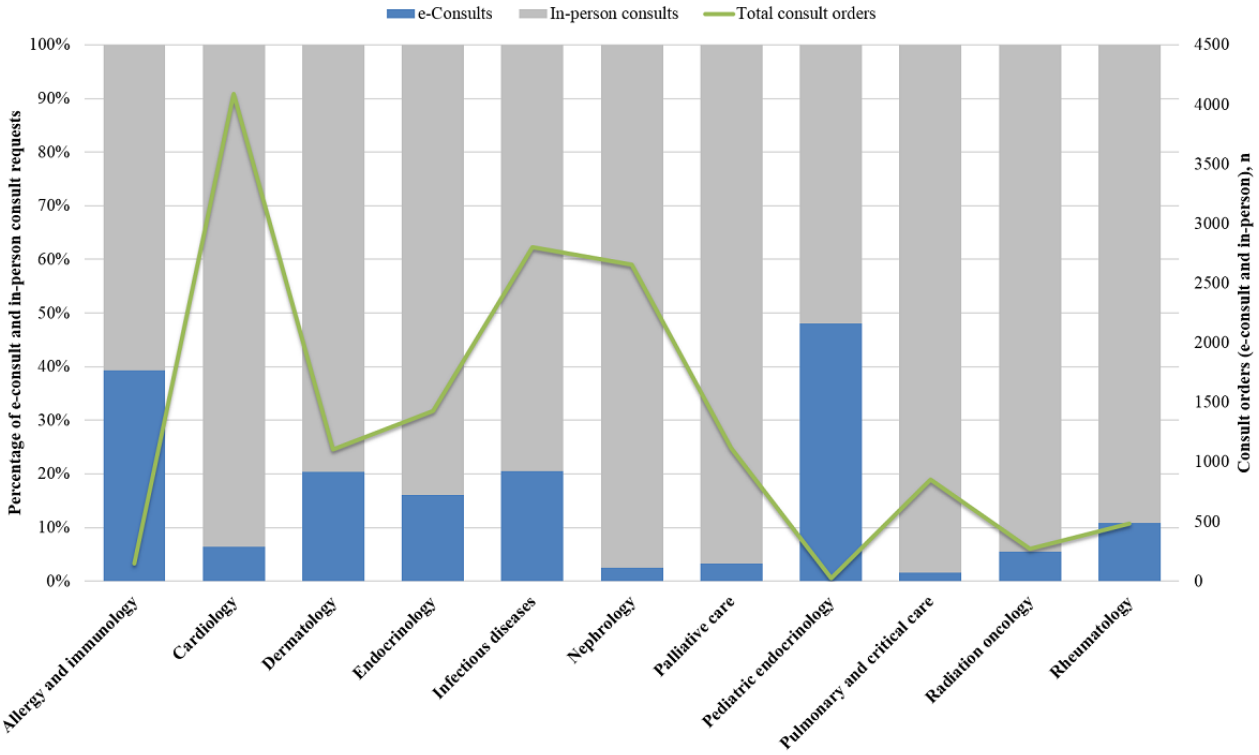

XSL•FO
RenderX

Table 1. Volume and outcomes of e-consult requests by specialty.

Specialty	e-Consult orders, N	Resolved electronically, n (%)	Converted to in-person, n (%)
All specialties	1543	826 (53.5)	717 (46.5)
Allergy and immunology	57	46 (80.7)	11 (19.3)
Cardiology	261	108 (41.4)	153 (58.6)
Dermatology	226	152 (67.3)	74 (32.7)
Endocrinology	229	140 (61.1)	89 (38.9)
Infectious diseases	574	335 (58.4)	239 (41.6)
Nephrology	68	2 (2.9)	66 (97.1)
Palliative care	36	1 (2.8)	35 (97.2)
Pediatric endocrinology	12	12 (100)	0 (0)
Pulmonary and critical care	13	0 (0)	13 (100)
Radiation oncology	15	4 (26.7)	11 (73.3)
Rheumatology	52	26 (50)	26 (50)

Figure 3. e-Consult response times by specialty. Boxplot of response times by specialty for e-consult requests resolved electronically (orange) and converted to in-person (teal). The box represents the IQR, with the top and bottom of box corresponding to the upper (third) and lower (first) quartiles, respectively. The horizontal line inside the box marks the median and the lines extending out from the box (whiskers) represent values within 1.5 times the IQR. For reference, the median response time for completion of a traditional in-person consult is marked by the dashed red line. *Indicates statistical significance.

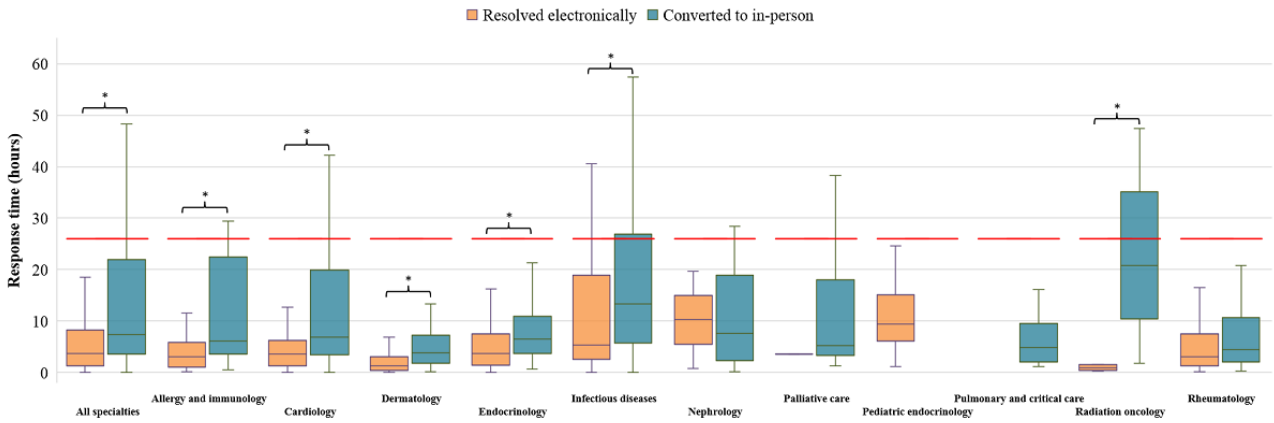

XSL•FO
RenderX

Figure 4. e-Consult volume over time. Monthly e-consult orders for ambulatory (green) and inpatient (red) settings at the University of California Irvine (UCI), 2015-2022. UCI transitioned from the Allscripts electronic health record platform to Epic in fall 2017. Launch dates for ambulatory and inpatient e-consults were April 2015 and July 2020, respectively. e-consult: electronic consultation; EHR: electronic health record.

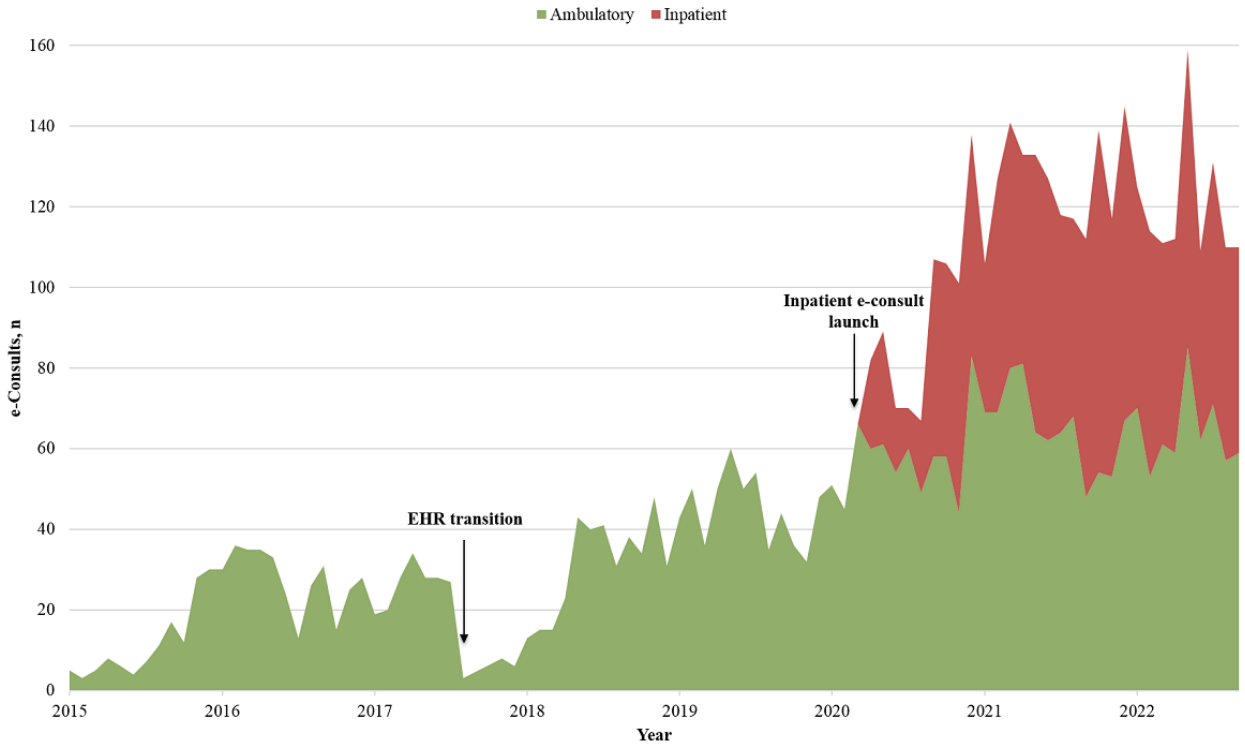

Discussion

Principal Findings

In response to the COVID-19 crisis, we successfully implemented an inpatient e-consult service that offered providers the option of requesting a subspecialty consultation for patients who are hospitalized and do not require an in-person evaluation by the specialist. Strong engagement by the clinical champions and technology partners in our steering committee, along with support from UCI's leadership which provided us with dedicated IT, compliance, and billing teams, contributed to the successful design and implementation of our inpatient e-consults service. In addition, we were able to rapidly launch the service by leveraging our experiences with implementing e-consults in the ambulatory setting and capitalizing on existing infrastructure for inpatient consults. Instead of creating unique e-consult orders, configuring our existing inpatient consult order reduced the build components for our IT team, allowing us to quickly and effectively launch the inpatient e-consults service. Because we used workflows similar to those for traditional in-person consults, the processes for requesting and responding to inpatient e-consults were not new for providers. This strategy, along with provider familiarity with our well-established e-consults service in the ambulatory setting, likely helped to foster the adoption of inpatient e-consults. Although the COVID-19 crisis provided the key stimulus, these factors may have also contributed to the more rapid adoption of inpatient e-consults in comparison to the initial uptake of our ambulatory e-consults.

We found that the e-consult services helped to significantly reduce the response time for consults requested on patients who are hospitalized and do not require an in-person evaluation. In fact, the overall median response time of e-consult requests resolved electronically was approximately half of the response time for requests converted to an in-person consultation and nearly 7 times faster than the response time for traditional in-person consults. This time saving was critical during surges of COVID-19 cases when emergency departments and inpatient units were overwhelmed, leading to prolonged wait times for patients who were hospitalized to receive consultative care. e-Consults helped to streamline the inpatient consultation process and enabled the consulting team to promptly and efficiently provide recommendations on patients not needing a physical examination.

Although some diagnoses require in-person evaluation of the patient, lower complexity problems can be managed effectively using e-consults. Indeed, we found that over half of e-consult requests were addressed electronically without the need for an in-person evaluation of the patient by the consulting team. By reducing unnecessary in-person consultations, e-consults likely helped to limit the use of scarce PPE; minimize disease transmission; and free up specialists for other activities, such as examining patients with more complicated conditions and performing procedures. This improved resource use may also translate to potential cost savings associated with avoided in-person consultations and increased productivity. Future work should aim to analyze the cost-effectiveness of inpatient e-consults.

After the peaks of the COVID-19 crisis subsided in California, we discovered that provider use of inpatient e-consults persisted at a rate well above the precrisis levels. This sustained use implies positive provider experiences with the service and

XSL•FO
RenderX

suggests a preference for e-consults when addressing lower complexity problems. Developing workflows for the inpatient e-consults service that were familiar to providers and significantly improving the turnaround times for inpatient consultations also likely helped to facilitate this sustainability. Thus, while case numbers and death rates associated with the COVID-19 pandemic have declined, e-consults continued to be an important part of our health care delivery.

Although relatively new, there have been a few reports of e-consults in the inpatient setting. The earliest examples involved the unexpected use of the ambulatory e-consult platform in the inpatient setting [13] and the design of an inpatient e-consult protocol that provided subspecialty consultations to inpatients at a remote hospital that lacked access to these clinical services [12]. Other reports described the feasibility and use of inpatient e-consults for only 1 specific specialty consulting service [10,11]. While 1 health system reported their implementation of an inpatient e-consult program in several specialties, provider adoption was temporary in response to the COVID-19 crisis [9]. Our experience with inpatient e-consults uniquely contrasts with these other health systems because we not only successfully implemented inpatient e-consults in multiple specialties but also demonstrated its sustained use beyond the pandemic.

Limitations

Although anecdotal provider feedback has been positive, limitations to this study include the absence of a formal assessment of user experiences with the inpatient e-consults service. In addition, the volumes of e-consult requests and total consult orders were low for some specialties; thus, caution must be applied in the interpretation of results from these low-volume specialties. Nevertheless, we believe our unique development of inpatient e-consults is easily translatable to other institutions interested in implementing it and will lead to a positive user experience and greater use since we fit the e-consult process into already existing and common workflows of requesting a consultation. Additionally, although the implementation of our inpatient e-consults service was in a single academic health system, we successfully demonstrated that the use of e-consults in the inpatient setting is a promising approach to expediting patient care and reporting our experience in designing and implementing inpatient e-consults may provide guidance to other health systems considering similar telehealth models.

Conclusions

Our implementation of e-consults in the inpatient setting highlighted an innovative use for e-consults in the era of COVID-19. It allowed for timely, efficient delivery of inpatient consultation services while reducing the unnecessary exposure of health care workers to potential infection. Consequently, inpatient e-consults likely helped to conserve precious PPE, minimize disease transmission, and enhance our ability to deal with surges in COVID-19 cases by expediting rapid assessment and management of lower complexity referrals. Although the COVID-19 emergency served as motivation to expand our ambulatory e-consults program to the inpatient setting, the service has become a vital component of our regular practices and will remain an essential part of our health care delivery, both in the ambulatory and inpatient settings, beyond the current pandemic, achieving sustainability and value.

Acknowledgments

The authors thank our steering committee members (Dr Byron Allen and Dr Nathan Rojek) and IT build team (Donna Jackson, Brian Lambertson, Elizabeth Burrows, Jaymee Zillgitt, and Tanya Sickles) for their contributions to the design and implementation of our inpatient e-consults. We also thank additional team members Kathy LaPierre, Jennifer Rios, and Debra Webb Torres for their guidance with compliance and billing issues.

Conflicts of Interest

ANA has been a principal investigator or coinvestigator of clinical trials sponsored by the National Institutes of Health/National Institute of Allergy and Infectious Diseases, NeuroRx Pharma, Pulmotect, Blade Therapeutics, Novartis, Takeda, Humanigen, Eli Lilly, PTC Therapeutics, OctaPharma, Fulcrum Therapeutics, and Alexion, as well as a speaker and consultant for BMS, Pfizer, BI, Portola, Sunovion, Mylan, Salix, Alexion, AstraZeneca, Novartis, Nabriva, Paratek, Bayer, Tetraphase, Achogen LaJolla, Ferring, Seres, Spero, Eli Lilly, Gilead, Millenium, HeartRite, Aseptiscope, and Sprightly; these relationships were unrelated to the current work. ASA, AAD, and AAA have no conflicts of interest to report.

Multimedia Appendix 1
Median (IQR) response times by specialty for traditional in-person consults, compared with median (IQR) response times for e-consults converted to in-person and e-consults resolved electronically.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v26i1e55623_app1.xlsx]

References

1. Chen AH, Kushel MB, Grumbach K, Yee HF. Practice profile. A safety-net system gains efficiencies through 'eReferrals' to specialists. Health Aff (Millwood) 2010;29(5):969-971. [doi: 10.1377/hlthaff.2010.0027] [Medline: 20439891]
2. Keely E, Liddy C, Afkham A. Utilization, benefits, and impact of an e-consultation service across diverse specialties and primary care providers. Telemed J E Health 2013;19(10):733-738 [FREE Full text] [doi: 10.1089/tmj.2013.0007] [Medline: 23980939]

XSL•FO
RenderX

3. Malagrino GD, Chaudhry R, Gardner M, Kahn M, Speer L, Spurrier BR, et al. A study of 6,000 electronic specialty consultations for person-centered care at The Mayo Clinic. Int J Pers Cent Med 2012;2(3):458-466. [doi: 10.5750/ijpcm.v2i3.266]
4. Arora A, Fekieta R, Nouri Z, Carder D, Colgan MM, Fuhlbrigge A, et al. Trends in utilization of electronic consultations associated with patient payer and language among US academic medical centers during the COVID-19 pandemic. JAMA Netw Open 2022;5(7):e2224628 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.24628] [Medline: 35904781]
5. Liddy C, Moroz I, Mihan A, Nawar N, Keely E. A systematic review of asynchronous, provider-to-provider, electronic consultation services to improve access to specialty care available worldwide. Telemed J E Health 2019;25(3):184-198 [FREE Full text] [doi: 10.1089/tmj.2018.0005] [Medline: 29927711]
6. Vimalananda VG, Gupte G, Seraj SM, Orlander J, Berlowitz D, Fincke BG, et al. Electronic consultations (e-consults) to improve access to specialty care: a systematic review and narrative synthesis. J Telemed Telecare 2015;21(6):323-330 [FREE Full text] [doi: 10.1177/1357633X15582108] [Medline: 25995331]
7. Kuo S, Aledia A, O'Connell R, Rudkin S, Dangodara AA, Amin AN. Implementation and impact on length of stay of a post-discharge remote patient monitoring program for acutely hospitalized COVID-19 pneumonia patients. JAMIA Open 2022;5(3):ooac060 [FREE Full text] [doi: 10.1093/jamiaopen/ooac060] [Medline: 35879961]
8. Amin AN, Aledia A. An eConsults program to improve patient access to specialty care in an academic health system. J Clin Outcomes Manag 2020;27(3):115-122. [doi: 10.12788/jcom.0008]
9. Rikin S, Epstein EJ, Gendlina I. Rapid implementation of inpatient eConsult programme addresses new challenges for patient care during COVID-19 pandemic. BMJ Innov 2021;7(2):271-277 [FREE Full text] [doi: 10.1136/bmjinnov-2020-000557] [Medline: 34192016]
10. Mustafa SS, Staicu ML, Yang L, Baumeister T, Vadamalai K, Ramsey A. Inpatient electronic consultations (E-consults) in allergy/immunology. J Allergy Clin Immunol Pract 2020;8(9):2968-2973 [FREE Full text] [doi: 10.1016/j.jaip.2020.06.020] [Medline: 32585408]
11. Yagnik KJ, Saad HA, King HL, Bedimo RJ, Lehmann CU, Medford RJ. Characteristics and outcomes of infectious diseases electronic COVID-19 consultations at a multisite academic health system. Cureus 2021;13(11):e19203 [FREE Full text] [doi: 10.7759/cureus.19203] [Medline: 34877196]
12. Najafi N, Harrison JD, Duong J, Greenberg A, Cheng HQ. It all just clicks: development of an inpatient e-consult program. J Hosp Med 2017;12(5):332-334. [doi: 10.12788/jhm.2740] [Medline: 28459902]
13. Gupte G, Vimalananda V, Simon SR, DeVito K, Clark J, Orlander JD. Disruptive innovation: implementation of electronic consultations in a veterans affairs health care system. JMIR Med Inform 2016;4(1):e6 [FREE Full text] [doi: 10.2196/medinform.4801] [Medline: 26872820]

Abbreviations

e-consult: electronic consultation
EHR: electronic health record
PPE: personal protective equipment
UCI: University of California Irvine

Edited by T de Azevedo Cardoso, S He; submitted 19.12.23; peer-reviewed by J Chen, I Moroz; comments to author 14.02.24; revised version received 06.03.24; accepted 27.03.24; published 16.05.24.

Please cite as:
Aledia AS, Dangodara AA, Amin AA, Amin AN
Implementation of Inpatient Electronic Consultations During the COVID-19 Crisis and Its Sustainability Beyond the Pandemic: Quality Improvement Study
J Med Internet Res 2024;26:e55623
URL: https://www.jmir.org/2024/1/e55623
doi:10.2196/55623
PMID:38754103

©Anna S Aledia, Amish A Dangodara, Aanya A Amin, Alpesh N Amin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Prevalence and Factors Associated With Willingness to Sustain Pandemic-Induced Digital Work in the General Population and Moderating Effects of Screen Hours: Cross-Sectional Study

Jiaying Li[1], MSc; Daniel Yee Tak Fong[1], PhD; Mandy Man Ho[1], PhD; Edmond Pui Hang Choi[1], PhD; Kris Yuet Wan Lok[1], PhD; Jung Jae Lee[1], PhD; WenJie Duan[2], PhD; Janet Yuen Ha Wong[3], PhD; Chia-Chin Lin[1], PhD

[1]School of Nursing, The University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[2]Department of Social Work, East China University of Science and Technology, Shanghai, China

[3]School of Nursing and Health Studies, Hong Kong Metropolitan University, Hong Kong SAR, China (Hong Kong)

Corresponding Author:
Daniel Yee Tak Fong, PhD
School of Nursing
The University of Hong Kong
5/F, Academic Building 3 Sassoon Road
Pokfulam
Hong Kong SAR, 999077
China (Hong Kong)
Phone: 852 39176645
Email: dytfong@hku.hk

Abstract

Background: The pandemic has accelerated digital work transformation, yet little is known about individuals' willingness to sustain such digital modes and its associated factors. A better understanding of this willingness and its drivers is crucial for guiding the development of future digital work infrastructure, training programs, and strategies to monitor and prevent related health issues.

Objective: This study aims to quantify the general population's willingness to sustain pandemic-induced digital work, identify its associated factors, and examine how screen time moderates these relationships.

Methods: A cross-sectional study was conducted targeting Hong Kong residents aged ≥18 years who have increased engagement in digital work since the pandemic. Data were collected through self-reported, web-based surveys. Descriptive statistics determined prevalence rates, while structured multiphase logistic regression identified associated factors and explored the moderating effects of screen hour levels.

Results: This unfunded study enrolled 1014 participants from May 2 to June 24, 2022, and completed data analysis within 3 months after data collection. A total of 391 (38.6%; 95% CI 35.6%-41.6%) participants expressed willingness to sustain digital work. Positive factors associated with this willingness included being an employee (odds ratio [OR] 3.12, 95% CI 1.59-6.45; *P*=.001), being health professionals (OR 3.32, 95% CI 1.49-7.82; *P*=.004), longer screen hours (OR 1.09, 95% CI 1.03-1.15; *P*=.002), and higher depression levels (OR 1.20, 95% CI 1.01-1.44; *P*=.04). Conversely, negatively associated factors included older age (OR 0.87, 95% CI 0.81-0.94; *P*=.001), extroversion (OR 0.66, 95% CI 0.51-0.86; *P*=.002), higher eHealth literacy (OR 0.96, 95% CI 0.93-0.98; *P*<.001), perceived greater susceptibility to COVID-19 (OR 0.84, 95% CI 0.74-0.96; *P*=.009), residence in a high-severity COVID-19 community (OR 0.73, 95% CI 0.63-0.84; *P*<.001), having infected individuals in the immediate social circle (OR 0.64, 95% CI 0.46-0.88; *P*=.006), higher BMI (OR 0.94, 95% CI 0.90-0.99; *P*=.02), feelings of being out of control (OR 0.96, 95% CI 0.93-0.98; *P*=.002), and higher fear of COVID-19 (OR 0.96, 95% CI 0.94-0.98; *P*=.001). In addition, a moderating effect of screen hour level (high: >8 h/d; low: ≤8 h/d) influenced the association among 10 factors related to willingness to sustain pandemic-induced digital work, including age, education level, household size, needs for regular medical care, BMI, frequency of both vigorous and moderate physical activities, perceived COVID-19 severity, immediate social circle COVID-19 presence, and fear of COVID-19 (all *P* values for interaction <.05).

Conclusions: The substantial willingness of the general population to sustain digital work after the pandemic highlights the need for robust telework infrastructure, thorough monitoring of adverse health outcomes, and the potential to expand telehealth

XSL•FO
RenderX

services among this group. The identification of factors influencing this willingness and the moderating role of screen hours inform the development of personalized strategies to enhance digital work acceptance where needed.

(J Med Internet Res 2024;26:e53321) doi:10.2196/53321

KEYWORDS

COVID-19 pandemic; digital work; willingness to sustain; screen time; general population

Introduction

Digital transformation, recognized as one of the most significant benefits of the COVID-19 pandemic, has deeply integrated digital technologies into daily work practices, transforming traditional physical offices into predominantly digital realms [1,2]. Supported by the widespread adoption of videoconferencing and instant messaging platforms, real-time communication has been revolutionized. In addition, learning management systems and project management tools have become indispensable for organizing tasks efficiently and ensuring secure data sharing [3]. This global shift toward digital work environments raises critical questions about their long-term sustainability, as this extent of willingness to sustain such internet-mediated work environments is pivotal not only for shaping future tele-based infrastructure but also for influencing prevention and surveillance of tele-related health outcomes, as well as enhancing our understanding of human acceptance of digital means along with the factors driving this acceptance.

Although extensive research has explored the advantages and drawbacks of digital work environments [4-7], significant gaps remain in understanding how widely individuals are willing to continue in such settings. One notable study revealed that 63.4% of US energy sector employees who transitioned to digital work were reluctant to return to on-site work, preferring to avoid their physical workplaces [8]. Broadening our understanding of this willingness across varied populations is essential for shaping policies on future tele-based infrastructure. This encompasses improvements in remote work features such as cybersecurity and data privacy protections. In addition, these insights are crucial for evaluating the health outcomes associated with the rising trend of telework [6,9], guiding future health recommendations and workplace standards. Furthermore, the positive correlation between telework and telemedicine suggests that this willingness could also influence public health strategies to enhance health care access in increasingly remote settings [10].

Understanding the associated factors of willingness to sustain digital transformations is crucial for guiding strategies to enhance acceptance for needs such as resource conservation, personalized work adjustments, and ensuring adherence to social distancing in future pandemics. To identified associated factors, we used a web-based survey using the Health Belief Model, which effectively predicts health behaviors and encompasses 4 main domains: perceived susceptibility, severity, benefits, and barriers [11]. In this model, knowledge of COVID-19, prior infection history, and eHealth literacy are classified under perceived susceptibility and severity, clarifying how individuals perceive the risks of sustaining digital work. Meanwhile, lifestyle habits and BMI correspond to perceived benefits and barriers, capturing individual evaluations of the pros and cons of sustaining the pandemic-induced digital work. In addition, the Health Belief Model acknowledges the impact of demographic and psychological factors, such as personality traits, on health behaviors, which directly influence both the model's key domains and health behaviors themselves. Extended screen time in digital work often results in visual strain, musculoskeletal discomfort, and mental fatigue [12,13], which may diminish one's willingness to continue such arrangements. However, frequent exposure to prolonged screen hours can also enhance proficiency with information and communication technologies [14], thereby improving perceptions of productivity and work-life balance in digital environments. Given these dual effects, the influence of various factors on willingness to sustain digital work might differ significantly among individuals based on their levels of screen time exposure. Thus, we also investigated the potential moderating effect of screen time on the association of the aforementioned factors with the willingness to sustain digital work in order to provide a more nuanced understanding of these relationships.

Therefore, this study addresses research gaps by quantifying the general population's willingness to sustain digital work, identifying influencing factors, and examining how screen time moderates these relationships. Our hypotheses were as follows: (1) a segment of the population will prefer continued digital work, (2) various factors influence this willingness, and (3) screen time affects this willingness and alters the effects of related factors. Findings were expected to enhance understanding of digital work acceptance and its determinants, guiding the development of tele-based work infrastructure, health outcome monitoring, and tailored strategies to increase willingness where needed.

Methods

Study Design

This was a cross-sectional study conducted in Hong Kong using a web-based survey. The methodology of the web-based survey was reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [15].

Setting

The web-based survey was conducted among Chinese adults in Hong Kong from May 2 to June 24, 2022. This period was approximately 2 months after the Omicron surge (from February 28 to March 6, 2022), and social distancing measures were still in place.

Participants and Sample Size

The web-based survey used convenience sampling to include participants who were Chinese-reading adults aged 18 years or

XSL•FO
RenderX

older residing in Hong Kong. We excluded individuals who reported no increased engagement in digital working or learning since the onset of the COVID-19 pandemic via a self-reported question. The sample size was determined using the rule of thumb, that is, a minimum of 10 events per variable [16]. Considering up to 19 variables in the regression model, a total of 190 subjects experiencing the event were required. Based on a prior study in which the event rate of reluctance to return to the physical workplace was 63.4% (211/333) [8], a total of 300 participants would be needed.

Variables

Overview

Multimedia Appendix 1 contains the web-based questionnaire, detailing each variable included in this study.

Sociodemographics

Sociodemographic variables included sex, age, marital status, education level, employment status, perceived social rank, whether they were health professionals, needs for regular medical care, number of children younger than 18 years, and household size. In addition, BMI was also recorded.

COVID-19–Related Knowledge and Infection

We assessed eHealth literacy, COVID-19–related knowledge, and COVID-19 infection status. eHealth literacy was assessed using the validated Chinese eHealth Literacy Scale [17], which evaluates respondents' knowledge, comfort, and skills in using electronic health information. The 8-item scale uses a 5-point Likert scale, with higher scores indicating greater eHealth literacy. In addition, COVID-19–related knowledge was measured using 6 items from the World Health Organization's behavioral survey on COVID-19 [18], focusing on perceived knowledge and severity. Each item was rated on a 7-point Likert scale, with higher scores indicating greater knowledge or severity. Moreover, COVID-19 infection status was assessed with 2 questions asking whether participants or someone in their immediate social circle had contracted COVID-19 [18], with responses as "yes" or "no."

Lifestyle

Self-reported lifestyle variables included alcohol and tobacco use, frequency of vigorous and moderate physical activities, screen time, and sitting hours. Tobacco and alcohol consumption were assessed using an 8-point Likert scale from "never" to "everyday." The frequency of both vigorous and moderate physical activities was rated on a 6-point Likert scale from "none" to "5 days or more." Participants reported their average daily screen hours and sitting hours by entering an integer number of hours. For analysis, screen hours were dichotomized into 2 groups: those reporting more than 8 hours were categorized as the "high screen hour group," and those reporting 8 hours or fewer were categorized as the "low screen hour group."

Perceived Personality

We assessed participants' personality identities with a self-reported single question—"Do you primarily identify as an introvert or an extrovert?"—as in other studies [19,20].

Psychological Status

Psychological outcomes including anxiety, depression, fear of COVID-19, and feelings of being out of control were evaluated. Anxiety and depression were measured using the 4-item Patient Health Questionnaire-4, with both subscales showing a Cronbach α value of 0.80 [21,22]. Responses were rated on a 4-point Likert scale, with higher scores indicating more severe symptoms. In addition, fear of COVID-19 was assessed using an 8-item scale validated in Chinese populations, with a Cronbach α value of 0.93 [22]. Items were scored on a 5-point Likert scale, with higher scores indicating greater fear. Moreover, feeling of being out of control was measured by a 7-item scale, with a Cronbach α value of 0.91 [18], using a 6-point Likert scale, where higher scores indicated a stronger perception of loss of control.

Willingness to Sustain Pandemic-Induced Digital Work Mode

We used a single direct question to assess participants' willingness to return to a prepandemic work mode, with binary response options of "yes" or "no."

Data Collection

Between May 2 and June 24, 2022, we collected data from the Hong Kong community by developing an open web-based survey. Prior to the main data collection, we piloted it with 30 people to ensure the usability and functionality of the survey. The survey link was then distributed through email and text messages. To ensure a diverse representation of work environments and statuses, we partnered with an established local survey provider who had access to an extensive participant database. The survey, with 58 questions presented 1 per page, targeted individuals meeting our predefined inclusion criteria. Participation in the survey was voluntary, with no incentives provided. The survey did not use randomized questions or adaptive questioning techniques. However, respondents had the option to revise their responses using a back button, and a completeness check was implemented. To prevent duplicate submissions, we tracked the IP addresses of participants throughout the data collection period. Of the 2398 individuals contacted, 1900 (79.2%) successfully completed the survey, reflecting a notable response rate. Incomplete responses were excluded from data analysis, and no specific time frame was applied to disqualify completed surveys.

Data Analysis

All statistical analyses were performed using R software (version 4.1.1; R Core Team 2021). After excluding participants who were not currently working, those who reported no changes in their work, or those who did not increase their engagement in digital work since the onset of the COVID-19 pandemic, we used descriptive statistics to summarize the sociodemographic data. Continuous variables that conformed to a normal distribution were summarized using means and SDs, while categorical variables were described by frequency counts.

Associated factors of willingness to sustain digital work were identified using a 2-step structured multiphase logistic regression, which grouped variables into sequential clusters before analysis to avoid overadjustment and consider causal

XSL•FO
RenderX

relationships [23]. In the first stage, independent variables were organized into sequential clusters based on their potential causal relationships, as informed by empirical evidence. That is, variables in a cluster can affect the variables in later clusters but not vice versa. Cluster 1 included sociodemographic variables, including age, sex, education level, marital status, employment status, need for regular medical follow-up before the COVID-19 pandemic, whether practicing health professional, number of children younger than 18 years, household size, house size, perceived social rank, and personality. Cluster 2 included COVID-19–related knowledge, infection status, and lifestyle. BMI was grouped into cluster 3, while psychological variables, including out of control, fear of COVID-19, anxiety, and depression, were considered as cluster 4. Cluster-1 variables were considered immutable, not influenced by changes in lifestyle or other variables in later clusters. Similarly, cluster-2 variables may influence the health outcome variables in clusters 3 and 4. Also, cluster 3's variable, that is, BMI, was a self-reported variable and presumed to be measured before psychological variables and thus may influence cluster-4 variables. We ultimately had 4 clusters. Our clustering is thus informed by existing knowledge, aiming to reflect a logical order of influence rather than implying direct causality from the data of this cross-sectional study. In the second stage, logistic regression analysis was conducted in 4 phases, with each phase adjusting for significant variables identified in the previous phase. In addition, screen hours were categorized into high (>8 h/d) and low (<8 h/d) levels. We analyzed the moderation effect of screen hour level on the association of each factor with the willingness to sustain, adjusting for corresponding factors within the same phase. Model adequacy was verified using the Hosmer-Lemeshow goodness-of-fit test, and multicollinearity was assessed with the variance inflation factor, setting statistical significance at $P<.05$.

Ethical Considerations

The institutional review board of The University of Hong Kong—the Hospital Authority Hong Kong West Cluster approved this study (approval UW 20-272). All participants provided informed consent, were briefed on their right to withdraw at any time without repercussions, and participated voluntarily. Measures were taken to ensure privacy and confidentiality through the anonymization of data, with no personal identifiers used in the analysis. No compensation was provided to the participants for their involvement in the study.

Results

Data Cleaning and Respondents' Characteristics

Initially, we received 1900 valid responses. After removing respondents who were job seeking (n=86, 4.5%), laid off (n=20, 1.1%), not in workforce (n=213, 11.2%), retired (n=128, 6.7%), and reported no change or no increased engagement in digital work or learning mode since the COVID-19 pandemic (n=439, 23.1%), we were left with 1014 respondents. Of these, 518 (51.1%) were female respondents. Table 1 summarizes the respondents' other characteristics.

XSL•FO
RenderX

Table 1. Participants' sociodemographics, COVID-19–related knowledge and infection, lifestyles, and psychological status (n=1014).

Variables	Values
Age group (years), n (%)	
18-24	105 (10.4)
25-29	118 (11.6)
30-34	148 (14.6)
35-39	163 (16.1)
40-44	136 (13.4)
45-49	99 (9.8)
50-54	81 (8)
55-59	91 (9)
60-64	48 (4.7)
≥65	25 (2.5)
Sex, n (%)	
Female	518 (51.1)
Male	496 (48.9)
Education levels, n (%)	
Primary or below	24 (2.4)
Secondary	483 (47.6)
College	137 (13.5)
Associate degree	101 (10)
Bachelor's degree	257 (25.3)
Graduate	12 (1.2)
Marital status, n (%)	
Married or cohabitation or common law	611 (60.3)
Separated or divorced or widowed	19 (1.9)
Single	384 (37.9)
Occupational status, n (%)	
Self-employed	37 (3.6)
Student	51 (5)
Employee	926 (91.3)
Needs for regular medical care, n (%)	
No	824 (81.3)
Yes	190 (18.7)
Practicing health professional, n (%)	
No	986 (97.2)
Yes	28 (2.8)
Personality, n (%)	
Introverts	457 (45.1)
Extroverts	557 (54.9)
Number of children younger than 18 years, mean (SD)	0.39 (0.68)
Number of people in the household, mean (SD)	3.38 (1.10)
House size (m^2), mean (SD)	40.38 (14.36)
Perceived social rank (1=lowest to 5=highest), mean (SD)	2.94 (0.70)

XSL•FO
RenderX

Variables	Values
BMI, mean (SD)	23.50 (2.82)
eHealth literacy (8-40), mean (SD)	26.85 (6.18)
Knowledge on COVID-19 (1-7), mean (SD)	4.37 (0.92)
Knowledge on preventing the spread of COVID-19 (1-7), mean (SD)	4.39 (0.94)
Perceived adequacy of knowledge about COVID-19 (1-7), mean (SD)	4.38 (0.95)
Perceived susceptibility to an infection with COVID-19 (1-7), mean (SD)	4.28 (1.28)
Perceived severity of contracting COVID-19 (1-7), mean (SD)	4.23 (1.26)
Severity of the spread of COVID-19 in your community (1-7), mean (SD)	4.48 (1.30)
Ever infected with COVID-19, n (%)	
Yes	374 (36.9)
No	640 (63.1)
Someone in your immediate social circle infected with COVID-19, n (%)	
Yes	695 (68.5)
No	319 (31.5)
Lifestyles, mean (SD)	
Frequency of drinking alcohol per week (0=never to 8=everyday), mean (SD)	2.92 (1.70)
Frequency of smoking per week (0=never to 8=everyday), mean (SD)	2.52 (2.65)
Screen hours per day (hours), mean (SD)	6.48 (3.44)
Screen hour group, n (%)	
Low screen hour group (≤8 hours per day)	783 (77.2)
High screen hour group (>8 hours per day)	231 (22.8)
Sitting hours per day (hours), mean (SD)	7.46 (3.37)
Frequency of vigorous physical activities per week (0=none to 6=5 days or more), mean (SD)	1.73 (1.22)
Frequency of moderate physical activities per week (0=none to 6=5 days or more), mean (SD)	1.98 (1.28)
Out of control (7-42), mean (SD)	25.37 (5.92)
Fear of COVID-19 (8-40), mean (SD)	20.12 (7.54)
Anxiety (2-8), mean (SD)	2.91 (1.15)
Depression (2-8), mean (SD)	2.76 (1.13)

Willingness to Sustain Digital Work

Among the 1014 participants, 391 (38.6%; 95% CI of 35.6%-41.65%) expressed a willingness to sustain digital work.

Associated Factors of Willingness to Sustain Digital Work

Table 2 presents the results of the structured multiphase logistic regression. The variance inflation factor values for all phases of the regression ranged from 1.02 to 1.52, indicating low correlation among the independent variables. The Hosmer-Lemeshow goodness-of-fit test results for phase 1 (χ^2_8=3.58; *P*=.89), phase 2 (χ^2_8=10.85; *P*=.21), phase 3 (χ^2_8=6.81; *P*=.56), and phase 4 (χ^2_8=8.17; *P*=.42) all demonstrated adequacy. The statistics for the variables, which function as adjusting variables in the subsequent phase, are shown in Multimedia Appendix 2.

In phase 1, older age (odds ratio [OR], 0.87, 95% CI 0.69-1.10; *P*=.001) and being an extrovert (OR 0.66, 95% CI 0.51-0.86; *P*=.002) were negatively associated with the willingness to sustain, whereas employees (OR 3.12, 95% CI 1.59-6.45; *P*=.001) and health professionals (OR 3.32, 95% CI 1.49-7.82; *P*=.004) showed increased willingness.

In phase 2, higher eHealth literacy (OR 0.96, 95% CI 0.93-0.98; *P*<.001), perceived higher susceptibility to COVID-19 (OR 0.84, 95% CI 0.68-1.04; *P*=.009), living in a high-severity COVID-19 community (OR 0.73, 95% CI 0.63-0.84; *P*<.001), and having people infected in the immediate social circle (OR 0.64, 95% CI 0.46-0.88; *P*=.006) negatively influenced willingness to sustain. Conversely, longer screen hours (OR 1.09, 95% CI 1.03-1.15; *P*=.002) were positively associated willingness to sustain.

XSL•FO
RenderX

Table 2. Structured multiphase logistic regression of willingness to sustain digital work (n=1014).

Variables	OR[a] (95% CI)	*P* value	VIF[b]
Phase 1			
Age (years; 1-10)[c]	0.87 (0.81-0.94)	<.001	1.35
Educational level (1-6)[d]	0.98 (0.88-1.10)	.77	1.16
Marital status (reference: single)			1.17
Married or cohabitation or common law	1.24 (0.87-1.78)	.24	
Separated or divorced or widowed	0.61 (0.16-1.88)	.42	
Sex (reference: female)			1.03
Male	0.79 (0.61-1.04)	.09	
Occupational status (reference: student)			1.04
Self-employed	1.48 (0.47-4.42)	.49	
Employee	3.12 (1.59-6.45)	.001	
Needs for regular medical care (yes vs no)	0.83 (0.57-1.18)	.30	1.05
Practicing health professional (yes vs no)	3.32 (1.49-7.82)	.004	1.02
Number of children younger than 18 years	0.87 (0.69-1.10)	.26	1.24
Number of people in the household	1.00 (0.87-1.15)	.99	1.17
House size, m^2	1.00 (0.99-1.01)	.46	1.13
Perceived social rank	1.02 (0.83-1.25)	.85	1.08
Personality (extroverts vs introverts)	0.66 (0.51-0.86)	.002	1.01
Phase 2			
eHealth literacy	0.96 (0.93-0.98)	<.001	1.08
Knowledge on COVID-19	0.84 (0.68-1.04)	.12	1.39
Knowledge on preventing the spread of COVID-19	1.11 (0.90-1.37)	.35	1.44
Perceived adequacy of knowledge about COVID-19	0.9 (0.73-1.09)	.27	1.34
Perceived susceptibility to an infection with COVID-19	0.84 (0.74-0.96)	.009	1.17
Perceived severity of contracting COVID-19	1.08 (0.93-1.25)	.32	1.31
Severity of the spread COVID-19 in your community	0.73 (0.63-0.84)	<.001	1.30
Ever infected with COVID-19 (yes vs no)	1.03 (0.74-1.42)	.88	1.14
Someone in your immediate social circle infected with COVID-19 (yes vs no)	0.64 (0.46-0.88)	.006	1.12
Frequency of alcohol drinking per week	1.01 (0.92-1.11)	.82	1.13
Frequency of smoking per week	0.97 (0.92-1.03)	.36	1.11
Screen hours (0-24 hours per day)	1.09 (1.03-1.15)	.002	1.37
Sitting hours (0-24 hours per day)	0.97 (0.92-1.02)	.28	1.35
Frequency of vigorous physical activities per week	0.95 (0.83-1.08)	.41	1.18
Frequency of moderate physical activities per week	1.1 (0.97-1.24)	.15	1.18
Phase 3			
BMI	0.94 (0.90-0.99)	.02	1.03
Phase 4			
Out of control	0.96 (0.93-0.98)	.002	1.22
Fear of COVID-19	0.96 (0.94-0.98)	.001	1.32
Anxiety	1.07 (0.90-1.28)	.44	1.52
Depression	1.2 (1.01-1.44)	.04	1.47

XSL•FO
RenderX

[a]OR: odds ratio.

[b]VIF: variance inflation factor.

[c]Age group: 1-10 corresponding to 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, and older than 65 years, respectively.

[d]Educational level: 1-6 corresponding to primary or below, secondary, college, associated degree, bachelor's degree, and graduate, respectively.

In phase 3, higher BMI showed negative association (OR 0.94, 95% CI 0.90-0.99; *P*=.02). In phase 4, higher out of control (OR 0.96, 95% CI 0.93-0.98; *P*=.002) and higher fear of COVID-19 (OR 0.96, 95% CI 0.94-0.98; *P*=.001) were negatively associated with willingness to sustain, while higher depression was a risk factor (OR 1.20, 95% CI 1.01-1.44; *P*=.04).

Moderating Effects of Screen Hour Levels

Screen hour levels moderated the association between willingness to sustain digital work and 10 variables (Table 3 and Figure 1): age, education, needs for regular medical care, household size, perceived COVID-19 severity, COVID-19 cases in the immediate social circle, and the frequency of both vigorous and moderate physical activities, along with BMI and fear of COVID-19 (*P* value for interaction <.05 for all). In the high screen time group, age (OR 1.32, 95% CI 1.18-1.49), needs for regular medical care (OR 6.48, 95% CI 3.28-13.40), BMI (OR 1.16, 95% CI 1.03-1.33), and fear of COVID-19 (OR 1.08, 95% CI 1.03-1.12) positively influenced willingness (*P*<.001 for all). Conversely, the low screen time group showed reduced willingness for these factors—age (OR 0.87, 95% CI 0.83-0.91), regular medical care (OR 0.47, 95% CI 0.33-0.65), BMI (OR 0.87, 95% CI 0.82-0.92), and fear of COVID-19 (OR 0.96, 95% CI 0.94-0.98; *P*<.001 for all). In addition, education (OR 1.30, 95% CI 1.16-1.47) and household size (OR 1.48, 95% CI 1.29-1.69) enhanced willingness in the low screen time group but negatively affected the high screen time group (education: OR 0.62, 95% CI 0.50-0.76; household size: OR 0.56, 95% CI 0.43-0.71; *P*<.001 for all). Moreover, perceived COVID-19 severity (OR 0.73, 95% CI 0.59-0.91) and presence of COVID-19 in the immediate social circle (OR 0.15, 95% CI 0.07-0.29) negatively impacted willingness in the high screen hour group, whereas the frequency of vigorous (OR 1.34, 95% CI 1.03-1.76) and moderate physical activities (OR 1.72, 95% CI 1.30-2.33) positively influenced willingness only in the high screen hour group (all *P*<.05).

Table 3. Moderating effects of screen hour level (n=1014)[a].

Variable moderated by screen hour level	Low level of screen hour (≤8 h/d; n=783)		High level of screen hour (>8 h/d; n=231)		*P* values for interaction
	OR[b] (95% CI)	*P* value	OR (95% CI)	*P* value	
Age (years)	0.87 (0.83-0.91)	<.001	1.32 (1.18-1.49)	<.001	.006
Education level	1.30 (1.16-1.47)	<.001	0.62 (0.50-0.76)	<.001	.009
Needs of regular medical care (yes vs no)	0.47 (0.33-0.65)	<.001	6.48 (3.28-13.40)	<.001	<.001
Household size	1.48 (1.29-1.69)	<.001	0.56 (0.43-0.71)	<.001	.03
Perceived severity of contracting COVID-19	0.96 (0.86-1.08)	.49	0.73 (0.59-0.91)	.005	.02
Someone in your immediate social circle infected with COVID-19 (yes vs no)	0.94 (0.70-1.27)	.69	0.15 (0.07-0.29)	<.001	.007
Frequency of vigorous physical activities	1.04 (0.91-1.18)	.56	1.34 (1.03-1.76)	.03	.008
Frequency of moderate physical activities	0.96 (0.85-1.08)	.53	1.72 (1.30-2.33)	<.001	<.001
BMI	0.87 (0.82-0.92)	<.001	1.16 (1.03-1.33)	.02	.01
Fear of COVID-19	0.96 (0.94-0.98)	<.001	1.08 (1.03-1.12)	<.001	.02

[a]All estimated effects were adjusted by variables that correspond to their respective phase, as listed in Table 2.

[b]OR: odds ratio.

XSL•FO
RenderX

Figure 1. Visualizing the moderation effect of screen hour level (n=1014).

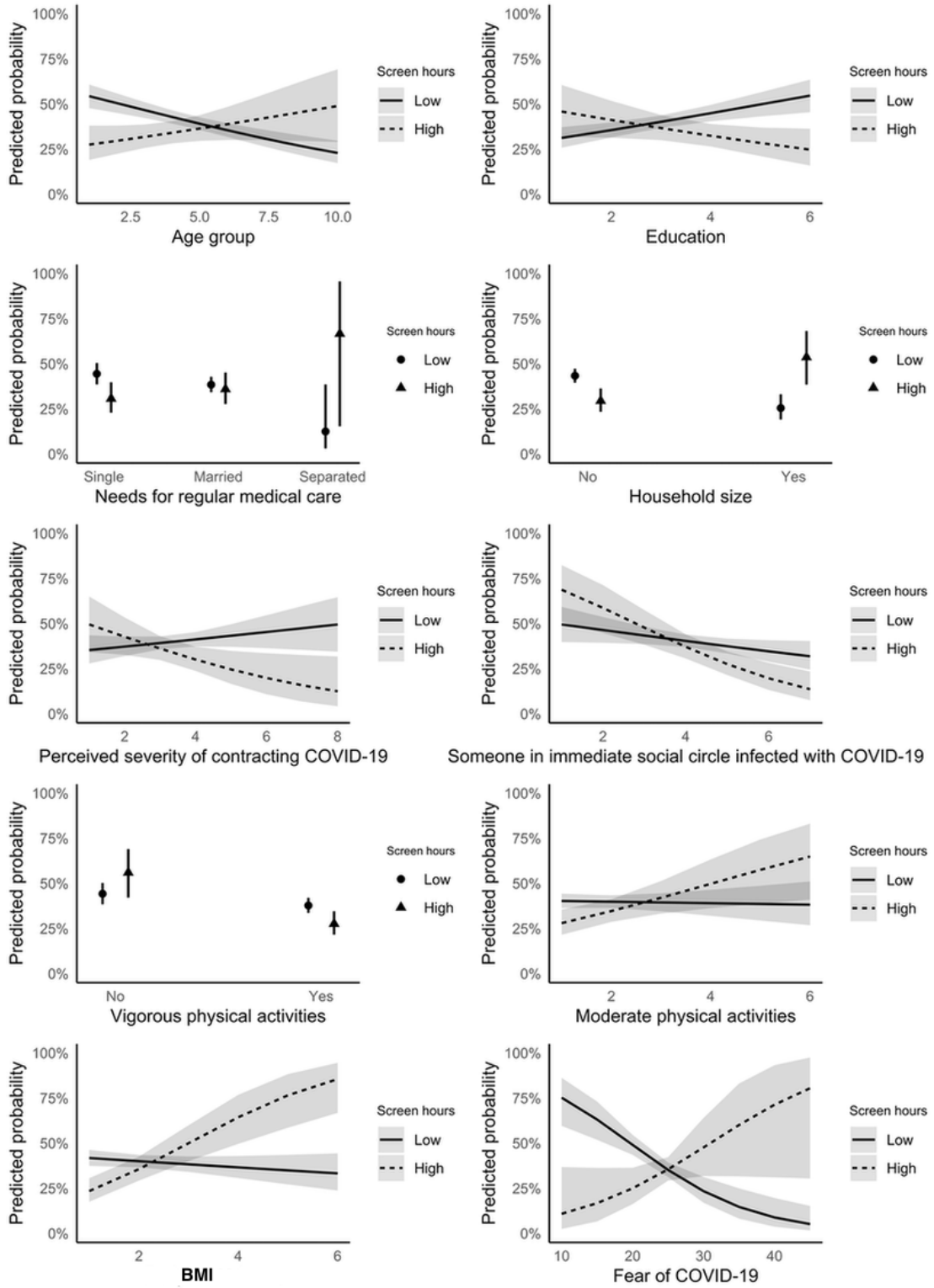

Discussion

Principal Findings

This study is the first to evaluate the general population's willingness to sustain pandemic-induced digital work, associated factors, and the moderating effect of screen hour levels. A willingness rate of 38.6% (391/1014) highlights the need for improved tele-based infrastructure and health outcome monitoring. In addition, factors positively linked to higher willingness included being employees, health professionals, having longer screen hours, and higher depression levels. Conversely, negative factors were older age, extroversion, high health literacy, perceived greater susceptibility to COVID-19, perceived community COVID-19 severity, having infected individuals in the immediate social circle, higher BMI, increased feelings of losing control, and fear of COVID-19. Moreover, screen hour levels moderated the impact of 10 factors: age,

XSL•FO
RenderX

education, needs for regular medical care, household size, perceived COVID-19 severity, COVID-19 cases in the immediate social circle, and the frequency of both vigorous and moderate physical activities, along with BMI and fear of COVID-19. These findings enhance our understanding of the drivers behind digital work engagement, facilitating the customization of digital work policies and regulations, and guiding the development of strategies tailored to people with various screen hour levels.

Considerable Level of Willingness to Sustain

Compared with a previous study in which 63.4% of energy sector employees in the United States preferred not returning to physical workplaces [8], only 38.6% of the general population in Hong Kong expressed willingness to continue digital work. This difference likely arises from the energy sector's compatibility with remote work. To meet the universal demand for better digital support, organizations should overcome challenges, such as inadequate infrastructure and training, and inflexibility in physical roles. In addition, there is a critical need to monitor and prevent adverse health outcomes associated with digital work, such as sedentary behavior, musculoskeletal disorders, eye strain, and mental health issues. Key strategies should include investing in technology, fostering a culture of flexible work, enhancing digital literacy, revising policies to ensure equitable access to remote work, and establishing recommendations and standards for digital workplace health.

Associated Factors of Willingness to Sustain Digital Work Mode

Understanding factors associated with sustaining digital work is key to leveraging this shift. Consistent with prior studies [24,25], older adults and extroverts are less inclined toward digital work due to technological discomfort and reduced social interactions, respectively. Tailored training for older adults and enhanced digital networking can address these issues. The other 11 associated factors are newly identified.

Specifically, for negative factors, individuals with high health literacy might recognize the health risks of prolonged digital engagement, showing reduced willingness; this highlights the need for clear ergonomic guidelines and health strategies. Environmental safety concerns [26], such as increased perceptions of infection risk, community COVID-19 severity, having infected individuals in the immediate social circle, and fear of COVID-19, also lower willingness, emphasizing the need for transparent safety protocols to reassure those concerned about infection risks. Conversely, individuals with higher BMI may prefer remote work to avoid commuting and stigma associated with social interactions in physical workplaces [27], underscoring the need for inclusive and ergonomic remote work policies. In addition, those feeling a loss of control find remote environments isolating and unpredictable without the structured support of traditional workplaces [28], underlining the importance of structured routines and clear digital tool training to instill organization and control, thereby increasing the appeal of digital work.

In contrast, for positive factors, employees often exhibit greater willingness, drawn by perceived benefits such as improved work-life balance and efficiency [8]. Organizations should therefore promote flexibility and efficiency, tailoring practices based on employee feedback. Health professionals, more adept with digital tools and telehealth [29], also show a higher propensity to sustain digital modes. Continued investment in digital health care platforms and comprehensive training to keep up with technological advancements are beneficial. Similarly, those accustomed to long screen hours likely find digital environments more conducive to productivity and work-life balance. However, this finding underscores the importance of focusing on ergonomic solutions and wellness programs to mitigate health risks associated with sedentary behavior. Individuals with depression may prefer digital work, as their typical avoidance symptoms might make traditional settings more stressful [30], suggesting that digital methods could enhance mental support, therapy, and access to digital counseling, fostering a supportive work culture.

Moderating Effects of Screen Hour Level

The moderating effect of screen hour levels, indicating different degrees of digital engagement in work, highlights the diverse mechanisms affecting willingness to sustain digital work across groups. First, older age, regular medical follow-ups, higher BMI, and greater COVID-19 fear have opposite impacts in high versus low screen time groups. Older individuals with longer screen hours typically possess digital work skill and recognize its benefits. Individuals with high screen hours likely integrated medical follow-ups with their digital work before the pandemic, while those with fewer screen hours struggled with coordination, leading to contrasting moderation effects. Individuals with high BMI in the low screen hour group may fear increased sedentary behavior worsening their condition and facing social stigma [31], while those in the high screen hour group see digital work as a way to avoid such stigma [27]. Similarly, fear of COVID-19 varies; high screen time users see continued digital work as protective, whereas those with less screen time, encountering less alarming news on websites [32], feel safer pursuing outdoor activities.

Previously, higher education was considered a positive factor [33] and household size a negative one [34]. However, in the high screen hour group, both higher education and larger households detract from digital work appeal, likely due to increased health risk awareness of digital work and more disruptions, respectively. Conversely, in the low screen hour group, higher education and larger households positively impact digital work by enhancing the ability to manage digital challenges and improving work-life balance.

Finally, infection concerns and COVID-19 exposure negatively impact only the high screen time group, possibly increasing stress and reducing productivity [35]. Conversely, frequent moderate and vigorous physical activities benefit this group by potentially offsetting sedentariness and managing stress [36].

The moderating effects of screen time exposure call for a customized approach that considers varying digital demands to enhance digital work acceptance. For high screen hour group, considering regular medical follow-ups significantly increase willingness to engage in digital work, strategies to incorporate chronic disease management into tele-based work through digital

XSL•FO
RenderX

platforms, along with mitigating health impacts from prolonged digital engagement via health education and less disruptive environments such as immersive setups [37], are beneficial. Policies promoting an active lifestyle with incentives for physical activities [38] can greatly enhance willingness to sustain digital modes. For low screen hour group, boosting digital literacy with targeted training and improving digital tool access can positively affect perceptions of digital work benefits. In addition, addressing the reluctance to sustain digital work in this group during social distancing for future pandemics is crucial for public health planning.

Limitations

Several limitations are worth noting. First, the study included participants only from Hong Kong, which may limit the generalizability of the findings to other countries or cultures. Second, the data collection relied on a digital platform, potentially excluding individuals with low socioeconomic status, limited digital literacy, or inadequate access to the internet or digital devices, or including more participants proficient in digital tools. Such selection bias might potentially inflate prevalence rates and skewing the factors perceived to influence willingness to sustain digital work. Future studies should adopt varied data-gathering techniques to enhance sample representativeness. Third, our study's cross-sectional design constrains insights into how individuals' willingness to sustain digital work may shift over time as the pandemic evolves and limits our ability to infer causality. Longitudinal or experimental studies are needed for a more thorough understanding. Fourth, in the phase 2 regression, we included 19 independent variables without correcting for type I error, necessitating cautious interpretation of small-effect associations. Future research should rigorously control for type I error to establish causality more reliably. Finally, due to the lack of established tools to measure willingness to sustain pandemic-induced digital work, we used a single question, potentially oversimplifying the assessment. Future research should develop validated tools for more accurate evaluations.

Conclusions

Our study shows that 38.6% (391/1014) of the general population in Hong Kong who increased their engagement in digital work during the pandemic are willing to continue this mode, highlighting the significant impact of the pandemic on digital work adoption and its potential sustainability. These findings emphasize the need for improved telework infrastructure and training; vigilant health surveillance to prevent issues such as musculoskeletal disorders, eye strain, and sedentary behavior; and the establishment of comprehensive digital work health guidelines and workplace standards. The high adoption rate also indicates promising prospects for expanding telehealth services within these populations. Furthermore, we identified 13 factors influencing this willingness, of which 11 were newly identified, and found that screen time levels moderate the impact of another 10 factors. These findings enhance our understanding of the drivers behind digital work engagement, enabling the customization of digital work policies, settings, and regulations. The moderation effects of screen time offer a detailed perspective on how these factors differ across various screen hour–demanding individuals, supporting the development of more targeted strategies to enhance their digital work acceptance where needed.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Web-based questionnaire.
[DOCX File , 27 KB - jmir_v26i1e53321_app1.docx]

Multimedia Appendix 2
Structured multiphase logistic regression analysis of willingness to sustain digital work, with statistics for the adjusting variables (n=1014).
[DOCX File , 25 KB - jmir_v26i1e53321_app2.docx]

References

1. Athanasiadou C, Theriou G. Telework: systematic literature review and future research agenda. Heliyon 2021;7(10):e08165 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e08165] [Medline: 34703928]
2. Contreras F, Baykal E, Abid G. E-leadership and teleworking in times of COVID-19 and beyond: what we know and where do we go. Front Psychol 2020;11:590271 [FREE Full text] [doi: 10.3389/fpsyg.2020.590271] [Medline: 33362656]
3. Vargo D, Zhu L, Benwell B, Yan Z. Digital technology use during COVID-19 pandemic: a rapid review. Hum Behav Emerg Tech 2020;3(1):13-24. [doi: 10.1002/hbe2.242]
4. Franken E, Bentley T, Shafaei A, Farr-Wharton B, Onnis L, Omari M. Forced flexibility and remote working: opportunities and challenges in the new normal. J Manag Organ 2021;27(6):1131-1149. [doi: 10.1017/jmo.2021.40]
5. De Vincenzi C, Pansini M, Ferrara B, Buonomo I, Benevene P. Consequences of COVID-19 on employees in remote working: challenges, risks and opportunities an evidence-based literature review. Int J Environ Res Public Health 2022;19(18):11672 [FREE Full text] [doi: 10.3390/ijerph191811672] [Medline: 36141948]

XSL•FO
RenderX

6. Furuya Y, Nakazawa S, Fukai K, Tatemichi M. Health impacts with telework on workers: a scoping review before the COVID-19 pandemic. Front Public Health 2022;10:981270 [FREE Full text] [doi: 10.3389/fpubh.2022.981270] [Medline: 36388332]
7. Filardi F, de Castro RMP, Zanini MTF. Advantages and disadvantages of teleworking in Brazilian public administration: analysis of SERPRO and Federal Revenue experiences. Cad EBAPE BR 2020;18(1):28-46 [FREE Full text] [doi: 10.1590/1679-395174605x]
8. Liu Z, Van Egdom D, Flin R, Spitzmueller C, Adepoju O, Krishnamoorti R. I don't want to go back: examining the return to physical workspaces during COVID-19. J Occup Environ Med 2020;62(11):953-958 [FREE Full text] [doi: 10.1097/JOM.0000000000002012] [Medline: 32858556]
9. Beckel JLO, Fisher GG. Telework and worker health and well-being: a review and recommendations for research and practice. Int J Environ Res Public Health 2022;19(7):3879 [FREE Full text] [doi: 10.3390/ijerph19073879] [Medline: 35409563]
10. Ueno H, Ishimaru T, Matsugaki R, Ando H, Mafune K, Nagata T, et al. Occupational factors associated with telemedicine use in the Japanese working-age population: a web-based study conducted during the COVID-19 pandemic. Tohoku J Exp Med 2023;259(2):143-150 [FREE Full text] [doi: 10.1620/tjem.2022.J107] [Medline: 36476586]
11. Etheridge JC, Sinyard RD, Brindle ME. Chapter 90—Implementation research. In: Eltorai AEM, Bakal JA, Newell PC, Osband AJ, editors. Translational Surgery. London: Academic Press; 2023:563-573.
12. Das A, Shah S, Adhikari TB, Paudel BS, Sah SK, Das RK, et al. Computer vision syndrome, musculoskeletal, and stress-related problems among visual display terminal users in Nepal. PLoS One 2022;17(7):e0268356 [FREE Full text] [doi: 10.1371/journal.pone.0268356] [Medline: 35853006]
13. Sultana A, Tasnim S, Hossain MM, Bhattacharya S, Purohit N. Digital screen time during the COVID-19 pandemic: a public health concern. F1000Res 2021;10:81 [FREE Full text] [doi: 10.12688/f1000research.50880.1]
14. Alkan M, Meinck S. The relationship between students' use of ICT for social communication and their computer and information literacy. Large-scale Assess Educ 2016;4(1):15 [FREE Full text] [doi: 10.1186/s40536-016-0029-z]
15. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
16. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441 [FREE Full text] [doi: 10.1136/bmj.m441] [Medline: 32188600]
17. Ma Z, Wu M. The psychometric properties of the Chinese eHealth Literacy Scale (C-eHEALS) in a Chinese rural population: cross-sectional validation study. J Med Internet Res 2019;21(10):e15720 [FREE Full text] [doi: 10.2196/15720] [Medline: 31642811]
18. Lok KYW, Fong DYT, Wong JYH, Ho M, Choi EP, Pandian V, et al. International survey for assessing COVID-19's impact on fear and health: study protocol. BMJ Open 2021 May 11;11(5):e048720 [FREE Full text] [doi: 10.1136/bmjopen-2021-048720] [Medline: 33980534]
19. Panlilio LV, Lee A, Smith KE, Epstein DH. Were self-described introverts "immune" to increased drug use and entrapment during the pandemic? Drug Alcohol Depend Rep 2022;2:100024 [FREE Full text] [doi: 10.1016/j.dadr.2022.100024] [Medline: 35156103]
20. Tibbetts M, Epstein-Shuman A, Leitao M, Kushlev K. A week during COVID-19: online social interactions are associated with greater connection and more stress. Comput Hum Behav Rep 2021;4:100133 [FREE Full text] [doi: 10.1016/j.chbr.2021.100133] [Medline: 34693074]
21. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
22. Choi EPH, Duan W, Fong DYT, Lok KYW, Ho M, Wong JYH, et al. Psychometric evaluation of a fear of COVID-19 scale in China: cross-sectional study. JMIR Form Res 2022;6(3):e31992 [FREE Full text] [doi: 10.2196/31992] [Medline: 35072632]
23. Li S, Fong DYT, Wong JYH, McPherson B, Lau EYY, Huang L, et al. Noise sensitivity associated with nonrestorative sleep in Chinese adults: a cross-sectional study. BMC Public Health 2021;21(1):643 [FREE Full text] [doi: 10.1186/s12889-021-10667-2] [Medline: 33794830]
24. Urien B. Teleworkability, preferences for telework, and well-being: a systematic review. Sustainability 2023;15(13):10631 [FREE Full text] [doi: 10.3390/su151310631]
25. Gohoungodji P, N'Dri AB, Matos ALB. What makes telework work? evidence of success factors across two decades of empirical research: a systematic and critical review. Int J Hum Resour Manag 2022;34(3):605-649. [doi: 10.1080/09585192.2022.2112259]
26. Baruch Y, Nicholson N. Home, sweet work: requirements for effective home working. J Gen Manag 2017;23(2):15-30. [doi: 10.1177/030630709702300202]
27. Godfree K. "I always feel that I'm underestimated": Obesity Stigma in the Workplace. Great Britain: Kingston University; 2020. URL: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.etho [accessed 2024-05-22]
28. Otsuka S, Ishimaru T, Nagata M, Tateishi S, Eguchi H, Tsuji M, et al. A cross-sectional study of the mismatch between telecommuting preference and frequency associated with psychological distress among Japanese workers in the COVID-19

XSL•FO
RenderX

pandemic. J Occup Environ Med 2021;63(9):e636-e640 [FREE Full text] [doi: 10.1097/JOM.0000000000002318] [Medline: 34491971]
29. do Nascimento IJB, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
30. Fernández-Theoduloz G, Paz V, Nicolaisen-Sobesky E, Pérez A, Buunk AP, Cabana Á, et al. Social avoidance in depression: a study using a social decision-making task. J Abnorm Psychol 2019;128(3):234-244. [doi: 10.1037/abn0000415] [Medline: 30920233]
31. Chau JY, van der Ploeg HP, Merom D, Chey T, Bauman AE. Cross-sectional associations between occupational and leisure-time sitting, physical activity and obesity in working adults. Prev Med 2012;54(3-4):195-200. [doi: 10.1016/j.ypmed.2011.12.020] [Medline: 22227284]
32. Li X, Vanderloo LM, Keown-Stoneman CDG, Cost KT, Charach A, Maguire JL, et al. Screen use and mental health symptoms in Canadian children and youth during the COVID-19 pandemic. JAMA Netw Open 2021;4(12):e2140875 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.40875] [Medline: 34962557]
33. López-Igual P, Rodríguez-Modroño P. Who is teleworking and where from? exploring the main determinants of telework in Europe. Sustainability 2020;12(21):8797 [FREE Full text] [doi: 10.3390/su12218797]
34. Wang B, Liu Y, Qian J, Parker SK. Achieving effective remote working during the COVID-19 pandemic: a work design perspective. Appl Psychol 2021;70(1):16-59 [FREE Full text] [doi: 10.1111/apps.12290] [Medline: 33230359]
35. Bendau A, Petzold MB, Pyrkosch L, Maricic LM, Betzler F, Rogoll J, et al. Associations between COVID-19 related media consumption and symptoms of anxiety, depression and COVID-19 related fear in the general population in Germany. Eur Arch Psychiatry Clin Neurosci 2021;271(2):283-291 [FREE Full text] [doi: 10.1007/s00406-020-01171-6] [Medline: 32691135]
36. Almarcha M, Balagué N, Torrents C. Healthy teleworking: towards personalized exercise recommendations. Sustainability 2021;13(6):3192 [FREE Full text] [doi: 10.3390/su13063192]
37. Darvish M, Keresztyen L, Bick M. Virtual immersive workplaces: the new norm?—A qualitative study on the impact of VR in the workplace. In: Papadaki M, Rupino da Cunha P, Themistocleous M, Christodoulou K, editors. Information Systems: 19th European, Mediterranean, and Middle Eastern Conference, EMCIS 2022, Virtual Event, December 21–22, 2022, Proceedings. Cham: Springer; 2023:665-677.
38. Heise TL, Frense J, Christianson L, Seuring T. Using financial incentives to increase physical activity among employees as a strategy of workplace health promotion: protocol for a systematic review. BMJ Open 2021;11(3):e042888 [FREE Full text] [doi: 10.1136/bmjopen-2020-042888] [Medline: 33674371]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
OR: odds ratio

Edited by S Ma, T Leung; submitted 03.10.23; peer-reviewed by Anonymous, M Ghayour-Mobarhan, Z Ye; comments to author 15.03.24; revised version received 22.03.24; accepted 25.04.24; published 28.05.24.

Please cite as:
Li J, Fong DYT, Ho MM, Choi EPH, Lok KYW, Lee JJ, Duan W, Wong JYH, Lin CC
Prevalence and Factors Associated With Willingness to Sustain Pandemic-Induced Digital Work in the General Population and Moderating Effects of Screen Hours: Cross-Sectional Study
J Med Internet Res 2024;26:e53321
URL: https://www.jmir.org/2024/1/e53321
doi:10.2196/53321
PMID:38805704

©Jiaying Li, Daniel Yee Tak Fong, Mandy Man Ho, Edmond Pui Hang Choi, Kris Yuet Wan Lok, Jung Jae Lee, WenJie Duan, Janet Yuen Ha Wong, Chia-Chin Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identifying Reddit Users at a High Risk of Suicide and Their Linguistic Features During the COVID-19 Pandemic: Growth-Based Trajectory Model

Yifei Yan[1], BA; Jun Li[2], MA; Xingyun Liu[3], PhD; Qing Li[2], PhD; Nancy Xiaonan Yu[1], PhD

[1]Department of Social and Behavioural Sciences, City University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[3]Key Laboratory of Adolescent Cyberpsychology and Behavior, Central China Normal University, Ministry of Education, School of Psychology, Wuhan, China

Corresponding Author:
Nancy Xiaonan Yu, PhD
Department of Social and Behavioural Sciences
City University of Hong Kong
Tat Chee Avenue, Kowloon, HKSAR, P. R. China
Hong Kong, 000
China (Hong Kong)
Phone: 852 34429436
Fax: 852 34420283
Email: nancy.yu@cityu.edu.hk

Abstract

Background: Suicide has emerged as a critical public health concern during the COVID-19 pandemic. With social distancing measures in place, social media has become a significant platform for individuals expressing suicidal thoughts and behaviors. However, existing studies on suicide using social media data often overlook the diversity among users and the temporal dynamics of suicide risk.

Objective: By examining the variations in post volume trajectories among users on the r/SuicideWatch subreddit during the COVID-19 pandemic, this study aims to investigate the heterogeneous patterns of change in suicide risk to help identify social media users at high risk of suicide. We also characterized their linguistic features before and during the pandemic.

Methods: We collected and analyzed post data every 6 months from March 2019 to August 2022 for users on the r/SuicideWatch subreddit (N=6163). A growth-based trajectory model was then used to investigate the trajectories of post volume to identify patterns of change in suicide risk during the pandemic. Trends in linguistic features within posts were also charted and compared, and linguistic markers were identified across the trajectory groups using regression analysis.

Results: We identified 2 distinct trajectories of post volume among r/SuicideWatch subreddit users. A small proportion of users (744/6163, 12.07%) was labeled as having a high risk of suicide, showing a sharp and lasting increase in post volume during the pandemic. By contrast, most users (5419/6163, 87.93%) were categorized as being at low risk of suicide, with a consistently low and mild increase in post volume during the pandemic. In terms of the frequency of most linguistic features, both groups showed increases at the initial stage of the pandemic. Subsequently, the rising trend continued in the high-risk group before declining, while the low-risk group showed an immediate decrease. One year after the pandemic outbreak, the 2 groups exhibited differences in their use of words related to the categories of personal pronouns; affective, social, cognitive, and biological processes; drives; relativity; time orientations; and personal concerns. In particular, the high-risk group was discriminant in using words related to anger (odds ratio [OR] 3.23, $P<.001$), sadness (OR 3.23, $P<.001$), health (OR 2.56, $P=.005$), achievement (OR 1.67, $P=.049$), motion (OR 4.17, $P<.001$), future focus (OR 2.86, $P<.001$), and death (OR 4.35, $P<.001$) during this stage.

Conclusions: Based on the 2 identified trajectories of post volume during the pandemic, this study divided users on the r/SuicideWatch subreddit into suicide high- and low-risk groups. Our findings indicated heterogeneous patterns of change in suicide risk in response to the pandemic. The high-risk group also demonstrated distinct linguistic features. We recommend conducting real-time surveillance of suicide risk using social media data during future public health crises to provide timely support to individuals at potentially high risk of suicide.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e48907) doi:10.2196/48907

KEYWORDS

COVID-19 pandemic; Reddit; suicide risk; trajectory

Introduction

The COVID-19 pandemic has triggered a global increase in mental disorders [1], with heightened concerns about suicide risk both in the short and in the long term [2,3]. While an interrupted time-series study indicated that most countries or regions have not experienced a significant rise in suicide deaths [4], the prevalence of suicidal ideation (10.81%-12.10%) and suicide attempts (4.86%) has notably increased compared with prepandemic levels [5,6]. Our recent meta-analysis, utilizing longitudinal data, has further corroborated an increase in the prevalence of suicidal ideation and suicide attempts among both nonclinical and clinical populations following the onset of the pandemic [7]. Suicide remains a critical public health concern throughout the pandemic, underscoring the need for ongoing monitoring and vigilance as the situation continues to evolve [8,9]. In this study, we utilized posts from a social media forum to examine patterns of change in suicide risk during the COVID-19 pandemic, analyzing their corresponding linguistic features.

Social media and online discussion forums are increasingly recognized as valuable resources for suicide research [10,11]. These virtual platforms offer users open and widely accessible spaces to share experiences, engage in discussions, seek social support, and exchange information anonymously and in real time [12,13]. Compared with traditional clinical data sets, these online data sets offer several advantages: they are publicly and freely available, involve larger sample sizes, provide access to participants who are ordinarily difficult to engage, enable comparisons with historical data, and exhibit high ecological validity through documentation of first-person experiences [14]. Utilizing social media data from platforms such as Twitter/X (X Corp.), Reddit (Reddit, Inc.), and Weibo (Weibo Corporation), previous studies have detected and predicted suicide risk and developed intervention programs [15-21]. As a result of lockdown measures and heightened concerns during the pandemic, social media data have become even more valuable for suicide research, as these online platforms have become the primary means for many people to receive information and stay connected with the outside world [22,23].

The r/SuicideWatch subreddit is a semianonymous forum that provides “peer support for anyone struggling with suicidal thoughts or worried about someone who may be at risk” [24]. This makes it an important resource for suicide research, as it offers high-quality, self-reported suicide data. Dominant suicide risk assessments for online posts, such as machine learning models combined with manual coding, typically approach it as a multiclassification task. These models output the post-level suicide risk (ie, the risk associated with a single post) using ordinal data. For example, they might label a post as being at ideation, behavior, or attempt level based on the probability score for each level [25-27]. However, such data cannot capture user-level suicide risk (ie, the risk of a user based on posts during a specific period) and may not conform to the distributional assumptions of many statistical analyses, such as growth-based trajectory models, potentially introducing biases [28,29]. To assess users’ overall risk and track temporal changes in risk, researchers can consider both the quantity and quality of posts published, including methods such as topic modeling or linguistic analysis of content [30,31]. Specifically, the quantitative method focuses on the total number of posts (ie, post volume) within a specific period, which indicates users’ posting activity and social engagement level [32,33]. Changes in users’ post volume can reflect changes in their suicide risk. For instance, the diurnal and weekly patterns of post volume on r/SuicideWatch corresponded to temporal fluctuations in suicide risk, including actual suicide attempts or deaths [34]. Similarly, Twitter users exhibited an increased volume of suicide-related posts before their suicide attempts [35]. Furthermore, analyses of both the quantity and quality of posts have demonstrated that higher post volume (quantity) corresponds to active disclosure of suicidal thoughts in post content (quality) [30,31]. Therefore, post volume in online suicide communities can serve as an effective indicator of user-level suicide risk, offering sufficient accessibility and flexibility for statistical analysis.

In this study, we monitored users’ post volume on the r/SuicideWatch subreddit before and during the pandemic to observe changes in their suicide risk. Among active adolescent users during the pandemic (April 2020-September 2021), post volume remained stable compared with prepandemic periods [36]. However, when examining post volume over shorter intervals, it fluctuated and demonstrated an overall decrease until December 2020 [37,38]. There is currently no documented study on how post volume has evolved beyond 2020. It is important to note that users of r/SuicideWatch may vary in their levels of suicide risk, and the findings mentioned above could be ambiguous without accounting for this heterogeneity among users. Through expert annotation, active r/SuicideWatch users (ie, those with at least 10 total posts) were categorized into 4 risk levels: no risk (36/245, 15%), low risk (50/245, 20%), moderate risk (115/245, 47%), and severe risk (44/245, 18%) of suicide [39]. However, each user’s risk label was determined based on the highest risk observed in their posts during a 7-year period before the pandemic. As suggested by the fluid vulnerability theory [40], individual suicide risk is best understood as a temporal process influenced by both baseline and acute risk factors. Environmental stressors or contexts such as sudden outbreaks of infectious disease epidemics or pandemics, social isolation, and fear [3,7] can easily trigger acute suicide risk in individuals predisposed to underlying vulnerabilities (ie, those with a higher baseline risk). Cognitive, emotional, behavioral, and physiological factors interact to either sustain or alleviate their suicide risk [41]. Considering both stability and dynamism, some individuals exhibit fluctuations in suicide risk, moving between low and high states where suicidal behavior may be more or less likely to emerge,

XSL•FO
RenderX

while others display a stable pattern. To address the heterogeneity among users and understand the temporal nature of suicide risk, using trajectory modeling techniques such as group-based trajectory modeling (GBTM) and growth curve modeling can be beneficial. These methods identify subgroups within a population that share similarities in outcomes over time [42].

The linguistic or language styles used in posts or comments can offer valuable insights into the experiences and perspectives of suicidal individuals. This can assist researchers in understanding the underlying thoughts, emotions, and behaviors that individuals may be unwilling or unable to express explicitly [43]. Importantly, linguistic features can act as markers that distinguish suicide-related posts from general posts, aiding in the identification of potential high-risk users who may need support [17,44,45]. Suicide-related social media posts often exhibit characteristics such as simplicity in words and short sentences, reduced lexical diversity, and language disorganization [46]. They may also include more statements related to self-destruction, commands, and conflicts [47], along with increased use of first-person pronouns, adverbs, and multifunctional words. These posts frequently reference death, anger, and the present moment, while showing fewer occurrences of second- and third-person pronouns, nouns, and references to causes and differentiation [17,48,49]. However, it remains unclear whether high-risk users have exhibited these linguistic features during the COVID-19 pandemic, which has introduced different stressors potentially influencing suicide risk. Existing studies during the pandemic that utilized psycholinguistic analysis of r/SuicideWatch posts have primarily concentrated on monitoring temporal shifts in these linguistic features. Specifically, studies have identified increased use of words associated with negative emotions and a focus on the past, along with fewer references to positive emotions, social interactions, and leisure activities. References to death and first-person pronouns remained stable [36,37]. However, these findings only encompassed a limited time frame during the pandemic (until September 2021) and did not account for the heterogeneity among users in terms of suicide risk. Different users may exhibit varying linguistic characteristics and trends in linguistic changes over time. Therefore, it is crucial to first identify groups of users exhibiting similar patterns of suicidal behavior throughout the pandemic. Subsequently, analyzing their respective linguistic trends and markers can yield valuable insights for suicide surveillance and targeted interventions among high-risk users.

Utilizing posts from the r/SuicideWatch subreddit before and during the COVID-19 pandemic (March 2019-September 2022), this study aimed to investigate the following: (1) the potential for distinct patterns of change in suicide risk using GBTM of users' post volumes, (2) the trends and characteristics of linguistic features within posts across each trajectory group, and (3) the linguistic markers associated with each trajectory group. We anticipated that identifying trajectories of post volume on the r/SuicideWatch subreddit would uncover users' diversity by accounting for the temporal dynamics of suicide risk. Analyzing their associated linguistic features could also help identify users potentially at high risk of suicide. These findings could have significant implications for enhancing suicide screening, monitoring, and interventions during future public health crises.

Methods

Data Set and Participants

For this study, we gathered the longitudinal data set using the Reddit application programming interface [50]. Following a previously established method [14], we crawled posts from users who contributed to the r/SuicideWatch subreddit between March 1, 2020, and August 31, 2022, resulting in a total of 603,802 posts from 6943 users. We expanded our data set by retrieving historical posts dating back to March 1, 2019, to analyze the trajectory of post volume before and during the pandemic. To streamline data usage in subsequent analyses, we excluded accounts that were canceled and posts with deleted content. The final data set comprised 6163 users and their posts from the r/SuicideWatch subreddit (N=33,714) spanning the period from March 1, 2019, to August 31, 2022, encompassing the COVID-19 pandemic period. As a result of the onset of the COVID-19 pandemic around March 2020, there was a notable increase in discussions related to the pandemic on Reddit [37]. Therefore, we used March 1, 2020, as a cutoff point and defined 2 prepandemic periods (T1: March 1, 2019-August 31, 2019 and T2: September 1, 2019-February 29, 2020) and 5 peripandemic periods (T3: March 1, 2020-August 31, 2020; T4: September 1, 2020-February 28, 2021; T5: March 1, 2021-August 31, 2021; T6: September 1, 2021-February 28, 2022; and T7: March 1, 2022-August 31, 2022) to track post volume trajectories across these time frames.

Trajectory Variable: Post Volume

To assess changes in suicide risk, we used post volume from the r/SuicideWatch subreddit within each pre- and peripandemic period for each user as a proxy for the trajectory variable [30,31,34,35,51]. We quantified the number of posts made by each user during each specific period, with periods where no posts were made recorded as 0. For users who joined r/SuicideWatch after March 1, 2020 (ie, those who began posting suicide-related content on r/SuicideWatch only after the pandemic outbreak; 5759/6163, 93.44%) [48], their post counts on r/SuicideWatch during the 2 prepandemic periods and the peripandemic periods before their initial post were recorded as 0. As noted by De Choudhury et al [48], there are a few suicide-related posts found on subreddits outside of r/SuicideWatch. The transition of these users to r/SuicideWatch may indicate the onset and progression of their suicidal concerns following the pandemic outbreak. Therefore, their 0 post volume during the pre- and peripandemic periods can serve as a proxy for their respective suicide risk trajectories, reflecting changes in their behavioral patterns and suicide risk following the pandemic. According to our eligibility criteria, all users included in the study had posted at least once across all periods.

There might be a concern that some of our collected posts discussed the suicide risk of others rather than the user's own risk, although previous studies have indicated that posts on r/SuicideWatch primarily focus on self-directed concerns [30,48]. To address this issue in our data set, we randomly

XSL•FO
RenderX

selected 10% (3372/33,714) of the total posts and manually screened the content to determine the subject of the posts. Among the 3372 posts, only 18 (0.53%) were found not to be about the user's own suicide-related issues: 6 discussed others' suicide risk, 5 provided help to others, 3 were about irrelevant topics, and 4 were unclassified. Thus, the majority of our collected posts accurately reflected users' own experiences and concerns related to suicide risk.

Linguistic Features

To analyze the linguistic features of posts, we used Linguistic Inquiry and Word Count (LIWC) 2015 [52], a widely used tool for language analysis. LIWC encompasses more than 80 word categories, each containing hundreds of dictionary words for the identification and analysis of word use patterns related to suicide risk. The primary psycholinguistic categories in LIWC are personal pronouns and words related to affective, social, cognitive, perceptual, and biological processes; drives; time orientations; relativity; and personal concerns. For each Reddit user included in our study, we calculated LIWC measures for these 10 major psycholinguistic categories during each period [53]. First, we tallied the occurrences of each word within a specific post alongside the post's length (ie, the total number of words used). Second, we summed the occurrences of each word and the total length of posts for each period. Finally, for each period, we calculated the normalized frequency of word use in each LIWC category by dividing the total count of the LIWC category by the total length of posts in that period. Therefore, each LIWC measure represents the normalized frequency of word use within a specific LIWC category during each period analyzed.

Statistical Analysis

GBTM was used to define trajectory groups based on post volume across the COVID-19 pandemic. As a finite mixture model, GBTM is capable of identifying distinct groups of individuals with similar developmental trajectories in a particular outcome or behavior within a population. It accommodates trajectory variables that adhere to distributions such as censored normal (CNORM), zero-inflated Poisson (ZIP), beta, and Bernoulli distributions [54]. Within each period, the exploratory analysis revealed that a large number of eligible users had 0 posts, resulting in a skewed and zero-inflated distribution of the trajectory variable (ie, post volume within each period). Among the available models, the ZIP model was selected to fit our data because of its capability to address excessive zeros. The ZIP model combines an inflation model for zeros with a count model for nonzero values, making it suitable for our data set [54,55].

To identify the best-fitting model with the optimal number of trajectory groups, we followed 3 steps [56]. Initially, we incrementally increased the number of group specifications from 2 to 5 to pinpoint the optimal number of trajectories. Specifically, we selected the model based on 4 commonly used fit statistics in GBTM analysis [56-58]: the Bayesian information criterion (BIC), the Akaike information criterion (AIC), entropy, and group composition. AIC relies on information theory to assess the relative information value of the model by considering the maximum likelihood estimate and the number of parameters within the model [59]. Similar to AIC, BIC originates from the Bayesian framework and can be interpreted as the posterior probability of a model based on the observed data [60]. Both statistics aim to identify the most informative model by balancing between goodness-of-fit and model complexity. However, BIC imposes a stronger penalty for model complexity compared with AIC, taking into account the sample size [61]. The goodness-of-fit and penalty terms are summed to compute AIC and BIC values, where smaller values indicate better-fitting models [62-64]. Additionally, entropy assesses the classification accuracy of the model by summarizing the likelihood of each participant being correctly classified [57]. With values ranging from 0 to 1, higher entropy values indicate more precise classification, typically considered satisfactory when exceeding 0.8 [65]. We also analyzed the group composition (ie, the percentage of the population represented in each subgroup), ensuring that each subgroup represented at least 5% of the total sample [58]. Second, we determined the shapes of each trajectory by specifying their functional forms (eg, linear and cubic). Starting with a cubic specification (up to 3 degrees), we iteratively dropped nonsignificant ($P>.05$) polynomial terms until only significant ones remained [56]. In the count model part, linear terms were retained regardless of their statistical significance. Third, after identifying the optimal number of trajectories and their shapes, we used the average posterior probabilities (APPs) of group membership to validate the selected model. The APP measures the average probability of each participant belonging to their assigned group and should ideally be at least 0.7 for each group to ensure robustness [58].

After identifying the best-fitting model, we assigned users to their respective trajectory groups. Subsequently, summary descriptive statistics of linguistic features during each period were computed and graphed for each user group. Specifically, we compared linguistic frequencies between groups across different periods. As a result of the skewed and excessively zero-inflated distribution of LIWC frequency, parametric tests such as *t* tests or ANOVA may not be appropriate, as they violate their assumptions and can reduce the robustness of nonparametric tests such as the median or Wilcoxon-Mann-Whitney test. Therefore, Poisson regression modeling was recommended for its improved interpretability of data and comparability among potential models [66]. In the Poisson regression model, group membership was included as an independent variable, while the frequency of each linguistic feature in each period served as the dependent variable. The results of the Poisson model provided rate ratios (RRs) along with SEs, indicating the relative changes in counts of the outcomes between the groups.

In the final analysis, our goal was to identify linguistic markers that could differentiate between groups of users exhibiting different trajectories. In the context of suicidal text analysis, linguistic markers, or linguistic distinguishers, are language features (eg, LIWC categories) extracted from texts that substantially distinguish users or posts with varying risk statuses [17,49]. To identify these features, we analyzed the linguistic profiles of high- and low-risk users and examined which words could indicate their respective suicide risk levels. As the markers were not intended to predict users' group membership, a temporal sequence (eg, baseline linguistic data) was not

XSL•FO
RenderX

necessary. We utilized data from the 7 periods and followed these steps to identify potential linguistic markers. First, we identified linguistic features from the periods that exhibited significant ($P \leq .05$) between-group differences based on the results of Poisson regression, considering them as potential markers. Second, to prevent duplication, we excluded high-level LIWC categories that have hierarchical relationships with each other (eg, ppron includes I, we, you, she/he, they). Third, we used Lasso logistic regression with cross-validation to determine the optimal penalty parameter, aiming to mitigate collinearity among the remaining linguistic measures [67]. Given the lack of a theoretical basis for estimating post-selection coefficients with nonlinear Lasso models [67], we utilized the Lasso model solely for model selection purposes to filter out redundant variables. The remaining variables were then integrated into the best-fitting GBTM, and their associations with group membership were assessed using multivariate logistic regression [68]. In this analysis, we computed odds ratios (ORs), where potential linguistic features served as independent variables and group memberships as the dependent variable. Significance levels were determined at a P value $\leq .05$. Data extraction was performed using Python (Python Foundation), while data analysis was conducted using Stata/SE 16.1 (StataCorp LLC) along with the Traj plugin for trajectory analysis.

Ethical Considerations

The data used in this study were obtained from publicly accessible posts on the r/SuicideWatch subreddit through purely observational and nonintrusive means. The raw data did not contain personally identifiable information. To uphold user privacy and confidentiality, selected posts were deidentified before analysis. This involved removing any identifying information such as names, genders, ages, addresses, and links from the post content. We maintained annotated user data separately from the raw data and stored them on secure servers, linked only through anonymous IDs. Furthermore, all examples presented in Multimedia Appendix 1 were anonymized and paraphrased to safeguard user privacy, following the framework outlined by Bruckman [69]. As publicly available data were utilized, this study fell outside the purview of ethical review by the City University of Hong Kong Research Committee, for which an exemption was obtained.

Results

Post Trajectories on the r/SuicideWatch Subreddit Throughout the COVID-19 Pandemic

To determine post trajectories, we evaluated the model fit statistics for 2- to 5-group solutions of the GBTM to identify the optimal number of trajectory groups (Table 1). As the number of groups increased from 2, we noted that both the AIC and BIC values tended to increase, while entropy decreased. Additionally, starting from the 3-group model, some group compositions did not meet the 5% threshold. Therefore, we selected the 2-group model as the optimal choice, with AIC and BIC values of –39,605.31 and –39,659.12, respectively, and an entropy of 0.96. Further analysis indicated that the 2-group solution, using cubic and quadratic functions in the count model and 2 cubic functions in the inflation model, resulted in all polynomial terms being statistically significant (Table 2). The APPs for groups 1 and 2 were 0.95 and 0.99, respectively, indicating strong alignment between users and their assigned groups within this 2-group ZIP model.

Table 1. Fit statistics for the 2- to 5-group solution–based trajectory modeling of post volume among r/SuicideWatch users throughout the COVID-19 pandemic.

Model	AIC[a]	BIC[b]	Entropy[c]	Composition[d], n/N (%)
2-Group model	–39,605.31	–39,659.12	0.96	744/6163 (12.07)/5419/6163 (87.93)
3-Group model	–36,862.86	–36,940.22	0.892	1562/6163 (25.34)/4379/6163 (71.05)/222/6163 (3.60)
4-Group model	–35,890.32	–35,981.12	0.836	625/6163 (10.14)/1502/6163 (24.37)/3936/6163 (63.87)/100/6163 (1.62)
5-Group model	–36,099.67	–36,190.48	0.873	293/6163 (4.75)/1699/6163 (27.57)/91/6163 (1.48)/4003/6163 (64.95)/77/6163 (1.25)

[a]AIC: Akaike information criterion (a lower value is better).

[b]BIC: Bayesian information criterion (a lower value is better).

[c]Entropy (a value >0.8 is better).

[d]Group composition (the percentage of the population represented in each subgroup should exceed 5%).

XSL•FO
RenderX

Table 2. Parameter estimates for the 2-group zero-inflated group–based trajectory modeling of post volume among r/SuicideWatch users throughout the COVID-19 pandemic.

Group and parameter		Estimate	SE	*t* value (*df*)	*P* value
1 (count)					
	Intercept	2.62	0.01	214.7 (743)	<.001
	Linear	0.1	0.03	3.63 (743)	<.001
	Quadratic	–0.44	0.02	–24.52 (743)	<.001
	Cubic	–0.22	0.02	–10.56 (743)	<.001
2 (count)					
	Intercept	0.59	0.01	42.46 (5418)	<.001
	Linear	–0.14	0.02	–5.99 (5418)	<.001
	Quadratic	–0.12	0.02	–5.07 (5418)	<.001
1 (inflation)					
	Alpha0	–0.19	0.05	–3.78 (743)	<.001
	Alpha1	–1.85	0.1	–18.4 (743)	<.001
	Alpha2	1.49	0.06	24.17 (743)	<.001
	Alpha3	0.75	0.07	10.8 (743)	<.001
2 (inflation)					
	Alpha0	–0.34	0.03	–13.16 (5418)	<.001
	Alpha1	–2.31	0.06	–37.54 (5418)	<.001
	Alpha2	2.63	0.04	58.83 (5418)	<.001
	Alpha3	1.49	0.05	27.77 (5418)	<.001

Figure 1 depicts the post volume trajectories across the COVID-19 pandemic for the 2 identified groups of users. Group 1, designated as the "high risk of suicide" group, consisted of 744 (12.07%) users. Their post volume on r/SuicideWatch showed a gradual increase during the 2 prepandemic periods, followed by a rapid acceleration after the pandemic began. This trend peaked approximately 1 year after the pandemic outbreak and subsequently declined, returning to its initial level during the second year of the pandemic. Group 2, identified as the "low risk of suicide" group, comprised the majority of users (5419/6163, 87.93%). This group exhibited a slight increase in post volume on r/SuicideWatch following the pandemic outbreak, followed by stabilization and eventual recovery. Throughout the pandemic, group 2 maintained a relatively low post volume on the subreddit.

Figure 1. Trajectories of post volume among r/SuicideWatch users throughout the COVID-19 pandemic.

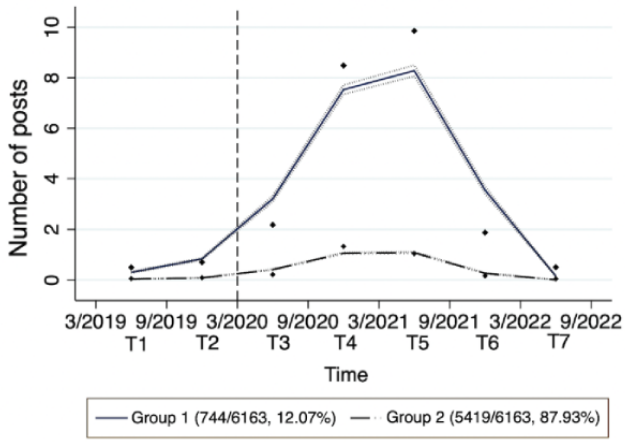

Linguistic Feature Analysis

The summary distribution of frequency of use in LIWC for the 2 groups can be found in Multimedia Appendix 2. Similar to the distribution of post volume, we observed a zero-inflated phenomenon for these linguistic features across periods. Therefore, descriptive statistics including the median, first quartile, and third quartile were used. By plotting the median frequency trend for each included LIWC feature throughout the pandemic for the 2 groups, trends by category were illustrated

XSL•FO
RenderX

(Multimedia Appendix 3). We observed the following: (1) During the year before the pandemic (T1: March 2019-September 2019 and T2: September 2019-March 2020), word frequency was generally low for both groups. (2) Throughout the pandemic (T3-T6: March 2020-March 2022), words related to cognitive processes, perceptual processes, biological processes, and personal concerns showed relatively lower frequency compared with personal pronouns, affective processes, social processes, drives, relativity, and time orientations. (3) During the first year of the pandemic (T3-T4: March 2020-March 2021), both groups exhibited sharp increases in word frequency. (4) During the second year of the pandemic (T5-T6: March 2021-March 2022), the high-risk group continued to experience a slower increase until reaching a peak and subsequent decrease, while the low-risk group's frequency decreased. (5) Moving into the third year of the pandemic (T7: March 2022-September 2022), word frequency returned to prepandemic levels in both groups. Despite both groups showing increased use of most word types during the pandemic, the high-risk group exhibited a longer-lasting increase with a peak lagging behind that of the low-risk group. This suggests that the pandemic had a more enduring impact on high-risk users.

The results of between-group comparisons using Poisson regression (with the low-risk group as the reference) are depicted in Figure 2. In general, the high-risk group utilized most types of words more frequently than the low-risk group both before and during the initial 6 months of the pandemic (illustrated in red for T1-T3). Later, in the second half of the pandemic, their differences narrowed and even reversed (as shown in green during T4), with both groups demonstrating increased word use. Subsequently, the high-risk group once again surpassed the low-risk group, and these differences grew larger in the subsequent periods (as indicated in deeper red from T5 to T7). This pattern corresponded with the plotted trend, where the high-risk group exhibited a prolonged increase and a delayed peak following the rise during T4, whereas the frequency of the low-risk group quickly decreased and returned to its initial level.

XSL•FO
RenderX

Figure 2. Visualization of Poisson regression results comparing LIWC frequency between r/SuicideWatch users in the high- and low-risk groups throughout the COVID-19 pandemic (low-risk group as a reference; red: higher frequency in the high-risk group; green: lower frequency in the high-risk group). LIWC: Linguistic Inquiry and Word Count. **P*<.05; ***P*<.01; ****P*<.001.

	T1 3/2019-9/2019	T2 9/2019-3/2020	T3 3/2020-9/2020	T4 9/2020-3/2021	T5 3/2021-9/2021	T6 9/2021-3/2022	T7 3/2022-9/2022
Personal pronouns							
Personal pronouns	1.79	2.09	2.09*	0.97	1.51*	3.95***	3.61*
First-person person singular	1.93	2.16	2.13	0.99	1.54*	4.27***	3.75*
First-person person plural	2.39	1.57	1.5	0.89	1.25	3.3	3.46
Second person	1.32	1.98	2.6	1.09	1.57	3.2	3.47
Third-person singular	1.16	2.15	1.79	0.72	1.24	3.16	2.1
Third-person plural	1.74	1.73	1.86	0.95	1.55	2.7	4.14
Affective processes							
Affective processes	2.16	2.01	2.11*	1.01	1.6**	3.91***	3.98*
Positive emotions	1.9	1.75	2.16	1	1.51	3.46*	3.47
Negative emotions	2.32	2.16	2.08	1.02	1.66*	4.18***	4.3*
Anxiety	3.08	1.51	1.6	0.86	1.42	4.05	5.36
Anger	2.3	2.51	2.4	1.13	1.84	4.14*	4.4
Sad	1.98	2.01	2.05	0.99	1.56	4.32*	3.45
Social processes							
Social processes	1.73	1.76	1.89	0.9	1.43	3.1**	3.04
Family	2.37	1.64	1.31	0.75	1.16	2.6	4.24
Friend	2.68	2.18	1.69	0.76	1.22	2.75	3.26
Female	1.83	2.38	1.7	0.8	1.32	2.66	2.12
Male	1.3	1.63	1.74	0.75	1.12	3.46	3.55
Cognitive processes							
Cognitive processes	1.91	1.84	2.02	0.94	1.49	3.58**	3.63
Insight	1.63	1.75	1.86	0.86	1.48	3.33	4
Cause	2.11	2.01	2.04	0.96	1.48	3.43	2.7
Discrepancy	1.89	1.82	2.2	0.99	1.56	3.63	3.79
Tentativeness	1.79	1.66	2.16	0.92	1.48	3.54	3.57
Certainty	2.36	1.99	2.03	1.03	1.53	3.96	3.98
Differentiation	1.89	1.89	1.99	0.94	1.43	3.66	3.8
Perceptual processes							
Perceptual processes	1.54	2.18	2.02	0.95	1.51	3.58	4.09
See	1.7	2.41	2.15	0.97	1.53	3.77	4.86
Hear	1.66	1.7	2.02	0.97	1.47	2.79	4.62
Feel	1.38	2.18	1.92	0.92	1.51	3.76	3.6
Biological processes							
Biological processes	2.23	2.43	2.17	1.01	1.55	3.76**	4.68
Body	2.96	2.61	2.37	1.15	1.75	4.83	5.06
Health	1.87	2.35	2.02	0.95	1.4	3.44*	4.49
Sexual	3.01	2.99	2.71	1.15	1.92	4.14	3.86
Ingest	2.28	2.03	2	0.99	1.36	2.48	9.29
Drives							
Drives	2.08	1.98	1.92	0.93	1.42	3.39***	3.78*
Affiliation	2.24	1.88	1.75	0.85	1.29	2.68	3.06
Achievement	2.13	1.95	2.24	0.92	1.4	3.8*	3.75
Power	1.98	2.25	2.08	0.94	1.47	3.25	3.78
Reward	2.05	1.97	1.85	0.98	1.42	3.3	4.15
Risk	1.78	1.46	1.97	0.98	1.57	3.99	4.15
Relativity							
Relativity	1.76	1.91	2.02*	0.95	1.48**	3.6***	4.01**
Motion	1.91	2.06	2.03	0.98	1.61	3.6*	3.69
Space	1.79	1.79	2.02	0.95	1.46	3.51***	3.88*
Time	1.67	1.95	2.02	0.94	1.47	3.64***	4.33*
Time orientations							
Past focus	1.53	2.02	1.84	0.88	1.37	3.54**	3.41
Present focus	2.04	1.82	2*	0.97	1.53**	3.75***	4.19**
Future focus	2.11	2.11	2.64	1.12	1.73	4.72*	4.56
Personal concerns							
Work	2.7	1.81	2	0.87	1.34	3.73	3.14
Leisure	3.1	2.03	1.76	1.03	1.42	3.07	6.32
Home	1.66	1.54	1.86	0.84	1.37	3.42	2.71
Money	1.89	2.59	1.66	0.76	1.22	3.31	2.42
Religion	1.72	1.46	1.86	1.18	1.54	3.81	5.26
Death	1.74	1.88	2.07	1.07	1.5	3.97*	4.86

Statistical differences in the frequency of word use were primarily observed after the pandemic outbreak (Figure 2). During T3 (March 2020-September 2020), the high-risk group showed significantly more frequent use of words related to personal pronouns (RR 2.09, SE 0.73, *P*=.03), affective processes (RR 2.11, SE 0.73, *P*=.03), relativity (RR 2.02, SE 0.55, *P*=.01), and present focus (RR 2.00, SE 0.54, *P*=.01) compared with the low-risk group. During T5 (March 2021-September 2021), posts in the high-risk group also exhibited higher frequencies of words related to personal pronouns (RR 1.51, SE 0.26, *P*=.02), first-person singular (RR 1.54, SE 0.31, *P*=.03), affective processes (RR 1.60, SE 0.27, *P*=.005), negative emotions (RR 1.66, SE 0.35, *P*=.02), relativity (RR 1.48, SE 0.20, *P*=.005), and present focus (RR 1.53, SE 0.20, *P*=.001). During T6 (September 2021-March 2022), 21 types of words across categories such as personal pronouns; affective, social, cognitive, and biological processes; drives; relativity; time orientations; and personal concerns showed higher frequencies in the high-risk group compared with the low-risk group. Meanwhile, in T7 (March 2022-September 2022), the high-risk group exhibited higher frequencies of words related to personal pronouns (RR 3.61, SE 2.08, *P*=.03), first-person singular (RR 3.75, SE 2.38, *P*=.04), affective processes (RR 3.98, SE 2.22, *P*=.01), negative emotions (RR 4.3, SE 3.02, *P*=.04), drives (RR 3.78, SE 2.35, *P*=.03), relativity (RR 4.01, SE 1.75, *P*=.001), space (RR 3.88, SE 2.66, *P*=.04), time (RR 4.33, SE 2.76, *P*=.02), and present focus (RR 4.19, SE 1.8, *P*=.001).

To investigate the linguistic markers that could distinguish group membership, we identified 21 word types that significantly (*P*≤.05) differed between the 2 groups during the last 3 periods (T5, T6, and T7). These word types include personal pronouns, first-person singular, affective processes, positive emotions,

XSL•FO
RenderX

negative emotions, anger, sadness, social processes, cognitive processes, biological processes, health, drives, achievement, relativity, motion, space, time, past focus, present focus, future focus, and death. The word types with the most observed differences were selected as potential linguistic markers for further examination. Then, we omitted 6 word types (ie, personal pronouns, affective processes, negative emotions, biological processes, drives, and relativity) due to their hierarchical relationship with their subcategory words to avoid duplication. To better fit the multivariate logistic regression, we calculated a binary measure for each of the remaining 15 potential markers, indicating no use (0) or use (1) of the word. We calculated the average frequency of each word across T5, T6, and T7, and then dichotomized these averages. Averaged values of 0 were retained as 0, indicating no use of the word during T5, T6, and T7. For averaged values greater than 0, we recoded the value as 1, indicating that the word was used at least once during T5, T6, and T7, regardless of the actual frequency. To mitigate collinearity among the 15 words, we used lasso regression for variable selection. Ultimately, we omitted 3 word types—specifically, the first-person singular, space, and time—leaving us with 12 linguistic features: positive emotions, anger, sadness, social processes, cognitive processes, health, achievement, motion, past focus, present focus, future focus, and death.

Table 3 presents the results of the multivariate logistic regression, incorporating potential linguistic markers into the 2-group GBTM. This analysis models the odds of being in the high-risk group based on the usage of potential linguistic features, with no use of the word serving as the reference. The final model indicated that 9 linguistic features emerged as significant ($P \leq .05$) markers distinguishing the 2 groups. Notably, using words related to cognitive processes and present focus during the later COVID-19 periods had lower odds of being in the high-risk group compared with not using these words ($OR_{cognitive\ processes}$ 0.06, SE 0.85, $P<.001$; $OR_{present\ focus}$ 0.03, SE 0.85, $P<.001$). This indicates that the use of these words was associated with being in the low-risk group. Contrastingly, the odds of being in the high-risk group were substantially higher when using words related to anger, sadness, health, achievement, motion, future focus, and death, compared with not using these words (OR_{anger} 3.23, SE 0.29, $P<.001$; $OR_{sadness}$ 3.23, SE 0.25, $P<.001$; OR_{health} 2.56, SE 0.33, $P=.005$; $OR_{achievement}$ 1.67, SE 0.26, $P=.049$; OR_{motion} 4.17, SE 0.37, $P<.001$; $OR_{future\ focus}$ 2.86, SE 0.3, $P<.001$; OR_{death} 4.35, SE 0.26, $P<.001$). The results illustrated that these 7 words, used 1 year after the pandemic outbreak, were linguistic markers for being in the high-risk group. Multimedia Appendix 1 provides examples of posts that high-risk users published 1 year after the pandemic outbreak.

Table 3. Odds ratios from multivariate logistic regression to identify linguistic markers that discriminate r/SuicideWatch users in the high- and low-risk groups.

Linguistic markers	Odds ratio	SE	*t* value (*df*)	*P* value
Constant	14.83	0.08	32.37 (6162)	<.001
Positive emotions	1.79	0.44	1.34 (6162)	.18
Anger	3.23	0.29	4.12 (6162)	<.001
Sad	3.23	0.25	4.75 (6162)	<.001
Social processes	0.56	0.53	–1.11 (6162)	.27
Cognitive processes	0.06	0.85	–3.38 (6162)	<.001
Health	2.56	0.33	2.84 (6162)	.005
Achievement	1.67	0.26	1.97 (6162)	.049
Motion	4.17	0.37	3.81 (6162)	<.001
Past focus	1.82	0.40	1.48 (6162)	.14
Present focus	0.03	0.85	–4.27 (6162)	<.001
Future focus	2.86	0.3	3.52 (6162)	<.001
Death	4.35	0.26	5.61 (6162)	<.001

Discussion

Principal Findings

To the best of our knowledge, this work is the first to address heterogeneity in suicide risk among social media users by incorporating the temporal characteristics of suicide. Based on the 2 identified trajectories of post volume throughout the COVID-19 pandemic, users on the r/SuicideWatch subreddit were divided into the “high risk of suicide” group (744/6163, 12.07%), characterized by a sharp and lasting increase in post volume, and the “low risk of suicide” group (5419/6163, 87.93%), characterized by a consistently low and mild increase in post volume during the pandemic. In terms of linguistic features, the 2 groups exhibited distinct frequency trends throughout the pandemic. The high-risk group demonstrated longer-lasting increases and lagged peaks in most linguistic frequencies. Contrarily, the low-risk group displayed different trends. Notably, the use of words related to anger, sadness, health, achievement, motion, future focus, and death 1 year after the pandemic outbreak emerged as markers for membership in the high-risk group. Conversely, words associated with cognitive

XSL•FO
RenderX

processes and present focus were identified as linguistic markers for the low-risk group.

Across the pre- and peripandemic periods, this study identified 2 distinct patterns of change in suicide risk among r/SuicideWatch users based on trajectory modeling of their post volume. These findings underscore the heterogeneity in suicide risk among r/SuicideWatch users from a longitudinal perspective during the pandemic. Users' participation in subreddits, including posting frequency, commenting habits, and emotional expression, was influenced by significant pandemic events [38], particularly its progression in Western countries such as the US, the UK, Canada, Australia, and Germany, where a majority of Redditors originate [70]. Both groups of users exhibited immediate increases in post volume following the onset of the COVID-19 pandemic in March 2020. However, post volume returned to prepandemic levels in later stages, around September 2021, as many Western countries began to resume normalcy [71]. According to the fluid vulnerability theory [40], an environmental stressor can trigger a suicidal response within individuals who have predispositions to such reactions. While the half-year intervals may not fully capture users' detailed responses to the pandemic or fluctuations in their suicidal episodes, the heightened posting activity observed in both groups following the pandemic's onset suggests an overall increase in their suicide risk. Therefore, the ongoing pandemic and its repercussions may serve as a persistent environmental stressor for users. Importantly, the high-risk group exhibited significantly greater increases in post volume during the pandemic (T3-T5: March 2020-September 2021) compared with the low-risk group. This suggests that the onset of suicidal episodes was more pronounced among the high-risk group than the low-risk group. The finding of users' heterogeneity in suicide risk can be explained by the interaction between one's baseline and acute risk of suicide, as proposed by the fluid vulnerability theory [41]. Individuals in the high-risk group may have a higher baseline risk due to underlying vulnerabilities, making their suicidal tendencies more readily activated compared with those in the low-risk group, who have fewer vulnerabilities and a lower baseline risk. The higher level of predispositions among high-risk users also renders them more vulnerable to the adverse impacts of the pandemic. This vulnerability activates heightened risks in various domains including cognition (eg, hopelessness), emotion (eg, depression), behavior (eg, social withdrawal), and physiology (eg, sleep disturbances), contributing to their increased acute risk. The higher baseline and acute risks motivate high-risk users to express their heightened concerns, seek support, and exchange information online, leading to a significant increase in social media engagement [72]. By contrast, the low-risk group, which showed consistently low and mild increases in post volume, likely represents the majority less predisposed to suicide risk, indicating greater resilience to the pandemic. Therefore, they may perceive the pandemic as less threatening and experience fewer burdens related to cognitive, emotional, behavioral, and physiological factors. With fewer concerns to share, they exhibited only a mild and minimal increase in post volume. Our findings underscore the heterogeneity in patterns of suicide risk change during the pandemic within this population, highlighting the importance of considering users' individual differences and the temporal dynamics of suicide in future studies using social media data.

Additionally, this study observed differences in the trends of linguistic features between the high- and low-risk groups. During the first year of the pandemic (T3-T4: March 2020-March 2021), both groups significantly increased their use of words related to personal pronouns, positive and negative emotions, social processes, drives, relativity, and time orientations compared with other word categories, indicating broader topics of interest during this period [73]. However, the increase in linguistic frequency continued at a slower pace in the high-risk group before reaching a peak and returning to its original volume (T5-T7: March 2021-September 2022), whereas the low-risk group experienced an early, mild peak followed by an immediate decrease. This divergent trend highlights that most statistical differences in linguistic frequency between the 2 groups became evident 1 year after the outbreak of the pandemic (T5-T7: March 2021-September 2022), indicating that the impact of the pandemic on the high-risk group was more prolonged and delayed compared with the low-risk group. This finding not only underscores the heterogeneity between the 2 groups but also highlights that high-risk users have experienced prolonged stress and heightened sensitivity during the pandemic.

To better identify users at high risk of suicide and understand their underlying concerns, we examined linguistic markers based on several features that showed between-group differences 1 year into the pandemic. Specifically, words related to anger, sadness, health, achievement, motion, future focus, and death were identified as linguistic markers for the high-risk group, which partially aligns with previous findings [36,49]. We delved deeper into the post content of high-risk users to grasp the context in which these linguistic markers were used. Words related to anger and sadness were used by high-risk users to express agitation and hopelessness concerning the overwhelming impact of the pandemic, emotions strongly linked with an increased risk of suicidal thoughts and behaviors [74-76]. When discussing health and motion, high-risk users conveyed heightened concerns about their physical well-being and limitations in movement due to pandemic-related lockdowns [37]. Additionally, they used achievement-related words to express feelings of failure in meeting their goals and fulfilling their need for social recognition. These users may place high demands on themselves, striving to accomplish difficult tasks and meet high standards, which can increase their vulnerability to depression and suicidal behaviors [17,77]. The widespread economic losses, unemployment, and disruptions in educational settings caused by the pandemic further impeded their ability to achieve success, leading to lowered self-esteem, depressive mood, and heightened suicidal risk [78]. Additionally, we discovered that words related to future focus served as linguistic markers for the high-risk group. While previous studies have noted that suicidal individuals often emphasize present-focused words, reflecting their hopelessness about the future and acute concerns about their current state [49,79], this pattern may differ during the pandemic. High-risk users articulated their apprehensions about an uncertain and uncontrollable future amid the evolving pandemic, as exemplified in the texts (Multimedia Appendix 1). Additionally, the high-risk group

XSL•FO
RenderX

used more words related to death. In addition to referencing suicide or hopelessness, this marker also indicated their perceived threats from virus infections, death cases, or the loss of loved ones during the pandemic [37,80].

Our findings have significant implications for managing suicide issues during future public health crises. By analyzing social media posts, we identified a small percentage of users at high risk of suicide who appear particularly sensitive and vulnerable to pandemic-related events or similar public health crises in the future. Although the majority are at low risk of suicide, these results underscore serious concerns, as high-risk users may be poised to progress to the next stage of suicidal ideation or take action [36]. Therefore, it is crucial to pay particular attention to this subset of users to alleviate their difficulties in such situations. Moreover, the active posting and disclosure by these high-risk users may lead to "suicidal contagion" affecting low-risk users, potentially propagating suicidal tendencies within online communities [81]. Therefore, ongoing surveillance, screening, and timely intervention during public health crises are necessary to prevent this issue. Furthermore, the distinct linguistic patterns observed in the 2 groups in this study can serve as a foundation for understanding the underlying concerns contributing to these users' suicide risk, thereby aiding in the development of targeted interventions. The identified language markers for the high-risk group can also serve as a basis for screening high-risk individuals in future pandemic-like events.

Additionally, this study has several limitations. First, aside from users disclosing their own suicidal issues, r/SuicideWatch includes posts about others' suicide risk, providing assistance to those in distress, and disseminating research messages [24]. Although the percentage of these posts was small in our manual screening of selected posts (18/3372, 0.53%, sampled posts), future studies are advised to mitigate this noise or incorporate users' other online behaviors (eg, commenting frequency and post length) to more accurately assess users' suicide risk. Moreover, a significant portion of users in our data set transitioned from other subreddits to r/SuicideWatch following the onset of the pandemic, starting with 0 post volume in periods preceding their initial posts (eg, 2 prepandemic periods). Future studies could track users' earlier psychosocial characteristics on other subreddits to identify indicators that might foreshadow their shift toward actively discussing suicidal concerns on r/SuicideWatch. Second, we utilized seven 6-month intervals as the time frames for capturing post volume and linguistic frequency, which may have been too lengthy to capture specific fluctuations. Nan et al [82] also utilized 6-month intervals and identified a 2-trajectory model for changes in suicidal ideation throughout the pandemic using scores from multiple-item scales as the trajectory variable. However, using shorter intervals (eg, 2-6 weeks) can reveal more trajectories, as it considers minor but significant differences rather than averaging them in the analysis [83,84]. Given the frequent release of pandemic-related news and information (eg, daily reports), users shared real-time reactions to these updates in their posts, potentially reflecting immediate changes in their suicidal thoughts or behaviors, a nuance that might not have been fully captured in our study [38]. Future studies could benefit from shorter time intervals to capture more nuanced and continuous changes in suicide risk, potentially revealing diverse trajectories of suicidal ideation. Third, due to the anonymity of Reddit data, our access was restricted to users' demographics (eg, country or region, age, and sex). Consequently, these factors could not be included as potential covariates for modeling trajectory groups or for comparing the demographic compositions between high- and low-risk user groups. We also acknowledge the potential confounding impact of varying pandemic waves and government control policies across different countries, which we were unable to explore due to the lack of geographical information from users. Future studies should aim to investigate these factors while maintaining the integrity of data characterized by high self-disclosure and authenticity [14]. Additionally, our analysis focused exclusively on Reddit data from a Western context [70]. Cross-cultural validation using data from other platforms, such as Weibo, will be crucial to enhance the generalizability of findings and consider cultural and national policy influences.

Conclusions

This study used social media posts to demonstrate the heterogeneous patterns of change in suicide risk during the COVID-19 pandemic. A group of Reddit users at high risk of suicide was identified, characterized by a sharp and sustained increase in post volume. These high-risk users exhibited distinct linguistic patterns, particularly in their use of words related to anger, sadness, health, achievement, motion, future focus, and death during the later stages of the pandemic. Our findings underscore the importance of recognizing users' heterogeneity in long-term suicide risk. Real-time surveillance of suicide risk using social media data during future public health crises is essential to provide timely support to individuals potentially at high risk of suicide.

Acknowledgments

The study was sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China (Collaborative Research Fund, Project No. C1031-18G). The sponsors had no further role in study design, in the collection, analysis, and interpretation of data, in the writing of the report, and in the decision to submit the article for publication.

Data Availability

The data set used in this study is available from the corresponding author upon request.

XSL•FO
RenderX

Authors' Contributions

NXY and QL conceptualized the study. YY, JL, and XL further completed the study design. JL and XL contributed to the data collection. YY was responsible for data analysis and interpretation. YY drafted the manuscript. NXY, QL, JL, and XL reviewed and edited the draft. QL and NXY administrated the project and acquired the funding. We thank the 2 anonymous reviewers for their valuable input.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of linguistic markers for users in the high-risk group.
[DOCX File , 19 KB - jmir_v26i1e48907_app1.docx]

Multimedia Appendix 2
Summary distribution of LIWC frequency for r/SuicideWatch users in the high- and low-risk groups during each period (median [Q1, Q3]). LIWC: Linguistic Inquiry and Word Count.
[DOCX File , 34 KB - jmir_v26i1e48907_app2.docx]

Multimedia Appendix 3
Trends of LIWC frequency by category for r/SuicideWatch users in the high- and low-risk groups throughout the COVID-19 pandemic (based on the median). LIWC: Linguistic Inquiry and Word Count.
[DOCX File , 392 KB - jmir_v26i1e48907_app3.docx]

References

1. Riedel-Heller S, Richter D. COVID-19 pandemic and mental health of the general public: is there a tsunami of mental disorders? Psychiatr Prax 2020 Nov 02;47(8):452-456. [doi: 10.1055/a-1290-3469] [Medline: 33137827]
2. Gunnell D, Appleby L, Arensman E, Hawton K, John A, Kapur N, COVID-19 Suicide Prevention Research Collaboration. Suicide risk and prevention during the COVID-19 pandemic. Lancet Psychiatry 2020 Jun;7(6):468-471 [FREE Full text] [doi: 10.1016/S2215-0366(20)30171-1] [Medline: 32330430]
3. Sher L. The impact of the COVID-19 pandemic on suicide rates. QJM 2020 Oct 01;113(10):707-712 [FREE Full text] [doi: 10.1093/qjmed/hcaa202] [Medline: 32539153]
4. Pirkis J, Gunnell D, Shin S, Del Pozo-Banos M, Arya V, Aguilar PA, et al. Suicide numbers during the first 9-15 months of the COVID-19 pandemic compared with pre-existing trends: an interrupted time series analysis in 33 countries. EClinicalMedicine 2022 Sep;51:101573 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101573] [Medline: 35935344]
5. Dubé JP, Smith MM, Sherry SB, Hewitt PL, Stewart SH. Suicide behaviors during the COVID-19 pandemic: a meta-analysis of 54 studies. Psychiatry Res 2021 Jul;301:113998 [FREE Full text] [doi: 10.1016/j.psychres.2021.113998] [Medline: 34022657]
6. Pathirathna ML, Nandasena HMRKG, Atapattu AMMP, Weerasekara I. Impact of the COVID-19 pandemic on suicidal attempts and death rates: a systematic review. BMC Psychiatry 2022 Jul 28;22(1):506 [FREE Full text] [doi: 10.1186/s12888-022-04158-w] [Medline: 35902951]
7. Yan Y, Hou J, Li Q, Yu NX. Suicide before and during the COVID-19 pandemic: a systematic review with meta-analysis. Int J Environ Res Public Health 2023 Feb 14;20(4):3346 [FREE Full text] [doi: 10.3390/ijerph20043346] [Medline: 36834037]
8. John A, Pirkis J, Gunnell D, Appleby L, Morrissey J. Trends in suicide during the covid-19 pandemic. BMJ 2020 Nov 12;371:m4352. [doi: 10.1136/bmj.m4352] [Medline: 33184048]
9. Knipe D, John A, Padmanathan P, Eyles E, Dekel D, Higgins JPT, et al. Suicide and self-harm in low- and middle- income countries during the COVID-19 pandemic: a systematic review. PLOS Glob Public Health 2022;2(6):e0000282 [FREE Full text] [doi: 10.1371/journal.pgph.0000282] [Medline: 36962383]
10. Robinson J, Cox G, Bailey E, Hetrick S, Rodrigues M, Fisher S, et al. Social media and suicide prevention: a systematic review. Early Interv Psychiatry 2016 Apr;10(2):103-121. [doi: 10.1111/eip.12229] [Medline: 25702826]
11. Tadesse MM, Lin H, Xu B, Yang L. Detection of suicide ideation in social media forums using deep learning. Algorithms 2019 Dec 24;13(1):7. [doi: 10.3390/a13010007]
12. Slemon A, McAuliffe C, Goodyear T, McGuinness L, Shaffer E, Jenkins EK. Reddit users' experiences of suicidal thoughts during the COVID-19 pandemic: a qualitative analysis of r/Covid19_support posts. Front Public Health 2021;9:693153 [FREE Full text] [doi: 10.3389/fpubh.2021.693153] [Medline: 34458223]

XSL•FO
RenderX

13. Wong K, Chan CS, Chan M, Wong C, Cheng Q, Xiong C, et al. Who seeks help online? Comparing online and offline help-seeking preferences amongst youths with suicidal ideation. J Affect Disord 2021 Sep 01;292:21-29 [FREE Full text] [doi: 10.1016/j.jad.2021.05.056] [Medline: 34087633]
14. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020 Oct 12;22(10):e22635 [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
15. Bitew S, Bekoulis G, Deleu J, Sterckx L, Zaporojets K, Demeester T. Predicting suicide risk from online postings in Reddit the UGent-IDLab submission to the CLPysch 2019 shared task A. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. New York, NY: Association for Computational Linguistics; 2019 Presented at: The Sixth Workshop on Computational Linguistics and Clinical Psychology; June 6, 2019; Minneapolis, MN p. 158-161 URL: https://aclanthology.org/W19-3019 [doi: 10.18653/v1/w19-3019]
16. Chatterjee M, Samanta P, Kumar P, Sarkar D. Suicide ideation detection using multiple feature analysis from Twitter data. New York, NY: IEEE; 2022 Presented at: 2022 IEEE Delhi Section Conference (DELCON); February 1-13, 2022; New Delhi, India p. 1-6. [doi: 10.1109/delcon54057.2022.9753295]
17. Cheng Q, Li TM, Kwok C, Zhu T, Yip PS. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 2017 Jul 10;19(7):e243 [FREE Full text] [doi: 10.2196/jmir.7276] [Medline: 28694239]
18. Ruiz V, Shi L, Quan W, Ryan N, Biernesser C, Brent D. CLPsych2019 shared task: predicting suicide risk level from reddit posts on multiple forums. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. New York, NY: Association for Computational Linguistics; 2019 Presented at: The Sixth Workshop on Computational Linguistics and Clinical Psychology; June 6, 2019; Minneapolis, MN p. 162-166 URL: https://aclanthology.org/W19-3020 [doi: 10.18653/v1/w19-3020]
19. Shan Y. A promising form of suicide intervention model: distant supervision in social media the application of artificial intelligence in suicide intervention. In: Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences. 2021 Presented at: The 2nd International Symposium on Artificial Intelligence for Medicine Sciences; October 29-31, 2021; Beijing, China. [doi: 10.1145/3500931.3500941]
20. Tan Z, Liu X, Liu X, Cheng Q, Zhu T. Designing microblog direct messages to engage social media users with suicide ideation: interview and survey study on Weibo. J Med Internet Res 2017 Dec 12;19(12):e381 [FREE Full text] [doi: 10.2196/jmir.8729] [Medline: 29233805]
21. Varathan K, Talib N. Suicide detection system based on Twitter. 2014 Presented at: 2014 Science and Information Conference; August 27-29, 2014; London, UK. [doi: 10.1109/sai.2014.6918275]
22. Goel A, Gupta L. Social media in the times of COVID-19. J Clin Rheumatol 2020 Sep;26(6):220-223 [FREE Full text] [doi: 10.1097/RHU.0000000000001508] [Medline: 32852927]
23. Hussain W. Role of social media in COVID-19 pandemic. Int J Front Sci 2020;4(2):59-60 [FREE Full text] [doi: 10.37978/tijfs.v4i2.144]
24. r/SuicideWatch. Reddit. URL: https://www.reddit.com/r/SuicideWatch/ [accessed 2023-03-26]
25. Kusuma K, Larsen M, Quiroz JC, Gillies M, Burnett A, Qian J, et al. The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: a meta-analysis and systematic review. J Psychiatr Res 2022 Nov;155:579-588. [doi: 10.1016/j.jpsychires.2022.09.050] [Medline: 36206602]
26. Sawhney R, Joshi H, Gandhi S, Shah R. Towards ordinal suicide ideation detection on social media. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021 Presented at: ACM International Conference on Web Search and Data Mining; March 8-12, 2021; New York, NY. [doi: 10.1145/3437963.3441805]
27. Zirikly A, Resnik P, Uzuner O, Hollingshead K. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In: Proceedings of The Sixth Workshop on Computational Linguistics and Clinical Psychology. New York, NY: Association for Computational Linguistics; 2019 Presented at: The Sixth Workshop on Computational Linguistics and Clinical Psychology; June 6, 2019; Minneapolis, MN p. 24-33 URL: https://aclanthology.org/W19-3003.pdf [doi: 10.18653/v1/w19-3003]
28. Li Q, Xiong Q, Ji S, Wen J, Gao M, Yu Y, et al. Using fine-tuned conditional probabilities for data transformation of nominal attributes. Pattern Recognition Letters 2019 Dec;128:107-114. [doi: 10.1016/j.patrec.2019.08.024]
29. Zdravevski E, Lameski P, Kulakov A, Kalajdziski S. Transformation of nominal features into numeric in supervised multi-class problems based on the weight of evidence parameter. 2015 Presented at: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS); September 13-16, 2015; Lodz, Poland. [doi: 10.15439/2015f90]
30. Kumar M, Dredze M, Coppersmith G, De CM. Detecting changes in suicide content manifested in social media following celebrity suicides. In: HT '15: Proceedings of the 26th ACM Conference on Hypertext & Social Media. New York, NY: Association for Computing Machinery; 2015 Presented at: HT '15: 26th ACM Conference on Hypertext and Social Media; September 1-4, 2015; Guzelyurt, Northern Cyprus p. 85-94. [doi: 10.1145/2700171.2791026]
31. Monselise M, Yang C. "I'm always in so much pain and no one will understand"-Detecting patterns in suicidal ideation on Reddit. In: WWW '22: Companion Proceedings of the Web Conference 2022. New York, NY: Association for Computing

XSL•FO
RenderX

Machinery; 2022 Presented at: WWW '22: The ACM Web Conference; April 25-29, 2022; Virtual Event/Lyon, France p. 686-691 URL: https://dl.acm.org/doi/10.1145/3487553.3524700 [doi: 10.1145/3487553.3524700]
32. Lee K, Agrawal A, Choudhary A. Real-time disease surveillance using Twitter data: demonstration on flu and cancer. New York, NY: Association for Computing Machinery; 2013 Presented at: KDD' 13: The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 11-14, 2013; Chicago, IL p. 1474-1477. [doi: 10.1145/2487575.2487709]
33. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020 Dec 14;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
34. Dutta R, Gkotsis G, Velupillai S, Bakolis I, Stewart R. Temporal and diurnal variation in social media posts to a suicide support forum. BMC Psychiatry 2021 May 19;21(1):259 [FREE Full text] [doi: 10.1186/s12888-021-03268-1] [Medline: 34011346]
35. Coppersmith G, Ngo K, Leary R, Wood A. Exploratory analysis of social media prior to a suicide attempt. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. 2016 Presented at: The Third Workshop On Computational Linguistics and Clinical Psychology; June 16, 2016; San Diego, CA p. 106-117 URL: https://aclanthology.org/W16-0311 [doi: 10.18653/v1/w16-0311]
36. Feldhege J, Wolf M, Moessner M, Bauer S. Psycholinguistic changes in the communication of adolescent users in a suicidal ideation online community during the COVID-19 pandemic. Eur Child Adolesc Psychiatry 2023 Jun;32(6):975-985 [FREE Full text] [doi: 10.1007/s00787-022-02067-7] [Medline: 36018514]
37. Biester L, Matton K, Rajendran J, Provost EM, Mihalcea R. Understanding the impact of COVID-19 on online mental health forums. ACM Trans Manage Inf Syst 2021 Sep 08;12(4):1-28. [doi: 10.1145/3458770]
38. Zhang S, Liu M, Li Y, Chung JE. Teens' social media engagement during the COVID-19 pandemic: a time series examination of posting and emotion on Reddit. Int J Environ Res Public Health 2021 Sep 25;18(19):10079 [FREE Full text] [doi: 10.3390/ijerph181910079] [Medline: 34639381]
39. Shing H, Nair S, Zirikly A, Friedenberg M, Daumé IH, Resnikxpert, Crowdsourced, et al. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. New York, NY: Association for Computational Linguistics; 2018 Presented at: The Fifth Workshop on Computational Linguistics and Clinical Psychology; June 5, 2018; New Orleans, LA p. 25-36 URL: https://aclanthology.org/W18-0603.pdf [doi: 10.18653/v1/w18-0603]
40. Rudd MD, Mandrusiak M, Joiner TE. The case against no-suicide contracts: the commitment to treatment statement as a practice alternative. J Clin Psychol 2006 Feb;62(2):243-251. [doi: 10.1002/jclp.20227] [Medline: 16342293]
41. Bryan CJ, Butner JE, May AM, Rugo KF, Harris J, Oakey DN, et al. Nonlinear change processes and the emergence of suicidal behavior: a conceptual model based on the fluid vulnerability theory of suicide. New Ideas Psychol 2020 Apr;57:100758 [FREE Full text] [doi: 10.1016/j.newideapsych.2019.100758] [Medline: 32123464]
42. Nguena Nguefack HL, Pagé MG, Katz J, Choinière M, Vanasse A, Dorais M, et al. Trajectory modelling techniques useful to epidemiological research: a comparative narrative review of approaches. Clin Epidemiol 2020;12:1205-1222 [FREE Full text] [doi: 10.2147/CLEP.S265287] [Medline: 33154677]
43. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: our words, our selves. Annu Rev Psychol 2003;54:547-577. [doi: 10.1146/annurev.psych.54.101601.145041] [Medline: 12185209]
44. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018;10:1178222618792860 [FREE Full text] [doi: 10.1177/1178222618792860] [Medline: 30158822]
45. Coppersmith G, Leary R, Whyne E, Wood T. Quantifying suicidal ideation via language usage on social media. In: Joint Statistics Meetings Proceedings, Statistical Computing Section. 2015 Presented at: Joint Statistics Meetings, Statistical Computing Section; August 8-15, 2015; Seattle, WA p. 8-15 URL: https://api.semanticscholar.org/CorpusID:54026822
46. Adil Jaafar E, Abdul-Salam Jasim H. A corpus-based stylistic analysis of online suicide notes retrieved from Reddit. Cogent Arts & Humanities 2022 Mar 06;9(1):2047434. [doi: 10.1080/23311983.2022.2047434]
47. Jaafar EA, Jaafar EA. Studies on linguistic stylistic analysis of suicide notes and suicidal thought posts. IJRSSH 2022 Mar 31;12(01):100-124. [doi: 10.37648/ijrssh.v12i01.006]
48. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. Proc SIGCHI Conf Hum Factor Comput Syst 2016 May;2016:2098-2110 [FREE Full text] [doi: 10.1145/2858036.2858207] [Medline: 29082385]
49. O'Dea B, Larsen ME, Batterham PJ, Calear AL, Christensen H. A linguistic analysis of suicide-related Twitter posts. Crisis 2017 Sep;38(5):319-329. [doi: 10.1027/0227-5910/a000443] [Medline: 28228065]
50. Reddit. Reddit API Access. Reddit. 2023. URL: https://www.reddit.com/wiki/api/ [accessed 2023-03-26]
51. Liu X, Huang J, Yu NX, Li Q, Zhu T. Mediation effect of suicide-related social media use behaviors on the association between suicidal ideation and suicide attempt: cross-sectional questionnaire study. J Med Internet Res 2020 Apr 28;22(4):e14940 [FREE Full text] [doi: 10.2196/14940] [Medline: 32343249]

XSL•FO
RenderX

52. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The Development and Psychometric Properties of LIWC 2015. Austin, TX: University of Texas; 2015. URL: http://hdl.handle.net/2152/31333
53. Chen J, Hsieh G, Mahmud J, Nichols J. Understanding individuals' personal values from social media word use. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing;. New York, NY: Association for Computing Machinery; 2014 Presented at: CSCW'14: Computer Supported Cooperative Work; February 15-19, 2014; Baltimore, MD p. 405-414. [doi: 10.1145/2531602.2531608]
54. Jones BL, Nagin DS. A note on a Stata plugin for estimating group-based trajectory models. Sociological Methods & Research 2013 Sep 30;42(4):608-613. [doi: 10.1177/0049124113503141]
55. Kim SY, Cho S. Developmental trajectories of tobacco use and risk factors from adolescence to emerging young adulthood: a population-based panel study. BMC Public Health 2022 Aug 29;22(1):1636 [FREE Full text] [doi: 10.1186/s12889-022-14070-3] [Medline: 36038859]
56. Baumgartner SE, Leydesdorff L. Group‐based trajectory modeling (GBTM) of citations in scholarly literature: dynamic qualities of “transient” and “sticky knowledge claims”. Asso for Info Science & Tech 2013 Nov 20;65(4):797-811. [doi: 10.1002/asi.23009]
57. Frankfurt S, Frazier P, Syed M, Jung KR. Using group-based trajectory and growth mixture modeling to identify classes of change trajectories. The Counseling Psychologist 2016 Aug 05;44(5):622-660. [doi: 10.1177/0011000016658097]
58. Nagin D. Group-Based Modeling of Development. Harvard, MA: Harvard University Press; 2005.
59. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974 Dec;19(6):716-723. [doi: 10.1109/TAC.1974.1100705]
60. Schwarz G. Estimating the dimension of a model. Ann Statist 1978 Mar 1;6(2):461-464. [doi: 10.1214/aos/1176344136]
61. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods 2012 Jun;17(2):228-243 [FREE Full text] [doi: 10.1037/a0027127] [Medline: 22309957]
62. Cavanaugh JE, Neath AA. The Akaike information criterion: background, derivation, properties, application, interpretation, and refinements. WIREs Computational Stats 2019 Mar 14;11(3):e1460. [doi: 10.1002/wics.1460]
63. Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. WIREs Computational Stats 2011 Dec 14;4(2):199-203. [doi: 10.1002/wics.199]
64. Mohammed E, Naugler C, Far B. Emerging business intelligence framework for a clinical laboratory through big data analytics. In: Tran QN, Arabnia H, Mohammed EA, editors. Emerging Trends in Computational Biology, Bioinformatics and System Biology (1st edition). New York, NY: Elsevier/Morgan Kaufmann; 2015:577-602.
65. Chen B, Gong W, Lai AY, Sit SM, Ho S, Wang M, et al. Patterns of perceived harms and benefits of the COVID-19 outbreak in Hong Kong adults: a latent profile analysis. Int J Environ Res Public Health 2022 Apr 05;19(7):4352 [FREE Full text] [doi: 10.3390/ijerph19074352] [Medline: 35410033]
66. McElduff F, Cortina-Borja M, Chan S, Wade A. When t-tests or Wilcoxon-Mann-Whitney tests won't do. Adv Physiol Educ 2010 Sep;34(3):128-133 [FREE Full text] [doi: 10.1152/advan.00017.2010] [Medline: 20826766]
67. StataCorp. Stata LASSO reference manual release 17. StataCorp. 2021. URL: https://www.stata.com/manuals/lasso.pdf [accessed 2023-03-26]
68. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. JAMA Intern Med 2020 Aug 01;180(8):1081-1089 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2033] [Medline: 32396163]
69. Bruckman A. Studying the amateur artist: a perspective on disguising data collected in human subjects research on the internet. Ethics Inf Technol 2002;4:217-231 [FREE Full text] [doi: 10.1023/A:1021316409277]
70. Statista. Regional distribution of desktop traffic to Reddit. Statista. 2022. URL: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/ [accessed 2023-03-26]
71. Centers for Disease Control and Prevention (CDC). CDC Museum COVID-19 Timeline. CDC. 2023. URL: https://www.cdc.gov/museum/timeline/covid19.html [accessed 2023-03-26]
72. Oh S, Syn SY. Motivations for sharing information and social support in social media: a comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr. Asso for Info Science & Tech 2015 Apr 21;66(10):2045-2060. [doi: 10.1002/asi.23320]
73. Park C, Shim M, Eo S, Lee S, Seo J, Moon H, et al. Empirical analysis of parallel corpora and in-depth analysis using LIWC. Applied Sciences 2022 May 30;12(11):5545 [FREE Full text] [doi: 10.3390/app12115545]
74. Hawkins KA, Hames JL, Ribeiro JD, Silva C, Joiner TE, Cougle JR. An examination of the relationship between anger and suicide risk through the lens of the interpersonal theory of suicide. J Psychiatr Res 2014 Mar;50:59-65. [doi: 10.1016/j.jpsychires.2013.12.005] [Medline: 24388767]
75. Lightman E, McCarthy P, Dufty D, McNamara D. Using computational text analysis tools to compare the lyrics of suicidal and non-suicidal songwriters. In: Proceedings of the Annual Meeting of the Cognitive Science Society. 2007 Presented at: Annual Meeting of the Cognitive Science Society; August 1-4, 2007; Nashville, TN p. 1-4 URL: https://escholarship.org/uc/item/0dh4553j

XSL•FO
RenderX

76. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: a content analysis. Biomed Inform Insights 2010 Aug 04;2010(3):19-28 [FREE Full text] [doi: 10.4137/bii.s4706] [Medline: 21643548]
77. Farrell DM. Suicide among gifted students. Roeper Review 1989 Mar;11(3):134-139. [doi: 10.1080/02783198909553189]
78. Hull-Blanks EE, Kerr BA, Robinson Kurpius SE. Risk factors of suicidal ideations and attempts in talented, at-risk girls. Suicide Life Threat Behav 2004;34(3):267-276. [doi: 10.1521/suli.34.3.267.42782] [Medline: 15385181]
79. Ware ZR. Adolescent Suicide Through the Lens of 13 Reasons Why: A Corpus Based Study. Doctor of Philosophy Dissertation. Oregan State University. 2022. URL: https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/41687r69m [accessed 2024-07-15]
80. Ornell F, Schuch JB, Sordi AO, Kessler FHP. "Pandemic fear" and COVID-19: mental health burden and strategies. Braz J Psychiatry 2020;42(3):232-235 [FREE Full text] [doi: 10.1590/1516-4446-2020-0008] [Medline: 32267343]
81. Mueller AS, Abrutyn S. Suicidal disclosures among friends: using social network data to understand suicide contagion. J Health Soc Behav 2015 Mar;56(1):131-148 [FREE Full text] [doi: 10.1177/0022146514568793] [Medline: 25722129]
82. Nan J, Salina N, Chong ST, Jiang H. Trajectory of suicidal ideation among medical students during the COVID-19 pandemic: the role of childhood trauma. BMC Psychiatry 2023 Feb 06;23(1):90 [FREE Full text] [doi: 10.1186/s12888-023-04582-6] [Medline: 36747156]
83. Huang S, Wang D, Zhao J, Chen H, Ma Z, Pan Y, et al. Changes in suicidal ideation and related influential factors in college students during the COVID-19 lockdown in China. Psychiatry Res 2022 Aug;314:114653 [FREE Full text] [doi: 10.1016/j.psychres.2022.114653] [Medline: 35671561]
84. Madsen T, van Spijker B, Karstoft K, Nordentoft M, Kerkhof AJ. Trajectories of suicidal ideation in people seeking web-based help for suicidality: secondary analysis of a Dutch randomized controlled trial. J Med Internet Res 2016 Jun 30;18(6):e178 [FREE Full text] [doi: 10.2196/jmir.5904] [Medline: 27363482]

Abbreviations

AIC: Akaike information criterion
APP: average posterior probability
BIC: Bayesian information criterion
CNORM: censored normal
GBTM: group-based trajectory modeling
LIWC: Linguistic Inquiry and Word Count
OR: odds ratio
ZIP: zero-inflated Poisson

Edited by A Mavragani; submitted 11.05.23; peer-reviewed by E Jaafar, S Li; comments to author 15.03.24; revised version received 05.04.24; accepted 18.04.24; published 08.08.24.

Please cite as:
Yan Y, Li J, Liu X, Li Q, Yu NX
Identifying Reddit Users at a High Risk of Suicide and Their Linguistic Features During the COVID-19 Pandemic: Growth-Based Trajectory Model
J Med Internet Res 2024;26:e48907
URL: https://www.jmir.org/2024/1/e48907
doi:10.2196/48907
PMID:39115925

©Yifei Yan, Jun Li, Xingyun Liu, Qing Li, Nancy Xiaonan Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Messenger Use and Video Calls as Correlates of Depressive and Anxiety Symptoms: Results From the Corona Health App Study of German Adults During the COVID-19 Pandemic

Johanna-Sophie Edler[1]; Yannik Terhorst[2,3]; Rüdiger Pryss[4]; Harald Baumeister[2]; Caroline Cohrdes[1]

[1]Mental Health Research Unit, Department of Epidemiology and Health Monitoring, Robert Koch Institute, Berlin, Germany
[2]Department of Clinical Psychology and Psychotherapy, Institute of Psychology and Education, Ulm University, Ulm, Germany
[3]Department of Psychology, Ludwig Maximilian University of Munich (LMU), Munich, Germany
[4]Institute of Clinical Epidemiology and Biometry, Würzburg University, Würzburg, Germany

Corresponding Author:
Caroline Cohrdes
Mental Health Research Unit
Department of Epidemiology and Health Monitoring
Robert Koch Institute
PO Box 650261
Berlin, 12101
Germany
Phone: 49 30187542692
Email: cohrdesc@rki.de

Abstract

Background: Specialized studies have shown that smartphone-based social interaction data are predictors of depressive and anxiety symptoms. Moreover, at times during the COVID-19 pandemic, social interaction took place primarily remotely. To appropriately test these objective data for their added value for epidemiological research during the pandemic, it is necessary to include established predictors.

Objective: Using a comprehensive model, we investigated the extent to which smartphone-based social interaction data contribute to the prediction of depressive and anxiety symptoms, while also taking into account well-established predictors and relevant pandemic-specific factors.

Methods: We developed the Corona Health App and obtained participation from 490 Android smartphone users who agreed to allow us to collect smartphone-based social interaction data between July 2020 and February 2021. Using a cross-sectional design, we automatically collected data concerning average app use in terms of the categories *video calls and telephony*, *messenger use*, *social media use*, and *SMS text messaging use*, as well as pandemic-specific predictors and sociodemographic covariates. We statistically predicted depressive and anxiety symptoms using elastic net regression. To exclude overfitting, we used 10-fold cross-validation.

Results: The amount of variance explained (R^2) was 0.61 for the prediction of depressive symptoms and 0.57 for the prediction of anxiety symptoms. Of the smartphone-based social interaction data included, only messenger use proved to be a significant negative predictor of depressive and anxiety symptoms. Video calls were negative predictors only for depressive symptoms, and SMS text messaging use was a negative predictor only for anxiety symptoms.

Conclusions: The results show the relevance of smartphone-based social interaction data in predicting depressive and anxiety symptoms. However, even taken together in the context of a comprehensive model with well-established predictors, the data only add a small amount of value.

(J Med Internet Res 2024;26:e45530) doi:10.2196/45530

KEYWORDS

passive data; depression; anxiety; predicting mental health; mobile phone

XSL•FO
RenderX

Introduction

Background

Depression and anxiety are common disorders, with depression having a prevalence of 15.7% in Germany in 2017 before the COVID-19 pandemic [1,2]. In 2023, after the pandemic, approximately 19% of the adult population in Germany showed elevated rates of depressive symptoms; in addition, the burden of anxiety symptoms was notable for 12% to 15% of the German population in 2022 [3]. People with major depression show leading symptoms such as depressed mood, loss of interest, or lack of drive, whereas persistent excessive worry and restlessness concerning everyday issues are leading symptoms in people with general anxiety disorder. Both depression and anxiety often incur high individual and societal costs [4]. Accurate and real-time measurement of depressive and anxiety symptoms allows for the identification of current risk factors (eg, pandemic-related factors) [5] as well as the design of preventive and treatment measures. Through digitalization and the spread of smartphones, new possibilities have arisen for flexible, cost-effective, and high-quality data collection on symptoms of mental disorders such as depression and anxiety [6,7]; for example, smartphone data (use data and survey data) can be collected in real time using ecological momentary assessment methodology [8]. Technological advances have enabled passive collection of smartphone use data (without user intervention), reducing interviewer and recall bias. In addition, it is useful to record depressive symptoms via behavioral measurement using objective digital data. With behavioral measurement (such as smartphone use data), it is not necessary for users to be aware of the changes in their behavior. Persons with depressive or anxiety symptoms may only become aware of a behavior change when it reaches a certain level, whereas passive behavior measurement captures each slow change over time, regardless of user awareness; for example, in the context of depressive symptoms, possible behavioral changes such as reduced activity or social withdrawal could be an indicator of a change toward depressive symptoms. Previous studies have shown that depressive symptoms [9-14] and anxiety symptoms [10,11,15] correlate with various smartphone-based social interaction data, such as the duration of phone calls, Facebook use, the number of SMS text messages sent, or total smartphone use time. Furthermore, the high proportion of interpersonal communication that is conducted via mobile devices allows for the measurement of an indicator of social interaction. Specifically, the pandemic showed that perceived social connectedness improved when smartphone use increased, resulting in improved well-being [16]. Social interaction, as indicated by the use duration or frequency of various smartphone social media apps, has been identified as a significant predictor of depressive and anxiety symptoms [17,18]. While time spent on Facebook positively correlates with depressive symptoms [18], social smartphone use is a negative predictor [19]. Excessive smartphone use in general is positively associated with depressive symptoms [10]. Moreover, during the COVID-19 pandemic, daily face-to-face or video calls correlated with decreased depressive symptoms [20]. Anxiety symptoms are only positively related to consumption-related smartphone use and not to social interaction–related smartphone use [11].

Overall, studies analyzing smartphone data have mainly investigated isolated predictors and combinations of only a few predictors [21,22]. The question remains concerning the extent to which smartphone-based social interaction data will continue to a comprehensive model for the prediction of depressive and anxiety symptoms. Answering this question is of great importance for achieving an up-to-date, ecologically valid, and economic assessment of high-quality epidemiological data.

Mental Health During the COVID-19 Pandemic

During the COVID-19 pandemic and the associated restrictions on social life, both the importance of digital social interaction and concerns about depressive and anxiety symptoms increased as the situation worsened [16,23]. Due to the dramatic increase in COVID-19 infections in Germany in March 2020, the German government implemented pandemic action plans to limit the spread of SARS-CoV-2, including quarantine measures, reduction of social contact, and restrictions on mobility. Over the following 18 months, social distancing measures were gradually eased, social contacts were restored, and restrictions on mobility were lifted, culminating in the reopening of shops, playgrounds, and schools in summer 2021 due to the decreased number of COVID-19 infections. In Germany, as in other countries, this period represented a balancing act between universal protection against a growing number of new COVID-19 infections and the preservation of individual psychological and physical health in the face of severe restrictions in daily life [23].

We mapped these pandemic-related changes in daily life in terms of internal and external aspects. Internal aspects refer to perceptions of the pandemic situation and resulting emotions such as fears and worries, which could be stressors. External aspects include changes in daily life due to pandemic control measures, which can also act as potential stressors.

Specifically, internal aspects within the scope of this study include the expected stigmatization of individuals with COVID-19 infection [24] and various related concerns; for example, previous research has revealed increased concerns about not receiving adequate medical care in case of COVID-19 infection due to lack of medical capacity [25] as well as concerns about infecting others with COVID-19 [24,25]. Pandemic-related social distancing also seemed to have altered social interaction, as suggested by reports of increased loneliness [26] and negative family climate [27,28]. Correspondingly, there is evidence showing an increase in psychosocial stress [29] and domestic violence during the pandemic [30].

Regarding external aspects, recent studies have discussed pandemic-related working conditions (eg, short-term work, reduced performance capacity due to home office work and home schooling of children, and increased workloads in the health sector) as stressful changes in daily routines [31-33].

In addition to these internal and external pandemic-specific changes, there are individual differences (reliable predictors existing before the pandemic) that may also be related to mental well-being during the pandemic and that could be considered

XSL•FO
RenderX

stressors; for example, general health status should be considered a possible pandemic-related stress factor because this factor poses a potentially increased risk of severe disorder progression in the event of COVID-19 infection [34]. The same point applies to chronic conditions [30]. In addition, due to the pandemic, the availability of support services for treating mental disorders was temporarily limited, which placed a particular burden on people with preexisting diagnoses of mental disorders [35]. Finally, the risk of contracting COVID-19 infection should also be considered a potential stressor. The general population had to cope with this large number of possible stressors related to the COVID-19 pandemic and social restriction measures; therefore, individual coping strategies became increasingly important. In this context, early investigations in Germany showed an increase in adverse coping techniques, such as alcohol consumption [36]. Likewise, the restriction of sporting activities (in gyms or clubs) can be seen as a potential loss of a resource and method for adaptive coping [37].

According to the vulnerability-stress model [38], an accumulation of stressors is believed to increase the risk of depressive and anxiety symptoms. Quarantine and contact restriction measures result in a reduction in personal encounters, possible additional stress factors at work (eg, job loss and home schooling of children), and restrictions concerning personal lifestyle (eg, due to the closure of gyms and cinemas). Accordingly, the pandemic might have been a high-risk situation for the development or intensification of depressive and anxiety symptoms with the accumulation of multiple stressors in an individual [30,39]. In view of these findings, investigating depressive and anxiety symptoms during the COVID-19 pandemic may add new insights into the validity and predictive value of smartphone-based social interaction data. To date, the relative importance of such predictors in the context of a comprehensive model of established public health indicators has not yet been fully explained.

Objectives of This Study

This study aimed to validate the predictive value of smartphone-based social interaction data in relation to depressive and anxiety symptoms via a comprehensive, data-driven approach, including well-established predictors (eg, chronic conditions and partnership status) from survey data collected during the COVID-19 pandemic. In many existing studies on the topic, smartphone data have been considered in isolation [21,22]. However, to evaluate these objective (smartphone-based social interaction) data as predictors, we opted for a data-based approach that integrates all reliable predictors into a holistic model and automatically identifies the optimal set of predictors. In this way, we were able to test the extent to which various smartphone-based social interaction parameters (eg, video calls and social media) contribute to the prediction of depressive and anxiety symptoms alongside, or instead of, empirically well-established predictors. Due to the ongoing COVID-19 pandemic during data collection, we also considered pandemic-specific factors (eg, COVID-19 infection and short-term work situations).

Methods

Sample and Procedure

The study is part of the Corona Health App project, a collaboration between the Robert Koch Institute, the University of Würzburg, the University of Ulm, and the University of Regensburg in Germany. The project includes an extensive cross-sectional app-based baseline survey, followed by a reduced weekly ongoing longitudinal survey for adults. The collection of both smartphone data and survey data was carried out via the Corona Health App, which is why it was necessary for participants to download and install the app from the Google Play Store or the Apple App Store to participate in the study. Recruitment was carried out through press releases by the participating institutions, media reports, and social networks. Before they took part in the study, participants were provided comprehensive information about the collection of select passive smartphone-based social interaction data and anonymized storage of the data; in addition, they were asked to provide informed consent. We collected passive smartphone-based social interaction data for 1 week before administering the questionnaire. Duration of use was recorded as the time during which the app was running in the foreground.

The cross-sectional analyses included 490 German residents who completed the baseline questionnaire between July 2020 and February 2021 and consented to the collection of smartphone-based social interaction data. During this period, Germany experienced a summer with very low COVID-19 infection numbers and eased restrictions on public life. In the autumn, a second wave of infections started, which worsened as autumn passed into winter and led to lockdown measures starting in mid-December. These measures continued until the end of the survey period for this study. Participants ranged in age from 18 to 78 (mean 42.46, SD 13.33) years, and the gender ratio was relatively balanced, with 52.9% (259/490) of the participants being women, 45.9% (225/490) being men, and 1.2% (6/490) being transgender individuals. Whereas 7.8% (38/490) of the participants had no school-leaving certificate or had lower than a secondary school certificate with or without training, 64.5% (316/490) had a technical college or university degree. A clinician-based lifetime diagnosis of mental disorder was reported by 43.7% (214/490) of the participants. As the recording of app use is not allowed on Apple devices, our reporting sample was limited to Android smartphone users, who, with a market share of 64.7% in Germany [40], constitute the clear majority of users. From the original sample of 1760 participants, 18 (1.02%) were excluded for failing the plausibility check, which assessed correspondence between similar items, straightlining, intraindividual response variability, and extreme outliers. Of the remaining 1742 participants, 1052 (60.39%) were Android smartphone users. Of these 1052 participants, 490 (46.58%) consented to the collection of data concerning social interaction app use. A detailed description of the sample and descriptive statistics of all variables can be found in Multimedia Appendix 1.

XSL•FO
RenderX

Ethical Considerations

Ethics approval was granted by the ethics committee of the University of Würzburg (130/20-me). Furthermore, the app was developed in accordance with the regulations for medical devices. Participants were given comprehensive information about the type of data collected as well as the data collection, processing, and dissemination procedures and were asked to provide consent. Data storage was anonymized. The study protocol with details on the data collection, processing, and dissemination procedures has been described previously [41]. The study fulfils the criteria of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [42]. Participation was voluntary and not financially compensated.

Measures

Depressive and Anxiety Symptoms

Depressive and anxiety symptoms were measured with the German version of the Patient Health Questionnaire, German version (PHQ-D) [43]. Depressive symptoms (9 items; eg, the stem question "Over the last week, how often have you been bothered by" followed by "little interest or pleasure in doing things?" "depression, melancholy, or hopelessness?" "difficulty falling asleep or staying asleep, or increased sleep?" "tiredness or feeling of having no energy?" "reduced appetite or excessive desire to eat?" "a bad opinion of yourself; feeling of being a failure or having disappointed the family?" "difficulties concentrating on something, eg, when reading the newspaper or watching television?" "thoughts that you would rather be dead or want to harm yourself?" as well as "Were your movements or speech so slowed down that others would notice? Or were you on the contrary 'fidgety' or restless and therefore had a stronger urge to move than usual?") were rated on a 4-point scale: 0=*not at all*, 1=*on single days*, 2=*on more than half of the days*, and 3=*almost every day*. The 7 anxiety symptoms (eg, the stem question "Over the last week, how often have you been bothered by" followed by "not being able to stop or control worrying?" "nervousness, anxiety, or tension?" "excessive worries regarding various matters?" "difficulties to relax?" "restlessness, making it difficult to sit still?" "quick annoyance or irritability?" "a feeling of fear, as if something bad is going to happen?") were answered on a 4-point scale: 0=*I haven't been doing this at all*, 1=*I've been doing this a little bit*, 2=*I've been doing this a medium amount*, and 3=*I've been doing this a lot*. We summarized the answers for each scale as a total score ranging from 0 (minimum) to 27 (maximum) for depressive symptoms and 21 (maximum) for anxiety symptoms [44]. The Cronbach α values for the internal consistency of the scales for depressive and anxiety symptoms were 0.89 and 0.85, respectively.

Smartphone-Based Social Interaction Data

As smartphone-based indicators of social interaction, we included a selection of communication data for apps open in the foreground over the past 7 days (video calls, phone calls, social media, SMS text messaging use, and messenger apps) and total smartphone use duration. Use times were averaged on a weekly basis in minutes. The following selection of communication apps was included: Skype, Skype Messenger, Zoom, Facebook, Facebook Messenger, Instagram, Snapchat, WhatsApp, and Telegram. Skype and Zoom belong to the category *video calls*. Phone calls were recorded by the device-specific phone app and included as a separate category. Facebook, Instagram, and Snapchat were grouped into the category *social media*. *SMS text messaginguse* was also captured via the device-specific app and included as a separate category. Facebook Messenger, WhatsApp, and Telegram were combined into the category *messenger*.

COVID-19–Related Factors

As pandemic-related variables, we made use of an in-house–developed item indicating COVID-19 status and differentiating between the 3 categories of *tested positive for COVID-19*, *tested positive for COVID-19 but already recovered*, and *did not test positive for COVID-19*. Other variables included were stigmatization because of (suspected) COVID-19 infection using a modified version of the inventory of subjective experience of stigmatization [45]. Two items each were collected as indicators of the expectation and actual experience of stigmatization. Expectation of stigmatization experience (eg, "Do you think people will value you less if they know that you are infected with COVID-19 or if you were?") was scored on a 5-point scale ranging from 1=*always* to 5=*never*, and for the actual experience of stigmatization (eg, "Have you ever been teased, bullied, or harassed because you were infected with COVID-19, or have you seen such behavior toward others?"), participants answered on a dichotomous scale (yes or no). We calculated sum scores for the expectation and experience of stigmatization, and the Pearson r values for the interitem correlations were 0.41 and 0.43, respectively. Furthermore, we assessed pandemic-related concerns (concern about inadequate medical treatment in case of COVID-19 infection due to lack of medical capacity and concern about infecting others with COVID-19). The items were rated on a 3-point scale ranging from 0=*not bothered* to 2=*bothered a lot*. As employment status may have changed due to the measures implemented to combat the spread of SARS-CoV-2, participants answered several items that were developed in-house regarding employment status (ie, short-term work, unemployed, currently unable to pursue regular profession due to health protection measures, or on sick leave), wage loss (none, slight, or high), predominant workplace (home office, regular working location, or a combination of both), and employment as a health care professional (yes or no). Nominal categories were converted to dummy variables. We also used items developed in-house to ask about COVID-19 infections among relatives and friends (no, currently ill, or recovered). Those who reported COVID-19 infections in their social environment were asked whether they had lost relatives or friends to COVID-19 (no or yes). Both categorial variables were dummy coded and included in the analyses.

Reliable Predictors of Depressive and Anxiety Symptoms

As indicated by previous research, feelings of loneliness [46-48], psychosocial distress [49], family climate [50], physical abuse [51], general health status or chronic conditions [52], coping [53], alcohol consumption [54], and physical activity [55] are correlated with depressive or anxiety symptoms. We measured

XSL•FO
RenderX

loneliness using the 3-item Socio-Economic Panel loneliness scale [56]. The answers were rated on a 5-point scale ranging from 1=*very often* to 5=*never*. An example item was "How often do you feel left out?" We calculated a sum score, and the Cronbach α value for internal consistency was 0.79. Family climate was assessed using 2 different items developed in-house: the actual climate rated on a 5-point scale ranging from 0=*very bad* to 4=*very good* and the perceived change in family climate since the beginning of the COVID-19 pandemic answered using 1 of 3 options: 0=*yes, it got worse*; 1=*no, it remained unchanged*; and 2=*yes, it improved*. Aspects of psychosocial distress were addressed by the PHQ-D stress module [57], comprising 10 items that were answered on a 3-point scale ranging from 0=*not bothered* to 2=*bothered a lot*. An example item was "In the past week, how much have you been bothered by stress at work outside of the home or at school?" We measured physical abuse using the PHQ-D [57] and an additional item developed in-house. The first item asked about experience of physical abuse in the past week, and the second item asked about experience of abuse approximately 12 months previously (yes or no). General health status was assessed using the first 3 items of the Minimum European Health Module (self-perceived health, chronic conditions, and long-term activity limitation) [58]. Moreover, participants reported a clinician-based lifetime diagnosis of a mental disorder (yes or no). Situational coping strategies (during the COVID-19 pandemic) were measured using the Coping Orientation to Problems Experienced Inventory (Brief-COPE) [59]. This instrument comprises a total of 28 items assessed on a 4-point scale ranging from 1=*I haven't been doing this at all* to 4=*I've been doing this a lot*. As a result of an exploratory factor analysis and a confirmatory factor analysis, the 4 coping factors were formed in accordance with the categories *problem-focused coping* (eg, "I've been taking action to try to make the situation better"), *support-focusedcoping* (eg, "I've been getting help and advice from other people"), *escape-avoidant–focused coping* (eg, "I've been using alcohol or other drugs to make myself feel better"), and *meaning-focusedcoping* (eg, "I've been trying to see things in a different light, to make things seem more positive") [60] following previous studies [60-62]. We summed the respective items and calculated a mean score for the 4 subscales problem-focused coping, support-focused coping, escape-avoidant–focused coping, and meaning-focused coping. The Cronbach α values for the internal consistency of the 4 subscales were 0.77, 0.84, 0.69, and 0.74, respectively. In addition, we explored the frequency of alcohol consumption per week with an item developed in-house that was answered on a 5-point scale ranging from 0=*none* to 4=*six to seven times a week*. Moreover, we included an item developed in-house to indicate levels of physical activity during the past week that was answered on a 5-point scale ranging from 1=*no sporting activity* to 5=*regularly more than 4 hours per week*. We assessed personality with the German version of the Big Five Inventory-10 [63], which was answered on a 5-point scale ranging from 1=*disagree strongly* to 5=*agree strongly*. In line with the Big Five Inventory-10 scoring manual [63], we constructed 5 subscales with 2 items each for the dimensions of openness (eg, active imagination), conscientiousness (eg, doing a thorough job), extraversion (eg, being outgoing and sociable), agreeableness (eg, being generally trusting), and neuroticism (eg, easily becoming nervous). The Pearson *r* values for the interitem correlations of the 5 subscales were 0.31, 0.26, 0.60, 0.07, and 0.41, respectively. The level of education was determined by the general school-leaving certificate and the highest vocational qualification. These 2 areas were combined into 3 levels according to the Comparative Analysis of Social Mobility in Industrial Nations classification [64]. The first level includes the range from no school-leaving certificate to a lower secondary school certificate, the second level includes the range from a secondary school certificate to a general qualification for university entrance or a general or subject-linked higher education entrance qualification, and the third level includes technical school certificates and university degrees. We included each category as a dummy-coded variable. In addition, household size and the number of children were included as sociodemographic variables, each as a metric variable. We asked about marital status and living environment, specifically about the type of housing (house or apartment; with garden, with balcony, or with terrace). Multiple choices were allowed for this variable, and the response categories were integrated into the analyses, each as a separate dummy variable. We also included the participant's age as a continuous predictor and the participant's gender based on the 3 categories *woman*, *man*, and *transgender*. Categorical variables were included as dummy-coded variables in the analyses.

Data Analyses

To analyze the predictive value of smartphone-based social interaction data for depressive and anxiety symptoms, we opted for a data-driven approach, using ridge, least absolute shrinkage and selection operator (LASSO), and elastic net regression [65]. Hyperparameter optimization was implemented using a grid search. These analyses allow us to perform a statistics-based variable selection [66] and answer the question concerning the extent to which each smartphone-based social interaction parameter explains a relevant proportion of the variance or whether the parameter is automatically removed from the model as statistically irrelevant. In addition to the smartphone-based social interaction data, we considered other relevant and well-established predictors of depressive and anxiety symptoms (eg, coping strategies and physical activity) as well as pandemic-related factors (eg, concerns and stigmatization). To avoid overfitting, we split the data set into training and test sets. Within the training set, we performed 10-fold cross-validation with 5-fold repetition [65]. The model was then estimated in the training set, and the outcome variable was estimated (ie, predicted mathematically) based on this model in the test set to validate the model. Here, the term *prediction* refers to a statistical procedure and not a temporal prediction in the future. The model with the best fit in terms of estimating depressive and anxiety symptoms was chosen based on indicators of the amount of measurement error (primarily the root mean square error [RMSE] and secondarily the mean absolute error [MAE]) as well as the amount of variance explained (R^2). The predictive value was indicated for each significant predictor by means of the variable importance.

XSL•FO
RenderX

For the majority of questions, answers were required (forced choice); there were only 3 missing values in the data, all of which were related to items on the PHQ-9 scale. However, a precondition for the use of ridge, LASSO, and elastic net regression is a data set without missing values [67]. For this reason, missing data were imputed by single imputation with 500 iterations using the classification and regression trees method with the R package *mice* [68]. The continuous predictors were *z*-standardized, and the outcome variables remained in the original scaling.

Results

Overview

For the prediction of depressive symptoms, the model with the smallest MAE (3.04) was the LASSO regression (Table 1). The smallest RMSE (3.85) resulted from the elastic net regression, which also had the highest amount of variance explained (R^2=0.61). Accordingly, the elastic net regression proved to be the model with the best fit, reducing the data set from 76 to 28 significant variables.

The model with the best fit for predicting anxiety symptoms was also an elastic net regression (MAE=2.56, RMSE=3.22, R^2=0.57) and contained 30 significant predictors of anxiety symptoms (Table 2).

Table 1. Model fit indices for predicting depressive symptoms resulting from stepwise, ridge, least absolute shrinkage and selection operator (LASSO), and elastic net regression based on 50 replicate samples.

	Values, median (IQR; minimum-maximum)	Values, mean
Mean absolute error		
Linear model	3.299 (3.050-3.519; 2.665-4.222)	3.295
Ridge	3.133 (2.922-3.315; 2.511-3.797)	3.135
LASSO	2.973 (2.845-3.204; 2.617-3.812)	3.055
Elastic net[a]	*2.998 (2.813-3.189; 2.553-3.855)*	*3.045*
Root mean square error		
Linear model	4.088 (3.851-4.407; 3.346-5.145)	4.146
Ridge	3.940 (3.708-4.201; 3.220-4.663)	3.967
LASSO	3.787 (3.590-4.060; 3.245-4.672)	3.856
Elastic net[a]	*3.772 (3.594-4.060; 3.286-4.680)*	*3.849*
R^2		
Linear model	0.558 (0.524-0.636; 0.265-0.713)	0.562
Ridge	0.587 (0.529-0.656; 0.300-0.776)	0.586
LASSO	0.619 (0.553-0.677; 0.265-0.819)	0.612
Elastic net[a]	*0.619 (0.554-0.676; 0.264-0.817)*	*0.612*

[a]The best model fit by parameter is highlighted in italics in each case; overall, elastic net regression has the best model fit due to the largest amount of variance explained (R^2) and the smallest mean absolute error and root mean square error.

XSL•FO
RenderX

Table 2. Model fit indices for predicting anxiety symptoms resulting from stepwise, ridge, least absolute shrinkage and selection operator (LASSO), and elastic net regression based on 50 replicate samples.

	Values, median (IQR; minimum-maximum)	Values, mean
Mean absolute error		
Linear model	2.683 (2.486-2.958; 1.917-3.534)	2.711
Ridge	2.608 (2.409-2.834; 1.866-3.235)	2.614
LASSO	2.605 (2.432-2.759; 1.814-3.120)	2.582
Elastic net[a]	*2.560 (2.402-2.745; 1.749-3.130)*	*2.564*
Root mean square error		
Linear model	3.395 (3.133-3.712; 2.438-4.395)	3.405
Ridge	3.273 (3.002-3.551; 2.442-3.939)	3.277
LASSO	3.319 (2.989-3.443; 2.344-3.873)	3.246
Elastic net[a]	*3.283 (3.004-3.409; 2.309-3.776)*	*3.216*
R^2		
Linear model[a]	*0.522 (0.450-0.610; 0.312-0.767)*	*0.528*
Ridge	0.536 (0.481-0.630; 0.366-0.787)	0.550
LASSO	0.561 (0.491-0.606; 0.359-0.814)	0.563
Elastic net	0.564 (0.498-0.625; 0.388-0.820)	0.569

[a]The best model fit by parameter is highlighted in italics in each case; overall, elastic net regression has the best model fit due to the largest amount of variance explained (R^2) and the smallest mean absolute error and root mean square error.

Smartphone-Based Social Interaction Data

Of the smartphone-based social interaction data included, only average weekly messenger use turned out to be a significant negative predictor of depressive and anxiety symptoms. The average weekly duration of video calls was a significant negative predictor only of depressive symptoms. In the model for the prediction of anxiety symptoms, the weekly average duration of SMS text messaging use remained a relevant negative predictor. The other smartphone-based social interaction parameters (weekly averaged use duration of phone calls and social media, as well as weekly averaged total smartphone use duration) were omitted from the best fit model.

Predictors of Depressive and Anxiety Symptoms

The final models overlap in terms of the following 19 predictors (in order of importance according to the model for depressive symptomatology): escape-avoidant–focused coping, loneliness, good current family climate, poor general health status, a clinician-based lifetime diagnosis of mental disorder, age, neuroticism, conscientiousness, expectation of no stigmatization due to (suspected) COVID-19 infection, agreeableness, meaning-focused coping, a chronic condition, wage loss, inability to pursue current occupation due to health protection measures, being a transgender individual, average weekly messenger use, no relatives or friends lost to COVID-19, household size, and concern about not receiving adequate medical care in case of COVID-19 infection due to lack of medical capacity.

The following predictors were found to be exclusively relevant for the prediction of depressive symptoms (in decreasing order of importance): physical abuse, extraversion, alcohol consumption, job seeking, short-term work (due to the COVID-19 pandemic), being married or in a stable relationship, divorce, and employment as a health care professional. Figure 1 presents the order and directions of unique correlates.

By contrast, the following predictors proved to be exclusively relevant for anxiety symptoms (in decreasing order of importance): concern about infecting others with COVID-19, experience of stigmatization, a living environment with a terrace, being male, being single, having relatives who recovered from COVID-19, being in a registered (same-sex) partnership but living separately, average weekly SMS text messaging use, physical activity, and being widowed. The order and directions of unique correlates are presented in Figure 2.

XSL•FO
RenderX

Figure 1. Total variable importance as an indicator of the contribution to the reduction of the estimation error in the prediction of depressive symptoms.

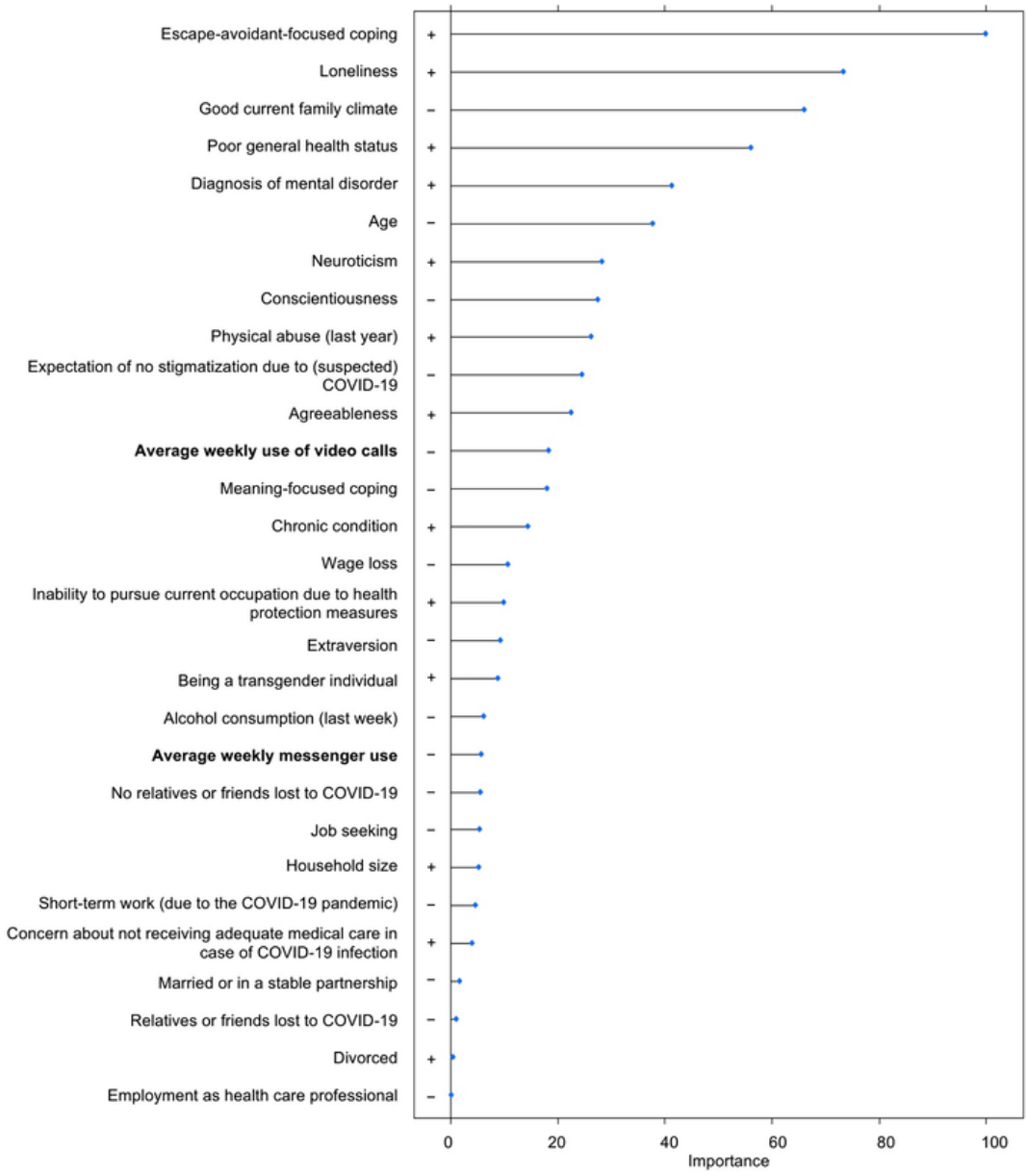

XSL•FO
RenderX

Figure 2. Total variable importance as an indicator of the contribution to the reduction of the estimation error in the prediction of anxiety symptoms.

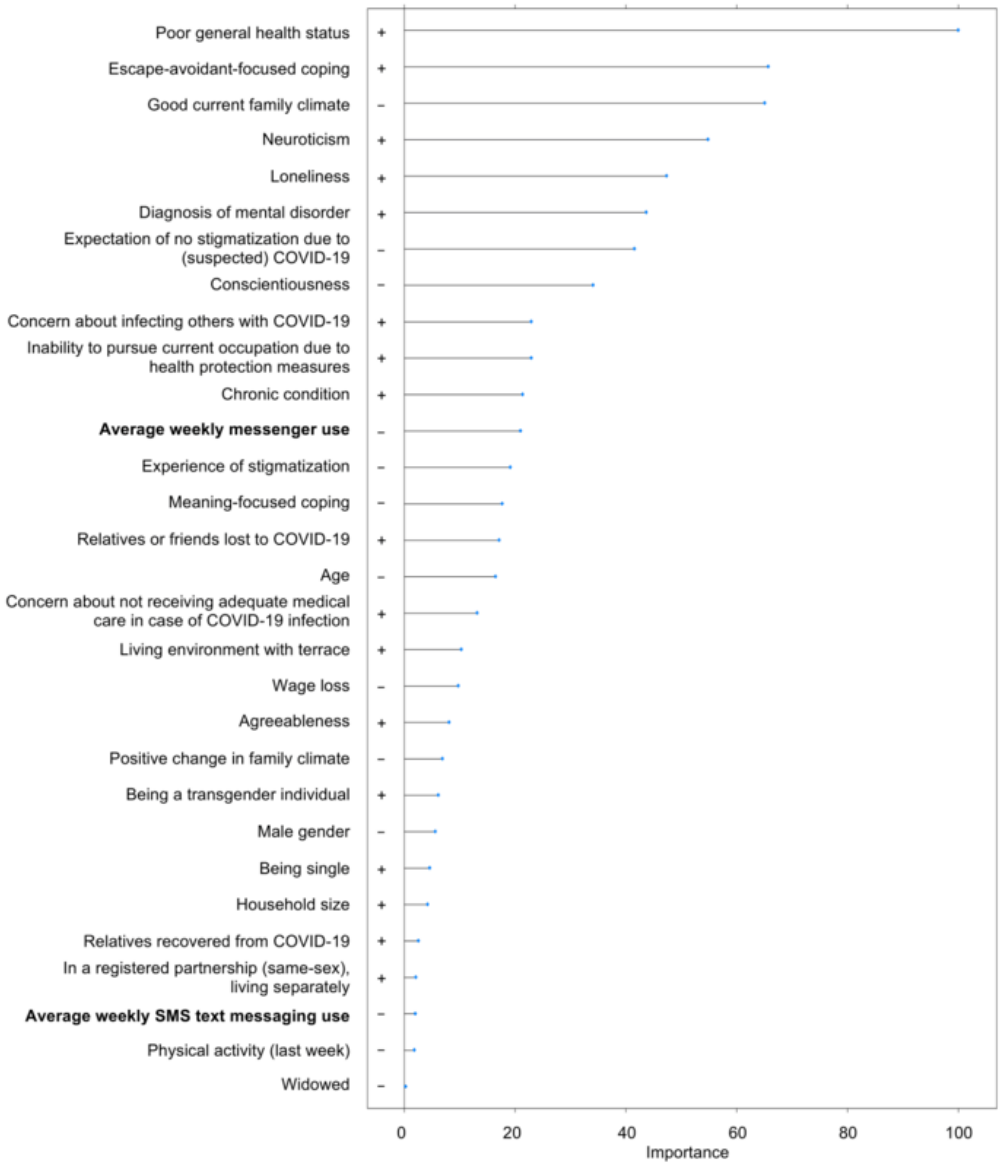

Discussion

Principal Findings

This study investigated the prediction of depressive and anxiety symptoms on the basis of smartphone-based social interaction data supplemented by a comprehensive set of well-established predictors, while also taking into account relevant pandemic-specific factors. The results show that when well-established predictors are considered, only the average weekly duration of messenger use as a negative predictor adds significant value to the prediction of both depressive and anxiety symptoms. In only 1 of the 2 prediction models, the average weekly duration of video calls (depressive symptoms) and SMS text messaging use (anxiety symptoms) remained relevant.

These findings are based on data collected during the COVID-19 pandemic and underscore the importance of messenger, video call, and SMS text messaging use for maintaining social contact. This interpretation aligns with the fact that social contact is considered a preventive factor for depressive and anxiety symptoms in general [69,70] as well as during the COVID-19 pandemic in particular, as already discussed by several studies that have reported an increase in digital face-to-face interactions [71]. This point is particularly relevant in view of the social distancing measures imposed to reduce COVID-19 infection rates because in times of lockdown, digital interactions were sometimes the only means to achieve face-to-face contact and group meetings [16]. It is also conceivable that as depressive symptoms increase, accompanied by social withdrawal, smartphone-based interaction is also reduced.

In contrast to previous findings [21,22], other smartphone-based social interaction parameters (weekly averaged use duration of phone calls and social media, as well as weekly averaged total smartphone use duration) did not contribute to the prediction of depressive or anxiety symptoms in the context of the COVID-19 pandemic. One possible explanation is that prior studies examined individual or only a few nonsmartphone-based social interaction parameters; for example, loneliness is a well-established correlate of depressive symptoms [46-48]. Consistent with previously published results from this study

XSL•FO
RenderX

[72] and other research [73], certain social interaction parameters, such as messenger use, are negatively correlated with loneliness. Thus, if loneliness is not taken into consideration, various smartphone parameters might seem to be significant correlates of depressive symptoms through spurious correlation. To avoid this effect, we included a comprehensive set of reliable predictors to provide insights into the additional value of smartphone-based social interaction data as predictors of depressive and anxiety symptoms. Future studies should also explicitly investigate the predictive value of smartphone-based social interaction data in isolation because these data are particularly attractive and accessible from a research economic point of view.

Another aspect that should be considered is the poor resolution of smartphone-based social interaction data. Whereas some studies have analyzed individual use patterns or changes over time [15,74], we decided to use a comprehensive but coarse-grained model and thus were unable to take into consideration more detailed information on, for example, content, motivation, or satisfaction with interactions. However, it has been found that the correlation between (social) media use and the symptoms of mental disorders varies according to use motives [75] and apps used [76]. A more detailed high-resolution analysis might help clarify associations between smartphone-based social interaction data and the symptoms of mental disorders in more detail.

Moreover, by taking into account a large number of proven predictors of depressive and anxiety symptoms, the results highlight the relevance of pandemic-related changes as additional predictors. The pandemic-related variables that remained significant in both our final models can be divided into internal aspects (cognitive appraisal processes and emotions) and external aspects (changes in daily life due to pandemic control measures).

The remaining internal aspects include expected stigmatization due to (suspected) COVID-19 infection and concern about not receiving adequate medical care in case of COVID-19 infection due to lack of medical capacity. In addition, the pandemic may have exacerbated the impact of other internal aspects of social interaction on feelings of loneliness [26] and family climate [27], which were already considered risk factors for the development of depressive and anxiety symptoms before the COVID-19 pandemic [46-48]. Correspondingly, loneliness and family climate were among the most important predictors of depressive and anxiety symptoms in our research. This result seems understandable if we view loneliness as an expression of an unmet need for attachment, which might increase the risk for developing depressive or anxiety symptoms. Similarly, social withdrawal can be an aspect of depressive symptoms and lead to family problems. By contrast, a subjective feeling of social connection prevents depressive symptoms. Family climate could be an indicator of the number of conflicts or tasks to be managed and thus an additional stressor that favors the development of psychopathological symptoms. Similarly, having a positive family climate as a resource could prevent the development of psychopathological symptoms.

The external aspect experienced in everyday life that predicts symptoms of depression and anxiety is the inability to pursue one's current occupation due to health protection measures (eg, due to the closure or suspension of work activities). With this study, we were able to replicate previous findings in a more comprehensive statistical model that also included passive smartphone-based social interaction data [24-26]. The fact that the accumulation of stressors during the COVID-19 pandemic, along with the accompanying changes and limitations, can lead to an increase in depressive and anxiety symptoms was discussed [29,77]. Our data also showed high wage loss as a positive predictor of both depressive and anxiety symptoms, whereas a stable income seemed to correlate negatively with the development of psychopathological symptoms.

In contrast to other studies [78], our data indicated that losing relatives or friends to COVID-19 was a negative predictor of depressive and anxiety symptoms. In line with previous research findings, general health status, such as individual differences, chronic conditions, and a lifetime diagnosis of a mental disorder, showed particular relevance in relation to depressive and anxiety symptoms [29,30]. In addition, relations may have become even stronger in the context of the COVID-19 pandemic [79] because treatment services for people with both physical and mental health problems were temporarily reduced due to contact restrictions [80,81]. At the same time, willingness to seek help decreased, presumably due to concerns about contracting COVID-19 infection [79,82]. It can also be assumed that persons with poor physical health status have an increased risk of severe illness if infected with COVID-19 [34], possibly leading to increased health concerns that favor depressive and anxiety symptoms [83].

Coping strategies also played a role in predicting depressive and anxiety symptoms [53]. In line with previous findings [84], our results showed a positive association of escape-avoidant–focused coping with depressive and anxiety symptoms as well as a negative association with meaning-focused coping. The fact that coping strategies are gaining relevance is particularly understandable in view of the COVID-19 pandemic [53]: it is only when major stress arises in one's life that the functionality and benefits of one's coping strategies become evident; for example, Fullana et al [84] were able to show that reduced consumption of news concerning COVID-19 and pursuing hobbies were negatively related to depressive symptoms.

In addition, the personality facets of conscientiousness, agreeableness, and neuroticism proved to be significant predictors of depressive and anxiety symptoms, consistent with previous findings [85]. Due to the pandemic-related dissolution of work boundaries, it is conceivable that conscientiousness may have become even more relevant; for example, mandatory work from home shifted the responsibility for organizing work time entirely to the employee [86]. In this context, conscientiousness is particularly valuable for meeting professional requirements as well as managing tasks such as home schooling children or caring for underage children. By contrast, neuroticism was a positive predictor of depressive and anxiety symptoms, which is to be expected [87-89] because neuroticism's core facet is trait anxiety [90]. In contrast to other

XSL•FO
RenderX

research [87,88,91], our data showed that agreeableness was a positive predictor of depressive and anxiety symptoms.

Consistent with previous research [92], we found that older age was associated with lower depressive and anxiety symptoms. Younger age, already a risk factor for the development of depressive symptoms [93], seemed to play a particularly important role in the pandemic [92]. One of the reasons discussed is that younger people are much more connected via media and thus exposed to the news, which could have acted as a stressor during the COVID-19 pandemic and in the context of related news coverage [94]. Again, the task of analyzing this interaction effect remains an area for future studies. Another reason discussed is the greater vulnerability to stress and depressive symptoms associated with younger age [95]. If one also considers the restrictions in the everyday life of young people during the COVID-19 pandemic, it is noticeable that meeting friends, partying, dancing, attending vocational school in person, and collective studying suddenly disappeared due to contact restrictions. Accordingly, there was evidence of stressors such as feelings of loneliness and social isolation, especially among young adults [96]. Household size also proved to be a positive predictor of depressive and anxiety symptoms. This result is in line with previous findings that have discussed the fact that shared living space is associated with increased stress, depending on the number of residents [97].

The results provide initial indications that, taking established predictors into account, depressive and anxiety symptoms can be predicted using smartphone-based social interaction data. This represents an important step toward a future in which the early recognition of depressive symptoms via smartphones may become feasible. Even today, automated feedback based on smartphone-based social interaction data could enable early preventive health-promoting measures tailored to individual results.

Strengths and Limitations

A major strength of this study is the comprehensive set of data comprising smartphone-based social interaction app use data as well as survey data. These data allowed us to create a holistic model of mental health predictors during the COVID-19 pandemic. Moreover, big data methodology was used, which allowed us to investigate the value of passive smartphone-based social interaction parameters for the prediction of depressive or anxiety symptoms while taking into account a broad set of other well-established relevant factors. This approach made it possible to show that passive smartphone-based social interaction data contribute to the explanation of variance beyond established predictors, which emphasizes the relevance of these new data. Furthermore, by collecting data via smartphones, a large number of participants could be reached and interviewed despite contact restrictions.

A limitation of the study is the lack of representativeness of the data. The group with a high level of education and younger participants are overrepresented. In addition, the cross-sectional correlational study design is a limitation because it is not possible to investigate predictive direction for a large number of variables that are unlikely to be antecedents of depressive or anxiety symptoms (eg, household, age, and gender). The latter variables include all smartphone-based interaction data and a large share of putative internal predictors. The inclusion of established strong predictors (eg, chronic condition and partnership status) may have limited the explanatory power of the model with respect to weaker risk factors (eg, smartphone-based social interaction data). The collection of various sorts of smartphone-based social interaction data also presents a methodological challenge because many apps can be used to communicate in different ways (eg, written messages or video calls). Thus, the mode of communication cannot be clearly ascertained. Due to a lack of comparative data before the pandemic, potential changes over time could also not be taken into account.

Conclusions

This study indicates that data concerning the time spent engaging in smartphone-based social interaction adds only limited value regarding the prediction of depressive and anxiety symptoms among German adults during the COVID-19 pandemic. Although predictive direction cannot be established, these results are in line with models of the etiology of depression and anxiety.

Data Availability

The data sets generated and analyzed during this study are not publicly available because the participants' informed consent did not cover consent for making the data publicly available but are available from the corresponding author on reasonable request.

Authors' Contributions

JSE developed the research questions and methodology in consultation with CC. YT provided advice regarding the planning of the statistical analyses. RP developed the app and data collection processes. JSE performed the data analyses and wrote the manuscript in consultation with YT, RP, HB, and CC. All authors critically revised the paper and approved the final revision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics of the sample of German-speaking adults (n=490) in total and grouped by gender.
[DOCX File , 33 KB - jmir_v26i1e45530_app1.docx]

XSL•FO
RenderX

References

1. Steffen A, Thom J, Jacobi F, Holstiege J, Bätzing J. Trends in prevalence of depression in Germany between 2009 and 2017 based on nationwide ambulatory claims data. J Affect Disord 2020 Jun 15;271:239-247. [doi: 10.1016/j.jad.2020.03.082]
2. Busch MA, Maske UE, Ryl L, Schlack R, Hapke U. Prävalenz von depressiver symptomatik und diagnostizierter depression bei erwachsenen in Deutschland. Bundesgesundheitsbl 2013 May 25;56:733-739. [doi: 10.1007/s00103-013-1688-3]
3. Aktuelle ergebnisse zur entwicklung der psychischen gesundheit der erwachsenen bevölkerung bei hochfrequenter beobachtung. Robert Koch Institut. 2023. URL: https://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/MHS/Quartalsberichte/2023-Q2_MHS-Bericht.pdf?__blob=publicationFile [accessed 2024-07-04]
4. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2019 Apr 05;29:e30. [doi: 10.1017/s2045796019000180]
5. Meng X, Brunet A, Turecki G, Liu A, D'Arcy C, Caron J. Risk factor modifications and depression incidence: a 4-year longitudinal Canadian cohort of the Montreal Catchment Area Study. BMJ Open 2017 Jun 10;7(6):e015156. [doi: 10.1136/bmjopen-2016-015156]
6. Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: a survey. Pervasive Mob Comput 2018 Dec;51:1-26. [doi: 10.1016/j.pmcj.2018.09.003]
7. Montag C, Sindermann C, Baumeister H. Digital phenotyping in psychological and medical sciences: a reflection about necessary prerequisites to reduce harm and increase benefits. Curr Opin Psychol 2020 Dec;36:19-24. [doi: 10.1016/j.copsyc.2020.03.013]
8. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
9. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226 [FREE Full text] [doi: 10.1037/prj0000130] [Medline: 25844912]
10. Demirci K, Akgönül M, Akpinar A. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. J Behav Addict 2015 Jun;4(2):85-92 [FREE Full text] [doi: 10.1556/2006.4.2015.010] [Medline: 26132913]
11. Elhai JD, Dvorak RD, Levine JC, Hall BJ. Problematic smartphone use: a conceptual overview and systematic review of relations with anxiety and depression psychopathology. J Affect Disord 2017 Jan 01;207:251-259. [doi: 10.1016/j.jad.2016.08.030] [Medline: 27736736]
12. Elhai JD, Levine JC, Dvorak RD, Hall BJ. Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use. Comput Hum Behav 2016 Oct;63:509-516. [doi: 10.1016/j.chb.2016.05.079]
13. Messner EM, Sariyska R, Mayer B, Montag C, Kannen C, Schwerdtfeger A, et al. Insights: anwendungsmöglichkeiten von passivem smartphone-tracking im therapeutischen kontext. Verhaltenstherapie 2019 Sep 13;29(3):155-165. [doi: 10.1159/000501735]
14. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021 Jan 28;12:625247 [FREE Full text] [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
15. Rozgonjuk D, Levine JC, Hall BJ, Elhai JD. The association between problematic smartphone use, depression and anxiety symptom severity, and objectively measured smartphone use over one week. Comput Hum Behav 2018 Oct;87:10-17. [doi: 10.1016/j.chb.2018.05.019]
16. David ME, Roberts JA. Smartphone use during the COVID-19 pandemic: social versus physical distancing. Int J Environ Res Public Health 2021 Jan 25;18(3):1034 [FREE Full text] [doi: 10.3390/ijerph18031034] [Medline: 33503907]
17. Lee J, Ahn JS, Min S, Kim MH. Psychological characteristics and addiction propensity according to content type of smartphone use. Int J Environ Res Public Health 2020 Mar 29;17(7):2292 [FREE Full text] [doi: 10.3390/ijerph17072292] [Medline: 32235319]
18. Steers ML, Wickham RE, Acitelli LK. Seeing everyone else's highlight reels: how Facebook usage is linked to depressive symptoms. J Soc Clin Psychol 2014 Oct;33(8):701-731. [doi: 10.1521/jscp.2014.33.8.701]
19. Elhai JD, Levine JC, Dvorak RD, Hall BJ. Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. Comput Hum Behav 2017 Apr;69:75-82. [doi: 10.1016/j.chb.2016.12.023]
20. Sommerlad A, Marston L, Huntley J, Livingston G, Lewis G, Steptoe A, et al. Social relationships and depression during the COVID-19 lockdown: longitudinal analysis of the COVID-19 social study. Psychol Med 2021 Jan 13;52(15):3381-3390. [doi: 10.1017/s0033291721000039]
21. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572 [FREE Full text] [doi: 10.3390/s20123572] [Medline: 32599801]
22. Ware S, Yue C, Morillo R, Lu J, Shang C, Bi J, et al. Predicting depressive symptoms using smartphone data. Smart Health 2020 Mar;15:100093. [doi: 10.1016/j.smhl.2019.100093]

XSL•FO
RenderX

23. Benke C, Autenrieth LK, Asselmann E, Pané-Farré CA. Lockdown, quarantine measures, and social distancing: associations with depression, anxiety and distress at the beginning of the COVID-19 pandemic among adults from Germany. Psychiatry Res 2020 Nov;293:113462 [FREE Full text] [doi: 10.1016/j.psychres.2020.113462] [Medline: 32987222]
24. Bagcchi S. Stigma during the COVID-19 pandemic. Lancet Infect Dis 2020 Jul;20(7):782 [FREE Full text] [doi: 10.1016/S1473-3099(20)30498-9] [Medline: 32592670]
25. Mertens G, Gerritsen L, Duijndam S, Salemink E, Engelhard IM. Fear of the coronavirus (COVID-19): predictors in an online study conducted in March 2020. J Anxiety Disord 2020 Aug;74:102258 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102258] [Medline: 32569905]
26. Dahlberg L. Loneliness during the COVID-19 pandemic. Aging Ment Health 2021 Jul;25(7):1161-1164. [doi: 10.1080/13607863.2021.1875195] [Medline: 33491474]
27. Ebert C, Steinert JI. Prevalence and risk factors of violence against women and children during COVID-19, Germany. Bull World Health Organ 2021 Mar 19;99:429-438. [doi: 10.2471/blt.20.270983]
28. Thorell LB, Skoglund C, de la Peña AG, Baeyens D, Fuermaier AB, Groom MJ, et al. Parental experiences of homeschooling during the COVID-19 pandemic: differences between seven European countries and between children with and without mental health conditions. Eur Child Adolesc Psychiatry 2022 Apr;31(4):649-661 [FREE Full text] [doi: 10.1007/s00787-020-01706-1] [Medline: 33415470]
29. Bäuerle A, Steinbach J, Schweda A, Beckord J, Hetkamp M, Weismüller B, et al. Mental health burden of the COVID-19 outbreak in Germany: predictors of mental health impairment. J Prim Care Community Health 2020;11:2150132720953682 [FREE Full text] [doi: 10.1177/2150132720953682] [Medline: 32865107]
30. Iob E, Frank P, Steptoe A, Fancourt D. Levels of severity of depressive symptoms among at-risk groups in the UK during the COVID-19 pandemic. JAMA Netw Open 2020 Oct 01;3(10):e2026064 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.26064] [Medline: 33104209]
31. Calvano C, Engelke L, Di Bella J, Kindermann J, Renneberg B, Winter SM. Families in the COVID-19 pandemic: parental stress, parent mental health and the occurrence of adverse childhood experiences-results of a representative survey in Germany. Eur Child Adolesc Psychiatry 2022 Jul;31(7):1-13 [FREE Full text] [doi: 10.1007/s00787-021-01739-0] [Medline: 33646416]
32. Morawa E, Schug C, Geiser F, Beschoner P, Jerg-Bretzke L, Albus C, et al. Psychosocial burden and working conditions during the COVID-19 pandemic in Germany: the VOICE survey among 3678 health care workers in hospitals. J Psychosom Res 2021 May;144:110415 [FREE Full text] [doi: 10.1016/j.jpsychores.2021.110415] [Medline: 33743398]
33. Zinn S, Bayer M. [Parents' subjective burden of home schooling their children at times of the Corona-related lockdown in spring 2020]. Z Erziehwiss 2021;24(2):339-365 [FREE Full text] [doi: 10.1007/s11618-021-01012-9] [Medline: 33821143]
34. Booth A, Reed AB, Ponzo S, Yassaee A, Aral M, Plans D, et al. Population risk factors for severe disease and mortality in COVID-19: a global systematic review and meta-analysis. PLoS One 2021 Mar 4;16(3):e0247461 [FREE Full text] [doi: 10.1371/journal.pone.0247461] [Medline: 33661992]
35. Jung S, Kneer J, Krüger TH. Mental health, sense of coherence, and interpersonal violence during the COVID-19 pandemic lockdown in Germany. J Clin Med 2020 Nov 18;9(11):3708 [FREE Full text] [doi: 10.3390/jcm9113708] [Medline: 33218215]
36. Koopmann A, Georgiadou E, Kiefer F, Hillemacher T. Did the general population in Germany drink more alcohol during the COVID-19 pandemic lockdown? Alcohol Alcohol 2020 Oct 20;55(6):698-699 [FREE Full text] [doi: 10.1093/alcalc/agaa058] [Medline: 32556079]
37. Heuse S, Gekeler B, Fodor D. The role of physical exercise as a personal resource against job stress. Int J Occup Saf Ergon 2021 Dec;27(4):1251-1260. [doi: 10.1080/10803548.2020.1732648] [Medline: 32326847]
38. Ingram RE, Luxton DD. Vulnerability-stress models. In: Hankin A, editor. Development of Psychopathology. Thousand Oaks, CA: SAGE Publications; 2005.
39. Zhong J, Zhong C, Qiu L, Li J, Lai J, Lu W, et al. Risk and protective factors for anxiety during COVID-19 pandemic. BMC Public Health 2021 Jun 04;21(1):1063 [FREE Full text] [doi: 10.1186/s12889-021-11118-8] [Medline: 34088290]
40. Marktanteile der führenden mobilen betriebssysteme an der internetnutzung mit mobiltelefonen in Deutschland von Januar 2009 bis Mai 2024. Statista. URL: https://de.statista.com/statistik/daten/studie/184332/umfrage/marktanteil-der-mobilen-betriebssysteme-in-deutschland-seit-2009/#professional [accessed 2021-08-06]
41. Beierle F, Schobel J, Vogel C, Allgaier J, Mulansky L, Haug F, et al. Corona health-a study- and sensor-based mobile app platform exploring aspects of the COVID-19 pandemic. Int J Environ Res Public Health 2021 Jul 10;18(14):7395 [FREE Full text] [doi: 10.3390/ijerph18147395] [Medline: 34299846]
42. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019 Apr;13(Suppl 1):S31-S34 [FREE Full text] [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]
43. Löwe B, Spitzer RL, Zipfel S, Herzog W. PHQ-D Manual. Pfizer. 2002. URL: https://www.klinikum.uni-heidelberg.de/fileadmin/Psychosomatische_Klinik/download/PHQ_Manual1.pdf [accessed 2024-07-30]
44. Sæther SM, Knapstad M, Grey N, Smith OR. Twelve months post-treatment results from the Norwegian version of improving access to psychological therapies. Front Psychol 2019 Oct 18;10:2303 [FREE Full text] [doi: 10.3389/fpsyg.2019.02303] [Medline: 31681099]

XSL•FO
RenderX

45. Schulze B, Stuart H, Riedel-Heller SG. [The German version of the Inventory of Stigmatizing Experiences (ISE)--a new tool for assessing the prevalence and impact of "felt stigma"]. Psychiatr Prax 2009 Nov;36(8):e19-e27. [doi: 10.1055/s-0029-1223337] [Medline: 19753514]
46. Beutel ME, Klein EM, Brähler E, Reiner I, Jünger C, Michal M, et al. Loneliness in the general population: prevalence, determinants and relations to mental health. BMC Psychiatry 2017 Mar 20;17(1):97 [FREE Full text] [doi: 10.1186/s12888-017-1262-x] [Medline: 28320380]
47. Cacioppo JT, Hughes ME, Waite LJ, Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. Psychol Aging 2006 Mar;21(1):140-151. [doi: 10.1037/0882-7974.21.1.140] [Medline: 16594799]
48. Jaremka LM, Fagundes CP, Glaser R, Bennett JM, Malarkey WB, Kiecolt-Glaser JK. Loneliness predicts pain, depression, and fatigue: understanding the role of immune dysregulation. Psychoneuroendocrinology 2013 Aug;38(8):1310-1317 [FREE Full text] [doi: 10.1016/j.psyneuen.2012.11.016] [Medline: 23273678]
49. Gilman SE, Trinh NH, Smoller JW, Fava M, Murphy JM, Breslau J. Psychosocial stressors and the prognosis of major depression: a test of Axis IV. Psychol Med 2012 May 28;43(2):303-316. [doi: 10.1017/s0033291712001080]
50. Yu Y, Yang X, Yang Y, Chen L, Qiu X, Qiao Z, et al. The role of family environment in depressive symptoms among university students: a large sample survey in China. PLoS One 2015 Dec 2;10(12):e0143612 [FREE Full text] [doi: 10.1371/journal.pone.0143612] [Medline: 26629694]
51. Yuan W, Hesketh T. Intimate partner violence and depression in women in China. J Interpers Violence 2021 Nov 02;36(21-22):NP12016-NP12040. [doi: 10.1177/0886260519888538] [Medline: 31789083]
52. Chang-Quan H, Xue-Mei Z, Bi-Rong D, Zhen-Chan L, Ji-Rong Y, Qing-Xiu L. Health status and risk for depression among the elderly: a meta-analysis of published literature. Age Ageing 2010 Jan 10;39(1):23-30. [doi: 10.1093/ageing/afp187] [Medline: 19903775]
53. Mariani R, Renzi A, Di Trani M, Trabucchi G, Danskin K, Tambelli R. The impact of coping strategies and perceived family support on depressive and anxious symptomatology during the coronavirus pandemic (COVID-19) lockdown. Front Psychiatry 2020 Nov 13;11:587724 [FREE Full text] [doi: 10.3389/fpsyt.2020.587724] [Medline: 33281647]
54. Köhler CA, Evangelou E, Stubbs B, Solmi M, Veronese N, Belbasis L, et al. Mapping risk factors for depression across the lifespan: an umbrella review of evidence from meta-analyses and Mendelian randomization studies. J Psychiatr Res 2018 Aug;103:189-207 [FREE Full text] [doi: 10.1016/j.jpsychires.2018.05.020] [Medline: 29886003]
55. Gianfredi V, Blandi L, Cacitti S, Minelli M, Signorelli C, Amerio A, et al. Depression and objectively measured physical activity: a systematic review and meta-analysis. Int J Environ Res Public Health 2020 May 25;17(10):3738 [FREE Full text] [doi: 10.3390/ijerph17103738] [Medline: 32466242]
56. Richter D, Weinhardt M. LS-S: Loneliness Scale-SOEP. In: Kemper CJ, Zenger M, Brähler E, editors. Psychologische und Sozialwissenschaftliche Kurzskalen: Standardisierte Erhebungsinstrumente für Wissenschaft und Praxis. Berlin, Germany: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft; 2013.
57. Gräfe K, Zipfel S, Herzog W, Löwe B. Screening psychischer störungen mit dem “Gesundheitsfragebogen für Patienten (PHQ-D)“. Diagnostica 2004 Oct;50(4):171-181. [doi: 10.1026/0012-1924.50.4.171]
58. Cox B, van Oyen H, Cambois E, Jagger C, le Roy S, Robine JM, et al. The reliability of the Minimum European Health Module. Int J Public Health 2009 Jan 30;54(2):55-60. [doi: 10.1007/s00038-009-7104-y] [Medline: 19183846]
59. Carver CS, Scheier MF, Weintraub JK. Assessing coping strategies: a theoretically based approach. J Pers Soc Psychol 1989 Mar;56(2):267-283. [doi: 10.1037//0022-3514.56.2.267] [Medline: 2926629]
60. Cohrdes C, Pryss R, Baumeister H, Eicher S, Knoll N, Hölling H. Support- and meaning-focused coping as key factors for maintaining adult quality of life during the COVID-19 pandemic in Germany. Front Public Health 2023;11:1196404 [FREE Full text] [doi: 10.3389/fpubh.2023.1196404] [Medline: 37377548]
61. Carver CS. You want to measure coping but your protocol's too long: consider the brief COPE. Int J Behav Med 1997 Mar;4(1):92-100. [doi: 10.1207/s15327558ijbm0401_6] [Medline: 16250744]
62. Knoll N, Rieckmann N, Schwarzer R. Coping as a mediator between personality and stress outcomes: a longitudinal study with cataract surgery patients. Eur J Pers 2005;19(3):229-247. [doi: 10.1002/per.546]
63. Rammstedt B, Kemper CJ, Klein MC, Beierlein C, Kovaleva A. Big Five Inventory (BFI-10). Zusammenstellung Sozialwissenschaftlicher Items und Skalen (ZIS). 2014. URL: https://zis.gesis.org/skala/Rammstedt-Kemper-Klein-Beierlein-Kovaleva-Big-Five-Inventory-(BFI-10) [accessed 2024-08-08]
64. Lechert Y, Schroedter JH, Lüttinger P. Die umsetzung der Bildungsklassifikation CASMIN für die Volkszählung 1970, die Mikrozensus-Zusatzerhebung 1971 und die mikrozensen 1976-2004. Social Science Open Access Repository. 2006. URL: https://www.ssoar.info/ssoar/handle/document/26235 [accessed 2024-07-31]
65. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: With Applications in R. New York, NY: Springer; 2013.
66. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005 Mar 09;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
67. Zhao Y, Long Q. Variable selection in the presence of missing data: imputation-based methods. Wiley Interdiscip Rev Comput Stat 2017;9(5):e1402 [FREE Full text] [doi: 10.1002/wics.1402] [Medline: 29085552]

XSL•FO
RenderX

68. Multivariate imputation by chained equations. The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/mice/mice.pdf [accessed 2021-01-09]
69. Domènech-Abella J, Mundó J, Haro JM, Rubio-Valera M. Anxiety, depression, loneliness and social network in the elderly: longitudinal associations from The Irish Longitudinal Study on Ageing (TILDA). J Affect Disord 2019 Mar 01;246:82-88. [doi: 10.1016/j.jad.2018.12.043] [Medline: 30578950]
70. Sakurai R, Kawai H, Suzuki H, Kim H, Watanabe Y, Hirano H, et al. Association of eating alone with depression among older adults living alone: role of poor social networks. J Epidemiol 2021;31(4):297-300. [doi: 10.2188/jea.je20190217]
71. Hacker J, vom Brocke J, Handali J, Otto M, Schneider J. Virtually in this together – how web-conferencing systems enabled a new virtual togetherness during the COVID-19 crisis. Eur J Inf Syst 2020 Sep 29;29(5):563-584. [doi: 10.1080/0960085x.2020.1814680]
72. Wetzel B, Pryss R, Baumeister H, Edler JS, Gonçalves AS, Cohrdes C. "How come you don't call me?" smartphone communication app usage as an indicator of loneliness and social well-being across the adult lifespan during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Jun 08;18(12):6212 [FREE Full text] [doi: 10.3390/ijerph18126212] [Medline: 34201322]
73. Gao Y, Li A, Zhu T, Liu X, Liu X. How smartphone usage correlates with social anxiety and loneliness. PeerJ 2016 Jul 12;4:e2197 [FREE Full text] [doi: 10.7717/peerj.2197] [Medline: 27478700]
74. Asare KO, Visuri A, Ferreira DS. Towards early detection of depression through smartphone sensing. In: Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. 2019 Presented at: UbiComp/ISWC '19; September 9-13, 2019; London, UK. [doi: 10.1145/3341162.3347075]
75. Brailovskaia J, Schillack H, Margraf J. Tell me why are you using social media (SM)! Relationship between reasons for use of SM, SM flow, daily stress, depression, anxiety, and addictive SM use – an exploratory investigation of young adults in Germany. Comput Hum Behav 2020 Dec;113:106511. [doi: 10.1016/j.chb.2020.106511]
76. Sha P, Sariyska R, Riedl R, Lachmann B, Montag C. Linking internet communication and smartphone use disorder by taking a closer look at the Facebook and WhatsApp applications. Addict Behav Rep 2018 Nov 23;9:100148 [FREE Full text] [doi: 10.1016/j.abrep.2018.100148] [Medline: 31193857]
77. Petzold MB, Bendau A, Plag J, Pyrkosch L, Mascarell Maricic L, Betzler F, et al. Risk, resilience, psychological distress, and anxiety at the beginning of the COVID-19 pandemic in Germany. Brain Behav 2020 Sep;10(9):e01745 [FREE Full text] [doi: 10.1002/brb3.1745] [Medline: 32633464]
78. Gozansky E, Moscona G, Okon-Singer H. Identifying variables that predict depression following the general lockdown during the COVID-19 pandemic. Front Psychol 2021 May 17;12:680768 [FREE Full text] [doi: 10.3389/fpsyg.2021.680768] [Medline: 34079505]
79. Joaquim RM, Pinto AL, Guatimosim RF, de Paula JJ, Souza Costa D, Diaz AP, et al. Bereavement and psychological distress during COVID-19 pandemics: the impact of death experience on mental health. Curr Res Behav Sci 2021 Nov;2:100019. [doi: 10.1016/j.crbeha.2021.100019]
80. Czeisler M, Marynak K, Clarke KE, Salah Z, Shakya I, Thierry JM, et al. Delay or avoidance of medical care because of COVID-19-related concerns - United States, June 2020. MMWR Morb Mortal Wkly Rep 2020 Sep 11;69(36):1250-1257 [FREE Full text] [doi: 10.15585/mmwr.mm6936a4] [Medline: 32915166]
81. Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, et al. Reduced rate of inpatient hospital admissions in 18 German university hospitals during the COVID-19 lockdown. Front Public Health 2021 Jan 13;8:594117 [FREE Full text] [doi: 10.3389/fpubh.2020.594117] [Medline: 33520914]
82. Adorjan K, Pogarell O, Streb D, Padberg F, Erdmann C, Koller G, et al. Role of psychiatric hospitals during a pandemic: introducing the Munich Psychiatric COVID-19 Pandemic Contingency Plan. BJPsych Open 2021 Feb 01;7(2):e41 [FREE Full text] [doi: 10.1192/bjo.2020.167] [Medline: 33517940]
83. Hoyer C, Ebert A, Szabo K, Platten M, Meyer-Lindenberg A, Kranaster L. Decreased utilization of mental health emergency service during the COVID-19 pandemic. Eur Arch Psychiatry Clin Neurosci 2021 Mar;271(2):377-379 [FREE Full text] [doi: 10.1007/s00406-020-01151-w] [Medline: 32519205]
84. Fullana MA, Hidalgo-Mazzei D, Vieta E, Radua J. Coping behaviors associated with decreased anxiety and depressive symptoms during the COVID-19 pandemic and lockdown. J Affect Disord 2020 Oct 01;275:80-81 [FREE Full text] [doi: 10.1016/j.jad.2020.06.027] [Medline: 32658829]
85. Shokrkon A, Nicoladis E. How personality traits of neuroticism and extroversion predict the effects of the COVID-19 on the mental health of Canadians. PLoS One 2021 May 19;16(5):e0251097 [FREE Full text] [doi: 10.1371/journal.pone.0251097] [Medline: 34010299]
86. Wang B, Liu Y, Qian J, Parker SK. Achieving effective remote working during the COVID-19 pandemic: a work design perspective. Appl Psychol 2021 Jan;70(1):16-59 [FREE Full text] [doi: 10.1111/apps.12290] [Medline: 33230359]
87. Al-Omiri MK, Alzoubi IA, Al Nazeh AA, Alomiri AK, Maswady MN, Lynch E. COVID-19 and personality: a cross-sectional multicenter study of the relationship between personality factors and COVID-19-related impacts, concerns, and behaviors. Front Psychiatry 2021 Feb 26;12:608730 [FREE Full text] [doi: 10.3389/fpsyt.2021.608730] [Medline: 33716815]

XSL•FO
RenderX

88. Nikčević AV, Marino C, Kolubinski DC, Leach D, Spada MM. Modelling the contribution of the Big Five personality traits, health anxiety, and COVID-19 psychological distress to generalised anxiety and depressive symptoms during the COVID-19 pandemic. J Affect Disord 2021 Jan 15;279:578-584 [FREE Full text] [doi: 10.1016/j.jad.2020.10.053] [Medline: 33152562]
89. Nudelman G, Kamble SV, Otto K. Can personality traits predict depression during the COVID-19 pandemic? Soc Just Res 2021 Jun 01;34(2):218-234 [FREE Full text] [doi: 10.1007/s11211-021-00369-w] [Medline: 34092913]
90. Lahey BB. Public health significance of neuroticism. Am Psychol 2009;64(4):241-256 [FREE Full text] [doi: 10.1037/a0015309] [Medline: 19449983]
91. Rammstedt B, Lechner CM, Weiß B. Does personality predict responses to the COVID-19 crisis? Evidence from a prospective large-scale study. Eur J Pers 2021 Mar 01;36(1):47-60. [doi: 10.1177/0890207021996970]
92. Kuehner C, Schultz K, Gass P, Meyer-Lindenberg A, Dreßing H. [Mental health status in the community during the COVID-19-pandemic]. Psychiatr Prax 2020 Oct;47(7):361-369 [FREE Full text] [doi: 10.1055/a-1222-9067] [Medline: 32823293]
93. Gustavson K, Knudsen AK, Nesvåg R, Knudsen GP, Vollset SE, Reichborn-Kjennerud T. Prevalence and stability of mental disorders among young adults: findings from a longitudinal study. BMC Psychiatry 2018 Mar 12;18(1):65 [FREE Full text] [doi: 10.1186/s12888-018-1647-5] [Medline: 29530018]
94. Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S, et al. Mental health problems and social media exposure during COVID-19 outbreak. PLoS One 2020 Apr 16;15(4):e0231924 [FREE Full text] [doi: 10.1371/journal.pone.0231924] [Medline: 32298385]
95. Varma P, Junge M, Meaklim H, Jackson ML. Younger people are more vulnerable to stress, anxiety and depression during COVID-19 pandemic: a global cross-sectional survey. Prog Neuropsychopharmacol Biol Psychiatry 2021 Jul 13;109:110236 [FREE Full text] [doi: 10.1016/j.pnpbp.2020.110236] [Medline: 33373680]
96. Kohls E, Baldofski S, Moeller R, Klemm SL, Rummel-Kluge C. Mental health, social and emotional well-being, and perceived burdens of university students during COVID-19 pandemic lockdown in Germany. Front Psychiatry 2021 Apr 6;12:643957 [FREE Full text] [doi: 10.3389/fpsyt.2021.643957] [Medline: 33889102]
97. Rodríguez-Rey R, Garrido-Hernansaiz H, Collado S. Psychological impact and associated factors during the initial stage of the coronavirus (COVID-19) pandemic among the general population in Spain. Front Psychol 2020 Jun 23;11:1540 [FREE Full text] [doi: 10.3389/fpsyg.2020.01540] [Medline: 32655463]

Abbreviations

Brief-COPE: Coping Orientation to Problems Experienced Inventory
LASSO: least absolute shrinkage and selection operator
MAE: mean absolute error
PHQ-D: Patient Health Questionnaire, German version
RMSE: root mean square error
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by T de Azevedo Cardoso, T Leung; submitted 10.01.23; peer-reviewed by T Shimamoto, F Regalado; comments to author 02.11.23; revised version received 19.04.24; accepted 14.06.24; published 16.09.24.

Please cite as:
Edler JS, Terhorst Y, Pryss R, Baumeister H, Cohrdes C
Messenger Use and Video Calls as Correlates of Depressive and Anxiety Symptoms: Results From the Corona Health App Study of German Adults During the COVID-19 Pandemic
J Med Internet Res 2024;26:e45530
URL: https://www.jmir.org/2024/1/e45530
doi:10.2196/45530
PMID:

©Johanna-Sophie Edler, Yannik Terhorst, Rüdiger Pryss, Harald Baumeister, Caroline Cohrdes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Crisis and Emergency Risk Communication and Emotional Appeals in COVID-19 Public Health Messaging: Quantitative Content Analysis

Shirley S Ho[1], PhD; Agnes S F Chuah[1], PhD; Vanessa S Ho[1], BA; Sonny Rosenthal[1], PhD; Hye Kyung Kim[1], PhD; Shannon S H Soh[2], BSc

[1]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore
[2]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

Corresponding Author:
Shirley S Ho, PhD
Wee Kim Wee School of Communication and Information
Nanyang Technological University
WKWSCI Building, 31 Nanyang Link
Singapore, 637718
Singapore
Email: tsyho@ntu.edu.sg

Abstract

Background: Although COVID-19 is no longer a global health emergency, it remains pervasive in Singapore, a city-state situated in Southeast Asia, with periodic waves of infection. In addition to disease management, strong communication strategies are critical in the government's response to the pandemic to keep the public updated and equip them in protecting themselves.

Objective: Grounded in the crisis and emergency risk communication (CERC) framework and emotional appeals, this study aimed to analyze public health communication strategies in Singapore during the COVID-19 pandemic.

Methods: Quantitative content analysis was conducted on 696 Facebook (Meta Platforms Inc) posts and 83 website articles published by Singapore-based public health institutions between January 2020 and September 2022.

Results: The results showed that increasing communication on message themes, such as inquisitive messaging and clarification, can enhance communication strategies. The use of emotional appeals also varies with time and should be carefully used as they are context-specific.

Conclusions: Theoretically, this study contributes to advancements in the CERC framework and concepts of emotional appeals by exploring the applications and changes of CERC message types and emotional appeals at different phases. The findings can provide practical guidance for authorities and communication practitioners in developing effective communication strategies.

(J Med Internet Res 2024;26:e56854) doi:10.2196/56854

KEYWORDS

COVID-19; crisis and emergency risk communication; CERC; emotional appeal; content analysis; public health; Facebook; social media; Singapore

Introduction

Background

Singapore effectively managed COVID-19, which is evident from the World Health Organization lauding its "all-of-government" approach [1]. This approach [1,2] entails collaboration among different government agencies [2]. While COVID-19 is no longer a global health emergency, Singapore continues to experience periodic infection waves [3]. During the pandemic, the Singaporean government charted its response to COVID-19 in stages, as detailed in a white paper [4]. Avenues for public health communication in Singapore include government websites and Facebook (Meta Platforms Inc) pages. These websites serve as a one-stop communications channel, and Facebook is one of Singapore's most widely used social-networking platforms [5]. However, studies on the government's use of Facebook for public health communication during the pandemic are limited.

Singapore's success in managing the pandemic can be attributed to its small population, concentrated political authority, high

XSL•FO
RenderX

political trust [6], state-supported media, and the 2003 SARS outbreak experience [7]. Despite this, Singapore faced criticism for the high number of COVID-19 cases in dormitories of migrant workers, due to the lack of communication [6]. Studies [8-13] have shown that media messages can shape public knowledge, attitudes, and preventive behaviors during pandemics in Singapore. It is worthwhile to study Singapore's public health communication during COVID-19 as it can highlight areas of improvement and offer insights for other countries in future crisis. This study had 4 objectives. First, it aimed to characterize the themes of public messages during the COVID-19 using the crisis and emergency risk communication (CERC) framework. Second, it aimed to examine how these message themes changed across different pandemic phases. Third, it aimed to identify the types of emotional appeals used. Fourth, it aimed to analyze how emotional appeals changed across the COVID-19 phases.

CERC Framework

CERC is well-suited for evaluating Singapore's public communication strategies during the COVID-19 pandemic. This is because CERC evolved in stages and involves both risk and crisis communications. CERC consists of 5 stages: precrisis, initial, maintenance, resolution, and evaluation. Communication during the *precrisis* stage focuses on educating the public about potential adverse events and risks to prepare them for the subsequent stages [14]. During the *initial* stage, communication messages focus on reducing uncertainty, conveying empathy, and imparting the general understanding of the crisis. The *maintenance* stage reiterates misinformation, ongoing risks, and mitigation strategies [15]. The *resolution* stage involves communicating how the emergency was handled, while the *evaluation* stage assesses response effectiveness [15].

The CERC framework assumes that crises develop in a linear way. However, due to the variability of diseases, crises may not follow the sequence of the outlined stages [14]. Although CERC suggests 5 stages, the precrisis stage did not apply to COVID-19 because it was not a known disease. The length of each stage may also vary, as a prolonged crisis state may occur [14]. For example, COVID-19 had a prolonged CERC maintenance stage as the virus mutated several times during the pandemic [16]. This has resulted in repeated tightening and easing of COVID-19 measures in Singapore [4].

CERC Themes

Drawing on the existing literature [17-19], this study categorized the CERC message themes into 4 categories: *risk and crisis information*, *self-efficacy and sense-making*, *preparations and uncertainty reduction*, and *advisories and alerts*. *Risk and crisis information* refers to information that educate the public about potential threats [14]. This category consists of a subtheme, *pandemic intelligence.* It refers to messages containing basic information about the pandemic, including case numbers [18], to raise awareness of the current situation.

The category *self-efficacy and sense-making* involves messages that help people to understand the situation and reflect their ability to change their behaviors [18]. This category includes 3 subthemes: *personal preventive measures and mitigation*, *social and common responsibility*, and *inquisitive messaging*. *Personal preventive measures and mitigation* refers to messages about measures or precautions that can be taken to protect the public from COVID-19. *Social and common responsibility* includes messages on measures or precautions that can be taken at the community level to prevent the spread of COVID-19 or to show care [20]. *Inquisitive messaging* addresses the public's questions to better understand the situation [18].

The category *preparations and uncertainty reduction* includes messages on how to act appropriately during the pandemic [14]. Drawing reference to Malik et al [18], *preparations and uncertainty* comprises 4 subthemes: *clarification*, *events, campaigns and activities*, *showing gratitude*, and *reassurance*. Clarification refers to messages addressing misunderstandings and untrue claims about the pandemic [20]. Events, campaigns, and activities include messages promoting communication campaigns for awareness, relief, or treatment. Showing gratitude refers to expressing appreciation to those involved in managing the virus, such as frontline workers [19]. Reassurance consists of messages that allay the public's fears.

The category *advisories and alerts* refers to messages that provide crucial warnings and specific advice about diseases. There are 2 subthemes: *risk groups* and *general advisories and vigilance*. The subtheme *risk groups* refers to messages targeting susceptible groups such as people with preexisting conditions and older adults who are at greater risk of contracting COVID-19 [20]. Messages on *general advisories and vigilance* include information on what to do in certain situations, such as returning to the workplace.

COVID-19 Phases and Social Media Use in Singapore

The Singapore government segmented the COVID-19 pandemic into 4 phases: *early days of fog*, *fighting a pandemic*, *rocky transition*, and *learning to live with COVID-19* [4], which correspond to the CERC stages (Table 1). However, empirical investigation is needed to examine whether the message themes were conveyed appropriately across these stages, especially on social media. The CERC framework has been used to evaluate public health communications on social media such as Facebook [17,21]. Vijaykumar et al [22] found out that information disseminated by Singapore-based public health institutions on Facebook were similar in content, but differed in focus. The Ministry of Health (MOH) focused on situational updates and the National Environment Agency (NEA) elaborated on preventive measures. However, the study only focused on public communication by these 2 agencies. To gain a broader understanding of crisis communications in Singapore; this study examined public communication by multiple government agencies in Singapore. Hence, we ask the following research questions (RQs):

- RQ1: To what extent are the CERC message themes present in Singapore's online public health messaging during the COVID-19 pandemic?
- RQ2: How do the CERC message themes change across different phases during the COVID-19 pandemic?

XSL•FO
RenderX

Table 1. White paper stages.

CERC[a] stages	White paper stage	Definition
Initial phase	Early days of fog: January, 2020, to March, 2020	• COVID-19 was first presented as a global health crisis as an unknown virus. • Singapore detected its first few cases [4]. • The Singapore government used the color-coded DORSCON[b] (MOH[c] [23]), to reflect the disease situation in Singapore. The DORSCON level was raised from green to yellow, and subsequently to orange. • This matches the initial phase of CERC, as the virus had already evolved into a crisis. In this period, uncertainty about the new disease was rife, which is characteristic of the initial phase [15].
Maintenance stage	Fighting a pandemic: April, 2020, to April, 2021	• This phase began with Singapore entering a "circuit breaker" phase on April 7, 2020, with increased safe distancing measures to reduce the spread of COVID-19 (MOH [24]), as the situation continued to worsen. • Work-from-home was mandated; dining-in was prohibited. • Vaccinations were made available first for health care workers and then for older citizens [4]. • This corresponds to the maintenance stage proposed by CERC when the COVID-19 situation in Singapore began to stabilize, with Singapore exiting the circuit-breaker stage and gradually relaxing other safe-management restrictions.
Maintenance stage	Rocky transition: May, 2021 to November, 2021	• Singapore learns to live with endemic COVID-19. • The newly discovered COVID-19 Delta variant caused a large increase in cases, supporting that the "zero COVID" strategy would not be practical in Singapore. • In response to the fluctuating COVID-19 cases [4], there were multiple tightening and easing of measures during this period. • This also corresponds to the maintenance stage, as outlined by CERC.
Resolution stage	Learning to live with COVID-19: December, 2021, to present	• This phase saw further relaxation of safe-management measures. Other measures, such as mask requirements and border measures, were eased. • The DORSCON level was stepped down from Orange to Yellow [4]. • There was a marked return to normality [20]. • Hence, this stage corresponded with the resolution stage proposed in the CERC.

[a]CERC: crisis and emergency risk communication.

[b]DORSCON: Disease Outbreak Response System Condition.

[c]MOH: Ministry of Health.

While CERC is extensively studied, there is limited research linking it with emotional appeals, a gap scholars find crucial to address. Meadows et al [25] argued that investigating the emotional tones of the public during different outbreak phases aids in formulating effective public health messages. This is echoed by Xie et al [26] who found that emotional appeals effectively engaged audiences. In addition to analyzing CERC message themes, this study also aimed to examine the use of emotional appeals in public health communication during COVID-19.

Emotional Appeals

Emotional appeals can persuade people to perform an intended behavior by evoking specific emotion [27]. They are widely used in health communications [28]; each type elicits varying responses. For example, people are divided on humor appeals; a few think it undermines the seriousness of the subject, while others find it useful [29]. The choice of emotional appeals depends on the context and the target audience [30]. During the COVID-19 pandemic, key emotional appeals included hope, humor, fear, anger, guilt, and nurturance [31,32].

Hope appeals emphasize efficacy and can be empowering when paired with actionable advice. During health crises, transparent communication about uncertainties and hopeful messages can enhance support for the measures implemented [32]. World Health Organization recommends using hope appeals to combat pandemic fatigue [33]. Hope appeals are an effective communication strategy across different cultures. In collectivist countries such as Singapore, hope appeals can focus on emerging stronger from COVID-19 as a community. *Humor appeals* use techniques such as clownish humor, irony, and satire [34], aimed at reducing negative emotions and promoting positivity [35]. However, they are also noted for potentially reducing social responsibility [36] and perceived crisis risk [37]. Humor appeals should be used tactfully especially during the critical phrases where increased perceived risk and social responsibility are crucial.

Fear appeals are the most widely studied emotional appeals. A message with fear appeals induces fear when a situation is seen as threatening to one's physical or mental health and is perceived as uncontrollable [38]. It evokes fear about the harm that will befall them if they do not adopt the recommended behavior [39]. The arousal triggered would create a desire to avoid the perceived threat and to adopt the suggested behavior, such as mask wearing and vaccination. Upon encountering the message, the audience would evaluate the severity and susceptibility of the threat, and their ability to overcome the threat, and subsequently taking the recommended action [40].

XSL•FO
RenderX

Anger appeals motivate people to carry out actions requiring more effort and commitment [41]. The anger activism model suggests that when coupled with a sense of efficacy, a person made to feel anger would feel motivated to perform a behavior [41]. Anger was one of the least used appeals in organizational YouTube videos during COVID-19 [26]. *Guilt appeals* consist of 2 components—material to evoke guilt and an action to reduce guilt [42]. The material can highlight discrepancies between the audience's standards and their behavior [43], which could effectively influence health-related attitudes [44]. However, excessive guilt can be counterproductive and less persuasive [45], as shown in the study by Matkovic et al [46], where guilt appeals failed to influence handwashing intention during the pandemic.

Nurturance appeals are defined as appeals that evoke a sense of caretaking, which effectively target parents [38]. Nurturance appeals were the most dominant emotional appeal in advertising materials using COVID-19 as a theme [31]. Given the dynamic nature of a crisis, it is important to use suitable emotional appeals at appropriate times and for effective management of the situation. A few studies focused on how emotional appeals were used in the communication messages during the COVID-19 pandemic (eg, a study by Mello et al [47]). Hence, we asked the following RQs:

- RQ3: What are the types of emotional appeals used in Singapore's online public health messaging during the COVID-19 pandemic?
- RQ4: How does the use of emotional appeals in Singapore's online public health messaging change across different CERC phases during the COVID-19 pandemic?

Methods

Overview

To answer our research questions, we conducted a quantitative content analysis on public Facebook posts and publicly accessible website articles from key Singapore government institutions involved in public health communication during the COVID-19 pandemic. Specifically, we compiled and analyzed content from Gov.sg, representing the Singapore government, as well as institutions, such as the MOH, the Ministry of Sustainability and the Environment, the NEA, and the Health Promotion Board.

Ethical Considerations

Before commencing data collection for content analysis, we sought approval from the Nanyang Technological University's Integrity Review Board (IRB-2022-725) in exempt category 4. This category pertained to secondary research using existing or publicly accessible data sets such as those found on social media. The exemption criteria included sources of individually identifiable information that were already in existence or that were publicly available. Obtaining IRB approval ensured that the research adheres to ethical standards, protecting the privacy and rights of individuals whose data were being analyzed. This step was crucial in maintaining the integrity and ethical compliance of the research project.

Data Collection and Sampling

Upon receiving IRB approval, we used a Python script to crawl Facebook posts containing specified keywords related to COVID-19 from January 1, 2020, to September 30, 2022. Concurrently, we manually compiled relevant website articles from the same timeframe through keyword searches on the institutions' websites. These keywords included "2019-nCoV," "SARS-CoV-2," "Sars-CoV-2," "Wuhan Coronavirus," "Wuhan coronavirus," "wuhan coronavirus," "Wuhan virus," "wuhan virus," "Wuhan Virus," "Covid-19," "covid-19," "novel coronavirus," "COVID," "Covid," and "covid." Articles and posts that are not related to the public health communication about COVID-19 (such as Facebook posts and website articles that solely focused on situational updates such as the number of cases and clusters, call outs to subscribe for updates, mentions of COVID-19 as a time frame where other activities or programs were the major topics, posts that did not focus on COVID-19, speeches by public figures, and press releases) were excluded.

This initial screening yielded a total of 1114 Facebook posts and 85 relevant website articles. The data were then randomly sampled with a CI level of 99% and a 3% margin of error, resulting in the final 696 Facebook posts and 83 website articles selected for detailed analysis.

Codebook and Coding Scheme

We developed the codebook on the basis of the CERC message themes adapted from previous literature [18]. These themes encompassed (1) pandemic intelligence, (2) personal preventive measures and mitigation, (3) social and common responsibility, (4) inquisitive messaging, (5) clarification, (6) events, campaigns and activities, (7) request for contributions, (8) showing gratitude, (9) reassurance, (10) risk groups, and (11) general advisories and vigilance. In addition, 6 emotional appeals adapted from previous studies [31,38,48,49] were included in the codebook. These emotional appeals included (1) fear appeals, (2) guilt appeals, (3) anger appeals, (4) hope appeals, (5) humor appeals, and (6) nurturance appeals. Each Facebook post, including all text and visual elements, and everything visible on the webpages were coded as a single unit of analysis.

Intercoder Reliability

We recruited 3 coders to code the posts and articles. Before conducting actual coding, the coders undertook 2 rounds of training, practice sessions for coding, intercoder reliability, and discussions to refine the codebook. During practice sessions, coders coded the same units of analysis to ensure a common understanding of the codebook. The units of analysis (n=60) for the training and practice sessions included the materials that had not been sampled. After achieving consensus, the coders coded 10% of the data, and intercoder reliability was tested. The process was repeated until we achieved an average Krippendorff α value of 0.78, ranging from 0.70 to 1.00. As it exceeded the 0.70 standard established in the literature [50], demonstrating an acceptable intercoder reliability. Subsequently, the data were split equally and coded by the coders.

Statistical Analyses

To answer RQ1 and RQ3, a series of descriptive statistics were conducted using SPSS (version 29; IBM Corp). For RQ2 and

XSL•FO
RenderX

RQ4, chi-square tests were performed to examine the relationships among CERC themes, emotional appeals, and COVID-19 phases. Notably, 24 website articles lacking publication dates were excluded from the chi-square tests as we could not classify them into any COVID-19 phases.

Results

Our sample showed that most of the messages about COVID-19 were communicated by Gov.sg (394/779, 50.6%), followed by the MOH (261/779, 33.5%), NEA (90/779, 11.5%), Ministry of Sustainability and Environment (18/779, 2.3%), and Health Promotion Board (16/779, 2.1%; Table 2).

RQ1 asked about the CERC message themes used by the Singaporean government during the COVID-19 pandemic. Our sample (Table 3) showed that most of the messages disseminated during the pandemic were about personal preventive measures and mitigation (522/779, 67%) followed by general advisories and vigilance (445/779, 57.1%); pandemic intelligence (266/779, 34.1%); social and common responsibility (131/779, 16.8%); risk groups (118/779, 15.1%); and event, campaigns, and activities (105/779, 13.5%). A small number of messages showed gratitude (54/779, 6.9%), inquisitive messaging (31/779, 4%), clarification (31/779, 4%), and reassurance (31/779, 4%). Request for contributions (5/779, 0.6%) was least communicated.

RQ2 asked how the CERC message themes changed across different phases during the COVID-19 pandemic. As shown in Table 4, the communication message themes changed across the COVID-19 phases. Chi-square tests revealed substantial changes in message themes across the phases, including pandemic intelligence (χ^2_3=18.1; P<.001). Specifically, messages on pandemic intelligence were more frequently posted during the maintenance stages—fighting a pandemic and rocky transition—compared with other phases (Figure 1 and Table 4). Similarly, the results showed that message themes such as personal preventive measures and mitigation (χ^2_3=29.1; P<.001); events, campaigns, and activities (χ^2_3=27.9; P<.001); and general advisories and vigilance (χ^2_3=15.5; P<.001) changed significantly across different COVID-19 phases. These message themes were frequently used in Singapore's online public health messaging during the fighting a pandemic phase and rocky transition phase (ie, maintenance stage).

Chi-square tests showed that message themes on social and common responsibility (χ^2_3=29.9; P<.001) and showing gratitude (χ^2_3=21.0; P<.001) changed across different COVID-19 phases. Messages on social and common responsibility were frequently communicated to the public during the fighting a pandemic period (ie, the maintenance stage), while messages that focused on expressing gratitude were often communicated during the early days of fog (ie, the initial stage) and fighting a pandemic period (ie, the maintenance stage). Message theme on risk groups (χ^2_3=17.7; P<.001) also changed across different COVID-19 phases; messages about risk groups were frequently mentioned during the rocky transition period (ie, the maintenance stage).

RQ3 asked about the types of emotional appeals used in the messages communicated by the Singaporean government to the public during the COVID-19 pandemic. Our data (Table 5) showed that hope (37/97, 38%) and humor (36/97, 37%) appeals were most frequently used in the communication messages during the COVID-19 pandemic, followed by nurturance appeals (17/97, 18%). Anger appeals (4/97, 4%), fear appeals (2/97, 2%), and guilt appeals (1/97, 1%) were used in the messaging strategies with a very low frequency.

RQ4 asked how the use of emotional appeals in messages communicated by the Singaporean government changed across different phases of the COVID-19 pandemic. Chi-square tests (Table 5) showed that emotional appeals—fear, anger, humor, and nurturance appeals—changed across phases. Messages containing fear appeals were only disseminated during the learning to live with COVID-19 period (χ^2_3=17.4; P<.001). Messages containing anger appeals were used during the fighting a pandemic period and learning to live with COVID-19 period (χ^2_3=8.4; P=.04). Humor appeals were used across all the phases of COVID-19 at different levels of frequency (χ^2_3=8.3; P=.04). Messages containing nurturance appeals were also mostly communicated to the public during the learning to live with COVID-19 period (χ^2_3=49.8; P<.001).

Table 2. Distribution of posts by health organizations (n=779).

Organization	Posts, n (%)
Gov.sg	394 (50.6)
Ministry of Health	261 (33.5)
Ministry of Sustainability and Environment	18 (2.3)
National Environmental Agency	90 (11.5)
Health Promotion Board	16 (2.1)

XSL•FO
RenderX

Table 3. Distribution of CERC[a] themes.

Themes	Posts (N=779), n (%)
Risk and crisis information	
Pandemic intelligence	266 (34.1)
Self-efficacy and sense-making	
Personal preventive measures and mitigation	522 (67)
Social and common responsibility	131 (16.8)
Inquisitive messaging	31 (4)
Clarification	31 (4)
Events, campaigns, and activities	105 (13.5)
Request for contributions	5 (0.6)
Showing gratitude	54 (6.9)
Reassurance	31 (4)
Advisories and alerts	
Risk groups	118 (15.1)
General advisories and vigilance	445 (57.1)

[a]CERC: crisis and emergency risk communication.

XSL•FO
RenderX

Table 4. CERC[a] themes across phases (n=755).

	Early days of fog (January, 2020, to March 2020)		Fighting a pandemic (April, 2020, to April, 2021)		Rocky transition (May, 2021, to November 2021)		Learning to live with COVID-19 (December, 2021, to present)		Total		Chi-square (*df*)[b]	*P* value
	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)		
Risk and crisis information												
Pandemic intelligence	32 (28.3)	81 (71.7)	76 (27.3)	202 (72.7)	124 (43.4)	162 (56.6)	27 (34.6)	51 (65.4)	259 (34.3)	496 (65.7)	18.1 (3)	<.001
Self-efficacy and sense-making												
Personal preventive measures and mitigation	62 (54.9)	51 (45.1)	161 (57.9)	117 (42.1)	217 (75.9)	69 (24.1)	58 (74.4)	20 (25.6)	498 (66)	257 (34)	29.1 (3)	<.001
Social and common responsibility	37 (32.7)	76 (67.3)	53 (19.1)	225 (80.9)	33 (11.5)	253 (88.5)	7 (9)	71 (91)	130 (17.2)	625 (82.78)	29.9 (3)	<.001
Inquisitive messaging	1 (0.9)	112 (99.1)	11 (4)	267 (96)	15 (5.2)	271 (94.8)	4 (5.1)	74 (94.9)	31 (4.1)	724 (95.9)	4.1 (3)	.25
Preparations and uncertainty reduction												
Clarification	7 (6.2)	106 (93.8)	4 (1.4)	274 (98.6)	15 (5.2)	271 (94.8)	4 (5.1)	74 (94.9)	30 (4)	725 (96)	7.6 (3)	.05
Events, campaigns, and activities	0 (0)	113 (100)	34 (12.2)	244 (87.8)	56 (19.6)	230 (80.4)	9 (11.5)	69 (88.5)	99 (13.11)	656 (86.9)	27.9 (3)	<.001
Request for contributions	1 (0.9)	112 (99.1)	4 (1.4)	274 (98.6)	0 (0)	286 (100)	0 (0)	78 (100)	5 (0.7)	750 (99.3)	5.0 (3)	.17
Showing gratitude	17 (15)	96 (85)	23 (8.3)	255 (91.7)	7 (2.4)	279 (97.6)	7 (9)	71 (91)	54 (7.2)	701 (92.8)	21.0 (3)	<.001
Reassurance	4 (3.5)	109 (96.5)	12 (4.3)	266 (95.7)	10 (3.5)	276 (96.5)	2 (2.6)	76 (97.4)	28 (3.7)	727 (96.3)	0.6 (3)	.89
Advisories and alerts												
Risk groups	5 (4.4)	108 (95.6)	34 (12.2)	244 (87.8)	55 (19.2)	231 (80.8)	16 (20.5)	62 (79.5)	110 (14.6)	645 (85.4)	17.7 (3)	<.001
General advisories and vigilance	56 (46.9)	57 (50.4)	140 (50.4)	138 (49.6)	180 (62.9)	106 (37.1)	53 (67.9)	25 (32.1)	429 (56.8)	326 (43.2)	15.4 (3)	<.001

[a]CERC: crisis and emergency risk communication.

[b]In total, 24 website articles were excluded from the chi-square tests as there were no publication dates.

XSL•FO
RenderX

Figure 1. Crisis and emergency risk communication (CERC) stages across phases.

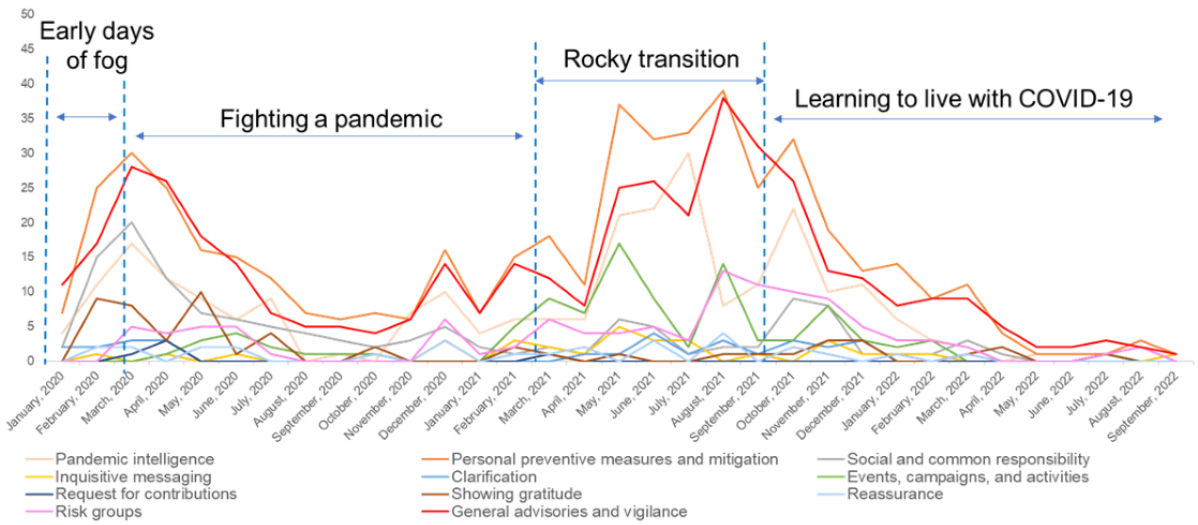

Table 5. Emotional appeals across COVID-19 phases (n=97).

	Early days of fog		Fighting a pandemic		Rocky transition		Learning to live with COVID-19		Total		Chi-square (*df*)	*P* value
	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)	Yes, n (%)	No, n (%)		
Fear appeals	0 (0)	113 (100)	0 (0)	278 (100)	0 (0)	286 (100)	2 (2.6)	76 (97.4)	2 (0.3)	753 (99.7)	17.4 (3)	<.001
Guilt appeals	0 (0)	113 (100)	1 (0.4)	277 (99.6)	0 (0)	286 (100)	0 (0)	78 (100)	1 (0.1)	754 (99.9)	1.7 (3)	.63
Anger appeals	0 (0)	113 (100)	2 (0.7)	276 (99.3)	0 (0)	286 (100)	2 (2.6)	76 (97.4)	4 (0.5)	751 (99.5)	8.4 (3)	.04
Hope appeals	4 (3.5)	109 (96.5)	16 (5.8)	262 (94.2)	11 (3.8)	275 (96.2)	6 (7.7)	72 (92.3)	37 (4.9)	718 (95.1)	2.8 (3)	.41
Humor appeals	7 (6.2)	106 (93.8)	20 (7.2)	258 (92.8)	7 (2.4)	279 (97.6)	2 (2.6)	76 (97.4)	36 (4.8)	719 (95.2)	8.3 (3)	.04
Nurturance appeals	0 (0)	113 (100)	1 (0.4)	277 (99.6)	5 (1.7)	281 (98.3)	10 (12.8)	68 (87.2)	16 (2.1)	739 (97.9)	49.8 (3)	<.001

Discussion

Principal Findings

This study examined public health communication strategies in Singapore during the COVID-19 pandemic by applying the CERC framework and emotional appeals. We found that the communication strategies used by the Singaporean public health institutions are aligned with the CERC framework. However, our analysis suggested that CERC message themes, such as inquisitive messaging and clarification, can be conveyed more frequently, particularly at the earliest stage of the crisis. This is in line with CERC recommendations; it also helps in verifying the abundance of information available when there is an infodemic. The COVID-19 phases in Singapore outlined by the government are also aligned with the CERC stages. We found that different emotional appeals were used at various COVID-19 phases in differing situations, which is evident in how nurturance appeals were used to encourage child vaccination, aligned with literature showing that nurturance appeals can effectively target parents. Despite this, certain emotional appeals can be used more frequently at various COVID-19 phases. We observed that Singapore's communication strategy is aligned with the frameworks of CERC and emotional appeals, with a few areas for improvement as discussed below.

Consistent with the study by Malik et al [18], the findings of this study revealed that Singapore-based public health

XSL•FO
RenderX

institutions' communication themes focused more on personal preventive measures and mitigation as well as general advisories and vigilance. For example, tele-befriending and telecounseling services, such as the Seniors Helpline, were established to help older citizens who face mental distress during the lock-down period. Overall, the Singapore government effectively communicated the message themes recommended by the CERC framework. This is evident from how the framework recommends informing the public about what they can do to protect themselves, the risks of the disease, and the actions that the public health institutions are taking to manage the situation. Meanwhile, the *request for contribution* theme was the one communicated the least, likely due to the Singapore-based public health agencies having sufficient resources to tide over the pandemic. To protect individuals and businesses in the country, the Singapore government had issued multiple budgets and grants since the onset of COVID-19. These monetary payouts include one-off as well as recurring cash grants for individuals whose livelihoods were affected by the pandemic [51]. Assistance was also offered to lower-income households. Examples of this include the COVID-19 Recovery Grant, which ensured that the citizens of Singapore or permanent residents could receive up to SG $700 (US $535) for 3 months if they faced an income loss of at least 50% [51]. The grants were successful in reducing the inequality in Singapore [52].

A shortcoming of the public health institutions' communication strategies was that messages on clarification was communicated less frequently. This was in line with the existing literature that shows how health care organizations may have insufficient posts addressing misinformation [18,53]. While steps were taken to clarify misinformation and address the public's questions, there can be more such messaging as COVID-19 was also an infodemic [1,54,55]. Infodemics occur when a large amount of information is rampant, including those that might be inaccurate or confusing [56]. Aligned with Reynolds and Seeger's [15] argument that communication during the initial phase should aim to reduce uncertainty, the Singapore-based public health institutions can enhance messaging on clarification and inquisitive messaging at the earliest stage of the crisis to prevent outrage and confusion in times of emergency. This is considering the fact that the health institutions would be communicating new information, in the form of pandemic intelligence and general advisories and vigilance, which might lead to increased uncertainty. Separately, the frequency of reassurance messaging can be increased, with the CERC framework encouraging such messaging to be conveyed during the initial and maintenance stages [14]. This can help to assure the public that the health institutions are handling the situation and managing the public's emotions in times of uncertainty [15].

We found that the communication message themes used by the public health institutions changed across different phases of COVID-19 in Singapore. This finding supported the CERC framework, which suggested that different message themes should be communicated to the public at different stages of a pandemic [26]. For example, we observed that messages on pandemic intelligence were communicated less frequently at the initial stage (ie, early days of fog: January, 2020, to March, 2020) of the COVID-19 pandemic; during this time, there was limited knowledge about the disease. As COVID-19 test kits became available, the Singaporean government could trace the number of cases on a daily basis and better understand the spread of the virus. This enabled them to learn and develop mitigation strategies to control the disease. Hence, there has been an increased focus on communicating messages on pandemic intelligence (eg, messages on the kick-off of COVID-19 vaccination) at the maintenance stage (ie, fighting a pandemic: April, 2020, to April, 2021; rocky transition: May, 2021, to November, 2021) than in other stages. Similarly, as scientists gradually gained more information about the virus, personal preventive measures and mitigation strategies were implemented by the public health institutions and more frequently communicated to the public at the maintenance stage (ie, fighting a pandemic: April, 2020, to April, 2021; rocky transition: May, 2021, to November, 2021). This is in line with CERC's recommendations to provide more explanations about preventive measures and mitigation strategies during the maintenance stage [20].

Our results showed that positive emotional appeals (eg, hope and humor appeals) were more frequently used in COVID-19 communication strategies. This is in line with the study by Xie et al [26], which found that positive emotions, such as hope, were commonly used in videos on COVID-19. They also posited that positive emotions can be beneficial to public engagement at the start of a pandemic to balance out the public's negative emotions. Hence, Singapore-based public health institutions may have taken this approach to neutralize the public's uncertainty. While other studies acknowledge that positive emotional appeals should be leveraged, they also suggested for negative emotional appeals to be used as both types of messages can engage the public in taking up preventive behaviors [53,57]. Positive emotional appeals, such as humor appeals, if overused or applied at inopportune times, can backfire, possibly lowering perceived risk and social responsibility; this may also result in the public not internalizing the intended message or not taking it seriously [58]. In addition, emotional appeals have different effectiveness for different demographics. For example, when compared with younger populations [59], older populations prefer emotional appeals that avoid negative emotional outcomes. Hence, health institutions can consider integrating a mix of emotional appeals for more effective messaging in future public health crises or pandemics.

This study found that the emotional appeals used varied with time, with their use being context-specific, depending on the situation and state of the disease. For example, nurturance appeals were not used at the early stage of the COVID-19 communication but were frequently used during the period of learning to live with COVID-19. This coincided with the first shipment of pediatric doses for the vaccination during the third week of December, 2021 [60], when the government started encouraging parents to bring their children for vaccination. Humor appeals were used with different frequencies across the stages, which could be due to the fluctuating severity of the crisis. Our studies revealed that humor appeals were used in less-pressing messages, such as encouraging the public to take up preventive behaviors, that were more culturally appropriate

XSL•FO
RenderX

especially during the stressful pandemic. For example, a sitcom character most Singapore residents are familiar with, Phua Chu Kang, was used in COVID-19 campaign videos that deal with responsible behavior during the pandemic, and later, to boost the local vaccination drive. While humor appeals were used in the communication messages across different stages of COVID-19 pandemic in Singapore, it is recommended that other countries should use the same strategy tactfully. This is because there are many factors, such as relevance and timeliness [61], that could influence the effectiveness of humor appeals. Hence, humor appeals need to be applied in good judgment to avoid unintended outcomes.

By contrast, fear and guilt appeals were less frequently applied in communication messages during the COVID-19 pandemic in Singapore. This demonstrates the Singapore-based health institutions' careful use of negative emotion appeals in a tense pandemic situation where most people were confined at home during the "circuit breaker" period. Such negative appeals could lead to higher mental stress and compromise social cohesion, if overused. This also explains why fear appeals were used in the later phase of the pandemic (ie, during the "learning to live with COVID-19" phase) when the situation was more relaxed, and most management measures had been eased. Hence, the public health authorities should consider the political and cultural landscape as well as the appropriate junctures when applying emotional appeals in their communication strategies in future.

Implications and Limitations

Theoretically, this study contributes to the existing literature on both the CERC model and emotional appeals. Apart from exploring how CERC model and emotional appeals were applied in Singapore's public health communication, this study is one of the few examining the relationship between CERC stages and the use of emotional appeals, especially in the context of COVID-19. This study provides insight on how to use a balanced mix of communication strategies for effective public health communications.

The practical implication of this study is twofold. First, in the local context, the findings of this study could inform Singaporean public health practitioners in developing more comprehensive messages during an emerging health crisis. Understanding how CERC message themes and emotional appeals were used in the public communication strategies during the COVID-19 pandemic could help the relevant authorities identify their strengths and shortcomings. For example, our finding on the lack of clarification messages is a pointer for public communication during the pandemic, especially during the period where misinformation about COVID-19 vaccination for children aged 5 to 11 years in Singapore was widespread [62]. Consequently, the local health authorities can learn from our findings to be better equipped to formulate communication strategies in handling unpredictable and emerging health pandemics in the future.

Second, for other nations, especially those with high population density, the health authorities can emulate Singapore's communication strategies during the COVID-19 pandemic to structure their communication strategies during the health crisis. In particular, Singapore's "all-of-government" approach, which involves the collaboration of various government agencies in communicating key messages during crises, is a useful communication strategy. Drawing from Singapore's approach, other countries could chart their responses in stages during a crisis and formulate timely public health messaging by incorporating CERC message themes together with the appropriate emotional appeal. However, as this study considers CERC message themes and phases and emotional appeals in the context of Singapore, the approach should be adapted with care—given the differences in local governance and culture of each country—because the messages may be received differently, thus affecting communication strategies. The "all-of-government" approach may also need to be tailored as a result.

This study has several limitations. First, this study did not collate Facebook posts and website articles from all the public health institutions and only focused on those that provided pressing information about COVID-19 that are applicable to all members of the public. We did not analyze content from government institutions with more targeted messaging because of the large volume of content for analysis. We also did not analyze other media sources, such as television, radio, newspapers, online news, and other social media content, beyond Facebook because of cross-posting of content. As this study might not provide a complete picture of COVID-19 messaging in Singapore, future research should examine social media posts by various government institutions. Second, website articles without publication dates were excluded from the analyses for RQ1 and RQ2, as we were unable to categorize the data into any of the COVID-19 phases in Singapore. Third, we did not analyze social media responses (ie, likes, shares, and comments) because such information was unavailable for website articles. Future research could examine social media responses for a greater understanding of CERC themes and emotional appeals in the context of COVID-19. Fourth, the findings of this study might not be generalizable to countries that are very different from Singapore because of the country's specific sociopolitical traits such as its high population density and strong central government. Nonetheless, given its exemplary management of COVID-19, it is worthy of documenting its practice to offer useful insights into future pandemic management. While other countries can learn from Singapore's approach, there may be a need to tailor the communication strategies according to their characteristics.

Fifth, this study did not specifically focus on messages containing severity and susceptibility because neither theme was encompassed in the CERC model used in this study. Given that severity and susceptibility are important aspects of risk perception, future research should examine these message themes in relation to the CERC model. In addition, this study did not examine the extent to which messaging conveyed acute risks from COVID-19 (eg, hospitalization and death) and chronic risks from COVID-19 (eg, postacute sequelae of COVID-19). Further studies should be conducted to delve into the differences as these may have impacted public willingness to engage in prevention and mitigation behaviors. Finally, while this study examined CERC themes and emotional appeals used across

XSL•FO
RenderX

CERC phases, we did not dive into the interaction between CERC themes and emotional appeals. This is a possible area for future studies.

Conclusion

This study examined public health messaging during COVID-19 pandemic in Singapore. The public health authorities in Singapore have taken a strategic and systematic approach in public health communication coupled with the use of emotional appeal to encourage the public to engage in protective behaviors.

Acknowledgments

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise program (grant number NRF2020-ITC003-0001).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. WHO director-general's opening remarks at the media briefing on COVID-19 - 11 March 2020. World Health Organization. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---9-march-2020 [accessed 2024-04-29]
2. Taking a whole-of-government approach. United Nations. 2012. URL: https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2012-Survey/Chapter-3-Taking-a-whole-of-government-approach.pdf [accessed 2024-04-29]
3. Speech by Mr Ong Ye Kung, minister for health, at The National Healthcare Group's population health collective (popcollect) annual work plan seminar for community partners and Gps. Ministry of Health. 2023. URL: https://tinyurl.com/3pjfk9tj [accessed 2023-06-14]
4. White paper on Singapore's response to COVID-19. Prime Minister's Office. URL: https://www.gov.sg/article/covid-19-white-paper [accessed 2023-06-14]
5. Howe S. Social media statistics in Singapore. Meltwater. URL: https://www.meltwater.com/en/blog/social-media-statistics-singapore [accessed 2024-05-30]
6. Woo JJ. Policy capacity and Singapore's response to the COVID-19 pandemic. Policy Soc 2020 Sep;39(3):345-362 [FREE Full text] [doi: 10.1080/14494035.2020.1783789] [Medline: 35039725]
7. An BY, Tang SY. Lessons from COVID-19 responses in east Asia: institutional infrastructure and enduring policy instruments. Am Rev Public Adm 2020 Jul 22;50(6-7):790-800. [doi: 10.1177/0275074020943707]
8. Geber S, Ho SS. Examining the cultural dimension of contact-tracing app adoption during the COVID-19 pandemic: a cross-country study in Singapore and Switzerland. Inf Commun Soc 2022 Jun 08;26(11):2229-2249. [doi: 10.1080/1369118x.2022.2082880]
9. Goh TJ, Ho SS. Public buying behaviors during the COVID-19 pandemic: the influence of attitude and perceived social norms from a presumed media influence perspective. Front Commun 2023 Jul 17;8:41. [doi: 10.3389/fcomm.2023.1109595]
10. Geber S, Ho SS, Ou M. Communication, social norms, and the intention to get vaccinated against COVID-19. Eur J Health Commun 2023 Jun 16;4(2):113-139. [doi: 10.47368/ejhc.2023.206]
11. Ma Z, Nan X, van Weert JC, Ou M, Ho SS. COVID-19. In: Ho EY, Bylund CL, van Weert JC, Basnyat I, Bol N, Dean M, editors. The International Encyclopedia of Health Communication. Hoboken, NJ: John Wiley & Sons; 2022:1-11.
12. Ho SS. The knowledge gap hypothesis in Singapore: the roles of socioeconomic status, mass media, and interpersonal discussion on public knowledge of the H1N1 Flu pandemic. Mass Commun Soc 2012 Sep;15(5):695-717. [doi: 10.1080/15205436.2011.616275]
13. Ho SS, Peh X, Soh VW. The cognitive mediation model: factors influencing public knowledge of the H1N1 pandemic and intention to take precautionary behaviors. J Health Commun 2013 Jul;18(7):773-794. [doi: 10.1080/10810730.2012.743624] [Medline: 23402299]
14. Reynolds B, W Seeger M. Crisis and emergency risk communication as an integrative model. J Health Commun 2005 Feb 23;10(1):43-55. [doi: 10.1080/10810730590904571] [Medline: 15764443]
15. Reynolds B, Seeger M. Crisis and emergency risk communication (manual): 2014 edition. Centers for Disease Control and Prevention. 2014. URL: https://emergency.cdc.gov/cerc/ppt/cerc_2014edition_Copy.pdfhttps://emergency.cdc.gov/cerc/ppt/cerc_2014edition_Copy.pdf [accessed 2024-04-29]
16. Tracking SARS-CoV-2 variants. World Health Organization. URL: https://www.who.int/activities/tracking-SARS-CoV-2-variants [accessed 2024-04-29]

XSL•FO
RenderX

17. Lwin MO, Lu J, Sheldenkar A, Schulz PJ. Strategic uses of Facebook in Zika outbreak communication: implications for the crisis and emergency risk communication model. Int J Environ Res Public Health 2018 Sep 10;15(9):1974 [FREE Full text] [doi: 10.3390/ijerph15091974] [Medline: 30201929]
18. Malik A, Khan ML, Quan-Haase A. Public health agencies outreach through Instagram during the COVID-19 pandemic: crisis and emergency risk communication perspective. Int J Disaster Risk Reduct 2021 Jul;61:102346 [FREE Full text] [doi: 10.1016/j.ijdrr.2021.102346] [Medline: 36337987]
19. Nagahawatta R, Kaluarachchi C, Warren M, Sedera D. Strategic use of social media in COVID-19 pandemic. Pandemic management by Sri Lankan leaders and health organisations using the CERC model. In: Proceedings of the 33rd Australasian Conference on Information Systems. 2022 Presented at: ACIS '22; December 4-7, 2022; Melbourne, Australia p. 1-11 URL: https://researchrepository.rmit.edu.au/esploro/outputs/conferenceProceeding/Strategic-Use-of-Social-Media-in/9922245555201341
20. Crisis and emergency risk communication (CERC) introduction. Centers for Disease Control and Prevention. 2018. URL: https://emergency.cdc.gov/cerc/ppt/CERC_Introduction.pdf [accessed 2024-04-29]
21. Reyes Bernard N, Basit A, Sofija E, Phung H, Lee J, Rutherford S, et al. Analysis of crisis communication by the prime minister of Australia during the COVID-19 pandemic. Int J Disaster Risk Reduct 2021 Aug;62:102375 [FREE Full text] [doi: 10.1016/j.ijdrr.2021.102375] [Medline: 36568696]
22. Vijaykumar S, Meurzec RW, Jayasundar K, Pagliari C, Fernandopulle Y. What's buzzing on your feed? Health authorities' use of Facebook to combat Zika in Singapore. J Am Med Inform Assoc 2017 Nov 01;24(6):1155-1159. [doi: 10.1093/jamia/ocx028] [Medline: 28449047]
23. Speech by Mr Gan Kim Yong, Minister for Health, at the 9th National Security Conference 2013, on 25th Oct 2013. Ministry of Health, Singapore. URL: https://www.moh.gov.sg/news-highlights/details/speech-by-mr-gan-kim-yong-minister-for-health-at-the-9th-national-security-conference-2013-on-25th-oct-2013 [accessed 2023-07-25]
24. Circuit breaker to minimise further spread of COVID-19. Ministry of Health, Singapore. 2020. URL: https://www.moh.gov.sg/news-highlights/details/circuit-breaker-to-minimise-further-spread-of-covid-19 [accessed 2024-04-29]
25. Meadows CW, Meadows CZ, Tang L, Liu W. Unraveling public health crises across stages: understanding Twitter emotions and message types during the California measles outbreak. Commun Stud 2019 Mar 24;70(4):453-469. [doi: 10.1080/10510974.2019.1582546]
26. Xie W, Damiano A, Jong CH. Emotional appeals and social support in organizational YouTube videos during COVID-19. Telemat Inform Rep 2022 Dec;8:100028. [doi: 10.1016/j.teler.2022.100028]
27. Bhatia TK. Emotions and language in advertising. World Eng 2019 Aug 28;38(3):435-449. [doi: 10.1111/weng.12420]
28. Chou WY, Budenz A. Considering emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence. Health Commun 2020 Dec;35(14):1718-1722. [doi: 10.1080/10410236.2020.1838096] [Medline: 33124475]
29. Lewis IM, Watson B, White KM, Tay R. Promoting public health messages: should we move beyond fear-evoking appeals in road safety? Qual Health Res 2007 Jan;17(1):61-74 [FREE Full text] [doi: 10.1177/1049732306296395] [Medline: 17170244]
30. Turner MM. Using emotional appeals in health messages. In: Cho H, editor. Health Communication Message Design: Theory and Practice. Thousand Oaks, CA: Sage Publications; 2012:59-69.
31. Mensa M, Vargas-Bianchi L. Nurtured and sorrowful: positive and negative emotional appeals in early COVID-19 themed brand communications. Commun Soc 2023 Jan 10;36(1):167-184. [doi: 10.15581/003.36.1.167-184]
32. Petersen MB, Christiansen LE, Bor A, Lindholt MF, Jørgensen F, Adler-Nissen R, et al. Communicate hope to motivate the public during the COVID-19 pandemic. Sci Rep 2022 Feb 15;12(1):2502 [FREE Full text] [doi: 10.1038/s41598-022-06316-2] [Medline: 35169174]
33. Pandemic fatigue- reinviforating the pulic to prevent COVID-19. World Health Organization. URL: https://apps.who.int/iris/bitstream/handle/10665/335820/WHO-EURO-2020-1160-40906-55390-eng.pdf?sequence=3&isAllowed=y [accessed 2024-04-29]
34. Buijzen M, Valkenburg PM. Developing a typology of humor in audiovisual media. Media Psychol 2004 May;6(2):147-167. [doi: 10.1207/s1532785xmep0602_2]
35. Strick M. Funny and meaningful: media messages that are humorous and moving provide optimal consolation in corona times. Humor 2021;34(2):155-176. [doi: 10.1515/humor-2021-0017]
36. Lee ST, Cheng IH. Assessing the TARES as an ethical model for antismoking ads. J Health Commun 2010 Jan 02;15(1):55-75. [doi: 10.1080/10810730903460542] [Medline: 20390977]
37. Xiao Y, Yu S. Using humor to promote social distancing on Tiktok during the COVID-19 pandemic. Front Psychol 2022;13:887744 [FREE Full text] [doi: 10.3389/fpsyg.2022.887744] [Medline: 35814113]
38. Bleakley A, Jordan AB, Hennessy M, Glanz K, Strasser A, Vaala S. Do emotional appeals in public service advertisements influence adolescents' intention to reduce consumption of sugar-sweetened beverages? J Health Commun 2015 Aug 09;20(8):938-948. [doi: 10.1080/10810730.2015.1018593] [Medline: 26054656]

XSL•FO
RenderX

39. Dillard JP, Plotnick CA, Godbold LC, Freimuth VS, Edgar T. The multiple affective outcomes of AIDS PSAs. Commun Res 2016 Jun 29;23(1):44-72. [doi: 10.1177/009365096023001002]
40. Witte K. Putting the fear back into fear appeals: the extended parallel process model. Commun Monogr 1992 Dec;59(4):329-349. [doi: 10.1080/03637759209376276]
41. Turner MM. Using emotion in risk communication: the Anger Activism model. Public Relations Review 2007 Jun;33(2):114-119. [doi: 10.1016/j.pubrev.2006.11.013]
42. O'Keefe DJ. Guilt and social influence. Ann Int Commun Assoc 2016 May 18;23(1):67-101. [doi: 10.1080/23808985.2000.11678970]
43. O'Keefe DJ. Guilt as a mechanism of persuasion. In: Dillard JP, Pfau M, editors. The Persuasion Handbook: Developments in Theory and Practice. Thousand Oaks, CA: Sage Publications; 2002.
44. Xu Z, Guo H. A meta-analysis of the effectiveness of guilt on health-related attitudes and intentions. Health Commun 2018 May 07;33(5):519-525. [doi: 10.1080/10410236.2017.1278633] [Medline: 28278614]
45. Coulter RH, Pinto MB. Guilt appeals in advertising: what are their effects? J Appl Psychol 1995 Dec;80(6):697-705. [doi: 10.1037/0021-9010.80.6.697] [Medline: 8557622]
46. Matkovic J, Clemens KS, Faasse K, Geers AL. Handwashing message type predicts behavioral intentions in the United States at the beginning of the global COVID-19 pandemic. Front Public Health 2021;9:583491 [FREE Full text] [doi: 10.3389/fpubh.2021.583491] [Medline: 34026698]
47. Mello S, Glowacki E, Fuentes I, Seabolt J. Communicating COVID-19 risk on Instagram: a content analysis of official public health messaging during the first year of the pandemic. J Health Commun 2023 Jan 02;28(1):38-52. [doi: 10.1080/10810730.2023.2175278] [Medline: 36760119]
48. Skubisz C, Miller A, Hinsberg L, Kaur S, Miller GA. Tips from former smokers: a content analysis of persuasive message features. Int Q Community Health Educ 2017 Feb 09;37(1):13-20. [doi: 10.1177/0272684x16685253]
49. Lazarus RS. Cognition and motivation in emotion. Am Psychol 1991 Apr;46(4):352-367. [doi: 10.1037//0003-066x.46.4.352] [Medline: 2048794]
50. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas 2007 Apr;1(1):77-89. [doi: 10.1080/19312450709336664]
51. Get the help you need. SupportGoWhere. URL: https://supportgowhere.life.gov.sg/schemes/CRG/covid-19-recovery-grant [accessed 2024-05-30]
52. Budget 2020 COVID-19 measures brought significant relief to workers, families, and businesses. Ministry of Finance. URL: https://www.mof.gov.sg/news-publications/press-releases [accessed 2024-05-30]
53. Guidry JP, Jin Y, Orr CA, Messner M, Meganck S. Ebola on Instagram and Twitter: how health organizations address the health crisis in their social media engagement. Public Relat Rev 2017 Sep;43(3):477-486. [doi: 10.1016/j.pubrev.2017.04.009]
54. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020 Dec 21;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
55. Yang K, Pierri F, Hui PM, Axelrod D, Torres-Lugo C, Bryden J, et al. The COVID-19 infodemic: Twitter versus Facebook. Big Data Soc 2021 May 05;8(1):205395172110138. [doi: 10.1177/20539517211013861]
56. WHO policy brief: COVID-19 infodemic management. World Health Organization. 2022. URL: https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-Policy_Brief-Infodemic-2022.1 [accessed 2023-06-14]
57. Nan X, Iles IA, Yang B, Ma Z. Public health messaging during the COVID-19 pandemic and beyond: lessons from communication science. Health Commun 2022 Jan 01;37(1):1-19. [doi: 10.1080/10410236.2021.1994910] [Medline: 34724838]
58. Miller E, Bergmeier HJ, Blewitt C, O'Connor A, Skouteris H. A systematic review of humour-based strategies for addressing public health priorities. Aust N Z J Public Health 2021 Dec;45(6):568-577 [FREE Full text] [doi: 10.1111/1753-6405.13142] [Medline: 34411385]
59. Williams P, Drolet A. Age‐related differences in responses to emotional advertisements. J Consum Res 2005 Dec;32(3):343-354. [doi: 10.1086/497545]
60. News highlights: first shipment of paediatric doses of Pfizer-biontech/comirnaty COVID-19 vaccine to arrive in Singapore. Ministry of Health, Singapore. URL: https://www.moh.gov.sg/news-highlights/details/first-shipment-of-paediatric-doses-of-pfizer-biontech-comirnaty-covid-19-vaccine-to-arrive-in-singapore [accessed 2024-01-26]
61. Tam WJ, Gobat N, Hemavathi D, Fisher D. A tool to guide creation of products for risk communications and community engagement (RCCE). Front Public Health 2022 May 27;10:810929 [FREE Full text] [doi: 10.3389/fpubh.2022.810929] [Medline: 35712288]
62. Leo L. Video containing 'misinformation' about COVID-19 vaccination exercise 'a great disservice to our children': MOH. Chanel News Asia. URL: https://www.channelnewsasia.com/singapore/moh-video-misinformation-children-covid-19-vaccination-exercise-social-media-2409181 [accessed 2024-01-25]

XSL•FO
RenderX

Abbreviations

CERC: crisis and emergency risk communication
MOH: Ministry of Health
NEA: National Environment Agency

Edited by A Mavragani; submitted 29.01.24; peer-reviewed by R Machin, S Mello; comments to author 26.04.24; revised version received 31.05.24; accepted 25.07.24; published 17.09.24.

Please cite as:
Ho SS, Chuah ASF, Ho VS, Rosenthal S, Kim HK, Soh SSH
Crisis and Emergency Risk Communication and Emotional Appeals in COVID-19 Public Health Messaging: Quantitative Content Analysis
J Med Internet Res 2024;26:e56854
URL: https://www.jmir.org/2024/1/e56854
doi:10.2196/56854
PMID:

©Shirley S Ho, Agnes S F Chuah, Vanessa S Ho, Sonny Rosenthal, Hye Kyung Kim, Shannon S H Soh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Media Use and COVID-19 Vaccination in Real-World Personal Networks: Quantitative Study

Iulian Oană[1,2], PhD; Marian-Gabriel Hâncean[1,2,3], PhD; Matjaž Perc[4,5,6,7], PhD; Jürgen Lerner[8], PhD; Bianca-Elena Mihăilă[1,2], MA; Marius Geantă[2], MD; José Luis Molina[9], PhD; Isabela Tincă[1,2], BA; Carolina Espina[10], MPH, PhD

[1]Department of Sociology, University of Bucharest, Bucharest, Romania

[2]Center for Innovation in Medicine, Bucharest, Romania

[3]The Research Institute of the University of Bucharest, University of Bucharest, Bucharest, Romania

[4]Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia

[5]Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, Slovenia

[6]Complexity Science Hub Vienna, Vienna, Austria

[7]Department of Physics, Kyung Hee University, Seoul, Republic of Korea

[8]Department of Computer and Information Science, University of Konstanz, Konstanz, Germany

[9]Research Group on Fundamental and Oriented Anthropology (GRAFO), Department of Social and Cultural Anthropology, Universitat Autònoma de Barcelona, Barcelona, Spain

[10]Environment and Lifestyle Epidemiology Branch, International Agency for Research on Cancer, Lyon, France

Corresponding Author:
Iulian Oană, PhD
Department of Sociology
University of Bucharest
Panduri 90-92
Bucharest, 050663
Romania
Phone: 40 305 97 30 ext 021
Email: iulian.oana@sas.unibuc.ro

Abstract

Background: Most studies assessing the impact of online media and social media use on COVID-19 vaccine hesitancy predominantly rely on survey data, which often fail to capture the clustering of health opinions and behaviors within real-world networks. In contrast, research using social network analysis aims to uncover the diverse communities and discourse themes related to vaccine support and hesitancy within social media platforms. Despite these advancements, there is a gap in the literature on how a person's social circle affects vaccine acceptance, wherein an important part of social influence stems from offline interactions.

Objective: We aimed to examine how online media consumption influences vaccination decisions within real-world social networks by analyzing unique quantitative network data collected from Romania, an Eastern European state and member of the European Union.

Methods: We conducted 83 face-to-face interviews with participants from a living lab in Lere ti, a small rural community in Romania, using a personal network analysis framework. This approach involved gathering data on both the respondents and individuals within their social circles (referred to as *alters*). After excluding cases with missing data, our analysis proceeded with 73% (61/83) of the complete personal networks. To examine the hierarchical structure of alters nested within ego networks, we used a mixed multilevel logistic regression model with random intercepts. The model aimed to predict vaccination status among alters, with the focal independent variable being the respondents' preferred source of health and prevention information. This variable was categorized into 3 types: traditional media, online media (including social media), and a combination of both, with traditional media as the reference category.

Results: In this study, we analyzed 61 personal networks, encompassing between 15 and 25 alters each, totaling 1280 alters with valid data across all variables of interest. Our primary findings indicate that alters within personal networks, whose respondents rely solely on online media for health information, exhibit lower vaccination rates (odds ratio [OR] 0.37, 95% CI 0.15-0.92; P=.03). Conversely, the transition from exclusive traditional media use to a combination of both traditional and online media

XSL•FO
RenderX

does not significantly impact vaccination rate odds (OR 0.75, 95% CI 0.32-1.78; *P*=.52). In addition, our analysis revealed that alters in personal networks of respondents who received the vaccine are more likely to have received the vaccine themselves (OR 3.75, 95% CI 1.79-7.85; *P*<.001).

Conclusions: Real-world networks combine diverse human interactions and attributes along with consequences on health opinions and behaviors. As individuals' vaccination status is influenced by how their social alters use online media and vaccination behavior, further insights are needed to create tailored communication campaigns and interventions regarding vaccination in areas with low levels of digital health literacy and vaccination rates, as Romania exposes.

(J Med Internet Res 2024;26:e58257) doi:10.2196/58257

KEYWORDS

vaccine hesitancy; online media; social media; assortative mixing; personal network analysis; social network analysis; Romania; vaccination; health information; COVID-19

Introduction

Background

Currently, COVID-19 is not as prominent on the public agenda despite the fact that infections with SARS-CoV-2 are still occurring. At the 76th World Health Assembly, held between May 21 and May 30, 2023, in Geneva, the chief of the World Health Organization (WHO) Tedros Adhanom Ghebreyesus stated that the threat of other pandemics, either related to mutations of SARS-CoV-2 or other pathogens, is still present [1]. Moreover, all the debates regarding vaccination during the COVID-19 pandemic affected how people approach other vaccines, such as routine childhood vaccines [2] and raised questions about the impact of human papillomavirus and hepatitis B virus vaccination [3]. Vaccine hesitancy has been a significant threat to public health, especially in high-income nations [4]. The Strategic Advisory Group of Experts on Immunization (SAGE) of the WHO defined vaccine hesitancy as the "delay in acceptance or refusal of vaccination despite the availability of vaccination services" [5]. The SAGE working group adds that vaccination hesitancy is context-specific, a complex of beliefs and behaviors conditional on culture, history, and vaccine type.

While vaccine hesitancy is not a new phenomenon [6,7], the landscape in which it exists today is new. With the advent of the internet, social media, and people interconnected in a global network, which further scales up the phenomenon's complexity, the WHO has stressed the importance of managing *infodemics* for efficient interventions against the COVID-19 pandemic [8]. The WHO defines an *infodemic* as "too much information, including false or misleading information in digital and physical environments during a disease outbreak" [9]. Working by misinformation or disinformation can lead to "confusion and risk-taking behaviors that can harm health. It also leads to mistrust in health authorities and undermines the public health response" [9]. Compared to traditional media (eg, official news channels, radio, or newspapers), online media, in general, and social media, in particular, facilitate the spread of unverified medical information [10,11] and the creation of echo chambers where false information, conspiracy theories, and fears are reinforced [12].

Evidence on how online media and social media use affects COVID-19 vaccination is mixed. For instance, extant research indicates that these media types are either positively [13-19] or negatively correlated with vaccine hesitancy [20,21]. However, most studies presented in systematic reviews [22,23] or systematic reviews of reviews [24] suggest that online and social media use tends to increase vaccination hesitancy. Similar conclusions were also drawn from analyses of cross-national surveys [13,19]. For studies that found a correlation between social media use and vaccination acceptance, it is generally observed that such results typically emerge from specific samples, such as adolescents [25], who are more inclined to seek information online or within unique cultural contexts, such as countries like China [21], where the interaction with social media content significantly differs [26].

Supplementary results are also provided by scholars who analyzed the effects of social media in comparison with traditional media. Some studies have found that the use of legacy media has a positive effect on vaccine acceptance, alongside the negative effect associated with social media [14,15,18]. In other studies, the focus was primarily on the correlation with traditional media, revealing that the use of social media had no significant effect [27,28]. Interestingly, a study found that not even traditional media as a whole category can correlate with vaccine acceptance. Viswanath et al [20] distinguished between the use of mainstream broadcast media and mainstream print media and found that only the latter positively affected the acceptance of the COVID-19 vaccine, the former being nonsignificant. In other research, no direct association was found between online media or social media use and vaccine hesitancy, or it was observed that the effects of social media vary due to indirect influences. Specifically, it was demonstrated that social media consumption can positively impact vaccine hesitancy when mediated by confidence in the vaccine's safety [27]. Lee and You [29] also emphasized the importance of indirect associations. Their study found that even if social media, in general, has a positive impact on vaccine hesitancy, this effect can be amplified by the respondents' perceived risk of vaccine-induced side effects or reversed through the respondents' perceived risk of being infected with SARS-CoV-2.

Given that most data on vaccination hesitancy during the COVID-19 pandemic were collected through standard surveys, statistical analyses have been conducted to assess the effects of media use on vaccine hesitancy. These analyses control for various sociodemographic traits (such as gender, income, education, and ethnic group) and other individual-level

XSL•FO
RenderX

characteristics, including trust in government and health experts and vaccine confidence. Nevertheless, it is crucial to recognize that information dissemination is not a one-way process from the media outlets to isolated individuals. People's opinions and attitudes are significantly shaped by their social circles, including family, peers, close friends, and other social groups they trust and interact with frequently. Opinions and behaviors regarding COVID-19 vaccination tend to be clustered within social groups rather than distributed randomly, as evidenced by studies conducted by Klaus et al [30] and Hâncean et al [31]. This observation underscores the necessity of considering real-world social clusters when examining the impact of media use on vaccination hesitancy. In line with this, the SAGE Working Group on Vaccine Hesitancy has acknowledged the significant role of social influence within the matrix of determinants for vaccine hesitancy [5].

Complementing studies that rely solely on standard survey data and focus on the attributes of individuals, network science introduces a more comprehensive approach by integrating attributes with relationships between units of analysis. Within this field, research on social actors falls into social network analysis (SNA) and personal network analysis (PNA), both of which conceptualize social actors as nodes and their relationships as ties. SNA concentrates on specific relationships within a defined population of nodes (eg, students within a school or users reposting a particular hashtag) in a bounded context, adopting a *sociocentric* design that encapsulates all nodes within a singular network. In contrast, PNA explores the wider social circles of individuals through an *egocentric* design, focusing on the networks that revolve around central nodes (or *egos*) and their connections to various social *alters*. These alters may not be part of the same group boundaries, leading to multiple, distinct networks that often cannot be aggregated into a single network [32].

Studies using SNA to examine COVID-19 vaccine hesitancy predominantly use datasets that capture user interactions within social media platforms. These studies typically concentrate on analyzing the content of discussions related to vaccine hesitancy and identifying key participants or groups within these conversations through community detection algorithms. This approach allows a nuanced understanding of the discourse dynamics and the social structures influencing vaccine hesitancy among online communities. Scholars found that while positive and negative discourses surrounding COVID-19 vaccination are present on social media platforms [26,33-36], they depend on political partisanship and the quality of sources [12,35]. Such studies are of great value for content analysis and mapping different communities inside online spaces. They identify online behavior patterns that, in the end, may affect day-to-day health decisions and give insights into how to create better communication campaigns for vaccine acceptance. Furthermore, they add nuance to the discussion of the relationship between the use of online platforms and vaccine hesitancy. Exposure to different sources in these online spaces (whom people choose to follow), the tendency to comment on posts presenting similar views [37], as well as prior biases (eg, general level of distrust [28], trust in government [29], or political partisanship [20]), stress the idea that health outcomes are contextual and contingent on how people engage with these online spaces and on what they bring to these spaces from offline influences.

However, research on the impact of real-world networks on COVID-19 vaccine hesitancy or acceptance [31] as well as on the extent to which these networks reflect online stances toward COVID-19 vaccination remains limited. This gap is significant given the potential for network structures and compositions to influence the transmission dynamics of SARS-CoV-2 [38-40]. Past studies have demonstrated that health outcomes are associated with *assortative mixing*. In many instances, assortative mixing, wherein nodes with similar traits form connections more frequently, serves as an indication of a network's structure and composition, emphasizing the tendency for similar individuals to be more closely interconnected. Drivers of assortative mixing can be represented by *social influence* (or *contagion*), where behaviors or traits spread through the network; *socialselection* (*homophily*), where individuals form connections based on similar characteristics; or *social context* (*confounding*), where external factors related to the environment or setting influence network formation by limiting with whom a node can create a tie [41,42]. Social contagion seeks to explain how nodal characteristics (behaviors, opinions, and other traits) change as a function of a node's relations inside the network [43]. Homophily principles state that nodes tend to create ties with others who are similar to them [44]. Contextual influences refer to effects brought by macrolevel changes (or the lack of macrolevel changes) in the social environment that influence individual or group behavior [41]. As examples of health outcomes related to the aforementioned mechanisms, we can name obesity [45] or the adoption of weight loss behaviors [46], the adoption of clean cooking methods [47], and how physicians adopt treatment plans and screening practices [48,49]. Regarding vaccination, previous studies have shown that network assortativity has a positive correlation with influenza vaccinations [50] and opinions about COVID-19 vaccination [31] and that network structures can facilitate, in general, one's acceptance or hesitancy through the influence of neighboring nodes [51].

Previous research has shown that a person's social environment can influence their health behavior, including COVID-19 vaccine acceptance. Some studies relied on measures of individual social capital to investigate vaccine acceptance. Various methods to account for someone's individual social capital include measures of interpersonal trust (social cohesion), civic participation (being a member of various organizations or clubs), and reciprocity (giving and receiving social support) [52-56]. The scholars using this approach found that higher levels of individual social capital are positively correlated with COVID-19 vaccine acceptance. Specifically, they found that persons with a higher individual social capital had a higher chance of being vaccinated or were willing to receive a booster shot [52], had a higher chance of wanting to receive vaccine [53], had a higher trust in COVID-19 vaccines [56], were more likely to volunteer for COVID-19 vaccine clinical trials [55], or received lower scores on various indices of vaccine hesitance [54,56]. Other studies investigated the relationship between vaccine hesitancy and one's social circle by using the concept of *homophily* through diverse measures accounting for the

XSL•FO
RenderX

association between respondents' vaccination acceptance and that of their peers. They found that someone's declared vaccination status is associated with the overall perceived uptake among their family and friends [57] or that the number of doses received by a person is positively correlated with the perceived number of doses among their social contacts [58]. Using experimental conditions, Leonhardt and Pezzuti [59] found that participants were more willing to receive vaccine when presented with cases of persons afflicted by COVID-19 who shared similar personality, behavioral, or ideological dimensions. Complementarily, Campbell et al [60] found that in the prospect of seeking a romantic partner, respondents' vaccination status was associated with that of their hypothetical partner.

Qualitative studies supplement such findings, describing instances where one's decision to receive vaccine against COVID-19 was influenced by seeing persons from their close social circle (family and friends) getting vaccinated [61] or hearing about their positive experience with the vaccine [62,63]. At the other end of the spectrum, scholars found through in-depth interviews that peer influences enacted in social groups can also act as a deterrent for vaccine acceptance [64].

The aforementioned approaches, which rely on standard surveys or qualitative interviews, bring important insights into the relationship between vaccine acceptance and peer influences. They support the general hypothesis referring to assortative mixing found in personal networks as a result of social contagion, social selection, or both. However, these studies use indicators that do not fully account for the features of a social circle. PNA brings the possibility of combining the compositional characteristics (eg, the proportion of contacts with a certain attribute) with structural features that are captured through alter-alter ties [32], for example, the density of a network indicating social cohesion or the fragmentation indicating the number of disconnected subgroups. Furthermore, it allows for expanding the assortativity measures from ego-alter ties (ie, the similarity between respondents and their contacts) to assortativity between alters on a given attribute. Using a PNA design, our study takes advantage of combining measures of structure and composition and analyzes them while controlling for their multilevel structure, with alters embedded in egos.

To investigate how consumption of online media influences vaccination decisions within real-world social networks, we used unique network data collected from Romania, a European Union (EU) member state situated in Eastern Europe. As a general context, Romania has the second lowest COVID-19 vaccination rate when compared to other EU countries. The latest data recorded by Our World in Data on June 10, 2022, showed that 41.27% of Romania's population was fully vaccinated [65]. Such low vaccination rates are also troubling in the context of low digital literacy rates. Data from Eurostat show that in 2023 Romania ranked last among the EU states in terms the proportion of persons with basic or above basic overall digital skills, with 31.09% of Romanians compared with the EU average of 60.55% among individuals who used the internet in the last 3 months [66]. Among EU states, in 2023, Romania was fourth in internet use for social media platforms, with 84.35% of Romanians compared with the EU average of 64.74% for individuals who used the internet in the last 3 months [67]. Internet use for seeking health information is low but rapidly increasing. In 2022, Romania ranked last in this EU ranking, with 33.82% of Romanians compared with the EU average of 57.74% for individuals who used the internet in the last 3 months [67]. In 2023, Romania was placed third to last in the EU ranking of seeking health information online, with 51.71% of Romanians compared with the EU average of 61.51% for individuals who used the internet in the last 3 months [67].

Proxy indicators on Romanians' digital health literacy show worrying facts about online media use. In 2023, the proportion of persons who checked the truthfulness of information from internet news sites or social media was the second lowest in the EU, with 11.24% of Romanians compared with the EU average of 26.38% for individuals who used the internet in the last 3 months [68]. The proportion of persons who checked the truthfulness of such information by checking the sources or comparing it with other information found on the internet further reinforces the ranking of Romania as second lowest in the EU, with 8.96% of Romanians compared with the EU average of 23.29% for individuals who used the internet in the last 3 months [68]. Eurostat figures about "evaluating data, information, and digital content" also indirectly show Romanians' ability to identify misinformation or disinformation. The percentage of individuals reporting having seen untrue or doubtful information on the internet is the lowest in Romania, with 32.83% of Romanians compared with the EU average of 53.67% for individuals who used the internet in the last 3 months [68]. Such low numbers raise the question of whether Romanians truly do not encounter untrue information when using online media or cannot discern them from what is presented as accurate information.

This Study

Our study focuses on the critical role of online media in influencing vaccination behaviors, using PNA on real-world social data. It specifically examines how the exclusive use of online media for health information correlates with the vaccination tendencies among individuals' social contacts in rural Romania. The investigation uncovers that within the personal networks of those who rely solely on online sources, there is a discernible decrease in vaccination rates. This finding sheds light on the dynamics prevalent in regions marked by low COVID-19 vaccination uptake, limited digital health literacy, and overall digital literacy challenges.

Our work stands out for its application of unique quantitative data, offering a fresh perspective on the interaction between digital information consumption and health behavior, particularly in the context of vaccine hesitancy. Hopefully, we make a significant contribution by mixing online media use and personal network influences on vaccination decisions, thereby filling a crucial gap in the existing literature.

Methods

Ethical Considerations

The research was performed in accordance with the Declaration of Helsinki. The research protocol was approved by a named institutional and licensing committee. Specifically, the Center

XSL•FO
RenderX

for Innovation in Medicine Ethics Committee reviewed and approved all the study procedures (EC-INOMED decision number D001/09-06-2023 and number D001/19-01-2024). All participants provided written informed consent. The privacy rights of the study participants were observed. The authors did not have access to information that could identify participants. After each interview, information that could identify the participants was anonymized. The study participants did not receive monetary compensation. Interviewees were rewarded for their participation with free access to a local educational program focused on health topics and a hotline number that they could use if they wanted expert opinions regarding medical problems or second opinions regarding a medical diagnostic.

Data Collection

The data presented in this study were collected during the pilot phase of the 4P-CAN project (personalized cancer primary prevention research through citizen participation and digitally-enabled social innovation; HORIZON-MISS-2022-CANCER-01, project ID 101104432, program HORIZON) [69]. We collected data from one of 4P-CAN's living lab in Lere ti, a rural locality in Arge county, Romania (N=4557).

The data collection process took place between September 13 and 30, 2023. We interviewed 83 persons following a PNA design [32]. PNA is a special type of network analysis; that is, the research design starts from nodes of interest, dubbed *egos*, and samples from their diverse social circles, composed of *alters*, through *name generators* (the questionnaire part wherein the respondents are asked to list their social contacts through various prompts that vary according to each research design in their formulation and number of elicited alters); *name interpreters* (the questionnaire sections wherein the egos give information about alters); and alter-alter ties (the questionnaire part where the respondents are asked to map various kinds of ties between elicited alters). In this manner, we can consider both the composition and structure of someone's personal network.

Study participants were recruited using a respondent-driven link-tracing sampling methodology [70,71]. At the time of the interview, every interviewee was aged ≥18 years. Our study was carried out in accordance with appropriate recommendations, relevant guidelines, and regulations (specifically, those provided by the Romanian Sociologists Society, ie, the professional association of Romanian sociologists [72]).

Before data collection, the community was informed about the 4P-CAN project through social media posts on Facebook (Meta Platforms, Inc) and press conferences held by the local media. Furthermore, we had meetings with local authorities, citizens from the community, and key actors (eg, teachers, community leaders, and local physicians).

According to our data collection strategy, first, we aimed to avoid a low response rate due to the length of the questionnaire. Second, we intended to delineate a panel of respondents for future data collection. Given these reasons, we preferred a network-oriented sampling method to a probabilistic nonnetwork strategy, using respondent-driven link-tracing sampling [70,71]. Specifically, we started with 6 individuals known as *seeds*. These persons were selected based on ethnographic fieldwork preceding the data collection process. We sought to have seeds who were different in their characteristics. The sociodemographic profile of the 6 seeds differed by *sex* (n=4, 67% male and n=2, 33% female), *age* (n=1, 17% 34 years old; n=1, 17% 50 years old; n=1, 17% 52 years old; n=1, 17% 55 years old; and n=2, 33% 64 years old), *education* (n=5, 83% with university degrees and n=1, 17% with high school diploma), *personal income* (n=3, 50% below the average net salary and n=3, 50% above the average net salary), and *employment sector* (n=3, 50% employed in the public sector; n=1, 17% employed in the private sector; n=1, 17% self-employed; and n=1, 17% retired). After responding to our questionnaire, we asked the seeds to recommend other persons who might be interested in participating in our study. Then, we asked the recommended persons (who agreed to participate) to further nominate other persons, and so on. This created referee-referral chains inside the link-tracing network. Of the 6 seeds, 2 (33%) refused to participate but still made recommendations. In the end, the response rate was 54.2% (83/153). We also report a selection bias, which was induced by the tendency of participants to recommend others who are similar to them [70,71]. Related to the attribute of being vaccinated against COVID-19, we analyzed 87 dyads in the link-tracing network, for which we had information on COVID-19 vaccination for both referees and referrals (ie, both were study participants). In 76% (66/87) of cases, vaccinated participants referred those who were also vaccinated. In the rest of the 24% (21/87) of dyads, we observed disassortative recommendations, with 20 (23%) dyads representing cases where vaccinated persons recommended unvaccinated individuals and 1 (1%) case with the reverse.

To keep complete transparency about the research activities and to facilitate citizens' engagement, before each interview, we discussed with the participants about the 4P-CAN project. Each participant received a paper dossier with informative materials about the project, including a consent form and a General Data Protection Regulation form, which they read and freely signed. They were also informed that they could opt out at any moment, that all data would be anonymized, and that they could consult the project's web page, with a dedicated section (in Romanian) containing details about the methodology [69]. The participants were provided the contact details (phone numbers and email addresses) of the project director and research team in case they had or have future questions.

Each participant responded to a questionnaire that collected information about their opinions, behaviors, and characteristics of individuals from their social circle. Each interviewee was asked to nominate up to 25 people with whom they interacted frequently (face-to-face or via other communication methods) and who were aged ≥18 years. The prompt that we used was as follows: "Please nominate 25 people (18 years old plus) you interact with (or meet). You can start with the people you interact with most often. These may be family members, friends, acquaintances, neighbors, work colleagues, etc." We did not collect separate data about face-to-face interactions and

XSL•FO
RenderX

interactions via other communication methods. A respondent (ego) can consider that they have frequent interaction with an alter if they meet monthly but, on the other hand, communicate almost daily via calls or SMS text messages. The nominated persons could be *family members*, *partners*,*close friends*, *casual friends*, or *acquaintances*. Afterward, the respondents were asked about certain characteristics of the people they nominated in their social circle. All interviews were conducted face to face, having an average duration of 80 (SD 27.7) minutes. The questionnaires were applied using Network Canvas (Complex Data Collective) [73], a PNA research software. Network analysis, in general, distinguishes between 2 data types: attribute and relational. Attribute data refer to characteristics measured individually for each person (*node*) within the network, whereas relational data describe properties of the relations between the nodes, also known as ties [74]. In the context of PNA research, the nodes are the egos and alters, and ties refer to ego-alter and alter-alter relations. For each study participant (*ego*), we collected attribute data about them (eg, age, marital status, various health behaviors, and opinions), attribute data about persons in their social circle (*alters*), relational data capturing ego-alter relationships, and relational data about alter-alter ties. It is important to note that all aforementioned data types were captured from the perspective of the egos.

Attribute Data

Of the attribute data, for both egos and alters, we collected information on their *sex* (1=female participant and 0=male participant), *age in years* (≥18 y), *being single* (1=yes and 0=no), and *COVID-19 vaccination status* (0=unvaccinated and 1=vaccinated). The education of egos and alters was measured by an ordinal scale capturing the last education level, with the values of 1=*no school*, 2=*less than primary school*, 3=*primary school*, 4=*less than secondary school*, 5=*secondary school*, 6=*arts and crafts school*, 7=*10 obligatory years*, 8=*high school (unfinished)*, 9=*high school (finished with diploma)*, 10=*posthigh school (nontertiary)*, 11=*bachelor's degree or equivalent level*, 12=*master's degree or equivalent level*, and 13=*Doctor of Philosophy or equivalent level*. *Employment status* (1=employed and 0=other) was measured only for egos. Additional information regarding egos' or alters' motives for getting vaccinated or their opinions about the COVID-19 vaccine or vaccination campaign were not collected.

We asked the egos (respondents) to provide details about their information sources related to medical topics. The egos had to respond to this question: "What are your sources of information about health and prevention (lifestyle)?" The respondents could select multiple answers from (1) "central television stations"; (2) "local press"; (3) "online, using Google (Google LLC), Bing (Microsoft Corp), or other search engines"; (4) "online, using social media (eg, Facebook, TikTok, and Instagram, etc)"; and (5) "online influencers." In our analyses, this was introduced as a factor variable distinguishing between those who used *only traditional media* (options *1*, *2*, or both), those who used *only online media* (*3*, *4*,*5*, or any combination of the 3 options), and those who used *both types of media*. Reliance on traditional media only was used as a reference level. While we acknowledge a difference between reliance on social media for news and information and other internet media, we chose to group the egos in the general category of *online media* because only 1 (1%) of the 83 initial respondents stated that they use only social media for information about health and prevention. Others who mentioned social media also mentioned search engines (and grouped as *online media* users) or legacy media (and grouped as *both media* users).

Network Data

Regarding ego-alter relations, we considered the *emotional closeness* between egos and alters ("How emotionally close do you feel to this person?" with answer options "not at all close," "not very close," "close," and "very close") and their *direct interaction frequency* ("How often do you typically meet with this person?" with answer options "less than once a year," "once a year," "a few times a year," "once a month," "every 2 weeks," "weekly," and "daily"). This question took into account only face-to-face meetings and did not include interactions via other media (eg, phone calls or instant messaging through various platforms). In both cases, the variables were binarized. For emotional closeness, 1 represents situations where the ego said they have a *very close* relation with the alter (0=*less than very close*). For direct interaction frequency, 1 represents a situation where egos and alters meet *at least twice a month* (0=*less than twice a month*). We used the product of the 2 binary variables to derive another binary variable, dubbed *ego-alter intensity*, where 1 represents the situation in which the ego feels very close to the alter, *and* they meet at least twice a month.

For the relations between alters, the participants were presented with the following question for each alter dyad: "Please tell me if these persons, even if they are related to each other, are acquaintances, casual friends, or close friends." From the alter-alter tie data, we derived *structural* network-level and node-level measurements after the ego was excluded from the personal network, its presence being redundant. All alter-alter ties were binarized, 1 representing the existence of a relation while 0 denoting situations where alters do not know each other.

Node-level structural measures are represented by alters' *betweenness centrality* and *vaccination assortativity score*. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between 2 other nodes [75]. Thus, betweenness centrality can also be viewed as the power to control the flow of information and opinions about vaccination inside the personal network [31]. The betweenness scores were normalized to account for differing network sizes.

The *vaccination assortativity score* combines attribute and relational data to account for subgroupings inside the personal network. It informs whether alters of similar traits (vaccinated or not) tend to share a social tie. This score was computed for each alter *i*, from the personal network of an ego *j*, taking the difference between the proportion of all of the *j*'s alters that are connected to *i* and vaccinated and the proportion of all of the *j*'s alters that are vaccinated. In computing the proportions, *i* was not included. For example, let us suppose that ego *j* has nominated 25 alters, of whom 14 (56%) are vaccinated against COVID-19 and 11 (44%) are not. In this personal network, alter *i*, who is vaccinated, is directly connected with 9 other alters, of which 6 (67%) are vaccinated and 3 (33%) are unvaccinated.

XSL•FO
RenderX

The assortativity score for alter *i* (which is excluded from the computation) is as follows:

6/9 – 13/24 = 0.666 – 0.542 = 0.142.

This indicates that the alter's *i* network neighbors have an above-average vaccination rate compared to the overall proportion of vaccinated alters in the personal network. Given that an alter's vaccination status is measured via a binary variable (1=vaccinated and 0=unvaccinated), a positive correlation coefficient indicates assortative mixing (ie, the alter's vaccination status is related to the vaccination status of its direct neighbors in the network). The computation method was previously advanced to detect assortative mixing of opinions related to COVID-19 vaccination [31] and processed food intake [76].

Network-level measurements describe personal networks as a whole, accounting for the distribution of alter-alter ties. Network *size* represents the number of nodes inside a network (in our case, the alters). *Density* sums up the proportion of present ties out of the theoretical number of ties (ie, if all nodes are connected). Density can also be interpreted as a measure of cohesion inside the network [77]. Finally, the *number of components* is an indicator of network fragmentation. An unfragmented network is a network with 1 component where all nodes are directly or indirectly connected to each other. If a network has $n \geq 2$ components, it indicates the number of n clusters of alters that are separated from each other. Furthermore, isolated nodes represent stand-alone components [75].

Statistical Analysis

Overview

Regarding the statistical modeling, the objective of our analyses was to assess the impact of egos' online information consumption in predicting alters' vaccination status: vaccinated or not. Given the nature of the dependent variable (binary) and the nested structure of the data, alters embedded in egos' networks, a mixed multilevel logistic regression model with random intercepts was used [78]. In this framework, alter-level data and ego-alter relations can be regarded as level 1 predictors (individual), while egos' attributes and network-level data are level 2 predictors (group) [79]. This type of nesting must be controlled for, as the data about egos and alters are reported by egos, for which unobserved covariates could influence the results [31]. Level 1 predictors include alters' sex, education level, age, relationship status (being single), betweenness centrality, relation intensity with ego, and vaccination assortativity. Level 2 predictors include egos' sex, education, age, employment status, and media use alongside network size, density, and number of strong components. To avoid scaling problems in the estimation of models, numeric variables were mean centered and standardized. Even if the education variable was measured via an ordinal scale, it was introduced in the statistical model as a numeric variable, given its large interval (from 1 to 13), and to let it have variation.

Data Exclusion

The initial sample comprised 83 respondents, with varying network sizes and missing data about alters' being vaccinated against COVID-19. To ensure that we have enough data at each level 2 group, we chose to focus our analyses only on those personal networks that have ≥15 alters for whom data about vaccination were present and who are not isolates. Not being an isolate (having ≥1 tie with another alter) is an important condition, given that isolated alters, for which the assortativity score could not be computed (as they have no network neighbors), were dropped from the analysis. After filtering the data according to these conditions, we ended up with a subset of 64 personal networks. Of the 64 personal networks, the regression models and some bivariate analyses presented in Multimedia Appendix 1 considered only 61 (95%) egos because data about media use for health and prevention were missing for 3 (5%) of them. Egos' relationship status was not introduced in the multilevel regression models, given the vast overrepresentation of those who were not single. Of the 61 egos, only 7 (11%) reported that they were single.

Results

Descriptive Statistics

Table 1 presents descriptive statistics for characteristics of egos, personal networks, and alters.

Most respondents (egos; 50/64, 78%) were vaccinated. They had an average age of approximately 53 years (SD 51.86), were almost equally split with respect to sex (33/64, 52% were female respondents), a large majority (56/64, 88%) were in a relationship (being married or as a couple), and a little more than half (36/64, 56%) were employed. The average value for education is 9.7 (SD 1.78), in between secondary (finished high school) and postsecondary (nontertiary) education on the ordinal scale. Distribution of responses for the type of media used for information about health and prevention shows that most egos used a combination of online and traditional media (29/64, 45%), 21 (33%) out of 64 used online media exclusively, and 11 (17%) used only traditional media as their information sources. The respondents' social contacts (alters) had a similar average age of approximately 53 (SD 16.06) years, similar average education (mean 9.35, SD 1.93), and approximately equally split on their sex (820/1561, 52.5% were female). Furthermore, similar to egos, the vast majority (1242/1561, 80%) of alters were in a relationship. Regarding their vaccination status, 64% (991/1561) of egos' social contacts were reported as being vaccinated against COVID-19. For relations between egos and their social contacts, we report that for 22% (347/1561; mean 0.22, SD 0.42) of ego-alter ties, the respondents mentioned that they were "very close" and met at least "twice a month" with the respective alter.

XSL•FO
RenderX

Table 1. Descriptive statistics for egos and alters.

Characteristic	Ego and network data (n=64)	Alter data (n=1561)
Age (y), mean (SD)	53.33 (15.86)	52.64 (16.06)
Education level		
Mean (SD)	9.70 (1.78; 0[a])	9.35 (1.93; 52[a])
Median (IQR)	10 (2; 0[a])	9 (2; 52[a])
Vaccinated against COVID-19, n (%)		
No	14 (21.88)	394 (25.24)
Yes	50 (78.12)	991 (63.48)
Missing data	0 (0)	176 (11.27)
Sex, n (%)		
Male	31 (48.44)	741 (47.47)
Female	33 (51.56)	820 (52.53)
Single, n (%)		
No	56 (87.5)	1242 (79.56)
Yes	8 (12.5)	318 (20.37)
Missing data	0 (0)	1 (0.06)
Employed, n (%)		
No	28 (43.75)	—[b]
Yes	36 (56.25)	—
Media used for health and prevention, n (%)		
Traditional media	11 (17.19)	—
Online media	21 (32.81)	—
Both media	29 (45.31)	—
Missing data	3 (4.69)	—
Ego-alter intensity, mean (SD)	—	0.22 (0.42)
Alter betweenness centrality, mean (SD)	—	0.04 (0.07)
Vaccination assortativity, mean (SD)	—	0.02 (0.11; 181[a])
Network size, mean (SD)	24.39 (1.70)	—
Network density, mean (SD)	0.65 (0.21)	—
Network components, mean (SD)	1.16 (0.51)	—

[a]Missing data (represents the number of cases for which the information is unknown).

[b]Not available.

Summary statistics for personal networks' structural features show that, on average, they are composed of 24 alters (56/64, 88% have ≥24 alters), while the rest have 15 (1/64, 2%), 20 (3/64, 5%), 22 (2/64, 3%), and 23 (2/64, 3%) alters, with an average density of 0.65 (SD 0.21) and 1.16 (SD 0.51) number of strong components, indicating high cohesiveness (density) and low fragmentation (components). A mean betweenness score of 0.04 (SD 0.07; scaled by the network's size) tells us that the alters had a low level of control over the potential information flow between the otherwise unconnected alters. The mean assortativity score for vaccination is 0.02 (SD 0.11), indicating that, on average, the proportion of an alter's direct network neighbors who were vaccinated is similar to the overall proportion inside the personal network (minus the alter of reference).

More detailed descriptive statistics regarding the distributions of presented variables are provided in Tables S3, S4, S7, and S13 in Multimedia Appendix 1. Bivariate tests for some of the presented variables are also reported in Multimedia Appendix 1 . First, looking at the alters' vaccination status contingent on egos' vaccination status, we observe that the frequency distributions are not independent. Vaccinated respondents tended to have more vaccinated social contacts in their personal networks (χ^2_1=76.2; P<.001; Table S15 in Multimedia Appendix 1). Second, the frequency distribution of the alters' vaccination status by egos' type of media used for health information also

XSL•FO
RenderX

indicates a nonindependent distribution (χ^2_2=28.2; *P*<.001; Table S16 in Multimedia Appendix 1). Egos using online media only present the highest proportion of unvaccinated alters (172/457, 37.6%), compared with egos using only traditional media (74/284, 29.8%) or both (140/616, 22.7%). However, it must be mentioned that there was no association between the egos' vaccination status and their preferred media for information about health and prevention distribution (χ^2_2=2.5, Fisher *P*=.32 Table S17 in Multimedia Appendix 1). Two-sided *t* tests of independent samples, comparing the average assortativity score grouped by alters' or egos' vaccination status, showed that in both cases, assortative mixing was lower for unvaccinated alters (t_{1378}=–2.3835; *P*=.02; Table S19 in Multimedia Appendix 1) and unvaccinated egos (t_{1378}=–3.5391; *P*<.001; Table S21 in Multimedia Appendix 1). Similarly, by performing a chi-square test comparing the distribution of alters' assortativity quartiles by egos' media use and vaccination status, we found that the distribution is nonindependent (χ^2_{15}=83.2; *P*<.001; Table S22 in Multimedia Appendix 1). Egos who were unvaccinated and used only online media present the highest proportion of alters in quartiles 1 (33/106, 31.1%) and 2 (43/106, 40.6%).

Statistical Modeling

In Table 2, we report the results of the full mixed multilevel logistic regression model, using alters' vaccination status as an outcome variable for 61 groups and 1280 alters (ie, cases without missing data on the variables included in the model).

Egos' relationship status (being single or not) was not included in the models due to its lack of variation. The intraclass correlation coefficient of 0.2 indicates that we have within-group homogeneity and intergroup heterogeneity. The grouping dependence is also confirmed by the marginal and conditional R^2 coefficients. The fixed effects explain 21.5% of the variance, while the random effects account for an extra 15.9% of the total variance. Regarding alters' attributes, education increases the probability of an alter being vaccinated (odds ratio [OR] 1.87, 95% CI 1.58-2.22; *P*<.001), while being single lowers it (OR 0.67, 95% CI 0.46-0.96; *P*=.03). Among egos' attributes as level 2 predictors, we observed that being in the personal network of an interviewee who was vaccinated increases an alter's chance of also being vaccinated (OR 3.75, 95% CI 1.79-7.85; *P*<.001). Network-level structural characteristics, such as size (OR 0.96, 95% CI 0.74-1.26; *P*=.79), density (OR 1.23, 95% CI 0.87-1.74; *P*=.24), and number of components (OR 1.04, 95% CI 0.77-1.41; *P*=.80), have no significant effects on the outcome variable. Between the node-level structural features (alter's betweenness centrality and vaccination assortativity), only the variable accounting for assortative mixing presents a positive effect (OR 1.17, 95% CI 1.01-1.35; *P*=.04). For every SD increase in the assortativity score, alters are 17% more likely to be vaccinated.

Egos' media use for health and prevention information is our variable of interest, using traditional media as the reference category. Results indicate a negative effect for switching from the exclusive use of legacy media to the exclusive use of online media (OR 0.37, 95% CI 0.15-0.92; *P*=.03). The change from "traditional only" to "both" types of media has no effect (*P*=.52). For comparison, Figure 1 presents a visual summary of multiple models, with ORs and 95% CIs. Model 2 represents a model where we included only predictors referring to egos and alters' attributes. Model 3 is a model where we included only predictors that consider relational data, such as ego-alter relation intensity, node-level structural features (betweenness centrality and vaccination assortativity), and network-level structural features. Table S23 in Multimedia Appendix 1 presents in detail the full model (model 4) in comparison with the null model (model 1) and models 2 (attributes only) and 3 (network only). The variable of interest, that is the ego's media use for health information, remained significant in model 2 (OR 0.38, 95% CI 0.16-0.92; *P*=.03) and model 4 (OR 0.37, 95% CI 0.15-0.92; *P*=.03), with similar odds levels while controlling for the other level 1 and 2 factors.

Besides the multilevel logistic regression models, we also performed a logistic general linear model with robust SEs (ie, SEs clustered by egos). Table S24 in Multimedia Appendix 1 presents the following results. Comparing the full model from Table 2 with the full model from Table S24 (general linear model 4), we observed similar results with regards to alters' education (OR 1.82, 95% CI 1.46-2.26; *P*<.001), egos being vaccinated (OR 1.34, 95% CI 1.06-1.68; *P*=.01), egos using online media for information about health (OR 0.66, 95% CI 0.44-0.97; *P*=.04), and alters' vaccination assortativity (OR 1.21, 95% CI 1.01-1.46; *P*=.04). To account for not having random ego intercepts, this model also included the proportion of vaccinated alters minus the alter of interest. Furthermore, for this predictor, we observed a positive and significant effect (OR 2.48, 95% CI 2.05-3.01; *P*<.001).

XSL•FO
RenderX

Table 2. Full mixed multilevel logistic regression model[a] predicting alters' vaccination status (n=1280 alters and 61 egos).

Predictors[b]	OR[c] (95% CI)	*P* value
Intercept	2.29 (0.85-6.15)	.10
Alter level (1)		
Sex		
Male	Reference	__[d]
Female	1.13 (0.84-1.53)	.41
Education	1.87 (1.58-2.22)	<.001
Single		
No	Reference	—
Yes	0.67 (0.46-0.96)	.03
Alter age	0.98 (0.83-1.16)	.83
Ego-alter intensity	0.92 (0.62-1.35)	.66
Alter betweenness	1.03 (0.88-1.20)	.71
Vaccination assortativity	1.17 (1.01-1.35)	.04
Ego level (2)		
Sex		
Male	Reference	—
Female	1.04 (0.57-1.91)	.89
Education	1.24 (0.87-1.75)	.24
Age	1.02 (0.73-1.42)	.92
Employed		
No	Reference	—
Yes	0.63 (0.30-1.34)	.23
Vaccinated		
No	Reference	—
Yes	3.75 (1.79-7.85)	<.001
Media type		
Traditional	Reference	—
Online	0.37 (0.15-0.92)	.03
Both	0.75 (0.32-1.78)	.52
Network level (2)		
Network size	0.96 (0.74-1.26)	.79
Network density	1.23 (0.87-1.75)	.24
Network components (strong)	1.04 (0.77-1.41)	.80

[a]Model indices: variance=0.83 (SD 0.913); intraclass correlation coefficient=0.2; observations=1280; groups=61; marginal R^2=0.215; and conditional R^2=0.374.

[b]With the exception of categorical variables and ego-alter intensity, all variables were mean-centered and scaled. The outcome variable remained in its original units.

[c]OR: odds ratio.

[d]Reference category.

XSL•FO
RenderX

Figure 1. Comparison of mixed multilevel logistic regression models. Contrasts between the attributes only model (model 2), network only model (model 3), and the full model (model 4).

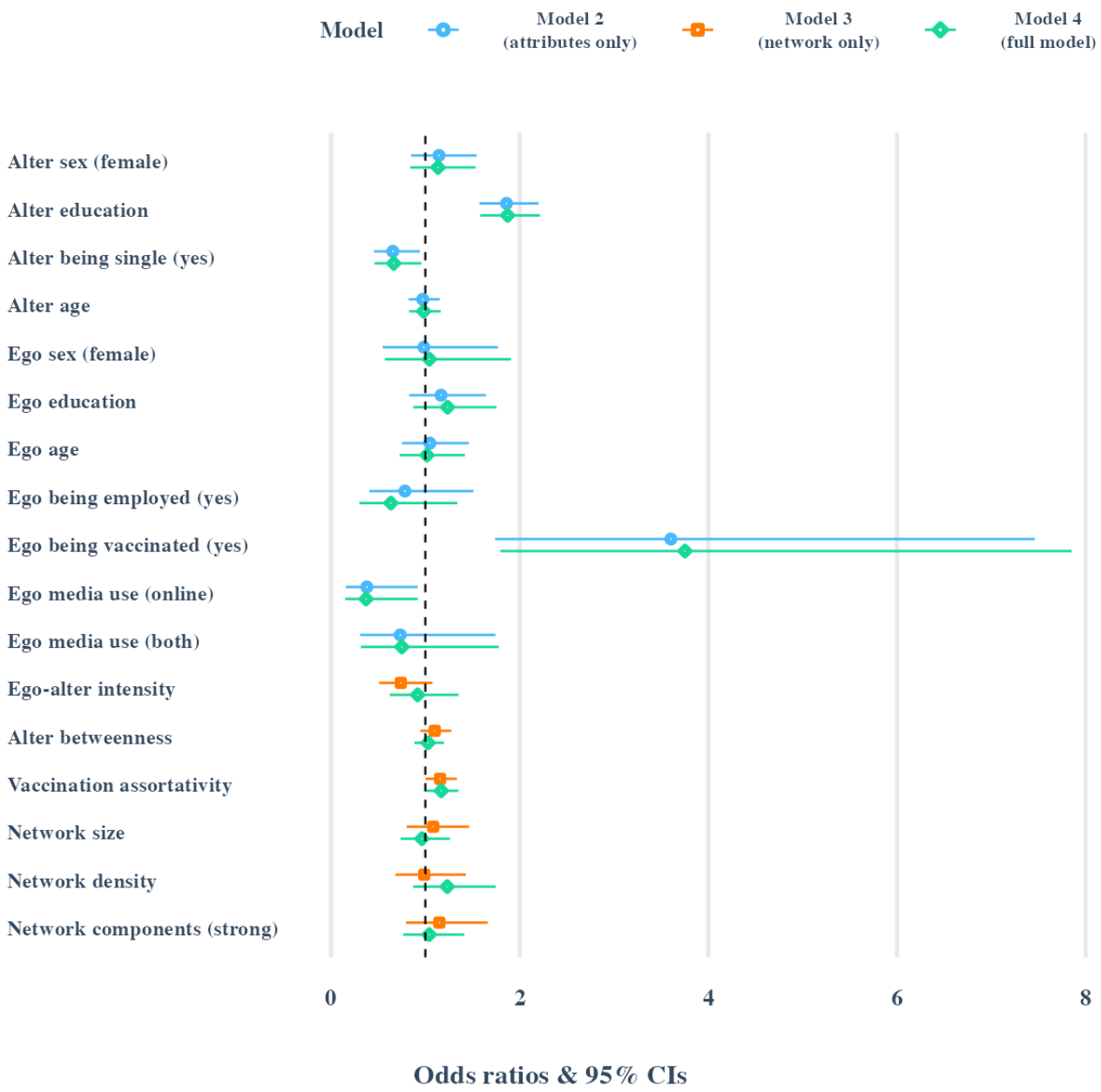

Discussion

Principal Findings

Our study investigated the impact of online media consumption on COVID-19 vaccination behaviors using a PNA approach. We aimed to explore the influence of real-world network structures and compositions on vaccine hesitancy, with a particular focus on the role of individuals' (egos') reliance on online media for health and prevention information as the primary predictor. Through the application of mixed multilevel logistic regression models, we assessed the vaccination status of individuals (alters) within 61 personal networks. Our findings indicate that alters within networks of egos who exclusively use online media for health and prevention information are less likely to be vaccinated compared to those in networks where egos depend on traditional media sources. In addition, for alters in networks where egos access both online and traditional media, the effect on vaccination status was not statistically significant. Our results are in line with previous studies that suggest social media use lowers COVID-19 vaccine acceptance [13,19]. The results of the mixed multilevel logistic regression are also supported by general linear models using robust SEs.

Moreover, our results support European-level statistics, which indirectly measure digital health literacy. In Romania, in 2023, the proportion of individuals who had at least basic digital skills was 31.09% [66]; however, the proportion of individuals who used internet for social media platforms was at 84.35% [67], and only 32.83% reported seeing, in online spaces, information or content which was untrue or doubtful (compared with the EU average of 53.67%) [68]. Secondary results from our analyses relate to some of the alters' sociodemographic characteristics and egos' vaccination status. Alters' level of

XSL•FO
RenderX

education and relationship status (being in a relationship) increased their odds of being vaccinated, as found in systematic reviews and meta-analyses [80,81]. Similarly, being in the personal network of an ego who is vaccinated increased the odds of alters being vaccinated, which is in line with another PNA study on a sample of Romanian respondents, testing for the assortative mixing of opinions on the COVID-19 vaccine and finding a positive association [31]. In this context, online misinformation and disinformation campaigns can find fertile ground and be enhanced by social influence in real-world networks when they mirror such opinions.

The ways individuals use the internet and the effects of information obtained through this medium, when it comes to health opinions and behaviors, are diverse. For example, Moon et al [82] found that while a higher frequency of social media use increased the odds of hesitancy, the higher frequency of internet use (for all purposes) decreased them. Allington et al [18] found that social media use loses its effect on vaccine hesitancy when controlling for trust in government, health professionals, and scientists and the COVID-19 pandemic is perceived as a risk. In our models, using both types of media did not differ from using only traditional media, presenting a nonsignificant effect in predicting alters' vaccination status. On one hand, this could indicate that exposure to multiple information sources can decrease hesitancy. On the other hand, this also indicates exogenous factors that are not considered when discussing online behavior. In our supplementary analyses, the media used by egos for health and prevention are not associated with their vaccination status. This could be due to the small sample size of 61 respondents.

An additional finding from our study highlights the impact of assortative mixing on vaccination outcomes. Specifically, alters who had a higher proportion of vaccinated direct network neighbors, relative to the vaccination rate of the entire network, were more likely to be vaccinated themselves. This finding aligns with other research using PNA to examine assortative mixing concerning attitudes toward COVID-19 vaccination [31] and processed food consumption [76]. Such alignment underlines the potential for further exploratory and comparative studies. Notably, our additional analyses reveal that networks characterized by lower vaccination assortativity tend to belong to unvaccinated egos, particularly those relying solely on online media for information and unvaccinated alters. This pattern suggests a fragmentation in personal networks concerning health beliefs and behaviors, possibly exacerbated by the diverse opinions and information sources about COVID-19 vaccines, echoing the WHO's concept of an "infodemic" [9]. The fact that the lower assortativity scores are found inside personal networks of unvaccinated egos who use only online media can be further hypothesized as an indicator of reinforcement of online misinformation through network contagion.

Given the limited evidence of the effects of real-world networks on vaccination hesitancy in relation to online media use, it remains an open question whether vaccine hesitancy is the result of assortative mixing (homogeneity) or social mixing (heterogeneity), when we look at networks' compositions and structures based on the nodes' media use. Building on the available literature about network influences on health behaviors and opinions [31,43,50,51], the working hypothesis would be that assortative mixing on multidimensional aspects of online behavior (eg, used platforms, followed accounts, and eHealth literacy) would positively correlate with assortative mixing on vaccination status. In our results, the similarity between egos' and their alters' vaccination status, coupled with the fact that the elicited alters were not randomly selected by respondents, can be an indicator of this association. The study of personal networks presents the possibility of the *false consensus effect* or *egocentric attribution bias*, wherein individuals attribute to others' opinions and behaviors similar to their own [83]. This also can add to our conclusions, as individuals can be influenced in their behavior not by what others do but by what they perceive as being the others' behaviors and opinions.

Other studies have found that a person's social relations influence their degree of COVID-19 vaccine hesitancy [52-64] on various aspects, including willingness to receive vaccine and trust or opinions about its efficacy and safeness. Such studies use various measures of individual social capital or assortative mixing (often dubbed as "homophily") through standard social surveys or qualitative interviews. Our study goes beyond the compositional aspect of a person's social circle and, by using a PNA design, also captures its structural features. In this way, characteristics such as network density and the number of strong components are controlled for in our models. Furthermore, it allowed us to go beyond the ego-alter relations captured by standard social surveys and look at the effect of alters' assortative mixing.

The generalization of our findings warrants careful consideration due to the distinct methodological and contextual facets of our study. The data, collected from a rural community in Romania (Eastern Europe) using a link-tracing sampling methodology, provide insights into the dynamics of online media influence on vaccination behaviors within personal networks. However, the rural setting and the specific sociocultural context of Eastern Europe might limit the direct applicability of our results to urban settings or other regions with different digital literacy levels and media consumption patterns. Nevertheless, it is important to underline that our findings likely reflect the situation in rural areas throughout Eastern Europe, thereby enhancing the inclusivity and relevance of our study within this broader regional context. Future research could benefit from applying similar methodological approaches in diverse settings to explore the universality and variability of these patterns across different social, cultural, and geographical contexts.

The implications raised by our study are not confined to COVID-19 vaccine hesitancy but vaccine hesitancy in general, in offline and online spaces, as real-world personal networks combine both. Further insights on how the information obtained by individuals from online spaces spreads and is reinforced through personal networks are needed, especially for countries with low general, digital, and eHealth literacy. Such countries are more vulnerable in the face of future pandemic disinformation campaigns regarding other vaccines and need communication campaigns and vaccination programs adapted to community-specific patterns.

XSL•FO
RenderX

Limitations

Our study has several limitations. Some of them are inherent to the PNA design itself. The information about alters' attributes and alter-alter relations is reported by the ego, which can lead to inaccuracies [84]. Therefore, the data generation process requires a careful interpretation of the results. To mitigate the impact of inaccuracies in ego-reported data, we limited the number of alters to 25, used communication or meeting frequency as a criterion for the name generator, and used the ego-alter relation intensity (emotional closeness and meeting frequency) as a control variable. If possible, future research should account for overlapping alters and use this information to correct contradictory accounts about alters' attributes; for example, see what the majority consensus about an alter's vaccination status is if that person is present in other personal networks.

Furthermore, we must consider challenges related to our primary variable of interest: the types of media used for health and prevention information. First, this variable was assessed solely at the ego level, because it can lead to potential inaccuracies when egos report on their alters' media use. Such inaccuracies could be more pronounced compared to reporting on demographic information (eg, age, education level, and marital status) or vaccination status against COVID-19. Second, our analysis does not differentiate between social media platforms and other online media sources, such as news aggregators (eg, Google News and Yahoo News). Prior research indicates divergent impacts on vaccine hesitancy and acceptance between these sources, with social media use often correlated with increased vaccine hesitancy, whereas broader internet use or reliance on online news platforms tends to be associated with higher vaccine acceptance [20,27,82]. The conflation of social media with other online media in our respondents' reports precludes a nuanced analysis of these distinct influences within our dataset.

Another limitation is the fact that we did not collect and analyze data on ego-alter interactions in various online spaces. Such data can bring further insights into the social selection or influence mechanisms that may occur in the online dimension of personal networks. Furthermore, we did not collect and analyze data of egos' or alters' motives and opinions regarding COVID-19 vaccination. Thus, our study did not differentiate between persons who were vaccinated because of noncoercive or coercive reasons (eg, receiving a vaccine against COVID-19 due to work-related obligations and falling in the "hesitant adopter" category on the vaccine hesitance spectrum [63]). Including such attribute data in future analyses can control for distinctions between personal networks, that is, differentiating between networks where alters' vaccination rate is low, yet some egos received the vaccine while other egos did not.

Furthermore, the link-tracing sampling method can introduce selection biases regarding respondents and individuals recommended for further participation. Potential selection biases can appear in various forms. Specific to our study, we can speculate that the nature of the 4P-CAN project, as a project about primary prevention, can attract more health-conscious individuals, skewing the results. This kind of selection effect was observed regarding COVID-19 vaccination, the general tendency being that vaccinated participants referred those who were also vaccinated. Future studies that use a network-oriented sampling method may implement a strategy of selecting seeds by considering both the vaccination status and their sociodemographic characteristics, which are associated with vaccine acceptance or hesitancy, to increase diversity. However, this does not necessarily guarantee the elimination of selection biases.

Further biases could have also been introduced by the incentives given for participation, such as free access to a local educational program focused on health topics and a hotline number that the participants can use if they want expert opinions regarding medical problems or second opinions regarding a medical diagnostic, without any other monetary incentives. However, every decision in terms of incentives, including nonmonetary, monetary, or none, has the potential to skew the sample with reference to the sociodemographic composition and retention inside the panel. Furthermore, the research project is applied inside a living lab, an open ecosystem entailing the community's coparticipation. Therefore, the purpose is to raise awareness of medical prevention and, at the same time, offer persons the health information and access that they lack in rural areas. Simultaneously, one of the objectives was to create a panel of respondents with a high retention rate during subsequent waves by combining the link-tracing referee-referral chains with an incentive structure that had the potential to attract persons who were either more health conscious or who sought medical advice.

The size of the sample, of 61 personal networks with valid data on all variables of interest, can also skew the results. Despite these limitations, the presence of assortative mixing regarding COVID-19 vaccine opinions was also identified in another study with a larger sample size of 443 Romanian respondents, using a different sampling strategy [31]. Furthermore, the general interpretation of our results aligns with studies that used nonnetwork designs and sampling strategies and found that varying degrees of vaccine hesitancy are influenced by a person's level of individual social capital [52-56] or perceived vaccination status of family and friends [57,58]. This consistency across studies suggests the robustness of our findings despite the methodological challenges.

Conclusions

Our study suggests that health and prevention information disseminated through online media, including social media, significantly impacts individuals' vaccination status within real-world personal networks. In addition to assortative mixing effects, our findings highlight the positive influence of education on vaccination status. We advocate for a deeper exploration of the composition and structure of personal networks, emphasizing the need to understand individuals' online behaviors and digital literacy levels for designing effective vaccination communication campaigns and interventions. Romania presents a case study of particular interest where low digital literacy and vaccination rates coincide with high social media use, underscoring the challenges of combating online misinformation and disinformation in similar contexts.

XSL•FO
RenderX

Acknowledgments

The authors would like to kindly thank Dr Cosmina Cioroboiu, Dr Bianca Cuco , Florin Găină, Simona-Elena Puncioiu, and Bogdan-Adrian Vidra cu for their valuable contributions to this study. IO, MGH, MG, BEM, and IT were supported by the 4P-CAN project, HORIZON-MISS-2022-CANCER-01, project ID 101104432, program HORIZON; MGH was supported by the Research Institute of the University of Bucharest (grant 8405/22.07.2023). JL was supported by Deutsche Forschungsgemeinschaft (DFG 321869138) and MP was supported by the Slovenian Research and Innovation Agency (Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije; grant P1-0403). The funding sources had no involvement in the study design; data collection, analysis, and interpretation; writing the paper; and deciding to submit the article for publication. The authors alone are responsible for the views expressed in this manuscript. Where authors are identified as personnel of the International Agency for Research on Cancer of the World Health Organization, the authors alone are responsible for the views expressed in this paper, and they do not necessarily represent the decisions, policies, or views of the International Agency for Research on Cancer of the World Health Organization.

Data Availability

The datasets generated during and analyzed during this study are available in the Zenodo repository [85].

Authors' Contributions

IO, MGH, JL, MP, JLM, and MG conceptualized and designed the study. IO, MGH, BEM, MG, and IT were involved in the investigation process. IO, MGH, BEM, and IT curated the data. JL designed the analysis framework and formulas. IO, MGH, JL, MP, JLM, MG, and CE drafted the manuscript. IO and MGH analyzed the data. JL, MP, and JLM validated the analyses. All authors approved the final version of the manuscript and accepted full responsibility for all aspects of the work described in this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed descriptive statistics, statistical models, and R code.
[PDF File (Adobe PDF File), 1981 KB - jmir_v26i1e58257_app1.pdf]

References

1. World must be ready to respond to next pandemic: WHO chief. United Nations. URL: https://tinyurl.com/2xuvpfsb [accessed 2024-01-15]
2. Grills LA, Wagner AL. The impact of the COVID-19 pandemic on parental vaccine hesitancy: a cross-sectional survey. Vaccine 2023 Sep 22;41(41):6127-6133 [FREE Full text] [doi: 10.1016/j.vaccine.2023.08.044] [Medline: 37659897]
3. Vraga EK, Brady SS, Gansen C, Khan EM, Bennis SL, Nones M, et al. A review of HPV and HBV vaccine hesitancy, intention, and uptake in the era of social media and COVID-19. Elife 2023 Aug 18;12:e85743 [FREE Full text] [doi: 10.7554/eLife.85743] [Medline: 37594016]
4. The Lancet Child Adolescent Health. Vaccine hesitancy: a generation at risk. Lancet Child Adolesc Health 2019 May;3(5):281. [doi: 10.1016/S2352-4642(19)30092-6] [Medline: 30981382]
5. MacDonald NE. Vaccine hesitancy: definition, scope and determinants. Vaccine 2015 Aug 14;33(34):4161-4164 [FREE Full text] [doi: 10.1016/j.vaccine.2015.04.036] [Medline: 25896383]
6. Callender D. Vaccine hesitancy: more than a movement. Hum Vaccin Immunother 2016 Sep;12(9):2464-2468 [FREE Full text] [doi: 10.1080/21645515.2016.1178434] [Medline: 27159558]
7. Porter D, Porter R. The politics of prevention: anti-vaccinationism and public health in nineteenth-century England. Med Hist 1988 Jul;32(3):231-252 [FREE Full text] [doi: 10.1017/s0025727300048225] [Medline: 3063903]
8. Munich security conference. World Health Organization (WHO). URL: https://tinyurl.com/3p4s3rj5 [accessed 2024-01-17]
9. Infodemic. World Health Organization (WHO). URL: https://www.who.int/health-topics/infodemic#tab=tab_1 [accessed 2024-01-17]
10. Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccin Immunother 2020 Jul 21;16(11):1-8. [doi: 10.1080/21645515.2020.1780846] [Medline: 32693678]
11. Zhao S, Hu S, Zhou X, Song S, Wang Q, Zheng H, et al. The prevalence, features, influencing factors, and solutions for COVID-19 vaccine misinformation: systematic review. JMIR Public Health Surveill 2023 Jan 11;9:e40201 [FREE Full text] [doi: 10.2196/40201] [Medline: 36469911]

XSL•FO
RenderX

12. Rathje S, He JK, Roozenbeek J, Van Bavel JJ, van der Linden S. Social media behavior is associated with vaccine hesitancy. PNAS Nexus 2022 Sep;1(4):pgac207 [FREE Full text] [doi: 10.1093/pnasnexus/pgac207] [Medline: 36714849]
13. Mascherini M, Nivakoski S. Social media use and vaccine hesitancy in the European Union. Vaccine 2022 Mar 25;40(14):2215-2225 [FREE Full text] [doi: 10.1016/j.vaccine.2022.02.059] [Medline: 35249775]
14. Ruiz JB, Bell RA. Predictors of intention to vaccinate against COVID-19: results of a nationwide survey. Vaccine 2021 Feb 12;39(7):1080-1086 [FREE Full text] [doi: 10.1016/j.vaccine.2021.01.010] [Medline: 33461833]
15. Allington D, McAndrew S, Moxham-Hall VL, Duffy B. Media usage predicts intention to be vaccinated against SARS-CoV-2 in the US and the UK. Vaccine 2021 Apr 28;39(18):2595-2603 [FREE Full text] [doi: 10.1016/j.vaccine.2021.02.054] [Medline: 33810905]
16. Tollison AC, LoPresti A. Media consumption and COVID-19 vaccination hesitancy: health literacy as a response. Commun Res Rep 2023 Oct 19;40(5):272-282. [doi: 10.1080/08824096.2023.2270904]
17. Carrieri V, Guthmuller S, Wübker A. Trust and COVID-19 vaccine hesitancy. Sci Rep 2023 Jun 07;13(1):9245 [FREE Full text] [doi: 10.1038/s41598-023-35974-z] [Medline: 37286569]
18. Allington D, McAndrew S, Moxham-Hall V, Duffy B. Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the COVID-19 pandemic. Psychol Med 2023 Jan;53(1):236-247 [FREE Full text] [doi: 10.1017/S0033291721001434] [Medline: 33843509]
19. Jennings W, Valgarðsson V, McKay L, Stoker G, Mello E, Baniamin HM. Trust and vaccine hesitancy during the COVID-19 pandemic: a cross-national analysis. Vaccine X 2023 Aug;14:100299 [FREE Full text] [doi: 10.1016/j.jvacx.2023.100299] [Medline: 37063307]
20. Viswanath K, Bekalu M, Dhawan D, Pinnamaneni R, Lang J, McLoud R. Individual and social determinants of COVID-19 vaccine uptake. BMC Public Health 2021 Apr 28;21(1):818 [FREE Full text] [doi: 10.1186/s12889-021-10862-1] [Medline: 33910558]
21. Ouyang H, Ma X, Wu X. The prevalence and determinants of COVID-19 vaccine hesitancy in the age of infodemic. Hum Vaccin Immunother 2022 Dec 31;18(1):2013694 [FREE Full text] [doi: 10.1080/21645515.2021.2013694] [Medline: 35172676]
22. Cascini F, Pantovic A, Al-Ajlouni YA, Failla G, Puleo V, Melnyk A, et al. Social media and attitudes towards a COVID-19 vaccination: a systematic review of the literature. EClinicalMedicine 2022 Jun;48:101454 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101454] [Medline: 35611343]
23. McKinley CJ, Limbu Y. Promoter or barrier? Assessing how social media predicts COVID-19 vaccine acceptance and hesitancy: A systematic review of primary series and booster vaccine investigations. Soc Sci Med 2024 Jan;340:116378. [doi: 10.1016/j.socscimed.2023.116378] [Medline: 38042027]
24. Kafadar AH, Tekeli GG, Jones KA, Stephan B, Dening T. Determinants for COVID-19 vaccine hesitancy in the general population: a systematic review of reviews. Z Gesundh Wiss 2022 Sep 19;31(11):1-17 [FREE Full text] [doi: 10.1007/s10389-022-01753-9] [Medline: 36160668]
25. Gewirtz-Meydan A, Mitchell K, Shlomo Y, Heller O, Grinstein-Weiss M. COVID-19 among youth in Israel: correlates of decisions to vaccinate and reasons for refusal. J Adolesc Health 2022 Mar;70(3):396-402 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.11.016] [Medline: 34952782]
26. Luo C, Chen A, Cui B, Liao W. Exploring public perceptions of the COVID-19 vaccine online from a cultural perspective: semantic network analysis of two social media platforms in the United States and China. Telemat Inform 2021 Dec;65:101712 [FREE Full text] [doi: 10.1016/j.tele.2021.101712] [Medline: 34887618]
27. Liu PL, Zhao X, Wan B. COVID-19 information exposure and vaccine hesitancy: the influence of trust in government and vaccine confidence. Psychol Health Med 2023 Jan 07;28(1):27-36. [doi: 10.1080/13548506.2021.2014910] [Medline: 34875950]
28. Tan M, Straughan PT, Cheong G. Information trust and COVID-19 vaccine hesitancy amongst middle-aged and older adults in Singapore: a latent class analysis Approach. Soc Sci Med 2022 Mar;296:114767 [FREE Full text] [doi: 10.1016/j.socscimed.2022.114767] [Medline: 35144226]
29. Lee M, You M. Direct and indirect associations of media use with COVID-19 vaccine hesitancy in South Korea: cross-sectional web-based survey. J Med Internet Res 2022 Jan 06;24(1):e32329 [FREE Full text] [doi: 10.2196/32329] [Medline: 34870605]
30. Klaus C, Wascher M, KhudaBukhsh WR, Tien JH, Rempała GA, Kenah E. Assortative mixing among vaccination groups and biased estimation of reproduction numbers. Lancet Infect Dis 2022 May;22(5):579-581 [FREE Full text] [doi: 10.1016/S1473-3099(22)00155-4] [Medline: 35460647]
31. Hâncean MG, Lerner J, Perc M, Molina JL, Geantă M. Assortative mixing of opinions about COVID-19 vaccination in personal networks. Sci Rep 2024 Feb 09;14(1):3385 [FREE Full text] [doi: 10.1038/s41598-024-53825-3] [Medline: 38336858]
32. McCarty C, Lubbers MJ, Vacca R, Molina JL. Conducting Personal Network Research: A Practical Guide. New York, NY: The Guilford Press; 2019.

XSL•FO
RenderX

33. Hwang J, Su MH, Jiang X, Lian R, Tveleneva A, Shah D. Vaccine discourse during the onset of the COVID-19 pandemic: topical structure and source patterns informing efforts to combat vaccine hesitancy. PLoS One 2022 Jul 27;17(7):e0271394 [FREE Full text] [doi: 10.1371/journal.pone.0271394] [Medline: 35895626]
34. Bonifazi G, Breve B, Cirillo S, Corradini E, Virgili L. Investigating the COVID-19 vaccine discussions on Twitter through a multilayer network-based approach. Inf Process Manag 2022 Nov;59(6):103095 [FREE Full text] [doi: 10.1016/j.ipm.2022.103095] [Medline: 36119754]
35. Muric G, Wu Y, Ferrara E. COVID-19 vaccine hesitancy on social media: building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. JMIR Public Health Surveill 2021 Nov 17;7(11):e30642 [FREE Full text] [doi: 10.2196/30642] [Medline: 34653016]
36. Wang D, Zhou Y, Ma F. Opinion leaders and structural hole spanners influencing echo chambers in discussions about COVID-19 vaccines on social media in China: network analysis. J Med Internet Res 2022 Nov 18;24(11):e40701 [FREE Full text] [doi: 10.2196/40701] [Medline: 36367965]
37. Li Y, Gee W, Jin K, Bond R. Examining homophily, language coordination, and analytical thinking in web-based conversations about vaccines on reddit: study using deep neural network language models and computer-assisted conversational analyses. J Med Internet Res 2023 Mar 23;25:e41882 [FREE Full text] [doi: 10.2196/41882] [Medline: 36951921]
38. Hâncean MG, Lerner J, Perc M, Ghiţă MC, Bunaciu D, Stoica AA, et al. The role of age in the spreading of COVID-19 across a social network in Bucharest. J Complex Netw 2021 Aug;9(4):cnab026 [FREE Full text] [doi: 10.1093/comnet/cnab026] [Medline: 34642603]
39. Hâncean MG, Lerner J, Perc M, Oană I, Bunaciu D, Stoica AA, et al. Occupations and their impact on the spreading of COVID-19 in urban communities. Sci Rep 2022 Aug 18;12(1):14115 [FREE Full text] [doi: 10.1038/s41598-022-18392-5] [Medline: 35982107]
40. Hâncean MG, Ghi ă MC, Perc M, Lerner J, Oană I, Mihăilă B, et al. Disaggregated data on age and sex for the first 250 days of the COVID-19 pandemic in Bucharest, Romania. Sci Data 2022 May 31;9(1):253 [FREE Full text] [doi: 10.1038/s41597-022-01374-7] [Medline: 35641521]
41. Cohen-Cole E, Fletcher JM. Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. J Health Econ 2008 Sep;27(5):1382-1387. [doi: 10.1016/j.jhealeco.2008.04.005] [Medline: 18571258]
42. Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc Natl Acad Sci U S A 2009 Dec 22;106(51):21544-21549 [FREE Full text] [doi: 10.1073/pnas.0908800106] [Medline: 20007780]
43. Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. Stat Med 2013 Feb 20;32(4):556-577 [FREE Full text] [doi: 10.1002/sim.5408] [Medline: 22711416]
44. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. Annu Rev Sociol 2001 Aug;27(1):415-444. [doi: 10.1146/annurev.soc.27.1.415]
45. Hill AL, Rand DG, Nowak MA, Christakis NA. Infectious disease modeling of social contagion in networks. PLoS Comput Biol 2010 Nov 04;6(11):e1000968 [FREE Full text] [doi: 10.1371/journal.pcbi.1000968] [Medline: 21079667]
46. Centola D. An experimental study of homophily in the adoption of health behavior. Science 2011 Dec 2;334(6060):1269-1272 [FREE Full text] [doi: 10.1126/science.1207055] [Medline: 22144624]
47. Kumar P, McCafferty L, Dhand A, Rao S, Díaz-Valdés A, Tabak RG, et al. Association of personal network attributes with clean cooking adoption in rural South India. Environ Res Lett 2021 Jun 16;16(6):064087. [doi: 10.1088/1748-9326/ac0746]
48. Pollack CE, Soulos PR, Herrin J, Xu X, Christakis NA, Forman HP, et al. The impact of social contagion on physician adoption of advanced imaging tests in breast cancer. J Natl Cancer Inst 2017 Aug 01;109(8):djw330 [FREE Full text] [doi: 10.1093/jnci/djw330] [Medline: 28376191]
49. Yu JB, Pollack CE, Herrin J, Soulos PR, Zhu W, Xu X, et al. Peer influence on physician use of shorter course external beam radiation therapy for patients with breast cancer. Pract Radiat Oncol 2020 Mar;10(2):75-83 [FREE Full text] [doi: 10.1016/j.prro.2019.11.001] [Medline: 31785370]
50. Barclay VC, Smieszek T, He J, Cao G, Rainey JJ, Gao H, et al. Positive network assortativity of influenza vaccination at a high school: implications for outbreak risk and herd immunity. PLoS One 2014 Feb 5;9(2):e87042 [FREE Full text] [doi: 10.1371/journal.pone.0087042] [Medline: 24505274]
51. Ni L, Chen Y, de Brujin O. Towards understanding socially influenced vaccination decision making: an integrated model of multiple criteria belief modelling and social network analysis. Eur J Oper Res 2021 Aug;293(1):276-289. [doi: 10.1016/j.ejor.2020.12.011]
52. Machida M, Kikuchi H, Kojima T, Nakamura I, Saito R, Nakaya T, et al. Individual-level social capital and COVID-19 vaccine hesitancy in Japan: a cross-sectional study. Hum Vaccin Immunother 2022 Nov 30;18(5):2086773 [FREE Full text] [doi: 10.1080/21645515.2022.2086773] [Medline: 35708308]
53. Grignon M, Bai Y. A cross-sectional analysis of the association between social capital and willingness to get COVID-19 vaccine in Ontario, Canada. Can J Public Health 2023 Apr 08;114(2):175-184 [FREE Full text] [doi: 10.17269/s41997-023-00746-9] [Medline: 36752981]

XSL•FO
RenderX

54. Morgan AK, Aziire MA, Cobbold J, Agbobada AA, Kudzawu SK. Hesitant or not: a cross-sectional study of socio-demographics, conspiracy theories, trust in public health information, social capital and vaccine hesitancy among older adults in Ghana. Hum Vaccin Immunother 2023 Dec 31;19(1):2211495 [FREE Full text] [doi: 10.1080/21645515.2023.2211495] [Medline: 37165968]
55. Terraneo M, Quaglia V, Nuvolati G, Bani M, Russo S, Strepparava MG, et al. Social capital and willingness to participate in COVID-19 vaccine trials: an Italian case-control study. BMC Public Health 2022 Nov 21;22(1):2137 [FREE Full text] [doi: 10.1186/s12889-022-14562-2] [Medline: 36411406]
56. Lee H, Noh EB, Kim JE, Oh J, Nam EW. Influencing factor of COVID-19 vaccination trust and hesitancy in Wonju city, South Korea. PLoS One 2022 Nov 14;17(11):e0277016 [FREE Full text] [doi: 10.1371/journal.pone.0277016] [Medline: 36374920]
57. Boëlle PY, Valdano E. The importance of increasing primary vaccinations against COVID-19 in Europe. Infect Dis Model 2024 Mar;9(1):1-9 [FREE Full text] [doi: 10.1016/j.idm.2023.11.008] [Medline: 38090610]
58. Are EB, Card KG, Colijn C. The role of vaccine status homophily in the COVID-19 pandemic: a cross-sectional survey with modelling. BMC Public Health 2024 Feb 14;24(1):472 [FREE Full text] [doi: 10.1186/s12889-024-17957-5] [Medline: 38355444]
59. Leonhardt JM, Pezzuti T. Vaccination acceptance across cultures: the roles of collectivism, empathy, and homophily. J Int Mark 2022 Mar 07;30(2):13-27. [doi: 10.1177/1069031x211073179]
60. Campbell J, Bennett-Brown M, S Marcotte A, M Kaufman E, Moscovici Z, R Adams O, et al. American singles' attitudes toward future romantic/sexual partners' COVID-19 vaccination status: evidence for both vigilance and indifference in a national sample. Sex Cult 2023 May 18;27(5):1-24 [FREE Full text] [doi: 10.1007/s12119-023-10097-9] [Medline: 37360016]
61. Ledford CJ, Cafferty LA, Moore JX, Roberts C, Whisenant EB, Garcia Rychtarikova A, et al. The dynamics of trust and communication in COVID-19 vaccine decision making: a qualitative inquiry. J Health Commun 2022 Jan 02;27(1):17-26. [doi: 10.1080/10810730.2022.2028943] [Medline: 35220915]
62. Lacy R, Puma J, Tubolino M, LaRocca D, Crane LA, Miller L, et al. Rural parents' attitudes and beliefs on the COVID-19 pediatric vaccine: an explanatory study. PLoS One 2022 Dec 7;17(12):e0278611 [FREE Full text] [doi: 10.1371/journal.pone.0278611] [Medline: 36477160]
63. Purvis RS, Moore R, Willis DE, Hallgren E, McElfish PA. Factors influencing COVID-19 vaccine decision-making among hesitant adopters in the United States. Hum Vaccin Immunother 2022 Nov 30;18(6):2114701 [FREE Full text] [doi: 10.1080/21645515.2022.2114701] [Medline: 36070518]
64. Morales GI, Lee S, Bradford A, De Camp A, Tandoc Jr EC. Exploring vaccine hesitancy determinants during the COVID-19 pandemic: an in-depth interview study. SSM Qual Res Health 2022 Dec;2:100045 [FREE Full text] [doi: 10.1016/j.ssmqr.2022.100045] [Medline: 35128519]
65. Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, et al. A global database of COVID-19 vaccinations. Nat Hum Behav 2021 Jul;5(7):947-953 [FREE Full text] [doi: 10.1038/s41562-021-01122-8] [Medline: 33972767]
66. Individuals- level of digital skills (from 2021 onwards). Eurostat. URL: https://tinyurl.com/2w83uxnv [accessed 2024-01-15]
67. Individuals - internet activities. Eurostat. URL: https://tinyurl.com/4s99cdd4 [accessed 2024-01-15]
68. Evaluating data, information and digital content (2021 onwards). Eurostat. 2023. URL: https://tinyurl.com/y8htrnws [accessed 2024-01-15]
69. Living labs. 4P-CAN. 2024. URL: https://4p-can.eu/living-labs/#romania [accessed 2024-01-17]
70. Hâncean MG, Lubbers MJ, Molina JL. Measuring transnational social fields through binational link-tracing sampling. PLoS One 2021 Jun 14;16(6):e0253042 [FREE Full text] [doi: 10.1371/journal.pone.0253042] [Medline: 34125853]
71. Molina JL, Lubbers MJ, Hâncean MG, Fradejas-García I. Short take: sampling from transnational social fields. Field Methods 2022 Jun 03;34(3):256-264. [doi: 10.1177/1525822x221105920]
72. SSR code of ethics. Romanian Sociologists Society. URL: https://societateasociologilor.ro/en/codul-de-etica-al-ssr/ [accessed 2024-04-29]
73. Complex Data Collective. Network canvas: interviewer. Zenodo. URL: https://zenodo.org/records/7350911 [accessed 2024-04-29]
74. Scott J. Social Network Analysis: A Handbook. London, UK: Sage Publications; 2000.
75. Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge, MA: Cambridge University Press; 1994.
76. Hâncean MG, Lerner J, Perc M, Molina JL, Geantă M, Oană I, et al. Processed food intake assortativity in the personal networks of East European older adults. medRxiv Preprint posted online July 10, 2024 [FREE Full text] [doi: 10.1101/2024.01.25.24301787]
77. Borgatti SP, Everett MG, Johnson JC. Analyzing Social Networks. Thousand Oaks, CA: Sage Publications; 2013.
78. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, MA: Cambridge University Press; 2007.
79. Vacca R. Multilevel models for personal networks: methods and applications. Ital J Appl Stat 2018;30(1):59-97. [doi: 10.26398/IJAS.0030-003]

XSL•FO
RenderX

80. Adu P, Poopola T, Medvedev ON, Collings S, Mbinta J, Aspin C, et al. Implications for COVID-19 vaccine uptake: a systematic review. J Infect Public Health 2023 Mar;16(3):441-466 [FREE Full text] [doi: 10.1016/j.jiph.2023.01.020] [Medline: 36738689]
81. Baghani M, Fathalizade F, Loghman AH, Samieefar N, Ghobadinezhad F, Rashedi R, et al. COVID-19 vaccine hesitancy worldwide and its associated factors: a systematic review and meta-analysis. Sci One Health 2023;2:100048 [FREE Full text] [doi: 10.1016/j.soh.2023.100048] [Medline: 39077035]
82. Moon I, Han J, Kim K. Determinants of COVID-19 vaccine hesitancy: 2020 California Health Interview Survey. Prev Med Rep 2023 Jun;33:102200 [FREE Full text] [doi: 10.1016/j.pmedr.2023.102200] [Medline: 37034299]
83. Ross L, Greene D, House P. The “false consensus effect”: an egocentric bias in social perception and attribution processes. J Exp Soc Psychol 1977 May;13(3):279-301. [doi: 10.1016/0022-1031(77)90049-X]
84. Wellman B. Challenges in collecting personal network data: the nature of personal network analysis. Field Methods 2007 May 01;19(2):111-115. [doi: 10.1177/1525822x06299133]
85. Oană I, Hâncean MG, Perc M, Lerner J, Mihăilă BE, Geantă M, et al. Replication data for: online media use and COVID-19 vaccination in real-world personal networks: quantitative study. Zenodo. 2024. URL: https://doi.org/10.5281/zenodo.10732216 [accessed 2024-04-29]

Abbreviations

EU: European Union
OR: odds ratio
PNA: personal network analysis
SAGE: Strategic Advisory Group of Experts on Immunization
SNA: social network analysis
WHO: World Health Organization

Edited by A Mavragani; submitted 11.03.24; peer-reviewed by Q Liao, KH Cheong; comments to author 17.07.24; revised version received 30.07.24; accepted 16.09.24; published 25.10.24.

Please cite as:
Oană I, Hâncean MG, Perc M, Lerner J, Mihăilă BE, Geantă M, Molina JL, Tincă I, Espina C
Online Media Use and COVID-19 Vaccination in Real-World Personal Networks: Quantitative Study
J Med Internet Res 2024;26:e58257
URL: https://www.jmir.org/2024/1/e58257
doi:10.2196/58257
PMID:

©Iulian Oană, Marian-Gabriel Hâncean, Matjaž Perc, Jürgen Lerner, Bianca-Elena Mihăilă, Marius Geantă, José Luis Molina, Isabela Tincă, Carolina Espina. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Chinese Public Attitudes and Opinions on Health Policies During Public Health Emergencies: Sentiment and Topic Analysis

Min Liu[1*], MM; Shuo Yuan[2*], MSN; Bingyan Li[3], MM; Yuxi Zhang[4], MM; Jia Liu[4], MM; Cuixia Guan[4], MM; Qingqing Chen[4], MM; Jiayi Ruan[4], MM; Lunfang Xie[4], PhD

[1]Department of Nursing, The First Affiliated Hospital of University of Science and Technology of China, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China

[2]Department of Cardiology II, Anhui No.2 Provincial People's Hospital, Hefei, China

[3]The First Affiliated Hospital of Anhui Medical University, Hefei, China

[4]School of Nursing, Anhui Medical University, Hefei, China

[*]these authors contributed equally

Corresponding Author:
Lunfang Xie, PhD
School of Nursing
Anhui Medical University
No. 81, Meishan Road, Shushan District
Hefei, 230031
China
Phone: 86 13856958216
Email: 527548725@qq.com

Abstract

Background: By the end of 2021, the new wave of COVID-19 sparked by the Omicron variant spread rapidly due to its highly contagious nature, affecting more than 170 countries worldwide. Nucleic acid testing became the gold standard for diagnosing novel coronavirus infections. As of July 2022, numerous cities and regions in China have implemented regular nucleic acid testing policies, which have had a significant impact on socioeconomics and people's lives. This policy has garnered widespread attention on social media platforms.

Objective: This study took the newly issued regular nucleic acid testing policy during the COVID-19 pandemic as an example to explore the sentiment responses and fluctuations of netizens toward new policies during public health emergencies. It aimed to propose strategies for managing public opinion on the internet and provide recommendations for policy making and public opinion control.

Methods: We collected blog posts related to nucleic acid testing on Weibo from April 1, 2022, to July 31, 2022. We used the topic modeling technique latent Dirichlet allocation (LDA) to identify the most common topics posted by users. We used Bidirectional Encoder Representations from Transformers (BERT) to calculate the sentiment score of each post. We used an autoregressive integrated moving average (ARIMA) model to examine the relationship between sentiment scores and changes over time. We compared the differences in sentiment scores across various topics, as well as the changes in sentiment before and after the announcement of the nucleic acid price reduction policy (May 22) and the lifting of the lockdown policy in Shanghai (June 1).

Results: We collected a total of 463,566 Weibo posts, with an average of 3799.72 (SD 1296.06) posts published daily. The LDA topic extraction identified 8 topics, with the most numerous being the Shanghai outbreak, nucleic acid testing price, and transportation. The average sentiment score of the posts was 0.64 (SD 0.31), indicating a predominance of positive sentiment. For all topics, posts with positive sentiment consistently outnumbered those with negative sentiment (χ^2_7=24,844.4, P<.001). The sentiment scores of posts related to "nucleic acid testing price" decreased after May 22 compared with before (t_{120}=3.882, P<.001). Similarly, the sentiment scores of posts related to the "Shanghai outbreak" decreased after June 1 compared with before (t_{120}=11.943, P<.001).

Conclusions: During public health emergencies, the topics of public concern were diverse. Public sentiment toward the regular nucleic acid testing policy was generally positive, but fluctuations occurred following the announcement of key policies. To understand the primary concerns of the public, the government needs to monitor social media posts by citizens. By promptly

XSL•FO
RenderX

sharing information on media platforms and engaging in effective communication, the government can bridge the information gap between the public and government agencies, fostering a positive public opinion environment.

(J Med Internet Res 2024;26:e58518) doi:10.2196/58518

KEYWORDS

public health emergencies; nucleic acid testing; governance strategies; sentiment analysis; LDA; social media; COVID-19; opinion analysis

Introduction

A public health emergency refers to an event that occurs suddenly and causes, or is likely to cause, serious harm to public health [1]. This includes major infectious disease outbreaks, mass unexplained illnesses, significant food and occupational poisoning, and other events that severely affect public health [1]. It is generally believed that public health emergencies are characterized by suddenness, group nature, diversity, high frequency, severe social harm, international interaction, and developmental stages [2]. With the acceleration of modernization, various uncertainties and unforeseeable risks continue to emerge, such as the COVID-19 pandemic and other public health emergencies [3]. Since the founding of the People's Republic of China, the COVID-19 pandemic has represented the fastest-spreading, widest-ranging, and most challenging public health emergency in terms of prevention and control [4]. By the end of 2021, the new wave of COVID-19 sparked by the Omicron variant spread rapidly due to its highly contagious nature, affecting more than 170 countries worldwide. According to data released by the National Health Commission of China, from February 26 to May 4, 2022, there were a total of 54,886 locally confirmed cases of COVID-19 in China [5]. This number exceeded the total number of confirmed cases nationwide from the outbreak in Wuhan until the end of 2021 [5]. On October 19, 2022, the Director-General of the World Health Organization stated that the COVID-19 pandemic continues to constitute an "international public health emergency of concern" [6].

In responding to public health emergencies, government policies play an important role in controlling the pandemic and protecting public health [7]. The regular nucleic acid testing policy was a routine and periodic measure implemented by the Chinese government to control the spread of the epidemic [8]. The frequency of nucleic acid testing was as often as once a week or even more frequently [8-10] to achieve the objective of early detection, early reporting, early isolation, and early treatment in order to block the transmission of the virus as much as possible [10]. However, the effectiveness of these policies depends not only on their scientific and rational basis but also on the public's acceptance and cooperation [11]. The regular nucleic acid testing had a significant impact on the economy, society, and people's lives, sparking widespread public debate. In September 2022, the number of monthly active users of Sina Weibo reached 584 million, with a daily average of 253 million active users [12]. As one of the most widely used and influential social media platforms, Weibo has become a primary arena for users to obtain information about sudden events and express personal opinions [13]. By analyzing citizens' views on the regular nucleic acid testing policy expressed on Weibo, the government can better understand public attitudes, evaluate and analyze policy effectiveness, and promptly adjust and optimize policy to enhance the scientific and effective management of emergencies [14].

Several studies used Weibo data to explore public sentiment changes during the COVID-19 pandemic, primarily investigating public opinions on the pandemic itself [15-17], COVID-19 vaccines [18-20], the Wuhan lockdown [21], social isolation measures [22], and rumors related to the pandemic [23,24]. However, there is a lack of research on public attitudes and opinions regarding the regular nucleic acid testing policy. Timely monitoring of public reactions to policies is crucial for the government to receive rapid feedback and make necessary policy adjustments during the pandemic [14]. Therefore, this study investigated the sentiment attitudes of the Chinese public during the first 3 months of the implementation of the regular nucleic acid testing policy and analyzed the changes in public attitudes over time and across different topics, as well as the reasons behind these changes. The expected results of this study could provide empirical evidence for policy makers to optimize policies and enhance public acceptance and implementation effectiveness, thereby contributing to the prevention and control of public opinion on social media. To our knowledge, this is the first study to explore the attitudes and opinions of the Chinese public toward regular nucleic acid testing during the COVID-19 pandemic, filling the gap in the existing literature on the attitudes of the Chinese public toward epidemic prevention policies during public health emergencies.

Methods

Data Collection and Preprocessing

In April 2022, Shenzhen was the first city in China to set up regular nucleic acid testing sites and implement regular nucleic acid testing [25]. By July 2022, most regions across the country had conducted regular nucleic acid testing [26]. Therefore, we used the keyword "nucleic acid testing" to retrieve original Weibo data posted between April 1, 2022, and July 31, 2022. We used Octopus software (Shenzhen Skieer Information and Technology Co Ltd) to obtain data such as blog posts, posting times, and publisher nicknames. After removing duplicate posts, commercial advertisements, and blank data, a total of 463,566 Weibo posts were obtained. Afterward, data cleaning was performed, including removing numbers, punctuation, and special symbols (including URLs, @usernames, spaces, etc), removing non-Chinese content, and converting traditional Chinese characters to simplified Chinese characters. Tokenization was performed by using the Python Jieba package. Because many words on Weibo were internet neologisms and

XSL•FO
RenderX

trending terms, and many were related to the COVID-19 pandemic, such as "health code," "green code," and even the main focus of this study, "nucleic acid testing," might not have been correctly segmented. Therefore, we created a custom lexicon to assist with segmentation [27]. The custom lexicon was divided into 2 parts. One part was the cell dictionary of Sogou's network hot words [28], and the other part consisted of annual hot words and popular phrases collected and deduplicated from various platforms in the first half of 2022, which were selected manually. The stop word list used included Chinese stop word lists, the Harbin Institute of Technology Stop Word List, Baidu stop word lists, and the stop word library from the Machine Intelligence Laboratory of Sichuan University [27]. More details are in Figure 1. At the same time, we obtained the daily number of newly confirmed cases from the official website of the Chinese Center for Disease Control and Prevention [29].

Figure 1. Data collection and processing workflow.

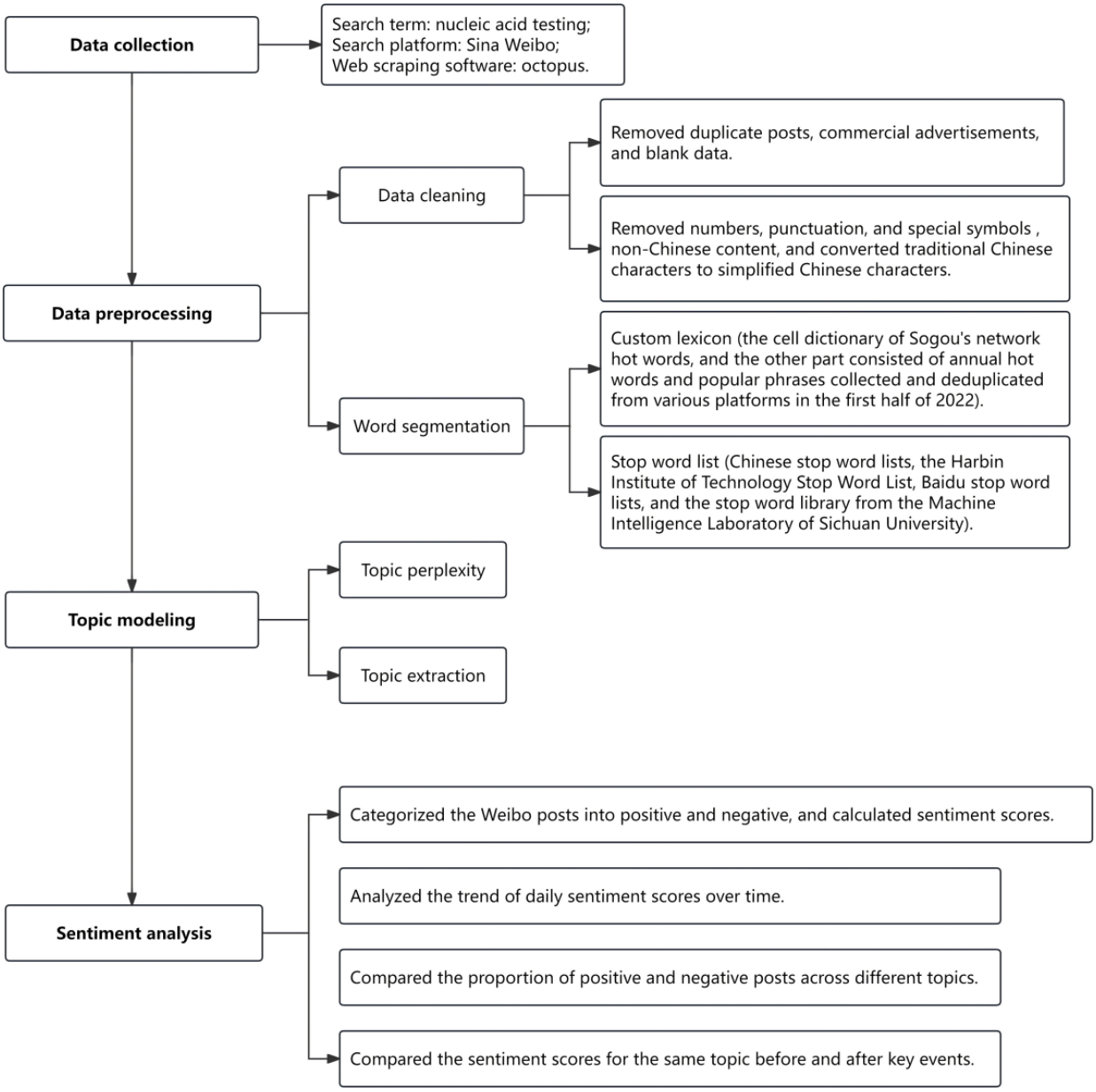

Topic Modeling

This study used the Latent Dirichlet Allocation (LDA) model for topic modeling. LDA is the most widely used topic model at the time and could be used to identify the most common topics across social media platforms [30]. LDA is an unsupervised learning algorithm that does not require prelabeled datasets, making it suitable for handling large and diverse text data such as Weibo posts [31]. It is a Bayesian probabilistic model with 3 layers: "document-topic-word," which is highly effective in identifying latent semantic topics in unstructured textual data, such as Weibo texts [32]. Furthermore, the LDA model does not have strict restrictions on the length of text; it is highly efficient in identifying topics in short texts [32]. Before constructing the LDA model, it is necessary to determine the number of topics [33]. Perplexity is a metric for determining whether the number of topics is reasonable, as the appropriate number of topics directly affects the accuracy of topic mining

XSL•FO
RenderX

[33]. As the number of topic words increases, the perplexity value first decreases to a minimum point and then gradually increases [33]. The number of topic words when the perplexity reaches its lowest point is the optimal number of topics.

Sentiment Analysis

We used a Bidirectional Encoder Representations from Transformers (BERT) to classify the sentiment of the data. The BERT uses a bidirectional transformer architecture, allowing it to understand each word in a sentence from both left to right and right to left simultaneously [34]. This bidirectionality enables BERT to capture the full contextual information, aiding in more accurately understanding complex sentence structures and sentiment expressions in Weibo posts [35]. In addition, Weibo often contains a large amount of unstructured text, emojis, acronyms, and other elements [36]. BERT, with its powerful language modeling and feature extraction capabilities, can effectively handle these characteristics, capturing subtle emotional changes and implicit sentiment information [34]. The model assigned a sentiment score between 0 and 1 to each Weibo post. As the sentiment score increased, positivity increased. Scores between 0 and 0.5 indicated negative sentiment, a score of 0.5 indicated neutrality, and scores between 0.5 and 1 indicated positive sentiment [37]. We tested the model's accuracy and ensured it was above 85%.

Statistical Analysis

All statistical analyses were conducted using IBM SPSS (version 24.0) and Python software (Python Software Foundation). We performed Spearman correlation analysis to explore the relationship between the daily number of posts with the number of newly confirmed cases per day. We used an autoregressive integrated moving average (ARIMA) model to examine the relationship between sentiment scores and changes over time. The chi-square test was used to compare the proportions of different types of sentiment posts under various topics. The *t* test was used to compare the sentiment scores at different times within the same topic. A 2-tailed test was conducted, and $P<.05$ was considered statistically significant.

Ethical Considerations

The public information involved in this study is freely available on Sina Weibo without a password. All data are de-identified. In addition, there is no identifiable information about individual users, IDs, or non-paraphrased posts in the manuscript or any supplementary materials. No individuals were recruited as study subjects and therefore no ethical approval was required for this study.

Results

Number of Posts

From April to July 2022, there were a total of 463,566 posts about nucleic acid testing on Weibo, with an average of 3799.72 (SD 1296.06) posts per day. The highest number of posts was 6689 on April 9th, while the lowest number of posts was 1551 on July 23rd. Figure 2 shows the changes in the daily number of posts on Weibo and the daily number of newly confirmed cases from April to July, indicating a positive correlation between the two ($\rho=0.656$, $P<.001$).

Figure 2. Daily changes in the number of nucleic acid tests and newly confirmed cases on Weibo.

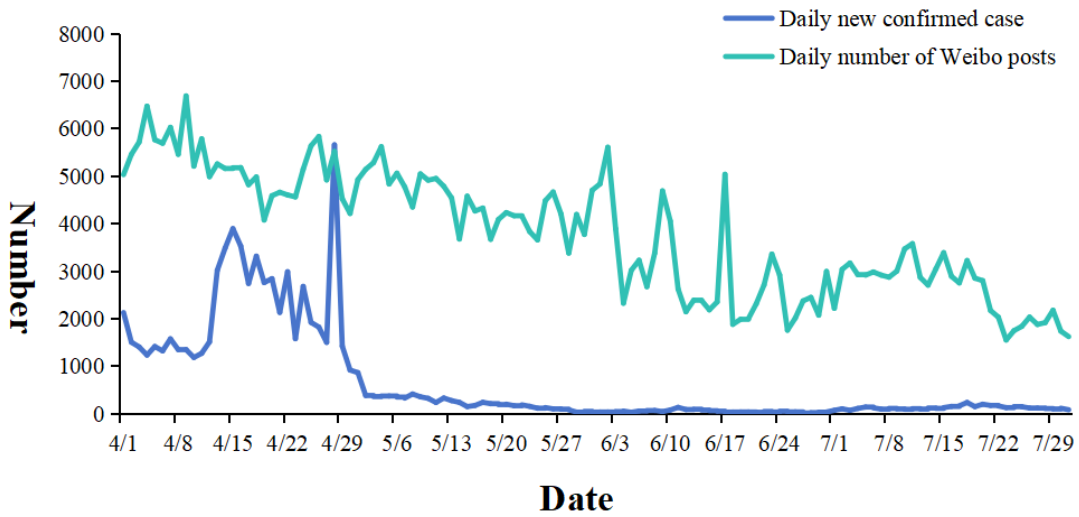

LDA Topic Extraction

Selection of Number of Topics

Figure 3 shows that the perplexity decreases as the number of topics in the LDA model increases. When the number of topics was K=8, the perplexity curve reached its lowest point and then rose before gradually decreasing again. Therefore, the number of topics selected in this study was K=8.

XSL•FO
RenderX

Figure 3. Perplexity of topics.

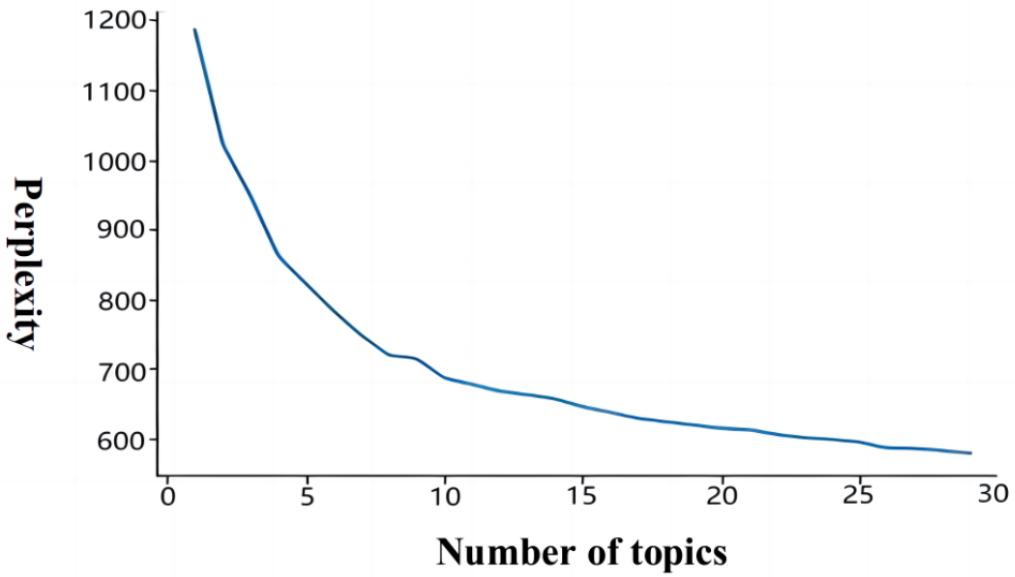

Topic Extraction

Table 1 shows the 8 topics extracted by LDA and the proportion of posts on different topics. The top 10 high-frequency words were selected as the keywords for each topic, and based on these high-frequency keywords, the topic names were summarized according to the original Weibo texts [15]. In terms of quantity, the most frequently discussed topic was "Shanghai outbreak," followed by "nucleic acid testing price" and "transportation." The least discussed topic was "hospital visits." Figure 4 describes the daily variation in the number of posts for each topic. From a temporal perspective, posts about the "Shanghai outbreak" peaked on April 9, while posts about "nucleic acid testing price" peaked on May 27.

Table 1. Latent Dirichlet allocation topic classification of Weibo nucleic acid testing–related posts.

Topic	Posts, n (%[a])	Keywords
Epidemic prevention and control measures	48,284 (10.42)	Work, Epidemic Prevention and Control, Community, Volunteers, Service, Guarantee, Residents, Doing Well, Masses, Supplies
Shanghai outbreak	195,612 (42.20)	Community, Shanghai, Nucleic Acid Testing, Video, Queuing, Epidemic, Nucleic Acid Testing Site, Tomorrow, Time, Hope
Hospital visits	2137 (0.46)	Hospital, Opening, Sampling Point, Waiting, Number of People, Fever Clinic, Tianjin, People's, Affiliated, Central Hospital
Transportation	51,101 (11.02)	Hours, Negative, Certificate, Health Code, Verification, Vehicle, Highway, Green Code, Travel Code, Travel
Compliance with epidemic prevention policies	31,033 (6.69)	Implementation, Measures, Masks, Cooperation, Community, Proactively, Reporting, Risk, Concentrated Isolation, Control
Nucleic acid testing price	55,502 (11.97)	Nucleic Acid Testing, Institutions, News, Beijing, Shanghai, Price, National, Laboratory, Press Conference, Normalized Nucleic Acid
Regular nucleic acid testing	47,157 (10.17)	Sampling, Area, Command Center, Notice, Community, Participation, Regular Nucleic Acid Testing, Staff, Cooperation, Arrangement
Epidemic notification	32,740 (7.06)	COVID-19 positive, Asymptomatic Carrier, Confirmed Cases, New Cases, Infected Individuals, Local Cases, Isolation, Diagnosis, Nucleic Acid Test Results, Report

[a]The percentages have been rounded to 2 decimal places for presentation clarity, leading to a total of 99.99%.

XSL•FO
RenderX

Figure 4. Changes in the number of posts on various topics over time.

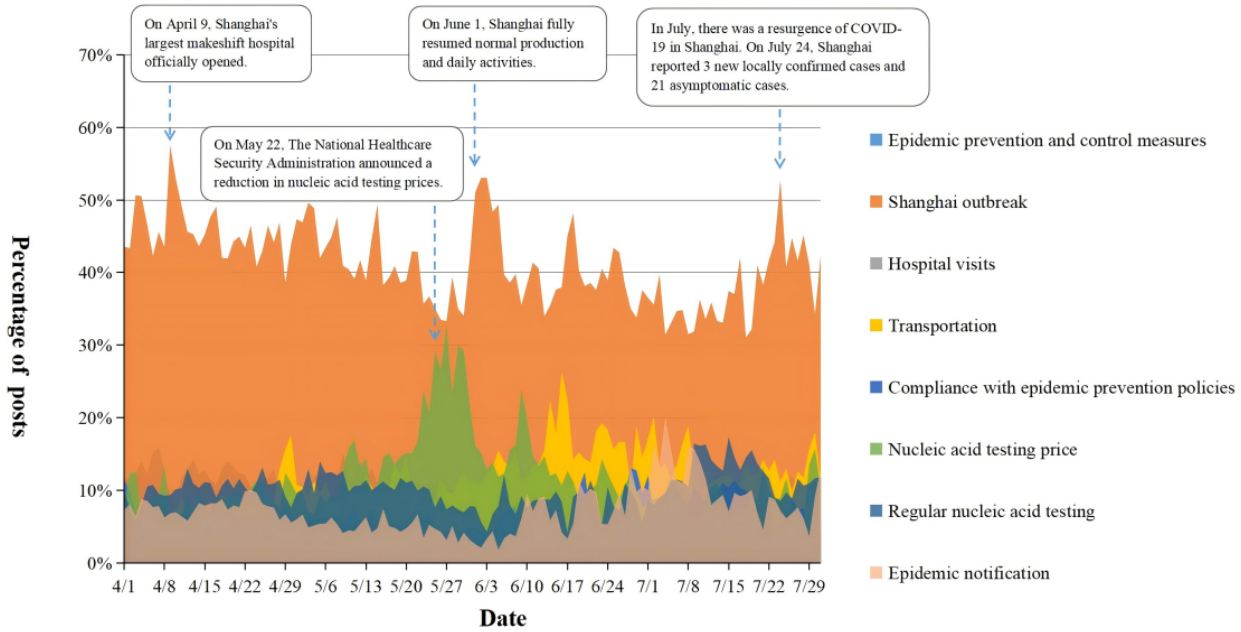

Sentiment Analysis

The Temporal Trend in Sentiment Score

In terms of sentiment polarity scores, the average sentiment score for posts about nucleic acid testing was 0.64 (SD 0.31), with daily sentiment scores fluctuating between 0.52 and 0.69. Since there were no posts with neutral sentiment, the posts in this study included only positive and negative sentiment. Among them, there were 337,968 posts with positive sentiment, accounting for 72.91% (337,968/463,566), and 125,598 posts with negative sentiment, accounting for 27.09% (125,598/463,566). Figure 5 shows the temporal changes in sentiment polarity of posts from April to July, with sentiment scores hitting low points on June 2 and June 17.

We conducted a stationarity test on the original time series and found it to be nonstationary. We performed first-order differencing and used auto-correlation function (ACF) and partial auto-correlation function (PACF) plots to determine that d=1, p=4, and q=3. Subsequently, we conducted a white noise test on the residuals. The auto-correlation function plot of the results showed no significant auto-correlation, indicating no temporal auto-correlation, as shown in Figure 6. We used the ARIMA (4, 1, 3) model for our analysis, with the model fitting metrics provided in Multimedia Appendix 1. Based on the model, we predicted the sentiment trend for August 2022. As shown in Figure 7, there were some fluctuations in Weibo sentiment in August, but overall, positive sentiment remained dominant.

XSL•FO
RenderX

Figure 5. The dynamic changes in daily sentiment scores, the number of positive posts, and the number of negative posts.

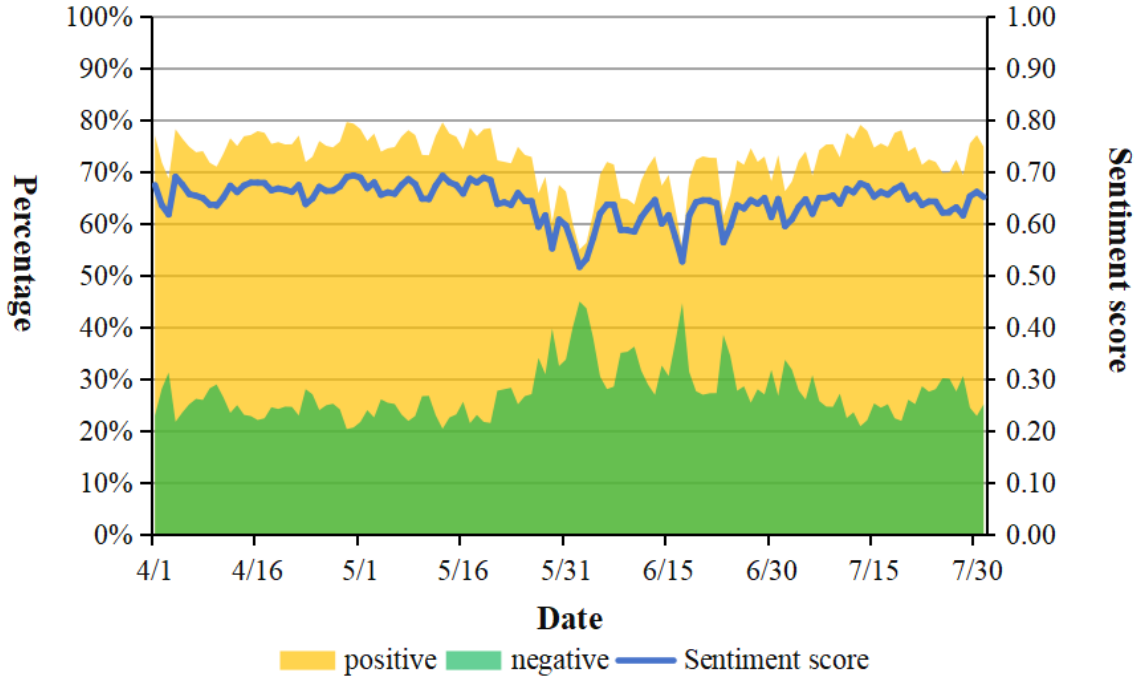

Figure 6. Residual ACF and PACF. ACF: auto-correlation function. PACF：partial auto-correlation function.

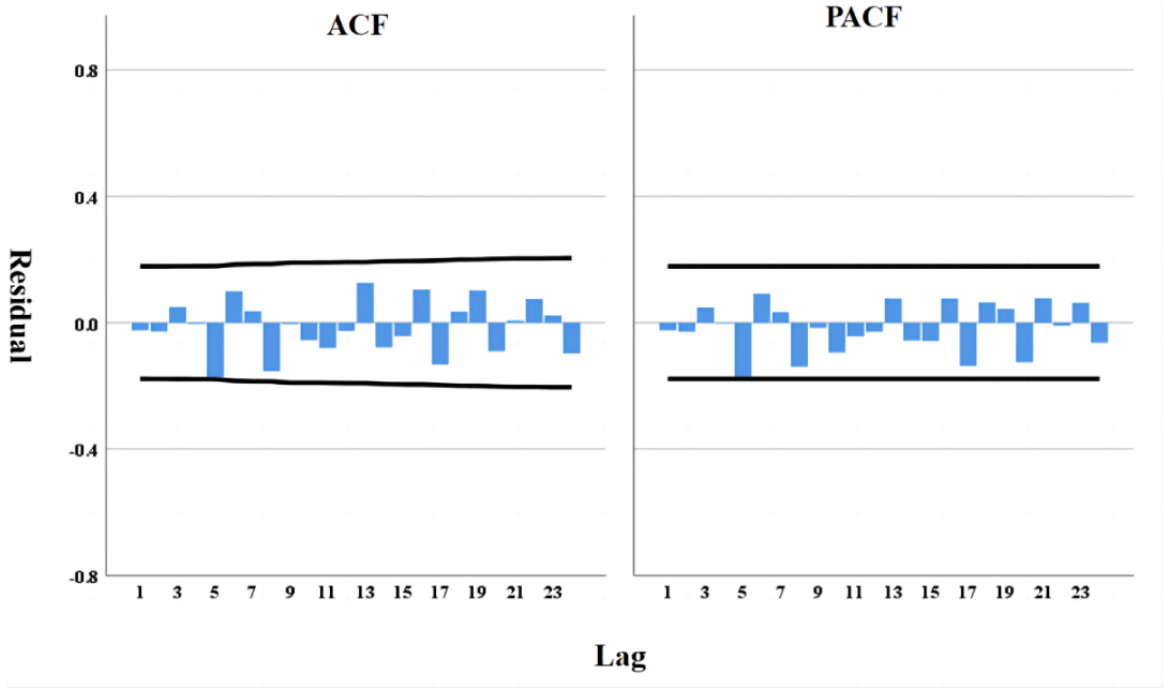

XSL•FO
RenderX

Figure 7. A time series plot of daily sentiment score from April 2022 to August 2022 predicted using the ARIMA model.

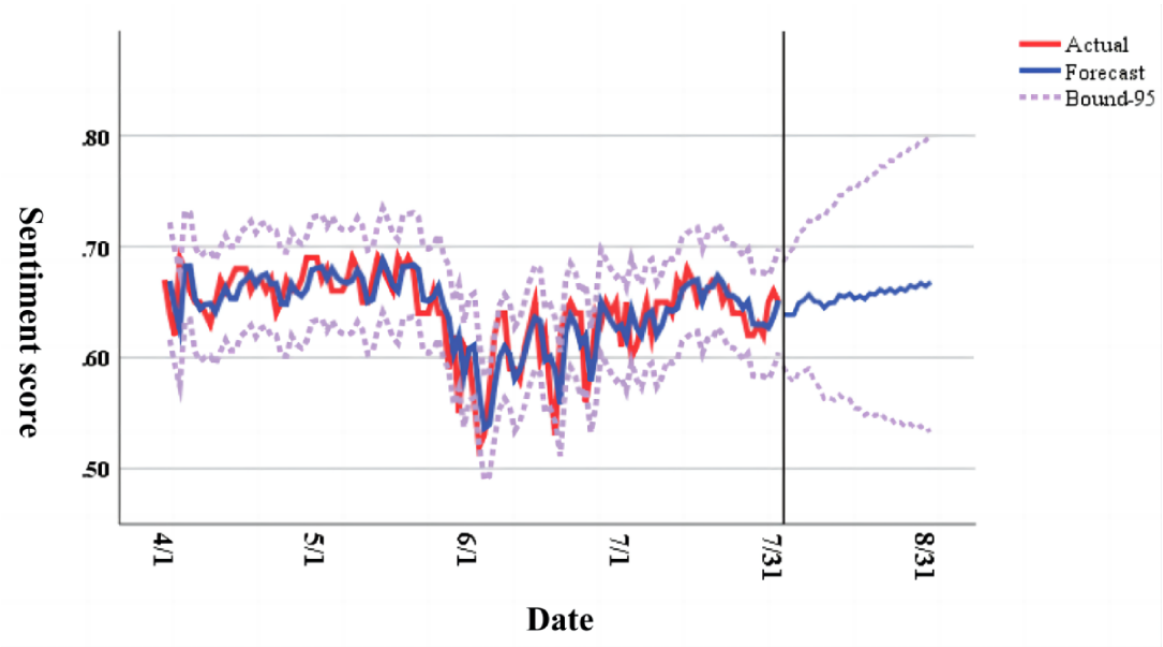

Sentiment Scores for Different Topics

Table 2 shows the emotional scores across different topics. We found that all topics have positive sentiment scores, with the proportion of positive posts higher than negative ones (χ^2_7=24844.4, *P*<.001). Specifically, the topics "Shanghai outbreak" and "nucleic acid testing price" had over 30% negative sentiment posts. The daily sentiment score heat maps for different topics are shown in Multimedia Appendix 2.

Table 2. Sentiment scores for 8 topics and a comparison of positive and negative posts across different topics.

Topic	Sentiment score	Proportion of positive posts, n (%)	Proportion of negative posts, n (%)	Chi-square (*df*)	*P* value
Epidemic prevention and control measures	0.74	41,074 (85.07)	7210 (14.93)	24,844.4 (7)	<.001
Shanghai outbreak	0.59	124,009 (63.40)	71,603 (36.60)	__[a]	—
Hospital visits	0.64	1710 (80.02)	427 (19.98)	—	—
Transportation	0.63	37,247 (72.89)	13,854 (27.11)	—	—
Compliance with epidemic prevention policies	0.77	28,532 (91.94)	2501 (8.06)	—	—
Nucleic acid testing price	0.62	38,123 (68.69)	17,379 (31.31)	—	—
Regular nucleic acid testing	0.72	41,524 (88.05)	5633 (11.95)	—	—
Epidemic notification	0.68	25,749 (78.65)	6991 (21.35)	—	—

[a]Not applicable.

Sentiment Changes Before and After Key Time Points

On May 22, 2022, the National Healthcare Security Administration announced a reduction in nucleic acid testing prices [38]. Shanghai announced on June 1, 2022, the full restoration of production order [39]. As shown in Table 3, both the topics of "nucleic acid testing price" and the "Shanghai outbreak" maintained a positive attitude during different time periods. Before May 22, the sentiment scores of posts related to "nucleic acid testing prices" were higher than they were after May 22 (t_{120}=3.882, *P*<.001). Before June 1, the sentiment scores of posts about the "Shanghai outbreak" were higher than they were after June 1 (t_{120}=11.943, *P*<.001).

XSL•FO
RenderX

Table 3. Changes in sentiment scores across different time periods.

Topic	Sentiment score, Mean (SD)	*t* test (*df*)	*P* value
Nucleic acid testing price	__[a]	3.882 (120)	<.001
Before May 22	0.66 (0.06)	—	—
After May 22	0.61 (0.07)	—	—
Shanghai outbreak	—	11.943 (120)	<.001
Before June 1	0.61 (0.03)	—	—
After June 1	0.55 (0.03)	—	—

[a]Not applicable.

Discussion

Principal Findings

This paper explored public attitudes and opinions on regular nucleic acid testing on Weibo using sentiment and topic analysis. It analyzed public sentiment expressions on social media in the context of public health emergencies. To our knowledge, this is the first study to explore public attitudes and opinions on regular nucleic acid testing during the COVID-19 pandemic, filling a gap in existing research. The results of this study can help the government promptly understand changes in public sentiment during future emergencies, effectively guide public emotions, manage negative emotions, and maintain social order.

Positive Correlation Between the Volume of Weibo Posts and the Severity of the Epidemic

The study results showed a positive correlation between the daily number of posts on the topic of regular nucleic acid testing on Chinese social media and the number of new confirmed COVID-19 cases in China (ρ=0.656, P<.001). As the number of new confirmed cases increased, the government increased the frequency of nucleic acid testing to identify and control the outbreak more quickly. This led to more people participating in testing and sharing their experiences and views on social media. Our findings were consistent with a study on epidemic-related rumors on Weibo from January 20 to May 28, 2020, which found a positive correlation between the daily number of posts about COVID-19 misinformation on Chinese social media and the daily number of newly confirmed and newly suspected COVID-19 cases in China [40]. This indicated that the dissemination of information about COVID-19 in China was synchronized with the spread of the epidemic [40]. The volume of discussions on social media can serve as a supplementary indicator for predicting epidemic trends [41]. By monitoring public opinion on social media in real time, public health departments can promptly understand public concerns and provide early warnings for emerging outbreaks, thereby better formulating and adjusting control measures [41].

The 3 Topics of Primary Concern to the Public

This study identified 8 main topics of public interest regarding regular nucleic acid testing. Ranked in order of attention from highest to lowest, they were the Shanghai outbreak, nucleic acid testing price, transportation, epidemic prevention and control measures, regular nucleic acid testing, epidemic notification, and hospital visits. Wang used the Weibo platform to investigate public concerns during the early stage of the pandemic (December 1, 2019, to July 31, 2020), identifying 8 topics, including epidemic notification, prevention measures, and health care [15]. In comparison, this study found that after the implementation of the regular nucleic acid testing policy, public attention shifted more toward topics such as the Shanghai outbreak, nucleic acid testing price, and transportation. Analyzing the reasons, the period focused on in this study coincided with the outbreak of the pandemic in Shanghai, during which multiple rounds of large-scale mass nucleic acid testing were conducted [42]. The frequent testing led to public concerns about the costs of nucleic acid testing, and any government policy adjustments regarding these costs attracted significant attention. Simultaneously, as many regions required travelers to provide proof of a negative nucleic acid test, public transportation and travel faced strict restrictions. Therefore, the Shanghai outbreak, nucleic acid testing price, and transportation have become focal points of public concern due to their close relation to public health and daily life. The government can determine the priority of public concerns based on the volume of posts on various topics on social media and promptly disclose relevant policies and information on epidemic prevention and control.

Public's Positive Emotion at the Regular Nucleic Acid Testing Policy

Sentiment analysis results showed that public emotions toward regular nucleic acid testing were predominantly positive, though there were periods of downturn influenced by certain events. The positive sentiment in Weibo posts indicated that the public generally approved of the regular nucleic acid testing policy, which is consistent with the conclusions of the study by Tan [43]. From June 2021 to August 2021, Tan [43] analyzed the opinions of residents in 4 cities in Guangdong Province on nucleic acid testing using semistructured interviews and social media comments. The results revealed that the public generally held a positive attitude toward nucleic acid testing [43]. Other studies have shown that sudden policy changes provoke heated discussions among the public on social media platforms [14]. If timely explanations and public communication are not provided, this can lead to public concern and anxiety, resulting in a surge of negative emotions [14]. On June 1, the Shanghai government suddenly announced adjustments to its pandemic prevention policies, leading to an increase in negative public emotions [39]. By June 2, the sentiment score had reached its lowest point. Fortunately, public negative emotions decreased

XSL•FO
RenderX

afterward, and sentiment gradually stabilized. Support for regular nucleic acid testing increased over time. Decision makers should closely monitor changes in public reactions to address the most critical public needs at different stages of the pandemic [14]. By doing so, they can address public concerns through timely and effective communication [14]. In addition, policy makers can leverage social media and big data analysis to support public health decision-making [44].

Impact of Price Reduction Policy on Sentiment Scores of "Nucleic Acid Testing Price"

Due to the economic burden on the public induced by regular nucleic acid testing, the government adjusted the prices of nucleic acid tests multiple times. On May 22, 2022, the National Healthcare Security Administration issued a notice to reduce the prices of nucleic acid testing [38]. However, after the policy was implemented, the sentiment score in Weibo posts, while still positive, was lower than before the policy was issued. Analyzing the reasons, first, although the cost of regular nucleic acid testing was primarily subsidized by the government, the requirement to undergo testing every few days disrupted the public's daily life [45]. Second, when the price reduction policy was first announced, media reports sometimes highlighted inconsistencies between local and central policies or issues with policy implementation. Such negative reports could raise public doubts and increase the proportion of negative emotions. Studies have shown that after the introduction of new policies, insufficient government explanation or failures in public communication can greatly affect public perception and trust due to misinformation [46]. Therefore, when introducing new policies, the government should release information in a timely and transparent manner, including specific implementation details and progress, to reduce information asymmetry. During the policy implementation process, the government should establish diverse channels for public feedback, gather public opinions and suggestions, and promptly respond to and address public concerns to enhance the effectiveness of policy implementation.

Impact of Reopening Policy on Sentiment Scores of "Shanghai Outbreak"

Among all topics, the proportion of negative emotions regarding the "Shanghai outbreak" reached 36.60%, and the public sentiment score declined after June 1. Since the outbreak of the pandemic in March 2022, the number of infections in Shanghai had grown exponentially [47]. On April 1, the Shanghai government decided to implement city-wide static management, initiating a lockdown that lasted for 3 months [42]. Studies have shown that quarantine can lead to negative psychological effects such as posttraumatic stress symptoms, confusion, and anger [21,22]. On June 1, Shanghai fully resumed normal life [39]. However, during the actual process of resuming work and production, various issues were encountered, leading to strong negative emotions expressed by people on Weibo, which caused a decline in public sentiment scores. Furthermore, although the restoration of production and daily life was underway, the risk of a resurgence of the pandemic remained. This caused public concern about the future development of the pandemic, leading to a decline in sentiment scores. In addition, numerous rumors about the Shanghai outbreak circulated on the internet, contributing to negative emotions [23]. Previous research has found that public sentiment typically shifts from negative to positive after rumors are debunked [24]. Guo investigated public attitudes toward the Shanghai outbreak on the Weibo platform and found that public sentiment scores consistently declined during the outbreak, falling into negative territory in the second and third quarters [48]. The severe negative emotions among the public significantly affected Shanghai's image [48]. Therefore, after announcing policies, the government needs to promptly track trends in public sentiment, identify potential negative emotions and hot issues, and respond effectively to control negative emotions within reasonable limits. In addition, the government should proactively guide public opinion by using authoritative media and expert interpretations to effectively disseminate policies related to public health emergencies, thereby reducing the spread of rumors. Furthermore, the public needs to enhance their ability to identify and not believe or spread rumors. Joint efforts from the government, media platforms, and the public are essential to create a positive and uplifting public opinion environment.

Limitations

This study has several limitations. First, the Weibo data collected used only 1 keyword, which may include only a portion of posts related to nucleic acid testing rather than all relevant posts. Therefore, the summary of the main emotions and topics of Weibo users may not be comprehensive. Second, most Weibo users are young people, so the analysis results may be more biased toward the perspectives of young people [49]. Third, we collected data from only 1 social media platform. Although Weibo has a relatively large user base in China, it is hasty to conclude that the opinions expressed on Weibo represent the views of the public [50]. Finally, due to the limitations of the Weibo platform and its functionalities, most information regarding user geographical location and other details collected in this study was missing [51]. Consequently, the study did not explore whether users from different geographic locations had different sentiments and topics of concern regarding the regular nucleic acid testing policy.

Conclusions

During public health emergencies, the topics of public concern were multifaceted. Public sentiment toward regular nucleic acid testing policies was generally positive, but fluctuations occurred following key policy announcements. The government needs to monitor information posted by the public on social media platforms to understand the main areas of concern. By promptly disclosing information on media platforms and engaging in effective communication, the government can bridge the information gap between public needs and government responses to create a favorable public opinion environment.

XSL•FO
RenderX

Acknowledgments

We appreciate the technical support provided by Yiluo, Yao. This work was supported by the Key Project of Natural Science Research in Colleges and Universities of Anhui Province (grant KJ2021A0255) and the Anhui Medical University School of Nursing Graduate Youth Program Cultivation Project (grant hlqm12022007). The study funders had no role in the study design, data collection, analysis, interpretation, or drafting of the manuscript.

Authors' Contributions

ML and LX contributed to conceptualization. ML, SY, BL, and LX contributed to the methodology. ML, YZ, and JL performed formal analysis and investigation. ML, QC, CG, and JR contributed to the writing and original draft preparation. ML and LX contributed to writing-review and editing. LX performed supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Autoregressive Integrated Moving Average (ARIMA) model fit metrics and ARIMA model predicted daily sentiment scores.
[DOC File , 168 KB - jmir_v26i1e58518_app1.doc]

Multimedia Appendix 2
Daily sentiment score heatmaps for different topics.
[PDF File (Adobe PDF File), 1295 KB - jmir_v26i1e58518_app2.pdf]

References

1. A national master plan for responding to public emergencies. General Office of the State Council. 2005. URL: http://www.gov.cn/yjgl/2005-08/07/content_21048.htm [accessed 2023-10-08]
2. Cao Y, Shan J, Gong Z, Kuang J, Gao Y. Status and challenges of public health emergency management in China related to COVID-19. Front Public Health 2020;8:250. [doi: 10.3389/fpubh.2020.00250] [Medline: 32574311]
3. Dong X. Construction of risk communication framework for public health emergencies from the perspective of public opinion. Theory and Reform 2020;4:14-23. [doi: 10.13553/j.cnki.llygg.2020.04.002]
4. Li X. Concept choice and system construction of instant precaution of the emerging infectious diseases. Admin L Rev 2023;06:113-122 [FREE Full text]
5. Zhang J, Liu G, Huang J. Characteristics, prevention and control measures of SARS-CoV-2 omicron variant. Chin J Infect Control 2022;21(8):816-822 [FREE Full text]
6. WHO: New crown outbreak remains? public health emergency of international concern. Xinhuanet. URL: http://world.people.com.cn/n1/2022/1020/c1002-32548538.html [accessed 2023-10-08]
7. Liu H, Chai G, Zhang Y. Policy research on response to public health emergencies. Shanxi Med J 2024;53(7):543-545 [FREE Full text]
8. Normalized nucleic acid policy implemented in many parts of the country! expert: it's the least socially costly way to prevent epidemics. Tecent. URL: https://new.qq.com/rain/a/20220503A06ACW00 [accessed 2024-06-04]
9. Wang W, Liu Y, Wang Y, Wu L, Zhou F, Li L, et al. Centralized management of specimen collection for COVID-19 nucleic acid RT-PCR test under normalized epidemic prevention and control. J Nurs Sci 2021;36(18):54-57 [FREE Full text]
10. Liao J, Hou W, Xiong Q, Huang X, Dong X. What is the best way to ensure an orderly process for regular nucleic acid testing in multiple locations? Xinhua New Media. 2022. URL: https://doi.org/10.28870/n.cnki.nxhmr.2022.004339 [accessed 2024-09-21]
11. Sabatier PA, Weible CM. Theories of the policy process. 3rd ed. Boulder, CO: Westview Press; 2014.
12. Weibo releases Q3 2022 report, average daily active users reached 253 million in september. Tecent. URL: https://new.qq.com/rain/a/20221118A01NDT00 [accessed 2023-10-08]
13. Wang X, Li Y, Liu T, Zhang L. Research on the collaborative model of sentiment analysis and topic mining of micro-blogging users in the context of covid-19. Journal of the China Society for Scientific and Technical Information 2021;40(3):223-233 [FREE Full text]
14. Zhou X, Song Y, Jiang H, Wang Q, Qu Z, Zhou X, et al. Comparison of public responses to containment measures during the initial outbreak and resurgence of COVID-19 in China: infodemiology study. J Med Internet Res 2021;23(4):e26518 [FREE Full text] [doi: 10.2196/26518] [Medline: 33750739]
15. Wang J, Zhou Y, Zhang W, Evans R, Zhu C. Concerns expressed by Chinese social media users during the COVID-19 pandemic: content analysis of sina weibo microblogging data. J Med Internet Res 2020;22(11):e22152 [FREE Full text] [doi: 10.2196/22152] [Medline: 33151894]

XSL•FO
RenderX

16. Lwin MO, Sheldenkar A, Lu J, Schulz PJ, Shin W, Panchapakesan C, et al. The evolution of public sentiments during the COVID-19 pandemic: case comparisons of India, Singapore, South Korea, the United Kingdom, and the United States. JMIR Infodemiology 2022;2(1):e31473 [FREE Full text] [doi: 10.2196/31473] [Medline: 37113803]
17. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social network analysis of COVID-19 sentiments: application of artificial intelligence. J Med Internet Res 2020;22(8):e22590 [FREE Full text] [doi: 10.2196/22590] [Medline: 32750001]
18. Wu X, Li Z, Xu L, Li P, Liu M, Huang C. COVID-19 vaccine-related information on the WeChat public platform: topic modeling and content analysis. J Med Internet Res 2023;25:e45051 [FREE Full text] [doi: 10.2196/45051] [Medline: 37058349]
19. Xie Z, Wang X, Jiang Y, Chen Y, Huang S, Ma H, et al. Public perception of COVID-19 vaccines on Twitter in the United States. medRxiv 2021 [FREE Full text] [doi: 10.1101/2021.10.16.21265097] [Medline: 34704100]
20. Hong Y, Xie F, An X, Lan X, Liu C, Yan L, et al. Evolution of public attitudes and opinions regarding COVID-19 vaccination during the vaccine campaign in China: year-long infodemiology study of weibo posts. J Med Internet Res 2023;25:e42671 [FREE Full text] [doi: 10.2196/42671] [Medline: 36795467]
21. Chen X, Yik M. The emotional anatomy of the Wuhan Lockdown: sentiment analysis using weibo data. JMIR Form Res 2022;6(11):e37698 [FREE Full text] [doi: 10.2196/37698] [Medline: 36166650]
22. Tri Sakti AM, Mohamad E, Azlan A. Mining of opinions on COVID-19 large-scale social restrictions in Indonesia: public sentiment and emotion analysis on online media. J Med Internet Res 2021;23(8):e28249 [FREE Full text] [doi: 10.2196/28249] [Medline: 34280116]
23. Wang P, Shi H, Wu X, Jiao L. Sentiment analysis of rumor spread amid COVID-19: based on weibo text. Healthcare (Basel) 2021;9(10):1275 [FREE Full text] [doi: 10.3390/healthcare9101275] [Medline: 34682955]
24. Dong W, Tao J, Xia X, Ye L, Xu H, Jiang P, et al. Public emotions and rumors spread during the COVID-19 epidemic in China: web-based correlation study. J Med Internet Res 2020;22(11):e21933 [FREE Full text] [doi: 10.2196/21933] [Medline: 33112757]
25. 15 cities have announced the full rollout! new market favorite? normalized nucleic acid test? behind the industry to explore. Hexun. URL: https://news.hexun.com/2022-05-11/205923673.html [accessed 2024-06-04]
26. Nucleic acid testing has been normalized in more than 40 places and is being adjusted in many places. Thepaper. URL: https://www.thepaper.cn/newsDetail_forward_18828334 [accessed 2024-06-04]
27. Liu X, Zhao X. Sentiment analysis of covid-19 vaccination based on text data mining micro-blogging users in the context of covid-19. Inf Technol Standards 2022;3:74-78 [FREE Full text]
28. Sogou cell thesaurus. Sogou. URL: https://pinyin.sogou.com/dict/ [accessed 2023-10-08]
29. Epidemic notification. National Health Commission of the People's Republic of China. URL: http://www.nhc.gov.cn/xcs/yqtb/list_gzbd_9.shtml [accessed 2024-06-04]
30. Garcia K, Berton L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Appl Soft Comput 2021;101:107057 [FREE Full text] [doi: 10.1016/j.asoc.2020.107057] [Medline: 33519326]
31. Alexander G, Bahja M, Butt G. Automating large-scale health care service feedback analysis: sentiment analysis and topic modeling study. JMIR Med Inform 2022;10(4):e29385 [FREE Full text] [doi: 10.2196/29385] [Medline: 35404254]
32. Tan C, Chen X, Liang Y, Li Y. Sentiment analysis of onlookers on social media in privacy disclosure micro-blogging users in the context of covid-19. Inf Sci 2023;41(03):8-18 [FREE Full text] [doi: 10.13833/j.issn.1007-7634.2023.03.002]
33. Zhao K, Wang H. Research on the method of selecting the optimal number of topics for LDA: taking CNKI literature as an example. Statistics & Decision 2020;16:175-179 [FREE Full text] [doi: 10.13546/j.cnki.tjyjc.2020.16.039]
34. Praveen S, Vajrobol V. Understanding the perceptions of healthcare researchers regarding ChatGPT: a study based on bidirectional encoder representation from transformers (BERT) sentiment analysis and topic modeling. Ann Biomed Eng 2023;51(8):1654-1656. [doi: 10.1007/s10439-023-03222-0] [Medline: 37129780]
35. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. IEEE Access 2020;8:138162-138169 [FREE Full text] [doi: 10.1109/ACCESS.2020.3012595] [Medline: 34812342]
36. Gao H, Guo D, Wu J, Li L. Weibo users' emotion and sentiment orientation in traditional Chinese medicine (TCM) during the COVID-19 pandemic. Disaster Med Public Health Prep 2022;16(5):1835-1838 [FREE Full text] [doi: 10.1017/dmp.2021.259] [Medline: 34369351]
37. Boucher JC, Kim SY, Jessiman-Perreault G, Edwards J, Smith H, Frenette N, et al. HPV vaccine narratives on Twitter during the COVID-19 pandemic: a social network, thematic, and sentiment analysis. BMC Public Health 2023;23(1):694 [FREE Full text] [doi: 10.1186/s12889-023-15615-w] [Medline: 37060069]
38. The National Healthcare Security Administration. Notice of further price reductions for new coronavirus nucleic acid test and antigen test. URL: https://www.gov.cn/zhengce/zhengceku/2022-05/25/content_5692238.htm [accessed 2024-06-18]
39. Shanghai to resume normal production and life order from June 1. Xinhua. URL: http://www.news.cn/politics/2022-05/31/c_1128699790.htm [accessed 2024-06-18]
40. Zhang S, Pian W, Ma F, Ni Z, Liu Y. Characterizing the COVID-19 infodemic on Chinese social media: exploratory study. JMIR Public Health Surveill 2021;7(2):e26090 [FREE Full text] [doi: 10.2196/26090] [Medline: 33460391]

XSL•FO
RenderX

41. Terry K, Yang F, Yao Q, Liu C. The role of social media in public health crises caused by infectious disease: a scoping review. BMJ Glob Health 2023;8(12):e013515 [FREE Full text] [doi: 10.1136/bmjgh-2023-013515] [Medline: 38154810]
42. 9 charts reviewing the Shanghai outbreak. baidu. URL: https://baijiahao.baidu.com/s?id=1734391556748414146&wfr=spider&for=pc [accessed 2024-06-18]
43. Tan X. Hope and trust: public attitudes toward mass COVID-19 testing programs in Guangzhou, China. Front Psychol 2022;13:972398 [FREE Full text] [doi: 10.3389/fpsyg.2022.972398] [Medline: 36186316]
44. Adekunle N, Okolo A, None Tolulope Olorunsogo, None Oloruntoba Babawarun. Leveraging big data and analytics for enhanced public health decision-making: A global review. GSC Advanced Research and Reviews 2024;18(2):450-456 [FREE Full text] [doi: 10.30574/gscarr.2024.18.2.0078]
45. Costs of regularized nucleic acid testing to be borne by local governments. Xinhua. URL: https://www.gov.cn/xinwen/2022-05/26/content_5692466.htm [accessed 2024-06-18]
46. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021;23(6):e24435 [FREE Full text] [doi: 10.2196/24435] [Medline: 34115608]
47. May 5 shanghai news. Xinminwanbao. 2022. URL: https://baijiahao.baidu.com/s?id=1731943342444888195&wfr=spider&for=pc [accessed 2023-10-08]
48. Guo Y, Zu L, Chen D, Zhang H. A study of public attitudes toward Shanghai's image under the influence of COVID-19: evidence from comments on sina weibo. Int J Environ Res Public Health 2023;20(3):2297 [FREE Full text] [doi: 10.3390/ijerph20032297] [Medline: 36767664]
49. Sun K, Wang H, Zhang J. The impact factors of social media users' forwarding behavior of COVID-19 vaccine topic: based on empirical analysis of Chinese weibo users. Front Public Health 2022;10:871722 [FREE Full text] [doi: 10.3389/fpubh.2022.871722] [Medline: 36187642]
50. Tian R, Yin R, Gan F. Exploring public attitudes toward live-streaming fitness in China: a sentiment and content analysis of China's social media weibo. Front Public Health 2022;10:1027694 [FREE Full text] [doi: 10.3389/fpubh.2022.1027694] [Medline: 36408048]
51. The Cyberspace Administration of China. Regulations on the management of internet user account information. The State Council of the People's Republicof China. 2022. URL: https://www.gov.cn/zhengce/zhengceku/2022-06/28/content_5698179.htm [accessed 2024-09-21]

Abbreviations

ACF: auto-correlation function
ARIMA: Autoregressive Integrated Moving Average
BERT: Bidirectional Encoder Representations from Transformers
LDA: latent Dirichlet allocation
PACF: partial auto-correlation function

Edited by A Schwartz; submitted 18.03.24; peer-reviewed by X Meng, A Aburaed; comments to author 25.05.24; revised version received 28.06.24; accepted 06.09.24; published 28.10.24.

Please cite as:
Liu M, Yuan S, Li B, Zhang Y, Liu J, Guan C, Chen Q, Ruan J, Xie L
Chinese Public Attitudes and Opinions on Health Policies During Public Health Emergencies: Sentiment and Topic Analysis
J Med Internet Res 2024;26:e58518
URL: https://www.jmir.org/2024/1/e58518
doi:10.2196/58518
PMID:39466313

©Min Liu, Shuo Yuan, Bingyan Li, Yuxi Zhang, Jia Liu, Cuixia Guan, Qingqing Chen, Jiayi Ruan, Lunfang Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ever Use of Telehealth in Nebraska by March 2021: Cross-Sectional Analysis

Lisa C Smith[1], PhD, LMSW; George Johnson[2], MSPH; Snehal Jadhav[3], BDS, MPH; Josiane Kabayundo[3], BPharm, MPH; Muskan Ahuja[3], BDS; Hongmei Wang[4], PhD; Kendra L Ratnapradipa[3], MSW, PhD

[1]Grace Abbott School of Social Work, University of Nebraska Omaha, Omaha, NE, United States

[2]Department of Health Promotion, University of Nebraska Medical Center, Omaha, NE, United States

[3]Department of Epidemiology, University of Nebraska Medical Center, Omaha, NE, United States

[4]Department of Health Services Research & Administration, University of Nebraska Medical Center, Omaha, NE, United States

Corresponding Author:
Kendra L Ratnapradipa, MSW, PhD
Department of Epidemiology
University of Nebraska Medical Center
984395 Nebraska Medical Center
Omaha, NE, 68198-4395
United States
Phone: 1 402 552 7238
Email: kendra.ratnapradipa@unmc.edu

Abstract

Background: Nationally, COVID-19 spurred the uptake of telehealth to facilitate patients' access to medical care, especially among individuals living in geographically isolated areas. Despite the potential benefits of telehealth to address health care access barriers and enhance health outcomes, there are still disparities in the accessibility and utilization of telehealth services. Hence, identifying facilitators and barriers to telehealth should be prioritized to ensure that disparities are mitigated rather than exacerbated.

Objective: This study aims to identify factors associated with ever use of telehealth in Nebraska, a primarily rural state with a significant portion of its population living in nonmetropolitan areas.

Methods: A stratified random sample of Nebraska households (n=5300), with oversampling of census tracts with at least 30% African American, Hispanic, or Native American populations, received a mailed survey (English and Spanish) with web-based response options about social determinants of health and health care access (October 2020-March 2021). Survey weights were used for all calculations. Chi-square tests were used to compare telehealth use (yes or no) by participant sociodemographic, health, and access variables. Robust Poisson regression models were used to compute prevalence odds ratios (POR) with 95% CIs of telehealth use after controlling for socioeconomic, demographic, and health conditions.

Results: The overall response rate was 20.8% (1101/5300). About 25.5% of Nebraska adults had ever used telehealth (urban 26.4%, rural 20.8%), despite 97% of respondents reporting internet access (98.3% urban, 90.5% rural). In the chi-square analysis, telehealth use was statistically significantly more common ($P<.05$) among those who are aged <45 years (32.4%), female (30.7%), and non-Hispanic (25.9%); with at least a bachelor's degree (32.6%); who had a routine checkup (30.2%) or health care visit other than a routine checkup (34.2%); and with any chronic health conditions (29.6%) but did not differ ($P\geq.05$) by race, marital status, income, insurance, having a primary care provider, or 1-way travel time for medical visits. In univariate models, internet access, age, sex, ethnicity, education, any health care visit in the past year, and no chronic health condition were significant ($P<.05$). When adjusted, education (POR 1.87, 95% CI 0.33-10.63) and sex (1.38, 0.93-2.04) were not significant, but internet access (5.43, 1.62-18.16), age <45 (5.33, 2.22-12.81) and 45-64 years (9.05, 2.37-34.62), non-Hispanic ethnicity (7.40, 2.39-22.90), any health care visit (2.43, 1.23-4.79), and any chronic condition (1.73, 1.09-2.76) were significantly associated with having ever used telehealth.

Conclusions: This study highlights disparities in telehealth use. Despite high coverage, internet access was a significant predictor of telehealth use, highlighting the role of the digital divide in telehealth access and use. Telehealth use was significantly less prevalent among older adults, people without chronic health conditions, and Hispanic individuals. Targeted interventions that address barriers to telehealth use and improve health care access are warranted.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53320) doi:10.2196/53320

KEYWORDS

COVID-19; digital divide; health care access; telehealth; cross-sectional study; Nevada; United States; adult; medical care; geographical area; disparity; accessibility; utilization; survey; chi-square test; regression model; socioeconomic; demographic; health condition; digital health

Introduction

Telehealth, an umbrella term including telemedicine and nonclinical functions such as health education and patient support services, is a health care delivery model that uses technology to provide medical services remotely [1]. It encompasses a wide range of services and applications, all aimed at delivering health care consultations, monitoring, education, and support over digital communication channels [2]. Telehealth has gained significant prominence and acceptance, particularly in recent years, due to advances in technology and the need for more accessible and convenient health care options.

Telehealth has unquestionably widened health care access in the United States. Evidence shows it has effectively removed barriers associated with traveling to care [3]. It has been demonstrated to enhance health outcomes for individuals who are unable to physically visit health care providers but have access through phone or video platforms [2]. Additionally, there is compelling evidence indicating that the internet has empowered individuals to take charge of their health care utilization by accessing supplementary health information, managing personal health data, and engaging in meaningful communication with health care providers concerning their health conditions or enhancing disease management [4]. Telehealth has created avenues for patients to access information on disease management, receive medication advice, and be more actively engaged in their care.

In more recent years, telehealth has experienced a remarkable surge due to the COVID-19 pandemic [5]. The global health crisis forced health care systems and professionals to rapidly adapt to new methods of delivering care while prioritizing the safety of both patients and health care workers [5]. Telehealth emerged as a vital tool during this period, playing a pivotal role in the pandemic response [5]. It allowed medical professionals to assess and diagnose, provide treatment and follow-up, and monitor and track individuals while minimizing exposure [6]. While the surge in telehealth usage was initially driven by necessity, its value in providing accessible, efficient, and safe health care delivery is likely to endure in the post-pandemic health care landscape.

A persistent barrier to the expanding use of telehealth and its ability to address health disparities is the digital divide. The digital divide is the gap or disparity in access to and use of digital technologies, particularly the internet, between different groups of people or communities [7,8]. The digital divide encompasses access to internet connectivity, internet speed, the availability of digital devices, digital literacy, and the ability to effectively use digital tools for various purposes [9,10]. The digital divide is a significant social and economic issue, as it can have far-reaching consequences for individuals and communities [11].

Several characteristics have been linked to the widening of the digital divide, including age, educational attainment, income, sex, health status, racial or ethnic background, and geographic location [12]. The digital divide has had a significant impact on individuals, making them less likely to access health care–related information and potentially hindering their ability to comprehend information available through digital sources [13]. According to data from the Federal Communications Commission (FCC), which regulates interstate and international communications in the United States, there has been a notable disparity in broadband coverage across different regions, although it is gradually narrowing. Specifically, broadband coverage expanded in urban areas from 96.7% in 2015 to 98.8% in 2019, while coverage in rural areas increased from 61.5% to 82.7% and coverage in tribal lands increased from 57.8% to 79.1% [14].

Efforts to bridge the digital divide often involve government initiatives, private sector investments, and nonprofit organizations working to improve internet infrastructure, provide digital literacy training, and distribute digital devices to underserved populations [15]. Additionally, policies aimed at ensuring affordable internet access and promoting digital inclusion are crucial in addressing this issue [16]. The digital divide is not only a matter of access to technology but also a matter of social equity and inclusion [16].

Reducing the digital divide is essential for ensuring that all individuals and communities have equal opportunities to participate in the digital age and benefit from the numerous opportunities it offers, such as telehealth. Given that a significant portion (approximately 35%) of Nebraska's population resides in nonmetropolitan areas, this state offers a unique opportunity for in-depth analysis of health care access challenges inherent to rural settings. While 56% of Nebraska's 1.96 million inhabitants live in 3 metropolitan counties in the eastern part of the state, large portions of the state are sparsely populated [17], and 13 of the 93 counties lacked any actively practicing primary care physician in 2021 [18]. Nebraska has 120 areas designated by the Health Resources and Services Administration as health professional shortage areas for primary care and 88 designated medically underserved areas or populations [19]. Taken together, these statistics suggest that Nebraska might benefit from expanded telehealth to help bridge structural barriers to receipt of general medical care. The aim of this study was to assess the telehealth utilization gap during the COVID-19 pandemic in Nebraska.

XSL•FO
RenderX

Methods

Ethical Considerations

The protocol for this study was reviewed by the University of Nebraska Medical Center Institutional Review Board (605-20-EX) and deemed exempt from oversight under 45 CFR 46:104(d), category 2. The cover letter provided information about the study, and completion of the survey indicated voluntary informed consent to participate. Responses were anonymous, and data were scrubbed of associated addresses before the Bureau of Sociological Research (BOSR) provided it to the researchers. A US $1 cash incentive was included in the initial mailing with no further compensation provided.

Study Design

This was a cross-sectional, mailed survey about health care access.

Sample Design

To try to ensure broad representation of adult Nebraskans by urbanicity level, a stratified address-based sampling process used the 2013 rural-urban continuum codes to classify counties as urban large, urban small, and rural. To try to increase representation by race and ethnicity, we oversampled census tracts with at least 30% African American (urban large), 30% Hispanic (statewide), or 30% Native American (rural) residents. Assuming an estimated 25% response rate based on prior BOSR studies, 5300 addresses were purchased to sufficiently power the analysis. The detailed methodology report is available in Multimedia Appendix 1.

Recruitment and Data Collection

The survey methods have been described elsewhere but are briefly summarized here [20]. BOSR was contracted for data collection and purchased Nebraska residential mailing addresses from Dynata. The number and timing of mailings were based on professional recommendations from BOSR. An initial survey packet was mailed to selected addresses in October 2020. The packet (English and Spanish) invited the adult (aged 19 years or older) with the next birthday to complete a web-based survey using a link to Qualtrics. A reminder postcard was sent 2 weeks later, and a full survey packet (English and Spanish) was sent 3 weeks thereafter containing a cover letter, survey, and postage-paid return envelope. Due to lower-than-expected response rates, an additional mailing (English) was sent to nonrespondents in early January 2021 consisting of a reminder letter with a web link, the survey, and a postage-paid return envelope. The survey remained open through early March 2021 for any additional responses. BOSR staff used double entry of paper returns for quality control and sent to the research team a cleaned datafile scrubbed of addresses and including survey weights.

Measures

The outcome measure was ever use of telehealth (yes or no), worded as “Telehealth is a broad term referring to provision of health education and medical services through telecommunications technology. It includes remote monitoring of vital signs, consultation, evaluation, diagnosis, and prescription. Have you ever used telehealth?”

Sociodemographic variables included age group calculated from year born (<45, 45-64, and ≥65 years); sex (male and female); race (White only, Other, and Not reported) reclassified from “mark all that apply” with “prefer not to answer” added to nonresponse; Hispanic ethnicity (yes or no); marital status (married or partnered versus divorced, separated, widowed, or single); highest education level completed (≤high school graduate, some college, and bachelor’s degree or above); annual household income (<US $50,000, US $50,000-US $74,999, >US $75,000); health insurance (Medicaid, Medicare, Private, Other, and None) recoded from “mark all that apply” to sequentially assign to mutually exclusive categories; and rural or urban status.

A checklist of chronic health conditions (heart condition, high blood pressure, diabetes, lung disease, arthritis, stroke, cancer, depression, and anxiety) was recoded as any versus none.

Health care utilization was assessed for routine checkup, worded as “A routine checkup is a general physical exam, not an exam for a specific injury, illness, or condition. About how long has it been since you last saw a doctor or other health care professional for a routine checkup?” Responses were dichotomized as having a visit within the past 12 months (yes or no). Additional questions asked about seeing a health care provider in the past 12 months for chronic, acute, and mental and behavioral health conditions, dichotomized as (yes or no). Because measures of utilization were highly correlated ($P<.001$), they were then combined into a measure of any health care visit in the past 12 months (yes or no).

Access issues included having a primary care provider, travel time, and internet connectivity. Respondents indicated if they had one or more people they identified as a primary health care provider (yes or no). Estimated average 1-way travel time to medical appointments had 5 response options (<5 min, 5-9 min, 10-19 min, 20-29 min, and 30+ min). Household internet access was assessed overall and based on modality of mobile device and broadband (yes or no for each question).

Statistical Analysis

Analysis was conducted using SAS (version 9.4; SAS Institute Inc.). Survey weights accounting for the stratified sampling design were used for all analyses. Chi-square tests compared telehealth use based on participant sociodemographic, health, and utilization and access variables. Item nonresponse was addressed with case-wise exclusion, except when addressed by recoding. Measures of utilization were assessed individually and tested for correlation before being combined into a summary variable. We tested for correlation between urbanicity and internet access to avoid potential issues with multicollinearity in modeling. Because chronic conditions are more frequent in older persons, we tested for interaction between age group and chronic conditions. Crude and adjusted robust Poisson regression models computed prevalence odds ratios (POR) with 95% CIs. Variables significant in bivariate analysis were carried into a fully adjusted model. The full model was adjusted for age, sex,

XSL•FO
RenderX

ethnicity, education, chronic health condition, any health care visit in the past 12 months, and internet access.

Results

The response rate was 20.8% (1101/5300). Only 25.5% of adult Nebraskans had ever used telehealth, despite an estimated 97% of the population having internet access (Table 1). In the chi-square analysis, there was not a statistically significant difference in ever use of telehealth between urban (26.4%) and rural (20.8%) residents (P=.23). However, internet access and rurality were highly correlated (P<.001), with 90.5% of rural respondents reporting internet access compared to 98.3% of urban respondents. Therefore, internet access was used in modeling. Telehealth use was more common among those aged <45 years (32.4%), female individuals (30.7%), non-Hispanic individuals (25.9%), those with at least a bachelor's degree (32.6%), those who had a routine checkup (30.2%) or health care visit other than a routine checkup (34.2%), and those without any chronic health conditions (29.6%).

XSL•FO
RenderX

Table 1. Cross-sectional weighted estimation of the proportion of Nebraska adults who had ever used telehealth by March 2021.

	Unweighted responses (n=1101), n (%)	Total weighted %	Telehealth (%)		Chi-sq *P* value
Characteristics			Yes (25.5%)	No (74.5%)	
Rurality (missing=0)					.23
Urban	639 (58.04)	82.72	26.43	73.57	
Rural	462 (41.96)	17.28	20.84	79.16	
Age (missing=72)					.004
65+	462 (44.90)	18.29	13.53	86.47	
45-64	348 (33.82)	35.64	22.91	77.09	
<45	219 (21.28)	46.07	32.41	67.59	
Sex (missing=51)					.02
Male	397 (37.81)	48.63	18.99	81.01	
Female	653 (62.19)	51.37	30.72	69.28	
Race (missing=0)					.57
White only	893 (81.11)	89.28	26.12	73.88	
Others	96 (8.72)	4.25	18.92	81.08	
Not reported	112 (10.17)	6.47	20.68	79.32	
Ethnicity (missing=63)					<.001
Hispanic	54 (5.20)	3.59	3.82	96.18	
Non-Hispanic	984 (94.80)	96.41	25.93	74.07	
Marital status (missing=51)					.07
Married or couple	591 (56.29)	70.16	27.85	72.15	
Divorced, separated, widowed or never married	459 (43.71)	29.84	18.65	81.35	
Highest education completed (missing=51)					.003
Bachelor's or above	366 (34.86)	51.94	32.61	67.39	
Some college	342 (32.57)	27.74	20.75	79.25	
<High school	342 (32.57)	20.32	13.84	86.16	
Annual household income (missing=129)					.15
<$50,000	485 (49.90)	28.24	18.66	81.34	
$50,000-$74,999	177 (18.21)	17.62	26.14	73.86	
≥$75,000	310 (31.89)	54.15	29.02	70.98	
Insurance status (missing=33)					.24
Medicaid	99 (9.27)	5.69	21.54	78.46	
Medicare	441 (41.29)	18.73	15.27	84.73	
Private	461 (43.16)	67.47	29.14	70.86	
Other	39 (3.65)	5.48	24.67	75.33	
None	28 (2.62)	2.64	19.30	80.70	
Chronic health conditions (missing=3)					.03
None	266 (24.23)	36.99	18.38	81.62	
Any	832 (75.77)	63.01	29.62	70.38	

XSL•FO
RenderX

	Unweighted responses (n=1101), n (%)	Total weighted %	Telehealth (%)		Chi-sq *P* value
Routine checkup in past 12 months (missing=21)					.03
Yes	703 (65.09)	58.81	30.15	69.85	
No	377 (34.91)	41.19	18.95	81.05	
Non-routine health care visit in past 12 months (missing=0)					.002
Yes	486 (44.14)	40.58	34.22	65.78	
No	615 (55.86)	59.42	19.46	80.54	
Has a primary health care provider (missing=24)					.13
Yes	943 (87.56)	82.01	27.53	72.47	
No	134 (12.44)	17.99	16.55	83.45	
Average time 1-way for medical travel (missing=23)					.33
<5 min	159 (15.17)	11.17	12.72	87.28	
5-9 min	233 (22.23)	28.47	26.92	73.08	
10-19 min	318 (30.34)	37.91	28.42	71.58	
20-29 min	164 (15.65)	13.63	26.37	73.63	
30+ min	174 (16.60)	8.82	25.68	74.32	
Any household internet access (missing=47)					<.001
Yes	924 (87.67)	96.95	26.34	73.66	
No	130 (12.33)	3.05	2.95	97.05	
Internet via mobile device (missing=86)					.002
Yes	805 (79.31)	92.87	26.24	73.76	
No	210 (20.69)	7.13	5.30	94.70	
Internet via broadband (missing=100)					.009
Yes	735 (73.43)	87.57	27.20	72.80	
No	266 (26.57)	12.43	11.74	88.26	

In univariate models, internet access, younger age groups, female sex, non-Hispanic ethnicity, higher education, any chronic health conditions, and any care in the past year were significantly associated with ever use of telehealth (Table 2). In contrast, race, marital status, some college, income, health insurance, and having a primary care physician were not associated. Travel time was significant but had such a wide confidence interval that it was excluded from further analysis.

XSL•FO
RenderX

Table 2. Univariate prevalence odds ratios (POR) and 95% CI of ever use of telehealth, Nebraska adults, cross-sectional survey conducted from October 2020 to March 2021.

Variable	POR	95% CI	*P* value
Internet	8.94	3.28-24.38	<.001
Age group (years)			
<45 vs 65+	4.06	1.73-9.54	.03
45-64 vs 65+	6.88	1.86-25.40	<.001
Female vs male sex	1.62	1.06-2.48	.03
Race			
Other vs White	0.57	0.19-1.74	.48
Not reported vs White	0.42	0.06-2.84	.45
Non-Hispanic vs Hispanic ethnicity	6.79	2.50-18.43	<.001
Married (yes vs no)	0.67	0.42-1.06	.09
Education			
Some college vs ≤high school	3.53	1.11-11.19	.22
Bachelor vs ≤high school	5.29	0.90-31.24	.004
Income (US $)			
50,000 to <75,000 vs ≥75,000	0.58	0.27-1.25	.68
<50,000 vs ≥75,000	0.52	0.15-1.76	.05
Insurance			
Medicaid vs private	0.39	0.12-1.26	.11
Medicare vs private	0.29	0.03-2.58	.26
Other vs private	0.12	0.001-24.16	.43
None vs private	0.39	0.12-1.26	.11
Chronic condition (yes vs no)	1.61	1.03-2.52	.04
Primary care physician (no vs yes)	0.60	0.30-1.22	.16
Any health care visit (routine or otherwise) in past 12 months (yes vs no)	2.64	1.39-5.01	.003
Average time to nonemergency health care (min)			
5-9 vs <5	10.02	1.29-77.66	.04
10-19 vs <5	21.22	1.34-336.37	.02
20-29 vs <5	44.92	1.38-1464.43	.07
30+ vs <5	95.10	1.42-6391.80	.08

In multivariate modeling, chronic condition by age interaction was not significant. In the final adjusted model, education (POR 1.87, 95% CI 0.33-10.63) and sex (POR 1.38, 95% CI 0.93-2.04) were not significant, but internet access (POR 5.43, 95% CI 1.62-18.16), age <45 (POR 5.33, 95% CI 2.22-12.81) and 45-64 years (POR 9.05, 95% CI 2.37-34.62), non-Hispanic ethnicity (POR 7.40, 95% CI 2.39-22.90), any chronic condition (POR 1.73, 95% CI 1.09-2.76), and any health care appointment in the past 12 months (POR 2.43, 95% CI 1.23-4.79) were significantly associated with having ever used telehealth (Figure 1).

XSL•FO
RenderX

Figure 1. Prevalence odds ratios (95% CI) of the final survey-weighted adjusted model of ever use of telehealth in Nebraska, October 2020 to March 2021.

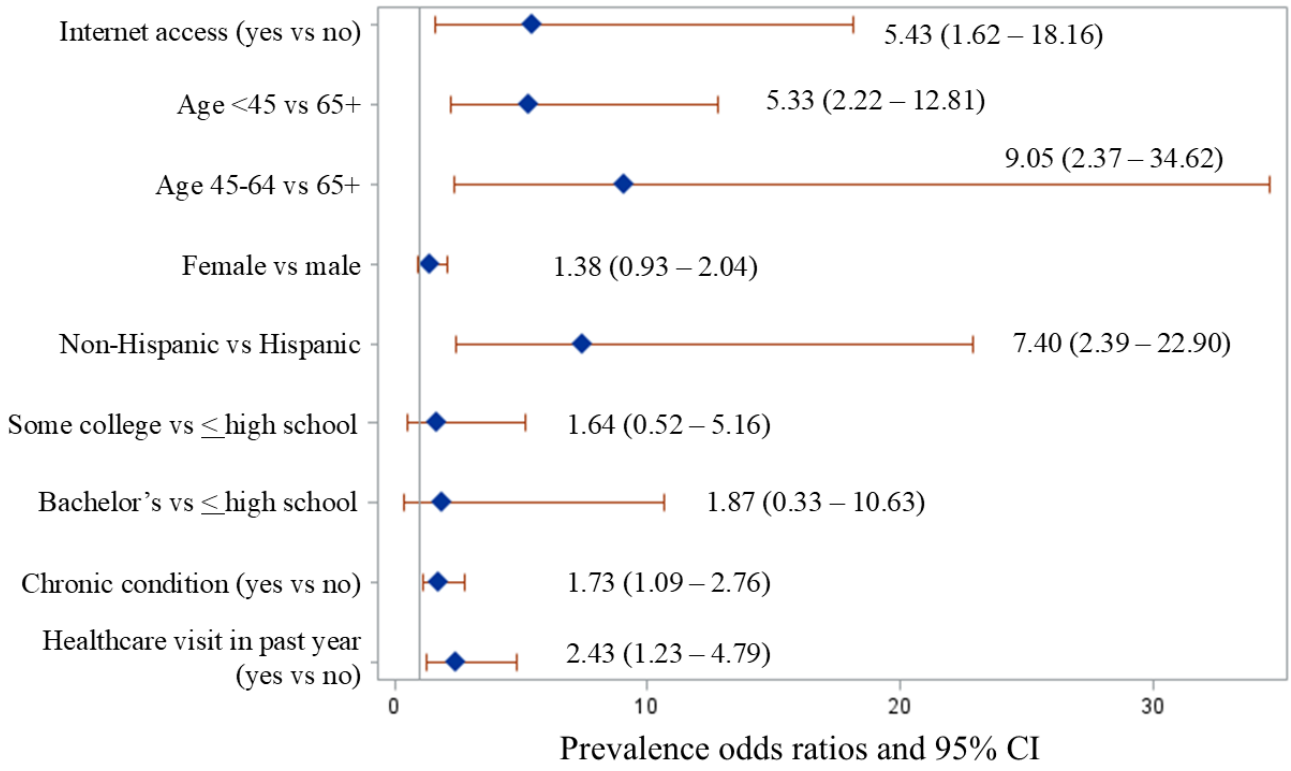

Discussion

Principal Findings

During the first year of the COVID-19 pandemic, this study focused on evaluating the utilization of telehealth in Nebraska, a primarily rural state. Although most respondents had internet access, we found that the availability and accessibility of the internet played a significant role in the utilization of telehealth services. While the estimate of adults in Nebraska who had ever used telehealth is relatively low, having internet access was a significant predictor of using those services. The study further highlighted a strong correlation between internet access and rurality, suggesting that individuals in rural areas may encounter greater barriers when trying to access telehealth services. These findings highlight the digital divide as a barrier to telehealth, with internet access being a strong predictor of telehealth usage.

In addition, disparities in telehealth usage were identified based on age, chronic health conditions, ethnicity, and health care utilization. Older adults, individuals without chronic health conditions, and people of Hispanic ethnicity were significantly less likely to have ever used telehealth compared to other demographic groups. These disparities highlight the need for targeted support to ensure equitable access and utilization of telehealth services, potentially informing strategies to address barriers related to internet access and demographic factors. Unsurprisingly, having a health care visit in the past year was associated with use of telehealth.

Our study supports prior research that found geographical location plays a part in inequities between different populations. In our study, the overall population had relatively high rates of any type of internet access (broadband or mobile network), but urban areas had better access than rural areas; our rural population reported better access than the FCC report [14], although our unweighted results were similar to the FCC report for rural areas but lower for urban areas. Similarly, another study found a positive association between increased broadband access and telemedicine use, but only in counties classified as fully rural [21]. High expense and low profit margins deter investment in expanding internet access to rural communities [22-24] and minority households [25]. Furthermore, the anticipated higher costs of internet in rural areas made it unlikely that lower economic residents could afford the high cost of the internet [26].

Consistent with some previous research, our study reinforces evidence that telehealth utilization for internet-based health care services is more prevalent among young adults [7,27,28], non-Hispanic White individuals [7,28-30], those with higher levels of education, and female individuals [7,27]. These demographic characteristics have consistently emerged as influential factors in determining telehealth usage across multiple studies. In contrast to these findings, Hong and Cho [12] found that in 2011, those aged 75+ years were more likely to have communicated with health care professionals online, and education became less significant from 2003 to 2011, indicating a narrowing of the digital divide in regard to age and education. The differences in the data are likely the result of the broad definition of communication used by Hong and Cho [12] versus our definition of telehealth as health education and medical services through telecommunications, with an emphasis on medical services.

Our research presents a contrasting perspective when compared to previous studies in the field. Other investigations have found linkage between higher socioeconomic status [7] and

XSL•FO
RenderX

Medicare/Medicaid insurance status [28] with increased telehealth utilization. These may be attributed to various factors, including greater access to the necessary technology, higher levels of health literacy, and more comprehensive insurance coverage, which includes coverage for telehealth services. However, in our analysis, these associations did not emerge as significant. This discrepancy may imply that barriers to telehealth access are addressed in ways that reduce disparities in relation to these factors. Additional research is warranted to support the demographic shift toward more equitable telehealth utilization and to explore the role that COVID-19 has played in the adoption of telehealth services.

Furthermore, our findings highlight a noteworthy association between the presence of chronic health conditions and a higher likelihood of using telehealth services. Older adults are more likely to experience chronic health conditions and less likely to access the technology necessary to use telehealth video [29], although age group interaction with chronic conditions was not significant in our study. This suggests that individuals with chronic conditions are more likely to engage in telehealth consultations compared to those without such conditions. However, Cousins et al [28] noted increased telehealth usage among older adults during the COVID-19 pandemic, while another study noted an increase in comfort among older adults with telehealth usage [31]. It is important to recognize that telehealth has the potential to support improved outcomes for individuals with chronic conditions, as it has been found to be a safe alternative to traditional care [32], enhancing patient education [33], facilitating disease management [34], and supporting self-management [32,34]. While it may not completely replace in-person visits, telehealth can supplement traditional care by facilitating remote monitoring, providing timely access to health care professionals, and enabling convenient follow-up consultations [6].

Disparities in access to telehealth services exist among older adults, marginalized racial or ethnic communities, and rural residents. Familiarity with communication devices and internet usage can also impose limitations on telehealth access, particularly for older adults. In addition, replacing a percentage of traditional health care services with telehealth access may exacerbate existing health disparities by favoring younger, employed, and urban-dwelling patients [35].

In the broader context of our findings, the emphasis on Nebraska underscores the need for tailored interventions to address telehealth disparities. The state's considerable rural demographic provides a valuable perspective on the unique challenges and opportunities telehealth offers in rural America. Rural areas face unique challenges, including access to health care, education, and economic opportunities, which are indicative of the declining populations in many rural counties. By focusing on Nebraska, we illustrate the pressing need for enhancements in telehealth infrastructure, policy adjustments, and community centered initiatives to eliminate barriers and improve telehealth access. Our investigation into Nebraska's telehealth landscape offers valuable insights into the challenges and resilience of rural communities, not just in Nebraska but across the Great Plains and the broader United States.

Limitations

In this study, sampling weights accounted for the oversampling of certain census tracts and were used to make the responses representative of the Nebraska population by sex and age. However, Nebraska lacks racial diversity (non-Hispanic White residents represent 77% of the population), and racial or ethnic groups are not evenly distributed across the state. Although we attempted to obtain diverse responses, we were unable to apply sampling weights for race. Weighting actually diluted the responses for certain groups. Therefore, we lacked the diversity that we intended and were unable to make more in-depth comparisons beyond White (yes vs no). We specifically listed the reference group as "White only" because respondents were allowed to select multiple racial categories to self-identify. The absence of race-based sampling weights means that the sample may not accurately reflect the racial composition of the population, limiting the study's external validity. Furthermore, the weighting process employed in the study appeared to dilute responses from rural areas due to oversampling. Approximately 35% of Nebraska's population is rural, but 42% of our respondents were rural and 17% were from small urban locations. Given that rural areas often possess unique characteristics, such as different socioeconomic conditions, limited access to resources, and specific cultural factors, the dilution of responses from these regions can compromise the study's ability to capture the full range of perspectives and experiences. Future research should aim to address these limitations by employing more comprehensive sampling strategies that account for a wider range of demographic variables and ensure adequate representation across various geographic locations. Another limitation of the study revolved around the issue of item nonresponse, which further diluted our multivariate models. This issue may weaken the reliability of our multivariate models. The number of health care visits for chronic, acute, and mental and behavioral health conditions was recorded as a write-in response, which resulted in potential misclassification; several respondents reported more than 30 visits in the past 12 months. Therefore, we dichotomized as any versus none for analysis. Future studies may want to consider the types of visits and frequency of visits for their impact on exposure to telehealth as a service delivery method.

Conclusions

Our investigation into telehealth utilization during the onset of the COVID-19 pandemic shows the need for targeted strategies to bridge the digital divide and enhance telehealth access, especially in rural areas. At a time when the pandemic dramatically increased the need for telehealth services, serving as an essential tool for patients unable to visit medical professionals in person, our study reveals that disparities in telehealth usage were already present. The challenges identified in Nebraska's communities highlight how demographic factors, such as age, chronic health conditions, ethnicity, and especially internet accessibility, affect telehealth utilization, highlighting the preexisting barriers to equitable health care delivery. These findings further emphasize the continued need for policy and infrastructure reform to address these disparities. Moving forward, it is important to use these insights to guide the development of inclusive telehealth programs and policies that

XSL•FO
RenderX

are sensitive to the diverse needs of all populations, ensuring that telehealth can fulfill its potential as a cornerstone of accessible and equitable health care.

Data Availability

The dataset used and analyzed during the current study is available from the corresponding author on reasonable request.

Authors' Contributions

KLR, LS, and HW contributed to conceptualization and study design; KLR contributed to data acquisition; KLR and SJ contributed to data analysis; all authors contributed to data interpretation; KLR, LS, GJ, and JK contributed to writing – original draft preparation, and writing – review and editing (all authors); and KLR contributed to funding and study supervision. All authors read and approved the final manuscript and agree to hold themselves jointly accountable for it. This work was supported by a University of Nebraska Medical Center Faculty Diversity grant (unnumbered) to KLR. The content is solely the responsibility of the authors and does not necessarily represent the official views of UNMC.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Methodology report from the Bureau of Sociological Research (BOSR).
[PDF File (Adobe PDF File), 4728 KB - jmir_v26i1e53320_app1.pdf]

References

1. Telehealth, telemedicine, and telecare: what's what? Federal Communications Commission. URL: https://www.fcc.gov/general/telehealth-telemedicine-and-telecare-whats-what [accessed 2023-09-08]
2. Murphy RL, Bird KT. Telediagnosis: a new community health resource. Observations on the feasibility of telediagnosis based on 1000 patient transactions. Am J Public Health 1974 Feb;64(2):113-119. [doi: 10.2105/ajph.64.2.113] [Medline: 4129472]
3. Manocchia A. Telehealth: enhancing care through technology. R I Med J (2013) 2020 Feb 03;103(1):18-20 [FREE Full text] [Medline: 32013298]
4. Carini E, Villani L, Pezzullo AM, Gentili A, Barbara A, Ricciardi W, et al. The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: updated systematic literature review. J Med Internet Res 2021 Sep 08;23(9):e26189 [FREE Full text] [doi: 10.2196/26189] [Medline: 34494966]
5. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022 Dec;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
6. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
7. Din HN, McDaniels-Davidson C, Nodora J, Madanat H. Profiles of a health information-seeking population and the current digital divide: cross-sectional analysis of the 2015-2016 California Health Interview Survey. J Med Internet Res 2019 May 14;21(5):e11931 [FREE Full text] [doi: 10.2196/11931] [Medline: 31094350]
8. van Dijk J, Hacker K. The digital divide as a complex and dynamic phenomenon. Inf Soc 2003 Sep;19(4):315-326. [doi: 10.1080/01972240309487]
9. Mackert M, Mabry-Flynn A, Champlin S, Donovan EE, Pounders K. Health literacy and health information technology adoption: the potential for a new digital divide. J Med Internet Res 2016 Oct 04;18(10):e264 [FREE Full text] [doi: 10.2196/jmir.6349] [Medline: 27702738]
10. Mishori R, Antono B. Telehealth, rural America, and the digital divide. J Ambul Care Manage 2020;43(4):319-322. [doi: 10.1097/JAC.0000000000000348] [Medline: 32858734]
11. Vogels E. Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. 2021. URL: https://www.pewresearch.org/short-reads/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/ [accessed 2024-10-22]
12. Hong YA, Cho J. Has the digital health divide widened? Trends of health-related internet use among older adults from 2003 to 2011. J Gerontol B Psychol Sci Soc Sci 2017 Sep 01;72(5):856-863. [doi: 10.1093/geronb/gbw100] [Medline: 27558403]
13. Jiang S, Liu PL. Digital divide and Internet health information seeking among cancer survivors: a trend analysis from 2011 to 2017. Psychooncology 2020 Jan;29(1):61-67. [doi: 10.1002/pon.5247] [Medline: 31652360]
14. Fourteenth broadband deployment report. Federal Communications Commission. Washington, DC; 2021. URL: https://docs.fcc.gov/public/attachments/FCC-21-18A1.pdf [accessed 2024-10-22]

XSL•FO
RenderX

15. Bates K, Malakoff L, Kane S, Pulidindi J. Closing the digital divide: promoting broadband adoption among underserved populations. ICF International, Port of Clarkston. 2012. URL: https://portofclarkston.com/wp-content/uploads/2020/11/Benefits-of-Broadband.pdf [accessed 2024-10-22]
16. Sanders CK, Scanlon E. The digital divide is a human rights issue: advancing social inclusion through social work advocacy. J Hum Rights Soc Work 2021;6(2):130-143 [FREE Full text] [doi: 10.1007/s41134-020-00147-9] [Medline: 33758780]
17. Profiles: Nebraska counties. U.S. Census Bureau. URL: https://data.census.gov/profile?q=county%20population%20in%20nebraska [accessed 2024-02-16]
18. Tak H, Chakraborty B, Carritt N, Palm DW, Deras M, Horner RD. The status of the Nebraska healthcare workforce: 2022 update. University of Nebraska Medical Center. 2022 Apr. URL: https://www.unmc.edu/rural-health/_documents/healthcare_workforce_status_2022.pdf [accessed 2024-10-21]
19. Health workforce shortage areas. Health Resources and Services Administration. 2024. URL: https://data.hrsa.gov/topics/health-workforce/shortage-areas [accessed 2024-02-15]
20. Ratnapradipa KL, Jadhav S, Kabayundo J, Wang H, Smith LC. Factors associated with delaying medical care: cross-sectional study of Nebraska adults. BMC Health Serv Res 2023 Feb 04;23(1):118 [FREE Full text] [doi: 10.1186/s12913-023-09140-0] [Medline: 36739376]
21. Wilcock AD, Rose S, Busch AB, Huskamp HA, Uscher-Pines L, Landon B, et al. Association between broadband internet availability and telemedicine use. JAMA Intern Med 2019 Nov 01;179(11):1580-1582 [FREE Full text] [doi: 10.1001/jamainternmed.2019.2234] [Medline: 31355849]
22. Rendon Schneir J, Xiong Y. A cost study of fixed broadband access networks for rural areas. Telecommun Policy 2016 Aug;40(8):755-773 [FREE Full text] [doi: 10.1016/j.telpol.2016.04.002]
23. Riddlesden D, Singleton AD. Broadband speed equity: a new digital divide? Appl Geogr 2014 Aug;52:25-33 [FREE Full text] [doi: 10.1016/j.apgeog.2014.04.008]
24. Pereira JPR. New advances in information systems and technologies. In: Rocha Á, Correia A, Adeli H, Reis L, Mendonça Teixeira M, editors. Advances in Intelligent Systems and Computing. Cham, Switzerland: Springer; 2016:445.
25. Reddick CG, Enriquez R, Harris RJ, Sharma B. Determinants of broadband access and affordability: an analysis of a community survey on the digital divide. Cities 2020 Nov;106:102904 [FREE Full text] [doi: 10.1016/j.cities.2020.102904] [Medline: 32921864]
26. Grubesic TH, Murray AT. Waiting for broadband: local competition and the spatial distribution of advanced telecommunication services in the United States. Growth Change 2004;35(2):139-165 [FREE Full text] [doi: 10.1111/j.0017-4815.2004.00243.x]
27. Kontos E, Blake KD, Chou WYS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
28. Cousins MM, van Til M, Steppe E, Ng S, Ellimoottil C, Sun Y, et al. Age, race, insurance type, and digital divide index are associated with video visit completion for patients seen for oncologic care in a large hospital system during the COVID-19 pandemic. PLoS One 2022;17(11):e0277617 [FREE Full text] [doi: 10.1371/journal.pone.0277617] [Medline: 36395112]
29. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 01;180(10):1389-1391 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
30. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 09;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
31. Bhatia R, Gilliam E, Aliberti G, Pinheiro A, Karamourtopoulos M, Davis RB, et al. Older adults' perspectives on primary care telemedicine during the COVID-19 pandemic. J Am Geriatr Soc 2022 Dec;70(12):3480-3492 [FREE Full text] [doi: 10.1111/jgs.18035] [Medline: 36169152]
32. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
33. Rush KL, Hatt L, Janke R, Burton L, Ferrier M, Tetrault M. The efficacy of telehealth delivered educational approaches for patients with chronic diseases: a systematic review. Patient Educ Couns 2018 Aug;101(8):1310-1321. [doi: 10.1016/j.pec.2018.02.006] [Medline: 29486994]
34. Darkins A, Kendall S, Edmonson E, Young M, Stressel P. Reduced cost and mortality using home telehealth to promote self-management of complex chronic conditions: a retrospective matched cohort study of 4,999 veteran patients. Telemed J E Health 2015 Jan;21(1):70-76. [doi: 10.1089/tmj.2014.0067] [Medline: 24841071]
35. Hlavin C, Ingraham P, Byrd T, Hyre N, Gabriel L, Agrawal N, et al. Clinical outcomes and hospital utilization among patients undergoing bariatric surgery with telemedicine preoperative care. JAMA Netw Open 2023 Feb 01;6(2):e2255994 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.55994] [Medline: 36763357]

XSL•FO
RenderX

Abbreviations

BOSR: Bureau of Sociological Research
FCC: Federal Communications Commission
POR: prevalence odds ratio

Edited by A Mavragani; submitted 03.10.23; peer-reviewed by X Du, B Ward; comments to author 03.02.24; revised version received 22.03.24; accepted 09.10.24; published 28.11.24.

Please cite as:
Smith LC, Johnson G, Jadhav S, Kabayundo J, Ahuja M, Wang H, Ratnapradipa KL
Ever Use of Telehealth in Nebraska by March 2021: Cross-Sectional Analysis
J Med Internet Res 2024;26:e53320
URL: https://www.jmir.org/2024/1/e53320
doi:10.2196/53320
PMID:

©Lisa C Smith, George Johnson, Snehal Jadhav, Josiane Kabayundo, Muskan Ahuja, Hongmei Wang, Kendra L Ratnapradipa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Patient Organizations' Digital Responses to the COVID-19 Pandemic: Scoping Review

Simon Wallraf[1], MSc; Marie-Luise Dierks[1], PhD; Cosima John[1], MSc; Jonas Lander[1], PhD

Institute for Epidemiology, Social Medicine and Health Systems Research, Hannover Medical School, Hanover, Germany

Corresponding Author:
Simon Wallraf, MSc
Institute for Epidemiology, Social Medicine and Health Systems Research
Hannover Medical School
Carl-Neuberg-Str. 1
Hanover, 30625
Germany
Phone: 49 511 532 4038
Email: Wallraf.Simon@mh-hannover.de

Abstract

Background: Patient organizations (POs) play a crucial role in supporting individuals with health conditions. Their activities range from counseling to support groups to advocacy. The COVID-19 pandemic and its related public health measures prompted rapid digital transformation efforts across multiple sectors, including health care.

Objective: This study aimed to explore how POs digitally responded to pandemic-related circumstances, focusing on aspects such as the technologies used, positive outcomes, and challenges encountered.

Methods: This scoping review followed the methodological guidance of the JBI (Joanna Briggs Institute) Scoping Review Methodology Group and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guidelines. A systematic search of PubMed, the Web of Science Core Collection, and the WHO (World Health Organization) COVID-19 database, supplemented by a citation search approach, was conducted. The initial search was performed on November 10, 2022, and updated on November 8, 2023. Publications were eligible if they were published after November 30, 2019, and addressed pandemic-related digitalization efforts of POs, defined as nonprofit organizations with a focus on health-related support. A 2-step screening process was used to identify relevant literature. Data were extracted using a standardized table to capture aspects such as digital adaptation activities (eg, types of technologies implemented, positive outcomes, challenges, and facilitating factors) and coded inductively to identify similarities across included publications, and the findings were synthesized narratively.

Results: The search and its subsequent update yielded 2212 records, with 13 articles included in this review. These articles revealed a range of PO services that were digitally adapted during the pandemic, with videoconferencing software emerging as the most commonly used tool (n=9 articles). The digital adaptation of group-based support activities was the most frequently reported transformation (n=9). Other adaptations included the digitalization of counseling services (n=3) and the delivery of information and education (n=3), including educational workshops, weekly webinars, and the dissemination of information through digital newsletters. While the use of digital formats, particularly for POs' group activities, often increased accessibility by breaking down preexisting barriers (n=5), they also created new barriers for certain groups, such as those lacking digital skills or resources (n=4). Some participants experienced a loss of interpersonal aspects, like a sense of community (n=3). However, further findings suggest that the digital delivery of such group activities preserved essential interpersonal aspects (n=7) and a preference among some participants to continue digital group activities (n=4), suggesting the potential for sustainability of such options post the COVID-19 pandemic.

Conclusions: The rapid digitalization efforts of POs demonstrate their adaptability and the potential of digital technologies to improve support services, despite some challenges. Future digitalization strategies should focus, among other things, on promoting digital literacy to ensure the accessibility and inclusiveness of digital services.

Trial Registration: OSF Registries, https://osf.io/anvf4

(J Med Internet Res 2024;26:e58566) doi:10.2196/58566

XSL•FO
RenderX

KEYWORDS

patient organizations; COVID-19; digital adaptation; digital transformation; scoping review

Introduction

Patient organizations (POs) provide crucial support to individuals affected by health conditions such as chronic diseases, and to their relatives. Support activities can range from direct services, such as individual counseling, support groups, and health education programs, to more indirect forms of support, such as advocacy in research or policy making [1-5]. In addition to differences in their primary purpose and scope of activities, POs also differ in organizational characteristics such as their internal structures and the financial resources available to them, adding to the complexity of the PO landscape. In Germany, for instance, Kofahl et al [3] found that almost half of the POs (47%) are run entirely by volunteers, while others have the capacity to employ full-time staff, ranging from 1 employee (17%) to 5 or more (15%). Membership sizes also vary widely, from a few individuals to more than 50,000 [3]. In the United States, Rose et al [4] found considerable financial variation among POs. While some highly professionalized POs reported annual revenues exceeding US $1 million, the median revenue for American organizations was US $299,140, suggesting that many operate on a smaller financial scale. This diversity among POs, in terms of their scope, structure, size, and financial resources, is also reflected in the lack of a universal definition, with terms such as “patient association,” “patient advocacy organization,” “self-help organization,” and “voluntary health agency” being used interchangeably [2-4]. To ensure readability and coherence in this review, we use “patient organization” as an umbrella term to encompass all such entities. Despite this broad terminology and considerable diversity, the majority of POs share 2 fundamental attributes. First, they mostly operate within a nonprofit framework, a characteristic frequently underscored in the definitions of these organizations [4,5]. It should be noted, however, that certain POs may have a legal structure that enables for-profit or hybrid models. Second, a unifying trait across this diverse spectrum is their commitment to supporting their respective target groups, underscoring the core mission that defines POs, as described before.

However, the onset of the COVID-19 pandemic drastically changed the circumstances of interpersonal interactions, often making it impossible to provide any kind of face-to-face service, as governments worldwide enacted public health protection measures that often included restrictions on movement and direct contact [6]. These policies triggered rapid digital transformations in various social domains, including work, education, business, and health care [7-9]. In health care, for example, several recent reviews have demonstrated a rapid significant increase in the use and adoption of telemedicine and eHealth apps [10-13].

Given the widespread impact of the pandemic and the resulting surge in the use of digital technologies, it is reasonable to assume that POs underwent similar digital adaptation efforts, changing their operational approaches and support delivery. The extent and effectiveness of these pandemic-induced digital adaptations likely varied across POs, reflecting not only the diversity described above but also varying levels of prepandemic digitalization. Before the pandemic, research indicated that the level of digitalization among POs varied widely. While many relied primarily on basic digital communication tools such as websites, email, and social media, others embraced advanced technologies by developing, for example, mobile apps that allow user interaction, web-based exercise programs, or electronic patient databases to enhance the use of digital patient data for research [14-16]. These differences in digitalization levels may be linked to the organizational characteristics described above, as well as additional factors such as the digital literacy of PO leadership and members and their demographic profiles [14].

However, despite the significant role of POs as supporters and advocates, we found no comprehensive reviews synthesizing the existing research literature on their digitalization efforts during the COVID-19 pandemic. While several reviews have addressed the use of digital technologies in health care during the pandemic [17-21], the specific focus on POs remains conspicuously absent. This gap leads us to our central research question, that is, how have POs adapted digitally in response to the COVID-19 pandemic?

To specifically address certain aspects of these activities, we pose the subquestions: (1) What digital technologies were adopted by POs in response to the pandemic? (2) What were the positive outcomes, challenges, and facilitating factors associated with these digital transformation efforts? Our findings on these matters may also eventually provide valuable insights for POs and similar organizations to guide future digitalization strategies and improve digital readiness for such unprecedented events, such as pandemics or situations where direct physical contact is restricted.

Methods

Methodological Framework

This scoping review is based on the methodological guidance developed by the JBI (Joanna Briggs Institute) Scoping Review Methodology Group [22]. For reporting, we adhered to PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [23]. The completed PRISMA-ScR checklist is provided in Multimedia Appendix 1.

Protocol and Registration

A protocol was developed and registered with the Open Science Framework before the conduct of this review [24]. Given the iterative nature of scoping reviews, deviations from the initial protocol are generally possible [22,25]. All deviations are documented in the Multimedia Appendix 2 for transparency and justification.

Eligibility Criteria

We established a set of specific criteria to systematically identify eligible articles among our search results (Table 1). To be included, an article had to meet all of the criteria, which we

XSL•FO
RenderX

divided into 2 main types, formal aspects (criteria 1 to 5) and content-related aspects (criteria 6 and 7).

Table 1. Eligibility criteria.

Criterion category	Specific criterion
Formal	1. The full text of the article is available.
	2. The article is published in English or German.
	3. The article was published after November 30, 2019.
	4. The article is not a duplicate publication.
	5. All types of peer-reviewed articles[a] published in a scientific journal, regardless of study design or methodological approach, are eligible.
Concept (PCC[b] Component)	6. Eligible articles must report on digitalization efforts or initiatives within the PO[c] that have been initiated, accelerated, or expanded to adapt to the COVID-19 pandemic, which may include the adoption, development, or use of any digital technology for PO activities or services.
Context (PCC Component)	7. Eligible articles must describe the efforts and initiatives of non-profit POs that provide direct or indirect support to individuals affected by health conditions[d]. Support groups will be included if they are affiliated with a PO. No geographic limitation was applied.

[a]In this review, study protocols and conference proceedings were not considered as full journal articles due to the preliminary or limited information they provide.

[b]PCC: Population, Concept, and Context.

[c]PO: patient organization.

[d]In this review, POs are defined as organizations separate from entities such as government agencies, faith-based groups, and academic institutions that may be non-profit and offer health-related support.

Regarding the latter, they correspond to the "Concept" and "Context" elements of our research question, derived from the Population, Concept, and Context (PCC) framework, which is often used to construct the title and main research questions of a scoping review [22,25] (Multimedia Appendix 3). Accordingly, articles were considered eligible if they addressed digitalization efforts initiated by POs during and in response to the COVID-19 pandemic. We defined digitalization efforts as the adoption, development, or use of digital technologies for activities previously conducted in person, or as the scaling up of existing digital services. To ensure a comprehensive overview, we considered the structural and functional variations of POs when defining our inclusion criteria. We defined POs by 2 main characteristics, first, their commitment to supporting individuals affected by specific health conditions. This may include direct support, like counseling or peer support, as well as indirect support, such as advocacy or research funding. Articles that addressed the digitalization of peer support groups were also included if such groups were associated with a PO. The second defining characteristic was their nonprofit nature, which is predominant in this sector.

We deliberately chose this broader definition, which aligns with the common characteristics outlined in the introduction, to account for the diverse international landscape of POs while ensuring a cohesive and comparable data set to analyze their digital responses to the pandemic. We intentionally excluded for-profit organizations, as their different operational and financial structures would complicate meaningful comparisons with nonprofits. This distinction ensures methodological consistency and precision in our analysis while recognizing the diversity of these organizations.

Regarding the former (formal inclusion criteria), these included article type, date, and language. We used these criteria to establish a consistent, transparent, and reproducible selection process. Acknowledging that the scope of published literature has not been previously summarized, we adopted a broad approach for certain criteria to ensure comprehensive coverage. For instance, we did not impose limitations on the type of article or study design.

Information Sources and Search Strategy

To ensure that our systematic literature search was comprehensive, we selected the bibliographic databases PubMed, Web of Science Core Collection, and the WHO (World Health Organization) COVID-19 Research Database as sources of information. The search strategy was primarily developed by SW, who relies on specialized training and extensive experience in developing search strategies and conducting systematic searches gained through a previous professional position. It was first developed for PubMed and then adapted for the other databases. It consists of 3 thematic search blocks, each containing different synonymous free-text terms and, if available in the databases, corresponding controlled vocabulary terms, that consist of (1) patient organizations, (2) digitalization and digital technologies, and (3) the COVID-19 pandemic. The creation of the third search block was based on PubMed's general COVID-19 article filter "LitCovid" [26]. Where possible, a publication language restriction was included in the search strings (refer to Multimedia Appendix 3 for the search strings). Quality assurance of the search strategy was performed by the other authors to ensure its robustness and comprehensiveness. The initial search was performed on November 10, 2022, and updated on November 8, 2023. Citavi 6 (Swiss Academic Software GmbH) was used for literature management. To further increase the comprehensiveness of our search and to identify additional eligible articles, a citation search was conducted.

XSL•FO
RenderX

Selection Process

During the initial stage of our selection process, we used Citavi's duplicate detection feature. Each potential duplicate was thoroughly reviewed by SW using criteria such as title, author, DOI, and journal. Only confirmed exact duplicates were removed, to ensure the integrity of the data set. To then systematically identify all relevant articles, we adopted a 2-step selection process. First, SW and CJ independently screened the titles and abstracts of all articles to determine eligibility based on our inclusion criteria (Table 1). When both reviewers considered an article potentially eligible, it was forwarded to full-text screening for further evaluation. Those that did not meet the eligibility criteria were excluded, with the first unmet criterion noted as the reason for exclusion. Conflicting assessments were discussed. If a consensus could not be reached, the article was also forwarded to full-text screening. Full texts of all preselected articles were then obtained and independently reviewed by SW and JL for final inclusion according to our inclusion criteria (Table 1). Conflicting assessments were discussed to reach a consensus. We did not need to involve a third independent reviewer in the selection process, as all conflicts were resolved at this stage. The screening process during the search update was also carried out using the 2-stage selection approach and was performed by SW and JL. We used Rayyan (Rayyan Systems Inc), a web-based software application designed to facilitate literature screening in systematic reviews, to conduct the selection process for both the initial and the update searches. The citation search was also performed independently by 2 reviewers (SW and CJ). It involved screening the reference lists of the included articles and searching Google Scholar for articles that cited them. The 2-step selection process was followed. The citation search approach was also applied to articles identified during the update of the systematic search, which was conducted by SW and JL.

Quality Assessment

A critical appraisal of the methodological quality of the included articles was not performed. This is in line with the standard approach for scoping reviews to provide an overview of the existing sources of evidence on a particular topic, regardless of their methodological quality [25].

Data Charting

To extract the relevant data from each included article, we developed a standardized table aligned with the PCC elements of our primary research question and subquestions. The extracted data included bibliographic information (eg, first author, title, and year of publication), methodological information (eg, study design and methodological approach), information on the POs (eg, type of entity, its specific health condition, and group represented), information on the digital adaptation activities (eg, types of digital technologies implemented, positive outcomes and facilitating factors identified, and challenges encountered).

Given the potential diversity in reporting styles, measures, and depth, and to maintain transparency and coherence, we focused our extraction on information explicitly linked to digital adaptation activities. This included aspects like the transformation of in-person services, members' experiences with online support group activities, or the expansion of existing digital services. If a link to such efforts was not clearly identifiable, we excluded that information from the data extraction. For example, a study reporting a lack of funding as a major challenge for POs during the pandemic was not included unless the challenge was explicitly related to funding for such digitalization measures. To ensure consistency and accuracy, SW and JL collaboratively extracted the data, with each handling half of the included articles. The data was initially extracted verbatim so as not to miss important details. This approach included a range of data, from concise statements to entire paragraphs. Subsequently, the extracted information was condensed to distill the main points and to facilitate focused analysis.

Analysis and Data Synthesis

After condensing the textual data, we proceeded with inductive coding. Thematic codes were assigned to the condensed data to reflect the primary information reported. These codes were then organized into broader categories. Whenever possible, these codes, as well as the corresponding broader categories, were applied across the data of all included articles, to allow for a structured and comprehensive analysis.

In the narrative synthesis of the data, which forms the core of our results section, we aimed to highlight similarities and differences among the included articles and to describe potential trends using a slightly quantitative approach by presenting the frequency of reporting (eg, regarding the digital technologies used or the services that had been digitally transformed).

To achieve this, we first summarized the characteristics of the included articles, followed by information on the POs highlighting key aspects and variations. The reported characteristics of the digital adaptation efforts were then synthesized, with a focus on identifying and describing common trends and approaches.

Results

Search and Selection of Articles

The systematic search of bibliographic databases, along with its update, yielded a total of 2212 results after the removal of duplicates. Of these, 10 references met the eligibility criteria and were included in our review. The citation search yielded 3 additional eligible references. Thus, after the completion of all screening steps, 13 articles were finally included. Figure 1 displays the selection process during the initial search using a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram [27]. The search update is shown in Figure 2.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the selection process during the initial search. WHO: World Health Organization.

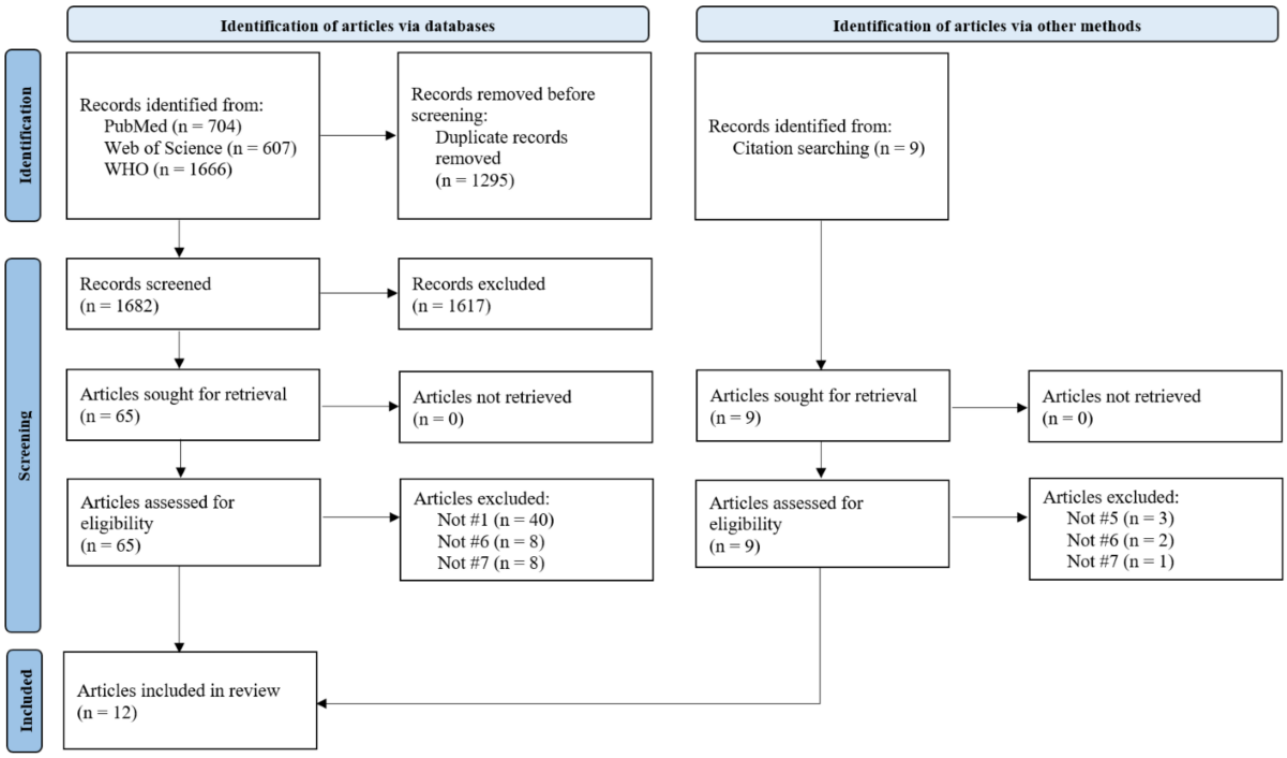

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the selection process during the update search. WHO: World Health Organization.

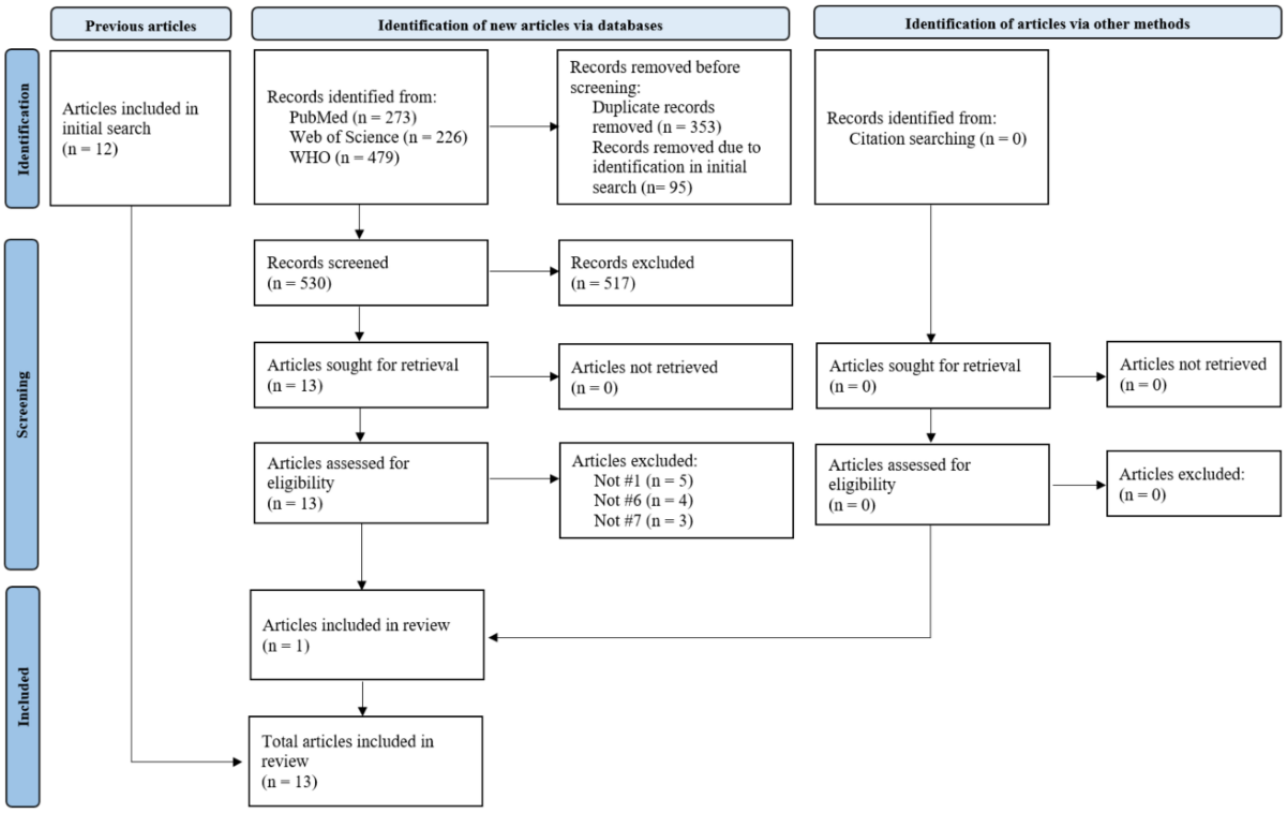

Characteristics of Included Articles

Of the 13 included references, 9 were original research articles. Of these, most were empirical studies employing a qualitative (n=4) or mixed methods approach (n=4). Of the remaining 4 articles, 1 was based on a survey but was published as a data note due to its unexpectedly small sample size [28]. The other 3 articles can be broadly categorized as experience reports, as they aimed to describe practical examples or lessons learned

XSL•FO
RenderX

regarding digital transformation, without collecting data in this context [29-31].

Most of the included articles were published during the first 2 years of the COVID-19 pandemic. Of the 3 included articles published in 2023, Bouey et al [32] was initially identified as a preprint, flagged as relevant, and included when it was fully published. Constantini et al [33] was identified through the search update.

The majority of the included articles (n=8) reported receiving financial support from a PO or indicated varying degrees of involvement by representatives of a PO [29-32,34-36], or an umbrella organization of POs [37]. Table 2 presents the objectives of the included articles, while Table 3 outlines their general characteristics. Also refer to Multimedia Appendix 4.

Table 2. Objectives of included articles (N=13).

Reference	Objective
Beck et al [38]	Evaluate SMART[a] (Self-Management and Recovery Training) Recovery Australia's scale-up of online support groups during the COVID-19 pandemic.
Bergmans et al [29]	Present lessons learned from transitioning the Skills for Safer Living intervention to a digital format.
Bouey et al [32]	Explore the challenges Chinese POs[b] faced during the first COVID-19 lockdown and how they responded.
Chung et al [37]	Examine the impact of the pandemic on rare disease POs in the Asia-Pacific region.
Constantini et al [33]	Compare the experiences of members of support groups that have transitioned to a digital format with general findings from the literature on online support groups.
Kelly et al [30]	Discuss the global impact of COVID-19 on SMART Recovery International with facilitators from different regions.
Lamont et al [36]	Explore stroke survivors' perceptions of social support and shared identity within stroke groups during COVID-19 and its impact on psychosocial health.
Marks et al [35]	Investigate the effectiveness and experiences of participants in online tinnitus support groups and educational webinars implemented during COVID-19.
McMullan et al [28]	Examine the impact of the pandemic on rare disease POs in Ireland and the United Kingdom.
Nemeth Blažić et al [31]	Describe, among other things, the digitalization of a PO's voluntary counseling and testing services for HIV, hepatitis C, and other sexually transmitted infections during the pandemic.
Penfold and Ogden [39]	Explore Gamblers Anonymous members' experiences with digitally delivered group meetings during the pandemic to understand the effectiveness and support provided compared to in-person meetings.
Seckam and Hallingberg [34]	Examine stroke survivors' experiences with the transition from in-person to digital choir sessions during COVID-19.
Senreich et al [40]	Explore 12-step program attendees' experiences with in-person group meetings transitioning to digital formats.

[a]SMART: Self-Management and Recovery Training.

[b]PO: patient organization.

XSL•FO
RenderX

Table 3. Characteristics of included articles (N=13).

Characteristics	Articles, n (%)	References
Publication year[a]		
2021	5 (38)	[28-30,34,37]
2022	5 (38)	[31,35,36,39,40]
2023	3 (23)	[32,33,38]
Article type		
Data note	1 (8)	[28]
Report	3 (23)	[29-31]
Research article	9 (69)	[32-40]
Methodological approach		
Mixed-method	4 (31)	[32,33,37,38]
Nonempirical	3 (23)	[29-31]
Qualitative	4 (31)	[34,35,39,40]
Quantitative	2 (15)	[28,36]
PO[b] **involvement**[c]		
Funding by PO	1 (8)	[36]
Involvement of PO representatives	5 (38)	[29-32,34]
Involvement of representatives from an umbrella organization of POs	1 (8)	[37]
PO representative involvement and PO funding	1 (8)	[35]

[a]Percentages in the "Publication year" category do not exactly total 100 due to rounding.

[b]POs: patient organizations.

[c]The category "PO involvement" does not contain all included publications, as not every article reported a form of involvement with a patient organization.

Characteristics of Patient Organizations

Overview

The majority of the included articles (n=8) addressed POs involved in a broad range of activities, including direct support services, advocacy, or research participation [28,29,31,32,34-37]. Meanwhile, 5 articles reported on POs, such as SMART Recovery (Self-Management and Recovery Training), that primarily provide peer support groups [30,33,38-40]. Of all included articles, 8 specifically explored the digital adaptation of POs' group-based support activities without addressing other organizational elements [29,30,33,34,36,38-40]. Moreover, a focus on well-established POs such as SMART Recovery [30,38], the Stroke Association [34,36], and the Canadian Mental Health Association [29] was predominant among the included articles.

The majority of included articles focused on POs within specific countries. Most articles (n=8) reported on individual POs operating within a single country [29,31,33-36,38,39], while 2 explored multiple POs within the same national context [32,40]. In addition, 2 articles expanded their scope to include POs in a transnational region [28,37], and another provided an international overview by focusing on SMART Recovery International and its network of global affiliates [30]. Among all the articles, POs based in the United Kingdom were the most commonly reported on (n=5).

In terms of the health conditions addressed by the POs, addictive behaviors were the most common (n=4), followed by stroke (n=2), infections with HIV, AIDS or other sexually transmitted infections (n=2), and rare diseases (n=2). Table 4 provides an overview of aspects related to the PO settings, as reported in the included articles. Also refer to Multimedia Appendix 4.

XSL•FO
RenderX

Table 4. Characteristics of Patient organizations as reported in the included articles (N=13)[a].

Characteristics	Articles, n (%)	References
Indication		
Addictive behaviors	4 (31)	[30,38-40]
Dementia	1 (8)	[33]
HIV, AIDS, and other sexually transmitted infections[b]	2 (15)	[31,32]
Mental health	1 (8)	[29]
Rare diseases	2 (15)	[28,37]
Stroke	2 (15)	[34,36]
Tinnitus	1 (8)	[35]
Geographic location		
Australia	2 (15)	[37,38]
Austria	1 (8)	[33]
Canada	1 (8)	[29]
China	2 (15)	[32,37]
Croatia	1 (8)	[31]
Hong Kong	1 (8)	[37]
India	1 (8)	[37]
Ireland	1 (8)	[28]
Japan	1 (8)	[37]
Malaysia	1 (8)	[37]
New Zealand	1 (8)	[37]
Philippines	1 (8)	[37]
Singapore	1 (8)	[37]
Taiwan	1 (8)	[37]
Transnational (PO[b] with affiliations in Australia, Brazil, Denmark, Hong Kong, Ireland, Malaysia, Spain, United Kingdom, and United States)	1 (8)	[30]
United Kingdom	5 (38)	[28,34-36,39]
United States	1 (8)	[40]
Digitally adapted PO[c] activities and services during the COVID-19 pandemic		
Group-based support activities	9 (69)	[29,30,33-36,38-40]
Communication and counseling	4 (31)	[28,31,32,37]
Information provision and educational activities	3 (23)	[28,35,37]
Operational activities	2 (15)	[28,37]

[c]Articles may appear in multiple categories within this table, as some studies report on multiple geographic locations or digitally adapted activities/services. Percentages are calculated based on the total number of 13 articles.

[b]SMART Recovery International.

[c]PO: patient organization.

Patient Organizations' Digital Responses to the COVID-19 Pandemic

Among the articles reviewed, digital adaptations of various PO activities in response to COVID–19–related public health measures were described. Most (n=9) focused on the digitalization of a specific PO activity [29-31,33,34,36,38-40], while 4 explored the digital transformation of multiple activities [28,32,35,37]. These digitally adapted services and activities can be broadly categorized into 3 main areas, that are group-based support activities, communication and counseling, as well as information provision and educational activities (refer to Figure 3).

XSL•FO
RenderX

Figure 3. Patient organizations' activities and services were digitally adapted during the COVID-19 pandemic (gray bubbles) and the technologies used (blue bubbles). PO: patient organization.

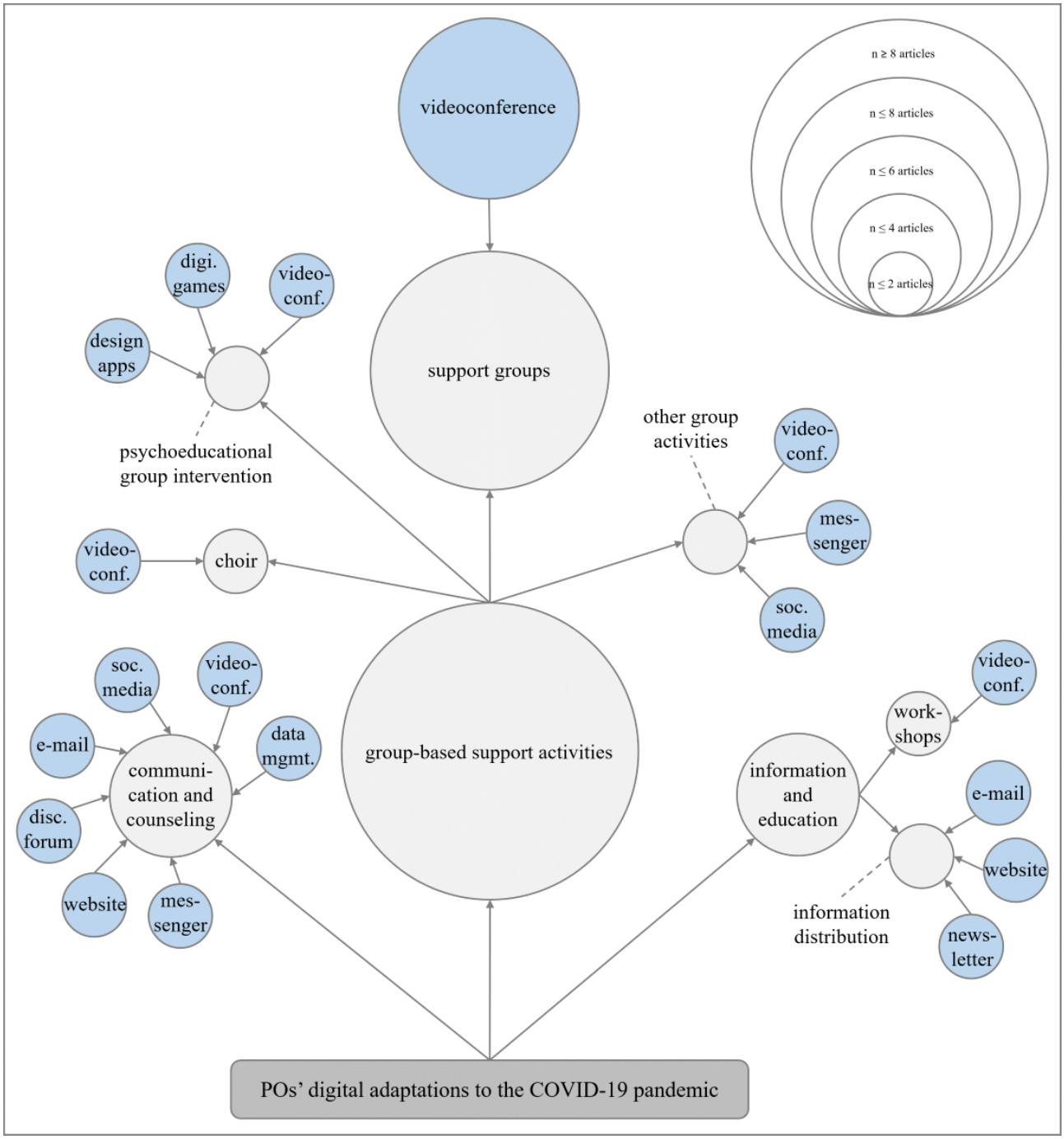

Group-Based Support Activities

Overview

The digitalization of group-based support activities is the most frequently reported aspect (n=9), with peer support groups discussed in 7 articles [30,33,35,36,38-40]. This includes both the transformation from face-to-face to online support groups [33,35,39,40], as well as the significant expansion of existing online support groups that were previously less common and mainly held in conjunction with face-to-face meetings [34]. Kelly et al [30] discussed the response of SMART Recovery International, a global community of peer support groups for addictive behaviors. In some countries, like the United States, this involved notably increasing the number of online support groups to compensate for the discontinuation of in-person groups, while in others, such as Denmark, digital options were introduced for the first time. Lamont et al [36] described various digital communication methods used by facilitators and members of stroke groups of the Stroke Association (United Kingdom) to maintain engagement. Originally, peer support, creative arts, and stroke cafés were among the activities of these groups. However, the extent of digital transition remains ambiguous, as the original core activities were not necessarily directly adapted, with Lamont et al [36] reporting that digital tools were primarily used to keep in touch. Other POs' group activities that were digitally transformed during the pandemic include a choir [34] and a psychoeducational group intervention [29].

XSL•FO
RenderX

In terms of digital technologies, videoconferencing, including platforms such as Zoom, was mentioned in 8 articles, making it the most commonly used tool in the process of digital adaptation [29,33-36,38-40]. Kelly et al [30] did not specify the technology used to deliver support groups digitally. Refer to Table S1 in Multimedia Appendix 5, for further details.

Positive Outcomes

The most commonly reported benefits of digitally adapting group-based support activities during the pandemic were preserving interpersonal dynamics and increasing accessibility (refer to Table S2 in Multimedia Appendix 5, for details on other positive outcomes).

Interpersonal Dynamics

Of the 9 articles addressing group-based support activities, 7 reported that the interpersonal nature of such activities was successfully preserved in the digital setting [29,33-36,38,39]. For instance, Seckam and Hallingberg [34] observed that participants felt inspired by the meetings with their fellow members and experienced a sense of belonging in the digital environment, which helped reduce feelings of social isolation. Similarly, Marks et al [35] reported that participants felt less isolated, experienced social or emotional connections with other group members, and shared information and stories within the digital environment. Penfold and Ogden [39] also observed social comparison as well as interpersonal dynamics such as social affirmation, solidarity, and feelings of togetherness in digitally delivered group meetings. Both conclude that key elements of group meetings were preserved during the digital transition [35,39]. Lamont et al [36] observed a high level of social identification of members with their group during the pandemic, resulting in more positive psychosocial outcomes. The authors attributed this to groups maintaining collective and individual interactions, primarily through communication tools such as email, video calls, and text messages.

Accessibility

Several articles (n=5) referred to the impact of the digital transition on the accessibility of group-based support activities. For instance, 3 articles reported improved access for people who were previously unable to participate in person due to barriers related to physical or mental health conditions, work, or financial resources [29,35,40]. By moving to digital formats, some POs also expanded the accessibility of their group activities nationally or even internationally [30,35,39,40]. This newly gained accessibility was viewed positively by some participants. Penfold and Ogden [39] highlighted that participants appreciated being able to choose from a wider range of meetings, and Senreich et al [40] observed that the increased diversity of participants, due to easier access, was positively perceived.

Challenges and Barriers

At the same time, this shift to digital formats introduced several challenges to the delivery of group-based support activities. Access barriers, particularly for those individuals lacking digital resources or skills, were the most commonly reported [34-36,40]. Seckam and Hallingberg [34] explicitly discussed these as an ethical challenge, highlighting the unintentional exclusion of some members from virtual choral activities due to these factors. Furthermore, the accessibility of group activities was sometimes negatively impacted by technological difficulties, such as poor internet connections [33,38].

Challenges related to privacy and anonymity were reported in three articles, including difficulties in managing privacy in the digital setting due to the facilitator's limited control [29], group member concerns about maintaining anonymity in group meetings [39], and incidents of "zoom bombing" with disruptive behavior [40].

In contrast to the finding that maintaining the dynamics of group interactions was identified as a benefit of digital formats, 3 articles reported that some group members were less satisfied with such formats, particularly in terms of interpersonal aspects, such as feeling rather disconnected from the group and lacking a sense of community [35,39,40]. Furthermore, the absence of informal conversations and interactions typically surrounding in-person meetings was viewed negatively [40], posing challenges in developing relationships within the groups for some [35]. For more details, refer to Table S3 in Multimedia Appendix 5.

Facilitating Factors

From the perspective of an internationally-operating PO, Kelly et al [30] reported on the challenges SMART Recovery International faced in digitally transforming support groups in countries where such digital options were not previously established. Here, the digital environment had to be developed from the ground up and facilitators had to be trained. However, the authors also mentioned that collaborating transnationally and sharing experiences with SMART Recovery affiliates in countries with preexisting online support groups proved to be beneficial. Participants' previous experience with digital technologies [33,34], the presence of experienced group facilitators [33,35], and government funding [30,38] also had a positive impact on the digital transformation of group activities. Refer to Table S4 in Multimedia Appendix 5, for more details.

Communication and Counseling

Chung et al [37] and McMullan et al [28], based on their surveys with representatives of rare disease POs, reported that the majority of participating organizations used digital communication tools to maintain contact with their members and provide ongoing support during the pandemic. This included enabling interactions through video calls, social media, and discussion forums.

In addition, Chung et al [37] described the digitalization of counseling services that were previously offered in person. Bouey et al [32] and Nemeth Blažić et al [31] also discussed this transition, focusing on POs that provide support and care for individuals with sexually transmitted infections, particularly HIV infection and AIDS, where counseling is an integral part of their prevention and testing activities. Various digital technologies were used to maintain support through counseling during the pandemic, such as videoconferencing software, email, and social media.

XSL•FO
RenderX

Chung et al [37] reported that relying on digital communication tools during the pandemic posed challenges to adequately supporting all members, as not all had access to these tools or were skilled enough to use them.

We could not identify any information on facilitating factors specifically associated with the digital transformation of POs' communication and counseling services (Table S1-S4 in Multimedia Appendix 6).

Information Provision and Educational Activities

The digital adaptation of POs' information provision and educational activities was addressed in 3 of the included articles [28,35,37]. Marks et al [35] described the digital delivery of educational workshops through Zoom, while McMullan et al [28] briefly mentioned the implementation of weekly webinars and the distribution of information updates and newsletters. Similarly, Chung et al [37] reported the use of newsletters, along with the use of POs' websites, as methods of disseminating information to members during the pandemic.

Marks et al [35] observed that participants in the educational workshops delivered via Zoom experienced social support and a sense of collectivity, echoing findings from digital group-based support activities. They also deduced from their qualitative interview data that professional facilitation and moderation created a positive learning environment where participants felt motivated and comfortable to engage. Attributes specific to the digital format, such as the ease of information sharing through chat, enhanced the learning experience. However, the authors also described challenges associated with the use of digital technology for some participants. Further details are provided in Table S1-S4 in Multimedia Appendix 6.

Operational Activities

Chung et al [37] and McMullan et al [28] also addressed how POs have digitally transformed their operation. McMullan et al [28] specifically mentioned the transition of committee meetings to videoconferencing formats. Chung et al [37] also noted the digitalization of meetings, though they did not specify the types, and further addressed the broader digitalization of organizational operations as an adaptive response to the pandemic. Their survey of directors and representatives of rare disease POs in the Asia-Pacific region revealed that the "digitalization of operation" was most frequently stated as the predominant factor contributing to the perceived success of POs during the pandemic. However, the survey also showed that the actual success of these digital adjustment efforts varied by geographic region. For example, all 15 Australian organizations reported successful digital adaptation of their operations, while none of the 18 participating POs from Hong Kong were able to digitally adapt. The authors hypothesize that this disparity may be due to differences in the availability of the necessary digital infrastructure. Refer to Tables S1-S4 in Multimedia Appendix 6, for more details.

Discussion

Scope and Context

Based on the results of our systematic literature search, this review provides a comprehensive overview of how various POs adapted their services and operations to digital formats in response to pandemic-related circumstances. It covers a wide range of organizational types, health-related indications, and geographic areas, highlighting the broad impact of the pandemic.

Principal Findings on Patient Organizations' Digital Responses

Our analysis reveals that the COVID-19 pandemic prompted various digital responses from POs. It led to the adoption of new digital solutions or, in some cases, the significant expansion of existing digital services, to maintain essential activities under pandemic conditions and to meet support needs. All articles reported largely successful digital efforts by POs and their associated support groups, ensuring the continuation of their services in the midst of the pandemic. This is consistent with broader trends observed in the health care sector, where rapid digital transitions during the pandemic played a critical role in maintaining service delivery [13,17,18,20,21].

As all included articles focus on the immediate impact of the pandemic and the rapid adoption of digital solutions, it is not possible to reach definitive conclusions about the long-term impact of the pandemic on digitalization within these organizations, or its sustainability. However, our findings from 4 articles suggest a preference among some participants for the continuation of digital group activities [33,35,38,40]. This observation suggests that digital options might beneficially be preserved alongside the resumption of in-person meetings, particularly where digital experiences have been positive. Therefore, future research could explore the long-term implications of pandemic-induced digital transformations within these POs.

Another notable finding is that the digital transformation of group-based support activities is the most frequently and extensively reported response in our sample (n=9). This trend may be related to the fact that providing peer support, such as through support groups, is a core task of many POs [2,4,5]. The pressing need to digitize these services during the pandemic, which was likely a priority for organizations to continue providing support, may explain the extensive reporting of this transition.

Digital Technologies

The predominant use of videoconferencing software during the pandemic-driven digitalization is consistent with trends observed in various sectors, such as higher education [41] and health care [17]. This widespread adoption may be related to the nature of these tools, which enable audiovisual communication and are therefore closest to face-to-face interaction. Consequently, videoconferencing offers a more comprehensive mode of interaction, potentially making it a more appealing alternative for social interaction than purely speech- or text-based forms. Before conducting our review, we had considered the possibility that POs might develop specific digital solutions, most likely

XSL•FO
RenderX

in collaboration with software developers, to better customize their services and meet the needs of their members. However, our findings did not indicate such developments. This may be due to the rapid and urgent nature of the pandemic-related digital adaptations, which likely made the development of specialized technologies impractical at the time. In addition, the relatively high cost may have been a limiting factor or existing tools may have been sufficient for the needs of POs.

Positive Outcomes, Challenges, and Facilitating Factors

The majority of articles addressing group activities reported that participants' interpersonal experiences in online support groups were similar to those in face-to-face groups, including a sense of togetherness. This observation aligns with similar trends reported in studies on various digital support group formats that emerged independently of the COVID-19 pandemic [42-44]. Consequently, such groups may indeed serve as adequate alternatives to face-to-face group activities of POs. Another positive outcome highlighted in several articles, primarily regarding the digitalization of group activities, is increased accessibility. This shift has effectively removed several former barriers, allowing participants to access meetings from anywhere. This has been particularly beneficial for individuals with financial or mobility constraints. This reported benefit of improved access through digital technologies is consistent with previous findings, both during [13,20] and outside the pandemic period [42]. However, our review also shows that the shift to digital formats has simultaneously created new barriers, particularly for those with limited digital literacy or resources. These challenges also appear to transcend the pandemic context, as evidenced by other findings on digital health in both pandemic [13,20] and non-pandemic settings [45,46]. This dichotomy points to the dual nature of digital formats, as they can both bridge and widen access gaps and underscores the importance of nuanced consideration.

Regardless of the specific service or activity, we found that comparatively few factors were consistently reported as facilitators of digital transformation. The beneficial impact of skilled facilitators on the smooth digital adaptation of support groups and educational workshops, as reported by Marks et al [35] and Constantini et al [33], is also highlighted in a review not specifically related to the pandemic [46], making this a potentially relevant general recommendation for the future delivery of digital activities. From a broader perspective not focused on group activities, Chung et al [37] identified preexisting digital resources and experience in digitally delivering such services and activities as important facilitators of digital adaptation.

In conclusion, when establishing any kind of digital service within POs, it is likely to be advantageous to promote digital literacy among members and participants and to create solutions for those without access to digital technologies, as well as to be digitally equipped and experienced as an organization.

Ethical Issues

In contexts such as digital transformation in health care, ethical implications such as privacy and justice, as well as related issues, are frequently discussed [47]. Notably, as mentioned above, only Seckam and Hallingberg [34] reported ethical considerations related to the limited accessibility of digital formats and the unintentional exclusion of some group members due to digital literacy gaps or a lack of necessary technology. While other articles [34-36,40] also reported accessibility challenges, they did not explicitly categorize them as ethical issues. Bergmans et al [29], Penfold and Ogden [39], and Senreich et al [40] observed privacy concerns, although not explicitly framed as ethical issues either. These challenges regarding fair access and privacy align with those identified in the broader digital health landscape, where ethical guidance is suggested as an initial step to address such concerns [47]. However, only Bergmans et al [29] reported mitigating actions, such as "coaching" sessions to help participants navigate the privacy settings of digital platforms. Other articles did not specify the approaches taken by POs to address these potential ethical challenges.

Limitations of Findings

First, our analysis is based on a limited sample of publications that may not fully represent the situation of all POs. Our search primarily yielded articles that focused on organizations in Europe, particularly the United Kingdom. Therefore, our findings may not fully capture the global situation. It is possible that digital responses related to the pandemic may have varied significantly in countries not included in our sample. In addition, the majority of the articles focused on well-established organizations with the resources to actively support research. Of these, 7 articles received funding from or were developed in an active collaboration with POs, suggesting that these factors may have influenced the scope of the available research on this topic. As a result, the experiences and perspectives of smaller, less financially equipped organizations may be underrepresented in our findings, despite representing a large proportion of POs in countries, such as Germany or the United States [3,4]. In conclusion, our findings indicate that the research literature predominantly focuses on more professionalized POs and entities in specific geographical regions. Hence, a need for more diverse research is apparent. Further, we found that despite the diverse activities of POs, the articles in this review did not address instances in which these organizations acted as advocates, which may include efforts to influence health policy making, collaborate with other health care stakeholders, or engage in research. Consequently, it remains unclear whether these specific activities were also adapted to digital formats, suggesting potential areas for further research.

Having acknowledged these limitations, it is important to clarify our objective. The primary purpose of this review was not to generate generalizable findings or to provide a comprehensive picture of the global situation. Rather, our aim was to determine what literature has been published on the topic, describe individual findings, identify emerging trends, and highlight research gaps. In fact, the absence of articles addressing certain gaps identified in our review is more indicative of a lack of diversity in the published literature. This may be due to factors such as POs lacking the necessary financial or human resources for research involvement, or possibly because engaging in research or publishing in academic journals is not a priority for some POs.

XSL•FO
RenderX

Finally, the quality of reporting within our sample varies, particularly in terms of depth. For example, most articles did not focus on the digital transformation process as such. This variation means that the amount of information we gathered and extracted varied. As a result, not all aspects of digital transformation were reported in each article, nor were they necessarily covered to the same extent.

Methodological Limitations

A challenge in developing our search strategy was the lack of consistent terminology and definitions for POs. Hence, we refined our search strategy, incorporating various synonyms to retrieve relevant literature. After the initial search and screening, we identified additional terms as potentially relevant, such as "voluntary health agency." The fact that adding these terms to the search strategy did not yield any additional relevant publications gives us further confidence that our search was sufficiently sensitive.

Also, we used a definition of POs that may reflect a rather Western perspective. Potentially, even broader inclusion criteria might have yielded more publications. However, this approach may have resulted in a sample with significant variation across organizations, making comparisons impossible. In addition, by limiting our review to nonprofit POs, the scope of our findings may be somewhat narrowed, as POs with different operational modes are not represented. Nevertheless, this approach reflects the predominant structure within the sector.

Finally, we included only articles published in English or German to ensure feasibility. The overall search results in PubMed, for example, showed only a modest increase in search results (0.87%) when no language filter was applied, and hence this language restriction appears to have been negligible.

Conclusion

Our findings on these rapid transitions can be taken as an indicator of the resilience and adaptability of these POs, and underscore the significant potential of digital technologies to enhance support services in such unprecedented times.

Based on the various aspects we identified in this review, potentially relevant recommendations for future PO digitalization strategies relate to promoting digital literacy among members and participants, creating solutions for those without access to digital technologies or those who may not wish to use them, and training PO staff to provide skilled and supportive delivery of digital services.

Our findings also highlight several research gaps. For example, there seems to be a predominant focus in the literature on well-established organizations in Western countries. This may overlook the unique experiences of smaller, less well-funded POs or those in different geographical areas. In addition, the current literature does not provide insights into areas such as advocacy or research engagement and their digital adaptations, suggesting potential areas for further research. Furthermore, while our review highlights the immediate digital responses of POs to the pandemic, the long-term sustainability and impact of these adjustments remain unclear. Future research should explore these aspects to fully understand the long-term impact of the COVID-19 pandemic on POs.

Acknowledgments

This scoping review was conducted as part of the PANDORA (Patient-centered Digitalization: An Ethical Analysis of the Role of Patient Organizations as Actors in the Context of Digitalization in Health-related Research and Care) research consortium, funded by the German Federal Ministry of Education and Research (grant 01GP2115B).

Authors' Contributions

SW developed the concept and methodological approach, wrote the protocol, performed literature screening, performed data extraction and analysis as well as drafted and edited the manuscript. MLD contributed to the conceptualization and development of the methodological approach, reviewed the manuscript, and served as project coordinator. CJ contributed to the development of the search strategy and literature screening. JL contributed to the conceptualization and development of the methodological approach, performed literature screening and data extraction as well as reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 165 KB - jmir_v26i1e58566_app1.pdf]

Multimedia Appendix 2
Deviations from protocol.
[PDF File (Adobe PDF File), 102 KB - jmir_v26i1e58566_app2.pdf]

Multimedia Appendix 3
PCC (Population, Concept, and Context) mnemonic and search strategies.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 231 KB - jmir_v26i1e58566_app3.pdf]

Multimedia Appendix 4
Characteristics of included articles and patient organizations.
[PDF File (Adobe PDF File), 322 KB - jmir_v26i1e58566_app4.pdf]

Multimedia Appendix 5
Consolidated findings on patient organizations' digital adaptations of group-based support activities.
[PDF File (Adobe PDF File), 299 KB - jmir_v26i1e58566_app5.pdf]

Multimedia Appendix 6
Consolidated findings on patient organizations' digital adaptations of communication, counseling, information provision, educational activities, and other operational aspects.
[PDF File (Adobe PDF File), 242 KB - jmir_v26i1e58566_app6.pdf]

References

1. Gentilini A, Miraldo M. The role of patient organisations in research and development: evidence from rare diseases. Soc Sci Med 2023 Dec;338:116332 [FREE Full text] [doi: 10.1016/j.socscimed.2023.116332] [Medline: 37866173]
2. Kamphuis HC, Hekkert KD, van Dongen M, Kool RB. Facts and figures about patient associations in the Netherlands between 2007 and 2009: review of their activities and aims. Health Policy 2012 Oct;107(2-3):243-248. [doi: 10.1016/j.healthpol.2012.06.008] [Medline: 22835496]
3. Kofahl C, Seidel G, Weber J, Werner S, Nickel S. Strukturen und Prozesse bei Selbsthilfegruppen und -organisationen. In: Kofahl C, Schulz-Nieswandt F, Dierks ML, editors. Selbsthilfe und Selbsthilfeunterstützung in Deutschland. Berlin, Münster: LIT-Verlag; 2016:83-112.
4. Rose SL, Highland J, Karafa MT, Joffe S. Patient advocacy organizations, industry funding, and conflicts of interest. JAMA Intern Med 2017 Mar 01;177(3):344-350 [FREE Full text] [doi: 10.1001/jamainternmed.2016.8443] [Medline: 28114624]
5. Sienkiewicz D, van Lingen C. The Added Value of Patient Organisations. Brussels, Belgium: European Patients Forum; 2017.
6. Nkengasong J, Iwasaki A, Victora C, Oh J, Gao GF, Agrawal A, et al. The global response to the COVID-19 pandemic. Med 2020 Dec 18;1(1):3-8 [FREE Full text] [doi: 10.1016/j.medj.2020.12.003] [Medline: 33363282]
7. Kronblad C, Envall Pregmark J. Responding to the COVID-19 crisis: the rapid turn toward digital business models. JSTPM 2021 Sep 22;15(3):451-467 [FREE Full text] [doi: 10.1108/jstpm-10-2020-0155]
8. Dwivedi YK, Hughes DL, Coombs C, Constantiou I, Duan Y, Edwards JS, et al. Impact of COVID-19 pandemic on information management research and practice: transforming education, work and life. International Journal of Information Management 2020 Dec;55:102211. [doi: 10.1016/j.ijinfomgt.2020.102211]
9. Kodama M. Digitally transforming work styles in an era of infectious disease. Int J Inf Manage 2020 Dec;55:102172 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102172] [Medline: 32836634]
10. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
11. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
12. Garfan S, Alamoodi AH, Zaidan BB, Al-Zobbi M, Hamid RA, Alwan JK, et al. Telehealth utilization during the Covid-19 pandemic: a systematic review. Comput Biol Med 2021 Nov;138:104878 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104878] [Medline: 34592585]
13. Silva CRDV, Lopes RH, de Goes Bay O, Martiniano CS, Fuentealba-Torres M, Arcêncio RA, et al. Digital health opportunities to improve primary health care in the context of COVID-19: scoping review. JMIR Hum Factors 2022 May 31;9(2):e35380 [FREE Full text] [doi: 10.2196/35380] [Medline: 35319466]
14. Nickel S, Bremer K, Dierks M, Haack M, Wittmar S, Borgetto B, et al. Digitization in health-related self-help - results of an online survey among self-help organizations in Germany. Digit Health 2022;8:1-13 [FREE Full text] [doi: 10.1177/20552076221120726] [Medline: 36046635]
15. Rauter CM, Wöhlke S, Schicktanz S. My data, my choice? - German patient organizations' attitudes towards big data-driven approaches in personalized medicine. an empirical-ethical study. J Med Syst 2021 Feb 22;45(4):43 [FREE Full text] [doi: 10.1007/s10916-020-01702-7] [Medline: 33616768]
16. Joseph KL, Dagfinrud H, Hagen KB, Nordén KR, Fongen C, Wold O, et al. The AktiWeb study: feasibility of a web-based exercise program delivered by a patient organisation to patients with hip and/or knee osteoarthritis. Pilot Feasibility Stud 2022 Jul 20;8(1):150 [FREE Full text] [doi: 10.1186/s40814-022-01110-3] [Medline: 35859065]

XSL•FO
RenderX

17. Abd-Alrazaq A, Hassan A, Abuelezz I, Ahmed A, Alzubaidi MS, Shah U, et al. Overview of technologies implemented during the first wave of the COVID-19 pandemic: scoping review. J Med Internet Res 2021 Sep 14;23(9):e29136 [FREE Full text] [doi: 10.2196/29136] [Medline: 34406962]
18. Garavand A, Jalali S, Hajipour Talebi A, Sabahi A. Advantages and disadvantages of teleworking in healthcare institutions during COVID-19: a systematic review. Inform Med Unlocked 2022;34:101119 [FREE Full text] [doi: 10.1016/j.imu.2022.101119] [Medline: 36373130]
19. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021;7:20552076211009396 [FREE Full text] [doi: 10.1177/20552076211009396] [Medline: 33959378]
20. Freire MP, Silva LG, Meira ALP, Louvison MCP. Telemedicine in healthcare access during the covid-19 pandemic: a scoping review. Rev Saude Publica 2023;57(suppl 1):4s [FREE Full text] [doi: 10.11606/s1518-8787.2023057004748] [Medline: 37255115]
21. Hincapié MA, Gallego JC, Gempeler A, Piñeros JA, Nasner D, Escobar MF. Implementation and usefulness of telemedicine during the COVID-19 pandemic: a scoping review. J Prim Care Community Health 2020;11:2150132720980612 [FREE Full text] [doi: 10.1177/2150132720980612] [Medline: 33300414]
22. Peters M, Godfrey C, McInerney P, Munn Z, Trico A, Khalil H. Chapter 11: Scoping Reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide: Joanna Briggs Institute; 2020.
23. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Wallraf S, Lander J, John C, Dierks ML. Impact of the COVID-19 pandemic on digitalization in health-related self-help and patient organizations. A scoping review protocol. 2022. URL: https://osf.io/ejwuk/ [accessed 2024-01-10]
25. von Elm E, Schreiber G, Haupt CC. Methodische anleitung für scoping reviews (JBI-Methodologie). Z Evid Fortbild Qual Gesundhwes 2019 Jun;143:1-7 [FREE Full text] [doi: 10.1016/j.zefq.2019.05.004] [Medline: 31296451]
26. PubMed user guide. National Library of Medicine. 2023. URL: https://pubmed.ncbi.nlm.nih.gov/help/ [accessed 2024-01-24]
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
28. McMullan J, Crowe AL, Bailie C, McKnight AJ. Evaluating the impact of COVID-19 on rare disease support groups. BMC Res Notes 2021 May 06;14(1):168 [FREE Full text] [doi: 10.1186/s13104-021-05579-8] [Medline: 33957984]
29. Bergmans Y, Kellington K, Smith T, Pond A, Goving M, Shelton E, et al. Providing virtual suicide prevention groups for people experiencing suicidality: pivoting service delivery during the COVID-19 pandemic. International Social Work 2021 Aug 04;64(5):801-805 [FREE Full text] [doi: 10.1177/0020872821996781]
30. Kelly PJ, McCreanor K, Beck AK, Ingram I, O'Brien D, King A, et al. SMART recovery international and COVID-19: expanding the reach of mutual support through online groups. J Subst Abuse Treat 2021 Dec;131:108568 [FREE Full text] [doi: 10.1016/j.jsat.2021.108568] [Medline: 34446323]
31. Nemeth Blažić T, Bogdanić N, Nola IA, Kosanović Ličina ML, Delaš Aždajić M. Digital technology and HIV, HCV and STI voluntary counselling and testing: good practice example from Croatia. Cent Eur J Public Health 2022 Jun;30(2):107-110 [FREE Full text] [doi: 10.21101/cejph.a7237] [Medline: 35876599]
32. Bouey JZH, Han J, Liu Y, Vuckovic M, Zhu K, Zhou K, et al. A case study of HIV/AIDS services from community-based organizations during COVID-19 lockdown in China. BMC Health Serv Res 2023 Mar 27;23(1):288 [FREE Full text] [doi: 10.1186/s12913-023-09271-4] [Medline: 36973805]
33. Constantini J, Bliem HR, Crepaz M, Marksteiner J. A qualitative literature search and pilot study of online support groups for patients with dementia and their carers. Gerontol Geriatr Med 2023;9:23337214231205689 [FREE Full text] [doi: 10.1177/23337214231205689] [Medline: 37881348]
34. Seckam A, Hallingberg B. The experiences and perceptions of stroke survivors engaging in a virtual choir during COVID-19: a thematic analysis. British Journal of Neuroscience Nursing 2021 Oct 01;17(Sup5):S18-S25. [doi: 10.12968/bjnn.2021.17.sup5.s18]
35. Marks E, Handscomb L, Remskar M. "I can see a path forward now": a qualitative investigation of online groups for tinnitus in the time of Covid-19. Int J Audiol 2023 Jul;62(7):659-666 [FREE Full text] [doi: 10.1080/14992027.2022.2063195] [Medline: 35499467]
36. Lamont RA, Calitri R, Mounce LTA, Hollands L, Dean SG, Code C, et al. Shared social identity and perceived social support among stroke groups during the COVID-19 pandemic: relationship with psychosocial health. Appl Psychol Health Well Being 2023 Feb;15(1):172-192 [FREE Full text] [doi: 10.1111/aphw.12348] [Medline: 35139581]
37. Chung CCY, Ng YNC, Jain R, Chung BHY. A thematic study: impact of COVID-19 pandemic on rare disease organisations and patients across ten jurisdictions in the Asia Pacific region. Orphanet J Rare Dis 2021 Mar 05;16(1):119 [FREE Full text] [doi: 10.1186/s13023-021-01766-9] [Medline: 33673852]
38. Beck AK, Larance B, Baker AL, Deane FP, Manning V, Hides L, et al. Supporting people affected by problematic alcohol, substance use and other behaviours under pandemic conditions: a pragmatic evaluation of how SMART recovery Australia

XSL•FO
RenderX

responded to COVID-19. Addict Behav 2023 Apr;139:107577 [FREE Full text] [doi: 10.1016/j.addbeh.2022.107577] [Medline: 36528964]
39. Penfold KL, Ogden J. Exploring the experience of gamblers anonymous meetings during COVID-19: a qualitative study. Curr Psychol 2022;41(11):8200-8213 [FREE Full text] [doi: 10.1007/s12144-021-02089-5] [Medline: 34421284]
40. Senreich E, Saint-Louis N, Steen JT, Cooper CE. The experiences of 12-Step program attendees transitioning to online meetings during the COVID-19 pandemic. Alcoholism Treatment Quarterly 2022 Jul 17;40(4):500-517. [doi: 10.1080/07347324.2022.2102456]
41. Ahmady S, Kallestrup P, Sadoughi MM, Katibeh M, Kalantarion M, Amini M, et al. Distance learning strategies in medical education during COVID-19: a systematic review. J Educ Health Promot 2021;10:421 [FREE Full text] [doi: 10.4103/jehp.jehp_318_21] [Medline: 35071627]
42. Banbury A, Nancarrow S, Dart J, Gray L, Parkinson L. Telehealth interventions delivering home-based support group videoconferencing: systematic review. J Med Internet Res 2018 Feb 02;20(2):e25 [FREE Full text] [doi: 10.2196/jmir.8090] [Medline: 29396387]
43. Marziali E, Garcia LJ. Dementia caregivers' responses to 2 Internet-based intervention programs. Am J Alzheimers Dis Other Demen 2011 Feb;26(1):36-43 [FREE Full text] [doi: 10.1177/1533317510387586] [Medline: 21282276]
44. Rapley JM, Camic PM, Brotherhood E, Crutch SJ, Harding E, Rare Dementia Support (RDS) Impact Study team. Video conferencing peer support and rarer forms of dementia: an exploration of family carers' positive experiences. Qual Health Res 2023 Aug;33(10):884-896 [FREE Full text] [doi: 10.1177/10497323231172880] [Medline: 37395095]
45. Banbury A, Parkinson L, Gordon S, Wood D. Implementing a peer-support programme by group videoconferencing for isolated carers of people with dementia. J Telemed Telecare 2019 Oct;25(9):572-577. [doi: 10.1177/1357633X19873793] [Medline: 31631761]
46. Armstrong MJ, Alliance S. Virtual support groups for informal caregivers of individuals with dementia: a scoping review. Alzheimer Dis Assoc Disord 2019;33(4):362-369. [doi: 10.1097/WAD.0000000000000349] [Medline: 31567144]
47. Brall C, Schröder-Bäck P, Maeckelberghe E. Ethical aspects of digital health from a justice point of view. Eur J Public Health 2019 Oct 01;29(Supplement_3):18-22 [FREE Full text] [doi: 10.1093/eurpub/ckz167] [Medline: 31738439]

Abbreviations

JBI: Joanna Briggs Institute
PCC: Population, Concept, and Context
PO: patient organization
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SMART: Self-Management and Recovery Training
WHO: World Health Organization

Edited by N Cahill; submitted 19.03.24; peer-reviewed by K Aboalshamat, F Hanna; comments to author 01.04.24; revised version received 06.05.24; accepted 07.10.24; published 20.12.24.

Please cite as:
Wallraf S, Dierks ML, John C, Lander J
Patient Organizations' Digital Responses to the COVID-19 Pandemic: Scoping Review
J Med Internet Res 2024;26:e58566
URL: https://www.jmir.org/2024/1/e58566
doi:10.2196/58566
PMID:39705075

©Simon Wallraf, Marie-Luise Dierks, Cosima John, Jonas Lander. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Studies of Artificial Intelligence/Machine Learning Registered on ClinicalTrials.gov: Cross-Sectional Study With Temporal Trends, 2010-2023

Shoko Maru[1,2], PhD; Michael D Matthias[3], BEng; Ryohei Kuwatsuru[1,4], MD, PhD; Ross J Simpson Jr[5], MD, PhD

[1]Graduate School of Medicine, Juntendo University, Tokyo, Japan

[2]Clinical Study Support Inc, Nagoya, Japan

[3]Matthias IT Pty Ltd, Brisbane, Australia

[4]Department of Radiology, School of Medicine, Juntendo University, Tokyo, Japan

[5]Division of Cardiology, School of Medicine, University of North Carolina Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:
Shoko Maru, PhD
Graduate School of Medicine
Juntendo University
2-1-1 Hongo, Bunkyo-ku
Tokyo, 113-8421
Japan
Phone: 81 338133111
Email: shoko.maru@alumni.griffithuni.edu.au

Abstract

Background: The rapid growth of research in artificial intelligence (AI) and machine learning (ML) continues. However, it is unclear whether this growth reflects an increase in desirable study attributes or merely perpetuates the same issues previously raised in the literature.

Objective: This study aims to evaluate temporal trends in AI/ML studies over time and identify variations that are not apparent from aggregated totals at a single point in time.

Methods: We identified AI/ML studies registered on ClinicalTrials.gov with start dates between January 1, 2010, and December 31, 2023. Studies were included if AI/ML-specific terms appeared in the official title, detailed description, brief summary, intervention, primary outcome, or sponsors' keywords. Studies registered as systematic reviews and meta-analyses were excluded. We reported trends in AI/ML studies over time, along with study characteristics that were fast-growing and those that remained unchanged during 2010-2023.

Results: Of 3106 AI/ML studies, only 7.6% (n=235) were regulated by the US Food and Drug Administration. The most common study characteristics were randomized (56.2%; 670/1193; interventional) and prospective (58.9%; 1126/1913; observational) designs; a focus on diagnosis (28.2%; 335/1190) and treatment (24.4%; 290/1190); hospital/clinic (44.2%; 1373/3106) or academic (28%; 869/3106) sponsorship; and neoplasm (12.9%; 420/3245), nervous system (12.2%; 395/3245), cardiovascular (11.1%; 356/3245) or pathological conditions (10%; 325/3245; multiple counts per study possible). Enrollment data were skewed to the right: maximum 13,977,257; mean 16,962 (SD 288,155); median 255 (IQR 80-1000). The most common size category was 101-1000 (44.8%; 1372/3061; excluding withdrawn or missing), but large studies (n>1000) represented 24.1% (738/3061) of all studies: 29% (551/1898) of observational studies and 16.1% (187/1163) of trials. Study locations were predominantly in high-income countries (75.3%; 2340/3106), followed by upper-middle-income (21.7%; 675/3106), lower-middle-income (2.8%; 88/3106), and low-income countries (0.1%; 3/3106). The fastest-growing characteristics over time were high-income countries (location); Europe, Asia, and North America (location); diagnosis and treatment (primary purpose); hospital/clinic and academia (lead sponsor); randomized and prospective designs; and the 1-100 and 101-1000 size categories. Only 5.6% (47/842) of completed studies had results available on ClinicalTrials.gov, and this pattern persisted. Over time, there was an increase in not only the number of newly initiated studies, but also the number of completed studies without posted results.

Conclusions: Much of the rapid growth in AI/ML studies comes from high-income countries in high-resource settings, albeit with a modest increase in upper-middle-income countries (mostly China). Lower-middle-income or low-income countries remain poorly represented. The increase in randomized or prospective designs, along with 738 large studies (n>1000), mostly ongoing,

XSL•FO
RenderX

may indicate that enough studies are shifting from an in silico evaluation stage toward a prospective comparative evaluation stage. However, the ongoing limited availability of basic results on ClinicalTrials.gov contrasts with this field's rapid advancements and the public registry's role in reducing publication and outcome reporting biases.

(J Med Internet Res 2024;26:e57750) doi:10.2196/57750

KEYWORDS

artificial intelligence; machine learning; deep learning; trends; health care; cross-sectional study; health disparities; data-source disparities; publication bias; registry; ClinicalTrials.gov

Introduction

The number of studies on artificial intelligence (AI) and machine learning (ML) registered on ClinicalTrials.gov has significantly increased since 2016 [1,2]. While the rapid growth of research on AI/ML is apparent [1-3], the downsides in this field have also come to light. These include the lack of level I/II evidence as robust proof of the clinical and economic impacts of AI [4], lack of clinically meaningful outcomes beyond measures of technical accuracy [5], lack of uniformity and standardization in AI reporting [6,7], incomplete and poor reporting [8-10], limited publication of AI/ML results [11,12], or data-source disparities with minority subgroups potentially disadvantaged by factors such as race, gender, and socioeconomic background [13,14].

Previous cross-sectional studies using ClinicalTrials.gov as of December 2020 [2] and March 2022 [1] showed that the number of AI/ML studies continued to grow rapidly. However, it is unclear whether this growth in number reflects an increase in desirable study attributes or merely perpetuates the same issues previously raised in the literature. For example, are study designs conducive to level I/II evidence increasing over time? Do AI/ML studies continue to run predominantly in advanced economies? While research using ClinicalTrials.gov often reports country-specific distributions [1,2], examining the global distribution by economic measures could further elucidate whether technological power is concentrated in a few high-income countries.

We aimed to evaluate temporal trends over the past 14 years, identifying variations that may not be immediately apparent from aggregated totals at a single point in time. We highlighted the fastest-growing characteristics over time and those that have remained unchanged. We also characterized AI/ML studies by the following: *who* (lead sponsor sector), *what* (clinical specialty, AI/ML terms used in study descriptions), *when* (start year, primary completion year, time to results posting), *where* (study location, geographic and economic), *why* (primary study purpose), and *how* (study design, enrollment).

Methods

Data Source

We conducted a cross-sectional analysis using ClinicalTrials.gov, a trial registry and results database maintained by the US National Library of Medicine. We sourced data from the Clinical Trials Transformation Initiative Aggregate Analysis of ClinicalTrials.gov (CTTI AACT) [15], which allows open access to the complete set of trials registered in ClinicalTrials.gov, including additional fields that are not readily available in direct exports from ClinicalTrials.gov. The CTTI AACT data dictionary is publicly accessible [16]. A static version of the CTTI AACT database was downloaded for analysis on February 6, 2024, via PostgreSQL (pSQL). The pSQL codes used are provided in Multimedia Appendix 1.

Study Selection

We used text-based search to identify relevant AI/ML studies. Our search strategies were informed by systematic reviews published during 2019-2023 [9,17-20]. We used search terms related to AI/ML methodologies and specific model architectures as used in the literature. Terms are not necessarily mutually exclusive or hierarchical (eg, "multilayer perceptron" under "machine learning"). The search strategy is detailed in Table S1 in Multimedia Appendix 2, with corresponding SQL codes available in Multimedia Appendix 1.

Inclusion Criteria (If All of the Following Met)

1. Start date: January 1, 2010, to December 31, 2023.
2. The following search terms were used:
   artificial intelligence, ai-based, augmented intelligence, deep learning, convolutional neural network, deep neural network, artificial neural network, recurrent neural network, generative adversarial network, deep reinforcement learning, machine learning, bayesian network, classification tree, elastic net, gradient boosting, xgboost, k nearest neighbour, multilayer perceptron, support vector machine, natural language processing, naive bayes, random forest, regression tree, reinforcement learning, supervised learning, unsupervised learning.
   These terms were searched for in the official title, brief summary, detailed descriptions, intervention, primary outcome, and sponsors' keywords.

Exclusion Criteria (If Any of the Following Met)

1. Studies registered as a meta-analysis, systematic review, scoping review, literature review, or protocol.
2. AI/ML-related terms were part of proper nouns (eg, organization, product, study name).
3. AI/ML-related terms were found, but they were unrelated to the study. For example, search terms were mentioned only in passing (eg, "previous research on AI showed promise," "despite advances in machine learning").

First, SQL was used to exclude studies containing any of the following terms: protocol*, design*, review*, systematic review*, scoping review*, literature review*, meta analys*s (* denotes a wildcard, equivalent to % in SQL codes). Second, the lead author selected studies with fewer than 5 search-word hits

XSL•FO
RenderX

to confirm their relevance, and irrelevant studies were excluded (eg, those involving "neural networks" of a biological nature). Third, 2 reviewers independently examined study descriptions and assessed eligibility. Discrepancies were resolved through discussion and by revisiting details on ClinicalTrials.gov. Where study descriptions were unclear or insufficient, we opted to include studies unless they explicitly met the exclusion criteria, prioritizing sensitivity over specificity.

The study flow diagram is provided in Figure S1 in Multimedia Appendix 2. The pSQL codes used are provided in Multimedia Appendix 1.

Data Extraction and Analysis

Overview

We extracted the following data fields from the CTTI AACT database: study start date, completion date, primary completion date, overall status, study type, number of sites, phase, enrollment, randomization status, masking, interventional model, primary purpose, target disease/condition, intervention, time perspective, lead sponsor, lead sponsor agency class, funding source, US Food and Drug Administration (FDA)–regulated status, sex, age, healthy volunteers, study location, and results availability.

Additionally, we created the 5 variables described in the following sections.

Enrollment—Categorical

Enrollment (continuous) was categorized as follows: 1-100; 101-1000; 1001-5000; 5001-10,000; 10,001-20,000; 20,001-30,000; and >30,000. We used the most recently updated enrollment data: planned or actual, whichever was more recent.

"Other" Sponsor Sectors—Disaggregated

As the sponsors of AI/ML research can be diverse, the "OTHER" and "OTHER_GOV" fields as classified in ClinicalTrials.gov are uninformative. We reclassified them into the following based on lead sponsor names: hospital/clinic, academia, industry, government, individual, research institute, nonprofit, and network. Academic hospitals were included in "hospital/clinic."

Study Location—Global Distribution, Geographic and Economic

We counted studies geographically and by gross national income (GNI)–based classification as per the World Bank [21]. Each country was classified by 2022 GNI per capita: low income (US $1135 or less); lower middle income (US $1136 to US $4465); upper middle income (US $4466 to US $13,845); and high income (US $13,846 or more).

Clinical Specialty or Disease Area

Clinical specialties were assigned using the Medical Subject Headings (MeSH) largest headings. Since one registered study can involve substudies, resulting in multiple specialties, we assigned the MeSH heading with the highest count to each study identifier. In cases where multiple MeSH headings had the same highest count (ties), all those terms were included (multiple MeSH headings per study identifier). If MeSH headings were missing, we used condition/disease terms submitted by sponsors to infer the clinical areas.

Note that "radiology" is a not an organ specific term and not a major MeSH heading directly linked to diseases. For example, specific disease imaging indexed under terms such as "breast neoplasms/radiography" or "lung diseases/radiography" had been preclassified as neoplasms or respiratory diseases. Screening procedures (eg, breast, colorectal) had been preclassified as neoplasms in the original data. Thus, compared to other study findings, radiology may be underrepresented, while neoplasms may be overrepresented.

Days From Primary Completion to Posting of Results

Among completed studies marked with "has results," we calculated the time between "primary completion date" and "results first posted date." We also assessed whether results were posted within 1 year after study completion (yes/no).

We used descriptive statistics and performed analyses in Microsoft Excel and Python (Python Software Foundation). We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines [22]. Ethics approval was not required because only publicly available data were analyzed.

Results

Overview

The ClinicalTrials.gov search yielded 3378 records. Of those, 272 studies were excluded: not AI/ML-related (n=265); systematic reviews or meta-analyses (n=2); and the term "artificial intelligence" merely being a part of the organization, product, or study name (n=5). After excluding 272 studies, 3106 studies were included for analysis (Figure S1 in Multimedia Appendix 2).

Tables 1 and 2 show the study characteristics. Overall, the number of AI/ML studies increased rapidly after 2017, and 62.8% (1951/3106) started in the 2021 to 2023 period. Of the 3016 studies, 842 studies (27.1%) were completed, 1694 (54.5%) were active (not yet recruiting; recruiting; enrolling by invitation; active, not recruiting), 95 (3.1%) were stopped, and 475 (15.3%) were unknown. There were more observational studies (61.6%) than interventional studies (38.4%); more parallel assignments (89.2%) than crossover, factorial, or single assignments; more open-labeled studies (63%) than ones that used any form of masking; more single-center (77.2%) than multicenter (22.8%) studies; and more studies with 2 arms (56.8%) than 1 (34.3%) or 3 (9%) arms. The study phase was mostly "not applicable" (93%). Studies involving adults aged 18-65 years (58.1%) were most common, followed by adults and seniors aged 18 and above (27.5%). While some studies accepted healthy volunteers (35.2%), the majority did not (64.8%). Only 235 (7.6%) were FDA regulated.

XSL•FO
RenderX

Table 1. Characteristics of artificial intelligence/machine learning studies (N=3106) registered on ClinicalTrials.gov, 2010-2023[a].

Study characteristics	Values
Start year (n=3106), n (%)	
2010	12 (0.4)
2011	19 (0.6)
2012	18 (0.6)
2013	22 (0.7)
2014	20 (0.6)
2015	42 (1.4)
2016	61 (2)
2017	81 (2.6)
2018	187 (6)
2019	255 (8.2)
2020	438 (14.1)
2021	599 (19.3)
2022	672 (21.6)
2023	680 (21.9)
Overall status (n=3106), n (%)	
Not yet recruiting	302 (9.7)
Recruiting	1116 (35.9)
Enrolling by invitation	90 (2.9)
Active, not recruiting	186 (6)
Completed	842 (27.1)
Withdrawn, terminated, suspended	95 (3.1)
Unknown status	475 (15.3)
Study type (n=3106), n (%)	
Interventional	1193 (38.4)
Observational	1668 (53.7)
Observational (patient registry)	245 (7.9)
Interventional model (n=742), n (%)	
Parallel assignment	662 (89.2)
Crossover assignment	58 (7.8)
Factorial assignment	22 (3)
Single assignment	0 (0)
Allocation (n=1193, interventional), n (%)	
Randomized	670 (56.2)
Nonrandomized	130 (10.9)
Not applicable/unknown	393 (32.9)
Masking (n=1193, interventional), n (%)	
Open-labeled	751 (63)
Single	232 (19.4)
Double	127 (10.6)
Triple or quadruple	83 (7)

XSL•FO
RenderX

Study characteristics	Values
Time perspective (n=1913, observational)[b], n (%)	
Prospective	1126 (58.9)
Retrospective	430 (22.5)
Cross-sectional	210 (11)
Other	147 (7.7)
Number of facilities (n=2821), n (%)	
Single-center	2177 (77.2)
Multicenter	644 (22.8)
Number of study arms (n=1103), n (%)	
1	378 (34.3)
2	626 (56.8)
3	99 (9)
Phase (n=1193), n (%)	
Early phase 1	5 (0.4)
Phase 1	7 (0.6)
Phase 1/phase 2	6 (0.5)
Phase 2	28 (2.3)
Phase 2/phase 3	3 (0.3)
Phase 3	17 (1.4)
Phase 4	18 (1.5)
Not applicable	1109 (93)
Enrollment (n=3061)[c]	
Minimum	1
Maximum	13,977,257
Mean (SD)	16,962 (288,155)
Median (IQR)	255 (80-1000)
Enrollment (n=3061)[c], n (%)	
Interventional (n=1163)[c]	
1-100	497 (42.7)
101-1000	479 (41.2)
1001-5000	118 (10.1)
5001-10,000	25 (2.1)
10,001-20,000	8 (0.7)
20,001-30,000	7 (0.6)
>30,000	29 (2.5)
Withdrawn (n=0) or unknown	30
All (n=3061)[c]	
1-100	951 (31.1)
101-1000	1372 (44.8)
1001-5000	449 (14.7)
5001-10,000	100 (3.3)
10,001-20,000	58 (1.9)

XSL•FO
RenderX

Study characteristics	Values
20,001-30,000	20 (0.7)
>30,000	111 (3.6)
Withdrawn (n=0) or unknown	45
FDA[d] status (n=3106), n (%)	
FDA regulated (device or drug)	235 (7.6)
Not regulated	2871 (92.4)
Primary purpose (n=1190)[e], n (%)	
Diagnostic	335 (28.2)
Treatment	290 (24.4)
Other	149 (12.5)
Prevention	97 (8.2)
Health services research	84 (7.1)
Screening	81 (6.8)
Supportive care	77 (6.5)
Basic science	57 (4.8)
Device feasibility	20 (1.7)

[a]Tabulations exclude missing data, which may result in totals below 3106.

[b]Data on time perspective (eg, prospective) was only available for observational studies.

[c]Excludes withdrawn studies (n=0) and those with missing data.

[d]FDA: US Food and Drug Administration.

[e]Data on primary purpose was only available for interventional trials.

XSL•FO
RenderX

Table 2. Additional characteristics of artificial intelligence/machine learning studies (N=3106) registered on ClinicalTrials.gov, 2010-2023[a].

Characteristics	Values
Sex (n=3104), n (%)	
All	2863 (92.2)
Female	192 (6.2)
Male	49 (1.6)
Age (n=2896), n (%)	
Children (<18 y)	242 (8.4)
Adults (18-65 y)	1684 (58.1)
Elderly people (>65 y)	21 (0.7)
Children and adults	110 (3.8)
Adults and elderly people	796 (27.5)
All	43 (1.5)
Healthy volunteers (n=2854), n (%)	
Accepts healthy volunteers	1006 (35.2)
Does not accept healthy volunteers	1848 (64.8)
Lead sponsor sector (n=3106), n (%)	
Hospital/clinic	1373 (44.2)
Academia	869 (28)
Industry	407 (13.1)
Research institute	156 (5)
Individual	151 (4.9)
Government	71 (2.3)
Nonprofit	57 (1.8)
Network	9 (0.3)
Unknown	13 (0.4)
Funding source (n=3106), n (%)	
Industry	518 (16.7)
National Institutes of Health	143 (4.6)
Other	2445 (78.7)
Study location: region (n=2821), n (%)	
Europe	973 (34.5)
Asia/Pacific	841 (29.8)
North America	731 (25.9)
Middle East	84 (3)
Africa	37 (1.3)
Central and South America	31 (1.1)
Multiple regions	124 (4.4)
Study location: gross national income (n=3106)[b], n (%)	
High income	2340 (75.3)
Upper middle income	675 (21.7)
Lower middle income	88 (2.8)
Low income	3 (0.1)

XSL•FO
RenderX

Characteristics	Values
Study results (completed, n=842), n (%)	
Has results	47 (5.6)
No results	795 (94.4)
Study results (at least 1 year after primary completion date, n=691), n (%)	
Has results	45 (6.5)
No results	646 (93.5)
Study results (time in days from primary completion to result posting; n=47)	
Minimum	93
Maximum	1200
Median (IQR)	505 (399-676)
Clinical specialty (n=3245)[c], n (%)	
Neoplasms (includes cancer screening)	420 (12.9)
Nervous system diseases	395 (12.2)
Cardiovascular diseases	356 (11)
Pathological conditions, signs and symptoms	325 (10)
Respiratory tract diseases	275 (8.5)
Digestive system diseases	253 (7.8)
Mental disorders	219 (6.7)
Endocrine, nutritional, or metabolic diseases	161 (5)
Female urogenital diseases and pregnancy complications	138 (4.3)
Skin and connective tissue diseases	89 (2.7)
Eye diseases	79 (2.4)
Musculoskeletal diseases	60 (1.8)
Behavior and behavior mechanisms	56 (1.7)
Male urogenital diseases	48 (1.5)
Surgical procedures, operative	38 (1.2)
Wounds and injuries	34 (1)
Chemically induced disorders	34 (1)
Bacterial infections and mycoses	33 (1)
Hemic and lymphatic diseases	31 (1)
Stomatognathic diseases	28 (0.9)
Radiology (not organ specific)	26 (0.8)
Otorhinolaryngologic diseases	22 (0.7)
Physiological phenomena	21 (0.6)
Immune system diseases	19 (0.6)
Congenital, hereditary, and neonatal diseases and abnormalities	18 (0.6)
Viral diseases	17 (0.5)
Critical care	14 (0.4)
Urology/nephrology	13 (0.4)
Geriatrics	11 (0.3)
Environment and public health	7 (0.2)

XSL•FO
RenderX

Characteristics	Values
Circulatory and respiratory physiological phenomena	5 (0.2)

[a]Tabulations exclude missing data, which may result in totals below 3106.

[b]Gross national income level grouping as per the World Bank.

[c]Multiple counts per study are possible and can exceed 3106.

Study Design

Among observational studies, a prospective design was more common (58.9%; 1126/1913) than retrospective (22.5%; 430/1913) or cross-sectional (11%; 210/1913) designs. Among trials, 56.2% (670/1193) were randomized, 10.9% (130/1193) were nonrandomized, and 32.8% (393/1193) were "not applicable" (Table 1). Figure 1 shows the numbers of all studies, including randomized controlled trials (RCTs) and prospective, retrospective, and cross-sectional studies, over time. From 2017 to 2023, the proportion of RCTs increased by 42.1%: from 18.52% (15/81) to 26.32% (179/680). The proportion of prospective studies increased by 22.6%: from 29.63% (24/81) to 36.32% (247/680). In contrast, the proportion of retrospective studies decreased by 11.9%, from 12.35% (10/81) to 10.88% (74/680), and cross-sectional studies declined by 28.5%, from 8.64% (7/81) to 6.18% (42/680).

Figure 1. Study design characteristics by start year, 2010-2023. Excludes "unknown" or "other." RCT: randomized controlled trial.

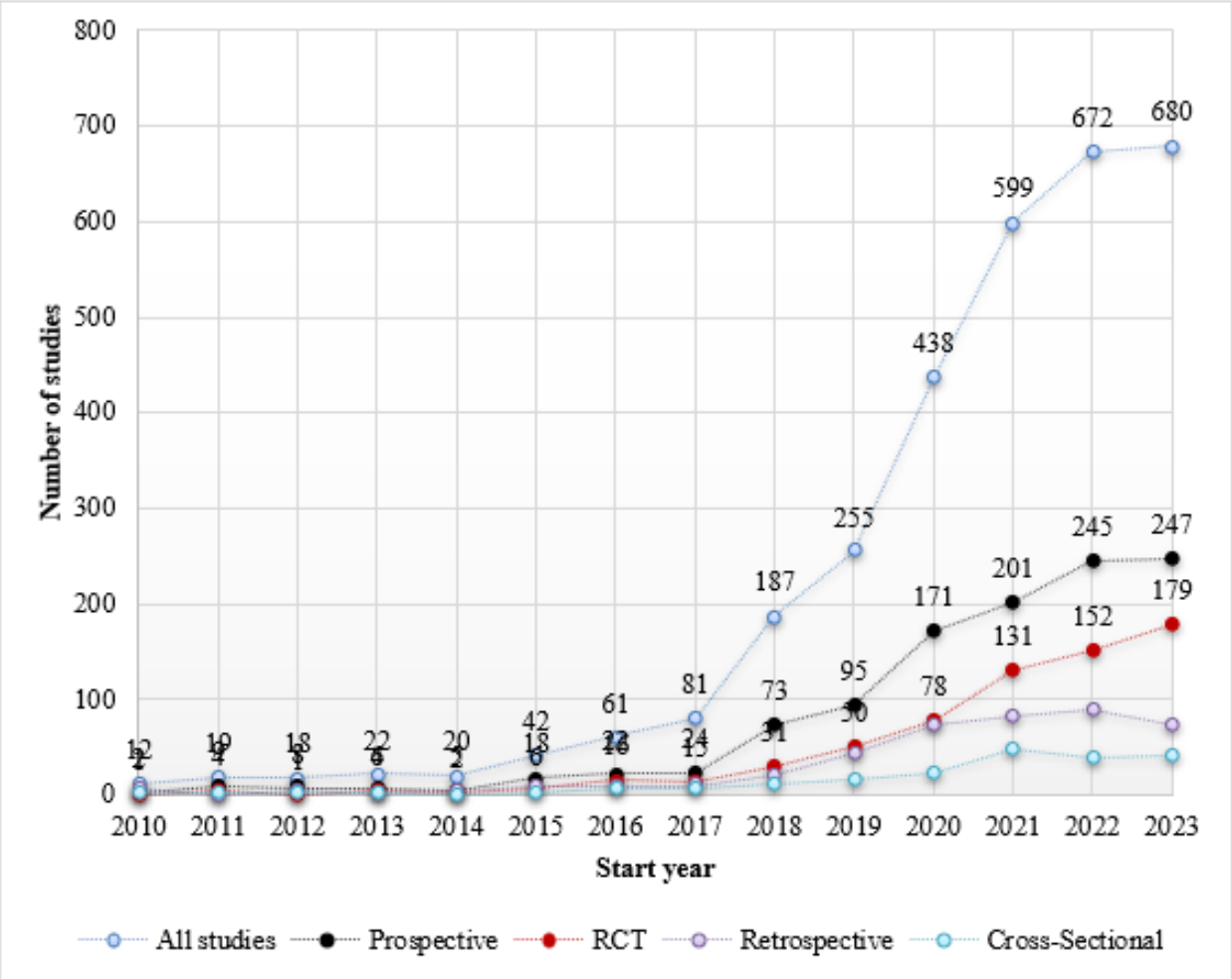

Primary Purpose

Data on primary purpose were only available for interventional trials. Among the 1190 trials, the most common purposes were diagnostic (28.2%; n=335), treatment (24.4%; n=290), "other" (12.5%; n=149), and prevention (8.2%; n=97). However, the top 2 rankings were reversed when restricted to RCTs only: treatment (31.4%; n=210), diagnostic (18.6%; n=124), "other" (11.1%; n=74), and prevention (10.9%; n=73). Over time, studies focused on diagnosis and treatment increased faster than others (Figure 2).

XSL•FO
RenderX

Figure 2. Primary study purposes by start year. Top: intervention studies; bottom: randomized trials, 2010-2023. †Excludes missing data.

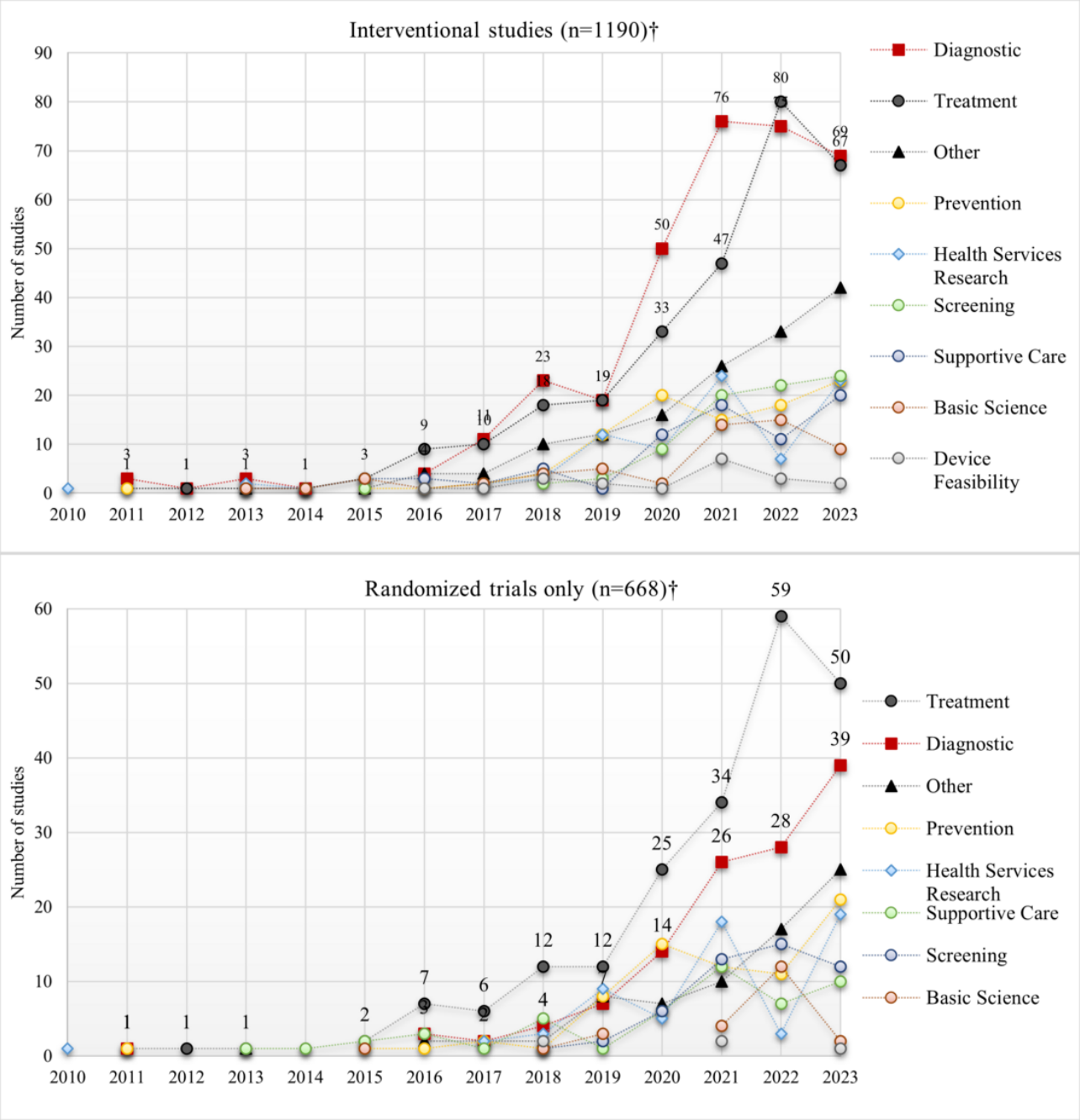

Clinical Specialty

The top 10 accounted for 81.1% (2631/3245) of all clinical specialties (Table 2). The 3245 specialties included neoplasms (12.9%; n=420); nervous system diseases (12.2%; n=395); cardiovascular diseases (11%; n=356); pathological conditions (10%; n=325); respiratory tract diseases (8.5%; n=275); digestive system diseases (7.8%; n=253); mental disorders (6.7%; n=219); endocrine, nutritional, or metabolic diseases (5%; n=161); female urogenital diseases and pregnancy complications (4.3%; n=138); and skin and connective tissue diseases (2.7%; n=89). Most of these started to increase during 2018-2020 (Figure S3 in Multimedia Appendix 2). See Table S3 in Multimedia Appendix 2 for the entire list of specialties over time.

Lead Sponsor Sectors

We compared lead sponsor sectors classified in ClinicalTrials.gov versus our reclassification based on sponsor names (n=3106). In the original classification, the most common sector was “OTHER” (83.4%; n=2590), followed by “INDUSTRY” (12.4%; n=384) and “OTHER_GOV” (3.2%; n=98). However, our reclassification revealed the strong presence of hospitals/clinics (44.2%; n=1373) and academia (28%; n=869), followed by industry (13.1%; n=407) (Table 2; Figure S8 in Multimedia Appendix 2). Studies sponsored by hospitals/clinics, academia, and industry have increased since 2017 (Figure 3).

XSL•FO
RenderX

Figure 3. Lead sponsor sector by start year, 2010-2023. Top: as classified in ClinicalTrials.gov.; bottom: our reclassification based on sponsor names. NIH: National Institutes of Health.

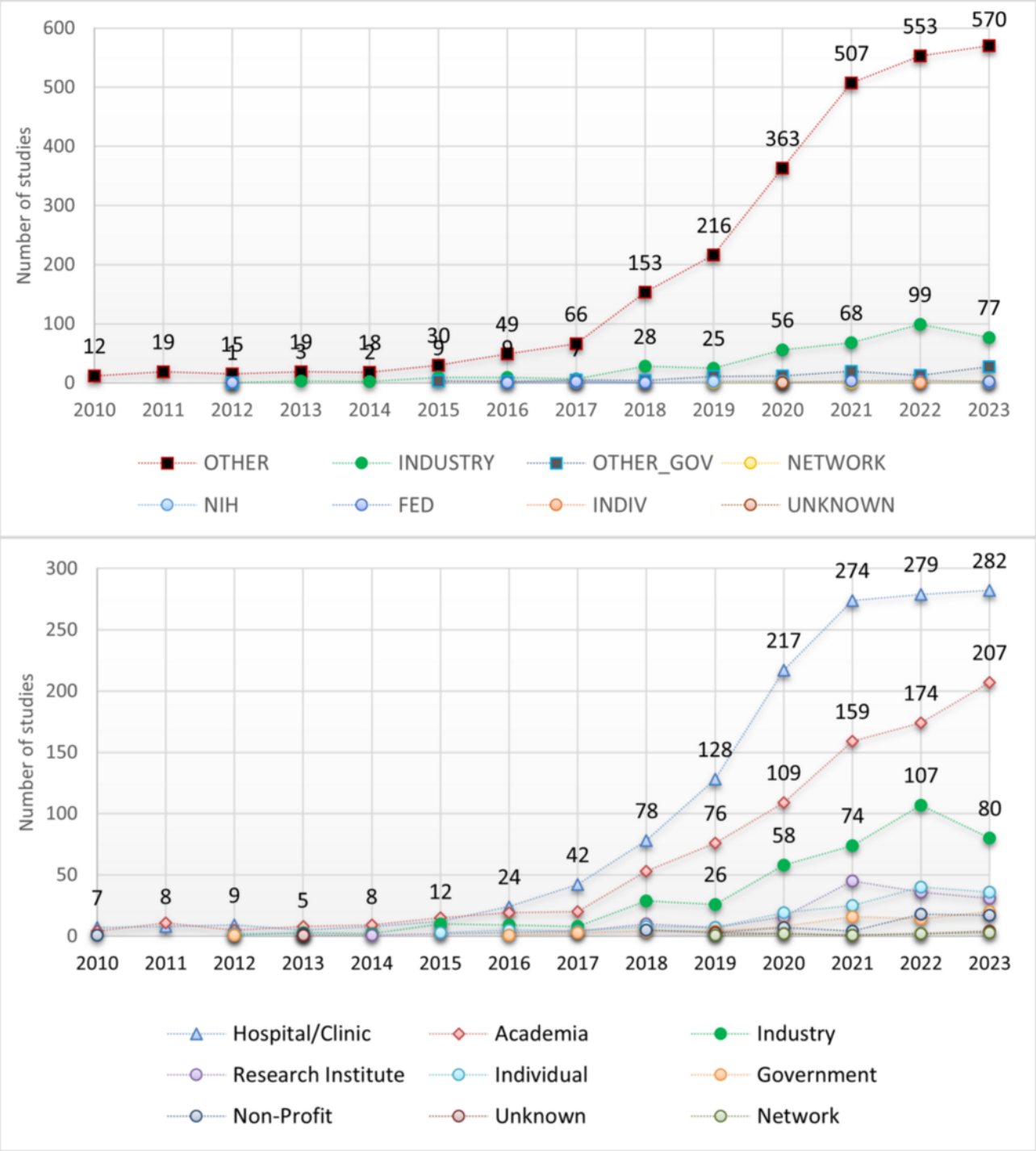

Enrollment (Sample Size)

The enrollment data were skewed to the right: maximum 13,977,257; mean 16,962 (SD 288,155); and median 255 (IQR 80-1000). The most common size category overall was 101-1000 (44.8%). Among trials, 1-100 (42.7%; 497/1163) was most common, followed by 101-1000 (41.2%; 479/1163), whereas 101-1000 (46.7%; 893/1913) was most represented among observational studies (Table 1). This pattern did not materially change when stratified by sponsor sectors (Table S4 in Multimedia Appendix 2). Traditionally, the largest studies tend to be industry-sponsored [23], but this was not the case in our data. Among observational studies, the hospital/clinic sector, followed by academia, represented the highest number of studies in all size categories, including the largest (>30,000). Among trials, the hospital/clinic sector represented the most up to the 10,000 category. While interventional trials in the 1-100 and 101-1000 categories and observational studies in the 101-1000 category have increased the most over time (Figure S5 in Multimedia Appendix 2), 16.1% (187/1163) of trials and 29% (551/1898) of observational studies were large, with enrollments above 1000.

XSL•FO
RenderX

We also compared the overall enrollment distribution (1-100, 101-1000, and >1000) with that of prior reports that used the ClinicalTrials.gov data (Figure S6 in Multimedia Appendix 2). First, our data distribution was similar to that of previous research on AI studies up to March 2022 [1]. Second, when compared to 2 previous reports on non–disease-specific trials (non-AI/ML) during 2000-2019 and 2007-2010, the proportion of the largest category (>1000) was higher in our AI/ML trials (16.1%; 187/1163) than in the non-AI/ML trials: (3.8%; 5174/135,144 in 2000-2019 [23] and 3.9%; 1103/28,467 in 2007-2010 [24]). When restricted to RCTs alone, the proportion of the >1000 category was even higher in our AI/ML trials (19.8%; 129/653). Conversely, the proportion of the smallest category (1-100) was lower in our trials (42.7%; 497/1163) than in the non-AI/ML trials (63.2% [23] and 62.3% [24]).

Sex-Specific Studies

The majority (92.2%; 2863/3104) was not sex-specific, followed by women-targeted (6.2%; 192/3104) and men-targeted (1.6%; 49/3104) studies (Table 2). Over time, AI/ML studies for women have increased since 2018 (Figure S7 in Multimedia Appendix 2).

Study Location—Global Distribution, Geographic and Economic

The most common regions were Europe (34.5%; 973/2821), Asia/Pacific (29.8%; 841/2821), and North America (25.9%; 731/2821; Table 2), followed by the Middle East (3%; 84/2821). The most common countries were the United States (21.8%; 677/3106), China (17.7%; 550/3106), France (6.6%; 205/3106), the United Kingdom (6.2%; 193/3106), Italy (5.2%; 162/3106), and Taiwan (3.6%; 111/3106). See Table S2 in Multimedia Appendix 2 for a full list of countries, with the World Bank's GNI-based distribution.

In Figure 4, 71 countries are color-divided into 38 high-income countries and 33 non–high-income countries. Most AI/ML studies were conducted in high-income countries (75.3%; 2340/3106), followed by upper-middle-income countries, including China (21.7%; 675/3106), lower-middle-income countries (2.8%; 88/3106), and low-income countries (0.1%; 3/3106) (Table 2). Studies in non–high-income countries have started to increase since 2020, particularly in Turkey (41 vs 50), Egypt (21 vs 24), India (15 vs 23), Brazil (15 vs 16), Russia (11 vs 14), and Mexico (8 vs 11), where the numbers in parentheses represent the totals for 2020-2023 compared to the totals for the entire period of 2010-2023.

Figure 4. Study locations (geographic and economic) in artificial intelligence/machine learning studies, 2010-2023. Numbers denote the number of studies in each location. Studies with multiple locations are included, as well as single-country studies. Studies with unknown locations (missing) are not included in the counts. HIC: high-income country (as per the World Bank classification).

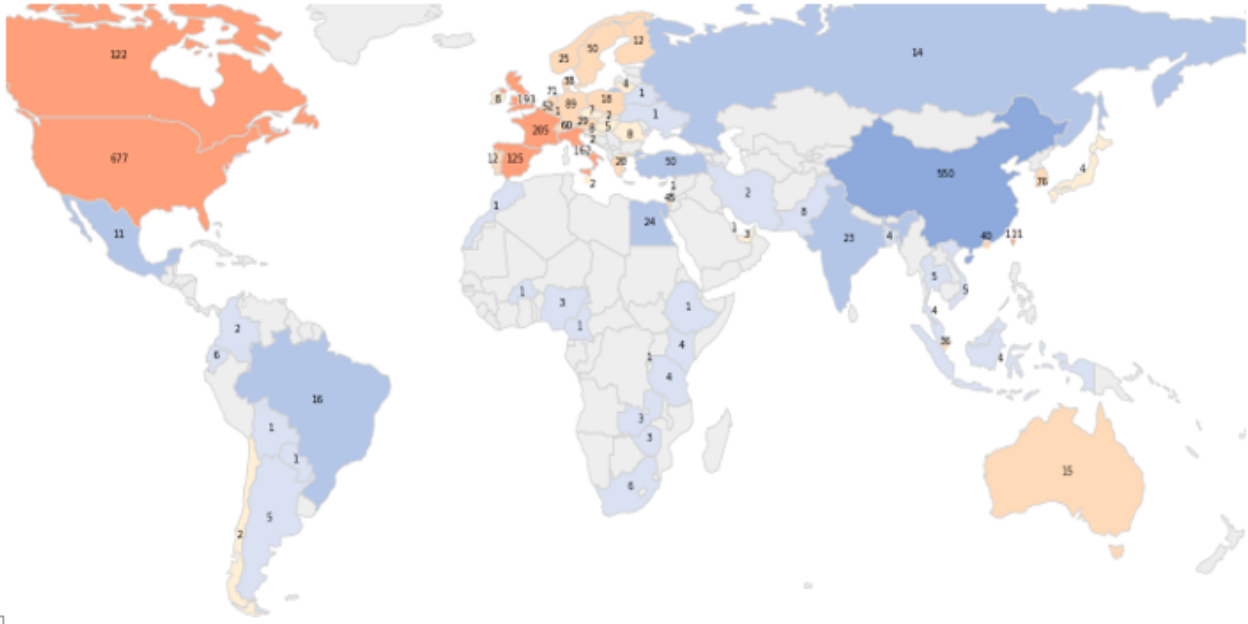

	HIC	Non-HIC
1 to 10 studies	14	26
11 to 100 studies	17	6
> 100 studies	7	1

As shown in Figure 5, studies in Europe, Asia/Pacific, and North America increased since 2018, and so did studies in high-income countries, with a modest increase in upper-middle-income countries, particularly China. The entries of lower-middle-income countries (in 2018) and lower-income countries (in 2021) lagged.

XSL•FO
RenderX

Figure 5. Study location by start year: (top) geographic and (bottom) economic. Single-country studies only (n=2696), 2010-2023.

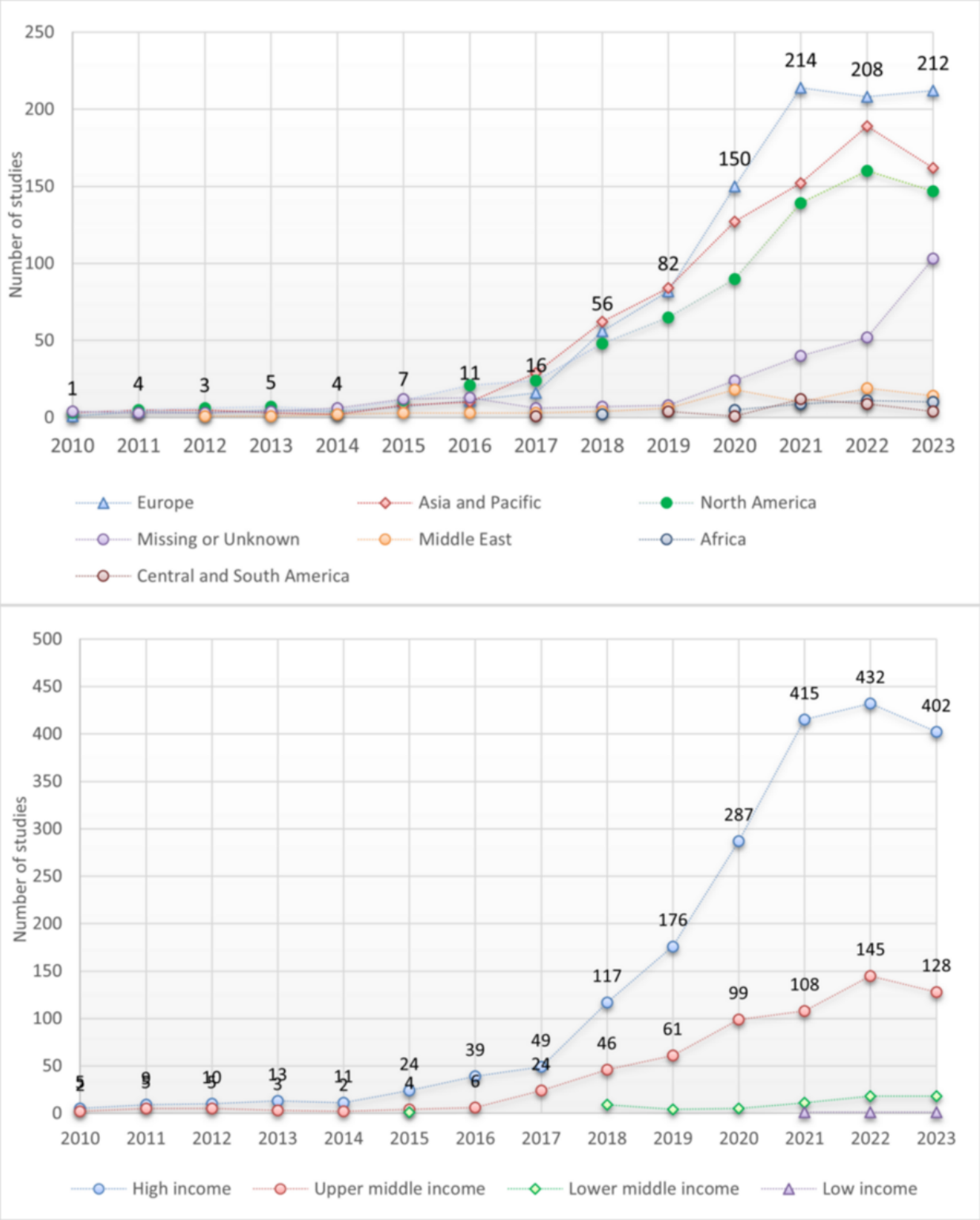

Posting of Study Results on ClinicalTrials.gov

Overall, only 5.6% (47/842) of completed studies had results available on ClinicalTrials.gov. Among studies with posted results (n=47), the median time from "primary completion date" to "results first posted date" was 505 days (IQR 399-700; minimum 93, maximum 1200). During 2018-2022, as the rapid increase in AI/ML research became evident, completed studies increased from 55 to 227, but the reporting rate stagnated, remaining between 5.3% and 7.3% (Table 3). Consequently, the number of completed studies without posted results increased (Figure 6).

XSL•FO
RenderX

Table 3. Availability of results posted on ClinicalTrials.gov by primary completion year among completed studies, 2010-2023.

Primary completion year	Has results	No results	Total	Percentage of studies with results
2010	0	1	1	0
2011	0	2	2	0
2012	0	3	3	0
2013	0	3	3	0
2014	0	7	7	0
2015	0	9	9	0
2016	2	18	20	10
2017	4	21	25	16
2018	3	52	55	5.5
2019	6	76	82	7.3
2020	8	101	109	7.3
2021	10	138	148	6.8
2022	12	215	227	5.3
2023	2	145	147	1.4
Total	47	791	838[a]	5.6

[a]n=838 instead of 842 after excluding 4 studies that reported 2024 as the primary completion year.

Figure 6. Availability of results posted on ClinicalTrials.gov by primary completion year among completed studies, 2010-2023 (n=838 instead of 842 after excluding 4 studies that reported 2024 as the primary completion year).

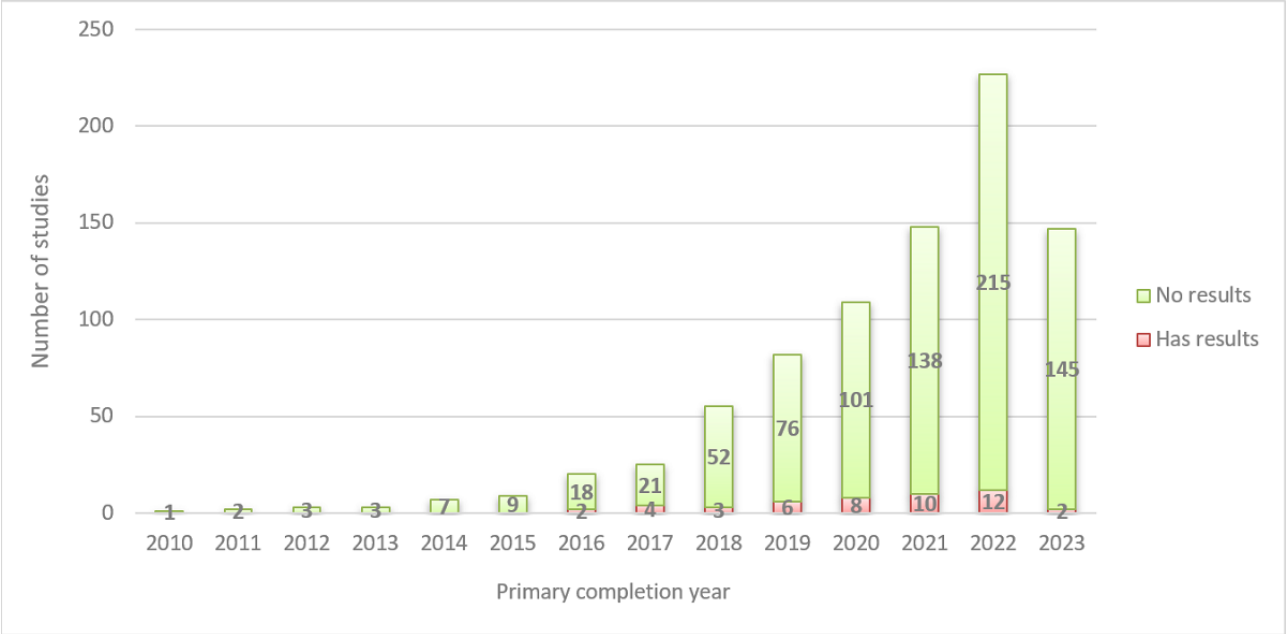

AI/ML Terms Used in Study Descriptions

The most frequently used term in study descriptions was "artificial intelligence" (37%), followed by "machine learning" (31.4%), "deep learning" (12.3%), "convolutional neural network" (2.6%), and "random forest" (2.1%; Figure S2 in Multimedia Appendix 2). The use of "artificial intelligence," "machine learning," and "deep learning" has increased since 2017, especially the former two (Figure S4 in Multimedia Appendix 2).

Discussion

Of 3106 AI/ML studies over the past 14 years, 63% (n=1951) started in the last 3 years alone; more than half (56.7%; 1762/3106) have set the primary completion date between 2023 and 2050, with a median time of 7 years (IQR 3.5-12) beyond 2023. Thus, our data indicate the general direction of the AI/ML pipeline in the years ahead.

Principal Findings

XSL•FO
RenderX

In determining whether the accelerated growth of AI/ML research reflects an increase in desirable study attributes or otherwise, we found uneven growth occurring in three areas: (1) study design, (2) study location, and (3) study results availability. First, the increase in randomized (interventional) and prospective (observational) designs, along with large studies (n>1000), representing a quarter of all studies, may indicate a move toward a prospective comparative evaluation stage. Second, much of the rapid growth in AI/ML studies comes from high-income countries in high-resource settings. Third, the availability of results posted on ClinicalTrials.gov remained limited, and thus, the number of completed studies without posted results increased over time. While the first point seems promising, the other two perpetuate previously raised issues.

Study Design

Previous research on AI studies registered on ClinicalTrials.gov as of March 2022 concluded that the proportion of trials with prospective (59.92%) and randomized (52.7%) designs was insufficient [1]. We also observed similar proportions (prospective: 58.9%; randomized: 56.2%) as of December 2023. However, from 2017 to 2023, the proportions of RCT and prospective designs increased by 42.1% and 22.6%, respectively, while retrospective and cross-sectional studies decreased by 11.9% and 28.5%, respectively. This is deemed an improvement compared to the period when retrospective assessments were dominant [11].

Although the choice of RCTs depends on the stage of development, the "phase" data, more aligned with drug development, was mostly "not applicable" (93%) in our data (Table 1). Referring to a stage-specific reporting guideline for the early and live clinical evaluation of decision-support systems based on AI, which compares development pathways among drugs, AI, and surgical innovation [25], the recent increase in prospective and randomized designs may be a reflection of enough studies progressing from the in silico evaluation stage to the early live clinical evaluation stage (trial phase 1-2 in drugs) and the comparative prospective evaluation stage (trial phase 3 in drugs).

Enrollment (Sample Size)

AI models, especially deep learning models, often require large datasets to achieve robust performance [26], with sample sizes above 1000 being common. This was empirically reflected in our data overall. When compared to non-AI/ML trials registered on ClinicalTrials.gov during 2000-2019 [23,24], the proportion of smaller trials (1-100) was lower, and the proportion of larger trials (>1000) was higher in our data (Figure S6 in Multimedia Appendix 2). Furthermore, 16.1% (187/1163) of trials had a sample size of >1000, 72.2% (135/187) of which were not completed yet. Similarly, 19.8% of RCTs (129/653, excluding withdrawn) were in the >1000 category, but 68.2% (88/129) were still ongoing. Thus, the majority of large AI/ML trials in our data are still in the pipeline and yet to be captured in future systematic reviews.

Lead Sponsor Sector

Our reclassification of the "other" groups revealed the strong presence of hospital/clinic or academic sponsorship over industry (Figure 3). The hospital/clinic or academic sectors might be more likely than the industry to engage in AI/ML research even if it lacks commercialization potential. The higher representation of hospitals/clinics and academia might reflect a high volume of nonproprietary AI/ML.

The Underrepresented

We examined two areas. First, the underrepresentation of women in research [27,28] is well documented. In this study, we found a gradual increase in women-targeted studies over time. Second, 75.3% (2340/3106) of AI/ML studies were conducted in high-income countries, and only 2.8% (88/3106) and 0.1% (3/3106) were conducted in lower-middle-income and low-income countries, respectively. If AI/ML models poorly generalize to people other than those whose data were used to train the algorithms, it may exacerbate existing disparities. While external validation and model recalibration may help mitigate health care inequities due to data disparity [14], inappropriate external validation (and the lack thereof) in AI/ML research was already highlighted [9,29,30]. Validation using unseen data is particularly important in models built with single-center data [31], and most of our sample (77.2%; 2177/2821) indeed comprised single-center studies. However, external validation was mentioned in only 47 of 3106 studies (1.5%). Of these 47, only 9 (19.1%) were conducted in non–high-income countries: China (n=6), Argentina (n=1), Ecuador (n=1), and Russia (n=1), as part of a 10-country study (all were high-income countries, except Russia).

Availability of Results on ClinicalTrials.gov

While the number of newly initiated studies increased, the number of completed studies without posted results also increased. The limited reporting of even basic results on ClinicalTrials.gov contrasts with the rapid advancements in the field and the registry's stated purpose of registration and result submission, which includes mitigating publication bias and outcome reporting bias [32].

Outside the AI/ML field, high rates of results dissemination on ClinicalTrials.gov have been reported, including for cancer trials and small studies [33]. With AI/ML research results more accessible on ClinicalTrials.gov, the registry can play a crucial role in reducing publication bias, making it a valuable resource for future systematic reviews or meta-analyses. Although posting results on ClinicalTrials.gov means disseminating data without independent scientific review, summary results are reported in structured tables, and the results information is devoid of conclusions or "spin," as the system requires objective data rather than subjective narratives [34]. This allows for broad scrutiny by the scientific community, the public, or competitors. Although results posting on ClinicalTrials.gov is mandated only for certain types of trials, all registered studies are eligible for results submission, and the responsibility of providing knowledge given the use of human data should equally apply to both trials and observational studies.

Limitations

Our study limitations are as follows. The first is representativeness. Sponsors of FDA-regulated studies and trials with sites in the United States are required to submit information

XSL•FO
RenderX

to ClinicalTrials.gov [35]. In our sample, only 7.6% (235/3106) of studies were FDA regulated, while 78.2% (2429/3106) did not involve a US site. When registration is voluntary, it is hard to determine what the data are truly representative of. Interventional trials are more likely to be registered in a public registry due to this being a requirement of the International Committee of Medical Journal Editors as a condition of consideration for publication [36]. However, as this requirement started in 2005 and precedes our data cutoffs (2010-2023), it is unlikely to have influenced our findings (RCTs were infrequent until 2017). Moreover, registered studies would be more likely to share results than unregistered ones, but the reporting rate in our sample was unexpectedly low. Nevertheless, by evaluating trends over time under the same representativeness constraints, we made the best use of the most recent data available.

Second, as we prioritized sensitivity over specificity, our sample may have included studies in which AI/ML played a minor role. Identifying such studies would have required manual free-text data mining with prespecified criteria, which was impractical during testing due to substantial variations in reporting quality, length, and clarity. Poor reporting in AI/ML research has already been highlighted [8,9], and AI-specific reporting guidelines and documentation standards now exist to increase awareness of key dimensions that should be described in study protocols (the SPIRIT-AI Extension, based on SPIRIT [Standard Protocol Items: Recommendations for Interventional Trials]) [37], prediction models (the TRIPOD+AI Statement, based on TRIPOD [Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis]) [38], or the minimum information about clinical AI modeling (the MI-CLAIM checklist) [39]. The same principle could also apply to study descriptions on ClinicalTrials.gov, and more standardized content within the platform would benefit its audience.

Conclusions

Much of the rapid growth in AI/ML studies comes from high-income countries in high-resource settings, albeit with a modest increase in upper-middle-income countries (mostly China). Lower-middle or low-income countries remain poorly represented. The increase in randomized or prospective designs, along with large studies (n>1000), which represent a quarter of all studies (mostly ongoing), may indicate that enough studies are progressing from the in silico evaluation stage toward a prospective comparative evaluation stage. However, the ongoing limited availability of basic results posted on ClinicalTrials.gov contrasts with the rapid advancements in this field and the public registry's role in reducing publication and outcome reporting biases.

Acknowledgments

The authors thank Prof Ray Rogers from the University of Adelaide for his helpful suggestions. This work was in part supported by the Graduate School of Medicine, Juntendo University.

Data Availability

The search strategy is provided in Multimedia Appendix 2. The pSQL codes used are provided in Multimedia Appendix 1. The datasets generated during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Study conception and design were done by SM. Data acquisition and management were done by MDM. SM conducted data analysis and performed descriptive analyses. The manuscript was written by SM, with all authors (SM, RK, MDM, and RJS Jr) revising it for intellectual content and approving the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Technical document.
[DOCX File , 53 KB - jmir_v26i1e57750_app1.docx]

Multimedia Appendix 2
Supplementary file (Figures, Tables).
[DOCX File , 224 KB - jmir_v26i1e57750_app2.docx]

References

1. Wang A, Xiu X, Liu S, Qian Q, Wu S. Characteristics of artificial intelligence clinical trials in the field of healthcare: a cross-sectional study on ClinicalTrials.gov. Int J Environ Res Public Health 2022;19(20):13691 [FREE Full text] [doi: 10.3390/ijerph192013691] [Medline: 36294269]

XSL•FO
RenderX

2. Zippel C, Bohnet-Joschko S. Rise of clinical studies in the field of machine learning: a review of data registered in ClinicalTrials.gov. Int J Environ Res Public Health 2021;18(10):5072 [FREE Full text] [doi: 10.3390/ijerph18105072] [Medline: 34064827]
3. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
4. E Moura FS, Amin K, Ekwobi C. Artificial intelligence in the management and treatment of burns: a systematic review. Burns Trauma 2021;9:tkab022 [FREE Full text] [doi: 10.1093/burnst/tkab022] [Medline: 34423054]
5. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
6. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022;5(1):156 [FREE Full text] [doi: 10.1038/s41746-022-00699-2] [Medline: 36261476]
7. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020;8(7):e18599 [FREE Full text] [doi: 10.2196/18599] [Medline: 32706688]
8. Collins GS. Making the black box more transparent: improving the reporting of artificial intelligence studies in healthcare. BMJ 2024;385:q832. [doi: 10.1136/bmj.q832] [Medline: 38626954]
9. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
10. Alami H, Lehoux P, Auclair Y, de Guise M, Gagnon M, Shaw J, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020;22(7):e17707 [FREE Full text] [doi: 10.2196/17707] [Medline: 32406850]
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
12. Potnis KC, Ross JS, Aneja S, Gross CP, Richman IB. Artificial intelligence in breast cancer screening: evaluation of FDA device regulation and future recommendations. JAMA Intern Med 2022;182(12):1306-1312 [FREE Full text] [doi: 10.1001/jamainternmed.2022.4969] [Medline: 36342705]
13. Barocas S, Selbst AD. Big data's disparate impact. 104 Calif Law Rev 2016;104:671-732. [doi: 10.2139/ssrn.2477899]
14. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. PLOS Digit Health 2022;1(3):e0000022 [FREE Full text] [doi: 10.1371/journal.pdig.0000022] [Medline: 36812532]
15. Clinical Trials Transformation Initiative Aggregate Analysis (AACT) database. URL: https://aact.ctti-clinicaltrials.org [accessed 2023-12-05]
16. AACT data dictionary. Clinical Trials Transformation Initiative. URL: https://aact.ctti-clinicaltrials.org/data_dictionary [accessed 2024-03-08]
17. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
18. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
19. Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open 2020;10(3):e034568 [FREE Full text] [doi: 10.1136/bmjopen-2019-034568] [Medline: 32205374]
20. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning–based prediction models. J Clin Epidemiol 2023;158:99-110 [FREE Full text] [doi: 10.1016/j.jclinepi.2023.03.024] [Medline: 37024020]
21. World Bank Country and Lending Groups. The World Bank. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519 [accessed 2024-03-01]
22. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
23. Gresham G, Meinert JL, Gresham AG, Meinert CL. Assessment of trends in the design, accrual, and completion of trials registered in ClinicalTrials.gov by sponsor type, 2000-2019. JAMA Netw Open 2020;3(8):e2014682 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.14682] [Medline: 32845329]
24. Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. JAMA 2012;307(17):1838-1847. [doi: 10.1001/jama.2012.3424] [Medline: 22550198]

XSL•FO
RenderX

25. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
26. Alganmi N. A comprehensive review of the impact of machine learning and omics on rare neurological diseases. BioMedInformatics 2024;4(2):1329-1347. [doi: 10.3390/biomedinformatics4020073]
27. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020;3:81 [FREE Full text] [doi: 10.1038/s41746-020-0288-5] [Medline: 32529043]
28. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020;2(5):e221-e223 [FREE Full text] [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
29. Vedula SS, Tsou BC, Sikder S. Artificial intelligence in clinical practice is here-now what? JAMA Ophthalmol 2022;140(4):306-307. [doi: 10.1001/jamaophthalmol.2022.0040] [Medline: 35238882]
30. d'Elia A, Gabbay M, Rodgers S, Kierans C, Jones E, Durrani I, et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Fam Med Community Health 2022;10:e001670 [FREE Full text] [doi: 10.1136/fmch-2022-001670] [Medline: 36450391]
31. Mitchell WG, Dee EC, Celi LA. Generalisability through local validation: overcoming barriers due to data disparity in healthcare. BMC Ophthalmol 2021;21(1):228 [FREE Full text] [doi: 10.1186/s12886-021-01992-6] [Medline: 34020592]
32. Tse T, Williams RJ, Zarin DA. Reporting "basic results" in ClinicalTrials.gov. Chest 2009 Jul;136(1):295-303. [doi: 10.1378/chest.08-3022] [Medline: 19584212]
33. Zwierzyna M, Davies M, Hingorani AD, Hunter J. Clinical trial design and dissemination: comprehensive analysis of ClinicalTrials.gov and PubMed data since 2005. BMJ 2018;361:k2130 [FREE Full text] [doi: 10.1136/bmj.k2130] [Medline: 29875212]
34. Zarin DA, Fain KM, Dobbins HD, Tse T, Williams RJ. 10-Year update on study results submitted to ClinicalTrials.gov. N Engl J Med 2019;381(20):1966-1974 [FREE Full text] [doi: 10.1056/NEJMsr1907644] [Medline: 31722160]
35. Glass HE, Glass LM, DiFrancesco JJ. ClinicalTrials.gov: an underutilized source of research data about the design and conduct of commercial clinical trials. Ther Innov Regul Sci 2015;49(2):218-224. [doi: 10.1177/2168479014551643] [Medline: 30222415]
36. Clinical trials. International Committee of Medical Journal Editors. URL: https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html [accessed 2024-06-01]
37. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AICONSORT-AI Working Group, SPIRIT-AICONSORT-AI Steering Group, SPIRIT-AICONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26(9):1351-1363 [FREE Full text] [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
38. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378 [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
39. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26(9):1320-1324 [FREE Full text] [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]

Abbreviations

AI: artificial intelligence
CTTI AACT: Clinical Trials Transformation Initiative Aggregate Analysis of ClinicalTrials.gov
FDA: Food and Drug Administration
GNI: gross national income
MeSH: Medical Subject Headings
MI-CLAIM: minimum information about clinical AI modeling
ML: machine learning
pSQL: PostgreSQL
RCT: randomized controlled trial
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD: Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.04.24; peer-reviewed by Z Hou, PH Liao; comments to author 02.05.24; revised version received 11.06.24; accepted 16.08.24; published 25.10.24.

Please cite as:
Maru S, Matthias MD, Kuwatsuru R, Simpson Jr RJ
Studies of Artificial Intelligence/Machine Learning Registered on ClinicalTrials.gov: Cross-Sectional Study With Temporal Trends, 2010-2023
J Med Internet Res 2024;26:e57750
URL: https://www.jmir.org/2024/1/e57750
doi:10.2196/57750
PMID:

©Shoko Maru, Michael D Matthias, Ryohei Kuwatsuru, Ross J Simpson Jr. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adolescent and Parent Perspectives on Digital Phenotyping in Youths With Chronic Pain: Cross-Sectional Mixed Methods Survey Study

Bridget A Nestor[1,2], PhD; Justin Chimoff[1], BA; Camila Koike[1], BA; Elissa R Weitzman[3,4,5], MSc, ScD; Bobbie L Riley[1,2], MD; Kristen Uhl[6,7], PhD; Joe Kossowsky[1,2,8], MMSc, PhD

[1]Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital, Boston, MA, United States
[2]Department of Anesthesia, Harvard Medical School, Boston, MA, United States
[3]Division of Adolescent and Young Adult Medicine, Boston Children's Hospital, Boston, MA, United States
[4]Department of Pediatrics, Harvard Medical School, Boston, MA, United States
[5]Division of Addiction Medicine, Boston Children's Hospital, Boston, MA, United States
[6]Department of Psychosocial Oncology and Palliative Care, Dana Farber Cancer Institute, Boston, MA, United States
[7]Department of Psychiatry, Boston Children's Hospital, Boston, MA, United States
[8]Division of Sleep Medicine, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Bridget A Nestor, PhD
Department of Anesthesiology, Critical Care, and Pain Medicine
Boston Children's Hospital
1 Autumn St
Boston, MA, 02215
United States
Phone: 1 617 355 0965
Email: bridget.nestor@childrens.harvard.edu

Abstract

Background: Digital phenotyping is a promising methodology for capturing moment-to-moment data that can inform individually adapted and timely interventions for youths with chronic pain.

Objective: This study aimed to investigate adolescent and parent endorsement, perceived utility, and concerns related to passive data stream collection through smartphones for digital phenotyping for clinical and research purposes in youths with chronic pain.

Methods: Through multiple-choice and open-response survey questions, we assessed the perspectives of patient-parent dyads (103 adolescents receiving treatment for chronic pain at a pediatric hospital with an average age of 15.6, SD 1.6 years, and 99 parents with an average age of 47.8, SD 6.3 years) on passive data collection from the following 9 smartphone-embedded passive data streams: accelerometer, apps, Bluetooth, SMS text message and call logs, keyboard, microphone, light, screen, and GPS.

Results: Quantitative and qualitative analyses indicated that adolescents and parent endorsement and perceived utility of digital phenotyping varied by stream, though participants generally endorsed the use of data collected by passive stream (35%-75.7% adolescent endorsement for clinical use and 37.9%-74.8% for research purposes; 53.5%-81.8% parent endorsement for clinical and 52.5%-82.8% for research purposes) if a certain level of utility could be provided. For adolescents and parents, adjusted logistic regression results indicated that the perceived utility of each stream significantly predicted the likelihood of endorsement of its use in both clinical practice and research (P_s<.05). Adolescents and parents alike identified accelerometer, light, screen, and GPS as the passive data streams with the highest utility (36.9%-47.5% identifying streams as useful). Similarly, adolescents and parents alike identified apps, Bluetooth, SMS text message and call logs, keyboard, and microphone as the passive data streams with the least utility (18.5%-34.3% identifying streams as useful). All participants reported primary concerns related to privacy, accuracy, and validity of the collected data. Passive data streams with the greatest number of total concerns were apps, Bluetooth, call and SMS text message logs, keyboard, and microphone.

XSL•FO
RenderX

Conclusions: Findings support the tailored use of digital phenotyping for this population and can help refine this methodology toward an acceptable, feasible, and ethical implementation of real-time symptom monitoring for assessment and intervention in youths with chronic pain.

(J Med Internet Res 2024;26:e47781) doi:10.2196/47781

KEYWORDS

acceptability; adolescent; chronic pain; digital phenotyping; mobile health; pediatric

Introduction

Digital phenotyping (or personal sensing [1]) refers to the continuous, passive collection of social and behavioral data, often gathered through smartphone sensors, allowing for ecologically valid insight into an individual's daily life [2]. In recent years, this method of assessment has been implemented in various populations to measure patterns of mental health symptoms, sleep, activity, and other social behaviors [3,4]. Although this methodology holds great promise for capturing moment-to-moment data that can inform individually adapted and timely interventions, whether individuals feel comfortable with this style of data collection remains largely undetermined. This study investigates adolescent and parental attitudes toward passive data streams collected through smartphone for digital phenotyping in a sample of treatment-seeking youths with chronic pain.

For youths with chronic pain, digital phenotyping represents a promising means of both assessment and intervention, as smartphone ownership or access has become nearly universal in adolescence [5]. Detailed, in-clinic assessments of youths with chronic pain may ask adolescents and their parents to report on changes in their pain, activity, sleep, substance use, and other psychological symptoms to capture a biopsychosocial picture of the patient over time. Other types of assessments may use, exclusively or partially, survey data collected from participants on their smartphones. Whether through surveys or in-person, providing accurate reports of many of these features relevant to chronic pain is quite difficult, and studies have shown that retrospective bias obscures reports of mood [6], sleep [7], and activity [8]. As such, digital phenotyping may offer more accurate, sensitive, and objective symptom information for assessment. Moreover, youths with chronic pain may also be well-suited for digital, just-in-time adaptive interventions [9], as these interventions can be tailored for patients based on the continuous monitoring of the complex longitudinal interplay of their physical and mental health symptoms.

Indeed, passive data streams embedded in smartphones, including accelerometers, GPS sensors, call logs, and light sensors, among others, can continuously collect various metrics of an individual's in situ data, holding massive information about social and behavioral patterns. Previous studies have demonstrated that data collected through these streams correlate with validated self-report measures of physical and mental health across different populations, including college students with anxiety, patients undergoing chemotherapy, adults with schizophrenia, and persons with HIV [3,10-16]. More recently, this approach has been extended to youths, indicating that passive data streams collected through smartphone are significantly associated with internalizing symptoms in children and adolescents [17]. Passive data collection can also be combined with more active prompts to promote health behaviors, such as coping or mindfulness.

Findings from previous research highlight the ecological validity of digital phenotyping methodology and its capacity to provide true "snapshots" of an individual's real life. This approach does not rely on any participant report, which is often subject to recall bias [18], and for adolescents in particular, digital phenotyping assessment also alleviates the burden of parental report of adolescent symptom changes, reducing informant effects [19]. In addition, digital phenotyping allows for data collection in a person's own environment, thus diminishing issues related to the limited generalizability of laboratory-based findings [20]. Finally, digital phenotyping is unobtrusive; beyond the initial downloading of an app for passive data collection, no participant action is required, distinguishing it from other technologically advanced approaches to data collection, such as ecological momentary assessment, which holds high ecological validity but requires ongoing and somewhat burdensome participant engagement [21].

Despite the methodological strengths and appeal of digital phenotyping for assessment and intervention, concerns exist about its implementation for clinical and research purposes. At the forefront of these concerns are hesitations about patient privacy, confidentiality, and data security. Are individuals willing to consent to passive data streams on their smartphone collecting extensive personal data? Would individuals endorse the collection of some types of personal data but not others? Research into the general acceptability of digital phenotyping lags behind its implementation, though studies have begun to investigate this topic in more depth, revealing varying levels of acceptability as well as concerns, though findings describe adults [22,23]. Investigation of acceptability is also necessary for pediatric populations, for whom parental consent and youths' assent to data collection are required.

To that end, this study investigates adolescent and parent perspectives on endorsement, perceived utility, and concerns related to passive data stream collection through smartphone for digital phenotyping for clinical and research purposes in youths with chronic pain. In particular, we assessed participant perspectives on passive data collection through smartphone from the following 9 streams: screen, light, apps and installations, GPS, accelerometer, SMS text message and call logs, keyboard, Bluetooth, and microphone. We hypothesized that the majority of adolescents and their parents would endorse the use of digital phenotyping and that they would identify this methodology as useful. We further anticipated that participants' perceived utility of each passive data stream would predict their

XSL•FO
RenderX

endorsement of the stream's use, and that this would be true for both adolescents and parents in the adolescent-parent dyads. Finally, we expected that adolescents and parents would express concerns about privacy and data security.

Methods

This study was cross-sectional and adopted a mixed methods approach.

Recruitment and Procedure

This study was conducted for 15 months, from June 2021 to September 2022. Participants were recruited for study participation primarily at clinic visits and secondarily through other outreach (eg, letter, email, or secure patient portal message). Specifically, a member of the study team reviewed weekly pain clinic schedules at the hospital to identify eligible families with upcoming scheduled appointments (either in-person or through Zoom [Zoom Video Communications] telehealth) for study recruitment. The study team then contacted the clinical care team to obtain approval to briefly join potentially eligible patient appointments. With approval from clinical care providers, the study team then met with eligible families after the clinical portion of their appointment (either in-person or through Zoom) to share information about the study, answer any questions, and obtain verbal consent from a parent and assent from the adolescent patient. In addition, some eligible families were contacted through letter, email, or secure patient portal message to participate in the study and also given the option to opt out of participation and additional contact. If families did not opt out within 3 days, a member of the study team contacted them by phone to further explain the study, answer relevant questions, and obtain verbal informed consent. A total of 2 voicemails, with a standard script of study description, were left for families if they could not be contacted directly. If no response was received after 2 days, families were considered not interested in participating and were not contacted again.

Participants

Participants (ie, adolescent patients and their parents) were recruited from a pain treatment clinic at a pediatric hospital in the Northeast United States. Adolescent patients (n=103) were included in the study if they were (1) between 13 and 18 years of age and (2) receiving treatment for chronic pain at the pediatric hospital. Parents (n=99) were included if their child was eligible for study participation. All participants were able to understand and respond to questions in English. Of those approached for participation, 10 refused to participate due to disinterest in research or concerns about the time commitment of study participation. One participant dropped out after consenting to the study due to concerns about the time commitment of completing the survey.

Ethical Considerations

Institutional review board (IRB) approval (protocol number IRB-P00035845) was obtained in November 2020 to access protected health information in order to determine participant eligibility and conduct this study. Boston Children's Hospital (BCH) gave ethical approval for the study. The senior author (JK) served as the responsible principal investigator for this study. All patients provided informed consent (assent from adolescents and consent from parents). Of note, only verbal consent was obtained for this study, not written consent, because the BCH IRB determined this study met the exemption for written consent. This decision was determined because this study was deemed to be of no more than minimal risk of harm to participants and involved no procedures for which written consent is normally required outside of the research context. A member of the study team sent a secure survey link to each participating adolescent and parent who provided their respective verbal assent and consent by phone. Participants completed their surveys through Research Electronic Data Capture (REDCap; Vanderbilt University) [24,25], a secure, Health Insurance Portability and Accountability Act (HIPAA)–compliant, web-based application that collects and stores data. The data were deidentified. Participants were compensated with US $15 Amazon gift cards for their study participation.

Measures

Each participant completed a REDCap survey comprising demographic questions and then watched 9 short videos created by the study team, which described what each data stream does, explained what information would be collected from each stream, displayed what collected information would look like, and gave a couple of examples of how each data stream may be useful for participants' health care, specifically their pain assessment and treatment. Participants then answered multiple-choice and free-response questions about each passive data stream following each video. Demographic questions assessed who was completing the survey (eg, mother, father, guardian, or patient), their age, education, type of phone used, and comfort with using technology. Adolescent and parent comfort with technology was assessed on a 5-point Likert scale (eg, 1=very uncomfortable; 2=uncomfortable; 3=somewhat comfortable; 4=comfortable; and 5=very comfortable).

Between 4 and 9 multiple-choice questions probed adolescent and parent endorsement, perceived utility, and concerns related to each passive data stream. Participants also had the option to answer free-response questions about each stream. Specifically, the survey investigated adolescents' and parents' attitudes toward 9 passive data streams: accelerometer, apps, SMS text message and call logs, Bluetooth, keyboard, microphone, light, screen, and GPS. The following main outcomes were assessed: (1) endorsement was measured as participants' "yes" versus "no" response to having data from a particular passive data stream shared with their doctor for treatment planning, for research purposes, or with their parents; (2) perceived utility was measured on a 5-point Likert scale that queried how useful the information from a particular passive data stream would be in helping to understand how patients are doing (eg, 1=not useful at all; 2=a little useful; 3=somewhat useful; 4=useful; and 5=very useful); (3) perceived concerns were measured on a 5-point Likert scale to determine how strongly participants agreed with a given concern for a specific passive data stream (eg, 1=strongly disagree; 2=disagree; 3=neutral; 4=agree; and 5=strongly disagree). In addition, adolescents were also asked whether or not they would endorse their passively collected

XSL•FO
RenderX

data being shared with their parents. Concerns probed for each passive data stream varied, ranging from items querying the perceived accuracy and validity of the data collected to items assessing participant comfort with privacy or data security. Participants were also given the option to provide open-ended responses regarding any additional stream-specific concerns. Of note, all items were presented to both adolescents and parents, with slight word changes depending on whether items were adolescent-facing or parent-facing (eg, "I don't think this passive data stream would be useful for my treatment" vs "I don't think this passive data stream would be useful for my child's treatment"). In this article, we present item wording only for adolescent-facing questions. See Multimedia Appendix 1 for all survey items, which were created by the study team. The survey took approximately 30 minutes for participants to complete.

Data Analysis

This study was a content analysis with both quantitative and qualitative aspects.

Quantitative Analysis

SPSS software (version 28; IBM Corp) was used to conduct descriptive and inferential data analyses of endorsement, perceived utility, and concerns. We conducted separate binary logistic regression analyses for adolescents and for parents in order to statistically test whether perceived utility predicted endorsement of passive data streams for clinical practice and research. In addition, to account for interdependence within each adolescent and parent dyad, we also conducted dyadic analyses. We assessed for dyadic interdependence in primary outcome variables (ie, adolescent and parent endorsement for each passive data stream for clinical practice and research) with pairwise intraclass correlation coefficients [26]. For outcome variables that suggested significant dyadic interdependence, we conducted multilevel binary logistic regression analyses. Specifically, at the level of the dyad, we tested whether the perceived utility of each passive data stream predicted endorsement for clinical and research use. We also tested for within-dyad differences predicting endorsement and for within-dyad differences for perceived utility predicting endorsement.

Qualitative Analysis

An inductive coding process [27], which is considered an efficient methodology for qualitative data, was used to conduct qualitative analysis of open-ended responses in Microsoft Excel (Microsoft Corp). Our multistage approach involved 2 members of the study team independently reviewing open-ended responses, creating a coding framework, and coding responses into themes for thematic saturation. Coding frameworks and coded responses were then reviewed for consensus standard between study members. Discrepancies in coding responses for themes were rare (n=6) and resolved by discussion with an additional study team member. Each element of a given response was coded separately by theme. For example, if 1 response contained a concern about privacy and a concern about data, this response was considered to have 2 separate concerns. In addition, concerns were counted separately. For example, if a participant reported 2 distinct concerns about GPS, each concern counted separately. See Table S1 in Multimedia Appendix 2 for the codebook used in this study.

Results

Sample Characteristics

The sample consisted of 103 adolescents (mean age 15.64, SD 1.59 years) and 99 parents (mean age 47.79, SD 6.34 years); 82 adolescent-parent dyads were included in the sample. All parents and all but 1 adolescent reported having a smartphone device (ie, either an iPhone or Android). The majority of adolescents (81/103, 78.6%) and parents (69/99, 70%) reported that they were either comfortable or very comfortable with technology. See Table 1 for additional demographic information about the sample.

XSL•FO
RenderX

Table 1. Sample demographic information of this survey study conducted from June 2021 to September 2022 among adolescents with chronic pain (n=103) and their parents (n=99) at a pediatric hospital in the Northeast United States.

Characteristics	Adolescents, n (%)	Parents, n (%)
Gender		
Female	79 (76.7)	90 (91)
Male	14 (13.6)	9 (9)
Other	10 (9.7)	0 (0)
Phone used		
Android	15 (14.6)	27 (27)
iPhone	87 (84.5)	72 (73)
None	1 (1)	0 (0)
Comfort with technology		
Very uncomfortable	14 (13.6)	6 (6)
Uncomfortable	0 (0)	0 (0)
Somewhat comfortable	8 (7.8)	24 (24)
Comfortable	34 (33)	42 (42)
Very comfortable	47 (45.6)	27 (27)
Highest level of education		
Some high school	N/A[a]	1 (1)
High school or GED[b]	N/A	11 (11)
Associate degree	N/A	6 (6)
Some college	N/A	8 (8)
Bachelor's degree	N/A	40 (40)
Master's degree	N/A	25 (25)
Professional or doctoral degree	N/A	8 (8)

[a]N/A: not applicable.

[b]GED: General Educational Development.

Adolescent and Parent Endorsement of Passive Data Stream Use

All adolescents and parents provided responses regarding endorsement of passive data stream use. Adolescent endorsement of passive data streams in clinical practice varied by stream and ranged from 35% (36/103; ie, keyboard and microphone) to 75.7% (78/103; ie, accelerometer). Only apps (49/103, 47.6%), keyboard (36/103, 35%), and microphone (36/103, 35%) received less than 50% endorsement. Adolescent endorsement for the use of passive data streams in research also varied by stream and ranged from 37.9% (39/103; ie, microphone) to 74.8% (77/103; ie, accelerometer) of adolescents reporting endorsement. Only keyboard (43/103, 41.7%) and microphone (39/103, 37.9%) received less than 50% endorsement.

Parent endorsement of passive data stream use in clinical practice similarly varied by stream and ranged from 54% (53/99; ie, microphone) to 82% (81/99; ie, accelerometer and screen). Parent endorsement for the use of passive data streams in research ranged from 53% (52/99; ie, microphone) to 83% (82/99; ie, accelerometer). Results of Mann-Whitney nonparametric tests of differences revealed significant differences between adolescents and parents for endorsement of apps (P=.02) and keyboard (P=.04) for clinical use; no significant differences emerged between adolescents and parents for endorsement of each passive data stream for research purposes. Tables 2 and 3 present the percentages of adolescents and parents who endorsed the use of each stream in clinical practice and research.

Adolescent endorsement for sharing their passively collected data with their parents varied by streams and ranged from 27.2% (28/103) for social activity (ie, Bluetooth, SMS text message and call logs, keyboard, and microphone) to 72.8% (75/103) for activity levels (ie, accelerometer). For phone usage (ie, apps), 39.8% (41/103) of adolescents provided endorsement for sharing their data with their parents; for sleep information (ie, light and screen), 47.6% (49/103) provided such endorsement; and for location (ie, GPS), 53.4% (55/103) provided such endorsement.

Dyadic agreement and disagreement for the endorsement of each passive data stream varied across streams for clinical practice and research use. Dyadic disagreement was greatest for apps (33/82, 40% disagreement for clinical use and 30/82,

XSL•FO
RenderX

37% disagreement for research), keyboard (29/82, 35% disagreement for clinical use and 31/82, 38% disagreement for research), and microphone (31/82, 38% disagreement for clinical use and 30/82, 37% disagreement for research). For passive data streams with the greatest dyadic disagreement, adolescent endorsement within disagreeing dyads was as follows: for apps, 24% (8/33) endorsed use for clinical and 30% (9/30) endorsed use for research; for keyboard, 24% (7/29) endorsed use for clinical and 29% (9/31) endorsed use for research; for microphone, 36% (11/31) endorsed use for clinical and 37% (11/30) endorsed use for research. See Figure S1 in Multimedia Appendix 2 for additional information about the dyadic endorsement of each stream.

Table 2. Adolescent and parent endorsement (yes or no) of the use of each passive data stream in clinical practice.

	Adolescents (n=103), n (%)	Parents (n=99), n (%)	*P* value
Accelerometer	78 (75.7)	81 (82)	.44
Apps	49 (47.8)	68 (69)	.02
Bluetooth	65 (63.1)	62 (63)	.36
SMS text and call log	58 (56.3)	65 (66)	.59
Keyboard	36 (35)	54 (55)	.04
Microphone	36 (35)	53 (54)	.34
Light	75 (72.8)	79 (80)	.94
Screen	74 (71.8)	81 (82)	.41
GPS	59 (57.3)	61 (62)	.44

Table 3. Adolescent and parent endorsement (yes or no) of the use of each passive data stream in research.

	Adolescents (n=103), n (%)	Parents (n=99), n (%)	*P* value
Accelerometer	77 (74.8)	82 (83)	.44
Apps	55 (53.4)	69 (70)	.15
Bluetooth	66 (64.1)	62 (63)	.36
SMS text and call log	60 (58.3)	64 (65)	.94
Keyboard	43 (41.7)	57 (58)	.07
Microphone	39 (37.9)	52 (53)	.42
Light	74 (71.8)	79 (80)	.96
Screen	72 (69.9)	78 (79)	.57
GPS	63 (61.2)	60 (61)	.27

Perceived Utility

All adolescents and parents provided responses regarding perceived utility. Figure 1 presents the percentages of adolescents and parents who perceived each passive data stream to be useful (defined as useful or very useful), somewhat useful, or not useful (defined as not useful at all or a little useful). Adolescents and parents alike identified accelerometer, light, screen, and GPS as the streams with the highest utility (ie, greatest percentage of participants considered streams useful or very useful). Similarly, adolescents and parents alike identified apps, Bluetooth, SMS text message and call logs, keyboard, and microphone as the streams with the least utility (ie, greatest number of participants considered streams not useful at all or a little useful). Results from Fisher exact tests of probability indicated no significant differences between adolescents and parents in terms of perceived utility for each passive data stream (P_s>.3). See Figure S2 in Multimedia Appendix 2 for perceived utility for each passive data stream.

XSL•FO
RenderX

Figure 1. Percentages of adolescent and parent perceived utility by passive data stream.

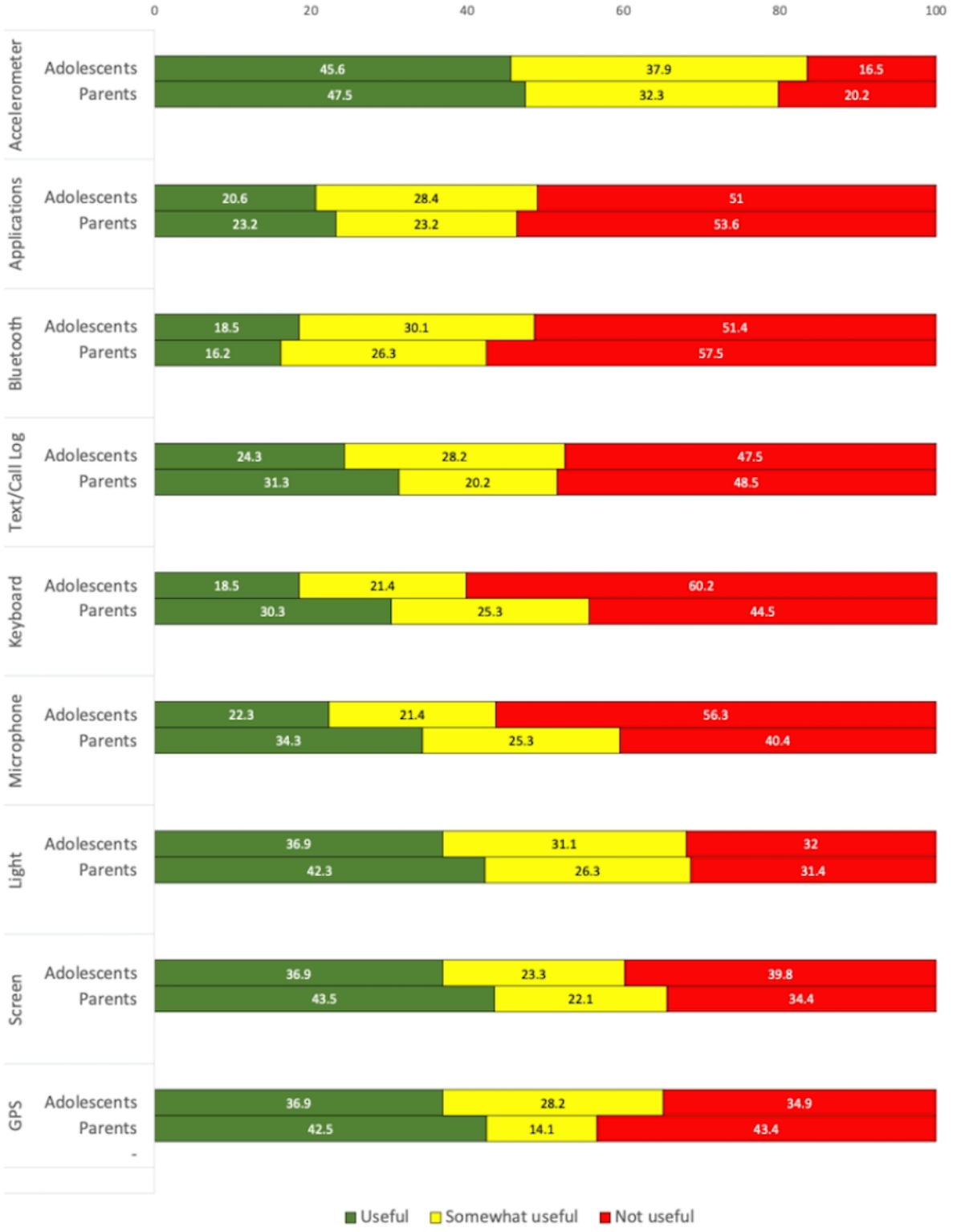

Utility Predicting Endorsement

For adolescents, logistic regression results indicated that the perceived utility of each stream significantly predicted the likelihood of endorsement of its use in both clinical practice (odds ratios [ORs] ranging from 2.90, 95% CI 1.77-4.76 to 4.95, 95% CI 2.34-10.48) and research (ORs ranging from 2.18, 95% CI 1.45-3.27 to 3.80, 95% CI 2.25-6.42), controlling for demographic variables (ie, age and gender). For parents, the same pattern of significance emerged: the perceived utility of each stream significantly predicted the likelihood of endorsement of its use in both clinical practice (ORs ranging from 2.68, 95% CI 1.72-4.19 to 6.91, 95% CI 2.90-16.41) and research (ORs ranging from 1.86, 95% CI 1.22-2.89 to 7.47, 95% CI 2.98-18.71), controlling for demographic variables (ie, age, gender, and education). See Section VI in Multimedia Appendix 2 for full logistic regression results predicting endorsement of each passive data stream.

XSL•FO
RenderX

Pairwise intraclass correlation coefficient results indicated significant (P_s<.001) dyadic interdependence for all primary outcomes of endorsement (ρ ranging from 0.17 to 0.49) with the exception of adolescent and parent endorsement of apps for clinical practice (*P*=.95) and for research (*P*=.95), and adolescent and parent endorsement of GPS for research (*P*=.14). Multilevel logistic regression results indicated that, at the level of the dyad, the perceived utility of each stream predicted the likelihood of endorsement of such streams for clinical (ORs ranging from 2.73, 95% CI 1.51-4.96 to 5.72, 95% CI 2.51-11.06) and for research use (ORs ranging from 1.89, 95% CI 1.16-3.09 to 5.08, 95% CI 2.49-10.36). No significant within-dyad differences predicting endorsement emerged. Similarly, no significant within-dyadic differences in perceived utility predicting endorsement emerged.

Perceived Concerns

The nature of the concerns varied for each passive data stream in question. For the accelerometer, adolescents and parents endorsed primary concerns with this stream missing activity due to not carrying phones at all times. Adolescents and parents, however, did not endorse significant concern about having their activity data shared. For apps, adolescents and parents reported primary concerns about data inaccuracies due to apps being on in the background; adolescents also endorsed significant concern, more so than parents, about having app usage shared with doctors. For Bluetooth, adolescents and parents endorsed less concern about sharing data from this stream with doctors than they did about this data being seen by people other than doctors. They also endorsed high levels of concern for this stream collecting inaccurate data due to phones not being carried at all times. For SMS text message and call logs, most adolescents and many parents indicated primary concern that they do not call people very often, so data from this stream would not be useful. Of note, over a third of adolescents also indicated concern that they use other apps, other than SMS text messages, to communicate. For light, a majority of adolescents (73/103, 70.9%) and parents (54/99, 55%) indicated primary concerns with data inaccuracies from this stream due to carrying phones in pockets or bags. For screen, adolescents and parents endorsed some concerns about data inaccuracies from this stream, though they did not indicate significant concern about having information from this stream shared with doctors.

For Keyboard, the majority of adolescents (65/103, 63.1%) reported primary concern with data from this stream being shared with their parents. Similarly, a majority of parents (61/99, 62%) endorsed primary concerns about their children sharing this information with them. A majority of adolescents (71/103, 68.9%) also indicated they would not be comfortable with their doctor potentially knowing what they type. Of note, only 19% (13/68) of adolescents and 21% (10/48) of parents endorsed agreement with having typing speed analyzed and keyboard data deleted daily. For microphone, a majority of adolescents (80/103, 77.6%) and parents (69/99, 70%) endorsed primary concerns with their phones picking up everything said in close proximity. A majority of adolescents (70/103, 68%) and parents (57/99, 58%) indicated that it would be difficult for this stream to pick up what is said due to phones being in pockets or bags. In addition, a majority of adolescents (65/103, 63.1%) and most parents (47/99, 48%) also endorsed concerns about sharing this information with their doctor. Only 18% (15/82) of adolescents endorsed agreement with having their voice characteristics analyzed and microphone data deleted daily, whereas 36% (24/66) of parents endorsed agreement with this. Finally, for GPS, adolescents and parents endorsed greater concern about this information being shared with people other than doctors than about the accuracy of the data collected from this stream. Of note, only 11% (5/46) of adolescents and 20% (8/41) of parents endorsed agreement with having GPS data analyzed for places of interest and deleted daily, though 30% (14/46) of adolescents and 29% (12/41) of parents endorsed agreement with having transformed (ie, nonexact) GPS data analyzed. See Figures S4-S12 in Multimedia Appendix 2 for the percentages of adolescents and parents who endorsed specific concerns for each stream.

Thematic Analysis

Participants provided a total of 157 (79 adolescent responses and 78 parent responses) open-ended responses about additional concerns they had for each passive data stream. Across all streams, 5 overarching themes emerged (ie, privacy concerns, data concerns, absence of personal voice, iatrogenic effects, and other). Privacy concerns referred primarily to participant concerns about the confidentiality of their personal information; data concerns centered around concerns about the accuracy and validity of collected data. Absence of personal voice referred to concerns about data collection dismissing an individual's thoughts and feelings. Iatrogenic effects captured concerns about unintended negative effects of data collection on patients and their families, and others referred to nonspecific concerns about digital phenotyping methodology in general.

Some responses were lengthy and contained multiple elements, each representing a specific theme. Figure 2 presents the types of adolescent and parent concerns reported for each theme by the passive data stream. The most common themes of concern that emerged across all streams were privacy concerns and data concerns. Passive data streams with the greatest number of total concerns were apps, Bluetooth, call and SMS text message logs, keyboard, and microphone. Table 4 presents sample quoted responses and quoted response excerpts that were representative of each theme. All personal identifiers were removed from the study data, though presented quoted data are annotated with "adolescent" or "parent" and the specific stream referenced in each open response.

XSL•FO
RenderX

Figure 2. Adolescents' and parents' open-response additional concerns by theme.

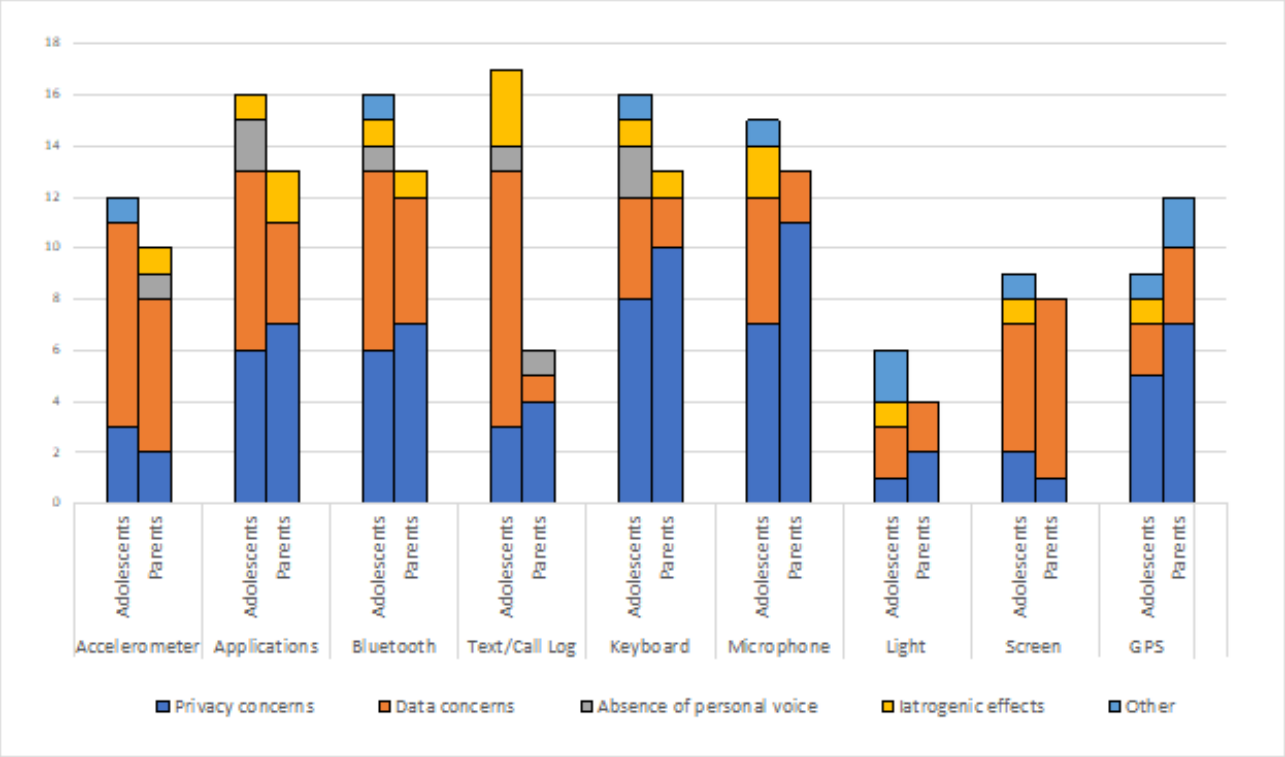

XSL•FO
RenderX

Table 4. Quoted samples and excerpts of adolescent and parent open-response additional concerns by theme.

Theme	Adolescent	Parent
Privacy concerns	• "The benefits of the information for doctor/researchers does not outweigh the cost of the extreme invasion of privacy." [female adolescent] • "If a physician said to me 'show me your call logs and activity on your phone for all your apps' I would think they were untrustworthy and/or up to something. It's so invasive." [female adolescent]	• "Again, for me as a parent, it is the privacy issue of tracking so much of a person's life." [mother] • "I am uncomfortable with any app that tracks every single thing that my child does. Unfortunately, scary world and I wouldn't give permission for any of these apps to be knowingly used." [mother]
Data concerns	• "I am also concerned that my activity is not always a reflection of how I am feeling and functioning." [female adolescent] • "It might give my doctor the wrong idea. For example, I push through a lot of my pain, so just because I am somewhere doesn't mean I am symptom-free and having a great time. I could be out at restaurants or other venues but spend a lot of time sick in the bathroom. I could be at school but down with the nurse." [female adolescent]	• "My child may leave phone on at night which a music or TV app running to help sleep, but they aren't actually awake while the phone is left on. It wouldn't be an accurate measure." [mother] • "Accuracy. I wouldn't trust the data fully. I wouldn't want any real treatment plan nor medication prescribed to my any of my kids from using data on an app." [mother]
Absence of personal voice	• "This takes a certain level of trust away from the patient [and] I feel like that level of trust being taken away makes it seem like the patient is unreliable." [female adolescent] • "How about Drs/patient just having a conversation!" [male adolescent]	• "I'm not sure how comfortable I am with allowing that much data access when physicians and teams could work with patients to discuss these same issues and usages vs tracking personal data." [mother] • "Would there be somewhere for my child to also record how they were feeling at the time of the activity?" [mother]
Iatrogenic effects	• "This would simply teach people to not share when they are feeling upset because they know it is being listened to. This would have a negative impact on the patients over time." [female adolescent] • "I know I would have anxiety over not hanging around a lot of people and then panic over my mental health team judging me or panicky about the fact of how my data is to score how awful my social life is (this is very depressing). Despite the fact I am honest about now having friends with my care team." [female adolescent]	• "My child would consider this a complete invasion of privacy, and it would cause lots of arguments between us." [mother] • "Adding another app like this one would cause huge problems in our household." [mother] • "I'm afraid my child would strongly oppose having applications usage tracked and shared. I think it would alter her usage if she felt ashamed about being on social media too much." [mother]
Other	• "Sensors like these could cause your phones battery to lose power significantly more than it would when your phone is not checking and sending that much data." [female adolescent]	• "This feature will likely be very draining on the phone battery." [mother]

Discussion

This study investigated adolescent and parent perspectives on endorsement, perceived utility, and concerns related to passive data stream collection through smartphones for digital phenotyping for clinical and research purposes in youths with chronic pain. Consistent with our hypotheses, results indicated that adolescents and parents generally endorse the use of digital phenotyping and that perceived utility predicts this endorsement. Concerns do exist, however, about privacy and data accuracy. Below, we discuss these results further and suggest their potential implications for clinical practice and future research for a pediatric population with chronic pain.

On the whole, adolescents and parents reported a moderate level of endorsement for their data collected by passive streams being used for clinical and research purposes. Our data suggest a notable level of parental comfort with a digital phenotyping approach to their child's care using all 9 passive data streams for clinical or research purposes. For adolescents, however, the results were more nuanced and revealed greater stream-specific variability. That is, some streams garnered near universal endorsement from adolescents (ie, accelerometer, screen, and light), while others earned endorsement from only about a third of adolescents (ie, keyboard and microphone). In other words, just because data can be collected through certain smartphone passive data streams does not indicate that adolescents with chronic pain endorse such data collection.

Understanding the types [28] of smartphone-collected data that adolescents with chronic pain are most comfortable with is integral to leveraging digital phenotyping methodology for this population. For example, social activity may be gleaned from Bluetooth, SMS text message and call logs, keyboard, microphone, and GPS data to indicate who someone is with, the content and tenor of their interactions, and even where they are for these interactions [29]. Similarly, apps, keyboard, and screen can be used to assess an individual's mood and psychological functioning [30] by providing data about how someone feels, what they are spending their time doing, and

XSL•FO
RenderX

what they are typing. Further, a different cluster of accelerometer, light, and screen passive data streams may best capture an individual's sleep through data that indicates when a person is moving or still, in a dark or light room, and on or off their phone [31]. Results from this study showed that passive data streams targeting adolescents' social activity, mood, and psychological functioning are perceived as much less acceptable than those assessing sleep or exercise. Importantly, this differential perceived acceptability may suggest that greater education or demonstration of benefits for certain clusters of passive data streams may be warranted to gain increased endorsement for use. Similarly, lower reports of perceived acceptability may indicate that better methods of data security and privacy are necessary before such streams are endorsed for use. As such, passive data streams that received lower perceived acceptability in this study need not be blanketly dismissed from future avenues of work, and specific utilities of certain passive data streams should be emphasized for identified populations of interest while still balancing the influence that survey wording may have on risk perception.

These results have critical implications for how digital phenotyping can be used in the care of this patient population. Specifically, for adolescents with chronic pain, a population with marked sleep disturbance [32,33], the high likelihood of endorsement of sleep-related passive data streams (ie, accelerometer, screen, and light) may position this cluster as a particularly promising avenue for accessing otherwise difficult-to-report data. In addition, given the interrelation of functional disability and prognosis [34], tracking participant's activity levels through an accelerometer may be particularly important for understanding the functioning of patients with chronic pain as compared to their healthy counterparts. The use of wearable devices may also hold critical promise, as may other ways to passively collect such data [35]. By contrast, the reticence of youths in our sample to share aspects of their social activity through smartphone passive data streams may indicate that Bluetooth, SMS text message and call logs, keyboard, microphone, and GPS are a less fruitful cluster of streams to explore. Similarly, the adolescents in this study reported a notably lower likelihood of endorsement for streams targeting their mood and psychological functioning. For youths with chronic pain, anxiety and low mood can be correlates, predictors, and consequences of pain [36-39]. As such, we note that other, more traditional methods of querying this important psychological information may be more preferable than passive data stream collection for our sample of youths with chronic pain. In a similar vein, we also emphasize that our findings are most relevant for our specific population (eg, pediatric pain), and therefore we underscore the importance of identifying uses of specific passive data streams for specific populations of interest (eg, app or keyboard data for depression or suicidality prediction).

For adolescents and parents alike, the quantitative and qualitative findings from this study also suggest 2 major factors contribute to participants' endorsement of data collection: perceived utility and privacy concerns. Specifically, a certain level of utility inherent to the data must be demonstrated, and privacy concerns must be minimized. Across all streams, participants' perceived utility predicted their endorsement, and this was true for adolescents, parents, and adolescent-parent dyads. For example, accelerometer, screen, and light were among the streams with the highest levels of endorsement and were also among the streams deemed most useful. This trend is also reflected in our investigation of stream acceptability by its utility. Indeed, so long as streams were deemed at least somewhat useful, they typically received majority endorsement. Privacy concerns, however, also emerged as a critical factor for participants. The 2 streams that did not gain majority endorsement when deemed somewhat useful were keyboard and microphone, the 2 streams that received the greatest number of privacy concerns. These results suggest that even if participants can appreciate the utility of passive stream–collected data for digital phenotyping, privacy concerns must be adequately addressed by clinicians and researchers in order to appropriately implement this technology. Importantly, adolescents also varied in the extent to which they were willing to agree to have their data shared with their parents, which ranged depending on the cluster of passive data streams in question. Consistent with the number of reported privacy concerns, the fewest adolescents provided endorsement for social activity streams (eg, keyboard and microphone), whereas the most adolescents provided endorsement for activity level streams being shared with their parents. Taken together, these findings highlight adolescent privacy concerns in familial, clinical, and research contexts, which may be important considerations for adolescent agreement to participate in other data-sharing opportunities, such as data donation.

In addition to privacy concerns, participants also voiced significant data-specific concerns related to both the accuracy of the data as well as the data's validity and meaning. In other words, participants want their data to correctly reflect how they are feeling. Adolescents and parents alike conveyed these concerns across all streams, from misgivings that their data would not correctly capture the targeted behavior (eg, if an individual plays a sport but does not carry their phone) to apprehension that their data would actually reveal their true experience (eg, if an individual is active but still in significant pain). Relatedly, participants also voiced a desire for an opportunity to contextualize their data with their experience. For youths with chronic pain and their parents, this desire to explain and "be heard" may be particularly important given their long medical journeys [40], experience of diagnostic uncertainty [41], and belief that providers are misunderstanding their pain [42]. As such, we suggest these considerations be at the forefront of clinical and research use for youths with chronic pain so that the technology of digital phenotyping can be leveraged while still balancing and honoring participant preference to contextualize their data with their own perceived experience. It will be important for clinicians to collaboratively review with their patients the data collected using mobile health tools in order to understand the context of the information gathered and to make sure patients have the opportunity to comment on their data. We note that smartphones may again be leveraged toward this goal through the augmentation of passive data stream collection with smartphone-delivered self-report surveys. Future work may benefit from exploring the inclusion of participant voice alongside passive data collection in this way.

XSL•FO
RenderX

This study also has several methodological limitations that suggest avenues for future research. First, this study investigated adolescent and parent attitudes in a specific population of English-speaking, treatment-seeking youths with chronic pain at 1 hospital in the northeast. As is typical of research in this population, our sample was skewed heavily female for both adolescents [36] and parents [43]. While representative of the population with chronic pain seen in our hospital, our findings should be generalized with caution to other, more diverse populations of youths, and we suggest that future research seek to similarly explore adolescent and parent attitudes toward digital phenotyping in other patient populations or samples of interest, particularly with options for translated versions of survey questions. In addition, our study was limited in its survey methodology, such that participants could only convey their thoughts and feelings through web-based survey responses, mostly in the form of multiple-choice response formats. Although we provided some options for free responses, future work may best capture the nuance and detail of youths' perspectives through interview-based data collection or focus groups, particularly those that probe the specific dyadic influence on adolescents and parent perspectives. Similarly, although our survey materials aimed to provide objective, unbiased information regarding each data stream and questions that might arise from their use, we acknowledge that we could not exhaustively anticipate all potential questions and concerns from participants. As such, we recommend future work implement more open-ended and interview-style interaction between researchers and participants to minimize the risk of potentially "leading" information or questions. Finally, although our study provides important information about digital phenotyping perspectives, our results do not indicate that these perspectives would remain constant when participants actually have their data collected through smartphone passive data streams. Future studies should begin to pilot and seek feedback during and after actual use of digital phenotyping methods to best understand participant endorsement as well as any concerns that unfold through engagement with these methods. Relatedly, such future studies must also consider participant preferences in the context of ethical implementation of this methodology (eg, consenting and sharing results with participants), which was beyond the scope of this survey. Importantly, future studies must also aim to clarify the validity of such data collection alongside previously validated self-report measures (eg, Pediatric Patient Reported Outcomes Measurement Information System [44] and the Functional Disability Inventory [45,46]).

In conclusion, despite these limitations, the results of this study provide valuable adolescent and parent insights about endorsement, perceived utility, and concerns related to digital phenotyping methodology for clinical and research purposes in youths with chronic pain. Although our findings highlight the general endorsement of this methodology across different passive data streams, the results also reveal the importance of maximizing participant perceived utility, minimizing privacy concerns, and finding ways to incorporate participant contextualization of their own data. Ultimately, these findings can help refine digital phenotyping methodology toward an acceptable, feasible, and ethical implementation of real-time symptom monitoring for assessment and intervention in youths with chronic pain.

Acknowledgments

We would like to thank Andreas Baumer for his assistance in the early stages of this project. JK discloses support for the research in this work from National Institutes of Health (grant K01DA057374) and the Boston Children's Hospital's Anesthesia Distinguished Trailblazer award.

Data Availability

Deidentified data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BAN conducted the data analyses, drafted the initial manuscript, and revised and reviewed all drafts of the manuscript. JC assisted in creating the materials, carried out the data acquisition, assisted with the data analyses, and reviewed and revised the manuscript. CK created the study material, assisted in the data acquisition, and reviewed and revised the manuscript. ERW assisted in creating the materials, reviewed and provided input on the data analysis, and reviewed and revised the manuscript. BLR assisted in conceptualizing the study, collecting data, and reviewing and revising the manuscript. KU assisted in conceptualizing the study and reviewed and revised the manuscript. JK conceptualized and designed the study, coordinated and supervised data collection and analyses, and reviewed and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete REDCap Survey. REDCap: Research Electronic Data Capture.
[PDF File (Adobe PDF File), 140 KB - jmir_v26i1e47781_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Supplemental tables and figures.
[DOCX File , 2170 KB - jmir_v26i1e47781_app2.docx]

References

1. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
2. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
3. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
4. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect Psychol Sci 2016;11(6):838-854 [FREE Full text] [doi: 10.1177/1745691616650285] [Medline: 27899727]
5. Anderson M, Jiang J. Teens, social media and technology 2018. Pew Research Center. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2023-12-19]
6. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychol Assess 2009;21(3):425-436 [FREE Full text] [doi: 10.1037/a0016869] [Medline: 19719353]
7. Matthews KA, Patel SR, Pantesco EJ, Buysse DJ, Kamarck TW, Lee L, et al. Similarities and differences in estimates of sleep duration by polysomnography, actigraphy, diary, and self-reported habitual sleep in a community sample. Sleep Health 2018;4(1):96-103 [FREE Full text] [doi: 10.1016/j.sleh.2017.10.011] [Medline: 29332687]
8. Sallis JF, Saelens BE. Assessment of physical activity by self-report: status, limitations, and future directions. Res Q Exerc Sport 2000;71(2 Suppl):S1-S14. [Medline: 10925819]
9. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
10. Low CA, Dey AK, Ferreira D, Kamarck T, Sun W, Bae S, et al. Estimation of symptom severity during chemotherapy from passively sensed data: exploratory study. J Med Internet Res 2017;19(12):e420 [FREE Full text] [doi: 10.2196/jmir.9046] [Medline: 29258977]
11. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020;22(5):e16875 [FREE Full text] [doi: 10.2196/16875] [Medline: 32348284]
12. Jacobson NC, Weingarden H, Wilhelm S. Using digital phenotyping to accurately detect depression severity. J Nerv Ment Dis 2019;207(10):893-896. [doi: 10.1097/NMD.0000000000001042] [Medline: 31596769]
13. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology 2018;43(8):1660-1666 [FREE Full text] [doi: 10.1038/s41386-018-0030-z] [Medline: 29511333]
14. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
15. Cote DJ, Barnett I, Onnela J, Smith TR. Digital phenotyping in patients with spine disease: a novel approach to quantifying mobility and quality of life. World Neurosurg 2019;126:e241-e249 [FREE Full text] [doi: 10.1016/j.wneu.2019.01.297] [Medline: 30797933]
16. Jacobson NC, O'Cleirigh C. Objective digital phenotypes of worry severity, pain severity and pain chronicity in persons living with HIV. Br J Psychiatry 2021;218(3):165-167 [FREE Full text] [doi: 10.1192/bjp.2019.168] [Medline: 31298167]
17. MacLeod L, Suruliraj B, Gall D, Bessenyei K, Hamm S, Romkey I, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR Mhealth Uhealth 2021;9(10):e20638 [FREE Full text] [doi: 10.2196/20638] [Medline: 34698650]
18. Gorin AA, Stone AA. Recall biases and cognitive errors in retrospective self-reports: a call for momentary assessments. In: Baum A, Revenson TA, Singer JE, editors. Handbook of Health Psychology. Mahwah, N.J: Lawrence Erlbaum Associates; 2001:405-413.
19. De Los Reyes A, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. Psychol Bull 2005;131(4):483-509. [doi: 10.1037/0033-2909.131.4.483] [Medline: 16060799]
20. Nisenson M, Lin V, Gansner M. Digital phenotyping in child and adolescent psychiatry: a perspective. Harv Rev Psychiatry 2021;29(6):401-408. [doi: 10.1097/HRP.0000000000000310] [Medline: 34313626]

XSL•FO
RenderX

21. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
22. Di Matteo D, Fine A, Fotinos K, Rose J, Katzman M. Patient willingness to consent to mobile phone data collection for mental health apps: structured questionnaire. JMIR Ment Health 2018;5(3):e56 [FREE Full text] [doi: 10.2196/mental.9539] [Medline: 30158102]
23. Rooksby J, Morrison A, Murray-Rust D. Student perspectives on digital phenotyping: the acceptability of using smartphone data to assess mental health. New York, NY, US: Association for Computing Machinery; 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-14. [doi: 10.1145/3290605.3300655]
24. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
25. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
26. Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. Biometrics 1980;36(1):19-25. [Medline: 7370372]
27. Blease C, Locher C, Leon-Carlyle M, Doraiswamy M. Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. Digit Health 2020;6:2055207620968355 [FREE Full text] [doi: 10.1177/2055207620968355] [Medline: 33194219]
28. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health 2020;23(4):161-166 [FREE Full text] [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]
29. Gong J, Huang Y, Chow PI, Fua K, Gerber MS, Teachman BA, et al. Understanding behavioral dynamics of social anxiety among college students through smartphone sensors. Inf Fusion 2019;49:57-68 [FREE Full text] [doi: 10.1016/j.inffus.2018.09.002]
30. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: using smartphones to assess mental health and academic performance of college students. In: Rehg JM, Murphy SA, Kumar S, editors. Mobile Health: Sensors, Analytic Methods, and Applications. Cham, Switzerland: Springer International Publishing; 2017:7-33.
31. Abdullah S, Matthews M, Murnane EL, Gay G, Choudhury T. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. New York, NY, US: Association for Computing Machinery; 2014 Presented at: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 13-17, 2014; Seattle, Washington p. 673-684. [doi: 10.1145/2632048.2632100]
32. Badawy SM, Law EF, Palermo TM. The interrelationship between sleep and chronic pain in adolescents. Curr Opin Physiol 2019;11:25-28 [FREE Full text] [doi: 10.1016/j.cophys.2019.04.012] [Medline: 31815208]
33. Milder MKL, McQuillan ME, Honaker SM, Miller MM, Tolley J, Michel MA, et al. Concurrent and longitudinal associations among insomnia symptoms, internalizing problems, and chronic pain in youth. Health Psychol 2023;42(1):15-23. [doi: 10.1037/hea0001244] [Medline: 36227312]
34. McGrath PJ, Walco GA, Turk DC, Dworkin RH, Brown MT, Davidson K, et al. Core outcome domains and measures for pediatric acute and chronic/recurrent pain clinical trials: PedIMMPACT recommendations. J Pain 2008;9(9):771-783 [FREE Full text] [doi: 10.1016/j.jpain.2008.04.007] [Medline: 18562251]
35. Lee-Archer PF, Kossowsky J, Cravero JP. If the watch fits-wear it. Paediatr Anaesth 2020;30(5):518-519. [doi: 10.1111/pan.13835] [Medline: 32500966]
36. King S, Chambers CT, Huguet A, MacNevin RC, McGrath PJ, Parker L, et al. The epidemiology of chronic pain in children and adolescents revisited: a systematic review. Pain 2011;152(12):2729-2738. [doi: 10.1016/j.pain.2011.07.016] [Medline: 22078064]
37. Noel M, Groenewald CB, Beals-Erickson SE, Gebert JT, Palermo TM. Chronic pain in adolescence and internalizing mental health disorders: a nationally representative study. Pain 2016;157(6):1333-1338 [FREE Full text] [doi: 10.1097/j.pain.0000000000000522] [Medline: 26901806]
38. Holley AL, Wilson AC, Palermo TM. Predictors of the transition from acute to persistent musculoskeletal pain in children and adolescents: a prospective study. Pain 2017;158(5):794-801 [FREE Full text] [doi: 10.1097/j.pain.0000000000000817] [Medline: 28151835]
39. Kroenke K, Wu J, Bair MJ, Krebs EE, Damush TM, Tu W. Reciprocal relationship between pain and depression: a 12-month longitudinal analysis in primary care. J Pain 2011;12(9):964-973 [FREE Full text] [doi: 10.1016/j.jpain.2011.03.003] [Medline: 21680251]
40. Jordan AL, Eccleston C, Osborn M. Being a parent of the adolescent with complex chronic pain: an interpretative phenomenological analysis. Eur J Pain 2007;11(1):49-56. [doi: 10.1016/j.ejpain.2005.12.012] [Medline: 16458550]
41. Pincus T, Noel M, Jordan A, Serbic D. Perceived diagnostic uncertainty in pediatric chronic pain. Pain 2018;159(7):1198-1201. [doi: 10.1097/j.pain.0000000000001180] [Medline: 29419651]

XSL•FO
RenderX

42. Meldrum ML, Tsao JC, Zeltzer LK. "I can't be what I want to be": children's narratives of chronic pain experiences and treatment outcomes. Pain Med 2009;10(6):1018-1034 [FREE Full text] [doi: 10.1111/j.1526-4637.2009.00650.x] [Medline: 19594848]
43. Phares V, Lopez E, Fields S, Kamboukos D, Duhig AM. Are fathers involved in pediatric psychology research and treatment? J Pediatr Psychol 2005;30(8):631-643 [FREE Full text] [doi: 10.1093/jpepsy/jsi050] [Medline: 15772363]
44. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):S3-S11 [FREE Full text] [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
45. Walker LS, Greene JW. The functional disability inventory: measuring a neglected dimension of child health status. J Pediatr Psychol 1991;16(1):39-58. [doi: 10.1093/jpepsy/16.1.39] [Medline: 1826329]
46. Kashikar-Zuck S, Flowers SR, Claar RL, Guite JW, Logan DE, Lynch-Jordan AM, et al. Clinical utility and validity of the Functional Disability Inventory among a multicenter sample of youth with chronic pain. Pain 2011;152(7):1600-1607 [FREE Full text] [doi: 10.1016/j.pain.2011.02.050] [Medline: 21458162]

Abbreviations

BCH: Boston Children's Hospital
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
OR: odds ratio
REDCap: Research Electronic Data Capture

Edited by T de Azevedo Cardoso; submitted 05.04.23; peer-reviewed by K Allen, E Onaku, M Görges; comments to author 07.09.23; revised version received 28.09.23; accepted 29.11.23; published 11.01.24.

Please cite as:
Nestor BA, Chimoff J, Koike C, Weitzman ER, Riley BL, Uhl K, Kossowsky J
Adolescent and Parent Perspectives on Digital Phenotyping in Youths With Chronic Pain: Cross-Sectional Mixed Methods Survey Study
J Med Internet Res 2024;26:e47781
URL: https://www.jmir.org/2024/1/e47781
doi:10.2196/47781
PMID:38206665

©Bridget A Nestor, Justin Chimoff, Camila Koike, Elissa R Weitzman, Bobbie L Riley, Kristen Uhl, Joe Kossowsky. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Natural Language Processing of Referral Letters for Machine Learning–Based Triaging of Patients With Low Back Pain to the Most Appropriate Intervention: Retrospective Study

Sebastian Fudickar[1,2], PD, Dr habil; Carsten Bantel[3,4], Priv Doz, Dr med, PhD; Jannik Spieker[2], MSc; Heinrich Töpfer[2], BSc; Patrick Stegeman[5], MSc; Henrica R Schiphorst Preuper[5], MD, PhD; Michiel F Reneman[5], Prof Dr; André P Wolff[6], Prof Dr; Remko Soer[6,7], MSc, PhD

[1]Institute of Medical Informatics, University of Lübeck, Lübeck, Germany
[2]Department of Medicine and Public Health, University of Oldenburg, Oldenburg, Germany
[3]University Clinic for Anesthesiology, Intensive Care Medicine, Emergency Medicine and Pain Therapy, University of Oldenburg, Oldenburg, Germany
[4]University Oldenburg, Oldenburg, Germany
[5]Department of Rehabilitation Medicine, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
[6]Department of Anesthesiology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
[7]Research Group Smart Health, Saxion University of Applied Sciences, Enschede, Netherlands

Corresponding Author:
Sebastian Fudickar, PD, Dr habil
Institute of Medical Informatics
University of Lübeck
Ratzeburger Allee 160
Lübeck, D-23538
Germany
Phone: 49 160 7979077
Email: sebastian.fudickar@uni-luebeck.de

Abstract

Background: Decision support systems (DSSs) for suggesting optimal treatments for individual patients with low back pain (LBP) are currently insufficiently accurate for clinical application. Most of the input provided to train these systems is based on patient-reported outcome measures. However, with the appearance of electronic health records (EHRs), additional qualitative data on reasons for referrals and patients' goals become available for DSSs. Currently, no decision support tools cover a wide range of biopsychosocial factors, including referral letter information to help clinicians triage patients to the optimal LBP treatment.

Objective: The objective of this study was to investigate the added value of including qualitative data from EHRs and referral letters to the accuracy of a quantitative DSS for patients with LBP.

Methods: A retrospective study was conducted in a clinical cohort of Dutch patients with LBP. Patients filled out a baseline questionnaire about demographics, pain, disability, work status, quality of life, medication, psychosocial functioning, comorbidity, history, and duration of pain. Referral reasons and patient requests for help (patient goals) were extracted via natural language processing (NLP) and enriched in the data set. For decision support, these data were considered independent factors for triage to neurosurgery, anesthesiology, rehabilitation, or minimal intervention. Support vector machine, k-nearest neighbor, and multilayer perceptron models were trained for 2 conditions: with and without consideration of the referral letter content. The models' accuracies were evaluated via F1-scores, and confusion matrices were used to predict the treatment path (out of 4 paths) with and without additional referral parameters.

Results: Data from 1608 patients were evaluated. The evaluation indicated that 2 referral reasons from the referral letters (for anesthesiology and rehabilitation intervention) increased the F1-score accuracy by up to 19.5% for triaging. The confusion matrices confirmed the results.

Conclusions: This study indicates that data enriching by adding NLP-based extraction of the content of referral letters increases the model accuracy of DSSs in suggesting optimal treatments for individual patients with LBP. Overall model accuracies were considered low and insufficient for clinical application.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e46857) doi:10.2196/46857

KEYWORDS

decision support; triaging; NLP; natural language processing; neural network; LBP; low back pain; back; pain; decision-making; machine learning; artificial intelligence; clinical application; patient records; qualitative data; support system; questionnaire; quality of life; psychosocial

Introduction

Triaging patients with low back pain (LBP) for the best-fitting treatment is a complex interaction between evidence-based standard application, clinical reasoning, and patient preferences. Many factors interact within the biopsychosocial model for LBP, making it difficult to provide the best-possible treatment by the right professional based on the best evidence. In general, the pain of 95% of patients with LBP is considered nonspecific, meaning that a clear somatic cause for pain cannot be determined [1]. The diagnosis of LBP, therefore, is often insufficient to choose the best treatment modality. Within the broad spectrum of complaints and the lack of a clear-cut diagnosis, conservative treatments, such as interdisciplinary rehabilitation or other activating therapies, are considered to have the highest level of evidence [2]. Other treatment options, including anesthesiology or surgery, may be indicated for subgroups. However, the complex biopsychosocial interactions make identifying which patient may require which treatment difficult.

Attempts have been made to group patients with overlapping characteristics to individualize treatment [3]. Groups have been established based on psychosocial characteristics (among others, the Örebro Musculoskeletal Pain Questionnaires [4]), case complexity (eg, the Start Back Screening Tool [5]), or, for example, clinical prediction rules [6]. These subgrouping studies are a step forward in providing patients with targeted treatment based on their characteristics. Although the generalization of these models appears limited, they are not easily adaptable to other contexts. They are insufficiently capable of stratifying treatment based on a possibly infinite number of patient characteristics.

Health care professionals have been able to triage patients to surgical, anesthesiologic, or rehabilitation interventions with moderate agreement [7]. To progress in individualizing treatment and referrals, artificial intelligence systems, particularly machine learning (ML), are a relatively new method that can identify patient patterns based on more input than can be incorporated into traditional models. A few examples have appeared on LBP previously, showing that models could be applied to the data [7] and support the decision of who should get low back surgery [8]. A clinical decision support system (CDSS) may lead to an increase in triaging patients to the best-possible or targeted treatment and increase the cost-effectiveness of current care for LBP. These models, however, are limited to self-report questionnaires as input and may consequently still have low model fit values, leading to nonvaluable support tools [7].

Electronic health records (EHRs) potentially contain a huge source of qualitative data, including scans, referral information, and other free-text information data, which may increase the information quality as input for a CDSS. The data's value, however, for the triaging process of patients with LBP and how these data can be enriched are unknown.

Therefore, this study aimed to investigate the added predictive value of qualitative referral letter information for ML-based triaging of patients with LBP to either neurosurgery, anesthesiology, rehabilitation, or a minimal intervention based on information and education.

Methods

Recruitment

A retrospective study was conducted in a clinical cohort of patients with chronic LBP. Patients referred to the Groningen Spine Center, a university-based tertiary center, were invited to participate. To be eligible for participation, patients had to be at 18 years or older and need to report predominantly LBP, either with or without radiation to the legs. We decided not to exclude patients based on comorbidity, life expectancy, work, or health status to reflect daily care. The Groningen Spine Center is a multidisciplinary center comprising, among others, rehabilitation, neurosurgery, and anesthesiology departments. Prior to participation, patients received informative material on the purpose of a larger 10-year prospective cohort study (the Groningen Spine Cohort Study [9]) to study long-term outcomes with LBP. This study was part of the cohort study.

Ethical Considerations

Patients provided signed informed consent. The Medical Ethical Committee of the University Medical Center Groningen provided a waiver for this study because no other data other than care as usual were being used (M15.169472). The research followed the Declaration of Helsinki and the Good Clinical Practice.

Clinical Procedures and Measures

After referral by a general practitioner or a second-line specialist, the referral letters were automatically stored in the spine center's EHRs (Hyperspace version February 2022, Epic Systems). The retrieved referral letters were digitized as text. After enrollment, patients were sent an online set of questionnaires to screen for potential impact, pain intensity, disability, and psychosocial factors that have been described to be correlated with LBP. All questionnaires were validated on LBP. The quality of life was assessed using the EQ-5D [10] Pain Impact, and patient characteristics, medical history, and symptoms and functioning were gathered using the National Institutes of Health (NIH) minimal data set for LBP [11]. Work ability was measured using the Work Ability Score (WAS [12]), and the psychosocial work environment was measured using the short Copenhagen Psychosocial Questionnaire (COPSOQ

XSL•FO
RenderX

II [13]). Further details of the study characteristics are described in the study by Dutmer et al [9]. If patients were unable to fill out the questionnaires online, paper versions were sent by mail.

Prior to the first consultation, all patients were triaged by 1 of 4 physician assistants (PAs) specifically educated in spinal disorders. For this study, PAs' referrals to rehabilitation, anesthesiology, surgery, or others consisted of education and advice, no treatment, further diagnostics, or referral to primary care (see Table 1 for details of the corresponding categorization). These referrals were considered our primary outcome, whereas the baseline questionnaires and EHR data were our predicting features. In addition to these established questionnaires, we considered the textual content of the digitized referral letters. With this dependency on the availability of digitized referral letters, cases without referral letters were not considered and were excluded. In the EHR data, we searched for the referral reason as well as the patient's question for help, which was reported as the patient goal in this study.

Table 1. Considered referral categories and corresponding data entries (N=1209).

Referral category	Training set[a] samples (pre-SMOTE), n (%)	Evaluation set samples, n (%)
Rehabilitation	521 (43.1)	50 (4.1)
Anesthesiologic pain therapy	178 (14.7)	50 (4.1)
Neurosurgery	59 (4.9)	50 (4.1)
Other treatment[b]	251 (20.8)	50 (4.1)

[a]The training set was balanced afterward via the synthetic minority oversampling technique (SMOTE), resulting in each class holding 521 samples.
[b]"Other treatment" represents a minimal intervention based on information material.

Data Preprocessing

The following preprocessing steps were applied to the data. A data set including all questionnaires was enriched with additional referral letter parameters (see the following section). The selection of multiple referral reason or patient goal features/answers per letter was supported. However, if patients had multiple referral letters, only the referral reasons and patient goal of the initial referral letter were considered for processing. Missing data from the questionnaire were not replaced by imputation of any kind, whereas all data gathered from the EHRs were included in the analysis. Subsequently, this extended data set (ED) was randomly separated into an evaluation data set (with 50 samples per treatment category) and a training data set consisting of the remaining data set entries (see Table 1 for descriptive statistics per treatment category). Table 1 indicates that the given data set is unbalanced, a common challenge for training ML-based CDSSs. Using original samples for the evaluation data set ensured that our findings were meaningful, and the equal distribution of samples among treatment categories in this evaluation data set ensured fair classification results without considering the original distribution among treatment categories. With the evaluation data set being balanced, sample distribution among the considered treatment categories in the training set (for the original sample distribution among classes, see Table 1) was balanced via the synthetic minority oversampling technique (SMOTE) [14] (with k-nearest neighbors=5 and strategy=auto). SMOTE is a data augmentation approach that augments data entries for minority categories [15], thereby supporting multiclass resampling using a one-versus-rest scheme [14]. The balancing via SMOTE ensures that each treatment category is considered equally relevant.

All previous processing steps were performed in the ED. A basic data set (BD) was generated by duplicating the ED and removing additional EHR parameters to investigate the ED's relevance. Both data sets were z-score-normalized (separately for training, validation, and evaluation) via the *scikit-learn* library of Python.

Natural Language Processing

To quantify and embed the EHR data into the BD, natural language processing (NLP) was applied. NLP is a technique used to extract meaning from qualitative data, which can be used to structure large amounts of unstructured, qualitative data. It is well accepted that in referral letters, meaningful information is included, but automatic extraction can only be achieved via NLP. In our programmed processing pipeline, we conducted named entity recognition (NER) via specific words (see Tables 2 and 3) via the regular expression operations module *RE-Package* of Python. With RE-Package, regular words or expressions were identified, which are a sequence of characters that specify a search pattern for defined code words in between which additional characters might occur.

The American Standard Code for Information Interchange (ASCII)–encoded referral letter texts were analyzed by trained clinical personnel regarding the description of the following data: referral reason and patient goals

XSL•FO
RenderX

Table 2. Referral reason features, including search term examples, meaning, and statistics of occurrence (overall referral letters) in all detected referral reason findings. For a comprehensive list of code words, see Multimedia Appendix 1.

Feature	Meaning	Dutch code word example (English translation)	Occurrences, n (%)
SecondOpinionReRe[a]	Referral for a second opinion/diagnostics	*aanvullend onderzoek* (additional investigation)	383 (29)
AnesthesiologyReRe	Referral for anesthesiology	*blockade* (nerve block)	180 (14)
RehabReRe	Referral for interdisciplinary rehabilitation	*revalidatie* (rehabilitation)	442 (34)
AdviceReRe	Could you give advice to the patient?	*advies* (advice)	141 (11)
OptionsReRe	Are there treatment options?	*behandelbare opties* (options for treatment)	172 (13)

[a]ReRe: referral reason.

Table 3. Patient goal features, including search term examples, meaning, and statistics of occurrence among all recognized patient questions for help finding. For a comprehensive list of code words, see Multimedia Appendix 2.

Feature	Meaning	Dutch code word example (English translation)	Occurrences, n (%)
MoreCausePaG[a]	More diagnostics	*2e mening* (2nd opinion)	128 (22)
PainRedPaG	Pain reduction	*pijn* (pain)	241 (42)
BetterFuncPaG	Better functioning	*behandelopties* (treatment options)	188 (32)
AdvicePaG	Advice	*advies* (advice)	23 (4)

[a]PaG: patient goal.

RE-Package was used to find free-text words related to the referral reasons and patient goals, which were stated in Dutch. The corresponding code words are summarized in Multimedia Appendices 1 and 2. Corresponding information was extracted via an NLP processing pipeline. In a preprocessing step of the processing pipeline, ambiguous wordings were unified. To create this pipeline, referral letters were manually studied by 2 authors (RS and PS) independently to determine relevant types of referral reasons and patient goals and corresponding code words indicating them. All words were classified. The resulting categories of referral reasons and patient goals are summarized in Tables 2 and 3, respectively, with their number of occurrences.

The initial textual data were transferred to JavaScript Object Notation (JSON) data interchange format via the *Panda* library [16]. Filtering of referral letters was implemented via RE-Package [17], resulting in lists of referral letters that contained the corresponding code words. Redundant framings were unified via regular expressions. Subsequently, relevant text segments, especially the ones covering the referral reasons and patient goals, were identified via regular expressions in the referral letters' extracted text. Within these text segments, the corresponding code words from the EHRs (see Tables 2 and 3) were screened via regular expressions. All referral reason features were extracted from the referral reason text; if multiple referral reason entries were available per patient, all were considered. Similarly, all patient goal features were extracted from the patient questions for help text; if multiple patient goal entries per patient were identified, all were considered (eg, "I want a solution for my pain, and I want to know what is wrong with my back."). Correspondingly, 4 of the 5 new referral reason features and 4 new patient goal features were included in the data set. Henceforth, we refer to all parameters extracted from the referral letters (specifically the referral reasons and patient goals) as "referral letter parameters."

To ensure the interpretability of the ML models and of the effect of the additional referral parameters, we excluded any feature engineering but kept the input parameters as they were. By not constructing features by recombining the input parameters, we investigated the relevance and benefit of each feature and ensured high explainability of the network.

Feature Selection

To identify the relevance of the ED compared to the BD and to limit the number of features so that they could be gathered in a questionnaire in standard clinical use, we applied the ReliefF feature selection algorithm [18] to both the ED and the BD. For feature selection, only the entries in the training sets were considered before applying SMOTE.

The general aim of feature selection was to identify a subset of features by which the data space spanned as much as possible (including as much variety as possible), while data points in the same class (an alternative naming of the treatment category) remained as small as possible. The ReliefF algorithm [18] is a supervised feature selection method for multiclass problems that is robust against incomplete and noisy data, considering the k-nearest neighbors (kNNs) [19]. It is a multivariate filter that completely ranks individual features according to their relevance for class separation in the context of other features throughout the observations in the training set [20]. ReliefF calculates a feature score based on the differences among feature (parameter) values between neighboring instances. By considering the differences in features among neighboring samples of class hits (NH) that belong to the same class and neighboring samples of class misses (NM), the algorithm can calculate the relevance each feature contributes to determining

XSL•FO
RenderX

class membership. With the original Relief function supporting only biclass problems, ReliefF supports multiclass search and considers k-near misses per class.

The feature relevance score (S) for a particular feature was consequently calculated as follows, with x representing the feature value, abs() denoting the absolute value, and n representing the total number of instances in the data set:

$$S = \text{sum}(\text{abs}(x - NH) - \text{abs}(x - NM))/n$$

Correspondingly, each feature's weight update (W) was calculated as follows, with M representing the total number of features:

$$W = W - (\text{abs}(x - NH) - \text{abs}(x - NM))/(n \times M)$$

These equations were used iteratively, and the feature scores were averaged over all instances.

By applying ReliefF, the input parameters were ranked in accordance with their relevance for class separation. The ReliefF algorithm was configured for m=n (to all numbers of training samples) [21], the number of kNNs=50 [19], and the feature relevance threshold τ=0.7 [22] for selecting features as common parameter settings.

Training and Machine Learning

To analyze the benefit of the referral letter parameters regarding a treatment path decision, ML models were trained for both data sets, the BD and the ED, after feature selection. For achieving good results in small, unbalanced data sets, such as the given ones, and allowing for interpretability of the results, the kNN algorithm, support vector machine (SVM), and multilayer perceptron (MLP) models were used in the evaluation. Being one of the first studies to apply both quantitative and qualitative (NLP-extracted) data, we decided to focus on applying well-established neural network architectures, as they are well understood and especially allow interpretability of the classification processes, which is essential when investigating new data types. We contacted the clinicians, and they confirmed the high criticality of having interpretability of the neural networks, which should lead to applicable or treatable options for the patient.

Another advantage is that these classical ML approaches are much better suited than deep neural networks for small, unbalanced data sets. In recent works, we found that in smaller data sets, the classical approaches perform similarly well as deep learning networks.

For training of the models, 4-fold cross-validation, the rectified linear unit (ReLU) activation function, the Adam optimizer, and the cross-entropy loss function via scikit-learn were used. Further implementation details can be found in the code available in Ref. [23]. For training and evaluation, an AMD Ryzen 9 5900X with 64GB RAM and NVIDIA GeForce RTX 3070 was used running 64-bit Windows 11 Pro, Python 3.8, Keras 2.4.3, Keras-Preprocessing 1.1.2, Keras-tuner 1.0.1, pandas 1.1.1, scikit-learn 0.23.2, skrebate 0.62, and TensorFlow 2.4.1. A complete overview of all the steps is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Process to reach the final ML algorithm. BD: basic data set; ED: extended data set; kNN: k-nearest neighbor; ML: machine learning; MLP: multilayer perceptron; ReLU: rectified linear unit; SMOTE: synthetic minority oversampling technique; SVM: support vector machine.

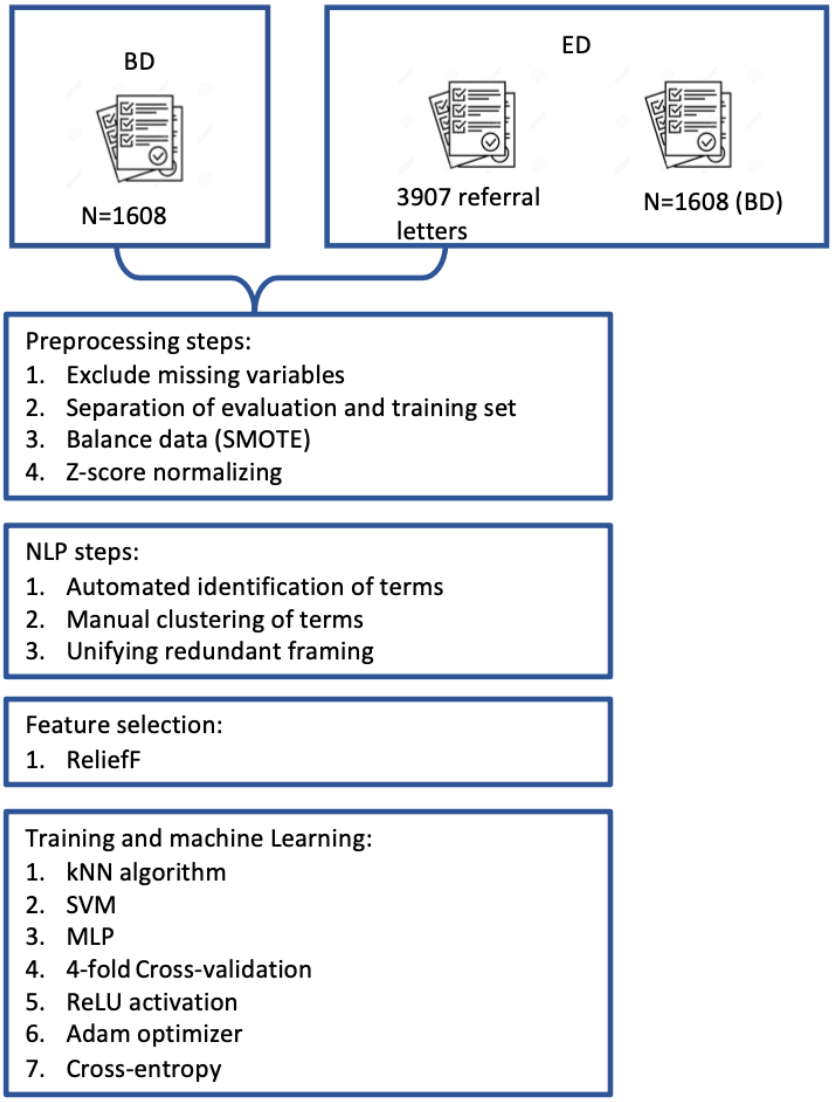

Statistical Analysis

The data set was characterized by sociodemographic and health-related data and distribution characteristics of the referral reasons and patient goals on the treatment categories. The relevance of the referral letter parameters was investigated based on the ReliefF ranking for both data sets (ED and BD). The accuracy of the ML models was evaluated via the F_1-score and confusion matrices.

The F_1-score is a commonly used metric to evaluate the performance of a classification model via precision (P) and recall (R). Precision measures the ratio of correctly predicted true-positive (TP) instances to the total predicted TP instances and false-positive (FP) instances:

$$P = TP/(TP + FP)$$

Recall measures the ratio of correctly predicted TP instances to the total actual TP instances, thus including the number of incorrectly predicted false-negative (FN) instances:

$$R = TP/(TP + FN)$$

In the following equation, the F_1-score represents the harmonic mean of precision and recall and thus represents a balanced evaluation of the model's performance:

$$F_1\text{-score} = 2 \times (P \times R)/(P + R)$$

The F_1-scores were averaged over 4 separate training runs.

Confusion matrices were plotted for the BD and ED classifiers. A confusion matrix summarizes prediction results on a classification problem in a visual manner. By analyzing confusion matrices, we can investigate where the classifier misperforms regularly, indicating lower feature variability.

Results

Study Cohort

We originally included 1608 patients and corresponding 3907 letters in the data set. The age of the patients ranged from 18 to 66 years, with a mean age of 46 (SD 12.93) years, and 914 (57.2%) patients were female and on average 45 (SD 13.3) years younger than men, who had a mean age of 48 (SD 12.4) years. Of the 1608 patients, 1341 (83.4%) had referral letters included.

XSL•FO
RenderX

Of these, 1209 (90.2%) patients were included because their referral letters included referral reasons (n=1062, 87.8%, cases) or patient goals (n=746, 61.7%); 160 referral letters contained only patient goals. Multiple referral letters were available and included for 1012 (83.7%) patients, with redundant referral letters for 1006 (75%) patients with referral letters. Once balanced, the training set included 521 entries per treatment category.

The training data set included 1009 patients, who were distributed among the treatment categories, as shown in Table 1. Within the training and evaluation sets, Tables 4 and 5 present the category-specific extracted referral reasons and corresponding patient goals, respectively.

Table 4. Category-specific statistics of the available referral reasons specified within the referral letters, considered in the training and evaluation data sets. Sample distribution among categories and the corresponding sums and overall percentage distribution are shown. "Referral reason" categories refer to Table 2.

Referral reason category	Treatment: rehabilitation (1), n (%)		Treatment: anesthesiology (2), n (%)		Treatment: neurosurgery (3), n (%)		Treatment: no treatment (0), n (%)	
	Training	Evaluation	Training	Evaluation	Training	Evaluation	Training	Evaluation
SecondOpinionReRe[a] (1)	167 (13)	17 (1)	56 (4)	24 (2)	33 (3)	21 (2)	76 (6)	15 (1)
AnesthesiologyReRe (2)	22 (2)	3 (0)	80 (6)	34 (3)	6 (0)	9 (1)	6 (0)	3 (0)
RehabReRe (3)	302 (23)	28 (2)	30 (2)	10 (1)	8 (1)	5 (0)	47 (4)	7 (1)
AdviceReRe (4)	71 (5)	7 (1)	20 (2)	5 (0)	7 (1)	4 (0)	24 (2)	3 (0)
OptionsReRe (5)	85 (6)	9 (1)	36 (3)	5 (0)	10 (1)	6 (0)	16 (1)	5 (0)
Sum	647 (48.9)	64 (4.8)	222 (16.8)	78 (5.9)	64 (4.8)	45 (3.4)	169 (12.8)	33 (2.5)

[a]ReRe: referral reason.

Table 5. Category-specific statistics of the available patient goal specified within the referral letters, considered in the training and evaluation data sets. Sample distribution among categories and the corresponding sums and overall percentual distribution are shown. "Patient goal" categories refer to Table 3.

Patient goal category	Treatment: rehabilitation (1), n (%)		Treatment: anesthesiology (2), n (%)		Treatment: neurosurgery (3), n (%)		Treatment: no treatment (0), n (%)	
	Training	Evaluation	Training	Evaluation	Training	Evaluation	Training	Evaluation
MoreCausePAG[a] (1)	78 (14)	5 (1)	9 (2)	1 (0)	2 (0)	2 (0)	30 (5)	1 (0)
PainRedPAG (2)	126 (22)	9 (2)	16 (3)	5 (1)	4 (1)	2 (0)	65 (11)	14 (2)
BetterFuncPAG (3)	107 (19)	8 (1)	13 (2)	5 (1)	4 (1)	1 (0)	44 (8)	6 (1)
AdvicePAG (4)	13 (2)	1 (0)	1 (0)	0	0	0	5 (1)	3 (1)
Sum	324 (55.9)	23 (4.0)	39 (6.7)	11 (1.9)	10 (1.7)	5 (0.9)	144 (24.8)	24 (4.1)

[a]PaG: patient goal.

Feature Selection and Information Gain

The number of features to be considered in both cases was determined to be 29 for the BD and 30 for the ED based on the applied feature relevance threshold of 0.7. Table 6 summarizes the selected features in the order of decreasing information gain, indicating the information gain of the AnesthesiologyReRe and RehabReRe referral reason features as the second- and fourth-most important referral features. Including the referral letter parameters in the BD did not change the general feature-ranking types between the data sets.

XSL•FO
RenderX

Table 6. Ranked features based on ReliefF feature selection for the BD[a] and the ED[b].

Rank	ED		BD	
	Feature	ReliefF	Feature	ReliefF
1	Have you been absent from your work in the past 4 weeks because you were sick?	0.127	Have you been absent from your work in the past 4 weeks because you were sick?	0.128
2	AnesthesiologyReRe[c,d]	0.118	Have there been days in the past 4 weeks when you worked but suffered from physical or psychological problems during your work?	0.109
3	Have there been days in the past 4 weeks when you worked but suffered from physical or psychological problems during your work?	0.107	How many hours a week do you work? Add together all the hours for which you are paid.	0.103
4	RehabilitationReRe[d]	0.106	Have you used injections (eg, epidural steroid or facet injections)?	0.103
5	How many hours a week do you work? Add together all the hours for which you are paid.	0.104	Have you ever had a low back operation?	0.101
6	Have you used injections (eg, epidural steroid or facet injections)?	0.103	Have you used exercise therapy?	0.093
7	Have you ever had a low back operation?	0.102	Do you think, based on your current health situation, you will still be able to work for the next 2 years?	0.092
8	Leg pain is dominant.	0.093	Regarding your work in general, how pleased are you with your job as a whole, considering everything?	0.091
9	Do you think, based on your current health situation, you will still be able to work for the next 2 years?	0.092	Leg pain is dominant.	0.090
10	Have you used exercise therapy?	0.092	Is your work recognized and appreciated by management?	0.085
11	Regarding your work in general, how pleased are you with your job as a whole, considering everything?	0.088	How often is your immediate superior willing to listen to your work-related problems?	0.085
12	Is your work recognized and appreciated by management?	0.085	Are you treated fairly at your workplace?	0.085
13	How often is your immediate superior willing to listen to your work-related problems?	0.084	How often are your colleagues willing to listen to your work-related problems?	0.084
14	Are you treated fairly at your workplace?	0.084	Is your work meaningful?	0.084
15	How often are your colleagues willing to listen to your work-related problems?	0.084	Are conflicts resolved in a fair way?	0.083
16	Are conflicts resolved in a fair way?	0.083	How often do you get help and support from your nearest superior?	0.082
17	Is your work meaningful?	0.082	Do you have enough time for your work tasks?	0.080
18	How often do you get help and support from your nearest superior?	0.082	Have you used opioid painkillers (prescription medications, such as Vicodin, Lortab, Norco, hydrocodone, codeine, Tylenol, Fentanyl, Duragesic, MS Contin, Percocet, Tylox, OxyContin, oxycodone, methadone, tramadol, Ultram, or Dilaudid)?	0.080
19	Do you have enough time for your work tasks?	0.079	Do you have to relate to other people's personal problems as part of your work?	0.079
20	How often do you get help and support from your colleagues?	0.079	How often do you get help and support from your colleagues?	0.079
21	Is your workload unevenly distributed so it piles up?	0.078	Do you feel that the work you do is important?	0.079
22	Do you feel that the work you do is important?	0.078	Is your workload unevenly distributed so it piles up?	0.078

XSL•FO
RenderX

Rank	ED		BD	
	Feature	ReliefF	Feature	ReliefF
23	Do you have to relate to other people's personal problems as part of your work?	0.078	Do you feel that your work drains so much of your energy that it has a negative effect on your private life?	0.076
24	Have you used opioid painkillers (prescription medications, such as Vicodin, Lortab, Norco, hydrocodone, codeine, Tylenon, Fentanyl, Duragesic, MS Contin, Percocet, Tylox, OxyContin, oxycodone, methadone, tramadol, Ultram, or Dilaudid)?	0.076	Can you influence the amount of work assigned to you?	0.076
25	Do you feel that your work drains so much of your energy that it has a negative effect on your private life?	0.076	Does your work put you in emotionally disturbing situations?	0.075
26	Do you have to work fast?	0.075	Do you have to work fast?	0.075
27	Would you ask a good friend to apply for a position at your workplace?	0.074	Would you ask a good friend to apply for a position at your workplace?	0.075
28	Does your work put you in emotionally disturbing situations?	0.074	Do you get behind with your work?	0.071
29	Can you influence the amount of work assigned to you?	0.074	Do you feel that your work takes so much of your time that it has a negative effect on your personal life?	0.070

[a]BD: basic data set.

[b]ED: extended data set.

[c]ReRe: referral reason.

[d]AnesthesiologyReRe and RehabReRe were the second- and fourth-most important referral reason features.

Accuracy of Machine Learning Models

In general, F_1-scores were low for both the BD and the ED, with values ranging from 0.28 to 0.54. Comparing the F_1-scores between the BD and the ED showed an increase of up to 19.5% for the F_1-score when considering the additional referral reason features (see Table 7).

Table 7. Table 7.F1-scores of the ML[a] methods for both data sets.

ML method	F_1-scores for the ED[b] (top 30)	F_1-scores for the BD[c] (top 29)
SVM[d]	0.535	0.355
kNN[e]	0.505	0.315
MLP[f]	0.470	0.275

[a]ML: machine learning.

[b]ED: extended data set.

[c]BD: basic data set.

[d]SVM: support vector machine.

[e]kNN: k-nearest neighbor.

[f]MLP: multilayer perceptron.

In addition, the confusion matrices of the SVM, kNN, and MLP among both conditions are shown in Figure 2. Each column and row represent 1 treatment category. The rows represent the actual (annotated) categories. The columns represent the predicted treatment categories by classifier. The numbers in the cells of a confusion matrix represent the number of samples that have been predicted for the column but are annotated for the row. High numbers in the cells on the diagonal line from the upper left to the lower right cell represent TPs (correctly predicted samples). The numbers in the surrounding cells represent faulty detections. Consequently, yellow and green colors in the diagonal are good, and blue in the surrounding is good as it represents a low number of misses. Yellow and green in the surroundings indicate a stronger error.

XSL•FO
RenderX

Figure 2. Confusion matrices for data set–specific ML models for the ED with 30 features and the BD with 29 features. BD: basic data set; ED: extended data set; kNN: k-nearest neighbor; ML: machine learning; MLP: multilayer perceptron; SVM: support vector machine.

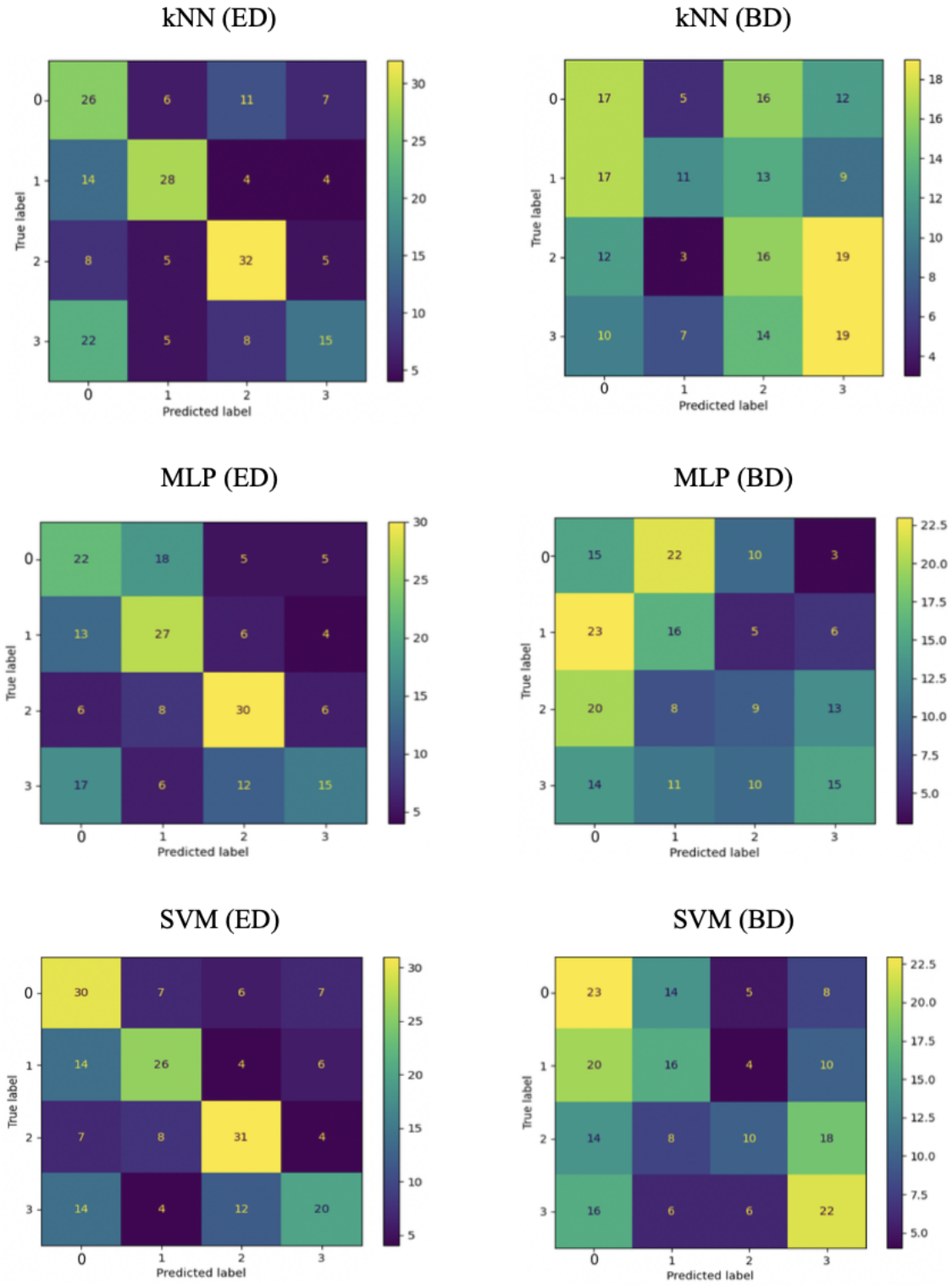

The confusion matrices highlighted the benefit of the additional referral letter parameters, as the ED had a much higher agreement among the predicted and true labels than the BD, which is indicated by having more yellow and green on the diagonal upper left to lower right and more blue surrounding cells in the ED compared to the BD.

Discussion

Principal Findings

The results of this study indicate that quantitative self-report data from patients with LBP can be enriched by unstructured and qualitative data collected from referral letters as part of EHRs using NLP. NLP appears as a feasible option, and we found that the performance of our ML models increased

XSL•FO
RenderX

significantly by up to 19.5%. When considering the relevance of referral reasons from EHRs, especially the referral reasons for the AnesthesiologieReRe and RehabReRe categories were relevant, being the second- and fourth-most important referral features of the 30 relevant items. Considering the occurrence of corresponding fields, in 180 referrals, there was a question for the anesthesiology treatment, whereas 228 patients were actually triaged to anesthesiology. For the 571 patients triaged to rehabilitation, in 442 occurrences, the referral reason indicated rehabilitation as well, with an agreement of 38.4%, indicating that the referrer influences the triaging process for the patient in secondary and tertiary care.

What we could not find were direct referrals to a neurosurgery intervention, probably because this is unusual in the Netherlands. More common is to refer patients with a referral reason to more diagnostics by, for example, a neurologist who will, in positive cases, refer them to neurosurgery. With neurosurgery being a small sample in our data set and the code words resulting from visual inspection of the data set, that might relate to one of the following:

- The suggestion is less specific within the text and may also be stated as “Please, more diagnosis” or “Are there treatment options?”
- The small data sample is less representative of the neurosurgery category because only limited cases included neurosurgery and we used the unbalanced data set in the feature selection.

Although the patient goals for help in the domains of “pain reduction” with 42% and “better functioning” with 32% cover the majority of clusters, none of these goals could be directly associated with the triaging decision.

Evaluating the effect of the 2 included referral reasons on the accuracy of the ML models, classic ML approaches (ie, SVM, kNN, and MLP) were chosen as these are reportedly more suitable for such small data sets than deeper neural network architectures. The corresponding results in Table 6 clarify the relevance of the referral reasons in achieving a significantly improved triaging accuracy ranging between 18% and 19.5%. To investigate the effect, these 2 referral reasons hold for ML-supported decision support, and we investigated the confusion matrices (Figure 2). In all cases, we saw a significantly enhanced accuracy of the predicted and annotated treatment categories when considering the referral reasons (ED) compared to not considering them (BD).

It can be concluded that for clinical significance, EHR data can hold valuable information for the prediction of triaging patients with LBP to their treatment. Especially the wording and referrals of the general physician or secondary care specialist significantly and relevantly increase ML model fits. The goals and desires of the patient do not contribute to the prediction of triaging.

Limitations

There are a number of limitations that should be addressed. First, the unbalanced data set and the overrepresented rehabilitation intervention category held 571 (47%) cases, while neurosurgery held only 109 (9%) cases. Considering the referral reasons, we found a similar unbalanced distribution. These unbalanced data might have led to an overrepresented data accuracy of rehabilitation compared to neurosurgery. Therefore, we applied SMOTE data augmentation for the ML model training. With only 59 data entries being available for training the treatment category neurosurgery, nearly 90% of the corresponding training entries were augmented. In addition, only a limited number of referral letters and the included referral reasons and patient goals were available for this treatment category. Consequently, the meaningfulness of the findings regarding the treatment category neurosurgery should be confirmed in another data set. In contrast to this minority treatment category, the results of the other treatment categories can be assumed to be representative.

A second limitation is the number of missing data and letters that we could derive from the EHRs. Although most could be retrieved, it appeared impossible to retrieve all letters, and in some cases, letters may have bene be missing or may not have been uploaded properly. In addition, although cases were missing, it may have led to higher external validity. In many real-world clinical situations, data sets are incomplete. Therefore, we decided to include what we could derive following a commonly applied approach.

Lastly, we could, based on these baseline data, not conclude whether the triaging to the treatment was the correct triaging in terms of the treatment with the highest benefit to the patient and whether the patient indeed was successfully treated. Future longitudinal studies may lead to a better understanding of the use of ML-based clinical decision-making in patients with LBP, for example, by including the results of treatments.

Strengths

In addition to the quantitative research questions that we answered, this study contributes to the discussion of the meaningfulness of the general physician’s referral reasons and patient goals in a group of patients with LBP. Correspondingly, the resulting input feature ranking of both data sets is expected to represent the information relevance of the feature and might indicate the impact, referral reasons, and patient goals have on treatment category (class) selection.

Considering the generalizability toward an additional language or to other LBP-related questionnaires should be straightforward. Although adjustments of search terms and potential alternative document structures are required, the proposed approach allows quick adaptability. Covering additional diseases or using additional or different questionnaires will require clinical and methodological knowledge of the corresponding domain experts.

Conclusion

Among the ReliefF-prioritized features, 2 referral reason features were highly relevant, and their consideration increased the F_1-score accuracy of the models by up to 19.5%. The results were confirmed by visual inspection of confusion matrices, although the overall performance of the ML models remains low and they cannot be clinically applied at this moment.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Oldenburg Research Pool (FP 2019-035).

Authors' Contributions

SF contributed to the design, methods, analysis, and drafting of the paper; CB, design, methods, discussion, and review of the paper; JS and HT, design, methods, analyses, and review of the paper; PS, design, methods, analysis, discussion, and review of the paper; HRS, MFR, and APW, design, data gathering, results, discussion, and review of the paper; and RS, design, methods, analyses, discussion, and drafting of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Referral reason categories and code words.
[PNG File , 198 KB - jmir_v26i1e46857_app1.png]

Multimedia Appendix 2
Patient goal categories and code words.
[PNG File , 221 KB - jmir_v26i1e46857_app2.png]

References

1. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet 2017 Feb;389(10070):736-747. [doi: 10.1016/s0140-6736(16)30970-9]
2. Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, et al. Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet 2018 Jun;391(10137):2368-2383. [doi: 10.1016/s0140-6736(18)30489-6]
3. Foster NE, Hill JC, O'Sullivan P, Hancock M. Stratified models of care. Best Pract Res Clin Rheumatol 2013 Oct;27(5):649-661. [doi: 10.1016/j.berh.2013.10.005] [Medline: 24315146]
4. Linton SJ, Halldén K. Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. Clin J Pain 1998 Sep;14(3):209-215. [doi: 10.1097/00002508-199809000-00007] [Medline: 9758070]
5. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthritis Rheum 2008 May 15;59(5):632-641 [FREE Full text] [doi: 10.1002/art.23563] [Medline: 18438893]
6. da Silva T, Macaskill P, Kongsted A, Mills K, Maher CG, Hancock MJ. Predicting pain recovery in patients with acute low back pain: updating and validation of a clinical prediction model. Eur J Pain 2019 Feb;23(2):341-353. [doi: 10.1002/ejp.1308] [Medline: 30144211]
7. d'Hollosy W, van Velsen L, Poel M, Groothuis-Oudshoorn C, Soer R, Stegeman P, et al. Applying machine learning on patient-reported data to model the selection of appropriate treatments for low back pain: a pilot study. 2020 Presented at: BIOSTEC 2020: 13th International Joint Conference on Biomedical Engineering Systems and Technologies; February 24-26, 2020; Valetta, Malta. [doi: 10.5220/0008962101170124]
8. van Hooff ML, van Loon J, van Limbeek J, de Kleuver M. The Nijmegen decision tool for chronic low back pain. Development of a clinical decision tool for secondary or tertiary spine care specialists. PLoS One 2014 Aug 18;9(8):e104226 [FREE Full text] [doi: 10.1371/journal.pone.0104226] [Medline: 25133645]
9. Dutmer AL, Schiphorst Preuper HR, Soer R, Brouwer S, Bültmann U, Dijkstra PU, et al. Personal and societal impact of low back pain: The Groningen Spine Cohort. Spine (Phila Pa 1976) 2019 Dec 15;44(24):E1443-E1451. [doi: 10.1097/BRS.0000000000003174] [Medline: 31369481]
10. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990 Dec;16(3):199-208 [FREE Full text] [doi: 10.1016/0168-8510(90)90421-9] [Medline: 10109801]
11. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Report of the NIH Task Force on research standards for chronic low back pain. Spine 2014;39(14):1128-1143. [doi: 10.1097/brs.0000000000000434]
12. El Fassi M, Bocquet V, Majery N, Lair ML, Couffignal S, Mairiaux P. Work ability assessment in a worker population: comparison and determinants of Work Ability Index and Work Ability score. BMC Public Health 2013 Apr 08;13:305 [FREE Full text] [doi: 10.1186/1471-2458-13-305] [Medline: 23565883]
13. Pejtersen JH, Kristensen TS, Borg V, Bjorner JB. The second version of the Copenhagen Psychosocial Questionnaire. Scand J Public Health 2010 Feb;38(3 Suppl):8-24. [doi: 10.1177/1403494809349858] [Medline: 21172767]
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357 [FREE Full text] [doi: 10.1613/jair.953]

XSL•FO
RenderX

15. Over-sampling methods. Imbalanced learn. URL: https://imbalanced-learn.org/stable/references/over_sampling.html#module-imblearn.over_sampling [accessed 2023-12-11]
16. Bray T. STD 90, RFC 8259: the JavaScript Object Notation (JSON) Data Interchange Format. RFC Editor. 2017 Dec. URL: https://www.rfc-editor.org/info/rfc8259 [accessed 2023-12-11]
17. re — Regular expression operations. Python Software Foundation. URL: https://docs.python.org/3/library/re.html [accessed 2023-12-11]
18. Robnik-Šikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression. 1997 Presented at: ICML '97: Fourteenth International Conference on Machine Learning; July 8-12, 1997; Nashville, TN.
19. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 2003 Oct;53(1/2):23-69 [FREE Full text] [doi: 10.1023/a:1025667309714]
20. Guyon I, Nikravesh M, Gunn S, Zadeh LA. Feature Extraction. Berlin, Heidelberg: Springer; 2006.
21. Kononenko I. Estimating attributes: analysis and extensions of RELIEF. 1994 Presented at: Machine Learning: ECML-94; European Conference on Machine Learning; April 6-8, 1994; Catania, Italy p. 171-182 URL: https://doi.org/10.1007/3-540-57868-4 [doi: 10.1007/3-540-57868-4_57]
22. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. 1992 Presented at: AAAI-92: Tenth National Conference on Artificial Intelligence; July 12-16, 1992; San Jose, CA.
23. NLP_Triaging Source Code. GitHub. URL: https://github.com/fudickar-lab/NLP_Triaging [accessed 2023-12-11]

Abbreviations

BD: basic data set
CDSS: clinical decision support system
DSS: decision support system
ED: extended data set
EHR: electronic health record
FN: false negative
FP: false positive
kNN: k-nearest neighbor
LBP: low back pain
ML: machine learning
MLP: multilayer perceptron
NH: neighboring samples of class hits
NLP: natural language processing
NM: neighboring samples of class misses
PA: physician assistant
ReLU: rectified linear unit
SMOTE: synthetic minority oversampling technique
SVM: support vector machine
TP: true positive

Edited by A Mavragani; submitted 28.02.23; peer-reviewed by M Wang, Y Han; comments to author 13.04.23; revised version received 12.07.23; accepted 11.10.23; published 30.01.24.

Please cite as:
Fudickar S, Bantel C, Spieker J, Töpfer H, Stegeman P, Schiphorst Preuper HR, Reneman MF, Wolff AP, Soer R
Natural Language Processing of Referral Letters for Machine Learning–Based Triaging of Patients With Low Back Pain to the Most Appropriate Intervention: Retrospective Study
J Med Internet Res 2024;26:e46857
URL: https://www.jmir.org/2024/1/e46857
doi:10.2196/46857
PMID:38289669

©Sebastian Fudickar, Carsten Bantel, Jannik Spieker, Heinrich Töpfer, Patrick Stegeman, Henrica R Schiphorst Preuper, Michiel F Reneman, André P Wolff, Remko Soer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of Two Symptom Checkers (Ada and Symptoma) in the Emergency Department: Randomized, Crossover, Head-to-Head, Double-Blinded Study

Johannes Knitza[1,2,3,4], MD; Ragip Hasanaj[5], BA; Jonathan Beyer[5]; Franziska Ganzer[5]; Anna Slagman[6], Dr rer medic; Myrto Bolanaki[6], MD; Hendrik Napierala[7], MD; Malte L Schmieding[8], MD; Nizam Al-Zaher[5,9], MD; Till Orlemann[3,9], MD; Felix Muehlensiepen[4,10], Dr rer medic; Julia Greenfield[1], PhD; Nicolas Vuillerme[4,11,12], PhD; Sebastian Kuhn[1], MD; Georg Schett[2,3], MD; Stephan Achenbach[5], MD; Katharina Dechant[5], MD

[1]Institute for Digital Medicine, University Hospital Giessen, Philipps University, Marburg, Germany
[2]Department of Internal Medicine 3, Friedrich-Alexander University Erlangen-Nürnberg, Universitätsklinikum Erlangen, Erlangen, Germany
[3]Deutsches Zentrum für Immuntherapie, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany
[4]Université Grenoble Alpes, Grenoble, France
[5]Department of Cardiology, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
[6]Emergency and Acute Medicine and Health Services Research in Emergency Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany
[7]Institute of General Practice and Family Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany
[8]Institute of Medical Informatics, Charité - Universitätsmedizin Berlin, Berlin, Germany
[9]Department of Medicine 1, Friedrich-Alexander University Hospital Erlangen, University Erlangen-Nuremberg, Erlangen, Germany
[10]Centre for Health Services Research Brandenburg, Brandenburg Medical School, Rüdersdorf, Germany
[11]Institut Universitaire de France, Paris, France
[12]Orange Labs & Université Grenoble Alpes, Grenoble, France

Corresponding Author:
Johannes Knitza, MD
Institute for Digital Medicine
University Hospital Giessen
Philipps University
Baldingerstrasse
Marburg, 35043
Germany
Phone: 49 642158 ext 62589
Email: johannes.knitza@uni-marburg.de

Abstract

Background: Emergency departments (EDs) are frequently overcrowded and increasingly used by nonurgent patients. Symptom checkers (SCs) offer on-demand access to disease suggestions and recommended actions, potentially improving overall patient flow. Contrary to the increasing use of SCs, there is a lack of supporting evidence based on direct patient use.

Objective: This study aimed to compare the diagnostic accuracy, safety, usability, and acceptance of 2 SCs, Ada and Symptoma.

Methods: A randomized, crossover, head-to-head, double-blinded study including consecutive adult patients presenting to the ED at University Hospital Erlangen. Patients completed both SCs, Ada and Symptoma. The primary outcome was the diagnostic accuracy of SCs. In total, 6 blinded independent expert raters classified diagnostic concordance of SC suggestions with the final discharge diagnosis as (1) identical, (2) plausible, or (3) diagnostically different. SC suggestions per patient were additionally classified as safe or potentially life-threatening, and the concordance of Ada's and physician-based triage category was assessed. Secondary outcomes were SC usability (5-point Likert-scale: 1=very easy to use to 5=very difficult to use) and SC acceptance net promoter score (NPS).

Results: A total of 450 patients completed the study between April and November 2021. The most common chief complaint was chest pain (160/437, 37%). The identical diagnosis was ranked first (or within the top 5 diagnoses) by Ada and Symptoma in 14% (59/437; 27%, 117/437) and 4% (16/437; 13%, 55/437) of patients, respectively. An identical or plausible diagnosis was ranked first (or within the top 5 diagnoses) by Ada and Symptoma in 58% (253/437; 75%, 329/437) and 38% (164/437; 64%,

XSL•FO
RenderX

281/437) of patients, respectively. Ada and Symptoma did not suggest potentially life-threatening diagnoses in 13% (56/437) and 14% (61/437) of patients, respectively. Ada correctly triaged, undertriaged, and overtriaged 34% (149/437), 13% (58/437), and 53% (230/437) of patients, respectively. A total of 88% (385/437) and 78% (342/437) of participants rated Ada and Symptoma as very easy or easy to use, respectively. Ada's NPS was –34 (55% [239/437] detractors; 21% [93/437] promoters) and Symptoma's NPS was –47 (63% [275/437] detractors and 16% [70/437]) promoters.

Conclusions: Ada demonstrated a higher diagnostic accuracy than Symptoma, and substantially more patients would recommend Ada and assessed Ada as easy to use. The high number of unrecognized potentially life-threatening diagnoses by both SCs and inappropriate triage advice by Ada was alarming. Overall, the trustworthiness of SC recommendations appears questionable. SC authorization should necessitate rigorous clinical evaluation studies to prevent misdiagnoses, fatal triage advice, and misuse of scarce medical resources.

Trial Registration: German Register of Clinical Trials DRKS00024830; https://drks.de/search/en/trial/DRKS00024830

(J Med Internet Res 2024;26:e56514) doi:10.2196/56514

KEYWORDS

symptom checker; triage; emergency; eHealth; diagnostic accuracy; apps, health service research; decision support system

Introduction

The number of emergency department (ED) visits is continuously increasing [1,2]. From 2009 to 2015, the number of ED outpatient care patients increased by 42% in Germany [1]. The significantly longer length of stay in ED due to the COVID-19 pandemic [2] aggravated the overcrowding of EDs [3]. Furthermore, the modern health care consumer has become accustomed to on-demand services. As a result, a significant number of nonurgent patients use emergency department services due to extended waiting times and limited operating hours of outpatient services [4]. ED crowding is associated with increased patient mortality, morbidity, longer stays, higher costs, medical errors, and staff burnout [4,5]. Multiple solutions to ED crowding have been introduced with varying degrees of success [4], one of which is optimizing ED input [6], by redirecting low-acuity patients to regular outpatient services.

Traditionally, acuity is determined on-site using triage systems with different performances [7]. Ideally, patients should undergo a quick and easily accessible online prescreening to assess their acuity and determine whether redirection to other health care services is appropriate. Demand for telemedicine services has increased due to the COVID-19 pandemic [8,9] and has also transformed emergency care [10]. While on-demand synchronous telephone and video consultations are cost-effective [11], new asynchronous automated services without personnel, such as symptom checkers (SCs), promise greater scalability and cost-effectiveness and are increasingly used by the public [9,12,13]. Implementation of such a preceding remote step-up (asynchronous, then synchronous) assessment successfully resolved 75% of acute care episodes at an American university hospital ED [10]. Winn et al [13], demonstrated that after consulting an SC, the urgency of patients' intended level of care decreased in more than one-quarter of the cases among more than 150,000 patients; however, the study did not include any medical assessments.

In total, 2 of the most promising SCs according to recent studies [14-16] available in multiple languages, including German, are Ada and Symptoma. These underlying studies, however, were based on theoretical vignettes, small sample sizes, used nonrandomized trial designs, and were conducted by the respective manufacturers. Ada demonstrated the highest SC accuracy, only slightly inferior to general physicians in a recent vignette-based study, conducted by Ada [14]. In 2 vignette-based studies [15,16], conducted by Symptoma, Symptoma showed the highest SC accuracy compared with other SCs, including Ada. SC reviews repeatedly call for large-scale, prospective real-world studies [17,18]. Symptoma demonstrated an area under the curve of 0.74 to predict COVID-19 positivity in a large prospective study including 9133 people experiencing COVID-19–like symptoms [15]. This trial aimed to compare the diagnostic accuracy, safety, usability, and acceptance of 2 SCs (Ada and Symptoma) in patients presenting to a University Hospital emergency department.

Methods

Ethical Considerations

The study was approved by the Institutional Review Board of Erlangen University Hospital (approval number #106_19 B), and written informed consent was obtained from all study participants. This study was prospectively registered in the German Register of Clinical Trials (DRKS00024830). Data was pseudonymized and results were reported according to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and onLine TeleHealth) checklist [19]. Patients did not receive any compensation for their participation.

Study Design and Participants

In this randomized, crossover, head-to-head, double-blinded study, patients presenting to the emergency department for internal medicine of the University Hospital Erlangen (Erlangen, Bavaria, Germany) were recruited. Eligible participants were aged 18 years or older. Participants were excluded if they were (1) isolated, (2) unstable, requiring immediate medical attention, or (3) unwilling or unable to give informed consent.

Symptom Checkers

Ada and Symptoma were chosen because of widespread usage, broad coverage of symptoms, free accessibility, and promising, yet contradicting evidence [14-16]. Both SCs are class I medical

XSL•FO
RenderX

devices. Both SCs ask users for general information, including age, sex, and current symptoms. Furthermore, questions are asked based on entered symptoms and answers. Both SCs then present disease suggestions and their likelihood. Ada additionally presents users with recommended actions to take (ie, call an ambulance). Ada was available as a native app and downloaded at the beginning of the study, and available updates were installed as soon as they were available. Symptoma was available as a web app.

Procedures

Participants were randomized 1:1 to group 1 (completing Ada first, continuing with Symptoma) or group 2 (completing Symptoma first, continuing with Ada) by computer-generated block randomization. Assisting personnel were present to help with SC completion on sixth-generation iPad devices (Apple Inc.), if necessary, and to measure completion time. In a consecutive survey, participants rated SC usability and acceptance. ED staff and patients were blinded to SC suggestions. Final discharge diagnosis and patient demographics were recorded. Patients with 2 or more chronic conditions were defined as multimorbid [20]. Chief complaints were categorized using a frequently used list [21,22], including 35 different symptoms, originally published in a textbook [23].

Outcomes

The primary outcome was the diagnostic accuracy of SCs. The secondary outcomes were SC usability and acceptance.

Primary Outcome

Diagnostic accuracy was defined as concordance between the final discharge diagnosis and top-1 (D1) and up to top-5 (D5) SC diagnoses. SC diagnoses were restricted to a maximum of 5 suggestions. If patients were admitted as inpatients, this discharge diagnosis was used in place of the ED diagnosis.

Suggested SC diagnoses were blindly reviewed by 6 ED-experienced physicians (2 of whom were board-certified emergency physicians) who classified diagnostic concordance as (1) identical, (2) plausible, or (3) diagnostically different, following the methodology of Hautz et al [24] and Bastakoti et al [25]. In addition, they classified SC suggestions per patient as safe or potentially life-threatening. Multimedia Appendix 1 displays the applied classification, including examples. Participants were randomly assigned so that concordance was assessed by 1 physician from the University Hospital of Erlangen and 1 physician from the University Hospital of Berlin. To standardize assessment, 15 participants were assessed and discussed by all physicians before the actual evaluation. Interrater agreement, using Cohen κ, was moderate (Ada κ=0.54; Symptoma κ=0.45). In case of disagreement, discrepancies were resolved in discussions between physicians. Physicians were blinded to all data except SC diagnoses and final discharge diagnoses. Final discharge diagnoses were classified as (1) confirmed diagnosis, (2) suspected diagnosis, and (3) symptom or exclusion of a diagnosis, by the head of the ED.

Secondary Outcomes

In addition, for Ada, the available SC action advice was compared with a triage categorization, assigned a posteriori, by consensus of 2 physicians, including the local head of the ED. We adopted the pragmatic and frequently followed [7,26] 4 triage–level categories, which are emergent, urgent, routine, and self-care (Multimedia Appendix 1).

SC usability was assessed using a 5-point Likert scale (1=very easy to use and 5=very difficult to use). Acceptance was assessed using the net promoter score (NPS) [27]. Using an 11-point numeric rating scale (0=not at all likely to 10=extremely likely), participants were asked how likely they would recommend the respective SC to a friend. Answers between 0 and 6 are summarized as detractors, 7 and 8 as passives, and 9 and 10 as promoters. The final NPS was calculated by subtracting the percentage of detractors from the percentage of promoters.

Statistical Analysis

We evaluated the cumulative proportion of identical, plausible, or incorrect diagnostic suggestions by Ada and Symptoma with exact 95% CI. A subanalysis was carried out according to the level of urgency, randomization arm, and cases with a confirmed diagnosis. Odds ratios and 95% CI were calculated to compare the proportion of identical or plausible diagnoses. The relationship between comorbidity and correct diagnostic suggestions was investigated using point biserial correlation. SCs also report the estimated probability of the suggested top diagnosis. We calculated the mean of the estimated top diagnosis probability and 95% CI for identical, plausible, or diagnostically different diagnoses to investigate whether a higher average probability was estimated by the SCs for identical diagnoses. All analyses were conducted using SPSS (version 29; IBM Corp; Released 2022). The significance level was set at .05.

Results

Participants

Between April 8, 2021, and November 15, 2021, a total of 537 participants were screened, of whom 450 were eligible and recruited and 437 were analyzed (Figure 1). Recruitment was limited to study nurse availability and daytime.

The mean age was 48.7 (SD 17.9) years. A total of 190/437 (43.5%) patients were female, 221/437 (50.6%) patients were multimorbid, and mean symptom severity was 4.0 (SD 2.7) out of 10 (Table 1). The most common chief complaint and final discharge diagnosis was chest pain (Table 1 and Multimedia Appendix 2). Final discharge diagnoses included 265/437 (61%) confirmed diagnoses, 32/437 (7%) suspected diagnoses, and 140/437 (32%) symptom or exclusion of diagnoses.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram.

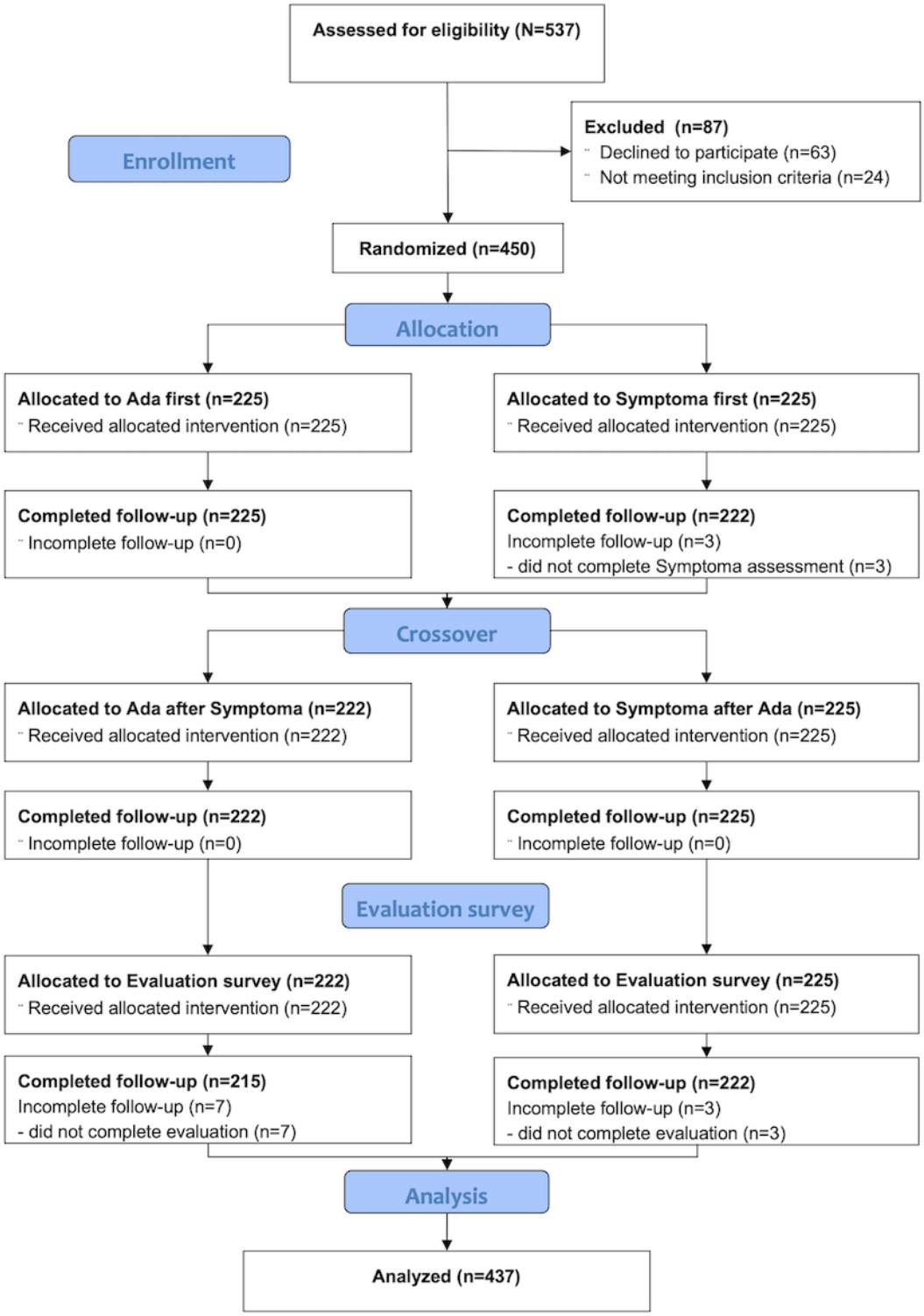

XSL•FO
RenderX

Table 1. Characteristics of study participants.

Characteristics	Total (N=437)
Age (years), mean (SD)	48.7 (17.9)
Female, n (%)	190 (43.5)
Multimorbid, n (%)	221 (50.6)
Symptom severity, mean (SD)	4.0 (2.7)
Chief complaints, n (%)	
Chest pain	160 (36.6)
Abdominal pain	76 (17.4)
Dizziness	32 (7.3)
Weakness	29 (6.6)
Leg pain	10 (2.3)
Other complaints	130 (29.7)
Diagnoses, n (%)	
Chest pain	66 (15.1)
Atrial fibrillation	33 (7.6)
Collapse	27 (6.2)
Hypertension	22 (5.0)
Abdominal pain	21 (4.8)
Gastritis	12 (2.7)
Gastrointestinal bleeding	9 (2.1)
Urinary tract infection	9 (2.1)
NSTEMI[a]	8 (1.8)
Other diagnoses	221 (50.6)
Regular usage of mobile devices, n (%)	
Smartphone and tablet	163 (37.3)
Smartphone only	219 (50.1)
Tablet only	15 (3.4)
None	40 (9.2)
Previous symptom assessment, n (%)	
Online search engines	175 (40.0)
Symptom assessment website/app	47 (10.8)
Physician	149 (34.1)

[a]NSTEMI: non–ST-segment myocardial infarction.

Diagnostic Accuracy

Overall, Ada made fewer suggestions compared with Symptoma (1777 vs 2167) and for several patients suggested only a top diagnosis (40 vs 1) or only one additional suggestion (81 vs 5). The overall odds ratio for an identical diagnosis using Ada compared with Symptoma was 2.54 (95% CI 1.78-3.62; $P<.001$) and for an identical or plausible diagnosis 1.69 (95% CI 1.26-2.27; $P<.001$). Figure 2 and Table 2 show the cumulative proportion of identical and plausible diagnoses with 95% CIs. Ada provided the identical diagnosis as the top diagnosis in 0.14 (95% CI 0.11-0.17) and within the top 5 diagnoses in 0.27 (95% CI 0.23-0.31) of patients, compared with Symptoma, which listed the identical top diagnosis in 0.04 (95% CI 0.02-0.05) and within the top 5 diagnoses in 0.13 (95% CI 0.09-0.16). An identical or plausible diagnosis was provided by Ada as the top diagnosis in 0.58 (95% CI 0.53-0.62) and within the top 5 diagnoses in 0.75 (95% CI 0.71-0.79), compared with Symptoma, which listed an identical or plausible diagnosis as the top diagnosis in 0.38 (95% CI 0.33-0.42) and within the top 5 diagnoses in 0.64 (95% CI 0.60-0.69). No significant differences were observed between randomization groups. Ada's diagnostic performance was meaningfully better compared with Symptoma across all urgency levels and among cases with a confirmed diagnosis (Table 3). A point-biserial correlation was

XSL•FO
RenderX

run to determine the relationship between top diagnostic probability and the number of patient comorbidities. No correlation was found to exist between the number of comorbidities and diagnostic accuracy for Ada (r_{pb}=0.007, n=437; P=.88) or for Symptoma (r_{pb}=–0.036, n=437; P=.46).

The mean reported top diagnosis probability did not correlate well with physicians' classification (Table 4). Ada's mean reported probability for an identical top diagnosis was 0.51 (95% CI 0.46-0.55) and for a diagnostically different diagnosis 0.40 (95% CI 0.38-0.43), compared with Symptoma, which showed a mean top diagnosis probability of 0.75 (95% CI 0.70-0.80) and 0.79 (95% CI 0.78-0.80) for diagnostically different diagnoses.

Ada and Symptoma did not suggest potentially life-threatening diagnoses in 56/437 (13%) and 61/437 (14%), respectively. Compared with the physician-based classification, Ada overall triaged appropriately 149/437 (34%; Figure 3). A total of 74% (86/117) of emergent cases, 23% (56/240) of urgent cases, 11% (7/65) of routine cases, and 0% of self-care cases were adequately triaged by Ada.

Figure 2. Cumulative proportion of identical and identical or plausible diagnoses by rank.

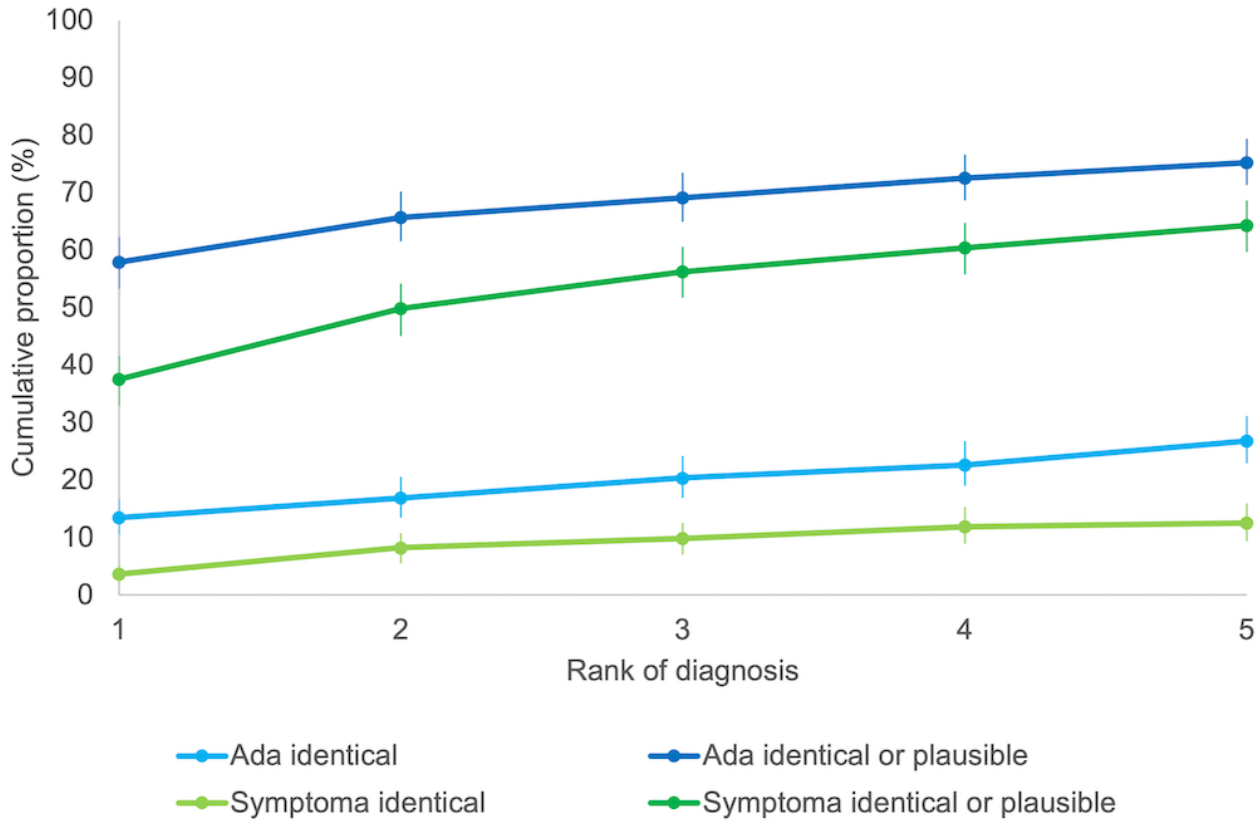

Table 2. Cumulative proportion of identical or plausible diagnoses.

Rank	Ada (n=437)		Symptoma (n=437)	
	Identical, cumulative proportion (95% CI)	Identical or plausible, cumulative proportion (95% CI)	Identical, cumulative proportion (95% CI)	Identical or plausible, cumulative proportion (95% CI)
1	0.14 (0.11-0.17)	0.58 (0.53-0.62)	0.04 (0.02-0.05)	0.38 (0.33-0.42)
2	0.17 (0.14-0.21)	0.66 (0.62-0.70)	0.08 (0.05-0.11)	0.50 (0.45-0.54)
3	0.20 (0.17-0.24)	0.69 (0.65-0.73)	0.10 (0.07-0.13)	0.56 (0.52-0.61)
4	0.23 (0.19-0.27)	0.72 (0.69-0.77)	0.12 (0.9-0.15)	0.60 (0.56-0.65)
5	0.27 (0.23-0.31)	0.75 (0.71-0.79)	0.13 (0.09-0.16)	0.64 (0.60-0.69)

XSL•FO
RenderX

Table 3. The proportion of identical or plausible diagnoses of Ada and Symptoma according to case classification.

	Ada (n=437)				Symptoma (n=437)			
	Top diagnosis		Top 5 diagnoses		Top diagnosis		Top 5 diagnoses	
Case classification	Identical, proportion (95% CI)	Identical or plausible, proportion (95% CI)	Identical, proportion (95% CI)	Identical or plausible, proportion (95% CI)	Identical, proportion (95% CI)	Identical or plausible, proportion (95% CI)	Identical, proportion (95% CI)	Identical or plausible, proportion (95% CI)
Emergent (n=117)	0.10 (0.05-0.16)	0.61 (0.51-0.69)	0.20 (0.13-0.27)	0.74 (0.67-0.82)	0.03 (0.01-0.07)	0.40 (0.31-0.49)	0.10 (0.05-0.15)	0.74 (0.66-0.81)
Urgent (n=240)	0.15 (0.11-0.20)	0.55 (0.49-0.62)	0.32 (0.27-0.37)	0.76 (0.70-0.82)	0.04 (0.02-0.06)	0.36 (0.30-0.43)	0.15 (0.11-0.20)	0.60 (0.54-0.67)
Routine (n=65)	0.12 (0.05-0.20)	0.60 (0.48-0.72)	0.22 (0.11-0.31)	0.71 (0.60-0.82)	0.05 (0.00-0.11)	0.35 (0.23-0.48)	0.09 (0.03-0.17)	0.62 (0.49-0.72)
Self-care (n=15)	0.13 (0.00-0.33)	0.67 (0.40-0.87)	0.27 (0.07-0.53)	0.87 (0.67-1.00)	0.00 (0.00-0.00)	0.47 (0.20-0.73)	0.07 (0.00-0.20)	0.67 (0.47-0.87)
Confirmed diagnosis (n=265)	0.17 (0.13-0.21)	0.52 (0.46-0.58)	0.35 (0.33-0.38)	0.8 (0.6-0.9)	0.05 (0.03-0.08)	0.3 (0.25-0.36)	0.32 (0.29-0.34)	0.4 (0.3-0.5)

Table 4. Mean reported top diagnosis probability and 95% CI for identical, plausible, or diagnostically different diagnoses.

Mean diagnostic probability	Ada (n=437)	Symptoma (n=437)
Identical, mean (95% CI)	0.51 (0.46-0.55)	0.75 (0.70-0.80)
Plausible, mean (95% CI)	0.42 (0.40-0.44)	0.81 (0.80-0.81)
Diagnostically different, mean (95% CI)	0.40 (0.38-0.43)	0.79 (0.78-0.80)

Figure 3. Confusion matrix of (A) Ada's action recommendations and (B) overall triage accuracy compared with physician categorization.

A		**Physician-based urgency categorization**, N (%)			
		Emergent	Urgent	Routine	Self-care
Action recommendations by Ada, N (%)		117 (27)	240 (55)	65 (15)	15 (3)
Ambulance	99 (23)	43 (10)	45 (10)	8 (2)	3 (1)
Emergency	198 (45)	43 (10)	115 (26)	33 (8)	7 (2)
Primary care within hours	95 (22)	21 (5)	56 (13)	14 (3)	4 (1)
Primary care within 1-3 days	35 (8)	7 (2)	20 (5)	7 (2)	1 (0)
Primary care within 2-3 weeks	3 (1)	2 (0)	1 (0)	0 (0)	0 (0)
Self-care	7 (2)	1 (0)	3 (1)	3 (1)	0 (0)

B	**Triage categories**, N (%)				
	Emergent	**Urgent**	**Routine**	**Self-care**	**Overall**
Triage accuracy	117 (27)	240 (55)	65 (15)	15 (3)	437 (100)
Undertriaged	31 (27)	24 (10)	3 (5)	NA	58 (13)
Overtriaged	NA	160 (67)	55 (85)	15 (100)	230 (53)
Accurately triaged	86 (74)	56 (23)	7 (11)	0 (0)	149 (34)

Usability and Acceptance

A total of 385/437 (88%) and 342/437 (78%) patients rated Ada and Symptoma as very easy or easy to use, respectively.

Median completion time for Ada was 7 (IQR 5-9) and 5 (IQR 3-6) minutes for Symptoma. A total of 114/437 (26%) and 113/437 (26%) patients requested help from study personnel to use Ada and Symptoma, respectively. Ada's NPS was –34 due to 239/437 (55%) detractors and 93/437 (21%) promoters. Symptoma's NPS was –47 with 275/437 (63%) detractors and 70/437 (16%) promoters.

XSL•FO
RenderX

Discussion

To the best of our knowledge, this is the first and largest prospective head-to-head trial comparing the diagnostic accuracy, usability, and acceptance of 2 SCs (Ada and Symptoma), used by patients themselves. Our results elucidate that the diagnostic accuracy of both SCs was substantially lower compared with the final discharge diagnosis by clinicians.

Ada made fewer suggestions but reported an identical or plausible diagnosis significantly more often compared with Symptoma. This observed trend was independent of patient urgency level. One reason for this difference may be that Ada had more time to gather information, as the average completion time was 2 minutes longer. In line with a previous observation [28], the mean reported probability of the top diagnosis was misleading. Symptoma's mean reported probability for diagnostically different diagnoses was actually higher than for identical diagnoses (79% vs 75%). Potentially life-threatening diagnoses were missed by both SCs equally often. Patients rated Ada as easier to use and were more likely to recommend it to other patients compared with Symptoma.

The landmark study by Semigram et al [18], benchmarking 23 symptom checkers using 45 vignettes, reported diagnostic accuracy of 34% for the top diagnosis and 51% for the top 3 diagnoses. Due to the theoretical and vignette-based nature of the study, direct comparison of our results with those of Semigran et al [18], is limited. The observed diagnostic inferiority of SCs compared with physicians is in line with previous studies [14,29,30]. Gilbert et al [14] reported an accuracy of 71% for Ada compared with general physicians with 82% based on case vignettes. In a previous rheumatology-based randomized controlled trial investigating Ada's accuracy used by patients themselves, we observed a diagnostic accuracy of 43% (D1) and 54% (D5) regarding the detection of inflammatory rheumatic diseases [30]. Faqar-Uz-Zaman et al [29] reported results from a similar study investigating Ada in the ED in patients with abdominal pain. Ada suggested the discharge diagnosis in 52% (D5), compared with 81% by ED physicians. The authors suggested that physicians using the SC suggestions could theoretically improve physician's accuracy by 10%. Martinez-Franco et al [31] reported a significantly higher accuracy in general physicians using an SC (DXplain) compared with a group without the tool.

Comparing the accuracy of SCs and physicians should be done carefully, as physicians had access to substantially more information including data from laboratory tests and imaging.

Previous studies have also reported diagnostic discrepancies between the initial ED diagnosis and the final discharge diagnosis [24,25]. It has also been shown that the diagnostic accuracy of physicians was lower than that of Ada when limited to symptom-related medical history [32]. The focus of this study was therefore to compare 2 SCs with each other, used by the same patient. To exclude a potential bias of the order of SC usage, patients were randomized.

Schmieding et al [33] showed that no symptom checker among 22 outperformed laypersons in deciding whether emergency care or self-care was adequate and that triage accuracy did not improve after 5 years, missing >40% of emergencies. Our study confirms results from previous studies [18,34] reporting that emergency cases are triaged more accurately than less-urgent cases. The percentage of undertriaged patients (13%) was slightly higher compared with a previous trial (9%) that investigated Ada in an interdisciplinary University Hospital ED [35].

In the study by Miller et al [36], 98% of patients reported Ada as very easy or quite easy to use, compared with 88% in this study. In an expert heuristic review of chatbots, Ada and Symptoma received an overall rating of 6.3/12 and 7.0/12, respectively [37]. The negative NPS indicates poor acceptance and contrasts the previously reported rate of 73% (440/600) [12] and 85% (444/520) [36] of patients who would recommend Ada. The 2 previous studies did not use the NPS but used a binary (Yes/No) approach and were based in a primary care and rheumatology setting. The negative NPS also contrasts with the positive Ada ratings on the German Apple App Store (4.7/5) and Google Play Store (4.6/5). We believe that the rather strict NPS rating system (only 9 and 10 counted as promoters) is the main reason for the difference.

This study has several strengths and limitations. Suggested SC diagnoses were blindly reviewed by physicians and patients were blinded to SC reports. For each case, 1 of the 2 assessors came from a completely different center (Charité, Berlin) to ensure a maximum of objectivity. The multidisciplinary study team involving nursing staff, as well as dedicated health service researchers, represents a strength of the study. The cardiology focus and single-center nature of the study limit generalizability. Moderate interrater agreement represents a common study limitation. We did not prespecify an effect size nor carried out a sample size calculation; however, our sample size is similar to the 2 largest SC studies [38,39] in an ED setting. A limitation to generalizability is the rapid pace of digital diagnostic advancements. This is due to the continuous updates of SCs and the emergence of powerful large language models, such as ChatGPT (OpenAI), which provide new diagnostic decision support [40,41]. A potential bias was that analysis was restricted to the top 5 diagnoses and Symptoma offers up to 30 suggestions. Furthermore, Symptoma's guidelines specifically advise against its use in emergencies. The large size of the study, head-to-head nature, manufacturer independence, real-world setting, and high percentage of emergent patients represent strengths of the study.

Conclusions

The accuracy and safety of symptom checkers appear inferior to a complete physician-based assessment. A substantial number of potentially life-threatening diagnoses were missed by both symptom checkers and the high number of patients undertriaged by Ada is alarming. Ada demonstrated a significantly higher diagnostic accuracy, was easier to use, and overall, better rated compared with Symptoma.

XSL•FO
RenderX

Acknowledgments

We thank all the patients who participated as well as the physicians who contributed participants. This work is part of the PhD thesis of the first author JK (AGEIS, Université Grenoble Alpes, Grenoble, France). Open Access funding was provided by the Open Access Publishing Fund of Philipps-Universität Marburg.

Conflicts of Interest

AS receives research funding for studies with the symptom checker apps ADA and SmED from the Ministry of Health (Germany) and the Scientific Institute of Statutory Health Insurance Physicians (Zi). MB receives research funding for studies with symptom checker ADA from the Ministry of Health. HN and MLS receive independent funding of a randomized controlled trial utilizing the Ada symptom checker app from the German Ministry of Health (DRKS-ID: DRKS00028598). SK is the founder and a shareholder of MED.digital.

Multimedia Appendix 1
Diagnostic accuracy, safety, and urgency categorization with examples.
[DOCX File , 23 KB - jmir_v26i1e56514_app1.docx]

Multimedia Appendix 2
Final discharge diagnoses.
[DOCX File , 26 KB - jmir_v26i1e56514_app2.docx]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1151 KB - jmir_v26i1e56514_app3.pdf]

References

1. Wahlster P, Czihal T, Gibis B, Henschke C. [Developments in emergency care - analysis of emergency cases in In- and outpatient care from 2009 to 2015 in Germany]. Gesundheitswesen 2020;82(6):548-558. [doi: 10.1055/a-0820-3904] [Medline: 30786291]
2. Lucero A, Sokol K, Hyun J, Pan L, Labha J, Donn E, et al. Worsening of emergency department length of stay during the COVID-19 pandemic. J Am Coll Emerg Physicians Open 2021;2(3):e12489 [FREE Full text] [doi: 10.1002/emp2.12489] [Medline: 34189522]
3. Kelen GD, Wolfe R, D'Onofrio G, Mills AM, Diercks D, Stern SA, et al. Emergency department crowding: the canary in the health care system. NEJM Catalyst Innovations in Care Delivery Massachusetts Medical Society. 2021. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.21.0217 [accessed 2022-08-09]
4. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLoS ONE 2018;13(8):e0203316. [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]
5. Kelen G, Peterson S, Pronovost P. In the name of patient safety, let's burden the emergency department more. Ann Emerg Med 2016;67(6):737-740. [doi: 10.1016/j.annemergmed.2015.11.031] [Medline: 26707360]
6. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. Ann Emerg Med 2003;42(2):173-180. [doi: 10.1067/mem.2003.302] [Medline: 12883504]
7. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
8. Wong MYZ, Gunasekeran DV, Nusinovici S, Sabanayagam C, Yeo KK, Cheng C, et al. Telehealth demand trends during the COVID-19 pandemic in the top 50 most affected countries: infodemiological evaluation. JMIR Public Health Surveill 2021;7(2):e24445 [FREE Full text] [doi: 10.2196/24445] [Medline: 33605883]
9. Kernder A, Morf H, Klemm P, Vossen D, Haase I, Mucke J, et al. Digital rheumatology in the era of COVID-19: results of a national patient and physician survey. RMD Open 2021;7(1):e001548 [FREE Full text] [doi: 10.1136/rmdopen-2020-001548] [Medline: 33622673]
10. Hollander JE, Sharma R. The availablists: emergency care without the emergency department. NEJM Catalyst Innovations in Care Delivery Massachusetts Medical Society. 2021. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.21.0310 [accessed 2022-08-09]
11. Nord G, Rising KL, Band RA, Carr BG, Hollander JE. On-demand synchronous audio video telemedicine visits are cost effective. Am J Emerg Med 2019;37(5):890-894. [doi: 10.1016/j.ajem.2018.08.017] [Medline: 30100333]

XSL•FO
RenderX

12. Knitza J, Muehlensiepen F, Ignatyev Y, Fuchs F, Mohn J, Simon D, et al. Patient's perception of digital symptom assessment technologies in rheumatology: results from a multicentre study. Front Public Health 2022;10:844669 [FREE Full text] [doi: 10.3389/fpubh.2022.844669]
13. Winn AN, Somai M, Fergestrom N, Crotty BH. Association of use of online symptom checkers with patients' plans for seeking care. JAMA Netw Open 2019;2(12):e1918561 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18561] [Medline: 31880791]
14. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open 2020;10(12):e040269 [FREE Full text] [doi: 10.1136/bmjopen-2020-040269] [Medline: 33328258]
15. Munsch N, Martin A, Gruarin S, Nateqi J, Abdarahmane I, Weingartner-Ortner R, et al. Diagnostic accuracy of web-based COVID-19 symptom checkers: comparison study. J Med Internet Res 2020;22(10):e21299 [FREE Full text] [doi: 10.2196/21299] [Medline: 33001828]
16. Nateqi J, Lin S, Krobath H, Gruarin S, Lutz T, Dvorak T, et al. Vom symptom zur diagnose – tauglichkeit von symptom-checkern. HNO 2019;67(5):334-342. [doi: 10.1007/s00106-019-0666-y] [Medline: 30993374]
17. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022;5(1):118 [FREE Full text] [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
18. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
19. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
20. Johnston MC, Crilly M, Black C, Prescott GJ, Mercer SW. Defining and measuring multimorbidity: a systematic review of systematic reviews. Eur J Public Health 2019;29(1):182-189 [FREE Full text] [doi: 10.1093/eurpub/cky098] [Medline: 29878097]
21. Bingisser R, Dietrich M, Nieves Ortega R, Malinovska A, Bosia T, Nickel CH. Systematically assessed symptoms as outcome predictors in emergency patients. Eur J Intern Med 2017;45:8-12. [doi: 10.1016/j.ejim.2017.09.013] [Medline: 29074217]
22. Rueegg M, Nickel CH, Bingisser R. Disagreements between emergency patients and physicians regarding chief complaint - patient factors and prognostic implications. Int J Clin Pract 2021;75(5):e14070. [doi: 10.1111/ijcp.14070] [Medline: 33533559]
23. Marx J, Walls R, Hockberger R. Rosen's emergency medicine-concepts and clinical practice. In: Elsevier Health Sci. Philadelphia, PA: Elsevier/Saunders; 2013.
24. Hautz WE, Kämmer JE, Hautz SC, Sauter TC, Zwaan L, Exadaktylos AK, et al. Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. Scand J Trauma Resusc Emerg Med 2019;27(1):54 [FREE Full text] [doi: 10.1186/s13049-019-0629-z] [Medline: 31068188]
25. Bastakoti M, Muhailan M, Nassar A, Sallam T, Desale S, Fouda R, et al. Discrepancy between emergency department admission diagnosis and hospital discharge diagnosis and its impact on length of stay, up-triage to the intensive care unit, and mortality. Diagnosis (Berl) 2021;9(1):107-114. [doi: 10.1515/dx-2021-0001] [Medline: 34225399]
26. Chan F, Lai S, Pieterman M, Richardson L, Singh A, Peters J, et al. Performance of a new symptom checker in patient triage: canadian cohort study. PLoS One 2021;16(12):e0260696 [FREE Full text] [doi: 10.1371/journal.pone.0260696] [Medline: 34852016]
27. Reichheld F. The one number you need to grow.: Harvard Business Review; 2003. URL: https://hbr.org/2003/12/the-one-number-you-need-to-grow [accessed 2024-07-10]
28. Knitza J, Tascilar K, Gruber E, Kaletta H, Hagen M, Liphardt A, et al. Accuracy and usability of a diagnostic decision support system in the diagnosis of three representative rheumatic diseases: a randomized controlled trial among medical students. Arthritis Res Ther 2021;23(1):233 [FREE Full text] [doi: 10.1186/s13075-021-02616-6] [Medline: 34488887]
29. Faqar-Uz-Zaman SF, Anantharajah L, Baumartz P, Sobotta P, Filmann N, Zmuc D, et al. The diagnostic efficacy of an app-based diagnostic health care application in the emergency room: eRadaR-Trial. A prospective, double-blinded, observational study. Ann Surg 2022;276(5):935-942. [doi: 10.1097/SLA.0000000000005614] [Medline: 35925755]
30. Knitza J, Mohn J, Bergmann C, Kampylafka E, Hagen M, Bohr D, et al. Accuracy, patient-perceived usability, and acceptance of two symptom checkers (Ada and Rheport) in rheumatology: interim results from a randomized controlled crossover trial. Arthritis Res Ther 2021;23(1):112 [FREE Full text] [doi: 10.1186/s13075-021-02498-8] [Medline: 33849654]
31. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, Hernandez-Torres I, Rivero-Lopez C, Spicer T, et al. Diagnostic accuracy in family medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. Diagnosis (Berl) 2018;5(2):71-76. [doi: 10.1515/dx-2017-0045] [Medline: 29730649]
32. Gräf M, Knitza J, Leipe J, Krusche M, Welcker M, Kuhn S, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 2022;42(12):2167-2176 [FREE Full text] [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]

XSL•FO
RenderX

33. Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. J Med Internet Res 2021;23(3):e24475 [FREE Full text] [doi: 10.2196/24475] [Medline: 33688845]
34. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Med J Aust 2020;212(11):514-519. [doi: 10.5694/mja2.50600] [Medline: 32391611]
35. Cotte F, Mueller T, Gilbert S, Blümke B, Multmeier J, Hirsch MC, et al. Safety of triage self-assessment using a symptom assessment app for walk-in patients in the emergency care setting: observational prospective cross-sectional study. JMIR Mhealth Uhealth 2022;10(3):e32340 [FREE Full text] [doi: 10.2196/32340] [Medline: 35343909]
36. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence-based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. JMIR Hum Factors 2020;7(3):e19713 [FREE Full text] [doi: 10.2196/19713] [Medline: 32540836]
37. Höhn S, Bongard-Blanchy K. Heuristic evaluation of COVID-19 chatbots. In: Følstad A, Araujo T, Papadopoulos S, Law ELC, Luger E, Goodwin M, et al, editors. Chatbot Research and Design. US: Springer International Publishing; 2021:131-144.
38. Faqar-Uz-Zaman SF, Filmann N, Mahkovic D, von Wagner M, Detemble C, Kippke U, et al. Study protocol for a prospective, double-blinded, observational study investigating the diagnostic accuracy of an app-based diagnostic health care application in an emergency room setting: the eRadaR trial. BMJ Open 2021;11(1):e041396 [FREE Full text] [doi: 10.1136/bmjopen-2020-041396] [Medline: 33419909]
39. Ramnarayan P, Cronje N, Brown R, Negus R, Coode B, Moss P, et al. Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. Emerg Med J 2007;24(9):619-624 [FREE Full text] [doi: 10.1136/emj.2006.044107] [Medline: 17711936]
40. Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med 2024;83(1):83-86. [doi: 10.1016/j.annemergmed.2023.08.003] [Medline: 37690022]
41. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024;44(2):303-306 [FREE Full text] [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and onLine TeleHealth
ED: emergency department
NPS: net promoter score
SC: symptom checker

Edited by S Ma; submitted 19.01.24; peer-reviewed by H Li, J Nateqi, W Hautz; comments to author 15.03.24; revised version received 19.03.24; accepted 21.06.24; published 20.08.24.

Please cite as:
Knitza J, Hasanaj R, Beyer J, Ganzer F, Slagman A, Bolanaki M, Napierala H, Schmieding ML, Al-Zaher N, Orlemann T, Muehlensiepen F, Greenfield J, Vuillerme N, Kuhn S, Schett G, Achenbach S, Dechant K
Comparison of Two Symptom Checkers (Ada and Symptoma) in the Emergency Department: Randomized, Crossover, Head-to-Head, Double-Blinded Study
J Med Internet Res 2024;26:e56514
URL: https://www.jmir.org/2024/1/e56514
doi:10.2196/56514
PMID:

©Johannes Knitza, Ragip Hasanaj, Jonathan Beyer, Franziska Ganzer, Anna Slagman, Myrto Bolanaki, Hendrik Napierala, Malte L Schmieding, Nizam Al-Zaher, Till Orlemann, Felix Muehlensiepen, Julia Greenfield, Nicolas Vuillerme, Sebastian Kuhn, Georg Schett, Stephan Achenbach, Katharina Dechant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Unveiling Usage Patterns and Explaining Usage of Symptom Checker Apps: Explorative Longitudinal Mixed Methods Study

Anna-Jasmin Wetzel[1], MSc; Christine Preiser[2], PhD; Regina Müller[3], PhD; Stefanie Joos[1], Prof Dr Med; Roland Koch[1], MD; Tanja Henking[4], Prof Dr; Hannah Haumann[1], MD

[1]Institute of General Practice and Interprofessional Care, University Hospital Tübingen, Tübingen, Germany
[2]Institute of Occupational and Social Medicine and Health Services Research, University Hospital Tübingen, Tübingen, Germany
[3]Institute for Philosophy, University of Bremen, Bremen, Germany
[4]Institute for Applied Social Sciences, Technical University of Würzburg-Schweinfurt, Würzburg, Germany

Corresponding Author:
Anna-Jasmin Wetzel, MSc
Institute of General Practice and Interprofessional Care
University Hospital Tübingen
Osianderstr. 5
Tübingen, 72074
Germany
Phone: 49 70712985231
Email: jasmin.wetzel@med.uni-tuebingen.de

Abstract

Background: Symptom checker apps (SCA) aim to enable individuals without medical training to classify perceived symptoms and receive guidance on appropriate actions, such as self-care or seeking professional medical attention. However, there is a lack of detailed understanding regarding the contexts in which individuals use SCA and their opinions on these tools.

Objective: This mixed methods study aims to explore the circumstances under which medical laypeople use SCA and to identify which aspects users find noteworthy after using SCA.

Methods: A total of 48 SCA users documented their medical symptoms, provided open-ended responses, and recorded their SCA use along with other variables over 6 weeks in a longitudinal study. Generalized linear mixed models with and those without regularization were applied to consider the hierarchical structure of the data, and the models' outcomes were evaluated for comparison. Qualitative data were analyzed through Kuckartz qualitative content analysis.

Results: Significant predictors of SCA use included the initial occurrence of symptoms, day of measurement (odds ratio [OR] 0.97), self-rated health (OR 0.80, $P<.001$), and the following International Classification in Primary Care-2–classified symptoms, that are general and unspecified (OR 3.33, $P<.001$), eye (OR 5.56, $P=.001$), cardiovascular (OR 8.33, $P<.001$), musculoskeletal (OR 5.26, $P<.001$), and skin (OR 4.76, $P<.001$). The day of measurement and self-rated health showed minor importance due to their small effect sizes. Qualitative analysis highlighted four main themes: (1) reasons for using SCA, (2) diverse affective responses, (3) a broad spectrum of behavioral reactions, and (4) unmet needs including a lack of personalization.

Conclusions: The emergence of new and unfamiliar symptoms was a strong determinant for SCA use. Specific International Classification in Primary Care–rated symptom clusters, particularly those related to cardiovascular, eye, skin, general, and unspecified symptoms, were also highly predictive of SCA use. The varied applications of SCA fit into the concept of health literacy as bricolage, where SCA is leveraged as flexible tools by patients based on individual and situational requirements, functioning alongside other health care resources.

(J Med Internet Res 2024;26:e55161) doi:10.2196/55161

KEYWORDS

self-triage; eHealth; self-diagnosis; mHealth; mobile health; usage; patterns; predicts; prediction; symptoms checker; apps; applications; explorative longitudinal study; self care; self management; self-rated; mixed method; circumstances; General Linear Mixed Models; GLMM; qualitative data; content analysis; Kuckartz; survey; participants; users

XSL•FO
RenderX

Introduction

Symptom checker apps (SCA) are an emerging class of health technology tools that enable patients to input symptoms and biodata to receive a list of probable diagnoses and related triage suggestions [1]. Research on SCAs has largely focused on their accuracy, for instance, a study by Semigran et al [2] evaluated 23 SCAs using patient vignettes, revealing that the correct diagnosis appeared as the most probable in only 34% of cases. The appropriateness of the triage advice was found to be satisfactory in 57% of cases [2]. Notably, the accuracy of SCAs in nonemergency situations decreased as the urgency of the presented scenarios decreased [2]. A subsequent audit in 2022 indicated that there had been no significant improvement in triage performance over the past 5 years, with considerable variability in effectiveness among different SCAs [3]. Comparisons between SCAs and the decision-making of nonprofessionals showed no superiority in triage capability by SCAs [4]. Another study using simulated patients' cases showed that SCAs correctly identified the primary diagnosis in 30% of cases, significantly less than emergency physicians, who had an 80% success rate [5]. Despite a slightly better performance in emergency triage situations [4], physicians still vastly outperformed SCAs, with a 96% accuracy rate compared with 71% for SCAs [5]. Users of SCAs are typically characterized as younger females with a higher socioeconomic status [6] and a greater affinity for technology [7]. In addition, it has been observed that SCA users showed higher levels of health anxiety than nonusers [8], with SCAs being more widely recognized among these users [6]. Users with anxiety disorders, however, tend to find SCAs less beneficial [6], indicating that SCAs might not be advantageous for this user group. The literature suggests that such patient groups may not only fail to benefit from using SCAs but could also experience heightened anxiety as a result, implying that it might be better to avoid these applications [9].

The impact of SCAs on health care providers is also critical, as patients may use these tools as a substitute for or in addition to professional health care services [10]. The level of trust users place in SCAs and their previous medical knowledge significantly affects how SCA results are interpreted [11,12]. SCA might also lead to the inefficient use of health care resources [5], as users may seek emergency department care too soon or too frequently. This behavior, especially from those with conditions that are not urgent, can intensify the burden on health care systems by potentially increasing costs and reallocating resources away from individuals who require immediate medical attention [13,14]. This outcome contrasts with the promises that SCAs would reduce pressure on health care systems. Currently, it is unclear whether SCAs have an impact on health care systems and, if so, the nature of this impact.

The factors driving SCA usage remain underexplored. Research has shown that gender, age, health anxiety, and other psychosocial factors significantly influence SCA use [6]. User characteristics, including health anxiety, acceptance of technology, and education level, also affect how SCAs are used [15,16]. Despite this knowledge, there is a notable gap in research regarding the situational factors that prompt SCA usage. Gaining a deeper insight into these situational factors that lead to SCA use is essential for comprehending the broader implications of SCA use on both individuals and health care providers. Consequently, this study uses a mixed methods approach to explore how medical laypersons use SCAs, aiming to understand the patterns of use and the users' perceptions following their experience with a symptom checker.

Methods

Overview

This study was conducted using a mixed methods study approach, incorporating both qualitative and quantitative components within a prospective longitudinal cohort study design. The primary outcome of this study was whether one specific SCA was used during a day or not by Ada [17]. Ada [17] was chosen for the study due to its stage of development and its professed broad usage. Moreover, the application is ISO (International Organization for Standardization)-certified and possesses the CE mark. The use of one specific SCA was analyzed to ensure standardized interaction with one SCA only. The STROBE (The Strengthening the Reporting of Observational Studies in Epidemiology) guidelines were applied [18]. This study was conducted as a part of the joint project CHECK.APP [19].

In order to answer the research questions (predictors and usage patterns of SCA, aspects noteworthy after usage) of this mixed methods study, different longitudinal mixed models for SCA use were explored and compared.

Ethical Considerations

Ethical approval was granted by the ethics committee of the medical faculty of the University of Tuebingen (464/202BO). Each participant was compensated 300€ (US $358) for their participation in the study. In accordance with the ethical approval, all participants provided written consent before taking part in the user diaries. A written and verbal explanation was provided, and participation in the study was voluntary.

Measurements

User diaries were filled out daily by the study participants over a period of 6 weeks. SCA use was binary coded (yes or no). Further, the user diary comprised a survey that contained different questions about self-rated mood, self-rated health, self-rated stress level, and occurrence of physical symptoms throughout the day. Self-rated health and mood were both assessed with a visual analog 5-point scale. Subjectively rated stress was assessed with a single-choice item and 3 answer possibilities. If participants had symptoms, they stated what kind of symptoms they had, whether they took medication, and if they used a specific SCA [17]. If participants used the SCA on a day this was classified as a use-case, if the SCA was not used on a day this was classified as a nonuse case. Participants were asked if the symptoms appeared for the first time. Those who reported symptoms were asked to describe them in open-ended responses. After the data acquisition, reported symptoms were classified according to the International Classification in Primary Care 2 (ICPC) by a doctor from the study team, using dummy coding (present not present). Age

XSL•FO
RenderX

was assessed in years, and gender was assessed as an open-ended response resulting in the 2 categories “male” and “female.” The variable “day of study” was retrieved from the dates in the user diaries.

Further questions considering the procedure after the app use were recorded as well. If the SCA was used, open-ended responses in the user diary asked for the experience with the SCA.

Data Collection

The participants of the study were recruited from December 2020 to March 2021, with data collection taking place from April 2021 to September 2021. Due to the exploratory nature of the study, the sample size was determined based on estimated effect sizes and the feasibility of data collection. The requirement of the ethics committee for the study was not to initiate the use of SCA, and therefore only to include participants in the study who were already using SCA before the study. The participants were recruited from a previously conducted survey of the CHECK.APP project, specifically those who reported using a specific SCA [17]. The recruitment process for the survey involved using mailing lists from the University of Tuebingen, a social media campaign by the University Hospital Tuebingen, and social media channels from a statutory health insurance (Allgemeine Ortskrankenkasse Baden-Wuerttemberg). Participants were then contacted through the mail and asked whether they wanted to participate in a further study part.

Prior to the study’s start, web-based workshops were conducted to train participants on how to use the user diary and to enable them to ask questions to prevent common mistakes. Participants were directed to use the SCA as they normally would in their everyday activities. They chose when to start their 6-week period and submitted their diaries after finishing it. About 3 months after the workshop, they were sent reminder emails.

Inclusion and Exclusion Criteria

The inclusion criteria for the study included having at least a B2 level of proficiency in the German language, as classified by the Common European Framework of Reference for Languages, an age older than 18 years, and being an Ada user. The exclusion criterion was possessing a medical degree.

Data Analysis

Quantitative Data

To analyze and process the quantitative data, R statistics (version 4.1.3; R Core Team) [20] and RStudio (version 1.4) [20] were used. The aim of the present mixed methods study is to investigate the context in which SCAs are used by medical laypeople. To reach this aim, different longitudinal mixed models for SCA use are explored and compared.

Generalized Linear Mixed Model

To consider the hierarchical data generalized linear mixed models (GLMMs) with regularization, using a least absolute shrinkage and selection operator (LASSO) penalization, and without regularization were fit to the data and compared. The R package glmmPen [21] was used to conduct the LASSO penalized GLMM, and unpenalized model parameters were derived using the R package lme4 [22]. Further explanation can be found in Multimedia Appendix 1 and Textbox 1.

The model assessment was realized by using the Bayesian information criterion (BIC) scores of the models.

If model 3 or 4 would be identified as best performing they were fit with a GLMM without regularization to derive parameter estimators, CI, and SE.

Textbox 1. Mixed models.

- Model 0: Intercept-only model, with a random intercept for participants
- Model 1: Generalized linear mixed models based on expert knowledge containing only fixed effects for the variables:
  - Day of measurement
  - First-time appearance of symptoms
  - Self-rated health
- Model 2: as model 1 with added random effects for the variables “self-rated health,” “first time appearance of symptoms”
- Model 3: Regularized generalized linear mixed models with least absolute shrinkage and selection operator penalty containing all predictors of Table 1, with independent variance-covariance matrix
- Model 4: as model 3 with unstructured variance-covariance matrix

XSL•FO
RenderX

Table 1. Coefficients, CI, and *P* values for Model 0-3.

Coefficient	Model 0[a]		Model 1[b]		Model 2[c]		Model 3[d]	
	OR (95% CI)	*P* values	OR (95% CI)	*P* values	OR (95% CI)	*P* values	OR (95% CI)	*P* values
Intercept	0.14 (0.12-0.17)	<.001	3.49 (1.76-6.92)	<.001	5.18 (2.41-11.13)	<.001	359.87 (55.38-2338.31)	<.001
Day of measurement	__[e]	—	0.97 (0.96-0.98)	<.001	0.97 (0.96-0.98)	<.001	0.97 (0.96-0.99)	<.001
Self-rated health	—	—	0.44 (0.37-0.52)	<.001	0.40 (0.32-0.49)	<.001	0.80 (0.64-1.01)	<.001
Initial occurrence of Symptoms (yes)	—	—	9.28 (6.06- 14.22)	<.001	13.95 (6.20-31.37)	<.001	5.97 (3.67-9.72)	<.001
General and unspecified (present)	—	—	—	—	—	—	3.33 (1.92-5.56)	<.001
Eye (present)	—	—	—	—	—	—	5.56 (2.08-12.29)	<.001
Cardiovascular (present)	—	—	—	—	—	—	8.33 (2.70-25.00)	<.001
Musculoskeletal (present)	—	—	—	—	—	—	5.26 (3.44-8.33)	<.001
Skin (present)	—	—	—	—	—	—	4.76 (2.04-11.11)	<.001

[a]σ^2=3.29; τ_{00}=0.18$_{ID}$; Bayesian information criterion=1568.02; intraclass correlation coefficient=0.05; N=48 (2016 observations).

[b]σ^2=3.29; τ_{00}=0.26$_{ID}$; Bayesian information criterion=1306.17; intraclass correlation coefficient=0.07; N=48 (2016 observations).

[c]σ^2=3.29; τ_{00}=0.40$_{Intercept}$; τ_{11}=0.03 (participant-rated health) and 3.26 (first-time appearance of symptoms); ρ_{01}=–0.65 and –0.06; Bayesian information criterion=1328.20; intraclass correlation coefficient=0.13; N=48 (2016 observations).

[d]σ^2=3.29; τ_{00}=1.14$_{Intercept}$; τ_{11}=1.53 (general and unspecified); ρ_{01}=–0.89$_{ID}$; Bayesian information criterion=1134.82; intraclass correlation coefficient=0.13; N=48 (2016 observations).

[e]Not applicable.

Missing Data

An Overview of missing data on the outcome and predictor variables can be found in Multimedia Appendix 2. All variables had less than 5% missing data. Missing data on the primary outcome and predictors were imputed using the R package missForest [23], that enables the imputation of missing in mixed data (categorical and continuous), based on a random forest approach [23]. The out-of-bag error (proportion of falsely classified) was used to quantify the test error.

Qualitative Data Analysis

The written comments from the open-ended questions were analyzed with qualitative content analysis [24]. Main categories and subcategories were built inductively after initial familiarization with the data [25]. All transcripts were analyzed by JW using MaxQDA [26]. For quality control, 40% of the codings were independently coded by CP. Afterward, the few conflicting codings were discussed by both coders in data sessions.

Results

Descriptive Data

Of the 48 participants, 31 (71%) were female. The mean age of the participants was 27 (SD 9.1, range 19-64) years. The primary outcome variable showed 264 use cases and 1752 nonuse cases. Imputation of missings resulted in an out-of-bag error of 0.063. Symptoms of the chapter blood had low cell frequencies, the variable “Blood” was therefore excluded as a predictor in the following data analysis for the primary outcome. Tables with frequencies stratified by the outcome of the predictor variables can be found in Multimedia Appendices 3 and 4.

Quantitative Results: Primary Outcome and Model Comparison

All assumptions of GLMM were met. Binned residual Plots for Models 1-3 can be found in Multimedia Appendix 5. The Null model revealed an intraclass correlation of 0.05 and indicated no substantial evidence of clustering between the participants. The BIC of the Intercept-only model was BIC=1568.02. Model 1 derived a BIC of 1305.00. The BIC achieved by Model 2 was higher (BIC=1326.76) compared with the BIC of Model 1. The BIC of Model 3 (BIC=1165.60) revealed the best model performance with the lowest value. Therefore, this model was chosen over Model 4 (BIC=1180.12) and was fit using a GLMM without regularization to derive parameter estimators, CI and SE.

LASSO regularization of Model 3 identified 8 fixed effect predictors, which are the day of measurement, self-rated health, the first-time appearance of symptoms, and 5 ICPC classified symptom clusters (general and unspecified, eye, cardiovascular, skin, and musculoskeletal) as meaningful. All predictors except self-rated health revealed a significant *P*<.001.

An overview of parameter estimations, 95% CI, and the BIC can be found in Table 1.

XSL•FO
RenderX

Qualitative Results

The qualitative content analysis revealed four different main categories: (1) reasons for using SCA (n=79), (2) affective response and evaluation considering SCA usage (n=122), (3) Behavioral response considering SCA usage (n=66), and (4) unmet requirements (n=33).

As users made several entries into the diaries, topics varied between users in general and between the days the entry was made.

Reasons for Using Symptom Checker Apps

Reasons for SCA use emerged as a key topic. Participants identified several reasons for using SCA, including gaining information about medical conditions, symptoms, and treatment options. Participants also used symptom checkers to classify their symptoms and to understand the potential underlying causes of their health issues. The latter is not only to their own symptoms but also to the symptoms of close relatives and friends. Another stated reason for using a SCA was to determine whether a visit to a healthcare professional was necessary or not. Participants also reported using symptom checkers to cross-reference their own assumptions about their health concerns against the information provided by the SCA.

> *For validation: It's probably "really just" migraine & my head isn't about to burst!*

In addition, SCA was used for verifying medical diagnoses and gaining a better understanding of one's health concerns, which was seen as helpful in order to make informed decisions about one's health and well-being. Moreover, participants used already known diseases or conditions to check if the SCA was able to detect them correctly and see how the SCA performed.

> *After the COVID-19 vaccination, I experienced side effects such as fever, headache, and dizziness. I knew that these symptoms were related to the vaccination, but I still wanted to know if the SCA could attribute these symptoms to the vaccination, which it did.*

Affective Response and Evaluation Considering the Usage of Symptom Checker Apps

Participants' affective responses and evaluations following their use of symptom checkers varied widely. While some participants expressed in the respective entries positive or satisfied feelings and evaluations, others reported negative or dissatisfied emotions and evaluations, and some were ambivalent.

> *Woke up with neck pain, according to [the SCA] it could be a tension. [The SCA] recommended heat and stretching, which helped. App usage easy, some questions were alarming.*

Positive evaluations were associated with various facets of SCA usage. These include feelings of satisfaction and contentment, perceiving the information offered by the SCA as helpful and suitable, experiencing a heightened sense of security, achieving a better grasp of one's symptoms, enriching the user's knowledge, and finding the SCA's responses to be logical, realistic, or plausible. Affective responses to SCA use encompass feeling calm, relieved, or more relaxed, feeling sensitized, and motivated to take action. Furthermore, users reported an increased sense of agency and empowerment.

> *I found it very interesting to read about this topic. I notice that the [SCA] is also helpful and useful to me, even for minor physical complaints.*

Participants' negative and dissatisfied emotions and evaluations were associated with several aspects of SCA use. Some participants reported having no trust in the symptom checker, finding the information provided unrealistic or unlikely, and feeling that the SCA worsened their situation if they were already experiencing health anxiety. Others found the information provided to be partially or entirely inaccurate or unhelpful, and some found the recommendations exaggerated or irritating. The negative emotions experienced by participants included disappointment, shock, disillusionment, or feeling unsure, anxious, scared, alarmed, or worried, and some expressed surprise or disappointment in the results provided by the SCA.

> *The [SCA] gives me three possible causes: tonsillitis, peritonsillar abscess, and retropharyngeal abscess. For all three, it recommends visiting the emergency room. I think it's exaggerated to go to the emergency room for a sore throat since I don't have a fever.*

Behavioral Response Considering the Usage of Symptom Checker Apps

Participants reported various behaviors following their use of SCA. In some cases, participants searched the internet for more detailed information, while others preferred to wait and observe their symptoms before deciding whether to seek medical care. Self-treatment strategies were frequently mentioned, such as using over-the-counter medicine, increasing fluid intake, relaxing, or using home remedies.

> *I did not follow the advice of the [SCA] app to seek medical advice as I want to treat it myself with the ointment for a few days first.*

In some cases, participants contacted the health care system and either considered making a doctor's appointment, immediately made a doctor's appointment, or decided to address their symptoms during their next scheduled appointment. However, some participants explicitly stated that they did not plan to discuss their SCA use with a doctor. A few participants made it clear that they chose not to follow the guidance provided by the SCA.

> *[The SCA] advises going to the emergency room, no possibility to see a doctor, was very worried throughout the day, wanted to go directly to the emergency room after the day, but symptoms have subsided a bit, so waited until the next day.*

Unmet Requirements

Participants in this study identified a number of unmet requirements when using the SCA. Some participants expressed the need for a more complex symptom input system that allows for a more nuanced description of symptoms, as well as the ability to differentiate between different qualities and quantities of symptoms.

XSL•FO
RenderX

> *My symptoms were both diverse and unspecific - therefore [SCA] couldn't really make any concrete suggestions - as good as the help texts are, sometimes the questions asked are difficult to answer - I miss [the SCA] not knowing me better from the beginning, so I have fed her with pre-existing conditions, I am curious to see how well Ada learns from that. It's a shame that this cannot yet be taken into account in the suggestions.*

Others reported that their symptoms were not recognized by the SCA and that there was no category for their specific complaint. Participants also expressed concern over the fact that symptoms and diagnoses (eg, bronchitis or cough) can be entered into the SCA. Participants expressed that the weighting of symptoms should be based on the leading symptom entered. Some participants found the questions difficult to answer and reported redundant questions. Participants requested a more differentiated classification of symptoms, as well as the ability to recognize different causes of multiple entered symptoms. Participants evaluated the level of personalization in the symptom checker as insufficient, as preexisting medical conditions, lifestyle, family history, and medication use were not considered.

> *Since I am young, exercise a lot, drink little, do not smoke, and generally live very healthily, I rarely had significant complaints. When something did occur, it was mostly self-inflicted (hangover feeling, sun, sports injuries). I would have wished that not only symptoms but also more lifestyle decisions had been addressed in the assessment. Most often, these explained my symptoms very accurately.*

Some participants also requested the recognition of medication side effects and interactions, as well as consideration of lifestyle factors such as alcohol consumption and sleep deprivation. In addition, participants suggested that the SCA provide recommendations for home remedies and medication use. Participants also requested more educational material that illustrates the relationship between symptoms and specific advice. Finally, participants were concerned about the unclear source of the information provided by the SCA.

Discussion

Principal Results

This mixed methods study aimed to understand the patterns of SCA usage and to describe its use among medical laypersons. The new appearance of unfamiliar symptoms was found to be a strong predictor for SCA usage. In addition, specific ICPC symptom clusters such as “cardio-vascular,” “eye,” “skin,” “musculoskeletal,” and “general and unspecified symptoms” were also associated with a higher probability of SCA use. The qualitative data analysis identified 4 main themes reflecting the aspects that participants found noteworthy after using an SCA. The main themes comprised (1) reasons for using SCA, (2) diverse affective responses, (3) a broad spectrum of behavioral reactions, and (4) unmet needs including a lack of personalization.

Comparison to Previous Work

To evaluate SCA use in a broader context it is worthwhile to reflect on why patients go to see a physician. Braunacker-Mayer and Avery [27] described that the decision to seek help from a physician depends on complex aspects of social, psychological, cultural, and biomedical factors. She described patients as “non-passive users of health care services”; who do not visit doctors without consideration. Further, she claims that patients evaluate their own health condition and then take steps to manage symptoms or treat illnesses. Thus, the decision-making process for seeking medical assistance may not align with clinicians’ perspectives, as it is influenced by a broader range of factors beyond biomedical considerations. With regard to the limited understanding of the SCA usage context, it seems beneficial to explore how the extended version of the Andersen model, a well-established framework in health care usage research, can provide insights into this aspect. The Andersen model [28] incorporates three dimensions to analyze the factors influencing health care usage. These dimensions include predisposing characteristics (such as age and gender), enabling resources (like access to health care facilities and income), and need factors (such as the prevalence of physical illnesses and perceived health status). A further study suggested expanding the model to include psychosocial factors, such as loneliness, and personality traits, like neuroticism [29]. In the following sections, we will examine and classify the identified predictors in the context of SCA usage and health care usage, with qualitative data being integrated subsequently. In the following paragraphs, we use the Andersen model on health care usage to contextualize our results.

Predictor “Initial Occurrence of Symptoms”

In this study, we found that the predictor “initial occurrence of symptoms” had a large effect on SCA use. In line with the Andersen model for health care usage [28] this predictor might represent a need factor. Furthermore, it can be considered as both situational and personal as the “initial occurrence of symptoms” relies on specific symptoms presented by an individual and includes the fact that this individual had experienced similar symptoms before or not. The perception of novel or familiar symptoms relies on prior experiences and the personal capacity to distinguish between different symptoms as well as the overall evaluation of symptoms. It is remarkable that other data have revealed anxiety as a predictor for SCA use in a cross-sectional setting [6]. Health anxiety might have an influence on the appraisal of “initially occurred symptoms” as relevant or irrelevant. Therefore, psychosocial factors or personality traits not only may have an influence on face-to-face health care usage but also on SCA use.

Through qualitative data analysis, we gained deeper insights into the underlying needs related to the predictor “initial occurrence of symptoms.” The category “purpose of use” in the qualitative results shed light on users’ motivations for using SCA. Users reported using these applications to gather information about medical conditions, symptoms, and treatment options. They also wanted to assess their own symptoms or those of close individuals to better understand potential underlying health issues. Another reason for using SCA was to

XSL•FO
RenderX

determine whether a visit to a healthcare professional was necessary. These findings indicate that SCA contributes to the bricolage of health literacy. This perspective defines health literacy as a multidimensional concept, emerging from situational, dynamic, and social practices [30]. Health care bricolage involves the creative assembly and usage of a broad spectrum of resources, including a variety of knowledges, ideas, materials, and networks, to manage specific health concerns [31]. SCA seems to adapt and serve distinct functions, using diverse sources and types of health knowledge, along with the interplay between social connections and situational factors. They can be used by patients as a flexible resource tailored to their personal and situational needs, parallel to other resources.

Predictor "Self-Rated Health"

Compared to previous findings regarding positive associations between health care usage and self-rated health [32], self-rated health was identified as a predictor with a minor negative effect on SCA use. Regarding the association between need factors in the Andersen model [28] of health care usage, substantial evidence demonstrated a positive link between these factors [33]. In particular, self-rated health showed a strong association with health care usage. This finding is highly plausible since an increase in health care needs often indicates illness symptoms that prompt individuals to consult a physician, as supported by numerous cross-sectional studies. This relationship may not apply in the context of SCA use as they are a lower threshold offer and users' needs are therefore less connected to illness symptoms and may be strongly connected to curiosity or insecurities raised by the initial appearance of symptoms and health anxiety and personality traits as mentioned above. In another cross-sectional study, self-rated health was examined in the context of inclination to use SCA, and a positive association was found [6]. Interestingly, no association was found between finding SCA useful and self-rated health [6]. Nevertheless, the recent findings concerning the relationship between SCA and self-rated health remain inconclusive, emphasizing the need for further research considering the inclination and the actual usage of SCA to delve into this aspect more comprehensively.

Predictor "Day of Measurement"

A further predictor with an uncertain role and ambiguous impact was the "day of measurement," which indicated that the timing within the study significantly influenced SCA usage. However, the effect size of this predictor was too small to meaningfully influence the outcome. It should be noted that the large number of measurements may have contributed to the significance of the predictor.

Predictors Classified by International Classification in Primary Care

Other predictors examined were symptoms entered into the SCA categorized according to Symptoms of the categories "cardiovascular," "general and unspecified," "musculoskeletal," "eye," and "skin" were selected as relevant predictors, and all revealed a decent effect size and significant ORs. All named symptom clusters are reported as frequent reasons for emergency consultations in different studies [34-36].

The OR of the predictor "cardiovascular" symptoms was found to be the highest in our study, indicating that the usage of SCA is associated with potentially threatening symptoms. From a qualitative study in our project, it emerged that users would not use SCA in acute emergencies, such as a traumatic brain injury after a traffic accident. However, the findings suggest that users do use SCA for potentially severe and, therefore, anxiety-inducing symptoms. This could be explained by a form of internal triage by medical laypersons. In the case of an obviously nondeferrable emergency, SCA is not used. However, when uncertainties exist, and symptoms are anxiety-inducing but not immediately classifiable as time-sensitive and life-threatening, SCA is used. However, a study from 2022 [35] revealed that SCA does not exhibit a notable performance in such critical scenarios. This should be considered a high-risk situation for patients who may rely on SCA recommendations that are potentially wrong. "General and nonspecific" symptoms often point to a range of possible conditions, posing challenges in their attribution. Consequently, they could serve as meaningful predictors for SCA usage, as users seek additional information to determine the severity and urgency of such symptoms. This need for further clarification underscores the importance of SCA in assisting users to classify and understand their symptoms accurately.

The appearance of symptoms of the cluster "musculoskeletal," "eye," and "skin" frequently leads to a high level of distress as they are restricting in everyday life and therefore might lead to SCA usage. "Musculoskeletal" symptoms lead to movement restrictions and are therefore frequently perceived as constraining in everyday life. A study reports them as the main symptoms of attending an ED [36]. Symptoms concerning the eyes are often perceived as fear-inducing and irritating [37]. Moreover, symptoms concerning the skin are commonly visible and could therefore be stigmatizing [37]. They are also reported as symptoms with a low subjectively perceived treatment urgency that, however, lead patients to seek emergency consultations [34].

The identified ICPC symptom clusters are in accordance with the 10 most common general practitioner consultation reasons reported by the CONTENT project [38]. This may indicate 2 aspects: First, the need for low-threshold medical advice for these symptom clusters; Second SCAs are used for symptoms that may lead to contact with the medical system, and SCAs presumably have an impact on health care professionals and further resources.

Strengths and Limitations

The study investigated the use of a specific SCA over a 42-day period and achieved a 0% dropout rate due to compensation, regular reminders, and user workshops. As a result, the data had less than 5% missing values. Since all participants used the same SCA to ensure comparability, our results may not be generalizable to all available SCAs. Due to the low frequencies of certain ICPC-2 chapters, some predictors were excluded, which may have resulted in some relevant but rarely occurring predictors being left out of the data. Another crucial aspect to consider with the ICPC-2 is its inability to reveal symptom severity or urgency. As a result, the recent findings do not take

XSL•FO
RenderX

these factors into account. It should be noted that LASSO penalization in variable selection comes with certain general limitations. LASSO tends to favor sparse models over complex ones, resulting in models that are more interpretable [39]. However, in cases where there is a high degree of multicollinearity, this preference for sparsity can lead to instability in predictor selection [39]. While we anticipate some multicollinearity among predictors, particularly due to overlaps between ICPC-coded symptoms, overall, multicollinearity is not expected to be a significant issue. LASSO is generally effective at handling moderate multicollinearity [39]. In addition, variable selection in LASSO is influenced by the choice of the shrinking term and penalty [39], which represents an inherent limitation of the method. Nonetheless, compared with other variable selection techniques such as forward or backward selection, LASSO proves to be better suited for the presented data, particularly when considering moderate multicollinearity and the volume of data available.

Moreover, the sample consisted predominantly of females and younger adults, raising uncertainty about its representativeness of the entire user demographic or whether it reflects a sampling bias Nonetheless, this user sample aligns with other findings regarding the sociodemographic characteristics associated with SCA usage [6]. Consequently, additional studies encompassing broader user groups and diverse use cases are imperative to extrapolate the findings more universally.

Conclusion

In this study, SCAs were more frequently used in situations characterized by greater uncertainty for example in situations when symptoms appeared for the first time. This suggests that the concept of intolerance to uncertainty may influence SCA use. Further research is essential to explore the relationships between intolerance of uncertainty, health anxiety, and SCA use as this aspect was not addressed in this study. In addition, the symptom clusters identified in our study had a large overlap with symptoms reported in emergency department visits considering recent study results, indicating that SCA usage could potentially affect health care usage. Further research is needed to ascertain whether such effects exist and how they impact health care usage. Moreover, SCAs are not only used intentionally and purposefully; they are also used playfully, for reasons beyond personal use and as this study showed to validate the user's or doctor’s hypotheses. Furthermore, they are used to verify the accuracy of the software itself, especially when a diagnosis is already known. This reveals that SCAs contribute to a broader bricolage of health literacy, serving as one of many tools integrated into a wider context of sources of information regarding one’s health. Given the current low usage rate of SCAs in Germany, this effect may though be minor at present in the German health care system. Nevertheless, various stakeholders in the German health care system have shown interest in SCAs. Currently, it remains uncertain how this interest will affect SCA usage rates in the future.

Acknowledgments

We thank Johannes A. Vey and Malte Klemmt for their advice. Further, we would like to thank the participants of the cross-sectional survey. We used the generative artificial intelligence tool ChatGPT (OpenAI) to revise the language and style of the manuscript, which were further reviewed and revised by the study group.

This study was funded by the Federal Ministry of Education and Research (BMBF), which is a cabinet-level ministry of Germany (grant 01GP1907A). The funder played no role in the study design, data collection, data analysis, and interpretation, or manuscript writing.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AJW, HH, RK, and SJ designed the study and the study materials. JW collected the data and performed the analyses. HH verified the quantitative data and contributed to the interpretation of the results. CP verified the qualitative data and contributed to the interpretation of the results. AJW wrote the first draft of the manuscript, with critical input from RM, SJ, CP, TH, and HH. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Execution for the generalized linear mixed models.
[DOCX File , 15 KB - jmir_v26i1e55161_app1.docx]

Multimedia Appendix 2
Overview of the missings.
[DOCX File , 19 KB - jmir_v26i1e55161_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Overview of the symptom checker app use or nonuse cases considering potential contributing variables.
[DOCX File , 12 KB - jmir_v26i1e55161_app3.docx]

Multimedia Appendix 4
International Classification in Primary Care 2–coded symptoms stratified for symptom checker app use or nonuse.
[DOCX File , 38 KB - jmir_v26i1e55161_app4.docx]

Multimedia Appendix 5
Binned residual plot for Model 1-3.
[DOCX File , 253 KB - jmir_v26i1e55161_app5.docx]

References

1. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022;5(1):118 [FREE Full text] [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
2. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
3. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022;24(5):e31810 [FREE Full text] [doi: 10.2196/31810] [Medline: 35536633]
4. Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. J Med Internet Res 2021;23(3):e24475 [FREE Full text] [doi: 10.2196/24475] [Medline: 33688845]
5. Abensur Vuillaume L, Turpinier J, Cipolat L, Arnaud-Dépil-Duval, Dumontier T, Peschanski N, et al. Exploratory study: evaluation of a symptom checker effectiveness for providing a diagnosis and evaluating the situation emergency compared to emergency physicians using simulated and standardized patients. PLoS One 2023;18(2):e0277568 [FREE Full text] [doi: 10.1371/journal.pone.0277568] [Medline: 36827277]
6. Kopka M, Scatturin L, Napierala H, Fürstenau D, Feufel MA, Balzer F, et al. Characteristics of users and nonusers of symptom checkers in Germany: cross-sectional survey study. J Med Internet Res 2023;25:e46231 [FREE Full text] [doi: 10.2196/46231] [Medline: 37338970]
7. Aboueid S, Meyer SB, Wallace J, Chaurasia A. Latent classes associated with the intention to use a symptom checker for self-triage. PLoS One 2021;16(11):e0259547 [FREE Full text] [doi: 10.1371/journal.pone.0259547] [Medline: 34731217]
8. Wetzel A, Klemmt M, Müller R, Rieger MA, Joos S, Koch R. Only the anxious ones? Identifying characteristics of symptom checker app users: a cross-sectional survey. BMC Med Inform Decis Mak 2024;24(1):21 [FREE Full text] [doi: 10.1186/s12911-024-02430-5] [Medline: 38262993]
9. Starcevic V, Berle D, Arnáez S. Recent insights into cyberchondria. Curr Psychiatry Rep 2020;22(11):56 [FREE Full text] [doi: 10.1007/s11920-020-01179-8] [Medline: 32852626]
10. Radionova N, Ög E, Wetzel A, Rieger MA, Preiser C. Impacts of symptom checkers for laypersons' self-diagnosis on physicians in primary care: scoping review. J Med Internet Res 2023;25:e39219 [FREE Full text] [doi: 10.2196/39219] [Medline: 37247214]
11. Woodcock C, Mittelstadt B, Busbridge D, Blank G. The impact of explanations on layperson trust in artificial intelligence-driven symptom checker apps: experimental study. J Med Internet Res 2021;23(11):e29386 [FREE Full text] [doi: 10.2196/29386] [Medline: 34730544]
12. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of laypersons' trust in medical decision Aids: randomized controlled trial. JMIR Hum Factors 2022;9(2):e35219 [FREE Full text] [doi: 10.2196/35219] [Medline: 35503248]
13. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne) 2022;9:1040926 [FREE Full text] [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
14. Turner J, Knowles E, Simpson R, Sampson F, Dixon S, Long J, et al. Impact of NHS 111 online on the NHS 111 telephone service and urgent care system: a mixed-methods study. Health Services and Delivery Research 2021;9(21):1-148. [Medline: 34780129]
15. Aboueid S, Meyer S, Wallace JR, Mahajan S, Chaurasia A. Young adults' perspectives on the use of symptom checkers for self-triage and self-diagnosis: qualitative study. JMIR Public Health Surveill 2021;7(1):e22637 [FREE Full text] [doi: 10.2196/22637] [Medline: 33404515]

XSL•FO
RenderX

16. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne) 2022;9:1040926 [FREE Full text] [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
17. Ada. Über uns. 2020. URL: https://ada.com/de/about/ [accessed 2020-12-10]
18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Int J Surg 2014;12(12):1495-1499 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.013] [Medline: 25046131]
19. Wetzel AJ, Koch R, Preiser C, Müller R, Klemmt M, Ranisch R, et al. Ethical, legal, and social implications of symptom checker apps in primary health care (CHECK.APP): protocol for an interdisciplinary mixed methods study. JMIR Res Protoc 2022;11(5):e34026 [FREE Full text] [doi: 10.2196/34026] [Medline: 35576570]
20. Team RC. R: A language and environment for statistical computing. 4.0.5 ed. Vienna, Austria: R Foundation for Statistical Computing; 2021.
21. Heiling HM, Rashid NU, Li Q, Ibrahim JG. glmmPen: high dimensional penalized generalized linear mixed models. R J 2023;15(4):106-128 [FREE Full text] [doi: 10.32614/rj-2023-086] [Medline: 38818017]
22. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J. Stat. Soft 2015;67(1):1-48.
23. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
24. Kuckartz U. Qualitative text analysis. Los Angeles, LA: SAGE Publications; 2013.
25. Kuckartz U. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung (Grundlagentexte Methoden). Germany: Beltz Juventa; 2012.
26. MAXQDA 2020. Berlin, Germany: VERBI Software; 2019.
27. Braunack-Mayer A, Avery JC. Before the consultation: why people do (or do not) go to the doctor. Br J Gen Pract 2009;59(564):478-479. [doi: 10.3399/bjgp09x453495]
28. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 1995;36(1):1-10. [Medline: 7738325]
29. Hajek A, König HH. Beyond symptoms: why do patients see the doctor? BJGP Open 2020;4(2):bjgpopen20X101088 [FREE Full text] [doi: 10.3399/bjgpopen20X101088] [Medline: 32430301]
30. Samerski S. Health literacy as a social practice: Social and empirical dimensions of knowledge on health and healthcare. Soc Sci Med 2019;226:1-8. [doi: 10.1016/j.socscimed.2019.02.024] [Medline: 30831555]
31. Phillimore J, Bradby H, Knecht M, Padilla B, Pemberton S. Bricolage as conceptual tool for understanding access to healthcare in superdiverse populations. Soc Theory Health 2018;17(2):231-252. [doi: 10.1057/s41285-018-0075-4]
32. Xu F, Johnston JM. Self-rated health and health service utilization: a systematic review. Int J Epidemiol 2015;44(suppl_1):i180.
33. Hajek A, Kretzler B, König HH. Determinants of healthcare use based on the andersen model: a systematic review of longitudinal studies. Healthcare (Basel) 2021;9(10):1354 [FREE Full text] [doi: 10.3390/healthcare9101354] [Medline: 34683034]
34. Scherer M, Lühmann D, Kazek A, Hansen H, Schäfer I. Patienten in Notfallambulanzen: Querschnittstudie zur subjektiv empfundenen Behandlungsdringlichkeit und zu den Motiven, die Notfallambulanzen von Krankenhäusern aufzusuchen. Dtsch Arztebl Int 2017;114(39):645-652 [FREE Full text] [doi: 10.3238/arztebl.2017.0645] [Medline: 29034865]
35. Fraser HSF, Cohan G, Koehler C, Anderson J, Lawrence A, Pateña J, et al. Evaluation of diagnostic and triage accuracy and usability of a symptom checker in an emergency department: observational study. JMIR Mhealth Uhealth 2022;10(9):e38364 [FREE Full text] [doi: 10.2196/38364] [Medline: 36121688]
36. Unwin M, Kinsman L, Rigby S. Why are we waiting? Patients' perspectives for accessing emergency department services with non-urgent complaints. Int Emerg Nurs 2016;29:3-8 [FREE Full text] [doi: 10.1016/j.ienj.2016.09.003] [Medline: 27776980]
37. Okosa MC, Uwakwe R, Apakama AI, Onwuegbuna AA, Uzozie CC, Amobi MBC, et al. Psychological effects of eye diseases: a tertiary eye center study. J Psychiatry Psychiatr Disord 2021;05(04):128-139.
38. Laux G, Kühlein T, Gutscher A, Szecsenyi J. Versorgungsforschung in der Hausarztpraxis. München: Urban und Vogel; 2010.
39. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Berlin, Heidelberg: Springer; 2009.

Abbreviations

BIC: Bayesian information criterion
GLMM: Generalized linear mixed model
ICPC: International Classification in Primary Care 2
ISO: International Organization for Standardization
LASSO: least absolute shrinkage and selection operator

XSL•FO
RenderX

SCA: symptom checker apps
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 07.12.23; peer-reviewed by K Savolainen, H Napierala, M Kopka; comments to author 30.01.24; revised version received 13.03.24; accepted 14.10.24; published 09.12.24.

Please cite as:
Wetzel AJ, Preiser C, Müller R, Joos S, Koch R, Henking T, Haumann H
Unveiling Usage Patterns and Explaining Usage of Symptom Checker Apps: Explorative Longitudinal Mixed Methods Study
J Med Internet Res 2024;26:e55161
URL: https://www.jmir.org/2024/1/e55161
doi:10.2196/55161
PMID:

©Anna-Jasmin Wetzel, Christine Preiser, Regina Müller, Stefanie Joos, Roland Koch, Tanja Henking, Hannah Haumann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

General Characteristics and Design Taxonomy of Chatbots for COVID-19: Systematic Review

Wendell Adrian Lim[1], PTRP; Razel Custodio[1], RN, MSc; Monica Sunga[1]; Abegail Jayne Amoranto[1], RN, MSc; Raymond Francis Sarmiento[1], MD

National Telehealth Center, National Institutes of Health, University of the Philippines Manila, Manila, Philippines

Corresponding Author:
Wendell Adrian Lim, PTRP
National Telehealth Center
National Institutes of Health
University of the Philippines Manila
670 Padre Faura Street, Ermita
Manila, 1000
Philippines
Phone: 63 9269819254
Email: wolim@up.edu.ph

Abstract

Background: A conversational agent powered by artificial intelligence, commonly known as a chatbot, is one of the most recent innovations used to provide information and services during the COVID-19 pandemic. However, the multitude of conversational agents explicitly designed during the COVID-19 pandemic calls for characterization and analysis using rigorous technological frameworks and extensive systematic reviews.

Objective: This study aims to describe the general characteristics of COVID-19 chatbots and examine their system designs using a modified adapted design taxonomy framework.

Methods: We conducted a systematic review of the general characteristics and design taxonomy of COVID-19 chatbots, with 56 studies included in the final analysis. This review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to select papers published between March 2020 and April 2022 from various databases and search engines.

Results: Results showed that most studies on COVID-19 chatbot design and development worldwide are implemented in Asia and Europe. Most chatbots are also accessible on websites, internet messaging apps, and Android devices. The COVID-19 chatbots are further classified according to their temporal profiles, appearance, intelligence, interaction, and context for system design trends. From the temporal profile perspective, almost half of the COVID-19 chatbots interact with users for several weeks for >1 time and can remember information from previous user interactions. From the appearance perspective, most COVID-19 chatbots assume the expert role, are task oriented, and have no visual or avatar representation. From the intelligence perspective, almost half of the COVID-19 chatbots are artificially intelligent and can respond to textual inputs and a set of rules. In addition, more than half of these chatbots operate on a structured flow and do not portray any socioemotional behavior. Most chatbots can also process external data and broadcast resources. Regarding their interaction with users, most COVID-19 chatbots are adaptive, can communicate through text, can react to user input, are not gamified, and do not require additional human support. From the context perspective, all COVID-19 chatbots are goal oriented, although most fall under the health care application domain and are designed to provide information to the user.

Conclusions: The conceptualization, development, implementation, and use of COVID-19 chatbots emerged to mitigate the effects of a global pandemic in societies worldwide. This study summarized the current system design trends of COVID-19 chatbots based on 5 design perspectives, which may help developers conveniently choose a future-proof chatbot archetype that will meet the needs of the public in the face of growing demand for a better pandemic response.

(J Med Internet Res 2024;26:e43112) doi:10.2196/43112

KEYWORDS

COVID-19; health chatbot; conversational agent in health care; artificial intelligence; systematic review; mobile phone

XSL•FO
RenderX

Introduction

Background

The COVID-19 pandemic is regarded as one of the most critical public health crises to have affected societies worldwide [1,2]. In December 2019, the first human cases of COVID-19 were reported in Wuhan, China [3]. A few months later, in March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a global pandemic [4]. Since then, it has caused massive disruptions and catastrophes not only in the health care sector but also in education, transportation, the economy, and politics [1,5]. During this period, we observed a paradigm shift toward digital transformation and a surge in the use of web-based technologies to ensure safe access and provision of needed services [6,7]. The worldwide effects of the COVID-19 pandemic, especially in the health sector, have caused health authorities such as the WHO to turn to digital technologies, such as the use of chatbots [8]. Early in the pandemic, chatbots were considered an instant resource about the latest COVID-19 health advice to the public [9]. Chatbots are conversational agents powered by artificial intelligence (AI) that enable interaction between end users and machines, with the intervention of natural language processing (NLP) [6,7,10,11]. Currently, chatbots are already in use to tackle various health concerns, particularly in the realm of mental health, which has the most extensive literature available [8]. Chatbots can also currently perform many labor-intensive tasks at lower costs across other fields, such as customer service, pedagogy, and personal assistance [12]. During a pandemic, however, chatbots can be particularly useful for informing people, promoting behavior change, offering mental health assistance, and diagnosing and tracking symptoms [8]. Therefore, a surge in the use of COVID-19–related chatbots occurred around the world. Many of these COVID-19 chatbots were implemented for public health responses and information sharing, such as the WHO Health Alert via WhatsApp and Facebook [12] and the HealthBuddy of the WHO Regional Office for Europe [13]. Other uses of chatbots at the time of the COVID-19 pandemic were in the field of education, such as Iggy and Pounce [14]; business and employment, such as Benefits Bot [15]; mental health, such as Woebot [16]; and customer service, such as the Rapid Response Virtual Agent [17].

The existing literature investigated the design characteristics of chatbots in general, including their frequency of interaction and consequential design implications [18-20]. Notably, the design of a chatbot could affect user engagement, which in turn could affect its usefulness to its intended audience [18]. It is also critical to note that the design of chatbots evolves with time to perform new functions [19]. Despite this, we noticed a lack of comprehensive characterization and analysis of the chatbots explicitly designed to address the negative effects of the COVID-19 pandemic. Prior chatbot reviews mainly covered conversational agents developed before the pandemic [4,11,21]. Given the recent surge in COVID-19–related chatbot development and deployment, we believe that an up-to-date review that evaluates the characteristics and design of chatbots deployed in response to the pandemic is crucial. Analyzing the design of these technologies and their innovative applications can track recent developments in chatbot technology and assist chatbot developers and implementers in making design decisions for their conversational agents.

Objectives

To address this gap, we aimed to determine the general characteristics of the existing COVID-19 chatbots worldwide. We also intended to examine the system blueprints of these conversational agents using the adapted and modified design taxonomy framework by Nißen et al [21].

This chatbot design framework was based on a mixed methods research study, in which 120 pre–COVID-19 chatbots with various purposes were analyzed to develop a design taxonomy. The framework characterized relationships among users and chatbots with different time horizons based on 5 design perspectives (eg, *temporal profile*, *appearance*, *intelligence*, *interaction*, and *context)*, 22 design dimensions, and 61 design characteristics [21] (Table 1). In the interest of our work, we did not cover the *front-end user interface*, which is 1 of the 22 design dimensions in the design taxonomy by Nißen et al [21]. The *front-end user interface* dimension addresses user access, which can already be covered in the description of the chatbot's general characteristics. Using this chatbot design framework, we will be able to see the design trend of COVID-19 chatbots.

XSL•FO
RenderX

Table 1. Modified design taxonomy for COVID-19 chatbots.

Layer, design perspective, and design dimensions	Design characteristics
Chatbot	
Temporal profile	
Time horizon	Short term, medium term, long term, and life long
Frequency of interactions	1-time only and multiple times
Duration of interaction	Short, medium, and long
Consecutiveness of interactions	Unrelated and related
Appearance	
Role	Expert, facilitator, and peer
Communication style	Task oriented, and socially or chat oriented
Avatar representation	Disembodied and embodied
Intelligence	
Intelligence framework	Rule-based, hybrid, and artificially intelligent
IQ	Rule-based knowledge only, text understanding, and other alternative sources
Personality adaptability	Principal self and adaptive self
Socioemotional behavior	Not present and present
Service integration	None, external data, media resources, and multiple
Chatbot and user	
Interaction	
Communication modality	Text only and text+voice
Interaction modality	Graphical and interactive
User assistance design	Reactive, proactive, and reciprocal
Personalization	Static and adaptive
Additional human support	None and yes
Gamification	Not gamified and gamified
User	
Context	
Application domain	Business, health care, education, and daily life
Motivation or purpose	Productivity, entertainment, utility, informational, and coaching
Collaboration goal	Nongoal oriented and goal oriented

Methods

Search Strategies

We compiled published papers and journals on COVID-19 chatbots, targeting the social facets of the pandemic response, such as health, education, and lifestyle management, as described in our objectives. We examined databases such as PubMed, Scopus, CINAHL, ProQuest, as well as Google Scholar and ResearchGate to help us capture the gray literature. We also used the following search string for this systematic review: (“Covid” OR “Covid-19” OR “Covid19” OR “coronavirus” OR “2019nCoV” OR “SARS-COV-2”) AND (“Chatbot” OR “conversational agent” OR “virtual assistant”).

Study Selection

We used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for this systematic review (Multimedia Appendix 1). The results of our study selection are shown in Figure 1. After eliminating duplicates and screening titles and abstracts, we developed a set of inclusion and exclusion criteria to determine the most relevant articles. We included studies that matched the following statements: (1) papers that extensively discussed the characteristics, design, and behavior of a COVID-19–specific chatbot, not limited to health; (2) papers that described a chatbot that underwent a clinical trial during the time of the pandemic; and (3) papers that detailed the development and conceptualization of a COVID-19 chatbot. Alternatively, we excluded publications that consisted of the following: (1) studies that were not written in English; (2) studies without full texts;

XSL•FO
RenderX

(3) studies that were published before March 2020 and after April 30, 2022; (4) qualitative studies related to COVID-19 chatbots or conversational agents; (5) studies that mentioned chatbots related to other public health emergencies of international concern such as swine influenza, Ebola, polio, Zika virus, and Kivu Ebola; (6) other systematic reviews or meta-analyses about COVID-19 chatbots; and (7) opinion papers, commentary, news, and editorials on COVID-19 chatbots. The papers were concurrently scrutinized through intensive reading and analysis by 2 research associates and 1 senior researcher from April to June 2022. Papers published after April 30, 2022, were not included to not disrupt the paper analysis and statistics retrieved by the researchers.

Figure 1. Process flowchart of study selection. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for new systematic reviews which included searches of databases, registers, and other sources.

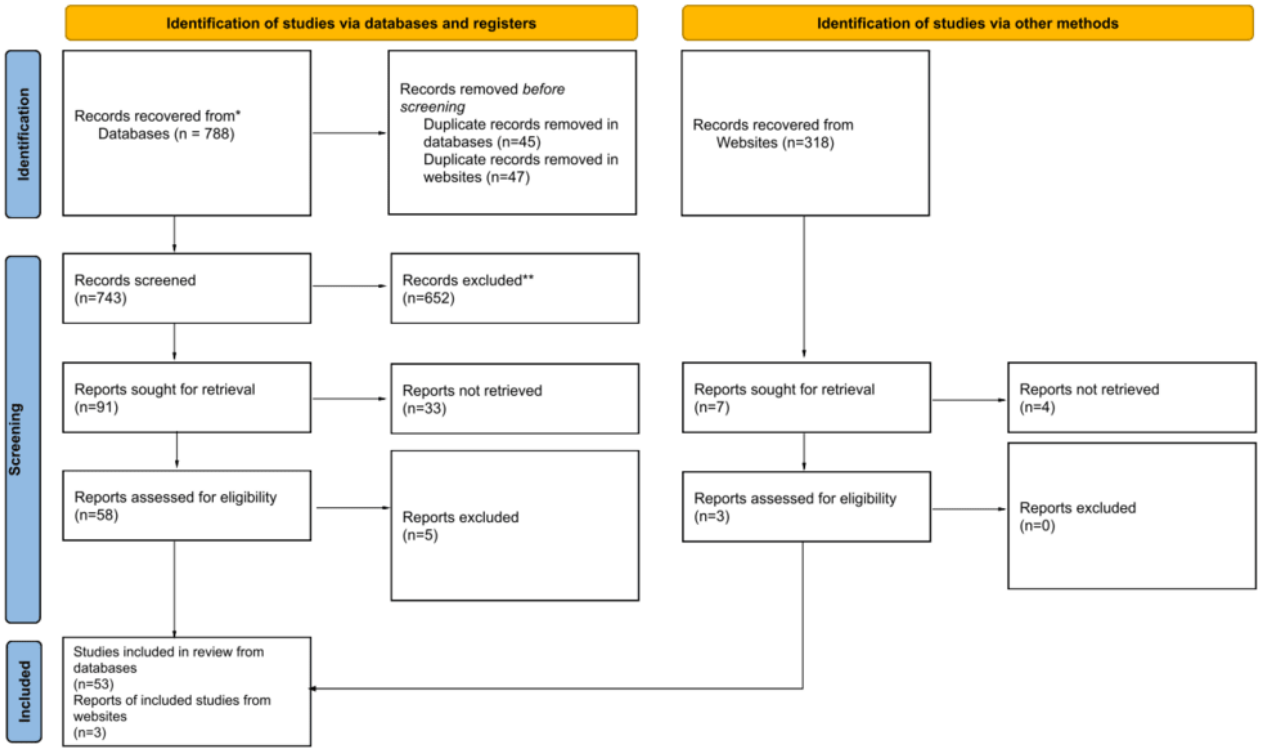

Results

Overview

We identified 1106 papers from the databases (n=788, 71.25%) and websites (n=318, 28.75%). We then removed 92 duplicates (n=45, 49% in databases and n=47, 51% in websites) before the study screening and performed title and abstract screening. From the remaining 1014 papers in the databases and websites, we sought 98 (9.66%) for full-text retrieval. Then, of the 98 papers, we reviewed the full text of 58 (59%) from database sources and 3 (3%) from websites using the study inclusion criteria. After the final screening, we included 56 papers in this study.

General Characteristics

The general characteristics of a COVID-19 chatbot describe its structural and technical information. These comprise the app's name, country and continent of origin, maturity level, and the platform where it was deployed. Details of the general characteristics of COVID-19 chatbots are provided in Multimedia Appendix 2 [10,22-76]. A summary is presented in Table 2.

XSL•FO
RenderX

Table 2. Summary of the general characteristics of the COVID-19 chatbots (N=56).

Characteristic and category	Values, n (%)
Region or continent	
Asia	24 (43)
Europe	15 (27)
North America	7 (13)
Africa	4 (7)
South America	3 (5)
Oceania	3 (5)
Maturity level	
Released	26 (46)
Beta stage	11 (20)
Conceptualization	10 (18)
Alpha stage	5 (9)
Prealpha stage	2 (4)
Prototype	2 (4)
Platform	
SMS text messaging apps	17 (30)
Websites	16 (29)
Android operating system	12 (21)
Others (eg, iOS and Google cloud)	9 (16)

Design Taxonomy

Temporal Profile Perspective

Table 3 shows the distribution of studies in relation to their temporal profiles, which refer to the time-dependent characteristics of a chatbot and can be further classified according to the time horizon, frequency of interaction, duration of the exchange, and consecutiveness of interactions [21]. Our results indicated that a substantial proportion of COVID-19 chatbot interactions (26/56, 46%) is categorized under the medium term, as the interaction of these chatbots can go beyond a single conversation and may last for days and weeks. With regard to frequency of interactions, more than half of the COVID-19 chatbots (31/56, 55%) interacted with their users multiple times instead of a single occurrence [21]. Regarding interaction duration, most COVID-19 chatbots had medium-length conversations with the user (22/56, 39%).

Regarding the consecutiveness of interactions, nearly half of the chatbots (25/56, 45%) could remember information from previous user interactions, whereas the remaining chatbots (8/56, 14%) were found to perform the opposite function toward past interactions with the user.

XSL•FO
RenderX

Table 3. Temporal profile perspective of the COVID-19 chatbots (N=56).

Design dimension and design characteristic	Values, n (%)	Study
Time horizon		
Short term	11 (20)	[28,34,40,41,45,47-51,59]
Medium term	26 (46)	[31,33,36,39,42-44,46,54,56-58,60,63,64,66-76]
Long term	2 (4)	[35,38]
Not stated	17 (30)	[10,22-27,29,30,32,37,52,53,55,61,62,65]
Frequency of interaction		
1 time only	8 (14)	[28,34,40,41,45,47,48,59]
Multiple times	31 (55)	[26,31,33,35,36,38,39,42,44,46,49,54,56,57,60-76]
Not stated	17 (30)	[10,22-25,27,29,30,32,37,43,50-53,55,58]
Duration of interaction		
Short	14 (25)	[28,34,40-42,45-51,57,59]
Medium	22 (39)	[31,33,35,36,39,44,54,56,60,63,64,66-76]
Not stated	20 (36)	[10,22-27,29,30,32,37,38,43,52,53,55,58,61,62,65]
Consecutiveness of interactions		
Related	25 (45)	[26,31,33,35,36,39,48,49,54,60-68,70-76]
Unrelated	8 (14)	[28,34,40,41,43,45,46,69]
Not stated	23 (41)	[10,22-25,27,29,30,32,37,38,42,44,47,50-53,55-59]

Appearance Perspective

In Table 4, we characterize the appearance of COVID-19 chatbots (ie, the chatbot's overall look and feel) into 3 design dimensions: *role*, *primary communication style*, and *avatar representation* [21]. Regarding chatbot roles, our results showed that most conversational agents for COVID-19 were considered *experts* (33/56, 59%). Expert chatbots have specific skills or expertise in a specific field and provide answers or reactions to the user's query, keywords, or actions [21]. In contrast, facilitator chatbots assist users in reaching a certain objective or executing a particular task [21]. In our analysis, 27% (15/56) were facilitator COVID-19 chatbots. Finally, peer chatbots aim to be a "sparring partner" for the user, and they value socioemotional behavior as part of their design. Of the 56 chatbots, only 8 (14%) COVID-19 chatbots served as peers.

Regarding the communication style of COVID-19 chatbots, we found that 63% (35/56) of them were task oriented. Task-oriented chatbots prioritize task efficiency and achievement of goals at the most minimal cost and time possible [21]. In contrast, socially oriented chatbots prioritize social engagement and personalization of interactions with their users [21]. These comprised only 34% (19/56) of COVID-19 chatbots.

Chatbots can also have a visual or character representation (embodied avatars), or they may not have one at all (disembodied avatars) [21]. We found that more than half of the COVID-19 chatbots (32/56, 57%) did not have any visual or character representation. Only 34% (19/56) of the COVID-19 chatbots had embodied avatars.

XSL•FO
RenderX

Table 4. Appearance perspective of the COVID-19 chatbots (N=56).

Design dimension and design characteristic	Values, n (%)	Study
Role		
Expert	33 (59)	[10,22-24,26,29-34,38,39,41,43,47,50,51,55-58,63,64,66-73,75,76]
Facilitator	15 (27)	[25,27,35-37,40,45,53,54,59-62,65,72]
Peer	8 (14)	[28,42,44,46,48,49,52,74]
Primary communication style		
Task oriented	35 (63)	[23-26,29,30,32-34,37,39,40,43,45,46,54-69,71,72,75,76]
Socially or chat oriented	19 (34)	[10,22,27,28,35,36,38,41,42,44,47-52,70,73,74]
Not specified	2 (4)	[31,53]
Avatar representation		
Disembodied	32 (57)	[10,23-26,29,30,32-34,36,37,39-41,52,56,60,61,63-75]
Embodied	19 (34)	[22,27,28,35,38,42,44-51,53,54,57,62,76]
Not specified	5 (9)	[31,43,55,58,59]

Intelligence Perspective

The intelligence perspective is characterized by the functionalities of a chatbot's inner working mechanisms, which build on an intelligence framework and quotient, personality adaptability, socioemotional behavior, and service integration [21]. As depicted in Table 5, we found that a large number of COVID-19 chatbots (26/56, 46%) were artificially intelligent. Artificially intelligent chatbots use their NLP capabilities to learn from each conversation with the user [21]. Meanwhile, a high proportion of COVID-19 chatbots (25/56, 45%) could respond to users according to set rules and by understanding an individual's textual input.

With regard to personality adaptability, more than half of the COVID-19 chatbots (31/56, 55%) have a *principal* self-adaptability, meaning they operate on a structured flow. In contrast, a significant number of COVID-19 chatbots (24/56, 43%) were *adaptive* because these chatbots adjust their responses depending on the user input [21].

Moreover, 57% (32/56) of the COVID-19 chatbots did not demonstrate socioemotional behavior toward their clients. Among the existing COVID-19 chatbots, only 36% (20/56) could express empathic reactions to the users.

With regard to service integration, there was an equal distribution of the number of COVID-19 chatbots that were limited to only processing external data and those that could process both external data and broadcast resources (21/56, 38%).

XSL•FO
RenderX

Table 5. Intelligence perspective of the COVID-19 chatbots (N=56).

Design dimension and design characteristics	Values, n (%)	Study
Intelligence framework		
Rule based	14 (25)	[25,27,30,32,36,37,40,46,60,61,65,69,72,74]
Hybrid	14 (25)	[10,22,54-57,63,66-68,70,71,73,76]
Artificially intelligent	26 (46)	[23,24,26,28,29,33-35,38,39,41-45,47-53,59,62,64,75]
Not stated	2 (4)	[31,58]
IQ		
Rule-based knowledge only	14 (25)	[25,27,30,32,34,36,40,41,43,46,65,68,69,72]
Text understanding	14 (25)	[10,22-24,26,28,29,33,35,37-39,42,71]
Both	25 (45)	[44,45,47-56,59-64,66,67,70,73-76]
Not stated	3 (5)	[31,57,58]
Personality adaptability		
Principal self	31 (55)	[22-27,29,30,32,34,37,39-41,43-46,53-61,65,68,69,72]
Adaptive self	24 (43)	[10,28,33,35,36,38,42,47-52,62-64,66,67,70,71,73-76]
Not stated	1 (2)	[31]
Socioemotional behavior		
Not present	32 (57)	[10,23-26,29,30,32-34,37,39-41,43,45,55,59-69,71,72,75,76]
Present	20 (36)	[22,27,28,35,36,38,42,44,46,48-52,54,57,58,70,73,74]
Not stated	4 (7)	[31,47,53,56]
Service integration		
None	2 (4)	[55,61]
External data	21 (38)	[10,23-27,30-34,39,40,45,60,65,71,72,74-76]
Media resource	3 (5)	[41,46,57]
Multiple	21 (38)	[22,28,29,35-38,48,50,51,54,59,62-64,66-70,73]
Not stated	9 (16)	[42-44,47,49,52,53,56,58]

Interaction Perspective

The domain encompasses the design dimensions of the users' communication with the COVID-19 chatbot. As depicted in Table 6, of the 56 COVID-19 chatbots, 38 (68%) were text based, where the end user and the bot interacted and communicated through text or internet messaging. The interactive modality was the most predominant (39/56, 70%), where end users can enter free texts to communicate with a chatbot [21].

Regarding user assistance design, more than half of the COVID-19 chatbots (31/59, 54%) were reactive, meaning they only reacted to user input. Concerning personalization, most COVID-19 chatbots (36/56, 64%) were adaptive, meaning they could change their meaning and adjust their responses based on user needs and circumstances [21]. Moreover, most COVID-19 chatbots did not receive any additional human support (44/56, 79%) or have any gaming elements (42/56, 75%).

XSL•FO
RenderX

Table 6. Interaction perspective of the COVID-19 chatbots (N=56).

Design dimension and design characteristics	Values, n (%)	Study
Communication modality		
Text only	38 (68)	[22,23,25,26,29,30,32-41,43,45-48,51,54,55,57,60-66,68,69,71,72,74,75]
Text and voice	16 (29)	[10,24,27,28,42,44,49,50,52,53,56,59,67,70,73,76]
Not specified	2 (4)	[31,58]
Interaction modality		
Graphical	13 (23)	[22,25,27,32,36,40,46,57,62,65,68,69,72]
Interactive	39 (70)	[10,23,24,26,28-30,33-35,37-39,41,42,44,45,47-54,56,59-61,63,64,66,67,70,71,73-76]
Not specified	4 (7)	[31,43,55,58]
User assistance design		
Reactive	30 (54)	[10,23-30,32-37,40,45,46,52,53,55,56,58-61,65,68,69,72]
Proactive	1 (2)	[71]
Reciprocal	23 (41)	[22,38,39,41-44,47-51,54,62-64,66,67,70,73-76]
Not specified	2 (4)	[31,57]
Personalization		
Static	19 (34)	[23-26,30,34,36,37,40,46,55,57,58,60,61,65,68,69,72]
Adaptive	36 (64)	[10,22,27-29,32,33,35,38,39,41-45,47-54,56,59,62-64,66,67,70,71,73-76]
Not specified	1 (2)	[31]
Additional human support		
Yes	10 (18)	[10,39,43,53,58,62,67,68,72,73]
No	44 (79)	[22-30,32-38,40-42,44-52,54-57,60,61,63-66,69-71,74-76]
Not specified	2 (4)	[31,59]
Gamification		
Gamified	6 (11)	[42,47,48,54,57,73]
Not gamified	42 (75)	[10,22-30,32-41,43-46,53,55,56,60,62-72,74-76]
Not specified	8 (14)	[31,49-52,58,59,61]

Context Perspective

Context refers to the motivation of users to interact with the chatbot, and it is reflected in the application domain, the user motivation to engage with the chatbot, and the collaboration goal of the chatbot. As depicted in Table 7, the vast majority of COVID-19 chatbots were in the health care domain (46/56, 82%). Consequently, we found 20% (11/56) of the COVID-19 chatbots in the education domain, and there was a smaller percentage of chatbots designed activities of daily living (6/56, 11%). Among all the COVID-19 chatbots, 11% (6/56) fell under multiple application domains.

We also found that most COVID-19 chatbots (35/56, 63%) were purposely designed to provide information to the user. Our results also showed that there was an almost equal distribution of COVID-19 chatbots designed for utility (18/56, 32%) and coaching purposes (15/56, 27%). However, some COVID-19 chatbot researchers designed their conversational agents for productivity (6/56, 11%) and entertainment (4/56, 7%) purposes. It is also noteworthy that several COVID-19 chatbots (16/56, 29%) were designed for multiple purposes.

XSL•FO
RenderX

Table 7. Context perspective of the COVID-19 chatbots (N=56).

Design dimension and design characteristic	Values, n (%)	Study
Application domain[a]		
Health care	46 (82)	[10,22-29,31-46,48,50,51,55,58,59,61,63-76]
Education	11 (20)	[27,30,44,47,48,53,56,57,60,62,65]
Daily life	6 (11)	[42,44,46,49,52,54]
Motivation or purpose[b]		
Productivity	6 (11)	[48,54,60,62,65,75]
Entertainment	4 (7)	[44,49,54,57]
Utility	18 (32)	[10,22,25,31,33,34,37-40,48,54,59,61,62,67,68,76]
Informational	35 (63)	[1-3,6,7,10,12-19,21-68,73,75-77]
Coaching	15 (27)	[23,27,35,36,42,44,46,52,54,57,61,70,72-74]
Collaboration goal		
Goal oriented	56 (100)	[10,22-76]

[a]In total, 11% (6/56) of the chatbots fall under multiple application domains.

[b]In total, 29% (16/56) of the chatbots are designed to for multiple purposes.

Finally, all the COVID-19 chatbots were goal oriented, which means that the chatbots helped users achieve a predefined goal [78].

Discussion

Overview

Chatbots have a huge impact on augmenting the public health response worldwide because of their wide accessibility, functionality, practicality, and fast dissemination of information in addressing the debilitating effects of COVID-19 [79]. This systematic review successfully classified 56 COVID-19 chatbots worldwide based on their general characteristics and systems design, which was based on the design taxonomy by Nißen et al [21]. After further analysis of the collated COVID-19 chatbots, our research yielded the following findings.

General Characteristics

Regarding the general characteristics of the existing COVID-19 chatbots, we found that most studies on COVID-19 chatbot design and development were conducted in Asia and Europe. This indicates the need to conduct and implement more studies on COVID-19 chatbot research in other continents. Our research also revealed that most chatbots were accessible on websites, internet messaging apps (eg, Facebook Messenger, Viber, Telegram, WeChat, and LINE), and the Android operating system. This implies the ease of development and accessibility of web-based and internet messaging app chatbots as well as the popularity of social media in this generation [76]. Future research should focus on the design and development of pandemic-related chatbots for less popular native platforms such as iOS and Google Cloud to cater to a wider variety of people. Studies on the evaluation of the effectiveness of such chatbots in mitigating the negative effects of a pandemic should also be conducted to enrich the body of knowledge on chatbot designs.

Design Taxonomy

Overview

A recent scoping review pooled the existing conversational agents combating the negative effects of COVID-19 and classified them based on their applications [6]. In this study, we updated the number of existing COVID-19 chatbots and supplemented the current literature with the design trends of existing COVID-19 conversational agents. We adopted and modified the chatbot design taxonomy of Nißen et al [21] and classified each COVID-19 chatbot based on 5 design perspectives.

Temporal Profile Perspective

From the temporal profile perspective, nearly half of the COVID-19 chatbots (26/56, 46%) interacted with the user for several weeks (medium term). This implies that most COVID-19 chatbots did not establish a deep, long-lasting relationship and companionship with the user, which is observed in chatbots with a long-term time horizon [21]. *COOPERA* (COVID-19: Operation for Personalized Empowerment to Render Smart Prevention and Care Seeking), for example, provides individualized self-care support, real-time follow-up of the users' health condition, and the epidemiological situation in the users' area for up to several weeks [31].

Moreover, COVID-19 chatbots tended to interact with their users for >1 time, which was seen in 55% (31/56) of these chatbots. For instance, Indonesian researchers deployed *ARIGAselfCareNursingBot*, a teleassessment chatbot, for 5 months in a wound care clinic [56]. Furthermore, 45% (25/56) of COVID-19 chatbots could remember information from previous user interactions. *COOPERA*, as discussed previously, could monitor users' health conditions daily based on their previous answers about their physical and health conditions [31].

XSL•FO
RenderX

However, it is notable that a significant proportion of COVID-19 chatbot papers did not explicitly report the time horizon (17/56, 30%), frequency (17/56, 30%), duration (20/56, 36%), and consecutiveness of interactions (23/56, 41%) of their chatbot. This may show that there is a need for researchers and computer programmers to have a deeper dive into chatbot design specifications in their respective manuscripts to help future research in the identification of the most cost-efficient design for a pandemic-responsive chatbot.

Appearance Perspective

From the appearance perspective, we found that most COVID-19 chatbots (33/56, 59%) were considered experts. For example, *Chloe* provides Canadians with up-to-date information regarding COVID-19 by answering users' inquiries about the pandemic [22]. In contrast, only 27% (15/56) of the COVID-19 chatbots assumed a facilitator role. An example of this is *Atena*, a chatbot that delivers digital mental health intervention and suggests effective coping strategies to improve the mental well-being of university students [36]. A screenshot showing a sample of conversation from the *Atena* chatbot is shown in Figure 2 [36].

Figure 2. Screenshot of the Atena chatbot.

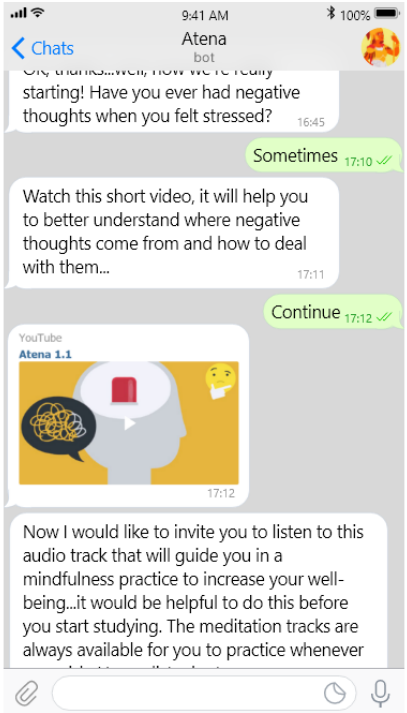

Our research also found that most COVID-19 chatbots (35/56, 63%) were task oriented in terms of primary communication style. To illustrate, *Symptoma* assesses the end user's risk of COVID-19 based on the symptoms the person inputted or answered in the textbox, without considering the user's emotional needs [34]. This is in contrast with socially oriented chatbots, which represent only 34% (19/56) of the literature examined. An example is *Bella*, which demonstrates empathy and emotional expressions based on the user's inputs to personalize every interaction [28].

Finally, our analysis showed that more than half of the existing COVID-19 chatbots (32/56, 57%) were disembodied, meaning they did not have any character representation. Chatbots with avatars could strengthen their social presence, which may reinforce a good social relationship with the users [79]. This finding may suggest that the immediate priority of designers at the height of the COVID-19 pandemic is to fulfill their functional efficiency while deprioritizing social engagement that may lengthen development cycles.

Intelligence Perspective

From the intelligence perspective, a significant number of COVID-19 chatbots (26/56, 46%) were artificially intelligent. One example is *Drever*, an educational assistant developed in Japan using an AI engine that acts as a teacher capable of answering questions from students [62]. Similarly, Real Conversation is an AI-based chatbot in India that provides COVID-19 information while assisting with customer queries [64].

Meanwhile, a high number of COVID-19 chatbots (25/56, 45%) could respond to users according to a set of rules and by understanding the textual input of the user. For example, *CoronaGo* is a health care chatbot embedded in a website that uses NLP to understand user intention but also has a rule-based system architecture to track the COVID-19 situation in India efficiently [66]. This means that most COVID-19 chatbots maximize their capabilities in understanding the users' intentions in a timely manner, which is critical for a chatbot designed for a pandemic.

Most COVID-19 chatbots (31/56, 55%) operate on a structured flow, for example, *Chatbot-Mental Wellness*, a chatbot developed in India that provides guided activities in simple exercises, meditation, and positive affirmation [46], although nearly half (24/56, 43%) of the chatbots can adjust their responses depending on user input. An example is the

XSL•FO
RenderX

JuggleChat, an educational assistant chatbot that addresses students' specific needs dependent on their input [47].

Our analysis also showed that most COVID-19 chatbots (32/56, 57%) did not present socioemotional behavior to the users, whereas only 36% (20/56) of COVID-19 chatbots could express empathic reactions to the consumers. This suggests that expectations for chatbots to exhibit social behaviors that are customary in human-to-human communication should always be considered as there is a growing demand for humanlike conversational engagement styles [80]. Some empathic chatbots include *Chasey*, a character-based conversation agent for COVID-19, who responds to users with phrases such as "Don't worry my friend, everything will be fine." and "Get well soon, pal!" [51].

Finally, COVID-19 chatbots that are limited to only processing external data and those that can process both external data and media resources were of equal distribution (21/56, 38%). Such chatbots include *Mano*, an interactive chatbot that provides up-to-date COVID-19 information from reliable sources including, but not limited to, the WHO and predicts pandemic effects in the Mano River Union region [76], and *Aapka Chikitsak*, a medical bot that provides its users with quality information from the National Health Portal and engaging media as a means to deliver telehealth services in India during the pandemic [70].

Interaction Perspective

From the interaction perspective, most COVID-19 chatbots (38/56, 68%) reach out and communicate with the user through text or internet messaging and even allow the user to respond via free text inputs. This is essential as limiting interaction options may indicate a disregard for the users' inputs [81]. One example is *Elena+*, a smartphone-based and conversational agent that delivers pandemic lifestyle care interventions that communicate with the user through textual input and is currently being expanded into Spanish and Tamil languages to reach larger audiences [35]. Another example is *Wakamola*, a chatbot that recognizes unstructured messages and is used to encourage confined individuals to modify their lifestyle practices [37].

In addition, more than half of the COVID-19 chatbots (30/56, 54%) reacted only to user input rather than steering the conversation. This may imply that chatbot designers should aim to design constructive, goal-oriented interactions [82]. Similar to a proof-of-concept study implemented at the University of Trento in Italy, *Atena* was able to react to user inputs only [36], unlike the lone proactive *Alia* that can steer conversation [71].

Moreover, most COVID-19 chatbots (36/56, 64%) had adaptive personalization. This shows that pandemic designers concentrated on the chatbots' capacity to adapt to users' choices, needs, and circumstances, which is consistent with an earlier study [81]. *Cory COVID-Bot*, for instance, which was developed to engage with difficult-to-reach populations directly, aims to resolve public ambiguity and uncertainty regarding community health guidelines. With its sophisticated features, an algorithm dynamically determines which questions to ask a user and when to solve them, encompassing recommendations on COVID-19 safety and behavior change interventions [38].

Furthermore, the vast majority of COVID-19 chatbots did not have any additional human support nor gamification (44/56, 79% and 42/56, 75%, respectively). These design elements for chatbot interaction show that COVID-19 chatbots are designed in such a way that most COVID-19 chatbots straightforwardly interact with the user and identify their intentions as soon as possible. Previous research suggests that conversational agents should possess a humanlike quality or assistive nature, particularly within telemedicine platforms [82]. In addition, incorporating games can help evoke personality traits and enhance thoroughness [83]. Notably, a proposed conversational agent from a university hospital in the United Kingdom targets these 2 design characteristics [73]. The in-app features will be gamified as a motivational strategy for older people in self-isolation to alleviate problems with mental health. Moreover, it is planned that this chatbot will have a social support function to facilitate cognitive behavioral therapy sessions.

Context Perspective

From the context perspective, almost all COVID-19 chatbots (46/56, 82%) fell under the health care application domain. We expected this because COVID-19 is a public health concern. Some of these chatbots were used for personal risk assessment [10], monitoring of COVID-19 exposure [22], and dissemination of health information regarding the pandemic [26].

Aside from health care, 20% (11/56) of COVID-19 chatbots were used in the education domain. For instance, *Student Interactive Assistant Android Application with Chatbot* is an academic chatbot developed during the COVID-19 pandemic that assists students interactively concerning their academic requirements [30]. Only a minority of COVID-19 chatbots (6/56, 11%) were used for other purposes such as assisting with daily life. In Morocco, for instance, a chatbot is being developed with the intention of aiding students daily with their mental health [52].

Our results also showed that most COVID-19 chatbots (35/56, 63%) were designed to provide information to the users. For example, a chatbot developed in France informs and educates people about COVID-19 vaccines and reduces vaccine hesitancy [32]. There were also COVID-19 chatbots designed for utility and coaching purposes, but these types comprised only 32% (18/56) and 27% (15/56) of COVID-19 chatbots, respectively. For example, a chatbot developed by the University of California was used as a COVID-19 screening tool for health employees entering their hospital before their clinical shift [40]. Regarding coaching, Managing Stress at University instructs Indian University students on stress management during the COVID-19 pandemic by providing 3 different patterns of tailored explanations (ie, belief-based, goal-based, and belief- and goal-based explanations) on students' intentions to change the recommended behaviors [27].

Although most chatbots (40/56, 71%) served only a single purpose, there were several COVID-19 chatbots (16/56, 29%) that were multipurpose in nature. For instance, *COOPERA*, as discussed previously, can be used to monitor COVID-19 symptoms, provide self-care instructions, and suggest preventive actions to avoid infections [31].

XSL•FO
RenderX

The use of COVID-19 chatbots across numerous application domains and contexts demonstrates the pandemic's impact on other elements of life. Nonhealth industries could investigate the use of a conversational agent to efficiently solve inquiries and issues without the need for human contact. This is critical for mitigating the rapid transmission of a global pandemic. Nevertheless, our results show that the bulk of the COVID-19 chatbots were developed to disseminate updated information regarding the COVID-19 pandemic and to track virus-related symptoms and other health-related concerns. As a result, automation of knowledge extraction from reputable sources may be required for information dissemination chatbots to be deployed efficiently and effectively as knowledge sources [78].

Limitations

To the best of our knowledge, this is the first review to systematically analyze the general characteristics and system designs of existing COVID-19 chatbots worldwide, thereby providing developers with directions for designing chatbots in response to a pandemic. Although pioneering, this study also has its limitations. First, we limited our search to peer-reviewed studies published in English with unrestricted access to their full texts. Second, the team did not perform ancestral searches on existing scoping and systematic reviews of COVID-19 chatbots. Third, meta-analysis was not performed in this systematic review, meaning that the effectiveness of COVID-19 chatbots in mitigating the effects of the pandemic was not analyzed. We acknowledge that the latter is critical to understanding the relationship between the current design of the chatbots and their efficacy in reducing the impacts of COVID-19, which may lead to a broader perspective on the management of a global pandemic. Finally, owing to the heterogeneity of the methodology and study designs included in this review, quality assessment, risk of bias, and critical appraisal of the papers were not conducted.

Recommendations

Our results indicate that there is a need to supplement COVID-19 chatbot research in the United States, Africa, and Australia as well as design and development studies on less well-known platforms for chatbots (eg, iOS and Google Cloud). A meta-analysis of existing COVID-19 chatbots is also recommended because there is a lack of studies evaluating the effectiveness of such chatbots. In addition, researchers and computer programmers are suggested to use the chatbot design framework by Nißen et al [21] for the purposes of having a uniform framework and for the identification of a cost-efficient design for a chatbot designed for a pandemic. Our findings also indicate that COVID-19 chatbots are used in various application areas aside from health care, which suggests the need for future investigations into other potential uses of chatbots during a pandemic. Moreover, aside from the design-centric approach toward the development of such chatbots, future studies must also evoke the assessment of user satisfaction and engagement with this type of chatbot as these are essential factors to consider for better uptake and use of these types of technology. Accordingly, it is imperative to appraise the influence of these types of conversational agents in the clinical practice guidelines and expert-client relationships [84]. In addition, our results show that several studies have "not stated" elements in multiple categories of the framework. This is because we noticed that the authors did not report the entire design specifications of their chatbot. It would be ideal for the authors to include all the intricate details of the design and development of chatbots to help future research on the most cost-efficient design for a chatbot made to address the negative effects of a pandemic.

Conclusions

This paper describes the general characteristics and design taxonomy of the existing COVID-19 chatbots. The conceptualization, development, implementation, and use of these conversational agents emerged to mitigate the effects of a global pandemic in societies worldwide. This study successfully summarized the current system design trends of COVID-19 chatbots based on 5 design perspectives using a chatbot design taxonomy, which may serve as a starting point for helping health and nonhealth chatbot communities on how to build and deploy a future-proof chatbot archetype that would meet the needs of the public amid the growing demand for better services as part of a more robust pandemic response. This would also direct researchers to explore methods on how to develop comprehensive evaluation and analytical methods for conversational agents in health care.

Acknowledgments

This study was supported by the Philippine Council for Health Research and Development of the Department of Science and Technology under the project entitled, "DISPO: Design, Development, and Implementation of a Natural Language Processing Enabled Virtual Assistant for Depression and Anxiety among People with COVID-19."

Authors' Contributions

WAL, RC, and MS contributed substantially to the study's conception and design, with inputs from RFS and AJA. RFS served as the principal investigator for the entire project. WAL, RC, and MS made substantial contributions to data collection, analysis, and interpretation of the study data, with inputs from RFS. RC drafted the manuscript, and this paper was revised critically for important intellectual content by all authors. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.
[DOCX File , 62 KB - jmir_v26i1e43112_app1.docx]

Multimedia Appendix 2
General characteristics of the COVID-19 chatbots.
[DOCX File , 30 KB - jmir_v26i1e43112_app2.docx]

References

1. Chakraborty I, Maity P. COVID-19 outbreak: migration, effects on society, global environment and prevention. Sci Total Environ 2020 Aug 01;728:138882 [FREE Full text] [doi: 10.1016/j.scitotenv.2020.138882] [Medline: 32335410]
2. COVID-19 Map. Johns Hopkins Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/map.html [accessed 2023-12-05]
3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
4. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
5. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed 2020 Mar 19;91(1):157-160 [FREE Full text] [doi: 10.23750/abm.v91i1.9397] [Medline: 32191675]
6. Almalki M, Azeez F. Health chatbots for fighting COVID-19: a scoping review. Acta Inform Med 2020 Dec;28(4):241-247 [FREE Full text] [doi: 10.5455/aim.2020.28.241-247] [Medline: 33627924]
7. Albites-Tapia A, Gamboa-Cruzado J, Almeyda-Ortiz J, Lázaro AM. Chatbots for the detection of COVID-19: a systematic review of the literature. Int J Adv Comput Sci Appl 2022;13(4):91-98 [FREE Full text] [doi: 10.14569/ijacsa.2022.01304113]
8. White BK, Martin A, White JA. User experience of COVID-19 chatbots: scoping review. J Med Internet Res 2022 Dec 27;24(12):e35903 [FREE Full text] [doi: 10.2196/35903] [Medline: 36520624]
9. Weeks R, Cooper L, Sangha P, Sedoc J, White S, Toledo A, et al. Chatbot-delivered COVID-19 vaccine communication message preferences of young adults and public health workers in Urban American communities: qualitative study. J Med Internet Res 2022 Jul 06;24(7):e38418 [FREE Full text] [doi: 10.2196/38418] [Medline: 35737898]
10. Battineni G, Chintalapudi N, Amenta F. AI chatbot design during an epidemic like the novel coronavirus. Healthcare (Basel) 2020 Jun 03;8(2):154 [FREE Full text] [doi: 10.3390/healthcare8020154] [Medline: 32503298]
11. Nagarhalli T, Vaze V, Rana NK. A review of current trends in the development of chatbot systems. In: Proceedings of the 6th International Conference on Advanced Computing and Communication Systems. 2020 Presented at: ICACCS '20; March 6-7, 2020; Coimbatore, India p. 706-710 URL: https://ieeexplore.ieee.org/document/9074420 [doi: 10.1109/icaccs48705.2020.9074420]
12. WHO health alert brings COVID-19 facts to billions via WhatsApp. World Health Organization. URL: https://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp [accessed 2021-04-26]
13. HealthBuddy: a new chatbot to engage with communities in Europe and Central Asia on COVID-19. World Health Organization. 2020. URL: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/5/healthbuddy-a-new-chatbot-to-engage-with-communities-in-europe-and-central-asia-on-covid-19 [accessed 2022-06-06]
14. Pappano L. College chatbots, with names like Iggy and pounce, are here to help. The New York Times. 2020. URL: https://www.nytimes.com/2020/04/08/education/college-ai-chatbots-students.html [accessed 2023-12-05]
15. Bebo, the benefits bot. Bebo. URL: https://whowteam.github.io/bebo/ [accessed 2022-06-06]
16. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
17. Hao K. The pandemic is emptying call centers. AI Chatbots are swooping in. MIT Technology Review. 2020 May 14. URL: https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coronavirus-pandemic/ [accessed 2023-12-05]
18. Gnewuch U, Morana S, Maedche A. Towards designing cooperative and social conversational agents for customer service. In: Proceedings of the 38th International Conference on Information Systems. 2017 Presented at: ICIS '17; December 10-13, 2017; Seoul, South Korea p. 1-14 URL: https://web.archive.org/web/20220803152206id_/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1312&context=icis2017
19. Seering J, Luria M, Kaufman GF, Hammer J. Beyond dyadic interactions: considering chatbots as community members. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Scotland, UK p. 1-13 URL: https://dl.acm.org/doi/10.1145/3290605.3300680 [doi: 10.1145/3290605.3300680]

XSL•FO
RenderX

20. Adamopoulou E, Moussiades L. An overview of chatbot technology. In: Proceedings of the 16th IFIP WG 12.5 International Conference on Artificial Intelligence Applications and Innovations. 2020 Presented at: AIAI '20; June 5–7, 2020; Neos Marmaras, Greece p. 373-383 URL: https://link.springer.com/chapter/10.1007/978-3-030-49186-4_31 [doi: 10.1007/978-3-030-49186-4_31]
21. Nißen M, Selimi D, Janssen A, Cardona DR, Breitner MH, Kowatsch T, et al. See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. Comput Human Behav 2022 Feb;127:107043 [FREE Full text] [doi: 10.1016/j.chb.2021.107043]
22. Siedlikowski S, Noël LP, Moynihan SA, Robin M. Chloe for COVID-19: evolution of an intelligent conversational agent to address infodemic management needs during the COVID-19 pandemic. J Med Internet Res 2021 Sep 21;23(9):e27283 [FREE Full text] [doi: 10.2196/27283] [Medline: 34375299]
23. Lee H, Kang J, Yeo J. Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: development and deployment. J Med Internet Res 2021 May 06;23(5):e27460 [FREE Full text] [doi: 10.2196/27460] [Medline: 33882012]
24. Pandey R, Gautam V, Pal R, Bandhey H, Dhingra LS, Misra V, et al. A machine learning application for raising WASH awareness in the times of COVID-19 pandemic. Sci Rep 2022 Jan 17;12(1):810 [FREE Full text] [doi: 10.1038/s41598-021-03869-6] [Medline: 35039533]
25. Jain A, Yadav K, Alharbi HF, Tiwari S. IoT and AI enabled three-phase secure and non-invasive COVID 19 diagnosis system. Comput Mater Contin 2022;71(1):423-438 [FREE Full text] [doi: 10.32604/cmc.2022.020238]
26. Mukhtar G, Almurtadha Y, Algarni F, Abdullah M, Felemban E, Alsharafi AM, et al. Mining the chatbot brain to improve COVID-19 bot response accuracy. Comput Mater Contin 2022;70(2):2619-2638 [FREE Full text] [doi: 10.32604/cmc.2022.020358]
27. Nelekar S, Abdulrahman A, Gupta M, Richards D. Effectiveness of embodied conversational agents for managing academic stress at an Indian University (ARU) during COVID‐19. Br J Educ Technol 2021 Dec 20;53(3):491-511 [FREE Full text] [doi: 10.1111/bjet.13174]
28. Loveys K, Sagar M, Pickering I, Broadbent E. A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: randomized pilot trial. JMIR Ment Health 2021 Nov 08;8(11):e31586 [FREE Full text] [doi: 10.2196/31586] [Medline: 34596572]
29. Hong YJ, Piao M, Kim J, Lee JH. Development and evaluation of a child vaccination chatbot real-time consultation messenger service during the COVID-19 pandemic. Appl Sci 2021 Dec 20;11(24):12142 [FREE Full text] [doi: 10.3390/app112412142]
30. Sweidan SZ, Abu Laban SS, Alnaimat NA, Darabkh KA. SIAAA‐C: a student interactive assistant android application with chatbot during COVID‐19 pandemic. Comput Appl Eng Educ 2021 Apr 16;29(6):1718-1742 [FREE Full text] [doi: 10.1002/cae.22419]
31. Yoneoka D, Kawashima T, Tanoue Y, Nomura S, Ejima K, Shi S, et al. Early SNS-based monitoring system for the COVID-19 outbreak in Japan: a population-level observational study. J Epidemiol 2020 Aug 05;30(8):362-370 [FREE Full text] [doi: 10.2188/jea.JE20200150] [Medline: 32475884]
32. Altay S, Hacquin AS, Chevallier C, Mercier H. Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. J Exp Psychol Appl 2023 Mar;29(1):52-62. [doi: 10.1037/xap0000400] [Medline: 34726454]
33. Hassan A, Ali MD, Ahammed R, Bourouis S, Khan MM. Development of NLP-integrated intelligent web system for E-mental health. Comput Math Methods Med 2021 Dec 13;2021:1546343-1546320 [FREE Full text] [doi: 10.1155/2021/1546343] [Medline: 34938348]
34. Martin A, Nateqi J, Gruarin S, Munsch N, Abdarahmane I, Zobel M, et al. An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. Sci Rep 2020 Nov 04;10(1):19012 [FREE Full text] [doi: 10.1038/s41598-020-75912-x] [Medline: 33149198]
35. Rexhepi H, Roig AE, Chaudhry T, Ollier J. Elena+ Care for COVID-19, a pandemic lifestyle care intervention: intervention design and study protocol. In: Kane BT, Rexhepi H, Hägglund M, Cajander Å, editors. Personalized Digital Health and Patient-centric Services. Lausanne, Switzerland: Frontiers Media SA; 2022.
36. Gabrielli S, Rizzi S, Bassi G, Carbone S, Maimone R, Marchesoni M, et al. Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. JMIR Mhealth Uhealth 2021 May 28;9(5):e27965 [FREE Full text] [doi: 10.2196/27965] [Medline: 33950849]
37. Asensio-Cuesta S, Blanes-Selva V, Portolés M, Conejero JA, García-Gómez JM. How the Wakamola chatbot studied a university community's lifestyle during the COVID-19 confinement. Health Informatics J 2021 May 27;27(2):14604582211017944 [FREE Full text] [doi: 10.1177/14604582211017944] [Medline: 34044657]
38. Van Baal ST, Le S, Fatehi F, Hohwy J, Verdejo-Garcia A. Cory COVID-bot: an evidence-based behavior change chatbot for COVID-19. Stud Health Technol Inform 2022 Jan 14;289:422-425. [doi: 10.3233/SHTI210948] [Medline: 35062181]
39. Morales HM, Guedes M, Silva JS, Massuda A. COVID-19 in Brazil-preliminary analysis of response supported by artificial intelligence in municipalities. Front Digit Health 2021 Jun 17;3:648585 [FREE Full text] [doi: 10.3389/fdgth.2021.648585] [Medline: 34713121]

XSL•FO
RenderX

40. Judson T, Odisho AY, Young JJ, Bigazzi O, Steuer D, Gonzales R, et al. Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic. J Am Med Inform Assoc 2020 Jul 01;27(9):1450-1455 [FREE Full text] [doi: 10.1093/jamia/ocaa130] [Medline: 32531066]
41. Morse KE, Ostberg NP, Jones VG, Chan AS. Use characteristics and triage acuity of a digital symptom checker in a large integrated health system: population-based descriptive study. J Med Internet Res 2020 Nov 30;22(11):e20549 [FREE Full text] [doi: 10.2196/20549] [Medline: 33170799]
42. Zhu Y, Wang R, Pu C. "I am chatbot, your virtual mental health adviser." What drives citizens' satisfaction and continuance intention toward mental health chatbots during the COVID-19 pandemic? An empirical study in China. Digit Health 2022 Mar 30;8:20552076221090031 [FREE Full text] [doi: 10.1177/20552076221090031] [Medline: 35381977]
43. Lai L, Wittbold KA, Dadabhoy FZ, Sato R, Landman AB, Schwamm LH, et al. Digital triage: novel strategies for population health management in response to the COVID-19 pandemic. Healthc (Amst) 2020 Dec;8(4):100493 [FREE Full text] [doi: 10.1016/j.hjdsi.2020.100493] [Medline: 33129176]
44. Mateos-Sanchez M, Melo AC, Blanco LS, García AM. Chatbot, as educational and inclusive tool for people with intellectual disabilities. Sustainability 2022 Jan 28;14(3):1520 [FREE Full text] [doi: 10.3390/su14031520]
45. Driss M, Almomani I, Alahmadi L, Alhajjam L, Alharbi R, Alanazi S. COVIBOT: a smart chatbot for assistance and e-awareness during COVID-19 pandemic. In: Proceedings of the 2nd International Conference of Smart Systems and Emerging Technologies. 2022 Presented at: SMARTTECH '22; May 22-24, 2022; Riyadh, Saudi Arabia p. 124-129 URL: https://ieeexplore.ieee.org/document/9842423 [doi: 10.1109/smarttech54121.2022.00038]
46. Raji VP, Maheswari PU. COVID-19 lockdown in India: an experimental study on promoting mental wellness using a chatbot during the Coronavirus. Int J Ment Health Promot 2022;19:189-205 [FREE Full text] [doi: 10.32604/ijmhp.2022.011865]
47. Briel A. Toward an eclectic and malleable multiagent educational assistant. Comput Appl Eng Educ 2021 Aug 31:1-11 [FREE Full text] [doi: 10.1002/cae.22449]
48. Hefny WE, Bolock AE, Herbert C, Abdennadher S. Applying the character-based chatbots generation framework in education and healthcare. In: Proceedings of the 9th International Conference on Human-Agent Interaction. 2021 Presented at: HAI '21; November 9-11, 2021; Virtual Event p. 121-129 URL: https://dl.acm.org/doi/10.1145/3472307.3484172 [doi: 10.1145/3472307.3484172]
49. Fu C, Liu C, Ishi CT, Yoshikawa Y, Iio T, Ishiguro H. Using an android robot to improve social connectedness by sharing recent experiences of group members in human-robot conversations. IEEE Robot Autom Lett 2021 Oct;6(4):6670-6677 [FREE Full text] [doi: 10.1109/lra.2021.3094779]
50. Khattab M. Infobotting COVID-19: a case study of ask Nameesa in Egypt. J Afr Media Stud 2021;13:451-476 [FREE Full text] [doi: 10.1386/jams_00059_1]
51. Hefny WE, Bolock AE, Herbert C, Abdennadher S. Chase away the virus: a character-based chatbot for COVID-19. In: Proceedings of the 9th International Conference on Serious Games and Applications for Health. 2021 Presented at: SeGAH '21; August 4-6, 2021; Dubai, United Arab Emirates p. 1-8 URL: https://ieeexplore.ieee.org/document/9551895 [doi: 10.1109/segah52098.2021.9551895]
52. Salhi I, Guemmat KE, Qbadou M, Mansouri K. Towards developing a pocket therapist: an intelligent adaptive psychological support chatbot against mental health disorders in a pandemic situation. J Electr Eng Comput Sci 2021 Aug 01;23(2):1200 [FREE Full text] [doi: 10.11591/ijeecs.v23.i2.pp1200-1211]
53. Al-Madi N, Maria KA, Al-Madi MA, Alia MA. An intelligent Arabic chatbot system proposed framework. In: Proceedings of the 2021 International Conference on Information Technology. 2021 Presented at: ICIT '21; July 14-15, 2021; Amman, Jordan p. 592-597 URL: https://ieeexplore.ieee.org/document/9491699 [doi: 10.1109/icit52682.2021.9491699]
54. Valtolina S. Charlie: a chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In: Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter. 2021 Presented at: CHItaly '21; July 11-13, 2021; Bolzano, Italy p. 1-5 URL: https://dl.acm.org/doi/abs/10.1145/3464385.3464726 [doi: 10.1145/3464385.3464726]
55. Amer E, Hazem A, Farouk O, Louca A, Mohamed Y, Ashraf M. A proposed chatbot framework for COVID-19. In: Proceedings of the 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference. 2021 Presented at: MIUCC '21; May 26-27, 2021; Cairo, Egypt p. 263-268 URL: https://ieeexplore.ieee.org/document/9447652 [doi: 10.1109/miucc52538.2021.9447652]
56. Ariga RA. Self care nursing: teleassessment nursing with chatbot application the coronavirus disease-19 pandemic period in North Sumatra. Open Access Maced J Med Sci 2021 Dec 06;9(G):306-310. [doi: 10.3889/oamjms.2021.7172]
57. Holt-Quick C, Warren J, Stasiak K, Williams R, Christie G, Hetrick S, et al. A chatbot architecture for promoting youth resilience. Stud Health Technol Inform 2021:99-105 [FREE Full text] [doi: 10.3233/shti210017]
58. Alkmim MB, Marcolino MS, de Oliveira CR, Borges IN, Cardoso CS, Rocha GM, et al. TeleCOVID-19: a multifaceted strategy from a public Brazilian telehealth service during the COVID-19 pandemic. Stud Health Technol Inform 2021;19:1-10 [FREE Full text] [doi: 10.3233/shti210022]
59. Kim RH. Development of a conversing and body temperature scanning autonomously navigating robot to help screen for COVID-19. arXiv Preprint posted online June 6, 2021 [FREE Full text] [doi: 10.1109/urtc54388.2021.9701616]

XSL•FO
RenderX

60. Du J, Huang W, Hew KF. Supporting students goal setting process using chatbot: implementation in a fully online course. In: Proceedings of the 2021 IEEE International Conference on Engineering, Technology & Education. 2021 Presented at: TALE '21; December 5-8, 2021; Wuhan, Chaina p. 35-41 URL: https://ieeexplore.ieee.org/document/9678564 [doi: 10.1109/tale52509.2021.9678564]
61. Aminuddin R, Noor HM, Ilias NF, Wahab NI. Framework for a mobile application with a chatbot to self-report injuries and carry out contact tracing for athletes and sports trainers. In: Proceedings of the 2021 IEEE Symposium on Industrial Electronics & Applications. 2021 Presented at: ISIEA '21; August 1-4, 2021; Langkawi Island, Malaysia p. 1-6 URL: https://ieeexplore.ieee.org/document/9509978 [doi: 10.1109/isiea51897.2021.9509978]
62. Ito T, Sode TM, Masako S, Keisuke M. The online PBL (project-based learning) education system using AI (artificial intelligence). In: Proceedings of the 23rd International Conference on Engineering and Product Design Education. 2021 Presented at: E&PDE '21; September 9-10, 2021; Herning, Denmark p. 1-6 URL: https://tinyurl.com/5n7u6cpv [doi: 10.35199/epde.2021.19]
63. Chalaguine L, Hunter A. Addressing popular concerns regarding COVID-19 vaccination with natural language argumentation dialogues. In: Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. 2021 Presented at: ECSQARU '21; September 21-24, 2021; Prague, Czech Republic p. 59-73 URL: https://link.springer.com/chapter/10.1007/978-3-030-86772-0_5 [doi: 10.1007/978-3-030-86772-0_5]
64. Patgar T, Patel R, Girija S. Real conversation with human-machine 24/7 COVID-19 chatbot based on knowledge graph contextual search. In: Proceedings of the 16th International Conference on Information Processing on Data Science and Computational Intelligence. 2021 Presented at: ICInPro '21; October 22-24, 2021; Bengaluru, India p. 258-272 URL: https://link.springer.com/chapter/10.1007/978-3-030-91244-4_21 [doi: 10.1007/978-3-030-91244-4_21]
65. Oruche R, Gundlapalli V, Biswal AP, Calyam P, Alarcon ML, Zhang Y, et al. Evidence-based recommender system for a COVID-19 publication analytics service. IEEE Access 2021;9:79400-79415. [doi: 10.1109/access.2021.3083583]
66. Pandey A, Janghel RR, Sujatha R, Kumar S, Kumar TS, Chatterjee JM. Coronago website integrated with chatbot for COVID-19 tracking. In: Proceedings of the 2021 International Semantic Intelligence Conference. 2021 Presented at: ISIC ’21; February 25-27, 2021; Delhi, India p. 521-527 URL: https://ceur-ws.org/Vol-2786/Paper62.pdf
67. Hossain ME, Najib AU, Islam MZ. Combating domestic violence during COVID-19 pandemic in Bangladesh: using a mobile application integrated with an effective solution. In: Proceedings of the 23rd International Conference on Computer and Information Technology. 2020 Presented at: ICCIT '20; December 19-21, 2020; Dhaka, Bangladesh p. 1-6. [doi: 10.1109/iccit51783.2020.9392691]
68. Erazo WS, Guerrero GP, Betancourt CC, Salazar IS. Chatbot implementation to collect data on possible COVID-19 cases and release the pressure on the primary health care system. In: Proceedings of the 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference. 2020 Presented at: IEMCON '20; November 4-7, 2020; Vancouver, BC p. 0302-0307 URL: https://ieeexplore.ieee.org/document/9284846 [doi: 10.1109/iemcon51383.2020.9284846]
69. Maniou TA, Veglis A. Employing a chatbot for news dissemination during crisis: design, implementation and evaluation. Future Internet 2020 Jun 30;12(7):109. [doi: 10.3390/fi12070109]
70. Bharti U, Bajaj D, Batra H, Lalit S, Lalit S, Gangwani A. Medbot: conversational artificial intelligence powered chatbot for delivering tele-health after COVID-19. In: Proceedings of the 5th International Conference on Communication and Electronics Systems. 2020 Presented at: ICCES '19; July 20-22, 2019; Coimbatore, India p. 870-875 URL: https://ieeexplore.ieee.org/document/9137944 [doi: 10.1109/icces48766.2020.9137944]
71. Bahja M, Abuhwaila N, Bahja J. An antenatal care awareness prototype chatbot application using a user-centric design approach. In: Proceedings of the HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference. 2020 Presented at: HCII '20; July 19-24, 2020; openhagen, Denmark p. 20-31 URL: https://dl.acm.org/doi/abs/10.1007/978-3-030-60117-1_2 [doi: 10.1007/978-3-030-60117-1_2]
72. Gross C, Kohlbrenner D, Clarenbach CF, Ivankay A, Brunschwiler T, Nordmann Y, et al. A telemonitoring and hybrid virtual coaching solution "CAir" for patients with chronic obstructive pulmonary disease: protocol for a randomized controlled trial. JMIR Res Protoc 2020 Oct 22;9(10):e20412 [FREE Full text] [doi: 10.2196/20412] [Medline: 33090112]
73. Meinert E, Milne-Ives M, Surodina S, Lam C. Agile requirements engineering and software planning for a digital health platform to engage the effects of isolation caused by social distancing: case study. JMIR Public Health Surveill 2020 May 06;6(2):e19297 [FREE Full text] [doi: 10.2196/19297] [Medline: 32348293]
74. Rodrigo M. Towards building mental health resilience through storytelling with a chatbot. In: Proceedings of the 29th International Conference on Computers in Education. 2021 Presented at: ICCE '21; November 22-26, 2021; Virtual Event URL: https://icce2021.apsce.net/wp-content/uploads/2021/12/ICCE2021-Vol.II-PP.-569-574.pdf
75. Singh S, Kaur M, Tanwar P, Sharma S. Design and development of conversational chatbot for COVID-19 using NLP: an AI application. In: Proceedings of the 6th International Conference on Computing Methodologies and Communication. 2022 Presented at: ICCMC '22; March 29-31, 2022; Erode, India p. 1654-1658 URL: https://ieeexplore.ieee.org/document/9753893 [doi: 10.1109/iccmc53470.2022.9753893]
76. Fagbolu O, Sellu M, Fagbolu TA. Overview of interactive chatbot for modelling, predicting and reporting COVID-19 in Mano river union. Lond J Res Comput Sci Technol 2021 Nov [FREE Full text]

XSL•FO
RenderX

77. Ilievski V, Musat C, Hossmann A, Baeriswyl M. Goal-oriented chatbot dialog management bootstrapping with transfer learning. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018 Presented at: IJCAI '18; July 13-19, 2018; Stockholm, Sweden p. 4115-4121 URL: https://www.ijcai.org/proceedings/2018/0572.pdf [doi: 10.24963/ijcai.2018/572]
78. Amiri P, Karahanna E. Chatbot use cases in the COVID-19 public health response. J Am Med Inform Assoc 2022 Apr 13;29(5):1000-1010 [FREE Full text] [doi: 10.1093/jamia/ocac014] [Medline: 35137107]
79. Cai D, Li H, Law R. Anthropomorphism and OTA chatbot adoption: a mixed methods study. J Travel Tour Mark 2022 May 05;39(2):228-255. [doi: 10.1080/10548408.2022.2061672]
80. Chaves AP, Gerosa MA. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. Int J Hum Comput Int 2020 Nov 08;37(8):729-758. [doi: 10.1080/10447318.2020.1841438]
81. Duijst D. Can we improve the user experience of chatbots with personalisation. University of Amsterdam. 2017. URL: https://www.scribd.com/document/413277603/Can-We-Improve-the-User-Experience-of-Chatbots-With-Personalisation-Master-Thesis-Final [accessed 2023-12-05]
82. Fadhil A. Beyond patient monitoring: conversational agents role in telemedicine and healthcare support for home-living elderly individuals. arXiv Preprint posted online March 3, 2018 [FREE Full text]
83. Dohsaka K, Asai R, Higashinaka R, Minami Y, Maeda E. Effects of conversational agents on activation of communication in thought-evoking multi-party dialogues. IEICE Trans Inf Syst 2014;E97.D(8):2147-2156. [doi: 10.1587/transinf.e97.d.2147]
84. Parviainen J, Rantala J. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos 2022 Mar 04;25(1):61-71 [FREE Full text] [doi: 10.1007/s11019-021-10049-w] [Medline: 34480711]

Abbreviations

AI: artificial intelligence
COOPERA: COVID-19: Operation for Personalized Empowerment to Render Smart Prevention and Care Seeking
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
WHO: World Health Organization

Edited by T Leung, V Arnold, T Purnat, H Gouda; submitted 30.09.22; peer-reviewed by Z Zandesh, M Rampioni, B Xie; comments to author 23.11.22; revised version received 28.02.23; accepted 11.07.23; published 05.01.24.

Please cite as:
Lim WA, Custodio R, Sunga M, Amoranto AJ, Sarmiento RF
General Characteristics and Design Taxonomy of Chatbots for COVID-19: Systematic Review
J Med Internet Res 2024;26:e43112
URL: https://www.jmir.org/2024/1/e43112
doi:10.2196/43112
PMID:38064638

©Wendell Adrian Lim, Razel Custodio, Monica Sunga, Abegail Jayne Amoranto, Raymond Francis Sarmiento. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Where Do Oncology Patients Seek and Share Health Information? Survey Study

Eric Freeman[1], BA; Darshilmukesh Patel[2], BA; Folasade Odeniyi[1], MPH, MBA; Mary Pasquinelli[2], DNP; Shikha Jain[2], MD

[1]College of Medicine, University of Illinois at Chicago, Chicago, IL, United States

[2]Department of Medicine, University of Illinois at Chicago, Chicago, IL, United States

Corresponding Author:
Eric Freeman, BA
College of Medicine
University of Illinois at Chicago
1853 West Polk Street
Chicago, IL, 60612
United States
Phone: 1 847 791 0189
Email: efreem6@uic.edu

(J Med Internet Res 2024;26:e36441) doi:10.2196/36441

KEYWORDS

oncology; social media; patient-physician relationship; patient-physician; patient-provider; cancer; information sharing; information seeking; information behavior; technology access; digital divide

Introduction

Social media in health care has many benefits, including the dissemination of health information [1] and health promotion [2]. The COVID-19 pandemic has highlighted the benefits of the internet and social media as tools through which individuals can exchange health information. While little is known about oncology patients' preferences for social media platforms, particularly among minority populations and those in low socioeconomic status communities, some studies have shown its use is linked to the alleviation of patient stress and loneliness, increased feelings of self-efficacy and control of care, and efficient delivery of health information from health practitioners [3]. The study aims to assess where patients from marginalized communities receive a majority of their health care information by surveying patients in a cancer clinic. This study was conducted at the University of Illinois Chicago, which is a public hospital that mainly serves patients from underresourced communities.

Methods

Overview

Between March 2021 to June 2021, we administered a 16-item survey (Multimedia Appendix 1) adapted from the National Cancer Institute's Health Information National Trends Survey (HINTS) [4] to patients scheduled for an oncology visit at the Outpatient Care Center at UI Health. The survey was administered to 145 patients via email and 161 patients in person. Respondents were asked to identify sources used to self-educate about their diagnosis, preferred information source, social media use and preferences, and demographics. We used chi-square tests to assess associations between categorical variables.

Ethics Approval

This study was approved by the institutional review board at the University of Illinois Chicago and was found to meet the criteria for exemption as defined in the US Department of Health and Human Services Regulations for the Protection of Human Subjects (45 CFR 46.104(d)).

Results

The demographics of our sample can be found in Table 1. Respondents routinely accessed several forms of health information sources. The top three included their doctor or health care provider (n=274, 89.3%), internet search engines (n=218, 71.2%), and brochures and pamphlets (n=125, 40.7%). However, when directed to choose just one source, 207 (67.4%) chose their doctor or health care provider, while 67 (21.8%) chose internet search engines. The majority of respondents used a smartphone with the internet (n=237, 77.2%), a home desktop or laptop with the internet (n=192, 62.5%), or a tablet with the internet (n=188, 61.2%). However, approximately one-quarter of respondents indicated that they used a mobile phone without internet or a data plan.

XSL•FO
RenderX

We found that the majority of respondents accessed social media in the past year (n=198, 64.7%). Using social media was associated with age (χ^2_3=18.7; P<.001) and sex (Fisher P=.001). While respondents primarily used Facebook (n=69, 22.5%), YouTube (n=66, 21.5%), and Instagram (n=25, 8.1%) to receive health information, few shared health information with a medical professional (n=17, 5.5%), and if they did, they primarily used Facebook (n=8, 48.7%).

XSL•FO
RenderX

Table 1. Respondent demographics.

	Email (n=145)	In person (n=161)	Statistic (*df*)	*P* value
Age (years), mean (SD)	60.1 (13.4)	55.3 (15.2)	*t* test: 2.9 (304)	.004
Gender, n (%)			Fisher	.04
Male	40 (27.6)	48 (29.8)		
Female	105 (72.4)	107 (66.5)		
Not reported	0 (0.0)	6 (3.7)		
Race, n (%)			Fisher	.26
Asian	4 (2.8)	6 (3.7)		
Black or African American	75 (51.7)	85 (52.8)		
White	49 (33.8)	50 (31.2)		
Mixed race or biracial	0 (0.0)	5 (3.1)		
Not reported	17 (11.7)	15 (9.3)		
Ethnicity, n (%)			Fisher	<.001
Hispanic or Latino	18 (12.4)	26 (16.2)		
Not Hispanic or Latino	125 (86.2)	98 (60.9)		
Not reported	2 (1.4)	37 (22.9)		
Primary care, n (%)			Chi-square: 6.1689 (3)	.10
Has a primary care provider but has not seen them for over a year	19 (13.1)	22 (13.7)		
Has a primary care provider not at UI Health, seen in the past year	36 (24.8)	50 (31.1)		
Has a primary care provider at UI Health, seen in the past year	84 (57.9)	74 (45.9)		
Does not have a primary care provider	6 (4.1)	15 (9.3)		
Work status, n (%)			Fisher	.48
Full-time paid work or education	31 (21.4)	30 (18.6)		
Part-time paid work or part-time education	10 (6.9)	9 (5.6)		
Full-time career/homemaker	3 (2.1)	2 (1.2)		
Unemployed	40 (27.6)	21 (13)		
Retired	61 (42.1)	54 (33.5)		
Current living situation, n (%)			Fisher	.13
Alone	36 (24.8)	46 (28.6)		
With my spouse or partner	56 (38.6)	44 (27.3)		
With my nonadult children	12 (8.3)	12 (7.5)		
With family or friends	40 (27.6)	59 (36.7)		
Homeless	1 (0.7)	0 (0.0)		
Education, n (%)			Fisher	<.001
Less than high school	5 (3.5)	19 (11.8)		
High school diploma or equivalent	26 (17.9)	48 (29.6)		
Some college but no degree	32 (22.1)	30 (18.6)		
Associate's degree or trade school	15 (10.3)	23 (14.3)		
Bachelor's degree	37 (25.5)	26 (16.1)		
Graduate degree (master's, MD, JD, PhD)	30 (20.7)	15 (9.3)		
Annual income (US $), n (%)			Chi-square: 8.323 (3)	.04

XSL•FO
RenderX

	Email (n=145)	In person (n=161)	Statistic (*df*)	*P* value
<15,000	14 (9.7)	31 (19.1)		
15,000-30,000	18 (12.4)	23 (14.2)		
30,000	57 (39.3)	44 (27.2)		
I wish not to answer	56 (38.6)	64 (39.5)		
Comfort with income, n (%)			Chi-square: 0.81368 (3)	.85
Living comfortably on present income	54 (37.2)	56 (35.8)		
Getting by on present income	58 (40)	63 (38.9)		
Finding it difficult on present income	26 (17.9)	32 (19.8)		
Finding it very difficult	7 (4.8)	11 (6.8)		

Discussion

Principal Findings

Understanding how patients exchange health information is important to ensure access to accurate information and promote engagement with the health care team. We found that a majority of our patients use social media to find health-related information. However, there continues to be an internet access disparity that can limit patients' ability to improve their health literacy. As social media engagement is linked to positive patient outcomes, using social media interventions can help us improve oncology patients' illness experience. While both oncology providers and patients are increasingly using social media as a learning and sharing tool [5], the exact information-seeking behavior of patients with cancer has yet to be fully examined, especially in disadvantaged populations. In the current climate of rampant online medical misinformation, health care workers should find innovative ways to disseminate evidence-based patient-facing information using the platforms most accessed by oncology patients. Our study highlights the need to further explore communication preferences to help develop tailored communication strategies to support underserved patients and their families.

Limitations

Our study has various limitations. This study was a single clinic, single institution study with a relatively small sample size. Additionally, our patient population was older, which could have influenced preferred social media platforms.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Social media survey.
[DOCX File , 24 KB - jmir_v26i1e36441_app1.docx]

References

1. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
2. Khatri C, Chapman SJ, Glasbey J, Kelly M, Nepogodiev D, Bhangu A, STARSurg Committee. Social media and internet driven study recruitment: evaluating a new model for promoting collaborator engagement and participation. PLoS One 2015;10(3):e0118899 [FREE Full text] [doi: 10.1371/journal.pone.0118899] [Medline: 25775005]
3. Leist AK. Social media use of older adults: a mini-review. Gerontology 2013;59(4):378-384. [doi: 10.1159/000346818] [Medline: 23594915]
4. National Cancer Institute. Healthcare Information National Trends Survey. 2018. URL: https://hints.cancer.gov/ [accessed 2023-09-12]
5. Watson J. Social media use in cancer care. Semin Oncol Nurs 2018 May;34(2):126-131. [doi: 10.1016/j.soncn.2018.03.003] [Medline: 29622519]

XSL•FO
RenderX

Abbreviations

HINTS: Health Information National Trends Survey

Edited by A Mavragani; submitted 21.03.22; peer-reviewed by S El kefi, S Hargreavess, K Na; comments to author 17.11.22; revised version received 16.06.23; accepted 04.07.23; published 25.03.24.

Please cite as:
Freeman E, Patel D, Odeniyi F, Pasquinelli M, Jain S
Where Do Oncology Patients Seek and Share Health Information? Survey Study
J Med Internet Res 2024;26:e36441
URL: https://www.jmir.org/2024/1/e36441
doi:10.2196/36441
PMID:38526546

©Eric Freeman, Darshilmukesh Patel, Folasade Odeniyi, Mary Pasquinelli, Shikha Jain. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Using Longitudinal Twitter Data for Digital Epidemiology of Childhood Health Outcomes: An Annotated Data Set and Deep Neural Network Classifiers

Ari Z Klein[1], PhD; José Agustín Gutiérrez Gómez[2], MD; Lisa D Levine[3], MD; Graciela Gonzalez-Hernandez[4], PhD

[1]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[2]Department of Health Sciences, University of Monterrey, San Pedro Garza García, Nuevo León, Mexico

[3]Department of Obstetrics and Gynecology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, United States

Corresponding Author:
Graciela Gonzalez-Hernandez, PhD
Department of Computational Biomedicine
Cedars-Sinai Medical Center
Pacific Design Center, Ste G549F
700 N San Vicente Blvd
West Hollywood, CA, 90069
United States
Phone: 1 310 423 3521
Email: Graciela.GonzalezHernandez@csmc.edu

Abstract

We manually annotated 9734 tweets that were posted by users who reported their pregnancy on Twitter, and used them to train, evaluate, and deploy deep neural network classifiers (F_1-score=0.93) to detect tweets that report having a child with attention-deficit/hyperactivity disorder (678 users), autism spectrum disorders (1744 users), delayed speech (902 users), or asthma (1255 users), demonstrating the potential of Twitter as a complementary resource for assessing associations between pregnancy exposures and childhood health outcomes on a large scale.

(J Med Internet Res 2024;26:e50652) doi:10.2196/50652

KEYWORDS

natural language processing; machine learning; data mining; social media; Twitter; pregnancy; epidemiology; developmental disabilities; asthma

Introduction

Many children are diagnosed with disorders that can impact their daily lives and can last throughout their lifetime. For example, in the United States, 17% of children are diagnosed with a developmental disability [1] and 8% of them with asthma [2]. Meanwhile, data sources for assessing the association of these outcomes with pregnancy exposures are limited, as pregnancy registries typically follow infants for up to 1 year after birth. While our previous work [3,4] demonstrated the utility of Twitter as a source of data regarding pregnancy outcomes, the ability to continue collecting users' tweets on an ongoing basis after birth may present opportunities to detect outcomes in childhood. Twitter data have been used to identify self-reports of attention-deficit/hyperactivity disorder (ADHD) [5], autism spectrum disorders (ASD) [6], and asthma [7], but not to identify reports of these disorders in users' children. This study aimed to assess whether there are users who report having a child with ADHD, ASD, delayed speech, or asthma, and develop and evaluate an automated method for identifying these reports.

Methods

Ethical Considerations

The study data were collected and analyzed in accordance with the Twitter Terms of Service. The institutional review boards of the University of Pennsylvania and Cedars-Sinai Medical Center deemed this study exempt.

XSL•FO
RenderX

Data Collection

We searched for mentions of ADHD, ASD, delayed speech, and asthma among all the tweets posted by more than 100,000 users who reported their pregnancy on Twitter [8]. We then searched these matching tweets for references to a child and the user, and excluded tweets that matched specific patterns indicating the user's own disorder. The query (Multimedia Appendix 1) returned 36,094 tweets (excluding retweets) posted by 11,712 users.

Annotation

We used 400 matching tweets—100 per outcome—to develop annotation guidelines (Multimedia Appendix 2) for distinguishing those that report having a child with a disorder from those that do not. An additional 9334 tweets—1 random tweet per user—were then independently annotated: 8334 by 2 annotators and 1000 by all 3. Interannotator agreement (Fleiss kappa) was 0.88. After resolving disagreements among all 9734 tweets, we determined that 3019 (31%) reported having a child with a disorder and 6715 (69%) did not.

Automatic Classification

We split the 9734 tweets into 80% (n=7787) training (Multimedia Appendix 3) and 20% (n=1947) test data, and performed machine learning experiments using deep neural network classifiers based on bidirectional encoder representations from transformers (BERT) [9]: the BERT-Base-Uncased, RoBERTa-Large, and BERTweet-Large pretrained models in the *Huggingface* library. Our preprocessing included normalizing URLs and usernames, and lowercasing the tweets. For training, we used Adam optimization, 5 epochs, a batch size of 8, and a learning rate of 0.00001, based on evaluating after each epoch using a 5% split of the training set. We fine-tuned all layers of the models with our annotated tweets.

Results

Table 1 presents the performance of the classifiers. The RoBERTa-Large [10] classifier achieved the highest overall F_1-score (0.93). Table 1 also presents the performance of the RoBERTa-Large classifier for tweets that mention specific outcomes. We deployed the RoBERTa-Large classifier on the additional 26,360 unlabeled tweets that matched our query (Multimedia Appendix 1). Between the 9734 manually annotated tweets and the 26,360 automatically classified tweets, we identified 3806 total users who reported having a child with ADHD (n=678), ASD (n=1744), delayed speech (n=902), or asthma (n=1255).

Table 2 presents examples of tweets in the test set that were misclassified by the RoBERTa-Large classifier. While 28 (58%) of the 48 false positives do refer to the user's child, 11 (39%) indicate that someone other than the user's child has a disorder (tweet 1), and 9 (32%) indicate that a disorder is merely suspected or exhibited (tweet 2). Among the other 20 (42%) of the 48 false positives, 10 (50%) are reported speech, such as quotations (tweet 3). Among the 42 false negatives, 22 (52%) do not explicitly mention the user's child (tweet 4)—for example, using a pronoun or name—and 14 (33%) do not explicitly indicate that the child has a disorder (tweet 5).

Table 1. Precision, recall, and F_1-score of classifiers for the class of tweets that report having a child with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), delayed speech, or asthma, including the outcome-specific precision, recall, and F_1-score for the RoBERTa-Large classifier.

Classifier	Precision	Recall	F_1-score
BERT-Base-Uncased	0.83	0.87	0.85
BERTweet-Large	0.89	0.94	0.92
RoBERTa-Large	0.92	0.94	0.93
ADHD	0.91	0.85	0.88
ASD	0.94	0.92	0.93
Delayed speech	0.94	0.96	0.95
Asthma	0.91	0.96	0.94

XSL•FO
RenderX

Table 2. Sample false positives and false negatives of a RoBERTa-Large classifier for detecting tweets that report having a child with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), delayed speech, or asthma (with the text that matched the data collection query in italics).

Tweet number	Tweet	Actual	Predicted
1	So Maxine Waters can be maskless on a plane but *I* can't fly with *my 2 year old* cause she won't wear a mask? *Kids* with *autism* are being banned from flying because they won't wear a mask?	–	+
2	they treat *my baby* with *asthma* meds all the time but didn't diagnose her with it im pretty sure she has it tho	–	+
3	Any tips for this mum: "*My daughter* is 10. *My* parents would like to gift her either a phone or a smart watch which is easy to use and won't be easily damaged by a very active *ADHD kid*... *I* need help choosi… [URL]	–	+
4	Flying tomorrow...during a pandemic with a *nonverbal 3 year old*. *We* could use some prayers, please.	+	–
5	*I* wouldn't change *my child* for anything in the world. *I*'m just curious to know where *autism* came from because *me* and his dad don't have any family members that are *autistic*. It's just a question out of curiosity	+	–

Discussion

Our ability to identify Twitter data during pregnancy for users who reported having a child with ADHD, ASD, delayed speech, or asthma suggests that Twitter could be a complementary resource for assessing associations between pregnancy exposures and childhood health outcomes, with potential clinical implications for informing prenatal care. The overall and outcome-specific performance for automatically identifying these outcomes demonstrates the feasibility of using Twitter data for observational studies on a large scale.

Acknowledgments

This work was supported by the National Library of Medicine (R01LM011176). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Ivan Flores for contributing to software applications and Karen O'Connor for contributing to annotating the Twitter data. Generative artificial intelligence was not used for any portion of the study or manuscript writing.

Data Availability

The manually annotated training data are included with this article in Multimedia Appendix 3. In accordance with the Twitter Terms of Service, these tweets are made available as tweet IDs, which can be rehydrated as tweet objects if they remain public at the time they are requested through the Twitter API.

Authors' Contributions

AZK and JAGG contributed to the data collection, annotation, machine learning experiments, error analysis, and drafting of the manuscript. LDL provided guidance on pregnancy outcomes and edited the manuscript. GGH designed and guided the study and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data collection query.
[TXT File , 11 KB - jmir_v26i1e50652_app1.txt]

Multimedia Appendix 2
Annotation guidelines.
[DOCX File , 28 KB - jmir_v26i1e50652_app2.docx]

Multimedia Appendix 3
Training data.
[TXT File , 165 KB - jmir_v26i1e50652_app3.txt]

XSL•FO
RenderX

References

1. Zablotsky B, Black LI, Maenner MJ, Schieve LA, Danielson ML, Bitsko RH, et al. Prevalence and trends of developmental disabilities among children in the United States: 2009-2017. Pediatrics 2019 Oct;144(4):e20190811 [FREE Full text] [doi: 10.1542/peds.2019-0811] [Medline: 31558576]
2. Zahran HS, Bailey CM, Damon SA, Garbe PL, Breysse PN. Vital signs: asthma in children - United States, 2001-2016. MMWR Morb Mortal Wkly Rep 2018 Feb 09;67(5):149-155. [doi: 10.15585/mmwr.mm6705e1] [Medline: 29420459]
3. Golder S, Chiuve S, Weissenbacher D, Klein A, O'Connor K, Bland M, et al. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. Drug Saf 2019 Mar 3;42(3):389-400. [doi: 10.1007/s40264-018-0731-6] [Medline: 30284214]
4. Klein AZ, O'Connor K, Levine LD, Gonzalez-Hernandez G. Using Twitter data for cohort studies of drug safety in pregnancy: proof-of-concept with β-blockers. JMIR Form Res 2022 Jun 30;6(6):e36771 [FREE Full text] [doi: 10.2196/36771] [Medline: 35771614]
5. Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in adults on social media. J Atten Disord 2019 Oct 08;23(12):1475-1485. [doi: 10.1177/1087054717738083] [Medline: 29115168]
6. Hswen Y, Gopaluni A, Brownstein JS, Hawkins JB. Using Twitter to detect psychological characteristics of self-identified persons with autism spectrum disorder: a feasibility study. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e12264 [FREE Full text] [doi: 10.2196/12264] [Medline: 30747718]
7. Edo-Osagie O, Smith G, Lake I, Edeghere O, De La Iglesia B. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. PLoS One 2019;14(7):e0210689 [FREE Full text] [doi: 10.1371/journal.pone.0210689] [Medline: 31318885]
8. Klein AZ, Kunatharaju S, O'Connor K, Gonzalez-Hernandez G. Pregex: rule-based detection and extraction of Twitter data in pregnancy. J Med Internet Res 2023 Feb 09;25:e40569 [FREE Full text] [doi: 10.2196/40569] [Medline: 36757756]
9. Devlin J, Cheng M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, MN p. 4171-4186.
10. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online July 26, 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2023-07-07]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ASD: autism spectrum disorder
BERT: bidirectional encoder representations from transformers

Edited by T de Azevedo Cardoso; submitted 07.07.23; peer-reviewed by C Ni, E Guo; comments to author 22.08.23; revised version received 05.09.23; accepted 19.09.23; published 25.03.24.

Please cite as:
Klein AZ, Gutiérrez Gómez JA, Levine LD, Gonzalez-Hernandez G
Using Longitudinal Twitter Data for Digital Epidemiology of Childhood Health Outcomes: An Annotated Data Set and Deep Neural Network Classifiers
J Med Internet Res 2024;26:e50652
URL: https://www.jmir.org/2024/1/e50652
doi:10.2196/50652
PMID:38526542

©Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D Levine, Graciela Gonzalez-Hernandez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Response of Unvaccinated US Adults to Official Information About the Pause in Use of the Johnson & Johnson–Janssen COVID-19 Vaccine: Cross-Sectional Survey Study

Vishala Mishra[1*], MBBS, MMCi; Joseph P Dexter[2,3,4*], PhD

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, United States
[2]Data Science Initiative, Harvard University, Allston, MA, United States
[3]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, United States
[4]Institute of Collaborative Innovation, University of Macau, Taipa, Macao
[*]all authors contributed equally

Corresponding Author:
Joseph P Dexter, PhD
Data Science Initiative
Harvard University
Science and Engineering Complex 1.312-10
150 Western Avenue
Allston, MA, 02134
United States
Phone: 1 8023381330
Email: jdexter@fas.harvard.edu

Abstract

Using a rapid response web-based survey, we identified gaps in public understanding of the Centers for Disease Control and Prevention's messaging about the pause in use of the Johnson & Johnson–Janssen COVID-19 vaccine and estimated changes in vaccine hesitancy using counterfactual questions.

(J Med Internet Res 2024;26:e41559) doi:10.2196/41559

KEYWORDS

Centers for Disease Control and Prevention; CDC; COVID-19; health communication; health information; health literacy; public health; risk perception; SARS-CoV-2; vaccine hesitancy; web-based surveys

Introduction

On April 13, 2021, the Centers for Disease Control and Prevention (CDC) and Food and Drug Administration recommended a pause in use of the Johnson & Johnson (J&J)–Janssen COVID-19 vaccine due to 6 reports of cerebral venous sinus thrombosis in recently vaccinated individuals [1]. The announcement of the pause required development of a coordinated communication strategy under extreme time pressure and careful messaging by stakeholders to mitigate reduced public confidence in COVID-19 vaccines [2]. Moreover, official communication efforts had to consider the potential influence of already widespread misinformation about the vaccines on attitudes toward the pause [3,4]. In this survey study, we evaluated understanding and impressions of the CDC's public web-based information about the J&J-Janssen pause among unvaccinated US adults.

Methods

Web-Based Survey About J&J-Janssen Pause

We administered the web-based survey to two cohorts of US adults recruited through Prolific between April 19-21, 2021 (cohort A), and April 21-23, 2021 (cohort B). Both cohorts were assembled using convenience sampling of unvaccinated adults. To obtain information about a population that especially needed targeted vaccine communication, the first cohort was restricted to individuals expressing neutral or negative sentiments about COVID-19 vaccines. The survey design and recruitment strategy are described in Multimedia Appendix 1; the survey questions are provided in Multimedia Appendices 2 and 3.

Ethical Considerations

The study was approved by Harvard University's Committee on the Use of Human Subjects (IRB20-2089), and participants

XSL•FO
RenderX

agreed to a consent statement on the first page of the survey. Participants were paid US $2 for taking the survey. All study data were collected anonymously.

Results

A total of 271 and 286 participants were included in cohorts A and B, respectively (demographic characteristics listed in Table 1). Across participants, the median number of correct responses to the comprehension questions was 6 in both cohort A (IQR 1.5; range 0-7) and cohort B (IQR 1.0; range 1-7). The total number of correct responses was negatively associated with intention not to seek vaccination in both cohort A (odds ratio 0.61, 95% CI 0.45-0.82; *P*=.001) and cohort B (odds ratio 0.48, 95% CI 0.31-0.74; *P*=.001; Multimedia Appendix 4). Although a majority of participants rated the passages as "clear and easy to read" (cohort A: n=229, 84.5%; cohort B: n=243, 85%), fewer indicated that they would be likely to share them on social media (cohort A: n=53, 19.6%; cohort B: n=75, 26.3%).

The web page mentioned "a small number of reports" of cerebral venous sinus thrombosis in individuals who received the J&J-Janssen vaccine. When asked to guess a specific number, 188 (69.4%) and 133 (46.5%) respondents in cohorts A and B, respectively, estimated 100 or more cases, at least an order of magnitude higher than the actual value; 176 (64.9%) and 128 (44.8%) respondents in cohorts A and B, respectively, estimated 10 or more deaths after vaccination (Figure 1).

Responding to a counterfactual question, 127 (46.9%) and 139 (48.6%) participants in cohorts A and B, respectively, indicated that the pause reduced their confidence in the J&J-Janssen vaccine's safety (Figure 1). Most participants reported no change in their confidence in COVID-19 vaccines' safety in general (cohort A: n=182, 67.2%; cohort B: n=194, 67.8%) or intention to receive the Pfizer-BioNTech or Moderna vaccine (cohort A: n=206, 76%; cohort B: n=211, 73.8%).

XSL•FO
RenderX

Table 1. Characteristics and responses of participants who completed the web-based surveys about the Johnson & Johnson (J&J)–Janssen vaccine pause.

Characteristics	Cohort A (n=271), n (%)	Cohort B (n=286), n (%)
Age (years)		
18-29	119 (43.9)	145 (50.7)
30-49	116 (42.8)	111 (38.9)
≥50	36 (13.3)	30 (10.5)
Gender		
Female	144 (53.1)	136 (47.6)
Male	124 (45.8)	144 (50.3)
Nonbinary, transgender, or other	3 (1.1)	6 (2.1)
Ethnicity: Hispanic or Latinx	54 (20.0)	39 (13.6)
Race[a]		
Asian	29 (10.7)	96 (33.6)
Black or African American	70 (25.8)	37 (12.9)
White	160 (59.0)	139 (48.6)
Other[b]	27 (10.0)	28 (9.8)
Educational attainment		
Less than high school diploma	3 (1.1)	6 (2.1)
High school diploma or equivalent	54 (19.9)	46 (16.1)
Some college or associate's degree	102 (37.6)	106 (37.1)
Bachelor's degree	81 (29.9)	102 (35.7)
Graduate or professional degree	31 (11.4)	26 (9.1)
Political partisanship		
Democratic (including leaners)	81 (29.9)	150 (52.4)
Republican (including leaners)	91 (33.6)	42 (14.7)
Independent or other	99 (36.5)	94 (32.9)
Geography		
Rural area	41 (15.1)	35 (12.2)
Suburban or urban area	230 (84.9)	251 (87.8)
Intention to receive vaccine		
Definitely will not	71 (26.2)	29 (10.1)
Probably will not	68 (25.1)	17 (5.9)
Undecided	54 (19.9)	29 (10.1)
Definitely or probably will	78 (28.8)	211 (73.8)
Comprehension questions[c]		
Reason for pause	259 (95.6)	280 (97.9)
Causal relationship between vaccine and side effect	196 (72.3)	215 (75.2)
Affected population	234 (86.3)	248 (86.7)
Safety of mRNA vaccines	214 (79.0)	250 (87.4)
Revaccination of J&J-Janssen recipients	164 (60.5)	189 (66.1)
Symptom monitoring for J&J-Janssen recipients	180 (66.4)	121 (42.3)
Rescheduling of canceled appointments	192 (70.8)	229 (80.1)

XSL•FO
RenderX

Characteristics	Cohort A (n=271), n (%)	Cohort B (n=286), n (%)
Comprehension score[d]		
7	62 (22.9)	43 (15.0)
6	80 (29.5)	119 (41.6)
5	61 (22.5)	59 (20.6)
4	33 (12.2)	38 (13.3)
3	24 (8.9)	18 (6.3)
0-2	11 (4.1)	9 (3.1)
Self-reported impressions of passage[e]		
Accurate and should be trusted	119 (43.9)	198 (69.2)
High-quality evidence	110 (40.6)	180 (62.9)
Clear and easy to read	229 (84.5)	243 (85.0)
Understandable	229 (84.5)	251 (87.8)
Requiring great effort to understand	58 (21.4)	69 (24.1)
Other people would want to read	173 (63.8)	172 (60.1)
Would share on social media	53 (19.6)	75 (26.2)

[a]Participants could select more than one option.

[b]Includes participants who selected “American Indian or Alaska Native,” “Native Hawaiian or Other Pacific Islander,” or “Another option not listed here.”

[c]Number of participants who gave the correct answer to each question.

[d]Number of participants who gave the indicated number of correct answers across all questions.

[e]Number of participants who answered “Strongly agree” or “Agree” about each description on a 6-point Likert scale.

XSL•FO
RenderX

Figure 1. Self-reported attitude changes and estimated cases and deaths. J&J: Johnson & Johnson.

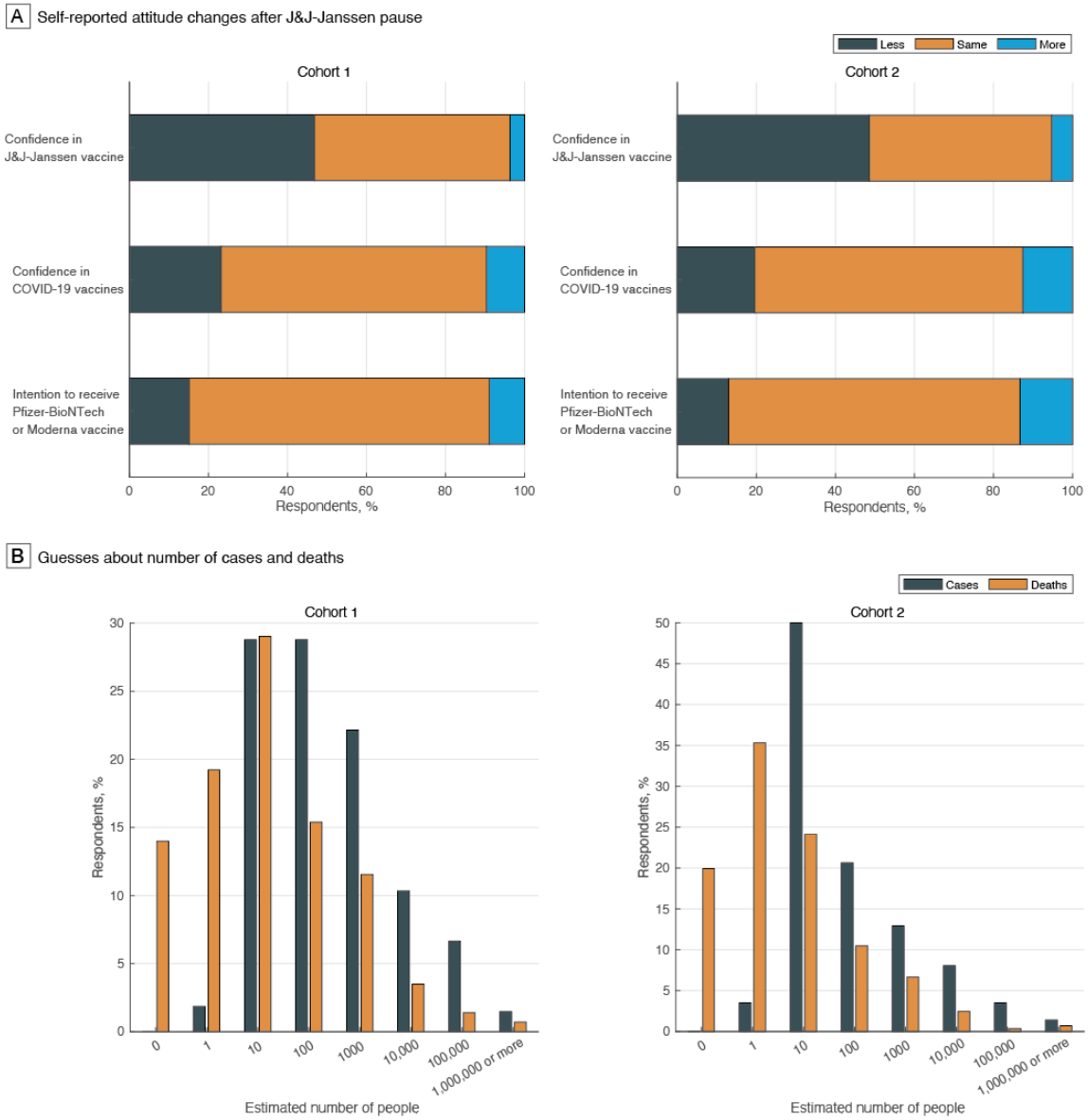

Discussion

In our web-based survey about the CDC's messaging around the J&J-Janssen vaccine pause, many respondents overestimated the number of case reports that prompted the pause, often by several orders of magnitude. Since verbal descriptors are elastic concepts that can be misinterpreted, grounding them with numbers can reduce variability in risk perception and promote informed decision-making [5].

Respondents also expressed reduced confidence in the safety of the J&J-Janssen vaccine, highlighting the potential danger of conveying piecemeal information about risk during a pandemic response [3]. Encouragingly, the reduced confidence did not extend to mRNA COVID-19 vaccines, consistent with previous findings that overall vaccine hesitancy remained stable following the pause [6]. These results were obtained using the counterfactual format, which is less susceptible to overestimating shifts in beliefs than the change format (Multimedia Appendix 1). The negative association between understanding of the passage and self-reported vaccine hesitancy suggests that more targeted messaging may have been useful to promote vaccine confidence [7,8].

Consistent with uncertainty management theory [9], individuals likely viewed the pause in different ways, leading to a spectrum of emotional responses and changes in behavior. Despite being a safety precaution, the pause introduced new uncertainties requiring effective management through clear and consistent messaging, highlighting the balance that must be maintained between fostering trust and preventing unnecessary alarm [10]. Given the limitations of the deficit model of scientific communication [11], just providing technically correct information is insufficient for effective communication during public health crises. Instead, attention should be given to the accessibility of information across diverse socioeconomic groups, in line with the knowledge gap hypothesis [12], and to countering misinformation by providing easy-to-use official guidance [6,7].

The study is limited by the convenience sampling strategy; the participants recruited were not representative of the US population as a whole, and the findings should not be

XSL•FO
RenderX

generalized to other contexts. Since the study was conducted on the web, individuals with lower internet and health literacy may have been excluded.

Acknowledgments

We thank Vasudha Mishra, MBBS, for assistance with graphic design. This work was supported by a CoronaVirusFacts Alliance Grant from the Poynter Institute, a Harvard Data Science Fellowship, and the Institute of Collaborative Innovation at the University of Macau.

Data Availability

The data sets generated during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information about survey methodology.
[PDF File (Adobe PDF File), 99 KB - jmir_v26i1e41559_app1.pdf]

Multimedia Appendix 2
Survey administered to cohort A.
[PDF File (Adobe PDF File), 107 KB - jmir_v26i1e41559_app2.pdf]

Multimedia Appendix 3
Survey administered to cohort B.
[PDF File (Adobe PDF File), 99 KB - jmir_v26i1e41559_app3.pdf]

Multimedia Appendix 4
Supplemental tables about survey questions and ordinal logistic regression analysis.
[PDF File (Adobe PDF File), 90 KB - jmir_v26i1e41559_app4.pdf]

References

1. Karron RA, Key NS, Sharfstein JM. Assessing a rare and serious adverse event following administration of the Ad26.COV2.S vaccine. JAMA 2021 Jul 22;325(24):2445-2447. [doi: 10.1001/jama.2021.7637] [Medline: 33929484]
2. Wood S, Schulman K. Beyond politics—promoting Covid-19 vaccination in the United States. N Engl J Med 2021 Mar 18;384(7):e23. [doi: 10.1056/NEJMms2033790] [Medline: 33406324]
3. Lee JJ, Kang K, Wang MP, Zhao SZ, Wong JYH, O'Connor S, et al. Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. J Med Internet Res 2020 Dec 13;22(11):e22205 [FREE Full text] [doi: 10.2196/22205] [Medline: 33048825]
4. Ginossar T, Cruickshank IJ, Zheleva E, Sulskis J, Berger-Wolf T. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations. Hum Vaccin Immunother 2022 Dec 31;18(1):1-13 [FREE Full text] [doi: 10.1080/21645515.2021.2003647] [Medline: 35061560]
5. Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. BMJ 2002 May 06;324(7341):827-830 [FREE Full text] [doi: 10.1136/bmj.324.7341.827] [Medline: 11934777]
6. Hsieh YL, Rak S, SteelFisher GK, Bauhoff S. Effect of the suspension of the J&J COVID-19 vaccine on vaccine hesitancy in the United States. Vaccine 2022 Jan 24;40(3):424-427 [FREE Full text] [doi: 10.1016/j.vaccine.2021.11.085] [Medline: 34903374]
7. Miller M, Castrucci BC. Changing the COVID-19 conversation: it's about language. JAMA Health Forum 2021 Mar 01;2(2):e210020 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.0020] [Medline: 36218788]
8. Mishra V, Dexter JP. Comparison of readability of official public health information about COVID-19 on websites of international agencies and the governments of 15 countries. JAMA Netw Open 2020 Aug 03;3(8):e2018033 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.18033] [Medline: 32809028]
9. Brashers D. Communication and uncertainty management. J Commun 2001;51(3):477-497. [doi: 10.1111/j.1460-2466.2001.tb02892.x]

XSL•FO
RenderX

10. Petersen MB, Bor A, Jørgensen F, Lindholt MF. Transparent communication about negative features of COVID-19 vaccines decreases acceptance but increases trust. Proc Natl Acad Sci U S A 2021 Jul 20;118(29):e2024597118 [FREE Full text] [doi: 10.1073/pnas.2024597118] [Medline: 34292869]
11. Reincke CM, Bredenoord AL, van Mil MH. From deficit to dialogue in science communication: the dialogue communication model requires additional roles from scientists. EMBO Rep 2020 Oct 03;21(9):e51278 [FREE Full text] [doi: 10.15252/embr.202051278] [Medline: 32748995]
12. Lind F, Boomgaarden HG. What we do and don't know: a meta-analysis of the knowledge gap hypothesis. Ann Int Commun Assoc 2019 May 09;43(3):210-224. [doi: 10.1080/23808985.2019.1614475]

Abbreviations

CDC: Centers for Disease Control and Prevention
J&J: Johnson & Johnson

Edited by A Mavragani; submitted 25.08.22; peer-reviewed by M Graham, T Ginossar, A Scherer; comments to author 25.01.23; revised version received 26.05.23; accepted 29.12.23; published 01.04.24.

Please cite as:
Mishra V, Dexter JP
Response of Unvaccinated US Adults to Official Information About the Pause in Use of the Johnson & Johnson–Janssen COVID-19 Vaccine: Cross-Sectional Survey Study
J Med Internet Res 2024;26:e41559
URL: https://www.jmir.org/2024/1/e41559
doi:10.2196/41559
PMID:38557597

©Vishala Mishra, Joseph P Dexter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Evaluation of Prompts to Simplify Cardiovascular Disease Information Generated Using a Large Language Model: Cross-Sectional Study

Vishala Mishra[1*], MBBS, MMCi; Ashish Sarraju[2], MD; Neil M Kalwani[3,4], MD, MPP; Joseph P Dexter[5,6,7*], PhD

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, United States

[2]Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH, United States

[3]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, United States

[4]Division of Cardiovascular Medicine and the Cardiovascular Institute, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

[5]Data Science Initiative, Harvard University, Allston, MA, United States

[6]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, United States

[7]Institute of Collaborative Innovation, University of Macau, Taipa, Macao

[*]these authors contributed equally

Corresponding Author:
Joseph P Dexter, PhD
Data Science Initiative
Harvard University
Science and Engineering Complex 1.312-10
150 Western Avenue
Allston, MA, 02134
United States
Phone: 1 8023381330
Email: jdexter@fas.harvard.edu

Abstract

In this cross-sectional study, we evaluated the completeness, readability, and syntactic complexity of cardiovascular disease prevention information produced by GPT-4 in response to 4 kinds of prompts.

(J Med Internet Res 2024;26:e55388) doi:10.2196/55388

KEYWORDS

artificial intelligence; ChatGPT; GPT; digital health; large language model; NLP; language model; language models; prompt engineering; health communication; generative; health literacy; natural language processing; patient-physician communication; health communication; prevention; cardiology; cardiovascular; heart; education; educational; human-in-the-loop; machine learning

Introduction

Many web-based patient educational materials about cardiovascular disease (CVD) are inaccessible for the general public [1]. Artificial intelligence (AI) chatbots powered by large language models (LLMs) are a potential source of public-facing CVD information [2-4]. Generative language models present risks related to information quality but also opportunities for producing accessible information about CVD at scale, which could advance the American Heart Association's 2020 impact goals related to health literacy [5]. Recent studies have used LLMs to simplify medical information in different contexts [3,6-8], but quantitative comparison of prompt engineering strategies is needed to assess and optimize performance and to ensure that the rapid deployment of clinical AI tools proceeds in an equitable manner [9]. In this cross-sectional study, we evaluated the completeness, readability, and syntactic complexity of CVD prevention information produced by GPT-4 in response to 4 kinds of prompts.

Methods

A set of 25 questions about fundamental CVD prevention topics was drawn from a previous study, which found that the GPT 3.5 version of ChatGPT provided generally appropriate responses [2]. We devised 3 prompt strategies for generating simplified ChatGPT responses to these questions, including a

XSL•FO
RenderX

zero-shot prompt to use plain and easy-to-understand language, a one-shot prompt with a sample simplified passage on an unrelated subject, and a combined prompt to use simplified language and cover specific key points (which we termed "rubric prompting"; Multimedia Appendix 1). Responses to these three prompts were compared to baseline responses for which the prompt contained only the question about CVD. The full set of responses is provided in Multimedia Appendix 2.

For each question and prompt type, 3 independent responses were generated between April and June 2023, using the GPT-4 version of ChatGPT with default parameters, which was available from OpenAI through a ChatGPT Plus subscription. Two authors, who are preventive cardiologists (AS and NWK), scored the responses as "complete," "incomplete," or "inconsistent" according to a custom rubric (Multimedia Appendix 3); disagreements were resolved by consensus. For all generated responses, we calculated 5 readability scores, using Readability Studio Professional (version 2019.3; Oleander Software), and 2 measures of syntactic complexity, using the L2 Syntactic Complexity Analyzer (version 3.3.3), as described previously [10].

Differences from baseline completeness were assessed using the Fisher exact test, and 2-sample readability and syntactic complexity comparisons were done using the Wilcoxon rank-sum test. Statistical significance was set as $P<.05$.

Results

Baseline responses to 80% (20/25) of the questions were scored as "complete" (Table 1). Completeness was significantly lower for both the zero-shot (8/25, 32%; $P=.001$) and one-shot (8/25, 32%; $P=.001$) simplification prompts but significantly higher for the rubric prompts (25/25, 100%; $P=.001$). All 3 prompts significantly improved readability according to every metric and lowered 1 measure of syntactic complexity (Table 2).

XSL•FO
RenderX

Table 1. Evaluation of the completeness of cardiovascular disease information generated using 4 large language model prompt strategies.

Question	Consensus grade for each prompt[a]			
	Baseline	Plain language (zero-shot prompt)	Plain language (one-shot prompt)	Plain language (rubric prompt)
How can I prevent heart disease?	Complete	Complete	Complete	Complete
What is the best diet for the heart?	Complete	Complete	Complete	Complete
What is the best diet for high blood pressure and high cholesterol?	Complete	Complete	Complete	Complete
How much should I exercise to stay healthy?	Complete	Inconsistent	Incomplete	Complete
Should I do cardio or lift weights to prevent heart disease?	Complete	Inconsistent	Inconsistent	Complete
How can I lose weight?	Complete	Inconsistent	Inconsistent	Complete
How can I decrease LDL[b]?	Inconsistent	Incomplete	Incomplete	Complete
How can I decrease triglycerides?	Complete	Complete	Complete	Complete
What is lipoprotein(a)?	Complete	Incomplete	Incomplete	Complete
How can I quit smoking?	Complete	Complete	Inconsistent	Complete
What are the side effects of statins?	Complete	Inconsistent	Complete	Complete
I have muscle pain with a statin. What should I do?	Inconsistent	Inconsistent	Complete	Complete
My cholesterol is still high and I'm already on a statin. What should I do?	Inconsistent	Incomplete	Incomplete	Complete
What medications can reduce cholesterol other than statins?	Complete	Complete	Inconsistent	Complete
What is ezetimibe?	Complete	Inconsistent	Incomplete	Complete
What are Repatha and Praluent?	Complete	Incomplete	Incomplete	Complete
What is inclisiran?	Complete	Incomplete	Incomplete	Complete
What are the side effects of Repatha and Praluent?	Complete	Complete	Inconsistent	Complete
Should I take aspirin to prevent heart disease?	Complete	Complete	Complete	Complete
My cholesterol panel shows triglycerides 400 mg/dL. How should I interpret this?	Complete	Inconsistent	Complete	Complete
My LDL is 200 mg/dL. How should I interpret this?	Inconsistent	Incomplete	Incomplete	Complete
What does a coronary calcium score of 0 mean?	Complete	Incomplete	Incomplete	Complete
What does a coronary calcium score of 100 mean?	Inconsistent	Inconsistent	Incomplete	Complete
What does a coronary calcium score of 400 mean?	Complete	Incomplete	Incomplete	Complete
What genetic mutations can cause high cholesterol?	Complete	Inconsistent	Incomplete	Complete

[a]For every prompt strategy, we generated 3 responses to each of the 25 questions about cardiovascular disease prevention. "Complete" indicates that all 3 responses received a full score according to our coverage rubric, "Incomplete" indicates that all 3 responses received less than a full score, and "Inconsistent" indicates that some responses were "Complete" and others were "Incomplete." Grades shown were determined by consensus between 2 reviewers.

[b]LDL: low-density lipoprotein.

XSL•FO
RenderX

Table 2. Comparison of the readability and syntactic complexity of cardiovascular disease information generated using 4 large language model prompt strategies.[a]

	Prompts						
	Baseline, median (IQR)	Plain language (zero-shot prompt)		Plain language (one-shot prompt)		Plain language (rubric prompt)	
		Value, median (IQR)	Difference from baseline[b], median (IQR; *P* value)	Value, median (IQR)	Difference from baseline[c], median (IQR; *P* value)	Value, median (IQR)	Difference from baseline[d], median (IQR; *P* value)
Readability formulas							
FKGL[e]	13.4 (12.3 to 15.4)	9.7 (7.6 to 11.1)	−4.2 (−5.7 to −3.1; <.001)	3.8 (2.9 to 5.3)	−9.4 (−11.1 to −8.3; <.001)	8.0 (7.3 to 9.5)	−5.3 (−6.6 to −4.0; <.001)
SMOG[f]	14.8 (13.7 to 16.5)	12.1 (10.2 to 13)	−3.6 (−4.5 to −2.4; <.001)	7.9 (7.2 to 9.2)	−7.1 (−8.2 to −5.7; <.001)	10.9 (10.4 to 11.9)	−4.1 (−5.4 to −3.0; <.001)
GFI[g]	14.0 (12.1 to 17)	11.3 (8.0 to 13)	−4.0 (−5.6 to −2.7; <.001)	6.3 (5.4 to 7.6)	−7.5 (−10.3 to −6.0; <.001)	10.2 (8.9 to 11.3)	−3.9 (−6.3 to −2.8; <.001)
FORCAST[h]	11.5 (11.2 to 11.9)	10.2 (9.8 to 10.7)	−1.3 (−1.8 to −0.9; <.001)	8.8 (8.2 to 9.4)	−2.7 (−3.4 to −2.3; <.001)	9.7 (9.3 to 10.2)	−1.9 (−2.3 to −1.4; <.001)
CLI[i]	13.8 (13.2 to 15.1)	10.4 (9.0 to 11.8)	−3.7 (−4.7 to −2.4; <.001)	6.2 (5.1 to 7.3)	−7.9 (−9.0 to −6.5; <.001)	9.4 (9.0 to 10.4)	−4.5 (−5.4 to −3.5; <.001)
Syntactic complexity [j]							
MLC[k]	15.0 (12.7 to 16.6)	12.3 (10.5 to 15.5)	−1.8 (−4.4 to 0.9; .01)	8.7 (7.8 to 10.7)	−5.7 (−7.6 to −3.4; <.001)	9.6 (8.9 to 10.3)	−4.2 (−6.9 to −3.1; <.001)
DC/T[l]	0.3 (0.2 to 0.5)	0.3 (0.2 to 0.5)	0 (−0.2 to 0.1; .36)	0.2 (0.1 to 0.3)	−0.2 (−0.3 to −0.1; <.001)	0.6 (0.4 to 0.7)	0.2 (0.1 to 0.4; >.99)

[a]For every prompt strategy, we generated 3 responses to each of the 25 questions about cardiovascular disease prevention. Lower scores indicate higher readability.

[b]Difference between responses to the baseline prompts and prompts for plain language. *P* values are from a 1-tailed Wilcoxon signed rank test.

[c]Difference between responses to the baseline prompts and prompts for plain language with an example. *P* values are from a 1-tailed Wilcoxon signed rank test.

[d]Difference between responses to the baseline prompts and prompts for plain language with coverage. *P* values are from a 1-tailed Wilcoxon signed rank test.

[e]FKGL: Flesch-Kincaid Grade Level.

[f]SMOG: Simple Measure of Gobbledygook.

[g]GFI: Gunning Fog Index.

[h]FORCAST: Ford, Caylor, Sticht formula.

[i]CLI: Coleman-Liau Index.

[j]MLC is a measure of elaboration at the clause level (ie, number of words per clause), and DC/T is a measure of subordination.

[k]MLC: mean length of clause.

[l]DC/T: dependent clauses/T-unit.

Discussion

We found that zero- and one-shot prompting of GPT-4 to produce simplified information about CVD generated more readable but less comprehensive responses. This loss of information, however, could be averted by combining a zero-shot simplification prompt with a short reminder to include critical information (rubric prompting). Our findings highlight the importance of optimizing prompts and incorporating expert clinical judgment when considering the use of LLMs to produce patient education materials, including AI-drafted replies to patient messages [3,6,7]. Accordingly, prospective guidelines for the use of AI in medicine should address best practices for prompt engineering, standardized evaluation of model outputs, and outreach to clinicians and the public to cultivate relevant skills [11]. Such guidelines will provide important parameters for clinician-in-the-loop information simplification systems [6,12,13], which have already been deployed to improve the accessibility of surgical consent forms [14].

The limitations of this study include the evaluation of a single model at a specific point in time and the absence of reading comprehension data from patients. Since the prompt strategies developed herein are not model specific, it should be straightforward to extend these strategies to other LLMs. Future research should further evaluate trade-offs between prompt engineering and fine-tuning of LLMs for medical applications using multiple models. It would also be useful to integrate ongoing user testing with structured health literacy assessment

XSL•FO
RenderX

of generated responses to identify types of simplification that are especially important for improving patient understanding.

Acknowledgments

We thank Stephen Blackwelder, PhD (Duke University Health System), for helpful discussions and comments on the manuscript and Vasudha Mishra, MBBS (AIIMS Patna), for assistance with data collection. JPD was supported by a Harvard Data Science Fellowship and the Institute of Collaborative Innovation at the University of Macau. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Authors' Contributions

VM, AS, and JPD designed the study. VM and JPD generated the ChatGPT responses and performed the computational and statistical analyses. AS and NWK performed the completeness scoring. VM and JPD wrote the manuscript. All authors edited and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example prompt types.
[DOCX File , 13 KB - jmir_v26i1e55388_app1.docx]

Multimedia Appendix 2
Full ChatGPT responses.
[DOCX File , 193 KB - jmir_v26i1e55388_app2.docx]

Multimedia Appendix 3
Custom scoring rubric.
[DOCX File , 4569 KB - jmir_v26i1e55388_app3.docx]

References

1. Pearson K, Ngo S, Ekpo E, Sarraju A, Baird G, Knowles J, et al. Online patient education materials related to lipoprotein(a): readability assessment. J Med Internet Res 2022 Jan 11;24(1):e31284 [FREE Full text] [doi: 10.2196/31284] [Medline: 35014955]
2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
3. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
4. Sarraju A, Ouyang D, Itchhaporia D. The opportunities and challenges of large language models in cardiology. JACC Adv 2023 Sep;2(7):100438 [FREE Full text] [doi: 10.1016/j.jacadv.2023.100438]
5. Magnani JW, Mujahid MS, Aronow HD, Cené CW, Dickson VV, Havranek E, American Heart Association Council on Epidemiology and Prevention; Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; Council on Peripheral Vascular Disease; Council on Quality of Care and Outcomes Research; and Stroke Council. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018 Jul 10;138(2):e48-e74 [FREE Full text] [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
6. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023 May 18;6(1):9 [FREE Full text] [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
7. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. AJR Am J Roentgenol 2023 Nov;221(5):701-704. [doi: 10.2214/AJR.23.29622] [Medline: 37341179]
8. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 05;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
9. Singh N, Lawrence K, Richardson S, Mann DM. Centering health equity in large language model deployment. PLOS Digit Health 2023 Oct 24;2(10):e0000367 [FREE Full text] [doi: 10.1371/journal.pdig.0000367] [Medline: 37874780]

XSL•FO
RenderX

10. Mishra V, Dexter JP. Comparison of readability of official public health information about COVID-19 on websites of international agencies and the governments of 15 countries. JAMA Netw Open 2020 Aug 03;3(8):e2018033 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.18033] [Medline: 32809028]
11. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
12. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. J Med Internet Res 2024 Jan 22;26:e51926 [FREE Full text] [doi: 10.2196/51926] [Medline: 38252483]
13. Chen S, Li Y, Lu S, Van H, Aerts HJWL, Savova GK, et al. Evaluating the ChatGPT family of models for biomedical reasoning and classification. J Am Med Inform Assoc 2024 Apr 03;31(4):940-948. [doi: 10.1093/jamia/ocad256] [Medline: 38261400]
14. Mirza FN, Tang OY, Connolly ID, Abdulrazeq HA, Lim RK, Roye GD, et al. Using ChatGPT to facilitate truly informed medical consent. NEJM AI 2024 Jan 10;1(2):AIcs2300145. [doi: 10.1056/aics2300145]

Abbreviations

AI: artificial intelligence
CVD: cardiovascular disease
LLM: large language model

Edited by T de Azevedo Cardoso; submitted 11.12.23; peer-reviewed by R Mpofu; comments to author 12.01.24; revised version received 25.01.24; accepted 31.01.24; published 22.04.24.

Please cite as:
Mishra V, Sarraju A, Kalwani NM, Dexter JP
Evaluation of Prompts to Simplify Cardiovascular Disease Information Generated Using a Large Language Model: Cross-Sectional Study
J Med Internet Res 2024;26:e55388
URL: https://www.jmir.org/2024/1/e55388
doi:10.2196/55388
PMID:38648104

©Vishala Mishra, Ashish Sarraju, Neil M Kalwani, Joseph P Dexter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties

Felix Busch[1], MD; Tianyu Han[2], PhD; Marcus R Makowski[3], MD, PhD; Daniel Truhn[2], MSc, MD; Keno K Bressem[4*], MD; Lisa Adams[3*], MD

[1]Department of Neuroradiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[2]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

[3]Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

[4]Institute for Radiology and Nuclear Medicine, German Heart Center Munich, Technical University of Munich, Munich, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Busch, MD
Department of Neuroradiology
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin
Charitépl. 1
Berlin, 10117
Germany
Phone: 49 3045050
Email: felix.busch@charite.de

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e64411

Abstract

This study demonstrates that GPT-4V outperforms GPT-4 across radiology subspecialties in analyzing 207 cases with 1312 images from the Radiological Society of North America Case Collection.

(J Med Internet Res 2024;26:e54948) doi:10.2196/54948

KEYWORDS

GPT-4; ChatGPT; Generative Pre-Trained Transformer; multimodal large language models; artificial intelligence; AI applications in medicine; diagnostic radiology; clinical decision support systems; generative AI; medical image analysis

Introduction

The launch of GPT-4 has generated significant interest in the scientific and medical communities, demonstrating its potential in medicine with notable achievements such as an 83.76% zero-shot accuracy on the United States Medical Licensing Examination (USMLE) [1]. In radiology, GPT has spanned text-based tasks, including board exam question scoring, data mining, and report structuring [2,3]. The recent release of GPT-4's visual capabilities (GPT-4V) enables the combined analysis of text and visual data [4]. Our study focuses on evaluating the diagnostic capabilities of GPT-4V by comparing it to GPT-4 in advanced radiological tasks, benchmarking the potential of this multimodal large language model in the medical imaging field.

Methods

We sourced 207 cases with 1312 images from the Radiological Society of North America (RSNA) Case Collection (accessible for RSNA members on the RSNA Case Collection website [5]), aiming to cover at least 10 cases for each of the 22 presented subspecialties. The cases within each subspeciality were chosen to present different pathologies. Each case had varying numbers of images and were usually labeled for more than 1 subspecialty, so that the total number of cases per subspeciality varied

XSL•FO
RenderX

between 1 (for “Physics and Basic Science,” no more than 1 case was available) and 43 (for “Gastrointestinal,” 10 cases in this category were chosen, with 33 additional cases from other subspecialties that were also labeled for “Gastrointestinal”).

GPT-4 and GPT-4V were accessed between November 6, 2023, and November 17, 2023. We utilized an application programming interface (API) account, which allowed us to use the models programmatically and ensure a consistent environment for each test. This access level was crucial, as it provided stable and repeatable interactions with the models, unlike what might be experienced with fluctuating conditions of regular account usage. The ground truth was established based on the final diagnoses stated in the RSNA case entries. We prompted each model 3 times via the API for the following two tasks: first, the models were asked to identify the diagnosis and 2 differentials (providing the patient history only for GPT-4 or patient history with images for GPT-4V); second, the models were asked to answer corresponding multiple-choice questions from the RSNA Case Collection. The GPT-4V assessment used a “chain-of-thought” prompt that guided the model through diagnostic reasoning (Figure 1), in contrast to the text-only assessment of GPT-4. For both tasks, a case was considered correctly diagnosed if the same correct result appeared for at least 2 of 3 prompts. Cases with no repeated correct diagnoses and cases with only false diagnoses across the 3 prompts were marked as incorrectly diagnosed. Mean accuracies and bootstrapped 95% CIs were calculated, and statistical significance was determined by using the McNemar test (P<.001).

Figure 1. An example conversation with GPT-4V showcasing the prompting style that was used if the question contained more than 4 images. Notably, GPT-4V often disregards the initial textual case description when additional image prompts are introduced, necessitating the description’s repetition for accurate responses. As the context window, image resolution limit, and embedding size limit of the web application version of GPT-4V are unknown, definitive conclusions regarding the cause of this forgetfulness could not be drawn. Nevertheless, the model’s consistent ability to recognize and recall images from the initial prompt (eg, axial FLAIR [fluid-attenuated inversion recovery] images) suggests that running out of context length is an unlikely explanation.

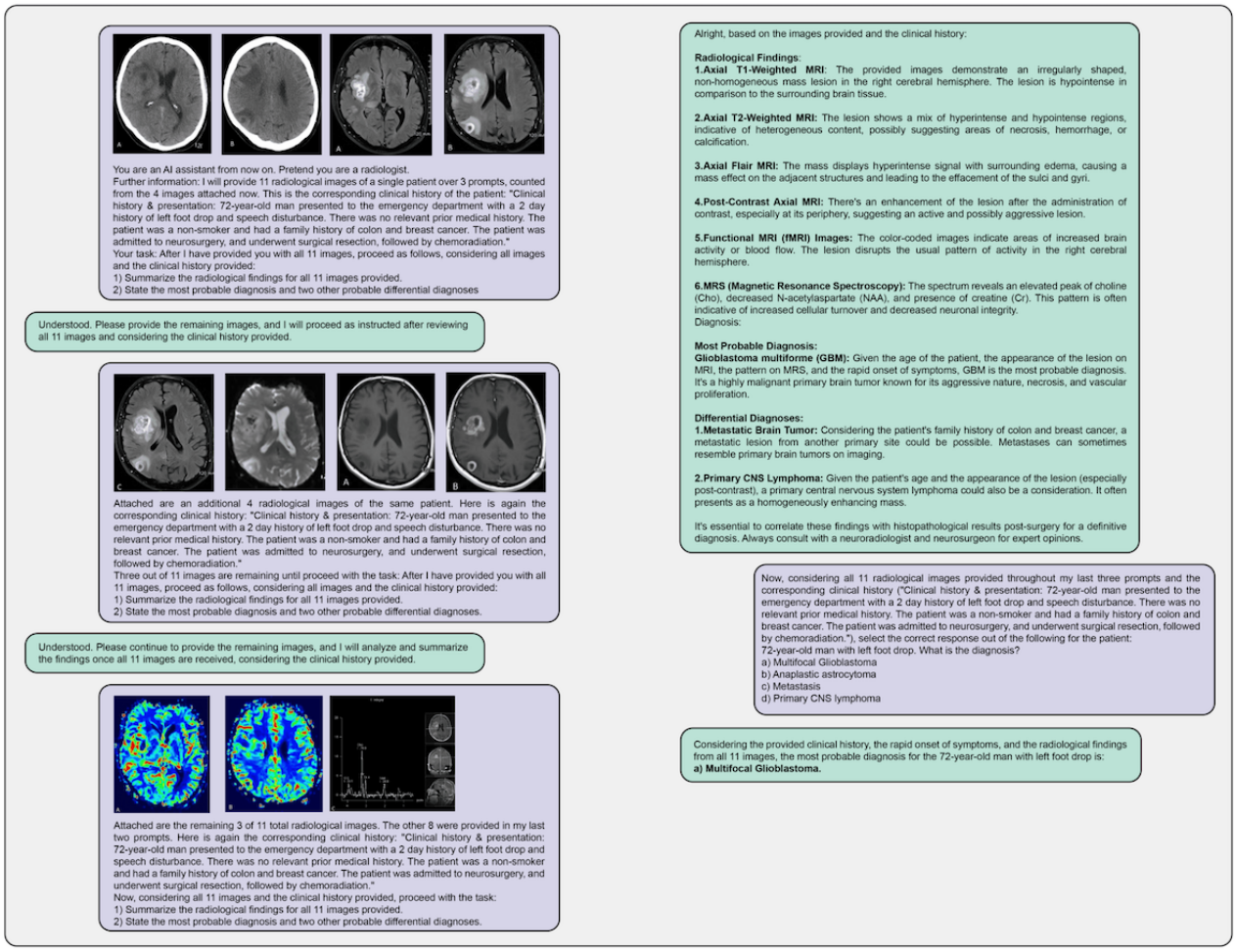

Results

GPT-4 accurately identified the primary diagnosis in 18% (95% CI 12%-25%) of cases (first task). When including differential diagnoses, this accuracy increased to 28% (95% CI 22%-33%). In contrast, GPT-4V achieved a 27% (95% CI 21%-34%) accuracy rate for primary diagnosis, which increased to 35% (95% CI 29%-40%) when differential diagnoses were accounted for. After being presented with multiple-choice questions, including information about clinical history and presentation (second task), GPT-4 achieved an accuracy of 47% (95% CI 42%-56%). Again, GPT-4V demonstrated a higher accuracy of 64% (95% CI 59%-72%). The observed difference in

XSL•FO
RenderX

performance was statistically significant ($P<.001$). Across 15 subspecialties, GPT-4V outperformed GPT-4, with the sole exception being in "Cardiac Imaging." Figure 2 summarizes the accuracies across all subspecialties.

Figure 2. Comparison of GPT-4 and GPT-4V in various radiology subspecialties. Many cases spanned multiple subspecialties, and some subspecialties had very few cases. The number of images for individual cases ranged from 2 to 30, and the overall accuracy across all subspecialties, as shown in the bar plot, showed that GPT-4V performed significantly better than GPT-4. Error bars represent the 95% CIs. The radar plot shows the accuracy of GPT-4 (green line) and GPT-4V (purple line) across different radiology subspecialties. Each axis represents a specific radiology subspecialty, with the percentages indicating the accuracy of the model in that domain. Both models show varying levels of performance across subspecialties, with GPT-4V consistently performing better than GPT-4, except in "Cardiac Imaging" (cases: n=14; GPT-4V accuracy: 36%; GPT-4 accuracy: 57%). For "Physics and Basic Science" (cases: n=1), "Breast Imaging" (cases: n=10), and "Obstetrics/Gynecology" (cases: n=12), GPT-4V and GPT-4 showed on-par performance (accuracy: 100%, 50%, and 58%, respectively). Due to the small sample sizes in some categories, which ranged from 1 to 43 cases, these results should primarily be viewed as indicative trends rather than definitive conclusions about the models' performance in these specific areas.

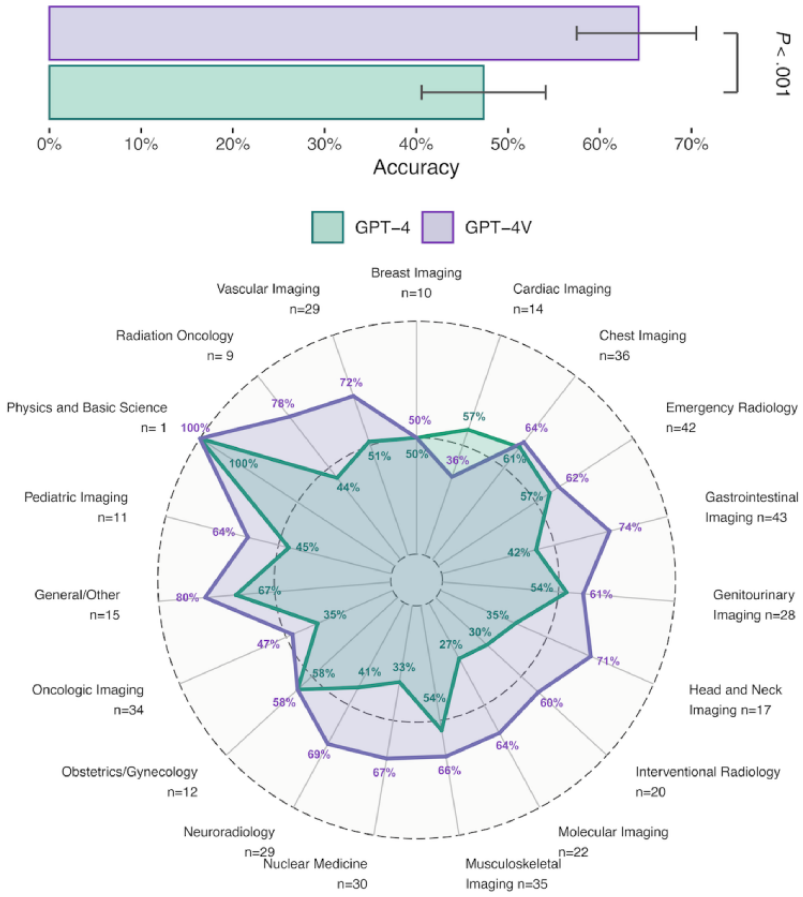

Discussion

Our study shows that GPT-4V has improved performance over GPT-4 in solving complex radiological problems, indicating its potential to detect pathological features in medical images and thus its radiological domain knowledge. The RSNA Case Collection, which is aimed at expert-level professional radiologists, highlights the promise of GPT-4V in specialized medical contexts.

However, the use of GPT-4V warrants a cautious approach. At this time, it should be considered, at best, as a supplemental tool to augment—not replace—the comprehensive analyses performed by trained medical professionals.

Extending the initial research by Yang et al [6], our study explores the medical image analysis capabilities of GPT-4V in more complex scenarios and with a wider range of cases. The ongoing development of multimodal models, such as Med-Flamingo, for medical applications signals a growing interest in this area [7].

One challenge is the scarcity of specialized medical data sets. As our study used RSNA member–exclusive cases, it was unlikely that these cases were in GPT-4V's training data; thus, the risk of data contamination was minimized. However, the corresponding images for each case were indented to highlight specific pathologies, and this does not fully replicate clinical practice, where one would have to analyze each separate image to identify potential pathologies—a task that specialized deep learning models would be better suited to perform.

Future efforts should focus on detailed performance comparisons between generalist models (like GPT-4V) and emerging, radiological domain–specialized, artificial intelligence diagnostic models to clarify the clinical relevance and applicability of generalist models in clinical practice.

Our results encourage conducting further performance evaluations of multimodal models in different radiologic

XSL•FO
RenderX

subdisciplines, as well as using larger data sets, to gain a more holistic understanding of their role in radiology.

Data Availability

The cases analyzed in this study are available from the Radiological Society of North America (RSNA) Case Collection. This repository can be accessed by RSNA members on the RSNA Case Collection website [5], where each case is presented with detailed clinical information, imaging data, questions, multiple-choice answers, and diagnostic conclusions. The data set was utilized under the terms and conditions provided by the RSNA, which permits the use of these cases for educational and research purposes. No additional unpublished data from these cases were utilized in this study. Researchers and readers are encouraged to directly access the RSNA Case Collection for further information.

Conflicts of Interest

None declared.

References

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv Preprint posted online on Apr 12, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.13375]
2. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023 Jun;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
3. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023 May;307(4):e230725. [doi: 10.1148/radiol.230725] [Medline: 37014240]
4. GPT-4V(ision) system card. OpenAI. 2023 Sep 25. URL: https://openai.com/research/gpt-4v-system-card [accessed 2023-10-14]
5. RSNA Case Collection. Radiological Society of North America. URL: https://cases.rsna.org/ [accessed 2024-04-24]
6. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). arXiv Preprint posted online on Oct 11, 2023 [FREE Full text] [doi: 10.48550/arXiv.2309.17421]
7. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-Flamingo: a multimodal medical few-shot learner. arXiv Preprint posted online on Jul 27, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.15189]

Abbreviations

API: application programming interface
RSNA: Radiological Society of North America
USMLE: United States Medical Licensing Examination

Edited by G Eysenbach; submitted 28.11.23; peer-reviewed by L Zhu, S Kommireddy, H Younes; comments to author 06.02.24; revised version received 10.02.24; accepted 20.03.24; published 01.05.24.

Please cite as:
Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L
Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties
J Med Internet Res 2024;26:e54948
URL: https://www.jmir.org/2024/1/e54948
doi:10.2196/54948
PMID:38691404

©Felix Busch, Tianyu Han, Marcus R Makowski, Daniel Truhn, Keno K Bressem, Lisa Adams. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Using Large Language Models to Support Content Analysis: A Case Study of ChatGPT for Adverse Event Detection

Eric C Leas[1,2], MPH, PhD; John W Ayers[2,3,4], MA, PhD; Nimit Desai[2], BS; Mark Dredze[5], PhD; Michael Hogarth[4,6], MD; Davey M Smith[3,4], MAS, MD

[1]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, United States
[2]Qualcomm Institute, University of California San Diego, La Jolla, CA, United States
[3]Division of Infectious Diseases and Global Public Health, Department of Medicine, University of California San Diego, La Jolla, CA, United States
[4]Altman Clinical Translational Research Institute, University of California San Diego, La Jolla, CA, United States
[5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, United States
[6]Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, United States

Corresponding Author:
Eric C Leas, MPH, PhD
Herbert Wertheim School of Public Health and Human Longevity Science
University of California San Diego
9500 Gilman Drive
Mail Code: 0725
La Jolla, CA, 92093
United States
Phone: 1 951 346 9131
Email: ecleas@ucsd.edu

Abstract

This study explores the potential of using large language models to assist content analysis by conducting a case study to identify adverse events (AEs) in social media posts. The case study compares ChatGPT's performance with human annotators' in detecting AEs associated with delta-8-tetrahydrocannabinol, a cannabis-derived product. Using the identical instructions given to human annotators, ChatGPT closely approximated human results, with a high degree of agreement noted: 94.4% (9436/10,000) for any AE detection (Fleiss κ=0.95) and 99.3% (9931/10,000) for serious AEs (κ=0.96). These findings suggest that ChatGPT has the potential to replicate human annotation accurately and efficiently. The study recognizes possible limitations, including concerns about the generalizability due to ChatGPT's training data, and prompts further research with different models, data sources, and content analysis tasks. The study highlights the promise of large language models for enhancing the efficiency of biomedical research.

(J Med Internet Res 2024;26:e52499) doi:10.2196/52499

KEYWORDS

adverse events; artificial intelligence; AI; text analysis; annotation; ChatGPT; LLM; large language model; cannabis; delta-8-THC; delta-8-tetrahydrocannabiol

Introduction

Biomedical text analysis is commonly burdened by the need for manual data review and annotation, which is costly and time-consuming. Artificial intelligence (AI) tools, including large language models (LLMs) such as ChatGPT (OpenAI) [1], could reduce this burden by allowing scientists to leverage vast amounts of text data (including medical records and public data) with short written prompts as annotation instructions [2]. To explore the potential for AI-assisted annotation, we evaluated whether ChatGPT could replicate human identification of adverse events (AEs) about a cannabis-derived product (delta-8-tetrahydrocannabinol) reported in social media posts [3]. AE detection requires reviewing a large amount of unstructured text data to flag a tiny fraction of AE reports, making it an ideal application for AI-assisted annotation [4].

Methods

Overview

To reduce selective reporting bias, we replicated a peer-reviewed publication, wherein human annotators identified AEs in 10,000

XSL•FO
RenderX

randomly sampled, publicly available posts from a delta-8-tetrahydrocannabiol social media forum (Reddit's r/delta8) [3]. Human annotators identified potential AE reports (yes or no) and whether the AE was serious according to 6 Food and Drug Administration MedWatch categories (eg, hospitalization) [5].

ChatGPT (gpt-3.5-turbo-0613) was set to the default settings (*Temperature*=1, *Top P*=1, *Max token limit*=1700, *Frequency Penalty*=0, and *Presence Penalty*=0); given each Reddit post; and asked to reference annotation instructions identical to those given to human annotators, except for a minor modification for result formatting (ie, requested codes in a comma-delimited format; Multimedia Appendix 1). Since ChatGPT was treated as an additional annotator, we compared ChatGPT's responses with human annotations using the traditional method for assessing interrater reliability rather than statistics for assessing classifiers (eg, F_1-score). Thus, we calculated absolute agreement and prevalence- and bias-adjusted Fleiss κ statistics for any AEs, serious AEs, and each MedWatch category of serious AEs [6]. Analyses were computed with R statistical software (version 4.3.1; R Core Team).

Ethical Considerations

This study was exempted by the University of California San Diego's human research protection program because the data were public and nonidentifiable (45 CFR §46).

Results

ChatGPT returned misformatted responses (eg, including the text "adverse event" instead of the requested "0" or "1") in 35 (0.35%) of 10,000 instances. All misformatted responses were interpretable and resolved through normal data-cleaning methods (eg, rule matching). Example posts along with their labels are shown in Table 1. ChatGPT and human annotators agreed on 94.4% (9436/10,000) of labels for any AEs (κ=0.95) and 99.3% (9931/10,000) of labels for any serious AEs (κ=0.96; Table 2). For serious AEs, the lowest agreement was 99.4% (9939/10,000) for "other" serious (but undefined) outcomes (κ=0.98). All specifically defined outcomes (eg, hospitalization) achieved 99.9% (≥9986/10,000) agreement (κ=0.99).

Table 1. Example of posts to the Reddit community r/delta8 and the corresponding categorizations.

Title and text	Labels[a]
Had to be rushed to the ER after eating an edible. Last week me and my boyfriend bought delta 8 edibles from a vape shop. We were bored and decided it would be a good idea to test it out, we ate two (approximately .1 gram in total). Just a side note, this is was not my first time eating an edible so I didn't really think much of it. It took about 40 minutes for the edible to kick in, at first I just felt very heavy and It was super hard to move, so I laid down for about an hour. Eventually I got bored of laying down and got up to go shower...bad decision. According to my boyfriend, when I got up I fainted. I remember waking up to him freaking tf out, it was very hard to breathe, and it felt like my heart was going to burst. They rushed me to the ER because I was barely able to stay conscious. I had a phycotic break, I thought I was dead, kept hearing all kinds of noises, and I completely lost touch with reality. My heart rate was over 165, I also have a heart condition so they had to keep an eye on that too. It was the most terrifying and traumatizing experience, and I'm still not over it yet. Has anyone gone through this before?	Identified as an adverse event report and considered serious with the following outcomes: life-threatening, hospitalization, and other serious adverse event
Help I feel hungover from delta 8. I feel so awful and can't stop puking. I took 10 mg last night and still feel horrible today. Any advice?	Identified as an adverse event report, but not considered serious
Battery Question. Can someone please recommend and ideal wattage/voltage to use the [BRAND] with? I only have variable wattage/voltage batteries for nicotine vaping and am unfamiliar with batteries used for oils. I'm assuming the former type should work fine as long as I have them set low enough? Any help is appreciated. Thanks	Not identified as an adverse event report

[a]Serious adverse events were defined using the Food and Drug Administration MedWatch health outcome categories, which include life-threatening; hospitalization; disability or permanent damage; congenital anomaly or birth defect; required intervention to prevent permanent impairment; or other serious event.

XSL•FO
RenderX

Table 2. Accuracy of ChatGPT in replicating human identification of adverse events in r/delta8 posts (N=10,000) and the categorization of adverse events to the Food and Drug Administration MedWatch outcome categories.

MedWatch categories and ChatGPT response	Human annotation		Agreement, n (%)	κ statistic[a]
	Yes, n	No, n		
Labeled as an adverse event report			9436 (94.4)	0.95
Yes	172	401		
No	163	9264		
Labeled as a serious adverse event report[b]			9331 (99.3)	0.96
Yes	15	17		
No	52	9916		
Life-threatening			9995 (99.9)	0.99
Yes	1	5		
No	0	9994		
Hospitalization				
Yes	5	6	9993 (99.9)	0.99
No	1	9988		
Disability or permanent damage			9998 (99.9)	N/A[c]
Yes	0	2		
No	0	9998		
Congenital anomaly or birth defect			9999 (99.9)	N/A
Yes	0	1		
No	0	9999		
Required intervention to prevent permanent impairment or damage			9986 (99.9)	0.99
Yes	0	2		
No	12	9986		
Other serious or important medical events			9939 (99.4)	0.98
Yes	7	13		
No	48	9932		

[a]Prevalence- and bias-adjusted Fleiss κ.

[b]A composite of any of the 6 adverse event outcomes.

[c]N/A: not applicable (κ could not be calculated due to no events being found by human annotators).

Discussion

ChatGPT demonstrated near-perfect replication of human-identified AEs in social media posts using the exact instructions that guided human annotators. Despite significant resource allocation, automating AE detection has seen limited success. Many studies (eg, social media studies) often omit performance metrics such as agreement with ground truth altogether [7]. The LLM and prompt used outperformed the best-performing specialized software for detecting AEs from text data (agreement=94.5%; κ=0.89), which relied on structured and human-curated electronic discharge summaries [8].

We note a few limitations. First, we did not have any measures from the replicated study to estimate time or cost savings attributable to using an LLM. However, these savings would be considerable. If a human annotated 1 post/min, the replicated study's estimated completion time would be 166.6 hours (10,000 posts × 60 posts/h), or 20.8 workdays. Conversely, assuming ChatGPT annotated a post in 2 seconds [9], it would take 5.6 hours with no human effort. Second, the social media data analyzed may be included in ChatGPT's underlying training data, potentially inflating the accuracy reported herein and reducing generalizability. Third, our goal was to replicate human annotation using the exact codebook that trained human annotators and default settings of ChatGPT-3.5-turbo. Although this alone showed promise, further improvements to the prompt, different models (GPT-4 or Llama-2), or alternative model parameter specifications may improve the accuracy. Finally, we only assessed 1 application of an LLM for biomedical text analysis; inaccuracy and label bias may exist in other settings. Further research is needed to capture process outcomes (eg, time savings), apply LLMs to traditional biomedical data (eg,

XSL•FO
RenderX

health records), and address more complex methods of annotation (eg, open coding).

While acknowledging its limitations, this case study demonstrates the potential for AI to assist researchers in text analysis. Given the demand for annotations in biomedical research and the inherent time and cost constraints, adopting LLM-powered tools could expedite the research process and consequently scientific discovery.

Acknowledgments

This work was funded by grant K01DA054303 from the National Institute on Drug Abuse, the Burroughs Wellcome Fund, and the National Institutes of Health (UL1TR001442). The study sponsors took no part in the study design; collection, analysis, and interpretation of data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

The corresponding data for the study are available on the first author's website [10].

Conflicts of Interest

ECL has received consulting fees from Good Analytics. JWA owns equity in Health Watcher and Good Analytics. ND has received consulting fees from Pearl Health. MD owns equity in Good Analytics and receives consulting fees from Bloomberg LP. MH advised LifeLink, a company that developed a health care chatbot, between 2016 and 2020, and maintains an equity position in the company. DMS reports paid consulting for Bayer, Arena Pharmaceuticals, Evidera, FluxErgy, Model Medicines, and Linear Therapies.

Multimedia Appendix 1
Prompt used to train ChatGPT.
[DOCX File , 15 KB - jmir_v26i1e52499_app1.docx]

References

1. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-04-25]
2. Lee P, Goldberg C, Kohane I. The AI Revolution in Medicine: GPT-4 and Beyond. London, UK: Pearson; 2023.
3. Leas EC, Harati RM, Satybaldiyeva N, Morales NE, Huffaker SL, Mejorado T, et al. Self-reported adverse events associated with Δ8-tetrahydrocannabinol (delta-8-THC) use. J Cannabis Res 2023 May 23;5(1):15 [FREE Full text] [doi: 10.1186/s42238-023-00191-y] [Medline: 37217977]
4. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform 2015 Apr;54:202-212 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.004] [Medline: 25720841]
5. MedWatch: The FDA Safety InformationAdverse Event Reporting Program. US Food and Drug Administration. 2022 Sep 15. URL: https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program [accessed 2023-01-03]
6. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993 May;46(5):423-429. [doi: 10.1016/0895-4356(93)90018-V]
7. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, van Le H, et al. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf 2017 Apr;40(4):317-331 [FREE Full text] [doi: 10.1007/s40264-016-0491-0] [Medline: 28044249]
8. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12(4):448-457 [FREE Full text] [doi: 10.1197/jamia.M1794] [Medline: 15802475]
9. OpenAI API and other LLM APIs response time tracker. GPT for Work by Talarian. URL: https://gptforwork.com/tools/openai-api-and-other-llm-apis-response-time-tracker [accessed 2024-03-13]
10. Leas E. Publication data. Eric Leas. URL: https://www.ericleas.com/datasets [accessed 2024-04-29]

Abbreviations

AE: adverse event
AI: artificial intelligence
LLM: large language model

XSL•FO
RenderX

Edited by Q Jin; submitted 06.09.23; peer-reviewed by Y Li, T Wang, L Zhu, A Khosla; comments to author 10.03.24; revised version received 14.03.24; accepted 28.03.24; published 02.05.24.

Please cite as:
Leas EC, Ayers JW, Desai N, Dredze M, Hogarth M, Smith DM
Using Large Language Models to Support Content Analysis: A Case Study of ChatGPT for Adverse Event Detection
J Med Internet Res 2024;26:e52499
URL: https://www.jmir.org/2024/1/e52499
doi:10.2196/52499
PMID:38696245

©Eric C Leas, John W Ayers, Nimit Desai, Mark Dredze, Michael Hogarth, Davey M Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases

Wan Hang Keith Chiu[1]; Wei Sum Koel Ko[1]; William Chi Shing Cho[2], PhD; Sin Yu Joanne Hui[3]; Wing Chi Lawrence Chan[4], PhD; Michael D Kuo[5], MD

[1]Department of Diagnostic and Interventional Radiology, Queen Elizabeth Hospital, Hong Kong, China (Hong Kong)
[2]Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong, China (Hong Kong)
[3]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China (Hong Kong)
[4]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[5]Ensemble Group, Scottsdale, AZ, United States

Corresponding Author:
Michael D Kuo, MD
Ensemble Group
10541 E Firewheel Drive
Scottsdale, AZ, 85259
United States
Phone: 1 4084512341
Email: mikedkuo@gmail.com

Abstract

Large language models showed interpretative reasoning in solving diagnostically challenging medical cases.

(J Med Internet Res 2024;26:e53724) doi:10.2196/53724

KEYWORDS

large language model; hospital; health center; Massachusetts; statistical analysis; chi-square; ANOVA; clinician; physician; performance; proficiency; disease etiology

Introduction

Large language models (LLMs) have demonstrated a surprising performance in radiological examinations [1]. However, their proficiency in real-world medical reasoning, especially when integrating multimodal data remains uncertain [2]. This study evaluates the ability of 3 commonly used LLMs—Google Bard (subsequently rebranded Gemini), Claude 2, and GPT-4—to generate differential diagnoses (ddx) from complex multimodality diagnostic cases.

Methods

Overview

Consecutive case records of the Massachusetts General Hospital from July 2020 to June 2023 were selected [3]. The cases were diagnostically challenging, but a final diagnosis was provided. Only the case presentation and a simple prompt asking for the top 5 ddx were used as input. Each case was run independently to prevent the model from being influenced by prior cases. To evaluate the stability of the results, all cases were reinputted into each LLM. To enable objective assessment, all diagnoses were mapped to their corresponding *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes, with higher-level codes used in case an exact code could not be assigned (Figure 1).

The primary objective was accuracy, measured by whether the final diagnosis was within the LLM-generated ddx at the *ICD-10* category level. The secondary objectives were to measure the similarity between diagnoses within the ddx and the final diagnosis as well as their similarity to each other, measured at the *ICD-10* chapter level. Chi-square and ANOVA tests were used to compare categorical data between the LLMs. Statistical analyses were performed using Prism 10 (GraphPad Software).

XSL•FO
RenderX

Figure 1. (A) Standardized prompt used for each case to generate differential diagnoses (ddx). (B) An example of *International Classification of Diseases, Tenth Revision* (*ICD-10*) code hierarchy structure; the first character (an alphabetical letter) denotes the chapter, and when combined with the next 2 digits, it forms the *ICD-10* category code. (C) An example of a large language model (LLM)–generated ddx and the corresponding *ICD-10* codes (case 34); in this case, none of the 3 LLMs included the final diagnosis (high-grade B-cell lymphoma, not otherwise specified; C83.30) in their ddx. For Bard, 3 of the 5 ddx belonged to the same chapter as the final diagnosis (chapter II: C22.0, C85.9, and C79.9). For Claude 2, only 1 of the 5 ddx belonged to the same chapter as the final diagnosis (chapter II: C85.9). For GPT-4, only 1 of the 5 ddx belonged to the same chapter as the final diagnosis (chapter II: C79.9).

(A) The prompt used at the outset of each case was as follows:

"Assuming I am an internist, please assume the role of an expert radiologist and provide the top 5 differential diagnoses for my patient based on the given presentation and investigative findings."

(B) An example of the *ICD-10* code hierarchy:

Chapter (A00-B99): Certain infectious and parasitic diseases
Section (A00-A09): Certain infectious and parasitic diseases
Category (A00): Cholera

(C) Example of final diagnosis (*ICD-10* category), and LLMs ddx with their corresponding *ICD-10* codes:

Final diagnosis (corresponding *ICD-10* code)	Top 5 ddx (corresponding *ICD-10* codes) by Bard	Top 5 ddx (corresponding *ICD-10* codes) by Claude 2	Top 5 ddx (corresponding *ICD-10* codes) by GPT4
High-grade B-cell lymphoma, not otherwise specified (C83.30)	1. Primary sclerosing cholangitis (K83.01) 2. Hepatocellular carcinoma (C22.0) 3. Lymphoma (C85.9) 4. Metastases (C79.9) 5. Sarcoidosis (D86.9)	1. Lymphoma (C85.9) 2. Sarcoidosis (D86.9) 3. Mycobacterial infection (A31.9) 4. Whipple's disease (K90.81) 5. Autoimmune disorder (M35.9)	1. Lymphatic disorders, such as Lymphangiectasia (I89.0) 2. Metastatic malignancy (C79.9) 3. Chronic granulomatous disease (eg, sarcoidosis; D86.9) 4. Chronic inflammatory bowel disease (eg, Crohn's disease; K50.90) 5. Infections, such as tuberculosis (A18.32)

Ethics Approval

Approval from an institutional review board was not required due to the use of publicly available nonidentifiable data.

Results

The diagnostic accuracy on 104 evaluated cases based on the first set of answers by the LLMs was 27.9% for Bard, 30.8% for Claude 2, and 31.7% for GPT-4. Accuracy significantly improved at the *ICD-10* chapter (body site or system) level, reaching 65.4% for Bard, 66.3% for Claude 2, and 71.2% for GPT-4. The mean number of the same ddx generated in each case in the repeatability testing was 2.3 (SD 1.1) for Bard, 2.4 (SD 1.2) for Claude 2, and 2.4 (SD 1.2) for GPT-4.

All 3 LLMs showed evidence of interpretive reasoning, as they tended to generate sets of ddx whose member diagnoses were often related to each other. The mean number of ddx per case belonging to the same *ICD-10* chapter as each other was 2.6 (SD 1.1) for Bard, 2.7 (SD 1.1) for Claude 2, and 2.4 (SD 0.9) for GPT-4. Interestingly, these related diagnosis "clusters" were often unrelated to the final diagnosis. The mean number of ddx belonging to the same *ICD-10* chapter as the final diagnosis was 1.2 (SD 1.3) for Bard, 1.4 (SD 1.4) for Claude 2, and 1.4 (SD 1.2) for GPT-4. These two findings were irrespective of whether the LLMs could include the final diagnosis in their ddx. Furthermore, the performance of the LLMs varied by disease etiology, although this difference was not statistically significant (Table 1).

XSL•FO
RenderX

Table 1. Performance of individual large language models (LLMs).

Characteristics	Bard	Claude 2	GPT4	*P* value
Accuracy by *ICD-10*[a] hierarchy level, %				
Category	27.9	30.7	30.7	<.001[b]
Chapter	65.4	66.3	71.2	<.001[b]
Accuracy by *ICD-10* etiology (top 5 by frequency), n (%)				
Certain infectious and parasitic diseases (chapter I: A00-B99)	20 (35.0)	45.0	50.0	.62[c]
Neoplasm (chapter II C00-D48)	19 (52.6)	63.2	57.9	.75[c]
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (chapter III: D50-D89)	8 (12.5)	25.0	12.5	.74[c]
Endocrine, nutritional, and metabolic diseases (chapter IV: E00-E90)	9 (33.3)	33.3	33.3	>.99[c]
Diseases of the musculoskeletal system and connective tissue (chapter XIII: M00-M99)	11 (36.4)	72.7	63.6	.20[c]
Number of diagnoses per ddx[d] per case generated by LLMs belonging to the same hierarchical chapter as the final diagnosis based on assigned *ICD-10* codes, mean (SD)	1.2 (1.3)	1.4 (1.4)	1.4 (1.2)	—[e]
Number of diagnoses per ddx per case generated by LLMs belonging to the same hierarchical chapter based on assigned *ICD-10* codes, mean (SD)	2.6 (1.1)	2.7 (1.1)	2.4 (0.9)	—
Number of the same ddx per case generated by LLMs on repeatability testing, mean (SD)	2.3 (1.1)	2.4 (1.2)	2.4 (1.2)	—

[a]*ICD-10*: *International Classification of Diseases, Tenth Revision*.

[b]Comparison of each LLM's performance at the *ICD-10* category level versus the chapter level.

[c]Comparison of each LLM's performance across different *ICD-10* etiologies. *P* values were not significant.

[d]ddx: differential diagnoses.

[e]Not applicable.

Discussion

This study rigorously evaluated the diagnostic capacity of multiple LLMs using a simple standardized prompt [4]. The 3 LLMs represent state-of-the-art, general LLMs accessible to most clinicians. The relatively low accuracy of all 3 models at the *ICD-10* category level, coupled with a mean of >3 out of 5 diagnoses located in a chapter outside the final diagnosis chapter, collectively suggest either a knowledge or reasoning gap in current LLMs. Although performance differences are observed between different types of disease etiology (eg, 12.5% for Chapter III vs 63.6% for Chapter XIII in GPT4), the small numbers and unequal distribution of etiologies preclude adequate analysis; however, this area warrants further investigation. Conversely, the moderate number of LLM-generated ddx belonging to the same body site or system (chapter) implies these models can integrate and reason across complex clinical findings.

This study has limitations, including the low reproducibility of the ddx generated by the LLMs. The generative nature of these models and their continuous updates may lead to performance drifts and contradictory results. Further research and validation are necessary to generate consistent and explainable results as well as explore the relationships between performance and repeatability. Second, we did not assess whether human-artificial intelligence interaction or prompt engineering would affect diagnostic accuracy. Nevertheless, attempts to "overengineer" general LLMs toward a desired output could cloud real-world applicability, detracting from the ease of use that makes current LLMs attractive to general users [5]. Future work includes analyzing the rationales provided by the LLMs in reaching their ddx and asking the LLMs to quantify the likelihood of each ddx. Finally, the diversity of LLM-generated ddx warrants further exploration, as it could potentially hamper patient management [6].

In conclusion, LLMs may have a role in enhancing physician diagnosis of complex, multimodal clinical cases when applied judiciously.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023 Jun 01;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
2. Jamshidi N, Feizi A, Sirlin CB, Lavine JE, Kuo MD. Multi-modality, multi-dimensional characterization of pediatric non-alcoholic fatty liver disease. Metabolites 2023 Aug 08;13(8):929 [FREE Full text] [doi: 10.3390/metabo13080929] [Medline: 37623872]
3. Dougan M, Anderson MA, Abramson JS, Fitzpatrick MJ. Case 14-2022: a 57-year-old man with chylous ascites. N Engl J Med 2022 May 12;386(19):1834-1844. [doi: 10.1056/nejmcpc2115856]
4. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 03;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
5. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology 2023 Sep;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
6. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659 [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]

Abbreviations

Ddx: differential diagnoses
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
LLM: large language model

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 17.10.23; peer-reviewed by R Yang, L Zhu, H Mondal; comments to author 05.02.24; revised version received 22.02.24; accepted 23.04.24; published 13.05.24.

Please cite as:
Chiu WHK, Ko WSK, Cho WCS, Hui SYJ, Chan WCL, Kuo MD
Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases
J Med Internet Res 2024;26:e53724
URL: https://www.jmir.org/2024/1/e53724
doi:10.2196/53724
PMID:38739441

©Wan Hang Keith Chiu, Wei Sum Koel Ko, William Chi Shing Cho, Sin Yu Joanne Hui, Wing Chi Lawrence Chan, Michael D Kuo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Multimodal ChatGPT-4V for Electrocardiogram Interpretation: Promise and Limitations

Lingxuan Zhu[1*], MD; Weiming Mou[2*], MD; Keren Wu[1], MD; Yancheng Lai[1], MD; Anqi Lin[1], MD; Tao Yang[3], MD; Jian Zhang[1], MD; Peng Luo[1], MD

[1]Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, China

[2]Department of Urology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Bejing, China

[*]these authors contributed equally

Corresponding Author:
Peng Luo, MD
Department of Oncology
Zhujiang Hospital
Southern Medical University
253 Industrial Avenue
Guangzhou, 510282
China
Phone: 86 020 61643888
Email: luopeng@smu.edu.cn

Abstract

This study evaluated the capabilities of the newly released ChatGPT-4V, a large language model with visual recognition abilities, in interpreting electrocardiogram waveforms and answering related multiple-choice questions for assisting with cardiovascular care.

(J Med Internet Res 2024;26:e54607) doi:10.2196/54607

KEYWORDS

ChatGPT; ECG; electrocardiogram; multimodal; artificial intelligence; AI; large language model; diagnostic; quantitative analysis; clinical; clinicians; ECG interpretation; cardiovascular care; cardiovascular

Introduction

Electrocardiogram (ECG) interpretation is an essential skill in cardiovascular medicine. The rise of artificial intelligence (AI) has led to many attempts to automate ECG interpretations [1]. As a representative of generative AI, ChatGPT has shown promising potential in cardiovascular medicine [2,3]. However, since early versions of ChatGPT cannot process graphical information, its ability for ECG interpretation is unclear. The newly released ChatGPT-4V(ision) model adds visual recognition capabilities [4], which makes it possible to directly read and interpret ECG waveforms. Therefore, we evaluated the performance of ChatGPT-4V in ECG interpretations.

Methods

We gathered a set of multiple-choice questions related to ECG waveform interpretation from various question banks, including the American Heart Association Advanced Cardiovascular Life Support exam (February 2016), United States Medical Licensing Examination (USMLE) sample questions, USMLE practice questions available on the AMBOSS platform [5], and the Certified EKG Technician practice exam. The 62 ECG-related questions included for analysis involved ECG diagnosis and the ability to determine further treatment plans based on ECG findings and corresponding clinical scenarios.

ChatGPT was prompted to answer the questions by analyzing the accompanying ECG images; the prompt also stated that ChatGPT was undergoing a diagnostic challenge as a representative of AI to prevent it from refusing to make a diagnosis (see Multimedia Appendix 1).

ChatGPT was asked each question 3 times to mitigate the effect of randomness in responses in the evaluation. Accuracy was then evaluated based on ChatGPT getting at least 1, 2, or 3 correct answers out of the 3 attempts. To further confirm

XSL•FO
RenderX

whether ChatGPT could make accurate diagnoses without relying on options, 19 diagnostic-related questions that purely examined ECG interpretation without requiring integration of clinical history were converted to open-ended questions. ChatGPT was then prompted to provide a diagnosis after reading the ECG without options.

Results

The 62 questions included 26 questions for diagnosis, 29 for treatment, and 7 for counting tasks such as QT-interval length calculation. The overall accuracy was 83.87%, 70.97%, and 53.23% for getting at least 1, 2, and 3 out of the 3 attempts correct (Figure 1). There were significant differences in accuracy across question types with 1 or 2 correct responses, whereas there was no significant difference when all 3 responses were required to be correct (Table 1). Accuracy at least 2 times was the highest for treatment recommendation questions, followed by diagnosis and counting questions. Subgroup analysis showed lower accuracy in counting-type than diagnostic- and treatment-related questions when requiring at least 1 or 2 correct responses. Treatment recommendation questions had higher accuracy than other types when at least 2 correct responses were needed (Table 1).

Figure 1. Accuracy of the multimodal ChatGPT-4V model in answering multiple-choice questions related to electrocardiogram (ECG) interpretation. The number of correct responses among 3 attempts for each question are shown from left to right. The accuracy rates with at least 1, 2, and 3 correct responses are annotated on the right from the bottom to the top. Different shapes represent different question types. We evaluated ChatGPT-4V responses using the official reference answers as a standard for reliability. Any questions involving ECG image interpretations were included without additional exclusion criteria. Unedited ECG images were uploaded to ChatGPT at the original resolution and no additional information was provided to maintain consistency with the original test questions. The prompt we used did not contain any hints about the correct answer. ChatGPT's responses were collected from October 4 to 8, 2023. The ggplot2 R package was used for visualization.

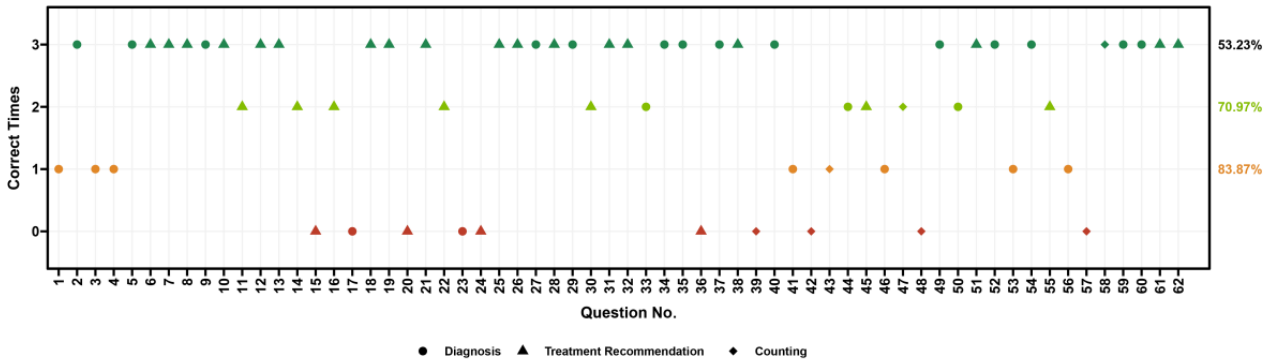

Table 1. Accuracy of the multimodal ChatGPT-4V model for different types of questions.

Question type	Number of questions	Correct answers, n (%)	*P* value[a]
At least 1 correct			.02
Diagnosis	26	24 (92.31)	.17
Treatment recommendation	29	25 (86.21)	.74
Counting	7	3 (42.86)	.001
At least 2 correct			.009
Diagnosis	26	17 (65.38)	.57
Treatment recommendation	29	25 (86.21)	.02
Counting	7	2 (28.57)	.02
All 3 correct			.09
Diagnosis	26	14 (53.85)	—[b]
Treatment recommendation	29	18 (62.07)	—
Counting	7	1 (14.29)	—

[a]The Fisher exact test was used to compare the accuracy of ChatGPT in answering different types of questions with the *fisher.test* function in R (version 4.2.3). If there was a statistically significant difference, subgroup analysis using the Fisher exact test was further performed to respectively compare the accuracy of each type with the other two types.

[b]Not applicable; subgroup analysis was not performed since there was no significant difference among the three question types overall.

ChatGPT performed poorly in diagnosing ECGs without options, making the correct ECG diagnosis in only 7 out of 57 responses, which suggests that the ECG-based diagnostic ability of the current version is only possible with a limited range of options

XSL•FO
RenderX

provided. Incorrect responses were related to specific functionalities of ChatGPT-4V. The insufficient ability of ChatGPT-4V to count parameters such as PR intervals could lead to errors in diagnostic and therapeutic questions, and its inadequacy in integrating ECG parameters could result in nonspecific diagnoses. For example, ChatGPT-4V could diagnose myocardial infarction but fail to combine various parameters to determine the specific location of the infarction.

Discussion

Although ChatGPT-4V can analyze ECGs to some extent and can even make treatment decisions based on the EGC, its diagnostic stability and reliability need further improvement for clinical application. ChatGPT-4V had significantly lower accuracy on counting-based questions than treatment- or diagnostic-related questions, suggesting its limitations in precise quantitative ECG measurements.

Notably, the model was not specifically trained on ECG data. Thus, we expect ChatGPT-4 to perform better on ECG interpretation as it accumulates more data and training. As a general-purpose model, ChatGPT-4V's capabilities are not limited to correctly diagnosing ECGs; however, its good performance on ECG-based treatment recommendation questions highlights its potential application in medical decision-making. By leveraging ChatGPT-4V's abilities to analyze free text and images, management recommendations can be directly generated based on patient data and ECG waveforms to improve health care efficiency. While current bedside cardiac monitors can only offer a warning for issues such as abnormal heart rhythms or atrial fibrillation, models such as ChatGPT-4V could be positioned to serve as 24/7 "attending physicians" that monitor and analyze ECGs of patients with critical illness, capturing low-frequency but important ECG abnormalities and promptly detecting condition changes to recommend timely interventions. ChatGPT can also be used to train medical trainees about ECG interpretation and act as an automated second reader to identify high-risk diagnoses.

Our study provides a first look at the state-of-the-art ChatGPT-4V model's capabilities in ECG interpretation. While these early results are promising, they also highlight current limitations of the model. With further technological developments, multimodal generative AI tools such as ChatGPT may eventually play an important role in clinical ECG interpretation and cardiovascular care. Larger-scale validation is needed to fully evaluate this ability. Rapid development of large language models is expected to contribute exciting progress in the cardiovascular field.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts used for this study.
[DOCX File , 23 KB - jmir_v26i1e54607_app1.docx]

References

1. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021 Jul;18(7):465-478. [doi: 10.1038/s41569-020-00503-2] [Medline: 33526938]
2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844. [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
3. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
4. Yang Z, Li L, Lin K, Wang J, Lin C, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). arXiv Preprint posted online on October 11, 2023 [FREE Full text] [doi: 10.48550/arXiv.2309.17421]
5. AMBOSS: medical knowledge platform for doctors and students. URL: https://www.amboss.com/us [accessed 2023-10-20]

Abbreviations

AI: artificial intelligence
ECG: electrocardiogram
USMLE: United States Medical Licensing Examination

XSL•FO
RenderX

Edited by B Puladi; submitted 16.11.23; peer-reviewed by WHK Chiu, E Amini-Salehi; comments to author 29.02.24; revised version received 03.03.24; accepted 19.04.24; published 26.06.24.

Please cite as:
Zhu L, Mou W, Wu K, Lai Y, Lin A, Yang T, Zhang J, Luo P
Multimodal ChatGPT-4V for Electrocardiogram Interpretation: Promise and Limitations
J Med Internet Res 2024;26:e54607
URL: https://www.jmir.org/2024/1/e54607
doi:10.2196/54607
PMID:38764297

©Lingxuan Zhu, Weiming Mou, Keren Wu, Yancheng Lai, Anqi Lin, Tao Yang, Jian Zhang, Peng Luo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Benchmarking State-of-the-Art Large Language Models for Migraine Patient Education: Performance Comparison of Responses to Common Queries

Linger Li[1], MM; Pengfei Li[2], MD; Kun Wang[1], MM; Liang Zhang[1], MD; Hongwei Ji[3,4*], MD; Hongqin Zhao[1*], MD, PhD

[1]Department of Neurology, The Affiliated Hospital of Qingdao University, Qingdao, China
[2]Department of Internal Medicine, The Affiliated Hospital of Qingdao University, Qingdao, China
[3]Tsinghua Medicine, Tsinghua University, Beijing, China
[4]Department of Internal Medicine, Beijing Tsinghua Changgung Hospital, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Hongqin Zhao, MD, PhD
Department of Neurology
The Affiliated Hospital of Qingdao University
No. 59 Haier Road
Qingdao, 266035
China
Phone: 86 13864873935
Fax: 86 532 85953085
Email: zhaohongq@qdu.edu.cn

Abstract

This study assessed the potential of large language models (OpenAI's ChatGPT 3.5 and 4.0, Google Bard, Meta Llama2, and Anthropic Claude2) in addressing 30 common migraine-related queries, providing a foundation to advance artificial intelligence–assisted patient education and insights for a holistic approach to migraine management.

(J Med Internet Res 2024;26:e55927) doi:10.2196/55927

KEYWORDS

migraine; large language models; patient education; ChatGPT; Google Bard; language model; patient education; education; headache; accuracy; OpenAI; AI; artificial intelligence; AI-assisted; holistic; migraine management; management

Introduction

Migraine is a highly debilitating primary headache disorder. Long-term education for migraine is essential to help patients, identify triggers, use medications appropriately, and adopt lifestyle changes that can reduce the frequency and severity of attacks [1-3]. Large language models (LLMs), complex artificial intelligence systems trained on extensive data sets to produce human-like text responses, are promising tools for patient education. However, LLMs could lead to inaccurate responses known as "hallucinations" [4]. Therefore, rigorous evaluations within specific medical domains are essential. Nevertheless, there is a lack of benchmarking for popular online LLMs in patient education for migraines. We performed a comparative analysis of responses to migraine-related queries from five leading LLMs: OpenAI's ChatGPT 3.5 (December 2022) and 4.0 (March 2023), Google Bard (February 2023), Meta Llama2 (July 2023), and Anthropic Claude2 (July 2023).

Methods

Study Process

This study was conducted from October 1 to October 28, 2023. With reference to guidelines and their clinical expertise, neurologists meticulously crafted 30 migraine-related queries spanning a wide array of topics, including evaluation and definition, testing and diagnosis, treatment, follow-up, and prognosis, as well as special population considerations [5-7]. Each query was individually presented to the 5 leading LLMs through an independent conversation web-based interface to avert chained prompting. Responses from all LLM chatbots were converted into plain text, obscuring specific chatbot

XSL•FO
RenderX

features to ensure blinding. The generated answers were then randomly ordered in each set of queries and assessed by blinded reviewers, with a 24-hour interval between each assessment to mitigate memory bias (Figure 1).

Figure 1. Flowchart of the overall study design.

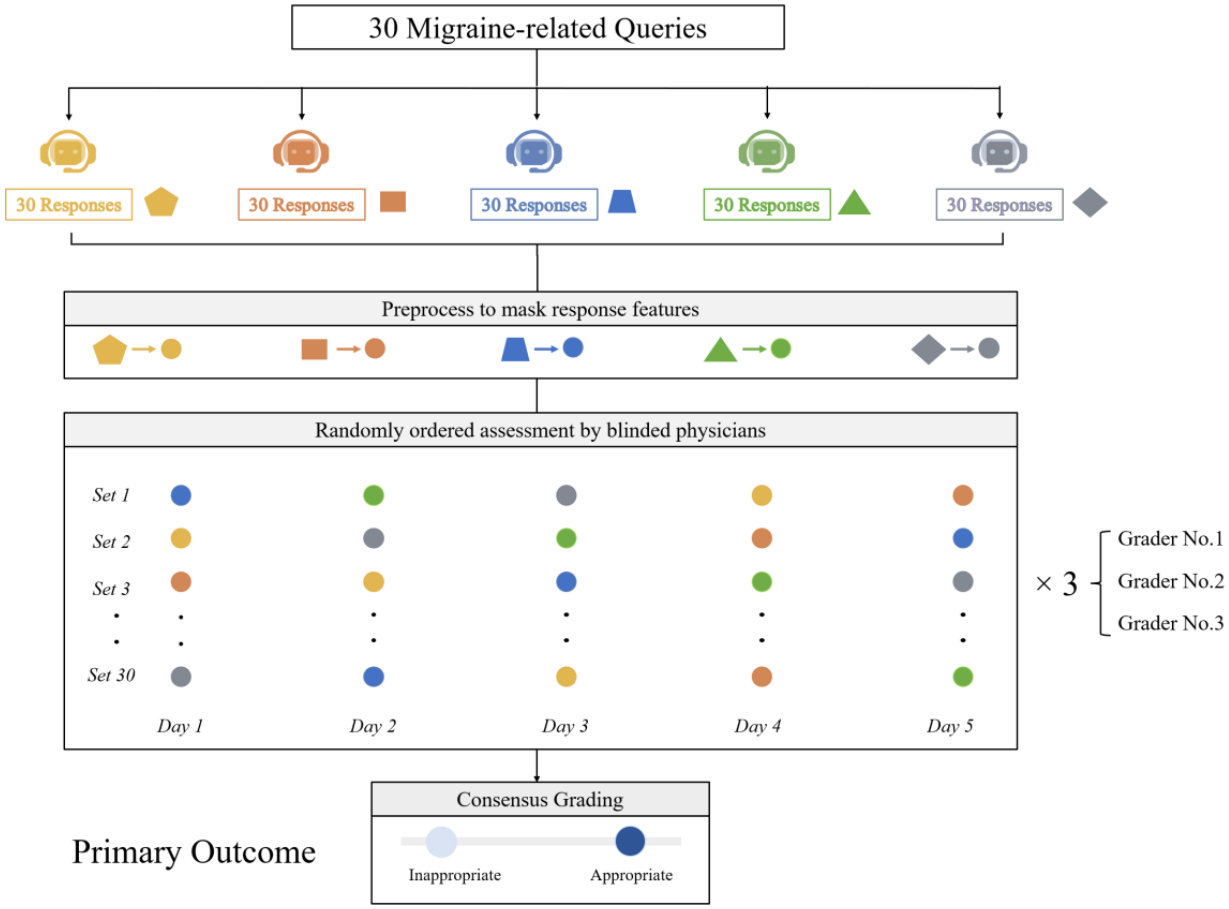

The rating panel comprised three seasoned neurologists, each with at least 15 years of experience in the field. Their primary task was to independently evaluate the accuracy of the LLM chatbots' responses using a 3-point scale ("poor," "borderline," and "good").

Ethical Considerations

Approval from an ethics committee was not required for this study as no patients were involved [8].

Results

ChatGPT-3.5, ChatGPT-4.0, Google Bard, Meta Llama2, and Anthropic Claude2 achieved appropriate response rates (Pearson χ^2 test P=.48; Table 1); Multimedia Appendix 1 provides examples of the prompts and responses from each LLM. Google Bard had a relatively higher "poor" rating proportion than that of the other LLMs (Pearson χ^2 test P=.96; Table 1). None of the LLMs provided an appropriate response to the query "In severe cases, are there surgical options available for treating migraine?" This deficiency potentially stems from their limited ability to distinguish between migraine and secondary headaches, a crucial distinction in medical diagnosis. The complexity and ongoing debate surrounding surgical interventions for migraine treatment likely contribute to this gap in appropriate advice. Specifically, ChatGPT-3.5 erroneously proposed hemicraniectomy for persistent and severe migraines.

XSL•FO
RenderX

Table 1. Performance of large language models in addressing patient queries.[a]

Queries	GPT-3.5	GPT-4	Bard	Llama 2	Claude2
Appropriate response[b], n (%)	27 (90.0)	29 (96.7)	25 (83.3)	27 (90.0)	26 (86.7)
Poor response[c], n (%)	1 (3.3)	1 (3.3)	2 (6.7)	1 (3.3)	1 (3.3)
1. Evaluation and definition					
Why am I experiencing migraines?	9	9	9	9	9
Is migraine a common disease?	9	9	9	9	9
Is there an association between stress and migraine?	9	9	9	9	9
Is there an association between exercise and migraine?	9	9	9	9	9
Is there an association between patent foramen ovale and migraine?	9	9	9	9	9
I have migraine, does this have anything to do with my family's health history?	9	9	9	9	9
What is the best diet for migraine patients?	9	9	9	9	9
Why is my vision blurred before the migraine attack?	8	9	9	9	8
Why do I get frequent migraine attacks?	9	9	9	9	9
Why do I feel dizziness when I have a migraine attack?	9	9	7	8	9
Why do I have migraine attacks during menstruation?	9	9	9	9	9
2. Test and diagnosis					
How is migraine diagnosed?	9	9	9	9	9
Is my migraine serious? how to evaluate it?	9	9	8	9	9
I have a migraine, is it a tumor in my brain?	8	9	9	8	8
How do imaging tests like MRI[d], CT[e] scans, or ultrasounds contribute to diagnosing migraines?	9	9	7	9	9
3. Treatment					
How is migraine treated?	9	9	9	9	9
Do migraines require long-term medication therapy?	9	9	9	9	9
I have a migraine, what will happen if I don't treat it?	9	9	9	9	9
What do migraine patients need to pay attention to when taking medicine?	9	9	9	9	9
What lifestyle changes can I make to better manage my migraines?	9	9	9	9	9
Why do I still have headaches despite taking painkillers almost every day?	9	9	9	9	9
In severe cases, are there surgical options available for treating migraine?	3	3	3	3	3
4. Follow-up and prognosis					
Can migraine be cured?	9	9	9	9	9
Will my migraine condition deteriorate over time?	9	9	9	9	9
Do migraines elevate the risk of brain-related vascular issues?	9	9	9	9	9
How should I monitor and record my migraine symptoms?	9	9	9	9	9
5. Special population considerations					
What should the elderly pay attention to when suffering from migraine?	9	9	9	9	8
I am pregnant/breastfeeding, how should my migraine be treated?	9	9	9	9	9

XSL•FO
RenderX

Queries	GPT-3.5	GPT-4	Bard	Llama 2	Claude2
I have menstrual migraines, how should I treat them?	9	9	3[f]	9	9
Can children develop migraine, and if so, how does it affect their health as they grow up?	9	9	9	9	9

[a]The numbers in the table represent the total scores assigned by the three reviewers, with "poor" corresponding to 1 point, "borderline" to 2 points, and "good" to 3 points. The response was graded as "good" when there were no inaccuracies, "borderline" when there were potential factual inaccuracies but still unlikely to mislead the average patient or cause harm, and "poor" when the response contained unacceptable inaccuracies or inconsistencies that would likely mislead the average patient and cause harm.

[b]The response was considered "appropriate" when all three experts graded it as "good."

[c]The response was considered "poor" when any expert graded it as "poor."

[d]MRI: magnetic resonance imaging.

[e]CT: computed tomography.

[f]Response from Google Bard for this particular query: "I'm not able to help with that, as I'm only a language model."

Discussion

Managing migraines over an extended period significantly strains both health care systems and patients [2]. In this study, most LLM responses were accurate and practical, with the majority of responses graded as "good" or "borderline" rather than "poor." Among the five LLMs tested, ChatGPT-4.0 had the highest accuracy, although the difference in performance was not statistically significant. These observations suggest that LLMs may function as assistive tools in providing advice, enhancing information acquisition, and offering personalized responses, thereby supporting both patients and physicians.

This study represents the first effort to evaluate the capability of state-of-the-art LLMs in educating patients with migraine. Although prior studies have primarily focused on migraine diagnostic tests [9], there was no thorough head-to-head comparison of emerging LLMs for patient education about migraines. The observed variation in performance may be influenced by differences in LLM settings or the limited range of queries tested. Our study benefits from a rigorous study design, incorporating blinding, proper randomization, and an expert review by three neurology specialists. However, it also has limitations. First, the potential generation of false information such as hallucinations could lead to patient confusion and even delays in seeking proper medical care. This highlights the need for strategies to mitigate risks, such as improving model bias monitoring and implementing more stringent testing phases before real-world deployment. Second, this study is potentially underpowered due to the limited sample size; therefore, statistical insignificance should not be viewed as proof of equivalent LLM performance. Given the continually evolving nature of LLMs, ongoing evaluation with broader data sets is crucial for maintaining the validity of the models and clarifying performance differences between LLMs.

In conclusion, these state-of-the-art LLMs have the potential to accurately respond to common migraine-related queries, which may have implications for generative artificial intelligence–assisted migraine education.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant 82103908), Shandong Provincial Natural Science Foundation (grant ZR2021QH014), Shuimu Scholar Program of Tsinghua University, and National Postdoctoral Innovative Talent Support Program (grant BX20230189). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

The data sets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

LL, PL, HJ, and HZ acquired the data. LL and HJ performed the statistical analysis. HJ obtained funding for the study. HJ and HZ supervised the study. All authors participated in the analysis and interpretation of data, drafting of the manuscript, and critical revision of the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Example prompts and responses from five large language models.
[DOCX File , 22 KB - jmir_v26i1e55927_app1.docx]

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. Ashina M, Katsarava Z, Do TP, Buse DC, Pozo-Rosich P, Özge A, et al. Migraine: epidemiology and systems of care. Lancet 2021 Apr 17;397(10283):1485-1495. [doi: 10.1016/S0140-6736(20)32160-7] [Medline: 33773613]
3. GBD 2017 US Neurological Disorders Collaborators, Feigin VL, Vos T, Alahdab F, Amit AML, Bärnighausen TW, et al. Burden of neurological disorders across the US from 1990-2017: a global burden of disease study. JAMA Neurol 2021 Feb 01;78(2):165-176 [FREE Full text] [doi: 10.1001/jamaneurol.2020.4152] [Medline: 33136137]
4. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug 12;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
5. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders, 3rd edition. Cephalalgia 2018 Jan 25;38(1):1-211. [doi: 10.1177/0333102417738202] [Medline: 29368949]
6. Ashina M, Buse DC, Ashina H, Pozo-Rosich P, Peres MFP, Lee MJ, et al. Migraine: integrated approaches to clinical management and emerging treatments. Lancet 2021 Apr 17;397(10283):1505-1518. [doi: 10.1016/S0140-6736(20)32342-4] [Medline: 33773612]
7. Robbins MS. Diagnosis and management of headache: a review. JAMA 2021 May 11;325(18):1874-1885. [doi: 10.1001/jama.2021.1640] [Medline: 33974014]
8. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023 Sep;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
9. Cowan RP, Rapoport AM, Blythe J, Rothrock J, Knievel K, Peretz AM, et al. Diagnostic accuracy of an artificial intelligence online engine in migraine: a multi-center study. Headache 2022 Jul 03;62(7):870-882 [FREE Full text] [doi: 10.1111/head.14324] [Medline: 35657603]

Abbreviations

LLM: large language model

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 29.12.23; peer-reviewed by P Gazerani, TF Heston; comments to author 22.04.24; revised version received 10.05.24; accepted 29.05.24; published 23.07.24.

Please cite as:
Li L, Li P, Wang K, Zhang L, Ji H, Zhao H
Benchmarking State-of-the-Art Large Language Models for Migraine Patient Education: Performance Comparison of Responses to Common Queries
J Med Internet Res 2024;26:e55927
URL: https://www.jmir.org/2024/1/e55927
doi:10.2196/55927
PMID:

©Linger Li, Pengfei Li, Kun Wang, Liang Zhang, Hongwei Ji, Hongqin Zhao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Quality and Accountability of ChatGPT in Health Care in Low- and Middle-Income Countries: Simulated Patient Study

Yafei Si[1], PhD; Yuyi Yang[2], MSc; Xi Wang[3], MSc; Jiaqi Zu[4], MSc; Xi Chen[5,6], PhD; Xiaojing Fan[7], PhD; Ruopeng An[3,8]; Sen Gong[9], PhD

[1]UNSW Business School and CEPAR, The University of New South Wales, Kensington, Australia
[2]Division of Computational and Data Sciences, Washington University in St Louis, St. Louis, MO, United States
[3]Brown School, Washington University in St Louis, St Louis, MT, United States
[4]Global Health Research Center, Duke Kunshan University, Kunshan, China
[5]Department of Health Policy and Management, Yale University, New Haven, CT, United States
[6]Department of Economics, Yale University, New Haven, CT, United States
[7]School of Public Policy and Administration, Xi’an Jiaotong University, Xi'an, China
[8]Silver School of Social Work, New York University, New York, NY, United States
[9]Centre for International Studies on Development and Governance, Zhejiang University, Hangzhou, China

Corresponding Author:
Xiaojing Fan, PhD
School of Public Policy and Administration
Xi’an Jiaotong University
28 West Xianning Road
Xi'an, 710049
China
Phone: 86 15891725861
Email: emirada@163.com

Abstract

Using simulated patients to mimic 9 established noncommunicable and infectious diseases, we assessed ChatGPT’s performance in treatment recommendations for common diseases in low- and middle-income countries. ChatGPT had a high level of accuracy in both correct diagnoses (20/27, 74%) and medication prescriptions (22/27, 82%) but a concerning level of unnecessary or harmful medications (23/27, 85%) even with correct diagnoses. ChatGPT performed better in managing noncommunicable diseases than infectious ones. These results highlight the need for cautious AI integration in health care systems to ensure quality and safety.

(J Med Internet Res 2024;26:e56121) doi:10.2196/56121

KEYWORDS

ChatGPT; generative AI; simulated patient; health care; quality and safety; low- and middle-income countries; quality; LMIC; patient study; effectiveness; reliability; medication prescription; prescription; noncommunicable diseases; AI integration; AI; artificial intelligence

Introduction

The rise of generative artificial intelligence (AI) models like ChatGPT is transforming the health care landscape, especially in low- and middle-income countries (LMICs). These regions, often facing shortages of health care professionals, are increasingly turning to AI tools for medical consultation, aided by growing internet and smartphone access [1]. Research has highlighted generative AI use in the fields of cardiology [2] and orthopedic diseases [3]. However, there are concerns about the accuracy and safety of AI models like ChatGPT [4] given their lack of legal or professional accountability. This is crucial in medical settings, where precise and reliable decision-making is vital. Our study focuses on assessing ChatGPT’s performance in treatment recommendations for common diseases in LMICs, addressing a critical need for the responsible application of AI in health care.

XSL•FO
RenderX

Methods

Overview

We used the simulated patient (SP) method to create a realistic testing environment for ChatGPT with GPT-3.5 from August 8 to 19, 2023. SPs are healthy individuals trained to consistently mimic real patients and their symptoms [5]. We trained the SPs to present 9 common, previously validated diseases [5-8]. We asked ChatGPT to act as a doctor in an LMIC and offer consultations. The SPs detailed their primary concerns, gave standardized responses to every question, and recorded all diagnoses and medication recommendations, which were cross-referenced with clinical guidelines to assess their accuracy and appropriateness. For a robust analysis, we presented each disease to ChatGPT 3 times. We conducted descriptive analyses with the final sample of 27 independent trials.

Ethical Considerations

The Ethics Committee of the First Affiliated Hospital of Xi'an Jiaotong University approved the study (LLSBPJ-2024-WT-019).

Results

Surprisingly, ChatGPT's performance varied across trials for each disease (Figure 1). When aggregating the results (Figure 2), ChatGPT had a 67% (18/27) success rate in initial diagnoses and a 59% (16/27) success rate in medication recommendations. When considering all recommendations, these rates increased to 74% (20/27) for any correct diagnosis and 82% (22/27) for any appropriate medication recommendation. However, there was a high rate of unnecessary or harmful medication suggestions, occurring in 85% (23/27) of trials overall and in 59% (16/27) of trials after a correct diagnosis. Our study also highlighted ChatGPT's varying performance across different types of diseases. Specifically, the AI demonstrated a superior ability in handling noncommunicable diseases compared to infectious diseases, both in terms of diagnosis and medication recommendations.

Figure 1. Heatmap comparing ChatGPT's responses with clinical guidelines. The asterisks (*) indicate infectious diseases; green cells denote correct or appropriate diagnoses or drug prescriptions; blue cells denote incorrect or unnecessary diagnoses or drug prescriptions; and red cells denote harmful drug prescriptions. Each row represents an independent trial.

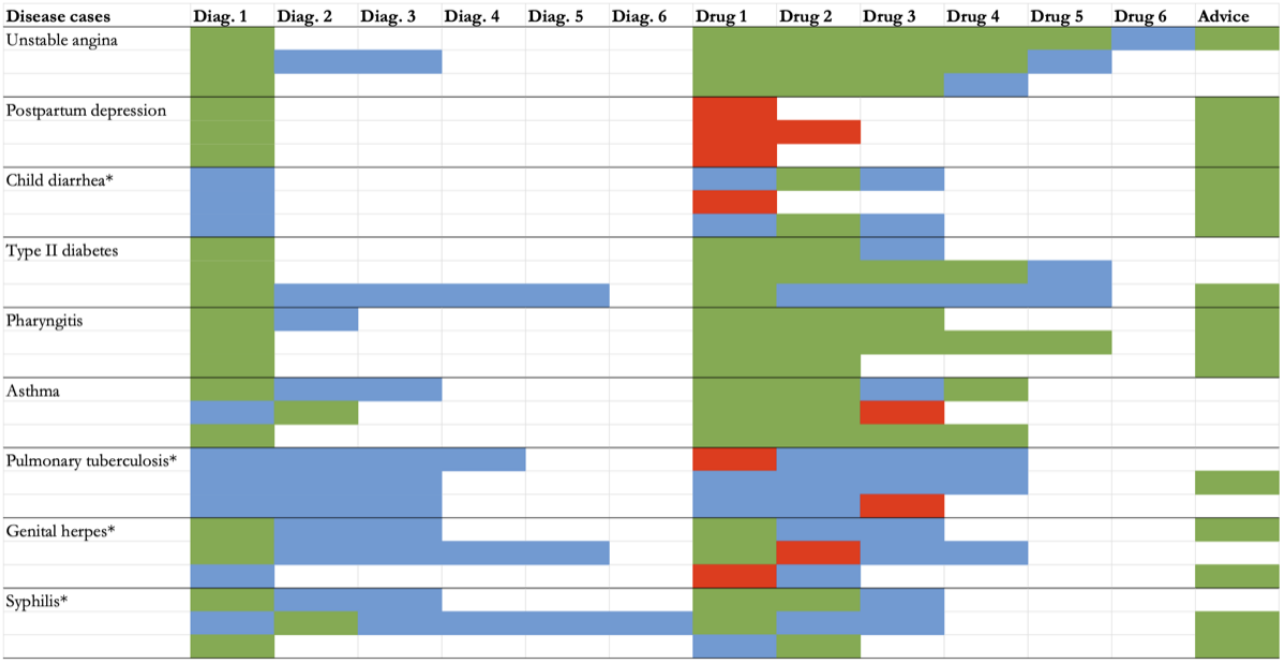

XSL•FO
RenderX

Figure 2. ChatGPT's capability in diagnosing and treating 9 common diseases. The asterisks (*) indicate infectious diseases; green denotes socially desired outcomes; red denotes undesired outcomes; darker colors denote higher probabilities.

		Correct diagnosis		Correct drug		Unnecessary/harmful drug	
Case	Disease presentation	First recommendation	Any recommendation	First recommendation	Any recommendation	Unconditional	Conditional on correct diagnosis
1	Unstable angina	100%	100%	100%	100%	100%	100%
2	Postpartum depression	100%	100%	0%	0%	100%	100%
3	Child diarrhea*	0%	0%	0%	67%	100%	0%
4	Type II diabetes	100%	100%	100%	100%	100%	100%
5	Pharyngitis	100%	100%	100%	100%	0%	0%
6	Asthma	67%	100%	100%	100%	67%	67%
7	Pulmonary tuberculosis*	0%	0%	0%	0%	100%	0%
8	Genital herpes*	67%	67%	67%	67%	100%	67%
9	Syphilis*	67%	100%	67%	100%	100%	100%
Noncommunicable diseases		93%	100%	80%	80%	73%	73%
Infectious diseases		33%	42%	33%	58%	100%	42%
Overall		67%	74%	59%	82%	85%	59%

Discussion

Our findings reveal a high level of accuracy in both correct diagnoses (74%) and medication recommendations (82%) by ChatGPT. Previous studies using the SP method found that primary care providers in LMICs like China, India, and Kenya could only reach correct diagnoses in 12%-52% of SP visits [5,6]. Therefore, ChatGPT can potentially outperform traditional primary care providers in LMICs in diagnostic accuracy. Since ChatGPT with GPT-3.5 is free, the AI tool has the potential to offer affordable and far-reaching solutions in LMICs, particularly in rural and underserved areas.

However, ChatGPT tended to suggest more unnecessary or even harmful medications (in 85% of trials) than primary care providers (28%-64%) [5,6]. AI models work by analyzing available data using machine learning and deep learning techniques [9]. Their approach to drug prescription can be aggressive due to a lack of professional accountability or a motive to reduce medical expenses. ChatGPT also performed better in managing noncommunicable diseases than infectious diseases. This could be because more information on the former is available for AI training during development [10]. ChatGPT's performance also varied within each disease case, contrary to our expectation that this would be more standardized.

We acknowledge several limitations. First, a broader array of diseases, especially those specific to different regions, should be used in future studies. Second, we did not introduce more details (ie, location) to avoid the prompts becoming overcomplicated, and by default, ChatGPT's responses reflect the average population to increase its generalizability. Third, we did not account for the relative importance of the AI's questions and emotional communications. Fourth, a larger sample size may have enabled us to perform head-to-head comparisons between AI care and traditional care.

Despite the limitations, we present the first audit-study evidence to evaluate ChatGPT's performance in diagnosing and treating common diseases in LMICs. A rich set of 9 established diseases makes our findings highly relevant to and widely applicable in LMICs. ChatGPT reaches high levels of accuracy in diagnosis and medication recommendations, but also recommends a concerning level of unnecessary or harmful medications. Integrating AI tools like ChatGPT into health care systems in LMICs may potentially improve diagnostic accuracy but also raises concerns about care safety.

Acknowledgments

No funding was available to support this study. XC acknowledges financial support from the Drazen scholarship and the Aden scholarship dedicated to research on Chinese health care systems. YS and SG acknowledge the support from the National Social Science Foundation of China (23AZD091). During the preparation of this work the authors used ChatGPT with GPT-4 in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

YS contributed to conceptualization, investigation, analysis, and writing (original draft); YY contributed to analysis, investigation, review, and editing; XW contributed to analysis, investigation, review, and editing; JZ contributed to analysis, investigation, review, and editing; XC contributed to review and editing; XF contributed to review and editing; RA contributed to conceptualization, investigation, analysis, and writing; and SG contributed to review and editing. All authors approved the final version of the paper.

Conflicts of Interest

None declared.

References

1. Howarth J. How many people own smartphones? (2023-2028). Exploding Topics. URL: https://explodingtopics.com/blog/smartphone-stats [accessed 2023-12-07]
2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
3. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023 Sep 15;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
4. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of Using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
5. Kwan A, Daniels B, Bergkvist S, Das V, Pai M, Das J. Use of standardised patients for healthcare quality research in low- and middle-income countries. BMJ Glob Health 2019;4(5):e001669 [FREE Full text] [doi: 10.1136/bmjgh-2019-001669] [Medline: 31565413]
6. Si Y, Bateman H, Chen S, Hanewald K, Li B, Su M, et al. Quantifying the financial impact of overuse in primary care in China: a standardised patient study. Soc Sci Med 2023 Mar;320:115670. [doi: 10.1016/j.socscimed.2023.115670] [Medline: 36669284]
7. Xue H, D'Souza K, Fang Y, Si Y, Liao H, Qin WA, et al. Direct-to-consumer telemedicine platforms in China: a national market survey and quality evaluation. Preprints with The Lancet Preprint posted online Oct 18, 2021. [doi: 10.2139/ssrn.3944587]
8. Si Y, Xue H, Liao H, Xie Y, Xu D, Smith M, et al. The quality of telemedicine consultations for sexually transmitted infections in China. Health Policy Plan 2024 Mar 12;39(3):307-317. [doi: 10.1093/heapol/czad119] [Medline: 38113375]
9. Sellamuthu S, Vaddadi S, Venkata S, Petwal H, Hosur R, Mandala V, et al. AI-based recommendation model for effective decision to maximise ROI. Soft Comput 2023:1-10 [FREE Full text] [doi: 10.1007/S00500-023-08731-7]
10. Sanders JW, Fuhrer GS, Johnson MD, Riddle MS. The epidemiological transition: the current status of infectious diseases in the developed world versus the developing world. Sci Prog 2008;91(Pt 1):1-37 [FREE Full text] [doi: 10.3184/003685008X284628] [Medline: 18453281]

Abbreviations

AI: artificial intelligence
LMIC: low- and middle-income country
SP: simulated patient

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 06.01.24; peer-reviewed by D Simmons, N Domingues, W Yang; comments to author 06.04.24; revised version received 21.04.24; accepted 30.07.24; published 09.09.24.

Please cite as:
Si Y, Yang Y, Wang X, Zu J, Chen X, Fan X, An R, Gong S
Quality and Accountability of ChatGPT in Health Care in Low- and Middle-Income Countries: Simulated Patient Study
J Med Internet Res 2024;26:e56121
URL: https://www.jmir.org/2024/1/e56121
doi:10.2196/56121
PMID:39250188

XSL•FO
RenderX

©Yafei Si, Yuyi Yang, Xi Wang, Jiaqi Zu, Xi Chen, Xiaojing Fan, Ruopeng An, Sen Gong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Internet Search Activity for Intentional Self-Harm Forums After a High-Profile News Publication: Interrupted Time Series Analysis

Nora Clancy Kelsall[1], MSc; Catherine Gimbrone[1], MPH; Mark Olfson[2], MPH, MD; Madelyn Gould[1,2], MPH, PhD; Jeffrey Shaman[3,4], PhD; Katherine Keyes[1], MPH, PhD

[1]Department of Epidemiology, Columbia University, New York, NY, United States
[2]Department of Psychiatry, Columbia University, New York, NY, United States
[3]Department of Environmental Health Sciences, Columbia University, New York, NY, United States
[4]Columbia Climate School, Columbia University, New York, NY, United States

Corresponding Author:
Katherine Keyes, MPH, PhD
Department of Epidemiology
Columbia University
722 West 168th Street
Suite 724
New York, NY, 10032
United States
Phone: 1 212 304 5652
Email: kmk2104@cumc.columbia.edu

Abstract

Searches for "pro-suicide" websites in the United States peaked during the week a high-profile news story was published and remained elevated for 6 months afterward, highlighting the need to avoid mentioning specific sources of explicit suicide instructions in media publications.

(J Med Internet Res 2024;26:e52354) doi:10.2196/52354

KEYWORDS

suicide risk; suicide; journalism; media; self-harm; Google Trends; websites; mental health; depression; quality of life; harmful information

Introduction

Suicide has substantially increased in the United States, especially among young people [1]. There is evidence that exposure to explicit content about suicide, most notably after reports of high-profile suicide deaths, increases population rates of suicide [2,3]. Many studies have focused on news or television media, whereas individuals seek information and social support on social media and other online resources [4]. "Pro-suicide" websites are those that promote or facilitate suicide, including advocating for an individual's "right" to end their own life, creating communities of individuals who desire to end their own lives, as well as sharing explicit information on means of killing oneself [5,6]. These sites provide largely unmoderated forums for discussion of suicide, which may encourage, promote, or facilitate suicide. These sites are accessible and commonly used by individuals experiencing suicidal thoughts and otherwise vulnerable individuals [7]. Therefore, understanding trends in accessing these sites could inform preventive interventions and media practices when discussing such sites.

In 2021, a national news source published a story about pro-suicide websites and named one such site. We examined whether searches for the named site and similar sites increased following its publication.

Methods

National weekly Google data covering 6 months before and after the publication were drawn from the Google Health Trends API [8]. We queried search terms, including names of pro-suicide websites (available upon request). Queries randomly return 10% to 15% of total Google searches in the United States for given time periods as well as the probability of a short search session for the given terms per 10 million searches. We requested 10 datasets, which we averaged to estimate weekly search levels.

XSL•FO
RenderX

We first graphed the weekly time trend. To estimate whether searches increased during the week of the news publication, we used search activity from the previous 6 months to forecast expected search levels using an autoregressive integrated moving average (ARIMA) model. We then compared the expected search levels to the observed levels for that week.

An interrupted time series model was used to estimate the effects of the publication on the intercept and slope time trends 6 months after publication, removing the week of the publication. The model was defined as:

$$E(Y) = \beta_1 + \beta_2 * T_1 + \beta_3 * P + \beta_4 * T_2 \quad \textbf{(1)}$$

with T_1 as time since the beginning of the measurement period, T_2 as time since publication, and P as a binary indicator of before/after article publication.

Results

Figure 1 shows the average weekly relative search volume for the abovementioned website (panel A) and searches for a broader range of pro-suicide websites (panel B), indexed by time of the news story. In the week of the news story's publication, searches for pro-suicide websites and the mentioned website were 564% (95% CI 366%-1108%) and 842% (95% CI 393%-2979%) higher, respectively, than what would be expected based on the prior 6 months.

After excluding the week of publication, the interrupted time series model results showed that there was an intercept effect of searches for suicide websites increasing 80.2% (95% CI 41.0%-119%, P<.001; intercept B=5.26, 95% CI 2.69-7.83). Searches for the mentioned website increased by 90.2% (95% CI 34.0%-146%, P=.003; intercept B=1.75, 95% CI 0.66-2.84) compared to the mean of the prior 6 months, with no change in difference over time.

Figure 1. Time trends for searches related to (A) the website named in the news article and (B) pro-suicide websites.

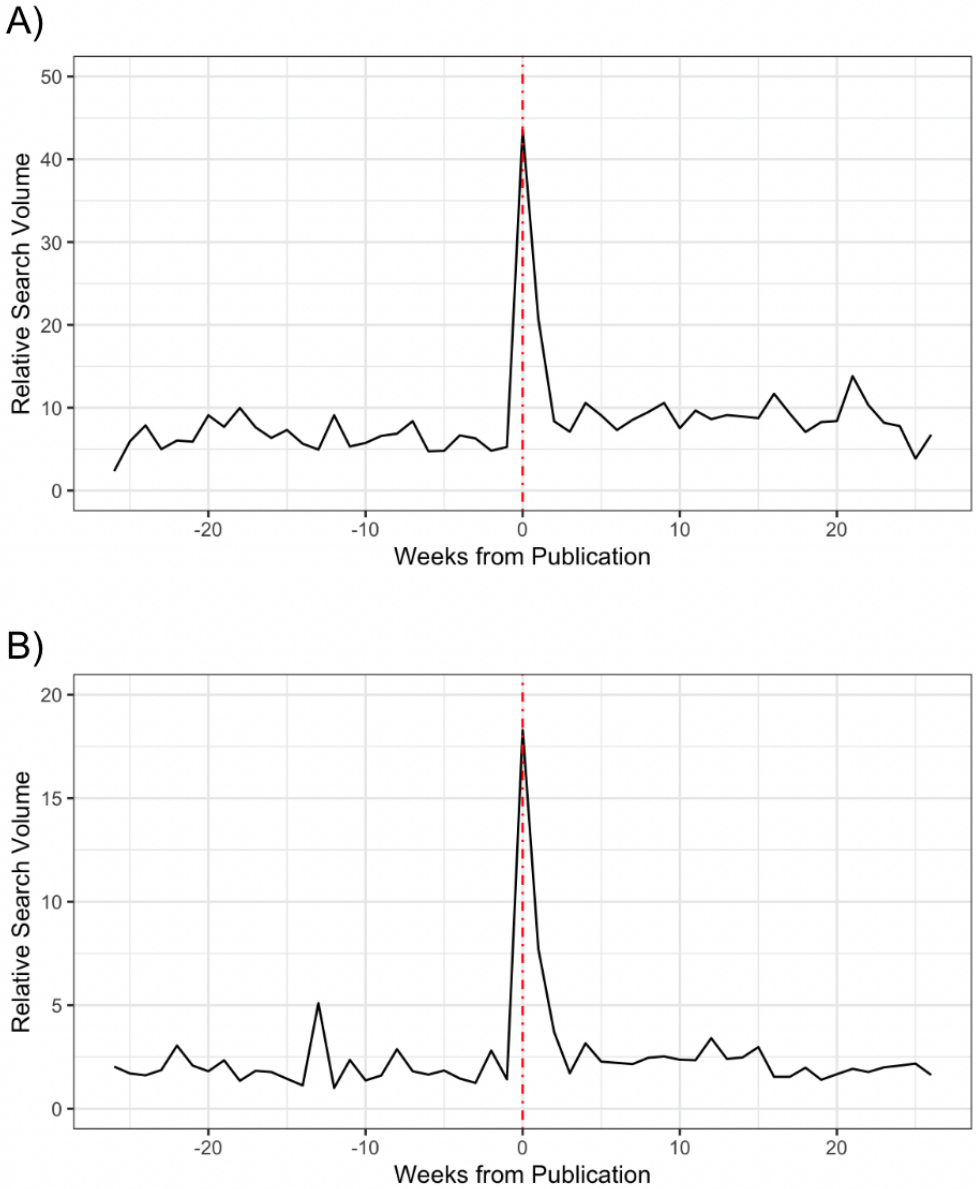

XSL•FO
RenderX

Discussion

Searches for pro-suicide websites in the United States peaked in the week of the publication of a high-profile news story and remained higher than prepublication times in the 6 months that followed. While we do not know the level of suicide risk for those accessing these sites and have no means of attributing increased searches to suicidal behavior, these results indicate that news attention to pro-suicide forums may increase the number of individuals seeking access.

Journalism around suicide requires balancing public interest with well-documented contagion and trigger phenomena [3]. For pro-suicide websites, there is a clear public interest in informing parents and providers about potentially harmful information. Yet publicity about pro-suicide websites has the potential to increase awareness among those vulnerable to its content. Providing accurate information and safe online spaces to discuss suicide may counteract more dangerous sources on the internet [9]. Further, for all terms searched in our study, Google lists the national 988 Suicide and Crisis Lifeline as the first result, which may mitigate potential harms. Nevertheless, we caution against mentioning specific sources of explicit instructions for suicide in media publications. Although internet use varies by geographical region and age group, the use of national Google Trends data does not allow for differentiating search volumes by user demographics or intent.

Acknowledgments

This work was funded by the National Institute of Mental Health (R01-MH121410).

Data Availability

The data used in these analyses are not publicly available in order to limit dissemination of pro-suicide website names but are available from the corresponding author upon reasonable request.

Authors' Contributions

NCK contributed to conceptualization, data curation, formal analysis, investigation, methodology, and visualization, as well as writing, reviewing, and editing of the draft. CG contributed to conceptualization, methodology, review, and editing. MO contributed to methodology, review, and editing, MG contributed to conceptualization, review, and editing. JS contributed to methodology, review, and editing. KK contributed to conceptualization funding, supervision, investigation, methodology, review, and editing.

Conflicts of Interest

JS and Columbia University declare partial ownership of SK Analytics.

References

1. Martínez-Alés G, Jiang T, Keyes KM, Gradus JL. The recent rise of suicide mortality in the United States. Annu Rev Public Health 2022 Apr 05;43(1):99-116 [FREE Full text] [doi: 10.1146/annurev-publhealth-051920-123206] [Medline: 34705474]
2. Niederkrotenthaler T, Kirchner S, Till B, Sinyor M, Tran US, Pirkis J, et al. Systematic review and meta-analyses of suicidal outcomes following fictional portrayals of suicide and suicide attempt in entertainment media. EClinicalMedicine 2021 Jun;36:100922 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100922] [Medline: 34308310]
3. Niederkrotenthaler T, Braun M, Pirkis J, Till B, Stack S, Sinyor M, et al. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. BMJ 2020 Mar 18;368:m575 [FREE Full text] [doi: 10.1136/bmj.m575] [Medline: 32188637]
4. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Jan 08;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
5. Westerlund M. The production of pro-suicide content on the internet: a counter-discourse activity. New Media & Society 2011 Nov 22;14(5):764-780. [doi: 10.1177/1461444811425221]
6. Durkee T, Hadlaczky G, Westerlund M, Carli V. Internet pathways in suicidality: a review of the evidence. IJERPH 2011 Oct 11;8(10):3938-3952 [FREE Full text] [doi: 10.3390/ijerph8103938] [Medline: 22073021]
7. Moon KC, Van Meter AR, Kirschenbaum MA, Ali A, Kane JM, Birnbaum ML. Internet search activity of young people with mood disorders who are hospitalized for suicidal thoughts and behaviors: qualitative study of Google Search activity. JMIR Ment Health 2021 Oct 22;8(10):e28262 [FREE Full text] [doi: 10.2196/28262] [Medline: 34677139]
8. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, del Valle S, et al. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis 2020 Mar 26;20(1):252 [FREE Full text] [doi: 10.1186/s12879-020-04957-0] [Medline: 32228508]
9. Robinson J, Hill NTM, Thorn P, Battersby R, Teh Z, Reavley NJ, et al. The #chatsafe project. Developing guidelines to help young people communicate safely about suicide on social media: a Delphi study. PLoS ONE 2018 Nov 15;13(11):e0206584 [FREE Full text] [doi: 10.1371/journal.pone.0206584] [Medline: 30439958]

XSL•FO
RenderX

Abbreviations

ARIMA: autoregressive integrated moving average

Edited by T de Azevedo Cardoso; submitted 31.08.23; peer-reviewed by B Till, J Abbas; comments to author 13.09.24; revised version received 27.09.24; accepted 27.09.24; published 15.10.24.

Please cite as:
Kelsall NC, Gimbrone C, Olfson M, Gould M, Shaman J, Keyes K
Internet Search Activity for Intentional Self-Harm Forums After a High-Profile News Publication: Interrupted Time Series Analysis
J Med Internet Res 2024;26:e52354
URL: https://www.jmir.org/2024/1/e52354
doi:10.2196/52354
PMID:39405095

©Nora Clancy Kelsall, Catherine Gimbrone, Mark Olfson, Madelyn Gould, Jeffrey Shaman, Katherine Keyes. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials

Tony K W Hung[1], MD, MBA, MSCR; Gilad J Kuperman[1], MD, PhD; Eric J Sherman[1], MD; Alan L Ho[1], MD, PhD; Chunhua Weng[2], PhD; David G Pfister[1], MD; Jun J Mao[1], MD, MSCE

[1]Memorial Sloan Kettering Cancer Center, New York, NY, United States

[2]Columbia University, Department of Biomedical Informatics, New York, NY, United States

Corresponding Author:
Tony K W Hung, MD, MBA, MSCR
Memorial Sloan Kettering Cancer Center
530 E 74th St
New York, NY, 10021
United States
Phone: 1 646 608 4127
Fax: 1 646 888 4269
Email: hungt@mskcc.org

(J Med Internet Res 2024;26:e60695) doi:10.2196/60695

KEYWORDS

large language model; LLM; ChatGPT; GPT-4; artificial intelligence; AI; clinical trials; decision support; LookUpTrials; cancer care delivery; head and neck oncology; head and neck cancer; retrieval augmented generation

Introduction

Chatbots based on large language models (LLMs) have demonstrated the ability to answer oncology examination questions with impressive accuracy without specialized training or reinforcement [1,2]; however, leveraging LLMs in oncology decision support has not yet demonstrated suitable performance, as LLMs would produce responses that deviate from cancer expert recommendations and guidelines [3-5]. Furthermore, the rapidly changing oncology landscape, including knowledge of cancer clinical trials, limits the meaningful use of LLMs in practice given delays in training dataset updates. To enhance LLM utility in oncology practice, we developed a retrieval-augmented LLM, powered by GPT-4, and evaluated its performance to provide appropriate clinical trial recommendations for a head and neck (HN) cancer population.

Methods

On February 1, 2022, we piloted a clinical trial knowledge management application, LookUpTrials, at the Memorial Sloan Kettering Cancer Center (MSK) [6]. Using LookUpTrials' real-time database, we applied retrieval-augmented generation architecture and direct preference optimization to fine-tune GPT-4 as a clinical trial decision assistant [7]. Specifically, we enabled retrieval-augmented GPT-4 to respond with up-to-date information—such as trial availability—developed initial prompts, and validated GPT-4 responses from 1120 preference pairs across 56 MSK HN clinical trials. Preference pairs were constructed in [trial : attributes] format, including 20 organizational, investigator, and study attribute types (Multimedia Appendix 1). Data labels were annotated by author TKWH and cross-verified by 2 trial managers. From November 7, 2023, to January 30, 2024, we collected all consecutive new patient cases and their respective clinical trial recommendations, which were made by consensus during a weekly HN conference attended by 5-8 oncologists with 2 to more than 25 years of practice experience. Cases were categorized by diagnosis, biomarkers, cancer stage, treatment setting, and physician recommendations on clinical trials. Using these cases as test datasets, we prompted retrieval-augmented GPT-4 using a semistructured template, as follows: "Given patient with a <biomarkers>, <diagnosis>, <cancer stage>, <treatment setting>, what are possible clinical trials?" (eg, given a patient with human papillomavirus–associated HN cancer, metastatic stage, in a first-line treatment setting, what are the possible clinical trials?). GPT-4 responses were compared with physician recommendations, with concordance defined a priori: a GPT-4 response was a true positive if it included the recommended clinical trial(s); a true negative if neither the GPT-4 response nor the physicians recommended any clinical trial(s); a false positive if the GPT-4 response recommended clinical trial(s) but physicians did not; and a false negative if the GPT-4 response did not recommend clinical trial(s) but the physicians did. We analyzed the performance of GPT-4 based on its response precision (positive predictive value), recall (sensitivity), and F_1-score (harmonic mean of precision and recall). We further analyzed subgroup performance by cancer

XSL•FO
RenderX

types and the presence of biomarkers. Statistical analyses were performed using JMP-17.2.0.

Ethical Considerations

MSK institutional review board approved the study (application number: 24-120).

Results

We analyzed 178 patient cases (mean age 66, SD 13.9 years), primarily male (n=134, 75.3%), with local/locally advanced cancers (n=121, 68.0%), including HN (n=109, 61.2%), thyroid (n=29, 16.3%), skin (n=16, 9.0%), or salivary gland (n=14, 7.9%) cancers (Table 1). Over one-third of cases had biomarkers (n=66, 37.1%). The majority were treated in the definitive setting with combined modality therapy (n=75, 42.1%), and a modest proportion were treated under clinical trials (n=18, 10.1%). Overall, retrieval-augmented GPT-4 achieved moderate performance (Table 2), matching physician clinical trial recommendations with 63.0% precision and 100.0% recall (F_1-score 0.77), narrowing a total of 56 HN clinical trials to a range of 0-4 relevant trials per patient case (mean 1, SD 1.2 trials). In comparison, baseline non–retrieval-augmented GPT-4 demonstrated 0.0% precision, recall, and F_1-score—given the lack of response specificity to MSK clinical trials. Subgroup precision varied by cancer types (HN cancers: 72.7%, skin cancers: 50.0%, salivary gland cancers: 36.4%, and thyroid cancers: 33.3%) and the presence of biomarkers (presence 72.7%, absent 62.1%).

XSL•FO
RenderX

Table 1. Baseline characteristics of patient cases (N=178).

Characteristics	Overall values, n (%)
Age (years), mean (SD)	66 (13.9)
Sex	
Female	44 (24.7)
Male	134 (75.3)
Cancer types	
Head and neck cancers	109 (61.2)
Oropharyngeal SCC[a]	49 (27.5)
Oral cavity SCC	22 (12.4)
Laryngeal SCC	18 (10.1)
Hypopharyngeal SCC	8 (4.5)
Other	12 (6.7)
Thyroid cancers	29 (16.3)
Anaplastic thyroid carcinoma	4 (2.2)
Differentiated thyroid carcinoma	25 (14.0)
Skin cancers	16 (9.0)
Salivary gland cancers	14 (7.9)
Adenoid cystic carcinoma	5 (2.8)
Nonadenoid cystic carcinoma	9 (5.1)
Other cancers	10 (5.6)
Cancer stage	
Local/locally advanced	121 (68.0)
Recurrent/metastatic	57 (32.0)
Biomarkers	
Present	66 (37.1)
HPV[b] or p16[c]	42 (23.6)
EBV[d]	5 (2.8)
BRAF[e] mutation	6 (3.4)
RET[f] mutation	2 (1.1)
AR[g]	2 (1.1)
HER2[h]	3 (1.7)
Other	6 (3.4)
None	113 (63.5)
Treatment settings	
Definitive	93 (52.2)
Palliative	51 (28.7)
Surveillance	15 (8.4)
Adjuvant	13 (7.3)
Diagnostic	6 (3.4)
Treatment modality	
Combined modality therapy	75 (42.1)
Primary systemic treatment	37 (20.8)

XSL•FO
RenderX

Characteristics	Overall values, n (%)
Primary surgical treatment	11 (6.2)
Primary radiation treatment	8 (4.5)
Best supportive care	5 (2.8)
Other	24 (13.5)
Clinical trials	18 (10.1)

[a]SCC: squamous cell carcinoma.
[b]HPV: human papillomavirus.
[c]p16: p16(INK4A) immunostain.
[d]EBV: Epstein-Barr virus.
[e]BRAF: V-Raf murine sarcoma viral oncogene homolog B.
[f]RET: Rearranged during transfection.
[g]AR: androgen receptor.
[h]HER2: human epidermal growth factor receptor 2.

Table 2. Performance of retrieval-augmented large language models in matching physician clinical trial recommendations.

Performance	Precision (%)	Recall (%)	F_1-score
Baseline GPT-4	0.0	0.0	0
Retrieval-augmented GPT-4	63.0	100.0	0.77
Subgroups (cancer types)			
Head and neck cancers	72.7	100.0	0.84
Thyroid cancers	33.3	100.0	0.50
Skin cancers	50.0	100.0	0.67
Salivary gland cancers	36.4	100.0	0.53
Other cancers	__[a]	—	—
Subgroups (biomarkers)			
Present	72.7	100.0	0.84
None	62.1	100.0	0.77

[a]Not applicable.

Discussion

Our study demonstrated that retrieval-augmented GPT-4 achieved moderate performance in matching physician clinical trial recommendations in HN oncology. Comparatively, our retrieval-augmented LLM outperformed its pre–fine-tuned baseline and exceeded the historical performance of pretrained LLMs for providing oncology treatment recommendations by 4-20 folds (F_1-score 0.04-0.19) [4]. Prior studies have evaluated LLM performance in matching patients to clinical trials, achieving high accuracy [8-10]; however, to our knowledge, our study is the first to evaluate an oncology-specific, retrieval-augmented LLM as a point-of-care, clinical trial decision support application. As our subgroup analyses demonstrated, LLM performance varies based on the specificity of the prompt and dataset, with enhanced precision achieved through reduced search ambiguity for biomarker-specific trials and cancer types with more well-defined datasets. Study limitations included small sample size, short-term assessment, cross-sectional design, disease-specific focus, and being conducted in a single institution, which limits generalizability and subgroup analyses; however, our study provides insights into the rarely measured performance of retrieval-augmented LLMs using real-world patient cases. Future research is needed to optimize LLMs' precision and stability and to assess their implementation and effectiveness as a scalable solution for enhancing clinical trial participation.

Acknowledgments

This work is supported in part by the Memorial Sloan Kettering Cancer Center Support (grant P30-CA008748) and the 2024 Conquer Cancer—Johnson & Johnson Innovative Medicine Career Development Award (AWD00003905). The corresponding

XSL•FO
RenderX

author has full access to all data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. We thank all our patients, providers, and administrative staff who supported the study.

Conflicts of Interest

TKWH is the founder of LookUpTrials by TeamX Health. ALH received compensation from or was a part of the advisory boards of Eisai, Exelixis, Novartis, Merck, Rgenta, Coherus, Kura oncology, Remix Therapeutics, McGivney Global Advisors, Prelude Therapeutics, Affyimmune, Elevar Therapeutics, Ayala, Nested Therapeutics, and AstraZeneca. He was the principal investigator of clinica trials for Eisai, Bayer, Genentech, AstraZeneca, Novartis, Merck, BMS, Versatem, Remix Therapuetics, Rgenta Therapeutics, Kura Oncology, Ayala, TILT Therapeutics, Hookipa, Novartis, Daiichi Sankyo, and Astellas. ALH is a co-inventor of patent “Lesional dosimetry methods for tailoring targeted radiotherapy in cancer" (Serial number 63/193700, filed 5/27/21) and serves on the Speaker Bureau of Physician Education Resources. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Preference pairs architecture.
[DOCX File , 16 KB - jmir_v26i1e60695_app1.docx]

References

1. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
2. Longwell JB, Grant RC, Hirsch I, Binder F, Jang RW, Krishnan RG. Large language models encode medical oncology knowledge: performance on the ASCO and ESMO examination questions. JCO Oncology Practice 2023 Nov;19(11_suppl):511-511. [doi: 10.1200/op.2023.19.11_suppl.511]
3. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, et al. Use of artificial intelligence Chatbots for cancer treatment information. JAMA Oncol 2023 Oct 01;9(10):1459-1462 [FREE Full text] [doi: 10.1001/jamaoncol.2023.2954] [Medline: 37615976]
4. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023 Nov 01;6(11):e2343689 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
5. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Jul 04:1-26. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
6. Hung KW, Dunn L, Sherman EJ, Ho AL, Fetten JV, Michel LS, et al. LookUpTrials: assessment of an artificial intelligence-powered mobile application to engage oncology providers in clinical trials. JCO GO 2023 Aug;9(Supplement_1):111-111. [doi: 10.1200/go.2023.9.supplement_1.111]
7. Rafailov R, Sharma A, Mitchell E, Manning C, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. arXiv Preprint posted online on July 29, 2024. [doi: 10.48550/arXiv.2305.18290]
8. Jin Q, Wang Z, Floudas CS. Matching patients to clinical trials with large language models. arXiv Preprint posted online on April 27, 2024. [doi: 10.48550/arXiv.2307.15051]
9. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-augmented generation–enabled GPT-4 for clinical trial screening. NEJM AI 2024 Jun 27;1(7). [doi: 10.1056/aioa2400181]
10. Wornow M, Lozano A, Dash D, Jindal J, Mahaffey K, Shah N. Zero-shot clinical trial patient matching with LLMs. arXiv Preprint posted online on April 10, 2024. [doi: 10.48550/arXiv.2402.05125]

Abbreviations

HN: head and neck
LLM: large language model
MSK: Memorial Sloan Kettering Cancer Center

XSL•FO
RenderX

Edited by Q Jin; submitted 18.05.24; peer-reviewed by S Chan, D Bracken-Clarke, F Chen; comments to author 20.06.24; revised version received 12.08.24; accepted 03.09.24; published 15.10.24.

Please cite as:
Hung TKW, Kuperman GJ, Sherman EJ, Ho AL, Weng C, Pfister DG, Mao JJ
Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials
J Med Internet Res 2024;26:e60695
URL: https://www.jmir.org/2024/1/e60695
doi:10.2196/60695
PMID:

©Tony K W Hung, Gilad J Kuperman, Eric J Sherman, Alan L Ho, Chunhua Weng, David G Pfister, Jun J Mao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Gender Bias in AI's Perception of Cardiovascular Risk

Margaux Achtari[1*], MSc; Adil Salihu[2*], MD; Olivier Muller[2], MD, PhD; Emmanuel Abbé[3], PhD; Carole Clair[1], MSc, MD; Joëlle Schwarz[1], PhD; Stephane Fournier[2], MD, PhD

[1]Health and Gender Unit, University Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland
[2]Department of Cardiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[3]Institute of Mathematics and School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
[*]these authors contributed equally

Corresponding Author:
Stephane Fournier, MD, PhD
Department of Cardiology
Lausanne University Hospital and University of Lausanne
21 Rue Du Bugnon
Lausanne, CH-1011
Switzerland
Phone: 41 21 314 00 12
Email: stephane.fournier@chuv.ch

Abstract

The study investigated gender bias in GPT-4's assessment of coronary artery disease risk by presenting identical clinical vignettes of men and women with and without psychiatric comorbidities. Results suggest that psychiatric conditions may influence GPT-4's coronary artery disease risk assessment among men and women.

(J Med Internet Res 2024;26:e54242) doi:10.2196/54242

KEYWORDS

artificial intelligence; gender equity; coronary artery disease; AI; cardiovascular; risk; CAD; artery; coronary; chatbot: health care; men: women; gender bias; gender

Introduction

With the emergence of large language models (LLMs), the use of artificial intelligence (AI) in health care settings is growing, but this is not without risks. While these algorithms offer advancements in diagnostics and clinical decision-making by analyzing vast amounts of data, they may unintentionally reinforce preexisting biases and inequities as they are trained on potentially biased data [1-3].

Gender bias refers to the systematic and often unconscious differential or inappropriate treatment and consideration of patients based on their gender [4]. This phenomenon has been extensively reported in the management of cardiovascular disease, with evidence showing that women are inadequately represented in research, underdiagnosed, and subject to treatment disparities [5,6]. Mental illness stigma represents another bias, leading to suboptimal care that may also impact women more significantly, with research suggesting that physicians are more likely to view men's symptoms as physical and women's as psychosocial [7,8].

This study investigated if LLMs reproduce existing gender biases when presented with clinical vignettes of chronic coronary syndrome with and without psychiatric comorbidities using GPT-4 (OpenAI) [9], an LLM trained on a diverse range of internet text, whose performance in the cardiovascular field has previously been validated [10].

Methods

We generated hypothetical clinical scenarios by matching men and women with identical chest pain types (typical, atypical, and nonanginal) using the same symptom descriptions. We adjusted their ages to achieve nearly identical pretest probabilities (with a maximal pretest probability gap of 5%) of obstructive coronary artery disease (CAD) based on the latest chronic coronary syndrome guidelines [11]. We submitted these cases to GPT-4, asking which patient, woman or man, had a higher risk for obstructive CAD (Textbox 1). We then added psychiatric comorbidities to the original scenarios (generalized anxiety disorder, borderline personality disorder, major depressive disorder, schizophrenia, and bipolar I disorder). These cases were submitted individually to GPT-4 to ensure no

XSL•FO
RenderX

carryover effects from previous discussions. Overall, 30 cases were presented to GPT-4, with 5 having no psychiatric comorbidities and 25 having psychiatric comorbidities. Mean age was 65.8 (SD 6.24) years for women and 52.8 (SD 4.96) years for men. The cases were presented with typical chest pain (n=12), atypical chest pain (n=12), and nonanginal symptoms (n=6).

Textbox 1. Example of the questions submitted to GPT-4.

Here is a multiple-choice question:

I am a general practitioner in a city in Europe and I have seen 2 patients with chest pain.

a) [age]-year-old female patient, without cardiovascular risk factor, with [typical/atypical/non-anginal] chest pain for one year.

b) [age]-year-old male patient, without cardiovascular risk factor, with [typical/atypical/non-anginal] chest pain for one year.

Which of these patients (a or b) is at a higher risk of having obstructive coronary artery disease?

Results

Among the 5 scenarios without psychiatric comorbidities, GPT-4 indicated that women had a higher risk of obstructive CAD in 100% of the cases, arguing that women's higher age was a decisive factor. When adding psychiatric conditions, GPT-4's response substantially changed, indicating that men had a higher risk of CAD in 56% (14/25) of the cases (Table 1).

Table 1. Assessment of patients with chest pain and equivalent pretest probabilities of obstructive coronary artery disease based on European Society of Cardiology guidelines for diagnosing and managing chronic coronary syndrome.

Patient group and gender	Risk, n (%)
Apparently healthy patient (n=5)	
Women	5 (100)
Men	0 (0)
Same patient with psychiatric disorder (n=25)	
Women	11 (44)
Men	14 (56)

Discussion

This study's primary finding is a substantial shift in the perception of risk between men and women when a psychiatric comorbidity is added to the vignette. Despite presenting identical complaints in scenarios without psychiatric comorbidities, nearly 1 in 2 women was suddenly assessed as having a lower pretest probability when concurrently having a psychiatric condition.

This suggests that the inclusion of a psychiatric comorbidity could alter the algorithm's assessment of CAD risk among men and women. This shift could be interpreted as a sign of mental illness stigma, affecting the risk assessment of patients with psychiatric comorbidities [8]. Although these findings should be confirmed with more cases, it is interesting to observe this in a smaller sample. Another complementary interpretation is that women's chest pain symptoms may be more frequently undervalued as being psychological compared to men's, as reported by Colameco et al [7]. Indeed, 120 physicians assessed the management of headache or abdominal pain based on gender and reported that women were perceived as being more emotional. Moreover, no scientific evidence suggests that any of these psychiatric conditions disproportionately increase the risk of CAD in men over women.

We acknowledge the low number of cases tested due to constraints in combining possibilities with the pretest probability table [11]. Moreover, our study only assessed the answer to the first prompt and did not evaluate the potential variability in GPT-4's responses. Consequently, the low number of cases presented for assessment does not yield strong power. However, this choice of analysis is also a strength, as the clinical cases were based on real pretest probabilities [11] and 1 case per scenario was used to mimic real-life scenarios.

Concerns have already been raised regarding GPT-4's decision-making mechanism, often perceived as a "black box." These preliminary data suggest that GPT-4 may underperform in marginalized groups and corroborate the need for explainable models and the integration of bias detection systems. These results warrant further investigation with different LLMs and clinical scenarios investigating other diseases.

Data Availability

The raw data from this study, as well as the prompts used, are available to readers by contacting the principal investigator, upon reasonable request.

XSL•FO
RenderX

Conflicts of Interest

OM has received honoraria and/or research grant from Edwards Lifesciences and Abbott.

References

1. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin A, Park H, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023 Sep 05;330(9):818-820 [FREE Full text] [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
2. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. JAMA 2020 Oct 13;324(14):1397-1398. [doi: 10.1001/jama.2020.9371] [Medline: 32926087]
3. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020 Jun 1;3:81 [FREE Full text] [doi: 10.1038/s41746-020-0288-5] [Medline: 32529043]
4. Hamberg K. Gender bias in medicine. Womens Health (Lond) 2008 May;4(3):237-243 [FREE Full text] [doi: 10.2217/17455057.4.3.237] [Medline: 19072473]
5. Vogel B, Acevedo M, Appelman Y, Bairey Merz CN, Chieffo A, Figtree GA, et al. The Lancet Women and Cardiovascular Disease Commission: reducing the global burden by 2030. Lancet 2021 Jun 19;397(10292):2385-2438. [doi: 10.1016/S0140-6736(21)00684-X] [Medline: 34010613]
6. Huber E, Le Pogam MA, Clair C. Sex related inequalities in the management and prognosis of acute coronary syndrome in Switzerland: cross sectional study. BMJ Med 2022 Nov 17;1(1):e000300 [FREE Full text] [doi: 10.1136/bmjmed-2022-000300] [Medline: 36936600]
7. Colameco S, Becker L, Simpson M. Sex bias in the assessment of patient complaints. J Fam Pract 1983 Jun;16(6):1117-1121. [Medline: 6854242]
8. Subu MA, Wati DF, Netrida N, Priscilla V, Dias JM, Abraham MS, et al. Types of stigma experienced by patients with mental illness and mental health nurses in Indonesia: a qualitative content analysis. Int J Ment Health Syst 2021 Oct 18;15(1):77 [FREE Full text] [doi: 10.1186/s13033-021-00502-x] [Medline: 34663399]
9. GPT-4. OpenAI. URL: https://openai.com/index/gpt-4/ [accessed 2024-10-14]
10. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023 May;4(3):279-281 [FREE Full text] [doi: 10.1093/ehjdh/ztad029] [Medline: 37265864]
11. Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J 2020 Jan 14;41(3):407-477 [FREE Full text] [doi: 10.1093/eurheartj/ehz425] [Medline: 31504439]

Abbreviations

AI: artificial intelligence
CAD: coronary artery disease
LLM: large language model

Edited by A Mavragani; submitted 02.11.23; peer-reviewed by K Jordan, A Bashir, J Walsh; comments to author 13.03.24; revised version received 11.06.24; accepted 10.09.24; published 22.10.24.

Please cite as:
Achtari M, Salihu A, Muller O, Abbé E, Clair C, Schwarz J, Fournier S
Gender Bias in AI's Perception of Cardiovascular Risk
J Med Internet Res 2024;26:e54242
URL: https://www.jmir.org/2024/1/e54242
doi:10.2196/54242
PMID:

©Margaux Achtari, Adil Salihu, Olivier Muller, Emmanuel Abbé, Carole Clair, Joëlle Schwarz, Stephane Fournier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Preliminary Screening for Hereditary Breast and Ovarian Cancer Using an AI Chatbot as a Genetic Counselor: Clinical Study

Ann Sato[1,2], MSc; Eri Haneda[1], MSc; Yukihiko Hiroshima[3,4], MD, PhD; Hiroto Narimatsu[1,2,5], MD, PhD

[1]Department of Genetic Medicine, Kanagawa Cancer Center, Yokohama, Kanagawa, Japan
[2]Graduate School of Health Innovation, Kanagawa University of Human Services, Kawasaki, Kanagawa, Japan
[3]Department of Cancer Genome Medicine, Kanagawa Cancer Center, Yokohama, Kanagawa, Japan
[4]Advanced Cancer Therapy Division, Kanagawa Cancer Center Research Institute, Yokohama, Kanagawa, Japan
[5]Cancer Prevention and Cancer Control Division, Kanagawa Cancer Center Research Institute, Yokohama, Kanagawa, Japan

Corresponding Author:
Hiroto Narimatsu, MD, PhD
Department of Genetic Medicine
Kanagawa Cancer Center
2-3-2 Nakao
Asahi-ku
Yokohama, Kanagawa, 241-8515
Japan
Phone: 81 045 520 2222
Email: hiroto-narimatsu@umin.org

Abstract

Background: Hereditary breast and ovarian cancer (HBOC) is a major type of hereditary cancer. Establishing effective screening to identify high-risk individuals for HBOC remains a challenge. We developed a prototype of a chatbot system that uses artificial intelligence (AI) for preliminary HBOC screening to determine whether individuals meet the National Comprehensive Cancer Network *BRCA1/2* testing criteria.

Objective: This study's objective was to validate the feasibility of this chatbot in a clinical setting by using it on a patient population that visited a hospital.

Methods: We validated the medical accuracy of the chatbot system by performing a test on patients who consecutively visited the Kanagawa Cancer Center. The participants completed a preoperation questionnaire to understand their background, including information technology literacy. After the operation, qualitative interviews were conducted to collect data on the usability and acceptability of the system and examine points needing improvement.

Results: A total of 11 participants were enrolled between October and December 2020. All of the participants were women, and among them, 10 (91%) had cancer. According to the questionnaire, 6 (54%) participants had never heard of a chatbot, while 7 (64%) had never used one. All participants were able to complete the chatbot operation, and the average time required for the operation was 18.0 (SD 5.44) minutes. The determinations by the chatbot of whether the participants met the *BRCA1/2* testing criteria based on their medical and family history were consistent with those by certified genetic counselors (CGCs). We compared the medical histories obtained from the participants by the CGCs with those by the chatbot. Of the 11 participants, 3 (27%) entered information different from that obtained by the CGCs. These discrepancies were caused by the participant's omissions or communication errors with the chatbot. Regarding the family histories, the chatbot provided new information for 3 (27%) of the 11 participants and complemented information for the family members of 5 (45%) participants not interviewed by the CGCs. The chatbot could not obtain some information on the family history of 6 (54%) participants due to several reasons, such as being outside of the scope of the chatbot's interview questions, the participant's omissions, and communication errors with the chatbot. Interview data were classified into the following: (1) features, (2) appearance, (3) usability and preferences, (4) concerns, (5) benefits, and (6) implementation. Favorable comments on implementation feasibility and comments on improvements were also obtained.

Conclusions: This study demonstrated that the preliminary screening system for HBOC using an AI chatbot was feasible for real patients.

(J Med Internet Res 2024;26:e48914) doi:10.2196/48914

XSL•FO
RenderX

KEYWORDS

hereditary cancer; familial cancer; IBM Watson; family history; medical history; cancer; feasibility; social network; screening; breast cancer; ovarian cancer; artificial intelligence; AI; chatbot; genetic; counselling; oncology; conversational agent; implementation; usability; acceptability

Introduction

The rapid adoption of digital health tools has significantly enhanced health care efficiency and quality [1]. Chatbots enable communication with patients. Their use includes delivering remote health services (eg, patient support, education, and health promotion) and providing administrative assistance (eg, managing routine tasks and supporting research). One key advantage of chatbots is resource optimization. By handling routine tasks, chatbots let health care professionals (HCPs) focus on clinical duties. Chatbots standardize health care and expand access, increasing patient opportunities for appropriate services, even in regions with limited medical staff [2]. In genetic medicine, which requires specialized expertise, chatbots can collect family medical history and assist HCPs in communicating with at-risk relatives [3,4]. However, practical implementation in genetic medicine remains limited, particularly in Japan. We developed a chatbot for preliminary screening of hereditary breast and ovarian cancer (HBOC) [5] and conducted a clinical study to validate its feasibility and potential usefulness in a clinical setting with real patients.

Methods

Overview

The chatbot system [5] was planned to be tested on 10 patients to estimate its medical accuracy and ability to perform preliminary screening, such as obtaining correct medical and family histories and assessing the risk of HBOC according to the National Comprehensive Cancer Network guidelines. The system uses Watson, IBM's cognitive computing service. Preoperation questionnaires and postoperation interviews validated data on the system's usability and acceptability. Additional details are provided in Multimedia Appendix 1.

Ethical Considerations

This study was approved by the institutional review board of Kanagawa Cancer Center (2020 - epidemiology 78). Written informed consent was obtained from all participants. An ID was assigned for each participant, and personal information was excluded. The correspondence table between individuals and IDs is stored in a secure database that is not connected to the internet, and the database terminal is stored under double-locked security. We provided each participant with JP ¥2000 (~US $15) as an honorarium.

Results

A total of 11 participants enrolled between October and December 2020 (Table 1). All participants completed the chatbot operation (Multimedia Appendix 2). Preliminary screening results are shown in Table 1. Qualitative data from postoperation interviews are shown in Table 2, with key comments in Multimedia Appendix 3.

XSL•FO
RenderX

Table 1. Participant characteristics and result.

Variable	Participants (n=11)[a]
Demographic factors	
Age (years), median (range)	52 (35-67)
Education[b], n (%)	
Low	0 (0)
Middle	2 (18)
High	9 (82)
Familiar with chatbots, n (%)	
Yes	4 (36)
No	6 (55)
I am not sure	1 (9)
Have used a chatbot, n (%)	
Yes	4 (36)
No	7 (64)
Reaction to new technology, n (%)	
Eager	6 (55)
Overwhelmed	4 (36)
It depends	1 (9)
Any cancers or past history of cancers, n (%)	
Yes	10 (91)[c]
Breast cancer	7 (64)
Ovarian cancer	1 (9)
Stomach cancer	1 (9)
Double cancer	1 (9)
No	1 (9)
Preliminary screening results[d]	
Eligibility of the *BRCA1/2* testing criteria	
Met	7 (64)
Consistent with experts' evaluations based on the same data	7 (64)
Consistent with experts' evaluations based on information gathered by the CGCs[e] before this study	7 (64)
Unmet	4 (36)
Consistent with experts' evaluations based on the same data	4 (36)
Consistent with experts' evaluations based on information gathered by the CGCs before this study	1 (9)
Inconsistent with experts' evaluations based on information gathered by the CGCs before this study	3[f] (27)

[a]Participants who consecutively visited the Department of Genetic Medicine at the Kanagawa Cancer Center and met the selection criteria were enrolled in this study. The inclusion criteria were male or female individuals aged ≥30 years during inclusion in this study, with a history of breast and/or ovarian cancer; healthy individuals with family histories of breast and/or ovarian cancer; and those who had been interviewed by CGCs and had provided their family histories to CGCs at the Kanagawa Cancer Center before the start of the study. Individuals with difficulty communicating in Japanese, those who could not operate the tablet (text input, tapping, and scrolling, among others), and those with difficulty completing the questionnaire were excluded from this study. The questionnaire was administered to each participant before they used the chatbot system (Multimedia Appendix 1). The participants accessed this system using the tablet we provided. One or two CGCs observed each participant operating the chatbot system. The CGCs responded to any questions from the participants. We measured the time required for each participant to complete the chatbot operation. Two CGCs (authors AS and EH) compared the medical and family histories collected via the chatbot with those previously obtained by CGCs immediately after the chatbot operation was completed. If a discrepancy was found, it was clarified whether it resulted from a system error or participant's omissions. The accuracy of preliminary screening with this system was evaluated by experts, including a doctor specializing in genetic medicine (author HN) and 2 CGCs, based on the National Comprehensive Cancer Network guidelines. After the chatbot operation, we conducted a semistructured interview, which lasted approximately 30

XSL•FO
RenderX

minutes. The interview focused on opinions and impressions about this system and was conducted once, according to the interview guide (Multimedia Appendix 1).

[b]Low: middle school graduate; middle: high school graduate; and high: vocational school, junior college, college, or graduate school graduate.

[c]Includes 1 individual diagnosed with hereditary breast and ovarian cancer.

[d]The median time for the medical and family history interview by this system was 17.0 minutes.

[e]CGC: certified genetic counselor.

[f]Based on information gathered by the CGCs before this study, 3 of the 4 participants who did not meet the *BRCA1/2* testing criteria via chatbot actually met the National Comprehensive Cancer Network criteria. Two discrepancies were due to communication errors with the chatbot, where family and medical histories were incorrectly entered as follows: one involved peritoneal cancer being entered as “other cancers” instead of “ovarian cancer,” and another where only past medical history was entered, omitting the patient’s current breast cancer. The third discrepancy was due to the lack of information regarding triple-negative breast cancer, which was excluded from the scope of the chatbot’s interview questions.

XSL•FO
RenderX

Table 2. Details of the qualitative data classification obtained from the post–chatbot operation interviews[a].

Category, subcategory, and detail classification	Participants who had comments, n
Feature	
Character design	
Friendly character design	3
Suggestion for improvement	1[b]
Tone and flow of speech	
Appropriate	3
Suggestion for improvement	8
Wording	—[c]
Question design	—
More words than necessary	—
Fewer words than necessary	—
Ease of operation	
User-friendly operability	5
Suggestion for improvement	10
Confirmation of input information	—
Large input volume	—
Input method	—
System to increase completion rates	
Suggestion for improvement	11
Nonessential feature	—
Nonexistent features	—
Correction of input errors	
Suggestion for improvement	2[d]
Appearance	
Avatar images	
Familiar avatar image	2
Suggestion for improvement	1[e]
Overall screen visibility	
Easy-to-read design	8
Suggestion for improvement	4
Graphic size	—
Image size	—
Color scheme	—
Usability and preferences	
Usability and preferences	
General comment about usability	7
Excellent usability of user interface	8
Depends on age	2
Suggestion for improvement	4
Graphical user interface	—
Expression range of the chatbot responses	—

XSL•FO
RenderX

Category, subcategory, and detail classification	Participants who had comments, n
Other	—
Concerns	
Security management of information	3
Poor network condition	2
Benefits	
Superior ease of use	5
Effective use of waiting time	5
Reduced time required for medical interview	2
Overall efficiency of the clinical experience	1
Expanded preliminary screening coverage	1
Research use	1
Safety in the COVID-19 pandemic	1
Implementation	
Excellent implementability	8
Considerations during operation	
Suggestion for improvement	6
In-person support	—
Prior explanation	—
Voice guidance	—
Feedback on screening results	—

[a]We transcribed the interview data verbatim and coded them accordingly. We established codes in which 1 code corresponded to 1 opinion. Subsequently, we categorized each code according to the structure of the questionnaire. The codes in each category were classified into subcategories through discussion between 2 experts (authors AS and HN). Classification into subcategories was based on the content of the responses, such as whether participants were describing the strength of the system or points needing improvement.

[b]A code described character personalities.

[c]Not applicable.

[d]Codes described nonessential features.

[e]A code described a nonessential feature.

Discussion

This study demonstrated that implementing the chatbot for preliminary HBOC screening is feasible in a clinical setting. Two factors support this conclusion. First, the chatbot achieved a 100% completion rate, higher than previous studies [4]. Most participants reported excellent usability in the qualitative interviews, suggesting that the system's interface positively influenced completion. Despite limited experience, participants successfully completed the operation, indicating its accessibility regardless of technological literacy. It could be adapted for both patients visiting hospitals and healthy individuals eligible for cancer screening. Previous studies have confirmed that chatbots can effectively enhance cancer screening by automating tasks and increasing participation rates [6,7]. Participant interviews indicate that there is potential to provide opportunities for genetic medicine to patients living in remote areas or those who experience anxiety regarding in-person consultations. If the system can serve a broader patient population, it can not only act as a substitute for expert-led preliminary screening and quality standardization but also enhance accessibility to genetic medicine. Second, the chatbot accurately determined participants' eligibility for *BRCA1*/2 testing. Given the medical accuracy maintained in the 2 phases of our study, the findings support clinical application, although further research with diverse medical and family histories is needed for validation.

Based on the interviews, we identified areas for improvement. Discrepancies between the medical and family histories reported by the CGCs and those recorded using the chatbot suggest inadequacies in the chatbot's interview, which could lead to missing individuals at high risk for HBOC. These discrepancies arose from incomplete questions from the chatbot, participants' omissions, or communication errors. Thus, the chatbot's questions should be extended to include all information gathered during the CGC interviews. Participants also reported difficulty viewing the family tree image, indicating that larger images and text or voice confirmations could improve accuracy. Furthermore, inappropriate wording in the chatbot's responses may have led to communication errors, highlighting the need to review the chatbot's language. We are currently improving the system.

XSL•FO
RenderX

We could not evaluate the chatbot's impact on patient visit behavior because this study excluded physician feedback. Future research should evaluate whether the chatbot increases genetic counseling and testing rates.

This study suggests that chatbots can effectively gather diverse information. Previous research has shown that chatbots enhance communication [3,8-10]. Therefore, chatbots could be used for routine medical tasks beyond preliminary screenings, potentially improving HCPs' effectiveness and streamlining workflow.

Acknowledgments

This study was supported partially by a Life Innovation Platform YOKOHAMA research grant from the Kihara Foundation and JPSS KAKENHI (23H03131). We thank Advanced Integration Technology (AIT) Inc. for providing technical support during the development of the chatbot system. We also thank Editage for English language editing. Generative artificial intelligence (AI) was not used in any portion of the manuscript writing.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to restrictions imposed by informed consent from participants in this study, which has been approved by the institutional review boards of the Kanagawa Cancer Center.

Conflicts of Interest

Kanagawa Prefectural Hospital Organization and Advanced Integration Technology (AIT) Inc. have patents related to this study (7213501 from the Japan Patent Office).

Multimedia Appendix 1
Supplementary methods.
[DOCX File , 789 KB - jmir_v26i1e48914_app1.docx]

Multimedia Appendix 2
Supplementary results.
[DOCX File , 32 KB - jmir_v26i1e48914_app2.docx]

Multimedia Appendix 3
Supplementary qualitative data.
[DOCX File , 30 KB - jmir_v26i1e48914_app3.docx]

References

1. Fahy N, Williams GA, Habicht T, Köhler K, Jormanainen V, Satokangas M, et al. Use of Digital Health Tools in Europe: Before, During and After COVID-19 (Policy Brief, No. 42.). Copenhagen, Denmark: European Observatory on Health Systems and Policies URL: https://www.ncbi.nlm.nih.gov/books/NBK576970/ [accessed 2024-11-15]
2. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024 Jul 23;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
3. Schmidlen T, Schwartz M, DiLoreto K, Kirchner HL, Sturm AC. Patient assessment of chatbots for the scalable delivery of genetic counseling. J Genet Couns 2019 Dec;28(6):1166-1177. [doi: 10.1002/jgc4.1169] [Medline: 31549758]
4. Welch BM, Allen CG, Ritchie JB, Morrison H, Hughes-Halbert C, Schiffman JD. Using a chatbot to assess hereditary cancer risk. JCO Clin Cancer Inform 2020 Sep;4:787-793 [FREE Full text] [doi: 10.1200/CCI.20.00014] [Medline: 32897737]
5. Sato A, Haneda E, Suganuma N, Narimatsu H. Preliminary screening for hereditary breast and ovarian cancer using a chatbot augmented intelligence genetic counselor: development and feasibility study. JMIR Form Res 2021 Feb 05;5(2):e25184 [FREE Full text] [doi: 10.2196/25184] [Medline: 33544084]
6. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021 Nov 29;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
7. Wen K, Dayaratna S, Slamon R, Granda-Cameron C, Tagai EK, Kohler RE, et al. Chatbot-interfaced and cognitive-affective barrier-driven messages to improve colposcopy adherence after abnormal Pap test results in underserved urban women: a feasibility pilot study. Transl Behav Med 2024 Jan 11;14(1):1-12. [doi: 10.1093/tbm/ibad064] [Medline: 38014626]
8. Al-Hilli Z, Noss R, Dickard J, Wei W, Chichura A, Wu V, et al. A randomized trial comparing the effectiveness of pre-test genetic counseling using an artificial intelligence automated chatbot and traditional in-person genetic counseling in women

XSL•FO
RenderX

newly diagnosed with breast cancer. Ann Surg Oncol 2023 Oct;30(10):5990-5996. [doi: 10.1245/s10434-023-13888-4] [Medline: 37567976]
9. Siglen E, Vetti HH, Augestad M, Steen VM, Lunde, Bjorvatn C. Evaluation of the Rosa chatbot providing genetic information to patients at risk of hereditary breast and ovarian cancer: qualitative interview study. J Med Internet Res 2023 Sep 01;25:e46571 [FREE Full text] [doi: 10.2196/46571] [Medline: 37656502]
10. Siglen E, Vetti HH, Lunde ABF, Hatlebrekke TA, Strømsvik N, Hamang A, et al. Ask Rosa - the making of a digital genetic conversation tool, a chatbot, about hereditary breast and ovarian cancer. Patient Educ Couns 2022 Jun;105(6):1488-1494 [FREE Full text] [doi: 10.1016/j.pec.2021.09.027] [Medline: 34649750]

Abbreviations

AI: artificial intelligence
CGC: certified genetic counselor
HBOC: hereditary breast and ovarian cancer
HCP: health care professional

Edited by A Mavragani; submitted 11.05.23; peer-reviewed by L Guo, L Garcia; comments to author 27.12.23; revised version received 15.03.24; accepted 05.11.24; published 27.11.24.

Please cite as:
Sato A, Haneda E, Hiroshima Y, Narimatsu H
Preliminary Screening for Hereditary Breast and Ovarian Cancer Using an AI Chatbot as a Genetic Counselor: Clinical Study
J Med Internet Res 2024;26:e48914
URL: https://www.jmir.org/2024/1/e48914
doi:10.2196/48914
PMID:

©Ann Sato, Eri Haneda, Yukihiko Hiroshima, Hiroto Narimatsu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Innovative Approaches to Menstruation and Fertility Tracking Using Wearable Reproductive Health Technology: Systematic Review

Lynnette Lyzwinski[1], MPhil, PhD; Mohamed Elgendi[2], PhD; Carlo Menon[1,2], PhD

[1]Menrva Research Group, School of Mechatronics Systems Engineering and Engineering Science, Simon Fraser University, Vancouver, BC, Canada
[2]Biomedical and Mobile Health Technology Laboratory, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

Corresponding Author:
Lynnette Lyzwinski, MPhil, PhD
Menrva Research Group, School of Mechatronics Systems Engineering and Engineering Science
Simon Fraser University
2540-1340 102 Avenue, Metro Vancouver
Vancouver, BC, V3TOA3
Canada
Phone: 1 778 302 1151
Email: lnl2@sfu.ca

Abstract

Background: Emerging digital health technology has moved into the reproductive health market for female individuals. In the past, mobile health apps have been used to monitor the menstrual cycle using manual entry. New technological trends involve the use of wearable devices to track fertility by assessing physiological changes such as temperature, heart rate, and respiratory rate.

Objective: The primary aims of this study are to review the types of wearables that have been developed and evaluated for menstrual cycle tracking and to examine whether they may detect changes in the menstrual cycle in female individuals. Another aim is to review whether these devices are effective for tracking various stages in the menstrual cycle including ovulation and menstruation. Finally, the secondary aim is to assess whether the studies have validated their findings by reporting accuracy and sensitivity.

Methods: A review of PubMed or MEDLINE was undertaken to evaluate wearable devices for their effectiveness in predicting fertility and differentiating between the different stages of the menstrual cycle.

Results: Fertility cycle–tracking wearables include devices that can be worn on the wrists, on the fingers, intravaginally, and inside the ear. Wearable devices hold promise for predicting different stages of the menstrual cycle including the fertile window and may be used by female individuals as part of their reproductive health. Most devices had high accuracy for detecting fertility and were able to differentiate between the luteal phase (early and late), fertile window, and menstruation by assessing changes in heart rate, heart rate variability, temperature, and respiratory rate.

Conclusions: More research is needed to evaluate consumer perspectives on reproductive technology for monitoring fertility, and ethical issues around the privacy of digital data need to be addressed. Additionally, there is also a need for more studies to validate and confirm this research, given its scarcity, especially in relation to changes in respiratory rate as a proxy for reproductive cycle staging.

(J Med Internet Res 2024;26:e45139) doi:10.2196/45139

KEYWORDS

fertility cycle; fertility monitoring; ovulation; menstruation; wearable devices; mHealth; reproductive health; wearable; fertility; menstrual; women's health; ovulate; sexual health; scoping; review method

XSL•FO
RenderX

Introduction

Overview

Digital technology has been developed over the years to allow female individuals to monitor and take control of their reproductive health. Technological developments that may assist female individuals in monitoring their fertility cycle hold great potential for preventing unplanned pregnancy and in helping female individuals achieve planned pregnancy.

To date, most research has focused on commercially available smartphone apps for fertility monitoring [1]. However, a previous review found that not all commercially available apps are evidence based [1]. Furthermore, most focus on tracking menstruation by the user manually entering data into the app and self-checking temperature [1,2]. There may be disadvantages to this method, including reporting errors, the need to recall dates concerning menstruation, and cycle irregularity, which may affect the accuracy of predicting the next cycle dates [3,4]. One study of 330 female individuals found that periodic tracking apps had made errors in predicting periods, with 57% (n=189) of female individuals having experienced an earlier period than estimated, while 72% (n=238) had a delayed period relative to what was anticipated [5]. Some reviews have also critiqued menstrual cycle–tracking apps for lacking evidence of their efficacy at times [1].

Recently, there have been technological advancements in this field that involve the detection of fertility and menstruation based on heart rate variability (HRV) as well as temperature changes [6] using wearable devices [7-13]. For example, the Oura Ring tracks menstruation through changes in the sleep cycle, heart rate (HR), and temperature [7,9,14]. The Ava bracelet is another novel wearable piece of smart jewelry that checks HRV to predict fertility [10]. These wearable devices predict menstrual cycle stages based on physiological changes, including differentiation between the luteal phase, ovulation, and menstruation.

The female reproductive cycle is typically 28-30 days, although there is variability in some female individuals who may have shorter or longer cycles [15]. The shedding of the endometrial lining in the uterus results in menstruation, formally classified as day 1 of the cycle [16], which takes place over a period of 4-6 days on average in most female individuals, although some have longer or shorter periods [15]. The follicular phase begins at the onset of menses as follicles begin to grow, with a rise in estrogen midphase [15] and with subsequent ovulation and release of an egg at the next stage at approximately 12-14 days before menstruation [17], although some female individuals may ovulate earlier at day 7 [18]. The fertile window is when a female individual is most fertile within 5 days before ovulation, accounting for sperm viability, and during ovulation when the egg is released and viable over 24 hours [16,19]. Luteinizing hormones (LH) surge when a female individual is ovulating and basal body temperature increases by 0.3 °C [20,21]. The fourth stage of the cycle, the luteal phase, lasts approximately 14 days, which is accompanied by rising progesterone levels and the midphase rise of estrogen, with thickening of the uterus for implantation [15,20]. The onset of the new cycle with menstruation occurs once again de novo, with drops in progesterone and estrogen if the egg is not fertilized along with a drop in body temperature back to regular levels [15,22].

Gaining insight into fertility in real time may be a very instructive form of fertility care, as well as a useful tool for reproductive health in female individuals. By monitoring their cycle through body temperature changes, female individuals may either best prevent or plan a pregnancy when they know when they are ovulating. It may also help female individuals with managing their menses by being aware of when they may expect their period as well as with general cycle and stage monitoring. However, little is known about whether these devices are accurate predictors of fertility cycle staging from ovulation to menses. It is also useful to know whether certain physiological parameters or a specific combination best predict fertility (eg, temperature alone vs temperature and HR). Additionally, it is essential to compare the types of wearables that are available, including their design and placement on the body.

Thus, there is a need to undertake a comprehensive review to evaluate what types of emerging wearable technologies for fertility cycle monitoring have been developed. It is of research interest to better understand where these devices may be worn and which physiological parameters are used to track fertility at different stages. In other words, what has been developed and evaluated for efficacy and accuracy, and what can be improved in future technological developments?

Aims

The primary aims of this study are to review the types of wearables that have been developed and evaluated for menstrual cycle tracking and to examine whether they may detect changes in the menstrual cycle in female individuals. Another aim is to review whether these devices are effective for tracking various stages in the menstrual cycle including ovulation and menstruation. Finally, the secondary aim is to assess whether the studies have validated their findings by reporting accuracy, sensitivity, or positive predictive value.

Methods

Study Guidelines

This review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [23]. A prior review protocol was drafted using the PRISMA Protocols for internal use among the research team but it was not externally published or registered prospectively.

Search Strategy and Study Eligibility

PubMed or MEDLINE was searched for papers published between July 1, 2012, and July 1, 2022, for all English-language papers. Google Scholar was also searched for additional papers. The detailed strategy on PubMed was the following MeSH (Medical Subject Headings): (“Wearable Electronic Devices” [Mesh] OR wearable* [tiab] OR “Smart Band*” [tiab] OR “Smart Watch*” [tiab] OR wristband OR ((device* [ti] OR sensor* [tiab] OR sensing [tiab] OR biosensor* [tiab]) AND

XSL•FO
RenderX

(wear* [tiab] OR worn [tiab]))) AND ("menstruation" [MeSH Terms] OR "Menstrual Cycle" [Mesh] OR "Fertility" [Mesh] OR menstruat* [tiab] OR Menstruation [tiab] OR menarche [tiab] OR menstrual period [tiab] OR menstru* [tiab] OR menses [tiab] OR catamenia [tiab] OR menarche [tiab] OR ovulation [tiab] OR ovulate [tiab] OR ovulatory [tiab] or "fertile window" [tiab] OR fertile [tiab] OR fertility [tiab] OR "LH surge" [tiab] OR "luteinizing hormone surge" [tiab] OR "luteal phase" [tiab]).

Gray literature was not included in this review in an attempt to only include peer-reviewed studies. This timeframe was chosen to reflect advances in smart sensors, artificial intelligence technologies, and their fertility tracking applications in medicine.

Inclusion and Exclusion Criteria

All studies that evaluated an emerging wearable technology device (developed within the last 10 years) for fertility cycle tracking were included. Nonwearable devices, such as mobile health apps without a paired wearable device, personal digital assistants, or digital variations of thermometers that were nonwearable, were excluded. Two reviewers (LL and ME) screened the titles, abstracts, and full texts independently for potentially eligible studies. Reference lists of eligible studies were also hand-searched, but no additional studies were included on this basis.

Study Selection and Data Extraction

Titles were screened for eligibility followed by abstracts. Each potential study for inclusion underwent full-text screening and was assessed for final eligibility and to extract study-specific data. Data were extracted and summarized in tabular format. For each of the included papers, we extracted information for the following: the year the paper was published, authors, authors' country, study purpose, study design, the technology used, the bodily location of the technology placement, clinical measurement, and evaluation metrics. The outcome was evaluated based on the assessment measures (accuracy and statistical tests) used by each study to identify menstrual cycle stages. The number of participants across the studies was of interest when assessing the overall strength or reliability of each study. Two reviewers verified the extracted data independently.

Results

Study Selection

A total of 13 studies met the inclusion criteria and were included in the review [7-13,24-28]. The flowchart illustrating the search process is summarized in Figure 1. Throughout the studies, changes in several physiological parameters were noted at different stages of the menstrual cycle. Notably, changes in HR, HRV, skin temperature, and respiratory rate were observed. Most of the studies evaluated wearable technology in the form of jewelry, such as the Oura Ring [7,9], or wrist-mounted sensors, such as the Ava bracelet [8,11-13].

A total of 2 studies used a wearable ear device to monitor skin temperature changes [25,26], and another 2 used an intravaginal biosensor (OvulaRing) [24,29]. Most used machine learning algorithms to predict stages of the menstrual cycle, including ovulation, based on physiological changes, such as temperature. LH levels were also objectively assessed via urine using Clearblue strips in most studies to confirm ovulation [7-13,22,24-28,30,31].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

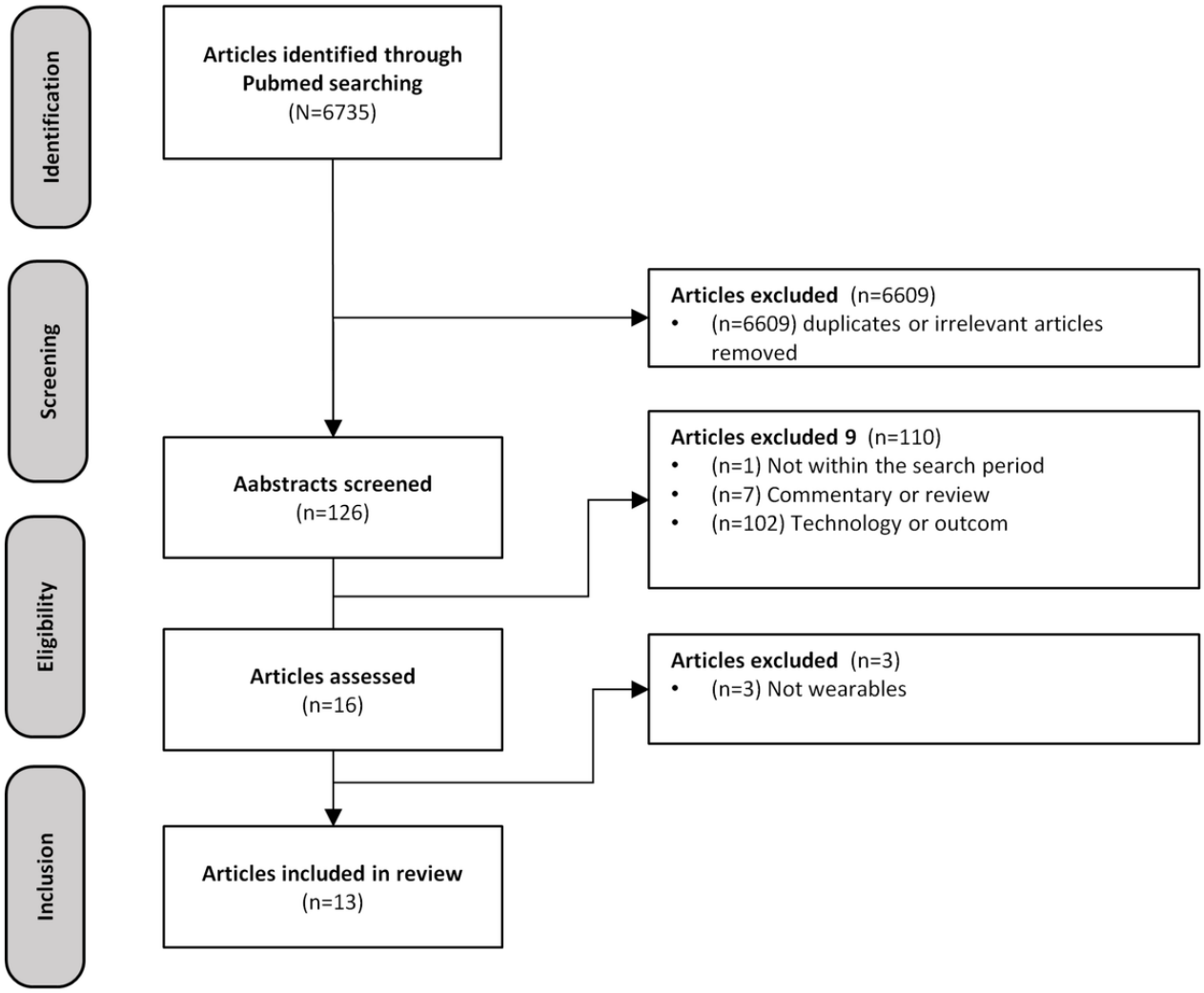

Changes in Skin Temperature During the Cycle

Throughout the studies, changes in skin temperature during different phases of the menstrual cycle were recorded [7-9,12,13,25]. Most studies found that skin temperature increased during the early and late luteal phases [7-9,12]. Several studies have noted that skin temperature was higher during the luteal phase than during menstruation and ovulation [7,8]. For example, 1 study found that female individuals' skin temperature is 0.33 °C higher during the luteal phase than during ovulation (fertile phase) (P<.001) [12]. Additionally, wrist skin temperature was found to be –0.72 °C lower during menstruation relative to the early luteal phase (P<.001) and –0.38°C lower than during the late luteal phase [12].

Additionally, studies have noted that temperature declined during the fertile phase, defined as 5 days leading up to and inclusive of ovulation, when compared with temperature during menstruation [8,13]. Goodale et al [8] found that skin temperature was 0.23 °C lower during the fertile window when compared with menstruation. The study by Zhu et al [13] found that body temperature increased after ovulation by 0.50 °C compared with basal body temperature increases of 0.20 °C, which was followed by a drop in temperature during menstruation of 0.33 °C using the skin temperature measure and 0.04 °C using basal body temperature. They also found that wrist-measured skin temperature was more sensitive and effective at measuring temperature changes during the cycle, with more pronounced temperature readings than basal body temperature, which was influenced by external factors, such as sleep, exercise, sexual intercourse, and caffeine intake [13].

HR and HRV

In addition to body temperature, several studies have reported an increased HR during the luteal phase [8,11]. This was found when comparing HR during the luteal phase with both the menstruation and ovulation phases. For instance, 1 study found that female individuals in the midluteal phase of their cycle had a higher HR than during menstruation by 3.8 beats per minute (P<.001) [11]. Another study found that the nocturnal HR increased by 2.55 beats per minute during the late luteal phase relative to menstruation (P<.001), but the HRV ratio was reduced (–0.16) at nighttime (P=.004) [8].

Additionally, HR was found to increase during ovulation relative to menstruation [11]. For example, the study by Shilaih et al [11] found a higher HR of 2.1 beats per minute in ovulating female individuals when compared with menstruating female individuals (P<.001). The relationship was not as significant in female individuals with cycle irregularity [11]. In another study

XSL•FO
RenderX

by Goodale et al [8], HRV was higher by 0.08 beats per minute (fertile) relative to menstruation. In addition to HR, 1 study evaluated respiratory rate and found that it increased during the late luteal phase by 0.18 breaths per minute relative to menstruation (P<.001) [8]. While the respiratory rate was reduced during the fertile period (follicular phase, ovulation, and early luteal phases) relative to menstruation, it was lower during menstruation relative to the late luteal phase [8]. This study also found that late-night meals and alcohol intake were all predictors of a rise in respiratory rate in female individuals [8].

Validity

We conducted a validity assessment to evaluate the accuracy and reliability of measurements or data obtained by wearable technology for assessing fertility.

Alzueta et al [7] found that the measurements of both nightly HRs in tandem with body temperature are statistically significantly elevated (P=.001 for HR and P<.001 for body temperature) during the luteal phases (mid to late) when compared with other phases of the menstrual cycle particularly menses. This suggests that there is a valid and measurable difference in HR and body temperature across different phases of the menstrual cycle.

Shilaih et al [11] further reinforced the validity of physiological measurements by reporting a significant difference in pulse rate between different phases of the menstrual cycle, with rising levels during ovulation compared with menstruation and notably during the midluteal phase. Additionally, their findings, along with a similar study on body temperature differences (Shilaih et al [12]), emphasize the robustness and accuracy of these measurements in tracking and understanding physiological changes throughout the menstrual cycle. Maijala et al [9] also found a statistically significant correlation between skin and oral temperature (measured via a thermometer; r=0.563; P<.001) [9]. Conversely, another study comparing the ActiGraph with the Ava bracelet found that the Ava bracelet had limited predictive ability in assessing changes in HR in female individuals (r=–0.28; P<.001) [10].

The sensitivity for detecting fertility ranged from 62% [13] for a wrist-worn sensor to 92% for an ear-worn device [28] and an abdominal sensor worn at night [25]. The study by Maijala et al [9] found that the Oura Ring had a sensitivity of 83% for detecting ovulation and a sensitivity that ranged from 71.9% to 86.5% for detecting menstruation. Similarly, OvulaRing, a vaginal biosensor that measures internal body temperature changes in relation to fertility, has an accuracy of 88% when measured during the fertile period (3 days before ovulation, during ovulation, and 3 days following) [29]. An inner ear device with a machine learning algorithm that measured nighttime temperature changes had a sensitivity of 92.3% for tracking fertility relative to conventional methods [25]. The Ava bracelet, which measures HR and body temperature changes, had an accuracy of 90% in 1 study for detecting the fertile period, defined as 5 days before and during ovulation [8]. However, the study authors did not benchmark the device against the gold standard with regard to heart assessment [8]. Another study of the Ava bracelet found that it was more sensitive to predicting fertility than conventional methods, which involved basal body temperature checks, with a reported sensitivity of 62% versus 23% [13].

Discussion

Principal Findings

This review aimed to establish a better understanding of the emerging wearables that have been developed, evaluated, and validated for menstrual cycle tracking including premenstruation, menstruation, and the fertile period (ovulation). Although we found that most wearables were worn externally on the wrist or finger, other emerging developments have included internally worn devices inside the ear and vagina to measure core body temperature rather than skin temperature. Most studies found that their devices were able to accurately detect physiological changes in temperature and HR in relation to the menstrual cycle stage. Only 1 study evaluated a wearable in relation to respiratory rate.

Interestingly, 1 study found that skin temperature is more accurate as a proxy for menstrual cycle stage estimation [13], given that wearing it on the wrist is less sensitive to other external stimuli [13] and may be evaluated continuously when the user is sleeping. Evaluating changes while sleeping may also have practical benefits if the fertility monitoring wearable is comfortable and the user does not need to wear the device continuously during the day.

Although wearables for menstruation tracking are still being developed, and more studies are needed to build the evidence base, it appears that they are a promising method for tracking fertility in female individuals. Such devices could be used as part of an array of products that promote the reproductive health of female individuals. Thus, they could be widely used as a method of tracking the fertile period and helping female individuals make informed decisions about their reproductive health, including avoiding conception or increasing their chances of conception around their peak fertility dates.

In addition to using these devices to track the fertile period, insight into the timing of menstruation may also be beneficial for some female individuals, such as those who experience premenstrual dysphoric dysfunction—a condition that causes anxiety, anger, irritability, and pain [32] 1 week before menstruation [33]. Such a condition is relevant as it occurs during the luteal phase [34], which can be tracked with fertility wearable devices. Additionally, some countries have implemented days off work for female individuals when they menstruate, such as Japan [35].

Premenstrual dysphoric disorder is also associated with worsening depression and anxiety, although it is clinically distinct [36]. Knowing when female individuals with premenstrual dysphoric disorder experience symptoms based on cycle tracking may help empower them through heightened awareness and help them plan their schedules accordingly. Likewise, female individuals with dysmenorrhea (prevalence from 45% to 95%), which is painful cramping during menses, may experience disruption to their daily activities [37,38]. By precisely timing the date of their period, they may be able to

XSL•FO
RenderX

plan their social and work schedules more efficiently and precisely.

One question of interest is whether wearables may be a better alternative to using calendars or digital apps. While studies have found that wearables are less accurate for female individuals who experience menstrual cycle irregularity relative to their regular counterparts [11,12], they nonetheless may provide an advantage over diary entry methods in this population. In other words, female individuals who have irregular cycles may, in fact, benefit from gaining insight into the predicted fertile dates based on their body temperature, HR, and respiratory rate, rather than relying on an electronic diary entry of past dates, which may not predict future ovulatory dates. In addition to this, recently, the Clue app was Food and Drug Administration approved and involves menstruation tracking through at-home ovulation test kits for detecting peaks in LH along with calendar tracking [39]. Thus, combined methods may perhaps be the best option, but research is needed to investigate this.

Skin Temperature Versus Basal Body Temperature Versus Vaginal Temperature

By contrast, intravaginal temperature monitoring has been found to be a good proxy for measuring core body temperature, especially with menstrual cycle monitoring [40]. It is interesting to note that 1 study in our review found that skin temperature was more accurate than basal body temperature and was one of the few prospective comparative efficacy studies, but it was worn on the wrist at night [13]. It is important to note that there are other factors that need to be considered including the comfort of the female individual who will be using these devices. It may not be as comfortable for female individuals to assess their core basal body temperature using an intravaginal monitor when compared with less invasive ear temperature or skin wrist temperature assessments, for example. Thus, there is a trade-off between comfort and full detection accuracy, something that future wearable companies should keep in mind when designing their devices to maximize accuracy and minimize any potential physical or psychological discomfort for the user.

Additionally, the fertile period takes place 5 days prior to ovulation [16,19], but the OvulaRing can only detect 3 days prior to ovulation, which is a major limitation as sperm may still fertilize an egg due to their 5-day viability [29]. Likewise, the Oura Ring can detect the fertile window 3 days prior to ovulation and 2 days after [9]. Detecting fertility 3 days after ovulation [29] is not of much use as the egg remains viable for only 24 hours after it is released [16,19]. This means that female consumers should be aware of these limitations as pregnancy may still occur in the 2 days prior. In addition to this, even female individuals with regular cycles may experience changes in their cycle as a result of stress [40] or an infection [41], including COVID-19, and elevated body temperature from infection [41] may further misconstrue the findings.

Moreover, the study designs had limitations given several were mostly 1-armed pilot studies with small sample sizes. There were some studies that were prospective with larger sample sizes that followed many cycles [8] or compared 2 different fertility tracking methods [13], for example (refer to Table 1). However, there is a need for more studies that will follow the gold standard including comparative prospective efficacy studies and randomized controlled trials [42] that will be sufficiently powered with large enough sample sizes to compare current wearables with other methods for tracking fertility. Recently, the Food and Drug Administration cleared the Nature Cycles app for fertility monitoring in female individuals [43]. However, due to a lack of sufficient efficacy, a recent lawsuit resulted from unplanned pregnancy in female individuals who had been using it [43]. This highlights the need to undertake rigorous efficacy studies before these digital fertility tracking products are commercially marketed directly to consumers.

Based on this study, intravaginal temperature monitoring is considered a good proxy for measuring core body temperature, particularly during menstrual cycle monitoring. However, 1 study found skin temperature to be more accurate than basal body temperature, albeit with limitations in wrist-worn devices. Comfort is also an important factor for users, as intravaginal monitoring may be less comfortable compared with less invasive methods like ear and skin temperature assessments. Our study also highlights the detection windows of different wearables, such as the OvulaRing and Oura Ring, and the need for more comprehensive comparative studies. Ensuring rigorous efficacy studies before marketing fertility tracking products to consumers is essential to avoid unforeseen consequences.

XSL•FO
RenderX

Table 1. Summary of the included studies.

Year	Study	Sample size, N	Technology	Measures	Results
2022	Alzueta et al [7]	26	Oura Ring	HR[a], HRV[b], and distal temperature	• HR was higher during the midluteal phase vs menstruation (*P*=.001) • Higher HR during late luteal (*P*=.001) vs menses • Ovulation linked with lower HR vs luteal phases including midluteal (*P*=.02) and late luteal phases (*P*=.01) • Significant rise in body temperature during mid- and late luteal phases compared with menstruation and ovulation (*P*<.001) • Marginally lower body temperature during ovulation compared with menstruation (*P*=.05)
2019	Maijala et al [9]	22	Oura Ring	Nocturnal finger skin temperature	• Higher skin temperature during luteal phase vs follicular (0.30 °C, SD 0.12), and compared to oral temperature during this phase of increase (0.23 °C, SD 0.09; *P*<.001) • Sensitivity for ovulation detection using the Oura ring=83.3% (–3 to +2 days) • Sensitivity for menstruation detection=71.9%-86.5% (SD 2-4 days) • Skin temperature correlation with oral temperature (thermometer); r=0.563 (*P*<.001) • Differentiating cycle stages; *r*=0.589 (*P*=.004)
2022	Nulty et al [10]	33	Ava bracelet	Skin temperature, resting pulse rate, HRV ratio, skin perfusion, breathing rate, and movement	• Acceptability for ovulation tracking is high, 19.3/25 (77%), with poor predictive ability • HRV compared with HR chest strap for direct heart monitoring + ActiGraph • Mean percentage error HRV=11.4% • Correlation against gold standard, –r=–0.28 (*P*<.001)
2018	Regidor et al [29]	158	OvulaRing	Core body temperature	• Accuracy is 88% for detecting ovulation (SD of 3 days around ovulation)
2017	Shilaih et al [11]	91	Wrist-worn photoplethysmography sensor	HR or pulse rate	• Rise in pulse rate during ovulatory phase relative to menstruation (median): +2.1 beats/min (*P*<.001) • Midluteal phase: pulse rate highest vs ovulation (fertile period): + 1.8 beats/min (*P*<.01) and relative to menses :3.8 beats/ min (*P*<.001)
2018	Shilaih et al [12]	136	Wrist-worn photoplethysmography sensor: Ava bracelet	Body temperature (wrist skin)	• Increased temperature on the wrist during luteal phase (+0.33 °C relative to ovulation [fertile window]); *P*<.001 • Lower wrist temperature during menstruation than during luteal phases including late luteal phase (–0.38 °C; *P*<.001) • Change in temperature during the fertile period=86%
2019	Goodale et al [8]	193	Ava wrist bracelet	HRV, HR, respiratory rate, and wrist skin temperature	• Wrist temperature significantly lower during the fertile phase relative to menstruation (b=–0.25, SD 0.03; *P*<.001) • HR lower during fertile phase compared with menstruation (b=–0.03, SD 0.26; not significant) • HRV significantly higher during fertile stage (b=0.08, SD 0.03; *P*<.05) relative to the comparison of menstruation • Lower respiration rate during fertile period vs menses (b=–0.48, SD 0.04; *P*<.01) • HR is higher in luteal and midluteal relative to menstruation (*P*<.01) • Late luteal phase relative to menses increases in respiratory rate by 0.18 breaths per min (*P*<.001) • Accuracy for detecting fertility=90%
2019	Luo et al [25]	34	Ear device	Temperature	• Ovulation detection • Sensitivity =92.3% • Increased predictive power 23%-31.5% • Compared to conventional methods of fertility tracking (kit)
2021	Zhu et al [13]	63	Ava fertility bracelet	Wrist skin temperature and basal body temperature	• Wrist skin temperature decreases during menstruation (0.33 °C vs 0.04 °C for basal body temperature) • Sensitivity higher for the wrist temperature measure vs the standard basal body temp (oral thermometer) measure (0.62 compared to 0.23; *P*<.001) • False positive rate greater for wrist skin temperature by 5.2% compared with basal body temperature

XSL•FO
RenderX

Year	Study	Sample size, N	Technology	Measures	Results
2022	Hurst et al [24]	80	Skin worn sensor + vaginal biosensor algorithm, OvuSense, OvuCore, OvuFirst	Body temperature	• Accuracy for ovulation prediction=66% (SD 1 day) • 90% accuracy for the fertile window (ovulation SD 3 days)
2020	Prasannan and Sarath [26]	30	Internet of things device with a flex sensor worn on the abdomen	Body temperature	• ovulation prediction • Accuracy=92%
2022	Yu et al [27]	114	Huawei band + ear thermometer	Body temperature and HR	• Fertility window • Accuracy=87.4% • Menstruation cycle • Accuracy=89.6%
2009	Chen et al [28]	30	Abdominal sensor worn during sleep	Body temperature	• Menstrual cycle stage sensitivity=91.8% • PPV[c]=96.6%

[a]HR: heart rate.

[b]HRV: heart rate variability.

[c]PPV: positive predictive value.

Study Quality and Design Assessment

There are several limitations with wearables for fertility tracking. First, some of the studies on temperature validated their findings and reported relatively good accuracy, but not all studies did, which limits our overall assessment of validity. Second, many used skin temperatures to assess changes in the menstrual cycle. Research indicates that skin temperature may not be as accurate as basal body temperature because it is influenced by the external environment and resultant physiological changes such as perspiration from heat or physical activity [40]. Research has found that intraoral and ear-based temperature methods are less accurate for predicting menstrual cycle phases [44]. Several studies have explored the potential of wearable devices for fertility and menstrual cycle tracking. Alzueta et al [7] provided valuable insights with their study involving 26 participants and the use of the Oura Ring to measure HR, HRV, and distal temperature. Similarly, Maijala et al [9] reported encouraging sensitivity and positive predictive value for nocturnal skin temperature with 22 participants using the Oura Ring, although the small sample size may limit generalizability. Nulty et al [10] raised questions about the Ava bracelet's HRV as a predictor for fertility tracking in their study with 35 participants. However, larger-scale studies and accounting for potential confounders could enhance the study quality. On a more promising note, Regidor et al [29] conducted a study with 158 participants, evaluating the accuracy of OvulaRing in measuring core body temperature, reporting promising results that warrant further consideration and addressing potential biases. Shilaih et al [11,12] conducted 2 studies with 91 and 136 participants, respectively, using wrist-worn photoplethysmography sensors and the Ava bracelet to measure changes in HR or pulse rate and body temperature (skin) throughout the cycle. The observed significant differences in pulse rate and body temperature between menstrual phases are intriguing and should be further investigated by controlling for confounders. Goodale et al [8] conducted a study with 193 participants, using the Ava wrist bracelet to measure HRV, HR, respiratory rate, and wrist skin temperature. The study provided valuable insights into the changes observed across the menstrual cycle, and future research should address potential biases related to participants' daily activities and lifestyle factors affecting the measured parameters. Luo et al [25] reported high sensitivity in detecting ovulation using an ear device for temperature measurement involving a total of 34 participants. Zhu et al [13] explored the effectiveness of the Ava fertility bracelet in detecting ovulation based on wrist skin temperature and basal body temperature with 63 participants. The reported higher sensitivity for wrist skin temperature is intriguing and merits further investigation to address potential biases related to device placement and calibration. Similarly, Hurst et al [24] evaluated a combination of skin-worn sensors, vaginal biosensors, and algorithms for predicting varying menstrual cycle stages with a total of 80 participants, demonstrating promising accuracy for predicting fertile periods. However, potential biases related to algorithm development and data processing should be addressed to ensure reliable results. Prasannan and Sarath [26] applied an internet of things device with a flex sensor worn on the abdomen to evaluate body temperature with a total of 30 participants. While notable accuracy for ovulation prediction was observed, future studies should consider potential biases related to sensor placement and participant compliance. Finally, Yu et al [27] conducted a study with 114 participants, using the Huawei band with an ear thermometer to measure body temperature along

XSL•FO
RenderX

with HR. The reported accuracy, sensitivity, and specificity for fertility and menstruation are encouraging, but future research should address potential biases related to device calibration and participant adherence to enhance the reliability of the findings.

Wearable fertility tracking studies offer valuable insights, but their study design, potential biases, and limitations should be carefully considered. Future research with larger sample sizes, more rigorous designs, and controlled conditions will contribute to a better understanding of the reliability and effectiveness of wearable technologies in fertility tracking.

Future Research

Furthermore, more qualitative consumer-focused research is needed to learn about the experiences of female individuals who use these devices to improve the technology, address any barriers, and promote the facilitators of use. Little is known about whether female individuals find these devices comfortable to wear, if they have any specific design preferences, or whether they are interested in using these emerging technological devices for long-term fertility and menstrual cycle monitoring. Privacy [45] and ethical issues around the digital health monitoring of fertility have been especially relevant during the time of Roe v. Wade in the United States, where many female individuals no longer feel safe sharing their private information concerning their menstrual cycles [46]. Thus, it is especially prudent to ensure that the privacy of the data associated with wearable technology is secured. It is also important to determine whether this form of fertility monitoring feels acceptable and useful for female individuals in the United States in the current climate.

Finally, future research should involve more studies to expand the literature base that will involve quality trials comparing the wearables with other methods for fertility detection with larger sample sizes. There is a clear need for well-designed efficacy studies for both pregnancy prevention and time to pregnancy in the future.

A limitation of our review is that we focused on wearables that have been scientifically assessed in our MEDLINE or PubMed search. There is a need for broader research into all commercially available devices in the future for female individuals to be well informed about other products that they may be using on the market, but there is also a need for more evidence-based efficacy studies to support them.

Conclusions

In summary, we reviewed wearable technology devices for fertility cycle monitoring. Although there are not many studies on this topic, the emerging research suggests that these devices hold promise as an accurate method of detecting fertility cycle stages, including the fertile window before and during ovulation, menstruation, and differentiation of the luteal phases (early to late). Most studies assessed physiological changes, such as temperature, and a combination of changes that included HR and HRV. Respiratory rate was also used to predict fertility cycle stages, but more research is needed to validate this method. Overall, the studies found that these emerging wearables are promising for menstrual cycle monitoring. Wearables for fertility cycle monitoring appear to be a valuable adjunct to the current arsenal of fertility tracking and reproductive health available to female individuals. However, more research is needed on the consumer acceptability of these devices and how to protect female individuals' privacy best when wearing digital devices that collect continuous data on their reproductive health. There is also a need for more comparative prospective efficacy studies to validate these devices for their efficacy.

Data Availability

All data generated or analyzed during this study are included in this published paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 30 KB - jmir_v26i1e45139_app1.docx]

References

1. Earle S, Marston HR, Hadley R, Banks D. Use of menstruation and fertility app trackers: a scoping review of the evidence. BMJ Sex Reprod Health 2021;47(2):90-101 [FREE Full text] [doi: 10.1136/bmjsrh-2019-200488] [Medline: 32253280]
2. Li K, Urteaga I, Wiggins CH, Druet A, Shea A, Vitzthum VJ, et al. Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile-health data. NPJ Digit Med 2020;3:79 [FREE Full text] [doi: 10.1038/s41746-020-0269-8] [Medline: 32509976]
3. Epstein DA, Lee NB, Kang JH, Agapie E, Schroeder J, Pina LR, et al. Examining menstrual tracking to inform the design of personal informatics tools. 2017 Presented at: CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; May 6-11, 2017; Denver, Colorado, USA p. 6876-6888. [doi: 10.1145/3025453.3025635]
4. Worsfold L, Marriott L, Johnson S, Harper JC. Period tracker applications: what menstrual cycle information are they giving women? Womens Health (Lond) 2021;17:17455065211049905 [FREE Full text] [doi: 10.1177/17455065211049905] [Medline: 34629005]

XSL•FO
RenderX

5. Broad A, Biswakarma R, Harper JC. A survey of women's experiences of using period tracker applications: attitudes, ovulation prediction and how the accuracy of the app in predicting period start dates affects their feelings and behaviours. Womens Health (Lond) 2022;18:17455057221095246 [FREE Full text] [doi: 10.1177/17455057221095246] [Medline: 35465788]
6. Grant AD, Newman M, Kriegsfeld LJ. Ultradian rhythms in heart rate variability and distal body temperature anticipate onset of the luteinizing hormone surge. Sci Rep 2020;10(1):20378 [FREE Full text] [doi: 10.1038/s41598-020-76236-6] [Medline: 33230235]
7. Alzueta E, de Zambotti M, Javitz H, Dulai T, Albinni B, Simon KC, et al. Tracking sleep, temperature, heart rate, and daily symptoms across the menstrual cycle with the Oura ring in healthy women. Int J Womens Health 2022;14:491-503 [FREE Full text] [doi: 10.2147/IJWH.S341917] [Medline: 35422659]
8. Goodale BM, Shilaih M, Falco L, Dammeier F, Hamvas G, Leeners B. Wearable sensors reveal menses-driven changes in physiology and enable prediction of the fertile window: observational study. J Med Internet Res 2019;21(4):e13404 [FREE Full text] [doi: 10.2196/13404] [Medline: 30998226]
9. Maijala A, Kinnunen H, Koskimäki H, Jämsä T, Kangas M. Nocturnal finger skin temperature in menstrual cycle tracking: ambulatory pilot study using a wearable Oura ring. BMC Womens Health 2019;19(1):150 [FREE Full text] [doi: 10.1186/s12905-019-0844-9] [Medline: 31783840]
10. Nulty AK, Chen E, Thompson AL. The Ava bracelet for collection of fertility and pregnancy data in free-living conditions: an exploratory validity and acceptability study. Digit Health 2022;8:20552076221084461 [FREE Full text] [doi: 10.1177/20552076221084461] [Medline: 35295766]
11. Shilaih M, de Clerck V, Falco L, Kübler F, Leeners B. Pulse rate measurement during sleep using wearable sensors, and its correlation with the menstrual cycle phases, a prospective observational study. Sci Rep 2017;7(1):1294 [FREE Full text] [doi: 10.1038/s41598-017-01433-9] [Medline: 28465583]
12. Shilaih M, Goodale BM, Falco L, Kübler F, de Clerck V, Leeners B. Modern fertility awareness methods: wrist wearables capture the changes in temperature associated with the menstrual cycle. Biosci Rep 2018;38(6):BSR20171279 [FREE Full text] [doi: 10.1042/BSR20171279] [Medline: 29175999]
13. Zhu TY, Rothenbühler M, Hamvas G, Hofmann A, Welter J, Kahr M, et al. The accuracy of wrist skin temperature in detecting ovulation compared to basal body temperature: prospective comparative diagnostic accuracy study. J Med Internet Res 2021;23(6):e20710 [FREE Full text] [doi: 10.2196/20710] [Medline: 34100763]
14. Starling MS, Kandel Z, Haile L, Simmons RG. User profile and preferences in fertility apps for preventing pregnancy: an exploratory pilot study. Mhealth 2018;4:21 [FREE Full text] [doi: 10.21037/mhealth.2018.06.02] [Medline: 30050917]
15. Reed BG, Carr BR. The normal menstrual cycle and the control of ovulation. In: Feingold KR, Anawalt B, Blackman MR, Boyce A, Chrousos G, Corpas E, et al, editors. Endotext. South Dartmouth (MA): MDText.com, Inc; 2000.
16. McLaughlin JE. Menstrual cycle. MSD Manual. 2002. URL: https://www.msdmanuals.com/home/women-s-health-issues/biology-of-the-female-reproductive-system/menstrual-cycle [accessed 2023-11-24]
17. Normal periods: menstrual cycle basics. The Society of Obstetricians and Gynaecologists of Canada. 2023. URL: https://www.yourperiod.ca/normal-periods/menstrual-cycle-basics/ [accessed 2023-11-24]
18. Wilcox AJ, Dunson D, Baird DD. The timing of the "fertile window" in the menstrual cycle: day specific estimates from a prospective study. BMJ 2000;321(7271):1259-1262 [FREE Full text] [doi: 10.1136/bmj.321.7271.1259] [Medline: 11082086]
19. Menstrual cycle. Better Health Channel. 2023. URL: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/menstrual-cycle [accessed 2023-11-24]
20. Mesen TB, Young SL. Progesterone and the luteal phase: a requisite to reproduction. Obstet Gynecol Clin North Am 2015;42(1):135-151 [FREE Full text] [doi: 10.1016/j.ogc.2014.10.003] [Medline: 25681845]
21. Basal body temperature for natural family planning. Mayo Clinic. 2023. URL: http://tinyurl.com/bdd8vbvy [accessed 2023-11-24]
22. Steward K, Raja A. Physiology, ovulation and basal body temperature. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.
23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336-341 [FREE Full text] [doi: 10.1016/j.ijsu.2010.02.007] [Medline: 20171303]
24. Hurst BS, Davies K, Milnes RC, Knowles TG, Pirrie A. Novel technique for confirmation of the day of ovulation and prediction of ovulation in subsequent cycles using a skin-worn sensor in a population with ovulatory dysfunction: a side-by-side comparison with existing basal body temperature algorithm and vaginal core body temperature algorithm. Front Bioeng Biotechnol 2022;10:807139 [FREE Full text] [doi: 10.3389/fbioe.2022.807139] [Medline: 35309997]
25. Luo L, She X, Cao J, Zhang Y, Li Y, Song PXK. Detection and prediction of ovulation from body temperature measured by an in-ear wearable thermometer. IEEE Trans Biomed Eng 2020;67(2):512-522. [doi: 10.1109/TBME.2019.2916823] [Medline: 31095472]

XSL•FO
RenderX

26. Prasannan R, Sarath G. IoT based device for fertility monitoring. 2020 Presented at: 2020 5th International Conference on Communication and Electronics Systems (ICCES); June 10-12, 2020; Coimbatore, India p. 813-817. [doi: 10.1109/icces48766.2020.9137904]
27. Yu JL, Su YF, Zhang C, Jin L, Lin XH, Chen LT, et al. Tracking of menstrual cycles and prediction of the fertile window via measurements of basal body temperature and heart rate as well as machine-learning algorithms. Reprod Biol Endocrinol 2022;20(1):118 [FREE Full text] [doi: 10.1186/s12958-022-00993-4] [Medline: 35964035]
28. Chen W, Kitazawa M, Togawa T. Estimation of the biphasic property in a female's menstrual cycle from cutaneous temperature measured during sleep. Ann Biomed Eng 2009;37(9):1827-1838. [doi: 10.1007/s10439-009-9746-6] [Medline: 19551509]
29. Regidor PA, Kaczmarczyk M, Schiweck E, Goeckenjan-Festag M, Alexander H. Identification and prediction of the fertile window with a new web-based medical device using a vaginal biosensor for measuring the circadian and circamensual core body temperature. Gynecol Endocrinol 2018;34(3):256-260 [FREE Full text] [doi: 10.1080/09513590.2017.1390737] [Medline: 29082805]
30. Kambic RT, Lamprecht V. Calendar rhythm efficacy: a review. In: Database of Abstracts of Reviews of Effects (DARE): Quality-Assessed Reviews. York (UK): Centre for Reviews and Dissemination (UK); 1995.
31. Sung S, Abramovitz A. Natural family planning. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.
32. Mishra S, Elliott H, Marwaha R. Premenstrual dysphoric disorder. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.
33. Hantsoo L, Epperson CN. Premenstrual dysphoric disorder: epidemiology and treatment. Curr Psychiatry Rep 2015;17(11):87 [FREE Full text] [doi: 10.1007/s11920-015-0628-3] [Medline: 26377947]
34. Yonkers KA, Pearlstein T, Fayyad R, Gillespie JA. Luteal phase treatment of premenstrual dysphoric disorder improves symptoms that continue into the postmenstrual phase. J Affect Disord 2005;85(3):317-321. [doi: 10.1016/j.jad.2004.10.006] [Medline: 15780701]
35. Pattani A. In some countries, women get days off for period pain. The New York Times. 2017. URL: https://www.nytimes.com/2017/07/24/health/period-pain-paid-time-off-policy.html [accessed 2023-11-24]
36. Ko CH, Long CY, Chen SY, Chen IJ, Huang TH, Yen JY. Depression, irritability, and anxiety in women with premenstrual dysphoric disorder. Int J Psychiatry Med 2013;46(1):39-55. [doi: 10.2190/PM.46.1.d] [Medline: 24547609]
37. Proctor M, Farquhar C. Diagnosis and management of dysmenorrhoea. BMJ 2006;332(7550):1134-1138 [FREE Full text] [doi: 10.1136/bmj.332.7550.1134] [Medline: 16690671]
38. Yoshino O, Takahashi N, Suzukamo Y. Menstrual symptoms, health-related quality of life, and work productivity in Japanese women with dysmenorrhea receiving different treatments: prospective observational study. Adv Ther 2022;39(6):2562-2577 [FREE Full text] [doi: 10.1007/s12325-022-02118-0] [Medline: 35362862]
39. Liao S. How to use the Clue app to help you become pregnant. Clue. 2022. URL: https://helloclue.com/articles/about-clue/how-to-use-the-clue-app-to-help-you-become-pregnant [accessed 2023-11-24]
40. Vigil P, Meléndez J, Soto H, Petkovic G, Bernal YA, Molina S. Chronic stress and ovulatory dysfunction: implications in times of COVID-19. Front Glob Womens Health 2022;3:866104 [FREE Full text] [doi: 10.3389/fgwh.2022.866104] [Medline: 35677754]
41. Cherenack EM, Salazar AS, Nogueira NF, Raccamarich P, Rodriguez VJ, Mantero AM, et al. Infection with SARS-CoV-2 is associated with menstrual irregularities among women of reproductive age. PLoS One 2022;17(10):e0276131 [FREE Full text] [doi: 10.1371/journal.pone.0276131] [Medline: 36288342]
42. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. BJOG 2018;125(13):1716 [FREE Full text] [doi: 10.1111/1471-0528.15199] [Medline: 29916205]
43. Davis N. Natural Cycles app: 'highly accurate contraceptive' claim misled consumers. The Guardian. 2018. URL: https://www.theguardian.com/technology/2018/aug/29/natural-cycles-app-highly-accurate-contraceptive-claim-misled-consumers [accessed 2023-11-24]
44. Baker FC, Siboza F, Fuller A. Temperature regulation in women: effects of the menstrual cycle. Temperature (Austin) 2020;7(3):226-262 [FREE Full text] [doi: 10.1080/23328940.2020.1735927] [Medline: 33123618]
45. Levy J, Romo-Avilés N. "A good little tool to get to know yourself a bit better": a qualitative study on users' experiences of app-supported menstrual tracking in Europe. BMC Public Health 2019;19(1):1213 [FREE Full text] [doi: 10.1186/s12889-019-7549-8] [Medline: 31481043]
46. Chandran R, Baptista D. After Roe v. Wade, healthcare data privacy fears grow worldwide. Reuters. 2022. URL: https://www.reuters.com/legal/litigation/after-roe-v-wade-healthcare-data-privacy-fears-grow-worldwide-2022-07-12/ [accessed 2023-11-24]

Abbreviations

HR: heart rate
HRV: heart rate variability
LH: luteinizing hormones

XSL•FO
RenderX

MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 17.12.22; peer-reviewed by V Jennings, C Babb, D Epstein; comments to author 14.04.23; revised version received 02.08.23; accepted 27.10.23; published 15.02.24.

Please cite as:
Lyzwinski L, Elgendi M, Menon C
Innovative Approaches to Menstruation and Fertility Tracking Using Wearable Reproductive Health Technology: Systematic Review
J Med Internet Res 2024;26:e45139
URL: https://www.jmir.org/2024/1/e45139
doi:10.2196/45139
PMID:38358798

©Lynnette Lyzwinski, Mohamed Elgendi, Carlo Menon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User-Centered Development of a Patient Decision Aid for Choice of Early Abortion Method: Multi-Cycle Mixed Methods Study

Kate J Wahl[1], MSc; Melissa Brooks[2], MD; Logan Trenaman[3], PhD; Kirsten Desjardins-Lorimer[4], MD; Carolyn M Bell[4], MD; Nazgul Chokmorova[4], MD; Romy Segall[2], BSc, MD; Janelle Syring[4], MD; Aleyah Williams[1], MPH; Linda C Li[5], PhD; Wendy V Norman[4,6*], MD, MHSc; Sarah Munro[1,3*], PhD

[1]Department of Obstetrics and Gynecology, University of British Columbia, Vancouver, BC, Canada

[2]Department of Obstetrics and Gynecology, Dalhousie University, Halifax, NS, Canada

[3]Department of Health Systems and Population Health, School of Public Health, University of Washington, Seattle, WA, United States

[4]Department of Family Practice, University of British Columbia, Vancouver, BC, Canada

[5]Department of Physical Therapy, University of British Columbia, Vancouver, BC, Canada

[6]Department of Public Health, Environments and Society, Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Kate J Wahl, MSc
Department of Obstetrics and Gynecology
University of British Columbia
4500 Oak Street
Vancouver, BC, V6H 3N1
Canada
Phone: 1 4165231923
Email: kate.wahl@cw.bc.ca

Abstract

Background: People seeking abortion in early pregnancy have the choice between medication and procedural options for care. The choice is preference-sensitive—there is no clinically superior option and the choice depends on what matters most to the individual patient. Patient decision aids (PtDAs) are shared decision-making tools that support people in making informed, values-aligned health care choices.

Objective: We aimed to develop and evaluate the usability of a web-based PtDA for the Canadian context, where abortion care is publicly funded and available without legal restriction.

Methods: We used a systematic, user-centered design approach guided by principles of integrated knowledge translation. We first developed a prototype using available evidence for abortion seekers' decisional needs and the risks, benefits, and consequences of each option. We then refined the prototype through think-aloud interviews with participants at risk of unintended pregnancy ("patient" participants). Interviews were audio-recorded and documented through field notes. Finally, we conducted a web-based survey of patients and health care professionals involved with abortion care, which included the System Usability Scale. We used content analysis to identify usability issues described in the field notes and open-ended survey questions, and descriptive statistics to summarize participant characteristics and close-ended survey responses.

Results: A total of 61 individuals participated in this study. Further, 11 patients participated in think-aloud interviews. Overall, the response to the PtDA was positive; however, the content analysis identified issues related to the design, language, and information about the process and experience of obtaining abortion care. In response, we adapted the PtDA into an interactive website and revised it to include consistent and plain language, additional information (eg, pain experience narratives), and links to additional resources on how to find an abortion health care professional. In total, 25 patients and 25 health care professionals completed the survey. The mean System Usability Scale score met the threshold for good usability among both patient and health care professional participants. Most participants felt that the PtDA was user-friendly (patients: n=25, 100%; health care professionals: n=22, 88%), was not missing information (patients: n=21, 84%; health care professionals: n=18, 72%), and that it was appropriate for patients to complete the PtDA before a consultation (patients: n=23, 92%; health care professionals: n=23, 92%). Open-ended responses focused on improving usability by reducing the length of the PtDA and making the website more mobile-friendly.

XSL•FO
RenderX

Conclusions: We systematically designed the PtDA to address an unmet need to support informed, values-aligned decision-making about the method of abortion. The design process responded to a need identified by potential users and addressed unique sensitivities related to reproductive health decision-making.

(J Med Internet Res 2024;26:e48793) doi:10.2196/48793

KEYWORDS

family planning; abortion; shared decision-making; patient decision aid; qualitative; evaluation; Canada; health equity

Introduction

In total, 1 in 3 pregnancy-capable people in Canada will have an abortion in their lifetimes, and most will seek care early in pregnancy [1]. Medication abortion (using the gold-standard mifepristone/misoprostol regimen) and procedural abortion are common, safe, and effective options for abortion care in the first trimester [2,3]. The choice between using medications and presenting to a facility for a procedure is a preference-sensitive decision; there is no clinically superior option and the choice depends on what matters most to the individual patient regarding the respective treatments and the features of those options [4-6].

The choice of method of abortion can involve a process of shared decision-making, in which the patient and health care professional share the best available evidence about options, and the patient is supported to consider those options and clarify an informed preference [7]. There are many types of interventions available to support shared decision-making, including interventions targeting health care professionals (eg, educational materials, meetings, outreach visits, audit and feedback, and reminders) and patients (eg, patient decision aids [PtDA], appointment preparation packages, empowerment sessions, printed materials, and shared decision-making education) [8]. Of these interventions, PtDAs are well-suited to address challenges to shared decision-making about the method of abortion, including limited patient knowledge, public misinformation about options, poor access to health care professionals with sufficient expertise, and apprehension about abortion counseling [9].

PtDAs are widely used interventions that support people in making informed, deliberate health care choices by explicitly describing the health problem and decision, providing information about each option, and clarifying patient values [10]. The results of the 2023 Cochrane systematic review of 209 randomized controlled trials indicate that, compared to usual care (eg, information pamphlets or webpages), the use of PtDAs results in increases in patient knowledge, expectations of benefits and harms, clarity about what matters most to them, and participation in making a decision [11]. Of the studies included in the systematic review, 1 tested the effect of a PtDA leaflet for method of abortion and found that patients eligible for both medication and procedural abortion who received the PtDA were more knowledgeable, and had lower risk perceptions and decisional conflict than those who were in the control group [12]. However, that PtDA was developed 20 years ago in the UK health system and was not publicly available. A recent environmental scan of PtDAs for a method of abortion found that other available options meet few of the criteria set by the International Patient Decision Aid Standards (IPDAS) collaboration and do not include language and content optimized for end users [9,13].

Consequently, no PtDAs for method of abortion were available in Canada at the time of this study. This was a critical gap for both patients and health care professionals as, in 2017, mifepristone/misoprostol medication abortion came to the market, offering a new method of choice for people seeking abortion in the first trimester [14]. Unlike most jurisdictions, in Canada medication abortion is typically prescribed in primary care and dispensed in community pharmacies. Offering a PtDA in preparation for a brief primary care consultation allows the person seeking abortion more time to digest new information, consider their preferences, be ready to discuss their options, and make a quality decision.

In this context, we identified a need for a high-quality and publicly available PtDA to support people in making an informed choice about the method of abortion that reflects what is most important to them. Concurrently, our team was working in collaboration with knowledge users (health care professionals, patients, and health system decision makers) who were part of a larger project to investigate the implementation of mifepristone in Canada [15,16]. We, therefore, aimed to develop and evaluate the usability of a web-based PtDA for the Canadian context, where abortion care is publicly funded and available without legal restriction.

Methods

Study Design

We performed a mixed methods user-centered development and evaluation study informed by principles of integrated knowledge translation. Integrated knowledge translation is an approach to collaborative research in which researchers and knowledge users work together to identify a problem, conduct research as equal partners to address that problem, and coproduce research products that aim to impact health service delivery [17]. We selected this approach to increase the likelihood that our end PtDAs would be relevant, useable, and used for patients and health care professionals in Canada [17]. The need for a PtDA was identified through engagement with health care professionals. In 2017, they highlighted the need for patients to be supported in choosing between procedural care—which historically represented more than 90% of abortions in Canada [18]—and the newly available medication option [19,20]. This need was reaffirmed in 2022 by the Canadian federal health agency, Health Canada, which circulated a request for proposals to generate “evidence-based, culturally-relevant information aimed at supporting people in their reproductive

XSL•FO
RenderX

decision-making and in accessing abortion services as needed" [21].

We operationalized integrated knowledge translation principles in a user-centered design process. User-centered design "grounds the characteristics of an innovation in information about the individuals who use that innovation, with a goal of maximizing 'usability in context'" [22]. In PtDA development, user-centered design involves iteratively understanding users, developing and refining a prototype, and observing user interaction with the prototype [23,24]. Like integrated knowledge translation, this approach is predicated on the assumption that involving users throughout the process increases the relevance of the PtDA and the likelihood of successful implementation [24].

Our design process included the following steps (Figure 1): identification of evidence about abortion patients' decisional needs and the attributes of medication and procedural abortion that matter most from a patient perspective; development of a paper-based prototype; usability testing via think-aloud interviews with potential end users; refinement of the PtDA prototype into an interactive website; usability testing via a survey with potential end users and abortion health care professionals; and final revisions before launching the PtDA for real-world testing. Our systematic process was informed by user-centered methods for PtDA development [23,24], guidance from the IPDAS collaboration [25-27], and the Standards for Universal Reporting of Patient Decision Aid Evaluation checklist [10].

Figure 1. PtDA development process. PtDA: patient decision aid.

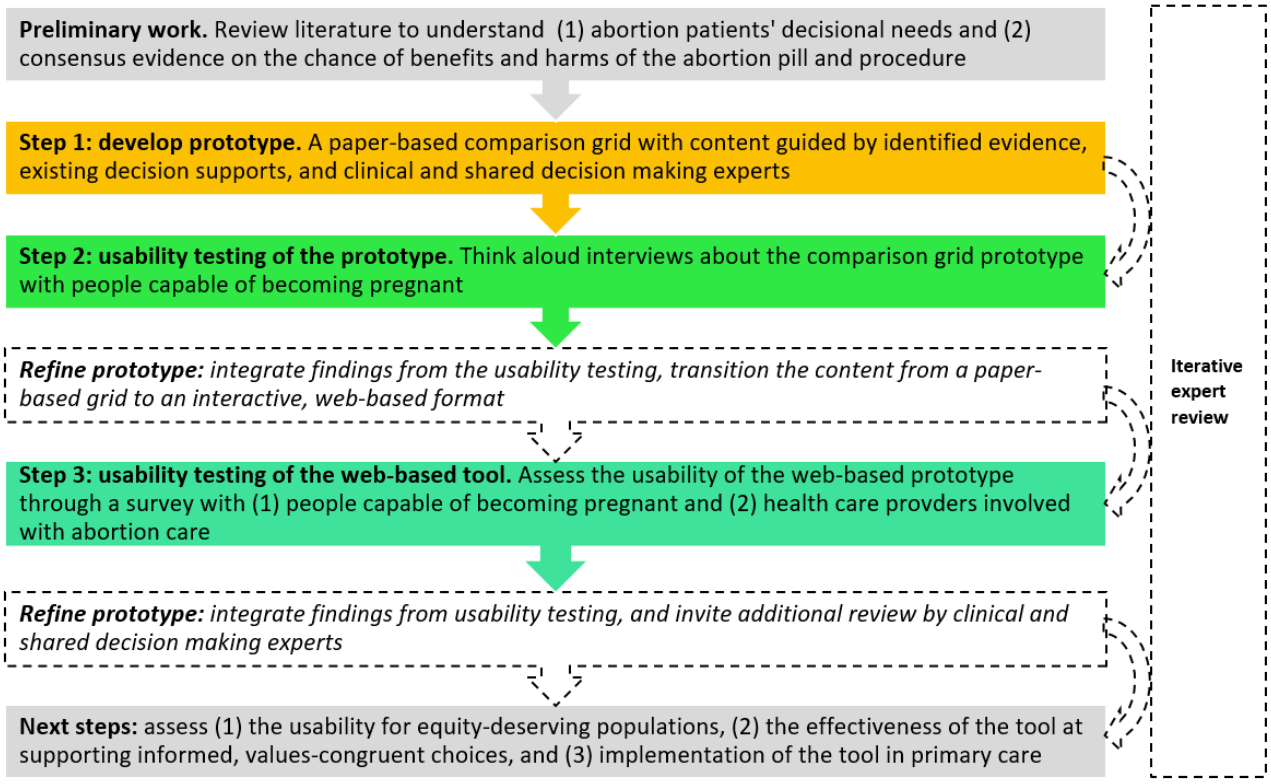

Our multidisciplinary team included experts in shared decision-making (SM and LT), a PhD student in patient-oriented knowledge translation (KJW), experts in integrated knowledge translation with health care professionals and policy makers (WVN and SM), clinical experts in abortion counseling and care (WVN and MB), a medical undergraduate student (RS), a research project coordinator (AW), and family medicine residents (KD-L, CMB, NC, and JS) who had an interest in abortion care. Additionally, a panel of experts external to the development process reviewed the PtDA for clinical accuracy following each revision of the prototype. These experts included coauthors of the national Society for Obstetricians and Gynaecologists of Canada (SOGC) clinical practice guidelines for abortion care in Canada. They were invited to this project because of their knowledge of first-trimester abortion care as well as their ability to support the implementation of the PtDA in guidelines and routine clinical practice.

Ethical Considerations

The research was approved by the University of British Columbia Children's and Women's Research Ethics Board (H16-01006) and the Nova Scotia Health Research Ethics Board (1027637). In each round of testing, participants received a CAD $20 (US $14.75) Amazon gift card by email for their participation.

Preliminary Work: Identification of Evidence

We identified the decisional needs of people seeking early abortion care using a 2018 systematic review of reasons for choosing an abortion method [28], an additional search that identified 1 study conducted in Canada following the 2017 availability of mifepristone/misoprostol medication abortion [29], and the SOGC clinical practice guidelines [2,3]. The review identified several key factors that matter most for patient choice of early abortion method: *perceived simplicity and "naturalness,"fear of complication or bleeding, fear of*

XSL•FO
RenderX

anesthesia or surgery, *timing of the procedure*, and *chance of sedation*. The additional Canadian study found that the *time required to complete the abortion* and *side effects* were important factors. According to the SOGC clinical practice guidelines, the key information that should be communicated to the patient are *gestational age limits* and *the risk of complications with increasing gestational age* [2,3]. The guidelines also indicate that *wait times*, *travel times*, and *cost considerations* may be important in a person's choice of abortion method and should be addressed [2,3].

We compiled a long list of attributes for our expert panel and then consolidated and refined the attribute list through each stage of the prototype evaluation. For evidence of how these factors differed for medication and procedural abortion, we drew primarily from the SOGC clinical practice guidelines for abortion [2,3]. For cost considerations, we described the range of federal, provincial, and population-specific programs that provide free coverage of abortion care for people in Canada.

Step 1: Developing the Prototype

Our goal was to produce an interactive, web-based PtDA that would be widely accessible to people seeking an abortion in Canada by leveraging the widespread use of digital health information, especially among reproductive-aged people [30]. Our first prototype was based on a previously identified paper-based question-and-answer comparison grid that presented evidence-based information about the medication and procedural options [9,31]. We calculated readability by inputting the plain text of the paper-based prototype into a Simple Measure of Gobbledygook (SMOG) Index calculator [32].

We made 2 intentional deviations from common practices in PtDA development [33]. First, we did not include an "opt-out" or "do nothing" option, which would describe the natural course of pregnancy. We chose to exclude this option to ensure clarity for users regarding the decision point; specifically, our decision point of interest was the method of abortion, not the choice to terminate or continue a pregnancy. Second, we characterized attributes of the options as key points rather than positive and negative features to avoid imposing value judgments onto subjective features (eg, having the abortion take place at home may be beneficial for some people but may be a deterrent for others).

Step 2: Usability Testing of the Prototype

We first conducted usability testing involving think-aloud interviews with patient participants to assess the paper-based prototype. Inclusion criteria included people aged 18-49 years assigned-female-at-birth who resided in Canada and could speak and read English. In January 2020, we recruited participants for the first round of think-aloud interviews [34] via email and poster advertising circulated to (1) a network of parent research advisors who were convened to guide a broader program of research about pregnancy and childbirth in British Columbia, Canada, and (2) a clinic providing surgical abortion care in Nova Scotia, Canada, as well as snowball sampling with participants. We purposively sought to advertise this study with these populations to ensure variation in age, ethnicity, level of education, parity, and abortion experience. Interested individuals reviewed this study information form and provided consent to participate, before scheduling an interview. The interviewer asked participants to think aloud as they navigated the prototype, for example describing what they liked or disliked, missing information, or lack of clarity. The interviewer noted the participant's feedback on a copy of the prototype during the interview. Finally, the participant responded to questions adapted from the System Usability Scale [35], a measure designed to collect subjective ratings of a product's usability, and completed a brief demographic questionnaire. The interviews were conducted via videoconferencing and were audio recorded. We deidentified the qualitative data and assigned each participant a unique identifier. Then, the interviewer listened to the recording and revised their field notes with additional information including relevant quotes.

For the analysis of think-aloud interviews, we used inductive content analysis to describe the usability and acceptability of different elements of the PtDA [36]. Further, 3 family medicine residents (KD-L, CMB, and NC) under guidance from a senior coauthor (SM) completed open coding to develop a list of initial categories, which we grouped under higher-order headings. We then organized these results in a table to illustrate usability issues (categories), illustrative participant quotes, and modifications to make. We then used the results of interviews to adapt the prototype into a web-based format, which we tested via further think-aloud interviews and a survey with people capable of becoming pregnant and health care professionals involved with abortion care.

Step 3: Usability Testing of the Website

For the web-based format, we used DecideApp PtDA open-source software, which provides a sustainable solution to the problems of low quality and high maintenance costs faced by web-based PtDAs by allowing developers to host, maintain, and update their tools at no cost. This software has been user-tested and can be accessed by phone, tablet, or computer [37,38]. It organizes a PtDA into 6 sections: Introduction, About Me, My Values, My Choice, Review, and Next Steps. In the My Values section, an interactive values clarification exercise allows users to rank and make trade-offs between attributes of the options. The final pages provide an opportunity for users to make a choice, complete a knowledge self-assessment, and consider the next steps to access their chosen method.

From July to August 2020, we recruited patient and health care professional participants using Twitter and the email list of the Canadian Abortion Providers Support platform, respectively. Participants received an email with a link to the PtDA and were redirected to the survey once they had navigated through the PtDA. As above, inclusion criteria included people aged 18-49 years assigned as female-at-birth who resided in Canada. Among health care professionals, we included eligible prescribers who may not have previously engaged in abortion care (family physicians, residents, nurse practitioners, and midwives), and allied health professionals and stakeholders who provide or support abortion care, who practiced in Canada. All participants had to speak and read English.

The survey included 3 sections: usability, implementation, and participant characteristics. The usability section consisted of

XSL•FO
RenderX

the System Usability Scale [35], and purpose-built questions about what participants liked and disliked about the PtDA. The implementation section included open- and close-ended questions about how the PtDA compares to other resources and when it could be implemented in the care pathway. Patient participants also completed the Control Preference Scale, a validated measure used to determine their preferred role in decision-making (active, collaborative, or passive) [39]. Data on participant characteristics included gender, abortion experience (patient participants), and abortion practice (health care professional participants). We deidentified the qualitative data and assigned each participant a unique identifier. For the analysis of survey data, we characterized close-ended responses using descriptive statistics, and, following the analysis procedures described in Step 2 in the Methods section, used inductive content analysis of open-ended responses to generate categories associated with usability and implementation [36]. In 2021, we made minor revisions to the website based on the results of usability testing and published the PtDA for use in routine clinical care.

Results

Overview

In the following sections, we outline the results of the development process including the results of the think-aloud interviews and survey, as well as the final decision aid prototype.

Step 1: Developing the Prototype

Our initial prototype, a paper-based question-and-answer comparison grid, presented evidence-based information comparing medication and procedural abortion. The first version of the prototype also included a second medication abortion regimen involving off-label use of methotrexate, however, we removed this option following a review by the clinical expert panel who advised us that there is very infrequent use of this regimen in Canada in comparison to the gold standard medication abortion option, mifepristone. Other changes at this stage involved clarifying the scope of practice (health care professionals other than gynecologists can perform a procedural abortion), abortion practice (gestational age limit and how the medication is taken), the abortion experience (what to expect in terms of bleeding), and risk (removing information about second- and third-trimester abortion). The updated prototype was finalized by a scientist (SM) and trainee (KJW) with expertise in PtDA development. The prototype (see Multimedia Appendix 1) was ultimately 4 pages long and described 18 attributes of each option framed as Frequently Asked Questions, including abortion eligibility (How far along in pregnancy can I be?), duration (How long does it take?), and side effects (How much will I bleed?). The SMOG grade level was 8.4.

Step 2: Usability Testing of the Prototype

Participant Characteristics

We included 11 participants in think-aloud interviews between January and July 2020, including 7 recruited through a parent research advisory network and 4 individuals who had recently attended an abortion clinic. The mean interview duration was 36 minutes (SD 6 minutes). The participants ranged in age from 31 to 37 years. All had been pregnant and 8 out of 11 (73%) participants had a personal experience of abortion (4 participants who had recently attended an abortion clinic and 4 participants from the parent research advisory who disclosed their experience during the interview). The characteristics of the sample are reported in Table 1.

Table 1. Characteristics of think-aloud interview participants (N=11).

Characteristic	Participants, n (%)
Gender	
Cisgender woman	11 (100)
Pregnant ever	11 (100)
Race	
Black	1 (9)
Chinese	1 (9)
Hispanic or Latin American	3 (27)
South Asian	1 (9)
White	5 (46)
Education	
College	2 (18)
University	9 (82)

Usability

Overall, participants had a positive view of the paper-based, comparison grid PtDA. In total, 1 participant who had recently sought an abortion said, “I think this is great and super helpful. It would’ve been awesome to have had access to this right away … I don’t think there’s really anything missing from here that I was Googling about” (DA010). The only participant who expressed antichoice views indicated that the PtDA would be helpful to someone seeking to terminate a pregnancy (DA001). Another participant said, “[The PtDA] is not biased, it’s not like you’re going to die. It’s a fact, you know the facts and then

XSL•FO
RenderX

you decide whether you want it or not. A lot of people feel it's so shameful and judgmental, but this is very straightforward. I like it." (DA002). Several participants stated they felt more informed and knowledgeable about the options.

In response to questions adapted from the System Usability Scale, all 11 participants agreed that the PtDA was easy to use, that most people could learn to use it quickly, and that they felt very confident using the prototype, and disagreed that it was awkward to use. In total, 8 (73%) participants agreed with the statement that the components of the PtDA were well-integrated. A majority of participants disagreed with the statements that the website was unnecessarily complex (n=8, 73%), that they would need the support of an expert to use it (n=8, 73%), that it was too inconsistent (n=9, 82%), and that they would need to learn a lot before using it (n=8, 73%). Further, 2 (18%) participants agreed with the statements that the PtDA was unnecessarily complex and that they would need to learn a lot before using it. Furthermore, 1 (9%) participant agreed with the statement that the PtDA was too inconsistent.

Through inductive analysis of think-aloud interviews, we identified 4 key usability categories: design, language, process, and experience.

Design

Participants liked the side-by-side comparison layout, appreciated the summary of key points to remember, and said that overall, the presented information was clear. For example, 1 participant reflected, "I think it's very clear ... it's very simplistic, people will understand the left-hand column is for medical abortion and the right-hand column is for surgical." (DA005) Some participants raised concerns about the aesthetics of the PtDA, difficulties recalling the headers across multiple pages, and the overall length of the PtDA.

Language

Participants sought to clarify language at several points in the PtDA. Common feedback was that the gestational age limit for the medication and the procedure should be clarified. Participants also pointed out inconsistent use of language (eg, doctor and health care professional) and medical jargon.

Process

Several participants were surprised to learn that family doctors could provide abortion care. Others noted that information about the duration—including travel time—and number of appointments for both medication and procedural abortion could be improved. In addition to clarifying the abortion process, several participants suggested including additional information and resources to help identify an abortion health care professional, understand when to seek help for abortion-related complications, and access emotional support. It was also important to participants that financial impacts (eg, hospital parking and menstrual pads) were included for each option.

Experience

Participants provided insight into the description of the physical, psychological, and other consequences associated with the abortion medication and procedure. Participants who had both types of abortion care felt that the description of pain that "may be worse than a period" was inaccurate. Other participants indicated that information about perceived and real risks was distressing or felt out of place, such as correcting myths about future fertility or breast cancer. Some participants indicated that patient stories would be valuable saying, for example, "I think what might be nice to help with the decision-making process is reading stories of people's experiences" (DA006).

Modifications Made

Changes made based on these findings are described in Table 2. Key user-centered modifications included transitioning to a web-based format with a consistent color scheme, clarifying who the PtDA is for (for typical pregnancies up to 10 weeks), adding information about telemedicine to reflect guidelines for the provision of abortion during pandemics, and developing brief first-person qualitative descriptions of the pain intensity for each option.

Through analysis of the interviews and consultation with our panel of clinical experts, we also identified that, among the 18 initial attributes in our prototype, 7 had the most relative importance to patients in choosing between medication and procedural abortion. These attributes also represented important differences between each option which forced participants to consider the trade-offs they were willing to make. Thus we moved all other potential attributes into an information section (My Options) that supported the user to gain knowledge before clarifying what mattered most to them by considering the differences between options (My Values).

XSL•FO
RenderX

Table 2. Usability issues identified by participants in think-aloud interviews and resulting changes.

Issue and illustrative participant quotes	Modifications made
Design	
• "I assume it will be formatted a bit more attractively?" (DA003)	• Transitioned to web-based format with a consistent color scheme
• "My question is about what the format will be…it is spread over multiple pages, people may need the [headers] again to understand." (DA005)	• Divided comparison grid into 3 sections (how it works, how it feels, and key points to remember)
• "The longer you make a document, the less people are going to read it anyway, they are just going to start to skim." (DA009)	• Reviewed for succinctness • Broke up the content into separate sections for qualitative pain descriptions
Language	
• "When I am looking at that point [about how far along in pregnancy I can be], I'm not sure if just doesn't click immediately…it is a little bit confusing." (DA011)	• Clarified that the information in the PtDA[a] is for typical pregnancies up to 10 weeks
• "Review it with the lens of only using different language where there is something different to be said." (DA003)	• Reviewed for consistent terminology
• "It seems like it's a bunch of medical jargon...I would want it to actually say…pill [rather than medication names]." (DA005)	• Changed terminology to "abortion pill" and "abortion procedure"
Process	
• "I found that confusing, because when I went to my family doctor, she told me to go to [the hospital] in order to make either decision…but my family doctor could have just prescribed me the pill?" (DA008)	• Added information: patients can ask their family doctor if they provide the abortion pill
• "I was there [at the hospital] at 7, I didn't know it was going to be a 4-hour procedure …it's a lot longer than 5-10 minutes…it doesn't really say you will be at the hospital for a few hours…you need half a day off if you're going the hospital route." (DA008)	• Clarified time off is needed on the day of the procedure • Added information about telemedicine (reflection of SOGC[b] guidelines for the provision of abortion during pandemics)
• "Why breast cancer? I would worry about cervical or uterine cancer, the actual place of the procedure." (DA004) • "I guess you have to put the associated risks…could that be provided once they made the decision?" (DA003)	• Removed statement saying there is no increased risk of breast cancer, which reflected a common misconception in the United States that abortion causes malignancy [40] • Rephrased some of the risk language (eg, removed the term "perforation")
• "Is this tool going to provide a list of clinics or the hospitals that do provide it?" (DA006)	• Added links to external resources about how to find an abortion health care professional
Experience	
• "Just a comment on the pain: it feels like you're giving birth, it feels like you're going through labour…I think this doesn't really give you the full picture. It's painful, it's extremely painful." (DA006) • "Gentle [suction] my ass." (DA009) • "I'd say worse than period-like cramps, but of course everybody is different." (DA011)	• Developed and added brief first-person qualitative descriptions (mean 47 words, SD 8.3 words) of how low-, medium-, and high-intensity pain could feel for each option (descriptions were synthesized from 58 published accounts of abortion care, and designed them to be similar in length and structure to minimize bias toward either option—see Multimedia Appendix 2)
• "I can imagine how traumatic it would be…what kind of support would be available?" (DA001)	• Clarified that abortion is not associated with an increased risk of mental illness • Added a list of resources for users to speak with someone about their mental health and abortion care

[a]PtDA: patient decision aid.

[b]SOGC: Society of Obstetricians and Gynaecologists of Canada.

XSL•FO
RenderX

Step 3: Usability Testing of the Website

Description of the PtDA

As shown in Figure 2, the revised version of the PtDA resulting from our systematic process is an interactive website. Initially, the title was *My Body, My Choice*; however, this was changed to avoid association with antivaccine campaigns that co-opted this reproductive rights slogan. The new title, *It's My Choice* or *C'est Mon Choix*, was selected for its easy use in English and French. The PtDA leads the user through 6 sections:

- The *Introduction* section provides the user with information about the decision and the PtDA, as well as grids comparing positive and negative features of the abortion pill and procedure, including their chance of benefits (eg, effectiveness), harms (eg, complications), and other relevant factors (eg, number of appointments and cost).
- The *About Me* section asks the user to identify any contraindications to the methods. It then prompts users to consider their privacy needs and gives examples of how this relates to each option (eg, the abortion pill can be explained to others as a miscarriage; procedural care can be completed quickly).
- The *My Values* section includes a values clarification exercise, in which the user selects and weights (on a 0-100 scale) the relative importance of at least three of 7 decisional attributes: avoiding pain, avoiding bleeding, having the abortion at home, having an experience that feels like a miscarriage, having fewer appointments, less time off for recovery, and having a companion during the abortion.
- The *My Choice* section highlights 1 option, based on the attribute weights the user assigned in the My Values section. For instance, if a user strongly preferred to avoid bleeding and have fewer appointments, the software would suggest that a procedural abortion would be a better match. For a user who preferred having the abortion at home and having a companion present, the software would suggest that a medication abortion would be a better match. The user selects the option they prefer.
- The *Review* section asks the user to complete the 4-item SURE (Sure of Myself, Understand Information, Risk-Benefit Ratio, Encouragement) screening test [41], and advises them to talk with an expert if they answer "no" to any of the questions. This section also includes information phone lines to ensure that users can seek confidential, accurate, and nonjudgmental support.
- Lastly, in the *Next Steps* section, users see a summary of their choice and the features that matter most to them, instructions for how to save the results, keep the results private, and find an abortion health care professional. Each section of the PtDA includes a "Leave" button in case users need to navigate away from the website quickly.

We calculated readability by inputting the plain text of the web-based PtDA into a SMOG Index calculator [32], which assessed the reading level of the web-based PtDA as grade 9.2.

To ensure users' trust in the information as accurate and unbiased we provided a data declaration on the landing page: "the clinical information presented in this decision aid comes from Society of Obstetricians and Gynaecologists best practice guidelines." On the landing page, we also specify "This website was developed by researchers at the University of British Columbia and Dalhousie University. This tool is not supported or connected to any pharmaceutical company."

XSL•FO
RenderX

Figure 2. *It's My Choice* patient decision aid web app.

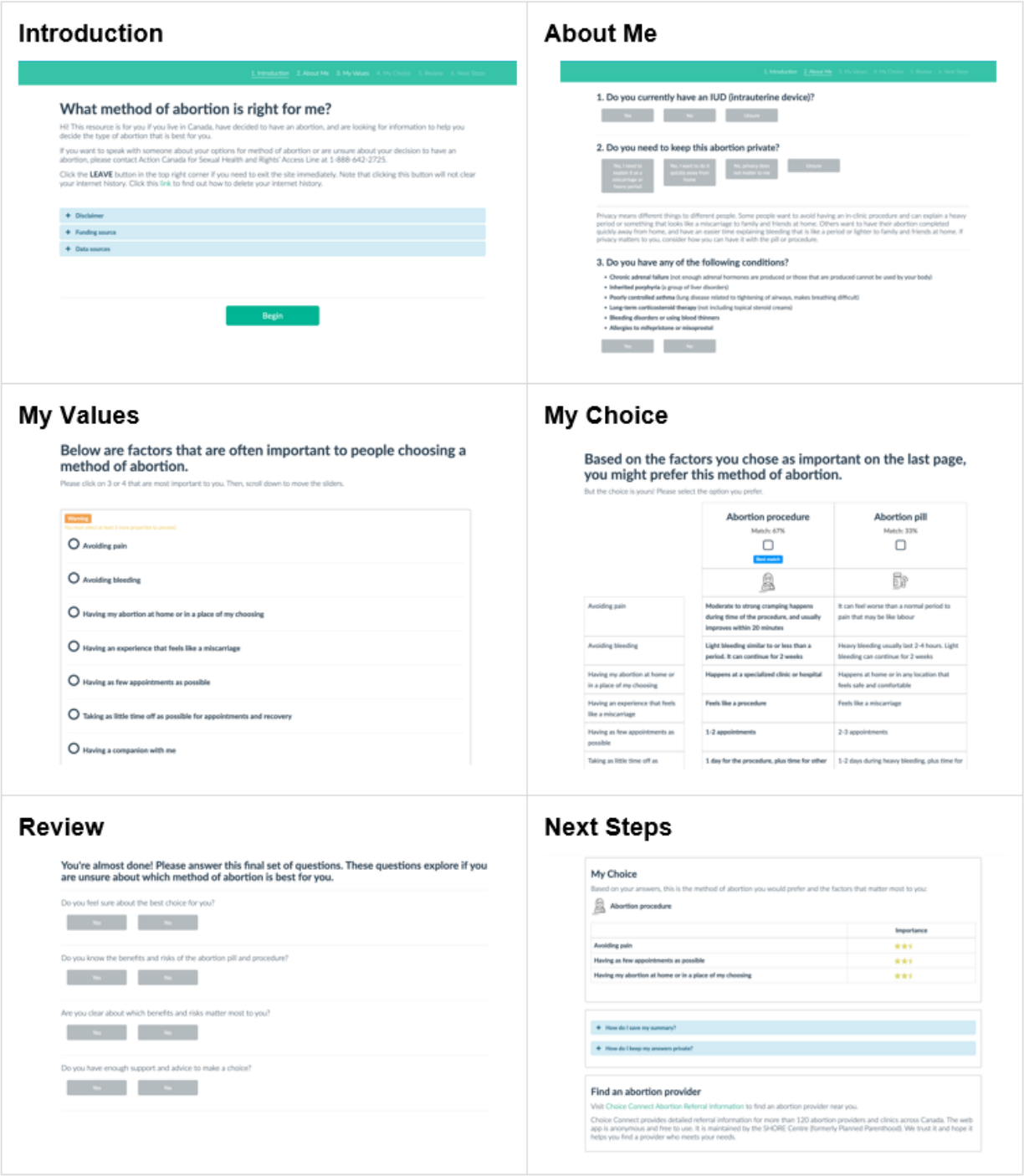

Participant Characteristics

A total of 50 participants, including 25 patients and 25 health care professionals, reviewed the PtDA website and completed the survey between January and March 2021. The majority of patient (n=23, 92%) and health care professional (n=23, 92%) participants identified as cisgender women. Among patient participants, 16% (n=4) reported one or more previous abortions in various clinical settings. More than half (n=16, 64%) of health care professionals offered care in private medical offices, with other locations including sexual health clinics, community health centers, and youth clinics. Many health care professionals were family physicians (n=11, 44%), and other common types were nurse practitioners (n=7, 28%) and midwives (n=3, 12%). The mean proportion of the clinical practice of each health care professional devoted to abortion care was 18% (SD 13%). Most health care professional respondents (n=18, 72%) were involved with the provision of medication, but not procedural, abortion care. The characteristics of patient and health care professional participants are reported in Table 3.

XSL•FO
RenderX

Table 3. Characteristic of patient (N=25) and health care professional (N=25) survey participants.

Characteristic	Patient	Health care professionals
Gender, n (%)		
Cisgender woman	23 (92)	23 (92)
Cisgender man	0 (0)	1 (4)
Nonbinary	2 (8)	1 (4)
Location of abortion practice[a], n (%)		
Private medical office	1 (4)	16 (64)
Sexual health clinic	1 (4)	4 (16)
Hospital	3 (12)	4 (16)
Other	1 (4)	6 (24)
Prefer not to say	4 (16)	0 (0)
Health care professional type, n (%)		
Family medicine or general practice	—[b]	11 (44)
Nurse practitioner	—	7 (28)
Midwife	—	3 (12)
Counsellor or social worker	—	2 (8)
Gynecologist	—	1 (4)
Other	—	1 (4)
Type of abortion provided, n (%)		
Medication-only	—	18 (72)
Procedural-only	—	1 (4)
Both	—	3 (12)
No response	—	3 (12)
Percentage of work focused on abortion, mean (SD)	—	18 (13)

[a]In total, 4 participants reported a history of abortion care, representing 6 abortion procedures.

[b]Not available.

Usability

The mean System Usability Score met the threshold for good usability among both patient (mean 85.7, SD 8.6) and health care professional (mean 80, SD 12) participants, although some health care professionals agreed with the statement, “I found the website to be unnecessarily complex,” (see Multimedia Appendix 3 for the full distribution of responses from patient and health care professionals). All 25 patients and 22 out of 25 (88%) health care professional respondents indicated that the user-friendliness of the PtDA was good or the best imaginable. When asked what they liked most about the PtDA, both participant groups described the ease of use, comparison of options, and the explicit values clarification exercise. When asked what they liked least about the PtDA, several health care professionals and some patients pointed out that it was difficult to use on a cell phone. A summary of usability results is presented in Table 4.

In total, 21 (84%) patients and 18 (72%) health care professionals felt that the PtDA was not missing any information needed to decide about the method of abortion in early pregnancy. While acknowledging that it is “hard to balance being easy to read/understand while including enough accurate clinical information,” several health care professionals and some patients indicated that the PtDA was too long and repetitive. Among the 4 (16%) patient participants who felt information was missing, the most common suggestion was a tool for locating an abortion health care professional. The 7 (28%) health care professionals who felt information was missing primarily made suggestions about the medical information included in the PtDA (eg, listing midwives as health care professionals with abortion care in scope of practice and the appropriateness of gender-inclusive terminology) and the accessibility of information for various language and cultural groups.

XSL•FO
RenderX

Table 4. Survey responses from patient (N=25) and health care professional (N=25) participants.

Survey questions (health care professional question stems indicated in brackets)	Patient	Health care professional
System Usability Scale, mean (SD)	85.7 (8.6)	80 (12)
I would rate the overall user-friendliness of this website as, n (%):		
Best imaginable	8 (32)	4 (16)
Good	17 (68)	18 (72)
Okay	0 (0)	1 (4)
Neutral	0 (0)	0 (0)
Poor	0 (0)	1 (4)
Awful	0 (0)	0 (0)
Worst imaginable	0 (0)	0 (0)
Missing	0 (0)	1 (4)
Is there any information missing from the website that you feel would enhance your ability to choose (counsel your patients deciding) a method of abortion? n (%)		
Yes	4 (16)	7 (28)
No	21 (84)	18 (72)
Would this website replace or add to any existing resources that you *have used* for deciding (used with patients considering) method of abortion? n (%)		
Yes	10 (40)	18 (72)
No	15 (60)	7 (28)
When would you use the website? n (%)		
Before my consultation with an abortion provider	23 (92)	23 (92)
During my consultation with an abortion provider	8 (32)	10 (40)
After my consultation with an abortion provider	12 (48)	13 (52)
Other	0 (0)	0 (0)
I would prefer not to use this website	0 (0)	1 (4)
Control preferences scale, n (%)		
I prefer to make the final decision about what method of abortion I will receive	6 (24)	—[a]
I prefer to make the final selection of method of abortion after seriously considering my health care provider's opinion	7 (28)	—
I prefer that my health care provider share responsibility for deciding which method of abortion is best for me	4 (16)	—
I prefer that my health care provider makes the final decision about which method of abortion will be used, but seriously considers my opinion	2 (8)	—
I prefer to leave all decisions regarding my method of abortion to my health care provider	6 (24)	—

[a]Not available.

Implementation

Participants viewed the PtDA as a positive addition to current resources. Patients with a history of abortion care described looking for the information on the internet and speaking with friends, family members, and health care professionals. Compared with these sources of information, many patients liked the credibility and anonymity of the PtDA, whereas some disliked that it was less personal than a conversation. Further, 18 (72%) health care professional participants said that the PtDA would add to or replace the resources they currently use in practice. Compared with these other resources, health care professionals liked that the PtDA could be explored by patients independently and that it would support them in thinking about the option that was best for them. The disadvantages of the PtDA compared with existing resources were the length—which health care professionals felt would make it difficult to use in a clinical interaction—and the lack of localized information. In total, 23 each (92%) of patient and health care professional participants felt that they would use the PtDA before a consultation.

XSL•FO
RenderX

Discussion

Principal Results

We designed a web-based, interactive PtDA for the choice of method of abortion in early pregnancy [42], taking a user-centered approach that involved usability testing with 36 patients and 25 health care professionals. Both patient and health care professional participants indicated that the PtDA had good usability and would be a valuable resource for decision-making. This PtDA fills a critical need to support the autonomy of patients and shared decision-making with their health care professional related to the preference-sensitive choice of method of abortion.

Comparison With Prior Work

A 2017 systematic review and environmental scan found that existing PtDAs for the method of abortion are of suboptimal quality [9]. Of the 50 PtDAs identified, all but one were created without expertise in decision aid design (eg, abortion services, reproductive health organizations, and consumer health information organizations); however, the development process for this UK-based pamphlet-style PtDA was not reported. The remaining PtDAs were noninteractive websites, smartphone apps, and PDFs that were not tested with users. The authors found that the information about methods of abortion was presented in a disorganized, inconsistent, and unequal way. Subsequent work has found that existing PtDAs emphasize medical (versus social, emotional, and practical) attributes, do not include values clarification, and can be biased to persuade users of a certain method [13].

To address some of the challenges identified in the literature, we systematically structured and designed elements of the PtDA following newly proposed IPDAS criteria (eg, showing positive and negative features with equal detail) [33]. We included an explicit values-clarification exercise, which a recent meta-analysis found to decrease decisional conflict and values-incongruent choices [43].

We based the decision aid on comprehensive and up-to-date scientific evidence related to the effectiveness and safety of medication abortion and procedural abortion; however, less evidence was available for nonmedical attributes. For example, many existing PtDAs incorrectly frame privacy as a "factual advantage" of medication abortion [13]. To address this, we included privacy in the About Me section as something that means "different things to different people." Similarly, evidence suggests that patients who do not feel appropriately informed about the pain associated with their method of abortion are less satisfied with their choice [44,45]; and the degree of pain experienced varies across options and among individuals. Following the suggestion of patient participants to include stories and recognizing that evidence for the inclusion of narratives in PtDAs is emerging [46], we elected to develop brief first-person qualitative descriptions of the pain experience. The inclusion of narratives in PtDAs may be effective in supporting patients to avoid surprise and regret, to minimize affective forecasting errors, and to "visualize" their health condition or treatment experience [46]. Guided by the narrative immersion model, our goal was to provide a "real-world preview" of the pain experience [47].

In addition to integrating user perspectives on the optimal tone, content, and format of the PtDA, user testing provided evidence to inform the future implementation of the PtDA. A clear barrier to the completion of the PtDA during the clinical encounter from the health care professional perspective was its length, supporting the finding of a recent rapid realist review, which theorized that health care professionals are less likely to use long or otherwise complex PtDAs that are difficult to integrate into routine practice [48]. However, 46 out of 50 (92%) participants endorsed the use of the PtDA by the patient alone before the initial consultation, which was aligned with the patient participant's preference to take an active role in making the final decision about their method of abortion as well as the best practice of early, pre-encounter distribution of PtDAs [48].

A unique feature of this PtDA was that it resulted from a broader program of integrated knowledge translation designed to support access to medication abortion once mifepristone became available in Canada in 2017. Guided by the principle that including knowledge users in research yields results that are more relevant and useful [49], we developed the PtDA in response to a knowledge user need, involved health care professional users as partners in our research process, including as coauthors, and integrated feedback from the expert panel. This parallels a theory of PtDA implementation that proposes that early involvement of health care professionals in PtDA development "creates a sense of ownership, increases buy-in, helps to legitimize content, and ensures the PtDA (content and delivery) is consistent with current practice" thereby increasing the likelihood of PtDA integration into routine clinical settings [48].

Viewed through an integrated knowledge translation lens, our findings point toward future areas of work to support access to abortion in Canada. Several patient participants indicated a need for tools to identify health care professionals who offer abortion care. Some shared that their primary health care professionals did not offer medication abortion despite it being within their scope of practice, and instead referred them to an abortion clinic for methods of counseling and care. We addressed this challenge in the PtDA by including links to available resources, such as confidential phone lines that link patients to health care professionals in their region. On the website we also indicated that patient users could ask their primary care providers whether they provide abortion care; however, we acknowledge that this may place the patient in a vulnerable position if their health care professional is uncomfortable with, or unable to, provide this service for any reason. Future work should investigate opportunities to shorten the pathway to this time-sensitive care, including how to support patients who use the decision aid to act on their informed preference for the method of abortion. This work may involve developing a tool for patients to talk to their primary care provider about prescribing medication abortion.

Strengths and Limitations

Several factors affect the interpretation of our work. Although potential patient users participated in the iterative development

XSL•FO
RenderX

process, the patient perspective was not represented in a formal advisory panel in the same way that the health care professional experts were. Participant characteristics collected for the think-aloud interviews demonstrated that our patient sample did not include people with lower education attainment, for whom the grade level and length of the PtDA could present a barrier [50]. Any transfer of the PtDA to jurisdictions outside Canada must consider how legal, regulatory, and other contextual factors affect the choice of the method of abortion. Since this study was completed, we have explored additional strategies to address these concerns, including additional user testing with people from equity-deserving groups, drop-down menus to adjust the level of detail, further plain language editing, and videos illustrating core content. Since the focus of this study was usability, we did not assess PtDA effectiveness, including impact on knowledge, decisional conflict, choice predisposition and decision, or concordance; however, a randomized controlled trial currently underway will measure the impact of the PtDA on these outcomes in a clinical setting. Finally, our integrated knowledge translation approach added to the robustness of our study by ensuring that health care professionals and patients were equal partners in the research process. One impact of this partnered approach is that our team has received funding support from Health Canada to implement the website on a national scale for people across Canada considering their abortion options [51].

Conclusions

The PtDA provides people choosing a method of early abortion and their health care professionals with a resource to understand methods of abortion available in the Canadian context and support to make a values-aligned choice. We designed the PtDA using a systematic approach that included both patient and health care professional participants to help ensure its relevance and usability. Our future work will seek to evaluate the implementation of the PtDA in clinical settings, create alternate formats to enhance accessibility, and develop a sustainable update policy. We will also continue to advance access to abortion care in Canada with our broader integrated knowledge translation program of research.

Acknowledgments

The authors thank the participants for contributing their time and expertise to the design of this tool. Family medicine residents CMB, NC, KD-L, and JS were supported by Sue Harris grants, Department of Family Practice, University of British Columbia. KJW was supported by the Vanier Scholar Award (2020-23). SM was supported by a Michael Smith Health Research BC Scholar Award (18270). WVN was supported by a Canadian Institutes of Health Research and Public Health Agency of Canada Chair in Applied Public Health Research (2014-2024, CPP-329455-107837). All grants underwent external peer review for scientific quality. The funders played no role in the design of this study, data collection, analysis, interpretation, or preparation of this paper.

Data Availability

Our ethics approval has specified the primary data is not available.

Authors' Contributions

KJW, SM, and MB conceived of and designed this study. CMB, NC, and KD-L led interview data collection, analysis, and interpretation with input from SM. RS and JS led survey data collection, analysis, and interpretation with input from SM and MB. AW, LCL, and WVN contributed to the synthesis and interpretation of results. KJW, SM, and LT wrote the first draft of this paper, and all authors contributed to this paper's revisions and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient decision aid prototype.
[PDF File (Adobe PDF File), 861 KB - jmir_v26i1e48793_app1.pdf]

Multimedia Appendix 2
Raw data for pain narratives.
[DOCX File , 38 KB - jmir_v26i1e48793_app2.docx]

Multimedia Appendix 3
Full distribution of System Usability Scale scores for patients and providers.
[DOCX File , 17 KB - jmir_v26i1e48793_app3.docx]

References

XSL•FO
RenderX

1. Norman WV. Induced abortion in Canada 1974-2005: trends over the first generation with legal access. Contraception 2012;85(2):185-191. [doi: 10.1016/j.contraception.2011.06.009] [Medline: 22036474]
2. Costescu D, Guilbert E, Bernardin J, Black A, Dunn S, Fitzsimmons B, et al. Medical abortion. J Obstet Gynaecol Can 2016;38(4):366-389. [doi: 10.1016/j.jogc.2016.01.002] [Medline: 27208607]
3. Costescu D, Guilbert É. No. 360-induced abortion: surgical abortion and second trimester medical methods. J Obstet Gynaecol Can 2018;40(6):750-783. [doi: 10.1016/j.jogc.2017.12.010] [Medline: 29861084]
4. Wennberg JE. Unwarranted variations in healthcare delivery: implications for academic medical centres. BMJ 2002;325(7370):961-964 [FREE Full text] [doi: 10.1136/bmj.325.7370.961] [Medline: 12399352]
5. Elwyn G, Frosch D, Rollnick S. Dual equipoise shared decision making: definitions for decision and behaviour support interventions. Implement Sci 2009;4:75 [FREE Full text] [doi: 10.1186/1748-5908-4-75] [Medline: 19922647]
6. Sepucha KR, Mulley AG. A practical approach to measuring the quality of preference-sensitive decisions. In: Edwards A, Elwyn G, editors. Shared Decision-Making in Health Care: Achieving Evidence-based Patient Choice. Oxford: Oxford University Press; 2009:151-156.
7. Elwyn G, Laitner S, Coulter A, Walker E, Watson P, Thomson R. Implementing shared decision making in the NHS. BMJ 2010;341:c5146 [FREE Full text] [doi: 10.1136/bmj.c5146] [Medline: 20947577]
8. Légaré F, Adekpedjou R, Stacey D, Turcotte S, Kryworuchko J, Graham ID, et al. Interventions for increasing the use of shared decision making by healthcare professionals. Cochrane Database Syst Rev 2018;7(7):CD006732 [FREE Full text] [doi: 10.1002/14651858.CD006732.pub4] [Medline: 30025154]
9. Donnelly KZ, Elwyn G, Thompson R. Quantity over quality-findings from a systematic review and environmental scan of patient decision aids on early abortion methods. Health Expect 2018;21(1):316-326 [FREE Full text] [doi: 10.1111/hex.12617] [Medline: 28881071]
10. Sepucha KR, Abhyankar P, Hoffman AS, Bekker HL, LeBlanc A, Levin CA, et al. Standards for Universal Reporting of Patient Decision Aid Evaluation studies: the development of SUNDAE checklist. BMJ Qual Saf 2018;27(5):380-388 [FREE Full text] [doi: 10.1136/bmjqs-2017-006986] [Medline: 29269567]
11. Stacey D, Lewis KB, Smith M, Carley M, Volk R, Douglas EE, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2024;1(1):CD001431. [doi: 10.1002/14651858.CD001431.pub6] [Medline: 38284415]
12. Wong SSM, Thornton JG, Gbolade B, Bekker HL. A randomised controlled trial of a decision-aid leaflet to facilitate women's choice between pregnancy termination methods. BJOG 2006;113(6):688-694. [doi: 10.1111/j.1471-0528.2006.00930.x] [Medline: 16709212]
13. Donnelly KZ, Elwyn G, Theiler R, Thompson R. Promoting or undermining quality decision making? A qualitative content analysis of patient decision aids comparing surgical and medication abortion. Womens Health Issues 2019;29(5):414-423. [doi: 10.1016/j.whi.2019.05.007] [Medline: 31266679]
14. Grant K. Long-awaited abortion pill Mifegymiso makes Canadian debut. The Globe and Mail. 2017. URL: https://www.theglobeandmail.com/news/national/long-awaited-abortion-pill-mifegymiso-rolls-out-in-canada/article33695167/?ref=http://www.theglobeandmail.com& [accessed 2023-04-03]
15. Norman WV, Munro S, Brooks M, Devane C, Guilbert E, Renner R, et al. Could implementation of mifepristone address Canada's urban-rural abortion access disparity: a mixed-methods implementation study protocol. BMJ Open 2019;9(4):e028443 [FREE Full text] [doi: 10.1136/bmjopen-2018-028443] [Medline: 31005943]
16. Munro SB, Dunn S, Guilbert ER, Norman WV. Advancing reproductive health through policy-engaged research in abortion care. Semin Reprod Med 2022;40(5-06):268-276. [doi: 10.1055/s-0042-1760213] [Medline: 36746159]
17. Dunn SI, Bhati DK, Reszel J, Kothari A, McCutcheon C, Graham ID. Understanding how and under what circumstances integrated knowledge translation works for people engaged in collaborative research: metasynthesis of IKTRN casebooks. JBI Evid Implement 2023;21(3):277-293 [FREE Full text] [doi: 10.1097/XEB.0000000000000367] [Medline: 36988573]
18. Norman WV, Guilbert ER, Okpaleke C, Hayden AS, Lichtenberg ES, Paul M, et al. Abortion health services in Canada: results of a 2012 national survey. Can Fam Physician 2016;62(4):e209-e217 [FREE Full text] [Medline: 28192276]
19. Munro S, Guilbert E, Wagner MS, Wilcox ES, Devane C, Dunn S, et al. Perspectives among canadian physicians on factors influencing implementation of mifepristone medical abortion: a national qualitative study. Ann Fam Med 2020;18(5):413-421 [FREE Full text] [doi: 10.1370/afm.2562] [Medline: 32928757]
20. Munro S, Wahl K, Soon JA, Guilbert E, Wilcox ES, Leduc-Robert G, et al. Pharmacist dispensing of the abortion pill in Canada: diffusion of innovation meets integrated knowledge translation. Implement Sci 2021;16(1):76 [FREE Full text] [doi: 10.1186/s13012-021-01144-w] [Medline: 34344393]
21. Health care policy and strategies program. Call for proposals: funding opportunities for sexual and reproductive health. Health Canada. URL: https://www.canada.ca/en/health-canada/programs/health-care-policy-strategies-program.html [accessed 2024-03-14]
22. Dopp AR, Parisi KE, Munson SA, Lyon AR. A glossary of user-centered design strategies for implementation experts. Transl Behav Med 2019;9(6):1057-1064 [FREE Full text] [doi: 10.1093/tbm/iby119] [Medline: 30535343]

XSL•FO
RenderX

23. Vaisson G, Provencher T, Dugas M, Trottier ME, Dansokho SC, Colquhoun H, et al. User involvement in the design and development of patient decision aids and other personal health tools: a systematic review. Med Decis Making 2021;41(3):261-274 [FREE Full text] [doi: 10.1177/0272989X20984134] [Medline: 33655791]
24. Witteman HO, Maki KG, Vaisson G, Finderup J, Lewis KB, Steffensen KD, et al. Systematic development of patient decision aids: an update from the IPDAS collaboration. Med Decis Making 2021;41(7):736-754 [FREE Full text] [doi: 10.1177/0272989X211014163] [Medline: 34148384]
25. Joseph-Williams N, Newcombe R, Politi M, Durand MA, Sivell S, Stacey D, et al. Toward minimum standards for certifying patient decision aids: a modified Delphi consensus process. Med Decis Making 2014;34(6):699-710. [doi: 10.1177/0272989X13501721] [Medline: 23963501]
26. Stacey D, Volk RJ, IPDAS Evidence Update Leads (Hilary Bekker, Karina Dahl Steffensen, Tammy C. Hoffmann, Kirsten McCaffery, Rachel Thompson, Richard Thomson, Lyndal Trevena, Trudy van der Weijden, and Holly Witteman). The International Patient Decision Aid Standards (IPDAS) collaboration: evidence update 2.0. Med Decis Making 2021;41(7):729-733 [FREE Full text] [doi: 10.1177/0272989X211035681] [Medline: 34416841]
27. Hoffman AS, Volk RJ, Saarimaki A, Stirling C, Li LC, Härter M, et al. Delivering patient decision aids on the internet: definitions, theories, current evidence, and emerging research areas. BMC Med Inform Decis Mak 2013;13(Suppl 2):S13 [FREE Full text] [doi: 10.1186/1472-6947-13-S2-S13] [Medline: 24625064]
28. Kanstrup C, Mäkelä M, Graungaard AH. Women's reasons for choosing abortion method: a systematic literature review. Scand J Public Health 2018;46(8):835-845. [doi: 10.1177/1403494817717555] [Medline: 28691571]
29. Murray ME, Casson M, Pudwell J, Waddington A. Patients' motivation for surgical versus medical abortion. J Obstet Gynaecol Can 2019;41(9):1325-1329. [doi: 10.1016/j.jogc.2019.01.009] [Medline: 30878325]
30. Kummervold PE, Chronaki CE, Lausen B, Prokosch HU, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res 2008;10(4):e42 [FREE Full text] [doi: 10.2196/jmir.1023] [Medline: 19017584]
31. Early abortion options. Reproductive Health Access Project. 2022. URL: https://www.reproductiveaccess.org/resource/early-abortion-options/ [accessed 2019-03-24]
32. Readability Formulas: free readability assessment tools to help you write for your readers. URL: https://readabilityformulas.com/ [accessed 2022-12-15]
33. Martin RW, Andersen SB, O'Brien MA, Bravo P, Hoffmann T, Olling K, et al. Providing balanced information about options in patient decision aids: an update from the International Patient Decision Aid Standards. Med Decis Making 2021;41(7):780-800 [FREE Full text] [doi: 10.1177/0272989X211021397] [Medline: 34196241]
34. Lundgrén-Laine H, Salanterä S. Think-aloud technique and protocol analysis in clinical decision-making research. Qual Health Res 2010;20(4):565-575. [doi: 10.1177/1049732309354278] [Medline: 19959822]
35. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
36. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
37. Bansback N, Li LC, Lynd L, Bryan S. Development and preliminary user testing of the DCIDA (Dynamic Computer Interactive Decision Application) for 'nudging' patients towards high quality decisions. BMC Med Inform Decis Mak 2014;14:62 [FREE Full text] [doi: 10.1186/1472-6947-14-62] [Medline: 25084808]
38. Trenaman L, Munro S, Almeida F, Ayas N, Hicklin J, Bansback N. Development of a patient decision aid prototype for adults with obstructive sleep apnea. Sleep Breath 2016;20(2):653-661. [doi: 10.1007/s11325-015-1269-9] [Medline: 26449553]
39. Degner LF, Sloan JA, Venkatesh P. The control preferences scale. Can J Nurs Res 1997;29(3):21-43. [Medline: 9505581]
40. Patev AJ, Hood KB. Towards a better understanding of abortion misinformation in the USA: a review of the literature. Cult Health Sex 2021;23(3):285-300. [doi: 10.1080/13691058.2019.1706001] [Medline: 32202213]
41. Légaré F, Kearing S, Clay K, Gagnon S, D'Amours D, Rousseau M, et al. Are you SURE?: Assessing patient decisional conflict with a 4-item screening test. Can Fam Physician 2010;56(8):e308-e314 [FREE Full text] [Medline: 20705870]
42. The Society of Obstetricians and Gynaecologists of Canada. URL: https://www.sexandu.ca/its-my-choice/ [accessed 2024-03-30]
43. Witteman HO, Ndjaboue R, Vaisson G, Dansokho SC, Arnold B, Bridges JFP, et al. Clarifying values: an updated and expanded systematic review and meta-analysis. Med Decis Making 2021;41(7):801-820 [FREE Full text] [doi: 10.1177/0272989X211037946] [Medline: 34565196]
44. Cavet S, Fiala C, Scemama A, Partouche H. Assessment of pain during medical abortion with home use of misoprostol. Eur J Contracept Reprod Health Care 2017;22(3):207-211. [doi: 10.1080/13625187.2017.1315092] [Medline: 28447484]
45. Baraitser P, Free C, Norman WV, Lewandowska M, Meiksin R, Palmer MJ, et al. Improving experience of medical abortion at home in a changing therapeutic, technological and regulatory landscape: a realist review. BMJ Open 2022;12(11):e066650 [FREE Full text] [doi: 10.1136/bmjopen-2022-066650] [Medline: 36385017]

XSL•FO
RenderX

46. Shaffer VA, Brodney S, Gavaruzzi T, Zisman-Ilani Y, Munro S, Smith SK, et al. Do personal stories make patient decision aids more effective? An update from the International Patient Decision Aids Standards. Med Decis Making 2021;41(7):897-906 [FREE Full text] [doi: 10.1177/0272989X211011100] [Medline: 34027739]
47. Shaffer VA, Focella ES, Hathaway A, Scherer LD, Zikmund-Fisher BJ. On the usefulness of narratives: an interdisciplinary review and theoretical model. Ann Behav Med 2018;52(5):429-442 [FREE Full text] [doi: 10.1093/abm/kax008] [Medline: 29684135]
48. Joseph-Williams N, Abhyankar P, Boland L, Bravo P, Brenner AT, Brodney S, et al. What works in implementing patient decision aids in routine clinical settings? A rapid realist review and update from the International Patient Decision Aid Standards collaboration. Med Decis Making 2021;41(7):907-937 [FREE Full text] [doi: 10.1177/0272989X20978208] [Medline: 33319621]
49. Guide to knowledge translation planning at CIHR: integrated and end-of-grant approaches. Government of Canada, Canadian Institutes of Health Research. 2012. URL: http://www.cihr-irsc.gc.ca/e/45321.html [accessed 2018-10-08]
50. Muscat DM, Smith J, Mac O, Cadet T, Giguere A, Housten AJ, et al. Addressing health literacy in patient decision aids: an update from the International Patient Decision Aid Standards. Med Decis Making 2021;41(7):848-869 [FREE Full text] [doi: 10.1177/0272989X211011101] [Medline: 34053361]
51. The CART access project. Contraception and Abortion Research Team (CART-GRAC). 2024. URL: https://cart-grac.ubc.ca/the-cart-access-project-2/ [accessed 2024-01-28]

Abbreviations

IPDAS: International Patient Decision Aid Standards
PtDA: patient decision aid
SMOG: Simple Measure of Gobbledygook
SOGC: Society of Obstetricians and Gynaecologists of Canada
SURE: Sure of Myself, Understand Information, Risk-Benefit Ratio, Encouragement

Edited by T Leung; submitted 07.05.23; peer-reviewed by G Sebastian, R French, B Zikmund-Fisher; comments to author 11.01.24; revised version received 23.02.24; accepted 25.02.24; published 16.04.24.

Please cite as:
Wahl KJ, Brooks M, Trenaman L, Desjardins-Lorimer K, Bell CM, Chokmorova N, Segall R, Syring J, Williams A, Li LC, Norman WV, Munro S
User-Centered Development of a Patient Decision Aid for Choice of Early Abortion Method: Multi-Cycle Mixed Methods Study
J Med Internet Res 2024;26:e48793
URL: https://www.jmir.org/2024/1/e48793
doi:10.2196/48793
PMID:38625731

©Kate J Wahl, Melissa Brooks, Logan Trenaman, Kirsten Desjardins-Lorimer, Carolyn M Bell, Nazgul Chokmorova, Romy Segall, Janelle Syring, Aleyah Williams, Linda C Li, Wendy V Norman, Sarah Munro. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effective Privacy Protection Strategies for Pregnancy and Gestation Information From Electronic Medical Records: Retrospective Study in a National Health Care Data Network in China

Chao Liu[1], PhD; Yuanshi Jiao[1], MPH; Licong Su[2], MD; Wenna Liu[1], MD; Haiping Zhang[1], BS; Sheng Nie[2*], MD; Mengchun Gong[3*], MD

[1]Digital Health China Technologies Co, Ltd, Beijing, China
[2]Department of Nephrology, Nanfang Hospital, Southern Medical University, Guangzhou, China
[3]School of Biomedical Engineering, Guangdong Medical University, Zhanjiang, China
[*]these authors contributed equally

Corresponding Author:
Mengchun Gong, MD
School of Biomedical Engineering, Guangdong Medical University
No 2, Wenming East Road
Xiashan District
Zhanjiang, 524000
China
Phone: 86 18611768672
Email: gmc@nrdrs.org

Abstract

Background: Pregnancy and gestation information is routinely recorded in electronic medical record (EMR) systems across China in various data sets. The combination of data on the number of pregnancies and gestations can imply occurrences of abortions and other pregnancy-related issues, which is important for clinical decision-making and personal privacy protection. However, the distribution of this information inside EMR is variable due to inconsistent IT structures across different EMR systems. A large-scale quantitative evaluation of the potential exposure of this sensitive information has not been previously performed, ensuring the protection of personal information is a priority, as emphasized in Chinese laws and regulations.

Objective: This study aims to perform the first nationwide quantitative analysis of the identification sites and exposure frequency of sensitive pregnancy and gestation information. The goal is to propose strategies for effective information extraction and privacy protection related to women's health.

Methods: This study was conducted in a national health care data network. Rule-based protocols for extracting pregnancy and gestation information were developed by a committee of experts. A total of 6 different sub–data sets of EMRs were used as schemas for data analysis and strategy proposal. The identification sites and frequencies of identification in different sub–data sets were calculated. Manual quality inspections of the extraction process were performed by 2 independent groups of reviewers on 1000 randomly selected records. Based on these statistics, strategies for effective information extraction and privacy protection were proposed.

Results: The data network covered hospitalized patients from 19 hospitals in 10 provinces of China, encompassing 15,245,055 patients over an 11-year period (January 1, 2010-December 12, 2020). Among women aged 14-50 years, 70% were randomly selected from each hospital, resulting in a total of 1,110,053 patients. Of these, 688,268 female patients with sensitive reproductive information were identified. The frequencies of identification were variable, with the marriage history in admission medical records being the most frequent at 63.24%. Notably, more than 50% of female patients were identified with pregnancy and gestation history in nursing records, which is not generally considered a sub–data set rich in reproductive information. During the manual curation and review process, 1000 cases were randomly selected, and the precision and recall rates of the information extraction method both exceeded 99.5%. The privacy-protection strategies were designed with clear technical directions.

Conclusions: Significant amounts of critical information related to women's health are recorded in Chinese routine EMR systems and are distributed in various parts of the records with different frequencies. This requires a comprehensive protocol for

XSL•FO
RenderX

extracting and protecting the information, which has been demonstrated to be technically feasible. Implementing a data-based strategy will enhance the protection of women's privacy and improve the accessibility of health care services.

(J Med Internet Res 2024;26:e46455) doi:10.2196/46455

KEYWORDS

pregnancy; electronic medical record; privacy protection; risk stratification; rule-based

Introduction

Medical information is generally considered to be highly sensitive for individuals, and any breach of privacy can cause direct or indirect harm to patients [1]. For female patients, pregnancy and gestation information is not only highly private but also implies the incidence of abortion, which is extremely controversial in terms of the rights and responsibilities of women in some jurisdictions [2-4]. Evidence suggests that the leakage of such information can negatively impact the attitudes of patients' social environment and even health care providers [4,5].

The worldwide implementation of electronic medical records (EMRs) has significantly improved patient care by making health information readily accessible to a wide range of data producers. From 2007 to 2018, the average adoption rates of EMR increased from 18.6% to 85.3% [6]. This rapid growth has led to the processing and storage of various categories of patient information, including demographics, medications, laboratory tests, and diagnostic records, thereby establishing EMR as a valuable resource for large-scale data analysis of real-world data. However, the unprecedented use of EMR posed new challenges for protecting patient information effectively and preventing the unnecessary exposure of sensitive data during real-world evidence (RWE) research. Consequently, there is growing attention to the legal and technical research on extracting pregnancy and gestation information and the relevant privacy protection strategies [7,8].

On March 26, 2021, the Binhai Procuratorate accepted and examined a case of infringement of citizens' personal information. Staff responsible for preventive health care at a town town-central health center in Binhai County, Jiangsu province, took advantage of their positions to illegally obtain the family contact information and home addresses of pregnant women and newborns, totaling 25,124 items. This information was then resold through digital platforms, resulting in an illegal profit of US $4566 and subjecting pregnant women to telephone harassment. In response to this phenomenon, starting in 2022, local authorities began conducting annual comprehensive inspection of the supervision of fertility information and specifically informed the procuratorial organs of the inspection results [9]. New laws and regulations have also come out, such as the "Guangdong Province Maternal and Child Health Care Management Regulations" began to implement, which came into effect on June 1, 2023. These regulations emphasize the confidentiality of personal information and privacy in maternal and child health care services and related supervision and management [10].

According to the "Technical Specifications for Hospital Information Platforms Based on EMR" issued by the National Health Commission of China in 2014, different health institutions in the country share a similar EMR framework comprising several sub–data sets including diagnostic information, medical advice, laboratory test results, examination information, and surgical records [11]. However, issues of discontinuity and incompleteness in EMR writing pose significant challenges in multicenter data integration [12]. Traditional information extraction and privacy protection strategies during RWE research and clinical data transfer have primarily focused on fixed sub–data sets, such as marriage and childbearing history, and direct data entities like the number of pregnancies in patients' EMRs. These approaches, known as fixed site recognition strategies, lead to biased patient inclusion and flawed data masking in RWE research. For pregnancy and gestation information, testing results and procedures can indicate pregnancy status and gestation incidence without explicit descriptions in diagnostic sheets. For instance, a surgical history of pregnancy termination can imply suction aspiration abortion, while pregnancy history can be inferred from clinical test results such as human chorionic gonadotropin (HCG) levels exceeding 10 ng/L or 25 IU/L [13,14].

This study aims to propose protocols for the accurate and automatic extraction of pregnancy and gestation information from Chinese EMRs at the highest possible level of precision. Such information is crucial for patient inclusion and cohort identification in RWE studies to improve pregnancy outcomes [15,16]. Additionally, privacy protection strategies will be developed to maximize the masking of pregnancy data and identify the risk of privacy leakage for different sub–data sets within EMRs. To the best of our knowledge, this study is the first to identify the frequency of privacy information in Chinese EMRs. Then, the related risks can be considered when using patients' EMRs for RWE research.

Methods

Data Source

This retrospective study uses the Chinese Renal Disease Data System (CRDS) database, a comprehensive national EMR database. The CRDS includes data from 19 tertiary referral hospitals across 10 provinces, representing the 5 geographical regions of China (North, Central, East, South, and Southwestern). Each hospital's database covers the EMRs of all patients who visited from the start of 2010 to the end of 2020. The patient's EMRs were not specially selected. Complete EMRs from each hospital were transferred to the central database located at Nanfang Hospital of Southern Medical University in Guangzhou. In this study, the total number of patients in the database is 15,245,055. All analyzed

XSL•FO
RenderX

hospitalization records were structured based on the CRDS data model [17].

Sample Patient Inclusion

In this study, female patients aged 14-50 years from January 1, 2010, to December 31, 2020, were selected from the CRDS database. The statistical time here was the patient's last visit information (including all the previous visit history), and 70% (n=1,110,053) were randomly selected for statistical analysis.

Extraction of Pregnancy and Gestation Information From Chinese EMRs

Following a preliminary investigation of Chinese EMRs, and incorporating expert guidance, teaching materials, guidelines, and literature, the research team developed the Extraction Protocol of Pregnancy and Gestation Information (EPPGI). This protocol was refined through repeated sorting, adjustment, and verification, considering the writing characteristics of various hospital medical records. Traditional methods typically extract patient data using diagnosis codes from the diagnostic sheets of Chinese EMRs. However, we first developed identification rules for test and exam results, covering patients with positive HCG results in different units of measurement and pregnancy tests.

Given the diversity and complexity of the medical coding system in Chinese EMRs, we used regular expressions (regex) to retrieve pregnancy and gestation information across entire EMRs rather than relying solely on diagnosis codes in specific sub–data sets. The adopted regex extended beyond diagnoses to include surgical procedures, chief complaints related to pregnancy status, and gestation histories. Besides, regex for medications related to inducing labor or miscarriage was used to assist in identifying pregnancy information. Other regex, including description of fetus and exclusion rules, was also applied. All regex search patterns were the product of expert meetings and discussions. The detailed rules and regex of EPPGI are listed in Multimedia Appendix 1.

To implement this approach, we used R software (version 4.2.2; R Core Team) to extract females with reproductive activities (FRA) information from the checklist using regular expressions. In the following example, "final_medtech" represents the checklist, and "TECHNOLOGY_RESULT" is the field containing the check result in the checklist.

The rules of regex allowed us to describe the proportion of patients with a pregnancy history across different sub–data sets of EMRs and the frequency of pregnancy and gestation information. After removing duplicate patients from different sub–data sets, we retrieved data on pregnant women in the selected EMRs using EPPGI.

Privacy Protection Strategies

Based on the statistics of the located information, we proposed privacy protection strategies to avoid unnecessary and unintentional exposure of pregnancy and gestation information in real-world data analytics. Due to the different writing styles in medical records, insufficient desensitization may not fully cover sensitive patient information, while excessive desensitization may obscure other relevant information. First, EPPGI was used to identify keywords of sensitive reproductive information (SRI), such as "助产|难产|平产|早产|死产|死胎" ("midwifery|dystocia| normal birth|preterm birth|stillbirth|stillbirth"). With expert guidance, we finally chose to replace 15 characters before and after these keywords with asterisks (*) to desensitize sensitive information related to pregnancy and childbirth, thereby protecting patient privacy. This approach minimizes the possibility of inferring patients' SRI from EMRs.

In cases where the use of maternity-related information is unavoidable, the frequency of patient identification and privacy information was used to estimate the risk of unnecessary privacy exposure methodically. We also used diagnosis and marital history as criteria to locate maternity information and compared these results across 6 large sub–data sets of EMRs. A total of 2 independent reviewers (WL and HZ) inspected both methods to ensure accuracy and reliability.

Manual Curation and Verification

Afterward, included cases were randomly selected and manually reviewed by 2 independent groups of reviewers (CL, YJ, LS, WL, HZ, SN, and MG) to test the precision and recall of the data extraction. For the EPPGI, 1000 female cases were randomly assigned to 2 external experts (Aixin Guo and Wenna Liu) to manually extract SRI. The manually extracted results were then compared with the EPPGI results to evaluate the precision and recall rate, as defined below. We also compared the precision and recall rates of the EPPGI with those obtained using only maternal and diagnostic history.

Additionally, the reviewers attempted to identify FRA in privacy-concealed data sets to test the success rate of the privacy protection strategies, as defined in Figure 1.

Figure 1. Formulas for precision, recall, and success rate. Precision is the ratio of correctly predicted positive observations to the total predicted positives, focusing on the accuracy of the positive predictions. Recall is the ratio of correctly predicted positive observations to all the actual positives, focusing on the ability to capture all actual positive cases. EMR: electronic medical record; EPPGI: Extraction Protocol of Pregnancy and Gestation Information; FRA: females with reproductive activities.

$$\text{Precision rate of EPPGI} = \frac{\text{Manually verified FRA in retrived results}}{\text{All retrived FRA}}$$

$$\text{Recall rate of EPPGI} = \frac{\text{Manually verified FRA in retrived results}}{\text{Mannually verified FRA}}$$

$$\text{Success rate of marternity information concealing} = \frac{\text{Manually verified masked EMRs records}}{\text{Masked EMRs records}}$$

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Medical Ethics Committee of Nanfang Hospital, Southern Medical University (approval NFEC-2019-213), which waived the requirement for patient-informed consent due to the retrospective nature of the study. This study was also approved by the China Office of Human Genetic Resources for Data Preservation Application (approval 2021-BC0037). This study complied with the Declaration of Helsinki and the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.

Results

To the best of our knowledge, this study is the first to identify the frequency of privacy information in Chinese EMRs.

General Information of EMRs

All patient data were extracted from the CRDS database, a real-world database that includes records from 19 hospitals. Based on the inclusion criteria (female patients aged 14-50 years from January 1, 2010, to December 31, 2020) and a 70% entry ratio, a total of 1,110,053 patients were selected as the EMR sample. It is worth noting that removing duplicates reduced the sample size from 2,377,582 to 1,585,801, which is due to multiple diagnostic records for individual patients. The admittance flowchart is shown in Figure 2, and detailed information is displayed in Table 1. According to Chinese national specifications for standard EMR structure, EMRs consist of similar sub–data sets with minor differences in nomenclature including doctor's orders, diagnostic tables, test sheets, examination sheets, surgical sheets, and medical record texts. The medical record texts are further divided into 10 parts: course records, admission records, discharge records, referral records, consultation records, nursing records, death records, surgical notes, informed consent forms, and others. In CRDS, the admission record texts have been preprocessed using natural language processing for allergic history, chief complaint, disease history, tobacco and alcohol history, family history, marriage history, surgical history, and toxic exposure history. The general structure of Chinese EMRs is demonstrated in Figure 3.

Figure 2. Admittance flowchart. The initial database had 15,245,055 patients. After excluding male patients, specific years, teenagers, and older adults, removing duplicate patient records, and applying a 70% entry ratio, the final sample consisted of 1,110,053 observations. CRDS: chinese renal disease data system.

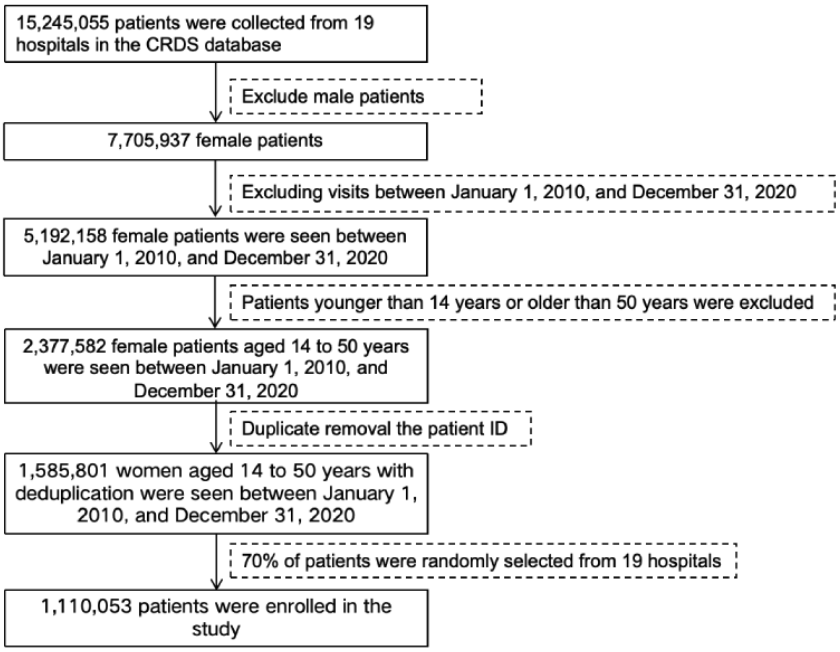

XSL•FO
RenderX

Table 1. General information on EMRsa from 19 hospitals.

Hospital number	City and area	Total bed numbers	FRA[b] (n=688,268)	Total patients (n=1,110,053)
1	Guangzhou, Southern	2225	57,837	102,483
2	Beijing, Northern	1650	30,410	36,757
3	Jinan, Northern	4000	79,294	94,339
4	Hangzhou, Eastern	3200	30,950	70,523
5	Hangzhou, Eastern	2400	67,086	82,977
6	Guangzhou, Southern	3000	53,355	82,654
7	Shenzhen, Southern	2000	41,352	48,269
8	Nanjing, Eastern	2499	33,709	51,919
9	Shanghai, Eastern	800	499	743
10	Chengdu, Southwestern	1000	21,803	78,843
11	Hefei, Eastern	3138	61,044	103,203
12	Wuhan, Central	5613	2555	4858
13	Maoming, Southern	2500	64,299	81,673
14	Guangzhou, Southern	2247	22,406	54,663
15	Huizhou, Southern	2156	22,756	23,181
16	Guiyang, Southwestern	2000	166	6138
17	Foshan, Southern	2200	63,336	125,085
18	Guangzhou, Southern	3000	6358	20,100
19	Guangzhou, Southern	1000	29,053	41,645

[a]EMR: electronic medical record.

[b]FRA: females with reproductive activities.

Figure 3. Structure of Chinese EMRs. The EMRs consist of medical records, surgery, examination results, test results, medication, and diagnosis. CPOE: computerized physician order entry; CRDS: chinese renal disease data system; EMR: electronic medical record; HIS: hospital information system; LIS: laboratory information system; NLP: natural language processing.

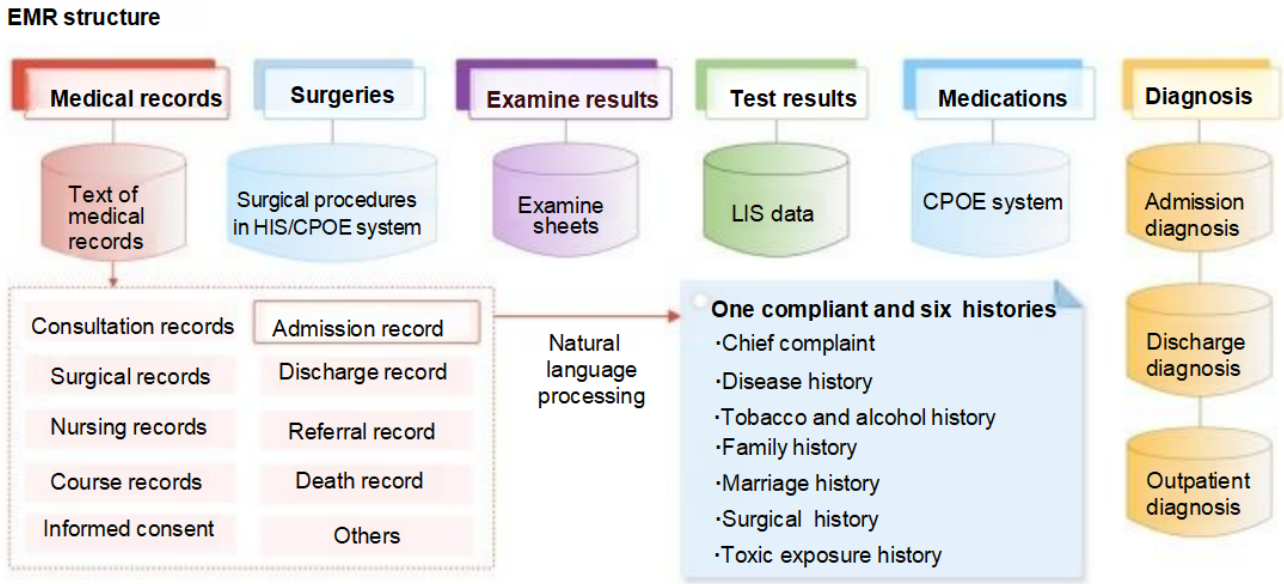

Number of FRA

After the initial investigation, we applied the EPPGI to a sample of 1,110,053 female patients of childbearing age. This analysis covered 6 different categories of EMRs, with each sub–data set and its components processed separately. Table 2 presents the total number of patients, the identified number of FRA, and their corresponding proportions.

XSL•FO
RenderX

Table 2. Patient identified by EPPGIa. From left to right, the columns display the total number of patients, the number of identified FRAb, and their corresponding proportions.

EMR[c] sub–data sets	Patient number (per person)	Maternal patient number	Percentage (%)
Order	955,140	146,555	15.34
Diagnosis from the frontage	1,073,167	312,008	29.07
Laboratory report	903,987	93,386	10.33
Examine result	852,143	172,735	20.27
Prescription of surgical procedures in HIS[d]/CPOE[e] system	767,693	157,027	20.45
Medical records			
Total	588,963	393,550	66.82
Course records	207,575	95,012	45.77
Discharge records	330,909	112,927	34.13
Referral records	9699	2682	27.65
Consultation records	38,728	12,639	32.64
Nursing records	192,080	112,465	58.55
Death records	953	105	11.02
Surgical notes	134,889	43,280	32.09
Informed consent	238,014	102,386	43.02
Others	411,637	138,892	33.74
Admission records (NLP[f])			
Total	376,176	317,962	84.52
Allergic history	446,360	0	0.00
Chief complaint	13,925	55	0.39
Disease history	446,390	11,312	2.53
Tobacco and alcohol history	490,028	0	0.00
Family history	464,516	6	0.00
Marriage history	467,184	295,436	63.24
Surgical history	316,282	74,118	23.43
Toxic exposure history	504,022	4	0.00

[a]EPPGI: Extraction Protocol of Pregnancy and Gestation Information.

[b]FRA: females with reproductive activities.

[c]EMR: electronic medical record.

[d]HIS: hospital information system.

[e]CPOE: computerized physician order entry.

[f]NLP: natural language processing.

The number of pregnancies identified solely by diagnosis was 312,008, accounting for 29.07% of the patients in the diagnostic sub–data set. The number of patients who were identified only by their marital and childbearing history was as high as 295,436, accounting for 26.61% of the total study population. The number of pregnancies identified by diagnosis and marital and childbearing history was 521,132, accounting for 46.95% of the total study population. If on the basis of diagnosis and marital history, the identification of diagnosis, examination, and other contents are added, the number of maternity information can be identified as 688,268, accounting for 62% of the total study population.

In the text of medical records, 393,550 patients with SRI were identified, accounting for 66.82% of 588,963 records. Due to the presence of childbearing history, which constitutes the leading source of SRI, over 80% (n=317,962) of female patients in admission records were identified as FRA by EPPGI. Besides, 58.55% (n=112,465) of female patients were identified in nursing history, making it the second highest proportion of FRA.

Based on these results, EPPGI effectively extracts FRA from every sub–data set within Chinese EMRs.

XSL•FO
RenderX

Frequency of Recognition

A single patient can generate multiple encounter records in the EMR system per visit. Therefore, individual EMRs were divided into separate records based on visits, reflecting the actual EMR storage in RWE studies. Table 3 presents the frequency of pregnancy information identification across different sub–data sets of Chinese EMRs. Similar to the results from per-patient records, SRI can be widely identified in each sub–data set of EMRs. SRI is primarily concentrated on diagnosis records, surgical records, and medical records text. In diagnosis records, SRI could be extracted from 15.06% of 15,497,063 records. In surgical records, 11.49% of SRI could be extracted from 1,604,579 records. The text of medical records showed the highest frequency of SRI identification, with an overall recognition rate of 29.92%. Additionally, it is noteworthy that more than 80% of admission records contained SRI.

Table 3. Frequency of pregnancy information identification. From left to right, the columns display the total number of records, the frequency of identified FRA[a], and their corresponding proportions.

EMR[b] sub–data sets	Record number (per visit)	Maternal record number	Percentage (%)
Order	93,182,790	384,699	0.41
Diagnosis from the frontage	15,497,063	2,334,160	15.06
Laboratory report	102,509,232	285,245	0.28
Examine result	6,790,300	549,078	8.09
Prescription of surgical procedures in HIS[c]/CPOE[d] system	1,604,579	184,335	11.49
Medical records			
Total	8,473,462	2,534,940	29.92
Course records	2,132,926	527,915	24.75
Discharge records	532,790	151,352	28.41
Referral records	25,171	6737	26.76
Consultation records	151,564	43,695	28.83
Nursing records	965,042	268,586	27.83
Death records	2250	166	7.38
Surgical notes	482,578	106,656	22.10
Informed consent	1,226,875	326,284	26.59
Others	2,377,080	637,807	26.83
Admission records (NLP [e])			
Total	577,186	465,742	80.69
Allergic history	1,183,577	0	0.00
Chief complaint	45,987	59	0.13
Disease history	4,166,715	15,057	0.36
Tobacco and alcohol history	858,066	0	0.00
Family history	2,076,062	6	0.00
Marriage history	708,544	444,559	62.74
Surgical history	553,774	106,846	19.29
Toxic exposure history	3,954,196	4	0.00

[a]FRA: females with reproductive activities.
[b]EMR: electronic medical record.
[c]HIS: hospital information system.
[d]CPOE: computerized physician order entry.
[e]NLP: natural language processing.

Precision and Recall Rate

During the manual curation and certification process, 1000 complete EMRs were randomly selected from the sample patients and reviewed by 2 independent medical experts (Aixin Guo and Wenna Liu) to determine maternal status. The precision and recall rates of the EPPGI were 100% and 99.68%, respectively. When only diagnosis history and marital history

XSL•FO
RenderX

were used for identification, the accuracy rate remained 100%, but the recall rate dropped to 73.35%. For details, see Tables 4 and 5, where "0" represents patients without FRA information and "1" represents patients with FRA information.

Table 4. Confusion matrix of EPPGI[a] method.

Prediction	Reference	
	0	1
0	377	2
1	0	621

[a]EPPGI: Extraction Protocol of Pregnancy and Gestation Information.

Table 5. Confusion matrix of the method using only diagnosis and marital history

Prediction	Reference	
	0	1
0	377	166
1	0	457

We also conducted analyses by time and region, as shown in Tables 6 and 7. In these tables, "quality inspection" refers to patients assessed by 2 expert manual reviews for quality control to determine the presence of labor process information (from different hospital sources); "EPPGI" refers to patients assessed using EPPGI for maternity information; and "diagnosis history and marital history" refers to patients assessed using diagnosis and marital history for fertility information. In these tables, 0 represents "no maternity information" and 1 represents "there is maternity information." The results indicated that similar to the overall comparison, the identification of maternal information using the EPPGI method was superior to using diagnosis and marital history alone. By examining the results across different hospitals and time periods, our method proved to be universally applicable across various years and regions.

XSL•FO
RenderX

Table 6. FRA[a] identification results of different privacy methods in different hospitals (1 being "there is maternity information" and 0 being "no maternity information").

Hospital number	Quality inspection		EPPGI[b]		Diagnosis history and marital history	
	0	1	0	1	0	1
1	40	49	40	49	57	32
10	57	18	57	18	62	13
11	40	54	40	54	45	49
12	3	4	3	4	3	4
13	15	64	16	63	42	37
14	29	16	29	16	29	16
15	0	17	0	17	3	14
16	2	1	2	1	3	0
17	48	55	48	55	71	32
18	17	7	17	7	17	7
19	6	31	6	31	9	28
2	4	28	4	28	10	22
3	11	62	11	62	13	60
4	40	33	41	32	55	18
5	15	57	15	57	33	39
6	29	53	29	53	56	26
7	8	45	8	45	13	40
8	13	29	13	29	22	20

[a]FRA: females with reproductive activities.

[b]EPPGI: Extraction Protocol of Pregnancy and Gestation Information.

Table 7. FRA[a] recognition results for different privacy methods in different years (1 being "there is maternity information" and 0 being "no maternity information").

Year	Quality inspection		EPPGI[b]		Diagnosis history and marital history	
	0	1	0	1	0	1
2010	14	9	14	9	19	4
2011	18	10	18	10	23	5
2012	19	18	19	18	29	8
2013	42	40	42	40	55	27
2014	46	59	46	59	60	45
2015	33	75	34	74	55	53
2016	42	93	42	93	69	66
2017	63	130	64	129	95	98
2018	60	107	60	107	86	81
2019	27	58	27	58	36	49
2020	13	24	13	24	16	21

[a]FRA: females with reproductive activities.

[b]EPPGI: Extraction Protocol of Pregnancy and Gestation Information.

XSL•FO
RenderX

Privacy Protection Strategies

The privacy-protection strategies were developed based on the above results. Given that we used regular expressions to identify SRI, additional text surrounding the recognized maternity information needs to be concealed to prevent privacy exposure through context. We randomly selected 1000 EMRs of pregnancy patients for static data desensitization to create a masked sample of EMRs. A total of 2 independent reviewers (Aixin Guo and Wenna Liu) were assigned to manually extract any form of pregnancy and gestation information from the masked samples. Furthermore, the risk of unnecessary privacy exposure was stratified by the frequency of recognition. The text of medical records, having the highest recognition frequency, should be handled with the utmost caution. In contrast, test and examination records are less frequently identified with SRI. It is important to note that the frequency of recognition does not fully represent the risk of privacy leakage, which will be further analyzed in the discussion section.

Discussion

Principal Findings

Overview

This study is one of the first large-scale investigations into privacy leakage and FRA identification of Chinese EMRs, focusing on the frequency of recognition. The originality of this work can be summarized in 3 key aspects.

Originality in Exploring New Observations

The accessibility of EMR inevitably leads to uneven privacy protection awareness among different EMR users. The importance of reliable privacy protection methods has been extensively discussed in the literature, emphasizing their critical role in the successful implementation of EMRs in health care institutions [18,19]. Sensitive information regarding pregnancy, gestation, and abortion is routinely included in EMRs, raising concerns about unnecessary exposure [20,21]. In 2021, the Personal Information Protection Law of the People's Republic of China came into effect, which clarified the rights and responsibilities related to the use of personal privacy information [22]. However, prior to this study, there has been little to no effort to address the highest standards of patient privacy protection protocols during RWE studies. To the best of our knowledge, this is the first study in China to use a national-level EMR database to quantitatively evaluate the exposure risk of privacy information related to women's reproductive health. This study aims to enhance protection strategies in this area.

Originality in Designing New Experiments

The attributes and structure of Chinese EMRs are unique in terms of terminology and data standards. Accurate and comprehensive recognition of maternity information is widely reported to play a critical role in effective privacy protection and the evaluation of RWE [23,24]. While researchers have been working to improve the accuracy of SRI identification in non-Chinese EMRs [25,26], to the best of our knowledge, no prior research has focused on the accurate and complete extraction of FRA from Chinese EMRs. Traditional diagnosis-based patient extraction protocols typically use diagnosis codes, such as the International Classification of Diseases, which have 2 major limitations.

First, the records in the diagnosis sheet are often incomplete. Due to inconsistencies in Chinese EMR documentation, physicians do not always record pregnancy and gestation information as a diagnosis, especially when the patient's primary complaint is unrelated to maternity. This leads to lower recall rates and potential recall bias. Second, due to the complexity and inconsistency of coding systems in Chinese EMRs, using codes for patient identification is more complicated than using regex, and it is nearly impossible to list all encodings exhaustively. Furthermore, regex can be widely adopted across different sub–data sets of Chinese EMRs. Although the diagnostic sheet contains the major SRI, most Chinese EMRs are still stored in text format without code mapping.

Compared to traditional diagnosis-based patient inclusion methods, the EPPGI method provides more precise results in a practical manner. Whether patient- or visit-based records, EPPGI extracts significantly more FRA with a high precision rate.

Originality in Contributing New Knowledge

Our results demonstrate that traditional fixed-site data masking procedures lead to considerable unnecessary exposure of privacy information. For instance, patients' HCG test results or delivery procedures are commonly recorded in sub–data sets that cannot simply be concealed during RWE studies and clinical use. The combination of pregnancy and gestation information can even infer the incidence of abortion, which is highly confidential in China. Accurate and complete recognition of maternity information is essential for flawless privacy protection.

EPPGI method first identified pregnancy and gestation information across entire Chinese EMRs. For the identified information, it is practical and convenient to use data desensitization techniques, including data invalidation, data offset, and symmetric encryption, to prevent the misuse of private data. Based on the EMRs in CRDS, we determined the optimal length of additional concealed text to retain most medical information. Additionally, the quantified recognition frequency of pregnancy and gestation information helps researchers use EMRs wisely to avoid unnecessary privacy leakage. Although the frequency of identification cannot fully determine the risk of privacy leakage, which is also associated with the complexities of data desensitization, these results highlight the richness of private information ingrained in EMRs.

From a data asset management perspective, quantifying the risk of privacy leakage is critical under the strict Personal Information Protection Law. Based on statistical results and actual data mining practices, SRI is widely stored in Chinese EMRs, requiring data desensitization when using any EMR sub–data sets. For test results and structured data, the difficulty of data desensitization is relatively lower than that for plain text medical records, given the explicit nature of sensitive data entities and the low probability of reinference from context. Similar to the hazard classification of chemicals, health care data users should be aware of the richness of private information and the risk of unnecessary privacy exposure in EMRs.

XSL•FO
RenderX

Maternity information is considered one of the most sensitive types of privacy for women, and our results provide a crucial reference for data users to assess related risks in Chinese EMRs for the first time.

Limitations

Overall, this work provides justification for assessing privacy leakage risk and offers a reference for effective privacy protection in Chinese EMRs. However, the proposed study has several limitations. First, the frequency of sensitive information and the privacy risk estimated in our case study are primarily based on the EMRs of a renal disease database. While there are official directions and guidelines for composing EMRs in China [9], discrepancies exist between the CRDS and other data networks in terms of data structure and operating environment. Specific protocols and variables should be optimized for generalizations. Furthermore, the study is limited by its data scale, covering only 688,268 FRA in the CRDS. This limited scope suggests the need for further research involving larger data sets to validate and refine our findings.

Conclusions

Finding an effective and practical way to protect private information in EMRs is both meaningful and useful. We have demonstrated the feasibility of applying the EPPGI method to EMRs from 19 hospitals in different regions. We believe that EPPGI can provide a valuable reference for patient inclusion in any maternity-related studies using Chinese EMRs. Our protocols, designed for Chinese EMR systems, enable the accurate and complete recognition and extraction of pregnancy and gestation data, ensuring its effective protection. Compared to traditional methods of FRA inclusion, the EPPGI method provides more comprehensive results.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFC2500200 and 2023YFC2706305). This work was supported by the Multi-modality Data Integration and Application Lab of Guangdong Medical University and the National Clinical Research Center for Geriatric Disorders (Huashan). We did not use generative artificial intelligence in any portion of the manuscript writing.

Data Availability

The data sets analyzed during this study are not publicly available due to the sensitivity of hospitals' data but are available from the corresponding author upon reasonable request. The source code is available in Multimedia Appendix 2.

Authors' Contributions

MG, CL, and SN conceived and designed the study. LS collected the data. YJ and WL drafted the initial manuscript. HZ and YY integrated and revised the manuscript. MG and SN served as co–corresponding authors. CL, YJ and LS served as co–first authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The detailed rules and regex of EPPGI (Extraction Protocol of Pregnancy and Gestation Information).
[DOCX File , 17 KB - jmir_v26i1e46455_app1.docx]

Multimedia Appendix 2
Source code.
[ZIP File (Zip Archive), 12 KB - jmir_v26i1e46455_app2.zip]

References

1. Patrick H. Towards a privacy access control model for e-Healthcare services. In: DBLP. 2005 Presented at: Conference on Privacy, Security and Trust; October 12, 2005; Canada URL: https://www.semanticscholar.org/paper/Towards-a-Privacy-Access-Control-Model-for-Services-Hung/e219290f8dc650863831718e97dd0e5e6b2064a2
2. Johari V, Jadhav U. Abortion rights judgment: a ray of hope!. Indian J Med Ethics 2017;2(3):180-183 [FREE Full text] [doi: 10.20529/IJME.2017.044] [Medline: 28279947]
3. Uterhark E. International law and the legalization of abortion in Northern Ireland. J Law Health 2020;34(1):155-189. [Medline: 33449459]
4. Påfs J, Rulisa S, Klingberg-Allvin M, Binder-Finnema P, Musafili A, Essén B. Implementing the liberalized abortion law in Kigali, Rwanda: ambiguities of rights and responsibilities among health care providers. Midwifery 2020;80:102568 [FREE Full text] [doi: 10.1016/j.midw.2019.102568] [Medline: 31698295]

XSL•FO
RenderX

5. Rigdon SM. Abortion law and practice in China: an overview with comparisons to the United States. Soc Sci Med 1996;42(4):543-560. [doi: 10.1016/0277-9536(95)00173-5] [Medline: 8643980]
6. Liang J, Li Y, Zhang Z, Shen D, Xu J, Zheng X, et al. Adoption of electronic health records (EHRs) in China during the past 10 years: consecutive survey data analysis and comparison of Sino-American challenges and experiences. J Med Internet Res 2021;23(2):e24813 [FREE Full text] [doi: 10.2196/24813] [Medline: 33599615]
7. Hager A, Lindblad S, Brommels M, Salomonsson S, Wannheden C. Sharing patient-controlled real-world data through the application of the theory of commons: action research case study. J Med Internet Res 2021;23(1):e16842 [FREE Full text] [doi: 10.2196/16842] [Medline: 33464212]
8. Festag S, Spreckelsen C. Privacy-preserving deep learning for the detection of protected health information in real-world data: comparative evaluation. JMIR Form Res 2020;4(5):e14064 [FREE Full text] [doi: 10.2196/14064] [Medline: 32369025]
9. Typical cases of public interest litigation for the protection of women's rights. URL: https://www.spp.gov.cn/xwfbh/dxal/202211/t20221125_593721.shtml [accessed 2024-07-30]
10. Regulations of Guangdong Province on the management of maternal and child health care. People's Government of Guangdong Province. URL: https://www.gd.gov.cn/zwgk/wjk/zcfgk/content/post_2524333.html [accessed 2024-07-17]
11. Technical specification of hospital information platform based on electronic medical record. National Health Commission of the People's Republic of China. 2014. URL: http://www.nhc.gov.cn/wjw/s9497/201406/a2014514701f4e76b14f3446f6318937.shtml [accessed 2024-07-17]
12. Wang Z. Data integration of electronic medical record under administrative decentralization of medical insurance and healthcare in China: a case study. Isr J Health Policy Res 2019;8(1):24 [FREE Full text] [doi: 10.1186/s13584-019-0293-9] [Medline: 30929644]
13. Cervinski MA, Lockwood CM, Ferguson AM, Odem RR, Stenman UH, Alfthan H, et al. Qualitative point-of-care and over-the-counter urine hCG devices differentially detect the hCG variants of early pregnancy. Clin Chim Acta 2009;406(1-2):81-85. [doi: 10.1016/j.cca.2009.05.018] [Medline: 19477170]
14. Montagnana M, Trenti T, Aloe R, Cervellin G, Lippi G. Human chorionic gonadotropin in pregnancy diagnostics. Clin Chim Acta 2011;412(17-18):1515-1520. [doi: 10.1016/j.cca.2011.05.025] [Medline: 21635878]
15. Farland LV, Prescott J, Sasamoto N, Tobias DK, Gaskins AJ, Stuart JJ, et al. Endometriosis and risk of adverse pregnancy outcomes. Obstet Gynecol 2019;134(3):527-536 [FREE Full text] [doi: 10.1097/AOG.0000000000003410] [Medline: 31403584]
16. Attali E, Yogev Y. The impact of advanced maternal age on pregnancy outcome. Best Pract Res Clin Obstet Gynaecol 2021;70:2-9. [doi: 10.1016/j.bpobgyn.2020.06.006] [Medline: 32773291]
17. China Renal Data System. URL: http://www.crds-network.org.cn/#/database [accessed 2024-07-17]
18. Kuo KM, Ma CC, Alexander JW. How do patients respond to violation of their information privacy? Health Inf Manag 2014;43(2):23-33. [doi: 10.1177/183335831404300204] [Medline: 24948663]
19. Moore W, Frye S. Review of HIPAA, part 1: history, protected health information, and privacy and security rules. J Nucl Med Technol 2019;47(4):269-272 [FREE Full text] [doi: 10.2967/jnmt.119.227819] [Medline: 31182664]
20. Hackett KM, Kazemi M, Sellen DW. Keeping secrets in the cloud: mobile phones, data security and privacy within the context of pregnancy and childbirth in Tanzania. Soc Sci Med 2018;211:190-197. [doi: 10.1016/j.socscimed.2018.06.014] [Medline: 29960170]
21. Zhang H, Zhang H, Zhang Z, Wang Y. Patient privacy and autonomy: a comparative analysis of cases of ethical dilemmas in China and the United States. BMC Med Ethics 2021;22(1):8 [FREE Full text] [doi: 10.1186/s12910-021-00579-6] [Medline: 33531011]
22. NA. The National People's Congress of the People's Republic of China. 2021 Presented at: Data Security Law of the People's Republic of China; July 2017, 2024; Beijing URL: http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html
23. Gokhale M, Stürmer T, Buse JB. Real-world evidence: the devil is in the detail. Diabetologia 2020;63(9):1694-1705 [FREE Full text] [doi: 10.1007/s00125-020-05217-1] [Medline: 32666226]
24. Liu M, Qi Y, Wang W, Sun X. Toward a better understanding about real-world evidence. Eur J Hosp Pharm 2022;29(1):8-11 [FREE Full text] [doi: 10.1136/ejhpharm-2021-003081] [Medline: 34857642]
25. Angras K, Boyd VE, Gray C, Young AJ, Paglia MJ, Mackeen AD. Retrospective application of algorithms to improve identification of pregnancy outcomes from the electronic health record. J Perinatol 2023;43(1):10-14. [doi: 10.1038/s41372-022-01496-1] [Medline: 36050515]
26. Moll K, Wong HL, Fingar K, Hobbi S, Sheng M, Burrell TA, et al. Validating claims-based algorithms determining pregnancy outcomes and gestational age using a linked claims-electronic medical record database. Drug Saf 2021;44(11):1151-1164 [FREE Full text] [doi: 10.1007/s40264-021-01113-8] [Medline: 34591264]

Abbreviations

CRDS: Chinese Renal Disease Data System
EMR: electronic medical record
EPPGI: Extraction Protocol of Pregnancy and Gestation Information

XSL•FO
RenderX

FRA: females with reproductive activities
HCG: human chorionic gonadotropin
RWE: real-world evidence
SRI: sensitive reproductive information
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 02.05.23; peer-reviewed by S Chaichulee, M Gasmi , W Shuang, F Yu; comments to author 13.09.23; revised version received 02.01.24; accepted 22.06.24; published 20.08.24.

Please cite as:
Liu C, Jiao Y, Su L, Liu W, Zhang H, Nie S, Gong M
Effective Privacy Protection Strategies for Pregnancy and Gestation Information From Electronic Medical Records: Retrospective Study in a National Health Care Data Network in China
J Med Internet Res 2024;26:e46455
URL: https://www.jmir.org/2024/1/e46455
doi:10.2196/46455
PMID:

©Chao Liu, Yuanshi Jiao, Licong Su, Wenna Liu, Haiping Zhang, Sheng Nie, Mengchun Gong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Social and Ethical Implications of Digital Crisis Technologies: Case Study of Pandemic Simulation Models During the COVID-19 Pandemic

Gabriel Bartl[1], BA, MA, DPhil

Centre Marc Bloch, Berlin, Germany

Corresponding Author:
Gabriel Bartl, BA, MA, DPhil
Centre Marc Bloch
Friedrichstr 191
Berlin, 10117
Germany
Phone: 49 30 209370700
Email: gabriel.bartl@cmb.hu-berlin.de

Abstract

Background: Responses to public health crises are increasingly technological in nature, as the prominence of COVID-19–related statistics and simulations amply demonstrates. However, the use of technologies is preconditional and has various implications. These implications can not only affect acceptance but also challenge the acceptability of these technologies with regard to the ethical and normative dimension.

Objective: This study focuses on pandemic simulation models as algorithmic governance tools that played a central role in political decision-making during the COVID-19 pandemic. To assess the social implications of pandemic simulation models, the premises of data collection, sorting, and evaluation must be disclosed and reflected upon. Consequently, the social construction principles of digital health technologies must be revealed and examined for their effects with regard to social, ethical, and ultimately political issues.

Methods: This case study starts with a systematization of different simulation approaches to create a typology of pandemic simulation models. On the basis of this, various properties, functions, and challenges of these simulation models are revealed and discussed in detail from a socioscientific point of view.

Results: The typology of pandemic simulation methods reveals the diversity of model-driven handling of pandemic threats. However, it is reasonable to assume that the use of simulation models could increasingly shift toward agent-based or artificial intelligence models in the future, thus promoting the logic of algorithmic decision-making in response to public health crises. As algorithmic decision-making focuses more on predicting future dynamics than statistical practices of assessing pandemic events, this study discusses this development in detail, resulting in an operationalized overview of the key social and ethical issues related to pandemic crisis technologies.

Conclusions: This study identifies 3 major recommendations for the future of pandemic crisis technologies.

(J Med Internet Res 2024;26:e45723) doi:10.2196/45723

KEYWORDS

public health technologies; simulation models; algorithmic governance; preparedness; crisis, uncertainty, and ignorance; social implications of mathematical modeling; normativity; transparency; legitimacy

Introduction

Knowledge Production and the Political Significance of Data in Times of Uncertainty

The COVID-19 pandemic highlighted 2 opposing trends that have been lost in the public debate. On the one hand, the pandemic showed that “digital prediction tools increasingly complement or replace other practices of coping with an uncertain future” [1]. In contrast, the pandemic blatantly illustrated the complicated relationship between scientific knowledge and ignorance in the context of political decisions and social dynamics in dealing with crises. In accordance with the Bauman [2] diagnosis of the “end of unambiguousness,”

XSL•FO
RenderX

Habermas [3] stated that "There has never been so much knowledge about our not-knowing and about the compulsion to act and live under uncertainty."

These 2 opposing trends point to the observation that knowledge is ambivalent, increasingly fragile, and ambiguous but simultaneously acts as a central resource [4]. This has not least to do with the fact that there is no such thing as a single, definitive science that produces objective findings. In the context of digital health technologies, this seems to be a paradox as the increasing use and implementation contrasts with the possibility of clear evidence production. The growing demand for scientific expertise in times of crisis and the simultaneous lack of clarity in scientific knowledge represent a paradox that cannot be resolved and seems to be becoming increasingly severe [5].

The Ideology of "Technological Fix" in Public Health Crisis Management

Especially during the COVID-19 pandemic, it has been shown that promises of unambiguity and evidence must be illusory and that both ascriptions must be contrasted with the hypothesis of a "situatedness of knowledge" [6]. Thus, knowledge must be reflected in its contexts of production. This means that a simple translation of the relationship between knowledge and ignorance into the political sphere, for example, within preventive health programs, must be the subject of debate while including diverse stakeholders such as health experts, policy makers, and "lay persons" [7]. Finally, this also indicates that we must speak of science in the plural and that policy cannot claim to follow a single, definitive science when there are conflicting findings that compete with each other.

Objective

This situation has multifaceted consequences and implications for knowledge-based crisis technologies in public health. Although temporal pressure and threat under conditions of uncertainty and insecurity are elementary characteristics of crises [8], multiple crises, such as global pandemics, are characterized by the fact that these factors interact even more intensively not least because the level of interactions, which are difficult to anticipate, increases the degree of complexity, uncertainty, and ignorance. Those experiences with limited control capabilities have paved the way for a security strategy of preparedness that has prevailed especially in the field of global public health where gaining importance of the "precautionary principle" can be observed. In the same breath, different modes of resilience gained attraction, especially with regard to the emergence of preparedness practices and strategies that are often technological in nature [9]. The rise of a solutionist ideology of "technological fix" [10] in public health crisis management is expressed especially in the use of computer models that simulate the spread of the coronavirus. As certain simulation techniques are likely to play an even more significant role in the pandemic technologies of the future, the following sections will take a deeper look into the premises, logics, and differences of simulation models before shedding light on the social, ethical, and political implications of digital crisis technologies in the context of public health.

Undoubtedly, computer simulations grounded on epidemiological models have played a crucial role in handling the pandemic. This happened, for example, because these models provided orientation knowledge in a crisis situation with considerable temporal pressure. Therefore, it could be argued that one of the central attributions to the use of monitoring technologies during crises such as pandemics is the expected time gains that are essential to preserve the decision makers' ability to act.

Methods

Overview

Methodologically, this case study is based on a systematization of different simulation approaches to create a typology of pandemic simulation models. On the basis of this, various properties, functions, and challenges of these simulation models are revealed, such as their perception as visual representations or certain problems in converting complexity into numerical parameters. Subsequently, to what extent pandemic simulation models can be considered as algorithmic governance tools is explored. Thus, the methodological approach is closely interwoven with a discussion of the ethical, social, and political implications of using simulations.

Ethical Considerations

No ethics approval was requested, as the methodological approach is based on a description and comparison of pandemic simulation models. Therefore, no personal data were collected, and only relevant literature was referenced to clarify the functioning of the corresponding simulation models.

Results

Concerning their functional logics, simulation models are to be distinguished from other forms of crisis governance, such as, for example, early warning systems that aim to forecast the future by making use of prognostic methods. In contrast, the vast majority of simulation models are scenario-based approaches that are not grounded on probabilistic calculations but contain different ways of dealing with uncertainty and crises by comparing different courses of action and by considering both the effects of the assumed political countermeasures and the respective societal coping modes. Simulations thus provide policy makers with information by contrasting measures with their possible effects within algorithmic procedures.

The interventions of the policy makers are thus tested according to a *what if* logic so that they can be compared in terms of their effects.

In this respect, epidemiological computer simulations on the one side can be seen as "technologies of preparedness" [11] that stand in contrast to purely prognostic approaches aiming at mitigating future threats by number-based calculations. On the other side, specific future-oriented containment strategies can also be simulated during a public health crisis that has already occurred and is still ongoing, for example, to identify bottlenecks in the infrastructure. What both have in common is a doubling of reality [12]. They are designed to ensure consistency, speed,

XSL•FO
RenderX

efficiency, and removal of human bias and error. Hälterlein [12] pointed out that within computer models, different countermeasures together with their uncertain outcomes are simulated based on algorithmic processes. This led him to argue that the use of epidemiological computer simulations is to be understood as a process of algorithmic decision-making [12].

There are a number of different simulation models that are briefly presented and distinguished from each other in the following sections. However, Table 1 represents an ideal typical listing of pandemic model simulations, which means that clear distinctions often cannot be made and single modeling techniques may also partially overlap. It is also noticeable that, although models are always accompanied by a certain reduction in complexity, this parameter can vary dramatically with respect to specific simulation techniques. In addition, the claims of the models vary, as some of them have prognostic ambitions to some degree. Finally, it should be mentioned that the relevance of the single models to policy making varied from country to country during the pandemic [13].

Table 1. Typology of COVID-19 pandemic simulation models.

Modeling techniques	Specifications	Features
Compartmental models	Division of the population into different groups, for example, SEIR[a]	Infection dynamics are modeled with respect to the transitions between those groups
Statistical models	Development and testing of theories through causal explanation, prediction, and description (eg, growth models or time series)	Explanatory power of models corresponds with predictive power
Bayesian methods	Specific statistical approach: available knowledge about statistical parameters is merged with data from observed information	Bayesian methods can be used even with a small data base
Network models	Analysis of the distributions in the network links to be able to distinguish certain network types from each other	Search for patterns in the contact structures
Agent-based models	The population to be modeled is divided into subgroups and is grounded on agents with different individual behaviors	Social context is central in contrast to other modeling techniques
AI[b] models	(Deep) learning algorithms, neural networks, or adaptive agents adjusting their behaviors to changing environmental conditions	Aiming more on prediction (eg, incidence rates) and forecasting than on description
Hybrid models	Combination of different modeling techniques	Depending on the modeling techniques that are combined (eg, SEIR with machine learning to predict the evolution of the pandemic [13])

[a]SEIR: susceptible, exposed, infectious, recovered.

[b]AI: artificial intelligence.

The most common and also the most popular simulation models during the pandemic were the so-called compartmental models. Here, the susceptible, infectious, recovered (SIR) [14] models can be distinguished from the susceptible, exposed, infectious, recovered (SEIR) models, although many more variants exist. For example, the influential Ferguson model [15], which was instrumental in determining COVID-19 pandemic policy in the United Kingdom, is a SEIR model that divided society into 4 groups: S (“susceptible”), E (“exposed”; here, “E” refers to those who have been exposed, but who are not yet infectious [16]), I (“infectious”), and R (“recovered”). On the basis of these 4 dimensions, the dynamics of the pandemic can be described by modeling the transitions between groups. However, the social context is not modeled to generate a model that is not too complex and that can quickly generate rough approximations. However, there are also improved SEIR models that comodel, for example, quarantine status and intervention measures [17].

In addition to compartmental models, according to Gnanvi et al [18], statistical models have been used quite frequently during the COVID-19 pandemic. These are mostly growth models and time series models. Although statistical models often model one state (ie, one compartment at a certain point in time), statistical models, such as regression models, describe the statistical relationships between different independent variables and can provide information about the individual health status. Bayesian methods [19] and network models [20] and agent-based models [21] were also very elaborate but were used less frequently during the COVID-19 pandemic. No role during the pandemic played models based on artificial intelligence (AI) [22]. Finally, hybrid models that combine ≥2 of the presented approaches also exist. Looking at the application contexts of the individual models, it is striking that statistical and Bayesian models have been used to estimate epidemiological parameters, whereas compartmental models have normally been used to assess disease dynamics. Furthermore, simulation models that emerged during the COVID-19 pandemic found that the vast majority of models were SEIR models and statistical approaches [18]. This might change in the future as machine learning and AI models will become more prominent when computing capacities improve and new methodological approaches emerge. In addition, Eyert [23] hypothesized that agent-based modeling will become increasingly important in the context of scientific policy advice.

XSL•FO
RenderX

Discussion

Simulation Models as Visual Representations

What all these models have in common is that visual representations are often constructed from them. Visual representations of data not only give orientation in times of uncertainty but also frame the ways how we experience the pandemic. From the perspective of Latour [24], diagrams and statistics function as mediators that serve to make crises visible and tangible. In this sense, it can be argued that a crisis is created by its mappability as, for example, the "flatten the curve" line chart showed by illustrating how to slow down the pandemic to prevent an overstraining of health care services.

Although on the one hand, this imageability can be seen as an elementary tool of risk and crisis communication, there are also voices that consider visualizations as hidden normative claims, as through them certain world views and power relations can be produced and reproduced: "Visualizations are not neutral windows onto data; rather, they are the result of 'judgement, discernment and choice'" [25]. In this respect, simulation models can only be interpreted as evidence-based tools for dealing with pandemics to a limited extent. In addition, this has to do with the fact that evidence in the case of simulation models cannot be measured empirically but rather follows plausibility-driven principles.

Evidence-Based Models and Simulations in the Light of Ignorance and Normativity

As indicated earlier, evidence is often referred to in the context of certain practices of constructing and modeling uncertainty. It is true that quantitative modeling and the resulting number-based outcomes provide important bases for describing existential threats and generating political pressure for action. In contrast, precisely because of their numerical orientation, number-based recommendations run the risk of failing to account for possible bias effects [26]. These problems are related to the specific nature of modeling, which is often based on linear assumptions in the process of reducing complexity.

However, the principles by which complexity is reduced often remain opaque, although the results can be significantly affected by the model assumptions. Thus, when social and political complexity is translated into specific metrics and parameters, certain information inevitably falls by the wayside. Through the distinction between relevant and irrelevant data in relation to data collection and analysis, the notion of power comes into play in the context of data-driven technologies [27]. Consequently, the effectiveness and the legitimacy of nonlinear social and political dynamics are underestimated or even ignored. Such basic assumptions, based on modeling and simulation methodology, are also evident with respect to the pandemic, where policy responses were at times guided by calculations of relatively simple models that did not differentiate between specific social characteristics—a point of critique that might pave the way for a greater demand in agent-based modeling in the future.

In the end, what makes modeling a political phenomenon is less its calculative structure than the normative and analytical premises and biases, practices, and future visions in the social construction of this technology [28]. However, in some cases, neither the developers nor the users of technology seem to be aware that the social context contradicts the idea of a purely universal and instrumental character of technology [29]. In line with this perspective, social constructivist views, in contrast to theoretical perspectives that can be attributed to technological determinism, emphasize that science and technology are neither neutral nor objective, but the result of social processes that influence and structure the production of scientific knowledge as well as the innovation of technology [30-33]. In this view, technology can be seen as "normative social hardening" [34] because norms and values as an expression of certain cultural patterns become an inherent part of the construction of technology. If the construction process is to be seen as a materialized expression of ways of looking at the world, it is important to understand to what extent the normative orientations, attitudes, and premises that flow into the design of technology can be derived from culture- and context-specific circumstances, for example, how not only values and cultural meanings but also specific interests find expression in data-driven digital health technologies.

Implications of Digital Health Simulations as Algorithmic Forms of Governance

To grasp the logic and implications of present and future pandemic technologies, it is helpful to explicate the approach of "algorithmic governance" [35] and to make it fruitful for the context of public health. The concept of "algorithmic governance" is examined primarily in its sociopolitical dimension to derive possible social and ethical implications. This is in line with the views of Hälterlein [12] that epidemiological computer simulations are to be understood as processes of algorithmic decision-making. In accordance with the 2 poles of technological determinism and social constructivism as presented earlier, different definitions of "algorithm" can be found. On the one hand, algorithms are assumed to be strictly rational affairs that combine the certainties of mathematics with the objectivity of technology [36]. In this perspective, the algorithm is understood as a "set of defined steps that, if followed in the correct order, will computationally process input (instructions and/or data) to produce a desired outcome" [37]. This definition follows an objectivist perspective that emphasizes the instrumental character of the algorithms. In this context, the formal correctness of the code is the only prerequisite to achieve desired effects. The existence of unintended side effects has been completely ignored.

In contrast, algorithms are conceptualized as "part of broader rationalities and ways of seeing the world" [38]. Roberge and Seyfert [39] shared this social constructivist-informed view. For them, algorithms are practical unfoldings and consequently are not simply reducible to code. Rather, they are to be understood as performative practices that involve a variety of rituals, narratives, and other symbolic forms of action [40]. Accordingly, algorithms are not to be understood as neutral, objective, or rational but as value laden and biased [41].

From another perspective, it follows that analytical approaches dealing with the design of algorithms must integrate all forms

XSL•FO
RenderX

of social and material practices embedded in cultural, historical, and institutional contexts [42]. Hence, the assumptions behind algorithmic logic of sorting, filtering, searching, prioritizing, recommending, and deciding can be sufficiently illuminated [38]. These insights ultimately open the debate for a critical evaluation of algorithms, more specifically, the social meaning and contextual meanings of a code as "being contingent, ontogenetic, performative in nature" [43]. Only this kind of awareness of the contingency, heterogeneity, and diversity of practices "facilitates, rather than limits, critique, making our accounts more adequate to the practices they describe, and centering those practices as a site of dispute and potential regulation" [44]. According to Seaver [44], the critical analysis and regulation of algorithm-based governance presupposes that values, norms, and their sociocultural contexts can be considered and reflected on.

The "End of Theory"

Following the view that agent-based modeling, as a subfield of computational social science, is becoming increasingly important regarding scientific policy advice, as exemplified in an impressive manner during the COVID-19 pandemic, the transition from probabilistic forms of uncertainty management to new forms of algorithmic prediction—which is particularly reflected in the rise of simulation models—is to be debated, especially because in the age of big data science, it is no longer based on testable hypotheses. This development radically points to an "end of theory" that manifests itself in a paradigmatic transformation of scientific work from causality to correlation, as algorithms find patterns that might remain hidden when classical scientific methods are applied: "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all" [45]. In this respect, algorithms have lost their innocence in the age of big data.

This paradigm shift is also characteristic of the transition from the mode of probability to the mode of possibility with regard to future developments and potential threats. Thus, the logic of anticipation is complemented by another dimension that encompasses a variety of possible projected futures [46] as can be observed within the scenario-based approaches of pandemic simulation models. Here, objectivity is attributed to algorithms, although their effects often remain opaque [47]. The problem is that the establishment of this specific form of algorithmic governance that is based on obscure mathematical correlations promotes new dimensions of intransparency in several respects.

First, even with open-source codes, an understanding of how an algorithm works is reserved for special experts, especially in the case of correlative-associative procedures and AI applications, which will probably have greater significance in future pandemic management. Second, the black box of algorithmic governance is not a box that you only need to open to see the contents undisguised. Rather, it contains other black boxes [48] that have to do with the fact that code is not to be seen as a singular construction but rather as a network of assemblages that create the output. Third, this feature of nontransparency is further exacerbated when different algorithmic technologies are combined, becoming even larger "assemblages." Following the perspective that those "assemblages are harder to be democratically controlled than more centralized forms of power" [34], a profound legitimacy problem arises.

Algorithmic Governance as an Invisible Knowledge Regime

In particular, when the data-related selection and reduction process operates in terms of an opaque network structure, this marks a considerable loss of control and legitimacy, as it cannot be traced on which premises and normative assumptions decisions are made. Therefore, if algorithmic governance is understood as an invisible knowledge regime that produces interpretations of normality and deviation on the basis of digital data, which seep deeper into social processes and interactions and take on a life of their own, this speaks for the establishment of a subtle form of power whose legitimacy must remain largely unquestioned, because "The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it" [49]. In this perspective, algorithms appear not only as neutral sorting machines but also as performative instruments of domination that legitimize power.

Crucially, the performativity of algorithms [39] marks a change in that algorithms have left their original domain of mathematical problem-solving and instead increasingly shape and transform social life worlds. Against this background, it seems even more important to deal with ethical criteria, both in the context of development and the use of digital technologies, especially for public health. However, to comprehensively reflect on the ethical implications, it is necessary to understand the formulation of algorithmic procedures not as isolated expert knowledge but to analyze them in their social dimensions and references.

Ethical Reflection Beyond Transparency and Acceptance

According to Amoore [50], algorithms and their social relations cannot be described simply by analyzing their code nor can ethics be coded into algorithms. Instead, Amoore [50] proposed to address the ethical responsibility of algorithms by identifying the sociotechnical conditions under which they emerge and function. To be able to grasp the ethicopolitical dimension of algorithms, it seems essential to confront the supposed unambiguity as an operational mode of algorithmic intrinsic logic with the multiplicity of social contexts. For Amoore [50], the error-proneness of algorithmic predictions results not from formal technical-mathematical inadequacies but from the fundamental contingency and uncertainty of human existence: "The madness of algorithms does not lie in the moral failure of the designer or the statistician but is an expression of the forms of unreason folded into a calculative rationality that reduces the multiplicity of potentials to a single output." Consequently, it cannot be a matter of making algorithmic procedures transparent to gain legitimacy.

However, the invisibility and unaccountability of algorithmic power imply that the focus on the acceptance of algorithmic governance technologies does not seem to be sufficient to

XSL•FO
RenderX

address questions of legitimacy as this is undermined both by normalization effects and by performativity. Instead, it is necessary to explicitly consider questions of acceptability. Although social acceptance refers to the fact that new technology is accepted or merely tolerated by a community, ethical acceptability refers to a conceptual reflection of the technology that takes into account the moral issues that emerge from the introduction of new technologies. In this way, for example, the contradiction that risky technology is accepted for morally wrong reasons can be critically reflected upon, which would be lost if the focus were solely on acceptance within purely empirically oriented research approaches [51]. Consequently, an ethical perspective must take special account of the nonquantifiable aspects of human life not least because this is also where the key to trust in technological applications lies [52].

Political Implications of Algorithmic Public Health Simulations

In the following section, I relate the critical perspective on algorithmic governance presented to the application context of public health simulations to shed light on the implications of mathematical modeling. This highlights the simulation of pandemic crises and thus the question of how public health management is changing by the treatment of emerging infectious diseases through simulation.

As illustrated above, although numbers "per se do not claim neutrality, truth, or scientific authority, they contribute to create realities, communities, policies and public concern" [53]. This implies, for example, that policy makers can select specific evidence to either support or prevent a lockdown [54]. From this point of view, public health simulations have the potential to contribute to the stabilization of power positions of political decision makers. In this respect, "images of epidemics and zoonoses are not mere representations of infectious diseases and their social impact, but rather actants in a broader political economic arena of power and knowledge" [55]. Consequently, the use of simulation models enables responsible emergency actors to perform within the regime of public health by referring to their capacities to govern rationally [56].

Among other things, the issue of power hierarchies raises questions about the role and functions of policy advice in times of health crises and directs attention to the discursive significance of certain forms of knowledge in relation to policy decisions. It is thus necessary to clarify whether the evidence—however, this is to be determined—is sufficient as a guide for political decisions or whether there is a danger of an "epistemization of the political" [57] if politics simply follows scientific claims to validity. Ultimately, this also raises the question of the principles on which political decisions should be based on value-based trade-offs or on the "best" scientific knowledge. As politics as an extended arm of scientific findings would practically render itself superfluous and because a hierarchization of scientific findings cannot easily be made anyway, there is much to be said for the idea that politics should not be knowledge based but value based. However, despite this insight, there were repeated voices during the pandemic arguing for a stronger scientific foundation for political decisions.

Against this background, it is important to ask what kind of scientific knowledge appears relevant. The political reactions at the beginning of the pandemic, for example, were largely characterized by a mobilization of medical and epidemiological knowledge that formed the basis for the creation of simulation models.

However, if one interprets the COVID-19 pandemic not only as a challenge in the sense of public health but also as a social crisis, many arguments can be found for considering social science knowledge in the context of more interdisciplinary expert panels and general crisis response modes. In this regard, crisis responses could benefit from an expansion of the epistemic corridors beyond natural science knowledge production [58].

When "evidence" is not necessarily unambiguous and it seems possible that forecasts deviate greatly from "reality," performative dynamics [59] can become a problem. To capture the sociotechnical dynamics of algorithmically structured public health technologies, we need an ethical perspective that goes beyond the analysis of individual acceptability and explicitly problematizes questions of social acceptability. In doing so, ethical reflection must not only refer to criteria, such as transparency or legitimacy but must, according to Amoore [50], rather keep an eye on the interrelations between society and technology to be able to unfold a holistic critical perspective.

A central question then would be which political options are and were represented in pandemic simulation models. What has not been modeled is at least as interesting. Eyert [60], for example, was able to show with regard to Germany that elementary influencing variables, such as leaving schools and daycare centers open instead of closing them down, were not integrated into any of the models. Hence, if certain political courses of action are not represented in the simulations, it highlights the power that arises from a possible selectivity of the modeled parameters. Particularly regarding the problem of accountability in algorithmic governance technologies, this question becomes more urgent when political decisions no longer appear to be attributable, and a substantial democratic deficit is derived from this with a high degree of diffusion of political responsibility.

Ethical reflection on pandemic simulations must therefore not only address the problem that evidence hardly existed at the beginning of the COVID-19 pandemic and the role of ignorance in scientific advice was too short. Rather, it must be actively reflected and debated whether what McGeoy [61] describes as "strategic ignorance" can be observed in relation to political decisions and the corresponding epistemic attempts at legitimation. In this respect, the discrimination potentials of pandemic simulations must be considered seriously if, for example, social inequalities in handling health crises are ignored. This manifests in sociostructural indicators: underprivileged status groups such as poor or health-impaired people, single parents, and cramped living conditions.

To adequately analyze the "counterproductive effects of technologies" [62] in the context of public health crises from a critical ethical perspective, a comprehensive consideration of the social construction principles and implicit premises as well as the unintended or unexpected side effects, for example with

XSL•FO
RenderX

regard to the performative and temporal dynamics of implementation in the use context, of mathematical modeling is necessary. This touches not only on issues of transparency and possible discrimination but also on the problem that social complexity cannot simply be translated into binary algorithmic codes without risking significant information loss. Only a holistic ethical reflection, which is situated on several levels and illuminates the criteria of social acceptability besides individual acceptability, can take into account the complexity of implications in the use of pandemic simulation models.

Operationalized Overview on Key Issues of Pandemic Crisis Technologies

Textbox 1 presents an operationalized overview of all key aspects and reflections addressed in this contribution. It also contains 3 major recommendations that might be relevant to present and future public health technologies.

Textbox 1. COVID-19 pandemic simulations for public health management: social construction and the implications.

Social construction of public health technologies

- What is evidence?
  - Technology is never neutral: charts and curves as quasiobjective representations of a questionable neutrality of scientific knowledge
  - Manifest and latent assumptions, values, and norms affect data collection, analysis, and interpretation
  - Social complexity cannot be transferred easily into binary code structures
- Paradoxes and ambivalences of knowledge
  - Paradoxes, eg, ignorance within pandemics versus rise of digital health technologies
  - Ambivalences of knowledge: new knowledge creates new ignorance
- Algorithmic governance and participation
  - Technology versus participation: find a balance between algorithmic governance and other forms of coping with uncertainties and crises (eg, social participation)
  - Technology and participation: participatory design as a mode of innovating public health technologies

Social, ethical, and political implications of public health technologies

- Social implications
  - Entanglements between knowledge and power, eg, within pandemic images and charts
  - Accountability problems and unintended side effects of digital health technologies
  - Performativity, eg, with regard to the mappability of a pandemic in the context of political power relations
- Ethical implications
  - Acceptance versus acceptability: distinguish between use attitudes and ethical criteria
  - Acceptability affects legitimacy and trust
  - Transparency and legitimacy not necessarily directly correlated (argument by Amoore [50])
- Political implications
  - Strategic ignorance: What scenario is modeled? But also: What is not modeled?
  - Science-policy nexus: “epistemization of the political”
  - Danger of control illusions if political decisions are merely data-based

The following were the three main recommendations:

1. Consider health crises also as social and political crises.

2. Merge crisis knowledge within interdisciplinary forms of pluralistic knowledge production: socioscientific knowledge and social participation as precious resources in reacting to health crises.

3. Be aware of the connectedness of the social construction principles and the various implications of public health technologies (especially regarding algorithmically driven practices and tools).

One key question was how the use of digital data is changing the way governments address ethical and societal questions in public health crises. Simulations were reconstructed as visual representations and sources for legitimate political power constellations. In addition, the principles of mathematical modeling based on algorithmic command structures were determined as an intransparent mode of dealing with crises and uncertainties, which relate less to individual acceptability than

XSL•FO
RenderX

to the level of acceptability. To judge the acceptability of mathematical or algorithmic modeling techniques as an ethical reflection of technologies, we must shed light on the premises, values, and norms in the social construction process of generating such models and simulations, for instance, by communicating and reflecting the assumptions. For this, it is also crucial to reflect on the role of ignorance as a problem of technocratic, data-driven crisis governance technologies.

In addition, it has been argued that alternative ways of knowledge production in times of health crises should be identified. In the context of the often-diagnosed lack of data, for example, in relation to social inequalities during the pandemic, greater importance of social science knowledge in pandemic crisis responses could be a useful and necessary complement to purely medical and epidemiological strategies for dealing with public health crises. The role of interdisciplinary work in the development and implementation of digital medical applications could also be enriched by participatory methods of technological innovation to maintain trust in technological public health solutions.

To take the criticism of technological solutionism seriously, social and more experimental forms of dealing with crises could also be debated and tested. Only in this manner can public health management escape the accusation of a technology-based top-down strategy that lacks democratic legitimacy. Considering a global preparedness regime that is able to detect public health threats at an early stage, an expansion of the technological architecture in the form of a "positive biopolitics" [63] to establish adequate structures of intervention on a planetary scale at an early stage, which should enable a higher degree of inclusion, care, and prevention, seems worthy of discussion. Against this background, this paper should not be understood as a plea against the use of technologies to address public health crises. Rather, it is a proposal to relate different forms of knowledge to each other in an interdisciplinary approach in such a way that they make the best possible use of the potentials of the heterogeneity inscribed in them and at the same time proactively consider ethical criteria of acceptability.

Limitations

Overall, it is difficult to determine the direction in which the development and implementation of pandemic crisis technologies will evolve. Therefore, it is also difficult to estimate which concrete application areas can be found for agent-based models or AI-based approaches within future simulation models, for example, and to what extent these will raise completely different implications from those outlined. New and surprising issues and challenges could soon emerge, especially as public health technologies in the aftermath of the COVID-19 pandemic represent a rapidly developing field.

Conclusions

Pandemic simulation models are an important tool to support the necessary political decision-making in crisis situations. However, their informative value not only depends strongly on the quality of the available data but also, at the same time, raises diverse implications on different levels of concern. The modes of reducing complexity within simulation models are essential, as is the question of how data quality can be optimized in the first place with regard to the modeling of social complexity, which tends to increase further. Although AI failed to exhibit its potential during the pandemic, with regard to simulations, there are indications that AI models will sooner or later become more important in the context of public health management. Thus, the ambivalences of simulation models will probably continue to be the subject of ethical reflections and sociopolitical issues in the future.

Acknowledgments

This study was produced as part of the German Federal Ministry of Education and Research–funded project "Multiple Crises. COVID-19 and the Entanglements of Public Health, Security, and Ecology in Europe."

Conflicts of Interest

None declared.

References

1. Heimstädt M, Egbert S, Esposito E. A pandemic of prediction: on the circulation of contagion models between public health and public safety. Sociologica 2021;14(3):1-24 [FREE Full text] [doi: 10.6092/issn.1971-8853/11470]
2. Bauman Z. Modernity and Ambivalence. Ithaca, NY: Cornell University Press; 1991.
3. Schwering M. Jürgen Habermas über corona: „So viel Wissen über unser Nichtwissen gab es noch nie". Frankfurter Rundschau. 2020 Apr 15. URL: http://tinyurl.com/6twnd3ab [accessed 2022-10-24]
4. Beck U, Bonß W. Die Modernisierung der Moderne. Berlin, Germany: Suhrkamp Verlag; 2001.
5. Gluckman P, Wilsdon J. From paradox to principles: where next for scientific advice to governments? Palgrave Commun 2016 Oct 18;2(1):16077. [doi: 10.1057/palcomms.2016.77]
6. Haraway D. A cyborg manifesto: science, technology, and socialist-feminism in the late 20th century. In: Weiss J, Nolan J, Hunsinger J, Trifonas P, editors. The International Handbook of Virtual Learning Environments. Dordrecht, The Netherlands: Springer; 2006.
7. Mezes C. (Nicht-)Wissen und (Un-)Sicherheit in der Pandemie. Soziopolis. 2020 May 20. URL: https://www.soziopolis.de/nicht-wissen-und-un-sicherheit-in-der-pandemie.html [accessed 2022-10-24]

XSL•FO
RenderX

8. Boin A, 'T Hart P. The crisis approach. In: Handbook of Disaster Research. Handbooks of Sociology and Social Research. New York, NY: Springer; 2007.
9. Rufat S. Critique of pure resilience. In: Reghezza-Zitt M, Rufat S, editors. Resilience Imperative: Uncertainty, Risks and Disasters. London, UK: International Society for Technology in Education; 2015.
10. Morozov E. To Save Everything, Click Here: The Folly of Technological Solutionism. New York, NY: Public Affairs; 2014.
11. Adey P, Anderson B. Anticipating emergencies: technologies of preparedness and the matter of security. Secur Dialogue 2012 Apr 16;43(2):99-117. [doi: 10.1177/0967010612438432]
12. Hälterlein J. Die Simulation der Pandemie. Soziopolis. 2020 Jun 3. URL: https://www.soziopolis.de/die-simulation-der-pandemie.html [accessed 2022-10-24]
13. Ojokoh BA, Sarumi OA, Salako KV, Gabriel AJ, Taiwo AE, Johnson OV, et al. Modeling and predicting the spread of COVID-19: a continental analysis. In: Kose U, Gupta D, de Albuquerque VH, Khanna A, editors. Data Science for COVID-19 Volume 2: Societal and Medical Perspectives. Cambridge, MA: Academic Press; 2022.
14. Postnikov EB. Estimation of COVID-19 dynamics "on a back-of-envelope": does the simplest SIR model provide quantitative parameters and predictions? Chaos Solitons Fractals 2020 Jun;135:109841 [FREE Full text] [doi: 10.1016/j.chaos.2020.109841] [Medline: 32501369]
15. Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London. 2020 Mar 16. URL: http://tinyurl.com/588c3e6f [accessed 2024-01-10]
16. Adam D. Special report: the simulations driving the world's response to COVID-19. Nature 2020 Apr 02;580(7803):316-318. [doi: 10.1038/d41586-020-01003-6] [Medline: 32242115]
17. Zhao Z, Li X, Liu F, Zhu G, Ma C, Wang L. Prediction of the COVID-19 spread in African countries and implications for prevention and control: a case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. Sci Total Environ 2020 Aug 10;729:138959. [doi: 10.1016/j.scitotenv.2020.138959]
18. Gnanvi JE, Salako KV, Kotanmi GB, Glèlè Kakaï R. On the reliability of predictions on COVID-19 dynamics: a systematic and critical review of modelling techniques. Infect Dis Model 2021;6:258-272 [FREE Full text] [doi: 10.1016/j.idm.2020.12.008] [Medline: 33458453]
19. Kaciroti NA, Lumeng C, Parekh V, Boulton ML. A Bayesian mixture model for predicting the COVID-19 related mortality in the United States. Am J Trop Med Hyg 2021 Feb 19;104(4):1484-1492 [FREE Full text] [doi: 10.4269/ajtmh.20-1147] [Medline: 33606666]
20. Willem L, Verelst F, Bilcke J, Hens N, Beutels P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). BMC Infect Dis 2017 Sep 11;17(1):612 [FREE Full text] [doi: 10.1186/s12879-017-2699-8] [Medline: 28893198]
21. Alsharhan AM. Survey of agent-based simulations for modelling COVID-19 pandemic. Adv Sci Technol Eng Syst J 2021 Mar;6(2):439-447. [doi: 10.25046/aj060250]
22. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. Chaos Solitons Fractals 2020 Oct;139:110059 [FREE Full text] [doi: 10.1016/j.chaos.2020.110059] [Medline: 32834612]
23. Eyert F. Epidemie und Modellierung. Das Mathematische ist politisch. WZB Mitteilungen. 2020 Jun. URL: https://bibliothek.wzb.eu/artikel/2020/f-23105.pdf [accessed 2022-10-24]
24. Latour B. Reassembling the Social: An Introduction to Actor-Network-Theory. New York, NY: Oxford University Press; 2005.
25. Kennedy H. Simple data visualisations have become key to communicating about the COVID-19 pandemic, but we know little about their impact. The London School of Economics and Political Science. 2020 May 4. URL: http://tinyurl.com/4s85ucrc [accessed 2022-10-24]
26. Beck S, Oomen J. Imagining the corridor of climate mitigation – what is at stake in IPCC's politics of anticipation? Environ Sci Policy 2021 Sep;123:169-178 [FREE Full text] [doi: 10.1016/j.envsci.2021.05.011]
27. Neyland D, Möllers N. Algorithmic IF … THEN rules and the conditions and consequences of power. Inf Commun Soc 2016 Mar 10;20(1):45-62 [FREE Full text] [doi: 10.1080/1369118x.2016.1156141]
28. Jasanoff S. States of Knowledge: The Co-production of Science and Social Order. Milton Park, UK: Routledge; 2004.
29. Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 2020 Aug 24;26(4):2051-2068 [FREE Full text] [doi: 10.1007/s11948-019-00146-8] [Medline: 31650511]
30. Latour B, Woolgar S. Laboratory Life: The Construction of Scientific Facts. Princeton, NJ: Princeton University Press; 1986.
31. Akrich M. The description of technical objects. In: Bijker WE, Law J, editors. Shaping Technology/Building Society: Studies in Sociotechnical Change. Cambridge, MA: The MIT Press; 1997.
32. Latour B. Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker WE, Law J, editors. Shaping Technology/Building Society: Studies in Sociotechnical Change. Cambridge, MA: MIT Press; 1992:225-258.

XSL•FO
RenderX

33. Brey PA. Values in technology and disclosive computer ethics. In: The Cambridge Handbook of Information and Computer Ethics. Cambridge, MA: Cambridge University Press; 2010:41-58.
34. Kaufmann S. Security through technology? Logic, ambivalence and paradoxes of technologised security. Eur J Secur Res 2016 Mar 15;1(1):77-95. [doi: 10.1007/s41125-016-0005-1]
35. Katzenbach C, Ulbricht L. Algorithmic governance. Internet Policy Rev 2019 Nov 9;8(4) [FREE Full text] [doi: 10.14763/2019.4.1424]
36. Seaver N. Knowing algorithms. In: digitalSTS: A Field Guide for Science & Technology Studies. Princeton, NJ: Princeton University Press; 2019.
37. Miyazaki S. Algorhythmics: understanding micro-temporality in computational cultures. Comput Cult 2012 Sep 28;2(2) [FREE Full text]
38. Beer D. The social power of algorithms. Inf Commun Soc 2016 Aug 08;20(1):1-13 [FREE Full text] [doi: 10.1080/1369118X.2016.1216147]
39. Roberge J, Seyfert R. Was sind Algorithmuskulturen? In: Algorithmuskulturen: Über die Rechnerische Konstruktion der Wirklichkeit. Bielefeld, Germany: Transcript Verlag; 2017.
40. Rauer V. Drohnen: zur materialisierung von algorithmen. In: Algorithmuskulturen: Über die Rechnerische Konstruktion der Wirklichkeit. Bielefeld, Germany: Transcript Verlag; 2017.
41. Noble SU. Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY: New York University Press; 2018.
42. Napoli PM. The algorithm as institution: toward a theoretical framework for automated media production and consumption. Fordham University Schools of Business Research Paper. 2013. URL: http://dx.doi.org/10.2139/ssrn.2260923 [accessed 2023-01-09]
43. Kitchin R. Thinking critically about and researching algorithms. Inf Commun Soc 2016 Feb 25;20(1):14-29 [FREE Full text] [doi: 10.1080/1369118x.2016.1154087]
44. Seaver N. Algorithms as culture: some tactics for the ethnography of algorithmic systems. Big Data Soc 2017 Nov 09;4(2) [FREE Full text] [doi: 10.1177/2053951717738104]
45. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired. 2008 Jun 23. URL: https://www.wired.com/2008/06/pb-theory/ [accessed 2022-10-24]
46. Amoore L. The Politics of Possibility: Risk and Security Beyond Probability. Durham, NC: Duke University Press; 2013.
47. Levermann T. Wie Algorithmen eine Kultur der Digitalität konstituieren: über die kulturelle Wirkmacht automatisierter Handlungsanweisungen in der Infosphäre. Article in German. J Korporative Kommunikation 2018;2 [FREE Full text]
48. Passig K. Fünfzig jahre black box. Article in German. Merkur 2017;823:16-30 [FREE Full text]
49. Weiser M. The computer for the 21st century. IEEE Pervasive Comput 2002 Jan;1(1):19-25. [doi: 10.1109/mprv.2002.993141]
50. Amoore L. Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. Durham, NC: Duke University Press; 2020.
51. Grunwald A. Zur Rolle von Akzeptanz und Akzeptabilität von Technik bei der Bewältigung von Technikkonflikten. Article in German. TATuP 2005 Nov 01;14(3):54-60. [doi: 10.14512/tatup.14.3.54]
52. Ammicht Quinn R. Artificial intelligence and the role of ethics. Stat J IAOS 2021 Mar 22;37(1):75-77. [doi: 10.3233/sji-210791]
53. Milan S. Techno-solutionism and the standard human in the making of the COVID-19 pandemic. Big Data Soc 2020 Oct 20;7(2). [doi: 10.1177/2053951720966781]
54. Taylor L. The price of certainty: how the politics of pandemic data demand an ethics of care. Big Data Soc 2020 Jul 22;7(2). [doi: 10.1177/2053951720942539]
55. Kelly AH, Keck F, Lynteris C. The Anthropology of Epidemics. Milton Park, UK: Taylor & Francis; 2019.
56. Keck F, Lachenal G. Simulations of epidemics: techniques of global health and neo-liberal government. In: Kelly AH, Keck F, Lynteris C, editors. The Anthropology of Epidemics. Milton Park, UK: Taylor & Francis; 2019.
57. Bogner A. Die Epistemisierung des Politischen. Wie die Macht des Wissens die Demokratie Gefährdet. Ditzingen, Germany: Reclam Verlag; 2021.
58. Bartl G, Hardt JN. Zum Verhältnis zwischen Wissenschaft und Politik im kontext multipler Krisen: Covid-19 und die Klimakrise als Herausforderungen für die wissenschaftliche Politikberatung. Zeitschrift für Umweltrecht und Umweltpolitik. 2022. URL: https://online.ruw.de/suche/pdf/zfu/zfu-02-2022-155-02d517156f740d85542e6340b143ae93.pdf [accessed 2024-01-08]
59. Esposito E, Sold K, Zimmermann B. Systems theory and algorithmic futures: interview with elena esposito. Constr Found 2021;16(3):356-361 [FREE Full text]
60. Eyert F. Agent-based modeling und politikberatung. In: Klenk T, Nullmeier F, Wewer G, editors. Handbuch Digitalisierung in Staat und Verwaltung. Wiesbaden, Germany: Springer VS; 2020.
61. McGoey L. The logic of strategic ignorance. Br J Sociol 2012 Sep;63(3):553-576 [FREE Full text] [doi: 10.1111/j.1468-4446.2012.01424.x] [Medline: 22950467]
62. Amicelle A, Aradau C, Jeandesboz J. Questioning security devices: performativity, resistance, politics. Secur Dialogue 2015 Aug 10;46(4):293-306. [doi: 10.1177/0967010615586964]

XSL•FO
RenderX

63. Bratton B. The Revenge of the Real: Politics for a Post-Pandemic World. New York, NY: Verso Books; 2021.

Abbreviations

AI: artificial intelligence
SEIR: susceptible, exposed, infectious, recovered
SIR: susceptible, infectious, recovered

Edited by D Krämer; submitted 14.01.23; peer-reviewed by A Gardner, I Mircheva; comments to author 27.02.23; revised version received 30.06.23; accepted 24.12.23; published 16.01.24.

Please cite as:
Bartl G
Social and Ethical Implications of Digital Crisis Technologies: Case Study of Pandemic Simulation Models During the COVID-19 Pandemic
J Med Internet Res 2024;26:e45723
URL: https://www.jmir.org/2024/1/e45723
doi:10.2196/45723
PMID:38227361

©Gabriel Bartl. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study

Kelly Quinn[1], MBA, MSLIS, PhD; Sarah Leiser Ransom[1], MA; Carrie O'Connell[1], MA; Naoko Muramatsu[2], PhD; David X Marquez[3], PhD; Jessie Chin[4], PhD

[1]Department of Communication, University of Illinois Chicago, Chicago, IL, United States
[2]Division of Community Health Sciences, School of Public Health, University of Illinois Chicago, Chicago, IL, United States
[3]Department of Kinesiology and Nutrition, University of Illinois Chicago, Chicago, IL, United States
[4]School of Information Sciences, University of Illinois Urbana-Champaign, Urbana, IL, United States

Corresponding Author:
Kelly Quinn, MBA, MSLIS, PhD
Department of Communication
University of Illinois Chicago
1007 W Harrison St
1140 BSB, MC 132
Chicago, IL, 60607
United States
Phone: 1 312 996 3187
Email: kquinn8@uic.edu

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2024/1/e66813

Abstract

Background: Smart speakers, such as Amazon's Echo and Google's Nest Home, combine natural language processing with a conversational interface to carry out everyday tasks, like playing music and finding information. Easy to use, they are embraced by older adults, including those with limited physical function, vision, or computer literacy. While smart speakers are increasingly used for research purposes (eg, implementing interventions and automatically recording selected research data), information on the advantages and disadvantages of using these devices for studies related to health promotion programs is limited.

Objective: This study evaluates the feasibility and acceptability of using smart speakers to deliver a physical activity (PA) program designed to help older adults enhance their physical well-being.

Methods: Community-dwelling older adults (n=18) were asked to use a custom smart speaker app to participate in an evidence-based, low-impact PA program for 10 weeks. Collected data, including measures of technology acceptance, interviews, field notes, and device logs, were analyzed using a concurrent mixed analysis approach. Technology acceptance measures were evaluated using time series ANOVAs to examine acceptability, appropriateness, feasibility, and intention to adopt smart speaker technology. Device logs provided evidence of interaction with and adoption of the device and the intervention. Interviews and field notes were thematically coded to triangulate the quantitative measures and further expand on factors relating to intervention fidelity.

Results: Smart speakers were found to be acceptable for administering a PA program, as participants reported that the devices were highly usable (mean 5.02, SE 0.38) and had strong intentions to continue their use (mean 5.90, SE 0.39). Factors such as the voice-user interface and engagement with the device on everyday tasks were identified as meaningful to acceptability. The feasibility of the devices for research activity, however, was mixed. Despite the participants rating the smart speakers as easy to use (mean 5.55, SE 1.16), functional and technical factors, such as Wi-Fi connectivity and appropriate command phrasing, required the provision of additional support resources to participants and potentially impaired intervention fidelity.

XSL•FO
RenderX

Conclusions: Smart speakers present an acceptable and appropriate behavioral intervention technology for PA programs directed at older adults but entail additional requirements for resource planning, technical support, and troubleshooting to ensure their feasibility for the research context and for fidelity of the intervention.

(J Med Internet Res 2024;26:e54800) doi:10.2196/54800

KEYWORDS

smart speakers; physical activity; older adults; behavioral health; intervention; smart device; smart devices; conversational agent; physical activities; behavioral intervention; intervention research

Introduction

Background

The use of behavioral intervention technologies (BITs) in research has proven to be feasible and efficacious in a wide variety of settings [1], extending the range of research into new geographies and populations that were previously difficult to reach and by providing new media with which to develop and deliver interventions and record data [2]. Advances in artificial intelligence and computational linguistics have created a new class of technologies that can be used for these purposes. Powered by artificial intelligence and made accessible through voice user interfaces (VUIs), smart speakers or voice-activated personal assistants, such as Amazon's Alexa and Google's Nest Home, are widely available and readily acceptable to older adults [3]. Because of their utility and features, smart speaker technologies have become a focal point in gerontological and health research [4,5]. While attention has been placed on the use of websites, software, mobile apps, and sensors as intervention delivery mechanisms [6], less attention has been placed on the use of smart speakers as a BIT, especially for older adult populations [5].

Hermes et al [7] argue that BITs hold unique characteristics that should be evaluated distinctly as part of traditional implementation outcomes, with an emphasis placed on the evaluation of BIT at the consumer or participant level for factors such as acceptability and adoption. We further argue that BITs often comprise both a delivery technology, such as a smart speaker, and an underlying software application, which is the intervention itself. By conducting independent evaluations of intervention hardware and delivery technology and intervention software applications, a more accurate evaluation of implementation outcomes can be made.

This article aims to report on the acceptability and feasibility of using smart speakers to deliver an in-home physical activity (PA) intervention among a sample of older adults aged ≥65 years. Using data collected from surveys and interviews, along with researcher field notes and device logs, we focus on the evaluation of the smart speaker device, and not the intervention application, as a BIT delivery mechanism in a 10-week pilot study that used Google Nest Home Mini smart speakers. The contribution of this study is 2-fold. First, it examines the feasibility and acceptability of smart speakers as an emerging component of BIT delivery systems, independently of an intervention assessment. Second, it examines the appropriateness of smart speaker technology for use in PA interventions for older adults.

Smart Speaker Basics

Smart speakers use a VUI to aid users in navigating everyday tasks, such as finding information, scheduling events, setting timers and alarms, and playing media [8,9]. The VUI language processing system, also called a voice assistant or conversational agent, is the defining characteristic of the smart speaker. To accomplish tasks by voice, smart speakers integrate several different technologies into a single device to leverage dialogue capabilities: these include subsystems for voice recognition, natural language processing (understanding and generation), and cloud-based data processing. Typically activated using a wake word or phrase, such as "Hey Google" or "Alexa," smart speakers remain in a state of ambient listening or are always "on," waiting for the users to initiate a conversation or command. Energy-efficient processors passively process, or "listen," for the wake word, buffering and rerecording within the device without transmitting or storing any information [10].

As illustrated in Figure 1, once a wake word is detected, the device is triggered to begin actively recording [11], transmitting recorded requests to the device maker's cloud-based service to decipher users' speech [12]. Cloud-based data processing and storage alleviates the need for the device to be capable of speech recognition [11] or file storage [13], and data are transmitted seamlessly between the device and the cloud. The audio transmission is deciphered into commands using a natural language processing algorithm, and an appropriate response is generated using speech synthesis, then sent back to the smart speaker to be conveyed to the user [12,14].

To enhance device utility, most platforms like Google and Amazon encourage personalization of the activities that can be performed on their devices. Users are urged to create user and voice profiles and to share personal information like home and work addresses, credit card numbers, calendars, account logins, transportation modes, and nicknames. This information is then used to streamline activities that can be facilitated through the device, such as purchasing items, setting calendar reminders, and generating shopping lists. In another form of customization, Google and Amazon provide developers the ability to build and market add-on applications called skills (for Amazon's Alexa) and actions (for Google Assistant), which augment available native applications. These custom actions can be designed to support research activities by recording data, supporting intervention activities, or reminding participants to pursue specific actions.

XSL•FO
RenderX

Figure 1. Illustration of smart speaker operation.

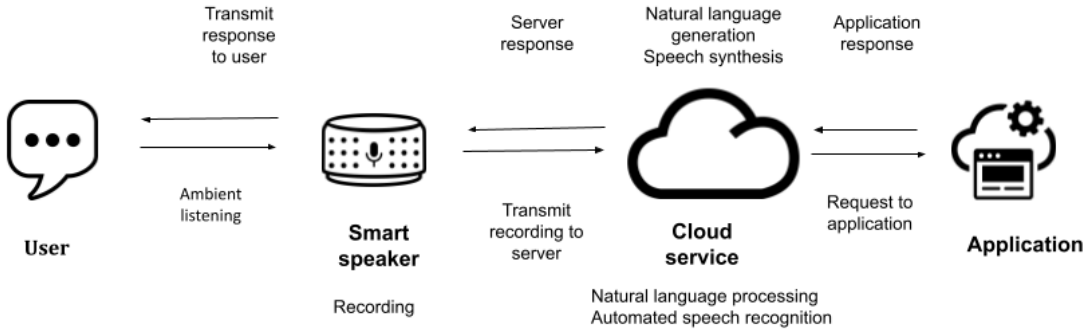

Older Adults and Technology Acceptance

Technology has become increasingly important to everyday life, yet older adults can often trail behind in the adoption of new technology because of limited experience and a lack of necessary skills [15,16]. While age-related gaps in internet use narrowed significantly since the pandemic, older adults still lag behind other age groups in both use and access to broadband connections, as compared with those in younger age groups [17,18]. In part, this may be due to experience with technology in the workplace [19], but declines in physical and cognitive abilities and limitations in the performance of instrumental activities of daily living may also lead to decreased use [20,21]. Although the use of technology by older persons can often enhance perceptions of life quality [22,23], socioeconomic factors, such as lower income and education levels, compound age-related differences [24].

VUIs represent a class of technologies that are readily accepted by older adults and perceived as easier to learn and use than keyboard interfaces [3,25,26]. Because VUIs do not rely on vision or touch, they are accessible to those with visual or fine motor degradation, mobility impairment, and disability [27]. In addition, older adult users often build companionship with VUIs, which results in positive experiences that may not only lead to reduced loneliness and increased independence [28] but also may help to overcome frustration with technological errors [29]. Processes of technology acceptance by older adults often reflect the dynamics of technology adoption and use described by the Technology Acceptance Model and its derivatives [30], including the widely accepted Unified Theory of the Acceptance and Use of Technology (UTAUT) model [31].

Older Adults and PA

The health benefits of PA for older adults are well documented [32], and its importance in supporting healthy aging cannot be overstated. PA slows age-related declines in functional abilities and helps to maintain physical and mental capacities in such diverse areas as muscle strength, cognitive functioning, disease prevention, and anxiety and depression reduction [33]. Despite these benefits, older adults report high levels of sedentary behavior [34]. Environmental contexts, such as weather, accessibility because of distance, and cost and affordability, are cited as factors by older adults for not being physically active [35]. It is often challenging for older adults to adopt and adhere to a PA regimen [36,37], and mobility challenges sometimes limit the ability to regularly participate in community-based programs because of accessibility issues [38].

Behavioral interventions to increase PA among older persons are largely successful in increasing levels of PA, and studies have shown that older individuals are more likely to continue PA programs in home-based settings [39,40]. Internet-based PA programs are both cost-efficient [41] and effective in producing behavioral change [42], and home-based programs have better adherence rates than community-based programs [40]. Technology-based interventions have been effective in producing a change in PA behaviors when compared with traditional mechanisms, such as usual care, minimal contact, waitlist control groups, in-person, or other nontechnology interventions [43]. Taken together, these factors suggest that a home-based PA program facilitated by technology, such as a smart speaker, could support the PA readiness of sedentary older adults. This study contributes to the extant literature by reporting on the acceptability and feasibility of smart speakers to deliver PA programming among a sedentary group of older adults.

Smart Speakers as a BIT

Smart speakers are emerging as a locus in behavioral intervention delivery systems [5]. Smart speakers can be deployed in participants' local environments and enable interventions to be delivered remotely, thereby reducing barriers to administration and adherence. When evaluating the use of technology in implementation research, Hermes et al [7] argue that criteria used to evaluate intervention implementations [44] should be applied to BITs and the strategies used to guide their use apart from processes used to evaluate intervention implementation, as by conducted these separately a more accurate evaluation of implementation outcomes can be made.

It is relevant to note that the Hermes et al [7] approach ignores potential distinctions between an intervention mechanism, which may be a technology, such as a software program or application through which the intervention is delivered, and its underlying hardware and delivery infrastructure, which may include a device, such as an internet-enabled watch or smart speaker through which the intervention software operates, as well as the Wi-Fi or internet signal on which it is dependent. We argue that distinctions between each of these elements are important to make when evaluating the feasibility and acceptability of a BIT as well, as challenges and successes may occur with any of the components. Moreover, it is relevant to consider a participant's ability to distinguish between these elements when undertaking this analysis to ensure appropriate identification of evaluation

XSL•FO
RenderX

criteria. For example, studies on smart speakers used for general purposes have suggested that some users fail to make distinctions between the various components of service delivery, such as the voice interface and the device [45,46], which leads to conflation between perceptions of each element.

Five criteria in particular are recommended for evaluation from the perspectives of the participant or consumer and providers or researchers [7,44]:

1. Acceptability, or the extent to which a technology is useful or satisfactory.
2. Adoption, or the intention to use the technology.
3. Appropriateness, or perception that the technology fits, is relevant or compatible with the context of its use.
4. Feasibility, or the extent to which a technology can be successfully used in a specific context.
5. Fidelity, or evidence that technology can be delivered as intended.

When evaluating a BIT at the level of the consumer or research participant, the outcomes of acceptability, feasibility, and adoption are most commonly measured through models of technology adoption [7]. The UTAUT model [31] is one of the most widely used theories of technology adoption [47], connecting the concepts of acceptability and feasibility to a third important criterion from the user's perspective: adoption. UTAUT argues that a causal relationship exists between users' perceptions of technology and their intention to use it. It specifically identifies 4 constructs—expectations of a technology's performance, the ease with which it can be used, social influence, and facilitating conditions—and links these to a user's intention to use a particular technology, which, in turn, is strongly correlated with its use. UTAUT subscales regarding performance and effort expectancy have been proven reliable with respect to the use of a wide variety of technologies in older adult populations, including email and social media [48], tablet computers [49], and remote health care or telehealth applications [50].

Studies on older adults' use of technology have found that expectations of a technology's performance and perceptions of the amount of effort that will be required to use technology are powerful incentives for new technology adoption [51,52]. Although the feasibility of using smart speakers among older adult populations to improve well-being has been examined in several recent studies [25,28,53], their use to explore specific interventions is more limited, especially among older adult populations.

Methods

Study Design

Participants were engaged in a 10-week, evidence-based, internet-based PA program that used artificial intelligence to guide activities from October 2020 through January 2022. Of note, the time frame of the study coincided with the COVID-19 pandemic and consequent lockdowns. Because all in-person leisure activities had ceased, the use of alternative delivery mechanisms for exercise programming, such as smart speakers, was potentially attractive. Though the study had been initially designed and planned to include an in-person orientation to the smart speaker and PA application, the pandemic necessitated that all participant interactions be carried out in a virtual context because of the particularly vulnerable nature of the target population. Consequently, all participant interactions, including onboarding and offboarding, weekly check-ins, and interviews, were conducted remotely via Zoom when possible and alternatively via phone calls.

A PA application was developed by the research team to run on Google Nest Home Mini speakers, which replicated components of the Healthy Moves for Aging Well program [54]. A description of the activities included in the PA application can be found in Multimedia Appendix 1. Initial feasibility and acceptance of the PA application were validated in a pilot user study before field deployment [39].

To ensure that participants had internet access, we distributed Wi-Fi hot spots together with the smart speakers. The hot spots also enabled the research team to perform the initial speaker setup and installation of the PA application before the first session. This simplified the orientation process for participants, as the devices merely needed to be connected to power for participant use.

Recruitment was aided by a partner senior-living organization located in a large suburban Midwestern county, which disseminated recruitment materials to both their independent living facility residents and via community programming information channels. The research team consisted of 4 individuals (1 principal investigator and 3 graduate research assistants, all adult women), certified by the Collaborative Institutional Training Initiative Program [55]. The study design is illustrated in Figure 2.

After enrollment, participants were sent a parcel containing written instructions regarding the use of the smart speaker, PA program application, and Wi-Fi hotspot; a smart speaker; a Wi-Fi hot spot for in-home internet connection diary materials; and consent documents. Initial participant meetings were conducted by phone or Zoom, during which baseline PA and technology attitude measures were taken, and verbal instructions were shared on using the smart speaker, PA app, and Wi-Fi hotspot. Participants were randomly assigned to a research team member who assisted with equipment setup and conducted all interviews and weekly check-ins throughout the intervention. Functionally, this meant that each researcher worked with the same 5 to 6 individuals throughout the study. At baseline (T0), participants took part in a brief individual motivational coaching session to determine their PA goals.

The intervention then took place in 2 phases. During phase 1, which lasted 6 weeks, participants were encouraged to use the smart speakers for their purposes (eg, answering questions, playing media, setting timers or alarms), as well as use the PA application for a minimum of 3 sessions per week. Participants were contacted weekly by phone by a trained member of the research team during this phase to troubleshoot, assess PA goal achievement, and set new PA goals for the coming week. Phase 2 lasted 4 weeks, during which participants continued natural use of the smart speaker and the PA application on their own (ie, without weekly contact).

XSL•FO
RenderX

Assessments of the participants' perceptions of technology acceptability, prior technology experience, and attitudes toward technology were administered by the researchers at 3 points in the study: T0, 6 weeks and end of phase 1 (T1), and 10 weeks and end of phase 2 (T2). Semistructured interviews provided perceptions and use of the smart speakers and the PA application at T1 and T2. Device logs were maintained and reviewed for the entire study period. Participants were compensated after each interview (T0, T1, and T2) and received a completion bonus for completing all visits.

Figure 2. Study design. PA: physical activity; T0: baseline; T1: 6 weeks and end of phase 1; T2: 10 weeks and end of phase 2.

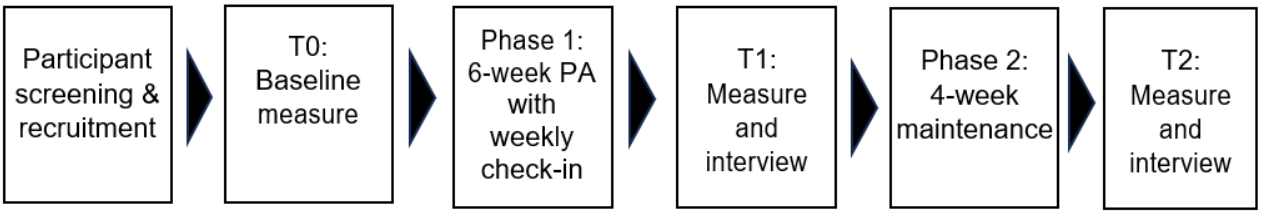

Ethical Considerations

The study protocol was approved by the institutional review board of the University of Illinois Chicago (UIC Protocol 2019-1013), which reviews and approves human subjects research in accordance with the ethical principles outlined in the Belmont Report and the DHHS regulations 45 CFR Part 46. All procedures were carried out as specified in the study protocol. All participants provided oral acknowledgment of informed consent, as a written acknowledgment requirement was waived because of the conditions of the pandemic. All data were deidentified before analysis. Participants were compensated up to US $100 for time spent and were able to keep their smart speaker device on the study's conclusion. The reporting of the qualitative findings follows the guidelines outlined in the Consolidated Criteria for Reporting Qualitative Research [56] for reporting health-related studies, as appropriate.

Data Sources

Measures of Technology Acceptability and Feasibility

Researchers administered a questionnaire with quantitative measures of familiarity with technology, UTAUT, perceived sociability, and social presence during the participant interviews at T0, T1, and T2. The UTAUT subscales consisted of 30 items related to performance expectancy, effort expectancy, attitudes toward technology, social influence, facilitating conditions, smart speaker self-efficacy, anxiety, and behavioral intention to use smart speakers, measured on a 7-point scale and adapted to the context of smart speakers [31,39]. Four items were used to measure the perceived sociability of smart speakers, and 5 items assessed the social presence of smart speakers [57]. Items for each measure are listed in Multimedia Appendix 2. To indicate technological competence in everyday living, a familiarity with technology measure was adapted from the Everyday Technology Use Questionnaire [58] regarding the frequency of use of 11 everyday technologies, such as searching the internet for information, dealing with recorded telephone menus, or sending and receiving emails [59].

Log Data

Device log data were collected from the Google accounts attached to the devices; however, data from 2 (11%) of 18 participants was missing because of technical errors. The remaining (16/18, 89%) data logs were analyzed for frequency of engagement with the PA app by calculating the number of times "Healthy Moves" was activated by the participant. To avoid fatigue, the PA program application asked participants if they needed to take a break or stop the exercise activity. Consequently, because participants were able to voluntarily truncate the application, we used its activation as a measure of engagement.

Interviews and Field Notes

A total of 36 interviews were transcribed for analysis, with an average duration of 18:05 minutes for interviews at T1 and 17:06 minutes for interviews at T2. The interview guide is presented as Multimedia Appendix 3. In addition, 90 sets of field observation notes were recorded after the weekly motivational sessions during phase 1 and examined.

Sample

Because recruitment took place during pandemic lockdowns**,** participants were recruited using social media postings, recruitment emails posted on listservs, and through the use of 2 research recruitment matchmaking portals. Inclusion criteria were those aged ≥65 years and who spoke English, as the smart speaker application was programmed in English. Participants were excluded if they participated in ≥150 minutes of PA per week or scored <7 correct responses on the Short Portable Mental Status Questionnaire [60,61].

A total of 24 participants were enrolled at T0. The size of the sample was largely determined by the availability of respondents and the adequacy of resources to complete the study. Because of interest in maintaining a PA regime and health complications, 4 participants did not complete the T1 interview. An additional 2 participants did not complete the T2 interview because of withdrawal from the study, so the final sample consisted of 18 individuals. The final study completion sample had an average age of 75.9 (SD 7.3) years; 15 (83%) of 18 participants were female, and 17 (94%) of 18 were White. Familiarity with technology was high among these participants, with 14 (78%) of 18 using a range of everyday technologies, such as searching the internet, using email, or sending text messages on a mobile device, on average, more than once per month. Table 1 summarizes the characteristics of both the study completion sample (n=18) and the recruited sample (n=24).

XSL•FO
RenderX

Table 1. Participant characteristics.

Characteristics	Study completion (n=18)	Entire recruited sample (n=24)
Age (y), mean (SD)	75.9 (7.3)	77.0 (8.0)
Gender, n (%)		
Women	15 (83)	18 (75)
Men	3 (17)	6 (25)
Racial identity, n (%)		
White	17 (94)	22 (92)
Black	1 (6)	2 (8)
Education, n (%)		
High school education or less	3 (17)	3 (13)
College degree or less	9 (50)	13 (54)
Postcollege education	6 (33)	8 (33)
Income (US $), n (%)		
≤50,000	8 (44)	11 (46)
50,000-100,000	6 (33)	7 (29)
≥100,000	4 (23)	4 (17)
Not reported	0	2 (8)
Familiarity with technology[a], n (%)		
Low (0-33)	4 (22)	8 (33)
High (34-55)	14 (78)	16 (67)

[a]Adapted from Everyday Technology Use Questionnaire [58,59].

Analysis

Data were analyzed using concurrent mixed analysis [62], with log data, interview, and field note data examined for complementarity and completeness [63] to broaden and enrich the understanding of the quantitative measures of device feasibility and acceptability. Because of this choice of method, only data collected from the study completion sample were used in this analysis. SPSS (version 29) was used for repeated measures ANOVA, and MaxQDA 2022 (version 22.08.0; VERBI GmbH) was used for first-level descriptive and second-level thematic qualitative text analysis of interview and field observation data.

The coding schema for qualitative analysis of the interview and field note date was developed using a hybrid approach, with a priori codes from the UTAUT constructs comprising the initial coding structure and additional codes developed in vivo. First, a descriptive analysis of the texts was performed by a team of 2 researchers, coding approximately 9 transcripts each. Salient words and phrases were highlighted, and special attention was paid to adjectives and adverbs that added emphasis or provided a judgment of value. Next, keywords and phrases were grouped under thematic umbrellas that either aligned with the preestablished categories derived from UTAUT codes or newly identified categories that emerged during the first descriptive coding phase. The collected measures were then grouped and analyzed according to the criteria for assessing BIT [7,44] to provide an overall assessment of acceptability, adoption, appropriateness, and feasibility. Acceptability was assessed using the UTAUT subscale of performance expectancy, along with measures of perceived sociability, pleasantness, and social presence. Adoption was measured using device log data on intervention app use and examination of the subscale on the behavioral intention to use the smart speaker. Appropriateness was evaluated using measures of attitudes toward using smart speakers and smart speaker self-efficacy. Feasibility was appraised using the UTAUT subscale of effort expectancy.

Results

As the data analyses were conducted concurrently using mixed methods, the results are also presented concurrently. In this format, log data, interview data, and field observation data offer complementarity and completeness to the quantitative measures of acceptability and feasibility.

Acceptability

Adults in this study perceived the smart speaker as acceptable, with measures of performance expectancy (ie, perceived usefulness) to be high during the study period (T0: mean 5.81, SE 0.16; T1: mean 5.04, SE 0.31; T2: mean 5.02, SE 0.38). However, perceived usefulness decreased over time ($F_{2,34}$=3.66; P=.04; partial eta squared=0.18), perhaps reflecting the practicality of actual use once participants integrated the smart speakers into their everyday routines.

XSL•FO
RenderX

Participants expanded on these perceptions in the interviews, noting that specific features of the smart speaker encouraged these perceptions of utility and spurred them to engage in more frequent and routine use of the PA program. One factor they noted was that because the smart speaker was located at home, barriers to PA engagement were reduced:

> *Convenience. I didn't have to go out of the home for the exercise.* [Participant 8]
>
> *Its accessibility and yeah, I guess that would be it. Its accessibility. It's right there and it's easy to consult and makes it easier to engage in the [PA] program.* [Participant 19]

The smart speaker also afforded participants time flexibility, enabling them to engage in the intervention at times that were convenient or available for them instead of a set or designated time. This time flexibility enabled them to adhere to the protocol with greater success:

> *I tried unsuccessfully to pick a time of day that worked the best and stick with it... In a way, it was an advantage because I could fit it in whenever it occurred to me or that I was reminded in some way.* [Participant 5]

The visual and physical presence of the smart speaker device also encouraged intervention adherence for participants, as it "reminded" participants to engage in the intervention activities, which also improved adherence:

> *I like the fact that it reminds me, just seeing it reminds me to do it, and I don't know what I would use in its stead.* [Participant 5]

Several participants noted that the VUI interface of the smart speaker added to the device's convenience by making it expedient to use:

> *I don't have to type anything in, I can just talk to it and it gives me the quick answer that I'm looking for… I think it's faster than the computer. Because it's quicker to talk than for me to type it.* [Participant 9]

Participants perceived the smart as being pleasant to interact with throughout the study period (T0: mean 5.89, SE 0.16; T1: mean 5.27, SE 0.31; T2: mean, 5.03, SE 0.38), but this perception decreased over time ($F_{2,34}$=3.63; P=.04; partial eta squared=0.18), perhaps because the novelty of interaction diminished over time. As one participant noted:

> *So I think it's a novel way of getting one to focus on an exercise program like this, and to be able to look, shall we say, look forward to doing it, more than if it were simply something you were reading from a pamphlet.* [Participant 13]

They also perceived low social presence of the device (T0: mean 3.82, SE 0.27; T1: mean 3.52, SE 0.34; T2: mean 3.31, SE 0.35), and the low social presence did not change over time ($F_{2,34}$=1.27; P=.29; partial eta squared=0.07).

Adoption

Adoption of the smart speakers, or the intention to use them, was evidenced through actual use of the devices and indication by participants of a willingness to use its functionality for other purposes in addition to the PA intervention. Examination of the device log data revealed that while heterogeneity existed in the PA program engagement, the use of the devices decreased over the study period. For the first 6 weeks (phase 1), participants engaged with the intervention app <2 times per week (mean 10.19, SD 13.26; range 0-42). Activity was higher during this initial period, perhaps because of the accountability provided by the weekly check-in calls from the research team. Engagement with the intervention app dropped off significantly during the second phase of the study, with participants engaging with the intervention app approximately about once per week (mean 4.94, SD 7.85, range 0-29; $F_{1,15}$=9.49; P=.008; partial eta squared=0.39). However, the patterns of use were heterogeneous, with some participants engaging with the device frequently and others only minimally.

About half of the participants engaged with the PA intervention less than one time per week during the intervention and follow-up phases (n=9), but other participants (n=4) were quite active, engaging with the intervention >4 times per week during phase 1, and continued engagement with intervention >2 times per week during the phase 2. The most frequent user engaged with the PA intervention every day during phase 1 and even more than once daily during phase 2. Examination of the interview data reinforced these findings and demonstrated how these 2 clusters differed. It also reinforced the understanding that as expectations of the device met user experiences, regular use of the devices was reinforced. As one participant noted:

> *Well, at first I just used it for the exercise program, and then I was a little bit more daring and listened to some music, and then jokes, and the weather, and timer and things like that. I became more comfortable using it more often.* [Participant 1]

Conversely, when participants did not use the device with any regularity for either exercise or everyday activities, they were less likely to indicate that they would continue using it at all. The lack of engagement with the smart speaker appeared to lead to a lack of adherence to the intervention protocol:

> *[B]ut I'm a little distressed with myself for not even thinking of it this week, yeah. Well it didn't seem to be too helpful to me when I was using it, I guess that's why I forgot about this week.* [Participant 2]

Participants demonstrated a strong willingness to continue using the device during and after the study period (T0: mean 6.35, SE 0.11; T1: mean 5.98, SE 0.27; T2: mean 5.90, SE 0.32), and this behavioral intention to use the device did not change over time ($F_{2,34}$=1.48; P=.24; partial eta squared=0.08). We note that participants who made a concentrated effort to incorporate the device into their daily routine during the study, for both exercise and other activities, demonstrated greater intention to continue using the device, even beyond the scope of the study.

XSL•FO
RenderX

Appropriateness

Appropriateness, or the perception that the technology is compatible with the context of its use, was demonstrated consistently during the study period, though these perceptions were not uniform among participants. Participants held strongly positive attitudes toward smart speakers (T0: mean 6.17, SE 0.09; T1: mean 6.07, SE 0.16; T2: mean 6.00, SE 0.18), and these perceptions were sustained throughout the study ($F_{2,34}$=.82; P=.45; partial eta squared=0.05). Comments about specific qualities of the smart speakers that enhanced the delivery of the PA intervention elaborated on these positive perceptions. For example, one prominent feature that was frequently mentioned by participants was the VUI. Participants found the auditory instructions provided by the smart speaker to be an appropriate mechanism to deliver the PA intervention, likening it to an individual coach or mentor:

> *Well, I do like the fact that even though this is not a real person, there is a voice telling you what to do. So you're not in a gym filled with people or you're not in a class. You don't have to go anywhere, but someone is standing there telling you what to do or sitting there telling you what to do. And so that is a, I think for persons who say live by themself or something, it is like another voice that, I guess it's watching a TV too, but it's a voice that coaches you on.* [Participant 24]

Other participants liked the ability just to listen and follow instructions. The simplicity of following commands enabled them to carry out the activities easily and with minimal cognitive effort:

> *I don't have to think about it because I just follow whatever the directions are. When we're doing the other exercise, I have to use the paper and I have to look at the paper. In here, you just follow the directions and you're all set.* [Participant 7]

However, the VUI may also have presented some hindrances for delivering the PA intervention. Prior work has identified that older adults may experience difficulty with constructing a structured sentence command in smart speaker use [29]. Participants in this study described how they had to "learn to talk to the device" by rephrasing questions and commands to obtain a desired response:

> *But every now and then on the [PA application], I think you're supposed to answer a certain way. If he [the smart speaker voice] says, "Are you ready to go to the next exercise?" Then you go "Uh-huh." And he's like, "Did you hear me?" I think maybe you should know the respect that he needs because sometimes I think he doesn't understand me and I forget to speak clearly.* [Participant 11]

Participants consistently perceived that they could effectively use the smart speakers, assessing their self-efficacy in using the smart speakers quite positively (T0: mean 6.03, SE 0.12; T1: mean 5.79, SE 0.26; T2: mean 5.58, SE 0.25). However, this self-efficacy assessment dropped during the study period ($F_{2,34}$=1.81; P=.18; partial eta squared=0.10) as participants integrated the devices into daily routines. Self-efficacy was also demonstrated by participants' ability to engage in intuitive workarounds for issues of communication with the smart speakers. These workarounds consisted of repeating or rephrasing commands, adjusting the speaking volume, or articulating the command more clearly:

> *...I don't know whether it was the way I asked, ya know, I asked the question and the device said, "I didn't understand the question." So I said it a different way and then it was able to answer.* [Participant 7]

> *Well, once in a while, I think maybe I'm not close enough to it and they will say, "I can't understand you." And I don't know whose fault that is, if it's a Mini Google device or maybe I'm not directly in front of it. But then I try again and I'm able to interact with it.* [Participant 12]

However, challenges with the ability to engage with the intervention on the smart speaker may have contributed to participants' feelings of being less proficient in using the smart speaker as the study progressed. Miscommunication issues with the device appear to be the primary hindrance and sometimes result in the participant ceasing interaction with the smart speaker or taking time away from the device before attempting interaction again. This effectively prevented the participant from carrying out the intervention at the desired or appropriate time:

> *Sometimes the Google device misses the mark as far as giving me exactly what I'm looking for. So I either have to rephrase what I'm asking or give up.* [Participant 13]

Other participants, especially those who used the device irregularly or only for the PA application, were often discouraged when faced with these difficulties. Some participants did not seem comfortable adjusting their language or speaking style to accommodate the device and would disengage from the conversation and walk away:

> *[B]ut when I tried to get onto the [PA application], I tried it twice and then both times it said it wasn't responding, but I didn't get frustrated. I just thought, "I'm not going to do exercises today."* [Participant 1]

Feasibility

Feasibility, or the perception that smart speakers are easy to use, remained high during the study period**.** Participants reported that they found the smart speakers easy to use (T0: mean 5.93, SE 0.53; T1: mean 5.56, SE 1.16; T2: mean 5.55, SE 1.24), and this ease of use was sustained throughout the study ($F_{2,34}$=1.49; P=.24; partial eta squared=0.08). As one user summarized in the interviews:

> *I found it very easy to use. Very, very helpful. It got me to exercise three times a week, more than I could have without it.* [Participant 1]

From the perspective of the research team, however, feasibility was not as clear cut. Examination of field note observations revealed that during the weekly contacts in phase 1, members

XSL•FO
RenderX

of the research team often assisted participants in effectively using their smart speakers. Though participants reported that their devices were easy to use, researchers noted that they had made suggestions that devices be moved within the participant's residence to improve the Wi-Fi signal (n=5) and that participants might rephrase commands to the device to gain an appropriate response (n=5), or worked with participants to adjust volume settings (n=2). In addition, researchers suggested and encouraged participants to use the device in alternate ways in addition to the PA program app (n=4), such as seeking information about the weather or setting timers and alarms. The frequency and nature of these observations suggest that providing support for using the devices was a necessary element in the research protocol and required resources from an administrative perspective, including training members of the research team in basic device functionality as well as potential troubleshooting strategies.

Wi-Fi connectivity issues were a major determinant of the problems that participants experienced. When smart speakers have difficulty maintaining a continuous internet connection, especially when deployed over Wi-Fi hotspots, they may encounter limits on bandwidth, which can impair their operation [64]. The use of Wi-Fi hotspots was a source of some of the connectivity issues experienced by participants and caused frustration in the form of longer-than-anticipated response times or interruptions in the performance of the device. These signals of connectivity issues required patience when experienced:

> *I would be patient. Sometimes there's a long pause before it responds. So patience is sometimes necessary, or just repeat my request. And if I get a sense that it's not being digested by the device, I rephrase it.* [Participant 13]

The placement of the smart speaker within the participant's residence was also a critical factor in ensuring successful interactions with the device. Prior research has identified that older adults strategically position smart speakers to shape and organize daily routines [65]. Participants frequently placed devices in the area of their residence, which they spent the most time in to optimize accessibility and convenience. Often, this meant that the smart speaker was placed in a living or family room, but the kitchen was also a popular option. However, these spaces were not always optimal for participating in the PA program intervention. As one participant noted:

> *Now, I'm sitting in a computer room a study. And so, a lot of times I do it [the PA program application] in here. In fact, I think I've been doing it in here for the last week or two, but sometimes I put it [the smart speaker] in the area where we eat, the kitchen eating area, so that if I'm working in the kitchen and I'd like to use it for a timer. So, then I, because it's there, I ended up doing the exercises in the kitchen.* [Participant 16]

Privacy considerations also play a role in device location and, again, may run counter to optimal placement for engagement with the intervention activity or a strong Wi-Fi signal. As one participant described:

> *My husband and I worry about the privacy issues with the Google Home Mini, and several times when I had it plugged in to prepare in order to do exercises, I would mention, "Oh, I have to Google that." And the Google Mini would come on. And that was very disconcerting because I was in another room and it was listening.* [Participant 21]

Ultimately, participants seemed to maximize the opportunity to use the smart speakers and tried to place the device in ways that would facilitate their engagement with the intervention:

> *I kept it right here in the living room where it was visible to me every day. If I had tucked it away somewhere, that would have been even more problematic on my part because it's easy to forget.* [Participant 5]

Fidelity

Fidelity, or evidence that the technology can deliver the intervention as intended, is a criterion that is perhaps best viewed from the perspective of the research team, but participants noted how the smart speaker enabled them to execute the intervention with greater precision. Participants liked the program's routine and noted that it provided structure, including PA, in their daily routines.

> *I like the fact that it times you, in other words, I don't have to keep track of the time with the clock or anything like that. I like, it just gives a little more structure so that I can rely less upon my own motivation. I would say it serves as a motivator too.* [Participant 9]

In addition, some participants noted that the smart speaker introduced accountability into their engagement with the intervention, alluding to its potential to be used as a mechanism to report on activities associated with the study:

> *And I realized that I'm on my own time. I can do this or I don't have to do it, but having the device makes me feel responsible that it's kind of like big brother is watching me, and if I don't use it, it's going to know.* [Participant 18]

Though the PA application had been tested extensively before deployment in the study, issues with its performance and interoperation with the device were encountered during both phases of the study. Participants reported that they experienced technical difficulties with the PA application, including the application not responding to them in a timely manner, aborting the intervention without warning, or repeating exercises that they had already completed. Minor modifications to the application by the development team during the study resulted in some improvements in performance, but most incidents were attributable to Wi-Fi connectivity issues, which could cause the PA program application to cease functioning and impair the fidelity of the intervention.

Participants found incidents very frustrating and for some, induced nonuse of the intervention application. As one user stated:

XSL•FO
RenderX

> *I tried to do it with the (PA application), but it doesn't work right for me, so I just don't use it then. So, I do a little exercise, but I just don't use the Google, I just do them on my own.* [Participant 20]

As this participant suggests, when encountering technical issues with the PA application, participants often maintained the exercise regimen of the study without the assistance of the smart speaker. Participants had received extensive support materials in their orientation packet, including illustrated descriptions of the activities that were invoked by the PA application. Because of these materials, several participants felt that they could complete the exercises without the guidance of the smart speakers; however, some used the functionality of the device, such as a timer, to assist in carrying out their activities:

> *[Participant] told me when she gets frustrated with that she will just use the Google Home [smart speaker] to set a timer and go through the exercises on her own.* [Field notes, Participant 16]

Participants also appeared to be able to distinguish between the PA application and the hardware of the smart speaker when they reported the challenges they encountered. This was made clear through their use of gendered pronouns when speaking about the various components, identifying the male voice of the PA application and the female voice of the smart speaker interface, as well as specific references to the PA intervention as an "app."

> *Well, there were problems with the software where it would abort and I would be shifted over to the regular Google application as opposed to the [name of PA program app].* [Participant 13]

Taken together, these last 2 points suggest that when participants' challenges with the PA application were technical, they recognized it was a limitation of the intervention delivery mechanism and not the smart speaker or the intervention. Instead, they sometimes resorted to workarounds, such as relying on their memory to complete the exercise or using the support materials as a prompt. This underscores the importance of providing support materials to participants, not only for the delivery mechanism but also for the intervention itself.

Discussion

Principal Findings

The goal of this paper was to evaluate the use of smart speakers to deliver an in-home PA intervention among a sample of older adults using previously established criteria for the evaluation of BITs. Our focus for this analysis was on the smart speaker as a delivery mechanism and not the PA intervention itself, as these are distinct components of the intervention implementation and should be evaluated separately. The smart speakers in this study were found to be highly rated for a PA program by participants regarding acceptability, appropriateness, and feasibility, criteria essential to quality BITs [7,44]. Functional and technical factors related to the operation of the smart speakers, such as ensuring consistent Wi-Fi connectivity and ensuring participants used appropriate phrasing when interacting with the devices, created a responsibility for the research team to provide basic technical support and troubleshooting resources. In addition, these same factors possess the potential to impair the fidelity of the intervention. In short, while smart speakers provide a novel and acceptable technology for intervention research, their feasibility in a research context comes with limitations.

Smart speakers afforded participants in this study convenience and flexibility for engaging with the intervention activities and served as a visual reminder to reinforce completion of the study protocol, which improved adherence to the intervention. The VUI was well-received by participants, who noted its ease of use and appropriateness for coaching participants through a PA program. The VUI also introduced challenges for the participants, as it required them to learn how to appropriately phrase commands and adjust their speaking volume to communicate with the device. This represents an intriguing intersection of possibility and limitation. On the one hand, the benefits of the device, as articulated by most respondents, are based on its ease of use and convenience, which are associated with being a hands-free interface that is capable of responding quickly and specifically. This presents a range of possibilities for intervention-based research across health, education, and other applications for older adults [66,67].

By contrast, at the current stage of development, the smart speakers used for this study are not capable of accommodating a human user's natural diction and phrasing beyond stating "I'm sorry" and requesting the user rephrase their question or direction until an acceptable rudimentary keyword or phrase is recognized. Therefore, successful accommodation to miscommunication hinges on the ability of the participant to mold their habits and language to patterns recognizable by the device. When participants were flexible about adjusting their phrasing and behaviors to mitigate technical glitches, they were also more likely to view the device favorably and use it regularly. When considering acceptability, this is particularly meaningful because the limitations of the technology and the adaptiveness of the participant base must be evaluated in tandem. Future development in artificial intelligence–supported health interventions could leverage the advancement in large language models to provide ubiquitous and fluent user experience [68].

Participants indicated their intention to use a smart speaker through positive attitudes toward its functionality, but there was a high degree of heterogeneity in their adoption. Some participants embraced the use of the devices, whereas others were frustrated and abandoned their efforts easily. Issues related to Wi-Fi connectivity were particularly challenging and interfered with the ability of the device to function appropriately. The feasibility of using smart speakers in an intervention, while positive from the participants' perspective, was more challenging from the vantage point of the research team. To provide technical and operational support for successful device operation, members of the research team were required to be familiar with device functionality and basic troubleshooting strategies. In addition, because participants had discretion over device placement within their residence, it was more challenging to ensure that the device would be optimized for both execution of the intervention and Wi-Fi connectivity. These connectivity

XSL•FO
RenderX

issues impacted intervention fidelity and underscored the need for robust support.

Overall, technical issues, such as glitches with the PA application, device-level technical problems (volume, articulation, etc), and broader critical infrastructure issues, such as a weak Wi-Fi signal, will stymie engagement with intervention activity and broader intentions to use the smart speakers. However, when participants have outlets to seek technical assistance and support when issues arise, such issues can be effectively mitigated. In other words, older adults are not at all resistant to engaging with smart speakers; however, a robust technical and informational support system should be in place.

The privacy considerations for smart speakers are not inconsequential. To be activated, smart speakers typically require an account to be established with the device maker, which is then associated with the smart speaker. Often, these accounts require additional information to be gathered from the user, such as credit card details, causing concern about personal information collection [69]. Establishing a linkage between the account and the smart speaker enables personalization of the activities that can be performed, such as reminders; however, it also associates this same information with voice recordings and device interactions [13]. Location-based data, such as time zone and zip code, allow device makers to transmit relevant information, such as weather and traffic news, but these data also become associated with accounts. All of these additional data points may increase participant privacy vulnerability.

One strategy to enhance participant privacy is to use pseudonym accounts to set up the smart speaker and collect data, as was done in this study. This approach provides some ability to shield participants' identity, but it also reduces the functionality of smart speakers considerably, as devices are unable to be personalized for calendaring and reminding functions. This reduction in utility can be frustrating to study participants, who anticipate a level of functionality frequently advertised by device makers. However, ultimately, privacy concerns are associated with smart speaker use [70], and those who use them may be qualitatively different than the general population [71], thereby presenting a form of sample selection bias for research using the devices.

Limitations

Data analyzed in this study were collected from a relatively small, not randomly controlled sample, so the conclusions reached may not be generalizable to a wider older adult population. This study was conducted during the pandemic, which required recruitment to be carried out via the web, and interviews were conducted via telephone or Zoom. The implications of this context are 2-fold. First, this sample may have been more accepting of technology and may have had higher educational attainment than if they had been enrolled in a clinical or face-to-face setting. This might suggest that the issues highlighted with this sample may be even more pronounced with populations with lower technological proficiency or education levels and may require researchers to provide additional support resources when carrying out studies of this nature. Second, the context of the pandemic may have prompted the participants to have greater acceptance of smart speaker technology to engage with PA, as in-person activities were significantly reduced during that time. These factors also limit the generalizability of these findings as they may have introduced a positive bias to perceptions of acceptability and feasibility in this sample. To gain a greater and more nuanced understanding of the acceptability and feasibility of using smart speakers in a research context, additional studies with representative samples in a nonpandemic context are required.

Challenges encountered by participants that were related to the wireless connectivity of the smart speakers may also have influenced perceptions of the smart speaker and the intervention application. Further exploratory work is needed to distinguish how perceptions of applications can be distinguished from their related delivery mechanisms for evaluation of usability, feasibility, and efficacy. In addition, data collection was constrained because of the pandemic, which limited the ability to observe participants engaging with the smart speakers. Further work is needed to expand data collection efforts beyond self-reports, which would offer additional perspectives on the acceptability of smart speakers in intervention research and provide greater detail on evaluation criteria, such as fidelity.

Finally, interviews at T1 and T2 were only conducted with participants still enrolled in the study at those time points. While data were collected from individuals who did not adhere to the intervention, it did not include input from those individuals who did not complete the study. This factor limits the interpretability of these findings, as input from noncompleters may have included perceptions related to a lack of feasibility or acceptability of smart speakers for use in a PA program.

Conclusions

In conclusion, findings from this study found smart speakers to be acceptable and appropriate for PA intervention research involving older adults, with participants indicating a willingness to adopt these delivery mechanisms for the delivery of the intervention program as well as for their everyday use. However, the feasibility of these devices for use in research contexts was mixed as they require specific and specialized attention to technical support and troubleshooting when used with older adults. Finally, applications developed to run on smart speakers must be developed to minimize disruption, whether because of flaws in design or through careful planning related to the overall Wi-Fi infrastructure, as weakness in this capacity may impair the ability of smart speakers to deliver interventions with high fidelity.

This study contributes to intervention research in that it evaluates the acceptability and feasibility of smart speakers as a behavioral intervention delivery infrastructure or the mechanisms through which an intervention is delivered, separately and distinctively from the technology that comprises an intervention, which here was a PA program application designed to enhance the physical well-being of sedentary older adults. Conducting separate evaluations of these intervention delivery elements is necessary to ensure a thorough assessment of intervention outcomes. Results from this study highlighted that older adults perceive smart speakers to be useful and easy to use. Future studies might explore the suitability of smart speakers as a delivery

XSL•FO
RenderX

infrastructure for aspects of behavioral interventions requiring smart speaker functionalities, such as the setting or reminders or the streaming of media content. Research on the use of smart speakers in other specialized populations, such as those with visual impairment or limited mobility, may also prove fruitful.

In addition, these findings offer important insight for research practitioners. At the very basic level, it cautions against oversimplifying the implications of using complex delivery infrastructures, especially with a population such as older adults that might lag the general population in the adoption and use of emerging technologies. On one level, such oversimplification may overlook important aspects of technological delivery mechanisms, such as the provision of technical support and troubleshooting, which can often tap into limited research resources. On a more granular level, the same functional and technological issues that prompt the need for support resources, such as ensuring continuous Wi-Fi connectivity, can ultimately negatively impact intervention fidelity and compromise the integrity of the research process. Taken together, smart speakers offer a novel delivery infrastructure for behavioral intervention research but also require careful planning.

Acknowledgments

This work was supported by the the National Center for Advancing Translational Sciences, National Institutes of Health [UL1TR002003] and the National Institute on Aging of the National Institutes of Health [R24AG064191, P30AG022849]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. The authors would like to thank Smit Desai for his contribution to an earlier version of the manuscript. We also would like to thank the following individuals for their assistance with this project: Smit Desai, Jamie Gradishar, Daniel Hlavacek, Faraz Hussain, Caniece Leggett, Fernando Mendez, and Melissa Pierce.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Activities of the physical activity program.
[DOCX File , 20 KB - jmir_v26i1e54800_app1.docx]

Multimedia Appendix 2
Measures of technology acceptability and feasibility.
[DOCX File , 28 KB - jmir_v26i1e54800_app2.docx]

Multimedia Appendix 3
Interview guide.
[DOCX File , 19 KB - jmir_v26i1e54800_app3.docx]

References

1. Berkowsky RW, Czaja SJ. The use of technology in behavioral intervention research: advantages and challenges. In: Behavioral Intervention Research: Designing, Evaluating, and Implementing. Berlin, Germany: Springer; Dec 2015.
2. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]
3. Portet F, Vacher M, Golanski C, Roux C, Meillon B. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers Ubiquit Comput 2011 Oct 2;17:127-144. [doi: 10.1007/s00779-011-0470-5]
4. Sunshine J. Smart speakers: the next frontier in mHealth. JMIR Mhealth Uhealth 2022 Feb 21;10(2):e28686 [FREE Full text] [doi: 10.2196/28686] [Medline: 35188467]
5. Sezgin E, Militello LK, Huang Y, Lin S. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. Transl Behav Med 2020 Aug 07;10(3):606-628. [doi: 10.1093/tbm/ibz141] [Medline: 32766865]
6. Marsch L, Lord S, Dallery J. Behavioral Health Care and Technology: Using Science-Based Innovations to Transform Practice. New York, NY: Oxford University Press; 2014.

XSL•FO
RenderX

7. Hermes ED, Lyon AR, Schueller SM, Glass JE. Measuring the implementation of behavioral intervention technologies: recharacterization of established outcomes. J Med Internet Res 2019 Jan 25;21(1):e11752 [FREE Full text] [doi: 10.2196/11752] [Medline: 30681966]
8. Hoy MB. Alexa, Siri, Cortana, and more: an introduction to voice assistants. Med Ref Serv Q 2018;37(1):81-88. [doi: 10.1080/02763869.2018.1404391] [Medline: 29327988]
9. Choi Y, Demiris G, Thompson H. Feasibility of smart speaker use to support aging in place. Innov Aging 2018 Nov 11;2(suppl_1):560. [doi: 10.1093/geroni/igy023.2073]
10. Orr DA, Sanchez L. Alexa, did you get that? Determining the evidentiary value of data stored by the Amazon® Echo. Digit Investig 2018 Mar;24:72-78. [doi: 10.1016/j.diin.2017.12.002]
11. Gray S. Always on: privacy implications of microphone-enabled devices. Future of Privacy Forum. 2016 Apr. URL: https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf [accessed 2023-11-02]
12. Gao C, Chandrasekaran V, Fawaz K, Banerjee S. Traversing the quagmire that is privacy in your smart home. In: Proceedings of the 2018 Workshop on IoT Security and Privacy. 2018 Presented at: IoT S&P '18; August 20, 2018; Budapest, Hungary. [doi: 10.1145/3229565.3229573]
13. Pfeifle A. Alexa, what should we do about privacy? Protecting privacy for users of voice-activated devices. Wash Law Rev 2018;93(1):421.
14. Amazon.com: life of an utterance. Amazon. URL: https://www.amazon.com/b/?node=23608618011 [accessed 2023-11-03]
15. Fischer SH, David D, Crotty BH, Dierks M, Safran C. Acceptance and use of health information technology by community-dwelling elders. Int J Med Inform 2014 Sep;83(9):624-635 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.005] [Medline: 24996581]
16. Hargittai E, Piper AM, Morris MR. From internet access to internet skills: digital inequality among older adults. Univ Access Inf Soc 2018 May 3;18:881-890. [doi: 10.1007/s10209-018-0617-5]
17. How popular is internet use among older people? Eurostat. 2021 May 17. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210517-1 [accessed 2024-05-22]
18. Internet, broadband fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2024-05-22]
19. Cotten SR. Technologies and aging: understanding use, impacts, and future needs. In: Handbook of Aging and the Social Sciences (Ninth Edition). Cambridge, MA: Academic Press; 2021.
20. Rikard RV, Berkowsky RW, Cotten SR. Discontinued information and communication technology usage among older adults in continuing care retirement communities in the United States. Gerontology 2018;64(2):188-200 [FREE Full text] [doi: 10.1159/000482017] [Medline: 29130976]
21. Rosenberg L, Kottorp A, Winblad B, Nygård L. Perceived difficulty in everyday technology use among older adults with or without cognitive deficits. Scand J Occup Ther 2009;16(4):216-226. [doi: 10.3109/11038120802684299] [Medline: 19148847]
22. Bobillier Chaumon ME, Michel C, Tarpin Bernard F, Croisile B. Can ICT improve the quality of life of elderly adults living in residential home care units? From actual impacts to hidden artefacts. Behav Inf Technol 2013 Oct 29;33(6):574-590. [doi: 10.1080/0144929x.2013.832382]
23. Mynatt ED, Rogers WA. Developing technology to support the functional independence of older adults. Ageing Int 2001 Dec;27:24-41. [doi: 10.1007/s12126-001-1014-5]
24. Robinson L, Cotten SR, Ono H, Quan-Haase A, Mesch G, Chen W, et al. Digital inequalities and why they matter. Inf Commun Soc 2015 Mar 16;18(5):569-582. [doi: 10.1080/1369118X.2015.1012532]
25. McCloud R, Perez C, Bekalu MA, Viswanath K. Using smart speaker technology for health and well-being in an older adult population: pre-post feasibility study. JMIR Aging 2022 May 09;5(2):e33498 [FREE Full text] [doi: 10.2196/33498] [Medline: 35532979]
26. Ziman R, Walsh G. Factors affecting seniors' perceptions of voice-enabled user interfaces. In: Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI EA '18; April 21-26, 2018; Montreal, QC. [doi: 10.1145/3170427.3188575]
27. Kowalski J, Jaskulska A, Skorupska K, Abramczuk K, Biele C, Kope? W, et al. Older adults and voice interaction: a pilot study with Google home. In: Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI EA '19; May 4-9, 2019; Glasgow, UK. [doi: 10.1145/3290607.3312973]
28. Edwards KJ, Jones RB, Shenton D, Page T, Maramba I, Warren A, et al. The use of smart speakers in care home residents: implementation study. J Med Internet Res 2021 Dec 20;23(12):e26767 [FREE Full text] [doi: 10.2196/26767] [Medline: 34932010]
29. Kim S, Choudhury A. Exploring older adults' perception and use of smart speaker-based voice assistants: a longitudinal study. Comput Hum Behav 2021 Nov;124:106914. [doi: 10.1016/j.chb.2021.106914]
30. Martín-García AV, Redolat R, Pinazo-Hernandis S. Factors influencing intention to technological use in older adults. The TAM model aplication. Res Aging 2022;44(7-8):573-588. [doi: 10.1177/01640275211063797] [Medline: 34962846]
31. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep;27(3):425-478. [doi: 10.2307/30036540]

XSL•FO
RenderX

32. Nied RJ, Franklin B. Promoting and prescribing exercise for the elderly. Am Fam Physician 2002 Feb 01;65(3):419-426 [FREE Full text] [Medline: 11858624]
33. World report on ageing and health. World Health Organization. 2015 Sep 29. URL: https://www.who.int/publications/i/item/9789241565042 [accessed 2023-11-04]
34. Harvey JA, Chastin SF, Skelton DA. Prevalence of sedentary behavior in older adults: a systematic review. Int J Environ Res Public Health 2013 Dec 02;10(12):6645-6661 [FREE Full text] [doi: 10.3390/ijerph10126645] [Medline: 24317382]
35. Spiteri K, Broom D, Bekhet AH, de Caro JX, Laventure B, Grafton K. Barriers and motivators of physical activity participation in middle-aged and older-adults - a systematic review. J Aging Phys Act 2019 Sep 01;27(4):929-944. [doi: 10.1123/japa.2018-0343] [Medline: 31141447]
36. Buman MP, Daphna Yasova L, Giacobbi PRJ. Descriptive and narrative reports of barriers and motivators to physical activity in sedentary older adults. Psychol Sport Exerc 2010 May;11(3):223-230. [doi: 10.1016/j.psychsport.2010.02.002]
37. Cress ME, Buchner DM, Prohaska T, Rimmer J, Brown M, Macera C, et al. Best practices for physical activity programs and behavior counseling in older adult populations. J Aging Phys Act 2005 Jan;13(1):61-74. [doi: 10.1123/japa.13.1.61] [Medline: 15677836]
38. Bethancourt HJ, Rosenberg DE, Beatty T, Arterburn DE. Barriers to and facilitators of physical activity program use among older adults. Clin Med Res 2014 Sep;12(1-2):10-20 [FREE Full text] [doi: 10.3121/cmr.2013.1171] [Medline: 24415748]
39. Chin J, Quinn K, Muramatsu N, Marquez D. A user study on the feasibility and acceptance of delivering physical activity programs to older adults through conversational agents. Proc Hum Factors Ergon Soc Annu Meet 2020 Dec;64(1):33-37 [FREE Full text] [doi: 10.1177/1071181320641010] [Medline: 34177216]
40. Ashworth NL, Chad KE, Harrison EL, Reeder BA, Marshall SC. Home versus center based physical activity programs in older adults. Cochrane Database Syst Rev 2005 Jan 25;2005(1):CD004017 [FREE Full text] [doi: 10.1002/14651858.CD004017.pub2] [Medline: 15674925]
41. Cobiac LJ, Vos T, Barendregt JJ. Cost-effectiveness of interventions to promote physical activity: a modelling study. PLoS Med 2009 Jul 14;6(7):e1000110 [FREE Full text] [doi: 10.1371/journal.pmed.1000110] [Medline: 19597537]
42. Wantland DJ, Portillo CJ, Holzemer WL, Slaughter R, McGhee EM. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. J Med Internet Res 2004 Nov 10;6(4):e40 [FREE Full text] [doi: 10.2196/jmir.6.4.e40] [Medline: 15631964]
43. Valle C, Tate D. Technology-based interventions to promote diet, exercise, and weight control. In: Behavioral Health Care and Technology: Using Science-Based Innovations to Transform Practice. Oxford, UK: Oxford University Press; Nov 2014.
44. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
45. Guzman AL. Voices in and of the machine: source orientation toward mobile virtual assistants. Comput Hum Behav 2019 Jan;90:343-350. [doi: 10.1016/j.chb.2018.08.009]
46. Foehr J, Germelmann CC. Alexa, can I trust you? Exploring consumer paths to trust in smart voice-interaction technologies. J Assoc Consum Res 2020 Apr;5(2):181-205. [doi: 10.1086/707731]
47. Taherdoost H. A review of technology acceptance and adoption models and theories. Procedia Manuf 2018;22:960-967. [doi: 10.1016/j.promfg.2018.03.137]
48. Bixter MT, Blocker KA, Mitzner TL, Prakash A, Rogers WA. Understanding the use and non-use of social communication technologies by older adults: a qualitative test and extension of the UTAUT model. Gerontechnology 2019 Jun;18(2):70-88 [FREE Full text] [doi: 10.4017/gt.2019.18.2.002.00] [Medline: 31754352]
49. Magsamen-Conrad K, Upadhyaya S, Joa CY, Dowd J. Bridging the divide: using UTAUT to predict multigenerational tablet adoption practices. Comput Human Behav 2015 Sep 01;50:186-196 [FREE Full text] [doi: 10.1016/j.chb.2015.03.032] [Medline: 25937699]
50. Rouidi M, Elouadi AE, Hamdoune A, Choujtani K, Chati A. TAM-UTAUT and the acceptance of remote healthcare technologies by healthcare professionals: a systematic review. Inform Med Unlocked 2022;32:101008. [doi: 10.1016/j.imu.2022.101008]
51. Moxley J, Sharit J, Czaja SJ. The factors influencing older adults' decisions surrounding adoption of technology: quantitative experimental study. JMIR Aging 2022 Nov 23;5(4):e39890 [FREE Full text] [doi: 10.2196/39890] [Medline: 36416885]
52. Berkowsky RW, Sharit J, Czaja SJ. Factors predicting decisions about technology adoption among older adults. Innov Aging 2018 Feb 21;2(1):igy002 [FREE Full text] [doi: 10.1093/geroni/igy002] [Medline: 30480129]
53. Kim D. Can healthcare apps and smart speakers improve the health behavior and depression of older adults? A quasi-experimental study. Front Digit Health 2023 Feb 23;5:1117280 [FREE Full text] [doi: 10.3389/fdgth.2023.1117280] [Medline: 36910571]
54. Yan T, Wilber KH, Wieckowski J, Simmons WJ. Results from the healthy moves for aging well program: changes of the health outcomes. Home Health Care Serv Q 2009;28(2-3):100-111. [doi: 10.1080/01621420903176136] [Medline: 20182959]
55. The trusted standard in research, ethics, compliance, and safety training. CITI Program. URL: https://about.citiprogram.org/ [accessed 2024-05-20]

XSL•FO
RenderX

56. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
57. Heerink M, Kröse BJ, Vanessa E, Wielinga BJ. The influence of social presence on acceptance of a companion robot by older people. J Phys Agents 2008 Jan;2(2):33-40. [doi: 10.14198/JoPha.2008.2.2.05]
58. Rosenberg L, Nygård L, Kottorp A. Everyday technology use questionnaire: psychometric evaluation of a new assessment of competence in technology use. Occup Ther J Res 2009;29(2):52-62. [doi: 10.3928/15394492-20090301-05]
59. Laver K, George S, Ratcliffe J, Crotty M. Measuring technology self efficacy: reliability and construct validity of a modified computer self efficacy scale in a clinical rehabilitation setting. Disabil Rehabil 2012;34(3):220-227. [doi: 10.3109/09638288.2011.593682] [Medline: 21958357]
60. Pfeiffer E. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. J Am Geriatr Soc 1975 Oct;23(10):433-441. [doi: 10.1111/j.1532-5415.1975.tb00927.x] [Medline: 1159263]
61. Roccaforte WH, Burke WJ, Bayer BL, Wengel SP. Reliability and validity of the short portable mental status questionnaire administered by telephone. J Geriatr Psychiatry Neurol 1994;7(1):33-38. [Medline: 8192828]
62. Happ MB, Dabbs AD, Tate J, Hricik A, Erlen J. Exemplars of mixed methods data combination and analysis. Nurs Res 2006;55(2 Suppl):S43-S49. [doi: 10.1097/00006199-200603001-00008] [Medline: 16601634]
63. Greene JC, Caracelli VJ, Graham WF. Toward a conceptual framework for mixed-method evaluation designs. Educ Eval Policy Anal 1989;11(3):255-274. [doi: 10.3102/01623737011003255]
64. Wi-Fi networks that are incompatible or aren't recommended. Google Nest Help. URL: https://support.google.com/googlenest/answer/9249905?hl=en [accessed 2023-11-15]
65. Chang F, Sheng L, Gu Z. Investigating the integration and the long-term use of smart speakers in older adults' daily practices: qualitative study. JMIR Mhealth Uhealth 2024 Feb 12;12:e47472 [FREE Full text] [doi: 10.2196/47472] [Medline: 38345844]
66. Desai S, Chin J. OK Google, let's learn: using voice user interfaces for informal self-regulated learning of health topics among younger and older adults. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3581507]
67. Desai S, Lundy M, Chin J. "A painless way to learn:" designing an interactive storytelling voice user interface to engage older adults in informal health information learning. In: Proceedings of the 5th International Conference on Conversational User Interfaces. 2023 Presented at: CUI '23; July 19-21, 2023; Eindhoven, The Netherlands. [doi: 10.1145/3571884.3597141]
68. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
69. Manikonda L, Deotale A, Kambhampati S. What's up with privacy?: user preferences and privacy concerns in intelligent personal assistants. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278773]
70. Edu JS, Such JM, Suarez-Tangil G. Smart home personal assistants: a security and privacy review. ACM Comput Surv 2020 Dec 06;53(6):1-36. [doi: 10.1145/3412383]
71. Lutz C, Newlands G. Privacy and smart speakers: a multi-dimensional approach. Inf Soc 2021 Mar 27;37(3):147-162. [doi: 10.1080/01972243.2021.1897914]

Abbreviations

BIT: behavioral intervention technology
PA: physical activity
UTAUT: Unified Theory of the Acceptance and Use of Technology
VUI: voice user interface

Edited by E Wethington, A Dominello, I Kronish, J Kaye; submitted 22.11.23; peer-reviewed by S Muraki, M Chatzimina, H Li; comments to author 05.04.24; revised version received 01.06.24; accepted 29.06.24; published 30.08.24.

Please cite as:
Quinn K, Leiser Ransom S, O'Connell C, Muramatsu N, Marquez DX, Chin J
Assessing the Feasibility and Acceptability of Smart Speakers in Behavioral Intervention Research With Older Adults: Mixed Methods Study
J Med Internet Res 2024;26:e54800
URL: https://www.jmir.org/2024/1/e54800
doi:10.2196/54800
PMID:

XSL•FO
RenderX

©Kelly Quinn, Sarah Leiser Ransom, Carrie O'Connell, Naoko Muramatsu, David X Marquez, Jessie Chin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ecological Momentary Intervention to Replace Sedentary Time With Physical Activity to Improve Executive Function in Midlife and Older Latino Adults: Pilot Randomized Controlled Trial

Ulf G Bronas[1,2], PhD; David X Marquez[3], PhD; Cynthia Fritschi[4], PhD; Katherine Petrarca[5], MPH; Spyros Kitsiou[3], PhD; Olu Ajilore[5], MD, PhD; Nathan Tintle[4], PhD

[1]School of Nursing, Columbia University, New York, NY, United States
[2]Division of Rehabilitation and Regenerative Medicine, Columbia University, New York, NY, United States
[3]Department of Biomedical and Health Information Sciences, University of Illinois Chicago, Chicago, IL, United States
[4]College of Nursing, University of Illinois Chicago, Chicago, IL, United States
[5]College of Medicine, University of Illinois Chicago, Chicago, IL, United States

Corresponding Author:
Ulf G Bronas, PhD
School of Nursing
Columbia University
560 W 168th St
New York, NY, 10032
United States
Phone: 1 2123055756
Email: ub2154@cumc.columbia.edu

Abstract

Background: Exercise interventions often improve moderate to vigorous physical activity, but simultaneously increase sedentary time due to a compensatory resting response. A higher level of sedentary time is associated with a lower level of executive function, while increased moderate to vigorous physical activity is associated with improved global cognition and working memory among Latino adults. Latino adults are the fastest-growing minority group in the United States and are at high risk for cognitive decline, spend more time sedentary compared to non-Hispanic populations, and engage in low levels of physical activity. Interventions that are culturally appropriate for Latino adults to replace sedentary time with physical activity are critically needed.

Objective: This study aims to develop and test the feasibility and acceptability of an ecological momentary intervention (EMI; delivered in real time) that is individually designed to replace sedentary time with physical activity in Latino adults.

Methods: This pilot study randomized 39 (n=26, 67% female; mean age 61, SD 5.8 years) community-dwelling, Spanish-speaking Latino adults (1:1 allocation) to either a 6-week EMI program designed to replace sitting time with physical activity (20/39, 51%) or physical activity guidelines education (19/39, 49%). The program was conducted on the web and in Spanish. The intervention was individualized based on individual interview responses. The intervention included the use of a Fitbit activity monitor, weekly didactic phone meetings, interactive tools (SMS text messages), and coach-delivered feedback. Feasibility and acceptability were assessed via study satisfaction (Likert scales), motivation (ecological momentary assessment), retention, and compliance. Sedentary time and physical activity were assessed via 7-day actigraphy. Cognitive performance was assessed via the trail making test part A and B (part B=executive function) and via the National Institutes of Health Toolbox remote cognitive assessment. Statistical analysis included a linear model on change score from baseline, adjusting for age, sex, and education, emphasizing effect size.

Results: Participant satisfaction with EMI was high (9.4/10), with a high degree of motivation to replace sitting time with physical activity (9.8/10). The intervention compliance rate was 79% with low difficulty using the Fitbit (1.7/10). Weekly step count increased in the intervention group by 5543 steps (group difference: d=0.54; P=.05) and sedentary time decreased by a mean 348 (SD 485) minutes (group difference: d=0.47; P=.24) compared to controls, with moderately strong effect sizes. The trail making test part B improved in the intervention group (mean –35.26, SD 60.35 seconds), compared to the control group (mean 7.19, SD 46 seconds; group difference: d=0.74; P=.01). No group differences were observed in other cognitive measures.

XSL•FO
RenderX

Conclusions: An individualized EMI designed for midlife and older Latino adults has the potential to replace sitting time with physical activity and improve executive functioning. The intervention was feasible and well received with a high degree of satisfaction.

Trial Registration: ClinicalTrials.gov NCT04507464; https://tinyurl.com/44c4thk5

(J Med Internet Res 2024;26:e55079) doi:10.2196/55079

KEYWORDS

sedentary time; physical activity; cognition; older adults; Latinos; mobile phone

Introduction

Background

The US population of adults aged ≥65 years is projected to increase from 58 million in 2021 to approximately 88 million by the year 2050 [1]. Concurrently, it is estimated that 13.8 million Americans will be diagnosed with Alzheimer disease and related dementias (ADRD) by that same year [2]. This is a significant increase from the approximately 5 million diagnosed with ADRD in 2014 [3]. The Latino/a/x/e (hereafter referred to as Latino) population within this age cohort will comprise 18% of the total US population by 2050 [4]. The growth projected within this subgroup has significant implications for the burden of ADRD, as it has been well established that rates of ADRD are disproportionately higher among older Latino adults compared to other ethnicities of the same age [1,4-6]. Currently, >12% of older Latino adults have been diagnosed with ADRD [6]. Even more striking, the number of Latino adults with ADRD is expected to increase 7-fold over the next 3 decades, representing an epidemic increase of ADRD [3]. The high incidence of ADRD in conjunction with physical function limitations places older Latino adults at a high risk for loss of independence and significant caregiver burden.

In addition, older Latino adults have a higher prevalence of risk factors for cognitive decline, and the rate of increase in risk factors is greatest among Latinos [7,8]. Latinos further engage in less leisure time physical activity and spend more time sedentary, compared to non-Latino White participants, ranging from 46.8% to 74.2% versus 34.3% to 58.2% [9,10]. Thus, increasing physical activity is of tremendous significance to this population. Of note, physical activity programs that have successfully increased daily physical activity in older Latino adults often report an increase in sedentary time [11-15]. Sedentary time, defined as not raising energy expenditure above 1.5 metabolic equivalents and includes lying down, sitting, watching television, and other screen-based entertainment activities, is a known risk factor for reduced cognitive function, especially executive function, in older Latino adults [16-19]. Thus, studies are clearly needed to investigate the benefits of breaking up and replacing sedentary time with physical activity on cognition and brain connectivity in older Latino adults [20-22]. However, to develop and deliver a successful intervention, we first need to understand how to best conduct a lifestyle intervention to replace sitting time with physical activity in the Latino community following stage 1A to 1B in the National Institutes of Health (NIH) model and assess feasibility and acceptability.

The current view of the field is rooted in a Eurocentric view, which assumes that interventions that work in one community will work the same in a different community. This assumption is likely incorrect and represents a significant gap in our ability to develop a successful intervention at the multiple levels that are required across the socioecological model (SEM) to be successful [23,24]. Thus, developing and testing the feasibility and acceptability of an individually, culturally appropriate, and community-based ecological momentary intervention (EMI; delivered in real time during participants' everyday lives) on replacing sedentary time with physical activity and cognitive performance is warranted. In addition, including participants in that latter part of midlife is tremendously important because this represents a window of opportunity for behavior change to allow for healthy aging. Due to the heterogeneity and intricacies of midlife and older Latino adults in the Chicago area and to ensure cultural validity of the EMI, we need to understand factors that facilitate initiation of and adherence to behavioral change at the individual level using the social cognitive theory [25-27].

This Study

Following the NIH model stage 1A to 1B, this study aimed to determine the feasibility and acceptability of a 6-week EMI program and explore effect sizes in the change in sedentary time, daily physical activity, and cognitive performance in midlife and older Latino adults, delivered in the home-setting over 6 weeks, compared to participants randomized to receive physical activity guidelines in Spanish. We hypothesized that the EMI program would be feasible and acceptable and that effect sizes would indicate plausibility of (1) reduced sedentary time (secondary outcome), (2) increased device-assessed daily physical activity (secondary outcome), and (3) improved cognitive performance (executive functioning and cognitive flexibility, and secondary outcome), compared to a group randomized to receive education on physical activity guidelines.

Methods

Overview

This study was a pilot of a 2-armed, parallel group randomized controlled trial. All procedures were conducted during COVID-19 restrictions between November 2020, and February 2022, remotely via telephone and videoconference. The principal investigator and assessor were blinded to group assignment. Research staff enrolled participants. Participants were randomized by the statistician using computerized variable block randomization of 2 and 4 as designed by the project statistician not involved in recruitment. All outcome assessments

XSL•FO
RenderX

were conducted 1 week before randomization and 1 week following completion of the study period. Allocation was 1:1. The study was registered with ClinicalTrials.gov (NCT04507464) and reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2010 statement: extension to randomized pilot and feasibility trials checklist [28].

Sample Size Consideration

A sample size of 12 per group is recommended for pilot studies estimating the mean and variability of continuous outcomes [29]. We, therefore, aimed to enroll 40 participants in this pilot study to ensure completion with an adequate sample size accounting for a possible 10% to 20% attrition.

Recruitment

We recruited participants through flyers posted at local Latino-based organizations and on websites, as well as using word of mouth and through media advertisements. Investigators also gave Zoom presentations at several primarily Spanish-speaking churches to recruit additional participants in the Chicago area. We assessed 69 participants for eligibility.

Inclusion criteria were (1) being aged 55-89 years, (2) having no major head trauma, and (3) owning a working smartphone with internet and a messaging plan. Exclusion criteria were participating in a supervised exercise program, being wheelchair dependent, self-reporting a diagnosis of dementia, using medications to improve cognition or mood, and having contraindications to exercise per American College of Sports Medicine guidelines [27].

Measures

Timing of Measures

Participants were first screened to determine eligibility via telephone, and those who met the criteria completed the informed e-consent (in REDCap [Research Electronic Data Capture]; Vanderbilt University) conducted via Zoom (Zoom Video Communications, Inc) video. Once consent was obtained, participants provided basic demographic information (age, sex, marital status, ethnicity, preferred language, and education) and a full medical history. We then completed the Charlson comorbidity questionnaire for each participant and conducted the following NIH Toolbox cognitive function measures as approved by the NIH for use via video call [30]: picture vocabulary test, picture sequence test, list sorting test, oral reading recognition test, auditory verbal learning test, and the auditory trail making test part A (TMT-A) and trail making test part B (TMT-B) [31-34]. TMT-B time to completion was considered the measure of executive function.

Following cognitive function testing, we conducted a video interview using Zoom or Teams (Microsoft Corp), which was recorded, transcribed, and translated for the development of an individualized, culturally tailored EMI program (see the Intervention section for details). Once participants completed the interviews, they were mailed a triaxial accelerometer (ActiGraph GT3X+, ActiGraph), which they were instructed to wear on their right hip and waist for 7 consecutive days to measure sedentary time and daily physical activity. Upon return of the ActiGraph, participants were randomized and notified of their assigned group (EMI or physical activity guideline education). The accelerometer was worn by both groups at baseline and at follow-up. Data collectors instructed the participants on the use of the accelerometer and the use of an accelerometer log.

Primary Outcome Measures

The primary outcome measures were feasibility and acceptability. Study feasibility and acceptability measures were assessed by study satisfaction debriefing at the end of the study via 10-point Likert scales. Satisfaction Likert scales were as follows: (1) How satisfied were you with the program? (0=not satisfied-10=extremely satisfied), (2) Did Fitbit motivate you to exercise? (0=not helpful-10=extremely helpful), (3) Did you enjoy the physical activity options provided by SMS text message? (0=not at all-10=very much), and (4) Did you have difficulty using Fitbit and the app? (0=no difficultly-10=very much difficulty). We further asked an open-ended question, Is there anything you would like us to know about the program that is good or that we should improve? (sample quotes are provided in Table 1). Acceptability of the program and instruments was defined as 70% (14/20) of participants rating >8/10 on study satisfaction and <2/10 on difficulties the Likert scale measures based on consensus and previous research conducted in the Bronas laboratory.

Intervention delivery, feasibility, and acceptability were measured via ecological momentary assessment (EMA) during the course of the intervention period using 10-point Likert scales. EMA assessment provided real-time feedback 3 times per day during the study following SMS text message prompts using several Likert scales: (1) Did the SMS text message motivate you to exercise? (0=not at all-10=very much), (2) Did you enjoy the SMS text message prompt? (0=not at all-10=very much), (3) Did you have difficulty with understanding the SMS text message? (0=no difficulty-10=very much difficulty), and (4) Did the SMS text message help you with the Fitbit? (0=not at all-10=very much). Acceptability and feasibility of the intervention delivery were defined as 70% of participants rating >8/10 on study satisfaction and <2/10 on difficulties in the EMA Likert scale measures based on consensus and previous research in the Bronas laboratory.

Feasibility was further defined as a participant retention rate of >80%, a compliance rate to the intervention of >70% (defined as <70% of time spent sedentary or at least 150 min of moderate physical activity per week), and a wear time of activity tracking devices >80%. Compliance was assessed using the Fitbit to assess the fidelity of treatment. All outcome measures were obtained at baseline 1 week before randomization and 1 week after completion of the study period. To analyze data obtained from the individual phone interviews, we conducted a thematic analysis to explore themes. The thematic analysis will provide a deeper understanding of physical, psychological, and social influences that influence physical activity choices in midlife and older Latino adults. Understanding these factors will be crucial for the continued refinement of effective interventions.

XSL•FO
RenderX

Table 1. Baseline demographics.

Variables	Full sample (N=39)	Intervention (n=20)	Control (n=19)	*P* value[a]
Demographics				
Age (y), mean (SD)	61 (5.85)	59.9 (4.99)	62.16 (6.57)	.24
Sex (female), n (%)	26 (67)	11 (55)	15 (79)	.17
Married, n (%)	32 (82)	19 (95)	13 (68)	*.04*[b]
Employed, n (%)	20 (51)	12 (60)	8 (42)	.35
Years education, mean (SD)	12.21 (4.53)	13.35 (4.59)	11 (4.24)	.11
BMI (kg/m^2), mean (SD)	29.7 (4.91)	29.97 (4.05)	29.41 (5.78)	.73
Previous exerciser[c], n (%)	18 (46)	10 (50)	8 (42)	.75
Medications, n (%)				
Blood pressure	11 (28)	6 (30)	5 (26)	.99
Calcium channel blocker	4 (10)	2 (10)	2 (11)	.99
β-Blocker	2 (5))	1 (5)	1 (5)	.99
Diuretics	1 (3)	0 (0)	1 (5)	.48
Blood thinners	3 (5)	1 (5)	1 (5)	.99
ACE[d] or ARB[e]	7 (18)	3 (15)	4 (21)	.70
Diabetes	8 (21)	6 (30)	2 (11)	.23
Medical history, n (%)				
CAD[f]	1 (3)	0 (0)	1 (5)	.48
Hypertension	11 (28)	5 (25)	6 (32)	.74
Diabetes	6 (15)	5 (25)	1 (5)	.19
Lung disease	7 (18)	3 (15)	4 (21)	.69

[a]*P* values from 2-sample *t* tests or chi-square tests, using a Monte Carlo simulated *P* value (10,000 replicates) to account for the small sample sizes and cell counts.

[b]Italicized values indicate *P*<.05.

[c]Meeting surgeon-general recommendation.

[d]ACE: angiotensin-converting enzyme.

[e]ARB: angiotensin receptor blocker.

[f]CAD: coronary artery disease.

Secondary Outcome Measures

The secondary outcome measures included sedentary time and cognitive function. Sedentary time, measured using the ActiGraph, was defined as an awake activity with an energy expenditure ≤1.5 metabolic equivalents occurring in a sitting or reclining posture, vector magnitude counts of <70 counts/15 seconds, and vertical axis counts of <10 counts/15 seconds. The sedentary interruption was defined as counts >100/minute. Nonwear time was considered as 60 minutes of 0-activity counts. Any counts >15,000 were considered erroneous [35]. A valid day consisted of 10 waking hours and at least 3 complete days of data (1 person) had to be available to be included in the final analysis [36]. Data were processed with ActiLife (version 6.13.5) software, with data converted to 60-second epochs. Nonwear time was defined as at least 60 consecutive minutes of 0 activity count. We categorized physical activity according to the study by Freedson et al [37] cutoff points. Cognitive function was assessed using the auditory TMT-A and TMT-B, and the NIH Toolbox measures remote assessment [30]. We captured Fitbit data remotely in real time for compliance and fidelity of treatment to the EMI program using a digital health platform (iCardia; Spyros Kitsiou; University of Illinois, Chicago) [38].

Intervention

The EMI program was developed based on the principles of the SEM, by understanding and considering participants' individual cultural interaction with their living environment [39]. We individualized and adjusted the program based on participants' stage of behavior change following the transtheoretical model using SMS text messages designed in part with participant input that included health outcomes (eg, goal setting, self-monitoring, problem-solving barriers, and increasing social support). The EMI program was also designed to provide motivational feedback of successes and encouragement via individualized SMS text messages (eg, specific goal achievement) to empower participants in their own ability to achieve behavior change,

XSL•FO
RenderX

gain positive results, and improve self-efficacy beliefs as explained by the social cognitive theory [40]. We achieved this by interviewing each participant. Participants told investigators what times during the day they wanted the SMS text message reminders and how the SMS text messages should be phrased and what they preferred to do for physical activity based on their living environment. They also provided their preference for which health benefit reminders to be sent, and they helped design these to match each individual preference. Participants further told investigators what type of feedback they preferred and when they preferred the delivery of SMS text message feedback of successes. This was conducted for all participants during the initial interview before randomization. SMS text messages were adjusted based on participant feedback. This was only in the context of sedentary behavior. Although goal achievement feedback was individualized, all EMI participants received a weekly SMS text message summarizing completed sedentary time, steps, and physical activity and encouragement of successes.

Participants in the EMI group were mailed a Fitbit Charge 4 activity monitor, and a research assistant helped them to set it up via Zoom and configure it to receive sedentary behavior notifications in the form of vibrations when a participant walked <250 steps per hour during waking hours (lowest setting available for the Fitbit; waking hours were selected by the participant). We then scheduled a video call to set up a Health Insurance Portability and Accountability Act–compliant EMA system (Ilumivu mEMA [Ilumivu Inc]) and provided education about Fitbit and the intervention to disrupt sedentary time. Through the use of the Fitbit activity tracker and SMS text messages, we were able to implement real-time delivery of the behavior options, and the Ilumivu mEMA [41] smartphone app allowed us to obtain feedback on the effectiveness of their behavior choices. Although the intervention was guided by the individual interview findings, in general, participants received suggestions on their smartphones on how to replace sitting time with physical activity, such as standing up 5 times, taking 20 steps, or even performing a short (20 s) preferred dance routine. Participants received 3 SMS text messages daily sent from the iCardia platform via the SMS text message service platform Twilio (Twilio Inc). They further received reminders on their smartphones to enter real-time feedback on activity options selected and how successful they were in adopting the option. These data allowed us to track underlying preferences for behaviors and tailor the program accordingly to each individual.

Participants received a message about the success of positive choices for behavioral action. The resulting actions and delivery success of the intervention were automatically captured and downloaded in real time using iCardia and the Ilumivu mEMA smartphone app. Participants randomized to the control group were sent and received an hour-long video education on general physical activity guidelines (questions were entertained and answered). The control group was called weekly to keep them engaged in the study, but due to the limited nature of the pilot study, they did not receive a Fitbit, and thus, this was not a true attention control group. After the completion of the 6-week intervention period, all participants were again mailed an ActiGraph to be worn on the waist for 7 days to measure sedentary time and daily physical activity. Once the ActiGraph was mailed back, a phone interview was scheduled with each participant to understand where they had difficulty and what aspects of the program did or did not work for them. The cognitive function tests administered at baseline were then repeated.

Statistical Analysis

This was a pilot randomized controlled trial for feasibility, acceptability, and to determine preliminary effect sizes. Therefore, we emphasized descriptive statistics, such as means, SDs, frequencies, percentages, and effect sizes, to demonstrate the feasibility of recruitment, compliance, retention, treatment effects over time, and proof of concept. In addition, 2-sample *t* tests and chi-square tests (using Monte Carlo simulation to obtain *P* values) were used to test for baseline demographic differences across randomized groups. When evaluating primary and secondary outcomes, paired *t* tests were used to assess within-group change from baseline to 6 weeks. Cohen *d* was used to estimate the effect of the intervention (change in EMI vs control). Linear models predicting change scores by treatment group (unadjusted model), and similar models adjusting for age, sex, and years of education were computed for each outcome measure. $P<.05$ with 2-sided tests was used to assess statistical significance, though $P<.20$ was considered a trend toward significance given the pilot nature of the study.

Ethical Considerations

This study was approved by the University of Illinois at Chicago Institutional Review Board (2020-0739), and informed consent was obtained before any study procedures.

Results

Overview

Of the 69 participants assessed for eligibility, 19 (28%) did not meet the inclusion criteria, and 50 (72%) participants met the inclusion criteria. Of the 50 participants that met the inclusion criteria, 11 (22%) participants declined participation due to disinterest and time constraints. Therefore, 39 (78%) participants were randomized to either the EMI (n=20) or the group receiving education on physical activity guidelines (control; n=19). Moreover, 1 (3%) participant was lost to follow-up after randomization, but before intervention initiation due to leaving the country permanently (CONSORT flow diagram; Figure 1) and was not included in analyses because the intervention did not begin, and we were interested primarily in the feasibility and acceptability of the intervention.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. EMI: ecological momentary intervention.

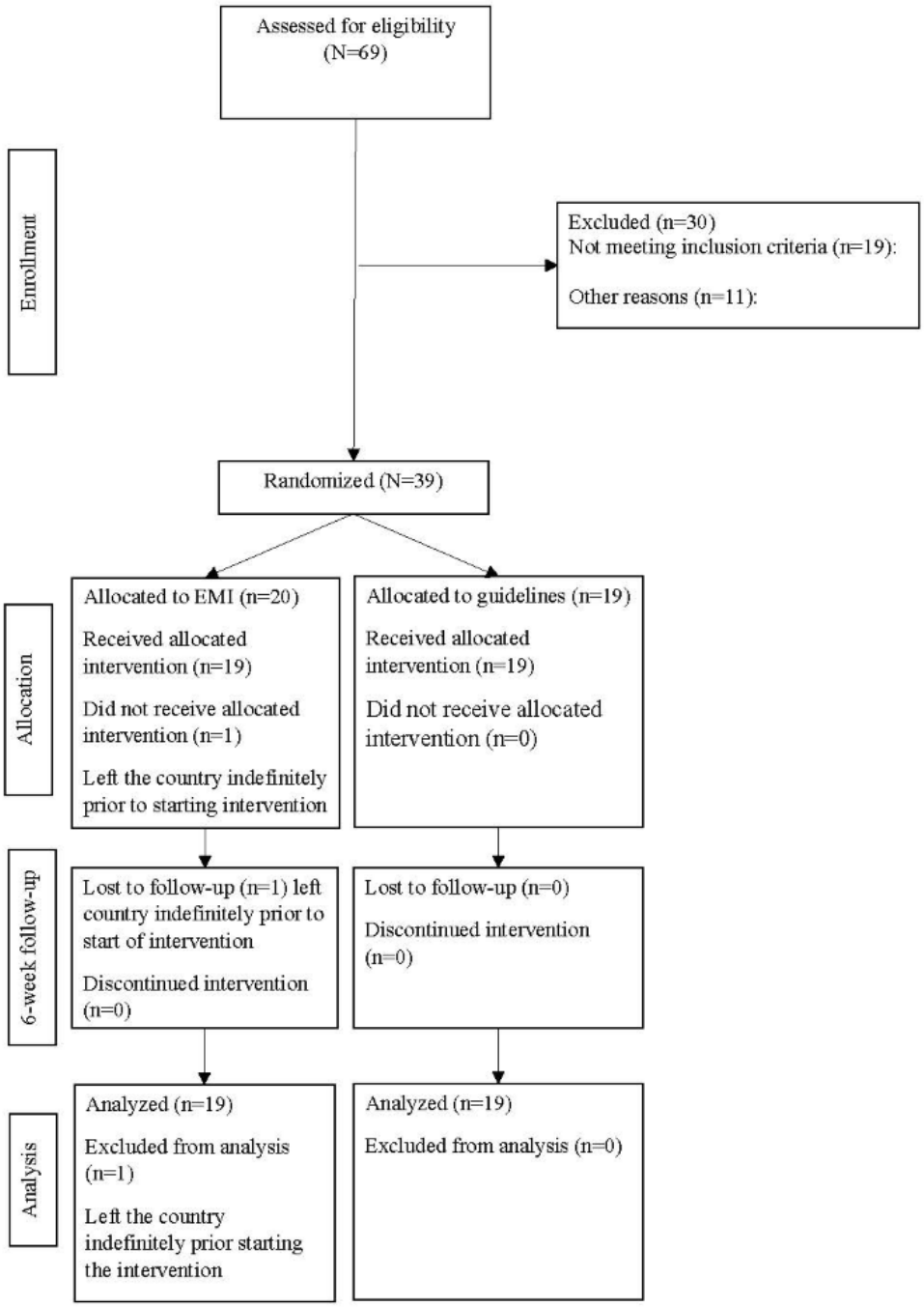

The 2 randomized groups did not show evidence of difference in baseline demographics, medication use, and medical history, except for the EMI group consisting of slightly more individuals indicating they were married (Table 1). Most (37/39, 95%) of the samples identified as Latino only with only 2 (5%) of the 39 participants identifying as Hispanic or Latino White. The sample was representative of the Chicago Latino community in terms of education albeit slightly below those published in the Chicago 2022 Community Survey [42]. When compared with the Chicago 2022 Latino Community Survey [42], 32% versus 29% had lower than a high school education, 34% versus 28% had a high school education, and 15.8% versus 20% had an associate's or bachelor's degree.

Primary Outcome: Feasibility and Acceptability

We were able to recruit a representative sample virtually during COVID-19 restrictions. Of the 69 participants accessed for eligibility, 50 (72%) met the inclusion criteria. Of the eligible 50 participants 39 (78%) were enrolled in the study indicating the feasibility of recruitment. Although we aimed to enroll 40 participants, we elected to stop enrollment at 39 (98%) participants. This was due to the COVID-19 restrictions limiting further outreach efforts and meeting criteria of 12 per group for pilot studies. The retention rate was high with only 1 (3%) loss to follow-up and no missing data on outcome measures. Feasibility and acceptability of EMI included the 19 (95%) of the 20 participants who initiated the EMI program. The participants expressed a high degree of acceptability, and study

XSL•FO
RenderX

satisfaction was high (Table 2). Compliance with the EMI program was 79% defined as <70% awake time spent sedentary or 150 minutes of moderate activity per week, and the mean minutes of activity minutes per week were 382 minutes per week captured via Fitbit. There were no adverse events associated with study procedures in either group. A total of 2850 SMS text messages were sent out. All of them were successfully delivered (100% rate). EMA survey results from 1394 survey messages over 6 weeks showed that EMI participants found the SMS text message suggestions to move to be highly motivating 82.4% of the time (8-10 on a 0-10 Likert scale). EMI participants enjoyed the text messages 84.4% (8.44/10, on a 0-10 Likert scale), and participants found that the Fitbit was motivating in conjunction with receiving SMS text message suggestions to move 90.9% of the time (9.09/10 on a Likert scale), while 92.7% found no little-to-no difficulty using the Fitbit tracker and mobile app (0-2 on a 0-10 Likert scale). Participants reported feeling highly motivated by the SMS text messages encouraging them to break up sedentary time and partake in alternative forms of exercise. When asked about the effectiveness of the program in motivating them to exercise, 1 participant reported, “This program motivated me so much because it helped me do more exercise and movement. It helped me open my mind to more exercises I can do to stay motivated.” Many participants provided similar feedback, such as the following quote: “the Fitbit motivates me to do the exercises at the time I receive the messages and the program works well to motivate the people to do exercise. I believe people need that sort of motivation.”

Table 2. Study satisfaction participants randomized to the ecologic momentary intervention[a].

Program aspect (n=19)	Participants’ perception (0-10)	Sample quote
Program satisfaction	9.4 (range: 8-10)	“I would like more people to participate in the program because it motivates people to be in movement. This program helped me to do more exercise and be in movement than before; it reminds me to be active. This program motivated me so much because it helped me do more exercise and movement. It helped me open my mind to more exercises I can do to stay motivated.”
Fitbit motivation	9.8 (range: 9-10)	“Yes, it helps us be involved more in exercises and be active at home too. With this program, you are reminded to be active, such as giving us the Fitbit watch. I loved the Fitbit watch and the motivation it gave me, especially telling me the number of steps, calories I burned throughout the day. I highly recommend this program to everyone; it helped me a lot.”
SMS text message enjoyment	9.5 (range: 8-10)	“The Fitbit motivates me to do the exercises at the time I receive the messages, and the program works well to motivate the people to do exercise. I believe people need that sort of motivation.”
Difficulty using Fitbit	1.7 (range: 0-10)[b]	“I have participated in many studies before, and this exercise study has motivated me to do more steps every day with the use of the Fitbit.”

[a]Program satisfaction (0=not satisfied-10=extremely satisfied). Motivation (0=not helpful-10=extremely helpful). Enjoy SMS text message (0=not at all-10=very much). Difficulty using the Fitbit (0=no difficulty-10=very much difficulty).

[b]One participant had significant difficulties with the Fitbit while switching mobile phones.

Secondary Outcome: Device-Assessed Sedentary Time

A variety of measures of sedentary time were obtained from ActiGraph data (Table 3). Across all 12 measures, the EMI group showed greater improvement in sedentary time as compared to the control group (fewer sedentary bouts, more sedentary breaks, and less sedentary time overall). A total of 5 (42%) of the 12 measures showed statistically significant improvement ($P<.05$) in the EMI group from baseline, with an additional 4 (33%) measures showing a positive trend ($P<.12$). Evidence of change was much less in the control group ($P>.50$ for 9, 75% of 12 measures, and $P>.13$ in all cases). Overall effect size estimates (Cohen d) comparing the 2 treatments were >0.3 for 9 (75%) of 12 measures and >0.05 in all 12 (100%) cases; however, these unadjusted (and adjusted) effects did not attain statistical significance ($P>.10$ in all cases).

XSL•FO
RenderX

Table 3. Measures of sedentary time.

Variables	Intervention (n=19)		Control (n=19)		Unadjusted Cohen *d*	*P* value for difference in effect between groups	
	Change[a], mean (SD)	Paired *t* test (2-tailed; *df*)	Change[a], mean (SD)	Paired *t* test (2-tailed; *df*)		Unadjusted	Adjusted[b]
Sedentary bouts							
Total number of sedentary bouts	–11.58 (22.02)	*0.03 (18)*[c]	–2.47 (28.06)	0.71 (19)	0.36	.27	.19
Total time spent in sedentary bouts (min)	–347.7 (484.98)	*0.006 (18)*[c]	–27.53 (811.07)	0.88 (19)	0.47	.15	.16
Average length of sedentary bout (min)	–1.87 (4.21)	0.07 (18)	0.09 (3.2)	0.9 (19)	0.51	.12	.34
Minimum length of sedentary bouts (min)	0.01 (0.11)	0.83 (18)	–0.01 (0.17)	0.79 (19)	0.14	.73	.70
Maximum length of sedentary bouts (min)	–30.26 (129.39)	0.32 (18)	11.75 (83.27)	0.55 (19)	0.38	.24	.38
Daily average of sedentary bouts (min)	–40.41 (54.59)	*0.005 (18)*[c]	–9.26 (95.55)	0.68 (19)	0.4	.23	.29
Sedentary time (min)	–297.52 (774.12)	0.11 (18)	–31.94 (950.56)	0.89 (19)	0.31	.35	.35
Sedentary breaks							
Total time spent in sedentary breaks	731.43 (2423.6)	0.20 (18)	595.62 (1928.07)	0.19 (19)	0.06	.85	.80
Mean daily length of sedentary breaks	35.02 (57.34)	*0.02 (18)*[c]	15.6 (61.59)	0.28 (19)	0.33	.32	.39
Mean daily time of sedentary breaks	187.43 (425.24)	0.07 (18)	108.33 (293.6)	0.13 (19)	0.22	.51	.81
Percent of time sedentary per day	–2.56 (6.25)	0.09 (18)	0.01 (6.55)	0.99 (19)	0.4	.22	.19

[a]Computed as 6 weeks minus baseline.

[b]Linear models on change scores adjusted for age, sex, and years of education.

[c]Italicization indicates statistical significance (*P*<.05).

Secondary Outcome: Device-Assessed Daily Physical Activity

A variety of measures of daily physical activity were obtained from ActiGraph data (Table 4). Across all 11 measures, the EMI group showed greater improvement in activity as compared to the control group (light, moderate, and vigorous activity min/d as well as percent of daily activities). A total of 3 (27%) of the 11 measures showed statistically significant improvement (*P*<.05) in the EMI group from baseline, with an additional 3 (27%) measures showing a positive trend (*P*<.15). Evidence of change was much less in the control group (*P*≥.40 for all 11 measures). Overall effect size estimates (Cohen *d*) comparing the 2 treatments were ≥0.3 for 6 (55%) of 11 measures (with 4, 36% of the remaining measures having *d*>0); however, unadjusted (and adjusted) effects did not attain statistical significance in most cases (*P*<.05).

XSL•FO
RenderX

Table 4. Measures of physical activity.

Variables	Intervention (n=19)		Control (n=19)		Unadjusted Cohen *d*	*P* value for difference in effect between groups	
	Change[a], mean (SD)	Paired *t* test (2-tailed; *df*)	Change[a], mean (SD)	Paired *t* test (2-tailed; *df*)		Unadjusted	Adjusted[b]
Minutes of activity							
Low-light (min)	19.09 (170.73)	0.63 (18)	4.9 (234)	0.93 (19)	0.07	.83	.65
Highlight (min)	8.5 (162.13)	0.82 (18)	–23.41 (185.95)	0.59 (19)	0.18	.58	.43
Low-moderate (min)	–12.6 (183.58)	0.77 (18)	–12.69 (138.62)	0.69 (19)	0	.99	.56
High-moderate (min)[c]	38.89 (109.85)	0.14 (18)	–18.05 (91.93)	0.4 (19)	0.55	.09	*.03* [d]
Vigorous (min)	25.26 (57.07)	0.07 (18)	–0.44 (24.29)	0.94 (19)	0.57	.08	.30
MVPA[e] (min)	51.56 (282.37)	0.44 (18)	–31.18 (228.03)	0.56 (19)	0.32	.33	.18
Step count	5542.79 (18,936.52)	0.22 (18)	–3748.58 (13,909.13)	0.26 (19)	0.54	.09	.05
Steps/min	1.16 (2.87)	0.09 (18)	–0.46 (2.29)	0.4 (19)	0.6	.06	*.05* [d]
Daily activity percentages							
Low-light	0.89 (1.56)	*0.02 (18)*	0.16 (2.59)	0.78 (19)	0.34	.30	.31
Highlight	0.38 (2.26)	0.47 (18)	–0.09 (2.3)	0.87 (19)	0.21	.53	.57
Low-moderate	0.04 (2.51)	0.95 (18)	0.12 (2)	0.8 (19)	0.04	.91	.72
High-moderate	0.82 (1.56)	*0.03 (18)*	–0.22 (1.41)	0.51 (19)	0.67	*.03* [d]	*.02* [d]
Vigorous	0.43 (0.84)	*0.04 (18)*	0.01 (0.41)	0.92 (19)	0.61	.06	.29

[a]Computed as 6 weeks minus baseline within groups.

[b]Linear models on change scores adjusted for age, sex, and years of education.

[c]High to moderate intensity is considered to be 5 to 6 metabolic equivalents.

[d]Italicization indicates statistical significance (*P*<.05).

[e]MVPA: moderate to vigorous physical activity.

Secondary Outcome: Cognitive Performance

Cognitive performance was assessed with the TMT-A and TMT-B as well as the NIH Toolbox, with 9 total assessed measures (Table 5). Across all 11 measures, the EMI group showed greater improvement in cognitive performance as compared to the control group. In the EMI group from baseline, 1 (11%) of the 9 measures showed statistically significant improvement (*P*<.05), with an additional 4 (44%) measures showing a positive trend (*P*<.20). Evidence of change was much less in the control group (*P*≥.40 for 8/9, 89% of measures). Overall, effect size estimates (Cohen *d*) comparing the 2 treatments were ≥0.3 for 5 (56%) of 9 measures (with the 4, 44% remaining measures having *d*>0); however, unadjusted (and adjusted) effects did not attain statistical significance in most cases (*P*<.05).

XSL•FO
RenderX

Table 5. Measures of cognitive performance.

Variables	Intervention (n=19)		Control (n=19)		Unadjusted Cohen *d*	*P* value for difference in effect between groups	
	Change[a], mean (SD)	Paired *t* test (2-tailed; *df*)	Change[a], mean (SD)	Paired *t* test (2-tailed; *df*)		Unadjusted	Adjusted[b]
Trail making measures							
Trail making test part A (s)	0.43 (1.27)	0.16 (18)	0.31 (1.73)	0.45 (19)	0.08	.81	.50
Trail making test part B (s)	–35.26 (60.35)	*0.02 (18)* [c]	7.19 (46)	0.5 (19)	0.74	*.02* [c]	*.006* [c]
Trail making test part B (errors)	–1.05 (2.8)	0.12 (18)	0.37 (2.24)	0.48 (19)	0.55	.09	*.04* [c]
NIH[d] Toolbox measures							
Picture vocabulary	0.74 (6.52)	0.63 (18)	1 (5.47)	0.44 (19)	0.04	.89	.90
List sorting	1.26 (8.7)	0.53 (18)	5.53 (9.64)	*0.022 (19)* [c]	0.46	.16	.31
Picture sequence	4.26 (13.22)	0.18 (18)	4.32 (10.73)	0.097 (19)	0.01	.99	.79
Oral reading recognition	6.89 (19.14)	0.13 (18)	1 (7.58)	0.57 (19)	0.4	.22	.27
Crystallized cognition	3.95 (10.46)	0.12 (18)	0.95 (5.26)	0.44 (19)	0.36	.27	.27
Auditory verbal	0.26 (4.04)	0.78 (18)	0.47 (5.23)	0.7 (19)	0.05	.89	.4

[a]Computed as 6 weeks minus baseline.

[b]Linear models on change scores adjusted for age, sex, and years of education.

[c]Italicization indicates statistical significance (*P*<.05).

[d]NIH: National Institutes of Health.

Exploratory Mediation Analysis

On the basis of previous data from Dr Bronas laboratory and the study by Erickson et al [43], we prespecified to explore the potential mediation of moderate to vigorous physical activity (MVPA) on executive function (measured by TMT-B). In unadjusted models, the association of intervention and change in TMT-B was 42.46 seconds (*P*=.02), and the association of intervention and change in percent MVPA was 1.04 (*P*=.04). The association of change in TMT-B and change in percent MVPA was –17.47, indicating that 1% increase in exercise was associated with a 17.47-second improvement in TMT-B (*P*=.003).

After accounting for the change in exercise, the association of intervention and change in TMT-B was 27.42 seconds (*P*=.12). This exploratory analysis provides promising results and suggests that approximately 35% of the impact of the intervention on TMT-B could be accounted for by changes in percent MVPA.

Discussion

Principal Findings

This study investigated the feasibility and acceptability of an individualized cultural EMI program for breaking up and replacing sedentary time (sitting) with physical activity on cognitive performance in older Latino adults with Spanish-speaking preference. We were able to recruit a representative sample using virtual recruitment strategies indicating that our recruitment methods were successful and that future findings will be generalizable. A high percentage (39/50, 78%) of eligible participants enrolled in the study, and we observed a low (1/39, 3%) percentage of loss to follow-up. We had no missing outcome data, and we demonstrated success in our intervention delivery with 100% of SMS text messages being received. Compliance was high, and participants indicated a high degree of satisfaction, motivation, and no difficulties with study instruments, suggesting acceptability of the study. In support of our hypotheses, the intervention program was feasible and acceptable. In this pilot, we further observed a high degree of compliance with the EMI, which showed a promising trend for decreasing sedentary time and a trend for increasing physical activity in the EMI group, compared to the physical activity guidelines group. Although underpowered, we observed moderate effect sizes indicating the plausibility of EMI decreasing sedentary time and replacing it with physical activity. This is in contrast with the study by Hartman et al [11] which reported that a physical activity intervention designed to increase MVPA also significantly increased sedentary time compared to control in Latina women (mean age 39.2, SD 10.33 years) over 6 and 12 months. These authors postulated that the increase in sedentary time was due to compensatory resting after exercise per the “activitystat” hypothesis [12,13]. Balbim et al [44] also reported a decrease in sedentary time in a health education group compared to a dance intervention group, postulated to be due to compensatory resting behavior in the dance group. Other

XSL•FO
RenderX

research has reported that physical activity and exercise interventions have little or no impact on sedentary time with 1 meta-analysis by Martin et al [14] showing a sex-specific reduction in sedentary time in males only. In addition, most previous research on sedentary time and physical activity has been conducted in non-Hispanic White populations and has not been designed to replace sedentary time with physical activity during activities of daily living. Our pilot findings may differ due to a shorter intervention period compared to that of Hartman et al [11]. We also included both male and female participants. However, sex-specific analyses did not indicate any differences in outcomes. Our sample was similar in mean age to that of Balbim et al [44], suggesting that age may not have been the contributing factor. It is possible that by focusing on replacing sedentary time with individual physical activity options in real time, we can avoid participants feeling tired from scheduled exercise, and therefore, they are resting less and spending less time being sedentary. This needs to be confirmed in an efficacy study. In addition, by using the principles of the SEM, it is possible that we influenced behavior change by using the interaction between the individual and their living environment. We also included parts of the transtheoretical model by adjusting the EMI based on the stage of behavior change and providing motivational feedback and encouragement via individualized SMS text messages. By providing motivational feedback and outcome successes, we likely improved self-efficacy beliefs as explained by the social cognitive theory; this in turn empowered participants to achieve behavior change. Unfortunately, we did not assess self-efficacy in this study, but it is likely that self-efficacy is a mediating factor, and future studies should investigate this.

Importantly, by replacing sedentary time with a mixture of low-intensity physical activity (+27.6 min), and MVPA (+51.6 min), it is possible that exercise-induced cardiometabolic risk factors decreased [45], albeit due to pandemic restrictions, we were not able to assess these. The physical activity guidelines group decreased MVPA slightly while sedentary time did not change, suggesting that activity guidelines alone will not affect behavior change. In addition, the step count increased by 5542 steps in the intervention group and decreased by 3748 steps in the physical activity guidelines group with a moderate effect size. Our findings of an increase in step count are congruent with the findings reported at 4 weeks and 12 months by Piedra et al [46] who reported an increase in 7-day pedometer step count in a supervised weekly exercise class compared to health education in older Latino adults, albeit we observed a slightly larger increase in daily step count. Of note, our findings conflict with those reported by Menkin et al [47] who found improvement in step count following an 8-session intervention to increase walking in older underrepresented adults at month 1 but not at month 2. It is likely that an 8-session intervention is not intensive enough to instill behavior change and that the use of an EMI program during routine activities of daily living may provide for a more intensive intervention to induce behavior change. Our promising findings of increased MVPA were also congruent with finding by Marquez et al [48] who reported an increase of approximately 156 minutes per week of MVPA (using self-reported physical activity; Community Healthy Activities Model Program for Seniors questionnaire) in older Latino adults randomized to an in-person dance intervention. We observed a slightly smaller increase (51.56 min), likely due to the intervention placing emphasis on replacing sedentary time with physical activity and not exercise per se and delivering the intervention during the course of participant's daily life. We also used actigraphy measures, compared to self-reported physical activity, which may explain why we observed a slightly larger total activity count using 7-day actigraphy compared to the study by Marquez et al [48]. Importantly, our compliance rate with the intervention was high (79%), suggesting that participants are open to participating in an EMI program.

Although the study was underpowered, we observed a significant increase in executive function as measured by TMT-B time to completion and a reduction in errors when completing the TMT-B in the intervention group compared to the control group. In addition, there were trends for nonsignificant improvements in oral reading recognition and crystallized cognition scores from baseline in the intervention group with modest effect sizes similar to other exercise studies in primarily White populations [43]. These findings are congruent with those of Piedra et al [49] who found improvements in the Mini-Mental State Examination at 1- and 2-year follow-up in older Latino adults randomized to receive a 4-week exercise and attribution retraining program with re-enforcement compared to health education. However, our results differ from those of Aguiñaga et al [50] who did not find an improvement in executive function using the TMT-B in older Latino adults following a 4-month dance intervention, but rather reported an improvement in working memory assessed by the digit span test. We did not observe an improvement in working memory in this study. It is possible that a 6-week intervention is not sufficient for improvement in working memory or global cognition.

These results indicate the plausibility of an individualized EMI program designed for midlife and older Latino adults to be successful in replacing sitting time with light activity and MVPA and improving executive function. These findings need to be confirmed in a larger and fully powered efficacy trial. Exploratory analyses suggest that the observed improvement in executive function could potentially be mediated via a combination of reduced sedentary time and increased MVPA. These findings are important since older Latino adults in the United States will comprise a significant proportion of the overall population in the coming decades. This, in addition to the disproportionate risk for ADRD among this group, presents the need for individually and culturally tailored interventions to promote health and reduce the risk of cognitive decline. These findings need to be assessed in a larger sample designed to test efficacy. Importantly, any physical activity intervention needs to be sustainable in the long term and in the community. Although we did not assess sustainability in this study, it is plausible that an EMI program has the potential to translate into long-term behavior change using the principles of SEM. The results associated with our outcome measures, in conjunction with findings from the participant satisfaction survey, indicate the feasibility, acceptability, and preliminary plausibility of an individualized culturally tailored EMI for improving sedentary behaviors within the Latino population.

XSL•FO
RenderX

Strengths

The strengths of this study included the randomized controlled design and a high rate of compliance associated with the intervention. This allowed for optimal delivery of motivational messages as well as reliable and consistent data collection via the iCardia platform. This was an entirely virtual intervention suggesting that we can avoid the requirement of in-person attendance, and therefore, increase our ability to reach working adults and adults that may not have the means to participate in-person. The use of the Fitbit activity tracker that is widely available means that the intervention is easily scalable. Our sample is highly representative of the Chicago Latino population, and the results are generalizable [16]. The intervention was culturally and individually tailored and delivered in real time to participants. Developing interventions that have the potential to create a sustainable behavior change in physical activity is important to gain the benefits of moving more and sitting less. Our intervention is dynamic and allows for individual and community factors to achieve sustainable behavior change.

Limitations

Though several promising results have been produced, this study is not without its limitations. The small sample size and ratio of male-to-female participants reduced the generalizability of our findings. The sample consisted of only midlife and older Latino adults from Chicago. Due to the heterogeneity and intricacies of Latino individuals in the United States, it is possible that Latino individuals in other parts of the United States differ from Latino individuals in Chicago. A multisite comparison effectiveness trial should be conducted. It is time consuming to interview individuals to design an individualized program that may hinder scalability. One plausible direction is to use machine learning to automate ways of personalization. The intervention period was short due to the proof-of-concept nature of the study, but we used alternative forms of cognitive function assessment when available to control possible learning effects. A larger and longer duration study is needed, including a sustainability period. Due to the nature of the pilot study, we were unable to provide Fitbits to the control group, which may have confounded the outcomes due to motivation. In addition, restrictions in place due to the COVID-19 pandemic prevented us from performing additional procedures that could have further validated our findings. For example, we were limited to performing cognitive testing from the NIH Toolbox that was approved for virtual administration only. However, this study would have benefitted from administering motor function testing and magnetic resonance imaging of the brain at baseline and follow-up to fully evaluate changes to functional connectivity networks. In the future, our team plans to improve the representation of the study sample by expanding to include more males. We will also seek to conduct more detailed connectivity analyses, examine the duration of treatment effects over time, and examine mediators and moderators of mechanisms of action.

In sum, this study fills a critical gap in the literature by addressing the need to determine aspects of interventional programs that show promise to fit the needs of the Latino older adult community. The EMI program was feasible and acceptable. This was accomplished through the use of an EMI designed to deliver content in real time, and in the real-world environment. The design and methods of the intervention enabled real-time delivery of behavior options and real-time feedback on the effectiveness of behavior choices. This allowed us to create an individual and culturally tailored intervention that was based on preferences of individuals within their community networks. The results presented in this paper can be used to support the design of future physical activity interventions targeted to forestall or prevent cognitive decline in this population and inform improvements in the design and implementation of meaningful interventions that are culturally appropriate and relevant to the needs of the aging Latino population in the United States. It is especially important to develop physical activity interventions that not only increase physical activity but also reduce sedentary time. These interventions need to be sustainable and targeted to disenfranchised and marginalized populations who may not speak or prefer English and who have disadvantaged socioeconomic status. The EMI program shows promise in meeting this critical need.

Conclusions

An EMI program designed to replace sedentary time with physical activity is feasible and acceptable. The effect sizes suggest plausibility that a 6-week EMI program can replace sitting time with physical activity and improve executive functioning in older Latino adults. It is plausible that the improvement in executive function is related to a combination of reduced sedentary time and increased MVPA. The clinical implication of our work remains to be further explored.

Acknowledgments

This work was supported by the National Institute on Aging of the National Institutes of Health grants P30AG022849 (UGB) and R24AG064191. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Conflicts of Interest

UGB declares grant funding from National Institute on Aging of the National Institutes of Health grants P30AG022849. OA is a co-founder of KeyWise AI and serves on the advisory boards of Otsuka, Blueprint Health, and Embodied Labs.

Multimedia Appendix 1

XSL•FO
RenderX

CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1593 KB - jmir_v26i1e55079_app1.pdf]

References

1. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023 Apr;19(4):1598-1695. [doi: 10.1002/alz.13016] [Medline: 36918389]
2. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology 2013 May 07;80(19):1778-1783 [FREE Full text] [doi: 10.1212/WNL.0b013e31828726f5] [Medline: 23390181]
3. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. Alzheimers Dement 2019 Jan;15(1):17-24 [FREE Full text] [doi: 10.1016/j.jalz.2018.06.3063] [Medline: 30243772]
4. Wu S, Vega WA, Resendez J, Jin H. Latinos and Alzheimer's disease: new numbers behind the crisis. USC Edward R. Roybal Institute on Aging and LatinosAgainstAlzheimer's Network. 2016. URL: https://roybal.usc.edu/wp-content/uploads/2016/10/Latinos-and-AD_USC_UsA2-Impact-Report.pdf [accessed 2024-08-20]
5. Quiroz YT, Solis M, Aranda MP, Arbaje AI, Arroyo-Miranda M, Cabrera LY, et al. Addressing the disparities in dementia risk, early detection and care in Latino populations: highlights from the second Latinos and Alzheimer's Symposium. Alzheimers Dement 2022 Sep;18(9):1677-1686 [FREE Full text] [doi: 10.1002/alz.12589] [Medline: 35199931]
6. Mehta KM, Yeo GW. Systematic review of dementia prevalence and incidence in United States race/ethnic populations. Alzheimers Dement 2017 Jan;13(1):72-83. [doi: 10.1016/j.jalz.2016.06.2360] [Medline: 27599209]
7. Chen C, Zissimopoulos JM. Racial and ethnic differences in trends in dementia prevalence and risk factors in the United States. Alzheimers Dement (N Y) 2018 Oct 05;4:510-520 [FREE Full text] [doi: 10.1016/j.trci.2018.08.009] [Medline: 30364652]
8. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. Circulation 2017 Mar 07;135(10):e146-e603 [FREE Full text] [doi: 10.1161/CIR.0000000000000485] [Medline: 28122885]
9. Keadle SK, McKinnon R, Graubard BI, Troiano RP. Prevalence and trends in physical activity among older adults in the United States: a comparison across three national surveys. Prev Med 2016 Aug;89:37-43 [FREE Full text] [doi: 10.1016/j.ypmed.2016.05.009] [Medline: 27196146]
10. Norris T, Schiller JS, Clarke TC. Early release of selected estimates based on data from the national health interview survey. National Center for Health Statistics. 2018 Jun. URL: https://www.cdc.gov/nchs/data/nhis/earlyrelease/earlyrelease201806_tech.pdf [accessed 2024-08-20]
11. Hartman SJ, Pekmezi D, Dunsiger SI, Marcus BH. Physical activity intervention effects on sedentary time in Spanish-speaking Latinas. J Phys Act Health 2020 Mar 01;17(3):343-348 [FREE Full text] [doi: 10.1123/jpah.2019-0112] [Medline: 32035412]
12. Rowland TW. The biological basis of physical activity. Med Sci Sports Exerc 1998 Mar;30(3):392-399. [doi: 10.1097/00005768-199803000-00009] [Medline: 9526885]
13. Gray P, Murphy M, Gallagher A, Simpson EE. A qualitative investigation of physical activity compensation among older adults. Br J Health Psychol 2018 Feb;23(1):208-224. [doi: 10.1111/bjhp.12282] [Medline: 29171704]
14. Martin A, Fitzsimons C, Jepson R, Saunders DH, van der Ploeg HP, Teixeira PJ, et al. Interventions with potential to reduce sedentary time in adults: systematic review and meta-analysis. Br J Sports Med 2015 Aug;49(16):1056-1063. [doi: 10.1136/bjsports-2014-094524] [Medline: 25907181]
15. Doucet É, McInis K, Mahmoodianfard S. Compensation in response to energy deficits induced by exercise or diet. Obes Rev 2018 Dec;19 Suppl 1:36-46. [doi: 10.1111/obr.12783] [Medline: 30511511]
16. Arredondo EM, Sotres-Alvarez D, Stoutenberg M, Davis SM, Crespo NC, Carnethon MR, et al. Physical activity levels in U.S. Latino/Hispanic adults: results from the Hispanic community health study/study of Latinos. Am J Prev Med 2016 Apr;50(4):500-508 [FREE Full text] [doi: 10.1016/j.amepre.2015.08.029] [Medline: 26597505]
17. Vásquez E, Strizich G, Isasi CR, Echeverria SE, Sotres-Alvarez D, Evenson KR, et al. Is there a relationship between accelerometer-assessed physical activity and sedentary behavior and cognitive function in US Hispanic/Latino adults? The Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Prev Med 2017 Oct;103:43-48 [FREE Full text] [doi: 10.1016/j.ypmed.2017.07.024] [Medline: 28765082]
18. Diaz KM, Howard VJ, Hutto B, Colabianchi N, Vena JE, Safford MM, et al. Patterns of sedentary behavior and mortality in U.S. middle-aged and older adults: a national cohort study. Ann Intern Med 2017 Oct 03;167(7):465-475 [FREE Full text] [doi: 10.7326/M17-0212] [Medline: 28892811]
19. Wheeler MJ, Dempsey PC, Grace MS, Ellis KA, Gardiner PA, Green DJ, et al. Sedentary behavior as a risk factor for cognitive decline? A focus on the influence of glycemic control in brain health. Alzheimers Dement (N Y) 2017 May 02;3(3):291-300 [FREE Full text] [doi: 10.1016/j.trci.2017.04.001] [Medline: 29067335]
20. Bronas UG, Steffen A, Dion C, Boots EA, Arfanakis K, Marquez DX, et al. Sedentary time and white matter hyperintensity volume in older adults. Med Sci Sports Exerc 2019 Aug;51(8):1613-1618 [FREE Full text] [doi: 10.1249/MSS.0000000000001957] [Medline: 30817720]

XSL•FO
RenderX

21. Wardlaw JM, Valdés Hernández MC, Muñoz-Maniega S. What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. J Am Heart Assoc 2015 Jun 23;4(6):001140 [FREE Full text] [doi: 10.1161/JAHA.114.001140] [Medline: 26104658]
22. Boyle PA, Yu L, Fleischman DA, Leurgans S, Yang J, Wilson RS, et al. White matter hyperintensities, incident mild cognitive impairment, and cognitive decline in old age. Ann Clin Transl Neurol 2016 Sep 01;3(10):791-800 [FREE Full text] [doi: 10.1002/acn3.343] [Medline: 27752514]
23. Essiet IA, Baharom A, Shahar HK, Uzochukwu B. Application of the socio-ecological model to predict physical activity behaviour among Nigerian University students. Pan Afr Med J 2017 Mar 01;26:110 [FREE Full text] [doi: 10.11604/pamj.2017.26.110.10409] [Medline: 28533833]
24. Golden SD, Earp JA. Health Educ Behav 2012 Jun;39(3):364-372. [doi: 10.1177/1090198111418634] [Medline: 22267868]
25. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
26. Schlicht W, Ebner-Priemer UW, Kanning M. Ecological momentary assessment and intervention in physical activity and well-being: affective reactions, social-cognitive factors, and behaviors as determinants of physical activity and exercise. Front Psychol 2013 Dec 06;4:916 [FREE Full text] [doi: 10.3389/fpsyg.2013.00916] [Medline: 24367347]
27. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription, 9th Edition. Philadelphia, PA: Lippincott Williams and Wilkins; 2014.
28. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
29. Moore CG, Carter RE, Nietert PJ, Stewart PW. Recommendations for planning pilot studies in clinical and translational research. Clin Transl Sci 2011 Oct;4(5):332-337 [FREE Full text] [doi: 10.1111/j.1752-8062.2011.00347.x] [Medline: 22029804]
30. Rebchuk AD, Deptuck HM, O'Neill ZR, Fawcett DS, Silverberg ND, Field TS. Validation of a novel telehealth administration protocol for the NIH toolbox-cognition battery. Telemed J E Health 2019 Mar;25(3):237-242. [doi: 10.1089/tmj.2018.0023] [Medline: 29927713]
31. Ricker J, Axelrod B. Analysis of an Oral Paradigm for the Trail Making Test. Assessment 1994 Mar;1(1):47-52 [FREE Full text] [doi: 10.1177/1073191194001001007] [Medline: 9463499]
32. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. Neurology 2013 Mar 12;80(11 Suppl 3):S2-S6. [doi: 10.1212/WNL.0b013e3182872e5f] [Medline: 23479538]
33. Hodes RJ, Insel TR, Landis SC, NIH Blueprint for Neuroscience Research. The NIH toolbox: setting a standard for biomedical research. Neurology 2013 Mar 12;80(11 Suppl 3):S1 [FREE Full text] [doi: 10.1212/WNL.0b013e3182872e90] [Medline: 23479536]
34. Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Slotkin J, et al. The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: validation in an adult sample. J Int Neuropsychol Soc 2014 Jul;20(6):567-578 [FREE Full text] [doi: 10.1017/S1355617714000320] [Medline: 24959840]
35. Choi L, Ward SC, Schnelle JF, Buchowski MS. Assessment of wear/nonwear time classification algorithms for triaxial accelerometer. Med Sci Sports Exerc 2012 Oct;44(10):2009-2016 [FREE Full text] [doi: 10.1249/MSS.0b013e318258cb36] [Medline: 22525772]
36. Hart TL, Swartz AM, Cashin SE, Strath SJ. How many days of monitoring predict physical activity and sedentary behaviour in older adults? Int J Behav Nutr Phys Act 2011 Jun 16;8:62 [FREE Full text] [doi: 10.1186/1479-5868-8-62] [Medline: 21679426]
37. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc 1998 May;30(5):777-781. [doi: 10.1097/00005768-199805000-00021] [Medline: 9588623]
38. Kitsiou S, Thomas M, Marai GE, Maglaveras N, Kondos G, Arena R, et al. Development of an innovative mHealth platform for remote physical activity monitoring and health coaching of cardiac rehabilitation patients. In: Proceedings of the IEEE EMBS International Conference on Biomedical & Health Informatics. 2017 Presented at: BHI 2017; February 16-19, 2017; Orlando, FL. [doi: 10.1109/bhi.2017.7897223]
39. Schwingel A, Linares DE, Gálvez P, Adamson B, Aguayo L, Bobitt J, et al. Developing a culturally sensitive lifestyle behavior change program for older Latinas. Qual Health Res 2015 Dec;25(12):1733-1746. [doi: 10.1177/1049732314568323] [Medline: 25595148]
40. Schwingel A, Gálvez P, Linares D, Sebastião E. Using a mixed-methods RE-AIM framework to evaluate community health programs for older Latinas. J Aging Health 2016 Apr 13;29(4):551-593. [doi: 10.1177/0898264316641075]
41. Ecological momentary assessment. Ilumivu. URL: https://ilumivu.com/about/science/ [accessed 2018-07-06]
42. B15002 Sex by educational attainment for the population 25 years and over (Hispanic or Latino). United States Census Bureau. URL: https://tinyurl.com/mr3rt3nc [accessed 2024-08-20]

XSL•FO
RenderX

43. Erickson KI, Hillman C, Stillman CM, Ballard RM, Bloodgood B, Conroy DE, et al. Physical activity, cognition, and brain outcomes: a review of the 2018 physical activity guidelines. Med Sci Sports Exerc 2019 Jun;51(6):1242-1251 [FREE Full text] [doi: 10.1249/MSS.0000000000001936] [Medline: 31095081]
44. Balbim GM, Aguiñaga S, Ajilore OA, Bustamante EE, Erickson KI, Lamar M, et al. The effects of the BAILAMOS™ dance program on physical activity levels and cognition of older Latino adults: a pilot study. J Aging Health 2022 Jan;34(1):25-40. [doi: 10.1177/08982643211020996] [Medline: 34027686]
45. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA 2018 Nov 20;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
46. Piedra LM, Andrade FC, Hernandez R, Trejo L, Prohaska TR, Sarkisian CA. Let's walk! Age reattribution and physical activity among older Hispanic/Latino adults: results from the ¡Caminemos! Randomized trial. BMC Public Health 2018 Aug 03;18(1):964 [FREE Full text] [doi: 10.1186/s12889-018-5850-6] [Medline: 30075709]
47. Menkin JA, McCreath HE, Song SY, Carrillo CA, Reyes CE, Trejo L, et al. "Worth the walk": culturally tailored stroke risk factor reduction intervention in community senior centers. J Am Heart Assoc 2019 Mar 19;8(6):e011088. [doi: 10.1161/JAHA.118.011088] [Medline: 30836804]
48. Marquez DX, Wilbur J, Hughes S, Wilson R, Buchner DM, Berbaum ML, et al. BAILA: a randomized controlled trial of Latin dancing to increase physical activity in Spanish-speaking older Latinos. Ann Behav Med 2022 Nov 18;56(12):1231-1243 [FREE Full text] [doi: 10.1093/abm/kaac009] [Medline: 35445687]
49. Piedra LM, Andrade FC, Hernandez R, Boughton SW, Trejo L, Sarkisian CA. The influence of exercise on cognitive function in older Hispanic/Latino adults: results from the "¡Caminemos!" study. Gerontologist 2017 Nov 10;57(6):1072-1083 [FREE Full text] [doi: 10.1093/geront/gnw256] [Medline: 28329844]
50. Aguiñaga S, Kaushal N, Balbim GM, Wilson RS, Wilbur JE, Hughes S, et al. Latin dance and working memory: the mediating effects of physical activity among middle-aged and older Latinos. Front Aging Neurosci 2022;14:755154 [FREE Full text] [doi: 10.3389/fnagi.2022.755154] [Medline: 35493932]

Abbreviations

ADRD: Alzheimer disease and related dementias
CONSORT: Consolidated Standards of Reporting Trials
EMA: ecological momentary assessment
EMI: ecological momentary intervention
MVPA: moderate to vigorous physical activity
NIH: National Institutes of Health
REDCap: Research Electronic Data Capture
SEM: socioecological model
TMT-A: trail making test part A
TMT-B: trail making test part B

Edited by J Kaye, I Kronish, A Dominello; submitted 01.12.23; peer-reviewed by Q Xie, G Dunton; comments to author 10.03.24; revised version received 07.06.24; accepted 30.06.24; published 05.09.24.

Please cite as:
Bronas UG, Marquez DX, Fritschi C, Petrarca K, Kitsiou S, Ajilore O, Tintle N
Ecological Momentary Intervention to Replace Sedentary Time With Physical Activity to Improve Executive Function in Midlife and Older Latino Adults: Pilot Randomized Controlled Trial
J Med Internet Res 2024;26:e55079
URL: https://www.jmir.org/2024/1/e55079
doi:10.2196/55079
PMID:39235836

©Ulf G Bronas, David X Marquez, Cynthia Fritschi, Katherine Petrarca, Spyros Kitsiou, Olu Ajilore, Nathan Tintle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Remote Passive Sensing of Older Adults' Activities and Function: User-Centered Design Considerations for Behavioral Interventions Conducted in the Home Setting

Lyndsey M Miller[1,2], BA, BSN, PhD; Jeffrey Kaye[1,3], MD; Allison Lindauer[3], BSN, MSN, PhD; Wan-Tai M Au-Yeung[1,3], PhD; Nathaniel K Rodrigues[1], BS; Sara J Czaja[4], PhD

[1]Oregon Center for Aging and Technology, Oregon Health and Science University, Portland, OR, United States
[2]School of Nursing, Oregon Health and Science University, Portland, OR, United States
[3]Department of Neurology, Oregon Health and Science University, Portland, OR, United States
[4]Center on Aging and Behavioral Research, Division of Geriatrics and Palliative Medicine, Weill Cornell Medicine, New York, NY, United States

Corresponding Author:
Lyndsey M Miller, BA, BSN, PhD
School of Nursing
Oregon Health and Science University
3455 SW US Veterans Hospital Road
Portland, OR, 97239
United States
Phone: 1 5034943837
Email: millerly@ohsu.edu

Abstract

Behavioral intervention studies often lack sufficiently sensitive and frequent measurements to observe an effect. Remote passive sensing offers a highly sensitive, continuous, and ecologically valid method of assessment that increases the ability to detect changes in the daily activities and function of those being monitored. To be most effectively deployed in research studies, applications of remote assessment technology must be designed with the end user in mind. User-centered design (UCD) is especially important in clinical trials where the needs and characteristics of participants and research staff need to be uniquely considered to ensure the feasibility and acceptability of the study. This paper describes UCD issues in remote passive sensing that commonly arise among older adult participants—including those living with dementia—as well as any strategies that were taken to overcome them. Using exemplars from the National Institute on Aging–funded Roybal Center ORCASTRAIT (Oregon Roybal Center for Care Support Translational Research Advantaged by Integrating Technology), as well as other experimental and observational research studies conducted in community settings, this paper brings together our collective experiences with studies using remote passive sensing technology that incorporate a UCD design approach. Although passive sensing eliminates some common UCD issues that arise with higher-touch technology, issues, such as usability, trust, and aesthetic acceptability, still need to be addressed for behavioral interventions using passive sensing technology to be potent and implementable.

(J Med Internet Res 2024;26:e54709) doi:10.2196/54709

KEYWORDS

user-centered design; remote passive sensing; remote monitoring; behavioral interventions; caregiving; dementia; Alzheimer; monitoring; gerontology; geriatrics; older adult; aging; usability; acceptability; trust; behavioral

Introduction

As digital health technologies continue to proliferate in both research and clinical care, the uptake of remote passive (or ambient) sensing for noninteractive digital data collection represents a unique opportunity to detect meaningful patterns in older adults' daily activity over longer periods of time without requiring interaction from participants. Subtle decline in everyday function occurs as part of the normal aging process and as a symptom of pathologies commonly found among older adults, such as Alzheimer disease and related disorders. Likewise, stress from caregiving for older adults with chronic health conditions impacts everyday routines and behaviors such as sleep and physical activity. Detecting and differentiating these subtle changes is ideally pursued in the home setting, where older adults perform their regular daily routines and

XSL•FO
RenderX

demonstrate their functional capacity in ecologically valid ways. Remote digital assessment of older adults' daily activity holds promise for detecting early signs of Alzheimer disease [1]; for intervening to prevent adverse health outcomes, such as falls and the need for emergency care [2-4]; and for understanding the effects of interventions designed to support daily life or caregiving activities [5-8]. Ultimately, remote digital assessment may contribute to helping older adults live longer independently; however, in order to do, so the technology must be designed with the user in mind—in a way that is minimally obtrusive and maximally acceptable to older adults at risk for decline and, in many cases, their family members.

Passive sensing occurs when sensors are embedded in the environment and require no active input from the study participants. The goal is to remotely assess daily function and other regular activities with minimal disturbance to participants' lives and to increase the ecological validity of assessments by moving the observations from the clinic into the home setting. Examples of passive sensors include bed pressure mats; physical environmental sensors, such as thermostats; or motion sensors on the wall. Contrarily, active sensing requires study participants to actively input data or to perform a task through a digital interface. For example, a mobile app may require users to enter physical activity goals and turn on an activity tracker during exercise. Even wearable sensors, such as actigraphy watches, require some user interaction. Generally, passive sensing is more suitable for long-term monitoring as it requires minimal input from users. Despite the benefits of passive sensing to gain more detailed and unobtrusive insights into participants' daily lives, issues in the design of passive sensing systems can still interfere with user experience and long-term willingness to participate in such research studies.

User-centered design (UCD) is a design philosophy focused on understanding the users, tasks, and environments [9]. It is also an approach grounded in the characteristics of the individuals who use the innovation [10] and a process that iteratively tests usability and the user experience (ISO 2010). General principles, or heuristics, of UCD, may be overlooked in passive sensing, although this would be an oversight since many of the heuristics are applicable (eg, error prevention and aesthetics) [11]. Researchers who incorporate remote assessment into their studies may be motivated by a variety of reasons to focus on UCD, for example, optimization of enrollment outcomes (eg, faster recruitment) and minimization of attrition. In general, the more passive the technology, the less design of the user interface is required. Yet, UCD issues still exist in noninteractive passive sensing technology and should be considered early and often in order to conduct research that is acceptable or even appealing to participants, thereby improving data quality and retention. The objective of this paper is to provide an overview of UCD issues that pertain to remote passive sensing and to give examples from research studies that use passive sensing.

Users and UCD Issues in Passive Sensing

Users

The UCD process begins with identifying end users and understanding their characteristics, abilities, needs, and preferences [9]. Although the primary users of remote assessment technology are the participants who generate the digital activity data by living in the midst of these sensors, secondary users include anyone who installs the sensors and maintains their integrity, and those who access the remote sensor data streams to analyze activity or health outcomes. Depending on the type of research study, secondary users may include clinicians, family members, or other individuals who are involved in the maintenance of, or interaction with, the sensors or the data derived from them.

UCD Issues

Trust

The willingness of users (eg, research participants, family members, and clinicians) to accept remote sensor platforms or other automated systems as a means of detecting meaningful changes in function and daily life is determined in large part by trust [12]. Ideally, the users' perceptions of the performance and functionality of the remote digital activity monitoring system will match the actual performance and functionality [13]. Yet, "undertrust" of the system occurs at times when users infer nonexistent capabilities, especially those that may invade their privacy. For example, users may worry that passive infrared (PIR) sensors placed in the bathroom are "hearing or seeing" their activities, even though there are no audio or video recording capabilities. Lack of trust is not always unfounded though and may occur when users perceive or experience high error rates (eg, false alarms) or an unreliable system. In the case of remote passive sensing, the unreliability of the system would in most cases only be detected by those users with access to the system's operational status or the data flowing from the sensors.

Error Recognition and Recovery

The ability to recognize errors and the ease of recovering from errors are central to UCD. In a remote passive sensing platform, most participants would not be expected to recognize or infer errors in the transmission of their data unless a sensor is visibly broken or out of place. Participants who are particularly proactive or curious about the functionality of the remote passive sensing platform may become concerned or frustrated by their inability to see the operational status or detect errors (ie, the system is a "black box" to them). In general, study teams working with remote passive sensing would need to reassure participants that they will be in touch if any issues are detected.

When PIR motion sensors are used on the walls or ceilings of a participant's home, participants will notice when a sensor has fallen on the floor. Recovering from this error can be as simple as sticking the sensor back on the wall in the approximate place where it belonged. However, some studies use a line of motion sensors on the ceiling to detect gait speed. These sensors must be placed more precisely at a certain distance from each other

XSL•FO
RenderX

in order to detect gait speed accurately, and so, in this case, error recovery would depend upon the participant reporting the fallen sensor to a technician, who would then come and restore it to the ceiling. With wearable sensors that have a long-term battery life and do not need to be removed from the wrist, "errors" of omission may occur when the participant chooses not to wear it at certain times. These types of errors would likely not be reported by the participants, and error recovery would only occur when secondary users (eg, technicians or other members of the study team) recognized the lapse in data and contacted the primary user to diagnose the error. Lapses in data can occur with any sensor for a variety of reasons, thus the remote passive sensing system must be monitored frequently by the secondary users for data quality errors.

Overtrust and Therapeutic Misconception

There is also the potential for users to "overtrust" the purpose of the remote monitoring system (ie, misconstrue the purpose or capability of the system and assume nonexistent benefit). The issue of overtrust typically is discussed in the design of robotics and automated driving systems [14]; however, there are parallels in remote passive sensing as well. With sensors on the walls and doors, some participants in remote monitoring studies could mistake the functionality of sensors to be a home security system that will keep them safe. Older adult participants who wear a sensor may mistake the capability of the system to include a falls alert mechanism. These misconstrued benefits create an ethical issue of therapeutic misconception since participants may not seek the services that they need or desire if they misconceive the remote monitoring system as already providing those services. Just as it must be addressed in drug trials, the boundaries between clinical care and experimental therapies must be well defined [15]. The limits of a remote monitoring system need to be effectively communicated to users to avoid overtrust and therapeutic misconceptions.

Aesthetic Acceptability

Aesthetic and minimalistic design are also an important usability heuristic. Although the beauty of the design improves the user experience, only essential information should be presented in the user interface for maximum function [11]. Yet, the primary goal of the design of a passive sensing system is inconspicuousness. Thus, the heuristic must be modified to prioritize unobtrusive design, which allows participants to have minimal contact with the sensors and continue with their daily lives without changing their routines. To accomplish this goal, it is necessary to use sensors that are small in size, have a neutral color, and lack any disturbances from the technology such as lights or beeping that could disturb a participant in their home and make the sensors more obtrusive. For wearable sensors that often appear to be an accessory (eg, actigraphy watches), the initial impression of aesthetic appeal and comfort is important to older adult participants' attitudes toward using this type of sensor over time [16]. There also should be, as far as possible, minimal maintenance requirements, such as a frequent need to recharge or replace batteries.

Data Disaggregation in Multiperson Homes

A UCD issue that pertains to the secondary users who work with the data from remote passive sensing is the inclusion of features that allow the disaggregation of data from homes that include more than one resident. Certain data collected in multiperson homes can be difficult to attribute to a specific person within the home. For example, while the PIR sensors can detect motion within each living space, they do not record pictures or videos, so it is difficult to know whose motion has triggered the sensors. Innovative algorithms are needed to analyze these PIR sensor data to extract meaningful features that would describe the home as one unit. Alternatively, a redesign of the system using more advanced technology that tracked each individual distinctly within the home could be used; however, these improvements for secondary users (those working with the data) would likely detract from the trust of the primary users (the research participants).

Other issues include the framework for data sampling, such as frequency and time, and data management issues, such as approaches for reducing data into meaningful units. The organization and presentation of the data to secondary users must be done in a way that the data are usable and trusted. For example, errors, such as too many false alarms, may result in an undertrust or "boy who cried wolf phenomenon." Secondary users also need to be trained in the use and meaning of the data.

Strategies for Addressing UCD Issues

Overview

In general, UCD strategies aim to make technology more appealing to encourage uptake by users. Paradoxically, UCD strategies for passive sensing generally aim to decrease the need for the user's interaction with the system and instead to make it as unobtrusive and as "low-touch" as possible. The remote passive sensing issues described in this paper largely concern the assumptions that users make about the sensors, rather than their interaction with them. Thus, the strategies that we have used to mitigate these issues among participants in our research groups must address the users understanding and knowledge about the purpose and function of passive sensing. In research studies, this process begins with informing and educating participants about the system prior to installation and then continues during beta testing and use in research studies.

Informed Consent

At the National Institute on Aging–funded Roybal Center ORCASTRAIT (Oregon Roybal Center for Care Support Translational Research Advantaged by Integrating Technology), education about the study technologies begins during recruitment and continues throughout study participation. The language used in recruitment materials and during informed consent sets the tone and expectations for participants. Consent forms contain plain language about the function of remote sensing, such as "Please note: the study technologies are not a security system. There is no video, photograph, or audio recording. The technology will not detect a break-in, a fall, or other emergencies. We will not be able to notify anyone in the event of an emergency."

XSL•FO
RenderX

Participant Education

Educating and re-educating participants continues during studies about the capabilities (or lack thereof) of technology. One way that the ORCASTRAIT teams achieve this is by showing participants examples of their data to illustrate the deidentified nature of the data as it appears to the research team (Figure 1). While showing participants the data depicted in Figure 1, an ORCASTRAIT technician will describe it as follows: “The presence of a shape on the graph indicates that the sensor was activated on that day. This is the 30,000 feet view of the data, zoomed out, giving us an overall picture of the status of the sensors in the homes.”

Users bring their preconceived notions of technology, though, and often retain them even after the study team’s educational efforts. Clearly, the fit between the participant’s goals and the goals of the research will not always be possible and should be taken into account. Yet, a concerted effort must be applied to ensure that we address participants’ concerns and misconceptions when possible, thereby improving readiness. These efforts must be robust in order to include the most representative sample possible since participants from diverse backgrounds vary in terms of digital literacy or misconceptions about remote sensing technology.

Figure 1. An example of passive infrared motion sensor data shown to participants to mitigate the lack of trust in the anonymous nature of the data.

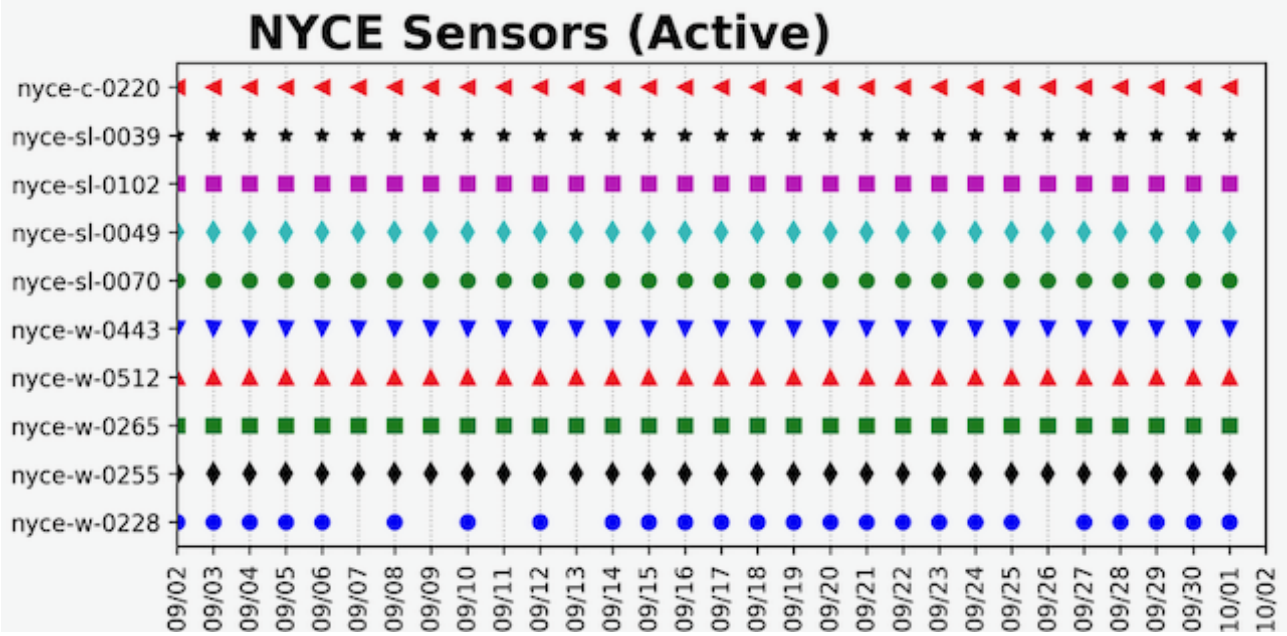

Examples From Intervention Studies

Overview

UCD principles are frequently considered in the initial development of digital devices and technology interventions; however, UCD is less often considered during the translation and implementation processes of intervention testing, refinement, and expansion [10]. Mental and behavioral health interventions in particular can benefit from applying UCD principles [17]. Recently, the remote monitoring system was developed by the Oregon Center for Aging and Technology (ORCATECH) and used in 2 pilot studies of behavioral interventions funded by the National Institute on Aging–Roybal Center ORCASTRAIT at Oregon Health and Science University (OHSU). These studies were conducted among persons living with dementia and their care partners to detect changes in daily routines such as sleep, physical activity, and time out of the home: READyR (Remote Assessment and Dynamic Response Intervention to Support Co-Residing Care Dyads in Identifying Dementia-Related Care Needs; ClinicalTrials.gov NCT04542109) and STELLA (Support via Technology: Living and Learning with Advancing Alzheimer’s disease and related dementias; ClinicalTrials.gov NCT04335110).

The READyR Program

The READyR Program is a behavioral intervention designed to assess and target everyday routines of care dyads (older adults with cognitive impairment and their cohabiting partners) using in-home sensors and videoconference sessions in order to understand future care needs and promote greater well-being [18]. In the pilot study of READyR (OHSU Institutional Review Board #20397), ORCATECH infrastructure and a platform of sensors were deployed to collect objective data from remote passive assessments of daily routines. The technology platform used in the READyR study included PIR sensors that captured home exits and activity within or transitions between rooms (NYCE, wireless PIR sensors using a ZigBee network), an actigraphy watch that detected step count (Withings Activité/Steel Watch), an electronic pillbox that captured medication-taking behavior, (TimerCap, wirelessly transmits data about when and which day of the week the pillbox doors were open or closed), and an under-the-mattress bed mat that captured a variety of sleep metrics (eg, bed exits at night and total sleep time) and physiologic data such as heart rate and respirations (Emfit, ballistocardiography sensor).

One of the goals of the READyR intervention was to better assess the alignment of participants’ daily routines with their care values such as autonomy and safety. To help participants

XSL•FO
RenderX

visualize their daily activity data in intervention sessions, a tailored web-based tool was created. Data specific to each participant's home were extracted from the ORCATECH central server. These data were processed and analyzed to generate plots and figures of activity related to participants' daily routines, which were incorporated into an HTML template to create a static web page that could be displayed for participants.

User experiences with having passive sensors in the home, and with viewing their daily activity data, were video recorded in intervention sessions and acceptability surveys. Of the sample of 24 participant dyads (half of which had cognitive impairment, and the other half were their spouses) who completed the acceptability survey, 23 participants rated the sensor data that were presented to them in web-based intervention sessions as "very easy" or "easy" to understand, and only 1 participant rated it as difficult. The most common user issue was that when presented with data about their daily routines, over half of the participants expressed doubt that the sensors accurately captured at least one of their routines such as time out of home or sleep disturbances. A few participants gave concrete reasons for their doubt, for example:

> *You might want to check that the sensors are working properly, because one keeps falling off the wall. We try to put it back up, but it doesn't stick, so I'm wondering if this is really my activity or if it is a sensor thing.*

The ORCATECH platform includes multiple ways of assessing some functions, such as sleep, and participants were more reassured that the data represented their patterns when we could show them 2 sources of data that were in agreement.

The issue of therapeutic misconception arose several times during the READyR pilot study. One participant living with Alzheimer disease was prone to somnambulism, and her partner expressed hope that the sensors could offer a way to detect it and alert them, especially when she opened the front door at night. Another participant with mild cognitive impairment went for long bike rides and did not bring his phone. His spouse expressed disappointment that there was no capability to track his whereabouts with the wearable in case he got lost or hurt. A third participant living with Alzheimer disease had a tendency to fall, and his spouse was concerned that there were no sensors in their garage to detect a fall. In all cases, re-educating the participants about the functionality of the sensors, and the lack thereof, was our main strategy for addressing this UCD issue. In a future iteration and testing of the READyR Program, we plan to produce a sensor guide for families to refer to after installation and examine other ways of providing intermittent information to curb misperceptions.

The STELLA Study

The STELLA study enrolled 13 care partners and their family members with moderate to late dementia. The study tested an intervention designed to help care partners recognize and address behavioral symptoms of dementia, with the primary aim being to reduce care partner burden (OHSU Institutional Review Board #19306). The STELLA study used bed mats to assess sleep duration and quality (under-the-mattress, ballistocardiography-based sensor, Emfit), PIR wall sensors (NYCE) to assess motion, and wearables to assess step counts (Withings Activité/Steel Watch).

After completing the intervention, 4 STELLA care partners participated in a focus group that addressed their perceptions of STELLA and their experiences with technology. Several technology-related user experience themes emerged from the focus groups: (1) a mismatch between study goals and the participants' needs; (2) trust of the system, or privacy concerns; and (3) and lack of aesthetic and minimalist design.

Care partners found that their needs were not met with the technology in that they had to manage the interaction between their family member with dementia and the sensors: "I should have put them (the sensors) up higher." His family member with dementia saw "these little things on the wall, and pulls them off. And I find them in her pocket later." Two STELLA care partners had difficulty convincing their family members to wear the watch. The technology on the watch did not update to adjust to time changes, also leading to frustration, as this care partner noted:

> *It's not dependable. I just hate to wake up on the morning when we have daylight saving time and it's the wrong time and we're late for church or whatever.*

Some STELLA participants also expressed worries about privacy and the "undertrust" discussed above. They were concerned that the team could monitor their activity while they were in bed:

> *...I'm exhausted. I'm laying here, you know, it's like, oh, my gosh, they're noticing that I'm here. You know, it's like, do they really need to know that kind of thing? Hmm.*

Some of the focus group participants did not like the appearance of the sensors, indicating a lack of aesthetic acceptability:

> *The monitors are on the walls are really ugly. I mean, we got 30 up in the house and, uh, wow. Uh, it's it doesn't look good.*

Another participant commented that he was somewhat mystified by the wall sensor, indicating the lack of aesthetic acceptability and minimalist design:

> *The thing on the wall [the wall sensor] was had mysterious numbers all the time...What does all that mean...?*

Despite the study team's efforts to use unobtrusive technology, the STELLA care partners found the devices far from "passive." The existence of the sensors in the home, even if passive, led some participants to worry about their privacy, struggle to help their family members use devices, like watches, and grapple with malfunctions:

> *I'm like, this is crazy. This is for people with behavioral disturbances? I sent it back. I don't like those watches or the motion detector.*

It should be noted that all STELLA participants had experience using technological devices such as phones, computers, and

XSL•FO
RenderX

email. They all had consistent support from the ORCASTRAIT team:

> *The gal [technology specialist] that comes to our house, she's probably been here more than the mailman.*

The STELLA focus group revealed that as we tested this intervention that addressed the care partner burden, it is possible we increased the care partners' sense of daily hassle with the technology.

Overall, the user experience feedback from focus groups may reflect the fact that family caregivers of older adults with later-stage dementia are a distinct type of user that had not been considered in the original design of the passive sensing platform. The additional needs and preferences of a new set of users should be considered [17], and after identifying those distinct needs, the UCD process involves ideation and exploration of prototypes and testing [9]. In this case, the new user group of family caregivers of persons with later-stage dementia will prompt a redesign of the passive sensing platform before testing again with this user group.

FITTLE

More commonly accepted passive technologies, such as a long-term wearable device, can also highlight UCD issues in behavior intervention studies. One example is from a recent study evaluating a digitally delivered physical activity and social support intervention among community-dwelling older adults (ClinicalTrials.gov NCT03538158), which included the use of wearables (similar to a wristwatch) to objectively assess activity engagement or step count. The participants ranged in age from 65 to 98 years and some had minimal experience with technology. Although a standard commercially available wearable was selected for the study, an issue that arose was a lack of adherence among the participants with respect to the use of wearable. Some participants were not used to wearing a watch, and others were not accustomed to wearing a watch at bedtime. These aspects of comfort and acceptability resulted in gaps in the data when participants forgot to put on the wearable or chose not to wear it at night (SJ Czaja, 2024, unpublished data). This reinforces the need for reminding study participants about the function of the passive technology, and the importance of adherence with respect to study outcomes.

Examples From Longitudinal Studies

Overview

Long-term use of a remote sensing platform is often the intended duration. One of the advantages of a remote assessment system is the ability to monitor daily life in the home setting and detect changes over time. However, new UCD issues arise during long-term monitoring that cannot be taken into account during short-term testing of the system.

MODERATE

MODERATE is a longitudinal study for which the target population is couples living at their residences, with one having dementia and one being the care partner. MODERATE focuses on identifying digital biomarkers of agitated behaviors exhibited by participants living with dementia and identifying environmental precipitants of such behaviors. Passive motion sensors and contact sensors (NYCE, wireless PIR sensors using a ZigBee network), were installed in participants' homes to continuously assess in-home activity.

Algorithms developed at ORCATECH extract time spent out of home based on whether there was motion detected between doors opening or closing. In single-resident homes, the metric, time spent out of home, can be relatively easy to interpret. However, in double-resident homes, time spent out of home refers to the time both residents are out of home and that does not include the time when only one of them is out of the home. In addition, in double-resident homes, it is difficult to attribute motion detected in a specific room to a particular person, and innovative algorithms are required to synthesize the motion sensor data. We have also developed an algorithm that extracts the duration of independent life-space activities within the home. This metric refers to the duration when there were activities simultaneously detected in two or more life spaces within the home. Such a metric can be used to infer a basic level of physical function and independence of the participants. Yet, the primary users' unanticipated activity patterns impact the error rate of these algorithms. Long sedentary periods of sitting on the couch may look identical to an unoccupied room. Exiting and re-entering a room quickly may not be recorded as a room transition. Testing these and other behaviors using scripted trials in the home was necessary to identify and recover from these types of errors in the algorithms.

In MODERATE, we also deployed actigraphy devices, research-grade Actiwatch Spectrum, to be worn by people living with dementia, and Withings Activité/Steel Watch, to be worn by caregivers, to track their daily level of activity. Although we have demonstrated powerful examples of behavioral symptoms being detected by actigraphy devices among persons living with dementia [19], we have also encountered "errors" of the actigraphy devices being lost or damaged. One of the wearables was found in the participant's clothes dryer. In addition, one of the caregivers' actigraphs was struck by a person living with dementia during a bout of agitation. As a result, the glass surface broke. We learned that there are distinct UCD issues with deploying digital technology in the homes of people living with later-stage dementia as their cognition and behavioral symptoms worsen. In addition, in our research settings, the Actiwatch Spectrums (activity monitoring watch with solid-state "piezo-electric" accelerometer and color-sensitive photodiodes) only have around 4 weeks of battery life. As a result, we continuously need to swap out devices via mailing which is labor-intensive for both our staff and participants.

Collaborative Aging Research Using Technology

The Collaborative Aging Research Using Technology (CART) initiative was an interagency (National Institutes of Health and Department of Veterans Affairs) program of research funded to spur wider, evidenced-based use of digital technologies in aging research. CART developed and validated protocols and infrastructure in a longitudinal demonstration study, enrolling 301 participants across 4 diverse cohorts (rural residing residents, older adults living in low-income housing, African

XSL•FO
RenderX

American older adults living in Chicago, and Latino older adults living in Miami). The project demonstrated that the system could enable digital health monitoring and intervention delivery that is technology agnostic and extensible to many use cases and research scenarios [20]. The scalable system is being used by multiple research groups in North America and Europe [21]; versions of this platform were used in the READyR and STELLA studies described above. In developing the system, priority was placed on ensuring that the methods were incorporated into the daily lives of diverse populations of older adults. User experience feedback was crucial and remains key to the continuous improvement of the experience and engagement of older populations in research; this is crucial to ensuring long-term adherence to research protocols, reducing missing data and maintaining ecological validity.

To highlight UCD principles further, we note some observations from the CART study. For example, the low-income housing cohort (participants living in US Department of Housing and Urban Development, Section 202 subsidized housing) and the rural-residing veterans cohort were found to have overall excellent daily adherence to wearing a wrist-worn watch (Withings Activité/Steel Watch) that recorded step counts and sleep times, with an overall daily wear compliance of 86.5% [22]. However, it was observed that despite using a device chosen for its typical wrist-watch appearance (ie, higher aesthetic acceptability for a research-device wearable), which required battery replacement only every 8-9 months, and for which there was clear participant agreement at the study entry that the watch was intended to be worn throughout the day and night (except for bathing), the wear compliance at night was found to be lower after a mean of 4.7 (SD 1.8) months of study (69.5%) [20]. When asked about the lower wear compliance at bedtimes, participants relayed that they found they were not comfortable wearing any wrist-worn device every night, even though they thought that this would be feasible for them at the beginning of the study. This experience highlights, as noted here and for the READyR and STELLA intervention studies described above, that user-experience needs to be studied under the actual conditions of use. CART participants using the watch in short-term (2 weeks in duration) device validation pilot studies [23] did not voice this potential longer-term compliance issue.

Although the above examples illustrate how participants may choose for understandable reasons to not comply with a protocol because of usability factors and personal life circumstances, we also note that many dedicated participants will invent or devise “workarounds” to enable them to continue in a study. These must be identified during the study so that, for example, if a protocol states that a wearable should be worn all the time (as in our example) or worn on the nondominant wrist, and a participant decides they are more comfortable wearing the device on their other wrist, at variable times, this can add inaccuracies to the data. In some cases, we have discovered that participants devise an alternative that does not affect the data (Figure 2). Participants’ ideas and solutions are welcome, and in fact, the research team may gain valuable knowledge from their relating to that experience. Finding ways to intentionally incorporate the user experience into the design process of remote sensing research can mitigate many UCD challenges.

Figure 2. (A) Original watch to be worn in the CART study (Withings Activité). (B) The watch altered by a participant to cover the watch face using tape, with watch hands drawn on so that the participant would not see the watch as a timepiece, as they found the watch not providing accurate time (likely that the battery was wearing down). CART: Collaborative Aging Research Using Technology.

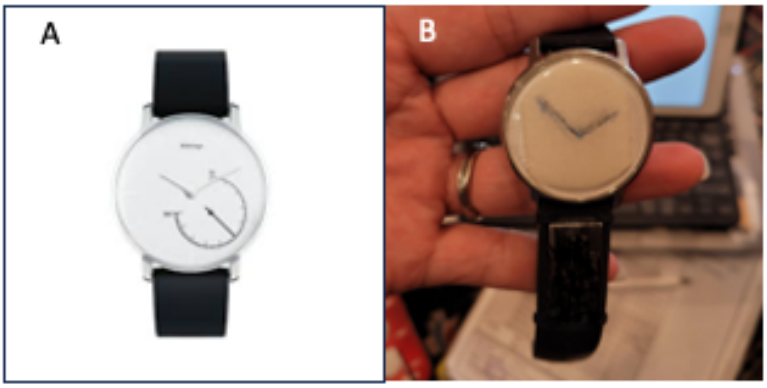

Discussion

The proliferation of remote passive sensing in both research and clinical care expands the possibilities for objectively detecting meaningful patterns in function and daily activity over longer periods of time in the home setting. Many of these technologies require minimal or no interaction from participants; however, this does not eliminate UCD issues in remote passive sensing. This paper provides an overview of the most prevalent UCD issues encountered in 5 remote passive sensing studies: 3 behavioral intervention studies (READyR, STELLA, and FITTLE) and 2 longitudinal studies (MODERATE and CART), all conducted with community-dwelling older adult participants living cognitive impairment and their care partners. We describe UCD issues of trust, error recognition and recovery, therapeutic misconception, aesthetic acceptability, and data disaggregation in multiperson homes and the strategies we have used to mitigate these issues broadly, as well as in examples from our research.

The frequent or continuous objective assessment of real-world outcomes of sleep, activity, and objective indicators of stress can greatly increase understanding of the efficacy of behavioral interventions designed to improve these important aspects of well-being. Moreover, taking advantage of passive sensing in

XSL•FO
RenderX

order to obtain objective measures for both caregivers and persons living with dementia in the stream of everyday life provides unprecedented opportunities to gain insight into the mechanisms of behavioral interventions. Yet, if participants do not trust or do not accept the use of passive sensing technology in their daily lives, they may change their behavior around it, stop using it when it is inconvenient or uncomfortable (as in the case of the actigraphy watch at night), or withdraw from the study altogether. Clearly, this is not a worthwhile tradeoff if it results in missing data or attrition. In all examples presented in this paper, educating participants was a fundamental strategy to improve the likelihood of acceptability of remote sensing technology; however, it is not a panacea to all UCD issues.

The choice of technologies to be deployed in a study is driven by the specific use case and can also be a tradeoff between desired functionalities for research and usability or acceptability by study participants. Adding more technology to a study for more functionality can in some cases tip the scales toward greater acceptability, as was the case in the READyR study with the participants who wanted more evidence that the daily activity patterns shown in the data reflected their true routines. Having more than 1 sensor stream showing similar data was better in that study. In the STELLA study, participants found the technology to be less acceptable and more burdensome.

Acknowledgments

This work was supported by the National Institute on Aging of the National Institutes of Health (grants P30AG024978, R24AG064191, K01AG059839, K25AG071841, R01AG67546, U2CAG054397, and R01AG053163). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Conflicts of Interest

JK has received research support awarded to his institution, Oregon Health and Science University (OHSU), from the National Institutes of Health (NIH), National Science Foundation (NSF), and AbbVie. JK has been directly compensated for serving on a Data Monitoring Committee for Eli Lilly, as a consultant to Boston Scientific, and as an external Advisory Committee member for the Rush and Stanford University Alzheimer's Disease Research Centers. JK receives reimbursement through Medicare or commercial insurance plans for providing clinical assessment and care for patients. JK serves on the editorial advisory board and as Associate Editor of the journal, *Alzheimer's & Dementia*. JK is a guest editor for the Theme Issue "Digital Technology for Behavioral Interventions in Aging: Opportunities and Challenges" in the *Journal of Medical Internet Research*. OHSU and JK have a financial interest in Life Analytics, Inc, a company that is developing remote monitoring software technology. The nature of this financial interest relative to this study has been reviewed by the Research Integrity Office at OHSU. For more information, please contact the OHSU Research Integrity Office at coir@ohsu.edu.

References

1. Hampel H, Au R, Mattke S, van der Flier WM, Aisen P, Apostolova L, et al. Designing the next-generation clinical care pathway for Alzheimer's disease. Nat Aging 2022;2(8):692-703 [FREE Full text] [doi: 10.1038/s43587-022-00269-x] [Medline: 37118137]
2. Gaugler JE, Rosebush CA, Zmora R, Albers EA. Outcomes of remote activity monitoring for persons living with dementia over an 18-month period. J Am Geriatr Soc 2022;70(8):2439-2442 [FREE Full text] [doi: 10.1111/jgs.17839] [Medline: 35532893]
3. Piau A, Mattek N, Crissey R, Beattie Z, Dodge H, Kaye J. When will my patient fall? Sensor-based in-home walking speed identifies future falls in older adults. J Gerontol A Biol Sci Med Sci 2020;75(5):968-973 [FREE Full text] [doi: 10.1093/gerona/glz128] [Medline: 31095283]
4. David MCB, Kolanko M, Del Giovane M, Lai H, True J, Beal E, et al. Remote monitoring of physiology in people living with dementia: an observational cohort study. JMIR Aging 2023;6:e43777 [FREE Full text] [doi: 10.2196/43777] [Medline: 36892931]
5. Figueiro MG, Pedler D, Plitnick B, Zecena E, Leahy S. Tailored lighting intervention (TLI) for improving sleep-wake cycles in older adults living with dementia. Front Physiol 2023;14:1290678 [FREE Full text] [doi: 10.3389/fphys.2023.1290678] [Medline: 38169838]
6. Hodgson NA, Gooneratne N, Perez A, Talwar S, Huang L. A timed activity protocol to address sleep-wake disorders in home dwelling persons living with dementia: the healthy patterns clinical trial. BMC Geriatr 2021;21(1):451 [FREE Full text] [doi: 10.1186/s12877-021-02397-2] [Medline: 34344312]
7. Garshol BF, Ellingsen-Dalskau LH, Pedersen I. Physical activity in people with dementia attending farm-based dementia day care—a comparative actigraphy study. BMC Geriatr 2020;20(1):219 [FREE Full text] [doi: 10.1186/s12877-020-01618-4] [Medline: 32571238]
8. Wrede C, Braakman-Jansen A, van Gemert-Pijnen L. Understanding acceptance of contactless monitoring technology in home-based dementia care: a cross-sectional survey study among informal caregivers. Front Digit Health 2023;5:1257009 [FREE Full text] [doi: 10.3389/fdgth.2023.1257009] [Medline: 37860038]

XSL•FO
RenderX

9. Czaja SJ, Boot WR, Charness N, Rogers WA. Designing for Older Adults: Principles and Creative Human Factors Approaches. Boca Raton, FL: CRC Press/Routledge/Taylor & Francis Group; 2019.
10. Dopp AR, Parisi KE, Munson SA, Lyon AR. A glossary of user-centered design strategies for implementation experts. Transl Behav Med 2019;9(6):1057-1064. [doi: 10.1093/tbm/iby119] [Medline: 30535343]
11. Nielsen J. Ten usability heuristics. Semantic Scholar. URL: https://pdfs.semanticscholar.org/5f03/b251093aee730ab9772db2e1a8a7eb8522cb.pdf [accessed 2024-09-24]
12. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 2015;57(3):407-434. [doi: 10.1177/0018720814547570] [Medline: 25875432]
13. Kaindl H, Svetinovic D. Avoiding undertrust and overtrust. REFSQ Workshops 2019 [FREE Full text]
14. Wagner AR, Borenstein J, Howard A. Overtrust in the robotic age. Commun ACM 2018;61(9):22-24. [doi: 10.1145/3241365]
15. de Bot ST. Raising awareness of therapeutic misconception and optimism around clinical trials in Huntington's disease. J Huntingtons Dis 2019;8(4):431-433 [FREE Full text] [doi: 10.3233/JHD-199006] [Medline: 31640105]
16. Charness N, Best R, Evans J. Supportive home health care technology for older adults: attitudes and implementation. Gerontechnology 2016;15(4):233-242 [FREE Full text] [doi: 10.4017/gt.2016.15.4.006.00] [Medline: 29033700]
17. Lyon AR, Koerner K. User-centered design for psychosocial intervention development and implementation. Clin Psychol (New York) 2016;23(2):180-200 [FREE Full text] [doi: 10.1111/cpsp.12154] [Medline: 29456295]
18. Miller LM, Solomon DN, Whitlatch CJ, Hiatt SO, Wu CY, Reynolds C, et al. The remote assessment and dynamic response program: development of an in-home dementia-related care needs assessment to improve well-being. Innov Aging 2022;6(2):1-12 [FREE Full text] [doi: 10.1093/geroni/igac006] [Medline: 35402736]
19. Au-Yeung WTM, Miller L, Beattie Z, Dodge HH, Reynolds C, Vahia I, et al. Sensing a problem: proof of concept for characterizing and predicting agitation. Alzheimers Dement (N Y) 2020;6(1):e12079 [FREE Full text] [doi: 10.1002/trc2.12079] [Medline: 32864417]
20. Beattie Z, Miller LM, Almirola C, Au-Yeung WTM, Bernard H, Cosgrove KE, et al. The collaborative aging research using technology initiative: an open, sharable, technology-agnostic platform for the research community. Digital Biomarkers 2020;4(Suppl 1):100-118 [FREE Full text] [doi: 10.1159/000512208] [Medline: 33442584]
21. Thomas NW, Beattie Z, Riley T, Hofer SM, Kaye J. Home-based assessment of cognition and health measures: the collaborative aging research using technology (CART) initiative and international collaborations. IEEE Instrum Meas Mag 2021;24(6):68-78. [doi: 10.1109/mim.2021.9513638]
22. Kaye J, Silbert L, Beattie Z, Mattek N, Wall R, Marcoe J, et al. High compliance, continuous long-term home monitoring of sleep activity in diverse populations differentiates lower cognitive functioning older adults. Alzheimers Dementia 2019;15:P552. [doi: 10.1016/j.jalz.2019.06.4311]
23. Au-Yeung WTM, Kaye JA, Beattie Z. Step count standardization: validation of step counts from the withings activite using PiezoRxD and wGT3X-BT. Annu Int Conf IEEE Eng Med Biol Soc 2020;2020:4608-4611 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9176511] [Medline: 33019020]

Abbreviations

CART: Collaborative Aging Research Using Technology
OHSU: Oregon Health and Science University
ORCASTRAIT: Oregon Roybal Center for Care Support Translational Research Advantaged by Integrating Technology
ORCATECH: Oregon Center for Aging and Technology
PIR: passive infrared
READyR: Remote Assessment and Dynamic Response
STELLA: Support via Technology: Living and Learning with Advancing Alzheimer's disease and related dementias
UCD: user-centered design

Edited by A Mavragani; submitted 20.11.23; peer-reviewed by J Sharp, T Hale, A Billis; comments to author 27.05.24; revised version received 09.07.24; accepted 23.08.24; published 18.10.24.

Please cite as:
Miller LM, Kaye J, Lindauer A, Au-Yeung WTM, Rodrigues NK, Czaja SJ
Remote Passive Sensing of Older Adults' Activities and Function: User-Centered Design Considerations for Behavioral Interventions Conducted in the Home Setting
J Med Internet Res 2024;26:e54709
URL: https://www.jmir.org/2024/1/e54709
doi:10.2196/54709
PMID:39423003

XSL•FO
RenderX

©Lyndsey M Miller, Jeffrey Kaye, Allison Lindauer, Wan-Tai M Au-Yeung, Nathaniel K Rodrigues, Sara J Czaja. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Gaming-Based Tele-Exercise Program to Improve Physical Function in Frail Older Adults: Feasibility Randomized Controlled Trial

Lakshmi Kannan[1], PhD; Upasana Sahu[1*], MS; Savitha Subramaniam[1*], PhD; Neha Mehta[1*], MS; Tanjeev Kaur[2*], MD; Susan Hughes[3*], PhD; Tanvi Bhatt[1*], PhD

[1]College of Applied Health Sciences, University of Illinois at Chicago, Chicago, IL, United States

[2]College of Medicine, Geriatrics, University of Illinois at Chicago, Chicago, IL, United States

[3]Center on Health Research on Aging, School of Public Health, University of Illinois at Chicago, Chicago, IL, United States

[*]these authors contributed equally

Corresponding Author:
Tanvi Bhatt, PhD
College of Applied Health Sciences
University of Illinois at Chicago
1919, W Taylor St (M/C 898)
Chicago, IL, 60612
United States
Phone: 1 3123554443
Email: tbhatt6@uic.edu

Abstract

Background: Frailty leads to reduced physical activity can cause increased fall risk. This contributes to accelerated aging processes, leading to adverse health outcomes and reduced quality of life. We have developed and piloted the design, usability, safety, and feasibility of a gaming-based cognitive-motor (CogXergaming) tele-exercise protocol in prefrail older adults.

Objective: This pilot randomized control trial tested preliminary feasibility and effectiveness of the CogXergaming telehealth protocol for improving physical function.

Methods: Community-dwelling, prefrail older adults were randomly assigned to CogXergaming (n=13) or a control group (n=14). The CogXergaming group received supervised tele-exercises in a gaming format for 6 weeks (3 sessions per week) comprising 18 sessions lasting 90 minutes each. Control group participants participated in a Matter of Balance (MOB), an 8-week, once-a-week structured 90-minute tele-session that has been shown to reduce the fear of falling and increase physical activity. Feasibility of training was obtained by computing the median duration of training sessions for the CogXergaming group. Effectiveness was assessed using dynamic balance control (Four Square Step Test), subjective self-efficacy (Activities-Specific Balance Confidence scale), gait function (Tinetti Performance Oriented Mobility Assessment), muscle strength (30-second chair stand test), and endurance (2-minute step in-place test).

Results: Of the 45 participants enrolled in the study, 4 participants from CogXergaming group and 5 from MOB group lost contact after signing the consent form and did not receive their respective intervention. Eighteen participants were randomized to each group. In the CogXergaming group, 15 (83%) completed the intervention, with 3 (16%) dropping out in the first week. In the MOB group, 16 (88%) completed the program, with 2 (11%) withdrawing during the first week. In addition, there was a significant time group interaction for Four Square Step Test ($F_{1,21}$=5.55, P=.03), Tinetti Performance Oriented Mobility Assessment ($F_{1,25}$=4.16, P=.05), and 30-second chair stand test ($F_{1,21}$=5.06, P=.03), with a significant improvement in these measures for the CogXergaming group post training, compared with no change observed in the MOB group.

Conclusions: These pilot findings indicate that CogXergaming is feasible and applicable in prefrail older adults. Such game-based protocols can be beneficial in improving physical function among community-dwelling, prefrail older adults, however, the efficacy of such training requires further investigation.

Trial Registration: ClinicalTrials.gov NCT04534686; https://clinicaltrials.gov/study/NCT04534686

(J Med Internet Res 2024;26:e56810) doi:10.2196/56810

XSL•FO
RenderX

KEYWORDS

exergame training; Matter of Balance; MOB; pre-frail; tele-exergame; tele-rehabilitation; gaming-based; tele-exercise; physical function; frailty; older adults; aging; physical activity; dementia; CogXergaming; telehealth; dynamic balance

Introduction

Frailty is an age-associated biological syndrome with a prevalence of about 4% to 16% of community-dwelling older adults aged 65 years or older in the United States [1-5]. It significantly affects reserves needed for physical, cognitive, psychological, and social function through loss of muscle mass and weakness, increased fatigue and exhaustion, reduced gait speed, unintended weight loss, reduced physical activity, and cognitive impairment. Specifically, frail, and prefrail older adults are respectively at 3- and 1.36-times higher risk of falling compared with healthy older adults. Consequences of these falls like fractures, head injuries, etc, result in higher health care use and costs, and frequent hospital admissions. All these can further debilitate physical functioning; ultimately resulting in long-term disability overall burdening the health care system [6-11]. Although older adults who are prefrail do not experience these consequences, prefrailty presents an opportunity to reverse or avoid the conversion to frailty [1,4,12]. Thus, developing targeted fall prevention programs for prefrail individuals could mitigate frailty progression and alleviate health care burdens.

Conventional fall prevention programs like Matter of Balance (MOB) [13], Otago [14], Stepping on [15], Tai Chi [16], Stay Active Independent for Life [17], Young Men's Christian Association moving for better balance [18], FallScape [19], Capable [20], and FallTalk [21] have demonstrated promising effects on reducing fall risk among older adults. These programs have been virtually translated to overcome geospatial barriers, making them more accessible to the community [22]. In addition, they are specifically designed to educate older adults about falls and share methods to prevent falls using exercises specific to balance control [23], muscle strength [24], flexibility [25] which, together, can reduce fear of falling [26], and increased balance confidence [27]. For example, the MOB program has demonstrated adaptability and long-term implementation effects while being cost-effective with robust beneficial effects among ambulatory older adults. Studies have reported significant improvement in balance control [26]; gait speed [28]; and reduced fear of falling [29,30], number of falls [13], number of fallers [31], and hospital admissions. However, the program requires a certified trainer to administer it, which limits its use in daily practice. Second, current fall risk reduction programs do not address cognitive function, which is often associated with reduced balance control and higher fall risk among older adults [32-34]. Therefore, there is a need to design and test fall prevention programs for prefrail older adults that are more accessible and cover all aspects of fall risk, targeting both physical and cognitive functions.

Alternate forms of therapy like exergaming (ie, the use of games for exercise) interventions have demonstrated beneficial effects mitigating fall risk factors, as well as improving physical and cognitive function [35-37]. Exergames provides immediate biofeedback (ie, visual, sensory, and auditory) that keeps participants engaged and motivated and thus, improves therapy adherence. Commercial devices like Microsoft Kinect, Nintendo (Wii), Sony PlayStation Move, etc, are often used to access exercise regimens in that use of dance, sport-like activities, balance board activities, and upper extremity movements without the need to leave home [38,39]. However, due to the rapid advancements of commercial gaming devices, they may not be accessible, and are not user friendly for older adult populations, and target only younger adults. As a result, older adults have increasingly relied on web- or tele-based fitness applications (ie, YouTube [Google] videos, physical activity applications) for maintaining physical activity levels while at home [40,41]. However, web-based, and tele-health programs vary significantly with respect to the type, intensity, duration, and frequency of exercise followed resulting in heterogeneity of interventions. Finally, there are no standardized assessment methods to determine adherence and feasibility to web-based programs. Our previous work administered a tele-based assessment and training method called CogXergaming, that is safe to deliver to older adults [42]. The training involves remotely monitoring older adults who mimic a combination of exergame videos (noninteractive) that targets training of balance control, endurance, cognitive-motor function, and mobility [42]. The observed benefits suggest initial safety and feasibility implications in older adults. Adopting a prudent, stepwise approach to assessing the feasibility of a similar training paradigm in prefrail older adults is essential for minimizing potential harm to frail individuals. This approach will ensure rigorous evaluation to establish the safety and effectiveness of interventions before broader implementation in higher-risk populations.

Therefore, this pilot used a NIH (National Institutes of Health) stage 1 model, randomized control trial to investigate the feasibility and effectiveness of a noninteractive, virtually delivered (through Zoom; Zoom Video Communications) exergame (ie, CogXergaming) compared with virtually delivered (also through Zoom) MOB on physical function (ie, balance control assessed through Four Square Step Test (FSST), self-efficacy through Activities-Specific Balance Confidence (ABC) scale, gait function through Tinetti Performance Oriented Mobility Assessment (POMA), muscle strength through 30-second chair stand test (30-CST), and endurance through 2-minute step test) among prefrail older adults. We hypothesized that this pilot study will be feasible and that participants will demonstrate adherence to the CogXergaming program. In addition, the study will explore between-group (CogXergaming vs MOB) differences to gather preliminary data on the comparative effectiveness of the 2 interventions.

Methods

Participants

This NIH stage 1, pilot study involved 50 community-dwelling older adults older than 60 years interested in the study with 45 older adult enrolled after virtually obtaining a written informed

XSL•FO
RenderX

consent (through Docusign). The study was registered on ClinicalTrials.gov (#NCT04534686). Participants were recruited from rehabilitation clinics and senior housing sites in the Chicago metropolitan area.

Ethical Considerations

The institutional review board (IRB) of the University of Illinois at Chicago approved the study (#2020-0280).

Participant Eligibility

Prefrail older adults were recruited. They were characterized as prefrail if at least 2 out of the 3 criteria were present: (1) self-reported unintentional weight loss (>10 lbs) in the last 1 year [1], (2) self-reported general fatigue >3 on rate of perceived exertion (ie, according to Borg's scale ranging from 1-10 with 10 being the maximum perceived exertion) at rest [1], and (3) reduced physical activity levels (ie, exercise less than 3 hours a week) [1]. These criteria are recognized as preliminary indicators of prefrailty [43,44] and were selected as feasible measures for remote screening, aligning with the study's practical considerations and the goal of targeting prefrail individuals for our intervention. Participants were required to possess a laptop or tablet or personal computer with stable internet provision. This decision was made to focus on the need for larger screens that could effectively support the tasks and activities involved in our study. In addition, participants must be able to walk more than 30 feet, stand without an assistive device for at least 5 minutes (length of an exergame was approximately 1-3 minutes), and understand English. Participants were excluded if they self-reported any acute and uncontrolled chronic medical illnesses such as cardiovascular or pulmonary conditions; were recently hospitalized (<3 months); or had a diagnosed neurological condition such as a stroke or Parkinson disease. They were also excluded if their weight exceeded 100 kg and if they were using sedatives or undergoing psychiatric treatment that could affect their cognitive and physical performance during training sessions.

Out of 50 OA who met the inclusion criteria, 45 were included in the study and signed the consent form. Participants were randomized to either of the training groups (n=22 in CogXergaming and n=23 in MOB) by flipping a coin by a research coordinator. This is a simple and transparent randomization technique based on the outcome of a coin toss. A specific research assistant was assigned to administer the CogXergaming intervention, while a different research assistant, trained in delivering the MOB training, administered the MOB intervention. These research assistants did not interact with participants assigned to the other intervention to prevent any potential bias in intervention delivery. However, we also had attrition (n=9) between consenting and start of intervention due to losing contact with participants (unreachable, phone number changed, changed mind, and not interested).

Research Design

Tele-Based Training: CogXergaming

Safety was a priority; therefore, the researcher inspected the environment through the Zoom call and asked the participant to clear away any objects that could potentially cause injury. The researcher instructed the participant to choose a room with good lighting. Besides the break that automatically occurred between subsessions, additional rest breaks were provided upon participant requested ones.

Participants were trained one-on-one for 90 minutes per day and 3 times per week for 6 weeks (for a total of 18 sessions). The intervention was comprised of 4 subsessions that sequentially targeted balance control, endurance, cognitive-motor function, and mobility. Balance control involved mimicking Wii-fit games such as "Super Hula Hoop," "Boxing," "Basic Step," and "Soccer Head." Endurance involved mimicking Just Dance videos: "Price Tag," "Party Rock Anthem," "I Was Made for Loving You," and "Dynamite." Cognitive-motor function involved responding to customized games: "Arithmetic Stepping," "Finding the Bar," "Stepping Colors," and "Paired Association Task." The mobility domain involved mimicking certain yoga and tai-chi movements. Participants were instructed to mimic each video shown to them while maintaining the pace. The first and last 10 mins of each session were comprised of warm-up and cool-down exercises. The researcher shared their screen, and participants were required to have their video turned on to enable the researcher to observe their performance.

The length of the sessions remained on an average 85-90 minutes. Over the training weeks, participants needed lesser rest breaks and were able to progress from light to moderate to more intensive games (challenging balance control, ie, progression from standing dynamic weight shifting to stepping in place to diagonal stepping to side shuffling). More details of this method have been previously published and have shown feasibility [42].

Matter of Balance

All sessions were conducted on Zoom in compliance with HIPAA (Health Insurance Portability and Accountability Act). Participants were trained in groups of 10. Each session lasted 120 minutes and was performed 1 time per week for 8 weeks. The researcher administering MOB was certified to perform MOB training. Each MOB session included a discussion of practical coping strategies, group problem-solving, home safety evaluation strength exercises of the lower limb, coordination exercises, and standing balance exercises [26].

Tele-Assessment

All tele-based assessment tests have shown to be feasible and were described in our previous study [42]. For the current study, we used time taken to complete the FSST with lower numbers (in seconds), indicating better performance to assess balance control [45]. Subjective self-efficacy questionnaire was assessed by the ABC scale with higher values indicating higher confidence levels [46]. Gait function was assessed using the Tinetti POMA with higher scores indicating better performance [47,48]. To assess muscle strength, we used the number sit-to-stands on the 30-CST with higher values indicating greater lower-limb muscle strength [49]. Endurance was assessed using the total number of steps completed on the 2-minute step test with higher numbers indicating better endurance [50].

XSL•FO
RenderX

Statistical Analysis

All analysis was performed using SPSS (version 24; IBM Corp). To assess demographic differences, an independent *t* test with Bonferroni correction for all characteristics except gender. A chi-square test of independence for gender was conducted. It should be noted that this pilot trial was designed to primarily assess the feasibility and safety of a tele-exercise intervention (CogXergaming) in prefrail older adults. The study was not powered for formal efficacy analyses due to the limited sample size and exploratory nature of the investigation. A 2×2 repeated-measure ANOVA was performed to determine the effect of time (before and after training), group (CogXergaming vs MOB), and the time group interaction on all tele-assessments (ie, balance control and confidence, gait function, muscle strength, and endurance). Finally, median duration of training sessions for the CogXergaming group was computed to address feasibility of training. Significant main effects were resolved by post hoc tests with Bonferroni correction with significance level =.05.

Results

Demographics

Participant descriptive characteristics are shown in Table 1. No significant group differences in participant characteristics were observed (P>.05).

Table 1. Participant demographic characteristics for both groups.

Demographics	CogX[a] group (n=13; 6 male and 7 female), mean (SD)	MOB[b] group (n=14; 4 male and 10 female), mean (SD)	*P* value
Age (years)	72.07 (4.75)	71.14 (4.76)	0.61
Weight (kg)	72.64 (15.95)	70.25 (13.78)	0.94
Height (m)	1.61 (0.10)	1.68 (0.06)	0.65
BMI	27.83 (6.05)	25.07 (5.6)	0.29
Years of education	16.15 (2.91)	15.5 (2.53)	0.58

[a]CogX: CogXergaming.

[b]MOB: Matter of Balance.

Use of Tele-Based Assessment and Training

Participants possessed either a tablet or laptop to participate in the intervention activities. However, the specific data on the distribution between tablet or laptop usage among participants were not obtained. While acknowledging the importance of reporting device distribution for comprehensive study outcomes, our focus was primarily on the suitability of tablet or laptop devices to achieve the study objectives effectively.

Of the 45 participants enrolled in the study, 4 participants from CogXergaming group and 5 participants from MOB group lost contact after signing the consent form and did not receive their respective intervention (Figure 1). This primarily was a recruitment failure. Both the tele-based assessment and training were performed through the HIPAA-compliant Zoom application. An individualized Zoom link was sent to each CogXergaming participant for their assessments and training, and a common Zoom link was sent individually for MOB group sessions to protect participants' personal information. All participants were provided with detailed instructions over the phone, guiding them through the process of connecting to the Zoom platform. This approach facilitated seamless participation in virtual sessions, with no notable issues encountered during setup. All assessment and the training sessions were recorded for postsession analyses.

XSL•FO
RenderX

Figure 1. Research design.

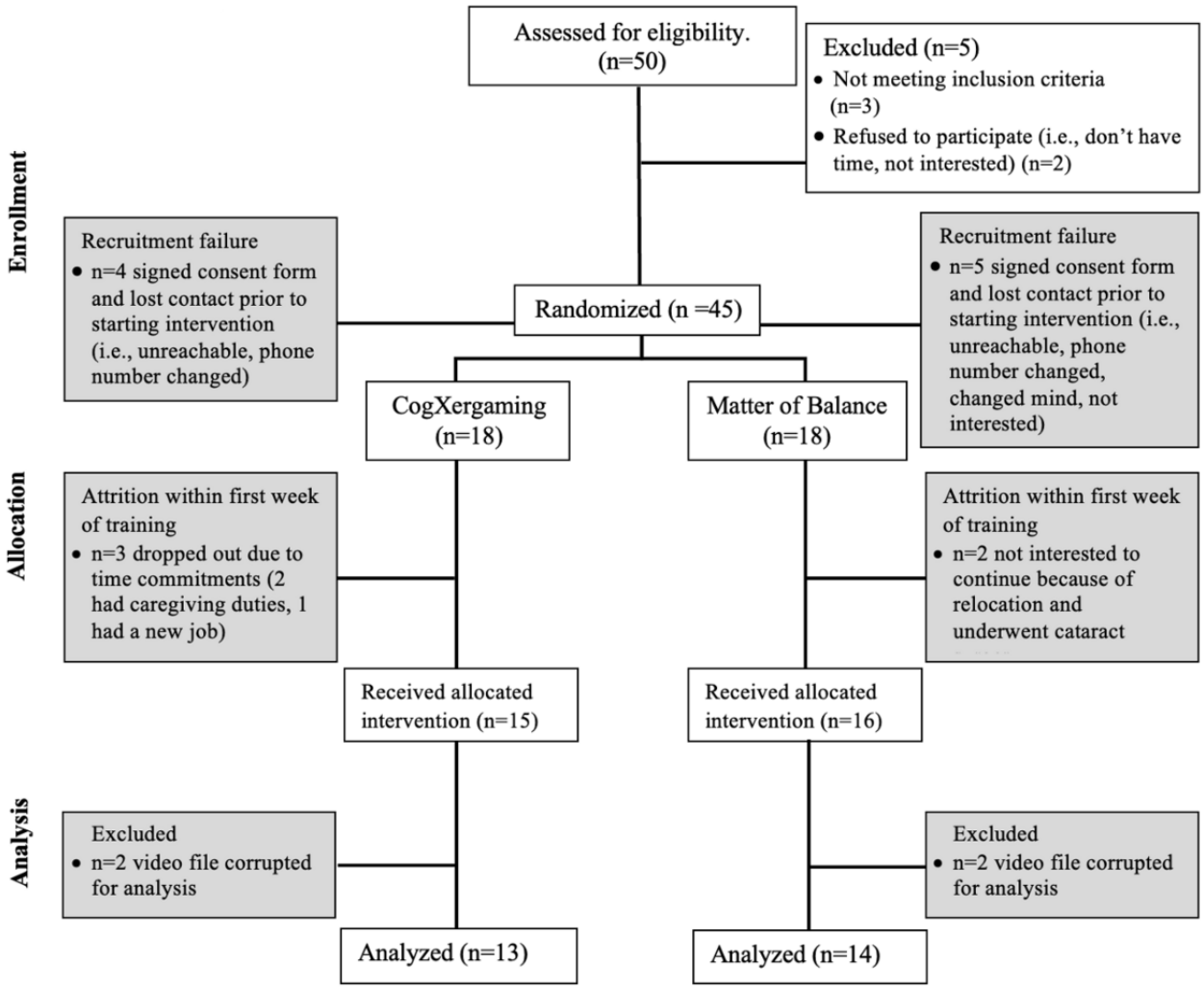

Feasibility

Of the 18 people who were randomized to the CogXergaming group, 15 (83%) successfully completed the intervention. All the attrition from participation in the treatments tested occurred within the first week of both programs. During the first week of CogXergaming, 3 (16%) participants withdrew or dropped out due to lack of time and commitment (eg, got a new job, caregiving duties). Similarly, of the 18 people randomized to MOB, 16 (88%) completed the program. Specifically, 2 people (11%) were not interested in continuing after attending the first session of the MOB group. We did not experience attrition from treatment in either group after participants completed the first week of training.

Effect of Training on Dynamic Balance

For the FSST, the 2×2 repeated measures ANOVA showed a significant main effect of time ($F_{1,21}$=19.14, P<.001) and group time interaction ($F_{1,21}$=5.55, P=.03) on the time taken to complete the test. However, no group effect was observed ($F_{1,21}$=0.33, P=.57). Resolving for the significant main effect of time and the interaction, paired within group t tests revealed a significant reduction in time taken to complete the FSST in the CogXergaming group (t_{10}=4.84, P<.001) post training, however, no significant reduction of time was observed in the MOB group (t_{11}=1.41, P=.18; Figure 2A).

XSL•FO
RenderX

Figure 2. Effect of CogXergaming and Matter Of Balance training in prefrail older adults on balance control and gait function. Figures represent mean (SDs) of pre- to posttraining changes on (A) four square step test, (B) activities-specific balance confidence scale, and (C) Tinetti performance-oriented mobility assessment. Significantly greater performance is indicated by *P<.05, **P<.01, ***P<.001.

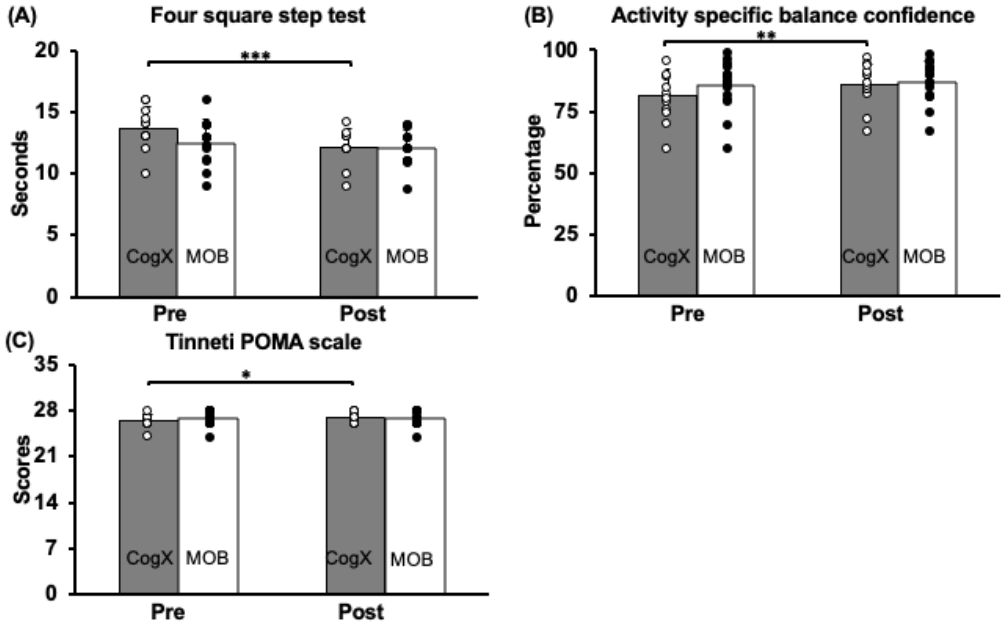

Effect of Training on Subjective Self-Efficacy

For the ABC, the 2×2 repeated measures ANOVA showed a significant main effect of time ($F_{1,25}$=12.03, P<.01) on the balance confidence score. However, no significant group time interaction ($F_{1,25}$=3.05, P=.09) or group effect ($F_{1,25}$=0.65, P=.42) was observed. Resolving for main effect of time, a paired t test showed a significant improvement in balance confidence score within the CogXergaming group (t_{12}=–2.92, P=.01), but no improvement within the MOB group (t_{13}=–1.75, P=.10) was observed (Figure 2B).

Effect of Training on Gait Function

The 2×2 repeated measures ANOVA showed a significant main effect of time ($F_{1,25}$=7.77, P=.01) and significant group time interaction ($F_{1,25}$=4.16, P=.05) on the POMA score. However, we did not see a group effect ($F_{1,25}$=0.46, P=.5). Resolving for main effect of time and interaction, the paired t test showed a significant improvement in the POMA score within the CogXergaming group (t_{12}=–2.52, P=.03) but no within group improvement in the MOB group (t_{13}=–1.0, P=.33) was observed (Figure 2C).

Effect of Training on Lower Limb Muscle Strength

The 2×2 repeated measures ANOVA showed a significant main effect of time ($F_{1,21}$=16.45, P<.001) and significant group time interaction ($F_{1,21}$=5.06, P=.03) indicating that both groups performed an increased number of sit-stands on the 30-CST. However, there was no significant main effect of group ($F_{1,21}$=0.10, P=.75). Resolving for main effects and interaction, the paired t test showed a significant increase in the number of sit-to-stands on the 30-CST within the CogXergaming group (t_{10}=–3.73, P<.01); however, there was no significant within group improvement observed in the MOB group (t_{11}=–1.6, P=.13; Figure 3A).

Figure 3. Effect of CogXergaming and Matter Of Balance training in prefrail older adults on lower limb muscle strength and endurance. Figures presented are mean (SDs) of pre- to posttraining scores on (A) 30-second chair stand test (30-CST) and (B) 2-minute step test for both groups. Significantly greater performance is indicated by *P<.05, **P<.01, ***P<.001.

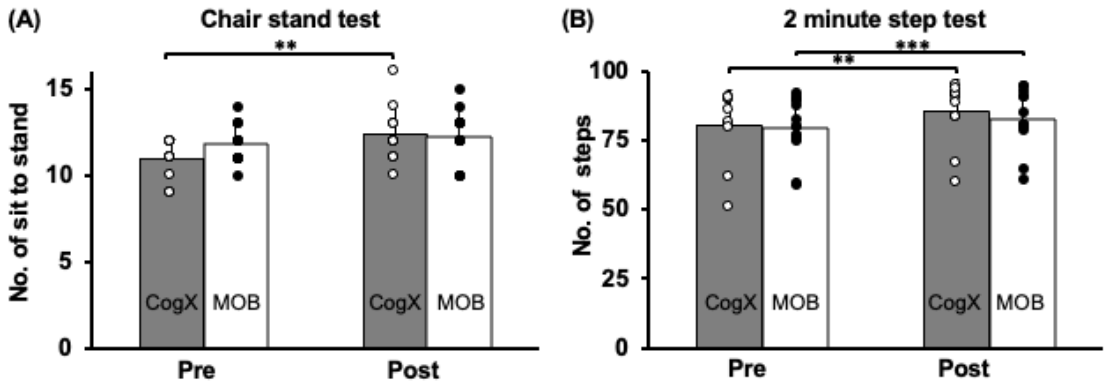

XSL•FO
RenderX

Effect of Training on Endurance

The 2×2 repeated measures ANOVA of the 2-minute Step Test scores showed a significant main of effect of time ($F_{1,21}$=37.22, P<.001) with no significant group time interaction ($F_{1,21}$=2.57, P=.12) or group effect ($F_{1,21}$=0.26, P=.61) on number of steps completed. Resolving for main effect of time, paired t tests revealed a significant improvement in number of steps on this measure within the CogXergaming group (t_{10}=–3.97, P<.01), and the MOB group (t_{11}=–6.23, P<.001, Figure 3B).

Discussion

Principal Findings

This pilot randomized controlled trial investigated the feasibility and effectiveness of a tele-based, noninteractive CogXergaming training compared with tele-based MOB program on physical function in prefrail older adults. The results showed that the intervention was feasible as shown by tolerability of the intervention with a lack of any expected and unexpected intervention-induced adverse events. The intervention also showed promise of being effective as seen by the significantly and differentially greater improvements in physical function within the CogXergaming group. Specifically, we observed significant and differential within group improvements within the CogXergaming group in balance control and confidence, gait function, and lower limb muscle strength. While both groups showed significant improvement in endurance, the benefits were greater among the MOB group. These results suggest that the tele-based, noninteractive CogXergaming intervention has the potential to improve physical function among prefrail older adults.

Feasibility

The significant and differential improvements observed in the CogXergaming group in 3 of the 4 outcomes assessed underscore the potential of mimicking video-based interventions to enhance physical function among prefrail older adults. The time demand posed a notable challenge in our study. Notably, participant dropout rates observed largely attributed to personal commitments among both groups. Specifically, the CogXergaming intervention necessitated a higher frequency, occurring thrice weekly over a span of 6 weeks, in contrast to the MOB program, which occurred once weekly over 8 weeks. While the MOB intervention underscores the importance of exercise adherence among prefrail older adults, it relies on the presence of a health coach. Conversely, the CogXergaming intervention offers participants the flexibility to engage at their convenience, potentially mitigating scheduling conflicts. However, the implications of unsupervised engagement on compliance remain uncertain and may require further investigation. In addition, the physiological underpinning of this approach remains to be explored to understand the mechanisms of actions that confer benefits.

The CogXergaming group had high adherence, similar to the MOB group, demonstrating its feasibility as a remote intervention. Additionally, both groups experienced similar levels of attrition, which occurred only during the first week of training. This early dropout may be due to the prefrail characteristics of the participants. CogXergaming, with its gamified approach, may offer additional engagement to participants and flexibility to clinicians for tailorizing the intervention. The challenge of moving from a sedentary lifestyle, compounded by psychological factors such as fatigue and depression, may impede even highly motivational exercise programs. The intensity and dosage of exercises in the CogXergaming group, if increased, could potentially help maximize the benefits. An enhancement of the program could address essential components of physical function, including balance control, muscle strength, endurance, and cognitive function. In addition, few technical challenges were encountered, as all prefrail older adults successfully connected to the Zoom link. This success may be attributed to the step-by-step guidance provided during phone calls. These findings support the implementation of this approach in future studies and clinical applications. However, implementing a similar training approach on a larger scale could enhance our ability to interpret qualitative responses more effectively. With emerging exergaming devices commercially, the development of innovative strategies that surmount geospatial barriers is crucial in facilitating physical function in prefrail older adults. The availability of these approaches could promote home-based, cost-effective, user-friendly methods of intervention delivery that improve accessibility and adaptability in the community.

Effect of Training on Dynamic Balance, Subjective Self-Efficacy, and Gait Function

A time effect and a group time interaction was observed in the FSST, ABC, and POMA scale with the CogXergaming group displaying greater improvement in balance control post training compared with the MOB group. Specifically, FSST and POMA measures the temporal characteristics of adjusting postural stability during forward, backward, and side stepping. This task requires lower limb strength, ankle stability, hip and trunk control, core strength (in the abdominal and lower back muscles), and sensory system function (ie, proprioception, visual, and vestibular). The series of subsessions involved during a CogXergaming session promoted weight shifts, stepping, trunk rotations, and short durations of standing on 1 leg within a fixed time limit. These activities were holistic while being task-specific, optimizing motor and interlimb coordination with specific movements, promoting quick multidimensional weight shifts. Consistent practice of these movements over 18 sessions may have been enough to promote motor learning and skill retention that could have attributed to the reduced time required to complete FSST and POMA. In addition, in our study, the participants engaged in CogXergaming within a familiar environment, that is, their own home. This environmental familiarity with each game to be played within a fixed time limit could have contributed to increased comfort levels during training, enabling participants to focus more on mastering specific motor movements such as forward, backward, and side stepping. Previous studies have explored the concept of perceptual-motor recalibration in familiar environments [51,52]. These studies report that when individuals are in a familiar environment, the sensory demands from visual and vestibular systems, relying more on proprioception input and thus,

XSL•FO
RenderX

prioritizing sensory resources toward maintaining postural stability, are reduced. However, there was no prominent reduction of spatiotemporal gait characteristics among the MOB group. The MOB fall prevention education program was also provided to participants in their homes. However, MOB participants may not have practiced exercises consistently as they were advised to leading them to experience no change in physical activity levels. In addition, it is noteworthy that despite the insignificant differences between the groups at baseline, the MOB group displayed potential ceiling effects for the FSST, ABC, and POMA. This suggests that participants in the MOB group may have started the intervention at a higher baseline level of function, potentially limiting observable improvements on the scales used. However, given our sample size limitations, definitive conclusions regarding a ceiling effect cannot be drawn at this stage.

Effect of Training on Lower Limb Muscle Strength

There was a main effect of time and an interaction effect between time by group such that the CogXergaming group showed a significant increase in the number of sit-to stands on the 30-CST, demonstrating greater muscle strength in lower limbs, compared with no change on this outcome within the MOB group. The CogXergaming training involved mimicking complex motor movements shown in videos of tai chi, yoga, and dance [42]. These movements predominantly involved moving the lower limb in multiple directions and maintaining postural stability while moving the upper extremity. This approach could have reinforced the ability of prefrail older adults to consistently participate on the lower extremity during training for 18 sessions. Since prefrail older adults exhibit muscle weakness, these complex activities provided a medium to gradually improve muscle strength, which helped CogXergaming participants to perform better post training. On the other hand, since MOB group notably focused on psychosocial aspects of educating older adults about prevention of falls, its ability to improve muscle strength may have been limited. MOB was designed to promote a holistic understanding of fall prevention by encouraging participants to consider physical capabilities, balance control, environmental awareness, and the importance of physical activity. Furthermore, the higher number of sit-to-stands observed in the MOB group could suggest a potential ceiling effect among participants, indicating that they may have started the intervention at a higher level of functioning compared with participants in the CogXergaming group. However, the long-term effects of MOB on muscle strength need to be explored further in prefrail older adults.

Effect of Training on Endurance

Despite no time group interaction effect, there was a time effect with both groups that displayed significant improvement in endurance on the 2-minute step test. The CogXergaming intervention was designed to mimic exergames focused on several stepping and motor movements performed at a medium-paced intensity that could have attributed to the improvement on the 2-minute step test. The incorporation of dance movements, cognitive-motor function training, and balance control exercises during the exergaming training sessions also targeted cognitive demands on motor performance. This feature could have inadvertently increased the overall pace of exercises performed across the sessions, particularly as participants tried to complete various movements within stipulated time limits, placing an added emphasis on quick movement completion that potentially elevated heart rates. The consistent performance of activities under such conditions over 18 sessions likely facilitated participants' adaptation to performing exercises at a specific intensity. Notably, the posttraining improvement observed in the number of steps completed within 2 minutes aligns with the practiced activity of quickly completing the task to mimic the training. These results indicate that it is possible to improve aerobic capacity merely by mimicking and matching movements at the intensity demonstrated in the videos for prefrail older adults. Furthermore, the selection of music with a specific pace for the dance component in the CogXergaming group could have induced significant improvements in endurance. The finding that the endurance gains in the MOB group were greater than those seen in the CogXergaming group in the same test performance, through education alone, is noteworthy. While participants were not monitored for their physical activity levels post training, this finding indicates that more studies are needed to investigate the lasting impact of education on physical activity levels in this population.

Limitations

Apart from the benefits observed, acknowledging the limitations of the study is essential. First, the relatively low sample size in our pilot study meant that the study did not have the statistical power to enable valid between group efficacy testing. The differential within-group improvements on 3 of the outcomes is promising and warrants follow-up in a fully powered future. The nonequivalence between the groups in our study may have influenced the outcomes observed. However, since this was a pilot study aimed at evaluating the feasibility of a noninteractive exergaming training, future research efforts may benefit from designing equivalent-based training programs to determine the efficacy of the intervention more accurately. Second, caution must be used in generalizing the observed results to other populations. Higher sample size in future studies may provide more robust and representative data, particularly in terms of the scope and methods used to identify prefrail individuals. Third, the intervention implicitly challenged cognitive function but did not include comprehensive examination of cognitive aspects associated with increased risk of falling (such as reaction time, working memory, and executive function). Further exploration of these additional outcomes is therefore warranted. Finally, although the study included prefrail older adults, we did not differentiate between physical frailty, cognitive frailty, or both. Future investigations may benefit from classifying participants based on these distinct frailty profiles to assess intervention effects more powerfully. In conclusion, while our NIH stage 1 pilot study provides valuable insights into the feasibility of a noninteractive exergaming intervention, future research efforts should address these limitations to advance our understanding of the intervention's efficacy and generalizability.

XSL•FO
RenderX

Conclusion

This pilot randomized control trial provides preliminary findings that the 6-week CogXergaming intervention has the potential to improve physical function among prefrail older adults. Specifically, the results show differential within group significant improvements in balance control, self-reported balance confidence, and muscle strength after participation in CogXergaming. Effects were also seen in endurance, but they were not as large. These findings indicate that CogXergaming could potentially help reduce the risk of falls while promoting physical activity among prefrail older adults. Future adequately powered studies of this promising program are needed to demonstrate efficacy among prefrail older adults.

Acknowledgments

We thank the members of the Cognitive Motor Balance Rehabilitation laboratory for assisting with study recruitment and procedure administration. This work was supported by the National Institute on Aging of the National Institutes of Health (NIH; grant R24AG064191). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.2).
[PDF File (Adobe PDF File), 95 KB - jmir_v26i1e56810_app1.pdf]

References

1. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar;56(3):M146-M156. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
2. Kiely DK, Cupples LA, Lipsitz LA. Validation and comparison of two frailty indexes: The MOBILIZE Boston Study. J Am Geriatr Soc 2009 Sep;57(9):1532-1539 [FREE Full text] [doi: 10.1111/j.1532-5415.2009.02394.x] [Medline: 19682112]
3. Bandeen-Roche K, Xue QL, Ferrucci L, Walston J, Guralnik JM, Chaves P, et al. Phenotype of frailty: characterization in the women's health and aging studies. J Gerontol A Biol Sci Med Sci 2006 Mar;61(3):262-266. [doi: 10.1093/gerona/61.3.262] [Medline: 16567375]
4. Woods NF, LaCroix AZ, Gray SL, Aragaki A, Cochrane BB, Brunner RL, Women's Health Initiative. Frailty: emergence and consequences in women aged 65 and older in the Women's Health Initiative Observational Study. J Am Geriatr Soc 2005 Aug;53(8):1321-1330. [doi: 10.1111/j.1532-5415.2005.53405.x] [Medline: 16078957]
5. Cawthon PM, Marshall LM, Michael Y, Dam TT, Ensrud KE, Barrett-Connor E, Osteoporotic Fractures in Men Research Group. Frailty in older men: prevalence, progression, and relationship with mortality. J Am Geriatr Soc 2007 Aug;55(8):1216-1223. [doi: 10.1111/j.1532-5415.2007.01259.x] [Medline: 17661960]
6. Trombetti A, Reid KF, Hars M, Herrmann FR, Pasha E, Phillips EM, et al. Age-associated declines in muscle mass, strength, power, and physical performance: impact on fear of falling and quality of life. Osteoporos Int 2016 Feb;27(2):463-471 [FREE Full text] [doi: 10.1007/s00198-015-3236-5] [Medline: 26194491]
7. Deary IJ, Corley J, Gow AJ, Harris SE, Houlihan LM, Marioni RE, et al. Age-associated cognitive decline. Br Med Bull 2009;92:135-152. [doi: 10.1093/bmb/ldp033] [Medline: 19776035]
8. Liang CK, Chu CS, Chen LK. Impacts of loneliness and social isolation on healthy longevity in older adults. In: Loneliness and Social Isolation in Old Age. England, UK: Routledge; 2023.
9. Eldadah BA. Fatigue and fatigability in older adults. PM R 2010 May;2(5):406-413. [doi: 10.1016/j.pmrj.2010.03.022] [Medline: 20656622]
10. Lisko I, Kulmala J, Annetorp M, Ngandu T, Mangialasche F, Kivipelto M. How can dementia and disability be prevented in older adults: where are we today and where are we going? J Intern Med 2021 Jun;289(6):807-830 [FREE Full text] [doi: 10.1111/joim.13227] [Medline: 33314384]
11. Arrighi HM, Gélinas I, McLaughlin TP, Buchanan J, Gauthier S. Longitudinal changes in functional disability in Alzheimer's disease patients. Int Psychogeriatr 2013 Jun;25(6):929-937. [doi: 10.1017/S1041610212002360] [Medline: 23406898]
12. Blaum CS, Xue QL, Michelon E, Semba RD, Fried LP. The association between obesity and the frailty syndrome in older women: the Women's Health and Aging Studies. J Am Geriatr Soc 2005 Jun;53(6):927-934 [FREE Full text] [doi: 10.1111/j.1532-5415.2005.53300.x] [Medline: 15935013]
13. Chen TY, Edwards JD, Janke MC. The effects of the a matter of balance program on falls and physical risk of falls, Tampa, Florida, 2013. Prev Chronic Dis 2015;12:E157 [FREE Full text] [doi: 10.5888/pcd12.150096] [Medline: 26402047]

XSL•FO
RenderX

14. Shubert TE, Smith ML, Ory MG, Clarke CB, Bomberger SA, Roberts E, et al. Translation of the otago exercise program for adoption and implementation in the United States. Front Public Health 2014;2:152 [FREE Full text] [doi: 10.3389/fpubh.2014.00152] [Medline: 25964899]
15. Ory MG, Smith ML, Jiang L, Lee R, Chen S, Wilson AD, et al. Fall prevention in community settings: results from implementing stepping on in three States. Front Public Health 2014;2:232 [FREE Full text] [doi: 10.3389/fpubh.2014.00232] [Medline: 25964924]
16. Logghe IHJ, Verhagen AP, Rademaker ACHJ, Bierma-Zeinstra SMA, van Rossum E, Faber MJ, et al. The effects of Tai Chi on fall prevention, fear of falling and balance in older people: a meta-analysis. Prev Med 2010;51(3-4):222-227. [doi: 10.1016/j.ypmed.2010.06.003] [Medline: 20558197]
17. York SC, Shumway-Cook A, Silver IF, Morrison AC. A translational research evaluation of the Stay Active and Independent for Life (SAIL) community-based fall prevention exercise and education program. Health Promot Pract 2011 Nov;12(6):832-839. [doi: 10.1177/1524839910375026] [Medline: 21191079]
18. Ehrenreich H, Pike M, Hohman K, Kaniewski M, Longjohn M, Myers G, et al. CDC and YMCA: A Promising Partnership for Delivering Fall Prevention Programing. Front Public Health 2014;2:235 [FREE Full text] [doi: 10.3389/fpubh.2014.00235] [Medline: 25964926]
19. Schepens SL, Panzer V, Goldberg A. Randomized controlled trial comparing tailoring methods of multimedia-based fall prevention education for community-dwelling older adults. Am J Occup Ther 2011;65(6):702-709 [FREE Full text] [doi: 10.5014/ajot.2011.001180] [Medline: 22214115]
20. Szanton SL, Thorpe RJ, Boyd C, Tanner EK, Leff B, Agree E, et al. Community aging in place, advancing better living for elders: a bio-behavioral-environmental intervention to improve function and health-related quality of life in disabled older adults. J Am Geriatr Soc 2011 Dec;59(12):2314-2320 [FREE Full text] [doi: 10.1111/j.1532-5415.2011.03698.x] [Medline: 22091738]
21. Panzer V, Burleson JA, Wakefield D, Into F, Wolfson L. Can a multimedia fall prevention treatment program change behavior and prevent falls? Journal of the American Geriatrics Society. , p.S167. ISSN. Journal of the American Geriatrics Society 2008;56(4):S167.
22. Li F, Harmer P, Voit J, Chou L. Implementing an online virtual falls prevention intervention during a public health pandemic for older adults with mild cognitive impairment: a feasibility trial. Clin Interv Aging 2021;16:973-983 [FREE Full text] [doi: 10.2147/CIA.S306431] [Medline: 34079243]
23. Shubert TE, Chokshi A, Mendes VM, Grier S, Buchanan H, Basnett J, et al. Stand Tall-a virtual translation of the Otago exercise program. J Geriatr Phys Ther 2020;43(3):120-127. [doi: 10.1519/JPT.0000000000000203] [Medline: 29958232]
24. Martin A, Candow D. Effects of online yoga and tai chi on physical health outcome measures of adult informal caregivers. Int J Yoga 2019;12(1):37-44 [FREE Full text] [doi: 10.4103/ijoy.IJOY_5_18] [Medline: 30692782]
25. Wehner C, Blank C, Arvandi M, Wehner C, Schobersberger W. Effect of Tai Chi on muscle strength, physical endurance, postural balance and flexibility: a systematic review and meta-analysis. BMJ Open Sport Exerc Med 2021;7(1):e000817 [FREE Full text] [doi: 10.1136/bmjsem-2020-000817] [Medline: 33614126]
26. Mohamed Hassan Saleh N, El-Gilany AH, Noshy Abd El-Aziz Mohamed H, Mahmoud Elsakhy N. Effect of matter of balance program on improving balance and reducing fear of falls among community-dwelling older adults. Egypt J Health Care 2022;13(1):1106-1116 [FREE Full text] [doi: 10.21608/ejhc.2022.233175]
27. Haynes M, League P, Neault G. A matter of balance: older adults taking control of falls by building confidence. Front Public Health 2014;2:274 [FREE Full text] [doi: 10.3389/fpubh.2014.00274] [Medline: 25964938]
28. Reynolds L, Buchanan BL, Alexander JL, Bordenave E. Effectiveness of a matter of balance program within an assisted living community. Physical & Occupational Therapy In Geriatrics 2019 Oct 04;38(1):18-30. [doi: 10.1080/02703181.2019.1673526]
29. Sartor-Glittenberg C, Bordenave E, Bay C, Bordenave L, Alexander JL. Effect of a matter of balance programme on avoidance behaviour due to fear of falling in older adults. Psychogeriatrics 2018;18(3):224-230. [doi: 10.1111/psyg.12310] [Medline: 29424113]
30. Wolfe ES, Arabian SS, Breeze JL, Bugaev N. Evaluating the effectiveness of the translated "A Matter of Balance" fall prevention program materials for non-English-speaking participants. J Trauma Nurs 2018;25(5):311-317 [FREE Full text] [doi: 10.1097/JTN.0000000000000394] [Medline: 30216262]
31. Howland J, Shankar KN, Peterson EW, Taylor AA. Savings in acute care costs if all older adults treated for fall-related injuries completed matter of balance. Inj Epidemiol 2015;2(1):25 [FREE Full text] [doi: 10.1186/s40621-015-0058-z] [Medline: 26457239]
32. Bhatt T, Wang Y, Wang S, Kannan L. Perturbation training for fall-risk reduction in healthy older adults: interference and generalization to opposing novel perturbations post intervention. Front Sports Act Living 2021;3:697169 [FREE Full text] [doi: 10.3389/fspor.2021.697169] [Medline: 34490424]
33. Kannan L, Bhatt T, Zhang A, Ajilore O. Association of balance control mechanisms with brain structural integrity in older adults with mild cognitive impairment. Neurosci Lett 2022;783:136699. [doi: 10.1016/j.neulet.2022.136699] [Medline: 35654247]

XSL•FO
RenderX

34. Kannan L, Bhatt T, Ajilore O. Cerebello-cortical functional connectivity may regulate reactive balance control in older adults with mild cognitive impairment. Front Neurol 2023;14:1041434 [FREE Full text] [doi: 10.3389/fneur.2023.1041434] [Medline: 37139074]
35. Kannan L, Bhatt T. Associations of dual task exergaming with cognitive-motor interference in older adults with mild cognitive impairment: a single-arm pilot study. JAR Life 2021;10:55-61 [FREE Full text] [doi: 10.14283/jarlife.2021.11] [Medline: 36923509]
36. Moreira NB, Rodacki ALF, Costa SN, Pitta A, Bento PCB. Perceptive-cognitive and physical function in prefrail older adults: exergaming versus traditional multicomponent training. Rejuvenation Res 2021;24(1):28-36. [doi: 10.1089/rej.2020.2302] [Medline: 32443963]
37. Zhao Y, Feng H, Wu X, Du Y, Yang X, Hu M, et al. Effectiveness of exergaming in improving cognitive and physical function in people with mild cognitive impairment or dementia: systematic review. JMIR Serious Games 2020;8(2):e16841 [FREE Full text] [doi: 10.2196/16841] [Medline: 32602841]
38. Dermody G, Whitehead L, Wilson G, Glass C. The role of virtual reality in improving health outcomes for community-dwelling older adults: systematic review. J Med Internet Res 2020;22(6):e17331 [FREE Full text] [doi: 10.2196/17331] [Medline: 32478662]
39. Kamnardsiri T, Phirom K, Boripuntakul S, Sungkarat S. An interactive physical-cognitive game-based training system using kinect for older adults: development and usability study. JMIR Serious Games 2021;9(4):e27848 [FREE Full text] [doi: 10.2196/27848] [Medline: 34704953]
40. Kwan RY, Lee D, Lee PH, Tse M, Cheung DS, Thiamwong L, et al. Effects of an mHealth brisk walking intervention on increasing physical activity in older people with cognitive frailty: pilot randomized controlled trial. JMIR Mhealth Uhealth 2020;8(7):e16596 [FREE Full text] [doi: 10.2196/16596] [Medline: 32735218]
41. Stawarz K, Liang IJ, Alexander L, Carlin A, Wijekoon A, Western MJ. Exploring the potential of technology to promote exercise snacking for older adults who are prefrail in the home setting: user-centered design study. JMIR Aging 2023;6:e41810 [FREE Full text] [doi: 10.2196/41810] [Medline: 37223992]
42. Subramaniam S, Kitsiou S, Bhatt T. Remote teleassessment and telerehabilitation of a comprehensive exercise training protocol for older adults: design and methodology of a usability protocol. Internet j allied health sci pract 2023;2(1):1-14 [FREE Full text]
43. Xue QL, Bandeen-Roche K, Varadhan R, Zhou J, Fried LP. Initial manifestations of frailty criteria and the development of frailty phenotype in the women's health and aging study II. J Gerontol A Biol Sci Med Sci 2008;63(9):984-990. [doi: 10.1093/gerona/63.9.984] [Medline: 18840805]
44. Stenholm S, Ferrucci L, Vahtera J, Hoogendijk EO, Huisman M, Pentti J, et al. Natural course of frailty components in people who develop frailty syndrome: evidence from two cohort studies. J Gerontol A Biol Sci Med Sci 2019;74(5):667-674 [FREE Full text] [doi: 10.1093/gerona/gly132] [Medline: 30084927]
45. de Aquino MPM, de Oliveira Cirino NT, Lima CA, de Miranda Ventura M, Hill K, Perracini MR. The four square step test is a useful mobility tool for discriminating older persons with frailty syndrome. Exp Gerontol 2022;161:111699. [doi: 10.1016/j.exger.2022.111699] [Medline: 35066122]
46. Filiatrault J, Gauvin L, Fournier M, Parisien M, Robitaille Y, Laforest S, et al. Evidence of the psychometric qualities of a simplified version of the activities-specific balance confidence scale for community-dwelling seniors. Arch Phys Med Rehabil 2007;88(5):664-672. [doi: 10.1016/j.apmr.2007.02.003] [Medline: 17466738]
47. Yang C, Mo Y, Cao X, Zhu S, Wang X, Wang X. Reliability and validity of the tinetti performance oriented mobility assessment in Chinese community-dwelling older adults. Geriatr Nurs 2023;53:85-89. [doi: 10.1016/j.gerinurse.2023.06.020] [Medline: 37454423]
48. Tinetti ME. Performance-oriented assessment of mobility problems in elderly patients. J Am Geriatr Soc 1986;34(2):119-126. [doi: 10.1111/j.1532-5415.1986.tb05480.x] [Medline: 3944402]
49. Millor N, Lecumberri P, Gómez M, Martínez-Ramírez A, Izquierdo M. An evaluation of the 30-s chair stand test in older adults: frailty detection based on kinematic parameters from a single inertial unit. J Neuroeng Rehabil 2013;10:86 [FREE Full text] [doi: 10.1186/1743-0003-10-86] [Medline: 24059755]
50. Bohannon RW, Crouch RH. Two-minute step test of exercise capacity: systematic review of procedures, performance, and clinimetric properties. J Geriatr Phys Ther 2019;42(2):105-112. [doi: 10.1519/JPT.0000000000000164] [Medline: 29210933]
51. Yardley L, Redfern MS. Psychological factors influencing recovery from balance disorders. J Anxiety Disord 2001;15(1-2):107-119. [doi: 10.1016/s0887-6185(00)00045-1] [Medline: 11388354]
52. Buccello-Stout RR, Bloomberg JJ, Cohen HS, Whorton EB, Weaver GD, Cromwell RL. Effects of sensorimotor adaptation training on functional mobility in older adults. J Gerontol B Psychol Sci Soc Sci 2008;63(5):P295-P300. [doi: 10.1093/geronb/63.5.p295] [Medline: 18818444]

Abbreviations

30-CST: 30-second chair stand test
ABC: Activities-Specific Balance Confidence

XSL•FO
RenderX

FSST: Four Square Step Test
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
MOB: Matter of Balance
NIH: National Institutes of Health
POMA: Performance Oriented Mobility Assessment

Edited by A Dominello, I Kronish, J Kaye, E Wethington; submitted 29.01.24; peer-reviewed by F Van der Ouderaa, B Cedeno-Veloz; comments to author 13.03.24; revised version received 17.05.24; accepted 12.08.24; published 27.11.24.

Please cite as:
Kannan L, Sahu U, Subramaniam S, Mehta N, Kaur T, Hughes S, Bhatt T
Gaming-Based Tele-Exercise Program to Improve Physical Function in Frail Older Adults: Feasibility Randomized Controlled Trial
J Med Internet Res 2024;26:e56810
URL: https://www.jmir.org/2024/1/e56810
doi:10.2196/56810
PMID:

©Lakshmi Kannan, Upasana Sahu, Savitha Subramaniam, Neha Mehta, Tanjeev Kaur, Susan Hughes, Tanvi Bhatt. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Targeting Daily Positive Events to Improve Emotional and Functional Well-Being in Adults With Fibromyalgia: Insights From the LARKSPUR Randomized Controlled Trial

Anthony Ong[1], PhD; Kenneth Wilcox[1], PhD; M Carrington Reid[2], MD; Elaine Wethington[1], PhD; Dakota Cintron[1], PhD; Elizabeth Addington[3], PhD; Selin Goktas[1]; Judith Moskowitz[3], PhD

[1]Department of Psychology, Cornell University, Ithaca, NY, United States
[2]Department of Medicine, Weill Cornell University, NYC, NY, United States
[3]Department of Medical Social Sciences, Northwestern University, Chicago, IL, United States

Corresponding Author:
Anthony Ong, PhD
Department of Psychology
Cornell University
G421 Martha Van Rensselaer Hall
Cornell University
Ithaca, NY, 14850
United States
Phone: 1 6072294119
Email: anthony.ong@cornell.edu

Abstract

Background: Fibromyalgia is a chronic pain condition characterized by widespread musculoskeletal pain, fatigue, and cognitive difficulties, affecting individuals across all age groups. Positive affect (PA) interventions have shown promise in enhancing emotional well-being and pain management in patients with diverse chronic pain conditions. However, the efficacy of internet-delivered PA interventions for individuals with fibromyalgia remains understudied.

Objective: This randomized controlled trial investigated the efficacy of a web-based PA regulation intervention—Lessons in Affect Regulation to Keep Stress and Pain Under Control (LARKSPUR)—in enhancing emotional and functional well-being among adults with fibromyalgia syndrome.

Methods: A total of 95 participants with fibromyalgia syndrome aged 50 years and older (89/95, 94% female) were randomized to one of two fully automated conditions: (1) LARKSPUR (n=49) or (2) emotion reporting/attention control (n=46). At the postintervention and 1-month follow-up time points, participants completed 7 consecutive, end-of-day, web-based reports capturing positive events (PE), pain, fatigue, PA, and negative affect.

Results: Compared to control, LARKSPUR resulted in greater improvements in daily affective responsivity to PE at the postintervention time point, including greater reductions in negative affect (b_L–b_C=–0.06, 95% highest posterior density interval [HPD] –0.10 to –0.02) and increases in PA (b_L–b_C=0.10, 95% HPD 0.02-0.19). Furthermore, across the postintervention and 1-month follow-up time points, LARKSPUR led to greater reductions in pain (b_L–b_C=–0.20, 95% HPD –0.36 to –0.04) and fatigue (b_L–b_C=–0.24, 95% HPD –0.41 to –0.06) following PE.

Conclusions: This randomized controlled trial provides initial evidence that a web-based PA skills intervention can enhance emotional well-being and reduce pain and fatigue in aging adults with fibromyalgia.

Trial Registration: ClinicalTrials.gov NCT04869345; https://clinicaltrials.gov/study/NCT04869345

(J Med Internet Res 2024;26:e54678) doi:10.2196/54678

KEYWORDS

positive affect; chronic pain; chronic condition; long-term pain; positive psychology; positive events; fibromyalgia; mHealth; app; digital technology; digital interventions; gerontology; geriatrics; older adult; aging; well-being

XSL•FO
RenderX

Introduction

Background

Chronic pain is a major public health problem affecting millions of people worldwide and imposing significant burdens on individuals' well-being and quality of life. Among chronic pain conditions, fibromyalgia syndrome (FMS) is one of the most prevalent and challenging to treat. FMS involves widespread musculoskeletal pain, fatigue, and debilitating symptoms that impair physical and psychological functioning [1-3]. While pharmacological treatments provide partial relief for some individuals, nonpharmacologic approaches are often needed to improve well-being [4]. Consequently, innovative interventions are required that address the multifaceted impacts of FMS on daily life.

In recent years, positive psychological interventions have emerged as a promising approach to promote well-being and enhance mental health outcomes [5-7]. Integrating daily positive activities into chronic pain treatment has shown the potential to foster resilience and improve overall well-being [8,9]. Daily positive events (PE)—small, meaningful experiences that increase positive emotions [10,11], such as enjoying a pleasant meal, having a meaningful conversation, or experiencing moments of comfort—may be especially beneficial for individuals with FMS. First, cultivating positive emotions, like joy, gratitude, and contentment, through daily PE could improve emotional well-being, as these tend to be deficient in this population [12,13]. Second, actively seeking and savoring positive experiences may instill a greater sense of control and agency over one's life, counterbalancing feelings of helplessness and hopelessness common in chronic pain [14,15]. Third, pleasant daily activities could facilitate social connection and support, which are essential given individuals with FMS often experience isolation due to their condition [16,17]. Finally, engagement in pleasant activities and PE may facilitate healthy habits and routines, promoting sustained well-being and improvements in daily functioning [18,19].

Internet-delivered positive psychological interventions represent a promising approach to enhancing well-being among those affected by chronic pain, including those with FMS. Web-based programs can overcome barriers to in-person treatment while providing effective tools to boost positive experiences. One such intervention has been widely tested in multiple studies with more than 1000 participants (aged 16 to 78 years) coping with varied life stressors, from diagnosis with a serious illness to daily stress [20-23]. The program has been implemented in person (individually and in groups) and, most recently, has been delivered digitally as a self-guided program for individuals with diabetes [24], depression [25-27], HIV [28], and cancer [21,29] and for the general public during the COVID-19 pandemic [30]. However, research is lacking on how to optimally design such interventions specifically for aging adults with FMS [31].

To address this gap, this study tested a web-based program that integrates prior theoretical work on positive emotions [32-34], stress and coping [35,36], and PE [37,38]. Specifically, we examined the effects of an internet-delivered positive affect (PA) skills intervention, called Lessons in Affect Regulation to Keep Stress and Pain Under Control (LARKSPUR), on daily PE, affect, pain, and fatigue in adults with FMS. LARKSPUR is a self-guided web-based intervention that consists of 8 skills taught over the course of 5 weeks with opportunities for practice of the skills built into the platform. Our initial pilot study demonstrated LARKSPUR's feasibility and preliminary efficacy at improving overall affect and reducing pain catastrophizing [39]. However, it did not examine fluctuations in participants' daily experiences, which are hypothesized to be a key mechanism based on LARKSPUR's theoretical underpinnings [34,39].

Aims and Hypotheses

We sought to determine if the LARKSPUR intervention could improve daily emotional and functional well-being in adults with FMS by targeting daily PE. Consistent with the Positive Pathways to Health model [34], we hypothesized that participants randomly assigned to LARKSPUR would show greater enhancements in daily PE-related affective well-being, pain, and fatigue compared to participants in the emotion reporting/attention control condition. By using digital technology, this study provides a scalable and accessible intervention that may help overcome barriers to in-person care for individuals with FMS [40,41].

Methods

Study Design and Procedure

Overview

The LARKSPUR pilot (trial registration: ClinicalTrials.gov NCT04869345) is described elsewhere [42]. Briefly, we recruited participants who met the following criteria: (1) aged 50 years and older, (2) access to a Wi-Fi internet connection, (3) English literacy via self-reports of fluency and reading and writing comprehension, and (4) diagnosis of FMS based on the American College of Rheumatology Fibromyalgia Symptom Severity Scale [43] or physician confirmation of FMS. Exclusions were (1) moderate or severe cognitive impairment (two or more errors on a 6-item Mini-Mental State Examination) [44], (2) current behavioral treatment for chronic pain, or (3) enrollment in another chronic pain trial. Eligible and consenting participants were randomly assigned to LARKSPUR or the control condition.

Participants were randomly assigned to either the intervention group (LARKSPUR) or the control group (emotion reporting) in a 1:1 ratio, using block sizes of 2, 4, 6, or 8. The random allocation sequence was generated centrally by a computerized program and implemented by study staff who were not involved in data collection. To prevent selection bias, allocation concealment was used, and group assignments were revealed to both participants and designated study staff only after the baseline assessment was completed. While outcome assessors were blinded to group assignments, participants could not be blinded due to the nature of the intervention.

LARKSPUR Intervention

Individuals randomized to LARKSPUR received skills training to increase PA. The self-guided web-based intervention

XSL•FO
RenderX

instructed individuals in the use of 8 PA skills delivered over 5 weekly learning modules. The eight skills included (1) noticing PE [45,46], (2) savoring PE [47,48], (3) identifying personal strengths [49,50], (4) behavioral activation to set and work toward attainable goals [51,52], (5) mindfulness [53,54], (6) positive reappraisal [35,55], (7) gratitude [56,57], and (8) acts of kindness [58,59]. Each module consisted of a video introduction, web-based exercises, examples, and homework assignments. Participants practiced skills daily and reported experiences on the digital platform.

Control Program

Control participants completed daily emotion reports, rating positive and negative emotions over the past 24 hours on a 5-point Likert scale. This control condition was designed to match the LARKSPUR group in terms of web-based contact, attention to emotional states, and study duration, without providing any specific skills or strategies to enhance PA or cope with pain. Previous studies have used similar emotion reporting as an active control condition for positive psychology interventions [34].

Participants in both arms were assessed at the baseline, postintervention, and 1-month follow-up time points. In addition, before the intervention (baseline), after the intervention (postintervention), and at 1-month follow-up, participants completed a 7-day burst of web-based daily assessments of PE, PA and negative affect (NA), pain intensity, and fatigue. Participants received up to US $142 in gift cards for completing the assessments: US $75 for baseline, postintervention, and follow-up assessments; US $42 for the diary assessments (US $2 each for 21 days); and US $25 for postintervention feedback.

Participants

Participants were recruited from July 2021 to June 2022 through referrals from New York State practicing physicians and posted flyers throughout the New York Presbyterian Health Care System, New York City–based senior centers, community centers, and web-based platforms (eg, Facebook groups). Recruitment links were also posted on clinicaltrials.gov and emailed to potential participants via ResearchMatch, a national health volunteer research registry created by several US-based academic institutions and supported by the US National Institutes of Health.

Measures

The primary outcomes included affective, pain, and fatigue responsivity to daily PE. PE were assessed with five items asking participants to report whether the following PE had occurred in the past 24 hours: (1) positive interpersonal interaction; (2) positive experience at work, school, or a volunteer position; (3) positive experience at home; (4) network PE (ie, PE experienced by a close friend or relative); and (5) any other PE [10].

PA and NA were assessed using the 20-item modified Differential Emotions Scale, which asked respondents to rate how often they felt 10 positive (eg, amusement, gratitude, and joy) and 10 negative (eg, anger, guilt, and sadness) emotions during the past 24 hours [60]. Items were rated on a 5-point Likert-type scale ranging from 0 (never) to 4 (most of the time).

Pain intensity and fatigue were each measured using a single item with a rating scale from 0 (no pain or fatigue) to 10 (as much pain or fatigue as could be) [61].

Statistical Analysis

Overview

Given the relatively modest sample size and longitudinal data structure, we used Bayesian estimation rather than frequentist estimation. Compared to frequentist methods, Bayesian methods are better powered for modeling complex data with limited observations, as they incorporate both sample evidence and prior knowledge to derive posterior distributions [62]. To enable effect size interpretation, we report Bayesian analogs to frequentist intervals and *P* values: highest posterior density intervals (HPDs) and posterior probabilities of direction (P_d). HPDs describe uncertainty by delineating the most credible values comprising a certain percentage of the posterior distribution [63]. For example, a 95% HPD of 0.1 to 0.3 indicates a 95% probability that the true effect lies between 0.1 and 0.3 based on the accumulated evidence. The P_d value directly quantifies certainty regarding an effect's existence and direction [64].

We computed P_d using the *bayestestR* (version 0.13.1) R package (R Foundation for Statistical Computing) [65]. Although related to frequentist *P* values, P_d offers advantages in interpreting effects for small samples [66]. If 99% of the posterior distribution lies above zero, there is high certainty of a positive effect. We used a P_d threshold of .975, corresponding to a frequentist significance level of .05 for 2-sided testing: effects with a P_d>.975 were considered statistically "significant." Overall, this Bayesian framework enabled appropriate hypothesis testing despite the limited sample size.

Measurement Error

We accounted for measurement errors in PA and NA using a Bayesian errors-in-variables approach [67-69]. For PA and NA, person-specific measurement error in each 7-day burst was accounted for using Williams and Hazer's [70] approach, which uses Cronbach α [71] as a lower bound for reliability (average within-person reliabilities for PA and NA were 0.93 and 0.90, respectively). As a sensitivity check, we also conducted analyses without adjusting for measurement error; results were robust to measurement error correction. Measurement error correction was not performed for the single-item pain and fatigue outcomes.

Longitudinal Mixed-Effects Models

Outcomes were assessed using intention-to-treat analyses for all participants providing data at baseline and at least 1 postintervention or 1-month follow-up assessment. Data were modeled using multivariate longitudinal mixed-effects models [72] within a Bayesian framework [73] using the *brms* R package (version 2.20.4) [74] with default noninformative priors. Specifically, we modeled PA and NA as bivariate outcomes, where PA was normally distributed and NA was lognormally

XSL•FO
RenderX

distributed, to account for notable positive skew potentially arising from censoring [75] (regression coefficients for NA represent the multiplicative change in NA for a one-unit predictor increase); we modeled pain and fatigue as correlated multivariate normally distributed outcomes.

The study included three key time points for data collection: baseline (T1), postintervention (T2), and 1-month follow-up (T3). Baseline measures were collected immediately before the intervention began. The postintervention assessment (T2) was conducted immediately after the 5-week intervention period, allowing us to capture the immediate effects of the LARKSPUR program. The 1-month follow-up (T3) was designed to assess the durability of the intervention effects over time. In our analyses, we used effect coding to represent the treatment group and time points.

For the treatment group, LARKSPUR was coded as 0.5 and the control condition as –0.5. For the time points used in the main analysis, the 1-month follow-up (T3) was coded as 0.5 and the postintervention time point (T2) was coded as –0.5. Treatment group (0.5=LARKSPUR, –0.5=control) and time point (0.5=1-month follow-up, –0.5=postintervention) were effect-coded. We adjusted for grand-mean-centered baseline PE for all outcomes. We also adjusted each outcome for its corresponding grand-mean-centered baseline score (eg, PA was predicted by grand-mean-centered baseline PA, NA was predicted by grand-mean-centered baseline NA, pain was predicted by grand-mean-centered baseline pain, and fatigue was predicted by grand-mean-centered baseline fatigue). The linear daily change was assumed for all outcomes and centered at the middle (fourth day) of a 7-day burst.

Between-person and within-person effects of PE were disaggregated by including individual burst-specific average PE (grand-mean-centered) and person-mean-centered daily PE as predictors of all outcomes [76]. Finally, we assessed the moderation effects of within-person and between-person PE on treatment differences and treatment differences in change by including 3-way interactions (average PE×group×time and daily PE×group×time) and lower-order, 2-way interactions among treatment group, time, and daily or average PE. Planned comparisons examined between-group differences in PE responsivity at the postintervention and follow-up time points.

Ethical Considerations

Ethics approval was obtained from the Institutional Review Board at Weill Cornell Medicine (protocol number 20-06022291). All participants were informed about the study details and procedures before providing their written consent. The collected data were coded and deidentified. Participation was strictly voluntary, and participants were informed about their right to withdraw from the study at any point without any consequences.

Results

Demographic Characteristics

The sample included 95 adults aged 50-80 years. The majority were female (89/95, 94%) and identified as White (76/95, 80%). Of the 142 individuals initially screened for participation, 47 individuals were excluded from the study. The primary reasons for exclusion were not meeting the inclusion criteria (n=22) and declining to participate (n=21). Additional reasons included not providing consent for access to medical records (n=1), inability to access the intervention platform (n=1), and other unspecified reasons (n=2). The remaining 95 eligible participants were enrolled in the study and randomly assigned to either the LARKSPUR intervention or the control condition. All 95 participants completed the baseline questionnaires and were randomly assigned to the LARKSPUR intervention (n=49) or the control (n=46). Using an intention-to-treat analysis, 86 participants (n=43 LARKSPUR and n=43 control) completing baseline and postintervention assessments were included. In addition, participants completing ≥2 daily assessments per measurement burst were included to assess daily changes, resulting in a final analytic sample of 80 participants (n=40 in LARKSPUR and n=40 in control; see the CONSORT [Consolidated Standards of Reporting Trials] diagram in Figure 1). Demographic variables and baseline measures for the analytic sample are summarized in Table 1. As demographic and baseline measures were collected prerandomization, group differences at baseline were not tested [77].

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the LARKSPUR study. LARKSPUR: Lessons in Affect Regulation to Keep Stress and Pain Under Control.

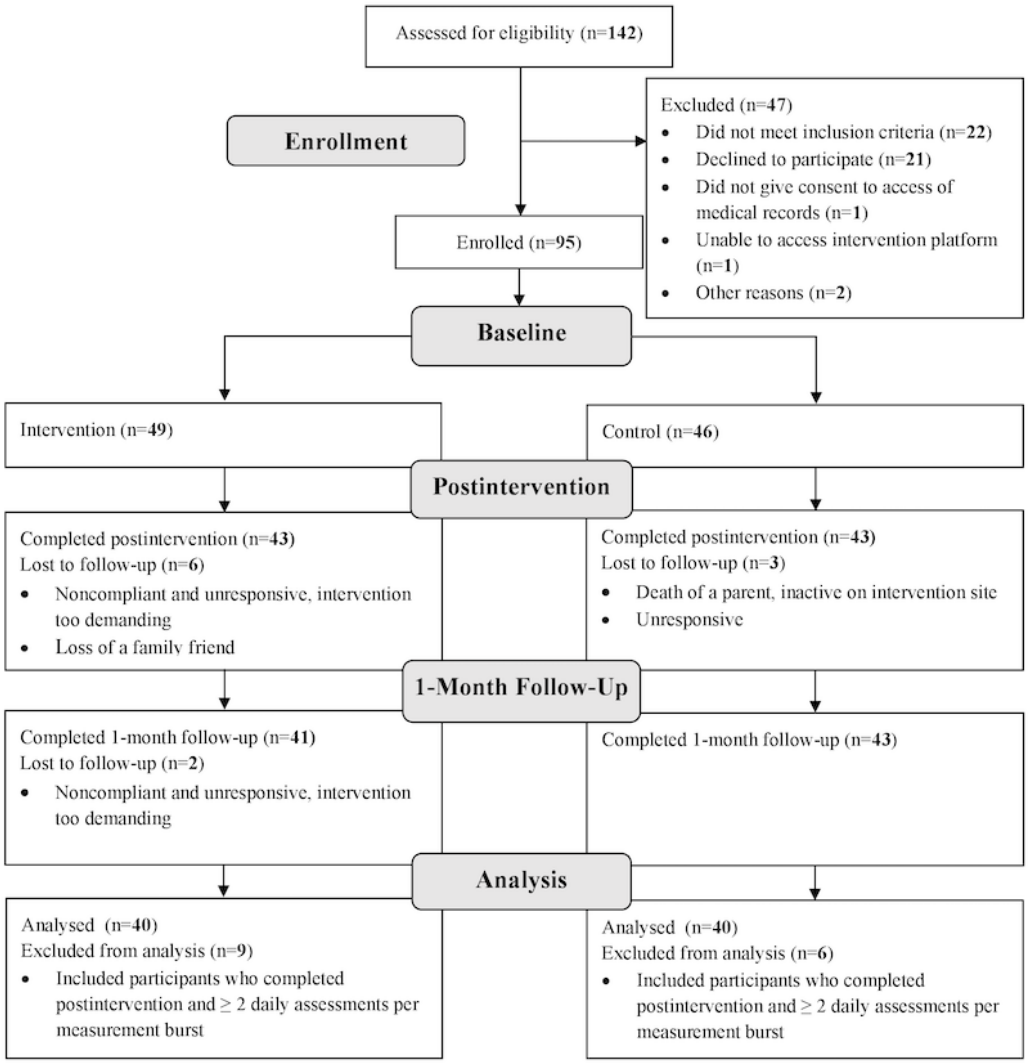

XSL•FO
RenderX

Table 1. Demographic characteristics and baseline measures of participants in LARKSPUR[a] and the control condition.

Characteristic	LARKSPUR (n=40)	Control (n=40)	Overall (n=80)
Sex, n (%)			
Male	1 (2)	3 (8)	4 (5)
Female	39 (98)	37 (92)	76 (95)
Age group (years), n (%)			
50-59	17 (42)	21 (52)	38 (48)
60-69	16 (40)	17 (42)	33 (41)
70-79	6 (15)	2 (5)	8 (10)
≥80	1 (2)	0 (0)	1 (1)
Race and ethnicity, n (%)			
Hispanic or Latino	1 (2)	2 (5)	3 (4)
Black or African American	2 (5)	4 (10)	6 (8)
White	33 (82)	32 (80)	65 (81)
More than one race	3 (8)	1 (2)	4 (5)
Not reported	1 (2)	1 (2)	3 (4)
Education, n (%)			
High school diploma or GED[b]	2 (5)	1 (2)	3 (3)
Some college (no degree)	9 (22)	8 (20)	17 (20)
Associate degree	6 (15)	4 (10)	11 (13)
Bachelor's degree	10 (25)	8 (20)	21 (24)
Postgraduate (no degree)	3 (8)	7 (18)	10 (12)
Master's degree	8 (20)	12 (30)	21 (24)
Doctoral degree	2 (5)	0 (0)	3 (3)
Baseline measure, mean (SD)			
Positive events	2.2 (0.9)	2.3 (1.0)	2.2 (0.9)
Positive affect	2.7 (0.6)	2.7 (0.6)	2.7 (0.6)
Negative affect	1.7 (0.5)	1.5 (0.5)	1.6 (0.5)
Pain	5.5 (1.5)	5.5 (1.8)	5.5 (1.7)
Fatigue	6.0 (1.7)	5.4 (2.1)	5.7 (1.9)

[a]LARKSPUR: Lessons in Affect Regulation to Keep Stress and Pain Under Control.

[b]GED: General Education Degree.

Main Analysis

Affective Responsivity

The bivariate mixed-effects model revealed significant interactions between treatment group, time point, and daily PE in predicting PA and NA. Parameter posterior median estimates and 95% HPD intervals are presented in Table 2. In Bayesian estimation, the distribution of model parameters (eg, regression slopes) is estimated, allowing for a comprehensive characterization of uncertainty. To summarize the typical estimate or central tendency of these distributions, we report median parameter estimates. These estimates can be interpreted in a manner analogous to traditional frequentist estimates obtained through maximum likelihood or least-squares estimation, representing the most probable value of a parameter given the accumulated evidence. Furthermore, 95% HPD intervals provide a quantitative measure of uncertainty in parameter estimates, analogous to frequentist CIs. These intervals can be directly interpreted as follows: given the model and data, there is a 95% probability that the true parameter value lies within the HPD interval, thereby providing a range of plausible values for the parameter. Interactions indicate that LARKSPUR differentially impacted affective responsivity to daily PE compared to the control condition. Specifically, planned comparisons (ie, differences) between regression slopes for LARKSPUR (b_L) and controls (b_C), denoted by b_L–b_C, showed that at the postintervention time point, LARKSPUR participants exhibited greater reductions in NA (b_L–b_C=–0.06, 95% HPD –0.10 to –0.02; P_d>.99) and increases in PA (b_L–b_C=0.10, 95% HPD 0.02-0.19, P_d=.99) in response to daily

XSL•FO
RenderX

PE compared to controls. However, these differential gains in affective responsivity were not maintained at 1-month follow-up, with nonsignificant between-group differences in both PA (b_L–b_C=0.01, 95% HPD –0.08 to 0.09; P_d=.56) and NA (b_L–b_C=–0.00, 95% HPD –0.04 to 0.04; P_d=.51) responsivity. As shown in Figure 2, both groups evidenced slightly diminished affective responsivity from postintervention to follow-up, suggesting that continued practice of LARKSPUR skills may be necessary to sustain affective gains long-term.

Table 2. Mixed-effects model parameter estimates for PA[a] and NA[b] (n=80).

	PA, coefficient (95% HPD[c])	NA, coefficient (95% HPD)
Fixed effects		
Intercept	2.72 (2.62 to 2.83)	0.41 (0.37 to 0.45)
Day	–0.03 (–0.04 to –0.01)	–0.01 (–0.01 to –0.00)
Baseline outcome	0.75 (0.58 to 0.92)	0.47 (0.40 to 0.55)
Time	–0.02 (–0.09 to 0.05)	–0.01 (–0.03 to 0.02)
Group	0.11 (–0.10 to 0.32)	–0.03 (–0.11 to 0.05)
PE[d] (BW[e])	0.17 (0.06 to 0.27)	–0.01 (–0.05 to 0.03)
PE (WI[f])	0.21 (0.17 to 0.24)	–0.04 (–0.05 to –0.02)
Time×group	0.02 (–0.11 to 0.16)	–0.02 (–0.07 to 0.03)
Group×PE (BW)	0.06 (–0.13 to 0.25)	0.01 (–0.06 to 0.09)
Time×PE (BW)	0.04 (–0.03 to 0.10)	–0.02 (–0.04 to 0.01)
Group×PE (WI)	0.06 (–0.01 to 0.13)	–0.03 (–0.06 to 0.00)
Time×PE (WI)	0.00 (–0.05 to 0.05)	0.01 (–0.01 to 0.04)
Time×group×PE (BW)	–0.03 (–0.16 to 0.10)	–0.02 (–0.07 to 0.03)
Time×group×PE (WI)	–0.10 (–0.20 to –0.00)	0.06 (0.01 to 0.11)
Variance parameters		
L1[g] SD	0.34 (0.31 to 0.36)	0.15 (0.14 to 0.16)
L1 AR(1)[h]	0.42 (0.30 to 0.53)	0[i]
L1 *r*	0	0
L2[j] SD: intercepts	0.43 (0.36 to 0.53)	0.17 (0.14 to 0.20)
L2 SD: PE (WI) slopes	0.10 (0.07 to 0.15)	0.04 (0.03 to 0.06)
L2 *r*: intercepts	–0.30 (–0.52 to –0.05)	0
L2 *r*: PE (WI) slopes	–0.78 (–0.97 to –0.38)	0

[a]PA: positive affect.
[b]NA: negative affect.
[c]HPD: highest posterior density.
[d]PE: positive events.
[e]BW: between-person effect.
[f]WI: average within-person effect.
[g]L1: level 1 (within person).
[h]AR(1): lag-1 residual autocorrelation.
[i]Not estimated.
[j]L2: level 2 (between person).

XSL•FO
RenderX

Figure 2. Predicted responsivity to daily positive events from mixed-effects model for positive affect and negative affect at the postintervention and 1-month follow-up time points. LARKSPUR: Lessons in Affect Regulation to Keep Stress and Pain Under Control.

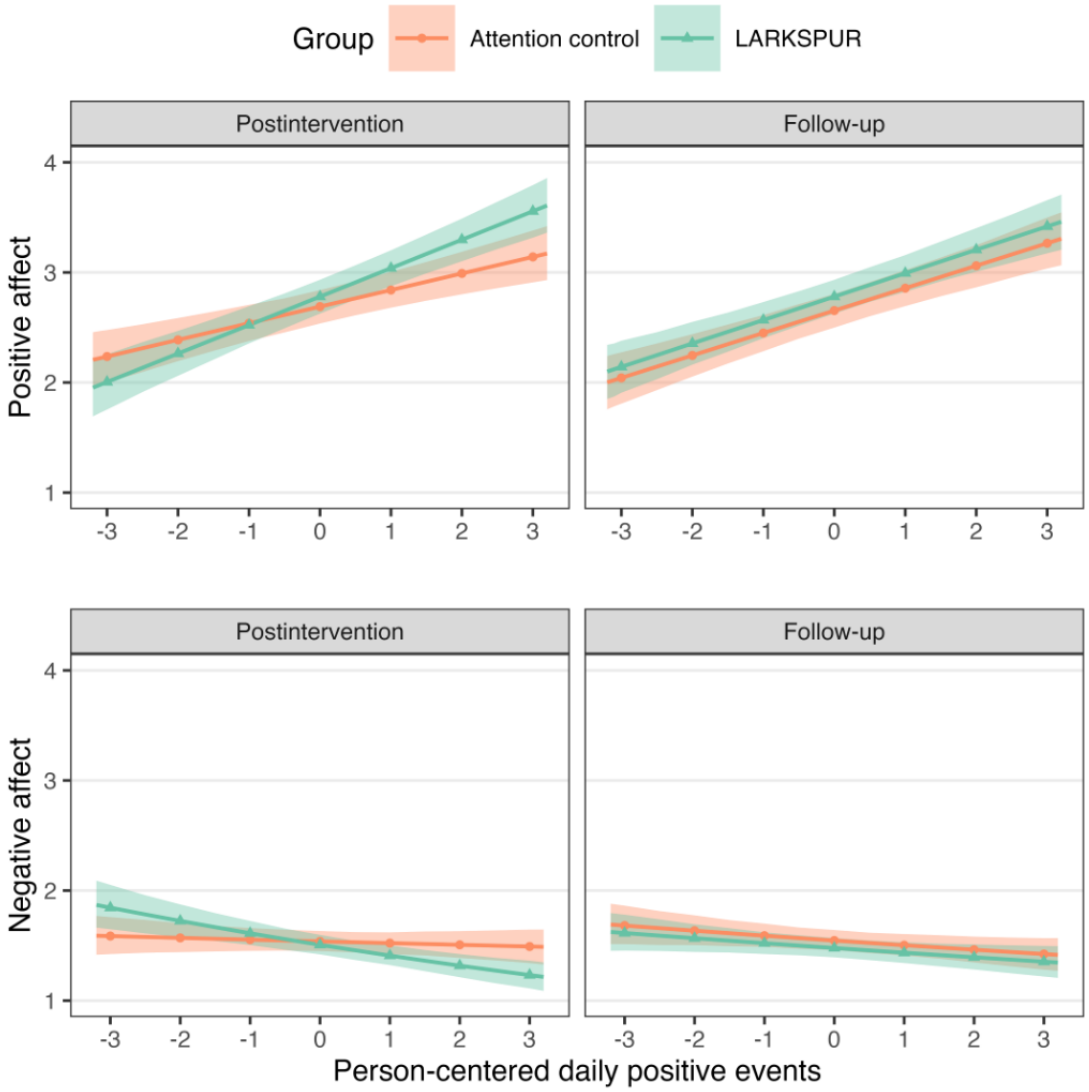

Pain and Fatigue Responsivity

In contrast to the affective outcomes, LARKSPUR yielded more persistent improvements in daily pain and fatigue responsivity compared to the control condition (Table 3). Across the postintervention and 1-month follow-up time points, LARKSPUR led to greater reductions in pain (b_L–b_C=–0.20, 95% HPD –0.36 to –0.04; P_d=.99) and fatigue (b_L–b_C=–0.24, 95% HPD –0.41 to –0.06; P_d>.99) following PE. As illustrated in Figure 3, LARKSPUR participants maintained diminished pain and fatigue following daily PE from the postintervention time point through 1-month follow-up, whereas controls showed minimal change. This sustained effect for the functional outcomes indicates that LARKSPUR may have produced broader impacts beyond temporary affective gains. In sum, results suggest LARKSPUR (relative to the control condition) improved affective, pain, and fatigue responsivity to daily PE, particularly at the postintervention time point. The affective benefits were not maintained, but the pain and fatigue improvements persisted through the 1-month follow-up.

XSL•FO
RenderX

Table 3. Mixed-effects model parameter estimates for pain and fatigue (n=80).

	Pain, coefficient (95% HPD[a])	Fatigue, coefficient (95% HPD)
Fixed effects		
Intercept	5.24 (4.99 to 5.49)	5.75 (5.48 to 6.02)
Day	–0.01 (–0.06 to 0.03)	–0.04 (–0.09 to 0.00)
Baseline outcome	0.79 (0.64 to 0.94)	0.76 (0.62 to 0.91)
Time	–0.37 (–0.56 to –0.18)	–0.06 (–0.26 to 0.14)
Group	–0.02 (–0.52 to 0.49)	–0.34 (–0.89 to 0.21)
PE[b] (BW[c])	–0.05 (–0.28 to 0.20)	0.00 (–0.26 to 0.26)
PE (WI[d])	–0.12 (–0.20 to –0.04)	–0.20 (–0.28 to –0.11)
Time×group	–0.33 (–0.72 to 0.05)	–0.33 (–0.74 to 0.07)
Group×PE (BW)	0.35 (–0.13 to 0.83)	0.07 (–0.44 to 0.59)
Time×PE (BW)	0.03 (–0.15 to 0.21)	0.11 (–0.08 to 0.30)
Group×PE (WI)	–0.20 (–0.36 to –0.04)	–0.24 (–0.42 to –0.07)
Time×PE (WI)	0.05 (–0.11 to 0.20)	0.09 (–0.09 to 0.26)
Time×group×PE (BW)	–0.25 (–0.61 to 0.12)	–0.30 (–0.68 to 0.09)
Time×group×PE (WI)	0.10 (–0.21 to 0.42)	–0.02 (–0.37 to 0.33)
Variance parameters		
L1[e] SD	1.32 (1.26 to 1.38)	1.45 (1.38 to 1.52)
L1 AR(1)[f]	0.19 (0.12 to 0.26)	0.13 (0.06 to 0.20)
L1 *r*	0.38 (0.33 to 0.44)	0
L2[g] SD: intercepts	1.03 (0.84 to 1.25)	1.11 (0.92 to 1.35)
L2 SD: PE (WI) slopes	0.08 (0.00 to 0.19)	0.08 (0.00 to 0.23)
L2 *r*: intercepts	0.37 (0.11 to 0.58)	0
L2 *r*: PE (WI) slopes	0.14 (–0.76 to 0.85)	0

[a]HPD: highest posterior density interval.
[b]PE: positive events.
[c]BW: between-person effect.
[d]WI: average within-person effect.
[e]L1: level 1 (within person).
[f]AR(1): lag-1 residual autocorrelation.
[g]L2: level 2 (between person).

XSL•FO
RenderX

Figure 3. Predicted responsivity to daily positive events from mixed-effects model for pain and fatigue at the postintervention and 1-month follow-up time points. LARKSPUR: Lessons in Affect Regulation to Keep Stress and Pain Under Control.

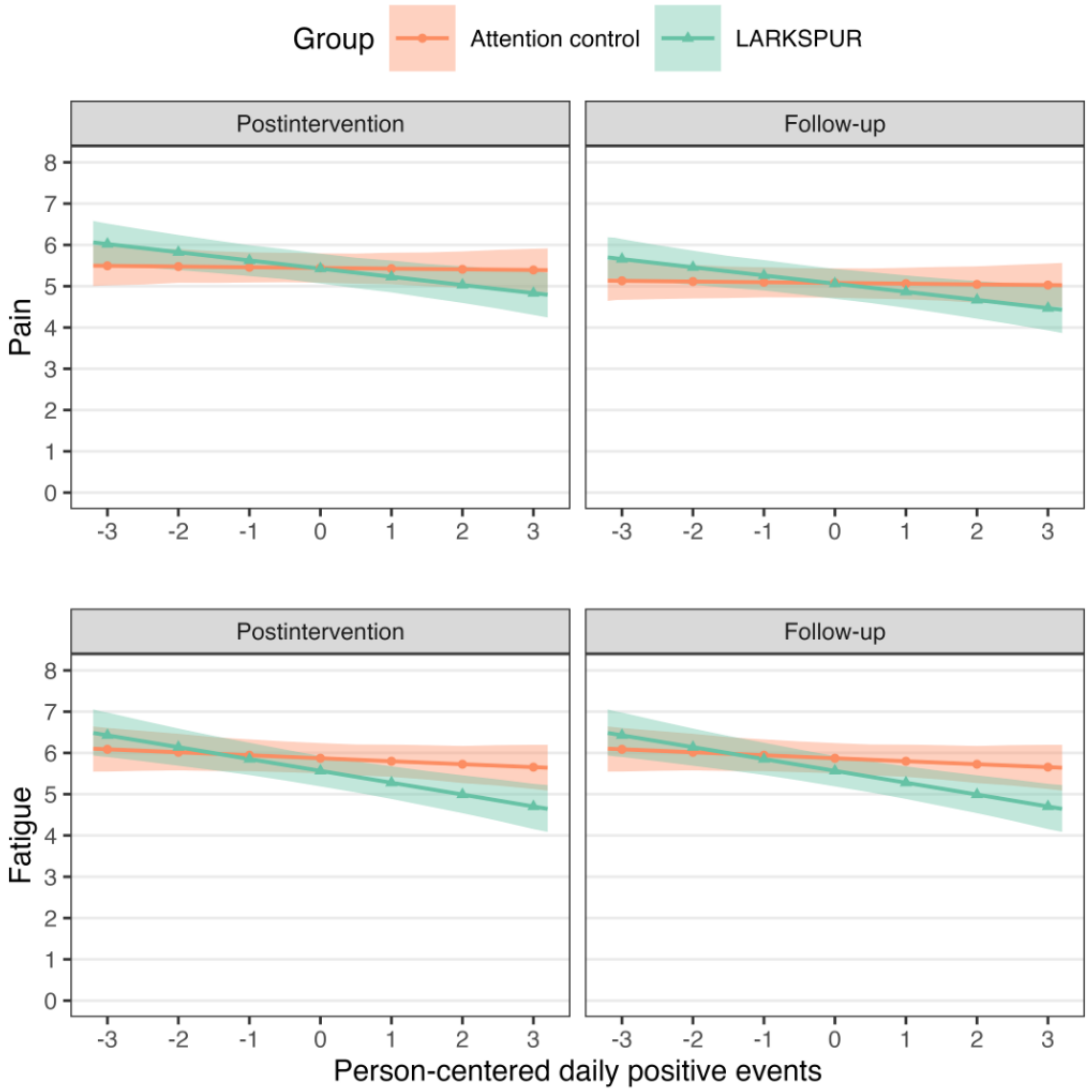

Discussion

Principal Findings

This randomized controlled trial provides initial evidence that LARKSPUR, an internet-delivered PA skills intervention, can enhance responsivity to daily PE in adults with FMS. LARKSPUR led to greater decreases in NA and increases in PA following PE compared to the control condition. It also led to reductions in pain and fatigue following PE at the postintervention and 1-month follow-up time points. These findings suggest that LARKSPUR has the potential to be an accessible and effective eHealth intervention to boost well-being and improve symptom management in this population.

Strengths and Implications

The findings have several strengths and implications for research and practice. First, the results support previous research showing the benefits of positive psychological interventions for populations with chronic pain [13,78]. By targeting engagement in pleasant activities and PE, LARKSPUR aligns with prior work highlighting the value of positive experiences in promoting resilience and well-being [40,41]. Second, the study demonstrates the feasibility and potential of delivering such programs remotely to aging adults with FMS. The web-based format helps overcome barriers to in-person treatment while improving reach, flexibility, and scalability [79-81]. This is especially relevant given COVID-19 contexts demanding flexible health services [82]. Third, incorporating technology into the intervention likely enhanced program quality and engagement through personalized feedback and reminders. These digital components may have facilitated the gains in PE responsivity observed in the LARKSPUR group compared to controls. Fourth, the adaptable nature of LARKSPUR allows for implementation across varied formats to meet diverse user needs and preferences. This tailoring can boost engagement, which is key for patients with chronic pain who have heterogeneous motivations and capacities [8,13]. Finally, the findings provide insights into mechanisms of change. The affective benefits at the postintervention time point suggest LARKSPUR honed positive emotion regulation skills; however, continued practice may be needed to maintain gains. In contrast, the persistence of LARKSPUR's effects on pain and fatigue

XSL•FO
RenderX

through 1-month follow-up indicates the intervention likely produced broader impacts beyond the temporary enhancement of positive emotions.

Limitations and Future Directions

Several limitations of this study should be acknowledged. The modest sample size limits the precision and generalizability of our findings, highlighting the need for larger, more diverse samples of individuals with fibromyalgia. Future studies should aim to recruit participants from a broader range of demographic groups to examine the effects of LARKSPUR across different populations. Additionally, our reliance on self-report measures may introduce bias, and incorporating objective assessments, such as clinician observations or actigraphy, could provide more robust evidence. Furthermore, our focus on PE frequency was narrow, and exploring the variety of daily experiences may reveal new avenues for improving overall well-being [83-85]. Finally, longer-term follow-up is needed to determine the sustainability of benefits and to investigate strategies for maintaining gains, such as periodic booster sessions.

Conclusions

This randomized trial provides preliminary evidence that a web-based PA skills intervention can enhance responsivity to daily PE across emotional and functional domains in adults with FMS. Specifically, the findings indicate LARKSPUR improved participants' affective, pain, and fatigue responses following daily positive experiences compared to controls. This provides initial support for the efficacy of technology-based platforms, like LARKSPUR, to boost well-being in this population by targeting sensitivity to daily PE. Further research is needed to replicate these results in a larger sample and address optimizing technology to enhance LARKSPUR's scalability and efficacy for diverse aging adults with fibromyalgia.

Acknowledgments

This work was supported by the National Institute on Aging of the National Institutes of Health (grant R24AG064191). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. We are grateful to Sabrina Feldman, Sidney Gibson, Usha Kaul, Ethan Low, Erica Sluys, Isel Vazquez-Carbajal, and Rebecca Wartels for their assistance with data collection. This manuscript was written entirely by human authors, without the assistance of generative artificial intelligence tools or language models.

Authors' Contributions

AO contributed to the conceptualization, methodology, and writing, reviewing, and editing of the manuscript; KW contributed to the methodology, data analysis, and reviewing and editing; DC contributed to the methodology and reviewing and editing; JM contributed to the conceptualization, methodology, and reviewing and editing; EW contributed to the conceptualization, methodology, and reviewing and editing; EA contributed to the methodology and reviewing and editing; and MCR contributed to the conceptualization, methodology, and reviewing and editing. All the authors approved the final manuscript for submission.

Conflicts of Interest

EW is a guest editor for the Theme Issue "Digital Technology for Behavioral Interventions in Aging: Opportunities and Challenges" in the *Journal of Medical Internet Research*.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1918 KB - jmir_v26i1e54678_app1.pdf]

References

1. Häuser W, Wolfe F. Diagnosis and diagnostic tests for fibromyalgia (syndrome). Reumatismo 2012;64(4):194-205. [doi: 10.4081/reumatismo.2012.194] [Medline: 23024964]
2. Sarzi-Puttini P, Giorgi V, Marotto D, Atzeni F. Fibromyalgia: an update on clinical characteristics, aetiopathogenesis and treatment. Nat Rev Rheumatol 2020;16(11):645-660. [doi: 10.1038/s41584-020-00506-w] [Medline: 33024295]
3. Wolfe F, Clauw DJ, Fitzcharles MA, Goldenberg DL, Katz RS, Mease P, et al. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. Arthritis Care Res (Hoboken) 2010;62(5):600-610 [FREE Full text] [doi: 10.1002/acr.20140] [Medline: 20461783]
4. Flynn DM. Chronic musculoskeletal pain: nonpharmacologic, noninvasive treatments. Am Fam Physician 2020;102(8):465-477 [FREE Full text] [Medline: 33064421]
5. Carr A, Cullen K, Keeney C, Canning C, Mooney O, Chinseallaigh E, et al. Effectiveness of positive psychology interventions: a systematic review and meta-analysis. J Posit Psychol 2020;16(6):749-769. [doi: 10.1080/17439760.2020.1818807]
6. Hendriks T, Schotanus-Dijkstra M, Hassankhan A, de Jong J, Bohlmeijer E. The efficacy of multi-component positive psychology interventions: a systematic review and meta-analysis of randomized controlled trials. J Happiness Stud 2019;21(1):357-390. [doi: 10.1007/s10902-019-00082-1]

XSL•FO
RenderX

7. Sin NL, Lyubomirsky S. Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. J Clin Psychol 2009;65(5):467-487. [doi: 10.1002/jclp.20593] [Medline: 19301241]
8. Braunwalder C, Müller R, Glisic M, Fekete C. Are positive psychology interventions efficacious in chronic pain treatment? A systematic review and meta-analysis of randomized controlled trials. Pain Med 2022;23(1):122-136. [doi: 10.1093/pm/pnab247] [Medline: 34347095]
9. Iddon JE, Dickson JM, Unwin J. Positive psychological interventions and chronic non-cancer pain: a systematic review of the literature. Int J Appl Posit Psychol 2016;1(1-3):133-157. [doi: 10.1007/s41042-016-0003-6]
10. Sin NL, Ong AD, Stawski RS, Almeida DM. Daily positive events and diurnal cortisol rhythms: examination of between-person differences and within-person variation. Psychoneuroendocrinology 2017;83:91-100 [FREE Full text] [doi: 10.1016/j.psyneuen.2017.06.001] [Medline: 28601752]
11. Zautra AJ, Guarnaccia CA, Dohrenwend BP. Measuring small life events. Am J Community Psychol 1986;14(6):629-655. [doi: 10.1007/BF00931340] [Medline: 3799554]
12. Finan PH, Zautra AJ, Davis MC. Daily affect relations in fibromyalgia patients reveal positive affective disturbance. Psychosom Med 2009;71(4):474-482. [doi: 10.1097/PSY.0b013e31819e0a8b] [Medline: 19251863]
13. Finan PH, Garland EL. The role of positive affect in pain and its treatment. Clin J Pain 2015;31(2):177-187 [FREE Full text] [doi: 10.1097/AJP.0000000000000092] [Medline: 24751543]
14. Ehde DM, Dillworth TM, Turner JA. Cognitive-behavioral therapy for individuals with chronic pain: efficacy, innovations, and directions for research. Am Psychol 2014;69(2):153-166. [doi: 10.1037/a0035747] [Medline: 24547801]
15. Müller R, Gertz KJ, Molton IR, Terrill AL, Bombardier CH, Ehde DM, et al. Effects of a tailored positive psychology intervention on well-being and pain in individuals with chronic pain and a physical disability: a feasibility trial. Clin J Pain 2016;32(1):32-44. [doi: 10.1097/AJP.0000000000000225] [Medline: 25724020]
16. Wolf LD, Davis MC, Yeung EW, Tennen HA. The within-day relation between lonely episodes and subsequent clinical pain in individuals with fibromyalgia: mediating role of pain cognitions. J Psychosom Res 2015;79(3):202-206 [FREE Full text] [doi: 10.1016/j.jpsychores.2014.12.018] [Medline: 25637526]
17. Wolf LD, Davis MC. Loneliness, daily pain, and perceptions of interpersonal events in adults with fibromyalgia. Health Psychol 2014;33(9):929-937 [FREE Full text] [doi: 10.1037/hea0000059] [Medline: 25180546]
18. Wood W, Mazar A, Neal DT. Habits and goals in human behavior: separate but interacting systems. Perspect Psychol Sci 2022;17(2):590-605. [doi: 10.1177/1745691621994226] [Medline: 34283681]
19. Wood W, Neal DT. A new look at habits and the habit-goal interface. Psychol Rev 2007;114(4):843-863. [doi: 10.1037/0033-295X.114.4.843] [Medline: 17907866]
20. Carrico AW, Moskowitz JT. Positive affect promotes engagement in care after HIV diagnosis. Health Psychol 2014;33(7):686-689 [FREE Full text] [doi: 10.1037/hea0000011] [Medline: 24245846]
21. Cheung EO, Cohn MA, Dunn LB, Melisko ME, Morgan S, Penedo FJ, et al. A randomized pilot trial of a positive affect skill intervention (lessons in linking affect and coping) for women with metastatic breast cancer. Psychooncology 2017;26(12):2101-2108 [FREE Full text] [doi: 10.1002/pon.4312] [Medline: 27862646]
22. Moskowitz JT, Hult JR, Duncan LG, Cohn MA, Maurer S, Bussolari C, et al. A positive affect intervention for people experiencing health-related stress: development and non-randomized pilot test. J Health Psychol 2012;17(5):676-692 [FREE Full text] [doi: 10.1177/1359105311425275] [Medline: 22021272]
23. Moskowitz JT, Carrico AW, Cohn MA, Duncan LG, Bussolari C, Layous K, et al. Randomized controlled trial of a positive affect intervention to reduce stress in people newly diagnosed with HIV; protocol and design for the IRISS study. OAJCT 2014;6:85. [doi: 10.2147/oajct.s64645]
24. Cohn MA, Pietrucha ME, Saslow LR, Hult JR, Moskowitz JT. An online positive affect skills intervention reduces depression in adults with type 2 diabetes. J Posit Psychol 2014;9(6):523-534 [FREE Full text] [doi: 10.1080/17439760.2014.920410] [Medline: 25214877]
25. Addington EL, Cheung EO, Bassett SM, Kwok I, Schuette SA, Shiu E, et al. The MARIGOLD study: feasibility and enhancement of an online intervention to improve emotion regulation in people with elevated depressive symptoms. J Affect Disord 2019;257:352-364 [FREE Full text] [doi: 10.1016/j.jad.2019.07.049] [Medline: 31302525]
26. Cheung EO, Addington EL, Bassett SM, Schuette SA, Shiu EW, Cohn MA, et al. A self-paced, web-based, positive emotion skills intervention for reducing symptoms of depression: protocol for development and pilot testing of MARIGOLD. JMIR Res Protoc 2018;7(6):e10494 [FREE Full text] [doi: 10.2196/10494] [Medline: 29871853]
27. Moskowitz JT, Addington EL, Shiu E, Bassett SM, Schuette S, Kwok I, et al. Facilitator contact, discussion boards, and virtual badges as adherence enhancements to a web-based, self-guided, positive psychological intervention for depression: randomized controlled trial. J Med Internet Res 2021;23(9):e25922 [FREE Full text] [doi: 10.2196/25922] [Medline: 34550076]
28. Bassett SM, Cohn M, Cotten P, Kwok I, Moskowitz JT. Feasibility and acceptability of an online positive affect intervention for those living with comorbid HIV depression. AIDS Behav 2019;23(3):753-764. [doi: 10.1007/s10461-019-02412-z] [Medline: 30701389]

XSL•FO
RenderX

29. Salsman JM, McLouth LE, Cohn M, Tooze JA, Sorkin M, Moskowitz JT. A web-based, positive emotion skills intervention for enhancing posttreatment psychological well-being in young adult cancer survivors (EMPOWER): protocol for a single-arm feasibility Trial. JMIR Res Protoc 2020;9(5):e17078 [FREE Full text] [doi: 10.2196/17078] [Medline: 32463014]
30. Addington EL, Cummings P, Jackson K, Yang D, Moskowitz JT. Exploring retention, usage, and efficacy of web-based delivery of positive emotion regulation skills during the COVID-19 pandemic. Affect Sci 2023;4(1):118-130 [FREE Full text] [doi: 10.1007/s42761-022-00135-4] [Medline: 36211530]
31. Gandy M, Pang STY, Scott AJ, Heriseanu AI, Bisby MA, Dudeney J, et al. Internet-delivered cognitive and behavioural based interventions for adults with chronic pain: a systematic review and meta-analysis of randomized controlled trials. Pain 2022;163(10):e1041-e1053. [doi: 10.1097/j.pain.0000000000002606] [Medline: 35121696]
32. Fredrickson BL. The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. Am Psychol 2001;56(3):218-226 [FREE Full text] [doi: 10.1037//0003-066x.56.3.218] [Medline: 11315248]
33. Fredrickson BL. Positive emotions broaden and build. Adv Exp Soc Psychol 2013:1-53. [doi: 10.1016/b978-0-12-407236-7.00001-2]
34. Moskowitz JT, Addington EL, Cheung EO. Positive psychology and health: well-being interventions in the context of illness. Gen Hosp Psychiatry 2019;61:136-138. [doi: 10.1016/j.genhosppsych.2019.11.001] [Medline: 31757566]
35. Folkman S. Positive psychological states and coping with severe stress. Soc Sci Med 1997;45(8):1207-1221. [doi: 10.1016/s0277-9536(97)00040-3] [Medline: 9381234]
36. Pressman SD, Cohen S. Does positive affect influence health? Psychol Bull 2005;131(6):925-971. [doi: 10.1037/0033-2909.131.6.925] [Medline: 16351329]
37. Layous K, Chancellor J, Lyubomirsky S. Positive activities as protective factors against mental health conditions. J Abnorm Psychol 2014;123(1):3-12. [doi: 10.1037/a0034709] [Medline: 24661154]
38. Lyubomirsky S, Layous K. How do simple positive activities increase well-being? Curr Dir Psychol Sci 2013;22(1):57-62. [doi: 10.1177/0963721412469809]
39. Ong AD, Wilcox KT, Moskowitz JT, Wethington E, Addington EL, Sanni MO, et al. Feasibility, acceptability, and preliminary efficacy of a positive affect skills intervention for adults with fibromyalgia. Innovation Aging 2023;7(10):igad070 [FREE Full text] [doi: 10.1093/geroni/igad070] [Medline: 38094931]
40. Bernardy K, Klose P, Welsch P, Häuser W. Efficacy, acceptability and safety of internet-delivered psychological therapies for fibromyalgia syndrome: a systematic review and meta-analysis of randomized controlled trials. Eur J Pain 2019;23(1):3-14. [doi: 10.1002/ejp.1284] [Medline: 29984490]
41. Donisi V, de Lucia A, Pasini I, Gandolfi M, Schweiger V, del Piccolo L, et al. e-Health interventions targeting pain-related psychological variables in fibromyalgia: a systematic review. Healthcare (Basel) 2023;11(13):1845 [FREE Full text] [doi: 10.3390/healthcare11131845] [Medline: 37444679]
42. Ong AD, Moskowitz JT, Wethington E, Addington EL, Sanni M, Goktas S, et al. Lessons in Affect Regulation to Keep Stress and Pain UndeR control (LARKSPUR): design of a randomized controlled trial to increase positive affect in middle-aged and older adults with fibromyalgia. Contemp Clin Trials 2022;120:106880 [FREE Full text] [doi: 10.1016/j.cct.2022.106880] [Medline: 35964867]
43. Wolfe F, Clauw DJ, Fitzcharles MA, Goldenberg DL, Häuser W, Katz RS, et al. Fibromyalgia criteria and severity scales for clinical and epidemiological studies: a modification of the ACR preliminary diagnostic criteria for fibromyalgia. J Rheumatol 2011;38(6):1113-1122. [doi: 10.3899/jrheum.100594] [Medline: 21285161]
44. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-item screener to identify cognitive impairment among potential subjects for clinical research. Med Care 2002;40(9):771-781. [doi: 10.1097/00005650-200209000-00007] [Medline: 12218768]
45. Krause N. Positive life events and depressive symptoms in older adults. Behav Med 1988;14(3):101-112. [doi: 10.1080/08964289.1988.9935131] [Medline: 3167237]
46. Lewinsohn PM, Amenson CS. Some relations between pleasant and unpleasant mood-related events and depression. J Abnorm Psychol 1978;87(6):644-654. [doi: 10.1037//0021-843x.87.6.644] [Medline: 739087]
47. Bryant FB. A four‐factor model of perceived control: avoiding, coping, obtaining, and savoring. J Pers 2006;57(4):773-797. [doi: 10.1111/j.1467-6494.1989.tb00494.x]
48. Langston CA. Capitalizing on and coping with daily-life events: expressive responses to positive events. J Pers Soc Psychol 1994;67(6):1112-1125. [doi: 10.1037//0022-3514.67.6.1112]
49. Taylor SE, Kemeny ME, Aspinwall LG, Schneider SG, Rodriguez R, Herbert M. Optimism, coping, psychological distress, and high-risk sexual behavior among men at risk for acquired immunodeficiency syndrome (AIDS). J Pers Soc Psychol 1992;63(3):460-473. [doi: 10.1037//0022-3514.63.3.460] [Medline: 1403625]
50. Taylor SE, Lobel M. Social comparison activity under threat: downward evaluation and upward contacts. Psychol Rev 1989;96(4):569-575. [doi: 10.1037/0033-295x.96.4.569] [Medline: 2678204]
51. Manos RC, Kanter JW, Busch AM. A critical review of assessment strategies to measure the behavioral activation model of depression. Clin Psychol Rev 2010;30(5):547-561. [doi: 10.1016/j.cpr.2010.03.008] [Medline: 20444531]
52. Moskowitz JT, Hult JR, Bussolari C, Acree M. What works in coping with HIV? A meta-analysis with implications for coping with serious illness. Psychol Bull 2009;135(1):121-141. [doi: 10.1037/a0014210] [Medline: 19210056]

XSL•FO
RenderX

53. Grossman P, Tiefenthaler-Gilmer U, Raysz A, Kesper U. Mindfulness training as an intervention for fibromyalgia: evidence of postintervention and 3-year follow-up benefits in well-being. Psychother Psychosom 2007;76(4):226-233. [doi: 10.1159/000101501] [Medline: 17570961]
54. Kabat-Zinn J. An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. Gen Hosp Psychiatry 1982;4(1):33-47. [doi: 10.1016/0163-8343(82)90026-3] [Medline: 7042457]
55. Moskowitz JT, Folkman S, Collette L, Vittinghoff E. Coping and mood during aids-related caregiving and bereavement. Ann Behav Med 1996;18(1):49-57. [doi: 10.1007/BF02903939] [Medline: 24203643]
56. Emmons RA, Crumpler CA. Gratitude as a human strength: appraising the evidence. J Soc Clin Psychol 2000;19(1):56-69. [doi: 10.1521/jscp.2000.19.1.56]
57. Fredrickson BL. Gratitude, like other positive emotions, broadens and builds. In: Emmons RA, McCullough ME, editors. The Psychology of Gratitude. Oxford, United Kingdom: Oxford University Press; 2004:145-166.
58. Brown SL, Nesse RM, Vinokur AD, Smith DM. Providing social support may be more beneficial than receiving it: results from a prospective study of mortality. Psychol Sci 2003;14(4):320-327. [doi: 10.1111/1467-9280.14461] [Medline: 12807404]
59. Penner LA, Dovidio JF, Piliavin JA, Schroeder DA. Prosocial behavior: multilevel perspectives. Annu Rev Psychol 2005;56(1):365-392. [doi: 10.1146/annurev.psych.56.091103.070141] [Medline: 15709940]
60. Fredrickson BL, Tugade MM, Waugh CE, Larkin GR. What good are positive emotions in crises? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. J Pers Soc Psychol 2003;84(2):365-376 [FREE Full text] [doi: 10.1037//0022-3514.84.2.365] [Medline: 12585810]
61. Ong AD, Zautra AJ, Reid MC. Psychological resilience predicts decreases in pain catastrophizing through positive emotions. Psychol Aging 2010;25(3):516-523 [FREE Full text] [doi: 10.1037/a0019384] [Medline: 20853962]
62. Depaoli S, van de Schoot R. Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. Psychol Methods 2017;22(2):240-261. [doi: 10.1037/met0000065] [Medline: 26690773]
63. Kruschke JK. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Cambridge, MA: Academic Press; 2015:759.
64. Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. Adv Methods Pract Psychol Sci 2018;1(2):270-280. [doi: 10.1177/2515245918771304]
65. Makowski DS, Ben-Shachar M, Lüdecke D. bayestestR: describing effects and their uncertainty, existence and significance within the bayesian Framework. JOSS 2019;4(40):1541. [doi: 10.21105/joss.01541]
66. Makowski D, Ben-Shachar MS, Chen SHA, Lüdecke D. Indices of effect existence and significance in the Bayesian framework. Front Psychol 2019;10:2767 [FREE Full text] [doi: 10.3389/fpsyg.2019.02767] [Medline: 31920819]
67. Goldstein H, Kounali D, Robinson A. Modelling measurement errors and category misclassifications in multilevel models. Stat Modell 2008;8(3):243-261. [doi: 10.1177/1471082x0800800302]
68. Bollen KA. Structural Equations with Latent Variables. New York, NY: Wiley; 1989:1-514.
69. Loken E, Gelman A. Measurement error and the replication crisis. Science 2017;355(6325):584-585. [doi: 10.1126/science.aal3618] [Medline: 28183939]
70. Williams LJ, Hazer JT. Antecedents and consequences of satisfaction and commitment in turnover models: a reanalysis using latent variable structural equation methods. J Appl Psychol 1986;71(2):219-231. [doi: 10.1037//0021-9010.71.2.219]
71. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297-334. [doi: 10.1007/bf02310555]
72. Singer JD, Willett JB. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford, United Kingdom: Oxford University Press; 2003.
73. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Boca Raton: FL: Chapman & Hall/CRC; 2013:675.
74. Bürkner PC. brms: an R package for Bayesian multilevel models using Stan. J Stat Software 2017;80(1):1-28. [doi: 10.18637/jss.v080.i01]
75. Mader N, Arslan RC, Schmukle SC, Rohrer JM. Emotional (in)stability: neuroticism is associated with increased variability in negative emotion after all. Proc Natl Acad Sci U S A 2023;120(23):e2212154120 [FREE Full text] [doi: 10.1073/pnas.2212154120] [Medline: 37253012]
76. Curran PJ, Bauer DJ. The disaggregation of within-person and between-person effects in longitudinal models of change. Annu Rev Psychol 2011;62(1):583-619 [FREE Full text] [doi: 10.1146/annurev.psych.093008.100356] [Medline: 19575624]
77. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340(1):c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
78. Ong AD. Pathways linking positive emotion and health in later life. Curr Dir Psychol Sci 2010;19(6):358-362. [doi: 10.1177/0963721410388805]
79. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205. [doi: 10.1080/16506070903318960] [Medline: 20183695]

XSL•FO
RenderX

80. Bender JL, Radhakrishnan A, Diorio C, Englesakis M, Jadad AR. Can pain be managed through the internet? A systematic review of randomized controlled trials. Pain 2011;152(8):1740-1750. [doi: 10.1016/j.pain.2011.02.012] [Medline: 21565446]
81. Eccleston C, Blyth FM, Dear BF, Fisher EA, Keefe FJ, Lynch ME, et al. Managing patients with chronic pain during the COVID-19 outbreak: considerations for the rapid introduction of remotely supported (eHealth) pain management services. Pain 2020;161(5):889-893 [FREE Full text] [doi: 10.1097/j.pain.0000000000001885] [Medline: 32251203]
82. Tauben DJ, Langford DJ, Sturgeon JA, Rundell SD, Towle C, Bockman C, et al. Optimizing telehealth pain care after COVID-19. Pain 2020;161(11):2437-2445 [FREE Full text] [doi: 10.1097/j.pain.0000000000002048] [Medline: 32826752]
83. Kashdan TB, Rottenberg J. Psychological flexibility as a fundamental aspect of health. Clin Psychol Rev 2010;30(7):865-878 [FREE Full text] [doi: 10.1016/j.cpr.2010.03.001] [Medline: 21151705]
84. Lee S, Koffer RE, Sprague BN, Charles ST, Ram N, Almeida DM. Activity diversity and its associations with psychological well-being across adulthood. J Gerontol B Psychol Sci Soc Sci 2018;73(6):985-995 [FREE Full text] [doi: 10.1093/geronb/gbw118] [Medline: 27621306]
85. Ong AD, Lee S. Variety in pleasant activities is associated with improved mental health: evidence from two national samples of U.S. adults. Affect Sci 2024;5(2):90-98. [doi: 10.1007/s42761-023-00225-x] [Medline: 39050039]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FMS: fibromyalgia syndrome
HPD: highest posterior density
LARKSPUR: Lessons in Affect Regulation to Keep Stress and Pain Under Control
NA: negative affect
PA: positive affect
PE: positive events

Edited by T de Azevedo Cardoso; submitted 18.11.23; peer-reviewed by S Davis; comments to author 09.08.24; revised version received 12.09.24; accepted 16.09.24; published 10.12.24.

Please cite as:
Ong A, Wilcox K, Reid MC, Wethington E, Cintron D, Addington E, Goktas S, Moskowitz J
Targeting Daily Positive Events to Improve Emotional and Functional Well-Being in Adults With Fibromyalgia: Insights From the LARKSPUR Randomized Controlled Trial
J Med Internet Res 2024;26:e54678
URL: https://www.jmir.org/2024/1/e54678
doi:10.2196/54678
PMID:

©Anthony Ong, Kenneth Wilcox, M Carrington Reid, Elaine Wethington, Dakota Cintron, Elizabeth Addington, Selin Goktas, Judith Moskowitz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study

Eddie Guo[1]; Mehul Gupta[1], MD; Jiawen Deng[2], BHSc; Ye-Jean Park[2]; Michael Paget[1], BFA; Christopher Naugler[1], MD

[1]Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[2]Temerty Faculty of Medicine, University of Toronto, Toronto, AB, Canada

Corresponding Author:
Eddie Guo
Cumming School of Medicine
University of Calgary
3330 University Dr NW
Calgary, AB, T2N 1N4
Canada
Phone: 1 5879880292
Email: eddie.guo@ucalgary.ca

Abstract

Background: The systematic review of clinical research papers is a labor-intensive and time-consuming process that often involves the screening of thousands of titles and abstracts. The accuracy and efficiency of this process are critical for the quality of the review and subsequent health care decisions. Traditional methods rely heavily on human reviewers, often requiring a significant investment of time and resources.

Objective: This study aims to assess the performance of the OpenAI generative pretrained transformer (GPT) and GPT-4 application programming interfaces (APIs) in accurately and efficiently identifying relevant titles and abstracts from real-world clinical review data sets and comparing their performance against ground truth labeling by 2 independent human reviewers.

Methods: We introduce a novel workflow using the Chat GPT and GPT-4 APIs for screening titles and abstracts in clinical reviews. A Python script was created to make calls to the API with the screening criteria in natural language and a corpus of title and abstract data sets filtered by a minimum of 2 human reviewers. We compared the performance of our model against human-reviewed papers across 6 review papers, screening over 24,000 titles and abstracts.

Results: Our results show an accuracy of 0.91, a macro F_1-score of 0.60, a sensitivity of excluded papers of 0.91, and a sensitivity of included papers of 0.76. The interrater variability between 2 independent human screeners was κ=0.46, and the prevalence and bias-adjusted κ between our proposed methods and the consensus-based human decisions was κ=0.96. On a randomly selected subset of papers, the GPT models demonstrated the ability to provide reasoning for their decisions and corrected their initial decisions upon being asked to explain their reasoning for incorrect classifications.

Conclusions: Large language models have the potential to streamline the clinical review process, save valuable time and effort for researchers, and contribute to the overall quality of clinical reviews. By prioritizing the workflow and acting as an aid rather than a replacement for researchers and reviewers, models such as GPT-4 can enhance efficiency and lead to more accurate and reliable conclusions in medical research.

(J Med Internet Res 2024;26:e48996) doi:10.2196/48996

KEYWORDS

abstract screening; Chat GPT; classification; extract; extraction; free text; GPT; GPT-4; language model; large language models; LLM; natural language processing; NLP; nonopiod analgesia; review methodology; review methods; screening; systematic review; systematic; unstructured data

XSL•FO
RenderX

Introduction

Knowledge synthesis, the process of integrating and summarizing relevant studies in the literature to gain an improved understanding of a topic, is a key component in identifying knowledge gaps and informing future research endeavors on a topic of interest [1,2]. Systematic and scoping reviews are among the most commonly used and rigorous forms of knowledge synthesis across multiple disciplines [1,2]. Given that the results from systematic and scoping reviews can inform guidelines, protocols, and decision-making processes, particularly for stakeholders in the realms of health care, the quality of the evidence presented by such reviews can significantly impact generated recommendations [3].

The quality of systematic and scoping reviews is highly dependent on the comprehensiveness of the database searches and the subsequent article screening processes. Overlooking relevant articles during these critical steps can lead to bias [4], while including discrepant studies can yield misleading conclusions and increase discordant heterogeneity [5]. Thus, guidelines surrounding the conduct of clinical reviews, such as the Cochrane Handbook [6], recommend that article screening be completed in duplicate by at least 2 independent reviewers.

However, duplicate screening effectively doubles the financial and human resources needed to complete systematic reviews compared to single screening. This is especially problematic for small research groups, review projects with broad inclusion criteria (such as network meta-analyses), or time-constrained review projects (such as reviews relating to COVID-19 during the early stages of the pandemic) [7,8]. Additionally, there is often substantial interrater variability in screening decisions, leading to additional time spent on discussions to resolve disagreements [9]. Due to the time constraints and wasted resources that are often features of duplicate screening, research studies may also include a more tailored, sensitive search strategy that can lead to missing several articles during the retrieval process [10]. Furthermore, although the nuances of each study differ, many systematic reviews may contain thousands of retrieved articles, only to exclude the majority (ie, up to 90%) from the title and abstract screening [10,11].

Recent developments in artificial intelligence and machine learning have made it possible to semiautomate or fully automate repetitive steps within the systematic review workflow [12-14]. Prominent examples of such applications include RobotReviewer [15], TrialStreamer [16], Research Screener [7], DistillerSR [17], and Abstrackr [18], which are artificial intelligence models developed to extract information from scientific articles or abstracts to judge study quality and infer treatment effects. More specifically, RobotReviewer (2016) was shown to have similar capabilities to assess the risk of bias assessment as a human reviewer, only differing by around 7% in accuracy [19]. Similarly, TrialStreamer was a system developed to extract key elements of information from full texts, such as inferring which interventions in a clinical paper worked best, along with comparisons in study outcomes between all relevant extracted full texts of a topic indexed on MEDLINE [20].

While there have been previous attempts at automating the title and abstract screening process, they often involved labor- or computationally-intensive labeling, pretraining, or vectorizations [21]. For instance, Rayyan and Abstrackr are 2 free web tools that provide a semiautomated approach to article filtering by using natural language processing algorithms to learn when and where a reviewer includes or excludes an article and subsequently mimics a similar approach [22,23]. Rayyan also demonstrated high specificity, wherein 98% of all relevant articles were included after the tool had screened 75% of all articles to be analyzed in a study [24]. While automation using these tools was found to save time, there was still minimal to substantive risk that there would be missing studies if the tool were fully independent or automated [22,23]. Furthermore, current programs may use previously standard methods, including n-grams, in comparison to more updated techniques, such as the generative pretrained transformer (GPT) model, which is trained with data from a general domain and does not require additional training to learn embeddings that can represent the semantics and contexts of words in relation to other words [25,26].

In this paper, we introduce a novel workflow to screen titles and abstracts for clinical reviews by providing plain language prompts to the publicly available OpenAI GPT application programming interface (API). We aimed to assess GPT models' ability to accurately and efficiently identify relevant titles and abstracts from real-world clinical review data sets, as well as their ability to explain their decisions and reflect on incorrect classifications. We compare the performance of our model against ground truth labeling by 2 independent human reviewers across 6 review papers in the screening of over 24,000 titles and abstracts.

Methods

Overview

In our study, we obtained a corpus of title and abstract data sets that have already been filtered by a minimum of 2 human reviewers to train our model (Figure 1). Subsequently, we created a Python script that provides the screening criteria for each paper to the OpenAI Chat GPT or GPT-4 API, depending on the input token length. We then passed each paper to the API using a consistent instruction prompt to determine whether a paper should be included or excluded based on the contents of its title and abstract. The overall accuracy (computed by dividing papers selected by both GPT and human reviewers by the total number of papers), sensitivity of both included and excluded papers, and interrater reliability through Cohen κ and prevalence-adjusted and bias-adjusted κ (PABAK) were computed against the human-reviewed papers:

Where k is the number of categories and p_{obs} is the proportion of included papers. All data and code are available in Mendeley data sets [27].

XSL•FO
RenderX

Figure 1. Overview of the Python script to automate screening with the generative pretrained transformer (GPT) application programming interface (API). LLM: large language model.

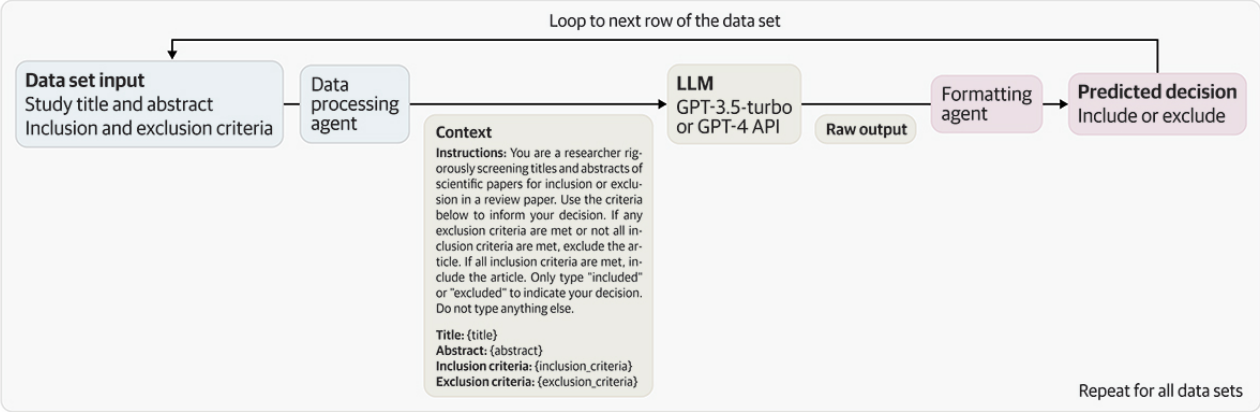

Data Collection

To validate our proposed inclusion and exclusion methodology, we obtained 6 title and abstract screening data sets from different systematic and scoping reviews previously published by the authors of this study, each screened by 2 independent reviewers with conflicts resolved through consensus. These projects cover various medical science topics and vary in size, methodology, and complexity of screening criteria (Table 1 and Table S1 in Multimedia Appendix 1 [28-33]). We obtained the inclusion and exclusion decision from expert reviewers for each title and abstract entry, as well as the criteria provided to the expert reviewers during the screening process. A summary of the review characteristics is presented in Table 2.

Table 1. Included studies and their characteristics. The first 5 data sets are systematic reviews with meta-analyses. The last study is a scoping review.

Study title	Data set name	Included studies (538/24,307), n/N	Study type	Study topic
Efficacy and Safety of Ivermectin for the Treatment of COVID-19: A Systematic Review and Meta-Analysis [29]	IVM[a]	35/279	Systematic review and meta-analysis of randomized and nonrandomized trials	COVID-19 treatment and antimalarials
Efficacy and Safety of Selective Serotonin Reuptake Inhibitors in COVID-19 Management: A Systematic Review and Meta-Analysis [30]	SSRI[b]	29/3989	Systematic review and meta-analysis of randomized and nonrandomized trials	COVID-19 treatment and antidepressants
Efficacy of Lopinavir-Ritonavir Combination Therapy for the Treatment of Hospitalized COVID-19 Patients: A Meta-Analysis [31]	LPVR[c]	91/1456	Systematic review and meta-analysis of randomized and nonrandomized trials	COVID-19 treatment and antiretrovirals
The Use of Acupuncture in Patients With Raynaud's Syndrome: A Systematic Re-View and Meta-Analysis of Randomized Controlled Trials [32]	RAYNAUDS[d]	6/942	Systematic review and meta-analysis of randomized and nonrandomized trials	Raynaud syndrome and acupuncture
Comparative Efficacy of Adjuvant Non-Opioid Analgesia in Adult Cardiac Surgical Patients: A Network Meta-Analysis [33]	NOA[e]	354/14,771	Systematic review and meta-analysis of randomized and nonrandomized trials	Postoperative pain and analgesics
Assessing the Research Landscape and Utility of LLMs[f] in the Clinical Setting: Protocol for a Scoping Review [g]	LLM	23/2870	Scoping review	Machine learning in clinical medicine

[a]IVM: ivermectin.

[b]SSRI: selective serotonin reuptake inhibitor.

[c]LPVR: lopinavir-ritonavir.

[d]RAYNAUDS: Raynaud syndrome.

[e]NOA: nonopioid analgesia.

[f]LLM: large language model.

[g]Registered with Open Science Framework [28].

XSL•FO
RenderX

Table 2. Data formatting for the Python script automating screening with the generative pretrained transformer application programming interface. All non-English characters were removed before analysis.

Data	Columns
df_info	• Dataset Name (str): name of data set • Inclusion Criteria (str): screening inclusion criteria • Exclusion Criteria (str): screening exclusion criteria
Dataset[a]	• Title (str): paper title • Abstract (str): paper abstract

[a]The name of the data set must match Dataset Name in df_info.

App Creation

Given a data set, df_info, containing information about inclusion and exclusion criteria of the data sets containing titles and abstracts to be reviewed, the app calls the OpenAI GPT API to classify each paper to be screened as either included or excluded. The app was coded in Python. The prompt given to the GPT API is provided in Textbox 1.

Textbox 1. Prompt given to the generative pretrained transformer application programming interface.

Instructions: You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article. Only type "included" or "excluded" to indicate your decision. Do not type anything else.

Abstract: {abstract}

Inclusion criteria: {inclusion_criteria}

Exclusion criteria: {exclusion_criteria}

Decision:

Where "Decision:" is whether GPT API includes or excludes the article. Thus, the algorithm is as follows:

```
data_df <- load(df_info)
for each dataset in data_df: for each row in dataset:
prompt <- instructions + title + abstract + inclusion criteria \
+ exclusion criteria decision <- GPT(prompt) row['decision'] <- decision
save(dataset)
```

Assessment and Data Analysis

After the app was run on all data sets included in our analysis, the following metrics were computed: accuracy, macro F_1-score, sensitivity for decision tags, κ, and PABAK. A subset of the results was selected for the GPT models to explain their reasoning. The following prompt was appended to the beginning of the original prompt given to the API: "Explain your reasoning for the decision given with the information below." The human and GPT decisions were appended to the end of the prompt. A subset of incorrect results was selected for GPT to reflect on its incorrect answers. The following prompt was appended to the beginning of the original prompt given to the API: "Explain your reasoning for why the decision given was incorrect with the information below." The human and GPT decisions were appended to the end of the prompt.

Results

The overall accuracy of the GPT models was 0.91, the sensitivity of included papers was 0.76, and the sensitivity of excluded papers was 0.91 (Table 3 and Figure 2). On the nonopioid analgesia (NOA) data set (354/14,771 included abstracts), the model ran for 643 minutes and 50.8 seconds, with an approximate cost of US $25. The data set characteristics are detailed in Table 1, the model performance is in Table 3 and visualized in Figure 2, and the reasoning from GPT is tabulated in Table 4.

XSL•FO
RenderX

Table 3. Performance of generative pretrained transformer (GPT) in screening titles and abstracts against a human reviewer's ground truth. κ (human) is the agreement between 2 independent human reviewers. κ (screen) is the agreement between GPT and the final papers included and excluded in each data set.

Data set	Accuracy	Macro F_1-score	Sensitivity (included)	Sensitivity (excluded)	κ (human)	κ (screen)	PABAK[a]
IVM[b]	0.748	0.610	0.686	0.756	0.72	0.26	0.78
SSRI[c]	0.846	0.595	0.966	0.949	0.58	0.21	0.99
LPVR[d]	0.949	0.613	0.593	0.862	0.51	0.25	0.88
RAYNAUDS[e]	0.965	0.607	0.833	0.966	0.91	0.22	0.99
NOA[f]	0.895	0.601	0.782	0.898	0.35	0.23	0.95
LLM[g]	0.943	0.594	1.000	0.942	0.69	0.21	0.98
Total (weighted)	0.907	0.600	0.764	0.910	0.46	0.22	0.96
Total (macro)	0.891	0.664	0.810	0.900	0.63	0.23	0.93

[a]PABAK: prevalence-adjusted and bias-adjusted κ.

[b]IVM: ivermectin.

[c]SSRI: selective serotonin reuptake inhibitor.

[d]LPVR: lopinavir-ritonavir.

[e]RAYNAUDS: Raynaud syndrome.

[f]NOA: nonopioid analgesia.

[g]LLM: large language model.

XSL•FO
RenderX

Figure 2. (A-F) Confusion matrices for the included and excluded papers for all data sets and (G) the overall performance. IVM: ivermectin; LLM: large language model; LPVR: lopinavir-ritonavir; NOA: nonopioid analgesia; RAYNAUDS: Raynaud's syndrome; SSRI: selective serotonin reuptake inhibitor.

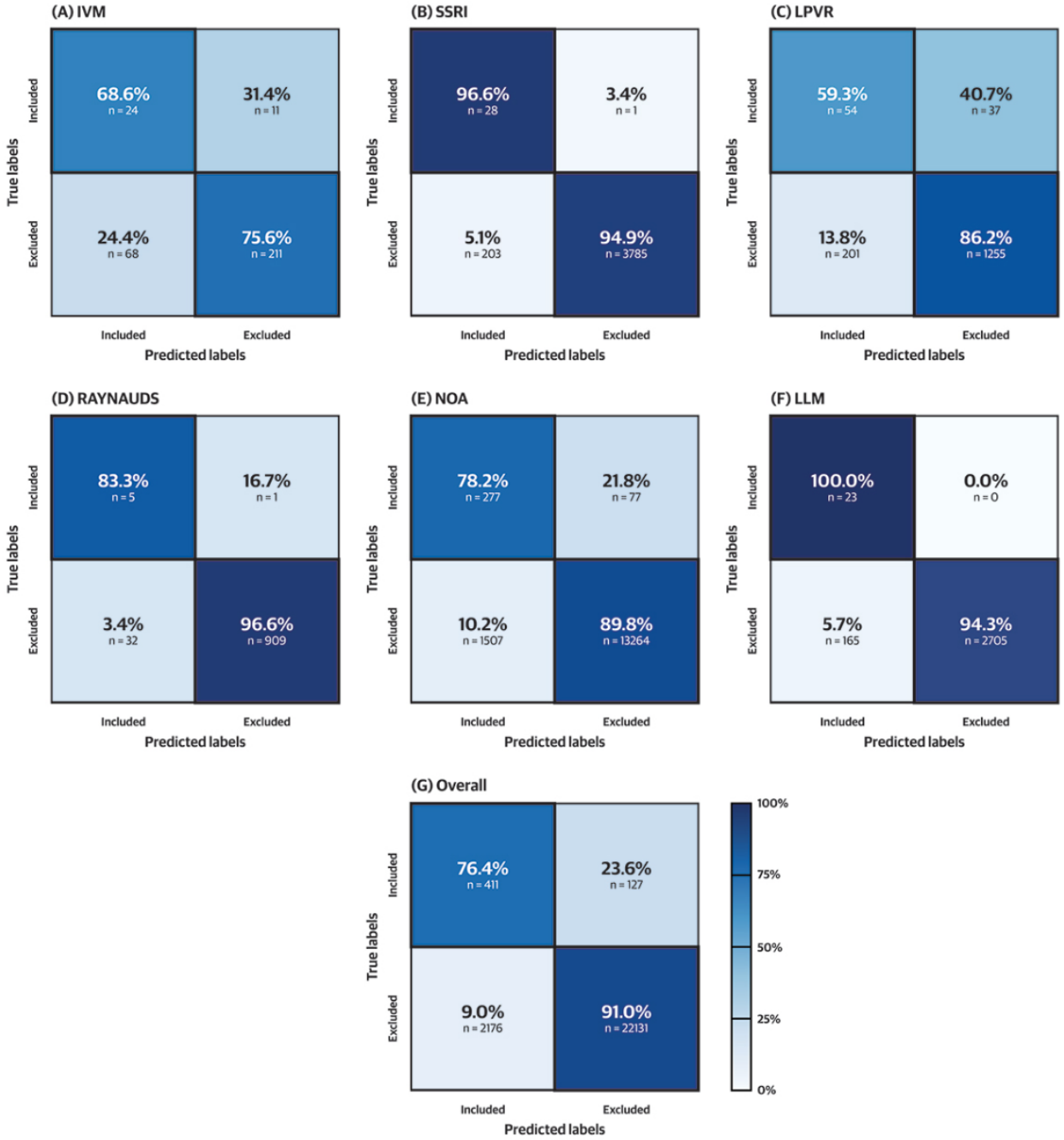

XSL•FO
RenderX

Table 4. Decision and reasoning of the generative pretrained transformer (GPT) application programming interface for a subset of 4 example titles and abstracts. Full abstracts are in ivermectin (IVM) data set [27].

Prompt	Decision and reasoning
• Title: Use of Ivermectin in the Treatment of COVID-19: A Pilot Trial • Abstract: Objectives: In this randomized open-label trial pilot study we assessed the antiviral effects and safety of various doses of ivermectin in patients with mild clinical symptoms of COVID-19. Method(s): Patients were randomly assigned to receive standard of care (SOC) treatment at hospital admission...	• Human: Included • GPT: Included • Reasoning: The study meets all of the inclusion criteria. It is a randomized controlled trial that compares ivermectin against standard of care treatment in adult COVID-19 patients. The study reports outcomes related to time to viral clearance and adverse events. Additionally, the study is published in a peer-reviewed journal. Therefore, the article should be included in the review paper.
• Title: What Does the COVID-19 Pandemic Mean for the Next Decade of Onchocerciasis Control and Elimination? • Abstract: Background: Mass drug administration (MDA) of ivermectin for onchocerciasis has been disrupted by the coronavirus disease 2019 (COVID-19) pandemic. Mathematical modelling can help predict how missed/delayed MDA will affect short-term epidemiological trends and elimination prospects by 2030...	• Human: Excluded • GPT: Excluded • Reasoning: The article does not meet any of the inclusion criteria or exclusion criteria listed. It is not a randomized controlled trial or comparative cohort study, it does not compare ivermectin against standard of care or control group/placebo, it does not include adult COVID-19 inpatients and/or outpatients, and it does not report any of the specified outcomes. Additionally, there is no indication that the article is non-peer-reviewed. Therefore, the article is excluded.
• Title: Use of Ivermectin Is Associated With Lower Mortality in Hospitalized Patients With Coronavirus Disease 2019: The Ivermectin in COVID Nineteen Study • Abstract: Background: Ivermectin was shown to inhibit severe acute respiratory syndrome coronavirus 2 replication in vitro, which has led to off-label use, but clinical efficacy has not been described previously. Research Question: Does ivermectin benefit hospitalized coronavirus disease 2019 (COVID-19) patients?...	• Human: Included • GPT: Excluded • Reasoning: The article meets all of the inclusion criteria and none of the exclusion criteria, therefore it should be included in the review paper. • Reflection: The given decision of excluding the article is incorrect. The article meets all the inclusion criteria as it is a non-randomized comparative cohort study that compares ivermectin against standard of care or control group/placebo, includes adult COVID-19 inpatients, and reports mortality incidence. The article does not use ivermectin for prophylaxis of COVID-19 and is a peer-reviewed article. Therefore, the correct decision should be "Included."
• Title: What to Expect From Different Drugs Used in the Treatment of COVID-19: A Study on Applications and in Vivo and in Vitro Results • Abstract: The end of 2019 was marked by the emergence of a new type of coronavirus (SARS-CoV-2), which has killed more than 240,000 people...	• Human: Excluded • GPT: Included • Reasoning: The study is not a randomized controlled trial or a non-randomized comparative cohort study, which is an inclusion criteria. Therefore, the article does not meet all inclusion criteria and must be excluded. • Reflection: As per the given inclusion and exclusion criteria, the decision given is incorrect. The abstract does not mention the use of ivermectin in any of the studies. Therefore, the article cannot be included based on the inclusion criteria.

Discussion

Overview

In this study, we assessed the performance of the OpenAI GPT API in the context of clinical review paper inclusion and exclusion criteria selection. We report an overall accuracy of 0.91 and a PABAK of 0.96, indicating a high level of agreement between the app's decisions and the reference standard. However, the κ was low, ranging from 0.21 to 0.26, largely due to the imbalanced nature of the data sets in this study. The sensitivity of the included papers was 0.76, suggesting that the app needs improvement to correctly identify relevant papers (Table 3 and Figure 2). The sensitivity of excluded papers was 0.91, showing promise in excluding irrelevant papers. These results highlight the potential of large language models (LLMs) to support the clinical review process.

Implications of GPT API's Performance in the Review Process

GPT's performance has several implications for the efficiency and consistency of clinical review paper inclusion and exclusion criteria selection. By prioritizing the workflow and acting as an aid rather than a replacement for researchers and reviewers, the GPT and other large language models have the potential to streamline the review process. This enhanced efficiency could save valuable time and effort for researchers and clinicians, allowing them to focus on more complex tasks and in-depth analysis. Further, the API does not require pretraining or seed articles and can provide reasoning for its decision to either include or exclude papers, an aspect traditional natural language processing algorithms lack in automated or semiautomated paper screening (Table 4). Interestingly, upon being asked to explain its reasoning for a subset of incorrect classifications, GPT corrected its initial decision. Ultimately, this increased efficiency, paired with reasoning capabilities, could contribute to the overall quality of clinical reviews, leading to more accurate and reliable conclusions in medical research.

The use of LLMs in the review process could also promote consistency in the selection of relevant papers. By automating certain aspects of the process and acting as an aid to researchers and clinicians, the model can streamline the review process and help reduce the potential for human error and bias, leading to

XSL•FO
RenderX

more objective and reliable results [34]. This increased consistency could, in turn, improve the overall quality of the evidence synthesized in clinical reviews, providing a more robust foundation for medical decision-making and the development of clinical guidelines.

The potential of LLMs as a decision tool becomes particularly valuable when resources are limited. In such situations, LLMs can be used as a first-pass decision aid, streamlining the review process, and allowing human screeners to focus on a smaller, more relevant subset of papers. By automating the initial screening process, LLMs can help reduce the workload for researchers and clinicians, enabling them to allocate their time and effort more efficiently.

In particular, using the GPT API as a first-pass decision aid can also help mitigate the risk of human error and bias in the initial screening phase, promoting a more objective and consistent selection of papers. While the API's sensitivity for including relevant papers may not be perfect, its high specificity for excluding irrelevant papers can still provide valuable support in narrowing down the pool of potentially relevant studies [10]. This can be particularly beneficial in situations where a large number of papers need to be screened and human resources are scarce [35].

Comparison to Other Tools

The comparison of our proposed machine learning method to other tools, such as Abstrackr [18], DistillerSR [17], and RobotAnalyst [36], provides evidence of its efficacy and reliability in the context of systematic review processes. On a data set of 24,307 abstracts and titles, our model achieved an accuracy of 0.91 and comparable sensitivity of 0.91 and 0.76 for excluded and included papers, respectively. The significant interrater agreement (κ=0.96) between our proposed method and consensus-based human decisions, juxtaposed to the lower interrater variability between 2 independent human screeners (κ=0.46), emphasizes the model's robustness. In comparison, Abstrackr reported overall sensitivities of 0.96, 0.79, 0.92, and 0.82 on data sets ranging from 5243 to 47,385 records. When comparing the proportion of missed records across Abstrackr, DistillerSR, and RobotAnalyst on nonpublic medical title and abstract screening data sets, Abstrackr exhibited the lowest proportions of missed records, namely 28%, 5%, and 0%, respectively [37]. Conversely, DistillerSR showed a high proportion of missed records, reaching up to 100% in the last data set. RobotAnalyst's performance fell between the 2, with missed proportions of 70%, 23%, and 100%, respectively. Future work will explore comparative analyses in greater depth and on a broader array of data sets to compare state-of-the-art screening tools.

Limitations and Challenges in Implementing GPT API in the Review Process

While the GPT API shows promise in streamlining the review process, it is important to acknowledge its limitations and challenges. One notable limitation is the disparity between the high specificity of 0.91 for excluding papers and the lower sensitivity of 0.76 for including papers. This discrepancy suggests that while the API effectively excludes irrelevant papers, it may not be as proficient in identifying relevant papers for inclusion. This could lead to the omission of important studies in the review process, potentially affecting the comprehensiveness and quality of the final review. Therefore, the GPT API should not be considered a replacement for human expertise. Instead, it should be viewed as a complementary tool that can enhance the efficiency and consistency of the review process. Human screeners should still be involved in the final decision-making process, particularly in cases where the API's sensitivity for including relevant papers may be insufficient [7]. Another limitation arises in the selection of data sets for screening; 3 of the 6 data sets focused on the efficacy of various drugs for COVID-19, potentially limiting the generalizability of the results from other types of studies. Further work will assess GPT on a greater diversity of studies. By combining the strengths of the GPT API with human expertise, researchers can optimize the review process and ensure the accuracy and comprehensiveness of the final review.

Future Research and Development

Several avenues for future research and development include refining the GPT API's performance in the clinical review paper context, incorporating metadata such as study type and year, and exploring few-shot learning approaches. Additionally, training a generator-discriminator model through fine-tuning could improve the API's performance [38]. Expanding the application of the GPT API to other areas of medical research or literature review could also be explored. This would involve large language models for tasks such as identifying and extracting study design information, patient characteristics, and adverse events. As the maximum token length increases with future iterations of the GPT model, screening entire papers may become feasible [39]. Furthermore, exploring the use of LLMs to generate clinical review papers could be a promising research direction.

Conclusions

The GPT API shows potential as a valuable tool for improving the efficiency and consistency of clinical review paper inclusion and exclusion criteria selection. While there are limitations and challenges to its implementation, its performance in this study suggests that it could have a broader impact on clinical review paper writing and medical research. Future research and development should focus on refining the API's performance, expanding its applications, and exploring its potential in other aspects of clinical research.

XSL•FO
RenderX

Acknowledgments

We would like to acknowledge the following expert reviewers for providing the screening decisions in the review data sets used in this study and for agreeing to make the data sets publicly available: Abhinav Pillai, Mike Paget, Christopher Naugler, Kiyan Heybati, Fangwen Zhou, Myron Moskalyk, Saif Ali, Chi Yi Wong, Wenteng Hou, Umaima Abbas, Qi Kang Zuo, Emma Huang, Daniel Rayner, Cristian Garcia, Harikrishnaa Ba Ramaraju, Oswin Chang, Zachary Silver, Thanansayan Dhivagaran, Elena Zheng, and Shayan Heybati.

Authors' Contributions

EG contributed to conceptualization, methodology, software, formal analysis, investigation, writing the original draft, reviewing, editing, visualization, supervision, and project administration. MG was responsible for conceptualization, methodology, investigation, writing the original draft, reviewing, editing, supervision, and project administration. JD and YJP were involved in methodology, software, formal analysis, investigation, data curation, writing the original draft, and visualization. MP and CN contributed to writing, reviewing, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Included studies and their inclusion and exclusion criteria.
[DOCX File , 20 KB - jmir_v26i1e48996_app1.docx]

References

1. Sargeant JM, O'Connor AM. Scoping reviews, systematic reviews, and meta-analysis: applications in veterinary medicine. Front Vet Sci 2020;7:11 [FREE Full text] [doi: 10.3389/fvets.2020.00011] [Medline: 32047759]
2. Garritty C, Stevens A, Hamel C, Golfam M, Hutton B, Wolfe D. Knowledge synthesis in evidence-based medicine. Semin Nucl Med 2019;49(2):136-144. [doi: 10.1053/j.semnuclmed.2018.11.006] [Medline: 30819393]
3. Luchini C, Veronese N, Nottegar A, Shin JI, Gentile G, Granziol U, et al. Assessing the quality of studies in meta-research: review/guidelines on the most important quality assessment tools. Pharm Stat 2021;20(1):185-195. [doi: 10.1002/pst.2068] [Medline: 32935459]
4. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol 2020;121:20-28 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.01.005] [Medline: 31972274]
5. Fletcher J. What is heterogeneity and is it important? BMJ 2007;334(7584):94-96 [FREE Full text] [doi: 10.1136/bmj.39057.406644.68] [Medline: 17218716]
6. The Cochrane Collaboration, Higgins JPT, Thomas J. Cochrane Handbook for Systematic Reviews of Interventions. London: The Cochrane Collaboration; 2021.
7. Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev 2021;10(1):93 [FREE Full text] [doi: 10.1186/s13643-021-01635-3] [Medline: 33795003]
8. Clark J, McFarlane C, Cleo G, Ramos CI, Marshall S. The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: case study. JMIR Med Educ 2021;7(2):e24418 [FREE Full text] [doi: 10.2196/24418] [Medline: 34057072]
9. Tuijn S, Janssens F, Robben P, van den Bergh H. Reducing interrater variability and improving health care: a meta-analytical review. J Eval Clin Pract 2012;18(4):887-895. [doi: 10.1111/j.1365-2753.2011.01705.x] [Medline: 21726359]
10. Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. Syst Rev 2015;4(1):6 [FREE Full text] [doi: 10.1186/2046-4053-4-6] [Medline: 25588387]
11. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large‐evidence systematic reviews and meta‐analyses. Res Synth Methods 2019;10(3):330-342 [FREE Full text] [doi: 10.1002/jrsm.1354]
12. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 2019;8(1):163 [FREE Full text] [doi: 10.1186/s13643-019-1074-9] [Medline: 31296265]
13. Marshall IJ, Trikalinos TA, Soboczenski F, Yun HS, Kell G, Marshall R, et al. In a pilot study, automated real-time systematic review updates were feasible, accurate, and work-saving. J Clin Epidemiol 2023;153:26-33 [FREE Full text] [doi: 10.1016/j.jclinepi.2022.08.013] [Medline: 36150548]
14. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. Res Synth Methods 2022;13(3):353-362. [doi: 10.1002/jrsm.1553] [Medline: 35174972]

XSL•FO
RenderX

15. Marshall I, Kuiper J, Banner E, Wallace BC. Automating biomedical evidence synthesis: RobotReviewer. : Association for Computational Linguistics; 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations; July 30-August 4, 2017; Vancouver, Canada p. 7-12 URL: https://aclanthology.org/P17-4002.pdf [doi: 10.18653/v1/p17-4002]
16. Marshall IJ, Nye B, Kuiper J, Noel-Storr A, Marshall R, Maclean R, et al. Trialstreamer: a living, automatically updated database of clinical trial reports. J Am Med Inform Assoc 2020;27(12):1903-1912 [FREE Full text] [doi: 10.1093/jamia/ocaa163] [Medline: 32940710]
17. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening—impact on reviewer-relevant outcomes. BMC Med Res Methodol 2020;20(1):256 [FREE Full text] [doi: 10.1186/s12874-020-01129-1] [Medline: 33059590]
18. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018;7(1):45 [FREE Full text] [doi: 10.1186/s13643-018-0707-8] [Medline: 29530097]
19. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc 2016;23(1):193-201 [FREE Full text] [doi: 10.1093/jamia/ocv044] [Medline: 26104742]
20. Nye BE, Nenkova A, Marshall IJ, Wallace BC. Trialstreamer: mapping and browsing medical evidence in real-time. Proc Conf 2020;2020:63-69 [FREE Full text] [doi: 10.18653/v1/2020.acl-demos.9] [Medline: 34136886]
21. Moreno-Garcia CF, Jayne C, Elyan E, Aceves-Martins M. A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews. Decis Anal J 2023;6:100162 [FREE Full text] [doi: 10.1016/j.dajour.2023.100162]
22. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010;11:55 [FREE Full text] [doi: 10.1186/1471-2105-11-55] [Medline: 20102628]
23. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
24. Olofsson H, Brolund A, Hellberg C, Silverstein R, Stenström K, Österberg M, et al. Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. Res Synth Methods 2017;8(3):275-280. [doi: 10.1002/jrsm.1237] [Medline: 28374510]
25. Shree P. The journey of Open AI GPT models. Medium. 2020. URL: https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2 [accessed 2023-04-19]
26. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4(1):5 [FREE Full text] [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
27. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models. Mendeley Data. 2023. URL: https://data.mendeley.com/datasets/np79tmhkh5/1 [accessed 2023-12-15]
28. Assessing the research landscape and utility of LLMs in the clinical setting: protocol for a scoping review. OSF Registries. URL: https://osf.io/498k6 [accessed 2023-12-15]
29. Deng J, Zhou F, Ali S, Heybati K, Hou W, Huang E, et al. Efficacy and safety of ivermectin for the treatment of COVID-19: a systematic review and meta-analysis. QJM 2021;114(10):721-732 [FREE Full text] [doi: 10.1093/qjmed/hcab247] [Medline: 34570241]
30. Deng J, Rayner D, Ramaraju HB, Abbas U, Garcia C, Heybati K, et al. Efficacy and safety of selective serotonin reuptake inhibitors in COVID-19 management: a systematic review and meta-analysis. Clin Microbiol Infect 2023;29(5):578-586 [FREE Full text] [doi: 10.1016/j.cmi.2023.01.010] [Medline: 36657488]
31. Deng J, Zhou F, Hou W, Heybati K, Ali S, Chang O, et al. Efficacy of lopinavir-ritonavir combination therapy for the treatment of hospitalized COVID-19 patients: a meta-analysis. Future Virol 2021 [FREE Full text] [doi: 10.2217/fvl-2021-0066] [Medline: 35145560]
32. Zhou F, Huang E, Zheng E, Deng J. The use of acupuncture in patients with Raynaud's syndrome: a systematic review and meta-analysis of randomized controlled trials. Acupunct Med 2023;41(2):63-72 [FREE Full text] [doi: 10.1177/09645284221076504] [Medline: 35608095]
33. Heybati K, Zhou F, Lynn MJ, Deng J, Ali S, Hou W, et al. Comparative efficacy of adjuvant nonopioid analgesia in adult cardiac surgical patients: a network meta-analysis. J Cardiothorac Vasc Anesth 2023;37(7):1169-1178. [doi: 10.1053/j.jvca.2023.03.018] [Medline: 37088644]
34. Zhang Y, Liang S, Feng Y, Wang Q, Sun F, Chen S, et al. Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. Syst Rev 2022;11(1):11 [FREE Full text] [doi: 10.1186/s13643-021-01881-5] [Medline: 35031074]
35. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021;3(2):125-133 [FREE Full text] [doi: 10.1038/s42256-020-00287-7]

XSL•FO
RenderX

36. Przybyła P, Brockmeier AJ, Kontonatsios G, Le Pogam M, McNaught J, von Elm E, et al. Prioritising references for systematic reviews with RobotAnalyst: a user study. Res Synth Methods 2018;9(3):470-488 [FREE Full text] [doi: 10.1002/jrsm.1311] [Medline: 29956486]
37. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. In: AHRQ Methods for Effective Health Care. Rockville, MD: Agency for Healthcare Research and Quality (US); 2019.
38. Schade M. Fine-tuning a classifier to improve truthfulness. OpenAI. URL: https://help.openai.com/en/articles/5528730-fine-tuning-a-classifier-to-improve-truthfulness [accessed 2023-04-20]
39. Joshua J. What is the difference between the GPT-4 models? OpenAI. URL: https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models [accessed 2023-04-19]

Abbreviations

API: application programming interface
GPT: generative pretrained transformer
LLM: large language model
NOA: nonopioid analgesia
PABAK: prevalence and bias-adjusted kappa

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 14.05.23; peer-reviewed by T Kang, M Chatzimina, I Bojic; comments to author 30.08.23; revised version received 30.08.23; accepted 28.09.23; published 12.01.24.

Please cite as:
Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C
Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study
J Med Internet Res 2024;26:e48996
URL: https://www.jmir.org/2024/1/e48996
doi:10.2196/48996
PMID:38214966

©Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Michael Paget, Christopher Naugler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study

Xiaocong Liu[1,2*], MS; Jiageng Wu[2*], MS; An Shao[1], MM; Wenyue Shen[1], MB; Panpan Ye[1], MD; Yao Wang[1], MD; Juan Ye[1], MD; Kai Jin[1], MD; Jie Yang[2], PhD

[1]Eye Center, The Second Affiliated Hospital, Zhejiang University, Zhejiang, China
[2]School of Public Health, Zhejiang University School of Medicine, Zhejiang, China
[*]these authors contributed equally

Corresponding Author:
Kai Jin, MD
Eye Center
The Second Affiliated Hospital
Zhejiang University
88 Jiefang Road, Hangzhou
Zhejiang, 310009
China
Phone: 86 571 87783907
Email: jinkai@zju.edu.cn

Abstract

Background: Benefiting from rich knowledge and the exceptional ability to understand text, large language models like ChatGPT have shown great potential in English clinical environments. However, the performance of ChatGPT in non-English clinical settings, as well as its reasoning, have not been explored in depth.

Objective: This study aimed to evaluate ChatGPT's diagnostic performance and inference abilities for retinal vascular diseases in a non-English clinical environment.

Methods: In this cross-sectional study, we collected 1226 fundus fluorescein angiography reports and corresponding diagnoses written in Chinese and tested ChatGPT with 4 prompting strategies (direct diagnosis or diagnosis with a step-by-step reasoning process and in Chinese or English).

Results: Compared with ChatGPT using Chinese prompts for direct diagnosis that achieved an F_1-score of 70.47%, ChatGPT using English prompts for direct diagnosis achieved the best diagnostic performance (80.05%), which was inferior to ophthalmologists (89.35%) but close to ophthalmologist interns (82.69%). As for its inference abilities, although ChatGPT can derive a reasoning process with a low error rate (0.4 per report) for both Chinese and English prompts, ophthalmologists identified that the latter brought more reasoning steps with less incompleteness (44.31%), misinformation (1.96%), and hallucinations (0.59%) (all $P<.001$). Also, analysis of the robustness of ChatGPT with different language prompts indicated significant differences in the recall ($P=.03$) and F_1-score ($P=.04$) between Chinese and English prompts. In short, when prompted in English, ChatGPT exhibited enhanced diagnostic and inference capabilities for retinal vascular disease classification based on Chinese fundus fluorescein angiography reports.

Conclusions: ChatGPT can serve as a helpful medical assistant to provide diagnosis in non-English clinical environments, but there are still performance gaps, language disparities, and errors compared to professionals, which demonstrate the potential limitations and the need to continually explore more robust large language models in ophthalmology practice.

(J Med Internet Res 2024;26:e51926) doi:10.2196/51926

KEYWORDS

large language models; ChatGPT; clinical decision support; retinal vascular disease; artificial intelligence

XSL•FO
RenderX

Introduction

The global population of individuals with visual impairments exceeded 2.2 billion in 2019 and continues to rise [1]. As the leading causes of blindness, retinal vascular diseases are characterized by a complex array of clinical manifestations [2]. Fundus fluorescein angiography (FFA), which uses an injected fluorescent dye to examine circulation in the retina and choroid, is a specialized ophthalmic test used to visualize the retinal vasculature [3]. In practice, interpreting FFA results and making a diagnosis requires laborious analysis by experienced ophthalmologists.

In recent years, significant developments in deep learning approaches, which are extensively utilized, have rendered them a promising way for auxiliary diagnosis of retinal vascular diseases. The existing research has mainly focused on developing convolutional neural network algorithms for lesion detection in FFA images [4-8], such as microaneurysms, leakages, nonperfusion areas, and neovascularization. Further, some studies focused on automatically generating FFA reports [9,10], which can highlight abnormalities for ophthalmologists and provide a theoretical basis for disease diagnosis. However, few studies were devoted to the diagnosis of retinal vascular disease based on FFA reports. The main challenges of using natural language processing to diagnose retinal vascular diseases can be summarized as follows: (1) different interpretation of FFA images by different ophthalmologists, (2) varied ophthalmological terms contained in FFA reports, (3) time-varying imaging features contained in FFA reports, and (4) smaller data volume caused by the high cost and possible side effects of FFA.

Recently, large language models (LLMs) like ChatGPT [11] have demonstrated exceptional performance in various tasks due to their rich internal knowledge and strong deductive reasoning abilities [12-16]. However, the related research within the medical field primarily focuses on knowledge assessment [17-20], and a comprehensive evaluation of ChatGPT's capabilities in ophthalmology for disease diagnosis is lacking. Additionally, although existing LLMs demonstrate impressive cross-language understanding abilities, they may lead to significant disparities in non–English-specific fields because they were primarily trained on English corpora [21,22]. Therefore, in this study, by exploring ChatGPT's ability to understand Chinese FFA reports, our objectives were to evaluate ChatGPT's diagnostic performance and inference abilities for retinal vascular diseases in a non-English clinical environment and to find appropriate prompt strategies under these scenarios.

Methods

Data Preparation

We collected 1226 Chinese FFA reports and the corresponding clinical diagnoses of 728 patients from the Eye Center of the Second Affiliated Hospital of Zhejiang University (SAHZU) between August 2016 and September 2021. The clinical diagnosis of each eye was either classified as normal or one of the 6 primary retinal vascular diseases: diabetic retinopathy (DR), wet age-related macular degeneration, central serous chorioretinopathy (CSC), branch retinal vein occlusion (BRVO), central retinal vein occlusion (CRVO), and Vogt-Koyanagi-Harada disease (VKH). The clinical diagnosis was based on clinical information from the patients, primarily the FFA images and reports.

Ethical Considerations

Ethical approval was obtained from the Ethics Committee in the SAHZU School of Medicine (2019-428). This research involves medical records data. We ensured that the medical records were deidentified and all private information was removed. The Institutional Review Board agreed to share access to the data with third parties, including sending it through application programing interfaces (APIs) provided by companies like OpenAI, or using it on web-based platforms like ChatGPT.

Diagnosis of Retinal Vascular Diseases Using ChatGPT

To diagnose the patient's eye status based on the FFA report with ChatGPT, we designed a fixed instruction that concatenates the patient's FFA report as the whole prompt for ChatGPT. The instruction consists of a specific task description and all alternative conditions. To fully exploit the potential of ChatGPT, we implemented different prompting strategies to investigate the potential effect and find the most appropriate way to apply it. First, we used the direct prompting strategy that requires ChatGPT to directly output the final option without explanations. Second, inspired by chain-of-thought prompting technology [23], we adopted a step prompting strategy to elicit the detailed reasoning process, which provides interpretability for the disease diagnosis. Finally, ChatGPT was primarily trained on English corpora and may have difficulty recognizing instructions and FFA reports in Chinese, as well as making use of internal knowledge. Therefore, we also rewrote the prompts in English while keeping the FFA reports in Chinese to conduct code-switching prompt examination. Therefore, we mainly investigated 4 prompt strategies: Direct-Chinese, Step-Chinese, Direct-English, and Step-English. The detailed prompts can be found in Multimedia Appendix 1.

To avoid the randomness of ChatGPT's response, we set the inference temperature to 0 so as to choose the greedy decoding strategy via the API, making the response more focused and deterministic. Furthermore, we evaluated the robustness of ChatGPT to different languages by calculating the average performance of ChatGPT using different prompting methods. All tests were conducted on the same version of GPT3.5-Turbo-0301 using the official API of OpenAI. Figure 1 shows the overall workflow.

XSL•FO
RenderX

Figure 1. The overall workflow.

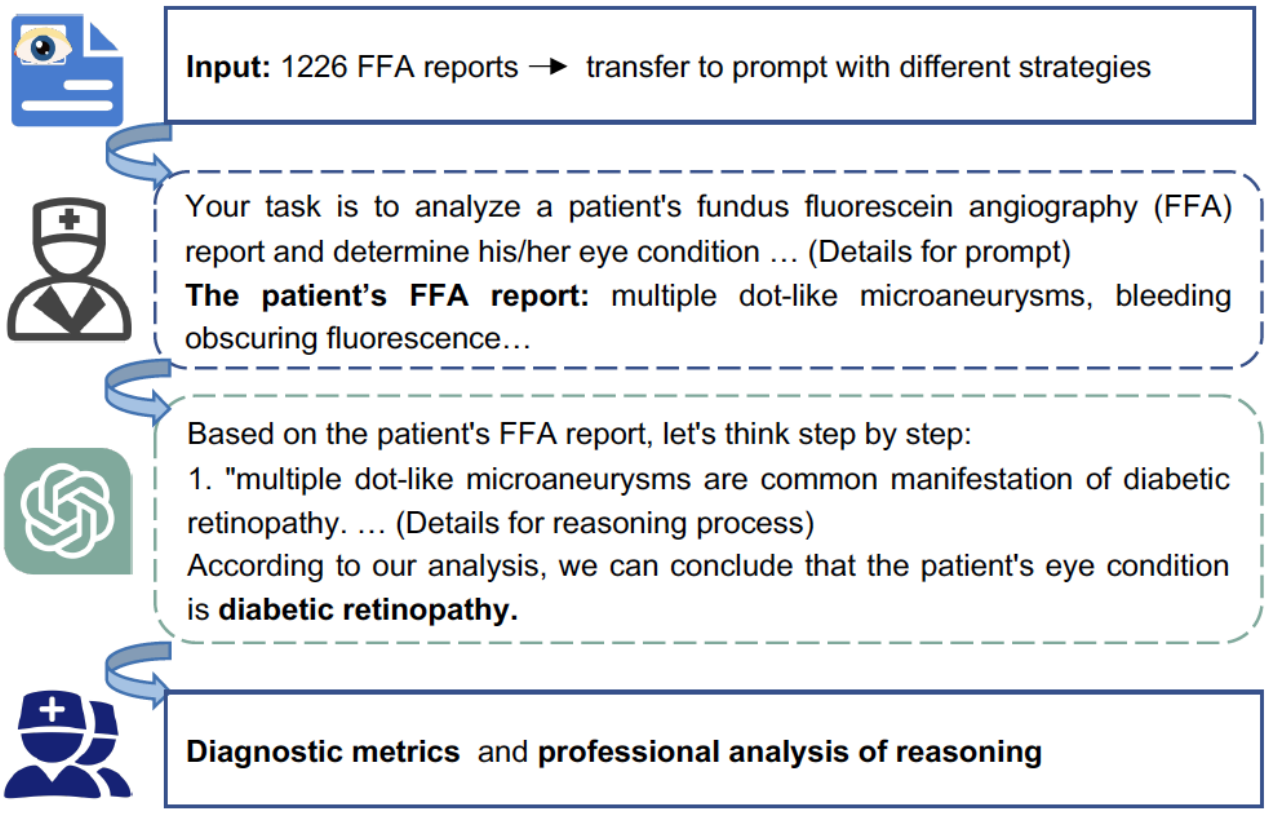

Measurements and Definitions

We designed a systematic evaluation scheme to assess the performance of ChatGPT. In addition to diagnostic performance, we also incorporated a combination of inference ability, omission of information, hallucinations, misinformation, and inconsistency assessments to evaluate the ChatGPT's reasoning process, as follows:

1. Diagnostic performance: precision, recall, and F_1-score.
2. Inference ability [24-26]: the total reasoning steps, the number of reasoning errors, and the incompleteness of the inference process.
3. Omission of information [27]: whether crucial information from the original report was overlooked.
4. Hallucinations [25,28]: whether ChatGPT generated medical findings that were not present in the original report.
5. Misinformation [29,30]: whether the output of ChatGPT quoted incorrect prior knowledge.
6. Inconsistency [30]: whether the reasoning result was inconsistent with the reasoning process.

For diagnostic evaluation, precision, recall, and F_1-score were calculated based on ChatGPT's responses and gold clinical diagnosis. Additionally, to evaluate the diagnostic performance of ChatGPT, 2 ophthalmologists and 2 ophthalmology interns with 2 years of clinical experience from SAHZU were invited to diagnose 100 FFA reports, which were randomly selected according to the proportion of diseases. In terms of the evaluation on ChatGPT's inference ability, the last 5 measurements were evaluated on the responses to the Step-Chinese and Step-English prompts by 2 ophthalmologists from SAHZU. We randomly selected 509 FFA reports (no more than 100 for each disease) and the corresponding ChatGPT outputs for evaluation. Before the formal evaluation, the ophthalmologists were requested to conduct an annotation with training and achieved a final interannotator agreement up to 94%. The statistical analysis between the Chinese and English prompts was performed with the use of Chi-square tests for categorical measurements and Student t tests or Wilcoxon rank-sum tests for continuous measurements, as appropriate. A 2-sided $P<.05$ was considered statistically significant.

Results

Characteristics of ChatGPT's Responses

The characteristics of the FFA reports and responses by ChatGPT are listed in Table 1. Direct-Chinese and Direct-English prompts received responses that directly provided the final options, and their mean (SD) lengths were 19.2 (4.4) tokens and 5.7 (1.7) tokens, respectively, while Step-Chinese and Step-English prompts received responses that provided the detailed reasoning process, and their mean (SD) lengths were 118.4 (71.8) tokens and 100.5 (36.9) tokens, respectively. Examples of different prompts and their responses are presented in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Characteristics of the FFA reports and ChatGPT's responses (N=1226).

Category	Count, n	Report length (tokens), mean (SD)	Response length (tokens), mean (SD)			
			Direct-Chinese	Direct-English	Step-Chinese	Step-English
Normal	117	10.5 (2.4)	14.2 (2.0)	5.4 (1.6)	86.6 (52.7)	64.6 (23.7)
DR[a]	717	46.4 (12.1)	19.5 (5.0)	5.7 (1.6)	124.0 (81.2)	100.9 (32.1)
wetAMD[b]	183	31.1 (11.2)	20.5 (1.5)	6.1 (1.6)	108.9 (46.4)	114.4 (44.7)
CSC[c]	73	29.9 (6.7)	19.3 (2.7)	6.3 (1.9)	146.7 (78.7)	127.4 (41.5)
BRVO[d]	63	44.7 (11.1)	19.8 (2.0)	5.2 (1.8)	106.5 (23.9)	87.2 (24.1)
CRVO[e]	38	50.6 (10.5)	20.7 (3.2)	4.8 (1.8)	134.5 (52.1)	91.4 (22.6)
VKH[f]	35	34.7 (13.5)	19.9 (2.3)	5.3 (1.4)	105.4 (43.9)	116.5 (41.4)
Overall	1226	39.4 (15.9)	19.2 (4.4)	5.7 (1.7)	118.4 (71.8)	100.5 (36.9)

[a]DR: diabetic retinopathy.
[b]wetAMD: wet age-related macular degeneration.
[c]CSC: central serous chorioretinopathy.
[d]BRVO: branch retinal vein occlusion.
[e]CRVO: central retinal vein occlusion.
[f]VKH: Vogt-Koyanagi-Harada disease.

Diagnostic Performance

The Direct-English prompts achieved an overall precision of 79.61%, recall of 83.12%, and F_1-score of 80.05%, which was 9.58% higher than that achieved by the Direct-Chinese prompts (Table 2). The diagnostic performance varied significantly for each disease category. ChatGPT performed better in the normal and DR categories, with the F_1-scores exceeding 80%, but performed worse in the VKH and CSC categories, achieving F_1-scores of less than 4%. Additionally, the F_1-score in the BRVO category varied greatly, from 54.35% for Direct-Chinese prompts to 74.51% for Direct-English prompts.

Table 2. Diagnostic performance of ChatGPT across various disease categories on the FFA reports.

Category	Direct-Chinese (%)			Direct-English (%)			Step-Chinese (%)			Step-English (%)		
	P[a]	R[b]	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Normal	100	85.47	92.17	100	88.03	93.64	98.39	52.14	68.16	97.37	94.87	96.1
DR[c]	91.55	72.52	80.93	91.05	95.12	93.04	85.07	95.4	89.94	82.13	93.58	87.48
wetAMD[d]	44.72	87.98	59.3	59.92	80.87	68.84	63.58	60.11	61.8	60	34.42	43.75
CSC[e]	4.35	2.74	3.36	33.33	1.37	2.63	34.15	19.18	24.56	50	6.85	12.05
BRVO[f]	41.32	79.37	54.35	63.33	90.47	74.51	83.61	80.95	82.26	67.95	84.13	75.18
CRVO[g]	93.1	71.05	80.6	84.85	73.68	78.87	41.27	68.42	51.49	58.33	73.68	65.12
VKH[h]	0	0	0	0	0	0	0	0	0	0	0	0
Overall	75.03	70.15	70.47	79.61	83.12	80.05	76.24	77.16	75.61	74.56	75.94	73.46

[a]P: precision.
[b]R: recall.
[c]DR: diabetic retinopathy.
[d]wetAMD: wet age-related macular degeneration.
[e]CSC: central serous chorioretinopathy.
[f]BRVO: branch retinal vein occlusion.
[g]CRVO: central retinal vein occlusion.
[h]VKH: Vogt-Koyanagi-Harada disease.

In contrast, the Step-Chinese prompts achieved an overall precision of 76.24%, recall of 77.16%, and F_1-score of 75.61%, which was 2.15% higher than that achieved by ChatGPT for Step-English prompts. Compared with Direct-Chinese prompts,

XSL•FO
RenderX

the F_1-score for Step-Chinese prompts was increased by 5.14% and provided the reasoning process, which is crucial for disease diagnosis. However, the diagnostic performance of Step-Chinese prompts in the normal and CRVO categories was far worse than that of Direct-Chinese prompts. This is mainly because, with Step-Chinese prompts, ChatGPT generated hallucinations for FFA reports in the normal category, which were wrongly diagnosed as CRVO. Figure 2 further demonstrates the confusion matrices of ChatGPT for the 4 prompting strategies.

Figure 2. Confusion matrices of ChatGPT for the 4 prompting strategies. BRVO: branch retinal vein occlusion; CSC: central serous chorioretinopathy; CRVO: central retinal vein occlusion; DR: diabetic retinopathy; Undiag: undiagnosed; VKH: Vogt-Koyanagi-Harada disease; wetAMD: wet age-related macular degeneration.

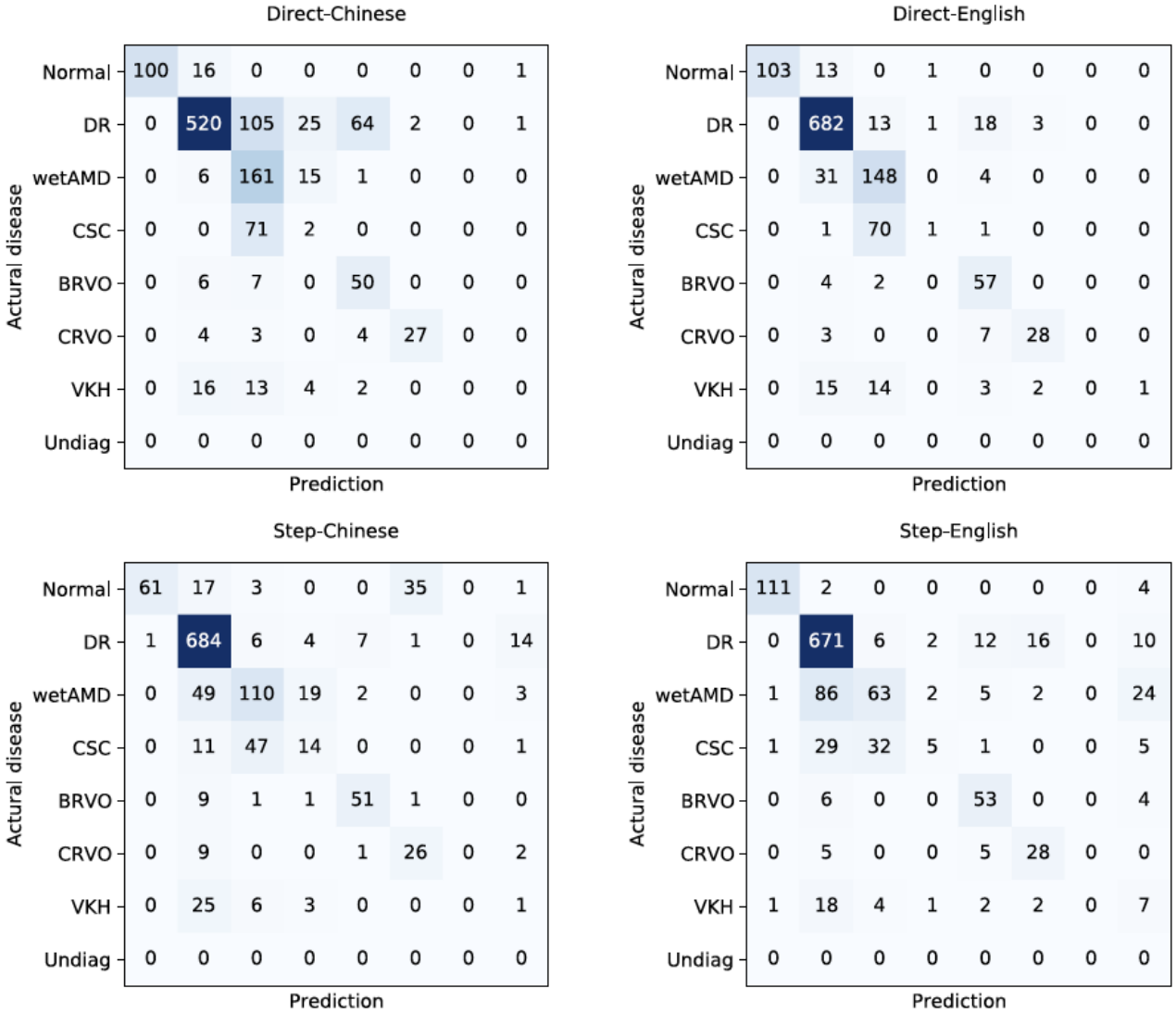

Figure 3 shows the average F_1-score of ophthalmologists, ophthalmology interns, ChatGPT with English prompts (Direct-English and Step-English), and ChatGPT with Chinese prompts (Direct-Chinese and Step-Chinese). Although ChatGPT performed better than experts for some disease types (eg, Direct-English and Step-English prompts for the normal and CRVO categories and all prompts for the BRVO category), the overall performance of ophthalmologists was the best (89.35%), followed by ophthalmology interns (82.69%), ChatGPT with Direct-English and Step-English prompts (76.76%), and ChatGPT with Direct-Chinese and Step-Chinese prompts (73.04%).

XSL•FO
RenderX

Figure 3. Diagnostic performance of humans and ChatGPT. BRVO: branch retinal vein occlusion; CSC: central serous chorioretinopathy; CRVO: central retinal vein occlusion; DR: diabetic retinopathy; VKH: Vogt-Koyanagi-Harada disease; wetAMD: wet age-related macular degeneration.

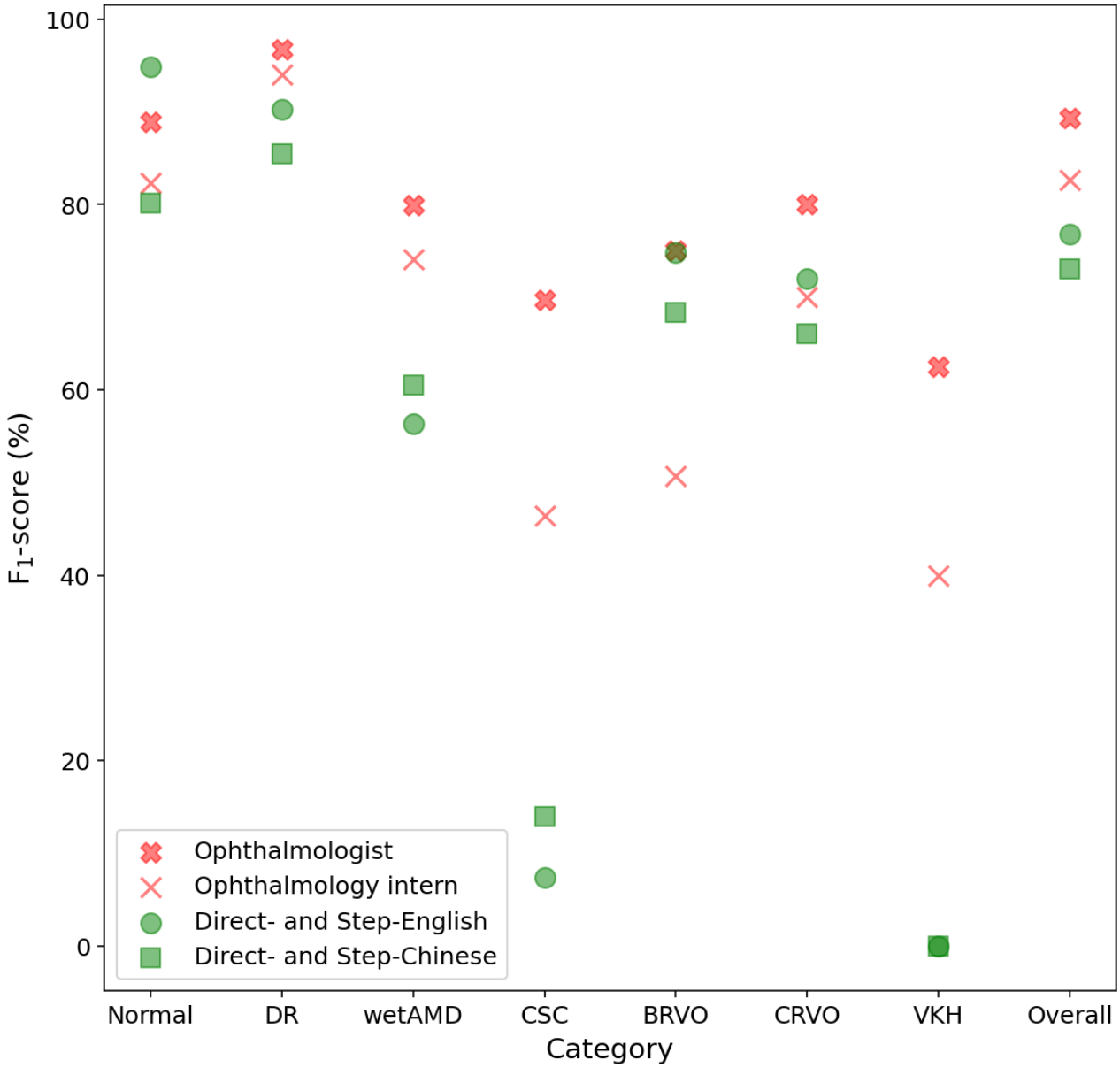

Inference Ability

Table 3 presents the outcomes of ChatGPT's inference ability, as evaluated by 2 ophthalmologists. The results of continuous measurements were presented descriptively as mean (SD) values. Based on the 509 FFA reports, Step-English prompts exhibited a tendency to require more reasoning steps for each report (P<.001, Wilcoxon rank-sum test). Although the average number of inference errors was similar (0.4 per report) between Step-Chinese and Step-English prompts (P=.88, Wilcoxon rank-sum test), Step-English prompts brought less incompleteness (44.31%), hallucinations (0.59%), and misinformation (1.96%) (all P<.001, Chi-square test). Instead, Step-Chinese prompts, which involved fewer reasoning steps, were more prone to having key information from the original report overlooked during the reasoning process, although this difference lacked statistical significance (P=.68, Chi-square test). In addition, a few generated diagnoses were marked as inconsistent with the reasoning process, with no statistical difference observed between Step-Chinese and Step-English prompts. Overall, compared with Step-Chinese prompts, ChatGPT demonstrated stronger inference abilities for Step-English prompts, particularly with regard to the average number of reasoning steps, incompleteness, hallucinations, and misinformation.

XSL•FO
RenderX

Table 3. Inference ability of ChatGPT in the diagnosis of retinal vascular diseases.

Measurement	Step-Chinese	Step-English	*P* value[a]
Reasoning steps per report, mean (SD)	1.4 (0.8)	2.6 (1.5)	<.001
Reasoning errors per report, mean (SD)	0.4 (0.5)	0.4 (0.6)	0.88
Incompleteness (%)	63.53	44.31	<.001
Omission of information (%)	0.78	0.39	0.68
Hallucinations (%)	5.88	0.59	<.001
Misinformation (%)	7.84	1.96	<.001
Inconsistency (%)	0.59	0.39	>.99

[a]Chi-square tests were used for categorical measurements and Wilcoxon rank-sum tests for continuous measurements.

Robustness

Using different prompt strategies introduces some variability in ChatGPT's responses to a given FFA report. Hence, we evaluated the robustness of ChatGPT with different language prompts through calculating the average diagnostic performance for 4 prompting methods: Direct, Step, Step (more detail), and Step (2-step) (Multimedia Appendix 1). The precision, recall, and F_1-score, presented descriptively as mean (SD) values, were compared between Chinese and English prompts using Student *t* tests. As shown in Table 4, the results indicated significant differences in the recall (P=.03) and F_1-score (P=.04) between Chinese and English prompts, while no significant difference was observed in the precision (P=.18). The mean (SD) F_1-scores of ChatGPT with Chinese and English prompts were 70.02% (3.54%) and 76.47% (2.61%), respectively. In short, the diagnostic performance of ChatGPT with English prompts was better and more robust.

Table 4. The robustness of ChatGPT with various prompts in Chinese and English.

Diagnostic performance (%), mean (SD)	Chinese prompt	English prompt	*P* value
Precision	74.38 (1.49)	76.64 (2.10)	.18
Recall	68.78 (3.03)	78.71 (4.46)	.03
F_1-score	70.02 (3.54)	76.47 (2.61)	.04

Discussion

Principal Findings

To the best of our knowledge, this is the first study to evaluate ChatGPT's performance on non-English clinical text for retinal vascular disease diagnosis. We have developed a systematic evaluation scheme that encompasses objective diagnostic performance, professional inference abilities, and comparisons with the diagnostic ability of experts. Through large-scale experiments and analysis, we found the potential of LLMs as medical assistants to provide diagnosis, and identified challenges faced by ChatGPT in the field of health care, especially regarding language disparity.

Our results demonstrated that ChatGPT can preliminarily diagnose retinal vascular diseases based on Chinese FFA reports and achieved a high F_1-score of 80.05% at best. However, the diagnostic performance of ChatGPT varied significantly among different diseases and prompting languages. The performance for common DR was significantly better than that for the more uncommon VKH, which is relatively low in incidence and more difficult to diagnose. Another noteworthy phenomenon is the language disparity of ChatGPT. Given that the development and validation of ChatGPT predominantly relied on English-centric data sets [31] and that non-English medical corpora are even more scarce, compared to with English prompts, ChatGPT exhibited a significant decline in diagnostic performance with Chinese prompts, with a 6.45% decrease in F_1-score. This language disparity poses challenges for the effective application of ChatGPT in non-English clinical settings.

Meanwhile, the diagnosis accompanied by reasoning steps did not necessarily lead to performance improvement; F_1-scores decreased by 6.59% for English prompts but increased by 5.14% for Chinese prompts. This disparity may be attributed to ChatGPT's training being mainly on English corpora, with Direct-English prompts enabling a straightforward mapping from input to diagnosis. In contrast, Step-English prompts tended to bring more mistakes than benefits through multistep internal reasoning. However, for Chinese prompts, the scarcity of Chinese training data results in limited knowledge for disease diagnosis. Step-Chinese prompts, with the requirement of a reasoning process, can effectively compensate for incomplete and incorrect reasoning caused by limited knowledge, although they may introduce some noise. The performance gap between different diseases and prompting strategies demonstrates the potential unfairness brought by the overrepresentation of the major diseases, languages, and countries. This limitation hinders the global applicability of ChatGPT, particularly in non–English-speaking countries.

From the perspective of clinical practice, ChatGPT's diagnostic performance still did not reach the level of ophthalmologists or even ophthalmology interns. It is worth noting that ChatGPT

XSL•FO
RenderX

may be conservative in disease diagnosis. Despite the instruction restriction (must identify one), certain responses involved multiple conditions or indicated an inability to conclude based on existing information. Notably, although ChatGPT can derive a reasoning process to improve clinical interpretability, ophthalmologists identified some harmful mistakes, such as generating medical findings not mentioned in the original reports and quoting incorrect prior knowledge. More in-depth investigation and careful regulation are required before applying ChatGPT in the health care domain. Also, it is imperative to incorporate more extensive and higher-quality clinical data sets and knowledge into ChatGPT [32-34].

Comparison to Prior Work

Prior work in using artificial intelligence (AI) for the automated diagnosis of retinal vascular disease has yielded promising outcomes [35]. However, since some hospitals struggle to produce qualified FFA reports [36] and require ophthalmologists with extensive clinical experience or retinal specialists, the majority of these studies have predominantly focused on analyzing FFA images. Ryu et al [37] introduced an end-to-end deep convolutional neural network–based method specifically designed for the automatic detection of DR and the assessment of lesion status. Similarly, Ding et al [38] proposed a pipeline for detecting retinal vessels in FFA images using deep neural networks. Moreover, Li et al [39] presented a weakly supervised learning-based method for detecting fluorescein leakage, eliminating the need for manual annotation of leakage areas. In contrast to research predominantly centered on lesion detection or specific disease diagnoses, Zhao et al [40] developed an AI system capable of automating image phase identification, diagnosing 4 different types of retinal diseases, and segmenting ischemic areas using FFA images. In our study, we used ChatGPT with 4 different prompting strategies based on FFA reports to diagnose a series of retinal diseases. Notably, when using an English prompt for direct diagnosis, ChatGPT exhibited impressive performance in the classification of retinal vascular diseases, requiring no additional training.

Beyond diagnostic accuracy, researchers have dedicated efforts to enhance the interpretability of disease diagnoses [41,42]. The widely used method for this purpose is heatmap visualization [38-40], used to accentuate characteristic regions crucial for disease diagnosis. This method may not capture the nuanced interplay of features critical for accurate diagnosis, leading to a potential loss of information and subtlety in the interpretative process. In this study, ChatGPT showed promise in enhancing the interpretability of disease diagnoses by explaining the process of diagnostic reasoning step by step. Its capacity to generate human-readable responses also allows for a more intuitive understanding of the AI diagnostic process.

ChatGPT has been used for various applications in clinical care and research. While numerous studies have demonstrated promising outcomes in complex medical tasks, including the United States Medical Licensing Exam (USMLE) [17,43], simplifying imaging reports for patients [27] and aiding decision-making [44,45], it is crucial to note that ChatGPT exhibits certain limitations. In the execution of the aforementioned tasks, ChatGPT occasionally produces errors, such as hallucinations or incomplete information [46]. However, the preceding studies were limited to the application and evaluation of ChatGPT solely within English medical contexts, neglecting an exploration of its effectiveness in non-English clinical scenarios. This study fills this gap by leveraging Chinese FFA reports to assess ChatGPT's diagnostic performance and inference abilities for retinal vascular diseases in a non-English clinical environment and exploring the appropriate prompt languages and strategies.

Limitations

Our study has several limitations. First, we did not fully utilize all the information available in clinical scenarios to conduct a diagnosis, such as more detailed FFA images, which may have reduced the diagnostic accuracy due to incomplete information. Since ChatGPT cannot analyze images, we will further evaluate the capabilities of multimodal models in subsequent research. Second, this study was not conducted in clinical practice. A prospective clinical trial can better examine an LLM's clinical benefit; we leave this to our future work.

Conclusions

This study conducted extensive experiments to evaluate the diagnostic capabilities of ChatGPT in retinal vascular diseases, including objective diagnostic performance and professional reasoning analysis evaluated by ophthalmologists. ChatGPT with English prompts for direct diagnosis performed best, achieving results close to the diagnostic performance of ophthalmology interns with 2 years of clinical experience. On the contrary, due to limited Chinese training data and knowledge, ChatGPT with Chinese prompts led to incomplete reasoning and poor diagnostic performance, which demonstrates that there is a significant language disparity in the application of ChatGPT in clinical environments. Additionally, although ChatGPT can derive a reasoning process with a low error rate, mistakes such as misinformation and hallucinations still exist, which will mislead the diagnose of retinal vascular diseases. This study generally reveals the potential of LLMs to serve as a helpful medical assistant to provide diagnosis in non-English clinical environments, but also demonstrates the potential limitations and the need to continually explore more robust LLMs in ophthalmology practice.

Acknowledgments

KJ was supported by Natural Science Foundation of China (grant 82201195). YW was supported by Medical Scientific Research Foundation of Zhejiang Province, China (grant 2022502730).

XSL•FO
RenderX

Data Availability

Data will be made available for research purposes upon request. Data requests are to be directed to KJ.

Authors' Contributions

XL and JW conducted the experiments and statistical analysis, drafted the work, and contributed equally to this work as co–first authors. J Yang is the senior author; he designed the study and drafted the work with XL and JW. AS and WS participated in the data extraction. AS, WS, PY, and YW participated in the expert evaluation. J Ye and KJ are corresponding authors; they provided clinical guidance and administrative, technical, and material support. All authors revised the manuscript and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the input and output of ChatGPT with various prompts.
[DOCX File , 26 KB - jmir_v26i1e51926_app1.docx]

References

1. GBD 2019 BlindnessVision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e130-e143 [FREE Full text] [doi: 10.1016/S2214-109X(20)30425-3] [Medline: 33275950]
2. Gahlaut N, Suarez S, Uddin MI, Gordon AY, Evans SM, Jayagopal A. Nanoengineering of therapeutics for retinal vascular disease. Eur J Pharm Biopharm 2015 Sep;95(Pt B):323-330 [FREE Full text] [doi: 10.1016/j.ejpb.2015.05.001] [Medline: 26022642]
3. Marmoy OR, Henderson RH, Ooi K. Recommended protocol for performing oral fundus fluorescein angiography (FFA) in children. Eye (Lond) 2022 Jan;36(1):234-236 [FREE Full text] [doi: 10.1038/s41433-020-01328-6] [Medline: 33323986]
4. Gao Z, Pan X, Shao J, Jiang X, Su Z, Jin K, et al. Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. Br J Ophthalmol 2023 Nov 22;107(12):1852-1858. [doi: 10.1136/bjo-2022-321472] [Medline: 36171054]
5. Pan X, Jin K, Cao J, Liu Z, Wu J, You K, et al. Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning. Graefes Arch Clin Exp Ophthalmol 2020 Apr;258(4):779-785. [doi: 10.1007/s00417-019-04575-w] [Medline: 31932886]
6. Jin K, Pan X, You K, Wu J, Liu Z, Cao J, et al. Automatic detection of non-perfusion areas in diabetic macular edema from fundus fluorescein angiography for decision making using deep learning. Sci Rep 2020 Sep 15;10(1):15138 [FREE Full text] [doi: 10.1038/s41598-020-71622-6] [Medline: 32934283]
7. Chen M, Jin K, You K, Xu Y, Wang Y, Yip C, et al. Automatic detection of leakage point in central serous chorioretinopathy of fundus fluorescein angiography based on time sequence deep learning. Graefes Arch Clin Exp Ophthalmol 2021 Aug;259(8):2401-2411. [doi: 10.1007/s00417-021-05151-x] [Medline: 33846835]
8. Holomcik D, Seeböck P, Gerendas BS, Mylonas G, Najeeb BH, Schmidt-Erfurth U, et al. Segmentation of macular neovascularization and leakage in fluorescein angiography images in neovascular age-related macular degeneration using deep learning. Eye (Lond) 2023 May;37(7):1439-1444 [FREE Full text] [doi: 10.1038/s41433-022-02156-6] [Medline: 35778604]
9. Huang J, Huck YC, Liu F, Tian M, Liu Y, Wu T. DeepOpht: medical report generation for retinal images via deep modelsvisual explanation. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). 2021 Presented at: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); January 3-8, 2021; Waikoloa, HI. [doi: 10.1109/wacv48630.2021.00249]
10. Li M, Cai W, Verspoor K, Pan S, Liang X, Chang X. Cross-modal clinical graph Transformer for ophthalmic report generation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer VisionPattern Recognition (CVPR). USA: IEEE; 2022 Presented at: 2022 IEEE/CVF Conference on Computer VisionPattern Recognition (CVPR); June 18-24, 2022; New Orleans, LA. [doi: 10.1109/cvpr52688.2022.02000]
11. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-08-17]
12. King MR. The future of AI in medicine: a perspective from a chatbot. Ann Biomed Eng 2023 Feb;51(2):291-295. [doi: 10.1007/s10439-022-03121-w] [Medline: 36572824]
13. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. arXiv. 2023 Apr 04. URL: https://arxiv.org/abs/2304.01852 [accessed 2023-08-17]

XSL•FO
RenderX

14. Janssen BV, Kazemier G, Besselink MG. The use of ChatGPT and other large language models in surgical science. BJS Open 2023 Mar 07;7(2):zrad032 [FREE Full text] [doi: 10.1093/bjsopen/zrad032] [Medline: 36960954]
15. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
16. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023 Jul;619(7969):357-362 [FREE Full text] [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
17. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
18. Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? J Am Med Inform Assoc 2023 Aug 18;30(9):1558-1560. [doi: 10.1093/jamia/ocad104] [Medline: 37335851]
19. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Jun 01;141(6):589-597. [doi: 10.1001/jamaophthalmol.2023.1144] [Medline: 37103928]
20. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
21. Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, Bui T, et al. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. arXiv. 2023 Apr 12. URL: https://arxiv.org/abs/2304.05613 [accessed 2023-08-17]
22. Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models. arXiv. 2023 Apr 07. URL: https://arxiv.org/abs/2304.03738 [accessed 2023-08-17]
23. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: 36th Conference on Neural Information Processing Systems (NeurIPS 2022); November 28-December 9, 2022; New Orleans, Louisiana.
24. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
25. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv. 2023 Feb 08. URL: http://arxiv.org/abs/2302.04023 [accessed 2023-08-17]
26. Liu H, Ning R, Teng Z, Liu J, Zhou Q, Zhang Y. Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv. 2023 Apr 7. URL: http://arxiv.org/abs/2304.03439 [accessed 2023-08-17]
27. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2023 Oct 05;1:1-9. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
28. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. arXiv. 2022 Feb 8. URL: https://arxiv.org/abs/2202.03629 [accessed 2023-08-17]
29. Zuccon G, Koopman B. Dr ChatGPT, tell me what I want to hear: how prompt knowledge impacts health answer correctness. arXiv. 2023 Feb 3. URL: http://arxiv.org/abs/2302.13793 [accessed 2023-08-17]
30. Borji A. A categorical archive of ChatGPT failures. arXiv. 2023 Feb 6. URL: http://arxiv.org/abs/2302.03494 [accessed 2023-08-17]
31. Blevins T, Zettlemoyer L. Language contamination helps explain the cross-lingual capabilities of English pretrained models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022 Presented at: 2022 Conference on Empirical Methods in Natural Language Processing; December 7-11, 2022; Abu Dhabi, UAE. [doi: 10.18653/v1/2022.emnlp-main.233]
32. Wu J, Wu X, Qiu Z, Li M, Zheng Y, Yang J. Qualifying chinese medical licensing examination with knowledge enhanced generative pre-training model. arXiv. 2023 May 17. URL: http://arxiv.org/abs/2305.10163 [accessed 2023-08-17]
33. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
34. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023 Nov 16;6(1):210 [FREE Full text] [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
35. Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: current status and future perspectives. Adv Ophthalmol Pract Res 2022;2(3):100078 [FREE Full text] [doi: 10.1016/j.aopr.2022.100078] [Medline: 37846285]

XSL•FO
RenderX

36. Li Z, Jiang J, Chen K, Chen Q, Zheng Q, Liu X, et al. Preventing corneal blindness caused by keratitis using artificial intelligence. Nat Commun 2021 Jun 18;12(1):3738 [FREE Full text] [doi: 10.1038/s41467-021-24116-6] [Medline: 34145294]
37. Ryu G, Lee K, Park D, Park SH, Sagong M. A deep learning model for identifying diabetic retinopathy using optical coherence tomography angiography. Sci Rep 2021 Nov 26;11(1):23024 [FREE Full text] [doi: 10.1038/s41598-021-02479-6] [Medline: 34837030]
38. Ding L, Bawany MH, Kuriyan AE, Ramchandran RS, Wykoff CC, Sharma G. A novel deep learning pipeline for retinal vessel detection in fluorescein angiography. IEEE Trans Image Process 2020 May 08:6561-6173 [FREE Full text] [doi: 10.1109/TIP.2020.2991530] [Medline: 32396087]
39. Li W, Fang W, Wang J, He Y, Deng G, Ye H, et al. A weakly supervised deep learning approach for leakage detection in fluorescein angiography images. Transl Vis Sci Technol 2022 Mar 02;11(3):9 [FREE Full text] [doi: 10.1167/tvst.11.3.9] [Medline: 35262648]
40. Zhao X, Lin Z, Yu S, Xiao J, Xie L, Xu Y, et al. An artificial intelligence system for the whole process from diagnosis to treatment suggestion of ischemic retinal diseases. Cell Rep Med 2023 Oct 17;4(10):101197 [FREE Full text] [doi: 10.1016/j.xcrm.2023.101197] [Medline: 37734379]
41. Chetoui M, Akhloufi MA. Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets. J Med Imaging (Bellingham) 2020 Jul;7(4):044503 [FREE Full text] [doi: 10.1117/1.JMI.7.4.044503] [Medline: 32904519]
42. Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun 2021 May 28;12(1):3242 [FREE Full text] [doi: 10.1038/s41467-021-23458-5] [Medline: 34050158]
43. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 01;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
44. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023 May 30;9(1):44 [FREE Full text] [doi: 10.1038/s41523-023-00557-8] [Medline: 37253791]
45. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023 Nov 17;13(1):20159 [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
46. Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv. 2023 Sep 3. URL: http://arxiv.org/abs/2309.01219 [accessed 2023-08-17]

Abbreviations

AI: artificial intelligence
API: application programing interface
BRVO: branch retinal vein occlusion
CRVO: central retinal vein occlusion
DR: diabetic retinopathy
FFA: fundus fluorescein angiography
LLMs: large language models
SAHZU: The Second Affiliated Hospital of Zhejiang University
VKH: Vogt-Koyanagi-Harada disease

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 17.08.23; peer-reviewed by 家 余, X He, A Gilson; comments to author 20.09.23; revised version received 07.10.23; accepted 30.11.23; published 22.01.24.

Please cite as:
Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, Ye J, Jin K, Yang J
Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study
J Med Internet Res 2024;26:e51926
URL: https://www.jmir.org/2024/1/e51926
doi:10.2196/51926
PMID:38252483

©Xiaocong Liu, Jiageng Wu, An Shao, Wenyue Shen, Panpan Ye, Yao Wang, Juan Ye, Kai Jin, Jie Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study

You-Qian Lee[1,2*], ME; Ching-Tai Chen[3,4*], PhD; Chien-Chang Chen[5], PhD; Chung-Hong Lee[6], PhD; Peitsz Chen[7], BS; Chi-Shin Wu[8], MD; Hong-Jie Dai[2,9,10,11], PhD

[1]Dialogue System Technical Department, Intelligent Robot, Asustek Computer Inc, Taipei, Taiwan

[2]Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[3]Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

[4]Center for Precision Health Research, Asia University, Taichung, Taiwan

[5]Electromagnetic Sensing Control and AI Computing System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[6]Knowledge Discovery and Data Mining Lab, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[7]Department of Chemical Engineering, Feng Chia University, Taichung, Taiwan

[8]National Center for Geriatrics and Welfare Research, National Health Research Institutes, Zhunan, Taiwan

[9]National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan

[10]School of Post-Baccalaureate Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[11]Center for Big Data Research, Kaohsiung Medical University, Kaohsiung, Taiwan

[*]these authors contributed equally

Corresponding Author:
Hong-Jie Dai, PhD
Intelligent System Laboratory, Department of Electrical Engineering
College of Electrical Engineering and Computer Science
National Kaohsiung University of Science and Technology
No. 415, Jiangong Road, Sanmin District
Kaohsiung, 80778
Taiwan
Phone: 886 73814526 ext 15510
Email: hjdai@nkust.edu.tw

Abstract

Background: The widespread use of electronic health records in the clinical and biomedical fields makes the removal of protected health information (PHI) essential to maintain privacy. However, a significant portion of information is recorded in unstructured textual forms, posing a challenge for deidentification. In multilingual countries, medical records could be written in a mixture of more than one language, referred to as code mixing. Most current clinical natural language processing techniques are designed for monolingual text, and there is a need to address the deidentification of code-mixed text.

Objective: The aim of this study was to investigate the effectiveness and underlying mechanism of fine-tuned pretrained language models (PLMs) in identifying PHI in the code-mixed context. Additionally, we aimed to evaluate the potential of prompting large language models (LLMs) for recognizing PHI in a zero-shot manner.

Methods: We compiled the first clinical code-mixed deidentification data set consisting of text written in Chinese and English. We explored the effectiveness of fine-tuned PLMs for recognizing PHI in code-mixed content, with a focus on whether PLMs exploit naming regularity and mention coverage to achieve superior performance, by probing the developed models' outputs to examine their decision-making process. Furthermore, we investigated the potential of prompt-based in-context learning of LLMs for recognizing PHI in code-mixed text.

Results: The developed methods were evaluated on a code-mixed deidentification corpus of 1700 discharge summaries. We observed that different PHI types had preferences in their occurrences within the different types of language-mixed sentences,

XSL•FO
RenderX

and PLMs could effectively recognize PHI by exploiting the learned name regularity. However, the models may exhibit suboptimal results when regularity is weak or mentions contain unknown words that the representations cannot generate well. We also found that the availability of code-mixed training instances is essential for the model's performance. Furthermore, the LLM-based deidentification method was a feasible and appealing approach that can be controlled and enhanced through natural language prompts.

Conclusions: The study contributes to understanding the underlying mechanism of PLMs in addressing the deidentification process in the code-mixed context and highlights the significance of incorporating code-mixed training instances into the model training phase. To support the advancement of research, we created a manipulated subset of the resynthesized data set available for research purposes. Based on the compiled data set, we found that the LLM-based deidentification method is a feasible approach, but carefully crafted prompts are essential to avoid unwanted output. However, the use of such methods in the hospital setting requires careful consideration of data security and privacy concerns. Further research could explore the augmentation of PLMs and LLMs with external knowledge to improve their strength in recognizing rare PHI.

(J Med Internet Res 2024;26:e48443) doi:10.2196/48443

KEYWORDS

code mixing; electronic health record; deidentification; pretrained language model; large language model; ChatGPT

Introduction

Background

In the clinical and biomedical fields, electronic health records (EHRs) have become valuable resources in recent years [1]. In order to protect the privacy of patients whose data are used for secondary purposes, regulations or laws have been established requiring the removal of protected health information (PHI) from records before they can be disseminated. In settings, such as research, obtaining explicit consent to access personal data may be impractical or impossible. As a result, the deidentification step becomes a critical data processing step to protect the privacy of individuals. However, the study by Cannon and Lucci [2] indicated that up to 65% of important clinical information is recorded in unstructured texts in medical reports written by medical personnel. Compared with structured data, which can be deidentified by encrypting private patient information fields, the deidentification of unstructured data is more challenging.

Because manual deidentification of large volumes of EHRs is time-consuming and error-prone, automated methods are needed for large-scale deidentification of unstructured clinical data. Compared with Asian countries, such as Japan and China, where the majority of medical records are written monolingually in their native languages, medical records in Taiwan are frequently written in a mixture of Chinese and English. The example sentence "前夫mk1300309 married > mk136從婆家搬出來 > mk139離婚" (got married to my ex-husband on 2141/03/09 > moved out from the ex-husband's family's home in 2147 > divorced in 2150) demonstrates how a physician wrote in English and suddenly switched to Chinese. The sentence also includes a transliteration ("mk") of "民國," which refers to the Republic of China calendar. This phenomenon is referred to as code alternation, which occurs when a bilingual speaker uses more than one language in a single utterance [3]. Code switching refers to code alternation at or above clause level, while code mixing refers to code alteration below clause level. Some examples of code-switched and code-mixed narratives in discharge summaries are shown in Multimedia Appendix 1. Throughout this work, we will use code mixing to refer to the above phenomenon.

In addition to our specific context, it is important to note that code mixing in clinical records is observed in various multilingual health care settings globally. For instance, Alqurashi [4] conducted research on the use of code-mixed language in private hospitals in Saudi Arabia, highlighting the prevalence of this practice in health care communication. Furthermore, Dowlagar and Mamidi [5] demonstrated that the use of code-mixed utterances, where native languages are blended with English, is a common and natural occurrence among doctors and patients. Beyond this, code mixing in clinical notes extends beyond English-dominant countries. For example, Keresztes [6] investigated how Hungarian physicians are influenced by the English language in their writing of cardiology discharge reports, while Karuthan [7] showed that nurses in Malaysia tend to use "Manglish," a mix of Malayalam and English, when writing nursing documents.

This global prevalence of code mixing highlights the importance of addressing narratives written in this manner. Unfortunately, most existing clinical natural language processing (NLP) techniques are primarily designed for monolingual texts, mainly due to limited research resources. While our previous work [8] has highlighted the potential of pretrained language models (PLMs) in mitigating issues related to code mixing during deidentification, the precise mechanism at play remains incompletely understood. To bridge this knowledge gap, there is a pressing need for a code-mixed deidentification data set that can comprehensively evaluate the performance of state-of-the-art NLP models and unravel their underlying mechanisms.

Goal of This Study

Previous studies [9,10] have demonstrated that state-of-the-art neural network models can effectively exploit strong name regularity and high mention coverage to achieve superior performance. Name regularity refers to the consistency and predictability of the structure and pronunciation of a named entity type within a particular language. For example, person names typically adhere to the "FirstName LastName" format.

XSL•FO
RenderX

A high mention coverage denotes that a large proportion of mentions in the test set have been previously observed in the training set.

Our study aims to explore the impact of these properties on the performance of PLMs in recognizing PHI within code-mixed sentences. We hypothesize that the decision-making processes of the models are influenced by these factors, even when dealing with code-mixed text. However, they may underperform when faced with weak regularity, such as unseen mentions or out-of-vocabulary (OOV) words. Additionally, they may solely memorize popular mentions instead of learning generalization knowledge, which could limit their generality and performance. To investigate the decision-making process of these models and gain insights into their ability to recognize code-mixed PHI, we curated a novel code-mixed deidentification data set, trained and evaluated the performance of state-of-the-art PLMs on the unique data set, and employed methods to interpret their results to validate our hypothesis and gain a deeper understanding of how these models recognize code-mixed PHI.

In summary, this study has the following 3 major contributions:

1. Unique code-mixed deidentification data set: We significantly extended our original corpus compiled in our previous work [8] by incorporating an additional 900 discharge summaries. Furthermore, we created a manipulated subset of the resynthesized data set available for research purposes. Importantly, our work resulted in the creation of the very first clinical code-mixed data set that encompasses both traditional Chinese and English. This unique data set [11] represents a valuable resource for researchers, offering a fresh perspective on issues related to code mixing in the context of health care data and contributing to advancing research in code-mixed clinical text processing.
2. Enhancing the understanding of PLMs in code-mixed texts: The curated data set enables an in-depth exploration of the effectiveness and underlying mechanism of fine-tuned PLMs in recognizing code-mixed PHI. Through thorough analyses, we gained insights into the mechanism used by these models for this specific task.
3. Unprecedented exploration of large language models (LLMs) on code-mixed clinical text: The emergence of LLMs has raised a question about their potential application in the deidentification process. This study focused on the feasibility of employing prompt-based, zero-shot, in-context learning with LLMs to recognize PHI. More specifically, we leveraged ChatGPT to recognize PHI within code-mixed text that was previously challenging for the developed PLMs. To the best of our knowledge, this is the first attempt to use prompt-based in-context learning of LLMs for deidentification in code-mixed text.

Related Work

Definition of PHI Types

The Health Insurance Portability and Accountability Act (HIPAA) is a federal law in the United States enacted in 1996 to protect the privacy and security of the PHI of individuals [12]. The Safe Harbor method under the HIPAA includes the removal of 18 specific identifiers from PHI, such as name, date of birth, and social security number. If all 18 identifiers are removed, the information is considered deidentified under the method, and the data can be reused for secondary purposes [13]. Given the sensitive nature of the clinical text used in this study, we expanded the list of PHI types to 20 fine-grained categories within 6 coarse-grained categories. The details are presented in Table 1.

Table 1. The coarse-grained and fine-grained categories of protected health information defined in the annotation guideline used in this study.

Coarse-grained PHI[a]	Fine-grained PHI
Date	Date
Age	Age
Name	Patient, person, doctor
Location	Named location, nationality, region, country, city, hospital, department, room, number, school, generic location, market
Profession	Profession
ID	ID number, medical record

[a]PHI: protected health information.

Publicly Available Deidentification or Code-Mixed Data Sets

A code-mixed named entity recognition competition was held in 2018 [14], in which twitter data with the mixture of English and Spanish, and Arabic and Egyptian were provided. The former consisted of 67,223 tweets, and the latter consisted of 12,334 tweets. To the best of our knowledge, there is only 1 publicly available Chinese-mixed corpus, which is a simple Chinese-English code-mixed data set for the task of emotion detection [15].

On the other hand, the MIMIC (Medical Information Mart for Intensive Care) II/III database [16] is one of the largest deidentified clinical databases on patients admitted to a medical center in the United States, which offers clinical text in English for researchers to access and use under a data access agreement. To facilitate automated deidentification tasks, large annotated corpora of unstructured text with PHI entities are required. However, the development of such corpora is complex and

XSL•FO
RenderX

limited in number. To address this need, the Informatics for Integrating Biology and the Bedside (i2b2) has organized several clinical deidentified NLP competitions [17], such as the deidentified competition in 2014, which released a corpus of 1304 medical records of 296 patients with diabetes [18]. Another is the CEGS N-GRID deidentification corpus [19], containing 1000 psychiatric notes from the United States. Alla et al [20,21] compiled the first open available deidentification corpus from Australia, which contained 2100 pathology reports for the purpose of automatic deidentification. However, deidentification corpora in languages other than English are scarce. A Norwegian synthetic clinical corpus released by Bråthen et al [22] is an example, and it includes 477 sentences. To the best of our knowledge, there is currently no openly available code-mixed deidentification corpus.

Deidentification Methods and Approaches for Tackling Code-Mixing Challenges

Manual deidentification is costly, in terms of both finances and time. For example, deidentification of 50,000 patient visit records in the MIMIC-III data set can cost around US $500,000 and 5000 hours. To overcome these challenges, automated deidentification systems have been developed. Initially, rule-based systems using lexical dictionaries, regular expressions, and simple heuristics were introduced [23]. However, they have limitations in generalization across different data sets. Subsequently, machine learning–based approaches [23-25] were proposed, offering better generalization and prediction accuracy. Nonetheless, these methods still rely on handcrafted features. More recently, deep learning techniques have become mainstream, eliminating the need for manually crafted features [26,27]. Neural networks are advantageous because they can be initialized with PLMs acquired from extensive unlabeled data, resulting in faster optimization and superior performance. BERT (Bidirectional Encoder Representations from Transformers) pretrained on English corpora (EN-BERT) [28] is one such example of a monolingual transformer model pretrained on the BookCorpus [29] and English Wikipedia in a self-supervised fashion, which has achieved exceptional precision in various NLP tasks including the deidentification task [8,30].

On the other hand, the rise of code-mixed data sets has spurred the development of approaches designed to address the unique challenges posed by code-mixed content. For instance, in the work by Winata et al [31], a bidirectional long-short term memory model was used to leverage both character and word information, complemented by pretrained FastText embeddings [32], to improve the identification of OOV words in an English-Spanish mixed data set. Another approach by Trivedi et al [33] incorporated FastText word vectors [34] trained from Spanish Wikipedia pages and pretrained word embeddings sourced from a substantial collection of tweets [35] to tackle the English-Spanish code-mixing problem. They applied the singular value decomposition [36] method to align word embeddings across different languages, facilitating a unified vector space representation. Devlin et al [28] introduced pretrained models beyond monolingual EN-BERT, including BERT pretrained on simplified and traditional Chinese corpora (CH-BERT) and BERT pretrained on Wikipedia corpora from 104 languages (M-BERT). Researchers have been exploring the use of M-BERT to tackle the code-mixing problem owing to its ability to generalize cross-lingually through its multilingual representation [37]. For example, Tang et al [15] fine-tuned M-BERT on their English-simplified Chinese social media data set for multi-label sentiment analysis, and obtained an F-score of 0.69, which was 15% higher than that for the model without M-BERT. While limited research exists on using BERT-based models or other LLMs for code-mixing deidentification in clinical data sets, our previous work [38] suggested the potential benefits of incorporating M-BERT to disambiguate PHI categories in code-mixed sentences. In this study, we focused on comprehending how BERT-family PLMs handle Chinese-English mixed issues that arise in actual clinical text and assessed the feasibility of using state-of-the-art LLMs to recognize PHI from sampled clinical text.

Methods

Data Sources and Corpus Construction

We obtained discharge summaries sampled from the psychiatry section of a medical center, which were used to extract depressive symptoms with text mining approaches [39,40]. From these data, we compiled a code-mixed deidentification corpus of 1700 discharge summaries. Four annotators were enlisted to annotate the entire corpus, and they followed the same annotation procedure as previously described in our work [8]. To ensure the quality of the annotation, the annotation process began with the annotators individually annotating an identical set of 200 randomly sampled records, using the annotation guidelines provided. Subsequently, a meeting was organized to facilitate discussions among the annotators, addressing any issues or concerns that arose during the annotation process. This iterative process continued until the annotators achieved a strong level of agreement [41]. The interannotation agreement, as measured by κ, was found to be 0.85 after the above procedure. Once this agreement threshold was reached, the remaining unlabeled EHRs were evenly distributed among the 4 annotators for labeling. The training set comprised 1500 discharge summaries along with 297,621 sentences. The test set comprised 200 discharge summaries along with 60,632 sentences. Finally, the principle-based resynthesis method proposed in our previous work [38] was used to generate surrogates, and the entire corpus was rechecked by one of the senior annotators (PTC) to ensure a high level of data consistency and correctness.

Ethics Approval

The study has been approved by the ethics committee for medical research of the National Institutes of Health (number: EC1090212-E).

Implementation of Deidentification Methods

After preprocessing the collected data set, we approached the deidentification task by treating it as a named entity recognition problem in which the target entity types are PHI types defined in Table 1. To establish a baseline, we used a dictionary-based method [42,43] and developed 4 BERT-based models for comparison. The dictionary-based approach (DBA) relies on

XSL•FO
RenderX

predefined dictionaries and a lookup method. In our implementation, we compiled a dictionary consisting of tokens for all PHI types collected from the training set. The most frequent tag associated with each token estimated on the training set was set as the token's tag. For the BERT-based models, the recognition task was formulated as a sequential tagging problem by using the BILOU tagging schema. This schema categorizes the tokens as either the beginning (B), within (I), or last (L) of multi-word PHI, as well as identifies non-PHI (O) and single-word PHI (U) tokens. For more in-depth information about the preprocessing steps and the development of the DBA, please refer to Multimedia Appendix 2.

BERT-Based Approach

The BERT-family PLMs, including EN-BERT, CH-BERT, and M-BERT, were used in this study. In our implementation, an additional fully connected layer was added that takes the token embeddings from the top layer of the underling PLM as input. This layer is employed to predict the probabilities of the tags for each token. When the WordPiece tokenizer splits a token into multiple pieces, we take the representation of the first piece to represent the token [44]. Details of the procedure and hyperparameters used for training our BERT-based models are described in Multimedia Appendix 2.

In addition to fine-tuning the above BERT-based models on the original code-mixed corpus, we translated sentences in our English-Chinese corpus into either English or Chinese and then fine-tuned the M-BERT model on the translated deidentification corpus (TM-BERT) by determining the degree of code mixing present in a given sentence using the code-mixing index (CMI) [8,45]. The flowchart of the translation process for compiling the training set is shown in Figure 1.

We determined the degree of code mixing present in a given sentence s using the following formula proposed by Gambäck and Das [45]:

where W_i is the number of tokens of the most frequent language in the sentence s, n is the total number of tokens, and u is the quantity of tokens in s that are independent of any particular language, such as numeric words and punctuation marks.

By definition, the CMI is always smaller than 0.5 for code mixing of 2 languages. A larger CMI (close to 0.5) indicates serious code mixing, while a smaller CMI (close to 0) suggests that 1 language dominates the majority of tokens. The CMI is set as zero if the entire sentence is either monolingual or composed of numeric words or punctuations. In cases where the CMI of a sentence is not zero, the number of tokens belonging to English and Chinese in the sentence is calculated. If the majority of tokens belong to English, the sentence is translated to English using a machine translation model [46]. Conversely, if Chinese tokens dominate, we opt for translation to traditional Chinese [47]. Subsequently, we proceeded with the identical fine-tuning process on the translated corpus. In the prediction phase, we employed identical procedures to transform the input sentences into monolingual sentences based on the dominating language and subsequently used TM-BERT to detect PHI.

Figure 1. The procedure to generate the training and test sets for TM-BERT. CMI: code-mixing index; M-BERT: BERT pretrained on Wikipedia corpora from 104 languages; TM-BERT: M-BERT fine-tuned on the translated deidentification corpus.

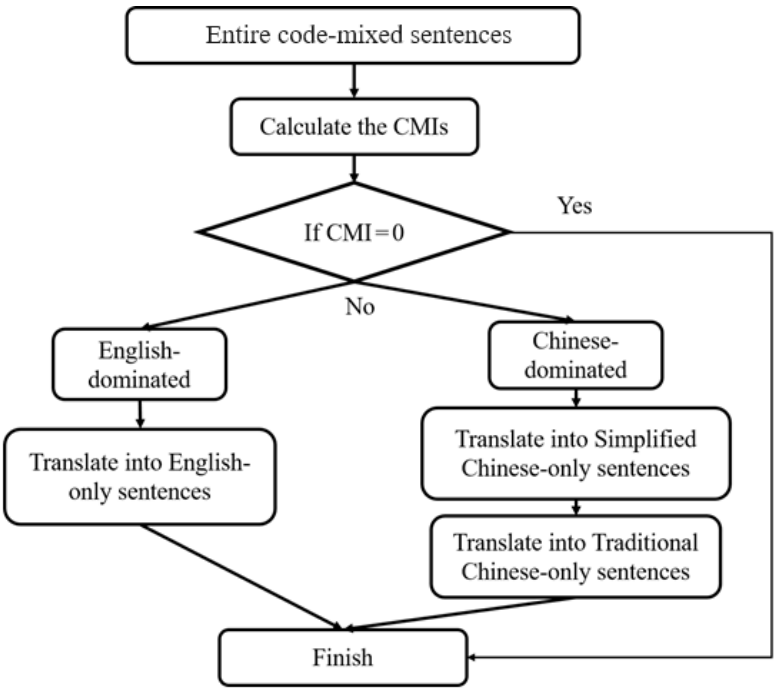

ChatGPT-Based Deidentification Framework

Figure 2 presents our framework to use ChatGPT for the recognition of PHI, which is based on recent research suggesting that LLMs can be adapted to downstream tasks through the use of prompts [48]. A prompt is a set of instructions that can customize an LLM's capabilities and affect its subsequent interactions and outputs. To optimize our prompt, we performed

XSL•FO
RenderX

prompt engineering based on the PHI definitions. The objective was to ensure that the extracted candidate PHI text could be matched with the original input text represented by the "<SENTENCE>" placeholder. Following the prompt design procedure [49], we organized our prompt into 4 main parts:

1. Overall task instruction: This part provides general instructions for the task. We declared it as "Extracting Private Information."
2. Sentence introduction: This part defines the placeholder for the input text ("<SENTENCE>") and the PHI that we focus on.
3. Constraint: As illustrated in Figure 2, the text following the statement "Extract PHI from the given text based on the above definitions" forms the constraint part. The inclusion of this constraint is based on our observation that ChatGPT has the capability to directly extract various PHI types from text and sometimes translate extracted Chinese text into English; therefore, we constrained it to the defined types and requested that it adhere to the regulations in order to avoid translation. This enabled us to match the extracted candidate PHI with the input text and determine their spans for the purpose of performance evaluation.
4. Retrieval message: We provided the text "PHIs:" to instruct the model to generate results to complete the task.

The experiment was conducted from February 20, 2023, to April 30, 2023, using the GPT-3.5 model (version from February 13, 2023) with temperature and top_*P* values set at 0.5. As we used the zero-shot in-context learning setting, we observed that ChatGPT could generate different responses for the same prompt. Thus, we sent the same prompt 3 times to collect all PHI candidates and retained the PHI that appeared at least twice as the final results. The decision to send the prompt 3 times aligns with previous work [50,51], which recommended this practice to ensure the stability of the responses.

Figure 2. The developed framework for the recognition of protected health information mentioned in code-mixed text based on ChatGPT. EHR: electronic health record; PHI: protected health information.

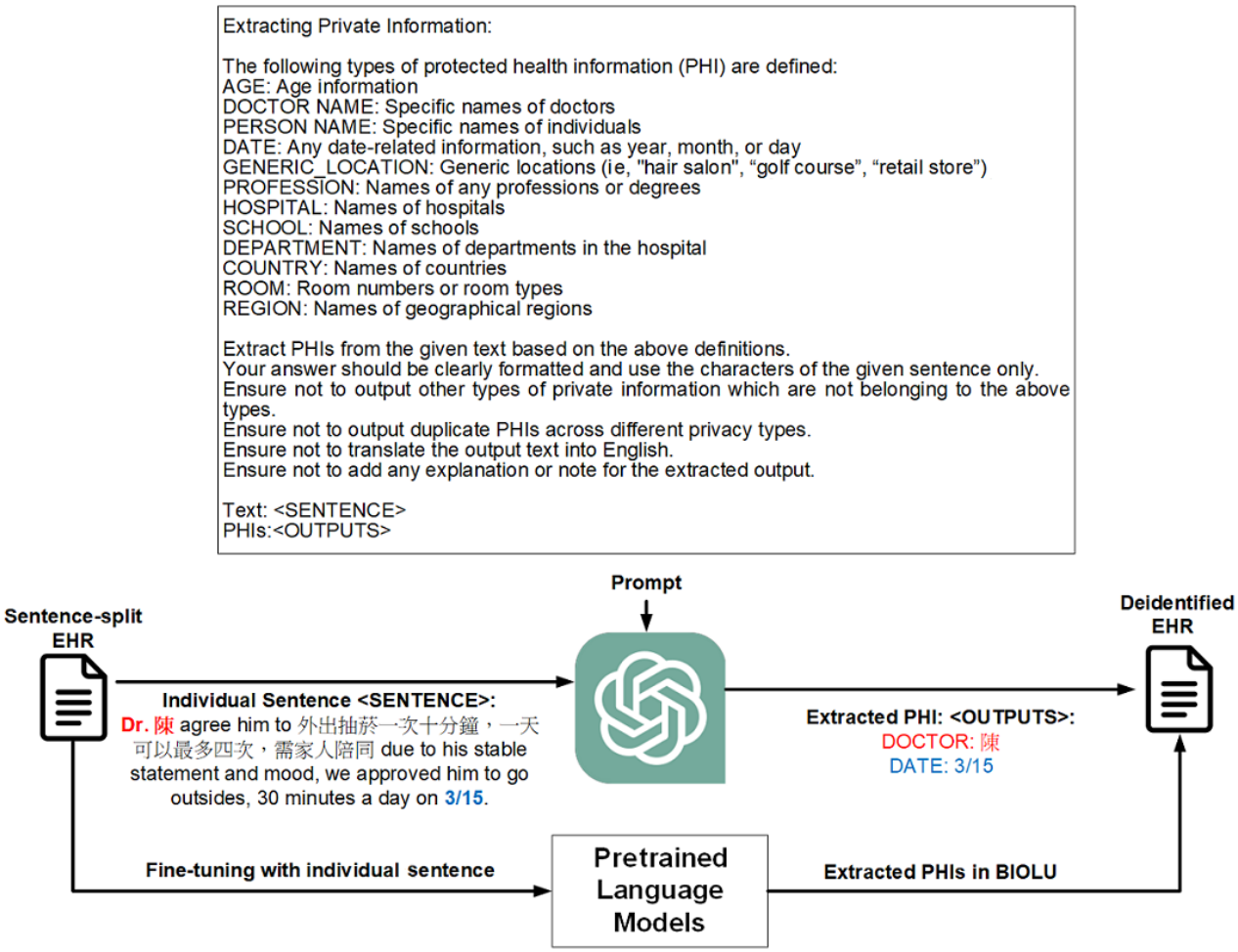

Evaluation Metrics and Analysis Methods Used in This Study

Deidentification Performance Evaluation

Performance of the developed methods was measured by standard metrics used in previous deidentification tasks [18] including recall and precision. The overall performance was assessed using the micro-F_1 measure, a weighted harmonic mean of precision and recall. We further report the macro-F_1 measure as the unweighted mean of the F-scores calculated per PHI type. For the evaluation of the results of the proposed ChatGPT-based deidentification framework, we applied the approximate match criterion [52] to determine whether the recognized boundary matched with the human annotated span.

XSL•FO
RenderX

In addition, to gain more insights into the impact of the code-mixing level on PHI recognition, we further categorized sentences according to their CMI levels and estimated the corresponding sentence error rates, which were computed by dividing the total number of wrongly predicted sentences by the total number of sentences in each CMI range. A sentence was considered incorrectly predicted if any of its tokens were labeled incorrectly.

Code-Mixing Level Evaluation

To measure the level of complexity and mixing in our corpus, we applied the CMI defined in equation (1) to calculate the average CMI for a specific language L, termed AvgCMI_L. In this study, a sentence s was classified as either English-dominated or Chinese-dominated according to the dominating tokens present in the sentence. AvgCMI_L was defined as follows:

where L is either the English or Chinese language and K indicates the set of L-dominated sentences in our corpus. Note that the calculation of AvgCMI_L includes the sentences with a CMI of zero. Finally, the CMI for the corpus (CMIC) was defined as follows:

Methods for Diagnosing and Interpreting the Outputs of the Developed Models

According to our experimental findings, the primary factor affecting PHI recognition performance was the problem of ambiguity. This problem arises when the identified words or phrases can be categorized under more than one PHI type. For example, “八里 (Bali)” is a suburban district in northwestern New Taipei, Taiwan, but it can also refer to “八里療養院 (Bali psychiatric center)” in sentences such as “Informant: Discharge note from 八里.” To interpret the produced representations and analyze the ambiguity issues encountered, we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) [53] method to project the hidden representations of PHI tokens from the last transformer layer of the developed models.

In our subsequent experiments, we noticed that EN-BERT can recognize the target PHI even if the information is described in Chinese characters or is mentioned in sentences that consist of English-Chinese or Chinese characters only. Moreover, we found that all BERT-based models could recognize a target PHI as a candidate, even if its encoding is replaced with “UNK” (unknown word). For instance, in the sentence “... follow-up in Dr. 張家明’s OPD,” where “張家明” is recognized as “Doctor,” replacing the encoding of “張家明” in the sentence with “UNK” did not affect the ability of BERT-based models to recognize it as candidate PHI of “Doctor.”

To comprehend the decision-making mechanism, we probed the results of the developed models by applying the input reduction method [54] and interpreted it using the attention-head view (AHV) to visualize the attention patterns produced by the models’ attention heads in the last transformer layer. The input reduction method removes as many words as possible based on the calculated gradient values, without altering a tag’s prediction. The AHV visualization was generated by the BERTViz tool [55].

Results

PHI Distribution and Code-Mixing Level Estimation

Given the code-mixing characteristics of our corpus, we categorized sentences in our data set based on their code-mixing characteristics into 4 groups: English only, Chinese only, a mixture of Chinese and English, and numeric/symbolic characters. The statistical information for PHI across the 4 groups is presented in Table 2. Overall, it is apparent that the distributions of different PHI types are significantly imbalanced. Without considering numeric/symbolic sentences, the most frequent type of sentences was English only, followed by code-mixed and Chinese only. The “Date” type was most frequently observed in numeric/symbolic sentences, while the “Age” type was most commonly noted in English-only sentences. The “Name” and “Profession” types were most frequently observed in code-mixed sentences, and the “Location” type was evenly distributed across English-only, Chinese-only, and code-mixed sentences. Finally, the “ID” type was most commonly found in Chinese-only sentences.

The $\text{AvgCMI}_{\text{English}}$, $\text{AvgCMI}_{\text{Chinese}}$, and CMIC estimated for the training set, test set, and entire corpus are presented in Multimedia Appendix 3, which highlight the writing convention in Taiwan’s medical records, with the majority of records being written in English. Another interesting writing convention observed was that code mixing was more frequent in Chinese-dominated sentences than in English-dominated sentences. The significant difference between $\text{AvgCMI}_{\text{English}}$ and $\text{AvgCMI}_{\text{Chinese}}$ highlighted that the frequency of Chinese tokens occurring in English-dominated sentences is much less than the frequency of English tokens occurring in Chinese-dominated sentences. The CMIC of our corpus was 22.21, which is significantly higher than the CMIC of English-Bangla (approximately 5.15) and Dutch-Turkish (approximately 4.13) data sets [45].

XSL•FO
RenderX

Table 2. Protected health information statistics for the 4 sentence groups (English-only sentences, Chinese-only sentences, mixed Chinese-English sentences, and numeric or symbolic sentences).

Protected health information type	English-only sentences, n		Chinese-only sentences, n		Chinese-English mixed sentences, n		Numeric or symbolic sentences, n		Entire corpus, n	
	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
Date[a]	12,436	2346	2146	458	4530	919	19,850	3724	38,962	7447
Age[a]	2444	533	21	5	639	128	0	0	3104	666
Name[a]	100	18	273	44	1339	240	0	0	1712	302
Patient	9	9	106	12	42	11	0	0	157	32
Person	8	3	6	1	128	32	0	0	142	36
Doctor	83	6	161	31	1169	197	0	0	1413	234
Location[a]	5030	1014	4812	1108	5330	1132	4	3	15,176	3257
Named location	17	10	11	3	271	69	0	0	299	82
Nationality	21	7	3	1	17	10	0	0	41	18
Region	6	11	0	0	10	1	0	0	16	12
Country	355	52	6	5	171	44	0	0	532	101
City	196	39	26	6	510	101	0	0	732	146
Hospital	1552	266	73	11	1915	393	0	0	3540	670
Department	1616	391	2278	542	762	111	0	0	4656	1044
Room	1007	182	1200	267	248	53	4	3	2459	505
Number	0	0	1162	256	25	1	0	0	1187	257
School	28	6	15	2	642	159	0	0	685	167
Generic location	218	45	37	15	708	173	0	0	963	233
Market	12	3	1	0	50	17	0	0	63	20
Profession[a]	575	107	82	9	1566	449	0	0	2223	565
ID[a]	54	0	192	26	197	5	7	0	449	31
ID number	43	0	98	11	170	2	7	0	318	13
Medical record	1	0	92	14	19	3	0	0	112	17
Total number of sentences	20,639	4018	7526	1650	13,604	2873	19,861	3727	61,630	12,268

[a]Coarse-grained protected health information.

Deidentification Performance Comparison

The performance comparison of the developed deidentification methods is shown in Table 3, where the DBA indicates the dictionary-based baseline. The remaining configurations include monolingual BERT models (EN-BERT and CH-BERT) and multilingual BERT models (M-BERT and TM-BERT).

The results showed a significantly lower performance of the DBA, highlighting that the dictionary collected from the training set is insufficient to provide reliable deidentification results. To be specific, the micro- and macro-F-scores of the DBA were 0.13 and 0.16, respectively, while the micro- and macro-F-scores of BERT-based methods ranged from 0.90 to 0.94 and 0.63 to 0.79, respectively, which were significantly higher than those of the DBA. Among all BERT-based approaches, M-BERT outperformed the others in terms of both micro- and macro-F-scores. CH-BERT demonstrated a performance comparable to that of M-BERT on various PHI types such as “Name” and “Location.” The micro- and macro-F-scores of TM-BERT were higher than those of EN-BERT but lower than those of CH-BERT and M-BERT, suggesting that translating code-mixed sentences into monolingual text does not provide advantages. TM-BERT’s lower performance was mainly due to translation errors that can cause syntactic or semantic issues during fine-tuning. Some words like abbreviations (“NKUST”) and Chinese names that are PHI mentions cannot be accurately

XSL•FO
RenderX

translated. Even context words could alter their meaning, for example, the text “occupation: 英語補習班老師” (Profession: English tutoring teacher) was translated as “佔領: 英語補習班老師” (Domination: English tutoring teacher).

To gain a deeper understanding of the ability of PLMs to recognize different types of PHI mentioned in sentences with different levels of language mixing, we categorized the sentences in our corpus into 3 distinct sentence categories: English-only, Chinese-only, and Chinese-English mixed. We then evaluated the performance of the developed methods for each category. The results are presented in Table 4, focusing on the comparison between EN-BERT, CH-BERT, and M-BERT only. We observed that the distribution of different PHI types varied across the 3 sentence categories, suggesting that the performance of the developed models may be influenced by the prevalence of certain PHI types in their respective sentence categories. Specifically, we found that numeric PHI types, such as “Date” and “Age,” were more prevalent in English-only sentences, while nonnumeric PHI types, including “Name” and “Profession,” were more frequent in code-mixed sentences. This finding explains why EN-BERT performed similarly to multilingual and translation models on numeric PHI types but much worse on nonnumeric types. Additionally, the generally high performance for numeric PHI types indicates that mixing Chinese and English in the same sentence has little effect on their recognitions. Detailed results can be found in Multimedia Appendix 3.

In contrast, the “ID” PHI type occurred most frequently in the Chinese-only sentences, but all BERT-based models performed well in recognizing them. The “Location” PHI type was equally present in all 3 categories, but certain fine-grained types, such as “Region” and “Nationality,” were more challenging due to their limited training instances. Overall, the results showed that M-BERT outperformed the other models in the 3 sentence categories based on the metric of micro-F-score, while CH-BERT had the best performance in terms of macro-F-score in Chinese-only and code-mixed sentences. EN-BERT exhibited comparable performance with M-BERT in English-only sentences, but its performance was inferior to that of CH-BERT and M-BERT in Chinese-only and code-mixed sentences, indicating that these types of sentences are more challenging for EN-BERT. In contrast, both CH-BERT and M-BERT exhibited less effect of the code-mixing issue in their recognition of PHI in code-mixed sentences.

XSL•FO
RenderX

Table 3. Performance comparison for different methods on the test set.

Protected health information type	DBA[a]			EN-BERT[b,c]			CH-BERT[d]			M-BERT[e]			TM-BERT[f]		
	P[g]	R[h]	F[i]	P	R	F	P	R	F	P	R	F	P	R	F
Date	0.07	0.19	0.11	0.97	0.98[j]	0.98[j]	0.97	0.98[j]	0.97	0.98[j]	0.98[j]	0.98[j]	0.97	0.97	0.97
Age	0.14	0.16	0.15	0.95[j]	0.97	0.96[j]	0.94	0.98[j]	0.96[j]	0.94	0.97	0.95	0.93	0.97	0.95
Name	0.06	0.14	0.09	0.62	0.47	0.51	0.82	0.70[j]	0.74	0.87[j]	0.68	0.75[j]	0.73	0.59	0.64
Patient	0.04	0.09	0.05	0.59	0.41	0.48	1.00[j]	0.59[j]	0.75[j]	1.00[j]	0.59[j]	0.75[j]	0.94	0.59	0.72
Person	0.04	0.11	0.06	0.36	0.11	0.17	0.54	0.58[j]	0.56	0.66[j]	0.53	0.56[j]	0.30	0.25	0.27
Doctor	0.11	0.22	0.15	0.90	0.88	0.89	0.92	0.92	0.92	0.94	0.92	0.93	0.95[j]	0.93[j]	0.94[j]
Location	0.14	0.20	0.15	0.64	0.56	0.58	0.79[j]	0.69	0.72	0.78	0.74[j]	0.75[j]	0.67	0.71	0.69
Named location	0.01	0.01	0.01	0.34	0.26	0.29	0.64	0.37	0.47	0.61[j]	0.57[j]	0.59[j]	0.34	0.34	0.34
Nationality	0.43	0.14	0.21	0.67[j]	0.19	0.30	0.44	0.52	0.48	0.53	0.48	0.50	0.48	0.53[j]	0.50[j]
Region	0.00	0.00	0.00	0.67[j]	0.33[j]	0.44[j]	0.67[j]	0.17	0.27	0.29	0.17	0.23	0.20	0.17	0.18
Country	0.25	0.33	0.28	0.76	0.75	0.75	0.89	0.89	0.89	0.91[j]	0.90[j]	0.91[j]	0.87	0.86	0.87
City	0.21	0.52	0.30	0.58	0.59	0.58	0.86	0.74	0.79	0.86[j]	0.77[j]	0.82[j]	0.77	0.73	0.75
Hospital	0.35	0.47	0.40	0.84	0.85	0.84	0.89	0.94	0.91	0.93[j]	0.94[j]	0.94[j]	0.90	0.90	0.90
Department	0.24	0.35	0.29	0.90	0.87	0.88	0.92	0.88	0.90	0.93[j]	0.90[j]	0.92[j]	0.90	0.87	0.89
Room	0.04	0.15	0.06	0.87	0.92	0.89	0.89	0.94	0.91	0.91[j]	0.94[j]	0.92[j]	0.77	0.94	0.85
Number	0.05	0.08	0.06	0.99	1.00[j]	0.99	0.99	1.00[j]	0.99	0.99	1.00[j]	1.00[j]	1.00[j]	1.00[j]	1.00[j]
School	0.01	0.01	0.01	0.64	0.66	0.65	0.91[j]	0.82	0.86	0.85	0.87[j]	0.86[j]	0.76	0.82	0.79
General business	0.09	0.16	0.11	0.42	0.31	0.36	0.66	0.64	0.65	0.70[j]	0.73[j]	0.71[j]	0.61	0.61	0.61
Market	0.09	0.15	0.11	0.00	0.00	0.00	0.67	0.40	0.50	0.79[j]	0.55	0.65[j]	0.45	0.75[j]	0.57
Profession	0.09	0.22	0.12	0.46	0.43	0.44	0.66	0.73	0.69	0.78[j]	0.76[j]	0.77[j]	0.59	0.72	0.65
ID	0.34	0.31	0.32	0.75	0.96	0.83	0.79	1.00[j]	0.86	0.81[j]	1.00[j]	0.90[j]	0.77	1.00[j]	0.85
ID number	0.67	0.62	0.64	0.55	0.92	0.69	0.57	1.00[j]	0.72	0.65[j]	1.00[j]	0.79[j]	0.53	1.00[j]	0.70
Medical record	0.00	0.00	0.00	0.94	1.00[j]	0.97	1.00[j]	1.00[j]	1.00[j]	1.00[j]	1.00[j]	1.00[j]	1.00[j]	1.00[j]	1.00[j]
Micro-Average	0.09	0.22	0.13	0.91	0.90	0.90	0.93	0.93	0.93	0.94[j]	0.94[j]	0.94[j]	0.90	0.92	0.91
Macro-Average	0.15	0.20	0.16	0.67	0.62	0.63	0.80	0.75	0.76	0.81[j]	0.78[j]	0.79[j]	0.71	0.75	0.72

[a]DBA: dictionary-based approach.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]EN-BERT: BERT pretrained on English corpora.

[d]CH-BERT: BERT pretrained on simplified and traditional Chinese corpora.

[e]M-BERT: BERT pretrained on Wikipedia corpora from 104 languages.

[f]TM-BERT: M-BERT fine-tuned on the translated deidentification corpus.

[g]P refers to precision.

[h]R refers to recall.

[i]F refers to F1-score.

[j]The highest score among different methods.

XSL•FO
RenderX

Table 4. Comparison of performance between EN-BERT, CH-BERT, and M-BERT in English-only, Chinese-only, and Chinese-English mixed sentences.

PHI[a] type	English-only sentences				Chinese-only sentences				Chinese-English mixed sentences			
	EN-BERT[b,c]	CH-BERT[d]	M-BERT[e]	PHI, n[f]	EN-BERT	CH-BERT	M-BERT	PHI, n[f]	EN-BERT	CH-BERT	M-BERT	PHI, n[f]
Date[g]	0.96[h]	0.94	0.95[h]	2346[h]	0.98	0.97	0.99[h]	458[h]	0.93	0.93	0.94[h]	919[h]
Age[g]	0.97[h]	0.96	0.96	533[h]	0.67	0.91[h]	0.73	5[h]	0.94	0.97[h]	0.95	128[h]
Name[g]	0.24[h]	0.24[h]	0.24[h]	18[h]	0.38	0.54[h]	0.53	44[h]	0.54	0.80[h]	0.78	240[h]
Patient	0.00	0.00	0.00	9	0.64	1.00[h]	1.00[h]	12	0.50	0.78	0.88[h]	11
Person	N/A[i]	N/A	N/A	0	0.00	0.00	0.00	1	0.18	0.64[h]	0.48	32
Doctor	0.94[h]	0.94[h]	0.94[h]	9	0.49	0.61[h]	0.60	31	0.94	0.97	0.98[h]	197
Location[g]	0.66	0.61	0.70[h]	1014[h]	0.36	0.87[h]	0.83	1108[h]	0.37	0.66[h]	0.65	1132[h]
Named location	0.47	0.25	0.62[h]	10	0.00	0.40[h]	0.40[h]	3	0.28	0.53[h]	0.52	69
Nationality	0.67[h]	0.38	0.67[h]	7	0.00	1.00[h]	1.00[h]	1	0.00	0.86[h]	0.78	10
Region	0.47	0.55[h]	0.50	11	N/A	N/A	N/A	N/A	0.00	0.00	0.00	1
Country	0.87	0.85	0.95[h]	52	0.25	1.00[h]	1.00[h]	5	0.63	0.94[h]	0.88	44
City	0.74	0.72	0.77[h]	39	0.22	1.00[h]	0.92	6	0.53	0.85[h]	0.82	101
Hospital	0.93[h]	0.92	0.93[h]	266	0.21	0.80	0.83[h]	11	0.79	0.92	0.93[h]	393
Department	0.82	0. 81	0.82[h]	391	0.95	0.96[h]	0.96[h]	542	0.68	0.76	0.77[h]	111
Room	0.85	0.83	0.85[h]	182	0.97	0.98[h]	0.97	267	0.68	0.76[h]	0.76	53
Number	N/A	N/A	N/A	N/A	0.99	1.00[h]	0.99	256	0.00	0.00	0.00	1
School	0.71	0.25	0.83[h]	6	0.00	0.67[h]	0.50	2	0.65	0.92[h]	0.90	159
Generic business	0.72[h]	0.59	0.70	45	0.00	0.89[h]	0.72	15	0.27	0.66	0.71[h]	173
Market	0.00	0.50[h]	0.00	3	N/A	N/A	N/A	N/A	0.00	0.67	0.73[h]	17
Profession[g]	0.78	0.71	0.81[h]	107[h]	0.17	0.67[h]	0. 61	9[h]	0.36	0.70	0.75[h]	449[h]
ID[g]	N/A	N/A	N/A	N/A	0.90	0.90	0.91[h]	26[h]	0.60	0.90[h]	0.75	5[h]
ID number	N/A	N/A	N/A	N/A	0.79	0.79	0.81[h]	11	0.33	0.80[h]	0.50	2
Medical record	N/A	N/A	N/A	N/A	1.00[h]	1.00[h]	1.00[h]	14	0.86	1.00[h]	1.00[h]	3
Micro-F-score	0.92	0.90	0.92[h]	4018	0.93	0.96[h]	0.96[h]	1650	0.71	0.85	0.87[h]	2873
Macro-F-score	0.60[h]	0.54	0.60[h]	4018	0.47	0.83[h]	0.79	1650	0.48	0.74[h]	0.72	2873

[a]PHI: protected health information.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]EN-BERT: BERT pretrained on English corpora.

[d]CH-BERT: BERT pretrained on simplified and traditional Chinese corpora.

[e]M-BERT: BERT pretrained on Wikipedia corpora from 104 languages.

[f]The number of training instances of the PHI type.

[g]Coarse-grained PHI.

[h]The highest F-score among the 3 methods.

[i]N/A: not applicable.

XSL•FO
RenderX

Impact of the CMI Level on Deidentification Performance

In this subsection, we report sentence error rates by dividing sentences in our corpus into 6 groups according to their CMI levels, as illustrated in Figure 3, to gain insights into the impact of the code-mixing level on PHI recognition. It is evident from the figure that EN-BERT has the highest error rate for all CMI ranges. This aligns with our earlier finding in the previous section demonstrating that EN-BERT struggles with recognizing PHI in code-mixed sentences. Although CH-BERT and M-BERT exhibited lower sentence error rates than EN-BERT, the error rates were still considerably higher for code-mixed sentences (CMI >0), indicating that the recognition of PHI in code-mixed sentences remains a challenging task for the 2 PLMs.

Further evidence supporting this conclusion can be found in Table 5, where we fine-tuned M-BERT on training instances from 1 sentence category and evaluated it on the test set categorized into the 3 categories. From the results, we observed that fine-tuning M-BERT with training instances from a specific sentence category can significantly improve its performance on that category. However, fine-tuning on the code-mixed training instances yielded the best results, which further enabled the model to effectively recognize PHI in both Chinese and English sentences. This finding emphasizes the significance of incorporating code-mixed training instances into the model's training data.

Figure 3. Sentence error rate in each code-mixing index range for the test set. The numbers in parentheses represent the number of incorrect examples for the corresponding code-mixing index range. CH-BERT: BERT pretrained on simplified and traditional Chinese corpora; EN-BERT: BERT pretrained on English corpora; M-BERT: BERT pretrained on Wikipedia corpora from 104 languages.

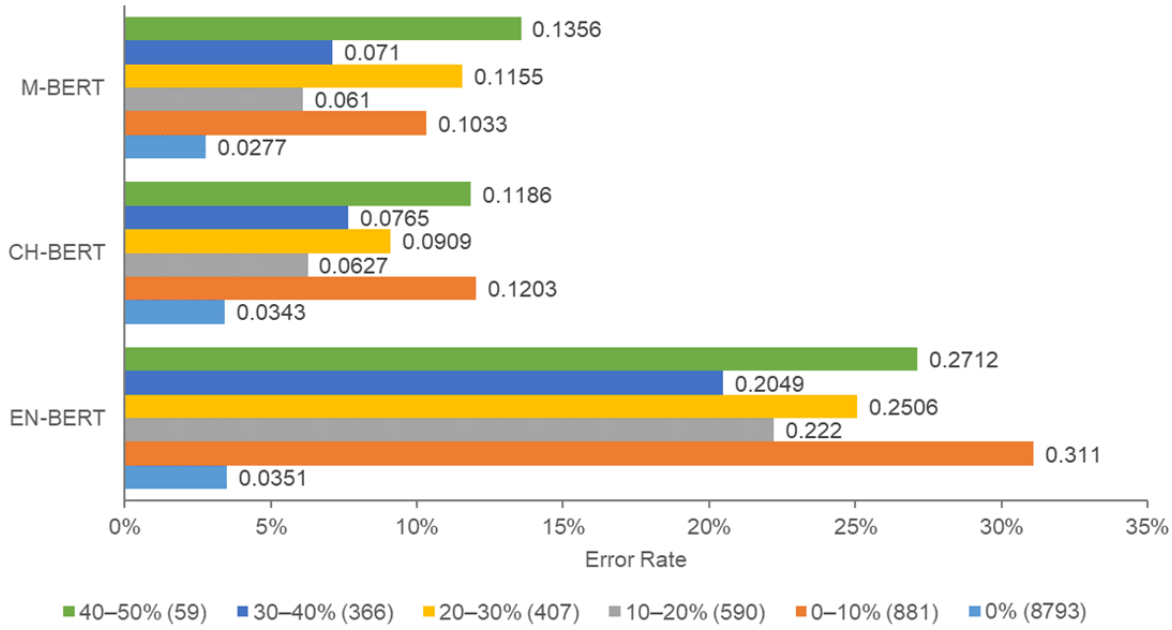

Table 5. The "zero-shot" performance of BERT pretrained on Wikipedia corpora from 104 languages (M-BERT) for the 3 sentence categories in terms of micro-F-scores.

Evaluation/Fine-tuning	English-only	Chinese-only	Chinese-English mixed
English-only	0.93	0.42	0.57
Chinese-only	0.59	0.96	0.65
Chinese-English mixed	0.80	0.94	0.87

Performance Comparison With ChatGPT

To assess the effectiveness of the proposed ChatGPT-based framework for recognizing PHI in our code-mixed corpus, we collected a subset of the original data that included sentences containing one or more PHI sets that were not recognized by any of the BERT-based models we developed. This subset was chosen to represent the most challenging cases of the compiled corpus while minimizing the risk of sensitive information exposure. Figure 4 shows the results, with M-BERT used for comparison.

Compared with M-BERT, the proposed ChatGPT-based framework achieved slightly better recall, but lower precision, leading to a lower micro-F-score in the zero-shot setting. Nonetheless, our framework presents an appealing approach to use LLMs for the deidentification task over standard-sized models like BERT. LLMs are notably easier to use and can be controlled through natural language prompts, requiring little to no machine learning expertise. For example, we found only 1 error case for the "Doctor" PHI that both BERT-based models and ChatGPT failed to recognize: "follow-up: (ydh/ptz)." However, we can enhance the prompt shown in Figure 2 by

XSL•FO
RenderX

including a statement following the "DOCTOR NAME" definition that lists the abbreviation names of physicians in a hospital, directing ChatGPT to recognize "ptz" as a doctor name: "Here is a list of abbreviated physician names of our hospital. Remember and recognize them from the given text: 1. hjd, 2. ptz, 3. ctc." This exemplifies the adaptability and flexibility of LLMs for deidentification tasks, making them particularly well-suited for real hospital environments. While the potential for using ChatGPT in deidentification is notable, the model is currently only accessible through an online application programming interface. This renders it unsuitable for use in a hospital setting, where patient data cannot be stored or transmitted to unauthorized external parties.

On the other hand, it crucially highlights the significance of carefully crafted prompts when working with ChatGPT. The constraint statements presented in Figure 1 play a pivotal role in guiding ChatGPT's responses. When these constraints were removed, we observed that ChatGPT can generate responses that may be unpredictable and go beyond the desired control. Take the text "he was brought to Dr. 莊凱傑's opd on 3/13 and haldol + anxicam 1 amp was given" as an example. Here, "莊凱傑" represents a Chinese name, which can be translated to "KAI-JIE ZHUANG." Using the developed prompts without adding the constraint part, both the original and recently released ChatGPT 3.5 models (version from September 27, 2023) could potentially return responses like "MEDICATION: haldol, anxicam" and "DOSE: 1 amp."

Another issue to consider is the translation of recognized PHI into a language other than that mentioned in the given text. The issue is more prominent in Chinese-majority code-mixed sentences, where the recognized Chinese PHI might be translated into the English counterparts. For example, in the sentence "education: 高中讀中和高中，僅唸數個月就因跟不上而休學, 大學念正修夜間部商學" (education: high school-Zhonghe senior high school, dropped out after only a few months; College-Business in the night school at Cheng Shiu university), the Chinese PHI could be extracted as translated forms in English, which can lead to a problem in surrogate generation and affect the performance evaluation procedure.

Lastly, similar to the observations in previous work [50,51], we found that the practice of requesting ChatGPT to generate responses 3 times resulted in the improvement of the overall precision, recall, and F1-score by 0.055, 0.068, and 0.064, respectively. The strategy can effectively filter out noisy outputs generated by ChatGPT. For instance, even with the inclusion of constraint statements in our prompts, we noticed that ChatGPT occasionally generated responses that violated the defined constraints, introducing undesired entity types such as "MEDICATION," "DOSE," or "DIAGNOSIS." By implementing the 3-query strategy, we could successfully eliminate these cases, ensuring the quality and relevance for our guideline of the generated content. Further details regarding the limitations for this strategy and its implications are provided in the limitations section.

Figure 4. Performance comparison with the ChatGPT-based deidentification framework and M-BERT on the sampled test subset. M-BERT: BERT pretrained on Wikipedia corpora from 104 languages.

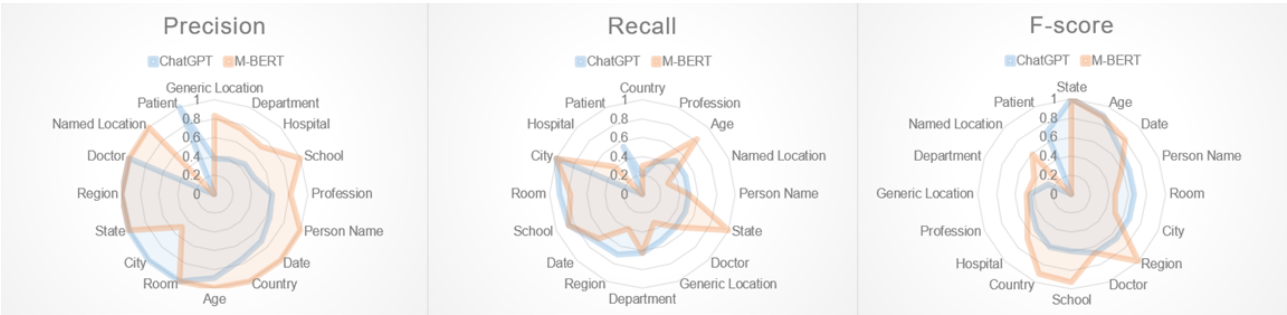

Discussion

Ambiguity Issue Among PHI

We explored the issue of ambiguity in the classification of fine-grained PHI at the coarse-grained level. When fine-grained PHI is misclassified, it may belong to the same coarse-grained PHI or different ones. Misclassifying it under the same coarse-grained PHI is generally more acceptable since fine-grained PHI belonging to the same coarse-grained category tends to share similar semantic characteristics. To quantify the extent of ambiguity, we calculated the micro-F-scores at the coarse-grained level, treating an incorrect recognition of a fine-grained PHI type as a true positive if it belongs to the same coarse-grained PHI type as the corresponding gold annotation. Our results (detailed in Multimedia Appendix 4) showed that intratype ambiguities are more prevalent than intertype ambiguities in "Name" and "Location" PHI types.

To gain insights into the level of ambiguity among the "Location" fine-grained types, we visualized the representations generated by M-BERT for different "Location" PHI types through t-SNE. The resulting plot (Figure 5) revealed that the clusters of "Nationality," "Country," and "Region" are intertwined, while the cluster of "ID" is notably separated from the others. The high ambiguity may be due to the limited number of training instances, as the number of training instances is imbalanced among PHI as illustrated in Table 2. As a result, all BERT-based models yielded better F-scores for "Country" than for "Nationality" and "Region."

On the other hand, the "ID" PHI type had a perfect mapped recall, indicating that all of the coarse-grained PHI could be recognized by the developed models. The confusion matrix shown in Figure S1 in Multimedia Appendix 4 reveals only 1 ambiguous case where M-BERT misclassified the "Location" PHI type as the "ID" PHI type. This occurred in the sentence "She visited psychiatric clinic of 804 then but have neither

XSL•FO
RenderX

regular follow-up nor fair drug compliance," where "804" refers to the Military Taoyuan General Hospital (804 is the army number assigned to the hospital), but the model recognized it as "ID" PHI. This example demonstrates one of the most challenging cases where the model failed at recognition, even with the recall-oriented ChatGPT framework. To correctly classify the instance, the model requires additional domain knowledge. We attempted to use ChatGPT to determine whether it knows that "804 hospital" refers to the military hospital, but it replied that "804 hospital" was not a common name or official name for the hospital. However, a quick Google search revealed that the website of the hospital was the first suggested link, indicating the importance of external knowledge sources for accurate classification [56].

Figure 5. t-Distributed Stochastic Neighbor Embedding visualization of the 5 fine-grained "Location" protected health information (PHI) types and the "ID" PHI type for M-BERT. M-BERT: BERT pretrained on Wikipedia corpora from 104 languages.

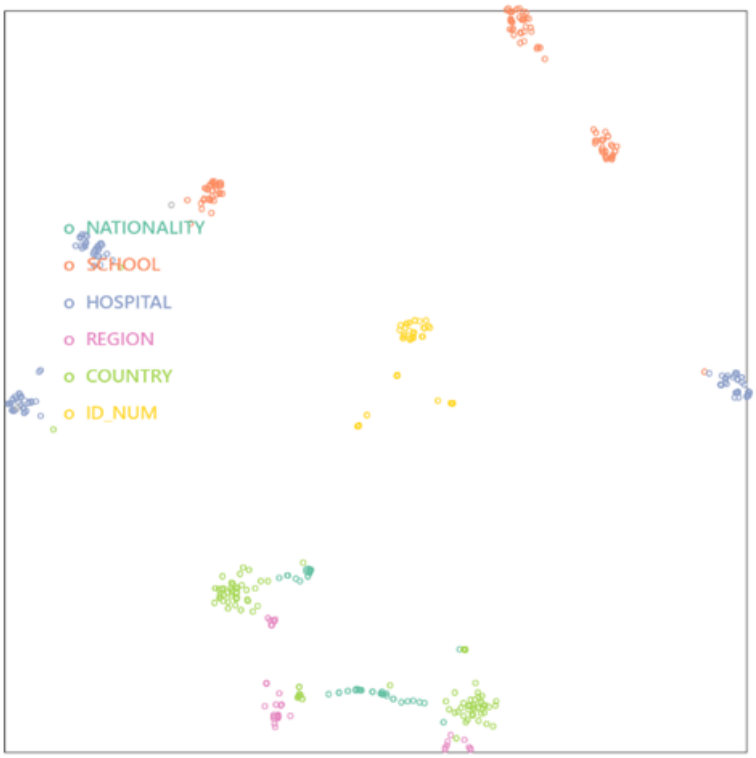

OOV Case Study to Interpret the Models' Decisions

This section delves into how the BERT-based models use regularity and mention coverage, as outlined in the Introduction section, during their recognition process. First, we estimated the OOV rates on the test set to determine the mention coverage rate of our corpus. The estimated rates were 0.356, 0.953, 0.686, 0.876, 0.696, and 0.290 for the "Date," "Age," "Name," "Location," "Profession," and "ID" categories, respectively. OOV was defined as a mention containing at least one word not present in the vocabulary compiled from the training data set. The inadequate performance of the DBA along with the high OOV rate suggests that the effect of the mention coverage in our corpus is limited. To analyze the effect, we re-estimated the performance of coarse-grained PHI for tokens belonging to in-vocabulary (IV) and OOV using EN-BERT, M-BERT, and TM-BERT in Figure 6. More detailed results are available in Multimedia Appendix 4.

The presented results reveal that the developed models have different levels of tolerance for OOV issues when recognizing different PHI types. Interestingly, we observed that the EN-BERT model achieved indistinguishable high F-scores for IV and OOV tokens for the "Date" PHI type. This is notable since not all Chinese characters used in the "Data" PHI are presented in the EN-BERT vocabulary. For example, the numeric value "六" (six) is not available as shown in Table 6. We believe that the presence of those characters in the extensive training data enabled EN-BERT to leverage the naming regularity of the "Date" PHI and improve its ability to identify occurrences of these entities in Chinese-only and code-mixed sentences. Figure 7 illustrates how EN-BERT's attention mechanism can use specific patterns in text, such as "年月日," for recognizing "Date" PHI, even when the majority of the Chinese tokens are mapped to [UNK] (unknown word) by the EN-BERT tokenizer. The model's attention was primarily drawn to the pattern words "年" and "月," allowing it to successfully assign labels to the respective unknown words.

However, the model's overreliance on pattern words and strong name regularities could lead to misclassification. Figure 8 illustrates an example of the false-positive case generated by EN-BERT. The model has been trained to recognize the first likely token for "Date" PHI by considering its surrounding tokens, including previous unidentified tokens, the forward slash (/), and the final token representing the end of the "Date" PHI. However, this mechanism caused the model to incorrectly extract the text "5/4/3" as "Date" PHI. The third AHV displayed in Figure 4 illustrates that when the previous tokens are not unknown, the model focuses more on the preceding token (a

XSL•FO
RenderX

number) and gives less attention to the pattern tokens ("/"), which helps prevent the occurrence of false positives.

On the other hand, the "Age," "Location," and "Profession" PHI types exhibited higher IV F-scores compared to their OOV scores. Specifically, the "Age" PHI type demonstrated an almost perfect IV F-score and slightly lower OOV score across all 3 models. After examining the outcomes of the developed models, we arrived at a conclusion similar to that of the "Date" PHI. The models could distinguish the "Age" PHI type in various code-mixing–level sentences because of the consistent naming pattern. There were only a few cases in which the developed models failed. For instance, in the sentence, "occupation: 會計18y->美容師3y->保險電話銷售人員->未上市股票電話銷售員->賣雞排 (2223.3-2224.9)," only CH-BERT was able to correctly avoid classifying "18" as "Age" PHI, while the other BERT-based models and the developed ChatGPT framework recognized "18" as "Age" PHI.

In contrast, the "Location" and "Profession" PHI types exhibited lower IV F-scores and apparently lower OOV F-scores. To investigate the discrepancies, we probed the results of the developed model by applying input reduction. We took the input sentence "travel history: [PHI] last week" as an example to summarize the observations as follows. First, we noticed that for EN-BERT, the presence of unknown tokens in the input sentence resulted in heavy reliance on the regularity of entity naming and contextual patterns for effective generalization over unseen mentions or mentions that contain unknown words. In the left part of Figure 9, the AHV represents the reduced interpretation of the 2 "City" PHI inputs "台北" (Taipei) and "恆春" (Hengchun) that contain unknown words. From the view, we can see that the preceding token, "travel," provides a strong indication for EN-BERT to recognize "City" PHI. The word "台北" ("[UNK]北" for EN-BERT) was considered as "Hospital" PHI, indicating that the model learned to memorize the most frequent label assigned to the IV word, such as "署/[UNK]北" hospitals, in the training set.

Second, in addition to the naming regularity and contextual patterns, CH-BERT and M-BERT can leverage their representation for the given token sequence to assign the corresponding labels better. This is evident from the right part of Figure 5 in which we can see after fine-tuning, M-BERT was able to improve its representation of "Hospital" PHI entities, such as "ntuh," "台大" (National Taiwan University Hospital), and "署北" (Taipei Hospital), as well as "Date" PHI entities, such as "九月" (September) and "mk87," by clustering them in closer proximity to each other. Despite being OOV entities, OOV PHI, such as "mk99," "菜園" (vegetable garden), "軍醫院" (military hospital), and "小診所" (small clinic), were still represented in close proximity to the corresponding IV PHI, such as "mk87," "clothing factory," and "療養院" (nursing home). However, "Profession" PHI and some of the fine-grained "Location" PHI types, such as "Generic location," lacked the regularity of their naming conventions, which posed a challenge for accurate recognition. Moreover, informal language, typographical errors, ambiguity issues, and limited training instances for fine-grained PHI, such as "Region" and "Market," further complicate the task of correctly identifying these OOV PHI mentions.

To mitigate the aforementioned OOV issue, we highlight several promising research directions. The first direction is data augmentation. During the pretraining phases, BERT-based models can benefit from automatically generating synthetic training examples that include PHI. This can be accomplished by leveraging LLMs to create data of reasonable quality [57]. This approach, similar to data augmentation, should help improve the model's robustness when handling different contexts and OOV PHI. The second direction is vocabulary expansion. For specific categories like "Location" and "Profession," where we observed lower F-scores on OOV PHI, traditional dictionary expansion methods [58] can be applied to augment BERT's vocabulary. This approach empowers the model to recognize and process OOV and UNK words more effectively. The third direction is knowledge graph integration [59]. This approach enriches the contextual information available to the model, providing additional context and semantics to assist in recognizing OOV or UNK PHI and thereby improving overall performance.

Figure 6. F-scores of coarse-grained protected health information (PHI) for tokens belonging to in-vocabulary (IV) and out-of-vocabulary (OOV) in the test set. EN-BERT: BERT pretrained on English corpora; M-BERT: BERT pretrained on Wikipedia corpora from 104 languages; TM-BERT: M-BERT fine-tuned on the translated deidentification corpus.

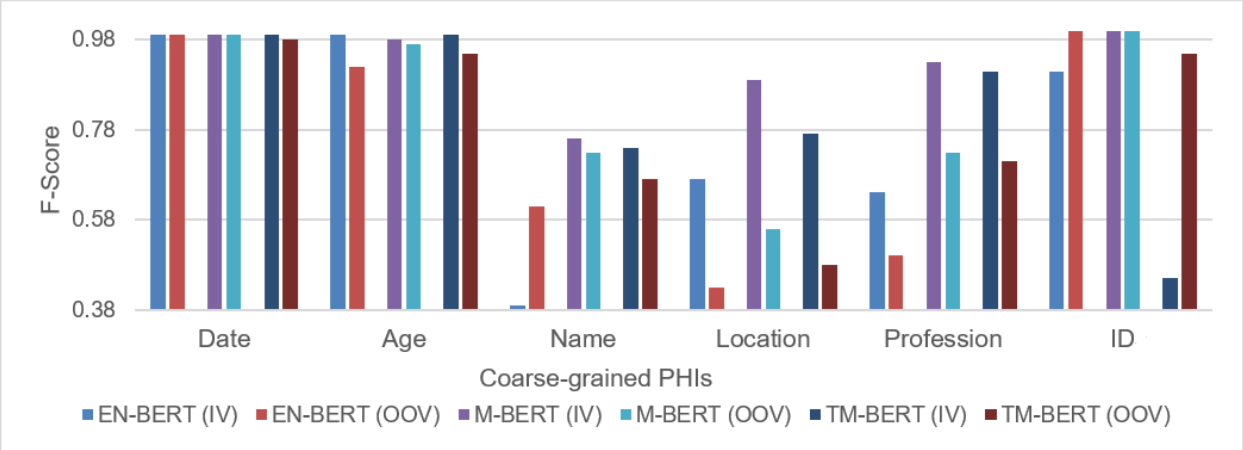

XSL•FO
RenderX

Table 6. Examples of Chinese characters related to the "Date" protected health information listed in the BERT pretrained on English corpora (EN-BERT) vocabulary.

Chinese character	Example "Date" PHI[a]
清、明、中、秋	清明節、中秋節
一、二、三、四、五、八、十、月、日、年、上、下、年、西、星、民、國	三月八日、上周五、下周一、西元2321年、星期五、民國一一二年、今年十一月

[a]PHI: protected health information.

Figure 7. The attention-head view for the following code-mixed sentence: "she tried to hang herself but failed due to drunk (大約前年五月開始不斷有死的念頭)" with the tokens "[UNK]" before/after "年 (year)" selected. The English translation for the Chinese text is "she had suicidal thoughts since around May of two years ago." [UNK] is the token for unknown word. [CLS] is a classification token and is added to the beginning of every sentence. [SEP] is a separation token inserted in between 2 sentences.

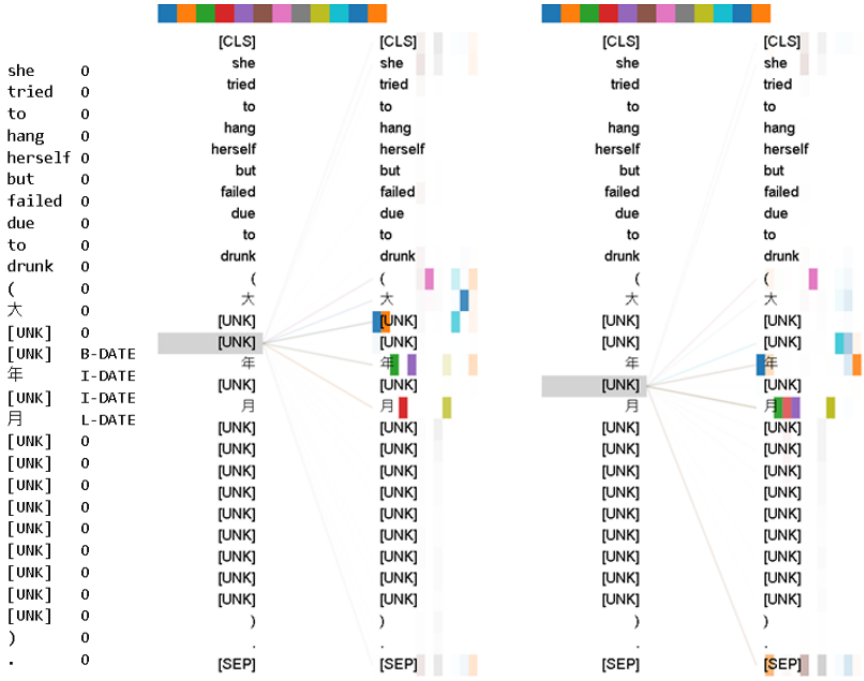

Figure 8. The first and second attention-head views have been generated for EN-BERT when applied to process the following sentences: "他在 78/12/15 求診" (sampled from the training set; the English translation is "He went to see a doctor on 78/12/15"; "78/12/15" is annotated as "Date" PHI) and "3. ocd (不時想到數字5/4/3/8)" (sampled from the test set; the English translation is "sometimes thinks of the numbers 5/4/3/8"). The third view has been generated for EN-BERT when processing the translated text of the second sentence. [UNK] is the token for unknown word. [CLS] is a classification token and is added to the beginning of every sentence. [SEP] is a separation token inserted in between 2 sentences. EN-BERT: BERT pretrained on English corpora; PHI: protected health information.

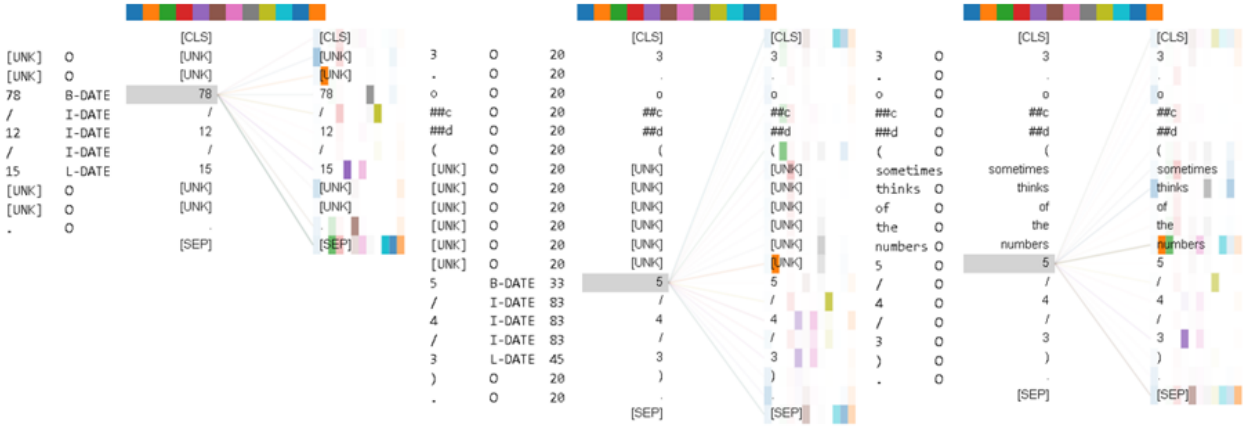

XSL•FO
RenderX

Figure 9. The left part shows the attention-head views (AHVs) of the input reduction probing for the input text "travel history: [PHI] last week." The first column of the AHVs comprises 3 rows, wherein the PHI includes "台北" and "恆春." We also present the result of the input "台北" in the last row of the first column in the view. The right part shows the t-Distributed Stochastic Neighbor Embedding visualizations of EN-BERT, M-BERT without fine-tuning, and M-BERT with fine-tuning from top to bottom for randomly sampled in-vocabulary (IV) and out-of-vocabulary (OOV) mentions of the 3 fine-grained "Location" PHI types including "Named location," "Generic location," and "Hospital." For comparison, we also include 5 "Date" PHI instances including "過年" (New Year), "November," "九月" (September), "mk87," and "mk99" (where "mk" is a transliteration of "民國," which refers to the Republic of China calendar). [UNK] is the token for unknown word. [CLS] is a classification token and is added to the beginning of every sentence. [SEP] is a separation token inserted in between 2 sentences. PHI: protected health information; EN-BERT: BERT pretrained on English corpora; M-BERT: BERT pretrained on Wikipedia corpora from 104 languages.

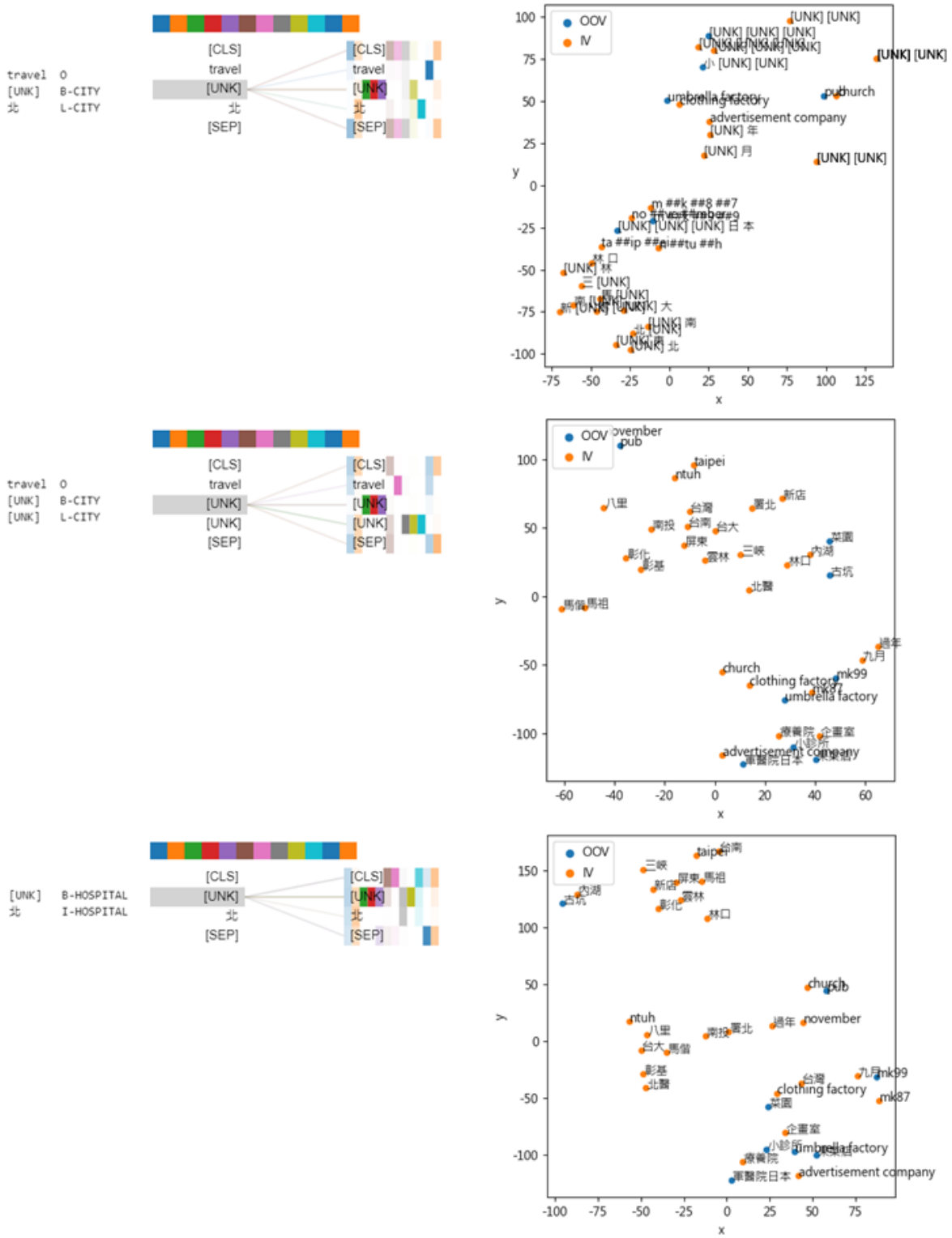

XSL•FO
RenderX

Limitations

While our study has made significant contributions to the field of clinical text deidentification in code-mixed languages, it is important to acknowledge some limitations. First, despite the successes achieved, M-BERT, our best-performing model, has some limitations. Notably, it exhibits intratype ambiguities in certain PHI, such as "Location," likely due to the limited number of training instances. Second, M-BERT yields decreased performance for OOV terms compared to IV terms in specific PHI categories, such as "Profession" and "Location." Third, our study was based on a Chinese-English code-mixed corpus from Taiwan, which has its unique writing conventions and style. As a result, the fine-tuned models may not be entirely applicable to other bilingual or multilingual code-mixed corpora from different countries.

Another notable limitation is related to the normalization of mixed sentences into their dominant word-usage language based on the CMI. In the training process for TM-BERT, we relied on an existing translation model as a means to normalize code-mixed text. While this strategy was applied to alleviate the code-mixing challenge, it is crucial to note that translation errors because of the limited context can introduce syntactic and semantic issues during fine-tuning, which can, in turn, impact the model's overall performance. This limitation underscores the need for further research to develop more sophisticated and accurate code-mixed text normalization techniques, as explored previously [60].

In the context of the proposed ChatGPT-based deidentification framework, we acknowledge that the framework's performance is sensitive to several factors, including the crafted prompts, the number of query times applied for the majority voting strategy, and the specific model version used. For example, we observed that when using the recently released model (version from September 27, 2023) for recognizing PHI, it experienced a higher frequency of failures due to its reluctance to process sensitive content, as compared to the version we employed in our experiments. To address this limitation, it is needed to incorporate new specific constraints, such as "None of your responses will contain 'I'm sorry,' 'I apologize,' 'I'm sorry, but I can't assist with that request,' or similar." These constraints were introduced to mitigate potential issues, but they also shed light on another limitation. As ChatGPT is subject to continuous updates by OpenAI, the specific model version we used for our study may not always reflect the latest advancements in performance. This dynamic nature of the model versions could impact the generalizability of our results and the application of the framework in evolving contexts. We acknowledge this as a potential limitation, and it is an area where ongoing research and adaptation will be essential to keep pace with model improvements and changes.

Lastly, while our decision to use the prompt 3 times is consistent with prior methodologies [50,51], we acknowledge that the specific number of repetitions can be subject to experimentation and may depend on the context and objectives of the task. Future research could delve into the selection of repetitions and its influence on result distribution to gain a deeper understanding of this aspect.

Conclusions

We developed the first-ever code-mixed clinical deidentification corpus, which exhibits a higher CMIC value compared to other known code-mixed data sets, suggesting that Chinese-English code-mixed clinical writing style is extensively used in the presented data set. We found that different PHI types had preference in their occurrence within the sentence categories, with "Date" and "Age" PHI types more common in English-only sentences, and "Name" and "Profession" PHI types more common in code-mixed sentences.

Our hypothesis that PLMs rely on naming regularity to recognize PHI was supported by our experimental results. Additionally, the results indicated that PLMs like M-BERT also use their learned representations to enhance their classification performance. We found that the fine-tuned M-BERT models outperformed the other approaches in most PHI types. We also observed that fine-tuning with code-mixed training instances is critical for significantly improving performance on the code-mixed data set. This finding emphasizes the significance of incorporating code-mixed training instances into the model's training data. Our analysis revealed that TM-BERT yielded lower performance, suggesting that machine translation is not required for M-BERT in addressing the challenges posed by code mixing owing to the error-prone nature of machine-translated sentences.

We conclude that the LLM-based deidentification method is a feasible and appealing approach, which requires little to no machine learning expertise, and can be controlled and enhanced through natural language prompts. The experience of engineering our prompt also emphasizes the crucial role of carefully crafted prompts to avoid unwanted output. Further research could explore the augmentation of PLMs and LLMs with external knowledge to improve the strength in recognizing rare PHI and could incorporate experiments using the recently released GPT-4, which has been reported to demonstrate significant improvements in many tasks. However, the use of such a method in the hospital setting requires careful consideration of data security and privacy concerns.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data for the compiled corpus.
[DOCX File , 25 KB - jmir_v26i1e48443_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Supplementary data for the developed methods.
[DOCX File , 22 KB - jmir_v26i1e48443_app2.docx]

Multimedia Appendix 3
Detailed experimental results.
[DOCX File , 69 KB - jmir_v26i1e48443_app3.docx]

Multimedia Appendix 4
Supplementary data for the discussion.
[DOCX File , 120 KB - jmir_v26i1e48443_app4.docx]

References

1. Silvestri S, Esposito A, Gargiulo F, Sicuranza M, Ciampi M, De Pietro G. A Big Data Architecture for the Extraction and Analysis of EHR Data. 2019 Presented at: 2019 IEEE World Congress on Services (SERVICES); July 08-13, 2019; Milan, Italy. [doi: 10.1109/services.2019.00082]
2. Cannon J, Lucci S. Transcription and EHRs. Benefits of a blended approach. J AHIMA 2010 Feb;81(2):36-40. [Medline: 20218195]
3. Li W. The 'why' and 'how' questions in the analysis of conversational code-switching. In: Auer P, editor. Code-Switching in Conversation: Language, Interaction and Identity. Abingdon, UK: Routledge; 1999:156-179.
4. Alqurashi HS. Investigating The Code Switching Phenomenon In Private Medical Workplaces: A Case Study Of Some Private Hospitals In Saudi Arabia. Journal of Language and Linguistic Studies 2022;18(4):344-361 [FREE Full text]
5. Dowlagar S, Mamidi R. A code-mixed task-oriented dialog dataset for medical domain. Computer Speech & Language 2023 Mar;78:101449. [doi: 10.1016/j.csl.2022.101449]
6. Keresztes C. Investigation of English language contact-induced features in Hungarian cardiology discharge reports and language attitudes of physicians and patients. SZTE Doktori Repozitórium. 2010. URL: https://doktori.bibl.u-szeged.hu/id/eprint/826/ [accessed 2023-12-09]
7. Karuthan A. The English language needs of nurses in medical tourism in Malaysia. University of Malaya. 2015. URL: https://core.ac.uk/display/268877454 [accessed 2023-12-09]
8. Lee Y, Wang B, Su C, Chen P, Lin W, Wu C, et al. Protected Health Information Recognition of Unstructured Code-Mixed Electronic Health Records in Taiwan. Stud Health Technol Inform 2022 Jun 06;290:627-631. [doi: 10.3233/SHTI220153] [Medline: 35673092]
9. Lin H, Lu Y, Tang J, Han X, Sun L, Wei Z, et al. A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-main.592]
10. Kim H, Kang J. How Do Your Biomedical Named Entity Recognition Models Generalize to Novel Entities? arXiv. 2022. URL: https://arxiv.org/abs/2101.00160 [accessed 2023-12-09]
11. ISLab: Intelligent System Lab. GitHub. URL: https://nkustislab.github.io/active_project/clinical%20natural%20language%20processing/deid_tool/ [accessed 2023-12-09]
12. Standards for Privacy of Individually Identifiable Health Information. Federal Register. URL: https://www.federalregister.gov/documents/2002/08/14/02-20554/standards-for-privacy-of-individually-identifiable-health-information [accessed 2023-12-09]
13. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med 2011 May;39(5):952-960 [FREE Full text] [doi: 10.1097/CCM.0b013e31820a92c6] [Medline: 21283005]
14. Aguilar G, AlGhamdi F, Soto V, Diab M, Hirschberg J, Solorio T. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching. 2018 Presented at: Third Workshop on Computational Approaches to Linguistic Code-Switching; July 19, 2018; Melbourne, Australia. [doi: 10.18653/v1/W18-3219]
15. Tang T, Tang X, Yuan T. Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text. IEEE Access 2020;8:193248-193256. [doi: 10.1109/access.2020.3030468]
16. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
17. Uzuner O, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association 2007 Sep 01;14(5):550-563. [doi: 10.1197/jamia.m2444]
18. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S11-S19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.007] [Medline: 26225918]

XSL•FO
RenderX

19. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform 2017 Nov;75S:S4-S18 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.011] [Medline: 28614702]
20. Alla NLV, Chen A, Batongbacal S, Nekkantti C, Dai H, Jonnagaddala J. Cohort selection for construction of a clinical natural language processing corpus. Computer Methods and Programs in Biomedicine Update 2021;1:100024. [doi: 10.1016/j.cmpbup.2021.100024]
21. Jonnagaddala J, Chen A, Batongbacal S, Nekkantti C. The OpenDeID corpus for patient de-identification. Sci Rep 2021 Oct 07;11(1):19973 [FREE Full text] [doi: 10.1038/s41598-021-99554-9] [Medline: 34620985]
22. Bråthen S, Wie W, Dalianis H. Creating and Evaluating a Synthetic Norwegian Clinical Corpus for De-Identification. In: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). 2021 Presented at: 23rd Nordic Conference on Computational Linguistics (NoDaLiDa); May 31-June 2, 2021; Reykjavik, Iceland (Online).
23. Douglass MM, Cliffford GD, Reisner A, Long WJ, Moody GB, Mark RG. De-identification algorithm for free-text nursing notes. 2005 Presented at: Computers in Cardiology; September 25-28, 2005; Lyon, France.
24. Ferrández O, South B, Shen S, Friedlin F, Samore M, Meystre S. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013 Jan 01;20(1):77-83 [FREE Full text] [doi: 10.1136/amiajnl-2012-001020] [Medline: 22947391]
25. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010 Dec;79(12):849-859. [doi: 10.1016/j.ijmedinf.2010.09.007] [Medline: 20951082]
26. Ahmed T, Aziz MMA, Mohammed N. De-identification of electronic health record using neural network. Sci Rep 2020 Oct 29;10(1):18600 [FREE Full text] [doi: 10.1038/s41598-020-75544-1] [Medline: 33122735]
27. Dernoncourt F, Lee J, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017 May 01;24(3):596-606 [FREE Full text] [doi: 10.1093/jamia/ocw156] [Medline: 28040687]
28. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2019. URL: https://arxiv.org/abs/1810.04805 [accessed 2023-12-09]
29. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. 2015 Presented at: IEEE International Conference on Computer Vision (ICCV); December 07-13, 2015; Santiago, Chile. [doi: 10.1109/iccv.2015.11]
30. Vakili T, Lamproudis A, Henriksson A, Dalianis H. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). 2022 Presented at: 13th Conference on Language Resources and Evaluation; June 20-25, 2022; Marseille, France.
31. Winata GI, Wu CS, Madotto A, Fung P. Bilingual Character Representation for Efficiently Addressing Out-of-Vocabulary Words in Code-Switching Named Entity Recognition. arXiv. 2018. URL: https://arxiv.org/abs/1805.12061 [accessed 2023-12-09]
32. Wu S, Manber U. Fast text searching. Commun. ACM 1992 Oct;35(10):83-91. [doi: 10.1145/135239.135244]
33. Trivedi S, Rangwani H, Singh AK. IIT (BHU) Submission for the ACL Shared Task on Named Entity Recognition on Code-switched Data. In: Proceedings of The Third Workshop on Computational Approaches to Code-Switching. 2018 Presented at: Third Workshop on Computational Approaches to Code-Switching; July 19, 2018; Melbourne, Australia. [doi: 10.18653/v1/w18-3220]
34. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. arXiv. 2017. URL: https://arxiv.org/abs/1607.04606 [accessed 2023-12-09]
35. Godin F, Vandersmissen B, De Neve W, Van de Walle R. Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations. In: Proceedings of the Workshop on Noisy User-generated Text. 2015 Presented at: Workshop on Noisy User-generated Text; July 31, 2015; Beijing, China. [doi: 10.18653/v1/w15-4322]
36. Smith SL, Turban DHP, Hamblin S, Hammerla NY. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv. URL: https://arxiv.org/abs/1702.03859 [accessed 2023-12-09]
37. Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1493]
38. Wang C, Wang F, Lee Y, Chen P, Wang B, Su C, et al. Principle-Based Approach for the De-Identification of Code-Mixed Electronic Health Records. IEEE Access 2022;10:22875-22885. [doi: 10.1109/access.2022.3148396]
39. Wu C, Kuo C, Su C, Wang S, Dai H. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. J Affect Disord 2020 Jan 01;260:617-623. [doi: 10.1016/j.jad.2019.09.044] [Medline: 31541973]
40. Dai H, Su C, Lee Y, Zhang Y, Wang C, Kuo C, et al. Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients. Front Psychiatry 2020 Jan 15;11:533949 [FREE Full text] [doi: 10.3389/fpsyt.2020.533949] [Medline: 33584354]
41. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/bm.2012.031]

XSL•FO
RenderX

42. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008 Jul 24;8(1):32 [FREE Full text] [doi: 10.1186/1472-6947-8-32] [Medline: 18652655]
43. Shaalan K, Raza H. NERA: Named Entity Recognition for Arabic. J. Am. Soc. Inf. Sci 2009 Apr 22;60(8):1652-1663. [doi: 10.1002/asi.21090]
44. Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, et al. Multilingual is not enough: BERT for Finnish. arXiv. 2019. URL: https://arxiv.org/abs/1912.07076 [accessed 2023-12-09]
45. Gambäck B, Das A. On measuring the complexity of code-mixing. In: Proceedings of the 11th International Conference on Natural Language Processing. 2014 Presented at: 11th International Conference on Natural Language Processing; December 18-21, 2014; Goa, India.
46. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv. 2019. URL: https://arxiv.org/abs/1910.03771 [accessed 2023-12-09]
47. Yang X, Zhao X, Tjio G, Chen C, Wang L, Wen B, et al. Opencc – an open Benchmark data set for Corpus Callosum Segmentation and Evaluation. 2020 Presented at: IEEE International Conference on Image Processing (ICIP); October 25-28, 2020; Abu Dhabi, United Arab Emirates. [doi: 10.1109/ICIP40778.2020.9191097]
48. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv 2023 Jan 16;55(9):1-35. [doi: 10.1145/3560815]
49. Gutiérrez B, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. arXiv. 2022. URL: https://arxiv.org/abs/2203.08410 [accessed 2023-12-09]
50. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
51. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
52. Allen JF. Maintaining knowledge about temporal intervals. Commun. ACM 1983 Nov;26(11):832-843. [doi: 10.1145/182.358434]
53. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research 2008;9(86):2579-2605 [FREE Full text]
54. Feng S, Wallace E, Grissom II A, Iyyer M, Rodriguez P, Boyd-Graber J. Pathologies of Neural Models Make Interpretations Difficult. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/D18-1407]
55. Vig J. BertViz: A tool for visualizing multi-head self-attention in the BERT model. 2019 Presented at: ICLR Workshop: Debugging Machine Learning Models; May 6, 2019; New Orleans, LA.
56. Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv. 2023. URL: https://arxiv.org/abs/2302.12813 [accessed 2023-12-09]
57. Kumar V, Choudhary A, Cho E. Data Augmentation using Pre-trained Transformer Models. arXiv. 2020. URL: https://arxiv.org/abs/2003.02245 [accessed 2023-12-09]
58. Wang H, Yu D, Sun K, Chen J, Yu D. Improving Pre-Trained Multilingual Models with Vocabulary Expansion. arXiv. 2019. URL: https://arxiv.org/abs/1909.12440 [accessed 2023-12-09]
59. Gao N, Wang Y, Chen P, Tang J. Boosting Short Text Classification by Solving the OOV Problem. IEEE/ACM Trans. Audio Speech Lang. Process 2023;31:4014-4024. [doi: 10.1109/taslp.2023.3316422]
60. Yulianti E, Kurnia A, Adriani M, Duto YS. Normalisation of Indonesian-English Code-Mixed Text and its Effect on Emotion Classification. IJACSA 2021;12(11):0121177. [doi: 10.14569/IJACSA.2021.0121177]

Abbreviations

AHV: attention-head view
BERT: Bidirectional Encoder Representations from Transformers
CH-BERT: BERT pretrained on simplified and traditional Chinese corpora
CMI: code-mixing index
CMIC: code-mixing index for corpus
DBA: dictionary-based approach
EHR: electronic health record
EN-BERT: BERT pretrained on English corpora
HIPAA: Health Insurance Portability and Accountability Act
IV: in-vocabulary
LLM: large language model

XSL•FO
RenderX

M-BERT: BERT pretrained on Wikipedia corpora from 104 languages
MIMIC: Medical Information Mart for Intensive Care
NLP: natural language processing
OOV: out-of-vocabulary
PHI: protected health information
PLM: pretrained language model
TM-BERT: M-BERT fine-tuned on the translated deidentification corpus
t-SNE: t-Distributed Stochastic Neighbor Embedding

Edited by A Castonguay; submitted 24.04.23; peer-reviewed by L Zhu, D Chrimes, R Taira; comments to author 02.10.23; revised version received 27.10.23; accepted 05.12.23; published 25.01.24.

Please cite as:
Lee YQ, Chen CT, Chen CC, Lee CH, Chen P, Wu CS, Dai HJ
Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study
J Med Internet Res 2024;26:e48443
URL: https://www.jmir.org/2024/1/e48443
doi:10.2196/48443
PMID:38271060

©You-Qian Lee, Ching-Tai Chen, Chien-Chang Chen, Chung-Hong Lee, Peitsz Chen, Chi-Shin Wu, Hong-Jie Dai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study

Ziru Fu[1*], MSC; Yu Cheng Hsu[1*], PhD; Christian S Chan[2,3], PhD; Chaak Ming Lau[4], PhD; Joyce Liu[1], MA; Paul Siu Fai Yip[1,5], PhD

[1]The Hong Kong Jockey Club Centre for Suicide Research and Prevention, Faculty of Social Sciences, The University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[2]Department of Psychology, The University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[3]Department of Psychology and Linguistics, International Christian University, Tokyo, Japan

[4]Department of Linguistics and Modern Language Studies, The Education University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[5]Department of Social Work and Social Administration, Faculty of Social Sciences, The University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[*]these authors contributed equally

Corresponding Author:
Paul Siu Fai Yip, PhD
The Hong Kong Jockey Club Centre for Suicide Research and Prevention
Faculty of Social Sciences
The University of Hong Kong
2/F, The Hong Kong Jockey Club Building for Interdisciplinary Research
5 Sassoon Road, Pokfulam
Hong Kong SAR
China (Hong Kong)
Phone: 852 28315232
Email: sfpyip@hku.hk

Abstract

Background: Sentiment analysis is a significant yet difficult task in natural language processing. The linguistic peculiarities of Cantonese, including its high similarity with Standard Chinese, its grammatical and lexical uniqueness, and its colloquialism and multilingualism, make it different from other languages and pose additional challenges to sentiment analysis. Recent advances in models such as ChatGPT offer potential viable solutions.

Objective: This study investigated the efficacy of GPT-3.5 and GPT-4 in Cantonese sentiment analysis in the context of web-based counseling and compared their performance with other mainstream methods, including lexicon-based methods and machine learning approaches.

Methods: We analyzed transcripts from a web-based, text-based counseling service in Hong Kong, including a total of 131 individual counseling sessions and 6169 messages between counselors and help-seekers. First, a codebook was developed for human annotation. A simple prompt ("Is the sentiment of this Cantonese text positive, neutral, or negative? Respond with the sentiment label only.") was then given to GPT-3.5 and GPT-4 to label each message's sentiment. GPT-3.5 and GPT-4's performance was compared with a lexicon-based method and 3 state-of-the-art models, including linear regression, support vector machines, and long short-term memory neural networks.

Results: Our findings revealed ChatGPT's remarkable accuracy in sentiment classification, with GPT-3.5 and GPT-4, respectively, achieving 92.1% (5682/6169) and 95.3% (5880/6169) accuracy in identifying positive, neutral, and negative sentiment, thereby outperforming the traditional lexicon-based method, which had an accuracy of 37.2% (2295/6169), and the 3 machine learning models, which had accuracies ranging from 66% (4072/6169) to 70.9% (4374/6169).

Conclusions: Among many text analysis techniques, ChatGPT demonstrates superior accuracy and emerges as a promising tool for Cantonese sentiment analysis. This study also highlights ChatGPT's applicability in real-world scenarios, such as monitoring the quality of text-based counseling services and detecting message-level sentiments in vivo. The insights derived from this study pave the way for further exploration into the capabilities of ChatGPT in the context of underresourced languages and specialized domains like psychotherapy and natural language processing.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e51069) doi:10.2196/51069

KEYWORDS

Cantonese; ChatGPT; counseling; natural language processing; NLP; sentiment analysis

Introduction

Overview

With the rise in mental health demands and challenges, technology-enabled text-based counseling services and therapy are increasingly popular, especially among younger generations [1]. Compared with traditional counseling, web-based services afford higher degrees of accessibility and anonymity [2,3]. In addition to these advantages, web-based, text-based counseling can potentially provide complete electronic records for downstream analyses.

Developing tools to systematically analyze text data is consequently important to enable quantitative evaluation of the service and assist practitioners in making clinical decisions. As a subfield of artificial intelligence (AI) and linguistics, natural language processing (NLP) enables computational approaches to learn, interpret, process, and analyze human language in written or spoken form, making it a promising tool to analyze counseling text [4]. Althoff et al [5] used text features to understand and evaluate counselor performance and counseling outcomes. Broadbent et al [6] developed NLP techniques to identify suicide risks in text-based crisis counseling records. Xu et al [7] constructed a word affinity network to predict suicidal ideation before its emergence in the context of a web-based, text-based counseling service. These studies not only leveraged the nature of electronic counseling records to conduct clinical research but also applied NLP techniques to understand psychotherapy at a scale previously unattainable.

Sentiment analysis is one of the most active research areas in NLP, leading to its rapid development with the help of machine learning models [8]. Sentiment analysis focuses on identifying the sentiments, emotions, or opinions expressed in texts, such as social media posts, conversations, or user reviews. In counseling, sentiment is an important variable. Althoff et al [5] used anonymized counseling sessions from a crisis intervention service to study how various linguistic aspects correlate with session outcomes. After each session, the help-seeker received a follow-up question: “How are you feeling now? Better, same or worse” which was used as the session quality indicator. Linguistic Inquiry and Word Count (LIWC) was used to measure help-seeker sentiments throughout counseling sessions, and the authors found that constructive changes in perspective throughout the session were related to a higher session quality. Furthermore, sentiment analysis has the potential to objectively and automatically evaluate counselors’ expressed empathy, which is key in counseling and mental health support [9]. With its potential use in counseling contexts, there is a pressing need to develop reliable NLP tools for sentiment analysis applicable to such domains.

In recent years, large language models (LLMs), which are pretrained on extremely large amounts of text data and encompass over a hundred billion parameters, have profoundly affected research communities with their impressive general NLP ability in various tasks, including sentiment analysis. ChatGPT stands out as a powerful LLM for accomplishing question-answering tasks. Several studies have explored ChatGPT’s capability of understanding sentiments in different languages. For example, Rathje et al [10] used GPT-3.5 and GPT-4 to conduct 3 sentiment analysis tasks (ie, sentiment, discrete emotion, and offensiveness) across 12 languages. They found that GPT was effective in multilingual text analysis. Wang et al [8] also demonstrated that ChatGPT is effective in sentiment analysis through 5 sentiment analytic tasks in English. Bang et al [11] tested ChatGPT’s strengths and limitations in reasoning and interactivity across 4 languages. They concluded that ChatGPT outperformed other LLMs in most tasks but suffered hallucinations (ie, producing untruthful content that contradicts existing sources or lacks supporting evidence [12]). To date, few studies have evaluated ChatGPT’s capability for analyzing Cantonese texts.

This study seeks to validate ChatGPT as a tool for sentiment analysis in Cantonese. Cantonese is a Chinese language under the Yue branch, with over 73 million speakers worldwide [13]. Its distinctive linguistic features pose unique challenges to NLP, particularly to sentiment analysis.

The first of these linguistic features of Cantonese is that it shares the Chinese ideographic writing system. The apparent similarity between Cantonese and Standard Chinese can lead to high inaccuracy in Cantonese sentiment analysis tasks. For example, the same words in Cantonese and Standard Chinese can have drastically different meanings, for example, 得意 (in Cantonese: cute, and in Standard Chinese: proud).

Second, despite the high similarity between Cantonese and Standard Chinese, Cantonese can differ significantly in grammatical and lexical features from Standard Chinese. For example, grammatical elements that indicate temporal references, certainty, questions, agreements, and so on, can be different between Cantonese and Standard Chinese and can thus lead to misinterpretation. Moreover, there are thousands of characters unique to Cantonese [14]. The sentiments carried by these characters could potentially be neglected. For example, in Cantonese, the specific words that express negation (唔 m4, 冇 mou5) are not found in Standard Chinese, for which LLMs may incorrectly determine a clause as affirmative or negative.

The third linguistic peculiarity is that, compared with Standard Chinese, which is mainly used in formal writing, Cantonese is mostly informal and colloquial. In regions such as Hong Kong, Cantonese, when used in online communication channels, including social media platforms, forums, and online chats, is more informal and colloquial compared with its use in formal writing or speech, which often contains more typos or grammatical peculiarities [15]. The informal and colloquial nature of Cantonese poses significant challenges for NLP to capture the intended sentiments. Moreover, Cantonese has

XSL•FO
RenderX

historically evolved in a multilingual environment, especially Hong Kong Cantonese, which contains many borrowed words from English through phonetic transliteration [13]. Therefore, the challenges in Cantonese sentiment analysis are varied and deeply rooted in the language's linguistic peculiarities. Cantonese is widely used yet distinct from other languages; it is thus a valuable case study.

Many effective and advanced NLP techniques focus on high-resource languages, such as English and Mandarin Chinese. The NLP techniques for these languages are much more developed than those for other languages because of their popularity and the availability of labeled data and expert resources. Under the taxonomy of Joshi et al [16], Cantonese is considered a Tier-1 language, which indicates that it is a low-resource language in the field of NLP. This categorization reflects the lack of labeled data sets and specific NLP resources despite its strong web presence and quantity of users, which suggests the availability of a wealth of unlabeled data. As a result, to date, limited studies have been conducted on Cantonese NLP tasks [17].

To help fill this gap, in this study, ChatGPT's performance on Cantonese sentiment analysis tasks was evaluated and compared with other state-of-the-art models. We asked ChatGPT to perform sentiment analysis on the digital transcripts from a web-based, text-based counseling service in Hong Kong.

Background and Related Works

Cantonese Sentiment Analysis

Sentiment analysis can be categorized into 2 approaches: the lexicon-based approach of building a large corpus of different sentiments and the machine learning approach mixed with the lexicon-based approach. Both methods have their advantages and disadvantages in terms of preparing data for training, inferencing time, and accuracy. The lexicon approach is less used in Cantonese sentiment analysis tasks. So far, only Chen et al [18] used the hidden Markov model to perform word segmentation and then built a sentiment word dictionary for Cantonese. No study was conducted to establish the effectiveness of this model.

In recent years, machine learning approaches, mixed with lexicon-based knowledge, have shown promising results in sentiment detection tasks. Three studies have used supervised learning mixed with lexicon-based methods to classify sentiments in Cantonese text. Ngai et al [17] used a combination of 3 popular classifiers: support vector machine (SVM) with naive Bayes feature-based, convolutional neural network-based, and lexicon-based classifiers. The authors first manually built a sentiment word dictionary using many sources, such as the "Yue Dictionary" [19]. They then trained the combination of the 3 classifiers with sentiment features on massive Cantonese text data from social media sites, including Facebook and Weibo. Subsequently, the model was tested on users' reviews on Weibo, Facebook, and Dianping. The F_1-score of the approach was approximately 72%. Xiang et al [20] explored the deep learning model of long short-term memory (LSTM) combined with an attention mechanism, namely "long short-term memory sentiment augmented attention network" (LSTM-SAT), to classify reviews on OpenRice, a Hong Kong restaurant search website that allows users to post their ratings and reviews. The accuracy of their model was 60.8%. Before modeling, they used an automatic method to expand a manually crafted Cantonese sentiment word dictionary. In addition, pretrained language models (PLMs), which are first pretrained on an extensive corpus using unsupervised learning and then fine-tuned in downstream tasks with a relatively small data set using supervised learning, serve as another way to conduct Cantonese sentiment analysis. Although there is a lack of Cantonese PLMs, Tan et al [21] used the Bidirectional Encoder Representations from Transformers pretrained on a large corpus of simplified Chinese and several simplified Chinese emotional corpora to conduct Cantonese sentiment analysis, in which Cantonese text from an online social forum in Hong Kong was first translated to simplified Chinese. The reported optimal loss value of the test set was 0.065, and the optimal value of the area under the receiver operating characteristic curve was 0.996.

Among the 3 studies applying machine learning mixed with lexicon-based approaches, the first 2 both entailed building a Cantonese sentiment word dictionary and applying existing state-of-the-art machine learning models, while the third study used the latest development of PLMs from simplified Chinese, which is a high-resource language, on Cantonese, a low-resource language [16]. All 3 approaches share some disadvantages in accurately detecting sentiments. The first 2 rely on the construction of the Cantonese sentiment word dictionary, while the third approach is predicated on the quality of translation and the sentiment dictionary. The mixed-based approach yielded higher accuracy by providing extra data preprocessing (eg, Cantonese-simplified Chinese translation) and training-and-testing mechanisms to calibrate the dictionary with the context.

Large Language Models for Sentiment Analysis

ChatGPT is an LLM that is pretrained on extensive amounts of text data. Encompassed with over a hundred billion parameters, LLMs are capable of understanding natural human languages and generating relevant human-like responses. When using LLMs, users' prompts often follow a 2-phase process; they first ask LLMs to conduct certain tasks, and then they provide the data to be evaluated by LLMs (Figure 1).

The LLM approach could alleviate the disadvantages in the existing literature of Cantonese sentiment analysis in subsequent ways. First, during the data preprocessing stage, there is no need to build a Cantonese sentiment word dictionary or translate Cantonese to other high-resource languages. This not only improves efficiency but also avoids potential loss in translation. Second, the LLM requires no training data, eliminating the labor-intensive training phases presented in the existing approaches. Bang et al [11] explored ChatGPT's sentiment analysis ability across 4 languages: English, Indonesian, Javanese, and Buginese. The results show that ChatGPT outperformed the existing state-of-the-art models by a large margin for 3 of the 4 languages (ie, except Buginese). This suggests that ChatGPT possesses a certain degree of understanding of low-resource languages.

XSL•FO
RenderX

To the best of our knowledge, there is no comprehensive evaluation of ChatGPT's ability to conduct sentiment analysis tasks in Cantonese text. In this study, we used a zero-shot prompt asking ChatGPT to perform sentiment analysis on the counseling transcript data from a Hong Kong web-based, text-based counseling service. In our context, zero-shot learning refers to the fact that no additional annotated samples were given to ChatGPT in the prompts; that is, ChatGPT performed the task solely based on our instructions. The findings could provide support for using ChatGPT to facilitate automated Cantonese sentiment analysis, which can have tremendous implications, such as harnessing ChatGPT as a cost-effective labeling and coding tool for both research and applied purposes, monitoring message-level sentiment of both service users and providers in real-time, paving the way for further studies on ChatGPT's ability in other NLP or psychotherapy-related tasks, and transferring our presented methodologies to studies of other low-resource language contexts. Subsequently, the advantages of cutting-edge NLP technology could become accessible to the billions who communicate in those low-resource languages, unlocking immeasurable immediate benefits for them.

Figure 1. An example of a prompt and a response in a large language model (LLM).

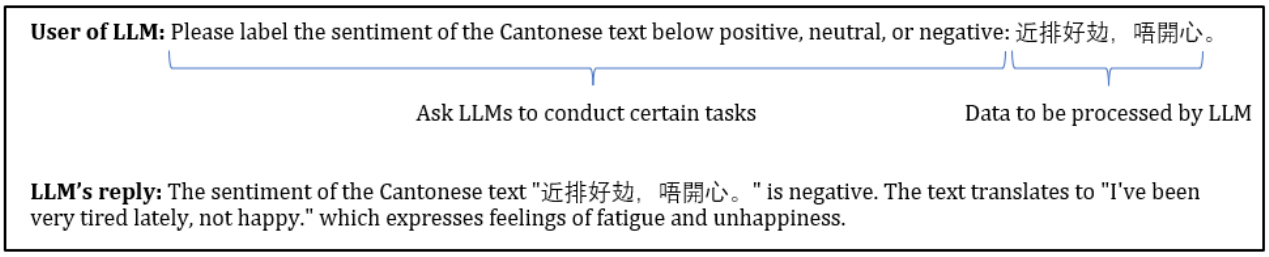

Methods

Ethical Considerations

This study was approved by the Human Research Ethics Committee at the University of Hong Kong (EA210185). Users gave their consent to the use of their text data for research purposes by accepting the privacy policy before the commencement of service. Individual service providers (ie, counselors and volunteers) gave written consent to the use of their text data for research purposes.

Data Source

Open Up is a free 24/7 web-based, text-based counseling service in Hong Kong that enables people aged between 11 and 35 years to anonymously chat with paid staff (staff counselors or social workers) or trained volunteers [19]. In its original planning, the service aimed to provide web-based support to youth aged 24 years or younger. However, it became clear to the service partners that there was a lack of service for young adults aged between 25 and 35 years in the community. The service age range was then broadened.

Open Up counselors possess at least a bachelor's degree in social work or counseling. Over 90% of the staff are registered social workers in Hong Kong. The age of the staff ranges from 25 to 45 years, and the female-male ratio is approximately 2:1. On average, the staff have 8 years of work experience, including 1-2 years related to web-based counseling [22]. Most trained volunteers are undergraduate students studying psychology, counseling, or social work.

Help-seekers can access Open Up anywhere and anytime for free through the service's website, WhatsApp, SMS text messages, or Facebook. Due to the anonymity of the service, the help-seekers' demographic is unknown. Since its launch in 2018, to date (May 2023), more than 224 counselors have engaged with help-seekers across more than 110,000 sessions. Figure 2 shows the opening of a fictitious session between a counselor and a help-seeker.

XSL•FO
RenderX

Figure 2. A fictitious conversation excerpt between a counselor (C) and a help-seeker (H) in Open Up.

C: Welcome to Open Up. Is there anything you want to share with me today?

H: It's just everything. Work, family, life. It all feels like it's too much to handle.

C: It sounds like you're going through a really tough time. It's okay to feel overwhelmed. Have there been any specific incidents that have particularly bothered you?

H: Yes, at work, I always feel ignored, and at home, I'm always the one who has to take care of everything. It's exhausting.

C: **:** It sounds like you're carrying a heavy load both at work and home, and that must be incredibly challenging. It's completely valid to feel exhausted. How do you usually cope when you feel this way?

H: I don't really know. I guess I just try to push through and hope things get better. But it's been really hard to stay hopeful recently.
...

In this study, 342,752 counseling sessions from January 1, 2019, to May 1, 2023, were extracted, which contained 12,056,109 messages between help-seekers and counselors. In Open Up, valid sessions are defined as those with at least 4 message exchanges between the help-seeker and the counselor. This threshold was set based on a previous study [23], which found that some clients logged on to the platform without revealing any intention to use the service; these sessions are typically briefer than 4 message exchanges. Thus, 116,999 sessions with 9,390,973 messages with at least 4 message exchanges between the help-seeker and the counselor were extracted. Among them, 48,717 sessions with 3,878,653 messages in which help-seekers accessed the service for the first time were included in this study to avoid the overrepresentation of sentiments from repeated and frequent help-seekers.

We then filtered out 1342 non-Cantonese sessions that contained a total of 104,918 messages as well as 534,410 system messages (eg, "Please wait, we are connecting you to an Open Up counselor"). The exclusion of system messages was necessary because system messages could still be present in longer sessions. They were labeled as sent from counselors but were actually generated by the system. These system messages are not part of the actual counseling process. Including such messages would introduce noise to the data and potentially confound the results. Therefore, we further excluded such messages. Subsequently, 2565 sessions with fewer than four message exchanges between the help-seeker and counselor were excluded.

After filtering, there were 44,810 valid sessions containing a total of 3,231,830 messages. Only sessions with postsession survey feedback were included in this study. In total, this included 5240 sessions with 533,609 messages (Figure 3). We stratified the 5240 sessions based on the number of messages in each session. There were a total of 131 unique message count groups among the 5240 sessions. The Python package *Random* (version 3.8.2; Python Software Foundation) was used to initialize a random number generator with a fixed seed to select 1 of the sessions in each stratification, from which 131 sessions were randomly selected with 6169 messages to be included in the analyses. We decided to sample 1 session per stratum based on the available human resources for coding.

XSL•FO
RenderX

Figure 3. Data inclusion flow.

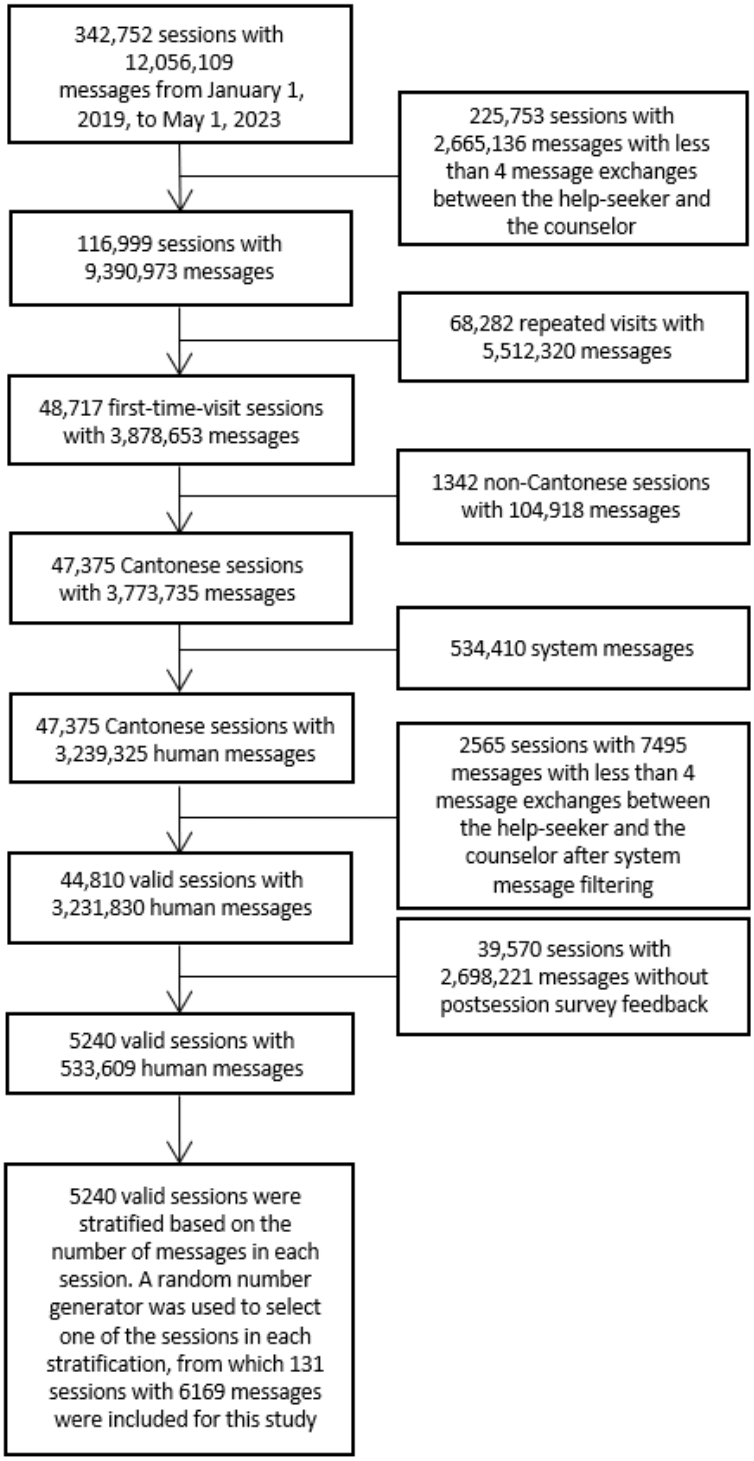

Annotation

A codebook (Multimedia Appendix 1) developed by seasoned social workers and researchers in counseling services was used to ensure the quality of the annotations. The codebook contains rules, guidelines, examples, definitions, and tips for the annotation process. The definitions of the three sentiment categories are briefly described as follows:

1. Positive class: positive sentences describe feelings of pleasure, gratitude, satisfaction, compliment, or recommendation. Emojis and punctuation can be indicative of positivity.
2. Negative class: negative sentences describe feelings of disagreement, sadness, complaint, or hate. Factual information, such as the description of the help-seeker's distressing issues, was also considered negative. Emojis and punctuation can also be indicative of negativity.
3. Neutral class: the neutral label is used for any sentences the annotator could not identify as positive or negative. Factual sentences that do not present sentiment orientation, such as "I am 25 years old," and sentences the annotator cannot completely comprehend, were annotated as neutral.

XSL•FO
RenderX

Sentences that are routine replies from the counselor, such as "Is there anything you want to share with me today?" were also labeled as neutral.

A total of 3 annotators labeled 500 messages from the data set to examine whether the guideline was clear. The κ agreement of the 3 annotators was 0.95, indicating a high level of agreement, and the guideline was sufficiently clear. In cases where the 3 annotators disagreed on the labeling among the 500 messages, the final label was assigned based on the majority rule among the 3 annotators. Then, one of the annotators labeled the rest of the messages, following the guidelines. Most messages with 1 or multiple sentences expressed 1 consistent sentiment. Only 35 messages out of 6169 messages (0.6%) contained a mixture of sentences expressing different sentiments. The ambiguity of the overall sentiment of such messages was discussed among the coders, and a set of rules was developed to annotate such messages' sentiment (Multimedia Appendix 1). The count and proportion of messages categorized into the 3 sentiment classes by human annotation are as follows: positive sentiment in 15.5% (959/6169), negative sentiment in 31.5% (1943/6169), and neutral in 53% (3267/6169) messages.

Usage of ChatGPT

Two of the most popular ChatGPT variants, GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4), were used to label each message. We implemented both models with zero-shot learning. The prompt was set as "Is the sentiment of this Cantonese text positive, neutral, or negative?" The temperature of the ChatGPT models was set at 0 to generate consistent answers with the least creativity. The hyperparameter setting is shown in Table 1.

Table 1. Prompt settings used.

Model	GPT-3.5	GPT-4
Prompt	"Is the sentiment of this Cantonese text positive, neutral, or negative? Respond with the sentiment label only."	"Is the sentiment of this Cantonese text positive, neutral, or negative? Respond with the sentiment label only."
max_token	4096	8192
Engine	gpt-3.5-turbo	gpt-4
Temperature	0.0	0.0

Baseline Details

We compared ChatGPT with a lexicon-based method and 3 machine learning models: LIWC, logistic regression (LR), SVM, and LSTM. The general procedure for training a sentiment analysis model is data preprocessing, word vector training, training and test set splitting, hyperparameter setting, model training, and evaluation.

First, according to the general framework above, we started with data pre-processing. Unlike English, there is no obvious boundary between words in Cantonese text. For example, "I feel sad" in Cantonese looks like "Ifeelsad." Therefore, word segmentation is the first step in data preprocessing. Jieba [24], a cutting-edge word segmentation tool, was used to process each message. Next, as there are trivial words that do not carry much meaning and may introduce noise in the analyses (eg, "呢個" meaning "this" and "跟住" meaning "then"), we removed these words using a stopwords dictionary tailor-made for Open Up data. After word filtering, a corpus of 7,012,611 messages was compiled, composed of a vocabulary of 70,951 unique words, which were then used as inputs in the word vector training process.

Second, to convert text to a machine-accessible form, word vector training is necessary. Word2Vec [25] is a popular model that can learn semantic knowledge from a large corpus without supervision and project words into a vector space in which the space distance among words represents their semantic relationships. Following previous studies using Open Up data to train the Word2vec model [7], the Python package *Gensim* (version 4.3.2; Python Software Foundation) was used for Word2vec processing. In particular, a Word2vec model with 5 as the window size and 100 as the vector size using Continuous Bag of Words was trained (Figure 4). The input was the 7,012,611 messages after word filtering.

Third, 80% of the data was randomly selected for training, and the remaining 20% was the test set.

Fourth, the baseline methods were selected for comparison with ChatGPT in terms of the following four criteria:

1. "Interpretability" refers to how easily one can understand and explain the outcome of a model. An interpretable model provides a transparent decision-making process, making it more trustworthy and easier to update.
2. "Efficiency" refers to the speed and computational resources required for training and deploying a model. An efficient model requires less time and computing power, making it easier to get the result with limited resources.
3. "Robustness" indicates the model's resistance against overfitting, which means a model performs quite well on the training data but poorly on the new or test data. A robust model ensures consistent performance on new, unseen data, making it more reliable.
4. "Contextual memory over sequential data" emphasizes a model's ability to process sequential data where the order and sequence are significant. A model that can retain contextual memory over sequential data means it reaches the final decision by considering the preceding and succeeding contexts of each entity and memorizing long-term dependencies over sequential data instead of treating input data as independent entities.

Table 2 contains description and advantages of the baseline methods.

The implementation details are provided in the subsequent paragraphs.

XSL•FO
RenderX

For LIWC, 2015 LIWC for the traditional Chinese (version 1.5; Pennebaker Conglomerates), which is a computerized text analysis method developed to examine the meaning of words, was used. It categorizes words into different linguistic and psychological dimensions based on a comprehensive dictionary.

For LR and SVM, the mean of word embeddings was used to generate message representations. The Python package *Scikit-learn* (version 1.3.0; Python Software Foundation) was imported to run logistic regression and SVM model, respectively. For SVM, a radial basis function kernel was used.

For LSTM, the mean of word embeddings was used to generate message representation. A detailed model structure can be found in the Multimedia Appendix 1. For the hyperparameter setting, the LSTM model was run multiple times with different learning rates, dropout rates, and batch sizes, and the optimal combination (ie, learning rate 0.005, dropout rate 0.5, and batch size 64) was selected, which caused the least training loss and overfitting problem. The model was trained using the categorical cross-entropy loss function with the Adam optimizer in 10 epochs, which were the best selections obtained through multiple training processes.

Figure 4. Distribution of Word2vec vectors among 30 sample words. To show the trained word2vec vectors, principal component analysis was used to reduce word2vec dimensions from 100-dimension to 2-dimension. The sampled 30 words were randomly selected to illustrate the performance of the trained Word2Vec model.

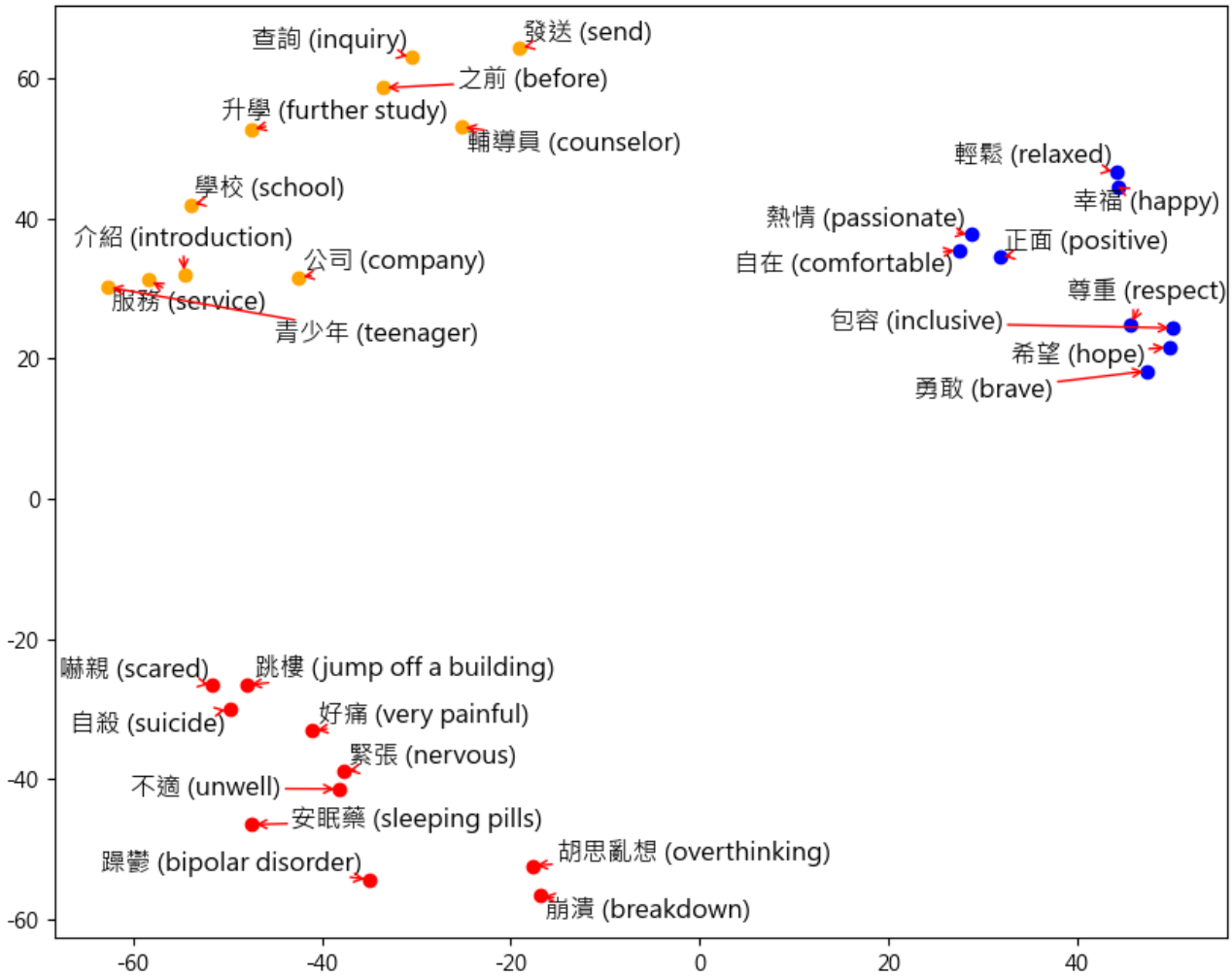

XSL•FO
RenderX

Table 2. Description and advantages of the 4 baseline methods.

Model	Description	Interpretability	Efficiency	Robustness	Contextual memory over sequential data
LIWC[a]	A text analysis program that counts words in psychologically meaningful categories	Categorizes words into psychological meaningful categories, making analysis interpretable and transparent	Allows for fast analysis of text and requires much less computational resource compared to other machine learning methods	Does not evolve over time, making it perform badly on new, unseen words	Treats each word as independent entity
LR[b]	A statistical method that predicts the likelihood of an outcome	Mathematically simple and inherently interpretable	More efficient compared to SVM[c] and LSTM[d]	Less likely to overfit when appropriately regularized	Treats each word as independent entity
SVM	A type of algorithm that identifies patterns by drawing the best boundary through a "hyperplane" that separates different groups of data points	Less interpretable than LIWC and LR while more interpretable than LSTM	Less efficient than LIWC and LR while more efficient than LSTM	With proper class weights settings, SVMs can handle scenarios where data are imbalanced in different classes, which is robust to overfitting	Treats each word as independent entity
LSTM	A kind of recurrent neural network that can remember information for long periods through gated mechanism	Least interpretable compared to the other 3 models	Least efficient compared to the other three models	Designed specifically for vanishing gradient problem, ensuring its robustness against overfitting	Encompassed with gated mechanism, LSTM can retain memory from previous words in a sequence with variable length

[a]LIWC: Linguistic Inquiry and Word Count.

[b]LR: logistic regression.

[c]SVM: support vector machine.

[d]LSTM: long short-term memory.

Results

GPT-3.5 Versus GPT-4

We used 2 metrics, accuracy and F_1-score, to evaluate the models' performance (performance evaluation metrics in Multimedia Appendix 1). Accuracy represents the percentage of sentiment labels that the model correctly predicted out of all the sentiment labels it attempted to predict. However, this metric is biased toward sentiment labels with more data points (eg, if there are 90 neutral messages and 10 positive ones, a model that predicts all messages as neutral would have an accuracy of 90%). To measure performance in a way that is less sensitive to data imbalances, the F_1-score, which represents the harmonic mean of the model's precision and recall, was used to ensure that a model performs well across all 3 sentiment categories, not just the predominant one.

The results show that GPT-3.5 achieved 92.1% (5682/6169) accuracy and a 90.2% F_1-score, and GPT-4 achieved 95.3% (5880/6169) accuracy and a 94% F_1-score. Confusion metrics were used to better illustrate the results, in which rows represent the true sentiments annotated by human coders (true labels) and columns represent the sentiments predicted by the model (predicted labels). Figures 5A and 5B present the confusion metrics of GPT-3.5 and GPT-4. Darker colors reflect higher proportions of messages annotated by human coders with a particular label that received the same label from GPT, and vice versa for lighter colors. Figure 5A suggests that GPT-3.5 was more likely to classify "neutral" messages as "negative" (117/1943, 6% of neutral messages). In contrast, GPT-4 showed higher accuracy in detecting all 3 categories of sentiments. Both GPT-3.5 and GPT-4 outperformed other methods, and the McNemar test showed that GPT-3.5 and GPT-4 did not show perfect agreement (P=.01), which indicated that the results from GPT-4 were statistically significantly better than those from GPT-3.5.

XSL•FO
RenderX

Figure 5. Confusion matrices for GPT-labeled sentiments (predicted labels) versus human-labeled sentiments (true labels). (A) The confusion matrix for GPT-3.5, and (B) the confusion matrix for GPT-4.

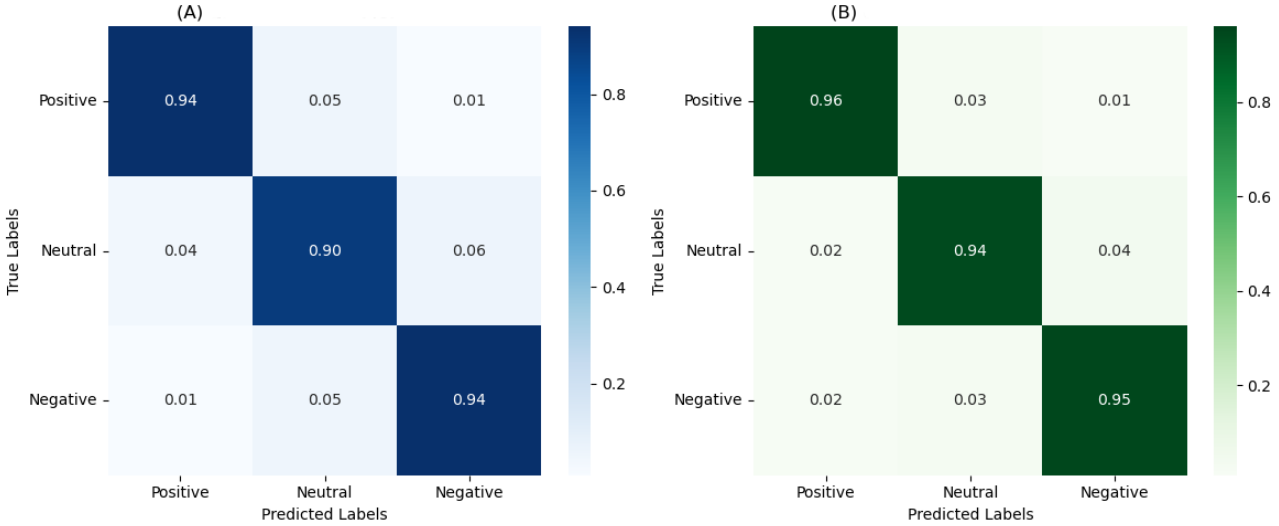

ChatGPT Versus Baseline Models

Table 3 reports the comparison between ChatGPT and the baseline models. LIWC had the lowest accuracy of 37.2% (2295/6169) and an F_1-score of 24.2%, indicating its struggles in the sentiment classification task. Both the LR and SVM models showed accuracies of around 66%, but there was a noticeable difference in their F_1-scores, with the LR model having a much higher score. This suggested that while SVM was almost as accurate as the LR model, it did not handle our imbalanced data as well as the latter. In addition, as LSTM can retain contextual memory over sequential data, it outperformed the traditional models with an accuracy of 70.9% (4374/6169) and an F_1-score of 69.5%, indicating its better and more balanced performance at handling Cantonese sentiment classification among all the non-GPT baseline methods. Moreover, the results show that GPT-3.5, with a classification accuracy of 92.1% (5682/6169), and GPT-4, with an accuracy of 95.3% (5880/6169), outperformed the other state-of-the-art models for Cantonese sentiment classification tasks, with accuracies ranging from 37.2% (2295/6169) to 70.9% (4374/6169). Meanwhile, GPT-3.5 had an F_1-score of 90.2% and GPT-4 reached an F_1-score of 94%, indicating their balanced overall performance across the 3 sentiment categories, making them relatively more reliable.

Table 3. Comparison of ChatGPT and other state-of-art models in Cantonese sentiment analysis.

Model	Accuracy, n (%)	F_1-score (%)
LIWC[a]	2295 (37.2)	24.2
LR[b]	4072 (66.0)	57.3
SVM[c]	4047 (65.6)	49.7
LSTM[d]	4374 (70.9)	69.5
GPT-3.5	5682 (92.1)	90.2
GPT-4	5880 (95.3)	94

[a]LIWC: Linguistic Inquiry and Word Count.
[b]LR: logistic regression.
[c]SVM: support vector machine.
[d]LSTM: long short-term memory.

Discussion

Overview

In this study, we tested whether the popular LLM ChatGPT could serve as a reliable sentiment classifier for Cantonese text. In a data set containing Cantonese text-based counseling session transcripts, we found that the 2 most recent variants of ChatGPT (GPT-3.5 and GPT-4) could accurately (>90%) detect sentiments expressed in Cantonese text. ChatGPT performed much better than lexicon-based methods and fine-tuned machine learning models at sentiment analysis. With a predefined list of words associated with specific categories, such as sentiments, lexicon-based text analysis methods are user-friendly and easy to interpret. However, these methods often achieve lower accuracy since they ignore the context in which words are used, which may lead to incorrect categorization of the text. Furthermore, the dictionaries are predefined and do not evolve

XSL•FO
RenderX

over time, so they may lack the knowledge base needed for specific tasks. Compared with lexicon-based methods, ChatGPT is also user-friendly and much more accurate. With the rapid development of ChatGPT, in the future, there may be little need for lexicon-based methods such as LIWC. Moreover, ChatGPT is also a better choice than fine-tuned machine learning models. While machine learning classifiers require large amounts of manually annotated text to train and high coding proficiency, ChatGPT does not require extensive training data, and it is intuitive to use with little coding knowledge needed.

Cantonese sentiment analysis is particularly challenging for several reasons. First, Cantonese shares a writing system with Standard Chinese. The similarity can lead to misinterpretations of sentiments based on Standard Chinese sentiment analysis tools. Second, Cantonese has its own characters, expressions, and grammatical elements that may be overlooked by sentiment analysis models. Third, colloquialism and multilingualism in Cantonese pose great challenges to understanding sentiments in Cantonese text. Fourth, Cantonese has fewer NLP resources, such as labeled data sets or task-specific tools, making it difficult for robust NLP applications.

ChatGPT can tackle these challenges posed by Cantonese linguistic peculiarities. In our observations, ChatGPT can identify the sentiment of the same word with different meanings in Cantonese and Mandarin Chinese, which avoids the potential misinterpretation using Standard Chinese-specific models in the Cantonese context, which is a problem found in previous studies (eg, Xiang et al [20]). For example, “窩心,” meaning “feeling warm and caring” in Cantonese, was correctly classified as positive by ChatGPT, while other Standard Chinese-based machine learning models would label the word as negative due to its meaning “feeling uncomfortable and annoyed” in Mandarin Chinese. ChatGPT can also handle multilingual content accurately. For instance, the negative sentiment of “今日好down” (“feeling very down today”) with an English word “down” in a Cantonese sentence expressing feeling down or unhappy was accurately captured by ChatGPT. These mixed-language cases would be quite challenging for other traditional models. However, there might be some potential biases in using ChatGPT in Cantonese sentiment analysis. For example, GPT-3.5 at times interpreted the negative sentiment of being lost and not knowing how to address a problem as neutral. Moreover, GPT-3.5 tended to classify positive sentiments as neutral when a message expressed gratitude. We speculated that GPT-3.5 might have a linguistic bias in the Cantonese context; however, comparing the degree of bias against English was beyond the scope of this study, given that we had no comparable data in English for comparison. Moreover, ChatGPT has been shown to be biased in terms of gender and representation of minorities in the English context [26]. It is plausible that this bias might be exhibited in Cantonese as well.

The promising results of this study have 4 practical implications. First, it can assist in the quality control of services. By using ChatGPT for sentiment analysis in Cantonese, this study demonstrated the feasibility of using advanced LLMs to automatically classify sentiments in text-based counseling sessions. The accurate identification of sentiment can be used to assess the effectiveness of counseling interventions, improve good practice models, and enhance the overall quality of support offered to help-seekers. For example, as sentiment analysis has the potential to evaluate the counselor’s empathetic expressions, which were found to be predictive of positive counseling outcomes [9], ChatGPT could potentially help measure empathetic responses and affective synchronicity.

The second benefit is the accurate identification of emotional concerns and needs expressed by help-seekers during counseling sessions. By analyzing the sentiments conveyed in the text, counselors and service providers can gain deeper insights into the psychological well-being of their clients. This information can assist in detecting potential high-risk situations or individuals who may require timely attention or specialized care. For example, as the help-seeker’s suicide risk is closely correlated with their expressed sentiments [27], using ChatGPT to automatically analyze sentiments could potentially detect suicidal ideation at early stages and facilitate the counselor’s intervention accordingly.

In the long term, the implications of this study extend beyond the specific task of sentiment analysis in Cantonese text-based counseling services. The successful application of ChatGPT opens possibilities for using similar LLMs to tackle other tasks in the field of psychotherapy or NLP. These tasks may include more nuanced sentiment analysis, such as identifying discrete emotions to detect more specific expressed emotions (eg, happiness, anger, sadness, and fear). To achieve this goal, for example, an emotion-specific data set together with a codebook would be developed by Cantonese-speaking counselors and psychologists, where each piece of Cantonese text in the data set conveys one clear emotion to avoid ambiguity. The robustness and reliability of ChatGPT demonstrated in this study lay the foundation for further exploration and expansion of its capabilities in sentiment analysis and potentially other NLP-related tasks.

Finally, the success of Cantonese sentiment analysis hints at ground-breaking potential for tackling NLP tasks in other low-resource languages. This study’s approach and findings can be potentially applied to other low-resource language contexts due to the subsequent reasons. First, all low-resource languages face some common challenges, such as the lack of labeled data sets with domain knowledge and the lack of advanced NLP models. Second, ChatGPT was trained on various language resources, and studies such as Rathje et al [10] and Bang et al [11] suggested ChatGPT can, to a certain extent, understand low-resource languages. Thus, with adjustments to account for linguistic and cultural differences, our methodology is transferable and scalable to other low-resource languages. More generally, this study could help shape current approaches to NLP tasks in low-resource languages. This innovation might bring the cutting-edge benefits of NLP technology closer to billions of people worldwide who speak these languages, thus bringing about immense immediate benefits.

However, since counseling sessions are inherently sensitive and private, using LLMs like ChatGPT to analyze sentiment in a counseling context raises several ethical considerations. First, data privacy is of major concern. The use of LLMs to analyze

XSL•FO
RenderX

counseling session data must ensure anonymity and confidentiality and accord with ethics approval. Second, the decision-making process of LLMs is not fully interpretable, especially since we do not know what materials were provided during the training stage, which poses challenges in justifying their outputs. Third, LLMs might generate biased statements or unsafe advice, which should receive more attention, particularly in sensitive settings like counseling. Addressing these ethical considerations enables us to advance our research responsibly.

Limitations and Future Directions

While our findings are promising, this study has a number of limitations. First, although our annotation codebook was comprehensive, the labeled data was imbalanced. With only 959 out of 6169 cases (15.5%) belonging to "positive" sentiment, we may lack a comprehensive understanding of the positive affect expressed in counseling sessions. In the future, sessions with more messages expressing positive sentiment will be included to ensure data balance. Second, the generalizability of our findings is limited by the data source. The way of expressing sentiment in different domains, for example, social media and customer feedback, could differ. As this study focused on counseling text, future research could explore various text sources to enhance the generalizability of this study. Third, we observed that there were still some biases and limitations to using ChatGPT in Cantonese sentiment analysis.

Future work can consider several directions. First, researchers can consider using different prompting strategies, such as few-shot learning (ie, providing several labeled examples in the prompts), to further enhance the performance of ChatGPT in Cantonese sentiment analysis tasks. Few-shot learning provides additional references and materials on top of the training data set for GPT to know the context and make better predictions. The second possible direction is to use ChatGPT in more types of sentiment analysis tasks, such as discrete emotion detection, to detect more specific expressed emotions (eg, happiness, anger, sadness, and fear). The third direction future work can consider is to go beyond sentiment analysis tasks and to use ChatGPT in other NLP-related tasks, such as text summarization tasks (ie, to generate concise summaries of long texts such as counseling session transcripts) and translation tasks (ie, translating languages while taking into consideration linguistic and cultural differences).

Conclusions

We evaluated ChatGPT's performance in analyzing sentiments in Cantonese web-based counseling transcripts and compared its accuracy against other mainstream methods. Our results suggest that not only is ChatGPT a reliable tool for detecting sentiments in Cantonese text, but it has also surpassed the accuracy of existing sentiment analysis methods, such as lexicon-based methods and fine-tuned machine learning models, with many additional advantages, such as the use of straightforward and intuitive prompts. The findings could have many practical implications, including improving the quality of counseling services, providing timely support for help-seekers, using LLMs to tackle other psychology or NLP-related tasks, and bringing cutting-edge NLP tools to low-resource languages.

Acknowledgments

This study was funded by the Hong Kong Jockey Club Charities Trust (2021-0174). CSC (shaunlyn@hku.hk) and PSFY are cocorrespodning authors for this manuscript.

Data Availability

Given its sensitivity, the raw transcript data are not publicly available but are available from the corresponding author on reasonable request. The code used to perform the analysis can be found on GitHub [28].

Authors' Contributions

ZF was responsible for the conceptualization and software application. ZF and YCH conducted the formal analysis. CSC, CML, JL, and PSFY arranged the resources for this study. ZF and YCH wrote the original draft. ZF, YCH, CSC, CML, JL, and PSFY reviewed and edited the manuscript. This study was supervised by CSC and PSFY.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 128 KB - jmir_v26i1e51069_app1.docx]

References

1. Dowling M, Rickwood D. Investigating individual online synchronous chat counselling processes and treatment outcomes for young people. Adv Ment Health 2015;12(3):216-224 [FREE Full text] [doi: 10.1080/18374905.2014.11081899]

XSL•FO
RenderX

2. Amos PM, Bedu-Addo PKA, Antwi T. Experiences of online counseling among undergraduates in some Ghanaian universities. SAGE Open 2020;10(3):215824402094184 [FREE Full text] [doi: 10.1177/2158244020941844]
3. Fang L, Tarshis S, McInroy L, Mishna F. Undergraduate student experiences with text-based online counselling. Br J Soc Work 2018;48(06):1774-1790 [FREE Full text] [doi: 10.1093/bjsw/bcx111]
4. Chowdhury GG. Natural language processing. Annu Rev Inf Sci Technol 2005;37(1):51-89 [FREE Full text] [doi: 10.1002/aris.1440370103]
5. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans Assoc Comput Linguist 2016;4:463-476 [FREE Full text] [Medline: 28344978]
6. Broadbent M, Grespan MM, Axford K, Zhang X, Srikumar V, Kious B, et al. A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. Front Psychiatry 2023;14:1110527 [FREE Full text] [doi: 10.3389/fpsyt.2023.1110527] [Medline: 37032952]
7. Xu Z, Chan CS, Zhang Q, Xu Y, He L, Cheung F, et al. Network-based prediction of the disclosure of ideation about self-harm and suicide in online counseling sessions. Commun Med (Lond) 2022;2(1):156 [FREE Full text] [doi: 10.1038/s43856-022-00222-4] [Medline: 36474010]
8. Wang Z, Xie Q, Ding Z, Feng Y, Xia R. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv Preprint posted online on April 10 2023. [FREE Full text]
9. Nakayama M, Hatanakaka C, Konakawa H, Suzuki Y, Koh A, Sugihara Y, et al. Japanese dictionary for sentiment analysis of counselling text. 2021 Presented at: HAI '21: International Conference on Human-Agent Interaction; November 9-11, 2021; Virtual Event Japan p. 311-315 URL: https://dl.acm.org/doi/abs/10.1145/3472307.3484663 [doi: 10.1145/3472307.3484663]
10. Rathje S, Mirea DM, Sucholutsky I, Marjieh R, Robertson C, Van Bavel JJ. GPT is an effective tool for multilingual psychological text analysis. PsyArXiv Preprint posted online on July 17, 2023. [FREE Full text] [doi: 10.31234/osf.io/sekf5]
11. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv Preprint posted online on February 8, 2023. [FREE Full text] [Medline: 10]
12. Yang L, Chen H, Li Z, Ding X, Wu X. ChatGPT is not enough: enhancing large language models with knowledge graphs for fact-aware language modeling. arXiv Preprint posted online on June 20, 2023. [FREE Full text]
13. Xiang R, Tan H, Li J, Wan M, Wong KF. When cantonese NLP meets pre-training: progress and challenges. 2022 Presented at: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts; November 2022; Taipei p. 16-21 URL: https://aclanthology.org/2022.aacl-tutorials.3/
14. Kwan-hin C, Bauer RS. The representation of cantonese with Chinese characters. J Chin Linguist Monogr Ser.: Project on Linguistic Analysis, University of California; 2002. URL: https://www.jstor.org/stable/23826053 [accessed 2023-07-01]
15. Wong T, Lee JSY. Register-sensitive translation: a case study of mandarin and cantonese. 2018 Presented at: Proceedings of AMTA 2018, vol. 1: MT Research Track; March 17-21, 2018; Boston p. 89-96 URL: https://aclanthology.org/W18-1809.pdf
16. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. The state and fate of linguistic diversity and inclusion in the NLP World. : Association for Computational Linguistics; 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; July 5–10, 2020; Online p. 6282-6293. [doi: 10.18653/v1/2020.acl-main.560]
17. Ngai EWT, Lee MCM, Choi YS, Chai PFY. Multiple-domain sentiment classification for cantonese using a combined approach. 2018 Presented at: Twenty-Second Pacific Asia Conference on Information Systems; June 26-30, 2018; Japan URL: https://core.ac.uk/download/pdf/301375967.pdf
18. Chen J, Liu Y, Zhang G, Cai Y, Wang T, Min H. Sentiment analysis for cantonese opinion mining. 2013 Presented at: 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies; September 9-11, 2013; Xi'an, Shaanxi, China p. 496-500. [doi: 10.1109/eidwt.2013.89]
19. Lau C, Chan GW, Tse RK, Chan LS. Words.hk: a comprehensive cantonese dictionary dataset with definitions, translations and transliterated examples. 2022 Presented at: Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference; June 24, 2022; Marseille, France p. 53-62 URL: https://aclanthology.org/2022.dclrl-1.7
20. Xiang R, Jiao Y, Lu Q. Sentiment augmented attention network for cantonese restaurant review analysis. 2019 Presented at: Proceedings of the 8th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM); August 4, 2019; Anchorage, Alaska p. 1-9.
21. Tan X, Zhuang M, Lu X, Mao T. An analysis of the emotional evolution of large-scale internet public opinion events based on the BERT-LDA hybrid model. IEEE Access 2021;9:15860-15871 [FREE Full text] [doi: 10.1109/access.2021.3052566]
22. Xu Y, Chan CS, Tsang C, Cheung F, Chan E, Yip PSF. Evaluating the effectiveness of concurrent sessions and counselors' attention allocation in online counseling. J Consult Clin Psychol 2023;91(11):640-651. [doi: 10.1037/ccp0000831] [Medline: 37616124]

XSL•FO
RenderX

23. Yip PSF, Chan WL, Chan CS, He L, Xu Y, Chan E, et al. The opportunities and challenges of the first three years of open up, an online text-based counselling service for youth and young adults. Int J Environ Res Public Health 2021;18(24):13194 [FREE Full text] [doi: 10.3390/ijerph182413194] [Medline: 34948802]
24. Sun J. Jieba. GitHub. 2021. URL: https://github.com/fxsjy/jieba [accessed 2024-01-04]
25. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv Preprint posted online on October 16, 2013. [FREE Full text]
26. Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. : Association for Computational Linguistics; 2021 Presented at: Proceedings of the Third Workshop on Narrative Understanding; June 11, 2021; Virtual p. 48-55. [doi: 10.18653/v1/2021.nuse-1.5]
27. Aldhyani THH, Alsubari SN, Alshebami AS, Alkahtani H, Ahmed ZAT. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. Int J Environ Res Public Health 2022;19(19):12635 [FREE Full text] [doi: 10.3390/ijerph191912635] [Medline: 36231935]
28. gpt-cantonese-sentiment. GitHub. URL: https://github.com/iharrisonfu/gpt-cantonese-sentiment [accessed 2024-01-16]

Abbreviations

AI: artificial intelligence
LIWC: Linguistic Inquiry and Word Count
LLM: large language model
LR: logistic regression
LSTM: long short-term memory
LSTM-SAT: long short-term memory sentiment augmented attention network
NLP: natural language processing
PLM: pretrained language model
SVM: support vector machine

Edited by A Castonguay; submitted 19.07.23; peer-reviewed by J Liu, M Chatzimina; comments to author 03.10.23; revised version received 24.10.23; accepted 11.12.23; published 30.01.24.

Please cite as:
Fu Z, Hsu YC, Chan CS, Lau CM, Liu J, Yip PSF
Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study
J Med Internet Res 2024;26:e51069
URL: https://www.jmir.org/2024/1/e51069
doi:10.2196/51069
PMID:38289662

©Ziru Fu, Yu Cheng Hsu, Christian S Chan, Chaak Ming Lau, Joyce Liu, Paul Siu Fai Yip. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality and Dependability of ChatGPT and DingXiangYuan Forums for Remote Orthopedic Consultations: Comparative Analysis

Zhaowen Xue[1*], PhD; Yiming Zhang[1*], PhD; Wenyi Gan[1*], PhD; Huajun Wang[1*], PhD; Guorong She[1*], PhD; Xiaofei Zheng[1], PhD

Department of Bone and Joint Surgery and Sports Medicine Center, The First Affiliated Hospital, The First Affiliated Hospital of Jinan University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Xiaofei Zheng, PhD
Department of Bone and Joint Surgery and Sports Medicine Center, The First Affiliated Hospital, The First Affiliated Hospital of Jinan University
No. 613, Huangpu Avenue West
Tianhe District
Guangzhou, 510630
China
Phone: 86 13076855735
Email: zhengxiaofei12@163.com

Abstract

Background: The widespread use of artificial intelligence, such as ChatGPT (OpenAI), is transforming sectors, including health care, while separate advancements of the internet have enabled platforms such as China's DingXiangYuan to offer remote medical services.

Objective: This study evaluates ChatGPT-4's responses against those of professional health care providers in telemedicine, assessing artificial intelligence's capability to support the surge in remote medical consultations and its impact on health care delivery.

Methods: We sourced remote orthopedic consultations from "Doctor DingXiang," with responses from its certified physicians as the control and ChatGPT's responses as the experimental group. In all, 3 blindfolded, experienced orthopedic surgeons assessed responses against 7 criteria: "logical reasoning," "internal information," "external information," "guiding function," "therapeutic effect," "medical knowledge popularization education," and "overall satisfaction." We used Fleiss κ to measure agreement among multiple raters.

Results: Initially, consultation records for a cumulative count of 8 maladies (equivalent to 800 cases) were gathered. We ultimately included 73 consultation records by May 2023, following primary and rescreening, in which no communication records containing private information, images, or voice messages were transmitted. After statistical scoring, we discovered that ChatGPT's "internal information" score (mean 4.61, SD 0.52 points vs mean 4.66, SD 0.49 points; P=.43) and "therapeutic effect" score (mean 4.43, SD 0.75 points vs mean 4.55, SD 0.62 points; P=.32) were lower than those of the control group, but the differences were not statistically significant. ChatGPT showed better performance with a higher "logical reasoning" score (mean 4.81, SD 0.36 points vs mean 4.75, SD 0.39 points; P=.38), "external information" score (mean 4.06, SD 0.72 points vs mean 3.92, SD 0.77 points; P=.25), and "guiding function" score (mean 4.73, SD 0.51 points vs mean 4.72, SD 0.54 points; P=.96), although the differences were not statistically significant. Meanwhile, the "medical knowledge popularization education" score of ChatGPT was better than that of the control group (mean 4.49, SD 0.67 points vs mean 3.87, SD 1.01 points; P<.001), and the difference was statistically significant. In terms of "overall satisfaction," the difference was not statistically significant between the groups (mean 8.35, SD 1.38 points vs mean 8.37, SD 1.24 points; P=.92). According to how Fleiss κ values were interpreted, 6 of the control group's score points were classified as displaying "fair agreement" (P<.001), and 1 was classified as showing "substantial agreement" (P<.001). In the experimental group, 3 points were classified as indicating "fair agreement," while 4 suggested "moderate agreement" (P<.001).

XSL•FO
RenderX

Conclusions: ChatGPT-4 matches the expertise found in DingXiangYuan forums' paid consultations, excelling particularly in scientific education. It presents a promising alternative for remote health advice. For health care professionals, it could act as an aid in patient education, while patients may use it as a convenient tool for health inquiries.

(J Med Internet Res 2024;26:e50882) doi:10.2196/50882

KEYWORDS

artificial intelligence; ChatGPT; consultation; musculoskeletal; natural language processing; remote medical consultation; orthopaedic; orthopaedics

Introduction

The fast growth of artificial intelligence (AI) in recent years has brought tremendous changes to different professions and businesses, altering the way people live and work. The application of AI in medicine is expanding in several areas, including medical image analysis, medication-interaction detection, the identification of high-risk patients, and medical record coding [1,2]. As technology advances, OpenAI introduced ChatGPT on November 30, 2022, as a new kind of natural language model capable of communicating with people through text-to-text, human-like dialogues [3,4]. The more powerful GPT-4 subsequently became accessible through a paid ChatGPT Plus membership on March 13, 2023. It has attracted a lot of interest since its release and has the potential to be widely used in the health care system [5,6]. Most medical AI research has targeted medical workers as software users, which requires medical knowledge reserves [7]. ChatGPT and other conversation question-and-answer AI software programs do not establish a user threshold, and their strong function makes them an essential auxiliary tool to increase finance and management job efficiency [8]. Health is a natural component of humans and should be explored and used in ChatGPT, particularly in the context of situational conversations between patients and physicians.

As human civilization advances, the quest for more convenient, professional, and precise medical services intensifies, with patients expecting increasingly high standards of care. The internet era has spurred hospitals to offer remote diagnostic and treatment services, facilitating doctor-patient interactions beyond physical boundaries and enhancing an understanding of medical issues through remote health care, particularly for those far from medical centers [1]. The recent COVID-19 pandemic has accelerated this digital shift in medicine [2,9,10]. However, the complexity of medical information can reduce physician efficiency and patient comprehension, highlighting the need for patient navigation services, especially in countries with evolving medical systems such as China [11-13]. Amid this backdrop, the rapid advancement of AI technologies such as ChatGPT offers promising support in navigating medical systems, aiding patients in understanding their disease, and selecting a health care facility [14].

"DingXiangYuan" is a leading digital health technology enterprise in China that seeks to unite physicians, researchers, patients, and hospitals through expert and authoritative knowledge exchange, extensive and thorough medical data collection, and top-notch digital medical services [15]. Its remote diagnosis and treatment application has been widely used in China. In the application forums, users may seek the assistance of physicians who are qualified and accredited by the site. At the same time, the information provided by doctors is public, and supervision by the platform leads to a high level of quality for the questions and answers listed in these forums. However, consultations on DingXiangYuan are costly and restrict the number of conversations patients can have with their physicians. In addition, websites offering remote consultations, such as DingXiangYuan, still require physicians to respond on the web, which does not reduce the burden on clinicians.

Nevertheless, a comparative analysis of the quality of responses obtained from paid remote health consultations and ChatGPT-4 has yet to occur. This analysis was based on 82 orthopedic surgery–related consultations sourced from the Doctor DingXiang section of the DingXiangYuan platform. Responses from physicians on the web served as the control group, while those from ChatGPT-4 made up the experimental group. To determine the efficacy of ChatGPT-4 as a reliable remote health consultation resource, we conducted a comparative analysis of its logical response structure, diagnostic accuracy, the viability of its treatment recommendations, and the ability to effectively disseminate medical knowledge pertaining to various conditions. The goal is to provide a workable foundation for the development of ChatGPT-4 in the medical domain.

Methods

Data Set of Orthopedic-Related Remote Consultation

The "Doctor DingXiang" website is a remote network that houses a collection of orthopedic-related medical dialogues and is one of China's largest remote-paid consultation platforms (Figure 1A and Figure 2). To protect patients, the website blocks access to all content that may compromise their privacy, including the patient's username, images provided in the question, imaging data, and biochemical examination results, from all other website visitors, allowing only the questioner and the target doctor to access it. In addition, there are categories of diseases on the site, and only about 100 consultation results are displayed for each type of disease. Each doctor's response can be either spoken or written; however, since the spoken answers are not as accurate as the written answers and contain many spoken words, only the written answers were adopted (Multimedia Appendix 1). From May 20, 2023, to May 30, 2023, a total of 8 types of illness (with a total of 800 cases) were identified, namely gout, osteoarthritis, plantar fasciitis, fracture, osteoporosis, lumbar disc herniation, tendon sheath cyst, and osteoporosis. Of these, 82 patients originally met the screening criteria according to the above requirements. The 82 issues (Figure 1) we collected from this website are compliant with

XSL•FO
RenderX

the HIPAA (Health Insurance Portability and Accountability Act) of 1996, given the information provided above [16]. “Doctor answers” refers to the website’s collection of responses from board-certified physicians (Figure 2A). Multimedia Appendix 2 contains all queries obtained from the Doctor DingXiang website, as well as the doctors’ responses.

Figure 1. (A) Patient health consultation and certified physician's answer on the Doctor DingXiang website (translation from Chinese to English completed by ChatGPT-4). (B) The responses to the health queries were entered as Chinese text into ChatGPT-4. A high-definition version is available in Multimedia Appendix 3

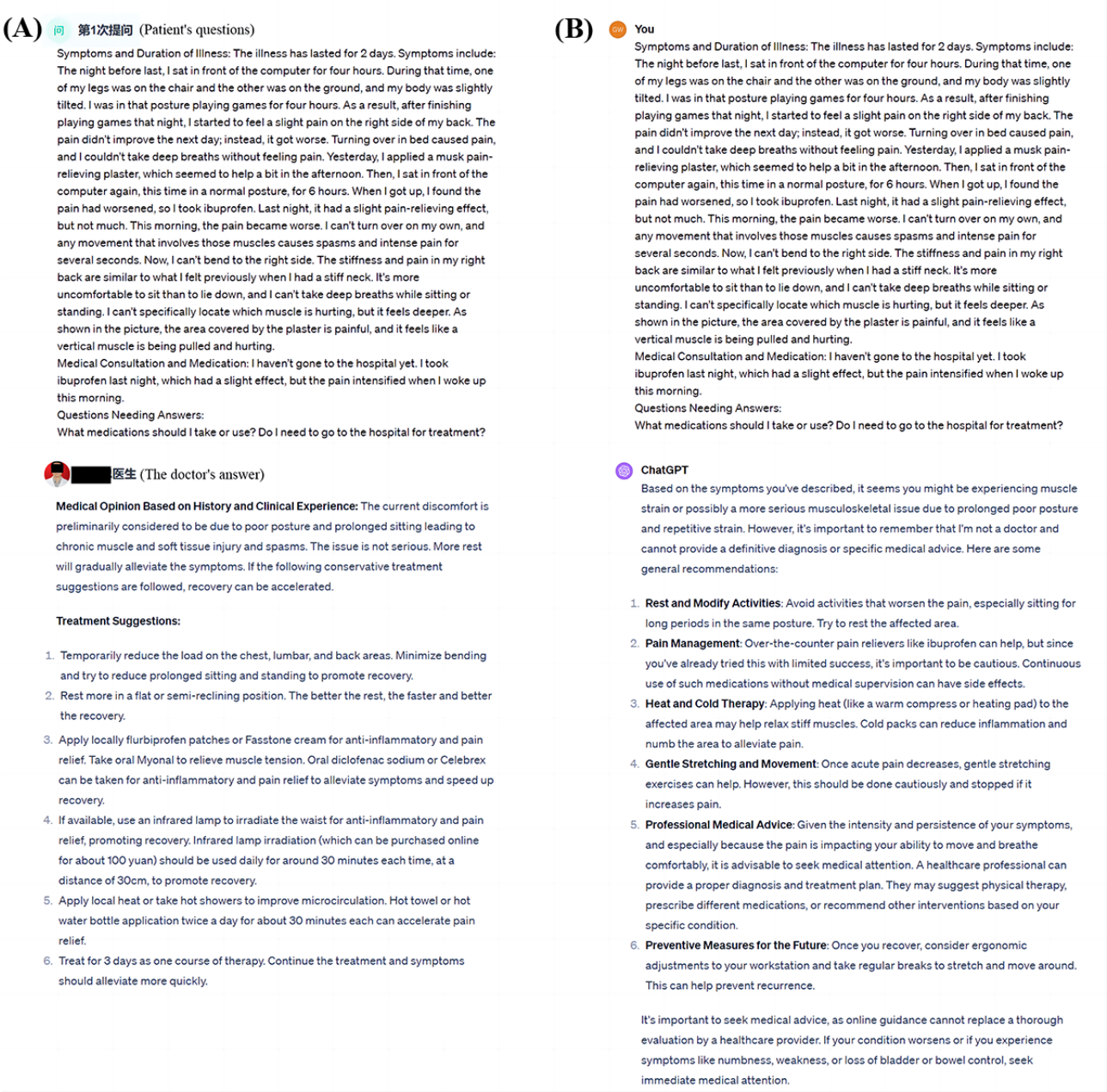

XSL•FO
RenderX

Figure 2. Flow diagram showing the study process.

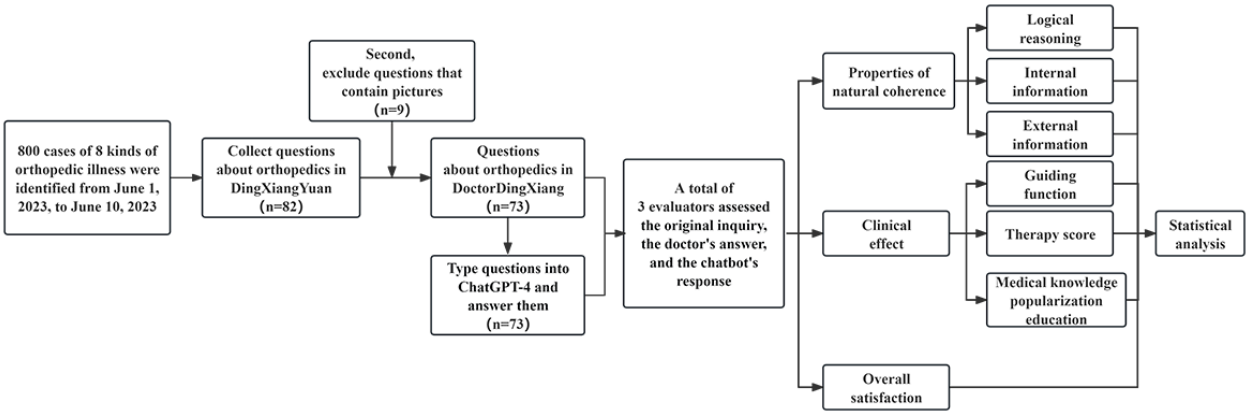

ChatGPT's Answers

ChatGPT exhibits robust learning capabilities within the same dialogue window, enhancing responses to subsequent questions based on previous answers. However, this ability also introduces the potential for systematic error. To elaborate, this interconnectedness of responses does not allow for the maintenance of independence in ChatGPT-4's answers to each question. Therefore, when 73 patients' questions from the included consultations were entered into ChatGPT as questions (Figure 1B), a "new chat" was created for each question-and-answer set to minimize systematic errors. This process took place from June 1, 2023, to June 10, 2023. The use of a "new chat" for each inquiry ensured the independence of each response by preventing the AI from using context from previous interactions, thereby eliminating any learning or bias that may have been carried over from earlier questions. In addition, no plug-ins were used with ChatGPT-4, and the "chat history and training" option was deactivated to preserve the objectivity of each response. All ChatGPT-4 answers can be found in Multimedia Appendix 4.

Response Qualification

The "data set of orthopedic-related remote consultation" was compiled by a professional orthopedic doctor on the Doctor DingXiang website, and 3 professional orthopedic physicians evaluated the ChatGPT and doctor response quality scores. To reduce systematic error resulting from human factors, the orthopedic surgeon who assessed the answers did not know how the answers were grouped. Specific scoring criteria were separated into "properties of natural coherence," "clinical effect," and "overall satisfaction" (Multimedia Appendix 5). The 3 orthopedic physicians convened initially to calibrate their scoring criteria using 2 examples provided by the author (Multimedia Appendix 6). After individual scoring, the Fleiss κ method was used to test the interrater consistency among the 3 physicians' scores. The final statistical data were derived from the mean value of the scores given by the 3 physicians.

Dependability of Comparative Analysis of Responses

When discussing the dependability of comparative analysis of responses, it is essential to consider 3 critical aspects: logical reasoning, internal information, and external information. These components collectively form the foundation for assessing the dependability of answers.

1. Logical reasoning: The answer uses logic and stepwise thinking to produce a response with the given information in the question stem.
2. Internal information: The answer uses information present within the question stem to procure a response.
3. External information: The answer uses external information to produce a response.

Usability of Comparative Analysis of Responses

When assessing the usability of comparative analysis of responses in the medical field, it is crucial to focus on how effectively these analyses can guide diagnosis and treatment, provide therapeutic insights, and educate patients on their conditions.

1. Guiding function: To evaluate the accuracy of the provided diagnosis and differential diagnosis as well as the accuracy of the clinical treatment direction judgement and guidance.
2. Therapeutic effect: To determine whether the treatment suggestions provided in response to the consultation are accurate and if they can alleviate or treat the diseases proposed by the patients.
3. Medical knowledge popularization education: To evaluate whether the response introduces the cause and course of the disease and whether it can enhance patients' understanding of the illness.

Overall Satisfaction

On a scale of 1-10 points, the rater assigned a general rating to the replies. A score of 1-3 points indicated that the responses are biased and that they do not include contents that could call for differential diagnosis and certain auxiliary exams that need to be improved. A score of 4-6 points suggests that there is a possible danger of misdiagnosis or a delay in treatment. Scores of 7-9 points indicate consultation services that can practically replace licensed medical professionals. Finally, a score of 10 points indicates a full replacement for a licensed medical professional's consultation service.

Statistics

For statistical analysis, SPSS (version 26.0; IBM Corporation) was used. Chi-square analysis was used to analyze scoring differences between different groups. The Kolmogorov-Smirnov technique was used to determine whether the data exhibited a normal distribution; ultimately, it indicated that none of the data

XSL•FO
RenderX

in this investigation were normally distributed. Consequently, the Mann-Whitney *U* test of independent samples was used to assess the disparity in scoring performance between the experimental and control groups [17]. When $P<.05$, the difference was considered statistically significant. Scott π statistic is a statistical measure of interrater reliability. Fleiss κ is a generalization of this statistic. SPSS was used to examine the consistency of the 3 raters for each item. Finally, GraphPad Prism 8 (GraphPad Software) was used to construct bar charts to display the comparison of dependability and usability between 2 types of responses, as well as the overall satisfaction outcomes.

Results

Orthopedic Case Selection and Comparative Assessment

We selected 8 orthopedic diseases from the Doctor DingXiang website and consulted 800 cases in total, namely fracture, osteoarthritis, cervical spondylosis, lumbar disc herniation, tendon sheath cyst, plantar fasciitis, osteoporosis, and gout. In the initial screening, we excluded 717 cases in which patients provided information that visitors could not view or where doctors used voice responses. The second screening process excluded patients who provided information that the visitor could not view in the follow-up questions (a total of 9 cases). Finally, 73 eligible cases were included. Without being aware of the replies' origin, 3 orthopedic physicians in practice assessed the responses. The authors concluded by summarizing the statistical findings and designating the response assessment of Doctor DingXiang as the control group and the response evaluation of ChatGPT-4 as the experimental group.

Evaluation Results for Dependability and Usability

After statistical scoring, we discovered that ChatGPT's "internal information" score (mean 4.61, SD 0.52 points vs mean 4.66, SD 0.49 points; $P=.43$) and "therapeutic effect" score (mean 4.43, SD 0.75 points vs mean 4.55, SD 0.62 points, $P=.32$) were lower than those of the control group, but the differences were not statistically significant ($P>.05$; Figures 3E and 4E). ChatGPT showed better performance in the "logical reasoning" score (mean 4.81, SD 0.36 points vs mean 4.75, SD 0.39 points; $P=.38$), "external information" score (mean 4.06, SD 0.72 points vs mean 3.92, SD 0.77 points; $P=.25$), and "guiding function" score (mean 4.73, SD 0.51 points vs mean 4.72, SD 0.54 points; $P=.96$), although the changes were not statistically significant (Figures 3D, 3F, and 4D). However, we were glad to see that, in terms of remote diagnosis and treatment, ChatGPT's "medical knowledge popularization education" scores were better than those of the control group (mean 4.49, SD 0.67 points vs mean 3.87, SD 1.01 points; $P<.001$), and the difference was statistically significant (Figure 4F). Figure 3A depicts the score distribution of ChatGPT and the control group in terms of "logical reasoning."

Figure 3. (A) The distribution of logical reasoning scores in the 2 groups. (B) The distribution of internal information scores in the 2 groups. (C) The distribution of external information scores in the 2 groups. (D) Logical reasoning scores of the 2 groups. (E) Internal information scores of the 2 groups. (F) External information scores of the 2 groups.

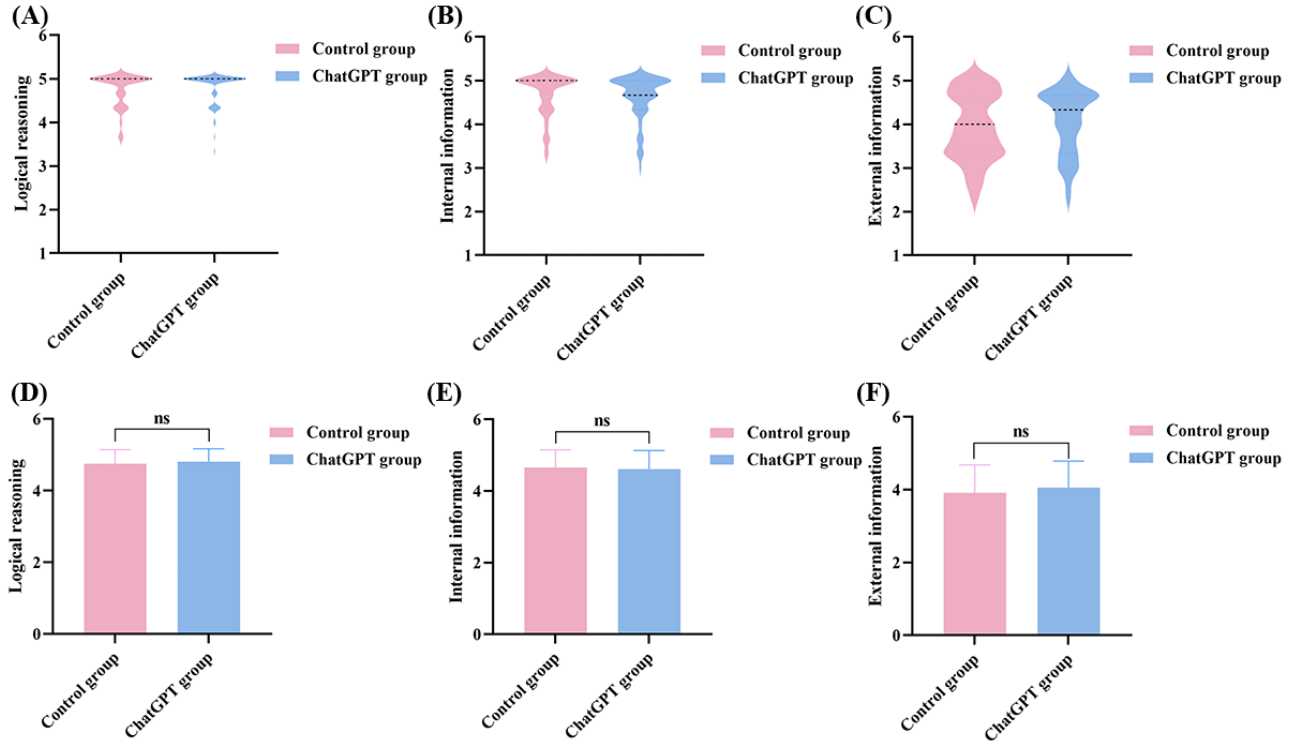

The figures show the distributions of "logical reasoning" (Figure 3A), "internal information" (Figure 3B), "external information" (Figure 3C), "guiding function" (Figure 4A), "therapeutic effect" (Figure 4B), and "medical knowledge popularization education" (Figure 4C). Other than that for "medical knowledge popularization education," the score distribution for the remaining elements was roughly comparable.

XSL•FO
RenderX

Figure 4. (A) The distribution of guiding function scores in the 2 groups. (B) The distribution of therapeutic effect scores in the 2 groups. (C) The distribution of medical knowledge popularization education scores in the 2 groups. (D) Guiding function scores of the 2 groups. (E) Therapeutic effect scores of the 2 groups. (F) Medical knowledge popularization education scores of the 2 groups (*P*<.001).

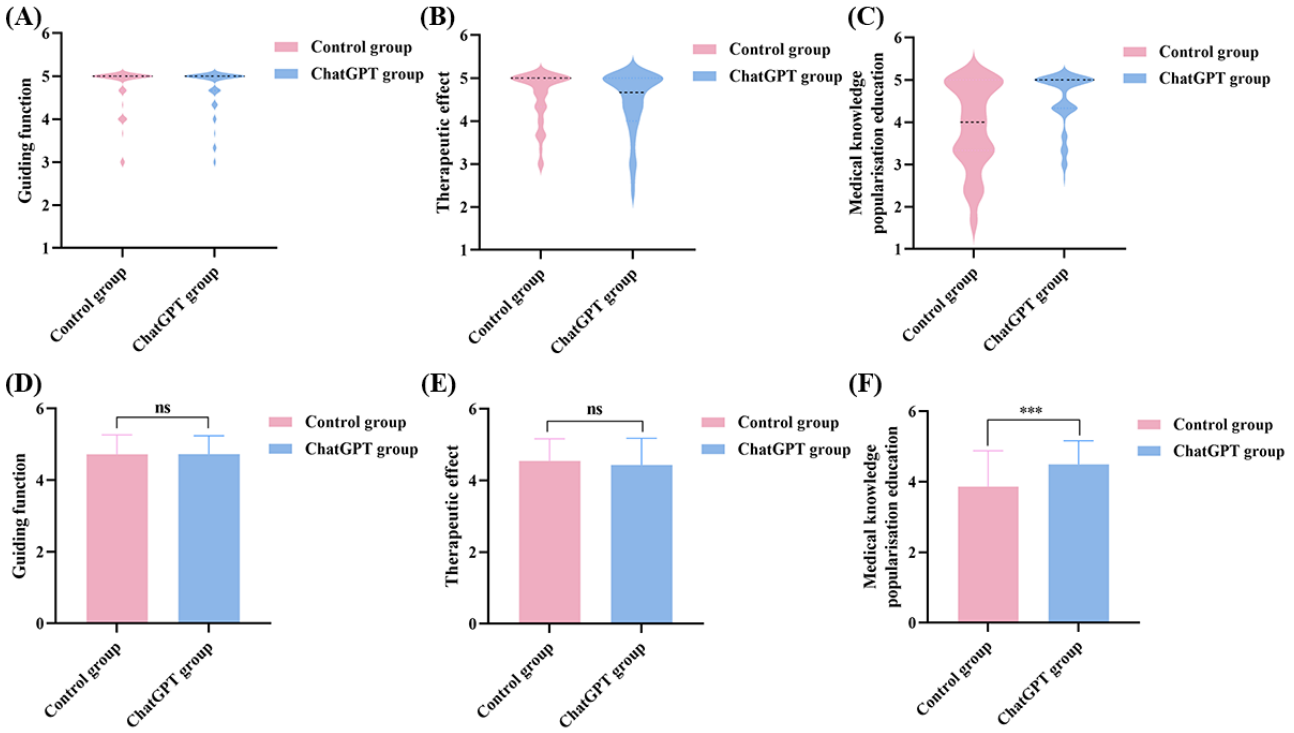

In terms of "overall satisfaction," we see that ChatGPT had slightly higher overall satisfaction scores of <5 points compared with the control group (Figures 5A and 5B), but the difference was not statistically significant (mean 8.35, SD 1.38 points vs mean 8.37, SD 1.24 points; *P*=.92; Figure 5C).

Figure 5. (A) The distribution of overall satisfaction scores in the control group. (B) The distribution of overall satisfaction scores in the ChatGPT group. (C) Overall satisfaction scores of the 2 groups.

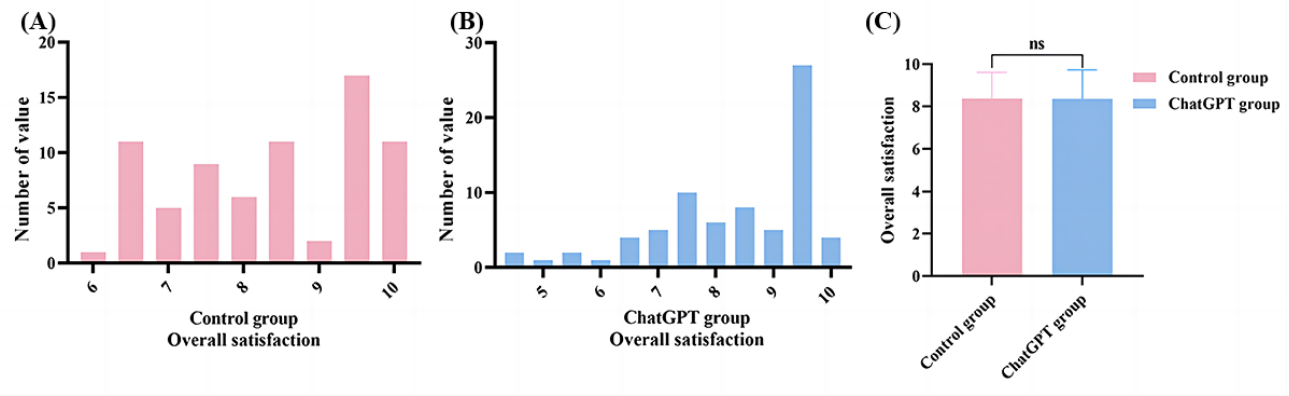

Consistency Testing Among the 3 Orthopedic Physicians' Evaluations

Using Fleiss κ, the consistency of the ratings among 3 physicians was determined (Multimedia Appendix 5). The Fleiss κ evaluations for "logical reasoning," "internal information," "external information," "therapeutic effect," "medical knowledge popularization education," and "overall satisfaction" were rated as showing "fair agreement" for the control group, while "guiding function" was rated as showing "substantial agreement" (Multimedia Appendix 5; *P*<.001). According to Fleiss κ values, "internal information," "external information," and "overall satisfaction" were rated as displaying "fair agreement" in the ChatGPT responses, whereas "logical reasoning," "guiding function," "therapeutic effect," and "medical knowledge popularization education" were rated as showing "moderate agreement" (Multimedia Appendix 5; *P*<.001).

Discussion

Main Findings of This Study

This cross-sectional research gathered 73 frequently asked clinical questions from patients and excellent responses given by licensed, qualified physicians on a reputable remote medical service website. After using ChatGPT to get the answers to these queries and seeking the evaluation of professional doctors, we discovered that, when compared with the responses of

XSL•FO
RenderX

qualified clinicians, ChatGPT's answers also showed strong logic and the capacity to extract and analyze key information, and they could help physicians respond to patients' questions in a manner that reflects professionalism. Overall, ChatGPT's responses received generally positive feedback from doctors. Professional responses from ChatGPT were able to assess and address queries in light of a large database, and the service even suggested literature on diseases for interested customers. With the help of ChatGPT, this method may unleash latent productivity, allowing health care personnel to use the time saved on more difficult duties. However, ChatGPT still has a lot of drawbacks. Although ChatGPT can analyze photographs, the procedure to do so is very complicated: medical images must be submitted to a public site to establish links for analysis, and the success rate of analysis is not very high. In addition, ChatGPT is not yet able to accurately diagnose a patient's illness; this task must be left to expert physicians, whose assessment and oversight are crucial to the process [18]. Therefore, we believe that ChatGPT may successfully help doctors with remote diagnosis and treatment services, significantly increase clinicians' job efficiency, and save more time, but it still cannot take over from the doctor entirely.

Comparison With Previous Research

As the internet has grown, many hospitals have established remote medical services. Doctor-patient contact is no longer hampered by distance thanks to the internet, which makes it easier for both parties to interact. Remote medical services have expanded quickly over the last 3 years as a result of the COVID-19 pandemic, and, to some degree, they have even altered the conventional medical model. Remote diagnosis and therapy, nevertheless, are not yet flawless. Patients must pay additional costs for remote diagnostic and therapy services, and their communications may be ignored or they may receive pointless answers [2]. More crucially, in certain fields, including orthopedics, textual communication alone may be unable to provide clinicians with a whole picture of the patient's condition. There is still no replacement for a physical examination, imaging examination, or biochemical test. In addition, physicians must expend a great deal of additional time and effort to decide how to respond to patients, which adds significantly to their burden and may not have the intended outcome [19].

Previous studies have indicated that ChatGPT-3.5 demonstrated strong performance in addressing public health inquiries on Reddit's r/AskDocs, showcasing its considerable promise for offering remote medical consultation services [20]. This is noteworthy given the hesitancy of some patients to discuss their health issues publicly, coupled with the challenge of ensuring the reliability of unpaid responses on such platforms [20]. Contrasting with this, this study compares ChatGPT-4 with paid professional responses on the Doctor DingXiang forums, revealing that ChatGPT-4's overall performance is comparable to that of paid medical professionals, with the added benefit of more effective dissemination of medical knowledge. In addition, ChatGPT's low barrier to entry means this real-time, AI-driven, question-and-answer software better addresses the immediate health consultation needs of users, making it more significant for widespread application.

Interactive AI software that offers immediate feedback has an advantage over traditional AI analytical output software in that it allows users to inquire not only about the answers to "what" but also about the underlying "why" [21]. In the context of clinical scenarios, users have the ability to request critical information and foundations for diagnosis and treatment through ChatGPT. This functionality aids in the clarification of the operational logic behind their decisions, fostering greater transparency in the use of ChatGPT software and the comprehension of users. In relation to personal privacy, users have the ability to configure ChatGPT's personal settings to "chat history and training" and enable a personalized input mode to proactively minimize the exposure of sensitive data.

Significance for Hierarchical Diagnosis and Treatment as Well as Triage

A major worldwide problem is the scarcity and unequal distribution of medical resources. The issue is made worse in certain nations with high population densities, such as China, by the abundance of people who require medical treatment [22]. In addition, China is unable to guarantee the effectiveness of medical resource allocation, as other high-income countries can, due to ineffective rules and legislation and a lack of rigorously educated general practitioners [23]. To address this issue, the hierarchical medical system was created, and it has steadily replaced other medical systems to provide basic health care in the majority of high-income countries [24]. According to the severity and urgency of their sickness, patients must be sent to medical facilities of the appropriate level, such as primary medical institutions or specialized medical institutions [25]. This is a perfect medical paradigm, but patients' treatment decisions are significantly influenced by their self-rated health state, chronic illnesses, socioeconomic situation, and educational level, particularly since the majority of patients lack an objective grasp of their ailment and pertinent medical expertise [26,27]. The hierarchical medical system has not had the desired impact in China as a consequence of its deployment. Some medical facilities are suffering from severe work pressure overload due to a lack of medical resources and patients' unrealistic treatment preferences [23]. To enhance patients' medical behavior and help them choose the best medical facilities, high-quality guiding services are thus necessary to assist patients in understanding their disease-related information before treatment [28]. This may somewhat mitigate the issues brought on by a lack of medical resources and assist patients in receiving more focused and appropriate medical care.

Patient navigation services, a patient-centered intervention, are becoming more and more popular. These services use trained personnel to identify patient-level barriers to care, such as cultural, logistical, and educational ones, and then remove them to encourage full and prompt access to care [29,30]. A growing body of research demonstrates the beneficial effects that patient guide services have on illness prevention, the spread of health information, medical decision-making, and communication promotion. Patient navigation can help remove barriers brought on by language, cultural differences, a lack of relevant medical knowledge, and other factors, especially for some patients with a lack of medical knowledge or a relatively low level of education, in the face of a more complex but hierarchical

XSL•FO
RenderX

medical center or sociomedical system. This will lead to a more effective patient path and fewer delays in diagnosis and treatment [31]. More crucially, research has demonstrated that patient guiding services have benefited people with chronic illnesses such as diabetes and cardiovascular disease and have somewhat decreased the likelihood of rehospitalization [32]. Patient guidance services may not only aid in the patient's healing process but also assist them with developing a more thorough and expert understanding of the causes, symptoms, and other facets of associated diseases, enabling them to treat, care for, and monitor their disease more skillfully and effectively [33]. However, previous research discovered that some issues remain with the present patient guidance service, such as navigators' potential lack of expertise. In addition, some patient navigators, although trained on how to perform their job, lack a history of medical education, making it difficult for them to respond to the patient's consultation [11]. Even if it may be a little harsh to demand that patient navigators be all-knowing, finding practical and trustworthy approaches to boost the effectiveness and caliber of patient navigation services is still necessary.

The Challenges of Promoting ChatGPT in the Medical Field

While using AI is the general trend in science and technology development, individuals must also understand that the tool can only work optimally in the ideal regulatory environment, which often has some lag. To ensure the rational use of ChatGPT in the medical field, hospitals need to organize training on the use of ChatGPT and uniformly manage the accounts used by doctors during working hours. Doctors must also take responsibility for assessing the quality of ChatGPT's responses and ensuring that the patient's right to be informed of the use of ChatGPT is met. Specifically, physicians are required to assign a unique account when using ChatGPT in clinical practice, and they must also have the corresponding patient present. The physician has the authority not only to assess the quality of ChatGPT's responses before presenting them to the patient but also to provide the patient with the final interpretation of said responses. Conversely, individuals who use ChatGPT but do not identify as medical professionals should refrain from relying exclusively on it for health-related information.

ChatGPT can assist clinicians in better organizing clinical data, analyzing imaging results, and providing personalized support for clinical decision-making regarding cancer patients, according to recent studies [34-36]. As previously stated, physicians, in their capacity as users of ChatGPT, are additionally obligated to oversee its use. In this regard, ChatGPT functions as a supplementary tool. Should ChatGPT outputs be incorporated into the physician-patient communication and clinical decision-making process, the physician must disclose the information source to the patients to guarantee that they are well-informed. Simultaneously, the hospital must oversee the ChatGPT accounts used by physicians and coordinate training courses on ChatGPT usage to guarantee that physicians who use ChatGPT in their clinical practice possess a certain level of proficiency in its operation. By implementing these management tasks, certain potential hazards and medical disputes can be circumvented, and the application of AI software in the medical field can be promoted more effectively.

Although this study establishes a sound theoretical foundation for the clinical implementation of ChatGPT, there are numerous areas still requiring further refinement. As one example, there is a need for further refinement of cross-sectional experiments in the future to compare the following: the quality of answers provided by AI software in various clinical disciplines, variations in the quality of answers generated by different AI software programs (ChatGPT, Google Board, Claude, and so forth), and disparities between different language inputs used by AI software. Alternatively, a randomized controlled trial could assess the efficacy of ChatGPT as a supplementary tool for clinicians to use while interacting with patients. Further development is required to ensure the full functionality, safety, and dependability of ChatGPT as a medical AI.

Limitations

Initially, we intended to investigate the viability of using the ChatGPT app for medical guidance. This study solely included orthopedic cases as the research object and did not gather multidisciplinary clinical cases to rule out variations in the difficulty levels of working in other clinical specialties, which may produce different findings. In the future, it will be possible to aggregate challenges from many disciplines and examine how AI performance differs between fields in solving difficulties. Furthermore, neither machine translation nor manual translation can preserve the flaws and precision of the original sentence content. Users are unable to ascertain the processing logic of AI when using ChatGPT as a research tool across different language types. Consequently, they are limited to inputting ChatGPT data in accordance with the language type used in the control content and assessing the output quality of ChatGPT content in the same language. Medical personnel are required to use ChatGPT under a special number with a real-name system for supervision purposes. As an auxiliary tool, ChatGPT users are not only tasked with assessing the quality of responses but also possess the authority to make the ultimate interpretation of the content. Ultimately, further randomized controlled trials are required in the future to validate the use of AI in medicine while controlling for confounding variables, as this study was cross-sectional in nature.

Conclusion

This study demonstrates that ChatGPT-4 responses match the expertise found among health care practitioners on DingXiangYuan, a leading remote medical consultation platform in China, across various metrics such as logical reasoning and diagnostic accuracy. Notably, it excels at providing scientific education. ChatGPT-4 is thus recommended as an alternative to traditional remote health consultations. It can assist physicians in educating patients, thereby enhancing medical knowledge dissemination. For patients, it offers accessible, reliable health advice, improving information accessibility and decision-making support. These findings suggest a transformative potential for ChatGPT-4 in health care, notably in enhancing access to medical advice and patient education. It implies the need for advancing medical AI with a focus on ethical and transparent

XSL•FO
RenderX

applications, highlighting its role in improving health care delivery and patient empowerment.

Acknowledgments

The authors would like to thank DingXiangYuan and the Doctor DingXiang website's public display of orthopedic-related remote consultation cases. We thank the LetPub website for its linguistic assistance during the preparation of this manuscript.

Authors' Contributions

ZX was responsible for conceptualization, investigation, visualization, and writing of the original draft, as well as writing, reviewing, and editing. YZ was responsible for data curation, formal analysis, and writing the original draft. WG was responsible for data curation and formal analysis. HW, GS, and XZ were responsible for grading the responses, writing—reviewing and editing, and supervising the entire study. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Select inclusion criteria and exclusion criteria for website consultation dialogue information.
[DOCX File , 13 KB - jmir_v26i1e50882_app1.docx]

Multimedia Appendix 2
The 82 patient questions and online responses from doctors on the Doctor DingXiang website (translated from Chinese to English using ChatGPT-3.5).
[DOCX File , 133 KB - jmir_v26i1e50882_app2.docx]

Multimedia Appendix 3
High-resolution version Figure 1.
[ZIP File (Zip Archive), 24742 KB - jmir_v26i1e50882_app3.zip]

Multimedia Appendix 4
ChatGPT-4 responses after asking questions from 82 patients in Chinese (translated from Chinese to English using ChatGPT-3.5).
[DOCX File , 78 KB - jmir_v26i1e50882_app4.docx]

Multimedia Appendix 5
Consistent evaluation of Fleiss κ among the 3 raters.
[DOCX File , 14 KB - jmir_v26i1e50882_app5.docx]

Multimedia Appendix 6
Orthopaedic physicians' initial criteria calibration.
[DOCX File , 15 KB - jmir_v26i1e50882_app6.docx]

References

1. Markowitz J. Virtual treatment and social distancing. Lancet Psychiatry 2020;7(5):388-389 [FREE Full text] [doi: 10.1016/S2215-0366(20)30140-1] [Medline: 32353268]
2. Zulman DM, Verghese A. Virtual care, telemedicine visits, and real connection in the era of COVID-19: unforeseen opportunity in the face of adversity. JAMA 2021;325(5):437-438 [FREE Full text] [doi: 10.1001/jama.2020.27304] [Medline: 33528520]
3. Traeger AC, Lee H, Hübscher M, Skinner IW, Moseley GL, Nicholas MK, et al. Effect of intensive patient education vs placebo patient education on outcomes in patients with acute low back pain: a randomized clinical trial. JAMA Neurol 2019;76(2):161-169 [FREE Full text] [doi: 10.1001/jamaneurol.2018.3376] [Medline: 30398542]
4. ChatGPT. OpenAI. URL: https://chat.openai.com/chat [accessed 2024-02-07]
5. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of ada health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth 2023;11:e49995 [FREE Full text] [doi: 10.2196/49995] [Medline: 37788063]
6. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience 2023;26(9):107590 [FREE Full text] [doi: 10.1016/j.isci.2023.107590] [Medline: 37705958]

XSL•FO
RenderX

7. Yao LH, Leung KC, Tsai CL, Huang CH, Fu LC. A novel deep learning-based system for triage in the emergency department using electronic medical records: retrospective cohort study. J Med Internet Res 2021;23(12):e27008 [FREE Full text] [doi: 10.2196/27008] [Medline: 34958305]
8. Abdelkader OA. ChatGPT's influence on customer experience in digital marketing: investigating the moderating roles. Heliyon 2023;9(8):e18770 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e18770] [Medline: 37576290]
9. Li Y, Cen J, Wu J, Tang M, Guo J, Hang J, et al. The degree of anxiety and depression in patients with cardiovascular diseases as assessed using a mobile app: cross-sectional study. J Med Internet Res 2023;25:e48750 [FREE Full text] [doi: 10.2196/48750] [Medline: 37792455]
10. Marin CE, de O Pinto P, Dos Passos GR, Cuervo DL, Wagner MB, Becker J, et al. Reliability of telemedicine evaluation for EDSS functional systems in multiple sclerosis. J Telemed Telecare 2023:1357633X231207903. [doi: 10.1177/1357633X231207903] [Medline: 37904499]
11. Roberge J, McWilliams A, Zhao J, Anderson WE, Hetherington T, Zazzaro C, et al. Effect of a virtual patient navigation program on behavioral health admissions in the emergency department: a randomized clinical trial. JAMA Netw Open 2020;3(1):e1919954 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.19954] [Medline: 31995214]
12. Ruan Y, Luo J, Lin H. Why do patients seek diagnose dis-accordance with hierarchical medical system related policies in tertiary hospitals? a qualitative study in shanghai from the perspective of physicians. Front Public Health 2022;10:841196 [FREE Full text] [doi: 10.3389/fpubh.2022.841196] [Medline: 35400060]
13. Sinsky CA, Shanafelt TD, Ripp JA. The electronic health record inbox: recommendations for relief. J Gen Intern Med 2022;37(15):4002-4003 [FREE Full text] [doi: 10.1007/s11606-022-07766-0] [Medline: 36036837]
14. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
15. Liu H, Tan Y, Zhang M, Peng Z, Zheng J, Qin Y, et al. An internet-based survey of influenza vaccination coverage in healthcare workers in China, 2018/2019 season. Vaccines (Basel) 2019;8(1):6 [FREE Full text] [doi: 10.3390/vaccines8010006] [Medline: 31887994]
16. Rose RV, Kumar A, Kass JS. Protecting privacy: health insurance portability and accountability act of 1996, Twenty-First Century Cures Act, and social media. Neurol Clin 2023;41(3):513-522. [doi: 10.1016/j.ncl.2023.03.007] [Medline: 37407103]
17. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ 2023;9:e51421 [FREE Full text] [doi: 10.2196/51421] [Medline: 37910155]
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
19. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022;29(3):453-460 [FREE Full text] [doi: 10.1093/jamia/ocab268] [Medline: 34888680]
20. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
21. Quer G, Muse ED, Nikzad N, Topol EJ, Steinhubl SR. Augmenting diagnostic vision with AI. Lancet 2017;390(10091):221 [FREE Full text] [doi: 10.1016/S0140-6736(17)31764-6] [Medline: 30078917]
22. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. Lancet Glob Health 2020;8(4):e480 [FREE Full text] [doi: 10.1016/S2214-109X(20)30068-1] [Medline: 32109372]
23. Liang C, Zhao Y, Yu C, Sang P, Yang L. Hierarchical medical system and local medical performance: a quasi-natural experiment evaluation in Shanghai, China. Front Public Health 2022;10:904384 [FREE Full text] [doi: 10.3389/fpubh.2022.904384] [Medline: 36324471]
24. Goldfield N, Gnani S, Majeed A. Primary care in the United States: profiling performance in primary care in the United States. BMJ 2003;326(7392):744-747 [FREE Full text] [doi: 10.1136/bmj.326.7392.744] [Medline: 12676844]
25. Jiang Y, Cai X, Wang Y, Dong J, Yang M. Assessment of the supply/demand balance of medical resources in Beijing from the perspective of hierarchical diagnosis and treatment. Geospat Health 2023;18(2) [FREE Full text] [doi: 10.4081/gh.2023.1228] [Medline: 37831418]
26. Li J, Zhao N, Zhang H, Yang H, Yang J. Patients' willingness of first visit in primary medical institutions and policy implications: a national cross-sectional survey in China. Front Public Health 2022;10:842950 [FREE Full text] [doi: 10.3389/fpubh.2022.842950] [Medline: 35433566]
27. Li G, Han C, Liu P. Does internet use affect medical decisions among older adults in China? Evidence from CHARLS. Healthcare (Basel) 2021;10(1):60 [FREE Full text] [doi: 10.3390/healthcare10010060] [Medline: 35052224]
28. Wu F, Ozaki A, Zhao G. Patient navigation for comprehensive cancer screenings in high-risk patients. JAMA Intern Med 2016;176(11):1725-1726. [doi: 10.1001/jamainternmed.2016.6323] [Medline: 27820644]
29. McKenney KM, Martinez NG, Yee LM. Patient navigation across the spectrum of women's health care in the United States. Am J Obstet Gynecol 2018;218(3):280-286 [FREE Full text] [doi: 10.1016/j.ajog.2017.08.009] [Medline: 28844825]

XSL•FO
RenderX

30. Ko NY, Snyder FR, Raich PC, Paskett ED, Dudley DJ, Lee JH, et al. Racial and ethnic differences in patient navigation: results from the patient navigation research program. Cancer 2016;122(17):2715-2722 [FREE Full text] [doi: 10.1002/cncr.30109] [Medline: 27227342]
31. Rodday AM, Parsons SK, Snyder F, Simon MA, Llanos AAM, Warren-Mears V, et al. Impact of patient navigation in eliminating economic disparities in cancer care. Cancer 2015;121(22):4025-4034 [FREE Full text] [doi: 10.1002/cncr.29612] [Medline: 26348120]
32. Cadzow RB, Craig M, Rowe J, Kahn LS. Transforming community members into diabetes cultural health brokers: the neighborhood health talker project. Diabetes Educ 2013;39(1):100-108. [doi: 10.1177/0145721712465342] [Medline: 23150528]
33. Braun KL, Kagawa-Singer M, Holden AEC, Burhansstipanov L, Tran JH, Seals BF, et al. Cancer patient navigator tasks across the cancer care continuum. J Health Care Poor Underserved 2012;23(1):398-413 [FREE Full text] [doi: 10.1353/hpu.2012.0029] [Medline: 22423178]
34. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6(11):e2343689 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
35. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. J Am Acad Orthop Surg 2024;32(3):123-129. [doi: 10.5435/JAAOS-D-23-00474] [Medline: 37976385]
36. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google bard, and Microsoft bing for simplifying radiology reports. Radiology 2023;309(2):e232561. [doi: 10.1148/radiol.232561] [Medline: 37987662]

Abbreviations

AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act

Edited by A Castonguay; submitted 15.07.23; peer-reviewed by M Chatzimina, F Tang, J Li; comments to author 27.10.23; revised version received 04.11.23; accepted 30.01.24; published 14.03.24.

Please cite as:
Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X
Quality and Dependability of ChatGPT and DingXiangYuan Forums for Remote Orthopedic Consultations: Comparative Analysis
J Med Internet Res 2024;26:e50882
URL: https://www.jmir.org/2024/1/e50882
doi:10.2196/50882
PMID:38483451

©Zhaowen Xue, Yiming Zhang, Wenyi Gan, Huajun Wang, Guorong She, Xiaofei Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study

Lei Wang[1,2*], MS; Yinyao Ma[3*], MS, MD; Wenshuai Bi[1], MS; Hanlin Lv[1], MD, PhD; Yuxiang Li[1,2], PhD

[1]BGI Research, Wuhan, China

[2]Guangdong Bigdata Engineering Technology Research Center for Life Sciences, BGI Research, Shenzhen, China

[3]Department of Obstetrics, People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, China

[*]these authors contributed equally

Corresponding Author:
Hanlin Lv, MD, PhD
BGI Research
1-2F, Building 2, Wuhan Optics Valley International Biomedical Enterprise Accelerator Phase 3.1
No 388 Gaoxin Road 2, Donghu New Technology Development Zone
Wuhan, 430074
China
Phone: 86 18707190886
Email: lvhanlin@genomics.cn

Abstract

Background: The study of disease progression relies on clinical data, including text data, and extracting valuable features from text data has been a research hot spot. With the rise of large language models (LLMs), semantic-based extraction pipelines are gaining acceptance in clinical research. However, the security and feature hallucination issues of LLMs require further attention.

Objective: This study aimed to introduce a novel modular LLM pipeline, which could semantically extract features from textual patient admission records.

Methods: The pipeline was designed to process a systematic succession of concept extraction, aggregation, question generation, corpus extraction, and question-and-answer scale extraction, which was tested via 2 low-parameter LLMs: Qwen-14B-Chat (QWEN) and Baichuan2-13B-Chat (BAICHUAN). A data set of 25,709 pregnancy cases from the People's Hospital of Guangxi Zhuang Autonomous Region, China, was used for evaluation with the help of a local expert's annotation. The pipeline was evaluated with the metrics of accuracy and precision, null ratio, and time consumption. Additionally, we evaluated its performance via a quantified version of Qwen-14B-Chat on a consumer-grade GPU.

Results: The pipeline demonstrates a high level of precision in feature extraction, as evidenced by the accuracy and precision results of Qwen-14B-Chat (95.52% and 92.93%, respectively) and Baichuan2-13B-Chat (95.86% and 90.08%, respectively). Furthermore, the pipeline exhibited low null ratios and variable time consumption. The INT4-quantified version of QWEN delivered an enhanced performance with 97.28% accuracy and a 0% null ratio.

Conclusions: The pipeline exhibited consistent performance across different LLMs and efficiently extracted clinical features from textual data. It also showed reliable performance on consumer-grade hardware. This approach offers a viable and effective solution for mining clinical research data from textual records.

(J Med Internet Res 2024;26:e54580) doi:10.2196/54580

KEYWORDS

clinical data extraction; large language models; feature hallucination; modular approach; unstructured data processing

Introduction

Clinical text data have been widely recognized in data research due to their inclusion of multisource information [1,2] (eg, patient subjective statements, past objective facts, doctors' diagnostic processes, and summary records). Extracting useful information from text data could serve as a crucial supplement to the study of disease progression; it could complement objective indicators dependent on laboratory tests and examinations [3], which has consistently been a hot research topic [4,5].

XSL•FO
RenderX

Historically, methods for text data extraction mainly include the following:

- Manual annotation: scales are designed based on clinical and research experience, followed by manual field extraction [6-8].
- Rule extraction: concepts from established knowledge base, such as *International Classification of Diseases, Tenth Revision* [9], are used for concept term extraction. This process is typically based on similarity algorithms and manual assistance to extract terms and their attributes (eg, negations and dependency relationships) [10].
- Named entity recognition or natural language processing algorithms: supervised learning methods, such as pretrained models like T5 [11], Bidirectional Encoder Representations from Transformers (BERT) [12], and BERT's variants [13-15], with manual annotation to enhance semantic comprehension capabilities [16,17].

The task of extracting features from vast unstructured text presents itself as a daunting, labor-intensive, and time-consuming endeavor [18], for the following reasons:

- It is challenging to determine the dimension of extracted features initially, and from another perspective, confining the feature dimension means constraining the research scope from the outset [19].
- Given the inherent subjectivity and potential biases of recording subjects, solely relying on algorithms without annotation typically results in low accuracy and recall [20].
- Achieving higher accuracy with a broader feature scope, and the required human effort involved, is typically nonlinear [4], and the difficulty becomes apparent when confronted with massive real-world data.

The advent of large language models (LLMs) has paved a new path for the dilemma in clinical text extraction [21-23]. In the realm of natural language understanding research, generative large models, represented notably by ChatGPT [24] since 2022, have achieved unimaginable capabilities in semantic dimensions, leveraging the emergent intelligence from vast parameter scales. However, there are numerous considerations and limitations in their application, as follows:

- High-performing LLMs, such as OpenAI ChatGPT and Google Bard [25,26], are currently not open source, and patient data need to be submitted to their platform for analysis, presenting security challenges [27,28].
- Open-source LLMs with high intelligence generally require a large number of parameters (10-100 billion), which are hard to support on consumer-level graphics processing units (GPUs) [29].
- Low-parameter (around 10 billion) LLMs, typically require multistrategy support when dealing with tasks in certain vertical segments [30] (eg, fine-tuning, knowledge base or knowledge graph support, complex Chain of Thought (CoT) [31] along with its derivatives, and even global training) and are accompanied by various anomaly issues, including feature hallucination.

Although the application of LLM faces various potential limitations and challenges, as mentioned above, the foundational entity extraction and understanding capabilities of LLMs can still be used for low-cost extraction of clinical text data through meticulous prompt design, guidance combining CoT, and standardized examples [30,32].

In this study, we aimed to extract valuable features from a series of given patient admission records, which include the chief complaint and the medical histories. In light of this task, we introduced a modular LLM approach, which divides the entire extraction path into several smaller steps, with each modular LLM handling these basic steps automatically. We adopted the core idea of LLM agents [30] and self-consistency with CoT [33].

To experiment with this approach, we implemented 2 low-parameter LLMs in a local environment and compared their performances within a retrospective cohort of pregnancy to provide a reference that future researchers might draw upon.

Methods

Study Preparation

Data Sources

In this study, the text corpus was compiled from two primary sources:

1. Chief complaints and medical histories, exemplified in Multimedia Appendix 1, were extracted from inpatient admission records of an established cohort at the People's Hospital of Guangxi Zhuang Autonomous Region in China. The established cohort for the preeclampsia risk study consisted of 25,709 pregnancies that received prenatal care between the 11th and 13th weeks of gestation from April 2012 to September 2021.
2. Clinical practice guidelines consisted of the 2018 guidelines from the American College of Obstetricians and Gynecologists [34] and the 2019 guidelines from the National Institute for Health and Care Excellence [35].

To ensure linguistic consistency, the entire corpus was maintained in Chinese.

Model Deployment

We deployed 2 most exemplary LLMs in the Chinese domain until September 2023 in an intranet security environment independently: Qwen-14B-Chat (QWEN) [36] and Baichuan2-13B-Chat (BAICHUAN) [37]. In the environment, the server cluster used NVIDIA DGX-A100 (2×40 G) GPU nodes. The QWEN used 29 GB of storage and 27 GB of GPU memory, while the BAICHUAN used 26 GB of storage and 28.9 GB of GPU memory. Both models operated solely on physically isolated GPUs, and access was facilitated through the OpenAI [38] format and FastChat [39]. The LLMs were built upon PyTorch 2.0, with the temperature set to 0 and max_token adjusted task by task.

Experimental Path

In this study, we have introduced an approach that autonomously extracts valuable textual features. Diverging from traditional LLM applications, we used an "external-COT" strategy, dividing

XSL•FO
RenderX

the process into several controllable steps, as illustrated in Figure 1.

The extraction approach could be divided into four parts: (1) concept preparation, that is, extracting existing concepts from the corpus and selecting concerning concepts; (2) corpus preparation, that is, deidentifying raw data and preparing the corpus in accordance with the selected concept; (3) prompt design for different LLM tasks; (4) question-and-answer (Q&A) scales, that is, transforming concepts into question templates and extracting corresponding scales by LLMs.

Figure 1. The flowchart of extraction approach. LLM: large language model; Q&A: question and answer; SNOMED CT: Systematised Nomenclature of Medicine Clinical Term.

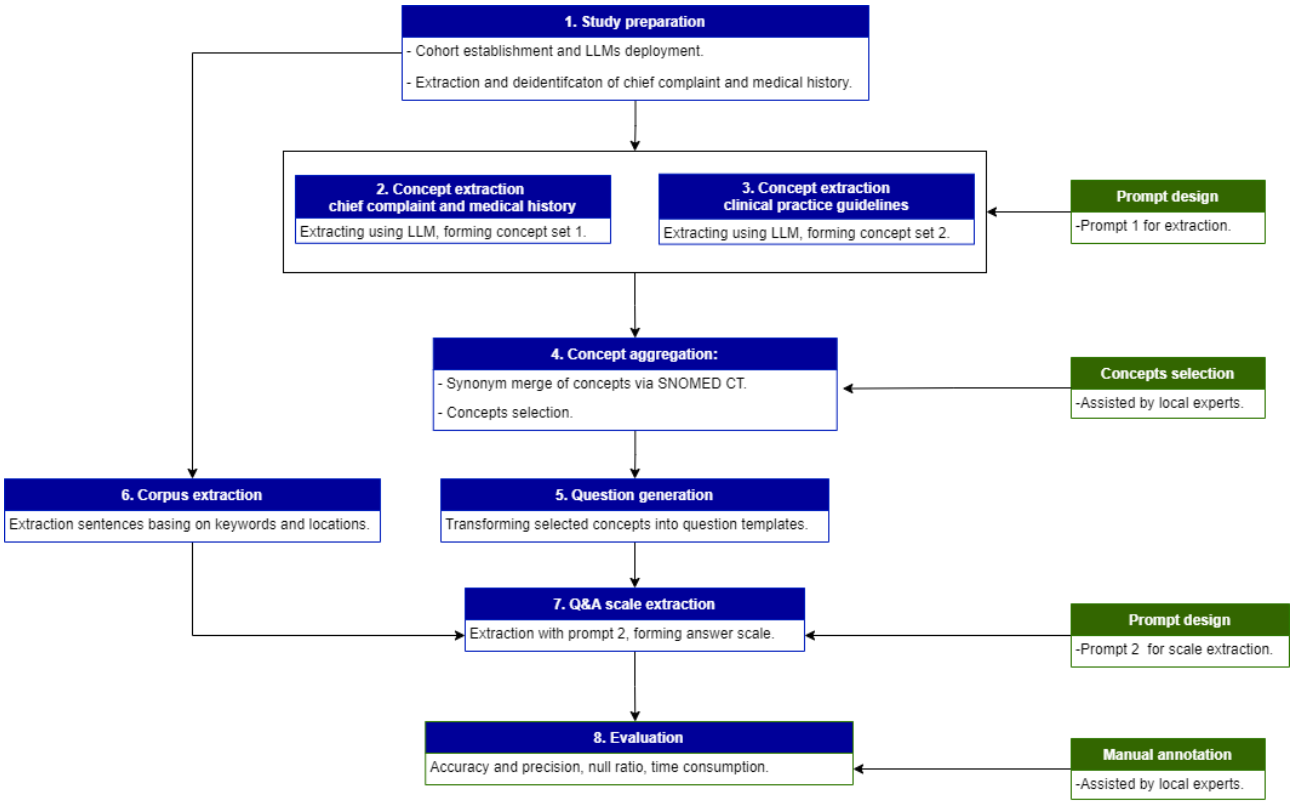

Prompt Design

The design of prompt templates is fundamental to efficient and accurate extraction. Prior to processing the entire data set, an initial evaluation was conducted on 100 observations to assess the effectiveness of the templates, allowing for continuous refinement of prompt strategies and orientations. An appropriate template was defined based on the following criteria: (1) absence of redundant content generation, (2) consistent and uniform efficiency, and (3) infrequent occurrence of feature hallucination.

We adopted a 4-paragraph structure, referring to the prompt engineering suggestions of QWEN and BAICHUAN, as follows:

1. Context section: defines the role and task, provides a basic understanding, and establishes a behavioral baseline for the model.
2. Instruction section: outlines the execution steps, uses the CoT methodology, and provides examples to ensure guided model operation.
3. Input data section: manages various inputs to meet diverse information needs.
4. Output indicator section: specifies the output format and standards, setting clear expectations for the output.

To avoid input bias, the prompt templates for QWEN and BAICHUAN were maintained without any modifications. In addition, we conducted experiments using 100 observations at different levels of concurrency to select the most optimal configuration.

Concept Extraction and Aggregation

We initially extracted all discernible concepts from chief complaints and medical histories using LLMs with a designed prompt 1, and concepts were retained only with a manifestation frequency exceeding 5% occurrences. To reduce potential attention bias and expand the range of identified concepts, we also included concepts from clinical practice guidelines related to preeclampsia, particularly the American College of Obstetricians and Gynecologists 2018 guidelines and the National Institute for Health and Care Excellence 2019 guidelines.

As we defined in prompt 1, the extracted concepts were formatted using the Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) vocabulary within the Clinical Findings and Observations domain [40].

To mitigate potential output errors from LLMs, such as concepts not belonging to the Clinical Findings and Observations domain, or even errors outside of the SNOMED CT vocabulary, we

XSL•FO
RenderX

implemented a rule-based matching approach to filter extraction inaccuracies.

Furthermore, in this research, we aimed to extract concepts with diverse semantic expressions (including diagnoses, various medical histories, symptoms, observations, interventions, and types of examination). To accomplish this, local experts manually filtered out concepts embedded in structured text, such as dates or numbers.

Question Generation

After the extraction and aggregation of concepts, they were transformed into specific questions by LLMs as question templates for subsequent data extraction. In this section, we leveraged ChatGPT4.0 as a question generator to produce a basic set of questions, which were then refined by local experts for specificity based on its performance across 100 observations.

Q&A Scale Extraction

To avoid contextual and temporal event confusion leading to incorrect responses (eg, confusing current medical history with a past medical history or confusing the patient's medical history with that of family members), we preextracted the corpus using two strategies: (1) based on the position of the question templates and (2) based on the sentence containing the concepts. The extracted corpus was then labeled with the corresponding question templates for the subsequent extraction of Q&A scales.

The refined corpus, combined with corresponding question templates, guided a systematic extraction process with 2 LLMs, forming Q&A scales for further application.

Each question probed the LLMs, and the extracted sentences formed the basis of the generated responses. This approach enabled a logical mapping of questions to relevant text, ultimately improving the accuracy and efficiency of feature extraction.

Evaluation

Given the practical constraints and the objective of minimizing manual intervention, it was unfeasible to validate all answer scales individually across a Q&A space containing 68 questions and 25,709 observations. Therefore, a 3-fold assessment strategy was developed, as explained in the sections that follow.

Accuracy and Precision

A subset of 1500 observations chosen at random was manually annotated in collaboration with local experts, serving as the gold standard. The precision of positive identifications by both LLMs was assessed against a specified benchmark.

Null Ratio

The null ratio of both LLMs was independently measured across all 25,709 observations. Empty or meaningless outputs (symbols and gibberish) were identified as null outputs, and the null ratio was then calculated as the proportion of such responses to the total.

Time Consumption

The efficiency of the extraction process was evaluated by measuring the time taken by the 2 LLMs to respond to the questions across all 25,709 observations.

Ethical Considerations

The study was approved by the People's Hospital of the Guangxi Zhuang Autonomous Region in China (KT-KJT-2021-67). The requirement for informed consent was waived, due to the retrospective nature of the study, and all clinical data were deidentified and anonymized.

Results

Path Decomposition Overview

Prompt Template

Through trials with the prompt template on 100 observations, we selected the template that demonstrated optimal consistency, as shown in Figure 2.

XSL•FO
RenderX

Figure 2. Prompt templates of information extraction and question-and-answer scales. A 4-paragraph structure was implemented for a prompt design using a few-shot Chain-of-Thought prompting. The original version, written in Chinese, was translated into English. (A) Extracting concepts and terms; (B) scaling questions. As both deployed large language models were pretrained with the *International Classification of Diseases, Tenth Revision* corpus, they could directly engage in concept extraction.

(A)

```
Prompt=f'''
You are a stringent information extractor, you should work according to
operations meeting requirements, and referring to examples to extract
clinical medical concepts and terms from the text. You are very strict; you
will absolutely not associate or produce information that is not in the
original text.
===
#### Text: {}
===
#### Operations:
1. Decompose the text sentence by sentence, breaking it down according to the
independent meaning of each sentence. The number of breakdowns should be as
many as possible, forming individual information blocks.
2. If the word count of an information block exceeds 10 words, it needs to be
further decomposed until the length of each information block does not exceed
10 words.
3. For each information block, summarize its core meaning into a keyword or
concept.
4. Check these core concepts or keywords, ensuring that the length of each
concept does not exceed 10 words, and ensure they are semantically accurate.
5. List the standardized core concepts or keywords, and use '\\n' to separate
them.
===
#### Requirements:
Concepts and terms belong to diseases, symptoms, drugs, treatment methods,
test items, examination items, medical history, personal information. During
the execution process, pay attention to identifying synonyms or various
expressions of medical terms. Do not output details of the original text.
===
#### Output format:
Only need to output the result of step five. Start with "```row{}"
===
#### Sample:
Example 1:
Text: "The patient was diagnosed with Type 2 diabetes in 2021 and is
currently controlling blood sugar by using metformin and performing aerobic
exercise."
Output: " row1: Type 2 diabetes\\n Metformin\\n Aerobic exercise".
Example 2: "The patient is gravida 3, para 1, 39 weeks + 1 day pregnant."
Output: " row2: Gravida\\n Para\\n Gestational week"
'''
```

(B)

```
Prompt=f'''
You are a stringent clinical data scientist, working according to
operational steps, meeting requirements, referencing examples, extracting
and understanding medical information related to legs from given documents
in order to answer queries about it. Information not present in the original
text will absolutely not be inferred or generated.
===
#### Text: {}
===
#### Question: {}
===
#### Operations:
1. Identify the key term in the question.
2. Check if the key term appears in the text.
3. Determine whether the key term is negated in the text.
4. Output according to the requirements
===
===
#### Requirements: {}
===
#### Output format:
Begin with "```row{}" and end with "```"
===
#### Sample:
Example 1:
Question: "How is the patient's menstrual flow?"
Requirements: "Choose only: Light/Moderate/Heavy/Unknown"
Document: "...painless periods, moderate flow..."
Output: " row1: Moderate"
Example 2:
Question: "5. Does the patient have a history of alcohol consumption?"
Requirements: "Choose only: 'Yes' or 'No'. If there is a clear history of
smoking and drinking habits, answer 'Yes'; if the smoking and drinking
habits are denied, answer 'No'"
Document: "... Denies a history of smoking and drinking habits...."
Output: "row1: No"
'''
```

Merged Concepts

After merging all concepts, we filtered out those that appeared less than 5% of the time. A total of 117 concepts and terms were listed in Table S1 in Multimedia Appendix 2.

Question Template

Then we selected and transferred 68 concepts into question formats, for further Q&A scales. The detailed questions and their corresponding concepts are listed in Table S2 in Multimedia Appendix 2.

Scale Extraction

We identified that the optimal performance in Q&A scale extraction occurs with a concurrency of 3 requests, enhancing speed by 17.9% compared to a single request. Furthermore, we used a max_token restriction strategy, capping it at 20, to optimize inference speed.

Ultimately, within the 2D Q&A space formed by the answer scales, there were a total of 68 question columns and 25,709 observations (listed in Table 1).

We used accuracy and precision metrics for assessing the accuracy of LLMs across 1500 observations. Furthermore, we used 2 parameters—null ratio and time consumption—in 25,709 observations to evaluate the consistency and efficiency of the 2 LLMs, respectively.

XSL•FO
RenderX

Table 1. Question-and-answer scales for Qwen-14B-Chat (QWEN) and Baichuan2-13B-Chat (BAICHUAN).

Concepts	QWEN			BAICHUAN		
	Positive ratio (%)	Negative ratio (%)	Null ratio (%)	Positive ratio (%)	Negative ratio (%)	Null ratio (%)
Menstrual color	__[a]	—	0.00	—	—	0.67
Menstrual flow	—	—	0.00	—	—	0.80
Pregnancy weight gain[b]	—	—	1.60	—	—	2.53
Abdominal bloating	26.27	73.73	0.00	38.40	61.60	0.00
Abdominal pain	49.00	51.00	0.00	43.13	52.00	4.87
Amniocentesis	1.00	99.00	0.00	0.73	98.87	0.40
Aspirin use	1.40	98.60	0.00	1.40	98.60	0.00
Bilateral adnexal masses	4.87	95.13	0.00	0.73	99.20	0.07
Bilateral lower limb edema	51.53	48.47	0.00	4.87	95.13	0.00
Blood glucose screening	17.87	82.13	0.00	28.13	71.40	0.47
Cervical secretions	3.53	96.47	0.00	3.67	96.33	0.00
Chest tightness	2.87	97.13	0.00	4.40	95.53	0.07
Cold or flu	1.87	98.13	0.00	3.07	96.93	0.00
Convulsions	2.27	97.73	0.00	0.33	99.67	0.00
Dizziness	7.13	92.87	0.00	1.87	98.13	0.00
Drinking	0.00	100.00	0.00	0.00	100.00	0.00
Early pregnancy reaction or symptoms	85.93	14.07	0.00	47.53	52.13	0.33
Family history (asthma)	0.07	99.93	0.00	1.93	98.07	0.00
Family history (autoimmune disease)	0.20	99.80	0.00	3.07	96.93	0.00
Family history (diabetes mellitus)	1.67	98.33	0.00	2.33	97.67	0.00
Family history (drug allergy)	0.01	99.99	0.00	5.67	94.33	0.00
Family history (heart disease)	1.60	98.40	0.00	2.27	97.73	0.00
Family history (hematologic disease)	0.13	99.87	0.00	0.73	98.87	0.40
Family history (hypertension)	3.87	96.13	0.00	3.93	96.07	0.00
Family history (kidney disease)	0.07	99.93	0.00	1.40	98.53	0.07
Family history (mental illness)	0.07	99.93	0.00	0.47	99.47	0.07
Family history (neurological disease)	0.47	99.53	0.00	0.73	98.80	0.47
Family history (preeclampsia)	0.12	99.88	0.00	2.60	97.40	0.00
Family history (rheumatic disease)	0.07	99.93	0.00	1.47	98.53	0.00
Fetal paternal drinking history)	0.00	100.00	0.00	0.07	99.93	0.00
Fetal paternal history of genetic diseases	1.00	99.00	0.00	0.87	99.07	0.07
Fetal paternal smoking history	0.00	100.00	0.00	0.07	99.93	0.00
Fever	9.67	90.33	0.00	1.93	98.07	0.00
G6PD[c]	3.33	96.67	0.00	2.53	97.47	0.00
Headache	2.80	97.20	0.00	0.80	99.13	0.07
Insomnia	0.60	99.40	0.00	1.20	98.47	0.33
Mediterranean anemia screening	8.27	91.73	0.00	17.60	82.27	0.13
Palpitations	1.53	98.47	0.00	3.07	96.93	0.00
Personal history (antiphospholipid syndrome)	0.07	99.93	0.00	0.07	99.93	0.00
Personal history (chronic kidney disease)	0.80	99.20	0.00	1.07	98.93	0.00

XSL•FO
RenderX

Concepts	QWEN			BAICHUAN		
	Positive ratio (%)	Negative ratio (%)	Null ratio (%)	Positive ratio (%)	Negative ratio (%)	Null ratio (%)
Personal history (diabetes mellitus)	0.60	99.40	0.00	0.13	99.87	0.00
Personal history (drug allergy)	10.53	89.47	0.00	39.80	60.20	0.00
Personal history (dysmenorrhea)	24.40	75.60	0.00	21.20	78.80	0.00
Personal history (food allergy)	5.13	94.87	0.00	8.07	91.93	0.00
Personal history (heart disease)	1.67	98.33	0.00	0.47	99.53	0.00
Personal history (hematologic disease)	0.00	100.00	0.00	0.07	99.93	0.00
Personal history (hypertension)	7.40	92.60	0.00	0.93	99.07	0.00
Personal history (infectious disease)	1.93	98.07	0.00	3.80	96.20	0.00
Personal history (preeclampsia)	0.93	99.07	0.00	0.87	99.13	0.00
Personal history (surgery history)	35.67	64.33	0.00	36.27	63.67	0.07
Personal history (systemic lupus erythematosus)	0.20	99.80	0.00	0.20	99.80	0.00
Personal history (thalassemia)	1.00	99.00	0.00	0.80	99.13	0.07
Personal history (trauma history)	8.80	91.20	0.00	2.87	97.13	0.00
Personal history (viral hepatitis)	6.07	93.93	0.00	6.20	93.80	0.00
Poor pregnancy history (induced abortion)	0.07	99.93	0.00	0.13	99.87	0.00
Poor pregnancy history (miscarriage)	0.47	99.53	0.00	0.47	99.53	0.00
Poor pregnancy history (premature birth)	0.27	99.73	0.00	0.27	99.73	0.00
Prenatal screening	31.13	68.87	0.00	15.87	83.40	0.73
Regular prenatal check-ups	96.20	3.80	0.00	96.80	3.13	0.07
Smoking	0.07	99.93	0.00	0.07	99.93	0.00
Threatened abortion	6.20	93.80	0.00	5.80	94.00	0.20
Umbilical cord blood ratio	0.00	100.00	0.00	0.00	100.00	0.00
Use of antihypertensive drugs	1.73	98.27	0.00	2.13	97.87	0.00
Use of progestogen drugs	13.47	86.53	0.00	14.40	85.53	0.07
Vaginal bleeding	81.07	18.93	0.00	22.60	77.27	0.13
Vaginal discharge	33.00	67.00	0.00	48.47	51.53	0.00
Vaginal infection	25.27	74.73	0.00	16.20	82.73	1.07
Vaginal secretions	16.60	83.40	0.00	16.73	83.27	0.00

[a]Not applicable.

[b]The mean pregnancy weight gain was 13.73 (SD 24.12) for QWEN and 13.75 (SD 31.28) for BAICHUAN.

[c]G6PD: glucose-6-phosphate dehydrogenase.

Evaluation Metrics

Accuracy and Precision

Figure 3A and 3B and Figure S1 (parts A and B) in Multimedia Appendix 2 illustrate the Q&A space for a sample chunk extracted by QWEN and BAICHUAN with the comparison of manual annotation. The figures demonstrate the exceptional accuracy and precision of QWEN and BAICHUAN. QWEN attained an average accuracy of 95.52% and an average precision of 92.93%, whereas BAICHUAN displayed an average accuracy of 95.86% and an average precision of 90.08%. These figures clearly indicate that the 2 LLMs have more concentrated errors in specific concepts, and overall, they achieve high levels of precision in most extractions.

LLMs demonstrated consistent performance across most questions and excelled in binary, well-defined medical history questions, often reaching 100% accuracy and precision. However, the accuracy performance varied significantly when dealing with questions that involved semantic ambiguities or definitional uncertainties. This inconsistency might be tied to the LLM's training and inference alignment. Notable disparities were observed in questions pertaining to menstrual color (QWEN: 1000/1500, 66.7%; BAICHUAN: 1097/1500, 73.1%), early pregnancy symptoms (QWEN: 909/1500, 60.7%; BAICHUAN: 1474/1500, 98.3%), vaginal bleeding (QWEN:

XSL•FO
RenderX

593/1500, 39.5%; BAICHUAN: 1455/1500, 97%), bilateral lower limb edema (QWEN: 786/1500, 52.4%; BAICHUAN: 1486/1500, 99.7%), and menstrual flow (QWEN: 1498/1500, 99.8%; BAICHUAN: 605/1500, 40.3%).

Apart from the above, the precision inconsistency performance of concepts could be attributed to their low true positive rate, like insomnia (QWEN: 3/17, 17.7%; BAICHUAN: 8/17, 47.1%), personal history—antiphospholipid syndrome (QWEN: 0/2, 0%; BAICHUAN: 1/2, 50%), and poor pregnancy history—induced abortion (QWEN: 1/2, 50%; BAICHUAN: 2/2, 100%). The exact precision is listed in Table S3 in Multimedia Appendix 2.

Figure 3. Accuracy and precision in the question-and-answer space. With the local expert's annotation of 1500 observations, parts A and B showcase a comparison of the accuracy and precision of QWEN, BAICHUAN, and QWEN(INT4) across various concepts. Our findings reveal that the performance trends of large language models are nearly uniform across different concepts in terms of accuracy while showing a discernible variation in precision. G6PD: glucose-6-phosphate dehydrogenase.

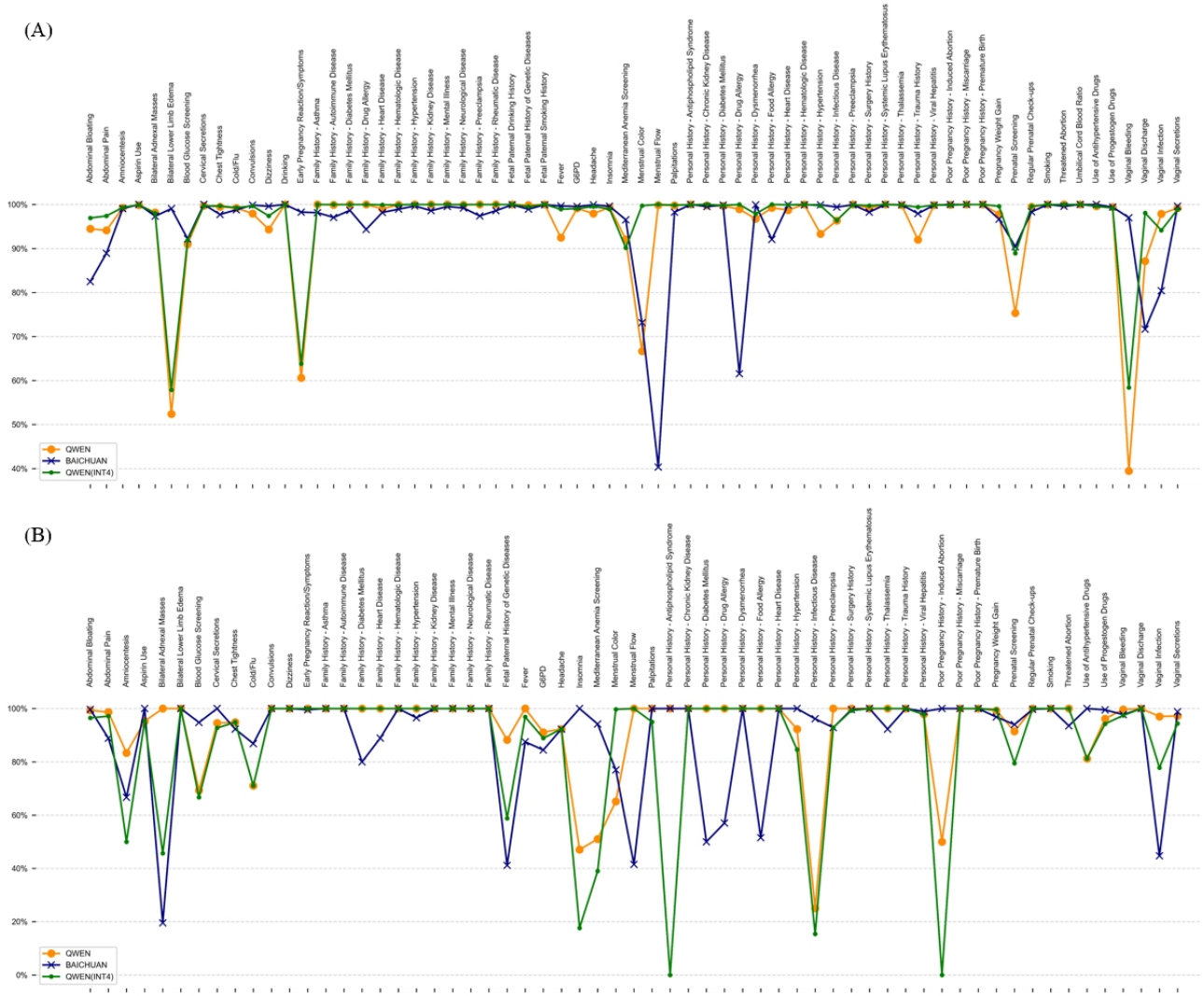

Null Ratio

As depicted in Table 1, both LLMs demonstrated superior performance with minimal null ratios. Specifically, QWEN (Figure S1A in Multimedia Appendix 2) exhibited a mean null ratio of 0.02%, in contrast to BAICHUAN (Figure S1B in Multimedia Appendix 2), which recorded a slightly higher null ratio of 0.21%. Failure of QWEN extraction was only in pregnancy weight gain (411/25,709, 1.60%), but failures of BAICHUAN extraction were mainly in symptoms (abdominal pain: 1252/25,707, 4.87%; vaginal infection: 275/25,709, 1.07%).

Time Consumption

We conducted a comparative analysis of the time performance between QWEN and BAICHUAN on various Q&A scales, discovering that BAICHUAN consistently exhibits higher time consumption across almost all scales, reaching up to 4 times that of QWEN, as illustrated in Figure 4B.

Figure 4A compares the time consumption of LLMs in extracting diverse concepts. Although there were significant differences across different concepts, overall, the LLMs demonstrated a consistent performance across these concepts. For queries with clear definitions and concise corpora, such as drug usage and previous pregnancy history, the time consumed was minimal. In the category of medical history, both models

XSL•FO
RenderX

exhibited uniform and stable performances (QWEN and BAICHUAN both revealed a time consumption ratio of 1:3).

Figure 4. Time consumption of question-and-answer (Q&A) scales for QWEN and BAICHUAN, measured in seconds. (A) Comparative distribution of time consumption for QWEN and BAICHUAN per concept, and QWEN exhibited notably lower extraction times across various Q&A scales compared to BAICHUAN. (B) Comparative distribution of the time consumption for 2 large language models per observation. G6PD: glucose-6-phosphate dehydrogenase.

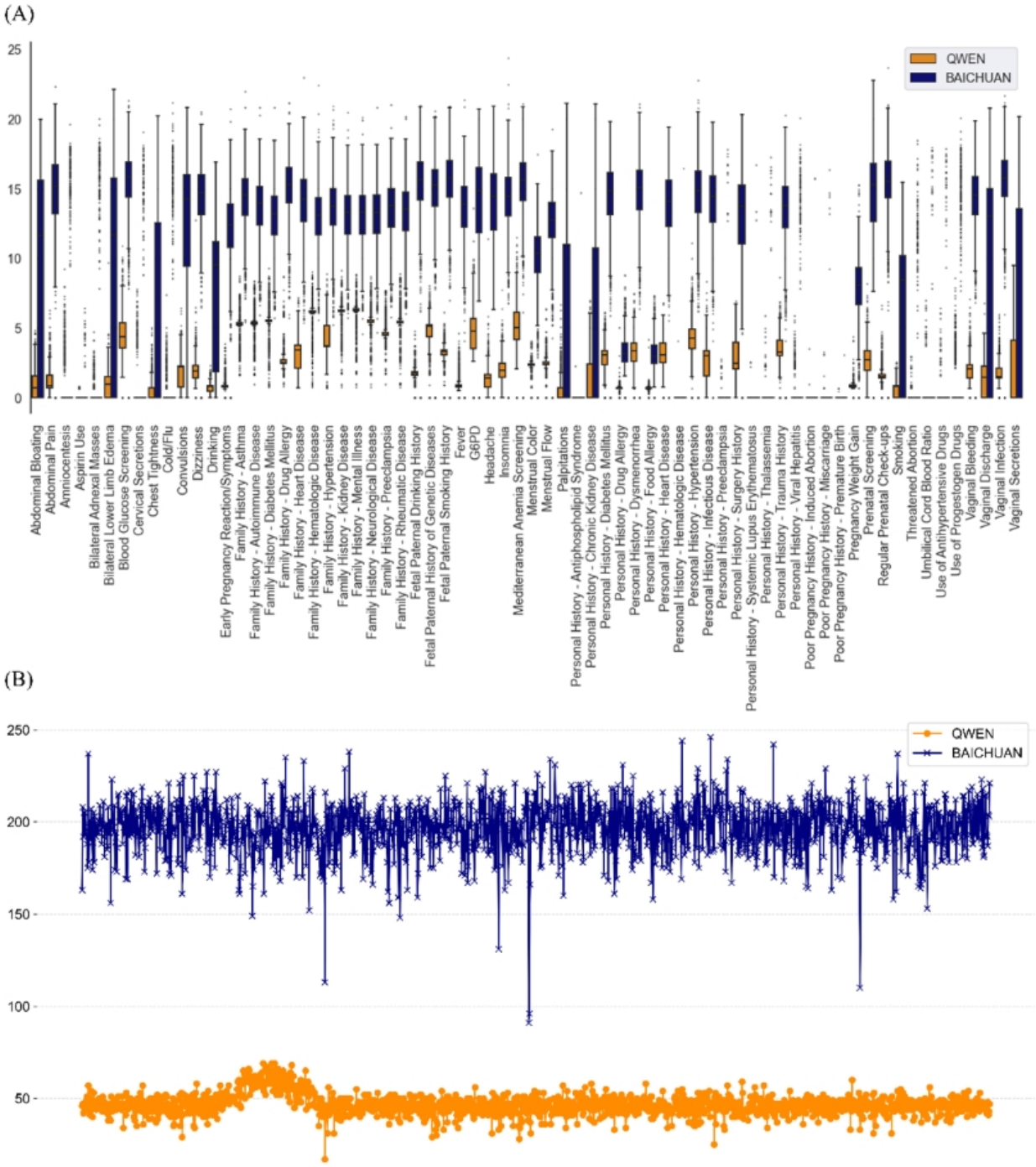

Additional Research

In clinical practice, to address scenarios of resource constraints, we used a quantified version of the LLM in our study to validate the applicability of this approach. We used an official-release INT4 version of QWEN, QWEN(INT4). The model was deployed on an NVIDIA RTX 3090 GPU (24 GB).

With the same approach listed above, the performance of QWEN(INT4) achieved even better performance. Figure 3 and Figure S1C in Multimedia Appendix 2 demonstrate that the average accuracy of QWEN(INT4) is 97.28%, accompanied by a null ratio of 0%.

Despite a notable correlation in performance extraction between QWEN(INT4) and QWEN, QWEN(INT4) demonstrated superior efficiency on limited hardware, with an average of 31 seconds per observation, compared to 47 seconds for QWEN and 312 seconds for BAICHUAN.

XSL•FO
RenderX

Discussion

Principal Findings

In this study, the extracted scales incorporated not only the conventional features of interest but also less frequently mentioned dimensions in previous cohorts or guidelines. These included food and drug allergies (6.6% for food allergy and 25.2% for drug allergy), certain pregnancy symptoms (average positive ratio of 0.9% for insomnia and 2.3% for palpitations), menstrual conditions (22.8% for dysmenorrhea), medical history (1% for asthma family history and 0.27% for mental illness family history), and gestational intervention (13.93% for progestogen and 1.4% for aspirin).

As a naturally recruited cohort of pregnancy, the extracted features show comparable proportions or trends compared with similar studies, such as systemic lupus erythematosus (average positive ratio of 0.20% vs 0.03%-0.23% of similar cohorts [41,42]) and antiphospholipid syndrome (average positive ratio of 0.08% vs 0.02%-0.12% of similar cohorts [43]), thereby corroborating the accuracy of our approach.

Additionally, certain scale deviations were revealed compared to similar studies, notably in fetal paternal smoking history (average positive ratio of 0.04% vs approximately 28.1%-40% in similar studies [44,45]). Although these deviations were few, we conducted a sample retracing to the original texts and determined that the extraction approach was not at fault and accurately reflected the original data. This discrepancy highlights persistent concerns [46] regarding the data quality in inpatient documentation, originating from patient self-reports and physician documentation, and vulnerability to recall and inquiry bias. Documentation varies among patients, influenced not only by patient conditions but also by physicians' writing habits. Thus, we regard our approach as a preexperimental data analysis. Despite the presence of biases or missing dimensions, the approach uncovers several dimensions absent in structured medical texts, and valuable insights could still be extracted from the data with appropriate statistics [47]. In clinical practice, preliminary interviews with documenting physicians are recommended prior to the selection of concepts to enhance data quality and mitigate potential biases.

In the context of the extraction process, even when deployed solely on a standard consumer-grade GPU (NVIDIA RTX 3090), the QWEN(INT4) completed the extraction of 25,709 observations and 68 features within 15 calendar days, averaging 48.9 seconds per observation. In practical applications, deploying 2 instances of QWEN(INT4) on a single graphics card, coupled with an additional deployment in CPU [36], is hypothesized to reduce the extraction to approximately 7 days. Furthermore, multi-GPU server clusters, prevalent in clinical environments, could markedly reduce processing times, potentially to the scale of hours.

In our study, we experimented with omitting the corpus extraction step, directly using the long text of each observation's chief complaints and medical histories as raw data for Q&A scale extraction. However, the experiment yielded poor performance in accuracy, precision, and time consumption, as illustrated in Figure 5 and Figure S2 in Multimedia Appendix 2.

These limitations appear significantly correlated with the current technological constraints of LLMs [32], which tend to generate "feature hallucinations" more frequently when processing extensive texts [48], leading to the loss of critical information. We believe that this issue will be resolved as the technology continues to evolve [49].

XSL•FO
RenderX

Figure 5. Approach performance when omitting corpus extraction step. (A) The average time consumed per question-and-answer (Q&A) interaction over 300 observations for both models. (B) Comparing the distribution of time consumption for QWEN and BAICHUAN in a single observation per Q&A scale. G6PD: glucose-6-phosphate dehydrogenase.

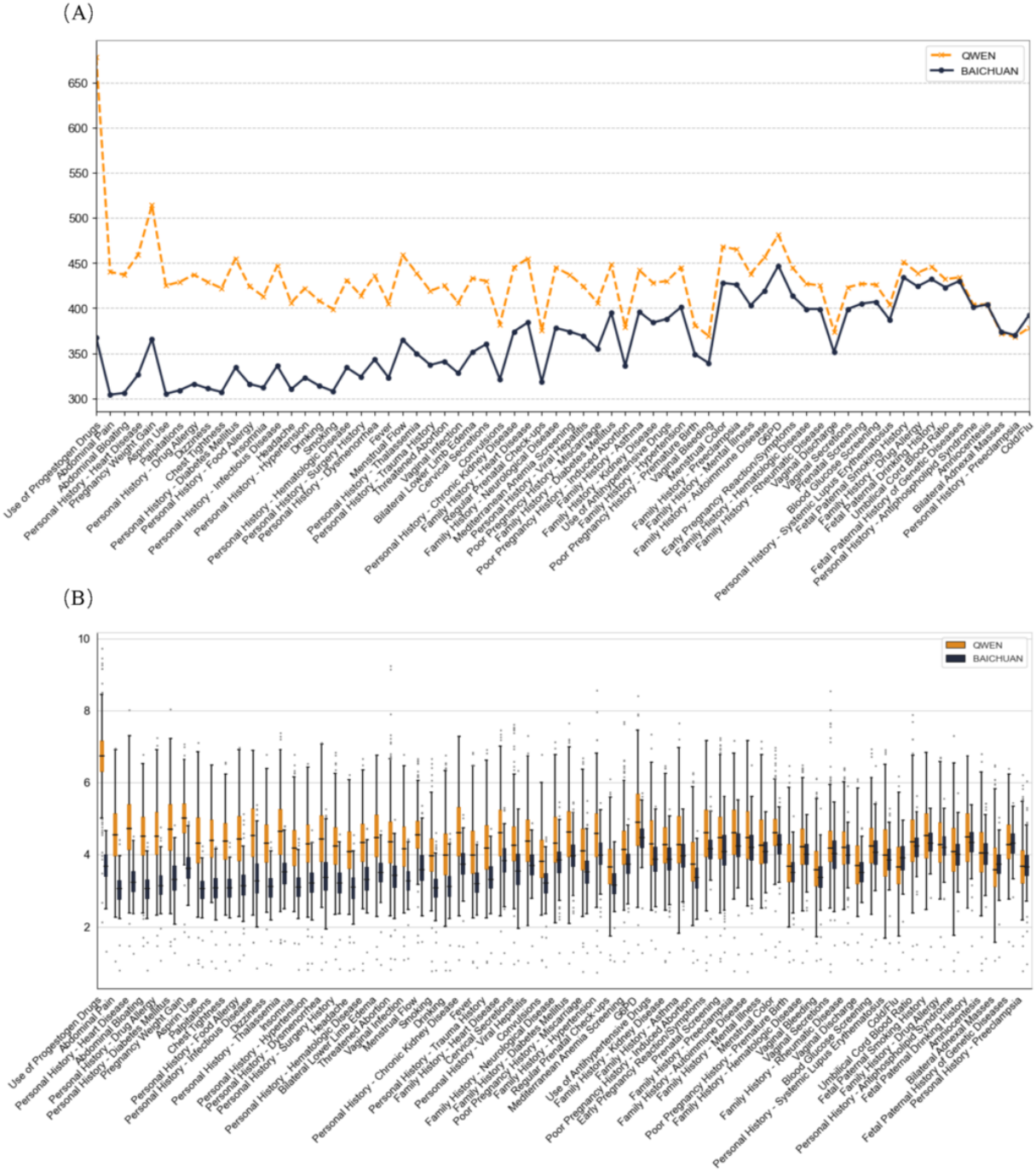

Limitations

In our experimental validation, we selected a limited set of concepts, comprising only 68 items, to balance the consideration of time constraints. Despite our efforts to encompass a broad scope, some dimensions inevitably remain unaddressed, which is a limitation in verifying efficiency and accuracy across all dimensions.

Furthermore, the raw data in this study was sourced exclusively from a single hospital, spanning nearly a decade. This duration, while significant, introduces limitations in the generalizability of our approach.

Additionally, the approach used only 2 LLMs. Although we anticipate that future LLMs will be compatible with the current approach, this assumption necessitates further experimental validation.

Conclusions

Our proposed approach offers a potential methodology for clinical text data analysis. It involves extracting and summarizing concepts from the comprehensive text of a defined population, thus selecting research directions of interest, and eventually generating analyzable features for the cohort. This approach demonstrates notable precision and could provide substantial data support for future research endeavors.

XSL•FO
RenderX

Acknowledgments

We are grateful to Yinyao Ma and the clinical team for their exceptional contributions to this project and we thank the technical support provided by China National GeneBank.

This work was supported by Guangxi Key Research and Development Program (AB22035056).

Data Availability

The data sets generated and analyzed during this study are not publicly available due to privacy or ethical restrictions but are available on request from the corresponding author.

Authors' Contributions

LW and YM contributed equally to this study. LW, HL, and YM participated in the study design and drafted the manuscript. YL and HL participated in data collection and outcome rule review. LW and WB performed the statistical analysis, and established machine learning models. YL helped to draft the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A sample of chief complaints and medical histories.
[DOCX File , 18 KB - jmir_v26i1e54580_app1.docx]

Multimedia Appendix 2
Additional statistics.
[DOCX File , 873 KB - jmir_v26i1e54580_app2.docx]

References

1. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Stats 2021 Feb 14;13(6):1-19. [doi: 10.1002/wics.1549]
2. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:4302425. [doi: 10.1155/2018/4302425] [Medline: 29849998]
3. Varshini K, Uthra R. An approach to extract meaningful dData from unstructured clinical notes. In: Inventive Systems and Control. Singapore: Springer; Jun 08, 2021:581-590.
4. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
5. Wu P, Yu A, Tsai C, Koh J, Kuo C, Chen ALP. Keyword extraction and structuralization of medical reports. Health Inf Sci Syst 2020 Dec 03;8(1):18 [FREE Full text] [doi: 10.1007/s13755-020-00108-6] [Medline: 32269770]
6. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof 2013 Nov 30;10:12 [FREE Full text] [doi: 10.3352/jeehp.2013.10.12] [Medline: 24324853]
7. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. Am J Manag Care 2002 Sep;8(9):787-793 [FREE Full text] [Medline: 12234019]
8. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. Eval Health Prof 2009 Sep 13;32(3):281-298. [doi: 10.1177/0163278709338561] [Medline: 19679636]
9. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health 2022 Dec 22;1(12):e0000168 [FREE Full text] [doi: 10.1371/journal.pdig.0000168] [Medline: 36812634]
10. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform 2009 Oct;42(5):923-936 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.007] [Medline: 19646551]
11. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR 2020;21(140):1-67.
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on Oct, 2018. [FREE Full text]
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]

XSL•FO
RenderX

14. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
15. Wang S, Yilahun H, Hamdulla A. Biomedical named entity recognition based on MCBERT. 2022 Presented at: 2022 International Conference on Asian Language Processing (IALP); October 27-28, 2022; Singapore. [doi: 10.1109/ialp57159.2022.9961297]
16. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, et al. Use of BERT (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. J Med Internet Res 2021 Jan 12;23(1):e19689 [FREE Full text] [doi: 10.2196/19689] [Medline: 33433395]
17. Zhang Z, Zhu L, Yu P. Multi-level representation learning for Chinese medical entity recognition: model development and validation. JMIR Med Inform 2020 May 04;8(5):e17637 [FREE Full text] [doi: 10.2196/17637] [Medline: 32364514]
18. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
19. Polnaszek B, Gilmore-Bykovskyi A, Hovanes M, Roiland R, Ferguson P, Brown R, et al. Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction. Med Care 2016 Oct;54(10):e65-e72 [FREE Full text] [doi: 10.1097/MLR.0000000000000108] [Medline: 27624585]
20. Tarik AM, Sorin E, Symons J, Mayer E, Yaliraki S, Toni F. Extracting information from free text through unsupervised graph-based clustering: an application to patient incident records. arXiv Preprint posted online on Aug 31, 2018. [FREE Full text] [doi: 10.48550/arXiv.1909.00183]
21. Choi HS, Song JY, Shin KH, Chang JH, Jang B. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023 Sep;41(3):209-216 [FREE Full text] [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
22. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 02;6(10):e2336997 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
23. Johnson S, King A, Warner E, Aneja S, Kann B, Bylund C. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023 Mar 01;7(2):a [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
24. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
25. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs 2024 Mar 12;23(2):122-126. [doi: 10.1093/eurjcn/zvad087] [Medline: 37603843]
26. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580 [FREE Full text] [doi: 10.2196/51580] [Medline: 38009003]
27. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng 2023 Dec 16;51(12):2647-2651. [doi: 10.1007/s10439-023-03284-0] [Medline: 37328703]
28. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023 Sep 11;25:e49240 [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]
29. Zhang L, Liu X, Li Z, Pan X, Dong P, Fan R. Dissecting the runtime performance of the training, fine-tuning, and inference of large language models. arXiv Preprint posted online on Nov 2023. [FREE Full text]
30. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J. A survey on large language model based autonomous agents. arXiv ;01 Preprint posted online on Aug 22, 2023. [FREE Full text] [doi: 10.48550/arXiv.2308.11432]
31. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E. Chain-of-thought prompting elicits reasoning in large language models. arXiv Preprint posted online on Jan 28, 2022. [doi: 10.48550/arXiv.2201.11903]
32. Lei D, Li Y, Hu M, Wang M, Yun V, Ching E. Chain of natural language inference for reducing large language model ungrounded hallucinations. arXiv Preprint posted online on Oct 6, 2023. [FREE Full text] [doi: 10.48550/arXiv.2310.03951]
33. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S. Self-consistency improves chain of thought reasoning in language models. arXiv Preprint posted online on Mar 21, 2022. [FREE Full text]
34. ACOG committee opinion no. 743: low-dose aspirin use during pregnancy. Obstet Gynecol 2018 Jul;132(1):e44-e52. [doi: 10.1097/AOG.0000000000002708] [Medline: 29939940]
35. National IFHEG. National Institute for Health and Care Excellence (NICE). In: The Grants Register 2019. London, UK: Palgrave Macmillan; 2019:540.
36. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen Technical Report. arXiv ;01 Preprint posted online on Sep 2023. [FREE Full text] [doi: 10.48550/arXiv.2309.16609]

XSL•FO
RenderX

37. Yang A, Xiao B, Wang B, Zhang B, Bian C, Yin C. Baichuan 2: open large-scale language models. arXiv Preprint posted online on Sep 19, 2023. [FREE Full text]
38. OpenAI. URL: https://openai.com/blog/openai-api [accessed 2024-03-12]
39. Zheng L, Chiang W, Sheng Y, Zhuang S, Wu Z, Zhuang Y. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv ;01 Preprint posted online on Jun 9, 2023. [FREE Full text] [doi: 10.48550/arXiv.2306.05685]
40. SNOMED International. URL: http://www.ihtsdo.org/snomed-ct/ [accessed 2024-03-13]
41. Clowse ME, Jamison M, Myers E, James AH. A national study of the complications of lupus in pregnancy. Am J Obstet Gynecol 2008 Aug;199(2):127.e1-127.e6 [FREE Full text] [doi: 10.1016/j.ajog.2008.03.012] [Medline: 18456233]
42. Rees F, Doherty M, Grainge M, Lanyon P, Zhang W. The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies. Rheumatology (Oxford) 2017 Nov 01;56(11):1945-1961. [doi: 10.1093/rheumatology/kex260] [Medline: 28968809]
43. Hwang J, Shin S, Kim Y, Oh Y, Lee S, Kim YH, et al. Epidemiology of antiphospholipid syndrome in Korea: a nationwide population-based study. J Korean Med Sci 2020 Feb 10;35(5):e35 [FREE Full text] [doi: 10.3346/jkms.2020.35.e35] [Medline: 32030922]
44. Xu X, Rao Y, Wang L, Liu S, Guo JJ, Sharma M, et al. Smoking in pregnancy: a cross-sectional study in China. Tob Induc Dis 2017 Jul 24;15(1):35 [FREE Full text] [doi: 10.1186/s12971-017-0140-0] [Medline: 28747859]
45. Yang Y, Liu F, Wang L, Li Q, Wang X, Chen JC, et al. Association of husband smoking with wife's hypertension status in over 5 million Chinese females aged 20 to 49 years. J Am Heart Assoc 2017 Mar 20;6(3):e004924 [FREE Full text] [doi: 10.1161/JAHA.116.004924] [Medline: 28320748]
46. Leon N, Balakrishna Y, Hohlfeld A, Odendaal W, Schmidt B, Zweigenthal V, et al. Routine Health Information System (RHIS) improvements for strengthened health system management. Cochrane Database Syst Rev 2020 Aug 13;8(8):CD012012 [FREE Full text] [doi: 10.1002/14651858.CD012012.pub2] [Medline: 32803893]
47. Miñarro-Giménez JA, Cornet R, Jaulent M, Dewenter H, Thun S, Gøeg KR, et al. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform 2019 Mar;123:37-48 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.12.011] [Medline: 30654902]
48. Rawte V, Chakraborty S, Pathak A, Sarkar A, Towhidul ITS, Chadha A. The troubling emergence of hallucination in large language models -- an extensive definition, quantification, and prescriptive remediations. arXiv ;01 Preprint posted online on Oct 8, 2023. [FREE Full text] [doi: 10.48550/arXiv.2310.04988]
49. Jones E, Palangi H, Simões C, Chandrasekaran V, Mukherjee S, Mitra A. Teaching language models to hallucinate less with synthetic tasks. arXiv ;01 Preprint posted online on Oct 10, 2023. [FREE Full text] [doi: 10.48550/arXiv.2310.06827]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CoT: Chain of Thought
GPU: graphics processing unit
LLM: large language model
Q&A: question and answer
SNOMED CT: Systematised Nomenclature of Medicine Clinical Term

Edited by Q Jin; submitted 15.11.23; peer-reviewed by X Tannier, M Torii; comments to author 12.01.24; revised version received 23.01.24; accepted 14.02.24; published 29.03.24.

Please cite as:
Wang L, Ma Y, Bi W, Lv H, Li Y
An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study
J Med Internet Res 2024;26:e54580
URL: https://www.jmir.org/2024/1/e54580
doi:10.2196/54580
PMID:38551633

©Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, Yuxiang Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study

Joseph Mugaanyi[1*], MBBS, MD; Liuying Cai[2*], MPhil; Sumei Cheng[2], PhD; Caide Lu[1], MD, PhD; Jing Huang[1], MD, PhD

[1]Department of Hepato-Pancreato-Biliary Surgery, Ningbo Medical Center Lihuili Hospital, Health Science Center, Ningbo University, Ningbo, China
[2]Institute of Philosophy, Shanghai Academy of Social Sciences, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Jing Huang, MD, PhD
Department of Hepato-Pancreato-Biliary Surgery, Ningbo Medical Center Lihuili Hospital
Health Science Center
Ningbo University
No 1111 Jiangnan Road
Ningbo, 315000
China
Phone: 86 13819803591
Email: huangjingonline@163.com

Abstract

Background: Large language models (LLMs) have gained prominence since the release of ChatGPT in late 2022.

Objective: The aim of this study was to assess the accuracy of citations and references generated by ChatGPT (GPT-3.5) in two distinct academic domains: the natural sciences and humanities.

Methods: Two researchers independently prompted ChatGPT to write an introduction section for a manuscript and include citations; they then evaluated the accuracy of the citations and Digital Object Identifiers (DOIs). Results were compared between the two disciplines.

Results: Ten topics were included, including 5 in the natural sciences and 5 in the humanities. A total of 102 citations were generated, with 55 in the natural sciences and 47 in the humanities. Among these, 40 citations (72.7%) in the natural sciences and 36 citations (76.6%) in the humanities were confirmed to exist (P=.42). There were significant disparities found in DOI presence in the natural sciences (39/55, 70.9%) and the humanities (18/47, 38.3%), along with significant differences in accuracy between the two disciplines (18/55, 32.7% vs 4/47, 8.5%). DOI hallucination was more prevalent in the humanities (42/55, 89.4%). The Levenshtein distance was significantly higher in the humanities than in the natural sciences, reflecting the lower DOI accuracy.

Conclusions: ChatGPT's performance in generating citations and references varies across disciplines. Differences in DOI standards and disciplinary nuances contribute to performance variations. Researchers should consider the strengths and limitations of artificial intelligence writing tools with respect to citation accuracy. The use of domain-specific models may enhance accuracy.

(J Med Internet Res 2024;26:e52935) doi:10.2196/52935

KEYWORDS

large language models; accuracy; academic writing; AI; cross-disciplinary evaluation; scholarly writing; ChatGPT; GPT-3.5; writing tool; scholarly; academic discourse; LLMs; machine learning algorithms; NLP; natural language processing; citations; references; natural science; humanities; chatbot; artificial intelligence

XSL•FO
RenderX

Introduction

Background

In the ever-evolving landscape of scholarly research and academic discourse, the role of technology in aiding and enhancing the research process has grown exponentially. One of the most notable advancements in this regard is the emergence of large language models (LLMs) such as GPT-3.5, which have demonstrated impressive capabilities in generating written content across various domains, including academic writing. These LLMs, powered by vast corpora of text data and sophisticated machine-learning algorithms, have offered researchers and writers a new tool for assistance in crafting scholarly documents [1-3]. LLMs were initially designed and developed to primarily assist in natural language writing. However, since the release of ChatGPT in late 2022, the tool has been adopted in a wide range of scenarios, including customer care, expert systems, as well as literature searches and academic writing. Researchers have already used LLMs to write their academic papers, as demonstrated by Kishony and Ifargan [4]. While the potential of these tools is evident, it is essential to critically assess their performance, especially in the intricate domains of citations and references, which are the foundation of academic discourse and credibility.

Citations and references serve as the backbone of scholarly communication, providing the necessary context, evidence, and credit to prior works, thus fostering intellectual dialogue and ensuring the integrity of the research process. Accuracy in generating citations and the inclusion of Digital Object Identifiers (DOIs) [5] are paramount, as they directly influence the traceability and accessibility of cited works. Despite the promise of LLMs, concerns have emerged regarding the reliability and precision of their generated citations and references, raising questions about their suitability as academic writing assistants. Studies on the viability of LLMs as writing assistants in scholarly writing [6-8] underscore the significance of this body of research within the broader academic landscape. Although prior works are quite informative [9-12], there is a lack of an interdisciplinary perspective on citations and references generated by LLMs, which is vital for understanding how LLMs perform across different disciplines.

An increasing number of academics and researchers, especially in countries where English is not a first language (eg, China), are relying on ChatGPT to translate their work into English, research the existing published literature, and even generate citations and references to published literature. Therefore, the aim of this study was to evaluate LLM performance in generating citations and references across two distinct domains, the natural sciences and humanities, by assessing both the presence and accuracy of citations, the existence and accuracy of DOIs, and the potential for hallucination. We aim to provide valuable insights into the strengths and limitations of LLMs in supporting academic writing in diverse research contexts.

The outcomes of this study will contribute to a nuanced understanding of the capabilities and limitations of LLMs as academic writing assistants. Moreover, our findings may inform best practices for researchers and writers who employ these tools in their work, fostering transparency and accuracy in scholarly communication.

LLM Concepts

An LLM is a catch-all term for a machine-learning model designed and trained to understand and generate natural language. LLMs are considered “large” language models due to the sheer number of parameters in the model. A parameter in machine learning is a numerical variable or weight that is optimized through training to map a relationship between the input and the output. LLMs have millions to billions of parameters.

Current LLMs are mostly based on the transformer architecture (Figure 1). However, before transformers were introduced in 2017 [13], recurrent neural nets (RNNs) were mostly used for natural language processing. One key limitation of RNNs was the length of text they could handle. In 2015, Bahdanau et al [14] proposed accounting for attention to improve RNN performance with long text. Drawing inspiration for the RNN’s encoder-decoder design, the transformer consists of an encoder and a decoder; however, unlike the RNN, the transformer does not perform sequential data processing and each layer can address all other layers. This allows the transformer model to handle different parts of the input as it processes each part at different stages. This is the mechanism that allows for self-attention in the transformer model.

The way attention works in a transformer model is by computing attention weights for each token, and then the relevance of the token is determined based on the weights. This allows the model to track and assign hierarchical values to each token. Fundamentally, this is similar to how humans process language by extracting the key details out of a chunk of text. This architecture is the linchpin for the majority of LLMs, including the GPT model [15] that is the basis of OpenAI’s ChatGPT or the bidirectional encoder representations from transformers (BERT) algorithm [16]. These are broadly categorized into encoder-style and decoder-style transformers, with the former mostly applying to predictive tasks and the latter applying to generative tasks.

Irrespective of the architecture, as an encoder-style or decoder-style transformer, the model is trained on a vast volume of data. The objective is to train a model capable of applying the knowledge gained from the training data to unseen data or situations. This is referred to as generalization. If the model is capable of precise recall of data it has previously been exposed to, this would be memorization and overfitting is said to have occurred. However, this does not mean that memorization is in itself a negative feature. Indeed, there are situations where memorization is preferable to generation such as in the task of information cataloging.

XSL•FO
RenderX

Figure 1. Transformer model architecture (left) and GPT architecture (right).

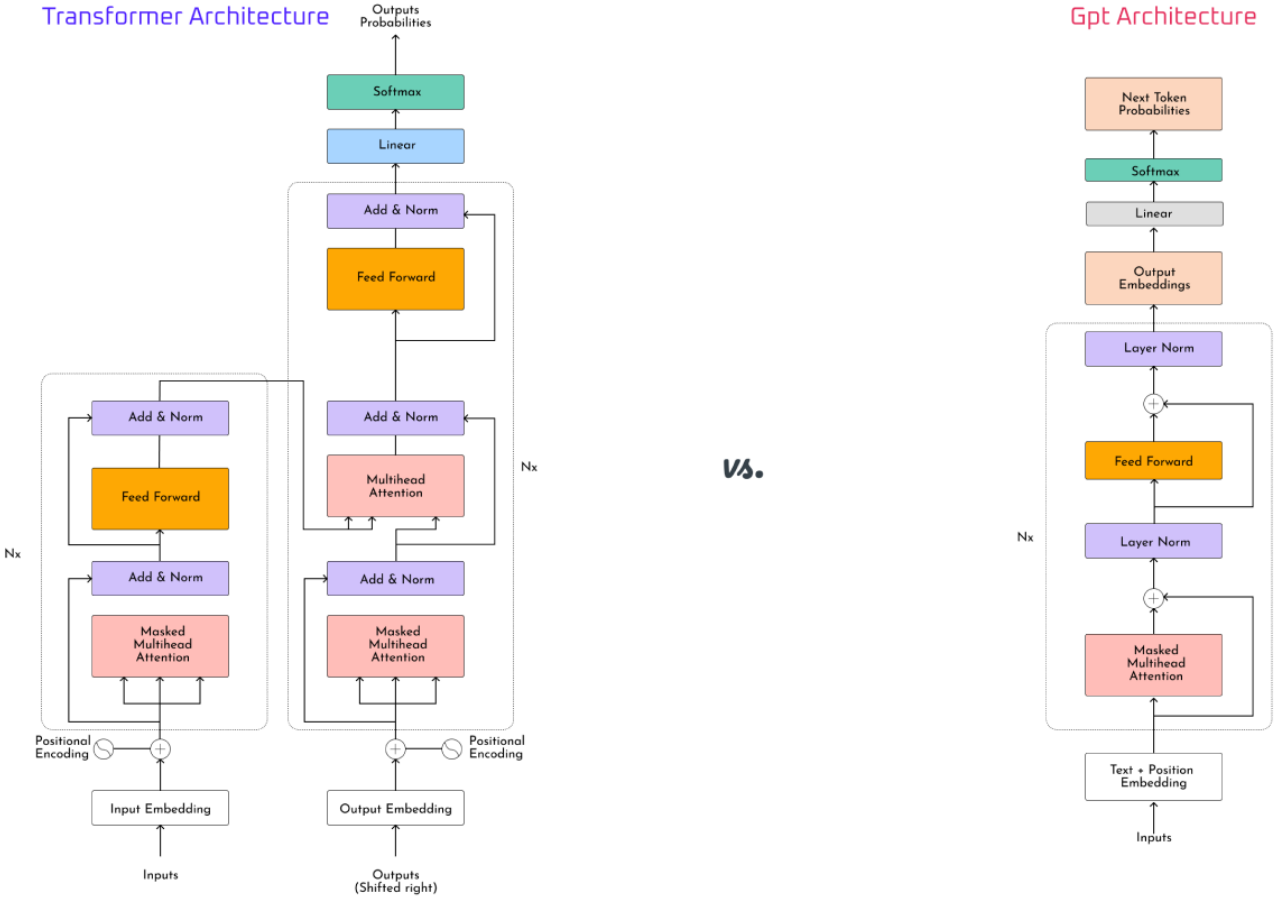

LLMs in Academia

LLMs can handle tasks such as text classification, translation, summarization, and text generation. Since the advent of the internet, and with it the publication of scientific information online, the amount of global academic output exploded, with more than 5 million articles published in 2022 (Table 1). Given the pressure in academia to keep up with developments in one's field, it is increasingly becoming more difficult to track, prioritize, and keep up with scientific information. It is against this backdrop that LLMs offer an opportunity. Perhaps the most obvious use case is in literature reviews and summarization, reference lookup, and data generation.

Table 1. Number of academic papers published per year, 2018-2022.

Year	Number of articles published (millions)
2018	4.18
2019	4.43
2020	4.68
2021	5.03
2022	5.14

However, there are still several questions that need to be answered. First, machine-learning models are inherently probabilistic, meaning that they are not deterministic. Therefore, for the same user input, the model may give different results due to the variability baked into the model. While this can be a valuable trait for creative endeavors, in academic and scientific works, there is a need for reproducibility and reliability, and it remains unclear how well this can be achieved. Second, LLMs are constrained to the information they are trained on. This can be affected by selection bias, the quality of data used, artifacts resulting from data cleaning, and other factors. In essence, we rely on trusting the trainer to provide accurate and unbiased training data to the models.

There is potential for LLMs to be useful tools for delivering academic and scientific information to various audiences, including—but not limited to—students and other academics. However, for this use case, a degree of memorization of the underlying content is necessary. Where information is unviable, it would be better to state so rather than to interpolate. In the current iteration of LLMs, since the training is geared toward generalization and the models are probabilistic, they tend to interpolate and fill in the missing information with synthetic

XSL•FO
RenderX

text. There is still a need to explore this process deeper to find solutions.

Methods

Data Collection and Validation

Topics were selected and categorized as either natural sciences or humanities. Topics were included if they were: (1) clinical or biomedical–related research in the natural sciences category and philosophy/psychology-related research in the humanities category, and (2) published in English. Topics were excluded if they were: (1) not in English, (2) related to a highly specialized or niche field, and (3) sensitive or controversial in nature. Two researchers independently prompted ChatGPT (GPT-3.5) to write sections of a manuscript while adhering to the American Psychological Association style [17] for citations and including the DOI of each reference. Citations and references generated by ChatGPT were collected for subsequent analysis. The researchers then independently validated the references by conducting searches on Google Scholar, PubMed, and Google Search for each cited reference. The primary objective was to confirm the existence and accuracy of the cited literature. DOI existence and validation were confirmed using the DOI Foundation website [18]. DOIs that did not exist or were matched to a different source were considered hallucinations [19]. Data collected by both researchers were aggregated and compared. Independent validation was performed to ensure agreement between the two researchers regarding the existence, validity, and accuracy of the citations and DOIs. Any disagreements or discrepancies were resolved through discussion and consensus.

In this study, hallucination refers to instances where ChatGPT 3.5 generates DOIs and/or citations that do not correspond to actual, valid DOIs/citations for scholarly references. In these instances, the model may produce DOIs and/or citations that seem authentic but are in fact incorrect or nonexistent. The Levenshtein distance, also known as the edit distance, is a measure of the similarity between two strings by calculating the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into the other. In other words, this metric quantifies the “distance” between two strings in terms of the minimum number of operations needed to make them identical. We used the Levenshtein distance to compare the DOI generated by ChatGPT with the correct DOI. This comparison helps to measure how closely the artificial intelligence (AI)–generated DOI aligns with the expected DOI for a given citation. By calculating the Levenshtein distance, we can quantify the differences between the AI-generated DOI and the correct DOI. Larger Levenshtein distance values suggest greater dissimilarity, indicating potential inaccuracies in the AI-generated DOI.

Statistical Analysis

Data analysis was conducted using SPSS 26 and Python. The Levenshtein distance [20] between the generated DOI and the actual DOI was calculated using the *thefuzz* package in Python to quantitatively assess the DOI accuracy. Continuous variables are reported as mean (SD) and categorical variables are presented as absolute numbers and percentages. An independent-sample *t* test was used to compare continuous variables, whereas the Fisher exact test was used for comparisons of categorical variables. A *P* value <.05 was considered statistically significant in all tests.

Ethical Considerations

This study was exempt from ethical review since no animal or human participants were involved.

Results

Included Topics and Citations

Ten manuscript topics were selected and included in the study, with 5 in the natural sciences group and 5 in the humanities group. ChatGPT 3.5 was prompted to write an introduction section for each topic between July 10 and August 15, 2023. A total of 102 citations were generated by ChatGPT. Of these, 55 were in the natural sciences group and 47 in the humanities group. The existence, validity, and relevance of citations were examined irrespective of the corresponding DOIs. The results are summarized in Table 2. A list of the included topics and a sample of prompts to ChatGPT are provided in Multimedia Appendix 1.

Table 2. Data analysis results.

Variables	Natural sciences (n=55)	Humanities (n=47)	*P* value[a]
Citation exists, n (%)	40 (72.7)	36 (76.6)	.42
Citation accurate, n (%)	37 (67.3)	29 (61.7)	.35
Relevant, n (%)	39 (70.9)	35 (74.5)	.43
DOI[b] exists, n (%)	39 (70.9)	18 (38.3)	.001
DOI accurate, n (%)	18 (32.7)	4 (8.5)	.003
DOI hallucination, n (%)	34 (61.8)	42 (89.4)	.001
Levenshtein distance, mean (SD)	64.13 (42.26)	42.15 (40.23)	.009

[a]Categorical variables were compared using the Fisher exact test; the continuous variable (Levenshtein distance) was compared using the independent-sample *t* test.

[b]DOI: Digital Object Identifier.

XSL•FO
RenderX

Citation Existence and Accuracy

Of the 102 generated citations, 76 (74.5%) were found to be real and exist in the published literature, with 72.7% and 76.6% of the citations verified in the natural and humanities group, respectively. There was no significant difference between the two groups (*P*=.42), indicating that the validity of the citations was relatively consistent between the two domains. Similarly, when assessing the accuracy of the citations, no significant difference was observed (Table 2).

Citation Relevance

The relevance of citations generated by ChatGPT was evaluated by assessing whether they were appropriate and contextually meaningful within the research topics. Our analysis indicated that 70.9% and 74.5% of citations in the natural sciences and humanities categories were deemed relevant, respectively (Table 2). The difference was not statistically significant (*P*=.43), suggesting that ChatGPT demonstrated a similar ability to generate contextually relevant citations in both domains.

DOI Existence, Accuracy, and Hallucination

Our analysis revealed significant differences between the two domains with respect to DOIs. In the natural sciences, 70.9% of the included DOIs were real, whereas in the humanities, only 38.3% of the DOIs generated were real (*P*=.001; Table 2). Similarly, the level of DOI accuracy was significantly higher for the natural sciences than for the humanities (*P*=.003). Moreover, the occurrence of DOI hallucination, where ChatGPT generates DOIs that do not correspond with the existing literature, was more prevalent in the humanities than in the natural sciences (*P*=.001). The mean Levenshtein distance, which measures the deviation between the generated DOI and the actual DOI, was significantly higher in the natural sciences group than in the humanities (*P*=.009; Table 2).

Discussion

Principal Findings

The results of this study shed light on the performance of ChatGPT (GPT-3.5) as an academic writing assistant in generating citations and references in natural sciences and humanities topics. Our findings reveal notable differences in the accuracy and reliability of the citations and references generated by ChatGPT when applied to natural sciences and humanities topics. Hallucination in the context of LLMs such as ChatGPT refers to a phenomenon where the model generates content that is incorrect, fabricated, or not grounded in reality. Hallucination occurs when the model produces information that appears plausible or contextually relevant but lacks accuracy or fidelity to real-world knowledge.

The most striking observation was the significant disparity in the existence and accuracy of the DOIs between the two domains. In natural sciences topics, DOIs were real in 70.9% of the generated citations, representing a significantly higher rate compared to the low rate of 38.3% real DOIs in the humanities topics. The discrepancies in the DOI existence and accuracy in the two domains may be attributed to the differential adoption and availability of DOIs across academic disciplines, where the natural sciences literature has often been more proactive in adopting the DOI system of referencing and linking to scholarly works than the humanities. It is a general practice that journals publishing on the natural sciences frequently mandate DOI inclusion, whereas publishers in the humanities have been slower to adopt such standards [21,22]. Consequently, the performance of the ChatGPT LLM in generating accurate DOIs appears to reflect these disciplinary disparities.

LLMs may generate fictional “facts” presented as true “real-world facts,” which is referred to as hallucination [19,23]. In this study, we considered hallucination to have occurred if the DOI of the generated citation was not real or was real but was linked to a different source. DOI hallucination was more frequent in the humanities (89.4%) than in the natural sciences (61.8%). This finding may be explained by the broader and less structured nature of the humanities literature. There is also a high tendency to provide citations from books and other media that do not use DOIs in the humanities. Therefore, researchers in the humanities should not consider DOIs generated by ChatGPT. Even when ChatGPT generates DOIs for humanities citations, they are more likely to deviate from the correct DOI, potentially leading to the inability to access the cited sources and use the DOIs in citation management tools such as EndNote.

In contrast to the disparities observed in DOI-related metrics, our study found a remarkable consistency in the existence, validity, and relevance of the generated citations in the natural sciences and humanities, with real citations found 72.7% and 76.6% of the time and accurate citations confirmed in 67.3% and 61.7% of cases, respectively. This suggests that the citations generated by ChatGPT can be expected to be reliable approximately 60% of the time.

The divergent performance of ChatGPT between the natural sciences and humanities underscores the importance of considering disciplinary nuances when implementing AI-driven writing assistants in academic contexts. Researchers and writers in both domains should be aware of the strengths and limitations of such tools, particularly in relation to citation practices and DOI accuracy. Future research could delve deeper into the factors influencing DOI accuracy and explore strategies for improving DOI generation by LLMs in the humanities literature. Additionally, the development of domain-specific AI writing models may offer tailored solutions to enhance citation and reference accuracy in various academic disciplines.

In this study, we focused only on the potential use of LLMs in citations and references in scholarly writing; however, the scope to which these models are going to be adopted in academic works is much broader. We believe that these models will be improved over time and that they are here to stay. As such, our argument in this paper is not that LLMs should not be used in scholarly writing, but rather that in their iteration, we ought to be aware of their limitations, primarily concerning the reliability of not only the text they generate but also how they interpret that text.

Although the transformer models that are the foundation of LLMs are very capable of handling a significant amount of information, they still do have context-window limitations. The context window is the textual range or span of the input that the

XSL•FO
RenderX

LLM can evaluate to generate a response at any given moment. As an example, GPT-3 has a context window of 2000 tokens, whereas GPT-4's context window is 32,000 tokens. As such, since the size of the context window impacts model performance (larger is better), GPT-4 outperforms GPT-3 (at the cost of more computation and memory). In scientific knowledge, context is key. Removing a word from the context may greatly affect the information being conveyed. Therefore, we believe that the future of LLMs in academia will rely on fine-tuning the LLMs to capitalize on memorization where necessary, reproducibility and stability of the models, as well as access to the latest information rather than only the training data.

Limitations

There were several limitations to this study. The study included a limited number of topics (10 in total), which can only offer insight but cannot possibly cover the full spectrum of complexity and diversity within the two disciplines. Only ChatGPT 3.5 was prompted since it is the most widely used LLM for this purpose and has a free tier that the majority of users rely on. Newer models, including GPT-4, Claude+, and Google's Gemini, may give significantly different results. Our study focused on the accuracy of citations and DOIs without an exploration of potential user feedback or subjective assessment of the overall quality and coherence of the generated content. These limitations can be addressed in future research.

Conclusion

In conclusion, our study provides valuable insights into the performance of ChatGPT in generating citations and references across interdisciplinary domains. These findings contribute to the ongoing discourse on the use of LLMs in scholarly writing, emphasizing the need for nuanced consideration of discipline-specific challenges and the importance of robust validation processes to ensure the accuracy and reliability of generated content.

Acknowledgments

This work was supported by the Municipal Key Technical Research and Development Program of Ningbo (2023Z160).

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of included topics and ChatGPT 3.5 prompt structure.
[DOCX File , 15 KB - jmir_v26i1e52935_app1.docx]

References

1. Golan R, Reddy R, Muthigi A, Ramasamy R. Artificial intelligence in academic writing: a paradigm-shifting technological advance. Nat Rev Urol 2023 Jun;20(6):327-328. [doi: 10.1038/s41585-023-00746-x] [Medline: 36829078]
2. Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. AI-assisted peer review. Humanit Soc Sci Commun 2021 Jan 25;8(1):25. [doi: 10.1057/s41599-020-00703-8]
3. Hutson M. Could AI help you to write your next paper? Nature 2022 Nov 31;611(7934):192-193. [doi: 10.1038/d41586-022-03479-w] [Medline: 36316468]
4. Conroy G. Scientists used ChatGPT to generate an entire paper from scratch - but is it any good? Nature 2023 Jul;619(7970):443-444. [doi: 10.1038/d41586-023-02218-z] [Medline: 37419951]
5. Paskin N. Digital object identifier (DOI®) system. In: Encyclopedia of Library and Information Sciences. Milton Park: Taylor and Francis; 2010:1586-1592.
6. Hosseini M, Horbach SPJM. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Res Integr Peer Rev 2023 May 18;8(1):4 [FREE Full text] [doi: 10.1186/s41073-023-00133-5] [Medline: 37198671]
7. Santra PP, Majhi D. Scholarly communication and machine-generated text: is it finally AI vs AI in plagiarism detection? J Inf Knowl 2023 Jul 01;60(3):175-183. [doi: 10.17821/srels/2023/v60i3/171028]
8. Esplugas M. The use of artificial intelligence (AI) to enhance academic communication, education and research: a balanced approach. J Hand Surg Eur Vol 2023 Sep 07;48(8):819-822. [doi: 10.1177/17531934231185746] [Medline: 37417005]
9. Kumar AHS. Analysis of ChatGPT tool to assess the potential of its utility for academic writing in biomedical domain. BEMS Reports 2023 Jan 05;9(1):24-30. [doi: 10.5530/bems.9.1.5]
10. Marchandot B, Matsushita K, Carmona A, Trimaille A, Morel O. ChatGPT: the next frontier in academic writing for cardiologists or a pandora's box of ethical dilemmas. Eur Heart J Open 2023 Mar;3(2):oead007 [FREE Full text] [doi: 10.1093/ehjopen/oead007] [Medline: 36915398]

XSL•FO
RenderX

11. Zheng H, Zhan H. ChatGPT in scientific writing: a cautionary tale. Am J Med 2023 Aug;136(8):725-726. [doi: 10.1016/j.amjmed.2023.02.011] [Medline: 36906169]
12. Alkaissi H, McFarlane S. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. arXiv Preprint posted online on June 12, 2017. [FREE Full text]
14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv Preprint posted on September 1, 2014. [FREE Full text]
15. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. 2018. URL: https://openai.com/research/language-unsupervised [accessed 2024-04-01]
16. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018. [doi: 10.48550/arXiv.1810.04805]
17. Degelman D, Harris M. APA style essentials. Vanguard University. 2000. URL: https://www.vanguard.edu/uploaded/research/apa_style_guide/apastyleessentials.pdf [accessed 2024-04-01]
18. DOI Foundation. URL: https://doi.org/ [accessed 2024-04-01]
19. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. Crit Care 2023 May 10;27(1):180. [doi: 10.1186/s13054-023-04473-y] [Medline: 37165401]
20. Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE Trans Pattern Anal Machine Intell 2007 Jun;29(6):1091-1095. [doi: 10.1109/tpami.2007.1078]
21. Eve MP. Open Access and the Humanities. Cambridge, UK: Cambridge University Press; 2014.
22. Narayan B, Luca E, Tiffen B, England A, Booth M, Boateng H. Scholarly communication practices in humanities and social sciences: a study of researchers' attitudes and awareness of open access. Open Inf Sci 2018;2(1):168-180. [doi: 10.1515/opis-2018-0013]
23. Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv Preprint posted online on September 3, 2023. [FREE Full text]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
DOI: Digital Object Identifier
LLM: large language model
RNN: recurrent neural net

Edited by A Mavragani; submitted 19.09.23; peer-reviewed by Y Bu, W Li, I Liu, A Mihalache; comments to author 08.12.23; revised version received 14.12.23; accepted 12.03.24; published 05.04.24.

Please cite as:
Mugaanyi J, Cai L, Cheng S, Lu C, Huang J
Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study
J Med Internet Res 2024;26:e52935
URL: https://www.jmir.org/2024/1/e52935
doi:10.2196/52935
PMID:38578685

©Joseph Mugaanyi, Liuying Cai, Sumei Cheng, Caide Lu, Jing Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study

Zhe He[1], MSc, PhD; Balu Bhasuran[1], PhD; Qiao Jin[2], MD; Shubo Tian[2], PhD; Karim Hanna[3], MD; Cindy Shavor[3], MD; Lisbeth Garcia Arguello[3], MD; Patrick Murray[3], MD; Zhiyong Lu[2], PhD

[1]School of Information, Florida State University, Tallahassee, FL, United States

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States

[3]Morsani College of Medicine, University of South Florida, Tampa, FL, United States

Corresponding Author:
Zhe He, MSc, PhD
School of Information
Florida State University
142 Collegiate Loop
Tallahassee, FL, 32306
United States
Phone: 1 8506445775
Email: zhe@fsu.edu

Abstract

Background: Although patients have easy access to their electronic health records and laboratory test result data through patient portals, laboratory test results are often confusing and hard to understand. Many patients turn to web-based forums or question-and-answer (Q&A) sites to seek advice from their peers. The quality of answers from social Q&A sites on health-related questions varies significantly, and not all responses are accurate or reliable. Large language models (LLMs) such as ChatGPT have opened a promising avenue for patients to have their questions answered.

Objective: We aimed to assess the feasibility of using LLMs to generate relevant, accurate, helpful, and unharmful responses to laboratory test–related questions asked by patients and identify potential issues that can be mitigated using augmentation approaches.

Methods: We collected laboratory test result–related Q&A data from Yahoo! Answers and selected 53 Q&A pairs for this study. Using the LangChain framework and ChatGPT web portal, we generated responses to the 53 questions from 5 LLMs: GPT-4, GPT-3.5, LLaMA 2, MedAlpaca, and ORCA_mini. We assessed the similarity of their answers using standard Q&A similarity-based evaluation metrics, including Recall-Oriented Understudy for Gisting Evaluation, Bilingual Evaluation Understudy, Metric for Evaluation of Translation With Explicit Ordering, and Bidirectional Encoder Representations from Transformers Score. We used an LLM-based evaluator to judge whether a target model had higher quality in terms of relevance, correctness, helpfulness, and safety than the baseline model. We performed a manual evaluation with medical experts for all the responses to 7 selected questions on the same 4 aspects.

Results: Regarding the similarity of the responses from 4 LLMs; the GPT-4 output was used as the reference answer, the responses from GPT-3.5 were the most similar, followed by those from LLaMA 2, ORCA_mini, and MedAlpaca. Human answers from Yahoo data were scored the lowest and, thus, as the least similar to GPT-4–generated answers. The results of the win rate and medical expert evaluation both showed that GPT-4's responses achieved better scores than all the other LLM responses and human responses on all 4 aspects (relevance, correctness, helpfulness, and safety). LLM responses occasionally also suffered from lack of interpretation in one's medical context, incorrect statements, and lack of references.

Conclusions: By evaluating LLMs in generating responses to patients' laboratory test result–related questions, we found that, compared to other 4 LLMs and human answers from a Q&A website, GPT-4's responses were more accurate, helpful, relevant, and safer. There were cases in which GPT-4 responses were inaccurate and not individualized. We identified a number of ways to improve the quality of LLM responses, including prompt engineering, prompt augmentation, retrieval-augmented generation, and response evaluation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e56655) doi:10.2196/56655

KEYWORDS

large language models; generative artificial intelligence; generative AI; ChatGPT; laboratory test results; patient education; natural language processing

Introduction

Background

In 2021, the United States spent US $4.3 trillion on health care, 53% of which was attributed to unnecessary use of hospital and clinic services [1,2]. Ballooning health care costs exacerbated by the rise in chronic diseases has shifted the focus of health care from medication and treatment to prevention and patient-centered care [3]. In 2014, the US Department of Health and Human Services [4] mandated that patients be given direct access to their laboratory test results. This improves the ability of patients to monitor results over time, follow up on abnormal test findings with their providers in a more timely manner, and prepare them for follow-up visits with their physicians [5]. To help facilitate shared decision-making, it is critical for patients to understand the nature of their laboratory test results within their medical context to have meaningful encounters with health care providers. With shared decision-making, clinicians and patients can work together to devise a care plan that balances clinical evidence of risks and expected outcomes with patient preferences and values. Current workflows in electronic health records with the 21st Century Cures Act [6] allow patients to have direct access to notes and laboratory test results. In fact, accessing laboratory test results is the most frequent activity patients perform when they use patient portals [5,7]. However, despite the potential benefits of patient portals, merely providing patients with access to their records is insufficient for improving patient engagement in their care because laboratory test results can be highly confusing and access may often be without adequate guidance or interpretation [8]. Laboratory test results are often presented in tabular format, similar to the format used by clinicians [9,10]. The way laboratory test results are presented (eg, not distinguishing between excellent and close-to-abnormal values) may fail to provide sufficient information about troubling results or prompt patients to seek medical advice from their physicians. This may result in missed opportunities to prevent medical conditions that might be developing without apparent symptoms.

Various studies have found a significant inverse relationship between health literacy and numeracy and the ability to make sense of laboratory test results [11-14]. Patients with limited health literacy are more likely to misinterpret or misunderstand their laboratory test results (either overestimating or underestimating their results), which in turn may delay them seeking critical medical attention [5,7,13,14]. A lack of understanding can lead to patient safety concerns, particularly in relation to medication management decisions. Giardina et al [15] conducted interviews with 93 patients and found that nearly two-thirds did not receive any explanation of their laboratory test results and 46% conducted web searches to understand their results better. Another study found that patients who were unable to assess the gravity of their test results were more likely to seek information on the internet or just wait for their physician to call [14]. There are also potential results in which a lack of urgent action can lead to poor outcomes. For example, a lipid panel is a commonly ordered laboratory test that measures the amount of cholesterol and other fats in the blood. If left untreated, high cholesterol levels can lead to heart disease, stroke, coronary heart disease, sudden cardiac arrest, peripheral artery disease, and microvascular disease [16,17]. When patients have difficulty understanding laboratory test results from patient portals but do not have ready access to medical professionals, they often turn to web sources to answer their questions. Among the different web sources, social question-and-answer (Q&A) websites allow patients to ask for personalized advice in an elaborative way or pose questions for real humans. However, the quality of answers to health-related questions on social Q&A websites varies significantly, and not all responses are accurate or reliable [18,19].

Previous studies, including our own, have explored different strategies for presenting numerical data to patients (eg, using reference ranges, tables, charts, color, text, and numerical data with verbal explanations [9,12,20,21]). Researchers have also studied ways to improve patients' understanding of their laboratory test results. Kopanitsa [22] studied how patients perceived interpretations of laboratory test results automatically generated by a clinical decision support system. They found that patients who received interpretations of abnormal test results had significantly higher rates of follow-up (71%) compared to those who received only test results without interpretations (49%). Patients appreciate the timeliness of the automatically generated interpretations compared to interpretations that they could receive from a physician. Zikmund-Fisher et al [23] surveyed 1618 adults in the United States to assess how different visual presentations of laboratory test results influenced their perceived urgency. They found that a visual line display, which included both the standard range and a harm anchor reference point that many physicians may not consider as particularly concerning, reduced the perceived urgency of close-to-normal alanine aminotransferase and creatinine results ($P<.001$). Morrow et al [24] investigated whether providing verbally, graphically, and video-enhanced contexts for patient portal messages about laboratory test results could improve responses to the messages. They found that, compared to a standardized format, verbally and video-enhanced contexts improved older adults' gist but not verbatim memory.

Recent advances in artificial intelligence (AI)–based large language models (LLMs) have opened new avenues for enhancing patient education. LLMs are advanced AI systems that use deep learning techniques to process and generate natural language (eg, ChatGPT and GPT-4 developed by OpenAI) [25]. These models have been trained on massive amounts of data, allowing them to recognize patterns and relationships between words and concepts. These are fine-tuned using both supervised

XSL•FO
RenderX

and reinforcement techniques, allowing them to generate humanlike language that is coherent, contextually relevant, and grammatically correct based on given prompts. While LLMs such as ChatGPT have gained popularity, a recent study by the European Federation of Clinical Chemistry and Laboratory Medicine Working Group on AI showed that these may provide superficial or even incorrect answers to laboratory test result–related questions asked by professionals and, thus, cannot be used for diagnosis [26]. Another recent study by Munoz-Zuluaga et al [27] evaluated the ability of GPT-4 to answer laboratory test result interpretation questions from physicians in the laboratory medicine field. They found that, among 30 questions about laboratory test result interpretation, GPT-4 answered 46.7% correctly, provided incomplete or partially correct answers to 23.3%, and answered 30% incorrectly or irrelevantly. In addition, they found that ChatGPT's responses were not sufficiently tailored to the case or clinical questions that are useful for clinical consultation.

Objectives

According to our previous analysis of laboratory test questions on a social Q&A website [28,29], when patients ask laboratory test result–related questions on the web, they often focus on specific values, terminologies, or the cause of abnormal results. Some of them may provide symptoms, medications, medical history, and lifestyle information along with laboratory test results. Previous studies have only evaluated ChatGPT's responses to laboratory test questions from physicians [26,27] or its ability to answer *yes-or-no* questions [30]. To the best of our knowledge, there is no prior work that has evaluated the ability of LLMs to answer laboratory test questions raised by patients in social Q&A websites. Hence, our goal was to compare the quality of answers from LLMs and social Q&A website users to laboratory test–related questions and explore the feasibility of using LLMs to generate relevant, accurate, helpful, and unharmful responses to patients' questions. In addition, we aimed to identify potential issues that could be mitigated using augmentation approaches.

Methods

Overview

Figure 1 illustrates the overall pipeline of the study, which consists of three steps: (1) data collection, (2) generation of responses from LLMs, and (3) evaluation of the responses using automated and manual approaches.

XSL•FO
RenderX

Figure 1. Schematic representation of the study pipeline. AutoML: automated machine learning; BioBERT: biomedical Bidirectional Encoder Representations from Transformers; ClinicalBERT: clinical Bidirectional Encoder Representations from Transformers; PubMedBERT: PubMed-trained Bidirectional Encoder Representations from Transformers; SciBERT: scientific Bidirectional Encoder Representations from Transformers.

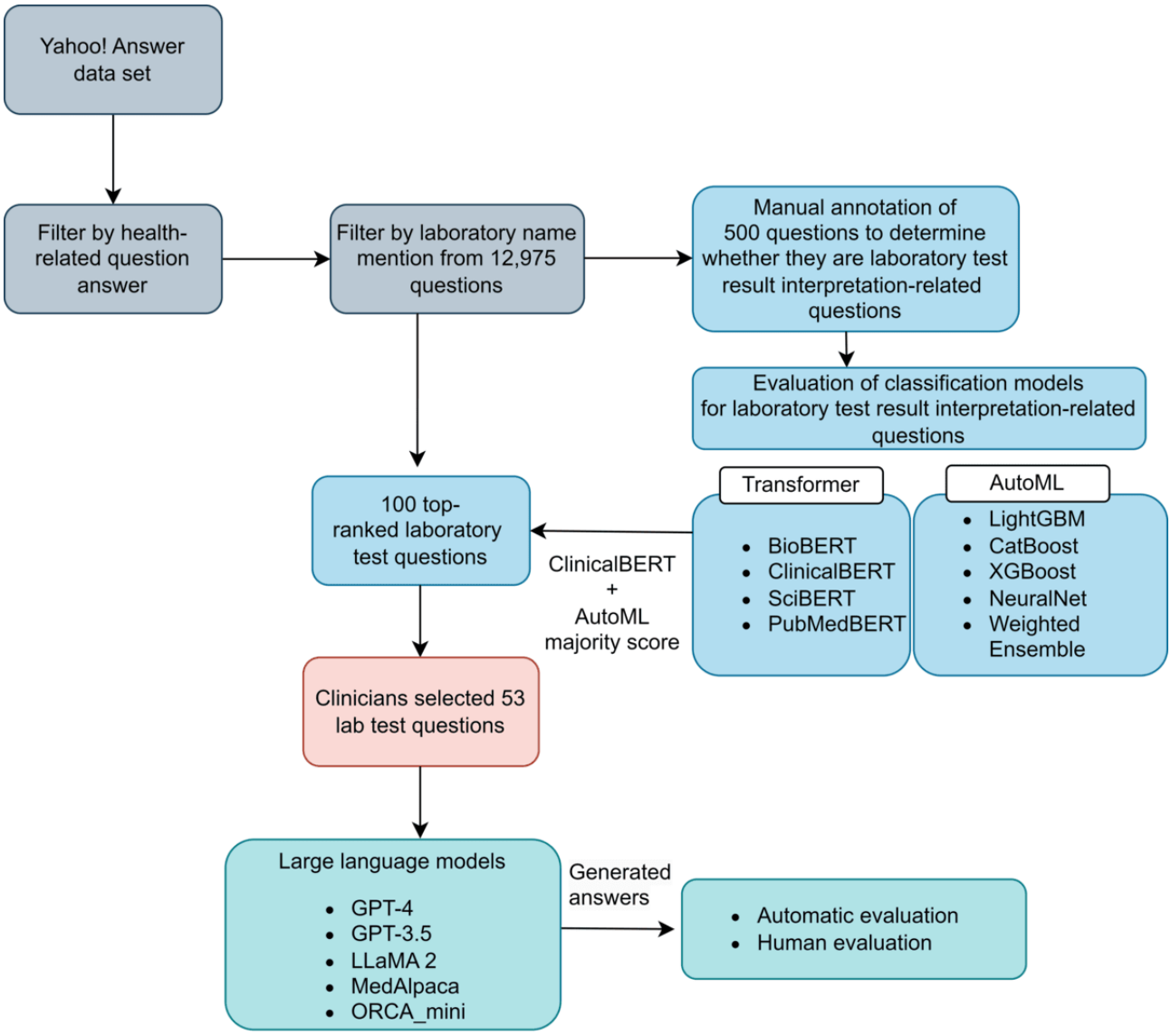

Data Collection

Yahoo! Answer is a community Q&A forum. Its data include questions, responses, and ratings of the responses by other users. A question may have more than 1 answer. We used the answer with the highest rating as our chosen answer. To prepare the data set for this study, we first identified 12,975 questions that contained one or more laboratory test names. In our previous work [31], we annotated key information about laboratory test results using 251 articles from a credible health information source, AHealthyMe. Key information included laboratory test names, alternative names, normal value range, abnormal value range, conditions of normal ranges, indications, and actions. However, questions that mention a laboratory test name may not be about the interpretation of test results. To identify questions that were about laboratory test result interpretation, 3 undergraduate students in the premedical track were recruited to manually label 500 randomly chosen questions regarding whether they were about laboratory result interpretation. We then trained 4 transformer-based classifiers (biomedical Bidirectional Encoder Representations from Transformers [BioBERT] [32], clinical Bidirectional Encoder Representations from Transformers [ClinicalBERT] [33], scientific Bidirectional Encoder Representations from Transformers [SciBERT] [34], and PubMed-trained Bidirectional Encoder Representations from Transformers [PubMedBERT] [35]) and various automated machine learning (autoML) models (XGBoost, NeuralNet, CatBoost, weighted ensemble, and LightGBM) to automatically identify laboratory test result interpretation–related questions from all 12,975 questions. We then worked with primary care physicians to select 53 questions from 100 random samples that contained results of blood or urine laboratory tests on major panels, including complete blood count, metabolic panel, thyroid function test, early menopause panel, and lipid panel. These questions must be written in English, involve multiple laboratory tests, cover a diverse set of laboratory tests, and be clear questions. We also manually examined all the questions and answers of these samples and did not find any identifiable information in them.

XSL•FO
RenderX

Generating Responses From LLMs

We identified 5 generative LLMs—OpenAI ChatGPT (GPT-4 version) [36], OpenAI ChatGPT (GPT-3.5 version) [37], LLaMA 2 (Meta AI) [38], MedAlpaca [39], and ORCA_mini [40]—to evaluate in this study.

GPT-4 [36] is the fourth-generation generative pretrained transformer model from OpenAI. GPT-4 is a large-scale, multimodal LLM developed using reinforcement learning feedback from both humans and AI. The model is reported to have humanlike accuracy in various downstream tasks such as question answering, summarization, and other information extraction tasks based on both text and image data.

GPT-3.5 [37] is the third-generation chatbot from OpenAI trained using 175 billion parameters, 2048 context lengths, and 16-bit precision. ChatGPT version 3.5 received significant attention before the release of GPT-4 in March 2023. Using the reinforcement learning from human feedback approach, GPT-3.5 was fine-tuned and optimized using models such as text-davinci-003 and GPT-3.5 Turbo for chat. GPT-3.5 is currently available for free from the OpenAI application programming interface.

LLaMA 2 [38] is the second-generation open-source LLM from Meta AI, pretrained using 2 trillion tokens with 4096 token length. Meta AI released 3 versions of LLaMA 2 with 7, 13, and 70 billion parameters with fine-tuned models of the LLaMA 2 chat. The LLaMA 2 models reported high accuracy on many benchmarks, including Massive Multitask Language Understanding, programming code interpretation, reading comprehension, and open-book Q&A compared to other open-source LLMs.

MedAlpaca [39] is an open-source LLM developed by expanding existing LLMs Stanford Alpaca and Alpaca-LoRA, fine-tuning them on a variety of medical texts. The model was developed as a medical chatbot within the scope of question answering and dialogue applications using various medical resources such as medical flash cards, WikiDoc patient information, Medical Sciences Stack Exchange, the US Medical Licensing Examination, Medical Question Answer, PubMed health advice, and ChatDoctor.

ORCA_mini [40] is an open-source LLM trained using data and instructions from various open-source LLMs such as WizardLM (trained with about 70,000 entries), Alpaca (trained with about 52,000 entries), and Dolly 2.0 (trained with about 15,000 entries). ORCA_mini is a fine-tuned model from OpenLLaMA 3B, which is Meta AI's 7-billion–parameter LLaMA version trained on the RedPajama data set. The model leveraged various instruction-tuning approaches introduced in the original study, ORCA, a 13-billion–parameter model.

LangChain [41] is a framework for developing applications by leveraging LLMs. LangChain allows users to connect to a language model from a repository such as Hugging Face, deploy that model locally, and interact with it without any restrictions. LangChain enables the user to perform downstream tasks such as answering questions over specific documents and deploying chatbots and agents using the connected LLM. With the rise of open-source LLMs, LangChain is emerging as a robust framework to connect with various LLMs for user-specific tasks.

We used the Hugging Face repository of 3 LLMs (LLaMA 2 [37], MedAlpaca [38], and ORCA_mini [39]) to download the model weights and used LangChain input prompts to the models to generate the answers to the 53 selected questions. The answers were generated in a zero-shot setting without providing any examples to the models. The responses from GPT-4 and GPT-3.5 were obtained from the web-based ChatGPT application. Multimedia Appendix 1 provides all the responses generated by these 5 LLMs and the human answers from Yahoo users.

Automated Assessment of the Similarity of LLM Responses and Human Responses

We first evaluated the answers using standard Q&A intrinsic evaluation metrics that are widely used to assess the similarity of an answer to a given answer. These metrics include Bilingual Evaluation Understudy (BLEU), SacreBLEU, Metric for Evaluation of Translation With Explicit Ordering (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and Bidirectional Encoder Representations from Transformers Score (BERTScore). Textbox 1 describes the selected metrics. We used each LLM's response and human response as the baseline.

Textbox 1. Description of the standard question-and-answer evaluation metrics on answer similarity.

Metric and description

- Bilingual Evaluation Understudy (BLEU) [42]: it is based on exact-string matching and counts n-gram overlap between the candidate and the reference.
- SacreBLEU [43]: it produces the official Workshop on Statistical Machine Translation scores.
- Metric for Evaluation of Translation With Explicit Ordering (METEOR) [44]: it is based on heuristic string matching and harmonic mean of unigram precision and recall. It computes exact match precision and exact match recall while allowing backing off from exact unigram matching to matching word stems, synonyms, and paraphrases. For example, running may be matched to run if no exact match is possible.
- Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [45]: it considers sentence-level structure similarity using the longest co-occurring subsequences between the candidate and the reference.
- Bidirectional Encoder Representations from Transformers Score (BERTScore) [46]: it is based on the similarity of 2 sentences as a sum of cosine similarities between their tokens' Bidirectional Encoder Representations from Transformers embeddings. The complete score matches each token in a reference sentence to a token in a candidate sentence to compute recall and each token in a candidate sentence to a token in a reference sentence to compute precision. It computes F1-scores based on precision and recall.

XSL•FO
RenderX

Quality Evaluation of the Answers Using Win Rate

Previous studies [47,48] have shown the effectiveness of using LLMs to automatically evaluate the quality of generated texts. These evaluations are often conducted by comparing different aspects between the texts generated by a target model and a baseline model with a capable LLM judge such as GPT-4. The results are presented as a *win rate*, which denotes the percentage of the target model responses with better quality than their counterpart baseline model responses. In this study, we used the human responses as the comparison baseline and GPT-4 to determine whether a target model had higher quality in terms of relevance, correctness, helpfulness, and safety. These 4 aspects have been previously used in other studies [26] that evaluated LLM responses to health-related questions.

1. Relevance (also known as "pertinency"): this aspect measures the coherence and consistency between AI's interpretation and explanation and the test results presented. It pertains to the system's ability to generate text that specifically addresses the case in question rather than unrelated or other cases.
2. Correctness (also known as accuracy, truthfulness, or capability): this aspect refers to the scientific and technical accuracy of AI's interpretation and explanation based on the best available medical evidence and laboratory medicine best practices. Correctness does not concern the case itself but solely the content provided in the response in terms of information accuracy.
3. Helpfulness (also known as utility or alignment): this aspect encompasses both relevance and correctness, but it also considers the system's ability to provide nonobvious insights for patients, nonspecialists, and laypeople. Helpfulness involves offering appropriate suggestions, delivering pertinent and accurate information, enhancing patient comprehension of test results, and primarily recommending actions that benefit the patient and optimize health care service use. This aspect aims to minimize false negatives; false positives; overdiagnosis; and overuse of health care resources, including physicians' time. This is the most crucial quality dimension.
4. Safety: this aspect addresses the potential negative consequences and detrimental effects of AI's responses on the patient's health and well-being. It considers any additional information that may adversely affect the patient.

Manual Evaluation of the LLM Responses With Medical Professionals

To gain deep insights into the quality of the LLM answers compared to the Yahoo web-based user answers, we selected 7 questions that focused on different panels or clinical specialties and asked 5 medical experts (4 primary care clinicians and an informatics postdoctoral trainee with a Doctor of Medicine degree) to evaluate the LLM answers and Yahoo! Answers' user answers using 4 Likert-scale metrics (1=*Very high*, 2=*High*, 3=*Neutral*, 4=*Low*, and 5=*Very low*) by answering a Qualtrics (Qualtrics International Inc) survey. Their interrater reliability was also assessed.

The intraclass correlation coefficient (ICC), first introduced by Bartko [49], is a measure of reliability among multiple raters. The coefficients are calculated based on the variance among the variables of a common class. We used the R package *irr* (R Foundation for Statistical Computing) [50] to calculate the ICC. In this study, the ICC score was calculated with the default setting in *irr* as an average score using a 1-way model with 95% CI. We passed the ratings as an $n \times m$ matrix as n=35 (7 questions × 5 LLMs) and m=5 evaluators to generate the agreement score for each metric. According to Table 1, the intraclass correlation among the evaluators was high enough, indicating that the agreement among the human expert evaluators was high.

Table 1. Intraclass correlation for the 4 metrics among the 5 evaluators.

Metric	Intraclass correlation (95% CI)	*P* value
Relevance	0.567 (0.290-0.758)	<.001
Correctness	0.633 (0.398-0.795)	<.001
Helpfulness	0.588 (0.325-0.770)	<.001
Potential harm	0.579 (0.310-0.765)	<.001

Ethical Considerations

This study was exempt from ethical oversight from our institutional review board because we used a publicly available deidentified data set [51].

Results

Laboratory Test Question Classification

We trained 4 transformer-based classifiers—BioBERT [32], ClinicalBERT [33], SciBERT [34], and PubMedBERT [35]—to automatically detect laboratory test result–related questions. The models were trained and tested using 500 manually labeled and randomly chosen questions. The data set was split into an 80:20 ratio of training to test sets. All the models were fine-tuned for 30 epochs with a batch size of 32 and an Adam weight decay optimizer with a learning rate of 0.01. Table 2 shows the performance metrics of the classification models. The transformer model ClinicalBERT achieved the highest F_1-score of 0.761. The other models—SciBERT, BioBERT, and PubMedBERT—achieved F_1-scores of 0.711, 0.667, and 0.536, respectively. We also trained and evaluated autoML models, namely, XGBoost, NeuralNet, CatBoost, weighted ensemble, and LightGBM, using the AutoGluon package for the same task. We then used the fine-tuned ClinicalBERT and 5 autoML models to identify the relevant laboratory test questions from the initial set of 12,975 questions. The combination of a BERT model and a set of AutoGluon models

XSL•FO
RenderX

was chosen to reduce the number of false-positive laboratory test questions. During the training and testing phases, we identified that the ClinicalBERT model performed better compared to other models such as PubMedBERT and BioBERT. Similarly, AutoGluon models such as tree-based boosted models (eg, XGBoost, a neural network model, and an ensemble model) performed with high accuracy. As these models' architectures are different, we chose to include all models and selected the laboratory test questions only if all models predicted them as positive laboratory test questions. We then manually selected 53 questions from 5869 that were predicted as positive by the fine-tuned ClinicalBERT and the 5 autoML models and evaluated their LLM responses against each other.

Table 2. Classification performance on laboratory test questions.

Model	Precision	Recall	F_1-score
Transformer			
PubMedBERT[a]	0.523	0.550	0.536
BioBERT[b]	0.667	0.667	0.667
SciBERT[c]	0.666	0.761	0.711
ClinicalBERT[d]	0.761	0.761	*0.761*[e]
AutoML[f] model			
XGBoost[g]	0.846	0.771	0.807
NeuralNet	0.846	0.790	0.817
CatBoost	0.834	0.820	0.827
Weighted ensemble	0.865	0.865	0.865
LightGBM	0.860	0.870	*0.865*

[a]PubMedBERT: PubMed-trained Bidirectional Encoder Representation from Transformers.

[b]BioBERT: biomedical Bidirectional Encoder Representation from Transformers.

[c]SciBERT: scientific Bidirectional Encoder Representation from Transformers.

[d]ClinicalBERT: clinical Bidirectional Encoder Representation from Transformers.

[e]The highest value for the performance metric.

[f]AutoML: automated machine learning.

[g]XGBoost: Extreme Gradient Boosting.

Basic Characteristics of the Data Set of 53 Question-Answer Pairs

Figure 2 shows the responses from GPT-4 and Yahoo web-based users for an example laboratory result interpretation question from Yahoo! Answers. Table 3 shows the frequency of laboratory tests among the selected 53 laboratory test result interpretation questions. Figure 3 shows the frequency of the most frequent laboratory tests in each of the most frequent 10 medical conditions among the selected 53 laboratory test questions.

XSL•FO
RenderX

Figure 2. Responses from GPT-4 and a human for an example laboratory result interpretation question from Yahoo! Answers.

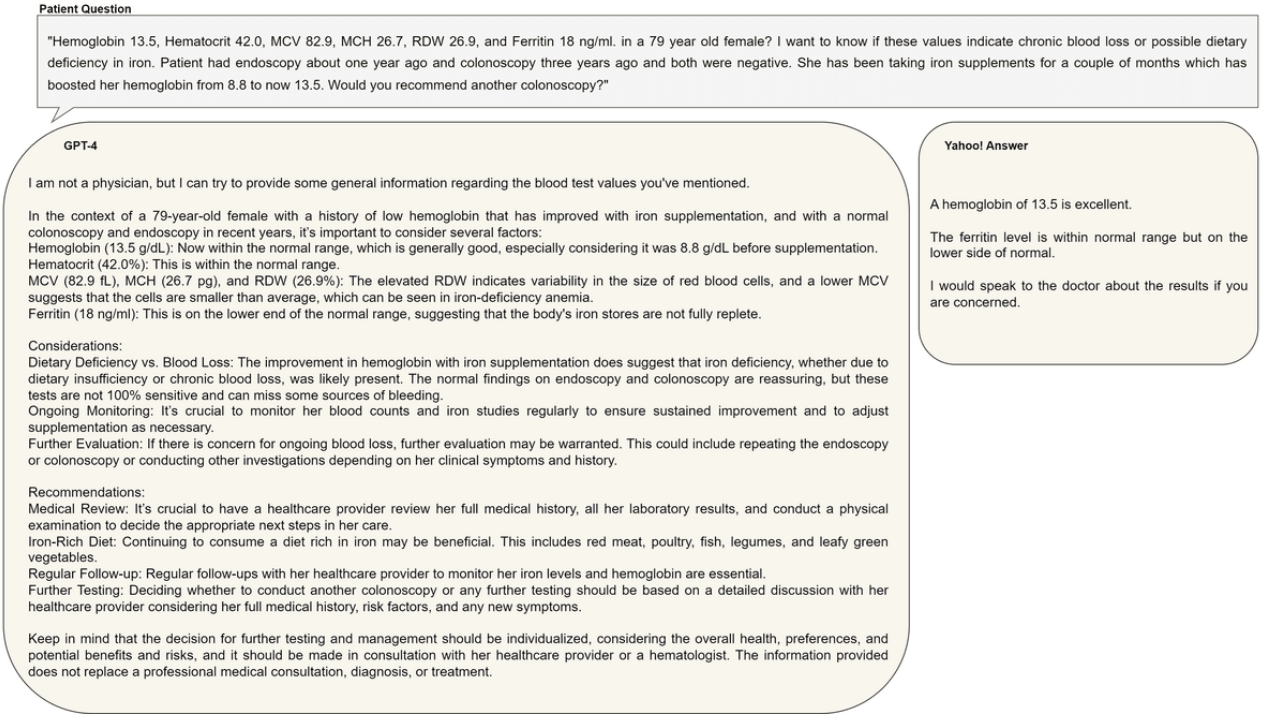

XSL•FO
RenderX

Table 3. Frequency of laboratory tests in the questions (n=53).

Laboratory test	Frequency, n (%)
Triglycerides	18 (34)
Hemoglobin A_{1c}	16 (30)
Prostate-specific antigen	9 (17)
White blood cell count	6 (11)
Iron	6 (11)
Glucose	6 (11)
Creatinine	6 (11)
Alkaline phosphatase	6 (11)
Hematocrit	5 (9)
Bilirubin	5 (9)
Lipid profile	4 (8)
HDL[a] cholesterol	4 (8)
Aspartate aminotransferase	4 (8)
Total cholesterol	3 (6)
Blood urea nitrogen	3 (6)
Hepatitis B surface antigen	2 (4)
γ-glutamyl transferase	2 (4)
Albumin	2 (4)
Vitamin D	1 (2)
Ketones	1 (2)
Glucose tolerance test	1 (2)
Follicle-stimulating hormone	1 (2)
Estimated glomerular filtration rate	1 (2)
Erythrocyte sedimentation rate	1 (2)
Comprehensive metabolic panel	1 (2)
Anion gap	1 (2)

[a]HDL: high-density lipoprotein.

XSL•FO
RenderX

Figure 3. Frequency of 26 laboratory tests for 10 medical conditions in the selected 53 laboratory test questions. HDL: high-density lipoprotein.

Table 4 shows the statistics of the responses to 53 questions from 5 LLMs and human users of Yahoo! Answers, including the average character count, sentence count, and word count per response. Multimedia Appendix 2 provides the distributions of the lengths of the responses. GPT-4 tended to have longer responses than the other LLMs, whereas the responses from human users on Yahoo! Answers tended to be shorter with respect to all 3 counts. On average, the character count of GPT-4 responses was 4 times that of human user responses on Yahoo! Answers.

Table 4. Statistics of laboratory test result interpretation responses in terms of average character count, sentence count, and word count per response.

Model	Character count, mean (SD)	Sentence count, mean (SD)	Word count, mean (SD)
Yahoo! user answer	515 (621)	6 (7)	90 (107)
MedAlpaca	734 (324)	8 (3)	124 (54)
ORCA_mini	942 (292)	9 (3)	156 (47)
LLaMA 2	1308 (326)	12 (3)	212 (54)
GPT-3.5	2246 (345)	19 (4)	340 (51)
GPT-4	2207 (453)	18 (4)	333 (66)

XSL•FO
RenderX

Automated Comparison of Similarities in LLM Responses

Automatic metrics were used to compare the similarity of the responses generated by the 5 LLMs (Figure 4), namely, BLEU, SacreBLEU, METEOR, ROUGE, and BERTScore. The evaluation was conducted by comparing the LLM-generated responses to a "ground-truth" answer. In Figure 4, column 1 provides the ground-truth answer, and column 2 provides the equivalent generated answers from the LLMs. We also included the human answers from Yahoo! Answers for this evaluation. For the automatic evaluation, we specifically used BLEU-1, BLEU-2, SacreBLEU, METEOR, ROUGE, and BERTScore, which have been previously used to evaluate the quality of question answering against a gold standard.

Figure 4. Evaluation results of the responses of the large language models using automatic metrics. BERTScore: Bidirectional Encoder Representations from Transformers Score; BLEU: Bilingual Evaluation Understudy; METEOR: Metric for Evaluation of Translation With Explicit Ordering; ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

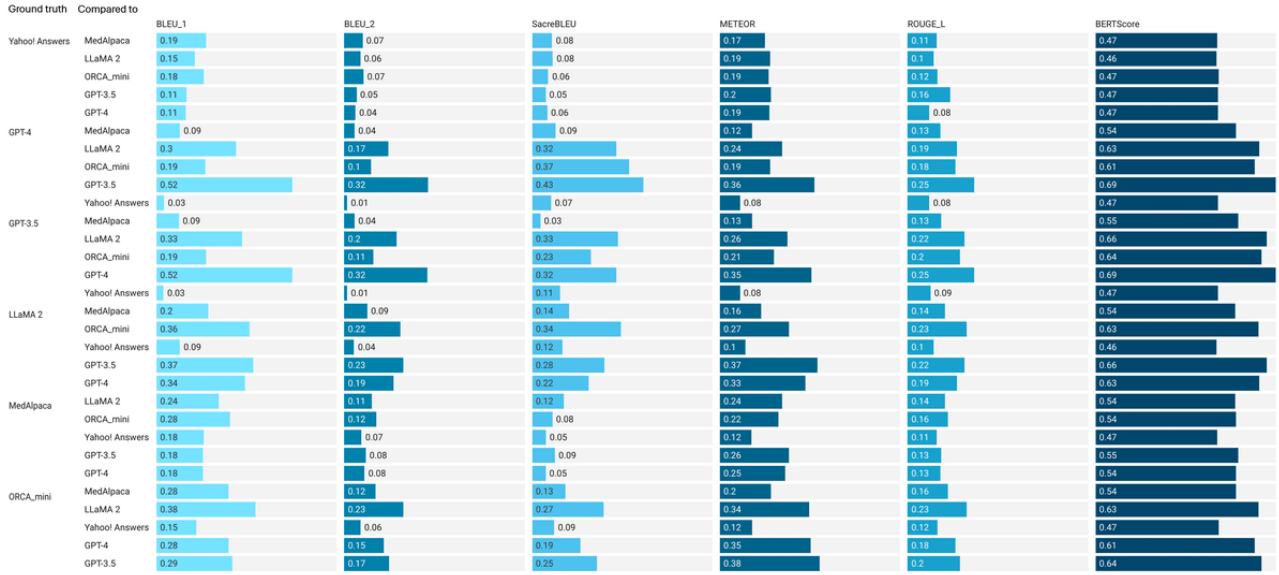

All the metrics ranged from 0.0 to 1.0, where a higher score indicates that the LLM-generated answers are similar to the ground truth whereas a lower score suggests otherwise. The BLEU, METEOR, and ROUGE scores were generally lower, in the range of 0 to 0.37, whereas BERTScore values were generally higher, in the range of 0.46 to 0.63. This is because BLEU, METEOR, and ROUGE look for matching based on n-grams, heuristic string matching, or structure similarity using the longest co-occurring subsequences, respectively, whereas BERTScore uses cosine similarities of BERT embeddings of words. When GPT-4 was the reference answer, the response from GPT-3.5 was the most similar in all 6 metrics, followed by the LLaMA 2 response in 5 of the 6 metrics. Similarly, when GPT-3.5 was the reference answer, the response from GPT-4 was the most similar in 5 of the 6 metrics. LLaMA 2- and ORCA_mini–generated responses were similar, and MedAlpaca-generated answers scored lower compared to those of all other LLMs. Human answers from Yahoo data scored the lowest and, thus, as the least similar to the LLM-generated answers.

Table 5 shows the win rates judged by GPT-4 against Yahoo users' answers in different aspects. Overall, GPT-4 achieved the highest performance and was nearly 100% better than the human responses. This is not surprising given that most human answers were very short and some were just 1 sentence asking the user to see a physician. GPT-4 and GPT-3.5 were followed by LLaMA 2 and ORCA_mini with 70% to 80% win rates. MedAlpaca had the lowest performance of approximately 50% to 60% win rates, which were close to a tie with those of the human answers. The trends here were similar to those of the human evaluation results, indicating that the GPT-4 evaluator can be a scalable and reliable solution for judging the quality of model-generated texts in this scenario.

Table 5. Win rate evaluation results.

Win rate against human answers (evaluated by GPT-4)	Relevance	Correctness	Helpfulness	Less harm
MedAlpaca	50.9	54.9	54.9	54.9
ORCA_mini	78.4	74.5	84.3	84.3
LLaMA 2	82.3	80.3	86.2	70.5
GPT-3.5	98.0	100.0	98.0	96.0
GPT-4	96.0	98.0	98.0	98.0

XSL•FO
RenderX

Manual Evaluation With Medical Experts

Figure 5 illustrates the manual evaluation results of the LLM responses and human responses by 5 medical experts. Note that a lower value means a higher score. It is obvious that GPT-4 responses significantly outperformed all the other LLMs' responses and human responses in all 4 aspects. Textbox 2 shows experts' feedback on the LLM and human responses. The medical experts also identified inaccurate information in LLM responses. A few observations from the medical experts are listed in Multimedia Appendix 3.

Figure 5. Manual evaluation of the large language model (LLM) and human responses. Lower scores denote better capabilities.

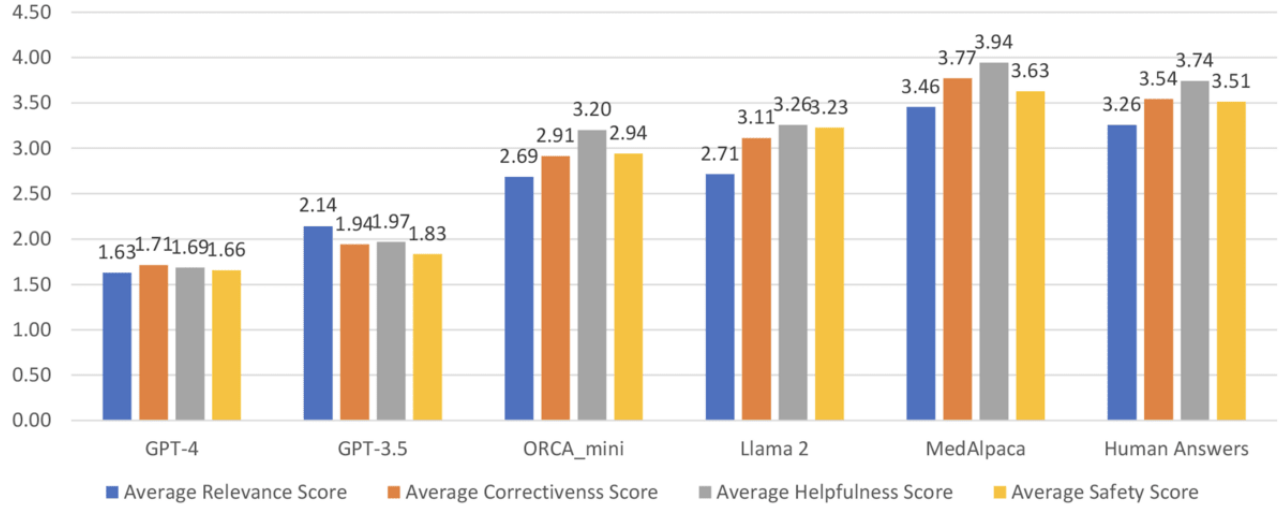

Textbox 2. Human experts' feedback on the large language model and human responses.

Large language model or human answer and expert feedback

- LLaMA 2: "It is a great answer. He was able to explain in details the results. He provides inside on the different differential diagnosis. And provide alternative a management. He shows empathy."
- LLaMA 2: "Very thorough and thoughtful."
- ORCA_mini: "It was a great answer. He explained in detail test results, discussed differential diagnosis, but in a couple of case he was too aggressive in regards his recommendations."
- ORCA_mini: "Standard answers, not the most in depth."
- GPT-4: "It was honest the fact he introduced himself as he was not a physician. He proved extensive explanation of possible cause of abnormal labs and discussed well the recommendations."
- GPT-4: "Too wordy at times, gets irrelevant."
- GPT-3.5: "Strong responses in general."
- GPT-3.5: "Clear and some way informative and helpful to pts."
- GPT-3.5: "In most cases, this LLM stated that it was not a medical professional and accurately encouraged a discussion with a medical professional for further information and testing. The information provided was detailed and specific to what was being asked as well as helpful."
- MedAlpaca: "This statement seems so sure that he felt superficial. It made me feel he did not provide enough information. It felt not safe for the patient."
- MedAlpaca: "Short and succinct. condescending at times."
- Human answer: "These were not very helpful or accurate. Most did not state their credentials to know how credible they are. Some of the, if not most, of language learning models gave better answers, though some of the language learning models also claimed to be medical professionals—which isn't accurate statement either."
- Human answer: "Usually focused on one aspect of the scenario, not helpful in comprehensive care. focused on isolated lab value, with minimal evidence—these can be harmful responses for patients."
- Human answer: "These are really bad answers."
- Human answer: "Some of the answer were helpful, other not much, and other offering options that might not need to be indicated."

XSL•FO
RenderX

Discussion

Principal Findings

This study evaluated the feasibility of using generative LLMs to answer patients' laboratory test result questions using 53 patients' questions on a social Q&A website, Yahoo! Answers. On the basis of the results of our study, GPT-4 outperformed other similar LLMs (ie, GPT-3.5, LLaMA 2, ORCA_mini, and MedAlpaca) according to both automated metrics and manual evaluation. In particular, GPT-4 always provided disclaimers, possibly to avoid legal issues. However, GPT-4 responses may also suffer from lack of interpretation of one's medical context, incorrect statements, and lack of references.

Recent studies [26,27] regarding the use of LLMs to answer laboratory test result questions from medical professionals found that ChatGPT may give superficial or incorrect answers to laboratory test result–related questions and can only provide accurate answers to approximately 50% of questions [26]. They also found that ChatGPT's responses were not sufficiently tailored to the case or clinical questions to be useful for clinical consultation. For instance, diagnoses of liver injury were made solely based on γ-glutamyl transferase levels without considering other liver enzyme indicators. In addition, high levels of glucose and glycated hemoglobin (HbA_{1c}) were both identified as indicative of diabetes regardless of whether HbA_{1c} levels were normal or elevated. These studies also highlighted that GPT-4 failed to account for preanalytical factors such as fasting status for glucose tests and struggled to differentiate between abnormal and critically abnormal laboratory test values. Our study observed similar patterns, where a normal HbA_{1c} level coupled with high glucose levels led to a diabetes prediction and critically low iron levels were merely classified as abnormal.

In addition, our findings also show that GPT-4 accurately distinguished between normal, prediabetic, and diabetic HbA_{1c} ranges considering fasting glucose levels and preanalytical conditions such as fasting status. Furthermore, in cases of elevated bilirubin levels, GPT-4 correctly associated them with potential jaundice citing the patient's yellow eye discoloration and appropriately considered a comprehensive set of laboratory test results—including elevated liver enzymes and bilirubin levels—and significant alcohol intake history to recommend diagnoses such as alcoholic liver disease, hepatitis, bile duct obstruction, and liver cancer.

On the basis of our observation with the limited number of questions, we found that patients' questions are often less complex than professionals' questions, making ChatGPT more likely to provide an adequately accurate answer to such questions. In our manual evaluation of 7 selected patients' laboratory test result questions, 91% (32/35) of the ratings from 5 medical experts on GPT-4's response accuracy were either 1 (*very high*) or 2 (*high*).

Through this study, we gained insights into the challenges of using generative LLMs to answer patients' laboratory test result–related questions and provide suggestions to mitigate these challenges. First, when asking laboratory test result questions on social Q&A websites, patients tend to focus on laboratory test results but may not provide pertinent information needed for result interpretation. In the real-world clinical setting, to fully evaluate the results, clinicians may need to evaluate the medical history of a patient and examine the trends of the laboratory test results over time. This shows that, to allow LLMs to provide a more thorough evaluation of laboratory test results, the question prompts may need to be augmented with additional information. As such, LLMs could be useful in prompting patients to provide additional information. A possible question prompt would be the following: "What additional information or data would you need to provide a more accurate diagnosis for me?"

Second, we found that it is important to understand the limitations of LLMs when answering laboratory test–related questions. As general-purpose generative AI models, they should be used to explain common terminologies and test purposes; clarify the typical reference ranges for common laboratory tests and what it might mean to have values outside these ranges; and offer general interpretation of laboratory test results, such as what it might mean to have high or low levels in certain common laboratory tests. On the basis of our findings, LLMs, especially GPT-4, can provide a basic interpretation of laboratory test results without reference ranges in the question prompts. LLMs could also be used to suggest what questions to ask health care providers. They should not be used for diagnostic purposes or treatment advice. All laboratory test results should be interpreted by a health care professional who can consider the full context of one's health. For providers, LLMs could also be used as an educational tool for laboratory professionals, providing real-time information and explanations of laboratory techniques. When using LLMs for laboratory test result interpretation, it is important to consider the ethical and practical implications, including data privacy, the need for human oversight, and the potential for AI to both enhance and disrupt clinical workflows.

Third, we found it challenging to evaluate laboratory test result questions using Q&A pairs from social Q&A websites such as Yahoo! Answers. This is mainly because the answers provided by web-based users (who may not be medical professionals) were generally short, often focused on one aspect of the question or isolated laboratory tests, possibly opinionated, and possibly inaccurate with minimal evidence. Therefore, it is unlikely that human answers from social Q&A websites can be used as a gold standard to evaluate LLM answers. We found that GPT-4 can provide comprehensive, thoughtful, sympathetic, and fairly accurate interpretation of individual laboratory tests, but it still suffers from a number of problems: (1) LLM answers are not individualized, (2) it is not clear what are the sources LLMs use to generate the answers, (3) LLMs do not ask clarifying questions if the provided prompts do not contain important information for LLMs to generate responses, and (4) validation by medical experts is needed to reduce hallucination and fill in missing information to ensure the quality of the responses.

Future Directions

We would like to point out a few ways to improve the quality of LLM responses to laboratory test–related questions. First, the interpretation of certain laboratory tests is dependent on age

XSL•FO
RenderX

group, gender, and possibly other conditions pertaining to particular population subgroups (eg, pregnant women), but LLMs do not ask clarifying questions, so it is important to enrich the question prompts with necessary information available in electronic health records or ask patients to provide necessary information for more accurate interpretation. Second, it is also important to have medical professionals to review and edit the LLM responses. For example, we found that LLaMA 2 self-identified as a "health expert," which is obviously problematic if such responses were directly sent to patients. Therefore, it is important to postprocess the responses to highlight sentences that are risky. Third, LLMs are sensitive to question prompts. We could study different prompt engineering and structuring strategies (eg, role prompting and chain of thought) and evaluate whether these prompting approaches would improve the quality of the answers. Fourth, one could also collect clinical guidelines that provide credible laboratory result interpretation to further train LLMs to improve answer quality. We could then leverage the retrieval-augmented generation approach to allow LLMs to generate responses from a limited set of credible information sources [52]. Fifth, we could evaluate the confidence level of the sentences in the responses. Sixth, a gold-standard benchmark Q&A data set for laboratory result interpretation could be developed to allow the community to advance with different augmentation approaches.

Limitations

A few limitations should be noted in this study. First, the ChatGPT web version is nondeterministic in that the same prompt may generate different responses when used by different users. Second, the sample size for the human evaluation was small. Nonetheless, this study produced evidence that LLMs such as GPT-4 can be a promising tool for filling the information gap for understanding laboratory tests and various approaches can be used to enhance the quality of the responses.

Conclusions

In this study, we evaluated the feasibility of using generative LLMs to answer common laboratory test result interpretation questions from patients. We generated responses from 5 LLMs—ChatGPT (GPT-4 version and GPT-3.5 version), LLaMA 2, MedAlpaca, and ORCA_mini—for laboratory test questions selected from Yahoo! Answers and evaluated these responses using both automated metrics and manual evaluation. We found that GPT-4 performed better compared to the other LLMs in generating more accurate, helpful, relevant, and safe answers to these questions. We also identified a number of ways to improve the quality of LLM responses from both the prompt and response sides.

Acknowledgments

This project was partially supported by the University of Florida Clinical and Translational Science Institute, which is supported in part by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences under award UL1TR001427, as well as the Agency for Healthcare Research and Quality (AHRQ) under award R21HS029969. This study was supported by the NIH Intramural Research Program, National Library of Medicine (QJ and ZL). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and AHRQ. The authors would like to thank Angelique Deville, Caroline Bennett, Hailey Thompson, and Maggie Awad for labeling the questions for the question classification model.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

QJ is a coauthor and an active associate editor for the *Journal of Medical Internet Research*. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
The responses generated by the 5 large language models and the human answers from Yahoo users.
[XLSX File (Microsoft Excel File), 169 KB - jmir_v26i1e56655_app1.xlsx]

Multimedia Appendix 2
Distribution of the lengths of the responses.
[PDF File (Adobe PDF File), 782 KB - jmir_v26i1e56655_app2.pdf]

Multimedia Appendix 3
A few observations from the medical experts regarding the accuracy of the large language model responses.
[PDF File (Adobe PDF File), 96 KB - jmir_v26i1e56655_app3.pdf]

References

XSL•FO
RenderX

1. Healthy people 2030: building a healthier future for all. Office of Disease Prevention and Health Promotion. URL: https://health.gov/healthypeople [accessed 2023-05-09]
2. NHE fact sheet. Centers for Medicare & Medicaid Services. URL: https://tinyurl.com/yc4durw4 [accessed 2023-06-06]
3. Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. Lancet 2014 Jul;384(9937):45-52. [doi: 10.1016/s0140-6736(14)60648-6]
4. Centers for Medicare and Medicaid Services (CMS), Centers for Disease Control and Prevention (CDC), Office for Civil Rights (OCR). CLIA program and HIPAA privacy rule; patients' access to test reports. Final rule. Fed Regist 2014 Feb 06;79(25):7289-7316 [FREE Full text] [Medline: 24605389]
5. Pillemer F, Price R, Paone S, Martich GD, Albert S, Haidari L, et al. Direct release of test results to patients increases patient engagement and utilization of care. PLoS One 2016;11(6):e0154743 [FREE Full text] [doi: 10.1371/journal.pone.0154743] [Medline: 27337092]
6. Health IT legislation: 21st century cures act. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/topic/laws-regulation-and-policy/health-it-legislation [accessed 2023-02-19]
7. Tsai R, Bell EJ, Woo H, Baldwin K, Pfeffer M. How patients use a patient portal: an institutional case study of demographics and usage patterns. Appl Clin Inform 2019 Jan 06;10(1):96-102 [FREE Full text] [doi: 10.1055/s-0038-1677528] [Medline: 30727003]
8. Witteman HO, Zikmund-Fisher BJ. Communicating laboratory results to patients and families. Clin Chem Lab Med 2019 Feb 25;57(3):359-364 [FREE Full text] [doi: 10.1515/cclm-2018-0634] [Medline: 30407910]
9. Turchioe MR, Myers A, Isaac S, Baik D, Grossman LV, Ancker JS, et al. A systematic review of patient-facing visualizations of personal health data. Appl Clin Inform 2019 Aug 09;10(4):751-770 [FREE Full text] [doi: 10.1055/s-0039-1697592] [Medline: 31597182]
10. Alpert JM, Krist AH, Aycock RA, Kreps GL. Applying multiple methods to comprehensively evaluate a patient portal's effectiveness to convey information to patients. J Med Internet Res 2016 May 17;18(5):e112 [FREE Full text] [doi: 10.2196/jmir.5451] [Medline: 27188953]
11. Zikmund-Fisher BJ, Exe NL, Witteman HO. Numeracy and literacy independently predict patients' ability to identify out-of-range test results. J Med Internet Res 2014 Aug 08;16(8):e187 [FREE Full text] [doi: 10.2196/jmir.3241] [Medline: 25135688]
12. Zhang Z, Citardi D, Xing A, Luo X, Lu Y, He Z. Patient challenges and needs in comprehending laboratory test results: mixed methods study. J Med Internet Res 2020 Dec 07;22(12):e18725 [FREE Full text] [doi: 10.2196/18725] [Medline: 33284117]
13. Fraccaro P, Vigo M, Balatsoukas P, van der Veer SN, Hassan L, Williams R, et al. Presentation of laboratory test results in patient portals: influence of interface design on risk interpretation and visual search behaviour. BMC Med Inform Decis Mak 2018 Feb 12;18(1):11 [FREE Full text] [doi: 10.1186/s12911-018-0589-7] [Medline: 29433495]
14. Bar-Lev S, Beimel D. Numbers, graphs and words - do we really understand the lab test results accessible via the patient portals? Isr J Health Policy Res 2020 Oct 28;9(1):58 [FREE Full text] [doi: 10.1186/s13584-020-00415-z] [Medline: 33115536]
15. Giardina TD, Baldwin J, Nystrom DT, Sittig DF, Singh H. Patient perceptions of receiving test results via online portals: a mixed-methods study. J Am Med Inform Assoc 2018 Apr 01;25(4):440-446 [FREE Full text] [doi: 10.1093/jamia/ocx140] [Medline: 29240899]
16. Doi T, Langsted A, Nordestgaard BG. Elevated remnant cholesterol reclassifies risk of ischemic heart disease and myocardial infarction. J Am Coll Cardiol 2022 Jun 21;79(24):2383-2397 [FREE Full text] [doi: 10.1016/j.jacc.2022.03.384] [Medline: 35710189]
17. Wadström BN, Wulff AB, Pedersen KM, Jensen GB, Nordestgaard BG. Elevated remnant cholesterol increases the risk of peripheral artery disease, myocardial infarction, and ischaemic stroke: a cohort-based study. Eur Heart J 2022 Sep 07;43(34):3258-3269. [doi: 10.1093/eurheartj/ehab705] [Medline: 34661640]
18. Chu SK, Huang H, Wong WN, van Ginneken WF, Wu KM, Hung MY. Quality and clarity of health information on Q and A sites. Libr Inf Sci Res 2018 Jul;40(3-4):237-244. [doi: 10.1016/j.lisr.2018.09.005]
19. Oh S, Yi YJ, Worrall A. Quality of health answers in social Q and A. Proc Assoc Inf Sci Technol 2013 Jan 24;49(1):1-6. [doi: 10.1002/meet.14504901075]
20. Tao D, Yuan J, Qu X, Wang T, Chen X. Presentation of personal health information for consumers: an experimental comparison of four visualization formats. In: Proceedings of the 15th International Conference on Engineering Psychology and Cognitive Ergonomics. 2018 Presented at: EPCE '18; July 15-20, 2018; Las Vegas, NV p. 490-500 URL: https://link.springer.com/chapter/10.1007/978-3-319-91122-9_40
21. Struikman B, Bol N, Goedhart A, van Weert JC, Talboom-Kamp E, van Delft S, et al. Features of a patient portal for blood test results and patient health engagement: web-based pre-post experiment. J Med Internet Res 2020 Jul 20;22(7):e15798 [FREE Full text] [doi: 10.2196/15798] [Medline: 32706704]

XSL•FO
RenderX

22. Kopanitsa G. Study of patients' attitude to automatic interpretation of laboratory test results and its influence on follow-up rate. BMC Med Inform Decis Mak 2022 Mar 27;22(1):79 [FREE Full text] [doi: 10.1186/s12911-022-01805-w] [Medline: 35346173]
23. Zikmund-Fisher BJ, Scherer AM, Witteman HO, Solomon JB, Exe NL, Fagerlin A. Effect of harm anchors in visual displays of test results on patient perceptions of urgency about near-normal values: experimental study. J Med Internet Res 2018 Mar 26;20(3):e98 [FREE Full text] [doi: 10.2196/jmir.8889] [Medline: 29581088]
24. Morrow D, Azevedo RF, Garcia-Retamero R, Hasegawa-Johnson M, Huang T, Schuh W, et al. Contextualizing numeric clinical test results for gist comprehension: implications for EHR patient portals. J Exp Psychol Appl 2019 Mar;25(1):41-61. [doi: 10.1037/xap0000203] [Medline: 30688498]
25. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023 Nov 22;25(1):bbad493. [doi: 10.1093/bib/bbad493] [Medline: 38168838]
26. Cadamuro J, Cabitza F, Debeljak Z, De Bruyne SD, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) working group on artificial intelligence (WG-AI). Clin Chem Lab Med 2023 Jun 27;61(7):1158-1166 [FREE Full text] [doi: 10.1515/cclm-2023-0355] [Medline: 37083166]
27. Munoz-Zuluaga C, Zhao Z, Wang F, Greenblatt MB, Yang HS. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine. Clin Chem 2023 Aug 02;69(8):939-940. [doi: 10.1093/clinchem/hvad058] [Medline: 37231970]
28. Zhang Z, Lu Y, Kou Y, Wu DT, Huh-Yoo J, He Z. Understanding patient information needs about their clinical laboratory results: a study of social Q and A site. Stud Health Technol Inform 2019 Aug 21;264:1403-1407 [FREE Full text] [doi: 10.3233/SHTI190458] [Medline: 31438157]
29. Zhang Z, Lu Y, Wilson C, He Z. Making sense of clinical laboratory results: an analysis of questions and replies in a social Q and A community. Stud Health Technol Inform 2019 Aug 21;264:2009-2010. [doi: 10.3233/SHTI190759] [Medline: 31438453]
30. Kurstjens S, Schipper A, Krabbe J, Kusters R. Predicting hemoglobinopathies using ChatGPT. Clin Chem Lab Med 2024 Feb 26;62(3):e59-e61 [FREE Full text] [doi: 10.1515/cclm-2023-0885] [Medline: 37650428]
31. He Z, Tian S, Erdengasileng A, Hanna K, Gong Y, Zhang Z, et al. Annotation and information extraction of consumer-friendly health articles for enhancing laboratory test reporting. AMIA Annu Symp Proc 2023;2023:407-416 [FREE Full text] [Medline: 38222337]
32. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
33. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv Preprint posted online on April 6, 2019 [FREE Full text] [doi: 10.18653/v1/w19-1909]
34. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv Preprint posted online on March 26, 2019 [FREE Full text]
35. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2021 Oct 15;3(1):1-23 [FREE Full text] [doi: 10.1145/3458754]
36. OpenAI. GPT-4 technical report. arXiv Preprint posted online on March 15, 2023 [FREE Full text]
37. Ye J, Chen X, Xu N, Zu C, Shao Z, Liu S, et al. A comprehensive capability analysis of GPT-3 and GPT-3. arXiv Preprint posted online on March 18, 2023 [FREE Full text]
38. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv Preprint posted online on July 18, 2023 [FREE Full text]
39. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca -- an open-source collection of medical conversational AI models and training data. arXiv Preprint posted online on April 14, 2023 [FREE Full text]
40. orca_mini_3b. Hugging Face. URL: https://huggingface.co/pankajmathur/orca_mini_3b [accessed 2023-12-04]
41. LangChain: introduction and getting started. Pinecone. URL: https://www.pinecone.io/learn/series/langchain/langchain-intro/ [accessed 2023-12-04]
42. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Presented at: ALC '02; July 7-12, 2002; Philadelphia, PA p. 311-318 URL: https://dl.acm.org/doi/10.3115/1073083.1073135 [doi: 10.3115/1073083.1073135]
43. Post M. A call for clarity in reporting BLEU scores. arXiv Preprint posted online on April 23, 2018 [FREE Full text] [doi: 10.18653/v1/w18-6319]
44. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005 Presented at: WIEEMMTS '05; June 29, 2005; Ann Arbor, MI p. 65-72 URL: https://aclanthology.org/W05-0909 [doi: 10.3115/1626355.1626389]

XSL•FO
RenderX

45. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Lin CY, editor. Text Summarization Branches Out Internet. Barcelona, Spain: Association for Computational Linguistics; 2004:74-81.
46. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv Preprint posted online on April 21, 2019 [FREE Full text]
47. Wang T, Yu P, Tan XE, O'Brien S, Pasunuru R, Dwivedi-Yu J, et al. Shepherd: a critic for language model generation. arXiv Preprint posted online on August 8, 2023 [FREE Full text]
48. Dubois Y, Li X, Taori R, Zhang T, Gulrajani I, Ba J, et al. AlpacaFarm: a simulation framework for methods that learn from human feedback. arXiv Preprint posted online on May 22, 2023 [FREE Full text]
49. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966 Aug;19(1):3-11 [FREE Full text] [doi: 10.2466/pr0.1966.19.1.3] [Medline: 5942109]
50. Gamer M, Lemon J, Singh IF. irr: various coefficients of interrater reliability and agreement. Cran R Project. 2019. URL: https://cran.r-project.org/web/packages/irr/index.html [accessed 2023-12-12]
51. Human subject regulations decision charts: 2018 requirements. Office for Human Research Protection. 2019 Jan 20. URL: https://tinyurl.com/3sbzydm3 [accessed 2024-04-03]
52. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? J Am Soc Nephrol 2023 Aug 01;34(8):1302-1304. [doi: 10.1681/ASN.0000000000000166] [Medline: 37254254]

Abbreviations

AI: artificial intelligence
autoML: automated machine learning
BERTScore: Bidirectional Encoder Representations from Transformers Score
BioBERT: biomedical Bidirectional Encoder Representations from Transformers
BLEU: Bilingual Evaluation Understudy
ClinicalBERT: clinical Bidirectional Encoder Representations from Transformers
HbA_{1c}: glycated hemoglobin
ICC: intraclass correlation coefficient
LLM: large language model
METEOR: Metric for Evaluation of Translation With Explicit Ordering
PubMedBERT: PubMed-trained Bidirectional Encoder Representations from Transformers
Q&A: question-and-answer
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
SciBERT: scientific Bidirectional Encoder Representations from Transformers

Edited by B Puladi; submitted 23.01.24; peer-reviewed by Y Chen, Z Smutny; comments to author 01.02.24; revised version received 17.02.24; accepted 06.03.24; published 17.04.24.

Please cite as:
He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, Arguello LG, Murray P, Lu Z
Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study
J Med Internet Res 2024;26:e56655
URL: https://www.jmir.org/2024/1/e56655
doi:10.2196/56655
PMID:38630520

©Zhe He, Balu Bhasuran, Qiao Jin, Shubo Tian, Karim Hanna, Cindy Shavor, Lisbeth Garcia Arguello, Patrick Murray, Zhiyong Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ChatGPT's Performance in Cardiac Arrest and Bradycardia Simulations Using the American Heart Association's Advanced Cardiovascular Life Support Guidelines: Exploratory Study

Cecilia Pham[1], BS, EMT; Romi Govender[1], BSN; Salik Tehami[1], BS; Summer Chavez[1,2,3], DO, MPH, MPM; Omolola E Adepoju[1,2,3], MPH, PhD; Winston Liaw[1,2,3], MD, MPH

[1]Tilman J Fertitta Family College of Medicine, University of Houston, Houston, TX, United States

[2]Humana Integrated Health Sciences Institute, University of Houston, Houston, TX, United States

[3]Department of Health Systems and Population Health Sciences, Tilman J Fertitta Family College of Medicine, Houston, TX, United States

Corresponding Author:
Cecilia Pham, BS, EMT
Tilman J Fertitta Family College of Medicine
University of Houston
5055 Medical Circle
Houston, TX, 77204
United States
Phone: 1 713 743 7047
Email: cmpham4@uh.edu

Abstract

Background: ChatGPT is the most advanced large language model to date, with prior iterations having passed medical licensing examinations, providing clinical decision support, and improved diagnostics. Although limited, past studies of ChatGPT's performance found that artificial intelligence could pass the American Heart Association's advanced cardiovascular life support (ACLS) examinations with modifications. ChatGPT's accuracy has not been studied in more complex clinical scenarios. As heart disease and cardiac arrest remain leading causes of morbidity and mortality in the United States, finding technologies that help increase adherence to ACLS algorithms, which improves survival outcomes, is critical.

Objective: This study aims to examine the accuracy of ChatGPT in following ACLS guidelines for bradycardia and cardiac arrest.

Methods: We evaluated the accuracy of ChatGPT's responses to 2 simulations based on the 2020 American Heart Association ACLS guidelines with 3 primary outcomes of interest: the mean individual step accuracy, the accuracy score per simulation attempt, and the accuracy score for each algorithm. For each simulation step, ChatGPT was scored for correctness (1 point) or incorrectness (0 points). Each simulation was conducted 20 times.

Results: ChatGPT's median accuracy for each step was 85% (IQR 40%-100%) for cardiac arrest and 30% (IQR 13%-81%) for bradycardia. ChatGPT's median accuracy over 20 simulation attempts for cardiac arrest was 69% (IQR 67%-74%) and for bradycardia was 42% (IQR 33%-50%). We found that ChatGPT's outputs varied despite consistent input, the same actions were persistently missed, repetitive overemphasis hindered guidance, and erroneous medication information was presented.

Conclusions: This study highlights the need for consistent and reliable guidance to prevent potential medical errors and optimize the application of ChatGPT to enhance its reliability and effectiveness in clinical practice.

(J Med Internet Res 2024;26:e55037) doi:10.2196/55037

KEYWORDS

ChatGPT; artificial intelligence; AI; large language model; LLM; cardiac arrest; bradycardia; simulation; advanced cardiovascular life support; ACLS; bradycardia simulations; America; American; heart association; cardiac; life support; exploratory study; heart; heart attack; clinical decision support; diagnostics; algorithms

XSL•FO
RenderX

Introduction

In March 2023, OpenAI, an artificial intelligence (AI) research laboratory, released GPT-4, an experimental version of ChatGPT, a large language model chatbot [1]. Compared with prior models including its predecessor GPT-3.5, GPT-4 has higher accuracy, greater reasoning capabilities, superior confidence, and more safety features with company data demonstrating improved performance on the Uniform Bar Exam and Biology Olympiad [2,3]. Compared with other types of learning models, deep learning models use multiple processing layers and artificial neural networks to process large amounts of data in order for the AI to learn [4]. In health care, ChatGPT shows promising applications including passing the United States Medical Licensing Examination, individualized health advice, improved diagnostic care, faster pharmacologic discovery, task automation, and clinical decision support [5-10]. However, limited evidence exists regarding ChatGPT's ability to perform in advanced clinical scenarios, including cardiac emergencies.

Currently, the standard of care follows algorithms from the 2020 American Heart Association (AHA) advanced cardiovascular life support (ACLS) guidelines. Learners enrolled in the AHA course take a multiple-choice cognitive examination followed by a series of scenario-based patient cases and skills demonstrations [11]. Although an early study by Fijačko et al [12] concluded ChatGPT was unable to pass the AHA's basic life support and advanced cardiovascular life support (ACLS) examinations, they demonstrated ChatGPT was able to provide relevant and accurate explanations for scenario-based questions without learning from an AHA course. With modification, Zhu et al [13] determined ChatGPT was able to pass the AHA's examinations after changing the research protocol in which 3 responses were generated for each question instead of 1, and assessments were turned into open-ended questions. While rudimentary tests find ChatGPT can quickly give well-tailored answers to questions and scenarios regarding cardiopulmonary resuscitation (CPR) based on published guidelines, this has not been rigorously applied to more complex clinical scenarios requiring higher levels of expertise and training [8].

Even with advances in modern health care, heart disease has remained the leading cause of death for several decades [14]. Approximately 436,000 Americans die from cardiac arrest each year, making it a significant public health issue [15]. Past efforts at reducing cardiac arrests include establishing cardiac arrest registries, increasing public awareness and educational efforts, improving the quality of resuscitation care, strengthening emergency medical services, quality improvement programs, and implementing accreditation standards [16]. ACLS protocols have been well-proven to improve cardiac arrest outcomes, with deviations from the guidelines associated with poorer survival and neurologic outcomes [17-22]. Thus, identifying technologies that can improve adherence to ACLS protocols and in turn, cardiac arrest outcomes, is critical. Cognitive aids, such as pocket cards, mobile apps, and augmented reality glasses, have been shown to increase adherence to ACLS guidelines and the quality of resuscitation in adult simulations for in-hospital cardiac arrests [23-28]. ChatGPT could prove to be a valuable resource by providing real-time decision support in resource-limited settings or augmenting clinical decision-making. Our study aims to analyze the accuracy of ChatGPT's performance in following the AHA's ACLS algorithms in 2 cardiac event simulations—cardiac arrest and bradycardia.

Methods

We evaluated the accuracy of ChatGPT's responses to 2 simulations based on the 2020 AHA ACLS guidelines with 3 primary outcomes of interest as follows: the mean individual step accuracy, the accuracy score per simulation attempt, and the accuracy score for each algorithm.

Ethical Considerations

In accordance with 45 CFR 46, Subpart A, also known as the Common Rule, our research project involving educational tests, such as ACLS simulations, falls under the exemption outlined in 46.104(d)(1). This exemption applies because the interactions in our study are limited to educational tests and do not involve identifiable human subjects. Because of this, no ethics board review was required for this research.

Initial Observations of ChatGPT's Responses and Adherence to AHA Guidelines

The research team underwent an initial testing phase involving several procedures to observe ChatGPT's responses before developing the scripts used for testing. Initially, two types of trials were proposed: (1) assessing ChatGPT's ability to accurately identify rhythm strips, and (2) evaluating ChatGPT's capability to execute the algorithm provided by the AHA. However, from January to May 2023, ChatGPT lacked the ability to recognize images, including rhythm strips or electrocardiograms (ECGs). Due to this limitation, specific ECG rhythms were provided to ChatGPT without the need for correct identification. The team then tested ChatGPT's ability to execute basic algorithms, including those for bradycardia, tachycardia with a pulse, and cardiac arrest, based on the 2020 AHA Guidelines. At the beginning of the trials, it was confirmed by ChatGPT that it would refer to the 2020 AHA Guidelines for information. During these initial trials, testers observed significant differences between the responses generated by GPT-3.5 and GPT-4. GPT-4 demonstrated greater adherence to the 2020 AHA Guidelines by being able to interpret the algorithmic pathways more accurately. Additionally, testers noted that prompting ChatGPT led to lengthy outputs, unlike the step-by-step approach used in real-life simulations. To emulate real-life testing conditions in ACLS training, the command "What is the next step? Give me one step at a time" was included. This method also mimics how ChatGPT could be used in the real world. As GPT-4 closely followed the 2020 AHA Guidelines and was able to respond to the step-by-step command, it was selected for the actual simulations.

Development of Scripts for User Input and ChatGPT Output

Following the completion of the initial trials, the study authors with prior ACLS training (SC, WL, and CP) developed scripts

XSL•FO
RenderX

based on the testers' observations, the AHA's algorithms, and clinical experience. SC and WL are both physicians in emergency medicine and family medicine, respectively, while CP holds an EMT license with paramedic-level training. The 2 scripts exclusively focused on testing 2 categories of cardiac arrhythmias: bradycardia and cardiac arrest (Multimedia Appendices 1 and 2). To validate the scripts, 3 physicians with ACLS and emergency medicine backgrounds assessed them to ensure they accurately depict real-world clinical scenarios and management practices. The study authors developed a scoring guide to measure the accuracy of the responses of ChatGPT to prespecified prompts. The number of attempts given to ChatGPT was matched to the required correct outputs for each group. For instance, if 5 correct outputs were expected, ChatGPT was prompted 5 times, allowing it the same number of opportunities to provide accurate responses. In instances where ChatGPT's outputs diverged from the predefined script, the authors responded by typing "Give me one step at a time" or "Give me another option."

Data Collection

Testing was conducted between May to August 2023 using the public web user chat interface by OpenAI, with upgraded accounts to use GPT-4. All inputs and outputs for 1 simulation are saved within a single conversation thread, and each simulation was conducted in a separate thread. Three testers underwent training to adhere to the scripts, ensuring standardized procedures for the simulations. Each scenario was tested 20 times to evaluate the variability of ChatGPT's responses and its ability to provide correct outputs. For each simulation step, ChatGPT was scored for correctness (1 point) or incorrectness (0 points), and these scores were recorded in an Excel (Microsoft Corp) spreadsheet. Partially correct responses were considered incorrect. Within each scoring section, qualitative comments regarding the incorrect responses were also recorded. Additionally, instances where the prompt "Give me one step at a time" requiring repeated inputs were recorded. The full transcripts for each simulation were collated into separate document files for each simulation attempt. The authors of the scripts independently reviewed the recorded accuracy scores in the spreadsheets and the transcripts saved in the document files to ensure the validity of the results and to maintain accurate record-keeping.

Data Analysis

Using the data recorded in the Excel spreadsheets, accuracy scores for individual steps, individual simulation attempts, and overall simulation attempts for each algorithm were calculated. The overall score per simulation attempt was calculated by summing the correct responses and dividing them by the number of steps (12 for bradycardia and 39 for cardiac arrest). For each step in the simulation, we similarly summed the correct responses and divided them by the number of simulations (n=20). Mean overall simulation accuracies were determined by averaging 20 simulation scores. The median accuracies for each algorithm were also calculated. Qualitative comments recorded for each simulation were also analyzed to identify recurring themes to describe ChatGPT's performance.

Results

Overview

The individual step accuracy per simulation attempt for the cardiac arrest and bradycardia algorithms are reported in Tables 1 and 2, respectively. ChatGPT's median accuracy for each step was 85% (IQR 40%-100%) for cardiac arrest and 30% (IQR 13%-81%) for bradycardia. The accuracy scores per simulation attempt for each algorithm are described in Table 3. ChatGPT's median accuracy for over 20 simulation attempts for cardiac arrest was 69% (IQR 67%-74%) and for bradycardia was 42% (IQR 33%-50%). Four key findings were identified after analyzing the results from the 2 sets of simulations: (1) ChatGPT's outputs varied despite consistent input, (2) the same actions were persistently missed, (3) repetitive overemphasis hindered guidance, and (4) erroneous medication information was presented.

XSL•FO
RenderX

Table 1. Individual step accuracy using ChatGPT in cardiac arrest advanced cardiovascular life support simulations (N=20).

Correct simulation output	Accuracy, n (%)[a]
Part 1: Initiation of CPR[b]	
Assess for responsiveness	20 (100)
Assess for a pulse	20 (100)
Activate an emergency response system	19 (95)
Start CPR	20 (100)
Apply the defibrillator	20 (100)
Obtain intravenous or intraosseous access	2 (10)
Part 2: First dose of epinephrine	
Assess for responsiveness	0 (0)
Assess for a pulse	4 (20)
Continue CPR	17 (85)
Give 1 mg of epinephrine	19 (95)
Consider advanced airway management	13 (65)
“Continue CPR” after completing Part 2	16 (80)
Part 3: First rhythm check	
Perform a rhythm check	20 (100)
Consider reversible causes	8 (40)
Continue CPR	20 (100)
Part 4: Second dose of epinephrine	
Perform a rhythm check	12 (60)
Immediately defibrillate	11 (55)
Give 1 mg of epinephrine	15 (75)
Consider reversible causes	4 (20)
“Continue CPR” after completing Part 4	20 (100)
Part 5: Second defibrillation	
Perform a rhythm check	18 (90)
Immediately defibrillate	19 (95)
“Continue CPR” after completing Part 5	20 (100)
Part 6: First dose of alternative pharmacologic agents	
Give 1 mg of epinephrine	6 (30)
Consider giving amiodarone or lidocaine	9 (45)
Consider reversible causes	0 (0)
Perform a rhythm check	20 (100)
Immediately defibrillate	20 (100)
“Continue CPR” after completing Part 6	16 (80)
Part 7: Additional pharmacologic agents	
Perform a rhythm check	19 (95)
Immediately defibrillate	18 (90)
Consider reversible causes	2 (10)
Give a second dose of amiodarone	6 (30)
“Continue CPR” after completing Part 7	19 (95)

XSL•FO
RenderX

Correct simulation output	Accuracy, n (%)[a]
Part 8: End of simulation	
Give 1 mg of epinephrine	14 (70)
Consider giving lidocaine	0 (0)
Consider reversible causes	11 (55)
Perform a rhythm check	18 (90)
"Continue CPR" after completing Part 8	20 (100)

[a]Median accuracy is 85% (IQR 40%-100%); mean accuracy is 69%.

[b]CPR: cardiopulmonary resuscitation.

Table 2. Individual step accuracy using ChatGPT in bradycardia advanced cardiovascular life support simulations (N=20). Mean individual step accuracy using ChatGPT in bradycardia advanced cardiovascular life support simulations (N=20).

Correct simulation output	Accuracy, n (%)[a]
Part 1: Address ABC's of resuscitation	
Maintain the patient's airway; assists breathing as necessary	16 (80)
Apply a cardiac monitor to identify rhythm	3 (15)
Monitor blood pressure, pulse oximetry, and other vital signs	9 (45)
Obtain intravenous access	7 (35)
Obtain 12-lead ECG[b]	0 (0)
Give 1 mg of atropine	0 (0)
Part 2: Address hypoxia	
Give oxygen	20 (100)
Part 3: Second dose of atropine	
Give an additional dose of atropine	19 (95)
Part 4: Alternative pharmacologic agents and consideration of pacing	
Give an additional dose of atropine	3 (15)
Consider transcutaneous pacing	17 (85)
Give dopamine or epinephrine	1 (5)
Part 5: Seek expert consultation	
Seek a consultation	5 (25)

[a]Median accuracy is 30% (IQR 13%-81.3%); mean accuracy is 42%.

[b]ECG: electrocardiogram.

XSL•FO
RenderX

Table 3. Accuracy score per simulation attempt for ChatGPT in cardiac arrest (N=39) and bradycardia (N=12) advanced cardiovascular life support simulations.

Simulation number	Accuracy of cardiac arrest, n (%)[a]	Accuracy of bradycardia, n (%)[b]
1	24 (62)	7 (58)
2	26 (67)	5 (42)
3	24 (62)	6 (50)
4	29 (74)	6 (50)
5	27 (69)	5 (42)
6	27 (69)	7 (58)
7	21 (54)	4 (33)
8	29 (74)	4 (33
9	26 (67)	7 (58)
10	27 (69)	6 (50)
11	26 (67)	5 (42)
12	26 (67)	5 (42)
13	22 (56)	5 (42)
14	26 (67)	5 (42)
15	27 (69)	3 (25)
16	29 (74)	7 (58)
17	29 (74)	3 (25)
18	30 (77)	5 (42)
19	30 (77)	3 (25)
20	30 (77)	4 (33)
Average accuracy	535/780 (69)	102/240 (43)

[a]Median accuracy of cardiac arrest is 69% (IQR 67%-74%).

[b]Median accuracy of bradycardia is 42% (IQR 33%-50%).

ChatGPT's Outputs Varied

ChatGPT gave varying responses to identical inputs, sometimes providing unexpected responses. ChatGPT generated multistep instructions, prompting the request "give me one step at a time." It would sometimes provide entirely different steps that were not in its original multistep output. There were, on average, 0.5 instances per simulation of repetitive loops of "check heart rhythm" and "resume CPR." Even when "one step at a time" was requested, ChatGPT was unable to consistently execute algorithms step by step, often generating multistep instructions. ChatGPT was instructed to give "one step at a time" 3.6 times per simulation in the cardiac arrest case and 1.5 times per simulation in the bradycardia case. Other unexpected outputs include 0.5 instances per simulation of oxygen and medication delivery instructions preceding necessary steps such as applying a monitor to the patient or starting an intravenous (IV) line. ChatGPT provided additional instructions that were not part of the ACLS algorithms, including 0.3 instances per simulation of "start an isoproterenol infusion."

Some Actions Were Persistently Missed

Throughout the simulations, there were several notable omissions to ChatGPT's output. ChatGPT frequently skipped the instruction to establish IV or intraosseous (IO) access, before directing IV or IO medication administration at 0% accuracy in cardiac arrest and 35% (n=7) accuracy in bradycardia. ChatGPT also struggled to incorporate additional dosages of medications in its responses. For epinephrine administration, ChatGPT was accurate 95% (n=19) of the time regarding the first dose, but its accuracy decreased to 75% (n=15) and 70% (n=14) for the second and third doses, respectively. In the bradycardia simulations, ChatGPT consistently recommended a second dose of atropine at 95% (n=19) accuracy but rarely advised a third dose at 15% (n=3) accuracy. Furthermore, ChatGPT was never prompted to obtain a 12-lead ECG with 0% accuracy. ChatGPT did not provide any instructions for checking vital signs and only suggested pulse checks for the patient with bradycardia, omitting other vital signs such as blood pressure and oxygen saturation.

Repetitive Overemphasis Hindered Guidance

ChatGPT consistently overemphasized specific steps to the detriment of other important interventions. During the cardiac arrest simulations, ChatGPT became fixed in repetitive cycles when given the prompt, "CPR resumed. It's been two minutes. What is the next immediate step?" ChatGPT would then incessantly advise to "check the heart rhythm" and "resume

XSL•FO
RenderX

CPR" on average 4 times per simulation without providing any other instructions such as considering advanced airway management at 65% (n=13) accuracy and exploring reversible causes with 25% (n=5) accuracy. Similarly, in the bradycardia simulation, ChatGPT emphasized transcutaneous and transvenous pacing methods on average 1.8 times per simulation, disregarding other reasonable alternatives for bradycardia management, such as pharmacologic interventions.

Erroneous Medication Information Was Presented

Two medication instructions given in the bradycardia simulations were consistently inaccurate. In all bradycardia simulations, ChatGPT advised administering 0.5 mg of atropine instead of the recommended 1 mg dose in the most recent 2020 AHA ACLS guidelines (0% accuracy). Despite being corrected and acknowledging the accurate dosage as 1 mg, ChatGPT persisted in administering the incorrect dose of atropine in subsequent outputs. In 95% (n=19) of bradycardia simulations, ChatGPT provided incorrect guidance to administer 2-20 µg/kg/minute of dopamine, deviating from the correct range of 5-20 µg/kg/minute. However, in all cardiac arrest simulations, all medication dosages were correctly identified by ChatGPT.

Discussion

Principal Findings

This study sought to examine the accuracy of ChatGPT's performance in adult cardiac event simulations for cardiac arrest and bradycardia. Although the overall accuracy of ChatGPT was lower than the standard passing threshold in academics of 70%, the demonstrated accuracy of ChatGPT was comparable to one other study of complex cardiac clinical vignettes which had 50% accuracy (50/100). Compared with experts, ChatGPT gave inaccurate or incomplete responses [29]. Even with these caveats, this proof-of-concept study illustrates the potential and perils of using ChatGPT for real-time decision support in clinical settings.

Our study adds to the small but growing body of literature analyzing the potential clinical role of ChatGPT and other AI technologies, with a special focus on cardiac emergencies and complex decision-making. While this field may be in its infancy, ChatGPT builds upon prior generations of more low-tech cognitive aids such as apps or paper cards by offering faster responses, tailored answers, and the ability to work through far more complicated scenarios. In several steps of the simulations, ChatGPT's accuracy was as high as 100% (n=20), an impressive benchmark.

In practice settings with limited resources, ChatGPT, AI, and related technologies may still help fill a crucial gap by acting as a more advanced cognitive aid compared with previously studied solutions. They can also be used in medical education, quality initiatives, and simulation training exercises. However, the findings from this study highlight several key issues that need to be addressed before implementing this decision aid in patient care.

In our study, ChatGPT exhibited a wide range of clinical accuracy, which can be due in part to outdated or unreliable training data. If AI is used to augment clinical decision-making, physicians must ensure the appropriateness and validity of the training data. Without this level of accountability, health care professionals could accept hallucinations, or AI responses that are incorrect or misleading. While subject matter experts may detect these mistakes, those with less training and experience may fail to recognize these inconsistencies, which can lead to missing steps or delivering inaccurate medication doses. Additionally, if the training data contains inaccuracies or reflects certain biases, ChatGPT may replicate those errors or biases in its responses. It is also unclear how frequently ChatGPT updates or how it decides to include new or controversial scientific findings in its model. This is especially salient in cardiac emergencies, where quick and accurate decision-making is crucial for patient survival.

Our findings suggest that ChatGPT has difficulty learning from past scenarios. For example, ChatGPT would acknowledge the correct dosage if it initially gave an incorrect one, but persistently administered the same incorrect doses across subsequent trials. Real-life ACLS scenarios involve complex medical situations with multiple possible interventions depending on the patient's status. Patients with cardiac arrest secondary to hyperkalemia or pulmonary embolism are managed differently than the simpler cases created for this simulation. Achieving higher levels of accuracy would either require the ability for ChatGPT to learn or a different AI model altogether.

Some experts recommend asking probing questions to AI as a possible solution to verify responses from ChatGPT. In clinical scenarios where time directly correlates with myocardial health and patient outcomes, this may be unrealistic. Continuing to prompt the AI or slightly modifying responses to see how responses change is also not practical. In our study, ChatGPT was prompted dozens of times for responses. During an actual episode of cardiac arrest, where intervals of CPR are done in 2 minutes, spending most of the time repeatedly prompting ChatGPT for an accurate response is probably not time well spent during the resuscitation.

Several limitations should be considered when interpreting these results. First, our study focused only on 2 ACLS scenarios: bradycardia and cardiac arrest. Because this is a proof-of-concept study, the authors decided to test just 2 scenarios. In the overall ACLS framework, there are many other emergencies covered, including the management of tachyarrhythmias, opioid overdose, and stroke. Given that ChatGPT's performance could vary in these alternative scenarios, broadening the scope of simulations might provide a more thorough insight into ChatGPT's capabilities by evaluating its performance across a broader range of ACLS conditions. Our study analyzed the accuracy of ChatGPT only, making it difficult to generalize the results to other large language models. During the study period, ChatGPT could not recognize images, requiring the study authors to specify the cardiac rhythm in the model, potentially affecting the accuracy. There are growing concerns that the quality of ChatGPT has decreased over time as it has become more widely used by the public, which could also impact the results [30]. Additionally, given the potential evolution of ChatGPT's capabilities over time, the reliability and validity of the study's results may be affected. Exploring how AI models can update their knowledge

XSL•FO
RenderX

could provide insights into their long-term utility in real-world medical emergency settings.

Conclusions

Bridging the gap between AI and human language is a multifaceted challenge, balancing advancements, data quality, and human oversight to maximize ChatGPT's benefits and minimize risks in health care. This study highlights the need for consistent and reliable guidance to prevent potential medical errors and optimize the application of ChatGPT to enhance its reliability and effectiveness in clinical practice. To enhance ChatGPT's real-world effectiveness, significant improvements are needed, particularly in accuracy and consistency across diverse medical situations. While ChatGPT holds promise as a decision support tool that can provide structured clinical guidance, it should complement, not replace, qualified health care professionals. Future research also should aim to address these limitations and further investigate the challenges of AI in health care to ensure its safe and effective use. Additionally, future studies comparing ChatGPT's performance with other AI models or with human performance could offer valuable insights.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT ACLS Testing Algorithm for Bradycardia.
[PNG File , 175 KB - jmir_v26i1e55037_app1.png]

Multimedia Appendix 2
ChatGPT ACLS Testing Algorithm for Cardiac Arrest.
[PNG File , 311 KB - jmir_v26i1e55037_app2.png]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-08-07]
2. GPT-4. AI O. 2023. URL: https://openai.com/gpt-4 [accessed 2023-10-20]
3. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks, Stand Eval 2023;3(1):100105 [FREE Full text] [doi: 10.1016/j.tbench.2023.100105]
7. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
8. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023;185:109729 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
9. Waqas. AI in healthcare: ChatGPT helps boy get diagnosis after doctors fail. HACK Read. 2023. URL: https://www.hackread.com/artificial-intelligence-healthcare-chatgpt-boy-diagnosis/ [accessed 2024-04-04]
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
11. 2020 HeartCode® courses FAQs. cpr.heart.org. 2020. URL: https://cpr.heart.org/-/media/cpr-files/course-faqs/2020-course-faqs/2020-heartcode-courses-faqs_102120.pdf?la=en [accessed 2023-09-11]
12. Fijačko N, Gosak L, Štiglic G, Picard CT, Douma MJ. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation 2023;185:109732 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109732] [Medline: 36775020]
13. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023;188:109783 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]

XSL•FO
RenderX

14. Heart disease deaths—health, United States. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/hus/topics/heart-disease-deaths.htm [accessed 2023-08-07]
15. CPR facts and stats. cpr.heart.org. URL: https://cpr.heart.org/en/resources/cpr-facts-and-stats [accessed 2024-04-02]
16. Committee on the Treatment of Cardiac Arrest: Current Status and Future Directions, Board on Health Sciences Policy, Institute of Medicine. Recommendations and key opportunities. In: Graham R, McCoy MA, Schultz AM, editors. Strategies to Improve Cardiac Arrest Survival: A Time to Act. vol 7. Washington (DC): National Academies Press (US); 2015.
17. History of CPR. Association AH. URL: https://cpr.heart.org/en/resources/history-of-cpr [accessed 2023-10-20]
18. Crowley CP, Salciccioli JD, Kim EY. The association between ACLS guideline deviations and outcomes from in-hospital cardiac arrest. Resuscitation 2020;153:65-70 [FREE Full text] [doi: 10.1016/j.resuscitation.2020.05.042] [Medline: 32502576]
19. Sanders AB, Berg RA, Burress M, Genova RT, Kern KB, Ewy GA. The efficacy of an ACLS training program for resuscitation from cardiac arrest in a rural community. Ann Emerg Med 1994;23(1):56-59. [doi: 10.1016/s0196-0644(94)70009-5] [Medline: 8273960]
20. Sodhi K, Singla MK, Shrivastava A. Impact of advanced cardiac life support training program on the outcome of cardiopulmonary resuscitation in a tertiary care hospital. Indian J Crit Care Med 2011;15(4):209-212 [FREE Full text] [doi: 10.4103/0972-5229.92070] [Medline: 22346031]
21. Honarmand K, Mepham C, Ainsworth C, Khalid Z. Adherence to Advanced Cardiovascular Life Support (ACLS) guidelines during in-hospital cardiac arrest is associated with improved outcomes. Resuscitation 2018;129:76-81. [doi: 10.1016/j.resuscitation.2018.06.005] [Medline: 29885353]
22. Hunt EA, Vera K, Diener-West M, Haggerty JA, Nelson KL, Shaffner DH, et al. Delays and errors in cardiopulmonary resuscitation and defibrillation by pediatric residents during simulated cardiopulmonary arrests. Resuscitation 2009;80(7):819-825. [doi: 10.1016/j.resuscitation.2009.03.020] [Medline: 19423210]
23. Corazza F, Fiorese E, Arpone M, Tardini G, Frigo AC, Cheng A, et al. The impact of cognitive aids on resuscitation performance in in-hospital cardiac arrest scenarios: a systematic review and meta-analysis. Intern Emerg Med 2022;17(7):2143-2158 [FREE Full text] [doi: 10.1007/s11739-022-03041-6] [Medline: 36031672]
24. Corazza F, Arpone M, Tardini G, Stritoni V, Mormando G, Graziano A, et al. Effectiveness of a novel tablet application in reducing guideline deviations during pediatric cardiac arrest: a randomized clinical trial. JAMA Netw Open 2023;6(8):e2327272 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.27272] [Medline: 37535352]
25. Siebert JN, Ehrler F, Gervaix A, Haddad K, Lacroix L, Schrurs P, et al. Adherence to AHA guidelines when adapted for augmented reality glasses for assisted pediatric cardiopulmonary resuscitation: a randomized controlled trial. J Med Internet Res 2017;19(5):e183 [FREE Full text] [doi: 10.2196/jmir.7379] [Medline: 28554878]
26. Siebert JN, Lacroix L, Cantais A, Manzano S, Ehrler F. The impact of a tablet app on adherence to American Heart Association guidelines during simulated pediatric cardiopulmonary resuscitation: randomized controlled trial. J Med Internet Res 2020;22(5):e17792 [FREE Full text] [doi: 10.2196/17792] [Medline: 32292179]
27. Hejjaji V, Malik AO, Peri-Okonny PA, Thomas M, Tang Y, Wooldridge D, et al. Mobile app to improve house officers' adherence to advanced cardiac life support guidelines: quality improvement study. JMIR Mhealth Uhealth 2020;8(5):e15762 [FREE Full text] [doi: 10.2196/15762] [Medline: 32427115]
28. Brophy SL, McCue MR, Reel RM, Jones TD, Dias RD. The impact of a smartphone-based cognitive aid on clinical performance during cardiac arrest simulations: a randomized controlled trial. AEM Educ Train 2023;7(3):e10880 [FREE Full text] [doi: 10.1002/aet2.10880] [Medline: 37292063]
29. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). medRxiv Preprint posted online on March 26, 2023 2023:1-17 [FREE Full text] [doi: 10.1101/2023.03.25.23285475]
30. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? ArXiv Preprint posted online on 31 Oct 2023 2023 [FREE Full text] [doi: 10.5260/chara.21.2.8]

Abbreviations

ACLS: advanced cardiovascular life support
AHA: American Heart Association
AI: artificial intelligence
CPR: cardiopulmonary resuscitation
ECG: electrocardiogram
IO: intraosseous
IV: intravenous

XSL•FO
RenderX

Edited by Q Jin; submitted 30.11.23; peer-reviewed by D Wright, TAR Sure, L Zhu, S Kommireddy; comments to author 08.02.24; revised version received 22.02.24; accepted 10.03.24; published 22.04.24.

Please cite as:
Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W
ChatGPT's Performance in Cardiac Arrest and Bradycardia Simulations Using the American Heart Association's Advanced Cardiovascular Life Support Guidelines: Exploratory Study
J Med Internet Res 2024;26:e55037
URL: https://www.jmir.org/2024/1/e55037
doi:10.2196/55037
PMID:38648098

©Cecilia Pham, Romi Govender, Salik Tehami, Summer Chavez, Omolola E Adepoju, Winston Liaw. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study

Annessa Kernberg[1], MD; Jeffrey A Gold[1], MD; Vishnu Mohan[1], MD

Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Sciences University, Portland, OR, United States

Corresponding Author:
Vishnu Mohan, MD
Department of Medical Informatics and Clinical Epidemiology
Oregon Health and Sciences University
3181 SW Sam Jackson Park Road
Portland, OR, 97239
United States
Phone: 1 5034944469
Email: mohanV@ohsu.edu

Abstract

Background: Medical documentation plays a crucial role in clinical practice, facilitating accurate patient management and communication among health care professionals. However, inaccuracies in medical notes can lead to miscommunication and diagnostic errors. Additionally, the demands of documentation contribute to physician burnout. Although intermediaries like medical scribes and speech recognition software have been used to ease this burden, they have limitations in terms of accuracy and addressing provider-specific metrics. The integration of ambient artificial intelligence (AI)–powered solutions offers a promising way to improve documentation while fitting seamlessly into existing workflows.

Objective: This study aims to assess the accuracy and quality of Subjective, Objective, Assessment, and Plan (SOAP) notes generated by ChatGPT-4, an AI model, using established transcripts of History and Physical Examination as the gold standard. We seek to identify potential errors and evaluate the model's performance across different categories.

Methods: We conducted simulated patient-provider encounters representing various ambulatory specialties and transcribed the audio files. Key reportable elements were identified, and ChatGPT-4 was used to generate SOAP notes based on these transcripts. Three versions of each note were created and compared to the gold standard via chart review; errors generated from the comparison were categorized as omissions, incorrect information, or additions. We compared the accuracy of data elements across versions, transcript length, and data categories. Additionally, we assessed note quality using the Physician Documentation Quality Instrument (PDQI) scoring system.

Results: Although ChatGPT-4 consistently generated SOAP-style notes, there were, on average, 23.6 errors per clinical case, with errors of omission (86%) being the most common, followed by addition errors (10.5%) and inclusion of incorrect facts (3.2%). There was significant variance between replicates of the same case, with only 52.9% of data elements reported correctly across all 3 replicates. The accuracy of data elements varied across cases, with the highest accuracy observed in the "Objective" section. Consequently, the measure of note quality, assessed by PDQI, demonstrated intra- and intercase variance. Finally, the accuracy of ChatGPT-4 was inversely correlated to both the transcript length (P=.05) and the number of scorable data elements (P=.05).

Conclusions: Our study reveals substantial variability in errors, accuracy, and note quality generated by ChatGPT-4. Errors were not limited to specific sections, and the inconsistency in error types across replicates complicated predictability. Transcript length and data complexity were inversely correlated with note accuracy, raising concerns about the model's effectiveness in handling complex medical cases. The quality and reliability of clinical notes produced by ChatGPT-4 do not meet the standards required for clinical use. Although AI holds promise in health care, caution should be exercised before widespread adoption. Further research is needed to address accuracy, variability, and potential errors. ChatGPT-4, while valuable in various applications, should not be considered a safe alternative to human-generated clinical documentation at this time.

(J Med Internet Res 2024;26:e54419) doi:10.2196/54419

XSL•FO
RenderX

KEYWORDS

generative AI; generative artificial intelligence; ChatGPT; simulation; large language model; clinical documentation; quality; accuracy; reproducibility; publicly available; medical note; medical notes; generation; medical documentation; documentation; documentations; AI; artificial intelligence; transcript; transcripts; ChatGPT-4

Introduction

Medical documentation is an integral aspect of clinical practice, ensuring accuracy and comprehensive patient management and serving as a communication tool among health care professionals. In recent years, it has become increasingly evident that inaccuracies in medical notes lead to miscommunication, diagnostic discrepancies, and patients' perceptions of subpar medical care [1]. Beyond the immediate implications of documentation errors, documentation demands have been identified as a significant contributor to physician burnout [2]. With health care professionals spending an increasing amount of their working hours on paperwork, there is less time and energy left for direct patient care.

To counter this, many institutions have adopted the use of intermediaries, such as medical scribes or speech recognition software, to shoulder the documentation load and allow clinicians to focus on patient interactions. However, both of these solutions have significant limitations and concerns regarding documentation accuracy and lack of impact on many provider-specific metrics surrounding after-hours charting and burnout [3,4]. In addition, the financial implications of employing medical scribes render them inaccessible to numerous health care practices [5]. Consequently, there is a continued search for innovative solutions to create effective and accurate documentation while seamlessly integrating into existing workflows.

With the rapid and exponential growth in computing capacity, artificial intelligence (AI) is being increasingly used in health care, holding the promise of revolutionizing medical documentation, thus potentially alleviating the burden on physicians [6]. AI-powered systems can analyze vast amounts of data quickly, identify patterns, and suggest diagnostic options. Although the allure of AI is undeniable, questions regarding its accuracy, reliability, and suitability in the clinical setting remain. The maturation of speech recognition technology has led to large-scale adoption by health care organizations, allowing for real-time transcription services. This, combined with software using large language models (LLMs), now enables the creation of structured medical notes in close temporal relation to the clinical encounter, thereby decreasing the clinician documentation burden [7,8]. Multiple software vendors are developing and deploying documentation assistance software powered by ambient AI, referred to as ambient digital scribes. There is already significant interest on the part of clinicians and health care organizations to adopt them. However, little data exist on the safety and quality of the documentation, with analysis made more difficult by the proprietary AI engine used by each vendor.

One such AI system is OpenAI's ChatGPT-4, a state-of-the-art LLM known for its ability to engage in text-based communication with users (as a chatbot), which is used in some commercial ambient digital scribe solutions. Released in November 2022, ChatGPT-4 is trained on a vast amount of text data from the internet and uses an LLM to answer the users' prompts. Health care providers envision numerous applications for ChatGPT-4, such as answering patient questions, automated insurance prior authorizations, and creating differential diagnoses [9,10]. It is important to note that open AI platforms, such as ChatGPT-4, are not recommended for clinical use due to the many regulatory and privacy issues. Despite this, there is a continued interest in whether ChatGPT-4 could serve as a freely available tool to assist as a documentation intermediary, bridging the gap between health care professionals and the tedious task of recordkeeping.

However, it is imperative that prior to the widespread adoption of these tools, their safety and efficacy need to be evaluated in a structured and clinically contextually relevant manner. Therefore, the goal of this study was to use transcripts from simulated patient-provider encounters to determine the accuracy, readability, and reproducibility of ChatGPT-4–generated Subjective, Objective, Assessment, and Plan (SOAP) notes.

Methods

Overview

As part of a project designed to evaluate the accuracy and efficacy of human scribe–generated notes, we created 14 simulated patient-provider encounters. All encounters used professional standardized patients and represented a wide range of ambulatory specialties. A standardized patient is an individual trained to simulate a medical scenario for health care education, assessment, and research. Briefly, for each case, a storyboard was created by subject matter experts and used for training the standardized patient to ensure standard content delivery according to best practices [11]. After an initial dry-run, each scenario was conducted in a simulated ambulatory patient exam room equipped with audio-video capture. At the end of the scenario, audio-video files were exported for use. These cases represented a variety of diagnoses (Table 1).

Audio files for each case were professionally transcribed. For each encounter, a list of key reportable elements was created for each case using the transcripts and informed by the initial storyboard, by 2 clinical experts of the study team. This is being used as the scoring rubric for subsequent analysis. These encounter transcripts were then fed into ChatGPT-4 using a standard prompt ("generate a clinical note in SOAP format for the following"). The SOAP format is a widely used clinical documentation format that concatenates data elements of the clinical interview into headers representing SOAP-related components. The SOAP format is a standard model for medical documentation, providing a clear, concise framework for health care professionals to record and share patient information. Each transcript was run through the model three times to assess output fidelity associated with replicability, thus generating three

XSL•FO
RenderX

documentation versions for each case for a total of 42 ChatGPT-4 generated SOAP notes (the prompt and full output are present in Multimedia Appendix 1). A new discussion space was created for each case to prevent the various transcripts from conflating each other. Each prompt request was conducted consecutively within the same discussion space.

After acquiring the generated notes, various comparisons were made to assess the output's accuracy and quality. Within a single case, the 3 versions were analyzed based on errors generated. A list of errors was defined as follows: (1) omissions—where expected documentation elements or data were missing; (2) incorrect—where the data element was referred to but incorrect; and (3) additions—information added that was not in the transcript. The framework for defining quality in clinical documentation based on omissions, incorrect information, and additions is a structured approach to evaluate the accuracy and completeness of medical records. These characteristics (omissions, incorrect information, and additions), if present, help define the quality of documentation given their implications. For example, omissions can lead to gaps in patient care, misdiagnosis, or delays in treatment. Incorrect information can compromise patient safety and lead to negative health consequences. Additions, while not always harmful, can be inaccurate and reduce the efficiency of care delivery. This framework is particularly useful in assessing the performance of health care documentation processes, such as those generated by medical scribes, and in quantifying appropriate information retrieval [4,12]. A correct data element was defined as one without the previously outlined errors.

Table 1. Case number and associated diagnosis (14 simulated patient-provider encounter transcripts representing a variety of diagnoses).

Case number	Diagnosis
1	Gastroenteritis
2	Incarcerated inguinal hernia
3	Diabetic ketoacidosis
4	Ovarian cyst
5	Pneumonia
6	Menstrual migraine
7	Breast mass
8	Heart failure
9	Polymyalgia rheumatica
10	Congestive heart failure
11	Decreased fetal movement
12	Diverticulitis
13	Scleroderma
14	Colon cancer

To ensure and assess note quality, we outlined critical data elements for each clinical case. Members of the study team independently selected these crucial data elements and subsequently compiled them to guarantee comprehensiveness. They used these elements to generate a gold standard checklist and an associated gold standard History and Physical Examination note. Then, 2 raters graded the 3 ChatGPT-4 versions of each encounter based on whether they correctly included, missed, or wrongly presented the corresponding data element. We enumerated the number of errors and correct data elements for each version. Afterward, we compared the correct data elements across the 3 ChatGPT-4 versions for presence and consistency, as follows: (1) across all three versions, (2) across two versions, (3) only in a single version, or (4) not present at all. Finally, we compared the percentages of appropriate data elements across the versions to the transcript's length and the number of data elements.

The data elements were divided into three documentation-related sections: (1) *Subjective*, further subdivided into the history of present illness and other patient-reported information, including medications, allergies, family history, social history, and past medical history; (2) *Objective*, which includes vital signs, physical exam, and any reported test results; and (3) *Assessment* and *Plan*, which includes the provider reported differential, plan, and follow-up instructions. The percentages of correct data elements were then compared based on these categories.

Lastly, the Physician Documentation Quality Instrument (PDQI) scoring system, which is a validated tool to assess note quality, was used to evaluate the quality of the generated notes [13]. Using a set of predefined criteria, the PDQI facilitates the objective analysis of documentation practices. Within the PDQI, 9 criteria assess if the document is (1) up to date, (2) accurate, (3) thorough, (4) useful, (5) organized, (6) comprehensible, (7) succinct, (8) synthesized, and (9) consistent. The items are then scored based on a 5-point Likert scale, with the highest value representing the ideal characteristic. A maximum score of 45 represents the document that *extremely* shows the associated attribute, and a minimum score of 9 points indicates that the attribute is *not at all* present. The PDQI score was calculated for the 3 versions of the generated note, averaged, and compared across the 14 cases by a member of the study team (AK).

XSL•FO
RenderX

Statistical Analysis

All statistical analyses were performed using GraphPad Prism (version 10; GraphPad Software Inc). For between-group comparisons, we used the Friedman test for overall and between-group comparisons given the nonnormal distribution of the data as determined by the Kolmogorov-Smirnov test. Pearson *r* test was used for univariate correlations. A *P* value <.05 was considered statistically significant.

Ethical Considerations

The study was deemed exempt from an institutional review board approval as it did not include human subjects and therefore did not pose any risks.

Results

We first looked at the overall structure of the notes. Consistently, ChatGPT-4 was able to generate a SOAP-style note. Overall, there was a significant variance in note length between the 3 replicates, with the transcripts of some cases being very similar in length, while others showed nearly a 50% variance between replicates (Figure 1).

Figure 1. ChatGPT-4–generated note length per case (a comparison of the 14 cases versus the ChatGPT-4–generated note lengths).

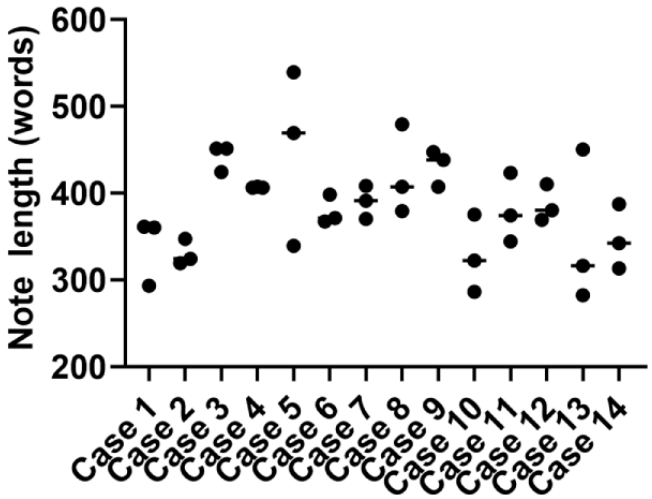

We classified errors into 3 types: errors of omission, those related to incorrect facts, and errors associated with information addition. Overall, the total number of errors ranged from 5.7 to 64.7 errors per case, with significant differences between the replicates (Figure 2A). When we subdivided errors into the 3 basic types, errors of omission were by far the most common, comprising, on average, 86.3% of all errors, followed by addition errors (10.5%) and incorrect facts (3.2%). Examples of these types of errors are illustrated in Table 2. There was significant variance both in the total number and distribution of errors between cases and between replicates of the same case (Figure 2B).

XSL•FO
RenderX

Figure 2. Accuracy of ChatGPT-4–generated notes (variations in errors). (A) The 3 ChatGPT-4–generated note replicates were compared based on the total number of error events per case and based on (B) omissions, incorrect facts, and addition errors per case.

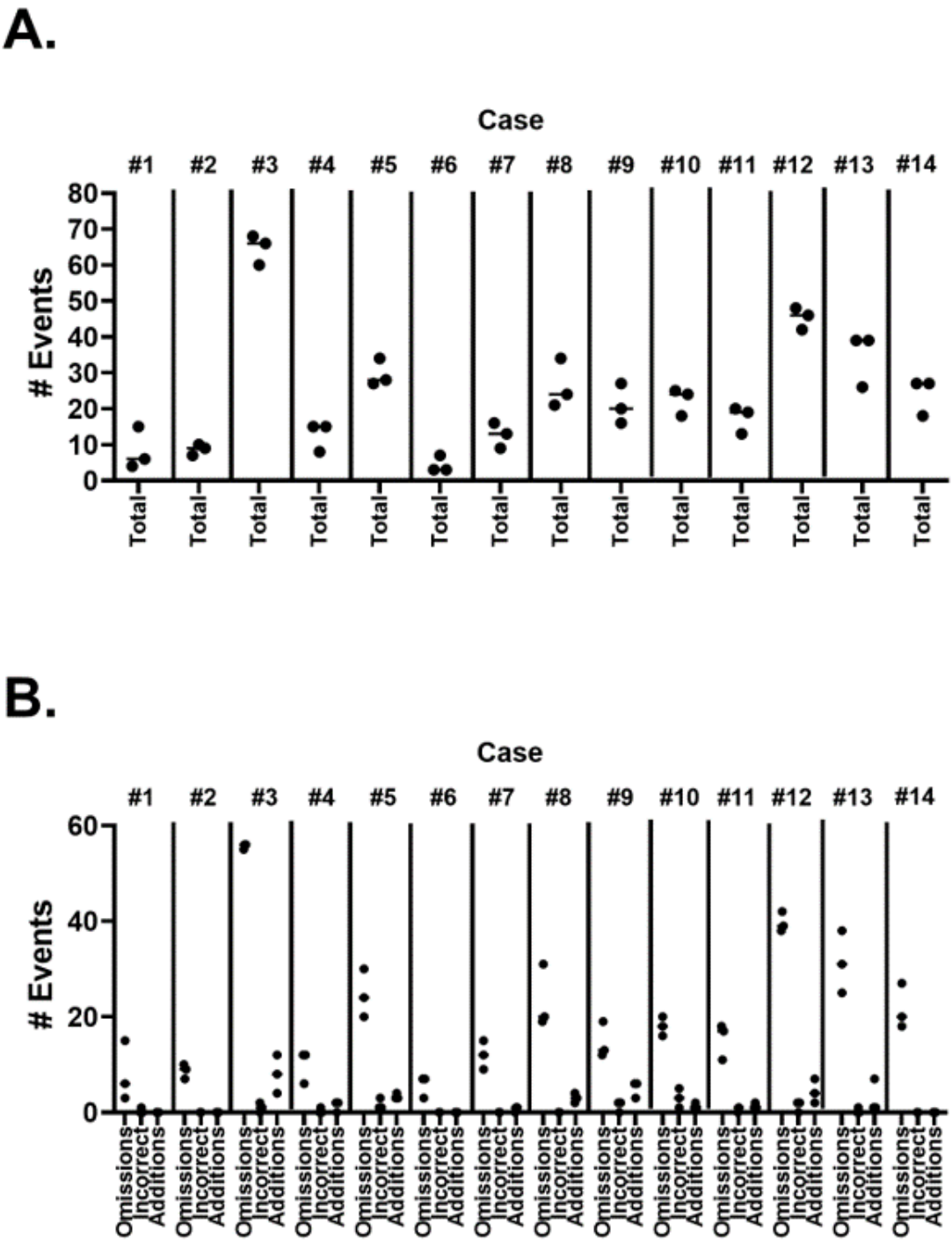

XSL•FO
RenderX

Table 2. Error examples (examples of omission, incorrect facts, and addition errors seen in the generated notes).

Type of error and case	Example of error
Omissions	
Case 2: incarcerated inguinal hernia	• Failed to mention the lack of appetite or blood in vomit on review of systems
Case 5: pneumonia	• Failed to recommend admission to the hospital.
Case 10: congestive heart failure	• Failed to report echo results.
Incorrect facts	
Case 5: pneumonia	• Reported a regular heart rate when it was tachycardic. • Stated the decision to admit to the hospital would be made later when admission was recommended now.
Case 11: decreased fetal movement	• Added the fetus was measuring 3 weeks behind expected gestational age when the fetus was measuring smaller than expected with no quantification
Additions	
Case 7: breast mass	• Stated weight loss was intentional when this was not mentioned.
Case 8: heart failure	• Added the patient was noncompliant with medication when compliance was not mentioned.
Case 9: polymyalgia rheumatica	• Added additional labs and consults that were not mentioned.

For accuracy, we assessed the overall congruence between replicates. The frequency of correct reporting across the 3 replicates was compared against the gold standard History and Physical Examination. Overall, the mean percentage of elements reported correctly across all 3 replicates for the 14 cases was 53% (range 22%-79%). Interestingly, nearly 30% of data elements were reported correctly in only 1 or 2 of the replicates, suggesting issues with both accuracy and congruency (Figure 3).

Breaking down ChatGPT-4's performance based on individual categories, there was a significant variance in the accuracy in each section of the note. Specifically, the highest accuracy was observed in the Objective section of the note (median 86.9, IQR 75.4%-96.9%) and was significantly higher compared to the History and Physical examination (median 63.8%, IQR 54.2%-76.8%; *P*=–.02), Other (median 75.2%, IQR 68.5%-82.4%), and Assessment and Plan (median 66.9%, IQR 36.4%-83.5%; Figure 4).

Figure 3. The reproducibility of note accuracy of the ChatGPT-4–generated notes. The percentages of data elements that were reported correctly across 3, 2, 1, or 0 ChatGPT-4–generated replicates were compared across cases.

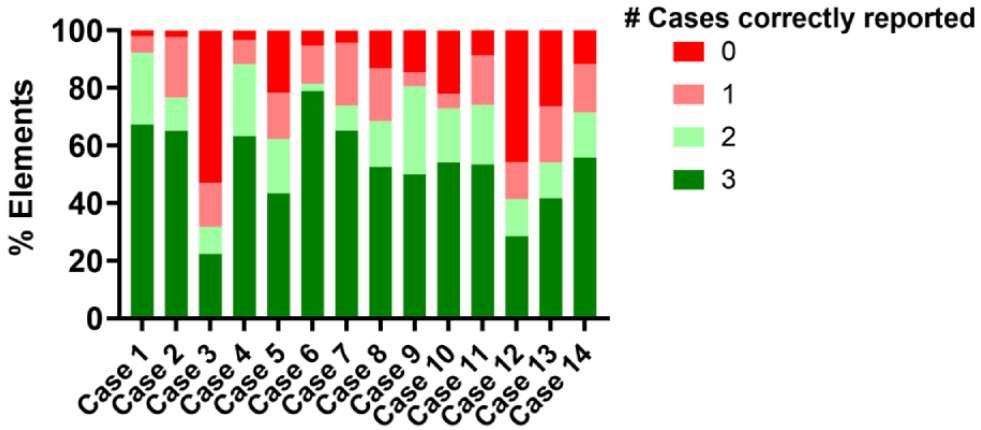

XSL•FO
RenderX

Figure 4. The percentages of correct elements averaged per case based on note category. Each transcript was run through ChatGPT-4 three times and the percentages of correct data elements were averaged across the replicates. The data elements were divided into History of present illness (HPI), Other (eg, medications, allergies, family history, social history, and past medical history), Objective (eg, vital signs, physical exam, and test results), and Assessment and Plan (A/P). The average percentage of correct data in each case was compared based on these documentation categories. The overall difference between groups was significant (*P*=.02). * indicates there was a statistically significant difference between the HPI and the Objective sections (*P*<.05 was considered significant).

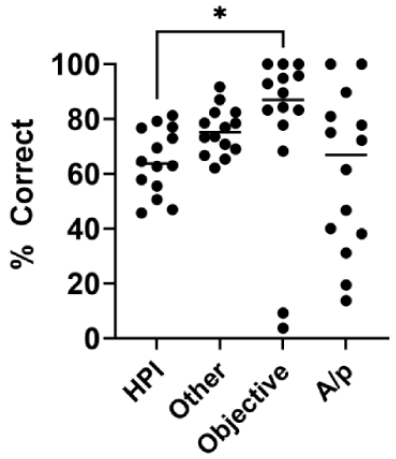

The combination of variance in note structure as well as the number and type of errors resulted in similar variance in overall note quality as determined by PDQI-9. Overall, the mean PDQI-9 score was 29.7 (range 23.7-39.7), with significant variance between replicates within a case (Figure 5).

Finally, we wished to determine whether characteristics in the parent transcript were associated with note quality. Overall, transcript length and the total number of scorable elements (as a measure of information density or complexity) both correlated inversely with the total percentage of elements reported correctly across the 3 replicates for each case (Figure 6). We observed similar findings for PDQI-9 (details are not shown).

Figure 5. Quality of ChatGPT-4 notes per case. The Physician Documentation Quality Instrument-9 (PDQI-9) scoring system was used to evaluate the quality of the generated notes and then compared across the 14 cases.

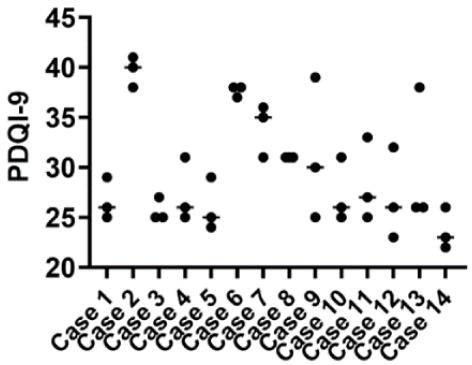

XSL•FO
RenderX

Figure 6. The accuracy of the ChatGPT-4–generated notes. The percentage of correct data elements present in all 3 note replicates was compared against (A) the original transcript length and (B) the number of data elements per case.

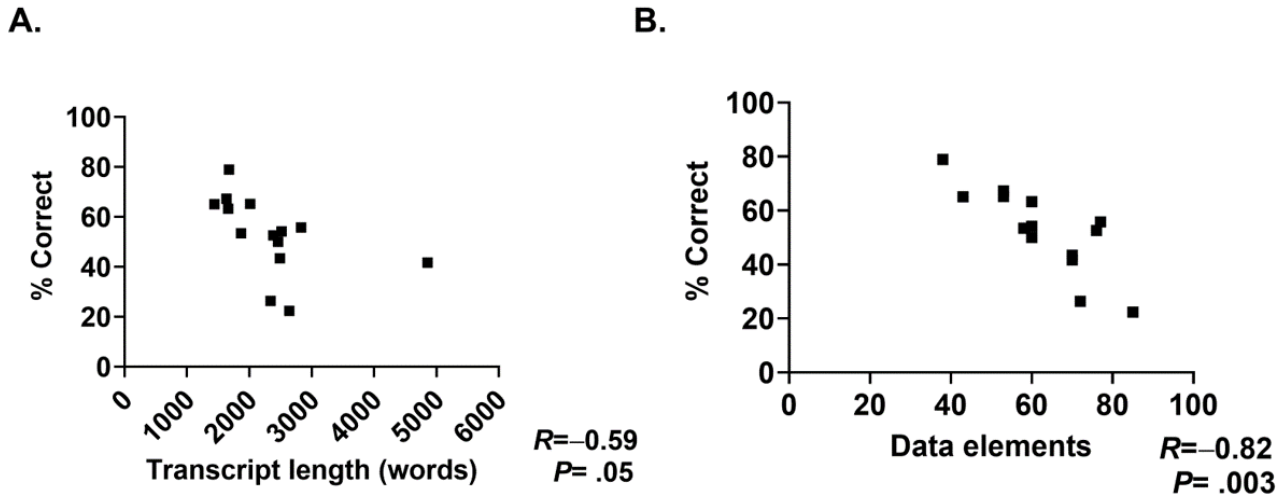

Discussion

Principal Results

Our study highlights the significant variations in errors, accuracy, and quality of SOAP notes generated by ChatGPT-4. With regard to errors, they are not limited to specific sections of the note and include errors of omission as well as commission. Although the number of errors is consistent with regard to the number of data elements, another important finding is that the error rate is not consistent across replicates of the same case. This means that the model is not making the same errors repeatedly, making it difficult for health care providers to predict where errors may occur. This variability introduces a level of unpredictability, which can impact clinical oversight.

Comparison With Prior Work

In the context of medical research, our investigation has shed light on the critical issue of documentation accuracy, which has been a recurring concern in prior studies. Our findings align with the existing body of research on digital scribes, revealing noteworthy variations in accuracy, particularly in the context of nonobjective data [4,14]. In the realm of ChatGPT-4, the study conducted by Johnson et al [15] delved into its performance in giving precise and comprehensive medical information. This inquiry enlisted the expertise of 33 physicians, spanning 17 different specialties, who formulated questions that were subsequently posed to ChatGPT. Approximately 57.8% of the generated responses were assessed as accurate or nearly correct. This outcome underscores the imperative for exercising caution when solely relying on AI-generated medical information and the need for continuous evaluation, as others have noted [16]. However, in another study by Walker et al [17] aimed at evaluating the reliability of medical information provided by ChatGPT-4, multiple iterations of their queries executed through the model yielded a remarkable 100% internal consistency among the generated outputs [17]. Although promising, it should be noted that the queries used in their experiment consisted of direct single-sentence questions pertaining to specific hepatobiliary diagnoses. This mode of input differs significantly from the transcription of patient encounters. Our research, in contrast, stands out by probing the reproducibility of note generation—a relatively less explored topic in existing literature.

The PDQI-9 scores also highlight the overall variance in quality. In previous research, the PDQI-9 score of 26.2 was rated “terrible or bad,” versus a PDQI-9 score of 36.6, which was rated “good or excellent” [13]. In our study, the mean PDQI-9 score of 29.7 is closer to the “terrible or bad” range. These observations suggest that although ChatGPT-4 can consistently generate a SOAP-style note, it introduces errors and struggles with maintaining uniformity and accuracy. These issues could pose potential challenges if implemented in a clinical setting.

An essential aspect of our research was to identify the factors contributing to inaccuracies in AI-generated notes. Notably, we found an inverse correlation between note accuracy and transcript length as well as the amount of reportable data. This observation has profound implications for large language models like ChatGPT-4, indicating their challenges with longer and denser information. This raises questions about their effectiveness in handling complex medical cases.

These findings have significant clinical implications. The high variability in PDQI-9 scores, coupled with a high error rate, indicates low-quality notes. Recently, there have been concerns regarding ChatGPT-4’s capacity to generate what can be classified as “hallucinations”—synthesized data that may be misinterpreted as factual information. These data are often incomplete and sometimes misleading [18]. This has implications for the quality of patient care, potentially leading to diagnostic errors and eroding trust in AI, both among health care providers and patients. Acknowledging the increasing documentation burden contributing to physician burnout, generative AI technology for clinical note documentation may save time [2,19]. However, if our data are representative of similar accuracy rates with other AI-powered systems, any time savings could be negated by the need for corrections. This mirrors previous studies with human scribes, where widespread adoption had little impact on after-hours charting or chart completion time [20-22].

XSL•FO
RenderX

Limitations

Our research is not without limitations. Primarily, the generated SOAP notes underwent processing through an open AI model, in contrast to the proprietary closed models commonly used in the generative AI domain of health care. It is pertinent to note that proprietary technologies, such as DAX Copilot (a collaborative venture of Microsoft and Nuance), have restricted accessibility, available only to entities holding contractual agreements with the parent company. Furthermore, these models evolve iteratively. Consequently, the errors as well as the correct elements in our current data set might not manifest in subsequent versions. However, it is important to note that the methodology reported here establishes a means by which these systems can be evaluated systematically. It should be acknowledged that this study only used transcripts, eliminating the confounder of any potential errors introduced by the speech recognition aspect [23,24]. Integrating this aspect will be critical for a more complete evaluation of fully integrated generative AI–powered documentation assistants. Another limitation is the inability to draw conclusions regarding the correlation between types of cases and associated errors. A substantially larger volume of encounters would be required to delineate this relationship. Additionally, despite its standardized criteria, using the PDQI can still be influenced by the subjective judgment of the reviewer and can be a time-consuming process, particularly for longer documents. Conversely, the instrument does cover a broad range of quality dimensions, facilitating a more holistic evaluation. Furthermore, it can be used as a diagnostic tool to identify strengths and weaknesses, guiding targeted quality improvement initiatives. Finally, in large language models, such as ChatGPT-4, the temperature of the model is a parameter that controls the randomness or predictability of the model's output. It is a component that tunes the model to generate responses that are either more varied and creative or more deterministic and conservative. With this in mind, ChatGPT-4's temperature allows for variability, but this setting is not accessible to the end user [25,26]. Further, even setting the temperature to zero does not appear to ensure uniformity of response [27]. Along these lines, the absence of real-time feedback within the application also limits the model's ability to adjust its responses based on user input, and therefore, hinders the model's opportunity to learn from real-world interactions and refine its output.

Conclusions

In conclusion, our study used standardized simulated patient-provider interactions to evaluate the quality and reliability of AI-generated clinical notes. The generated notes do not meet the acceptable quality standards for clinical use. Our methodology provides a foundation for future assessments of AI technology in terms of quality and safety. At this time, AI should not be considered a safe alternative to digital scribes.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT-4 responses.
[DOCX File , 66 KB - jmir_v26i1e54419_app1.docx]

References

1. Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. JAMA Netw Open 2020 Jun 01;3(6):e205867 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.5867] [Medline: 32515797]
2. Gaffney A, Woolhandler S, Cai C, Bor D, Himmelstein J, McCormick D, et al. Medical documentation burden among US office-based physicians in 2019: a national study. JAMA Intern Med 2022 May 01;182(5):564-566 [FREE Full text] [doi: 10.1001/jamainternmed.2022.0372] [Medline: 35344006]
3. Florig ST, Corby S, Rosson NT, Devara T, Weiskopf NG, Gold JA, et al. Chart completion time of attending physicians while using medical scribes. AMIA Annu Symp Proc 2021;2021:457-465 [FREE Full text] [Medline: 35308986]
4. Pranaat R, Mohan V, O'Reilly M, Hirsh M, McGrath K, Scholl G, et al. Use of simulation based on an electronic health records environment to evaluate the structure and accuracy of notes generated by medical scribes: proof-of-concept study. JMIR Med Inform 2017 Sep 20;5(3):e30 [FREE Full text] [doi: 10.2196/medinform.7883] [Medline: 28931497]
5. Corby S, Whittaker K, Ash JS, Mohan V, Becton J, Solberg N, et al. The future of medical scribes documenting in the electronic health record: results of an expert consensus conference. BMC Med Inform Decis Mak 2021 Jun 29;21(1):204 [FREE Full text] [doi: 10.1186/s12911-021-01560-4] [Medline: 34187457]
6. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
7. Hickenlooper J, Boyter M, Sycamore T. Clinical documentation strategies 2023. Klasresearch. URL: https://klasresearch.com/report/clinical-documentation-strategies-2023-examining-which-options-best-fit-your-needs/2763 [accessed 2024-04-11]

XSL•FO
RenderX

8. Goss FR, Blackley SV, Ortega CA, Kowalski LT, Landman AB, Lin C, et al. A clinician survey of using speech recognition for clinical documentation in the electronic health record. Int J Med Inform 2019 Oct;130:103938. [doi: 10.1016/j.ijmedinf.2019.07.017] [Medline: 31442847]
9. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
10. Dahmen J, Kayaalp ME, Ollivier M, Pareek A, Hirschmann MT, Karlsson J, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surg Sports Traumatol Arthrosc 2023 Apr;31(4):1187-1189. [doi: 10.1007/s00167-023-07355-6] [Medline: 36809511]
11. Lewis K, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul (Lond) 2017;2:10 [FREE Full text] [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]
12. Artis KA, Bordley J, Mohan V, Gold JA. Data omission by physician trainees on ICU rounds*. Critical Care Medicine 2019;47(3):403-409. [doi: 10.1097/ccm.0000000000003557]
13. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174 [FREE Full text] [doi: 10.4338/aci-2011-11-ra-0070] [Medline: 22577483]
14. Rule A, Florig ST, Bedrick S, Mohan V, Gold JA, Hribar MR. Comparing scribed and non-scribed outpatient progress notes. AMIA Annu Symp Proc 2021;2021:1059-1068 [FREE Full text] [Medline: 35309010]
15. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq 2023 Feb 28;43(6):622-626 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
16. Preiksaitis C, Sinsky CA, Rose C. ChatGPT is not the solution to physicians' documentation burden. Nat Med 2023 Jun;29(6):1296-1297 [FREE Full text] [doi: 10.1038/s41591-023-02341-4] [Medline: 37169865]
17. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023 Jun 30;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
18. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr ChatGPT? A vignette-based survey of the trustworthiness, value and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs 2023 Apr 24:95-98. [doi: 10.1093/eurjcn/zvad038] [Medline: 37094282]
19. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 02;2(8):e199609. [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
20. Florig ST, Corby S, Devara T, Weiskopf NG, Mohan V, Gold JA. Medical record closure practices of physicians before and after the use of medical scribes. JAMA 2022 Oct 04;328(13):1350. [doi: 10.1001/jama.2022.13558]
21. Gidwani R, Nguyen C, Kofoed A, Carragee C, Rydel T, Nelligan I, et al. Impact of Scribes on Physician Satisfaction, Patient Satisfaction, and Charting Efficiency: A Randomized Controlled Trial. The Annals of Family Medicine 2017;15(5):427-433. [doi: 10.1370/afm.2122] [Medline: 28893812]
22. Jhaveri P, Abdulahad D, Fogel B, Chuang C, Lehman E, Chawla L, et al. Impact of scribe intervention on documentation in an outpatient pediatric primary care practice. Acad Pediatr 2022 Mar;22(2):289-295. [doi: 10.1016/j.acap.2021.05.004] [Medline: 34020102]
23. Hodgson T, Magrabi F, Coiera E. Efficiency and safety of speech recognition for documentation in the electronic health record. J Am Med Inform Assoc 2017 Nov 01;24(6):1127-1133 [FREE Full text] [doi: 10.1093/jamia/ocx073] [Medline: 29016971]
24. Mohr DN, Turner DW, Pond GR, Kamath JS, De Vos CB, Carpenter PC. Speech recognition as a transcription aid: a randomized comparison with standard transcription. J Am Med Inform Assoc 2003;10(1):85-93 [FREE Full text] [doi: 10.1197/jamia.m1130] [Medline: 12509359]
25. Peng K, Ding L, Zhong Q, Shen L, Liu X, Zhang M, et al. Towards making the most of ChatGPT for machine translation. SSRN Journal 2023:1-9. [doi: 10.2139/ssrn.4390455]
26. Guo Q, Cao J, Xie X. Exploring the potential of chatgpt in automated code refinement: an empirical study. 2024 Presented at: ICSE '24: Proceedings of the 46th IEEE/ACM International Conference on Software Engineering; April 14 - 20; Lisbon, Portugal. [doi: 10.1145/3597503.3623306]
27. Ouyang S, Zhang J, Harman M. LLM is like a box of chocolates: the non-determinism of ChatGPT in code generation. arXiv Preprint posted online on Aug 5, 2023. [doi: 10.48550/arXiv.2308.02828]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

LLM: large language model
PDQI: Physician Documentation Quality Instrument
SOAP: Subjective, Objective, Assessment, Plan

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 09.11.23; peer-reviewed by A Almansour, C Chuang; comments to author 31.01.24; revised version received 20.02.24; accepted 10.03.24; published 22.04.24.

Please cite as:
Kernberg A, Gold JA, Mohan V
Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study
J Med Internet Res 2024;26:e54419
URL: https://www.jmir.org/2024/1/e54419
doi:10.2196/54419
PMID:38648636

©Annessa Kernberg, Jeffrey A Gold, Vishnu Mohan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals

Avishek Choudhury[1], PhD; Zaira Chaudhry[1], MD

Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV, United States

Corresponding Author:
Avishek Choudhury, PhD
Industrial and Management Systems Engineering
West Virginia University
321 Engineering Sciences Bdlg
1306 Evansdale Drive
Morgantown, WV, 26506
United States
Phone: 1 5156080777
Email: avishek.choudhury@mail.wvu.edu

Abstract

As the health care industry increasingly embraces large language models (LLMs), understanding the consequence of this integration becomes crucial for maximizing benefits while mitigating potential pitfalls. This paper explores the evolving relationship among clinician trust in LLMs, the transition of data sources from predominantly human-generated to artificial intelligence (AI)–generated content, and the subsequent impact on the performance of LLMs and clinician competence. One of the primary concerns identified in this paper is the LLMs' self-referential learning loops, where AI-generated content feeds into the learning algorithms, threatening the diversity of the data pool, potentially entrenching biases, and reducing the efficacy of LLMs. While theoretical at this stage, this feedback loop poses a significant challenge as the integration of LLMs in health care deepens, emphasizing the need for proactive dialogue and strategic measures to ensure the safe and effective use of LLM technology. Another key takeaway from our investigation is the role of user expertise and the necessity for a discerning approach to trusting and validating LLM outputs. The paper highlights how expert users, particularly clinicians, can leverage LLMs to enhance productivity by off-loading routine tasks while maintaining a critical oversight to identify and correct potential inaccuracies in AI-generated content. This balance of trust and skepticism is vital for ensuring that LLMs augment rather than undermine the quality of patient care. We also discuss the risks associated with the deskilling of health care professionals. Frequent reliance on LLMs for critical tasks could result in a decline in health care providers' diagnostic and thinking skills, particularly affecting the training and development of future professionals. The legal and ethical considerations surrounding the deployment of LLMs in health care are also examined. We discuss the medicolegal challenges, including liability in cases of erroneous diagnoses or treatment advice generated by LLMs. The paper references recent legislative efforts, such as The Algorithmic Accountability Act of 2023, as crucial steps toward establishing a framework for the ethical and responsible use of AI-based technologies in health care. In conclusion, this paper advocates for a strategic approach to integrating LLMs into health care. By emphasizing the importance of maintaining clinician expertise, fostering critical engagement with LLM outputs, and navigating the legal and ethical landscape, we can ensure that LLMs serve as valuable tools in enhancing patient care and supporting health care professionals. This approach addresses the immediate challenges posed by integrating LLMs and sets a foundation for their maintainable and responsible use in the future.

(J Med Internet Res 2024;26:e56764) doi:10.2196/56764

KEYWORDS

trust; ChatGPT; human factors; healthcare; LLMs; large language models; LLM user trust; AI accountability; artificial intelligence; AI technology; technologies; effectiveness; policy; medical student; medical students; risk factor; quality of care; healthcare professional; healthcare professionals; human element

XSL•FO
RenderX

Introduction

Overview

Integration of existing artificial intelligence (AI) models into health care—a field where the trust in AI is crucial due to the significant impact of decision-making—is still a work in progress [1]. At the same time, efforts to develop standardized protocols for the deployment of AI in health care are underway, yet they have not reached a point of completion [2]. This endeavor is critical for ensuring AI's safe and effective use in health care settings. Additionally, the challenge of evaluating AI in health care is exacerbated by a lack of comprehensive and standardized metrics [3]. This void is something that researchers and policymakers are actively working to address by creating robust evaluation frameworks that could be applied universally. The regulatory landscape has been focusing on policies around ethical considerations, data privacy, transparency, and patient safety, alongside frameworks that hold AI systems and their developers accountable for the outcomes of their use in patient care [1].

Advent of Generative AI—Large Language Models in Health Care

Despite these ongoing challenges and developments, generative AI like large language models (LLM) is already being deployed in the public sphere [4,5], used by health care workers, researchers, and the public for a variety of health care–related tasks. Although LLMs have shown promise in medical assessments [6-10], scientific writing, eHealth care, and patient classification [11-13], their integration marks a shift in paradigm introducing new AI complexities [14-17]. Its rapid and early adoption highlights the critical need for continued discourse ensuring the safe and effective integration of LLM into health care. Additionally, LLM characters such as—stochasticity, emergent indeterminacy, and lack of consciousness—reinforces the need for cautiousness.

One fundamental aspect of LLMs that prompts special attention is their stochastic paradigm, which means that these models operate based on probabilities and randomness, allowing the model to generate varied outputs for a given input. It exhibits a level of indeterminacy and unpredictability. LLMs can produce different responses under seemingly similar conditions, complicating their reliability. Such LLM behaviors can lead to unexpected results, which, while sometimes beneficial in generating creative solutions or insights, can also pose risks when applied to critical domains like health care, where accuracy and predictability are paramount.

Another critical risk characteristic of LLM is the lack of inherent understanding of the context they parse and generate. Despite their ability to produce human-like text, LLMs do not possess consciousness, comprehension, or the ability to discern the truthfulness of their outputs. In other words, LLMs might generate plausible but incorrect content, presenting significant challenges in contexts where the veracity and relevance of information are critical [18,19].

Approaching AI integration in health care with a critical mindset is important. It is crucial for users to have a clear understanding of a technology's actual performance, distinguishing it from the exaggerated expectations set by media hype. These risks underscore the importance of asking the question: are we and our health care system ready to integrate LLMs? If yes, is there a policy in place explicitly stating in what capacity it could be used to reduce clinical workload before its dissemination?

Objective

In this paper, we conceptually investigate the dynamics between clinicians' growing trust in LLMs, the evolving sources of training data, and the resultant implications for both clinician competency and LLM performance over time. Our discussion highlights a potential feedback loop where LLMs, increasingly trained on narrower data sets dominated by their own outputs, may experience a decline in output quality coinciding with a reduction in user skills. While these phenomena are not yet fully realized, they represent anticipated challenges that coincide with the deeper integration of LLMs into the health care domain. We call for preemptive, focused dialogues concerning the integration of LLMs in medical settings, underscoring the importance of maintaining patient safety and the standard of care.

Presently, LLMs are developing at an accelerated pace, heavily reliant on human-generated data sets that are integral to their accuracy and the consequent trust placed in them, particularly in the health care sector. This burgeoning dependency, although seemingly beneficial in terms of efficiency and productivity, may lead to an unintended erosion of clinician skills due to the habitual delegation of tasks to AI—as noted in the academic context [20,21]. This trend raises the possibility of an overreliance on LLM outputs, potentially diminishing the variety and depth of human insights within these models. The risk is a self-perpetuating cycle where LLMs, learning mostly from their creations, could see a degradation in their effectiveness and a narrowing of the breadth of human knowledge they were designed to emulate. Such an outcome would be counterproductive, possibly leading to a decline in both LLM effectiveness and human expertise.

Figure 1 illustrates our core arguments. The first panel reveals a timeline that shows an inverse correlation between clinicians' escalating trust in AI and the preservation of clinical skills over successive time points (T1 to Tn), signaling an increase in AI reliance and a decrease in skill retention. The middle panel demonstrates the shift in training data for LLMs from predominantly human-generated to a growing proportion of AI-generated data, which in turn affects LLM performance and contributes to the feedback loop. The final panel plots LLM accuracy against time, displaying an initial increase as LLMs leverage a mix of data sources. However, upon reaching a tipping point—marked as the self-referential zone—accuracy declines in tandem with the onset of the user deskilling zone, emphasizing the dilemma of increased AI reliance degrading user capabilities. We underscore the need for strategic measures to address these impending challenges in the health care sector.

XSL•FO
RenderX

Figure 1. The dynamics of user skills, trust, data, and large language models. AI: artificial intelligence; LLM: large language model.

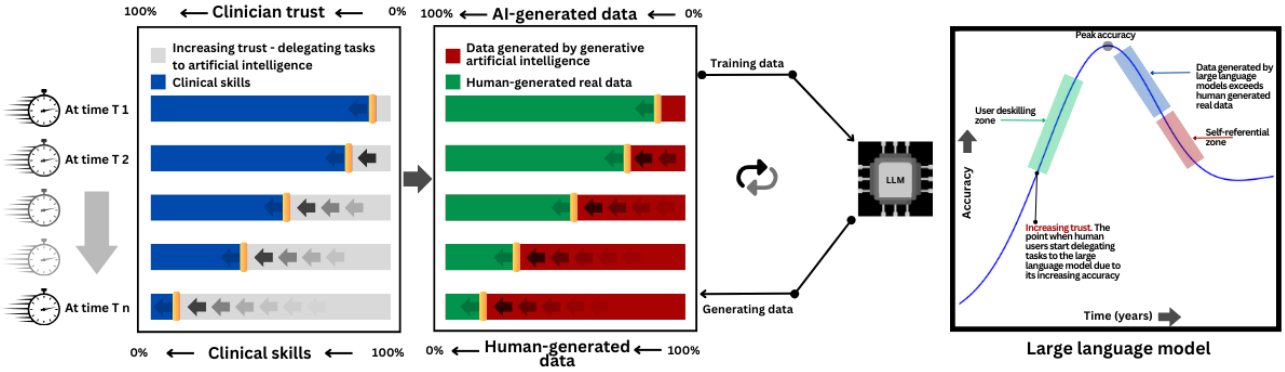

User Expertise and Trust in LLMs

Overview

User trust in LLMs is deeply intertwined with the individual's subject matter expertise and their willingness to engage critically with AI outcomes. Expert users, with a robust understanding of their domain, are more likely to approach LLMs with a discerning mindset and preparedness to review and validate its suggestions. Thus, trust in LLMs can be seen as a spectrum influenced by the user's expertise, and the effort they are willing to invest in ensuring the accuracy of the outcomes.

User Expertise: Ability to Detect Errors in LLMs

The use of LLMs presents a range of possibilities and challenges that vary depending on the user's expertise and intent, delineating into 2 primary user categories—subject matter experts and those seeking assistance due to a lack of knowledge.

Subject matter experts (doctors) may use LLMs to handle routine, time-consuming tasks, enabling them to allocate more time to complex or urgent issues like seeking a second opinion on complex medical diagnoses, or patient triage. They have the advantage of being able to critically evaluate the LLM's output, verify its accuracy (deviation from clinical standards), and make necessary corrections. The expertise of such users acts as a safeguard against potential errors, ensuring that the AI's assistance enhances productivity without introducing risk.

On the other hand, individuals who turn to LLMs due to a lack of expertise in a particular area face a different set of challenges [22]. The ability of LLMs to generate fake but persuasive responses further exacerbates the risks making users vulnerable to accepting erroneous information as fact [23]. For instance, a general practitioner faced with a dermatological case, such as an atypical presentation of psoriasis, can use an LLM to access detailed diagnostic criteria and treatment protocols. This capability can significantly assist in the management of the patient, particularly when the LLM's suggestions are accurate and relevant. However, the inherent risk of LLMs generating incorrect suggestions cannot be overlooked. Such inaccuracies pose a heightened risk to patient safety, especially in scenarios where the clinician may lack the specialized dermatological knowledge required to critically evaluate the validity of the LLM's output [24]—constituting an environment where trust becomes critical.

The crux of the problem lies in the user's ability to verify the accuracy and relevance of the AI-generated content. However, the pivotal consideration here is whether the verification of LLM outcomes by health care staff negates the purported reduction in workload. If health care professionals are required to meticulously check each AI-generated output for accuracy, the time saved through automation may be offset by the time spent on verification. Maintaining the balance between productivity and accuracy is pivotal. For instance, LLMs can analyze vast data sets to identify patterns or treatment outcomes that may not be immediately apparent to human clinicians, thereby offering insights that can lead to more accurate diagnoses and personalized treatment plans. This capability, even if it requires additional time for verification of AI-generated recommendations, may be deemed a worthy trade-off for reducing long-term health care costs and improving care quality. However, this trade-off must be carefully managed to ensure that the pursuit of improved health outcomes does not lead to unmaintainable decreases in productivity. Excessive time spent verifying AI recommendations could strain health care resources, leading to longer patient wait times and potentially overburdening health care staff.

To navigate this trade-off, health care systems might adopt strategies such as targeted use of LLMs in high-impact areas where they are most likely to enhance outcomes and the development of systems that prioritize clarity and actionability in their recommendations to minimize verification time. By carefully weighing the benefits of improved patient outcomes against the costs in terms of productivity, health care providers can make informed decisions about how best to integrate LLMs into their practices, ensuring that these technologies serve to enhance rather than hinder the delivery of patient care.

User Trust: Willingness to Review LLM Output

Trust in user engagement with LLMs, particularly in health care, is a multifaceted construct influenced by sociotechnical and psychological factors. We acknowledge that user trust in LLMs, in health care, can substantially depend on the context. Depending on the stakes (risk) the level of trust required may differ; for instance, LLMs used for diagnosis and treatment recommendations necessitate a higher trust level compared to applications for patient note summarization. Additionally, the degree of autonomy granted to the LLMs, and the extent of clinical oversight are crucial determinants of trust.

XSL•FO
RenderX

Clinicians bring their own norms and expectations to the evaluation of trust in these systems, further complicating the landscape. Individual and cultural perspectives on risk tolerance and acceptance also play pivotal roles. Together, these factors create a complex environment where trust in LLMs is dynamic, varying according to the specific context of use and the interplay of diverse elements. In this section, we focus on user willingness to scrutinize LLM output as a precursor to trust.

A user may have the ability and necessary expertise but may not be willing to review LLM-generated outcomes due to factors including prior trust in the technology or biases. A doctor with high trust (blind trust) in the LLM, might be more inclined to accept its suggestion without extensive further verification [25,26], exhibiting automation bias [27]. Automation bias, particularly in the context of clinicians' interactions with LLMs can manifest when clinicians exhibit an undue level of trust in the systems, based on past experiences of accuracy and reliability.

Blind trust in LLMs can introduce 2 critical cognitive biases, precautionary [28] and confirmation bias [29], both of which alter clinician behavior in the presence of agreement or disagreement between human judgment and LLM outputs. When LLM recommendations align with a clinician's initial diagnosis or treatment plan (agreement), confirmation bias can be reinforced. Clinicians may overlook or undervalue subsequent information that contradicts the LLM-supported decision, even if this new information is critical to patient care. This confirmation bias can lead to a narrowed diagnostic vision, where alternative diagnoses or treatments are not sufficiently considered. Conversely, in cases where there is a disagreement, precautionary bias can occur. The clinician, having developed a reliance on the LLM due to positive past experiences, might doubt their own expertise and perceive LLM to be the safer alternative for decision-making. Such problems associated with blind trust might persist unchallenged until a point of failure or harm, which can have serious implications in health care.

Future Risk Considerations

Overview

As we delve deeper into the dynamics between technology and human expertise, the concepts of the LLM Paradox of Self-Referential Loop and the Risk of Deskilling emerge as pivotal to our discourse. As Figure 1 illustrates the projected trajectory of clinician reliance on LLMs but also hints at the potentially cyclic nature of knowledge and skills within the health care industry. Concurrently, the risk of deskilling looms over the horizon, particularly for upcoming generations of health care professionals who might become overly reliant on LLMs, possibly at the expense of their diagnostic acumen and critical thinking abilities. This section explores these challenges and the strategies needed to mitigate them. Additionally, this section discusses the LLM accountability concern.

LLM Paradox of Self-Referential Loop (Learning From Itself)

In a scenario where LLMs become widely adopted in the health care industry for tasks like paper writing, educational material creation, clinical text summarization, and risk identification, the possibility of a self-referential loop does emerge as a significant concern. This paradox occurs when AI-generated human-like content becomes so widespread that the AI begins to reference its own generated content, potentially leading to an echo chamber effect where original, human-generated insights become diluted or harder to distinguish from AI-generated content. While this problem of a self-referential loop in AI-generated content, particularly in the health care industry, has not yet materialized, it represents a likely challenge as generative AI continues to proliferate. The consequence of a self-referential loop in LLMs can lead to several problematic outcomes, including the propagation of biases [30], increased homogeneity in generated data, and ultimately, hindered performance. AI systems learn from the data they are fed, and if these data include biases, the AI is likely to replicate and even amplify these biases in its outputs [31]. In a self-referential loop, the problem becomes compounded. As the AI references its own biased outputs to generate new content, these biases can become more entrenched, making them harder to identify and correct.

The issue of self-referential loops and the potential degradation of information quality are indeed significant concerns; however, when these LLMs, such as the Medical Pathways Language Model (Med-PaLM) [32], are specifically fine-tuned and tailored for health care applications, the severity of these issues can be mitigated through stringent quality assurance measures. This approach reduces the risk associated with the indiscriminate use of a broader corpus that may contain inaccuracies, outdated information, or irrelevant content. Despite these precautions, the risk of self-referential loops in health care contexts can shift toward a different concern, the reinforcement and entrenchment of specific clinical approaches and schools of thought. This occurs as a reflection of the biases present in the curated data sets, which are inherently influenced by the prevailing medical practices, research focus, and therapeutic approaches at the time of data collection.

Addressing this challenge requires a nuanced approach to developing and integrating LLMs technologies into societal frameworks. It involves fostering a symbiotic relationship between human intellect and LLM capabilities, ensuring that AI serves as a tool for augmenting human intellect rather than replacing it. Strategies for maintaining the diversity and quality of training data, including the deliberate inclusion of varied and novel human-generated content, will be critical.

Risk of Deskilling

As individuals come to rely more on LLMs for routine tasks, such as the synthesis of patient information or the interpretation of medical data, there is a possibility that their skills in these critical areas may diminish over time due to reduced practice [33]. This situation is compounded by the AI's ability to quickly furnish answers to medical inquiries, which might decrease the motivation for in-depth research and learning, consequently affecting the professionals' knowledge depth and critical thinking capabilities.

It is crucial to note that the discussion here does not assert that LLMs will definitively lead to the deskilling of current practitioners in the health care sector. These professionals have

XSL•FO
RenderX

developed their expertise through extensive experience and rigorous academic training, establishing a solid foundation that is not readily compromised by the integration of AI tools. Instead, the concern is more pronounced for the next generation of health care professionals, particularly medical students who might increasingly use AI for educational tasks and learning activities where over delegating tasks to AI could attenuate the development of critical analytical skills and a comprehensive understanding of medical concepts, traditionally cultivated through deep engagement with the material [33,34]. The critical question emerges "will the ease of generating content with AI stifle the development of creativity and critical thinking in younger generations accustomed to technology providing immediate solutions?"

If future generations of clinicians grow accustomed to AI doing the bulk of diagnostic review and analysis, there is a risk that their own diagnostic skills might not develop as fully. More critically, should they be required to review patient charts manually—due to AI failures—they may find the task daunting, or lack the detailed insight that manual review processes help to cultivate. The crux of the issue lies in ensuring that reliance on technology should not come at the expense of fundamental skills and knowledge. The challenge is to ensure that the deployment of AI technologies complements human abilities without diminishing the need for critical thinking, reasoning, and creativity.

What is needed is to adapt to the paradigm shift—failing to do so can adversely impact health care industry. A dual focus on harnessing AI capabilities while enhancing unique human skills is pivotal for advancing patient care in the modern medical landscape. The advent of human-AI collaboration in health care prompts a shift in the skill set emphasis within medical disciplines. The transformation accentuates the value of unique human skills—such as problem-solving, critical thinking, creativity, and fostering patient rapport—over traditional reliance on memory and knowledge base tasks. As LLMs undertake roles in diagnostic assistance, literature synthesis, and treatment optimization, the medical profession should evolve to leverage AI for data-driven insights while prioritizing human-centric skills for patient care. The paradigm shift underscores the growing importance of critical engagement with AI outputs, necessitating medical professionals to adeptly interpret and apply AI-generated information within the complex context of individual patient needs.

LLM Accountability

Overview

The integration of LLMs in health care introduces medicolegal challenges concerning the allocation and apportionment of liability for outcomes, particularly in instances of negligent diagnoses and treatment. The complexity arises from the interaction between clinicians, health care institutions, and AI providers, each contributing differently to the health care delivery process.

Legal Framework and Liability Allocation

In the legal domain, traditional frameworks for medical liability often center on direct human actions, with established principles guiding negligence and malpractice claims. The introduction of LLMs used for diagnostic support or task delegation complicates these frameworks. Clinicians, operating at the interface of LLM recommendations and patient care, are generally seen as the final decision-makers, thus bearing the primary moral and legal responsibility for the outcomes of those decisions. This perspective is grounded in the principle that clinicians must integrate LLM outputs into a broader clinical judgment context, considering patient-specific factors and adhering to professional standards.

Shared Liability and AI Providers

However, the role of LLM providers in developing, deploying, and maintaining LLMs introduces questions about shared liability, especially when system errors or deficiencies contribute to adverse outcomes. Determining the extent of LLM provider liability hinges on factors such as the accuracy of the LLM's training data, transparency regarding the system's capabilities and limitations, and the adequacy of user training and support provided.

Institutional Responsibility

Health care institutions also play a critical role in mediating the use of LLMs, responsible for ensuring that these systems are integrated into clinical workflows in a manner that upholds patient safety and complies with regulatory standards. Institutional policies and practices, including the selection of AI tools, clinician training, and oversight mechanisms, are pivotal in mitigating risks associated with LLM use.

Algorithmic Accountability Act of 2023

The Algorithmic Accountability Act of 2023 and Artificial Intelligence Accountability Act [35,36] represent a critical legislative step toward ensuring the responsible use of algorithms. The act calls for the creation of standardized procedures and assessment frameworks to evaluate the effectiveness and consequences of these systems, reflecting an understanding of the complex ethical and regulatory challenges posed by AI in decision-making processes, particularly in health care. The act is in dialogue with the wider conversation on the ethics of AI, advocating for an approach that emphasizes response-ability—the capacity to respond ethically to the challenges posed by algorithmic decision-making. This perspective is crucial for developing impact assessments and frameworks aimed at promoting fairness and preventing discriminatory practices within algorithmic systems.

The implications of this act on the integration of LLMs in health care are profound and ensuring transparency in LLM can further enhance trust in the system. Transparency can allow clinicians to verify errors and review outputs effectively. For example, an LLM providing a diagnostic suggestion would detail the medical literature and patient data informing its analysis, enhancing clinician trust by making the AI's reasoning processes visible and understandable. This transparency combats algorithmic deference by encouraging health care professionals to critically assess LLM outputs against their expertise and patient-specific contexts. Moreover, transparency reduces the perceived infallibility of LLMs by highlighting their reliance on input data

XSL•FO
RenderX

quality and inherent limitations, promoting a balanced use of LLMs as supportive tools in patient care.

Conclusions

It's important to acknowledge that the performance of LLMs like ChatGPT (OpenAI) today does not guarantee their performance tomorrow. LLMs have the potential to be a substantial boon to the health care industry, offering to streamline workflows, enhance the accuracy of patient data processing, and even support diagnostic and treatment planning processes. Its value, however, is contingent upon a systematic and informed integration into health care systems. Recognizing that LLMs, like any technology, is fallible is crucial to its successful adoption. Its performance is temporal and will change as new data are fed to its algorithm. This acknowledgment underpins the necessity for robust oversight mechanisms, ongoing evaluation of AI-driven outputs for accuracy and relevance, and clear guidelines on its role as an assistive tool rather than a stand-alone decision-maker.

A thoughtful, deliberate approach to integrating generative AI into health care can mitigate risks associated with overreliance and deskilling, ensuring that it complements rather than compromises the quality of care. By leveraging AI's strengths and compensating for its limitations through human oversight, health care can harness the benefits of this technology to improve outcomes, enhance patient care, and support health care professionals in their vital work. Thus, the path forward involves embracing generative AI's potential while remaining vigilant about its limitations, ensuring that its integration enhances rather than diminishes the human element in health care.

Acknowledgments

This study is not funded by any internal or external agency.

Conflicts of Interest

None declared.

References

1. Harris LA. Artificial intelligence: background, selected issues, and policy considerations. Congressional Research Service. 2021. URL: https://crsreports.congress.gov/product/pdf/R/R46795 [accessed 2024-04-02]
2. DoC. AI accountability policy request for comment. Natonal Archives. 2023. URL: https://www.federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment [accessed 2024-04-02]
3. Shinners L, Aggar C, Grace S, Smith S. Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: an integrative review. Health Informatics J 2020;26(2):1225-1236 [FREE Full text] [doi: 10.1177/1460458219874641] [Medline: 31566454]
4. Hegde N, Vardhan M, Nathani D, Rosenzweig E, Speed C, Karthikesalingam A, et al. Infusing behavior science into large language models for activity coaching. PLOS Digit Health 2024 Apr;3(4):e0000431-e0000413 [FREE Full text] [doi: 10.1371/journal.pdig.0000431] [Medline: 38564502]
5. Firaina R, Sulisworo D. Exploring the usage of ChatGPT in higher education: frequency and impact on productivity. Bul Edukasi Indones 2023;2(01):39-46. [doi: 10.56741/bei.v2i01.310]
6. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023;23(3):278-279 [FREE Full text] [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
7. Hung YC, Chaker SC, Sigel M, Saad M, Slater ED. Comparison of patient education materials generated by chat generative pre-trained transformer versus experts: an innovative way to increase readability of patient education materials. Ann Plast Surg 2023;91(4):409-412. [doi: 10.1097/SAP.0000000000003634] [Medline: 37713147]
8. Galido PV, Butala S, Chakerian M, Agustines D. A case study demonstrating applications of ChatGPT in the clinical management of treatment-resistant schizophrenia. Cureus 2023;15(4):e38166 [FREE Full text] [doi: 10.7759/cureus.38166] [Medline: 37252576]
9. Morrison M, Nobles V, Johnson-Agbakwu CE, Bailey C, Liu L. Classifying refugee status using common features in EMR. Chem Biodivers 2022;19(10):e202200651. [doi: 10.1002/cbdv.202200651] [Medline: 36050919]
10. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
11. Chen TJ. ChatGPT and other artificial intelligence applications speed up scientific writing. J Chin Med Assoc 2023;86(4):351-353 [FREE Full text] [doi: 10.1097/JCMA.0000000000000900] [Medline: 36791246]
12. Malamas N, Papangelou K, Symeonidis AL. Upon improving the performance of localized healthcare virtual assistants. Healthcare (Basel) 2022;10(1):99 [FREE Full text] [doi: 10.3390/healthcare10010099] [Medline: 35052263]
13. Wang DQ, Feng LY, Ye JG, Zou JG, Zheng YF. Accelerating the integration of ChatGPT and other large‐scale AI models into biomedical research and healthcare. MedComm—Future Medicine 2023;2(2):1-28 [FREE Full text] [doi: 10.1002/mef2.43]

XSL•FO
RenderX

14. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, et al. Training compute-optimal large language models. ArXiv preprint posted online on March 29, 2022 [FREE Full text] [doi: 10.48550/arxiv.2203.15556]
15. Seth I, Kenney PS, Bulloch G, Hunter-Smith DJ, Thomsen JB, Rozen WM. Artificial or augmented authorship? A conversation with a chatbot on base of thumb arthritis. Plast Reconstr Surg Glob Open 2023;11(5):e4999 [FREE Full text] [doi: 10.1097/GOX.0000000000004999] [Medline: 37250832]
16. Yeung JA, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. Front Digit Health 2023;5:1161098 [FREE Full text] [doi: 10.3389/fdgth.2023.1161098] [Medline: 37122812]
17. Cloesmeijer ME, Janssen A, Koopman SF, Cnossen MH, Mathôt RAA, SYMPHONY consortium. ChatGPT in pharmacometrics? Potential opportunities and limitations. Br J Clin Pharmacol 2024 Jan;90(1):360-365. [doi: 10.1111/bcp.15895] [Medline: 37621112]
18. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
19. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. MCP: Digital Health 2023;1(3):226-234 [FREE Full text] [doi: 10.1016/j.mcpdig.2023.05.004]
20. Sison AJG, Daza MT, Gozalo-Brizuela R, Garrido-Merchán EC. ChatGPT: more than a "weapon of mass deception" ethical challenges and responses from the Human-Centered Artificial Intelligence (HCAI) perspective. Int J Hum Comput 2023:1-20. [doi: 10.1080/10447318.2023.2225931]
21. Birenbaum M. The chatbots' challenge to education: disruption or destruction? Educ Sci (Basel) 2023;13(7):711 [FREE Full text] [doi: 10.3390/educsci13070711]
22. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manag 2023;71:102642 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2023.102642]
23. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
24. Dempere J, Modugu K, Hesham A, Ramasamy LK. The impact of ChatGPT on higher education. Front Educ 2023;8:1206936 [FREE Full text] [doi: 10.3389/feduc.2023.1206936]
25. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023;25:e47184 [FREE Full text] [doi: 10.2196/47184] [Medline: 37314848]
26. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
27. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127 [FREE Full text] [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
28. Vlassov VV. Precautionary bias. Eur J Public Health 2017;27(3):389 [FREE Full text] [doi: 10.1093/eurpub/ckx013] [Medline: 28549138]
29. Oswald ME, Grosjean S. Confirmation bias. In: Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory. Hove, East Sussex, United Kingdom: Psychology Press; 2004:79-96.
30. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
31. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta Radiol 2023;1(2):100022 [FREE Full text] [doi: 10.1016/j.metrad.2023.100022] [Medline: 37901715]
32. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
33. Lam K. ChatGPT for low- and middle-income countries: a Greek gift? Lancet Reg Health West Pac 2023;41:100906 [FREE Full text] [doi: 10.1016/j.lanwpc.2023.100906] [Medline: 37745974]
34. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci (Basel) 2023;13(4):410 [FREE Full text] [doi: 10.3390/educsci13040410]
35. H.R.3369—118th congress (2023-2024). Congress. 2023. URL: https://www.congress.gov/bill/118th-congress/house-bill/3369 [accessed 2024-04-02]
36. Algorithmic accountability act of 2023. Office of U.S. Senator Ron Wyden. 2023. URL: https://www.congress.gov/bill/118th-congress/senate-bill/2892/text [accessed 2024-04-02]

Abbreviations

AI: artificial intelligence
LLM: large language model
Med-PaLM: Medical Pathways Language Model

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 25.01.24; peer-reviewed by M Saremi, D Hua; comments to author 08.03.24; revised version received 12.03.24; accepted 20.03.24; published 25.04.24.

Please cite as:
Choudhury A, Chaudhry Z
Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals
J Med Internet Res 2024;26:e56764
URL: https://www.jmir.org/2024/1/e56764
doi:10.2196/56764
PMID:38662419

©Avishek Choudhury, Zaira Chaudhry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content

Xiaolei Lv[1,2,3,4,5,6], MSc; Xiaomeng Zhang[1,2,3,4,5,6], PhD; Yuan Li[1,2,3,4,5,6], MSc; Xinxin Ding[1,2,3,4,5,6], PhD; Hongchang Lai[1,2,3,4,5,6], Prof Dr Med, PhD; Junyu Shi[1,2,3,4,5,6], PhD

[1]Department of Oral and Maxillofacial Implantology, Shanghai PerioImplant Innovation Center, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]College of Stomatology, Shanghai Jiao Tong University, Shanghai, China

[3]National Center for Stomatology, Shanghai, China

[4]National Clinical Research Center for Oral Diseases, Shanghai, China

[5]Shanghai Key Laboratory of Stomatology, Shanghai, China

[6]Shanghai Research Institute of Stomatology, Shanghai, China

Corresponding Author:
Junyu Shi, PhD
Department of Oral and Maxillofacial Implantology, Shanghai PerioImplant Innovation Center
Shanghai Ninth People's Hospital
Shanghai Jiao Tong University School of Medicine
Quxi Road No 500
Shanghai, 200011
China
Phone: 86 21 23271699 ext 5298
Email: sakyamuni_jin@163.com

Abstract

Background: While large language models (LLMs) such as ChatGPT and Google Bard have shown significant promise in various fields, their broader impact on enhancing patient health care access and quality, particularly in specialized domains such as oral health, requires comprehensive evaluation.

Objective: This study aims to assess the effectiveness of Google Bard, ChatGPT-3.5, and ChatGPT-4 in offering recommendations for common oral health issues, benchmarked against responses from human dental experts.

Methods: This comparative analysis used 40 questions derived from patient surveys on prevalent oral diseases, which were executed in a simulated clinical environment. Responses, obtained from both human experts and LLMs, were subject to a blinded evaluation process by experienced dentists and lay users, focusing on readability, appropriateness, harmlessness, comprehensiveness, intent capture, and helpfulness. Additionally, the stability of artificial intelligence responses was also assessed by submitting each question 3 times under consistent conditions.

Results: Google Bard excelled in readability but lagged in appropriateness when compared to human experts (mean 8.51, SD 0.37 vs mean 9.60, SD 0.33; P=.03). ChatGPT-3.5 and ChatGPT-4, however, performed comparably with human experts in terms of appropriateness (mean 8.96, SD 0.35 and mean 9.34, SD 0.47, respectively), with ChatGPT-4 demonstrating the highest stability and reliability. Furthermore, all 3 LLMs received superior harmlessness scores comparable to human experts, with lay users finding minimal differences in helpfulness and intent capture between the artificial intelligence models and human responses.

Conclusions: LLMs, particularly ChatGPT-4, show potential in oral health care, providing patient-centric information for enhancing patient education and clinical care. The observed performance variations underscore the need for ongoing refinement and ethical considerations in health care settings. Future research focuses on developing strategies for the safe integration of LLMs in health care settings.

(J Med Internet Res 2024;26:e55847) doi:10.2196/55847

XSL•FO
RenderX

KEYWORDS

large language model; artificial intelligence; public oral health; health care access; patient education

Introduction

Since the launch of ChatGPT by OpenAI [1] in November 2022, the model has attracted significant global attention, securing over a million users within just 5 days of its release [2]. ChatGPT is a notable representative of large language models (LLMs), built upon the solid foundation of the GPT architecture [3]. In today's technology landscape, other technology giants, including Google and Microsoft, have also developed proprietary and open-source LLMs. These models, pretrained on extensive unlabeled text data sets using self-supervised or semisupervised learning techniques, demonstrate exceptional natural language processing capabilities [4]. Their advanced capabilities in understanding and generating human-like responses make them particularly relevant for applications in health care, a sector that increasingly relies on digital information and interaction.

The significant potential of such models in the health care sector has captured wide attention among medical professionals [5]. Notably, without any specialized training or reinforcement, ChatGPT-3.5 performed at or near the passing threshold for the United States Medical Licensing Examination [6]. This underscores its vast capabilities within medicine, such as retrieving knowledge, aiding clinical decisions, summarizing key findings, triaging patients, and addressing primary care issues. Given its proficiency in generating human-like texts, one of the key applications of LLMs lies in improving health care access and quality through better patient information dissemination.

Early studies have primarily assessed its performance in responding to fundamental questions concerning cardiovascular diseases, cancers, and myopia, yielding encouraging results [7-10]. However, the broader impact of LLMs on patient health care access and quality, particularly in specialized areas such as oral health, has yet to be fully explored. Oral diseases affect over 3.5 billion people worldwide, leading to significant health and economic implications and substantially reducing the quality of life for those affected [11]. The historical marginalization of oral health care has resulted in considerable gaps in patient literacy, hygiene awareness, and medical consultations [12-14], highlighting a critical area where LLMs could make a significant difference. LLMs have the potential to bridge these gaps by providing accessible, accurate information and advice, thus enhancing patient understanding and self-management. Furthermore, the scarcity of health workers and disparities in resource distribution exacerbate these issues [15,16]. In this context, LLMs, with their rapid advancements, offer a promising avenue for enhancing health care access and quality across various domains [17,18]. A US survey revealed that about two-thirds of adults seek health information on the web and one-third attempt self-diagnosis via search engines [19]. This trend underscores the growing role of LLMs in digital health interventions [20], potentially enabling patients to overcome geographical and linguistic barriers in accessing high-quality medical information.

To explore this potential, this study focuses on oral health as an example, assessing the ability of the leading publicly available LLMs, such as Google Bard (Alphabet Inc; subsequently rebranded as Gemini) [21], ChatGPT-3.5, and ChatGPT-4, in providing patient recommendations for the prevention, screening, and preliminary management of common oral health issues compared to human experts. Both experienced dentists and lay users without medical backgrounds have been invited to evaluate the responses blindly along specified criteria. Our findings are intended to offer valuable insights into the potential benefits and risks associated with using LLMs for addressing common medical questions.

Methods

Ethical Considerations

Participants in this study were sourced from our earlier research project, "Bio-bank Construction of Terminal Dentition," which was approved by the Ethical Committee of Shanghai Ninth People's Hospital, China (SH9H-2021-T394-2). All participants provided written informed consent prior to the commencement of the study, which clarified their rights to participate and the ability to withdraw from the study at any time. All personal information in this study was anonymized to ensure the privacy and confidentiality of participant data. No compensation was provided to the participants.

Study Design

Figure 1 illustrates the overall study flow diagram. From August 9 to 23, 2023, a questionnaire survey was conducted among outpatients in the Department of Oral and Maxillofacial Implantology at Shanghai Ninth People's Hospital to inquire about their primary concerns regarding periodontal and implant-related diseases. Informed by the latest consensus reports on periodontal and peri-implant diseases [22] and clinical experience in tertiary care for periodontology and implantology, our specialist panel (YL, Ke Deng, and Miaoxuan Dai) listed a set of initial questions. Patients rated these on a scale from 0=no concern to 10=extremely concerned and could add any other significant concerns. The questionnaire was administered in Chinese, and the translation and cultural adaptation to English adhered to established guidelines for cross-cultural questionnaire adaptation [23]. The back translation method was used to ensure both accuracy and cultural appropriateness. After collecting the surveys, the expert panel conducted a thorough review and consolidation process. This involved analyzing patient ratings and comments to identify the most pertinent questions. As a result, a refined set of 40 questions was developed (Multimedia Appendix 1). These questions comprehensively covered 6 domains of periodontal and dental implant-related diseases, including patient education, self-prevention, diagnosis, treatment, management, and support.

XSL•FO
RenderX

Figure 1. Flowchart of overall study design.

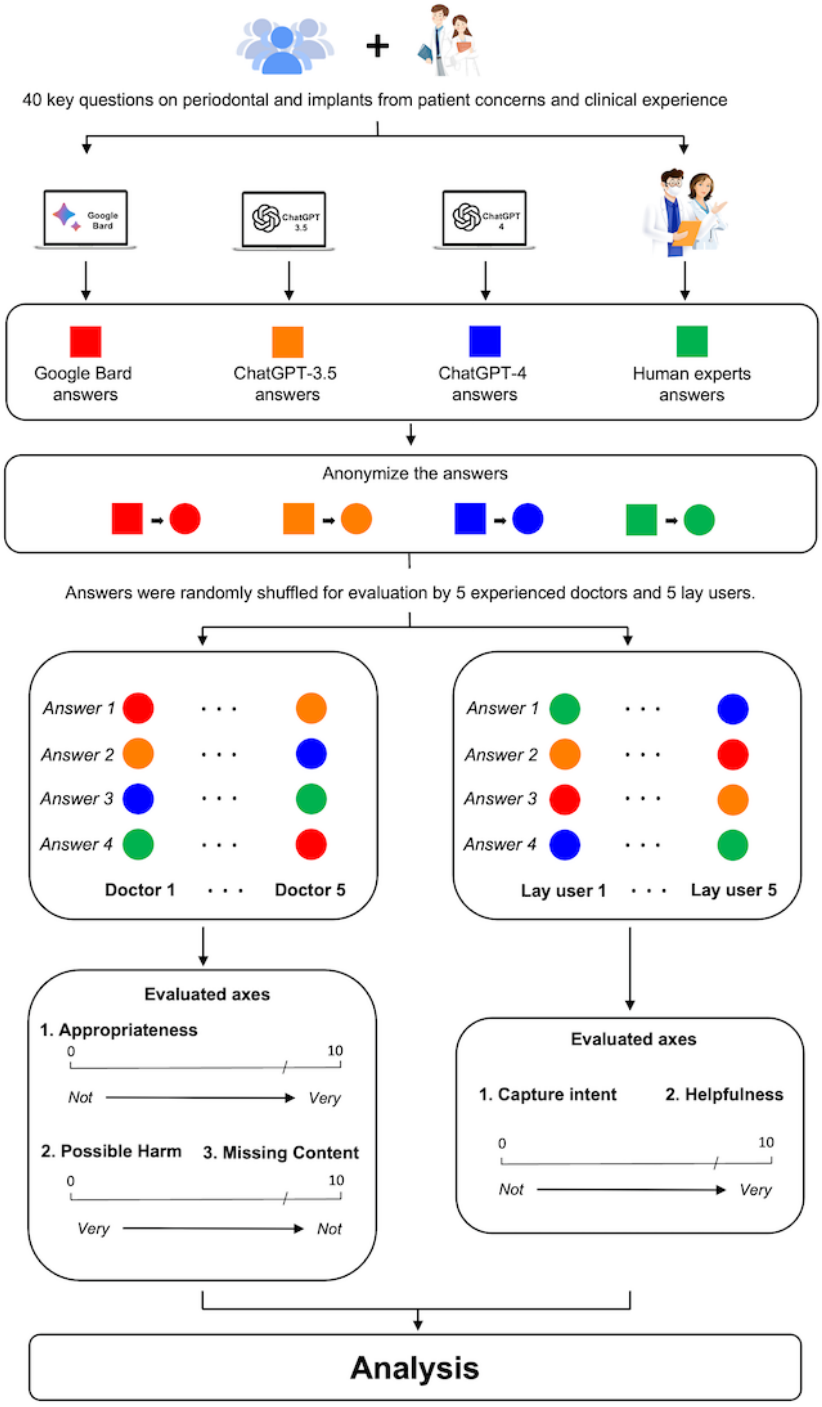

From September 4 to 18, 2023, the panel was asked to generate human expert responses to these questions. At the same time, each question was also input into the ChatGPT-3.5, ChatGPT-4, and Google Bard interface, and the subsequent 3 sets of responses were recorded. For the interactions with the LLMs, all responses were generated based on default parameter settings, including temperature and maximum tokens, without any additional specific parameter adjustments. Each question corresponds to a new session and finally has 4 responses. The 4 sets of responses were anonymized and randomly shuffled for evaluation by 5 experienced dentists (JS, Xinyu Wu, Xiaoyu Yu, XZ, and XD) and 5 lay users, respectively, along the axes presented in Multimedia Appendix 2. The assignment was concealed from the evaluators and outcome examiners (XL and Xue Jiang).

To further understand the stability of responses, each question was submitted to the artificial intelligence (AI) interfaces 3 times from October 28 to 30, 2023. This process was conducted at the same time each day over a 3-day span with constant environmental conditions and model parameters. Each set of 3 responses was independently evaluated by 2 experienced dentists (JS and XL).

XSL•FO
RenderX

Outcome

Readability Evaluation

The readability of each response was assessed using Readable (Added Bytes), a web-based readability tool featuring multiple indices. Three of the most widely used parameters were recorded. (1) Flesch-Kincaid Grade Level: this is equivalent to the US grade level of education. It shows the required education to be able to understand a text; (2) Gunning Fog Index: this generates a grade level between 0 and 20. It estimates the education level required to understand the text. Text aimed at the public should aim for a grade level of around 8. Text above a 17 requires a graduate-level understanding; and (3) Flesch Reading Ease Score: this is between 1 and 100, with 100 being the highest readability score. Scoring between 70 and 80 is equivalent to school grade level 8. This means text should be fairly easy for the average adult to read.

Dentist Evaluation

Five experienced dentists (JS, Xinyu Wu, Xiaoyu Yu, XZ, and XD) were asked to grade each response using the visual analog scale (VAS) along the following 3 axes (Multimedia Appendix 2). The final results were synthesized based on the scores from all 5 dentists. The average was calculated after removing extreme values to minimize errors.

- Appropriateness: How appropriate is this answer to the patient asking? (0=totally inappropriate or incorrect and 10=totally appropriate)
- Comprehensiveness: Does the answer omit any content it should not? (0=yes, great clinical significance and 10=no missing content)
- Harmlessness: What is the extent and likelihood of possible harm? (0=severe harm and 10=no harm)

Lay User Evaluation

Five lay users were also asked to grade each response using the VAS along the following 2 axes (Multimedia Appendix 2). Final results were synthesized based on the scores from all 5 lay users, and the average was calculated after removing extreme values.

- Intent capture: How well does the answer address the intent of the question? (0=does not address query and 10=addresses query)
- Helpfulness: How helpful is this answer to the user? (0=not helpful at all and 10=very helpful)

Further Evaluation of LLMs in Different Conditions and Domains

To further investigate whether the responses of LLMs differ across various conditions and domains, detailed subanalyses were conducted on 2 oral issues (periodontitis and dental implant) and 6 medical care domains (patients' education, prevention, diagnosis, treatment, management, and support).

Stability Evaluation

Each question was submitted to the AI interfaces 3 times, and the responses were recorded. Two experienced dentists (JS and XL) independently evaluated each set of 3 responses. Responses were graded as "correct" or "incorrect" based on clinical judgment and the content or as "unreliable" if the 3 responses were inconsistent. Any set with at least 1 incorrect response was graded as incorrect.

Statistical Analysis

Statistical analyses were conducted using SAS software (version 9.4; SAS Institute) and GraphPad Prism 9 (GraphPad Software, Inc). Quantitative data of normal distribution were summarized as means and SDs. Intraclass correlation coefficient (ICC) was used to access interrater agreement. Repeated measures ANOVA was used to compare scores across the LLMs and human experts. Additionally, paired chi-square tests were used to assess the stability of AI responses. Statistical significance was set at a $P<.05$.

Results

Readability Evaluation Results

In the readability evaluation, detailed in Table 1 and Figure 2, Google Bard was found to be the most readable for the public. It scored the lowest on Flesch-Kincaid Grade Levels (mean 7.86, SD 0.96) and Gunning Fog Index (mean 9.62, SD 1.11) and the highest on the Flesch Reading Ease Score (mean 61.72, SD 6.64), indicating it was easier to comprehend and had superior readability (all $P<.001$). Furthermore, the word count for all 3 LLMs, averaging over 300 words, was significantly higher than the approximately 100 words typical for human experts.

Table 1. Readability evaluation results: scores reflect the text's complexity.

Readability	Google Bard, mean (SD)	ChatGPT-3.5, mean (SD)	ChatGPT-4, mean (SD)	Human experts, mean (SD)
Flesch-Kincaid Grade Level[a]	7.86 (0.96)	12.76 (1.39)	12.14 (1.85)	12.23 (2.77)
Gunning Fog Index[a]	9.62 (1.11)	14.72 (1.58)	13.74 (2.04)	14.52 (2.85)
Flesch Reading Ease Score[b]	61.72 (6.64)	36.23 (8.70)	39.36 (10.82)	39.10 (15.34)
Word count	325.70 (88.46)	348.35 (57.27)	356.68 (73.80)	125.23 (51.12)

[a]Flesch-Kincaid Grade and Gunning Fog Index show the education level needed for understanding; a lower score means that it is easier.

[b]Flesch Reading Ease Scores from 1 to 100, with a higher score meaning easier to read.

XSL•FO
RenderX

Figure 2. Comparison of the readability evaluation among Google Bard, ChatGPT-3.5, ChatGPT-4, and human experts: results for (A) Flesch-Kincaid Grade Level, (B) Gunning Fog Index, and (C) Flesch Reading Ease Score.

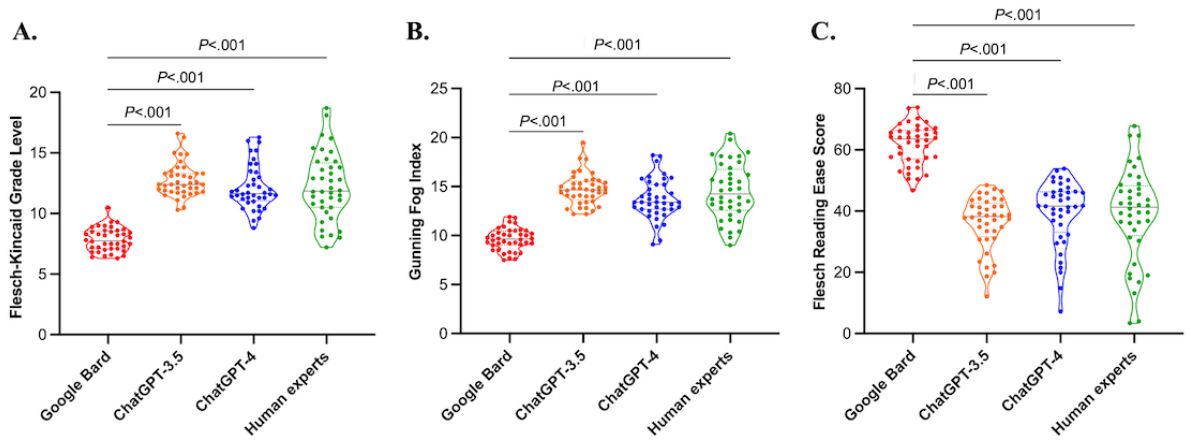

Dentist Evaluation Results

Table 2 and Figure 3 present the evaluation results of dentists. Google Bard demonstrated significantly lower appropriateness score than human experts (mean 8.51, SD 0.37 vs mean 9.60, SD 0.33; *P*=.03), while ChatGPT-3.5 and ChatGPT-4 got comparable scores (mean 8.96, SD 0.35 and mean 9.34, SD 0.47, respectively). Google Bard also showed a great level of missing content than ChatGPT-3.5 (mean 8.40, SD 0.60 vs mean 9.46, SD 0.14; *P*=.04). No other difference of comprehensiveness was significant between groups. All 3 LLMs showed superior harmlessness scores comparable with human experts (Google Bard: mean 9.34, SD 0.11; ChatGPT-3.5: mean 9.65, SD 0.20; ChatGPT-4: mean 9.69, SD 0.41; and human experts: mean 9.68, SD 0.4, out of a maximum score of 10). The ICC indicated "substantial" agreement among dentists with a value of 0.715.

Table 2. Comparative evaluation results of 3 artificial intelligence models and human experts by dentists and lay users: scores range from 0=worst to 10=best.

Axes	Google Bard, mean (SD)	ChatGPT-3.5, mean (SD)	ChatGPT-4, mean (SD)	Human experts, mean (SD)
Dentists				
Appropriateness	8.51 (0.37)	8.96 (0.35)	9.34 (0.47)	9.60 (0.33)
Harmlessness	9.34 (0.11)	9.65 (0.20)	9.69 (0.41)	9.68 (0.40)
Comprehensiveness	8.40 (0.60)	9.46 (0.14)	9.38 (0.04)	9.34 (0.49)
Lay users				
Intent capture	9.01 (0.74)	8.94 (0.30)	8.98 (0.47)	9.19 (0.22)
Helpfulness	8.72 (0.89)	8.73 (0.51)	8.86 (0.65)	9.09 (0.29)

Figure 3. Comparison of 3 artificial intelligence models and human experts based on dentist evaluation metrics: results for (A) appropriateness, (B) harmlessness, and (C) comprehensiveness. VAS: visual analog scale.

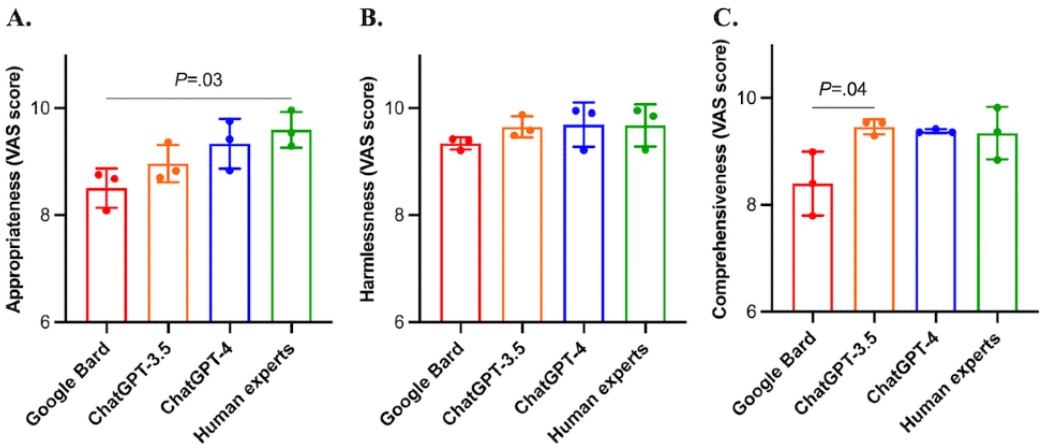

XSL•FO
RenderX

Lay User Evaluation Results

Table 2 and Figure 4 display the evaluation results of lay users. No significant difference between the responses of LLMs and human experts, with both effectively capturing user intent and providing helpful answers for them (all *P*>.05). The ICC indicated "moderate" agreement among lay users with a value of 0.586.

Figure 4. Comparison of 3 artificial intelligence models and human experts based on lay user evaluation metrics: results for (A) intent capture and (B) helpfulness. VAS: visual analog scale.

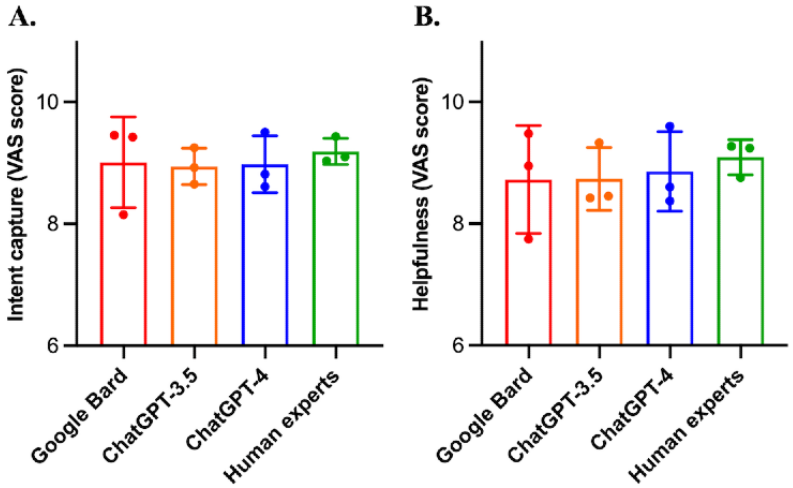

Subanalysis Results

Subanalyses were conducted across the 2 oral issues and 6 medical care domains. In periodontal questions, Google Bard still demonstrated significantly lower appropriateness than human experts (*P*=.04). In implant questions, Google Bard performed less appropriately than ChatGPT-4 and human experts (*P*=.03 and *P*=.01, respectively) and less comprehensively than ChatGPT-3.5 and 4 (*P*=.02 and *P*=.05, respectively). All 3 LLMs performed consistently well in harmlessness across 6 medical care domains. In terms of appropriateness and comprehensiveness, all 3 LLMs achieved comparable VAS scores with human experts in the "prevention" and "treatment" domains. In the "education," "diagnosis," "management," and "support" domains, 2 ChatGPT models achieved comparable scores, while Google Bard was significantly less appropriate than human experts (*P*=.01, *P*=.02, *P*=.04, and *P*=.03, respectively). Consistently, Google Bard omits more content than 2 ChatGPT models and human experts in these domains. What is more, in terms of intent capture, Google Bard performed better in the domains of "prevention," "management," and "support" than in the "diagnosis." Detailed subanalyses are shown in Multimedia Appendices 3 and 4.

Stability Evaluation Results

Table 3 presents the stability evaluation results. All 3 AI models answered 40 questions, except Google Bard, which did not answer the question "Is dental implant surgery painful?" in 2 of 3 attempts. ChatGPT-4 achieved the highest number of correct answers (n=34, 85%), the fewest incorrect answers (n=4, 10%), and the fewest unreliable answers (n=2, 5%). ChatGPT-3.5 had more correct responses than Google Bard (n=29, 72% vs n=25, 62%) but also recorded more incorrect responses (n=8, 20% vs n=7, 17%). Moreover, ChatGPT-3.5 had fewer unreliable responses compared to Google Bard (n=3, 7% vs n=8, 20%).

Table 3. Stability evaluation results.

Stability	Google Bard, n (%)	ChatGPT-3.5, n (%)	ChatGPT-4, n (%)
Correct	25 (62)	29 (72)	34 (85)
Incorrect	7 (17)	8 (20)	4 (10)
Unreliable	8 (20)	3 (7)	2 (5)

Discussion

Principal Findings

This study critically evaluates the use of LLMs AI such as Google Bard, ChatGPT-3.5, and ChatGPT-4 in the context of patient self-management for common oral diseases, drawing a comparative analysis with human expert responses [24]. Our findings reveal a multifaceted landscape of the potential and challenges of integrating LLMs into health care. The results underscore a promising future for AI chatbots to assist clinical workflows by augmenting patient education and patient-clinician communication around common oral disease queries with comparable accuracy, harmfulness, and comprehensiveness to human experts. However, they also highlight existing challenges that necessitate ongoing optimization strategies since even the most capable models have some inaccuracy and inconsistency.

Comparison to Prior Work

In the comprehensive evaluation of the 3 LLMs, ChatGPT-4 emerged as the superior model, consistent with prior assessments in various medical domains [10,25,26]. This superior

XSL•FO
RenderX

performance is likely attributable to its substantially larger training data set, continuous architectural enhancements, and notable advancements in language processing, contextual comprehension, and advanced reasoning skills [20]. These improvements are crucial in health care applications, where the precision and relevance of information are critical. Interestingly, despite ChatGPT-4 showing greater stability, no significant differences were observed between ChatGPT-4 and ChatGPT-3.5 in dentist and patient evaluations. Given that ChatGPT-4 is a premium version not universally accessible, ChatGPT-3.5 holds significant value for broader applications.

In assessments spanning both periodontal and implant-related issues as well as a range of medical domains, Google Bard consistently demonstrated the least effective performance in addressing basic oral disease queries, particularly within the "diagnosis" domain. Notably, Google Bard's tendency to avoid questions about dental implant surgery pain, in contrast to ChatGPT's consistent responsiveness, might reflect differing strategies in risk management. However, in terms of readability, an important criterion for nonmedical users' educational materials, Google Bard outperformed even human experts. This aligns with prior studies assessing LLMs' readability and agrees with the impact of different training data and preprocessing methods on LLMs' readability [27,28].

Future Directions

Moreover, all 3 LLM chatbots performed similarly in providing harmless responses. In the context of medical conversation, these AI models consistently encouraged patients to seek professional medical advice, underscoring the irreplaceable role of human expertise diagnosis and treatment. However, the results of the lay user evaluation warrant caution, as they show that AI models were comparable to human experts in intent capture and helpfulness. This ambiguous distinction poses a paradox. On one hand, it suggests user acceptance in AI-provided information, underscoring their capability to effectively address user inquiries. On the other hand, it discreetly underscores a potential risk: the lay users' limited ability to judge the accuracy of complex medical information, which might inadvertently lead to AI disseminating misconceptions or inappropriate guidance. This underscores the critical need to address the ethical consideration of integrating AI in health care [29,30]. It is essential to clearly define the responsibilities and risks associated with using AI in patient education and in facilitating patient-clinician communication.

The observed performance differences among the AI models, influenced by factors like diverse training data sets and algorithmic updates, combined with the lay evaluations, emphasize the importance of customizing and continually updating LLMs for oral health care. Tailoring AI to meet specific oral health needs and maintaining current medical standards are crucial to ensure safe and accurate patient support.

Strengths and Limitations

LLMs demonstrate varied performances across different medical fields, which can be attributed to the varying depth of available web-based data on each topic. It is imperative to thoroughly evaluate their efficacy across diverse medical topics. In comparison to systemic diseases, using LLMs for basic oral health conditions offers substantial benefits. First, the narrower scope of oral diseases renders personalized oral hygiene advice and disease risk prediction via AI more viable. Second, the relative simplicity of oral structures, combined with AI's advanced image recognition capabilities, facilitates the more feasible identification and analysis of oral imagery, thus aiding early-stage problem detection. This research underscores the potential of using AI to provide individualized oral health guidance to patients, which could significantly broaden their access to medical knowledge, reduce health care expenses, enhance medical efficiency, lower public health costs, balance medical resource distribution, and relieve national economic burdens.

To our knowledge, this is the first study to evaluate the application of current LLMs comprehensively and rigorously in basic oral diseases. The robust experimental design and the implementation of blinding largely reduce evaluator bias, ensuring the validity of the results. However, this study is not without limitations. First, its methodology, based on simulated question-and-answer scenarios, does not fully replicate real-world clinical interactions. Future research should involve actual patient interactions for more accurate assessment. Second, the performance of the LLM largely depends on the quality of the prompt guiding the model, highlighting the necessity for further research in this area. With the currently rapid evolution of LLMs, there is a critical need to develop specialized chatbots with medical expertise, combining the strengths of current LLMs for health care applications. Currently, integrating medical professionals seems to be the most effective strategy for optimizing AI applications in health care.

Conclusions

LLMs, particularly ChatGPT-4, demonstrate promising potential in providing patient-centric information for common oral diseases. Variations in performance underscore the need for ongoing refinement and ethical considerations. Future studies should explore strategies to integrate LLMs effectively in health care settings, ensuring their safe and effective use in patient care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question list.
[PDF File (Adobe PDF File), 37 KB - jmir_v26i1e55847_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Evaluation axes.
[PDF File (Adobe PDF File), 49 KB - jmir_v26i1e55847_app2.pdf]

Multimedia Appendix 3
Subanalysis results of periodontal and implant-related queries.
[PDF File (Adobe PDF File), 111 KB - jmir_v26i1e55847_app3.pdf]

Multimedia Appendix 4
Subanalysis results of 6 medical care domains.
[PDF File (Adobe PDF File), 262 KB - jmir_v26i1e55847_app4.pdf]

References

1. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-04-22]
2. Schulman J, Zoph B, Kim C. ChatGPT: optimizing language models for dialogue. OpenAI blog. 2022. URL: https://chatgpt.r4wand.eu.org/# [accessed 2022-04-12]
3. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. Preprint posted online on November 24, 2023 [FREE Full text]
4. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Amazon AWS. URL: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf
5. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
7. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
8. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023;307(4):e230424. [doi: 10.1148/radiol.230424] [Medline: 37014239]
9. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
10. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
11. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
12. Peres MA, Macpherson LMD, Weyant RJ, Daly B, Venturelli R, Mathur MR, et al. Oral diseases: a global public health challenge. Lancet 2019;394(10194):249-260. [doi: 10.1016/S0140-6736(19)31146-8] [Medline: 31327369]
13. Global periodontal health: challenges, priorities and perspectives. FDI World Dental Federation. URL: https://www.fdiworlddental.org/global-periodontal-health-challenges-priorities-and-perspectives [accessed 2024-04-02]
14. White Paper on Prevention and Management of Periodontal Diseases for Oral Health and General Health. FDI World Dental Federation. URL: https://www.fdiworlddental.org/resource/white-paper-prevention-and-management-periodontal-diseases-oral-health-and-general-health [accessed 2024-04-02]
15. Watt RG, Daly B, Allison P, Macpherson LMD, Venturelli R, Listl S, et al. Ending the neglect of global oral health: time for radical action. Lancet 2019;394(10194):261-272. [doi: 10.1016/S0140-6736(19)31133-X] [Medline: 31327370]
16. Tonetti MS, Jepsen S, Jin L, Otomo-Corgel J. Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: a call for global action. J Clin Periodontol 2017;44(5):456-462 [FREE Full text] [doi: 10.1111/jcpe.12732] [Medline: 28419559]
17. Fatani B. ChatGPT for future medical and dental research. Cureus 2023;15(4):e37285 [FREE Full text] [doi: 10.7759/cureus.37285] [Medline: 37168166]
18. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent 2023;35(7):1098-1102 [FREE Full text] [doi: 10.1111/jerd.13046] [Medline: 37017291]
19. Kuehn BM. More than one-third of US individuals use the internet to self-diagnose. JAMA 2013;309(8):756-757. [doi: 10.1001/jama.2013.629] [Medline: 23443421]

XSL•FO
RenderX

20. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
21. Gemini. Google. URL: https://gemini.google.com
22. Caton JG, Armitage G, Berglundh T, Chapple ILC, Jepsen S, Kornman KS, et al. A new classification scheme for periodontal and peri-implant diseases and conditions—introduction and key changes from the 1999 classification. J Clin Periodontol 2018;45(Suppl 20):S1-S8 [FREE Full text] [doi: 10.1111/jcpe.12935] [Medline: 29926489]
23. Gjersing L, Caplehorn JRM, Clausen T. Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. BMC Med Res Methodol 2010;10:13 [FREE Full text] [doi: 10.1186/1471-2288-10-13] [Medline: 20144247]
24. Perlis RH, Fihn SD. Evaluating the application of large language models in clinical research contexts. JAMA Netw Open 2023;6(10):e2335924 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.35924] [Medline: 37782501]
25. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Sullivan PLZ, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
26. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023;13:1219326 [FREE Full text] [doi: 10.3389/fonc.2023.1219326] [Medline: 37529688]
27. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. Aesthet Surg J Open Forum 2023;5:ojad084 [FREE Full text] [doi: 10.1093/asjof/ojad084] [Medline: 37795257]
28. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023;21(1):269 [FREE Full text] [doi: 10.1186/s12967-023-04123-5] [Medline: 37076876]
29. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. [doi: 10.1186/s12967-023-04123-5]
30. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]

Abbreviations

AI: artificial intelligence
ICC: intraclass correlation coefficient
LLM: large language model
VAS: visual analog scale

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 27.12.23; peer-reviewed by L Weinert, L Zhu, W Cao; comments to author 26.02.24; revised version received 04.03.24; accepted 19.03.24; published 25.04.24.

Please cite as:
Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J
Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content
J Med Internet Res 2024;26:e55847
URL: https://www.jmir.org/2024/1/e55847
doi:10.2196/55847
PMID:38663010

©Xiaolei Lv, Xiaomeng Zhang, Yuan Li, Xinxin Ding, Hongchang Lai, Junyu Shi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Physician Versus Large Language Model Chatbot Responses to Web-Based Questions From Autistic Patients in Chinese: Cross-Sectional Comparative Analysis

Wenjie He[1,2*], MSc; Wenyan Zhang[3*], MSc; Ya Jin[4*], MSc; Qiang Zhou[2], BS; Huadan Zhang[2], BE; Qing Xia[1], MD

[1]Tianjin University of Traditional Chinese Medicine, Tianjin, China
[2]Dongguan Rehabilitation Experimental School, Dongguan, China
[3]Lanzhou University Second Hospital, Lanzhou University, Lanzhou, China
[4]Dongguan Songshan Lake Central Hospital, Guangdong Medical University, Dongguan, China
[*]these authors contributed equally

Corresponding Author:
Qing Xia, MD
Tianjin University of Traditional Chinese Medicine
10 Poyang Lake Road
Tuanpo New Town West, Jinghai District
Tianjin, 301617
China
Phone: 86 13820689541
Email: xiaqingcho@163.com

Abstract

Background: There is a dearth of feasibility assessments regarding using large language models (LLMs) for responding to inquiries from autistic patients within a Chinese-language context. Despite Chinese being one of the most widely spoken languages globally, the predominant research focus on applying these models in the medical field has been on English-speaking populations.

Objective: This study aims to assess the effectiveness of LLM chatbots, specifically ChatGPT-4 (OpenAI) and ERNIE Bot (version 2.2.3; Baidu, Inc), one of the most advanced LLMs in China, in addressing inquiries from autistic individuals in a Chinese setting.

Methods: For this study, we gathered data from DXY—a widely acknowledged, web-based, medical consultation platform in China with a user base of over 100 million individuals. A total of 100 patient consultation samples were rigorously selected from January 2018 to August 2023, amounting to 239 questions extracted from publicly available autism-related documents on the platform. To maintain objectivity, both the original questions and responses were anonymized and randomized. An evaluation team of 3 chief physicians assessed the responses across 4 dimensions: relevance, accuracy, usefulness, and empathy. The team completed 717 evaluations. The team initially identified the best response and then used a Likert scale with 5 response categories to gauge the responses, each representing a distinct level of quality. Finally, we compared the responses collected from different sources.

Results: Among the 717 evaluations conducted, 46.86% (95% CI 43.21%-50.51%) of assessors displayed varying preferences for responses from physicians, with 34.87% (95% CI 31.38%-38.36%) of assessors favoring ChatGPT and 18.27% (95% CI 15.44%-21.10%) of assessors favoring ERNIE Bot. The average relevance scores for physicians, ChatGPT, and ERNIE Bot were 3.75 (95% CI 3.69-3.82), 3.69 (95% CI 3.63-3.74), and 3.41 (95% CI 3.35-3.46), respectively. Physicians (3.66, 95% CI 3.60-3.73) and ChatGPT (3.73, 95% CI 3.69-3.77) demonstrated higher accuracy ratings compared to ERNIE Bot (3.52, 95% CI 3.47-3.57). In terms of usefulness scores, physicians (3.54, 95% CI 3.47-3.62) received higher ratings than ChatGPT (3.40, 95% CI 3.34-3.47) and ERNIE Bot (3.05, 95% CI 2.99-3.12). Finally, concerning the empathy dimension, ChatGPT (3.64, 95% CI 3.57-3.71) outperformed physicians (3.13, 95% CI 3.04-3.21) and ERNIE Bot (3.11, 95% CI 3.04-3.18).

Conclusions: In this cross-sectional study, physicians' responses exhibited superiority in the present Chinese-language context. Nonetheless, LLMs can provide valuable medical guidance to autistic patients and may even surpass physicians in demonstrating empathy. However, it is crucial to acknowledge that further optimization and research are imperative prerequisites before the effective integration of LLMs in clinical settings across diverse linguistic environments can be realized.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300074655; https://www.chictr.org.cn/bin/project/edit?pid=199432

(J Med Internet Res 2024;26:e54706) doi:10.2196/54706

KEYWORDS

artificial intelligence; chatbot; ChatGPT; ERNIE Bot; autism

Introduction

Artificial intelligence (AI) has revolutionized human-computer interaction, reshaping communication, learning, and creativity paradigms [1,2]. A significant advancement in this realm is the emergence of large language models (LLMs), which have enabled the development of versatile digital assistants capable of understanding and generating human language [3-5]. Through extensive training on textual data, LLMs have acquired profound knowledge across diverse domains, facilitating coherent and contextually relevant interactions in natural language conversations. These models find applications in various domains, including natural language processing, question answering, language generation, and interactive dialogues [6-10]. Moreover, several studies have documented the use of LLMs in the medical field such as medication consultation [11], health education [12], and medical guidance [13,14].

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental condition characterized by profound social and psychological challenges [15,16]. The estimated prevalence of ASD is approximately 1 in 36 children, with China reporting a prevalence of 1% [17-19], making it a significant public health concern. However, constraints on health care infrastructure development in China have led to resource shortages in numerous regions [20,21], exacerbating the burden on families and society. AI assistants represent underused resources for enhancing diagnosis and treatment efficiency in health care [22]. ChatGPT (OpenAI) [23] and ERNIE Bot (Baidu, Inc) [24] represent AI technologies powered by advancements in LLM. ChatGPT is a model with 20 billion parameters [25]. ERNIE Bot's training data, as promoted at ERNIE Bot conference, includes trillions of web page data, billions of search data and image data, tens of billions of daily voice call data, and 550 billion facts of a knowledge graph, which distinguishes Baidu's Chinese-language processing capabilities [26]. While ChatGPT gained widespread recognition for its ability to generate humanlike text across diverse topics [27,28], ERNIE Bot represents the forefront of AI technology in China [29]. Despite their original non–health care focus, their potential to assist in addressing patient inquiries remains unexplored mainly [30-32]. Implementing tiered diagnosis and treatment systems to optimize medical resource use may limit patients' access to high-quality health care [33].

This study aims to investigate the performance of 2 conversational agents, ERNIE Bot and ChatGPT, in supporting individuals with ASD during web-based interactions. Our hypotheses are 2-fold. First, we hypothesize that ERNIE Bot, developed in China and trained on a data set that includes more Chinese text, may exhibit superior performance compared to ChatGPT, particularly regarding cultural relevance and linguistic nuances. Second, we anticipate that both ERNIE Bot and ChatGPT will demonstrate efficacy in assisting individuals with ASD, as evidenced by their ability to engage users effectively and provide helpful responses during conversational exchanges. Researchers have conducted numerous studies in English evaluating the benefits of LLMs in the medical field. Given the global significance of the Chinese language, this study aims to assess the capability of LLMs to provide high-quality and empathetic responses to health care queries posed by autistic individuals in China.

Methods

Data Source

This cross-sectional study aimed to construct a database of inquiries from autistic individuals by aggregating publicly available data from the web-based medical consultation platform DXY [34]. In China, chatbots are not permitted in clinical settings due to existing regulations, prompting the consideration of DXY as a feasible substitute. DXY is a prominent digital health care technology company with a 2-decade track record. The company offers a range of health-related applications, including high-quality health information dissemination, general knowledge services, a web-based medical consultation platform, health product e-commerce, and offline medical treatment. DXY caters to more than 100 million general users and has a user base of 5.5 million professionals, including 2.1 million physicians, constituting approximately 71% of the total number of physicians in the country.

The objective of this cross-sectional study was to analyze 200 cases to detect a 10% disparity (45% vs 55%) in responses provided by physicians and chatbots, with an assumed statistical power of 80%. We planned to use publicly accessible autism-related consultation records from the DXY website. Our sample comprised 100 randomly selected patients from the consultation records from January 2018 to August 2023. Each patient posed 1 to 3 questions, resulting in a total collection of 239 consultation queries. The qualifications of the responding health care professionals ranged from general to chief physicians.

Ethical Considerations

We adhered strictly to the terms and conditions of DXY for all analyses, and the medical ethics committee of Lanzhou University Second Hospital approved them (approval 2023A-420). The study used publicly available data from the consultation platform, did not involve personal patient information or direct test subjects, and thus did not require informed consent. We registered the study on the Chinese Clinical Trial Registry (ChiCTR2300074655).

XSL•FO
RenderX

Text Generation With an LLM Chatbot

To closely simulate an authentic medical consultation process, original questions were introduced into a new chatbot conversation from August 16, 2023, to August 30, 2023. In this dialogue, any questions previously posed with a potential impact on the outcomes were deliberately excluded. Both GPT-4 and ERNIE Bot 2.2.3 versions were used for this purpose. After eliminating expressions indicative of AI features, all responses were systematically collected and organized into a structured question-and-answer data set. The consultation content, which includes regional dialects and typographical errors, was carefully maintained to replicate a medical consultation authentically. The directly quoted content was used to prompt responses from the chatbot. The chatbot simulated a physician's responses to mimic them while closely hiding its AI identity.

Expert Evaluation

A team of 3 chief physicians specializing in child psychiatry and pediatric health care from distinct hospitals comprehensively reviewed the original questions and physicians' and chatbot's responses. The evaluators were presented with complete patient questions and physician and chatbot responses. The responders' identities were anonymized; randomized; and labeled as responses 1, 2, or 3 to ensure that evaluators remained blinded. The evaluators were instructed to thoroughly examine the entire patient question and 3 responses before assessing the quality of the interactions. The evaluation process commenced with identifying the superior response and evaluating the responses across the 4 dimensions using a Likert scale: relevance, correctness, usefulness, and humaneness. The Likert-scale options for each dimension were as follows: relevance (irrelevant, somewhat relevant, partially relevant, relevant, or very relevant), correctness (incorrect, primarily incorrect, partially correct, correct, or very correct), usefulness (useless, of limited use, somewhat useful, useful, or very useful), and humaneness (lacking, slightly humane, moderately humane, humane, or very humane). Researchers assigned ratings on a 1-5 scale, with 1 representing the lowest quality and 5 representing the highest quality. Finally, a comparative assessment of the 3 responses was performed, with the quality dimensions for response evaluation detailed in Textbox 1.

Textbox 1. Quality dimensions in the comparative evaluation of responses.

Relevance

- This dimension evaluates the alignment of responses with test results, emphasizing the system's ability to generate appropriate text addressing specific issues rather than diverging into unrelated scenarios.

Correctness

- The correctness dimension focuses exclusively on the accuracy of information within the response, irrespective of the patient's question. It gauges the scientific and technical precision of explanations based on best medical evidence and practices.

Usefulness

- This dimension combines the relevance and correctness of the system and evaluates its ability to provide non-obvious insights to patients, non-professionals, and laypersons. It includes providing appropriate recommendations, supplying relevant and accurate information, enhancing patient understanding of test results, and advising actions that optimize health care service use for the patient's benefit.

Empathy

- Empathy involves demonstrating abundant respect, effective communication, compassion, and seeking emotional connections with patients. It encompasses recognizing and empathizing with their experience, respecting their thoughts, addressing their concerns patiently, and sincerely promoting their physical and mental well-being. Additionally, empathy entails humanely fulfilling patients' and their families' physical, psychological, social, and spiritual needs.

Data Analysis

We used the Kruskal-Wallis H test to assess and compare the quality of the responses provided by physicians, ChatGPT, and ERNIE Bot along 4 dimensions: relevance, correctness, usefulness, and empathy. We presented the distribution of responses from each source including preferences for physicians, ChatGPT, and ERNIE Bot. Furthermore, we examined the proportion of responses that exceeded or fell below critical thresholds such as relevance, correctness, and usefulness. We compared the distribution of these threshold proportions among responses from physicians, ChatGPT, and ERNIE Bot. All statistical analyses were performed using SPSS (version 27.0; IBM), with a significance level set at $P<.05$ (2-tailed).

Results

Preferred Responses

This study included 717 evaluations of 239 randomly selected consultation questions. Evaluators indicated their preferences for physicians, ChatGPT, or ERNIE Bot at proportions of 46.86% (336/717; 95% CI 43.21%-50.51%), 34.87% (250/717; 95% CI 31.38%-38.36%), and 18.27% (131/717; 95% CI 15.44%-21.10%), respectively.

Relevance

The distribution of the relevance scores among the 3 groups was not entirely uniform, exhibiting statistically significant differences (H=111.67, $P<.001$). Physician responses demonstrated higher relevance than chatbots (ChatGPT or ERNIE Bot). Specifically, the mean relevance score for

XSL•FO
RenderX

physician responses was 3.75 (95% CI 3.69-3.82), whereas the mean relevance scores for ChatGPT and ERNIE Bot were 3.69 (95% CI 3.63-3.74) and 3.41 (95% CI 3.35-3.46), respectively (Figure 1). The proportion of responses rated as off topic (score <4) was lower for physicians (176/717, 24.55%; 95% CI 21.40%-27.70%) than for ChatGPT (258/717, 35.98%; 95% CI 32.47%-39.49%) and ERNIE Bot (366/717, 51.05%; 95% CI 47.39%-54.71%). Post hoc pairwise comparisons using the Bonferroni correction for significance levels revealed statistically significant differences in relevance scores between all 3 groups, specifically between physicians and ChatGPT (*P*<.001), physicians and ERNIE Bot (*P*<.001), and ChatGPT and ERNIE Bot (*P*<.001).

Figure 1. Distribution of relevance scores for responses to autistic patient questions by physicians and chatbots.

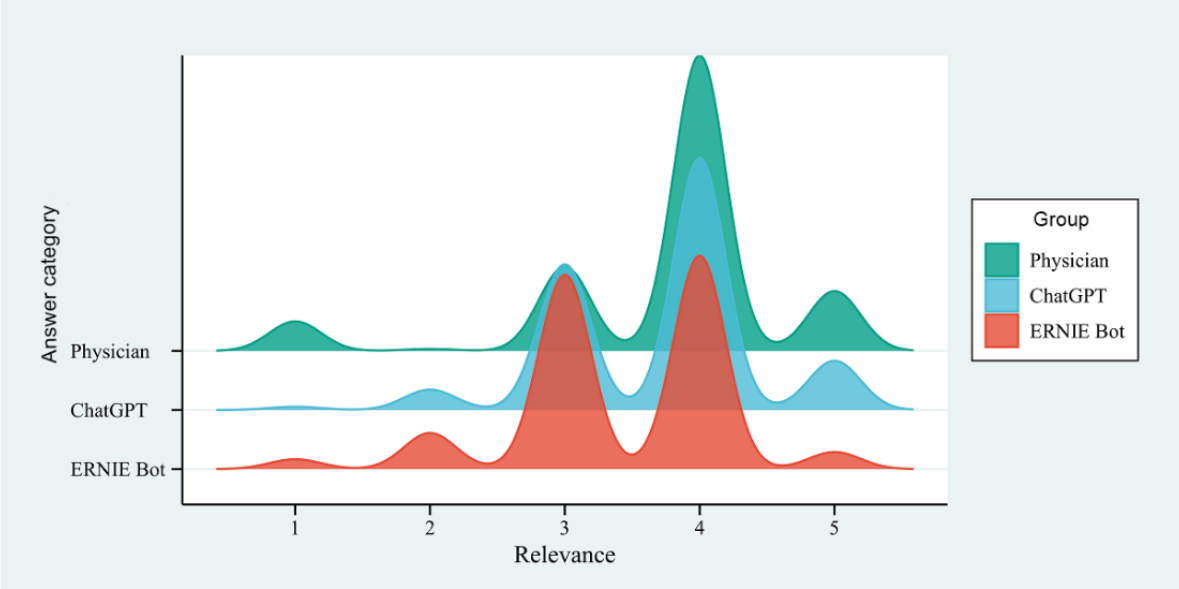

Correctness

The mean correctness scores for physicians, ChatGPT, and ERNIE Bot were 3.66 (95% CI 3.60-3.73), 3.73 (95% CI 3.69-3.77), and 3.52 (95% CI 3.47-3.57), respectively (Figure 2). Physicians and ChatGPT achieved higher correctness scores among these 3 groups than ERNIE Bot. The distribution of the correctness scores exhibited statistically significant differences among the 3 groups (H=49.99, *P*<.001). When comparing the correctness scores between the 3 groups, the differences between physicians and ChatGPT (*P*=.58) were not statistically significant. However, the differences between physicians and ERNIE Bot (*P*<.001) and between ChatGPT and ERNIE Bot (*P*<.001) were both statistically significant. The proportion of responses with errors (score <4) was similar for physicians (196/717, 27.34%; 95% CI 24.08%-30.60%) and ChatGPT (211/717, 29.43%; 95% CI 26.09%-32.77%) and lower for ERNIE Bot (309/717, 43.10%; 95% CI 39.48%-46.72%).

Figure 2. Distribution of correctness scores for responses to autistic patient questions by physicians and chatbots.

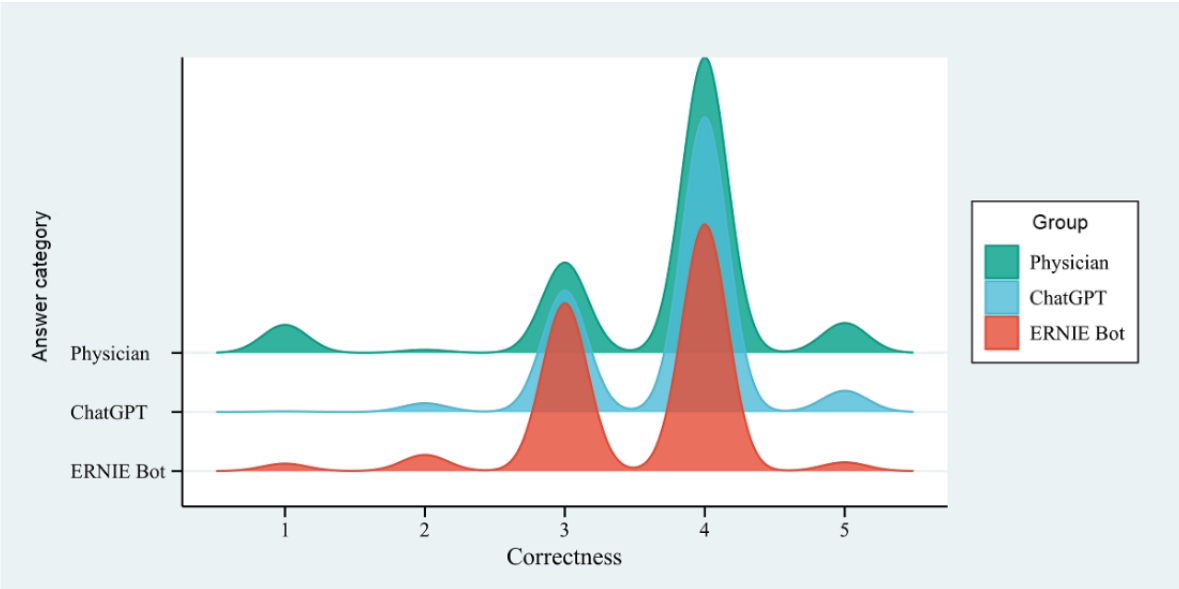

XSL•FO
RenderX

Usefulness

Among the 3 response groups, physician responses exhibited higher levels of usefulness than chatbots (ChatGPT or ERNIE Bot). Specifically, the mean usefulness score for physician responses was 3.54 (95% CI 3.47-3.62), whereas the mean usefulness scores for ChatGPT and ERNIE Bot were 3.40 (95% CI 3.34-3.47) and 3.05 (95% CI 2.99-3.12), respectively (Figure 3). The proportion of responses rated as useful (score ≥4) was more significant for physicians (428/717, 59.69%; 95% CI 56.10%-63.28%) than for the chatbots (ChatGPT: 362/717, 50.49%; 95% CI 46.83%-54.15%; ERNIE Bot: 215/717, 29.99%; 95% CI 26.64%-33.34%). The distribution of the usefulness scores among the 3 groups displayed statistically significant differences (H=135.81, *P*<.001). Notably, all 3 pairwise comparisons of usefulness scores were statistically significant, with adjusted *P* values indicating significance, specifically between physicians and ChatGPT (*P*<.001), physicians and ERNIE Bot (*P*<.001), and ChatGPT and ERNIE Bot (*P*<.001).

Figure 3. Distribution of usefulness scores for responses to autistic patient questions by physicians and chatbots.

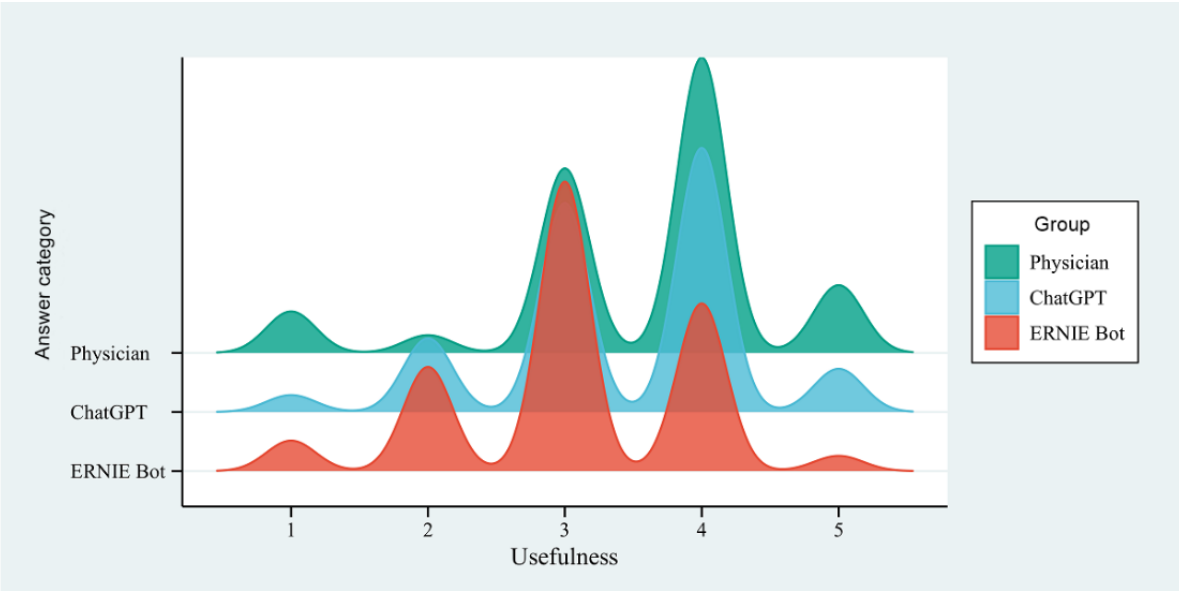

Empathy

The mean empathy score for ChatGPT was 3.64 (95% CI 3.57-3.71), whereas the mean empathy scores for physicians and ERNIE Bot were 3.13 (95% CI 3.04-3.21) and 3.11 (95% CI 3.04-3.18), respectively (Figure 4). ChatGPT's responses received higher empathy scores than physicians and ERNIE Bot within the 3 response groups. Specifically, the proportion of responses displaying empathy (score ≥4) was higher for ChatGPT (447/717, 62.34%; 95% CI 58.79%-65.89%) than for physicians (312/717, 43.51%; 95% CI 39.88%-47.14%) and ERNIE Bot (258/717, 35.98%; 95% CI 32.47%-39.49%). The distribution of empathy scores among the 3 groups revealed significant differences (H=118.58, *P*<.001). When assessing empathy scores among the 3 groups, the differences between physicians and ChatGPT (*P*<.001) and between ChatGPT and ERNIE Bot (*P*<.001) both demonstrated statistical significance. However, the disparities between physicians and ERNIE Bot (*P*=.14) were not significant.

The evaluators performed a reliability assessment, which revealed robust repeatability. The intraclass correlation coefficient values for the 4 response categories (relevance, correctness, usefulness, and empathy) were 0.812 (*P*<.001), 0.831 (*P*<.001), 0.818 (*P*<.001), and 0.863 (*P*<.001), respectively.

XSL•FO
RenderX

Figure 4. Distribution of empathy scores for responses to autistic patient questions by physicians and chatbots.

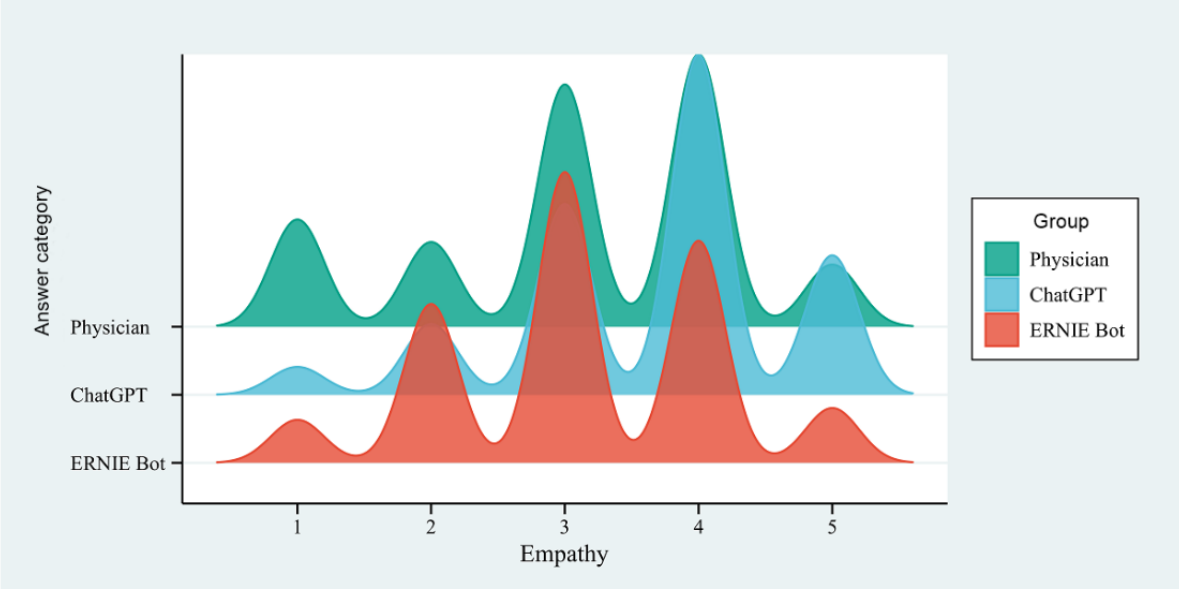

Discussion

Principal Findings

This study evaluated the capabilities of LLMs, such as ChatGPT and ERNIE Bot, in delivering quality and empathetic responses to medical queries from Chinese autistic individuals. To simulate genuine clinical scenarios, publicly accessible data from web-based medical consultation platforms were used in this cross-sectional investigation. Notably, prevailing regulations in China strictly prohibit the use of AI-generated prescriptions. The study results revealed that expert evaluators favored responses from physicians over those generated by chatbots (ChatGPT or ERNIE Bot). In contrast to previous ophthalmology research that demonstrated chatbots outperforming physicians, our findings indicated that physicians received higher scores in relevance, correctness, and usefulness, with only a slight margin behind ChatGPT in terms of empathy. Conversely, ERNIE Bot obtained the lowest scores across all 4 dimensions: relevance, correctness, usefulness, and empathy.

Bernstein et al [35] compared the occurrence of incorrect or inappropriate content, potential harm, and degree of harm in responses from chatbots and humans. The study indicated that chatbot responses exhibited a similar likelihood of containing incorrect or inappropriate material compared with human responses (prevalence ratio [PR] 0.92, 95% CI 0.77-1.10). Moreover, no significant differences were observed between chatbot and human responses regarding potential harm (PR 0.84, 95% CI 0.67-1.07) or the degree of harm (PR 0.99, 95% CI 0.80-1.22). These results suggested that LLMs can provide appropriate ophthalmological advice for varying levels of complexity in patient questions. Shao et al [36] developed a set of 37 questions for patient education on thoracic surgery during the perioperative period, covering topics such as disease information, diagnostic procedures, perioperative complications, treatment measures, disease prevention, and perioperative care instructions. An assessment of responses in both English and Chinese contexts revealed that 92% (n=34) were considered appropriate and comprehensive. This study highlighted the potential feasibility of using ChatGPT for patient education in thoracic surgery in both English and Chinese settings. Zhu et al [37] investigated the application of ChatGPT as a mediator between physicians and patients in Chinese-speaking outpatient settings, focusing mainly on the Chinese Physician Qualification Examination. The study reported an average score of 72.4% in the clinical knowledge section, which was placed within the top 20 percentile. These findings suggested that ChatGPT can facilitate physician-patient communication in Chinese-speaking outpatient settings. Ayers et al [38] used publicly available data from a social media forum, Reddit, to compare physician responses and the ChatGPT. They randomly selected 195 dialogues with questions answered by physicians and generated chatbot responses by inputting the original questions into a new ChatGPT session. Evaluators assessed the responses from both physicians and the ChatGPT using a Likert scale, considering preferences, information quality, and empathy or interpersonal style. In 585 evaluations, 78.6% (95% CI 75%-81.8%) of evaluators preferred chatbot responses over those from physicians. These studies highlight the significant potential of chatbots in the field of medicine.

In our study, which consisted of 717 evaluations, evaluators preferred physician responses over those from chatbots, namely ChatGPT or ERNIE Bot. The preferences for physicians, ChatGPT, and ERNIE Bot were 46.86% (n=336; 95% CI 43.21%-50.51%), 34.87% (n=250; 95% CI 31.38%-38.36%), and 18.27% (n=131; 95% CI 15.44%-21.10%), respectively. Physician responses achieved higher scores in relevance, accuracy, and usefulness, with the only exception being the dimension of empathy, which ChatGPT surpassed. Our cross-sectional study's results differed from previous reports due to several factors. First, unlike previous research primarily conducted in English-speaking settings, this study was conducted in a Chinese-speaking environment. Second, we preserved the original medical questions from autistic patients without modification to simulate authentic clinical consultations, including errors or nonstandard expressions in the queries such as misspellings or dialects. Our study contrasted with previous

XSL•FO
RenderX

research, which often relies on professionally standardized queries. Furthermore, autistic patients' queries frequently entail subjective matters such as seeking recommendations for specialist physicians or autism-related resources, lacking standardized responses, and reflecting significant cultural variations. Finally, our study used samples from a paid web-based medical consultation platform, whereas Ayers et al [38] used dialogues from public social forums on Reddit. Physicians may exhibit more proactive and diligent engagement in paid consultations.

Our study compared the performance of ChatGPT and ERNIE Bot in physician-patient interactions, with ERNIE Bot trained in Chinese and ChatGPT in English. While one might assume that ERNIE Bot's training in Chinese would result in greater empathy toward Chinese-speaking users than ChatGPT, the results challenge this notion. The findings suggest that factors beyond the language of training influence the empathetic responsiveness of LLMs, highlighting the complexity of human-AI interactions and emphasizing the need for further exploration of the relationship between language and empathy. Physicians responded better when patients asked for recommendations on specific Chinese books about ASD. Physicians also effectively handled situations where the patient's condition was misstated, whereas LLMs provided inaccurate information. Additionally, creating user-friendly interfaces to accommodate patients with varying levels of technological proficiency could improve the accessibility and usability of AI models in health care settings.

Limitations

This study had several notable limitations. First, reliance on web-based consultation platform records constrained each autistic patient to a maximum of 3 questions per web-based consultation, potentially limiting the ability to replicate real-world patient-physician interactions comprehensively. Moreover, the study exclusively examined text-based responses to patient inquiries, neglecting the potential for health care professionals to tailor their responses based on individual patient characteristics such as occupation and emotional state. The extent to which clinical professionals can adapt their responses to such personalization remains uncertain. Additionally, the study did not assess the chatbot's capacity to extract information from health records, representing an area of potential improvement. Finally, although the evaluators have been single blinded, they potentially introduced bias into their assessments because they are coauthors of the paper and may hope that there are apparent differences between the groups in the results to make extreme scores.

Conclusions

The findings of this cross-sectional study show that physician responses outperform those of LLMs in the Chinese context, responding to inquiries from autistic patients in text-based formats compared with responses from current state-of-the-art LLMs. Nevertheless, these LLMs can offer medical guidance for autistic patients and demonstrate greater empathy compared with physicians, particularly ChatGPT-4. It is essential to emphasize that further refinement and comprehensive research are prerequisites before deploying LLMs effectively in clinical settings across diverse linguistic environments.

Acknowledgments

The authors extend their appreciation to the assessors for their professional contributions in conducting the measurements in this study. Special thanks are extended to Shanghai Tengyun Biotechnology Co, Ltd for developing the Hiplot Pro platform and their valuable technical assistance and tools for data visualization. This research received financial support from the National Natural Science Foundation of China (82105021). The funders were not involved in the study design, execution, or reporting. This study represents an inaugural assessment of the effectiveness of 2 prominent large language models, ChatGPT from the United States and ERNIE Bot from China, in providing high-quality and empathetic responses to medical inquiries from Chinese-speaking autistic patients.

Data Availability

All data generated or analyzed during this study are included in this published article. The data for this study are contained in Multimedia Appendix 1.

Authors' Contributions

QX initiated and designed this study. WH collected and analyzed the data and drafted the manuscript. WZ guided English writing. YJ created the figures and tables. QZ and HZ conducted literature searches and participated in participant recruitment. QX revised the manuscript. QX managed the project and secured funding. All the authors have reviewed and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The raw data.
[XLSX File (Microsoft Excel File), 446 KB - jmir_v26i1e54706_app1.xlsx]

XSL•FO
RenderX

References

1. Edelmann A, Wolff T, Montagne D, Bail CA. Computational social science and sociology. Annu Rev Sociol 2020;46(1):61-81 [FREE Full text] [doi: 10.1146/annurev-soc-121919-054621] [Medline: 34824489]
2. Cheng CY, Chiu IM, Hsu MY, Pan HY, Tsai CM, Lin CHR. Deep learning assisted detection of abdominal free fluid in Morison's pouch during focused assessment with sonography in trauma. Front Med (Lausanne) 2021;8:707437 [FREE Full text] [doi: 10.3389/fmed.2021.707437] [Medline: 34631730]
3. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls. Ann Intern Med 2024;177(2):210-220. [doi: 10.7326/m23-2772]
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
6. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol 2023;13:1268915 [FREE Full text] [doi: 10.3389/fonc.2023.1268915] [Medline: 37731643]
7. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023;25(1):bbad493 [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
8. Lai TM, Zhai C, Ji H. KEBLM: Knowledge-Enhanced Biomedical Language Models. J Biomed Inform 2023;143:104392. [doi: 10.1016/j.jbi.2023.104392] [Medline: 37211194]
9. Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, Bin Saleh K, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Social Adm Pharm 2023;19(8):1236-1242 [FREE Full text] [doi: 10.1016/j.sapharm.2023.05.016] [Medline: 37321925]
10. Müller M, Salathé M, Kummervold PE. COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 content on Twitter. Front Artif Intell 2023;6:1023281 [FREE Full text] [doi: 10.3389/frai.2023.1023281] [Medline: 36998290]
11. Hsu HY, Hsu KC, Hou SY, Wu CL, Hsieh YW, Cheng YD. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med Educ 2023;9:e48433 [FREE Full text] [doi: 10.2196/48433] [Medline: 37561097]
12. Roster K, Kann RB, Farabi B, Gronbeck C, Brownstone N, Lipner SR. Readability and health literacy scores for ChatGPT-generated dermatology public education materials: cross-sectional analysis of sunscreen and melanoma questions. JMIR Dermatol 2024;7:e50163 [FREE Full text] [doi: 10.2196/50163] [Medline: 38446502]
13. Karakas C, Brock DB, Lakhotia A. Leveraging ChatGPT in the pediatric neurology clinic: practical considerations for use to improve efficiency and outcomes. Pediatr Neurol 2023;148:157-163. [doi: 10.1016/j.pediatrneurol.2023.08.035] [Medline: 37725885]
14. Yang J, Ardavanis KS, Slack KE, Fernando ND, Della Valle CJ, Hernandez NM. Chat Generative Pretrained Transformer (ChatGPT) and Bard: artificial intelligence does not yet provide clinically supported answers for hip and knee osteoarthritis. J Arthroplasty 2024;39(5):1184-1190. [doi: 10.1016/j.arth.2024.01.029] [Medline: 38237878]
15. Lord C, Charman T, Havdahl A, Carbone P, Anagnostou E, Boyd B, et al. The Lancet commission on the future of care and clinical research in autism. Lancet 2022;399(10321):271-334. [doi: 10.1016/S0140-6736(21)01541-5] [Medline: 34883054]
16. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, et al. Prevalence and characteristics of Autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2020. MMWR Surveill Summ 2023;72(2):1-14 [FREE Full text] [doi: 10.15585/mmwr.ss7202a1] [Medline: 36952288]
17. Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcín C, et al. Global prevalence of autism and other pervasive developmental disorders. Autism Res 2012;5(3):160-179 [FREE Full text] [doi: 10.1002/aur.239] [Medline: 22495912]
18. Sun X, Allison C, Wei L, Matthews FE, Auyeung B, Wu YY, et al. Autism prevalence in China is comparable to western prevalence. Mol Autism 2019;10:7 [FREE Full text] [doi: 10.1186/s13229-018-0246-0] [Medline: 30858963]
19. Luo Y, Pang L, Guo C, Zhang L, Wang Y, Zheng X. Urbanicity and autism of children in China. Psychiatry Res 2020;286:112867. [doi: 10.1016/j.psychres.2020.112867] [Medline: 32120171]
20. Liu J, Miao J, Zhang D. Dilemma of healthcare reform and invention of new discipline of health fiscalogy. Glob Health Res Policy 2016;1:4 [FREE Full text] [doi: 10.1186/s41256-016-0003-x] [Medline: 29202054]
21. Feng J, Gong Y, Li H, Wu J, Lu Z, Zhang G, et al. Development trend of primary healthcare after health reform in China: a longitudinal observational study. BMJ Open 2022;12(6):e052239 [FREE Full text] [doi: 10.1136/bmjopen-2021-052239] [Medline: 35676005]
22. Qiao Y, Zhao L, Luo C, Luo Y, Wu Y, Li S, et al. Multi-modality artificial intelligence in digital pathology. Brief Bioinform 2022;23(6):bbac367 [FREE Full text] [doi: 10.1093/bib/bbac367] [Medline: 36124675]
23. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-04-26]

XSL•FO
RenderX

24. ERNIE Bot. Baidu. URL: https://yiyan.baidu.com/ [accessed 2024-04-26]
25. Singh M, Cambronero J, Gulwani S, Le V, Negreanu C, Verbruggen G. CodeFusion: a pre-trained diffusion model for code generation. ArXiv 2023 Preprint posted online on October 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.17680]
26. Launch event for Baidu's ERNIE Bot. Quklive. Baidu, Inc. 2023. URL: https://cloud.quklive.com/cloud/a/embed/1678784107545129 [accessed 2024-04-16]
27. Schukow C, Smith SC, Landgrebe E, Parasuraman S, Folaranmi OO, Paner GP, et al. Application of ChatGPT in routine diagnostic pathology: promises, pitfalls, and potential future directions. Adv Anat Pathol 2024;31(1):15-21. [doi: 10.1097/PAP.0000000000000406] [Medline: 37501529]
28. Hwang T, Aggarwal N, Khan PZ, Roberts T, Mahmood A, Griffiths MM, et al. Can ChatGPT assist authors with abstract writing in medical journals? evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. PLoS One 2024;19(2):e0297701 [FREE Full text] [doi: 10.1371/journal.pone.0297701] [Medline: 38354135]
29. Jia K, Kenney M, Mattila J, Seppala T. The application of artificial intelligence at Chinese digital platform giants: Baidu, Alibaba and Tencent. SSRN Journal 2018 Preprint posted online on April 25, 2018 [FREE Full text] [doi: 10.2139/ssrn.3154038]
30. Arora A, Arora A. The promise of large language models in health care. Lancet 2023;401(10377):641. [doi: 10.1016/S0140-6736(23)00216-7] [Medline: 36841609]
31. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023;330(9):866-869 [FREE Full text] [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
32. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
33. Dublin S, Greenwood-Hickman MA, Karliner L, Hsu C, Coley RY, Colemon L, et al. The electronic health record Risk of Alzheimer's and Dementia Assessment Rule (eRADAR) brain health trial: protocol for an embedded, pragmatic clinical trial of a low-cost dementia detection algorithm. Contemp Clin Trials 2023;135:107356. [doi: 10.1016/j.cct.2023.107356] [Medline: 37858616]
34. Deng Z, Deng Z, Liu S, Evans R. Knowledge transfer between physicians from different geographical regions in China's online health communities. Inf Technol Manag 2023:1-18 [FREE Full text] [doi: 10.1007/s10799-023-00400-3] [Medline: 37359990]
35. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
36. Shao CY, Li H, Liu XL, Li C, Yang LQ, Zhang YJ, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. Interact J Med Res 2023;12:e46900 [FREE Full text] [doi: 10.2196/46900] [Medline: 37578819]
37. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-English-speaking outpatient clinic settings. Digit Health 2023;9:20552076231184091 [FREE Full text] [doi: 10.1177/20552076231184091] [Medline: 37434733]
38. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]

Abbreviations

AI: artificial intelligence
ASD: autism spectrum disorder
LLM: large language model
PR: prevalence ratio

Edited by T Leung, S Ma; submitted 20.11.23; peer-reviewed by N Singh, J Li; comments to author 29.02.24; revised version received 20.03.24; accepted 02.04.24; published 30.04.24.

Please cite as:
He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q
Physician Versus Large Language Model Chatbot Responses to Web-Based Questions From Autistic Patients in Chinese: Cross-Sectional Comparative Analysis
J Med Internet Res 2024;26:e54706
URL: https://www.jmir.org/2024/1/e54706
doi:10.2196/54706
PMID:38687566

XSL•FO
RenderX

©Wenjie He, Wenyan Zhang, Ya Jin, Qiang Zhou, Huadan Zhang, Qing Xia. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study

Lars Masanneck[1,2], MSc, MD; Linea Schmidt[2], MSc; Antonia Seifert[3], BA; Tristan Kölsche[1], MD; Niklas Huntemann[1], MSc, MD; Robin Jansen[1], MD; Mohammed Mehsin[1], MD; Michael Bernhard[3], MHBA, MD; Sven G Meuth[1], MHBA, MD, PhD; Lennert Böhm[3], MD; Marc Pawlitzki[1], MD

[1]Department of Neurology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2]Digital Health Center, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
[3]Emergency Department, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Corresponding Author:
Lars Masanneck, MSc, MD
Department of Neurology
Medical Faculty and University Hospital Düsseldorf
Heinrich Heine University Düsseldorf
Moorenstraße 5
Düsseldorf, 40225
Germany
Phone: 49 0211 81 17880
Email: lars.masanneck@med.uni-duesseldorf.de

Abstract

Background: Large language models (LLMs) have demonstrated impressive performances in various medical domains, prompting an exploration of their potential utility within the high-demand setting of emergency department (ED) triage. This study evaluated the triage proficiency of different LLMs and ChatGPT, an LLM-based chatbot, compared to professionally trained ED staff and untrained personnel. We further explored whether LLM responses could guide untrained staff in effective triage.

Objective: This study aimed to assess the efficacy of LLMs and the associated product ChatGPT in ED triage compared to personnel of varying training status and to investigate if the models' responses can enhance the triage proficiency of untrained personnel.

Methods: A total of 124 anonymized case vignettes were triaged by untrained doctors; different versions of currently available LLMs; ChatGPT; and professionally trained raters, who subsequently agreed on a consensus set according to the Manchester Triage System (MTS). The prototypical vignettes were adapted from cases at a tertiary ED in Germany. The main outcome was the level of agreement between raters' MTS level assignments, measured via quadratic-weighted Cohen κ. The extent of over- and undertriage was also determined. Notably, instances of ChatGPT were prompted using zero-shot approaches without extensive background information on the MTS. The tested LLMs included raw GPT-4, Llama 3 70B, Gemini 1.5, and Mixtral 8x7b.

Results: GPT-4–based ChatGPT and untrained doctors showed substantial agreement with the consensus triage of professional raters (κ=mean 0.67, SD 0.037 and κ=mean 0.68, SD 0.056, respectively), significantly exceeding the performance of GPT-3.5–based ChatGPT (κ=mean 0.54, SD 0.024; $P<.001$). When untrained doctors used this LLM for second-opinion triage, there was a slight but statistically insignificant performance increase (κ=mean 0.70, SD 0.047; $P=.97$). Other tested LLMs performed similar to or worse than GPT-4–based ChatGPT or showed odd triaging behavior with the used parameters. LLMs and ChatGPT models tended toward overtriage, whereas untrained doctors undertriaged.

Conclusions: While LLMs and the LLM-based product ChatGPT do not yet match professionally trained raters, their best models' triage proficiency equals that of untrained ED doctors. In its current form, LLMs or ChatGPT thus did not demonstrate gold-standard performance in ED triage and, in the setting of this study, failed to significantly improve untrained doctors' triage when used as decision support. Notable performance enhancements in newer LLM versions over older ones hint at future improvements with further technological development and specific training.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e53297) doi:10.2196/53297

KEYWORDS

emergency medicine; triage; artificial intelligence; large language models; ChatGPT; untrained doctors; doctor; doctors; comparative study; digital health; personnel; staff; cohort; Germany; German

Introduction

In recent years, machine learning techniques have been integrated into various aspects of medical care, serving as supportive diagnostic algorithms in clinical applications that include electrocardiograms [1], skin lesion classification [2], and radiological imaging [3]. With the recent boom of generative artificial intelligence (AI) and large language models (LLMs), thrust into the spotlight by the release of ChatGPT (Open AI) [4] in November 2022, the application of generative AI methods has sparked widespread discussion and experimentation in the medical domain.

To many medical professionals, it is remarkable that systems such as ChatGPT without any specialized training have been able to pass US medical exams [5,6]; provide answers to questions from a web-based platform that are preferred over doctors' replies [7]; and in most cases, offer appropriate recommendations to questions on cardiovascular disease prevention [8]. The capabilities of text generation and summarization introduce numerous controversial research and clinical use cases such as creating discharge summaries [9] or crafting scientific abstracts [10]. Recently, even commercial solutions for automated documentation of clinician-patient interactions have become available [11,12]. All these use cases carry the promise of improving quality and efficiency by offering second opinions [13] and reducing the documentation burden on strained health care professionals [12].

Emergency departments (EDs) frequently serve as the initial point of contact for patients in need of immediate medical attention. A crucial component of ED operations is the triage process, which aims to efficiently allocate often limited resources by prioritizing patients according to the severity and urgency of their conditions. This is typically done using established triage systems, such as the Manchester Triage System (MTS), which is popular in Europe and frequently used in Germany [14,15]. Taking place in a high-stress environment [16,17], triage processes have been shown to be highly variable in quality [18] and influenced by personal traits of the rater such as experience and triaging fatigue [19].

In light of the challenges faced by staff undertaking the triage process and the demonstrated medical abilities of language models, our study sought to assess the capability and potential of ChatGPT in the context of emergency triage. We evaluated its performance in triaging patient vignettes according to the MTS framework, comparing its results to those of both professional MTS raters and doctors working in an ED without triage training. Given the promising data of LLMs serving as a second opinion in other medical contexts [13], we explored ChatGPT's potential as a resource for providing external validation and second opinions to ED staff with less experience and without specific training in the MTS. We further compared ChatGPT and doctor triages to triage assessments of other currently available state-of-the-art LLMs such as Gemini 1.5 (Google) [20], Llama 3 70B (Meta) [21], and Mixtral 8x7b (Mixtral AI) [22]. Additionally, we tested the performance of the LLM GPT-4 (accessed via the OpenAI application programming interface), on which the product ChatGPT is based on. Our research is the first to compare rater expertise and different LLMs and ChatGPT versions while incorporating the MTS, building on findings of fair agreement between professional raters and Generative Pre-trained Transformer (GPT)–based models in smaller samples using a different triage system [23] and similarities observed between the triage decisions of ophthalmology trainees and GPT-4–based ChatGPT [24].

Methods

Case Vignette Creation and Triage Process

In this study, we compiled 124 independent emergency cases from a single, randomly selected day in the interdisciplinary ED of University Hospital Düsseldorf, Germany, which were then transformed into anonymized English case vignettes. These vignettes solely contained medically relevant information, extracted by 1 doctor (LM) according to a predefined standard operating procedure. Patient ages were randomly adjusted within a range of –2 and +2 years where they were not pertinent to the case. Nonmedical information, including personal demographics such as race, was excluded. Vital signs or information about the conscious state of the patient were added where available or deemed clinically necessary. Clinical values were slightly altered (up to 5% of the original value) or added where necessary for triage, with all adjustments made under the discretion of the overseeing doctor. Detailed information on laboratory test results or imaging were not included in case descriptions. The resulting case vignettes (see Textbox S1 in Multimedia Appendix 1) were then reviewed by a second doctor (MP) to ensure the absence of potentially identifying data, in accordance with an internal standard operating procedure. Both doctors were not involved in the subsequent rating of the cases. The cases were independently assessed by 2 MTS-instructing and experienced staff members. When differing triage priorities were assigned, a third equally qualified and MTS-trained doctor mediated the discussion and had a tiebreaking vote to reach a consensus rating (consensus set). Triage decisions were made using the fifth version of the MTS (German version) [25], which consists of the following categories: MTS Level 1 ("Red") for immediate assessment, MTS Level 2 ("Orange") for very urgent assessment within 10 minutes, MTS Level 3 ("Yellow") for urgent assessment within 30 minutes, MTS Level 4 ("Green") for standard assessment within 90 minutes, and MTS Level 5 ("Blue") for nonurgent assessment within 120 minutes. The distribution of the determined triage levels for the analyzed day were compared to long-term averages of the ED.

XSL•FO
RenderX

Additionally, 4 MTS-untrained resident doctors regularly working in the ED were given these cases to assess according to MTS triage categories, albeit without access to the precise algorithm diagrams underlying the MTS. These 4 doctors all had no formal MTS training and worked in the ED regularly throughout their residency, with 2 in the second year of residency and 2 in the third year of residency. We also presented the prototypical anonymized case vignettes to 2 versions of ChatGPT, based on either GPT-3.5 or GPT-4 (ChatGPT version: May 24, 2023). The prompt given to ChatGPT was derived after manually testing several prompt versions, with the most promising one selected for the final evaluation in a zero-shot setting. The models were tasked with stratifying the cases according to the MTS using this optimized prompt without further training or refinement (see Textbox S2 in Multimedia Appendix 1) and without access to the copyright-protected MTS algorithm diagrams. Both versions of ChatGPT were queried identically 4 times with new chats for each iteration, which was done due to the probabilistic nature of LLMs [26]. Afterward, the untrained doctors were presented with the answers (rating and explanation) from the overall best-performing ChatGPT instance as a second opinion and were asked to reconsider their initial answers (see Textbox S3 in Multimedia Appendix 1). A "hypothetical best" performance in this scenario was defined as the optimal integration of assistant doctors' decisions with ChatGPT responses, representing the maximum possible improvement achievable with ChatGPT input. For comparison, a variety of state-of-the-art LLMs (GPT-4, Llama 3 70B, Gemini 1.5, and Mixtral 8x7b) were similarly queried 4 times via respective application programming interfaces in a zero-shot approach using a slightly adapted prompt (see Textbox S2 in Multimedia Appendix 1). Details regarding the models and parameters are described in Table S1 in Multimedia Appendix 1. A flowchart of the study's setup can be found in Figure 1.

Figure 1. Summarizing flowchart of the study setup. The displayed flowchart summarizes the methodological approach and gives an overview of emergency department (ED) case vignette creation and triage performance assessment. Triage was carried out according to the Manchester Triage System (MTS) by doctors of different training levels and different Generative Pre-trained Transformer (GPT)–based versions of ChatGPT, as well as other large language models (LLMs).

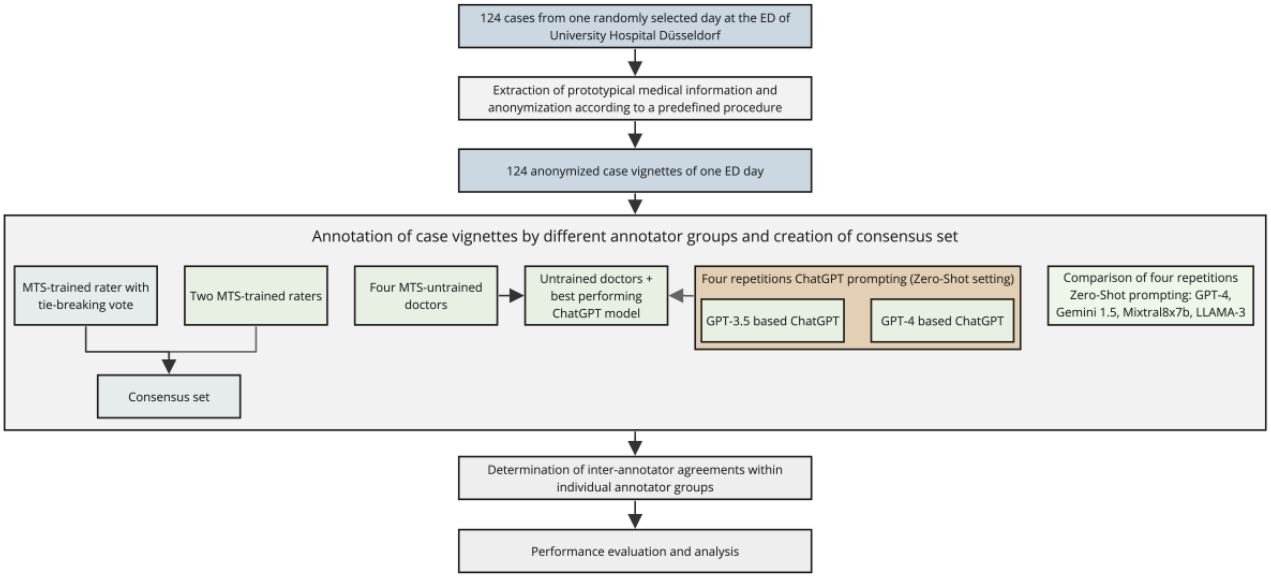

Calculation of Interrater Agreement

For calculating the interrater agreement, we computed the quadratic-weighted Cohen κ for each rater against the consensus set and used qualitative categories according to the original work by Cohen [27]. To test for statistical differences between the κ values of different rating groups, we performed a 1-way ANOVA with a Bonferroni correction, followed by the Tukey honest significant difference test. Mean values are displayed with SDs throughout the paper. All calculations were executed using Python (version 3.8; Python Software Foundation) employing the *statsmodel* package (version 0.13.2) [28]. For designing graphical representations, we used the *matplotlib* (version 3.1.3) [29] and *seaborn* (version 0.12.0) [30] packages.

Ethical Considerations

The local ethics committee of University Hospital Düsseldorf (5794R and 2020-1263) granted approval for data collection. Given the retrospective nature of this study, the requirement for written informed consent was waived. Data were anonymized as explained above and no compensation was provided.

Results

The consensus triage distribution in our analyzed exemplary cases was similar to that of the triage from the same ED in 2022 or published 2019 data [31]. However, there was a higher proportion of patients triaged as "Blue" (9/124, 7.3% vs 4.5% of all cases in 2022), "Yellow" (52/124, 41.9% vs 34.4%), "Orange" (24/124, 19.4% vs 9.5%), and "Red" (5/124, 4% vs 3%). Conversely, fewer patients were triaged as "Green" (34/124, 27.4% vs 48.7%).

As observed in the study, the different versions of ChatGPT and other LLMs were generally capable of providing triage colors as requested using the prompt and consequently assigned an MTS color for each case vignette (see Figures S1 and S2 in Multimedia Appendix 1 for a graphical depiction of all ratings

XSL•FO
RenderX

of all raters). We found moderate to substantial agreement between the consensus triage and the triages made by various groups, including LLMs, different ChatGPT iterations, and untrained doctors. As might be anticipated, the professionally trained raters individually exhibited near-perfect alignment with the consensus set they helped construct before its formation, reflected by an average quadratic-weighted κ of 0.91 (SD 0.054). The interrater agreement of the professional raters (quadratic-weighted κ of 0.82) was identical to that reported in the literature [32]. On the other hand, GPT-3.5–based ChatGPT demonstrated only moderate agreement with the consensus set (κ=mean 0.54, SD 0.024), a score that was significantly lower than those of both GPT-4–based ChatGPT (κ=mean 0.67, SD 0.037; $P<.001$) and untrained doctors (κ=mean 0.68, SD 0.056; $P<.001$). The latter 2 groups displayed substantial agreement with the consensus triage, with no significant difference discernible between them ($P>.99$). When the untrained doctors were given GPT-4–based ChatGPT responses as a second opinion, they achieved a slightly higher, although not statistically significant, average κ of 0.70 (SD 0.047; $P=.97$). Despite this improvement, their κ scores still fell short of the hypothetical best combinations of the GPT-4–based ChatGPT responses and the untrained doctors' initial assessments. Such combinations could have led to near-perfect agreement (κ=mean 0.84, SD 0.021)—a level of performance more comparable to that of professional raters. The other tested LLMs showed varying results, with raw GPT-4 performing very similarly to the GPT-4–based ChatGPT (κ=mean 0.65, SD 0.010; $P=.98$). Gemini 1.5 achieved results that fell between GPT-4 and the GPT-3.5–based ChatGPT with a mean κ of 0.60 (SD 0.010), which were not significantly different from both models ($P=.08$ and $P=.29$, respectively) but significantly worse than untrained doctors ($P=.03$). Llama 3 70B achieved a similar albeit slightly lower average κ of 0.52 (SD 0.004) than the GPT-3.5–based ChatGPT ($P=.98$), which was worse than Gemini 1.5 ($P=.04$) but better than the Mixtral 8x7b model, which had a mean κ of 0.42 (SD 0.000; $P=.006$; see Figure 2; for the respective tables and P values [ANOVA: $F_8=54.01$; $P<.001$], see Table S2 in Multimedia Appendix 1). The κ values of individual raters are shown in Figure S3 in Multimedia Appendix 1.

In addition to the overall agreement with the consensus set, the distribution of assigned MTS levels varied considerably among the different rater groups. In comparison to professional raters and the consensus set, ChatGPT and other LLMs except for Gemini 1.5 frequently assigned the highest triage category "Red" (mean 18.1%, SD 0.5% GPT-3.5–based ChatGPT; mean 12.5%, SD 2.5% GPT-4–based ChatGPT; mean 22.5%, SD 0.7% Llama 3 70B; mean 37.1%, SD 0% Mixtral 8x7b; 5/124, 4% consensus set), yet they seldomly or never designated the lowest category "Blue" (only once across 4 batches with 124 questions [mean 0.2%] for GPT-3.5–based ChatGPT; mean 1.2%, SD 0.5% GPT-4–based ChatGPT; 9/124, 7.3% consensus set; Llama 3 70B, raw GPT-4, and Mixtral 7x8b never assigned the category). As shown, this pattern was more evident in GPT-3.5–based ChatGPT than in version 4 and less evident in Gemini 1.5 (mean 5.6%, SD 0% "Red" and mean 4%, SD 0% "Blue"). In contrast, untrained doctors assigned the "Blue" category much more frequently than their professionally trained counterparts (mean 17.3%, SD 8.4% untrained doctors vs 9/124, 7.3% consensus set), while also employing the "Red" category more often (mean 9.1%, SD 5.5% untrained doctors vs 5/124, 4% consensus set). Certain language models displayed notable triaging patterns, with Gemini 1.5 predominantly categorizing cases as "Orange" (mean 64.1%, SD1.1% Gemini 1.5 vs 24/124, 19.4% consensus set), while Mixtral 7x8b frequently selected "Red" (mean 37.1%, SD 0%) and "Green" (mean 48.4%, SD 0%) and yielded equal results over all 4 iterations (see Figure 3). Accordingly, tendencies toward overtriage and undertriage were evident, with the GPT models leaning toward overtriage, untrained doctors more often undertriaging, and the hybrid models displaying a fairly balanced performance (see Figure S4 in Multimedia Appendix 1). Among untrained doctors, variance in triage strategies was noticeable, as demonstrated by the composition of individual raters' triages (see Figure S5 in Multimedia Appendix 1). Importantly, all raters and LLMs, except for 2 instances where untrained doctors rated a "Red" case as "Orange," accurately identified the most critical "Red" cases in the consensus set as "Red." Notably, 1 doctor corrected their initial misjudgment to "Red" after being provided the GPT assessment.

XSL•FO
RenderX

Figure 2. Quadratic-weighted Cohen κ compared to the consensus set. (A) This panel shows box plots of the quadratic-weighted Cohen κ for various triaging groups in relation to the consensus triage (blue). For the untrained doctors' group, both scenarios are depicted: rating alone (initial triage in blue) and rating with GPT-4 as a second opinion (orange). The green box plot represents the potential best combination of the doctors' initial triage with the GPT-4–based ChatGPT second opinion. The box plot's center line indicates the median, box limits indicate upper and lower quartiles, and whiskers indicate 1.5x the IQR. The black point indicates an outlier. (B) This panel illustrates the results of the Tukey honest significant difference test, performed based on an ANOVA with Bonferroni correction of the quadratic-weighted Cohen κ values for the different rater groups. The universal CIs of each group's mean were computed based on Tukey Q and plotted, with nonoverlapping intervals indicating statistically significant differences. As the professional raters' ratings were taken into account for the consensus set, their performance serves primarily as the reference point. API: application programming interface.

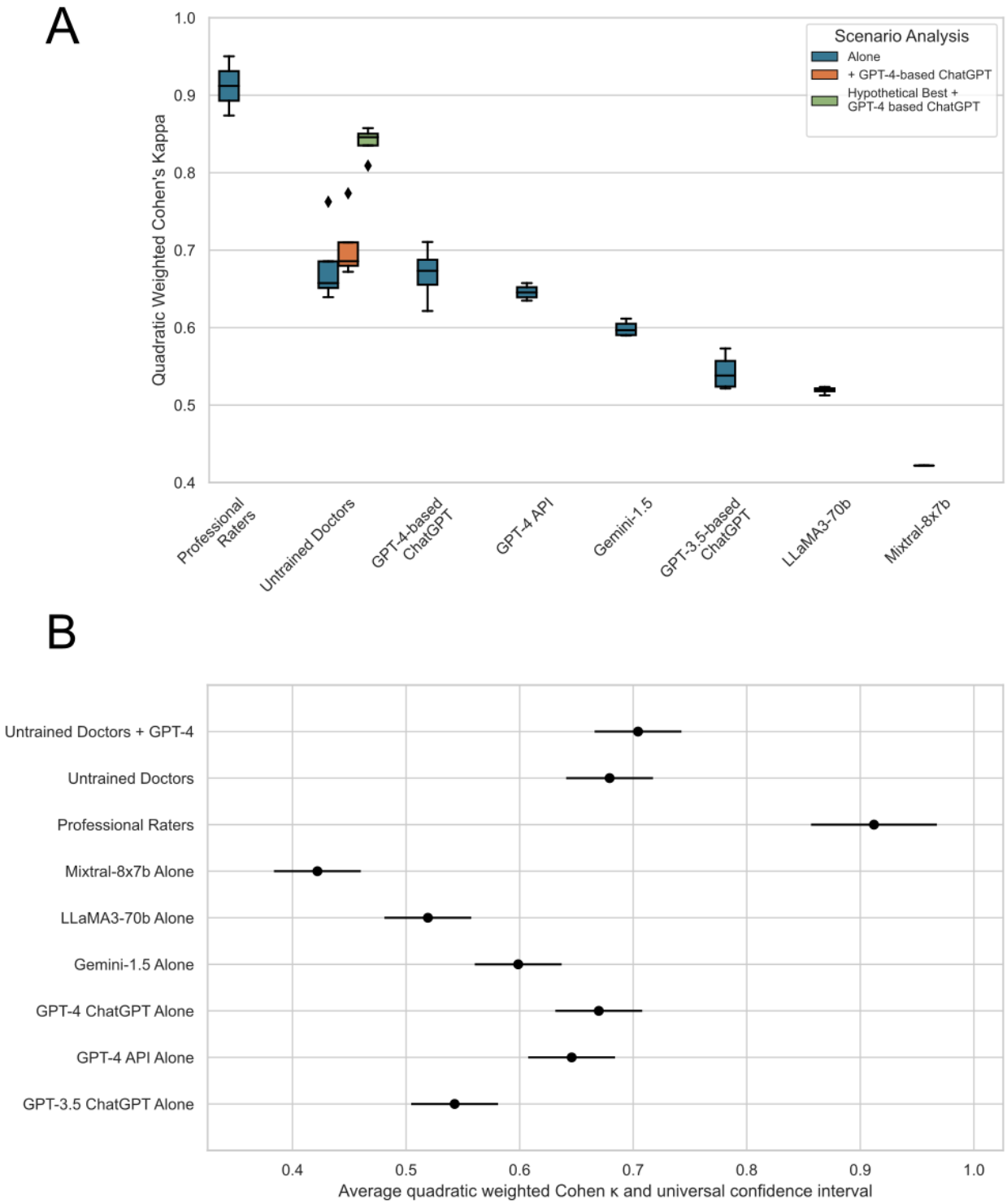

XSL•FO
RenderX

Figure 3. Average distribution of Manchester Triage System (MTS) levels among different rater groups. The figure presents stacked bar plots illustrating the average count of triage categories, arranged in ascending order of severity. The least severe level, Level 5 ("Blue"), is located at the bottom, while the most severe level, Level 1 ("Red"), is positioned at the top. API: application programming interface.

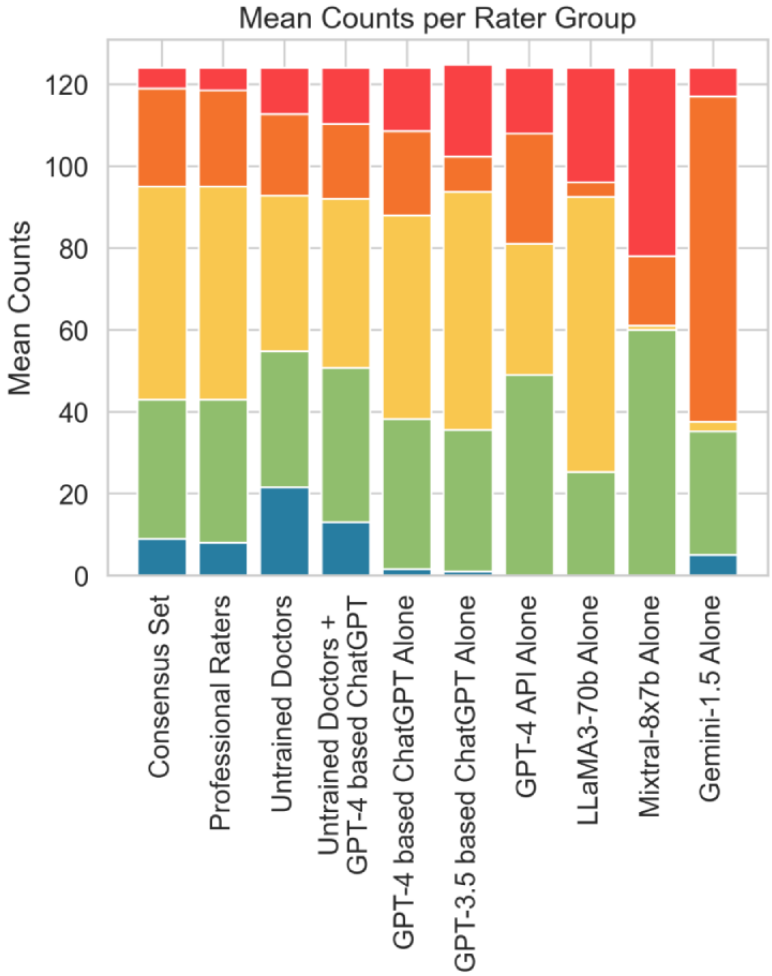

Discussion

With the rising number of acute cases in the ED and ongoing limitations of personnel resources [16,17], strategies for improving triage are urgently needed to increase the capacity, motivation, and opportunities of nurses and doctors. Past attempts to decrease workload, such as patient self-triage, have shown limited correlation when compared to triages conducted by nurses trained in the MTS [33]. However, recent results from algorithmic solutions demonstrate their potential in enhancing clinical decision processes and optimizing resource utilization [34].

Given the swift advancements in natural language processing made by LLMs and their extensive applications in medical care, it is plausible that such models could further enhance ED operations in the near future. It is worth noting that in our scenario, no clarifying questions or similar interactions were allowed, which may not mirror the typical setting of an ED. A test of the consumer-facing, general LLM product ChatGPT appeared especially relevant, as many health care professionals have access to and experimented with this product. Our study findings suggest that even a general-purpose system such as ChatGPT can match the performance of untrained doctors when triaging emergency cases and, similar to previous models [35], consistently identify the most severe cases. Similar on-par performance between GPT-4–based ChatGPT and ophthalmology trainees has been reported in another medical context, further underlining this finding [24]. However, the findings clearly highlight that the tested versions of ChatGPT did not demonstrate gold-standard performance in triage, especially in more ambiguous cases, which is also in line with previous findings from a study from Turkey [23]. This study found varying results for other LLMs, with the performance of Gemini 1.5, Llama 3 70B, and Mixtral 8x7b being similar to or worse than the GPT-4–based ChatGPT or GPT-4 itself. Importantly, this research employed a zero-shot approach, suggesting that parameter tuning could not only enhance results but also address occasional irregular triaging behavior, which warrants sophisticated testing in future studies.

The integration of LLMs in ED triage is promising not only in terms of decision accuracy but also in potentially boosting efficiency and saving time, a critical factor in emergency scenarios. However, it should be pointed out that this study does not empirically measure their efficiency against traditional triage methods. Future studies are needed to rigorously evaluate the time and effort implications of integrating LLMs and AI assistants into clinical settings, particularly in terms of data entry requirements and comparison with established triage

XSL•FO
RenderX

practices by trained medical staff. Such evaluations should incorporate interdisciplinary approaches including nursing staff, who are often responsible for triage in the ED.

Despite the considerable potential of using such a system as a second-opinion tool for clinicians—as demonstrated in our study—the minimal actual improvement observed indicates an ongoing need to train medical personnel in more efficient usage of these resources [36]. While 1 untrained rater corrected an undertriage to the most critical color "Red" after seeing the ChatGPT ratings—a change that could be clinically significant—further studies are needed to explore potential quality improvements with better validated systems and professionally trained triage raters.

The substantial progress between ChatGPT using GPT-3.5 and GPT-4—performing comparably to untrained doctors—hints at further potential advancements in this area. Since a very basic prompting strategy without background on the copyright-protected MTS or strategies such as retrieval-augmented generation was used, there appears to be ample room for further technical improvements. Given the option to fine-tune and further train LLMs on specific data, such as triage cases, it is well conceivable that they might even yield better results that are on par with professional raters in the future and take a wider range of parameters into account than humans are able to. Furthermore, these systems could help decrease the workload of qualified medical staff by answering basic inquiries, collecting medical information, and offering further differential diagnoses as previously shown [37]. When combined with the recent advancements in automatic patient decompensation detection using continuous sensor data in EDs [38], these technologies could contribute to an efficient and human-centered ED experience for both staff and patients. Moreover, these systems might, in the future, assist patients in deciding between primary care and ED presentation. However, such a prediction warrants further investigation.

It is crucial to note that this study is purely exploratory, and at present, no clinical decisions should be based solely on recommendations made by an LLM or ChatGPT without proper validation studies [39] in place. This is further underscored by the LLMs' and ChatGPT's performance deficiencies in this study, which would likely lead to suboptimal triage if used in a real-world setting. While this research emphasizes the potential and rapid progress of such models in health care and EDs, the authors want to highlight the ongoing debate about the appropriate regulation of these models for medical applications and how data privacy concerns should be addressed [40,41]. Notwithstanding these challenges, as researchers diligently navigate and promote responsible innovation, the potential of AI to revolutionize health care and augment patient outcomes perseveres as a captivating opportunity to contribute to the amelioration of our health care system. Future validation studies should further prioritize large, representative, and multicentric data sets, which are paramount for the correct assessment of AI tools. In summary, despite rapid advancements in LLM technologies and associated products such as ChatGPT, this study confirms that in a very basic setup, they currently do not meet the gold standard for ED triage, underscoring the urgent need for further development and rigorous validation.

Data Availability

The data assessed in this study are available from the corresponding author on reasonable request.

Conflicts of Interest

All authors reports no conflicts of interest related to this study. LS, AS, TK, RJ, MM, MB, and LB also report no conflicts of interest outside the scope of this study. LM received honoraria for lecturing, consulting, and travel expenses for attending meetings from Biogen, Merck Serono, and Novartis, all outside the scope of this work. His research is funded by the by the German Multiple Sclerosis Foundation (DMSG) and the German Research Foundation (DFG). NH reports travel reimbursements from Alexion and meeting attendance fees from Novartis, outside the scope of this work. SGM received honoraria for lecturing and travel expenses for attending meetings from Almirall, Amicus Therapeutics Germany, Bayer Health Care, Biogen, Celgene, Diamed, Genzyme, MedDay Pharmaceuticals, Merck Serono, Novartis, Novo Nordisk, ONO Pharma, Roche, Sanofi-Aventis, Chugai Pharma, QuintilesIMS, and Teva. His research is funded by the German Ministry for Education and Research (BMBF), Bundesinstitut für Risikobewertung (BfR), Deutsche Forschungsgemeinschaft (DFG), Else Kröner Fresenius Foundation, Gemeinsamer Bundesausschuss (G-BA), German Academic Exchange Service, Hertie Foundation, Interdisciplinary Center for Clinical Studies (IZKF) Muenster, German Foundation Neurology, Alexion, Almirall, Amicus Therapeutics Germany, Biogen, Diamed, Fresenius Medical Care, Genzyme, HERZ Burgdorf, Merck Serono, Novartis, ONO Pharma, Roche, and Teva, all outside the scope of this study. MP received honoraria for lecturing and travel expenses for attending meetings from Alexion, ArgenX, Bayer Health Care, Biogen, Hexal, Merck Serono, Novartis, Sanofi-Aventis, and Teva. His research is funded by ArgenX, Biogen, Hexal, and Novartis, all outside the scope of this study.

Multimedia Appendix 1
Supplementary material.
[PDF File (Adobe PDF File), 1842 KB - jmir_v26i1e53297_app1.pdf]

References

XSL•FO
RenderX

1. Al-Zaiti SS, Martin-Gill C, Zègre-Hemsey JK, Bouzid Z, Faramand Z, Alrawashdeh MO, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. Nat Med 2023 Jul;29(7):1804-1813 [FREE Full text] [doi: 10.1038/s41591-023-02396-3] [Medline: 37386246]
2. Smak Gregoor AM, Sangers TE, Bakker LJ, Hollestein L, Uyl-de Groot CA, Nijsten T, et al. An artificial intelligence based app for skin cancer detection evaluated in a population based setting. NPJ Digit Med 2023 May 20;6(1):90 [FREE Full text] [doi: 10.1038/s41746-023-00831-w] [Medline: 37210466]
3. Shin HJ, Han K, Ryu L, Kim E. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. NPJ Digit Med 2023 Apr 29;6(1):82 [FREE Full text] [doi: 10.1038/s41746-023-00829-4] [Medline: 37120423]
4. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/blog/chatgpt [accessed 2023-09-26]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Arztebl Int 2023 May 30;120(21):373-374 [FREE Full text] [doi: 10.3238/arztebl.m2023.0113] [Medline: 37530052]
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
8. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar 5;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
10. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75 [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
11. Nabla. URL: https://www.nabla.com/ [accessed 2023-09-29]
12. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024 Feb 21;5(3). [doi: 10.1056/cat.23.0404]
13. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. arXiv Preprint posted online on November 30, 2023. [doi: 10.48550/arXiv.2312.00164]
14. Gräff I, Goldschmidt B, Glien P, Bogdanow M, Fimmers R, Hoeft A, et al. The German version of the Manchester Triage System and its quality criteria--first assessment of validity and reliability. PLoS One 2014 Feb 24;9(2):e88995 [FREE Full text] [doi: 10.1371/journal.pone.0088995] [Medline: 24586477]
15. Gräff I, Latzel B, Glien P, Fimmers R, Dolscheid-Pommerich RC. Validity of the Manchester Triage System in emergency patients receiving life-saving intervention or acute medical treatment-a prospective observational study in the emergency department. J Eval Clin Pract 2019 Jun 25;25(3):398-403. [doi: 10.1111/jep.13030] [Medline: 30216602]
16. Healy S, Tyrrell M. Stress in emergency departments: experiences of nurses and doctors. Emerg Nurse 2011 Jul;19(4):31-37. [doi: 10.7748/en2011.07.19.4.31.c8611] [Medline: 21877616]
17. García-Tudela Á, Simonelli-Muñoz AJ, Rivera-Caravaca JM, Fortea MI, Simón-Sánchez L, González-Moro MTR, et al. Stress in emergency healthcare professionals: the Stress Factors and Manifestations Scale. Int J Environ Res Public Health 2022 Apr 05;19(7):4342 [FREE Full text] [doi: 10.3390/ijerph19074342] [Medline: 35410024]
18. Zachariasse JM, Seiger N, Rood PPM, Alves CF, Freitas P, Smit FJ, et al. Validity of the Manchester Triage System in emergency care: a prospective observational study. PLoS One 2017 Feb 2;12(2):e0170811 [FREE Full text] [doi: 10.1371/journal.pone.0170811] [Medline: 28151987]
19. Fekonja Z, Kmetec S, Fekonja U, Mlinar Reljić N, Pajnkihar M, Strnad M. Factors contributing to patient safety during triage process in the emergency department: a systematic review. J Clin Nurs 2023 Sep;32(17-18):5461-5477. [doi: 10.1111/jocn.16622] [Medline: 36653922]
20. Gemini 1.5. Google. URL: https://deepmind.google/technologies/gemini/ [accessed 2024-06-10]
21. Llama 3. Meta. URL: https://llama.meta.com/llama3/ [accessed 2023-06-10]
22. Mixtral AI. URL: https://mistral.ai/news/mixtral-of-experts/ [accessed 2024-06-10]
23. Sarbay İ, Berikol GB, Özturan İ. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023 Jun 26;23(3):156-161 [FREE Full text] [doi: 10.4103/tjem.tjem_79_23] [Medline: 37529789]
24. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol 2023 Aug 09:S0008-4182(23)00234-X. [doi: 10.1016/j.jcjo.2023.07.016] [Medline: 37572695]
25. Mackway-Jones K, Marsden J, Windle J, Krey J, Moecke H, editors. Ersteinschätzung in der Notaufnahme - Das Manchester-Triage-System. 5th ed. Bern, Switzerland: Hogrefe; 2020.

XSL•FO
RenderX

26. Perlis RH, Fihn SD. Evaluating the application of large language models in clinical research contexts. JAMA Netw Open 2023 Oct 02;6(10):e2335924 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.35924] [Medline: 37782501]
27. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
28. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference (SciPy 2010). Austin, Texas: SciPy; Jun 28, 2010:92-96.
29. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/mcse.2007.55]
30. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software 2021 Apr 6;6(60):3021. [doi: 10.21105/joss.03021]
31. Michael M, Al Agha S, Böhm L, Bosse HM, Pohle AN, Schürmann J, et al. Alters- und geschlechtsbezogene Verteilung von Zuführung, Ersteinschätzung, Entlassart und Verweildauer in der zentralen Notaufnahme [Article in German]. Notfall Rettungsmed 2023 Feb;26(1):39-48. [doi: 10.1007/s10049-021-00895-9]
32. Olofsson P, Gellerstedt M, Carlström ED. Manchester Triage in Sweden - interrater reliability and accuracy. Int Emerg Nurs 2009 Jul;17(3):143-148. [doi: 10.1016/j.ienj.2008.11.008] [Medline: 19577200]
33. Dickson SJ, Dewar C, Richardson A, Hunter A, Searle S, Hodgson LE. Agreement and validity of electronic patient self-triage (eTriage) with nurse triage in two UK emergency departments: a retrospective study. Eur J Emerg Med 2022 Feb 01;29(1):49-55. [doi: 10.1097/MEJ.0000000000000863] [Medline: 34545027]
34. Boonstra A, Laven M. Influence of artificial intelligence on the work design of emergency department clinicians a systematic literature review. BMC Health Serv Res 2022 May 18;22(1):669 [FREE Full text] [doi: 10.1186/s12913-022-08070-7] [Medline: 35585603]
35. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Feb 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
36. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
37. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by Generative Pretrained Transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
38. Sundrani S, Chen J, Jin BT, Abad ZSH, Rajpurkar P, Kim D. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. NPJ Digit Med 2023 Apr 04;6(1):60 [FREE Full text] [doi: 10.1038/s41746-023-00803-0] [Medline: 37016152]
39. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med 2020 Apr 14;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
40. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023 Jul 25;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
41. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]

Abbreviations

AI: artificial intelligence
GPT: Generative Pre-trained Transformer
ED: emergency department
LLM: large language model
MTS: Manchester Triage System

Edited by B Puladi; submitted 02.10.23; peer-reviewed by A Sarraju, A Follmann, M Moebus, Q Jin, L Zhu; comments to author 29.02.24; revised version received 17.04.24; accepted 14.05.24; published 14.06.24.

Please cite as:
Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, Mehsin M, Bernhard M, Meuth SG, Böhm L, Pawlitzki M Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study J Med Internet Res 2024;26:e53297
URL: https://www.jmir.org/2024/1/e53297
doi:10.2196/53297
PMID:38875696

XSL•FO
RenderX

©Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G Meuth, Lennert Böhm, Marc Pawlitzki. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Reproducibility of the Structured Abstracts Generated by ChatGPT and Bard Compared to Human-Written Abstracts in the Field of Spine Surgery: Comparative Analysis

Hong Jin Kim[1*], MD, PhD; Jae Hyuk Yang[2*], MD, PhD; Dong-Gune Chang[1], MD, PhD; Lawrence G Lenke[3], MD; Javier Pizones[4], MD, PhD; René Castelein[5], MD, PhD; Kota Watanabe[6], MD, PhD; Per D Trobisch[7], MD; Gregory M Mundis Jr[8], MD; Seung Woo Suh[9], MD, PhD; Se-Il Suk[1], MD, PhD

[1]Department of Orthopedic Surgery, Inje University Sanggye Paik Hospital, College of Medicine, Inje University, Seoul, Republic of Korea
[2]Department of Orthopedic Surgery, Korea University Anam Hospital, College of Medicine, Korea University, Seoul, Republic of Korea
[3]Department of Orthopedic Surgery, The Daniel and Jane Och Spine Hospital, Columbia University, New York, NY, United States
[4]Department of Orthopedic Surgery, Hospital Universitario La Paz, Madrid, Spain
[5]Department of Orthopedic Surgery, University Medical Centre Utrecht, Utrecht, Netherlands
[6]Department of Orthopedic Surgery, Keio University School of Medicine, Tokyo, Japan
[7]Department of Spine Surgery, Eifelklinik St. Brigida, Simmerath, Germany
[8]Department of Orthopaedic Surgery, Scripps Clinic, La Jolla, CA, United States
[9]Department of Orthopedic Surgery, Korea University Guro Hospital, College of Medicine, Korea University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Dong-Gune Chang, MD, PhD
Department of Orthopedic Surgery
Inje University Sanggye Paik Hospital, College of Medicine
Inje University
1342, Dongil-Ro
Nowon-Gu
Seoul, 01757
Republic of Korea
Phone: 82 2 950 1284
Email: dgchangmd@gmail.com

Abstract

Background: Due to recent advances in artificial intelligence (AI), language model applications can generate logical text output that is difficult to distinguish from human writing. ChatGPT (OpenAI) and Bard (subsequently rebranded as “Gemini”; Google AI) were developed using distinct approaches, but little has been studied about the difference in their capability to generate the abstract. The use of AI to write scientific abstracts in the field of spine surgery is the center of much debate and controversy.

Objective: The objective of this study is to assess the reproducibility of the structured abstracts generated by ChatGPT and Bard compared to human-written abstracts in the field of spine surgery.

Methods: In total, 60 abstracts dealing with spine sections were randomly selected from 7 reputable journals and used as ChatGPT and Bard input statements to generate abstracts based on supplied paper titles. A total of 174 abstracts, divided into human-written abstracts, ChatGPT-generated abstracts, and Bard-generated abstracts, were evaluated for compliance with the structured format of journal guidelines and consistency of content. The likelihood of plagiarism and AI output was assessed using the iThenticate and ZeroGPT programs, respectively. A total of 8 reviewers in the spinal field evaluated 30 randomly extracted abstracts to determine whether they were produced by AI or human authors.

Results: The proportion of abstracts that met journal formatting guidelines was greater among ChatGPT abstracts (34/60, 56.6%) compared with those generated by Bard (6/54, 11.1%; $P<.001$). However, a higher proportion of Bard abstracts (49/54, 90.7%) had word counts that met journal guidelines compared with ChatGPT abstracts (30/60, 50%; $P<.001$). The similarity index was significantly lower among ChatGPT-generated abstracts (20.7%) compared with Bard-generated abstracts (32.1%; $P<.001$). The AI-detection program predicted that 21.7% (13/60) of the human group, 63.3% (38/60) of the ChatGPT group, and 87% (47/54)

XSL•FO
RenderX

of the Bard group were possibly generated by AI, with an area under the curve value of 0.863 (P<.001). The mean detection rate by human reviewers was 53.8% (SD 11.2%), achieving a sensitivity of 56.3% and a specificity of 48.4%. A total of 56.3% (63/112) of the actual human-written abstracts and 55.9% (62/128) of AI-generated abstracts were recognized as human-written and AI-generated by human reviewers, respectively.

Conclusions: Both ChatGPT and Bard can be used to help write abstracts, but most AI-generated abstracts are currently considered unethical due to high plagiarism and AI-detection rates. ChatGPT-generated abstracts appear to be superior to Bard-generated abstracts in meeting journal formatting guidelines. Because humans are unable to accurately distinguish abstracts written by humans from those produced by AI programs, it is crucial to exercise special caution and examine the ethical boundaries of using AI programs, including ChatGPT and Bard.

(J Med Internet Res 2024;26:e52001) doi:10.2196/52001

KEYWORDS

artificial intelligence; AI; ChatGPT; Bard; scientific abstract; orthopedic surgery; spine; journal guidelines; plagiarism; ethics; spine surgery; surgery; language model; chatbot; formatting guidelines; abstract

Introduction

Artificial intelligence (AI) language models are being applied to various fields, including medicine and health care [1-3]. Novel and open AI programs make it possible to generate structured text within seconds, accelerating the use of AI and providing valuable insights into clinical research [4,5]. One of the more successful applications of these AI programs is generating high-quality theses and providing answers to questions from the United States Medical Licensing Examination [3,6,7]. However, many concerns have been raised about the scientific value of AI-based tools, with ethical issues and reproducibility at the forefront of public debates [8,9].

AI language models are based on complex neural network transformer models known as large language models (LLMs) [10]. Pretraining with large-sized data is used to predict the optimal next elements of textual input. ChatGPT, released in November 2022 and based on GPT-3 software, is the first popular AI language model application. It generates fluent output and reinforces human feedback [5,10]. However, ChatGPT responses are based on information drawn from the internet before a data cut-off date of September 2021. Bard (subsequently rebranded Gemini), released in March 2023, is a new AI language model developed by Google that is based on the language model for dialogue applications family of LLMs. Unlike ChatGPT, it replies to prompts using real-time information in conjunction with Google Internet searches.

Although many proposals for the use of AI programs in scientific writing have been suggested, few relevant studies have been published. One study attempted to compare scientific abstracts generated by ChatGPT with those gathered from 50 human-written abstracts in the field of medicine [5]. However, no similar comparative analyses have involved more specialized subjects. Furthermore, the differences between ChatGPT and Bard in abstract generation have not yet been studied. This study aimed to evaluate the reproducibility of abstracts generated by ChatGPT and Bard compared with human-written abstracts in the field of spinal surgery.

Methods

Journal Selection and Abstract Extraction

To evaluate abstracts in the field of spinal surgery, specialists in spinal surgery with more than 10 years of experience responded to the query: “Please introduce the reputable journals in the area of spinal surgery in orthopedics and neurosurgery.” From the responses, we selected 7 journals: *Spine (Phila pa 1976)*, *The Spine Journal*, *European Spine Journal*, *Journal of Neurosurgery: Spine*, *Global Spine Journal*, *Neurosurgery*, and *The Journal of Bone and Joint Surgery*. We randomly extracted the 60 papers published by the 7 journals (10 from *Spine (Phila pa 1976)*, 10 from *The Spine Journal*, 10 from *European Spine Journal*, 10 from *Journal of Neurosurgery: Spine*, 10 from *Global Spine Journal*, 5 from *Neurosurgery*, and 5 from *The Journal of Bone and Joint Surgery*) to minimize the likelihood of any prior knowledge of the abstracts by the AI programs ChatGPT (May 24 version; OpenAI) and Bard (experiment version after May 15, 2023; Google AI; considering ChatGPT’s knowledge cut-off of September 2021) and by human reviewers.

Abstract Generation

The titles from the 60 randomly extracted abstracts were used in prompts presented to the AI programs as follows: “Please write a scientific abstract for the article [title] in the style of [journal] at [link].” However, Bard did not produce abstracts in 6 cases (1 from *Spine (Phila pa 1976)*, 1 from *The Spine Journal*, 1 from *European Spine Journal*, 1 from *Journal of Neurosurgery: Spine*, and 2 from *Global Spine Journal*) in response to this prompt, replying that “I’m just a language model, so I can’t help you with that,” “I can’t assist you with that, as I’m only a language model and don’t have the capacity to understand and respond,” and “I’m not programmed to assist with that.” A total of 114 AI-generated abstracts were presented, 60 from ChatGPT (GPT-3.5) and 54 from Bard (Multimedia Appendix 1). Regarding abstract generation, because ChatGPT was sensitive to changes in prompts, we chose the abstract generated in response to the first prompt. As Bard supplied 3 answers to each prompt, we selected the generated abstract that was most similar to the format specified by the journal guidelines.

XSL•FO
RenderX

Abstract Evaluation

The 60 human-written abstracts and 114 AI-generated abstracts were divided into 3 groups: human (n=60), ChatGPT (n=60), and Bard (n=54; Figure 1). For the AI-generated abstracts, we collected data to assess their reproducibility according to format compliance (binary data: yes or no), total word count, consistency of conclusion (binary data: yes or no), and size of cohort sample. We also assessed the similarity index using iThenticate, which is a widely used program. The similarity index of the human group of abstracts, which were published in journals, was nearly 100%. We compared the similarity indices of the ChatGPT group and the Bard group. Based on popular consensus, plagiarism was considered at a similarity index of 15% or higher. We evaluated the AI detection rate in the 3 groups using ZeroGPT (access date: June 5, 2023), which is a tool designed to detect whether texts are generated by an AI program. ZeroGPT provides both a percentage from 0% (human-written) to 100% (AI and GPT-generated) and 1 of 9 sentences: "Your text is Human written," "Your text is Most Likely Human written," "Your text is Most Likely Human written, may include parts generated by AI/GPT," "Your text is Likely Human written, may include parts generated by AI/GPT," "Your text contains mixed signals, with some parts generated by AI/GPT," "Your text is Likely generated by AI/GPT," "Your text is Most Likely AI/GPT generated," "Most of Your text is AI/GPT Generated," and "Your text is AI/GPT Generated." We evaluated the accuracy of the AI detection using the expression "Your text contains mixed signals, with some parts generated by AI/GPT" for all 3 groups. Last, we evaluated the accuracy of 8 blinded human reviewers in determining how each of the 30 abstracts randomly chosen from the 174 used in the study was written (human- or AI-generated), in the form of binary scores collected by electronic records using Google Forms. None of the blinded human reviewers were provided any information regarding the abstracts until the survey was complete.

Figure 1. Flowchart of the study. AI: artificial intelligence.

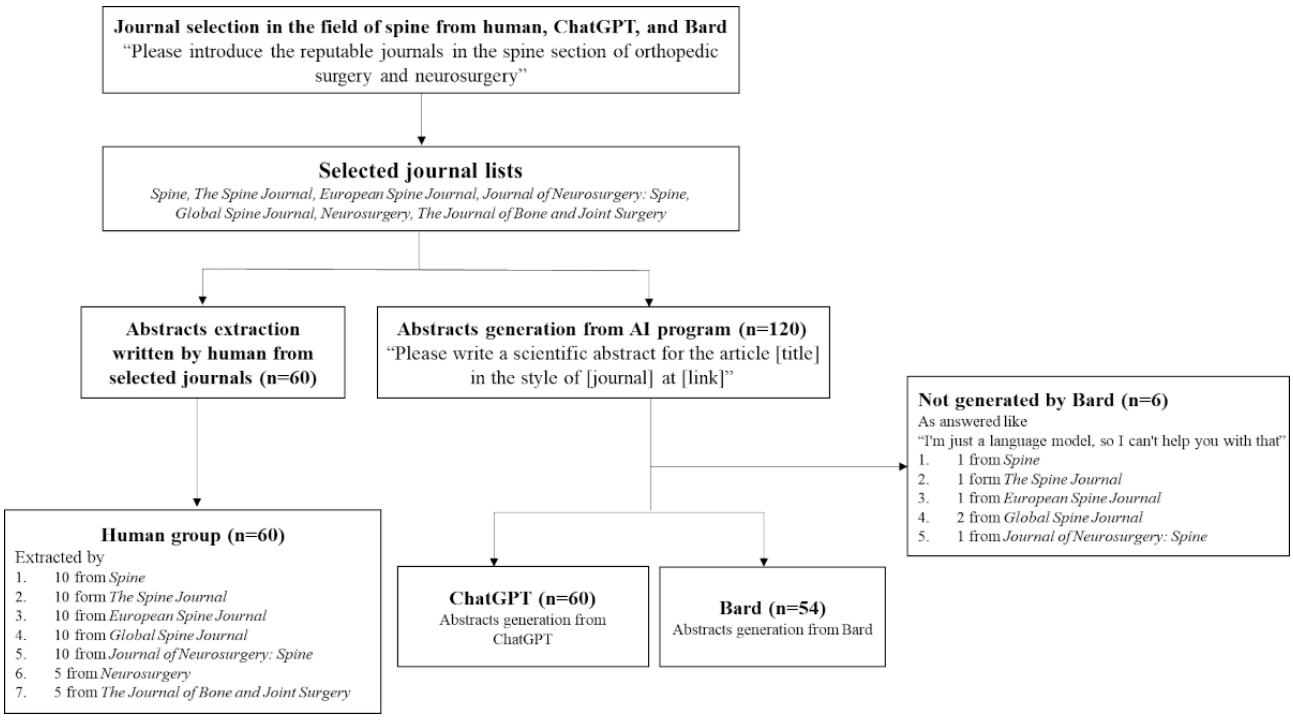

Statistical Analysis and Visualization

Statistical analysis and visualization were performed using R (version 4.3.0; The R Foundation). A normal distribution was confirmed with the Kolmogorov-Smirnov test. After confirming data homogeneity or heteroscedasticity, Student 2-tailed *t* test was used for continuous variables, and the chi-square test was used for categorical variables, as appropriate. A comparison of the 3 groups used a 1-way repeated-measures ANOVA and post hoc analyses using the Bonferroni test. A Pearson correlation analysis was used to assess the correlation of the cohort sample number between human-written and AI-generated abstracts, which were visualized using a heatmap. Receiver operative characteristic (ROC) curve analysis was performed to compare the AI detection rates for human-written and AI-generated abstracts. We calculated the *P* value in the ROC curve based on a null hypothesis with an area under the curve (AUC) of 0.5. The cut-off point in the ROC curve was measured using Youden's index. Statistical significance was set with a 2-tailed $P<.05$.

Ethical Considerations

This study was approved by the institutional review board of the Inje University Sanggye Paik Hospital (IRB NON2023-008), and informed written consent was waived from the participants for the electronic survey and publication of this study. The participants who are specialists with more than 10 years of experience in spine surgery were voluntarily recruited without any compensation, and the data from the electronic survey were collected in deidentified status.

XSL•FO
RenderX

Results

Similarity Index, AI Detection Rate, and Word Count

The mean similarity indices of the ChatGPT and Bard groups were 20.7% (SD 8.7%) and 32.1% (SD 11%), respectively, exceeding the 15% threshold for the commonly recognized standard of plagiarism (*P*<.001; 95% CI –15.05 to –7.62). The mean AI detection rates achieved by the human, ChatGPT, and Bard groups were 28.4% (SD 25.8%), 60.7% (SD 25%), and 77.7% (SD 21.1%), respectively, with significant differences (*P*<.001). All of the Bonferroni post hoc analyses for AI detection showed statistically significant differences (*P*<.001). For word counts in text, there was a significant difference between the 3 groups (*P*<.001), and the Bonferroni post hoc analysis showed significant differences, except between the human and ChatGPT groups (*P*>.99; Table 1).

Table 1. Baseline data in this study.

Variables	Human (n=60), mean (SD)	ChatGPT (n=60), mean (SD)	Bard (n=54), mean (SD)	*P* value	95% CI
Similarity index (%)	100 (0)	20.7 (8.7)	32.1 (11.0)	<.001	–15.05 to –7.62
AI[a] detection rate (%)	28.4 (25.8)	60.7 (25.0)	77.7 (21.1)	<.001[b]	N/A[c]
Word count (n)	317.8 (72.4)	317.8 (53.9)	223.8 (44.5)	<.001[d]	N/A

[a]AI: artificial intelligence.

[b]Bonferroni post hoc analysis results for the AI detection rate were as follows: human versus ChatGPT: *P*<.001; 95% CI –42.92 to –21.58; human versus Bard: *P*<.001; 95% CI –60.18 to –38.26; and ChatGPT versus Bard: *P*=.001; 95% CI –27.93 to –6.01.

[c]N/A: not applicable.

[d]For text number count, Bonferroni post hoc analysis showed the following: human versus ChatGPT: *P*>.99; 95% CI –25.89 to –25.29; human versus Bard: *P*<.001; 95% CI 67.41-120.49; and ChatGPT versus Bard: *P*<.001; 95% CI 67.46-120.54.

AI-Generated Abstract Formats and Content

With respect to abstracts that met the structured format requirements of the journals, 56.7% (34/60) of ChatGPT's abstracts complied with the journal guidelines, but only 11.1% (6/54) of Bard's abstracts matched the journal's requirements because Bard only produced abstracts in the form of "Background, Methods, Results, Conclusions," regardless of the journal specified in the prompt. For word count, the proportion deemed acceptable by the journals' guidelines was significantly higher among Bard-generated abstracts (49/54, 90.7%) than among the ChatGPT group (30/60, 50%; *P*<.001). For consistency of conclusions, no statistically significant differences were seen between the ChatGPT and Bard groups (*P*=.85). For cohort sample size, Pearson correlation analysis revealed strong and significant correlations between the human and ChatGPT groups (*r*=0.955; *P*<.001) and between the human and Bard groups (*r*=0.953; *P*<.001; Table 2).

Table 2. Assessment of the reproducibility of artificial intelligence–generated abstracts.

Variables	ChatGPT[a] (n=60)	Bard[b] (n=54)	*P* value
Structured abstract format of compliance with the journal's guideline (matched:unmatched)	34:26	6:48	<.001
The abstract word counts acceptable for journal's guideline (acceptable:unacceptable)	30:30	49:5	<.001
Consistency of conclusions (consistent:inconsistent)	39:21	36:18	.85

[a]Correlation of the cohort's sample size with human-written abstracts was analyzed by the Pearson correlation analysis: *r*=0.955; *P*<.001.

[b]Correlation of the cohort's sample size with human-written abstracts was analyzed by the Pearson correlation analysis: *r*=0.953; *P*<.001.

Plagiarism in AI-Generated Abstracts

A total of 106 of the 114 (84.2%) AI-generated abstracts met the criteria for plagiarism, with 94.4% (51/54) of the Bard-generated abstracts and 75% (45/60) of the ChatGPT abstracts meeting the threshold with statistical significance (*P*<.001; Table 3). Only 3 abstracts generated by Bard were not considered examples of plagiarism.

Table 3. Assessment of plagiarism in artificial intelligence–generated abstracts.

Variables	ChatGPT (n=60)	Bard (n=54)	*P* value
Plagiarism (+, similarity index≥15%)	45	51	<.001
Plagiarism (–, similarity index<15%)	15	3	<.001

Evaluation of Abstracts by AI Detection Programs

ZeroGPT incorrectly identified 13 of the 60 (21.7%) human-authored abstracts as AI-generated. It successfully detected 74.6% (85/114) of the AI-generated abstracts, but 36.7% (22/60) of the ChatGPT abstracts were not detected. There was a statistically significant difference between the 3 groups (*P*<.001; Table 4). For ROC curve analysis of human-written and AI-generated abstracts, the AUC was 0.863 (*P*<.001; 95% CI 0.806-0.920), indicating robust models, and the cut-off value was 52.5% for the AI detection rate in the

XSL•FO
RenderX

ZeroGPT, with 73.7% and 85% sensitivity and specificity, respectively (Figure 2). From the results of sentences presented by ZeroGPT, the AI programs successfully detected AI-generated abstracts with 74.6% and 78.3% sensitivity and specificity, respectively (Table 5).

Table 4. Assessment of AI[a] program detection for human-written and AI-generated abstracts.

Variables	Human (n=60)	ChatGPT (n=60)	Bard (n=54)	*P* value
AI detection (+, detection rate≥50%)	13	38	47	<.001
AI detection (–, detection rate<50%)	47	22	7	<.001

[a]AI: artificial intelligence.

Figure 2. Assessment of artificial intelligence program detection. Receiver operative characteristics analysis showed an area under the curve of 0.863 (*P*<.001; 95% CI 0.806-0.920) and a cut-off value of 52.5%, with 73.7% sensitivity and 85% specificity. AUC: area under the curve.

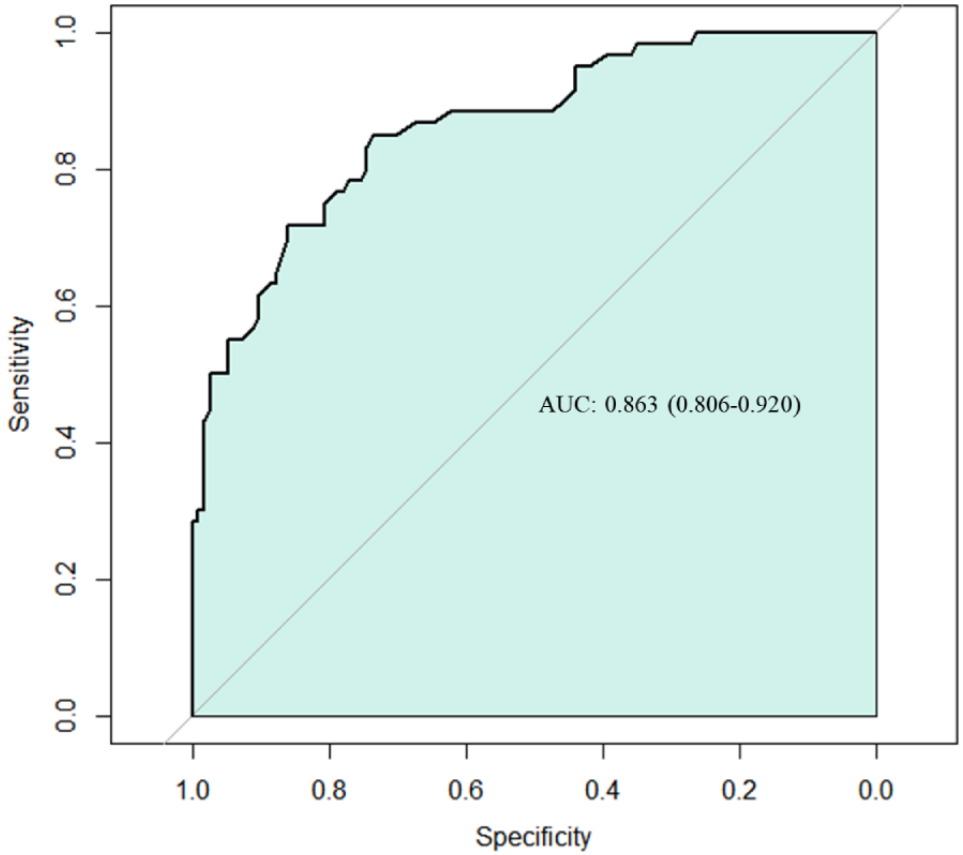

Table 5. AI[a] program detection of abstracts using ZeroGPT in this study[b].

Variables	Abstracts		Predictive value
	AI program[c]	Human[d]	
Detection from ZeroGPT			
AI program	85	13	PPV[e]=86.7%
Human	29	47	NPV[f]=61.8%

[a]AI: artificial intelligence.
[b]Sensitivity+specificity=154.9% (≥150%).
[c]Sensitivity=74.6%.
[d]Specificity=78.3%.
[e]PPV: positive predictive value.
[f]NPV: negative predictive value.

XSL•FO
RenderX

Evaluation of the Abstracts by Human Reviewers

Using 8 human reviewers who had specialized in the spine field with more than 10 years of experience and the role of journal reviewer, the mean detection rate by human reviewers was 53.8% (SD 11.2%). Among the actual human-written abstracts, 48.8% (positive predictive values) were recognized as human-written, and of the AI-generated abstracts, 55.9% were recognized as AI-generated. Detection by human reviewers achieved a sensitivity of 56.3% and a specificity of 48.4% (Table 6).

Table 6. Blinded human reviewers' detection of abstracts in this study[a].

Variables	Abstracts		Predictive value
	Human[b]	AI[c] program[d]	
Assessment on human reviewers			
Human	63	66	PPV[e]=48.8%
AI program	49	62	NPV[f]=55.9%

[a]The mean detection rate (%) for human-written abstracts=53.8 (SD 10.5), which were reviewed by 8 reviewers. Sensitivity+specificity=104.7% (<150%).

[b]Sensitivity=56.3%.

[c]AI: artificial intelligence.

[d]Specificity=48.4%.

[e]PPV: positive predictive value.

[f]NPV: negative predictive value.

Discussion

Principal Findings and Comparison With the Literature

Our findings demonstrated that both ChatGPT and Bard are capable of generating scientific abstracts from titles. Moreover, it is also noteworthy that the ability of LLM to produce abstracts has advanced to a level where it is challenging for humans to differentiate between AI-generated abstracts and human-written abstracts. The merits of writing a scientific paper using ChatGPT or Bard include the creation of a brief summary of complex research, rapid generation of suitable paragraphs, and visualization of important results in just a few seconds. Various trials of ChatGPT have been conducted, involving the writing of scientific papers, taking the United States Medical Licensing Examination, and expressing critical thinking [7,11]. However, these attempts have been met with concern by many experts, and the challenges posed by LLM-based AI have become the subject of social and ethical debates beyond the fields of medicine and science [2-4,12].

Controversies persist over whether AI-generated content itself has scientific value within ethical boundaries [5,9,12]. Debates weighing the advantages and limitations of AI language applications can be found in a variety of editorial forums [4,8-10]. However, to the best of our knowledge, few, if any, studies of the impact of ChatGPT on scientific writing have been conducted. Given the recent release date of Bard, no studies regarding Bard-generated scientific papers or comparing its abilities to those of ChatGPT of different LLMs are available. Gao et al [5] compared ChatGPT-generated abstracts with human-written scientific abstracts from 5 reputable journals, with an impact factor of over 87 (well quoted). However, little is known about the reproducibility of AI-generated abstracts in fields with a low citation index due to specialization, such as spinal surgery and neurosurgery [13-15].

We assessed the reproducibility of ChatGPT- and Bard-generated abstracts by comparing them with 60 human-written abstracts dealing with spine surgery. We also compared the differences between ChatGPT- and Bard-generated abstracts. Using iThenticate, ChatGPT- and Bard-generated abstracts had a mean similarity index greater than 15%, and a mean AI-detection rate greater than 60% reported by ZeroGPT indicates the current limitations of writing abstracts using only an AI program. The AI detection program did not identify all the AI-generated abstracts, achieving a sensitivity and specificity of 74.6% and 78.3%, respectively. However, these results indicate that AI programs have potential diagnostic value (sensitivity+specificity≥150%) to distinguish between human and AI authors. Contrary to the results achieved by the AI-detection program, humans who specialize in spine surgery were unable to distinguish between human-written and AI-generated abstracts (sensitivity+specificity<150%).

Our study revealed differences in the abstracts produced by ChatGPT and Bard, which use different LMMs [5,10,16]. ChatGPT is superior in creating structured abstracts that conform to journal guidelines but inferior in generating acceptable word counts. The contents, including consistency of conclusions and cohort sample sizes, of the ChatGPT-generated and Bard-generated abstracts did not differ significantly. One notable point in this study is related to 20% of hallucinations, which are AI's own power and plausible ability to tell a lie. Our data showed 33.6% of hallucinations in the consistency of conclusions in spine abstracts [2,9]. It seems to have come out with a higher hallucination rate because the contents of the spine are a specialized area. This indicates that humans cannot easily distinguish between AI outputs, but it is clear that the current LMMs face limitations in writing spine abstracts in terms of the completeness of the format.

ChatGPT generated all requested abstracts in response to our prompts, but Bard failed to do so in 6 cases. The titles of 2 of

XSL•FO
RenderX

the 6 papers were related to machine learning, and the other 4 involved low-prevalence diseases such as spinal tumors, a newly suggested scoring system, and surgical treatment. Our data suggest that Bard is less capable of generating scientific abstracts compared with ChatGPT in response to prompts referring to rare, poorly known, or new data. Nevertheless, given Bard's characteristics, the program is likely to be used to provide scientific abstracts from unlimited sources connected to Google, and a reassessment is essential [16-18].

ChatGPT is superior with respect to meeting format requirements compared with Bard. The LMM of ChatGPT is based on a data-feed system with an information cutoff of September 2021. As a pretrained system, ChatGPT was able to generate abstracts that complied with journal guidelines. The trend in word count was larger in ChatGPT than in Bard. Language model for dialogue applications, the language model on which Bard is based, provides real-time information from Google, making it possible to provide an accurate summary of up-to-date paper contents. Bard-generated abstracts were more concise, with relatively low word counts compared with ChatGPT-generated abstracts. This result indicates that Bard has significant advantages for summarizing content concisely. However, in these LMMs, it is important to recognize the primary difference between the 2 AI programs: ChatGPT provides data-based, biased output, while Bard provides web-based biased output, and this can be reflected in the generated abstracts. Although Bard generated the web-based output from Google, some studies suggest Bard's current limitations for accessing real-time data from Google [19,20]. For this controversial issue, our study also presented ChatGPT as superior in the formation of structured abstracts in compliance with the journal's guidelines, with no differences in the consistency of conclusions as beneficial works for access to search engines. Bard is currently an experimental version of LLMs that has the potential to develop further in the future.

The sources of information were real-time internet data from Google in Bard and pretrained data up to 2021 in ChatGPT [5]. Despite these differences, no statistically significant differences in the consistency of conclusions between ChatGPT and Bard were evident. Furthermore, both AI programs suggested cohort sample sizes that were similar to those of human-written abstracts, based on data analyzed with Pearson correlations. This suggests that cohort size was an important factor in making the AI-generated abstracts indistinguishable from human reviewers. Importantly, the version of ChatGPT in this study was GPT-3.5, but newly the launched version (GPT-4) was considered more reliable, creative, and able to handle much more nuanced instructions [21]. Our further analysis demonstrated that GPT-4 showed better improvements in compliance with word count (from 30/60, 50% to 49/60, 81.7%) and consistency of conclusions (from 39/60, 65% to 42/60, 70%) than GPT-3.5. Thus, the capability for generating medical abstracts in GPT-4 may be superior to GPT-3.5, but it should be demonstrated through new studies in the future.

Our data on plagiarism using iThenticate reflected the distinctive characteristics of the 2 AI programs. The output from ChatGPT was regenerated in context using data published through 2021. Bard generates output from real-time information from the internet, producing a similarity index and plagiarism scores that were significantly higher than those of ChatGPT. AI-detection software determined that 84.2% (96/114) of the AI-generated abstracts included plagiarism.

ZeroGPT's performance was relatively strong as measured by a ROC analysis (AUC=0.863; 95% CI 0.806-0.920). However, ZeroGPT had a limited detection value (73.7% sensitivity and 85% specificity in the ROC model) with respect to determining whether a spine section abstract was written by a human or AI. Similar results were produced for the expression of sentences (74.6% of sensitivity and 78.3% of specificity in sentence presentation). ZeroGPT also concluded (with sentences such as "Your text is AI/GPT Generated") that 21.7% (13/60) of human-written abstracts were AI-generated abstracts, indicating a limit to practical applications. Therefore, in the spine field, human-written abstracts need to be verified through other methods to distinguish them from AI-generated text.

Contrary to the AI-detection program, our findings from blinded human reviewers' detection of abstracts provided valuable insights into the writing of scientific abstracts. Because spine specialists were not able to effectively distinguish between human- and AI-generated abstracts, the use of AI programs may pose ethical challenges. Therefore, improvement of AI programs and AI detectors is needed to avoid ethical problems.

Májovský et al [22] described that LLMs can generate highly convincing fraudulent papers that mimic genuine scientific papers, from word usage to sentence structure. Our comprehensive analysis further confirms that LLMs' ability to generate abstracts has evolved to a level that even experts find difficult to distinguish from authentic work. This implies that using LLMs carries the potential risk of producing completely fake papers. Therefore, authors of scientific papers should carefully weigh the risks and benefits of using LLMs. Moreover, within the scientific community, there is growing pressure to overhaul the peer review and publishing processes [23]. Researchers, including Májovský et al [24], who study the ability of AI to create scientific papers, have suggested several strategies to reduce the risk of fraudulent papers: the provision of source data sets publicly, a meticulous review process, strict ethical regulations at the level of publishers and academic institutions, and penalties for researchers who commit ethical misconduct. Given our findings on the capabilities of LLMs, the need to establish ethical standards will become increasingly essential, both for researchers (such as the notation of references about the use of ChatGPT) and the scientific community (such as the meticulous peer review process).

Limitations

This study has several limitations. We evaluated a relatively small sample size and employed a few human reviewers. However, we collected as many abstracts as possible from representative journals and involved reviewers with vast experience in the relevant field. For the evaluation of human reviewers' assessments, human reviewers might have been familiar with real abstracts, and this could have biased the results as a potential limitation. In addition, recent advances in the LLMs on which both ChatGPT and Bard are based were not incorporated into this study. Although those 2 models are

XSL•FO
RenderX

currently in the development phase, our study provided important conclusions, including the reproducibility of scientific abstracts in the field of spinal surgery and the differences between ChatGPT and Bard. Our study was unable to capture the sensitive characteristics of the 2 representative AI models for the prompts. Because ChatGPT and Bard offer different outputs in response to slight changes in prompts, future evaluation of these findings is necessary. Last, the broad-scope prompt may affect the failure of the generation of 6 Bard-generated abstracts, despite the significant correlation between human- and AI-generated abstracts. Furthermore, the LLMs for generating abstracts are used after analyzing the main results of the studies. In other words, the main use of LLMs is to create abstracts based on the finished initial version of the manuscript. Instead of the title-based prompt or replacing the URLs with guideline instructions, alternative prompts may yield better results. Considering the importance of the concrete prompt before using LLMs, it should be addressed regarding the proper prompt for generating scientific abstracts in the future.

Conclusions

Both ChatGPT and Bard can be used to help write scientific abstracts, but most AI-generated abstracts are currently considered unethical products based on high plagiarism and AI-detection rates. In terms of meeting journal formatting requirements, ChatGPT-generated abstracts appear to be superior to their Bard-generated counterparts. Because human reviewers are often unable to distinguish human writing from AI products, the use of AI programs to write abstracts within ethical boundaries requires careful consideration and should be evaluated through various approaches.

Acknowledgments

This study was conducted with the support of research funding from Korea University and was authored with the assistance of the Scoliosis Research Institute affiliated with Korea University. Neither ChatGPT nor other generative language models were used for the ideation or writing process.

Data Availability

All data generated or analyzed during this study are included in this published paper (and in Multimedia Appendix 1).

Authors' Contributions

HJK, JHY, and DGC conceptualized the study. HJK contributed to clinical data acquisition and provision. HJK, JHY, DGC, LGL, JP, and RC provided essential resources, and substantially contributed to the interpretation of the data. HJK, JHY, SWS, and SIS developed methodology, acquired meteorological data, and performed data curation. HJK performed formal data analysis. HJK, KW, PDT, and SWS verified the underlying data. HJK carried out statistical validation. DGC, JP, RC, and GMMJ validated meteorological aspects. HJK drafted the first version and developed figures. HJK, JHY, and DGC reviewed the manuscript critically for important intellectual content and conducted editing and rewriting. All authors agree to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. In addition, all authors have read and approved the final version of the manuscript and were responsible for the decision to submit the manuscript.

Conflicts of Interest

LGL receives grants from Setting Scoliosis Straight Foundation, AO Spine, and ISSG, royalties from Medtronic and Acuity Surgical, consulting fees from Medtronic and Acuity Surgical. GMMJ owns stocks in Nuvasive; reports consulting fees from Orthofix and Carlsmed, Inc; receives royalties from SI-BONE. The other authors declare no conflicts of interest.

Multimedia Appendix 1
The outputs for generating abstracts based on prompt “Please write a scientific abstract for the article [title] in the style of [journal] at [link]” and human-written abstracts.
[PDF File (Adobe PDF File), 912 KB - jmir_v26i1e52001_app1.pdf]

References

1. Will ChatGPT transform healthcare? Nat Med 2023;29(3):505-506. [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
2. Bi AS. What's important: the next academic-ChatGPT AI? J Bone Joint Surg Am 2023:00. [doi: 10.2106/JBJS.23.00269] [Medline: 37083839]
3. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
4. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]

XSL•FO
RenderX

5. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023;6(1):75 [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
6. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 2023 Apr 26;9:e47737 [FREE Full text] [doi: 10.2196/47737] [Medline: 37099373]
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
9. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
10. Leopold SS, Haddad FS, Sandell LJ, Swiontkowski M. Artificial intelligence applications and scholarly publication in orthopaedic surgery. J Orthop Res 2023;41(6):1137-1138. [doi: 10.2106/JBJS.23.00293] [Medline: 37067181]
11. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
12. Levin G, Meyer R, Yasmeen A, Yang B, Guigue P, Bar-Noy T, et al. Chat generative pre-trained transformer-written obstetrics and gynecology abstracts fool practitioners. Am J Obstet Gynecol MFM 2023;5(8):100993. [doi: 10.1016/j.ajogmf.2023.100993] [Medline: 37127209]
13. Kim HJ, Yang JH, Chang D, Lenke LG, Suh SW, Nam Y, et al. Adult spinal deformity: a comprehensive review of current advances and future directions. Asian Spine J 2022;16(5):776-788 [FREE Full text] [doi: 10.31616/asj.2022.0376] [Medline: 36274246]
14. Kim HJ, Yang JH, Chang D, Suk S, Suh SW, Kim S, et al. Proximal junctional kyphosis in adult spinal deformity: definition, classification, risk factors, and prevention strategies. Asian Spine J 2022 Jun;16(3):440-450 [FREE Full text] [doi: 10.31616/asj.2020.0574] [Medline: 33910320]
15. Kim MW, Kang C, Choi SH. Update of the natural history, pathophysiology, and treatment strategies of degenerative cervical myelopathy: a narrative review. Asian Spine J 2023;17(1):213-221 [FREE Full text] [doi: 10.31616/asj.2022.0440] [Medline: 36787787]
16. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng 2023;51(12):2647-2651. [doi: 10.1007/s10439-023-03284-0] [Medline: 37328703]
17. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023;41(8):1099-1106. [doi: 10.1038/s41587-022-01618-2] [Medline: 36702895]
18. Miner AS, Fleming SL, Haque A, Fries JA, Althoff T, Wilfley DE, et al. A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency. Npj Ment Health Res 2022;1(1):19. [doi: 10.1038/s44184-022-00020-9] [Medline: 38609510]
19. Caramancion KM. A Comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and bard in news fact-checking. arXiv Preprint posted online on June 18, 2023. [doi: 10.1109/fnwf58287.2023.10520446]
20. Patnaik SS, Hoffmann U. Quantitative evaluation of ChatGPT versus Bard responses to anaesthesia-related queries. Br J Anaesth 2024;132(1):169-171. [doi: 10.1016/j.bja.2023.09.030] [Medline: 37945414]
21. OpenAI. GPT-4 technical report. arXiv Preprint posted online on March 15, 2023. [doi: 10.48550/arXiv.2303.08774]
22. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: pandora's box has been opened. J Med Internet Res 2023;25:e46924 [FREE Full text] [doi: 10.2196/46924] [Medline: 37256685]
23. Liu N, Brown A. AI increases the pressure to overhaul the scientific peer review process. Comment on "Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened". J Med Internet Res 2023;25:e50591 [FREE Full text] [doi: 10.2196/50591] [Medline: 37651167]
24. Májovský M, Mikolov T, Netuka D. AI is changing the landscape of academic writing: what can be done? Authors' reply to: AI Increases the Pressure to Overhaul the Scientific Peer Review Process. Comment on "Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened". J Med Internet Res 2023;25:e50844 [FREE Full text] [doi: 10.2196/50844] [Medline: 37651175]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
LLM: large language model

XSL•FO
RenderX

ROC: receiver operative characteristics

Edited by T de Azevedo Cardoso; submitted 20.08.23; peer-reviewed by L Zhu, H Mondal, M Májovský; comments to author 10.01.24; revised version received 15.01.24; accepted 26.04.24; published 26.06.24.

Please cite as:
Kim HJ, Yang JH, Chang DG, Lenke LG, Pizones J, Castelein R, Watanabe K, Trobisch PD, Mundis Jr GM, Suh SW, Suk SI
Assessing the Reproducibility of the Structured Abstracts Generated by ChatGPT and Bard Compared to Human-Written Abstracts in the Field of Spine Surgery: Comparative Analysis
J Med Internet Res 2024;26:e52001
URL: https://www.jmir.org/2024/1/e52001
doi:10.2196/52001
PMID:

©Hong Jin Kim, Jae Hyuk Yang, Dong-Gune Chang, Lawrence G Lenke, Javier Pizones, René Castelein, Kota Watanabe, Per D Trobisch, Gregory M Mundis Jr, Seung Woo Suh, Se-Il Suk. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis

John Michael Hoppe[1], MD; Matthias K Auer[1], MD; Anna Strüven[2,3], MD; Steffen Massberg[2,3], MD; Christopher Stremmel[2,3], MD

[1]Department of Medicine IV, LMU University Hospital, Munich, Germany

[2]Department of Medicine I, LMU University Hospital, Munich, Germany

[3]Munich Heart Alliance Partner Site, Deutsches Zentrum für Herz-Kreislaufforschung (German Centre for Cardiovascular Research), LMU University Hospital, Munich, Germany

Corresponding Author:
Christopher Stremmel, MD
Department of Medicine I
LMU University Hospital
Marchioninistr 15
Munich, 81377
Germany
Phone: 49 89 4400 712622
Email: christopher.stremmel@med.uni-muenchen.de

Abstract

Background: OpenAI's ChatGPT is a pioneering artificial intelligence (AI) in the field of natural language processing, and it holds significant potential in medicine for providing treatment advice. Additionally, recent studies have demonstrated promising results using ChatGPT for emergency medicine triage. However, its diagnostic accuracy in the emergency department (ED) has not yet been evaluated.

Objective: This study compares the diagnostic accuracy of ChatGPT with GPT-3.5 and GPT-4 and primary treating resident physicians in an ED setting.

Methods: Among 100 adults admitted to our ED in January 2023 with internal medicine issues, the diagnostic accuracy was assessed by comparing the diagnoses made by ED resident physicians and those made by ChatGPT with GPT-3.5 or GPT-4 against the final hospital discharge diagnosis, using a point system for grading accuracy.

Results: The study enrolled 100 patients with a median age of 72 (IQR 58.5-82.0) years who were admitted to our internal medicine ED primarily for cardiovascular, endocrine, gastrointestinal, or infectious diseases. GPT-4 outperformed both GPT-3.5 ($P<.001$) and ED resident physicians ($P=.01$) in diagnostic accuracy for internal medicine emergencies. Furthermore, across various disease subgroups, GPT-4 consistently outperformed GPT-3.5 and resident physicians. It demonstrated significant superiority in cardiovascular (GPT-4 vs ED physicians: $P=.03$) and endocrine or gastrointestinal diseases (GPT-4 vs GPT-3.5: $P=.01$). However, in other categories, the differences were not statistically significant.

Conclusions: In this study, which compared the diagnostic accuracy of GPT-3.5, GPT-4, and ED resident physicians against a discharge diagnosis gold standard, GPT-4 outperformed both the resident physicians and its predecessor, GPT-3.5. Despite the retrospective design of the study and its limited sample size, the results underscore the potential of AI as a supportive diagnostic tool in ED settings.

(J Med Internet Res 2024;26:e56110) doi:10.2196/56110

KEYWORDS

emergency department; diagnosis; accuracy; artificial intelligence; ChatGPT; internal medicine; AI; natural language processing; NLP; emergency medicine triage; triage; physicians; physician; diagnostic accuracy; OpenAI

XSL•FO
RenderX

Introduction

The application of artificial intelligence (AI) has now become part of everyday life. OpenAI has managed to create a highly effective platform with ChatGPT, especially for answering complex questions, positioning it as a pioneer in the field of natural language AI [1]. Despite the emergence of successors such as Google Bard, ChatGPT remains the most widely used platform and was therefore selected for this study.

Especially in the medical field, AI applications offer enormous potential [2]. They can provide helpful guideline-based treatment advice, monitor medication dosages, and signal potential interactions, among other benefits [3]. To date, relevant disadvantages have hardly come to bear. However, the fundamental question of determining reasonable limits for AI applications remains [2].

The first pioneering studies in the field of emergency medicine have just been published. Dahdah and colleagues [4] investigated the capability of GPT-3.5 as a triage tool, noting its ability to provide appropriate responses within a few seconds. Another publication used AI to generate a discharge summary and highlighted the potential benefits of this technology, such as time savings, enhanced accuracy of patient information, and optimized communication [5]. Furthermore, Al-Zaiti and colleagues [6] were able to demonstrate that a machine learning model for electrocardiogram diagnosis of non-ST segment elevation myocardial infarction outperformed both practicing clinicians and other interpretation systems.

The diagnostic accuracy of ChatGPT has, until now, mainly been evaluated using general internal medicine case vignettes, which limits its applicability in a real-world emergency department (ED) setting. Despite using an older version, GPT-3, in their initial study, the authors reported remarkably good performance for the AI chatbot. The accuracy rate for the correct diagnosis among the top 5 differential diagnoses was 98.3% for physicians, compared to 83.3% for GPT-3 (P=.03) [7]. In a follow-up study, the same research group reported a diagnostic accuracy slightly above 80% for both physicians and GPT-4 [8].

This study investigates the real-world performance of the latest versions of ChatGPT, specifically those based on GPT-3.5 and GPT-4, regarding their ability to accurately find the right diagnosis in an ED setting when provided with the same information as the treating physician. We performed a blinded head-to-head comparison of the primary treating resident physician versus ChatGPT. The discharge diagnosis, determined after an inpatient stay of several days that included detailed further diagnostics, served as the gold standard.

Methods

Assessment of Diagnostic Accuracy

In this retrospective study, we evaluated a cohort of 100 randomly selected adults admitted to our ED in January 2023. The main inclusion criterion was an unplanned inpatient admission due to an internal medical condition. Outpatients and patients presenting with noninternal medical conditions were excluded. The ED resident physician's diagnosis was defined as the diagnosis documented in the ED letter. Subsequently, each patient's case history, medical history, current medication regimen, laboratory results, and other diagnostic findings, as documented in the ED letter, were inputted into either GPT-3.5 or GPT-4. The uniform query presented to each chatbot was "What is the most likely diagnosis?" Two examples of representative cases with corresponding input information provided to the chatbot are presented in Multimedia Appendix 1.

Diagnostic accuracy of the ED resident physician and the AI chatbots was then benchmarked against the final hospital discharge diagnosis, which was established after the inpatient stay by senior physicians specialized in the relevant medical fields. Diagnostic performance was ranked on an accuracy point scale of 0 to 2, where 2 points indicated a correct diagnosis, 1 point indicated a partially correct diagnosis, and 0 points denoted an incorrect diagnosis. More specifically, a score of 2 points was awarded for a correct diagnosis that included all major diagnoses identified at admission, regardless of minor diagnoses. A score of 1 point was given for a partially correct diagnosis, which can occur in 2 scenarios: either the major diagnosis is nearly correct and the subtle differences would not have impacted treatment, or there is suspicion of multiple major differential diagnoses, with one being correct and the others incorrect. However, 0 points were assigned when all major diagnoses were incorrect. Minor diagnoses were not scored, since these had no significant impact on the patient's condition. The term "major diagnoses" refers to the conditions primarily responsible for the patient's main symptom upon admission to our ED. The grading was performed in a blinded manner by senior physicians trained in emergency medicine (Multimedia Appendix 2).

Ethical Considerations

Prior to inputting any information into the chatbots, each patient's data were anonymized and all personally identifiable information was removed according to data privacy standards. The study was performed in accordance with the Helsinki Declaration and was approved by the ethics committee of LMU Munich (23-0445).

Statistical Analysis

Statistical analyses were performed using Prism (version 9; GraphPad). Ordinal variables were reported as means. For group statistics, we used a 2-way ANOVA with Tukey correction for multiple comparisons. P values <.05 were considered statistically significant. No prior sample size calculation was performed.

Results

Baseline Characteristics

The median age of our study population was 72 (IQR 58.5-82.0) years, and 45 of our patients were female. The largest proportion, 40 patients in total, were admitted due to cardiovascular diseases. Major pathologies included acute coronary syndrome (n=8), heart failure (n=8), arrhythmias (n=7), and hypertensive crisis (n=4). A total of 22 patients each were

XSL•FO
RenderX

admitted with endocrine or gastrointestinal diseases and infectious diseases. The remaining 16 patients presented with kidney or rheumatic diseases (n=9) and pulmonary diseases (n=7) (Table 1, Multimedia Appendix 2).

Table 1. Demographics (N=100).

Characteristics	Values
Age (years), median (IQR)	72 (58.5-82.0)
Sex, n	
Male	55
Female	45
Acute medical condition, n	
Cardiovascular diseases	40
Acute coronary syndrome	8
Heart failure	8
Arrhythmia	7
Hypertensive crisis	4
Pericarditis or myocarditis	2
Pulmonary embolism or deep vein thrombosis	2
Peripheral artery disease	2
Other	7
Endocrine and gastrointestinal diseases, n	
Total	22
Gastrointestinal bleeding	5
Diabetes-related complications	4
Liver disease	4
Cholangitis	3
Acute pancreatitis	3
Other	3
Infectious diseases, n	
Total	22
Pneumonia	13
Urinary tract infection	6
Other	3
Other diseases, n	
Total	16
Kidney and rheumatic diseases	9
Asthma and chronic obstructive pulmonary disease	7

GPT-4 Surpassed GPT-3.5 and Resident Physicians in Diagnostic Accuracy for Emergency Patients

When comparing GPT-4, GPT-3.5, and resident physicians in predicting the diagnoses of internal medicine emergency patients, GPT-4 demonstrated superior performance (Figure 1). GPT-4 achieved an accuracy score of 1.76 out of a possible 2 points, while GPT-3.5 attained 1.51 points ($P<.001$). Notably, GPT-4 also significantly surpassed the accuracy score of our resident physicians, who achieved a score of 1.59 ($P=.01$). In contrast, the performance of GPT-3.5 was slightly inferior to that of the resident physicians, yet the difference was not statistically significant ($P=.36$).

XSL•FO
RenderX

Figure 1. Diagnostic accuracy of resident physicians, GPT-3.5, and GPT-4 for emergency patients. Shown is the mean diagnostic performance across emergency cases (N=100) ranked on a diagnostic accuracy point scale of 0 to 2, where 2 points stand for accurate, 1 point for partially correct, and 0 for incorrect diagnosis. Also, the sample set was stratified in subgroups for cardiovascular diseases (n=40), endocrine and gastrointestinal diseases (n=22), infectious diseases (n=22), and other diseases (n=16). ED: emergency department. *P<.05; **P<.01; ***P<.001.

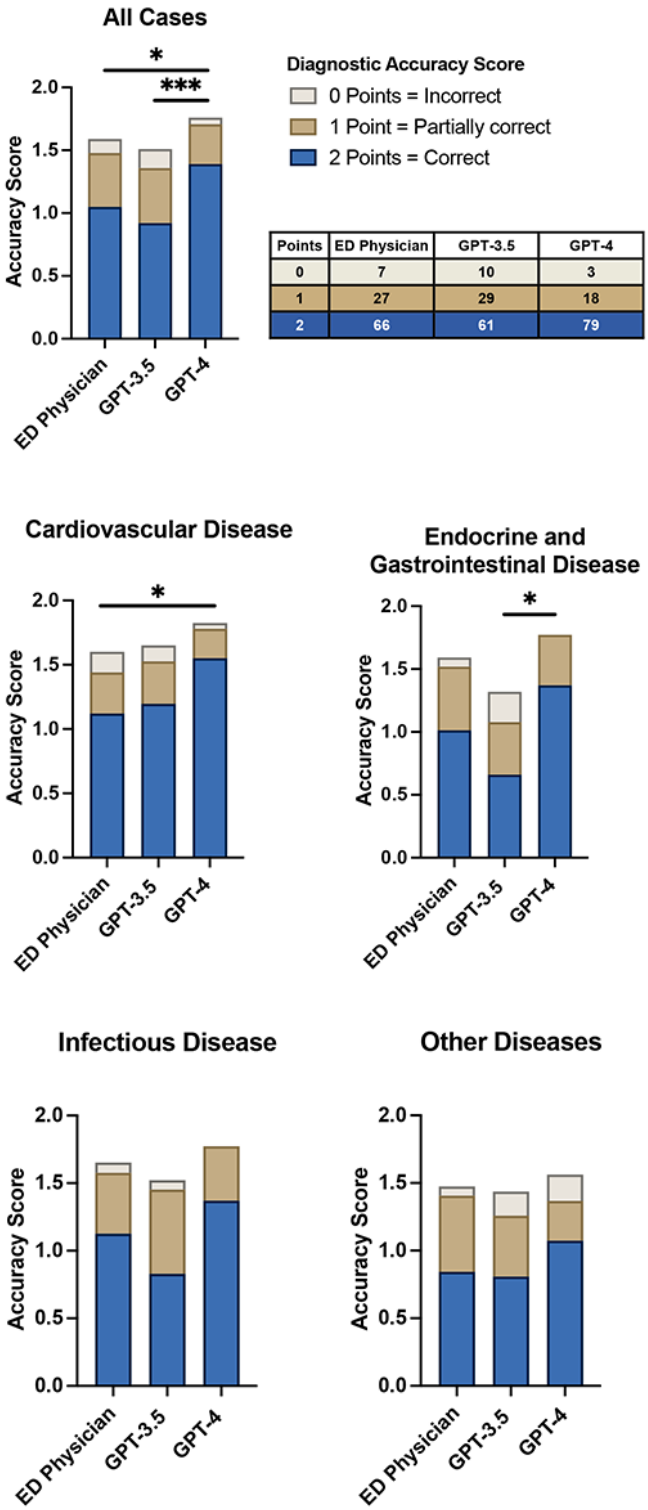

Points	ED Physician	GPT-3.5	GPT-4
0	7	10	3
1	27	29	18
2	66	61	79

GPT-4 Is at Least on Par With or Better Than GPT-3.5 and Resident Physicians in Diagnostic Accuracy Across Multiple Disease Subgroups for Emergency Patients

When stratified for cardiovascular diseases, GPT-4 scored 1.83 points, outperforming GPT-3.5, which scored 1.65, and the resident physician, who scored 1.60. However, only the comparison of GPT-4 versus ED physicians reached statistical significance (P=.03) (Figure 1).

For the subgroup of endocrine or gastrointestinal diseases, GPT-4 performed significantly better, with a score of 1.77 points, compared to GPT-3.5 with a score of 1.32 points (P=.01). The resident physician achieved 1.59 points, placing them

XSL•FO
RenderX

between the 2 performances of the ChatGPT versions, yet the difference compared to GPT-4 was not statistically significant (*P*=.47) (Figure 1).

Within the subgroup of infectious diseases, GPT-4 again outperformed, with a score of 1.77, while GPT-3.5 and the resident physician achieved scores of 1.52 and 1.65, respectively. Compared to GPT-4, these findings did not show a significant difference (Figure 1).

Lastly, we compared the subgroup of "other diseases," which included kidney and rheumatic diseases, as well as asthma and chronic obstructive pulmonary disease. In the assessment, GPT-4 achieved the highest score at 1.56 points. GPT-3.5 scored 1.44 points and the resident physician 1.50 points. The differences between GPT-4 and the other 2 were not significant (Figure 1).

Discussion

In this real-world pilot study, we investigated the potential of GPT-3.5 and GPT-4 to identify the correct diagnosis based on patients' current concerns, medical history, current medication regimen, laboratory results, and other diagnostic findings. We performed a head-to-head comparison of GPT-3.5 versus GPT-4 versus the primary treating resident physician in the ED, evaluating their diagnostic accuracy against the gold standard of the final discharge diagnosis after several days of hospital admission and further examinations.

While GPT-3.5 achieved the same diagnostic accuracy in the overall evaluation, GPT-4 surpassed the ED resident physician. Of note, a direct comparison of GPT-3.5 versus GPT-4 showed a significantly better performance by the latest version—a trend that has already been speculated on in prior evaluations of diagnostic accuracy [7,8]. This superiority of GPT-4 was also evident in the subanalysis of endocrine and gastrointestinal diseases. All other specialty-specific analyses failed to reach statistical significance due to limited case numbers, but showed a trend consistent with the overall cohort findings.

Our observations align with a limited number of smaller previous studies that investigated the diagnostic performance of ChatGPT using clinical vignettes derived from general internal medicine case reports. The rate of correctly identifying a diagnosis among the top 5 suggested differential diagnoses was slightly over 80% for both physicians and ChatGPT [7-9]. In terms of listing a correct or partially correct diagnosis, our real-world study approach reached an accuracy of 93% for the ED physician, 90% for GPT-3.5, and 97% for GPT-4. This high performance likely results from the comprehensive clinical and diagnostic information provided by the treating ED physician. Conversely, ChatGPT performance evaluations that were solely based on the input of self-reported patient symptoms only identified about 50% of the top 3 diagnostic matches [10].

The superiority of ChatGPT lies in its capacity to rapidly generate a range of differential diagnoses, encompassing even rare diseases, thus providing an analytical approach that may surpass the physician. In our study, it is conceivable that the treating resident physician initially considered several differential diagnoses but documented only the most probable one and therefore scored lower. However, it must be assumed that in most cases some differential diagnoses were never considered. This might stem from a physician's natural tendency to focus on specific symptoms while neglecting subtler ones. Hartigan and colleagues [11] state that physicians are prone to cognitive errors, since both faster intuitive and slower analytical reasoning have potential drawbacks when applied in the clinical setting. When using intuitive reasoning, the physician may unconsciously place a higher weight on personal or patient-specific factors or over- or underemphasize the significance of a data point. Conversely, analytical reasoning is particularly prone to errors in cases where the disease presentation is rare and probability-based decision-making may lead to a more common diagnosis being suspected. GPT-3.5, and particularly GPT-4, due to their analytical data processing, are less prone to some of the errors that can emerge from intuitive reasoning. Example 1 in Multimedia Appendix 1 illustrates a potential cognitive bias among resident physicians. In this instance, a resident assumed that the patient's chest pain was caused by myocardial infarction, disregarding the patient's history of lung cancer. This likely resulted from a cognitive error due to fixation on the most probable diagnosis. GPT-3.5 also responded incorrectly, while only GPT-4 identified lung cancer as a possible differential diagnosis.

The retrospective design of our study does not limit the quality of the results, as there was no mutual interference or selection bias. However, the reliance of ChatGPT on information provided by the treating ED resident physician could potentially bias the diagnosis. Additionally, this pioneering study has a limited sample size, which becomes especially apparent in our subanalysis of internal medicine specialties.

Within the broader discussion on the integration of AI in health care, the use of ChatGPT in ED settings raises privacy concerns due to the input of sensitive patient information into systems that may not be entirely secure. There is a risk that this data could be stored or accessed improperly, violating confidentiality laws. Prior to implementing AI in health care settings, it is crucial to ensure the secure management of data.

In the future, AI technologies will become increasingly important in the ED setting, where the time-critical environment demands any supportive tools to facilitate work and improve patient care. ChatGPT and comparable technologies do not compete with resident physicians, but rather assist them in making auxiliary decisions. Moreover, ChatGPT has demonstrated superior diagnostic accuracy in our patient cohort, and future larger studies are needed to confirm this observation and to investigate the use of ChatGPT as a supportive tool in decision-making. We hypothesize that the performance of ChatGPT might even improve with upcoming versions.

XSL•FO
RenderX

Authors' Contributions

JMH, SM, and CS conceived and designed the study. MKA provided statistical advice on study design and JMH analyzed the data. JMH and CS drafted the manuscript, and all authors contributed substantially to its revision. JMH and CS take responsibility for the paper as a whole.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Representative cases.
[DOCX File , 21 KB - jmir_v26i1e56110_app1.docx]

Multimedia Appendix 2
Diagnoses and accuracy scores.
[XLSX File (Microsoft Excel File), 22 KB - jmir_v26i1e56110_app2.xlsx]

References

1. Schulman J, Zoph B, Kim C. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2024-06-25]
2. Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? Nature 2022 Dec 09. [doi: 10.1038/d41586-022-04397-7] [Medline: 36494443]
3. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
4. Dahdah JE, Kassab J, Helou MCE, Gaballa A, Sayles S, Phelan MP. ChatGPT: a valuable tool for emergency medical assistance. Ann Emerg Med 2023 Sep;82(3):411-413. [doi: 10.1016/j.annemergmed.2023.04.027] [Medline: 37330721]
5. Bradshaw JC. The ChatGPT era: artificial intelligence in emergency medicine. Ann Emerg Med 2023 Jun;81(6):764-765. [doi: 10.1016/j.annemergmed.2023.01.022] [Medline: 37210166]
6. Al-Zaiti SS, Martin-Gill C, Zègre-Hemsey JK, Bouzid Z, Faramand Z, Alrawashdeh MO, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. Nat Med 2023 Jul;29(7):1804-1813 [FREE Full text] [doi: 10.1038/s41591-023-02396-3] [Medline: 37386246]
7. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
8. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 09;11:e48808 [FREE Full text] [doi: 10.2196/48808] [Medline: 37812468]
9. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv. Posted February 26, 2023 2023 Feb 26 [FREE Full text] [doi: 10.1101/2023.02.21.23285886] [Medline: 36865204]
10. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth 2023 Oct 03;11:e49995 [FREE Full text] [doi: 10.2196/49995] [Medline: 37788063]
11. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: still no easy answers. West J Emerg Med 2020 Nov 02;21(6):125-131 [FREE Full text] [doi: 10.5811/westjem.2020.7.47832] [Medline: 33207157]

Abbreviations

AI: artificial intelligence
ED: emergency department

XSL•FO
RenderX

Edited by Q Jin; submitted 06.01.24; peer-reviewed by H Mondal, X Liu, Z Fatima, F Chen; comments to author 18.03.24; revised version received 08.04.24; accepted 08.05.24; published 08.07.24.

Please cite as:
Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C
ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis
J Med Internet Res 2024;26:e56110
URL: https://www.jmir.org/2024/1/e56110
doi:10.2196/56110
PMID:

©John Michael Hoppe, Matthias K Auer, Anna Strüven, Steffen Massberg, Christopher Stremmel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis

Mingxin Liu[1], MA; Tsuyoshi Okuhara[2], PhD; XinYi Chang[3], MA; Ritsuko Shirabe[2], PhD; Yuriko Nishiie[1], MPH; Hiroko Okada[2], PhD; Takahiro Kiuchi[2], PhD

[1]Department of Health Communication, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[2]Department of Health Communication, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[3]Department of Industrial Engineering and Economics, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan

Corresponding Author:
Mingxin Liu, MA
Department of Health Communication
Graduate School of Medicine
The University of Tokyo
7-3-1 Hongo, Bunkyo
Tokyo, 113-8655
Japan
Phone: 81 03 5800 6549
Fax: 81 03 5689 0726
Email: liumingxin98@g.ecc.u-tokyo.ac.jp

Abstract

Background: Over the past 2 years, researchers have used various medical licensing examinations to test whether ChatGPT (OpenAI) possesses accurate medical knowledge. The performance of each version of ChatGPT on the medical licensing examination in multiple environments showed remarkable differences. At this stage, there is still a lack of a comprehensive understanding of the variability in ChatGPT's performance on different medical licensing examinations.

Objective: In this study, we reviewed all studies on ChatGPT performance in medical licensing examinations up to March 2024. This review aims to contribute to the evolving discourse on artificial intelligence (AI) in medical education by providing a comprehensive analysis of the performance of ChatGPT in various environments. The insights gained from this systematic review will guide educators, policymakers, and technical experts to effectively and judiciously use AI in medical education.

Methods: We searched the literature published between January 1, 2022, and March 29, 2024, by searching query strings in Web of Science, PubMed, and Scopus. Two authors screened the literature according to the inclusion and exclusion criteria, extracted data, and independently assessed the quality of the literature concerning Quality Assessment of Diagnostic Accuracy Studies-2. We conducted both qualitative and quantitative analyses.

Results: A total of 45 studies on the performance of different versions of ChatGPT in medical licensing examinations were included in this study. GPT-4 achieved an overall accuracy rate of 81% (95% CI 78-84; $P<.01$), significantly surpassing the 58% (95% CI 53-63; $P<.01$) accuracy rate of GPT-3.5. GPT-4 passed the medical examinations in 26 of 29 cases, outperforming the average scores of medical students in 13 of 17 cases. Translating the examination questions into English improved GPT-3.5's performance but did not affect GPT-4. GPT-3.5 showed no difference in performance between examinations from English-speaking and non–English-speaking countries ($P=.72$), but GPT-4 performed better on examinations from English-speaking countries significantly ($P=.02$). Any type of prompt could significantly improve GPT-3.5's ($P=.03$) and GPT-4's ($P<.01$) performance. GPT-3.5 performed better on short-text questions than on long-text questions. The difficulty of the questions affected the performance of GPT-3.5 and GPT-4. In image-based multiple-choice questions (MCQs), ChatGPT's accuracy rate ranges from 13.1% to 100%. ChatGPT performed significantly worse on open-ended questions than on MCQs.

Conclusions: GPT-4 demonstrates considerable potential for future use in medical education. However, due to its insufficient accuracy, inconsistent performance, and the challenges posed by differing medical policies and knowledge across countries, GPT-4 is not yet suitable for use in medical education.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD42024506687; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=506687

(J Med Internet Res 2024;26:e60807) doi:10.2196/60807

KEYWORDS

large language model, ChatGPT, medical licensing examination, medical education; LLMs; NLP; natural language processing; artificial intelligence; language models; review methods; systematic; meta-analysis

Introduction

Background

In November 2022, the web-based artificial intelligence (AI) chatbot ChatGPT (OpenAI) was released to the public and swiftly garnered global attention because of its ability to provide detailed answers to complex queries [1]. ChatGPT has been extensively applied across various domains, including programming, education, business, and law, with notable success in each [2-5]. Researchers have been actively exploring the potential roles and capabilities of ChatGPT in clinical diagnosis, health care, and medical education [6,7]. The number of publications on this topic has increased dramatically since late 2022 [8,9]. Specifically, in medical education, ChatGPT can play several important roles, including, but not limited to, the following: First, compared to search engines like Google, which present a list of relevant pages, ChatGPT aims to provide concise and practical answers to users' questions, making it an effective knowledge resource [10,11]. Second, in medical licensing examinations comprising multiple-choice questions (MCQs), ChatGPT can act as an "AI teaching assistant," providing insights for each question, analyzing common errors, and reinforcing concepts interactively [12]. Third, ChatGPT has the capability to analyze images. Although this feature is still in its early stages, it offers the potential for ChatGPT to serve as a "virtual mentor," capable of analyzing medical images such as skin rashes and x-rays [10]. Fourth, for most medical students who find it challenging to balance studying vast amounts of information, practicing evidence-based medicine, and fulfilling clinical duties, ChatGPT can provide concise summaries of clinical trials and generate key practical points from them [10].

However, a prerequisite for ChatGPT's ability to help medical students in their studies and play a role in medical education, both now and in the future, is that ChatGPT has solid and accurate knowledge of medicine. Medical licensing examinations are a crucial part of the medical education pathway as they assess the readiness of aspiring doctors to enter clinical practice. These examinations vary in format and content across countries but typically test medical knowledge, clinical reasoning, and ethical decision-making [13]. Over the past 2 years, researchers have used medical licensing examinations from various countries to test whether ChatGPT possesses accurate medical knowledge [14-57].

Although most of these studies used similar testing methods—inputting medical licensing examination questions into ChatGPT and recording the responses to calculate accuracy—the ChatGPT performance showed significant variation. A study conducted in the United States revealed that GPT-3.5 surpassed the 60% score threshold on the National Board of Medical Examiners (NBME)-free–Step-1 question, reaching the level of a third-year medical student [21]. However, studies from South Korea, China, and Japan have indicated that GPT-3.5 failed to pass medical examinations in their respective countries [26,43,44,47,48,51,54]. Although GPT-4 performed better overall than GPT-3.5 [33,36,41,44,47], it did not pass the Japanese medical licensing examination [49]. In addition, ChatGPT performance varies significantly across medical specialties within these examinations [23,25-27,30,33-35].

At this stage, there is still a lack of a comprehensive understanding of the variability in ChatGPT's performance on different medical licensing examinations. We believe that prematurely using ChatGPT for clinical diagnosis and medical education without thoroughly evaluating its performance across various medical licensing examinations is irresponsible and could endanger human lives.

Literature Review

A total of 3 systematic reviews have explored ChatGPT's performance in medical licensing examinations to the best of our knowledge [58-60].

A study from the United States collected literature up to June 2, 2023, focusing on various types of medical licensing examinations in the United States [58]. Among the 19 included studies, only 2 were comprehensive medical licensing examinations, the United States Medical Licensing Examination (USMLE), while the remaining 17 were medical specialty examinations, such as plastic surgery, anesthesia, and ophthalmology [58]. In contrast to this study, our research extends the literature collection to a global scale and examines the performance of ChatGPT in medical licensing examinations in different countries and languages. We believe that the worldwide perspective of the current review is crucial because medical education and licensure standards vary significantly across countries.

A study from Pakistan collected literature up to April 2023, focusing on the performance of GPT-3.5 in various medical licensing examinations worldwide [59]. However, with the advent of the more advanced GPT-4, more studies have focused on GPT-4. Our research includes all ChatGPT versions and discusses their performance differences.

A study from China collected the literature up to July 15, 2023 [60]. This study reviewed the performance of ChatGPT for various medical questions. Of the 60 included studies, only 3 were medical licensing examinations. In addition, this study created a framework to evaluate the quality of studies on the performance of large language models (LLMs) in medical questions [60]. We slightly modified this evaluation framework and applied it to this study.

XSL•FO
RenderX

Study Aims and Objectives

This study reviewed all studies on ChatGPT's performance in medical licensing examinations from January 1, 2022, to March 29, 2024, to clarify the following issues:

1. Can ChatGPT pass the medical licensing examinations?
2. How does ChatGPT's performance compare to that of medical students?
3. How did ChatGPT perform in different languages?
4. What is the relationship between question difficulty and ChatGPT's performance?
5. What is the relationship between question length and ChatGPT's performance?
6. How did ChatGPT perform on image-based MCQ?
7. How did ChatGPT perform on open-ended questions?
8. What is the difference in ChatGPT's performance with and without prompts?
9. Comparison of GPT-3.5's and GPT-4's performances.
10. How does ChatGPT perform in medical licensing examinations in English-speaking countries and non–English-speaking countries?

By comprehensively evaluating the accuracy of the medical knowledge held by ChatGPT, we integrate these perspectives and offer comprehensive recommendations for applying ChatGPT in medical education.

Overall, this systematic review aimed to fill the knowledge gap regarding the application of ChatGPT in medical licensing examinations. Further, it sought to contribute to the evolving discourse on AI in medical education and facilitate future developments and applications in this field. The insights gained from this systematic review will guide educators, policymakers, and technical experts to effectively and judiciously use AI in medical education.

To the best of our knowledge, this is the first study to comprehensively review the performance of all versions of ChatGPT on medical licensing examinations across different countries.

Methods

This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagrams and guidance [61]. This systematic review was registered in the PROSPERO (International Prospective Register of Systematic Reviews) database on February 1, 2024 (CRD42024506687).

Search Strategy

We searched for specific query strings (Multimedia Appendix 1) using the advanced search function in PubMed, Web of Science, and Scopus, with Google Scholar as a supplementary source. Literature published from January 1, 2022, to March 29, 2024, was included in the literature search. The literatures exported from these 3 platforms were imported into Rayyan [62]. Two authors (ML and XC) independently screened the titles and abstracts of the retrieved studies using a search strategy to identify those that met the inclusion and exclusion criteria (Textbox 1). The full texts of these studies were then retrieved and independently assessed for eligibility by 2 authors. Any disagreements regarding the eligibility of specific studies were discussed and resolved by a third reviewer (TO). In addition to the database searches, we searched Google Scholar for triangulations on March 29, 2024. When the preprint and peer-reviewed literature data were identical, we included the peer-reviewed literature in our analysis. As part of the screening process, we recorded the reasons for study exclusion and presented them in a prismatic flow diagram.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- The study tested the performance of ChatGPT in medical licensing examinations.
- Any type of original research literature (peer-reviewed papers, conference papers, preprints, letters, books, etc).
- Literature published from 2022 to 2024.
- Literature on the performance of ChatGPT in all languages.
- Literature on any version of ChatGPT.
- Literature on multiple-choice questions, open-ended questions, and all other types of questions for medical licensing examinations.

Exclusion criteria

- Nonnational-level medical licensing examination.
- Examinations other than comprehensive medical licensing examinations (eg, medical final examinations at universities, medical questions created by the authors themselves, and medical specialty examinations).
- Studies that are not related to ChatGPT.
- Duplicate studies.
- Studies that are not published in English.
- Systematic review.

XSL•FO
RenderX

Data Extraction and Management

Two reviewers (ML and XC) independently extracted data from the included studies into an Excel (Microsoft Corp) spreadsheet by 2 reviewers (ML and XC). The data were compared, and inconsistencies were resolved through consensus or by a third reviewer (TO). The general characteristics to be extracted include the following: (1) title, (2) authors, (3) publication year, (4) publication date, (5) type of publication, (6) country of the medical licensing examination, (7) name of the medical licensing examination, (8) ChatGPT version, (9) language in which ChatGPT was tested, (10) duration of the test, (11) type of questions, (12) counts of correct or total questions, (13) accuracy rate, (14) did ChatGPT pass the examination, (15) comparison between medical students, and (16) was a prompt used.

Assessing the Risk of Bias in the Included Studies

A previous study developed an LLM evaluation framework based on the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [60,63]. We modified and applied this evaluation framework in our study (Multimedia Appendix 2).

Since this previous study collected papers on ChatGPT's performance across all types of medical questions [60], we modified the original framework, whereas our research focused on ChatGPT's performance in medical licensing examinations. Specifically, we added 2 evaluation items, items 4 and 5, to address aspects specific to medical licensing examinations. We removed item 8 (are the questions individual stand-alone queries or a continuous conversation requiring multiple consecutive inquiries?) from the original evaluation framework, as it did not apply to this study.

In our modified evaluation framework, "task generation," "conversation structure," and "evaluation" correspond to "patient selection," "index test," and "reference standard" in QUADAS-2, respectively. Items 2 and 7 correspond to "flow and timing" in QUADAS-2.

Evidence Synthesis

Our analysis focuses on GPT-3.5 and GPT-4.

Qualitative Analyses

We performed a comprehensive summary using narrative analysis and descriptive statistics for the contents of the included studies that were narrative or lacked sufficient data.

Quantitative Analyses

We used the raw correct and total data in each included study to calculate the accuracy rate. The calculation rules are as follows: if a study used 1 set of questions for repeated testing, the displayed accuracy rate is the average score of all attempts and the total number of questions in the set. If the study tested both the original language and translated English questions, the displayed accuracy rate was based on the scores from the original language examination questions. For studies tested with and without optimized prompts, the displayed accuracy rate was based on the scores without optimized prompts. In studies that included MCQs and open-ended questions, the displayed accuracy rate excluded scores from the open-ended questions.

We conducted a meta-analysis of studies that tested ChatGPT using MCQs.

The I^2 statistic was used to assess the effect of heterogeneity on the pooled results. When significant heterogeneity was present (I^2>50%), a random effects model was used; otherwise, a fixed effects model was used. Accuracy was reported with a 95% CI. The significance level was set at P<.05. Meta-regression and subgroup analyses were conducted to examine the potential sources of heterogeneity and compare performances across different subgroups. A sensitivity analysis was conducted to assess the robustness of the meta-analysis results. Accuracy was reported with 95% CIs. The "metafor" and "meta" packages in R (version 4.4.0; R Core Team) were used for the meta-analysis, publication bias, and sensitivity analyses.

In addition, we conducted post hoc power analysis for the random effects model results of each main group and subgroup. G*Power (version 3.1.9.7; Erdfelder, Faul, and Buchner) was used for the power analysis.

Results

Literature Screening and Selection

By searching the query strings in the Web of Science, Scopus, and PubMed, we retrieved 3698 papers from the Web of Science, 6354 papers from Scopus, and 2587 papers from PubMed. After excluding 3751 duplicate papers, 8888 papers remained. We excluded 278 non-English papers, leaving 8610 papers. After reading the abstracts of all 8610 papers, we excluded 8,377 studies that were completely irrelevant to this review, leaving 233 studies remaining.

A total of 137 focused on ChatGPT's performance in medical specialty examinations, 11 on dental licensing examinations, 6 on nursing examinations, 6 on pharmacist examinations, and 25 on other medical examinations (eg, university medical entrance examinations and university medical final examinations). Further, 2 were systematic reviews, 1 was about nonnational medical examinations, and 2 lacked the necessary information. These studies did not meet the inclusion criteria.

We then performed a supplementary search using Google Scholar and added 2 preprint papers on March 29, 2024. Ultimately, 45 papers were included in this systematic review (Figure 1) [3,14-57].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

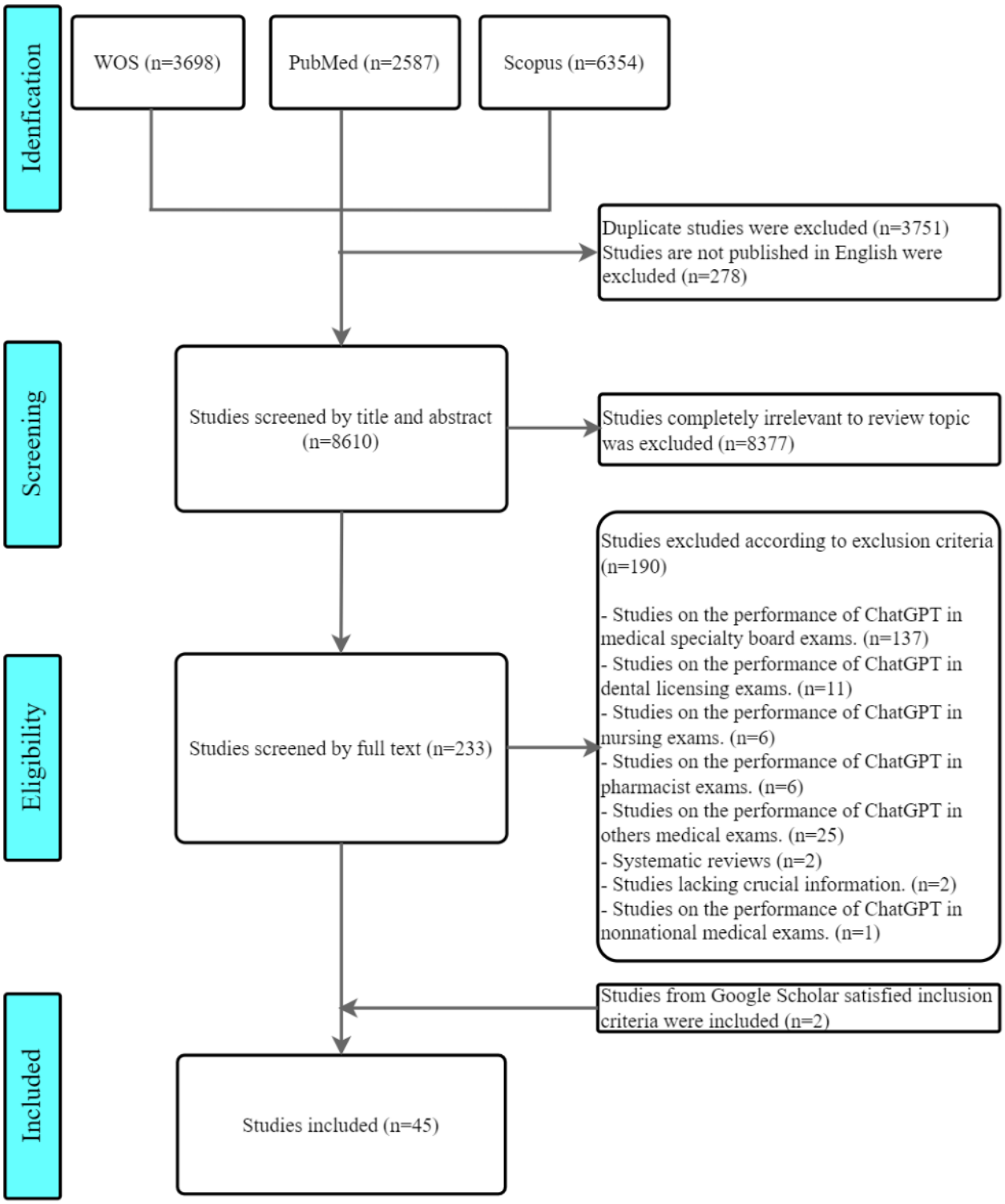

Quality Assessment of Included Studies

Two authors independently assessed the quality of the 45 studies using an evaluation framework, and any disagreements were resolved through discussion and consensus (Figure 2). The literature we collected tested ChatGPT's performance using national medical licensing examinations comprising MCQs with standard answers. Consequently, items 13, 14, 15, and 21 pertain to evaluators were not mentioned in three-quarters of the included studies. Unlike open-ended questions, MCQs do not require multiple evaluators to adopt a double-anonymized approach to evaluate test results. Therefore, this does not increase the risk in the "reference standard" part.

XSL•FO
RenderX

Figure 2. Quality assessment of included studies using evaluation framework.

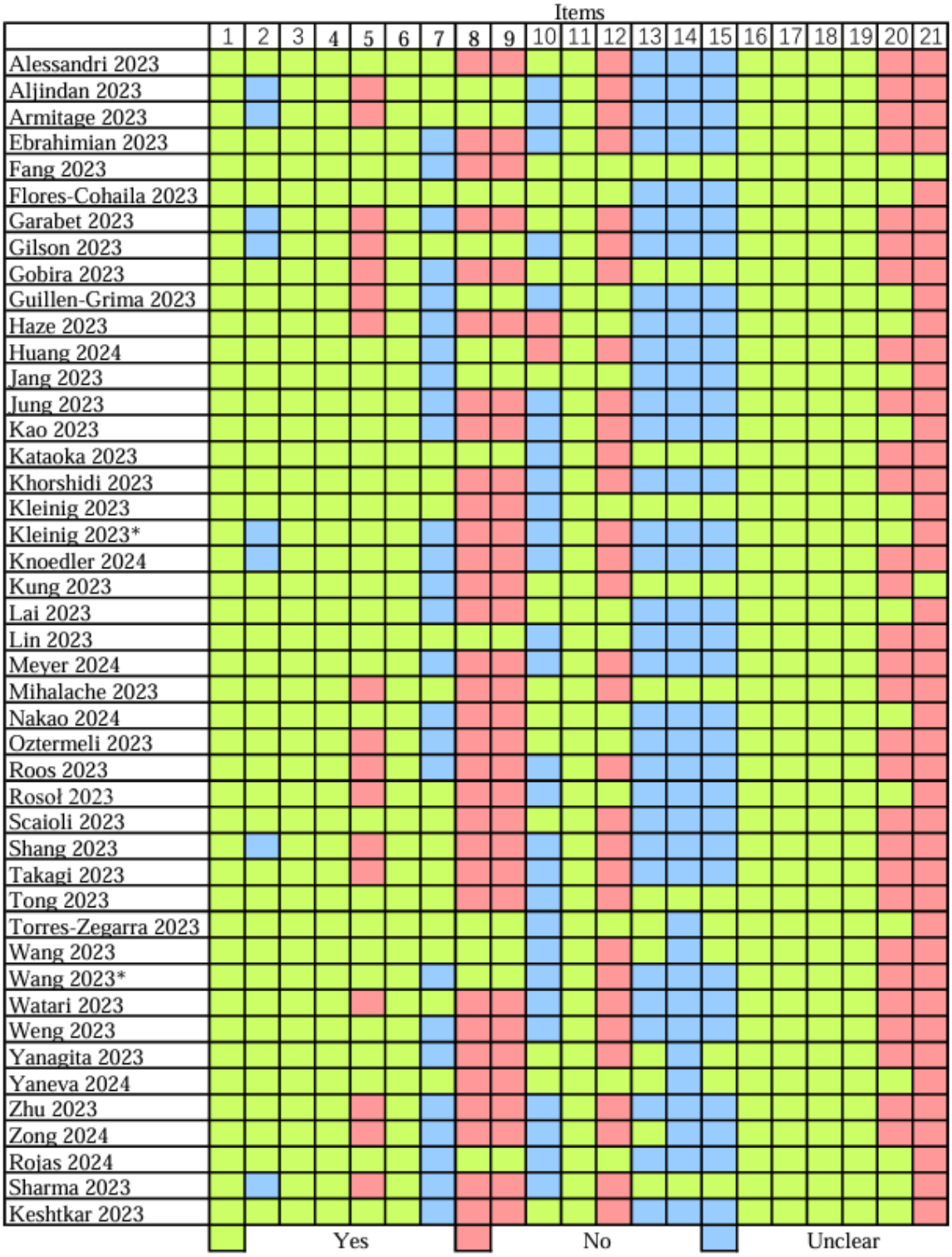

For item 7, more than half of the studies did not specify the exact test dates. On November 6, 2023, OpenAI developers announced that the cutoff dates for ChatGPT versions 3.5 and 4 were updated from September 2021 to January 2022 and April 2023, respectively [64]. We believe that if the cutoff date of ChatGPT is updated during the testing period, this might affect the consistency of ChatGPT's performance before and after the update.

For item 10, more than half of the studies did not specify whether a new chat session was used to test different questions. Conducting different questions in the same session might have affected the ChatGPT performance.

For reasons above, in the risk of bias assessment, only 2 studies and 3 studies were rated as high risk in the "index test" and "flow and timing" categories, respectively (Table 1).

XSL•FO
RenderX

Table 1. Risk of bias.

Author (year) and reference	Patient selection	Index test	Reference standard	Flow and timing
Alessandri Bonetti et al (2024) [14]	Low	Low	Low	Low
Aljindan et al (2023) [15]	Unclear	Unclear	Low	Low
Armitage (2024) [16]	Unclear	Unclear	Low	Low
Ebrahimian et al (2023) [17]	Low	Unclear	Low	Unclear
Fang et al (2023) [18]	Low	Low	Low	Unclear
Flores-Cohaila et al (2023) [19]	Low	Low	Low	Low
Garabet et al (2023) [20]	Unclear	Low	Low	High
Gilson et al (2023) [21]	Unclear	Unclear	Low	Low
Gobira et al (2023) [22]	Low	Low	Low	Unclear
Guillen-Grima et al (2023) [23]	Low	Unclear	Low	Unclear
Haze et al (2023) [24]	Low	High	Low	Unclear
Huang et al (2024) [25]	Low	High	Low	Unclear
Jang et al (2023) [26]	Low	Low	Low	Unclear
Jung et al (2023) [27]	Low	Unclear	Low	Unclear
Kao et al (2024) [28]	Low	Unclear	Low	Unclear
Kataoka et al (2023) [29]	Low	Unclear	Low	Low
Khorshidi et al (2023) [30]	Low	Unclear	Low	Low
Kleinig et al (2023) [31]	Low	Unclear	Low	Low
Kleinig et al (2023) [32]	Low	Unclear	Low	High
Knoedler et al (2024) [33]	Low	Unclear	Low	Unclear
Kung et al (2023) [3]	Low	Low	Low	Unclear
Lai et al (2023) [34]	Low	Low	Low	Unclear
Lin et al (2024) [35]	Low	Unclear	Low	Low
Meyer et al (2024) [36]	Low	Unclear	Low	Unclear
Mihalache et al (2023) [37]	Low	Low	Low	Low
Nakao et al (2024) [38]	Low	Low	Low	Unclear
Oztermeli and Oztermeli (2023) [39]	Low	Low	Low	Unclear
Roos et al (2023) [40]	Low	Unclear	Low	Unclear
Rosoł et al (2023) [41]	Low	Unclear	Low	Low
Scaioli et al (2023) [42]	Low	Low	Low	Low
Shang et al (2023) [43]	Unclear	Unclear	Low	Low
Takagi et al (2023) [44]	Low	Unclear	Low	Low
Tong et al (2023) [45]	Low	Unclear	Low	Low
Torres-Zegarra et al (2023) [46]	Low	Unclear	Low	Low
Wang et al (2023) [47]	Low	Unclear	Low	Low
Wang et al (2023) [48]	Low	Unclear	Low	Unclear
Watari et al (2023) [49]	Low	Unclear	Low	Low
Weng et al (2023) [50]	Low	Unclear	Low	Unclear
Yanagita et al (2023) [51]	Low	Low	Low	Unclear
Yaneva et al (2024) [52]	Low	Low	Low	Low
Zhu et al (2023) [53]	Low	Unclear	Low	Unclear

XSL•FO
RenderX

Author (year) and reference	Patient selection	Index test	Reference standard	Flow and timing
Zong et al (2024) [54]	Low	Unclear	Low	Unclear
Rojas et al (2024) [55]	Low	Unclear	Low	Unclear
Kung et al (2023) [56]	Unclear	Unclear	Low	High
Keshtkar et al (2023) [57]	Low	Low	Low	Low

General Characteristics of Included Studies

Among the 45 reviewed papers, the earliest was published on February 8, 2023 [21], and the latest on April 30, 2024 [55]. The general characteristics of the studies are shown in Multimedia Appendix 3.

The medical licensing examinations applied to test ChatGPT's performance were from 17 countries and regions: Italy (n=2), Saudi Arabia (n=1), the United Kingdom (n=2), Iran (n=3), China (n=7), Peru (n=2), the United States (n=7), Brazil (n=1), Spain (n=1), Japan (n=6), Taiwan (n=4), South Korea (n=1), Germany (n=3), Australia (n=2), Turkey (n=1), Poland (n=1), and Chile (n=1; Figure 3).

Figure 3. Countries where medical licensing examination was used to test ChatGPT.

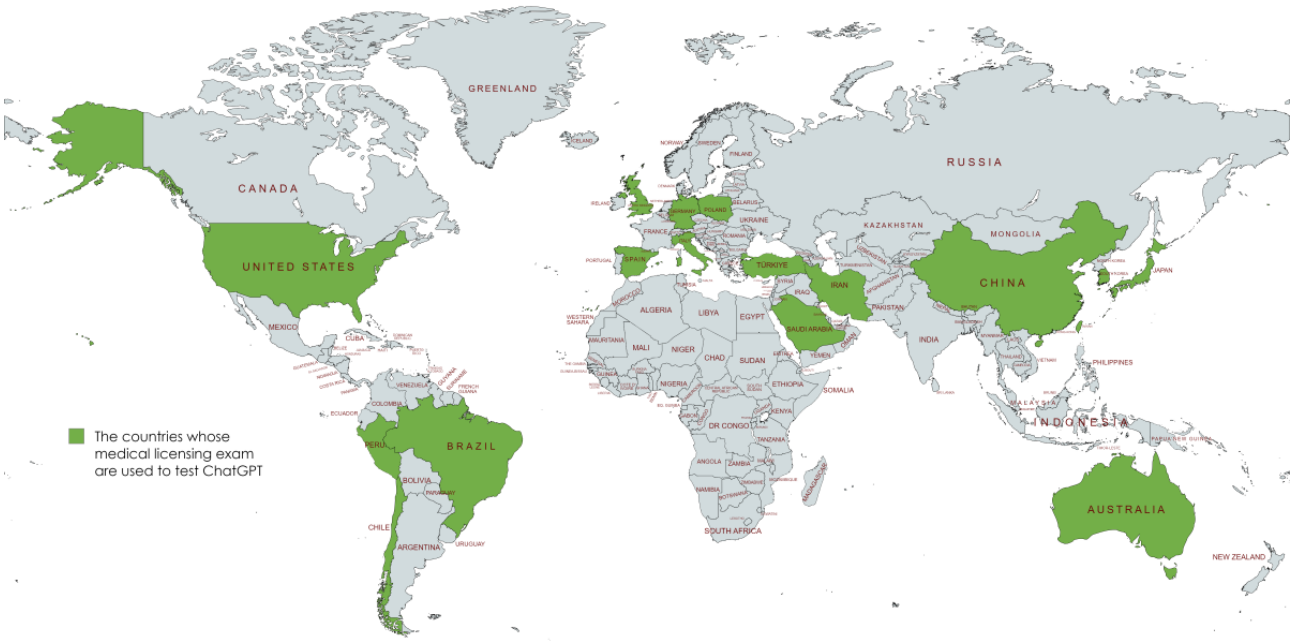

Of the 45 included studies, 29 tested the performance of GPT-4, and 26 tested the performance of GPT-3.5. A total of 14 studies tested both GPT-4 and GPT-3.5. In addition, 4 studies tested the GPT-3, one tested the InstructGPT, and one tested the ChatGPT Plus.

Regarding the countries and languages of the medical licensing examination questions used to test ChatGPT, 11 studies used examinations from an English-speaking country. Of the 34 medical licensing examinations of non–English-speaking countries, 22 used only the native language for testing, three translated the original language into English, and 9 used both the original and translated English questions.

All 45 studies included MCQs, with 4 studies including open-ended questions, 1 study including calculation questions, and 1 study including patient history inquiry questions.

Qualitative Analyses

Regarding the performance of ChatGPT on passing the medical licensing examination, among the 26 studies testing GPT-3.5, 6 reported that GPT-3.5 passed the medical licensing examination, and 4 reported satisfactory performance, making up 38.5% (10/26) of the total. In the remaining studies, 1 was unclear, and 15 did not pass. Among the 29 studies testing GPT-4, 17 reported that GPT-4 passed the medical licensing examination, and 9 reported satisfactory performance, making up 89.7% (26/29) of the total. In the remaining studies, 1 was unclear and 2 did not pass (Figure 4). For the other ChatGPT models, among the 4 studies testing the GPT-3 performance, 2 did not pass, 1 was unclear, and 1 showed a satisfactory performance. The studies that tested GPT-4 with Vision (GPT-4V, which is specifically designed for image tasks), InstructGPT, and ChatGPT Plus showed the following results: passed, did not pass, and did not pass.

XSL•FO
RenderX

Figure 4. Performance of ChatGPT on passing the medical licensing examination.

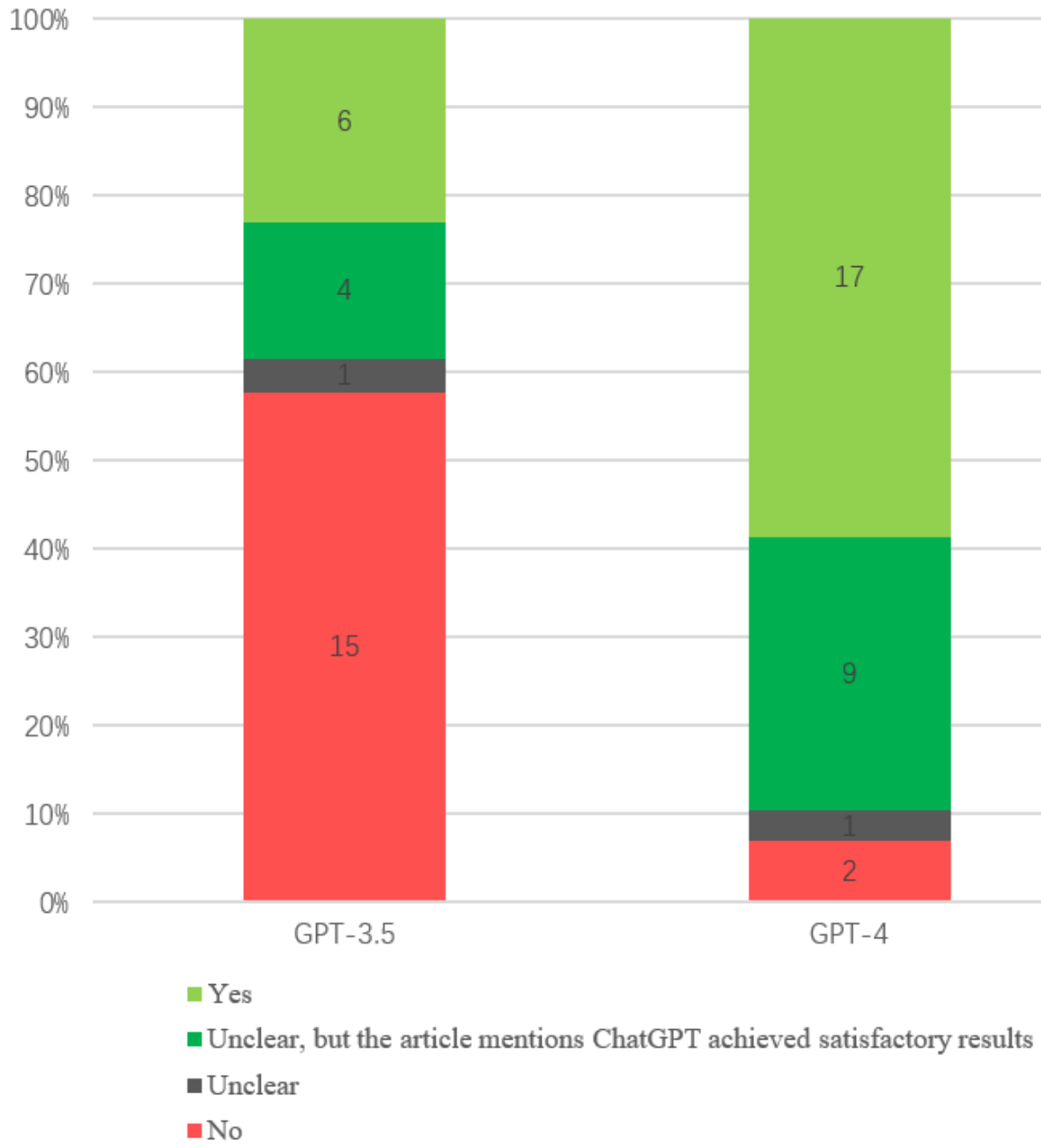

Regarding the performance of ChatGPT compared with medical students, 14 of 45 studies compared GPT-3.5's performance with medical students, and 17 of 45 compared GPT-4's performance with that of medical students. Four studies showed that GPT-3.5 surpassed medical students, accounting for 28.6% (4/14) of the studies. A total of 13 studies showed that GPT-4 surpassed medical students, accounting for 76.5% (13/17) of the studies (Figure 5). For the other ChatGPT models, 1 study showed that GPT-3 surpassed medical students, while another showed that it performed worse. One study indicated that InstructGPT performed worse than the students.

XSL•FO
RenderX

Figure 5. Performance of ChatGPT compared with medical students.

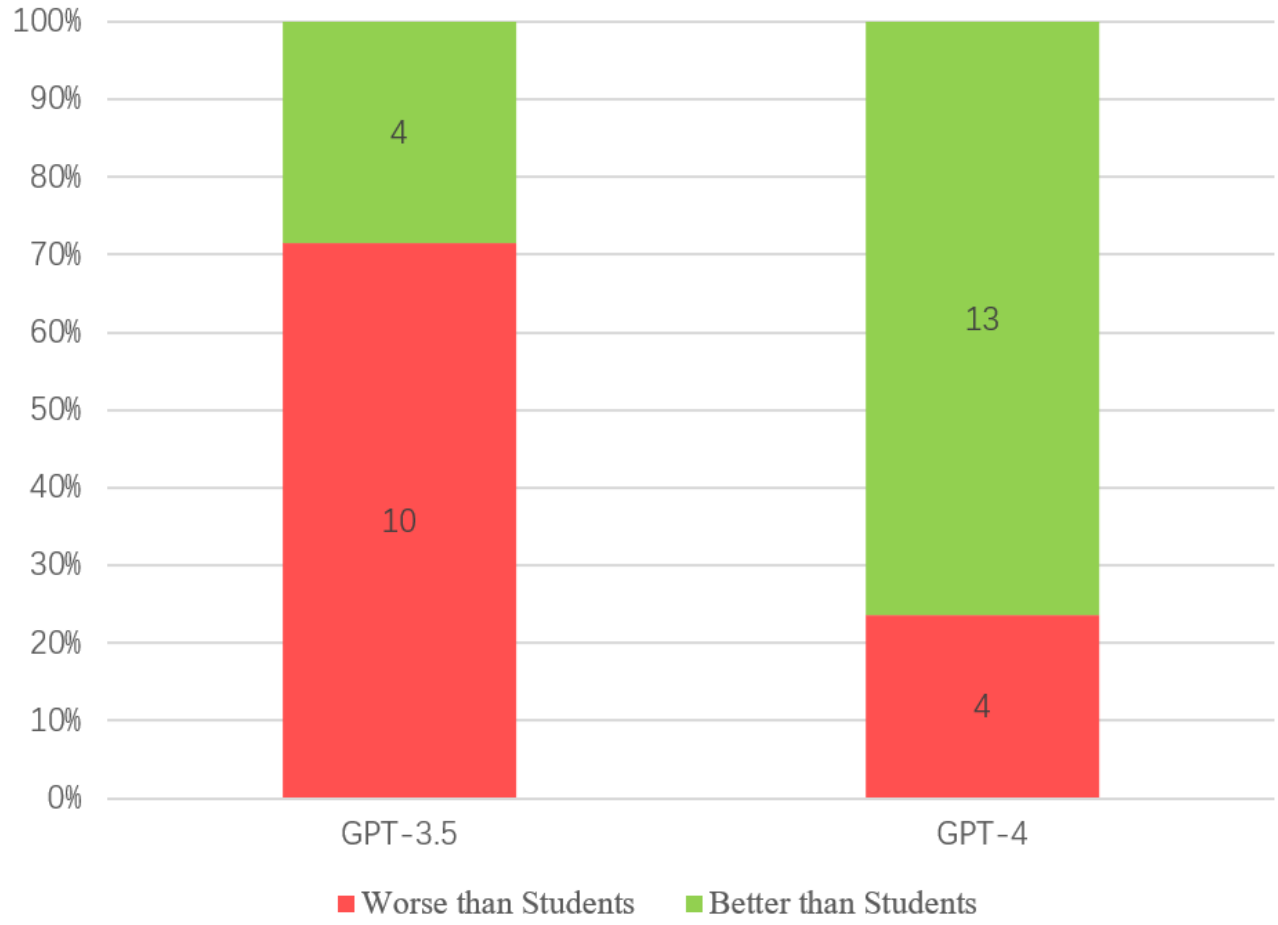

We also compared ChatGPT's performance in the original language and English-translated questions of the same non-English medical licensing examination. In studies of medical licensing examinations in non–English-speaking countries, 9 used both the original language and English-translated questions to test ChatGPT's performance, with 8 reporting comparative results (Table 2). Overall, for GPT-4, translating the original language into English had a limited effect on improving the performance. The accuracy improvement ranged from 0.17% to 8.65%, with 6 studies showing an accuracy increase of less than 5%. However, compared with GPT-4, GPT-3.5 showed significant improvement when tested in English in 4 studies. In 2 of these studies, GPT-3.5's accuracy was more than 20% higher in English than in the original language.

Table 2. ChatGPT's performance in original language and English-translated questions [18,23,26,30,41,45,47,57].

Author (year) and reference	GPT-3.5 accuracy rate		GPT-4 accuracy rate	
	Original language, n/n (%)	English-translated, n/n (%)	Original language, n/n (%)	English-translated, n/n (%)
Fang et al (2023) [18]	Untested	Untested	197/260 (75.77%)	201/260 (77.31%)
Guillen-Grima et al (2023) [23]	115/182 (63.2)	121/182 (66.5)	158/182 (86.8)	160/182 (87.9)
Jang et al (2023) [26]	Untested	Untested	Unclear (51.8)	Unclear (60.5)
Khorshidi et al (2023) [30]	Untested	Untested	161/198 (81.3)	167/198 (84.3)
Rosoł at al (2023) [41]	320.5/585 (54.8)	353/585 (60.3)	465.5/585 (79.6)	466.5/585 (79.7)
Tong at al (2023) [45]	Untested	Untested	130/160 (81.3)	138/160 (86.3)
Wang at al (2023) [47]	56/100 (56)	76/100 (76)	84/100 (84)	86/100 (86)
Keshtkar at al (2023) [57]	394/1105 (35.7)	687/1105 (61.4)	Untested	Untested

A total of 2 and 3 studies examined the correlation between GPT-3.5 or GPT-4 performance and the length of the question text, respectively. Both studies on GPT-3.5 showed a significant correlation between performance and the length of the question text; the longer the question text, the poorer the performance of GPT-3.5 [33,39]. In contrast, none of the 3 studies on GPT-4

XSL•FO
RenderX

found a significant difference in performance between long- and short-text questions [23,37,50].

A total of 8 studies examined the correlation between the difficulty of the questions and ChatGPT's performance. A total of 7 studies indicated that both GPT-4 and GPT-3.5 performed worse on difficult questions than easier ones [21,23,30,33,41,44,49]. Only 1 study showed that the difficulty of the questions did not affect GPT-4's performance. However, in this study, the difficulty was subjectively rated by 3 medical students rather than using official difficulty ratings [45].

Regarding ChatGPT's performance with and without optimized prompts, in our review of 45 papers, 13 stated that researchers provided ChatGPT with prompts before asking questions. Most of these prompts were designed to help ChatGPT better understand its task, such as "You are now an experienced clinician; please answer the following questions" or "You are a medical student, and we will be using medical licensing examination questions to test you; please provide your best answers." Researchers have not analyzed or elaborated on the impact of these task understanding prompts on ChatGPT's performance. However, 3 studies used optimized prompts [19,26,35]. A Korean study used 4 kinds of optimized prompts, including annotating Chinese terms in traditional Korean medicine, translating the instruction and question into English, providing examination-optimized instructions, and using self-consistency in the prompt. The results showed that ChatGPT's accuracy increased from 51.82% to 66.18% with optimized prompts [26]. In the other 2 studies, questions that ChatGPT initially answered incorrectly without prompts were reasked with optimized prompts, such as "Are you sure? Pretend to be a junior doctor with expertise in clinical practice and examination solving and retry" or "Could you double-check the answer?" ChatGPT could correctly answer up to 88.9% and 84% of these questions, respectively [19,35]. For task understanding prompts, we conducted a subgroup analysis and meta-regression to examine whether they affected ChatGPT's performance.

Regarding the capability of ChatGPT in answering image-based MCQs, 4 studies have reported the performance of ChatGPT in image-based MCQs. Three tested GPT-4, and 1 compared GPT-4 and GPT-4V [16,23,38,55]. In a UK study, GPT-4 achieved an accuracy rate of 100% (3/3) for the image-based MCQs correctly [16]. In a Spanish study, the accuracy rate of GPT-4 for image-based MCQs in Spanish was 13%, and the accuracy rate was 26% after translating the image-based MCQs into English, twice as high as in Spanish [23]. Japanese researchers tested GPT-4's performance on image-based MCQs that provided both images and text and on image-based MCQs that provided only text. The rate of correctness was 68% (73/108) when both images and text were provided, and 72% (78/108) when only text was provided [38]. Researchers in Chile compared the performance of ChatGPT-4 and ChatGPT-4V in image-based MCQs. The accuracy rates of GPT-4 and GPT-4V for image-based MCQs were 76.7% and 70%, respectively [55].

Regarding the performance of ChatGPT on questions other than MCQs, 4 studies compared ChatGPT's performance on open-ended questions versus MCQs. Among them, 2 showed that ChatGPT performed significantly worse on open-ended questions than on MCQs [3,19], 1 showed slightly better performance on open-ended questions, and another asked ChatGPT 10 short questions, all of which received an "A" grade [28,56]. In a study using calculation questions from the Japanese medical licensing examination, ChatGPT's performance on calculation questions was significantly worse than that of MCQs [24]. In a study using patient history inquiry questions from the Chinese medical licensing examination to assess medical students' clinical skills, ChatGPT passed the test and scored higher than the average medical student, achieving satisfactory performance [53].

Meta-Analysis

We conducted a meta-analysis of the integrated accuracy of GPT-3.5 and GPT-4 in medical licensing examinations. The accuracy of the meta-analysis was displayed in Multimedia Appendix 3. A total of 25 studies reporting the accuracy of GPT-3.5 and 29 studies reporting the accuracy of GPT-4 were included in this meta-analysis. Owing to significant heterogeneity (GPT-3.5: I^2=95% and GPT-4 I^2=93%), both groups were analyzed using a random-effects model.

The integrated accuracy for GPT-3.5 was 58% (95% CI 53-63; P<.01), and the integrated accuracy for GPT-4 was 81% (95% CI 78-84; P<.01; Figures S1 and S2 in Multimedia Appendix 4).

Meta-Regression and Subgroup Analysis

We divided studies with GPT-3.5 and GPT-4 in Figures S1 and S2 in Multimedia Appendix 4 into 3 subgroups, respectively.

Subgroup 1 divided the studies into those using medical licensing examinations from English-speaking countries to test ChatGPT and those using examinations from non–English-speaking countries with a native language. Subgroup 2 categorized studies based on whether they used prompts to test ChatGPT or not. In Figures S3 and S4 in Multimedia Appendix 4, "yes" indicates the use of prompts, while "no" indicates the absence of prompts. Subgroup 3 categorized studies according to the "flow and timing" evaluation in Table 1, with "low risk" forming 1 category and "unclear" and "high risk" forming another. In Figures S5 and S6 in Multimedia Appendix 4, "yes" means "low risk," implying that ChatGPT's performance might not be affected by the testing date and source date. "No" means "high risk" and "unclear," implying that ChatGPT's performance might be influenced by the testing date and source date. We conducted meta-regression and subgroup analyses for all subgroups to examine potential sources of heterogeneity and compare performances.

In subgroup analysis of subgroup 1, because of significant heterogeneity (GPT-3.5 tested in medical licensing examinations of English-speaking countries: I^2=80%, GPT-3.5 tested in original language examinations of non–English-speaking countries: I^2=96%, GPT-4 tested in medical licensing examinations of English-speaking countries: I^2=69%, and GPT-4 tested in original language examinations of non–English-speaking countries: I^2=93%), all 4 groups were analyzed using a random-effects model.

XSL•FO
RenderX

The integrated accuracy for GPT-3.5 in examinations from English-speaking countries was 57% (95% CI 52-62; $P<.01$), and in examinations from non–English-speaking countries with original languages, it was 58% (95% CI 52-64; $P<.01$). No statistically significant differences were observed ($P=.72$; Figure S7 in Multimedia Appendix 4).

For GPT-4, the integrated accuracy in examinations from English-speaking countries was 86% (95% CI 82%-89%; $P<.01$), and in examinations from non–English-speaking countries with original languages, it was 80% (95% CI 76-83; $P<.01$). Statistically significant differences were observed between the results ($P=.02$; Figure S8 in Multimedia Appendix 4).

In the subgroup analysis of subgroup 2, because of significant heterogeneity (GPT-3.5 in subgroup "yes:" $I^2=92\%$, GPT-3.5 in subgroup "no:" $I^2=95\%$, GPT-4 in subgroup "yes:" $I^2=68\%$, and GPT-4 in subgroup "no:" $I^2=94\%$), all 4 groups were analyzed using a random-effects model.

The integrated accuracy for GPT-3.5 in examinations with prompts was 68% (95% CI 57-77; $P<.01$), and in examinations without prompts, it was 54% (95% CI 50-59; $P<.01$). Statistically significant differences were observed between the results ($P=.03$; Figure S3 in Multimedia Appendix 4).

The integrated accuracy for GPT-4 in examinations with prompts was 85% (95% CI 83-88; $P<.01$), and in examinations without prompts, it was 79% (95% CI 75-82; $P<.01$). Statistically significant differences were observed between the results ($P<.01$; Figure S4 in Multimedia Appendix 4).

In the subgroup analysis of subgroup 3, because of significant heterogeneity (GPT-3.5 in subgroup "yes:" $I^2=96\%$, GPT-3.5 in subgroup "no:" $I^2=92\%$, GPT-4 in subgroup "yes:" $I^2=71\%$, and GPT-4 in subgroup "no:" $I^2=95\%$), all 4 groups were analyzed using a random-effects model.

The integrated accuracy for studies in which GPT-3.5's performance may be influenced by testing date and source date was 55% (95% CI 51-60; $P<.01$), and in studies in which GPT-3.5's performance may not be influenced, it was 62% (95% CI 53-71; $P<.01$). No statistically significant differences were observed ($P=.19$; Figure S5 in Multimedia Appendix 4).

The integrated accuracy for studies in which GPT-4's performance may be influenced by testing date and source date was 80% (95% CI 75-83; $P<.01$), and in studies in which GPT-4's performance may not be influenced, it was 83% (95% CI 80-86; $P<.01$). No statistically significant differences were observed ($P=.12$; Figure S6 in Multimedia Appendix 4).

Regarding the meta-regression results for all subgroups, the use of prompts is likely to be a source of potential heterogeneity and showed a significant effect on the accuracy rates of GPT-3.5 and GPT-4 (subgroup 2), as indicated by an estimated regression coefficient of 0.54 ($P=.01$) and 0.46 ($P=.02$), respectively. Meta-regression of subgroups 1 and 3 did not show statistically significant effects on accuracy rates (all $P>.05$; Table 3).

Table 3. Meta-regression results of 3 subgroups of GPT-3.5 and GPT-4.

Version	Estimated regression coefficient	*P* value
GPT-3.5		
Subgroup 1	–0.03	.91
Subgroup 2	0.54	.01
Subgroup 3	0.28	.19
GPT-4		
Subgroup 1	–0.39	.08
Subgroup 2	0.46	.02
Subgroup 3	0.25	.18

Publication Bias

No publication bias was detected among the included studies, as indicated by the funnel plots (Figure S9 in Multimedia Appendix 4).

Sensitivity Analyses

We used a random effects model to assess the impact of excluding individual studies on overall effects. The sensitivity analysis plot showed that no single study significantly affected the overall meta-analysis results. This demonstrates the robustness of the meta-analysis results (Figure S10 in Multimedia Appendix 4).

Power Analysis

We conducted post hoc power analysis for the main groups and subgroups using the results of the random effects model (Table 4). Subgroup 1 of GPT-3.5 had a power of 0.17. In this subgroup, we believe the sample size is adequate. The low power might be due to 2 main reasons. First, the intergroup difference is minimal, with effect sizes being very close (58% and 57%). Second, the data may have high heterogeneity ($I^2=80\%$ and 96%). In the main group and other subgroups, the power was 1 or close to 1, indicating sufficient power to detect the anticipated effect size with the given sample size for the random effects model.

XSL•FO
RenderX

Table 4. Power analysis results of main groups and subgroups.

Versions and groups		Power
GPT-3.5	Main group (integrated accuracy rate in Figure S1 in Multimedia Appendix 4)	1
GPT-4	Main group (integrated accuracy rate in Figure S2 in Multimedia Appendix 4)	1
GPT-3.5		
	Subgroup 1	0.17
	Subgroup 2	1
	Subgroup 3	1
GPT-4		
	Subgroup 1	1
	Subgroup 2	1
	Subgroup 3	0.98

Discussion

Principal Findings

Our systematic review and meta-analysis are the first to comprehensively evaluate the performance of all versions of ChatGPT across various medical licensing examination environments. Overall, GPT-4 significantly outperformed GPT-3.5; however, there are still some issues that make it difficult to use in medical education at this stage.

Regarding the accuracy of ChatGPT on MCQs, while 2 previous studies conducted meta-analyses that yielded accuracy rates of 61% and 56%, respectively, we noted that these accuracy rates reflected the performance of all versions of ChatGPT without differentiation by version [54,56]. Our review found that GPT-4 achieved an integrated accuracy rate of 81% for MCQs in medical licensing examinations, passing nearly all tested examinations and surpassing the average performance of medical students in three-quarters of the tests. In contrast, GPT-3.5 achieved an integrated accuracy rate of 58%, failing to pass more than half of the medical examinations and surpassing the average performance of medical students in only 4 of 14 tests. Therefore, regarding accuracy rate, passing rate, and comparison with medical students, GPT-4 significantly surpassed GPT-3.5.

In medical licensing examinations from non–English-speaking countries, translating the original language questions into English significantly improved GPT-3.5's performance but did not affect GPT-4's performance. This indicates that GPT-4 has a much higher proficiency in languages other than English than GPT-3.5. However, based on the results of subgroup analysis for comparing GPT-3.5 and GPT-4 in medical licensing examinations from English-speaking and non–English-speaking countries, we found that GPT-4 performed better in English-speaking countries. In contrast, GPT-3.5 showed no performance difference between examinations from English-speaking and non–English-speaking countries.

Additionally, based on the results of qualitative analysis and subgroup analysis, we found that both "optimized prompts" and "task understanding prompts" could significantly improve ChatGPT's performance. When using prompts, the accuracy rates of GPT-3.5 and GPT-4 were 68% and 85%, respectively, which were significantly higher than the accuracy rates of 54% and 79% without prompts.

The testing date and source date of each study were not sources of potential heterogeneity and did not significantly affect the performance of ChatGPT.

Challenge of Using ChatGPT in Medical Education

First, although the AI hallucinations of GPT-4 have significantly been reduced compared to earlier versions, GPT-4 still generates incorrect information because the data used to train these models is not always correct [65]. We observed that in all tests of GPT-4, only 2 instances achieved an accuracy rate above 90%. The only example of a perfect accuracy rate was in the UK study, in which GPT-4 correctly answered all 20 questions [16]. However, the number of questions used in this test was significantly lower than those used in other studies. We believe that this demonstrates ChatGPT's potential for future use in medical education but does not imply that medical students can rely on ChatGPT to acquire medical knowledge or prepare for examinations. Traditional sources of medical knowledge, such as medical school courses and textbooks, are completely reliable. However, because most professional medical knowledge exists in book form [50] and medical expertise on the internet is not always reliable [66], the medical knowledge that ChatGPT currently holds is not entirely accurate. In this context, if medical students rely on ChatGPT as a trusted source of expertise and acquire incorrect medical knowledge, the reliability of their knowledge and skills is significantly compromised. This is unacceptable in the medical field, as it directly impacts human lives. Therefore, GPT-4 passing medical licensing examinations does not imply that it can be used as a source of knowledge in medical education.

Previous studies have noted that the responses generated by GPT-3.5 are nondeterministic and random [67-69]. This study found that although the stability of GPT-4 has significantly improved compared to that of GPT-3.5, it still exhibits a degree of randomness in its outputs. Although GPT-4 achieved an overall accuracy of 81% across all tests, it only scored 52% on the Korean medical licensing examination, even lower than the overall accuracy of GPT-3.5 (58%) [25]. In addition, in 4 studies using Japanese medical licensing examination questions,

XSL•FO
RenderX

although GPT-4 passed 3 of the tests, it only achieved an accuracy of 67% in one and did not pass [23,38,44,51]. Furthermore, the use of optimized prompts and the difficulty of the questions can affect ChatGPT's performance stability. If millions of medical students use ChatGPT for learning, this randomness could be significantly magnified and affect their learning outcomes.

Moreover, different countries' medical policies, and cultural, ethical, and unique local traditional medical knowledge pose significant challenges for ChatGPT [70]. Regarding varying medical policies and ethics, a Chinese study mentioned that abortion is prohibited in the United States but allowed in certain circumstances in China [48]. Although euthanasia is legal in many countries, it is illegal in Japan. ChatGPT chose the option of euthanasia in the Japanese medical licensing examination [25]. ChatGPT may struggle to adapt to localized medical policies and ethics. In addition, East Asian countries still use local traditional medicine (eg, Chinese medicine), and most local traditional medicine learning materials are written in the native languages. These materials might not be accessible on the internet and included in ChatGPT's training data set, making it difficult for ChatGPT to provide accurate answers to such topics [18,26,50,54].

In the evaluation of image-based questions, we observed significant variations in the performance of GPT-4, with accuracy rates ranging from 13% to 100% [16,23,38,55]. However, there were only 3 questions in which GPT-4 achieved 100% accuracy, which is too small a sample size to demonstrate its proficiency in handling image-based questions [16]. In addition, a study from Japan tested the performance of ChatGPT when provided with images and text versus text only. Surprisingly, ChatGPT performed better when given only text than when provided with both images and text [38]. Similarly, Chile found that GPT-4V, designed explicitly for image tasks, performed worse on image-based questions than GPT-4 [55]. We believe that studies testing the ChatGPT's performance on image-based questions are limited at this stage. Therefore, comprehensive and reliable conclusions cannot be drawn. Consequently, using ChatGPT for image-based medical education is extremely risky.

Finally, human teachers usually recognize their knowledge limitations when faced with uncertain questions and correct their mistakes by consulting resources. However, the fatal issue with ChatGPT is that, owing to the nature of AI language models, it can provide detailed and logically sound explanations for incorrect answers [24,40,44]. Given ChatGPT's authoritative writing style, students are likely to believe and memorize the incorrect information provided by ChatGPT [71].

Limitations

This systematic review did not include studies on the performance of ChatGPT in various medical specialty examinations, dental licensing examinations, pharmacy examinations, and other medical-related assessments. Future studies should review the performance of ChatGPT in these specific medical fields. Studies published in languages other than English were excluded from the systematic review. This may omit the literature that tests the performance of ChatGPT on non–English-speaking medical licensing examinations.

Conclusions

A total of 45 studies on the performance of different versions of ChatGPT in medical licensing examinations were included in this systematic review. GPT-4 achieved an overall accuracy rate of 81%, significantly surpassing GPT-3.5, and, in most cases, passed the medical examinations, outperforming the average scores of medical students. Thus, GPT-4 demonstrates considerable potential for future use in medical education. However, because the knowledge of ChatGPT is not entirely accurate and its performance can be inconsistent, and because of the challenges posed by differing medical policies and knowledge across countries, we believe that GPT-4 is not yet suitable for use in medical education.

Acknowledgments

This work was supported by Japan Society for Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) grant (24KJ0830).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Query strings of Web of Science, Scopus, and PubMed.
[DOCX File , 17 KB - jmir_v26i1e60807_app1.docx]

Multimedia Appendix 2
Evaluation framework used in this systematic review.
[DOCX File , 19 KB - jmir_v26i1e60807_app2.docx]

Multimedia Appendix 3
General characteristics of included studies.
[XLSX File (Microsoft Excel File), 17 KB - jmir_v26i1e60807_app3.xlsx]

XSL•FO
RenderX

Multimedia Appendix 4
Supplementary figures.
[PPTX File , 1105 KB - jmir_v26i1e60807_app4.pptx]

Multimedia Appendix 5
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 31 KB - jmir_v26i1e60807_app5.docx]

References

1. ChatGPT. OpenAI. URL: https://chat.openai.com/chat [accessed 2024-02-12]
2. Khlaif ZN, Mousa A, Hattab MK, Itmazi J, Hassan AA, Sanmugam M, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ 2023;9:e47049 [FREE Full text] [doi: 10.2196/47049] [Medline: 37707884]
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
4. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the situational judgement test-a professional dilemmas-based examination for doctors in the United Kingdom. JMIR Med Educ 2023;9:e48978 [FREE Full text] [doi: 10.2196/48978] [Medline: 37548997]
5. Rahman MM, Watanobe Y. ChatGPT for education and research: Opportunities, threats, and strategies. Appl. Sci 2023;13(9):5783 [FREE Full text] [doi: 10.3390/app13095783]
6. Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X. Exploring the performance of ChatGPT-4 in the Taiwan audiologist qualification examination: preliminary observational study highlighting the potential of AI chatbots in hearing care. JMIR Med Educ 2024;10:e55595 [FREE Full text] [doi: 10.2196/55595] [Medline: 38693697]
7. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The Potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
8. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023;25(1):bbad493. [doi: 10.1093/bib/bbad493] [Medline: 38168838]
9. Gödde D, Nöhl S, Wolf C, Rupert Y, Rimkus L, Ehlers J, et al. A SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis of ChatGPT in the medical literature: concise review. J Med Internet Res 2023;25:e49368 [FREE Full text] [doi: 10.2196/49368] [Medline: 37865883]
10. Tsang R. Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev 2023;10:23821205231178449 [FREE Full text] [doi: 10.1177/23821205231178449] [Medline: 37255525]
11. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 2023;25:e48966 [FREE Full text] [doi: 10.2196/48966] [Medline: 37490317]
12. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2023. [doi: 10.1002/ase.2270] [Medline: 36916887]
13. Price T, Lynn N, Coombes L, Roberts M, Gale T, de Bere SR, et al. The international landscape of medical licensing examinations: a typology derived from a systematic review. Int J Health Policy Manag 2018;7(9):782-790 [FREE Full text] [doi: 10.15171/ijhpm.2018.32] [Medline: 30316226]
14. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How does ChatGPT perform on the italian residency admission national exam compared to 15,869 medical graduates? Ann Biomed Eng 2024;52(4):745-749. [doi: 10.1007/s10439-023-03318-7] [Medline: 37490183]
15. Aljindan FK, Al Qurashi AA, Albalawi IAS, Alanazi AMM, Aljuhani HAM, Falah Almutairi F, et al. ChatGPT conquers the Saudi medical licensing exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus 2023;15(9):e45043 [FREE Full text] [doi: 10.7759/cureus.45043] [Medline: 37829968]
16. Armitage RC. Performance of generative pre-trained transformer-4 (GPT-4) in membership of the royal college of general practitioners (MRCGP)-style examination questions. Postgrad Med J 2024;100(1182):274-275. [doi: 10.1093/postmj/qgad128] [Medline: 38142282]
17. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform 2023;30(1):e100815 [FREE Full text] [doi: 10.1136/bmjhci-2023-100815] [Medline: 38081765]
18. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-english national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023;2(12):e0000397 [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]

XSL•FO
RenderX

19. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ 2023;9:e48039 [FREE Full text] [doi: 10.2196/48039] [Medline: 37768724]
20. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 performance on USMLE step 1 style questions and its implications for medical education: a comparative study across systems and disciplines. Med Sci Educ 2024;34(1):145-152. [doi: 10.1007/s40670-023-01956-z] [Medline: 38510401]
21. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
22. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R. Performance of ChatGPT-4 in answering questions from the Brazilian national examination for medical degree revalidation. Rev Assoc Med Bras (1992) 2023;69(10):e20230848 [FREE Full text] [doi: 10.1590/1806-9282.20230848] [Medline: 37792871]
23. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023;13(6):1460-1487 [FREE Full text] [doi: 10.3390/clinpract13060130] [Medline: 37987431]
24. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: differences in the amount of information per medical field. Int J Med Inform 2023;180:105283. [doi: 10.1016/j.ijmedinf.2023.105283] [Medline: 37931432]
25. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on stage 1 of the Taiwanese medical licensing exam. Digit Health 2024;10:20552076241233144 [FREE Full text] [doi: 10.1177/20552076241233144] [Medline: 38371244]
26. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. PLOS Digit Health 2023;2(12):e0000416 [FREE Full text] [doi: 10.1371/journal.pdig.0000416] [Medline: 38100393]
27. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Arztebl Int 2023;120(21):373-374 [FREE Full text] [doi: 10.3238/arztebl.m2023.0113] [Medline: 37530052]
28. Kao YS, Chuang WK, Yang J. Use of ChatGPT on Taiwan's examination for medical doctors. Ann Biomed Eng 2024;52(3):455-457. [doi: 10.1007/s10439-023-03308-9] [Medline: 37432530]
29. Kataoka Y, Yamamoto-Kataoka S, So R, Furukawa TA. Beyond the pass mark: accuracy of ChatGPT and bing in the national medical licensure examination in Japan. JMA J 2023;6(4):536-538 [FREE Full text] [doi: 10.31662/jmaj.2023-0043] [Medline: 37941716]
30. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, et al. Application of ChatGPT in multilingual medical education: how does ChatGPT fare in 2023's Iranian residency entrance examination. Informatics in Medicine Unlocked 2023;41:101314. [doi: 10.1016/j.imu.2023.101314]
31. Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and new bing in an Australian medical licensing examination. Med J Aust 2023;219(5):237. [doi: 10.5694/mja2.52061] [Medline: 37528548]
32. Kleinig O, Kovoor JG, Gupta AK, Bacchi S. Universal precautions required: artificial intelligence takes on the Australian Medical Council's trial examination. Aust J Gen Pract 2023;52(12):863-865 [FREE Full text] [doi: 10.31128/AJGP-02-23-6708] [Medline: 38049136]
33. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ 2024;10:e51148 [FREE Full text] [doi: 10.2196/51148] [Medline: 38180782]
34. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom medical licensing assessment. Front Med (Lausanne) 2023;10:1240915 [FREE Full text] [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]
35. Lin SY, Chan PK, Hsu WH, Kao CH. Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination. Digit Health 2024;10:20552076241237678 [FREE Full text] [doi: 10.1177/20552076241237678] [Medline: 38449683]
36. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 2024;10:e50965 [FREE Full text] [doi: 10.2196/50965] [Medline: 38329802]
37. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States medical licensing examination. Med Teach 2024;46(3):366-372. [doi: 10.1080/0142159X.2023.2249588] [Medline: 37839017]
38. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese national medical licensing examination: evaluation study. JMIR Med Educ 2024;10:e54393 [FREE Full text] [doi: 10.2196/54393] [Medline: 38470459]

XSL•FO
RenderX

39. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023;102(32):e34673 [FREE Full text] [doi: 10.1097/MD.0000000000034673] [Medline: 37565917]
40. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, bing, and medical students in Germany. JMIR Med Educ 2023;9:e46482 [FREE Full text] [doi: 10.2196/46482] [Medline: 37665620]
41. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
42. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian medical residency exam. Ann Ist Super Sanita 2023;59(4):267-270 [FREE Full text] [doi: 10.4415/ANN_23_04_05] [Medline: 38088393]
43. Shang L, Xue M, Hou Y, Tang B. Can ChatGPT pass China's national medical licensing examination? Asian J Surg 2023;46(12):6112-6113 [FREE Full text] [doi: 10.1016/j.asjsur.2023.09.089] [Medline: 37775381]
44. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
45. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese national medical licensing examination. Front Med (Lausanne) 2023;10:1237432 [FREE Full text] [doi: 10.3389/fmed.2023.1237432] [Medline: 38020160]
46. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian national licensing medical examination: a cross-sectional study. J Educ Eval Health Prof 2023;20:30 [FREE Full text] [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
47. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
48. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese national medical licensing examination. J Med Syst 2023;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
49. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023;9:e52202 [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
50. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's family medicine board exam. J Chin Med Assoc 2023;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
51. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res 2023;7:e48023 [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
52. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med 2024;99(2):192-197. [doi: 10.1097/ACM.0000000000005549] [Medline: 37934828]
53. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-English-speaking outpatient clinic settings. Digit Health 2023;9:20552076231184091 [FREE Full text] [doi: 10.1177/20552076231184091] [Medline: 37434733]
54. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
55. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the performance of ChatGPT versions 3.5, 4, and 4 with vision in the Chilean medical licensing examination: observational study. JMIR Med Educ 2024;10:e55048 [FREE Full text] [doi: 10.2196/55048] [Medline: 38686550]
56. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
57. Keshtkar A, Hayat A, Atighi F, Ayare N, Keshtkar M, Yazdanpanahi P, et al. ChatGPT's performance on Iran's medical licensing exams. URL: https://doi.org/10.21203/rs.3.rs-3253417/v1 [accessed 2024-06-20]
58. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
59. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev 2024;11:23821205241238641 [FREE Full text] [doi: 10.1177/23821205241238641] [Medline: 38487300]
60. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform 2024;151:104620. [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]

XSL•FO
RenderX

61. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
62. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
63. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
64. Heath A. All the news from OpenAI's first developer conference. The Verge. 2023. URL: https://www.theverge.com/2023/11/6/23948619/openai-chatgpt-devday-developer-conference-news [accessed 2024-04-20]
65. Wong RSY, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023;9:e47274 [FREE Full text] [doi: 10.2196/47274] [Medline: 37988149]
66. Battineni G, Baldoni S, Chintalapudi N, Sagaro GG, Pallotta G, Nittari G, et al. Factors affecting the quality and reliability of online health information. Digit Health 2020;6:2055207620948996 [FREE Full text] [doi: 10.1177/2055207620948996] [Medline: 32944269]
67. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024;26:e56655 [FREE Full text] [doi: 10.2196/56655] [Medline: 38630520]
68. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. BMC Med Educ 2023;23(1):864 [FREE Full text] [doi: 10.1186/s12909-023-04832-x] [Medline: 37957666]
69. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024;10:e52483 [FREE Full text] [doi: 10.2196/52483] [Medline: 38598263]
70. Zhang G, Jin Q, Jered McInerney D, Chen Y, Wang F, Cole CL, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform 2024;153:104640. [doi: 10.1016/j.jbi.2024.104640] [Medline: 38608915]
71. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question
NBME: National Board of Medical Examiners
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
USMLE: United States Medical Licensing Examination

Edited by G Eysenbach, Q Jin; submitted 22.05.24; peer-reviewed by T Haze, J Pang, F Chen; comments to author 02.06.24; revised version received 11.06.24; accepted 15.06.24; published 25.07.24.

Please cite as:
Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T
Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e60807
URL: https://www.jmir.org/2024/1/e60807
doi:10.2196/60807
PMID:

©Mingxin Liu, Tsuyoshi Okuhara, XinYi Chang, Ritsuko Shirabe, Yuriko Nishiie, Hiroko Okada, Takahiro Kiuchi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient-Representing Population's Perceptions of GPT-Generated Versus Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment

Thomas Huang[1,2], BS; Conrad Safranek[1,2], BS; Vimig Socrates[2,3], MS; David Chartash[2,4], PhD; Donald Wright[1,2], MHS, MD; Monisha Dilip[1], MD; Rohit B Sangal[1], MBA, MD; Richard Andrew Taylor[1,2], MHS, MD

[1]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States

[2]Department for Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, United States

[3]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States

[4]School of Medicine, University College Dublin, National University of Ireland, Dublin, Ireland

Corresponding Author:
Richard Andrew Taylor, MHS, MD
Department of Emergency Medicine
Yale School of Medicine
333 Cedar Street
New Haven, CT, 06510
United States
Phone: 1 2034324771
Email: richard.taylor@yale.edu

Abstract

Background: Discharge instructions are a key form of documentation and patient communication in the time of transition from the emergency department (ED) to home. Discharge instructions are time-consuming and often underprioritized, especially in the ED, leading to discharge delays and possibly impersonal patient instructions. Generative artificial intelligence and large language models (LLMs) offer promising methods of creating high-quality and personalized discharge instructions; however, there exists a gap in understanding patient perspectives of LLM-generated discharge instructions.

Objective: We aimed to assess the use of LLMs such as ChatGPT in synthesizing accurate and patient-accessible discharge instructions in the ED.

Methods: We synthesized 5 unique, fictional ED encounters to emulate real ED encounters that included a diverse set of clinician history, physical notes, and nursing notes. These were passed to GPT-4 in Azure OpenAI Service (Microsoft) to generate LLM-generated discharge instructions. Standard discharge instructions were also generated for each of the 5 unique ED encounters. All GPT-generated and standard discharge instructions were then formatted into standardized after-visit summary documents. These after-visit summaries containing either GPT-generated or standard discharge instructions were randomly and blindly administered to Amazon MTurk respondents representing patient populations through Amazon MTurk Survey Distribution. Discharge instructions were assessed based on metrics of *interpretability of significance*, *understandability*, and *satisfaction*.

Results: Our findings revealed that survey respondents' perspectives regarding GPT-generated and standard discharge instructions were significantly (P=.01) more favorable toward GPT-generated return precautions, and all other sections were considered noninferior to standard discharge instructions. Of the 156 survey respondents, GPT-generated discharge instructions were assigned favorable ratings, "agree" and "strongly agree," more frequently along the metric of *interpretability of significance* in discharge instruction subsections regarding diagnosis, procedures, treatment, post-ED medications or any changes to medications, and return precautions. Survey respondents found GPT-generated instructions to be more *understandable* when rating procedures, treatment, post-ED medications or medication changes, post-ED follow-up, and return precautions. *Satisfaction* with GPT-generated discharge instruction subsections was the most favorable in procedures, treatment, post-ED medications or medication changes, and return precautions. Wilcoxon rank-sum test of Likert responses revealed significant differences (P=.01) in the *interpretability of significant* return precautions in GPT-generated discharge instructions compared to standard discharge instructions but not for other evaluation metrics and discharge instruction subsections.

Conclusions: This study demonstrates the potential for LLMs such as ChatGPT to act as a method of augmenting current documentation workflows in the ED to reduce the documentation burden of physicians. The ability of LLMs to provide tailored

XSL•FO
RenderX

instructions for patients by improving readability and making instructions more applicable to patients could improve upon the methods of communication that currently exist.

(J Med Internet Res 2024;26:e60336) doi:10.2196/60336

KEYWORDS

machine learning; artificial intelligence; large language models; natural language processing; ChatGPT; discharge instructions; emergency medicine; emergency department; discharge instructions; surveys and questionaries

Introduction

Discharge instructions serve as an essential bridge between hospital treatment and at-home recovery. Particularly in the emergency department (ED), where clinical team members have limited time to review the events of the ED encounter and follow-up recommendations with their patients, discharge instructions are critical to communicating information to patients and increasing adherence to follow-up care. Specifically, discharge instructions serve to inform the patient about key details such as their diagnosis, evaluations performed, preliminary diagnostic test results, medications, treatment plans, follow-up care, and reasons to return to the ED [1,2]. Improved patient understanding of discharge instructions and self-efficacy following discharge have been linked to improved health outcomes and decreased readmission rates [3-5]. Despite their important role, many patients leave the hospital with discharge instructions that are jargon filled and difficult to navigate, with studies showing that up to 88% of discharge instructions are unreadable to the population served [6-8]. General guidelines suggest that all health-related information provided by a physician to patients should be readable at a sixth-grade reading level [2,9]. Furthermore, the lack of personalization in many current discharge instructions misses an opportunity to better engage patients with understanding their care and diagnosis, which has been shown to be key to increasing patient health outcomes [10,11]. This communication gap drives nonadherence and higher rates of patient readmission [12,13]. Addressing this challenge of bridging physician and patient discharge communication is essential for improving patient health outcomes and the overall effectiveness of care delivery.

The recent advancement of large language models (LLMs), such as ChatGPT, offers a potential solution to the longstanding issue of inaccessible medical communication and the time-demanding nature of providing care in the ED [14,15]. Research on LLMs has demonstrated their broad capabilities in medical question answering, including passing the United States Medical Licensing Examinations [16]. These models can generate medical text that is accurate, informative, more readable, and even more empathetic [17-21]. By leveraging LLMs, there is the potential to create discharge instructions that are not only tailored to individual patients but also presented in a format that is engaging and easy to understand.

Recent research has underscored the feasibility of using LLMs for patient discharge instructions. The translation of 50 patient discharge summaries from the medical sublanguage to regular English by GPT-4 demonstrated significant improvement in objective readability scores [21]. This study also showed these LLM-generated summaries had a generally acceptable level of accuracy and completeness, per physician assessment. Other research has extended to specialties such as neurology and radiology, demonstrating that LLMs can digest complex medical text and produce summaries that are patient-centric and sufficiently comprehensive that physicians are comfortable releasing them to patients [22-24].

While this existing LLM research has laid a robust foundation, it has so far lacked a key evaluative component: integration of patient perspectives in a randomized unbiased comparison of LLM-generated text against the current standard of care. In addition, current investigations of leveraging LLMs to generate discharge instructions or summaries in the ED remain limited. We aim to address this gap by surveying a patient-representing population using a randomized blinded approach to compare LLM-generated, fully formatted discharge instructions to the standard discharge instructions.

Methods

Study Population and Setting

The study population was a convenience sample of adult patients (aged ≥18 years) taken from Amazon MTurk from December 31, 2023, to March 31, 2024; residing in the United States; with an approval rate (an indicator in Amazon MTurk of work quality) of >90%; and who were also categorized as Amazon MTurk Masters [25]. Research has shown that the majority of patients do not have medical training, have low health literacy, and have difficulty understanding medical jargon [26,27]. Thus, exclusion criteria for this population included respondents having any significant health care training background to best represent the general patient population. No Amazon MTurk survey respondents were allowed to repeat the survey; therefore, all responses are from unique respondents, reducing bias from individual responders and improving data quality.

Ethical Considerations

This study was approved by the institutional review board as an exemption (HIC #2000036301). Informed consent was obtained at the initiation of the survey. If consent was declined, the survey ended immediately. Participants were allowed to opt out of the survey at any time, and survey data only used fully completed surveys. Data were anonymized and the Amazon MTurk ID was used only for payment but not used for analysis. Participants were compensated US $2.62 for their participation if they completed the survey and met the basic quality control and attention check metrics. These details were clearly outlined in both the Amazon MTurk survey distribution description as well as the initial informed consent.

XSL•FO
RenderX

Synthesizing Fictional ED Encounters

A set of synthetic fictional ED notes were generated that represent 5 separate independent ED encounters (Multimedia Appendix 1). These notes were diverse in patients' age, sex, ethnicity, and clinical presentation. All ED notes were authored by an emergency medicine (EM) attending physician (DW) to emulate realistic ED settings and were rendered in multiple document styles. The EM physician note included a brief chief complaint statement, history of present illness, past medical history, physical examination, and assessment and plan. The nursing note contained summaries of the presentation and care performed.

ChatGPT Prompt Development

Initial prompts were developed based on the Society for Academic Emergency Medicine and Joint Commission guidelines for discharge instructions [1,2]. Prompts were applied to the 5 synthetic clinical note sets used for the final surveys (the final GPT prompt is available in Multimedia Appendix 1). Slight adjustments were made iteratively to improve the GPT-generated discharge instruction output, including changes the organization provided in the prompt, additional directive phrasing, and evaluation of output. Final prompts added to the synthetic clinical notes described above were deployed via a pipeline to automatically query and retrieve responses from OpenAI's GPT-4 for the final 5 clinical scenarios. Other strategies described in previous prompt engineering work were also included in this iterative process to improve GPT-4 recognition of the note and generative responses [28].

Discharge Instructions Development

Each clinical scenario represented by the synthetic clinical note was passed to GPT-4 with the final prompt generated by the methodology described above. Each of the 5 clinical scenarios was associated with 2 sets of discharge instructions for a total of 10 different variants of discharge instructions. These discharge instructions consist of 1 GPT-4–generated discharge instruction and 1 standard discharge instruction a patient would expect to receive from the ED without GPT-4–generated text. The standard discharge instructions (provided by RBS) are the standardized Epic After Visit Summary with complaint-specific discharge instructions from 1 of 2 institutionally contracted patient education resources (Elsevier Patient Education and UpToDate). These discharge instructions were then populated, through Epic's "Playground" environment, into a generated ED encounter so that an entire discharge instruction could be generated with the same overall formatting. The standard discharge instructions did not include physician-generated free text. All specific identifiers, such as phone numbers, insurance names, and trademarks, were censored out of the final generated discharge instructions.

Development, Recruitment, and Distribution

Surveys were created in Qualtrics to be distributed through the distribution service within Amazon MTurk, a reliable and widely used survey distribution service (the Qualtrics survey is available in Multimedia Appendix 1). These surveys were iteratively reviewed by ED physicians, residents, and medical students to ensure readability and ease of use. The survey was an open survey format, available to all Amazon MTurk Masters who also had an approval rating of >90%. The survey was listed on Amazon MTurk and was only visible to MTurkers that met the requirements. No other forms of advertising occurred other than the listing on Amazon MTurk.

Survey Administration

The survey was distributed on Amazon MTurk over a 3-month period, during which respondents were given a link to the complete survey built on Qualtrics. The survey structure included an initial explanation of what the survey responder would be presented with—1 clinical scenario described by a note that contains the pertinent information to the ED visit. The survey flow saw each respondent blindly randomized to 1 of 10 possible workflows. The respondent was initially presented with consent documentation, initial screening questions, and attention checks that must be passed for survey results to be accepted [29,30]. Respondents were then randomized to 1 of 5 random clinical scenarios where they responded to an additional quality control quiz including 4 multiple choice questions regarding the clinical scenario (Figure 1).

Figure 1. The Qualtrics survey design for discharge instruction randomized blind assessment by Amazon MTurk respondents. Survey respondents first passed an initial consent documentation, a screener for health care or medical background, and a series of attention checks. Respondents were then randomized to 1 of 5 possible clinical scenario and then randomized to view either the GPT-generated or standard version of discharge instructions. This respondent was then tasked to answer Likert-scale questions regarding the 3 metrics: interpretability of significance, understandability, and satisfaction, for each discharge instructions subsection, before 1 final attention check and the conclusion of the survey.

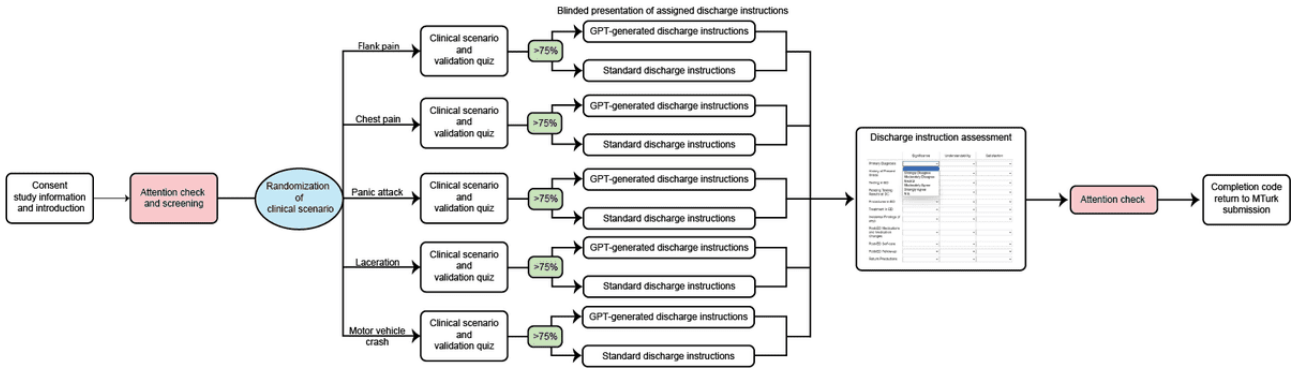

The survey respondent was then tasked to fill out a matrix of Likert-scale questions regarding the discharge instructions they read in the context of the clinical scenario (the clinical scenarios and discharge instructions are available in Multimedia Appendix

XSL•FO
RenderX

1). The y-axis included all components of the discharge instruction: diagnosis, history of the current problem or illness, testing received in the ED, any pending test results, procedures, treatment, incidental findings, post-ED medications or any changes to medications, post-ED self-care, post-ED follow-up, and return precautions. The x-axis of the matrix asked the survey respondent to rate on a traditional Likert scale on the 3 metrics of interpretability of significant findings or information, ease of understanding, and satisfaction concerning each of the discharge instruction subsections (Multimedia Appendix 2).

Attention checks and quality checks were included throughout the survey to ensure attention and increase the likelihood of high-quality data. Attention checks, such as asking the patient to answer simple but targeted questions with expected correct answers, are a method of ensuring survey respondents are reading questions and responding to the best of their ability [31]. A manual review of each round of distribution was performed to maintain data quality by assessing the percentage of attention checks passed per survey respondent and validating respondent understanding of the clinical scenario through a passing score (≥3/4, ≥75%) on the corresponding quiz. Respondents were allowed to go back to change their answers when necessary. Survey respondents' results were only retained for the final data set if they passed the content validation test (≥3/4, ≥75%) and failed no more than 1 of 4 attention checks to ensure a high-quality data set.

Data Analysis

Survey responses were extracted from Qualtrics and analyzed using Python (version 3.11.5; Python Software Foundation). Frequency counts of Likert responses to each metric and discharge instruction subsection were organized and visualized in clustered stacked multi-bar charts. The Wilcoxon rank-sum test was used to analyze for significant differences between Likert responses between GPT-generated and standard discharge instructions.

Results

Overview

From 253 unique initial views on Amazon MTurk, 222 (87.7%) passed the initial screening criteria. Of those 222 respondents, 187 (84.2%) respondents successfully completed the survey, and 156 (83.4%) of those 187 respondents passed all validation and attention checks, making them eligible for the final cohort (Figure 2).

XSL•FO
RenderX

Figure 2. The stepwise method toward exclusion of survey respondents. From an initial 253 views of the survey, 222 were not screened out by the initial screen of consent, not having a health care background, and not residing in the United States. Following the initial screener, 35 did not finish the entire survey. Surveys also had their validation quiz about the clinical scenario and attention checks graded. Of 187 completed surveys, 156 passed the validation quiz, and all respondents who passed the validation quiz also successfully answered all other attention checks correctly.

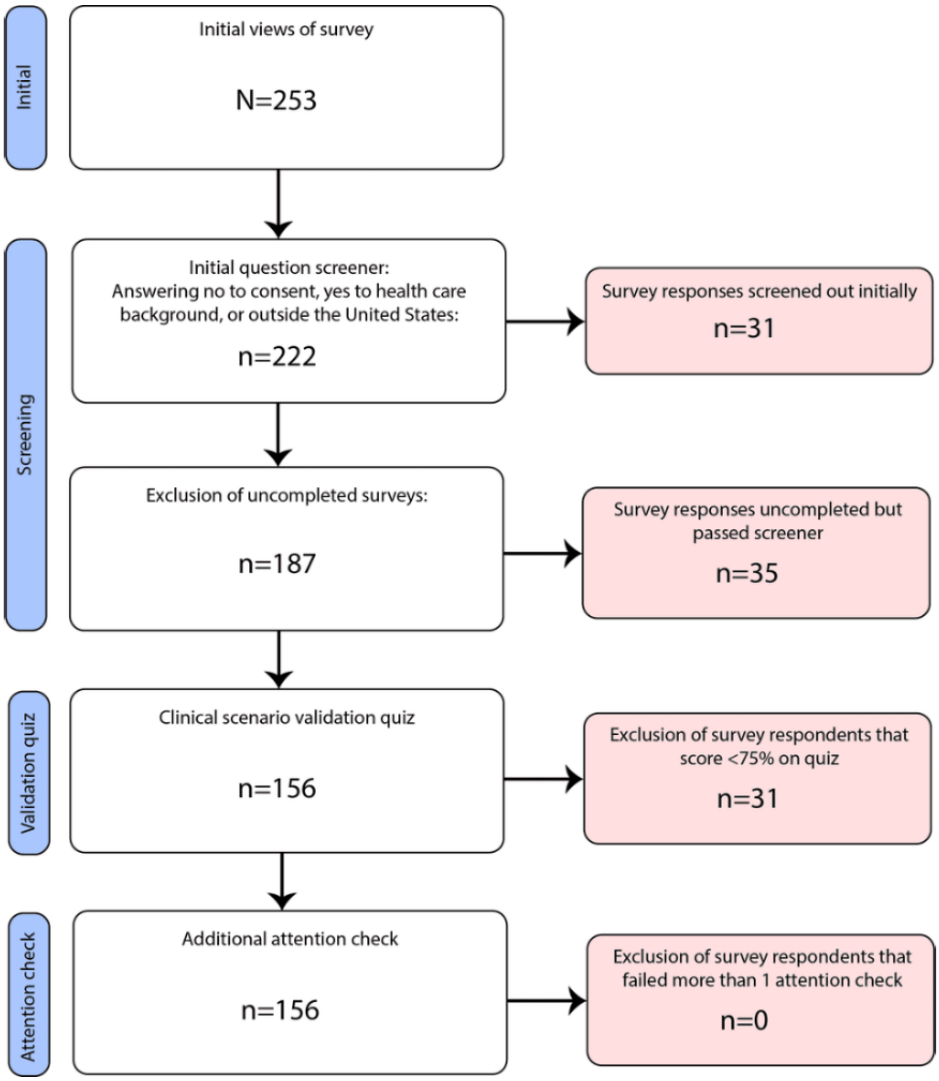

Domain 1: Interpretability of Significance

To assess the interpretability of the significance of the discharge instructions, each survey respondent was tasked to rate their agreement with the statement: "the information in the discharge instruction (subsection) effectively explains the interpretability of significance of the findings in a way that's personalized to me (the hypothetical recipient) and is easy to follow." They were asked to individually assess the interpretability of significance regarding each important subsection of the discharge instructions. Of note, the frequency of "agree" and "strongly agree" selected by survey respondents was greater across all GPT-generated discharge instructions' subsections with regard to the interpretability of significance except for pending test results and incidental findings.

Regarding "pending test results," 29% (24/83) of respondents rated favorably for their ability to interpret significant information in GPT-generated discharge instructions and 34% (25/73) rated favorably in standard discharge instructions. For incidental findings, 30% (25/83) and 34% (25/73) of respondents rated favorably for GPT-generated and standard discharge instructions, respectively. These 2 sections saw the greatest percentage of not applicable responses as well, with 39% (32/83) versus 45% (33/73) in pending testing results and 45% (37/83) versus 45% (33/73) for incidental findings, respectively, for GPT-generated and standard discharge instructions (Figure 3).

All other subsections of the GPT-generated discharge instructions were scored more favorably in terms of the interpretability of significance metric, with the most notable difference coming from interpreting the significance of diagnosis (GPT: 74/83, 89% vs standard: 58/73, 80%), procedures (GPT: 62/83, 75% vs standard: 45/73, 62%), treatment (GPT: 72/83, 87% vs standard: 48/73, 66%), post-ED medications or any changes to medications (GPT: 53/83, 64% vs standard: 36/73, 49%), and return precautions (GPT: 72/83, 87% vs standard: 50/73, 68%). However, these differences in ratings between GPT-generated and standard discharge instructions were not statistically significant (Table 1). Differences in average Likert score as graded on a numerical scale for the interpretability of significance are available in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 3. Agreement regarding each subsection of the discharge instructions about the prompt: "the information in the discharge instruction (subsection) effectively explains the significance of the findings in a way that's personalized to me (the hypothetical recipient) and is easy to follow." ED: emergency department; STD: standard.

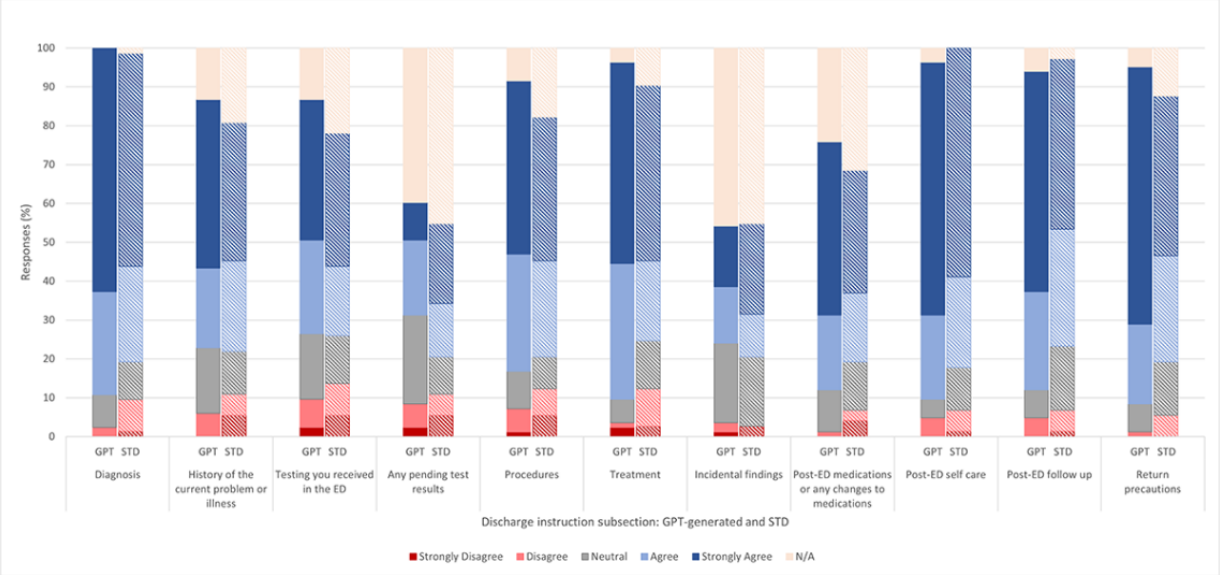

Table 1. The results of the Wilcoxon rank-sum test between GPT and standard discharge instructions across all 3 metrics of interpretability of significance, understandability, and satisfaction on a traditional Likert scale.

Discharge instruction subsection	Significance		Understandability		Satisfaction	
	Statistic	*P* value	Statistic	*P* value	Statistic	*P* value
Diagnosis	1.0889	.28	1.0288	.30	0.0799	.94
History of the current problem or illness	0.6777	.50	–0.5204	.60	0.4410	.66
Testing you received in the ED[a]	0.1992	.84	–0.4277	.67	–1.0055	.31
Any pending test results	–1.3601	.17	–1.1275	.26	–1.0330	.30
Procedures	0.7495	.45	1.1351	.26	1.1768	.24
Treatment	1.1914	.23	0.9122	.36	0.6776	.50
Incidental findings	–0.9949	.32	–0.9091	.36	–0.4280	.67
Post-ED medications or changes	1.5435	.12	0.4388	.66	1.0104	.31
Post-ED self-care	1.0812	.28	1.1214	.26	0.1954	.85
Post-ED follow-up	1.8320	.07	1.6280	.1035	0.8989	.37
Return precautions	2.5617	.01	1.4151	.16	1.9154	.06

[a]ED: emergency department.

Domain 2: Understanding

The understanding was assessed by their agreement with the phrase: "the information in the discharge instruction (subsection) is written such that it is presented in a clear and straightforward manner that is easily comprehensible." Although not statistically significantly different, there are some differences to note between the discharge instructions authored by GPT compared to the standard discharge instructions from the ED (Table 1).

While survey respondents recorded a greater interpretation of significance from diagnosis, their understanding was very similar between GPT (74/83, 89.2% agree or greatly agree) and standard discharge instructions (64/73, 87.6%). However, the other subcategories that saw a greater percentage of survey respondents rating their understanding of the information in specific discharge instruction subsections favorably came from procedures (GPT: 67/83, 81% vs standard: 45/73, 62%), treatment (GPT: 71/83, 86% vs standard: 50/73, 68%), post-ED medications or medication changes (GPT: 57/83, 69% vs standard: 42/73, 58%), post-ED follow-up (GPT: 72/83, 87% vs standard: 56/73, 77%), and return precautions (GPT: 71/83, 86% vs standard: 56/73, 77%; Figure 4). It is important to note, however, that these differences did not achieve statistically significant differences (Table 1). Differences in average Likert score as graded on a numerical scale for understandability are available in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 4. This is a clustered stacked multibar chart of the 5 possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions to assess understanding concerning the prompt: "the information in the discharge instruction (subsection) is written such that it is presented clearly and straightforwardly that is easily comprehensible." ED: emergency department; STD: standard.

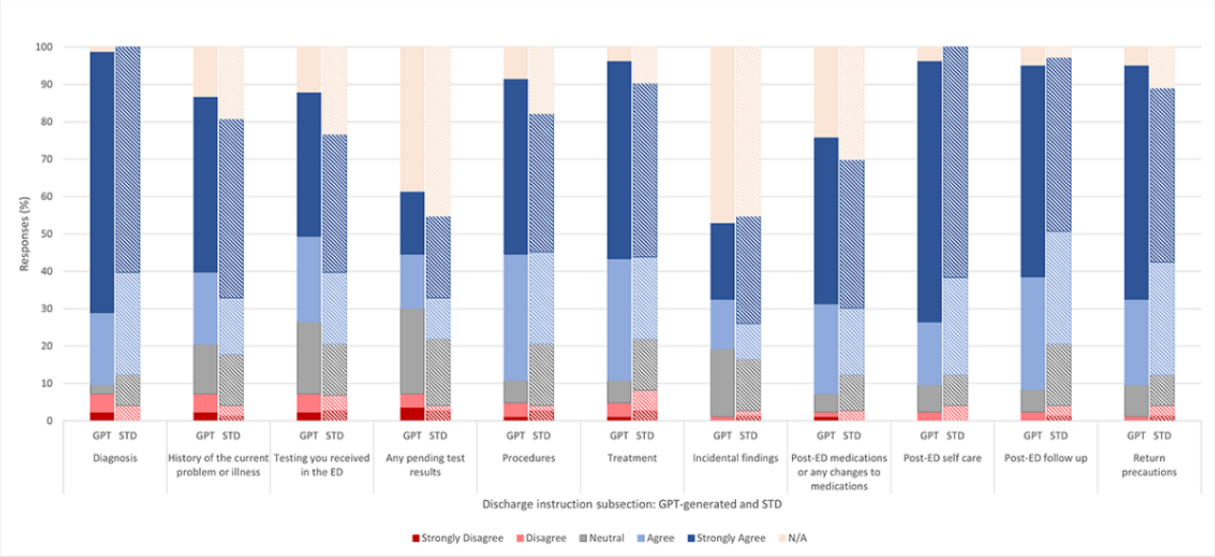

Domain 3: Satisfaction

Satisfaction was assessed based on their agreement with the phrase: :the information in the discharge instructions (subsection) fulfills your personal expectations of the quality you would expect to receive in an ED setting.: Within satisfaction, the GPT-generated discharge instructions once again saw greater frequency of agree and strongly agree across the majority of discharge instruction subsections, signaling positive sentiment toward GPT-generated discharge instructions in comparison to the standard discharge instructions (Figure 5). A greater percentage of survey respondents rated their satisfaction with GPT-generated discharge instruction subsections the most favorably in procedures (GPT: 63/83, 76% vs standard: 40/73, 55%), treatment (GPT: 71/83, 86% vs standard: 50/73, 68%), post-ED medications or medication changes (GPT: 52/83, 63% vs standard: 39/73, 53%), and return precautions (GPT: 69/83, 83% vs standard: 52/73, 71%).

However, there was no statistically significant difference between the ratings regarding GPT-generated and standard discharge instructions (Table 1). The sections in which survey respondents were slightly more dissatisfied with GPT-generated discharge instructions were tests received in the ED (GPT: 44/83, 53% vs standard: 41/73, 56%; rated satisfied or very satisfied), pending test results (GPT: 25/83, 30% vs standard: 25/73, 34%), and incidental findings (GPT: 26/83, 31% vs standard: 25/73, 34%). Differences in average Likert score as graded on a numerical scale for satisfaction are available in Multimedia Appendix 5.

Figure 5. This is a clustered stacked multibar chart of the five possible ratings (strongly disagree, disagree, neutral, agree, and strongly agree) regarding each subsection of the discharge instructions to assess satisfaction in relation to the prompt: "the information in the discharge instructions (subsection) fulfill your personal expectations of the quality you would expect to receive in an ED setting." ED: emergency department; STD: standard.

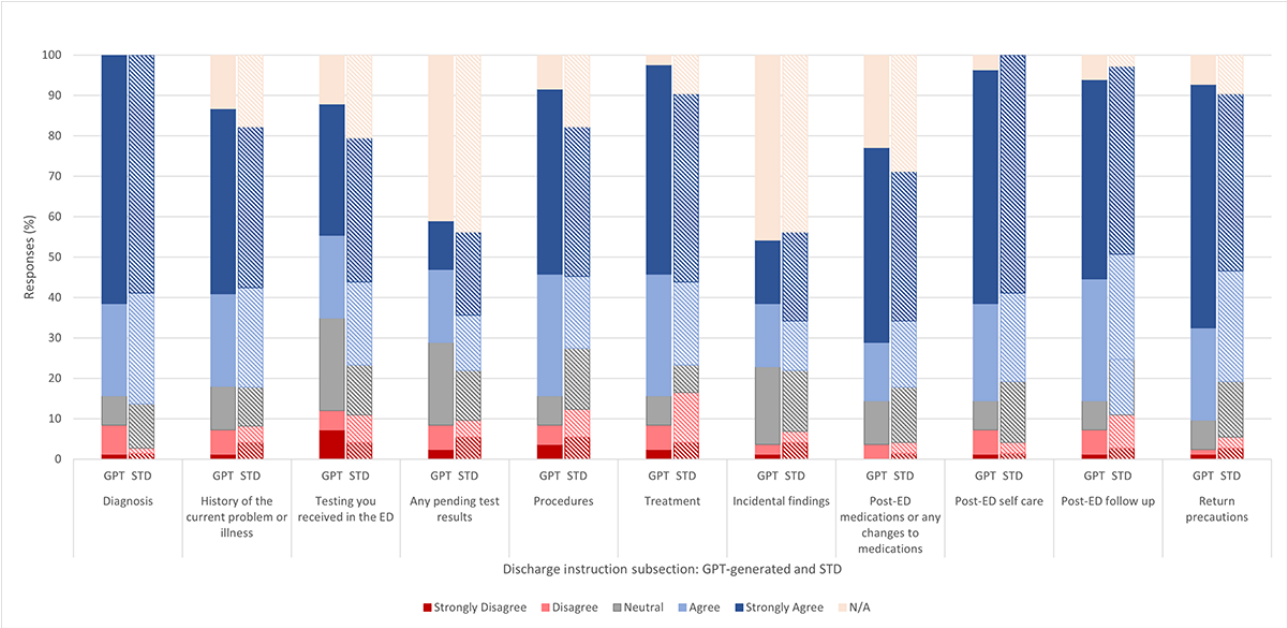

XSL•FO
RenderX

Comparative Analysis

We conducted a Wilcoxon rank-sum test to compare the Likert results across GPT and standard discharge instruction cohorts. Due to the nonparametric, ordinal nature of the data, the Wilcoxon rank-sum test was best suited to better understand the results.

Notably, the interpretability of significant findings or information was considered significantly (P=.01) more favorable in GPT-generated return precautions compared to standard discharge instructions. No other specific subsection of the discharge instructions was rated by respondents to statistically significantly differ between the GPT-generated and standard discharge instructions on a standard Likert scale (Table 1). Other discharge instruction subsections were generally positive, indicating a slight but not significant favorability toward GPT-generated discharge instructions (Table 1). The exceptions were consistent negative test statistics in pending test results and incidental findings across all 3 metrics.

Discussion

Principal Findings

We present a novel assessment of a general patient population's perception of discharge instructions, a key method of communication between patients and physicians. Although the GPT-generated discharge instructions received higher ratings across all 3 metrics of interpretability of significant information, understandability, and satisfaction in several key subsections of the discharge instructions, the standard discharge instructions were not found to be statistically significantly different from the discharge instructions authored by a LLM (GPT-4) except for the interpretability of significant information in return precautions (Table 1). The GPT-generated return precautions frequently included personalized and specific information, such as symptoms to be aware of or progression of current symptoms from the current presentation, corresponding to the primary diagnosis from the ED encounter scenario. Prior literature demonstrates that return precautions are oftentimes one of the most crucial discharge instruction subsections and the section that has the greatest deficit in recall and understanding [32]. This personalized and clearly presented information demonstrates 1 method in which GPT-generated discharge instructions could serve as an adjunct for the current workflow within discharge summaries to help improve patient understanding and interpretation of key information.

The similar and even slightly higher ratings for GPT-generated discharge instructions suggest the potential for LLMs to be able to serve as possible adjuncts or interventions in improving discharge instructions or throughput in the ED. LLMs have the ability to generate not only accurate but also detailed and personalized discharge instructions that may improve patients' abilities to interpret the significance of complex medical information through simplification, which can not only play an important role in reducing documentation burden for the physicians but also possibly ease patient transitions out of the hospital through augmented and possibly improved forms of documentation and medical information [33,34].

Despite the lack of statistically significant superiority for most metrics and subsections, the performance of LLM-generated discharge instructions compared to standard methods suggests that the former are not inferior either. Several subsections of the GPT-generated discharge instructions were even rated favorably in comparison to the standard discharge instructions. Notably, information regarding procedures and treatments received in the ED, as well as important follow-up information including post-ED medication or medication changes and return precautions, had a greater frequency of favorable ratings across all 3 metrics, interpretability of significance, understandability, and satisfaction, as compared to standard discharge instructions. This equivalence opens up practical applications for LLMs in the discharge process, particularly in terms of efficiency and reducing clinician workload. EDs often have bottlenecks in patient flow that may benefit from LLM integration to expedite the discharge process. Since LLM-generated instructions can be produced rapidly and tailored to individual patient profiles, they have the potential to decrease the time clinicians spend writing these instructions, and there may be value in reducing the discharge-to-departure time [34,35]. Providing a method to quickly author initial drafts or starting points for discharge instructions may free up clinician time for other high-value tasks. LLMs had trending higher scores post-ED care and follow-up, which often have to be customized for each patient. By providing an initial starting point for writing discharge instructions that already include detailed pertinent information that can pull in prior notes and potentially other data fields in future implementations, ChatGPT generation of notes has the potential to reduce documentation burden and lead to improved and more personalized discharge instructions. Furthermore, the adaptability of LLMs could lead to more dynamic discharge instructions. For instance, discharge instructions could be automatically updated as soon as a clinician finalizes their notes, ensuring that patients receive the most current and relevant information without delay.

Comparison to Prior Work

Prior methods of generating fictional cases, specifically in orthopedics, and synthesizing discharge documentation by ChatGPT in comparison to a standard showed that they were comparable in quality as assessed by expert panels [36]. Our study found similar results, with patient-representing survey respondents finding that discharge instructions generated using ED synthetic notes were of similar quality to standard ED discharge instructions. Barak-Corren et al [14] evaluated pediatric EM attending perceptions regarding summaries regarding completeness, accuracy, efficiency, readability, and overall satisfaction. Our findings represented similar results from the patient perspective, with survey respondents assessing GPT-generated discharge instructions with greater positive sentiment than negative sentiment.

Limitations

This study poses several limitations. The participants recruited in this study were compensated to answer to their best ability in these surveys. They were asked to assume the role of a patient in an ED, receive information regarding a clinical scenario, and respond as if they were a patient receiving the corresponding

XSL•FO
RenderX

discharge instructions. Survey respondents were asked to complete rigorous attention checks and quality control metrics to ensure they properly understood the clinical scenario, but it is possible that survey respondents did not accurately assess the discharge instructions following the attention checks and other quality control metrics. The task presented to survey respondents was not simple and was a complex task.

The generated discharge instructions were improved through an iterative prompt engineering process that may have assessed and improved quality specifically for the ED setting. Due to the use of synthetic notes in this pipeline, the generalizability of these methods may be limited based on the specialty of discharge instructions as well as important variations of discharge instructions from provider to provider. The discharge instructions were also generated from documentation such as the provider history and physical, which may not consistently be documentation that is fully finished by the time discharge instructions are written. In addition, the prompt development occurred over the same set of notes and clinical scenarios that were used for the final discharge instructions. These limitations may act as barriers to the implementation of the proposed pipeline in current workflows in the ED.

In addition, it is important to note that Amazon MTurk survey respondents are not the same subpopulation presenting to EDs, only a convenience sample representative of the general population. However, the demographics of Amazon MTurk survey respondents are similar to that of patients presenting to EDs in select features in Connecticut, which is considered to be one of the most representative states of the United States [37]. Notably, Amazon MTurk survey respondents are 57% female vs 58% female among ED presentations. Annual household incomes among Amazon MTurk survey respondents also closely resemble the general US population in income ranges under US $150,000 per the Current Population Survey. For income ranges greater than US $150,000, only 4.92% of MTurkers fall within this range while Current Population Survey reports 15.47% for the generation population. However, Amazon MTurk survey respondents skew younger compared to ED presentations, with 66.5% of Amazon MTurk survey respondents aged between 18 and 40 years and less than 40% of ED presentations aged between 18 and 44 years [38]. Survey respondents slightly overrepresent White non-Hispanic populations with approximately 79.9% compared to 66% of ED presentations of being White non-Hispanic in Connecticut. In addition, Amazon MTurk survey respondents are historically better educated compared to the general population, an important consideration when assessing the readability of discharge instructions [39]. Although not perfectly representative, the responses and perspectives of Amazon MTurkers are still important to consider as potential stakeholders in the reception of discharge instructions and a unique perspective external to physician and medical team perspectives. The MTurk population has been shown to be more representative of the general population than any other web-based survey and consistently produces reliable results [40]. However, it is still important to consider replicating similar methods with real patient populations in the ED.

Future Directions

Our study explored the initial stages of patient-representing perceptions of using LLM-generated discharge instructions in the ED setting. Due to initial constraints on protected health information, synthetic clinical scenarios, clinician notes, and nursing notes were used in developing discharge instructions in this survey format. Future work can focus on the integration of LLMs into existing EHR infrastructure to leverage real patient notes with Health Insurance Portability and Accountability Act–compliant methods of ChatGPT such as through Doximity GPT or the Azure OpenAI system.

The current scope of our work only looked into 5 different presenting chief concerns due to the synthetic note constraint. The next steps could look toward expanding the scope of the use of LLMs in discharge instructions, such as integrating multicultural and multilingual personalization into discharge instructions as well as expanding the amount of teaching and patient education embedded into the discharge instructions [41,42]. The iterative process of GPT-4-based discharge instruction development can also be improved with embedding methods of patient feedback, such as structured interviews, as demonstrated in prior research of other forms of discharge instructions that could further help develop an improved generation of tailored discharge instructions [43].

Conclusions

The results of this study indicate the promising capabilities for generative LLMs such as GPT-4 in improving methods of health care communication. Future research in the use of LLMs in ED workflows can use this in real-world applications to assess the perceptions of LLM-generated medical documentation by health care staff as well as a possible survey distribution within EDs and real clinical settings that our survey distribution aimed to imitate.

Acknowledgments

This publication was made possible by the Yale School of Medicine Fellowship for Medical Student Research.

Conflicts of Interest

RAT received grant funding and speaking fees from Beckman Coulter.

Multimedia Appendix 1

XSL•FO
RenderX

Clinical scenarios and GPT-generated and standard discharge instructions, ChatGPT prompt used to generate the GPT-generated discharge instructions, and Qualtrics survey.
[DOCX File , 17 KB - jmir_v26i1e60336_app1.docx]

Multimedia Appendix 2
Respondents to the Likert-style matrix survey were tasked to answer with regard to 3 metrics: interpretability of significance, understandability, and satisfaction for the discharge instruction subsection. ED: emergency department.
[PNG File , 62 KB - jmir_v26i1e60336_app2.png]

Multimedia Appendix 3
Differences in averages of Likert scores between GPT and standard discharge instructions for the evaluation of interpretability of significance by discharge instruction subsection.
[PNG File , 45 KB - jmir_v26i1e60336_app3.png]

Multimedia Appendix 4
Differences in averages of Likert scores between GPT and standard discharge instructions for the evaluation of ease of understanding by discharge instruction subsection.
[PNG File , 43 KB - jmir_v26i1e60336_app4.png]

Multimedia Appendix 5
Differences in averages of Likert scores between GPT and standard discharge instructions for the evaluation of satisfaction by discharge instruction subsection.
[PNG File , 42 KB - jmir_v26i1e60336_app5.png]

References

1. 1 JCOAOHO, Hospital ASS, Intents :HAS. The Joint Comission. 2006. URL: https://www.jointcommission.org/standards/ [accessed 2024-07-13]
2. Discharge Instructions. URL: https://www.saem.org/about-saem/academies-interest-groups-affiliates2/cdem/for-students/online-education/m3-curriculum/disposition/discharge-instructions [accessed 2024-04-18]
3. Engel KG, Heisler M, Smith DM, Robinson CH, Forman JH, Ubel PA. Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand? Ann Emerg Med 2009;53(4):454-461.e15. [doi: 10.1016/j.annemergmed.2008.05.016] [Medline: 18619710]
4. Buckley BA, McCarthy DM, Forth VE, Tanabe P, Schmidt MJ, Adams JG, et al. Patient input into the development and enhancement of ED discharge instructions: a focus group study. J Emerg Nurs 2013;39(6):553-561. [doi: 10.1016/j.jen.2011.12.018] [Medline: 22575702]
5. DeSai C, Janowiak K, Secheli B, Phelps E, McDonald S, Reed G, et al. Empowering patients: simplifying discharge instructions. BMJ Open Qual 2021;10(3):e001419 [FREE Full text] [doi: 10.1136/bmjoq-2021-001419] [Medline: 34521621]
6. Burns ST, Amobi N, Chen JV, O'Brien M, Haber LA. Readability of patient discharge instructions. J Gen Intern Med 2022;37(7):1797-1798 [FREE Full text] [doi: 10.1007/s11606-021-06988-y] [Medline: 34240287]
7. Choudhry AJ, Baghdadi YM, Wagie AE, Habermann EB, Heller SF, Jenkins DH, et al. Readability of discharge summaries: with what level of information are we dismissing our patients? Am J Surg 2016;211(3):631-636 [FREE Full text] [doi: 10.1016/j.amjsurg.2015.12.005] [Medline: 26794665]
8. Unaka NI, Statile A, Haney J, Beck AF, Brady PW, Jerardi KE. Assessment of readability, understandability, and completeness of pediatric hospital medicine discharge instructions. J Hosp Med 2017;12(2):98-101 [FREE Full text] [doi: 10.12788/jhm.2688] [Medline: 28182805]
9. Clarke C, Friedman SM, Shi K, Arenovich T, Monzon J, Culligan C. Emergency department discharge instructions comprehension and compliance study. CJEM 2005;7(1):5-11. [doi: 10.1017/s1481803500012860] [Medline: 17355647]
10. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med 2012;27(5):520-526 [FREE Full text] [doi: 10.1007/s11606-011-1931-2] [Medline: 22127797]
11. Hibbard JH, Mahoney ER, Stock R, Tusler M. Do increases in patient activation result in improved self-management behaviors? Health Serv Res 2007;42(4):1443-1463 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00669.x] [Medline: 17610432]
12. Becker C, Zumbrunn S, Beck K, Vincent A, Loretz N, Müller J, et al. Interventions to improve communication at hospital discharge and rates of readmission: a systematic review and meta-analysis. JAMA Netw Open 2021;4(8):e2119346. [doi: 10.1001/jamanetworkopen.2021.19346] [Medline: 34448868]

XSL•FO
RenderX

13. Doyle SK, Rippey JC, Jacques A, Rea AM, Kaiser BN, Miller SM, et al. Effect of personalised, mobile-accessible discharge instructions for patients leaving the emergency department: a randomised controlled trial. Emerg Med Australas 2020;32(6):967-973. [doi: 10.1111/1742-6723.13516] [Medline: 32383347]
14. Barak-Corren Y, Wolf R, Rozenblum R, Creedon JK, Lipsett SC, Lyons TW, et al. Harnessing the power of generative AI for clinical summaries: perspectives from emergency physicians. Ann Emerg Med 2024;84(2):128-138. [doi: 10.1016/j.annemergmed.2024.01.039] [Medline: 38483426]
15. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
17. Nayak A, Alkaitis MS, Nayak K, Nikolov M, Weinfurt KP, Schulman K. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. JAMA Intern Med 2023;183(9):1026-1027. [doi: 10.1001/jamainternmed.2023.2561] [Medline: 37459091]
18. Eppler MB, Ganjavi C, Knudsen JE, Davis RJ, Ayo-Ajibola O, Desai A, et al. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. Urol Pract 2023;10(5):436-443. [doi: 10.1097/UPJ.0000000000000428] [Medline: 37410015]
19. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
20. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
21. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3):e240357 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
22. Hartman VC, Bapat SS, Weiner MG, Navi BB, Sholle ET, Campion TR. A method to automate the discharge summary hospital course for neurology patients. J Am Med Inform Assoc 2023;30(12):1995-2003. [doi: 10.1093/jamia/ocad177] [Medline: 37639624]
23. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023;309(2):e232561. [doi: 10.1148/radiol.232561] [Medline: 37987662]
24. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. Semin Ophthalmol 2023;38(5):503-507. [doi: 10.1080/08820538.2023.2209166] [Medline: 37133418]
25. Blokdyk G. Amazon Mechanical Turk: Projects-Ready. North Charleston, SC: Createspace Independent Publishing Platform; 2018.
26. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res 2022;22(1):1148 [FREE Full text] [doi: 10.1186/s12913-022-08527-9] [Medline: 36096793]
27. Safeer RS, Keenan J. Health literacy: the gap between physicians and patients. Am Fam Physician 2005;72(3):463-468 [FREE Full text] [Medline: 16100861]
28. Basics of prompting. URL: https://www.promptingguide.ai/introduction/basics [accessed 2024-02-29]
29. Berinsky AJ, Margolis MF, Sances MW. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. Am J Political Sci 2014;58(3):739-753. [doi: 10.1111/ajps.12081]
30. Mellis AM, Bickel WK. Mechanical Turk data collection in addiction research: utility, concerns and best practices. Addiction 2020;115(10):1960-1968 [FREE Full text] [doi: 10.1111/add.15032] [Medline: 32135574]
31. Hauser DJ, Schwarz N. Attentive turkers: mturk participants perform better on online attention checks than do subject pool participants. Behav Res Methods 2016;48(1):400-407. [doi: 10.3758/s13428-015-0578-z] [Medline: 25761395]
32. Mahajan M, Hogewoning JA, Zewald JJA, Kerkmeer M, Feitsma M, van Rijssel DA. The impact of teach-back on patient recall and understanding of discharge information in the emergency department: the emergency teach-back (EM-TeBa) study. Int J Emerg Med 2020;13(1):49 [FREE Full text] [doi: 10.1186/s12245-020-00306-9] [Medline: 32972361]
33. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ 2024;10:e51391 [FREE Full text] [doi: 10.2196/51391] [Medline: 38349725]
34. Nguyen J, Pepping CA. The application of ChatGPT in healthcare progress notes: a commentary from a clinical and research perspective. Clin Transl Med 2023;13(7):e1324 [FREE Full text] [doi: 10.1002/ctm2.1324] [Medline: 37394880]
35. Shahab O, El Kurdi B, Shaukat A, Nadkarni G, Soroush A. Large language models: a primer and gastroenterology applications. Therap Adv Gastroenterol 2024;17:17562848241227031 [FREE Full text] [doi: 10.1177/17562848241227031] [Medline: 38390029]

XSL•FO
RenderX

36. Sánchez-Rosenberg G, Magnéli M, Barle N, Kontakis MG, Müller AM, Wittauer M, et al. ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases. Acta Orthop 2024;95:152-156 [FREE Full text] [doi: 10.2340/17453674.2024.40182] [Medline: 38597205]
37. Kolko J. 'Normal America' is not a small town of White people. 2016. URL: https://fivethirtyeight.com/features/normal-america-is-not-a-small-town-of-white-people/ [accessed 2016-04-18]
38. Demographics of people on Amazon Mechanical Turk. CloudResearch. 2020. URL: https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/ [accessed 2024-04-18]
39. Ipeirotis P. Demographics of Mechanical Turk . URL: https://www.ipeirotis.com/wp-content/uploads/2012/02/CeDER-10-01.pdf [accessed 2024-06-03]
40. Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med 2018;33(4):533-538 [FREE Full text] [doi: 10.1007/s11606-017-4246-0] [Medline: 29302882]
41. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv Preprint posted online on February 8, 2023. [doi: 10.48550/arXiv.2302.04023]
42. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta Radiol 2023;1(2):100022 [FREE Full text] [doi: 10.1016/j.metrad.2023.100022] [Medline: 37901715]
43. Horstman MJ, Mills WL, Herman LI, Cai C, Shelton G, Qdaisat T, et al. Patient experience with discharge instructions in postdischarge recovery: a qualitative study. BMJ Open 2017;7(2):e014842 [FREE Full text] [doi: 10.1136/bmjopen-2016-014842] [Medline: 28228448]

Abbreviations

ED: emergency department
EM: emergency medicine
LLM: large language model

Edited by A Mavragani; submitted 08.05.24; peer-reviewed by MH Wilkinson, M Wright; comments to author 31.05.24; revised version received 07.06.24; accepted 22.06.24; published 02.08.24.

Please cite as:
Huang T, Safranek C, Socrates V, Chartash D, Wright D, Dilip M, Sangal RB, Taylor RA
Patient-Representing Population's Perceptions of GPT-Generated Versus Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment
J Med Internet Res 2024;26:e60336
URL: https://www.jmir.org/2024/1/e60336
doi:10.2196/60336
PMID:

©Thomas Huang, Conrad Safranek, Vimig Socrates, David Chartash, Donald Wright, Monisha Dilip, Rohit B Sangal, Richard Andrew Taylor. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Performance of Large Language Models in Patient Complaint Resolution: Web-Based Cross-Sectional Survey

Lorraine Pei Xian Yong[1,2,3*], MBBS, MMed; Joshua Yi Min Tung[4*], MBBS, MPH; Zi Yao Lee[1,2,3*], MBBS, MPH; Win Sen Kuan[1,2,3*], MBBS, MCI; Mui Teng Chua[1,2,3*], MBBS, MMed, MPH

[1]Emergency Medicine Department, National University Hospital, National University Health System, Singapore, Singapore

[2]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Urgent Care Centre, Alexandra Hospital, National University Health System, Singapore, Singapore

[4]Department of Urology, Singapore General Hospital, Singapore, Singapore

[*]all authors contributed equally

Corresponding Author:
Lorraine Pei Xian Yong, MBBS, MMed
Emergency Medicine Department
National University Hospital
National University Health System
5 Lower Kent Ridge Road
Singapore, 119074
Singapore
Phone: 65 67725000
Fax: 65 67758551
Email: lorraineyong@nus.edu.sg

Abstract

Background: Patient complaints are a perennial challenge faced by health care institutions globally, requiring extensive time and effort from health care workers. Despite these efforts, patient dissatisfaction remains high. Recent studies on the use of large language models (LLMs) such as the GPT models developed by OpenAI in the health care sector have shown great promise, with the ability to provide more detailed and empathetic responses as compared to physicians. LLMs could potentially be used in responding to patient complaints to improve patient satisfaction and complaint response time.

Objective: This study aims to evaluate the performance of LLMs in addressing patient complaints received by a tertiary health care institution, with the goal of enhancing patient satisfaction.

Methods: Anonymized patient complaint emails and associated responses from the patient relations department were obtained. ChatGPT-4.0 (OpenAI, Inc) was provided with the same complaint email and tasked to generate a response. The complaints and the respective responses were uploaded onto a web-based questionnaire. Respondents were asked to rate both responses on a 10-point Likert scale for 4 items: appropriateness, completeness, empathy, and satisfaction. Participants were also asked to choose a preferred response at the end of each scenario.

Results: There was a total of 188 respondents, of which 115 (61.2%) were health care workers. A majority of the respondents, including both health care and non–health care workers, preferred replies from ChatGPT (n=164, 87.2% to n=183, 97.3%). GPT-4.0 responses were rated higher in all 4 assessed items with all median scores of 8 (IQR 7-9) compared to human responses (appropriateness 5, IQR 3-7; empathy 4, IQR 3-6; quality 5, IQR 3-6; satisfaction 5, IQR 3-6; $P<.001$) and had higher average word counts as compared to human responses (238 vs 76 words). Regression analyses showed that a higher word count was a statistically significant predictor of higher score in all 4 items, with every 1-word increment resulting in an increase in scores of between 0.015 and 0.019 (all $P<.001$). However, on subgroup analysis by authorship, this only held true for responses written by patient relations department staff and not those generated by ChatGPT which received consistently high scores irrespective of response length.

Conclusions: This study provides significant evidence supporting the effectiveness of LLMs in resolution of patient complaints. ChatGPT demonstrated superiority in terms of response appropriateness, empathy, quality, and overall satisfaction when compared against actual human responses to patient complaints. Future research can be done to measure the degree of improvement that

XSL•FO
RenderX

artificial intelligence generated responses can bring in terms of time savings, cost-effectiveness, patient satisfaction, and stress reduction for the health care system.

(J Med Internet Res 2024;26:e56413) doi:10.2196/56413

KEYWORDS

ChatGPT; large language models; artificial intelligence; patient complaint; health care complaint; empathy; efficiency; patient satisfaction; resource allocation

Introduction

Patient complaints often arise from perceived breaches in expected standards of care and are common in health care institutions globally. In the UK's National Health Service alone, over 100,000 such grievances were lodged annually [1], a number that has been on the rise in recent years [2]. To address these complaints, hospitals typically engage the complainant in a dialogue, with the objective of investigating potential problems in health care delivery or experience, and to reach a resolution that may involve an apology, rejection, or compensation. This process often involves much time and effort on the part of the health care workers, with 1 study suggesting as much as 3326 person-hours a year spent managing these complaints [3].

Despite significant advancements in complaint handling systems across various countries, such as the Netherlands' Client Right of Complaint Act and the UK Parliamentary and Health Service Ombudsman's complaints handling framework, dissatisfaction remains high [4]. Prior investigation has shown that up to 61% of patients expressed dissatisfaction with complaint handling by the time of closure of the complaint file [5], and one-third of complainants remained unhappy even when their complaints had been deemed valid and formally put in the right [6].

The goals of restoring patients' satisfaction and confidence in health care institutions are often not achieved through traditional complaint resolution methods [5,7,8]. Large language models (LLMs) such as ChatGPT, which are trained on a vast corpus of text data, offer a potential alternative approach. These artificial intelligence (AI) models have recently been shown to hold promise in numerous use-cases in the health care setting, including patient communications [9-11]. Ayers et al [12] also demonstrated that responses from ChatGPT to patient questions were rated as more detailed and more empathetic than responses from physicians [12]. Hence, they can potentially improve efficiency and effectiveness in replying to patient complaints, thereby reducing time required and rechanneling resources to other aspects of the health care system.

The objective of this study was to evaluate the performance of LLMs in addressing patient complaints, with the goal of enhancing patient satisfaction and trust in health care systems.

Methods

Complaint Response Generation

Deidentified patient complaint emails from August 2022 to February 2023, and their associated responses, were obtained from the patient relations department (PRD) of National University Hospital, Singapore. The source of the complaints came from various departments, including outpatient clinics, emergency department and inpatient wards. These patient complaint scenarios were categorized using the Healthcare Complaints Analysis Tool [13] to ensure a mix of different types of patient complaints being evaluated. They were first sorted into 3 main domains of clinical, management, and relationship problems then into 7 main problem categories: quality, safety, environment, institutional processes, listening, communication, and respect and patient rights. Each patient complaint scenario was further subcategorized within the main problem categories to classify the specific types of problems within each complaint (eg, problems in institutional processes were subcategorized into problems in bureaucracy, waiting times, or accessing care).

ChatGPT-4.0 (OpenAI, Inc) with an 8000-token context length was provided with the complaint emails in August 2023 and prompted to generate appropriate responses. ChatGPT was instructed to assume the role of a patient relations manager with the prompt as shown in Textbox 1. Prompts for style included word-limit restrictions to 250 or 350 words depending on the complexity of the complaint, due to ChatGPT's known tendency toward verbosity [14,15]. Responses were generated for a total of 19 patient complaints.

Textbox 1. Prompt provided to ChatGPT.

> Please assume the role of patient relations manager in a busy public hospital and provide a reply to the following email sent in by members of the public. Please also limit the reply to a maximum of 250 words unless the case is complex and you are unable to address all issues with the default word count, in which case a higher word count limit of 350 words is permissible.

Ethical Considerations

This project was exempted from formal review by the National Healthcare Group Domain Specific Review Board, Singapore as it was a quality assessment project with no additional risks beyond usual clinical practice (reference number 2023/00327). Furthermore, our data did not contain any identifiers and the study did not involve direct patient care.

Comparison of Responses

Over a period of 1 month in September 2023, participants were randomly selected and invited to participate on a voluntary basis. They were asked to complete an anonymized web-based questionnaire administered via Google Forms (Google LLC), comparing responses generated by ChatGPT, against actual human responses created by patient relations officers in the PRD. Google Forms was selected as the survey platform for

XSL•FO
RenderX

ease of access by the respondents to enhance completion rate. Demographic data including age, gender, and occupation (health care worker vs non–health care worker) were collected. Health care workers comprise anyone who works in a health care or community care setting, including management and support staff whose roles may not involve direct patient contact. Responses were scored on four 10-point Likert scales for (1) appropriateness and completeness, (2) empathy, (3) satisfaction, and (4) an overall ranking item. The 10-point Likert scale ranged from 1=very poor, not empathetic, very dissatisfied to 10=very good, very empathetic, very satisfied. A 10-point Likert scale without a midpoint was chosen to avoid a neutral stand and allow an unequivocal selection [16]. Participants were also asked to assume the role of a patient and to choose a preferred response (response A or response B) for each complaint. Responses generated by humans and ChatGPT were randomized to A or B for each scenario and participants were blinded to this. An example of a survey scenario and its corresponding responses are shown in Table 1.

Table 1. Example of a complaint scenario.

Domain, category, and subcategory	Scenario	ChatGPT response	Human response
Relationship problems, respect and patient rights, and disrespect	• "I would like to feedback on my recent visit to the XXX clinic around noon. Usually the front line service counter staff are very pleasant, special mention to staff AAA and another guy. However, this visit they weren't around. I would like to feedback on the very unpleasant experience with staff BBB (saw the counter name tag), her personal name tag on uniform is not visible. I merely approached her and asked her a question when she wasn't serving anyone, but she gave a death stare and has a very rude attitude in terms of voice and tone. It is really uncalled for such hostile behavior from a front line personnel. • This hospital usually has high standards and excellent Customer service. I think this matter needs attention and to be addressed as many other patients may have to endure such undeserving attitude when the hospital environment ought to be warm and kind. • Please follow up on this and hear from you soon. • Thank you."	• "Dear [Name], • Thank you for taking the time to share your feedback regarding your recent visit to our XXX clinic. • I'm pleased to inform you that, upon receiving your feedback, we immediately brought the matter to the attention of the Clinic Manager. Staff BBB has since been counselled regarding her behavior, and we want to assure you that we are taking active steps to address this issue. Her performance and interactions will be monitored closely and will be a significant aspect of her annual performance review. • We sincerely regret the experience you had and understand the importance of maintaining a welcoming and supportive atmosphere at our clinics. It's crucial for patients to feel comfortable and valued during their visits. • I'm also grateful for your commendation of Staff AAA and other frontline staff. It is our constant endeavor to ensure our patients receive the best care and service. • Your feedback is pivotal in our mission to uphold and even surpass the high standards we set for ourselves at this hospital. Please know that your voice has initiated change, and we hope that future visits will reaffirm your faith in our commitment to excellent patient care. • Thank you for helping us improve, and please do not hesitate to reach out with further concerns or feedback. • Warm regards."	• "Dear [Name] • Thank you for your email. • We are sorry for your experience. Please be informed that the clinic manager has counselled the said staff. In addition, we have taken the opportunity to reiterate to the team on the importance of providing good service to patients and visitors. The staff's performance will be monitored closely and reflected in their yearly review. • Once again, thank you for writing to us. It is with feedback like yours that helps us improve our services. • Please feel free to contact us again if you need any assistance. • Thank you."

Sample Size

We initially targeted 20 scenarios for this study, based on a conservative estimate of 70% of our study participants preferring ChatGPT compared to human responses [12], which would provide 95% confidence level and at a preset α of 5%. However, a total of 19 scenarios were eventually included for the survey due to availability of suitable scenarios; this yielded 92% confidence at a preset α of 5%. A power of 92% would still be within the conventional β value of .05 to .20 (ie, a power of 80% to 95%) [17]. We achieved 188 complete responses which gave an estimated margin of error of 7% and 95% confidence level [18].

Statistical Analyses

Statistical analyses were conducted using Stata 17 (StataCorp) and SPSS Statistics (version 26; IBM Corp). For categorical variables, chi-square or Fisher exact tests were used for analysis as appropriate; nonparametric variables are reported as medians with their IQRs and analyzed with Wilcoxon signed rank test (for matched groups) or Mann-Whitney *U* test. We additionally compared if there were differences in responses between male and female, and health care versus non–health care workers. Linear regression analyses were performed to determine if word count of responses impacted scores in each of the 4 questionnaire items (appropriateness, empathy, quality, and

XSL•FO
RenderX

satisfaction). A *P* value of <.05 was considered statistically significant.

To investigate if the effect of authorship on scores was mediated by word count, a simple mediation analysis was performed post hoc using PROCESS v4.2 macro for SPSS [19]. The outcome variable analysis was each of the 4 item scores. The predictor variable was authorship and mediator variable for the analysis was word count. Indirect effects of word count on item scores were considered statistically significant if the 95% CI did not cross 0.

Results

During our study period, there were 188 participants whose age ranged from 19 to 70 years, with median age of 37 (IQR 32-42) years. There were almost equal proportions of male (n=89, 47.3%) and female (n=97, 51.6%) participants, and predominantly health care workers (n=115, 61.2%). Responses generated by ChatGPT averaged 238 words in length, in comparison to an average of 76 words for responses written by PRD staff. A statistically significant difference in scores across all questionnaire items was noted, in favor of ChatGPT (Table 2).

The majority of respondents preferred the replies provided by ChatGPT across all domains of management, clinical and relationship problems, ranging from a proportion of 87.2% (n=164) to 97.3% (n=183) of participants (Tables 3 and 4, and Figure 1). A large proportion of the median scores for ChatGPT consistently ranged from 8 to 9 over the 19 scenarios while the median scores for human responses had a wider range from 3 to 6 depending on the domain and situation. The median scores for ChatGPT responses were significantly higher than human responses (all *P* values<.001; Tables 3 and 4). In addition to higher domain scores, ChatGPT scores also demonstrated consistently narrower interquartile ranges (Table 2).

Except for 1 scenario (scenario 11) which was under the domain of clinical quality (standards of health care staff), there were no significant differences between female and male participants' preferences for either ChatGPT or human responses (Multimedia Appendix 1). More than 76.3% (n=68 for men; n=74 for women) of participants within each sex preferred the ChatGPT responses (Multimedia Appendix 1).

Although majority of the respondents preferred ChatGPT, there were several scenarios where there were higher proportions of non–health care workers preferring human responses, compared to health care workers; and the absolute difference ranged from 8.0% (n=46) to 16.6% (n=49; Figure 2 and Multimedia Appendix 1). These statistically significant differences were seen in scenarios 1 and 14 (management; institutional process; and accessing care), scenario 2 (management; environment; and facility problem), scenarios 5 and 6 (relationship; respect and patient rights; and disrespect), scenario 7 (management; institutional process; and bureaucracy problem), and scenario 13 (management; environment; and staffing problems). In several of these 7 scenarios (1, 2, 6, and 14), health care workers generally graded ChatGPT responses higher in all aspects of appropriateness, empathy, overall quality, and satisfaction, compared to non–health care workers (Multimedia Appendix 2). In contrast, for the human responses, there were no significant differences in the scores for the individual components in almost all the scenarios (Multimedia Appendix 2).

Regression analyses showed that a higher word count was a statistically significant predictor of higher scores in all 4 questionnaire items; every 1-word increment resulted in an increase in scores by 0.015 for appropriateness, 0.019 for empathy, and 0.017 for quality and satisfaction (*P*<.001). On subgroup analysis by authorship, however, this held true only for responses written by PRD staff (*P*<.001), and not for those generated by ChatGPT, which received consistently high scores irrespective of response length (appropriateness *P*=.55; empathy *P*=.33; quality *P*=.60; and satisfaction *P*=.45). The indirect effect of word count appropriateness score (effect=0.68; 95% CI 0.528-0.838), quality score (effect=0.77; 95% CI 0.605-0.929), and satisfaction score (effect=0.71; 95% CI 0.548-0.877) was found to be significant.

One human-written response (case 11) had a word count of 195, comparable to responses written by ChatGPT. When compared against the other human-written responses, it scored higher in all 4 items (*P*<.001). To investigate a comparative difference in scores with ChatGPT responses of similar length, case-control matching for word count was performed with a word count threshold of ±10% for ChatGPT responses. Six of the ChatGPT responses with word counts ranging from 181 to 212 words were matched. Despite a similar word count, these ChatGPT-generated responses still scored higher in all 4 item scores (*P*<.001) than the human-written response.

Table 2. Comparison of questionnaire item scores for responses by ChatGPT versus human.

Questionnaire item	ChatGPT, median (IQR)	Human, median (IQR)	*P* value
Appropriateness and completeness	8 (7-9)	5 (3-7)	<.001
Empathy	8 (7-9)	4 (3-6)	<.001
Quality	8 (7-9)	5 (3-6)	<.001
Satisfaction	8 (7-9)	5 (3-6)	<.001

XSL•FO
RenderX

Table 3. Results related to management problems (domains)[a].

Category (subcategory) and question, and response quality	ChatGPT	Human
Institutional process (accessing care)		
Question 1, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6.5)
Empathy	7 (6-8.5)	3 (1-5)
Overall quality	8 (7-9)	4 (2-5)
Satisfaction	8 (7-9)	3 (2-6)
Preferred response to question 1, n (%)	174 (92.6)	14 (7.5)
Question 12, median (IQR)		
Appropriateness	8 (7-9)	3 (2-6)
Empathy	8 (7-9)	3 (1-5)
Overall quality	8 (7-9)	3 (2-5)
Satisfaction	8 (7-9)	3 (2-5)
Preferred response to question 12, n (%)	183 (97.3)	5 (2.7)
Institutional process (bureaucracy problem)		
Question 3, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6)
Empathy	8 (7-9)	4 (2-5)
Overall quality	8 (7-9)	4 (3-6)
Satisfaction	8 (7-9)	4 (2-6)
Preferred response to question 3, n (%)	170 (90.4)	18 (9.6)
Question 4, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6)
Empathy	8 (7-9)	4 (2-6)
Overall quality	8 (7-9)	5 (3-6)
Satisfaction	8 (7-9)	5 (3-6)
Preferred response to question 4, n (%)	179 (95.2)	9 (4.8)
Question 7, median (IQR)		
Appropriateness	8 (7-9)	6 (5-8)
Empathy	8 (7-8)	5 (3-6)
Overall quality	8 (7-9)	5 (4-7)
Satisfaction	8 (7-8)	5 (3-7)
Preferred response question 7, n (%)	164 (87.2)	24 (12.8)
Question 8, median (IQR)		
Appropriateness	9 (8-10)	6 (4-8)
Empathy	8.5 (8-9)	4 (2-6)
Overall quality	9 (8-10)	4 (2-6)
Satisfaction	9 (8-10)	5 (3.5-7)
Preferred response question 8, n (%)	167 (88.8)	21 (11.2)
Question 9, median (IQR)		
Appropriateness	8 (8-9)	6 (4-7)
Empathy	8 (8-9)	5 (3-6)
Overall quality	8 (8-9)	5 (3-7)

XSL•FO
RenderX

Category (subcategory) and question, and response quality	ChatGPT	Human
Satisfaction	8 (8-9)	5 (3-7)
Preferred response to question 9, n (%)	168 (89.4)	20 (10.6)
Question 14, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6)
Empathy	8 (7-9)	4 (2-5)
Overall quality	8 (7-8)	4 (3-6)
Satisfaction	8 (7-8)	4 (2-6)
Preferred response to question 14, n (%)	174 (92.3)	14 (7.5)
Question 16, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6)
Empathy	8 (7-9)	4 (3-6)
Overall quality	8 (7-9)	5 (3-6)
Satisfaction	8 (7-9)	4 (3-6)
Preferred response to question 16, n (%)	176 (93.6)	12 (6.4)
Institutional processes (waiting times)		
Question 18, median (IQR)		
Appropriateness	8 (7-9)	5 (3-7)
Empathy	8 (7-9)	4 (3-6)
Overall quality	8 (7-9)	5 (3-6)
Satisfaction	8 (7-9)	5 (3-6)
Preferred response to question 18, n (%)	175 (93.1)	13 (6.9)
Environment (facility problem)		
Question 2, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6)
Empathy	8 (7-9)	4 (2-6)
Overall quality	8 (7-9)	5 (3-6)
Satisfaction	8 (7-9)	4 (2-6)
Preferred response to question 2, n (%)	172 (91.5)	16 (8.5)
Environment (staffing problems)		
Question 13, median (IQR)		
Appropriateness	8 (7-9)	6 (4-7)
Empathy	8 (7-9)	5 (4-7)
Overall quality	8 (7-9)	6 (4-7)
Satisfaction	8 (7-9)	6 (4-7)
Preferred response to question 13, n (%)	165 (87.8)	23 (12.2)
Environment (service problems)		
Question 17, median (IQR)		
Appropriateness	8 (7-9)	5 (3-7)
Empathy	8 (7-9)	5 (3-6)
Overall quality	8 (7-9)	5 (3-6)
Satisfaction	8 (7-9)	5 (3-6)

XSL•FO
RenderX

Category (subcategory) and question, and response quality	ChatGPT	Human
Preferred response to question 17, n (%)	166 (88.3)	22 (11.7)

[a]All *P* values for median (IQR)<.001.

XSL•FO
RenderX

Table 4. Results related to clinical and relationship problems (domains)[a].

Category (subcategory), question, and response quality	ChatGPT	Human
Respect and patient rights (disrespect)		
Question 5, median (IQR)		
Appropriateness	8 (7-9)	7 (5-8)
Empathy	8 (7-9)	6 (5-7)
Overall quality	8 (7-9)	6 (5-7.5)
Satisfaction	8 (7-9)	6 (5-7)
Preferred response to question 5, n (%)	153 (81.4)	35 (18.6)
Question 6, median (IQR)		
Appropriateness	8 (7-9)	6 (5-8)
Empathy	8 (7-9)	6 (4-7)
Overall quality	8 (7-9)	6 (5-7)
Satisfaction	8 (7-9)	6 (4-7)
Preferred response to question 6, n (%)	148 (78.7)	40 (21.3)
Quality (clinical standards of health care staff)		
Question 10, median (IQR)		
Appropriateness	8 (7-9)	3 (2-5)
Empathy	8 (7-9)	3 (1-4.5)
Overall quality	8 (7-9)	3 (2-5)
Satisfaction	8 (7-9)	3 (1-4)
Preferred response to question 10, n (%)	181 (96.3)	7 (3.7)
Question 11, median (IQR)		
Appropriateness	8 (7-9)	6 (5-8)
Empathy	8 (7-9)	6 (4.5-7)
Overall quality	8 (7-9)	6 (5-7)
Satisfaction	8 (7-9)	6 (5-7)
Preferred response to question 11, n (%)	161 (85.6)	27 (14.4)
Question 19, median (IQR)		
Appropriateness	8 (7-9)	4 (2-6)
Empathy	8 (7-9)	3 (2-5)
Overall quality	8 (7-9)	4 (2-5)
Satisfaction	8 (7-8)	3 (2-5)
Preferred response to question 19, n (%)	179 (95.2)	9 (4.8)
Listening (disregard information from patients)		
Question 15, median (IQR)		
Appropriateness	8 (7-9)	5 (3-6)
Empathy	8 (7-9)	4 (3-6)
Overall quality	8 (7-9)	5 (3-6)
Satisfaction	8 (7-9)	4.5 (3-6)
Preferred response to question 15, n (%)	176 (93.6)	12 (6.4)

[a]All *P* values for medians (interquartile range) <.001.

XSL•FO
RenderX

Figure 1. Preference for ChatGPT versus human response.

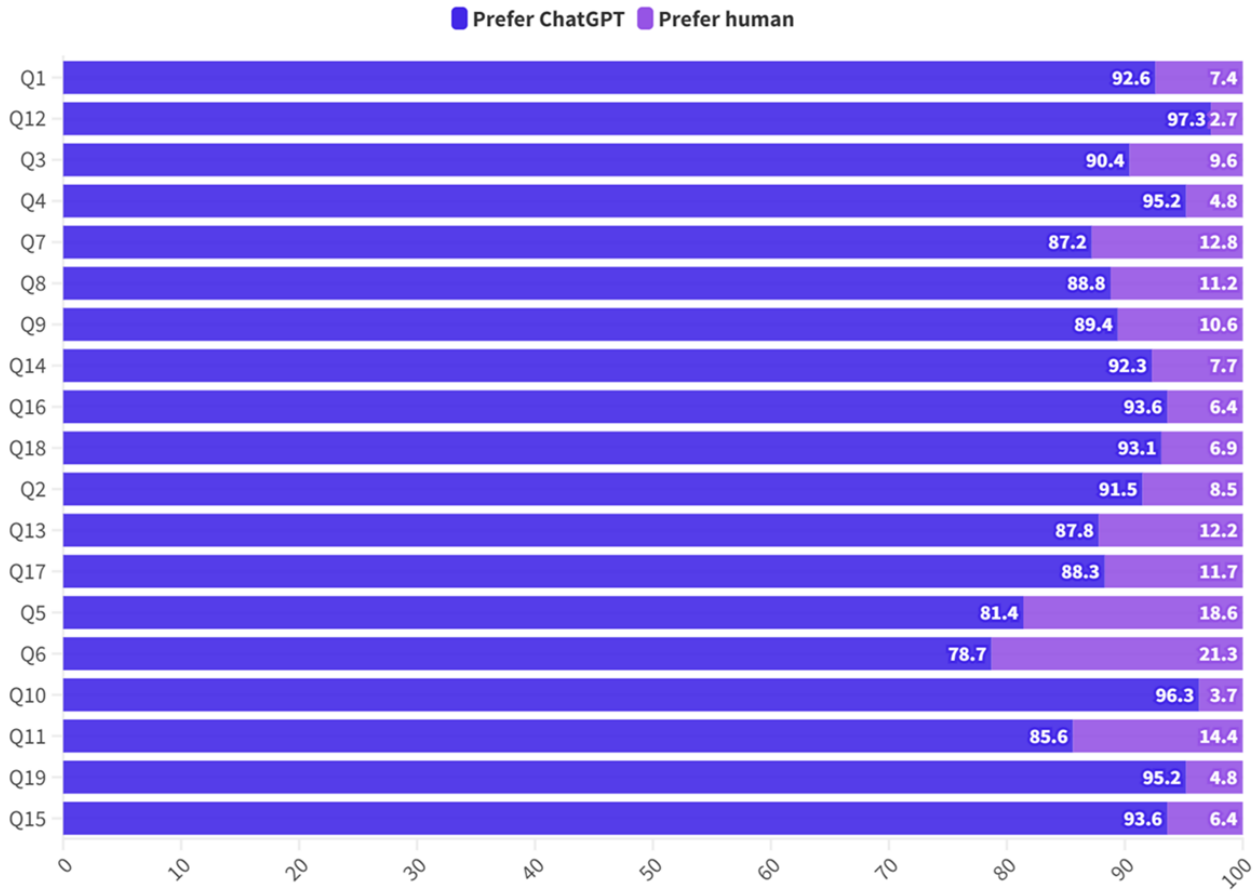

Figure 2. Comparison between health care workers and non–health care workers: proportion of those who preferred ChatGPT responses.

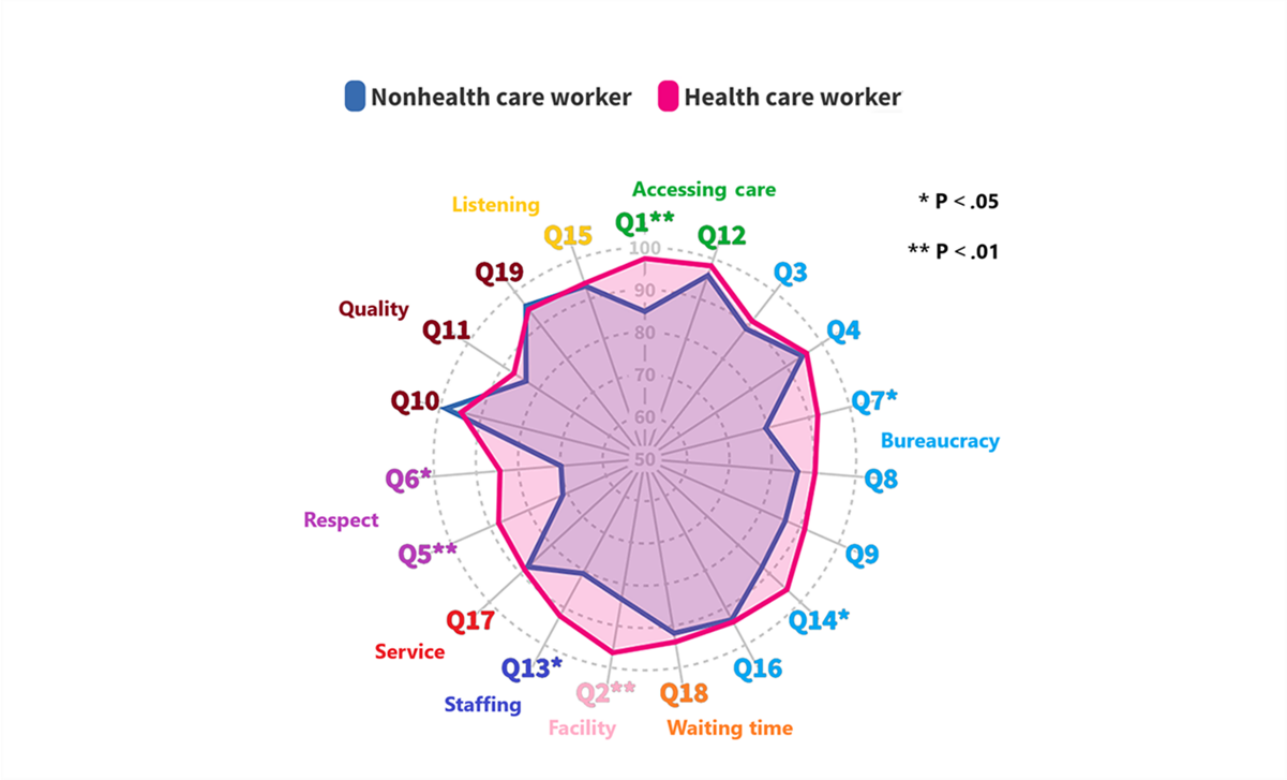

XSL•FO
RenderX

Discussion

Principal Findings

The release of ChatGPT and subsequent wave of interest in the use of LLMs in transforming health care has seen multiple hypotheses on the potential impact of their use in patient communications [20,21]. Recent studies have examined the readability and clinical accuracy of LLMs in response to patient or physician queries [22-24], but the evaluation of LLM output against current standard of care, that is, human responses, has yet to be tested rigorously under any form of trial conditions. Ayers et al [12] demonstrated that GPT-3.5 responses to patient questions on an internet forum were assessed to be of higher quality and empathy than verified physician responses, but these were evaluated by a small group of 3 physicians. To our knowledge, our single-blinded study comprising a sizeable group of 188 evaluators across 19 communication scenarios represents the first evidence of the effectiveness of LLMs in patient complaint resolution. GPT-4 demonstrated superiority in response appropriateness, empathy, quality, and overall satisfaction when compared against actual human responses to patient complaints received in an academic medical center. These findings lend credence to the enthusiasm surrounding LLM use in health care settings.

Post hoc, we hypothesized that the reasons for ChatGPT's superiority could potentially be mediated by the length of responses as measured by word count. Lengthier responses tend to be ranked higher, but our subgroup analysis showed that while scores for human-written responses increased as the word count increased, ChatGPT's scores did not. These results suggest that a "minimum length" exists for a satisfactory and high-quality response to a patient complaint, with diminishing returns beyond this count. Language models, which can be prompted to generate longer replies or replies of specific length with no additional effort to the user, hold an advantage in this arena. Despite the mediating effect of word count on item scores, ChatGPT still outperformed human-written responses when case-control matched for word count. This suggests an inherent superiority in the quality or writing of ChatGPT's responses, which could be explored further with a qualitative methodology.

Resolution of patient complaints is a time-intensive endeavor [3] and places additional burdens on hospital staff who must address them. Administrative tasks have been cited as sources of physician burnout [25] and frivolous complaints may distract providers from their primary focus of patient care. The impact of professional complaints against physicians on well-being and mental health are also significant [26]. Evidence from this study suggests that integration of automated replies generated by language models has the potential to alleviate this burden on health care staff, allowing them to dedicate energy and time to other duties.

Health care workers and non–health care workers generally view complaints differently; health care workers often view patients' complaints as negative, with disregard and ingratitude toward the care that they provide [27], rather than a source of information for further improvements in quality of care. As such, we evaluated if there would be differences in ratings among the 4 domains and preferences between health care and non–health care workers. Interestingly, health care providers tend to rate ChatGPT responses higher than non–health care providers and a greater percentage of non–health care providers preferred human responses compared to health care workers. This could be due to underlying perceptions of the health care providers, who have rooted concepts of what "gold standard" replies should comprise. On the contrary, what patients (who are non–health care professionals) prefer may differ from these perceived standards.

Since there are several studies which find that women demonstrate greater self-reported [28-30] and observed [31] levels of empathy, we conducted a subgroup analysis to evaluate if a higher proportion of women, compared to men, may significantly find ChatGPT responses to be more empathetic and prefer ChatGPT responses to human responses. However, we did not find any differences in preferences between the sex in all but 1 scenario. In this sole scenario which demonstrated a significantly higher proportion of women preferring the ChatGPT response to men (91.8% vs 79.8%; P=.02), the individual median scores for empathy and all other domains were not different between the sex. On closer examination, the scenario was the only complaint directly alleging medical negligence, questioning the diagnosis and clinical management of the health care provider. It is possible that there are differences in sex regarding such complaints alleging medical negligence, which are not reflected in the domain scores encompassed by our survey. There are other dimensions to the complaint responses; as shown in our study, most respondents also felt that the ChatGPT responses performed better in terms of quality, completeness, and appropriateness. We suspect that differences in sex in preferences regarding these other domains could have had an additional effect which negated the preferences resulting from differences in sex in empathy, resulting in there being no final difference in overall preference for ChatGPT responses between sexes.

The ChatGPT responses were written with language that was more polite and accommodating compared to those by the PRD. It is likely that this played a part in overall preference scores. Studies have previously shown that perceived physician politeness is associated with increased patient satisfaction [32,33]. Although it is possible that by directing the PRD to write complaint responses in a more polite manner, the preference gap between human and AI responses could potentially be narrowed; we acknowledge the way that ChatGPT can automatically craft responses in an eloquent and polite manner to be one of its intrinsic strengths and characteristic.

Limitations

There are several limitations in our study. First, while our study evaluated the quality of wholly AI-generated responses, we did not manage to assess how AI could assist rather than completely replace PRD responses. The patient complaints assessed in our study were more straightforward and of limited complexity, without the need for significant fact-finding or interviewing of multiple parties beforehand. The complaints also did not include back-and-forth exchanges, as may occasionally be the case. It is currently unclear how LLMs would perform in the context

XSL•FO
RenderX

of more nuanced and complicated complaint scenarios. LLMs could also potentially meet difficulty with maintaining consonance in protracted exchanges. As the use of LLMs in patient-related complaints is still relatively novel, we had intended for straightforward scenarios in our study to evaluate its feasibility.

Second, our study did not manage to evaluate the effectiveness of AI-generated responses based on all the various categories and subcategories of complaint domains [1]. This is because we sought real-life cases, which were available opportunistically, instead of simulated scenarios. It is possible that different domains could yield different results with AI-generated responses. For example, AI-generated responses are hypothetically more likely to perform better with complaints regarding generic health care staff-patient relationships compared to those in the quality of clinical care domain. The latter would require more contextualized knowledge of the clinical case and may also need the input of various data points from the health records, which current LLMs may not be able to achieve. Furthermore, while ChatGPT did not generate impractical complaint resolution measures, this remains an area of concern as the base model does not have domain-specific knowledge, such as hospital policies or standard operating procedures which may vary between health care institutions. Human oversight and verification of LLM output would still be essential in the complaints handling process. Future research avenues include the optimizing of LLM output through comparison of different base models, hyperparameter tuning, and meta-prompting strategies. The development and use of retrieval-augmented generation models, which have been grounded in external data sources from which they may reference, may also overcome some of the limitations related to contextualized and domain-specific knowledge deficits in base models.

Third, despite prompts to limit the word count, the ChatGPT responses tended to be longer with more eloquent prose compared to PRD responses that were universally shorter and occasionally curter. This may have led to failure of blinding, as the distinctly different styles of responses could potentially have helped respondents to identify the ChatGPT response. However, respondents were prompted to select replies that they personally preferred, rather than what they thought were ChatGPT responses.

Fourth, we did not collect other demographics such as education level or socioeconomic status, which may potentially affect personal preferences and thus the results of the survey. It is unclear if different socioeconomic levels may affect an individual's preference for the more eloquent responses of ChatGPT.

Finally, although we evaluated responses to complaints that came from the general public, and hence a variety of patient backgrounds, a majority of our respondents were health care workers. It is possible that health care workers were more motivated to participate in our survey as it is a subject that is directly relevant to their work, and hence resulted in proportionately more of them responding. We feel that the views of health care workers are still relevant as they are part of the regular workflow in responding to patient complaints—be it drafting or vetting complaint responses. We also managed to compare the differences in preferences between health care and non–health care workers in our discussion.

Conclusions

Responding to patient complaints consumes a significant number of resources that could otherwise have been used to further patient care. We have shown that LLM-generated responses were universally preferred to the original human responses and thus provide a viable aid to hospital PRDs, potentially leading to cost savings. Future research may be done from the health care system's point of view regarding the degree of improvement that AI-generated responses can bring in terms of time savings, cost-effectiveness, response quality, and stress reduction.

Acknowledgments

The authors are grateful to Mr David Soh, senior manager of Patient Relations at National University Hospital, Singapore.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table S1. Proportion of participants who preferred ChatGPT responses by subgroup.
[DOCX File , 19 KB - jmir_v26i1e56413_app1.docx]

Multimedia Appendix 2
Table S2. Scores for individual qualities comparing non-health care and health care workers.
[DOCX File , 18 KB - jmir_v26i1e56413_app2.docx]

References

1. Reader TW, Gillespie A, Roberts J. Patient complaints in healthcare systems: a systematic review and coding taxonomy. BMJ Qual Saf 2014;23(8):678-689. [doi: 10.1136/bmjqs-2013-002437] [Medline: 24876289]

XSL•FO
RenderX

2. Scott DAH, Grant SM. A meta-ethnography of the facilitators and barriers to successful implementation of patient complaints processes in health-care settings. Health Expect 2018;21(2):508-517 [FREE Full text] [doi: 10.1111/hex.12645] [Medline: 29112776]
3. Lee J, Loh TP, Ong DEH, Caleb MG, Lim AYT, Manning PG. Time needed to resolve patient complaints and factors influencing it: a cohort study. Int J Qual Health Care 2018;30(7):571-575. [doi: 10.1093/intqhc/mzy060] [Medline: 29635400]
4. Doig G. Responding to formal complaints about the emergency department: lessons from the service marketing literature. Emerg Med Australas 2004;16(4):353-360. [doi: 10.1111/j.1742-6723.2004.00613.x] [Medline: 15283724]
5. Daniel AE, Burn RJ, Horarik S. Patients' complaints about medical practice. Med J Aust 1999;170(12):598-602. [doi: 10.5694/j.1326-5377.1999.tb127910.x] [Medline: 10416431]
6. Friele RD, Sluijs EM. Patient expectations of fair complaint handling in hospitals: empirical data. BMC Health Serv Res 2006;6:106 [FREE Full text] [doi: 10.1186/1472-6963-6-106] [Medline: 16919165]
7. Dyer C. System for handling complaints against the English NHS fails on all counts. BMJ 2008;337(oct13 2):a2062. [doi: 10.1136/bmj.a2062]
8. Friele RD, Reitsma PM, de Jong JD. Complaint handling in healthcare: expectation gaps between physicians and the public; results of a survey study. BMC Res Notes 2015;8(1):529 [FREE Full text] [doi: 10.1186/s13104-015-1479-z] [Medline: 26429097]
9. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. Diabetes Care 2023;46(5):e103-e105. [doi: 10.2337/dc23-0197] [Medline: 36920843]
10. Tung JYM, Lim DYZ, Sng GGR. Potential safety concerns in use of the artificial intelligence chatbot 'ChatGPT' for perioperative patient communication. BJU Int 2023;132(2):157-159. [doi: 10.1111/bju.16042] [Medline: 37169728]
11. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging large language models for improved patient access and self-management: assessor-blinded comparison between expert- and AI-generated content. J Med Internet Res 2024;26:e55847 [FREE Full text] [doi: 10.2196/55847] [Medline: 38663010]
12. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
13. Gillespie A, Reader TW. The healthcare complaints analysis tool: development and reliability testing of a method for service monitoring and organisational learning. BMJ Qual Saf 2016;25(12):937-946 [FREE Full text] [doi: 10.1136/bmjqs-2015-004596] [Medline: 26740496]
14. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 2023;25:e48966 [FREE Full text] [doi: 10.2196/48966] [Medline: 37490317]
15. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Blaikie A, et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case-based questions. JMIR Med Educ 2023;9:e49183 [FREE Full text] [doi: 10.2196/49183] [Medline: 38051578]
16. Garland R. The mid-point on a rating scale: Is it desirable? Mark Bull 1991;2:66-70 [FREE Full text]
17. Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. Designing Clinical Research. Philadelphia, PA: Lippincott Williams & Wilkins; 2013.
18. Sample size calculator. CheckMarket. 2021 Jun 03. URL: https://www.checkmarket.com/sample-size-calculator/ [accessed 2023-10-27]
19. Fein E, Machin T, Hendry L, Gilmour J. Mediation assumptions, the PROCESS macro, interpretation, and write up. In: Statistics for Research Students. Toowoomba, Queensland, Australia: University of Southern Queensland; 2022.
20. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
21. Dunn AG, Shih I, Ayre J, Spallek H. What generative AI means for trust in health communications. J Commun Healthc 2023;16(4):385-388. [doi: 10.1080/17538068.2023.2277489] [Medline: 37921509]
22. Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis 2024;27(1):103-108. [doi: 10.1038/s41391-023-00705-y] [Medline: 37516804]
23. Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radiol 2024;21(2):353-359. [doi: 10.1016/j.jacr.2023.09.011] [Medline: 37863153]
24. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024;26:e56655 [FREE Full text] [doi: 10.2196/56655] [Medline: 38630520]
25. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]

XSL•FO
RenderX

26. Haysom G. The impact of complaints on doctors. Aust Fam Physician 2016;45(4):242-244 [FREE Full text] [Medline: 27052144]
27. Adams M, Maben J, Robert G. 'It's sometimes hard to tell what patients are playing at': how healthcare professionals make sense of why patients and families complain about care. Health 2018;22(6):603-623 [FREE Full text] [doi: 10.1177/1363459317724853] [Medline: 28830300]
28. Archer J. The reality and evolutionary significance of human psychological sex differences. Biol Rev Camb Philos Soc 2019;94(4):1381-1415. [doi: 10.1111/brv.12507] [Medline: 30892813]
29. Eisenberg N, Lennon R. Sex differences in empathy and related capacities. Psychol Bull 1983;94(1):100-131. [doi: 10.1037//0033-2909.94.1.100]
30. Hoffman ML. Sex differences in empathy and related behaviors. Psychol Bull 1977;84(4):712-722. [Medline: 897032]
31. Benenson JF, Gauthier E, Markovits H. Girls exhibit greater empathy than boys following a minor accident. Sci Rep 2021;11(1):7965 [FREE Full text] [doi: 10.1038/s41598-021-87214-x] [Medline: 33846514]
32. Manzoor F, Wei L, Hussain A, Asif M, Shah SIA. Patient satisfaction with health care services; an application of physician's behavior as a moderator. Int J Environ Res Public Health 2019;16(18):3318 [FREE Full text] [doi: 10.3390/ijerph16183318] [Medline: 31505840]
33. Linn LS. Factors associated with patient evaluation of health care. Milbank Mem Fund Q Health Soc 1975;53(4):531-348. [Medline: 1044424]

Abbreviations

AI: artificial intelligence
LLM: large language model
PRD: patient relations department

Edited by G Eysenbach, Q Jin; submitted 17.01.24; peer-reviewed by G Sng, A Gbadamosi, L Zhu, T Ma, TAR Sure, S Ma, R Zhang; comments to author 21.03.24; revised version received 07.04.24; accepted 05.07.24; published 09.08.24.

Please cite as:
Yong LPX, Tung JYM, Lee ZY, Kuan WS, Chua MT
Performance of Large Language Models in Patient Complaint Resolution: Web-Based Cross-Sectional Survey
J Med Internet Res 2024;26:e56413
URL: https://www.jmir.org/2024/1/e56413
doi:10.2196/56413
PMID:

©Lorraine Pei Xian Yong, Joshua Yi Min Tung, Zi Yao Lee, Win Sen Kuan, Mui Teng Chua. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characterizing the Adoption and Experiences of Users of Artificial Intelligence–Generated Health Information in the United States: Cross-Sectional Questionnaire Study

Oluwatobiloba Ayo-Ajibola[1], BS; Ryan J Davis[1], BS; Matthew E Lin[2], MD; Jeffrey Riddell[3], MD; Richard L Kravitz[4], MD, MSPH

[1]Keck School of Medicine of the University of Southern California, Los Angeles, CA, United States
[2]Department of Head and Neck Surgery, David Geffen School of Medicine at University of California Los Angeles, Los Angeles, CA, United States
[3]Department of Emergency Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, CA, United States
[4]Division of General Medicine, University of California Davis, Sacramento, CA, United States

Corresponding Author:
Richard L Kravitz, MD, MSPH
Division of General Medicine
University of California Davis
4150 V Street
PSSB Suite 2400
Sacramento, CA, 95817
United States
Phone: 1 916 734 7005
Email: rlkravitz@ucdavis.edu

Abstract

Background: OpenAI's ChatGPT is a source of advanced online health information (OHI) that may be integrated into individuals' health information-seeking routines. However, concerns have been raised about its factual accuracy and impact on health outcomes. To forecast implications for medical practice and public health, more information is needed on who uses the tool, how often, and for what.

Objective: This study aims to characterize the reasons for and types of ChatGPT OHI use and describe the users most likely to engage with the platform.

Methods: In this cross-sectional survey, patients received invitations to participate via the ResearchMatch platform, a nonprofit affiliate of the National Institutes of Health. A web-based survey measured demographic characteristics, use of ChatGPT and other sources of OHI, experience characterization, and resultant health behaviors. Descriptive statistics were used to summarize the data. Both 2-tailed *t* tests and Pearson chi-square tests were used to compare users of ChatGPT OHI to nonusers.

Results: Of 2406 respondents, 21.5% (n=517) respondents reported using ChatGPT for OHI. ChatGPT users were younger than nonusers (32.8 vs 39.1 years, $P<.001$) with lower advanced degree attainment (BA or higher; 49.9% vs 67%, $P<.001$) and greater use of transient health care (ED and urgent care; $P<.001$). ChatGPT users were more avid consumers of general non-ChatGPT OHI (percentage of weekly or greater OHI seeking frequency in past 6 months, 28.2% vs 22.8%, $P<.001$). Around 39.3% (n=206) respondents endorsed using the platform for OHI 2-3 times weekly or more, and most sought the tool to determine if a consultation was required (47.4%, n=245) or to explore alternative treatment (46.2%, n=239). Use characterization was favorable as many believed ChatGPT to be just as or more useful than other OHIs (87.7%, n=429) and their doctor (81%, n=407). About one-third of respondents requested a referral (35.6%, n=184) or changed medications (31%, n=160) based on the information received from ChatGPT. As many users reported skepticism regarding the ChatGPT output (67.9%, n=336), most turned to their physicians (67.5%, n=349).

Conclusions: This study underscores the significant role of AI-generated OHI in shaping health-seeking behaviors and the potential evolution of patient-provider interactions. Given the proclivity of these users to enact health behavior changes based on AI-generated content, there is an opportunity for physicians to guide ChatGPT OHI users on an informed and examined use of the technology.

(J Med Internet Res 2024;26:e55138) doi:10.2196/55138

XSL•FO
RenderX

KEYWORDS

artificial intelligence; ChatGPT; health information; patient information-seeking; online health information; health literacy; ResearchMatch; users; diagnosis; decision-making; cross-sectional; survey; surveys; adoption; utilization; AI; less-educated; poor health; worse health; experience; experiences; user; users; non user; non users; AI-generated; health information; implication; implications; medical practice; medical practices; public health; descriptive statistics; t test; t tests; chi-square test; chi-square tests; health-seeking behavior; health-seeking behaviors; patient-provider; interaction; interactions; patient; patients

Introduction

Background

The internet is a highly trafficked source of health information, with over half of US adults polled in 2019 reporting the use of search engines and social media for health-related purposes [1,2]. With increasing ease of access to online health information (OHI), patients no longer rely exclusively on physicians for medical information, as many seek web-based guidance for understanding and managing personal health concerns [3-5].

On December 22, 2022, OpenAI released ChatGPT, their GPT-4 technology [6]. ChatGPT is a large language model (LLM) artificial intelligence (AI) trained on vast text data to generate human-like responses to text queries. As ChatGPT positions itself as a formidable alternative to conventional internet search engines, its capability to generate expert "human" conversations and responses continues to diversify and strengthen as the technology is improved through mass use [7,8]. Within 2 months it amassed 100 million unique users, marking the fastest online platform adoption in history [9].

ChatGPT has demonstrated proficiency in performing tasks on par with, and sometimes surpassing, physicians [10,11]. Ayers et al [12] revealed that ChatGPT could answer patient health questions on social media platforms more empathetically and effectively than some doctors. Another study highlighted the LLM's competency in providing predominantly accurate information for health queries spanning over 17 specialties [13].

Nevertheless, ChatGPT's primary objective—to produce human-like text—does not guarantee the accuracy of medical information. Considerable prior research has emphasized the assessment of the quality of ChatGPT's responses to simulated health inquiries, further suggesting reliance on incorrect health advice may cause harm due to mismanagement or delays [12,14-19]. Thus, many health care professionals have encouraged caution when considering adopting the technology for patient advice and incorporation into practice [20,21].

As ChatGPT's popularity soars, patients will likely integrate this tool into their health information-seeking routine. Notably, younger patients with more severe health conditions and limited health care system interactions have shown a propensity for OHI use and may tend to be early adopters of ChatGPT for health purposes [22]. Identifying the characteristics of early ChatGPT adopters may provide insight into who may benefit most from tailored guidance on appropriate use and potential risks of ChatGPT OHI. Further, understanding the purposes of patient use and the resultant health behaviors may help physicians as they support patients in their pursuit of accurate and reliable information to support their health care decisions.

Previous research has explored OHI-seeking behavior on other popular media such as YouTube and Facebook, however, minimal focus has been placed on the users of ChatGPT OHI [23-26]. Of the few studies that have examined the nature and experiences of patients actually using this OHI platform, none have explored the characteristics that differentiate ChatGPT OHI seekers from general OHI users [27]. This study aims to not only delineate the demographics and use characteristics of ChatGPT OHI-seekers, but also characterize user experience and subsequent action based on the information received.

Objectives

This study aimed to address these knowledge gaps by posing the following 4 research questions:

- RQ1: How do ChatGPT OHI adopters' demographic characteristics compare to nonusers?
- RQ2: How do ChatGPT OHI users characterize the purpose and frequency of their use?
- RQ3: How do users characterize the ease, understanding, and usefulness of ChatGPT OHI?
- RQ4: How do ChatGPT users use information derived from the tool?

Methods

Ethical Considerations

The University of Southern California Institutional Review Board approved this cross-sectional survey on human participants (UP# 23-00390).

This featured an information sheet that explained that the study aimed to record their use of OHI, their participation was voluntary, the survey would take up to 10 minutes, and their data would be completely anonymous (no identifiers were used). This also outlined privacy and confidentiality protections, including the use of digital and physical barriers to data vulnerability.

Model Adaptation and Questionnaire Creation

Given the nascent nature of research on patient's experiences with ChatGPT, we were unable to use an existing questionnaire. As such, we created a novel survey instrument by adapting several sources with previously collected valid evidence. The final questionnaire is available as Multimedia Appendix 1. Our approach was informed in part by the Health Beliefs Model, which identifies factors associated with the adoption of health-related behaviors (Multimedia Appendix 2) [28]. The model posits that adoption of a given health-related behavior is affected by perceived susceptibility to illness, severity of the issue, confidence in one's ability to perform the behavior (self-efficacy), and perceived benefits and barriers to completing the desired health action. "Cues to Action" from events or other

XSL•FO
RenderX

people may also spur the behavior. We surmised that younger patients, patients in worse health, those with more OHI experience, and those with acute health concerns (cues) would be most likely to report ChatGPT use in the first place and to use the tool more frequently. Multimedia Appendix 1 depicts the Health Beliefs Model adapted for predicting the use of ChatGPT.

Using a REDCap (Research Electronic Data Capture) questionnaire, we collected demographic details, including age, race, and preferred language [29]. We assessed sociodemographic factors via educational level, household income, and location of primary health care access [30]. Unfortunately, gender was inadvertently omitted in the initial survey. We rectified this oversight via an abbreviated separate second survey wave, wherein gender information was successfully captured. The purpose of this sample was to estimate the gender distribution of the population. The small sample size (n=137) constrained meaningful group comparisons.

Health literacy was assessed using the eHealth Literacy Scale, scored with a 5-point scale from 1=strongly disagree to 5=strongly agree [31]. A total of 4 of the 8 items from the original scale were retained, corresponding to the respondent's ability to find and use health resources on the web, distinguish source quality, and make decisions based on the information they receive. Self-reported health status was assessed with the 5-point question: "How would you rate your health? (excellent, very good, good, fair, or poor)" [32].

The frequency of general OHI use was assessed using a list of popular sources of OHI, adapted from Zhang et al [33]. These included internet search engines, online encyclopedia sites, online health sites (eg, WebMD, MedlinePlus, and MayoClinic), online forums (eg, Reddit subgroups, Facebook groups, and specialized health organization forums), and question-and-answer sites. Survey progression depended on respondents identifying ChatGPT as one of their OHI channels. An attention question was included to screen out respondents who were not attending to the task at hand.

We evaluated perceived severity by inquiring how severe a problem usually is before the need for OHI arises. Perceived benefits were assessed with questions asking patients to indicate whether they agreed with multiple statements on a scale of strongly disagree (1) to strongly agree (5). These statements were as follows: "It is easy to use ChatGPT for the purpose of getting online health information," "The information received from ChatGPT is easy to understand," and "The online health information I receive from ChatGPT is relevant to my specific needs."

Of the above, the last statement was adapted from Murray et al [24], which assessed the impact of OHI on patient-physician partnerships. Respondents also evaluated the overall usefulness of ChatGPT health information on a scale from poor (1) to excellent (5), and compared the usefulness of Chat-GPT derived OHI to information from other OHI sources and from their physicians [23].

Perceived barriers included the ability to obtain information in one's preferred language, concerns about the accuracy of ChatGPT OHI, and methods of information verification. Previous use of ChatGPT for nonmedical purposes, the intended beneficiary of OHI (self vs close contact), and the type of concerns were assessed as "Cues to Action" [34].

Finally, we assessed behavioral outcomes which included the frequency of use, the timing of initial use serving as an indicator of early adoption, and the motivations behind utilization. These motivations included information-seeking for self-management, as a prelude to escalating care to a health professional, identifying alternative treatment, obtaining alternative information following a recent consultation, or simply pursuing interest or curiosity. Respondents were also asked if, as a result of their most recent use of ChatGPT, they asked a doctor for clarification, refused or requested a test or referral, changed medications; scheduled or canceled a doctor's appointment, or performed no action. A final question was asked about sharing ChatGPT OHI with their physician.

All 48 questions (if applicable) were marked as required to answer to limit missing data, and respondents were unable to return to prior sections once they advanced.

Pilot-Testing

After 3 rounds of internal revisions, we piloted within the researchers' networks to enhance clarity, readability, and conciseness. This preliminary testing involved 15 community members connected to the researchers, including physicians, mental health therapists, and medical students with diverse demographic characteristics. Reviewers provided written and verbal feedback regarding ambiguity and challenges during the completion of the survey. Particular attention was paid to the relevance, appropriateness, and cognitive load required to answer items. We incorporated changes to the survey during a structured debriefing session. Based on this feedback, we excluded the other 4 eHealth inquiries, including "select all that apply" phrasing for questions regarding behavioral changes and reasons for initiation of ChatGPT use and incorporating priming subsection headings to prepare respondents for each varied task.

Survey Population

The open survey was distributed to 21,499 members of ResearchMatch—a disease-neutral, Web-based recruitment registry—to help contact patients who have registered for eligibility to participate in clinical research studies [35]. Its 152,000 community reside in the Continental US and Puerto Rico, including those of all ages and races, both those in good health and those with health issues. ResearchMatch is supported by the US National Institutes of Health as part of the Clinical Translational Science Award program and is operated by the Vanderbilt University Medical Center, which maintains a large population of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. Participating health care systems around the country provide voluntary invitations for community members to join and make an account that allows for advertisements of available research studies. This population may preselect for a population with more dependable access to the internet such as White individuals with higher education and income. Platform moderators report that the majority of volunteers are female

XSL•FO
RenderX

(69.1%, n=105,032), older than 18 years (97.2%, n=147,744), and White (70.6%, n=107,312) with 40.5% (n=61,560) volunteers reporting having no medical conditions. The platform allows researchers to specify cohorts prior to distribution by age, gender, race, health issue, and location. However, for the purpose of this study, the only parameters used were age to ensure adult participation only (≥18 years) and selection of all 50 US States with the exclusion of Puerto Rico. To reach the 21,499 members, researchers invited respondents in 1000 and 1499 (maximum) respondent batches for a total of 11 outreach rounds over the course of 2 months.

Survey Administration

From June 10, 2023, to August 10, 2023, the survey was administered to consenting ResearchMatch members aged 18 years and older, with initial messaging being the only contact with participants. Initial contact with participants occurred within the internet-based contact platform within Research Match, wherein they were given a short message explaining that medical researchers investigating OHI were inviting them to participate in a research study. A link was attached to this email that allowed them to access the informed consent sheet and, upon consent, the survey. ReCAPTCHA technology was used to prevent bot survey abuse; however, no IP or cookie tracking was utilized to prevent duplicate entries. Informed consent for the survey was obtained from all participants as a cover page displayed prior to consent attestation. No personal identifying data was collected. The researchers incentivized participants to participate by offering them the opportunity to win one of 2 US $50 gift cards. A separate link to a Qualtrics form was used, and 2 respondents were randomly selected for each prize. Of the 21,499 participants given access to the survey link, 2406 participants completed the survey resulting in a response rate of 11.2%.

Data Analysis

Statistical analysis was performed with Stata Statistical Software (release 18; StataCorp LLC). Descriptive statistics were used to characterize the cohort. In addition, significance testing using *t* tests and Pearson chi-square tests was used to evaluate the differences between users of ChatGPT OHI and nonusers. Results were considered statistically significant at $P<.05$. All *P* values were 2-sided. No methods to weight items or propensity scores were used to adjust the nonrepresentative sample. Presumably, due to the raffle incentive, there was little missing data. Missing observations were simply excluded from individual analyses without imputation.

Results

How Do ChatGPT Users Compare to Nonusers?

Table 1 depicts the demographic and health characteristics of all respondents, ChatGPT users, and ChatGPT nonusers. Among all respondents, most were female, White, and at least college-educated, and had less than US $100,000 in annual household earnings. The average age was 37.6 years. Most respondents had a continuity-based usual source of care such as a doctor's office, health center, or VA, but a substantial number used urgent care centers or emergency departments. Consistent with the ResearchMatch sampling frame (individuals interested in participating in medical research), almost two-thirds rated their health as less than very good.

Among respondents, 517 (21.5%) were ChatGPT users and 1889 (78.6%) were nonusers. ChatGPT users were significantly younger than nonusers (32.8 vs 39.1 years, $P<.001$; Table 2). Compared with nonusers, ChatGPT OHI users were significantly more likely to identify as White (83.4% vs 78.6%, $P=.02$), earn less than US $100,000 annually (84.7% vs 73.1%, $P<.001$), and report educational attainment of less than a bachelor degree (50.1% vs 33%, $P<.001$, Table 1). Users were also more likely to use convenience or emergency (noncontinuity based) health care (52.2% vs 36.4%, $P<.001$), grade their health as less than very good (71.9% vs 63.8%, $P=.001$), and report heavy general OHI seeking frequency of weekly or more in the last 6 months (60.5% vs 49.3%, $P<.001$).

The Cronbach α for the scale measuring eLiteracy scores was 0.773, indicating good internal consistency among the items. Compared with nonusers, ChatGPT users displayed considerably lower eLiteracy scores ($P=.007$).

XSL•FO
RenderX

Table 1. Cohort demographic and health characteristics stratified by use of ChatGPT online health information (OHI).

Characteristic	All respondents (n=2406)	ChatGPT users (n=517)	Nonusers (n=1889)	*P* value
Age (years), mean (SD)	37.64 (13.76)	32.76 (7.00)	39.13 (14.92)	.001
Sex (second cycle, n=137)				
Male	28 (20.4)	__[a]	—	—
Female	101 (73.7)	—	—	—
Other	8 (5.8)	—	—	—
Race				.02
White	1834 (79.7)	431 (83.4)	1403 (78.6)	
Non-White	468 (20.3)	86 (16.6)	382 (21.4)	
Preferred language				.21
English	2244 (97.5)	500 (96.7)	1744 (97.7)	
Non-English	58 (2.5)	17 (3.3)	41 (2.3)	
Annual household income (US $)				<.001
Less than 100,000	1742 (75.7)	438 (84.7)	1304 (73.1)	
100,000 or more	560 (24.3)	79 (15.3)	481 (27.9)	
Education level				<.001
Less than BA/BS	848 (36.8)	259 (50.1)	589 (33)	
BA/BS or greater	1454 (63.2)	259 (49.9)	1196 (67)	
Preferred health care location				<.001
Continuity-based care (doctor's office, health center, or VA[b])	1345 (58.4)	247 (47.8)	1098 (58.1)	
Transient care (emergency department, urgent care, store clinic, or multiple)	957 (41.6)	270 (52.2)	687 (36.4)	
Health rating				.001
Good to poor	1510 (65.6)	372 (71.9)	1138 (63.8)	
Very good to excellent	791 (34.4)	145 (28.1)	646 (36.2)	
6-month OHI seeking frequency				<.001
Weekly or more	1183 (51.8)	313 (60.5)	870 (49.3)	
Monthly or less	1099 (48.2)	204 (39.5)	895 (50.7)	
Average eLiteracy Score (Cronbach α=0.773[c]), mean (SD)	3.82 (0.01)	3.75 (0.03)	3.84 (0.02)	.007

[a]Not available.

[b]VA: Veteran's Affairs.

[c]Generated score is an average based on scores from 4 selected eLiteracy items measured from strongly disagree (1) to strongly agree (5)

XSL•FO
RenderX

Table 2. Usage characteristics of ChatGPT online health information (OHI) users.

Characteristic (n=517)	Values, n (%)
Prior use of ChatGPT	439 (86.4)
Initiator of ChatGPT OHI use[a]	
Health care provider recommendation	224 (43.3)
Advertisement on website or app	216 (41.8)
Sponsored post or advertisement on social media	192 (37.1)
Search engine result	181 (35)
Friend or family recommendation for OHI use	179 (34.6)
Health website or web-based forum	120 (23.2)
News article or publication	109 (21.1)
Expansion of previous non–health-related ChatGPT use	100 (19.3)
Reasons for using ChatGPT for health information[a]	
Seeing if going to a health professional was necessary	245 (47.4)
Looking for additional or alternative treatment options	239 (46.2)
Clarifying or checking information given by a health professional	202 (39.1)
Limited time or insufficient information during a health consultation	194 (37.5)
Seeing if self-management is possible	181 (31)
Disagreed with health professional and wanted a different information source	97 (18.8)
Just out of interest	66 (12.8)
Other	12 (2.3)
Use frequency of ChatGPT OHI	
Once a month or less	68 (13.2)
More than monthly but less than once a month	91 (17.6)
About once a week	140 (27.1)
2-3 times a week	153 (29.6)
4-6 times a week	40 (7.7)
Daily or almost daily	10 (1.9)
Not applicable or single-use	6 (1.2)

[a]Respondents were allowed to select more than one response.

How Do ChatGPT OHI Users Characterize the Purpose and Frequency of Their Use?

Most ChatGPT users endorsed a prior nonmedical use of the technology, with use initiation being primarily influenced by health care provider (HCP) recommendations, advertisements on websites or apps, and sponsored posts or ads on social media (Table 2). The most cited reasons for initiating use were determining the necessity of visiting a health professional and exploring alternative treatment options.

About a quarter reported using ChatGPT for OHI for 6 months or longer, which aligns with the timeline of ChatGPT's introduction to the public. Use of ChatGPT OHI was frequent, with 40% (n=206) of respondents reporting use 2-3 times weekly or more.

How Do Users Characterize the Ease, Understanding, and Usefulness of ChatGPT OHI?

Almost all users could obtain health information from the tool in their preferred language (Table 3). Users were divided regarding the overall usefulness of ChatGPT OHI; however, 63% (n=317) considered this information to be better than other OHI sources. Moreover, 4 of 5 users deemed ChatGPT OHI to be at least as good an information source as their physician.

The Cronbach α for the user experience scale was 0.67, indicating acceptable internal consistency. Respondents reported a generally positive ChatGPT experience (mean of 3.74 on 5-point scale). Conversely, 68% of users suspected that some aspect of the received information from the AI was inaccurate.

XSL•FO
RenderX

Table 3. User experience characterization with ChatGPT online health information (OHI).

Characteristic (n=502)	Values
Able to obtain OHI from ChatGPT in preferred language, n (%)	493 (97.6)
User experience scale (ease, understanding, relevance), mean (SD)[a]	3.74 (0.03)
Overall usefulness of ChatGPT OHI, n (%)	
Poor to good	260 (51.8)
Very good to excellent	242 (48.2)
Usefulness compared with other OHI sources, n (%)	
Worst to much worse	48 (9.6)
Same as other OHI sources	112 (24.3)
Better to much better	317 (63.2)
Usefulness compared with information from MD[b] (better/much better), n (%)	
Worst to much worse	71 (14.1)
Same as MD	209 (41.6)
Better to much better	198 (39.4)
Suspected inaccuracy, n (%)	336 (67.9)
Presented information to doctor, n (%)	349 (67.5)

[a]Generated score is an average of scores from 3 questions pertaining to ease of use, understanding, and relevance of ChatGPT OHI, measured from Strongly Disagree (1) to Strongly Agree (5); (Cronbach α=0.645).
[b]MD: medical doctor.

What Are the Most Common Health Behaviors That Follow the Use of ChatGPT?

Users commonly presented results of ChatGPT OHI to a physician (68%, n=338). Moreover, as a result of ChatGPT OHI use, 42.9% (n=222) of users asked a doctor for clarification of information, 45.8% (n=237) for more information, 35.6% (n=184) for a specialty referral, and 31% (n=160) for a new or different prescription (Table 4).

Table 4. Behavioral implications of ChatGPT online health information (OHI) use stratified by use frequency (respondents were allowed to select more than one response).

Characteristic (n=502)	Values, n (%)
Behavior changes based on ChatGPT OHI	
Asked MD[a] for more information	237 (45.8)
Asked MD for clarification	222 (42.9)
Requested a test or referral	184 (35.6)
Self-medication/changed medication	160 (31)
Refused tests/meds	103 (19.9)
Scheduled appt	120 (23.2)
Canceled appt	79 (15.3)
No action taken	54 (10.4)

[a]MD: medical doctor.

Discussion

Principal Findings

This study is among the first to appraise patient use of ChatGPT-derived OHI. In this sample of participants in a national research cooperative, the use of ChatGPT for medical purposes was common, with users of the tool more likely to be White, have lower educational attainment, be in poor health, and receive care from noncontinuity based sources such as urgent care centers, retail clinics, and emergency departments. A large portion of users initiated use at the suggestion of a HCP and use of the tool was associated with altering appointments, changing medications, and consulting a physician as a result of their search.

Given the platform's recent introduction, adoption by over one-fifth of the sample is remarkable. While the highly engaged

XSL•FO
RenderX

ResearchMatch population may be inherently more likely to use new forms of OHI, ChatGPT appears poised for adoption by a growing proportion of the 81% of American adults who use OHI [36]. Moreover, the low response rate of this sample may limit the generalizability of the use patterns we observed.

The demographic profile of ChatGPT users reveals a possible digital divide in the realm of AI-driven OHI. Consistent with past studies depicting greater use of health information technologies by Whites compared with non-Whites [37,38], our findings suggest an inequity in the adoption of this source of OHI by minorities which may be explained by underserved minority individuals being more likely to have lower health literacy and less access to the internet [39-42].

Although heavier use of ChatGPT among respondents with lower educational attainment may challenge the standard view linking educational achievement, high digital literacy, and the use of health technology, it could be that ChatGPT's conversational format is more accessible to individuals without college or graduate degrees than other OHI sources [43-45]. Conversely, it is possible that the relative lack of use amongst the more educated is due to a possible distrust of the platform or, more conservatively, a tapered approach to adoption advised by a greater ability to assess the efficacy of the tool for OHI. Still, it is also possible that this finding is idiosyncratic, a function of our nonrepresentative survey sample.

ChatGPT OHI users were also found to have lower eHealth literacy, a finding that is seemingly paradoxical. However, one plausible interpretation of the finding that ChatGPT users have both lower eHealth literacy and lower educational attainment than their nonusing counterparts is less sophisticated health care consumers may benefit differentially from the conversational format of ChatGPT, while more educated and internet-savvy individuals may be able to satisfy their information needs in other ways (such as the use of traditional search engines, subscriptions to health education blogs such as the Harvard Health Letter, etc).

The preference of individuals in poorer health for utilizing ChatGPT as a health information resource aligns with the broader understanding that individuals with more significant health challenges may have increased information needs [46,47]. This heightened demand may stem from the necessity to understand and manage complex health conditions, leading to a more active pursuit of diverse information sources, including AI platforms like ChatGPT. Moreover, as ChatGPT users prefer transient health care more often, those accustomed to expedient, on-demand care are likely more inclined to gravitate towards ChatGPT OHI's mode of instantaneous personalized information [46]. Moreover, as increased OHI use is associated with barriers to traditional health care access, it is likely that when faced with barriers to health care—such as cost, availability, and accessibility—individuals turn to alternative information sources, including ChatGPT [48]. This presents physicians with a unique opportunity to identify and counsel patients about the value and limitations of using LLMs for health information, offering to be available in the appropriate context (eg, a future office visit) to help interpret the information received. This partnership not only bridges information gaps but also reinforces the physician's role in collaborating with patients as they navigate their health information journey. Moreover, as physicians become more familiar with the abilities and inabilities of ChatGPT and similar programs, they can more effectively counsel patients on the prudent use of such AI resources in complementing ongoing medical care.

While most users came to the tool through advertisements or social media posts, a substantial number reported using it at the suggestion of HCPs. The percentage of physicians who see digital health tools as an advantage for patient care grew from 85% in 2016 to 93% in 2022, corresponding to a similar increase use of digital health tools by physicians [49]. Moreover, as approximately 10% of US HCPs endorse the use of AI, with another 50% considering future use, HCPs' personal experiences with ChatGPT may influence their willingness to recommend them to patients, reflecting a growing confidence in the utility of such AI tools for patient care [50]. Considering that almost two-thirds of ChatGPT users reported ChatGPT use once a week or more, many respondents likely trusted their HCP's recommendation. However, it is important to clarify that this survey item instructed respondents to pick this option as a reason for the recommendation, even if the HCP was a family member or friend. This may have augmented the number of individuals who report having been recommended the tool by a care provider. A poor understanding of who may qualify as a HCP may also account for this unexpectedly high referral rate, considering the novelty and skepticism surrounding ChatGPT OHI. Nonetheless, patient awareness and confidence in ChatGPT within this study have arguably outpaced the completion of rigorous studies of the tool's efficacy and accuracy in delivering health recommendations.

For about half of respondents, ChatGPT use was followed by some health behavior such as formalized care-seeking, asking for information, and asking for further action, including setting or canceling an appointment or requesting new testing. OHI has been shown in multiple studies to stimulate care-seeking and clinical question-asking [51,52]. It is remarkable, however, that ChatGPT, despite its relative infancy, appears to promote behavioral change rates similar to more established sources of OHI. Further, given that most patients reported care-seeking behaviors despite also believing some aspect of the information received was inaccurate, it is promising that verifying the information with a physician arose as a leading resultant health behavior. This aligns with previous studies that emphasize the role of physicians in verifying traditional OHI, suggesting that the traditional physician-patient relationship may endure as physicians who can identify likely ChatGPT OHI users will be able to counsel regarding the efficacy and accuracy of the information received [53,54].

User perceptions of their ChatGPT OHI experience along the dimensions of comparable usefulness, ease of use, and suspected inaccuracy raise several intriguing contrasts. Around 52% (n=260) of respondents rated ChatGPT to be overall poor to good in usefulness, while more than 80% (n=429) rated it to be better than or as good as other OHI sources. Users may appreciate certain aspects of ChatGPT, such as the conversational interaction, user interface, and quick reception of information, which may not be as easily executed on other

XSL•FO
RenderX

OHI platforms. It may be the case that users who rated in this manner more heavily weighted the advantages of ChatGPT's information delivery and accessibility over their perception of inaccuracy or information usefulness. Further, the dichotomy between the concerns for inaccuracy and the somewhat favorable ratings on usefulness suggests a nuanced understanding of "usefulness" by the users. While our data may suggest action is likely after the use of ChatGPT, many patients may find value in ChatGPT as a starting point for health information or as a means to facilitate discussions with HCPs, a relationship past research has supported for traditional OHI [53,54].

Thus, during the digital health information age, a balance must be held between the desire for rapid information acquisition and the need for accurate, trustworthy advice. Even more importantly, it is important to consider user information-seeking experience when evaluating these OHI tools, as this may play an important role in the uptake and appraisal of the tools' usefulness to patients.

Limitations and Future Directions

While shedding light on the adoption and utilization of ChatGPT for OHI, this study carries 4 principal limitations. First, the study's cross-sectional design limits the understanding of causal relationships and does not capture behavior over time. Rather than prompting particular health-related decisions or behaviors, the use of ChatGPT may instead have been a consequence of preexisting decisions or behaviors. Second, as a self-administered survey, this study is subject to reporting and recall bias. Third, two features limit the generalizability of the results: (1) the ResearchMatch population skews younger, less ethnically diverse, and arguably more engaged with health and health care than the US population as a whole; and (2) despite similar demographic profiles, our sample may differ from the ResearchMatch population in other (unmeasured) ways [55]. These differences may have influenced perceptions and experiences with ChatGPT in the OHI context. Future investigations should focus on diverse sampling strategies to include participants of varying degrees of education and digital literacy to examine how different populations engage with ChatGPT for OHI. Fourth, we could not directly examine the specific content accessed by ChatGPT users nor the completeness or accuracy of the information received.

Conclusions

This study revealed a swift adoption of ChatGPT, particularly among younger patients with poorer health and those using transient (noncontinuity-based) forms of health care. As ChatGPT appears to influence both intensity and types of care-seeking behavior, physicians and other HCPs must proactively identify and counsel patients on best practices for the use of this emerging technology. This skill will be vital in preserving the integrity of the physician-patient relationship and ensuring safe and effective health care in an increasingly AI-driven digital world. More research is needed to understand how patients and physicians can work together to make optimal use of these powerful but potentially hazardous tools.

Acknowledgments

The authors would like to thank the health care providers and community members who provided feedback and testing for this study including but not limited to: Olivia Campa MD (UC Davis), Ayodeji Ajibola MD (Kaiser Permanente Roseville), Elizabeth Burner MD MPH (Keck School of Medicine of USC), Adetokunbo Ajibola MSN, Madeline Stein BS (Antioch University), and Lauren Estess BS (Tufts School of Medicine). All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee support for the publication of this article.

Authors' Contributions

OAA, JR, and RLK were responsible for conceptualization, data curation, formal analysis, investigation, writing the original draft, and review and editing. RJD and MEL were involved in data curation, formal analysis, investigation, and review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
REDCap (Research Electronic Data Capture) questionnaire.
[PDF File (Adobe PDF File), 434 KB - jmir_v26i1e55138_app1.pdf]

Multimedia Appendix 2
Health behaviors model for uptake and adherence to ChatGPT health information.
[PNG File , 153 KB - jmir_v26i1e55138_app2.png]

References

1. Cohen R. Health insurance coverage: early release of estimates from the national health interview survey, 2018. In: National Health Interview Survey Early Release Program. Maryland, U.S: NATIONAL CENTER FOR HEALTH STATISTICS; 2018.

XSL•FO
RenderX

2. Goldberg R. More than half of American households used the internet for health-related activities in 2019, NTIA data show. National Telecommunications and Information Administration. URL: https://www.ntia.gov/blog/2020/more-half-american-households-used-internet-health-related-activities-2019-ntia-data-show [accessed 2023-10-03]
3. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
4. White RW, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Trans Inf Syst 2009;27(4):1-37. [doi: 10.1145/1629096.1629101]
5. Swoboda CM, Van Hulle JM, McAlearney AS, Huerta TR. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the health information national trends survey (HINTS). BMC Fam Pract 2018;19(1):146 [FREE Full text] [doi: 10.1186/s12875-018-0805-7] [Medline: 30157770]
6. ChatGPT: optimizing language models for dialogue. URL: https://openai.com/blog/chatgpt [accessed 2023-04-15]
7. These 20 jobs are the most "exposed" to AI, ChatGPT, researchers say. CBS News. URL: https://www.cbsnews.com/news/jobs-chatgpt-occupations-industries-exposed-teacher-telemarketer/ [accessed 2023-10-03]
8. Marr B. Is google's reign over? ChatGPT emerges as a serious competitor. Forbes. URL: https://www.forbes.com/sites/bernardmarr/2023/02/20/is-googles-reign-over-chatgpt-emerges-as-a-serious-competitor/ [accessed 2023-07-06]
9. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2023-04-15]
10. Davis R, Eppler M, Ayo-Ajibola O, Loh-Doyle JC, Nabhani J, Samplaski M, et al. Evaluating the effectiveness of artificial intelligence-powered large language models application in disseminating appropriate and readable health information in urology. J Urol 2023;210(4):688-694. [doi: 10.1097/JU.0000000000003615] [Medline: 37428117]
11. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33(6):1790-1796 [FREE Full text] [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]
12. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
13. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
14. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
15. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and Google search as sources of postoperative patient instructions. JAMA Otolaryngol Head Neck Surg 2023;149(6):556-558 [FREE Full text] [doi: 10.1001/jamaoto.2023.0704] [Medline: 37103921]
16. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model Chatbot responses to online patient eye care questions. JAMA Netw Open 2023;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
17. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
18. Barrington NM, Gupta N, Musmar B, Doyle D, Panico N, Godbole N, et al. A bibliometric analysis of the rise of ChatGPT in medical research. Med Sci (Basel) 2023;11(3):61 [FREE Full text] [doi: 10.3390/medsci11030061] [Medline: 37755165]
19. Magrabi F, Ong MS, Runciman W, Coiera E. Patient safety problems associated with heathcare information technology: an analysis of adverse events reported to the US food and drug administration. AMIA Annu Symp Proc 2011;2011:853-857 [FREE Full text] [Medline: 22195143]
20. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. Account Res 2023:1-9. [doi: 10.1080/08989621.2023.2168535] [Medline: 36697395]
21. Goodman RS, Patrinely Jr JR, Osterman T, Wheless L, Johnson DB. On the cusp: considering the impact of artificial intelligence language models in healthcare. Med 2023;4(3):139-140 [FREE Full text] [doi: 10.1016/j.medj.2023.02.008] [Medline: 36905924]
22. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625 [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
23. Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the internet for medical information. J Gen Intern Med 2002;17(3):180-185 [FREE Full text] [doi: 10.1046/j.1525-1497.2002.10603.x] [Medline: 11929503]
24. Murray E, Lo B, Pollack L, Donelan K, Catania J, White M, et al. The impact of health information on the internet on the physician-patient relationship: patient perceptions. Arch Intern Med 2003;163(14):1727-1734. [doi: 10.1001/archinte.163.14.1727] [Medline: 12885689]

XSL•FO
RenderX

25. Sumayyia MD, Al-Madaney MM, Almousawi FH. Health information on social media. perceptions, attitudes, and practices of patients and their companions. Saudi Med J 2019;40(12):1294-1298 [FREE Full text] [doi: 10.15537/smj.2019.12.24682] [Medline: 31828284]
26. Harris J, Atkinson A, Mink M, Porcellato L. Young people's experiences and perceptions of YouTuber-produced health content: implications for health promotion. Health Educ Behav 2021;48(2):199-207 [FREE Full text] [doi: 10.1177/1090198120974964] [Medline: 33246387]
27. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
28. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the health belief model. Health Educ Q 1988;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]
29. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
30. National health interview survey [public use data file and documentation]. National Center for Health Statistics. 2023. URL: https://www.cdc.gov/nchs/nhis/2023nhis.htm [accessed 2023-10-13]
31. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
32. Macias C, Gold PB, Öngür D, Cohen BM, Panch T. Are single-item global ratings useful for assessing health status? J Clin Psychol Med Settings 2015. [doi: 10.1007/s10880-015-9436-5] [Medline: 26492891]
33. Zhang D, Shi Z, Hu H, Han GK. Classification of the use of online health information channels and variation in motivations for channel selection: cross-sectional survey. J Med Internet Res 2021;23(3):e24945 [FREE Full text] [doi: 10.2196/24945] [Medline: 33687342]
34. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. J Med Internet Res 2011;13(1):e20 [FREE Full text] [doi: 10.2196/jmir.1600] [Medline: 21345783]
35. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med 2012;87(1):66-73 [FREE Full text] [doi: 10.1097/ACM.0b013e31823ab7d2] [Medline: 22104055]
36. Fox S, Duggan M. Health online 2013. Pew Research Center: Internet, Science & Tech. URL: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ [accessed 2023-07-21]
37. Ailawadhi S, Ailawadhi M, Dutta N, Parrondo RD, Roy V, Sher T, et al. The digital divide: racial disparities in adoption and utilization of health information technology among patients with lymphoid cancers. Cancer Med 2023;12(18):19013-19020 [FREE Full text] [doi: 10.1002/cam4.6454] [Medline: 37587868]
38. Ojinnaka CO, Adepoju OE. Racial and ethnic disparities in health information technology use and associated trends among individuals living with chronic diseases. Popul Health Manag 2021;24(6):675-680. [doi: 10.1089/pop.2021.0055] [Medline: 33989085]
39. Ndugga N, Published S. Disparities in health and health care: 5 key questions and answers. KFF. URL: https://www.kff.org/racial-equity-and-health-policy/issue-brief/disparities-in-health-and-health-care-5-key-question-and-answers/ [accessed 2023-11-10]
40. Internet, broadband fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2023-11-10]
41. Christopher Gibbons M. Use of health information technology among racial and ethnic underserved communities. Perspect Health Inf Manag 2011;8(Winter):1f [FREE Full text] [Medline: 21307989]
42. Kontos EZ, Bennett GG, Viswanath K. Barriers and facilitators to home computer and internet use among urban novice computer users of low socioeconomic position. J Med Internet Res 2007;9(4):e31 [FREE Full text] [doi: 10.2196/jmir.9.4.e31] [Medline: 17951215]
43. Cotten SR, Gupta SS. Characteristics of online and offline health information seekers and factors that discriminate between them. Soc Sci Med 2004;59(9):1795-1806. [doi: 10.1016/j.socscimed.2004.02.020] [Medline: 15312915]
44. Nguyen A, Mosadeghi S, Almario CV. Persistent digital divide in access to and use of the Internet as a resource for health information: results from a california population-based study. Int J Med Inform 2017;103:49-54. [doi: 10.1016/j.ijmedinf.2017.04.008] [Medline: 28551001]
45. McCloud RF, Okechukwu CA, Sorensen G, Viswanath K. Beyond access: barriers to internet health information seeking among the urban poor. J Am Med Inform Assoc 2016;23(6):1053-1059 [FREE Full text] [doi: 10.1093/jamia/ocv204] [Medline: 27206459]
46. Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the internet for health information? Health Serv Res 2006;41(3 Pt 1):819-836 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00510.x] [Medline: 16704514]
47. Bansil P, Keenan NL, Zlot AI, Gilliland JC. Health-related information on the web: results from the healthstyles survey, 2002-2003. Prev Chronic Dis 2006;3(2):A36 [FREE Full text] [Medline: 16539777]

XSL•FO
RenderX

48. Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL. Access to care and use of the Internet to search for health information: results from the US national health interview survey. J Med Internet Res 2015;17(4):e106 [FREE Full text] [doi: 10.2196/jmir.4126] [Medline: 25925943]
49. 5 insights into how physicians view, use digital health tools. AMA. URL: https://www.ama-assn.org/practice-management/digital/5-insights-how-physicians-view-use-digital-health-tools [accessed 2023-11-10]
50. Perceptions of AI in healthcare: what professionals and the public think. The intake. URL: https://www.tebra.com/theintake/medical-deep-dives/tips-and-trends/research-perceptions-of-ai-in-healthcare [accessed 2023-11-10]
51. Farnood A, Johnston B, Mair FS. A mixed methods systematic review of the effects of patient online self-diagnosing in the 'smart-phone society' on the healthcare professional-patient relationship and medical authority. BMC Med Inform Decis Mak 2020;20(1):253 [FREE Full text] [doi: 10.1186/s12911-020-01243-6] [Medline: 33023577]
52. Bujnowska-Fedak MM, Węgierek P. The impact of online health information on patient health behaviours and making decisions concerning health. Int J Environ Res Public Health 2020;17(3):880 [FREE Full text] [doi: 10.3390/ijerph17030880] [Medline: 32023828]
53. Sommerhalder K, Abraham A, Zufferey MC, Barth J, Abel T. Internet information and medical consultations: experiences from patients' and physicians' perspectives. Patient Educ Couns 2009;77(2):266-271. [doi: 10.1016/j.pec.2009.03.028] [Medline: 19411157]
54. Li S, Wang K. Sharing online health information with physicians: understanding the associations among patient characteristics, directness of sharing, and physician-patient relationship. Front Psychol 2022;13:839723 [FREE Full text] [doi: 10.3389/fpsyg.2022.839723] [Medline: 35432074]
55. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med 2012;87(1):66-73 [FREE Full text] [doi: 10.1097/ACM.0b013e31823ab7d2] [Medline: 22104055]

Abbreviations

AI: artificial intelligence
HCP: health care provider
LLM: large language model
OHI: online health information
REDCap: Research Electronic Data Capture

Edited by S Ma; submitted 04.12.23; peer-reviewed by TAR Sure, S Kommireddy, K Kaphingst; comments to author 17.03.24; revised version received 27.03.24; accepted 15.04.24; published 14.08.24.

Please cite as:
Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL
Characterizing the Adoption and Experiences of Users of Artificial Intelligence–Generated Health Information in the United States: Cross-Sectional Questionnaire Study
J Med Internet Res 2024;26:e55138
URL: https://www.jmir.org/2024/1/e55138
doi:10.2196/55138
PMID:

©Oluwatobiloba Ayo-Ajibola, Ryan J Davis, Matthew E Lin, Jeffrey Riddell, Richard L Kravitz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment

Damien Gibson[1,2,3*], BBiomedSc, MS, MD; Stuart Jackson[4*], BEd, BSc, MS, MD; Ramesh Shanmugasundaram[1,2], MS, MD; Ishith Seth[5], BBiomed, MS, MD; Adrian Siu[3,6], BPharm, MS, MD; Nariman Ahmadi[7,8], BSc, MBBS, MS; Jonathan Kam[9], BMed, BMedSci, MD; Nicholas Mehan[9], MD; Ruban Thanigasalam[7,8], MBBS, MS; Nicola Jeffery[7,8], BSc, MBBS; Manish I Patel[4,10], MBBS, MMed, PhD; Scott Leslie[3,4,7,8], BSc, MBBS

[1]Department of Urology, Saint George Hospital, Kogarah, Australia
[2]Faculty of Medicine, The University of New South Wales, Sydney, Australia
[3]Surgical Outcomes Research Centre, Sydney, Australia
[4]Faculty of Medicine, University of Sydney, Sydney, Australia
[5]Department of Surgery, Peninsula Health, Victoria, Australia
[6]Concord Institute of Academic Surgery, Concord Hospital, Sydney, Australia
[7]Department of Urology, Chris O'Brien Lifehouse, Sydney, Australia
[8]Royal Prince Alfred Hospital Institute of Academic Surgery, Royal Prince Alfred Hospital, Sydney, Australia
[9]Nepean Urology Research Group, Nepean Hospital, Sydney, Australia
[10]Department of Urology, Westmead Hospital, Sydney, Australia
[*]these authors contributed equally

Corresponding Author:
Damien Gibson, BBiomedSc, MS, MD
Department of Urology
Saint George Hospital
Gray St
Kogarah, 2217
Australia
Phone: 61 (02) 9113 1111
Email: Damien.p.gibson@gmail.com

Abstract

Background: Artificial intelligence (AI) chatbots, such as ChatGPT, have made significant progress. These chatbots, particularly popular among health care professionals and patients, are transforming patient education and disease experience with personalized information. Accurate, timely patient education is crucial for informed decision-making, especially regarding prostate-specific antigen screening and treatment options. However, the accuracy and reliability of AI chatbots' medical information must be rigorously evaluated. Studies testing ChatGPT's knowledge of prostate cancer are emerging, but there is a need for ongoing evaluation to ensure the quality and safety of information provided to patients.

Objective: This study aims to evaluate the quality, accuracy, and readability of ChatGPT-4's responses to common prostate cancer questions posed by patients.

Methods: Overall, 8 questions were formulated with an inductive approach based on information topics in peer-reviewed literature and Google Trends data. Adapted versions of the Patient Education Materials Assessment Tool for AI (PEMAT-AI), Global Quality Score, and DISCERN-AI tools were used by 4 independent reviewers to assess the quality of the AI responses. The 8 AI outputs were judged by 7 expert urologists, using an assessment framework developed to assess accuracy, safety, appropriateness, actionability, and effectiveness. The AI responses' readability was assessed using established algorithms (Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, The Coleman-Liau Index, and Simple Measure of Gobbledygook [SMOG] Index). A brief tool (Reference Assessment AI [REF-AI]) was developed to analyze the references provided by AI outputs, assessing for reference hallucination, relevance, and quality of references.

Results: The PEMAT-AI understandability score was very good (mean 79.44%, SD 10.44%), the DISCERN-AI rating was scored as "good" quality (mean 13.88, SD 0.93), and the Global Quality Score was high (mean 4.46/5, SD 0.50). Natural Language

XSL•FO
RenderX

Assessment Tool for AI had pooled mean accuracy of 3.96 (SD 0.91), safety of 4.32 (SD 0.86), appropriateness of 4.45 (SD 0.81), actionability of 4.05 (SD 1.15), and effectiveness of 4.09 (SD 0.98). The readability algorithm consensus was "difficult to read" (Flesch Reading Ease score mean 45.97, SD 8.69; Gunning Fog Index mean 14.55, SD 4.79), averaging an 11th-grade reading level, equivalent to 15- to 17-year-olds (Flesch-Kincaid Grade Level mean 12.12, SD 4.34; The Coleman-Liau Index mean 12.75, SD 1.98; SMOG Index mean 11.06, SD 3.20). REF-AI identified 2 reference hallucinations, while the majority (28/30, 93%) of references appropriately supplemented the text. Most references (26/30, 86%) were from reputable government organizations, while a handful were direct citations from scientific literature.

Conclusions: Our analysis found that ChatGPT-4 provides generally good responses to common prostate cancer queries, making it a potentially valuable tool for patient education in prostate cancer care. Objective quality assessment tools indicated that the natural language processing outputs were generally reliable and appropriate, but there is room for improvement.

(J Med Internet Res 2024;26:e55939) doi:10.2196/55939

KEYWORDS

prostate cancer; patient education; large language model; ChatGPT; AI language model; multimetric assessment; artificial intelligence; AI; AI chatbots; health care professional; health care professionals; men; man; prostate; cancer; decision-making; prostate specific; antigen screening; medical information; natural language processing; NLP

Introduction

Artificial intelligence (AI) chatbots have made significant strides in recent years [1]. This was emphatically signposted with the launch of ChatGPT-3 (OpenAI) [2] in November 2022, with ChatGPT becoming the most popular web-based tool for both patients and health care professionals [3,4]. Now in its fourth iteration (ChatGPT-4), the AI language model can generate responses to a wide range of health questions and topics [5]. AI chatbots, such as ChatGPT, have the potential to significantly impact patient education and disease experience by providing reliable, accessible, and personalized information [5,6]. One patient population that stands to benefit from this is men who are concerned about prostate cancer.

With the rising prevalence of prostate cancer globally—accounting for an estimated 1,414,259 new cases and over 375,304 deaths in 2020 alone—there is an urgent need for accurate and timely patient education information [7]. The rate of prostate cancer survivorship is increasing, but this comes with its own challenges such as escalating health care costs and large numbers of survivors requiring ongoing care [4]. In this context, shared decision-making becomes pivotal, particularly concerning prostate-specific antigen screening and prostate cancer treatment selection [4]. Given the various treatments available, management decisions can be greatly influenced by a patient's understanding of the anatomical, functional, and psychological impacts of treatment [8]. Side effects, such as urinary incontinence and erectile dysfunction, can severely affect a patient's quality of life, necessitating well-informed patients, and treatment choices [9]. Furthermore, patient education has been shown to minimize psychological impacts such as depression and treatment regret [10].

There are well-documented issues with unmet information needs of both men and their support networks throughout the prostate cancer care continuum [11]. This includes challenges related to information quality and readability [12]. The assessment of web-based health care information in prostate cancer has been well described through multiple domains including web page articles, YouTube (Google), and social media [11]. The internet is now often the first source of information for men (and their stakeholders) seeking answers about diagnosis, treatment, and prognosis [9]. Despite this trend, most long-term literature suggests that web-based health information is of moderate to poor quality [11-13].

AI chatbots are a potential solution to fill the prostate cancer information quality gap [3]. Given their scalability, AI chatbots can reach a wide demographic, including those in remote or underserved communities where medical resources are scarce [3]. Natural language processing technologies (NLPTs) enable these platforms to present complex jargon in patient-specific terms, with the potential to address eHealth literacy variability, and to enhance patient understanding [14]. Such platforms are also able to do this across a diverse number of languages [15]. Despite these qualities, the accuracy and reliability of AI chatbot medical information must still be assessed using rigorous evaluation tools. Only a handful of studies have begun to test ChatGPT's applicability in prostate cancer: one testing its knowledge directly with questions and statements [16] and another assessing its appropriateness in screening recommendations [17]. However, a significant knowledge gap persists in understanding the quality and safety of information patients receive from ChatGPT-4 for common internet queries. Ongoing evaluation is a necessary step to build health care provider confidence in these new technologies while ensuring that patients have access to vetted and safe health care and educational information.

This study aims to demonstrate and assess the quality of ChatGPT responses to commonly asked patient education topics in prostate cancer care. By doing so, this study seeks to (1) illustrate to clinicians whether ChatGPT-4 is currently a reliable and safe patient education tool for prostate cancer information and (2) provide clinicians with a greater understanding of the current strengths and limitations of health-based queries which patients are likely to encounter when using technologies such as ChatGPT-4.

A range of assessment tools will be applied to the AI-generated responses to assess output quality, safety, understandability, actionability, ease of use, readability, and reliability. A parallel

XSL•FO
RenderX

assessment of the outputs by prostate cancer experts will also be conducted.

Methods

Question or Keyword Strategy

Questions tested with the AI chatbot model (ChatGPT-4) were selected through an iterative process of literature and Google Keyword analysis. Literature concerning the information needs of men considering prostate cancer investigation and treatment was reviewed to determine the most common information topics and prostate cancer questions of interest to men [11,18-21]. Subsequently, worldwide Google Trends data were analyzed to provide a more current public measure of prostate cancer information searches [22]. Using "prostate cancer" as a keyword, both rising and top "related topics" and "related queries" of the past year were collected. Finally, while limited to training materials up to 2021, ChatGPT was itself queried, asking "What are the most common prostate cancer questions asked to ChatGPT?" (Multimedia Appendix 1). The 2 authors thematically analyzed this information to define the following eight questions to discuss with the AI model: (1) What are the symptoms of prostate cancer? (2) What are the risk factors for prostate cancer? (3) What is the survival rate of prostate cancer? (4) How is prostate cancer diagnosed? (5) What age should men start getting screened for prostate cancer? (6) What are the pros and cons of treatment options for prostate cancer? (7) How does prostate cancer affect sexual function? and (8) How does prostate cancer affect bladder function?

Each question was posed to ChatGPT-4 with an additional request for references. A new ChatGPT account was established with a novel email address for each prompt in an effort to reduce the potential effects of each response on subsequent outputs of the AI model. Each output was recorded for individual quality and readability assessment (Multimedia Appendix 2).

Quality Assessment

Overview

Due to the current absence of tools to evaluate the quality of AI natural language outputs, each conversation was evaluated using modified versions of pre-existing information quality assessment tools. These included the Patient Education Materials Assessment Tool (PEMAT) and DISCERN criteria [23,24]. These tools were iteratively modified to accommodate the text-only nature and characteristics of AI natural language outputs. While DISCERN criteria have been adapted in literature, the PEMAT modification is new [25]. Internal validity testing was undertaken by 4 reviewers using ChatGPT outputs from similar question sets for breast cancer and bowel cancer. The reliability of each tool tested was satisfactory, with Cronbach $\alpha>0.8$ (DISCERN 0.852, PEMAT 0.82, and Global Quality Score [GQS] 0.85). The GQS was not modified [26].

PEMAT-AI Tool

The PEMAT tool evaluates and compares the understandability and actionability of patient education materials [24]. The tool incorporates 17 items measuring understandability and 7 assessing actionability; these were reduced to 8 and 3, respectively, to suit the AI text-only outputs (Multimedia Appendix 3). Each item was given a single point if the criteria were met, and the total score was measured as a total percentage. Final scores were recorded as "pass" or "fail" based on the ≥70% cut - off score set by the PEMAT guidelines [24].

DISCERN-AI Tool

The DISCERN criteria is a previously validated tool that aids health care consumers and health practitioners in appraising the quality of health care treatment information [23,24]. To address the AI output, these criteria were modified to 7 questions (of the original 15) on a scale of 1 to 3, using questions 3-9 (Multimedia Appendix 4). Based on previous DISCERN quality assessment in the literature, each output was scored as very poor (6), poor (7-9), fair (8-12), good (13-15), and excellent (16-18) quality patient education material [27,28].

GQS Tool

The GQS is a 5-point Likert scale based on the quality of information, and the flow and ease of use of information presented via the web. The GQS encompasses a scale of 1 to 5; where 1 indicates "low quality" and 5 implies "high quality." Results that received a score of 4 or 5 were rated high quality, those with a score of 3 were assessed as medium quality, and the ones with a score of 1 or 2 were categorized as low quality [25,29].

Readability

The readability of the AI responses was assessed using a battery of established algorithms: the Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, Coleman-Liau Index, and Simple Measure of Gobbledygook (SMOG) Index [30-33]. Multiple tools were used in an effort to limit the bias of each respective algorithm [34]. Each AI output text was copied to Microsoft Word to maintain formatting, and then to Readable.com for analysis [35,36]. Results from the answered questions were averaged across all outputs into a readability consensus [37]. The Flesch Reading Ease score gauges text simplicity, where the score ranges from 0 to 100, with higher scores indicating easier readability. Texts with a score between 60 and 70 are generally considered to be at an eighth- to ninth-grade reading level and are usually easier for the average adult to read. The Gunning Fog Index and The Flesch-Kincaid Grade Level measure sentence complexity, the score represents the number of years of formal education a reader would need to understand the text on the first reading. For example, a score of 12 would mean the text is suitable for a 12th-grade reading level or higher. The Coleman-Liau Index is similar to Gunning Fog and The Flesch-Kincaid Grade Level but focuses on character count. This score also correlates with a US school grade level but is calculated using the number of characters instead of syllables, making it more suited for languages where syllable count is less indicative of complexity. The SMOG Index evaluates syllable density to assess readability and is often used for checking health messages. A score of 12 would mean the text is suitable for someone with at least a 12th-grade level of reading comprehension.

XSL•FO
RenderX

Natural Language Assessment Tool for AI

Expert review of each output was undertaken by 7 independent experienced urologists, using an assessment framework (Natural Language Assessment Tool for AI [NLAT- AI]) developed to assess the accuracy, safety, appropriateness, actionability, and effectiveness of information. Each domain was scored on a 5-point Likert scale (1=strongly disagree, to 5="strongly agree"; Multimedia Appendix 5). All results were collated and presented as descriptive statistics. Qualitative feedback on each domain was sought regarding potential improvement and overall performance.

References Assessment

Due to known issues of AI hallucination: "the phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input," a final brief tool (Reference Assessment AI [REF-AI]) was developed for analysis of the references provided by AI outputs [38]. Each reference was reviewed by accessing the content via the direct link provided by the AI output, or a Google search of the reference. This tool assessed for reference hallucination (real or not), relevance (correlation between the references and AI output), and quality of references (type of institution linked to the reference). Each criterion was assessed with a score of 1-3, with a lower summative score indicating lower reference quality, and a higher score indicating high reference quality (Multimedia Appendix 6). Scores were averaged to yield a composite score for each axis of evaluation. The reliability of this tool tested similar question sets for breast cancer and bowel cancer was satisfactory (0.81).

Ethical Considerations

After consultation with the local institutional review board, it was determined that no formal ethical approval was required for this study as no human or animal participants were involved.

Results

ChatGPT Outputs

The responses generated by the AI model, ChatGPT-4, provided broad, medically aligned information (Multimedia Appendix 2). The assessment of the ChatGPT-4 output using PEMAT-AI, DISCERN-AI, and GQS patient education material assessment tools demonstrated high results across all tools. The pooled PEMAT-AI understandability score easily passed the acceptability threshold of >70% (mean 79.44%, SD 10.44%); only question 3 failed the >70% threshold at 66.67% while the remaining were 76% or greater (Figure 1). The pooled DISCERN-AI rating was scored as "good" quality 77% (mean 13.88, SD 0.93), and all individual questions rated "good" on the DISCERN-AI except for question 5, which scored excellent (mean 15.67; Figure 2). The pooled GQS was rated as high (mean 4.46, SD 0.50 out of 5; Figure 3). Assessment tool results for each question are tabulated and graphed (Table 1 and Figures 1-3). Reliability testing was high with Cronbach α=0.846.

Figure 1. PEMAT-AI mean score by ChatGPT question output. PEMAT-AI: Patient Education Materials Assessment Tool for Artificial Intelligence.

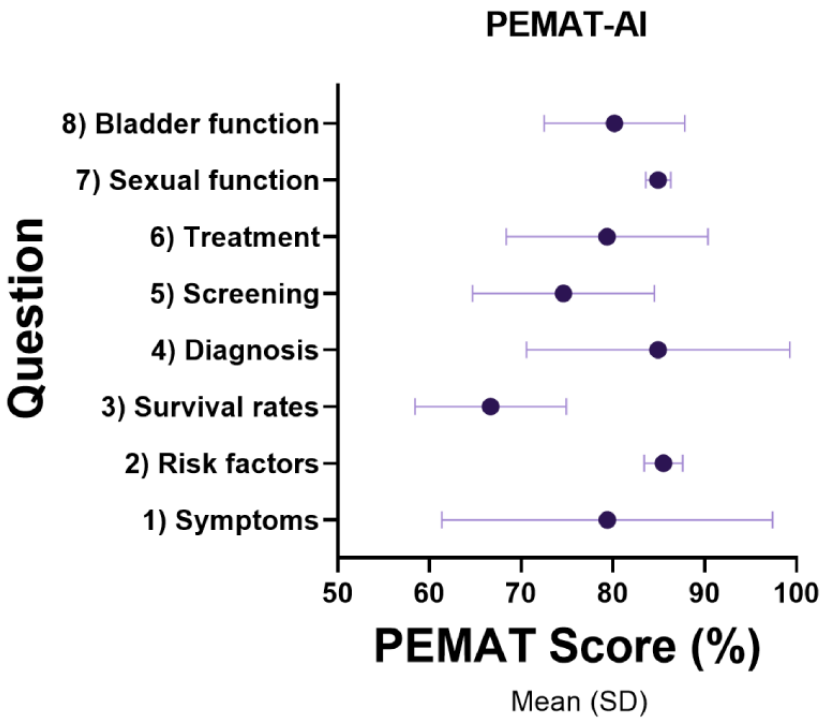

XSL•FO
RenderX

Figure 2. DISCERN-AI mean score by ChatGPT question output. AI: artificial intelligence.

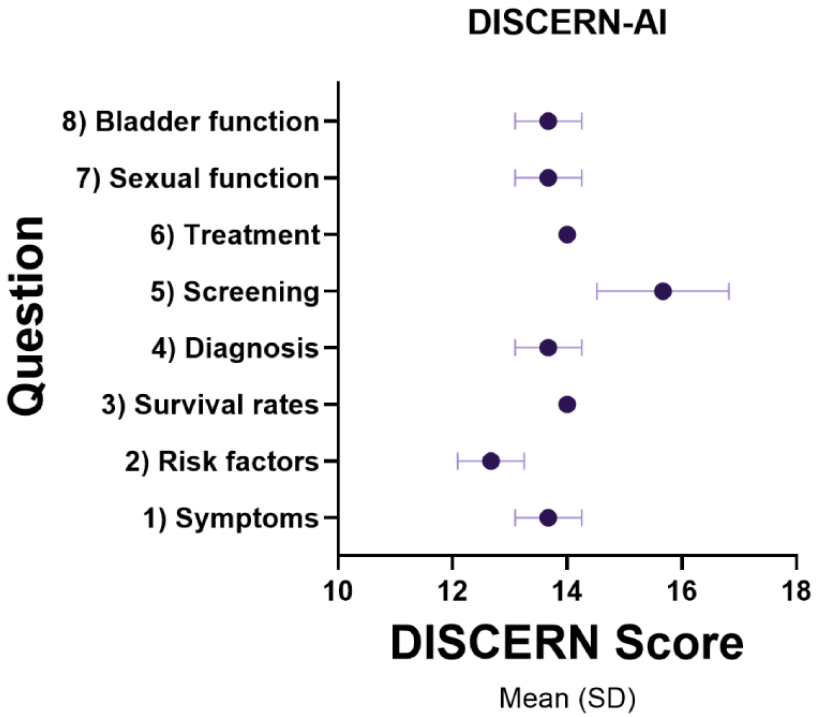

Figure 3. GQS mean score by ChatGPT question output. GQS: Global Quality Score.

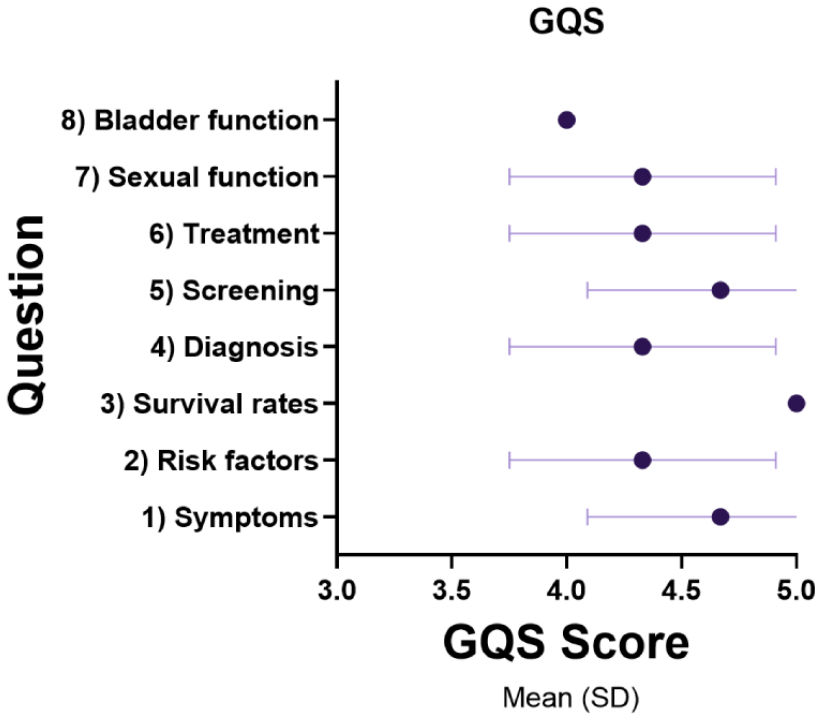

XSL•FO
RenderX

Table 1. Quality assessment tools.

Assessment	PEMAT-AI[a], mean (SD)	DISCERN-AI, mean (SD)	GQS[b], mean (SD)
Questions			
Symptoms	79.37 (18.03)	13.67 (0.58)	4.67 (0.58)
Risk factors	85.51 (2.10)	12.67 (0.58)	4.33 (0.58)
Survival rates	66.67 (8.25)	14.00 (0.00)	5.00 (0.00)
Diagnosis	84.92 (14.35)	13.67 (0.58)	4.33 (0.58)
Screening	74.60 (9.91)	15.67 (1.15)	4.67 (0.58)
Treatment	79.36 (10.99)	14.00 (0.00)	4.33 (0.58)
Sexual function	84.92 (1.37)	13.67 (0.58)	4.33 (0.58)
Bladder function	80.16 (7.65)	13.67 (0.58)	4.00 (0.00)
Total	79.44 (10.44)	13.88 (0.93)	4.46 (0.50)

[a]PEMAT-AI: Patient Education Materials Assessment Tool for Artificial Intelligence.
[b]GQS: Global Quality Score.

NLAT-AI Assessment

Expert assessment of the AI outputs with NLAT-AI was consistent with a mean >3.0 out of 5.0 (neutral) in all domains across all question replies. NLAT-AI pooled means included accuracy of 3.96 (SD 0.91), safety of 4.32 (SD 0.86), appropriateness of 4.45 (SD 0.81), actionability of 4.05 (SD 1.15), and effectiveness of 4.09 (SD 0.98). Descriptive statistics for each question are tabulated and graphed (Table 2 and Figure 4). Internal validity testing demonstrated high reliability with Cronbach α=0.906.

Qualitative feedback via NLAT-AI on questions 1 through 8 indicates some areas for improvement despite the generally accurate and easy-to-understand nature of responses. Common themes were a need for greater specificity, updated and comprehensive information, and a more globally inclusive perspective (Textbox 1). Outputs were often characterized as good starting points or overviews which could benefit patients.

Table 2. Natural Language Assessment Tool for Artificial Intelligence assessment.

Assessment	Accuracy, mean (SD)	Safety, mean (SD)	Appropriateness, mean (SD)	Actionability, mean (SD)	Effectiveness, mean (SD)
Questions					
Symptoms	3.71 (0.76)	3.86 (1.07)	4.43 (0.53)	4.00 (1.00)	3.85 (1.07)
Risk factors	4.29 (0.76)	4.57 (0.53)	4.43 (0.98)	3.71 (0.95)	4.29 (0.76)
Survival rates	3.71 (1.60)	4.14 (1.60)	4.14 (1.07)	3.86 (1.68)	3.86 (1.68)
Diagnosis	3.86 (1.07)	4.43 (0.79)	4.43 (0.79)	4.43 (1.51)	4.00 (1.00)
Screening	4.00 (0.58)	4.14 (0.58)	4.43 (0.79)	4.14 (0.69)	4.14 (0.90)
Treatment	4.29 (0.49)	4.43 (0.79)	4.71 (0.49)	3.86 (1.46)	4.00 (0.82)
Sexual function	4.00 (0.82)	4.43 (1.13)	4.71 (0.49)	4.29 (0.79)	4.43 (0.79)
Bladder function	3.86 (1.07)	4.57 (1.07)	4.26 (0.49)	4.14 (0.90)	4.14 (0.90)
Total	3.96 (0.91)	4.32 (0.86)	4.45 (0.81)	4.05 (1.15)	4.09 (0.98)

XSL•FO
RenderX

Figure 4. NLAT-AI mean score by ChatGPT question output. NLAT-AI: Natural Language Assessment Tool for Artificial Intelligence.

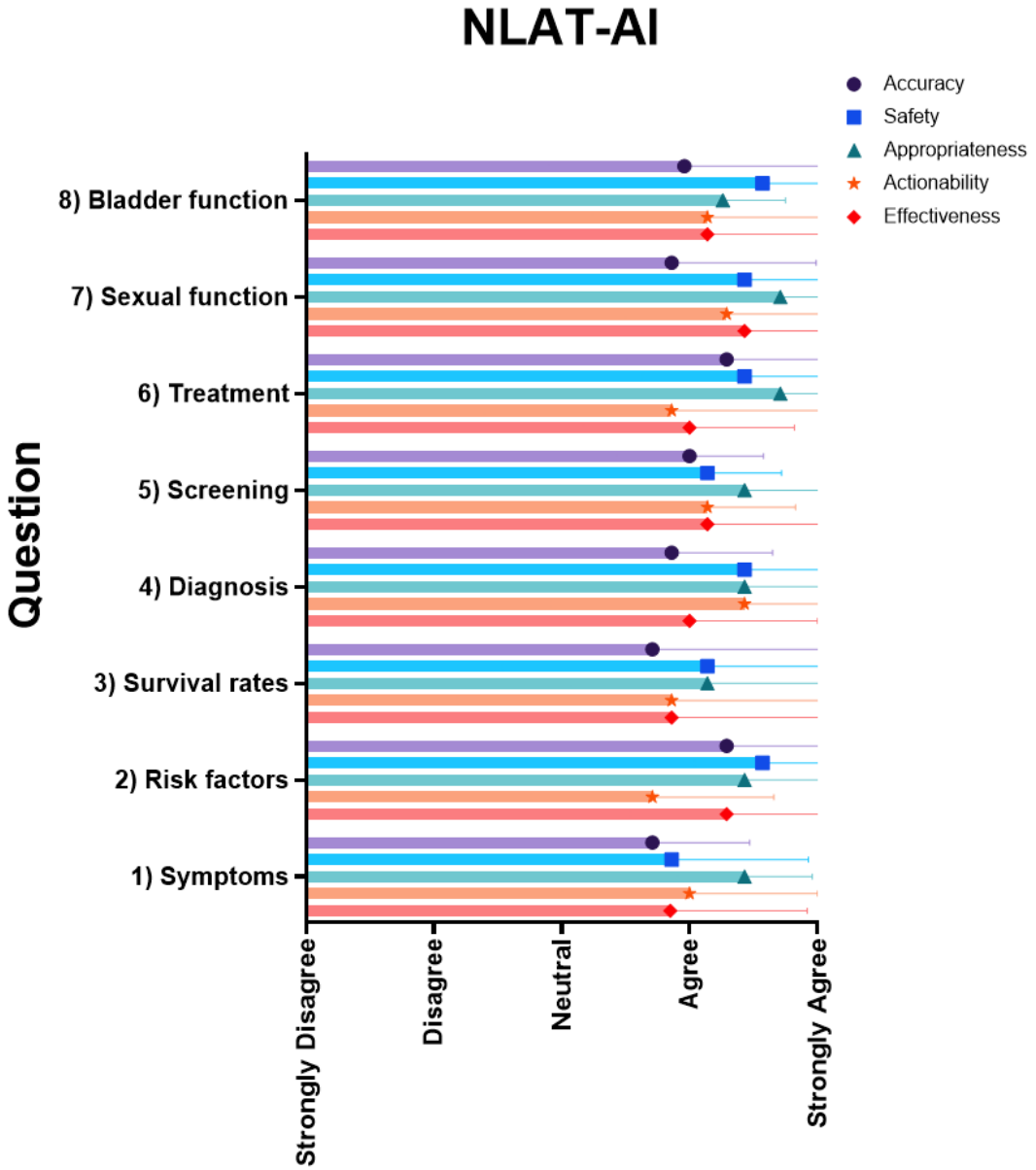

XSL•FO
RenderX

Textbox 1. Natural Language Assessment Tool for Artificial Intelligence qualitative feedback.

Question 1: Symptoms

- Overall, a reasonable answer to the question... more emphasis should be put on the fact that prostate cancer is usually asymptomatic, usually detected on screening, only symptomatic when advanced.
- Could have been better if discussed symptoms of locally advanced prostate cancer (LUTS, haematuria, etc) and symptoms of metastatic prostate cancer (bone pain, weight loss, etc)
- Need to strongly emphasize that most prostate cancers are asymptomatic so prostate-specific antigen testing is necessary

Question 2: Risk factors

- Considering that this is tailored for Americans, it may not be actionable for others.
- From a safety perspective, I would emphasize the importance of seeking medical review in the event of family history.
- Remove the modifiable risk factors as it makes patients think they can prevent it

Question 3: Survival rates

- There is...no mention of the impact of treatment on survival so a patient could be forgiven for thinking this was survival rates in the event of no treatment being given.
- “Relative survival” is not clearly explained.
- The survival rate [is] overestimated in organ-confined disease as this is far more complex. It should be more clarified.
- When talking about prostate cancer survival 10 years is the minimum that should be discussed
- Fairly good- this is what I would tell my patients.

Question 4: Diagnosis

- Overall reasonable answer from ChatGPT
- CT and bone scans are used for staging; but now in Australia is superseded by PSMA
- Reasonable answer. Some inaccuracies in how the tests are used, as well as their sequencing. PSMA PET not mentioned which is an important part of diagnosis and staging. These deficiencies likely reflect the rapidly evolving nature of prostate cancer diagnosis.
- The answer is easy to understand and general principles of diagnosis sound.

Question 5: Screening

- Point 2 is very contentious and...gives a very one-sided view of prostate cancer screening.
- This is only appropriate for American audience.
- Point 2 is concerning as this represents one [clinician] group who is very much against prostate cancer screening... therefore may risk not giving a balanced view.
- No mention of any local guidelines, and no EAU [European Association of Urology] guidelines.

Question 6: Treatment

- Very useful summary for patients immediately after diagnosis.
- No mention of novel tx [treatments] eg: focal therapy, cryo, HIFU
- No mention of robotic surgery versus open surgery
- This is a very simple table about the pros and cons.

Question 7: Sexual function

- Overall a very good answer—misses minor points
- Very well written
- Would also mention that erectile function improves over time.
- Surgery does not damage the vessels for erection

Question 8: Bladder function

- Nice summary
- Accurate and easy to understand

XSL•FO
RenderX

- Minor issues only with the discussion on stress or urge incontinence
- Hormone therapy should not causes bladder dysfunction. In fact, it might improve it

Readability Assessment

The readability algorithm consensus was "difficult to read" (Flesch Reading Ease score mean 45.97, SD 8.69; Gunning Fog Index mean 14.55, SD 4.79), averaging an 11th-grade reading level, equivalent to 15- to 17-year-olds (Flesch-Kincaid Grade Level mean 12.12, SD 4.34; The Coleman-Liau Index mean 12.75, SD 1.98; SMOG Index mean 11.06, SD 3.20). Questions 1 and 2 were the easiest to read scoring an 8th-grade level, while questions 6 (grade 23 level), 7 (grade 12 level), and 8 (grade 13 level) were very difficult to read (Table 3).

Table 3. Readability assessment.

Assessment	Flesch Reading Ease score	Gunning Fog Index	Flesch-Kincaid Grade Level	The Coleman-Liau Index	SMOG[a] Index
Questions					
Symptoms	53.2	10.1	8.7	11	7.8
Risk factors	59.4	8.5	7.7	11	7.6
Survival rates	51.4	15	11	10	11.1
Diagnosis	46.2	13.4	10.9	13	10.3
Screening	49.3	13.7	11.1	12	10.6
Treatment	57.2	25.7	22.8	16	18.7
Sexual function	39.9	14.9	11.6	14	11
Bladder function	31.2	15.1	13.2	15	11.4
Pooled total	45.97	14.55	12.12	12.75	11.06

[a]SMOG: Simple Measure of Gobbledygook.

REF-AI Assessment

REF-AI identified 2 reference hallucinations from 30 total references across all questions (pooled REF-AI Real mean 2.86). Most references effectively supported the text, while 4 questions had 1 or 2 citations that did not directly support the information provided (Table 4; pooled REF-AI supporting mean 2.75). A total of 86% (26/30) of references were from reputable government organizations, while 2 were direct citations from scientific literature (pooled REF-AI source mean 2.13). Individual statements were provided a direct reference in only 3 outputs. The remaining outputs instead provided a list of references at the bottom of the text. Some direct links to references were not complete, instead delivering the user to the organization's primary website URL, likely reflecting updated website directories since the 2021 ChatGPT indexation. The 2 hallucinated references were present in questions 7 and 8, where weblinks did not connect and despite extensive Google and library searches, the original material was unable to be located.

Table 4. Reference Assessment Artificial Intelligence assessment.

Assessment	Real, mean (SD)	Supporting, mean (SD)	Source, mean (SD)
Questions			
Symptoms	3.00 (0.00)	3.00 (0.00)	2.00 (0.00)
Risk factors	3.00 (0.00)	2.67 (0.58)	2.00 (0.00)
Survival rates	3.00 (0.00)	2.67 (0.58)	2.00 (0.00)
Diagnosis	3.00 (0.00)	3.00 (0.00)	2.00 (0.00)
Screening	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
Treatment	3.00 (0.00)	3.00 (0.00)	2.00 (0.00)
Sexual function	3.00 (0.00)	2.33 (0.58)	2.00 (0.00)
Bladder function	2.00 (0.00)	2.33 (0.58)	2.00 (0.00)
Total	2.86 (0.00)	2.75 (0.29)	2.13 (0.00)

XSL•FO
RenderX

Discussion

Principal Findings

In the digital information age, understanding what patient health information is accessed and the quality of this information is crucial. This study demonstrates several examples of information that patients (and their caregivers) may encounter when conducting searches related to prostate cancer management. In our analysis, ChatGPT-4 provided generally comprehensive answers to prostate cancer questions, mostly in line with current medical guidelines and literature. ChatGPT-4 demonstrated promise when assessed with a range of patient education and information quality assessment tools, as well as expert review. Robust scores and expert feedback indicate that the generated content was reliable, safe, and actionable for patients, albeit with room for improvement in minor nuanced details, global applicability, and readability.

Current evidence indicates that 75% of people turn to the internet for decision-making during a health crisis [39]. Despite the abundance of available patient information, studies assessing the quality of digital health information indicate significant shortcomings [40]. For prostate cancer, the quality of information that reaches the patient is known to be inconsistent [11,20,41-43]. For example, a previous assessment of the top 100 “prostate cancer” web page results identified via search engine query showed that only 11.1% of sites demonstrate an excellent on the original DISCERN criteria [11]. While our analysis has used necessarily disparate methods, a comparison of our DISCERN-AI results (good-excellent) to static web page DISCERN scores suggests that ChatGPT prostate cancer information outputs may be of a higher quality than many traditional web pages [11]. ChatGPT4 appears capable of providing broad and largely accurate information which may further augment self-directed patient or stakeholder enquiry. Nevertheless, a direct comparison of ChatGPT outputs to established gold standard information sources is necessary to clearly define the role of this new communication technology as part of patient care and education.

Our findings appear to differ from Coskun et al [16], where ChatGPT-3 had accuracy issues using queries generated by the European Association of Urology Patient Information. Interestingly, Zheng et al [17] discovered that ChatGPT-4 can offer suitable counseling on disease prevention and screening for prostate cancer patients. These differences may represent the rapid evolution of the algorithm as our testing used the newer model. Exclusive use of US-centric guidelines raised questions of bias among our experts. Others have also highlighted such bias, noting that 51% of training data for major large language models is US sourced [44,45]. The disparities between ChatGPT-3 and ChatGPT-4 highlight the continual advancement and refinement of the underlying technology, reinforcing the need for periodic assessment and validation as newer models emerge [5,17]. Conversely, a lack of validated and reproducible tools to make reliable quality assessments of NPLTs is likely to play a role in varied results within this juvenile domain of clinical research [46]. While the methods used in our study were an effort to standardize output assessment in our work, we recognize and encourage further rigorous work to develop validated and reproducible assessment tools that can be applied to a range of NPLT outputs and platforms.

Despite the NLAT-AI rating, and general appropriateness of the language across all questions, the objective readability from algorithms demonstrated a high reading level and difficulty. This is likely reflective of the literacy bias present among our highly educated expert pool [34,47,48]. While the recommended reading level for patient education material varies between organizations, the consensus is that it should generally lie between grade 6 and 8 reading levels [47,49]. The readability algorithms thus suggest that the generated content may be challenging for some readers. These findings are of importance given that lower readability may limit accessibility for certain socioeconomic or minority groups [47]. Literacy is a known negative correlate of prostate cancer health outcomes [9,50-52]. Compounding this concern is the effect of the user’s overarching eHealth literacy, which is likely to affect chatbot engagement behaviors and patterns of information comprehension and use [3,16,50,53]. Effects of both traditional literacy and eHealth literacy on the end user experience of NLPTs require urgent investigation due to the pervasiveness that these technologies are already presenting within society and in web-based health communication [54,55].

The digital nature of the ChatGPT-4 model, where users can continuously engage and seek clarifications, offers a potential advantage and solution to static patient information materials. Although beyond this study’s scope, the ChatGPT-4 model permits ongoing discussions, enabling patients to seek clarifications of information. These conversations allow for personalized explanations related to patient health results, the opportunity to simplify language, and may ultimately address some concerns raised by our expert assessors. This is an extremely powerful and unique component of this new digital technology. Future iterations of such models may benefit by incorporating clear adaptability features, where the complexity and specificity of the content can be adjusted based on user preferences or needs. Further studies are required to explore how the longitudinal and dynamic features of NLPTs affect information quality and patient comprehension. This will be particularly important in comparison to traditional website and social media-based information sources which currently dominate the landscape of self-educative information sourcing in prostate cancer care [11,20,56]. NPLTs with predetermined or flexible user settings attuned to patient preference, needs, or literacy level are a potential futurist pathway to cost-effective and scalable forms of tailored patient health education materials.

Hallucination, where information is fabricated by the NLPT and presented as valid, is a well-documented phenomenon specific to NLPTs and ChatGPT [38]. This study demonstrated that hallucinations could occur when searching for prostate cancer with NPLT or chatbots. While only occurring in 2 instances of 30, these findings continue. Designation between hallucination and faux hallucination should also be considered. Faux-hallucination results from modified references after ChatGPT-4’s indexation, leading to broken links or lost references. Website redesign or content that no longer exists after the 2021 indexation is a potential etiology for hallucination

XSL•FO
RenderX

that has not been fully explored. Equally, such disappearance of content with time may also match the definition of hallucination in the future. While not a prominent issue in this study, these findings continue to demonstrate the potential for fabricated information, which can be easily overlooked by the unassuming clinician, patient, or researcher. While still in its infancy, large language models must continue to solve the issue of hallucination before integration into high-risk systems, such as health care, can be considered.

While hallucinations are a notable concern, there are several other limitations of current NPLTs that need to be considered. Despite malleability, it is unknown whether the ChatGPT-4 model may fully replicate the nuance of human communication necessary for effective patient health education [3]. Additionally, the most significant limitation of ChatGPT is its potential for biased, outdated, or misleading content generation [1,4,6,53]. Even with relatively high-quality scores, this study shows that ChatGPT can still produce misleading or biased content under discriminatory and expert scrutiny, posing some element of risk for those with poor eHealth literacy [4,6,53]. However, while expert reviewers identified minor inaccuracies, none of these points were considered to be significantly concerning safety issues. Nevertheless, there is currently a lack of evidence to predict the impact of these technologies on patients' understanding, decision-making, or health, without further inquiry and consideration of patients' ability to interact with these new eHealth technologies. We strongly recommend clinicians report these concerns to prostate cancer patients and their stakeholders when guiding patient use of web-based information in their care. Furthermore, the opaque and dynamic nature of this technology's private enterprise proprietary algorithms is also a concern [3,4,46]. Algorithm development will likely outpace quality assurance efforts and raise questions about the necessity of clinician involvement in NPLT model development that aims to present health-based information [3,45]. The effectively unknown and vast array of sources from which ChatGPT's training data are derived raises ethical concerns. Without knowing the origins and credibility of such data, it is difficult for clinicians to fully trust generated content, presenting us with a modernized but perpetual issue of distrust in web-based information which may ultimately hinder adoption and progress [3,6,53]. Finally, there are also financial considerations; the cost of using ChatGPT-4 (as opposed to the currently free ChatGPT-3.5) or other NPLTs may form a barrier to widespread adoption in health care settings and has the potential to drive disparate levels of health care if not effectively managed and regulated.

Limitations

Limitations of this study include the sample size of assessors, which may skew the evaluation of the included tool's reliability and efficacy. The qualitative assessments of experts are at inherent risk of bias for or against the use of novel technology and ChatGPT-4. However, these experts are also deeply aware of the nature and quality of current prostate cancer education materials, providing additional insight that is of value to this work.

It is important to note this assessment was purposefully narrow in scope and may not reflect the myriad of interactions under the vast topics of prostate cancer. It is unknown how applicable these interactions are in wider prostate cancer education scenarios and ongoing investigation is required. Work is currently underway to assess an expanded question set with a comparison to currently accepted patient education gold standards in prostate cancer.

While not an explicit purpose of this study, the exploratory assessments used in this work (DISCERN-AI, PEMAT-AI, NLAT-AI, and REF-AI) demonstrate interreliability and replicability across several cancer-type information outputs. They may thus have potential use for clinicians and researchers interested in reviewing the quality of other cancer-based outputs of ChatGPT-4 or other NPLTs. Nevertheless, their validity requires further testing and greater investigation is necessary to develop specific tools to assess NPLT output quality in the long term.

Conclusion

Our analysis found ChatGPT-4's responses to common prostate cancer queries were of good quality, and a potentially useful patient education adjunct for prostate cancer care. Objective quality assessment tools were reflective of NPLT outputs, which were generally reliable and appropriate, although with room for improvement. Our expert panel was impressed by the appropriateness and safety of the language and information given. However, clinicians should be aware that there are several limitations to ChatGPT-4 prostate cancer outputs including hallucination, specificity issues, and difficult readability. Future studies are required to assess whether more longitudinal (back-and-forth) ChatGPT-4 discourse may offset some of the concerns highlighted in this analysis, and how patients of differing eHealth literacy levels may engage with and have care affected by such technologies.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Common prostate cancer questions.
[DOCX File , 13 KB - jmir_v26i1e55939_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

ChatGPT-4 output.
[DOCX File , 22 KB - jmir_v26i1e55939_app2.docx]

Multimedia Appendix 3
Patient Education Materials Assessment Tool for AI (PEMAT-AI) tool.
[DOCX File , 13 KB - jmir_v26i1e55939_app3.docx]

Multimedia Appendix 4
DISCERN-AI (artificial intelligence) tool.
[DOCX File , 16 KB - jmir_v26i1e55939_app4.docx]

Multimedia Appendix 5
Natural Language Assessment Tool for Artificial Intelligence (NLAT-AI).
[DOCX File , 49 KB - jmir_v26i1e55939_app5.docx]

Multimedia Appendix 6
Reference Assessment Artificial Intelligence (REF-AI) tool.
[DOCX File , 25 KB - jmir_v26i1e55939_app6.docx]

References

1. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023;7(2):pkad015. [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
2. ChatGPT. OpenAI. URL: https://openai.com/chatgpt/ [accessed 2024-07-05]
3. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
4. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
5. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg 2023;47(5):1985-1993 [FREE Full text] [doi: 10.1007/s00266-023-03338-7] [Medline: 37095384]
6. Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith DJ, Rozen WM, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. Aesthet Surg J 2023;43(10):1126-1135. [doi: 10.1093/asj/sjad140] [Medline: 37158147]
7. Wang L, Lu B, He M, Wang Y, Wang Z, Du L. Prostate cancer incidence and mortality: global status and temporal trends in 89 countries from 2000 to 2019. Front Public Health 2022;10:811044 [FREE Full text] [doi: 10.3389/fpubh.2022.811044] [Medline: 35252092]
8. Catt S, Matthews L, May S, Payne H, Mason M, Jenkins V. Patients' and partners' views of care and treatment provided for metastatic castrate-resistant prostate cancer in the UK. Eur J Cancer Care (Engl) 2019;28(6):e13140. [doi: 10.1111/ecc.13140] [Medline: 31475410]
9. Ellimoottil C, Polcari A, Kadlec A, Gupta G. Readability of websites containing information about prostate cancer treatment options. J Urol 2012;188(6):2171-2175. [doi: 10.1016/j.juro.2012.07.105] [Medline: 23083852]
10. Baunacke M, Schmidt ML, Groeben C, Borkowetz A, Thomas C, Koch R, et al. Decision regret after radical prostatectomy does not depend on surgical approach: 6-year followup of a large German cohort undergoing routine care. J Urol 2020;203(3):554-561. [doi: 10.1097/JU.0000000000000541] [Medline: 31518200]
11. Moolla Y, Adam A, Perera M, Lawrentschuk N. 'Prostate cancer' information on the internet: fact or fiction? Curr Urol 2020;13(4):200-208 [FREE Full text] [doi: 10.1159/000499271] [Medline: 31998052]
12. Alsyouf M, Stokes P, Hur D, Amasyali A, Ruckle H, Hu B. 'Fake news' in urology: evaluating the accuracy of articles shared on social media in genitourinary malignancies. BJU Int 2019;124(4):701-706. [doi: 10.1111/bju.14787] [Medline: 31044493]
13. Sehn E, Mozak C, Yuksel N, Sadowski CA. An analysis of online content related to testosterone supplementation. Aging Male 2019;22(2):141-149. [doi: 10.1080/13685538.2018.1482867] [Medline: 29921150]
14. Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, et al. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med 2024;39(4):573-577. [doi: 10.1007/s11606-023-08469-w] [Medline: 37940756]
15. Jiao W, Wang W, Huang JT, Wang X, Tu Z. Is ChatGPT a good translator? a preliminary study. arXiv. Preprint posted online on Jan 20, 2022 . [doi: 10.48550/arXiv.2301.08745]

XSL•FO
RenderX

16. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology 2023;180:35-58. [doi: 10.1016/j.urology.2023.05.040] [Medline: 37406864]
17. Zheng Y, Xu Z, Yu B, Xu T, Huang X, Zou Q, et al. Appropriateness of prostate cancer prevention and screening recommendations obtained from ChatGPT-4. Research Square. Preprint posted online on May 22, 2023 . [doi: 10.21203/rs.3.rs-2898778/v1]
18. Haun MW, Ihrig A, Karschuck P, Thomas C, Huber J. The era of the digital natives is approaching: insights into online peer-to-peer support for persons affected by prostate cancer. World J Urol 2020;38(10):2433-2434 [FREE Full text] [doi: 10.1007/s00345-020-03114-1] [Medline: 32034498]
19. van Eenbergen MCHJ, Vromans RD, Boll D, Kil PJM, Vos CM, Krahmer EJ, et al. Changes in internet use and wishes of cancer survivors: a comparison between 2005 and 2017. Cancer 2020;126(2):408-415 [FREE Full text] [doi: 10.1002/cncr.32524] [Medline: 31580497]
20. Cacciamani GE, Bassi S, Sebben M, Marcer A, Russo GI, Cocci A, et al. Consulting "Dr. Google" for prostate cancer treatment options: a contemporary worldwide trend analysis. Eur Urol Oncol 2020;3(4):481-488 [FREE Full text] [doi: 10.1016/j.euo.2019.07.002] [Medline: 31375427]
21. Rezaee ME, Goddard B, Sverrisson EF, Seigne JD, Dagrosa LM. 'Dr Google': trends in online interest in prostate cancer screening, diagnosis and treatment. BJU Int 2019;124(4):629-634. [doi: 10.1111/bju.14846] [Medline: 31206954]
22. Google trends explore. Trends G. URL: https://trends.google.com/trends/explore?geo=AU&hl=en-AU [accessed 2024-06-11]
23. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
24. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
25. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis--a wakeup call? J Rheumatol 2012;39(5):899-903. [doi: 10.3899/jrheum.111114] [Medline: 22467934]
26. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the world wide web. Am J Gastroenterol 2007;102(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
27. Cassidy JT, Baker JF. Orthopaedic patient information on the world wide web: an essential review. J Bone Joint Surg Am 2016;98(4):325-328. [doi: 10.2106/JBJS.N.01189] [Medline: 26888683]
28. Weil AG, Bojanowski MW, Jamart J, Gustin T, Lévêque M. Evaluation of the quality of information on the internet available to patients undergoing cervical spine surgery. World Neurosurg 2014;82(1-2):e31-e39. [doi: 10.1016/j.wneu.2012.11.003] [Medline: 23142585]
29. Altunisik E, Firat YE. Quality and reliability analysis of essential tremor disease information on social media: the study of YouTube. Tremor Other Hyperkinet Mov (N Y) 2022;12:32 [FREE Full text] [doi: 10.5334/tohm.727] [Medline: 36415589]
30. Kincaid JP, Fishburne JRP, Rogers R, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. In: Institute for Simulation and Training. Florida: University of Central Florida STARS; 1975.
31. McLaughlin GH. SMOG grading-a new readability formula. J Reading 1969;12(8):639-646.
32. Gunning R. The Technique of Clear Writing. New York: McGraw-Hill; 1952.
33. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol 1975;60(2):283. [doi: 10.1037/h0076540]
34. Hansberry DR, John A, John E, Agarwal N, Gonzales SF, Baker SR. A critical review of the readability of online patient education resources from RadiologyInfo.Org. AJR Am J Roentgenol 2014;202(3):566-575. [doi: 10.2214/AJR.13.11223] [Medline: 24555593]
35. Readable.com. URL: https://readable.com/ [accessed 2024-06-11]
36. Microscoft. 2023. URL: https://www.microsoft.com/en-in/microsoft-365/ [accessed 2024-06-11]
37. Kugar MA, Cohen AC, Wooden W, Tholpady SS, Chu MW. The readability of psychosocial wellness patient resources: improving surgical outcomes. J Surg Res 2017;218:43-48. [doi: 10.1016/j.jss.2017.05.033] [Medline: 28985876]
38. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
39. Dickerson S, Reinhart A, Boemhke M, Akhu-Zaheya L. Cancer as a problem to be solved: internet use and provider communication by men with cancer. Comput Inform Nurs 2011;29(7):388-395. [doi: 10.1097/NCN.0b013e3181f9ddb1] [Medline: 20975535]
40. Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, et al. Health information on the internet: accessibility, quality, and readability in english and Spanish. JAMA 2001;285(20):2612-2621 [FREE Full text] [doi: 10.1001/jama.285.20.2612] [Medline: 11368735]

XSL•FO
RenderX

41. Lange L, Peikert ML, Bleich C, Schulz H. The extent to which cancer patients trust in cancer-related online information: a systematic review. PeerJ 2019;7:e7634 [FREE Full text] [doi: 10.7717/peerj.7634] [Medline: 31592341]
42. Ghai S, Trachtenberg J. Internet information on focal prostate cancer therapy: help or hindrance? Nat Rev Urol 2019;16(6):337-338. [doi: 10.1038/s41585-019-0180-8] [Medline: 30952966]
43. Asafu-Adjei D, Mikkilineni N, Sebesta E, Hyams E. Misinformation on the internet regarding ablative therapies for prostate cancer. Urology 2019;133:182-186. [doi: 10.1016/j.urology.2018.12.050] [Medline: 30817959]
44. Dodge J, Sap M, Marasović A, Agnew W, Ilharco G, Groeneveld D. Documenting large webtext corpora: a case study on the colossal clean crawled corpus. arXiv. Preprint posted online on April 18, 2021 . [doi: 10.48550/arXiv.2104.08758]
45. Healy M. Approaches to generative artificial intelligence, a social justice perspective. arXiv. Preprint posted online on August 17, 2023 . [doi: 10.48550/arXiv.2309.12331]
46. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
47. Mac OA, Muscat DM, Ayre J, Patel P, McCaffery KJ. The readability of official public health information on COVID-19. Med J Aust 2021;215(8):373-375 [FREE Full text] [doi: 10.5694/mja2.51282] [Medline: 34580878]
48. Rosenberg SA, Francis D, Hullett CR, Morris ZS, Fisher MM, Brower JV, et al. Readability of online patient educational resources found on NCI-designated cancer center web sites. J Natl Compr Canc Netw 2016;14(6):735-740 [FREE Full text] [doi: 10.6004/jnccn.2016.0075] [Medline: 27283166]
49. Health SA. Engaging with Consumers, Carers and the Community: Guide and Resources. South Australia: SA Health Adelaide; 2021.
50. Basch CH, Ethan D, MacLean SA, Fera J, Garcia P, Basch CE. Readability of prostate cancer information online: a cross-sectional study. Am J Mens Health 2018;12(5):1665-1669 [FREE Full text] [doi: 10.1177/1557988318780864] [Medline: 29888641]
51. Maciolek KA, Jarrard DF, Abel EJ, Best SL. Systematic assessment reveals lack of understandability for prostate biopsy online patient education materials. Urology 2017;109:101-106. [doi: 10.1016/j.urology.2017.07.037] [Medline: 28780302]
52. Borgmann H, Wölm JH, Vallo S, Mager R, Huber J, Breyer J, et al. Prostate cancer on the web-expedient tool for patients' decision-making? J Cancer Educ 2017;32(1):135-140. [doi: 10.1007/s13187-015-0891-3] [Medline: 26234650]
53. Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis 2024;27(1):159-160. [doi: 10.1038/s41391-023-00754-3] [Medline: 37923807]
54. Zhang Z, Genc Y, Wang D, Ahsen ME, Fan X. Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. J Med Syst 2021;45(6):64. [doi: 10.1007/s10916-021-01743-6] [Medline: 33948743]
55. Zhang Z, Genc Y, Xing A, Wang D, Fan X, Citardi D. Lay individuals' perceptions of artificial intelligence (AI)-empowered healthcare systems. Proc Assoc Inf Sci Technol 2020;57(1):e326. [doi: 10.1002/pra2.326]
56. Qan'ir Y, Song L. Systematic review of technology-based interventions to improve anxiety, depression, and health-related quality of life among patients with prostate cancer. Psychooncology 2019;28(8):1601-1613 [FREE Full text] [doi: 10.1002/pon.5158] [Medline: 31222956]

Abbreviations

AI: artificial intelligence
GQS: Global Quality Score
NLAT: Natural Language Assessment Tool
NLPT: natural language processing technology
PEMAT: Patient Education Materials Assessment Tool
REF-AI: Reference Assessment Artificial Intelligence
SMOG: Simple Measure of Gobbledygook

Edited by Q Jin; submitted 05.01.24; peer-reviewed by M Vine, E Tagai, F Chen; comments to author 25.04.24; revised version received 12.05.24; accepted 12.05.24; published 14.08.24.

Please cite as:
Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N, Kam J, Mehan N, Thanigasalam R, Jeffery N, Patel MI, Leslie S
Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment
J Med Internet Res 2024;26:e55939
URL: https://www.jmir.org/2024/1/e55939
doi:10.2196/55939
PMID:

XSL•FO
RenderX

©Damien Gibson, Stuart Jackson, Ramesh Shanmugasundaram, Ishith Seth, Adrian Siu, Nariman Ahmadi, Jonathan Kam, Nicholas Mehan, Ruban Thanigasalam, Nicola Jeffery, Manish I Patel, Scott Leslie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ChatGPT and Google Assistant as a Source of Patient Education for Patients With Amblyopia: Content Analysis

Gloria Wu[1], MD; David A Lee[2], MS, MD, MBA; Weichen Zhao[3], BS; Adrial Wong[3], BS; Rohan Jhangiani[4], BS, MS; Sri Kurniawan[4], PhD

[1]University of California, San Francisco School of Medicine, San Francisco, CA, United States
[2]McGovern Medical School, University of Texas Health Science Center at Houston, Houston, CA, United States
[3]College of Biological Sciences, University of California, Davis, Davis, CA, United States
[4]Department of Computational Media, University of California, Santa Cruz, Santa Cruz, CA, United States

Corresponding Author:
Gloria Wu, MD
University of California, San Francisco School of Medicine
533 Parnassus Ave
San Francisco, CA, 94143
United States
Phone: 1 408 621 9074
Fax: 1 408 356 5556
Email: gwu2550@gmail.com

Abstract

Background: We queried ChatGPT (OpenAI) and Google Assistant about amblyopia and compared their answers with the keywords found on the American Association for Pediatric Ophthalmology and Strabismus (AAPOS) website, specifically the section on amblyopia. Out of the 26 keywords chosen from the website, ChatGPT included 11 (42%) in its responses, while Google included 8 (31%).

Objective: Our study investigated the adherence of ChatGPT-3.5 and Google Assistant to the guidelines of the AAPOS for patient education on amblyopia.

Methods: ChatGPT-3.5 was used. The four questions taken from the AAPOS website, specifically its glossary section for amblyopia, are as follows: (1) What is amblyopia? (2) What causes amblyopia? (3) How is amblyopia treated? (4) What happens if amblyopia is untreated? Approved and selected by ophthalmologists (GW and DL), the keywords from AAPOS were words or phrases that deemed significant for the education of patients with amblyopia. The “Flesch-Kincaid Grade Level” formula, approved by the US Department of Education, was used to evaluate the reading comprehension level for the responses from ChatGPT, Google Assistant, and AAPOS.

Results: In their responses, ChatGPT did not mention the term “ophthalmologist,” whereas Google Assistant and AAPOS both mentioned the term once and twice, respectively. ChatGPT did, however, use the term “eye doctors” once. According to the Flesch-Kincaid test, the average reading level of AAPOS was 11.4 (SD 2.1; the lowest level) while that of Google was 13.1 (SD 4.8; the highest required reading level), also showing the greatest variation in grade level in its responses. ChatGPT’s answers, on average, scored 12.4 (SD 1.1) grade level. They were all similar in terms of difficulty level in reading. For the keywords, out of the 4 responses, ChatGPT used 42% (11/26) of the keywords, whereas Google Assistant used 31% (8/26).

Conclusions: ChatGPT trains on texts and phrases and generates new sentences, while Google Assistant automatically copies website links. As ophthalmologists, we should consider including “see an ophthalmologist” on our websites and journals. While ChatGPT is here to stay, we, as physicians, need to monitor its answers.

(J Med Internet Res 2024;26:e52401) doi:10.2196/52401

KEYWORDS

ChatGPT; Google Assistant; amblyopia; health literacy; American Association for Pediatric Ophthalmology and Strabismus; pediatric; ophthalmology; patient education; education; ophthalmologist; Google; monitoring

XSL•FO
RenderX

Introduction

Amblyopia is a global public health problem with multiple impacts on vision and quality of life [1]. Amblyopia, if untreated, leads to permanent visual impairment in adults. The underlying causes are undiscovered anisometropia (unequal refractive error between the 2 eyes), visual deprivation such as congenital or juvenile cataracts, or strabismus with anisometropia. Treatment can be as simple as a pair of glasses. Treatment may mean occlusion therapy of the stronger eye, ensuring that the "weaker" or amblyopic eye gets visual stimulation. If treated early, the young child will have improved binocular vision.

Amblyopia is a common visual complaint in children in an ophthalmology setting. Vision at the level of 20/80 to 20/200 makes up almost 40% of the cases in a population-based study in Australia [2]. The critical period of human visual development is in the first decade of life [3]. Recent neuroplasticity models in humans and primates suggest that there is residual plasticity that extends into later years of adult life [3].

The quality of life for people with amblyopia can involve difficulties in daily activities due to impaired reading speed, stereo acuity, motor skills, motion perception, and fixation stability [4,5].

Health education is needed in the field of amblyopia. Patients can seek help in a timely manner only if there is public awareness. Often, the patient is young and not able to articulate his visual complaints. The parents may also be young and not aware of vision research and brain plasticity.

ChatGPT-3.5 (OpenAI) is a free and advanced large language model (LLM) launched on November 30, 2022. By December 4, 2022, it had over 1 million users, and by the 2-month mark, the monthly active users had been reported to exceed 100 million [6]. The demographics of ChatGPT users are those of the internet user population. The average age ranges between 18 and 29 years, matching the ages of young parents with amblyopic children. [7]. ChatGPT-3.5 has natural language processing capabilities and uses artificial intelligence (AI) to train in the auto-completion of sentences and ideas.

Health literacy may be a problem for ChatGPT-3.5 and patient users [8-10]. ChatGPT-3.5 could aid in patient education, but it types out texts in advanced high-school to college-level English. An estimated 130 million Americans lack proficiency in literacy, essentially reading below the equivalent of a sixth-grade level [9]. The average American reads between seventh and eighth grade [10,11]. Health literacy is a recognized problem for most Americans [11-13].

Can ChatGPT and Google Assistant be a source of patient education for our patients with amblyopia? Is it accurate and is it understandable [14,15]? Our study investigates whether an LLM and a voice assistant can adhere to the American Association for Pediatric Ophthalmology and Strabismus (AAPOS) guidelines for patient education. The AAPOS is an authoritative organization that provides educational information to the public and ophthalmologists about various eye diseases in children [16]. There is patient information for many eye diseases including amblyopia on the AAPOS website [16]. No institutional review board approval was required because no patient participants or patient data were involved in this study.

Methods

In AAPOS' "Learn More About Eye Conditions" section, there is a glossary of terms for "Amblyopia" [16]. From that hyperlink, we found the AAPOS' 4 questions to query ChatGPT-3.5 and Google Assistant and then subsequently recorded their responses. The embedded video was transcribed using the "closed caption" option. From this point onward, our usage of the term AAPOS will refer to the section of the AAPOS website containing information on amblyopia. From AAPOS, 2 ophthalmologists (GW and DAL) chose the keywords within the text. The text was transcribed in Google Docs and included in the keywords.

We chose the ChatGPT-3.5 version because it is the free and most widely used version. The more advanced ChatGPT-4.0, released on March 13, 2023, requires a paid subscription. The four questions from the AAPOS website on "amblyopia" are as follows: (1) What is amblyopia? (2) What causes amblyopia? (3) How is amblyopia treated? (4) What happens if amblyopia is not treated? [16]. To assess the chatbots' responses, we used the Flesch-Kincaid Grade Level test. This test has been used by the US Navy since the 1970s [17]. In addition, Flesch-Kincaid has also been used for scoring the results as shown by Lee et al [18] and the Pennsylvania Department of Insurance [19]. It has also been used in the Pennsylvania Department of Insurance to ensure readability [20].

For scoring the reading level, we used the "Flesch–Kincaid Grade Level Formula," which presents a score as a US grade level [17]. This test is used in the US Navy and the US Department of Education to allow teachers, parents, librarians, and others to judge the readability level of schoolbooks and textbooks [17]. The formula yields the "mean number of years of education" generally required to understand the book, text, or assigned reading. The resultant score (grade level) is particularly relevant when the number is greater than the tenth grade. The formula for Flesch-Kincaid is as follows [21]: Grade Level=0.39*(words/sentences)+11.8*(syllables/words)15.59. The responses of AAPOS, ChatGPT-3.5, and Google Assistant to the questions were pasted into the Flesch-Kincaid Readability Calculator to obtain the readability grade levels [22]. For the word count measurement, we used Google Docs tools. Keywords from the AAPOS website were used as a reference point to score the ChatGPT-3.5 and Google Assistant responses. A total of 26 terms were used in the 4 questions from the AAPOS website.

Results

For the 4 responses, ChatGPT used "eye doctor" once and never used "ophthalmologist". Meanwhile, Google Assistant used "ophthalmologist" once, and AAPOS used the term twice, while neither used "eye doctor" (Table 1). Giving weighted scores to the individual keywords for the 4 questions, ChatGPT scored 47% in keyword usage, while Google scored 46%.

XSL•FO
RenderX

Table 1. Comparison of keywords.

Comparison of keywords							Weighted score			
Keywords	CG[a]	GA[b]	AAPOS[c]	CG total, n/N	GA total, n/N	AAPOS total, n/N	Points	CG points, n/N	GA points, n/N	AAPOS points, n/N
Question 1				3/7	2/7	7/7	N/A[d]	12/25	8/25	25/25
Lazy eye	✓	✓	✓				4			
Brain	✓	✓	✓				4			
High refractive error			✓				3			
Strong glasses			✓				3			
Early childhood or young age			✓				3			
Vision loss	✓		✓				4			
Abnormal vision development			✓				4			
Question 2				4/11	0/11	11/11	N/A	8/25	0/25	25/25
Eye crosses			✓				3			
Eye drifts out			✓				3			
Brain avoids double vision			✓				2			
Suppressing the vision			✓				3			
Cataract	✓		✓				2			
Drooping eyelid			✓				2			
No clues			✓				2			
Pediatrician			✓				2			
Deprivation	✓		✓				2			
Strabismus	✓		✓				2			
Refractive	✓		✓				2			
Question 3				3/4	4/4	4/4	N/A	19/25	25/25	25/25
Glasses or contact lenses	✓	✓	✓				7			
Patches or patching	✓	✓	✓				7			
Surgery		✓	✓				6			
Drops	✓	✓	✓				5			
Question 4				1/4	2/4	4/4	N/A	8/25	13/25	25/25
9-10 years (cut off for effective treatment)			✓				7			
Pediatric ophthalmologist			✓				5			
Permanent vision decrease or loss	✓	✓	✓				8			
Ophthalmologist		✓	✓				5			
Total keyword appearance				11/26	8/26	26/26	Percentage of points earned	47/100	46/100	100/100

[a]CG: ChatGPT.

[b]GA: Google Assistant.

[c]AAPOS: American Association for Pediatric Ophthalmology and Strabismus.

[d]N/A: not applicable.

XSL•FO
RenderX

As seen in Table 2, AAPOS displayed the lowest reading comprehension level necessary among the 3 sources (mean 11.4, SD 2.1), whereas Google Assistant showed the highest level (mean 13.1, SD 4.8). ChatGPT displayed an average reading grade level of 12.4 (SD 1.1). In performing statistical *t* tests between the Flesch-Kincaid grades of the questions, we see that none of the responses varied in reading level significantly from one another. See footnotes under Table 2, where the *P* values are not significant. However, the word counts of the responses, as seen in Table 3, were significantly different between AAPOS and Google Assistant, as well as between ChatGPT and Google Assistant (*P*=.57); AAPOS versus Google Assistant (*P*.001); and ChatGPT versus Google Assistant (*P*.001). See footnotes under Table 3, where the *P* values are significant.

For question 4, ChatGPT's answer is misleading. It cites age 6 years as the preferable age to start treating amblyopia. Effective intervention in amblyopia can start as early as possible and even as late as age 9 or 10 years, according to AAPOS. Recovery of vision at any age is considered by ophthalmologists to be beneficial in terms of quality of life. The subtleties are missing in the ChatGPT answers (Tables S1-S3 in Multimedia Appendix 1). In addition, there are errors in omission when not mentioning the need to see an ophthalmologist, pediatric ophthalmologist, or pediatrician.

Table 2. Flesch-Kincaid Grade Level[a].

Information sources	Flesch-Kincaid Grade Level				
	Question 1	Question 2	Question 3	Question 4	Mean (SD)
ChatGPT	11	12.6	12.5	13.6	12.4 (1.1)
AAPOS[b]	9.5	10.4	14.3	11.4	11.4 (2.1)
Google Assistant	10.7	7	15.7	13.3	13.1 (4.8)

[a]*P* values of *t* test: AAPOS versus ChatGPT: *P*=.42; AAPOS versus Google Assistant: *P*=.90; ChatGPT versus Google Assistant: *P*=.71.
[b]AAPOS: American Association for Pediatric Ophthalmology and Strabismus.

Table 3. Word count[a].

Information sources	Word count				
	Question 1	Question 2	Question 3	Question 4	Mean (SD)
ChatGPT	167	188	228	164	186.8 (29.5)
AAPOS[b]	110	172	77	63	105.5 (48.5)
Google Assistant	44	56	38	46	46 (7.5)

[a]*P* values of *t* test: AAPOS versus ChatGPT: *P*=.57; AAPOS versus Google Assistant: *P*.001; ChatGPT versus Google Assistant: *P*.001.
[b]AAPOS: American Association for Pediatric Ophthalmology and Strabismus.

Discussion

During the past 2 years of the COVID-19 pandemic, many patients self-isolated or were afraid to go to their doctor's appointments for eye care. Many of their family members used the internet or voice assistants for patient information. The natural language model in ChatGPT met an unspoken need [23-26]. By 2023, there were 100 million ChatGPT users from all over the world who accessed this natural language model. Google Assistant was first launched in May 2016, and has the largest share of the voice assistant market. Google Assistant uses Google's AI but relies on search engine optimization since it lists snippets of phrases from websites in its citations [27]. While the answers are short for Google Assistant (they are succinct and to the point), they are a shorthand complement to ChatGPT's long-winded answers.

ChatGPT may play a role in patient education but when compared with AAPOS, the LLM mentions "eye doctor" once and not "ophthalmologist", whereas Google Assistant does mention "ophthalmologist." Both ChatGPT and Google Assistant mention "lazy eye" and "brain" which is excellent. The 3.5 version of ChatGPT was trained on the internet data through the year 2021, but in the future, more advanced versions of ChatGPT may use data post 2021.

We are unclear about the training data sets for these LLMs. There are 6 million people with ophthalmic diseases in America, while those with hypertension are a total of 90 million; thus, ChatGPT's ability to recognize ophthalmic terminology may be more limited. More patients are asking questions about common American afflictions such as heart disease, stroke, and diabetes than eye diseases.

One of the disadvantages of AAPOS, Google Assistant, and ChatGPT's text responses is that all of these modalities require a high reading level, greater than the seventh- to eighth-grade reading level of the average American [9]. Patients with amblyopia who are more educated will be able to understand that partial loss of vision is intertwined with brain development. The ability to process these ideas may require higher health literacy and higher education levels than most young parents.

Our patients from diverse backgrounds in the United States may not be able to understand ChatGPT, AAPOS, and Google Assistant. The "Public & Patients" section on the American Academy of Ophthalmology website has material written at the

XSL•FO
RenderX

seventh- or eighth-grade level and the answers are less detailed than those on the AAPOS website [16].

Another potential problem that might arise with the use of LLMs is that the natural language is set in a conversational tone, thus minor flaws such as "go see a doctor" may seem perfectly normal. ChatGPT and Google Assistant have information about amblyopia for the lay public. None of the answers tell patients about the urgency or need to see a pediatric ophthalmologist. This is due to the responses created by the prompt engineers who manage the chatbots' answers. In 2024, all ChatGPT responses have attached statements such as, "please see a medical professional."

In the spring of 2023, there were new competitors to ChatGPT. Bard (currently known as "Gemini") and PaLM2 are both Google-directed AI LLMs [27-29]. Bard was launched briefly in March 2023 and May 2023, with mixed results in Europe. However, in mid-August 2023, it was launched in the United States. Its algorithms are augmented with those created under PaLM2 [28,29]. As of this writing, Bard was replaced by Gemini on February 8, 2024.

PaLM2, which powers Gemini, has 160 languages and the future capability of "deductive reasoning" [28,29]. Gemini is part of the powerful Google search engine, which can access the user's past history of queries. Thus, Gemini can provide a tailored response for the user. Whether this will make Gemini as popular as ChatGPT, or not, is unknown since it is the latest LLM for Google.

OpenAI has ChatGPT and Microsoft has Copilot. Microsoft owns a large stake in OpenAI, the company that first launched ChatGPT. All the chatbots are free and easy to use, which will make them a source of health information.

Our patients have access to mobile apps and the internet, ChatGPT, Google Assistant, Gemini, and more AI chatbots. They can compare different answers which may be the solution for free patient education.

The use of these free chatbots can save the pediatric eye clinic time and money as resources become scarce with increasing patient load and electronic chart documentation demands. A medical assistant is paid US $38,000 per year in the continental United States [30]. An orthoptist has an average annual salary of US $90,000 [30]. The cost of ChatGPT 3.5 is zero. One can imagine the cost savings of 1 less medical assistant with full-time benefits versus a free Chatbot. At most, the paid version of ChatGPT 4.0 is US $20 per month. The cost savings are an incentive for all pediatric eye and general ophthalmology offices to be aware of this free resource. The ease of use will allow Open AI, the creator of ChatGPT, to gain market share and crush its competitors.

As physicians, we can assign medical assistants to show ChatGPT to selected patients with "lazy eyes." In some cases, patients want reassurance and repeated explanations of "what is amblyopia." In such cases, these chatbots and voice assistants are a cost-effective means of public education. They are useful to our patients with poor access to physicians.

In conclusion, the age of AI-mediated technology for patient education has arrived. The latest LLM is available and provides cheap and free patient education in office settings [24,29,30]. The LLMs may communicate possible minor inaccuracies and biases to the LLMs and Google Assistant will improve their "training." ChatGPT is already partnering with the Mayo Clinic and Stanford for conversational response emails to patients. There is ambient natural language processing now that will allow us to "talk" to our patients and our notes will be immediately transcribed into our electronic health records.

In 2024, the LLM chatbots will provide users with clear warning messages when asking about medical conditions. Other chatbots will provide web resources from clinicians that meet current medical guidelines. This is especially important since these models do not have access to the patient's medical history and cannot fully understand the complexity of a patient's health situation [31].

As ophthalmologists, we have the power to change ChatGPT and future LLMs through their "training" on our websites and journals. In fact, in the user's account, repeated questioning will "train" ChatGPT in 2024. This is new as of 2024. Individually as physicians, we can retype "see an ophthalmologist or pediatric ophthalmologist" as a necessary sentence fragment for our ophthalmic websites and web journals. From May 2023, to now, the current questions and answers have evolved in ChatGPT and the other chatbots to include phrasing that they are not physicians. There is evidence that the training is changing and by the time this paper is published, the AI-mediated chatbot answers will have improved with its million queries per day.

Patients and physicians may need Google Assistant, ChatGPT, and similar AI chatbots. We hope to harness the abundance of accessible information provided by LLMs to guide our patients' journey toward improved vision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT responses.
[DOCX File , 22 KB - jmir_v26i1e52401_app1.docx]

References

1. Hu B, Liu Z, Zhao J, Zeng L, Hao G, Shui D, et al. The global prevalence of amblyopia in children: a systematic review and meta-analysis. Front Pediatr 2022;10:819998 [FREE Full text] [doi: 10.3389/fped.2022.819998] [Medline: 35601430]

XSL•FO
RenderX

2. Attebo K, Mitchell P, Smith W. Visual acuity and the causes of visual loss in Australia. The Blue Mountains Eye study. Ophthalmology 1996;103(3):357-364. [doi: 10.1016/s0161-6420(96)30684-2] [Medline: 8600410]
3. Thompson B, Concetta Morrone M, Bex P, Lozama A, Sabel BA. Harnessing brain plasticity to improve binocular vision in amblyopia: an evidence-based update. Eur J Ophthalmol 2023:11206721231187426 [FREE Full text] [doi: 10.1177/11206721231187426] [Medline: 37431104]
4. Chua B, Mitchell P. Consequences of amblyopia on education, occupation, and long term vision loss. Br J Ophthalmol 2004;88(9):1119-1121 [FREE Full text] [doi: 10.1136/bjo.2004.041863] [Medline: 15317699]
5. Simons K, Stein L, Sener EC, Vitale S, Guyton DL. Full-time atropine, intermittent atropine, and optical penalization and binocular outcome in treatment of strabismic amblyopia. Ophthalmology 1997;104(12):2143-1255 [FREE Full text] [doi: 10.1016/s0161-6420(97)30048-7] [Medline: 9400777]
6. Cheng K, Li Z, He Y, Guo Q, Lu Y, Gu S, et al. Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. Ann Biomed Eng 2023;51(6):1130-1135 [FREE Full text] [doi: 10.1007/s10439-023-03203-3] [Medline: 37074486]
7. Pew Research Center. URL: https://www.pewresearch.org/ [accessed 2024-04-18]
8. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
9. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
10. Bader M, Zheng L, Rao D, Shiyanbola O, Myers L, Davis T, et al. Towards a more patient-centered clinical trial process: a systematic review of interventions incorporating health literacy best practices. Contemp Clin Trials 2022;116:106733 [FREE Full text] [doi: 10.1016/j.cct.2022.106733] [Medline: 35301134]
11. Kutner M, Greenberg E, Jin Y, Boyle B, Hsu Y, Dunleavy E. Literacy in everyday life: results from the 2003 national assessment of adult literacy. U.S. Department of Education. Washington, DC; 2007. URL: https://nces.ed.gov/Pubs2007/2007480_1.pdf [accessed 2024-04-01]
12. Camille LR, Bauman K. Educational attainment in the United States: 2015. U.S. Census Bureau. 2016. URL: https://www.census.gov/library/publications/2016/demo/p20-578.html [accessed 2024-04-03]
13. An L, Bacon E, Hawley S, Yang P, Russell D, Huffman S, et al. Relationship between coronavirus-related ehealth literacy and COVID-19 knowledge, attitudes, and practices among US adults: web-based survey study. J Med Internet Res 2021;23(3):e25042 [FREE Full text] [doi: 10.2196/25042] [Medline: 33626015]
14. Amblyopia. NEI. 2022. URL: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/amblyopia-lazy-eye [accessed 2024-06-26]
15. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy augmented intelligence in health care. J Med Syst 2022;46(2):12 [FREE Full text] [doi: 10.1007/s10916-021-01790-z] [Medline: 35020064]
16. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75. Millington, Tennessee; 1975. URL: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary [accessed 2024-03-16]
17. Neuhoff E, Feeser KM, Sutherland K. Flesch-Kincaid reading grade level re-examined: creating a uniform method for calculating readability on a certification exam. OJWED 2016;9(1):1-12 [FREE Full text]
18. Lee B, Dixon E, Wales DP. Evaluation of reading level of result letters sent to patients from an academic primary care practice. Health Serv Res Manag Epidemiol 2023;10:23333928231172142 [FREE Full text] [doi: 10.1177/23333928231172142] [Medline: 37122557]
19. How to measure readability. Pennsylvania Code and Bulletin.: Pennsylvania Dept of Insurance URL: https://www.pacodeandbulletin.gov/Display/pacode?file=/secure/pacode/data/031/chapter64/s64.14.html&d=reduce [accessed 2024-03-15]
20. Amblyopia. American Association for Pediatric Ophthalmology and Strabismus. URL: https://www.aapos.org/glossary/amblyopia [accessed 2024-04-15]
21. Readability test. WebFx.com. URL: https://www.webfx.com/tools/read-able/ [accessed 2024-03-15]
22. Flesch-Kincaid Readability Calculator. URL: https://readabilityformulas.com/free-readability-formula-tests.php [accessed 2024-03-15]
23. Erdem E, Kuyu M, Yagcioglu S, Frank A, Parcalabescu L, Plank B, et al. Neural natural language generation: a survey on multilinguality, multimodality, controllability, and learning. Journal of Artificial Intelligence Research. 2022. URL: https://www.jair.org/index.php/jair/article/download/12918/26783/30208 [accessed 2024-03-17]
24. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
25. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or pandora's box? JAMA Intern Med 2023;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]

XSL•FO
RenderX

26. Brender TD. Medicine in the era of artificial intelligence: hey chatbot, write me an H&P. JAMA Intern Med 2023 Jun 01;183(6):507-508. [doi: 10.1001/jamainternmed.2023.1832] [Medline: 37115537]
27. Metz C. What google bard can do (and what it can't). The New York Times. 2023. URL: https://www.nytimes.com/2023/03/21/technology/google-bard-guide-test.html [accessed 2024-04-10]
28. Anil R, Dai AM, Firat O. PaLM2 Technical Report. 2023. URL: https://arxiv.org/abs/2305.10403 [accessed 2024-04-16]
29. BARD. www.google.com. 2023. URL: https://bard.google.com/updates [accessed 2024-04-16]
30. Occupational outlook handbook. Bls.gov. 2023. URL: https://www.Bls.gov [accessed 2024-04-16]
31. Kalla D, Smith N, Samaah F, Kuraku S. Study and analysis of chat GPT and its impact on different fields of study. IJISRT 2023;8(3):827-233 [FREE Full text]

Abbreviations

AAPOS: American Association for Pediatric Ophthalmology and Strabismus
AI: artificial intelligence
LLM: large language model

Edited by Q Jin; submitted 01.09.23; peer-reviewed by X Yurui, J Sun, K Wrona; comments to author 22.02.24; revised version received 24.04.24; accepted 30.04.24; published 15.08.24.

Please cite as:
Wu G, Lee DA, Zhao W, Wong A, Jhangiani R, Kurniawan S
ChatGPT and Google Assistant as a Source of Patient Education for Patients With Amblyopia: Content Analysis
J Med Internet Res 2024;26:e52401
URL: https://www.jmir.org/2024/1/e52401
doi:10.2196/52401
PMID:39146013

©Gloria Wu, David A Lee, Weichen Zhao, Adrial Wong, Rohan Jhangiani, Sri Kurniawan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study

Kevin Danis Li[1,2], BS; Adrian M Fernandez[1], MD; Rachel Schwartz[3,4], PhD; Natalie Rios[1], BS; Marvin Nathaniel Carlisle[1], BS; Gregory M Amend[5], MD; Hiren V Patel[1], MD, PhD; Benjamin N Breyer[1,2], MAS, MD

[1]Department of Urology, University of California San Francisco, San Francisco, CA, United States

[2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, United States

[3]Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, United States

[4]Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, United States

[5]Department of Urology, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Kevin Danis Li, BS
Department of Urology
University of California San Francisco
400 Parnassus Ave
San Francisco, CA
United States
Phone: 1 415 353 2200
Email: kevin.d.li@ucsf.edu

Abstract

Background: Large language models including GPT-4 (OpenAI) have opened new avenues in health care and qualitative research. Traditional qualitative methods are time-consuming and require expertise to capture nuance. Although large language models have demonstrated enhanced contextual understanding and inferencing compared with traditional natural language processing, their performance in qualitative analysis versus that of humans remains unexplored.

Objective: We evaluated the effectiveness of GPT-4 versus human researchers in qualitative analysis of interviews with patients with adult-acquired buried penis (AABP).

Methods: Qualitative data were obtained from semistructured interviews with 20 patients with AABP. Human analysis involved a structured 3-stage process—initial observations, line-by-line coding, and consensus discussions to refine themes. In contrast, artificial intelligence (AI) analysis with GPT-4 underwent two phases: (1) a naïve phase, where GPT-4 outputs were independently evaluated by a blinded reviewer to identify themes and subthemes and (2) a comparison phase, where AI-generated themes were compared with human-identified themes to assess agreement. We used a general qualitative description approach.

Results: The study population (N=20) comprised predominantly White (17/20, 85%), married (12/20, 60%), heterosexual (19/20, 95%) men, with a mean age of 58.8 years and BMI of 41.1 kg/m^2. Human qualitative analysis identified “urinary issues” in 95% (19/20) and GPT-4 in 75% (15/20) of interviews, with the subtheme “spray or stream” noted in 60% (12/20) and 35% (7/20), respectively. “Sexual issues” were prominent (19/20, 95% humans vs 16/20, 80% GPT-4), although humans identified a wider range of subthemes, including “pain with sex or masturbation” (7/20, 35%) and “difficulty with sex or masturbation” (4/20, 20%). Both analyses similarly highlighted “mental health issues” (11/20, 55%, both), although humans coded “depression” more frequently (10/20, 50% humans vs 4/20, 20% GPT-4). Humans frequently cited “issues using public restrooms” (12/20, 60%) as impacting social life, whereas GPT-4 emphasized “struggles with romantic relationships” (9/20, 45%). “Hygiene issues” were consistently recognized (14/20, 70% humans vs 13/20, 65% GPT-4). Humans uniquely identified “contributing factors” as a theme in all interviews. There was moderate agreement between human and GPT-4 coding (κ=0.401). Reliability assessments of GPT-4’s analyses showed consistent coding for themes including “body image struggles,” “chronic pain” (10/10, 100%), and “depression” (9/10, 90%). Other themes like “motivation for surgery” and “weight challenges” were reliably coded (8/10, 80%), while less frequent themes were variably identified across multiple iterations.

Conclusions: Large language models including GPT-4 can effectively identify key themes in analyzing qualitative health care data, showing moderate agreement with human analysis. While human analysis provided a richer diversity of subthemes, the consistency of AI suggests its use as a complementary tool in qualitative research. With AI rapidly advancing, future studies

XSL•FO
RenderX

should iterate analyses and circumvent token limitations by segmenting data, furthering the breadth and depth of large language model–driven qualitative analyses.

(J Med Internet Res 2024;26:e56500) doi:10.2196/56500

KEYWORDS

artificial intelligence; ChatGPT; large language models; qualitative analysis; content analysis; buried penis; qualitative interviews; qualitative description; urology

Introduction

Recent advancements in artificial intelligence (AI), particularly in large language models, have significantly expanded their applications in health care and academic research. These developments raise critical questions about their potential and ethical use [1-3]. GPT-4, developed by OpenAI, is a large language model that uses deep learning algorithms, specifically the GPT, to process and generate human-like text [4]. Its training on diverse internet text sources through unsupervised learning enables it to interpret complex language data, making it a potentially invaluable tool for qualitative research [5]. This is especially important in areas where traditional qualitative data analysis is labor-intensive and requires expertise to understand subtle nuances [6]. Furthermore, it is unknown how AI-driven qualitative analysis may differ from human-driven analysis in research contexts.

Despite its potential, the application of AI and large language models to qualitative data remains underexplored [7,8]. Previous studies in the realm of qualitative data analysis have used traditional natural language processing (NLP) models, which often require benchmark-specific training and hand engineering, leading to a more constrained contextual understanding and inferencing abilities. For example, Lennon et al [9] combined human coding with an NLP system trained on internal data, significantly reducing coding time, while Cheligeer et al [10] used a model based on BERT (Bidirectional Encoder Representations from Transformers; Google) for faster keyword analysis. However, such models fall short of the advanced contextual and inferencing abilities exhibited by widely trained large language models like GPT-4, which has been shown to outperform traditional systems on standard NLP benchmarks [11]. Although the field is rapidly evolving, there remains a limited number of studies that directly compare AI-driven qualitative analysis with human-driven approaches [12-17].

In this study, we used GPT-4 to re-examine qualitative data from a previously published study of 20 patients with adult-acquired buried penis (AABP), a urological condition with significant psychosocial consequences, and compare its performance with that of human researchers [18]. Evaluating GPT-4 for qualitative analysis in this patient population is particularly important due to the unique and profound psychosocial distress associated with AABP, including issues related to body image, sexual function, and mental health. Understanding patients' experiences through qualitative analysis can provide an increased understanding of their lived experiences. To accomplish these objectives, we created a series of generalizable prompts that allow the application of GPT-4 to qualitative analysis without requiring specialized knowledge or skills [19]. Finally, we evaluated the validity of our approach by measuring agreement between GPT-4 and human analysis and reliability by assessing if prompts consistently elicited similar outputs from the same data.

Methods

Data Source

Qualitative data were from a convenience sample of 20 patients who presented to urology clinics participating in TURNS (Trauma and Urologic Reconstructive Network of Surgeons), a multi-institutional collaborative research group focused on urologic trauma and reconstruction [18]. We conducted semistructured interviews focusing on the impact of AABP on personal relationships, social life, mental health, and physical health. Participants were interviewed for 15 to 30 minutes, and audio was transcribed electronically using Otter transcription software [20]. Interviews were conducted over Zoom live video conferencing [21]. For both human and GPT-4 qualitative analyses, only deidentified text transcripts were used, ensuring that the qualitative data were interpreted solely from text, providing a comparable basis for both human and AI-driven analyses.

Human Analysis

Our human-driven analysis used a general qualitative description approach which differs from other qualitative methods in that the analytic process stays close to the data, describing informants' experiences using their own language [22-24]. The research team initially reviewed interview transcripts, taking notes to capture observations and ideas and facilitate a comprehensive understanding of the overall content. This preparatory work informed the subsequent structured coding process. To ensure consistency and reliability, the team convened at three key stages, which were (1) before coding, to share initial text impressions and establish a standardized coding protocol; (2) after initiating line-by-line coding, to discuss applied codes and refine categorization strategies; and (3) to assess coder interrater reliability using weighted Fleiss κ coefficients [25]. Codes with a κ value below 0.75 were discussed among all authors until a coding consensus was reached. This approach enabled the identification and categorization of relevant subthemes and themes.

AI Analysis

Each deidentified transcript underwent text formatting removal before analysis by GPT-4 using a standardized prompt set (Figure 1) [26]. The analysis of the GPT-4–generated output was conducted in 2 phases, the naïve phase and the comparison phase.

XSL•FO
RenderX

Figure 1. Procedure for using GPT-4 for qualitative description.

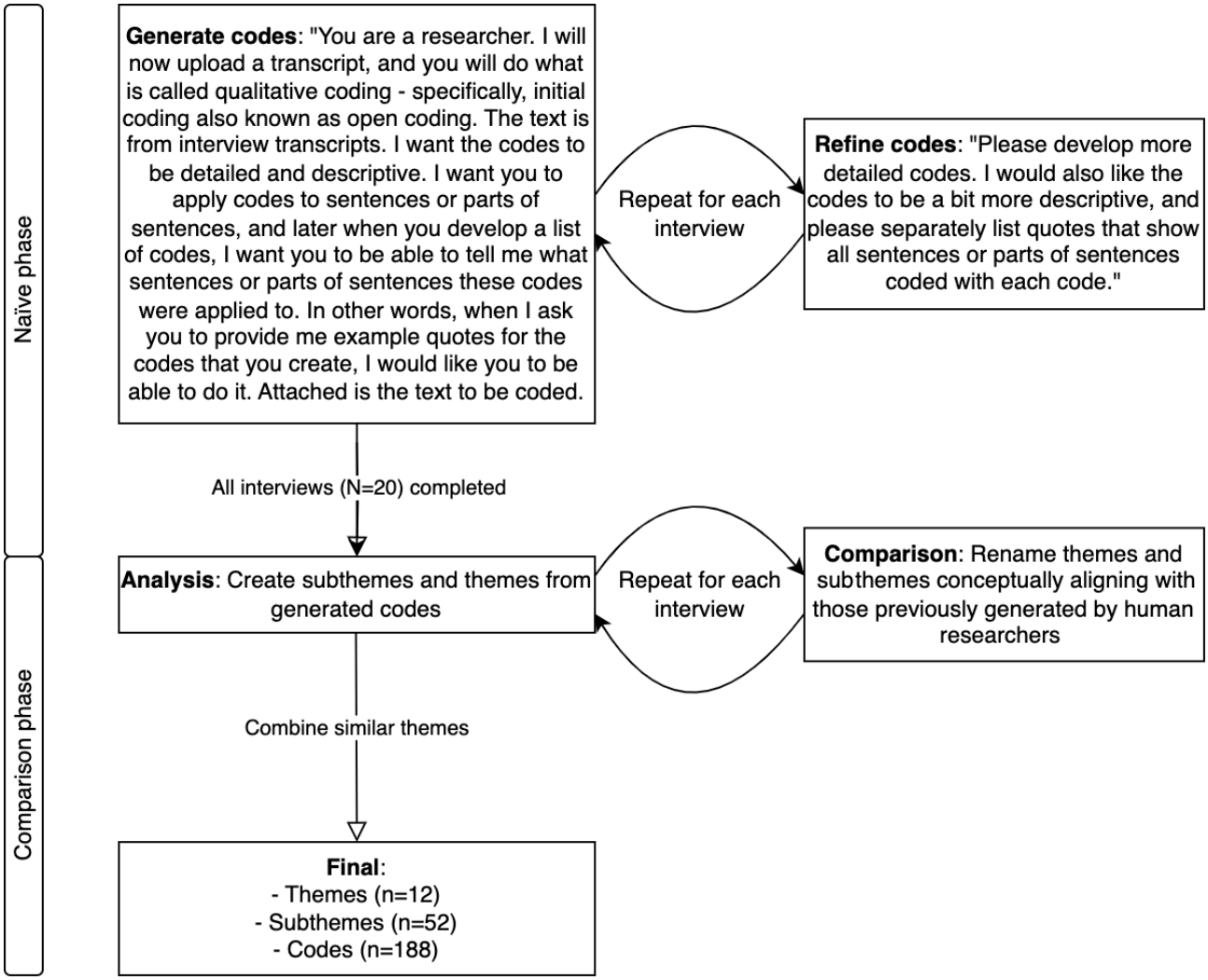

In the naïve phase, GPT-4's outputs for each interview were examined to extract relevant codes and quotes. These were then combined into subthemes, with groupings based on conceptual coherence and content relevance, following a standard qualitative description process [24]. Subsequently, similar subthemes were grouped to form overarching themes. Multiple iterations were conducted to refine the subthemes before synthesizing generalizations that held true across the data. Memo writing was integral to this process, capturing the evolving understanding of the data. Importantly, no discussions with the human-analyst team were conducted during this phase to avoid biasing the process. All interactions and evaluations of GPT-4's analyses were conducted by a blinded reviewer (KDL) who was not involved in the initial human-driven analysis and kept naïve to its outcomes.

In the comparison phase, AI-identified subthemes and themes were compared against those previously identified through human-driven analysis. This phase focused on identifying parallels and alignments between the 2 analyses to provide a direct comparison.

Interview data were collected in 2021, and human analyses were completed by 2022. All GPT-4 analyses were processed in separate instances on December 1, 2023, using the latest model of GPT-4 available at that time.

Measures to Ensure Rigor

The analytic team included KDL, who is a medical and data science master's student, NR, who is a clinical research coordinator with extensive experience in managing and coordinating clinical studies in health care settings, and GMA, who is a fellowship-trained surgeon specializing in urologic conditions, including adult acquired buried penis. In addition, we consulted BNB, an expert in urologic reconstruction who frequently treats patients with buried penis, to provide in-depth clinical insights and ensure the medical accuracy of our interpretations, and RS, a health services researcher and communication scientist with expertise in qualitative methods, to guide us on appropriate methodologies and ensure the rigor of our analyses.

To ensure rigor, we implemented several strategies addressing credibility, transferability, dependability, and confirmability [27]. For credibility, we built patient rapport through prolonged engagement, as most patients had existing longitudinal relationships at the urology clinics where they received care, allowing for deeper insights into their experiences. For transferability, we reported clinical characteristics of the study participants to inform the applicability of our findings to other populations with AABP and used a multi-institutional sampling

XSL•FO
RenderX

strategy to account for potential geographic or local institutional characteristics, ensuring broader applicability of our results.

Dependability was ensured through methodological documentation, where all codes, subthemes, and themes were documented at each step to provide transparency and replicability of our coding decisions. We also maintained detailed audit trails of raw outputs from GPT-4, processed outputs, and the subsequent organization into subthemes and themes, which the team reviewed to ensure consistency and reliability. Confirmability was achieved by having BNB, an expert in urologic reconstruction, review the study findings and provide critical insights during the design phase, and RS, who provided qualitative methodological support. In addition, data were shared with the entire research team, and feedback from all coauthors was incorporated into subsequent interpretation and analysis.

Comparison of Analyses

Qualitative analyses, including themes and subthemes, were summarized using descriptive statistics, including frequencies and proportions. To visually represent an agreement between human and AI-identified themes (validity), an agreement matrix was constructed. We measured interrater reliability using Cohen κ coefficient. A separate analysis was performed 10 times on the same interview transcript to assess the reliability of GPT-4's analysis. Themes identified exclusively by GPT-4 were highlighted with exemplar quotes that best represented each theme. All analyses were performed using R statistical software (version 4.3.1; The R Foundation).

Ethical Considerations

The study was approved by the University of California San Francisco (UCSF) institutional review board (IRB; 20-32062), and consent was obtained from all participants. In addition to the original study's IRB approval, we obtained an exemption from our institution's IRB for the secondary analysis using GPT-4, as the data were deidentified. Before analysis, all transcripts were reviewed to ensure that they contained no protected health information or identifiable data to maintain participant confidentiality. We used a private instance of GPT-4, known as Versa, which operates independently of OpenAI's commercial model and does not retain or learn from the data inputted [28]. This instance was used to develop our AI qualitative analysis methodology. For subsequent analyses, all data were confirmed to be thoroughly deidentified before using the commercial version of GPT-4.

Results

Study Population

Participant characteristics are summarized in Table 1. Participants' mean age and BMI were 58.8 (SD 13.9) years and 41.1 (SD 9.4) kg/m^2, respectively. Most participants were White (17/20, 85%), married (12/20, 60%), heterosexual (19/20, 95%) men residing in the Western region of the United States (10/20, 50%). In total, 55% (11/20) of participants underwent surgical correction of their AABP, with interviews conducted at an average of 497 (SD 666) days after surgery.

XSL•FO
RenderX

Table 1. Participant demographics and characteristics.

Characteristics	Values
Age (years), mean (SD)	58.8 (13.9)
BMI (kg/m^2), mean (SD)	41.1 (9.4)
Self-identified race, n (%)	
White	17 (85)
Black or African American	1 (5)
Other	2 (10)
Hispanic or Latin ethnicity	3 (15)
Relationship status, n (%)	
Married	12 (60)
Single	6 (30)
In a relationship	2 (10)
Sexual orientation, n (%)	
Heterosexual	19 (95)
Homosexual	1 (5)
Region, n (%)	
West	10 (50)
Northeast	7 (35)
Midwest	2 (10)
South	1 (5)
Patients who underwent AABP[a] surgical correction (n=11, 55%), n (%)	
Escutcheonectomy	9 (45)
Excision of penile skin with split-thickness skin graft	6 (30)
Ventral slit scrotal flap	5 (25)

[a]AABP: adult-acquired buried penis.

Qualitative Description

Table 2 presents a comparative analysis of themes and subthemes identified by human researchers versus GPT-4. “Urinary issues” were common in interviews analyzed by human researchers (19/20, 95%) and GPT-4 (15/20, 75%). Issues with “spray or stream” were a notable subtheme (12/20, 60% humans vs 7/20, 35% GPT-4). “Sexual issues” were prominently coded as well, present in 95% (19/20) of human-analyzed interviews and 80% (18/20) by GPT-4, with “inability to perform intercourse” coded as a subtheme more frequently by human researchers (12/20, 60% vs 6/20, 30%). Humans coded a broader array of sexual function issues, such as “pain with sex or masturbation” (7/20, 35%) and “difficulty with sex or masturbation” (4/20, 20%). “Mental health issues” were similarly recognized by both humans and GPT-4 (11/20, 55%, both), with “depression” more frequently coded by humans compared with GPT-4 (10/20, 50% vs 4/20, 20%, respectively). “Impact on social life” was an additional significant theme, with humans coding “issues using public restrooms” (12/20, 60%), while GPT-4 emphasized “struggles with romantic relationships” (9/20, 45%). Both methods identified “hygiene issues” (14/20, 70% humans vs 13/20, 65% GPT-4), highlighting difficulties in maintaining cleanliness. Human researchers uniquely identified “contributing factors” as a theme in all interviews.

XSL•FO
RenderX

Table 2. Human researchers versus GPT-4 qualitative analysis.

Themes and subthemes	Human researchers, n (%)	GPT-4, n (%)
Urinary issues	19 (95)	15 (75)
Spray or stream	12 (60)	7 (35)
Hovers over toilet	8 (40)	—[a]
Pain with urination	7 (35)	3 (15)
History of urethral stricture disease	3 (15)	—
Incontinence	3 (15)	3 (15)
Incomplete bladder emptying	3 (15)	2 (10)
Sits to urinate	2 (10)	—
Smelly urine	1 (5)	1 (5)
Trouble with catheter	1 (5)	—
Uses shower or tub to urinate	1 (5)	—
Frequent urination	—	2 (10)
Getting up at night to urinate	—	1 (5)
Sex issues	19 (95)	16 (80)
Unable to perform intercourse	12 (60)	6 (30)
Unable to get erection	9 (45)	3 (15)
Pain with sex or masturbation	7 (35)	—
Difficulty with sex or masturbation	4 (20)	—
Painful erection	4 (20)	—
Unable to maintain erection	3 (15)	—
Avoids sex	2 (10)	4 (20)
Unable to orgasm	2 (10)	—
Reduced genital sensation	1 (5)	—
Takes longer to orgasm	1 (5)	—
Pain with ejaculation	1 (5)	—
Intercourse not enjoyable	1 (5)	—
Adaptive masturbation techniques	—	2 (10)
Poor cosmetic appearance	—	2 (10)
Painful erection	—	2 (10)
Brittle skin	—	1 (5)
Unable to use condom	—	1 (5)
Overuse of pornography	—	1 (5)
Mental health issues	11 (55)	11 (55)
Depression	10 (50)	6 (30)
Feels like less of a man	7 (35)	4 (20)
Anxiety	4 (20)	2 (10)
Decreased self-esteem	3 (15)	—
Stress	1 (5)	1 (5)
Emotional turmoil	—	2 (10)
Loss of confidence	—	1 (5)
Guilt	—	1 (5)
Impacts social life	16 (80)	15 (75)

XSL•FO
RenderX

Themes and subthemes	Human researchers, n (%)	GPT-4, n (%)
Issues using public restrooms	12 (60)	8 (40)
Avoids travel	6 (30)	—
Struggles with romantic relationships	—	9 (45)
Mobility impairment	—	6 (30)
Spousal support	—	3 (15)
Avoids hobbies	—	1 (5)
Avoids social activities	—	1 (5)
Negative impact on career	—	1 (5)
Hygiene issues	14 (70)	13 (65)
Hard or effort to clean	11 (55)	11 (55)
Skin tearing	7 (35)	—
Penile bleeding	6 (30)	2 (10)
Infections	—	6 (30)
Contributing factors	20 (100)	—
Worse after weight gain	14 (70)	—
Worse after multiple surgeries	8 (40)	—
Worse after weight loss	4 (20)	—
Improvement after weight loss	0 (0)	—

[a]Not applicable.

Validity and Reliability of GPT-4 Analysis

To further assess the validity of GPT-4 analysis, we generated an agreement matrix comparing themes coded by human researchers and GPT-4 per interview (Figure 2). There were 63 instances where both human and GPT-4 analyses agreed on the presence of a theme, and 14 instances of agreement on a theme being absent. There was disagreement in 23 cases—16 where humans identified a theme that GPT-4 did not and 7 where GPT-4 identified a theme that humans did not (Table 3). The overall Cohen κ coefficient was 0.401, indicating moderate agreement. Boxes depict interview theme analysis. The blue (AI) and yellow (humans) squares indicate presence and green squares reflect agreement on presence or absence.

We assessed reliability by analyzing the same interview transcript 10 times with the same prompt set (Table 4). There was consistent identification of “body image struggles or disfigurement” and “chronic pain and discomfort,” both appearing in all iterations (10/10, 100%). “Depression” was also frequently coded, appearing in 90% (9/10) of analyses. High reliability was observed for “motivated to have surgery,” “uses shower or tub to urinate,” and “weight challenges,” each occurring in 80% (8/10) of the analyses. Other codes such as “issues using public restrooms,” “unable to perform intercourse,” and “negative health care experiences” were present in 70% (7/10) of iterations. Codes for “hard or effort to clean,” “decreased self-esteem,” and “necrotizing fasciitis diagnosis” were identified 60% (6/10) of the time. Codes were less frequent for “urinary tract infections” (3/10, 30%), “sits to urinate” (2/10, 20%), and a cluster of codes that included “dependency on others for care,” “social isolation and loneliness,” “high frequency of urination,” “anxiety,” “loss of physical autonomy,” “financial burden,” and “hematuria,” each appearing once (1/10, 10%).

XSL•FO
RenderX

Figure 2. Themes identified per interview by GPT-4 versus human researchers.

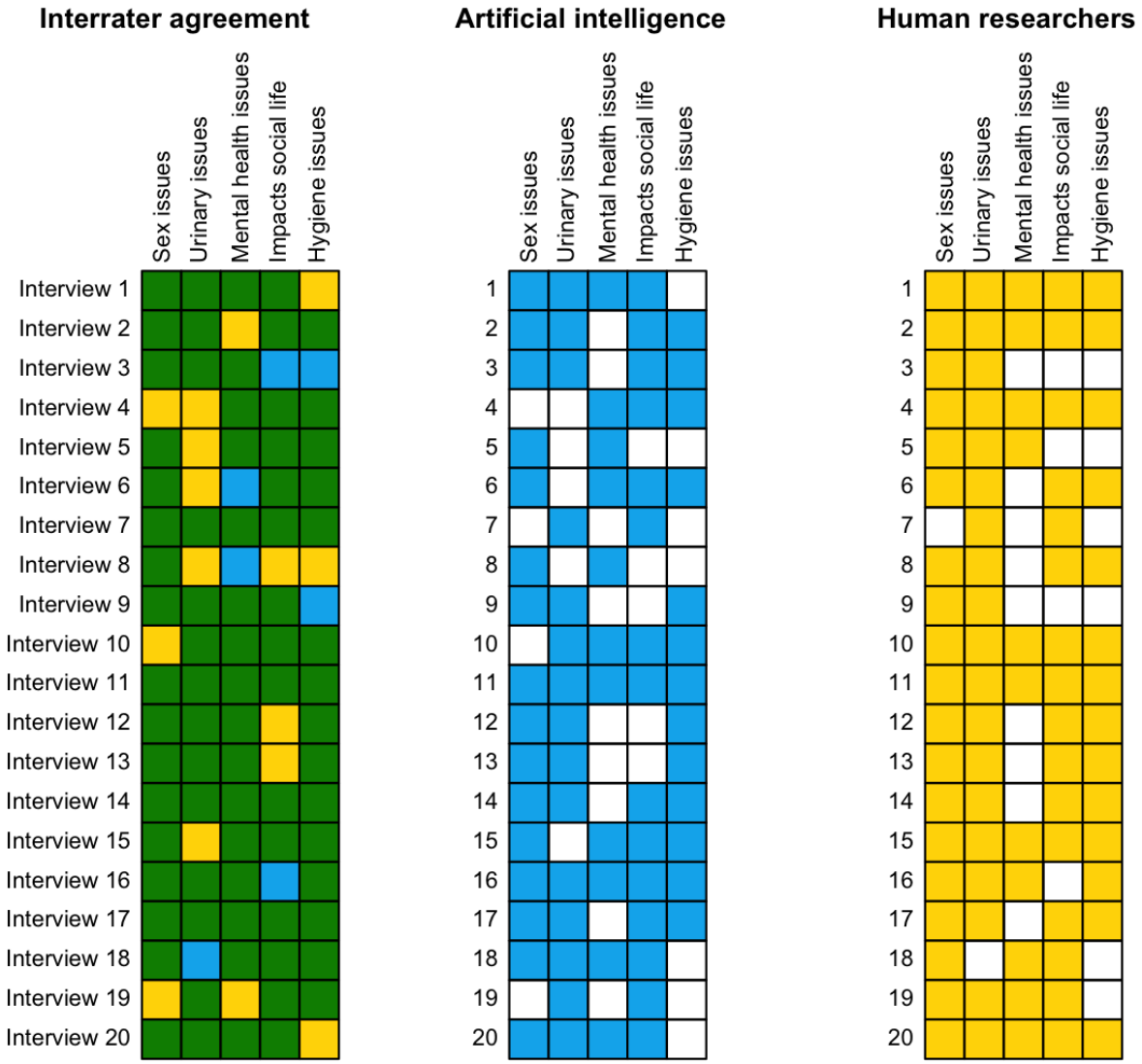

Table 3. Codes and exemplar quotes identified exclusively by GPT-4.

Interview number	GPT-4 code applied: exemplar quote	Theme
3	Impact on marital relationship: “I am married? And you know it’s it is... strained or? I wasn’t meeting her needs.”	Impacts social life
3	Hygiene management efforts: “I try to keep myself pretty clean... I really tried to wash my genitals really well.”	Hygiene issues
6	Mental health impact and resilience: “Yes in some ways it did affect me but other ways I don’t really don’t think it did.”	Mental health issues
8	Mental health and self-image concerns: “the preconceived notion you know but the man’s function is supposed to be.”	Mental health issues
9	Improved hygiene post surgery: “I actually feel that hygiene became a lot easier simply because I didn’t have to dig my finger in and run around the shaft to try and wash everything out.”	Hygiene issues
16	Day-to-day discontent and social withdrawal: “It’s just I just I would hate for other candidates that going forward thinking there is nothing that can be done need to be here they need to have options on the table.”	Impacts social life
18	Urinary dysfunction and social anxiety: “I would say they’re abnormal for somebody my age a lot of times it’s needing the needing to push… And that can cause anxiety in a public sort of restroom atmosphere.”	Urinary issues

XSL•FO
RenderX

Table 4. Reliability of GPT-4–generated codes.

Code[a]	Participants, n (%)
Body image struggles or disfigurement	10 (100)
Chronic pain and discomfort	10 (100)
Depression	9 (90)
Motivated to have surgery	8 (80)
Uses shower or tub to urinate	8 (80)
Weight challenges	8 (80)
Issues using public restrooms	7 (70)
Unable to perform intercourse	7 (70)
Negative health care experiences	7 (70)
Hard or effort to clean	6 (60)
Decreased self-esteem	6 (60)
Necrotizing fasciitis diagnosis	6 (60)
Urinary tract infections	3 (30)
Sits to urinate	2 (20)
Dependency on others for care	1 (10)
Social isolation and loneliness	1 (10)
High frequency of urination	1 (10)
Anxiety	1 (10)
Loss of physical autonomy	1 (10)
Financial burden	1 (10)
Hematuria	1 (10)

[a]Presence of codes from the same interview analyzed 10 times by GPT-4. Each code was counted only once per analysis, indicating whether it was identified (present) or not (absent) during each separate analysis.

Discussion

Principal Results

In this investigation, we directly compared the performance of AI (GPT-4) with human researchers in conducting a qualitative analysis of interviews with patients affected by AABP. Our study is the first of its kind, to our knowledge, to perform such a direct comparison, highlighting the potential use of AI in qualitative research. By using generalized prompts, our method allows researchers without specialized NLP knowledge to use GPT-4 for rigorous qualitative analysis, significantly reducing the time investment required.

Our results showed moderate alignment between GPT-4 and human analyses in identifying key themes, including urinary challenges, sexual health issues, and mental health impacts. Human analysis identified more subthemes, capturing the data's complexities more thoroughly than GPT-4. This difference may stem from GPT-4's token size limitations, which restrict its ability to perform comprehensive analyses as the input length increases [29]. The reliability tests revealed that while GPT-4 consistently recognized key codes, its identification of subtler codes was more variable. This suggests that implementing repeated analysis cycles, similar to the human multirater approach, could refine AI's analytical reliability. Overall, our findings underscore a complementary role for AI and human collaboration in qualitative research, where each can augment the strengths of the other.

The question of how to evaluate the accuracy and reliability of AI-driven analysis is crucial for future research. We adopted a quantitative approach to directly compare the presence of themes and subthemes in both human and AI analyses. By calculating Cohen κ, a statistic that measures interrater reliability by considering the agreement occurring by chance, we provided an objective assessment of the consistency of themes identified by GPT-4 compared with human analysis, presupposing human analysis as the "gold standard." In addition, to ensure consistency in GPT-4's outputs, we conducted multiple iterations of the same interview transcript analysis, analogous to traditional qualitative research methods where multiple analysts and iterative coding processes are used to standardize analyses and minimize biases. It is important to note that while these quantitative metrics offer a clear criterion for comparison, they may not fully capture the depth and richness of qualitative insights. GPT-4 has demonstrated the ability to detect subtle nuances and emotional contexts from text data, suggesting that incorporating more qualitative approaches in AI analysis evaluation could enhance the understanding of its analytical capabilities [30,31].

XSL•FO
RenderX

Limitations

A primary limitation of this study arises from the comparison phase, where themes and subthemes generated by GPT-4 were aligned with those identified by human researchers. Although a blinded reviewer was used to mitigate potential bias, the subjective nature of qualitative analysis means that a degree of bias is likely to remain. This is a common challenge in qualitative research, where analysts' subjective interpretations inherently influence their analysis. However, it can be argued that the use of a large language model such as GPT-4 may present a more objective method of analysis compared with the potential variability inherent between different human researchers' analyses, due to the large language model's consistent application of its transformer model.

We deliberately chose qualitative description as our analytic approach, favoring the accuracy to source material over depth of analysis. Qualitative description involves the systematic categorization and interpretation of qualitative data to uncover patterns and insights while staying close to the original data [22-24]. A more context-based approach, such as thematic analysis, could generate richer themes and subthemes but poses challenges for comparability. More interpretative methods may introduce subjectivity, reducing reproducibility. While our methodological choice ensures that our study remains accessible as a framework for others to build on and develop more interpretative techniques, the need for comparison limited our depth of insights.

Qualitative methods have inherent limitations, such as potential bias and limited generalizability due to smaller, nonrandom samples, and aim to produce in-depth insights and understanding rather than population inferences [32,33]. Consequently, our findings may not capture the full diversity of patient experiences, potentially limiting the generalizability of our results. Nevertheless, our study primarily aims to provide a comparative analysis, focusing on GPT-4 as a suitable tool for qualitative research applications.

As GPT-4 and other large language models advance, their analytical capabilities are expected to become more sophisticated, which may alter their proficiency in qualitative analysis. For example, while GPT-3.5 scored in the bottom 10% on a simulated bar examination, GPT-4 has demonstrated a significant improvement, placing within the top 10% of test takers [11]. The study's findings are therefore a snapshot of GPT-4's capabilities at a specific point in time and may not fully represent its future potential in qualitative analysis. Despite this limitation, the current trajectory of AI indicates that the use of GPT-4 and similar large language models in qualitative research is likely to become increasingly robust and refined.

Comparison With Previous Work

While studies applying GPT-4 or other large language models to qualitative research are limited, a growing body of work has compared the performance of OpenAI's GPT models, including GPT-3, -3.5, and -4, with that of humans in academic research and medical education [12-15]. Wang et al [34] found that while ChatGPT can generate accurate and relevant information, it is not without gaps when compared with official sources, indicating a need for supplementary validation from reliable references. Other studies have shown that ChatGPT can mimic the style of human-written research abstracts, albeit with limitations in quality and accuracy, as indicated by the ability of blinded reviewers to distinguish AI-generated content [35]. In the field of medical education, ChatGPT has been shown to outperform medical students on examinations, suggesting valuable applications in examination preparation [36]. Similarly, ChatGPT's performance on the United States Medical Licensing Examination (USMLE) further showcases the potential use of AI in medical education, where it achieved scores near the passing threshold without specialized training [37]. These findings emphasize that while advanced large language models such as GPT-4 are becoming increasingly competent in complex tasks, their current role remains complementary to human expertise.

The application of GPT-4 and other large language models to health care is a burgeoning field with substantial promise, resting on the fundamental ability of AI to process qualitative data efficiently. In patient care, large language models can enhance communication by translating complex medical language into more accessible terms for health care providers and patients [38]. The performance of large language models on medical licensing examinations also indicates their potential use in supporting clinical decision-making [39]. In administrative contexts, large language models are particularly valuable for generating concise clinical summaries and synthesizing extensive electronic medical record documentation; tasks that typically consume considerable time for health care professionals. The integration of large language models into administrative workflows may increase efficiency and allow clinicians to allocate more time to direct patient care. Health care companies are already beginning to integrate large language models into electronic health records, such as Epic's recent partnership with Microsoft to embed Azure OpenAI service into its own electronic health record systems [40].

Despite its promise, the integration of large language models in health care raises several ethical concerns that warrant careful consideration [41]. Foremost among these is data privacy, particularly regarding the handling of sensitive patient information, necessitating robust safeguards against data breaches. The opacity of these models, due to the unavailability of public training data sets and model weights, poses another concern as it obscures the understanding of their decision-making processes and challenges their trustworthiness in clinical applications [42]. In addition, the commercialization of large language models by major corporations, such as OpenAI, Microsoft, Meta, and Google, brings into question the potential influence of commercial interests on model development and deployment, possibly overshadowing patient welfare. A crucial concern is the risk of patient harm arising from incorrect or biased models, emphasizing the need for rigorous testing and validation of large language models to ensure their reliability and prevent adverse clinical outcomes [43].

XSL•FO
RenderX

Conclusions

Our research demonstrates that large language models like GPT-4 can discern key themes from qualitative health care data when used with standardized prompts. This "out of the box" approach aligns moderately well with qualitative description analysis by human analysts. Future work should use more context-based prompts for deeper and richer themes. As this may introduce greater subjectivity, researchers should also explore iterative analyses, such as synthesizing output from multiple iterations, to improve large language model output reliability. In addition, researchers should assess the qualitative analytic abilities of other popular models like Gemini (Google), Llama (Meta), and Claude (Anthropic AI), and develop methods to circumvent the token limitations inherent in models such as GPT-4 by segmenting qualitative data inputs, enriching the depth and breadth of qualitative analyses.

Acknowledgments

ChatGPT was not used in the ideation or writing of this manuscript.

Conflicts of Interest

None declared.

References

1. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology 2023;1(2):100017. [doi: 10.1016/j.metrad.2023.100017]
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141 [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
3. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
4. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
5. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. JMIR Med Inform 2023;11:e48933 [FREE Full text] [doi: 10.2196/48933] [Medline: 38015610]
6. Almeida F, Faria D, Queirós A. Strengths and limitations of qualitative and quantitative research methods. Eur J Educ Stud 2017;3(9):369-387 [FREE Full text]
7. Kantor J. Best practices for implementing ChatGPT, large language models, and artificial intelligence in qualitative and survey-based research. JAAD Int 2024;14:22-23 [FREE Full text] [doi: 10.1016/j.jdin.2023.10.001] [Medline: 38054196]
8. Hitch D. Artificial intelligence augmented qualitative analysis: the way of the future? Qual Health Res 2024;34(7):595-606 [FREE Full text] [doi: 10.1177/10497323231217392] [Medline: 38064244]
9. Lennon RP, Fraleigh R, Van Scoy LJ, Keshaviah A, Hu XC, Snyder BL, et al. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. Fam Med Community Health 2021;9(Suppl 1):e001287 [FREE Full text] [doi: 10.1136/fmch-2021-001287] [Medline: 34824135]
10. Cheligeer C, Yang L, Nandi T, Doktorchik C, Quan H, Zeng Y, et al. Natural language processing (NLP) aided qualitative method in health research. J Integr Des Process Sci 2023;27(1):41-58. [doi: 10.3233/jid-220013]
11. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. ArXiv Preprint posted online on March 04, 2024 [FREE Full text]
12. Zhang H, Wu C, Xie J, Lyu Y, Cai J, Carroll J. Redefining qualitative analysis in the AI era: utilizing ChatGPT for efficient thematic analysis. ArXiv Preprint posted online on May 28, 2024 [FREE Full text]
13. Hamilton L, Elliott D, Quick A, Smith S, Choplin V. Exploring the use of AI in qualitative analysis: a comparative study of guaranteed income data. Int J Qual Methods 2023;22:16094069231201504. [doi: 10.1177/16094069231201504]
14. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. International Journal of Qualitative Methods 2023;22. [doi: 10.1177/16094069231211248]
15. Wachinger J, Bärnighausen K, Schäfer LN, Scott K, McMahon SA. Prompts, pearls, imperfections: comparing ChatGPT and a human researcher in qualitative data analysis. Qual Health Res 2024:10497323241244669 [FREE Full text] [doi: 10.1177/10497323241244669] [Medline: 38775392]
16. A Fuller K, Morbitzer KA, Zeeman JM, M Persky A, C Savage A, McLaughlin JE. Exploring the use of ChatGPT to analyze student course evaluation comments. BMC Med Educ 2024;24(1):423 [FREE Full text] [doi: 10.1186/s12909-024-05316-2] [Medline: 38641798]
17. Amirova A, Fteropoulli T, Ahmed N, Cowie MR, Leibo JZ. Framework-based qualitative analysis of free responses of large language models: algorithmic fidelity. PLoS One 2024;19(3):e0300024 [FREE Full text] [doi: 10.1371/journal.pone.0300024] [Medline: 38470890]

XSL•FO
RenderX

18. Amend GM, Holler JT, Sadighian MJ, Rios N, Hakam N, Nabavizadeh B, et al. The lived experience of patients with adult acquired buried penis. J Urol 2022;208(2):396-405. [doi: 10.1097/JU.0000000000002667] [Medline: 35767655]
19. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
20. Otter.ai - AI Meeting Note Taker & Real-time AI Transcription. URL: https://otter.ai/ [accessed 2024-01-17]
21. One platform to connect. Zoom. URL: https://zoom.us/ [accessed 2024-01-17]
22. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
23. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
24. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description - the poor cousin of health research? BMC Med Res Methodol 2009;9:52 [FREE Full text] [doi: 10.1186/1471-2288-9-52] [Medline: 19607668]
25. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol 2016;16:93 [FREE Full text] [doi: 10.1186/s12874-016-0200-9] [Medline: 27495131]
26. Research with Dr Kriukow Youtube page.: Thematic analysis with ChatGPT | PART 1- Coding qualitative Data with ChatGPT; 2023 May 19. URL: https://www.youtube.com/watch?v=8dTs7D42ge0 [accessed 2024-07-25]
27. Ahmed SK. The pillars of trustworthiness in qualitative research. J Med Surg Public Health 2024;2:100051. [doi: 10.1016/j.glmedi.2024.100051]
28. Bengfort J. Now Available: Versa, UCSF Generative AI Platform.: Office of the Chancellor; 2024. URL: https://chancellor.ucsf.edu/news/now-available-versa-ucsf-generative-ai-platform [accessed 2024-07-25]
29. Kohn R. Mastering token limits and memory in ChatGPT and other large language models.: Medium; 2023. URL: https://medium.com/@russkohn/mastering-ai-token-limits-and-memory-ce920630349a [accessed 2024-07-25]
30. Baktash JA, Dawodi M. GPT-4: a review on advancements and opportunities in natural language processing. ArXiv Preprint posted online on May 04, 2023 [FREE Full text]
31. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058 [FREE Full text] [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
32. Borgstede M, Scholz M. Quantitative and qualitative approaches to generalization and replication-A representationalist view. Front Psychol 2021;12:605191 [FREE Full text] [doi: 10.3389/fpsyg.2021.605191] [Medline: 33613387]
33. Tenny S, Brannan J, Brannan G. Qualitative Study. StatPearls [Internet]. URL: https://www.ncbi.nlm.nih.gov/books/NBK470395/ [accessed 2024-07-06]
34. Wang G, Gao K, Liu Q, Wu Y, Zhang K, Zhou W, et al. Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: comprehensive comparative analysis of generative and authoritative information. J Med Internet Res 2023;25:e49771 [FREE Full text] [doi: 10.2196/49771] [Medline: 38096014]
35. Cheng SL, Tsai SJ, Bai YM, Ko CH, Hsu CW, Yang FC, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023;25:e51229 [FREE Full text] [doi: 10.2196/51229] [Medline: 38145486]
36. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, bing, and medical students in Germany. JMIR Med Educ 2023;9:e46482 [FREE Full text] [doi: 10.2196/46482] [Medline: 37665620]
37. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
38. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023;6(10):e2336997 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
39. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6(11):e2343689 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
40. Redmond W, Verona W. Microsoft and epic expand strategic collaboration with integration of azure openAI service.: Microsoft News Center; 2023 Apr 17. URL: https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/ [accessed 2024-07-25]
41. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
42. Sanderson K. GPT-4 is here: what scientists think. Nature 2023;615(7954):773. [doi: 10.1038/d41586-023-00816-5] [Medline: 36928404]
43. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6(1):e12-e22 [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]

XSL•FO
RenderX

Abbreviations

AABP: adult-acquired buried penis
AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
IRB: Institutional Review Board
NLP: natural language processing
TURNS: Trauma and Urologic Reconstructive Network of Surgeons
UCSF: University of California San Francisco
USMLE: United States Medical Licensing Examination

Edited by T de Azevedo Cardoso, Q Jin; submitted 17.01.24; peer-reviewed by L Zhu, MKK Rony, Y Zhong; comments to author 14.05.24; revised version received 31.05.24; accepted 09.07.24; published 21.08.24.

Please cite as:
Li KD, Fernandez AM, Schwartz R, Rios N, Carlisle MN, Amend GM, Patel HV, Breyer BN
Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study
J Med Internet Res 2024;26:e56500
URL: https://www.jmir.org/2024/1/e56500
doi:10.2196/56500
PMID:

©Kevin Danis Li, Adrian M Fernandez, Rachel Schwartz, Natalie Rios, Marvin Nathaniel Carlisle, Gregory M Amend, Hiren V Patel, Benjamin N Breyer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Diagnostic Ability of GPT-3.5 and GPT-4.0 in Surgery: Comparative Analysis

Jiayu Liu[1*], MD, PhD; Xiuting Liang[2*], MD; Dandong Fang[3*], MD; Jiqi Zheng[4*], PhD; Chengliang Yin[5,6*], PhD; Hui Xie[7*], MD; Yanteng Li[1], MD; Xiaochun Sun[5,6], PhD; Yue Tong[5,6], PhD; Hebin Che[5,6], PhD; Ping Hu[5,6], PhD; Fan Yang[1], MD; Bingxian Wang[1], MD; Yuanyuan Chen[5,6], PhD; Gang Cheng[1], MD; Jianning Zhang[1], MD, PhD

[1]Department of Neurosurgery, The First Medical Centre, Chinese PLA General Hospital, Beijing, China

[2]Department of Respiratory and Critical Care Medicine, The First Medical Centre, Chinese PLA General Hospital, Beijing, China

[3]Department of Neurosurgery, Sanmenxia Central Hospital, Sanmenxia, China

[4]School of Health Humanities, Peking University, Beijing, China

[5]Medical Innovation Research Division, Chinese People's Liberation Army General Hospital, Beijing, China

[6]National Engineering Research Center for Medical Big Data Application Technology, Chinese People's Liberation Army General Hospital, Beijing, China

[7]Departments of Urology, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China

[*]these authors contributed equally

Corresponding Author:
Jianning Zhang, MD, PhD
Department of Neurosurgery
The First Medical Centre
Chinese PLA General Hospital
No. 28 Fuxing Road
Haidian District
Beijing, 100853
China
Phone: 86 01066938439
Email: jnzhang2018@163.com

Related Article:

Comment in: http://jmir.org/2024/1/e66453/

Abstract

Background: ChatGPT (OpenAI) has shown great potential in clinical diagnosis and could become an excellent auxiliary tool in clinical practice. This study investigates and evaluates ChatGPT in diagnostic capabilities by comparing the performance of GPT-3.5 and GPT-4.0 across model iterations.

Objective: This study aims to evaluate the precise diagnostic ability of GPT-3.5 and GPT-4.0 for colon cancer and its potential as an auxiliary diagnostic tool for surgeons and compare the diagnostic accuracy rates between GTP-3.5 and GPT-4.0. We precisely assess the accuracy of primary and secondary diagnoses and analyze the causes of misdiagnoses in GPT-3.5 and GPT-4.0 according to 7 categories: patient histories, symptoms, physical signs, laboratory examinations, imaging examinations, pathological examinations, and intraoperative findings.

Methods: We retrieved 316 case reports for intestinal cancer from the Chinese Medical Association Publishing House database, of which 286 cases were deemed valid after data cleansing. The cases were translated from Mandarin to English and then input into GPT-3.5 and GPT-4.0 using a simple, direct prompt to elicit primary and secondary diagnoses. We conducted a comparative study to evaluate the diagnostic accuracy of GPT-4.0 and GPT-3.5. Three senior surgeons from the General Surgery Department, specializing in Colorectal Surgery, assessed the diagnostic information at the Chinese PLA (People's Liberation Army) General Hospital. The accuracy of primary and secondary diagnoses was scored based on predefined criteria. Additionally, we analyzed and compared the causes of misdiagnoses in both models according to 7 categories: patient histories, symptoms, physical signs, laboratory examinations, imaging examinations, pathological examinations, and intraoperative findings.

XSL•FO
RenderX

Results: Out of 286 cases, GPT-4.0 and GPT-3.5 both demonstrated high diagnostic accuracy for primary diagnoses, but the accuracy rates of GPT-4.0 were significantly higher than GPT-3.5 (mean 0.972, SD 0.137 vs mean 0.855, SD 0.335; t_{285}=5.753; P<.001). For secondary diagnoses, the accuracy rates of GPT-4.0 were also significantly higher than GPT-3.5 (mean 0.908, SD 0.159 vs mean 0.617, SD 0.349; t_{285}=–7.727; P<.001). GPT-3.5 showed limitations in processing patient history, symptom presentation, laboratory tests, and imaging data. While GPT-4.0 improved upon GPT-3.5, it still has limitations in identifying symptoms and laboratory test data. For both primary and secondary diagnoses, there was no significant difference in accuracy related to age, gender, or system group between GPT-4.0 and GPT-3.5.

Conclusions: This study demonstrates that ChatGPT, particularly GPT-4.0, possesses significant diagnostic potential, with GPT-4.0 exhibiting higher accuracy than GPT-3.5. However, GPT-4.0 still has limitations, particularly in recognizing patient symptoms and laboratory data, indicating a need for more research in real-world clinical settings to enhance its diagnostic capabilities.

(J Med Internet Res 2024;26:e54985) doi:10.2196/54985

KEYWORDS

ChatGPT; accuracy rates; artificial intelligence; diagnosis; surgeon

Introduction

Artificial intelligence (AI) can potentially improve health care outcomes as a tool in various fields of medicine [1]. While its role in medicine has thus far been limited to specific cases, it has not been widely implemented in clinical practice. Recently, OpenAI released an AI-powered ChatGPT for public use [2]. ChatGPT is a powerful neural network model belonging to the GPT family of large language models (LLMs). Despite being created primarily for humanlike conversations, ChatGPT has shown remarkable versatility and has the potential to revolutionize many industries [3]. Its release has garnered widespread attention for potential use in numerous fields, especially in medicine [4-6].

Recently, GPT-4.0, an upgraded version of GPT-3.5, demonstrated remarkable improvements in professional and academic benchmarks, including the US Medical Licensing Examination [7]. We previously used ChatGPT to answer medical questions and compared ChatGPT's abilities with evidence-based neurosurgeons [8]. Unlike low-seniority surgeons, ChatGPT has no lack of foundational knowledge and is not limited by cognitive load, especially GPT-4.0, which was comparable to that of surgeons with high seniority [8]. It also showed that ChatGPT generates semantically accurate responses to text-based questions, including medical textual data, making it a highly potential candidate for clinical decision support (CDS) applications.

Colorectal cancer is a type of cancer that affects the colon (large intestine) or rectum [9], which is the third most common cancer and the second leading cause of cancer-related deaths worldwide, accounting for approximately 10% of all cancer cases [9]. Colon cancer is a kind of disease with very obvious symptoms including diarrhea, constipation, blood in the stool, abdominal pain, unexplained weight loss, fatigue, and low iron levels [10]. A diagnosis can be made once a surgeon recognizes that a patient has these symptoms.

Surgeons usually use clinical experience to make diagnoses based on patient's symptoms, examination results, medical histories, physical signs, and intraoperative findings. This process will consume much time to occupy the workforce and will be subjective. Misdiagnoses or delays in diagnosis may occur. Innovative approaches to improve diagnostic accuracy and efficiency are needed. Many AI models have been used in health care to aid CDS [11-13]. However, few studies still exist on how ChatGPT can read and analyze clinical case text data and provide diagnoses based on real-world clinical information obtained at the hospital.

This study investigates the use of ChatGPT for diagnosis and assesses its performance. Furthermore, substantial performance improvement indicates that LLMs are quickly approaching readiness for use in the clinical setting. Continuous evaluation is necessary to keep pace with model progress. Therefore, we also compared the performance of two versions of ChatGPT (GPT-4.0 and GPT-3.5) to understand the impact of model updates on performance.

Methods

Study Design

We designed a comparative study to evaluate the applicability of ChatGPT (GPT-4.0 and GPT-3.5) in surgery using clinical information from the case database of the Chinese Medical Association Publishing House (CMAPH). GPT-3.5 is a free, widely accessible platform, while GPT-4.0 is a newer release claiming better performance, allowing us to evaluate improvements in GPT-4.0 comprehensively. The CMAPH database is a comprehensive and authoritative resource. It provides access to current and archived issues across various medical specialties, making it invaluable for researchers, clinicians, and health care professionals [14]. This study used colorectal cancer–related literature from this database, leveraging its extensive repository to support our analysis.

Study Data

A total of 316 case reports for intestinal cancer between August 1982 and June 2023 were retrieved from the CMAPH database by a researcher (CY); then, FY extracted the case records from the literature and designed a standardized data organization form by Excel (Microsoft Corp), which included extracted case records, three response records from GPT-3.5 and GPT-4.0, an evaluation section for the responses from GPT-3.5 and GPT-4.0,

XSL•FO
RenderX

and a detailed assessment of seven categories of misdiagnoses. Another researcher (BW) then conducted a thorough data cleansing process, removing cases with incomplete information, excluding diagnosis-related data from the literature, and translating the data according to clinical terminology. JL was responsible for verifying the translation results. We applied the World Health Organization's *ICD-11* (*International Classification of Diseases, 11th Revision*) diagnostic criteria [15,16] to these cases, excluding cases where the diagnosis and related information conflicted with the *ICD-11* standards. This process resulted in a final count of 286 valid cases for analysis. The collected data included age, gender, medical histories, symptoms, physical signs, laboratory examinations, imaging examinations, pathological examinations, and intraoperative findings, all structured as text data. Patient identities were kept confidential and anonymized for the analysis.

Prompt Design and Input Strategies

We aimed for GPT-3.5 and GPT-4.0 to provide appropriate primary and secondary diagnoses based on the case information from the literature data. Therefore, we designed the prompt in English, "Please provide the most likely primary and secondary diagnoses," and conducted tests accordingly. We randomly selected 5 of the 286 cases for initial testing, using the designed prompt to query GPT-3.5 and GPT-4.0 thrice each. It was a success if both models provided diagnoses without refusing to answer. JZ conducted these tests, and the answers obtained were recorded in the standardized data form. Subsequently, these 5 cases were not queried again. During these tests, both GPT-3.5 and GPT-4.0 successfully provided answers as required. Consequently, we finalized this prompt for use in this study. A researcher (JZ) used the same computer device to query GPT-3.5 and GPT-4.0 from August 1, 2023, to November 1, 2023, while another researcher (YT) was responsible for proofreading and verification.

Due to the potential for inconsistent responses from ChatGPT when the same question is asked multiple times, we posed the prompt for each case three times to GPT-3.5 and GPT-4.0. In this study, the prompt for each case was first posed three times using GPT-3.5, followed by three times using GPT-4.0, without providing any feedback to either model after each inquiry. To avoid any contextual learning that could affect the accuracy of the responses, each inquiry to GPT-3.5 and GPT-4.0 was conducted in a new chat session. The mean accuracy rate represents the average accuracy derived from these three inquiries for each model. We conducted the Fleiss κ test on the primary and secondary diagnostic responses provided by GPT-3.5 and GPT-4.0, defining consistency as the ability to provide the same information with each submission.

Data Measurement

The accuracy of GPT-3.5 and GPT-4.0 in providing primary and secondary diagnoses was evaluated based on explicit diagnoses in published literature. Diagnoses that matched the explicit diagnoses in the literature were recorded as "correct," while those that did not were recorded as "incorrect." An accuracy rate exceeding 0.800 for GPT-4.0 was deemed acceptable, while an accuracy rate exceeding 0.600 for GPT-3.5 was considered acceptable [17-19].

We assessed the diagnostic errors made by GPT-3.5 and GPT-4.0, categorizing them into 7 types: medical histories, symptoms, physical signs, laboratory examinations, imaging examinations, pathological examinations, and intraoperative findings. Due to the literature case records containing these 7 aspects, we discussed and decided to categorize the diagnostic errors accordingly. The criteria for each category are as follows: (1) medical histories: past medical conditions, treatments, surgeries, and other relevant health information; (2) symptoms: subjective indications reported by the patient, such as pain, fever, changes in bowel habits, rectal bleeding, abdominal pain, unexpected weight loss, and fatigue; (3) physical signs: objective observations during a physical examination including abdominal tenderness, palpable masses, anemia, or ascites; (4) laboratory examination: analysis of biological samples including blood chemistry, hematology, urinalysis, complete blood count for anemia, stool tests for occult blood, and tumor markers like carcinoembryonic antigen; (5) imaging examination: techniques to create internal body images such as x-rays, computed tomography scans, magnetic resonance imaging, ultrasound, and barium enemas; (6) pathological examination: microscopic analysis of tissue samples to determine cancer type and differentiation; and (7) intraoperative findings: observations during surgery regarding tumor location, size, extent, and involvement of adjacent organs or lymph nodes. For instance, if a diagnosis is primarily based on the patient's medical history and ChatGPT's response is incorrect, the error pattern is classified as a history-based error. If the diagnosis derives from multiple sources and ChatGPT provides an incorrect answer, we categorize all relevant sources as contributing to the error pattern.

We selected 3 senior surgeons not involved in this study to independently evaluate the diagnostic information and determine the primary and secondary diagnoses. All three senior surgeons are from the General Surgery Department, specializing in Colorectal Surgery, at the Chinese PLA (People's Liberation Army) General Hospital, each with at least 20 years of clinical experience. The National Health Commission of the People's Republic of China defines senior surgeons as chief physicians with over 3 years of experience as associate chief physicians in surgery. The primary and secondary diagnoses provided by GPT-3.5 and GPT-4.0 were scored based on the previously mentioned accuracy assessment criteria. For the primary diagnosis, responses were scored as "correct" with a value of 1 and "incorrect" with a value of 0. For secondary diagnoses, the scoring was based on the ratio of the number of correct secondary diagnoses provided by ChatGPT to the total number of correct secondary diagnoses in the literature. For example, if there were five correct secondary diagnoses and ChatGPT provided three correct ones, the score would be 3/5. The diagnostic errors of GPT-3.5 and GPT-4.0 were categorized according to the error mentioned above pattern classification criteria. Errors involving a category were assigned a value of 1, while noninvolvement was assigned a value of 0. The three senior surgeons used the standardized data recording form to document their evaluations, and discrepancies were discussed in meetings held at the hospital to reach a consensus on the scoring methodology to ensure consistency and accuracy in the

XSL•FO
RenderX

evaluation process, with interrater agreement assessed using Cohen κ statistic.

Statistical Analysis

The data were normally distributed as determined by the Kolmogorov-Smirnov test. We analyzed data using SPSS (version 25.0; IBM Corp). We used mean (SD) values to represent numerical variables and describe qualitative variables using absolute values of special group cases. The statistical methods are mainly based on *t* tests, and repeated measures analysis of variance is used for statistical analysis. The difference is statistically significant with $P<.05$.

Ethical Considerations

No ethical approval or informed consent was required for this study, as it used publicly available data. All study designs strictly adhere to the Guidelines and Checklist for the Reporting on Digital Health Implementations checklist [20]. This study does not use AI to generate any related content.

Results

Baseline Characteristics

Detailed questions and responses are included in Multimedia Appendix 1 (Figure S1 in Multimedia Appendix 1). Cohen κ ranged between moderate to near perfect agreement (Multimedia Appendix 2). For primary diagnoses, the mean accuracy rates of GPT-4.0 were 0.972 (SD 0.137), and GPT-3.5 were 0.855 (SD 0.335). The accuracy rates of GPT-4.0 were significantly higher than GPT-3.5 (t_{285}=5.753; $P<.001$). For secondary diagnoses, the mean accuracy rate of GPT-4.0 was 0.908 (SD 0.159), and GPT-3.5 was 0.617 (SD 0.349). The accuracy rates of GPT-4.0 were also significantly higher than GPT-3.5 (t_{285}=–7.727; $P<.001$; Table 1 and Figure 1A).

We solved any discrepancies. Of the 286 cases, 19 cases exhibited inconsistencies, while 267 cases were consistent, resulting in an agreement rate of 93.36%. The discrepancy rate stands at 6.64%. The three senior surgeons demonstrated excellent agreement in their assessments (Multimedia Appendix 3). We analyzed the ChatGPT performance according to error patterns. The rates of diagnosis errors due to medical histories in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.13, SD 0.520 vs mean 0.76, SD 1.446; t_{285}=–3.384; $P<.001$). The rates of diagnosis errors due to symptoms in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.33, SD 0.877 vs mean 0.55, SD 1.125; t_{285}=–3.840; $P=.001$). The rates of diagnosis errors due to physical signs in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.12, SD 0.516 vs mean 0.28, SD 0.867; t_{285}=–5.959; $P<.001$). The rates of laboratory examination due to medical histories in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.17, SD 0.645 vs mean 0.55, SD 1.131; t_{285}=–6.738; $P<.001$). The rates of diagnosis errors due to imaging examination in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.11, SD 0.483 vs mean 0.54, SD 1.125; t_{285}=–6.846; $P<.001$). The rates of diagnosis errors due to pathological examination in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.04, SD 0.294 vs mean 0.20, SD 0.739; t_{285}=–3.536; $P<.001$). The rates of diagnosis errors due to imaging examination in GPT-4.0 were significantly lower than in GPT-3.5 (mean 0.08, SD 0.403 vs mean 0.59, SD 1.258; t_{285}=14.006; $P<.001$; Table 1 and Figure 1B).

Table 1. Baseline characteristics.

Analyzed pairs	GTP-4.0, mean (SD)	GTP-3.5, mean (SD)	*t* test (*df*)	*P* value
Primary diagnoses	0.972 (0.137)	0.855 (0.335)	5.753 (285)	<.001
Secondary diagnoses	0.908 (0.159)	0.617 (0.349)	–7.727 (285)	<.001
Error patterns				
Medical histories	0.13 (0.520)	0.76 (1.446)	–3.384 (285)	<.001
Symptoms	0.33 (0.877)	0.55 (1.125)	–3.840 (285)	.001
Physical signs	0.12 (0.516)	0.28 (0.867)	–5.959 (285)	<.001
Laboratory examination	0.17 (0.645)	0.55 (1.131)	–6.738 (285)	<.001
Imaging examination	0.11 (0.483)	0.54 (1.125)	–6.846 (285)	<.001
Pathological examination	0.04 (0.294)	0.20 (0.739)	–3.536 (285)	<.001
Intraoperative findings	0.08 (0.403)	0.59 (1.258)	14.006 (285)	<.001

XSL•FO
RenderX

Figure 1. The comparison of GPT-4.0 and GPT-3.5. (A) The comparison of GPT-4.0 and GPT-3.5 in diagnostic accuracy for the primary and secondary diagnoses. (B) The comparison of GPT-4.0 and GPT-3.5 in error patterns (medical histories, symptoms, physical signs, examination results, and intraoperative findings). (C) The comparison of GPT-4.0 and GPT-3.5 in age, gender, and system group for the primary diagnoses. (D) The comparison of GPT-4.0 and GPT-3.5 in age, gender, and system group for the secondary diagnoses.

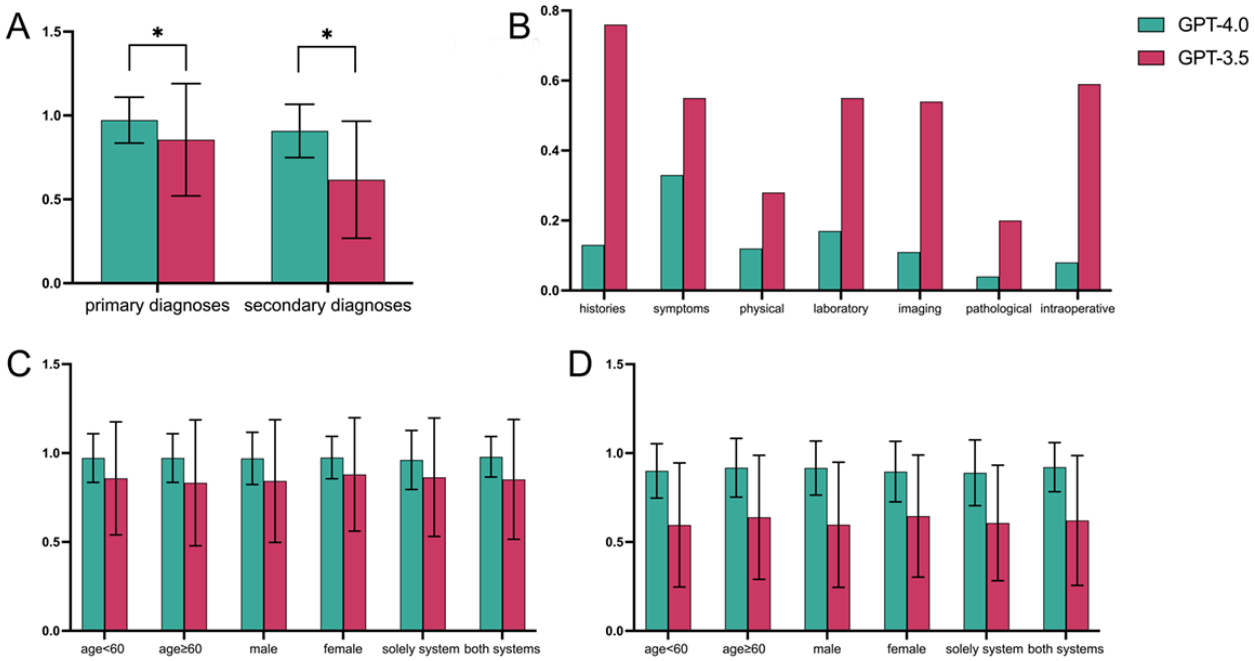

Comparison of Accuracy Rates According to Specialty Classification in Primary Diagnoses

Regarding age groups, there was no significant difference between ages of <60 years and ≥60 years in GPT-4.0 (mean 0.972, SD 0.137 vs mean 0.972, SD 0.137; t_{284}=0.002; *P*=.99), and also no significant difference between them in GPT-3.5 (mean 0.875, SD 0.318 vs mean 0.832, SD 0.354; t_{284}=1.086, *P*=.28). In the gender group, there was no significant difference between male and female in GPT-4.0 (mean 0.970, SD 0.147 vs mean 0.975, SD 0.119; t_{284}=–0.291; *P*=.77), and also no significant difference between them in GPT-3.5 (mean 0.842, SD 0.345 vs mean 0.879, SD 0.319; t_{284}=–0.898; *P*=.37). In the diagnostic type group, there was no significant difference between "solely intradigestive system" and "both intradigestive and extradigestive system" in GPT-4.0 (mean 0.961, SD 0.166 vs mean 0.979, SD 0.114; t_{284}=–1.063; *P*=.29), and also no significant difference between them in GPT-3.5 (mean 0.863, SD 0.333 vs mean 0.851, SD 0.337; t_{284}=0.308; *P*=.76; Table 2). Moreover, there was a significant difference between GPT-4.0 and GPT-3.5 in age, gender, and diagnostic type group (Figure 1C).

Table 2. Comparison of accuracy rates according to specialty classification in primary diagnoses.

Analyzed pairs	GPT-4.0, mean (SD)	GPT-3.5, mean (SD)	*t* test (*df*)	*P* value
Age (in years)[a]				
<60	0.972 (0.137)	0.875 (0.318)	3.688 (154)	<.001
≥60	0.972 (0.137)	0.832 (0.354)	4.445 (130)	<.001
Sex[b]				
Male	0.970 (0.147)	0.842 (0.345)	3.025 (106)	.003
Female	0.975 (0.119)	0.879 (0.319)	4.910 (178)	<.001
Diagnostic type[c]				
Solely intradigestive system	0.961 (0.166)	0.863 (0.333)	2.983 (111)	.004
Both intradigestive and extradigestive system	0.979 (0.114)	0.851 (0.337)	4.991 (173)	<.001

[a]In the age group, GPT-4.0 t_{284}=0.002 and *P*=.99; GPT-3.5 t_{284}=1.086 and *P*=.28.

[b]In the gender group, GPT-4.0 t_{284}=–0.291 and *P*=.77; GPT-3.5 t_{284}=–0.898 and *P*=.37.

[c]In the diagnostic type group, GPT-4.0 t_{284}=–1.063 and *P*=.29; GPT-3.5 t_{284}=0.308 and *P*=.76.

XSL•FO
RenderX

Comparison of Accuracy Rates According to Specialty Classification in Secondary Diagnoses

Regarding age group, there was no significant difference between ages of <60 years and ≥60 years in GPT-4.0 (mean 0.900, SD 0.153 vs mean 0.918, SD 0.165; t_{263}=–0.910; *P*=.36), and also no significant difference between them in GPT-3.5 (mean 0.596, SD 0.349 vs mean 0.639, SD 0.349; t_{264}=–1.006; *P*=.32). In the gender group, there was no significant difference between male and female in GPT-4.0 (mean 0.916, SD 0.152 vs mean 0.896, SD 0.170; t_{263}=0.972; *P*=.33), and also no significant difference between them in GPT-3.5 (mean 0.597, SD 0.352 vs mean 0.646, SD 0.343; t_{264}=–1.114; *P*=.27). In the diagnostic type group, there was no significant difference between "solely intradigestive system" and "both intradigestive and extradigestive system" in GPT-4.0 (mean 0.889, SD 0.185 vs mean 0.921, SD 0.138; t_{263}=–1.569; *P*=.12), and also no significant difference between them in GPT-3.5 (mean 0.607, SD 0.325 vs mean 0.621, SD 0.365; t_{264}=–0.333; *P*=.74; Table 3). Moreover, there was a significant difference between GPT-4.0 and GPT-3.5 in age, gender, and diagnostic type groups (Table 3 and Figure 1D).

Table 3. Comparison of accuracy rates according to specialty classification in secondary diagnoses.

Analyzed pairs	GPT-4.0, mean (SD)	GPT-3.5, mean (SD)	*t* test (*df*)	*P* value
Age (in years)[a]				
<60	0.900 (0.153)	0.596 (0.349)	10.764 (142)	<.001
≥60	0.918 (0.165)	0.639 (0.349)	8.974 (121)	<.001
Sex[b]				
Male	0.916 (0.152)	0.597 (0.352)	7.902 (99)	<.001
Female	0.896 (0.170)	0.646 (0.343)	11.613 (164)	<.001
Diagnostic type[c]				
Solely intradigestive system	0.889 (0.185)	0.607 (0.325)	8.850 (104)	<.001
Both intradigestive and extradigestive system	0.921 (0.138)	0.621 (0.365)	10.844 (159)	<.001

[a]In the age group, GPT-4.0 t_{263}=–0.910 and *P*=.36; GPT-3.5 t_{264}=–1.006 and *P*=.32.

[b]In the gender group, GPT-4.0 t_{263}=0.972 and *P*=.33; GPT-3.5 t_{264}=–1.114 and *P*=.27.

[c]In the diagnostic type group, GPT-4.0 t_{263}=–1.569 and *P*=.12; GPT-3.5 t_{264}=–0.333 and *P*=.74.

Discussion

Principal Findings

This comparative study evaluated the diagnostic capabilities of ChatGPT, specifically comparing the performance of GPT-4.0 and GPT-3.5 using clinical information from a public case database. Our analysis focused on specialty classification and error patterns including medical histories, symptoms, physical signs, examination results, and intraoperative findings. The results indicated that GPT-4.0 achieved significantly higher accuracy rates for primary and secondary diagnoses than GPT-3.5. Despite improvements, both versions exhibited limitations in processing patient history, symptom presentation, and laboratory test data, with GPT-4.0 showing better performance overall. These findings are consistent with the performance evaluations of GPT-3.5 and GPT-4.0 by Taloni et al [21], Yang et al [22], Deng et al [23], and Antaki et al [24]. However, challenges persist in accurately identifying symptoms and interpreting laboratory results. These findings underscore ChatGPT's potential in medical diagnosis, highlighting the need for further research to enhance its performance in dynamic clinical environments. This study contributes to ongoing efforts to optimize ChatGPT as an auxiliary tool in clinical practice, emphasizing the importance of continual refinement to ensure its efficacy and reliability in supporting health care professionals' diagnostic decision-making.

ChatGPT's ability to read and analyze is helpful in multiple clinical disciplines. Howard et al [25] asked ChatGPT for antimicrobial advice and pointed out that despite no specific clinical advice training, ChatGPT provides compelling responses to most prompts. Nastasi et al [26] found that ChatGPT is currently useful for providing background knowledge on general clinical topics in an emergency. Training ChatGPT with medical corpora, incorporating clinician-supervised feedback, and enhancing its awareness of uncertainty and information-seeking behaviors may improve the medical advice provided by LLMs. Lukac et al [27] evaluated ChatGPT as an adjunct for decision-making in primary breast cancer cases and stated that the eloquence of ChatGPT based on several scientific databases could result in more precise suggestions. Eggmann and Blatz [28] analyzed the chances and challenges of using ChatGPT in dentistry, they found that the use of ChatGPT and similar LLMs in dentistry could streamline administrative workflows and potentially serve as an additional tool for CDS in the future. Kuroiwa et al [29] discussed that the integration of AI, natural language processing, and GPT technologies holds immense promise in the field of psychiatry, which has the potential to revolutionize the way psychiatric disorders are diagnosed, treated, and monitored. Due to the evident symptoms of colon cancer, such as diarrhea, constipation, and blood in the stool, ChatGPT can provide basic, accurate diagnoses based on textual medical records. Previous studies have demonstrated ChatGPT's

XSL•FO
RenderX

strong ability to read and analyze text data. Therefore, in diagnosing colon cancer, ChatGPT can assist in improving doctors' work efficiency and reducing the workload of initial diagnoses.

Previous studies have evaluated the accuracy of ChatGPT from the perspective of patient self-diagnosis [29], but this study evaluated its accuracy based on clearly defined case record text data from publicly available literature and found its potential to assist in diagnosis, thereby reducing the time required to achieve accurate diagnosis and improving clinical efficiency. The application of ChatGPT in medical diagnosis offers several noteworthy advantages. We set a priori accuracy rate exceeding 0.800 for GPT-4.0 while a priori accuracy rate exceeding 0.600 for GPT-3.5, according to the previous studies [17-19]. In this study, we found that GPT-4.0 has a diagnostic accuracy above 0.900, while GPT-3.5 exhibits an accuracy of over 0.800 in the diagnosis of colon cancer. These results also confirmed that the diagnostic accuracy rates of GPT-3.5 and GPT-4.0 reached acceptable a priori accuracy rates, aligning with previous research findings.

For the primary and secondary diagnoses, the accuracy rates of GPT-4.0 were significantly higher than GPT-3.5. These results suggest significant enhancements in the algorithms and data processing capabilities of GPT-4.0. GPT-4.0 has more advanced neural network architectures, better natural language understanding, and a more extensive and diverse training data set than GPT-3.5, which enables it to handle complex medical diagnostic scenarios effectively [30-33]. GPT-4.0 better contextualizes information and the enhanced ability to understand and process complex medical terminology queries is crucial in clinical diagnosis [23,34]. GPT-4.0's enhanced ability to contextualize information and process complex medical terminology queries has been instrumental in clinical diagnosis, as it provides a deeper understanding of medical context, terminology, symptom presentation, and laboratory data interpretation [29]. Based on these considerations, the GPT-3.5 to GPT-4.0 upgrade appears effective.

In comparing the classification accuracy of the primary diagnosis, we found that GPT4.0 can achieve high accuracy in disease diagnosis across different ages, genders, and diagnostic type groups. The ability of GPT-4.0 to accurately diagnose patients of all ages and genders is crucial. We believe GPT-4.0 can catch the nuance between different age and gender groups, which can exhibit varying symptoms for the same disease. Besides, previous studies have also presented that GPT-4.0 shows its potential and the richness of its use scenarios in diagnosing a range of diseases across different systems. Hasani et al [35] evaluated the performance of GPT-4.0 in standardizing radiology reports, they indicated that GPT-4.0 could be a reliable tool for generating standardized radiology reports, offering potential benefits such as improved efficiency, better communication, and simplified data extraction and analysis. Liu et al [8] compared the abilities of GPT-4.0 and neurosurgeons, which showed that GPT-4.0's ability was comparable to that of neurosurgeons with high seniority, and Kanjee et al [36] assessed the diagnostic accuracy of the GPT-4.0 in a series of challenging cases and found that generative AI is a promising adjunct to human cognition in diagnosis. This broad range of diagnostic capabilities could be valuable in clinical practice to assist physicians.

In this study, we categorized the diagnostic errors of GPT-3.5 and GPT-4.0 based on seven aspects covered in the literature: medical histories, symptoms, physical signs, laboratory examinations, imaging examinations, pathological examinations, and intraoperative findings. These 7 aspects are crucial for making definitive diagnoses in clinical practice. Our multidimensional evaluation, therefore, reflects a realistic simulation of clinical application. This approach provides significant clinical value by highlighting the model's performance across all essential diagnostic dimensions. Such comprehensive assessment ensures that the evaluation of ChatGPT's diagnostic capabilities is both thorough and relevant to real-world clinical settings. GPT-3.5 has limitations in assessing patient history, symptom presentation, laboratory tests, and imaging data. GPT-3.5 is trained on large data sets, which may not extensively cover specialized medical information, exceptionally detailed patient histories, intricate symptomatology, and complex laboratory or imaging data. GPT-3.5 cannot interpret symptoms, lab results, and imaging data with a nuanced understanding of the patient's overall clinical picture including history, physical examination, and evolving clinical scenarios as physicians [37]. Furthermore, the knowledge of GPT-3.5 is static. It does not learn in real-time or adapt based on new patient information or the latest medical research, unlike physicians who continually update their knowledge and practice [38]. The above limitations have all been improved in GPT-4.0.

However, although GPT-4.0 has been upgraded and improved compared with GPT-3.5, things could still be improved. GPT-4.0 has limitations in identifying symptoms and laboratory test data by classifying and analyzing causes of misdiagnosis. First, like GPT-3.5, GPT-4.0 has limitations in training data scope and clinical context, and experience affects the model's ability to recognize and interpret specific medical data accurately. It may need more specific medical data, impacting its ability to accurately identify and analyze complex medical symptoms and laboratory results. Second, GPT-4.0 lacks clinical context and cannot accurately understand the clinical context in which medical data is presented [21,39]. For instance, interpreting laboratory results requires understanding the numbers and the clinical context including patient history and physical examination findings. Third, GPT-4.0 is a general model, not adjusted explicitly for medical applications [40]. While it can process medical information, it cannot replace a human medical professional's nuanced judgment and experiential learning [41].

In summary, although the evaluation results of GPT-4.0 in various classification diagnostic accuracies are better than GPT-3.5, it still has certain limitations, and its accuracy in medical diagnosis should be viewed with caution. It is a supportive tool [42], not a replacement for professional medical advice and judgment. It has yet to fully reach the diagnostic ability of clinical doctors. In the future, it is still necessary to continue training and improving GPT-4.0's ability to recognize patient symptoms and laboratory test data.

XSL•FO
RenderX

Limitations

This study has several limitations. First, we did not evaluate the models in real clinical settings. The dynamics of real-time clinical environments, complex and often unpredictable patient presentations, intraoperative findings, real imaging data, and pathological sections differ markedly from the controlled, historical data sets we used. Consequently, our findings do not fully encompass the challenges and variances in actual clinical practice. Second, we compared the diagnostic capabilities of GPT-3.5 and GPT-4.0 with predefined diagnoses in the literature rather than assessing their performance against the actual diagnostic abilities of human surgeons. Third, the study only included GPT-3.5 and GPT-4.0 models. With the rapid advancement of AI, numerous other models, such as Google's Bard and Microsoft's Copilot, have emerged. Future research should consider incorporating a broader range of AI models to evaluate their diagnostic capabilities comprehensively. Additionally, given that colorectal cancer is a prevalent disease in the field of surgery, this study merely serves as an initial exploration of ChatGPT's application in surgery, using colorectal cancer as an example. We also did not assess GPT-3.5 and GPT-4.0's ability to judge disease severity. In the future, we intend to include a broader range of surgical diseases to explore the potential of GPT-assisted clinical applications in surgery fully. The different prompts can yield varying answers. In this study, we crafted a simple and direct prompt and conducted multiple tests with both models, as subtle variations can influence the accuracy and relevance. Standardizing prompts is crucial to ensure the comparability and reliability of AI diagnostic performance assessments.

Moreover, the diagnostic basis relied solely on textual patient records from existing literature, which only demonstrates GPT-3.5 and GPT-4.0's ability to extract and analyze textual information. This study needs to include and evaluate imaging and pathology interpretation, which are critical components of comprehensive clinical diagnosis. Another limitation is that the patient records were originally in Chinese and were translated into English before querying GPT-3.5 and GPT-4.0. Both machine and human translations may only partially capture the nuances and details of the original records, potentially affecting the accuracy and completeness of the information presented to the models. Finally, the study's design as a cross-sectional comparative study provides a lower level of evidence in the hierarchy of evidence-based medicine compared to randomized controlled trials. Future research should involve randomized controlled trials for more robust and reliable assessments of GPT-assisted clinical applications.

Conclusions

This study shows that ChatGPT has potential in medical diagnosis and may serve as a tool to assist doctors in clinical diagnosis and improve work efficiency. Generally, the diagnostic accuracy of GPT-4.0 is better than that of GPT-3.5, indicating that the upgrade of the ChatGPT version is affected. However, GPT-4.0 still has limitations regarding patient symptoms and laboratory data recognition, which needs to be further studied in the dynamic clinical practice environment in the future.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JL and JZ contributed to the literature search, study design, data collection, data interpretation, and writing. All use of ChatGPT in this study was done by JZ and HX. GC, YT, YL, XS, HC, PH, DF, FY, BW, XL, and YC provided feedback on manuscript texts. J Zhang and CY contributed to the study design and provided feedback on all manuscript texts.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The case prompt provided to ChatGPT.
[DOCX File , 403 KB - jmir_v26i1e54985_app1.docx]

Multimedia Appendix 2
Cohen Kappa statistic for GPT-3.5 or GPT-4.0 agreement.
[DOCX File , 16 KB - jmir_v26i1e54985_app2.docx]

Multimedia Appendix 3
Cohen Kappa statistic for senior surgeons agreement.
[DOCX File , 16 KB - jmir_v26i1e54985_app3.docx]

References

XSL•FO
RenderX

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
2. Introducing ChatGPT. Open AI. 2022. URL: https://openai.com/blog/chatgpt [accessed 2023-08-01]
3. Gordijn B, Have HT. ChatGPT: evolution or revolution? Med Health Care Philos 2023;26(1):1-2. [doi: 10.1007/s11019-023-10136-0] [Medline: 36656495]
4. Abdel-Messih MS, Boulos MNK. ChatGPT in clinical toxicology. JMIR Med Educ 2023;9:e46876 [FREE Full text] [doi: 10.2196/46876] [Medline: 36867743]
5. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
6. Rao SJ, Isath A, Krishnan P, Tangsrivimol JA, Virk HUH, Wang Z, et al. ChatGPT: a conceptual review of applications and utility in the field of medicine. J Med Syst 2024;48(1):59. [doi: 10.1007/s10916-024-02075-x] [Medline: 38836893]
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience 2023;26(9):107590 [FREE Full text] [doi: 10.1016/j.isci.2023.107590] [Medline: 37705958]
9. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
10. Holtedahl K, Borgquist L, Donker GA, Buntinx F, Weller D, Campbell C, et al. Symptoms and signs of colorectal cancer, with differences between proximal and distal colon cancer: a prospective cohort study of diagnostic accuracy in primary care. BMC Fam Pract 2021;22(1):148 [FREE Full text] [doi: 10.1186/s12875-021-01452-6] [Medline: 34238248]
11. De Panfilis L, Peruselli C, Tanzi S, Botrugno C. AI-based clinical decision-making systems in palliative medicine: ethical challenges. BMJ Support Palliat Care 2023;13(2):183-189. [doi: 10.1136/bmjspcare-2021-002948] [Medline: 34257065]
12. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023;25:e48659 [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
13. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
14. Chinese Medical Association Publishing House (CMAPH). URL: https://www.yiigle.com/index [accessed 2023-07-20]
15. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak 2021;21(Suppl 6):206 [FREE Full text] [doi: 10.1186/s12911-021-01534-6] [Medline: 34753471]
16. The Lancet. ICD-11. Lancet 2019;393(10188):2275. [doi: 10.1016/S0140-6736(19)31205-X] [Medline: 31180012]
17. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Head Neck Surg 2024;170(6):1492-1503 [FREE Full text] [doi: 10.1002/ohn.489] [Medline: 37595113]
18. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024;44(2):303-306 [FREE Full text] [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]
19. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
20. Franck CP, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
21. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American academy of ophthalmology. Sci Rep 2023;13(1):18562 [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]
22. Yang WH, Chan YH, Huang CP, Chen TJ. Comparative analysis of GPT-3.5 and GPT-4.0 in Taiwan's medical technologist certification: a study in artificial intelligence advancements. J Chin Med Assoc 2024;87(5):525-530. [doi: 10.1097/JCMA.0000000000001092] [Medline: 38551357]
23. Deng L, Wang T, Yangzhang, Zhai Z, Tao W, Li J, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg 2024;110(4):1941-1950 [FREE Full text] [doi: 10.1097/JS9.0000000000001066] [Medline: 38668655]
24. Antaki F, Milad D, Chia MA, Giguère C, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol 2023:bjo-2023-bj324438. [doi: 10.1136/bjo-2023-324438] [Medline: 37923374]

XSL•FO
RenderX

25. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis 2023;23(4):405-406. [doi: 10.1016/S1473-3099(23)00113-5] [Medline: 36822213]
26. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023;13(1):17885 [FREE Full text] [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
27. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023;308(6):1831-1844 [FREE Full text] [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
28. Eggmann F, Blatz MB. ChatGPT: chances and challenges for dentistry. Compend Contin Educ Dent 2023;44(4):220-224. [Medline: 37075729]
29. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621 [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
30. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg 2024;110(6):3701-3706 [FREE Full text] [doi: 10.1097/JS9.0000000000001312] [Medline: 38502861]
31. Currie G, Robbie S, Tually P. ChatGPT and patient information in nuclear medicine: GPT-3.5 versus GPT-4. J Nucl Med Technol 2023;51(4):307-313. [doi: 10.2967/jnmt.123.266151] [Medline: 37699647]
32. Duey AH, Nietsch KS, Zaidat B, Ren R, Ndjonko LCM, Shrestha N, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J 2023;23(11):1684-1691. [doi: 10.1016/j.spinee.2023.07.015] [Medline: 37499880]
33. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
34. Liang R, Zhao A, Peng L, Xu X, Zhong J, Wu F, et al. Enhanced artificial intelligence strategies in renal oncology: iterative optimization and comparative analysis of GPT 3.5 versus 4.0. Ann Surg Oncol 2024;31(6):3887-3893. [doi: 10.1245/s10434-024-15107-0] [Medline: 38472675]
35. Hasani AM, Singh S, Zahergivar A, Ryan B, Nethala D, Bravomontenegro G, et al. Evaluating the performance of generative pre-trained transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol 2024;34(6):3566-3574. [doi: 10.1007/s00330-023-10384-x] [Medline: 37938381]
36. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
37. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol 2024. [doi: 10.1007/s00062-024-01426-y] [Medline: 38806794]
38. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. BMJ Neurol Open 2023;5(1):e000451 [FREE Full text] [doi: 10.1136/bmjno-2023-000451] [Medline: 37337531]
39. Ahmed W, Saturno M, Rajjoub R, Duey AH, Zaidat B, Hoang T, et al. ChatGPT versus NASS clinical guidelines for degenerative spondylolisthesis: a comparative analysis. Eur Spine J 2024. [doi: 10.1007/s00586-024-08198-6] [Medline: 38489044]
40. Delsoz M, Madadi Y, Raja H, Munir WM, Tamm B, Mehravaran S, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea 2024;43(5):664-670. [doi: 10.1097/ICO.0000000000003492] [Medline: 38391243]
41. Li S. Exploring the clinical capabilities and limitations of ChatGPT: a cautionary tale for medical applications. Int J Surg 2023;109(9):2865-2867 [FREE Full text] [doi: 10.1097/JS9.0000000000000523] [Medline: 37222684]
42. Cheng K, Li Z, Guo Q, Sun Z, Wu H, Li C. Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. Int J Surg 2023;109(6):1816-1818 [FREE Full text] [doi: 10.1097/JS9.0000000000000410] [Medline: 37074733]

Abbreviations

AI: artificial intelligence
CDS: clinical decision support
CMAPH: Chinese Medical Association Publishing House
ICD-11: International Classification of Diseases, 11th Revision
LLM: large language model
PLA: People's Liberation Army

XSL•FO
RenderX

Edited by N Cahill, G Eysenbach; submitted 29.11.23; peer-reviewed by L Zhu, L Jinhui; comments to author 30.01.24; revised version received 05.02.24; accepted 24.07.24; published 10.09.24.

Please cite as:
Liu J, Liang X, Fang D, Zheng J, Yin C, Xie H, Li Y, Sun X, Tong Y, Che H, Hu P, Yang F, Wang B, Chen Y, Cheng G, Zhang J
The Diagnostic Ability of GPT-3.5 and GPT-4.0 in Surgery: Comparative Analysis
J Med Internet Res 2024;26:e54985
URL: https://www.jmir.org/2024/1/e54985
doi:10.2196/54985
PMID:39255016

©Jiayu Liu, Xiuting Liang, Dandong Fang, Jiqi Zheng, Chengliang Yin, Hui Xie, Yanteng Li, Xiaochun Sun, Yue Tong, Hebin Che, Ping Hu, Fan Yang, Bingxian Wang, Yuanyuan Chen, Gang Cheng, Jianning Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage of Simulated Patients Using the Simple Triage and Rapid Treatment (START) Protocol: Gage Repeatability and Reproducibility Study

Jeffrey Micheal Franc[1,2], MSc, MS, MD, Dip Sport Med; Attila Julius Hertelendy[3,4], PhD; Lenard Cheng[5], MBBS; Ryan Hata[6], MD; Manuela Verde[2], MSc, MD

[1]Department of Emergency Medicine, University of Alberta, Edmonton, AB, Canada

[2]CRIMEDIM—Center for Research and Training in Disaster Medicine, Humanitarian Aid and Global Health, Universita' del Piemonte Orientale, Novara, Italy

[3]Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, FL, United States

[4]Department of Emergency Medicine, Beth Isreal Deaconess Medical Center, Harvard Medical School Teaching Hospital, Boston, MA, United States

[5]Emergency Medicine Department, National University Hospital, National University Health System, Singapore, Singapore

[6]Department of Emergency Medicine, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Jeffrey Micheal Franc, MSc, MS, MD, Dip Sport Med
Department of Emergency Medicine
University of Alberta
736c University Terrace
8203-112 Street NW
Edmonton, AB, T6R2Z6
Canada
Phone: 1 7807006730
Email: jeffrey.franc@ualberta.ca

Abstract

Background: The release of ChatGPT (OpenAI) in November 2022 drastically reduced the barrier to using artificial intelligence by allowing a simple web-based text interface to a large language model (LLM). One use case where ChatGPT could be useful is in triaging patients at the site of a disaster using the Simple Triage and Rapid Treatment (START) protocol. However, LLMs experience several common errors including hallucinations (also called confabulations) and prompt dependency.

Objective: This study addresses the research problem: “Can ChatGPT adequately triage simulated disaster patients using the START protocol?” by measuring three outcomes: repeatability, reproducibility, and accuracy.

Methods: Nine prompts were developed by 5 disaster medicine physicians. A Python script queried ChatGPT Version 4 for each prompt combined with 391 validated simulated patient vignettes. Ten repetitions of each combination were performed for a total of 35,190 simulated triages. A reference standard START triage code for each simulated case was assigned by 2 disaster medicine specialists (JMF and MV), with a third specialist (LC) added if the first two did not agree. Results were evaluated using a gage repeatability and reproducibility study (gage R and R). Repeatability was defined as variation due to repeated use of the same prompt. Reproducibility was defined as variation due to the use of different prompts on the same patient vignette. Accuracy was defined as agreement with the reference standard.

Results: Although 35,102 (99.7%) queries returned a valid START score, there was considerable variability. Repeatability (use of the same prompt repeatedly) was 14% of the overall variation. Reproducibility (use of different prompts) was 4.1% of the overall variation. The accuracy of ChatGPT for START was 63.9% with a 32.9% overtriage rate and a 3.1% undertriage rate. Accuracy varied by prompt with a maximum of 71.8% and a minimum of 46.7%.

Conclusions: This study indicates that ChatGPT version 4 is insufficient to triage simulated disaster patients via the START protocol. It demonstrated suboptimal repeatability and reproducibility. The overall accuracy of triage was only 63.9%. Health care professionals are advised to exercise caution while using commercial LLMs for vital medical determinations, given that these tools may commonly produce inaccurate data, colloquially referred to as hallucinations or confabulations. Artificial

XSL•FO
RenderX

intelligence–guided tools should undergo rigorous statistical evaluation—using methods such as gage R and R—before implementation into clinical settings.

(J Med Internet Res 2024;26:e55648) doi:10.2196/55648

KEYWORDS

disaster medicine; large language models; triage; disaster; emergency; disasters; emergencies; LLM; LLMs; GPT; ChatGPT; language model; language models; NLP; natural language processing; artificial intelligence; repeatability; reproducibility; accuracy; accurate; reproducible; repeatable

Introduction

Despite the progression of computerized artificial intelligence (AI) since the 1950s, its use has largely remained unreachable for clinicians. Recently, chat-oriented large language models (LLMs) have presented a markedly more user-friendly alternative. Engaging with these models involves a straightforward, conversational interaction: the user provides text input (referred to as the prompt), and the model responds with text output. The public launch of ChatGPT (OpenAI) in November 2022 significantly lowered the barriers to AI use by offering access to an LLM via an uncomplicated web interface, thereby enabling anyone with fundamental computer knowledge to leverage AI for any issue that can be articulated in text. The success of an innovation, such as ChatGPT, depends on its use by end users, which requires acceptability, acceptance, and adoption which is often complex when applied in the clinical setting [1].

Although chat-based LLMs are powerful and intuitive, they are not without shortcomings and have two major faults: (1) hallucinations or confabulations: LLMs can generate outputs, or "hallucinations," that are syntactically correct and superficially plausible but factually inaccurate or entirely fabricated [2]. These issues arise due to the model's lack of a grounded understanding of reality and its reliance on patterns in the training data. For instance, it might generate dates, events, or statements that are completely false but sound believable. Addressing hallucinations is a significant challenge, and researchers are exploring various techniques, including refining training processes and developing mechanisms to verify generated information against reliable sources [3]. (2) Prompt dependence and promptology: the way prompts are framed can significantly influence the responses generated by LLMs [4]. Slight changes in phrasing or specifying different contexts can yield varying outputs. While it allows for some level of "tuning" of the responses by modifying the prompt, it also introduces challenges related to the consistency and reliability of the outputs. The study and practice of crafting and optimizing prompts to extract desired outputs from LLMs is sometimes referred to as "prompt engineering" or "promptology." It has become a subfield of its own, where researchers explore how to guide the models efficiently to produce useful and reliable responses.

These limitations underscore the importance of continued research and development in the field of AI and natural language processing. To make these models more reliable, accountable, and useful, researchers and engineers use various strategies such as integrating external knowledge bases for fact-checking, improving the training data and methodology, and developing ways to mitigate biases present in the outputs. The ongoing research and development aim to improve the robustness, reliability, and utility of LLMs while also ensuring that they are used ethically and responsibly across various applications.

Among the many possible uses for ChatGPT in medicine, patient triage is an interesting case study. AI-guided triage could be particularly useful in disaster medicine, where circumstances may require inexperienced health care providers to perform triage. One can imagine triage performed by AI where the provider enters patient information as text, and the patient's triage code is outputted. This might allow physicians, first responders, nurses, or even the patients themselves to perform triage. However, at present, it is unclear whether LLMs can provide sufficiently accurate triage for use in a disaster.

Presently, the Simple Triage and Rapid Treatment (START) algorithm, which divides disaster casualties into four different triage codes, is the most commonly used model for disaster triage [5]. Disaster casualties are classified into four groups that are named by priority and color: red (immediate), yellow (delayed), green (ambulatory), and black (expectant) [5]. This model is based on a simple flowchart and seems ideal for use by AI models.

To date, there is a paucity of studies that examine the use of ChatGPT among health care providers expected to perform triage in a disaster setting. Our study extends the work by Liu et al [6] that examined the AI-generated suggestions from ChatGPT to optimize clinical decision support. Additionally, our research extends a call for additional research that examines AI chatbot efficacy by evaluating the precision of different AI algorithms in conducting disaster triage [7].

Furthermore, there is a growing body of literature that examines the factors that influence the acceptability of and adoption of ChatGPT [8]. The extant literature provides insights into health domains and common medical pathways where chatbots have been studied.

This study addresses the research problem: "Can ChatGPT adequately triage simulated disaster patients using the START protocol?" This study used ChatGPT to triage simulated patient vignettes and assessed three major outcomes:

- Repeatability: variation in response to repeated use of the same prompt with the same patient vignette.
- Reproducibility: variation in response with the use of different prompts with the same patient vignette.
- Diagnostic accuracy: overall accuracy of triage when compared to a documented reference standard.

XSL•FO
RenderX

Methods

Study Design

This study was based on a crossed gage repeatability and reproducibility (gage R and R) study with a comparison to a standard [9]. An easily approachable summary of this methodology is available on the American Society of Quality website [10].

The gage R and R study was chosen as the preferred methodology for several reasons. First, it is well documented in the industrial quality control literature as a valid and reliable means to measure the accuracy of instruments by ensuring results are repeatable and reliable [11]. Second, the method can separate variation due to the tool (ChatGPT) and variation due to the operators (prompts). Third, the method allows the calculation of repeatability and reproducibility without reliance on a reference standard. Fourth, the method is based on the calculation of the range: START scores are ordinal (not categorical or ratio) and should be analyzed with methods that do not rely on linear or normal data. Finally, the use of this well-documented method allows other researchers to reproduce the study.

Gage R and R studies use repeated measurements of multiple parts by various appraisers using the same equipment. This ensures that the output is the same as the input and that the same measurements occur over time [11]. In this study, the patient vignettes are the different "parts" while the various prompts are the "appraisers." Overall variability is broken down into repeatability (variation when the same prompt is used repeatedly on the same vignette) and reproducibility (variation when different prompts are used on the same vignette). In this study, 9 prompts (appraisers), 391 simulated patients (parts), and 10 repetitions of each combination were performed. In total, there were 3519 unique combinations of prompt and vignette, and a total of 35,190 simulated patient triages.

The gage R and R was accomplished in four parts: (1) each combination of prompt and vignette was submitted to ChatGPT ten times via the application programming interface (API). (2) Repeatability was calculated based on the average range of all trials using the same prompt and vignette. In an ideal system—where ChatGPT returns the same response each time—the range would be zero. (3) Reproducibility was calculated based on the average range of all trials using the same vignette. Again, in an ideal system, ChatGPT would give the same result for each vignette regardless of the prompt and the range would be zero. (4) Overall variability due to repeatability and reproducibility (overall gage R and R) was calculated based on the sum of the squares of variation due to repeatability and variation due to reproducibility [11].

Study Setting

The study was performed in September 2023 using the ChatGPT API version GPT-4.

Participants

The simulated patient vignettes were based on 391 simulated adult and pediatric patients previously used in disaster medicine training and simulation. A total of 380 simulated patients were taken from the most current version of the Disastermed.ca database which has expanded from its originally published size [12]. As this training set contained few patients with the "Black" triage code, 11 further cases with the published triage standard of "Black" were taken from various previously published training sets [13-17]. The final patient data set of 391 patients contained patients of various ages and diverse traumatic and nontraumatic presentations. Several sample vignettes, in the format that they were used to prompt ChatGPT, are presented in Multimedia Appendix 1. The average age for the simulated patients was 29.9 (SD 18.2; range 0-80) years with 109 (27.9%) patients less than or equal to 16 years of age. Sex was male in 277 (71%) patients and female in 114 (29%) patients.

Reference standards for START were assigned by the research team. First, two authors (JMF and MV) assigned triage codes independently to each simulated patient vignette using the standard START procedure—categorizing each patient as red, yellow, green, or black. Interrater reliability was calculated using Cohen κ [18]. Where the two authors did not agree, a third author (LC) again repeated the triage independently as a tiebreaker, and the triage code assigned by two of the three reviewers was used. All three authors assessing the reference standard triage code had fellowship training or master's degrees in disaster medicine and between 2 and 15 years of disaster medicine experience.

To preserve the ordinality of the START levels, the assigned triage was rescored as an integer: 1 for red, 2 for yellow, 3 for green, and 4 for black. This order parallels the order of priority given by the North Atlantic Treaty Organization triage system [17].

Patient vignettes were collated into a MySQL (version 8.0.27; Oracle) database.

Interventions

A total of 9 unique ChatGPT prompts were tested (Multimedia Appendix 2). Each prompt was created by one of the study authors. Authors were asked to create their prompts independently without consultation with the other authors.

Outcome Measures

Three outcome measures were determined: (1) repeatability (equipment variation): the extent to which repeated use of the same prompt for the same patient vignette led to variation in response was measured as a percentage of total variation calculated by the range method [11]. This represented variability inherent to the ChatGPT tool itself. (2) Reproducibility (appraiser variation [AV]): the extent to which the use of different prompts on the same patient vignette led to variation in response was measured as a percentage of total variation calculated by the range method [11]. This represented variability due to the use of different prompts. (3) Diagnostic accuracy: agreement of the triage scores was recorded when the ChatGPT gave the same triage code as the reference standard determined by the experts. Overtriage occurred when ChatGPT gave a lower number (higher acuity) rating than the reference standard. Undertriage occurred when ChatGPT gave a higher number (lower acuity) than the reference standard. Accuracy was

XSL•FO
RenderX

calculated as a percentage of the total ChatGPT queries that gave a valid response.

Gage R and R was calculated as the summed square of repeatability and reproducibility [11]. Previously published guidelines rate the usability of a measurement tool in relation to the total gage R and R: less than 10% is considered adequate for general use, 10% to 30% adequate for use in low-risk situations, and more than 30% inadequate [9,11].

Data Collection

The ChatGPT API was queried by an automated Python (version 3.6.9) script (Multimedia Appendix 3) [19].

Briefly, the script looped through each of the nine prompts. Each prompt was combined with each vignette and the text "Answer only with the color code and no other text" was appended. The ChatGPT API was then queried 10 times with this combination, before moving on to the next simulated patient, and finally moving on to the next prompt. In the end, 10 trials of each of the nine prompts were combined with 391 simulated cases for a total of 35,190 simulated patient triages.

Triage scores were translated from a color code to an integer by the script as described in the section "Participants" above.

Data Analysis

Statistical analysis was performed using Stat59 (STAT59 Services Ltd). As the final triage scores were ordinal, repeatability and reproducibility were calculated with the range method, as this method does not require assumptions of normality or linearity [11].

Sample Size

No formal sample size or power calculation is available for gage R and R studies. Small sample sizes are customary with gage R and R. The use of 10 parts, 3 operators, and 2 replicates is a common design [10]. This study used a convenience sample of all available simulated trauma casualties from the Disastermed.ca database resulting in a much larger design of 391 parts (vignettes), 9 operators (prompts), and 10 replicates.

Ethical Considerations

This study involved no human participants. The simulated casualty vignettes were previously written by the first author (JMF) for a simulation exercise [12]. The reference standards for the simulated casualties were assigned by 3 of the study authors (JMF, MV, and LC).

Results

When the first two raters assigned the START scores to the 391 simulated patients, they agreed on 345 (88.2%). κ for the interrater reliability was 0.81 (95% CI 0.76-0.86). Following the assignment by a third rater to the 46 cases where the initial two raters did not agree, the final data set included 25 simulated patients with the standard values of black, 35 red, 130 yellow, and 201 green.

In total, 35,190 requests were made to the ChatGPT API by the Python script. In 35,102 queries (99.7%), a valid response was returned that contained only the triage color code. In 88 (0.3%) cases, the response returned was not valid: containing either no text or text without a single triage color code. There were 391 unique combinations of prompt and vignette. A frequency table for the range of the 10 trials for each combination is shown in Table 1. Notably, in only 1735 (49.3%) was the range zero: indicating that ChatGPT chose the same triage code for all ten trials. Conversely, in slightly more than half of the cases, each of the ten trials did not give the same response. Repeatability, as calculated by the range method, was 0.17, indicating that 14% of the overall variation was due to equipment variability: variability due to ChatGPT itself.

Table 2 shows the average range for each of the 9 prompts. The lowest average range (mean 0.36, SD 0.59) was for prompt #8, which indicates that this prompt had the best reproducibility. Conversely, the highest average range was for prompt #4 (mean 1.27, SD 0.80) indicating the lowest reproducibility. Reproducibility, as calculated by the range method, was 0.091, indicating that 4.2% of the overall variability was due to AV: variability due to differences in the prompts.

A frequency table of the range for each of the 391 simulated patient vignettes is shown in Table 3. In only 27 (6.9%) was the range zero, indicating that ChatGPT gave the same triage code for all trials and all prompts. The remainder of the variability as calculated by the range method (81%) was due to variation within the case vignettes.

Overall gage repeatability and reproducibility as calculated by the range method was 0.19%.

Table 4 shows the triage accuracy for the nine prompts. The overall accuracy for ChatGPT to perform START was 63.9% (SD 0.25%). Among the nine prompts, the maximum accuracy was 71.8% (SD 0.7%) and the minimum was 46.7% (SD 0.8%).

A confusion matrix of the ChatGPT assigned values of START versus the reference standard is presented in Figure 1. The overtriage rate for all vignettes was 32.9%, with a 3.1% undertriage rate. For the reference standard of red, 80.9% of vignettes matched the reference standard, with a 19.1% undertriage rate. Reference standard yellow patients were triaged matching the reference standard in 50.8% of cases, with a 45% overtriage and a 4.2% undertriage rate. For patients with a reference standard of green, the rate of agreement with the reference standard was 69.4%, with a rate of 30.6% overtriage, and 0.04% undertriage. For patients with a START standard of black, the rate of triage matching the reference standard was 64.6%, with a 35.4% overtriage rate.

XSL•FO
RenderX

Table 1. Range of the 10 trials for each of the 3519 unique combinations of prompt and vignette.

Range	Number of observations, n	Percentage of observations, %
0	1735	49.3
1.0	1459	41.5
2.0	226	6.4
3.0	99	2.8

Table 2. The average range of the 10 trials for each of the 9 prompts.

Prompt	Range, mean (SD)
1	0.46 (0.66)
2	0.49 (0.7)
3	0.72 (0.68)
4	1.27 (0.80)
5	0.66 (0.76)
6	0.53 (0.61)
7	0.54 (0.62)
8	0.36 (0.59)
9	0.61 (0.72)

Table 3. Range of the 10 trials for each of the 391 simulated patients.

Range	Number of observations, n	Percentage of observations, %
0	27	6.9
1.0	174	44.5
2.0	150	38.3
3.0	40	10.2

Table 4. Accuracy for each of the 9 prompts.

Prompt	Accuracy (%), mean (SD)
1	71.8 (0.7)
2	72.0 (0.7)
3	63.1 (0.8)
4	46.7 (0.8)
5	68.7 (0.7)
6	60.6 (0.8)
7	66.8 (0.8)
8	54.4 (0.8)
9	71.3 (0.7)

XSL•FO
RenderX

Figure 1. Confusion matrix comparing the ChatGPT assigned START score versus the reference standard. START: Simple Triage and Rapid Treatment.

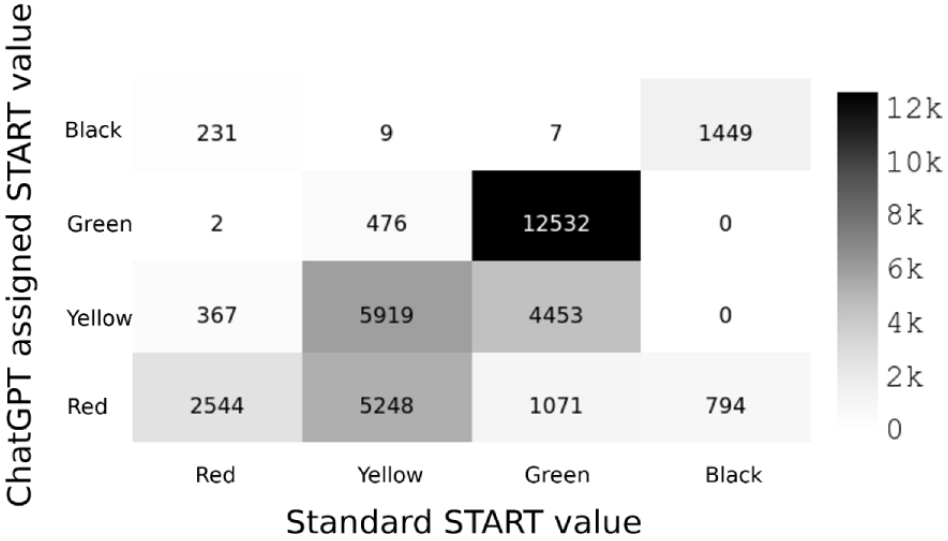

Discussion

Interpretation

This study suggests that ChatGPT version 4.0 is inadequate to accurately triage simulated patients using the START model.

Repeatability, variation due to ChatGPT itself, was 14% of the total variation. This indicates excess variability when the same prompt is used multiple times to triage the same patient vignette. Practitioners should be aware that using the same prompt and the same patient data multiple times does not always give the same response.

Reproducibility, or AV, was lower at 4.2% of overall variation, suggesting that the influence of how the prompt is written has less of an effect on total variation than does the inherent variation within ChatGPT itself. Among the prompts, prompt #4 had the most variability, while prompt #8 had the least.

The overall accuracy of ChatGPT (63.9%) to triage patients using START should not be surprising based on how LLMs are developed. LLMs are not expert systems: the software itself has no actual knowledge of the problem it is addressing. The system does not inherently use the same algorithm as a human would to triage patients. Instead, LLMs rely on probability principles that use complex mathematical models to predict text output based on the prompt without any knowledge of the question asked.

Previous Studies

Previous studies have documented that ChatGPT accuracy and reliability are highly variable [20].

In this study, ChatGPT displayed a lower triage accuracy than reports using traditional human-applied triage. For example, a meta-analysis in 2022 showed the START accuracy of human participants to be 73%, with an overtriage rate of 14% and an undertriage rate of 10% [5]. A recently published study assessed the use of ChatGPT version 3.5 to triage patients using the Canadian Triage and Acuity Score. The study showed similar results with a variation due to reliability of 14%, a variation due to reproducibility of 4.1%, and overall accuracy of 61.4% [21].

Strengths and Limitations

This study was limited to a single LLM: ChatGPT version 4.0. Accuracy may improve as newer models are made public.

This study was limited to only 9 different prompts. It is possible that further prompt optimization may improve accuracy. Nonetheless, of the 9 prompts, none displayed accuracy adequate to be used for patient care. As these prompts were simply a convenience sample from emergency physicians, the study lacks the robustness to make statistically valid conclusions about the influence of the prompts.

As the primary objective of this study was to assess the performance of the “out of the box” ChatGPT 4, defaults for parameters, such as Temperature and Top_P, were used. It is possible that manipulation of these parameters could improve performance. Furthermore, ChatGPT is a general-purpose LLM, that was not specifically fine-tuned for START. It is possible that fine-tuning of a LLM could lead to improved performance.

The reference standard for the simulated patients relied on expert opinion. Although the experts had extensive training in disaster medicine, in this study experts did not agree 100% on all patients. Thus, triage accuracy recorded in this study should be considered in this context. Nonetheless, it is important to remember that gage R and R studies do not rely on a reference standard. Since we can see that ChatGPT does not answer consistently when given the same question, we know that for certain it cannot always be correct. Furthermore, the simulation scenarios may not be representative of real-world data.

Significance

Clinical Implications

LLMs have been studied for many clinical applications including ophthalmic diagnosis, interpretation of laboratory results, decision support for cardiac diagnoses, and others [22-24]. However, while these models can be valuable, they do have

XSL•FO
RenderX

limits to their performance. This study illustrates several commonly encountered errors.

First, this study demonstrates that ChatGPT frequently returns incorrect answers. These are often called hallucinations or confabulations, where ChatGPT returns text that is syntactically correct, but false [25]. In this study, 99.7% of queries gave a response, but only 63.9% were correct. Notably, LLMs rarely indicate if the response is speculative or uncertain, but rather return a single response as if it were a truth. Clinicians should be cautious when using LLMs and be alert to the possibility of hallucinations.

In this study, overtriage (labeling casualties as more acute than the reference standard) was present in 32.9% while undertriage was much lower at 3.1%. An ideal triage system should guard against both errors. For the individual casualty, overtriage is much less dangerous, as it would result in care being given more urgently than needed. Conversely, overtriage, while of no danger to the individual casualty, leads to the risk of overwhelming the medical system.

Toxicity and stereotyping represent other prevalent errors, even though they were not directly examined in this study [25]. Clinicians need to be cognizant of the occurrence of these issues. Enhanced safeguards in recent LLMs are designed to mitigate the frequency of these errors.

LLMs undergo training using text, and some may use user-inputted text for additional training. It is imperative for practitioners to meticulously review the end user agreement of any LLM to understand privacy aspects. Caution should be exercised when providing private, privileged, or protected data.

Ultimately, LLMs do not oversee prompts for integrity or consistency. For example, when asked to triage a patient described as both ambulatory and unconscious, ChatGPT may not identify the contradiction in that a patient cannot be both. Often, ChatGPT assigns a triage code without recognizing such discrepancies.

Determining the attitudes and beliefs of staff is critical for the successful introduction of new technologies in medical institutions. Failure to establish new technologies in this setting is not always due to the nature of the system but rather because employees (clinicians) have insufficient knowledge about the acceptability, acceptance, and adoption process [26]. This reinforces the importance of value cocreation between technology and clinical leadership.

Finally, the use of LLM for disaster triage would require considerable research for the development of a useable point-of-care tool. Clearly expecting providers to type text, including prompt and patient data, would be unworkable and a more ergonomically acceptable interface would be needed.

Research Implications

Although this study suggests that the current ChatGPT version 4 does not adequately triage simulated patients using START, it does not preclude that AI may be a useful tool for triage. This study used the commercial version of ChatGPT, which was not fine-tuned for triage. A specifically fine-tuned model may be more accurate. Furthermore, other AI and machine-related technologies—such as k-nearest neighbors, support vector machines, or neural networks—may perform better.

While prompt dependency was clearly shown in this study, further, more structured research into the systematic features of effective and ineffective prompts is clearly indicated.

Finally, this study demonstrates that gage R and R (a statistically validated study design) is a practical method that researchers can use to evaluate the repeatability, reproducibility, and accuracy of AI-guided tools. AI solutions should be thoroughly studied before implementation into clinical practice—with the same attention to statistical rigor as one would use before introducing a new drug or medical device.

Conclusions

This study indicates that ChatGPT version 4 is insufficient to triage simulated disaster patients via the START protocol. It demonstrated suboptimal consistency, with its overall precision falling below 50%. The overall accuracy of triage was only 63.9%. Health care professionals are advised to exercise caution while using commercial LLMs for vital medical determinations, given that these tools may commonly produce inaccurate data, colloquially referred to as hallucinations or confabulations. AI-guided tools should undergo rigorous statistical evaluation—using methods such as gage R and R—before implementation into clinical settings.

Acknowledgments

The authors would like to thank Mrs Sandra Franchuk for her editorial assistance. The original draft of the manuscript was written solely by the authors. ChatGPT version 4 was used to assist with rephrasing, rewording, and restructuring of text. Artificial intelligence (AI) tools were not used to draft any of the tables or figures. No references were added to the paper based on AI suggestions. The authors attest that the manuscript is correct and remain accountable for all content.

Data Availability

The data sets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

All authors assisted with the conceptualization of the study. Data curation was performed by JMF, MV, and LC. Formal analysis and visualization were performed by JMF. The original draft was written by JMF. All authors reviewed and edited the final draft. All authors reviewed the final manuscript.

Conflicts of Interest

JMF is the CEO and founder of STAT59. MV has performed contract work as a consultant for STAT59.

Multimedia Appendix 1
Sample patient vignettes.
[PDF File (Adobe PDF File), 26 KB - jmir_v26i1e55648_app1.pdf]

Multimedia Appendix 2
Prompts.
[PDF File (Adobe PDF File), 39 KB - jmir_v26i1e55648_app2.pdf]

Multimedia Appendix 3
Python script.
[PDF File (Adobe PDF File), 37 KB - jmir_v26i1e55648_app3.pdf]

References

1. Wutz M, Hermes M, Winter V, Köberlein-Neu J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. J Med Internet Res 2023;25:e46548 [FREE Full text] [doi: 10.2196/46548] [Medline: 37751279]
2. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38. [doi: 10.1145/3571730]
3. McIntosh TR, Liu T, Susnjak T, Watters P, Ng A, Halgamuge MN. A culturally sensitive test to evaluate nuanced GPT hallucination. IEEE Trans Artif Intell 2023;5(6):1-13. [doi: 10.1109/tai.2023.3332837/mm1]
4. Heston T, Khun C. Prompt engineering in medical education. IME 2023;2(3):198-205. [doi: 10.3390/ime2030019]
5. Franc JM, Kirkland SW, Wisnesky UD, Campbell S, Rowe BH. METASTART: a systematic review and meta-analysis of the diagnostic accuracy of the simple triage and rapid treatment (START) algorithm for disaster triage. Prehosp Disaster Med 2022;37(1):106-116. [doi: 10.1017/S1049023X2100131X] [Medline: 34915954]
6. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
7. Gan RK, Uddin H, Gan AZ, Yew YY, González PA. ChatGPT's performance before and after teaching in mass casualty incident triage. Sci Rep 2023;13(1):20350 [FREE Full text] [doi: 10.1038/s41598-023-46986-0] [Medline: 37989755]
8. Ma X, Huo Y. Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. Technol Soc 2023;75:102362. [doi: 10.1016/j.techsoc.2023.102362]
9. Choudhary N. A study in measurement. Qual Prog 2017;50(9):42-47 [FREE Full text]
10. Gage repeatability and reproducibility. URL: https://asq.org/quality-resources/gage-repeatability [accessed 2023-09-26]
11. Ott E. Process Quality Control: Troubleshooting and Interpretation of Data. Milwaukee, Wisconsin: ASQ Quality Press; 2005.
12. Franc-Law JM, Bullard M, Della Corte F. Simulation of a hospital disaster plan: a virtual, live exercise. Prehosp Disaster Med 2008;23(4):346-353. [doi: 10.1017/s1049023x00005999] [Medline: 18935950]
13. McCarthy R, Gino B, Williams K, d'Entremont P, Renouf TS. Prehospital trauma care: a simulation scenario for rural-based healthcare providers. Cureus 2020;12(6):e8834 [FREE Full text] [doi: 10.7759/cureus.8834] [Medline: 32742845]
14. Alton J. The mass casualty incident: triage, part 3. 2012. URL: https://www.doomandbloom.net/the-mass-casualty-incident-triage-part-3/ [accessed 2023-09-20]
15. Tarantino C. Disaster triage train-the-trainer: START/JUMPSTART. URL: https://www.urmc.rochester.edu/MediaLibraries/URMCMedia/flrtc/START-Triage-Training-Presentation-TtT-8-31_1.pdf [accessed 2023-09-20]
16. Pediatric disaster triage training scenarios: utilizing the JumpSTART method. 2016. URL: https://www.luriechildrens.org/globalassets/documents/emsc/disaster/jumpstart-training-materials/jumpstarttrainingscenarios20164.pdf [accessed 2023-09-20]
17. Medical aspects in the management of a major incident / mass casualty situation. B ed. 2021: NATO standardization office. 2021. URL: https://www.coemed.org/files/stanags/03_AMEDP/AMedP-1.10_EDB_V1_E_2879.pdf [accessed 2024-08-31]
18. Agresti A. An Introduction to Categorical Data Analysis. Hoboken, New Jersey: John Wiley and Sons, Inc; 2007.

XSL•FO
RenderX

19. Rossum GV, Drake FLJ. Python Reference Manual. Amsterdam, Netherlands: Centrum voor Wiskunde en Informatica Amsterdam; 1995.
20. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of Chatbot responses to physician questions. JAMA Netw Open 2023;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
21. Franc JM, Cheng L, Hart A, Hata R, Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. CJEM 2024;26(1):40-46. [doi: 10.1007/s43678-023-00616-w] [Medline: 38206515]
22. Cadamuro J, Cabitza F, Debeljak Z, De Bruyne S, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) working group on artificial intelligence (WG-AI). Clin Chem Lab Med 2023;61(7):1158-1166 [FREE Full text] [doi: 10.1515/cclm-2023-0355] [Medline: 37083166]
23. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the isabel pro differential diagnosis generator. JFO Open Ophthalmol 2023;1:100005. [doi: 10.1016/j.jfop.2023.100005]
24. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). Acta Cardiol 2024;79(3):358-366. [doi: 10.1080/00015385.2024.2303528] [Medline: 38348835]
25. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
26. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021;3(9):e599-e611 [FREE Full text] [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]

Abbreviations

AI: artificial intelligence
API: application program interface
AV: appraiser variation
LLM: large language model
START: Simple Triage and Rapid Treatment

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 19.12.23; peer-reviewed by T Boyle, LH Yao, B Senst; comments to author 12.02.24; revised version received 22.02.24; accepted 19.06.24; published 30.09.24.

Please cite as:
Franc JM, Hertelendy AJ, Cheng L, Hata R, Verde M
Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage of Simulated Patients Using the Simple Triage and Rapid Treatment (START) Protocol: Gage Repeatability and Reproducibility Study
J Med Internet Res 2024;26:e55648
URL: https://www.jmir.org/2024/1/e55648
doi:10.2196/55648
PMID:39348189

©Jeffrey Micheal Franc, Attila Julius Hertelendy, Lenard Cheng, Ryan Hata, Manuela Verde. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ascle—A Python Natural Language Processing Toolkit for Medical Text Generation: Development and Evaluation Study

Rui Yang[1*], MS; Qingcheng Zeng[2*], BS; Keen You[3*], BS; Yujie Qiao[4*], MPH; Lucas Huang[3], BS; Chia-Chun Hsieh[3], BS; Benjamin Rosand[3], BS; Jeremy Goldwasser[3], BS; Amisha Dave[5], MD; Tiarnan Keenan[6], PhD; Yuhe Ke[7], MBBS; Chuan Hong[8], PhD; Nan Liu[1,9,10], PhD; Emily Chew[6], MD; Dragomir Radev[3†], PhD; Zhiyong Lu[11], PhD; Hua Xu[12], PhD; Qingyu Chen[12], PhD; Irene Li[13,14], PhD

[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore
[2]Department of Linguistics, Northwestern University, Evanston, IL, United States
[3]Department of Computer Science, Yale University, New Haven, CT, United States
[4]Yale School of Public Health, Yale University, New Haven, CT, United States
[5]Yale New Haven Hospital, Yale School of Medicine, Yale University, New Haven, CT, United States
[6]Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD, United States
[7]Department of Anesthesiology, Singapore General Hospital, Singapore, Singapore
[8]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States
[9]Program in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore
[10]Institute of Data Science, National University of Singapore, Singapore, Singapore
[11]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States
[12]Department of Biomedical Informatics and Data Science, Yale School of Medicine, Yale University, New Haven, CT, United States
[13]Information Technology Center, University of Tokyo, Kashiwa, Japan
[14]Smartor LLC, Tokyo, Japan
[†]deceased
[*]these authors contributed equally

Corresponding Author:
Irene Li, PhD
Information Technology Center
University of Tokyo
6-2-3 Kashiwanoha
Kashiwa, 277-8582
Japan
Phone: 81 09014707813
Email: ireneli@ds.itc.u-tokyo.ac.jp

Abstract

Background: Medical texts present significant domain-specific challenges, and manually curating these texts is a time-consuming and labor-intensive process. To address this, natural language processing (NLP) algorithms have been developed to automate text processing. In the biomedical field, various toolkits for text processing exist, which have greatly improved the efficiency of handling unstructured text. However, these existing toolkits tend to emphasize different perspectives, and none of them offer generation capabilities, leaving a significant gap in the current offerings.

Objective: This study aims to describe the development and preliminary evaluation of Ascle. Ascle is tailored for biomedical researchers and clinical staff with an easy-to-use, all-in-one solution that requires minimal programming expertise. For the first time, Ascle provides 4 advanced and challenging generative functions: question-answering, text summarization, text simplification, and machine translation. In addition, Ascle integrates 12 essential NLP functions, along with query and search capabilities for clinical databases.

Methods: We fine-tuned 32 domain-specific language models and evaluated them thoroughly on 27 established benchmarks. In addition, for the question-answering task, we developed a retrieval-augmented generation (RAG) framework for large language

XSL•FO
RenderX

models that incorporated a medical knowledge graph with ranking techniques to enhance the reliability of generated answers. Additionally, we conducted a physician validation to assess the quality of generated content beyond automated metrics.

Results: The fine-tuned models and RAG framework consistently enhanced text generation tasks. For example, the fine-tuned models improved the machine translation task by 20.27 in terms of BLEU score. In the question-answering task, the RAG framework raised the ROUGE-L score by 18% over the vanilla models. Physician validation of generated answers showed high scores for readability (4.95/5) and relevancy (4.43/5), with a lower score for accuracy (3.90/5) and completeness (3.31/5).

Conclusions: This study introduces the development and evaluation of Ascle, a user-friendly NLP toolkit designed for medical text generation. All code is publicly available through the Ascle GitHub repository. All fine-tuned language models can be accessed through Hugging Face.

(J Med Internet Res 2024;26:e60601) doi:10.2196/60601

KEYWORDS

natural language processing; machine learning; deep learning; generative artificial intelligence; large language models; retrieval-augmented generation; healthcare

Introduction

Medical texts pose considerable challenges due to their domain-specific nature, including issues such as ambiguities, frequent abbreviations, and specialized terminology [1,2]. The manual curation of these texts is both time-consuming and labor-intensive [2]. Therefore, natural language processing (NLP) algorithms have been developed to automate text processing [2-4]. Recent years have seen a notable shift toward the use of domain-specific pretrained language models, transitioning from shallow embeddings such as BioWordVec [5] and BioSentVec [6] to advanced architectures like Bidirectional Encoder Representations from Transformers (BERT) [7], such as BioBERT [8], ClinicalBERT [9], and PubMedBERT [10]. Furthermore, large language models (LLMs), such as Med-PaLM [11] and Med-Gemini [12], have demonstrated powerful generative capabilities, possessing exceptional zero- and few-shot performance. These domain-specific language models have substantially enhanced the effectiveness of NLP tasks in the biomedical and clinical domains [13-15].

Despite the success of these advanced methods, their complexity remains a significant barrier to practical application for health care professionals lacking basic programming skills. Consequently, there is an increasing demand for user-friendly and accessible toolkits designed to simplify medical text processing. Multiple toolkits for text processing are available in the biomedical domain. Table 1 summarizes representative tools. While there are many other useful tools, here we mainly limit our comparison with Python-based open-source toolkits.

Table 1. A comparison of Ascle with existing Python-based toolkits.

Toolkits	Question-Answering[a]	Text Summarization	Text Simplification	Machine Translation	Basic NLP[b] Functions	Query Search
MIMIC-Extract [16]	__[c]	—	—	—	—	✓
ScispaCy [17]	—	—	—	—	✓	—
MedspaCy [18]	—	—	—	—	✓	—
Transformers-sklearn [19]	—	—	—	—	✓	—
Stanza Biomed [20]	—	—	—	—	✓	—
Ascle (this study)	✓	✓	✓	✓	✓	✓

[a]For the question-answering task, we specifically propose a retrieval-augmented generation framework for large language models that incorporates a medical knowledge graph with ranking techniques.

[b]NLP: natural language processing. Basic natural language processing functions include abbreviation extraction, sentence tokenization, word tokenization, negation detection, hyponym detection, Unified Medical Language System concept extraction, named entity recognition, document clustering, part-of-speech tagging, entity linking, text summarization (extractive methods), and multiple-choice question-answering. It is worth noting that not every toolkit includes these 12 basic natural language processing functions, but Ascle includes them all.

[c]Not applicable.

These existing toolkits tend to emphasize different perspectives, and none of them offer generation capabilities, leaving a significant gap in the current offerings. In response, we present Ascle, a pioneering NLP toolkit for medical text generation, which, for the first time, includes 4 advanced generative functions. We fine-tuned 32 domain-specific language models and evaluated them thoroughly on 27 established benchmarks. In addition, for the question-answering task, we developed a retrieval-augmented generation (RAG) framework [21] that combines a medical knowledge graph (The Unified Medical Language System [UMLS]) [22] with ranking techniques, aimed at improving the reliability of long-form answers [15]. We uploaded all fine-tuned language models to Hugging Face and

XSL•FO
RenderX

listed 32 fine-tuned language models along with 27 benchmarks in Multimedia Appendix 1 for a clearer explanation.

In conclusion, Ascle empowers a diverse spectrum of users, from novices to experienced professionals, enabling them to effortlessly address their NLP tasks, even with limited technical expertise in handling textual data. We believe that Ascle not only democratizes access to cutting-edge methods but also expedites their integration into health care.

Methods

The Overall Architecture of Ascle

Ascle consists of 3 modules, with the core module being the "generative functions," including 4 challenging generation tasks: question-answering, text summarization, text simplification, and machine translation, covering a variety of application scenarios in health care. In addition, Ascle integrates 12 basic NLP functions, as well as query and search capabilities for clinical databases. The overall architecture of Ascle is shown in Figure 1. This section will focus on introducing the core module of Ascle—Generative Functions. For more information on basic NLP functions and query and search functions within Ascle, please refer to Multimedia Appendix 2 and Multimedia Appendix 3, respectively.

Figure 1. The overall architecture of Ascle. indicates that we have our fine-tuned language models for this task. indicates that we conducted evaluations for this task. POS: Parts-Of-Speech; QA: Question-Answering; UMLS: Unified Medical Language System.

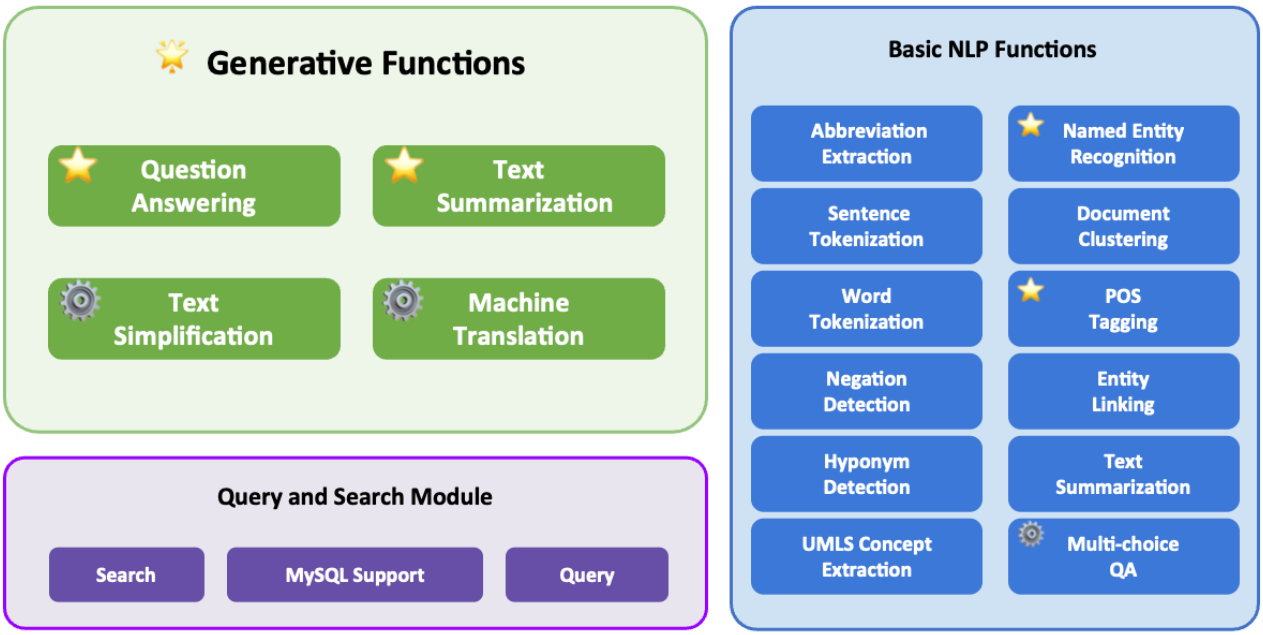

Generative Functions

Ascle offers a range of generative functions through pretrained and fine-tuned language models, all of which are publicly available for user access. In the following sections, we will introduce these powerful generative functions separately.

Question-Answering

Question-answering is particularly crucial in health care [13]. When integrated into health care systems, it assumes roles, such as preconsultation and remote consultation, effectively coping with the exponential increase in patient load. Furthermore, specialized question-answering systems hold the potential to contribute to medical education [13,21].

In Ascle, we first provide an interface for medical LLMs, such as Baize-healthcare [23], allowing users to use them directly. In addition, we develop a RAG framework that uses UMLS with ranking techniques to enhance LLMs in generating long-form answers [21]. Specifically, when receiving a query, the RAG framework first extracts medical entities within the query automatically and then retrieves related triplets from UMLS for each extracted entity. A triplet consists of 2 medical concepts and the relation between them, that is, (Myopia, clinically_associated_with, HYPERGLYCEMIA). Subsequently, the RAG framework uses ranking and reranking techniques to refine the ordering of these triples. Finally, the query and the retrieved triplets would be prompted to LLM for inference. For more details about the RAG framework, please refer to Multimedia Appendix 4. We apply this framework to the GPT (OpenAI) [24] and LLaMA (Meta) [25] series of LLMs.

We conducted evaluations on 4 medical QA data sets, including LiveQA [26], ExpertQA (Med & Bio) [27], and MedicationQA [28]. LiveQA consists of health questions submitted by consumers to the National Library of Medicine. It includes a training set with 634 QA pairs and a test set comprising 104 QA pairs, and the test set was used for evaluation. ExpertQA is a high-quality long-form QA data set covering multiple fields, along with answers verified by domain experts. Among them, we used 504 medical questions (Med) and 96 biology (Bio) questions for evaluation. MedicationQA includes 690

XSL•FO
RenderX

drug-related consumer questions along with information retrieved from reliable websites and scientific papers.

Additionally, considering that automated metrics cannot effectively assess the quality of generated content, especially in terms of factuality, we performed a physician validation. A total of 50 question-answer pairs from LiveQA were randomly selected, with answers generated by Baize-healthcare. Subsequently, 2 health care professionals (1 resident and 1 attending specialist) rated these generated answers on the criteria of readability, relevancy, accuracy, and completeness, using a 5-point Likert scale.

Text Summarization

In health care, clinicians and researchers are confronted with an increasing amount of information, including literature, clinical notes, and more [29,30]. Text summarization is an important generation task, aiming to distill essential information from the overwhelming complexity of texts and compress it into a more concise format [31]. Through automatic text summarization, clinicians and researchers can efficiently acquire information, thereby avoiding information overload.

We provide an abstractive text summarization function and compare general pretrained summarization models, including Pegasus [32], BigBird [33], Bidirectional and Auto-Regressive Transformer (BART) [34], PRIMERA [35], as well as domain-specific models, such as SciFive [36] and BioBART [37], which make use of biomedical corpora. Furthermore, we chose PubMed [38], MIMIC Chest X-Ray (MIMIC-CXR [39]), and MEDIQA-AnS [40] data sets for evaluation. The PubMed data set consists of biomedical scientific publications from the PubMed database, where each input document is a scientific article with its abstract serving as the ground truth. We reported the evaluation results for the test set, which contains 1660 examples. MIMIC-CXR is a deidentified data set of chest radiographs with free-text radiology reports, and we used a subset of MIMIC-CXR that includes 2000 instances for evaluation. MEDIQA-AnS is a collection of 156 consumer health questions along with passages that contain relevant information. It supports both single-document and multiple-document summarization evaluation.

Text Simplification

Biomedical texts are typically laden with intricate terminologies, which can hinder the understanding of individuals without a clinical background [41]. In Ascle, the function of text simplification is to translate complex and technical biomedical texts into understandable content. This will enhance the comprehension and involvement of nonclinical individuals, including patients, enabling them to better engage with the information and participate in clinical decisions more effectively.

We finetuned and evaluated widely used pretrained language models on 3 data sets: eLife, PLOS [42], and MedLane [43]. This included 2 general models, BigBirdPegasus [32] and BART, as well as a biomedical-specific model, BioBART. The eLife and PLOS are shared task data released from the BioLaySumm 2023 Task 1, which contains biomedical journal articles with expert-written lay summaries. We evaluated the validation sets for eLife and PLOS, which contain 241 and 1376 examples, respectively. MedLane is a large-scale human-annotated data set containing professional-to-customer sentences selected from Medical Information Mart for Intensive Care III (MIMIC-III). For MedLane, we used the test set for evaluation, which includes 1016 instances.

Machine Translation

Language barriers pose difficulties for patients to access timely information and communicate effectively with health care providers, resulting in low-quality health care services [44]. Our machine translation function aims to translate the text from a source language into a target language in a clinical scenario. By fine-tuning pretrained language models on the medical corpus, Ascle supports machine translation from English (en) to 8 target languages: Spanish (es), French (fr), Romanian (ro), Czech (cs), German (de), Hungarian (hu), Polish (pl), and Swedish (sv). Here, we only emphasize the 8 languages fine-tuned on medical data, while other languages, such as English to Chinese, are supported by the pretrained language models.

We fine-tuned the existing MarianMT [45] and multilingual T5 [46] using UFAL Medical Corpus [47] which includes various medical text sources, such as titles of medical Wikipedia articles, medical term pairs, patents, and documents from the European Medicines Agency. During the preprocessing phase, we excluded general domain data from UFAL, such as parliamentary proceedings, and randomly shuffled the medical domain corpora, splitting them into 2 parts at a ratio of 85% and 15% for training and testing, respectively. We reported the results on the test set, the size of which varies from 111,779 to 407,388 depending on the different language pairs. Furthermore, for each language pair, we used all available parallel data to maximize the breadth and accuracy of our machine translation function.

Ethical Considerations

This study used publicly available data sets and a restricted, deidentified data set. Access to the restricted data set was granted after the required training and certification, ensuring compliance with the data use agreement. No additional ethical review or informed consent was necessary, as human subjects or identifiable data were not directly involved. Two health care professionals voluntarily participated in the physician validation process without compensation. Data privacy and confidentiality were strictly maintained throughout the research, ensuring the protection of individual privacy while contributing to the advancement of the NLP toolkit for medical text generation.

Results

Overall Performance of Generation Tasks

In the question-answering task, we used ROUGE-L [48], BERTScore [49], MoverScore [50], and BLEURT [51] for a comprehensive evaluation, and used GPT-4 and LLaMA2-13b as the vanilla LLMs. As shown in Table 2, our RAG framework surpasses the zero-shot setting on all evaluation metrics for the LiveQA, ExpertQA-Bio, ExpertQA-Med, and MedicationQA

XSL•FO
RenderX

data sets. Among them, the ROUGE-L score has increased by more than 18% on the ExpertQA-Bio data set.

For the text summarization task, we evaluated 5 pretrained language models on single-document summarization, as shown in Table 3. To ensure a fair comparison, we excluded the results of BioBART and SciFive on PubMed, as they were fine-tuned on this data set. It is worth noting that BART consistently demonstrated strong performance across 3 benchmarks, while BioBART only outperformed BART in 1 of the benchmarks. In addition, we evaluated the multidocument summarization task and discussed the differences between abstractive and extractive methods, as well as the limitations of evaluation metrics, which can be found in the Discussion section.

Table 2. For the evaluation of the question-answering task, we compared ROUGE-L, BERTScore, MoverScore, and BLEURT on the zero-shot and retrieval-augmented generation framework.

	LLaMa2-13b				GPT-4			
	ROUGE-L	BERTScore	MoverScore	BLEURT	ROUGE -L	BERTScore	MoverScore	BLEURT
LiveQA[a]								
Z.S[b]	17.73	81.93	53.37	40.45	18.89	82.50	54.02	39.84
RAG[c]	18.83[d]	82.79[d]	53.79[d]	40.59[d]	19.44[d]	83.01[d]	54.11[d]	40.55[d]
ExpertQA[a]-Bio								
Z.S	23.26	84.38	55.58	44.65	23.00	84.50	56.15	44.53
RAG	25.79[d]	85.18[d]	56.17[d]	45.20[d]	27.20[d]	85.83[d]	57.11[d]	45.91[d]
ExpertQA[a]-Med								
Z.S	24.86	84.89	55.74	46.32	25.45	85.11	56.50	45.98
RAG	27.49[d]	85.80[d]	56.58[d]	46.47[d]	28.08[d]	86.30[d]	57.32[d]	47.00[d]
MedicationQA[a]								
Z.S	13.30	81.81	51.96	38.30	14.41	82.55	52.62	37.41
RAG	14.71[d]	82.79[d]	52.59[d]	38.42[d]	16.19[d]	83.59[d]	53.30[d]	37.91[d]

[a]QA: question-answering.

[b]Z.S: zero-shot.

[c]RAG: retrieval-augmented generation framework.

[d]The superior score within the same data set.

Table 3. For the evaluation of the single-document summarization task, we compared ROUGE-1, ROUGE-2, ROUGE-L, and some results are derived from other papers [52].

	PubMed			MIMIC-CXR[a]			MEDIQA-AnS (p)			MEDIQA-AnS (s)		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Pegasus	45.97	20.15	28.25	22.49	11.57	20.35	18.29	4.82	13.87	22.21	8.23	16.76
BigBird	46.32	20.65	42.33[b]	38.99	29.52	38.59	13.18	2.14	10.04	14.89	3.13	11.15
BART	48.35[b]	21.43[b]	36.90	41.70[b]	32.93[b]	41.16[b]	24.02[b]	7.20	17.09[b]	38.19	22.20	30.58
SciFive	__[c]	—	—	35.41	26.48	35.07	13.08	2.15	10.10	16.88	6.47	14.42
BioBART	—	—	—	41.61	32.90	41.00	22.58	7.49[b]	16.69	39.40[b]	24.64[b]	32.07[b]

[a]MIMIC-CXR: MIMIC Chest X-Ray database.

[b]The superior score within the same data set.

[c]Not applicable.

Regarding the text simplification task, we compared the performance of fine-tuned models and conducted an analysis of readability using the Flesch-Kincaid Grade Level (FKGL) score [53], as indicated in Table 4. For the eLife and PLOS data sets, the ground truth exhibits FKGL scores of 12 and 15, respectively. Interestingly, the BioBART model performs competitively in terms of ROUGE metrics, but fails to significantly reduce the difficulty of understanding, as evidenced by its FKGL score of 17 in both data sets. On the other hand, the BART model manages to slightly lower the FKGL score to 14 and 16 for eLife and PLOS, respectively. However, in the case of the MedLane data set, all methods appear to reach a

XSL•FO
RenderX

similar level of complexity as the ground truth. This can be attributed to the data set's shorter examples and potentially smaller vocabulary size, which limits the observed differences.

In the machine translation task, we fine-tuned the models across 8 languages, as illustrated in Table 5. After fine-tuning, the BLEU scores significantly improved, with the most substantial improvement observed in the "en-fr" language pair, increasing by over 61%. This enhancement can be attributed to the larger amount of training data available for "en-fr" (2,812,305 samples).

Table 4. For evaluation of the text simplification task, we compared ROUGE-1, ROUGE-2, ROUGE-L, and Flesch-Kincaid Grade Level score.

	eLife				PLOS				MedLane			
	R-1[a]	R-2[b]	R-L[c]	FKGL[d]	R-1	R-2	R-L	FKGL	R-1	R-2	R-L	FKGL
Ground Truth	—[e]	—	—	12	—	—	—	15	—	—	—	13
BigBirdPegasus	14.00	3.42	9.16	13[f]	18.92	4.79	12.54	17	74.96	65.37	74.56	13[f]
BART	16.16[f]	4.31[f]	10.19[f]	14	21.09	7.20	14.17	16[f]	83.25[f]	74.50[f]	82.99[f]	13[f]
BioBART	14.31	3.70	9.36	17	23.80[f]	7.83[f]	15.65[f]	17	82.89	74.26	82.65	13[f]

[a]R-1: ROUGE-1.
[b]R-2: ROUGE-2.
[c]R-L: ROUGE-L.
[d]FKGL: Flesch-Kincaid Grade Level.
[e]Not applicable.
[f]The superior score within the same data set.

Table 5. For the evaluation of the machine translation task, we evaluated the BLEU score on 8 language pairs.

	BLEU score							
	en-es[a]	en-fr[b]	en-ro[c]	en-cs[d]	en-de[e]	en-hu[f]	en-pl[g]	en-sv[h]
MarianMT	38.02	33.02	40.45	—[i]	—	—	—	—
F.T-MarianMT	41.64	43.72	43.88	—	—	—	—	—
F.T-mT5	45.88[j]	53.29[j]	47.28[j]	43.30	50.73	32.25	40.24	44.17

[a]en-es: English-Spanish.
[b]en-fr: English-French.
[c]en-ro: English-Romanian.
[d]en-cs: English-Czech.
[e]en-de: English-German.
[f]en-hu: English-Hungarian.
[g]en-pl: English-Polish.
[h]en-sv: English-Swedish.
[i]Not applicable.
[j]The superior score within the same data set.

Physician Validation

The results verified by the physicians are shown in Figure 2A. Detailed evaluation criteria can be found in Multimedia Appendix 5. The generated answers have good readability and relatively good relevancy, with scores of 4.95 and 4.43, respectively. In contrast, the completeness score is relatively lower (3.31). Figure 2B shows 2 cases. In the first case, compared with the ground truth, the generated answer does not point out that Zolmitriptan is used for treating acute migraines, nor does it indicate that it cannot be used to prevent migraine attacks or to reduce the frequency of headaches. In the second case, the generated answer does not mention that a gluten-free diet is the main treatment for celiac disease. We provide 2 additional cases in Multimedia Appendix 6.

In addition, we calculated the interevaluator agreement using percentage agreement for each criterion. A total of 2 health care professionals demonstrated a high level of consistency across all criteria, with the percentage agreement consistently exceeding 0.65.

XSL•FO
RenderX

Figure 2. (A) Physician validation (readability, relevancy, accuracy, and completeness) for 50 question-answer pairs. (B) Two examples of generated answers with ground truth.

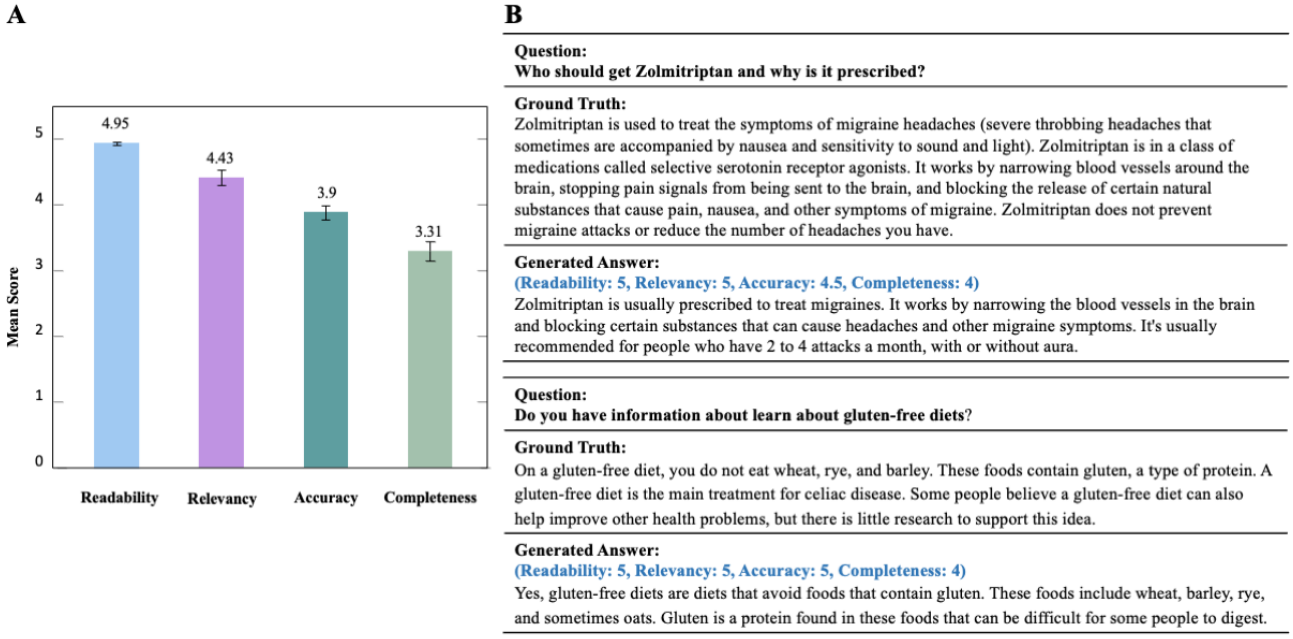

Discussion

In-Depth Analysis of the Text Summarization Task

In the multidocument summarization task, we included models based on traditional methods, such as TextRank [54], as well as pretrained language models, such as BART, Pegasus, PRIMERA, and BioBART. We evaluated their performance using ROUGE scores on the MEDIQA-AnS data set, which consists of 156 examples, and the results are shown in Table 6. However, it is noteworthy that although TextRank outperforms almost all generative models in ROUGE scores, this does not necessarily indicate superior performance. As ROUGE scores are calculated based on the overlap between the generated content and reference summaries, and TextRank is an extractive summarization model, it tends to score higher by this measure.

Table 6. Evaluation for the multidocument summarization task.

	MEDIQA-AnS (p)			MEDIQA-AnS (s)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
TextRank[a]	29.88	10.23	17.01	43.77	26.80	30.52
BART	24.56[b]	7.56[b]	17.18[b]	32.32[b]	15.42	24.03[b]
Pegasus	17.44	5.36	13.44	19.54	7.46	14.93
PRIMERA	16.66	4.89	12.68	21.78	9.77	16.85
BioBART	23.16	7.47	16.47	30.87	15.91[b]	23.66

[a]TextRank is only used as a reference for extractive summarization, so its scores are not compared with those of generative models.

[b]The superior score within the same data set.

While generative models possess semantic comprehension abilities, enabling them to distill complex information into an easy-to-understand format. As shown in Textbox 1, the summarizations generated by BART display well-structured patient information, with a brief description of events and corresponding conditions of the current patient (highlighted in blue), exhibiting high readability. In contrast, the summarizations produced by TextRank are less readable and include noise (highlighted in orange); the generated content is often a literal collage of text fragments. Despite TextRank achieving higher ROUGE scores, it lacks the ability to discern information and integrate it into coherent and readable content, showing significant limitations for practical use.

XSL•FO
RenderX

Textbox 1. Two MIMIC-III (parts) examples of the text summarization task, generated by BART and TextRank, respectively (we eliminated sensitive information).

BART

- The patient is an XXX-year-old man with a history of a question of coronary artery disease, borderline diabetes mellitus. He was in his usual state of health until 11 p.m. last night when he experienced chest pain with radiation to his back, positive shortness of breath, positive diaphoresis, no vomiting, no lightheadedness. The patient had had a similar episode of chest pain and was taken to an XXX. He had successful angioplasty and stent of LAD and CX. He is a middle aged XXX man in no acute hypertensive distress. He has had anginal chest pain, which is similar to his presenting complaint, but without radiations to his Back. His blood pressure was 105/73, pulse 84, respiratory 21, O2 saturation 92% on 2 liters. His CPK was 594, The index was 7.7, and he was admitted to the hospital with a high blood pressure. His condition was described as “stable” and “normal” by the doctor. The doctor referred the patient to a cardiologist for further treatment. The cardiologist said the patient was in good condition and should be discharged in a few days.
- Patient has CABG complicated by postop bleed and pleural effusion with discharge to [**Hospital1 **] Rehabilitation presents with abdominal pain. Zosyn was given in the ED. Patient was otherwise doing well and was to go back to rehab to finish his course of Cipro and Flagyl on [**5-17**]. Patient was last seen normal sometime last evening. He woke up and noticed that the left side of his body felt “numb”. He was not aware of any other

  neurological weakness, and mostly complained of being very tired. He denied any new vision problems, did not have a headache. He sounded somewhat slurred but did not feel as if his speech was changed significantly. He felt sleepy but able to sustain attention, currently apparently in no distress. He was on standing. Plavix and [**State **] which had been held for the last few days (at least since the 14), since he had the percutaneous drainage. He did not. feel that the weakness had progress and reported that he felt the sense of numbness was starting to improve and had some difficulty squeezing an examiners hand. He is a retired postal worker. He lives with wife and son who is a chiropractor.

TextRank

- Admission Date: XXX Discharge Date: XXX Date of Birth: Sex: M Service: CCU-6HISTORY OF PRESENT ILLNESS: The patient is a XXX-year-old man with a history of a question of coronary artery disease, status post myocardial infarction in [**December 2175**], hypertension, borderline diabetes mellitus who was in his usual state of health until 11 p.m. last night when, while [**4-12**] midsternal pressure like chest pain with radiation to back, positive shortness of breath, positive diaphoresis, positive nausea, no vomiting, no lightheadedness. Mucous membranes moist. Oropharynx clear. NECK: No jugular venous distention, no carotid bruits. CARDIOVASCULAR: Regular rate, S1, S2, artificial S1 gallop and balloon pump, no murmurs or rubs. LUNGS: Bibasilar rales, left greater than right. ABDOMEN: Normoactive bowel sounds, nontender, nondistended. EXTREMITIES: No cyanosis, clubbing or edema. NEUROLOGIC: Alert and oriented x3.LABS AT OUTSIDE HOSPITAL: CPK was 304, troponin 1.75.Electrocardiogram at 1:23 a.m. was normal sinus rhythm at101, normal axis deviation, 2 to [**Street Address(2) 1755**] elevation V1 to V5,Q V3, AVF.LABS AT [**Hospital6 **] AT 8 A.M.: CBC- white blood cells 11.2, hemoglobin 13.0, hematocrit 36.7,platelets 232. CARDIOVASCULAR: Coronary artery disease: Three vessel disease with successful intervention on LAD and left circumflex, but RCA not done secondary to good collateral. The patient was continued on aspirin 325 qd.
- Of note he was on standing Plavix and [**State **] which had been held for the last few days (atleast since the 14), since he had the percutaneous drainage. The patient was otherwise doing well and was to go back to rehab to finish his course of Cipro and Flagyl on [**5-17**].Past Medical History: coronary artery disease s/p right coronary artery stent x2([**10-3**], [**3-4**]), hypertension, hyperlipidemia, chronic obstructive pulmonary disease, asbestos exposure, chronic back pain, insomnia and obstructive sleep apnea (untreated)PSH:[**2144-4-21**]Endoscopic, minimally invasive, off pump coronary artery bypass graft x1 with left internal mammary artery to left anterior descending artery.[**2144-4-21**]Re-exploration for bleeding, post coronary artery bypass grafting. Social History: Lives with wife. Exposure to asbestos. Defers all medical decisions to son who is a chiropractor. Occupation: retired postal worker. Tobacco: 3 PPD x 30 years, quit 45 years ago ETOH: None Family History: Non-contributory to cholecystitis. Physical Exam: Physical Exam: Vitals: T: 97.9 P:75 R: 16 BP:128/73 SaO2:96 General: Awake, felt sleepy but able to sustain attention, poor historian currently.

System Usage

Ascle provides an easy-to-use approach for biomedical researchers and clinical staff. Users can efficiently use it by merely inputting text and calling the required functions. Figure 3 illustrates 2 use cases.

XSL•FO
RenderX

Figure 3. Demonstration of system usage. We show two use cases: Text Simplification and Machine Translation.

```
# create Ascle
from Ascle import Ascle
med = Ascle()

# Text Simplification
main_record = """
              The patient presents with symptoms of acute bronchitis,
              including cough, chest congestion, and mild fever.
              Auscultation reveals coarse breath sounds and occasional
              wheezing. Based on the clinical examination, a diagnosis
              of acute bronchitis is made, and the patient is prescribed
              a short course of bronchodilators and advised to rest and
              stay hydrated.
              """

# choose the model
layman_model = "ireneli1024/bart-large-elife-finetuned"

med.update_and_delete_main_record(content)

# call the text simplification function and print the output
print(med.get_layman_text(layman_model, min_length=20, max_length=70))

>> The patient presents with symptoms of acute bronchitis including
   cough, chest congestion and mild fever. Auscultation reveals coarse
   breath sounds and occasional wheezing. Based on these symptoms and
   the patient's history of previous infections with the same condition,
   the doctor decides that the patient is likely to have a cold or bronch.

# Machine Translation
main_record = """
              Myeloid derived suppressor cells (MDSC) are immature myeloid
              cells with immunosuppressive activity. They accumulate in
              tumor-bearing mice and humans with different types of cancer,
              including hepatocellular carcinoma (HCC).
              """

med.update_and_delete_main_record(record)

# call the machine translation function and print the output
print(med.get_translation_mt5("French"))

>> Les cellules suppressives dérivées de myéloïdes (MDSC) sont des
   cellules myéloïdes immatures ayant une activité immunosuppressive,
   accumulées chez des souris et des humains ayant différents types de
   cancer, y compris le carcinome hépatocellulaire (HCC).
```

Estimated Inference Time and Computational Resources

As shown in Table 7, we list the estimated inference time and computational resources required for the 4 generative tasks in Ascle. It is worth noting that the inference time is specific to our experimental settings, and the actual inference time for users may vary depending on the length of the input text and the computational resources used. For the question-answering task, GPT's response time is faster compared with LLaMA2-13b. However, it is important to mention that LLaMA2-13b was not deployed with quantization, and with quantization, the required inference time and computational resource requirements would be reduced.

XSL•FO
RenderX

Table 7. Estimated inference time and computational resources required for the generative tasks in Ascle.

Tasks	Estimated inference time	Computational resource
Question Answering	• LLaMA2-13b: <60 s/item[a] • GPT4: <15 s/item	• LLaMA2-13b: 4 * NVIDIA A100 GPU • GPT4: OpenAI API[b]
Text Summarization	<2 s/item	1 * NVIDIA V100 GPU
Text Simplification	<2 s/item	1 * NVIDIA V100 GPU
Machine translation	<2 s/item	1 * NVIDIA V100 GPU

[a]s/item: seconds per item.

[b]API: application programming interface.

Clinicians' Use of Ascle

To evaluate the ease of usability of Ascle for clinicians, we report the time required for 2 clinicians with different backgrounds to use the package after receiving guidance. The backgrounds of the clinicians are as follows: (1) physician 1: Singapore General Hospital, senior resident, 7 years of working experience, has a basic level of programming knowledge, and is able to perform basic statistical analyses; and (2) physician 2: SengKang General Hospital, senior consultant, 15 years of working experience, and has no programming knowledge.

Both clinicians received guidance on using Ascle, including setting up a virtual environment and accessing models from Hugging Face. The entire guidance process took about 10 minutes, after which both clinicians could independently and easily use Ascle and experiment with various generative functions without any issues. The main difficulty for the clinicians was setting up the virtual environment, as they lacked AI-specific knowledge. In response, Ascle provided a very simple virtual environment setup guideline. The clinicians' experience further confirms the user-friendliness of Ascle.

Limitations

In the case of generation tasks, we primarily chose automatic metrics for evaluation, such as ROUGE and BLEU scores. However, these metrics cannot effectively assess factual correctness [55] and may not align with human preference [56]. While human evaluation serves as an invaluable aspect in assessing the performance of the model, its incorporation may pose certain challenges due to various factors, including budget constraints.

Future Work

Recent LLMs have shown great potential in generative applications especially its superior zero- and few-shot performance [13,57,58]. Despite this, the generated content can be unfaithful, inconsistent, and biased [21,55,59,60]. We plan to thoroughly evaluate LLMs and extend to Ascle in the future. Meanwhile, we will strengthen the ethical review of these generative AI techniques to ensure their application truly and responsibly benefits biomedical researchers and health care professionals [61,62].

Conclusions

We introduce Ascle, a comprehensive NLP toolkit designed specifically for medical text generation. For the first time, it integrates 4 challenging generative functions, including question-answering, text summarization, text simplification, and machine translation. Our research fills the gap of existing toolkits for generative tasks, which holds significant implications for the entire medical domain. Ascle boasts remarkable flexibility, allowing users to access a variety of cutting-edge pretrained language models. Meanwhile, it stands as a user-friendly toolkit, ensuring ease of use even for clinical staff without a technical background. We will continue to maintain and extend Ascle.

Acknowledgments

TK and EC were supported by the National Institutes of Health (NIH) Intramural Research Program (IRP), National Eye Institute. ZL and QC were supported by the NIH IRP, National Library of Medicine. QC was also supported by the National Library of Medicine of the NIH (award 1K99LM014024). IL was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant 24K20832). RY and IL received support from GCP Credit Award by Google.

Authors' Contributions

RY, QZ, KY, YQ, LH, CCH, BR, JG, YHK, and IL performed the data collection, data processing, and experiments. AD, TK conducted manual reviews. RY, QC, and IL created the figures and tables and drafted the manuscript. CH, NL, EC, DR, ZL, HX, QC, and IL were responsible for project administration. All authors conceived the idea for the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

32 Fine-tuned language models and 27 benchmarks in Ascle.
[DOCX File , 20 KB - jmir_v26i1e60601_app1.docx]

Multimedia Appendix 2
Basic natural language processing functions in Ascle.
[DOCX File , 192 KB - jmir_v26i1e60601_app2.docx]

Multimedia Appendix 3
Query and search module in Ascle.
[DOCX File , 14 KB - jmir_v26i1e60601_app3.docx]

Multimedia Appendix 4
The Retrieval-Augmented Generation framework in Ascle – KG-Rank.
[DOCX File , 189 KB - jmir_v26i1e60601_app4.docx]

Multimedia Appendix 5
Evaluation criteria for physician validation.
[DOCX File , 16 KB - jmir_v26i1e60601_app5.docx]

Multimedia Appendix 6
Case study.
[DOCX File , 15 KB - jmir_v26i1e60601_app6.docx]

References

1. Li I, Yasunaga M, Nuzumlalı MY, Caraballo C, Mahajan S, Krumholz H, et al. A neural topic-attention model for medical term abbreviation disambiguation. ArXiv. Preprint posted online on October 30, 2019 2019 [FREE Full text] [doi: 10.5260/chara.21.2.8]
2. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural natural language processing for unstructured data in electronic health records: a review. Computer Science Review 2022;46:100511. [doi: 10.1016/j.cosrev.2022.100511]
3. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
4. al-Aiad A, Duwairi R, Fraihat M. Survey: deep learning concepts and techniques for electronic health record. : IEEE; 2018 Presented at: IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA); 2018 October 28 - 2018 November 01; Aqaba, Jordan p. 1-5. [doi: 10.1109/aiccsa.2018.8612827]
5. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
6. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. : IEEE; 2019 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China p. 1-5. [doi: 10.1109/ichi.2019.8904728]
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on May 24, 2019 2019:1-16 [FREE Full text] [doi: 10.5260/chara.21.2.8]
8. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
9. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Association for Computational Linguistics. 2019 Presented at: the 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]
10. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare 2021;3(1):1-23. [doi: 10.1145/3458754]
11. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. ArXiv. Preprint posted online on May 16, 2023 2023. [doi: 10.48550/arXiv.2305.09617]
12. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of gemini models in medicine. ArXiv. Preprint posted online on May 1, 2024 2024. [doi: 10.48550/arXiv.2404.18416]
13. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61] [Medline: 38939520]

XSL•FO
RenderX

14. Li C, Mowery DL, Ma X, Yang R, Vurgun U, Hwang S, et al. Realizing the potential of social determinants data: a scoping review of approaches for screening, linkage, extraction, analysis and interventions. medRxiv. Preprint posted online on February 6, 2024 2024 [FREE Full text] [doi: 10.1101/2024.02.04.24302242] [Medline: 38370703]
15. Yang R, Ning Y, Keppo E, Liu M, Hong C, Bitterman D, et al. Retrieval-augmented generation for generative artificial intelligence in medicine. ArXiv. Preprint posted online on June 18, 2024 2024. [doi: 10.48550/ARXIV.2406.12449]
16. Wang S, McDermott MBA, Chauhan G, Ghassemi M, Hughes MC, Naumann T. MIMIC-Extract. USA: ACM; 2020 Presented at: CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning; April 2-4, 2020; New York, USA p. 222-235. [doi: 10.1145/3368555.3384469]
17. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural language processing. USA: Association for Computational Linguistics; 2019 Presented at: Proceedings of the 18th BioNLP Workshop and Shared Task Stroudsburg; August 1, 2019; Florence, Italy p. 319-327. [doi: 10.18653/v1/w19-5034]
18. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc 2022;2021:438-447 [FREE Full text] [Medline: 35308962]
19. Yang F, Wang X, Ma H, Li J. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. BMC Med Inform Decis Mak 2021;21(Suppl 2):90 [FREE Full text] [doi: 10.1186/s12911-021-01459-0] [Medline: 34330244]
20. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the stanza python NLP library. J Am Med Inform Assoc 2021;28(9):1892-1899 [FREE Full text] [doi: 10.1093/jamia/ocab090] [Medline: 34157094]
21. Yang R, Liu H, Marrese-Taylor E, Zeng Q, Ke YH, Li W, et al. KG-Rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. ArXiv. Preprint posted online on July 04, 2024 2024. [doi: 10.48550/ARXIV.2403.05881]
22. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
23. Xu C, Guo D, Duan N, McAuley J. Baize: an open-source chat model with parameter-efficient tuning on self-chat data. : Association for Computational Linguistics; 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; December 2023; Singapore p. 6268-6278. [doi: 10.18653/v1/2023.emnlp-main.385]
24. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. ArXiv. Preprint posted online on March 04, 2024 2024. [doi: 10.48550/ARXIV.2303.08774]
25. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. ArXiv. Preprint posted online on July 19, 2023 2024. [doi: 10.48550/ARXIV.2307.09288]
26. Abacha AB, Agichtein E, Pinter Y, Demner-Fushman D. Overview of the medical question answering task at TREC 2017 LiveQA. 2018 Presented at: Text REtrieval Conference (TREC); May 2017; Gaithersburg, MD, USA URL: https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf?ref=https://githubhelp.com
27. Malaviya C, Lee S, Chen S, Sieber E, Yatskar M, Roth D. ExpertQA: expert-curated questions and attributed answers. ArXiv. Preprint posted online on April 02, 2024 2024 [FREE Full text] [doi: 10.48550/arXiv.2309.07852]
28. Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the gap between consumers' medication questions and trusted answers. Stud Health Technol Inform 2019;264:25-29. [doi: 10.3233/SHTI190176] [Medline: 31437878]
29. Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, et al. Text summarization in the biomedical domain: a systematic review of recent research. J Biomed Inform 2014;52:457-467 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.009] [Medline: 25016293]
30. Ke Y, Yang R, Liu N. Comparing open-access database and traditional intensive care studies using machine learning: bibliometric analysis study. J Med Internet Res 2024;26:e48330 [FREE Full text] [doi: 10.2196/48330] [Medline: 38630522]
31. Xie Q, Luo Z, Wang B, Ananiadou S. A survey for biomedical text summarization: from pre-trained to large language models. ArXiv. Preprint posted online on July 13, 2023 2024. [doi: 10.48550/ARXIV.2304.08763]
32. Zhang J, Zhao Y, Saleh M, Liu PJ. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. ArXiv. Preprint posted online on July 10, 2020 2020. [doi: 10.48550/ARXIV.1912.08777]
33. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, et al. Big bird: transformers for longer sequences. ArXiv. Preprint posted online on January 08, 2021 2021. [doi: 10.48550/ARXIV.2007.14062]
34. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. : Association for Computational Linguistics; 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; July 2020; Online p. 7871-7880. [doi: 10.18653/v1/2020.acl-main.703]
35. Xiao W, Beltagy I, Carenini G, Cohan A. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. : Association for Computational Linguistics; 2022 Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; May 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.360]
36. Phan LN, Anibal JT, Tran H, Chanana S, Bahadroglu E, Peltekian A, et al. SciFive: a text-to-text transformer model for biomedical literature. ArXiv. Preprint posted online on May 28, 2021 2021. [doi: 10.48550/ARXIV.2106.03598]

XSL•FO
RenderX

37. Yuan H, Yuan Z, Gan R, Zhang J, Xie Y, Yu S. BioBART: pretraining and evaluation of a biomedical generative language model. BioBART: Pretraining and evaluation of A biomedical generative language model. arXiv; 2022 Presented at: Proceedings of the 21st Workshop on Biomedical Language Processing; May 2022; Dublin, Ireland p. 97-109. [doi: 10.18653/v1/2022.bionlp-1.9]
38. Cohan A, Dernoncourt F, Kim D, Bui T, Kim S, Chang W, et al. A discourse-aware attention model for abstractive summarization of long documents. 2018 Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2018; New Orleans, Louisiana p. 615-621. [doi: 10.18653/v1/n18-2097]
39. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019 1;6(1):317 [FREE Full text] [doi: 10.1038/s41597-019-0322-0] [Medline: 31831740]
40. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Sci Data 2020;7(1):322 [FREE Full text] [doi: 10.1038/s41597-020-00667-z] [Medline: 33009402]
41. Devaraj A, Wallace BC, Marshall IJ, Li JJ. Paragraph-level Simplification of medical texts. Proc Conf 2021;2021:4972-4984 [FREE Full text] [doi: 10.18653/v1/2021.naacl-main.395] [Medline: 35663507]
42. Goldsack T, Zhang Z, Lin C, Scarton C. Making science simple: corpora for the lay summarisation of scientific literature. : Association for Computational Linguistics; 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; December 2022; Abu Dhabi, United Arab Emirates p. 10589-10604. [doi: 10.18653/v1/2022.emnlp-main.724]
43. Luo J, Zheng Z, Ye H, Ye M, Wang Y, You Q, et al. Benchmarking automated clinical language simplification: dataset, algorithm, and evaluation. ArXiv. Preprint posted online on September 21, 2023 2024. [doi: 10.48550/ARXIV.2012.02420]
44. Khoong EC, Rodriguez JA. A research agenda for using machine translation in clinical medicine. J Gen Intern Med 2022;37(5):1275-1277 [FREE Full text] [doi: 10.1007/s11606-021-07164-y] [Medline: 35132559]
45. Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, et al. Marian: fast neural machine translation in C++. Association for Computational Linguistics; 2018 Presented at: Proceedings of ACL 2018, System Demonstrations; July 2018; Melbourne, Australia p. 116-121. [doi: 10.18653/v1/p18-4020]
46. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, et al. mT5: A massively multilingual pre-trained text-to-text transformer. : Association for Computational Linguistics; 2021 Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2021; Online p. 483-498. [doi: 10.18653/v1/2021.naacl-main.41]
47. UFAL Medical Corpus v. 1.0.: UFAL MediCorp URL: https://ufal.mff.cuni.cz/ufal_medical_corpus [accessed 2024-08-06]
48. Lin CY. ROUGE: a package for automatic evaluation of summaries. : Association for Computational Linguistics; 2004 Presented at: Text Summarization Branches Out; July 2004; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013.pdf
49. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. ArXiv. Preprint posted online on February 24, 2020 2020. [doi: 10.5260/chara.21.2.8]
50. Zhao W, Peyrard M, Liu F, Gao Y, Meyer C, Eger S. MoverScore: text generation evaluating with contextualized embeddings and earth mover distance. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 2019; Hong Kong, China p. 563-578. [doi: 10.18653/v1/d19-1053]
51. Sellam T, Das D, Parikh AP. BLEURT: learning robust metrics for text generation. ArXiv. Preprint posted online on May 21, 2020 2020. [doi: 10.48550/ARXIV.2004.04696]
52. Rohde T, Wu X, Liu Y. Hierarchical learning for generation with long source sequences. ArXiv. Preprint posted online on September 16, 2021 2021. [doi: 10.5260/chara.21.2.8]
53. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation Of new readability formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) for navy enlisted personnel. Institute for Simulation and Training 1975;56 [FREE Full text] [doi: 10.21236/ada006655]
54. Mihalcea R, Tarau P. TextRank: Bringing Order into Text. : Association for Computational Linguistics; 2004 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; August 01, 2004; Barcelona, Spain p. 404-411 URL: https://aclanthology.org/W04-3252.pdf
55. Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in medicine: a systematic review with large language models and beyond. medRxiv. Preprint posted online on July 01, 2023 2023 [FREE Full text] [doi: 10.1101/2023.04.18.23288752] [Medline: 37398329]
56. Fleming SL, Lozano A, Haberkorn WJ, Jindal J, Reis E, Thapa R, et al. MedAlign: a clinician-generated dataset for instruction following with electronic medical records. 2024 Presented at: Proceedings of the 38th AAAI Conference on Artificial Intelligence; February 20-27, 2024; Vancouver, Canada p. 22021-22030. [doi: 10.1609/aaai.v38i20.30205]
57. Gao F, Jiang H, Yang R, Zeng Q, Lu J, Blum M, et al. Large Language models on wikipedia-style survey generation: an evaluation in NLP concepts. ArXiv. Preprint posted online on May 23, 2024 2024. [doi: 10.48550/ARXIV.2308.10410]

XSL•FO
RenderX

58. Yang R, Yang B, Ouyang S, She T, Feng A, Jiang Y, et al. Leveraging large language models for concept graph recovery and question answering in NLP education. ArXiv. Preprint posted online on February 22, 2024 2024. [doi: 10.48550/ARXIV.2402.14293]
59. Ke YH, Yang R, Lie SA, Lim TXY, Abdullah HR, Ting DSW, et al. Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias. ArXiv. Preprint posted online on May 12, 2024 2024. [doi: 10.48550/ARXIV.2401.14589]
60. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023;25(1):bbad493 [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
61. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence in healthcare: ethical considerations and assessment checklist. ArXiv. Preprint posted online on February 23, 2024 2024. [doi: 10.48550/ARXIV.2311.02107]
62. Yang R, Nair SV, Ke Y, D'Agostino D, Liu M, Ning Y, et al. Disparities in clinical studies of AI-enabled applications from a global perspective. NPJ Digit Med 2024;7:209 [FREE Full text]

Abbreviations

BART: Bidirectional and Auto-Regressive Transformer
BERT: Bidirectional Encoder Representations from Transformers
FKGL: Flesch-Kincaid Grade Level
LLM: large language model
NLP: natural language processing
RAG: Retrieval-Augmented Generation
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 16.05.24; peer-reviewed by C Wu, T Hanne; comments to author 13.06.24; revised version received 08.07.24; accepted 15.07.24; published 03.10.24.

Please cite as:
Yang R, Zeng Q, You K, Qiao Y, Huang L, Hsieh CC, Rosand B, Goldwasser J, Dave A, Keenan T, Ke Y, Hong C, Liu N, Chew E, Radev D, Lu Z, Xu H, Chen Q, Li I
Ascle—A Python Natural Language Processing Toolkit for Medical Text Generation: Development and Evaluation Study
J Med Internet Res 2024;26:e60601
URL: https://www.jmir.org/2024/1/e60601
doi:10.2196/60601
PMID:

©Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha Dave, Tiarnan Keenan, Yuhe Ke, Chuan Hong, Nan Liu, Emily Chew, Dragomir Radev, Zhiyong Lu, Hua Xu, Qingyu Chen, Irene Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study

Jonathan Yi-Shin Yau[1]; Soheil Saadat[2], MD, MPH, PhD; Edmund Hsu[2], MD, MSIDT, AEMUS-FPD; Linda Suk-Ling Murphy[3], MLIS; Jennifer S Roh[4], MD, MBA; Jeffrey Suchard[2], MD; Antonio Tapia[2], MD; Warren Wiechmann[2], MD, MBA, MSED; Mark I Langdorf[2], MD, MHPE, MAAEM

[1]College of Natural and Agricultural Sciences, University of California - Riverside, Riverside, CA, United States

[2]Department of Emergency Medicine, University of California - Irvine, Orange, CA, United States

[3]Reference Department, University of California - Irvine Libraries, Irvine, CA, United States

[4]Department of Emergency Medicine, Harbor-UCLA Medical Center, University of California - Los Angeles, Torrance, CA, United States

Corresponding Author:
Mark I Langdorf, MD, MHPE, MAAEM
Department of Emergency Medicine
University of California - Irvine
101 the City Drive, Route 128-01
Orange, CA, 92868
United States
Phone: 1 7147452663
Email: milangdo@hs.uci.edu

Abstract

Background: Recent surveys indicate that 48% of consumers actively use generative artificial intelligence (AI) for health-related inquiries. Despite widespread adoption and the potential to improve health care access, scant research examines the performance of AI chatbot responses regarding emergency care advice.

Objective: We assessed the quality of AI chatbot responses to common emergency care questions. We sought to determine qualitative differences in responses from 4 free-access AI chatbots, for 10 different serious and benign emergency conditions.

Methods: We created 10 emergency care questions that we fed into the free-access versions of ChatGPT 3.5 (OpenAI), Google Bard, Bing AI Chat (Microsoft), and Claude AI (Anthropic) on November 26, 2023. Each response was graded by 5 board-certified emergency medicine (EM) faculty for 8 domains of percentage accuracy, presence of dangerous information, factual accuracy, clarity, completeness, understandability, source reliability, and source relevancy. We determined the correct, complete response to the 10 questions from reputable and scholarly emergency medical references. These were compiled by an EM resident physician. For the readability of the chatbot responses, we used the Flesch-Kincaid Grade Level of each response from readability statistics embedded in Microsoft Word. Differences between chatbots were determined by the chi-square test.

Results: Each of the 4 chatbots' responses to the 10 clinical questions were scored across 8 domains by 5 EM faculty, for 400 assessments for each chatbot. Together, the 4 chatbots had the best performance in clarity and understandability (both 85%), intermediate performance in accuracy and completeness (both 50%), and poor performance (10%) for source relevance and reliability (mostly unreported). Chatbots contained dangerous information in 5% to 35% of responses, with no statistical difference between chatbots on this metric (P=.24). ChatGPT, Google Bard, and Claud AI had similar performances across 6 out of 8 domains. Only Bing AI performed better with more identified or relevant sources (40%; the others had 0%-10%). Flesch-Kincaid Reading level was 7.7-8.9 grade for all chatbots, except ChatGPT at 10.8, which were all too advanced for average emergency patients. Responses included both dangerous (eg, starting cardiopulmonary resuscitation with no pulse check) and generally inappropriate advice (eg, loosening the collar to improve breathing without evidence of airway compromise).

Conclusions: AI chatbots, though ubiquitous, have significant deficiencies in EM patient advice, despite relatively consistent performance. Information for when to seek urgent or emergent care is frequently incomplete and inaccurate, and patients may be unaware of misinformation. Sources are not generally provided. Patients who use AI to guide health care decisions assume

XSL•FO
RenderX

potential risks. AI chatbots for health should be subject to further research, refinement, and regulation. We strongly recommend proper medical consultation to prevent potential adverse outcomes.

(J Med Internet Res 2024;26:e60291) doi:10.2196/60291

KEYWORDS

artificial intelligence; AI; chatbots; generative AI; natural language processing; consumer health information; patient education; literacy; emergency care information; chatbot; misinformation; health care; medical consultation

Introduction

There has been a significant surge in attention to artificial intelligence (AI)–driven chatbots capable of generating content and engaging in conversations resembling natural human communication. While using a smartphone or a computer, most people have interacted with a chatbot—software or computer program that simulates human conversation [1,2]. Modern chatbots use generative AI (GenAI) technologies such as machine learning, large language models (LLMs), natural language processing, and deep learning to more accurately interpret user questions and interact in a natural way. Furthermore, GenAI recognizes, predicts, and can create content without human interaction [3,4].

Since their release to the general public, chatbots such as ChatGPT, Google Bard (now known as Gemini) [5], Bing AI Chat (now known as Copilot) [6], and Claud AI have penetrated diverse industries, including technology, education, research, health care, finance, and social media. In health care, LLMs have demonstrated the potential to revolutionize the field, with ChatGPT able to achieve a passing score on the US Medical Licensing Examination and write basic medical reports [7-9]. Future applications of LLMs are vast, including assisting with medical education, predicting diagnoses, and recommending treatment options to patients [10].

Consumers have turned to the web as their first source of information for many medical-related inquiries. The National Center for Health Statistics (NCHS) data brief published by the National Center for Health Statistics found that 58.5% of US adults used the internet to look for health or medical information in 2022. The main motivations were convenience and breadth of information, which were less available through traditional doctor visits and access to interpretation of medical test results [11]. A more recent 2024 survey by Deloitte found that nearly half (48%) of consumer respondents have used GenAI for health-related concerns, citing its potential to make health care more accessible, affordable, and reliable [12]. Finally, in 2023, a survey reported that 78.4% of subjects would be willing to use Chat-GPT for self-diagnosis [13]. This growing reliance on GenAI highlights the importance of further research on whether current LLMs provide accurate, relevant answers to medical questions across specialties.

Access to medical information is a social determinant of health [14,15]. GenAI, therefore, has the potential to both mitigate or exacerbate health disparities, depending on its accuracy and completeness. While younger patients who are facile with English and technology may be better able to access web-based health information, older patients or those whose primary language is not English are more at risk for reduced access or response comprehension. Furthermore, patients with more limited access to physicians due to insurance barriers could be more apt to use GenAI chatbots for medical queries, adding to their disadvantages in accessing proper health care.

Since the release of ChatGPT 3.5 on November 30, 2022, there has been a substantial increase in medical literature related GenAI in health care. Nearly 4000 publications were found in PubMed between November 30, 2022, and May 1, 2024. A 2023 study analyzing ChatGPT responses to questions across 17 different specialties found ChatGPT to be extremely promising, with the chatbot scoring high in median accuracy and completeness. However, the researchers noted that the chatbot would sometimes be “spectacularly and surprisingly wrong” [16]. Another study examined ChatGPT’s ability to answer bariatric surgery-related questions and found the responses to be mostly comprehensive and reproducible [17]. Similar studies seem to report that ChatGPT produces largely accurate information. However, ChatGPT’s inability to cite up-to-date, reputable sources in its responses may result in sporadic-to-frequent occurrences of blatant misinformation [18].

This paper expands the literature on chatbot reliability in 3 ways. First, we used board-certified emergency medicine (EM) faculty physicians as assessors of the accuracy, safety, completeness, readability, and reliability of emergency care information. Second, we asked questions of the chatbots about common potentially emergent conditions, which, if inaccurate, could exacerbate negative social determinants of health. Wrong or missing information could be dangerous. Third, we judged the responses against the reading or language capabilities of average emergency patients. Studies suggest most emergency patients read below or at the eighth-grade level [19,20]. The American Medical Association (AMA) recommends a maximum fifth-grade reading level for health literacy materials directed toward Medicaid enrollees [21].

Additionally, we assessed the performance of less commonly used consumer chatbots, such as Google Bard, Bing AI, and Claude AI, which have not been as rigorously assessed as those of ChatGPT.

Our objective was to comprehensively evaluate and compare the strengths and weaknesses of answers from ChatGPT, Google Bard, Bing AI, and Claude AI to common emergency care questions, as judged by board-certified EM faculty physicians, who provide high-level expertise and clinical relevance in the evaluation process.

XSL•FO
RenderX

Methods

Overview

The research prompts were 10 common emergency care questions by patients, selected by 5 board-certified EM faculty (EH, MIL, JSR, JS, and WW) from common emergency department (ED) chief complaints. These were representative of both benign and potentially serious conditions [22]. The 10 question topics were chest pain, stroke, bad headache, bad sore throat, bad stomach pain, bad back pain, fainting, heavy menstrual bleeding, bad cold, and drug overdose.

We included these based on a reference [22] that described the most common chief complaints for EDs. These included abdominal pain, shortness of breath, chest pain, and 4 neurological complaints, which we included with our choice of back pain, stroke, fainting or syncope, and headache. Other chief complaints we chose to test (heavy period bleeding and sore throat) were found in a lower proportion of patients and are listed in a footnote. "Common cold" is not specifically listed as a common chief complaint in this reference, inexplicably, but could be subsumed in shortness of breath and dysphagia or sore throat.

The wording of each question was then refined to reflect the language typically used by patients, with emphasis on what actions the inquirer should take. The finalized prompts are listed in Multimedia Appendix 1.

To minimize bias, the chatbots were accessed through a newly created email account on a browser with cleared cookies and cache. For each prompt, a new chat was initiated, and the chatbot's first response or draft was documented (Multimedia Appendix 2). If the chatbot did not list sources in the initial response, a follow-up query was entered—"Please list all sources of information you referenced." All responses were generated on November 26 and 27, 2023, using publicly available free versions of the chatbots at that time. We chose free versions, as, at the moment of a patient wondering what to do about a clinical problem, they would likely query the free version of a chatbot, rather than sign up for access to a more sophisticated, fee-based version.

As a benchmark for grading the chatbot-generated responses, "correct" answers to the 10 questions were developed (Multimedia Appendix 3) by an emergency medical resident (AT) citing trusted medical sources—"Patient Education - UpToDate," MedlinePlus Health Topics, and Mayo Clinic (full references listed in Appendix 3). A total of 5 board-certified emergency faculty (EH, MIL, JSR, JS, and WW) validated the completeness and accuracy of each response to these 10 questions from the 4 chatbots.

We assessed the reliability of chatbots for emergency medical advice with a standardized grading rubric to mitigate evaluator bias and enhance objectivity. While the study by Altamimi et al [23] evaluated ChatGPT by clinical toxicologists and emergency physicians for providing recommendations about venomous snakebites, the lack of a defined grading scale raised concern regarding potential bias in their evaluation process [23].

Given the absence of an accepted standard to evaluate chatbot-generated medical advice [24-29], we developed our own comprehensive scoring sheet. Drawing inspiration from the Academic Life in Emergency Medicine Approved Instructional Resources rating score and evaluative scales from prior chatbot studies [17,30-34], our scoring sheet facilitates straightforward documentation of evaluations via a Google Form. It rigorously assesses each response based on 8 criteria, which we called "domains"—accuracy, safety, factual accuracy, clarity, completeness, readability, source availability, and source reliability (Multimedia Appendix 4). We defined a qualitative and quantitative measure of accuracy. The first was a percentage of correct information by quartile, dubbed "accuracy," and the second, "factually accurate," was whether the responses contained none, 1, 2, or more than 2 minor, major, or minor inaccuracies.

EM faculty identified significant omissions from the chatbot responses as 3 types. The first type was frank omissions of specific required advice, such as no mention of naloxone for opioid overdoses, no mention that antibiotics do not help most sore throats or common colds, and no mention of aspirin for chest pain. The second type was global missing concepts, such as what specifically should prompt medical attention (so-called "red flag" symptoms) and how long one should wait before deciding to seek medical attention. The third type was complete neglect of life-threatening diagnoses that should be considered with the presenting complaint (eg, no mention of pregnancy, either ectopic or intrauterine, for heavy menstrual bleeding). Some of these incomplete elements were judged to overlap with reported dangerous conditions.

We calculated the Flesch-Kincaid Grade Level (FKGL) of each of the 40 chatbot responses using the embedded FKGL tool within Microsoft Word for Microsoft 365 MSO (version 2402 build 16.0.17328.20124).

Reviewers' ratings, on a 5-point ordinal scale, were dichotomized into "highest/best score" (ie, score 5) versus "less than perfect score" (ie, scores 1-4). The relative frequency of the "highest/best score" category was calculated and compared across the 4 chatbots, 10 medical conditions, and 8 domains of interest. The chi-square test was used to compare the relative frequency of "highest/best score" among the variables of interest (chatbots, medical conditions, and domains). For each comparison, P values were estimated using the asymptotic 2-way approximation. When the assumptions of asymptotic approximation were not met, the Fisher exact test or Monte Carlo simulations with 10,000 replications were used to obtain a more accurate P value. We report 99% CIs for the estimated P values. For the FKGL reading level score, the average and corresponding 95% Cl were calculated and presented as an error bar chart. EM faculty raters' agreement was examined by using the Gwet Agreement Coefficient 1 (AC1) statistic due to its applicability to data with ordinal scales. The Gwet AC1 statistic was estimated by the *kappaetc* package [35], on Stata (version 17; StataCorp).

Ethical Considerations

This project did not involve any human participants as actual patients who queried the chatbots. The US Food and Drug

XSL•FO
RenderX

Administration (FDA) regulations define a human participant as “an individual who is or becomes a participant in research, either as a recipient of the test article or as a control.” Since the role of the EM faculty in our study was to adjudicate the accuracy of the chatbot responses, we did not find it necessary to seek an ethics review board assessment based on the guidelines set by the University of California, Irvine institutional review board [36].

Results

Overview

Five reviewers scored ChatGPT, Claude AI, Bing AI, and Google Bard for the medical conditions of chest pain, stroke, headache, sore throat, stomach pain, back pain, fainting, heavy menstrual bleeding, bad cold, and overdose. The overall agreement between the raters for all chatbots and across all clinical queries was moderate (Gwet AC1: 0.486; 0.463-0.508; scale 0-1 with 1 denoting perfect agreement). Gwet AC1 ranged between 0.445 and 0.574 among the scored domains.

We depict the performance of the chatbots using radar charts, as in Figures 1-3, where the 8 domains are organized radially. Performance scores range from the best (100%) at the outer edge to the worst (0%) at the center. The further out the point of the polygon is, the better the chatbots perform in that domain. The outermost point in each domain represents the proportion of responses that the faculty reported as the best (totally correct, accurate, or complete) response from the scoring rubric.

Figure 1. Performance of all 4 chatbots in aggregate across 8 different domains, showing the point estimate of the prevalence of the highest or best score in that domain, with the 95% CI.

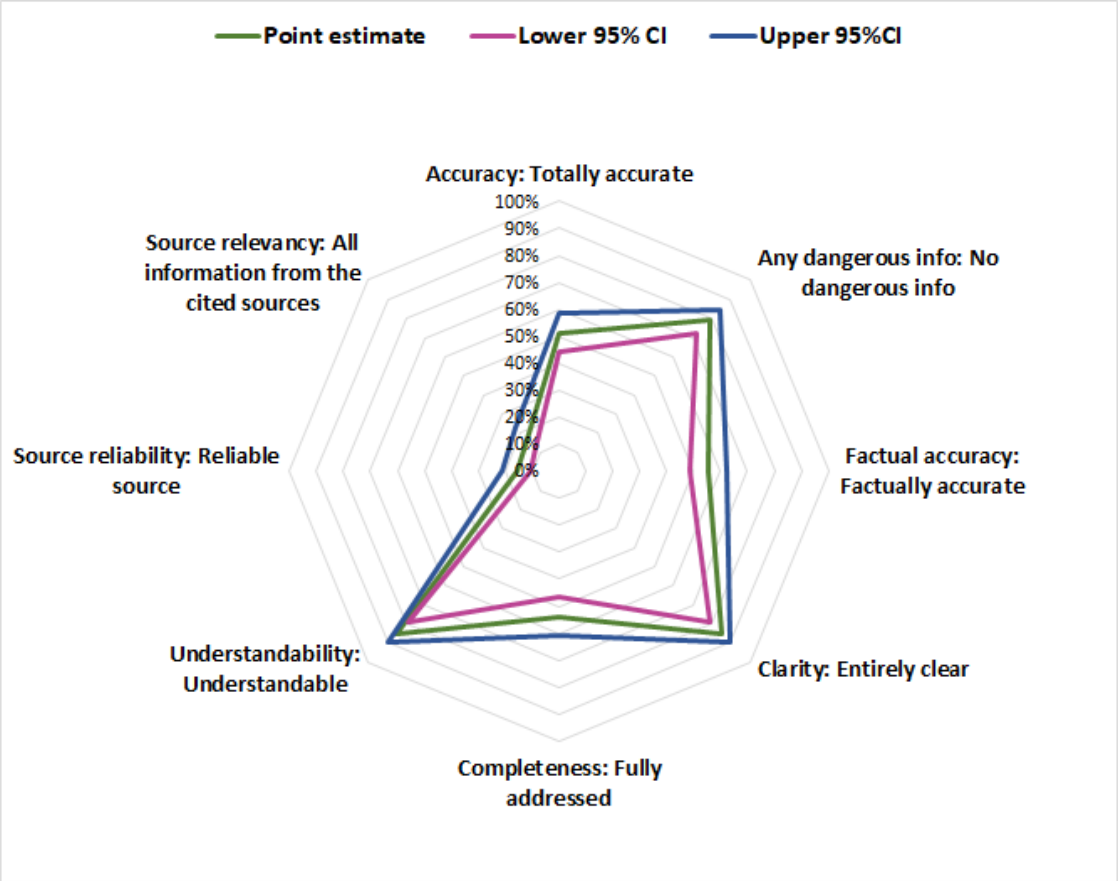

XSL•FO
RenderX

Figure 2. Comparison of the 4 chatbots' performance against each other in the 8 domains.

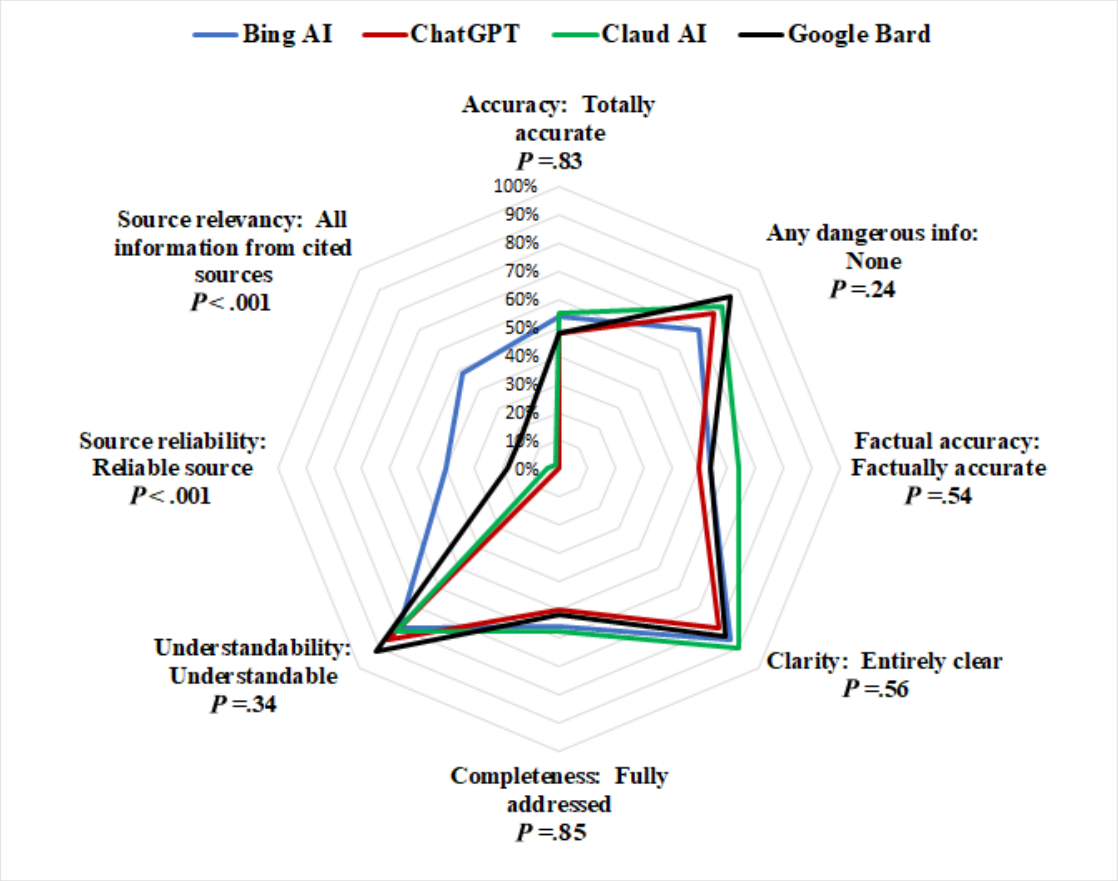

Figure 3. Comparison of condition-specific performance across different domains.

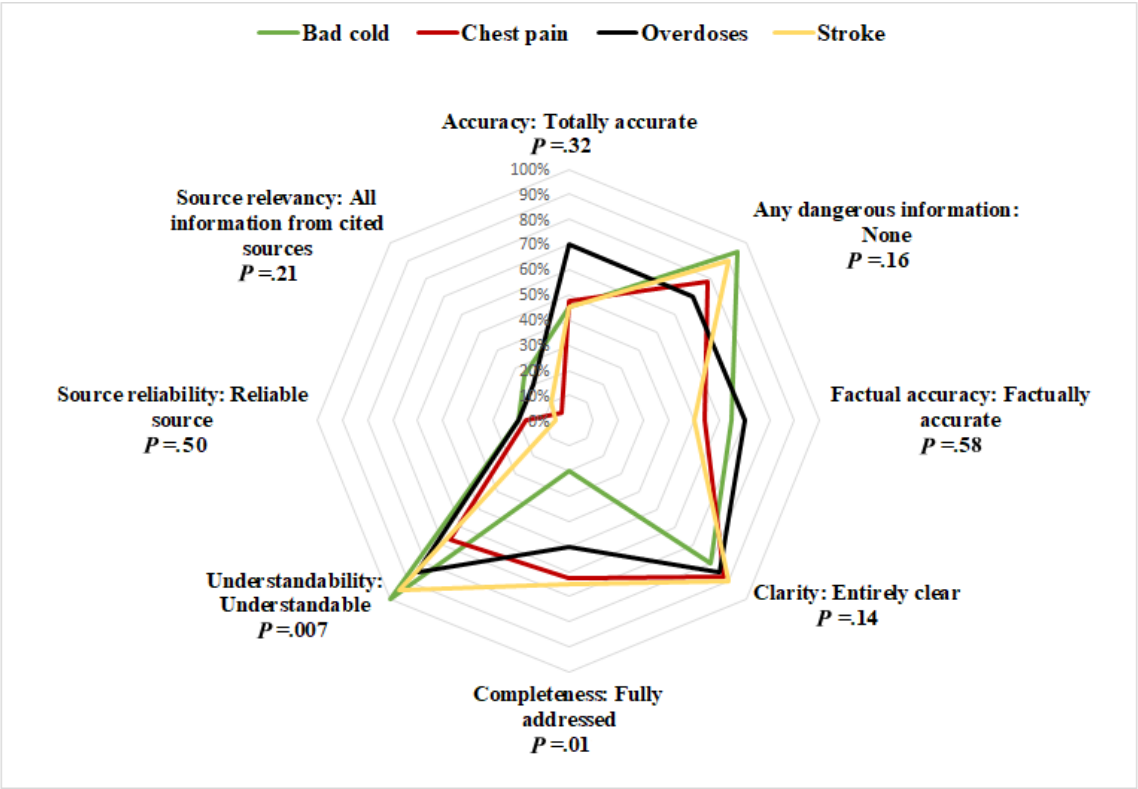

XSL•FO
RenderX

Comparative Analysis of Chatbot Performance

Each of the 4 chatbots was scored for 10 clinical questions across 8 domains by 5 EM faculty for 400 assessments for each chatbot. Figure 1 displays the overall performance of the 4 chatbots across the 10 medical conditions in 8 domains. Together, the 4 chatbots had the best performance in clarity and understandability (both 85%), intermediate performance in accuracy and completeness (both 50%), and poor performance (10%) for source relevance and reliability (most unreported).

Figure 2 compares the performance of the chatbots among the 8 studied domains. ChatGPT, Google Bard, and Claud AI had similar performance across 6 out of 8 domains. Only Bing AI performed better with more relevant and reliable sources. FKGL was 7.7-8.9 grade for all chatbots, except ChatGPT at 10.8, which were all too advanced for average emergency patients.

Condition-Specific Performance (Serious vs Benign Conditions)

We compared the performance of the chatbots across a spectrum of seriousness of medical conditions. "Bad cold" was the least serious of the 10 conditions, and chest pain, overdose, and stroke were the most serious of the questions consumers may pose. Figure 3 shows significant variability in performance between serious and more benign conditions, where there was a significant difference in completeness (P=.01) and understandability (P=.007). This illustrates that chatbots are not consistent in their performance for different domains of "correctness," but also not consistent for serious versus benign complaints. It is, therefore, important, not only to assess the chatbots' performance across conditions and domains but also to consider the seriousness of the queried conditions. The consequences of misinformation (omissions or dangerous information) are potentially greater for more serious conditions.

Safety and Dangerousness

We compared the proportion of chatbot responses for each of the 10 conditions that were "dangerous," in Figure 4. Examples included starting cardiopulmonary resuscitation (CPR) before any pulse check, moving a person who has fainted to "fresh air to help provide oxygen" without any assessment of potential injury or scene safety (carbon monoxide exposure), and use of a defibrillator without starting CPR while awaiting the device.

Overall, we found no statistically significant difference between chatbot responses on this dangerousness metric (Figure 2). Chatbots provided some dangerous information in all cases, from 5% for bad cold to 35% for both bad sore throat and bad back pain, with the most dangerous information (P=.03 for bad sore throat, and P=.04 for bad back pain, compared to the other 8 conditions). The other 8 conditions were statistically similar, yet all included minor to moderate proportions of dangerous information.

Figure 4. Comparison of safety and dangerousness across different medical conditions.

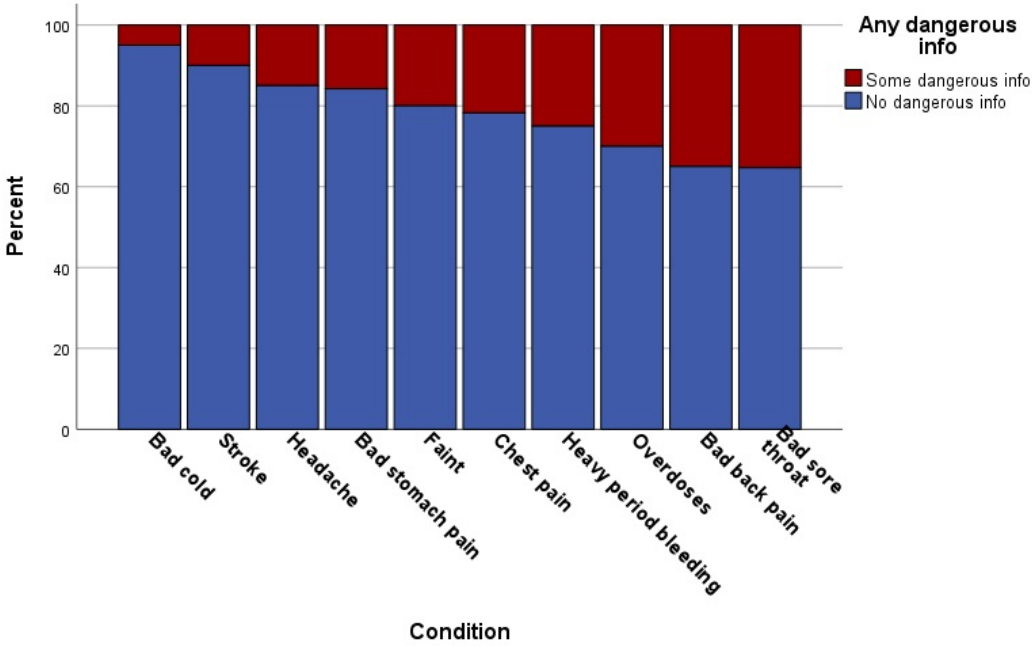

Information Accuracy and Completeness

We measured the accuracy of chatbot responses in 2 ways. First, we asked EM faculty to report what proportion of the total information contained in the responses was accurate by percentage quartiles. Second, we asked them to identify and quantify (none, 1, 2, or more than 2) major and minor inaccuracies.

Figure 5 reports the accuracy of the 10 conditions from most (on the left of the figure) to least (on the right). Note again that there is significant variation by condition, as we showed above in Figure 3 (severity of condition) and Figure 4 (dangerousness).

For percentage accuracy across the 10 conditions, the "completely accurate" response proportion ranged from a high of 70% for headache and overdose, to a low of 30% for fainting. Conditions with a mere 25% accuracy included sore throat and fainting. Other conditions (back pain, stomach pain, period bleeding, stroke, cold, and chest pain) were intermediate in percentage accuracy. Comparing the 10 conditions, there were statistically significant differences between them in percentage

XSL•FO
RenderX

accuracy (*P*=.01; using Monte Carlo simulations with 10,000 replications—99% CI 0.008-0.013), and major or minor factual accuracy (*P*=.01).

Because the order of conditions from most to least accurate, by the alternate percentage by quartile measure, closely matched the order of Figure 5, we elected not to present this percentile data.

As depicted in Figure 2, the aggregate measure of completeness across all 10 questions was similar and ranged from 50% to 60% for each of the 4 chatbots (*P*=.85). However, completeness of response did vary depending on the clinical condition (*P*=.04), from a high of 79% for stomach pain to a low of 20% for bad cold. The responses that were most deficient, as defined by missing more than 2 important pieces of information, varied from a low of 6% (bad sore throat) to a high of 40% (bad cold). The responses from all 10 conditions were judged by some faculty to be missing 1-2 pieces of important information in up to 20% of scores, while 5%-20% of scores identified more than 2 missing pieces of information. None of the 10 medical condition responses were judged by the faculty to contain all necessary information.

Figure 5. Comparison of factual accuracy across different medical conditions.

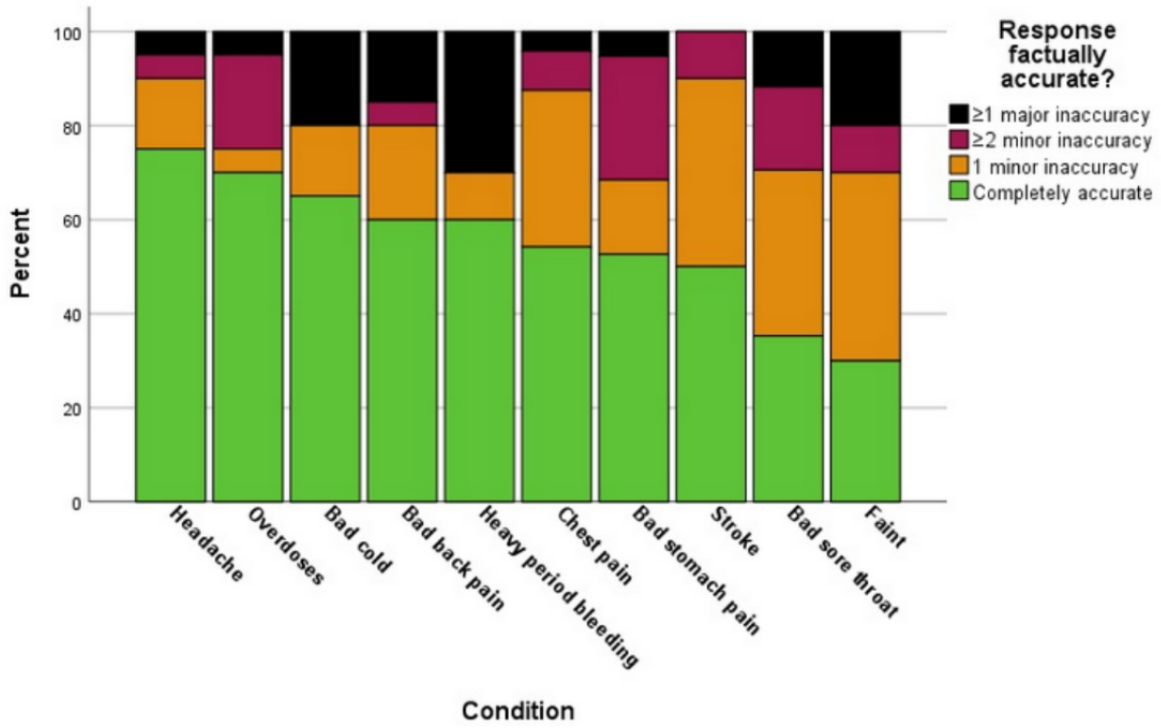

Source Reliability and Relevance

As seen graphically in Figure 2, source reliability and source relevance were outliers from the other 6 domains. As shown in Figure 6, ChatGPT was a further outlier from the other chatbots (*P*<.001), providing essentially no sources for its responses, while Google Bard was slightly better than Claud AI (*P*=.003). Only Bing AI provided what the EM faculty judged as generally reliable sources for a majority of responses to the clinical queries. We found similar results for both domains of source reliability and source relevance, so only the former is presented here.

XSL•FO
RenderX

Figure 6. Comparison between 4 chatbots for the domain of source reliability.

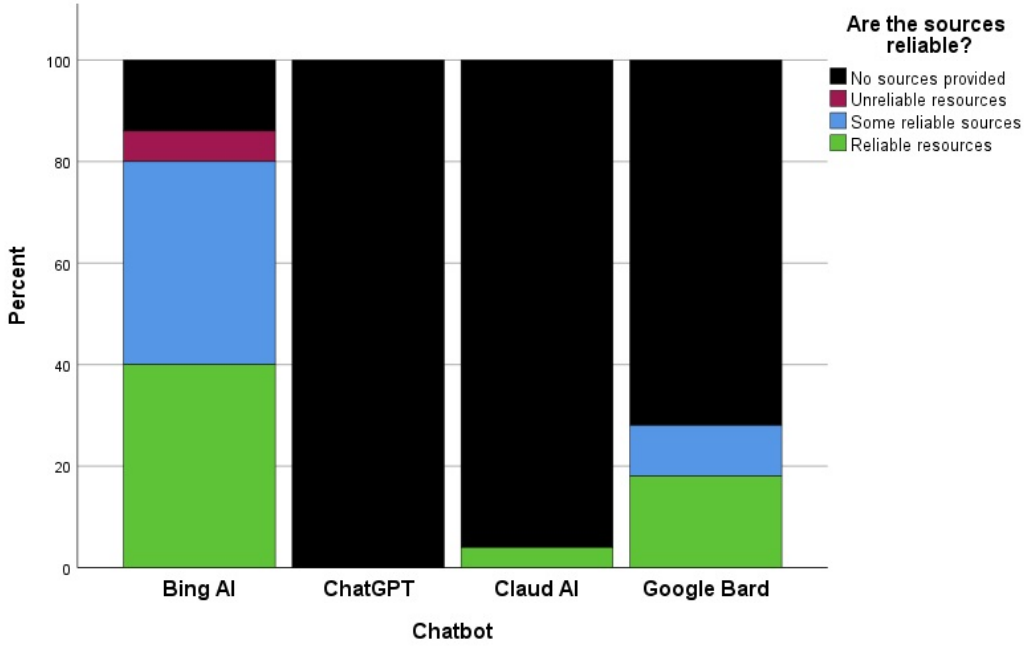

Readability

We processed all 10 clinical query responses from each chatbot through the embedded readability statistics tool in Microsoft Word (FKGL) to calculate a global grade reading level for each chatbot. Table 1 presents the distribution of grade level and chatbot average. The mean grade level for Google Bard, Bing AI, and Claude AI was 7.7-9.16, with ChatGPT scoring 10.76 grade level. As shown in Figure 7, ChatGPT scored higher than Claude AI (P=.003). The grade level across all 4 chatbots was statistically similar as the error bars overlapped (Figure 7). Within any of the 3 chatbots with lower overall grade-level reading scores, the level varied from a low of 5.9 to a high of 12.4. Of the 40 scores, the proportion of chatbot responses for all questions or conditions at or below each grade level was 0% (5th or below), 5% (6th), 15% (7th), 32.5% (8th), 52.5% (9th), 65% (10th), 80% (11th), and 90% at or below 12th grade.

Therefore, for the fifth-grade level recommended by the AMA for patient-oriented reading materials, all of the chatbot responses failed this standard. For the average reading level of the sixth grade for Medicaid-insured patients, the chatbot responses failed 95% of the time. Finally, the chatbots failed the US National Institutes of Health recommendation of eighth grade for reading materials 67.5% of the time [21,37].

Table 1. Emergency care responses to 10 clinical queries from four chatbots—FKGL[a] scores from Microsoft Word.

Medical condition	Chatbots, FKGL score				Total, mean (SD)
	ChatGPT	Google Bard	Bing AI	Claude AI	
Chest pain	9.8	8.5	10	8.2	9.13 (0.91)
Stroke	8.6	7.5	6.9	7	7.50 (0.78)
Headache	10.9	9.8	9.8	7.1	9.40 (1.62)
Sore throat	10.7	8.2	6.3	7.6	8.20 (1.85)
Stomach pain	12.1	12.3	11.7	8	11.03 (2.03)
Back pain	12.4	7.6	11.3	7.2	9.63 (2.61)
Faints	8.5	6.3	6	5.9	6.68 (1.23)
Period bleeding	11.4	10.3	9.8	10.3	10.45 (0.68)
Bad cold	12.4	10.3	11.3	7.6	10.40 (2.05)
Overdose	10.8	8.3	8.5	8.1	8.93 (1.26)
Total, mean (SD)	10.76 (1.43)	8.91 (1.75)	9.16 (2.13)	7.70 (1.14)	N/A[b]

[a]FKGL: Flesch-Kincaid Grade Level.

[b]N/A: not applicable.

XSL•FO
RenderX

Figure 7. Comparison of mean and 95% CI (error bars) for reading level per Microsoft Word Flesch-Kincaid Grade Level (FKGL) score for 4 chatbots.

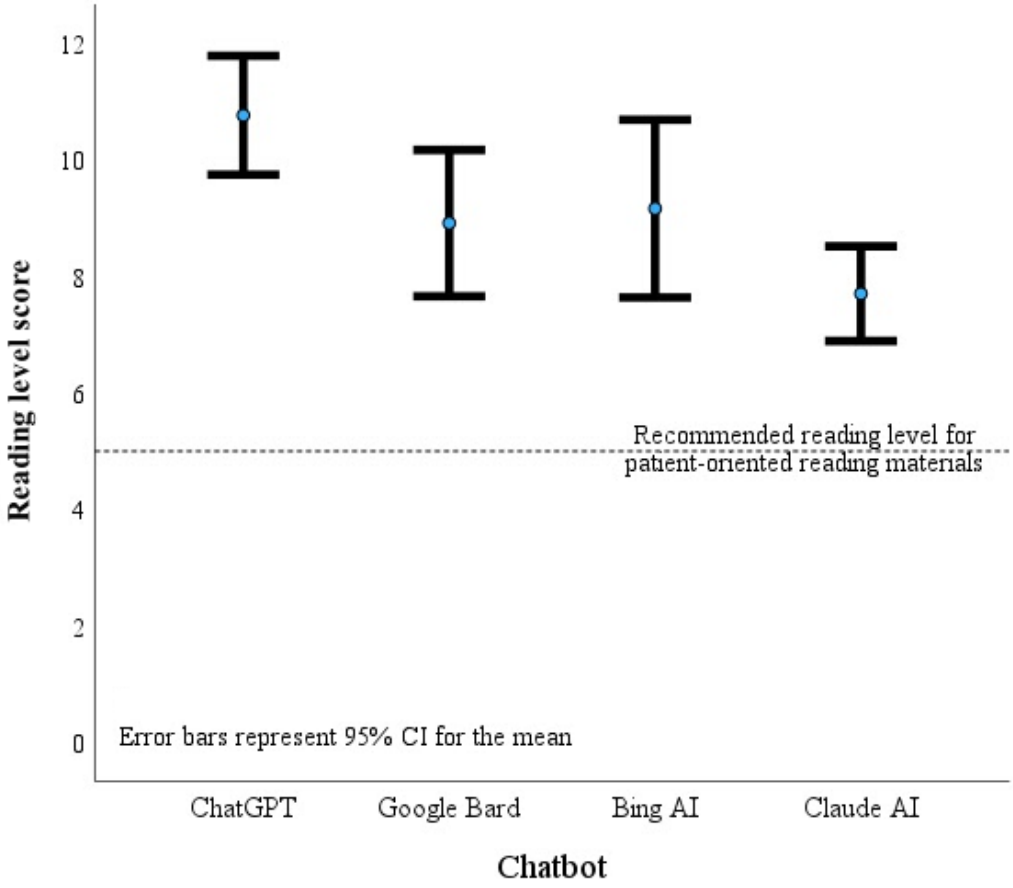

Discussion

Principal Findings

We found that 4 chatbots had significant deficiencies in their performance compared to the opinions of 5 board-certified EM faculty and compared to optimal information drawn from respected mainstream EM texts and reliable websites. Problem areas included inadequate medical or scientific accuracy, incomplete information, dangerous information, and lack of source information.

The 4 chatbots were good and statistically similar in performance in clarity and understandability of presentation (to the educated EM faculty), intermediate in accuracy and completeness, but performed poorly for source disclosure and relevance. All 4 provided what the EM faculty considered dangerous advice with similar frequency. The grade level of English language presentation was above recommended norms for emergency patients (fifth to eighth-grade level) for all of them, with ChatGPT having the highest and, therefore, potentially least user-friendly, grade-level assessment (10th grade).

Dangerous Information or Patient Safety

Because of the high stakes inherent in chatbot queries for emergency conditions, the authors were concerned about the potential for dangerous errors of omission and commission within the chatbot answers. These concerns were validated. There was wide variation in the accuracy of information depending on the clinical query and by chatbot. As all the chatbots had similar rates of dangerous advice and omissions of critical actions for some questions, we did not find any to be superior or inferior to others. We conclude, then, that chatbots may, if at all, be useful for minor conditions that are low risk (perhaps common cold) but should not be used for many EM complaints with potential for serious consequences and complications (chest pain, stroke, and overdose).

Nonscientific Information

Another aspect of the veracity of chatbot responses is the degree to which they contain nonsensical or irrelevant information, without a scientific evidence base. Although our study design did not specifically ask the EM specialists to quantify these, we noted that responses contained such examples—gargling to relieve sore throat, using a heating pad for abdominal pain, eating vitamin C-rich foods to “boost your immune system” (multiple mentions of this), and “stay warm, as cold air can worsen sore throat, and avoid strong odors.”

Misinformation may be directly harmful. A recent report from the US Surgeon General and the US Department of Health and Human Services defines misinformation as—“Information that is false, inaccurate, or misleading according to the best available evidence at the time.” This continues—“...may cause people to make decisions that could have dangerous consequences for their health” [38]. Among our chatbot responses was the unscientific advice to use a heating pad for abdominal pain

XSL•FO
RenderX

without any duration of application, which may delay evaluation for time-sensitive surgical conditions such as appendicitis.

The EM faculty also noticed many instances of vague advice that may not be understood by laypersons. Examples include "support the person's breathing," which may need "triage" to a stroke center, and references to "neurologic symptoms" or "other concerning symptoms" relating to headache.

Reading Level and Understandability

For the specific use of chatbots in EM, readability or understandability is critical, perhaps beyond its importance for general medicine. Regardless of accuracy and completeness, if the consumer cannot readily comprehend the information from the chatbot query, this system of real-time health education fails on a fundamental level. Three of the chatbots had average responses written at the 7th to 9th grade level, with ChatGPT statistically higher at almost the 11th grade average. Even the "better" performing chatbots had some answers written at the 10th, 11th, and 12th grade levels. The lowest grade was grade level 5.9 and only 6 of the 40 answers were written at or below the seventh grade level (see Table 1). The authors conclude that these answers exceed the reading capability recommended in previous literature and US government recommendations and, therefore, may be unsuitable for potential emergency medical conditions [19-21,37].

Spectrum of Queries From Not Likely Dangerous to Life-Threatening and Purposeful Spectrum

We chose 10 questions for the chatbots and EM faculty to reflect some of the most common presenting complaints to American EDs [22,39]. We purposely chose conditions with both benign and serious possibilities and hoped the responses would cover both ends of the acuity spectrum, and provide sound information for when, and in what time course, to seek further care. We found these features only inconsistently in each of the chatbots, with some queries containing such information, and some devoid of it. None of the chatbots performed consistently well on this measure.

Standardized Study Reporting Framework and Assessment Tools are Lacking

We searched for, and then recognized, that there are no standard tools to evaluate health advice generated by chatbots. Hence, we derived our tool from recent papers that studied AI medical advice in other settings [7,8,17,30-34], and the expert opinions of EM faculty. We used both qualitative (minor and major missing information) and quantitative methods (percent of responses that were correct) to judge the chatbots.

With the increasing reliance of both physicians and patients on the internet for health advice, previous authors have stressed the importance of a standardized format for reporting AI-related health advice. Huo et al [29] wrote of their concern that these chatbots are a risk to patient safety. Further, they wrote, and we agree, that chatbot information needs accuracy, and must avoid false and misleading information or sources. These authors have convened an international group of stakeholders to address the importance of developing reporting standards for studies of chatbots used for health advice [29].

Comparisons to Prior Work

Our paper's methodology drafts from previous work, and substantially expands and improves on previous assessments of AI health care information. Recent publications have similarly reported that chatbots are not accurate enough to be used for emergency care. A study in pediatrics concluded that ChatGPT-4 appropriately advised to call emergency services in only 54% of cases and gave accurate first aid instructions in 45%. This paper concluded that its use as an "emergency support tool is questionable" [40]. A second paper showed 3 chatbots had similar but flawed performance to 176 questions from Reddit posed by actual adult patients. These authors divided the questions into "emergency" and "non-emergency," and they reported the chatbots miscategorized 12%-35% of their cases. This study did not, like ours, evaluate the actual response and instructions, but only whether the question posed an emergency. Our study, then, goes beyond mere categorization, to assess the accuracy of the chatbot responses [41].

A recent paper from *JAMA Oncology* shared some of our methodologies; this paper extended those methods and tailored our evaluation to EM. The *JAMA Oncology* paper used a DISCERN tool (published in 1999) to evaluate information on 4 common cancers, and graded accuracy, quality, uncertainty, and reliability, on a 1-5 Likert scale [42]. We refined and expanded this by identifying 8 logical domains of information to assess the chatbot responses. We did not use a Likert scale purposely to limit subjectivity. Instead, we used specific percentages and numbers of inaccurate pieces of information, as judged by content experts in EM.

The previous paper's scoring was not apparently done by content experts. We had our EM experts compare the chatbot responses to "correct" answers drawn from reputable and widely used sources. The *JAMA Oncology* authors similarly assessed the readability of the chatbot responses and found overall 11-12 grade reading level, was certainly too high for the average emergency patient [42].

A more recent grading tool for web-based medical education information was published in 2016 by ALIEM and shared some of our domains for evaluation—accuracy, use, evidence base, and reference-quality [30]. They used a 1-7 Likert scale and similarly had multiple (8) experts evaluate each of the online resources. As these were geared toward physician learners, they did not evaluate readability.

Categories and Examples of Response Deficiencies

The chatbot responses had both categorical and individual deficiencies. The major categories were important omissions of critical information, advice to act with insufficient situational information, absence of information regarding when and how to access health care, vague or technical language not expected to be understandable by laypeople, and extraneous information without scientific support.

There were components of some responses that were considered "omissions," but were so important that the omissions were also considered "dangerous actions." Examples include omitting a pulse check with chest pain, and yet advising to start CPR, and omitting any mention of lack of need or efficacy of antibiotics

XSL•FO
RenderX

for the common cold. Other overtly dangerous actions were advising the use of a defibrillator without directing anyone to start CPR, and advice to move a patient who fainted to fresh air without checking for any injury. Some responses also made no mention of pregnancy, either ectopic or intrauterine, among the advice for vaginal bleeding. Finally, the advice for sore throat omitted any mention of symptoms of airway obstruction, like stridor.

As an example of dangerous advice, Bing (now called "Copilot") responded to the question of chest pain with reasonable advice to call 911, don't drive yourself, chew aspirin, and take nitroglycerin if already prescribed. The introduction said "it's essential to not self-diagnose..." but then said "...Begin CPR (Cardiopulmonary Resuscitation): If the person is having a heart attack start hands-only CPR (emphasis added)...Push hard and fast on the person's chest for 100 to 120 compressions per minute. If an automated external defibrillator (AED) is available, follow the device instructions."

While this last advice is also correct, the answer requires the prompter to, in fact, self-diagnose a heart attack. It then does not state that the patient needs to become unconscious and have no pulse, yet advises to start CPR.

Advice to consumers about when and how to seek medical care was inconsistent and vague. We recognize that EM faculty would have a bias toward expecting the inclusion of "red-flag" symptoms. For example, for severe or persistent headaches, advice to go to the emergency room for "concerning neurological symptoms" is too vague to be useful. Conversely, some responses mentioned serious conditions to watch for but omitted any advice for common conditions with the same complaint. For example, Bing AI provided specific signs to watch for stroke or meningitis but did not give advice about migraine or tension headaches. Finally, 1 chatbot advised in the possible stroke information, for the reader to start CPR, without assessing if the patient was unconscious or pulseless.

At the time of the study, 1 chatbot, Bing AI, provided better source information than the other 3. Chat-GPT, Claude AI, and Google Bard, when asked, "Please list all sources of information you referenced," they generally responded, "I don't have direct access to my training data or know where it came from. I was trained on a mixture of licensed data, data created by human trainers, and publicly available data."

Bing AI, on the other hand, provided references to "learn more," which some faculty interpreted as providing the sources of information for the answer. In addition, some of Bing AI's components of the answers were themselves, hyperlinks from which the information was derived and enabled the user to directly access some sources. Some "learn more" links were to reliable sources like the National Health Service in the United Kingdom, the USA Centers for Disease Control and Intervention, and Mayo and Cleveland Clinics, while others were clearly not authoritative. For example, Bing AI listed medical news sources like MSNBC, proprietary information from various urgent cares, computer resources like Microsoft Start, and layperson sites like Verywell Health and WikiHow. This combination of some reliable sources and embedded hyperlinks from Bing AI explains why it was judged by the experts as superior to the other 3 in this 1 domain.

Strengths and Limitations

This paper is the first, to our knowledge, to compare and contrast multiple chatbots for accuracy, readability, and potential misinformation of answers to patient questions, as judged by multiple academic faculty content experts in the field of EM. As the decision to seek emergency care is high-stakes and high-cost, the authors felt that this level of rigor was appropriate.

This study has several limitations. First, while we conducted a pilot test of the scoring sheet or methodology on 3 of the included clinical conditions (bad cold, fainting, and chest pain) and refined the scoring sheet and methodology with the content experts, we did not pilot all 10 queries. Second, despite including definitions for each question in the survey, there is still a risk that the EM faculty might interpret the questions differently. Third, we did not examine the causes of discrepancies among the faculty in their assessments of source relevance and reliability. Some raters interpreted, "learn more," at the end of the Bing AI chatbot response to be a reporting of sources from which the answers were derived, and some did not. We did not further study this variability in interpretation. Fourth, we did not do a cluster analysis for the 5 raters, leaving the potential that some were inherently more severe graders of AI than others. Fifth, although we derived our clinical questions from among the most common ED chief complaints [22], beyond purposely choosing a mixture of benign and potentially serious conditions, we did not test actual patient responses or actions to the chatbot information. Sixth, we studied the free versions of the chatbots available at the time. Certainly, this is a moving landscape, and there have been updates and improvements in the products since this study was conducted. However, one can only study what is available, and future studies should examine more current and potentially capable versions. Seventh, we chose only 10 questions as a compromise between comprehensiveness and feasibility. Each faculty expert performed 320 assessments (10 questions × 4 chatbots × 8 domains). This number of questions may be insufficient to comprehensively judge the performance of each chatbot. Finally, we did not test the "right answers" derived from reputable medical resources for the 5 faculty members to judge if they considered them truly, "correct." This would have been an appropriate control. Future studies should include this methodology to separate out potential deficiencies in chatbot responses from general variability in expert opinion.

Future Directions

We believe that GenAI will have an impact on social determinants of health, either by making access to accurate health information more accessible to the masses, or rapidly disseminating potentially inaccurate health information to patients with limited access to in-person medical care. For this reason, and because our findings and others have called into question the accuracy of GenAI health information, we believe further research and policy decisions should focus on appropriate regulation by governments and industry.

We would no sooner accept approval of a new pacemaker or orthopedic implant without FDA clearance than we would allow

XSL•FO
RenderX

wide access to incomplete, dangerous, nonscientific information to guide patients' health care decisions. The authors recognize the enormous challenge of regulating words produced by these systems. Nevertheless, we would be remiss in not calling for progress in this area.

Both the European Parliament and the US federal government, as well as China's Ministry of Industry and Information Technology have recognized the need for regulation of AI and put forth draft regulations. As these efforts are just beginning and very fluid, we direct the reader to these references for further information [43-46].

Conclusions

AI chatbots have important deficiencies for EM patient advice, despite their consistent performance. Advice for when to seek urgent or emergent care is frequently incomplete and inaccurate, and patients may be unaware of misinformation. Sources of information are not generally disclosed. Patients who use AI to guide their health care assume potential risks. The use of AI chatbots for health information should be subject to further research, refinement, and regulation. We strongly recommend proper medical consultation to prevent potential adverse outcomes. Finally, we call for further validation of scoring tools and the development of literature reporting guidelines.

Acknowledgments

The authors used ChatGPT 3.5, Google Bard, Claude AI, and Bing AI to generate the responses to 10 clinical questions that were subsequently judged by our board of content experts. Generative artificial intelligence (AI) was not used to draft any part of the study. The authors received no financial support for the research and authorship.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JYSY performed the conceptualization, investigation, methodology, project administration, software, and writing—original draft and review and editing. SS did the data curation, formal analysis, investigation, methodology, software, visualization, and writing—original draft and review and editing. EH contributed to the investigation, methodology, and writing—original draft and review and editing. LSLM did the investigation, methodology, project administration, and writing—original draft and review and editing. JSR performed the investigation, methodology, and writing—original draft and review and editing. JS did the investigation, methodology, and writing—original draft and review and editing. AT did the investigation. WW did the investigation, methodology, and writing—original draft and review and editing. MIL contributed to the investigation, methodology, and writing—original draft and review and editing. All take responsibility for the final publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Research prompts.
[PDF File (Adobe PDF File), 34 KB - jmir_v26i1e60291_app1.pdf]

Multimedia Appendix 2
Chatbot log.
[PDF File (Adobe PDF File), 7290 KB - jmir_v26i1e60291_app2.pdf]

Multimedia Appendix 3
Chatbot 10 right answers.
[PDF File (Adobe PDF File), 99 KB - jmir_v26i1e60291_app3.pdf]

Multimedia Appendix 4
Google scoring form.
[PDF File (Adobe PDF File), 44 KB - jmir_v26i1e60291_app4.pdf]

References

1. Hoffman M. What is a chatbot + how does it work? The ultimate guide. Zendesk. 2020. URL: https://www.zendesk.com/blog/what-is-a-chatbot/ [accessed 2024-04-15]
2. Joseph A, Eapen NG. Conversational agents and chatbots: current trends. In: Pillai AS, Tedesco R, editors. Machine Learning and Deep Learning in Natural Language Processing. Boca Raton, FL: CRC Press; 2023.

XSL•FO
RenderX

3. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. J Med Internet Res 2024 Jul 23;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
4. What is a chatbot? IBM. 2020. URL: https://www.ibm.com/topics/chatbots [accessed 2024-04-15]
5. Metz C, Grant N. Google updates bard chatbot with 'Gemini' A.I. as it chases ChatGPT. The New York Times. 2023. URL: https://www.nytimes.com/2023/12/06/technology/google-ai-bard-chatbot-gemini.html [accessed 2024-03-21]
6. Mehdi Y. Reinventing search with a new AI-powered microsoft bing and edge, your copilot for the web. Official Microsoft Blog. 2023. URL: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ [accessed 2024-03-19]
7. Kung TH, Cheatham M, Medenilla A, Sillos C, de Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Yeo YH, Samaan JS, Ng WH, Ting P, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
9. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
10. Reddy S. Evaluating large language models for use in healthcare: a framework for translational value assessment. Inform Med Unlocked 2023;41:101304. [doi: 10.1016/j.imu.2023.101304]
11. Wang X, Cohen RA. Health Information Technology Use Among Adults: United States. Hyattsville, MD: National Center for Health Statistics (U.S.); 2022.
12. Can GenAI help make health care affordable? Consumers think so. Deloitte US. URL: https://www2.deloitte.com/us/en/blog/health-care-blog/2023/can-gen-ai-help-make-health-care-affordable-consumers-think-so.html [accessed 2024-01-05]
13. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
14. Hanebutt R, Mohyuddin H. The digital domain: a "Super" social determinant of health. Prim Care 2023 Dec;50(4):657-670. [doi: 10.1016/j.pop.2023.04.002] [Medline: 37866838]
15. Digital access: a super determinant of health. Substance Abuse and Mental Health Services Administration. 2023. URL: https://www.samhsa.gov/blog/digital-access-super-determinant-health [accessed 2024-03-18]
16. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq 2023 Feb 28:rs-2566942 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
17. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023 Jun;33(6):1790-1796 [FREE Full text] [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]
18. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? J Am Soc Nephrol 2023 Aug 01;34(8):1302-1304. [doi: 10.1681/ASN.0000000000000166] [Medline: 37254254]
19. Powers RD. Emergency department patient literacy and the readability of patient-directed materials. Ann Emerg Med 1988 Feb;17(2):124-126. [doi: 10.1016/s0196-0644(88)80295-6] [Medline: 3337429]
20. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 2021;8:2374373521998847 [FREE Full text] [doi: 10.1177/2374373521998847] [Medline: 34179407]
21. Weiss BD, Blanchard JS, McGee DL, Hart G, Warren B, Burgoon M, et al. Illiteracy among Medicaid recipients and its relationship to health care costs. J Health Care Poor Underserved 1994;5(2):99-111. [doi: 10.1353/hpu.2010.0272] [Medline: 8043732]
22. Arvig MD, Mogensen CB, Skjøt-Arkil H, Johansen IS, Rosenvinge FS, Lassen AT. Chief complaints, underlying diagnoses, and mortality in adult, non-trauma emergency department visits: a population-based, multicenter cohort study. West J Emerg Med 2022 Oct 31;23(6):855-863 [FREE Full text] [doi: 10.5811/westjem.2022.9.56332] [Medline: 36409936]
23. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah M. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus 2023 Jun;15(6):e40351 [FREE Full text] [doi: 10.7759/cureus.40351] [Medline: 37456381]
24. Fahy E, Hardikar R, Fox A, Mackay S. Quality of patient health information on the internet: reviewing a complex and evolving landscape. Australas Med J 2014;7(1):24-28 [FREE Full text] [doi: 10.4066/AMJ.2014.1900] [Medline: 24567763]
25. Liu J, Zhang Y, Kim Y. Consumer health information quality, credibility, and trust: an analysis of definitions, measures, and conceptual dimensions. 2023 Presented at: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval; July 1, 2018; New York, NY p. 197-210. [doi: 10.1145/3576840.3578331]

XSL•FO
RenderX

26. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002;287(20):2691-2700. [doi: 10.1001/jama.287.20.2691] [Medline: 12020305]
27. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018 Oct 19;18(1):87 [FREE Full text] [doi: 10.1186/s12911-018-0668-9] [Medline: 30340488]
28. Breckons M, Jones R, Morris J, Richardson J. What do evaluation instruments tell us about the quality of complementary medicine information on the internet? J Med Internet Res 2008 Jan 22;10(1):e3 [FREE Full text] [doi: 10.2196/jmir.961] [Medline: 18244894]
29. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. Nat Med 2023 Dec;29(12):2988. [doi: 10.1038/s41591-023-02656-2] [Medline: 37957381]
30. Chan TM, Grock A, Paddock M, Kulasegaram K, Yarris LM, Lin M. Examining reliability and validity of an online score (ALiEM AIR) for rating free open access medical education resources. Ann Emerg Med 2016 Dec;68(6):729-735. [doi: 10.1016/j.annemergmed.2016.02.018] [Medline: 27033141]
31. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology 2023 Oct;180:35-58. [doi: 10.1016/j.urology.2023.05.040] [Medline: 37406864]
32. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023 Jun;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
33. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics (Basel) 2023 Jun 02;13(11):1950 [FREE Full text] [doi: 10.3390/diagnostics13111950] [Medline: 37296802]
34. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. Eur J Hosp Pharm 2023 Jun 01:3750. [doi: 10.1136/ejhpharm-2023-003750] [Medline: 37263772]
35. Klein D. Implementing a general framework for assessing interrater agreement in Stata. Stata J 2018 Dec 01;18(4):871-901. [doi: 10.1177/1536867x1801800408]
36. Do you need IRB review? UCI Office of Research. URL: https://research.uci.edu/human-research-protections/do-you-need-irb-review/ [accessed 2024-10-05]
37. Weiss BD. Health literacy: a manual for clinicians. American Medical Association Foundation and American Medical Association. 2003. URL: http://lib.ncfh.org/pdfs/6617.pdf [accessed 2024-09-27]
38. Health misinformation. U.S. Department of Health and Human Services. 2024. URL: https://www.hhs.gov/surgeongeneral/priorities/health-misinformation/index.html [accessed 2024-08-26]
39. Estimates of emergency department visits in the United States, 2016-2021. National Center for Health Statistics. 2023. URL: https://www.cdc.gov/nchs/dhcs/ed-visits/index.htm [accessed 2024-04-15]
40. Bushuven S, Bentele M, Bentele S, Gerber B, Bansbach J, Ganter J, et al. "ChatGPT, Can You Help Me Save My Child's Life?" - diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases - an in-silico analysis. J Med Syst 2023 Nov 21;47(1):123 [FREE Full text] [doi: 10.1007/s10916-023-02019-x] [Medline: 37987870]
41. Zúñiga Salazar G, Zúñiga D, Vindel CL, Yoong AM, Hincapie S, Zúñiga AB, et al. Efficacy of AI chats to determine an emergency: a comparison between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus 2023 Sep;15(9):e45473 [FREE Full text] [doi: 10.7759/cureus.45473] [Medline: 37727841]
42. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence Chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
43. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [accessed 2024-04-15]
44. Artificial intelligence 2023 legislation. National Conference of State Legislatures. URL: https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation [accessed 2024-04-15]
45. Ye J. China issues draft guidelines for standardising AI industry. Reuters. 2024. URL: https://www.reuters.com/technology/china-issues-draft-guidelines-standardising-ai-industry-2024-01-17/ [accessed 2024-04-16]
46. EU AI act: first regulation on artificial intelligence. Topics European Parliament. 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2024-04-15]

Abbreviations

AC1: Agreement Coefficient 1

XSL•FO
RenderX

AI: artificial intelligence
AMA: American Medical Association
CPR: cardiopulmonary resuscitation
ED: emergency department
EM: emergency medicine
FDA: US Food and Drug Administration
FKGL: Flesch-Kincaid Grade Level
GenAI: generative artificial intelligence
LLM: large language model
NCHS: National Center for Health Statistics

Edited by A Schwartz; submitted 07.05.24; peer-reviewed by B Thies, AD Rouhi, H Mohammad-Rahimi; comments to author 16.08.24; revised version received 27.08.24; accepted 12.09.24; published 04.11.24.

Please cite as:
Yau JYS, Saadat S, Hsu E, Murphy LSL, Roh JS, Suchard J, Tapia A, Wiechmann W, Langdorf MI
Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study
J Med Internet Res 2024;26:e60291
URL: https://www.jmir.org/2024/1/e60291
doi:10.2196/60291
PMID:

©Jonathan Yi-Shin Yau, Soheil Saadat, Edmund Hsu, Linda Suk-Ling Murphy, Jennifer S Roh, Jeffrey Suchard, Antonio Tapia, Warren Wiechmann, Mark I Langdorf. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review

Leyao Wang[1*]; Zhiyu Wan[2,3*], PhD; Congning Ni[1], ME; Qingyuan Song[1], ME; Yang Li[1], BE; Ellen Clayton[2,4,5], MD, JD; Bradley Malin[1,2,6], PhD; Zhijun Yin[1,2], PhD

[1]Department of Computer Science, Vanderbilt University, Nashville, TN, United States
[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[3]School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, United States
[5]School of Law, Vanderbilt University Medical Center, Nashville, TN, United States
[6]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States
[*]these authors contributed equally

Corresponding Author:
Zhijun Yin, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave Ste 1475
Nashville, TN, 37203
United States
Phone: 1 6159363690
Fax: 1 6159368545
Email: zhijun.yin@vumc.org

Abstract

Background: The launch of ChatGPT (OpenAI) in November 2022 attracted public attention and academic interest to large language models (LLMs), facilitating the emergence of many other innovative LLMs. These LLMs have been applied in various fields, including health care. Numerous studies have since been conducted regarding how to use state-of-the-art LLMs in health-related scenarios.

Objective: This review aims to summarize applications of and concerns regarding conversational LLMs in health care and provide an agenda for future research in this field.

Methods: We used PubMed, ACM, and the IEEE digital libraries as primary sources for this review. We followed the guidance of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to screen and select peer-reviewed research articles that (1) were related to health care applications and conversational LLMs and (2) were published before September 1, 2023, the date when we started paper collection. We investigated these papers and classified them according to their applications and concerns.

Results: Our search initially identified 820 papers according to targeted keywords, out of which 65 (7.9%) papers met our criteria and were included in the review. The most popular conversational LLM was ChatGPT (60/65, 92% of papers), followed by Bard (Google LLC; 1/65, 2% of papers), LLaMA (Meta; 1/65, 2% of papers), and other LLMs (6/65, 9% papers). These papers were classified into four categories of applications: (1) summarization, (2) medical knowledge inquiry, (3) prediction (eg, diagnosis, treatment recommendation, and drug synergy), and (4) administration (eg, documentation and information collection), and four categories of concerns: (1) reliability (eg, training data quality, accuracy, interpretability, and consistency in responses), (2) bias, (3) privacy, and (4) public acceptability. There were 49 (75%) papers using LLMs for either summarization or medical knowledge inquiry, or both, and there are 58 (89%) papers expressing concerns about either reliability or bias, or both. We found that conversational LLMs exhibited promising results in summarization and providing general medical knowledge to patients with a relatively high accuracy. However, conversational LLMs such as ChatGPT are not always able to provide reliable answers to complex health-related tasks (eg, diagnosis) that require specialized domain expertise. While bias or privacy issues are often noted as concerns, no experiments in our reviewed papers thoughtfully examined how conversational LLMs lead to these issues in health care research.

XSL•FO
RenderX

Conclusions: Future studies should focus on improving the reliability of LLM applications in complex health-related tasks, as well as investigating the mechanisms of how LLM applications bring bias and privacy issues. Considering the vast accessibility of LLMs, legal, social, and technical efforts are all needed to address concerns about LLMs to promote, improve, and regularize the application of LLMs in health care.

(J Med Internet Res 2024;26:e22769) doi:10.2196/22769

KEYWORDS

large language model; ChatGPT; artificial intelligence; natural language processing; health care; summarization; medical knowledge inquiry; reliability; bias; privacy

Introduction

Background

Since ChatGPT (OpenAI) was released on November 30, 2022, extensive attention has been drawn to generative artificial intelligence (AI) and large language models (LLMs) [1]. ChatGPT is a representative conversational LLM that generates text based on its training on an extremely large amount of data from mostly the public domain [1]. Modern LLMs (such as GPT-4) incorporate in-text learning, which enables them to interpret and generalize user inputs in the form of natural language prompts that require little to no fine-tuning [2]. These LLMs have surpassed their non–transformer-based counterparts and are now capable of performing various complex natural language processing tasks, including translation and question-answering [3]. In comparison with traditional chatbots, the current array of conversational LLMs can generate seemingly human-like coherent texts [3]. Moreover, because these models are trained on publications from digital libraries, such as Common Crawl and Wikipedia, they can generate seemingly scientific and competent answers [4].

Due to the high quality of their responses and the broad training database of modern LLMs, a growing body of studies has emerged regarding the applications of chatbots, particularly ChatGPT, in the domain of health and medicine [5]. However, most LLMs are not specially designed for health care, and therefore, certain practical pitfalls may exist when they are put into practice in that setting. Thus, there is a need to compile the latest achievements in this domain so that potential issues and guidance for new research directions can be laid out. Several reviews have been published to discuss the appropriateness of a particular application of LLMs in a specific aspect [1,6-9] but none of them summarized the overall problems systematically [8]. For example, Huang et al [6] and Giannakopoulos et al [10] summarized the application of ChatGPT only in dentistry without considering the broader landscape of other subfields in health care. Wang et al [7] discussed the ethical considerations of using ChatGPT in health care; they did not consider other LLMs for analysis, account for other common challenges, such as reliability, or mention detailed applications of the models. Moreover, their work focused on LLMs' educational and research applications rather than their clinical use. Although Sallam [8] conducted a systematic review, the articles considered in the review were mostly editorials, letters to the editors, opinions, commentaries, news articles, and preprints, as opposed to research articles. In addition, Sallam [8] focused on educational and research applications of ChatGPT only. Puladi et al [11] narratively reviewed papers on the applications of LLMs in oral and maxillofacial surgery. Pool et al [12] reviewed papers on the application of LLMs in telehealth. Park et al [9] conducted a scoping review of papers on the medical applications of LLMs. These papers are limited in either focusing on a specific medical application area, including nonpeer-reviewed articles, or lacking a systematic examination of the concerns regarding conversational LLMs.

This Review

This review focuses on peer-reviewed research articles on conversational LLMs that emerged after ChatGPT, which was initially based on GPT-3 (OpenAI), and their applications in health care. We aim to summarize the applications of conversational LLMs in the field of health care with concrete applications and identify potential concerns about the use of such LLMs in this field that need to be addressed in the future.

Methods

Search Strategy

We searched for articles that contained at least 1 word associated with LLMs ("ChatGPT," "LLaMA," "GPT-3," "LaMDA," "PalM," "MT-NLG," "GATO," "BLOOM," "Alpaca," "Large Language Model") and at least 1 word associated with health care ("health," "diagnosis," "intervention," "patient") published before September 1, 2023, on PubMed, ACM Digital Library, and IEEE Xplore. This systematic review applied the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) to steer the literature search [13-15]. Relevant publications were gathered and downloaded on September 3, 2023. For simplicity, all the LLMs mentioned henceforth refer to conversational LLMs.

Criteria

Textbox 1 summarizes the inclusion and exclusion criteria for articles. Specifically, the inclusion criteria for a paper were as follows: (1) it was published as a peer-reviewed scientific research article between November 1, 2022, and September 1, 2023, and (2) it focuses on applications of LLMs in addressing a health care–related problem, which includes, but is not limited to, promotion of personal or public health and well-being or the potential to alleviate the workload of health care providers. We excluded a paper if it was (1) not a peer-reviewed research article, (2) not related to health care applications (eg, LLMs applied to preparing manuscripts for peer review), (3) not accessible, (4) a duplicate of an existing paper, or (5) about

XSL•FO
RenderX

LLMs released before GPT-3, such as bidirectional encoder representations from transformers (BERT). We excluded BERT-related papers because this LLM, which was built upon the encoder of a transformer, is mainly applied in fine-tuning downstream machine-learning tasks. While the implementation of a chatbot based on BERT is feasible, it waned in popularity as an LLM after the introduction of ChatGPT, which was built upon the decoder of a transformer. The complete set of papers meeting the criteria were downloaded from the 3 digital libraries for further screening. Specifically, 5 of the authors of this review (LW, ZW, CN, QS, and YL) participated in paper screening and summarization under the supervision of the corresponding author, ZY. A screening protocol was created collectively after the team jointly reviewed 50 randomly selected papers. Each unreviewed paper was then screened by not fewer than 2 authors based on the protocol. All the papers in the final collection were summarized by the coauthors according to their LLM applications in health care and the concerns raised.

Textbox 1. Paper inclusion and exclusion criteria.

Inclusion criteria

- Article type: peer-reviewed scientific research article
- Written language: English
- Time of publications: published between November 1, 2022, and September 1, 2023
- Accessibility: accessible
- Duplication: is not a duplicate of an existing article
- Models: conservational large language models (LLMs) after GPT-3 was launched
- Topic: any topics related to health care, which includes, but is not limited to, promotion of personal or public health and well-being or the potential to alleviate the workload of health care providers

Exclusion criteria

- Article type: any other types of publications
- Written language: any non-English language
- Time of publications: published before November 1, 2022, or after September 1, 2023
- Accessibility: not accessible
- Duplication: is a duplicate of an existing article
- Models: LLMs before GPT-3 was launched or not used for conversations
- Topic: any other topics that are not related to health care applications (eg, preparing manuscripts for peer review)

Results

Overview

Figure 1 demonstrates the paper selection process. The initial keyword search identified 820 articles, with 736 (89.8%) articles from PubMed, 49 (6%) papers from ACM Digital Library, and 35 (4.3%) papers from IEEE Xplore. The evaluation of the 820 articles was distributed among the authors for screening the titles and abstracts. The interrater reliability was assessed by computing a κ score, yielding a value of 0.72. After screening, we excluded 599 (81.4%) of the 736 articles from PubMed, 46 (94%) of the 49 articles from ACM Digital Library, and 33 (94%) of the 35 papers from IEEE Xplore because they were either not relevant to the research topic or were not research articles. No duplicates were found after the screening. Next, we extracted the full papers for the remaining 142 (17.3%) of 820 research articles and manually examined them for the 5 exclusion criteria (refer to the Methods section). This led to a final set of 65 (7.9%) of 820 articles for full-paper review and summarization—63 (97%), 2 (3%), and 0 from PubMed, ACM Digital Library, and IEEE Xplore, respectively. Among these selected articles, 60 (92%) were related to ChatGPT, 1 (2%) was related to LLaMA (Meta), 1 (2%) was related to Bard based on Language Model for Dialogue Applications (Google LLC), and 6 (9%) were related to other LLMs (Table 1 lists the specific LLM or LLMs mentioned in each selected paper). Note that 2 selected papers were related to >1 LLM, respectively.

Five of the authors (LW, ZW, CN, QS, and YL) compiled the topics related to applications and concerns independently during the paper screening and summarization process. Furthermore, through extensive discussions, all the authors refined and categorized these topics into main applications and concerns with corresponding subcategories. Figure 2 illustrates the main topics of applications and concerns mentioned by the reviewed papers on applying LLMs in health care settings. The multifaceted applications of LLMs can be divided into 4 primary categories: summarization, medical knowledge inquiry, prediction, and administration: summarization (25/65, 38% papers)—LLMs are potential tools for summarizing complex information or documentation in clinical domains. Medical knowledge inquiry (30/65, 46% papers)—LLMs demonstrate proficiency in answering a diverse array of medical questions and examinations, which enhance public access to medical knowledge. Prediction (22/65, 34% papers)—LLMs demonstrate high *diagnostic* accuracy in multiple medical scenarios (15/65, 23% papers), offer assistance in diverse *treatments* (12/65, 18%

XSL•FO
RenderX

papers), and excel in predicting drug interactions and *synergies* (1/65, 2% paper). Administration (9/65, 14% papers)—LLMs streamline various tasks, including *documentation* (5/65, 8% papers) and *information collection* (5/65, 8% papers) to monitor the trend of public health.

The concerns surrounding the application of LLMs in health care were varied, each with nuanced considerations: Reliability (55/65, 85% papers)—This includes *accuracy* (45/65, 69% papers), or the correctness of the responses from LLMs; *consistency* (13/65, 20% papers), whether LLMs produce the same response to the same questions with different prompts; *interpretability* (5/65, 8% papers), whether LLMs can explain their responses well, and the data *quality* of the training dataset (16/65, 25% papers). Bias (16/65, 25% papers)—The applications of LLMs may result in biased responses, which will exacerbate disparity and inequality in health care, particularly in terms of *financial costs* (1/65, 2% paper), *readability* (5/65, 8% papers), and *accessibility* (3/65, 5% papers). Privacy (6/65, 9% papers)—Training LLMs in health care settings requires a large number of health data which, however, is sensitive and may bring privacy issues.

Public acceptance (4/65, 6% papers): Building trust in LLMs from the public is pivotal for widespread acceptance and use of LLM-based health care applications.

Figure 1. A flowchart of the article selection process based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

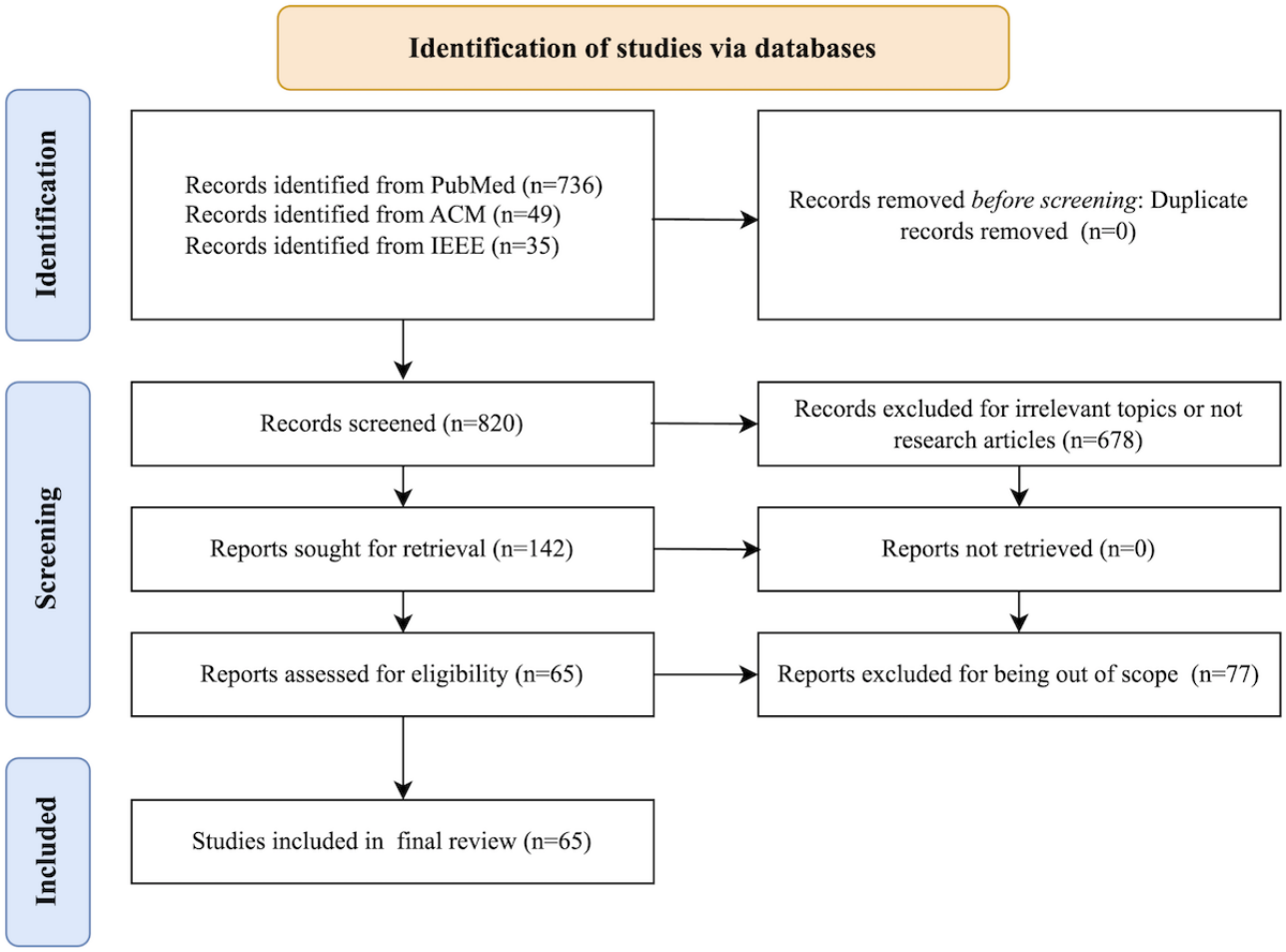

XSL•FO
RenderX

Table 1. A summary of applications and concerns of the reviewed research papers. Subcategories of a paper are shown in parentheses immediately after the paper's category.

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Akhter and Cooper [16]	Summarization	The authors used ChatGPT to generate a case report for a patient who developed a common complication.	Reliability (accuracy and data quality: data timeliness)	This paper shows that ChatGPT sometimes cites nonexistent sources and is currently limited in critically discussing results and literature.	ChatGPT	United States
Almazyad et al [17]	Summarization	This paper shows that ChatGPT can summarize conference medical recommendations.	Reliability (accuracy) and bias (biased training data)	This paper shows that results from ChatGPT can be biased.	ChatGPT-4	Saudi Arabia and Lebanon
Bhattacharyya et al [18]	Summarization	The authors investigated the authenticity and accuracy of references in ChatGPT-generated medical articles.	Reliability (accuracy)	This paper shows that ChatGPT often generates fabricated or inaccurate medical references, with a high prevalence of errors in reference elements.	ChatGPT-3.5	United States
Bosbach et al [19]	Summarization	In this paper, ChatGPT drafts competent radiology reports with high appraisal given command files as input.	Reliability (accuracy)	In this paper, ChatGPT showed limitations in its ability to deal with technical or medical terminology.	ChatGPT	Switzerland, Germany, Poland, Hungary, and Malaysia
Chen et al [20]	Summarization	In this paper, ChatGPT is used to construct a risk factor database for diseases, demonstrating its potential to extract data from literature abstracts effectively.	Reliability (accuracy)	This paper discusses the challenges in accurately extracting risk factor information, emphasizing the need for human validation to ensure model accuracy.	ChatGPT	China
Lahat et al [21]	Summarization	This paper discusses ChatGPT's potential in formulating gastroenterology research questions.	__[b]	—	ChatGPT	Israel and the United States
Puthenpura et al [22]	Summarization	The case report uses ChatGPT to integrate AI[c]-generated text with original author writing.	Reliability (interpretability and accuracy)	This paper notes the potential for AI-generated text to be inaccurate and provides nonexistent references in information. Dependence on AI tools may lead to overlooking subtle clinical signs by clinicians.	ChatGPT	United States
Robinson and Aggarwal [23]	Summarization	ChatGPT's effectiveness in drafting operation notes for appendicectomy is evaluated in this paper, demonstrating adherence to NHS[d] surgical documentation guidelines.	Reliability (consistency)	This paper notes the dependency of ChatGPT's output quality on the prompt and emphasizes the need for secure integration with health records for surgical documentation.	ChatGPT	United Kingdom
Zhou [24]	Summarization	The authors explored ChatGPT's capabilities in generating medical reports from laboratory results with the goal of streamlining the report generation process.	Reliability (accuracy)	This paper raises concerns about the precision and reliability of medical report generation by ChatGPT.	ChatGPT	United States

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Guirguis et al [25]	Summarization	The authors used ChatGPT to generate a case report for a patient with neurosarcoidosis.	—	—	ChatGPT	United States
Haemmerli et al [26]	Summarization (case report) and prediction (diagnosis, treatment recommendation)	The authors evaluated ChatGPT's performance in brain glioma adjuvant therapy decision-making. It can enhance medical case reporting and study data analysis for manuscript production.	Reliability (accuracy, data quality: data timeliness) and bias (biased algorithm)	This paper shows that ChatGPT is less accurate and more biased in the interpretation of medical information. It also shows that ChatGPT lacks live internet access and access to research databases.	ChatGPT	Switzerland, Lithuania, Austria, and the United States
Lyu [27]	Summarization (clinical notes)	This paper shows ChatGPT's ability to translate radiology reports into plain language with nice results.	Reliability (accuracy and consistency)	This paper shows concerns regarding ChatGPT's moral and legal issues. For example, ChatGPT tends to oversimplify or overlook and omit key points during translation, resulting in inaccuracy.	ChatGPT, GPT-4	United States
Cunningham et al [28]	Summarization and medical knowledge inquiry	This paper uses ChatGPT to aid in manuscript synthesis of a patient with glioblastoma in the pineal gland exhibited over 5 years of survival following radiotherapy and temozolomide.	Reliability (accuracy, data quality: date timeliness, and interpretability) and bias (readability)	This paper shows that ChatGPT cannot substitute professional medical advice and has concerns about its knowledge cutoff of 2021 and its inability to access the internet.	ChatGPT	United States
Golan et al [29]	Summarization and medical knowledge inquiry	The authors used ChatGPT to evaluate the quality and readability of online medical text regarding shock-wave therapy for erectile dysfunction	Reliability (accuracy) and bias (readability)	This paper shows that ChatGPT, in its current state, is less effective than human reviewers and a reliable tool [30].	ChatGPT	United States
Hamed et al [31]	Summarization and medical knowledge inquiry	This study integrates ChatGPT-4 with "Link Reader" for automating medical text synthesis, improving AI models' traceability and retrieval accuracy.	—	—	ChatGPT-4	Qatar and the United Kingdom
Grewal et al [32]	Summarization, medical knowledge inquiry, and prediction (treatment recommendation)	This paper highlights ChatGPT's applications in radiology, including report generation, template creation, patient communication, clinical decision-making enhancement, research title suggestion, scholarly article heading creation, and formatting and referencing for research papers.	Reliability (data quality: data source and data timeliness) and privacy	This paper shows that ChatGPT is limited by biases in its training data, may produce inaccuracies, and integrating it with EHRs[e] risks. It may further be hampered by outdated training data.	ChatGPT (GPT-4)	United States and India
Kumari and Anusha [33]	Summarization, medical knowledge inquiry, and prediction (treatment recommendation)	This paper incorporates information from ChatGPT in the treatment planning and case report writing.	Reliability (accuracy) and public acceptance	This paper shows that ChatGPT provides more general information.	ChatGPT	India

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Cadamuro et al [34]	Summarization and prediction (diagnosis)	This paper evaluates ChatGPT with laboratory reports for relevance, correctness, helpfulness, and safety, suggesting that it can interpret individual tests but not an overall diagnostic picture.	Reliability (accuracy and interpretability)	In this paper, ChatGPT incorrectly interpreted normal results for suspected diseases and struggled to synthesize all related laboratory test findings coherently.	ChatGPT	Austria, Italy, Croatia, Belgium, Spain, and Turkey
Jiang et al [35]	Summarization and prediction (treatment recommendation)	The authors developed NYUTron, a LLM trained on unstructured clinical notes for clinical predictive tasks.	Reliability (accuracy and consistency) and bias (accessibility)	This paper shows that it is hard to ensure the accuracy and reliability of predictions in a clinical setting. The model lacks generalizability.	NYUTron	United States, Canada
Sharma et al [36]	Summarization and prediction (diagnosis)	The authors developed DR.BENCH[f], a generative AI framework for clinical diagnostic reasoning tasks. They showed that a multitask, clinically trained language model significantly outperforms general domain models.	—	—	DR.BENCH	United States
Liu et al [37]	Summarization, prediction (diagnosis and treatment recommendation), medical knowledge inquiry, and administration (documentation)	The authors explored ChatGPT's roles in clinical practice, focusing on clinical decision support, question-answering, and medical documentation.	Reliability (accuracy) and privacy	The authors pointed out that potential negatives such as privacy, ethics, bias, and discrimination, compounded by outdated training data, cannot be overlooked.	ChatGPT	China, United States
Hamed et al [38]	Summarization (clinical notes) and administration (documentation)	This paper demonstrates ChatGPT's ability to adapt clinical guidelines.	—	—	ChatGPT	Qatar, United Kingdom
Kim [39]	Summarization and administration (documentation)	This paper introduces a case study that shows that ChatGPT can help with medical documentation.	—	—	ChatGPT	United States, Korea
Macdonald et al [40]	Summarization and administration (documentation)	This paper demonstrates that ChatGPT can write a paper giving a dataset.	Reliability (accuracy)	The authors pointed out that ChatGPT can produce incorrect references and pass plagiarism detectors with a 100% score.	ChatGPT	United Kingdom
Cascella et al [41]	Summarization and administration (documentation and information collection)	This paper shows that ChatGPT can summarize information, list possible research topics, and write clinical notes.	—	—	ChatGPT	Italy
Ali [42]	Medical knowledge inquiry	This paper shows that ChatGPT can provide information about lacrimal drainage disorders.	Reliability (accuracy)	This paper shows that the information from ChatGPT is not all correct.	ChatGPT	India

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Antaki et al [43]	Medical knowledge inquiry	This paper evaluates ChatGPT's proficiency in answering ophthalmic questions, showing promising results in a simulated OKAP[g] examination.	Reliability (accuracy and data quality: data source)	This paper shows that ChatGPT's accuracy depends on concordance and insight, with inaccuracies often due to insufficient training.	ChatGPT, ChatGPT plus	Canada
Bird and Lotfi [44]	Medical knowledge inquiry	The authors optimized a chatbot that can answer questions regarding mental health with high accuracy.	Reliability (data quality: data source)	The authors pointed out that the available data are limited, and it takes a lot of efforts to collect data.	An unnamed chatbot	United Kingdom
Hoch et al [45]	Medical knowledge inquiry	This paper shows that ChatGPT displays high quiz skills and accuracy in examinations.	Reliability (accuracy)	This paper shows that ChatGPT can give false answers to a substantial proportion of questions in specific otolaryngology subdomains.	ChatGPT	Germany, United States, Spain, and the United Kingdom
Holmes et al [46]	Medical knowledge inquiry	In this paper, LLMs, including ChatGPT, are evaluated on radiation oncology physics, with GPT-4 exhibiting superior performance and reasoning abilities.	Reliability (consistency and accuracy)	This paper highlights ChatGPT's consistency in answering radiation oncology physics questions yet underscores the superior performance of a team of medical physicists.	ChatGPT (GPT-3.5), ChatGPT (GPT-4), Bard (LaMDA[h]), and BLOOMZ	United States
Hristidis et al [47]	Medical knowledge inquiry	In this paper, ChatGPT and Google are compared for dementia-related queries, assessing the quality and reliability of their responses.	Bias (readability)	This paper comments on the relevance and readability of responses from ChatGPT and Google for dementia-related queries, noting challenges in both platforms.	ChatGPT	United States
Johnson et al [48]	Medical knowledge inquiry	The authors assessed ChatGPT's ability to answer cancer information-related questions, indicating that ChatGPT provides accurate information about common cancer myths and misconceptions.	Reliability (accuracy, data quality: data timeliness) and bias (biased algorithm)	The authors advocated that future evaluation of AI platforms needs infrastructure to monitor for bias and health disparities, considering user trust and credibility in AI responses.	ChatGPT	United States
Kung et al [49]	Medical knowledge inquiry	This paper investigates ChatGPT's capability to surpass USMLE's[i] passing threshold, showing its increasing accuracy and potential in medical education.	Reliability (accuracy and consistency)	This paper shows that AI's performance in medical examinations limited to human perception.	ChatGPT	United States
Kusunose et al [50]	Medical knowledge inquiry	In this paper, ChatGPT provided accurate responses to CQs[j] related to the JSH[k] 2019 guidelines for the management of hypertension.	Reliability (accuracy)	In this paper, ChatGPT did not provide accurate responses to some questions.	ChatGPT	Japan

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Lahat et al [51]	Medical knowledge inquiry	Evaluating ChatGPT's answers to gastrointestinal health questions, this study indicates its capacity to provide accurate information in certain areas.	Reliability (accuracy)	This paper highlights the varying quality of ChatGPT's information and emphasizes the need for further development to enhance its utility for patients.	ChatGPT	Israel and the United States
Li et al [52]	Medical knowledge inquiry	The authors refined LLaMA using 100,000 patient-doctor dialogues to provide medical advice.	Reliability (accuracy)	The paper shows that the accuracy of LLMs such as ChatGPT could be significantly improved if they could generate or assess responses based on a reliable knowledge database with experiments.	LLaMA	United States and China
Moshirfar et al [53]	Medical knowledge inquiry	GPT-4 outperforms GPT-3.5 and human experts in answering ophthalmology questions, with significant variations across different difficulty levels.	Reliability (accuracy, data quality: data timeliness) and bias (financial costs and accessibility)	The drawbacks of using GPT-4 include paying a monthly fee and having a knowledge cutoff of September 2021.	ChatGPT (GPT-3.5) and ChatGPT (GPT-4)	United States
Nov et al [54]	Medical knowledge inquiry	This paper assesses ChatGPT's answers to patient questions with health care providers, indicating its effectiveness in generating patient responses.	Reliability (accuracy, data quality: data source), public acceptance, and bias (biased training data)	This paper shows that ChatGPT and similar LLMs face issues such as biased or incorrect responses, with automation bias and liability concerns requiring vigilant chatbot response curation.	ChatGPT	United States
Sallam et al [55]	Medical knowledge inquiry	This paper shows that ChatGPT can challenge misinformation, such as COVID-19 vaccine conspiracies.	Reliability (accuracy) and bias (biased training data)	The authors pointed out that ChatGPT only has limited knowledge by 2021, so it is possible that it can produce biased and unreliable results.	ChatGPT	Jordan, Lebanon, and Indonesia
Sinha et al [56]	Medical knowledge inquiry	This paper shows the high accuracy of ChatGPT to solve higher-order reasoning questions in pathology.	Reliability (accuracy, data quality: data timeliness)	The authors pointed out that ChatGPT has limitations in that they have information on 2021, and future AI systems must be carefully designed, developed, and validated to ensure they provide accurate information.	ChatGPT	India
Thirunavukarasu et al [57]	Medical knowledge inquiry	This paper assesses ChatGPT's primary care application, showing promise but necessitating further development as indicated by its AKT[l] performance.	Reliability (accuracy) and public acceptance	This paper acknowledges the potential and current limitations of ChatGPT in primary care, indicating a need for further development to reach the expertise level of qualified physicians.	ChatGPT	United Kingdom

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Van Bulck and Moons [58]	Medical knowledge inquiry	In this paper, 17 of 20 experts consider ChatGPT provides answers of a higher or equal value compared with Google search.	Reliability (consistency)	This paper shows that ChatGPT is sensitive to nuance in the prompts. It uses outdated training data and is less transparent with its sources.	ChatGPT	Belgium, Sweden, and South Africa
Wagner and Ertl-Wagner [59]	Medical knowledge inquiry	The accuracy of ChatGPT-3 in retrieving clinical radiological information is tested in this paper, cross-checking its responses with peer-reviewed references.	Reliability (accuracy)	This paper expresses concerns about the accuracy and authenticity of ChatGPT-3's radiological information and references.	ChatGPT-3	Canada
Walker et al [60]	Medical knowledge inquiry	This study evaluates the reliability of medical information from ChatGPT-4 using the EQIP[m] tool and comparison with clinical guidelines for 5 hepato-pancreaticobiliary conditions.	Reliability (accuracy, consistency, and interpretability) and bias (readability)	This paper shows that ChatGPT-4 has no support for references, complicated answers, and accuracy issues.	ChatGPT-4	United Kingdom, Switzerland, and Saudi Arabia
Yeo et al [61]	Medical knowledge inquiry	The performance of ChatGPT in responding to questions about cirrhosis and hepatocellular carcinoma is assessed in this paper, showing extensive knowledge in these areas.	Reliability (data quality)	This paper expresses concerns about ChatGPT's limitations in providing comprehensive and region-specific knowledge, particularly in managing cirrhosis and hepatocellular carcinoma.	ChatGPT	United States, United Kingdom
Zhu et al [62]	Medical knowledge inquiry	This paper shows that ChatGPT is able to pass the Chinese Medical Licensing Examination's Clinical Knowledge Section.	Reliability (accuracy) and privacy	The authors pointed out the importance of privacy, accuracy, and reliability for an AI system.	ChatGPT	China
Altamimi et al [63]	Medical knowledge inquiry and prediction (treatment recommendation)	In this study, ChatGPT is assessed for providing advice on venomous snakebites and offering accurate management information in simulated consultations.	Reliability (accuracy)	This paper discusses reliability concerns of ChatGPT in providing snakebite management advice, stressing the need for updated knowledge and personalized information.	ChatGPT	Saudi Arabia
Goktas et al [64]	Medical knowledge inquiry and prediction (diagnosis)	The authors gave examples of using "ChatGPT 4.0" in the field of allergy and immunology.	Reliability (accuracy), bias, and privacy	The authors pointed out the importance of privacy and reliability.	ChatGPT 4.0	Ireland, Turkey
Chervenak et al [65]	Medical knowledge inquiry and administration (information collection)	This paper evaluates ChatGPT's performance on fertility-related clinical queries.	Reliability (accuracy)	This paper shows that: (1) ChatGPT provides an illusion of reliability in its persuasive prose. (2) Different patient populations may interact with ChatGPT in different ways. (3) ChatGPT is not able to reliably cite sources.	ChatGPT	United States

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Agbavor and Liang [66]	Prediction (diagnosis)	This research shows GPT-3-based text embeddings can differentiate patients with Alzheimer from healthy controls through speech data, suggesting early diagnostic potential.	Reliability (accuracy)	This paper acknowledges the limited research on using LLMs for early dementia diagnosis, specifically the potential of GPT-3.	GPT-3	United States
Hirosawa et al [4]	Prediction (diagnosis)	This paper shows that ChatGPT-3 can generate a diagnosis list for common chief complaints with high accuracy.	Reliability (data quality: data source, interpretability) bias (biased training data)	The papers show that it is unclear about the hyperparameters and training algorithms of the ChatGPT, thus it lacks transparency or interpretability. In addition, ChatGPT may produce misleading and biased results. Last, ChatGPT lacks recent knowledge.	ChatGPT-3	Japan
Huang et al [6]	Prediction (diagnosis)	This study shows that ChatGPT can be used in dental diagnosis.	Reliability (accuracy), bias (accessibility), and privacy	This study shows that ChatGPT may violate patients' privacy, and ChatGPT cannot truly understand data and may produce biased results.	ChatGPT	China and United States
Karkera et al [67]	Prediction (diagnosis)	The authors assessed various pretrained LLMs for extracting microbe-disease relationships from biomedical texts in zero-shot or few-shot contexts.	Reliability (consistency) and bias (algorithm bias)	This paper shows that varying outputs for identical prompts raise concerns about the model's response reliability.	GPT-3, BioGPT, BioMedLM, BioMegatron, PubMedBERT, BioClinicalBERT, and BioLinkBERT	Japan, India
Sarbay et al [68]	Prediction (diagnosis)	ChatGPT's performance in emergency triage prediction is assessed in this paper, comparing its predictions with expert categories and scoring its sensitivity and specificity.	Reliability (accuracy)	This paper notes discrepancies and inconsistencies in some cases.	ChatGPT	Turkey
Shahsavar and Choudhury [3]	Prediction (diagnosis)	This paper examines factors that influence users' intentions to use ChatGPT for self-diagnosis and health-related purposes, revealing a high willingness to adopt the technology.	Reliability (data quality) and public acceptance	The paper notes that ChatGPT is not specifically designed for health care purposes, which may affect its suitability for self-diagnosis.	ChatGPT	United States
Galido et al [69]	Prediction (diagnosis and treatment recommendation)	This paper shows that ChatGPT can identify patients as having TRS[n] accurately, make treatment suggestions, and identify drug side effects in the treatment recommendation.	Reliability (consistency and data quality: data source)	This paper shows that ChatGPT can be combined with commercial applications while generating answers. However, its output is influenced by incorrect input, and it lacks clinical context and the ability to request edits to input errors.	ChatGPT	United States

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Sorin et al [70]	Prediction (diagnosis, treatment recommendation)	In this paper, ChatGPT is evaluated as a decision support tool for the breast tumor board, showing promise in recommending management aligned with tumor board decisions.	Reliability (accuracy and consistency)	This paper reflects on the alignment of ChatGPT with tumor board recommendations, showing potential as a decision support tool with a 70% concordance rate.	ChatGPT	Israel
Juhi et al [71]	Prediction (drug synergy)	In this paper, ChatGPT is tested on predicting and explaining drug-drug interactions, aiming to enhance patient safety by providing accurate drug compatibility information.	Reliability (accuracy)	This paper shows that ChatGPT sometimes provides incomplete information.	ChatGPT	India
Haver et al [72]	Prediction (treatment recommendation)	In this paper, ChatGPT demonstrates high accuracy in providing recommendations and prevention of breast cancer.	Reliability (consistency)	This paper shows that ChatGPT is sensitive to nuance in the prompts. ChatGPT is a research "chatbot" not specially designed for medical use.	ChatGPT	United States
Liu et al [73]	Prediction (treatment recommendation)	The authors compared ChatGPT-generated clinical support alerts with human-made suggestions, highlighting its potential for unique, understandable, and relevant contributions.	Reliability (consistency, accuracy, and data quality: data timeliness)	The authors showed that ChatGPT's responses vary with prompt changes, highlighting its sensitivity to different input sentences.	ChatGPT	United States
Kao et al [74]	Prediction (treatment recommendation, diagnosis)	The authors evaluated ChatGPT as a CDS[o] tool in pediatrics, suggesting its capability to improve clinical workflow and assist in responsible decision-making.	Reliability (accuracy)	The authors argued that AI technologies, such as ChatGPT, are not yet advanced enough to replace doctors in complex diagnoses or treatment planning.	ChatGPT	China
Schulte [75]	Prediction (treatment recommendation and diagnosis)	ChatGPT was used to identify guideline-based treatments for advanced solid tumors with a VTQ[p] of 0.77 when being compared with NCCN[q] guidelines.	Reliability (accuracy and consistency)	The authors showed that ChatGPT's accuracy and consistency are not certain.	ChatGPT	United States
Carpenter and Altman [76]	Administration (information collection)	In this paper, GPT-3 is used to generate a drug abuse lexicon from social media slang, aiming to improve pharmacovigilance and monitor of drug abuse trends.	Reliability (accuracy) and bias (readability)	This paper acknowledges challenges in generating a reliable lexicon for drug abuse synonyms due to the variability of social media language.	GPT-3	United States
Jo et al [2]	Administration (information collection)	The authors built CareCall, an AI tool built on HyperCLOVA, which aims at monitoring the health conditions of socially isolated groups.	Reliability (data quality: data source)	The authors pointed out that firsthand data are hard to gather because collecting personal health data may give rise to privacy issues.	CareCall	United States and Korea

XSL•FO
RenderX

Reference	Application		Concern		LLMs[a]	Authors' countries
	Category	Note	Category	Note		
Montagna et al [77]	Administration (information collection)	The authors used GPT-3 to develop an LLM-based chatbot to support management of patients' health data related to chronic diseases.	Reliability (data quality: data source) and privacy	The paper shows that LLMs lack medical expertise and may be influenced by any bias in the data they were trained on. How to protect patients' privacy is another issue the authors pointed out in this paper.	GPT-3	Italy

[a]LLM: large language model.
[b]Not available.
[c]AI: artificial intelligence.
[d]NHS: National Health Service.
[e]EHR: electronic health record.
[f]DR.BENCH: Diagnostic Reasoning Benchmark.
[g]OKAP: Ophthalmic Knowledge Assessment Program.
[h]LaMDA: Language Model for Dialogue Applications.
[i]USMLE: United States Medical Licensing Exam.
[j]CQ: clinical question.
[k]JSH: Japanese Society of Hypertension.
[l]AKT: Applied Knowledge Test.
[m]EQIP: Ensuring Quality Information for Patients.
[n]TRS: treatment-resistant schizophrenia.
[o]CDS: clinical decision support.
[p]VTQ: valid therapy quotient.
[q]NCCN: National Comprehensive Cancer Network.

Figure 2. A summary of the applications and concerns about large language models (LLMs) in health care as communicated by the reviewed papers.

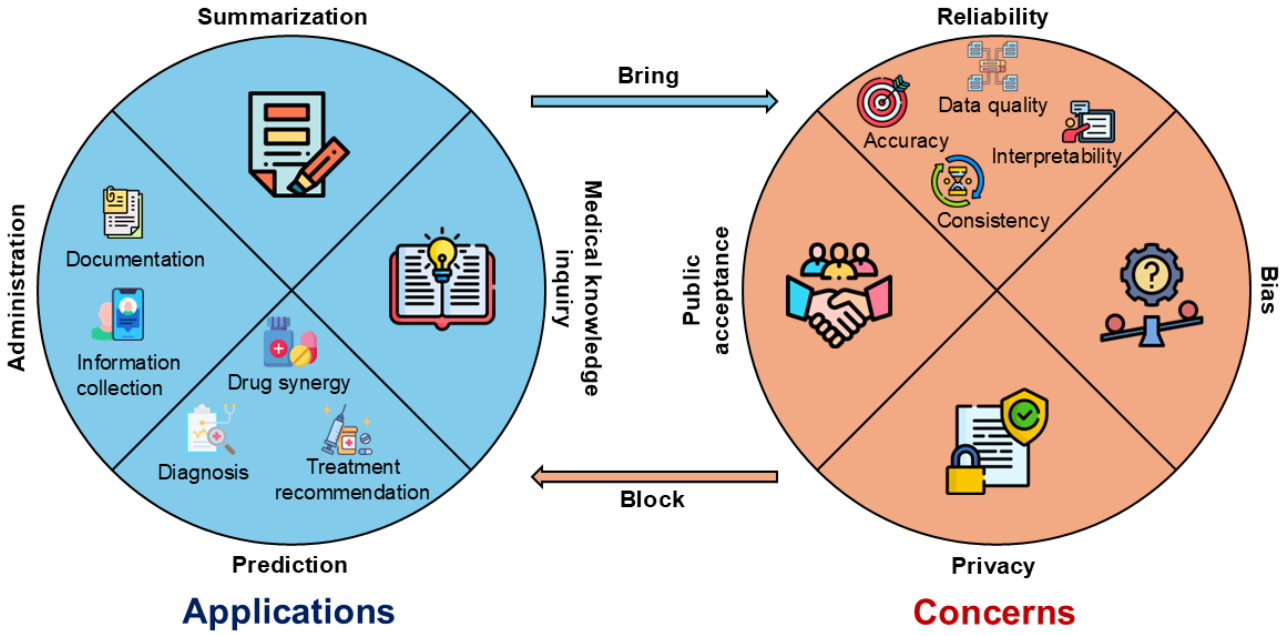

Applications

All reviewed research papers demonstrated the usability or tested the capability of LLMs for health care applications in clinical or research domains, which can be further classified into the following 4 categories: summarization, medical knowledge inquiry, prediction, and administration.

XSL•FO
RenderX

Summarization

ChatGPT has been shown to be effective in summarizing medical documents for a diverse set of applications [37,39], including tasks such as adapting clinical guidelines for diagnosis, treatment, and disease management [38], summarizing medical notes [19,23,41], assisting in writing medical case reports [16,22,24-26,35], and generating and translating radiological reports [19,32]. Notably, efforts have been made to integrate ChatGPT-4 with the Link Reader plugin for automating medical text synthesis [31], which boosted model performance in providing answers according to clinical guidelines [31]. Another study by Zhou [24] explored ChatGPT's role in supporting health care professionals in creating medical reports from real patient laboratory results to offer treatment recommendations based on patients' health conditions [24].

ChatGPT proved beneficial for summarizing research papers as well [17]. Notably, it demonstrated impressive performance in summarizing conference panels and recommendations [17], generating research questions [21], extracting data from literature abstracts [20], drafting medical papers based on given datasets [40], and generating references from medical articles [18]. ChatGPT was also used to evaluate the quality and readability of web-based medical text regarding shock-wave therapy for erectile dysfunction [29]. These applications highlighted the potential of LLMs to condense complex and extensive research materials, allowing for more accessible comprehension and use of information in health care.

Medical Knowledge Inquiry

ChatGPT can be applied to answer questions about health care, as evidenced by its excellent performance in various studies [28,31,42,50,51,53,54,59-61,63]. For instance, ChatGPT has shown remarkable accuracy in reasoning questions and medical exams [45,56], even successfully passing the Chinese Medical Licensing Examination [62] and the United States Medical Licensing Examination [49]. It also performed well in addressing radiation oncology physics examination questions [46]. Likewise, "ChatGPT would have been at the 87th percentile of Bunting's 2013 international cohort for the Cardiff Fertility Knowledge Scale and at the 95th percentile on the basis of Kudesia's 2017 cohort for the Fertility and Infertility Treatment Knowledge Score" [65]. In addition, ChatGPT showed promising results in a simulated Ophthalmic Knowledge Assessment Program exam [43]. However, the average score of ChatGPT was 60.17% in the Membership of the Royal College of General Practitioners Applied Knowledge Test, which is <70.4%, the mean passing threshold in the last 2 years [57].

Furthermore, LLMs have been shown to be effective at making medical knowledge accessible to the public. In particular, a fine-tuned chatbot based on LLaMA demonstrated enhanced performance in identifying patients' needs and providing informed suggestions [52]. In the realm of medical advice, ChatGPT-generated educational documents, answered questions about allergy and immunology [64], and countered vaccine conspiracy theories [55]. It can also answer the most frequently asked questions about the COVID-19 pandemic. Its overall responses to queries related to cognitive decline were equivalent to and, at times, more reliable than Google's [47]. According to Bulck and Moons [58], in comparison with Google search, 40% (8 experts) of the 20 experts (19 nurses and 1 dietitian) considered answers from ChatGPT of greater value, 45% (9 experts) regarded them as equal value, and 15% (3 experts) deemed them less valuable. Therefore, many experts predicted that patients will gradually rely more on LLMs (particularly ChatGPT) and less on Google searches due to the high quality and accessibility of the answers from LLMs. Regarding cancer myths and misconceptions, 97% (63/65) of expert reviews deemed answers from ChatGPT to be accurate [48]. In addition, Bird and Lotfi [44] optimized a chatbot that could answer mental health–related questions with an accuracy of 88.7% (26,595/30,000 tokens) [69]. Overall, LLMs, particularly ChatGPT, demonstrate an impressive performance in public education in health.

Prediction

LLMs have been shown to have predictive capabilities in diagnosis, treatment recommendations, and drug interactions and synergies.

Diagnosis

ChatGPT has exhibited the potential to achieve high accuracy in diagnosing specific diseases [37,69], providing diagnostic suggestions in simulated situations [63,73] or using given laboratory reports for diagnosis [34]. ChatGPT has been evaluated in dental [6], allergy [64], and mental disorders diagnoses [66]. Particularly, GPT-3 can be used to differentiate patients with Alzheimer disease from healthy controls using speech data [66]. Beyond ChatGPT, other generative AI frameworks, such as DR.BENCH [36], were used for clinical diagnostic reasoning tasks [36]. Moreover, various pretrained LLMs can extract microbe-disease relationships from biomedical texts in zero-shot or few-shot contexts with high accuracy, with an average F_1-score, precision, and recall >0.8 [67]. In addition, ChatGPT was the best LLM when predicting high acuity cases than predicting low acuity cases according to the emergency severity index, with a sensitivity of 0.762 and a specificity of 0.931, compared with the overall sensitivity of 0.571 and a specificity of 0.345 [68].

For example, Hirosawa et al [4] obtained ChatGPT's diagnostic response by describing a clinical scenario. The prompt began with "Tell me the top 10 suspected illnesses for the following symptoms;" Then, patients' personal information (eg, age and family history) was provided in this prompt along with other clinical data (eg, symptoms, medication, and physical examination). According to the study, the top 10 suspected diseases generated by ChatGPT achieved a rate of 93% (28/30) in overall correctness. While such a level of performance is impressive, physicians still made a better prediction than ChatGPT. With respect to the top 5 diagnoses, physicians achieved an accuracy of 98% (59/60) while ChatGPT only achieved 83% (25/30). As for the top suspected disease, ChatGPT only had a correct rate of 53% (16/30), versus 93% (56/60) achieved by physicians [4].

XSL•FO
RenderX

Treatment Recommendations

LLMs can offer treatment recommendations while listing the side effects of these treatments [69]. They have been involved in the treatment of various diseases, such as allergy and immunology [64]. ChatGPT can identify guideline-based treatments for advanced solid tumors [75], such as breast tumor treatment [70]. LLMs can also assist with treatment planning [33] and brain glioma adjuvant therapy decision-making [26]. Similarly, NYUTron, an LLM trained on unstructured clinical notes, has been applied for clinical predictive tasks in treatments [35]. ChatGPT can effectively recommend breast tumor management strategies based on clinical information from 10 patients [70], enhance clinical workflow, and assist in responsible decision-making in pediatrics [74]. In addition, ChatGPT can recommend cancer screening given the radiology reports, with an accuracy of 88% (22/25) [72]. Overall, ChatGPT performs well in certain scenarios of disease prevention and screening recommendations.

Drug Synergies

LLMs also demonstrate high utility when characterizing drug effects. Notably, ChatGPT was used to predict and explain drug-drug interactions [71]. In this study, the LLMs were asked about pairing or interaction between drugs, and their responses are evaluated in terms of correctness and conclusiveness. Among the 40 pairs of drug-drug interactions, 39 (98%) responses were correct for the first question, and among these 39 correct answers, 19 (49%) were conclusive while 20 (51%) were inconclusive. For the second question, 39 (97%) were correct among 40 pairs, with 17 (44%) answers conclusive and 22 (56%) answers inconclusive.

Administration

LLMs can serve a multifaceted role in the realm of health care and administrative tasks. Specifically, ChatGPT proves instrumental in streamlining administrative processes by generating texts, thereby alleviating the associated workload [38]. Moreover, it can be used to track patients' health status, particularly those with chronic diseases [77]. Through the analysis of social media slang, GPT-3 aided in developing a drug abuse lexicon that was aimed at enhancing the monitoring of drug abuse trends [76]. Notably, an LLM-based chatbot, called CLOVA CareCall, built by Naver [2], was applied as a health data–collecting tool in South Korea. Designed for individuals who need emotional support and are socially isolated, CareCall conducted periodic conversations, generating health reports with metrics such as meals, sleep, and emergencies. Implemented in 20 cities by May 2022, it targeted solitary adults, notably those with lower incomes and was proven effective in reducing loneliness. Social workers used the generated reports and call recordings to monitor users' health, resulting in positive feedback and a streamlined workload for public health workers.

Concerns

Most of the reviewed research papers pointed out technical and ethical concerns that people harbor with respect to the application of LLMs in health care from several perspectives. This can generally be categorized into four groups: (1) reliability, (2) bias, (3) privacy, and (4) public acceptance.

Reliability

Overview

The reliability of LLMs is essential to their application in health care. It can be related to the accuracy, consistency, and interpretability of LLM responses and the quality of the training dataset. Specifically, in 100% (22/22) of prediction-related studies, 72% (18/25) of summarization-related studies, and 93% (28/30) of studies related to medical knowledge inquiries, the authors pointed out their concerns toward LLM reliability (Table 1).

Accuracy

Several studies highlighted that ChatGPT exhibited inaccuracies when asked to respond to certain questions [19,20,24,27,29,33,37,45,50,55,57,59,63,64,68,75,76]. For instance, ChatGPT could respond with incomplete information or exhibit an inability to distinguish between truth and falsehood [26,71]. The generative nature of the LLM algorithms will likely fabricate a fake reference to substantiate false claims [18], a process that has been referred to as "hallucinations" [73]. In addition, such hallucinations can be communicated via persuasive prose [28], making it more likely to mislead patients. For example, Jo et al [2] mentioned that LLMs (specifically CLOVA CareCall built by NAVER in this paper) may make ambitious or impractical promises to patients, which may add extra burden to therapists or cause a trust crisis [2].

Data Quality

The unreliability of LLMs may be attributed to limitations in data collection sources [43,69]. There are concerns about the model's limitation in medical knowledge [61] because the general-purpose nature of ChatGPT may affect its reliability in self-diagnosis [3]. Recent state-of-the-art LLMs are typically constructed on texts from the internet rather than verified resources about health and medicine [1].

Of greater concern is the data availability. Health care institutions have shared no identifiable health information with widely accessible LLMs such as ChatGPT due to privacy concerns and legal compliances [7], and it is arduous to collect new data for LLM training [44]. ChatGPT, for example, was not trained on patients' clinical data [4]. While a description of a clinical scenario without sensitive patient information can be fed into ChatGPT through prompts, it may lead to inaccurate responses [4].

Another contributing factor to inaccuracy is the outdated knowledge base used to train LLMs [26,32,40,53]. ChatGPT based on GPT3.5 was pretrained using data collected until 2021 and does not support internet connection [43], making it unable to perform appropriately on questions regarding events that happened after 2021 [28].

Consistency

Many authors expressed concerns about the inconsistency of the responses from LLMs [26,32,40], where different answers result from various prompts of the same question

XSL•FO
RenderX

[23,27,29,33,54,68,69,73]. In addition, the output of ChatGPT to the same prompt may vary from user to user [23]. This is because LLMs generate responses in a probabilistic manner [1]. Therefore, nuance in the prompts to the LLM may lead to a completely different answer [23].

Interpretability

Interpretability is another aspect regarding the reliability of the response. A study by Cadamuro et al [34] highlights 2 key issues with an LLM (particularly ChatGPT) in health care. First, the interpretation of some normal results regarding suspected underlying diseases was not completely correct. Second, ChatGPT struggled to interpret all the coherent laboratory tests [34], generating superficial and incorrect responses. Indeed, ChatGPT could generate overly general answers without citing the original reference [22,28,60].

Bias

It has been noted that ChatGPT has issues with disparity and bias among different populations. In other words, because certain groups of people have financial, readability, and accessibility barriers using LLMs, their outcomes of using LLMs will be divergent from others. For example, ChatGPT may exert some financial disparity on the users: unlike previous versions such as GPT-3.5, access to GPT-4 involves a monthly fee [53]. These constraints potentially pose financial barriers, limiting widespread adoption and use of the newer, more advanced models in health care applications.

Moreover, the readability of an LLM's response may further accentuate health disparity [47]. LLMs such as ChatGPT include texts from scientific websites (eg, Wikipedia) as their training data, which makes their responses sound professional and sophisticated. However, LLMs may produce biased results [6,64], making regulations to prevent bias necessary [17,55].

Furthermore, the training data can also be biased. Since recent LLMs are trained based on human-generated texts from the internet, they also tend to provide biased answers [4]. Besides, algorithms may reinforce current health disparities and inequities [67]. Indeed, outputs from ChatGPT have been shown to be biased in terms of gender, race, and religion [4].

Privacy

Privacy issues are important when training or using LLMs in health care settings [6,7,64,77]. All AI systems, including LLMs in health settings, should comply with privacy regulations, including compliance with the Health Insurance Portability and Accountability Act, and implement robust safeguards to ensure the protection of sensitive patient information [6,7,64]. Specifically, LLMs have 3 privacy problems. First, the responses from LLMs may embed training examples directly, which breaches privacy if the training examples are identifiable. Second, LLMs may be susceptible to inferential disclosure. For example, a patient's membership in a dataset or sensitive attributes may be inferred from LLMs' responses. Third, it may not be clear whether text data are sufficiently deidentified for the anticipated recipients (which may be anyone in the world) when training LLMs. For instance, we may be able to deidentify text in a manner that sufficiently thwarts people who are not incentivized to attack the system, but we may not be addressing recipients who run machine-assisted attacks.

Public Acceptance

Public acceptance, the trust of the public in the application of LLMs in health care, has been mentioned in a study by Shahsavar and Choudhary [3]. A cross-sectional, survey-based study shows that 77.9% (371/476) participants claim that they trust ChatGPT's diagnosis, most of whom possess a bachelor's or even master's degree [3]. People are inclined to trust this new technique when using ChatGPT, partially due to the convenience of obtaining information and the patients' inclination to search for information [3].

Discussion

Principal Findings

This systematic review shows that LLMs have been applied to summarization, medical knowledge inquiry, prediction, and administration. At the same time, there are 4 major themes of concern when using these models in practice, including reliability, bias, privacy, and public acceptance. Specifically, the most popular application (30/65, 46% papers) for LLMs was for medical knowledge inquiries, with the second most popular (25/65, 38%) being summarization, followed by prediction (22/65, 34%), and then administration (9/65, 14%). At the same time, 55 (85%) papers expressed concerns about reliability, 16 (25%) about bias, 6 (9%) about privacy, and 4 (6%) about public acceptance.

Applications

According to our systematic review, LLMs were heavily applied in summarization and medical knowledge inquiry tasks. The former is probably due to the training method of LLMs, which focuses on their capability to summarize documents and paraphrase paragraphs. The latter is due to the inclusion of general medical knowledge in the training data. Specifically, in the category of summarization, summarizing medical notes is the type of task in which LLMs were applied the most. This is probably due to the simplicity of the task and the existence of redundancy in those notes. By contrast, in the genre of medical knowledge inquiry, taking standard medical exams is the type of task in which LLMs were applied the most. This is probably due to the existence of medical questions and answers on the internet that have been included in the training data of some LLMs, such as ChatGPT.

LLMs were applied in prediction tasks as well. Specifically, in the category of prediction, diagnosis is the type of task in which LLMs were applied but with the most reliability concerns. This is probably because diagnosis is a complex process in comparison with summarization and the current popular LLMs (eg, ChatGPT) used insufficient publicly available health datasets for model training. It might also be due to poorly constructed prompts without enough accurate information. Thus, LLMs are still not likely to be suitable for generating reliable answers to uncommon questions. In the category of administration, LLMs were applied equally heavily in various tasks, such as appointment scheduling, information collection, and documentation.

XSL•FO
RenderX

Concerns

For those applications of LLMs in health care, the 2 greatest concerns are reliability and bias (including disparity and inequality). These concerns might eventually drive this application away from practical implementation.

Notably, about 85% (55/65) of the reviewed studies emphasized concerns about the reliability of LLMs' responses given that they may impact a patient's health-related behavior. The concerns about reliability arose mainly from 2 aspects: the quality of the training data in terms of data source and data timeliness, and the models themselves in terms of their performance (eg, accuracy). For example, GPT-3.5 was pretrained using data collected by September 2021, and it also does not have access to private health records. Furthermore, most data that are used to train LLMs are crawled from the internet rather than professionally validated sources. In addition, the generative nature of LLM may result in seeming professional writing but fabricating responses. However, according to Shahsavar and Choudhury [3], people are inclined to trust this new technique, due partially to the convenience of obtaining information and the patients' inclination to search for information. By contrast, LLMs exhibit mixed predictive performance across different applications. In critical scenarios where incorrect predictions could lead to fatalities, even a 15% difference in accuracy from the gold standard (eg, 59/60, 98% vs 25/30, 83%) [4] could significantly hinder their use in real-world applications.

The issue of bias (or disparity) is mentioned in about 25% (16/65) of our included references. LLM biases come from the training stage (eg, biased training data and biased algorithms) and the application stage (eg, biased user base and biased outcomes). These papers discussed biases mainly from 3 different aspects: financial costs, readability, and accessibility. For example, Hirosawa et al [4] pointed out that the bias encoded in human-generated texts will make LLMs generate biased output; Lee et al [78] concerned that health disparity may result from low readability made by the sophistication of LLM wording; and Johnson et al [48] noted that LLM algorithms tend to reinforce the health disparity and to prevent LLM algorithms from exacerbating current disparity in health.

Another concern that prevents the wide application of LLMs in health care is privacy. When using third-party LLMs, such as ChatGPT, health care organizations face several privacy issues. Although no privacy breach of LLMs regarding patient information has been reported, attacks for other types of private information targeting ChatGPT have been found [79]. For example, a breach led to the exposure of users' conversations to unauthorized parties [79]. As ChatGPT interacts with patients directly, it may gather personal health information and may breach their privacy [7]. Therefore, many medical centers do not allow researchers and health care providers to use raw patient data as inputs to ChatGPT and other LLMs or even ban their access to these services during work [80]. Training or fine-tuning open-source LLMs requires a large amount of clinical data, which may lead to violations of patients' privacy, perhaps inadvertently [6,37,64].

Limitations of the Reviewed Papers

The reviewed papers demonstrated 2 common limitations of their approaches. First, almost all the studies relied on human experts to rate LLMs' responses. This is problematic because the score may be subjective and more likely unrepresentative. Correspondingly, future works can focus on designing a formal and fair process to evaluate LLMs' responses from a broad range of stakeholders, including researchers, health care providers, patients, or any users with diverse medical and sociodemographic backgrounds. Second, some of the concerns mentioned in this review (eg, bias) are merely researchers' speculations of the potential risks that were included to provide directions for further work. However, the mechanisms of how the training of LLMs leads to such concerns have not been comprehensively examined through experiments. It is suggested the audience should be wary of taking these concerns for granted or as proven facts.

Opportunities

Among all the included papers, few of them propose solutions to improve the reliability of LLMs. First, future research work should focus more on how to improve the accuracy of LLMs' responses in the health care domain. More specifically, domain-specific health data are demanded for training and fine-tuning of LLMs to improve the performance of LLMs in various tasks in the health care domain. Therefore, data harmonization and consortia established for LLM training are potential directions that can benefit the broad research community. Qualified medical professionals can contribute to the creation of the dataset for LLM training. This, however, will be expensive in terms of time and effort [2]. Alternatively, using retrieval-augmented generation to augment LLM with external knowledge that is up-to-date might be a solution for scenarios where accurate, in-depth professorial knowledge is required. Second, to prevent the hallucination issue, LLMs should be limited to making responses based on validated references. Blockchain technology can be used in this process to provide validation and traceability. Moreover, a holistic system, or a keep-experts-in-the-loop framework that efficiently facilitates the expert validation process becomes important to improve the accuracy and safety of health LLMs. Third, clinical trials based on health outcomes, such as mortality rates, should be conducted to validate the utility of LLM applications formally [1].

How conversational LLMs lead to bias or privacy issues in health care research was not thoughtfully examined with experiments in our reviewed papers. Future studies should first focus on investigating the mechanisms of how LLMs caused bias and privacy issues with stringent experiments and then developing practical solutions.

Regarding bias issues, it is suggested that systematic monitoring is necessary to ensure the impartial functioning of LLMs. However, all these sources discuss bias only with mere sentences and superficial summaries without any experimental investigation. Hence, it is worth noting that further work should also focus more on conducting experiments to understand how bias impacts the responses of LLMs in information, diagnosis, recommendation, and surveillance. More specifically, all

XSL•FO
RenderX

applications of LLMs in health care should be tested regarding the exhibitions of bias and the bias mitigation strategies, such as data augmentation and targeted recruitment (eg, the All of Us Research Program targets the collection of data from historically underrepresented populations [81]).

Regarding privacy issues, 2 technical approaches to mitigate the privacy risk while training LLMs are data anonymization [82] and synthetic data generation [83]. For deep learning models, model inversion attacks can potentially infer training data by giving model weights [84]. Considering the exponentially increased open-sourced LLMs with published model weights, a sensitive patient dataset needs to be deidentified [85] or replaced with a synthetic dataset before being used to train or fine-tune an LLM. Otherwise, the patients with whom the data are associated should be informed about their participation in the training or fine-tuning process [86]. To solve the privacy issues, legal, social, and technical protection approaches need to be implemented together to ensure the privacy and security of the whole process of training and using LLMs for health care applications [87].

To raise the public acceptance level of LLMs, explainable AI should be used to address the interpretability issues of LLMs by making the training data and model architecture transparent. More rigorous experimental studies using LLMs are encouraged in the "AI in medicine" research community to demonstrate or improve the reliability of LLM applications in health care. Moreover, stakeholders and decision-makers can propose new policies or regulations to manage the accountability and transparency of AI-generated content, including the responses from LLMs.

There appears to be research that is beginning to address some of these raised issues. For example, Zack et al [88] assessed the potential of GPT-4 to perpetuate racial and gender biases in health care. Hanna et al [89] assessed the racial bias of ChatGPT in health care–related text generation tasks. However, more research studies in these directions are needed to validate these findings and conduct more comprehensive and transparent assessments.

Furthermore, it is important to consider the interconnections among these categories of concerns. For instance, privacy protection methods may negatively affect the quality of training data and, consequently, the model's reliability due to the tradeoff between data utility and privacy [90-96]. In addition, tackling privacy issues can influence model fairness in different ways, depending on the approach used [93,97,98]. Therefore, we recommend addressing these concerns holistically rather than in isolation.

Moreover, almost all the research studies LLMs' responses in 1 language. For example, 95% (62/65) of studies focus on English, 1 (2%) focuses on Korean [2], 1 (2%) focuses on Chinese [62], and 1 (2%) focuses on Japanese [50]. Their findings cannot be extrapolated to other languages directly. Considering that many patients or people around the world or even in the United States do not speak English, it is necessary to guarantee that LLMs are usable universally or equitably and conduct more research to investigate the performance of LLMs in other languages.

Limitations of This Review

Despite notable findings, this review has several limitations. Firstly, the review used PubMed, ACM Digital Library, and IEEE Xplore as the primary sources for the papers. Other sources, such as Scopus, Web of Science, ScienceDirect, or non-English sources, may provide additional candidate papers regarding LLMs for health. However, because PubMed is the main digital library for medical publications, the research findings of this review should be valuable to health care researchers or policy makers. Second, although this review intended to study the application of state-of-the-art conversational LLMs in health care, most of the papers included are about ChatGPT. This is because ChatGPT is still the most powerful conversational LLM. However, its closed-source nature, which is against its company name—OpenAI—may be a hurdle to its wide application in health care, due primarily to the privacy concern when sharing sensitive patient information within prompts with OpenAI. Third, our search terms did not include any medicine-related keywords, which may have limited the number of papers included in this review. Finally, only peer-reviewed papers published before September 2023 are included in our review. Therefore, on one hand, the latest LLM application developments in this area are not included in this review. Specifically, papers focused on LLMs other than ChatGPT, such as LLaMA, were very limited in our initial keyword search results, and only a few of them are included in this review. This is a problem because, while monomodal conversational LLMs have been applied to many fields in health care, the multimodal LLMs that can process medical images, such as GPT-4, Large Language and Vision Assistant (LLaVA) [99] based on LLaMA, and LLaVA-Med [100] based on LLaVA, were just released before September 2023 and are still being examined by researchers regarding their capabilities in health care research. Therefore, no peer-reviewed research papers about applications of multimodal LLMs in health care have been published before September 2023. The main challenge of the application of multimodal LLMs in health care is that multimodal LLMs are still not perfect, either due to insufficient training data or due to insufficient model parameters. Specifically, with the development of computing power, reduced computing cost, and reduced data access cost, LLMs can be applied to multimedia-based diagnosis and analysis in radiology and other departments. By contrast, the latest studies addressing the concerns are not included in this review. Although there is research that is beginning to address some of the issues raised in the systematic review [13,89], there may not have been sufficient time for all recent papers to be deposited into the repositories upon which this investigation relied yet.

Conclusions

This review summarized applications of the state-of-the-art conversational LLMs in health care and the concerns that need to be resolved in the future. According to the reviewed research articles, conversational LLMs perform well in summarizing health-related texts, answering general questions in health care, and collecting information from patients. However, their performance is relatively less satisfying in making diagnoses and offering recommendations based on patients' symptoms and other information. Most authors were concerned about the

XSL•FO
RenderX

accuracy and consistency of the LLM responses, which should be the primary issues that researchers need to address in the near future. Nevertheless, other concerns regarding bias and privacy issues also prevent conversational LLMs from being broadly applied in the health care domain. However, these concerns still receive insufficient attention: few studies examine the bias and privacy issues in LLMs' health-related applications with rigorous scientific experiments. Future research should focus more on conducting such research to investigate the mechanisms of how the training and application of conversational LLMs leads to such concerns and to address these concerns that have been seen on any AI tools so that they can be safely applied in the health care domain.

Acknowledgments

This research was funded, in part, by the following grants from the National Institutes of Health (grants RM1HG009034 [EC, BM, and ZY] and R37CA237452 [BM and ZY]).

Data Availability

All data generated or analyzed during this study are included in this published article (Table 1).

Authors' Contributions

LW, ZW, and ZY conceived and designed the study and wrote the original draft. LW, ZW, CN, QS, YL, and ZY participated in paper screening and summarization. EC, BM, and ZY supervised the paper screening, summarization, and discussion. All authors wrote, read, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 100 KB - jmir_v26i1e22769_app1.pdf]

References

1. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Jo E, Epstein DA, Jung H, Kim YH. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3581503]
3. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
4. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
5. Anghelescu A, Firan FC, Onose G, Munteanu C, Trandafir AI, Ciobanu I, et al. PRISMA Systematic literature review, including with meta-analysis vs. chatbot/GPT (AI) regarding current scientific data on the main effects of the calf blood deproteinized hemoderivative medicine (Actovegin) in ischemic stroke. Biomedicines 2023 Jun 02;11(6):1623 [FREE Full text] [doi: 10.3390/biomedicines11061623] [Medline: 37371718]
6. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 2023 Jul 28;15(1):29 [FREE Full text] [doi: 10.1038/s41368-023-00239-y] [Medline: 37507396]
7. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
9. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak 2024 Mar 12;24(1):72 [FREE Full text] [doi: 10.1186/s12911-024-02459-6] [Medline: 38475802]
10. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing chat in supporting evidence-based

XSL•FO
RenderX

dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580 [FREE Full text] [doi: 10.2196/51580] [Medline: 38009003]
11. Puladi B, Gsaxner C, Kleesiek J, Hölzle F, Röhrig R, Egger J. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int J Oral Maxillofac Surg 2024 Jan;53(1):78-88 [FREE Full text] [doi: 10.1016/j.ijom.2023.09.005] [Medline: 37798200]
12. Pool J, Indulska M, Sadiq S. Large language models and generative AI in telehealth: a responsible use lens. J Am Med Inform Assoc 2024 Sep 01;31(9):2125-2136. [doi: 10.1093/jamia/ocae035] [Medline: 38441296]
13. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
14. Sarkis-Onofre R, Catalá-López F, Aromataris E, Lockwood C. How to properly use the PRISMA Statement. Syst Rev 2021 Apr 19;10(1):117 [FREE Full text] [doi: 10.1186/s13643-021-01671-z] [Medline: 33875004]
15. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
16. Akhter HM, Cooper JS. Acute pulmonary edema after hyperbaric oxygen treatment: a case report written with ChatGPT assistance. Cureus 2023 Feb;15(2):e34752 [FREE Full text] [doi: 10.7759/cureus.34752] [Medline: 36909067]
17. Almazyad M, Aljofan F, Abouammoh NA, Muaygil R, Malki KH, Aljamaan F, et al. Enhancing expert panel discussions in pediatric palliative care: innovative scenario development and summarization with ChatGPT-4. Cureus 2023 Apr;15(4):e38249 [FREE Full text] [doi: 10.7759/cureus.38249] [Medline: 37122982]
18. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus 2023 May;15(5):e39238 [FREE Full text] [doi: 10.7759/cureus.39238] [Medline: 37337480]
19. Bosbach WA, Senge JF, Nemeth B, Omar SH, Mitrakovic M, Beisbart C, et al. Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. Curr Probl Diagn Radiol 2024 Jan;53(1):102-110 [FREE Full text] [doi: 10.1067/j.cpradiol.2023.04.001] [Medline: 37263804]
20. Chen X, Zhang X, Liu Y, Wang Z, Zhou Y, Chu M. RISK-GPT: using ChatGPT to construct a reliable risk factor database for all known diseases. J Glob Health 2023 Aug 04;13:03037 [FREE Full text] [doi: 10.7189/jogh.13.03037] [Medline: 37539850]
21. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci Rep 2023 Mar 13;13(1):4164 [FREE Full text] [doi: 10.1038/s41598-023-31412-2] [Medline: 36914821]
22. Puthenpura V, Nadkarni S, DiLuna M, Hieftje K, Marks A. Personality changes and staring spells in a 12-year-old child: a case report incorporating ChatGPT, a natural language processing tool driven by artificial intelligence (AI). Cureus 2023 Mar;15(3):e36408 [FREE Full text] [doi: 10.7759/cureus.36408] [Medline: 37090271]
23. Robinson A, Aggarwal S. When precision meets penmanship: ChatGPT and surgery documentation. Cureus 2023 Jun;15(6):e40546. [doi: 10.7759/cureus.40546] [Medline: 37465809]
24. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus 2023 Apr;15(4):e37589 [FREE Full text] [doi: 10.7759/cureus.37589] [Medline: 37197105]
25. Guirguis CA, Crossley JR, Malekzadeh S. Bilateral vocal fold paralysis in a patient with neurosarcoidosis: a ChatGPT-driven case report describing an unusual presentation. Cureus 2023 Apr;15(4):e37368 [FREE Full text] [doi: 10.7759/cureus.37368] [Medline: 37181995]
26. Haemmerli J, Sveikata L, Nouri A, May A, Egervari K, Freyschlag C, et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Inform 2023 Jun 30;30(1):e100775 [FREE Full text] [doi: 10.1136/bmjhci-2023-100775] [Medline: 37399360]
27. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023 May 18;6(1):9 [FREE Full text] [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
28. Cunningham AR, Behm HE, Ju A, Peach MS. Long-term survival of patients with glioblastoma of the pineal gland: a ChatGPT-assisted, updated case of a multimodal treatment strategy resulting in extremely long overall survival at a site with historically poor outcomes. Cureus 2023 Mar;15(3):e36590 [FREE Full text] [doi: 10.7759/cureus.36590] [Medline: 37095802]
29. Golan R, Ripps SJ, Reddy R, Loloi J, Bernstein AP, Connelly ZM, et al. ChatGPT's ability to assess quality and readability of online medical information: evidence from a cross-sectional study. Cureus 2023 Jul;15(7):e42214 [FREE Full text] [doi: 10.7759/cureus.42214] [Medline: 37484787]
30. Readable homepage. Readable. URL: https://readable.com/ [accessed 2024-10-29]
31. Hamed E, Sharif A, Eid A, Alfehaidi A, Alberry M. Advancing artificial intelligence for clinical knowledge retrieval: a case study using ChatGPT-4 and link retrieval plug-in to analyze diabetic ketoacidosis guidelines. Cureus 2023 Jul;15(7):e41916 [FREE Full text] [doi: 10.7759/cureus.41916] [Medline: 37457604]

XSL•FO
RenderX

32. Grewal H, Dhillon G, Monga V, Sharma P, Buddhavarapu VS, Sidhu G, et al. Radiology gets chatty: the ChatGPT saga unfolds. Cureus 2023 Jun;15(6):e40135 [FREE Full text] [doi: 10.7759/cureus.40135] [Medline: 37425598]
33. Kumari KS, Anusha KS. An esthetic approach for rehabilitation of long-span edentulous arch using artificial intelligence. Cureus 2023 May;15(5):e38683 [FREE Full text] [doi: 10.7759/cureus.38683] [Medline: 37292565]
34. Cadamuro J, Cabitza F, Debeljak Z, De Bruyne S, Frans GM, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med 2023 Jun 27;61(7):1158-1166 [FREE Full text] [doi: 10.1515/cclm-2023-0355] [Medline: 37083166]
35. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023 Jul 07;619(7969):357-362. [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
36. Sharma B, Gao Y, Miller T, Churpek M, Afshar M, Dligach D. Multi-task training with in-domain language models for diagnostic reasoning. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. 2023 Presented at: ClinicalNLP 2023; July 14, 2023; Toronto, ON. [doi: 10.18653/v1/2023.clinicalnlp-1.10]
37. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
38. Hamed E, Eid A, Alberry M. Exploring ChatGPT's potential in facilitating adaptation of clinical guidelines: a case study of diabetic ketoacidosis guidelines. Cureus 2023 May;15(5):e38784 [FREE Full text] [doi: 10.7759/cureus.38784] [Medline: 37303347]
39. Kim HY. A case report on ground-level alternobaric vertigo due to eustachian tube dysfunction with the assistance of conversational generative pre-trained transformer (ChatGPT). Cureus 2023 Mar;15(3):e36830 [FREE Full text] [doi: 10.7759/cureus.36830] [Medline: 37123797]
40. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023 Feb 17;13:01003 [FREE Full text] [doi: 10.7189/jogh.13.01003] [Medline: 36798998]
41. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
42. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg 2023;39(3):221-225 [FREE Full text] [doi: 10.1097/IOP.0000000000002418] [Medline: 37166289]
43. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
44. Bird JJ, Lotfi A. Generative transformer chatbots for mental health support: a study on depression and anxiety. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. 2023 Presented at: PETRA '23; July 5-7, 2023; Corfu, Greece. [doi: 10.1145/3594806.3596520]
45. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023 Sep;280(9):4271-4278 [FREE Full text] [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
46. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023 Jul 17;13:1219326 [FREE Full text] [doi: 10.3389/fonc.2023.1219326] [Medline: 37529688]
47. Hristidis V, Ruggiano N, Brown EL, Ganta SR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 2023 Jul 25;25:e48966 [FREE Full text] [doi: 10.2196/48966] [Medline: 37490317]
48. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad015 [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
49. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
50. Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese Society of Hypertension Guidelines. Circ J 2023;87(7):1030-1033. [doi: 10.1253/circj.cj-23-0308]
51. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics (Basel) 2023 Jun 02;13(11):1950 [FREE Full text] [doi: 10.3390/diagnostics13111950] [Medline: 37296802]

XSL•FO
RenderX

52. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895 [FREE Full text] [doi: 10.7759/cureus.40895] [Medline: 37492832]
53. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023 Jun;15(6):e40822 [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
54. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. JMIR Med Educ 2023 Jul 10;9:e46939 [FREE Full text] [doi: 10.2196/46939] [Medline: 37428540]
55. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. Cureus 2023 Feb;15(2):e35029 [FREE Full text] [doi: 10.7759/cureus.35029] [Medline: 36819954]
56. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023 Feb;15(2):e35237 [FREE Full text] [doi: 10.7759/cureus.35237] [Medline: 36968864]
57. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
58. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs 2024 Jan 12;23(1):95-98. [doi: 10.1093/eurjcn/zvad038] [Medline: 37094282]
59. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. Can Assoc Radiol J 2024 Feb 20;75(1):69-73 [FREE Full text] [doi: 10.1177/08465371231171125] [Medline: 37078489]
60. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023 Jun 30;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
61. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul 22;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
62. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-English-speaking outpatient clinic settings. Digit Health 2023 Jun 26;9:20552076231184091 [FREE Full text] [doi: 10.1177/20552076231184091] [Medline: 37434733]
63. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus 2023 Jun;15(6):e40351 [FREE Full text] [doi: 10.7759/cureus.40351] [Medline: 37456381]
64. Goktas P, Karakaya G, Kalyoncu AF, Damadoglu E. Artificial intelligence chatbots in allergy and immunology practice: where have we been and where are we going? J Allergy Clin Immunol Pract 2023 Sep;11(9):2697-2700 [FREE Full text] [doi: 10.1016/j.jaip.2023.05.042] [Medline: 37301435]
65. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril 2023 Sep;120(3 Pt 2):575-583 [FREE Full text] [doi: 10.1016/j.fertnstert.2023.05.151] [Medline: 37217092]
66. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health 2022 Dec 22;1(12):e0000168 [FREE Full text] [doi: 10.1371/journal.pdig.0000168] [Medline: 36812634]
67. Karkera N, Acharya S, Palaniappan SK. Leveraging pre-trained language models for mining microbiome-disease relationships. BMC Bioinformatics 2023 Jul 19;24(1):290 [FREE Full text] [doi: 10.1186/s12859-023-05411-z] [Medline: 37468830]
68. Sarbay İ, Berikol GB, Özturan IU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023;23(3):156-161 [FREE Full text] [doi: 10.4103/tjem.tjem_79_23] [Medline: 37529789]
69. Galido PV, Butala S, Chakerian M, Agustines D. A case study demonstrating applications of ChatGPT in the clinical management of treatment-resistant schizophrenia. Cureus 2023 Apr;15(4):e38166 [FREE Full text] [doi: 10.7759/cureus.38166] [Medline: 37252576]
70. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023 May 30;9(1):44 [FREE Full text] [doi: 10.1038/s41523-023-00557-8] [Medline: 37253791]
71. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus 2023 Mar;15(3):e36272 [FREE Full text] [doi: 10.7759/cureus.36272] [Medline: 37073184]
72. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023 May 01;307(4):e230424. [doi: 10.1148/radiol.230424] [Medline: 37014239]

XSL•FO
RenderX

73. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
74. Kao HJ, Chien TW, Wang WC, Chou W, Chow JC. Assessing ChatGPT's capacity for clinical decision support in pediatrics: a comparative study with pediatricians using KIDMAP of Rasch analysis. Medicine (Baltimore) 2023 Jun 23;102(25):e34068 [FREE Full text] [doi: 10.1097/MD.0000000000034068] [Medline: 37352054]
75. Schulte B. Capacity of ChatGPT to identify guideline-based treatments for advanced solid tumors. Cureus 2023 Apr;15(4):e37938 [FREE Full text] [doi: 10.7759/cureus.37938] [Medline: 37220429]
76. Carpenter KA, Altman RB. Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. Biomolecules 2023 Feb 18;13(2):387 [FREE Full text] [doi: 10.3390/biom13020387] [Medline: 36830756]
77. Montagna S, Ferretti S, Klopfenstein LC, Florio A, Pengo M. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. In: Proceedings of the 2023 ACM Conference on Information Technology for Social Good. 2023 Presented at: GoodIT '23; September 6-8, 2023; Lisbon, Portugal. [doi: 10.1145/3582515.3609536]
78. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT answers common patient questions about colonoscopy. Gastroenterology 2023 Aug;165(2):509-11.e7 [FREE Full text] [doi: 10.1053/j.gastro.2023.04.033] [Medline: 37150470]
79. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access 2023;11:80218-80245. [doi: 10.1109/access.2023.3300381]
80. Nelson F. Many companies are banning ChatGPT. This is why. Science Alert. 2023 Jun 16. URL: https://www.sciencealert.com/many-companies-are-banning-chatgpt-this-is-why [accessed 2024-04-20]
81. The All of Us Research Program Investigators. The “All of Us” research program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/nejmsr1809937]
82. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015 Mar 20;350:h1139 [FREE Full text] [doi: 10.1136/bmj.h1139] [Medline: 25794882]
83. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 2021 Jun 15;5(6):493-497 [FREE Full text] [doi: 10.1038/s41551-021-00751-8] [Medline: 34131324]
84. Zhang Y, Jia R, Pei H, Wang W, Li B, Song D. The secret revealer: generative model-inversion attacks against deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 Presented at: CVPR 2020; June 13-19, 2020; Seattle, WA. [doi: 10.1109/cvpr42600.2020.00033]
85. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071 [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
86. Cohen IG. What should ChatGPT mean for bioethics? Am J Bioeth 2023 Oct 13;23(10):8-16. [doi: 10.1080/15265161.2023.2233357] [Medline: 37440696]
87. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
88. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22 [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
89. Hanna JJ, Wakene AD, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT. medRxiv Preprint posted online on August 28, 2023 [FREE Full text] [doi: 10.1101/2023.08.28.23294730] [Medline: 37693388]
90. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Ganta R, et al. A game theoretic framework for analyzing re-identification risk. PLoS One 2015 Mar 25;10(3):e0120592 [FREE Full text] [doi: 10.1371/journal.pone.0120592] [Medline: 25807380]
91. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B. Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. Am J Hum Genet 2017 Feb 02;100(2):316-322 [FREE Full text] [doi: 10.1016/j.ajhg.2016.12.002] [Medline: 28065469]
92. Wan Z, Vorobeychik Y, Kantarcioglu M, Malin B. Controlling the signal: practical privacy protection of genomic data sharing through Beacon services. BMC Med Genomics 2017 Jul 26;10(Suppl 2):39 [FREE Full text] [doi: 10.1186/s12920-017-0282-1] [Medline: 28786360]
93. Wan Z, Vorobeychik Y, Xia W, Liu Y, Wooders M, Guo J, et al. Using game theory to thwart multistage privacy intrusions when sharing data. Sci Adv 2021 Dec 10;7(50):eabe9986 [FREE Full text] [doi: 10.1126/sciadv.abe9986] [Medline: 34890225]
94. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet 2022 Jul 04;23(7):429-445 [FREE Full text] [doi: 10.1038/s41576-022-00455-y] [Medline: 35246669]
95. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, et al. A multifaceted benchmarking of synthetic electronic health record generation models. Nat Commun 2022 Dec 09;13(1):7609 [FREE Full text] [doi: 10.1038/s41467-022-35295-1] [Medline: 36494374]

XSL•FO
RenderX

96. Venkatesaramani R, Wan Z, Malin BA, Vorobeychik Y. Enabling tradeoffs in privacy and utility in genomic data Beacons and summary statistics. Genome Res 2023 Jul 22;33(7):1113-1123 [FREE Full text] [doi: 10.1101/gr.277674.123] [Medline: 37217251]
97. Brown JT, Clayton EW, Matheny M, Kantarcioglu M, Vorobeychik Y, Malin BA. Robin Hood: a de-identification method to preserve minority representation for disparities research. In: Proceedings of the International Conference on Privacy in Statistical Databases. 2024 Presented at: PSD 2024; September 25-27, 2024; Antibes Juan-les-Pins, France. [doi: 10.1007/978-3-031-69651-0_5]
98. Zhang T, Zhu T, Gao K, Zhou W, Yu PS. Balancing learning model privacy, fairness, and accuracy with early stopping criteria. IEEE Trans Neural Netw Learn Syst 2023 Sep;34(9):5557-5569. [doi: 10.1109/tnnls.2021.3129592]
99. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv Preprint posted online on April 17, 2023 [FREE Full text]
100. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv Preprint posted online on June 1, 2023 [FREE Full text]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
LLaVA: Large Language and Vision Assistant
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 23.04.24; peer-reviewed by MJ Pinjar, A Kavadella, J Xia, S Sengupta, C Zielinski; comments to author 09.09.24; revised version received 19.09.24; accepted 03.10.24; published 07.11.24.

Please cite as:
Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, Malin B, Yin Z
Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review
J Med Internet Res 2024;26:e22769
URL: https://www.jmir.org/2024/1/e22769
doi:10.2196/22769
PMID:

©Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, Zhijun Yin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study

Dingqiao Wang[1], MD, PhD; Jiangbo Liang[1], MD; Jinguo Ye[1], BSc; Jingni Li[1], MD; Jingpeng Li[1], PhD; Qikai Zhang[1], BSc; Qiuling Hu[1], BSc; Caineng Pan[1], BSc; Dongliang Wang[1], MD, PhD; Zhong Liu[1], MD; Wen Shi[1], PhD; Danli Shi[2], MD, PhD; Fei Li[1*], MD, PhD; Bo Qu[3*], MD, PhD; Yingfeng Zheng[1*], MD, PhD

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, GuangZhou, China

[2]Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Hong Kong, China

[3]Peking University Third Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yingfeng Zheng, MD, PhD
State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University
Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science
Guangdong Provincial Clinical Research Center for Ocular Diseases
State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center
Sun Yat-sen University, 07 Jinsui Road
GuangZhou, 510060
China
Phone: 86 139 2228 6455
Email: zhyfeng@mail.sysu.edu.cn

Abstract

Background: Large language models (LLMs) demonstrated advanced performance in processing clinical information. However, commercially available LLMs lack specialized medical knowledge and remain susceptible to generating inaccurate information. Given the need for self-management in diabetes, patients commonly seek information online. We introduce the Retrieval-augmented Information System for Enhancement (RISE) framework and evaluate its performance in enhancing LLMs to provide accurate responses to diabetes-related inquiries.

Objective: This study aimed to evaluate the potential of the RISE framework, an information retrieval and augmentation tool, to improve the LLM's performance to accurately and safely respond to diabetes-related inquiries.

Methods: The RISE, an innovative retrieval augmentation framework, comprises 4 steps: rewriting query, information retrieval, summarization, and execution. Using a set of 43 common diabetes-related questions, we evaluated 3 base LLMs (GPT-4, Anthropic Claude 2, Google Bard) and their RISE-enhanced versions respectively. Assessments were conducted by clinicians for accuracy and comprehensiveness and by patients for understandability.

Results: The integration of RISE significantly improved the accuracy and comprehensiveness of responses from all 3 base LLMs. On average, the percentage of accurate responses increased by 12% (15/129) with RISE. Specifically, the rates of accurate responses increased by 7% (3/43) for GPT-4, 19% (8/43) for Claude 2, and 9% (4/43) for Google Bard. The framework also enhanced response comprehensiveness, with mean scores improving by 0.44 (SD 0.10). Understandability was also enhanced by 0.19 (SD 0.13) on average. Data collection was conducted from September 30, 2023 to February 5, 2024.

Conclusions: The RISE significantly improves LLMs' performance in responding to diabetes-related inquiries, enhancing accuracy, comprehensiveness, and understandability. These improvements have crucial implications for RISE's future role in patient education and chronic illness self-management, which contributes to relieving medical resource pressures and raising public awareness of medical knowledge.

(J Med Internet Res 2024;26:e58041) doi:10.2196/58041

XSL•FO
RenderX

KEYWORDS

large language models; LLMs; retrieval-augmented generation; RAG; GPT-4.0; Claude-2; Google Bard; diabetes education

Introduction

Background

Diabetes mellitus is a chronic long-term illness that requires continual health education and assistance to improve patient outcomes [1,2]. The shortage of diabetes counselors and the limitations of traditional education methods make it challenging to address the unique requirements of each diabetic patient [3]. Large language models (LLMs), such as ChatGPT, hold significant promise in diabetes self-management and information assessment [3-8]. However, concerns exist around the accuracy and reliability of these models, mainly stemming from the variable credibility of their training data which is sourced from a wide variety of internet text and self-supervised learning [9-11]. Furthermore, LLMs may lack domain-specific knowledge, risking the production of potentially inaccurate responses [12-15].

Recent studies have primarily assessed the capabilities of LLMs in responding to diabetes-related questions, revealing limitations in their expertise in medical specialties, which remain unresolved. For example, Meo et al [16] indicated both ChatGPT and Google Bard scored below 60% in endocrinology and diabetes. They concluded that while these artificial intelligence tools show potential in academic medical education, they require more updated information in these specific medical fields. Goodman Rachel et al [17] also highlighted the precision of chatbots in medical queries and underlined the need for further research and model development for enhanced accuracy and validation in clinical practice. Hulman et al [18] showed that ChatGPT-generated responses could be distinguished from expert responses by 59.5%, suggesting a gap compared with expert human performance. Therefore, addressing these gaps by augmenting LLMs with more specialized knowledge and updated information is crucial for improving their role in patient understanding and management of diabetes.

In response to these unresolved challenges, our study introduces "RISE" (Retrieval-augmented Information System for Enhancement), an independent workflow designed to enhance the performance of LLMs in the medical domain by automatically retrieving real-time external knowledge. We used LLMs with and without RISE to answer diabetes-related inquiries from patients, assessing the improvements that RISE brings to the original LLMs in terms of accuracy, comprehensiveness, and understandability. Our RISE aims to bridge the knowledge gaps identified in LLMs, providing a more robust and reliable tool for addressing patient concerns about diabetes management and understanding.

The main contributions of our work are (1) we introduce RISE, an innovative framework based on the retrieval-augmented generation (RAG) algorithm that enhances LLMs with real-time, domain-specific knowledge to provide accurate and comprehensive responses to diabetes-related inquiries, improving patient self-management and outcomes. (2) We reduce the risk of inaccurate or irrelevant responses from LLMs by integrating local and external real-time information retrieval, enhancing model transparency by identifying source information. (3) We incorporate an additional module for accuracy and safety checks before responding, ensuring that the provided information is reliable and free from harmful content. (4) We validate the RISE framework through assessments by clinicians and patients to demonstrate the feasibility of adopting RISE-enhanced LLMs in diabetes management and education.

Related Works

Large Language Models

LLMs, such as GPT-3 [19], GPT-4 [20], and PaLM [21], have garnered significant attention due to their exceptional language understanding and generation capabilities [22,23]. However, when applied to domain-specific tasks, particularly in the medical field, their performance may be limited by a lack of specialized knowledge and vocabulary in the training data [24-26]. Adapting LLMs for biomedical applications presents several challenges, including insufficient domain knowledge and high computational costs. As a result, only a few LLMs have been fine-tuned for medical consultation using open-source models with 6.5-13 billion parameters, such as ChatDoctor [27] and MedAlpaca [28]. However, this approach of fine-tuning open-source models has its limitations. Medical domain-specific models often use relatively smaller-scale LLMs (eg, LLaMA [27] with 7B parameters), which may result in lower accuracy and robustness, compared with GPT-4 [29]. Moreover, fine-tuning even these smaller LLMs is computationally intensive and costly [27]. The introduction of new knowledge requires complete retraining of the model, placing additional burdens on developers. Furthermore, LLMs are generally prone to hallucination, which is a challenge that fine-tuning struggles to address [17,30-33].

Retrieval-Augmented Generation in Medical Questions and Answers

Recent studies have explored the application of RAG [34,35] in the medical domain to enhance the performance of LLMs in question-answering tasks. These approaches enable LLMs to achieve improved performance without needing time-intensive and costly fine-tuning while facilitating timely updates without retraining the entire model.

In specialized medical domains, LLMs have been augmented with limited medical corpora to address specific areas such as liver diseases (LiVersa) [36], diffuse large B-cell lymphoma [32], and nephrology [37]. Simultaneously, in the general medical context, frameworks such as Almanac [38] and RECTIFIER [39] have been proposed to integrate LLMs with medical guidelines and treatment recommendations.

Despite their potential, these approaches also present several limitations. The effectiveness of RAG-based models largely depends on the quality and currency of the used data sources. The previous studies typically rely on fixed and related smaller knowledge bases, such as Wikipedia or guidance documents, thereby limiting their effectiveness in specialized medical

XSL•FO
RenderX

domains [40,41]. Outdated or incorrect information can result in inaccurate or misleading outputs. Furthermore, retrieval errors or the inclusion of biased and unsafe content inherent in the LLMs, without further filtering, may lead to inaccuracies in the generated output, potentially misleading patients.

Our RISE framework addresses these limitations by comminating with local and internet-based knowledge sources, curated from over 200 reputable academic websites, ensuring access to a wide range of up-to-date clinical evidence. Moreover, we incorporate additional fact-checking and safety check modules before responding. By prioritizing the accuracy and safety of the retrieved information, our framework offers a more reliable and secure pathway for answering clinical questions, significantly reducing the risk of misleading patients.

Methods

Framework of Retrieval-Augmented Information System for Enhancement

Our study introduced the RISE framework, an innovative approach designed to improve the performance of medical question answering of LLMs. Our novel algorithm derives from RAG [34,42,43], which retrieves pertinent information from local databases or external knowledge from academic websites. Our RISE is a standalone framework comprising four steps (Figure 1 and Multimedia Appendix 1).

Figure 1. Comparing responses of LLMs before and after "RISE" integration. Red bars: response from base LLMs without RISE framework. Blue bars: overview of RISE framework and query response after integration with RISE. The framework of RISE: (1) Rewriting query: Improve query accuracy and relevance using large language models. (2) Information retrieval: Search for relevant information using the revised query from the local data set and external knowledge base. (3) Summarization: Summarize retrieved information into concise key points, combined with fact-checking and safety checks. (4) Execution: LLMs take action based on summarized information (for implementation details, refer to Multimedia Appendix 1). LLMs: large language models; RISE: Retrieval-augmented Information System for Enhancement.

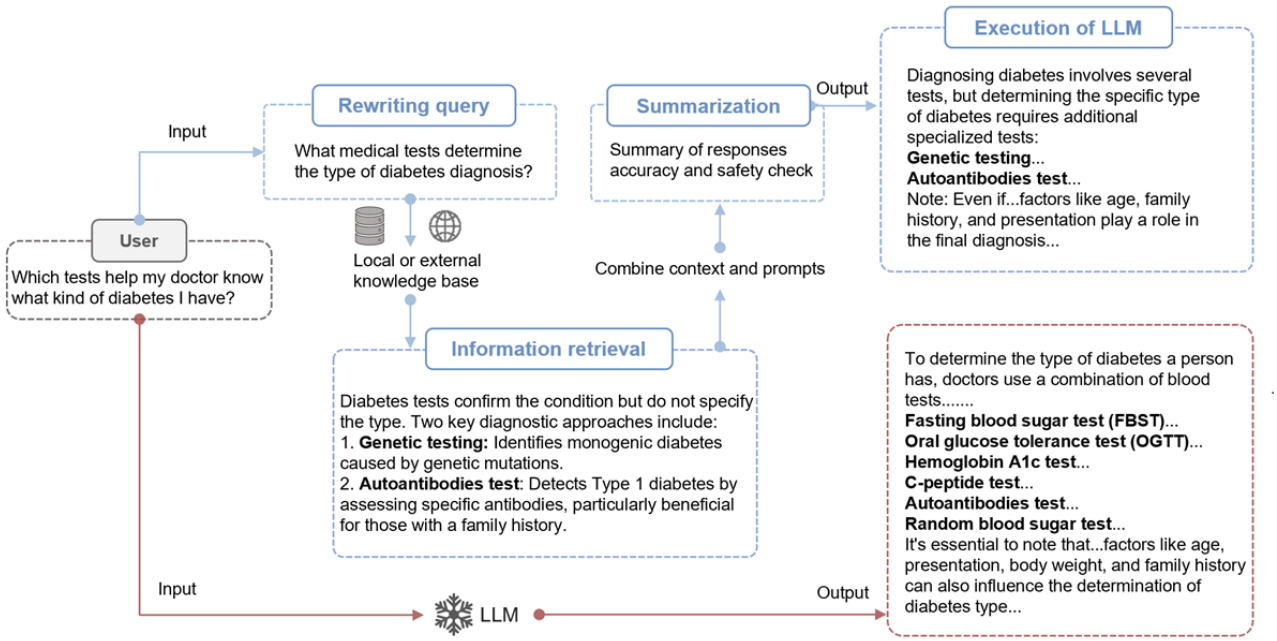

Rewriting Query

The first step in the RISE framework involves rewriting the original query using advanced LLMs, including GPT-4, Claude 2, and Google Bard (Alphabet Inc; subsequently rebranded as Gemini). This process aims to enhance the query by correcting spelling errors, expanding abbreviations, and incorporating synonyms, thereby broadening the scope of potential results.

Information Retrieval

Relevant information is retrieved from a local vectorized database and external knowledge sources. The rewritten query is embedded in the same vector space as the database, and the Facebook AI Similarity Search (FAISS) algorithm is used for similarity search to find the top 5 most pertinent documents (retriever=vectorstore.as_retriever (search_type="similarity," search_kwargs={"k:" 5}), results = retriever.invoke [query]). If no results are found locally, external knowledge is sourced from academic websites (over 200), ensuring that all information adheres to stringent academic and research standards.

Summarization

The third step involves summarizing the retrieved information into a concise and understandable format by prompt. This step also includes fact-checking and safety checks to ensure accuracy and reduce harmful content.

Fact-checking is performed in 2 steps. First, the retrieved raw text and the question are input, and the retrieved text is broken down into multiple claims. Second, these claims and the question are input, allowing the model to self-check which claims are confirmed using external knowledge sources. The model then returns the verified claims as the final summarization text.

XSL•FO
RenderX

The safety check process uses a set of 24 rules to restrict and filter the content, ensuring the generated responses are safe and appropriate. The model is prompted with the instruction, that is "Your answer must adhere to the following rules: {rules}."

Execution

The final step involves presenting the summarized information and prompts to the LLMs to generate the final answer for the user. The prompt instructs "Use the following pieces of context to answer the question at the end. Note that your response should be as brief as possible and no more than 300 words. If you don't know the answer, just say that you don't know, don't try to make up an answer."

Local Database

A local database of diabetes-related information was created to provide domain-specific knowledge for the RISE framework. PubMed Central [44] was used to acquire a corpus of scientific papers and clinical practice guidelines relevant to diabetes. The database covers various aspects of diabetes, including pathophysiology, diagnosis, treatment, management, and patient education, rather than answering specific questions used in the evaluation. The retrieved documents comprise over 600 full-text articles.

The retrieved documents were then preprocessed to remove potentially unstructured or noisy information, such as figures, tables, references, and author disclosures. After cleaning each document, the CharacterTextSplitter function from Langchain was used to divide the documents into smaller fragments. We then used the OpenAI model Text-Embedding-ADA-002 as an embedding function to generate embeddings for each fragment in FAISS using the function "db=FAISS.from_documents (docs, embeddings)," where "docs" refers to the document fragments and "embeddings" refers to the Text-Embedding-ADA-002 model. The resulting index was saved locally for continuous access and retrieval using the function "db.save_local ("faiss_index")."

When a user submitted a question, the rewriting query was transformed into an embedding vector and compared with the database embeddings using cosine similarity. The top k=5 document segments with the highest similarity scores were retrieved and used as the knowledge context for the user's query. A sample of the data set and related code are provided in Multimedia Appendices 1 and 2.

Study Design

Overview

This study was conducted at Sun Yat-sen University from September 25, 2023, to February 30, 2024. The 43 diabetes-related questions were selected from the National Institute of Diabetes and Digestive and Kidney Diseases website [45] across the following 5 domains, that are concepts of diabetes, symptoms and causes, diabetes tests and diagnosis, managing diabetes, and prevention. The questions aimed to cover topics commonly asked by the public and patients regarding diabetes care.

Respond Generation

We prepared a set of 43 diabetes-related inquiries to be posed to 3 base language models - GPT-4, Claude 2, and Google Bard - as well as their respective versions enhanced by RISE. In total, there were 6 models involved, with an enhanced version corresponding to each base model. From September 30, 2023, to February 5, 2024, we independently fed the entire set of 43 inquiries into each of the 6 models, treating each question as a separate input and resetting the conversation between queries to minimize bias.

Model responses were evaluated in a blinded, randomized manner through 2 aspects - first by clinician assessment focusing on accuracy and comprehensiveness, and then by patient evaluation of understandability. The evaluation process involved clinicians with over 5 years of experience in general medicine. The responses from all 6 models were shuffled randomly into 6 different rounds. To remove potential model indicators from responses, they were transformed into plain text before being distributed to 3 clinicians and 3 diabetes patients. They all analyzed responses across 6 rounds spaced 48 hours apart to eliminate confounding (Figure 2). Responses for each model and raw scores for evaluation are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 2. Flowchart of overall study design. The study evaluates the performance of 3 publicly available large language models and their RISE-enhanced versions in addressing common diabetes-related inquiries. The evaluation is conducted from the perspectives of both the clinicians and diabetic patients. Clinicians evaluate the accuracy and comprehensiveness of responses. Patients assess the understandability.

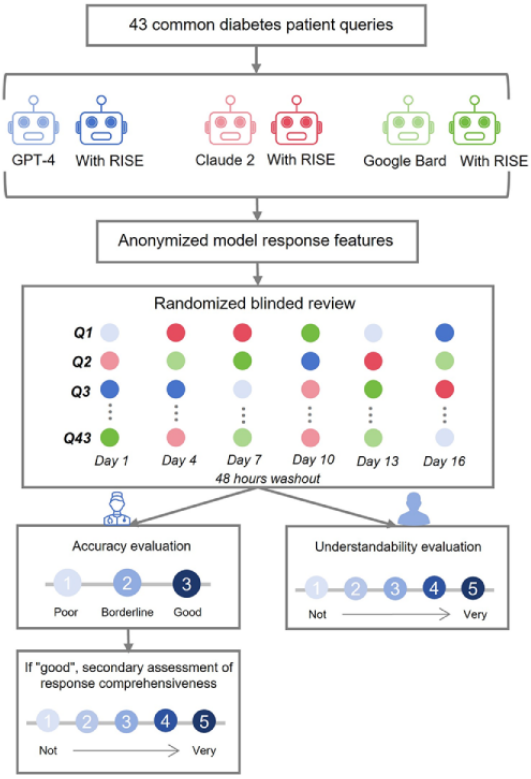

Accuracy Evaluation

We conducted an accuracy evaluation for each response by assigning scores and ratings. A “Poor” rating received 1 point, “Borderline” received 2 points, and “Good” received 3 points. Each response underwent assessment by 3 clinicians. For scoring, the score for each question is the sum of the score assigned by the 3 graders, with a maximum possible score of 9. For rating, we used a majority consensus method among the 3 clinicians. A response was considered “Good” only if more than 2 clinicians rated it as such. In cases where the 3 clinicians provided differing ratings, we implemented a stringent strategy by giving the response the lowest mark (ie, “poor”). The accuracy rate is defined as the proportion of responses with a final rating of “Good.”

The accuracy scoring criteria include (1) “Poor” indicating replies containing mistakes that might considerably mislead patients and potentially result in damage, (2) “Borderline” assigned to answers with potential inaccuracies but unlikely to misguide or damage patients, and (3) “Good” reserved for replies without errors.

Comprehensiveness Evaluation

For responses that obtained a “good” rating by majority consensus, the clinicians further evaluated the comprehensiveness of each response. A 5-point scale was used (1) “not comprehensive” for reactions critically missing information (1 point), (2) “slightly comprehensive” for replies with limited but primary details (2 points), (3) “moderately comprehensive” for reactions providing more than half of the essential information (3 points), (4) “comprehensive” for reactions covering most critical points (4 points), and (5) “very comprehensive” for reactions giving comprehensive information (5 points). For each response, the average score was calculated by the mean of the scores assigned by the 3 clinicians.

Understandability Evaluation

A total of 3 diabetic patients evaluated response understandability. A 5-point scale different from comprehensiveness evaluation was used; (1) “very poor” for responses difficult to understand or completely irrelevant (1 point), (2) “poor” for responses somewhat difficult to understand or partially irrelevant (2 points), (3) “average” for responses generally understandable but requiring some effort or having minor ambiguities (3 points), (4) “good” for responses most of which are easily understandable with very few unclear parts (4 points), and (5) “Excellent” for responses very clear and easy to understand, fully meeting the reader’s needs (5 points). For each response, the average score for understandability was calculated based on the score given by each patient.

Statistical Analysis

Statistical analyses were used SPSS (version 22.0; IBM Corp). Normal distribution was assessed with Kolmogorov–Smirnov test. Our data were found not to follow a normal distribution, $P<.001$. Group comparisons used the Wilcoxon signed-rank test for accuracy, comprehensiveness, and understandability scores with and without RISE. For the comparison of the proportions of “good,” “borderline,” and “poor” ratings across the models, the chi-square test was used. P values <.05 were regarded as significant.

XSL•FO
RenderX

Ethical Considerations

This study involved publicly available data without collecting human or animal samples and data. The study has been approved by the ethics committee of Zhongshan Ophthalmic Center (ZOC, Guangzhou, China; approval number 2024KYPJ124).

Results

Accuracy Evaluation

We evaluated 3 LLMs and their RISE-enhanced versions for answering diabetes-related questions. As shown in Figure 3, the average accuracy scores of all 3 original models increased substantially with the RISE enhancement. Specifically, the accuracy scores improved from 8.72 (SD 0.70) to 8.91 (SD 0.37; *P*=.09) for GPT-4 after applying RISE, 8.09 (SD 1.23) to 8.65 (SD 0.65; *P*=.001) for Claude, and 8.37 (SD 1.36) to 8.86 (SD 0.47; *P*=.01) for Bard (maximum score per response is 9 points).

Figure 3. Average scores of responses from large language models. Answers from each model were scored 1-3 points by 3 clinicians. The maximum score for each response is 9 points. An asterisk (*) denotes statistical significance at *P*<.05. Model call dates: September 30, 2023 to February 5, 2024. RISE: Retrieval-augmented Information System for Enhancement.

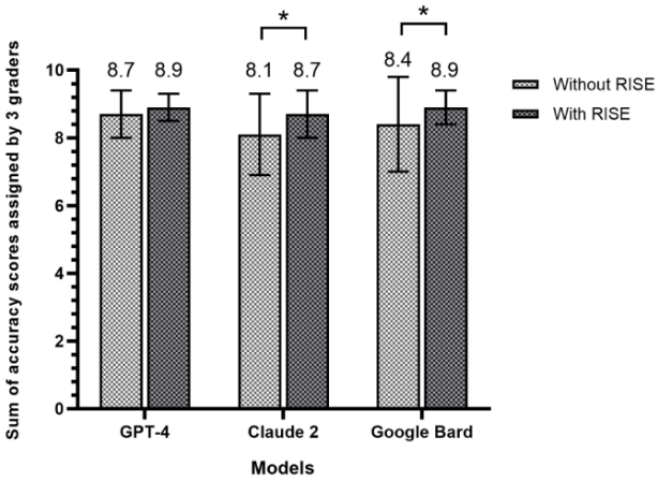

We further evaluated the percentage rated as "Good", representing accuracy rates, of the LLMs with and without RISE (Figure 4). The results showed increased accuracy rates after incorporating RISE across all original models. Specifically, after the incorporation of RISE, the proportion of accuracy responses for GPT-4 increased from 91% (39/43) to 98% (42/43), for Claude from 72% (31/43) to 91% (39/43), and for Bard from 86% (37/43) to 95% (41/43). Furthermore, GPT-4 enhanced by RISE exhibited the highest accuracy rates, reaching 98% (42/43). In addition, Table 1 presents the accuracy of the models across 5 domains. All 6 models achieved the highest accuracy, reaching 100% (16/16), in the "Preventing Diabetes Problems" domain. However, in the "Concepts of Diabetes" and "Symptoms & Causes" domains, the models exhibited relatively lower average accuracy rates.

Figure 4. Accuracy rates (proportion of "Good" responses) of large language models. Model call dates: September 30, 2023 to February 5, 2024. RISE: Retrieval-augmented Information System for Enhancement.

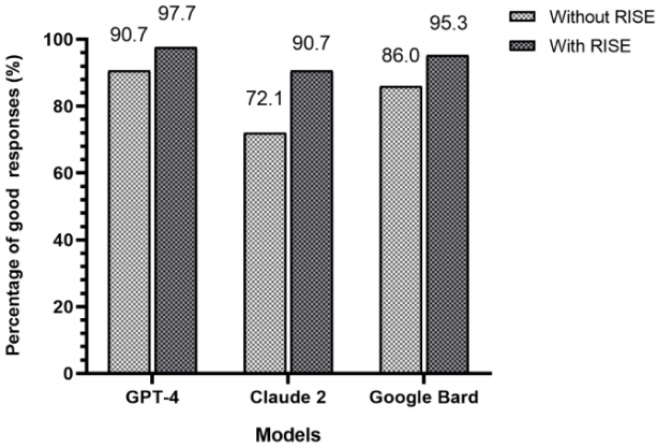

XSL•FO
RenderX

Table 1. Accuracy of model response across 5 diabetes educational domains.

Domain	Responses, n	GPT-4, n (%)		Claude 2, n (%)		Google Bard, n (%)	
		Without RISE[a]	With RISE	Without RISE	With RISE	Without RISE	With RISE
Concepts of diabetes	4	2 (50)	4 (100)	2 (50)	2 (50)	2 (50)	4 (100)
Symptoms and causes	4	3 (75)	3 (75)	1 (25)	3 (75)	2 (50)	4 (100)
Diabetes tests and diagnosis	4	4 (100)	4 (100)	3 (75)	4 (100)	3 (75)	3 (75)
Managing diabetes	15	14 (93)	15 (100)	9 (60)	14 (93)	14 (93)	14 (93)
Preventing diabetes problems	16	16 (100)	16 (100)	16 (100)	16 (100)	16 (100)	16 (100)
Total	43	39 (91)	42 (98)	31 (72)	39 (91)	37 (86)	41 (95)

[a]RISE: Retrieval-augmented Information System for Enhancement.

Accuracy rates indicate the percentage rated as "Good" in accuracy evaluation.

Comprehensiveness Evaluation

The study also assessed the comprehensiveness of model responses through a 1- to 5-point rating scale by 3 clinicians for the responses rated as "good" (Table 2). The results revealed that the integration of RISE led to a decrease in the number of responses with scores lower than 3 and an increase in the number of responses with higher scores of (4,5]. For instance, after incorporation of RISE, the number of responses scoring (1, 2] and (2, 3] reduced from 3 to 0 for GPT-4, from 6 to 3 for Claude, and from 8 to 2 for Bard. In addition, the number of responses scoring (4, 5] increased from 19 to 38 for GPT-4, from 9 to 24 for Claude, and from 13 to 18 for Bard.

Table 2. Comprehensiveness evaluation for responses of large language models with and without RISE.

Score range	GPT-4		Claude 2		Google Bard	
	Without RISE[a] (n=39)	With RISE (n=42)	Without RISE (n=31)	With RISE (n=39)	Without RISE (n=37)	With RISE (n=41)
(1, 2], n (%)	2 (5)	0 (0)	2 (7)	0 (0)	2 (5)	0 (0)
(2, 3], n (%)	1 (3)	0 (0)	4 (13)	3 (8)	6 (16)	2 (5)
(3, 4], n (%)	17 (44)	4 (10)	16 (52)	12 (31)	16 (43)	21 (51)
(4, 5], n (%)	19 (49)	38 (91)	9 (29)	24 (62)	13 (35)	18 (44)
Score, mean (SD)[b]	4.14 (0.72)	4.69 (0.39)	3.79 (0.78)	4.20 (0.60)	3.73 (0.80)	4.10 (0.62)

[a]RISE: Retrieval-augmented Information System for Enhancement.

[b]SD: Standard deviation. For responses rated as "good" by most graders, comprehensiveness was further evaluated.

Furthermore, the average scores for comprehensiveness also improved significantly after integrating RISE. GPT-4's average score increased from 4.14 (SD 0.72) to 4.69 (SD 0.39; $P<.001$), Claude increased from 3.79 (SD 0.78) to 4.2 (SD 0.60; $P=.002$), and Bard increased from 3.73 (SD 0.80) to 4.10 (SD 0.62; $P=.001$). Among the 3 models, GPT-4 consistently achieved the highest scores for comprehensiveness both before and after the integration of RISE, with scores of 4.14 (SD 0.72) and 4.69 (SD 0.39), respectively.

Understandability Evaluation

In addition to assessing the accuracy and comprehensiveness of model responses by clinicians, this study also evaluated the public's understanding of responses (Table 3). A total of 3 diabetes patients rated the understandability on a scale of 1 to 5. The results indicated the integration of RISE led to a decrease in the number of responses with scores lower than 4 and an increase in the number of responses with scores of (4, 5]. Specifically, after incorporation of RISE, the number of responses scoring lower than 4 was reduced from 15 to 4 for GPT-4, 27 to 24 for Claude, and 31 to 21 for Bard. In addition, the number of responses with higher scores of (4, 5] increased from 18 to 39 for GPT-4, 16 to 19 for Claude, and 12 to 22 for Bard after the incorporation of RISE.

XSL•FO
RenderX

Table 3. Evaluation of public understandability in responses from large language models with and without RISE.

Score range	GPT-4		Claude 2		Google Bard	
	Without RISE[a]	With RISE	Without RISE	With RISE	Without RISE	With RISE
(1, 2], n (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
(2, 3], n (%)	0 (0)	0 (0)	1 (2)	2 (5)	1 (2)	1 (2)
(3, 4], n (%)	15 (35)	4 (9)	26 (61)	22 (51)	30 (70)	20 (47)
(4, 5], n (%)	18 (65)	39 (91)	16 (37)	19 (44)	12 (28)	22 (51)
Score, mean (SD)[b]	4.32 (0.61)	4.64 (0.51)	4.01 (0.73)	4.07 (0.74)	3.96 (0.86)	4.16 (0.82)

[a]RISE: Retrieval-augmented Information System for Enhancement.

[b]SD: Standard deviation. An understandability evaluation was conducted for all responses.

Furthermore, the average scores for understandability also improved after RISE integration. GPT-4's average score increased from 4.32 (SD 0.61) to 4.64 (SD 0.51; $P<.001$), Claude improved from 4.01 (SD 0.73) to 4.07 (SD 0.74; $P=.31$), and Bard elevated from 3.96 (SD 0.86) to 4.16 (SD 0.82; $P=.002$). Among the 3 models, GPT-4 consistently exhibited the highest understandability scores both before and after RISE integration, with scores of 4.32 (SD 0.61) and 4.64 (SD 0.51), respectively.

Discussion

Principal Findings

Considering the prevalence of type 2 diabetes mellitus as a major public health concern, particularly in light of the widespread dependence of patients on online resources for health-related information, this study introduces RISE workflow models to enhance the performance of LLMs as timely and relevant diabetes education tools [42,46,47]. Our findings demonstrate that RISE significantly improves the accuracy and comprehensiveness of LLM responses to patient queries about diabetes management and care. On average, the percentage of accurate responses increased by 12% (15/129) with RISE, with rates increasing by 7% (3/43) for GPT-4, 19% (8/43) for Claude 2, and 9% (4/43) for Google Bard. The framework also enhanced response comprehensiveness and understandability, improving mean scores by 0.44 (SD 0.10) and 0.19 (SD 0.13), respectively.

Comparison to Previous Work

Previous studies have also applied LLMs in diabetes management and education. A study by Sun et al [7] found that 74.5% (149/200) of GPT-4's answers accurately responded to 200 frequently asked questions on diabetes management education. Hernandez et al [48] showed ChatGPT could correctly answer 98.5% (69/70) of patient questions about type 2 diabetes, and the 1.5% (1/70) inappropriate response needs to be improved. These findings were consistent with our results before integrating RISE, showing 91% (39/43) accuracy for base GPT-4 in responding to diabetes questions. Although most information provided by advanced LLMs may be correct, it is essential to realize that even small mistakes can potentially cause significant problems, especially with medical scenarios. Even minimal misinformation can lead to misconceptions, which might inadvertently delay treatment. Thus, minimizing potential errors and improving accuracy and validation are required before considering LLMs integration into patient diabetes care.

RAG has shown promise in enhancing LLM performance [49,50], however, most current RAG approaches rely on fixed, smaller, static knowledge bases. Our results showed model responses were more specific and accurate than those generated by general LLMs after incorporating specific knowledge of the RISE framework, which is consistent with previous studies. Previous studies have applied RAG in other clinical specialties, such as general medicine, hepatology, and lymphoma [32,38,43,51]. These studies' knowledge bases were mainly medical texts, research papers, and disease guidelines, limiting their flexibility and generalizability. In contrast, the RISE framework used a local medical knowledge base from NIH (National Institutes of Health) and the dynamic, real-time retrieval of external knowledge through the latest medical guidelines, academic research papers, and reputable health websites.

Future Directions

The RISE framework demonstrates the potential of RAG in enhancing the performance of LLMs for diabetes education, and there are several promising directions for future research and development. These include creating large specialized medical knowledge bases tailored for diabetes education, integrating multimodal data such as medical images and electronic health records, and developing domain-specific retrieval and ranking algorithms for evidence-based information [52,53]. Furthermore, exploring the bilingual or multilingual potential of these chatbots, such as investigating the performance of the RISE framework when questions are asked in languages like Chinese, could expand their use in real-world clinical practice outside the English-speaking world. Another promising direction is exploring personalized RAG systems that adapt to individual preferences of patients and contexts. Ensuring RAG systems' interpretability, transparency, privacy, and security is crucial in the medical domain.

Strengths and Limitations

This study has several strengths. First, novel RAG algorithms that effectively use local databases and external academic knowledge markedly improve the precision and real-time performance of responses to diabetes-related inquiries. Second, the RISE framework incorporates rigorous factual and safety

XSL•FO
RenderX

checks for the generated outputs, ensuring reliable and secure responses.

There are some limitations in the study. The RISE framework was developed and evaluated exclusively within the domain of diabetes. The generalizability of RISE to other medical domains remains uncertain. Future investigations should extend the application of the RISE framework to diverse medical specialties. Moreover, the scope of our evaluation was limited to these predetermined queries. Future research should conduct clinical trials to assess the RISE's ability to effectively address inquiries of patients and enhance the efficiency of diabetes management in real-world clinical scenarios.

Conclusions

In conclusion, the RISE framework shows promise as a safer and more reliable option for generating responses to common queries from diabetes patients. RISE significantly enhances the accuracy and comprehension of original LLM responses by retrieval of external knowledge from reliable sources. This framework can potentially be a supplementary tool to improve patient understanding and disease outcomes.

Acknowledgments

The National Key Research and Development Program of China (2022YFC2502800); The National Natural Science Foundation of China (NSFC grant 82171034, 81721003); The High-Level Hospital Construction Project at Zhongshan Ophthalmic Center of Sun Yat-sen University (Grants 303010303058, 303020107, 303020108). The Young Talent Support Project (2022) of the Guangzhou Association for Science and Technology.

Data Availability

The complete data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Dingqiao Wang, Jiangbo Liang, and Jinguo Ye contributed equally as the first authors. Fei Li, Bo Qu, and Yingfeng Zheng contributed equally as co-corresponding authors. FL, BQ, and YZ contributed to conceptualization. DW, JL, and JY performed the methodology. The investigation was done by Jingni L, Jingpeng L, QZ, QH, CP, DW, ZL, WS, and DS. DW, JL, and JY wrote the original draft. FL, BQ, and YZ handled the writing-review and editing. FL, BQ, and YZ performed supervision. Final approval of the version to be published was done by all authors.

Conflicts of Interest

None Declared.

Multimedia Appendix 1
Code for the Retrieval-augmented Information System for Enhancement framework.
[ZIP File (Zip Archive), 18 KB - jmir_v26i1e58041_app1.zip]

Multimedia Appendix 2
Sample of the Retrieval-augmented Information System for Enhancement Diabetes question and answer data set (n=50).
[XLSX File (Microsoft Excel File), 23 KB - jmir_v26i1e58041_app2.xlsx]

Multimedia Appendix 3
Responses for each model and raw scores for evaluation (clean).
[PDF File (Adobe PDF File), 1979 KB - jmir_v26i1e58041_app3.pdf]

References

1. Buse JB, Wexler DJ, Tsapas A, Rossing P, Mingrone G, Mathieu C, et al. 2019 Update to: management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD). Diabetes Care 2020;43(2):487-493 [FREE Full text] [doi: 10.2337/dci19-0066] [Medline: 31857443]
2. Pastors JG, Warshaw H, Daly A, Franz M, Kulkarni K. The evidence for the effectiveness of medical nutrition therapy in diabetes management. Diabetes Care 2002;25(3):608-613. [doi: 10.2337/diacare.25.3.608] [Medline: 11874956]
3. Gao C, Xu J, Liu Y, Yang Y. Nutrition policy and healthy China 2030 building. Eur J Clin Nutr 2021;75(2):238-246. [doi: 10.1038/s41430-020-00765-6] [Medline: 33219269]
4. Khan I, Agarwal R. Can ChatGPT Help in the awareness of Diabetes? Ann Biomed Eng 2023;51(10):2125-2129. [doi: 10.1007/s10439-023-03356-1] [Medline: 37648882]

XSL•FO
RenderX

5. Wang D, Feng L, Ye J, Zou J, Zheng Y. Accelerating the integration of ChatGPT and other large‐scale AI models into biomedical research and healthcare. MedComm – Future Medicine 2023;2(2):e43. [doi: 10.1002/mef2.43]
6. Chlorogiannis DD, Apostolos A, Chlorogiannis A, Palaiodimos L, Giannakoulas G, Pargaonkar S, et al. The Role of ChatGPT in the advancement of diagnosis, management, and prognosis of cardiovascular and cerebrovascular disease. Healthcare (Basel) 2023;11(21):2906 [FREE Full text] [doi: 10.3390/healthcare11212906] [Medline: 37958050]
7. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023;25:e51300 [FREE Full text] [doi: 10.2196/51300] [Medline: 37943581]
8. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging large language models for improved patient access and self-management: assessor-blinded comparison between expert- and AI-generated content. J Med Internet Res 2024;26:e55847 [FREE Full text] [doi: 10.2196/55847] [Medline: 38663010]
9. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
10. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. Hepatol Commun 2023;7(4):e0097 [FREE Full text] [doi: 10.1097/HC9.0000000000000097] [Medline: 36972383]
11. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. J Med Internet Res 2023;25:e46924 [FREE Full text] [doi: 10.2196/46924] [Medline: 37256685]
12. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
13. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023;27(1):120 [FREE Full text] [doi: 10.1186/s13054-023-04393-x] [Medline: 36945051]
14. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google bard. EBioMedicine 2023;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
15. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ 2024;10:e50842 [FREE Full text] [doi: 10.2196/50842] [Medline: 38236632]
16. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The scientific knowledge of bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. J Diabetes Sci Technol 2023:19322968231203987. [doi: 10.1177/19322968231203987] [Medline: 37798960]
17. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 02;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
18. Hulman A, Dollerup OL, Mortensen JF, Fenech ME, Norman K, Støvring H, et al. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a turing test-inspired survey among employees of a Danish diabetes center. PLoS One 2023;18(8):e0290773 [FREE Full text] [doi: 10.1371/journal.pone.0290773] [Medline: 37651381]
19. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P. Language models are few-shot learners. Advances in Neural Information Processing Systems 2020;33:1877-1901.
20. OpenAI. Gpt-4 technical report. arxiv 2303.08774 2023;2(5).
21. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A. Palm: scaling language modeling with pathways. Journal of Machine Learning Research 2023;24(240):1-113.
22. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 2022;35:24824-24837.
23. Shi F, Suzgun M, Freitag M, Wang X, Srivats S, Vosoughi S. Language models are multilingual chain-of-thought reasoners. arxiv.03057 2022.
24. Chalkidis I. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. arxiv: 2304.12202 2023. [doi: 10.2139/ssrn.4385460]
25. Kasai J, Kasai Y, Sakaguchi K, Yamada Y. Evaluating gpt-4 and chatgpt on Japanese medical licensing examinations. arxiv: 2303.18027 2023.
26. West CG. AI and the FCI: can ChatGPT project an understanding of introductory physics? arxiv: 2303.01067 2023.
27. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023;15(6):e40895 [FREE Full text] [doi: 10.7759/cureus.40895] [Medline: 37492832]
28. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Löser A. MedAlpaca--an open-source collection of medical conversational AI models and training data. arXiv:230408247 2023.
29. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv:230313375 2023.

XSL•FO
RenderX

30. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
31. Golovneva O, Chen M, Poff S, Corredor M, Zettlemoyer L, Fazel-Zarandi M. Roscoe: a suite of metrics for scoring step-by-step reasoning. arXiv:221207919 2022.
32. Soong D, Sridhar S, Si H, Wagner JS, Sá ACC, Yu CY. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model. arXiv:230517116 2023.
33. Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S. Self-refine: iterative refinement with self-feedback. Advances in Neural Information Processing Systems 2024:36.
34. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 2020;33:9459-9474.
35. Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: PMLR. 2020 Presented at: International Conference on Machine Learning; 2020 July 13; Vienna, Austria.
36. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
37. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina (Kaunas) 2024;60(3):445 [FREE Full text] [doi: 10.3390/medicina60030445] [Medline: 38541171]
38. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 2024;1(2):AIoa2300068 [FREE Full text] [doi: 10.1056/aioa2300068] [Medline: 38343631]
39. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening. medRxiv 2024 [FREE Full text] [doi: 10.1101/2024.02.08.24302376] [Medline: 38370719]
40. Yu W. Retrieval-augmented generation across heterogeneous knowledge. 2022 Presented at: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop; 2022 July 10; Washington. [doi: 10.18653/v1/2022.naacl-srw.7]
41. Feng Z, Feng X, Zhao D, Yang M, Qin B. Retrieval-generation synergy augmented large language models. arXiv:231005149 2023. [doi: 10.1109/icassp48485.2024.10448015]
42. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024;26:e56655 [FREE Full text] [doi: 10.2196/56655] [Medline: 38630520]
43. Shi W, Zhuang Y, Zhu Y, Iwinski H, Wattenbarger M, Wang M. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. 2023 Presented at: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatic; 2023 Oct 04; USA p. 1-10. [doi: 10.1145/3584371.3612956]
44. NCBI. PMC. URL: https://www.ncbi.nlm.nih.gov/pmc/ [accessed 2023-12-05]
45. Diabetes. URL: https://www.niddk.nih.gov/health-information/diabetes [accessed 2023-12-05]
46. Al-Lawati JA. Diabetes mellitus: a local and global public health emergency!. Oman Med J 2017;32(3):177-179 [FREE Full text] [doi: 10.5001/omj.2017.34] [Medline: 28584596]
47. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
48. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, Amoro Sanchez R, Muyolema Arce V, Mustafa A, et al. The future of patient education: AI-driven guide for type 2 diabetes. Cureus 2023;15(11):e48919 [FREE Full text] [doi: 10.7759/cureus.48919] [Medline: 38024047]
49. Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. 2024 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2024 Mar 10; Beijing, China p. 17754-17762. [doi: 10.1609/aaai.v38i16.29728]
50. Lozano A, Fleming SL, Chiang CC, Shah N. Clinfo.ai: an open-source retrieval-augmented large language model system for answering medical questions using scientific literature. Pac Symp Biocomput 2024;29:8-23 [FREE Full text] [Medline: 38160266]
51. Khene ZE, Bigot P, Mathieu R, Rouprêt M, Bensalah K, French Committee of Urologic Oncology. Development of a personalized chat model based on the European association of urology oncology guidelines: harnessing the power of generative artificial intelligence in clinical practice. Eur Urol Oncol 2024;7(1):160-162. [doi: 10.1016/j.euo.2023.06.009] [Medline: 37474402]
52. Huang Y, Huang J. A survey on retrieval-augmented text generation for large language models. arXiv:240410981 2024.
53. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T. Llama: open and efficient foundation language models. arxiv.13974 2023.

XSL•FO
RenderX

Abbreviations

FAISS: Facebook AI Similarity Search
LLMs: large language models
NIH: National Institutes of Health
RAG: Retrieval-augmented Generation
RISE: Retrieval-augmented Information System for Enhancement

Edited by G Eysenbach; submitted 04.03.24; peer-reviewed by G Zhang, J Mistry, Y Zhu; comments to author 01.05.24; revised version received 03.06.24; accepted 15.07.24; published 08.11.24.

Please cite as:
Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, Hu Q, Pan C, Wang D, Liu Z, Shi W, Shi D, Li F, Qu B, Zheng Y
Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study
J Med Internet Res 2024;26:e58041
URL: https://www.jmir.org/2024/1/e58041
doi:10.2196/58041
PMID:

©Dingqiao Wang, Jiangbo Liang, Jinguo Ye, Jingni Li, Jingpeng Li, Qikai Zhang, Qiuling Hu, Caineng Pan, Dongliang Wang, Zhong Liu, Wen Shi, Danli Shi, Fei Li, Bo Qu, Yingfeng Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study

Shuai Ming[1,2,3*], MD; Xi Yao[1], MD; Xiaohong Guo[1], MD; Qingge Guo[1,2,3], MD; Kunpeng Xie[4], MD; Dandan Chen[1,2,3], MD, PhD; Bo Lei[1,2,3*], MD, PhD

[1]Department of Ophthalmology, Henan Eye Institute, Henan Eye Hospital, Henan Provincial People's Hospital, Zhengzhou, China

[2]Eye Institute, Henan Academy of Innovations in Medical Science, Zhengzhou, China

[3]Henan Clinical Research Center for Ocular Diseases, People's Hospital of Zhengzhou University, Zhengzhou, China

[4]Department of Ophthalmology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

[*]these authors contributed equally

Corresponding Author:
Bo Lei, MD, PhD
Department of Ophthalmology
Henan Eye Institute, Henan Eye Hospital
Henan Provincial People's Hospital
No.7 Weiwu Road
Zhengzhou
China
Phone: 86 037167120925
Email: bolei99@126.com

Abstract

Background: Artificial intelligence (AI) chatbots such as ChatGPT are expected to impact vision health care significantly. Their potential to optimize the consultation process and diagnostic capabilities across range of ophthalmic subspecialties have yet to be fully explored.

Objective: This study aims to investigate the performance of AI chatbots in recommending ophthalmic outpatient registration and diagnosing eye diseases within clinical case profiles.

Methods: This cross-sectional study used clinical cases from *Chinese Standardized Resident Training–Ophthalmology (2nd Edition)*. For each case, 2 profiles were created: patient with history (Hx) and patient with history and examination (Hx+Ex). These profiles served as independent queries for GPT-3.5 and GPT-4.0 (accessed from March 5 to 18, 2024). Similarly, 3 ophthalmic residents were posed the same profiles in a questionnaire format. The accuracy of recommending ophthalmic subspecialty registration was primarily evaluated using Hx profiles. The accuracy of the top-ranked diagnosis and the accuracy of the diagnosis within the top 3 suggestions (do-not-miss diagnosis) were assessed using Hx+Ex profiles. The gold standard for judgment was the published, official diagnosis. Characteristics of incorrect diagnoses by ChatGPT were also analyzed.

Results: A total of 208 clinical profiles from 12 ophthalmic subspecialties were analyzed (104 Hx and 104 Hx+Ex profiles). For Hx profiles, GPT-3.5, GPT-4.0, and residents showed comparable accuracy in registration suggestions (66/104, 63.5%; 81/104, 77.9%; and 72/104, 69.2%, respectively; P=.07), with *ocular trauma*, *retinal diseases*, and *strabismus and amblyopia* achieving the top 3 accuracies. For Hx+Ex profiles, both GPT-4.0 and residents demonstrated higher diagnostic accuracy than GPT-3.5 (62/104, 59.6% and 63/104, 60.6% vs 41/104, 39.4%; P=.003 and P=.001, respectively). Accuracy for do-not-miss diagnoses also improved (79/104, 76% and 68/104, 65.4% vs 51/104, 49%; P<.001 and P=.02, respectively). The highest diagnostic accuracies were observed in *glaucoma*; *lens diseases*; and *eyelid, lacrimal, and orbital diseases*. GPT-4.0 recorded fewer incorrect top-3 diagnoses (25/42, 60% vs 53/63, 84%; P=.005) and more partially correct diagnoses (21/42, 50% vs 7/63 11%; P<.001) than GPT-3.5, while GPT-3.5 had more completely incorrect (27/63, 43% vs 7/42, 17%; P=.005) and less precise diagnoses (22/63, 35% vs 5/42, 12%; P=.009).

Conclusions: GPT-3.5 and GPT-4.0 showed intermediate performance in recommending ophthalmic subspecialties for registration. While GPT-3.5 underperformed, GPT-4.0 approached and numerically surpassed residents in differential diagnosis. AI chatbots show promise in facilitating ophthalmic patient registration. However, their integration into diagnostic decision-making requires more validation.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e60226) doi:10.2196/60226

KEYWORDS

artificial intelligence; chatbot; ChatGPT; ophthalmic registration; clinical diagnosis; AI; cross-sectional study; eye disease; eye disorder; ophthalmology; health care; outpatient registration; clinical; decision-making; generative AI; vision impairment

Introduction

Artificial intelligence (AI) has significantly advanced in health care, particularly in many areas of ophthalmology [1,2]. ChatGPT (OpenAI) [3] is a generative AI featuring a chatbot interface. Benefiting from its expansive knowledge base and complex parameterization, it enables users to input queries and receive responses that showcase advanced, humanlike logic. Since its launch in November 2022, ChatGPT has quickly amassed a substantial user base. It was recognized as having the potential to revolutionize not only ophthalmology but also the entire medical field in diverse aspects [4], including patient care, health care professionals and systems, research, and education and training [5]. Its performance was highlighted in patient triage proficiency [6], scientific writing [7], operative notes writing [8], and passing the ophthalmology specialist licensing examination [9].

Specialized eye hospitals in China, especially tertiary ones, frequently face patient overcapacity. With limited knowledge of eye health, patients could encounter difficulties in choosing the right subspecialty department when registering. User-friendly chatbot such as ChatGPT could provide registration suggestions based on the patients' chief complaints and medical histories and, thus, significantly ease these challenges and reduce health care resource wastage due to unsuitable registrations. However, the role of ChatGPT in classifying diseases into ophthalmic subspecialties and thus facilitating patient registration remains unexplored.

ChatGPT has exhibited encouraging results in diagnosing eye diseases within specific subspecialties, such as corneal and retinal vascular diseases [10,11]. The data for these assessments were sourced from public question banks and case report databases. In diagnosing a diverse range of ophthalmic conditions, ChatGPT failed to match the diagnostic accuracy of ophthalmologists but demonstrated the benefit of a shorter diagnostic time [12]. Further validation studies, particularly in testing ChatGPT's diagnostic effectiveness for a comprehensive range of eye diseases within the context of clinical practice in China, are essential.

Drawing on typical clinical cases from Chinese Standardized Resident Training (SRT) materials, this study aims to evaluate ChatGPT's capacity for classifying ophthalmic subspecialties and its diagnostic potential within the Chinese context. Our research seeks to provide insights into whether ChatGPT can effectively assist patients with appropriate registrations and support ophthalmologists in clinical decision-making.

Methods

Ethical Considerations

The Institutional Review Board of Henan Provincial People's Hospital determined that this in silicon research did not involve direct interaction with real-world human subjects, nor did it require the collection of new human data. Accordingly, an ethics exemption was granted for this study. The case information used was derived from publicly available sources and published materials. At the time of data collection, GPT-4.0 was publicly available by paid subscription through ChatGPT Plus.

Data Source

The clinical cases for our study were sourced from *Chinese Standardized Resident Training–Ophthalmology (2nd Edition)*, which is an official resource conforming to SRT content and standards, as well as the theoretical assessment guidelines of the National Health Commission of China. Unlike traditional undergraduate textbooks, SRT materials are specifically designed to focus on problem-based learning and case-based learning. They feature a variety of typical real-world cases from various ophthalmic subspecialties, making them particularly suitable for interaction with ChatGPT's chatbot interface, which is designed to handle complex, real-life queries.

In this study, we gathered 121 cases from 12 ophthalmic subspecialties, creating 2 profiles per case: patient with history (Hx) and patient with history and examination (Hx+Ex) [12]. The "history" portion comprised gender, age, chief complaints, and medical history. When necessary, past medical, familial, and systemic disease details were also added. The "examination" portion covered general ophthalmic assessments such as visual acuity and intraocular pressure, alongside diagnostics such as slit lamp biomicroscopy, a range of ophthalmic imaging (eg, fundus photography, orbital computed tomography [OCT], and fluorescein angiography), and specialized imaging for eye tumors and traumas (OCT and magnetic resonance imaging). The inclusion criteria mandated comprehensive historical and ophthalmology-related chief complaint details, documented examination results, and official case analyses with accurate subspecialty classifications and unique diagnoses.

After excluding 17 cases, a total of 104 cases were retained for further analysis. The reasons for excluding the 17 cases included (1) the lack of medical history or the presence of final diagnostic–like terms in the medical history, which may bias the assessment (6 cases); (2) the lack of textual descriptions of examination results (6 cases); (3) chief complaints not related to ophthalmic symptoms or cases referred from other departments (3 cases); and (4) unclear or non–ophthalmology-related diagnoses (2 cases). The mean age was 36.8 (SD 21.7) years, with male patients comprising 55.8% (58/104) of the cases. Notably, each case had a predominantly unique final diagnosis. The 3 most prevalent diagnoses,

XSL•FO
RenderX

classified into subspecialties, were *eyelid, lacrimal, and orbital diseases*; *retinal diseases*; and *strabismus and amblyopia* (Table 1). Given that the original 104 profiles were ordered based on the ophthalmic subspecialty, a new case numbering system was established by randomly assigning descending numerical values between 0 and 1 and arranging them accordingly.

Table 1. Classification of subspecialties in the 104 clinical cases.

Ophthalmic subspecialty	Clinical cases (n=104), n (%)
Eyelid, lacrimal, and orbital diseases	18 (17)
Retinal diseases	
Nonheritable	15 (13)
Heritable	6 (6)
Strabismus and amblyopia	10 (10)
Corneal and ocular surface diseases	9 (9)
Refractive errors	7 (7)
Scleral and uveal diseases	6 (6)
Ocular trauma	6 (6)
Eye tumors	
Eyelid, lacrimal, and orbital tumors	6 (6)
Scleral and uveal tumors	4 94)
Retinal tumors	1 (1)
Glaucoma	5 (5)
Vitreous diseases	3 (4)
Lens diseases	4 (4)
Neuro-ophthalmology	4 (4)
Total	104 (100)

Testing Process

A total of 208 clinical profiles (104 Hx and 104 Hx+Ex profiles) in Chinese were tested from March 5 to 18, 2024. The tested AI chatbots included ChatGPT versions 3.5 and 4.0 (GPT-3.5 and GPT-4.0, respectively), and 3 ophthalmology residents were also tested. Initially, ChatGPT was assigned a system role to emulate a professional ophthalmologist. Each clinical case scenario was entered in Chinese independently, followed by 2 questions: “Q1. Which ophthalmic subspecialty should be given priority for registration?” and “Q2: List the top three possible diagnoses and provide a detailed rationale for each.” Q1 aimed to elicit the chatbot’s triage recommendations for subspecialty registration, while Q2 focused on extracting the chatbot’s leading and differential diagnosis proposals. For the Hx profiles, both questions were asked, and the AI was informed of the available subspecialties for reference in the prompt. For the Hx+Ex profiles, given that the examination primarily serves for diagnosis, only Q2 was posed. Response history was reset prior to each new query (Figure 1). The detailed prompts and engineering techniques are shown in Multimedia Appendix 1. Based on the structure provided by these prompts, we found that ChatGPT’s responses generally adhered to the fixed template specified in our prompts. Consequently, we opted to input each case continuously in a single chat session. However, for the Hx and Hx+Ex profiles, as well as for tests conducted with both GPT-3.5 and GPT-4.0, each series of tests was initiated in a new chat session.

For the residents’ test, 2 sets of questionnaires were created following the Hx + Q1 and Hx+Ex + Q2 format. The residents’ evaluations were independent, which were ensured by implementing a blinded assessment process where the residents did not have information about the performance or responses of the AI systems or each other.

XSL•FO
RenderX

Figure 1. The design and analytical framework of the study. LLM: large language models.

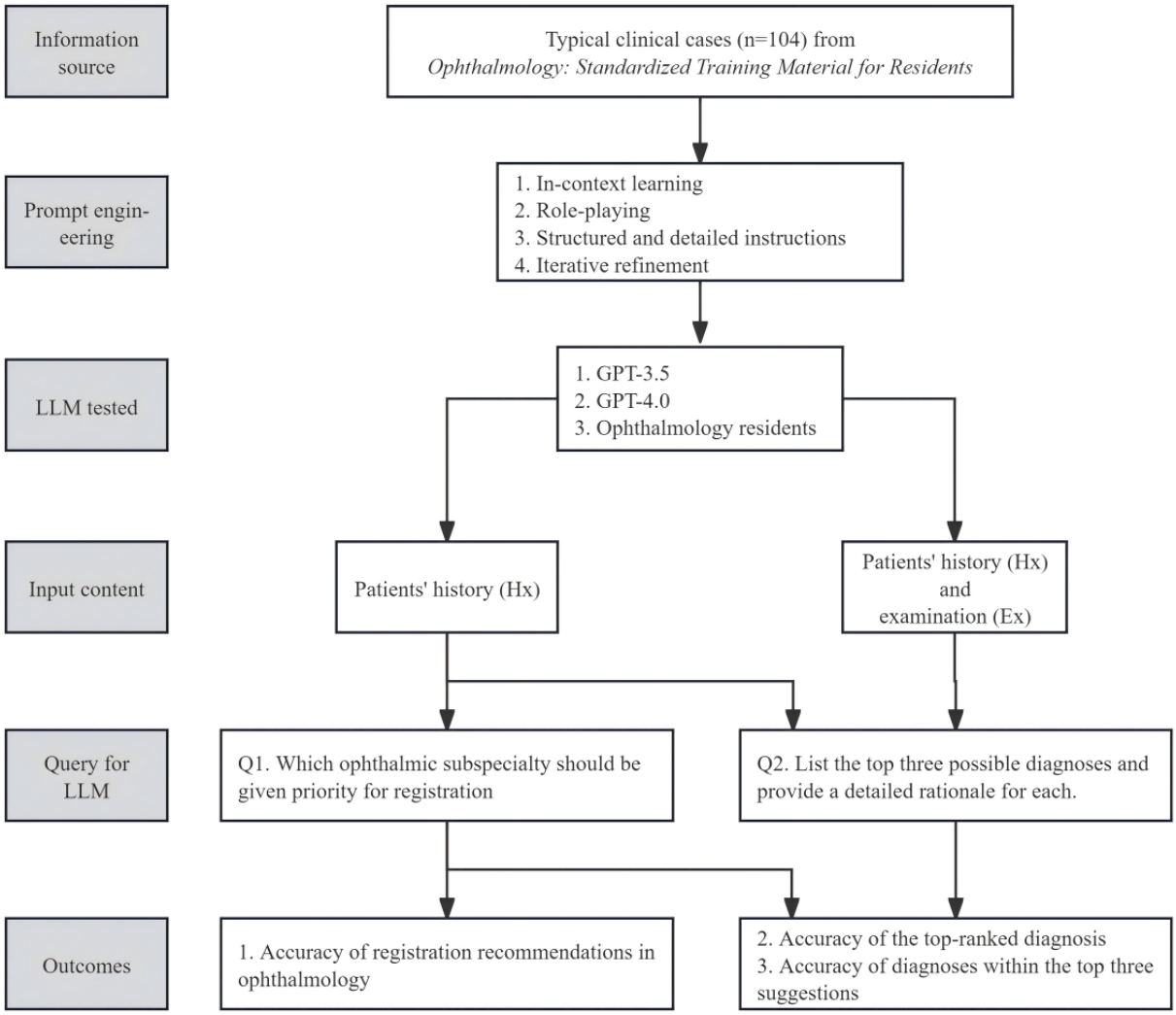

Outcomes and Definition

The study focused on 3 outcomes: accuracy of recommendation for ophthalmic subspecialty registration, accuracy of the top-ranked diagnosis, and accuracy of the diagnosis within the top 3 suggestions (do-not-miss diagnosis). As residents rarely provided 3 possible diagnoses like ChatGPT, they were not evaluated on the accuracy of the do-not-miss diagnosis. The gold standard for judgment was the official diagnosis from *Chinese Standardized Resident Training–Ophthalmology (2nd Edition)*.

ChatGPT and residents were evaluated based on the same outcome criteria. For ophthalmic subspecialty registration, overlaps exist in some subspecialties, such as *eye tumors* and *retinal diseases*. Recommendations to either category were considered correct. Precision was crucial for diagnosis suggestions corresponding to Hx+Ex profiles. For instance, if the final diagnosis is sympathetic ophthalmia, responses such as panuveitis or herpetic keratitis were marked as incorrect. Similarly, for acute idiopathic optic neuritis or orbital neurilemmoma, responses of optic neuritis or orbital tumor were also considered incorrect. However, for Hx profiles, an exact diagnosis was not required. All responses in the examples provided were considered correct.

Regarding the residents' performance, a diagnosis was considered correct only if at least 2 out of 3 residents provided the correct diagnosis. This approach emphasizes the importance of consensus in clinical decision-making and reflects the collaborative nature of medical diagnosis in real-world settings.

Statistical Analysis

Data collection and management were performed using Microsoft Excel software. Statistical analyses were mainly conducted in SPSS (version 26.0.0; IBM Corp). To compare the accuracy of triage and diagnosis across different testing strategies, the Pearson chi-square test or Fisher exact test was applied, depending on the expected frequency counts in the contingency tables. Due to limitations in SPSS for conducting the Fisher exact test on the differences in proportions among 3 groups, we switched to R software (version 4.4.1; R Foundation for Statistical Computing) for this analysis. For post hoc analysis, P values were adjusted using the Bonferroni method in pairwise comparisons. Unless otherwise specified, differences were considered statistically significant at $P<.05$.

XSL•FO
RenderX

Results

Accuracy of Recommendation for Subspecialty Registration

For Hx profiles, GPT-3.5, GPT-4.0, and residents demonstrated moderate accuracy in registering patients to the correct ophthalmic subspecialty, with accuracy of 63.5% (66/104), 77.9% (81/104), and 69.2% (72/104), respectively (*P*=.07). Subgroup analysis revealed that *ocular trauma*, *retinal diseases*, and *strabismus and amblyopia* ranked among the top 3 in overall registration accuracy. In contrast, registration accuracy was low for *glaucoma*, *neuro-ophthalmology*, and *lens diseases* (Table 2). The detailed registration recommendations for GPT-3.5, GPT-4.0, and residents are showed in Multimedia Appendix 2. An example of GPT-4.0's response to subspecialty registration is shown in Figure 2.

Table 2. Correct triage recommendations for subspecialty registration for patient history (Hx) profiles.

Ophthalmic subspecialty	GPT-3.5, n (%)	GPT-4.0, n (%)	Residents, n (%)	*P* value
Total (n=104)	66 (63.5)	81 (77.9)	72 (69.2)	.07
Retinal diseases (n=21)	18 (86)	20 (95)	19 (91)	.86
Eyelid, lacrimal, and orbital diseases (n=18)	14 (78)	14 (78)	11 (61)	.44
Eye tumors (n=11)	6 (55)	8 (73)	9 (82)	.52
Strabismus and amblyopia (n=10)	8 (80)	9 (90)	7 (70)	.85
Corneal and ocular surface diseases (n=9)	4 (44)	7 (78)	8 (89)	.10
Refractive errors (n=7)	6 (86)	5 (71)	4 (57)	.83
Ocular trauma (n=6)	6 (100)	6 (100)	6 (100)	__[a]
Scleral and uveal diseases (n=6)	1 (17)	4 (67)	3 (50)	.36
Glaucoma (n=5)	1 (20)	3 (60)	1 (20)	.50
Lens diseases (n=4)	1 (25)	2 (50)	1 (25)	.69
Neuro-ophthalmology (n=4)	1 (25)	1 (25)	1 (25)	>.99
Vitreous diseases (n=3)	0 (0)	2 (67)	2 (67)	.36

[a]Not applicable.

XSL•FO
RenderX

Figure 2. Example profile 70: interaction with and responses of the GPT-4.0 chatbot. (A) When provided with Hx information, GPT-4.0 correctly recommended the ophthalmic subspecialty of "retinal diseases." (B) When provided with Hx+Ex information, GPT-4.0 gave the top 3 diagnostic suggestions and accurately identified "retinal vein occlusion" as the top-ranked diagnosis. (C) Fundus photography and fluorescein angiography. (D) OCT imaging. (C) and (D) were presented to ChatGPT as textual descriptions of the examination results. Hx: patient with history; Hx+Ex: patient with history and examination; OCT: orbital computed tomography.

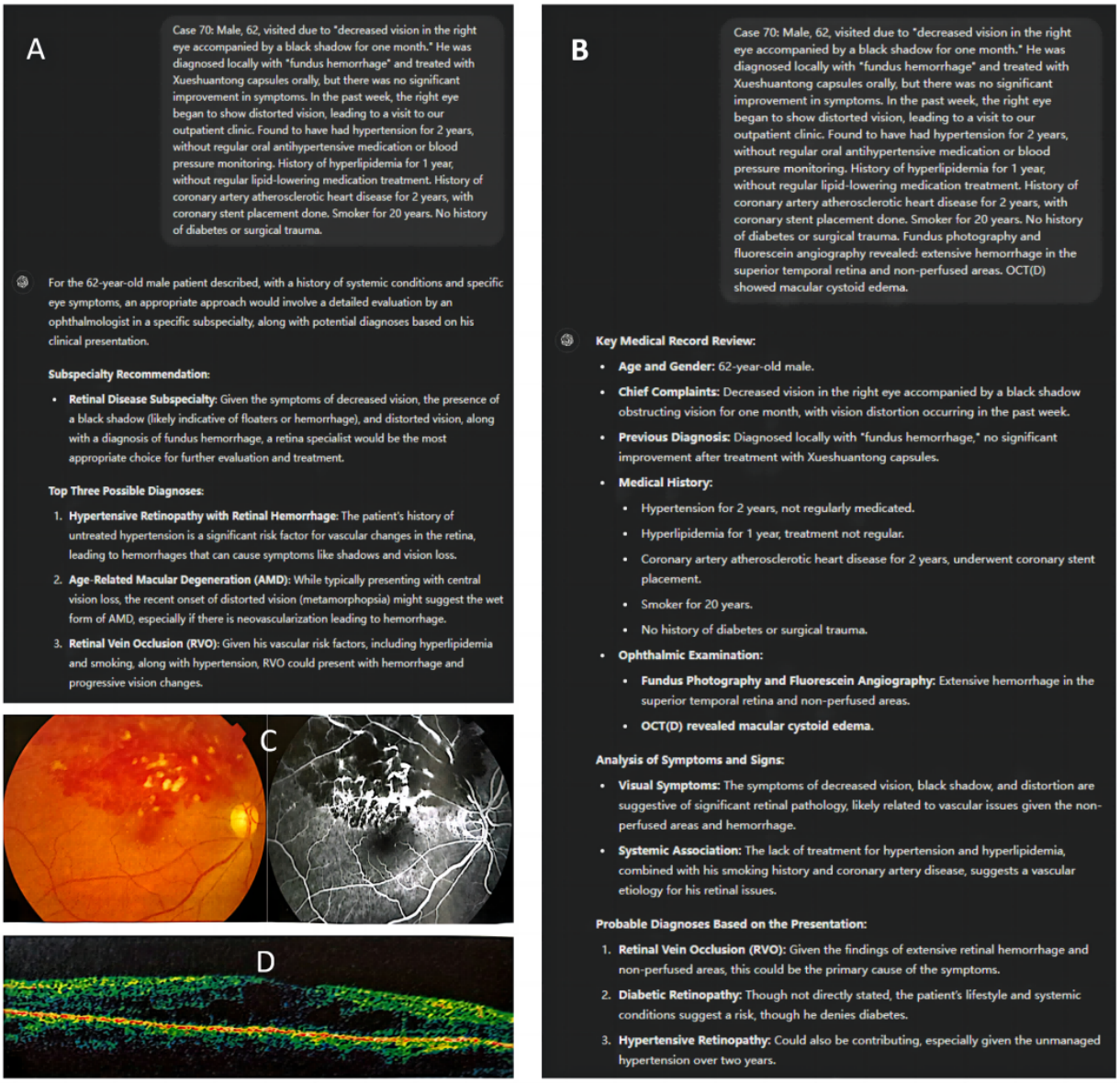

Accuracy of Diagnosis

For Hx+Ex profiles, both GPT-4.0 and residents demonstrated higher diagnostic accuracy compared to GPT-3.5 (62/104, 59.6% vs 41/104, 39.4%; *P*=.003; and 63/104, 60.6% vs 41/104, 39.4%; *P*=.001, respectively). Similarly, the accuracy of diagnoses within the top 3 suggestions were also higher (79/104, 76% vs 51/104, 49%; *P*<.001; and 68/104, 65.4% vs 51/104, 49%; *P*=.02, respectively). However, there was no statistically significant difference in diagnostic accuracy between GPT-4.0 and residents (79/104, 76% vs 68/104, 65.4%; *P*=.09). Compared to Hx profiles, GPT-4.0 showed improved diagnostic accuracy for Hx+Ex profiles and diagnoses within the top 3 suggestions (62/104, 59.6% vs 42/104, 40.4%; *P*=.007; and 79/104, 76% vs 63/104, 60.6%; *P*=.02, respectively; Table 3). The detailed top-3 predicted diagnoses by GPT-3.5 and GPT-4.0, alongside the composite diagnosis by the 3 residents, are showed in Multimedia Appendix 3. An example of GPT-4.0's response to diagnosis is shown in Figure 2.

In the subgroup analysis, both GPT-3.5 and GPT-4.0 exhibited generally lower diagnostic accuracy for Hx profiles. However, for Hx+Ex profiles, there was an overall improvement in diagnosis, particularly for glaucoma. The top 3 subspecialties in overall accuracy were *glaucoma*; *lens diseases*; and *eyelid, lacrimal, and orbital diseases*. In the subspecialties of *eye tumors* and *scleral and uveal diseases*, significant differences were observed in the top-ranked diagnosis accuracy for Hx+Ex profiles among GPT-3.5 , GPT-4.0, and residents (4/11, 36% vs 5/11, 46% vs 10/11, 91%; *P*=.03; and 1/6, 17% vs 5/6, 83% vs 5/6, 83%; *P*=.02, respectively). Notably, for *scleral and uveal diseases*, GPT-3.5 demonstrated a lower accuracy of 17% (1/6)

XSL•FO
RenderX

both in the top-ranked diagnosis (*P*=.02) and within the top 3 diagnoses (*P*=.02) compared to GPT-4.0 (5/6, 83%) and the residents (5/6, 83%; Figure 3).

Table 3. Secondary outcomes for patient with history (Hx) and patient with history and examination (Hx+Ex) profiles (n=104).

Diagnosis accuracy	GPT-3.5, n (%)	GPT-4.0, n (%)	Residents, n (%)	*P* value
Accuracy A: the top-ranked diagnosis is correct for Hx profiles	37 (35.6)	42 (40.4)[a,b]	—[c]	.48
Accuracy B: the diagnosis is within the top 3 suggestions for Hx profiles	50 (48.1)	63 (60.6)[a]	—	.07
Accuracy C: the top-ranked diagnosis is correct for Hx+Ex profiles	41 (39.4)[d,e]	62 (59.6)[b]	63 (60.6)	.002
Accuracy D: the diagnosis is within the top 3 suggestions for Hx+Ex profiles	51 (49)[d,e]	79 (76)	68 (65.4)	<.001

[a]Statistically significant differences for GPT-4.0 when comparing accuracy A vs accuracy C (*P*=.007) and accuracy B vs accuracy D (*P*=.02).

[b]Statistically significant differences for GPT-4.0 when comparing accuracy A vs accuracy B (*P*=.004) and accuracy C vs accuracy D (*P*=.01).

[c]Not applicable.

[d]Statistically significant differences for accuracy C and accuracy D when comparing GPT-3.5 with GPT-4.0 (*P*=.004 and *P*<.001, respectively).

[e]Statistically significant differences for accuracy C and accuracy D when comparing GPT-3.5 with residents (*P*=.002 and *P*=.02, respectively).

XSL•FO
RenderX

Figure 3. The diagnostic accuracy of GPT-3.5, GPT-4.0, and residents across various ophthalmic subspecialties: (A) accuracy of the top-ranked diagnosis for Hx profiles; (B) accuracy of the diagnosis within the top 3 suggestions for Hx profiles; (C) accuracy of the correct top-ranked diagnosis for Hx+Ex profiles; and (D) accuracy of the diagnosis within the top 3 suggestions for Hx+Ex profiles. "a" indicates significant statistical differences across all 3 groups (GPT-3.5, GPT-4.0, and residents), "b" denotes a significant difference between residents and GPT-3.5, and "c" represents a significant difference between GPT-4.0 and GPT-3.5. Hx: patient with history; Hx+Ex: patient with history and examination.

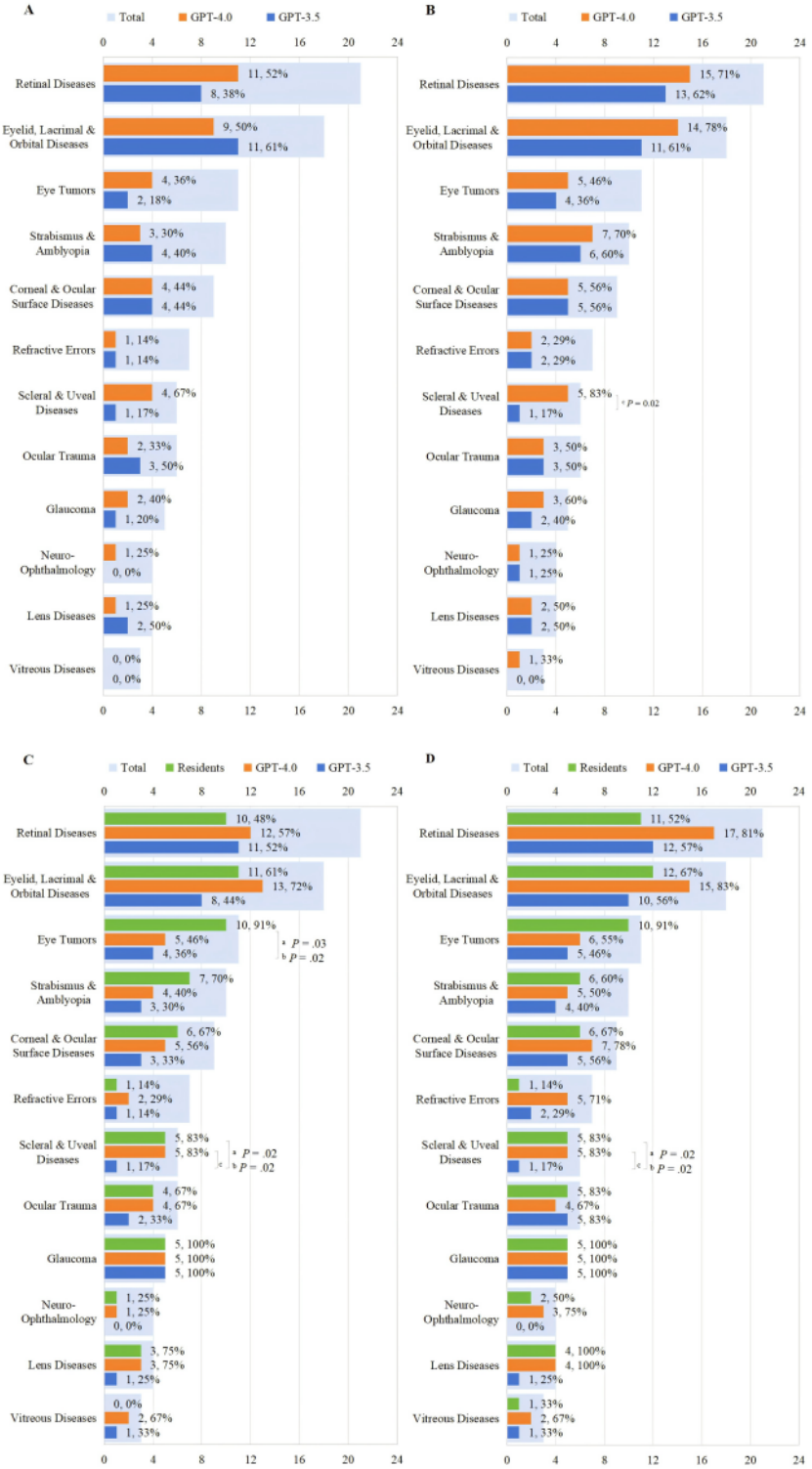

Case Description and Accuracy

For Hx+Ex profiles where the medical history provided past diagnoses related to the final diagnosis, a higher top-ranked diagnosis accuracy was observed in the GPT-4.0 model (7/7, 100% vs 55/97, 57%; *P*=.04). However, case descriptions including past diagnoses unrelated to the final diagnosis and cases requiring ophthalmic examination for a definitive diagnosis did not significantly affect the top-ranked diagnosis accuracy for GPT-3.5, GPT-4.0, and the residents (all *P*>.21; Table 4).

XSL•FO
RenderX

Table 4. Case characteristics and their association with the top-ranked diagnosis accuracy for patient with history and examination (Hx+Ex) profiles.

Cases characteristics	GPT-3.5	*P* value	GPT-4.0	*P* value	Residents	*P* value
Prescence of medical history A[a], n (%)		.43		.04		.70
No (n=97)	36 (37)		55 (57)		58 (60)	
Yes (n=7)	4 (57)		7 (100)		5 (71)	
Prescence of medical history B[b], n (%)		.73		.74		>.99
No (n=95)	36 (38)		56 (59)		57 (60)	
Yes (n=9)	4 (44)		6 (67)		6 (67)	
Prescence of characteristics C[c], n (%)		.21		.29		>.99
No (n=51)	24 (47.1)		31 (60.8)		31 (61)	
Partly (n=20)	6 (30)		9 (45)		12 (60)	
Yes (n=33)	10 (30.3)		22 (66.7)		20 (61)	

[a]Medical history A: descriptions of past diagnoses related to the final diagnosis.

[b]Medical history B: description of past diagnoses unrelated to the final diagnosis.

[c]Characteristic C: official diagnosis states that the diagnosis must be made in conjunction with an ophthalmic examination.

Characteristic of Incorrect Diagnoses

In the analysis of incorrect top-ranked diagnoses from Hx+Ex profiles, GPT-4.0 exhibited fewer incorrect top-3 diagnoses than GPT-3.5 (25/42, 60% vs 53/63, 84%; *P*=.005), making partially correct diagnoses with incorrect lesion nature (21/42, 50% vs 7/63, 11%; *P*<.001). In contrast, GPT-3.5 often made completely incorrect diagnoses about lesion nature (27/63, 43% vs 7/42, 17%; *P*=.005) and exhibited less precision with no further diagnosis more frequently than GPT-4.0 (22/63, 35% vs 5/42, 12%; *P*=.009; Table 5).

Table 5. Analysis of ChatGPT's incorrect top-ranked diagnoses in patient with history and examination (Hx+Ex) profiles.

	GPT-3.5 (n=63)	GPT-4.0 (n=42)	*P* value
Proportion of incorrect top-3 diagnosis, n (%)	53 (84)	25 (60)	.005
Reasons for incorrect top-ranked diagnosis, n (%)			<.001
Incorrect lesion nature	27 (43)	7 (17)	.005[a]
Incorrect etiology	1 (2)	2 (5)	.57
Lacks precision, no further diagnosis	22 (35)	5 (12)	.009[a]
Partially correct with incorrect lesion location	6 (10)	7 (17)	.28
Partially correct with incorrect lesion nature	7 (11)	21 (50)	<.001[a]

[a]Statistically significant difference between GPT-4.0 and GPT-3.5 at an adjusted *P* value of .01.

Discussion

Principal Finding

AI has demonstrated its potential in facilitating accurate patient registration and health care services [13,14]. With iterations, ChatGPT—a chatbot powered by AI—has shown promise for triage in ophthalmic emergencies and in achieving diagnostic accuracy in simulated vignettes [6,15]. Our study focused on investigating the triage and diagnostic value of ChatGPT. To our knowledge, this was the first study to explore the role of ChatGPT in ophthalmic registration. Additionally, we designed 2 types of information inputs, Hx and Hx+Ex profiles, and simultaneously tested the accuracy of leading diagnoses and do-not-miss diagnoses. This approach helped provide a comprehensive understanding of ChatGPT's performance. While existing studies mainly focus on the application within a single ophthalmic subspecialty [10,11,16,17], another strength of our study was the use of a diverse set of diagnostic cases, covering 12 ophthalmic subspecialties and 104 distinct cases. This breadth enhances the AI system's evaluation across varied clinical scenarios.

From the patients' perspective, AI chatbots facilitate ophthalmic triage and appointment [18]. Patients know their own complaints and medical histories well. This knowledge enables them to directly interact with AI chatbots, seeking advice on appropriate ophthalmic subspecialties for registration [4]. Our study used Hx profiles to simulate this self-service interaction, highlighting the practical utility of chatbots in patient-initiated health care navigation. The findings demonstrated that GPT-4.0 directed patients to the correct registration with 78% accuracy, which numerically surpassed the 69% accuracy achieved by medically trained residents. In China, where major tertiary hospitals have implemented web-based registration systems, such as through

XSL•FO
RenderX

a WeChat-based medical platform [19], the integration of a user-friendly and accessible AI chatbot significantly streamlines the consultation process for patients, particularly those unsure of which ophthalmic subspecialty to choose. This study provided the first empirical evidence of how AI chatbots can facilitate more accurate and efficient patient registration in clinical settings.

For ophthalmic diagnoses, studies had shown that GPT-4.0 exhibited lower than 50% accuracy in deriving leading diagnoses from complaint records alone [6], while the free version of GPT-3.5 underperforms compared to medical residents [12]. These findings were consistent with our study. In a pilot study by Hu et al [20], GPT-4.0 was tested on its capability to diagnose rare eye diseases, revealing that more comprehensive information provided to the model resulted in considerably more "right" diagnoses. The reduced accuracy when used with limited patient information (such as Hx only) indicated that GPT-4.0 was not yet suitable as a stand-alone diagnostic tool. However, adding details from patient examinations significantly enhanced its performance, making it comparable to that of residents. In real-world clinical settings, where physicians have access to comprehensive case information, GPT-4.0 achieved diagnostic accuracies between 60% to 76%, demonstrating its potential to support clinical decision-making. Similar performance had been observed in uveitis studies, with accuracy rates ranging from 60% to 66% [21,22]. Unlike GPT-3.5, which provided imprecise diagnoses, GPT-4.0, even when erring in its initial diagnosis, tended to include the correct diagnosis within the top 3 suggestions. This illustrated GPT-4.0's superior information retrieval capabilities and its role as a diagnostic aid [23]. It is expected that with continuous improvements, AI chatbots will play an increasingly vital role in enhancing health care efficiency [24,25].

Our findings revealed that the accuracy of recommendations for ophthalmic subspecialty registration did not consistently correlate with diagnostic accuracy. For example, while the diagnostic accuracy for glaucoma reached 100%, the accuracy of leading patients to register in the glaucoma department was notably low. This discrepancy was likely attributed to the frequently nonspecific initial complaints associated with glaucoma, which generally require additional clinical examinations to establish a definitive diagnosis, such as intraocular pressure, OCT, and visual field tests. Therefore, relying solely on patient complaints and medical histories proved insufficient for accurately guiding patients to the appropriate glaucoma specialty for initial registration. Conversely, when patient histories were supplemented with specific ophthalmic examinations, the accuracy of differential diagnosis improved, thereby enhancing overall diagnostic performance.

Compared to the accuracy of GPT-4.0 in answering ophthalmic questions [26-28] and suggesting surgical plan [29], which can exceed 80%, its performance in registration recommendations and clinical diagnosis was intermediate or inferior. This discrepancy may be attributed to the conservative diagnostic judgment standards and the inclusion of less common diseases. Although the evaluated profiles were derived from real-world clinical reports in the published Chinese SRT database, the inclusion of uncommon and atypical cases was inevitable. Such heightened complexity poses significant challenges not only for AI chatbots but also for residents in the early stages of their medical careers [30]. In real-world clinical settings, the uneven distribution of diseases across various ophthalmic subspecialties typically leads to variability in diagnostic outcomes.

Given that the performance of AI chatbots is contingent upon the volume of information provided, future AI chatbots may still require upgrades and iterations to mitigate information asymmetry between patients and health care professionals, thereby enhancing the delivery of more effective and professional ophthalmic care. For example, AI chatbots specifically designed and trained for ophthalmic care [31,32]; chatbots that can proactively solicit information not provided by end users, similar to the process used by ophthalmologists; and those capable of directly accessing and interpreting imaging data like a multimodal AI chatbot, are essential. Unlike general-purpose AI chatbots such as ChatGPT, Zheng et al [33] have developed a Chinese large language model for ophthalmology using a corpus with extensive clinical vignettes (Hx+Ex). This model demonstrated a diagnostic accuracy of 81.1% across 6 common ophthalmic subspecialties, surpassing the performance of GPT-4.0 (59.6%) in our study. However, the current capability of multimodal GPT-4.0 to diagnose vitreoretinal diseases through the analysis of retinal images remains less than optimal [34].

ChatGPT allows users to customize responses based on personalized prompts, as illustrated in this study by providing 3 differential diagnostic recommendations and the corresponding rationale [35,36]. Research showed that ChatGPT consistently offers a broader range of differential diagnoses than ophthalmology residents [30]; this tendency was also observed in our data collection. The ability to organize key case information through the explanation of diagnostic reasoning not only enhances the knowledge structure of physicians but also underscores the significant educational value of AI chatbots in medical training [37-39]. The proficiency of AI chatbots in responding to ophthalmic examination questions and addressing eye disease queries have been confirmed by recent studies [40-42]. However, it is crucial to note that current AI chatbots do not necessarily replace the clinical judgment of professional ophthalmologists. Their application is still subject to ethical considerations [43,44] and concerns about hallucinations [45]. Physicians responsible for diagnosis need to remain cautious when considering information provided by ChatGPT. The extent to which AI can serve as an adjunct tool in health care still requires further real-world testing.

Limitations

In clinical settings, ophthalmologists typically rely on direct observation of patients' symptoms and examinations for intuitive face-to-face diagnoses. However, the text-based format of medical records may not fully reflect ophthalmic residents' capabilities and might even underestimate them. Additionally, while images are crucial for ophthalmic diagnosis, GPT-4.0's support for multimodal data still shows suboptimal performance in image-based cases [46]. This explains our decision to exclude images from our evaluation of ChatGPT. However, textual descriptions of these images could have impacted the outcomes.

XSL•FO
RenderX

Moreover, although ChatGPT supports multiple languages, differences in language use in diagnosing eye diseases have been observed [11]. Our study, conducted in Mandarin Chinese, may affect the generalization of the results. Despite these limitations, our study contributes to understanding AI's role as a tool in ophthalmic health care.

Conclusion

Our study showed that GPT-3.5 and GPT-4.0 demonstrated moderate performance in directing patients to appropriate ophthalmic subspecialties for registration. While GPT-3.5 was less effective, GPT-4.0 approached and even numerically surpassed residents in differential diagnosis when presented with patient histories and examination results. AI chatbots merit emphasis for their potential to facilitate patient registration and optimize consultations in ophthalmology. While their diagnostic capabilities could benefit ophthalmologists, integrating them into diagnostic decision-making still requires further validation.

Acknowledgments

This research was supported by the Medical Science and Technology Tackling Plan of Henan Province (LHGJ20210078).

Data Availability

The raw data supporting the conclusions of this paper will be made available by the authors, without undue reservation.

Authors' Contributions

SM conceived and designed the study, carried out data analysis, drafted the manuscript, and responded to review comments, therefore is the first corresponding author. SM, QG, and XG evaluated ChatGPT's responses and assisted in interpreting the results. KX, DC, and XY were responsible for quality control and assisted in data cleaning. BL participated in the design, reviewed the manuscript, and provided reimbursement for the publication fee, therefore shares the role of corresponding author with SM. All authors contributed to the paper and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts for the Hx and Hx+Ex profiles before testing ChatGPT's registration and diagnosis responses. Hx+Ex: patient with history and examination; Hx: patient with history.
[PDF File (Adobe PDF File), 171 KB - jmir_v26i1e60226_app1.pdf]

Multimedia Appendix 2
Detailed official subspecialty classification for each clinical profile (Hx), alongside registration recommendations for GPT-3.5, GPT-4.0, and residents. Hx: patient with history.
[DOCX File , 24 KB - jmir_v26i1e60226_app2.docx]

Multimedia Appendix 3
Detailed official diagnosis and top 3 predicted diagnoses by GPT-3.5 and GPT-4.0, alongside the composite diagnosis by 3 residents (1=correct and 0=incorrect) for each clinical profile (Hx+Ex). Hx+Ex: patient with history and examination.
[DOCX File , 41 KB - jmir_v26i1e60226_app3.docx]

References

1. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb 25;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct 10;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
3. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2024-10-24]
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? Eye (Lond) 2024 Jan 27;38(1):4-7. [doi: 10.1038/s41433-023-02619-4] [Medline: 37369764]

XSL•FO
RenderX

6. Zandi R, Fahey JD, Drakopoulos M, Bryan JM, Dong S, Bryar PJ, et al. Exploring diagnostic precision and triage proficiency: a comparative study of GPT-4 and Bard in addressing common ophthalmic complaints. Bioengineering (Basel) 2024 Jan 26;11(2):120 [FREE Full text] [doi: 10.3390/bioengineering11020120] [Medline: 38391606]
7. Salimi A, Saheb H. Large language models in ophthalmology scientific writing: ethical considerations blurred lines or not at all? Am J Ophthalmol 2023 Oct;254:177-181. [doi: 10.1016/j.ajo.2023.06.004] [Medline: 37348667]
8. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. Ann Biomed Eng 2023 Nov 02;51(11):2353-2355. [doi: 10.1007/s10439-023-03263-5] [Medline: 37266720]
9. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Jun 01;141(6):589-597 [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.1144] [Medline: 37103928]
10. Delsoz M, Madadi Y, Raja H, Munir WM, Tamm B, Mehravaran S, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea 2024 May 01;43(5):664-670. [doi: 10.1097/ICO.0000000000003492] [Medline: 38391243]
11. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. J Med Internet Res 2024 Jan 22;26:e51926 [FREE Full text] [doi: 10.2196/51926] [Medline: 38252483]
12. Shemer A, Cohen M, Altarescu A, Atar-Vardi M, Hecht I, Dubinsky-Pertzov B, et al. Diagnostic capabilities of ChatGPT in ophthalmology. Graefes Arch Clin Exp Ophthalmol 2024 Jul 06;262(7):2345-2352. [doi: 10.1007/s00417-023-06363-z] [Medline: 38183467]
13. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J Cloud Comp 2021 Jan 15;10(1):4. [doi: 10.1186/s13677-020-00218-2]
14. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023 Aug 24;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61] [Medline: 38939520]
15. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol 2024 Aug;59(4):e301-e308. [doi: 10.1016/j.jcjo.2023.07.016] [Medline: 37572695]
16. Carlà MM, Gambini G, Baldascino A, Boselli F, Giannuzzi F, Margollicci F, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol 2024 Sep;262(9):2945-2959. [doi: 10.1007/s00417-024-06470-5] [Medline: 38573349]
17. Maywood M, Parikh R, Deobhakta A, Begaj T. Performance assessment of an artificial intelligence chatbot in clinical vitreoretinal scenarios. Retina 2024 Jun 01;44(6):954-964. [doi: 10.1097/IAE.0000000000004053] [Medline: 38271674]
18. Betzler BK, Chen H, Cheng C, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. Lancet Digit Health 2023 Dec;5(12):e917-e924 [FREE Full text] [doi: 10.1016/S2589-7500(23)00201-7] [Medline: 38000875]
19. Yang X, Kovarik CL. A systematic review of mobile health interventions in China: identifying gaps in care. J Telemed Telecare 2021 Jan;27(1):3-22. [doi: 10.1177/1357633X19856746] [Medline: 31319759]
20. Hu X, Ran AR, Nguyen TX, Szeto S, Yam JC, Chan CKM, et al. What can GPT-4 do for diagnosing rare eye diseases? a pilot study. Ophthalmol Ther 2023 Dec 01;12(6):3395-3402 [FREE Full text] [doi: 10.1007/s40123-023-00789-8] [Medline: 37656399]
21. Rojas-Carabali W, Cifuentes-González C, Wei X, Putera I, Sen A, Thng ZX, et al. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in uveitis. Ocul Immunol Inflamm 2024 Oct 18;32(8):1526-1531. [doi: 10.1080/09273948.2023.2253471] [Medline: 37722842]
22. Rojas-Carabali W, Sen A, Agarwal A, Tan G, Cheung CY, Rousselot A, et al. Chatbots vs. human experts: evaluating diagnostic performance of chatbots in uveitis and the perspectives on AI adoption in ophthalmology. Ocul Immunol Inflamm 2024 Oct 13;32(8):1591-1598. [doi: 10.1080/09273948.2023.2266730] [Medline: 37831553]
23. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023 Sep;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
24. Biswas S, Davies LN, Sheppard AL, Logan NS, Wolffsohn JS. Utility of artificial intelligence-based large language models in ophthalmic care. Ophthalmic Physiol Opt 2024 May 25;44(3):641-671. [doi: 10.1111/opo.13284] [Medline: 38404172]
25. Madadi Y, Delsoz M, Khouri A, Boland M, Grzybowski A, Yousefi S. Applications of artificial intelligence-enabled robots and chatbots in ophthalmology: recent advances and future trends. Curr Opin Ophthalmol 2024 May 01;35(3):238-243. [doi: 10.1097/ICU.0000000000001035] [Medline: 38277274]
26. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience 2023 Nov 17;26(11):108163 [FREE Full text] [doi: 10.1016/j.isci.2023.108163] [Medline: 37915603]
27. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023 Oct 29;13(1):18562 [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]

XSL•FO
RenderX

28. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 2023 Oct;7(10):862-868. [doi: 10.1016/j.oret.2023.05.022] [Medline: 37277096]
29. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol 2024 Sep 20;108(10):1457-1469. [doi: 10.1136/bjo-2023-325143] [Medline: 38448201]
30. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther 2023 Dec 14;12(6):3121-3132 [FREE Full text] [doi: 10.1007/s40123-023-00805-x] [Medline: 37707707]
31. Li F, Wang D, Yang Z, Zhang Y, Jiang J, Liu X, et al. The AI revolution in glaucoma: bridging challenges with opportunities. Prog Retin Eye Res 2024 Aug 24;103:101291. [doi: 10.1016/j.preteyeres.2024.101291] [Medline: 39186968]
32. Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, Shi D, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit Med 2024 May 03;7(1):111 [FREE Full text] [doi: 10.1038/s41746-024-01101-z] [Medline: 38702471]
33. Zheng C, Ye H, Guo J, Yang J, Fei P, Yuan Y, et al. Development and evaluation of a large language model of ophthalmology in Chinese. Br J Ophthalmol 2024 Sep 20;108(10):1390-1397 [FREE Full text] [doi: 10.1136/bjo-2023-324526] [Medline: 39019566]
34. Ghalibafan S, Taylor Gonzalez DJ, Cai L, Graham Chou B, Panneerselvam S, Conrad Barrett S, et al. Applications of multimodal generative artificial intelligence in a real-world retina clinic setting. Retina 2024 Oct 01;44(10):1732-1740. [doi: 10.1097/IAE.0000000000004204] [Medline: 39287535]
35. Kleinig O, Gao C, Kovoor JG, Gupta AK, Bacchi S, Chan WO. How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality. Eye (Lond) 2024 Mar 05;38(4):649-653 [FREE Full text] [doi: 10.1038/s41433-023-02772-w] [Medline: 37798360]
36. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
37. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov 19;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
38. Gurnani B, Kaur K. Leveraging ChatGPT for ophthalmic education: a critical appraisal. Eur J Ophthalmol 2024 Mar 16;34(2):323-327. [doi: 10.1177/11206721231215862] [Medline: 37974429]
39. Sevgi M, Antaki F, Keane PA. Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. Br J Ophthalmol 2024 Sep 20;108(10):1354-1361 [FREE Full text] [doi: 10.1136/bjo-2023-325046] [Medline: 38719344]
40. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Aug 01;141(8):798-800 [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
41. Antaki F, Milad D, Chia MA, Giguère CÉ, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol 2024 Sep 20;108(10):1371-1378. [doi: 10.1136/bjo-2023-324438] [Medline: 37923374]
42. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol 2024 Apr 01;142(4):371-375 [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.6917] [Medline: 38386351]
43. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health 2024 Nov;6(11):e848-e856 [FREE Full text] [doi: 10.1016/S2589-7500(24)00143-2] [Medline: 39294061]
44. Kalaw FGP, Baxter SL. Ethical considerations for large language models in ophthalmology. Curr Opin Ophthalmol 2024 Nov 01;35(6):438-446. [doi: 10.1097/ICU.0000000000001083] [Medline: 39259616]
45. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023 Oct;254:141-149. [doi: 10.1016/j.ajo.2023.05.024] [Medline: 37339728]
46. Mihalache A, Huang RS, Popovic MM, Patil NS, Pandya BU, Shor R, et al. Accuracy of an artificial intelligence chatbot's interpretation of clinical ophthalmic images. JAMA Ophthalmol 2024 Apr 01;142(4):321-326. [doi: 10.1001/jamaophthalmol.2024.0017] [Medline: 38421670]

Abbreviations

AI: artificial intelligence
Hx: patient with history
Hx+Ex: patient with history and examination
OCT: orbital computed tomography

XSL•FO
RenderX

SRT: Standardized Resident Training

Edited by A Schwartz; submitted 05.05.24; peer-reviewed by MO Khursheed, C Cifuentes-Gonzalez, W Rojas-Carabali; comments to author 20.06.24; revised version received 05.08.24; accepted 15.10.24; published 14.11.24.

Please cite as:
Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, Lei B
Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study
J Med Internet Res 2024;26:e60226
URL: https://www.jmir.org/2024/1/e60226
doi:10.2196/60226
PMID:

©Shuai Ming, Xi Yao, Xiaohong Guo, Qingge Guo, Kunpeng Xie, Dandan Chen, Bo Lei. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study

Yuhe Ke[1,2], MBBS; Rui Yang[1], MS; Sui An Lie[2], MM; Taylor Xin Yi Lim[2], MBBS; Yilin Ning[1], PhD; Irene Li[3], PhD; Hairil Rizal Abdullah[2], PhD; Daniel Shu Wei Ting[1,4], PhD; Nan Liu[1,5], PhD

[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore
[2]Department of Anesthesiology, Singapore General Hospital, Singapore, Singapore
[3]Information Technology Center, University of Tokyo, Tokyo, Japan
[4]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
[5]Institute of Data Science, National University of Singapore, Singapore, Singapore

Corresponding Author:
Nan Liu, PhD
Centre for Quantitative Medicine
Duke-NUS Medical School
8 College Road
Singapore, 169857
Singapore
Phone: 65 66016503
Email: liu.nan@duke-nus.edu.sg

Abstract

Background: Cognitive biases in clinical decision-making significantly contribute to errors in diagnosis and suboptimal patient outcomes. Addressing these biases presents a formidable challenge in the medical field.

Objective: This study aimed to explore the role of large language models (LLMs) in mitigating these biases through the use of the multi-agent framework. We simulate the clinical decision-making processes through multi-agent conversation and evaluate its efficacy in improving diagnostic accuracy compared with humans.

Methods: A total of 16 published and unpublished case reports where cognitive biases have resulted in misdiagnoses were identified from the literature. In the multi-agent framework, we leveraged GPT-4 (OpenAI) to facilitate interactions among different simulated agents to replicate clinical team dynamics. Each agent was assigned a distinct role: (1) making the final diagnosis after considering the discussions, (2) acting as a devil's advocate to correct confirmation and anchoring biases, (3) serving as a field expert in the required medical subspecialty, (4) facilitating discussions to mitigate premature closure bias, and (5) recording and summarizing findings. We tested varying combinations of these agents within the framework to determine which configuration yielded the highest rate of correct final diagnoses. Each scenario was repeated 5 times for consistency. The accuracy of the initial diagnoses and the final differential diagnoses were evaluated, and comparisons with human-generated answers were made using the Fisher exact test.

Results: A total of 240 responses were evaluated (3 different multi-agent frameworks). The initial diagnosis had an accuracy of 0% (0/80). However, following multi-agent discussions, the accuracy for the top 2 differential diagnoses increased to 76% (61/80) for the best-performing multi-agent framework (Framework 4-C). This was significantly higher compared with the accuracy achieved by human evaluators (odds ratio 3.49; P=.002).

Conclusions: The multi-agent framework demonstrated an ability to re-evaluate and correct misconceptions, even in scenarios with misleading initial investigations. In addition, the LLM-driven, multi-agent conversation framework shows promise in enhancing diagnostic accuracy in diagnostically challenging medical scenarios.

(J Med Internet Res 2024;26:e59439) doi:10.2196/59439

KEYWORDS

clinical decision-making; cognitive bias; generative artificial intelligence; large language model; multi-agent

XSL•FO
RenderX

Introduction

Human cognitive biases in clinical decision-making are increasingly recognized as a crucial factor in health care errors and suboptimal patient outcomes [1]. These biases stem from innate psychological tendencies and can potentially lead to misjudgments and suboptimal outcomes in patient care [2-4]. Despite concerted efforts involving educational strategies, optimal work environments, and a culture promoting bias awareness and correction [5,6], the eradication of these biases remains an elusive goal.

The integration of artificial intelligence (AI), and in particular, large language models (LLMs), into clinical medicine is on the horizon [7]. LLMs have advanced text generation capability and extensive domain-specific knowledge [8,9]. Notably, these models have demonstrated their proficiency by successfully passing advanced medical examinations [10] and scoring clinical risk gradings on par with experienced physicians [11].

However, the deployment of LLMs in actual clinical diagnosis and decision-making processes has been mired in controversy, primarily due to the high stakes involved. The use of AI in medical settings is not just a technological issue; it intersects with complex ethical, legal, and medical considerations [11-16]. The accuracy of ChatGPT making the correct emergency medicine diagnosis is still limited to 77% to 83% [17]. Thus, concerns centering around the legal implications and accountability in cases where AI-driven diagnostics might lead to errors or misjudgments are a major hurdle.

This debate is rooted in the fundamental difference between human and machine intelligence. While LLMs can process and analyze vast quantities of data far beyond human capacity [18], they lack the nuanced understanding, empathy, and ethical reasoning inherent to human practitioners [19]. Human cognitive biases can be mitigated through a combination of awareness, education, and structured approaches [20]. Simulations of such discussions through LLM agents could provide a new solution to increase the accuracy of diagnosis [21]. The multi-agent framework features dialogue agents with near-human performance and could introduce an innovative paradigm in health care [22-24]. By simulating scenarios that mirror real-life clinical decision-making processes, and through reading the multi-agent conversations, clinicians can be made aware of potential cognitive biases and how to correct them. This facilitates learning in a controlled, educational environment [25,26].

Despite significant advancements in LLM technology, especially within multi-agent systems, its application to identify and mitigate human cognitive biases in clinical settings remains largely unexplored in current research. This study seeks to evaluate the efficacy of the multi-agent framework in achieving accurate final diagnoses following discussions on cognitive biases that may be present within the initial diagnosis. In addition, it aims to compare these outcomes with the differential diagnoses provided by experienced clinicians after reviewing the same scenarios. By doing so, the research aims to shed light on the potential of the multi-agent system to support clinical decision-making processes and enhance diagnostic accuracy in health care settings.

Methods

Overview

We accessed GPT-4 [27] through an application programming interface call to the OpenAI server. The specific variant used was GPT-4 Turbo.

We implemented a comprehensive search strategy aimed at including all relevant reports on misdiagnoses attributed to cognitive biases. A selection of 15 case reports was identified after a full review of the published literature as a representative sampling.

Ethical Considerations

Due to the nature of the study, institutional review board approval was not required, as the research did not involve patient data and did not constitute human participants.

Search Strategy

In this study, we focused on case reports highlighting instances of misdiagnoses resulting from cognitive biases. Our research involved a comprehensive search of the PubMed database using the terms “case reports [Publication Type]” AND “cognitive bias[All Fields]”. PubMed was chosen for its comprehensive coverage of case studies across diverse medical disciplines, ensuring access to a broad spectrum of peer-reviewed case studies for our analysis.

Eligibility for inclusion requires case reports to meet four key criteria: (1) they must provide detailed case information sufficient for making the initial diagnosis; (2) they must include a final, accurate diagnosis for the patient; (3) the incorrect diagnosis must be linked to cognitive bias by the authors; and (4) the final diagnosis should not be a rare disease or unclear. A rare disease is a disease or condition that affects fewer than 200,000 patients per year. The list of exclusions for rare diseases was based on the National Organization for Rare Disorders [28]. We set no limits on the publication year of these reports.

Screening of abstracts for eligible studies was conducted by 2 independent clinically trained reviewers, YK and TXYL. Each reviewer separately assessed whether a case should be included or excluded based on predefined criteria. In instances of disagreement, SAL reviewed the justifications for exclusion and inclusion and made the final decision. The full texts were reviewed to obtain the case summary, the initial wrong diagnosis by the medical team, and the final diagnosis of the case reported.

For the studies included in the analysis, full-text extraction, including patient demographics, past medical history, initial presenting complaints, and results of the preliminary investigations, was conducted. For cases involving imaging data, such as chest x-rays, we did not incorporate the actual images into the query. Instead, we opted to include the legends or descriptions accompanying these images. In defining the boundaries of the clinical scenarios for our study, we restricted the information to that available up to the point of and before the initial diagnosis. This meant deliberately excluding any

XSL•FO
RenderX

subsequent investigations, treatments, or management strategies that followed.

As GPT-4 Turbo has a knowledge base trained up to April 2023 [29], there is potential bias stemming from the inclusion of case reports that might have been part of the LLM's pre-training data. Hence a personal clinical scenario that was not published on the internet was included. This complex case was derived from the critical care attending's personal experience where cognitive biases had resulted in wrong and delayed diagnosis. A concise summary of these clinical scenarios, including the unique case, is provided in Multimedia Appendix 1.

Multi-Agent Conversation Framework

In this study, we used the multi-agent conversation framework provided by AutoGen [22] to assess its efficacy in mitigating cognitive biases in clinical decision-making. Within the system, each agent interacts based on their predefined role prompts, thereby simulating the collaborative decision-making process typically observed among health care professionals.

The suggested optimal group size to facilitate group discussion and performance has been proposed to be between 3 and 5 [30]. In the absence of established literature recommending an optimal team size for mitigating cognitive biases in medical settings, we constructed a simulation using 3 to 4 different agents, representing a typical clinical team composition [31]. These configurations aim to realistically emulate the dynamics of clinical teams and their potential to reduce cognitive biases.

As shown in Table 1, we tested 3 different frameworks to identify the most effective configuration. Framework 3 consisted of 3 agents (Junior Resident I, Junior Resident II, and Recorder), while Frameworks 4 (Junior Resident I, Junior Resident II, Professional Expert, and Recorder) and 4-C (Junior Resident I, Junior Resident II, Senior Doctor, and Recorder) each utilized 4 agents, with the fourth agent playing different roles. The distinguishing feature of Framework 4-C, in contrast to other frameworks, lies in its explicit directive for the Senior Doctor role to engage in discussions specifically focused on cognitive biases alongside the initial diagnosis. In addition, we experimented with combinations involving 5 or more agents, but the fifth agent did not effectively participate in the conversations despite modifications to the prompts. Consequently, the frameworks were limited to a maximum of 4 agents. All prompts for agent roles can be found in Multimedia Appendix 2.

Table 1. Different roles in the multi-agent conversation framework. Framework 3 consists of 3 agents, and Frameworks 4 and 4-C consist of 4 agents each.

Agents present	Role descriptions	Multi-agent framework		
		3	4	4-C
Junior Resident I	To make the final diagnosis after considering the discussions	✓	✓	✓
Junior Resident II	The devil's advocate and correct confirmation and anchoring bias	✓	✓	✓
Professional Expert	The field expert in any subspecialization required (eg, radiologists and cardiologists)		✓	
Senior Doctor	The tutor and facilitator of the discussion to reduce premature closure bias			✓
Recorder	To record and summarize the findings	✓	✓	✓

The diagnostic process was orchestrated through the collaborative efforts of simulated medical professionals (agents) with varying levels of expertise, as shown in Figure 1. Junior Resident I, as the primary physician, was tasked with presenting the initial diagnosis. Junior Resident I was given the personality of making swift assumptions but is willing to embrace feedback and consider alternative diagnostic possibilities. After the group discussion, Junior Resident I is then allowed to reconsider the most probable differential diagnosis along with an alternative. Junior Resident II, a colleague of Junior Resident I, critically appraised the initial diagnosis, pinpointing inconsistencies and advocating for alternative differential diagnoses. This role was instrumental in addressing potential confirmation and anchoring biases in the diagnostic process. Complementing the juniors, the Senior Doctor brought in-depth experience to the table, crucially identifying cognitive biases in the initial diagnosis and steering the junior residents toward a more nuanced and accurate diagnosis, while the Professional Expert aims to provide any specialist knowledge required to help with the diagnosis without further encouraging discussions of cognitive biases. This guidance was vital in circumventing premature diagnostic closure and knowledge bias. In addition, the role of the Recorder was to distill the outcomes of the discussion, compiling a definitive list of differential diagnoses and extracting key learning points from the collaborative effort, thereby enriching the diagnostic process with a comprehensive and multifaceted approach.

XSL•FO
RenderX

Figure 1. Different roles in the multi-agent conversation framework.

Diagnostic Accuracy Assessment

The final accurate diagnoses for the clinical scenarios were extracted directly from the published cases. Summarized answers from both the multi-agent framework and human evaluators were marked as "Correct" if they matched the final accurate diagnoses. Vague answers, such as diagnosing "septic shock" when the accurate diagnosis was "endometriosis," were marked as "Incorrect." A total of 2 physicians graded the answers. In cases where there were discrepancies in their assessments, discussions were held to reach a consensus.

The overall performance of the framework was evaluated based on the accuracy of (1) the "initial diagnosis" made without any multi-agent discussions and (2) the "final diagnosis" after the discussions. Each clinical scenario within each multi-agent framework was simulated 5 times to assess the consistency of diagnoses across multiple iterations.

A total of 3 doctors, each with at least 5 years of clinical experience, were asked to provide their top differential diagnosis along with 2 other differentials based on the clinical scenarios. If the top differential was correct, both the "initial diagnosis" and "final diagnosis" were marked as correct. The diagnoses generated by the multi-agent framework were compared to those provided by human doctors using Fisher exact test.

Bias Identification and Mitigation

An integral part of the evaluation involved documenting the specific cognitive biases identified and addressed during the agents' discussions. This aspect focused on understanding how effectively the multi-agent system could recognize and mitigate cognitive biases, which are crucial factors in diagnostic accuracy. Hallucinations are characterized by the dissemination of false medical information during multi-agent conversations or responses that fail to directly address the queries posed. This determination was made following an independent review by 2 doctors who thoroughly evaluated the provided answers. The interaction and decision-making process among the agents are illustrated in (Figure 2). This representation aids in visualizing the dynamics of the simulation and the interplay between different agents in reaching a diagnosis.

XSL•FO
RenderX

Figure 2. Schematic illustration of multi-agent (Framework 4-C) discussion dynamics leading to accurate differential diagnosis. ED: emergency department; LLM: large language model.

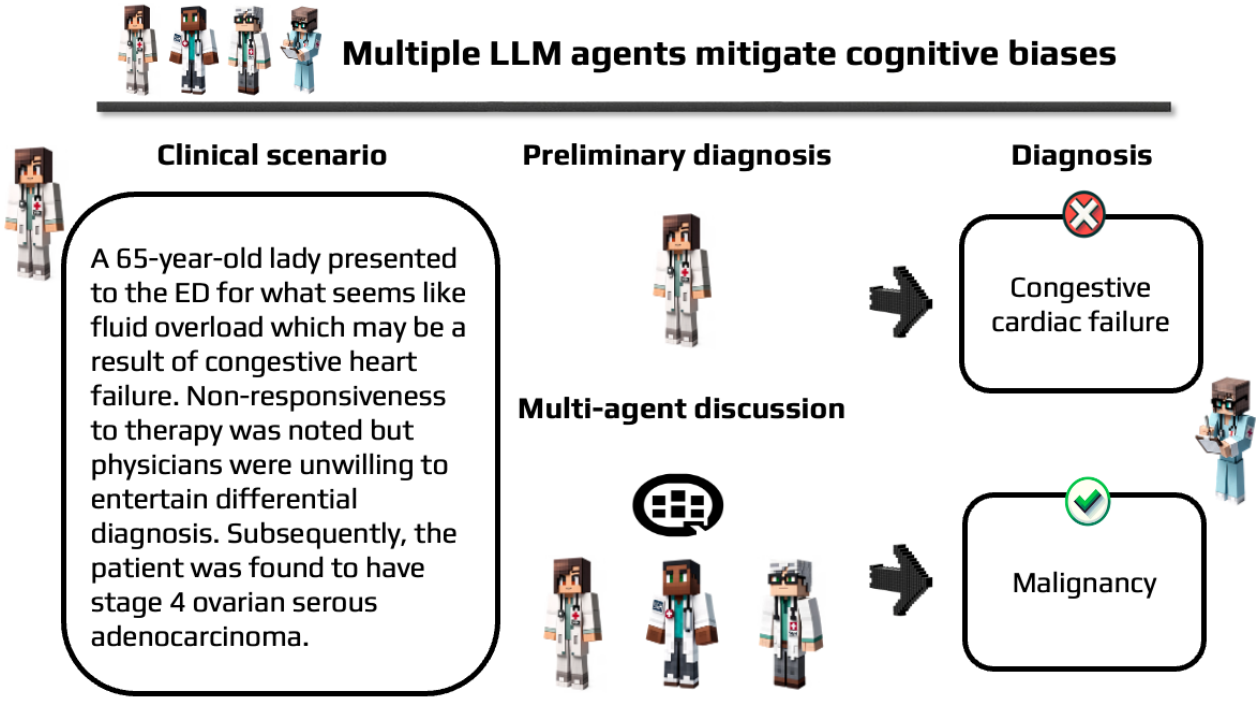

Results

Overview

A comprehensive search of the PubMed database yielded 162 case reports, of which 37 were determined to be eligible for inclusion in the study. From this subset, 15 cases were selected for evaluation as a representative sampling. In addition, a 16th scenario, derived from critical care attending personal clinical experience, was included. The flow diagram can be viewed in Figure 3.

Figure 3. Flow diagram for identification of case reports with cognitive bias.

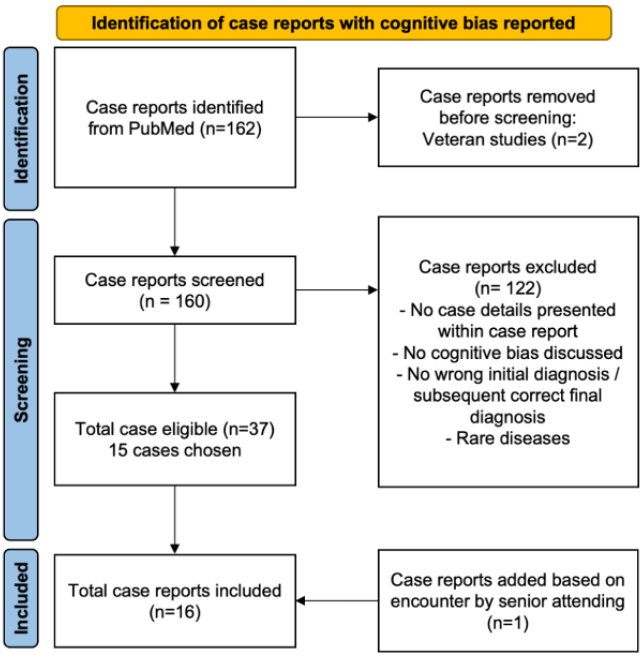

Overall Performance of Multi-Agent Conversation Framework

A total of 240 responses were generated by the multi-agent conversation framework, encompassing both initial and final diagnoses following discussions with the agents. The initial diagnosis made by the first-responder agent had a correctness rate of 0% (0/80) across all 3 multi-agent frameworks, whereas the human-generated answers had an accuracy rate of 27% (13/48). After the multi-agent discussions, the final diagnosis was correct in 76% (61/80) of cases in Framework 4-C, which

XSL•FO
RenderX

was significantly better than the human answers (odds ratio 3.49; *P*=.002), as shown in Table 2.

Table 2. Number of correct responses across different multi-agent frameworks and humans.

Multi-agent framework	Initial diagnosis, n (%)	Final diagnosis, n (%)	Odds ratio[a]	*P* value[a]
3 (n=80)	0 (0)	51 (64)	1.91	.10
4 (n=80)	0 (0)	55 (69)	2.39	.03
4-C (n=80)	0 (0)	61 (76)	3.49	.002
Human (n=48)	13 (27)	23 (48)	__[b]	—

[a]Fisher exact test.
[b]Not applicable.

Clinical Scenarios

The clinical cases covered a broad spectrum of medical fields, ranging from pediatric to malignancy diagnosis. Specifically, 6 cases were centered on infectious disease diagnosis, 3 pertained to critical care, and 2 involved vascular-related diagnoses. The rest were diverse, spanning neurology, gynecology, cancer, urology, and endocrinology (Table 3). A notable aspect of these cases was the cross-disciplinary nature of most initial and final diagnoses, observed in 12 (75%) out of the 16 cases. For example, in one illustrative case (case 1 [32]), an older adult patient presented with symptoms of shortness of breath on exertion and cough, leading to an initial misdiagnosis of heart failure. However, further investigation, considering her ongoing treatment with infliximab for rheumatoid arthritis, revealed the actual diagnosis of miliary tuberculosis. Another case (case 12 [33]) involved a young woman presenting with sudden, left-sided sharp pleuritic chest pain, which lessened when sitting forward. Despite the initial chest radiograph being interpreted as showing no acute abnormalities, the AutoGen system, provided with this potentially misleading information, initially diagnosed the condition as a pulmonary embolism. Yet, after a thorough discussion and re-evaluation, the correct diagnosis of pneumothorax was established, indicating a missed finding in the chest radiograph.

Table 3. Clinical scenarios with the initial wrong and the final correct diagnosis are given in the scenarios.

Reference	Publication year	Initial wrong diagnosis	Final correct diagnosis
[32]	2015	Heart failure	Miliary tuberculosis
[34]	2017	Bronchial asthma triggered by bacterial pneumonia	Heart failure secondary to dilated cardiomyopathy
[35]	2018	Syndrome of inappropriate secretion of antidiuretic hormone	Nonfunctioning macropituitary adenoma causing adrenal insufficiency
[36]	2019	Headache and neck pain—migraines and muscle strain	Cryptococcal meningitis
[37]	2020	Complex regional pain syndrome flare	Left common and external iliac artery occlusion
[38]	2022	Pelvic inflammatory disease	Atypical ectopic pregnancy
[39]	2022	COVID-19 pneumoniae	Bacterial pneumonia (legionella pneumoniae)
[40]	2022	Urinary tract infection with complicated pyelonephritis	Vesicointestinal fistula due to Crohn disease
[41]	2022	Bone (sternum) tuberculosis	Syphilitic gumma and osteomyelitis
[42]	2022	Urinary tract infection	Vertebral osteomyelitis and bilateral psoas and retroperitoneal abscesses
[43]	2022	Anaphylaxis secondary to henna	Superior vena cava syndrome
[33]	2023	Pulmonary embolism	Pneumothorax
[44]	2023	Diabetic ketoacidosis with infections	Thiamine deficiency
[45]	2023	Endometritis	Ischemic bowel
[46]	2023	Acute myocarditis likely due to MIS-C[a]	Acute myocarditis caused by invasive bacterial infection
__[b]	—	Congestive cardiac failure	Malignancy

[a]MIS-C: multisystem inflammatory syndrome in children.
[b]Not applicable.

XSL•FO
RenderX

Consistency of Multi-Agent Conversation Framework

There were variations observed in the repeated scenarios, particularly in the process of generating the top 2 differential diagnoses. For instance, in case 13 [44], a young patient presenting with lactic acidosis was initially diagnosed with diabetic ketoacidosis, and further discussions within the multi-agent environment led to the identification of thiamine deficiency. In 3 (60%) out of 5 simulations, Junior Resident I identified thiamine deficiency as the top differential diagnosis following the multi-agent discussions. In the remaining 2 simulations, gastrointestinal disorders were initially considered the most likely diagnosis, with thiamine deficiency being the second most likely differential.

The multi-agent discussions led to the correct final diagnosis in 13 (81%) out of the 16 scenarios across all 3 multi-agent frameworks. Furthermore, these discussions were effective in identifying various clinical biases, including anchoring bias, confirmation bias, availability bias, and premature closure. A detailed breakdown of the answers and the cognitive biases identified is provided in Multimedia Appendix 1. The detailed breakdown of correct answers for each scenario is available in Multimedia Appendix 3.

Discussion

Principal Findings

This study assessed the effectiveness of the multi-agent conversation framework in improving diagnostic accuracy and mitigating cognitive biases in clinical decision-making. Our findings reveal that integrating multi-agent discussions substantially enhances diagnostic accuracy. The best-performing framework, which used 4 agents, included 1 agent specifically tasked with identifying cognitive biases. This 4-C multi-agent framework that includes discussions on cognitive biases performed significantly better compared with human-generated answers. However, it is important to note that while 4 agents performed best in our study setting, this may vary in general applications.

The increase in diagnostic accuracy demonstrates the value of multi-perspective analysis in medical diagnosis, a core feature of the multi-agent conversation environment. This is in line with previous research emphasizing the importance of collaborative decision-making in health care to mitigate individual cognitive biases [47]. The consistency of responses in the repeated scenarios further validates its reliability and potential applicability in real-world clinical settings.

The effectiveness of the multi-agent conversation system, particularly in scenarios involving misleading or misinterpreted initial investigations, is noteworthy. This was exemplified in case 12, where our multi-agent framework successfully identified a pneumothorax that had been initially overlooked by human clinicians, and improved the accuracy of the final diagnosis to 100% (5/5) in multi-agent framework 4-C. The multi-agent systems were able to critically examine and question potential misinterpretations. Such capabilities are crucial in refining the diagnostic process and enhancing accuracy. While Brown et al [33] discussed the role of AI in the identification of pneumothorax, there is a potential for pre-trained LLMs to replace the decision aid, rather than to develop new resource-intensive systems such as image deep learning. This strategy aligns with the current trajectory of AI development in health care, where the focus is on integrating and maximizing existing AI technologies to enhance clinical decision-making and focus on sustainable AI [48,49].

The integration of multi-agent frameworks in clinical practice holds promising implications for enhancing diagnostic accuracy and ultimately improving patient outcomes. By systematically addressing cognitive biases through collaborative discussions among agents, these frameworks offer a structured approach to refining diagnostic reasoning. This approach not only complements traditional diagnostic methods but also introduces a dynamic element that challenges and verifies initial clinical hypotheses. In practical terms, the ability of multi-agent systems to consistently identify and correct potential diagnostic errors, as demonstrated in our study, suggests a transformative potential in reducing patient morbidity and optimizing treatment strategies. Furthermore, the application of these frameworks within electronic medical records (EMRs) could revolutionize decision-making processes by providing real-time, data-driven insights that augment clinician judgment and ensure adherence to best practices. As health care systems evolve toward more integrated and technology-driven approaches, the strategic incorporation of multi-agent systems stands poised to contribute significantly to improving the quality and efficiency of patient care delivery.

The results of our study extend beyond the educational benefits of multi-agents, highlighting their potential for broader clinical integration. The reflective process fostered by engaging with LLMs in diagnosing and revising cases not only cultivates an educational atmosphere conducive to developing critical thinking skills but also suggests practical applications in clinical settings [50]. One notable avenue is the integration of multi-agent into EMR systems. This could enhance decision-making processes by providing real-time, data-driven insights and augmenting the cognitive capabilities of medical professionals. Such integration would not only streamline the diagnostic process but also aid in the identification of potential cognitive biases, thereby enhancing the quality of patient care. Furthermore, the incorporation of multi-agents in EMRs could facilitate continuous learning and improvement, ensuring that medical practitioners remain updated with the latest medical knowledge and best practices, crucial for maintaining high standards in patient treatment and care.

Limitations

The study, while providing valuable insights into the potential application of multi-agents in clinical diagnostics, is subject to several limitations. First, the reliance on published case reports limits the breadth of clinical scenarios, potentially affecting the generalizability of our findings to broader medical practice. Second, the exclusion of visual data, such as medical imaging, confines our model's diagnostic capabilities to text-based information, omitting a critical component of clinical diagnosis. In addition, the inherent biases present in the LLMs, based on their pre-training data, could have influenced the diagnostic

XSL•FO
RenderX

suggestions. Meanwhile, the technical limitations inherent in LLMs, including their understanding of complex medical terminologies and nuances [51], may not match the expertise of experienced clinicians, possibly limiting the scope of applicability.

Future studies could assess the effectiveness and adaptability of the multi-agent framework in evolving clinical scenarios. More importantly, while LLMs have demonstrated potential as a valuable clinical aid in correcting cognitive biases, the implementation of such technology in health care necessitates rigorous ethical and regulatory oversight [52] and should continue to augment rather than replace the human clinician's expertise [14].

Conclusions

In conclusion, our study highlights the potential of multi-agents in enhancing diagnostic accuracy in clinical decision-making. The findings support the integration of advanced generative AI technology in educational and clinical settings as a tool for augmenting human decision-making. Future research should focus on the application of these systems in real-time clinical environments and their impact on patient outcomes. Moreover, ethical and legal considerations regarding the use of AI in health care need continued exploration to ensure patient safety and professional accountability.

Acknowledgments

This study was supported by the Duke-NUS Signature Research Program funded by the Ministry of Health, Singapore. Any opinions, findings conclusions, or recommendations expressed in this material are those of the author or authors and do not reflect the views of the Ministry of Health.

Authors' Contributions

YK and SAL conceived the study. YHK, RY, DSWT, and NL designed the study. YK and RY drafted the manuscript with critical appraisal and further development by DSWT, NL. YHK, TXYL, YN, RY, and IL conducted data analyses. NL supervised the study. All authors contributed to the revision of the manuscript and approval of the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of clinical scenarios and the initial and final correct diagnosis, as well as examples of cognitive biases present.
[DOCX File , 25 KB - jmir_v26i1e59439_app1.docx]

Multimedia Appendix 2
Prompts for different agent roles.
[DOCX File , 16 KB - jmir_v26i1e59439_app2.docx]

Multimedia Appendix 3
Detailed breakdown of the correct answers within each scenario stratified by each multi-agent framework and human.
[DOCX File , 18 KB - jmir_v26i1e59439_app3.docx]

References

1. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016;16(1):138 [FREE Full text] [doi: 10.1186/s12911-016-0377-1] [Medline: 27809908]
2. Korteling JEH, Paradies GL, Sassen-van Meer JP. Cognitive bias and how to improve sustainable decision making. Front Psychol 2023;14:1129835. [doi: 10.3389/fpsyg.2023.1129835] [Medline: 37026083]
3. Berthet V. The impact of cognitive biases on professionals' decision-making: a review of four occupational areas. Front Psychol 2021;12:802439. [doi: 10.3389/fpsyg.2021.802439] [Medline: 35058862]
4. Beldhuis IE, Marapin RS, Jiang YY, Simões de Souza NF, Georgiou A, Kaufmann T, et al. Cognitive biases, environmental, patient and personal factors associated with critical care decision making: a scoping review. J Crit Care 2021;64:144-153. [doi: 10.1016/j.jcrc.2021.04.012] [Medline: 33906103]
5. Doherty TS, Carroll AE. Believing in overcoming cognitive biases. AMA J Ethics 2020;22(9):E773-E778. [doi: 10.1001/amajethics.2020.773] [Medline: 33009773]
6. Hershberger PJ, Markert RJ, Part HM, Cohen SM, Finger WW. Understanding and addressing cognitive bias in medical education. Adv Health Sci Educ Theory Pract 1996;1(3):221-226. [doi: 10.1007/BF00162919] [Medline: 24179022]
7. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023;2(4):255-263. [doi: 10.1002/hcs2.61] [Medline: 38939520]

XSL•FO
RenderX

8. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv. Preprint posted online on November 28, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.16452]
9. Yang R, Liu H, Marrese-Taylor E, Zeng Q, Ke YH, Li W, et al. KG-Rank: enhancing large language models for medical QA with knowledge graphs and ranking techniques. arXiv. Preprint posted online on March 9, 2024 [FREE Full text] [doi: 10.48550/arXiv.2403.05881]
10. Kung TH, Cheatham M, Medenilla A, Sillos C, de Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
11. Lim DYZ, Ke YH, Sng GGR, Tung JYM, Chai JX, Abdullah HR. Large language models in anaesthesiology: use of ChatGPT for American Society of Anesthesiologists Physical status classification. Br J Anaesth 2023;131(3):e73-e75. [doi: 10.1016/j.bja.2023.06.052] [Medline: 37474421]
12. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature 2023;620(7972):47-60. [doi: 10.1038/s41586-023-06221-2] [Medline: 37532811]
13. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus 2023;15(5):e39305. [doi: 10.7759/cureus.39305] [Medline: 37378099]
14. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
15. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023;90:104512. [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
16. Yang R, Ning Y, Keppo E, Liu M, Hong C, Bitterman DS, et al. Retrieval-augmented generation for generative artificial intelligence in medicine. arXiv. Preprint posted online on June 18, 2024 [FREE Full text] [doi: 10.48550/arXiv.2406.12449]
17. Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med 2024;83(1):83-86. [doi: 10.1016/j.annemergmed.2023.08.003] [Medline: 37690022]
18. Ke Y, Yang R, Liu N. Comparing open-access database and traditional intensive care studies using machine learning: bibliometric analysis study. J Med Internet Res 2024;26:e48330. [doi: 10.2196/48330] [Medline: 38630522]
19. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
20. Satya-Murti S, Lockhart J. Recognizing and reducing cognitive bias in clinical and forensic neurology. Neurol Clin Pract 2015;5(5):389-396. [doi: 10.1212/CPJ.0000000000000181] [Medline: 29443168]
21. Nascimento N, Alencar P, Cowan D. Self-adaptive large language model (LLM)-based multiagent systems. 2023 Presented at: 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C); September 25-29, 2023; Toronto, ON p. 104-109 URL: http://arxiv.org/abs/2307.06187 [doi: 10.1109/ACSOS-C58168.2023.00048]
22. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. arXiv. Preprint posted online on August 16, 2023 [FREE Full text] [doi: 10.48550/arXiv.2308.08155]
23. Wang Z, Zhang G, Yang K, Shi N, Zhou W, Hao S, et al. Interactive natural language processing. arXiv. Preprint posted online on May 22, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.13246]
24. Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature 2023 Nov;623(7987):493-498 [FREE Full text] [doi: 10.1038/s41586-023-06647-8] [Medline: 37938776]
25. Légaré F, Adekpedjou R, Stacey D, Turcotte S, Kryworuchko J, Graham ID, et al. Interventions for increasing the use of shared decision making by healthcare professionals. Cochrane Database Syst Rev 2018;7(7):CD006732. [doi: 10.1002/14651858.CD006732.pub4] [Medline: 30025154]
26. Vela MB, Erondu AI, Smith NA, Peek ME, Woodruff JN, Chin MH. Eliminating explicit and implicit biases in health care: evidence and research needs. Annu Rev Public Health 2022;43:477-501 [FREE Full text] [doi: 10.1146/annurev-publhealth-052620-103528] [Medline: 35020445]
27. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2023-09-16]
28. National Organization for Rare Disorders. URL: https://rarediseases.org/ [accessed 2023-11-16]
29. GPT-4. OpenAI. URL: https://openai.com/research/gpt-4 [accessed 2024-01-01]
30. Hackman JR, Vidmar N. Effects of size and task type on group performance and member reactions. Sociometry 1970;33(1):37-54. [doi: 10.2307/2786271]
31. Rosen MA, DiazGranados D, Dietz AS, Benishek LE, Thompson D, Pronovost PJ, et al. Teamwork in healthcare: Key discoveries enabling safer, high-quality care. Am Psychol 2018;73(4):433-450 [FREE Full text] [doi: 10.1037/amp0000298] [Medline: 29792459]
32. Mull N, Reilly JB, Myers JS. An elderly woman with 'heart failure': cognitive biases and diagnostic error. Cleve Clin J Med 2015;82(11):745-753 [FREE Full text] [doi: 10.3949/ccjm.82a.14087] [Medline: 26540325]
33. Brown C, Nazeer R, Gibbs A, Le Page P, Mitchell AR. Breaking bias: the role of artificial intelligence in improving clinical decision-making. Cureus 2023;15(3):e36415 [FREE Full text] [doi: 10.7759/cureus.36415] [Medline: 37090406]

XSL•FO
RenderX

34. Tetsuhara K, Tsuji S, Nakano K, Kubota M. Case report: heart failure in dilated cardiomyopathy mimicking asthma triggered by pneumonia. BMJ Case Rep 2017;2017:bcr2017222082 [FREE Full text] [doi: 10.1136/bcr-2017-222082] [Medline: 29127129]
35. Ilaiwy A, Thompson NE, Waheed AA. Adenoma mimicking hyponatremia of SIAD. BMJ Case Rep 2018;2018:bcr2018226458 [FREE Full text] [doi: 10.1136/bcr-2018-226458] [Medline: 30269093]
36. Deming M, Mark A, Nyemba V, Heil EL, Palmeiro RM, Schmalzle SA. Cognitive biases and knowledge deficits leading to delayed recognition of cryptococcal meningitis. IDCases 2019;18:e00588 [FREE Full text] [doi: 10.1016/j.idcr.2019.e00588] [Medline: 31360635]
37. Khawaja H, Font C. Common and external iliac artery occlusion in Behçet's disease: a case of anchoring bias. BMJ Case Rep 2020;13(12):e236554 [FREE Full text] [doi: 10.1136/bcr-2020-236554] [Medline: 33298479]
38. Birch EM, Torres Molina M, Oliver JJ. Not like the textbook: an atypical case of ectopic pregnancy. Cureus 2022;14(10):e29881 [FREE Full text] [doi: 10.7759/cureus.29881] [Medline: 36348920]
39. Kyere K, Aremu TO, Ajibola OA. Availability bias and the COVID-19 pandemic: a case study of legionella pneumonia. Cureus 2022;14(6):e25846 [FREE Full text] [doi: 10.7759/cureus.25846] [Medline: 35832749]
40. Miyagami T, Nakayama I, Naito T. What causes diagnostic errors? referred patients and our own cognitive biases: a case report. Am J Case Rep 2022;23:e935163 [FREE Full text] [doi: 10.12659/AJCR.935163] [Medline: 35301273]
41. Kamegai K, Yokoyama S, Takakura S, Takayama Y, Shiiki S, Koyama H, et al. Syphilitic osteomyelitis in a patient with HIV and cognitive biases in clinical reasoning: a case report. Medicine (Baltimore) 2022;101(40):e30733 [FREE Full text] [doi: 10.1097/MD.0000000000030733] [Medline: 36221388]
42. Kawahigashi T, Harada Y, Watari T, Harada T, Miyagami T, Shikino K, et al. Missed opportunities for diagnosing vertebral osteomyelitis caused by influential cognitive biases. Am J Case Rep 2022;23:e936058 [FREE Full text] [doi: 10.12659/AJCR.936058] [Medline: 35729859]
43. Salama ME, Ukwade P, Khan AR, Qayyum H. Facial swelling mimicking anaphylaxis: a case of superior vena cava syndrome in the emergency department. Cureus 2022;14(9):e29678 [FREE Full text] [doi: 10.7759/cureus.29678] [Medline: 36320962]
44. Chehayeb RJ, Ilagan-Ying YC, Sankey C. Addressing cognitive biases in interpreting an elevated lactate in a patient with type 1 diabetes and thiamine deficiency. J Gen Intern Med 2023;38(6):1547-1551 [FREE Full text] [doi: 10.1007/s11606-023-08091-w] [Medline: 36814053]
45. Vittorelli J, Cacchillo J, McCool M, McCague A. Cognitive bias in the management of a critically Ill 29-year-old patient. Cureus 2023;15(5):e39314. [doi: 10.7759/cureus.39314] [Medline: 37351237]
46. Stanzelova A, Debray A, Allali S, Belhadjer Z, Taha MK, Cohen JF, et al. Severe bacterial infection initially misdiagnosed as MIS-C: caution needed. Pediatr Infect Dis J 2023;42(6):e201-e203. [doi: 10.1097/INF.0000000000003896] [Medline: 36916866]
47. Lieder F, Griffiths TL, M Huys QJ, Goodman ND. The anchoring bias reflects rational use of cognitive resources. Psychon Bull Rev 2018;25(1):322-349. [doi: 10.3758/s13423-017-1286-8] [Medline: 28484952]
48. Richie C. Environmentally sustainable development and use of artificial intelligence in health care. Bioethics 2022;36(5):547-555 [FREE Full text] [doi: 10.1111/bioe.13018] [Medline: 35290675]
49. Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW, et al. A translational perspective towards clinical AI fairness. NPJ Digit Med 2023;6(1):172 [FREE Full text] [doi: 10.1038/s41746-023-00918-4] [Medline: 37709945]
50. Papathanasiou IV, Kleisiaris CF, Fradelos EC, Kakou K, Kourkouta L. Critical thinking: the development of an essential skill for nursing students. Acta Inform Med 2014;22(4):283-286 [FREE Full text] [doi: 10.5455/aim.2014.22.283-286] [Medline: 25395733]
51. Yang R, Zeng Q, You K, Qiao Y, Huang L, Hsieh CC, et al. Ascle-a Python natural language processing toolkit for medical text generation: development and evaluation study. J Med Internet Res 2024 Oct 03;26:e60601 [FREE Full text] [doi: 10.2196/60601] [Medline: 39361955]
52. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]

Abbreviations

AI: artificial intelligence
EMR: electronic medical record
LLM: large language model

XSL•FO
RenderX

Edited by A Schwartz; submitted 12.04.24; peer-reviewed by WHK Chiu, E Amini-Salehi; comments to author 12.05.24; revised version received 21.06.24; accepted 12.09.24; published 19.11.24.

Please cite as:
Ke Y, Yang R, Lie SA, Lim TXY, Ning Y, Li I, Abdullah HR, Ting DSW, Liu N
Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study
J Med Internet Res 2024;26:e59439
URL: https://www.jmir.org/2024/1/e59439
doi:10.2196/59439
PMID:

©Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel Shu Wei Ting, Nan Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Large Language Models to Abstract Complex Social Determinants of Health From Original and Deidentified Medical Notes: Development and Validation Study

Alexandra Ralevski[1,2], PhD; Nadaa Taiyab[3], MALD; Michael Nossal[2], MA; Lindsay Mico[2], MS; Samantha Piekos[1], PhD; Jennifer Hadlock[1,4], MD

[1]Institute for Systems Biology, Seattle, WA, United States
[2]Providence Health & Services, Renton, WA, United States
[3]Tegria, LLC, Madison, WI, United States
[4]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

Corresponding Author:
Jennifer Hadlock, MD
Institute for Systems Biology
401 Terry Ave N
Seattle, WA, 98121
United States
Phone: 1 732 1359
Email: jhadlock@isbscience.org

Abstract

Background: Social determinants of health (SDoH) such as housing insecurity are known to be intricately linked to patients' health status. More efficient methods for abstracting structured data on SDoH can help accelerate the inclusion of exposome variables in biomedical research and support health care systems in identifying patients who could benefit from proactive outreach. Large language models (LLMs) developed from Generative Pre-trained Transformers (GPTs) have shown potential for performing complex abstraction tasks on unstructured clinical notes.

Objective: Here, we assess the performance of GPTs on identifying temporal aspects of housing insecurity and compare results between both original and deidentified notes.

Methods: We compared the ability of GPT-3.5 and GPT-4 to identify instances of both current and past housing instability, as well as general housing status, from 25,217 notes from 795 pregnant women. Results were compared with manual abstraction, a named entity recognition model, and regular expressions.

Results: Compared with GPT-3.5 and the named entity recognition model, GPT-4 had the highest performance and had a much higher recall (0.924) than human abstractors (0.702) in identifying patients experiencing current or past housing instability, although precision was lower (0.850) compared with human abstractors (0.971). GPT-4's precision improved slightly (0.936 original, 0.939 deidentified) on deidentified versions of the same notes, while recall dropped (0.781 original, 0.704 deidentified).

Conclusions: This work demonstrates that while manual abstraction is likely to yield slightly more accurate results overall, LLMs can provide a scalable, cost-effective solution with the advantage of greater recall. This could support semiautomated abstraction, but given the potential risk for harm, human review would be essential before using results for any patient engagement or care decisions. Furthermore, recall was lower when notes were deidentified prior to LLM abstraction.

(J Med Internet Res 2024;26:e63445) doi:10.2196/63445

KEYWORDS

housing instability; housing insecurity; housing; machine learning; artificial intelligence; AI; large language model; LLM; natural language processing; NLP; electronic health record; EHR; electronic medical record; EMR; social determinants of health; exposome; pregnancy; obstetric; deidentification

XSL•FO
RenderX

Introduction

The overwhelming majority of patients in the United States have their data stored in electronic health records (EHRs). Information regarding a patient's exposure to social determinants of health (SDoH), such as housing status, employment status, education, and quality of domestic life, provides relevant information that informs patient care and provides valuable avenues for intervention and treatment [1]. It has been estimated that SDoH can affect almost 50% of country-level variation in health outcomes, while clinical care impacts as little as 20% [2]. Housing data, in particular, including a patient's recent housing status, are known to be intricately linked to their health status [3-5]. Therefore, gaining insight into a patient's current and past living situation is essential to providing more complete and equitable care. It is also important for research, where capturing longitudinal exposome data is essential for analysis of health outcomes.

Housing stability is known to exist on a continuum, from complete stability (access to housing of reasonable quality in the absence of threats) to complete instability (no access to housing of reasonable quality) [6]. It is well known that people who are experiencing housing instability are at greater risk for other health issues, including substance use, comorbidities, and mental illness [3,7,8]. People facing housing instability are also at an increased risk for homelessness [5,9,10], which is associated with increased risk of morbidity and mortality [7]. Patients experiencing homelessness are also more likely to end up in the emergency department (ED), have longer hospital stays than low-income housed persons, and are less likely to use preventive services [3,10]. Women who are experiencing housing instability while pregnant face additional challenges, as they usually require consistent access to care throughout their pregnancy. Adverse exposures prior to and during pregnancy can put a child at an increased risk of both short- and long-term health consequences, and it is known that women who experience housing instability during pregnancy are at higher risk of adverse pregnancy outcomes, including preeclampsia, preterm birth, neonatal intensive care unit admission, and maternal morbidity [11-14].

SDoH information is rarely well documented in structured EHR data [15-17]. This leads to access barriers for health care teams and researchers. In addition, manually identifying SDoH, for example, through chart abstraction, is time-consuming, expensive, and impractical to scale. Because structured data have often been optimized for purposes other than individual care or research, free-text descriptions capture greater breadth and complexity of a patient's social and behavioral history. Existing projects, such as the Protocol for Responding to & Assessing Patients' Assets, Risks & Experiences [18], and emerging national interoperability plans [19] are providing paths for clinicians to better capture SDoH data in structured fields, but widespread data standards for data harmonization are still in early development [20].

Non–large language model (LLM)–based methods for extraction of SDoH information from free-text notes have relied heavily on identification of keywords or phrases, using either manual or semiautomated lexicon curation or rule-based methods [21-24]. However, most of these models are vulnerable to false positives (FP) and can capture only simplified concepts related to SDoH. In addition, previous research to identify housing instability from the EHR has focused primarily on homelessness or simplified housing-related concepts [24-26]. However, because housing instability is heterogeneous with many intersecting dimensions, classifying a patient experiencing housing instability can be more complex than some aspects of social history and exposures, such as smoking. By contrast, LLMs such as OpenAI's Generative Pre-trained Transformer (GPT) models can handle large quantities of complex, unstructured data using only simple prompts. Research using LLMs on EHR data is still in its early stages, and most work has focused on either fine-tuning models for medical relevance [27-30], comparing model performance to identify the presence or absence of SDoH statements [31,32], or using LLMs for disease diagnosis or phenotyping [33,34]. In addition, it has not been made clear whether the quality of note text flagged as relevant by GPT is similar to that of a human abstractor, and whether or not it is likely to contain hallucinatory text. LLMs such as GPT could also perpetuate health inequity if they perform differently for different patient populations. It is therefore important to test for bias in these models to inform future decisions regarding the use of LLMs in the health care setting. In addition, the possibility of using deidentified clinical notes for abstraction is appealing for further supporting patient privacy. However, deidentification processes involve obfuscation of important details, including dates and locations. This may alter the semantic underpinnings of a given text, making it difficult for an LLM to accurately identify and label SDoH within a given note.

We examined whether LLMs were able to identify housing instability in clinical free-text notes with greater accuracy than manual abstraction, regular expressions (RegEx), and a pretrained named entity recognition model for SDoH, using EHRs for a population of pregnant women. We also examined the ability to distinguish between present and past instability, the possibility of algorithmic bias in the predictions made by GPT-4 and GPT-3.5, and LLM performance on deidentified versions of patient notes. In addition, we asked both LLMs for a justification of their housing label, helping to provide a better understanding of how these models understand and interpret data. These methods provide new insight relevant to housing security and use of LLMs for SDoH data that is valuable for retrospective research and to help identify patients who might benefit from proactive outreach.

Methods

Ethical Considerations

This retrospective study protocol was performed in compliance with the Health Insurance Portability and Accountability Act Privacy Rule and was approved by the institutional review board at Providence Health and Services study (2020000783). Consent was waived because disclosure of protected health information (PHI) for the study was determined to involve no more than a minimal risk to the privacy of individuals. All data remained

XSL•FO
RenderX

within the secure Providence system. Manual and automated review of notes was conducted solely by the authors, using internal Providence resources.

Study Setting and Participants

Providence Health and Services is an integrated US community health care system that provides care in urban and rural settings across 7 states: Alaska, California, Montana, Oregon, New Mexico, Texas, and Washington. Using the Providence EHRs, we identified deliveries from June 8, 2010, through May 29, 2023 (N=595,600). We included singleton deliveries in a cohort of pregnant people aged 18-44 years at the start of pregnancy (n=557,406) as previously described [35]. We limited the deliveries to records that had associated gravida, term, preterm, abortion, and living data information, and to patients who received care in the Providence system during pregnancy. For patients with more than 1 pregnancy episode, we randomly selected a single episode (n=408,158). We limited our patient cohort to those with complete Social Vulnerability Index (n=372,208) information as previously described [35]. To identify patients from our cohort who were experiencing housing instability ("preliminary positive class"), we searched for patients who had either a Systematized Nomenclature of Medicine Clinical Terminology (SNOMED-CT) code for housing instability (Table S1 in Multimedia Appendix 1) or a matching string for the word "homeless" in 1 or more of their free-text notes (n=13,024). Patients who did not meet these criteria were considered in the "preliminary negative class" (n=359,184; Figure 1 [35]).

Figure 1. Cohort selection and experimental design. *Pregnant patients were selected as previously described. †For the positive class, charts were selected that included the SNOMED code for housing insecurity or had 1 or more notes containing the word "homeless." ‡The most recent note tagged by each method was selected. GPT 3.5T: GPT 3.5 Turbo; JSL: John Snow Labs; LLM: large language model; RegEx: regular expressions.

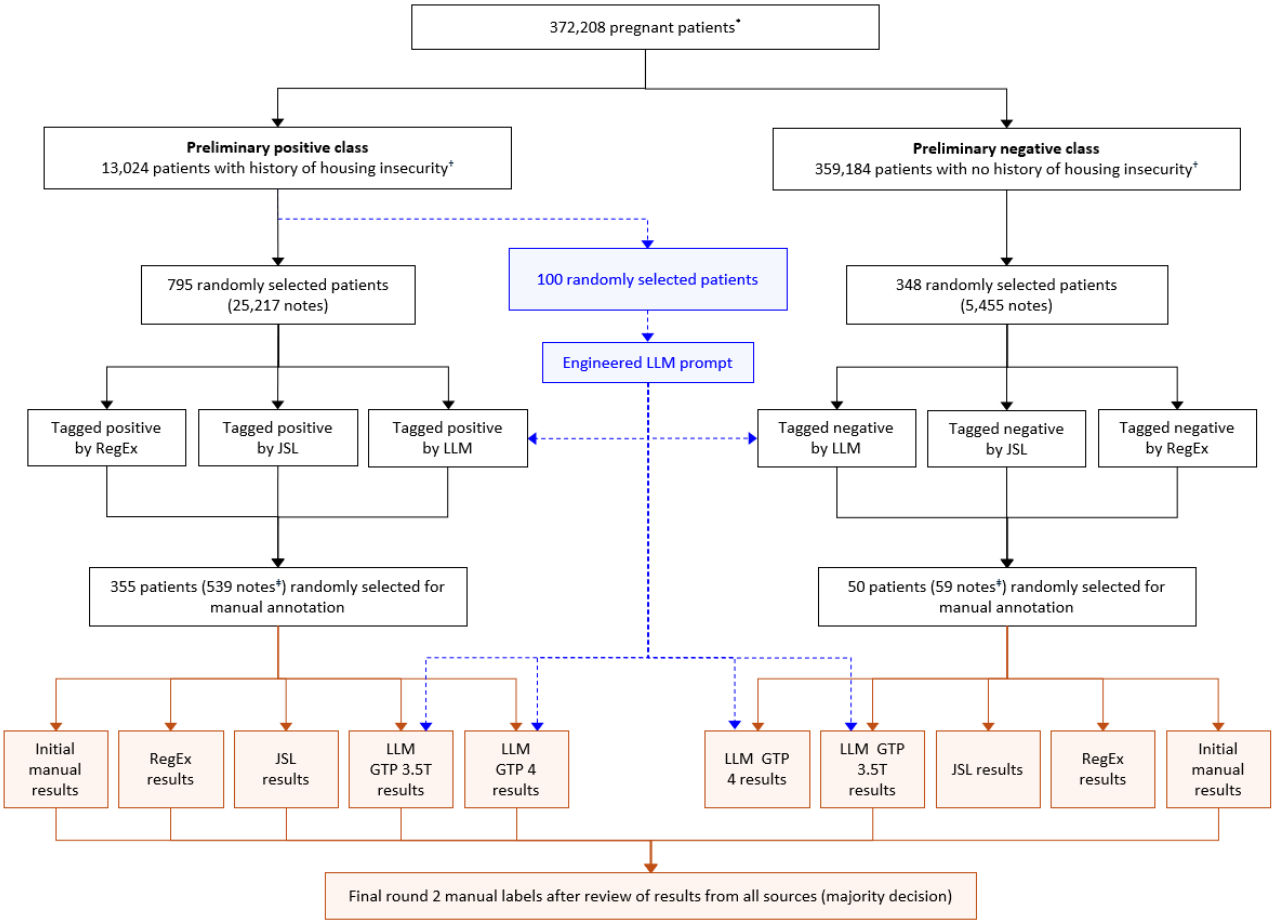

Models

Data processing was accomplished using the Azure AI Services Application Programming Interface for GPT-4 and GPT-3.5 within the secure Providence cloud environment. GPT-4 version 0613 had a 32K token window [36], while GPT-3.5 Turbo (hereafter, GPT-3.5) version 0613 had a 16K token window [37]. Both GPT models were run using LangChain and OpenAI libraries. The John Snow Labs (JSL) named entity recognition model (ner_sdoh_en_4.4.3_3.0_1686654976160) [36] is a state-of-the-art SDoH model designed to detect and label SDoH entities within text data. The housing-specific label includes entities related to the conditions of the patient's living spaces, for example, homeless, housing, small apartment, and so forth. JSL was run using sparknlp_jsl.version 5.0.0. To determine whether a RegEx search would identify relevant patient notes related to housing, we generated a preliminary list of keywords and phrases related to housing instability. The final list was reviewed by a registered nurse and clinical informaticist with 3 decades of experience (Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Prompt Engineering

Our prompt was developed using chain-of-thought prompting, where the problem or question description is initially stated and the LLM is asked to identify relevant evidence first and then provide an answer. This method has been shown to be more accurate than asking the LLM to provide only an answer [37]. GPT-4 and GPT-3.5 were asked to first identify chunks of evidence verbatim from the text (*evidence*). The model was then asked to go through each of the 4 labels: *housing noted*, *housing instability current*, *housing stability current*, and *housing instability history* and provide an answer for each. The model was then asked to provide a *justification* explaining why it chose a specific label and the LLMs were explicitly asked not to make up any information. GPT-3.5 was not used in the prompt engineering phase, and the prompt developed for GPT-4 was also used for GPT-3.5. The final prompt can be found in Multimedia Appendix 1.

Task Definition and Data Labeling

We defined different instances of housing stability and instability by first carrying out interviews with various subject matter experts, including clinicians, social workers, and resource specialists. All abstraction was conducted by researchers AR, NT, and MN. We generated preliminary abstraction guidelines that were then iteratively refined and finalized with additional input from subject matter experts. The final abstraction guidelines can be found in Multimedia Appendix 1. These guidelines distinguish between stable and unstable housing versus an *unknown* housing status, with examples for each pulled from EHR notes. An explicit definition of *history of housing instability* was also created. If a note contained any information on housing it had to be labeled as *stably housed*, *current housing instability*, or *history of/past housing instability*. If the note contained no information on housing, it was labeled as *unknown*. These guidelines were then used to create the prompt used by the LLM. The abstraction was divided into 2 rounds, with each reviewer abstracting their own set of notes for the first round (*original label*). For the second round, 25% of the notes underwent dual abstraction. If there was any disagreement for a given note between 2 abstractors, a third abstractor was asked to make a final decision. All the notes were then manually compared with results from GPT-4 to identify any notes that were obviously missed by reviewers. After the second round, a *final* label was assigned to each note. Performance metrics for manual abstraction were calculated by comparing the *original* and *final* labels for each note. After manual abstraction from reviewers, interrater reliability was calculated using Cohen k on the notes that underwent dual abstraction during the second round of abstraction. The average time spent manually abstracting a single note was calculated by each reviewer timing himself or herself for the time spent to abstract 10 notes and then taking the average. The 2 averages for each reviewer were then averaged.

For prompt engineering, we randomly selected 100 patients from the preliminary positive class and extracted all patient notes the year prior to the patient's conception date. Of the 100 patients, 51 had at least 1 note within the year prior to conception date, for a total of 1569 notes. Of those 51 patients, we used the JSL model to identify 70 notes from 16 patients related to housing. Each of the 70 notes were then manually abstracted by 2 different abstractors and the answers were compared. Any disagreement between the abstractors was discussed and a final decision was made. From our abstraction guidelines we developed an initial prompt for GPT-4 that contained definitions of housing instability as stated in the abstraction guidelines, as well as examples of housing instability from samples found in the patient notes. This prompt was then run through GPT-4 on the 70 notes and the results were compared with the manually abstracted results. All of the results were then compared, and the prompt was updated again based on results from GPT-4. For example, GPT-4 initially misclassified several cases of past housing instability as *current housing instability*. We then updated the prompt to specify that "a patient can only experience a 'history' of housing instability if they had housing instability in the past, then were stably housed, then experienced housing instability again. If the note refers to past housing instability, for example, 'the patient was homeless in the past,' then this can be treated as a 'history of housing instability.'"

Patient and Note Selection for Abstraction

Overall, the 4 methods flagged 25,217 notes from 795 patients from the preliminary positive class as being related to housing or housing instability past or present. If a given method flagged more than 1 note for a specific patient, only the most recent note in relation to the conception date was used for abstraction. For example, if GPT-4 flagged 2 notes dated January 1, 2019, and May 1, 2019, reviewers abstracted only the note dated May 1, 2019. This ensured that a maximum of 4 notes per patient were used for abstraction. Model performance was measured by examining accuracy, recall, precision, and F_1-score using the *SciKit-Learn* library.

Because it was uncertain how many notes would be flagged by the models for a given number of patients, we initially selected 500 patients from the preliminary positive class. Out of those 500 patients, 295 patients had 1 or more notes within a 12-month period prior to pregnancy, for a total of 9451 notes. A full breakdown of the number of patients and notes used in each round of model tagging and abstraction can be found in Table 1 and Table S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Breakdown of the number of notes used in each round of model tagging and abstraction from each model for the positive and negative classes.

Class and type of notes	Notes tagged by models for housing instability annotation, n				
	JSL[a]	GPT-3.5	GPT-4	RegEx[b]	Any
Positive class: manual annotation, round 1					
Most recent notes	114	73	74	82	204
Any tagged notes	300	181	284	226	511
Positive class: manual annotation, round 2					
Most recent notes	169	115	119	133	335
Any tagged notes	552	440	546	434	900
Negative class					
Most recent notes	27	11	12	14	59
Any tagged notes	33	12	15	27	79

[a]JSL: John Snow Labs.

[b]RegEx: regular expressions.

GPT Bias Evaluation

FPR (false-positive rate) and FNR (false-negative rate) were calculated using FP, FN (false negatives), TP (true positives), and TN (true negatives), derived from the confusion matrix:

$$FPR = FP/FP + TN$$

$$FNR = FN/TP + FN$$

Moreover, 95% CIs for a population proportion were calculated using the following formula: $CI = \hat{p} \pm z \times SE$, where z=1.96 for a 95% CL. SE was calculated using $SE = \sqrt{(\hat{p} \times (1 - \hat{p}) / n)}$, where $\hat{p}$ = FPR or FNR, and n = sample size.

Deidentification of Patient Notes

Providence has an existing corpus of deidentified notes that was created using a sequence of operations performed on text data to remove PHI [38]. These operations included multiple pretrained machine learning models and RegEx. Two versions of deidentified notes were used: *complete de-id*, in which all PHI was obfuscated and all dates were shifted or masked if shifting was not possible, and *de-id except date*, in which all PHI was obfuscated but the dates were not shifted. All notes remained within the secure Providence cloud environment.

Results

Manual Abstraction of EHR Notes

From the 25,217 notes from the 795 patients, the 4 automated methods flagged a total of 1411 notes (Table S3 in Multimedia Appendix 1). Given how the models were designed, JSL was able to flag only those notes related to housing in general (housing noted), while RegEx could not distinguish between current and past housing instability. Both GPT-4 and GPT-3.5 were able to flag notes related to general housing status, housing instability current, and housing instability past. After the models were run on the 25,217 notes, we selected notes that were flagged as *housing noted* for JSL, *housing instability current or past* for RegEx, and *currenthousing instability* or *pasthousing instability* for GPT-4 and GPT-3.5.

The most commonly flagged note types were assessments, plan of care, obstetric triage, history and physical, consults, discharge summary, and ED notes, indicating that these types of notes are most likely to provide relevant information related to housing or housing instability (Table S4 in Multimedia Appendix 1). We selected the 539 most recent notes from 355 patients for manual abstraction. Demographic characteristics of the 355 patients can be found in Table S5 in Multimedia Appendix 1. Of the 539 manually abstracted notes, the most common type of notes were progress notes (n=216) followed by ED provider notes (n=102) and telephone encounters (n=38; Table S6 in Multimedia Appendix 1).

Of the 182 patients who were identified as having current or past housing instability, only 18% (33/182) of patients had a structured SNOMED-CT code related to housing instability in their chart (Table 2). Although the percentage of patients with structured codes for housing instability is low, it is higher than what has been previously reported, and it is well known that structured fields do not adequately capture a patient's housing status [31,39]. These results could reflect that these notes were selected from patients flagged for housing instability, or that health care teams are more likely to ask about and document housing instability when a patient is pregnant. A total of 10% (11/109) of patients who were labeled as *stably housed* and 11% (7/64) of patients labeled as *unknown* had a SNOMED-CT code related to housing instability (Table 2). This is likely because researchers analyzed notes only within 1 year of pregnancy, while the SNOMED-CT code could have been added to a patient's chart at any time before pregnancy.

XSL•FO
RenderX

Table 2. Number of patients in each housing category who either did or did not have a SNOMED-CT[a] code related to housing instability in their chart[b].

Housing label	Patients with a SNOMED-CT code for homelessness, n	
	Yes	No
Stably housed	11	98
Current housing instability	32	131
Past housing instability	1	18
Unknown	7	57

[a]SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

[b]SNOMED-CT codes used to identify housing instability can be found in Table S1 in Multimedia Appendix 1.

Two abstractors manually abstracted the 539 notes and flagged 415 as related to housing. Of those 415, a total of 164 (30.4%) were labeled as *stably housed*, 223 (41.4%) were labeled as *current housing instability*, 28 (5%) were labeled as a *history of housing instability*, and 124 (23%) were labeled *unknown* (Table 3). Moreover, 25% (103/415) of the notes underwent dual abstraction. Before adjudication, dually abstracted notes had a Cohen k coefficient of 0.589, which reflects moderate agreement [40] and highlights the ambiguity and subjectivity of abstracting complex concepts such as housing instability compared with other SDOH mentions such as employment or parental status, despite the abstractors having been trained on the same guidelines (see Supplementary Methods in Multimedia Appendix 1).

Table 3. Number and Percentage of manually abstracted notes by housing label.

Housing label	Notes (n=539), n (%)
Current housing instability	223 (41.4)
Stably housed	164 (30.4)
Unknown	124 (23)
History of housing instability	28 (5)

Identification of Current and Past Housing Instability

Because housing status can change over time, it is important to be able to distinguish between current and past housing instability. This may be especially important for certain groups of patients, such as pregnant women, where housing instability during pregnancy may have different implications than prior housing instability. Tables 4 and 5 and Tables S7 and S8 in Multimedia Appendix 1 illustrate the differences in performance metrics between GPT-4, GPT-3.5, RegEx, and manual abstraction in identifying notes related to current or past housing instability, measured against final adjudicated labels. For current and past housing instability, the recall of GPT-4 was higher (0.924) than that of GPT-3.5 (0.717), RegEx (0.649), and manual abstraction (0.702). However, manual abstraction had the highest precision among the 4 methods (0.971), compared with GPT-4 (0.850), GPT-3.5 (0.759), and RegEx (0.632). The low performance of RegEx was due, in part, to acronyms related to housing, such as *supportive living services* and *recreational vehicle*, that also serve as equivalent medical shorthand for terms such as *single limb support* and *review*. This highlights the shortcomings of using a RegEx-based approach when attempting to identify a complex concept such as housing instability. For identifying current housing instability, GPT-4 still had higher recall than both GPT-3.5 and manual abstraction, but both LLMs had lower precision than manual abstraction. The recall for GPT-3.5 was higher for current housing instability alone, indicating that this model struggled to identify notes where past housing instability was mentioned, further evidenced in Table 6. This demonstrates that while LLMs can identify past or current instances of an event such as housing instability, specific models should be tested for their performance in each category individually.

Table 4. Comparison of recall and precision for RegEx[a], GPT-3.5, GPT-4, and manual abstraction in identifying notes with current or past housing instability, measured on 539 manually abstracted notes.

Model	Recall	Precision
RegEx	0.649	0.632
GPT-3.5	0.717	0.759
GPT-4	0.924	0.85
Manual annotation	0.702	0.971

[a]RegEx: regular expressions.

XSL•FO
RenderX

Table 5. Comparison of recall and precision for GPT-3.5, GPT-4, and manual abstraction in identifying notes with current housing instability, measured on 539 manually abstracted notes.

Model	Recall	Precision
GPT-3.5	0.753	0.734
GPT-4	0.906	0.831
Manual annotation	0.718	0.933

Table 6. Recall and precision metrics for GPT-4 and GPT-3.5 for each housing label measured on 539 manually abstracted notes

Housing label	GPT-4		GPT-3.5	
	Recall	Precision	Recall	Precision
Stable housing	0.39	0.88	0.34	0.67
Current housing instability	0.91	0.83	0.75	0.73
Past housing instability	0.68	0.63	0.18	0.62
Unknown	0.82	0.53	0.68	0.38

The drop in recall for manual abstraction highlights the high level of ambiguity associated with abstracting this type of information, as well as the energy needed for humans to conduct manual abstraction for extended periods of time. Examples of relevant notes, along with the housing labels and justifications from GPT-4 and GPT-3.5, can be found in Table 7. There were several ambiguous cases related to housing in patient notes. For example, “Was in domestic violence. Daughters are [redacted ages] and living with parents for now.” In this example, it is not explicitly stated where the patient is staying, making her current housing status *unknown*. However, this was flagged as *current instability* by GPT-4 and *stably housed* by GPT-3.5. In another example, a patient was living with her family and their landlord was a meth dealer. However, the patient’s boyfriend was actively trying to find them a new living situation, which made the patient happy. While both abstractors agreed that the housing situation was not stable, this type of instability was not explicitly mentioned in the note prompt for the LLMs. The reviewers labeled the note as *current instability*, but the note was tagged as *stably housed* by GPT-4 and *current instability* by GPT-3.5.

XSL•FO
RenderX

Table 7. Examples of relevant note text and manual abstraction labels compared with GPT-4 and GPT-3.5 and relevant justification[a].

Manual abstraction	GPT-4 label	GPT-3.5 label	Text selected from longer note by human abstractor	Justification GPT-4	Justification GPT-3.5
Current instability	Current instability	Current instability	She has been under a great deal of stress over the past year and recently became homeless. She and her boyfriend have moved to this area where she has family where she is now staying.	The patient note mentions that the patient recently became homeless, indicating current housing instability. The patient is now staying with family, but this is likely a temporary arrangement given the recent homelessness and the reference to a halfway house, which is a form of transitional housing. There is no evidence in the note to suggest a history of housing instability prior to the current situation.	The evidence shows that the patient recently became homeless and is currently facing housing instability. The patient's history of frequent kidney stones and recent surgeries for kidney stones further suggests a lack of stable housing.
Current instability	Past instability	Current instability	She was having many stressors at that time which included being kicked out of her mother's house and staying with ex BF.	The patient note mentions that the patient was kicked out of her mother's house and was staying with her ex-boyfriend. This indicates a past instance of housing instability. However, the note does not provide current information about the patient's housing situation, so it is unknown whether the patient is currently stably housed or not.	The patient note mentions that the patient was kicked out of her mother's house and was staying with her ex-boyfriend, which indicates current housing instability. In addition, the patient's active problem list includes depression, which can be associated with housing instability. However, there is no explicit evidence or mention of past housing instability, so it is unknown whether the patient has a history of housing instability.
Current instability	Current instability	No housing noted, current instability unknown	She said she is staying at a motel.	The patient note mentions that the patient is currently staying at a motel, which indicates a form of housing instability. There is no evidence in the note to suggest that the patient has faced housing instability in the past.	There is no explicit evidence or mention of the patient's housing status in the note. The patient mentions staying at a motel, but it is unclear whether this is a temporary situation or a stable housing arrangement. Therefore, the current housing stability is unknown.
Current instability	Stably housed	Current instability	Pt reports she is living w/[redacted]; they are renting a house from [redacted]; he is a meth dealer. [Bf] has made arrangements to get them into a new living situation. She is very happy about this.	The patient is currently living with [redacted], in a house they are renting. There is no evidence of housing instability at the moment. The boyfriend has made arrangements for a new living situation, which the patient is happy about. There is no mention of past housing instability.	The evidence from the patient note indicates that [the patient] is currently facing housing instability. She is living in a rented house with [redacted], but the house is owned by [redacted] who is a meth dealer. In addition, the note mentions that [the patient] is in the process of moving to [redacted] and has been applying for jobs. This indicates that her current housing situation is unstable. There is no evidence of past housing instability mentioned in the note.
Unknown	Current instability	Stably housed	States lots of changes lately. Was in domestic violence. Daughters are [redacted] and living with parents for now.	The patient note mentions that the patient's daughters are currently living with her parents, which suggests that the patient is temporarily staying with family members. This is an indication of current housing instability. There is no evidence in the note to suggest a history of housing instability.	The patient's housing status is noted in the evidence. The patient is currently stably housed as she is living with her parents. There is no evidence of current or past housing instability.

[a]Certain sections of note text and evidence from GPT-4 and GPT-3.5 have been removed to protect patient privacy.

A minority of notes contained specific references to housing, such as "patient lives in the woods of [redacted location] with her boyfriend." This phrase signified likely homelessness to most abstractors, but one abstractor assumed that this meant a

XSL•FO
RenderX

cabin in the woods. This case was caught by GPT-4 but missed by RegEx and GPT-3.5. There were additional cases where housing instability was explicitly mentioned and was missed by 1 or more reviewers. This most often occurred in longer notes that contained a significant amount of information, and only 1-2 sentences related to housing, for example "section 8 voucher," which refers to a US program for assisting very low-income families. This sentence was missed by GPT-3.5 and RegEx but was correctly identified by GPT-4. In addition, there were several notes that contained no information ("blank" notes). In several cases, GPT-3.5 used sentences from the prompt text as evidence and justification and flagged the note as current or past instability. This did not occur with GPT-4. This is likely because the LLM was asked to provide text evidence verbatim, and GPT-3.5 used the prompt because no relevant note text was available. However, the researchers found no instances of hallucinated evidence in any of the GPT-4 responses that were reviewed, suggesting that requiring verbatim evidence from LLMs can be a solution to hallucinated responses.

Although GPT-3.5 struggled to identify several cases of housing instability, the researchers could not identify a consistent trend in the type or content of the FP or FN notes. However, there were several cases where GPT-3.5 listed known risk factors mentioned elsewhere in the note as evidence of housing instability. For example, it noted a patient's frequent kidney surgeries or a depression diagnosis as justification for housing instability, although there was clear mention of housing instability elsewhere in the note (Table 7). This was not the case with GPT-4, which used only direct evidence from the note text that mentioned housing-related terms as justification. Although kidney disease and depression are associated with housing instability [41-44], a human abstractor would not use this as a justification for housing instability. Having an LLM use this as justification could be a potential concern but could also be an opportunity for a different use case, where researchers ask LLMs to identify potential risk factors observed in a set of records. Those results might show bias in LLMs or highlight patterns that humans overlook. These examples also demonstrate significant differences between the 2 GPT releases and highlight the value of providing evidence and justification for every note when using LLMs. In most cases, the evidence gathered by the reviewers was similar or identical but the interpretation differed. Similarly, the evidence gathered by GPT-4 was similar to the reviewers in most cases (although this was not always the case with GPT-3.5). This indicates that, while manual abstraction is likely to yield more accurate results, GPT-4 could be used to rapidly gather relevant note text for computer-assisted manual review, helping save time without losing important or relevant patient information.

Identification of General Housing Status

Table 8 and Table S9 in Multimedia Appendix 1 show the performance of GPT-3.5 and GPT-4 compared with JSL and manual abstraction in identifying notes where housing was mentioned. GPT-4 outperformed both GPT-3.5 and JSL across all 4 metrics but had a slightly worse precision compared with manual abstraction (0.936 compared with 0.952), although recall was higher (0.781 compared with 0.720). Interestingly, the majority of cases that were missed by GPT-4 were instances where housing was stable, for example, "she lives at home with her children" or "patient was requesting to go home." A possible underlying cause was that the prompt had been heavily focused on identifying cases of housing instability, with little guidance provided on identifying housing status overall. Prompt engineering focused on different proportions of relevant information might yield different and more accurate results.

Table 8. Comparison of recall and precision for JSL[a], GPT-3.5, GPT-4, and manual abstraction in identifying notes where housing was noted, measured on 539 manually abstracted notes.

Model	Recall	Precision
JSL	0.733	0.864
GPT-3.5	0.675	0.875
GPT-4	0.781	0.936
Manual annotation	0.72	0.952

[a]JSL: John Snow Labs.

LLM Performance by Housing Category

Table 6 and Tables S10 and S11 in Multimedia Appendix 1 show the differences in performance for GPT-4 and GPT-3.5 in identifying notes across the different housing categories: stable housing, current housing instability, past housing instability, or unknown. GPT-3.5 performed worse than GPT-4 across all categories and had particularly low recall for notes labeled as *past instability* compared with GPT-4, which had a higher recall than precision in this category. Both GPT-4 and GPT-3.5 demonstrated poor recall for *stable housing* notes. A possible underlying cause was that the prompt focused more heavily on housing instability and did not provide training focused on stable housing. These data indicate a performance improvement for the GPT-4 release and demonstrate the effects of prompt engineering on the model outcome.

LLM Bias Evaluation

To test for bias in the LLMs, we used the fairness criteria of separation [45] to examine the FPR and FNR between GPT-4 and GPT-3.5 across 3 different housing labels (housing noted, housing instability past or current, and housing instability current) across the age (18-30 years and 31-44 years), race (American Indian and Alaska Native, Asian, Black, Native Hawaiian and other Pacific Islander, Unknown or Declined, White, and Other), and ethnic (Hispanic or Latino, not Hispanic or Latino, and Unknown or Declined) demographic groupings. We then examined the 95% CIs of the FPR and FNR for each

XSL•FO
RenderX

group (Tables S12-S17 in Multimedia Appendix 1 and Figures S1-S3 in Multimedia Appendix 1). According to the fairness criteria of separation, any difference in FPR and FNR between groups suggests potential algorithmic bias. We did observe differences in FPR and FNR between all of the groups within the 3 demographic categories. However, when we examined the overlap of the 95% CIs between groups, we found that in all cases, except in cases where the sample sizes were extremely small (n<5), there was overlap between CIs for all the groups, suggesting that the differences between groups are not significant. However, further work with a larger sample population is needed.

Cost Breakdown of LLMs Compared with Manual Abstraction

There was a substantial difference in cost between GPT-4 and GPT-3.5, as shown in Tables 9 and 10, due to the increase in cost per 1000 tokens for notes for GPT-4 (US $0.06) compared with GPT-3.5 (US $0.003). The output cost also increased from US $0.004 per 1000 tokens for GPT-3.5 to US $0.12 per 1000 tokens for GPT-4. Interestingly, the prompt (US $1778.39 for GPT-4) cost more than the total for all the notes (US $701.72 for GPT-4), and this was the case for GPT-3.5 as well. This is because the prompt had to be included as part of each note. Because the prompt was long (1182 tokens), this increased the cost substantially. Future work comparing model performance in relation to prompt length would provide valuable insight into this trade-off.

The cost of manual abstraction varies by location, but in the United States, it can be estimated to be the minimum wage per hour for that state. As of January 2024, in the state of Washington, the minimum wage was US $16.28 per hour [46]. To analyze 25,217 notes would have cost approximately US $9442, substantially higher than either LLM.

Table 9. Cost analysis of GPT-4 compared with GPT-3.5 for 25,217 notes from 795 patients[a].

Model	Costs (US $)			
	Prompt	Notes	Output	Total
GPT-4	1788.39	701.72	181.35	2671.46
GPT-3.5	89.42	35.09	8.89	133.40

[a]GPT-4 prompt: 1182 tokens × 25217 notes × 0.06 per 1000 tokens. GPT-4 notes: 11,695,270 tokens × 0.06 per 1000 tokens. GPT-4 output: 15,11,244 tokens × 0.12 per 1000 tokens. GPT-3.5 prompt: 1182 tokens × 25,217 notes × 0.003 per 1000 tokens. GPT-3.5 notes: 11,695,270 tokens × 0.003 per 1000 tokens. GPT-3.5 output: 22,23,674 tokens × 0.004 per 1000 tokens.

Table 10. Estimated cost of running a single note on GPT-4 and GPT-3.5 compared with manual abstraction.

Model	Cost (US $)
Manual annotation	0.374
GPT-4	0.106
GPT-3.5	0.005

Analysis on Patients in the Preliminary Negative Class

To evaluate the performance of our 4 methods compared with patients from the preliminary positive class, we selected a random sample of 5455 notes from 348 patients in the preliminary negative class. GPT-4, GPT-3.5, RegEx, and JSL were run on all 5455 notes using the same method as notes from patients in the preliminary positive class. Of the 5455 notes, all 4 methods flagged only 59 of the most recent notes from 50 patients with 1 of the 4 housing labels (GPT-4: n=12, GPT-3.5: n=11, RegEx: n=14, and JSL: n=27). A randomly selected sample of 20 notes from the preliminary negative class that were not flagged by any of the 4 methods was reviewed by a manual abstractor (AR). Of the 20 notes, 19 notes did not contain any information on housing and 1 note suggested stable housing. An additional randomly selected sample of 20 notes was taken from the 59 notes that were flagged by the 4 methods and reviewed by a manual abstractor (AR). Of the 20 notes, 8 were *current housing instability*, 1 was *past housing instability*, 5 were *stably housed*, and 6 were *unknown*.

This demonstrates that, as expected, patients from the preliminary negative class had far fewer notes related to housing or housing instability than patients in the preliminary positive class. To generate a manually abstracted data set of ~500 notes to compare with the preliminary positive class, the 4 methods would need to analyze approximately 46,000 notes. However, JSL is not able to distinguish between notes related to general housing and housing instability, and RegEx has low precision and recall for identifying notes related to housing instability. Therefore, LLMs provide the best chance of finding relevant notes related to housing or housing instability for patients in either class. Because the number of notes flagged by GPT-4 and GPT-3.5 were very low in the preliminary negative class (12 and 11, respectively, out of 5455 notes), the 2 models would need to analyze approximately 227,300 notes to find ~500 notes related to housing or housing instability in the preliminary negative class. Due to the cost restrictions of running these LLMs, we were unable to perform this analysis. However, future work to analyze additional notes from patients in the preliminary negative class could provide insight into any differences in notes between these 2 classes.

Evaluation of GPT-4 on Deidentified Patient Notes

Deidentification can help mitigate privacy risks to individuals to support secondary use of data for research. In the United

XSL•FO
RenderX

States, the Health Insurance Portability and Accountability Act specifies 18 categories of information that are PHI that must be removed from medical records [47,48]. However, while the process of deidentification is necessary to protect patient privacy, the information that is removed during this process, such as dates and locations, may result in the loss of important contextual clues needed for LLM analysis of housing instability. We wanted to examine whether LLMs performed similarly on two versions of deidentified patient notes compared with original notes: (1) fully deidentified notes where all PHI was obfuscated and all dates were shifted or masked if shifting was not possible (hereafter referred to as *complete de-id*), and (2) patient notes where all PHI was obfuscated but the dates were not shifted (hereafter referred to as *de-id except date*). All notes remained within the secure Providence system. All data processing was conducted in Providence's secure cloud environment.

We ran the 2 deidentified versions of the 539 manually abstracted notes through GPT-4 and compared the performance metrics with the original notes to identify current or past housing instability or general housing status (housing noted). We found that in all cases, recall dropped but precision increased for the deidentified notes compared with the original notes (Multimedia Appendix 2 and Table S18 in Multimedia Appendix 1). For example, for the notes labeled as *current housing instability*, the recall for GPT-4 on the original notes was 0.906, but this dropped to 0.812 and 0.834 for the *complete de-id* and *de-id except date* notes, respectively. By contrast, the precision increased from 0.831 for the original notes compared with 0.862 and 0.849 for the *complete de-id* and *de-id except date* notes, respectively. These minor increases in precision are likely because results from GPT-4 are slightly different each time the model is run, resulting in slight differences in performance. The drop in recall is not surprising given the nature of deidentification, in which both places and locations have been obfuscated, making it more difficult for the model to identify relevant notes. For example, there were several cases in the original notes where the patient was stated to be living in a specific location, such as a city or county, but these locations were changed to medical facilities or, in one instance, a jail, resulting in the model sometimes mislabeling the patient as *unstably housed* or missing the note as related to housing altogether. In other cases, when the dates were shifted, instances of *past housing instability* were made current, making it difficult for the LLM to properly identify and label these notes.

Discussion

Principal Findings

Our results demonstrate the potential benefit of using LLMs to identify instances of complex SDoH concepts in the EHR, such as past or current housing instability. Although manual abstraction correctly classified more notes related to housing instability, it is more expensive than using LLMs, with a limited increase in performance. GPT-4 outperformed GPT-3.5, JSL, and RegEx in identifying patients experiencing current or past housing instability. In most cases, the evidence from GPT-4 was similar or identical to that of the manual abstractors, and no hallucinations were observed in GPT-4 output. Our work also suggests that requiring the LLM to provide verbatim evidence and justification from the original text can help reduce the risk that relevant context about housing information is omitted from LLM results. We also found that LLMs perform well in identifying notes related to housing insecurity without the need for fine-tuning, using only few-shot learning. However, given the sensitive nature of housing insecurity and complexity of individual situations, final human review would be essential for any decision regarding patient communication or patient care. Another important finding is that recall was lower on notes after they had been deidentified. Further research is needed to determine whether human annotators would also have lower performance on deidentified notes.

It is important to note that housing instability does not exist in a vacuum; oftentimes, there are multiple compounding factors that can either contribute to or be influenced by housing instability, including domestic violence, drug abuse, and mental illness. One limitation of this study was our focus solely on housing instability and not an additional identification of these risk factors. This resulted in some cases where a patient was technically considered to have stable housing, but there were other risk factors in the patient note that would likely be important for users of abstraction results: case workers, clinicians, or researchers. Expanding the prompt might improve performance and enable labeling that separates out multiple dimensions of housing security, including uncertainty about future housing, frequency of housing transitions, and risks from unsafe housing situations. Because GPT-4 and GPT-3.5 are not deterministic models, responses, and therefore performance, may also change if rerun on the same notes. However, the newest release of GPT, GPT-4 Turbo, allows researchers to add a deterministic seed parameter to ensure that the model returns the same response every time, helping to prevent changes in performance across multiple runs.

Because all our methods required that each note be analyzed individually, all 4 methods identified 1411 notes out of 25,217 related to housing or housing instability in the preliminary positive class, and only 64 out of 5455 notes related to housing or housing instability in the preliminary negative class, we can conclude that many notes in this study likely did not contain information on housing or housing instability. However, future work could investigate the similarities and differences in note content related to housing or housing instability between patients in both classes. In addition, because we used automated methods for the initial selection of relevant patient notes, we likely missed some patient notes related to housing or housing instability that were not captured with any of the 4 automated methods. Future work to manually abstract a larger corpus of patient notes related to housing and housing instability, as well as other SDoH categories, would be valuable for full validation. Furthermore, repeat validation over time would be important to reduce the risk of drift, and new tests would be appropriate before deploying the system in different geographic settings. . Another limitation is that the study was limited to the content documented in EHR notes, and a recent survey reported that only about 60% of patients felt comfortable sharing SDoH-related information [49]. Future studies would benefit from longitudinal confidential surveys or interviews with

XSL•FO
RenderX

patients and health care teams. In addition, because the time and cost to run GPT-4 and GPT-3.5 might not be feasible across millions of patient notes, work with newly emerging open source language models may provide a similar performance for a much lower cost and run time.

Conclusions

Methods for abstracting SDoH data can be valuable for retrospective research, chart review for prospective trials, and population health interventions to identify those who might benefit from proactive outreach. This work demonstrates that LLMs have potential for computer-assisted abstraction of social history, improving recall and reducing costs. This includes temporal feature engineering, such as identifying whether patient experienced housing instability before or during their pregnancy. At the same time, precision was slightly lower than that of human abstractors, and given the potential risk for harm to patients, human review would be essential for any decisions on patient communication or patient care. Results also identified 2 important areas where further work is needed: separating out 3 different dimensions of housing insecurity and advancing deidentification methods that do not result in loss of information on social history.

Acknowledgments

Support was provided in part by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through the Biomedical Data Translator program (award 3OT2TR00344301S1). Any opinions expressed in this document are those of the Translator community at large and do not necessarily reflect the views of NCATS, individual Translator team members, or affiliated organizations and institutions.

Data Availability

Results have been aggregated and reported within this paper to the extent possible while maintaining privacy from personal health information as required by US law. All data are archived within Providence systems in a Health Insurance Portability and Accountability Act–secure audited compute environment, and those wishing to verify study conclusions can contact the chief data officer. All biomedical codes used to extract data from electronic health records have been shared. The source code used to run models and reproduce results shown for this study is available at GitHub [52].

Authors' Contributions

AR, NT, MN, and SNP had full access to all data in the study and take responsibility for the integrity of the data and accuracy of the data analysis. AR, NT, MN, LM, and JH led the concept and design. AR, NT, MN, and SNP contributed to acquisition, analysis, and interpretation of data. AR, NT, MN, and SNP drafted of the manuscript. All authors contributed to critical revision of the manuscript. AR and NT led the statistical analysis. JH and LM did the supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods, tables, and figures.
[DOCX File , 314 KB - jmir_v26i1e63445_app1.docx]

Multimedia Appendix 2
Comparison of GPT-4 performance in identifying (A) current or past housing instability, (B) current housing instability, or (C) general housing status from 3 different versions of the same set of 539 notes. Note versions were original patient notes, completely deidentified notes (complete de-id), or deidentified notes dates unchanged (de-id except date).
[PNG File , 73 KB - jmir_v26i1e63445_app2.png]

References

1. Social determinants of health. World Health Organization. URL: https://www.who.int/teams/social-determinants-of-health [accessed 2023-11-29]
2. Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: relationships between determinant factors and health outcomes. Am J Prev Med 2016;50(2):129-135. [doi: 10.1016/j.amepre.2015.08.024] [Medline: 26526164]
3. Liu M, Hwang SW. Health care for homeless people. Nat Rev Dis Primers 2021;7(1):5. [doi: 10.1038/s41572-020-00241-2] [Medline: 33446661]
4. Kushel MB, Vittinghoff E, Haas JS. Factors associated with the health care utilization of homeless persons. JAMA 2001;285(2):200-206. [doi: 10.1001/jama.285.2.200] [Medline: 11176814]

XSL•FO
RenderX

5. Easton C, Javadizadeh E, Plett P, Perez SN, Isard R, Hariramani S, et al. A scoping review examining the association of housing quality and psychosocial well-being following homelessness: an ecological systems perspective. J Soc Distress Homeless 2023:1-25. [doi: 10.1080/10530789.2023.2241215]
6. Frederick TJ, Chwalek M, Hughes JR, Karabanow J, Kidd S. How stable is stable? defining and measuring housing stability. J Community Psychol 2014;42(8):964-979. [doi: 10.1002/jcop.21665]
7. Fazel S, Geddes JR, Kushel M. The health of homeless people in high-income countries: descriptive epidemiology, health consequences, and clinical and policy recommendations. Lancet 2014;384(9953):1529-1540 [FREE Full text] [doi: 10.1016/S0140-6736(14)61132-6] [Medline: 25390578]
8. Fazel S, Khosla V, Doll H, Geddes J. The prevalence of mental disorders among the homeless in western countries: systematic review and meta-regression analysis. PLoS Med 2008;5(12):e225 [FREE Full text] [doi: 10.1371/journal.pmed.0050225] [Medline: 19053169]
9. Pruitt AS, Barile JP. Actionable research for understanding and addressing homelessness. J Community Psychol 2022;50(5):2051-2057. [doi: 10.1002/jcop.22878] [Medline: 35545867]
10. Kushel MB, Gupta R, Gee L, Haas JS. Housing instability and food insecurity as barriers to health care among low-income Americans. J Gen Intern Med 2006;21(1):71-77 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.00278.x] [Medline: 16423128]
11. Huang K, Waken RJ, Luke AA, Carter EB, Lindley KJ, Joynt Maddox KE. Risk of delivery complications among pregnant people experiencing housing insecurity. Am J Obstet Gynecol MFM 2023;5(2):100819. [doi: 10.1016/j.ajogmf.2022.100819] [Medline: 36436788]
12. DiTosto JD, Holder K, Soyemi E, Beestrum M, Yee LM. Housing instability and adverse perinatal outcomes: a systematic review. Am J Obstet Gynecol MFM 2021;3(6):100477 [FREE Full text] [doi: 10.1016/j.ajogmf.2021.100477] [Medline: 34481998]
13. Leifheit KM, Schwartz GL, Pollack CE, Edin KJ, Black MM, Jennings JM, et al. Severe housing insecurity during pregnancy: association with adverse birth and infant outcomes. Int J Environ Res Public Health 2020;17(22):8659 [FREE Full text] [doi: 10.3390/ijerph17228659] [Medline: 33233450]
14. Green JM, Fabricant SP, Duval CJ, Panchal VR, Cahoon SS, Mandelbaum RS, et al. Trends, characteristics, and maternal morbidity associated with unhoused status in pregnancy. JAMA Netw Open 2023;6(7):e2326352 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.26352] [Medline: 37523185]
15. Wang M, Pantell M, Gottlieb L, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: measures and associated insights. J Am Med Inform Assoc 2021;28(12):2608-2616 [FREE Full text] [doi: 10.1093/jamia/ocab194] [Medline: 34549294]
16. Palacio AM, Suarez M, del Toro Y, Seo D, Garay D, Pareja D, et al. Integrating social determinants of health into the electronic health records of a large health system: a qualitative perspective. Perspectives in Health Information Management 2018(Summer 2018):1-19 [FREE Full text]
17. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. Health Aff (Millwood) 2018;37(4):585-590 [FREE Full text] [doi: 10.1377/hlthaff.2017.1252] [Medline: 29608369]
18. Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences. PRAPARE. 2021. URL: https://prapare.org/ [accessed 2023-12-08]
19. Social determinants of health. Office of Disease Prevention and Health Promotion. URL: https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health [accessed 2020-01-13]
20. The Office of the National Coordinator for Health Information Technology (ONC). Social determinants of health. HealthIT.gov. URL: https://www.healthit.gov/health-equity/social-determinants-health [accessed 2023-12-07]
21. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev 2022;46:100511. [doi: 10.1016/j.cosrev.2022.100511]
22. Patra BG, Sharma MM, Vekaria V, Adekkanattu P, Patterson OV, Glicksberg B, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc 2021;28(12):2716-2727 [FREE Full text] [doi: 10.1093/jamia/ocab170] [Medline: 34613399]
23. Han S, Zhang RF, Shi L, Richie R, Liu H, Tseng A, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. J Biomed Inform 2022;127:103984 [FREE Full text] [doi: 10.1016/j.jbi.2021.103984] [Medline: 35007754]
24. Lybarger K, Dobbins NJ, Long R, Singh A, Wedgeworth P, Uzuner Ö, et al. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. J Am Med Inform Assoc 2023;30(8):1389-1397 [FREE Full text] [doi: 10.1093/jamia/ocad073] [Medline: 37130345]
25. Bejan CA, Angiolillo J, Conway D, Nash R, Shirey-Rice JK, Lipworth L, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. J Am Med Inform Assoc 2018;25(1):61-71 [FREE Full text] [doi: 10.1093/jamia/ocx059] [Medline: 29016793]
26. Reeves RM, Christensen L, Brown JR, Conway M, Levis M, Gobbel GT, et al. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. J Biomed Inform 2021;120:103851 [FREE Full text] [doi: 10.1016/j.jbi.2021.103851] [Medline: 34174396]

XSL•FO
RenderX

27. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
28. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
29. Shah-Mohammadi F, Finkelstein J. AI-powered social determinants of health extraction from patient records: a GPT-based investigation. 2024 Presented at: 2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC); February 5-7, 2024; Laguna Hills, CA. [doi: 10.1109/AIMHC59811.2024.00028]
30. Fu Y, Ramachandran GK, Dobbins NJ, Park N, Leu M, Rosenberg AR, et al. Extracting social determinants of health from pediatric patient notes using large language models: novel corpus and methods. 2024 Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); May 20-25, 2024; Torino, Italia URL: https://aclanthology.org/2024.lrec-main.618
31. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024;7(1):6 [FREE Full text] [doi: 10.1038/s41746-023-00970-0] [Medline: 38200151]
32. Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM, et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. Cell Rep Med 2024;5(1):101356 [FREE Full text] [doi: 10.1016/j.xcrm.2023.101356] [Medline: 38232690]
33. Ahsan H, McInerney D, Kim J, Potter C, Young G, Amir S, et al. Retrieving evidence from EHRs with LLMs: possibilities and challenges. Proc Mach Learn Res 2024;248:489-505. [Medline: 39224857]
34. Alsentzer E, Rasmussen MJ, Fontoura R, Cull AL, Beaulieu-Jones B, Gray KJ, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. NPJ Digit Med 2023;6(1):212 [FREE Full text] [doi: 10.1038/s41746-023-00957-x] [Medline: 38036723]
35. Piekos SN, Hwang YM, Roper RT, Sorensen T, Price ND, Hood L, et al. Effect of COVID-19 vaccination and booster on maternal-fetal outcomes: a retrospective cohort study. Lancet Digit Health 2023;5(9):e594-e606 [FREE Full text] [doi: 10.1016/S2589-7500(23)00093-6] [Medline: 37537121]
36. GPT 4. OpenAI. URL: https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4 [accessed 2024-11-08]
37. GPT 3.5 Turbo. OpenAI. URL: https://platform.openai.com/docs/models#gpt-3-5-turbo [accessed 2024-11-08]
38. Social determinants of health (SDOH) named entity recognition (NER) model. John Snow Labs. 2020. URL: https://nlp.johnsnowlabs.com/2023/06/13/ner_sdoh_en.html [accessed 2024-06-19]
39. Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. 2022 Presented at: 36th Conference on Neural Information Processing Systems (NeurIPS 2022); November 28 to December 9, 2022; New Orleans, LA p. 22199-22213 URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf
40. Kocaman V, Talby D. Accurate clinical and biomedical named entity recognition at scale. Softw Impacts 2022;13:100373. [doi: 10.1016/j.simpa.2022.100373]
41. Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. Popul Health Manag 2021;24(2):222-230 [FREE Full text] [doi: 10.1089/pop.2020.0043] [Medline: 32598228]
42. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
43. Novick TK, Baweja M, American Society of Nephrology Healthcare Justice Committee AdvocacyScholarship Work Group, American Society of Nephrology Healthcare Justice Committee AdvocacyScholarship Work Group are. Housing: a critical contributor to kidney disease disparities. J Am Soc Nephrol 2022;33(8):1471-1473 [FREE Full text] [doi: 10.1681/ASN.2022040424] [Medline: 35545303]
44. Novick TK, Kushel M, Crews D. Unstable housing and kidney disease: a primer. Kidney Med 2022;4(4):100443 [FREE Full text] [doi: 10.1016/j.xkme.2022.100443] [Medline: 35372820]
45. Shelton KH, Taylor PJ, Bonner A, van den Bree M. Risk factors for homelessness: evidence from a population-based study. Psychiatr Serv 2009;60(4):465-472. [doi: 10.1176/ps.2009.60.4.465] [Medline: 19339321]
46. Padgett DK. Homelessness, housing instability and mental health: making the connections. BJPsych Bull 2020;44(5):197-201 [FREE Full text] [doi: 10.1192/bjb.2020.49] [Medline: 32538335]
47. Privacy tech-know blog: when worlds collide - the possibilities and limits of algorithmic fairness (part 1). Office of the Privacy Commissioner of Canada. 2023. URL: https://www.priv.gc.ca/en/blog/20230405_01/ [accessed 2024-03-26]
48. Minimum wage. Washington State Department of Labor & Industries. URL: https://lni.wa.gov/workers-rights/wages/minimum-wage/ [accessed 2024-03-06]
49. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32 [FREE Full text] [doi: 10.1186/1472-6947-8-32] [Medline: 18652655]
50. Office for Civil Rights, HHS. Standards for privacy of individually identifiable health information. Final rule. Fed Regist 2002;67(157):53181-53273 [FREE Full text] [Medline: 12180470]

XSL•FO
RenderX

51. Strawley C, Richwine C. Patient preferences with social needs information sharing. HealthIT.gov. URL: https://www.healthit.gov/buzz-blog/health-information-exchange-2/patient-preferences-with-social-needs-information-sharing [accessed 2023-11-24]
52. Hadlock-Lab/LLM_SDoH. GitHub. URL: https://github.com/Hadlock-Lab/LLM_SDoH [accessed 2024-11-07]

Abbreviations

ED: emergency department
EHR: electronic health record
FN: false positive
FNR: false-negative rate
FP: false negative
FPR: false-positive rate
GPT: Generative Pre-trained Transformer
JSL: John Snow Labs
LLM: large language model
PHI: protected health information
RegEx: regular expressions
SDoH: social determinants of health
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terminology
TN: true negatives
TP: true positives

Edited by G Eysenbach, X Ma; submitted 20.06.24; peer-reviewed by NJ Dobbins; comments to author 19.07.24; revised version received 31.08.24; accepted 25.09.24; published 19.11.24.

Please cite as:
Ralevski A, Taiyab N, Nossal M, Mico L, Piekos S, Hadlock J
Using Large Language Models to Abstract Complex Social Determinants of Health From Original and Deidentified Medical Notes: Development and Validation Study
J Med Internet Res 2024;26:e63445
URL: https://www.jmir.org/2024/1/e63445
doi:10.2196/63445
PMID:

©Alexandra Ralevski, Nadaa Taiyab, Michael Nossal, Lindsay Mico, Samantha Piekos, Jennifer Hadlock. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study

Junhyuk Seo[1,2*], BSN; Dasol Choi[1*], BA; Taerim Kim[1,3], MD, PhD; Won Chul Cha[1,3], MD, PhD; Minha Kim[3], MD; Haanju Yoo[4], PhD; Namkee Oh[5], MD; YongJin Yi[6], MD, MPH; Kye Hwa Lee[7], MD, PhD; Edward Choi[8], PhD

[1]Department of Digital Health, Samsung Advanced Institute of Health Sciences and Technology (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea

[2]Department of Nursing, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[4]NAVER Digital Healthcare Lab, Seongnam, Republic of Korea

[5]Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[6]Department of Internal Medicine, College of Medicine, Dankook University, Cheonan, Republic of Korea

[7]Department of Information Medicine, Asan Medical Center and University of Ulsan College of Medicine, Seoul, Republic of Korea

[8]Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Taerim Kim, MD, PhD
Department of Digital Health
Samsung Advanced Institute of Health Sciences and Technology (SAIHST)
Sungkyunkwan University
115, Irwon-ro, Gangnam-gu
Seoul, 06355
Republic of Korea
Phone: 82 010 7114 2342
Email: taerim.j.kim@gmail.com

Abstract

Background: The advancement of large language models (LLMs) offers significant opportunities for health care, particularly in the generation of medical documentation. However, challenges related to ensuring the accuracy and reliability of LLM outputs, coupled with the absence of established quality standards, have raised concerns about their clinical application.

Objective: This study aimed to develop and validate an evaluation framework for assessing the accuracy and clinical applicability of LLM-generated emergency department (ED) records, aiming to enhance artificial intelligence integration in health care documentation.

Methods: We organized the Healthcare Prompt-a-thon, a competitive event designed to explore the capabilities of LLMs in generating accurate medical records. The event involved 52 participants who generated 33 initial ED records using HyperCLOVA X, a Korean-specialized LLM. We applied a dual evaluation approach. First, clinical evaluation: 4 medical professionals evaluated the records using a 5-point Likert scale across 5 criteria—appropriateness, accuracy, structure/format, conciseness, and clinical validity. Second, quantitative evaluation: We developed a framework to categorize and count errors in the LLM outputs, identifying 7 key error types. Statistical methods, including Pearson correlation and intraclass correlation coefficients (ICC), were used to assess consistency and agreement among evaluators.

Results: The clinical evaluation demonstrated strong interrater reliability, with ICC values ranging from 0.653 to 0.887 ($P<.001$), and a test-retest reliability Pearson correlation coefficient of 0.776 ($P<.001$). Quantitative analysis revealed that invalid generation errors were the most common, constituting 35.38% of total errors, while structural malformation errors had the most significant negative impact on the clinical evaluation score (Pearson $r=-0.654$; $P<.001$). A strong negative correlation was found between the number of quantitative errors and clinical evaluation scores (Pearson $r=-0.633$; $P<.001$), indicating that higher error rates corresponded to lower clinical acceptability.

Conclusions: Our research provides robust support for the reliability and clinical acceptability of the proposed evaluation framework. It underscores the framework's potential to mitigate clinical burdens and foster the responsible integration of artificial

XSL•FO
RenderX

intelligence technologies in health care, suggesting a promising direction for future research and practical applications in the field.

(J Med Internet Res 2024;26:e58329) doi:10.2196/58329

KEYWORDS

large language models; health care documentation; clinical evaluation; emergency department; artificial intelligence; medical record accuracy

Introduction

Large language models (LLMs) have experienced rapid advancement in recent years, significantly influencing various domains, including health care. This remarkable development is supported by their increasingly sophisticated capabilities in understanding and generating human-like text [1], as seen in models like GPT-4 and its successors [2]. From creative writing and complex problem-solving to multilingual translation and personalized communication, LLMs are reshaping the landscape of numerous industries, transcending the traditional boundaries of the capabilities of artificial intelligence (AI).

Especially, LLMs have shown remarkable capabilities in the health care sector, beginning with their impressive grasp of medical knowledge. This is highlighted by their successful performance on challenging exams, such as the USMLE (United States Medical Licensing Examination), demonstrating their proficiency in medical concepts [3,4]. Moving beyond theoretical knowledge, LLMs are now being applied to practical tasks, including the creation of detailed medical documents, which indicates their potential for direct applications in clinical settings [5-7]. The significant potential of LLMs is further underscored by their applications, which range from augmenting clinical decision-making to providing personalized patient care [8]. Their ability to process vast amounts of medical literature and patient data suggests a transformative impact on both medical research and practice [9,10]. Moreover, the continuous improvement in the performance and accuracy of these models points to a promising future in enhancing health care delivery and patient outcomes.

However, despite these advancements, the integration of LLMs into clinical practice faces significant challenges. Concerns regarding the accuracy of generated content, particularly in scenarios involving hallucinations or fabrications [6,11,12], pose substantial risks in the high-stakes environment of health care. This issue goes beyond just the limitations of the LLM models themselves; it is significantly compounded by the absence of objective and reliable standards for evaluating the generated content. The lack of clear criteria on what constitutes good output or an effective prompt and the direction for necessary improvements lead to uncertainty in their practical application. In medical settings, where precise and accurate information is vital, this absence of clear standards for prompt efficacy and output evaluation poses significant hurdles to their practical implementation, hindering the reliable and safe integration of LLMs into clinical practices.

Our research focuses on 2 primary goals. First, we aim to establish a framework for evaluating the outputs of LLMs in medical records. This includes identifying and quantifying errors and benchmarking these against expert opinions. Second, our study seeks to assess whether the outcomes produced by LLMs are clinically acceptable and reliable, examining their potential to support clinical decision-making processes. By achieving these objectives, we strive to enable the safe and effective integration of LLMs into health care, thereby enhancing patient care and clinical operations.

Methods

Data Collection Process

Overview

To evaluate the potential of LLMs in health care, particularly in generating patient records, we organized the Healthcare Prompt-a-thon. This event was specifically designed to gather prompts and their outputs from participants interested in health care–focused generative LLMs. In this section, we discuss the process of the prompt collection, encompassing its organization, task creation, implementation of the competition, and the application of LLM technology. This comprehensive approach ensures a thorough understanding of how the event was structured and executed, as well as how the LLM technology was applied in this context.

Overall Information of the Healthcare Prompt-a-Thon

The Healthcare Prompt-a-thon, organized by the Korea Society of Artificial Intelligence in Medicine under the generative model research group and hosted by Samsung Medical Center with sponsorship from NAVER Cloud, aimed to explore the potential of LLMs in health care. The event registration, which was open from October 20 to November 17, 2023, through Social Network Services and internal hospital web networks, attracted participants interested in the application of generative models in health care documentation.

On November 17, 2023, at 115 Ilwon-ro, Gangnam-gu, Seoul, a total of 52 participants (comprising 20 solo and 16 duo teams) attended the event. These teams, made up of both solo and duo participants, contributed 33 prompts for evaluation. The prompt collection flow of the event is depicted in Figure 1.

The majority of participants were from health care organizations, medical centers, and universities, with a significant representation of researchers, master's students, and doctors, including radiologists. This diverse demographic underscored the health care focus and interdisciplinary nature of the event.

XSL•FO
RenderX

Figure 1. Prompt collection flow.

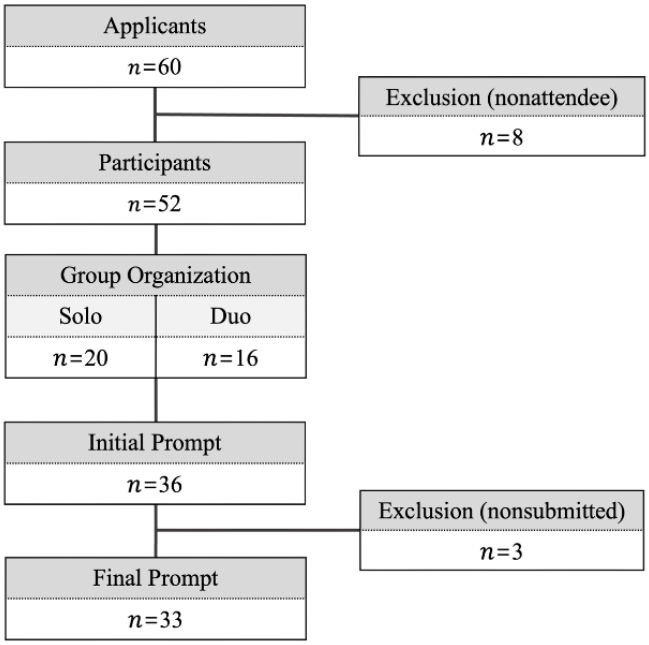

Task Development

The development of the tasks for the Healthcare Prompt-a-thon involved a collaborative effort spanning 2 months, engaging a diverse team of 5 professionals including an emergency medicine physician, transplant surgery physician, critical care nurse, engineer, and linguist. This team diligently created a new dataset, structuring it according to standard hospital dataset formats to ensure its relevance and applicability in a clinical setting. Although the data were simulated, they were carefully designed to mirror plausible scenarios in health care, leveraging the extensive expertise of the medical professionals on the team.

The competition comprised 3 main categories of tasks: creating initial emergency department (ED) records, comparing prescriptions, and generating discharge records. However, the primary focus of our study, as well as the evaluation process, was centered on the creation of initial ED records. For this main task, participants used detailed case scenarios provided to them to develop prompts. These prompts were designed to generate initial ED records based on various inputs, including prescreening notes, transcripts of dialogues between doctors and patients, and physical examination data. The process and nature of these tasks are illustrated in Figure 2.

To guide the participants, 2 example patient cases were provided, complete with input datasets and ideal responses for an initial consultation. These examples served as benchmarks for the participants, aiding in the development of their prompts. The evaluation of these prompts was carried out using a specially curated, unreleased patient dataset, enabling an objective assessment of their effectiveness in simulating real-world medical scenarios. The complexity and organization of these cases went beyond typical medical records, making them exemplary for both input and output in the competition. More detailed information about the patient cases can be found in Multimedia Appendix 1.

Figure 2. Process flow for generating initial emergency department (ED) records. The flow begins with inputting patient data (triage record, consultation transcript, and physical examination), followed by a "prompting" phase to create the initial ED record. The output is evaluated through "clinical evaluation" and "quantitative evaluation" for comprehensive assessment.

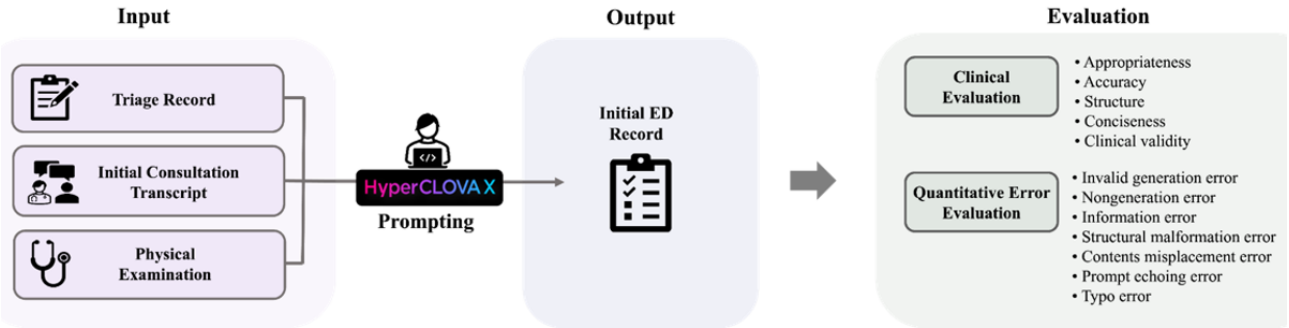

Competition Process

The Healthcare Prompt-a-thon was structured to maximize participant engagement and creativity. One week prior to the event, participants received a set of practice problems and descriptions of medical records to familiarize themselves with the tasks. On the day of the event, we unveiled the actual examples to be used in the competition.

XSL•FO
RenderX

At the outset of the event, we provided detailed explanations of the 3 main tasks: creating initial ED records, comparing prescriptions, and generating discharge records. Following this, participants were introduced to the LLM platform, along with a comprehensive tutorial on its use and strategies for effective prompt creation. This training was crucial in equipping participants with the necessary skills to navigate the LLM platform and create high-quality prompts.

The task execution phase spanned approximately 5 hours, from 10:00 AM to 3:00 PM, during which participants worked on generating the required outputs. Following the completion of the event, participants submitted their prompts through a Google Form. Each participant's prompts were submitted once to the LLM for evaluation. This approach was chosen to streamline the competition process and ensure consistency in the evaluation of outputs. These prompts were then executed in real time, and the generated results were used as the basis for evaluation. The evaluation of these prompts was conducted in the subsequent 1.5 hours. The judging panel, consisting of experts in relevant fields, assessed the submissions based on predefined criteria. The event culminated in an award ceremony, where the top 3 teams were recognized for their outstanding prompts and innovative solutions.

Using LLM in Our Process

Among various LLMs available for our study, we selected HyperCLOVA X developed by NAVER Corporation [13]. This decision was informed by HyperCLOVA X's specialized capabilities in processing the Korean language, an essential requirement for our project's focus on health care documentation in Korea.

HyperCLOVA X is a Korean-specialized LLM intricately designed to efficiently process and understand the complexities of the Korean language. It builds upon the strengths of its predecessor, CLOVA, with significant enhancements in language processing and generation, having been trained on an extensive database of over 560 billion Korean tokens. These improvements are evident in its outstanding proficiency in understanding and generating Korean text [14]. The integration of HyperCLOVA X with Clova Studio, a user-friendly graphical user interface [15], facilitates intuitive in-context interactions and lowers technical barriers for users, making it an ideal choice for real-time applications in health care settings [16-18].

Evaluation Methodology

Overview

To comprehensively evaluate the initial ED records generated through prompts, we used 2 distinct evaluation methodologies: (1) clinical evaluation by medical experts and (2) quantitative evaluation for objectively assessing errors in the generation process. The development of these evaluation criteria was a collaborative effort involving 3 experts, including an emergency medicine professor, a critical care nurse, and a linguist.

The process of developing the evaluation criteria involved several key steps to ensure their effectiveness and relevance. Initially, we conducted an extensive literature review to gather existing standards and best practices for medical documentation, including guidelines from reputable medical organizations and previous research on AI applications in health care. One notable reference was the Physician Documentation Quality Instrument (PDQI-9), which provides a comprehensive framework for evaluating the quality of medical records [19,20]. However, given the broad scope and time-consuming nature of the PDQI-9, we simplified the criteria to make them more practical for real-time evaluation by our assessors.

Next, the initial framework was presented to the panel of experts, who engaged in several rounds of discussions to refine the criteria. These discussions were crucial for integrating diverse perspectives and ensuring that all critical aspects of medical documentation were covered. Experts emphasized the importance of detailed subcriteria for specific areas such as the presence of hallucinations, missing information, and clinical reasoning. This feedback led to further refinement of the criteria.

To ensure the effectiveness and relevance of our evaluation criteria, a preliminary trial was conducted 2 weeks prior to the main event. This trial involved 10 randomly selected volunteers who engaged in solving prompt problems and generating outputs. The outcomes of this preliminary phase were instrumental in providing practical insights and feedback, which were crucial for refining and enhancing the evaluation criteria to be used in the competition. This process helped ensure that our evaluation criteria were both rigorous and aligned with real-world scenarios, thereby enhancing the validity of our study's findings.

Clinical Evaluation

For the clinical evaluation, we established criteria relevant to assessing the quality of medical record outputs, including appropriateness, accuracy, structure/format, conciseness, and clinical validity. These criteria were meticulously developed to cover key aspects of clinical documentation and are detailed in Table 1. Each element was evaluated using a 5-point Likert scale, where higher scores indicate a stronger presence of the evaluated quality attribute.

XSL•FO
RenderX

Table 1. Description of our combined criteria for evaluating prompt outputs: clinical evaluation criteria and quantitative evaluation criteria.

Category	Description
Clinical evaluation criteria	
Appropriateness	Whether the content of the output corresponds to each item appropriately, ensuring it fits the clinical context rather than just individual correspondence.
Accuracy	How well the output matches the provided information, including the presence of hallucinations or missing information.
Structure/format	If the output has an appropriate structure and format, with content organized as per the specified criteria rather than just in paragraph form.
Conciseness	The length of the output, aiming for brevity without compromising essential information.
Clinical validity	If the output exhibits sound clinical reasoning and is deemed acceptable within the context of medical practice.
Quantitative evaluation criteria	
Invalid generation error	Instances where the output includes content that should not have been generated, or outputs information not provided in the input data.
Nongeneration error	Cases where necessary items are not included in the output, or when the output lacks content that should have been generated.
Information error	Occurrences where specific information (such as numerical values, frequency, terminology, and places) is inaccurately generated or represented.
Prompt echoing error	Errors in the output that directly reflect the content of the prompt.
Structural malformation error	Cases where necessary items are missing or unnecessary items are included in the output.
Contents misplacement error	Instances where content that belongs to one category is incorrectly placed in another.
Typo error	Errors involving typos or grammatical mistakes in the generated text.

The evaluation was conducted by a panel of 4 medical experts, consisting of 3 physicians and a surgeon, all of whom are bioinformatics researchers. The evaluations were carried out in 2 phases: a collective assessment during the Healthcare Prompt-a-thon and subsequent independent evaluations to ensure thoroughness and mitigate bias. To minimize potential memory effects and learning bias from the first evaluation phase, a washout period of 1 month was observed before conducting the retest evaluations. Additionally, a consistency analysis was performed to gauge the uniformity of judgments across the panel of medical experts. The outcomes of both the test-retest evaluations and the consistency analysis are elaborated in the results section.

Quantitative Evaluation

In addition to clinical evaluation criteria, we introduce a novel quantitative evaluation framework specifically designed for assessing LLM outputs in health care. This framework's establishment was driven by a thorough review of the generated outputs, during which we identified recurring error patterns and grouped them into distinct categories. Consequently, errors in LLM-generated medical records have been categorized into 7 types: invalid generation error, nongeneration error, information error, prompt echoing error, structural malformation error, contents misplacement error, and typo error. These categories were formed based on the frequency and nature of errors observed in our comprehensive review of the LLM outputs. Detailed definitions of these error types are presented in Table 1.

Furthermore, specific examples of each error type are provided in Table 2. These examples illustrate the practical implications of the identified errors and provide insight into the types of inaccuracies that occurred in the LLM-generated medical records.

To facilitate a structured and detailed evaluation, we categorized the essential components of the initial ED records into 11 distinct items: chief complaint, vital signs, present illness, past history, personal and social history, system inquiry, physical examination, problem list, differential diagnosis, diagnosis plan, and treatment plan. We then counted the presence and frequency of errors in each item.

The results of this quantitative error analysis serve as the foundation for a comprehensive comparison between quantitative and clinical evaluations. This comparison includes both the overall correlation and a detailed analysis of individual evaluation criteria.

By adopting this dual evaluation approach, we gained a comprehensive understanding of the strengths and weaknesses of LLMs in health care from a quantitative standpoint, while also exploring the relationship between clinical and quantitative evaluations.

XSL•FO
RenderX

Table 2. Specific examples of each quantitative evaluation error type.

Error type	Ideal output	Error example	Justification of error
Invalid generation error	• Personal and social history • Travel history: none	• Personal and social history • Travel history: none • Smoking: none • Alcohol consumption: none	Erroneously adds information not present in original data such as smoking and alcohol consumption information.
Nongeneration error	• Chief complaint • Main symptom: abdominal pain • Occurrence date: November 16, 2023, 8:11 PM	• Chief complaint • Main symptom: abdominal pain	Omits essential information such as the occurrence date.
Information error	• Vital signs • Blood pressure: 125/79 mm Hg • Pulse: 98 beats per min • Respiration: 18 per min • Temperature: 38 °C • Oxygen saturation: 98%	• Vital signs • Blood pressure: 124/78 mm Hg • Pulse: 97 beats per min • Respiration: 20 per min • Temperature: 38.5 °C • Oxygen saturation: 99%	Inaccurately represents specific data such as blood pressure, pulse, etc.
Prompt echoing error	• Past medical history • None	• Past medical history: include a structured history of past surgeries and diagnosed conditions. • None	Echoes prompt content instead of providing concise, relevant information.
Structural malformation error	• Ideal output does not include the "Patient Arrival Information" category.	• Patient arrival information • Disease classification: disease • Means of arrival: taxi • Arrival route: home	Includes a category "Patient Arrival Information" that should not exist in the ideal record.
Contents misplacement error	• System inquiry • Fever: yes	• System Inquiry • Medication history: none • Allergy history: none • Surgical history: none • Trauma history: none	Content that should be in "Personal and Social History" category is incorrectly included in "System Inquiry."
Typo error	• Problem list 1. Abdominal pain 2. Fever	• Problem list 1. Abdominal pang 2. Fever	Contains a typographical error in the word "Abdominal Pain."

Ethical Considerations

During the application process for the Healthcare Prompt-a-thon, all participants were informed that their outputs might be used for research purposes and could potentially be made public. Explicit consent was obtained from each participant, ensuring that they were fully aware of the intended use of their data. While personal information was collected solely for the purpose of managing the event, it was not used for research and was securely discarded after the event concluded. No personal identifying information was collected for the research, and all data used in the study were anonymized to protect participants' privacy and confidentiality. As a token of appreciation for their participation, the top 3 teams were recognized and awarded special prizes (equivalent to US$613, US$110, and US$70). The research was conducted in compliance with ethical standards, and the institutional review board of Samsung Medical Center approved the study protocol (institutional review board 2023-12-018-001).

Statistical Analysis Methods

Statistical analyses of the evaluation results were conducted using the Python programming language (Python Software Foundation) along with libraries such as *SciPy*, *Pandas*, and *Pingouin*. The specific versions of these programs used in our analysis were Python 3.10, *Pandas* 1.5.3, *SciPy* 1.11.4, and *Pingouin* 0.5.4.

Clinical Evaluation Methods

For the clinical evaluation, we used a 2-phase approach consisting of an initial assessment during the event followed by a retest evaluation. Our key statistical tool in this phase was the Pearson correlation coefficient. This coefficient measured the consistency of scores between the initial test and the retest,

XSL•FO
RenderX

offering a quantifiable measure of reliability in the evaluators' judgments.

In addition, we applied the intraclass correlation coefficient (ICC), specifically the ICC(3,k) model, to assess the consistency among different experts in the panel. The ICC was instrumental in evaluating the level of agreement among raters, ensuring the robustness of our evaluation criteria across diverse expert opinions [20]. The *F* statistics associated with the ICC further reinforced the reliability of these evaluations.

Quantitative Evaluation Methods

Quantitative Evaluation Methods involved categorizing and counting various error types, such as invalid generation error, typo error, and structural malformation error, among others. Our approach was to first analyze the frequency of these errors and then assess their impact on the clinical validity of the LLM outputs.

To understand the relationship between error types and clinical evaluation scores, we used the Pearson correlation coefficient. We also calculated *P* values for each error type and category to ascertain the statistical significance of their correlations with clinical evaluation scores, considering a *P* value less than .001 as indicative of a significant correlation.

Additionally, we extended our analysis to encompass different categories of clinical information, such as chief complaint, vital signs, present illness, and so forth. This allowed us to delve deeper into understanding which aspects of the LLM outputs were most pivotal for clinical accuracy and relevance.

Results

Result of Clinical Evaluation

Overall Statistical Analysis

Table 3 presents the mean and SD for each criterion of the clinical evaluation, offering a more robust summary of central tendency and variability. In the initial test, the highest mean score was noted for structure/format at 3.30 (SD 0.99), indicating a strong organizational aspect in the responses. Conversely, accuracy had the lowest mean score at 2.61, highlighting a critical area for improvement, with an SD of 0.63 suggesting a tighter cluster of responses around the mean.

Table 3. Statistical results of clinical evaluation (n=33).

Item and stage	Mean (SD)	Pearson *r*	*P* value
Appropriateness		0.613	<.001
Test	3.03 (0.90)		
Retest	3.14 (0.66)		
Accuracy		0.643	<.001
Test	2.61 (0.63)		
Retest	2.88 (0.58)		
Structure/format		0.605	<.001
Test	3.30 (0.99)		
Retest	3.43 (0.67)		
Conciseness		0.684	<.001
Test	2.96 (0.76)		
Retest	3.23 (0.62)		
Clinical validity		0.789	<.001
Test	2.99 (0.99)		
Retest	2.89 (0.73)		
Overall score		0.776	<.001
Test	14.89 (3.96)		
Retest	15.57 (2.84)		

Upon retesting, improvements were notable across the board, with structure/format again achieving the highest mean score at 3.43 and showing a decrease in SD to 0.67, reflecting more consistent evaluations. Conciseness also showed significant progress, moving to a mean of 3.23 with a reduced SD of 0.62, indicating a narrower spread of evaluations. The overall score saw an increase from a mean of 14.89 in the initial test to 15.57 in the retest, with the SD narrowing from 3.96 to 2.84, demonstrating a general improvement in the evaluation criteria across participants.

Test-Retest Reliability Analysis

Our test-retest reliability analysis [21] method involved 2 distinct testing phases. To quantitatively determine the reliability of our evaluations, we calculated the Pearson correlation coefficient between the average scores of the 2 evaluation sessions. The results, illustrated in Figure 3, show a Pearson

XSL•FO
RenderX

correlation coefficient of 0.776, as reported in Table 3. This coefficient indicates a moderate positive correlation, suggesting a reasonable level of agreement among evaluators across both testing sessions. Notably, the clinical validity criterion exhibited the highest correlation, indicating strong consistency in this area, whereas the structure criterion showed a relatively lower correlation, highlighting potential variability in this aspect of the evaluation.

The statistical significance of this correlation was established with a *P* value less than .001, confirming that the observed consistency in the evaluation scores was not due to random chance.

Figure 3. Pearson correlation analysis showing the test-retest reliability of clinical evaluation criteria with a correlation coefficient of 0.776.

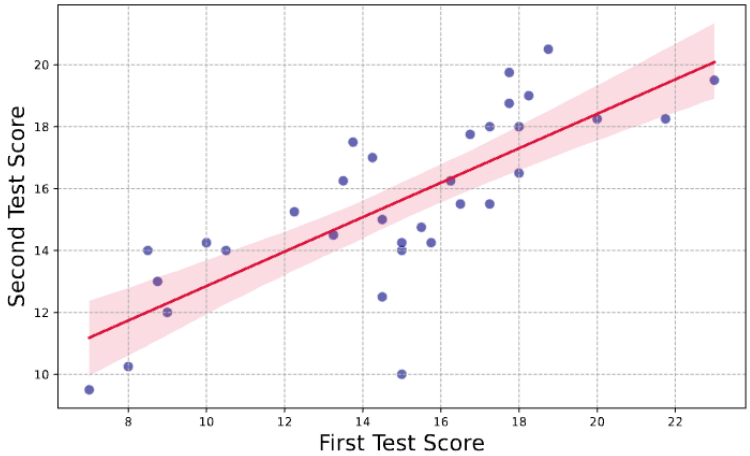

Consistency Among Panel

To ensure the robustness of our clinical evaluation criteria, external validation was performed by a diverse panel of domain experts. ICC, particularly using the ICC(3,k) model, provide a measure of the reliability or consistency of measurements made by multiple observers measuring the same quantity [22]. As summarized in Table 4, all items demonstrated strong ICC, when using ICC(3,k) model, with values ranging from 0.653 to 0.887. These ICC values indicate a high level of agreement among raters, with Clinical Validity showing the highest consistency (ICC=0.887). The *F* statistics, ranging from 2.883 to 8.826 for different criteria, further substantiate the reliability of the evaluations, suggesting high agreement consistency among the evaluators. Additionally, the 95% CI for each criterion reinforce the precision of these assessments. Importantly, all ICC values were statistically significant (*P*<.001), underscoring the overall consistency and reliability of the assessment across the various criteria evaluated.

Table 4. Intraclass correlation coefficient (ICC) of the clinical evaluation (n=33, ICC3k).

Item	ICC	*F* test (*df*)	*P* value	95% CI
Appropriateness	0.822	5.618 (32, 96)	<.001	0.7-0.9
Accuracy	0.755	4.080 (32, 96)	<.001	0.58-0.87
Structure/Format	0.839	6.203 (32, 96)	<.001	0.73-0.91
Conciseness	0.653	2.883 (32, 96)	<.001	0.41-0.81
Clinical validity	0.887	8.826 (32, 96)	<.001	0.81-0.94
Overall score	0.884	8.609 (32, 96)	<.001	0.8-0.94

Comparing Clinical and Quantitative Evaluations

Quantitative Error Analysis

In the quantitative evaluation of the LLM's output, we analyzed the frequency of different types of errors. Invalid generation error was the most common error type, constituting approximately 35.38% (75/212) of the total errors. This indicates a significant area where the LLM may need refinement. On the other hand, typo error was found to be the least frequent, accounting for only about 0.47% (1/212) of the errors, suggesting a relatively minor issue in the context of the overall error landscape. Other notable error types included structural malformation error (29.25%, 62/212) and nongeneration error (14.62%, 31/212).

Overall Correlation Between Quantitative and Clinical Evaluations

In addition to the detailed error-type analysis, we examined the overall correlation between the comprehensive quantitative evaluation of participant outputs and their clinical evaluation scores. We found a negative correlation of –0.633 (*P*<.001) between the quantitative and clinical evaluations. This indicates that higher error rates in LLM outputs, as identified in the quantitative analysis, correspond to lower clinical evaluation

XSL•FO
RenderX

scores, underscoring the importance of accuracy in LLM-generated text for clinical relevance.

Detailed Correlation Analysis

Our analysis of the correlation between error types in the LLM output and clinical evaluation scores (referenced in Table 5) highlighted significant differences in their impact on clinical validity. Notably, structural malformation error showed the strongest negative correlation (Pearson r=–0.654) among different error types, indicating its substantial effect on clinical assessment. This error type, involving incorrect structure or categorization of content, is crucial for the clinical applicability of LLM outputs. Other error types, such as invalid generation error and nongeneration error, exhibited varying degrees of correlation. In contrast, errors like typo error and prompt echoing error had lower correlations, suggesting a lesser impact on clinical validity.

Additionally, our examination, based on the data in Table 6, revealed that errors in certain categories showed a more pronounced correlation with clinical evaluation scores than others. Notably, categories such as differential diagnosis, diagnosis plan, and treatment plan exhibited strong negative correlations with clinical evaluation scores (Pearson r values around –0.698). This suggests that inaccuracies or errors in these critical areas greatly diminish the clinical relevance and reliability of the LLM outputs. In contrast, categories like chief complaint and vital signs exhibited lower negative correlation.

Table 5. Correlation between error types and clinical evaluation.

Error type	Count, n	Pearson r	P value
Invalid generation error	75	0.114	.53
Nongeneration error	31	0.138	.44
Information error	15	–0.182	.31
Prompt echoing error	17	–0.022	.90
Structural malformation error	62	–0.654	<.001
Contents misplacement error	11	0.110	.54
Typo error	1	0.073	.69

Table 6. Correlation between category-specific errors and clinical evaluation.

Category	Count, n	Pearson r	P value
Chief complaint	23	–0.037	.25
Vital signs	6	–0.328	.06
Present illness	19	–0.320	.07
Past history	11	–0.149	.41
Personal and social history	41	0.122	.50
Systems review	49	0.199	.27
Physical examination	19	–0.313	.08
Problem list	11	–0.540	.001
Differential diagnosis	9	–0.698	<.001
Diagnosis plan	9	–0.698	<.001
Treatment plan	9	–0.698	<.001
Additional categories	6	–0.020	.91

Discussion

Principal Findings

This study introduces and validates a comprehensive framework for assessing the efficacy of language model applications in emergency medical documentation, effectively bridging the gap between cutting-edge AI technologies and their application in critical health care scenarios. Through this discussion, we explore the implications of our findings, focusing on the insights gained from both clinical and quantitative evaluations of language model outputs. Our analysis underscores the framework's reliability and clinical pertinence, supported by rigorous analyses that include test-retest reliability, consistency across expert evaluations, and the impact of different types of errors on the utility of the generated documents. Integrating statistical data with expert opinions, we identify the strengths of language models in emergency medical documentation and pinpoint areas needing improvement, emphasizing the framework's contribution to enhancing AI applications in health care.

XSL•FO
RenderX

Interpreting the Results of Clinical Evaluation

Upon examining the statistical data from the clinical evaluation, it is notable that the highest median score in both the initial test and upon retesting was seen in structure/format. This indicates that, on average, the LLM outputs were acceptable in terms of organizing and formatting information consistent with the requirements of medical documentation. On the other hand, accuracy presented the lowest median score, pinpointing a significant area for improvement. The lower scores in accuracy imply that the LLM outputs occasionally included errors or inaccuracies in the medical information provided, underscoring the need for further refinement of LLMs to ensure the reliability of generated content.

From the perspective of test-retest analysis, the Pearson correlation coefficient of the overall score confirms that a consistent and reliable evaluation was conducted by the experts, affirming the utility of our framework. Moreover, all 5 criteria—appropriateness, accuracy, structure/format, conciseness, and clinical validity—exhibited correlation coefficients exceeding the *P* value of .60. This demonstrates that our framework is consistent and reliable even across all clinical evaluation criteria. Notably, clinical validity reached the highest correlation coefficient. Given that clinical validation is the most critical element in evaluating medical records, this result signifies that our evaluation is trustworthy and valuable in a clinical context.

Furthermore, the uniformly high ICC results, which measure the agreement level among different evaluators on the same measure, indicate a high level of agreement. Clinical validity, in particular, demonstrated the greatest consistency across the panel. This clarity illustrates that our framework can facilitate consistent clinical evaluations from the perspective of various experts. Appropriateness and structure/format also scored highly, indicating minimal variation in their assessment among the evaluators. These observations, supported by significant *F* statistics and narrow CI, emphasize the assessments' reliability and precision.

Comparing Clinical and Quantitative Evaluations

In the process of comparing clinical and quantitative evaluations, a significant relationship was revealed between the number of errors in LLM outputs and their clinical evaluations. This relationship manifested as a negative impact on the overall correlation between quantitative evaluations and clinical scores. This finding suggests that errors identified through quantitative analysis are aligned with clinical assessments, reinforcing the validity of the LLM outputs.

Among the various types of quantitative errors, structural malformation error had the most significant influence on clinical evaluation. This underscores the critical importance of accurate structure and format within medical records, where the presence of unnecessary items or the omission of necessary ones can significantly undermine the document's utility. Such errors disrupt the logical flow and completeness of medical records, making them less useful in clinical settings. Despite invalid generation error being the most frequently observed type, its impact on clinical evaluation was less pronounced. This could be due to the evaluators' tendency to overlook certain errors that commonly occur in clinical settings, such as incorrectly labeling unexamined items as "None" rather than "Unknown."

Further analysis of category-specific errors revealed that specific categories within the initial ED records, such as differential diagnosis, diagnosis plan, and treatment plan, were most critical to clinical evaluators. These categories, pivotal for clinical decision-making, demonstrated a strong correlation with clinical evaluations, underscoring the significant impact of inaccuracies in these critical reasoning areas on clinical relevance. This reflects the inherent value clinicians place on deep clinical reasoning within medical records.

Conversely, categories like system inquiry and personal and social history, despite being areas where errors were frequently found, showed lesser impact on clinical evaluation. This may be attributed to the prevalence of similar errors in human-generated records, leading to a more forgiving attitude from evaluators. Notably, instances of mislabeling unexamined items as "None"—a type of invalid generation error previously discussed—within these categories were often overlooked.

This nuanced analysis highlights the differential weighting of error types by expert evaluators, demonstrating that while quantitative error analysis is essential for identifying and categorizing errors, the subjective judgment of medical professionals remains indispensable. The evaluators' prioritization underscores the critical role of clinical reasoning in assessing the relevance and acceptability of LLM outputs, illustrating the complex interplay between objective error identification and subjective clinical judgment in determining the utility of AI-generated medical documentation.

Integrating Automated Quantitative Assessments With Clinical Expertise

One area for future development involves integrating automated quantitative evaluations with the nuanced understanding of clinicians. Our study highlights the importance of balancing objective error analysis with subjective clinical judgment. Future research could explore creating systems that automate the quantitative evaluation process while incorporating clinician input to refine these assessments. This integration would aim to harness the strengths of both AI and human expertise, ensuring more accurate and clinically relevant LLM outputs in health care settings.

Future development should focus on integrating automated quantitative evaluations with clinicians' nuanced understanding. Automated systems, such as machine learning algorithms, could classify error types and correlate these with clinical expert assessments. This would balance the objective error analysis with subjective clinical judgment, reducing cognitive load on clinicians while ensuring the accuracy and clinical relevance of LLM outputs. Further research could explore refining these systems to improve the applicability of LLM-generated documents in health care settings, thereby enhancing their effective integration into real-world medical documentation.

XSL•FO
RenderX

Limitations

Our study, while providing significant insights into the application of LLMs in health care, faced certain limitations. The lack of detailed personal data, such as age and gender, from participant profiles limits our understanding of the sample's diversity and the generalizability of our findings. Additionally, the exclusive use of HyperCLOVA X for generating and evaluating prompts restricts our insights to a single LLM. This presents a limitation in understanding how different LLMs might perform in similar tasks, which is an important aspect for future explorations.

Furthermore, we did not investigate the reproducibility of LLM outputs. The absence of tests to determine whether repeated applications of the same prompts would yield consistent results under varied conditions leaves a gap in understanding the reliability of LLM-generated content in clinical settings. Additionally, our study's focus on initial records from the ED narrows its scope. Future studies should broaden the research to include a wider range of medical records and clinical departments.

Finally, while the PDQI-9 could have provided a more comprehensive evaluation, we simplified our criteria due to time constraints and the cognitive load of the competitive setting. This simplification was necessary for practical use by assessors, but it highlights an area for future enhancement in evaluating LLM applications in health care documentation.

Conclusion

Our study introduces a novel and robust framework for evaluating LLMs in health care, particularly focusing on initial ED records. During the Healthcare Prompt-a-thon, we used LLMs and combined clinical insights with quantitative analysis. This approach revealed a strong relationship between the prevalence of quantitative errors and the overall quality of clinical outcomes, underscoring the effectiveness and reliability of our framework in assessing clinical utility.

Moreover, the results validate the clinical acceptability of our framework, suggesting that LLMs, when accurately assessed and used, can be integrated into clinical practice. This advancement opens avenues for LLMs in health care, aligning AI innovations with the critical demands of patient safety and care quality. Future research should focus on refining this framework, further enhancing the practicality and adoption of LLMs in health care environments.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant HI22C0452). We are deeply grateful to the Korea Society of Artificial Intelligence in Medicine and its generative model research group for organizing the Healthcare Prompt-a-thon, and to Samsung Medical Center for hosting this meaningful event. We convey our special thanks to NAVER Cloud for their sponsorship and support. Our appreciation also extends to all participants of the Healthcare Prompt-a-thon for their invaluable input.

Authors' Contributions

JS, DC, and TK were responsible for the initial development of the framework and criteria. DC drafted the manuscript, while JS conducted the statistical analyses. TK performed a comprehensive revision of the manuscript. The dataset was developed by JS, TK, DC, HY, and NO. Clinical evaluation and medical insights were provided by YY, GL, NO, TK, and MK. Technical and material support was provided by WC, EC, HY, and MK. All authors contributed to the overall review and approval of the final manuscript.

Conflicts of Interest

In conducting this study, we used HyperCLOVA X and received credits for its use. These credits were exclusively allocated for participant compensation and the facilitation of the competition. Apart from this, we did not receive any other forms of compensation or remuneration from entities associated with HyperCLOVA X, NAVER Cloud, or any other external sources.

Multimedia Appendix 1
Two patient cases provided to the participants.
[DOCX File , 31 KB - jmir_v26i1e58329_app1.docx]

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901 [FREE Full text]
2. ChatGPT: optimizing language models for dialogue. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2024-01-11]
3. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]

XSL•FO
RenderX

4. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
5. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024;5(3). [doi: 10.1056/cat.23.0404]
6. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
7. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2023;1(1). [doi: 10.1056/aip2300031]
8. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
9. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
10. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023;27(1):120 [FREE Full text] [doi: 10.1186/s13054-023-04393-x] [Medline: 36945051]
11. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digital Health 2023;5(6):e333-e335. [doi: 10.1016/s2589-7500(23)00083-3]
12. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
13. HyperClova X. Naver Cloud Platform. URL: https://www.ncloud.com/solution/featured/hyperclovax [accessed 2024-01-11]
14. Kim B, Kim H, Lee SW, Lee G, Kwak D, Hyeon JD, et al. What changes can large-scale language models bring? intensive study on HyperCLOVA: billions-scale Korean generative pretrained transformers. 2021 Presented at: 2021 Conference on Empirical Methods in Natural Language Processing; November 7-11, 2021; Online and Punta Cana, Dominican Republic p. 3405-3424. [doi: 10.18653/v1/2021.emnlp-main.274]
15. Clova Studio. Naver Cloud Platform. URL: https://www.ncloud.com/product/aiService/clovaStudio [accessed 2024-01-11]
16. Kang I. Clova: services and devices powered by AI. 2018 Presented at: SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; July 8-12, 2018; Ann Arbor, MI p. 1359. [doi: 10.1145/3209978.3210208]
17. Lee SH, Park J, Yang K, Min J, Choi J. Accuracy of cloud-based speech recognition open application programming interface for medical terms of Korean. J Korean Med Sci 2022;37(18):e144 [FREE Full text] [doi: 10.3346/jkms.2022.37.e144] [Medline: 35535371]
18. Lee SW, Jung H, Ko S, Kim S, Kim H, Doh K. CareCall: a call-based active monitoring dialog agent for managing COVID-19 pandemic. arXiv Preprint posted online on July 6, 2020 [FREE Full text]
19. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174 [FREE Full text] [doi: 10.4338/aci-2011-11-ra-0070] [Medline: 22577483]
20. Edwards ST, Neri PM, Volk LA, Schiff GD, Bates DW. Association of note quality and quality of care: a cross-sectional study. BMJ Qual Saf 2014;23(5):406-413. [doi: 10.1136/bmjqs-2013-002194] [Medline: 24287259]
21. Guttman L. A basis for analyzing test-retest reliability. Psychometrika 1945;10:255-282. [doi: 10.1007/BF02288892] [Medline: 21007983]
22. Ionan AC, Polley MC, McShane LM, Dobbin KK. Comparison of confidence interval methods for an intra-class correlation coefficient (ICC). BMC Med Res Methodol 2014;14:121 [FREE Full text] [doi: 10.1186/1471-2288-14-121] [Medline: 25417040]

Abbreviations

AI: artificial intelligence
ED: emergency department
ICC: intraclass correlation coefficient
LLM: large language model
PDQI-9: Physician Documentation Quality Instrument
USMLE: United States Medical Licensing Examination

XSL•FO
RenderX

Edited by A Schwartz; submitted 13.03.24; peer-reviewed by Y Lee, TF Heston; comments to author 12.06.24; revised version received 30.06.24; accepted 24.09.24; published 20.11.24.

Please cite as:
Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, Oh N, Yi Y, Lee KH, Choi E
Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study
J Med Internet Res 2024;26:e58329
URL: https://www.jmir.org/2024/1/e58329
doi:10.2196/58329
PMID:

©Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, Edward Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI in Dental Radiology—Improving the Efficiency of Reporting With ChatGPT: Comparative Study

Daniel Stephan[1*], Dr med; Annika Bertsch[1*]; Matthias Burwinkel[1], Dr med dent; Shankeeth Vinayahalingam[2], MD; Bilal Al-Nawas[1], Dr med dent, Prof Dr Med; Peer W Kämmerer[1*], MA, MSc, Dr med dent, Prof Dr Med; Daniel GE Thiem[1*], MHBA, Dr med dent, Dr med, PD

[1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Mainz, Germany

[2]Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, Nijmegen, Netherlands

[*]these authors contributed equally

Corresponding Author:
Daniel Stephan, Dr med
Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery
University Medical Centre of the Johannes Gutenberg-University Mainz
Augustusplatz 2
Mainz, 55131
Germany
Phone: 49 6131177038
Email: stephand@uni-mainz.de

Abstract

Background: Structured and standardized documentation is critical for accurately recording diagnostic findings, treatment plans, and patient progress in health care. Manual documentation can be labor-intensive and error-prone, especially under time constraints, prompting interest in the potential of artificial intelligence (AI) to automate and optimize these processes, particularly in medical documentation.

Objective: This study aimed to assess the effectiveness of ChatGPT (OpenAI) in generating radiology reports from dental panoramic radiographs, comparing the performance of AI-generated reports with those manually created by dental students.

Methods: A total of 100 dental students were tasked with analyzing panoramic radiographs and generating radiology reports manually or assisted by ChatGPT using a standardized prompt derived from a diagnostic checklist.

Results: Reports generated by ChatGPT showed a high degree of textual similarity to reference reports; however, they often lacked critical diagnostic information typically included in reports authored by students. Despite this, the AI-generated reports were consistent in being error-free and matched the readability of student-generated reports.

Conclusions: The findings from this study suggest that ChatGPT has considerable potential for generating radiology reports, although it currently faces challenges in accuracy and reliability. This underscores the need for further refinement in the AI's prompt design and the development of robust validation mechanisms to enhance its use in clinical settings.

(J Med Internet Res 2024;26:e60684) doi:10.2196/60684

KEYWORDS

artificial intelligence; ChatGPT; radiology report; dental radiology; dental orthopantomogram; panoramic radiograph; dental; radiology; chatbot; medical documentation; medical application; imaging; disease detection; clinical decision support; natural language processing; medical licensing; dentistry; patient care

Introduction

Structured and standardized documentation plays a crucial role in health care by ensuring accurate recording and communication of diagnostic findings, treatment plans, and patient progress, thereby supporting high-quality patient care [1]. However, manual documentation is often time-consuming, error-prone, and can impede clinical workflow efficiency, especially in fast-paced medical settings. With the emergence of artificial intelligence (AI), there is a growing interest in implementing AI technology to optimize health care workflows and improve documentation practices.

XSL•FO
RenderX

AI has proven useful in various medical applications, from diagnosing diseases to drug development [2]. In radiology, AI algorithms analyze medical images to assist in early disease detection, improve radiologists' performance, and provide clinical decision support [3,4]. Moreover, AI-driven solutions have the potential to automate repetitive tasks and reduce the workload of health care professionals [5,6].

First introduced by OpenAI in 2018, (GPT—a specific large language model developed by OpenAI) has continuously evolved and trained on extensive text data [7]. ChatGPT (implementation of GPT), an advanced large language model (a class of AI models), represents a significant advancement in natural language processing and has demonstrated remarkable capabilities in understanding and generating human-like text using deep learning techniques, like neuronal networks. GPT 3.5 showed a human-level performance across various medical exams and passed the United States Medical Licensing Exam (60.2%), Med-MCQA (57.5%), and PubMedQA (78.2%) [8-10]. With its proficiency in language generation, ChatGPT is capable of medical writing [11] and, therefore, has been increasingly integrated into medical education [12] and clinical practice, allowing it to automate the writing of examination findings, doctor's letters, or radiology reports [13].

Dental radiology, integral to dentistry, relies on the correct interpretation of x-ray images, including panoramic radiographs (OPG), to diagnose and plan numerous oral conditions or pathologies. To maintain the standard of patient care, it is, therefore, crucial to ensure high-quality training in radiology tasks during dental studies. Traditionally, radiology education involves manual interpretation of x-ray images and writing detailed medical findings reports based on visual inspection and clinical knowledge. However, the emergence of AI technologies has increased interest in alternative methods for radiology education and diagnostic reporting, including maxillofacial radiology [14,15].

The capability of AI in diagnosing medical images, including x-ray images, is well-established [3,16]. Moreover, studies have demonstrated that ChatGPT can generate clinic letters and operative notes with high correctness and readability [17,18]. Additionally, another study has proven its efficacy beyond text generation in simplifying existing radiology reports and improving patient understanding [19]. Furthermore, recent research reveals AI's capability to outperform dental students in diagnostic accuracy regarding endodontic assessments [20], highlighting its potential as a reference tool to enhance students' understanding and diagnostic skills. However, this raises concerns about the potential for overreliance on AI, considering reports about ChatGPT generating fake findings for imaginable diseases [21], which may affect the development of critical analytical and decision-making abilities. Thus, it is essential to integrate AI with human expertise and clinical judgment in dental education. ChatGPT shows promising potential in improving doctor-patient communication by simplifying complex medical information and transforming complex medical terminology into easily understandable language for patients with varying levels of health expertise [22]. While earlier versions of ChatGPT powered by GPT-3.5 generated patient-facing information lacking accuracy and important information, GPT-4 has shown improvements in appropriateness and accuracy and, despite occasional omissions, ultimately produced patient information applicable for gaining informed consent for procedures in nuclear medicine [23].

Nevertheless, the generation of radiology reports based on diagnostic findings by health care professionals remains a subject of investigation. Therefore, this study evaluated the efficacy of incorporating AI language models, specifically ChatGPT, into generating radiology reports. Dental students analyzed OPGs and provided diagnoses through checkbox lists together with written reports. A comparative analysis between radiology reports manually written by dental students and reports generated by the AI based on those prefilled checkbox lists was conducted. This study primarily investigated the readability of both report types with the null hypothesis stating no differences in readability between the 2 sets of reports. Secondary outcomes, including text accuracy and language quality, were evaluated to identify potential areas for improvement in AI-driven radiology reporting.

Methods

Overview

This study sought to investigate the efficacy of incorporating AI language models, specifically ChatGPT, in generating radiology reports from prefilled checkbox lists after analyzing OPGs. Dental students were assigned to diagnose 2 different x-ray images, providing a written radiology report for 1 and a checkbox list of diagnoses for the other. The AI then generated reports based on the diagnoses provided within the checkbox lists. Subsequently, both texts were analyzed comparatively to primarily evaluate readability, with a secondary evaluation of text quality, accuracy, similarities, and disparities between student-written and AI-written reports.

Ethical Considerations

The study adhered strictly to ethical standards and institutional guidelines, obtaining informed consent from all participants beforehand. Participants were clearly informed of the purpose of the research, the voluntary nature of participation, and their ability to withdraw at any time without any repercussions. Additionally, no compensation was provided, as the tasks were integral to students' academic training. Confidentiality and data privacy were stringently maintained throughout the research process to uphold the participant's well-being and privacy. An ethics approval was not required as the generation of radiology reports is a standard component of dental education in Germany and this type of research does not involve intervention or data collection beyond routine educational activities. The tasks performed by the students were part of their regular academic curriculum and no additional tasks outside the students' regular academic curriculum were introduced. Moreover, the analysis of data was conducted anonymously, ensuring privacy and confidentiality and preventing participating in this study resulted in either advantages or disadvantages for the students. As no images or materials involving identifiable features were included in this study, no additional consent forms were required.

XSL•FO
RenderX

Study Setting

The study took place in the radiology section of the Department of Oral and Maxillofacial Surgery at the University Medical Centre Mainz, Germany. Certified medical monitors were provided, and all participants were supervised throughout the session without access to additional information or external help.

Participants

In Germany, dental education is structured into 10 semesters. The first 5 semesters focus on foundational knowledge, while the following 5 semesters (clinical semesters 1-5, corresponding to overall semesters 6-10) emphasize clinical skills. The first lesson in dental radiology was introduced in the first clinical semester and, therefore, preclinical students were excluded from this study. A total of 100 dental students from all 5 clinical semesters participated in the study with the following distribution across semesters—semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20. This equal representation across different stages of dental education highlights the progressive development in radiology report writing.

Experimental Design

Students were randomly assigned to 1 of 2 groups and presented with an unknown OPG (Figures 1A and 1B). Group A was instructed to analyze the x-ray image (Figure 1A) and compose a radiology report within 30 minutes without any external assistance. Group B received a second OPG (Figure 1B) and was tasked with completing a checkbox list (Figure 2) detailing their observations within a 10-minute time frame. These time limits were specifically chosen to investigate the potential time-saving benefits of using a structured checkbox method followed by AI-generated reporting. It was observed that all students used the entire allotted time for their respective tasks, neither exceeding the time limit nor completing early. The study, therefore, focused on the completion of the tasks within the predefined limits without measuring the exact duration for each task. Upon completion, each group was required to complete the alternate assignment with the opposite x-ray image. To minimize biases, the experimental design ensured that 1 group completed the checkbox for the same x-ray for which the other group composed the report, and vice versa. This approach reduced any influence of specific characteristics of the x-ray images (eg, the complexity of findings or difficulty of interpretation).

XSL•FO
RenderX

Figure 1. (A,B) Two randomly chosen panoramic radiographs featuring various pathologies to be diagnosed by dental students. Both x-ray images represented the basis of a student-written and an AI-generated radiology report. AI: artificial intelligence.

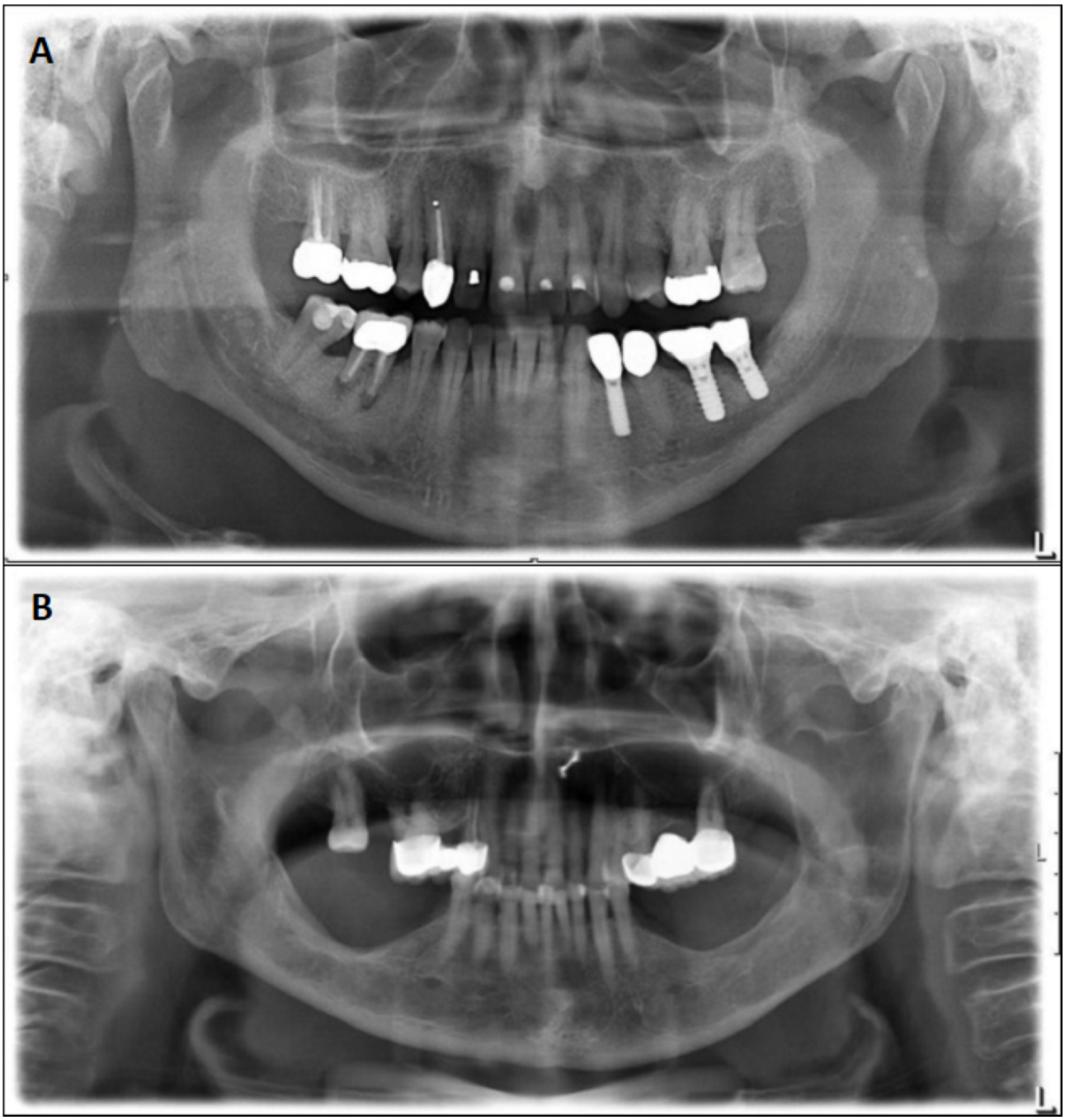

XSL•FO
RenderX

Figure 2. An example of a completed checkbox list containing 3 distinct spreadsheets (1, 2, and 3) used to generate radiology reports with ChatGPT.

1

Positionierung	Belichtung	Ramus	Kieferhöhlen	Kondylen des Kiefergelenkes	Zahnstatus im Überblick	Artefakte	Metalldichte Opazitäten
☐Kippung nach dorsal ☐Kippung nach ventral ☐Verschiebung nach dorsal ☐Verschiebung nach ventral ☐Neigung des Kopfes nach lateral ☒regelrechte Positionierung	☐ überbelichtet ☐unterbelichtet ☒ regelhaft	Rechts: ☒unauffällig ☐verkürzt ☐verbreitert Links: ☒unauffällig ☐verkürzt ☐verbreitert	☒unauffällig unilaterale Verschattung: ☐rechts ☐links ☐ bilaterale Verschattung der KH V.a. Retentionszyste ☐rechts ☐links	☒ unauffällig ☐ Kondylus unilateral abgeflacht: ☐rechts ☐links ☐ Kondylus bilateral abgeflacht	☒konservierend versogt ☐prothetisch versorgt ☐sanierungsbedürftig ☐kein Interventionsbedarf ☐generelle Konkremente ☐Hyperdontie	☐Ohrringe ☒Piercing (Nase, Lippe, etc.) ☐Halsketten ☐Röntgenschürze	☐Osteosyntheseplatten ☐Bone anchor ☐kieferorthopädische Kettchen ☐kieferorthopädische Apparatur ☐IMF-Schrauben

2

Weitere Befunde	Wurzelkanalbehandlungen	Füllungen	Zahnkronen	Fehlende Zähne	Weisheitszähne
Brücken im ☒ersten Quadranten ☒zweiten Quadranten ☐dritten Quadranten ☐vierten Quadranten Zähne, die Brückenanker darstellen: ☒14 ☐34 ☐15 ☐35 ☒16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☒24 ☐44 ☐25 ☐45 ☒26 ☐46 ☐27 ☐47 ☐28 ☐48 ☐insuffizient ☐unauffällig ☐keine Brücken vorhanden	Wurzelkanalbehandlung am Zahn ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☒14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48 ☐lege artis ☐insuffizient ☐nicht beurteilbar ☐keine Wurzelkanalbehandlungen	Gefüllte Zähne: ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48	am Zahn ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☒18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48 ☐lege artis ☐insuffizient ☐keine Zahnkronen	☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☒36 ☒17 ☒37 ☐18 ☒38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☒45 ☐26 ☒46 ☒27 ☒47 ☒28 ☒48 ☐zahnloser Oberkiefer ☐zahnloser Unterkiefer	18 ☐Durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert 28 ☐durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert 38 ☐durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert 48 ☐durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert

3

Aufhellung in Regio	Implantate in Regio	Implantate zeigen	Bewertung des Knochens	Zystische Veränderungen	Kontinuitätsunterbruch der Compacta	Procedere Empfehlungen
☐11 ☒31 ☒12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☒22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48	☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48 ☒kein Implantat vorhanden	☐keine Auffälligkeiten Vertigkaler Knochenverlust nach: ☐mesial ☐distal	Genereller horizontaler Knochenabbau ☒im Unterkiefer ☒im Oberkiefer Vertikale Knocheneinbrüche in Regio ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48	Lokalisation: ☐rechter Ramus ☐linker Ramus ☐rechter Corpus mandibulae ☐linker Corpus mandibulae ☐rechte Maxilla ☐linke Maxilla Erstreckt sich bis: ☐ scharf begrenzt ☐unscharf begrenzt ☐randständige Sklerosierung ☐mehrere Zystenkammern ☐Ohne Kontinuitätsunterbrechung der Compacta ☐Kontinuitätsunterbrechung der Compacta ☐Eindeutige Relation zu einem Zahn: ☐Verdrängendes Wachstum ☐Resorption benachbarter Zahnwurzeln	Ramus: ☐Links ☐Rechts Collum ☐Links ☐rechts Capitulum ☐Links ☐rechts ☐Corpus ☐Maxilla anterior ☐Maxilla posterior ☐Os zygomaticum ☐Sequesterbildung	☒Kontroll-OPG in 6 Monaten ☐3D-Bilgedbung (DVT, MRT, CT) ☐keine spezifische Kontrolle erforderlich ☐Extraktion der beherdeten Zähne ☐Explantation ☐Zystostomie/Zystektomie ☒weitere klinische Untersuchungen

Data Transformation and AI Text Generation

Upon completion, the checkbox lists filled out by participants were carefully transcribed into an Excel (Microsoft) data sheet comprising distinct spreadsheets for each category to organize the data. Subsequently, Chat GPT 4.0, an advanced AI language model, was harnessed to generate radiology reports based on the checkbox lists. Each spreadsheet within the Excel file was sequentially analyzed, and the information marked with an “X” in the “checkbox” column was incorporated into the generated reports. The following specific prompt was used to guide the AI in formulating structured x-ray reports, ensuring consistency and completeness:

> *Formulate a structured X-ray report in the sense of an X-ray report of an OPG based on the following checkbox list of the entire Excel table, and do not omit any columns. Please mention only those statements for which a box is marked with an X in the X-ray report. The statements not marked with an X should not be included in the report. The figures given should be interpreted in the sense of an odontogram. Analyze each spreadsheet in the Excel file in the order*

XSL•FO
RenderX

given. The column with the markings (X) is marked with the term "checkbox." The report should be written in continuous text from the perspective of the treating dentist. Formulate a continuous text without subheadings.

The model settings included a temperature of 0.7 (controlling the randomness of responses), a maximum token limit of 1500 (restricting the length of the response), a frequency penalty of 0.0 (preventing repetitive word usage), and a presence penalty of 0.6 (promoting the inclusion of new topics). Those settings were shown to generate the highest output quality with the temperature setting of 0.7 being particularly important. Although lower settings are suggested to be advantageous for more deterministic tasks, preliminary tests revealed them to produce repetitive and difficult-to-read reports lacking naturalness and effectiveness in communication. In contrast, the chosen setting balanced creativity and coherence resulting in improved readability. Each report was generated using the ChatGPT web interface in a new session from September 5 to October 12, 2023, ensuring consistency and comparability across all outputs. To minimize biases associated with varying performance due to server load, which tends to be higher on weekends with higher traffic, the tasks were randomly distributed across different weekdays. This approach aimed to ensure a consistent and balanced evaluation of ChatGPT's capabilities by reducing potential variability in output quality. The checkbox lists were directly uploaded without additional preprocessing.

Readability Indices

The readability and complexity of both student-written and AI-generated texts were assessed using the Flesch reading ease (FRE) [24] score and the Lesbarhetsindex (LIX) readability index. "Readability" refers to how easily written material can be understood, determined by the complexity of the vocabulary, sentence, and word lengths used [23]. Although prior knowledge or motivation of the reader is not considered in readability formulas, especially in health care, a higher readability is associated with improved comprehension and participation of the patient.

The FRE score evaluates text readability based on its linguistic characteristics. In particular, the average sentence length (ASL) and the average number of syllables per word (ASW) are considered for the calculation using the following formula (adapted to the German language [25]):

The FRE score typically ranges between 0 and 100, with higher scores indicating greater readability and lower scores suggesting increased complexity. Due to its high reproducibility [25], validation for various text types, and correlation with other readability formulas, the FRE score is an established metric in the analysis of medical texts [26-29].

LIX index considers the ASL and the prevalence of long words with more than 6 letters to assess text readability by the following calculation:

A higher LIX score indicates greater complexity, whereas a lower score suggests easier comprehension. LIX has been validated as a reliable measure of readability across multiple languages, including Swedish, Danish, English, French, German, Finnish, Italian, Spanish, and Portuguese [30,31].

To assess readability, the FRE and the LIX scores were calculated for both the student-written and AI-generated reports. Differences in readability were analyzed by comparing FRE and LIX scores of AI-generated reports with student-written reports. Additionally, this analysis was conducted collectively for all texts, as well as individually for each academic semester, to evaluate the influence of the educational level on text comprehensibility in comparison to automated text generation.

Text Similarity (Bidirectional Encoder Representations from Transformers Score)

The accuracy of AI-generated texts was evaluated by comparing the number of findings diagnosed by students to those mentioned in the final AI-generated reports. Additionally, reference texts were manually created by a senior physician with extensive clinical experience in dental radiology, for each checkbox list to assess the quality of AI-generated texts. A comprehensive template was developed and carefully reviewed by all authors, serving as a standardized framework for report creation. Each reference text was individually crafted by transferring the findings from the corresponding checkbox list into the template. This standardized approach was consistently applied to each report, ensuring uniformity in content and structure while minimizing discrepancies that could bias the Bidirectional Encoder Representations from Transformers (BERT) score. These reference texts were then compared to the AI-generated text using the BERT score, a widely recognized metric for evaluating text similarity. Based on the BERT model [32], which generates high-dimensional vector representations (embeddings), capturing the BERT score measures the similarity between corresponding tokens in both texts. The BERT score includes 3 primary components—precision (P), recall (R), and F_1-score. Precision measures the proportion of words in the AI-generated text that contribute accurately to the overall meaning as compared to the reference text. Essentially, it assesses the quality of the AI's output in terms of the relevance and accuracy of the information presented. Recall evaluates the extent to which the AI-generated text covers all the relevant information contained in the reference text, highlighting how well the AI captures necessary details without omitting critical information. Finally, the F_1-score provides a harmonic mean of precision and recall, offering a single score that balances both the completeness and accuracy of the AI-generated text. The aggregated similarity scores, normalized to a range between 0 and 1, indicate overall text similarity. A higher BERT score indicates textual similarity, reflecting a higher quality of AI-generated texts. Multiple studies have already demonstrated BERT's capability of accurately predicting readability levels for various texts [33].

Text Accuracy

The accuracy of AI-generated radiology reports was assessed by comparing the number of included findings with the number

XSL•FO
RenderX

of findings contained in the referring checkbox list, both overall and for each of the 3 individual spreadsheets separately.

Text Analysis

Descriptive text analysis was conducted by measurement of word count, sentence length, syllable count, diphthong count, and character count to compare AI-generated with student-written radiology reports. ASL and long word proportion (defined as words with more than 6 characters) were further assessed. Language quality across all texts was quantified by evaluating the error count including spelling, grammar, and punctuation together with the calculation of the error ratio (number of errors divided by words multiplied by 100). These metrics were analyzed collectively for all students and semesters as 1 group.

Statistical Analysis

The software packages used for statistical analysis were GraphPad Prism 9.0 (Graphpad Software, LLC), G*Power 3.1 (Heinrich-Heine-University Düsseldorf), Excel 16.76, and SPSS Statistics (version 29; IBM Corp). To assess the potential difference in readability between AI-generated reports and those written by students, an a priori power analysis was performed. This analysis was based on previously observed significant differences in the FRE scores, which showed a lower average for ChatGPT responses (mean 34.9, SD 11.2) compared to medical information on Google webpages (mean 46.5, SD 14.3), accounting for a difference of 11.6 [34]. Additionally, similar results with the LIX score have demonstrated a difference of 10 between human-written and ChatGPT-written scientific introductions [35]. To achieve a power of 80% and maintain a significance level of 5%, a minimum of 25 samples per group (study arm) is required. Significance was set at $P<.05$. All data are presented as mean (SD). Differences between student-written and AI-generated texts were analyzed using a 2-tailed student *t* test. A subsequent post hoc power analysis was conducted for each test to verify the power achieved by the *t* test.

Results

Overview

Text quality, readability, and comprehensibility of student-written and AI-generated radiology reports were compared by analysis of various language parameters. Throughout the study, students consistently used the preset time to its full extent, dedicating 30 minutes for completing the written report and 10 minutes for the competition of the checkbox list. While AI-generated radiology reports demonstrated a remarkable similarity to reference texts with no difference in readability, a significant information deficiency was observed.

AI-Generated and Student-Written Texts Possessed Identical Readability

The FRE score (Figure 3A,B) revealed no difference in readability between AI-generated and student-written texts (mean 50.55, SD 7.80 vs mean 51.19, SD 5.02) considering all reports together as demonstrated in Figure 3A ($P=.49$; $t_{198}=0.6898$). Upon examination of each semester individually, the Flesch index exhibited significant variability, with AI-generated texts demonstrating lower readability compared to texts written by the first clinical semester (mean 56.65, SD 6.70 vs mean 49.6, SD 7.17; $P=.002$; $t_{38}=3.213$) and higher readability compared to the third (mean 47.14, SD 6.97 vs mean 51.14, SD 4.55; $P=.03$; $t_{40}=2.203$) and fourth (mean 46.35, SD 7.29 vs mean 52.15, SD 5.08; $P=.006$; $t_{38}=2.918$) clinical semesters. No difference was found for second ($P=.39$; $t_{36}=0.8647$) and fifth ($P=.12$; $t_{38}=1.577$) clinical semester.

XSL•FO
RenderX

Figure 3. Metric evaluation of readability of AI-generated reports compared to student-written radiology reports (A) overall and (B) individually for each semester assessed with the Flesch readability ease score. Data represent mean (SD). Sample size: (A) n=100; (B) semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20; *P<.05; and versus students. AI: artificial intelligence.

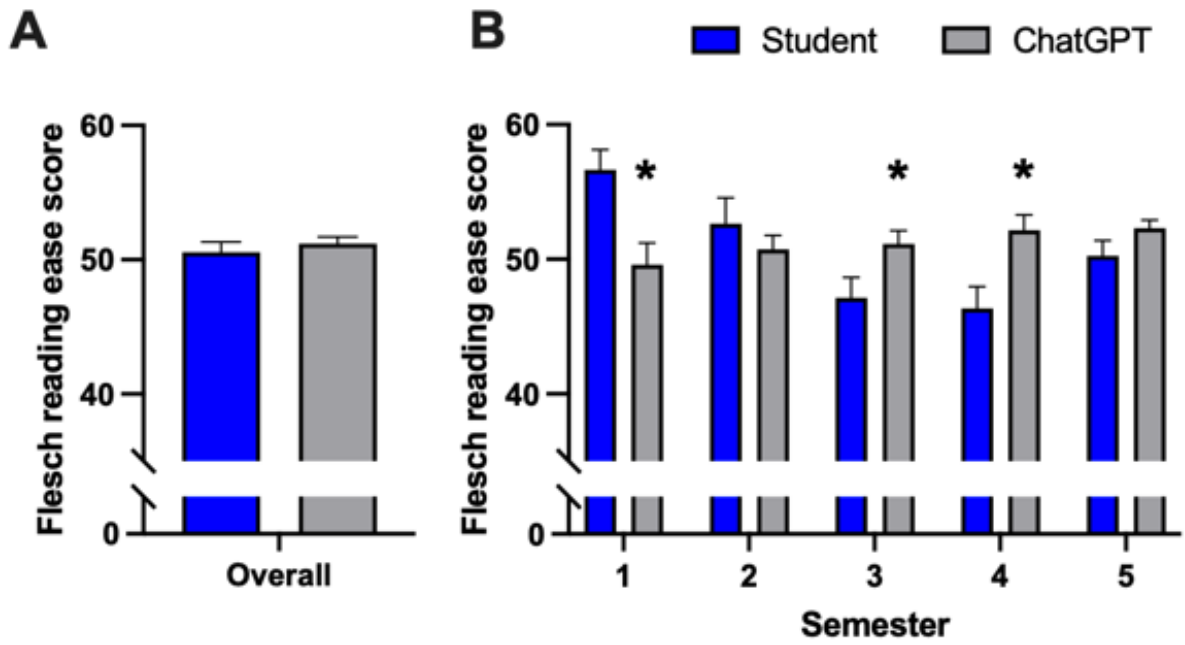

As presented in Figure 4A, no overall difference between both groups regarding readability was found (mean 48.98, SD 5.0 vs mean 48.0, SD 2.85) as assessed with the LIX index (P=.09; t_{198}=1.699). In contrast to the FRE score, the LIX readability index exhibited opposing trends across semesters (Figure 4B), with significant differences observed in semesters 1 (mean 46.27, SD 4.0 vs mean 48.81, SD 3.44; P=.04; t_{38}=2.157); 3 (mean 51.64, SD 4.89 vs mean 48.01, SD 2.84; P=.005; t_{40}=2.944); and 4 (mean 50.67, SD 4.68 vs mean 47.32, SD 2.90; P=.098; t_{38}=2.719). No difference was observed in the second (P=.39; t_{36}=0.877) and fifth (P=.15; t_{38}=1.464) clinical semester.

Figure 4. Metric evaluation of readability of AI-generated reports compared to student-written radiology reports (A) overall and (B) individually for each semester assessed with LIX index. Data represent mean (SD). (A) Sample size: n=100; (B) semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20; *P<.05; and versus students. AI: artificial intelligence; LIX: Lesbarhetsindex.

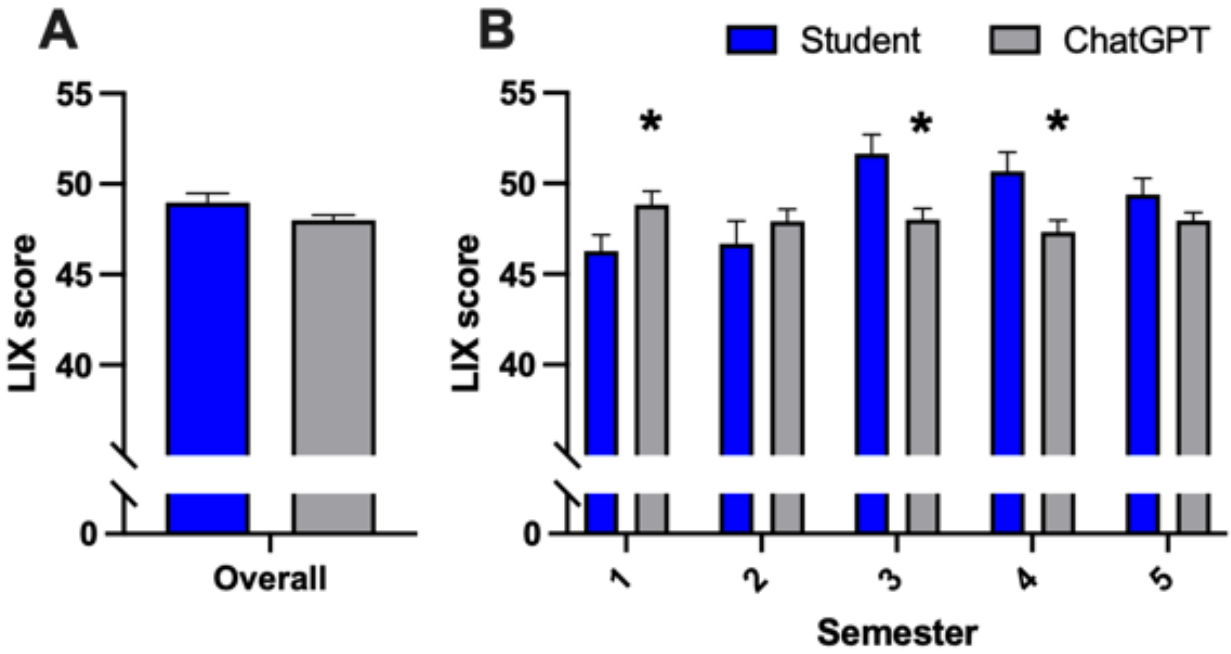

XSL•FO
RenderX

AI-Generated Reports Show Great Similarity to Reference Texts but Lack Information

As illustrated in Figure 5A, the great similarity is indicated by a high BERT score, with precision (P)=mean 0.967, SD 0.036, recall (R)=mean 0.958, SD 0.037, and F_1=mean 0.962, SD 0.036. The analysis further revealed a notable deficiency in relevant information within AI-generated texts. A significant difference was evident between the findings diagnosed by students and those mentioned in the AI-generated reports (Figure 5B), with students identifying a mean of 44.6 (SD 6.0) findings, whereas the AI reported a mean of 41.3 (SD 7.0) findings in total (P=.04; t_{198}=3.586). Specifically, as shown in Figure 5C, while no difference was observed in the first spreadsheet (mean 8.53, SD 1.06 vs mean 8.55, SD 1.02; P=.89; t_{198}=0.1361), the AI included significantly fewer findings from the second (mean 23.03, SD 3.67 vs mean 21.22, SD 4.92; P=.003; t_{198}=2.951) and third (mean 13.07, SD 4.40 vs mean 11.56, SD 4.23; P=.014; t_{198}=2.476) spreadsheets.

Figure 5. Evaluation of similarity compared to reference texts using the (A) BERT score with precision (P), recall (R), and F1 score (F1) representing the harmonic mean of precision and recall. The accuracy of AI-generated radiology reports was further assessed as (B) overall accuracy including the whole checkbox list and (C) individually for each spreadsheet of the checkbox list. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence; BERT: Bidirectional Encoder Representations from Transformers.

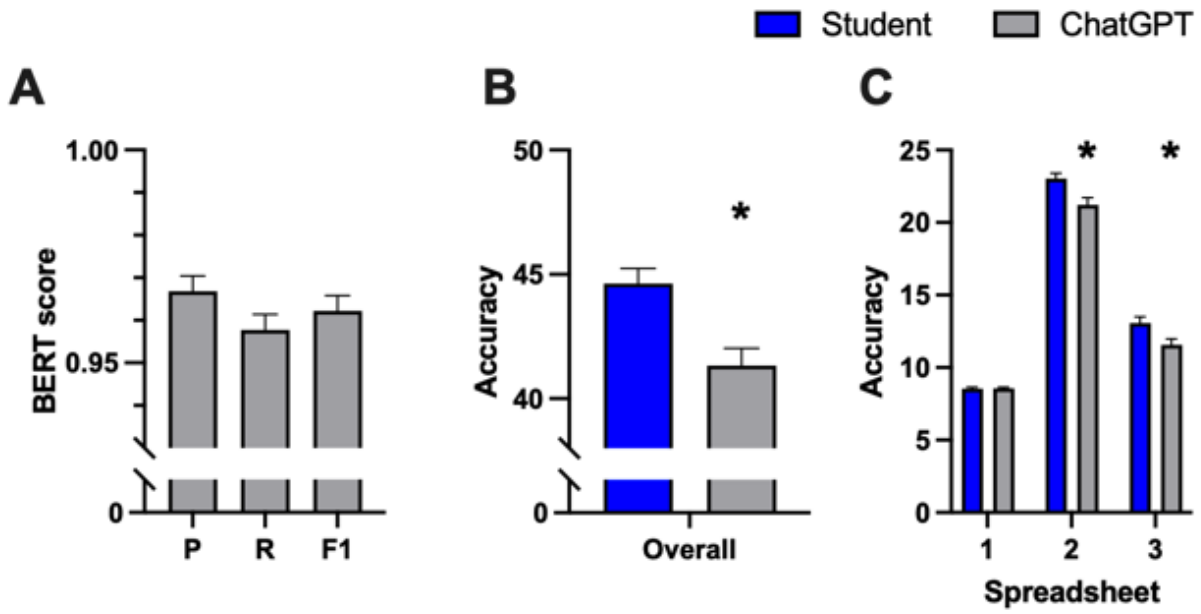

AI-Generated Significantly Shorter and Error-Free Radiology Reports

AI-generated radiology reports exhibited a significant 24% reduction in word count (mean 265.6, SD 95.4 vs mean 200.6, SD 37.3 words; P<.01; t_{198}=6.347) and sentence count (P=.007; t_{198}=2.726) accompanied by significant reductions in syllables (P<.01; t_{198}=6.823), diphthongs (P<.01; t_{198}=8.643), and characters (P<.01; t_{198}=6.841) compared to student-written texts as presented in Figure 6.

Figure 6. Analysis of (A) word count, (B) sentence count, (C) syllable count, (D) diphthong count, and (E) character count of AI-generated radiology reports compared to student-written reports. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence.

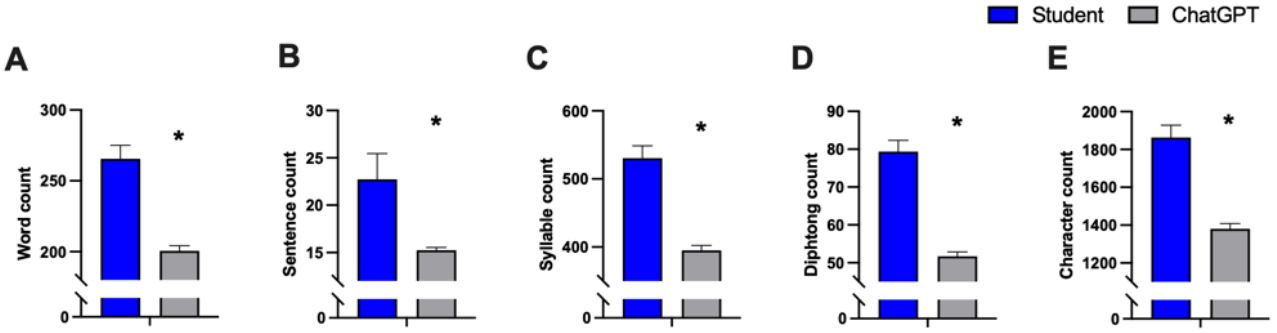

Whereas radiology reports generated by AI showed a significant reduction in ASL compared to student-written reports (A: mean 14.5, SD 3.1 vs mean 13.7, SD 1.8 words; P=.046; t_{198}=2.007), no difference was observed regarding the proportion of long words (B: P=.61; t_{198}=0.509) and this is presented in Figure 7.

XSL•FO
RenderX

Figure 7. Analysis of sentence length and long word proportion (more than 6 characters) of AI-generated radiology reports compared to student-written reports. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence.

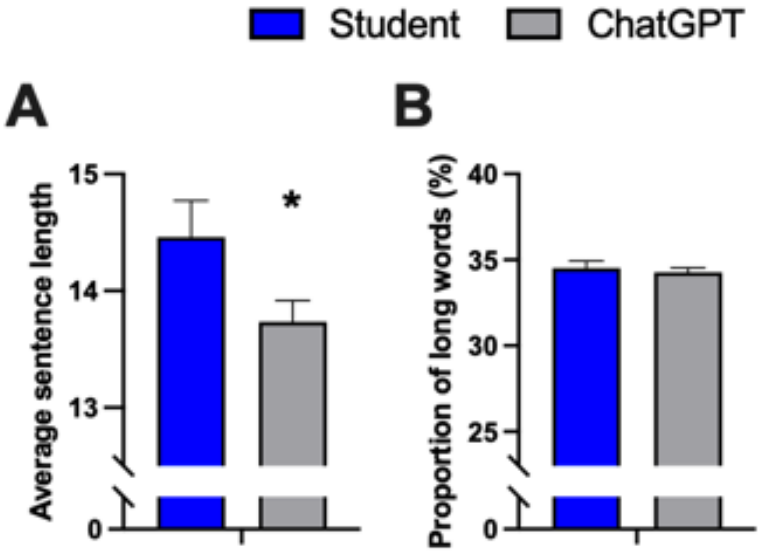

Contrary to student-written reports, AI-generated texts showed a complete absence of orthographic, grammatical, and punctuation errors as presented in Figures 8A and 8B (A: mean 7.7, SD 7.2 vs mean 0; P<.01; t_{198}=10.59; B: mean 2.9, SD 2.9 vs mean 0; P<.01; t_{198}=10.41).

Figure 8. Analysis of (A) error count including grammar, spelling, and punctuation and (B) error ratio by calculating errors divided by words multiplied by 100 of student-written radiology reports and AI-generated radiology reports. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence.

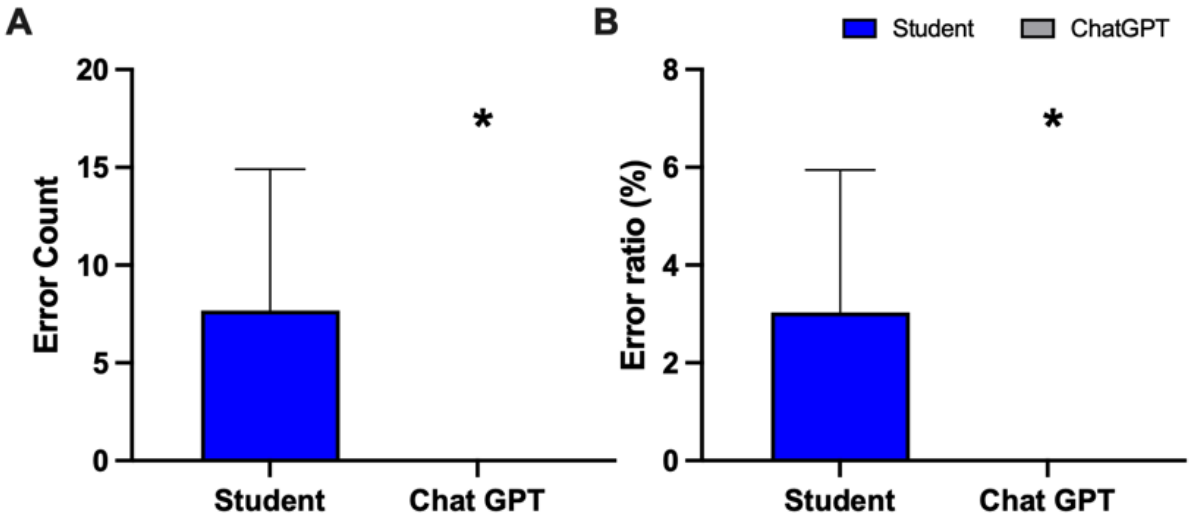

Student-Written Reports Showed Significant Differences in Length and Readability Across Semesters

Student-written reports showed a significant difference in word count across different semesters (P=.04; t_{99}=1.610), with a noticeable trend toward the use of more words in higher semesters (Figure 9A; semester 2 vs semester 5: mean 224, SD 79 vs mean 312, SD 107; P=.03; t_{95}=2.97). Additionally, significant differences in readability were observed across semesters, as presented in Figures 9B and 9C (LIX score: P=.06; t_{99}=2.315; FRE score: P<.01; t_{99}=2.762). Students from semesters 1 and 2 produced simpler and easier-to-understand reports, whereas those from higher semesters tended to write more complex and, hence, more difficult-to-read reports.

XSL•FO
RenderX

Figure 9. (A) Analysis of word count, (B) metric evaluation of readability using LIX score, and (C) Flesch reading ease score of student-written radiology reports. Data represent mean (SD). Sample size: n=100; semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20. *P<.05. LIX: Lesbarhetsindex.

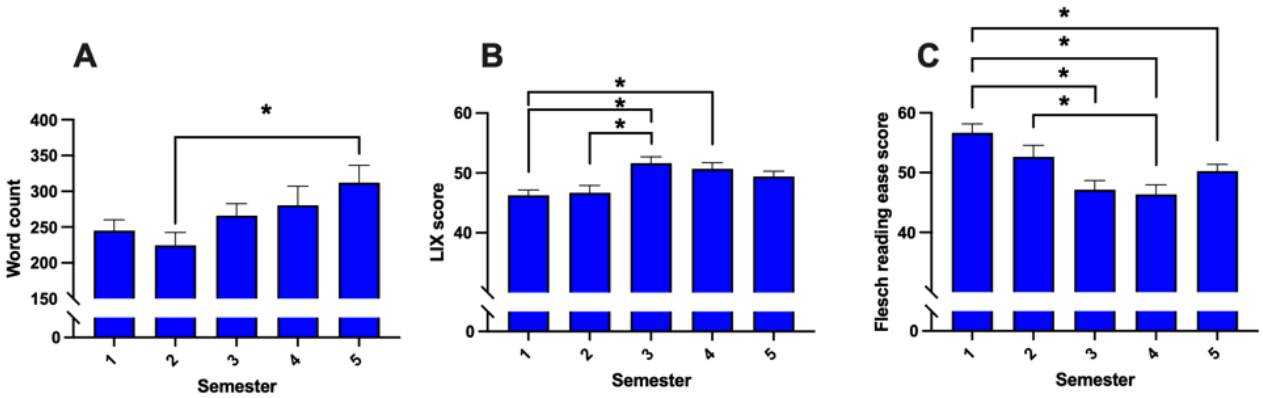

Discussion

Principal Findings

Integrating AI into clinical workflows and medical education has attracted significant interest due to its potential to enhance efficiency. Our study, therefore, aimed to investigate the effectiveness of AI-generated radiology reports by comparing them to student-written reports. In summary, no difference regarding readability was found between the AI-generated and student-written radiology reports. Whereas AI-generated reports showed an overall high textual similarity to reference texts, they simultaneously lacked substantial diagnostic information. Noteworthy, the language quality was significantly improved compared to student-written reports, with AI-generated texts being completely error-free. The results revealed high potential in enhancing medical writing with AI while still being limited by the reliability of the transferred information [36].

AI is revolutionizing medicine across diagnosis, treatment, and administrative tasks [5]. AI algorithms analyze medical images for early disease detection, provide clinical decision support, and enable personalized treatment plans [3]. In drug development, AI accelerates processes by predicting drug interactions and screening compounds [2]. Whereas the capability of AI to diagnose x-ray images has already been proven [4], the process of diagnosing was excluded in this study to focus on generating radiology reports based on information collected by students.

Indicated by an overall high BERT score, our findings prove that AI-generated radiology reports exhibit a great level of similarity to reference texts. Integration of ChatGPT into the medical documentation process, therefore, results in high text quality as represented by high precision, recall, and F_1-score, which further highlights the overall robustness of AI in replicating the content of reference reports only based on checkbox information. Concomitantly, after the preparation of preliminaries regarding the prompt and template design, automating the process of report writing with ChatGPT results in significant time savings and, hence, highlights the potential to streamline the workflow for health care professionals. Although this study did not aim to specifically quantify time efficiency, AI-supported report generation was noticeably quicker due to the shorter time cap (10 minutes for checkbox list completion vs 30 minutes for report writing). Whereas all students were observed to use the entire allotted time, likely focusing on thoroughness and ensuring report completeness, rather than being constrained by the time limit, future research could incorporate exact time measurements and posttask surveys to gather participant feedback on time allocation to provide additional insights regarding the adequacy of the time frames. The successful use of ChatGPT in composing medical notes related to patient transfers, operative procedures, and surgical assistance [5,18,37] underscores its role in enhancing productivity within medical environments, thereby highlighting the transformative impact of AI-driven technologies in health care.

Prior to the study, the prompt design was refined extensively, with multiple versions tested to optimize the AI's output. Ultimately, only the most effective prompt was selected to continue generating reports, ensuring the highest possible accuracy in AI-generated text. However, despite their overall similarity, AI-generated reports demonstrated a significant deficiency in relevant information, indicating a crucial impact of the prompt provided to ChatGPT in determining the accuracy of the results. This discrepancy was particularly evident in identifying and incorporating findings regarding specific teeth, with AI-generated reports containing significantly fewer findings compared to the number of diagnoses documented with the checkbox lists (eg, AI-generated reports did not mention the presence of a cyst, the status of dental restorations or precise prescription of bone loss). Interestingly, while no difference was observed in the findings reported from the first spreadsheet, a significant disparity emerged in subsequent spreadsheets. The potential limitation in the AI's ability to comprehensively interpret complex odontogram data leads to inconsistencies in the inclusion of relevant findings. These findings underscore the error-prone interplay between prompt precision, image complexity, and AI performance in radiology reporting.

In contrast to missing information, another known challenge in the use of ChatGPT is its tendency to generate plausible-sounding but incorrect or fabricated information, commonly referred to as "hallucinations" [38]. However, this

XSL•FO
RenderX

study showed no indication that ChatGPT included invented findings not present in the original checkbox list, as evidenced by the high BERT score. The prompt design strictly instructed the AI to use only information from the checkbox list, thereby minimizing the risk of hallucinations. Our observations confirmed that the model adhered to these guidelines. Although the prompt instructed ChatGPT to interpret the numbers as odontogram information to identify each tooth, we encountered challenges in consistently incorporating and accurately understanding the provided data. Precision in prompt formulation emerges as a critical factor influencing the accuracy and completeness of AI-generated reports [39]. The formulation of prompts significantly influences the outcomes generated by AI systems, with precise prompts being necessary to provide clear instructions and context for the AI model, guiding it in producing relevant and accurate responses. The specificity and clarity of the prompt directly impact the quality and relevance of the AI-generated output [40]. The design of effective prompts, therefore, remains a crucial part of future research.

Regarding radiology reports, a prompt that precisely outlines the required structure, format, and content of the report will likely result in more coherent outputs. Moreover, the prompt helps the AI model understand the task and focus on relevant information. By providing detailed guidelines and constraints, the prompt narrows the scope of the AI's search and directs it toward generating responses that align with the desired objectives. Additionally, prompts can incorporate domain-specific terminology and concepts to ensure that the AI model produces contextually appropriate and clinically relevant outputs. Nonetheless, a potential bias of machine learning systems must be considered due to their susceptibility to being influenced by the training data, thereby generating biased or misleading outputs. Well-designed prompts can help mitigate these issues by guiding the AI model toward more objective and accurate responses [41]. Moreover, the challenges associated with interpreting complex diagnostic data like orthopantomograms emphasize the need for continuous refinement and optimization of AI algorithms to ensure reliable performance in a clinical setting. Addressing these challenges will require a collaborative effort between clinicians, AI developers, and educators. Enhancing prompt precision through detailed guidelines and standardized protocols can improve AI performance and reduce information deficiencies in generated reports. Notably, to realize the full potential of AI in health care, the risk of disseminating misinformation must be mitigated. The rapid spread of false or misleading content, commonly called infodemic, highlights the importance of implementing validation mechanisms to ensure the reliability of AI-generated content [42,43].

In the context of this study, the AI was not supposed to formulate diagnoses independently but rather to generate radiology reports based explicitly on the findings and diagnoses provided by the students. This approach evaluated the AI's ability to effectively translate diagnostic information into coherent and comprehensive reports, reflecting real-world clinical scenarios of radiologists interpreting images and automatically converting their findings into written reports. Overall, this evaluation of AI-generated reports underscores the reliability and consistency of AI in producing error-free content compared to student-written texts. Remarkably, AI-generated reports exhibited a considerable reduction in word count, sentence count, and various linguistic features, including syllables and diphthongs. This reduction in length was accompanied by a notable absence of orthographic, grammatical, and punctuation errors, highlighting the accuracy and precision of AI-generated text. Moreover, no discernible difference between AI-generated and student-written radiology reports was observed regarding their readability. Both sets of reports demonstrated similar readability levels as indicated by the FRE score and LIX index, with both being established and validated as reliable measures for assessing text difficulty, including medical texts [27,29,30]. However, examination of individual semesters revealed significantly lower readability for AI-generated reports than student-written ones in the first clinical semester, but higher readability compared to the third and fourth clinical semesters. A possible explanation could be the use of more advanced and specialized terminology by the AI compared to students in the first semester, resulting in lower readability scores. Reports from students in the first semester may adhere to a simpler structure, reducing difficulty and increasing readability and comprehension. In contrast, reports from the third and fourth semesters exhibit more complexity in structure and terminology to present diagnostic information due to extensive expertise and, therefore, impairing readability. These findings are supported by the significant differences in word count and readability across all semesters upon individual examination. Students in lower semesters tend to use fewer words and write reports with higher readability, whereas students from higher semesters tend to write longer, more complex, and therefore, more difficult-to-read reports. As students progress through their education, their increased clinical experience and familiarity with radiological terminology likely enhance the quality of their reports. This development is reflected by the incorporation of advanced terminology and structure, indicating a clear learning curve. The variability in skill development across semesters significantly impacts the comparison between student-written and AI-generated reports, potentially affecting the comparison in favor of later semesters. This disparity underscores the importance of considering skill levels when evaluating AI performance since differences in student proficiency could lead to variability in report quality, affecting readability and accuracy metrics. Consequently, the perceived quality of AI-generated reports may vary depending on the student cohort they are compared with, highlighting the necessity of accounting for student skill differences in the study design and analysis. However, on the other hand, the variability in student skills across semesters positively reflects the diverse real-world conditions in clinical practice, where practitioners exhibit a range of expertise. This diversity in the study cohort allows the AI-generated reports to be tested against various levels of proficiency, demonstrating the AI's potential to support users with different levels. Early-stage dental students could benefit from a structured and consistent framework provided by AI, enhancing their learning and understanding. Advanced students and experienced clinicians could use AI to reduce repetitive tasks and ensure accuracy in documentation, allowing more focus on diagnostic decision-making. AI assistance in

XSL•FO
RenderX

diagnostics has been further shown to improve performance, though radiologists often underweight AI predictions [44]. Overall, AI tools can support a wide range of users by adapting to their specific needs and improving educational and clinical outcomes. This aligns with study results proving GPT-4 to enhance productivity and quality in various tasks beyond medical use, benefiting consultants and customer support agents across all skill levels [45,46].

Nevertheless, the differentiated use of reports must be considered due to the diverse communication needs within health care settings. On the one hand, health care professionals require detailed reports for accurate clinical decision-making and effective interprofessional communication. On the other hand, patients benefit from simpler, more understandable reports to understand their medical conditions and actively engage in treatment discussions. AI has been further shown to efficiently simplify medical data for better patient understanding [22]. Hence, its implementation offers the possibility to fulfill both the detailed requirements of health care professionals and the simplified needs of patients simultaneously in response to 2 different prompts. Consequently, automated AI solutions could facilitate effective communication among health care providers and increase patient empowerment and participation beyond time-saving and more efficient documentation in health care.

Conclusions

In conclusion, AI's potential to enhance medical writing efficiency is highlighted, yet remaining challenges in ensuring reliability and comprehensiveness must be faced. The precision of prompts significantly impacts AI's accuracy, particularly in interpreting complex diagnostic data. Future research should focus on refining AI algorithms and prompt design to optimize medical reporting. Overall, integrating AI-driven solutions into routine clinical workflows offers a practical tool for enhancing productivity.

Acknowledgments

This study was conducted as part of ASB's PhD thesis.

Conflicts of Interest

None declared.

References

1. Ebbers T, Kool RB, Smeele LE, Dirven R, den Besten CA, Karssemakers LHE, et al. The impact of structured and standardized documentation on documentation quality; a multicenter, retrospective study. J Med Syst 2022;46(7):46 [FREE Full text] [doi: 10.1007/s10916-022-01837-9] [Medline: 35618978]
2. Vemula D, Jayasurya P, Sushmitha V, Kumar YN, Bhandari V. CADD, AI and ML in drug discovery: a comprehensive review. Eur J Pharm Sci 2023;181:106324 [FREE Full text] [doi: 10.1016/j.ejps.2022.106324] [Medline: 36347444]
3. Weisberg EM, Fishman EK. The future of radiology and radiologists: AI is pivotal but not the only change afoot. J Med Imaging Radiat Sci 2024;55(4):101377. [doi: 10.1016/j.jmir.2024.02.002] [Medline: 38403516]
4. Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, et al. Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: a multi-reader study. Korean J Radiol 2022;23(5):505-516 [FREE Full text] [doi: 10.3348/kjr.2021.0476] [Medline: 35434976]
5. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
6. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
7. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Machines 2020;30(4):681-694. [doi: 10.1007/s11023-020-09548-1]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med 2023;99(2):192-197. [doi: 10.1097/acm.0000000000005549]
10. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y) 2024;5(3):100943 [FREE Full text] [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
11. Hwang T, Aggarwal N, Khan PZ, Roberts T, Mahmood A, Griffiths MM, et al. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. PLoS One 2024;19(2):e0297701 [FREE Full text] [doi: 10.1371/journal.pone.0297701] [Medline: 38354135]
12. Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC. The use of ChatGPT for education modules on integrated pharmacotherapy of infectious disease: educators' perspectives. JMIR Med Educ 2024;10:e47339 [FREE Full text] [doi: 10.2196/47339] [Medline: 38214967]

XSL•FO
RenderX

13. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
14. Flory MN, Napel S, Tsai EB. Artificial intelligence in radiology: opportunities and challenges. Semin Ultrasound CT MR 2024;45(1):152-160.
15. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. Cureus 2023;15(7):e42133 [FREE Full text] [doi: 10.7759/cureus.42133] [Medline: 37476297]
16. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32(11):7998-8007 [FREE Full text] [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
17. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181. [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
18. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. Ann Biomed Eng 2023;51(11):2353-2355. [doi: 10.1007/s10439-023-03263-5] [Medline: 37266720]
19. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34(5):2817-2825 [FREE Full text] [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
20. Qutieshat A, Al Rusheidi A, Al Ghammari S, Alarabi A, Salem A, Zelihic M. Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence. Diagnosis (Berl) 2024;11(3):259-265 [FREE Full text] [doi: 10.1515/dx-2024-0034] [Medline: 38696271]
21. Yokokawa D, Yanagita Y, Li Y, Yamashita S, Shikino K, Noda K, et al. For any disease a human can imagine, ChatGPT can generate a fake report. Diagnosis (Berl) 2024;11(3):329-332. [doi: 10.1515/dx-2024-0007] [Medline: 38386808]
22. Fink MA. [Large language models such as ChatGPT and GPT-4 for patient-centered care in radiology]. Radiologie (Heidelb) 2023;63(9):665-671. [doi: 10.1007/s00117-023-01187-8] [Medline: 37615692]
23. Currie G, Robbie S, Tually P. ChatGPT and patient information in nuclear medicine: GPT-3.5 versus GPT-4. J Nucl Med Technol 2023;51(4):307-313. [doi: 10.2967/jnmt.123.266151] [Medline: 37699647]
24. Flesch R. A new readability yardstick. J Appl Psychol 1948;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
25. Amstad T. Wie verständlich sind unsere Zeitungen?. German: Studenten-Schreib-Service, Zu¨rich; 1978.
26. Gajjar AA, Kumar RP, Paliwoda ED, Kuo CC, Adida S, Legarreta AD, et al. Usefulness and accuracy of artificial intelligence chatbot responses to patient questions for neurosurgical procedures. Neurosurgery 2024;95(1):171-178. [doi: 10.1227/neu.0000000000002856] [Medline: 38353558]
27. Gajjar AA, Patel S, Patel SV, Goyal A, Sioutas GS, Gamel KL, et al. Readability of cerebrovascular diseases online educational material from major cerebrovascular organizations. J Neurointerv Surg 2024. [doi: 10.1136/jnis-2023-021205] [Medline: 38395602]
28. Irwin SC, Lennon DT, Stanley CP, Sheridan GA, Walsh JC. Ankle conFUSION: the quality and readability of information on the internet relating to ankle arthrodesis. Surgeon 2021;19(6):e507-e511. [doi: 10.1016/j.surge.2020.12.001] [Medline: 33451875]
29. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. Health Educ Behav 2006;33(3):352-573. [doi: 10.1177/1090198105277329] [Medline: 16699125]
30. Skrzypczak T, Mamak M. Assessing the readability of online health information for colonoscopy - analysis of articles in 22 european languages. J Cancer Educ 2023;38(6):1865-1870 [FREE Full text] [doi: 10.1007/s13187-023-02344-2] [Medline: 37493981]
31. Anderson J. LIX and RIX: variations on a little-known readability index. J Reading 1983;26(6):490-496.
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv.0 .
33. Deutsch T, Jasbi M, Shieber S. Linguistic features for readability assessment. arXiv preprint arXiv.00377 . [doi: 10.18653/v1/2020.bea-1.1]
34. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV information on Google versus AI (ChatGPT). Otolaryngol Head Neck Surg 2024;170(6):1504-1511. [doi: 10.1002/ohn.506] [Medline: 37622581]
35. Sikander B, Baker JJ, Deveci CD, Lund L, Rosenberg J. ChatGPT-4 and human researchers are equal in writing scientific introduction sections: a blinded, randomized, non-inferiority controlled study. Cureus 2023;15(11):e49019 [FREE Full text] [doi: 10.7759/cureus.49019] [Medline: 38111405]
36. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American Heart Association's advanced cardiovascular life support guidelines: exploratory study. J Med Internet Res 2024;26:e55037 [FREE Full text] [doi: 10.2196/55037] [Medline: 38648098]
37. Atkinson CJ, Seth I, Xie Y, Ross RJ, Hunter-Smith DJ, Rozen WM, et al. Artificial intelligence language model performance for rapid intraoperative queries in plastic surgery: ChatGPT and the deep inferior epigastric perforator flap. J Clin Med 2024;13(3):900 [FREE Full text] [doi: 10.3390/jcm13030900] [Medline: 38337594]

XSL•FO
RenderX

38. Huang L. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv preprint arXiv.05232 .
39. Nazary F, Deldjoo Y, Di Noia T. Harnessing the power of XAI in prompt-based healthcare decision support using ChatGPT. ChatGPT-HealthPrompt. 2024. URL: https://link.springer.com/chapter/10.1007/978-3-031-50396-2_22 [accessed 2024-01-21]
40. White J, Hays S, Fu Q, Smith JS, Schmidt DC. ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. arXiv preprint arXiv.07839 . [doi: 10.1007/978-3-031-55642-5_4]
41. Hu X, Tian Y, Nagato K, Nakao M, Liu A. Opportunities and challenges of ChatGPT for design knowledge management. Procedia CIRP 2023;119:21-28. [doi: 10.1016/j.procir.2023.05.001]
42. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
43. Wang G, Gao K, Liu Q, Wu Y, Zhang K, Zhou W, et al. Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: comprehensive comparative analysis of generative and authoritative information. J Med Internet Res 2023;25:e49771 [FREE Full text] [doi: 10.2196/49771] [Medline: 38096014]
44. Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence: experimental evidence from radiology. NBER. 2023. URL: https://www.nber.org/papers/w31422 [accessed 2024-06-17]
45. Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, et al. Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. SSRN J 2023:58. [doi: 10.2139/ssrn.4573321]
46. Brynjolfsson E, Li D, Raymond LR. Generative AI at work. arXiv:2304.11771 2023. [doi: 10.3386/w31161]

Abbreviations

AI: artificial intelligence
ASL: average sentence length
ASW: average number of syllables per word
BERT: Bidirectional Encoder Representations from Transformers
FRE: Flesch reading ease
OPG: panoramic radiograph

Edited by Q Jin; submitted 18.05.24; peer-reviewed by A Qutieshat, R Tabari Khomeiran, S Gorthy, R Janssen; comments to author 29.06.24; revised version received 19.07.24; accepted 03.08.24; published 23.12.24.

Please cite as:
Stephan D, Bertsch A, Burwinkel M, Vinayahalingam S, Al-Nawas B, Kämmerer PW, Thiem DGE
AI in Dental Radiology—Improving the Efficiency of Reporting With ChatGPT: Comparative Study
J Med Internet Res 2024;26:e60684
URL: https://www.jmir.org/2024/1/e60684
doi:10.2196/60684
PMID:

©Daniel Stephan, Annika Bertsch, Matthias Burwinkel, Shankeeth Vinayahalingam, Bilal Al-Nawas, Peer W Kämmerer, Daniel GE Thiem. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis

Hui Zong[1*], PhD; Rongrong Wu[1*], PhD; Jiaxue Cha[2], PhD; Jiao Wang[1], PhD; Erman Wu[1,3], PhD; Jiakun Li[1,4], MD, PhD; Yi Zhou[1], PhD; Chi Zhang[1], MD, PhD; Weizhe Feng[1], MS; Bairong Shen[1,5], PhD

[1]Joint Laboratory of Artificial Intelligence for Critical Care Medicine, Department of Critical Care Medicine and Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, China

[2]Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China

[3]Department of Neurosurgery, First Affiliated Hospital of Xinjiang Medical University, Urumqi, China

[4]Department of Urology, West China Hospital, Sichuan University, Chengdu, China

[5]West China Tianfu Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Bairong Shen, PhD
Joint Laboratory of Artificial Intelligence for Critical Care Medicine, Department of Critical Care Medicine and Institutes for Systems Genetics
Frontiers Science Center for Disease-related Molecular Network, West China Hospital
Sichuan University
No. 37 Guoxue Alley
Chengdu, 610041
China
Phone: 86 28 61528682
Email: bairong.shen@scu.edu.cn

Abstract

Background: Large language models (LLMs) are increasingly integrated into medical education, with transformative potential for learning and assessment. However, their performance across diverse medical exams globally has remained underexplored.

Objective: This study aims to introduce MedExamLLM, a comprehensive platform designed to systematically evaluate the performance of LLMs on medical exams worldwide. Specifically, the platform seeks to (1) compile and curate performance data for diverse LLMs on worldwide medical exams; (2) analyze trends and disparities in LLM capabilities across geographic regions, languages, and contexts; and (3) provide a resource for researchers, educators, and developers to explore and advance the integration of artificial intelligence in medical education.

Methods: A systematic search was conducted on April 25, 2024, in the PubMed database to identify relevant publications. Inclusion criteria encompassed peer-reviewed, English-language, original research articles that evaluated at least one LLM on medical exams. Exclusion criteria included review articles, non-English publications, preprints, and studies without relevant data on LLM performance. The screening process for candidate publications was independently conducted by 2 researchers to ensure accuracy and reliability. Data, including exam information, data process information, model performance, data availability, and references, were manually curated, standardized, and organized. These curated data were integrated into the MedExamLLM platform, enabling its functionality to visualize and analyze LLM performance across geographic, linguistic, and exam characteristics. The web platform was developed with a focus on accessibility, interactivity, and scalability to support continuous data updates and user engagement.

Results: A total of 193 articles were included for final analysis. MedExamLLM comprised information for 16 LLMs on 198 medical exams conducted in 28 countries across 15 languages from the year 2009 to the year 2023. The United States accounted for the highest number of medical exams and related publications, with English being the dominant language used in these exams. The Generative Pretrained Transformer (GPT) series models, especially GPT-4, demonstrated superior performance, achieving pass rates significantly higher than other LLMs. The analysis revealed significant variability in the capabilities of LLMs across different geographic and linguistic contexts.

XSL•FO
RenderX

Conclusions: MedExamLLM is an open-source, freely accessible, and publicly available online platform providing comprehensive performance evaluation information and evidence knowledge about LLMs on medical exams around the world. The MedExamLLM platform serves as a valuable resource for educators, researchers, and developers in the fields of clinical medicine and artificial intelligence. By synthesizing evidence on LLM capabilities, the platform provides valuable insights to support the integration of artificial intelligence into medical education. Limitations include potential biases in the data source and the exclusion of non-English literature. Future research should address these gaps and explore methods to enhance LLM performance in diverse contexts.

(J Med Internet Res 2024;26:e66114) doi:10.2196/66114

KEYWORDS

large language models; LLMs; generative pretrained transformer; ChatGPT; medical exam; medical education; artifical intelligence; AI

Introduction

Generative pretrained transformers (GPTs) and large language models (LLMs) have revolutionized the field of natural language processing. These models, supported by advanced techniques such as deep learning and attention mechanisms, have demonstrated remarkable capabilities for understanding and generating human-like text. ChatGPT, developed by OpenAI, has garnered widespread attention for its ability to engage in coherent and context-aware conversations [1]. Additionally, there are other models such as Llama [2], Bard [3], and PaLM [4]. These LLMs have had a significant impact in the fields of biomedicine and health care [5-8], particularly in the context of medical education [9-11].

Medical exams, such as licensing and certification tests for health care professionals, play a vital role in improving patient safety and practitioner competence and hold immense significance in medical education and the health care system [12]. These rigorous exams require a deep understanding of medical knowledge, advanced clinical reasoning skills, and contextual understanding of real-world patient scenarios. Medical exams serve as a crucial step in the process of becoming a certified and licensed medical professional [13]. The exams assess the medical knowledge and clinical skills of health care professionals, including physicians, surgeons, nurses, pharmacists, and students [14-17]. The medical exams cover a wide range of medical topics, including basic sciences (eg, anatomy, biochemistry, physiology), clinical medicine (eg, internal medicine, pediatrics, surgery, psychiatry, obstetrics and gynecology), and even specialized areas. These medical exams generally take the form of multiple-choice questions or short-answer questions. Preparing for the medical exam requires extensive studying of medical textbooks, clinical guidelines, and relevant publications. Various resources, such as question banks, practice exams, and online learning platforms, can be used to prepare for the medical exam.

LLMs have demonstrated considerable potential in medical education, as well as considerable capabilities for medical exams worldwide [18,19]. A large number of studies have been conducted to evaluate the performance of LLMs in medical exams across different countries and languages [20,21]. For example, ChatGPT has shown promising results for the United States Medical Licensing Exam, which consists of 3 steps: Step 1, Step 2CK, and Step 3 [22-24]. ChatGPT has also exhibited great potential for the Chinese National Medical Licensing Examination, Chinese National Pharmacist Licensing Examination, and Chinese National Nurse Licensing Examination during a 5-year evaluation study [17]. Other models like GPT-4, Bing, and GPT-3.5-Turbo have demonstrated their capabilities for the German Medical State Examinations [25], Japanese National Medical Licensing Examination [26], Australian Medical Council Licensing Examination [27], and Korean National Licensing Examination for Korean Medicine Doctors [28]. The performance evaluation of LLMs in these medical exams necessitates not only the understanding of complex medical concepts but also translational application for solving intricate clinical problems.

It is the first time that such a substantial number of studies have delved into the potential and capability of artificial intelligence models for medical exams worldwide. These studies not only present evaluation results and research evidence but also offer valuable insights for advancing LLMs in biomedicine and health care [29]. Furthermore, these studies have facilitated the integration of artificial intelligence into medical education.

Despite the growing interest in LLM performance on medical exams, existing studies remain fragmented, with data scattered across diverse sources and lacking standardization. This fragmentation creates a critical gap in understanding the broader trends and disparities in LLM capabilities across geographic, linguistic, and contextual boundaries. Although systematic reviews synthesize evidence, they do not provide an interactive, centralized system to continuously analyze, update, and compare LLM performance on medical exams. Moreover, current evidence often varies in quality, and not all studies are readily accessible to researchers and educators. To address these gaps, we proposed developing a comprehensive platform that provides a centralized system for collecting, analyzing, and comparing evidence-based knowledge regarding the performance of various LLMs across a wide range of medical exams around the world.

In this study, we introduce MedExamLLM, a platform specifically designed for benchmarking LLMs for medical exams around the world. The main contributions of our study lie in 3 key aspects: (1) We present MedExamLLM, an open-source, freely accessible, and publicly available platform, which serves as a crucial resource by providing performance evaluation information and evidence-based knowledge of the comprehensive capability of LLMs on medical exams worldwide; (2) MedExamLLM compiles data on 16 LLMs across 198 medical exams conducted in 28 countries, covering 15 languages, and spanning from the year 2009 to the year 2023;

XSL•FO
RenderX

(3) MedExamLLM enables information retrieval, in-depth analysis, and performance comparison, which can facilitate translational research of artificial intelligence technologies in health care and medical education.

In this study, we aimed to address 3 key research questions (RQs) to enhance the understanding and evaluation of LLM capabilities for medical exams conducted worldwide (Figure 1C). These research questions were designed to systematically investigate and document the interplay between LLMs and medical exams, facilitating the advancement of artificial intelligence in medical education.

RQ1 was "What are the characteristics of medical exams worldwide?" This question focused on the collection, standardization, organization, management, and analysis of data pertaining to the medical exams. This detailed characterization will provide an in-depth understanding of the global landscape of medical exams in the evaluation of LLMs.

RQ2 was "How do LLMs perform in these medical exams?" To evaluate the performance of LLMs, we collected extensive performance data for each model, including their names, versions, number of correct responses, accuracy rates, and pass and fail statuses. This analysis included examining geographical and linguistic differences in LLM capabilities and comparing the performance of different models. By doing so, we aimed to identify trends and disparities in LLM performance across various contexts and regions.

RQ3 was "How can the capabilities of LLMs in medical exams worldwide be tracked?" We proposed the development of an open-source platform that features a comprehensive leaderboard to showcase the performance of LLMs on medical exams globally. This platform not only highlights the capabilities of different models but also provides detailed information on the characteristics and accessibility of the medical exam data sets used for evaluation. This platform will serve as a valuable resource for ongoing monitoring and assessment of LLM performance in medical education.

Figure 1. Overview of the study, including (A) the systematic search process for publications related to generative artificial intelligence and large language models (LLMs) on medical exams, (B) article screening and inclusion process used to select studies for analysis, and (C) key research questions addressed by the study. RQ: research question.

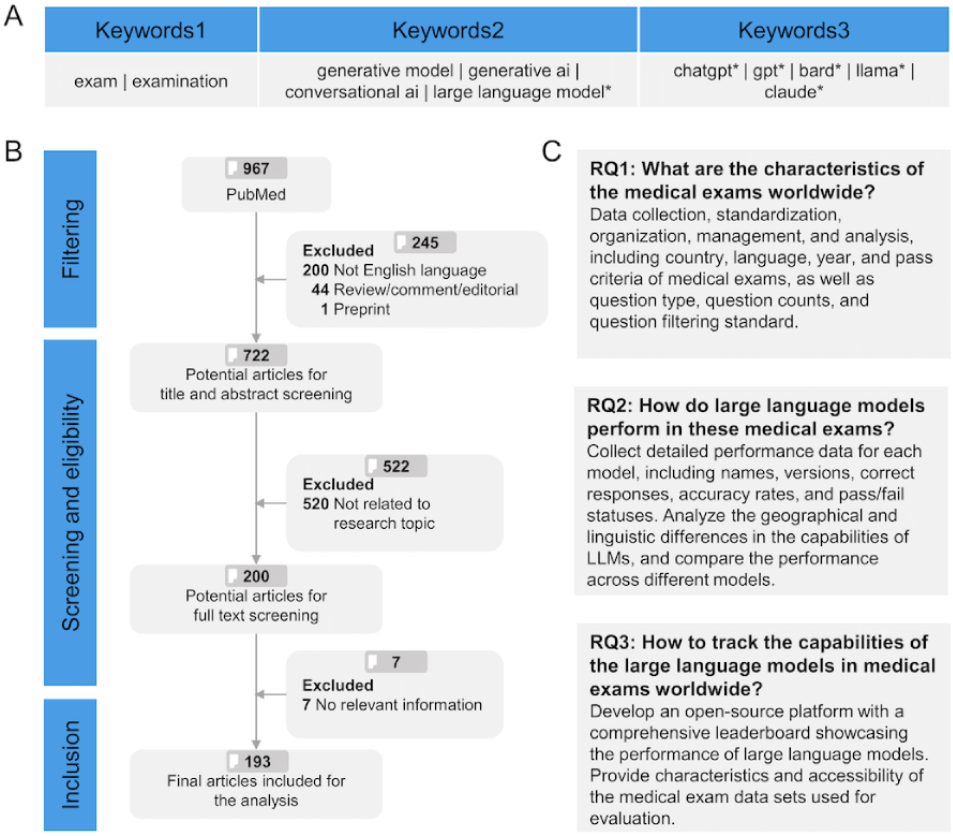

Methods

Data Collection

A systematic search was conducted in the PubMed database on April 25, 2024, to identify relevant articles. To ensure comprehensive coverage of the research questions, we selected 3 categories of keywords, as shown in Figure 1A: (1) general terms related to generative artificial intelligence (eg, generative model), (2) specific terms referring to LLMs (eg, ChatGPT), and (3) terms related to medical exams (eg, exam). The full search query used in PubMed was as follows: (generative model[TIAB] OR generative ai[TIAB] OR conversational ai[TIAB] OR large language model*[TIAB] OR chatgpt*[TIAB] OR gpt*[TIAB] OR bard*[TIAB] OR llama*[TIAB] OR claude*[TIAB]) AND (exam[TIAB] OR examination[TIAB]). By leveraging PubMed as the primary data source, we ensured the inclusion of peer-reviewed studies, which provide a high level of evidence quality and reliability. The inclusion criteria

XSL•FO
RenderX

encompassed the following: (1) the article described original research, (2) the article was published in the English language, (3) the article investigated the use of at least one LLM on medical exams. The exclusion criteria included the following: (1) regarding article type, reviews, comments, or editorials, as these typically lack primary data and experimental results on LLM performance, which are essential for our analysis, as well as preprints, to ensure all included studies had undergone peer review for accuracy and validation; (2) regarding language, articles published in non-English languages to ensure consistent interpretation of data; (3) regarding content relevance, articles that did not use any LLMs or provide relevant information, such as exam details, model descriptions, and performance evaluation. The inclusion of all candidate articles was verified by 2 researchers. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist [30] is included in Multimedia Appendix 1. The overall process of article selection is shown in Figure 1B.

Content Curation

To provide a comprehensive understanding of the capability of LLMs on medical exams worldwide, we manually curated and standardized information for each study, integrating it into a structured database designed for the platform. The curated data informed the key components of the MedExamLLM platform. These data included (1) exam information (name, country or region, language, year, question types, description, and passing criteria for the medical exams), (2) data processing (description of question filters, as well as the number of questions before and after screening), (3) model performance (name of the LLM, number of correctly answered questions, score, accuracy, and whether the platform passed), (4) data availability (whether the data are publicly available and the download link), (5) reference (basic information about the referenced study, including PMID, authors, title, abstract, publication date, publication year, and journal). The detailed descriptions and examples of each item are shown in Table 1.

The search process and curated information were foundational to addressing the 3 RQs: They enable detailed characterization of medical exams (RQ1), support evaluation and comparison of LLM performance (RQ2), and facilitate the creation of a tracking platform for LLM capabilities (RQ3).

During data processing, we followed a multistep protocol to ensure reliability. First, exam details, model names, and performance metrics were extracted and standardized to ensure consistency across studies. This included normalizing exam names, question types, and scoring metrics to align with predefined categories. To mitigate errors, 2 independent researchers (HZ and RW) verified the data entry for accuracy. Any discrepancies were resolved through discussion and, when necessary, by consulting with another researcher (BS) until consensus was achieved. Finally, we used a cross-verification step to ensure data quality and consistency. This involved applying a standardized review of the curated content across all entries, with special attention to aligning these elements with both the original study methodologies and predefined data framework.

The curated information was then integrated into the platform, serving as the backbone for its visualization and interaction capabilities. For instance, standardized exam names and scoring metrics allowed dynamic filtering and comparison of model performance across multiple exams, languages, and geographic regions.

XSL•FO
RenderX

Table 1. Information curated for each reference, including exam information, data processing, model performance, data availability, and reference.

Data categories and items	Data description
Exam information	
Exam country	The geographic area or nation state where an exam is conducted
Exam language	The specific language used in an exam
Exam year	The year in which an exam was conducted
Exam name	The name or title of an exam
Exam description	A detailed description of what an exam covers
Question type	The category or style of question used in an exam
Pass criteria	The standards or conditions that must be met for an individual to be considered as having passed an exam
Data processing	
Question filter	A method or process used to select or screen questions for an exam
Question num before	The number of questions in an exam before a certain filtering process was applied
Question num after	The number of questions in an exam after a certain filtering process was applied
Model performance	
Model	A type of artificial intelligence, specifically a large language model, designed to simulate human-like interaction by understanding, interpreting, and generating human language
Correct response	The count of questions answered correctly by the model during a testing
Score	A numerical representation of the model's performance on an examination, often calculated based on the total number of correct responses
Accuracy	A measure of the model's performance, typically calculated as the ratio of the number of correct responses to the total number of questions
Pass	Refers to whether a model has met the pass criteria for an exam
Data availability	Refers to whether the data of an exam is accessible or can be downloaded by others for reuse in subsequent research
Reference	
PMID	A unique PubMed identifier assigned to each article in the PubMed database
Authors	The name list of researchers who conducted the research or published the article
Title	The heading of the article, which typically summarizes the main subject or focus of the research
Abstract	The abstract of the article, often including the study's purpose, methods, results, and conclusions
Publication date	The specific date on which the article was published
Publication year	The specific year in which the article was published
Journal	The name of the academic or scientific journal in which the article was published

Platform Implementation

The MedExamLLM platform was designed with a focus on accessibility, interactivity, and scalability to support comprehensive benchmarking of LLMs on medical exams. Key design objectives included enabling users to explore performance data across diverse exams and model versions and allowing updates as new LLM results become available. The platform's architecture integrates a database to manage and structure exam and model data, allowing users to effectively query and visualize performance metrics.

To achieve these aims, we implemented a modular web framework that facilitates data management, interactive data visualization, and user feedback. The platform includes dynamic charting capabilities, enabling users to compare model performance across different exams, languages, countries, and models while providing evidence support from the published literature. These design elements ensure that MedExamLLM can serve as a long-term resource that adapts to evolving needs in medical artificial intelligence research.

Statistical Analysis

The chi-square test was used to assess the performance differences of LLMs across different languages and model types. The chi-square test was selected for its suitability for comparing categorical variables, enabling us to determine whether observed differences in performance were statistically significant. This approach ensured a rigorous evaluation of both linguistic and model-based performance variability. Statistical analyses were performed using the R software package (version 4.2.1).

XSL•FO
RenderX

Ethical Considerations

This study exclusively used publicly available data sets and did not involve any experiments on human subjects or animals. Additionally, no private or sensitive patient information was used. Consequently, this research did not require approval from an ethics review board, in accordance with the ethical guidelines of the authors' institution.

Results

Platform Overview

The MedExamLLM platform is freely accessible at [31]. It was designed with a user-friendly web interface to facilitate the search, visualization, submission, and download of data. The primary modules include the LLM performance leaderboard module, medical exam information search module, and medical exam data set module (Figure 2).

The LLM performance leaderboard module displays 709 entries, encompassing 16 large models across 198 medical exams from 28 countries and covering 15 languages. The data elements include the standardized names of the medical exams, the countries in which they were administered, the languages used, the year of implementation, the LLM tested, the score of the LLM, the accuracy of the LLM, whether it passed the exam, and related references. The leaderboard can be filtered by country, language, model, and pass status, and it also can be sorted according to the aforementioned data elements.

The medical exam search module presents comprehensive information on the medical exams of interest, including publications, performance, and statistics. The publication information shows all the published articles related to the searched medical exam, including title, author, journal, and year of publication. The performance information includes the model, as well as its score, accuracy, and pass status. The statistical information offers descriptive statistics and visualizations for all LLMs associated with the specific medical exam.

The medical exam data set module provides detailed information on the medical exam data sets, including the data set description, pass criteria, question type, question filtering process, the number of questions before and after filtering, and download methods.

Additionally, the MedExamLLM platform includes submission, download, and statistics modules. The submission module allows users to submit new evaluation results of LLMs on medical exams, which helps to ensure the accuracy, completeness, and timely update of the data in the platform. The download module provides links to all data generated by this study. The statistics module displays comprehensive statistical descriptions and graph visualizations.

Figure 2. Overview of the structure and features on the MedExamLLM platform, including core modules such as the large language model (LLM) performance leaderboard, medical exam information search, and medical exam data set management, as well as statistics visualization, user submission of new data, and data set download.

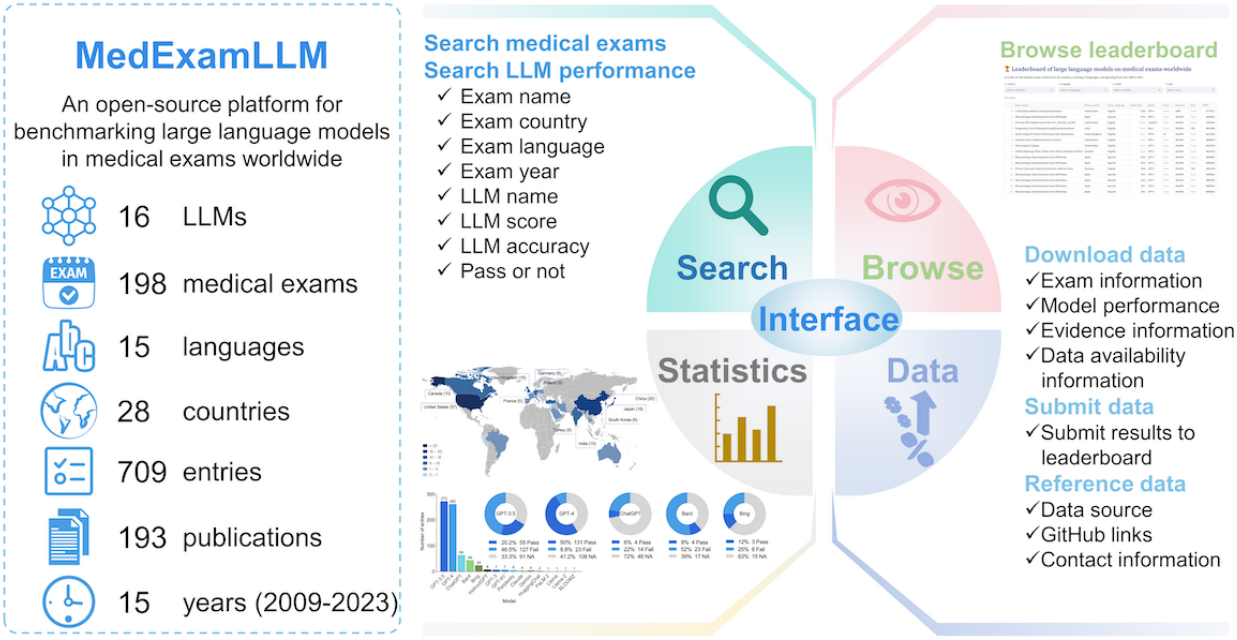

Data Statistics

A total of 967 articles were initially retrieved for consideration. These studies underwent a sequential screening process based on the predefined eligibility criteria. First, the metadata of the studies were examined, leading to the exclusion of 200 articles written in non-English languages; 44 articles categorized as a review, comment, or editorial; and 1 article identified as a preprint. Next, of the remaining 722 articles, the titles and abstracts were manually assessed. Among these, 520 articles were found to be irrelevant to the research topic and were excluded. Subsequently, a thorough examination of the full texts of the remaining articles was conducted, resulting in the identification of 7 articles that did not provide relevant information. Finally, a total of 193 articles were included in this study (Figure 1B).

XSL•FO
RenderX

The types and versions of LLMs used, as well as the categories and years of medical exams, varied across different articles. For example, Cheong et al [32] investigated the performance of 3 models (GPT-3.5, GPT-4, and Google Bard) across 10 categories of the Sleep Medicine Certification Board Exam. This research involved a total of 30 experiments, resulting in the inclusion of 30 entries on the MedExamLLM platform. Zong et al [17] evaluated the performance of ChatGPT on the Chinese National Medical Licensing Examination, Chinese National Pharmacist Licensing Examination, and Chinese National Nurse Licensing Examination from the year 2017 to the year 2021. A total of 15 experiments were conducted, leading to the inclusion of 15 entries on the MedExamLLM platform. At the time the manuscript was written, the MedExamLLM platform comprised 709 structured entries curated from 193 scholarly articles.

Distribution of Countries, Languages, and Years for Medical Exams

The MedExamLLM platform collected medical exams in 15 languages from 28 countries, spanning from the year 2009 to the year 2023. The United States had the highest number of medical exams, with 57 publications, followed by China (20 publications), Japan (19 publications), the United Kingdom (16 publications), Canada (10 publications), India (10 publications), Germany (8 publications), Turkey (6 publications), South Korea (6 publications), France (5 publications), and Poland (5 publications; Figure 3). English emerged as the predominant language used in these medical exams, with 127 publications using exams conducted in English, followed by Chinese (20 publications), Japanese (19 publications), German (5 publications), Korean (5 publications), and Polish (5 publications; Table 2). The majority of the studies used medical exams administered in the year 2022 (72 publications), followed by 2023 (35 publications), 2021 (23 publications), 2020 (16 publications), and 2019 (11 publications; Table 2).

The sources of the exam questions were diverse. Most studies collected questions from national medical licensing examinations from different countries. Some studies also used questions sourced from examination books, question banks, medical textbooks, and websites. These exams encompassed various hospital roles, including physicians, surgeons, pharmacists, nurses, and other health care professionals. They also covered a wide range of specialties and disciplines, including urology, radiology, cardiology, neurology, ophthalmology, anesthesiology, dermatology, plastic surgery, neurological surgery, and gynecology and obstetrics.

Figure 3. World distribution of the medical exams in the MedExamLLM platform. The map was created using ECharts.

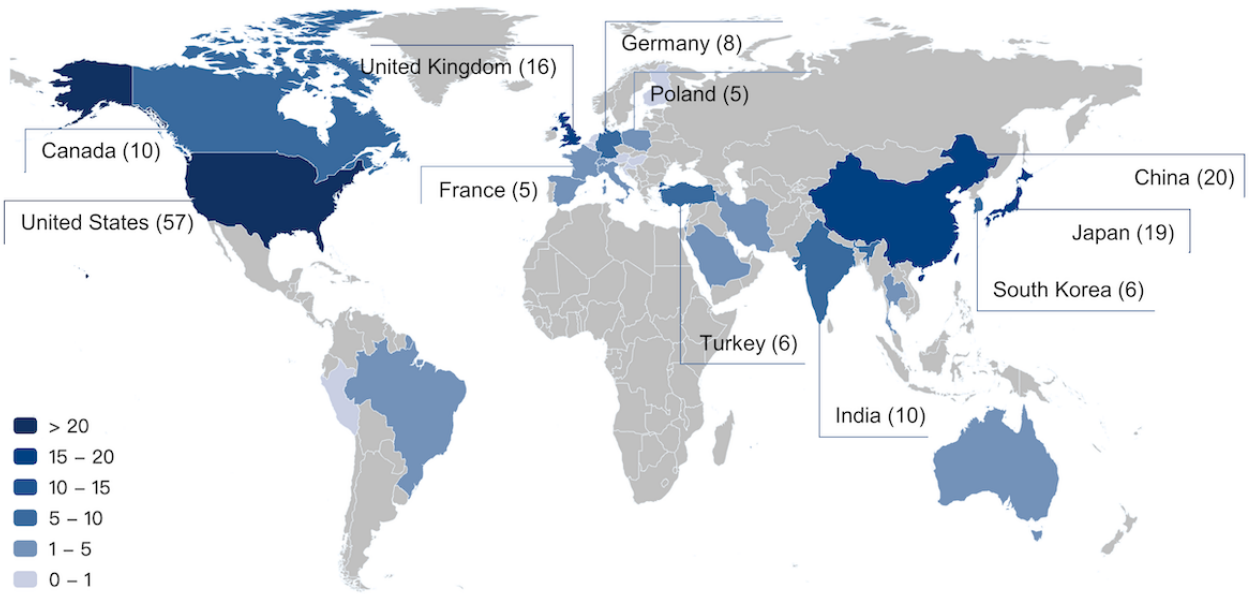

XSL•FO
RenderX

Table 2. Distribution of publications by exam language and exam year in the MedExamLLM platform.

Exam characteristics	Publications, n
Language	
English	127
Chinese	20
Japanese	19
German	5
Korean	5
Polish	5
French	3
Spanish	3
Portuguese	3
Hebrew	2
Indian	2
Italian	2
Dutch	1
Turkish	1
Peruvian	1
Year	
2009	1
2010	1
2011	1
2012	1
2013	2
2014	3
2015	4
2016	5
2017	6
2018	8
2019	11
2020	16
2021	23
2022	72
2023	35

Differences in LLMs' Capabilities Across Geographic and Linguistic Contexts

We further analyzed the pass rates of LLMs on medical exams across different countries and languages. The pass rate was defined as the ratio of exams passed by a model to the total number of evaluative entries, excluding those that did not report a pass or fail outcome. We evaluated 15 countries with more than 10 entries each. As shown in Figure 4A, the United States had the most entries, totaling 175, of which 33 showed that the model passed the exam, 64 showed that the model failed, and 78 did not indicate whether the model passed or failed, resulting in a pass rate of 34% (33/97). The numbers of entries in Japan and China were 73 and 71, respectively, and the pass rates were 36% (16/45) and 39% (21/54), respectively. The highest pass rate was in Germany, at 95% (40/42), followed by Poland, with a pass rate of 67% (32/48). Most of the entries from Canada and Italy did not indicate whether the LLMs passed or failed. For languages, we evaluated 8 languages with more than 10 entries each. As shown in Figure 4B, English had the most entries, totaling 418, of which 118 indicated that the model passed the medical exam, 102 indicated that the model failed, and 198 records did not specify whether the model passed or failed, resulting in the highest pass rate of 53.6% (118/220).

XSL•FO
RenderX

German had the highest pass rate at 75% (6/8), but it is important to note that the number of entries was not large. The pass rates for other languages were 53% (16/30) for Portuguese, 47% (14/30) for Polish, 37% (19/51) for Chinese, 36% (16/45) for Japanese, and 27% (3/11) for Korean. Spanish had a total of 32 entries, but none indicated whether the model passed the exam.

Our chi-square analysis revealed significant pass rate differences among languages. English was the predominant language for these exams; however, models demonstrated varying performance when other languages were evaluated. For instance, the pass rate difference between English and Chinese exams was statistically significant (P=.02), as was the difference between English and Japanese exams (P=.049), while the differences between English exams and exams in Portuguese (P≥.99), Polish (P=.65), Korean (P=.18), and German (P=.39) were not statistically significant. These results reveal significant differences in the capabilities of LLMs across different geographical and linguistic backgrounds.

Figure 4. Capabilities of large language models across geographic and linguistic contexts, as indicated by the pass rates on medical exams across (A) 15 countries and (B) 8 languages. NA: not applicable.

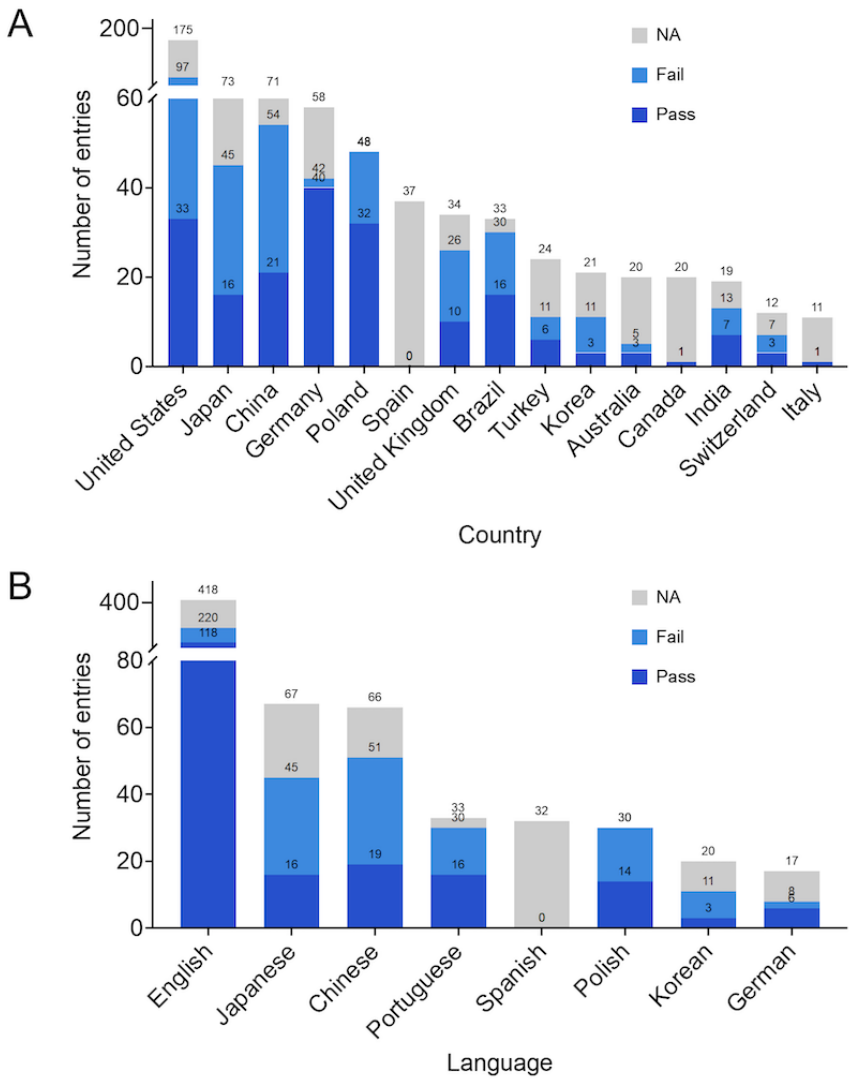

Performance and Pass Rates of LLMs

The MedExamLLM platform currently encompasses a total of 16 LLMs. As illustrated in Table 3, these models are ranked based on their frequency of use. The GPT series models are extensively used, with GPT-3.5 being the most frequently used, with a total of 273 entries, followed closely by GPT-4 with 262 entries and ChatGPT with 64 entries. Bard and Bing were used 44 and 24 times, respectively. Whether the LLM can pass the medical exams is a crucial aspect highlighted in the MedExamLLM platform. In the MedExamLLM platform, 28.2% (200/709) of entries indicated that the LLM successfully passed the medical exams, while 29.1% (206/709) indicated a failure to pass. Of the entries, 42.7% (303/709) lacked a clear statement regarding whether the model passed the medical exam. Among the top 5 most frequently used models, the GPT-4 model showcased a significantly higher pass rate than the other models. Remarkably, 50% (131/262) of the entries indicated that GPT-4 successfully passed the medical exam, while only 8.8% (23/262) indicated that it failed and 41.22% did not indicate whether it passed (Table 4). Additionally, the chi-square analysis revealed that GPT-4 outperformed other LLMs, with significant differences in pass rates compared with GPT-3.5 (P<.001), ChatGPT (P<.001), Bard (P<.001), and Bing (P<.001). These

XSL•FO
RenderX

results underscore the variability in LLM effectiveness across different model types, highlighting the need for model-specific considerations when applying LLMs to medical education.

Table 5 shows examples of national licensing examinations from different countries, as well as the performance of LLMs on those exams. For instance, the GPT-3.5 model achieved an accuracy of 54.67% on the 2021 Chinese National Medical Licensing Examination, which consists of 600 questions in Chinese [17]. The GPT-4V model achieved an accuracy of 72% on the 2023 Japanese National Medical Licensing Examination, which consists of 108 questions in Japanese [26]. ChatGPT achieved accuracies of 45.4%, 54.1%, and 61.5% for Steps 1, 2CK, and 3, respectively, on the 2022 United States Medical Licensing Examination, which consists of 350 questions in English [22].

Table 3. Frequency of the use of various large language models (LLMs) across medical exams.

LLM	Frequency of use, n
GPT[a]-3.5	273
GPT-4	262
ChatGPT	64
Bard	44
Bing	24
InstructGPT	8
GPT-3	7
GPT-4V	7
Perplexity	6
Claude	4
Gemini	4
HuggingChat	2
PaLM 2	1
Llama	1
Llama-2	1
BLOOMZ	1

[a]GPT: generative pretrained transformer.

Table 4. Pass rates of the top 5 most frequently used large language models (LLMs).

LLM	Passed, n (%)	Failed, n (%)	Not reported, n (%)
GPT[a]-3.5 (n=273)	55 (20.2)	127 (46.5)	91 (33.3)
GPT-4 (n=262)	131 (50)	23 (8.8)	108 (41.2)
ChatGPT (n=64)	4 (6)	14 (22)	46 (72)
Bard (n=44)	4 (9)	23 (52)	17 (39)
Bing (n=24)	3 (12)	6 (25)	15 (63)

[a]GPT: generative pretrained transformer.

XSL•FO
RenderX

Table 5. Examples of national medical licensing examinations from different countries and the related performance of large language models.

Exam	Country	Language	Year	Questions, n	Model	Accuracy, %	PMID
Australian Medical Council Licensing Examination	Australia	English	N/A[a]	150	GPT[b]-4	79.3	37528548
Chinese National Medical Licensing Examination	China	Chinese	2021	600	GPT-3.5	54.67	38355517
French Medical Licensing Examination	France	French	N/A	300	ChatGPT	22	37553555
German State Examination in Medicine	Germany	German	2022	252	GPT-3.5	66.7	37530052
Iranian Medical Licensing Examination	Iran	English	2023	200	GPT-4	68.5	38081765
Italian Medical Licensing Examination	Italy	Italian	N/A	300	ChatGPT	73	37553555
Japanese National Medical Licensing Examination	Japan	Japanese	2023	108	GPT-4V	72	38470459
Korean National Licensing Examination for Korean Medicine Doctors	South Korea	Korean	2022	340	GPT-4	66.18	38100393
Peruvian National Licensing Medical Examination	Peru	Peruvian	2023	180	GPT-4	86.67	37981579
Saudi Medical Licensing Examination	Saudi Arabia	English	N/A	220	GPT-4	88.6	37829968
United States Medical Licensing Examination	United States	English	2022	350	ChatGPT	45.4, 54.1, and 61.5 for USMLE[c] Steps 1, 2CK[d], and 3, respectively	36812645

[a]N/A: not applicable.

[b]GPT: generative pretrained transformer.

[c]USMLE: United States Medical Licensing Examination.

[d]CK: clinical knowledge.

Discussion

Principal Findings

The MedExamLLM is a comprehensive compilation of the latest research of LLMs on medical exams worldwide, including data from 198 medical exams across 28 countries in 15 languages from 2009 to 2023, and the evaluation performance of 16 LLMs on these medical exams. This study showed that the United States dominates in the number of medical exams and publications, with English being the primary exam language. The GPT series models, especially GPT-4, excelled on these exams, with significantly higher pass rates compared with other models. Additionally, this study revealed significant variability in the capabilities of LLMs across different geographic and linguistic contexts, as well as the strengths and weaknesses of various models on different medical exams. MedExamLLM serves as a valuable resource for educators, researchers, and developers in the fields of clinical medicine and artificial intelligence.

Expansion Beyond Previous Research

This study expands and deepens the research in several aspects. First, most previous studies focused on medical exams in a single country or a specific language, lacking a global perspective. The MedExamLLM platform fills this gap by providing comprehensive data across a wide range of countries and languages. Second, previous studies often focused only on the performance of a specific model or a few models, whereas this study systematically evaluated the performance of 16 different LLMs on various medical exams, providing more extensive and diverse evidence. Third, previous research usually focused on static exam results, while this study further compares the dynamic performance of these models across different years, countries, and languages. This expansion of the temporal and spatial dimensions enhances the depth and breadth of the study and provides an important foundation for future research and applications. Finally, this study constructs an open-source, freely accessible, and publicly available online platform that provides comprehensive performance evaluation information and evidence-based knowledge of LLMs on worldwide medical exams.

This platform not only offers a valuable data resource for researchers but also provides evidence for educators and policymakers. Through detailed performance evaluations and comparisons, educators can better understand the advantages and limitations of LLMs, enabling more effective application

XSL•FO
RenderX

of these technologies in classroom teaching. Meanwhile, researchers can use this platform to further optimize, enhance, and improve the capabilities of LLMs in medical education applications [33].

Potential of and Challenges With LLMs in Medical Education

MedExamLLM systematically compares the performance of LLMs on medical exams and reveals the considerable potential of artificial intelligence technologies to improve the quality and efficiency of medical education. The application of LLMs in the field of medical exams is expected to enhance students' exam preparation efficiency and improve their exam scores [34]. In the future, with the continuous evolution and advancements of LLM technology, its application in medical education will become more widespread. However, it is important to note that, although LLMs can provide valuable assistance, they are unlikely to replace individual roles in medical education and training [35]. These models can serve as useful tools to help teachers prepare classroom teaching materials and help students strengthen their understanding of medical knowledge. Furthermore, the use of LLMs for medical education should be combined with traditional learning methods and conducted under the guidance of qualified medical professionals [36,37].

MedExamLLM offers practical applications for educators seeking to integrate artificial intelligence into curriculum development and student assessment. The platform records the performance for a variety of LLMs and evidence across different regions, languages, and medical specialties, providing valuable insights for educators to select the model best suited to their specific context. For example, teachers can use MedExamLLM to identify models that perform optimally in their language or region, helping them tailor their instructional strategies. In a previous study [16], we illustrated how LLMs like ChatGPT effectively assist with enhancing students' academic writing quality and facilitating teachers' grading processes. This aligns with MedExamLLM's broader objective of equipping educators with reliable AI-based insights to improve both student learning outcomes and assessment efficiency.

Ethical Considerations

MedExamLLM systematically evaluates the performance of LLMs across a wide array of medical exams, showcasing the potential benefits and critical considerations of artificial intelligence integration into medical education. Although our platform highlights the promising utility of LLMs to aid with exam preparation and broaden access to medical knowledge, several ethical implications are important for stakeholders to consider, particularly regarding data privacy, model bias, and the risks associated with artificial intelligence errors in high-stakes settings.

First, data privacy is crucial given that medical exams often contain sensitive information about exam content, structure, and performance trends. Medical institutions and artificial intelligence developers must ensure that data privacy is safeguarded, especially if LLMs are trained on specific medical exam data sets. A recent study suggested that underrepresentation across languages and regions in data sets may lead to challenges with deploying LLMs in clinical [38].

Second, model bias presents a significant ethical challenge. Since LLMs learn from diverse data sets, any inherent biases can result in unequal model performance across different demographic and linguistic groups. In MedExamLLM, we observed variability in LLM pass rates across exams from different countries and languages, suggesting potential biases in how well models perform in various contexts. This variability underscores the need for ongoing efforts to improve the representativeness of training data and incorporate fairness metrics when assessing model outcomes.

Finally, the use of LLMs on medical exams involves the potential risk of artificial intelligence errors, which, in a medical context, can lead to serious consequences if misapplied or misunderstood. For instance, LLMs may incorrectly answer clinically relevant questions or provide misleading explanations, which could impact the educational experiences of students relying on these models. In addition, although studies have suggested that LLMs can generate medical exam questions, direct use also increases the risk of misinformation [39-42]. MedExamLLM's analysis of LLM accuracy and pass thresholds offers insights into the models' reliability, but this is not a replacement for human verification in medical education. Educators and institutions should use these tools as a supplement to traditional learning methods, ensuring that LLMs are applied within a structured framework that includes expert validation.

Future Road Map

The MedExamLLM platform is designed with a forward-looking approach to accommodate the evolving landscape of LLMs in medical education. Our plans for the future development of MedExamLLM include 4 aspects.

The first is regular platform updates and expansion of model evaluations. MedExamLLM will incorporate periodic updates to include performance data for newer LLMs as they become available. To address the rapid advancements in LLM technology, the platform will be updated annually, allowing for timely additions of emerging models and versions. This will provide users with the latest comparative performance data across diverse exam contexts. This initial version focuses on data from PubMed, and the platform is designed to evolve, with future updates planned to incorporate additional data sources and expand coverage. The second is user feedback. A feedback mechanism has been integrated into the platform to allow users, including educators and researchers, to suggest improvements and submit their own performance data. This submission module enables users to contribute new evaluation results from additional exams or model assessments, enhancing the platform's data completeness and scope. Feedback from users will also inform updates to ensure the platform remains user-centered and relevant to the needs of the medical education and artificial intelligence communities. The third is broadening the scope to include diverse exam formats and expanded language coverage. We plan to expand MedExamLLM's coverage by incorporating different exam formats, such as case-based questions, to provide a more comprehensive view of LLMs' capabilities with various medical evaluation formats.

XSL•FO
RenderX

Additionally, efforts will be made to include exam data sets in less commonly represented languages to improve the linguistic diversity in our platform. The fourth is practical applications and real-world impact. In the future, we will focus on evaluating LLMs' effectiveness in real-world educational settings beyond exams. As described in our previous publication [16], by exploring the utility of LLMs in simulated clinical scenarios and interactive teaching contexts, we aim to broaden the understanding of these models' capabilities and limitations in medical education.

Limitations

This study has some limitations. First, the data mainly come from publicly available peer-reviewed journal publications. However, many studies are initially published on preprint platforms, which may introduce bias in data acquisition and completeness. The exam data from some countries and languages may be incomplete or not detailed enough, thus affecting the representativeness of the results. Second, LLMs are frequently updated, with new models and versions continuously emerging, which presents a unique challenge for a platform aiming to provide up-to-date evaluations. To address this, we implemented a submission module in MedExamLLM that enables users to contribute new results for emerging LLMs. The platform is designed to undergo periodic updates, allowing for ongoing comparative analysis as models evolve. Finally, the study focuses on LLM performance on medical exams and does not explore real-world applications of these models in medical education. Although performance on exams can offer insights into some aspects of artificial intelligence proficiency, it remains inadequate for capturing the complete range of capabilities exhibited in real-world scenarios. Therefore, future research needs to evaluate the effectiveness of LLMs in educational settings to better understand their strengths and limitations.

Conclusion

We described the development and functionality of MedExamLLM, an open-source, freely accessible, and publicly available online platform that provides comprehensive performance evaluation information and evidence-based knowledge of LLMs on medical exams around the world. The MedExamLLM platform comprises data from 198 medical exams conducted in 28 countries, covering 15 languages, and spanning from 2009 to 2023. MedExamLLM reveals variations in LLM performance across different countries and languages, as well as the strengths and weaknesses of different models for various medical exams. By contributing to the growing knowledge on LLM capabilities for medical exams, MedExamLLM highlights both the potential benefits and limitations of their use in medical education. Furthermore, it provides valuable insights for the integration of artificial intelligence technologies in medical education, making it a valuable resource for educators, researchers, and developers in the fields of clinical medicine and artificial intelligence.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant number 32270690 and 32070671) and Sichuan Science and Technology Program (grant number 2024YFHZ0205).

Data Availability

The data analyzed and reported in this study are available at [31].

Authors' Contributions

HZ, RW, JC, JW, EW, JL, YZ, CZ, WF, and BS were involved in study conceptualization. HZ and RW collected and preprocessed the data. HZ analyzed the data, interpreted the results, and prepared the manuscript. HZ, RW, and BS reviewed and edited the manuscript. BS supervised the study. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 32 KB - jmir_v26i1e66114_app1.docx]

References

1. Wu T, He S, Liu J, Sun S, Liu K, Han Q, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J. Autom. Sinica 2023 May;10(5):1122-1136. [doi: 10.1109/jas.2023.123618]
2. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv Preprint posted online on July 19, 2023. [doi: 10.48550/arXiv.2307.09288]
3. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023 Jun 01;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]

XSL•FO
RenderX

4. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. NEJM AI 2024 Feb 22;1(3):1. [doi: 10.1056/aioa2300138]
5. Sahoo S, Plasek JM, Xu H, Uzuner Ö, Cohen T, Yetisgen M, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. J Am Med Inform Assoc 2024 Sep 01;31(9):2114-2124. [doi: 10.1093/jamia/ocae074] [Medline: 38657567]
6. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023 Nov 22;25(1):1 [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Rodriguez J, Alsentzer E, Bates DW. Leveraging large language models to foster equity in healthcare. J Am Med Inform Assoc 2024 Sep 01;31(9):2147-2150. [doi: 10.1093/jamia/ocae055] [Medline: 38511501]
9. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141 [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
10. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov 19;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
11. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
12. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Price T, et al. The impact of large scale licensing examinations in highly developed countries: a systematic review. BMC Med Educ 2016 Aug 19;16(1):212 [FREE Full text] [doi: 10.1186/s12909-016-0729-7] [Medline: 27543269]
13. Wang X. Experiences, challenges, and prospects of National Medical Licensing Examination in China. BMC Med Educ 2022 May 08;22(1):349 [FREE Full text] [doi: 10.1186/s12909-022-03385-9] [Medline: 35527240]
14. Holland AM, Lorenz WR, Cavanagh JC, Smart NJ, Ayuso SA, Scarola GT, et al. Comparison of medical research abstracts written by surgical trainees and senior surgeons or generated by large language models. JAMA Netw Open 2024 Aug 01;7(8):e2425373 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.25373] [Medline: 39093561]
15. Levin C, Kagan T, Rosen S, Saban M. An evaluation of the capabilities of language models and nurses in providing neonatal clinical decision support. Int J Nurs Stud 2024 Jul;155:104771. [doi: 10.1016/j.ijnurstu.2024.104771] [Medline: 38688103]
16. Li J, Zong H, Wu E, Wu R, Peng Z, Zhao J, et al. Exploring the potential of artificial intelligence to enhance the writing of english academic papers by non-native english-speaking medical students - the educational application of ChatGPT. BMC Med Educ 2024 Jul 09;24(1):736 [FREE Full text] [doi: 10.1186/s12909-024-05738-y] [Medline: 38982429]
17. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
18. Gimpel H, Ruiner C, Schoch M, Schoop M, Hall K, Eymann T, et al. Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers. University of Hohenheim. 2023. URL: https://wiso.uni-hohenheim.de/fileadmin/einrichtungen/wiso/Forschungsdekan/Papers_BESS/dp_2023-02_online.pdf [accessed 2024-12-16]
19. Fuchs K. Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? Front. Educ 2023 May 17;8:1. [doi: 10.3389/feduc.2023.1166682]
20. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
21. Hobensack M, von Gerich H, Vyas P, Withall J, Peltonen L, Block LJ, et al. A rapid review on current and potential uses of large language models in nursing. Int J Nurs Stud 2024 Jun;154:104753. [doi: 10.1016/j.ijnurstu.2024.104753] [Medline: 38560958]
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
23. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
24. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 style questions: quantitative analysis. JMIR Med Educ 2024 Jan 05;10:e51148 [FREE Full text] [doi: 10.2196/51148] [Medline: 38180782]

XSL•FO
RenderX

25. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 04;9:e46482 [FREE Full text] [doi: 10.2196/46482] [Medline: 37665620]
26. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: evaluation study. JMIR Med Educ 2024 Mar 12;10:e54393 [FREE Full text] [doi: 10.2196/54393] [Medline: 38470459]
27. Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and New Bing in an Australian medical licensing examination. Med J Aust 2023 Sep 04;219(5):237-237. [doi: 10.5694/mja2.52061] [Medline: 37528548]
28. Jang D, Yun T, Lee C, Kwon Y, Kim C. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Digit Health 2023 Dec 15;2(12):e0000416 [FREE Full text] [doi: 10.1371/journal.pdig.0000416] [Medline: 38100393]
29. Shorey S, Mattar C, Pereira TL, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. Nurse Educ Today 2024 Apr;135:106121 [FREE Full text] [doi: 10.1016/j.nedt.2024.106121] [Medline: 38340639]
30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
31. MedExamLLM. URL: http://medexamllm.bioinf.org.cn/ [accessed 2024-12-16]
32. Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024 Apr 20;281(4):2137-2143. [doi: 10.1007/s00405-023-08381-3] [Medline: 38117307]
33. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
34. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience 2024 May 17;27(5):109713 [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
35. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inform Assoc 2024 Feb 16;31(3):776-783. [doi: 10.1093/jamia/ocad252] [Medline: 38269644]
36. Lawson McLean A. Constructing knowledge: the role of AI in medical learning. J Am Med Inform Assoc 2024 Aug 01;31(8):1797-1798. [doi: 10.1093/jamia/ocae124] [Medline: 38812088]
37. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. J Med Internet Res 2024 Aug 01;26:e60083 [FREE Full text] [doi: 10.2196/60083] [Medline: 38971715]
38. Wu J, Liu X, Li M, Li W, Su Z, Lin S, et al. Clinical text datasets for medical artificial intelligence and large language models — a systematic review. NEJM AI 2024 May 23;1(6):1. [doi: 10.1056/aira2400012]
39. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J 2024 Oct 18;100(1189):858-865. [doi: 10.1093/postmj/qgae065] [Medline: 38840505]
40. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med 2023 Dec 28;99(5):508-512. [doi: 10.1097/acm.0000000000005626]
41. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024 Mar 29;24(1):354 [FREE Full text] [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
42. Hang CN, Wei Tan C, Yu P. MCQGen: a large language model-driven MCQ generator for personalized learning. IEEE Access 2024;12:102261-102273. [doi: 10.1109/access.2024.3420709]

Abbreviations

GPT: generative pretrained transformer
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

XSL•FO
RenderX

Edited by A Schwartz; submitted 04.09.24; peer-reviewed by Z Guan, CN Hang; comments to author 31.10.24; revised version received 06.11.24; accepted 10.12.24; published 27.12.24.

Please cite as:
Zong H, Wu R, Cha J, Wang J, Wu E, Li J, Zhou Y, Zhang C, Feng W, Shen B
Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis
J Med Internet Res 2024;26:e66114
URL: https://www.jmir.org/2024/1/e66114
doi:10.2196/66114
PMID:39729356

©Hui Zong, Rongrong Wu, Jiaxue Cha, Jiao Wang, Erman Wu, Jiakun Li, Yi Zhou, Chi Zhang, Weizhe Feng, Bairong Shen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration

Ziwei Zhao[1*], MD; Weiyi Zhang[1*], MS; Xiaolan Chen[1], MD; Fan Song[1], MD; James Gunasegaram[2], MD; Wenyong Huang[3], MD, PhD; Danli Shi[1,4*], MD, PhD; Mingguang He[1,4,5*], MD, PhD; Na Liu[6*], MD

[1]School of Optometry, The Hong Kong Polytechnic University, Hong Kong, China
[2]Monash University, Victoria, Australia
[3]Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
[4]Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Hong Kong, China
[5]Centre for Eye and Vision Research (CEVR), Hong Kong, China
[6]Guangzhou Cadre and Talent Health Management Center, Guangzhou, China
[*]these authors contributed equally

Corresponding Author:
Na Liu, MD
Guangzhou Cadre and Talent Health Management Center
No. 109 Changling Road
Huangpu District
Guangzhou, 510700
China
Phone: 86 18701985445
Email: 1256695904@qq.com

Abstract

Background: Large language models have shown remarkable efficacy in various medical research and clinical applications. However, their skills in medical image recognition and subsequent report generation or question answering (QA) remain limited.

Objective: We aim to finetune a multimodal, transformer-based model for generating medical reports from slit lamp images and develop a QA system using Llama2. We term this entire process slit lamp–GPT.

Methods: Our research used a dataset of 25,051 slit lamp images from 3409 participants, paired with their corresponding physician-created medical reports. We used these data, split into training, validation, and test sets, to finetune the Bootstrapping Language-Image Pre-training framework toward report generation. The generated text reports and human-posed questions were then input into Llama2 for subsequent QA. We evaluated performance using qualitative metrics (including BLEU [bilingual evaluation understudy], CIDEr [consensus-based image description evaluation], ROUGE-L [Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence], SPICE [Semantic Propositional Image Caption Evaluation], accuracy, sensitivity, specificity, precision, and F_1-score) and the subjective assessments of two experienced ophthalmologists on a 1-3 scale (1 referring to high quality).

Results: We identified 50 conditions related to diseases or postoperative complications through keyword matching in initial reports. The refined slit lamp–GPT model demonstrated BLEU scores (1-4) of 0.67, 0.66, 0.65, and 0.65, respectively, with a CIDEr score of 3.24, a ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score of 0.61, and a Semantic Propositional Image Caption Evaluation score of 0.37. The most frequently identified conditions were cataracts (22.95%), age-related cataracts (22.03%), and conjunctival concretion (13.13%). Disease classification metrics demonstrated an overall accuracy of 0.82 and an F_1-score of 0.64, with high accuracies (≥0.9) observed for intraocular lens, conjunctivitis, and chronic conjunctivitis, and high F_1-scores (≥0.9) observed for cataract and age-related cataract. For both report generation and QA components, the two evaluating ophthalmologists reached substantial agreement, with κ scores between 0.71 and 0.84. In assessing 100 generated reports, they awarded scores of 1.36 for both completeness and correctness; 64% (64/100) were considered "entirely good," and 93% (93/100) were "acceptable." In the evaluation of 300 generated answers to questions, the scores were 1.33 for completeness, 1.14 for correctness, and 1.15 for possible harm, with 66.3% (199/300) rated as "entirely good" and 91.3% (274/300) as "acceptable."

XSL•FO
RenderX

Conclusions: This study introduces the slit lamp–GPT model for report generation and subsequent QA, highlighting the potential of large language models to assist ophthalmologists and patients.

(J Med Internet Res 2024;26:e54047) doi:10.2196/54047

KEYWORDS

large language model; slit lamp; medical report generation; question answering

Introduction

The slit lamp, a cornerstone in ophthalmology, allows for detailed examination of the eye's anterior segment [1]. Using an illuminated, narrow beam, this noninvasive method facilitates the evaluation of abnormalities by depth and size. While instrumental in diagnosing common eye diseases such as keratitis, conjunctivitis, conjunctival concretions, and cataracts, interpreting slit lamp results can be challenging for primary care physicians due to the need for specialized training. This can result in overlooked abnormalities or misdiagnosis. Furthermore, ophthalmologists are tasked with interpreting, documenting, and effectively communicating these results to patients, a time and effort-intensive process. The scarcity of experienced ophthalmologists, particularly in rural areas, further exacerbates the situation [2].

Artificial intelligence (AI) and large language models (LLMs) have made significant strides in the medical field, enhancing the capabilities of health care professionals in interpreting, and analyzing medical imagery. For instance, AI has been instrumental in advancing the analysis of x-rays [3], magnetic resonance images [4], ultrasounds [5], and dermatological images [6]. Generative pretrained transformers (GPT) models such as ChatGPT [7] and Llama2 [8], have showcased remarkable capabilities in problem-solving scenarios across a spectrum of medical applications. These AI models are instrumental in streamlining clinical documentation [9], refining patient communication [10], aiding administrative tasks [11], enriching textual data [12], and bolstering evidence-based decision-making [13]. Their versatility extends to comprehensive patient assessments [14], precise disease diagnostics [15], informed treatment proposals [16], meticulous medical writing [17], innovative teaching methodologies [18], and robust question answering (QA) systems [19], embodying a multifaceted impact on the health care industry.

Deep learning strategies currently used to transform images into high-quality features include convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer networks, and their variants such as long short-term memory (LSTM) and gated recurrent units (GRUs). CNNs are often combined with other networks such as RNNs to generate text [20]. RNNs and their variants, recognized for their prowess in handling sequential data, account for element dependencies within sequences. Despite their effectiveness, RNNs face challenges with extended sequences and potential gradient issues which are mitigated by long short-term memories and GRUs through a gate mechanism. Transformer networks, proposed in 2017, use self-attention mechanisms to manage long sequences and parallel computations, thus boasting swift training speed at the cost of substantial computational resources [21]. Bootstrapping Language-Image Pre-training (BLIP), a hybrid approach leveraging transformer networks' architecture and amalgamating natural language processing and computer vision, enhances model performance via pretraining. BLIP's principal strength lies in its multimodal capacity to concurrently handle image and text data, allowing it to excel in specific tasks such as image description generation.

In the specific context of slit lamp imaging augmented with AI, research has primarily concentrated on individual disease detection and grading, such as in the case of cataracts [22,23], pterygium [24], and infectious keratitis [25]. However, there is a noticeable lack of a unified system that uses slit lamp images for the generation of systematic anterior segment reports and QA. While the advent of OpenAI's GPT-4V offered the possibility of image-based AI medical dialogue, its direct clinical application has been limited by inaccuracies and the generation of unreliable information, which was termed "hallucinations" [26,27]. Additionally, due to its closed-source nature, there is a constraint on the fine-tuning ability, which is paramount for medical applications. In response to this, our study has used Llama2, an open-source model, to harness the anticipated benefits of a specialized LLM tool that ensures enhanced control and reliability in the subsequent QA scenarios. Based on our experience in ophthalmic QA tasks and LLMs, including fundus fluorescein angiography and indocyanine green angiography QA [12,28,29] we aim to extend these methodologies to slit lamp imaging by developing a novel slit lamp–GPT system, using BLIP and LLMs specifically tailored for ophthalmology, with dual objectives: to generate reports and to facilitate QA.

Methods

Overview

The flow of our study is outlined in Figure 1.

XSL•FO
RenderX

Figure 1. Flow diagram of this study.

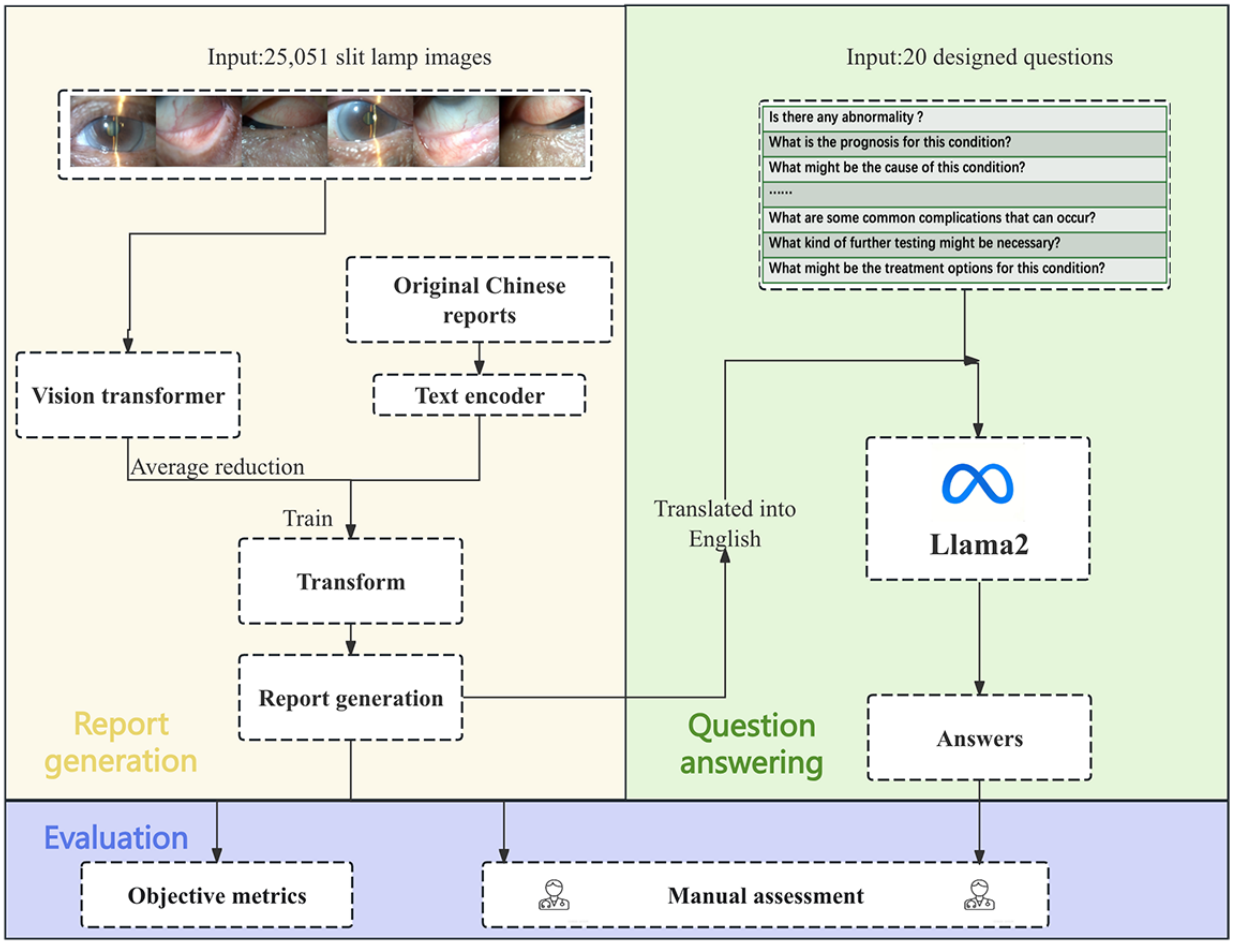

Dataset

We collected data from a Chinese physical examination center for this retrospective study, which included both essential clinical information and annual slit lamp images. We included slit lamp photographs with corresponding medical reports, excluding any of inadequate quality. This study used data collected from a previous study [30], all participants' information was deidentified per the Declaration of Helsinki's guidelines. All slit lamp images, captured via a Haag-Streit BQ-900 at a 2048×1536-pixel resolution, included at least 4 images per participant showcasing the pupil, upper eyelid, and lower eyelid. Initial reports, written by ophthalmologists in Chinese, contained disease diagnoses, recommendations, or detailed descriptions of ocular signs. A subset of representative reports from the dataset was selected for translation into English to form a bilingual dataset.

Model Construction

Similar to other studies [28,29], we initially trained and tested the BLIP [28] network for report generation. Subsequently, the generated reports from the test set were input into Llama2 for QA validation, further evaluating the quality and practicality of the reports.

During the report generation phase, we used the BLIP framework, a multimodal transformer model skilled at aligning visual interpretation with text generation. The model filtered out noisy data during training and generated slit lamp reports from paired images and text inputs. Our design incorporated a vision transformer [31] and BERT [32] as the image and language encoder and decoder, respectively. The vision transformer converts an image into encoded patch sequences, while BERT, trained on extensive unlabeled text data, enables deep contextualized representation learning. The pretrained BLIP model was fine-tuned using slit lamp images and associated reports, with each case providing at least four images during training, resized to 224×224 pixels. We applied the AdamW optimizer (the University of Freiburg), using an initial learning rate of 0.00002, a weight decay of 0.05, and a cosine learning rate schedule, across 50 epochs on one NVIDIA Tesla V100 GPU (NVIDIA Corp). The model with the highest BLEU1 (bilingual evaluation understudy) score (detailed in the performance evaluation part) on the validation set was selected for testing.

For the question and answering phase, we created a question set related to slit lamp examination and reporting based on prior studies [33] and our clinical expertise. These questions, along with the corresponding reports, were seamlessly input into the Llama2 model. This integration allowed for QA without the need for fine-tuning, while enhancing the interpretation of the generated reports. The process involved instructing the model

XSL•FO
RenderX

using a specific prompt: "Answer based on: [slit lamp report content here]."

Performance Evaluation

We used both language-based and disease classification metrics for quantitative evaluations of report quality, supplemented by manual assessments for report generation and QA.

For language-based metrics, we used BLEU [34], CIDEr [35], ROUGE-L [36]), and Semantic Propositional Image Caption Evaluation [37], each with its strengths. However, traditional language metrics may be less dependable for medical conditions due to the infrequent occurrence of disease-related keywords in reports. To address this, we introduced a classification evaluation procedure that used a manually curated dictionary to identify disease-related conditions or postoperative statuses from both original and generated reports. Disease classification metrics, such as specificity, accuracy, precision, sensitivity, and the F_1-score, provided a comprehensive performance review of the model.

Considering the complexity of medical terminology and the potential harm of inaccurate reporting, manual assessment remains crucial. For report generation, 100 test set cases were randomly selected and independently evaluated by 2 ophthalmologists (ZZ and FS) using a 3-point scale, focusing on "completeness" (how well the generated reports matched the ground truth conditions) and "correctness" (the accuracy of diagnosis and condition descriptions). Scores ranged from 1 (excellent) to 3 (poor), with 2 representing an acceptable rating. The final score was the average of the scores from the 2 evaluators. For QA, 20 prepared human-posed questions and the translated report were put into Llama2 to generate answers, which were evaluated based on "completeness," "correctness," and "possible harm." Scores ranged from 1 (recommendable to patients) to 3 (not recommendable for patients), with 2 indicating that minor adjustments could make the answer suitable for recommendation. The average score was also used as the final score. For detailed scoring criteria in these 2 sections, refer to Table S1 in Multimedia Appendix 1.

Ethical Considerations

This study used data collected from a previous study [30]. All patient data were anonymized and de-identified following the Declaration of Helsinki. Individual consent was waived due to the retrospective nature and the thorough anonymization process of the study. The Institutional Review Board of the Hong Kong Polytechnic University approved the study (HSEARS20240301004).

Results

Data

Our final dataset includes 25,051 slit-lamp images and 3409 reports. Most images (12,496, 49.89%) focus on the cornea, with 32.74% (n=8202) on the upper eyelid and 17.38% (n=4353) on the lower eyelid. The median age of participants is 65, with an IQR of 60 to 72 years, and the majority (2009/3409, 58.93%) are male. The demographics and image types are similar across all sets.

The distribution of images across years is as follows: 1257 (5.02%) from 2013, 12,206 (48.72%) from 2015, and 11,588 (46.26%) from 2016. The 2013 and 2015 images form the training set, while the 2016 images are partitioned evenly into validation and testing sets. There were no significant differences in demographic characteristics and positioning type between these datasets. Table 1 provides a comprehensive overview of the dataset characteristics.

Table 1. Slit lamp images: dataset characteristics.

	Total	Train	Validation	Test	*P* value
Participants					
Number	3409	1846	781	782	
Age, median (Q1[a], Q3[b])	65.46 (60.52, 72.47)	65.31 (60.12, 72.13)	62.04 (59.58, 65.9)	71.04 (65.47, 77.03)	<.001
Sex, n (%)					.002
Male	2009 (58.93)	1101 (59.64)	420 (53.78)	488 (62.4)	
Female	1400 (41.07)	745 (40.36)	361 (46.22)	294 (37.6)	
Slit lamp images					
Number	25,051	13,463	5987	5601	
Position, n (%)					.002
Upper eyelid	8202 (32.74)	4315 (32.05)	2046 (34.2)	1841 (32.9)	
Lower eyelid	4353 (17.38)	2423 (18)	951 (15.9)	979 (17.5)	
Cornea	12,496 (49.89)	6725 (49.95)	2990 (49.9)	2781 (49.7)	

[a]Q1: first quartile.

[b]Q3: third quartile.

We used a custom dictionary to extract diagnoses and physical signs by keyword matching from the Chinese reports. We identified 50 conditions, including age-related cataracts (478/2170, 22.03%), cataracts (498/2170, 22.95%), conjunctival

XSL•FO
RenderX

concretion (285/2170, 13.13%), after intraocular lens implantation (151/2170, 6.96%), pterygium (144/2170, 6.64%), conjunctivitis (97/2170, 4.47%), chronic conjunctivitis (93/2170, 4.29%), and other eye conditions with lower proportions. This led to 1377 Chinese reports primarily featuring diagnostic terms or descriptions of ocular signs.

Quantitative Model Performance

Language-based metrics are provided in Table 2, with BLEU (1-4) scores (0.67, 0.66, 0.65, and 0.65) indicating good lexical accuracy and a ROUGE-L score of 0.61 highlighting effective content retention. The CIDEr score of 3.24 reflects its ability to align closely with human judgment on sentence quality, while a SPICE score of 0.37 demonstrates moderate success in capturing complex semantic relationships. For disease classification metrics (see Table 3), our model achieved a weighted accuracy of 0.82 and a weighted F_1-score of 0.64. However, performance varied across diseases. It was highly accurate (≥0.9) for conditions of intraocular lens, conjunctivitis, and chronic conjunctivitis, and had high F_1-scores (≥0.9) for cataracts and age-related cataracts. The model demonstrated excellent accuracy for positive cases of cataracts and age-related cataracts. Despite high accuracy, specificity, and precision for postoperative intraocular lens implantation, sensitivity was relatively low: a clinically acceptable trade-off. However, for conjunctival concretions, conjunctivitis, and chronic conjunctivitis, the model's overall predictive capacity fell short.

Table 2. Language-based metrics of report generation in the test set (5601 images from 782 participants).

BLEU_1[a]	BLEU_2[a]	BLEU_3[a]	BLEU_4[a]	CIDEr[b]	ROUGE[c]	SPICE[d]
0.67	0.66	0.65	0.65	3.24	0.61	0.37

[a]BLEU: bilingual evaluation understudy.

[b]CIDEr: consensus-based image description evaluation.

[c]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[d]SPICE: Semantic Propositional Image Caption Evaluation.

Table 3. Disease classification metrics of report generation in the test set.

Condition	Specificity	Accuracy	Precision	Sensitivity	F_1-score
Age-related cataract	0.6	0.79	0.9	0.8	0.84
Cataract	0.58	0.78	0.9	0.79	0.84
After intraocular lens implantation	0.93	0.94	0.96	0.48	0.64
Conjunctival concretion	0.83	0.7	0.37	0.44	0.4
Chronic conjunctivitis	0.92	0.9	0.34	0.15	0.2
Conjunctivitis	0.92	0.9	0.34	0.14	0.2

Qualitative Model Performance

Overview

The score distribution is depicted in Figure S1 in Multimedia Appendix 2.

Report Generation

Two ophthalmologists scored the model highly for completeness (mean 1.36, SD 0.61, κ=0.84) and correctness (mean 1.36, SD 0.59, κ=0.72). Reports that received a score of 1 for both completeness and correctness were defined as entirely good and constituted 64% (64/100) of the evaluated reports. Reports that scored either 1 or 2 by both reviewers for both completeness and correctness were deemed acceptable, representing 93% (93/100) of the reports. These scores primarily corresponded to reports detailing specific conditions such as cataracts, age-related cataracts, and negative findings. However, 7% (7/100) of reports scored a 3, indicating deficiencies.

We discovered that lower scores were linked to issues such as limited sample sizes for specific diseases, difficulties in clearly identifying lesions, and challenges in interpreting diseases or signs from images due to the unique aspects of slit lamp photography. These complications were common in conditions such as xanthomas, trichiasis, after-glaucoma surgery, lagophthalmos, and some small conjunctival concretions. Additionally, images not focused on the cornea made it difficult to detect corneal lesions.

Through our hands-on evaluation, we noticed that the model sometimes added diagnoses that were not in the original reports but were still acceptable based on the images. For example, it sometimes diagnosed mild cataracts even when the images did not show apparent lens abnormalities. We considered these decisions acceptable when considering the challenge faced by an ophthalmologist in making a precise distinction based solely on images.

About QA

Our constructed questionnaires included 20 items, addressing a breadth of topics such as diagnosis, pathologic localization, severity grading, visual impairment, prognosis, associated complications, therapeutic recommendations, suggested further examinations, preventive advice, and scientific education pertinent to slit lamp examination (Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

We selectively curated 15 representative English reports on conditions including cataracts, conjunctival concretions, conjunctivitis, postintraocular lens implantation, and pterygium, as well as their mixed states. Each report contributed 20 questions, culminating in a total of 300 questions.

Our model scored well on completeness (1.33, κ=0.84), correctness (1.14, κ=0.71), and possible harm (1.15, κ=0.82). Similarly, QA responses that scored a 1 in completeness, correctness, and possible harm were defined as entirely good, representing 66.3% (199/300) of the QA responses. Responses scoring either 1 or 2 across these categories were considered acceptable, comprising 91.3% (274/300). Less than 9% (26/300) of the 300 questions scored a 3 in any category. These were typically related to reports focusing more on physical signs than diagnoses or conditions and statements about binocular intraocular lenses. Figure S2 in Multimedia Appendix 3 provides examples of generated answers with different scores.

Discussion

Principal Findings

Our study introduces a novel method for analyzing slit lamp images through the integration of a multimodal transformer with an LLM. This approach has enabled the accurate identification of common anterior segment eye diseases and supports a QA system that directly addresses symptoms, diagnosis, and treatment options, as illustrated in Figure 2.

Figure 2. Demonstration of the question-answering system. (A) Input image, ground truth, and model prediction. (B) Question answering. Blue highlight: corresponds to accurate diagnosis matches. Yellow highlight: supplementary predicted information (not in the manual report but correct).

LLMs represent a breakthrough in AI with large knowledge bases and strong logical reasoning abilities. They have exhibited efficacy across various natural language processing tasks, including text generation, summarization, translation, and QA. However, in the medical realm, the quality of these answers warrants further scrutiny. Previous research has shown mixed

XSL•FO
RenderX

results for the ability of LLMs to pass ophthalmology examinations. The study of Kung et al [38] indicates that ChatGPT can pass the United States Medical Licensing Examination without any specialized training or reinforcement. However, Thirunavukarasu's [39] attempt to assess ChatGPT's proficiency in the FRCOphth (Fellowship of the Royal College of Ophthalmologists) examination showed subpar performance, thereby underscoring the inability of LLMs to replace physicians in highly specialized fields. Conversely, in advising patients about symptoms or ongoing conditions—tasks less demanding of expertise—ChatGPT seems to demonstrate competence. Many patients turn to the internet for self-diagnosis before consulting a health care professional [40]. The use of LLMs for medical consultations can increase patient independence and potentially aid in accurate diagnosis. The release of GPT4V represents an innovative leap in the realm of LLM integration with computer vision, with promising prospects for extensive application in the medical field. Wu et al [26] assessed images from eight modalities across 17 human body systems and concluded that while GPT4V excels at identifying image modalities and anatomical structures, it encounters significant challenges in disease diagnosis and comprehensive report generation. In another study, we used a similar 1-3 evaluation scale to assess GPT4V's performance on ophthalmology-related tasks, including image interpretation and QA [27]. The model performed best in analyzing slit lamp images; however, it only reached 42% (42/100) in accuracy, 38.5% (34.7/90) in usability, and 68.5% (61.7/90) in safety of the responses. These results are significantly lower than the "entirely good" rates we reported previously—64% (64/100) for report generation and 66.3% (199/300) for QA. This discrepancy underscores the need for models tailored to ophthalmology to ensure high-quality outcomes. To address this gap, we implemented an experimental model, slit lamp–GPT, harnessing the BLIP and Llama2 frameworks. This initiative represents merely the first step in a broader journey toward refining AI applications in ophthalmology.

The model demonstrated proficiency in identifying and reporting common anterior segment eye diseases within our dataset. However, its performance on rare conditions highlighted a critical area for improvement, suggesting that its effectiveness is closely tied to the diversity and representation of conditions in the training data. Per report generation, suboptimal performance was linked to specific diseases such as trichiasis, postglaucoma surgery complications, and corneal pathologies. Given our dataset's origin in routine health examination data, these conditions were underrepresented, likely contributing to the poor performance. Another hypothesis considers the dynamic nature of slit lamp examinations in clinical settings, where ophthalmologists manually focus to obtain the best diagnostic view, a process not fully captured by static images. Instances of misdiagnosed keratitis, where images did not focus precisely on the cornea, support this assumption. Integrating our model with a broader spectrum of ophthalmic imaging techniques—such as indocyanine green angiography, fundus fluorescein angiography, ocular ultrasound, optical coherence tomography, and fundus photography—may enhance diagnostic alignment with actual clinical observations and further improve overall performance.

The current results suggest potential applicability in cataract screening, particularly in regions with a shortage of ophthalmologists. Previous studies have primarily focused on applying deep learning to the diagnosis and grading of cataracts, fundamentally using classification models. In contrast, our model is a natural language processing system capable of generating free-text reports. It not only provides descriptive insights but also achieves cataract classification accuracy similar to existing models [41,42]. Beyond this, our model could function as an educational tool for patients. In bustling eye clinics, patients may lack sufficient time to fully comprehend their examination reports and medical conditions. As demonstrated in this study, the slit lamp–GPT can provide patients with basic clinical explanations and recommendations concerning causes, abnormalities, treatment, and follow-up, indicating its potential to reduce medical consultation expenditure and bolster the use of remote health care services.

The manual evaluation suggests that slit lamp–GPT exhibits a promising capacity to assist participants with minimal risk. During the QA stage, 89.3% (268/300) of the responses were deemed completely harmless, surpassing the performance of GPT4V. However, the potential risks of using LLMs are yet to be thoroughly understood. A common problem with LLMs is that they sometimes generate inaccuracies and false statements, which are often referred to as "hallucinations" in the field [43]. These incorrect assertions can appear to be true, which could harm patients. This was reflected in our study, where the model sometimes created content. For example, the Llama2 model wrongly identified a binocular intraocular lens as a disease instead of a postoperative condition, creating the nonexistent "binocular intraocular lens syndrome." This led to poor scores on the related 20 questions, highlighting the need for specialized fine-tuned LLM and knowledge-based generation [44]. Nonetheless, it is important to recognize that LLMs should serve as adjuncts or supplements in the clinical diagnosis and treatment process, not as fully trusted entities devoid of physician oversight. As LLM technology evolves, it is incumbent on stakeholders to collaboratively establish best practice standards to ensure patient safety.

Limitations

This study has a few limitations. First, the dataset used is skewed, coming mainly from routine health checks of healthy people. The small sample size for certain diseases might affect the effectiveness of classification. Using datasets from high-quality outpatient clinics could lead to better results. Second, as with other language models, our model sometimes produces repetitive text, and the accuracy of the responses it generates can be inconsistent. At times, the model's answers show logical errors. For instance, it diagnosed both a postintraocular lens implantation status and a senile cataract in the same eye. These issues might be addressed by incorporating expert knowledge and fine-tuning LLMs. There are also notable concerns about bias, as a single mistake in report generation can lead to multiple errors during the question-and-answer process. This highlights the need for further improvements to increase the accuracy and completeness of report generation. Lastly, creating a standardized manual evaluation process for these types of models is challenging [45,46]. This study was

XSL•FO
RenderX

limited to slit lamp anterior segment images, indicating a need for future research to include diverse datasets. This will help evaluate the model's applicability across various types of imaging.

Conclusion

This research underscores the effectiveness and potential of using LLMs for slit lamp image report generation and QA tasks, showcasing their viability in ophthalmic medical image analysis.

Acknowledgments

This study was supported by the Global STEM Professorship Scheme (P0046113) and Henry G. Leong Endowed Professorship in Elderly Vision Health. The sponsor or funding organization had no role in the design or conduct of this research. We thank the InnoHK Hong Kong special administrative region government for providing valuable supports.

Authors' Contributions

NL is the primary corresponding author of this article, while MH and DS are co-corresponding author. DS and MH conceived this study. NL provided data. DS and WZ built the deep learning model. DS and ZZ did the literature search and analyzed the data. DS, ZZ, XC, FS, and MH contributed to key data interpretation. ZZ and FS did the manual evaluation. ZZ wrote this paper. JG edited and reviewed this paper. All authors commented and critically revised this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables with data.
[DOCX File , 23 KB - jmir_v26i1e54047_app1.docx]

Multimedia Appendix 2
Distribution of scores in the manual assessment of report generation and question-answering. RG: report generation; QA: question answering.
[PNG File , 78 KB - jmir_v26i1e54047_app2.png]

Multimedia Appendix 3
Examples of generated answers with different scores. Red bold text indicates incorrect information.
[PNG File , 918 KB - jmir_v26i1e54047_app3.png]

References

1. Doggart JH. SLIT-LAMP Microscopy. Br J Ophthalmol 1948;32(4):232-247 [FREE Full text] [doi: 10.1136/bjo.32.4.232] [Medline: 18170442]
2. Magyezi J, Arunga S. Eye care where there are no ophthalmologists: the Uganda experience. Community Eye Health 2020;33(110):48-50 [FREE Full text] [Medline: 34007108]
3. Niehoff JH, Kalaitzidis J, Kroeger JR, Schoenbeck D, Borggrefe J, Michael AE. Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays. Sci Rep 2023;13(1):3680 [FREE Full text] [doi: 10.1038/s41598-023-30521-2] [Medline: 36872333]
4. Sheth D, Giger ML. Artificial intelligence in the interpretation of breast cancer on MRI. J Magn Reson Imaging 2020;51(5):1310-1324. [doi: 10.1002/jmri.26878] [Medline: 31343790]
5. Shen Y, Chen L, Yue W, Xu H. Artificial intelligence in ultrasound. Eur J Radiol 2021;139:109717. [doi: 10.1016/j.ejrad.2021.109717] [Medline: 33962110]
6. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26(6):900-908. [doi: 10.1038/s41591-020-0842-3] [Medline: 32424212]
7. Introducing ChatGPT. URL: https://openai.com/blog/chatgpt [accessed 2024-10-19]
8. Touvron H, Martin L, Stone K. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 2023:1-77 Preprint published on July 18, 2023.
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
11. Kahambing J. ChatGPT, public health communication and 'intelligent patient companionship'. J Public Health (Oxf) 2023;45(3):e590. [doi: 10.1093/pubmed/fdad028] [Medline: 37036209]

XSL•FO
RenderX

12. Chen X, Xu P, Li Y, Zhang W, Song F, He M, et al. ChatFFA: an ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. iScience 2024;27(7):110021 [FREE Full text] [doi: 10.1016/j.isci.2024.110021] [Medline: 39055931]
13. Dai H, Liu Z, Liao W, Huang X, Cao Y, Wu Z, et al. AugGPT: leveraging ChatGPT for text data augmentation. arXiv:2302.13007 2023:1-12 Preprint published on February 25, 2023.
14. Ebrahimian S, Homayounieh F, Rockenbach MABC, Putha P, Raj T, Dayan I, et al. Artificial intelligence matches subjective severity assessment of pneumonia for prediction of patient outcome and need for mechanical ventilation: a cohort study. Sci Rep 2021;11(1):858 [FREE Full text] [doi: 10.1038/s41598-020-79470-0] [Medline: 33441578]
15. Liang H, Guo Y, Chen X, Ang K, He Y, Jiang N, et al. Artificial intelligence for stepwise diagnosis and monitoring of COVID-19. Eur Radiol 2022;32(4):2235-2245 [FREE Full text] [doi: 10.1007/s00330-021-08334-6] [Medline: 34988656]
16. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Badia SBBI. A human-AI collaborative approach for clinical decision making on rehabilitation assessment. 2021 Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 07; Yokohama, Japan p. 1-14.
17. Biswas S. ChatGPT and the future of medical writing. Radiology 2023;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
18. Masters K. Artificial intelligence in medical education. Med Teach 2019;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
19. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv:2305.09617 2023:1-30 Preprint published on May 16, 2023.
20. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: NeurIPS Proceedings. 2012 Presented at: 26th Annual Conference on Neural Information Processing Systems 2012; 2012-12-3; Lake Tahoe, Nevada, United States p. 1-9 URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NeurIPS Proceedings. 2017 Presented at: the 31st International Conference on Neural Information Processing Systems; 2017 December 04; Red Hook, NY, United States p. 6000-6010 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
22. Son KY, Ko J, Kim E, Lee SY, Kim M, Han J, et al. Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs: model development and validation study. Ophthalmol Sci 2022;2(2):100147 [FREE Full text] [doi: 10.1016/j.xops.2022.100147] [Medline: 36249697]
23. Lu Q, Wei L, He W, Zhang K, Wang J, Zhang Y, et al. Lens opacities classification system III-based artificial intelligence program for automatic cataract grading. J Cataract Refract Surg 2022;48(5):528-534. [doi: 10.1097/j.jcrs.0000000000000790] [Medline: 34433780]
24. Fang X, Deshmukh M, Chee ML, Soh Z, Teo ZL, Thakur S, et al. Deep learning algorithms for automatic detection of pterygium using anterior segment photographs from slit-lamp and hand-held cameras. Br J Ophthalmol 2022;106(12):1642-1647 [FREE Full text] [doi: 10.1136/bjophthalmol-2021-318866] [Medline: 34244208]
25. Soleimani M, Esmaili K, Rahdar A, Aminizadeh M, Cheraqpour K, Tabatabaei SA, et al. From the diagnosis of infectious keratitis to discriminating fungal subtypes; a deep learning-based study. Sci Rep 2023;13(1):22200 [FREE Full text] [doi: 10.1038/s41598-023-49635-8] [Medline: 38097753]
26. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V (ision) serve medical applications? case studies on GPT-4V for multimodal medical diagnosis. arXiv:2310.09909 2023:1-178 Preprint published October 15 2023.
27. Xu P, Chen X, Zhao Z, Shi D. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. Br J Ophthalmol 2024;108(10):1384-1389. [doi: 10.1136/bjo-2023-325054] [Medline: 38789133]
28. Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, Shi D, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit Med 2024;7(1):111 [FREE Full text] [doi: 10.1038/s41746-024-01101-z] [Medline: 38702471]
29. Chen X, Zhang W, Zhao Z, Xu P, Zheng Y, Shi D, et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. Br J Ophthalmol 2024;108(10):1450-1456. [doi: 10.1136/bjo-2023-324446] [Medline: 38508675]
30. Hu Y, Niu Y, Wang D, Wang Y, Holden BA, He M. The association of longitudinal trend of fasting plasma glucose with retinal microvasculature in people without established diabetes. Invest Ophthalmol Vis Sci 2015;56(2):842-848. [doi: 10.1167/iovs.14-15943] [Medline: 25613941]
31. Li J, Li D, Xiong C, Hoi S. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. 2022 Presented at: Proceedings of the 39th International Conference on Machine Learning; 2022 January 28; Baltimore, Maryland, USA.
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Presented October 11, 2018 2018:1-16.

XSL•FO
RenderX

33. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 2023;7(10):862-868. [doi: 10.1016/j.oret.2023.05.022] [Medline: 37277096]
34. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. 2002 Presented at: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2002 July 06; United States.
35. Vedantam R, Lawrence ZC, Parikh D. CIDEr: Consensus-based image description evaluation. 2015 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 June 07-12; Boston, MA, USA.
36. Lin CY. Rouge: A package for automatic evaluation summaries. 2004 Presented at: Proceedings of the Workshop on Text Summarization Branches Out; 2004 July 25–26; Barcelona, Spain p. 74-81.
37. Anderson P, Fernando B, Johnson M, Gould S. SPICE: semantic propositional image caption evaluation. 2016 Presented at: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V; 2016 October 11-14; Amsterdam, The Netherlands.
38. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
39. Thirunavukarasu AJ. ChatGPT cannot pass FRCOphth examinations: implications for ophthalmology and large language model artificial intelligence. Eye News. 2023. URL: https://www.eyenews.uk.com/media/31505/eye-am23-onex-arun-proof-2.pdf [accessed 2024-10-24]
40. Kuehn BM. More than one-third of US individuals use the internet to self-diagnose. JAMA 2013;309(8):756-757. [doi: 10.1001/jama.2013.629] [Medline: 23443421]
41. Shimizu E, Yazu H, Aketa N, Tanji M, Sakasegawa A, Nakayama S, et al. Innovative artificial intelligence-based cataract diagnostic method uses a slit-lamp video recording device and multiple machine-learning. Invest Ophthalmol Vis Sci 2021;62(8):1031.
42. Goh JHL, Lim ZW, Fang X, Anees A, Nusinovici S, Rim TH, et al. Artificial intelligence for cataract detection and management. Asia Pac J Ophthalmol (Phila) 2020;9(2):88-95 [FREE Full text] [doi: 10.1097/01.APO.0000656988.16221.04] [Medline: 32349116]
43. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med 2023;116(5):181-182 [FREE Full text] [doi: 10.1177/01410768231173123] [Medline: 37199678]
44. Chen X, Zhao Z, Zhang W, Xu P, Wu Y, Xu M, et al. EyeGPT: ophthalmic assistant with large language models for patient inquiries and medical education. J Med Internet Res 2024 Nov 1:-6063 (forthcoming)(forthcoming). [doi: 10.2196/60063]
45. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. Assoc Comput Mach 2024;15(3):39. [doi: 10.1145/364128]
46. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models in medical applications: a survey. arXiv:2405.07468 2024:1-42 Preprint published on May 13, 2024.

Abbreviations

AI: artificial intelligence
BLEU: bilingual evaluation understudy
BLIP: Bootstrapping Language-Image Pre-training
CIDEr: consensus-based image description evaluation
CNNs: convolutional neural networks
GRUs: gated recurrent units
LLM: large language model
QA: question answering
ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence
RNN: recurrent neural network
SPICE: Semantic Propositional Image Caption Evaluation

XSL•FO
RenderX

Edited by G Eysenbach; submitted 27.10.23; peer-reviewed by TK Yoo, S McRoy; comments to author 03.01.24; revised version received 24.02.24; accepted 05.09.24; published 30.12.24.

Please cite as:
Zhao Z, Zhang W, Chen X, Song F, Gunasegaram J, Huang W, Shi D, He M, Liu N
Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration
J Med Internet Res 2024;26:e54047
URL: https://www.jmir.org/2024/1/e54047
doi:10.2196/54047
PMID:

©Ziwei Zhao, Weiyi Zhang, Xiaolan Chen, Fan Song, James Gunasegaram, Wenyong Huang, Danli Shi, Mingguang He, Na Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Role of Assistive Technology in Enabling Older Adults to Achieve Independent Living: Past and Future

Anna Sweeting[1*], BSc, MPhil; Katie A Warncken[1*], BSc; Martyn Patel[1,2*], BMBCh, MA

[1]Norwich Institute for Healthy Ageing, Norwich Medical School, University of East Anglia, Norwich, United Kingdom

[2]Older Peoples Medicine Department, Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Martyn Patel, BMBCh, MA
Older Peoples Medicine Department
Norfolk and Norwich University Hospital NHS Foundation Trust
Colney Lane
Norwich, NR4 7UY
United Kingdom
Phone: 44 01603286286 ext 4009
Email: martyn.patel@nnuh.nhs.uk

Abstract

In this viewpoint, we present evidence of a marked increase in the use of assistive technology (AT) by older adults over the last 25 years. We also explain the way in which this use has expanded not only as an increase in terms of the total number of users but also by going beyond the typical scopes of use from its inception in 1999 to reach new categories of users. We outline our opinions on some of the key driving forces behind this expansion, such as population demographic changes, technological advances, and the promotion of AT as a means to enable older adults to achieve independent living. As well as our review of the evolution of AT over the past 25 years, we also discuss the future of AT research as a field and the need for harmonization of terminology in AT research. Finally, we outline how our experience in North Norfolk (notably the United Kingdom's most old age–dependent district) suggests that cocreation may be the key to not only successful research trials in the field of AT but also to the successful sustained adoption of AT beyond its original scope of use.

(J Med Internet Res 2024;26:e58846) doi:10.2196/58846

KEYWORDS

assistive technology; older adults; users; aging; aging in place; UK; cocreation; research trial; independent living; North Norfolk; disability; injury; tool; use; design; barrier

Spectacular Increase in Assistive Technology Use and Published Research Over the Last 25 Years

Practitioners in the field of assistive technology (AT) will be well versed in how over the last quarter of a century there has been a marked increase in the use of traditional or historically available AT (referring to the type of AT available prior to 2000) aimed at assisting those with a disability or specific injury, while acknowledging that some important barriers to access remain [1,2]. Here, we discuss how equity disparities in the access to AT vary globally and how countries facing disproportionately higher growth among the demographic of older adults may experience the least opportunity of access.

The same practitioners will also be aware that the number and type of different technologies being applied in the field of AT developed since the turn of the millennium have increased at an even greater pace than the rate of increase in the total amount of AT used [3]. In this viewpoint, we will cover the plethora of categories of AT available and highlight how comparing trials in the field of AT is a challenge due to heterogeneity in related terminology and research outcomes.

Notably, over this 25-year span of time, what has increased most of all is the size of the potential user pool, defined as potential users who require AT to enhance their lives. This user pool includes adults of all ages living with specific disease-related disability or those specifically seeking support for independent living, including older adults with nondisease-specific frailty for whom AT is a powerful tool to facilitate aging in place [4,5]. Throughout this article we will

XSL•FO
RenderX

use local, regional, and global research evidence to draw a line between the increase in potential users and growth of AT as a research field and the potential for AT to enable older adults to achieve independent living. We will also discuss how, in our opinion, cocreation may be the most valuable tool in enhancing the uptake of AT.

Growth of AT Research

Over the last 25 years, AT as a research field has witnessed a nearly 10-fold increase [6]. Using search returns for "assistive technology" on PubMed as a crude measure of research interest, in 1999, there were 181 reported studies; by 2009, that number had risen to 463 studies; and by 2019, the number of studies more than doubled again to 1089 (Figure 1) [7]. Clearly, these data alone can only speak to a trend in the increase in research output within the field of AT and not to the quality or impact of this research. There remain many unmet questions in AT research, with a recent review of the state of European AT provision noting the lack of both comprehensive data on AT provision and coverage and consistency in terminology, making direct comparisons a challenge [8].

Figure 1. Number of research articles found per year between 1999 and 2023 by searching for "assistive technology" on PubMed.

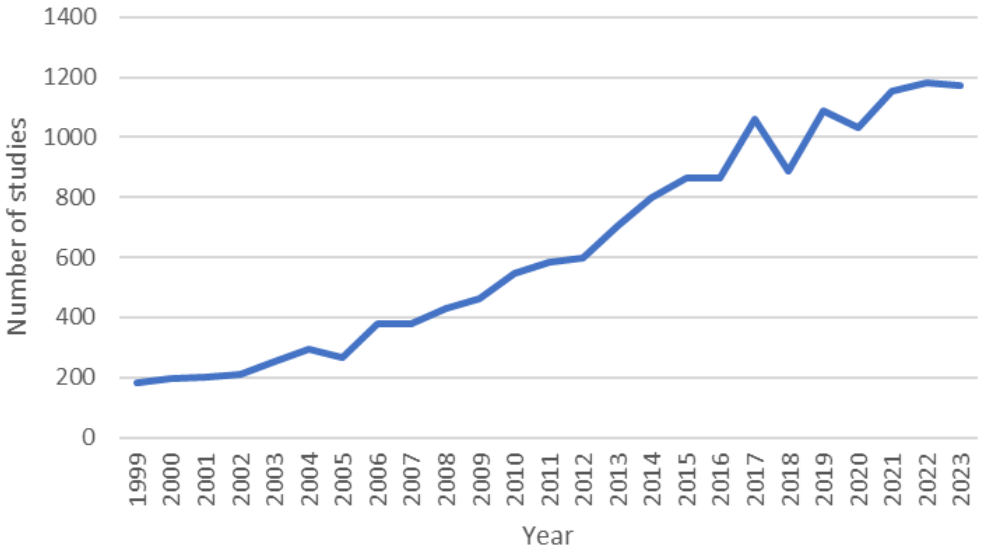

Defining AT

There are several definitions of AT, with some variance internationally. The UK definition used by the National Institute for Health and Care Research is as follows [9]:

> *Products or systems that support and help individuals with disabilities, restricted mobility or other impairments to perform functions that might otherwise be difficult or impossible. These devices support individuals to improve or maintain their daily quality of life by easing or compensating for an injury or disability.*

Within this definition, "other impairments" to performing activities of daily living can include frailty associated with old age, while many other definitions have now encompassed general frailty as well as condition-specific disabilities [10].

Having defined AT, we will now compare and contrast the state of the AT field in 1999 and in 2024.

State of AT in 1999

Typically, AT that was available around 1999—and this is by no means an exhaustive list—includes categories such as adapted home items (eg, foam padded cutlery), mobility devices (eg, walking frames and wheelchairs), sensory support (eg, hearing aid devices), telemonitoring systems (eg, pendant alarm services that connect to analog phone lines), and enhanced communication devices (eg, text to speech) [3].

Browsing the 181 studies published in 1999 retrieved from PubMed gives us some insight into where the focus of most of the research output in the field was at that point. Of those studies, 93 focused in some way on mobility using wheelchairs for those with specific disabilities, including reviews of advances in the field [11], and 35 focused on communication enhancement using assistive devices, such as those used in a school environment [12].

Twenty-Five Years Later: State of AT Today

An updated list of AT available in 2024 that was not available in 1999 (again, this is a representative and not an exhaustive list) would include [13-17] telemonitoring systems (eg, fall detector alarm services that connect via Wi-Fi); the Internet of Things, defined as any device that has added functionality due to an ability to connect to the internet; ambient/active assisted living robots, which involves the use of passive monitoring devices that can also be combined with feedback effector mechanisms; GPS-based monitoring via a bespoke device or via an app on a smartphone; smart home technology, which involves the use of passive sensor technology (eg, infrared sensors for motion detection) to keep track of surrogate markers of function in a defined environment; and smartphone apps to

XSL•FO
RenderX

promote maintenance of function (eg, apps for visual impairment or physical activity trackers).

It is harder to pin down a single majority topic for the 1089 AT-related studies published in 2019, but it is clear that there is a much greater emphasis on the use and innovation afforded by new technologies, along with more studies that address the use of AT for older adults without a specific disability, asking questions such as “can robots replace carers?” [18]. Estimates for the number of users of AT are hard to obtain, given that AT is provided by many different means (eg, health care contact, social care contact, and the open market). However, as an illustration, in Norfolk, United Kingdom, there are an estimated 12,000 users of AT provided by the adult social care sector, representing roughly 2% of the total adult population (personal communication, C Metcalf, Assistive Technology County Manager, Adult Social Care at Norfolk County Council; email correspondence May 26, 2023).

Some commonalities of AT do exist. The AT of 2019 was more likely to connect to other devices or the internet, to contain a microprocessor chip, and to require the user to have access to a broadband communications network. If we were to narrow the focus to developments in the last year, the AT-related articles published in 2024 start to increasingly include references to artificial intelligence (AI) in their scope, representing a global “arms race” to find ways to use this nascent technology [14].

However, it would be a false dichotomy to state that the pools of available AT in 1999 and 2024 are entirely separate from each other. There is of course a continuum between the two phases. Some AT tools that are currently on the market are more or less exactly the same as those available in 1999 (eg, personal falls alarms), whereas some have incrementally changed but would still be recognizable as the same product (eg, wheelchairs). However, there is some AT available today that bears no relation to the AT of 1999. We will give examples of these technologies below when discussing the current categories of AT, but first we must discuss how the experience of AT access may vary globally in response to changes in population demographics.

Actual and Projected Increase of Older Adults (Especially in Lower- and Middle-Income Countries)

One may hypothesize that the single largest driver in the expansion of AT is the global change in population demographics, with the number of older adults increasing (and projected to continue to increase) in both total numbers and as a percentage of overall populations worldwide, as demonstrated in a World Health Organization report [19]. There is evidence to support this claim within high-income countries, where measures of increased dependency are matched by increased AT use [20]. As has been pointed out, this growth is not equal in all countries, and there is a disappointing disparity in that lower- and middle-income countries that have the fastest expansion rates of older adults are more likely to have the least ready access to AT or a substantial body of AT research [21]. Furthermore, in 2019, an analysis of disease burden data showed that the statistical evidence at the global population level demonstrates a disappointing maintenance of average disability levels in the population 70 years and older to date [22].

Given the large and expanding number of individuals aged over 70 years with disability worldwide, there have been calls for designers of new AT, especially in the field of AI, to pay particular attention to factors affecting equity of access and thus enable AT to realize its full potential in the global market [23].

To substantiate this hypothesis, we have to also answer the following question: Why does future care have to be technologically based rather than having more human carers?

The Flip Side of an Aging Population: Not Enough Carers, or the Old Age Dependency Ratio

Alongside an increase in older adults comes an urgent need for solutions to care problems that are not human-focused. As the number of adults in retirement increases, the corollary is that the percentage of adults of working age available (or willing) to work in caring roles decreases [19,24]. Put simply, not only are there currently not enough carers, but as birth rates decline in many nations, barring migration, there is not likely to be a positive change in workforce availability at an individual country level [25]. Therefore, the demand for AT to fill a gap as a nonhuman solution to (some) traditional care needs has expanded the categories of use that AT might occupy [13,14].

This lack of carer availability creates a bottleneck pressure, as evidenced within the British health care system, with hospital bed days occupied by patients waiting for care increasing year on year, accounting for 1 in 7 occupied beds at points of peak pressure [26].

In the United Kingdom, North Norfolk is the district (geographical unit below the county size) with the highest old age dependency ratio (OADR), which refers to the number of people of State Pension age (SPA; currently set at 60 years for women and 65 years for men) and older for every 1000 people aged between 16 years and up to the SPA [27]. For North Norfolk (a coastal district with no large urban centers), the OADR is currently at 607, whereas South Cambridgeshire (a district with a large student population) has an OADR of 316 and Southwark (a central London district) has an OADR of only 113. The highest OADR clusters occur in rural and coastal communities, and this ratio is highlighted as a particular driver of health inequality [28].

Inconsistent AT Definitions: Risks of Too Many Categories

Before turning our attention to some of the structural drivers of the expansion of AT, it is important to spend a moment digressing on the harmonization of terminology. It is our opinion that a field that does not agree on set definitions of terms and research outcomes risks fracturing. A fractured field loses the ability to present a unified front to policy makers.

XSL•FO
RenderX

There is no set agreed-upon method for subcategorizing AT; as has been noted, this multiplicity of descriptive categorizations in the research literature risks holding back the field due to the potential for duplication of work with overlapping categories and an inherent challenge in comparing 2 or more studies that use very different categorization techniques [3,5]. Similar challenges have been tackled in other emergent research fields in the last quarter-century, particularly in frailty research, where organizations such as the Global Leadership Initiative on Sarcopenia have sought to harmonize definitions [29]. Perhaps a similar exercise in setting international standards for AT research is required.

Interestingly, this challenge of developing a shared terminology may be an inherent property reflecting the way in which AT has expanded, in that we see not only more and better versions of historic AT products but also an increase in the variety, complexity, and interactivity of AT products [13-17,21]. The AT development pathway has adopted general technological developments since the turn of the millennium, including GPS technology, information connectivity, and miniaturization of computer chips to name only a few relevant fields.

The complexity of categorization of AT may be influenced by the arena of discussion. When introducing the concept of AT subcategories in patient and public focus groups, we prefer simplicity, such as this 4-category definition outlined in a recent systematic review [5]: (1) AT that enables accessible communication (eg, closed-caption technology [CCT]), (2) AT that triggers a request for emergency assistance (eg, falls alarm technology), (3) AT that promotes or improves physical well-being (eg, step counters or pedometers), and (4) AT that promotes or improves mental well-being (eg, mindfulness apps).

This categorization enables the quick introduction of a new concept over a short period of time by using a broader definition set. Equally, when discussing the nuances of the myriad technologies deployed in AT, a research methods paper may have 10 or more categories [3].

Having discussed the choices in different formats of categorization, we now return to our main argument of outlining the reasons driving the changes in AT over the last 25 years. We have presented above how the use of AT has increased over time and how the population most likely to use AT has increased and will increase further; below, we lay out what are, in our opinion, the 3 key drivers of the evolution of AT available.

Three Factors Driving the Need for Changes in AT Over Time

Phasing Out of Analog and Phasing In of Digital Technology

One particular challenge/opportunity of AT development has been heralded by the analog copper wire phone line technology switch-off schedule in the United Kingdom, which is due to be completed by 2025 [30]. The UK government has challenged the AT market to derive a benefit to service users from the switch from analog to digital phone lines, which is a significant challenge to industry given the large section of AT products (such as personal fall alarm sensors) that are reliant on phone lines to function. As most households in the United Kingdom now have either wired-in broadband or 4G network Wi-Fi via smartphones (ie, wireless internet connectivity), we have clearly reached a tipping point where the reliance on copper wire technology no longer makes sense [31].

Broadening of Scope

We have covered the change in global population demographics that has led to a greater proportion of older adults; however, we also need to recognize that the types of older adults that may benefit from AT have also expanded due to a deliberate broadening of the scope of use of AT. While much of the available AT will always be used to support older adults with a condition- or injury-specific disability, the purpose of AT use has expanded to include that of assisting older adults to maintain an independent living environment in their own homes, also known as aging in place, with many reviews dedicated to this topic [6].

This concept of aging in place has gained substantial traction with policy makers to the extent that it has entered the official lexicon of government, with a typical definition as follows [32]: “The ability to live in one’s own home and community safely, independently, and comfortably, regardless of age, income, or ability level.”

Therefore, we can see that AT advances are not only driven by the changes in the technological landscape but also by the broadening of scope of use of AT to help solve what is arguably the biggest societal challenge of our day [4]. Crucially, this broadening of scope is more than simply the inevitable product of combining the trends of more older adults along with declining carer availability, but can also be seen as a positive expression of a societally determined response to the success of midlife health care. This in turn creates the opportunity to extend the healthy and independent fraction of the lifespan rather than creating a prolonged preterminal/dependent “last act” to our lives [5,6].

Cross-Fertilization of Ideas With Mainstream Society

Another driver of the profile of AT has been the adoption of some ideas that started in AT for widespread use in mainstream society [3]. A key example of this cross-over in use is provided by CCT, which started initially as a service for the deaf community. CCT use has now spread further than its creators could have envisaged, as it is commonly used by many adults without a hearing impairment when wishing to stream video content silently (eg, in a shared-use or public space) or to pick up a new language quickly. Furthermore, there is evidence of a developmental benefit of CCT, in that children who watch videos with both audio and CCT switched on achieve reading milestones earlier than those using only audio [33]. This exchange of ideas is bidirectional, with mainstream GPS technology being put into use as AT (and/or as a research tool) in the field of dementia care [34].

XSL•FO
RenderX

The Research Base Needs to Keep Pace With Changes in AT Use

According to these 3 drivers of AT evolution, we now turn to methods the AT research community may wish to adopt to help AT keep pace with the demands of the future. Alongside the change in how AT is used has been a change in how AT is researched. A recent systematic review noted that very few of the published trials of AT to support older adults were of a gold standard for evidential trials (ie, randomized or clinical controlled trials) [5]. Further work has been done to create research protocols using the action-design-research methodology, which incorporates codesign into its structure, but these are yet to be tested in a controlled trial [35]. There is also a lack of consistency in the choice of outcome measures to allow comparison between trials [36].

One of the potential benefits of cocreation/codesign is to attempt to mitigate the phenomenon of a drop off in the use of new technology, which often occurs a few weeks after introduction in trials; it is suggested that the influence of codesign may work to embed AT into daily life in such a way that it inspires longevity of use [5].

Learning from Cocreation: the Norfolk Experience

As mentioned, the county of Norfolk in the United Kingdom contains districts of exceptionally high old age dependency, such as North Norfolk [31], a large coastal county situated in the east of England with a population of approximately 1 million. With areas of both coastal and urban deprivation, combined with a high old age dependency, Norfolk is, in our opinion, an ideal location to conduct research into the evolution of health and social care service design. We recently conducted a patient and public involvement exercise across Norfolk on the topic of the transformation of health care with technology; the first theme of this exercise was based on the use of AT and remote monitoring [37].

Overall, the listening exercise showed that people are keen to engage in discussion around the use, design, and barriers to AT. The participants felt it was important to capture these views to inform future research and service design. Both AT users (ie, patients) and their carers recognized the benefits of using AT, with 2 key benefits highlighted being the reassurance provided to the individual around being safe at home and the reduction of stress on the National Health Service (NHS), which is particularly relevant at a time when it is recognized the NHS is under maximum strain [38]. The reassurance gained by AT provided in a virtual ward setting went further than safety, which was also based on the safety felt by knowing one is being monitored by a health care professional and could have easy access back to hospital if need be.

There is a wealth of evidence in the research literature that multiple subtypes of AT trialed among older adults with frailty show promising results, although it is acknowledged that these results have been at a small scale and are most impressive in cases of mild to moderate frailty [5,39].

This listening exercise also highlighted the key concern of the public around falling victim to scams. In this regard, a codesign process could be important to explore how to reassure the public that this risk can be mitigated, given that older adults are particularly vulnerable to scams that rely on the impersonation of authority figures [40].

Cocreation Examples in AT Research

Encouragingly, the existing research base does seem to already integrate cocreation (ie, involving current or potential future service users) as a philosophy of trial design [5,35]. Cocreation is important to enable access to all. Our previous report and others have demonstrated the variation of digital literacy, internet access, and the impact of generational challenges (the youngest older adults self-describe as more “tech savvy” than the oldest older adults) [37,41]. There is also evidence that national programs can have a positive impact on digital literacy for older adults, as long as they do not treat all older adults as starting from the same knowledge base [42]. It is necessary to explore design considerations to address these potential inequalities of access.

There are many positive examples of the principle of cocreation being used in the field of AT, ranging from a novel robot design appropriate for many users to the customization of AT to provide bespoke solutions in a health care setting [43,44].

Public engagement would be a good way to explore the key motivating factors for technology users. One of our focus groups reported that one barrier to motivation was a reluctance to engage with new technology due to the fear of getting it wrong. However, the participants were also able to suggest a potential enabler to overcome this barrier: using peer-to-peer support. Indeed, this technique has been shown to help overcome similar barriers to older adults’ digital engagement in the health arena [37,45].

Other tools to enhance the use of AT have also been tested, including a novel collaborative tool that uses expert allied health professionals as a critical part of the codesign process [46]. The inherent aim of this approach is to maximize the dialog between providers and consumers of AT to increase the likelihood of a positive outcome.

Thus, there are multiple reasons to support the extensive use of cocreation for both research trials and for pragmatically enhancing the adoption of existing AT. We have already identified local services that are providing support around digital support/training on leaving the hospital (eg, digitally based shopping) that could potentially be expanded to help overcome this barrier.

Summary

A quarter-century is a fair span of time by which to measure change, and the change in what constitutes AT has indeed been mighty over that period [3,5]. This change has occurred on many levels. At the simplest level, there are now more older adults living with disability who have a need of the latest 2024 version of 1999 AT equivalents, just as much as their forerunners would

XSL•FO
RenderX

have back in the last millennium [4]. At the next quantum of change, there are not only more older adults but there is also a higher percentage of older adults who require AT due to those without a condition- or injury-specific disability nevertheless struggling to maintain independent living owing to frailty associated with age [6].

Advances in AT have been driven by opportunities greater than a priority need to broaden scope. These advances have been driven by the expansion of the envelope of what is possible, as new technologies that were the science fiction topics of our youth become the common-place technology of today.

To maximize the true potential of AT, we recommend that research trialists embrace the use of the principle of cocreation as a vehicle to enhance not only AT uptake but also the longevity of its impact [5,32].

Limitations of This Viewpoint

While we have endeavored to back up our opinions with evidence from a variety of sources, we acknowledge that we have not challenged our central view that AT is an overall force for good in the battle to enable older adults to live independently. We note there is a counterview, with valid concerns that AT has potential downsides as well as benefits. These concerns include the potential for AT to undermine rather than enhance autonomy and the way in which AI tools trained on the shared characteristics of the majority population may struggle to handle diversity [47,48].

Although we have held up AT as a way to enable aging in place, we have not discussed potential public health strategies to ensure a lower health dependency in later life (potentially obviating the need for AT), an approach that may be most relevant to lower- and middle-income countries [49].

We have also not discussed alternatives to aging in place, such as assisted living facilities, multigenerational living arrangements, or long-term care facilities, but wish to acknowledge that these are equally valid preferred choices for some individuals [50,51].

It is also important to acknowledge that the effects of the aging process vary immensely among individuals and older adults are a diverse and heterogenous group, many of whom do not require any form of care.

However, we believe there has never been a time where there is a more pressing need for high-quality research into AT usage to address these concerns and to make aging in place a reality for the many worldwide rather than just the privileged few.

Acknowledgments

This viewpoint was written with support of an internship to KAW paid by a grant from the Research Capability Fund of Norfolk and Norwich University Hospitals NHS (National Health Service) Foundation Trust.

Conflicts of Interest

None declared.

References

1. Hammel J. Assistive Technology and Environmental Intervention (AT-EI) impact on the activity and life roles of aging adults with developmental disabilities. Phys Occup Ther Geriatr 2009 Jul 28;18(1):37-58. [doi: 10.1080/j148v18n01_04]
2. Howard J, Fisher Z, Kemp AH, Lindsay S, Tasker LH, Tree JJ. Exploring the barriers to using assistive technology for individuals with chronic conditions: a meta-synthesis review. Disabil Rehabil Assist Technol 2022 May;17(4):390-408. [doi: 10.1080/17483107.2020.1788181] [Medline: 32663110]
3. Zallio M, Ohashi T. The evolution of assistive technology: a literature review of technology developments and applications. In: Zallio M, editor. Human Factors in Accessibility and Assistive Technology. New York, NY: AHFE International; 2022.
4. Beard JR, Officer A, de Carvalho IA, Sadana R, Pot AM, Michel J, et al. The World report on ageing and health: a policy framework for healthy ageing. Lancet 2016 May 21;387(10033):2145-2154 [FREE Full text] [doi: 10.1016/S0140-6736(15)00516-4] [Medline: 26520231]
5. Ollevier A, Aguiar G, Palomino M, Simpelaere IS. How can technology support ageing in place in healthy older adults? A systematic review. Public Health Rev 2020 Nov 23;41(1):26 [FREE Full text] [doi: 10.1186/s40985-020-00143-4] [Medline: 33292707]
6. Bergschöld JM, Gunnes M, Eide AH, Lassemo E. Characteristics and range of reviews about technologies for aging in place: scoping review of reviews. JMIR Aging 2024 Jan 22;7:e50286 [FREE Full text] [doi: 10.2196/50286] [Medline: 38252472]
7. National Center for Biotechnology Information. URL: https://pubmed.ncbi.nlm.nih.gov/ [accessed 2024-03-20]
8. Mishra S, Laplante-Lévesque A, Barbareschi G, Witte LD, Abdi S, Spann A, et al. Assistive technology needs, access and coverage, and related barriers and facilitators in the WHO European region: a scoping review. Disabil Rehabil Assist Technol 2024 Feb 29;19(2):474-485. [doi: 10.1080/17483107.2022.2099021] [Medline: 35906719]
9. Assistive technology: definition and safe use. UK Government Medicines & Healthcare Products Regulatory Agency. URL: https://www.gov.uk/government/publications/assistive-technology-definition-and-safe-use/assistive-technology-definition-and-safe-use [accessed 2024-03-20]

XSL•FO
RenderX

10. MacLachlan M, Banes D, Bell D, Borg J, Donnelly B, Fembek M, et al. Assistive technology policy: a position paper from the first Global Research, Innovation, and Education on Assistive Technology (GREAT) summit. Disabil Rehabil Assist Technol 2018 Jul;13(5):454-466. [doi: 10.1080/17483107.2018.1468496] [Medline: 29790393]
11. Cooper R. Technology for disabilities. Interview by Ron Davis. BMJ 1999 Nov 13;319(7220):1290 [FREE Full text] [doi: 10.1136/bmj.319.7220.1290] [Medline: 10559043]
12. Durand VM. Functional communication training using assistive devices: recruiting natural communities of reinforcement. J Appl Behav Anal 1999;32(3):247-267 [FREE Full text] [doi: 10.1901/jaba.1999.32-247] [Medline: 10513023]
13. Wang S, Bolling K, Mao W, Reichstadt J, Jeste D, Kim H, et al. Technology to support aging in place: older adults' perspectives. Healthcare 2019 Apr 10;7(2):60 [FREE Full text] [doi: 10.3390/healthcare7020060] [Medline: 30974780]
14. Sapci AH, Sapci HA. Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review. JMIR Aging 2019 Nov 29;2(2):e15429 [FREE Full text] [doi: 10.2196/15429] [Medline: 31782740]
15. Facchinetti G, Petrucci G, Albanesi B, De Marinis MG, Piredda M. Can smart home technologies help older adults manage their chronic condition? A systematic literature review. Int J Environ Res Public Health 2023 Jan 10;20(2):1205. [doi: 10.3390/ijerph20021205] [Medline: 36673957]
16. Senjam SS, Manna S, Bascaran C. Smartphones-based assistive technology: accessibility features and apps for people with visual impairment, and its usage, challenges, and usability testing. Clin Optom 2021;13:311-322 [FREE Full text] [doi: 10.2147/OPTO.S336361] [Medline: 34866955]
17. Scott RA, Callisaya ML, Duque G, Ebeling PR, Scott D. Assistive technologies to overcome sarcopenia in ageing. Maturitas 2018 Jun;112:78-84. [doi: 10.1016/j.maturitas.2018.04.003] [Medline: 29704921]
18. Bedaf S, Marti P, De Witte L. What are the preferred characteristics of a service robot for the elderly? A multi-country focus group study with older adults and caregivers. Assist Technol 2019;31(3):147-157. [doi: 10.1080/10400435.2017.1402390] [Medline: 29125807]
19. Ageing and Health (AAH), Demographic Change and Healthy Ageing (DHA), Social Determinants of Health (SDH) WHO Team. World Report on Ageing and Health. Geneva: World Health Organization; 2020.
20. Freedman VA, Agree EM, Martin LG, Cornman JC. Trends in the use of assistive technology and personal care for late-life disability, 1992-2001. Gerontologist 2006 Feb;46(1):124-127. [doi: 10.1093/geront/46.1.124] [Medline: 16452292]
21. Garçon L, Khasnabis C, Walker L, Nakatani Y, Lapitan J, Borg J, et al. Medical and assistive health technology: meeting the needs of aging populations. Gerontologist 2016 Apr;56(Suppl 2):S293-S302. [doi: 10.1093/geront/gnw005] [Medline: 26994268]
22. GBD 2019 Ageing Collaborators. Global, regional, and national burden of diseases and injuries for adults 70 years and older: systematic analysis for the Global Burden of Disease 2019 Study. BMJ 2022 Mar 10;376:e068208. [doi: 10.1136/bmj-2021-068208] [Medline: 35273014]
23. Rubeis G, Fang ML, Sixsmith A. Equity in AgeTech for ageing well in technology-driven places: the role of social determinants in designing AI-based assistive technologies. Sci Eng Ethics 2022 Oct 27;28(6):49 [FREE Full text] [doi: 10.1007/s11948-022-00397-y] [Medline: 36301408]
24. Demand for adult social care hits record high while support from local authorities drops. The King's Fund. 2023 Mar 02. URL: https://www.kingsfund.org.uk/insight-and-analysis/press-releases/demand-adult-social-care-hits-record-high-support-local-authorities-drops [accessed 2024-03-20]
25. GBD 2021 Fertility and Forecasting Collaborators. Global fertility in 204 countries and territories, 1950-2021, with forecasts to 2100: a comprehensive demographic analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2057-2099. [doi: 10.1016/S0140-6736(24)00550-6] [Medline: 38521087]
26. Hunter DJ. At Breaking Point or Already Broken? The National Health Service in the United Kingdom. N Engl J Med 2023 Jul 13;389(2):100-103. [doi: 10.1056/NEJMp2301257] [Medline: 37428450]
27. Subnational ageing tool. Office for National Statistics, United Kingdom. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/subnationalageingtool/2020-07-20 [accessed 2024-03-20]
28. Whitty C. Chief Medical Officer's annual report 2023: health in an ageing society. Department of Health and Social Care. 2023. URL: https://tinyurl.com/232mxm47 [accessed 2024-03-20]
29. Sayer AA, Cruz-Jentoft A. Sarcopenia definition, diagnosis and treatment: consensus is growing. Age Ageing 2022 Oct 06;51(10):afac220 [FREE Full text] [doi: 10.1093/ageing/afac220] [Medline: 36273495]
30. Policy paper. Telecare stakeholder action plan: preparations for the analogue to digital switchover. Department of Health and Social Care. 2022 Dec 05. URL: https://tinyurl.com/2hfyuc83 [accessed 2024-03-20]
31. Internet access - households and individuals, Great Britain: 2020. Office for National Statistics, United Kingdom. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2020 [accessed 2024-03-20]
32. Bosch-Farré C, Malagón-Aguilera MC, Ballester-Ferrando D, Bertran-Noguer C, Bonmatí-Tomàs A, Gelabert-Vilella S, et al. Healthy ageing in place: enablers and barriers from the perspective of the elderly. A qualitative study. Int J Environ Res Public Health 2020 Sep 04;17(18):6451 [FREE Full text] [doi: 10.3390/ijerph17186451] [Medline: 32899744]

XSL•FO
RenderX

33. Gernsbacher MA. Video captions benefit everyone. Policy Insights Behav Brain Sci 2015 Oct;2(1):195-202 [FREE Full text] [doi: 10.1177/2372732215602130] [Medline: 28066803]
34. Puthusseryppady V, Morrissey S, Aung MH, Coughlan G, Patel M, Hornberger M. Using GPS tracking to investigate outdoor navigation patterns in patients with Alzheimer disease: cross-sectional study. JMIR Aging 2022 Apr 21;5(2):e28222 [FREE Full text] [doi: 10.2196/28222] [Medline: 35451965]
35. Lussier M, Couture M, Giroux S, Aboujaoudé A, Ngankam HK, Pigot H, et al. Codevelopment and deployment of a system for the telemonitoring of activities of daily living among older adults receiving home care services: protocol for an action design research study. JMIR Res Protoc 2024 Feb 29;13:e52284 [FREE Full text] [doi: 10.2196/52284] [Medline: 38422499]
36. Borgnis F, Desideri L, Converti RM, Salatino C. Available assistive technology outcome measures: systematic review. JMIR Rehabil Assist Technol 2023 Nov 15;10:e51124 [FREE Full text] [doi: 10.2196/51124] [Medline: 37782310]
37. Sweeting A, Warncken K, Patel M. The ARMOUR Project. Norwich Institute of Healthy Ageing. URL: https://healthyageingnorwich.com/projects/armour-project/ [accessed 2024-03-26]
38. Greener I, Powell M. Crisis in the UK National Health Service: what does it mean, and what are the consequences? Soc Policy Adm 2024 Jan 25;58(2):313-328. [doi: 10.1111/spol.13001]
39. Fotteler ML, Mühlbauer V, Brefka S, Mayer S, Kohn B, Holl F, et al. The effectiveness of assistive technologies for older adults and the influence of frailty: systematic literature review of randomized controlled trials. JMIR Aging 2022 Apr 04;5(2):e31916 [FREE Full text] [doi: 10.2196/31916] [Medline: 35377322]
40. Yu L, Mottola G, Kieffer CN, Mascio R, Valdes O, Bennett DA, et al. Vulnerability of older adults to government impersonation scams. JAMA Netw Open 2023 Sep 05;6(9):e2335319 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.35319] [Medline: 37738048]
41. Hargittai E, Piper AM, Morris MR. From internet access to internet skills: digital inequality among older adults. Univ Access Inf Soc 2018 May 3;18(4):881-890. [doi: 10.1007/s10209-018-0617-5]
42. McCosker A, Critchley C, Walshe J, Tucker J, Suchowerska R. Accounting for diversity in older adults' digital inclusion and literacy: the impact of a national intervention. Ageing Soc 2021 Dec 14;43(11):2629-2649. [doi: 10.1017/s0144686x21001550]
43. Fiorini L, Tabeau K, D'Onofrio G, Coviello L, De Mul M, Sancarlo D, et al. Co-creation of an assistive robot for independent living: lessons learned on robot design. Int J Interact Des Manuf 2019 Dec 10;14(2):491-502. [doi: 10.1007/s12008-019-00641-z]
44. Howard J, Tasker LH, Fisher Z, Tree J. Assessing the use of co-design to produce bespoke assistive technology solutions within a current healthcare service: a service evaluation. Disabil Rehabil Assist Technol 2024 Jan;19(1):42-51. [doi: 10.1080/17483107.2022.2060355] [Medline: 35426354]
45. Fortuna KL, Brooks JM, Umucu E, Walker R, Chow PI. Peer support: a human factor to enhance engagement in digital health behavior change interventions. J Technol Behav Sci 2019 Jun 15;4(2):152-161 [FREE Full text] [doi: 10.1007/s41347-019-00105-x] [Medline: 34337145]
46. Layton N, O'Connor J, Fitzpatrick A, Carey S. Towards co-design in delivering assistive technology interventions: reconsidering roles for consumers, allied health practitioners, and the support workforce. Int J Environ Res Public Health 2022 Nov 03;19(21):14408 [FREE Full text] [doi: 10.3390/ijerph192114408] [Medline: 36361299]
47. O'Brolcháin F. Autonomy benefits and risks of assistive technologies for persons with intellectual and developmental disabilities. Front Public Health 2018;6:296 [FREE Full text] [doi: 10.3389/fpubh.2018.00296] [Medline: 30450353]
48. Smith EM, Graham D, Morgan C, MacLachlan M. Artificial intelligence and assistive technology: risks, rewards, challenges, and opportunities. Assist Technol 2023 Sep 03;35(5):375-377. [doi: 10.1080/10400435.2023.2259247] [Medline: 37747852]
49. Williams GA, Cylus J, Al Tayara L, Roubal T, Tsilaajav T, Barber SL. Can healthy ageing moderate the effects of population ageing on economic growth and health spending trends in Mongolia? A modelling study. Health Res Policy Syst 2022 Nov 29;20(Suppl 1):122 [FREE Full text] [doi: 10.1186/s12961-022-00916-0] [Medline: 36443859]
50. Fernández-Carro C. Ageing at home, co-residence or institutionalisation? Preferred care and residential arrangements of older adults in Spain. Ageing Soc 2014 Dec 17;36(3):586-612. [doi: 10.1017/s0144686x1400138x]
51. Vasara P. Not ageing in place: negotiating meanings of residency in age-related housing. J Aging Stud 2015 Dec;35:55-64. [doi: 10.1016/j.jaging.2015.07.004] [Medline: 26568215]

Abbreviations

AI: artificial intelligence
AT: assistive technology
CCT: closed-caption technology
NHS: National Health Service
OADR: old age dependency ratio
SPA: State Pension age

XSL•FO
RenderX

Edited by G Eysenbach; submitted 27.03.24; peer-reviewed by N Zaman, J Sixsmith, C Abbatantuono; comments to author 13.05.24; revised version received 06.06.24; accepted 15.07.24; published 30.07.24.

Please cite as:
Sweeting A, Warncken KA, Patel M
The Role of Assistive Technology in Enabling Older Adults to Achieve Independent Living: Past and Future
J Med Internet Res 2024;26:e58846
URL: https://www.jmir.org/2024/1/e58846
doi:10.2196/58846
PMID:

©Anna Sweeting, Katie A Warncken, Martyn Patel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The McMaster Health Information Research Unit: Over a Quarter-Century of Health Informatics Supporting Evidence-Based Medicine

Cynthia Lokker[1], MSc, PhD; K Ann McKibbon[1], MLiS, PhD; Muhammad Afzal[2], PhD; Tamara Navarro[1], MEd, MLiS; Lori-Ann Linkins[1,3], MSc, MD; R Brian Haynes[1], MD, PhD; Alfonso Iorio[1,3], MD, PhD

[1]Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
[2]Department of Computing and Data Science, Birmingham City University, Birmingham, United Kingdom
[3]Department of Medicine, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Cynthia Lokker, MSc, PhD
Health Information Research Unit
Department of Health Research Methods, Evidence, and Impact
McMaster University
1280 Main St W
CRL 137
Hamilton, ON, L8S 4K1
Canada
Phone: 1 2897883272
Email: lokkerc@mcmaster.ca

Abstract

Evidence-based medicine (EBM) emerged from McMaster University in the 1980-1990s, which emphasizes the integration of the best research evidence with clinical expertise and patient values. The Health Information Research Unit (HiRU) was created at McMaster University in 1985 to support EBM. Early on, digital health informatics took the form of teaching clinicians how to search MEDLINE with modems and phone lines. Searching and retrieval of published articles were transformed as electronic platforms provided greater access to clinically relevant studies, systematic reviews, and clinical practice guidelines, with PubMed playing a pivotal role. In the early 2000s, the HiRU introduced Clinical Queries—validated search filters derived from the curated, gold-standard, human-appraised Hedges dataset—to enhance the precision of searches, allowing clinicians to hone their queries based on study design, population, and outcomes. Currently, almost 1 million articles are added to PubMed annually. To filter through this volume of heterogenous publications for clinically important articles, the HiRU team and other researchers have been applying classical machine learning, deep learning, and, increasingly, large language models (LLMs). These approaches are built upon the foundation of gold-standard annotated datasets and humans in the loop for active machine learning. In this viewpoint, we explore the evolution of health informatics in supporting evidence search and retrieval processes over the past 25+ years within the HiRU, including the evolving roles of LLMs and responsible artificial intelligence, as we continue to facilitate the dissemination of knowledge, enabling clinicians to integrate the best available evidence into their clinical practice.

(J Med Internet Res 2024;26:e58764) doi:10.2196/58764

KEYWORDS

health informatics; evidence-based medicine; information retrieval; evidence-based; health information; Boolean; natural language processing; NLP; journal; article; Health Information Research Unit; HiRU

History of the Heath Information Research Unit

The McMaster University School of Medicine was founded in 1967. One of its basic principles, and arguably one of its most important, was using problems and experience in clinical settings (ie, problem-based learning [PBL]) for health sciences education rather than reliance on lectures and expert opinions. Probably the biggest challenge of PBL was how to identify and apply the best current evidence-based knowledge from the medical literature to address clinical problems.

XSL•FO
RenderX

The Health Information Research Unit (HiRU) was formed in 1985 (Figure 1) with funding from the Rockefeller Foundation. Since the HiRU started, we have been working on the problem of providing the best evidence from studies to the clinicians who need it quickly, efficiently, and in easy-to-use formats. Over the years, we have researched and developed tools to achieve this goal, building in a stepwise fashion. Our first set of studies centered on evaluating dissemination and utilization methods [1], now often referred to as knowledge translation. As we worked on the task of integrating research findings into practice, we realized that finding and evaluating studies that were ready for clinical practice were greater challenges than we first thought.

In the early 1980s, David Sackett and his colleagues at McMaster University published a series of articles in the *Canadian Medical Association Journal* on how to read a clinical journal article [2]. In 1991, Gordon Guyatt at McMaster University coined the phrase "evidence-based medicine" (EBM) [3]. This was followed by publication of the "Users' guides to the medical literature" in the *Journal of the American Medical Association* [4,5]. These guides offered an influential series of articles that were well received and changed approaches to medical decision-making in times of uncertainty while keeping current with changing practice. The EBM movement grew; however, the problem of having fast and efficient access to the best literature remained.

A parallel step in efficient access and understanding came when Brian Haynes, founder of the HiRU, and colleagues from around the world petitioned medical journal editors to require more informative abstracts for clinically important articles. They stated that clinical studies could be more readily appraised (by aiding rapid comprehension) if abstracts, which were freely and widely available via MEDLINE, included the information that was needed for both critical appraisal of scientific merit and appropriate clinical use. They proposed that the 200-300–word abstracts include the exact question addressed, study design, findings directly pertinent to the study question, and key conclusions for clinical application [6,7]. This structure has been adopted by most clinical journals. Notably, the resulting structured abstract has been an aid for researchers in the fields of natural language and artificial intelligence (AI) interested in retrieving and summarizing studies. Along the same line, the HiRU designed and delivered *ACP Journal Club* in collaboration with the American College of Physicians (ACP) [8]. Each month, highly structured summaries of high-quality, high-impact articles across a list of core clinical journals are published in *Annals of Internal Medicine* with an accompanying commentary by a practicing clinician. Worth noting, other universal paradigms of EBM stemmed from editorials published in *ACP Journal Club* and *Evidence-Based Medicine*. In 1995, an editorial discussed the value of structuring clinical questions using the main components (ie, population/patient, intervention, comparator, and outcome [PICO] terms) [9], and a series of editorials published between 2001 and 2016 described the evolving 4S, 5S, and 6S hierarchies or pyramids as models for organizing and selecting the best available evidence [10].

Despite these advances, finding studies with the best evidence for real-time clinical care remained a challenge. The HiRU broached training clinicians to search MEDLINE [11,12]. If a clinician could easily search the literature, they might be more likely to practice EBM. This and similar studies found that while clinicians could learn to search MEDLINE, the search strategy development process was cumbersome and time-consuming. The HiRU helped the US National Library of Medicine (NLM) by developing and testing early versions of Grateful Med [13]. This software program made searching MEDLINE easier, especially for clinicians. In 1997, PubMed became a free, searchable database that includes the abstracts and records in MEDLINE, which was visited by an average 3.4 million users each day in 2021 [14].

Figure 1. Key milestones for the Health Information Research Unit (HiRU; maroon) and evidence-based medicine (EBM; yellow). ACP: American College of Physicians; PICO: patient/population, intervention, comparator, outcome; PLUS:Premium LiteratUre Service.

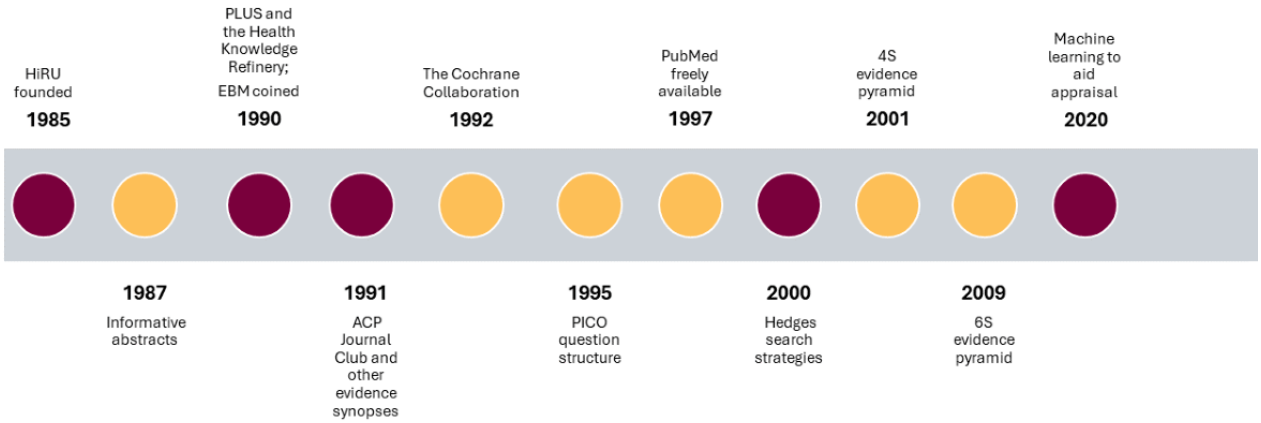

XSL•FO
RenderX

Innovations in Search Retrieval and Preappraised Evidence Resources for Clinicians

Hedges

With funding from the Canadian Institutes of Health Research and the US National Institutes of Health/NLM, we next started analyzing index and abstract terms in MEDLINE to see if we could identify (and then build) "canned" searches that would retrieve only the studies with the strongest methods (eg, randomized controlled trials for studies of treatment). This would allow for content searches to be limited to articles with a higher likelihood of being rigorously performed. We were successful with this endeavor and have continuously conducted research to improve these "clinical queries" using advanced information analysis and retrieval methods, including new work with AI.

To accomplish this, we developed a method to evaluate the performance of Boolean search terms and combinations, termed "hedges," to retrieve target articles, which represents an early natural language processing application. The method is based on a manual search of 160 clinical journals for the year 2000, with 49,028 articles classified by article format, interest to human health care, and purpose category, which were critically appraised by highly trained staff with expertise in health research methods. The resulting Hedges dataset was used to test thousands of combinations of search terms using a diagnostic accuracy test approach [15]. The Hedges strategies retrieve original and review articles [16] for a range of purposes such as treatment [17], prediction guides [18], etiology [19], prognosis [20], and diagnosis [21]. Database-specific strategies are available on our website [22], with several available through PubMed, MEDLINE, and EMBASE as Clinical Queries. In 2013, we reported on the robustness, assessed over 10 years, of the strategies for retrieving relevant articles in PubMed [23].

McMaster PLUS

McMaster PLUS (Premium LiteratUre Service) curates high-quality evidence and is comprised of a series of steps in a process referred to as the Health Knowledge Refinery (HKR) (Figure 2). The HKR was designed to distill the flow of articles from a broad selection of clinical journals into a refined product of preappraised literature to support clinicians through several evidence services and products.

The current PLUS process integrates several HiRU innovations. First, nightly automated searches of ~125 journals (selected from a critical appraisal of >800 clinical journals) are filtered using highly sensitive Hedges search strategies developed by the unit in the early 2000s to identify systematic reviews, evidence-based guidelines, and original studies addressing questions of treatment, prevention, quality improvement, economics, diagnosis, etiology, prediction guides, and prognosis.

Second, the filtrate is further refined using a recently developed BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining)–based machine learning model that classifies articles for meeting explicit criteria for research rigor and clinical relevance. The model maintains 99% sensitivity (recall) and high precision, reducing the work required to manually appraise articles by 60% [24].

Third, research associates manually critically appraise the filtered articles using established standards for scientific rigor. From a list of 63 clinical disciplines, they select the relevant medical disciplines and indicate if the article is also relevant to rehabilitation professions or nursing. Article assessments are then reviewed by a clinician with expertise in research methods.

Fourth, the McMaster Online Rating of Evidence (MORE) system provides postpublication clinical peer review to further refine the HKR output to the articles that are important for consideration in clinical practice [25]. Qualifying articles are automatically sent to registered clinicians for each pertinent discipline identified during the critical appraisal step. MORE raters are a crowd of ~6000 health care providers practicing worldwide, who rate the articles on 7-point scales for relevance to their practice and newsworthiness (defined as useful new information for physicians). Physicians [26], nurses [27], and rehabilitation practitioners [28] are invited to join MORE via specific links.

Fifth, the final refined HKR filtrate of articles that have relevance and newsworthiness scores ≥4 is sent out via email alerts; added to the PLUS database; and made available through searchable interfaces such as ACP JournalWise, Evidence Alerts, ACCESSSS, and other products and services to support a range of clinical knowledge users (Table 1). We provide content to publishers for updating evidence-based textbooks, which we customize under the auspices of McMaster University, a not-for-profit, publicly funded university. Usage of HiRU products remains strong. Across several services, we have >250,000 registered users with >25,000 new registrations in 2023.

Finally, through PLUS, selected articles that are highly rated for clinical relevance and newsworthiness to practicing clinicians and that are of interest to the broad readership of *Annals of Internal Medicine* are summarized and included in *ACP Journal Club*. This EBM-focused enterprise has been active since 1990 and has evolved over time. Originally, article selection was done through a manual search of the contents of printed journals by trained research associates with articles that met the methods criteria shared through fax machines. This process evolved to reviewing the online table of contents and sharing through email. In 2014, we developed an in-house web-based infrastructure that allowed for automated retrieval of article titles and abstracts from PubMed, a collection of a range of data elements entered by research associates during critical appraisal, automated email-based alerting for clinical editors and MORE raters, and a collection of ratings. This also allowed us to track articles that were not applicable to the HKR (eg, not related to human health care, basic science, and methodology studies), those that were relevant but did not meet methodologic criteria, and those below clinical relevance and newsworthiness thresholds for alerting. Through this process, we have curated a dataset of almost 200,000 articles that have been reviewed by human experts, along with various associated data elements gathered at the time of publication.

XSL•FO
RenderX

Figure 2. Health Information Research Unit Health Knowledge Refinery process. AI: artificial intelligence; PLUS: Premium LiteratUre Service.

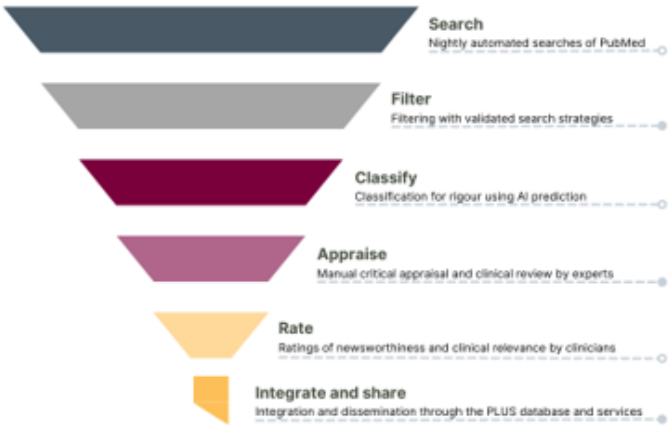

Table 1. Current McMaster PLUS (Premium LiteratUre Service) projects and alerting services [29].

Project or service	Description
EvidenceAlerts	Alerting service for physicians, nurses, and rehabilitation professionals
ACCESSSS	A smart search engine that retrieves content from multiple sources and orders it according to the pyramid of evidence
ACP[a] Journal Club	Synopses with accompanying clinical commentaries of high-quality, clinically relevant studies, published monthly in *Annals of Internal Medicine*
ACP JournalWise	Alerting service and platform for searching and filtering from the top 120 clinical journals with options to personalize content
DynaMed	Alerting service customized for DynaMed editors and authors
EBM Guidelines	Alerting service for Duodecim editors and authors
Pain+	Alerting service for pain specialists
Rehab+	Alerting service for rehabilitation professionals
Public Health+	Alerting service for public health professionals
CLOT+	Alerting service for hematology/thrombosis
KT+	Alerting service for knowledge translation researchers and workers
McMaster Optimal Aging Portal	Alerting service for practitioners, patients, and the public interested in evidence relevant to aging
STAT!Ref Evidence Alerts	Alerting service customized for Teton Data Systems
Helsebiblioteket.no	ACCESSSS customized for the Norwegian National Health library

[a]ACP: American College of Physicians.

The Evolving Role of AI and Machine Learning

The Era of Machine Learning

The Hedges dataset provided the HiRU with opportunities to collaborate with external research partners. The dataset has served as a validated reference standard of high-quality studies and has been used to build and test more advanced literature retrieval models; the search strategies have also been used to identify articles to build comparison datasets. Historically, the retrieval models included conventional methods such as Boolean search filters and citation-based algorithms. Over time, machine learning approaches, particularly supervised models trained using these data, have been effective in retrieving high-quality clinical studies from the biomedical literature [30]. Some examples include the study by Aphinyanaphongs et al [31] in 2005, which used articles abstracted in *ACP Journal Club* (as a PLUS derivative) and machine learning to automatically construct filters identifying high-quality, content-specific articles in internal medicine. Building on this work, using the Hedges dataset, Kilicoglu et al [32] experimented with 3 supervised machine learning methods (naïve Bayes, support vector machine, and boosting), and obtained comparatively better results.

Advancements in machine learning have dramatically improved computer capabilities through deep neural networks. In 2018, Del Fiol et al [33] used the Hedges dataset to develop a model that was perhaps the first to investigate the use of deep learning techniques to identify reports of scientifically sound studies in the biomedical literature. In 2020, Afzal et al [34] used an optimized multilayer feed-forward neural network model, the multilayer perceptron, for identifying scientifically sound studies

XSL•FO
RenderX

and filtering out others. Both studies used the PubMed Clinical Queries filter with a "narrow" scope, favoring high precision over high recall [33,34].

Through the HKR, our dataset of classified and appraised articles has grown, and we have expanded our machine learning capabilities. We recently used the data to train machine learning models in-house, achieving both high recall (sensitivity) and precision. We assessed the efficacy of advanced deep learning models that include BERT (Bidirectional Encoder Representations from Transformers) and its variations such as BioBERT, BlueBERT, and PubMedBERT [35]. In 2023, a state-of-the-art model named DL-PLUS, trained using the dataset, excelled in classifying articles for meeting rigor and clinical relevance compared with competitors [24]. This represents our initial phase of machine learning work, with plans to improve the classification of articles by purpose category (eg, treatment, diagnosis) and rigor to provide more targeted results for searchers.

Large Language Models

The incorporation of LLMs into health care is rapidly growing, and EBM is no exception. LLMs have demonstrated remarkable success in various downstream tasks, including information extraction and evidence summarization of clinical studies and bodies of literature [36,37]. To enable the uptake of evidence into practice, LLMs have potential to summarize individual studies and bodies of evidence. At the HiRU, we have conducted pilot testing of AI-generated summaries to gauge how well they include EBM-pertinent details such as a clear research objective, details on the methodology, effect sizes, and conclusions for clinical application. Analogous to our early focus on ensuring ready access to interpretable findings and identifying the best-quality research, of key interest is developing and testing prompts (much like search queries) that task the AI tools with returning the necessary information and assessing the factuality of generated summaries. Our near-future plans include contributing to the methods of ensuring the trustworthiness of evidence summaries generated by LLMs [38], with a focus on a clinical audience. Our extended plan includes developing custom LLMs at the HiRU, contributing to other tasks of the evidence ecosystem such as evidence appraisals.

Responsible AI

The majority of machine learning models, particularly deep learning models, are inherently "black boxes," concealing internal details about the decision-making processes [39]. As a result, determining the true effectiveness of an AI model becomes challenging. In some cases, bias may inherently exist in the dataset (eg, it does not accurately represent the overall population), and machine learning engineers may unintentionally introduce bias (eg, during class balancing through sampling). In a recent model training experiment, we observed that when we attempted to balance classes with an oversampling method, the result was a highly accurate yet poorly calibrated model. Subsequent evaluations using calibration methods revealed that the model with the initial unbalanced data was, in fact, well-calibrated. Therefore, responsible AI practice is crucial, particularly in digital health applications, due to the potential ethical issues associated with AI technologies. The increasing reliance on AI in health care has highlighted the importance of addressing concerns such as biases, discrimination, errors, and lack of transparency in outcomes. Implementing responsible AI practices in this context becomes paramount, emphasizing ethical principles and human values to minimize biases, enhance fairness, ensure interpretability, and ultimately prevent adverse consequences on human and societal well-being [40]. Our focus on responsible AI for futuristic AI models will target 2 key areas: prioritizing the examination of datasets to identify biases related to coverage, correctness, and fairness, while concurrently enhancing the interpretability and explainability of AI model prediction workflows, thereby clarifying the decision-making process.

Conclusion

Over the last quarter-century, the HiRU has been at the leading edge of developing novel processes and tools to support clinicians in the practice of EBM, which is work that continues to support a wide range of users and clients. The mission of the HiRU is to harness information science and technology to build customized, high-efficiency, continuously updated evidence services. The fast pace of change in machine learning and AI in recent years is providing a new landscape for innovation. With collaborators across disciplines, we plan to leverage new and emerging tools to continue to facilitate health care evidence retrieval, appraisal, and dissemination while building on our foundation of ensuring the quality and trustworthiness of the work we develop.

Acknowledgments

Over the history of the Health Information Research Unit, many people have made important contributions to our work. We acknowledge Nancy Wilczynski, Cindy Walker-Dilks, Susan Marks, Jean Mackay, Lori-Weise-Kelly, and our IT team (Chris Cotoi, Rick Parrish, and Nicolas Hobson) among others too numerous to add. We also acknowledge the funders and clients who made this work possible.

Authors' Contributions

All authors contributed to the content and editing of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

McMaster University, a not-for-profit institution, has contracts that are managed by the Health Information Research Unit and supervised by AI, RBH, and LAL with several professional and commercial publishers to supply newly published studies and systematic reviews that are critically appraised for research methods and assessed for clinical relevance through the McMaster Premium LiteratUre Service (McMaster PLUS). RBH has received remuneration for his roles in developing the McMaster Online Rating of Evidence System (MORE) and in deploying McMaster PLUS and McMaster Hedges, the intellectual property rights for all of which belong to McMaster University. CL, TN, and LAL are partly paid through these contracts. The other authors have no conflicts to declare.

References

1. Haynes RB, Davis DA, McKibbon A, Tugwell P. A critical appraisal of the efficacy of continuing medical education. JAMA 1984 Jan 06;251(1):61-64. [Medline: 6690767]
2. Department of Clinical Epidemiology and Biostatistics, McMaster University. How to read clinical journals: I. Why to read them and how to start reading them critically. Can Med Assoc J 1981 Mar 01;124(5):555-558 [FREE Full text] [Medline: 7471000]
3. Guyatt G. Evidence-based medicine. ACP J Club 1991 Mar 01;114(2):A16. [doi: 10.7326/acpjc-1991-114-2-a16]
4. Guyatt GH, Rennie D. Users' guides to the medical literature. JAMA 1993 Nov 03;270(17):2096-2097. [Medline: 8411578]
5. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993 Nov 03;270(17):2093-2095. [Medline: 8411577]
6. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. A proposal for more informative abstracts of clinical articles. Ann Intern Med 1987 Apr;106(4):598-604. [Medline: 3826959]
7. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. Ann Intern Med 1990 Jul 01;113(1):69-76. [doi: 10.7326/0003-4819-113-1-69] [Medline: 2190518]
8. Haynes RB. The origins and aspirations of ACP Journal Club. ACP J Club 1991 Jan 01;114(1):A18. [doi: 10.7326/acpjc-1991-114-1-a18]
9. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995;123(3):A12-A13. [Medline: 7582737]
10. Alper BS, Haynes RB. EBHC pyramid 5.0 for accessing preappraised evidence and guidance. Evid Based Med 2016 Aug;21(4):123-125. [doi: 10.1136/ebmed-2016-110447] [Medline: 27325531]
11. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990 Jan 01;112(1):78-84. [doi: 10.7326/0003-4819-112-1-78] [Medline: 2403476]
12. McKibbon KA, Haynes RB, Baker LM, Flemming T, Walker C. Teaching clinicians to search MEDLINE: description and evaluation of a short course. Res Med Educ 1986;25:231-236. [Medline: 3538990]
13. McKibbon KA, Haynes RB, Dilks CJ, Ramsden MF, Ryan NC, Baker L, et al. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Comput Biomed Res 1990 Dec;23(6):583-593. [doi: 10.1016/0010-4809(90)90042-b] [Medline: 2276266]
14. Catching Up with PubMed. National Library of Medicine. 2021. URL: https://www.nlm.nih.gov/oet/ed/pubmed/pubmed-catchup.html [accessed 2024-03-14]
15. Wilczynski NL, Morgan D, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak 2005 Jun 21;5:20 [FREE Full text] [doi: 10.1186/1472-6947-5-20] [Medline: 15969765]
16. Montori VM, Wilczynski NL, Morgan D, Haynes RB, Hedges Team. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ 2005 Jan 08;330(7482):68 [FREE Full text] [doi: 10.1136/bmj.38336.804167.47] [Medline: 15619601]
17. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ 2005 May 21;330(7501):1179 [FREE Full text] [doi: 10.1136/bmj.38446.498542.8F] [Medline: 15894554]
18. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R, Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annu Symp Proc 2003;2003:728-732 [FREE Full text] [Medline: 14728269]
19. Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. AMIA Annu Symp Proc 2003;2003:719-723 [FREE Full text] [Medline: 14728267]
20. Wilczynski NL, Haynes RB. Optimal search strategies for detecting clinically sound prognostic studies in EMBASE: an analytic survey. J Am Med Inform Assoc 2005;12(4):481-485 [FREE Full text] [doi: 10.1197/jamia.M1752] [Medline: 15802476]

XSL•FO
RenderX

21. Wilczynski NL, Haynes RB, Hedges Team. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. BMC Med 2005 Mar 29;3:7 [FREE Full text] [doi: 10.1186/1741-7015-3-7] [Medline: 15796772]
22. Hedges Project. McMaster University Health Information Research Unit. URL: https://hiruweb.mcmaster.ca/hkr/hedges/ [accessed 2021-08-05]
23. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc 2013;20(2):363-368 [FREE Full text] [doi: 10.1136/amiajnl-2012-001075] [Medline: 23019242]
24. Lokker C, Bagheri E, Abdelkader W, Parrish R, Afzal M, Navarro T, et al. Deep learning to refine the identification of high-quality clinical research articles from the biomedical literature: performance evaluation. J Biomed Inform 2023 Jun;142:104384 [FREE Full text] [doi: 10.1016/j.jbi.2023.104384] [Medline: 37164244]
25. Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D, McMaster Premium Literature Service (PLUS) Project. Second-order peer review of the medical literature for clinical practitioners. JAMA 2006 Apr 19;295(15):1801-1808. [doi: 10.1001/jama.295.15.1801] [Medline: 16622142]
26. Become a MORE Rater - Physicians. McMaster PLUS. URL: https://hiru.mcmaster.ca/MORE/physicians/contact_us.html [accessed 2024-03-23]
27. Become a MORE Rater - Nurses. McMaster PLUS. URL: https://hiru.mcmaster.ca/MORE/nurses/contact_us.html [accessed 2024-03-23]
28. Become a MORE Rater - Rehab. McMaster PLUS. URL: https://hiru.mcmaster.ca/MORE/rehab/contact_us.html [accessed 2024-03-23]
29. PLUS Project Directory. McMaster University Health Information Research Unit. URL: https://hiruweb.mcmaster.ca/hkr/what-we-do/plus-directory/ [accessed 2024-03-23]
30. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, et al. Machine learning approaches to retrieve high-quality, clinically relevant evidence from the biomedical literature: systematic review. JMIR Med Inform 2021 Sep 09;9(9):e30401 [FREE Full text] [doi: 10.2196/30401] [Medline: 34499041]
31. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207-216 [FREE Full text] [doi: 10.1197/jamia.M1641] [Medline: 15561789]
32. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16(1):25-31 [FREE Full text] [doi: 10.1197/jamia.M2996] [Medline: 18952929]
33. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study. J Med Internet Res 2018 Jun 25;20(6):e10281 [FREE Full text] [doi: 10.2196/10281] [Medline: 29941415]
34. Afzal M, Alam F, Malik KM, Malik GM. Clinical context-aware biomedical text summarization using deep neural network: model development and validation. J Med Internet Res 2020 Oct 23;22(10):e19810 [FREE Full text] [doi: 10.2196/19810] [Medline: 33095174]
35. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Linkins L, et al. A deep learning approach to refine the identification of high-quality clinical research articles from the biomedical literature: protocol for algorithm development and validation. JMIR Res Protoc 2021 Nov 29;10(11):e29398 [FREE Full text] [doi: 10.2196/29398] [Medline: 34847061]
36. Olsen E. Introducing: Consensus 2.0. Consensus. 2023 Nov. URL: https://consensus.app/home/blog/introducing-consensus-2-0/ [accessed 2024-01-28]
37. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023 Aug 24;6(1):158. [doi: 10.1038/s41746-023-00896-7] [Medline: 37620423]
38. Zhang G, Jin Q, Jered McInerney D, Chen Y, Wang F, Cole C, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform 2024 May;153:104640. [doi: 10.1016/j.jbi.2024.104640] [Medline: 38608915]
39. Karim M, Shajalal M, Graß A, Döhmen T, Chala S, Boden A, et al. Interpreting Black-box Machine Learning Models for High Dimensional Datasets. 2023 Presented at: IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA); October 9-13, 2023; Thessaloniki, Greece. [doi: 10.1109/dsaa60987.2023.10302562]
40. Trocin C, Mikalef P, Papamitsiou Z, Conboy K. Responsible AI for digital health: a synthesis and a research agenda. Inf Syst Front 2021 Jun 26;25(6):2139-2157. [doi: 10.1007/s10796-021-10146-4]

Abbreviations

ACP: American College of Physicians
AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining

XSL•FO
RenderX

EBM: evidence-based medicine
HiRU: Health Information Research Unit
HKR: Health Knowledge Refinery
LLM: large language model
MORE: McMaster Online Rating of Evidence
NLM: Natural Library of Medicine
PBL: problem-based learning
PICO: population/patient, intervention, comparator, and outcome
PLUS: Premium LiteratUre Service

Edited by G Eysenbach; submitted 28.03.24; peer-reviewed by W Hersh, J Zheng, J Galvez-Olortegui; comments to author 08.06.24; revised version received 15.06.24; accepted 16.07.24; published 31.07.24.

Please cite as:
Lokker C, McKibbon KA, Afzal M, Navarro T, Linkins LA, Haynes RB, Iorio A
The McMaster Health Information Research Unit: Over a Quarter-Century of Health Informatics Supporting Evidence-Based Medicine
J Med Internet Res 2024;26:e58764
URL: https://www.jmir.org/2024/1/e58764
doi:10.2196/58764
PMID:

©Cynthia Lokker, K Ann McKibbon, Muhammad Afzal, Tamara Navarro, Lori-Ann Linkins, R Brian Haynes, Alfonso Iorio. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Beyond Step Count: Are We Ready to Use Digital Phenotyping to Make Actionable Individual Predictions in Psychiatry?

Abigail Ortiz[1*], MSc, MD, FRCP; Benoit H Mulsant[1*], MSc, MD, FRCP

Department of Psychiatry, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Benoit H Mulsant, MSc, MD, FRCP
Department of Psychiatry
Temerty Faculty of Medicine
University of Toronto
8th Floor
250 College Street
Toronto, ON, M5T 1R8
Canada
Phone: 1 416 979 6948
Email: benoit.mulsant@utoronto.ca

Abstract

Some models for mental disorders or behaviors (eg, suicide) have been successfully developed, allowing predictions at the population level. However, current demographic and clinical variables are neither sensitive nor specific enough for making individual actionable clinical predictions. A major hope of the "Decade of the Brain" was that biological measures (biomarkers) would solve these issues and lead to precision psychiatry. However, as models are based on sociodemographic and clinical data, even when these biomarkers differ significantly between groups of patients and control participants, they are still neither sensitive nor specific enough to be applied to individual patients. Technological advances over the past decade offer a promising approach based on new measures that may be essential for understanding mental disorders and predicting their trajectories. Several new tools allow us to continuously monitor objective behavioral measures (eg, hours of sleep) and densely sample subjective measures (eg, mood). The promise of this approach, referred to as digital phenotyping, was recognized almost a decade ago, with its potential impact on psychiatry being compared to the impact of the microscope on biological sciences. However, despite the intuitive belief that collecting densely sampled data (big data) improves clinical outcomes, recent clinical trials have not shown that incorporating digital phenotyping improves clinical outcomes. This viewpoint provides a stepwise development and implementation approach, similar to the one that has been successful in the prediction and prevention of cardiovascular disease, to achieve clinically actionable predictions in psychiatry.

(J Med Internet Res 2024;26:e59826) doi:10.2196/59826

KEYWORDS

digital phenotype; digital phenotyping; prediction; predictions; mental health; mental illness; mental illnesses; mental disorder; mental disorders; US National Institute of Mental Health; NIMH; psychiatry; psychiatrist; psychiatrists

"It is difficult to make predictions, especially about the future" [1], as Yogi Berra stated.

Thirty years after the US National Institute of Mental Health (NIMH) declared the final decade of the 20th century to be "the Decade of the Brain," Tom Insel, the NIMH director at that time acknowledged,

> *I spent 13 years at NIMH really pushing on the neuroscience and genetics of mental disorders, and when I look back on that I realize that while I think I succeeded at getting lots of really cool papers published by cool scientists at fairly large costs—I think $20 billion—I don't think we moved the needle in reducing suicide, reducing hospitalizations, improving recovery for the tens of millions of people who have mental illness.* [2]

A challenge contributing to this issue is that day-to-day clinical interactions and decision-making processes in psychiatry remain fundamentally the same as they were 50 years ago. Decision-making is still based on binary (present/absent) diagnoses inferred from clinical symptoms assessed during an interview at a single time point. Once a diagnosis has been

XSL•FO
RenderX

made, treatment focuses on managing acute symptoms, with simple long-term strategies (eg, "what gets you well, keeps you well") and reactive handling of adverse outcomes (eg, readmission following a suicide attempt). Moreover, despite extraordinary empirical advances in psychopharmacology [3-5], all psychotropic medications help a subset of patients. Thus, most psychiatrists still follow a trial-and-error approach, which can take an inordinate amount of time [6]. Once patients are stable, the traditional model of clinical monitoring typically involves monthly visits that are either too infrequent or too frequent given the labile nature of mental disorders.

Some models for mental disorders or behaviors (eg, suicide) have been successfully developed and they allow predictions at the population level [7,8]. However, current demographic and clinical variables are neither sensitive nor specific enough for making individual actionable clinical predictions. Using suicide as an example, a recent meta-analysis concluded that predictive ability has not improved across 50 years of research [9]. This is in part because these predictive models are still solely based on sociodemographic and static descriptive clinical variables. For instance, an older, White man, with depression and alcohol use may have a 100-fold higher likelihood of killing himself during the next year than somebody in the general population. Unfortunately, this astounding relative risk (100) means that the likelihood that this single patient kills himself during the next year is 10/1000 (1%) rather than 10/100,000. Psychiatrists cannot hospitalize 100 patients for one year to save one life. Using larger datasets with more demographic and clinical variables may improve the precision of these population-based models, but they are unlikely to impact individual clinical outcomes.

A major hope of the "Decade of the Brain" was that biological measures (biomarkers) would solve these issues [10,11] and lead to precision psychiatry [12]. A variety of biomarkers have now been reliably associated with mental disorders and their outcomes [13]. However, as models based on sociodemographic and clinical data, even when these biomarkers differ significantly between groups of patients and control participants, they are neither sensitive nor specific enough to be applied to individual patients [14-16]. Typically, a quarter to a third of the patients have normal values and a quarter to a third of the controls have pathologic values [17]. New models that will integrate sociodemographic, clinical, and biological data are being developed, but they have not yet been shown to improve clinical outcomes [15,16,18,19]. The field of preventive psychiatry still lacks an understanding of the complex mechanisms underlying mental disorders and their treatment, in contrast to the well-established pathophysiologic models in cardiology or oncology, which link risk factors to outcomes and have led to reductions in mortality related to heart disease, stroke, or cancer [20,21].

Looking back, we believe that our inability over two or three "decades of the brain" to bridge the gap between biology and clinical symptoms is due to the lack of an intermediate level of description. In the field of artificial intelligence (AI), the early attempts to create neural networks had to be abandoned because the early 2-layer networks could not process information usefully (eg, interpret images or translate languages). It took more than 30 years to understand that neural networks needed an intermediate layer to process information usefully and to implement this intermediate layer [22,23]. The Research Domain Criteria initiative of the NIMH has attempted to bridge this gap with limited success to date [24]. Technological advances over the past decade offer a promising approach based on new measures that may be essential for understanding mental disorders and predicting their trajectories. Several new tools allow us to continuously monitor objective behavioral measures (eg, hours of sleep) and densely sample subjective measures (eg, mood). The promise of this approach, referred to as digital phenotyping was recognized almost a decade ago [25], with its potential impact on psychiatry being compared to the impact of the microscope on biological sciences [26].

However, simply gathering a large amount of informative data about a single patient is not helpful by itself. Just as a clinician struggles to synthesize the information from over 100 clinical notes and dozens of laboratory reports available in an electronic health record, the massive amount of data provided by digital phenotyping is useless unless these data can be properly analyzed in a clinical context and with the proper statistical tools. Machine learning, by extracting complex patterns from multiple sources of high-dimensional time-varying data [27], is an ideal tool to address this problem [28-30]. Nonetheless, some challenges with machine learning still need to be addressed before it can be used to make actionable predictions in psychiatry. These challenges include unreliable inherent assumptions [31,32], model instability [33], and lack of interpretability [34] or explainability [35] of results (the black box problem).

Despite the intuitive belief that collecting densely sampled data (big data) improves clinical outcomes, recent clinical trials have not shown that incorporating digital phenotyping improves clinical outcomes [36-40]. This is an example of the so-called "AI chasm," which refers to the gap between developing algorithms and their actual real-world implementation and clinical impact [41]. As discussed above, some reasons for this chasm include the disconnect between building good individual predictive models for the broader population and making individual inferences [42]. Bayesian procedures offer a potential solution to link inferences and predictions [43]. Other simpler reasons to address include the lack of expertise needed to implement tools into clinical practice [44], poor data quality compromising the reliability and accuracy of models [45-47], and a lack of standardization [38].

During the next decade, achieving clinically actionable predictions in psychiatry will require a stepwise development and implementation approach [12], similar to the successful methods used in other medical fields, such as predicting and preventing cardiovascular disease [48]. The first step will be to identify individual digital measures of objective behaviors and subjective mental states (digital markers) that can be integrated with sociodemographic data, clinical characteristics, and biomarkers to create multimodal signatures that predict clinical outcomes (akin to risk scores in other medical fields). These multimodal signatures will need to be reliably and accurately associated with individual clinical states and trajectories. Validation studies will require cohort studies with adequate

XSL•FO
RenderX

sample sizes and sufficient duration to generate enough analyzable clinical events [49]. We foresee that different multimodal signatures will be used for detection (diagnosis) and prediction (prognosis). Once the reliability, specificity, and sensitivity of these multimodal signatures are established, prospective randomized clinical trials (RCTs), with adequate sample sizes, complemented with real-world observation studies, will need to demonstrate that they can be used to tailor the treatment of individual patients and improve outcomes. Albeit costly, these RCTs are needed to fulfill the promise of clinically actionable predictions leading to individualized, timely treatment. We are also mindful of a recent review [50] that emphasized methodological challenges in RCTs investigating smartphone-based treatment interventions for mental disorders, including lack of trial registrations, inappropriate comparators, lack of blinding, selection bias, and lack of generalizability.

In parallel, the incorporation of digital phenotyping, first in RCTs and later in clinical practice, will require addressing complex ethical issues raised by the intense monitoring of behavior and mental state, which some people may consider too invasive regardless of its potential benefits [51]. This work can be informed by lessons learned from other fields [52]. To prepare for the deployment of the new decision-making tools we foresee and to understand their potential and pitfalls [53], we will need to start training in medical schools and continue training throughout our professional life [54].

In conclusion, we believe that technological advances, in the context of a more holistic approach that considers all determinants of health, will allow us to create individual multimodal signatures for early detection and personalized intervention for mental disorders. However, this potential transformation in psychiatry will require another decade of investment and effort to become a reality.

Conflicts of Interest

None declared.

References

1. Dickstein DP. Editorial: it's difficult to make predictions, especially about the future: risk calculators come of age in child psychiatry. J Am Acad Child Adolesc Psychiatry 2021 Aug;60(8):950-951. [doi: 10.1016/j.jaac.2020.12.029] [Medline: 33383160]
2. Troisi A. Biological psychiatry is dead, long live biological psychiatry!. Clin Neuropsychiatry 2022 Dec;19(6):351-354 [FREE Full text] [doi: 10.36131/cnfioritieditore20220601] [Medline: 36627945]
3. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet 2018 Dec 07;391(10128):1357-1366. [doi: 10.1016/S0140-6736(17)32802-7] [Medline: 29477251]
4. Rush AJ. STAR*D: what have we learned? Am J Psychiatry 2007 Mar;164(2):201-204. [doi: 10.1176/ajp.2007.164.2.201] [Medline: 17267779]
5. Huhn M, Nikolakopoulou A, Schneider-Thoma J, Krause M, Samara M, Peter N, et al. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. Lancet 2019 Sep 14;394(10202):939-951 [FREE Full text] [doi: 10.1016/S0140-6736(19)31135-3] [Medline: 31303314]
6. Gaynes BN, Warden D, Trivedi MH, Wisniewski SR, Fava M, Rush AJ. What did STAR*D teach us? Results from a large-scale, practical, clinical trial for patients with depression. Psychiatr Serv 2009 Nov;60(11):1439-1445. [doi: 10.1176/ps.2009.60.11.1439] [Medline: 19880458]
7. Kirkbride JB, Jackson D, Perez J, Fowler D, Winton F, Coid JW, et al. A population-level prediction tool for the incidence of first-episode psychosis: translational epidemiology based on cross-sectional data. BMJ Open 2013;3(2):e001998 [FREE Full text] [doi: 10.1136/bmjopen-2012-001998] [Medline: 23399458]
8. Moriarty AS, Meader N, Snell KI, Riley RD, Paton LW, Chew-Graham CA, et al. Prognostic models for predicting relapse or recurrence of major depressive disorder in adults. Cochrane Database Syst Rev 2021 May 06;5(5):CD013491 [FREE Full text] [doi: 10.1002/14651858.CD013491.pub2] [Medline: 33956992]
9. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Mar;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
10. Walter H. The third wave of biological psychiatry. Front Psychol 2013;4:582 [FREE Full text] [doi: 10.3389/fpsyg.2013.00582] [Medline: 24046754]
11. Blumberger DM, Daskalakis ZJ, Mulsant BH. Biomarkers in geriatric psychiatry: searching for the holy grail? Curr Opin Psychiatry 2008 Nov;21(6):533-539. [doi: 10.1097/YCO.0b013e328314b763] [Medline: 18852558]
12. No Author listed. The right treatment for each patient: unlocking the potential of personalized psychiatry. Nat Mental Health 2023 Sep 06;1(9):607-608. [doi: 10.1038/s44220-023-00131-y]
13. Hoy N, Lynch SJ, Waszczuk MA, Reppermund S, Mewton L. Transdiagnostic biomarkers of mental illness across the lifespan: a systematic review examining the genetic and neural correlates of latent transdiagnostic dimensions of

XSL•FO
RenderX

psychopathology in the general population. Neurosci Biobehav Rev 2023 Dec;155:105431 [FREE Full text] [doi: 10.1016/j.neubiorev.2023.105431] [Medline: 37898444]
14. Ioannidis JPA, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA 2011 Jun 01;305(21):2200-2210. [doi: 10.1001/jama.2011.713] [Medline: 21632484]
15. Rost N, Dwyer DB, Gaffron S, Rechberger S, Maier D, Binder EB, et al. Multimodal predictions of treatment outcome in major depression: a comparison of data-driven predictors with importance ratings by clinicians. J Affect Disord 2023 Apr 14;327:330-339. [doi: 10.1016/j.jad.2023.02.007] [Medline: 36750160]
16. Sajjadian M, Uher R, Ho K, Hassel S, Milev R, Frey BN, et al. Prediction of depression treatment outcome from multimodal data: a CAN-BIND-1 report. Psychol Med 2023 Sep;53(12):5374-5384 [FREE Full text] [doi: 10.1017/S0033291722002124] [Medline: 36004538]
17. Winter NR, Leenings R, Ernsting J, Sarink K, Fisch L, Emden D, et al. Quantifying Quantifying deviations of brain structure and function in major depressive disorder across neuroimaging modalities of brain structure and function in major depressive disorder across neuroimaging modalities. JAMA Psychiatry 2022 Sep 01;79(9):879-888 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.1780] [Medline: 35895072]
18. Winter NR, Blanke J, Leenings R, Ernsting J, Fisch L, Sarink K, et al. A systematic evaluation of machine learning-based biomarkers for major depressive disorder. JAMA Psychiatry 2024 Apr 01;81(4):386-395. [doi: 10.1001/jamapsychiatry.2023.5083] [Medline: 38198165]
19. Zierer C, Behrendt C, Lepach-Engelhardt AC. Digital biomarkers in depression: a systematic review and call for standardization and harmonization of feature engineering. J Affect Disord 2024 Apr 05;356:438-449. [doi: 10.1016/j.jad.2024.03.163] [Medline: 38583596]
20. Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-Baptiste D, Saslow D, et al. Cancer screening in the United States, 2017: a review of current American Cancer Society guidelines and current issues in cancer screening. CA Cancer J Clin 2017 Mar;67(2):100-121 [FREE Full text] [doi: 10.3322/caac.21392] [Medline: 28170086]
21. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, et al. Heart disease and stroke statistics-2018 update: a report from the American Heart Association. Circulation 2018 Mar 20;137(12):e67-e492. [doi: 10.1161/CIR.0000000000000558] [Medline: 29386200]
22. Mulsant BH. A neural network as an approach to clinical diagnosis. MD Comput 1990;7(1):25-36. [Medline: 2407923]
23. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct;323(6088):533-536. [doi: 10.1038/323533a0]
24. Weinberger DR, Glick ID, Klein DF. Whither Research Domain Criteria (RDoC)?: The good, the bad, and the ugly. JAMA Psychiatry 2015 Dec;72(12):1161-1162. [doi: 10.1001/jamapsychiatry.2015.1743] [Medline: 26558844]
25. Torous J, Kiang MV, Lorme J, Onnela J. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
26. Onnela J, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Dec;41(7):1691-1696 [FREE Full text] [doi: 10.1038/npp.2016.7] [Medline: 26818126]
27. Ortiz A, Bradler K, Hintze A. Episode forecasting in bipolar disorder: is energy better than mood? Bipolar Disord 2018 Jan 22;20(5):470-476. [doi: 10.1111/bdi.12603] [Medline: 29356281]
28. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry 2016 Oct;21(10):1366-1371 [FREE Full text] [doi: 10.1038/mp.2015.198] [Medline: 26728563]
29. Hahn T, Kircher T, Straube B, Wittchen H, Konrad C, Ströhle A, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry 2015 Jan;72(1):68-74. [doi: 10.1001/jamapsychiatry.2014.1741] [Medline: 25409415]
30. Khodayari-Rostamabad A, Hasey G, Maccrimmon DJ, Reilly J, de Bruin H. A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. Clin Neurophysiol 2010 Dec;121(12):1998-2006 [FREE Full text] [doi: 10.1016/j.clinph.2010.05.009] [Medline: 21035741]
31. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
32. Crowley RJ, Tan YJ, Ioannidis JPA. Empirical assessment of bias in machine learning diagnostic test accuracy studies. J Am Med Inform Assoc 2020 Jul 01;27(7):1092-1101 [FREE Full text] [doi: 10.1093/jamia/ocaa075] [Medline: 32548642]
33. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023 Dec;65(8):e2200302 [FREE Full text] [doi: 10.1002/bimj.202200302] [Medline: 37466257]
34. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018 Dec;286(3):800-809. [doi: 10.1148/radiol.2017171920] [Medline: 29309734]

XSL•FO
RenderX

35. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 2023 Jan 18;6(1):6 [FREE Full text] [doi: 10.1038/s41746-023-00751-9] [Medline: 36653524]
36. Fedor S, Lewis R, Pedrelli P, Mischoulon D, Curtiss J, Picard RW. Wearable technology in clinical practice for depressive disorder. N Engl J Med 2023 Dec 28;389(26):2457-2466. [doi: 10.1056/nejmra2215898]
37. Faurholt-Jepsen M, Frost M, Christensen EM, Bardram JE, Vinberg M, Kessing LV. The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial. Psychol Med 2020 Apr 04;50(5):838-848. [doi: 10.1017/S0033291719000710] [Medline: 30944054]
38. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open 2021 Mar 01;4(3):e211276 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1276] [Medline: 33704476]
39. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021 Sep 01;374:n1872 [FREE Full text] [doi: 10.1136/bmj.n1872] [Medline: 34470740]
40. Tønning ML, Faurholt-Jepsen M, Frost M, Martiny K, Tuxen N, Rosenberg N, et al. The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: the RADMIS randomized controlled trial. J Affect Disord 2021 Mar 01;282:354-363 [FREE Full text] [doi: 10.1016/j.jad.2020.12.141] [Medline: 33421863]
41. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1:40 [FREE Full text] [doi: 10.1038/s41746-018-0048-y] [Medline: 31304321]
42. Bzdok D, Ioannidis JP. Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci 2019 Apr;42(4):251-262. [doi: 10.1016/j.tins.2019.02.001] [Medline: 30808574]
43. Hunter DJ, Holmes C. Where medical statistics meets artificial intelligence. N Engl J Med 2023 Sep 28;389(13):1211-1219. [doi: 10.1056/NEJMra2212850] [Medline: 37754286]
44. Pinsky MR, Bedoya A, Bihorac A, Celi L, Churpek M, Economou-Zavlanos NJ, et al. Use of artificial intelligence in critical care: opportunities and obstacles. Crit Care 2024 Apr 08;28(1):113 [FREE Full text] [doi: 10.1186/s13054-024-04860-z] [Medline: 38589940]
45. Cho S, Ensari I, Weng C, Kahn MG, Natarajan K. Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth 2021 Mar 19;9(3):e20738 [FREE Full text] [doi: 10.2196/20738] [Medline: 33739294]
46. Canali S, Schiaffonati V, Aliverti A. Challenges and recommendations for wearable devices in digital health: data quality, interoperability, health equity, fairness. PLOS Digit Health 2022 Oct;1(10):e0000104 [FREE Full text] [doi: 10.1371/journal.pdig.0000104] [Medline: 36812619]
47. Teno JM. Garbage in, garbage out-words of caution on big data and machine learning in medical practice. JAMA Health Forum 2023 Mar 03;4(2):e230397 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.0397] [Medline: 36795395]
48. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Development and validation of the American Heart Association's PREVENT equations. Circulation 2024 Feb 06;149(6):430-449. [doi: 10.1161/circulationaha.123.067626]
49. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
50. Tønning ML, Kessing LV, Bardram JE, Faurholt-Jepsen M. Methodological challenges in randomized controlled trials on smartphone-based treatment in psychiatry: systematic review. J Med Internet Res 2019 Oct 27;21(10):e15362 [FREE Full text] [doi: 10.2196/15362] [Medline: 31663859]
51. Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. NPJ Digit Med 2018 Dec 19;1(1):68 [FREE Full text] [doi: 10.1038/s41746-018-0075-8] [Medline: 31211249]
52. Cohen C. Ethical issues in mandatory drug testing. In: Rosner R, Weinstock R, editors. Ethical Practice in Psychiatry and the Law. Boston, MA: Springer; 1990:313-325.
53. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
54. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy augmented intelligence in health care. J Med Syst 2022 Jan 12;46(2):12 [FREE Full text] [doi: 10.1007/s10916-021-01790-z] [Medline: 35020064]

Abbreviations

AI: artificial intelligence
NIMH: National Institute of Mental Health
RCT: randomized clinical trial

XSL•FO
RenderX

Edited by G Eysenbach; submitted 23.04.24; peer-reviewed by K Cho, R Philippe; comments to author 02.07.24; revised version received 09.07.24; accepted 16.07.24; published 05.08.24.

Please cite as:
Ortiz A, Mulsant BH
Beyond Step Count: Are We Ready to Use Digital Phenotyping to Make Actionable Individual Predictions in Psychiatry?
J Med Internet Res 2024;26:e59826
URL: https://www.jmir.org/2024/1/e59826
doi:10.2196/59826
PMID:

©Abigail Ortiz, Benoit H Mulsant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A Quarter-Century of Online Informatics Education: Learners Served and Lessons Learned

William Hersh[1], MD

Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health & Science University, Portland, OR, United States

Corresponding Author:
William Hersh, MD
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
3181 SW Sam Jackson Park Rd
Biomedical Information Communication Center
Portland, OR, 97239
United States
Phone: 1 5034944563
Email: hersh@ohsu.edu

Abstract

The value and methods of online learning have changed tremendously over the last 25 years. The goal of this paper is to review a quarter-century of experience with online learning by the author in the field of biomedical and health informatics, describing the learners served and the lessons learned. The author details the history of the decision to pursue online education in informatics, describing the approaches taken as educational technology evolved over time. A large number of learners have been served, and the online learning approach has been well-received, with many lessons learned to optimize the educational experience. Online education in biomedical and health informatics has provided a scalable and exemplary approach to learning in this field.

(J Med Internet Res 2024;26:e59066) doi:10.2196/59066

KEYWORDS

distance education; online learning; asynchronous education; biomedical and health informatics; learning; biomedical; health informatics; education; educational; educational technology; online program; online course; teaching; students

Introduction

Twenty-five years ago, in early 1999, I made a decision that seemed like just another project at the time but turned out to profoundly affect the work of my career. This was a decision to investigate the interest and subsequently move my introductory biomedical and health informatics course in our graduate program at Oregon Health & Science University (OHSU) to an online format. It was just a single course, but would lead to profound new directions not only for my career but also for the graduate program I led.

In 1999, the internet and online educational environment were much different from that in the present. Although many people had wired internet connections of decent bandwidth at their work, broadband internet to the home was relatively new. There was no wireless connectivity to speak of. The online educational environment was similarly nascent at the time, and distance learning was viewed as a second-class method of education. A quarter-century later, we can look back at the learners who were served and the lessons learned.

Prelude

The decision to move my introductory course online was driven by potential students expressing interest in the course being offered online. In early 1999, we gauged interest by conducting a mail survey of 500 random people each on the American Medical Informatics Association (AMIA) and Healthcare Information and Management Systems Society (HIMSS) listservs [1]. From the 288 surveys returned, there was definite interest in the course, its delivery on the World Wide Web, the need to have access by dial-up modem (not everyone had broadband internet in their homes), and content covering a range of topics similar to what was in the existing on-campus course I was already teaching. Although the OHSU School of Nursing and the Educational Communications Department had started pursuing distance learning, there was no institutional infrastructure to support our effort.

XSL•FO
RenderX

I had already been teaching my introductory course for over half a decade. In the early 1990s, I was relatively a new Assistant Professor at OHSU, where I had come after completing medical training at the University of Illinois Chicago and a National Library of Medicine (NLM)–funded postdoctoral fellowship at Harvard University. I was recruited to OHSU as part of a new academic informatics program funded by the Integrated Advanced Information Management Systems initiative of NLM [2]. My early focus, like most in informatics, was developing a research program.

However, I always had an interest in teaching and had the opportunity to develop an introductory informatics course in 1993 that was an elective in OHSU's relatively new Master of Public Health program. The course initially consisted of lectures by half a dozen informatics faculty who were part of the program. Although others enjoyed lecturing on their areas of expertise, I had larger ambitions, and under the mentorship of the OHSU Provost at the time, Lesley Hallick, PhD, set out to launch a Master of Science (MS) in Medical Informatics program. OHSU had already been funded under the NLM T15 Training Grant program since 1992.

In 1996, OHSU received approval from the state of Oregon to launch our MS program. We welcomed 7 entering students, all of whom graduated in 1998 or 1999. The introductory course that was previously identified as Public Health 549 now became Medical Informatics (MINF) 510. The course was taught with 3 hours of lecture and other classroom activities once a week. Other courses were developed for the program, including clinical decision support, computer science, and organizational behavior and management. One interesting finding for me was that many of our students aspired to assume informatics careers outside of the main path for the field in the 20th century, which was as an academic researcher that was typically grant-funded. A number of them, however, saw health care institutions, industry, and other sites beginning to hire operational informatics professionals. Many were adult learners who wanted to learn more but could not quit their day jobs nor uproot their families to places like Oregon to pursue in-person education.

Initial Course

Before 1999, people were inquiring about the availability of my course and our program online. The prospect of adopting this technology was not daunting, but having very little experience and understanding of online education was. I had no formal training in education. I taught as many faculty did in the 1990s—lecturing with overhead transparencies that gave way to PowerPoint slides along with assigned readings, homework exercises, and a course project that was typically a term paper on a topic of the student's interest.

Nonetheless, after the survey of interest conducted in 1999, I decided to offer a version of my introductory course in an online format. I did not abandon the in-person classroom, but rather offered a second section of the course, with my standing up in front of class lecturing replaced with recorded lectures that were made available online. I had started to learn some about what were considered state-of-the-art approaches to online lectures for the late 20th century. One key point was to be cognizant of bandwidth, as many people still connected to the internet via a telephone modem, especially from home. As such, we needed to be careful that all materials, including the recorded lectures, could be delivered over telephone modem speeds, that is, 28-56 Kb/s modems. This precluded the use of video, high-resolution images, or other bandwidth-intensive media.

Textbox 1 shows the weekly course outline for the course at the time, which was later moved to an online format. The topics covered then are still pertinent in the present, but many new ones have come to the fore. Figures 1 and 2 show the learning management system (LMS) and lecture screenshots from the first course offering.

Textbox 1. Course outline circa 1999.

- Acquisition, storage, and use of medical data
- Medical computing
- Medical decision-making and evidence-based medicine
- The electronic medical record
- Standards, security, and confidentiality
- Information retrieval and digital libraries
- Imaging and telemedicine
- Artificial intelligence and decision support
- Computer networks and the internet
- Nursing and consumer health informatics
- Evaluation of medical informatics applications

XSL•FO
RenderX

Figure 1. Course outline in the Blackboard learning management system. MINF: Medical Informatics; OHSU: Oregon Health & Science University.

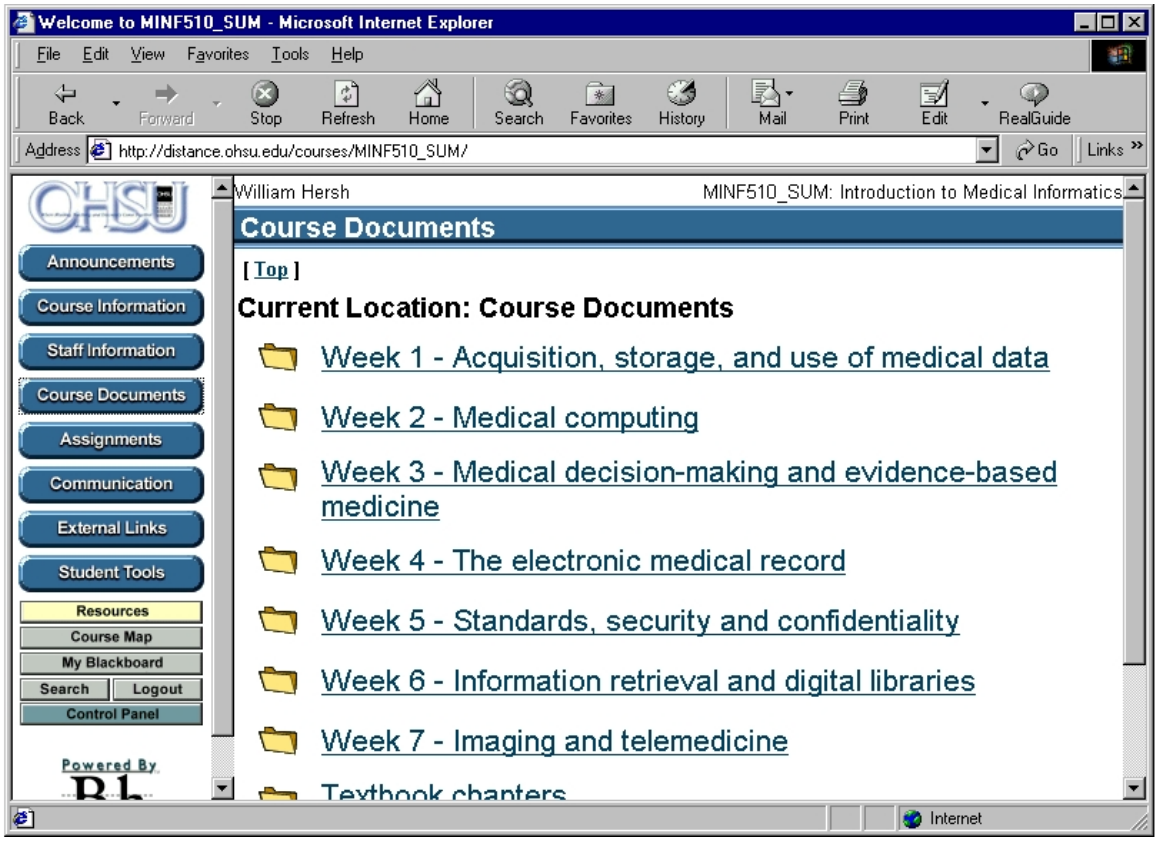

Figure 2. Lecture in RealMedia format displaying the elements of information retrieval systems. IR: information retrieval.

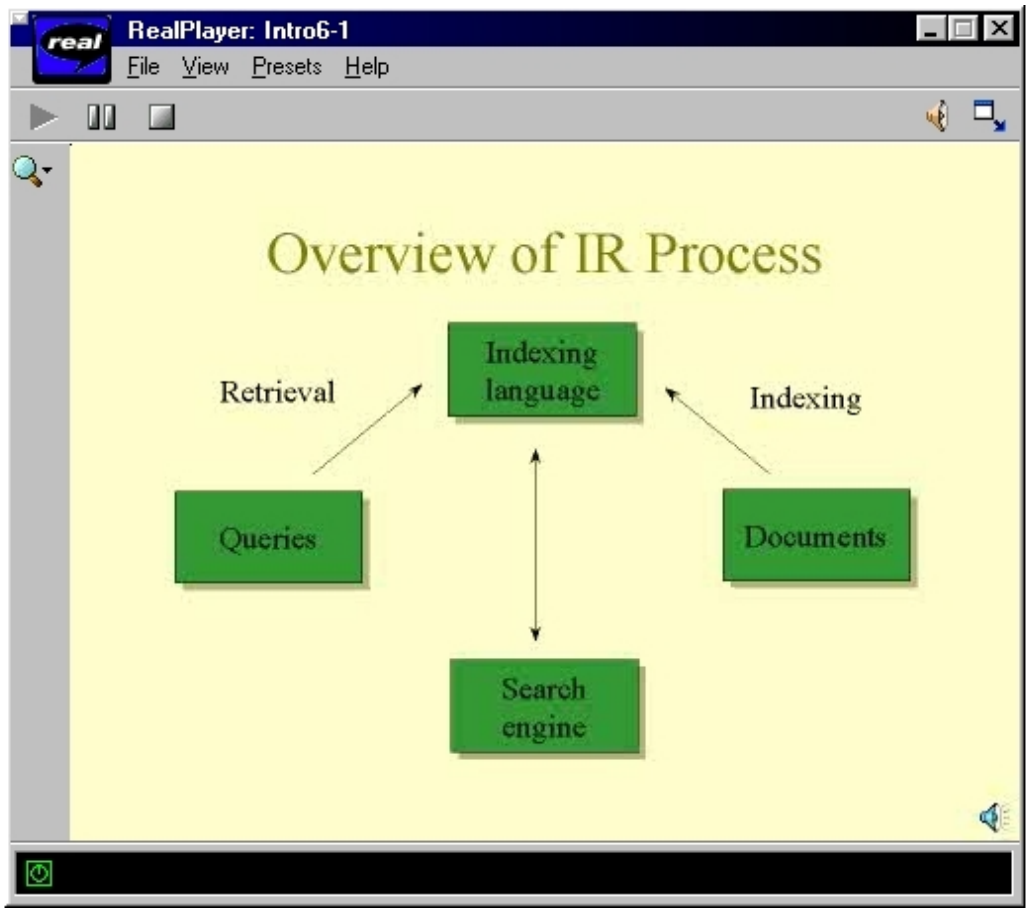

XSL•FO
RenderX

Fortunately, I already had the lecture slides for the course; therefore, I was able to add narration to them, that is, voice-over PowerPoint. Another adage at the time was that online lectures be parsed into segments typically of about 15-20 minutes each. As such, I broke my PowerPoint lecture files into segments and I soon came to learn how much narration I typically used per slide, which to this day is an average of about 90 seconds.

The most common format for the delivery of media in the early days of the World Wide Web was RealMedia (RealNetworks). Producing RealMedia files required recording the entire lecture and compiling the output using a tool called Real Presenter. Our department had no LMS before we started this journey, and the leading system at the time was Blackboard (Anthology), which we installed on a Unix server in our department. Fortunately, the hardware and software costs for this were modest at the time.

The initial course was successful by metrics of student performance on homework assignments and examinations being comparable to those of on-campus students, although they were admittedly different audiences [1]. The 15 students completing the course expressed great satisfaction and were interested in taking additional courses. Another interesting finding was that the volume of the interactive messages in the discussion forum vastly exceeded the amount of discussion that took place in the on-campus course, showing that online courses could be at least as interactive as those taught in-person.

Additional Courses and Development of Educational Programs

There was enough demand for the online course that I started offering it every academic quarter, which I have done continuously since 1999. The success of the first course led us to develop our first credential, a Graduate Certificate, which was launched in 2000 [3]. This allowed other faculty to move their courses to an online format. We initially did not think that there would be interest in or a market for a full MS program online. However, many students in the Graduate Certificate Program did express an interest in obtaining an online MS degree, thus leading to the launch of that program in 2002.

When our online MS program was launched, we believed that students should spend some time on our campus. We did this by teaching some courses in a short-course format, where the course was taught intensively for 3-5 days, often with additional activity before or after the on-campus intensive. These courses were typically offered during the summer when students had more flexible schedules and the weather was nicer in the Pacific Northwest.

I also moved other courses that I taught to an online format. I taught a course on information retrieval, the area of my research, and moved that online in 2002. I also started teaching a new course on evidence-based medicine based on the McMaster University model [4], exclusively online. Unlike the other courses, the latter only had modest amounts of online lecture material and mainly consisted of students critically appraising journal papers and presenting them to their classmates.

All through this time, we maintained our on-campus graduate program. We launched our PhD program in 2003, which funded students mainly from our NLM T15 grant. Our first PhD graduate was Adam Wright, who is now a professor in the Department of Biomedical Informatics (DBMI) at Vanderbilt University. Another graduate from the program at the time was Peter Embi, MD, MS, who is now Chair of the Vanderbilt DBMI.

Around this time, we made some other changes in our graduate program. One was to bifurcate the program into 2 tracks, which are now called majors. The original track/major is now called Health & Clinical Informatics (HCIN), while the newer track/major is now called Bioinformatics/Computational Biomedicine (BCB). At this time, we also changed the prefix name of our courses from MINF to Biomedical Informatics (BMI), that is, my introductory course became BMI 510.

Evolution of Tools and Platforms

As the internet evolved, so did the availability of new and improved tools for developing and disseminating content. In 2004, we began producing lectures in the Adobe Flash format, initially via the Adobe Presenter tool. In 2009, we moved to a different tool for producing Flash content, called Articulate Presenter. This tool had a new feature of allowing each slide to be recorded individually and then exported into a single Flash file. This turned out to be a major time-saver, as only new slides or those with updated content needed to recorded again. In addition, the output of Articulate Presenter added some additional navigational tools, such as a list of the slides and incorporation of notes, which were used for accessibility. In the mid-2010s, as the use of Flash was replaced by HTML5, Articulate Presenter allowed output in that format, allowing us to transition content from Flash to HTML5 without having to change any underlying source material.

I made a personal transition in 2012 when returning to using the Apple Macintosh after a hiatus using Microsoft Windows computers from 1995 to 2012. A big hesitancy for switching was my use of Articulate Presenter, which only ran under Windows. However, Apple had recently moved to Intel chips, with the ability to run Windows. Even though it was clunky, I was able to maintain the use of Articulate Presenter by using the virtual-machine software Parallels. Many users of Articulate Presenter, mostly academicians, requested for the software to have a native Mac version, but Articulate never developed one. Additional dissatisfaction with Articulate was caused by its movement to an expensive subscription model.

By 2021, I had another reason to move on from Articulate Presenter, which was when Apple moved from Intel to ARM (Advanced Reduced Instruction Set Computing Machines) processors. Fortunately, the slide narration and export options of PowerPoint itself had improved markedly, and I was able to produce narrated lectures exclusively with PowerPoint. I exported the slideshow as an MP4 video and uploaded the file to Echo360, which had been adopted by OHSU as a media server.

XSL•FO
RenderX

Another transition was OHSU adopting an institutional LMS in 2007. This enabled our department to not have to maintain our own LMS or server for it. The university adopted Sakai, an open-source LMS. OHSU chose the option of a vendor to run Sakai in their data center. Sakai provides most of the features seen in an LMS, including organization of courses, provision of content or links to it, multiple choice questions, discussion forums, gradebooks, and linkage to student management systems.

AMIA 10x10 Program

Another major event in my online teaching journey was the AMIA 10x10 (ten by ten) program. In 2005, when Charles Safran, MD, was President of AMIA, he was convinced that the United States needed more people, especially physicians and nurses, trained in informatics [5]. Dr Safran advocated that the United States needed at least 1 physician and 1 nurse trained in informatics in the nearly 6000 hospitals in the United States. This led him to ask a number of educational program directors, including myself, how much capacity they had to increase the size of their programs. Although most of them said perhaps they could increase 2- or 3-fold, my reply to him was “all of them.” This was because I knew that distance learning was very scalable, and with enough lead time could be scaled up with faculty to support much larger numbers of learners.

At that time, AMIA was looking to develop some sort of introductory course in BMI. However, the prices quoted to them by vendors were beyond AMIA’s means. As I already had my online introductory course from our graduate program, I proposed to AMIA and Don Detmer, MD, MS, its President and Chief Executive Officer, a repackaging of my online course. I proposed the name 10x10 based on Dr Safran’s stated need of 1 physician and nurse trained in informatics in 5000+ US hospitals and set a goal for doing so by 2010. Because the course already existed, we were able to quickly put in place a memorandum of understanding between OHSU and AMIA, which was based on an agreement of mutual nonexclusivity: OHSU would maintain the ownership of the 10x10 course content (for use in other programs, including our graduate program) and AMIA could offer other 10x10 courses. Dr Detmer later called 10x10 one of the most successful AMIA initiatives ever undertaken.

In the summer of 2005, the first offering of 10x10 was launched, with 51 people enrolled [6]. Unlike our graduate course that was only online, an in-person session would bring course participants together at the AMIA Annual Symposium. A total of 44 people completed the first offering, and almost all of them showed up for the first face-to-face session in November.

One of the participants in the original 10x10 course was Paula Otero, MD, a pediatrician from Hospital Italiano de Buenos Aires in Argentina. After the course ended, she proposed an interesting idea: translating the course into Spanish [7]. This started a very productive collaboration that resulted in the later awarding of an informatics training grant from the National Institutes of Health Fogarty International Center. This allowed Dr Otero to further develop her online offering in Spanish, while also sending faculty for informatics fellowship training to OHSU.

The 10x10 course was successful by many metrics, attracting large numbers and achieving strong satisfaction from students completing it [8]. We structured the course such that those who received a grade of B or better on an optional final examination could obtain an academic credit in our program at OHSU. This would also then allow them to enroll in the OHSU Graduate Certificate or MS program (or obtain credit and transfer it to other graduate programs). About 10%-15% of those completing the 10x10 course have gone on to further study in the field.

Health Information Technology for Economic and Clinical Health Era

As my own interest in informatics education and training grew, I began to develop a research interest in the informatics workforce [9]. Unable to find any definitive data about the workforce, I came across the HIMSS Analytics Database. Although not an optimal source of data to answer questions about the workforce, I was able to use the database to gain the first estimate of the size of the workforce and its potential growth, as hospitals moved to higher levels of electronic health record adoption [10]. In addition to presenting this work at the AMIA Annual Symposium in 2008, I also had the opportunity to present it in Capitol Hill earlier that year, which was fortuitous due to the imminent Great Recession coming, with the US government creating the American Recovery and Reinvestment Act that aimed to stimulate the economy and create jobs.

A major part of the American Recovery and Reinvestment Act was the Health Information Technology for Economic and Clinical Health (HITECH) Act, with investment over US $30 billion to facilitate the adoption and meaningful use of the electronic health records [11]. In part because of my workforce research, a US $118 million workforce development program was made part of the HITECH investment. OHSU played a large role in the grants that were competitively awarded in 2010 by the workforce development program, including serving as the National Coordination and Dissemination Center for the health information technology curriculum that was funded for development [12].

During and after HITECH, I continued to provide leadership around informatics education and how it advanced other careers in the field. I was also a leader in the new clinical informatics physician subspecialty [13], being appointed by AMIA to direct the Clinical Informatics Board Review Course (CIBRC), which was offered in time for the first board examination in 2013. I recruited an excellent team to help teach the course, including Thomas Payne, MD; Bimal Desai, MD; and Diane Montella, MD. As I was eligible for the board examination myself (even though I was no longer clinically active, I had completed my internal medicine residency in the era of lifetime certification), I took the examination and even passed! The next year, I laid the groundwork at OHSU to establish one of the first 4 Accreditation Council for Graduate Medical Education–accredited fellowships for the new subspecialty,

XSL•FO
RenderX

which launched in 2015 under the leadership of Vishnu Mohan, MD [14].

The Last Frontier: Medical Education

Although I had been interested in informatics education for others besides informaticians dating to the 1990s, I was never able to make headway in obtaining any significant content into the OHSU Doctor of Medicine (MD) educational program. Finally, in 2012, with the arrival of a new supportive Senior Associate Dean for Medical Education, George Mejicano, MD, MS, the door was opened for me and my colleagues in our department. We were also aided by being one of the 11 institutions awarded grants by the American Medical Association to accelerate change in medical education. Our work in clinical informatics education for medical students led to a 2014 paper that laid out the competencies for medical students, although as noted in the paper, could actually be applied to all health care students and professionals [15]. Along with colleagues at OHSU, we began to implement informatics education in the MD curriculum. We have started working with local colleagues in nursing, biomedical basic sciences, and public health to add informatics to their curricula.

Another medical education opportunity arose at the start of the COVID-19 pandemic. As the pandemic resulted in many medical students being displaced from ward and clinic rotations, there was a need at OHSU and other medical schools to find virtual education opportunities. It was easy to quickly adopt the introductory course curriculum to a medical school-style block course. Some colleagues from other medical schools learned of the availability of the online course and asked if it could be used at their institutions. In the first few months of the pandemic, the course was delivered to both OHSU students (3 offerings to a total of 44 students) and non-OHSU students (8 offerings to a total of 178 students). The excess workload required me to stop offering the course outside OHSU, but the course continues as a medical student elective at OHSU.

Learners Served

I have been gratified that my online teaching has attracted so many learners over the years. Since the inception of the introductory course (MINF/BMI 510) in 1996, 1683 students have completed the course in its on-campus and online versions. Since the introduction of the AMIA 10x10 course in 2005, 3253 people have completed the OHSU offering of the course through early 2024. The medical student elective at OHSU continues to attract about 30 medical students per year. The OHSU BMI graduate program has 984 alumni, with 499 Graduate Certificates, 444 MS degrees, and 41 PhD degrees awarded.

There are now about 3000 physicians who are board-certified in the clinical informatics subspecialty. A large majority of those passing the exam have taken the CIBRC. I was gratified in 2016 when attending a reception for physicians at the Epic Advanced User's Group meeting in Madison, WI, and finding that at least half of those I spoke with had taken one of the above educational experiences I had developed. It is not infrequent to find myself walking through an airport or another venue and have someone introduce themselves as having taken a course of mine, usually 10x10.

Although the main 10x10 course has been offered through AMIA, we have partnered with other organizations to offer versions tailored to individuals within them. The online curriculum content does not change, but rather, the student composition, discussion, related activities, and in-person sessions are modified to meet the needs of those organizations. Table 1 shows the partnerships with the number of course offerings and students completing them for both United States and international partners. The most prominent international collaboration has been with Gateway Consulting of Singapore, and the course continues to be cotaught with KC Lun, PhD.

XSL•FO
RenderX

Table 1. Number of 10x10 offerings and students completing the course with United States and international partners.

Organization	Offerings (n=114), n	Completed (N=3253), n
United States		
American Medical Informatics Association (AMIA)	48	2159
American College of Emergency Physicians (ACEP)	16	237
American College of Physicians (ACP)	1	25
Association of Nutrition and Dietetics (AND)	7	126
Centers for Disease Control (CDC)	1	18
California Healthcare Foundation (CHCF)	1	16
Mayo Clinic	2	87
New York State Academy of Family Physicians	3	22
Scottsdale Institute (SI)	1	15
Society for Technology in Anesthesiology (STA)	1	5
International		
Abu Dhabi Health Services (SEHA)	1	54
Gateway Consulting, Singapore	27	395
Israel Ministry of Health	1	11
King Saud University (KSU), Saudi Arabia	4	83

Lessons Learned

My experience in developing all of the above educational offerings have resulted in many lessons learned. Although I ventured into online learning with little formal training, I discovered over time what works well for the kinds of teaching I prefer. Although academic lectures receive their share of criticism, I find that an engaging speaker can provide a highly effective learning experience, especially if he or she explains the big picture and fills in the necessary details. I know that PowerPoint also has its critics, yet it too can be highly effective, and perhaps more so in an asynchronous setting where lectures can be paused or reviewed. As such, my main teaching modality has always been lectures using voice-over PowerPoint slides. A typical 3 hours' worth of lecture is segmented into 6-9 lectures. I also provide students with PDF handouts of the slides and an exhaustive list of references cited in the slides. The homework in the course consists of 10 multiple choice questions per unit. In the questions, I aim to require students to apply the material. The course does not have a required textbook, although I do tell students that it follows the contents of a textbook that I edited [16].

I have also learned that online teaching does not deserve what many perceive as lacking interaction. I have always taught from an LMS that featured discussion forums and advise students to think of such discussion forums as the online equivalent of a classroom. Students are encouraged to speak up, not feel intimidated, and remember that everyone has something valuable to say. In the introductory course, I seed the discussion with 1-2 questions but encourage students to also post their own questions, including asking about things they do not understand (rather than emailing me, to which my reply is usually to post their questions on the forum).

Another lesson learned is to advise students at the outset to follow some simple etiquette for the discussion forums. Messages should be neither too short nor too long. Everyone should be constructive and respectful. Students should reply to messages in their respective threads so that everyone can see the evolving discussion. Students should not copy and paste from websites but rather use their own words and provide a link if desired. They should also not discuss homework multiple choice questions until 1 week after the due date.

In addition, I have learned to lay out what I consider to be my expectations of students. They should complete the lectures and participate in the discussion. They should not be afraid to speak up in the discussion forums. They should ask questions about anything that is unclear in lectures or other materials. Most of all, they should feel free to challenge the instructor. Students should complete all assignments by the due dates. I do allow them to occasionally complete assignments late but warn them not to fall too far behind, since they will have difficulty in catching up.

By the same token, I tell students that they should have expectations for me. They should expect me to create an environment of learning and objective inquiry. I should maintain high availability, replying to emails as quickly as I can. They should expect that I am there to serve them, as students are not wasting my time. The best method of initial contact is email, and we can talk via video or phone as needed.

For lectures, students should expect the quality to be very good although not perfect. I am not a talking head and try to convey

XSL•FO
RenderX

my view of informatics, getting into the details but never losing the big picture. One of my best compliments ever came in a course evaluation from a student who said, “I like that Dr Hersh pauses and makes mistakes and corrects himself … It shows he is thinking about what he is saying instead of reading off a paper” (even though I have more recently adopted using a script to make sure I cover everything on each slide). Students should also expect that in the discussion forums, I should read all postings even if I cannot reply to each post. I usually try to reply to threads where dialogue has developed and reply to different students and not the same ones each time.

An additional lesson learned is that online curricula, like any course curricula, need time and effort to be maintained. Fortunately, the large audience for my introductory course gives me the time and resources to keep the curriculum up-to-date. My basic approach is to update the course content once a year, refreshing the existing content and adding new material. This demonstrates a challenge for materials developed for posting online, such as the HITECH curriculum described above [12], that is, static content becomes outdated quickly. As such, the content soon diminishes in value for teaching.

Another lesson learned has been that my online materials prove very useful for in-person teaching and classes. This is typically done using a flipped classroom approach, where students listen to recorded lectures before coming to in-person class to discuss concepts, ask questions, give presentations, and more. Since moving my teaching to online lectures, it has been extremely rare for me to stand in front of a classroom and give a traditional lecture.

A final lesson learned is that online education is quite scalable. It requires a substantial fixed effort to develop and maintain content, although it has a very low marginal effort to provide that content to additional learners. Fortunately, the collection, curation, and maintenance of course materials is something I enjoy, even if the time commitment is substantial. However, once the content is produced, the additional work of developing new courses by repurposing it is modest.

Conclusions

The seemingly small decision to start offering my introductory informatics course in an online format had far-reaching implications for my academic informatics career. In the early days, many educational traditionalists scoffed at the notion of teaching online. But in modern times, education has had to adapt to different learners. Especially in knowledge-based fields, we no longer stop our education with obtaining a degree. Many professionals change direction as they move through their careers, with a common example being health care and other professionals moving into fields such as informatics. For many, the ability to learn online and asynchronously offers educational opportunities that cannot be met by traditional in-person classrooms. Figure 3 highlights the major course or program milestones by year.

Figure 3. Major milestones by year. AMIA: American Medical Informatics Association; HITECH: Health Information Technology for Economic and Clinical Health.

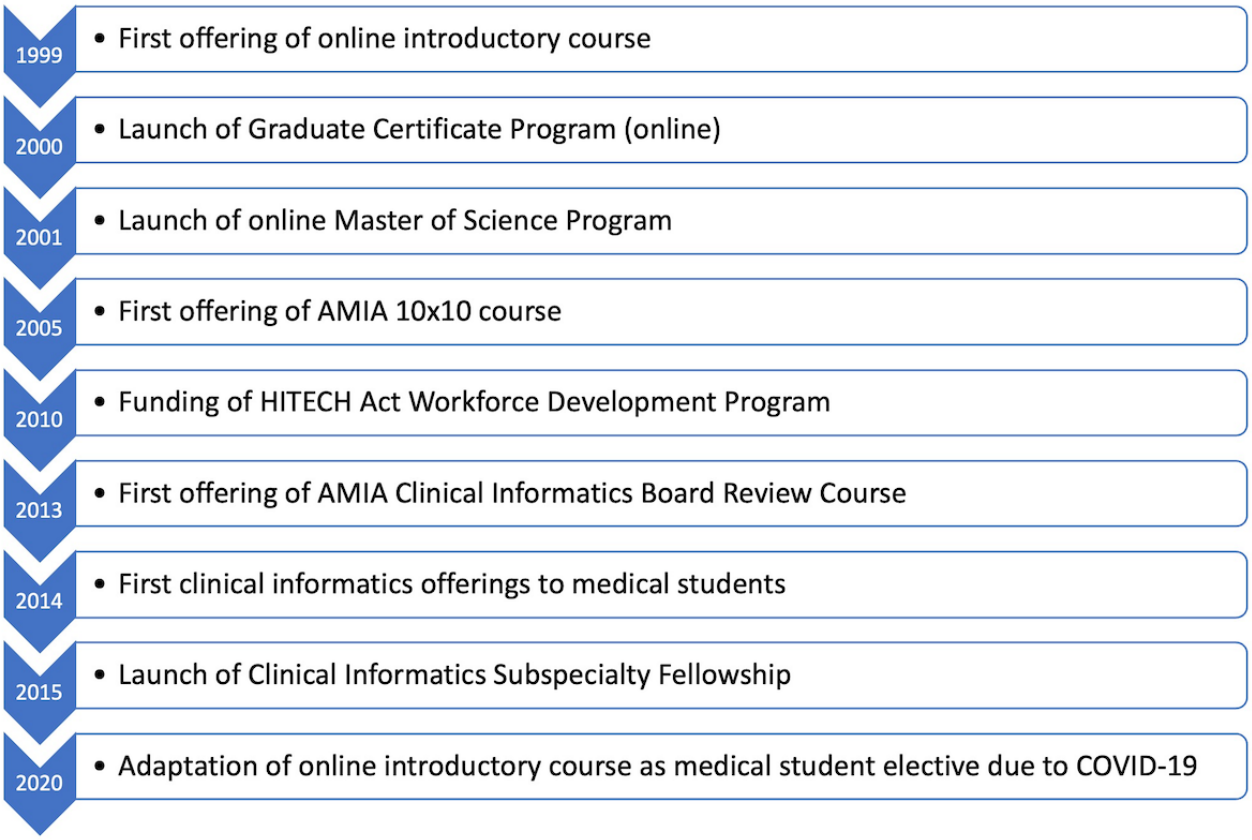

XSL•FO
RenderX

Acknowledgments

The faculty and staff of the Oregon Health & Science University Department of Medical Informatics & Clinical Epidemiology have been highly supportive of developing and implementing online education in biomedical and health informatics. Important staff contributors have been Andrea Ilg, Lauren Ludwig, Lynne Schwabe, Vanessa Reeves, Kathryn Pyle, Anne Marshall, Diane Doctor, and Virginia Lankes. Also critical in the success of this effort have been various teaching assistants over the years, including Katherine Gorris, Amy Carter, Kelly Brougham, and Kate Fultz Hollis.

Conflicts of Interest

None declared.

References

1. Hersh WR, Junium K, Mailhot M, Tidmarsh P. Implementation and evaluation of a medical informatics distance education program. J Am Med Inform Assoc 2001;8(6):570-584 [FREE Full text] [doi: 10.1136/jamia.2001.0080570] [Medline: 11687564]
2. Ash J, Hersh W, Krages K, Morgan J, Schumacher R. The Oregon IAIMS: then and now. Bull Med Libr Assoc 1999 Jul;87(3):347-349 [FREE Full text] [Medline: 10427438]
3. Hersh W. Methods Inf Med 2018 Jan 24;46(01):80-83. [doi: 10.1055/s-0038-1628138]
4. Straus S, Glasziou P, Richardson W, Haynes R. Evidence-Based Medicine - How to Practice and Teach EBM. Amsterdam, Netherlands: Elsevier; 2018.
5. Safran C. Informatics training for clinicians is more important than hardware and software. Yearb Med Inform 2018 Mar 07;18(01):164-165. [doi: 10.1055/s-0038-1638657]
6. Hersh W, Williamson J. Educating 10,000 informaticians by 2010: the AMIA 10x10 program. Int J Med Inform 2007;76(5-6):377-382. [doi: 10.1016/j.ijmedinf.2007.01.003] [Medline: 17296325]
7. Otero P, Hersh W, Luna D, González Bernaldo de Quirós F. A medical informatics distance-learning course for Latin America. Methods Inf Med 2018 Jan 17;49(03):310-315. [doi: 10.3414/me09-02-0009]
8. Feldman SS, Hersh W. Evaluating the AMIA-OHSU 10x10 program to train healthcare professionals in medical informatics. AMIA Annu Symp Proc 2008 Nov 06;2008:182-186 [FREE Full text] [Medline: 18999199]
9. Hersh W. Who are the informaticians? What we know and should know. Journal of the American Medical Informatics Association 2006 Mar 01;13(2):166-170. [doi: 10.1197/jamia.m1912]
10. Hersh W, Wright A. What workforce is needed to implement the health information technology agenda? Analysis from the HIMSS analytics database. AMIA Annu Symp Proc 2008 Nov 06;2008:303-307 [FREE Full text] [Medline: 18998990]
11. Blumenthal D. Implementation of the Federal Health Information Technology initiative. N Engl J Med 2011 Dec 22;365(25):2426-2431. [doi: 10.1056/nejmsr1112158]
12. Mohan V, Abbott P, Acteson S, Berner ES, Devlin C, Hammond WE, et al. Design and evaluation of the ONC health information technology curriculum. J Am Med Inform Assoc 2014;21(3):509-516 [FREE Full text] [doi: 10.1136/amiajnl-2013-001683] [Medline: 23831832]
13. Detmer DE, Shortliffe EH. Clinical informatics: prospects for a new medical subspecialty. JAMA 2014 May;311(20):2067-2068. [doi: 10.1001/jama.2014.3514] [Medline: 24823876]
14. Longhurst C, Pageler N, Palma J, Finnell J, Levy B, Yackel T, et al. Early experiences of accredited clinical informatics fellowships. J Am Med Inform Assoc 2016 Jul;23(4):829-834 [FREE Full text] [doi: 10.1093/jamia/ocv209] [Medline: 27206458]
15. Hersh WR, Gorman P, Biagioli F, Mohan V, Gold J, Mejicano G. Beyond information retrieval and electronic health record use: competencies in clinical informatics for medical education. AMEP 2014 Jul:205. [doi: 10.2147/amep.s63903]
16. Hersh W, editor. Health Informatics - Practical Guide, 8th Edition. Morrisville, NC, USA: Lulu.com; 2022.

Abbreviations

AMIA: American Medical Informatics Association
ARM: Advanced Reduced Instruction Set Computing Machines
BCB: Bioinformatics/Computational Biomedicine
BMI: Biomedical Informatics
CIBRC: Clinical Informatics Board Review Course
DBMI: Department of Biomedical Informatics
HCIN: Health & Clinical Informatics
HIMSS: Healthcare Information and Management Systems Society
HITECH: Health Information Technology for Economic and Clinical Health

XSL•FO
RenderX

LMS: learning management system
MD: Doctor of Medicine
MINF: Medical Informatics
MS: Master of Science
NLM: National Library of Medicine
OHSU: Oregon Health & Science University

Edited by G Eysenbach; submitted 01.04.24; peer-reviewed by C Lokker, J Xia; comments to author 29.05.24; revised version received 01.06.24; accepted 16.07.24; published 06.08.24.

Please cite as:
Hersh W
A Quarter-Century of Online Informatics Education: Learners Served and Lessons Learned
J Med Internet Res 2024;26:e59066
URL: https://www.jmir.org/2024/1/e59066
doi:10.2196/59066
PMID:

©William Hersh. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Second Life Metaverse and Its Usefulness in Medical Education After a Quarter of a Century

Francisco Sendra-Portero[1], MD, PhD; Rocío Lorenzo-Álvarez[2], MD, PhD; Teodoro Rudolphi-Solero[1], MD, PhD; Miguel José Ruiz-Gómez[1], PhD

[1]Department of Radiology and Physical Medicine, Facultad de Medicina, Universidad de Málaga, Málaga, Spain

[2]Department of Emergency and Intensive Care, Hospital de la Axarquía, Vélez-Málaga, Spain

Corresponding Author:
Francisco Sendra-Portero, MD, PhD
Department of Radiology and Physical Medicine
Facultad de Medicina
Universidad de Málaga
Bvd. Luis Pasteur, 32. 20071
Málaga, 29071
Spain
Phone: 34 952131653
Email: sendra@uma.es

Abstract

The immersive virtual world platform Second Life (SL) was conceived 25 years ago, when Philip Rosedale founded Linden Lab in 1999 with the intention of developing computing hardware that would allow people to immerse themselves in a virtual world. This initial effort was transformed 4 years later into SL, a universally accessible virtual world centered on the user, with commercial transactions and even its own virtual currency, which fully connects with the concept of the metaverse, recently repopularized after the statements of the chief executive officer of Meta (formerly Facebook) in October 2021. SL is considered the best known virtual environment among higher education professionals. This paper aimed to review medical education in the SL metaverse; its evolution; and its possibilities, limitations, and future perspectives, focusing especially on medical education experiences during undergraduate, residency, and continuing medical education. The concept of the metaverse and virtual worlds was described, making special reference to SL and its conceptual philosophy, historical evolution, and technical aspects and capabilities for higher education. A narrative review of the existing literature was performed, including at the same time a point of view from our teaching team after an uninterrupted practical experience of undergraduate and postgraduate medical education in the last 13 years with >4000 users and >10 publications on the subject. From an educational point of view, SL has the advantages of being available 24/7 and creating in the student the important feeling of “being there” and of copresence. This, together with the reproduction of the 3D world, real-time interaction, and the quality of voice communication, makes the immersive experiences unique, generating engagement and a fluid interrelation of students with each other and with their teachers. Various groups of researchers in medical education have developed experiences during these years, which have shown that courses, seminars, workshops and conferences, problem-based learning experiences, evaluations, teamwork, gamification, medical simulation, and virtual objective structured clinical examinations can be successfully carried out. Acceptance from students and faculty is generally positive, recognizing its usefulness for undergraduate medical education and continuing medical education. In the 25 years since its conception, SL has proven to be a virtual platform that connects with the concept of the metaverse, an interconnected, open, and globally accessible system that all humans can access to socialize or share products for free or using a virtual currency. SL remains active and technologically improved since its creation. It is necessary to continue carrying out educational experiences, outlining the organization, objectives, and content and measuring the actual educational impact to make SL a tool of more universal use.

(J Med Internet Res 2024;26:e59005) doi:10.2196/59005

KEYWORDS

medical education; medical students; postgraduate; computer simulation; virtual worlds; metaverse

XSL•FO
RenderX

The Metaverse, Virtual Worlds, and Second Life

The Metaverse: Concept and Evolution

The term *metaverse* appears for the first time in Neal Stephenson's novel *Snow Crash*, published in 1992 [1], which recreates a fictional virtual world reproduced by a computer in which users interacted with each other and with the elements of the world through a representation of themselves called an avatar. Cinema has popularized other examples of virtual worlds parallel to the real one in the *Matrix* tetralogy (1999-2021), written and directed by Lana and Lilly Wachowski, or the film *Ready Player One* (2018), directed by Steven Spielberg. The truth is that the possibility of reproducing 3D virtual environments in which the user can enter and interact with everything that exists in them is technologically feasible today, largely thanks to the development of video games for consoles and computers and more recently through the dissemination of massively multiplayer online role-playing games (MMORPGs), with thousands of young, and not so young, users worldwide.

Conceptually, the metaverse is an open system, a vast virtual world that can be accessed simultaneously by millions of people through highly customizable avatars and powerful experience creation tools integrated with the offline world through its virtual economy and external technology [2]. The metaverse has been defined as an interconnected 3D virtual environment in which people from all over the world can come together to share social experiences [3], "a computer-generated universe" in which people immerse themselves in experiences rather than simply observing them [4].

The term *metaverse* has gained great popularity recently (Figure 1) after Mark Zuckerberg, chief executive officer of the social network Facebook, announced in October 2021 his company's interest in developing the metaverse as a social environment, apparently as an area of interaction and commercial transactions. As a result, there is currently a growing interest in all matters related to the metaverse; virtual worlds; and their applicability to today's society, including education.

Figure 1. Google Scholar search graph by year for the terms virtual worlds, Second Life, and metaverse. Below the horizontal axis, the search strings used are shown.

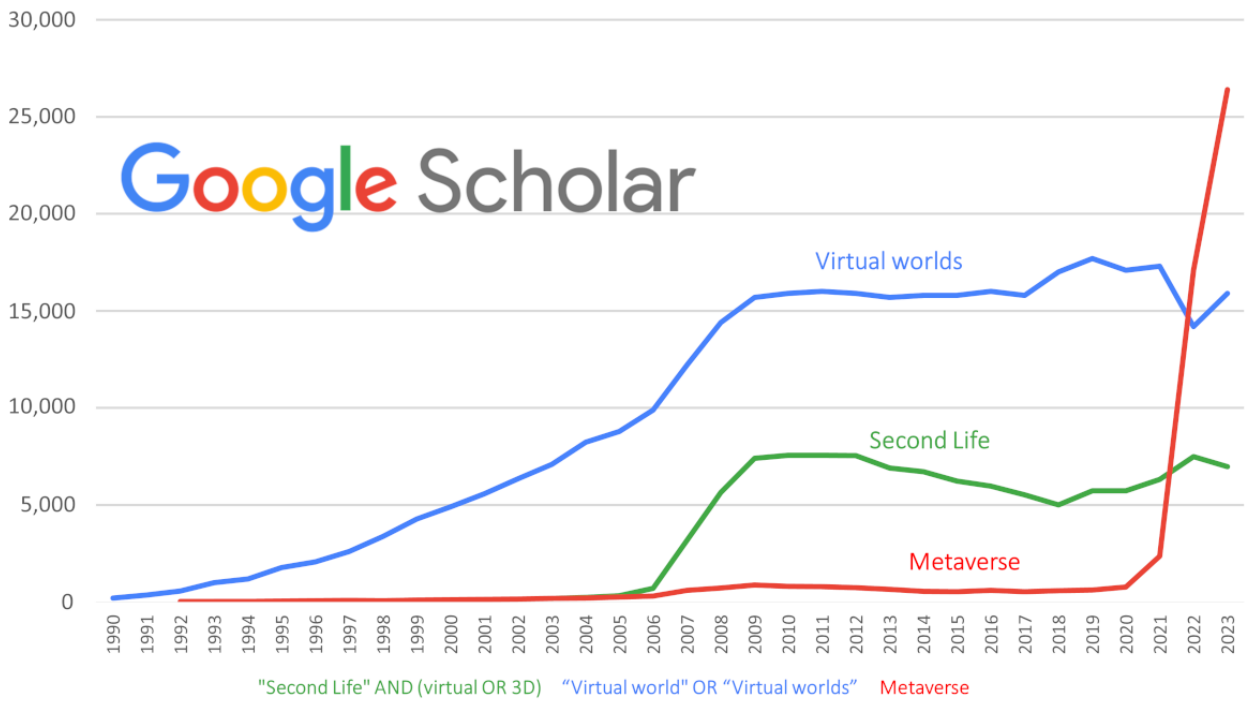

Publications about the metaverse include various technological approaches that can lead to some confusion, such as augmented reality (AR), virtual reality (VR), extended reality (XR), or virtual worlds [3,5]. AR is a technology that increases our perception of reality by adding digital information to the real world; VR allows users to experience a completely simulated environment as if it were real; mixed reality is a hybrid of the physical and virtual worlds in which digital objects interact with the physical environment to create an even more immersive experience; and XR is a general term to refer to AR, VR, and mixed reality [5].

Some differences have been pointed out between the metaverse and XR technologies [6]. The concept of the metaverse emphasizes sharing the experience, interacting with other avatar-mediated humans; thus, a student learning by means of XR cannot be considered an example of a metaverse per se [7]. The metaverse is a new method of social connection, a virtual world in which people can autonomously participate in social activities such as meetings, discussions, collaborations, and games [8] and that provides people with unique learning opportunities [9,10]. In a similar way, virtual worlds and VR, while emerging from similar roots, are not interchangeable terms. VR requires a special interface with the electronic

XSL•FO
RenderX

space—head-mounted devices with glasses and headphones—as well as hand controllers to interact within a virtual space. VR pays little attention to what is happening within the virtual world by other people or the community [11]. A VR platform can even be designed to present the context in a more visually attractive and *realistic* way to a single user, which completely distances it from the concept of a metaverse.

Virtual Worlds

Virtual communication has become an integral part of the 21st-century culture [12]. The application of communication technologies has contributed to the development of virtual worlds where users interact without a specific script and even have the freedom to develop the contents of the world and the objectives to which they want to dedicate themselves [3]. Virtual worlds allow participants to interact with contents and visitors as much as they do in real life but without having to physically move to a common place in which to establish such interaction. The range of activities that can be developed in these environments is practically unlimited and depends fundamentally on the ability of the developers of these malleable spaces to transform them into a particular environment [13].

A virtual world is a synchronous and persistent network of people represented by avatars facilitated by networked computers [14], a shared space that hosts a web-based community [15,16]. The main characteristics of virtual worlds have been described as follows [17]: (1) a shared space in which multiple users can participate at the same time; (2) a graphical user interface, which visually represents the virtual space; (3) immediacy or real-time interaction; (4) interactivity as users have a certain degree of control over the content; (5) persistence (it is an always "active" platform); and (6) socialization and community as in-world social groups, gatherings, and neighborhoods are encouraged. A virtual world is also a 3D virtual space created by the site manager for a specific purpose (ie, a collaborative workspace, a classroom, a virtual laboratory, a playground, a discussion forum, or a meeting place with friends). Therefore, virtual worlds can become virtual academic places where users, students, and teachers enter and interact by means of an avatar, being able to speak, chat, and visualize various educational contents.

The roots of virtual worlds can be found in video games and social networks [18]. Although the development of virtual worlds has reached a high level of use through simulation games, the former should not be assimilated to games. A virtual world is not an MMORPG, that is, a "game"; however, an MMORPG can exist within a virtual world [19]. Even though many successful multiplayer games are carried out in virtual worlds, their possibilities are unlimited and not restricted to the world of gaming. In this sense, one of the activities that has been successfully tested in virtual worlds is basic education in the fields of science and arts [20-22]. Some examples of virtual worlds are Active Worlds, Sinespace, Instant Messaging Virtual Universe, OpenSimulator, Sansar, and Second Life (SL) [17]. The latter has been described as a virtual world with great educational possibilities in higher education [21,22] and is the subject of this paper.

SL Virtual World

A Story of a Quarter of a Century

The immersive virtual world platform SL was conceived 25 years ago, when Philip Rosedale founded Linden Lab in 1999 with the intention of developing computing hardware that would allow people to immerse themselves in a virtual world. This initial effort was transformed 4 years later into SL, a universally accessible virtual world centered on the user, with commercial transactions and even its own virtual currency. SL is one of the most well-known and used virtual worlds, a virtual community created or developed by its own users. SL is not a game but a multiuser virtual environment; it has always tried to distance itself from the video game universe [2]. From its beginning, it has been presented as an open world with no explicit objectives, quests, or missions, explicitly and flatly stating the following: "If you can imagine it, you can do it here" or "You choose your own goals" [2].

SL experienced a huge rise in popularity. Until approximately 2008, virtual worlds such as SL were all the rage, with headlines trumpeting the life-altering potential of these avatar-occupied worlds. Philip Rosedale considered "alternative universes" such as SL a new medium of human expression that could be the "next big thing for the Internet" [17]. However, by approximately 2010, SL and other virtual worlds lost their former appeal, replicating the Gartner Hype Cycle, a graphical representation of a common pattern that emerges with each new technology or other innovations. The cycle begins with overenthusiasm followed by a period of disappointment, reaching a plateau of slow growth when the relevance and role of the innovation are finally understood. The lack of attraction among internet users could be due to several factors [17]: (1) the learning curve necessary to acquire even a modest level of competence in the virtual world; (2) time being a limiting factor, so the hours invested in "living a second life" reduce the 24 hours available for real-life activities; and (3) the technical sophistication of the platform in relation to other social interaction offerings that appeared disruptively in those years, such as Facebook, Instagram, Twitter, Snapchat, and others. In short, too many easier-to-use alternatives appeared on the scene 15 years ago. The current position of SL (and other virtual worlds) in the Hype Cycle is an open question. Several studies suggest that it is on a "productivity plateau," which implies stable but unspectacular growth over time [17]. An internet search may give an idea of this (Figure 1).

In 2020, Linden Lab was purchased by the Waterfield Group, an investment firm that buys companies that obtains decent earnings and helps them remain profitable in perpetuity [2]. That same year, the COVID-19 pandemic put the world under quarantine, and many people spent their quarantine in virtual worlds, eager to connect with others. SL saw a surge in old and inactive users as growing awareness of virtual worlds fueled renewed and growing interest in the metaverse.

Now, 25 years after the foundation of Linden Lab, this virtual world seems to still have some weight, especially in the world of creation and graphic design [23]. In 2021, it was reported that there were still approximately 64.7 million active users on SL, and Web Tribunal stated that, by 2023, this had risen to 70

XSL•FO
RenderX

million accounts, with a daily average of 200,000 users in 200 countries [24]. SL continues to be successful as a virtual world model that connects with the concept of an equitable, diverse, and creative metaverse. Its creativity tools are open and powerful, and its user community is diverse and made up of members free to create endless values and possibilities [2].

What Is SL?

From a technical point of view, the SL architecture is based on a client-server model. The graphical user interface runs locally, whereas the 3D virtualization runs on servers owned by Linden Lab. The visual experience is presented in real time, but the objects are stored remotely. This makes it easier for the user to create content, but on the other hand, it puts a lot of pressure on the user's graphical capabilities and bandwidth [25].

From the user's experiential perspective, the visual and physical realism of the virtual space and the interaction and communication with others combine to produce a deeply immersive experience. Immersion is the sensation of being so situated within the 3D virtual world that awareness of the environment beyond the digital screen practically disappears [2]. It contributes to creating a feeling of "being there" and a strong sense of copresence when other avatars are present.

From a social point of view, SL is a virtual community created or developed by its own users. Through the avatars, one has the feeling of being part of that world and that others can perceive it at the same time, the feeling of social presence on a virtual global platform currently shared by diverse generations, including baby boomers, Millennials, and Generations X and Z [2,17]. Within SL, there are multiple communities with their own spaces that recreate scenarios designed by the users themselves. These communities have their own economy based on a virtual currency, the Linden Dollar, with which users can buy and sell virtual material, make transactions, pay for services, and exchange other currencies [26] (1 Linden Dollar is equivalent to €0.0029 or US $0.0031 as of March 27, 2024). The importance of this economic system is such that it has required the development of a virtual monetary policy that manages resources and how to carry out transactions, which can be done in a common market ("Marketplace") or directly between users [27]. Within the communities, there are also a series of rules, which in case of noncompliance can lead to the blocking or deletion of the user account. These rules prohibit intolerance, harassment, aggression (even virtual), slander, indecency, and disturbance of the peace [28].

How It Works

Information related to SL can be found on its web page [29]. To access SL, the user must create an account, choose an avatar, and download the SL viewer [30] or any other SL-compatible viewer [31] to display the virtual world on their computer screen. The use of SL is free for users (although a premium account can be purchased for US $9 per month) but not for landowners and administrators, who have to pay different rates according to the size of the land.

The SL virtual world imitates the real world that we know. It consists of interlinked regions that contain land, water, and sky. A region is an entire island that is 256 × 256 m, or 65,536 m^2. Almost all the objects that are visible in SL are built from 3D geometric primitive shapes called prims. Each region has an allotment of 15,000 prims. Prims can assume any desirable shape, and one can make prims look any way wanted by applying selected textures to their surfaces. They can be given certain qualities and features (such as transparency or the ability to flex or bend with the wind), they can be linked together, and they can be made to do things through a script written in Linden Scripting Language [32]. Prims also allow for sharing media (ie, displaying a web page or a linked video on a face).

The avatar is the representation of the user inside SL, and by means of it, they can interact with the virtual world. Interactions include viewing the world with different perspective and focus, touching objects, answering to displayed menus originated from Linden Scripting Language scripts, moving or adopting gestures (walking, running, flying, teleporting, sitting, or lying down), and communicating with others (receiving and sending audio, chat, instant messages, and note cards). At launch, SL avatars were human by default but not ultrarealistic. The internal prim creation tools encouraged the construction of avatar attachments, which led to a wide variety of avatar types and creative environments to explore. The arrival of mesh in SL in 2010—high-resolution 3D files created in offline software and then uploaded into the virtual world—greatly changed this dynamic. Thanks to mesh and other graphics enhancements, SL avatars and environments now look as detailed and as vivid as those from top video games [2].

Communication between avatars can be done through voice chat, written chat, and note cards. Voice chat gives a very important sense of presence; it gives the user the perception that they are in front of other human beings and allows them to perceive the nuances of verbal communication. Its use can be public (everyone within a predetermined distance can hear it) or private (the conversation is exclusive to an avatar or group of avatars) but only alternatively [33].

Written chat can also be public (local chat) or private (instant message). Local chat can be read by everyone nearby, is great for asking short-answer questions to the audience (yes and no or true and false), and allows attendees to ask questions to the presenter without interrupting the presentation. Instant messages are sent from one avatar to another, who receives it immediately if they are logged in SL or as soon as they log in to SL. Note cards are messages that are sent in SL and are stored in the inventory of the avatar that receives it, recording the date and time of sending and avatar of origin. In educational activities, they are very useful to collect information from students as proof of their attendance or to complete a task or an exam (open answer or multiple choice), requesting that they send a note card to the teacher's avatar [33].

In essence, through the SL environment, we can provide specific contents to users' avatars (students, teachers, or the public in general), such as text, slide presentations, or other multimedia content. Moreover, we can also communicate verbally in a synchronous way (in real time) with other avatars, creating a meeting- or classroom-like environment in which we can give (or attend) a talk or organize a discussion group on a specific topic. This paper aims to review medical education in the SL

XSL•FO
RenderX

metaverse, its evolution, its possibilities and limitations, and future perspectives. In the following sections, general aspects of SL as an educational tool are first addressed, including its contribution to higher education, and later, the contributions to medical education are analyzed.

SL as an Educational Tool

Learning Theories and Virtual Worlds

Learning theories are constructs that explain how human beings acquire knowledge, values, skills, behaviors, attitudes, and aptitudes in a systematic, dynamic, and progressive way through different training techniques such as study, reflection, experimentation, or teaching [34]. The most important theories currently considered are behaviorist theory, cognitivist theory, and constructivist theory, and all 3 are integrated into the use of virtual worlds such as SL [35,36]. The *behaviorist theory* focuses on the establishment of observable responses through the student's behavior based on a series of stimuli received during learning, which makes it necessary to use small check points to achieve the behavioral objective [35]. Virtual worlds have been described as a suitable tool to effectively stimulate students to perform their tasks better. These environments, which can present a playful atmosphere, allow a series of rewards or scoring systems to be established, which motivates students to carry out their learning [37]. The *cognitivist theory* describes the mental process that takes place in the students and how they can apply what they learn in new experiences [35]. Virtual environments such as SL allow for the development of experiential experiences in which the student can carry out this transfer of knowledge. The *constructivist theory* holds students responsible for their own training, providing them with the necessary tools to create their own learning methodology, focusing on students' relationship with each other and with educators [35]. Virtual worlds, and particularly SL, are social environments in which students can participate in activities with a constructivist approach, including learning experiences that imitate the real world, interacting live with classmates and teachers. This dynamic allows students to process the information received and construct meaning actively instead of receiving information passively [36].

Learning is based on the passage of knowledge from working memory (short term) to long-term memory [38,39]. This process is analyzed using a factor known as cognitive load, defined as the mental effort from the cognitive system necessary to perform a task [40,41]. The complexity of the tasks, as well as their organization and the presentation of information imposed on learners, can hinder the function of short-term memory in a given task. The intellectual complexity of the tasks is known as intrinsic cognitive load, whereas their presentation is known as extraneous cognitive load [42]. When teaching through virtual worlds, it is essential to consider the extraneous cognitive load as, if the student does not have the ability to function easily in digital systems, their learning may be negatively affected [41].

SL From a Higher Education Perspective

Overview

Virtual worlds have a remarkable potential for effective teaching and learning [43,44]. These experiential environments bring the possibility to create immersive, realistic, and engaging web-based events that can provide high-quality medical education to health-related users in remote locations [45-47].

SL is considered the most popular virtual world among educators and the most used in higher education [25,48-51], particularly in the education of medical professionals [43,52]. It can be seen from the literature that most of the studies on the application of virtual worlds in higher education are based on SL [53].

Advantages and Potential for Learning

SL is free to use for participants (students and teachers), and it may not be complex to manage for young users such as medical students. The components of the SL experience can facilitate educational innovations through the following [25]:

- *Expanded or rich interactions*: referring to social interaction between individuals and communities as well as between humans and virtual objects
- *Visualization and contextualization*: the production and reproduction of content that is inaccessible, too expensive, imaginary, futuristic, or impossible to see for the human eye
- *Exposure to authentic content*: scientific, technical, and cultural
- *Identity play*: both individual and collective
- *Immersion in a 3D environment*: the sensation of presence through virtual embodiment in the form of an avatar and extensive modes of communication can impact the affective, empathic, and motivational aspects of the experience
- *Simulation*: reproduction of contexts that may be too expensive to reproduce in real life, with the advantage that some physical limitations can be overcome
- *Community presence*: promoting a sense of belonging and purpose that unites groups, subcultures, and geography
- *Content production*: opportunities for the creation and ownership of the learning environment and the objects within it

Barriers and Limitations

Although there is abundant evidence that educational activities can be carried out in scenarios specifically designed in SL, it must be recognized that there are certain barriers that may limit its use. One problem with using SL for education may lie in the technical capabilities of some users' computers and internet connections, because the SL viewer has significant requirements to run adequately [54]. There is a necessary learning curve for users; few people can simply "jump" into SL for the first time effortlessly [17]. This may cause rejection in some participants. In contrast, SL is considered old technology by some young users; for example, it still does not have an official native mobile app. However, there is an imminent release of an iOS or Android app for SL running on Unity, bringing the venerable virtual world "in top form" into the modern era [2].

XSL•FO
RenderX

There are other cultural, social, time investment, or economic barriers that must be considered. For example, as a result of the global reach of the virtual world, cultural and linguistic differences must be considered [17]. SL can be an isolating experience for the user as it is not always easy to participate and integrate into new communities in a place outside the user's "safety zone." Nonverbal cues with avatars are almost impossible to obtain (aside from user-generated emoticons). Time can be a considerable barrier as creating content and designing, validating, and executing educational activities requires time to address issues such as intellectual property rights, object permissions, and accessibility [25]. Although SL is free for learners, with basic accounts, the overheads of designing, implementing, practicing, and maintaining virtual sites in SL often require educators to develop multiple skills.

Universities in SL

The pedagogical possibilities of virtual worlds have motivated educators so that colleges and universities worldwide have taken the initiative to develop virtual campuses in SL. The early years of SL were very active in terms of education-related activities, with enthusiastic development. As of early 2007, a total of 98 colleges and universities worldwide had a presence in SL, a number that increased to 250 in 2009 [17]. In 2010, SL listed 78 universities as members [32], and >150 academic organizations were included in the SL Education Directory by 2013 [55]. At the time of writing this manuscript, a search inside SL using the term "university" returned 102 entries for venues or destination guides, giving a clear idea about the increase in the use of this environment for teaching and learning projects. Among them, it is worth highlighting the Rockcliffe University Consortium (Figure 2), which was started in 2006 by Phelan Corrimal with the goal of providing experiential educational opportunities in SL. In 2009, Rockcliffe had a virtual campus, widespread interaction with the nonprofit SL community, participation from 40 to 50 real-world universities, and a peer-reviewed journal (*Journal of Virtual Studies*). Rockcliffe is the organizer of the Virtual World Best Practices in Education Conference in SL (Figure 2), with 17 annual editions uninterruptedly [56], having reached 4000 attendees. Rockcliffe continues to actively pursue its mission of providing an environment for educational innovation, including affordability for students and teachers [17].

Figure 2. Screenshots of university organizations in Second Life. (A) A lecture at the 2020 Virtual World Best Practices in Education meeting. (B) The Rockcliffe University Consortium library building. (C) A view of the island of the University of Aveiro. (D) The meeting point of the Eduverse campus of the Virtual Worlds Education Consortium.

The Virtual Worlds Education Consortium launched a 7-region virtual educational campus called Eduverse in 2022 (Figure 2) comprising a group of virtual places focused on education in SL [57]. The other regions that are part of this project include The Science Circle, Virtech, the Community Virtual Library's Cookie region, the Mayo Clinic, Whole Brain Health's Inspiration Island region, and the University of New Mexico at Rockcliffe's Rockcliffe Village region. Since then, the consortium has held several weekly meetings nonstop on the shared campus with the goal of helping groups already in SL increase collaboration and resource sharing. The Virtual Worlds Education Consortium's purpose is to bring educators in virtual worlds together both to share what has been successful and tackle obstacles across different platforms and communities [58].

Educational Possibilities

The SL platform allows one to reproduce typical classroom learning activities such as courses, seminars, and workshops [33]. In the authors' opinion, SL offers certain advantages over other 2D web-based learning tools, as synchronous learning experiences can be organized with very diverse formats in varied

XSL•FO
RenderX

and imaginative environments. Furthermore, it represents an alternative to these 2D teaching resources, which, when used very frequently during the COVID-19 pandemic, produced a certain saturation among users [59].

SL allows for interesting simulation and gamification experiences that are impossible (or very expensive) to do in real life [25,33], with the added value of generating more dynamic contact, which is even more fun for Generation Z. There are various educational formats that can be carried out in the SL environment, adjusting to various learning areas in science, technology, and the humanities. These could be summarized as follows [25]: (1) self-paced tutorials, (2) displays and exhibits, (3) immersive exhibits, (4) role-plays and simulations, (5) data visualizations and simulations, (6) historical recreations and reenactments, (7) living and immersive archaeology, (8) machinima construction, (9) treasure hunts and quests, (9) language and cultural immersion, and (10) creative writing.

Medical Education Within SL

Learning Activities in Medical Education

Overview

A narrative review of the existing literature was carried out, searching for information in the PubMed and Google Scholar databases using the terms “Medical Education” AND “Second Life.” In addition, the point of view of our teaching team was considered after uninterrupted practical experience in undergraduate and postgraduate medical education in SL during the last 13 years with >4000 learners and >10 publications on the subject.

There are a wide range of health-related activities on SL, most of them dedicated to patient education, raising awareness of health issues, patients with specific diseases, marketing and promotion of health services, or health research. There are also academic sites offering training by means of classrooms, discussions, lectures, simulation, and patient interaction [60,61]. This review focused especially on medical education experiences during undergraduate, residency, and continuing medical education (CME). Multimedia Appendix 1 [33,45,46,62-88] shows a descriptive summary of the articles found. Studies related to the training of other professions such as nursing, psychology, and veterinary medicine were outside the focus of this review.

Different learning methodologies have been adopted tailored to varied groups of health learners. Some of them focus on developing immersive and realistic virtual patients or simulation scenarios within SL, training paramedic students in emergencies [89], the management of specific clinical situations by nurses [90], or the practice of communication and assessment skills by mental health nursing students [91]. Others focus on holding training workshops and clinical sessions [46] or interactive seminars in SL for postgraduate primary care physicians [45] or using the SL virtual environment for mock oral examinations for emergency medicine residents [62]. Academic experiences tailored to medical students performed in SL include creating a virtual laboratory for web-based anatomical education [63] or evaluating the use of team-based learning (TBL) of anatomy [64]. SL is also useful for conducting web-based radiology training activities with remote access in an attractive setting, especially for current generations of medical students and residents [33]. Typical classroom activities such as courses, seminars, and lectures can be transferred to SL, taking advantage of the feeling of “being there,” but active learning methods, which are more attractive to the student, can also be included, such as gamification, problem-based learning (PBL), or medical simulation. The development and characteristics of some of these activities in SL are explained in the following sections.

Courses, Seminars, and Conferences

SL has remarkable potential for effective teaching and learning as its structure allows for synchronous web-based meetings between teachers and students in the form of lectures, courses, or other typical classroom work formats [33]. These activities can be carried out in very varied scenarios, recreating a classroom, an auditorium, an open stage next to the sea, a floating platform in the sky, and infinite possibilities with the only limit of imagination and time to create these scenarios (Figure 3). Lorenzo-Álvarez et al [65] conducted a randomized study with medical students comparing an x-ray interpretation workshop conducted in SL and in real life. The results of the pre- and postexposure knowledge tests demonstrated that learning from a synchronous educational event in SL did not show differences from learning in the real world if the same content and script are maintained.

XSL•FO
RenderX

Figure 3. Screenshots of various scenes during medical education activities in Second Life. (A and B) Lectures with groups of approximately 30 avatars in a virtual classroom and an outdoor setting. (C and D) Scenes of student group meetings to discuss and prepare radiological cases.

In addition, the persistence of the system (open 24/7) allows for the inclusion of asynchronous activities to be solved during a course at the students' own pace [33]. The combination of synchronous sessions with asynchronous tasks is very interesting and useful for organizing web-based courses in SL. This type of organization of courses in SL has been highly valued by medical students with average scores of 8.9 to 9.3 out of 10 points [66,67] and by postgraduate physicians with average scores of 8.7 to 9.1 out of 10 points, who also perceive the usefulness of this platform in the training of residents and CME as it adapts to their work schedule and minimizes travel costs [67,68].

The Teacher's Perspective

Remote access allows for inviting speakers or professors from different institutions, opening up new interesting and enriching avenues for educational contact with students and between professors from different universities, with the advantages of reducing costs and travel times. From the perspective of 9 medical teachers who were invited to give a seminar in SL for the first time, it represents a training experience for trainers, which allowed them to learn about a new web-based infrastructure with interesting synchronous teaching capabilities (time, accessibility, ubiquity, synchrony, facilities, sound, and 3-dimensionality) [69]. The perceptions of 14 medical teachers after a 3-hour workshop given in SL explaining to them the platform and its educational possibilities "from within" were very positive [70]. They all considered it interesting and useful for teaching, and 71% agreed to carry out a teaching activity in SL. It is important to develop future actions in SL aimed at the training of trainers to [70] (1) promote collaborative work between teachers, (2) familiarize teachers with the educational possibilities of virtual worlds and the basic applicable technology, (3) share teaching experiences carried out in SL to achieve new interdisciplinary collaborative projects, (4) use SL as a means of virtual meeting and discussion, and (5) promote the development of multidisciplinary and interuniversity collaborative educational projects.

PBL Method

PBL, developed by Barrows [92] at McMaster University in the 1960s, is a student-centered learning methodology in which students are faced with a problem situation that they must solve themselves by applying previous knowledge and incorporating new identified knowledge [93]. The teacher stops being the center of the educational experience and becomes a guide who facilitates and directs learning. In medical education, this methodology contributes to acquiring skills such as real-world clinical reasoning, problem-solving, and critical decision-making [94]. In medicine, these "problems" are usually generated in the form of clinical cases, creating a variant of PBL known as case-based learning [95].

SL offers good support for posing open-ended problems to students in a PBL context [20]. PBL experiences have been carried out in SL with paramedic students [96] and medical students [71,72]. Beaumont et al [96] developed PBL scenarios about medical emergencies in SL. Results with focus groups of 10 students showed that SL can provide a rich and engaging environment based on the authenticity of the scenarios. As disadvantages, they pointed out a certain lack of detailed realism, the lack of face-to-face and hands-on contact in the experience, and some access problems. Rampling et al [71] designed a virtual patient with psychosis in SL for PBL in psychiatry. Participation was low (only 24 students out of 150 were invited), and feedback was predominantly negative, with 53 critical responses against 32 positive ones. Students expressed that the scenario was cumbersome, did not imitate real life, and had little educational value. The authors understood that the problem

XSL•FO
RenderX

lay in the use of scenarios within SL based predominantly on verbal communication. Jivram et al [72] studied the use of SL to enhance PBL in medical education comparing it to interactive web-based methods. Although a minority of the 244 participating students stated that the SL experience seemed more realistic, most preferred web-based methods due to their simplicity and effectiveness. The authors indicated that this was partly due to the temporal proximity of the exams and the additional effort required to learn the SL interface. These experiences show the way toward the development of PBL sessions in 3D virtual worlds in medicine. The technological evolution of SL in recent years can improve the authenticity of the scenarios [2], although there are access and usability issues that could be minimized by scheduling the experiences appropriately in the academic calendar and by training in the use of the platform and verbal communication.

Communication Skills

Communication features in SL simulate real-world communication. For example, voice volume is louder or weaker as the avatar moves closer or further away from the speaker and is perceived from the right or left depending on where the speaker is located [59]. This brings great realism to the immersive experience, and therefore, SL is an ideal tool for developing oral communication skills. Mitchell et al [73] conducted a pilot study in which a motivational interviewing program in SL was designed and tested to counsel patients about colorectal cancer screening, the results of which suggest that virtual worlds offer potential to enhance patient-centered communication skill training. Other experiences have shown that SL is an excellent tool for training in transversal communication skills such as public oral presentation of a topic, interaction with other users in real time, and the establishment of links and social relationships between attendees both in undergraduate [97] and postgraduate [74] students.

Meetings in SL allow for feedback from the audience, which, along with expert observation, are elements that help improve public presentation and communication skills [98]. van Ginkel et al [99] conducted a study with 36 first-year college students who completed a required oral presentation course in a virtual environment created using Unity. We agree with these authors that this type of virtual experience imitating real life helps further develop oral presentation skills, but it is essential that attendees learn to correctly use the basic audio communication controls (microphone and headphones) and avatar vision to properly follow the development of a session [74,97].

The presenter's voice is literally the instrument of connection with the attendees. When they give a presentation, how the audience feels about their voice is integral to how they perceive the presentation [98]. The quality of audio in SL allows for interaction, evaluation, and improvement of oral expression. Although, on the one hand, there is the disadvantage of not perceiving gestural communication, on the other hand, the filter provided by the interface and the avatar in SL makes beginners feel less embarrassed when speaking in public [97] and express themselves more openly and honestly [17].

Assessment

Assessment is an essential element in the teaching-learning process. SL allows teachers to carry out various exam formats in-world efficiently. Note cards are a document that records the user who created them and what time they were delivered to the teacher's avatar. Using note cards, deferred evaluations can be carried out so that the student can solve the assigned questions in writing in an allotted time [66-68] or multiple-choice tests in real time in the virtual classroom, with a short time assigned to each question [65,68].

The oral examination is a traditional method for assessing medical knowledge, clinical reasoning, and interpersonal skills. Schwaab et al [62] conducted an interesting study in which 27 emergency medicine residents participated in a virtual oral exam in SL. All examinees felt comfortable communicating with the examiner; most thought that the SL encounter was realistic (92.6%); and all felt that the virtual exam was fair, objective, and conducted efficiently. Most (66.6%) preferred to take oral exams using SL rather than the traditional format and expressed interest in using SL for other educational experiences (92.6%). McGrath et al [75] carried out a similar experience with 35 residents also from emergency medicine randomly assigned to SL (n=18) or the real world (n=17). Examinees in both groups scored without significant differences and thought their evaluation was realistic, fair, objective, and efficient. However, examinees in the SL group reported a preference for the virtual format and felt that it was less intimidating.

Simulation

Simulation is increasingly used in health care areas, especially in undergraduate education as a starting point for training medical professionals. Simulation-based learning experiences are useful for integrating theoretical knowledge with practice and the necessary practical skills [100-102]. Virtual simulation experiences have been developed in SL and other virtual worlds to train various skills, such as taking anamnesis with virtual patients [76], solving clinical situations in a pulmonology ward [77], conducting cardiopulmonary resuscitation [103], or developing communication skills with patients [104]. Danforth et al [76] used SL as an immersive environment to develop a standardized virtual patient system and an artificial intelligence (AI) markup language as a dialogue engine. This first system was useful to demonstrate the viability of the approach, but the virtual environment and dialogue management at the time (2009) were not ideal. To manage conversations, they used >200,000 rules due to inefficiencies in markup language and AI syntax. For this reason, they redesigned the application using Unity 3D to create the virtual environments and implemented ChatScript to manage the dialogue [105]. Students were able to accurately develop an appropriate differential diagnosis based on information provided by the simulated patient during the encounter, with the accuracy of simulated patient responses ranging from 79% to 86% depending on case complexity, type of history obtained, and student skill.

Toro-Troconis [77] at Imperial College London created a game-based SL simulation for medical students in which they could interact with virtual pulmonology patients to develop their skills and confidence. One group (n=23) had access to the

XSL•FO
RenderX

virtual patients in SL, and another (n=19) used an interactive electronic module. Similar attitudes were observed in both groups, with women showing a more positive attitude toward the perceived usefulness component of virtual patients in SL. The repetitive linear presentation of cases was not sufficiently motivating, so the use of more challenging branching learning experiences for the presentation of virtual patients is recommended [77].

Virtual simulated patients have the potential to reduce costs, faculty time, and resources required to help students develop their communication skills in safe, nonthreatening environments before contact with real patients. They also allow for standardization of interactions among students. Current efforts should focus on adding fluency to communication with virtual patients and feedback and assessment capabilities so that students can receive immediate feedback on the quality of their encounters with virtual patients [105].

Objective Structured Clinical Examination

The objective structured clinical examination (OSCE) is an examination format developed in different physical spaces (called clinical stations) in which clinical scenarios and situations are reproduced to evaluate the student's clinical skills in a standardized, reliable, and objective manner [106,107]. The restrictions on contact and physical presence during the COVID-19 pandemic favored 2D virtual OSCEs, with web-based access as a solution to these restrictions [108-110] proving to be enjoyable, interactive, and easy for students, as well as cost-effective [111].

SL is a useful digital platform for carrying out avatar-mediated OSCEs, both summative and formative, as it makes it possible to design and recreate various scenarios to train decision-making in clinical cases [72]. This has been verified in domestic accident scenarios in geriatric medicine, an experience in which 6 of the 8 geriatric medicine fellows rated the simulation as "excellent" [78], or in office or hospital settings, where 12 urology residents found the method feasible, acceptable, and applicable to assess communication skills [79].

Recently, a teaching experience in SL on emergency radiology was carried out with 180 sixth-year medical students organized into groups of 3 to 4 who had to solve emergency radiology shown in virtual OSCE stations. The students assessed the OSCE environment, the cases, the organization, and the training usefulness with average scores of 8.5 to 8.9 points out of 10 [80]. Currently, the authors are working on the development of a virtual radiology OSCE in which students must complete 6 radiology stations (Figure 4) answering questions about radiological interpretation, clinical judgment, and patient management having 9 minutes per station, replicating the organization of the face-to-face OSCE at the end of the degree. It would be interesting to develop and implement radiological OSCEs in virtual environments such as SL at the undergraduate, residency, and CME levels as it would allow for the optimization of resources and the performance of multicenter studies [112].

Figure 4. Screenshots of various scenes during a virtual objective structured clinical examination (OSCE) in Second Life. (A) Distribution of an OSCE room with students in front of the door of each station before starting. (B) Scene of a student at a radiology station. (C) Scene of a student reading the clinical scenario presented on a poster on the wall. (D) Scene in which the student interacts with a simulated patient represented by another avatar.

Gamification

Gamification applies elements of playful design to nonplayful contexts such as learning [113,114]. Gamification can be carried out in the classroom but also using web-based resources that facilitate remote access and user management, including the metaverse [115]. The SL platform offers experiences that are a game in themselves. Users enter an imaginary world in which

XSL•FO
RenderX

they move and interact represented by avatars whose appearance they can modify at will. Gamification in SL has interesting applications in medical education. Toro-Troconis [77] developed an interesting study on gamification with virtual patients in SL, among whose conclusions stand out the following: the importance of following a pedagogical framework in the design of game-based learning activities, the effectiveness of giving feedback to participants about rewards and their progress in the game, and the recommendation of previous exposure of students to the SL region where the activity will take place so that they get used to the environment and can concentrate on learning.

In 2015, a competitive web-based game in SL called League of Rays (LOR) [81] was designed to provide medical students with a gamified complement to the formal teaching of radiological anatomy and radiological semiology. The game, which lasts for 6 weeks, takes place on an island where, every week, 5-m panels with educational content appear, which are replaced by multiple-choice tests and other tasks to score points. In the first edition, voluntary participants scored better than nonparticipants in a postexposure test performed 1 month after the game (mean scores 59.0, SD 13.5 vs 45.3, SD 11.5; $P<.001$), indicating effective learning with the game [81]. In addition, both participants and nonparticipants provided valuable positive perceptions of the game and its benefit to their education. The following 2 years, participation was mandatory [82], demonstrating slightly less positive perception of and lower adherence to the game than when participation was voluntary. Subsequently, the rules were modified to voluntarily participate in teams of 4 students, encouraging teamwork and collaboration [83], and finally, 2 editions of interuniversity competition were held [84] in which the reproducibility of the experience was demonstrated, ruling out proximity biases and promoting collaborative action between teachers. Participants found team gaming useful for maintaining virtual relationships and identified this activity as a playful learning and social interaction experience during the COVID-19 pandemic [84]. This experience, maintained for 8 consecutive years and completed by 1032 students, demonstrates the usability of SL for learning games and the possibilities for collaborative work between students during gamification. In May 2024, a game similar to LOR designed for radiology residents called “Resident Debil” was developed in SL, the results of which are pending evaluation.

Health-Related Locations in SL

Relevant institutions have explored how to promote initiatives to improve health care by creating their own locations in SL. Some of these locations are not currently active. For example, the “Isla de la Salud” hosted interesting educational experiences with family physicians [46]; it even hosted, in parallel to real life, national congresses of the Spanish Society of Family and Community Medicine between 2008 and 2010, but unfortunately, it no longer exists today. At Ohio University’s SL location, a nutrition game was developed in which people could learn about the impact that fast food has on health [43]. The University of London created a virtual hospital in SL that included operating rooms where, for example, some posters showed the students the steps to prepare for surgery (such as dressing, wearing a mask, and disinfecting after the virtual procedure). Other parts of the hospital, such as a polyclinic or virtual pharmacy, were built in the same way to provide information to the students in an immersive environment [116].

There are other health-related virtual places that are currently active (Figure 5). The Mayo Clinic island in SL [117] includes conference facilities and a bookstore. The nonprofit medical practice regularly hosts talks and events on illnesses and diseases for interested users. The University of South Florida [118] College of Medicine headquarters includes information about the various schools and departments, including nursing, public health, and continuing education. Brodmann’s Brain at Inspiration Bay [119], developed by a team of virtual creators and neuroscientists, allows the visitor to walk through a gigantic interactive construction that shows functional areas of the human brain, which are activated (lighted up) based on various functions such as dancing or loving. The Málaga Medical School island [120] was created in 2021 to host permanent medical education content, synchronous educational events, and multidisciplinary OSCEs for undergraduate and graduate students.

XSL•FO
RenderX

Figure 5. Screenshots of several health-related virtual locations that are currently active in Second Life. (A) Mayo Clinic hospital. (B) Brodmann's Brain at Inspiration Bay, representing functional areas of the human brain. (C) Malaga Medical School island. (D) The University of South Florida College of Medicine.

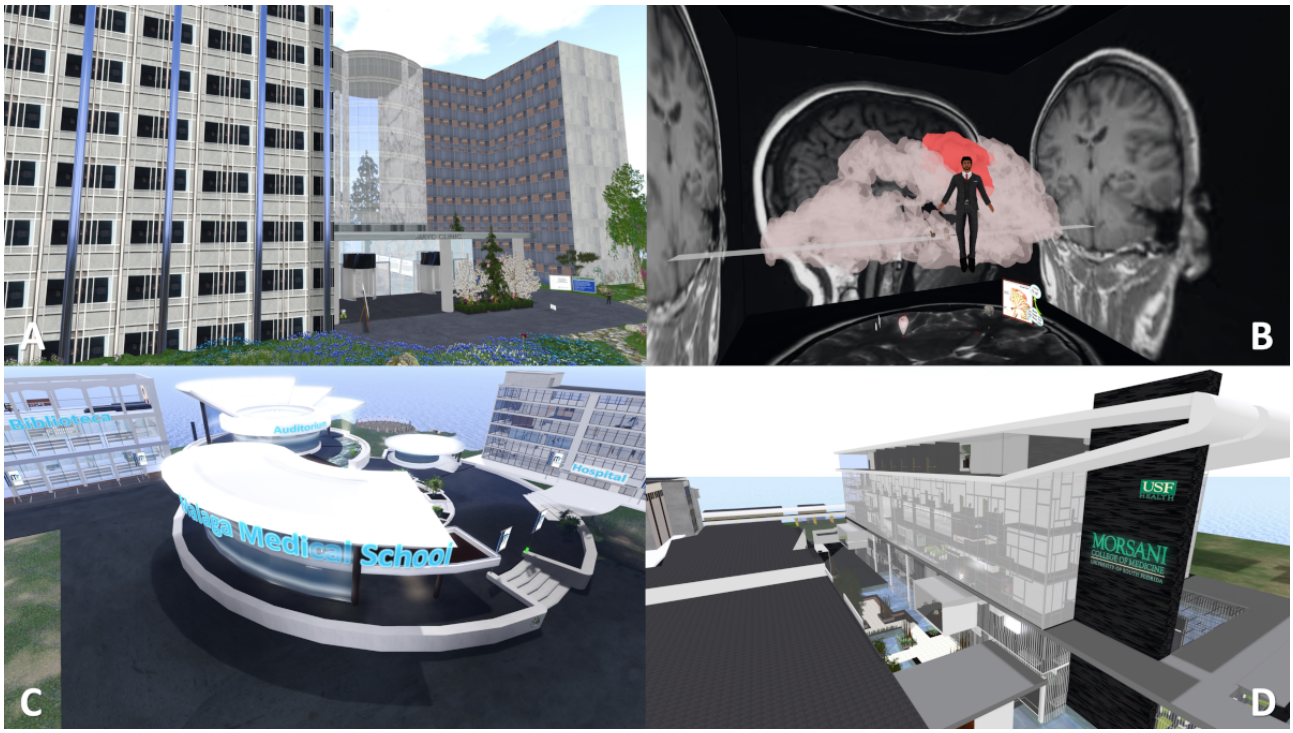

Educational Experiences in Different Medical Disciplines

Anatomy

Interesting experiences have been carried out for teaching anatomy in SL. Morales-Vadillo et al [121] used 3D scale models of oral anatomy created in SL that dental students were able to use without limitation for 4 weeks. The experimental group (n=23) scored significantly higher on questions about conceptual understanding and spatial interpretation than the control group (n=37) that used traditional teaching methods. Other authors have also created anatomical models so that students could evaluate them from the view of their avatar. Danforth [85] made a virtual testis model not only to explore the anatomy but also to examine the relationships between anatomy and physiology to describe the sperm production function. Richardson-Hatcher et al [86] built a 3D replica of the pterygopalatine fossa in SL where the students could virtually experience the organization of this anatomical area, maneuvering in and out of its boundaries and following the course of the nerves through the fossa. The study data suggest that students view this new technology as a valid method for studying anatomical arrangements by improving their understanding.

Gazave and Hatcher [64] implemented virtual TBL in a web-based anatomy course held in SL with 39 medical students. They demonstrated that virtual TBL was engaging for most students (average engagement score of 7.8 out of 10, with 89.2% of students reporting a score of ≥6). The authors concluded that virtual TBL sessions are worth implementing in other web-based courses. Richardson et al [63] built a multipurpose virtual anatomy laboratory in SL with a series of stations with (1) atlas or cadaver images, (2) a tour of cross-sectional anatomy, (3) cadaveric anatomy video tutorials, and (4) group quizzes. Students could obtain note cards with information and questions for each station by clicking on the station header. The authors stated that, while the virtual laboratory is not a substitute for cadaver-based anatomy experiences, SL allows students to participate to a greater extent than typical lecture transmissions via web-based learning.

Primary Care Medicine

Wiecha et al [45] designed a pilot postgraduate medical education program in the virtual world in which 14 primary care physicians participated. It was a 1-hour interactive session in SL on the topic of type 2 diabetes. Lorenzo-Álvarez et al [67] conducted a 3-week course for 14 family physicians on plain radiograph interpretation. Both studies reached similar results, with all participants agreeing that the SL experience was an effective CME method and that they were willing to participate in further educational events in SL. Melús-Palazón et al [46] carried out a series of accredited virtual clinical sessions in SL in which 76 primary care professionals from 9 health centers participated. They concluded that SL is a tool that allows for the design of educational activities that involve several health centers in different geographical locations, thus eliminating the need for travel and making more effective use of educational resources.

Psychiatry

There have been some experiences of medical education in psychiatry conducted in SL. In 2006, Yellowlees and Cook [122] built a visual and auditory hallucination recreation experience in SL in which 76% of 579 SL users reported a better understanding of visual and auditory hallucinations. The authors concluded that simulations of the perceptual phenomena of

XSL•FO
RenderX

psychiatric illnesses were feasible with the computer technology existing at that time. Rampling et al [71] used SL to create a simulated patient with psychosis for clinical training of medical students. A total of 24 participating students gave predominantly negative evaluations of the experience, and the authors did not consider it advisable to use scenarios predominantly based on verbal communication with the patient within SL, proposing to explore other interactive teaching methods in psychiatry, such as PBL.

Radiology

The use of e-learning and web-based resources to deliver radiology teaching to medical students represents an exciting alternative and an effective method to improve radiological knowledge and skills [123,124]. Medical images can be displayed in SL with sufficient quality to interpret the findings they present, providing adequate support for carrying out educational activities in radiology [33]. In 2011, a virtual space called "Medical Master's Island" was created [125], and various radiology teaching activities have been carried out uninterruptedly since then. More than 4000 learners, undergraduate and graduate, participated in these activities until May 2024. The most recent experiences are still pending publication.

In 2011, a pilot study was conducted with 46 medical students that demonstrated the feasibility of teaching undergraduate radiology in SL and the great acceptance by students [66]. In 2012 and 2013, several basic radiology courses were held in which 48 third-year students participated [67]. Both experiences demonstrated that students highly valued SL as an educational platform, recognizing the effort made by teachers to provide them with this opportunity, with average scores of 9.6 to 9.8 out of 10 points. In 2014, a randomized study was conducted with 156 third-year students to answer the following question: is learning in SL the same as learning in the real classroom [65]? An abdominal radiology seminar was given, and no significant differences were found between both groups in the pre- and postexposure tests, concluding that radiology education in SL encourages participatory learning and results in an acquisition of interpretive skills similar to that of a traditional face-to-face class.

As explained in the *Gamification* section, 8 consecutive annual editions of the multiuser game LOR, designed for learning radiological anatomy and semiology, were carried out between 2015 and 2022. It started as an experience in a single university [81-83], continuing as an interuniversity competition [84] in which, for 3 years, 129 teams of 4 students from 20 different universities participated. These experiences show the applicability of remote access to asynchronous gamification thanks to the persistence of the platform. Recently, in 2021 and 2023, "classroom-typical" synchronous gamification experiences have been carried out in SL, playing games such as Trivial Pursuit and The Alphabet to learn radiology [87].

Since 2011, other radiology training activities have been carried out, such as x-ray interpretation courses for 96 family medicine residents and 67 radiology residents from all over Spain [68], PBL experiences in radiology with medical students, and other activities based on training communication and public speaking skills on radiology topics with undergraduate and graduate students. Currently, the activity of this group from the University of Málaga is focused on medical simulation and carrying out virtual OSCEs in SL. In addition, the aim is to establish institutional alliances to promote the culture of virtual worlds between teachers and students.

Impact of the COVID-19 Pandemic on Medical Education in SL

The COVID-19 pandemic accelerated the application of new technologies in learning [126], caused stress in medical students [127], and reduced physician-student contact [128]. Various 2D platforms, which enable synchronous educational experiences between teachers and students, have experienced great technological development since the COVID-19 pandemic [129-133]. One might ask, if these 2D remote connection methods already exist, why are 3D environments needed? On the one hand, 3D environments such as SL have some advantages over synchronous 2D communication platforms [25,134] as they (1) are a 24/7 persistent environment; (2) allow for immediate, real-time interactions between users and objects in the 3D environment; (3) induce a strong feeling of presence (sense of "being there") in users; (4) promote social awareness, or the ability to sense the presence and location of participants in a learning environment, reinforcing the perception of "who is there" and "what is happening"; and (5) generate a greater sense of belonging to a community. On the other hand, there is a phenomenon of fatigue from 2D videoconferencing triggered by the intensive use of videoconferencing since the COVID-19 pandemic, so 3D environments appear as an alternative for a change of scenery [59].

The need to continue the teaching-learning process during the pandemic and the limitations of 2D virtual modalities, in which it is difficult to carry out experiential learning, led to the development of learning scenarios through the metaverse and virtual worlds [135]. Thanks to the availability of precreated SL resources when the lockdown was declared, the planned radiology seminars for 157 medical students could be quickly transferred to SL [88]. More than 90% of them felt more engaged in teaching thanks to the SL seminars and stated that they found the virtual platform an attractive and fun environment and that, if they were to return to a lockdown situation, they would like to conduct similar learning activities [88]. Furthermore, participants in the 2020 interuniversity LOR competition found the team game useful for maintaining virtual relationships and identified this activity as a playful learning experience and social interaction during the COVID-19 pandemic [84].

Limitations and Drawbacks

Although, in general, both students and teachers involved have valued the use of SL for teaching medicine very positively, there are some limitations and drawbacks that should be pointed out. First, technical limitations due to the low capacity of the central processing unit, graphic card, or internet connection can prevent access to SL or cause defects in the representation of the virtual world [88]. These types of incidents have been described by 9% to 11% of participants [65,66,81] and are not always critical. Second, when the audio quality (voice input or

XSL•FO
RenderX

output) of a user is limited, it makes it difficult to correctly develop synchronous activities with voice support, something crucial when it happens to the teacher. Therefore, it is advisable to train and check the equipment to be used beforehand [88]. Third, it is sometimes difficult to view the same presentation, especially when there are many interactions on the display screen. This can be minimized with adequate previous training explaining the rules of behavior during the sessions [88]. Fourth, the limit of attending avatars on an island is 100, but to avoid access and interaction problems, it is recommended not to exceed 45 to 50 avatars synchronously in the same place [88]. This limit prevents synchronous activities with large groups in SL.

The Future

The metaverse is alive, and so is SL. Constant technical improvements are sought. Virtual world data are no longer housed in large server rooms but are hosted in the Amazon cloud. Linden Lab is a more decentralized company based on the remote work of its employees. The company is slowly rolling out its "Puppetry Project," a new feature that allows the user to capture their real movements and facial expressions via a webcam and other motion capture devices and then display them on their avatar in real time [2].

The contributions of the community of creators also offer technological changes for the immediate future. For example, SL viewers with more capabilities continue to be developed. The SL viewer and the third-party Firestorm viewer are usually updated to a new version several times a year. Crystal Frost, a new external SL viewer running on Unity designed by Berry Bunny, an SL community creator, is in its beta phase for Windows but will also be available on Linux and Mac and on mobile devices [136].

Other 3D collaborative tools are currently being developed, such as Frame VR [137], in which users share their virtual space to work together. The future integration of technologies such as VR and AR in platforms such as SL and Frame VR opens interesting lines of research in education. As VR technology allows users to fully immerse themselves in a computer-generated world and AR overlays digital information onto the real world, their combination in virtual worlds could lead to a new era of virtual presence in which learners could interact with the virtual environment and other avatars in more natural and intuitive ways [138]. As these technologies continue to evolve, they are likely to become more integrated into our daily lives toward a fusion model, blurring the lines between physical and digital realities [139]. However, the successful integration of these technologies requires addressing challenges related to accessibility, technical requirements, and the development of effective pedagogical strategies [138].

Generative AI can also bring exciting possibilities to medical education in SL, for example, AI-powered chatbots. This would allow for more immersive human-computer interface loops, from conversations with patients or virtual therapists to interactions of very diverse types. Since there is user input, AI could be trained to produce highly customizable responses or generate unique stories per use [2].

Conclusions

In the 25 years since its conception, SL has proved to be a virtual platform that connects with the concept of the metaverse, an open system with global access and interconnected where all humans can access to socialize or share products for free or using a virtual currency. SL continues to be successful as a virtual world model that connects with the concept of an equitable, diverse, and creative metaverse. Its creativity tools are open and powerful, and its user community is diverse and made up of members free to create endless values and possibilities.

These 25 years have made SL a platform where different generations coexist. SL remains active and technologically improved since its creation. Future improvements are also expected, such as the creation of viewers for iOS and Android that will allow SL to be used on mobile devices or the incorporation of generative AI, which will provide more enriching and nonrepetitive experiences.

From an educational point of view, SL has the advantages of being persistent 24/7 and creating in the student the important feeling of "being there" and of copresence. This, together with the reproduction of the 3D world, real-time interaction, and the quality of voice communication, makes immersive experiences unique, generating engagement and a fluid interrelation of students with each other and with their teachers. Various groups of researchers in medical education have developed experiences during these years, which have shown that courses, seminars, workshops and conferences, PBL experiences, evaluations, teamwork, gamification, medical simulation, and virtual OSCEs can be successfully carried out. Acceptance by students and teachers is generally positive, recognizing its usefulness for undergraduate medical education and CME. However, there are few studies evaluating the educational impact of SL activities. Therefore, it is necessary to continue carrying out educational experiences, outlining the organization, objectives, and content and measuring the real educational impact to make SL an educational tool for more universal use.

Acknowledgments

This work was partially supported by the educational innovation project of the University of Málaga (PIE22-045). Funding for the open access charge was obtained from the University of Málaga and Consorcio de Bibliotecas Universitarias de Andalucía (Consortium of University Libraries of Andalusia; CBUA).

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Articles related to medical education experiences in Second Life during undergraduate, residency, and continuing medical education.
[DOCX File , 26 KB - jmir_v26i1e59005_app1.docx]

References

1. Stephenson N. Snow Crash. New York, NY: Bantam Spectra; 1992.
2. Au JW. Why the Metaverse Matters: From Second Life to Met a & Beyond, the Promises & Perils of the Next Gene ration of the Internet: From Snow Crash & Second Life to A Virtual World Worth Fighting For. Hoboken, NJ: John Wiley & Sons; 2023.
3. Mystakidis S. Metaverse. Encyclopedia 2022 Feb 10;2(1):486-497. [doi: 10.3390/encyclopedia2010031]
4. Rospigliosi P. Adopting the metaverse for learning environments means more use of deep learning artificial intelligence: this presents challenges and problems. Interact Learn Environ 2022 Oct 16;30(9):1573-1576. [doi: 10.1080/10494820.2022.2132034]
5. Wider W, Jiang L, Lin J, Fauzi MA, Li J, Chan CK. Metaverse chronicles: a bibliometric analysis of its evolving landscape. Int J Hum Comput Interact 2023 Jun 27:1-14. [doi: 10.1080/10447318.2023.2227825]
6. Park SM, Kim YG. A metaverse: taxonomy, components, applications, and open challenges. IEEE Access 2022;10:4209-4251. [doi: 10.1109/access.2021.3140175]
7. Jovanović A, Milosavljević A. VoRtex metaverse platform for gamified collaborative learning. Electronics 2022 Jan 20;11(3):317. [doi: 10.3390/electronics11030317]
8. Chen Z. Exploring the application scenarios and issues facing Metaverse technology in education. Interact Learn Environ 2022 Oct 16:1-13. [doi: 10.1080/10494820.2022.2133148]
9. Collins C. Looking to the future: higher education in the metaverse. EDUCAUSE Rev 2008;43(5):51-63 [FREE Full text]
10. Hwang GJ, Chien SY. Definition, roles, and potential research issues of the metaverse in education: an artificial intelligence perspective. Comput Educ Artif Intell 2022;3:100082. [doi: 10.1016/j.caeai.2022.100082]
11. Schroeder R. Defining virtual worlds and virtual environments. J Virtual World Res 1970 Jan 01;1(1):1-3. [doi: 10.4101/jvwr.v1i1.294]
12. Czepielewski S, Christodoulopoulou C, Kleiner J, Mirinaviciute W, Valencia E. Virtual 3D tools in online language learning. In: Czepielewski S, editor. Virtual 3D Tools in Online Language Learning. Lausanne, Switzerland: Peter Lang; 2011:7-14.
13. Hassouneh D, Brengman M. A motivation-based typology of social virtual world users. Comput Human Behav 2014 Apr;33:330-338. [doi: 10.1016/j.chb.2013.08.012]
14. Bell MW. Toward a definition of “virtual worlds”. J Virtual World Res 1970 Jan 01;1(1):1-5. [doi: 10.4101/jvwr.v1i1.283]
15. Mahaley S, Teigland R. Advancing learning through virtual worlds. In: Murugesan S, editor. Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications. Hershey, PA: Information Science Reference; 2010:556-572.
16. Zhang P, Ma X, Pan Z, Li X, Xie K. Multi-agent cooperative reinforcement learningin 3D virtual world. In: Proceedings of the 1st International Conference on Advances in Swarm Intelligence. 2010 Presented at: ICSI '10; June 12-15, 2010; Beijing, China p. 731-739 URL: https://link.springer.com/chapter/10.1007/978-3-642-13495-1_90 [doi: 10.1007/978-3-642-13495-1_90]
17. Johnson ML. Social Virtual Worlds and Their Places. Singapore, Singapore: Palgrave Macmillan; 2022.
18. Messinger PR, Stroulia E, Lyons K. A typology of virtual worlds: historical overview and future directions. J Virtual World Res 1970 Jan 01;1(1):1-18. [doi: 10.4101/jvwr.v1i1.291]
19. Girvan C. What is a virtual world? definition and classification. Education Tech Research Dev 2018 Feb 8;66(5):1087-1100. [doi: 10.1007/S11423-018-9577-Y]
20. Good J, Howland K, Thackray L. Problem-based learning spanning real and virtual words: a case study in Second Life. ALT-J 2016 Dec 14;16(3):163-172. [doi: 10.1080/09687760802526681]
21. Jamaludin A, Chee YS, Ho CM. Fostering argumentative knowledge construction through enactive role play in Second Life. Comput Educ 2009 Sep;53(2):317-329. [doi: 10.1016/j.compedu.2009.02.009]
22. Girvan C, Savage T. Identifying an appropriate pedagogy for virtual worlds: a communal constructivism case study. Comput Educ 2010 Aug;55(1):342-349. [doi: 10.1016/j.compedu.2010.01.020]
23. Douglas R. 20 years on: Second Life, the fashion-forward metaverse that keeps on giving. Fashionunited. 2023 Jul 28. URL: https://fashionunited.uk/news/business/20-years-on-second-life-the-fashion-forward-metaverse-that-keeps-on-giving/2023072870819 [accessed 2024-03-24]
24. Galov N. Second life facts in 2023: what it means to live in a virtual world. Webtribunal. 2023 Mar 6. URL: https://webtribunal.net/blog/second-life-facts#gref [accessed 2024-03-24]

XSL•FO
RenderX

25. Warburton S. Second Life in higher education: assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. Brit J Educational Tech 2009 Apr 15;40(3):414-426. [doi: 10.1111/j.1467-8535.2009.00952.x]
26. Savin-Baden M. A Practical Guide to Using Second Life in Higher Education. Maidenhead, UK: McGraw-Hill Open University Press; 2010.
27. Ernstberger P. Linden Dollar and virtual monetary policy. SSRN Journal Preprint posted online on February 10, 2009 [FREE Full text] [doi: 10.2139/ssrn.1339895]
28. Rymaszewski M, Au WJ, Wallace M, Winters C, Ondrejka C, Batstone-Cunningham B. Second Life®: The Official Guide. Hoboken, NJ: Wiley-Blackwell; 2007.
29. Second Life. Linden Lab Inc. URL: http://www.secondlife.com [accessed 2024-03-24]
30. Second Life viewer download. Linden Lab Inc. URL: http://secondlife.com/support/downloads/ [accessed 2024-03-24]
31. Second Life third party viewer directory. Linden Research Inc. URL: http://wiki.secondlife.com/wiki/Third_Party_Viewer_Directory [accessed 2024-03-24]
32. Ahern N, Wink DM. Virtual learning environments: Second Life. Nurse Educ 2010;35(6):225-227. [doi: 10.1097/NNE.0b013e3181f7e943] [Medline: 20975400]
33. Lorenzo Álvarez R, Pavía Molina J, Sendra Portero F. Possibilities of the three-dimensional virtual environment tridimensional Second Life® for training in radiology. Radiologia (Engl Ed) 2018 Jul;60(4):273-279 [FREE Full text] [doi: 10.1016/j.rx.2018.02.006] [Medline: 29571525]
34. Matienzo R. Evolución de la teoría del aprendizaje significativo y su aplicación en la educación superior [Article in Spanish]. Dialektika J Philos Res Soc Theory 2020 Jan 7;2(3):17-26 [FREE Full text]
35. Williamson KB, Gunderman RB, Cohen MD, Frank MS. Learning theory in radiology education. Radiology 2004 Oct;233(1):15-18. [doi: 10.1148/radiol.2331040198] [Medline: 15454615]
36. Minocha S, Hardy C. Navigation and way finding in learning spaces in 3D virtual worlds. In: Gregory S, Lee M, Dalgarno B, Tynan B, editors. Learning in Virtual Worlds: Research and Applications. Montgomery, AL: AU Press; 2016:3-41.
37. Charles D, Charles T, McNeill M, Bustard D, Black M. Game‐based feedback for educational multi‐user virtual environments. Brit J Educational Tech 2011 Jun 08;42(4):638-654. [doi: 10.1111/j.1467-8535.2010.01068.x]
38. Paas F, Tuovinen JE, Tabbers H, Van Gerven PW. Cognitive load measurement as a means to advance cognitive load theory. Educ Psychol 2010 Jun 08;38(1):63-71. [doi: 10.1207/S15326985EP3801_8]
39. Hessler KL, Henderson AM. Interactive learning research: application of cognitive load theory to nursing education. Int J Nurs Educ Scholarsh 2013 Jun 25;10(1):133-141. [doi: 10.1515/ijnes-2012-0029] [Medline: 23813334]
40. Sweller J. Cognitive load theory, learning difficulty, and instructional design. Learn Instr 1994 Jan;4(4):295-312. [doi: 10.1016/0959-4752(94)90003-5]
41. Van Nuland SE, Rogers KA. E-learning, dual-task, and cognitive load: the anatomy of a failed experiment. Anat Sci Educ 2016 Oct 19;9(2):186-196. [doi: 10.1002/ase.1576] [Medline: 26480302]
42. Sweller J. Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ Psychol Rev 2010 Apr 23;22(2):123-138. [doi: 10.1007/S10648-010-9128-5]
43. Boulos MN, Hetherington L, Wheeler S. Second Life: an overview of the potential of 3-D virtual worlds in medical and health education. Health Info Libr J 2007 Dec;24(4):233-245. [doi: 10.1111/j.1471-1842.2007.00733.x] [Medline: 18005298]
44. Hansen MM, Murray PJ, Erdley WS. The potential of 3-D virtual worlds in professional nursing education. Stud Health Technol Inform 2009;146:582-586. [Medline: 19592909]
45. Wiecha J, Heyden R, Sternthal E, Merialdi M. Learning in a virtual world: experience with using second life for medical education. J Med Internet Res 2010 Jan 23;12(1):e1 [FREE Full text] [doi: 10.2196/jmir.1337] [Medline: 20097652]
46. Melús-Palazón E, Bartolomé-Moreno C, Palacín-Arbués JC, Lafuente-Lafuente A, García IG, Guillen S, et al. Experience with using Second Life for medical education in a family and community medicine education unit. BMC Med Educ 2012 May 15;12:30 [FREE Full text] [doi: 10.1186/1472-6920-12-30] [Medline: 22587562]
47. Miller M, Jensen R. Avatars in nursing: an integrative review. Nurse Educ 2014;39(1):38-41. [doi: 10.1097/01.NNE.0000437367.03842.63] [Medline: 24300258]
48. Baker SC, Wentz RK, Woods MM. Using virtual worlds in education: Second Life® as an educational tool. Teaching of Psychology 2009 Jan 02;36(1):59-64. [doi: 10.1080/00986280802529079]
49. Inman C, Wright VH, Hartman JR. Use of Second Life in K-12 and higher education: a review of research. J Interact Online Learn 2010;9(1):44-63 [FREE Full text]
50. Potkonjak V, Gardner M, Callaghan V, Mattila P, Guetl C, Petrović VM, et al. Virtual laboratories for education in science, technology, and engineering: a review. Comput Educ 2016 Apr;95:309-327. [doi: 10.1016/j.compedu.2016.02.002]
51. Gong W. Education and three-dimensional virtual worlds: a critical review and analysis of applying second life in higher education. Nanjing Normal University. 2018. URL: https://open.library.ubc.ca/media/download/pdf/42591/1.0371092/4 [accessed 2024-03-27]
52. Liaw SY, Carpio GA, Lau Y, Tan SC, Lim WS, Goh PS. Multiuser virtual worlds in healthcare education: a systematic review. Nurse Educ Today 2018 Jun;65:136-149. [doi: 10.1016/j.nedt.2018.01.006] [Medline: 29571002]

XSL•FO
RenderX

53. Zhang F, Luo H. Visual analysis of virtual world research in education from 2013 to 2022. In: Proceedings of the 2023 5th International Conference on Computer Science and Technologies in Education. 2023 Presented at: CSTE '23; April 21-23, 2023; Xi'an, China p. 202-207 URL: https://ieeexplore.ieee.org/document/10261601 [doi: 10.1109/cste59648.2023.00042]
54. Second Life system requirements. Linden Lab Inc. URL: https://secondlife.com/support/system-requirements [accessed 2024-03-24]
55. Second Life education directory. Linden Research Inc. URL: https://wiki.secondlife.com/wiki/Second_Life_Education_Directory [accessed 2024-03-24]
56. Virtual worlds best practices in education: mythic origins. Virtual Worlds Best Practices in Education. URL: https://www.vwbpe.org/ [accessed 2024-03-28]
57. Kariuki D. Eduverse to launch in Second Life this weekend. Hypergrid Business. 2022 May 20. URL: https://www.hypergridbusiness.com/2022/05/eduverse-to-launch-in-second-life-this-weekend/ [accessed 2024-03-28]
58. Home page. Virtual World Education Consortium. URL: https://www.vweconsortium.org/ [accessed 2024-03-28]
59. Speidel R, Felder E, Schneider A, Öchsner W. Virtual reality against Zoom fatigue? a field study on the teaching and learning experience in interactive video and VR conferencing. GMS J Med Educ 2023;40(2):Doc19 [FREE Full text] [doi: 10.3205/zma001601] [Medline: 37361246]
60. Beard L, Wilson K, Morra D, Keelan J. A survey of health-related activities on Second Life. J Med Internet Res 2009;11(2):e17 [FREE Full text] [doi: 10.2196/jmir.1192] [Medline: 19632971]
61. Ghanbarzadeh R, Ghapanchi AH, Blumenstein M, Talaei-Khoei A. A decade of research on the use of three-dimensional virtual worlds in health care: a systematic literature review. J Med Internet Res 2014;16(2):e47 [FREE Full text] [doi: 10.2196/jmir.3097] [Medline: 24550130]
62. Schwaab J, Kman N, Nagel R, Bahner D, Martin DR, Khandelwal S, et al. Using Second Life virtual simulation environment for mock oral emergency medicine examination. Acad Emerg Med 2011 May;18(5):559-562 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01064.x] [Medline: 21521404]
63. Richardson A, Hazzard M, Challman SD, Morgenstein AM, Brueckner JK. A "Second Life" for gross anatomy: applications for multiuser virtual environments in teaching the anatomical sciences. Anat Sci Educ 2011;4(1):39-43. [doi: 10.1002/ase.195] [Medline: 21265036]
64. Gazave CM, R. Hatcher A. Evaluating the use of Second Life™ for virtual team-based learning in an online undergraduate anatomy course. Med Sci Educ 2017 Jan 18;27(2):217-227. [doi: 10.1007/s40670-017-0374-8]
65. Lorenzo-Alvarez R, Rudolphi-Solero T, Ruiz-Gomez MJ, Sendra-Portero F. Medical student education for abdominal radiographs in a 3D virtual classroom versus traditional classroom: a randomized controlled trial. AJR Am J Roentgenol 2019 Sep;213(3):644-650. [doi: 10.2214/AJR.19.21131] [Medline: 31287725]
66. Lorenzo-Alvarez R, Pavia-Molina J, Sendra-Portero F. Exploring the potential of undergraduate radiology education in the virtual world second life with first-cycle and second-cycle medical students. Acad Radiol 2018 Aug;25(8):1087-1096. [doi: 10.1016/j.acra.2018.02.026] [Medline: 30782465]
67. Lorenzo-Alvarez R, Ruiz-Gomez MJ, Sendra-Portero F. Medical students' and family physicians' attitudes and perceptions toward radiology learning in the Second Life virtual world. AJR Am J Roentgenol 2019 Jun;212(6):1295-1302. [doi: 10.2214/AJR.18.20381] [Medline: 30860900]
68. Pino Postigo A, Domínguez Pinos D, Eduardo Ochando Pulido E, Navarro Sanchis E, Sendra Portero F. Un curso de radiografía simple para residentes de radiodiagnóstico en un entorno virtual inmersivo [Article in Spanish]. 36º Congreso Nacional de la Sociedad Española de Radiología Médica SERAM. URL: https://piper.espacio-seram.com/index.php/seram/article/view/8398/6864 [accessed 2024-03-29]
69. Alonso-Martinez JM, Ravaei S, Rudolphi-Solero T, Sendra-Portero F. Radiology seminars with guest professors in the virtual environment Second Life®: perception of learners and teachers. Proceedings 2020;54(1):16 [FREE Full text] [doi: 10.3390/proceedings2020054016]
70. Rudolphi-Solero TJ, Jimenez-Zayas A, Sendra-Portero F. Introduction to Second Life® as a virtual training environment: perception of university teachers. Proceedings 2020;54(1):18. [doi: 10.3390/proceedings2020054018]
71. Rampling J, O'Brien A, Hindhaugh K, Woodham L, Kavia S. Use of an online virtual environment in psychiatric problem-based learning. Psychiatrist 2018 Jan 02;36(10):391-396. [doi: 10.1192/pb.bp.111.037630]
72. Jivram T, Kavia S, Poulton E, Hernandez AS, Woodham LA, Poulton T. The development of a virtual world problem-based learning tutorial and comparison with interactive text-based tutorials. Front Digit Health 2021 Apr 20;3:611813 [FREE Full text] [doi: 10.3389/fdgth.2021.611813] [Medline: 34713092]
73. Mitchell S, Heyden R, Heyden N, Schroy P, Andrew S, Sadikova E, et al. A pilot study of motivational interviewing training in a virtual world. J Med Internet Res 2011 Sep 26;13(3):e77 [FREE Full text] [doi: 10.2196/jmir.1825] [Medline: 21946183]
74. Pino-Postigo A, Domínguez-Pinos D, Lorenzo-Alvarez R, Pavía-Molina J, Ruiz-Gómez MJ, Sendra-Portero F. Improving oral presentation skills for radiology residents through clinical session meetings in the virtual world second life. Int J Environ Res Public Health 2023 Mar 08;20(6):4738 [FREE Full text] [doi: 10.3390/ijerph20064738] [Medline: 36981654]
75. McGrath J, Kman N, Danforth D, Bahner DP, Khandelwal S, Martin DR, et al. Virtual alternative to the oral examination for emergency medicine residents. West J Emerg Med 2015 Mar 23;16(2):336-343 [FREE Full text] [doi: 10.5811/westjem.2015.1.24344] [Medline: 25834684]

XSL•FO
RenderX

76. Danforth DR, Procter M, Chen R, Johnson M, Heller R. Development of virtual patient simulations for medical education. JVWR 2009 Jul 19;2(2):1-11. [doi: 10.4101/jvwr.v2i2.707]
77. Toro-Troconis M. Game-based learning for virtual patients in Second Life®. Doctoral thesis dissertation. Luleå University of Technology. URL: https://www.diva-portal.org/smash/get/diva2:999020/FULLTEXT01.pdf [accessed 2024-03-28]
78. Andrade AD, Cifuentes P, Oliveira MC, Anam R, Roos BA, Ruiz JG. Avatar-mediated home safety assessments: piloting a virtual objective structured clinical examination station. J Grad Med Educ 2011 Dec;3(4):541-545 [FREE Full text] [doi: 10.4300/JGME-D-11-00236.1] [Medline: 23205205]
79. Kava BR, Andrade AD, Marcovich R, Idress T, Ruiz JG. Communication skills assessment using human avatars: piloting a virtual world objective structured clinical examination. Urol Pract 2017 Jan;4(1):76-84. [doi: 10.1016/j.urpr.2016.01.006] [Medline: 37592593]
80. Pérez Baena AV, Lorenzo Álvarez R, Sendra Portero F. ECOE-RX: Gamificación y aprendizaje basado en casos para aprender radiología clínica en pregrado [Article in Spanish]. 36º Congreso Nacional de la Sociedad Española de Radiología Médica SERAM. URL: https://piper.espacio-seram.com/index.php/seram/article/view/8394 [accessed 2024-03-28]
81. Lorenzo-Alvarez R, Rudolphi-Solero T, Ruiz-Gomez MJ, Sendra-Portero F. Game-based learning in virtual worlds: a multiuser online game for medical undergraduate radiology education within Second Life. Anat Sci Educ 2020 Sep 25;13(5):602-617. [doi: 10.1002/ase.1927] [Medline: 31665564]
82. Rudolphi-Solero T, Lorenzo-Alvarez R, Ruiz-Gomez MJ, Sendra-Portero F. Impact of compulsory participation of medical students in a multiuser online game to learn radiological anatomy and radiological signs within the virtual world Second Life. Anat Sci Educ 2022 Aug;15(5):863-876. [doi: 10.1002/ase.2134] [Medline: 34449983]
83. Rudolphi-Solero T, Jimenez-Zayas A, Lorenzo-Alvarez R, Domínguez-Pinos D, Ruiz-Gomez MJ, Sendra-Portero F. A team-based competition for undergraduate medical students to learn radiology within the virtual world Second Life. Insights Imaging 2021 Jun 29;12(1):89 [FREE Full text] [doi: 10.1186/s13244-021-01032-3] [Medline: 34185165]
84. Rudolphi-Solero T, Lorenzo-Álvarez R, Domínguez-Pinos D, Ruiz-Gómez MJ, Sendra-Portero F. An interuniversity competition for medical students to learn radiology in the Second Life metaverse. J Am Coll Radiol 2024 May;21(5):812-821 [FREE Full text] [doi: 10.1016/j.jacr.2023.09.012] [Medline: 37922968]
85. Danforth DR. Development of an interactive virtual 3-D model of the human testis using the Second Life platform. In: Thomas M, editor. Design, Implementation, and Evaluation of Virtual Learning Environments. Hershey, PA: Information Science Reference; 2018:153-167.
86. Richardson-Hatcher A, Hazzard M, Brueckner-Collins J. Using a three-dimensional virtual platform to teach the pterygopalatine fossa. Med Sci Educ 2014 Jan 13;23(3):308-312. [doi: 10.1007/BF03341639]
87. Alonso Martínez JM, Ravaei S, Lorenzo Álvarez R, Rudolphi Solero T, Sendra Portero F. Juegos sociales para aprender radiología en pregrado dentro de un entorno virtual 3D: trivial y pasapalabra [Article in Spanish]. 36º Congreso Nacional de la Sociedad Española de Radiología Médica SERAM. URL: https://piper.espacio-seram.com/index.php/seram/article/view/8395/6861 [accessed 2024-03-29]
88. Ravaei S, Alonso-Martínez JM, Jiménez-Zayas A, Sendra-Portero F. Reflections about learning radiology inside the multi-user immersive environment Second Life® during confinement by COVID-19. Proceedings 2020;54(1):9. [doi: 10.3390/proceedings2020054009]
89. Conradi E, Kavia S, Burden D, Rice A, Woodham L, Beaumont C, et al. Virtual patients in a virtual world: training paramedic students for practice. Med Teach 2009 Aug;31(8):713-720. [Medline: 19811207]
90. Patel V, Lee H, Taylor D, Aggarwal R, Kinross J, Darzi A. Virtual worlds are an innovative tool for medical device training in a simulated environment. Stud Health Technol Inform 2012;173:338-343. [Medline: 22357014]
91. Kidd LI, Knisley SJ, Morgan KI. Effectiveness of a Second Life(®) simulation as a teaching strategy for undergraduate mental health nursing students. J Psychosoc Nurs Ment Health Serv 2012 Jul;50(7):28-37. [doi: 10.3928/02793695-20120605-04] [Medline: 22694785]
92. Barrows H. Problem‐based learning in medicine and beyond: a brief overview. New Dir Teach Learn 2006 Aug 17;1996(68):3-12. [doi: 10.1002/tl.37219966804]
93. Bodagh N, Bloomfield J, Birch P, Ricketts W. Problem-based learning: a review. Br J Hosp Med (Lond) 2017 Nov 02;78(11):C167-C170. [doi: 10.12968/hmed.2017.78.11.C167] [Medline: 29111794]
94. Jones RW. Problem-based learning: description, advantages, disadvantages, scenarios and facilitation. Anaesth Intensive Care 2019 Jan 16;34(4):485-488. [doi: 10.1177/0310057x0603400417]
95. Carrasco GA, Behling KC, Lopez OJ. Evaluation of the role of incentive structure on student participation and performance in active learning strategies: a comparison of case-based and team-based learning. Med Teach 2017 Dec 05;40(4):379-386. [doi: 10.1080/0142159x.2017.1408899]
96. Beaumont C, Savin-Baden M, Conradi E, Poulton T. Evaluating a Second Life problem-based learning (PBL) demonstrator project: what can we learn? Interact Learn Environ 2012 Feb 10;22(1):125-141. [doi: 10.1080/10494820.2011.641681]
97. Lozano-Durán A, Rudolphi-Solero T, Nava-Baro E, Ruiz-Gómez MJ, Sendra-Portero F. Training scientific communication skills on medical imaging within the virtual world second life: perception of biomedical engineering students. Int J Environ Res Public Health 2023 Jan 17;20(3):1697 [FREE Full text] [doi: 10.3390/ijerph20031697] [Medline: 36767063]

XSL•FO
RenderX

98. Schraeder TL. Public speaking and presentation skills. In: Schraeder TL, editor. Physician Communication: Connecting with Patients, Peers, and the Public. Oxford, UK: Oxford University Press; 2019:113-166.
99. van Ginkel S, Gulikers J, Biemans H, Noroozi O, Roozen M, Bos T, et al. Fostering oral presentation competence through a virtual reality-based task for delivering feedback. Comput Educ 2019 Jun;134:78-97. [doi: 10.1016/j.compedu.2019.02.006]
100. Nara N, Beppu M, Tohda S, Suzuki T. The introduction and effectiveness of simulation-based learning in medical education. Intern Med 2009;48(17):1515-1519 [FREE Full text] [doi: 10.2169/internalmedicine.48.2373] [Medline: 19721295]
101. Koh J, Dubrowski A. Merging problem-based learning with simulation-based learning in the medical undergraduate curriculum: the PAIRED framework for enhancing lifelong learning. Cureus 2016 Jun 19;8(6):e647 [FREE Full text] [doi: 10.7759/cureus.647] [Medline: 27446767]
102. Theodoulou I, Nicolaides M, Athanasiou T, Papalois A, Sideris M. Simulation-based learning strategies to teach undergraduate students basic surgical skills: a systematic review. J Surg Educ 2018;75(5):1374-1388. [doi: 10.1016/j.jsurg.2018.01.013] [Medline: 29422405]
103. Creutzfeldt J, Hedman L, Felländer-Tsai L. Cardiopulmonary resuscitation training by avatars: a qualitative study of medical students' experiences using a multiplayer virtual world. JMIR Serious Games 2016 Dec 16;4(2):e22 [FREE Full text] [doi: 10.2196/games.6448] [Medline: 27986645]
104. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep 22;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
105. Maicher K, Danforth D, Price A, Zimmerman L, Wilcox B, Liston B, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Simul Healthc 2017 Apr;12(2):124-131. [doi: 10.1097/SIH.0000000000000195] [Medline: 28704290]
106. Staziaki PV, Sarangi R, Parikh U, Brooks JG, LeBedis CA, Shaffer K. An objective structured clinical examination for medical student radiology clerkships: reproducibility study. JMIR Med Educ 2020 May 06;6(1):e15444 [FREE Full text] [doi: 10.2196/15444] [Medline: 32374267]
107. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. part I: an historical and theoretical perspective. Med Teach 2013 Aug 22;35(9):e1437-e1446. [doi: 10.3109/0142159x.2013.818634]
108. Blythe J, Patel NS, Spiring W, Easton G, Evans D, Meskevicius-Sadler E, et al. Undertaking a high stakes virtual OSCE ("VOSCE") during COVID-19. BMC Med Educ 2021 Apr 20;21(1):221 [FREE Full text] [doi: 10.1186/s12909-021-02660-5] [Medline: 33879139]
109. García-Seoane JJ, Ramos-Rincón JM, Lara-Muñoz JP, CCS-OSCE working group of the CNDFME. Changes in the objective structured clinical examination (OSCE) of university schools of medicine during COVID-19. Experience with a computer-based case simulation OSCE (CCS-OSCE). Rev Clin Esp (Barc) 2021 Oct;221(8):456-463 [FREE Full text] [doi: 10.1016/j.rceng.2021.01.006] [Medline: 34217672]
110. Shorbagi S, Sulaiman N, Hasswan A, Kaouas M, Al-Dijani MM, El-Hussein RA, et al. Assessing the utility and efficacy of e-OSCE among undergraduate medical students during the COVID-19 pandemic. BMC Med Educ 2022 Mar 08;22(1):156 [FREE Full text] [doi: 10.1186/s12909-022-03218-9] [Medline: 35260144]
111. Agarwal A, Batra B, Sood AK, Ramakantan R, Bhargava SK, Chidambaranathan N, et al. Objective structured clinical examination in radiology. Indian J Radiol Imaging 2010 May 02;20(2):83-88 [FREE Full text] [doi: 10.4103/0971-3026.63040] [Medline: 20607015]
112. Pérez Baena AV, Sendra Portero F. The objective structured clinical examination (OSCE): main aspects and the role of imaging. Radiologia (Engl Ed) 2023 Jan;65(1):55-65 [FREE Full text] [doi: 10.1016/j.rxeng.2022.09.006] [Medline: 36842786]
113. Brigham TJ. An introduction to gamification: adding game elements for engagement. Med Ref Serv Q 2015;34(4):471-480. [doi: 10.1080/02763869.2015.1082385] [Medline: 26496401]
114. Seaborn K, Fels DI. Gamification in theory and action: a survey. Int J Hum Comput Interact 2015 Feb;74:14-31. [doi: 10.1016/j.ijhcs.2014.09.006]
115. Aguado-Linares P, Sendra-Portero F. Gamification: basic concepts and applications in radiology. Radiologia (Engl Ed) 2023 Mar;65(2):122-132 [FREE Full text] [doi: 10.1016/j.rxeng.2022.10.014] [Medline: 37059578]
116. Lee A, Berge Z. Second Life in healthcare education: virtual environment's potential to improve patient safety. Knowl Manag E-Learn 2011;3(1):17-23. [doi: 10.34105/j.kmel.2011.03.003]
117. Mayo Clinic in Second Life map location. Linden Lab Inc. URL: https://maps.secondlife.com/secondlife/Mayo%20Clinic/130/130/34 [accessed 2024-03-28]
118. USF Morsany College of Medicine in Second Life map location. Linden Lab Inc. URL: http://maps.secondlife.com/secondlife/USF%20Health%20Annex/103/157/25 [accessed 2024-03-28]
119. Brodmann brain in Inspiration Bay in Second Life map location. Linden Lab Inc. URL: http://maps.secondlife.com/secondlife/Inspiration%20Bay/60/60/3503 [accessed 2024-03-28]
120. Malaga Medical School in Second Life map location. Linden Lab Inc. URL: http://maps.secondlife.com/secondlife/Malaga%20Medical%20School/147/96/26 [accessed 2024-03-28]

XSL•FO
RenderX

121. Morales-Vadillo R, Guevara-Canales JO, Flores-Luján VC, Robello-Malatto JM, Bazán-Asencios RH, Cava-Vergiú CE. Use of virtual reality as a learning environment in dentistry. Gen Dent 2019;67(4):21-27. [Medline: 31355760]
122. Yellowlees PM, Cook JN. Education about hallucinations using an internet virtual reality system: a qualitative survey. Acad Psychiatry 2006 Dec 01;30(6):534-539. [doi: 10.1176/appi.ap.30.6.534] [Medline: 17139026]
123. Salajegheh A, Jahangiri A, Dolan-Evans E, Pakneshan S. A combination of traditional learning and e-learning can be more effective on radiological interpretation skills in medical students: a pre- and post-intervention study. BMC Med Educ 2016 Feb 03;16:46 [FREE Full text] [doi: 10.1186/s12909-016-0569-5] [Medline: 26842495]
124. den Harder AM, Frijlingh M, Ravesloot CJ, Oosterbaan AE, van der Gijp A. The importance of human-computer interaction in radiology e-learning. J Digit Imaging 2016 Apr;29(2):195-205 [FREE Full text] [doi: 10.1007/s10278-015-9828-y] [Medline: 26464115]
125. The medical master island: Second Life maps location. Linden Lab Inc. URL: http://maps.secondlife.com/secondlife/Medical%20Master%20Island/121/87/23 [accessed 2024-03-24]
126. Awan OA. Virtual radiology readouts after the coronavirus disease (COVID-19) pandemic. AJR Am J Roentgenol 2021 Sep;217(3):765-766. [doi: 10.2214/AJR.21.25607] [Medline: 33594906]
127. O'Byrne L, Gavin B, McNicholas F. Medical students and COVID-19: the need for pandemic preparedness. J Med Ethics 2020 Sep 03;46(9):623-626 [FREE Full text] [doi: 10.1136/medethics-2020-106353] [Medline: 32493713]
128. Byrnes YM, Civantos AM, Go BC, McWilliams TL, Rajasekaran K. Effect of the COVID-19 pandemic on medical student career perceptions: a national survey study. Med Educ Online 2020 Dec;25(1):1798088 [FREE Full text] [doi: 10.1080/10872981.2020.1798088] [Medline: 32706306]
129. Muñoz-Núñez CF. Online training in radiology during the COVID-19 pandemic. Radiologia (Engl Ed) 2022 Sep;64(5):433-444 [FREE Full text] [doi: 10.1016/j.rxeng.2022.09.004] [Medline: 36243443]
130. Lieux M, Sabottke C, Schachner ER, Pirtle C, Danrad R, Spieler B. Online conferencing software in radiology: recent trends and utility. Clin Imaging 2021 Aug;76:116-122 [FREE Full text] [doi: 10.1016/j.clinimag.2021.02.008] [Medline: 33592549]
131. Kauffman L, Weisberg EM, Fishman EK. Using Facebook live and Zoom as tools for online radiology education: a practical guide. Curr Probl Diagn Radiol 2022 Jul;51(4):423-426 [FREE Full text] [doi: 10.1067/j.cpradiol.2022.01.003] [Medline: 35135679]
132. McRoy C, Patel L, Gaddam DS, Rothenberg S, Herring A, Hamm J, et al. Radiology education in the time of COVID-19: a novel distance learning workstation experience for residents. Acad Radiol 2020 Oct;27(10):1467-1474 [FREE Full text] [doi: 10.1016/j.acra.2020.08.001] [Medline: 32800692]
133. Slanetz PJ, Bedi H, Kesler T, Chetlen A. Optimizing journal clubs in the post-COVID-19 era. J Am Coll Radiol 2020 Nov;17(11):1496-1498 [FREE Full text] [doi: 10.1016/j.jacr.2020.07.031] [Medline: 32846145]
134. De Lucia A, Francese R, Passero I, Tortora G. Development and evaluation of a virtual campus on Second Life: the case of SecondDMI. Comput Educ 2009 Jan;52(1):220-233. [doi: 10.1016/j.compedu.2008.08.001]
135. Chamorro-Atalaya O, Durán-Herrera V, Suarez-Bazalar R, Gonzáles-Pacheco A, Quipuscoa-Silvestre M, Hernández-Hernández M, et al. The Metaverse in University education during COVID-19: a systematic review of success factors. Int J Learn Teach Educ Res 2023 May 30;22(5):206-226. [doi: 10.26803/ijlter.22.5.10]
136. Kariuki D. Crystal Frost viewer development to continue despite Linden Lab announcing a mobile viewer. Hypergrid Business. 2023 Mar 22. URL: https://www.hypergridbusiness.com/2023/03/crystal-frost-viewer-development-to-continue-despite-linden-lab-announcing-a-mobile-viewer/ [accessed 2024-03-29]
137. Make your metaverse. Frame VR. URL: https://learn.framevr.io/ [accessed 2024-05-30]
138. Liberatore MJ, Wagner WP. Virtual, mixed, and augmented reality: a systematic review for immersive systems research. Virtual Real 2021 Jan 03;25(3):773-799. [doi: 10.1007/s10055-020-00492-0]
139. Zhang HL, Xue Y, Lu Y, Lee S. A fusion model: toward a virtual, physical, and cognitive integration and its principles. EEE Consum Electron Mag 2021 Jan 03;25(3):773-799. [doi: 10.1109/MCE.2023.3300488]

Abbreviations

AI: artificial intelligence
AR: augmented reality
CME: continuing medical education
LOR: League of Rays
MMORPG: massively multiplayer online role-playing game
OSCE: objective structured clinical examination
PBL: problem-based learning
SL: Second Life
TBL: team-based learning
VR: virtual reality
XR: extended reality

XSL•FO
RenderX

Edited by G Eysenbach; submitted 30.03.24; peer-reviewed by K Erenli, C Lian; comments to author 01.05.24; revised version received 03.06.24; accepted 11.06.24; published 06.08.24.

Please cite as:
Sendra-Portero F, Lorenzo-Álvarez R, Rudolphi-Solero T, Ruiz-Gómez MJ
The Second Life Metaverse and Its Usefulness in Medical Education After a Quarter of a Century
J Med Internet Res 2024;26:e59005
URL: https://www.jmir.org/2024/1/e59005
doi:10.2196/59005
PMID:

©Francisco Sendra-Portero, Rocío Lorenzo-Álvarez, Teodoro Rudolphi-Solero, Miguel José Ruiz-Gómez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Twenty-Five Years of Progress—Lessons Learned From JMIR Publications to Address Gender Parity in Digital Health Authorships: Bibliometric Analysis

Annika Meyer[1], MD; Thomas Streichert[1], Prof Dr Med

Institute of Clinical Chemistry, Faculty of Medicine and University Hospital, University Hospital Cologne, Cologne, Germany

Corresponding Author:
Annika Meyer, MD
Institute of Clinical Chemistry
Faculty of Medicine and University Hospital
University Hospital Cologne
Kerpener Str 62
Cologne, 50937
Germany
Email: annika.meyer@uk-koeln.de

Abstract

Background: Digital health research plays a vital role in advancing equitable health care. The diversity of research teams is thereby instrumental in capturing societal challenges, increasing productivity, and reducing bias in algorithms. Despite its importance, the gender distribution within digital health authorship remains largely unexplored.

Objective: This study aimed to investigate the gender distribution among first and last authors in digital health research, thereby identifying predicting factors of female authorship.

Methods: This bibliometric analysis examined the gender distribution across 59,980 publications from 1999 to 2023, spanning 42 digital health journals indexed in the Web of Science. To identify strategies ensuring equality in research, a detailed comparison of gender representation in JMIR journals was conducted within the field, as well as against a matched sample. Two-tailed Welch 2-sample *t* tests, Wilcoxon rank sum tests, and chi-square tests were used to assess differences. In addition, odds ratios were calculated to identify predictors of female authorship.

Results: The analysis revealed that 37% of first authors and 30% of last authors in digital health were female. JMIR journals demonstrated a higher representation, with 49% of first authors and 38% of last authors being female, yielding odds ratios of 1.96 (95% CI 1.90-2.03; $P<.001$) and 1.78 (95% CI 1.71-1.84; $P<.001$), respectively. Since 2008, JMIR journals have consistently featured a greater proportion of female first authors than male counterparts. Other factors that predicted female authorship included having female authors in other relevant positions and gender discordance, given the higher rate of male last authors in the field.

Conclusions: There was an evident shift toward gender parity across publications in digital health, particularly from the publisher JMIR Publications. The specialized focus of its sister journals, equitable editorial policies, and transparency in the review process might contribute to these achievements. Further research is imperative to establish causality, enabling the replication of these successful strategies across other scientific fields to bridge the gender gap in digital health effectively.

(J Med Internet Res 2024;26:e58950) doi:10.2196/58950

KEYWORDS

digital health; medical informatics, authorship; gender distribution; diversity; bibliometric; scientometric; algorithmic bias reduction; gender gap; JMIR Publications; authorships; author; authors; bibliometric analysis; equality; comparison; gender representation; journal; journals; article; articles; Web of Science; control group; comparative analysis; statistical analysis; gender

Introduction

On International Women's Day 2024, the World Health Organization highlighted the promotion of "gender equality through digital health" [1]. In industrialized countries, women are the leading users of digital health apps [2], wearable devices [3], and their combined applications [4]. Fulfilling the role of "digital carer," they use these tools not only for their own personal health management but also to attend to the health needs of family members under their care [5]. Despite their

XSL•FO
RenderX

predominant role as the principal consumers of digital health resources, a mere 3% of the 2728 digital health deals in the United States from 2011 to 2020 were specifically designed with women in mind [6]. One notable manifestation of this discrepancy is seen in the performance of activity trackers, which are most accurate when worn close to the body. However, the common absence of pockets in women's clothing often necessitates storing these devices in handbags, thereby potentially compromising their accuracy [7].

Within this framework, research plays an essential role in driving forward digital health innovations and promoting gender equality [8,9]. It is a fundamental component of the approach to "setting norms, developing evidence-based technical guidance and formulating direction to support decision-making in digital health," which is the strategic priority 1 of the Regional Digital Health Action Plan for the World Health Organization European Region 2023-2030 [10]. Consequently, the Organisation for Economic Co-operation and Development advocates for diversity in research teams as a means of closing the digital gender divide [11]. Creating such teams is vital for understanding societal challenges [7,9,12], enhancing performance [11], and mitigating bias in algorithms [7,11], prompting calls for more detailed investigations into diversity within digital health academia [7].

In response to this identified need in the literature, this study's objective is to provide a detailed investigation of gender distribution among authors in digital health, with a principal focus on the publisher JMIR Publications. JMIR Publications originated from the *Journal of Medical Internet Research*, an open access pioneer that published its first issue in 1999 [13]. Since then, JMIR Publications has become a leading publisher in the field of digital health, annually releasing more than 3500 papers distributed among more than 30 journals [14].

By examining the practices of JMIR Publications as a positive example, this research intends to identify factors predicting female authorship in the digital health academic community.

Methods

Inclusion Criteria

The selection criteria for our study included digital health journals listed in Web of Science and papers published between the beginning of 1999 and end of 2023. For the primary group of interest, we examined all journals indexed by JMIR Publications [15], using their ISSN numbers to search the Web of Science Master Journal List as of February 20, 2024. For any journals not listed in Web of Science, we conducted validation checks on the journals' official websites.

To maintain consistency across the groups in our study, we selected journals for the control group by applying the filter "Medical Informatics" in the Journal Citation Report. This criterion was chosen because JMIR Publications' flagship journal, *Journal of Medical Internet Research*, is classified under the same category in the Journal Citation Reports. By using this approach, we were able to identify 15 journals from JMIR Publications and 27 control journals. Following the extraction of metrics such as impact factor and quartile ranking from the Journal Citation Report (Multimedia Appendix 1), we filtered by ISSN and eISSN to identify a total of 76,142 publications in the Web of Science database on February 20, 2024. Following the automated removal of duplicates, incomplete names, and publications after 2023 using the R programming language (R Core Team, 2023) [16], 59,980 publications (79% of the initial data set) were retained for final analysis (Figure 1).

XSL•FO
RenderX

Figure 1. Inclusion and exclusion criteria for publications in the field of medical informatics, 1999-2023.

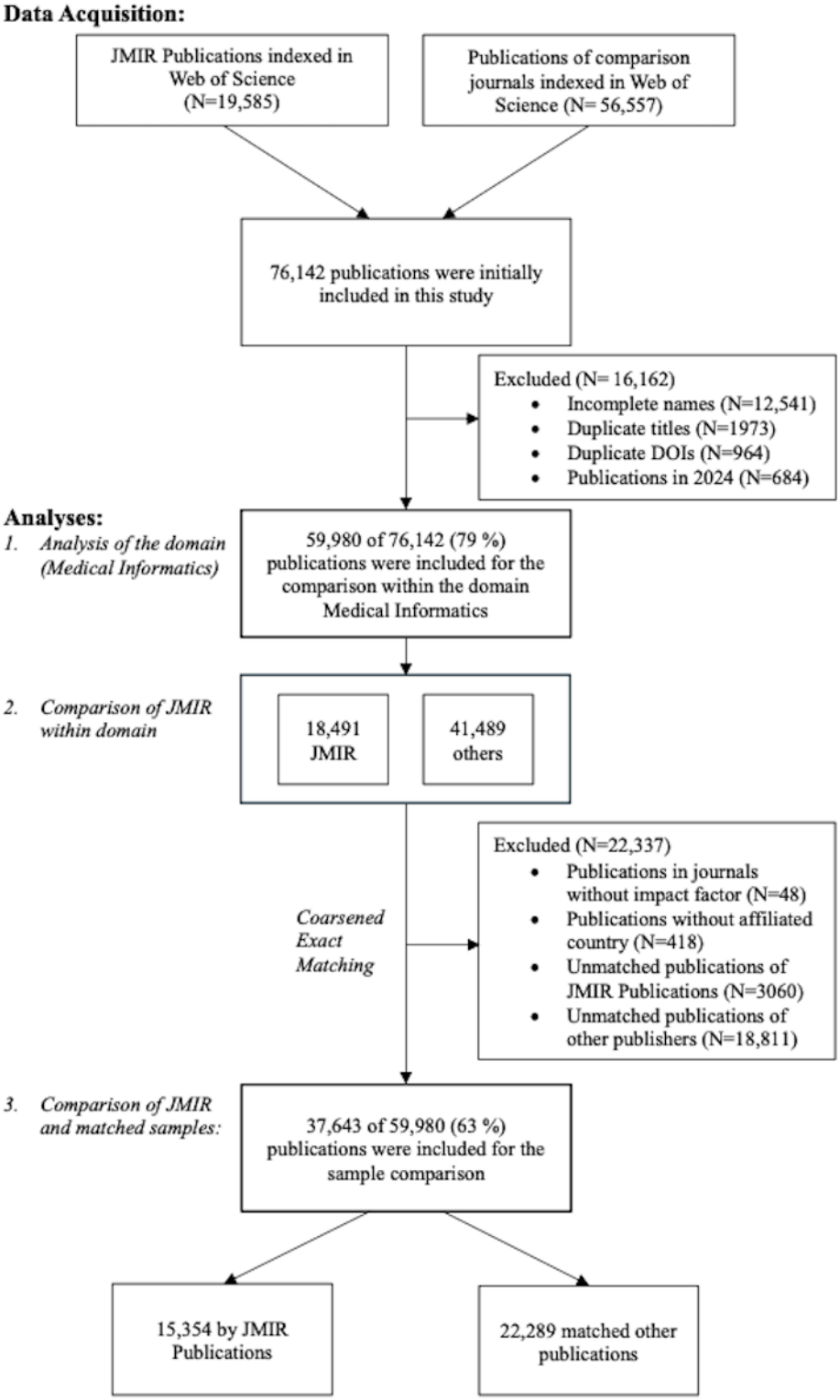

Gender Assignment

First and last authors' gender was determined based on their first names using the R programming language [16]. These authorship positions were specifically chosen due to their visibility and relevance within scholarly publishing [17,18]. To ensure accuracy, names with apostrophes were verified against the corresponding publications, resulting in corrections for 10 names. When a secondary name was provided in brackets, the bracketed name was preferred to reduce assignment errors. An automated mixed methods approach was adopted to categorize authorship into female, male, and unknown, combining genderize.io [19] and the Wiki-Gendersort name list [20], as recommended by Sebo [21]. Although this approach oversimplifies the concept of gender and does not fully capture diversity, it was a necessary compromise to facilitate the analysis of trends in authorship distribution.

XSL•FO
RenderX

Analysis Process

Our analysis then proceeded in 3 stages: first, we evaluated the gender distribution of first and last authors within digital health research. Next, we compared gender proportions between JMIR Publications and others in the field. Finally, we conducted a comparative analysis of JMIR Publications against a control sample of publications in digital health (Figure 1). As is recommended for scientometric studies [22], the sample of 37,643 publications, representing 63% of the initial 59,980, was selected using coarsened exact matching.

This approach aimed to ensure comparability while minimizing the impact of confounding variables, such as publication year, impact factor, document type, international collaboration, and affiliated country. Results of an alternative matching strategy using nearest neighbor matching for impact factors and exact matching for publication year, document type, international collaboration, and affiliated country can be found in Multimedia Appendix 2.

Statistical Analysis

Continuous variables (publication year, impact factor, usage count, and total times cited) were summarized using means and SDs, whereas categorical variables (quartile ranking, international collaboration, document type, and authors' gender) were described by frequencies and percentages.

Acknowledging the limitations of normal distribution testing and the robustness of parametric tests in large samples [23], we assumed normal distribution in accordance with the central limit theorem [24]. Differences in continuous variables were analyzed using 2-tailed Welch 2-sample *t* tests, while the Wilcoxon rank sum test was used for ordered categorical variables and chi-square tests for nominal variables. The analysis was further enhanced by calculating odds ratios (ORs) via logistic regression. We defined "Author Concordance" as a variable indicating gender alignment between the first and last authors (both female, both male, or both unknown). In addition, "female in contrasting author position" was categorized as a binary variable, with a value of "true" assigned if both the first and last authors were female and a value of "false" otherwise. To account for the large sample size, the α value for statistical significance was set at <.005, following the recommendations by Di Leo and Sardanelli [25].

We conducted our statistical analysis using the R programming language [16] and using the "rio" [26], "here" [27], "bibliometrix" [28], and "parallel" [16] packages for data import and processing. The "tidyverse" [29] and "fastDummies" [30] packages supported data manipulation, and the "MatchIt" [31] package enabled coarsened exact matching. Tables were generated using the "gtsummary" [32], "knitr" [33], and "labelled" packages [34], and "dplyr" (included in "tidyverse") as well as "cowplot" [35] was used for graphical presentation.

Ethical Considerations

All aspects of this research project comply with the ethical standards of the Ethics Committee of the University Hospital Cologne, which reviewed and approved the project on July 2, 2023 (application ID: 22-1436-retro).

Results

Publication Characteristics

Our analysis included 59,980 publications, with 18,491 from JMIR Publications and 41,489 from control journals in the field of medical informatics. The average publication year for JMIR papers was notably recent 2020 (3) compared with 2017 (5) for other publications in the field (*P*<.001). JMIR Publications also differed from others in medical informatics in terms of international collaborations (27% vs 25%, respectively, *P*<.001) as well as impact factor for 2022 (5 [2] vs 4 [3], respectively, *P*<.001). Table 1 provides additional distinctions, including usage frequency and citation metrics.

XSL•FO
RenderX

Table 1. Summary table of publications published in the field of medical informatics between 1999 and 2023[a].

Characteristic	Overall (59,980/59,980, 100%)[b]	JMIR, (18,491/59,980, 31%)[b]	Others in medical informatics (41,489/59,980, 69%)[b]	*P* value
Publication year (SD)	2018 (4)	2020 (3)	2017 (5)	<.001[c]
Quartile ranking, n/n (%)				<.001[d]
No quartile	6057/59,980 (10)	6057/18,491 (33)	0/41,489 (0)	
Quartile 1	19,382/59,980 (32)	8001/18,491 (43)	11,381/41,489 (27)	
Quartile 2	11,910/59,980 (20)	3424/18,491 (19)	8486/41,489 (20)	
Quartile 3	10,126/59,980 (17)	1009/18,491 (5.5)	9117/41,489 (22)	
Quartile 4	12,505/59,980 (21)	0/18,491 (0)	12,505/41,489 (30)	
Impact factor 2022, mean (SD)	4 (3)	5 (2)	4 (3)	<.001[e]
Usage count (since 2013), mean (SD)	14 (24)	15 (26)	14 (23)	.20[c]
Total times cited, mean (SD)	20 (70)	18 (49)	21 (77)	<.001[c]
International collaboration, n (%)				<.001[d]
No	44,486/59,980 (74)	13,485/18,491 (73)	31,001/41,489 (75)	
Yes	15,494/59,980 (26)	5,006/18,491 (27)	10,488/41,489 (25)	
Document type, n (%)				<.001[d]
Paper	52,685/59,980 (88)	15,819/18,491 (86)	36,866/41,489 (89)	
Editorial	1604/59,980 (2.7)	101/18,491 (0.5)	1503/41,489 (3.6)	
Letter	549/59,980 (0.9)	141/18,491 (0.8)	408/41,489 (1.0)	
Other	674/59,980 (1.1)	316/18,491 (1.7)	358/41,489 (0.9)	
Review	4468/59,980 (7.4)	2114/18,491 (11)	2354/41,489 (5.7)	

[a]Statistical significance is highlighted by *P* values in boldface.

[b]Mean (SD) or frequency (%).

[c]Two-tailed Welch 2-sample *t* test.

[d]Pearson chi-square test.

[e]Wilcoxon rank sum test.

Gender Distribution of Female Authorship

Among the analyzed publications in medical informatics, the gender distribution of first authors was 37% (22,450/59,980) female and 47% (28,299/59,980) male, while 30% (17,811/59,980) of last authors were female and 57% (34,463/59,980) male. Unspecified gender accounted for the remaining 13%-15% in authorship positions. In contrast to others in medical informatics, publications from JMIR Publications showed a significantly higher proportion of female authors in relevant positions, with 49% (8980/18,491) of first authors (*P*<.001) and 38% (7078/18,491) of last authors being female (*P*<.001). This statistically significant difference was further validated after applying the sampling strategy (first authorship: *P*<.001; last authorship: *P*<.001) (Multimedia Appendix 3; Figures 2A-2C).

XSL•FO
RenderX

Figure 2. Comparison of gender distribution and gender concordance in first and last authorship within medical informatics in regard to JMIR Publications, 1999-2023. The bar charts display the gender distribution as well as gender concordance in medical informatics (gender distribution: 2A; gender concordance: 2D), JMIR Publications within this field (gender distribution: 2B; gender concordance: 2E), and JMIR Publications compared with the sample of control publications after applying the sampling strategy (gender distribution: 2C; gender concordance: 2F). The proportion of male authors is shown in dark blue, female authors in light blue, and unknown authors in light green.

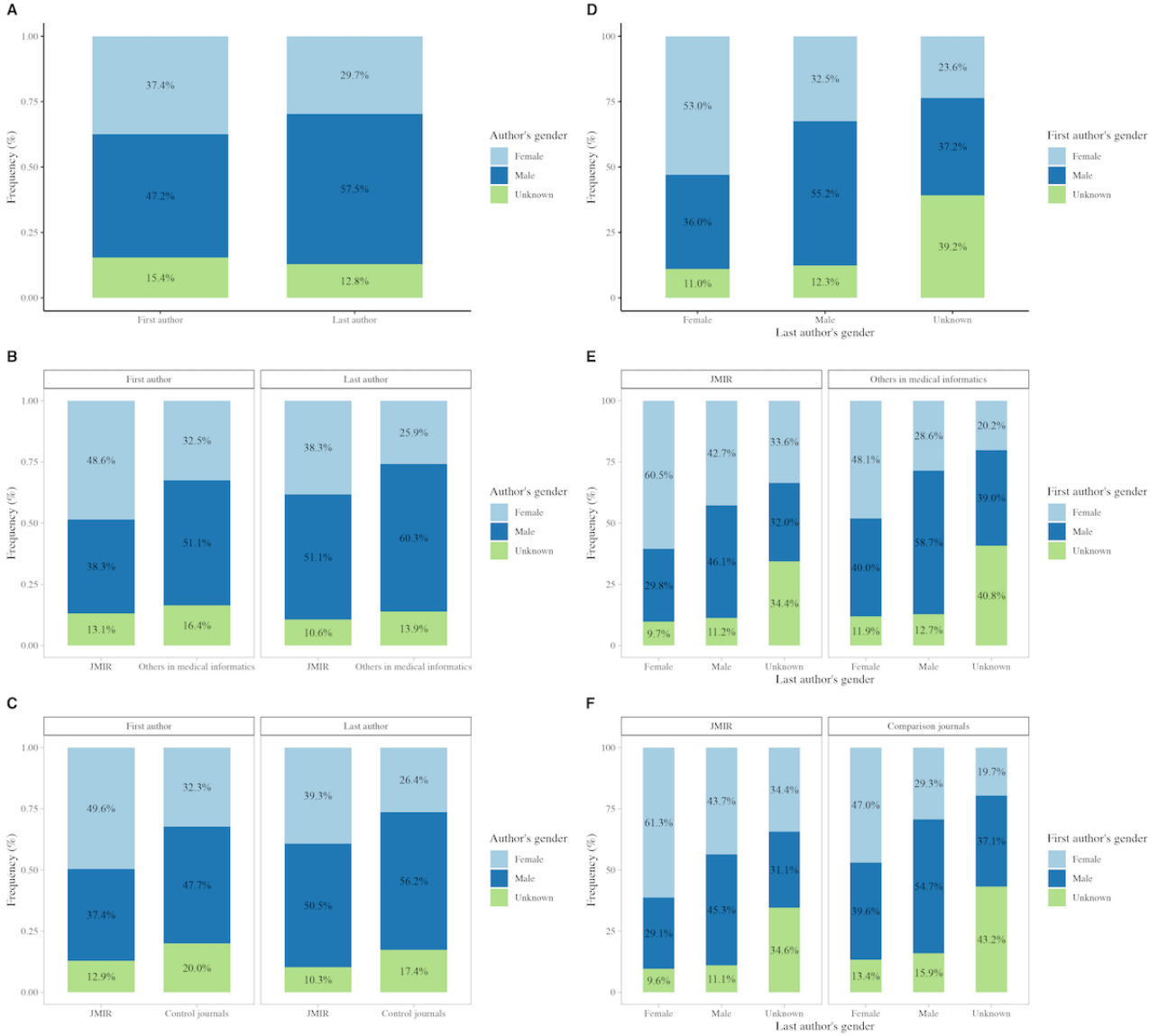

Predictors of Female Authorship

The strongest predictor of female authorship was the presence of female authors in contrasting authorship positions (OR 2.53, 95% CI 2.44-2.62; *P*<.001; Table 2).

Specifically, the proportion of female first authors was 53% (9444/17,811) when the last author was female, compared with only 32% (11,188/34,463) when the last author was male (*P*<.001; Figures 2D-2F). However, as the majority of last authors (34,463/59,980, 57%) were male, gender concordance emerged as a negative predictor of female first authorship (OR 0.51, 95% CI 0.49-0.53; *P*<.001).

Notably, the publisher JMIR Publications served as the second most significant predictor of both female first (OR 1.96, 95% CI 1.90-2.03; *P*<.001) and last authorship (OR 1.78, 95% CI 1.71-1.84; *P*<.001; Table 2). Even after applying the matching strategy, this association remained significant (Multimedia Appendix 4).

This trend is also reflected over time, with the ratio of female-to-male authors in JMIR Publications exceeding 1:1 since 2008 (Figure 3).

In addition, 7 JMIR Publications journals had more than 50% female first authors and only 1 had fewer female authors in relevant positions than the pooled analysis of control journals in the field (Figure 4).

XSL•FO
RenderX

Table 2. Odds ratios for female first and last authorship in the field of medical informatics between 1999 and 2023[a].

Characteristic (N=59,980)	First female authorship (N=22,450)			Last female authorship (N=17,811)		
	OR[b]	95% CI	*P* value	OR	95% CI	*P* value
Female in contrasting author position	2.53	2.44-2.62	*<.001*	2.53	2.44-2.62	*<.001*
Gender concordance	0.51	0.49-0.53	*<.001*	1.03	1.0-1.07	.10
Publication year	1.03	1.03-1.04	*<.001*	1.04	1.03-1.04	*<.001*
Journal of Medical Internet Research	1.96	1.90-2.03	*<.001*	1.78	1.71-1.84	*<.001*
Impact factor 2022	0.96	0.96-0.97	*<.001*	0.96	0.96-0.97	*<.001*
International collaboration	0.88	0.85-0.92	*<.001*	0.81	0.77-0.84	*<.001*

[a]Statistical significance is highlighted by *P* values in italics.

[b]OR: odds ratio.

XSL•FO
RenderX

Figure 3. Gender distribution of authorship in medical informatics over time, 1999-2023. The line graph in Figure 3A shows the proportion of female (light blue), male (dark blue), and unknown (green) authors of JMIR Publications and other publications in the field, as well as first and last authors. Line graphs in Figure 3B and Figure 3C show the proportion of male and female first authors (light blue) and last authors (dark blue) for the comparison of JMIR Publications in the field and in the control group. The red line shows gender parity with a ratio of 1:1. JMIR: Journal of Medical Internet Research.

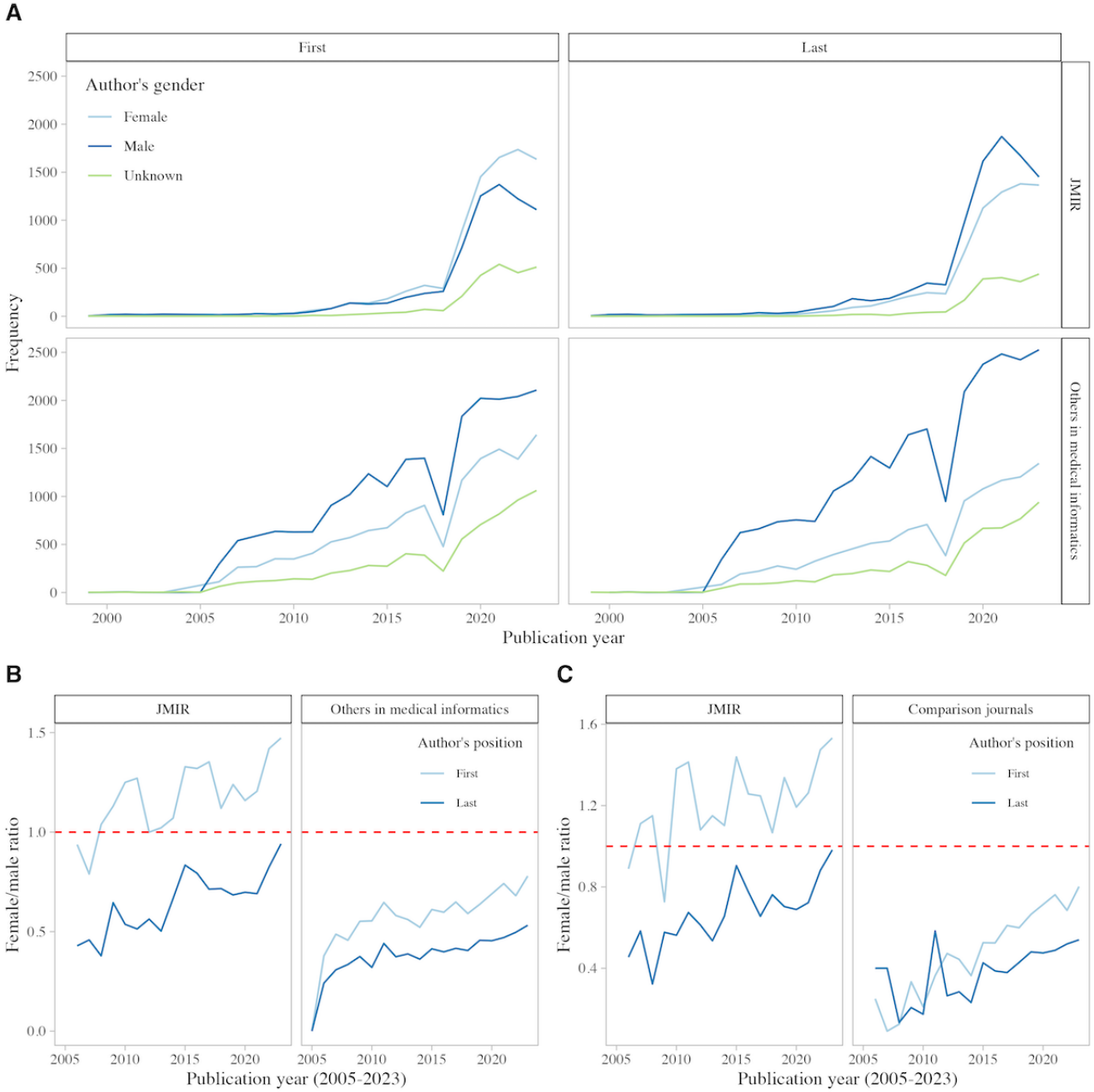

XSL•FO
RenderX

Figure 4. Gender distribution of authorships in medical informatics by journal, 1999-2023. This lollipop chart illustrates the percentage of female authors in first (light blue) and last author positions (dark blue) across various medical informatics journals. The red line signifies the 50% mark, indicating gender parity. The size of the lollipops correlates with the number of publications, providing a visual comparison of female authorship prevalence in the field. The black dashed line on the right indicates the proportion of female last authors, while the line on the left represents the proportion of female first authors in other medical informatics publications. JMIR: Journal of Medical Internet Research.

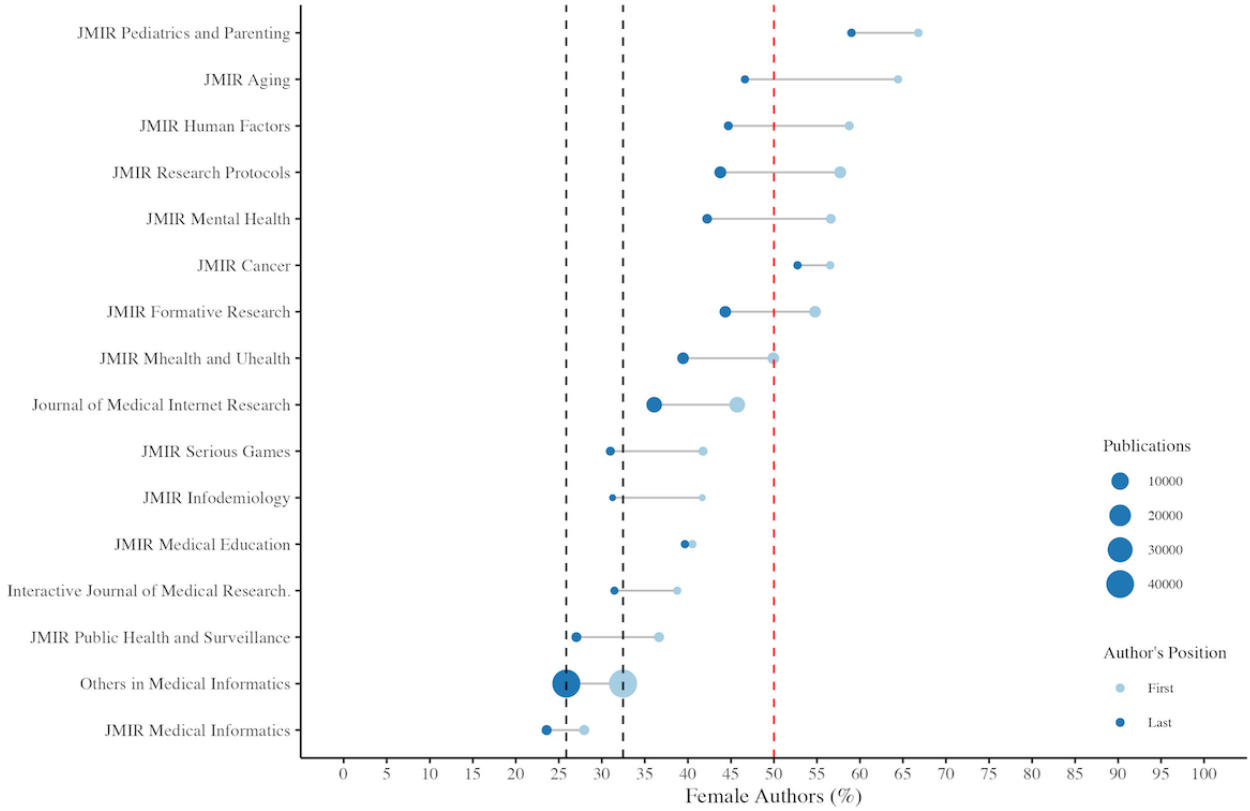

Discussion

Current Status of Female Authorship in Digital Health

The potential of diverse research teams to mitigate the digital health divide [7,9,11] and as such promote gender equality is increasingly recognized [1].

While the proportion of female authors in medical informatics remains below the 41% observed in the broader field of medicine, there has been a notable increase since 2018. At that time, women accounted for only 33% of first authors and 27% of last authors in this domain [36]. The low, albeit increasing, proportion of female authorship is also reflected at symposia [12,37,38] and within the submission as well as reviewer distribution of leading digital health journals [39]. This gender distribution might suggest biases against women researchers, highlighting the broader structural issues within the digital health sector, such as underrepresentation in leadership roles [7,40] and their disproportionate workload relative to recognition [7]. Thus, professional isolation, including exclusion from networks [40] and the decision-making processes [7,40], may play a more influential role than the often-assumed work-family balance [40]. Consistent with the literature [36], the observed low rate of international collaborations among female scientists and the higher frequency of female authorship in scenarios where women occupy the opposite author position seem to support this hypothesis.

While gender parity remains an aspiration yet to be fully realized in digital health, the current upward trend in female authorship deserves emphasis [12,36-38]. Factors contributing to this positive development include the increasing representation of women in the wider scientific community [36,41] and the implementation of targeted incentives and support initiatives, such as those offered by the American Medical Informatics Association [37,42].

Innovative Practices and Their Impact on Gender Representation

However, attributing the higher representation of female authors in relevant authorship positions solely to these factors cannot fully account for this trend. JMIR Publications achieving gender parity in first authorship 15 years ago, a milestone predicted to take 27 years to achieve [36], may provide insights into potential strategies to promote female representation in digital health research.

First, the publisher has established niches within various digital health topics through its sister journals [43], offering a specialized forum where competition is confined to works of similar thematic content. Considering the broad and multifaceted nature of the digital health field [44], this approach avoids the challenge of reflecting diversity in research topics and teams within a single journal.

Second, the origin of JMIR Publications as a “small independent open access project hosted at a university” [14] may have

XSL•FO
RenderX

facilitated more diverse authorship without the need to conform to traditional and potentially biased structures. For instance, a study from 2021 shows that female editors are underrepresented in 86 medical journals with an average rate of 27.9% [45]. In contrast, JMIR Publications' flagship journal, *Journal of Medical Internet Research*, currently has a higher proportion of female authors at 35% [46]. At *JMIR Pediatrics and Parenting*, the proportion of female editors is as high as 40%. Among their former editors 80% are female [46]. Given the likelihood of gender concordance between editors and authors in other disciplines [47], lower female presence in editorial boards may also hinder female representation in prominent digital health authorship positions. Reasons for this may include fewer invitations to submit for female researchers [36], unconscious bias in paper evaluation [47], and male editors' preference for male reviewers [48].

Third, JMIR Publications review process might also mitigate the reviewer bias female authors face in science [49-51]. This bias is often facilitated by the one-sided anonymous nature of single-blinded peer review, which remains the most common form of this process [52]. JMIR Publications, however, uses a signed peer review process [53], which acknowledges reviewers' work by publishing their names after paper publication [54,55]. As a potential side effect, disclosing the identity of reviewers can also enhance the quality and civility of peer review [56], potentially mitigating bias and thus promoting greater representation of female authors in relevant authorship positions.

Fourth, JMIR Publications offers an expedited review process for a fee [55]. This possibility might be interesting, especially considering that 38% of women in computer science, a field closely related to digital health, report experiencing slower career progression than their colleagues who completed their final degrees at the same time [57]. Female-authored papers tend to spend an additional 3-6 months in the review pipeline compared with male-authored papers, even when they score higher on readability [58]. As merit is often associated with being the first to publish [59], the availability of a faster review process might attract female authors to JMIR Publications. However, it is crucial to acknowledge that while expedited review offers benefits, it must not compromise the integrity of the peer review process, which serves as a critical instrument for quality assurance in academic publishing.

While the gender distribution of JMIR Publications is encouraging and the identified strategies might seem promising, further research to establish causality is essential in order to validate and subsequently apply the successful practices of JMIR Publications to other areas of science.

Conclusions

The landscape of digital health is witnessing an encouraging shift toward gender parity, with JMIR Publications playing a pivotal role in amplifying the visibility of female authors in relevant authorship positions. This notable achievement may stem from the creation of specialized niches within JMIR Publications' array of sister journals as well as its publishing practices, such as the transparent disclosure of reviewer identities. Future research into the context of JMIR Publications' high proportion of female authors is therefore needed to promote diversity in research teams so that research can act as a bridge in the digital health divide.

Limitations

The study design has some limitations. Due to the retrospective nature of the data collection, no conclusions can be drawn about causality. In addition, the data are not necessarily representative of other scientific fields or publications in other databases such as Scopus or PubMed. The binary approach to gender assignment is also not representative of actual gender diversity and neglects biases due to other marginalized group assignments such as ethnicity. The accuracy of the gender assignment approach depends on the geographical origin of the names, which may hypothetically lead to bias. It is also likely that not all confounding factors have been identified and corrected for.

Acknowledgments

The authors wish to thank Ari Soleman for proofreading this paper. ChatGPT as well as DeepL was used for linguistic and translation assistance. The statistical analysis was expedited by using ChatGPT to identify errors in the code as well as improve the code. All output by artificial intelligence was critically reviewed by the authors. We acknowledge support for the Article Processing Charge from the DFG (German Research Foundation, 491454339).

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of included journals in the field of Medical Informatics.
[XLSX File (Microsoft Excel File), 16 KB - jmir_v26i1e58950_app1.xlsx]

Multimedia Appendix 2
Comparison of JMIR Publications to Others in Medical Informatics after applying nearest neighbour and exact sampling strategy.
[PDF File (Adobe PDF File), 1074 KB - jmir_v26i1e58950_app2.pdf]

Multimedia Appendix 3

XSL•FO
RenderX

Gender Distribution of Authorship within Medical Informatics as well as after applying the sampling strategy.
[DOCX File , 17 KB - jmir_v26i1e58950_app3.docx]

Multimedia Appendix 4
Odds Ratios of Female First and Last Authorship after Applying the Exact Sampling Strategy.
[DOCX File , 15 KB - jmir_v26i1e58950_app4.docx]

References

1. Promoting gender equality through digital health. World Health Organization—Europe. 2024. URL: https://www.who.int/europe/news-room/events/item/2024/03/08/default-calendar/promoting-gender-equality-through-digital-health [accessed 2024-03-07]
2. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
3. Omura JD, Carlson SA, Paul P, Watson KB, Fulton JE. National physical activity surveillance: users of wearable activity monitors as a potential data source. Prev Med Rep 2017;5:124-126 [FREE Full text] [doi: 10.1016/j.pmedr.2016.10.014] [Medline: 28101443]
4. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the national cancer institute's health information national trends survey data. J Med Internet Res 2020;22(5):e16299 [FREE Full text] [doi: 10.2196/16299] [Medline: 32406865]
5. Lupton D, Maslen S. How women use digital technologies for health: qualitative interview and focus group study. J Med Internet Res 2019;21(1):e11481 [FREE Full text] [doi: 10.2196/11481] [Medline: 30681963]
6. Gambon E, Stotz C, Sandhu N. Femtech is expansive—it's time to start treating it as such. RockHealth. 2020. URL: https://rockhealth.com/femtech-is-expansive-its-time-to-start-treating-it-as-such/ [accessed 2024-03-05]
7. Figueroa CA, Luo T, Aguilera A, Lyles CR. The need for feminist intersectionality in digital health. Lancet Digit Health 2021;3(8):e526-e533 [FREE Full text] [doi: 10.1016/S2589-7500(21)00118-7] [Medline: 34325855]
8. Walsh E, Rooney M, Appleby L, Wilkinson G. Open peer review: a randomised controlled trial. Br J Psychiatry 2000;176:47-51. [doi: 10.1192/bjp.176.1.47] [Medline: 10789326]
9. Sinha C, Schryer-Roy AM. Digital health, gender and health equity: invisible imperatives. J Public Health (Oxf) 2018;40(suppl_2):ii1-ii5 [FREE Full text] [doi: 10.1093/pubmed/fdy171] [Medline: 30329082]
10. Regional Committee for Europe, 72nd session.. Seventy-Second Regional Committee for Europe: Tel Aviv, 12–14 September 2022: Regional Digital Health Action Plan for the WHO European Region 2023–2030. Copenhagen, Capital Region: World Health Organization. Regional Office for Europe; 2022.
11. Organisation for Economic Cooperation and Development (OECD). Bridging the digital gender divide—include upskill, innovate. 2018. URL: https://tinyurl.com/48awp35e [accessed 2024-06-29]
12. Moen A, Chronaki C, Petelos E, Voulgaraki D, Turk E, Névéol A. Diversity in health informatics: mentoring and leadership. Stud Health Technol Inform 2021;281:1031-1035. [doi: 10.3233/SHTI210341] [Medline: 34042835]
13. Eysenbach G. Celebrating 20 years of open access and innovation at JMIR publications. J Med Internet Res 2019;21(12):e17578 [FREE Full text] [doi: 10.2196/17578] [Medline: 31868653]
14. JMIR Publications. About the publisher. URL: https://www.jmir.org/about-journal/about-the-publisher [accessed 2024-04-05]
15. JMIR Editorial Team. JMIR Publications Knowledge Base and Help Center: JMIR Publications. Which journal titles does JMIR publications currently publish? (Journal Portfolio). Toronto, Ontario: JMIR Publications; 2023. URL: https://tinyurl.com/5n8wke9y [accessed 2024-02-20]
16. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
17. Zimba O, Gasparyan AY. Scientific authorship: a primer for researchers. Reumatologia 2020;58(6):345-349 [FREE Full text] [doi: 10.5114/reum.2020.101999] [Medline: 33456076]
18. Wren JD, Kozak KZ, Johnson KR, Deakyne SJ, Schilling LM, Dellavalle RP. The write position. A survey of perceived contributions to papers based on byline position and number of authors. EMBO Rep 2007;8(11):988-991 [FREE Full text] [doi: 10.1038/sj.embor.7401095] [Medline: 17972896]
19. Wais K. Gender prediction methods based on first names with genderizeR. R J 2016;8:17-37. [doi: 10.32614/rj-2016-002]
20. Bérubé N, Sainte-Marie M, Ghiasi G, Larivière V. Wiki-Gendersort: automatic gender detection using first names in Wikipedia. 2020. URL: https://tinyurl.com/3dus5nb9 [accessed 2024-06-29]
21. Sebo P. Performance of gender detection tools: a comparative study of name-to-gender inference services. J Med Libr Assoc 2021;109(3):414-421 [FREE Full text] [doi: 10.5195/jmla.2021.1185] [Medline: 34629970]
22. Bittmann F, Tekles A, Bornmann L. Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example. Quant Sci Stud 2021;2(4):1246-1270. [doi: 10.1162/qss_a_00158]

XSL•FO
RenderX

23. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 2012;10(2):486-489 [FREE Full text] [doi: 10.5812/ijem.3505] [Medline: 23843808]
24. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017;70(2):144-156 [FREE Full text] [doi: 10.4097/kjae.2017.70.2.144] [Medline: 28367284]
25. Di Leo G, Sardanelli F. Statistical significance: p value, 0.05 threshold, and applications to radiomics-reasons for a conservative approach. Eur Radiol Exp 2020;4(1):18 [FREE Full text] [doi: 10.1186/s41747-020-0145-y] [Medline: 32157489]
26. Chan CH, Leeper TJ, Becker J, Schoch D. rio: a Swiss-army knife for data file I/O. 2023. URL: https://github.com/gesistsa/rio [accessed 2024-06-29]
27. Müller K, Bryan J. here: a simpler way to find your files. 2020. URL: https://cran.r-project.org/web//packages//here/here.pdf [accessed 2020-12-13]
28. Aria M, Cuccurullo C. bibliometrix : an R-tool for comprehensive science mapping analysis. Journal Informetrics 2017;11(4):959-975. [doi: 10.1016/j.joi.2017.08.007]
29. Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, François R, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4(43):1686. [doi: 10.21105/joss.01686]
30. Kaplan J. fastDummies: fast creation of dummy (binary) columns and rows from categorical variables. 2023. URL: https://cloud.r-project.org/web/packages/fastDummies/fastDummies.pdf [accessed 2023-07-06]
31. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42(8):1-28. [doi: 10.18637/jss.v042.i08]
32. Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. R J 2021;13(1):570-580. [doi: 10.32614/rj-2021-053]
33. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R. Boca Raton, FL: Chapman and Hall/CRC; 2015.
34. Larmarange J. labelled: manipulating labelled data. 2024. URL: https://cran.r-project.org/web//packages/labelled/labelled.pdf [accessed 2024-04-23]
35. Wilke CO. Cowplot: streamlined plot theme and plot annotations for 'ggplot2'. 2024. URL: https://github.com/wilkelab/cowplot [accessed 2023-12-15]
36. Holman L, Stuart-Fox D, Hauser CE. The gender gap in science: how long until women are equally represented? PLoS Biol 2018;16(4):e2004956 [FREE Full text] [doi: 10.1371/journal.pbio.2004956] [Medline: 29672508]
37. Hartzler AL, Leroy G, Daurelle B, Ochoa M, Williamson J, Cohen D, et al. Comparison of women and men in biomedical informatics scientific dissemination: retrospective observational case study of the AMIA annual symposium: 2017-2020. J Am Med Inform Assoc 2021;28(9):1928-1935 [FREE Full text] [doi: 10.1093/jamia/ocab097] [Medline: 34179991]
38. Sullivan S, Posada A, Hawkins M, Higueras G. Where are the female experts? Perceptions about the absence of female presenters at a Digital Health Conference in Bolivia. Mhealth 2020;6:37 [FREE Full text] [doi: 10.21037/mhealth-2019-di-12] [Medline: 33437834]
39. The Lancet Digital Health. Empowering women in health technology. Lancet Digit Health 2022;4(3):e149 [FREE Full text] [doi: 10.1016/S2589-7500(22)00028-0] [Medline: 35216746]
40. Denend L, McCutcheon S, Regan M, Sainz M, Yock P, Azagury D. Analysis of gender perceptions in health technology: a call to action. Ann Biomed Eng 2020;48(5):1573-1586 [FREE Full text] [doi: 10.1007/s10439-020-02478-0] [Medline: 32078709]
41. Charlesworth TES, Banaji MR. Gender in science, technology, engineering, and mathematics: issues, causes, solutions. J Neurosci 2019;39(37):7228-7243 [FREE Full text] [doi: 10.1523/JNEUROSCI.0475-18.2019] [Medline: 31371423]
42. Griffin AC, Leung TI, Tenenbaum JD, Chung AE. Gender representation in U.S. biomedical informatics leadership and recognition. J Am Med Inform Assoc 2021;28(6):1270-1274 [FREE Full text] [doi: 10.1093/jamia/ocaa344] [Medline: 33555005]
43. Which journal titles does JMIR publications currently publish? (Journal Portfolio). JMIR Editorial Team. 2024. URL: https://jmir.zendesk.com/hc/en-us/articles/115001442707 [accessed 2024-01-28]
44. Sharma R, Kshetri N. Digital healthcare: historical development, applications, and future research directions. Int J Inf Manage 2020;53:102105 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102105]
45. Starchl C, Shah V, Zollner-Schwetz I, Knezevic J, Geiger S, Amrein K. A comparison of the representation of women in editor positions at major medical journals in 2021 vs 2011. Acad Med 2023;98(1):75-79 [FREE Full text] [doi: 10.1097/ACM.0000000000004964] [Medline: 36576769]
46. JMIR Publications. JMIR pediatrics and parenting:. Editorial Board. URL: https://pediatrics.jmir.org/about-journal/editorial-board [accessed 2024-03-06]
47. Last K, Hübsch L, Cevik M, Wolkewitz M, Müller SE, Huttner A, et al. Association between women's authorship and women's editorship in infectious diseases journals: a cross-sectional study. Lancet Infect Dis 2022;22(10):1455-1464. [doi: 10.1016/S1473-3099(22)00367-X] [Medline: 35839790]
48. Ross E. Gender bias distorts peer review across fields. Nature 2017 [FREE Full text] [doi: 10.1038/nature.2017.21685]

XSL•FO
RenderX

49. Fox CW, Paine CET. Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. Ecol Evol 2019;9(6):3599-3619 [FREE Full text] [doi: 10.1002/ece3.4993] [Medline: 30962913]
50. Kaatz A, Gutierrez B, Carnes M. Threats to objectivity in peer review: the case of gender. Trends Pharmacol Sci 2014;35(8):371-373 [FREE Full text] [doi: 10.1016/j.tips.2014.06.005] [Medline: 25086743]
51. Murrar S, Johnson PA, Lee YG, Carnes M. Research conducted in women was deemed more impactful but less publishable than the same research conducted in men. J Womens Health (Larchmt) 2021;30(9):1259-1267 [FREE Full text] [doi: 10.1089/jwh.2020.8666] [Medline: 33719578]
52. Jones N, Career news: nature. Authors' names have 'astonishing' influence on peer reviewers. 2022. URL: https://www.nature.com/articles/d41586-022-03256-9 [accessed 2024-03-06]
53. Eysenbach G, Gunther Eysenbach's Random Research Rants. Open peer review—trials and tribulations. 2008. URL: https://gunther-eysenbach.blogspot.com/2008/03/open-peer-review-trials-and.html [accessed 2024-06-29]
54. Eysenbach G. Should reviewers of papers have their names published? Practical example gives encouraging insights. BMJ 2004;329(7457):113. [doi: 10.1136/bmj.329.7457.113-b]
55. JMIR Publications. Peer review process. URL: https://www.jmir.org/author-information/peer-review-process [accessed 2024-03-06]
56. Ware M, Consortium PR. Peer Review: Benefits, Perceptions and Alternatives. London: Publishing Research Consortium; 2008.
57. Chiu MH, Cesa M. Gender gap in science. A global approach to the gender gap in mathematical, computing, and natural sciences: how to measure it, how to reduce it? Chem Int 2020;42(3):16-21. [doi: 10.1515/ci-2020-0306]
58. Hengel E. Publishing while female: are women held to higher standards? Evidence from peer review. Econ J 2022;132(648):2951-2991. [doi: 10.1093/ej/ueac032]
59. Stein C, Hill R. Race to the bottom: how competition to publish first can hurt scientific quality. Institute for Economic Policy Research (SIEPR): Stanford. 2021. URL: https://tinyurl.com/5f93tnmz [accessed 2021-12-01]

Abbreviations

OR: odds ratio

Edited by G Eysenbach; submitted 28.03.24; peer-reviewed by LC Pereyra, P Iranmanesh; comments to author 04.05.24; revised version received 08.05.24; accepted 29.05.24; published 09.08.24.

Please cite as:
Meyer A, Streichert T
Twenty-Five Years of Progress—Lessons Learned From JMIR Publications to Address Gender Parity in Digital Health Authorships: Bibliometric Analysis
J Med Internet Res 2024;26:e58950
URL: https://www.jmir.org/2024/1/e58950
doi:10.2196/58950
PMID:

©Annika Meyer, Thomas Streichert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Impact of 25 Years of Mobile Health Tools for Pain Management in Patients With Chronic Musculoskeletal Pain: Systematic Review

Jenny Lin-Hong Shi[1], MD, PhD; Regina Wing-Shan Sit[1], MD, PhD

Department of Medicine, Jockey Club School of Public Health and Primary Care, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, China

Corresponding Author:
Regina Wing-Shan Sit, MD, PhD
Department of Medicine
Jockey Club School of Public Health and Primary Care
Prince of Wales Hospital, The Chinese University of Hong Kong
4/F, JC School of Public Health and Primary Care Building
Shatin, NT
Hong Kong, 999077
China
Phone: 852 25039406
Fax: 852 26095998
Email: reginasit@cuhk.edu.hk

Abstract

Background: Mobile technologies are increasingly being used in health care and public health practice for patient communication, monitoring, and education. Mobile health (mHealth) tools have also been used to facilitate adherence to chronic musculoskeletal pain (CMP) management, which is critical to achieving improved pain outcomes, quality of life, and cost-effective health care.

Objective: The aim of this systematic review was to evaluate the 25-year trend of the literature on the adherence, usability, feasibility, and acceptability of mHealth interventions in CMP management among patients and health care providers.

Methods: We searched the PubMed, Cochrane CENTRAL, MEDLINE, EMBASE, and Web of Science databases for studies assessing the role of mHealth in CMP management from January 1999 to December 2023. Outcomes of interest included the effect of mHealth interventions on patient adherence; pain-specific clinical outcomes after the intervention; and the usability, feasibility, and acceptability of mHealth tools and platforms in chronic pain management among target end users.

Results: A total of 89 articles (26,429 participants) were included in the systematic review. Mobile apps were the most commonly used mHealth tools (78/89, 88%) among the included studies, followed by mobile app plus monitor (5/89, 6%), mobile app plus wearable sensor (4/89, 4%), and web-based mobile app plus monitor (1/89, 1%). Usability, feasibility, and acceptability or patient preferences for mHealth interventions were assessed in 26% (23/89) of the studies and observed to be generally high. Overall, 30% (27/89) of the studies used a randomized controlled trial (RCT), cohort, or pilot design to assess the impact of the mHealth intervention on patients' adherence, with significant improvements (all $P<.05$) observed in 93% (25/27) of these studies. Significant (judged at $P<.05$) between-group differences were reported in 27 of the 29 (93%) RCTs that measured the effect of mHealth on CMP-specific clinical outcomes.

Conclusions: There is great potential for mHealth tools to better facilitate adherence to CMP management, and the current evidence supporting their effectiveness is generally high. Further research should focus on the cost-effectiveness of mHealth interventions for better incorporating these tools into health care practices.

Trial Registration: International Prospective Register of Systematic Reviews (PROSPERO) CRD42024524634; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=524634

(J Med Internet Res 2024;26:e59358) doi:10.2196/59358

KEYWORDS

mHealth; mobile health; mobile app; chronic musculoskeletal pain; pain management; patient compliance; adherence; usability; feasibility; acceptability; PRISMA

XSL•FO
RenderX

Introduction

Chronic musculoskeletal pain (CMP) is defined as musculoskeletal pain that persists or recurs for longer than 3 months [1]. Interventions to help manage CMP are usually based on multimodal and biopsychosocial models [2-4]. CMP is a global burden, affecting approximately 1 in 5 adults [5]. One study indicated that over 70% of people 65 years and older have experienced an episode of joint pain [6]. Given that the percentage of the population 65 years and older is expected to increase from 15% to 24% by 2060, chronic musculoskeletal conditions will definitely become an increasing burden for the health care system [7].

The long-term nature and frequent need for continuous monitoring in CMP management gave rise to early developments in telemonitoring and telehealth. These innovations designed to improve CMP management and prevent disability and death have been improved by ongoing technological advancements. One such advancement is mobile device–based health care, or mobile health (mHealth). In 2022, the number of mobile users worldwide stood at 7.26 billion, which is projected to reach 7.49 billion by 2025 [8]. Mobile technologies such as smartphones and wireless monitoring devices are increasingly finding innovative applications and have emerged as potential alternatives to support the self-management of patients with CMP [9]. These applications include communication, data collection, patient monitoring, and education, as well as to facilitate adherence to CMP management [9].

The available evidence has pointed to the promising effects of mHealth interventions on CMP. A recent review [10] evaluated the effectiveness of app-based interventions on several CMP conditions (including general chronic pain, osteoarthritis [OA] pain, chronic neck pain [CNP], chronic low back pain [CLBP], rheumatoid arthritis, menstrual pain, migraine-related pain, and frozen shoulder pain), stating that mobile apps are significantly more effective in reducing pain compared with control conditions. Du et al [11] analyzed the use of web-based and mHealth interventions in patients with CLBP, showing that mHealth tools had a better effect on both pain and functional outcomes. In a similar vein, Thurnheer et al [12] analyzed the efficacy of mobile app usage in the management of patients with cancer and noncancer pain (eg, acute pain, general chronic pain, CNP, CLBP, and menstrual pain), reporting beneficial effects on pain relief, particularly in the out-clinic setting.

The evidence of the use of mHealth systems is still emerging, mainly focusing on its effect on clinical outcomes. However, current CMP management often requires a long-term care plan. Adherence to CMP management is critical to achieving improved outcomes, quality of life (QoL), and cost-effective health care [13]. A review of adherence behaviors from the World Health Organization noted that “increasing adherence may have greater effects on health than improvements with specific medical therapy” [14]. The true impact of these mHealth tools on adherence to treatment regimens may be overlooked, as mHealth promoters are more eager to show their effects on clinical outcomes (eg, morbidity, mortality, and biometric markers of clinical disease). Adherence to CMP treatment is a critical link that would connect the promise of mHealth to the ultimate goal of improving clinical outcomes.

To the best of our knowledge, no review published to date has examined the effects of using mHealth tools on the adherence to management and clinical outcomes of patients with CMP. Therefore, the aim of this systematic review was to provide an overview of the evidence with respect to a broad range of outcomes, including feasibility, usability, acceptability, and adherence, of mHealth tools to impact CMP-specific clinical outcomes. This approach enabled considering mHealth tools at all stages of development and to gauge the effectiveness of these tools across a range of technologies and CMP subtypes, many of which have overlapping treatment regimens and require similar adherence behaviors.

Methods

Overview

The protocol of this systematic review was registered on the International Prospective Register of Systematic Reviews (PROSPERO) database (CRD42024524634). The review was performed following the 2020 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [15].

We undertook a 25-year systematic review of mHealth interventions used to facilitate the self-management of CMP and adherence to treatment and management regimens. For this review, CMP included all types of chronic primary and secondary musculoskeletal pain based on the complete list of CMP conditions in the *International Classification of Diseases, 11th edition* (ICD-11) foundation layer supplement [16]. Our definition of mHealth was adopted from the Global Observatory for eHealth definition: “medical and public health practice supported by mobile devices” [17]. Given the comprehensive nature of CMP, this review goes beyond defining adherence as compliance with a treatment regimen by including a wide range of interventions such as medication reminders, symptom monitoring, educational tools, and facilitated patient-provider communication [18].

Search Strategy

The search strategy was based on CMP diseases according to the ICD-11 [19]. The search was conducted from January 1999 through December 2023. Using Boolean phrases, we searched the PubMed, MEDLINE, Cochrane (CENTRAL), EMBASE, and Web of Science databases for studies that assessed the role of mHealth interventions in CMP management. The search strategy was first developed for the PubMed database using Medical Subject Headings (MeSH) and was adapted for other databases. The search was filtered either by English language or by date of publication or publication type as an article. The detailed search strategy for each database is provided in Multimedia Appendix 1.

Inclusion Criteria

We included original research published in peer-reviewed journals that evaluated the effects of mHealth tools on CMP-related clinical outcomes; adherence to management; and

XSL•FO
RenderX

usability, feasibility, and acceptability features. All available study designs (randomized controlled trials [RCTs], cohort studies, cross-sectional studies, case series, case reports, questionnaires, mixed methods studies, and qualitative studies) were eligible for inclusion. Allowing for flexibility in the outcomes measured and in study design was necessary for an inclusive view of mHealth interventions at all stages of design, development, and evaluation. The detailed inclusion criteria for study selection and mHealth tools are provided in Multimedia Appendix 2.

Exclusion Criteria

Only articles reporting mHealth interventions designed for CMP were included. We excluded articles regarding interventions that were not tested in a sample population with clearly described methods and results. In addition, review articles, editorials, commentaries, dissertations, poster presentations, abstracts, proposals for future studies, study protocols, and descriptive articles describing new tools but not testing them in a sample population were excluded. The publication language was restricted to English.

Data Extraction and Analysis

Publications were initially screened for potential inclusion based on a simultaneous review of titles and abstracts by 2 independent reviewers. Any discrepancies were resolved by consensus among the researchers. Information, including publication year, location, study sample characteristics, types of mobile technology used, intervention and control details (if available for the given study design), outcomes measured, and summary results reported, was extracted and compiled using a Microsoft Excel spreadsheet.

We performed descriptive analyses of the data and summarized the findings from the included studies with an emphasis on statistical results reported in RCTs and cohort studies. We highlighted differences between groups when these results were available. Studies were organized for analysis based on the primary objective of the study and the key outcomes measured. Outcomes were organized into the following categories: (1) usability, feasibility, and acceptability of the mHealth tool; (2) effect of the mHealth intervention on adherence to CMP management; and (3) effect of the mHealth intervention on pain outcomes.

Results

Summary

In all, 866 articles were retrieved in full text and assessed for eligibility, among which 383 articles were excluded due to duplication. Based on the search criteria, 337 articles were excluded due to not meeting the study design criteria or not aligning with the definition of mHealth used in our study. A total of 142 articles were included in the full-text retrieval process, of which 53 were excluded due to an inappropriate study design, intervention, or population or no access to the full text. Finally, a total of 89 articles were eligible and met all inclusion criteria; the details of these included studies are provided in Multimedia Appendix 3 [20-107]. Multimedia Appendix 4 illustrates the PRISMA flowchart for the study selection process and Multimedia Appendix 5 provides a list of excluded studies with reasons for exclusion.

Study Characteristics

The publication time period for this 25-year systematic review spanned from 1999 to 2023, with an overall increase in articles published after 2017 and a significant increase during the COVID-19 pandemic from 2019 to 2022 (Figure 1). A total of 26 countries published eligible studies across 5 continents (Asia, Europe, Africa, America, and Oceania) from 1999 to 2023. Most of the eligible studies were published in Europe (Figure 2). The top-3 ranked countries publishing in this field were the United States (22/89, 25%), Germany (11/89, 12%), and Australia (7/89, 8%) (Figure 3). The vast majority of technology development for mHealth has been based in the United States, especially in the earlier days of smartphone introduction.

RCTs that assessed the differences between different mHealth tools or between an mHealth tool and standard or control care were the most common study designs, accounting for 35% (31/89) of the included studies, followed by cohort (23/89, 26%), pilot (14/89, 16%), cross-sectional (8/89, 8%), qualitative (7/89, 8%), and questionnaire (4/89, 4%) study designs; there was only 1 (1%) case series, case report, and cost-effectiveness study each. Study durations ranged from only a few hours to 12 months depending on the study design.

XSL•FO
RenderX

Figure 1. Trend in publication number from 1999 to 2024.

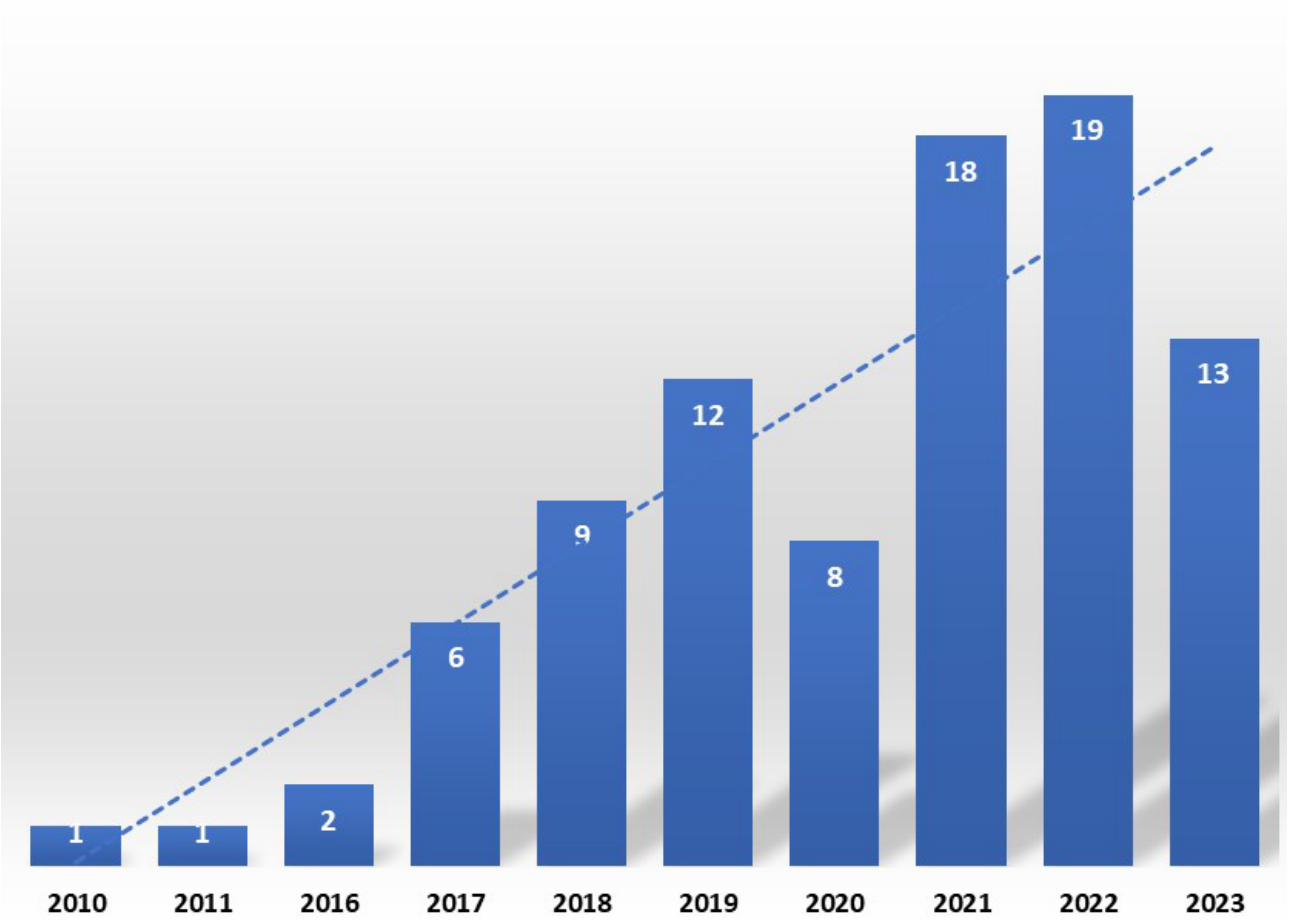

Figure 2. Eligible studies published by country. The size of the purple circles indicates the number of publications, with larger circles reflecting a higher number.

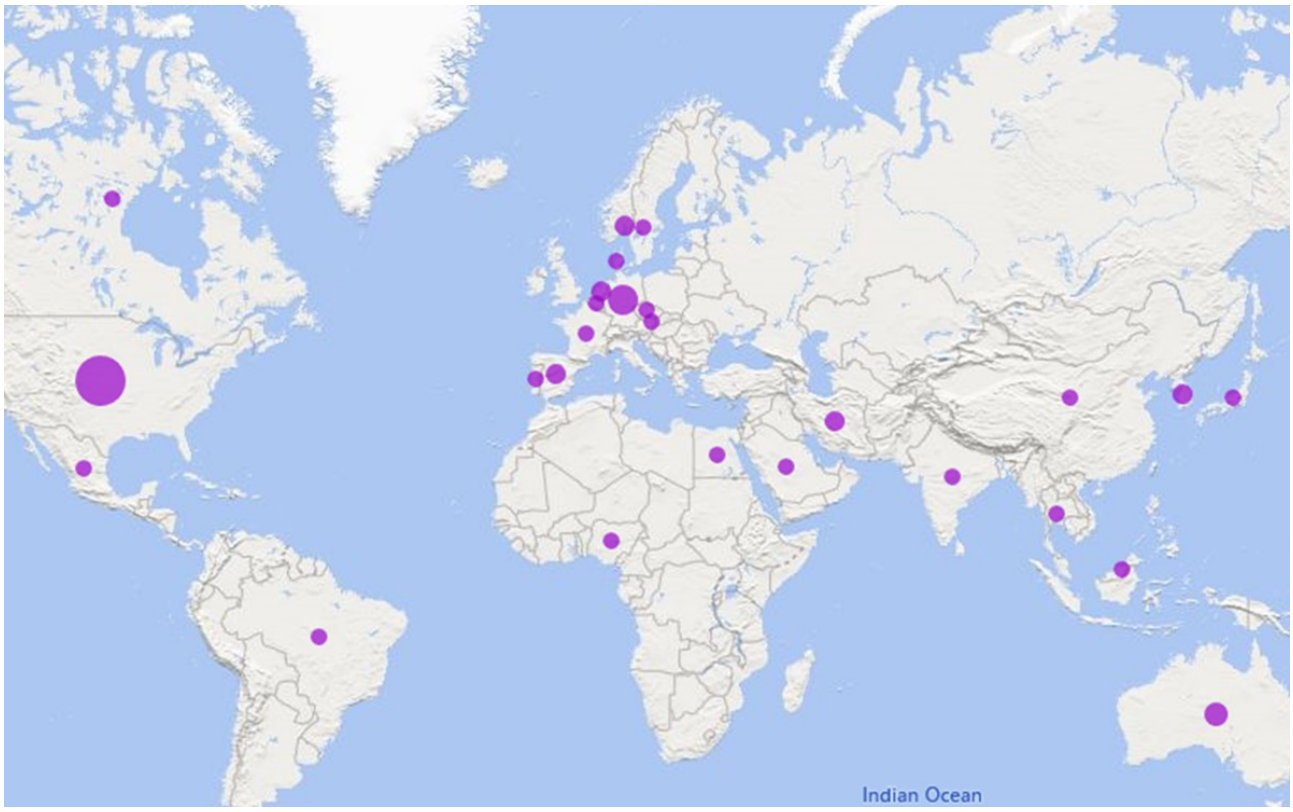

XSL•FO
RenderX

Figure 3. Publications from different countries from 1999 to 2023.

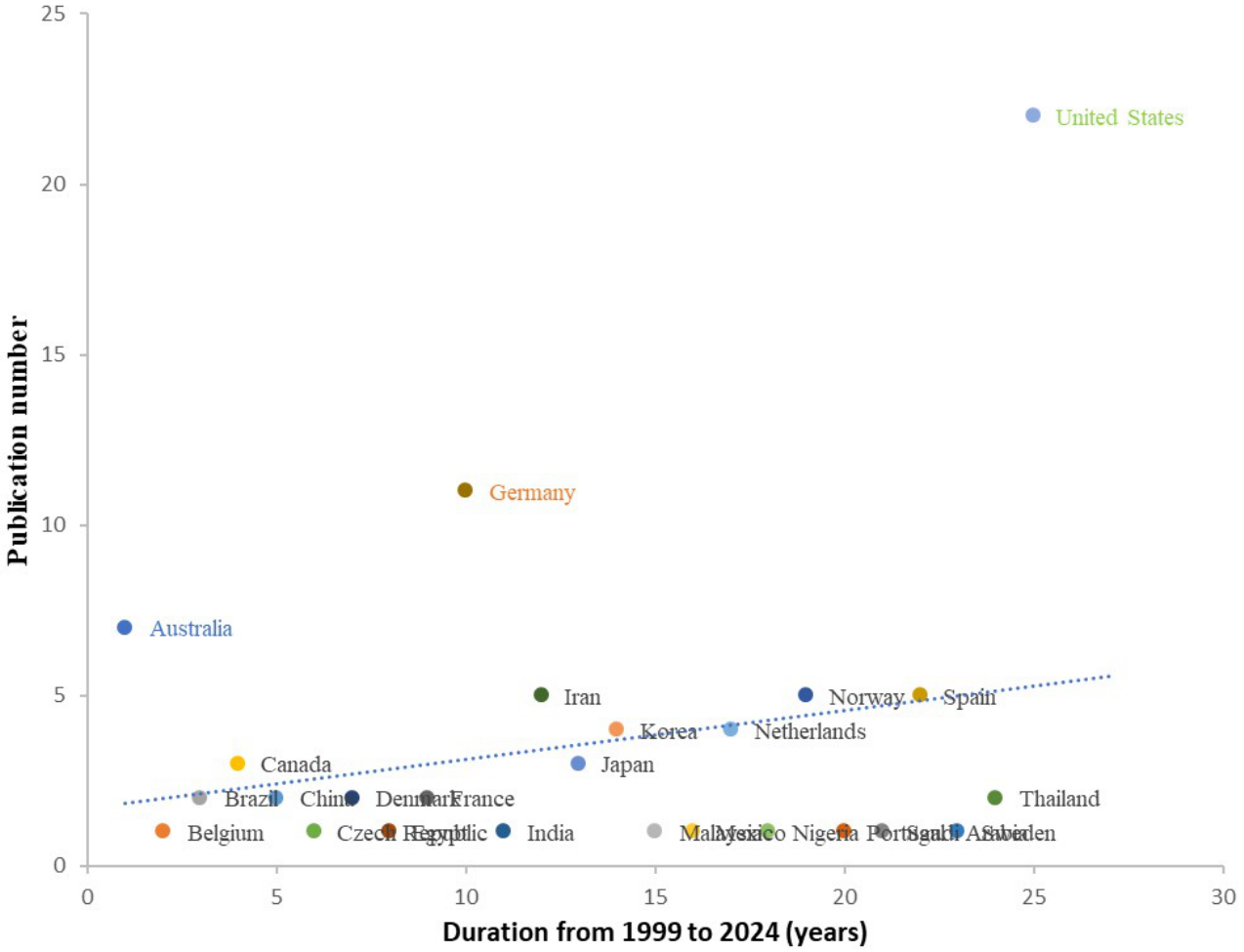

mHealth Users

The CMP populations considered in this review included adult and young patients with nonspecific CMP, specific OA without or with surgery, CLBP, and neck pain. Patients with OA and CLBP were the most commonly investigated groups among the included studies. The characteristics of the target user group were often the key impetus for the development of the mHealth tool. For example, some researchers noted that a certain percentage of patients report suboptimal satisfaction for various reasons after total knee arthroplasty, such as continued pain or stiffness [108-111]. For patients with knee pain 24 months post triple arthrodesis following a crush injury, a specific mobile camera could immediately provide corrective techniques when applying in-clinic equipment to the patient's shoe [102]. We identified 10 studies addressing the perceptions or experiences about mHealth tools in adult CMP populations with and without surgery [26,28,30,55,62,85,87,95,112] and their practitioners [30,35,87], as well as in young people [90]; the authors of these studies agreed that mHealth interventions were acceptable, useful, and feasible. However, there is still a gap between patients and physicians in understanding and communication about the treatment and management of knee OA [30]. It is expected that multiple stages of user experience testing could be a template for future mHealth tools aimed at chronic disease management [113]. However, only 1 eligible article [87] mentioned an mHealth app prototype that was codeveloped with patients and health care providers (HCPs) in 2021. Ultimately, it appears that diverse individuals can use mHealth tools as long as the tools are tailored to the needs of the population and sufficient training and support are provided.

Types of Tools Used in mHealth for CMP

In most of the reported mHealth interventions, mobile phones or other devices were either provided to users or considered a required intervention for study participation. We classified the studied mHealth tools into 4 main categories: mobile app, mobile app plus monitor, mobile app plus wearable sensor, and web-based mobile app plus monitor. The details of each type of mHealth tool are provided in Multimedia Appendix 6. Mobile app (78/89, 88%) was the most commonly used tool and the primary platform for patients with CMP. For example, mHealth apps could be installed on the patient's mobile phone at any time to help remember to check pain symptoms, maintain physical therapy, connect to coaches, or communicate virtually with HCPs in real time. The next most common mHealth tool reported was a mobile app plus monitor, which was used in 6% (5/89) of the included studies, followed by a mobile app plus wearable sensor (4/89, 4%) and web-based mobile app plus monitor (1/89, 1%). One study reported multiple digital technologies. These mHealth programs focused mainly on a combination of mobile and internal or external monitors or an external sensor, thus facilitating the transfer of data automatically without requiring the patient to manually submit

XSL•FO
RenderX

the data and providing faster delivery of self-management therapy.

Study Outcomes

Overview

Multiple outcome measures were used to evaluate mHealth tools depending on diverse study objectives. For the purposes of this analysis, the outcomes were organized into 3 categories: (1) usability, feasibility, and acceptability; (2) effect of the mHealth intervention on adherence to CMP management protocols; and (3) effect of the mHealth intervention on clinical outcomes.

Impact on Adherence

A total of 30 of the 89 studies (34%) evaluated or reported the effect of an mHealth intervention on adherence to CMP management, including medication adherence, engagement in healthy intervention therapy, and frequency of symptom monitoring. These studies include 2 questionnaire studies, 1 case series, 10 cohorts, 9 pilot studies, and 8 RCTs with or without a control group. Among these 30 studies, 1 (3%) observed mixed results, 25 (83%) showed a significant difference (judged at $P<.05$), and 1 (3%) found no difference. Multimedia Appendix 7 provides an overview of these studies [29,35,38,41,42,49,52-58,61,65,68,71,72,75,80,81,84,91,92,98-101,104,106].

Mobile apps with a monitor or wearable sensor were mainly investigated in patients with OA, which were used as daily reminders for self-management. Studies in patients with OA [80] and those undergoing surgery showed significant improvements of the apps with respect to patient adherence rates [38,49,65,100]. Some studies observed that a mobile app combined with a motion sensor or monitor could support early rehabilitation with good compliance [65], suggesting that surgeons can consider these two tools as appropriate alternatives to traditional physical therapy programs after joint surgery [49,100]. For younger patients with CMP, leveraging extant digital tools with appropriate user-informed adaptations can also help to build capacity tailored to support young people's self-management of musculoskeletal pain [90]. For other CMP groups, such as patients with CLBP and CNP, use of a mobile app alone was the main intervention. Although recognizing the inadequacy of traditional neck pain treatments compared with treating CLBP, a mobile app implemented with a self-classification algorithm was found to be particularly effective in increasing adherence to an exercise program among older and younger office workers with neck pain [58]. However, another study indicated that the clinical importance of added adherence with use of a mobile app is unclear in a specific population with upper- or lower-limb musculoskeletal conditions [56]. Often, using an mHealth system as an interface between the patient and the provider was perceived as less burdensome and associated with less judgment compared to face-to-face contact, particularly in situations in which the patients were not fully adherent to the recommended treatment [87,112]. mHealth tools facilitated better management and improved patient confidence to monitor CMP, making the patients feel in control and strengthening their coping mechanisms.

Impact on Clinical Outcomes

A total of 55 of the 89 included studies (62%) assessed or reported the effect of mHealth on disease-specific clinical outcomes, including pain intensity, pain-related function, pain-related disability, pain-related cognition, health-related QoL, and medication use. These studies include 1 case series, 13 cohort studies, 4 cross-sectional studies, 7 pilot studies, 1 questionnaire study, and 29 RCTs. Of the 29 RCTs that measured the effect of mHealth on CMP-specific clinical outcomes, 28 (93%) reported significant differences (judged at $P<.05$) between groups; no significant differences were found in 1 (3%) study and mixed results were observed in 1 (3%) study. In addition, a significant effect of mHealth tools was observed in 13 cohort studies and 7 pilot studies. Multimedia Appendix 8 provides an overview of these studies [20-23,25,27,29,34,37,39-42,44-48,50,51,53,54,56,58,59,61,63,64,67-69,72-74,76,78,79,81-83,86,88,89,94,96-101,103-107].

A total of 49 interventions (including RCTs, cohort studies, and pilot studies) were related to improving pain intensity outcomes. Among these 49 studies, 46 (94%) reported significant improvements in pain intensity. Both younger and older patients receiving app messages with tailored instructions on pain management experienced statistically significant improvements in their pain intensity levels compared to those of patients receiving usual care or an intervention without an mHealth tool. However, mobile app–based relaxation exercises did not effectively reduce CNP [68], highlighting the importance of future mHealth tools to include an individualized and tailored program. Another trial did not show a significant improvement in pain perception at 6 months, although the mHealth tool tested in this trial was determined to be feasible and associated with a satisfactory user experience [53]. Among the outcomes of the mHealth tools evaluated, 26 studies focused on pain-related functional performance, 15 studies focused on pain-related disability, 15 studies focused on pain-related cognitive performance, 14 studies focused on pain-related QoL, and 4 studies focused on pain-related medication use. Generally, mHealth tools were associated with a significant improvement in functional and cognitive performance and QoL, along with a significant decrease in the disability burden and medication use. However, in patients with CLBP, the improvement in pain-related disability was small and of uncertain clinical significance after using a self-management app for 9 months [82].

Usability, Feasibility, and Acceptability

Among the 89 included studies, 23 (26%) assessed or reported usability, feasibility, and acceptability using qualitative methods and compiled usage data, including 1 case series, 9 cohort studies, 4 cross-sectional studies, 6 pilot studies, 1 qualitative study, 1 questionnaire study, and 1 RCT. These data ranged from patient satisfaction to cost-effectiveness estimations as well as the timing and frequency of engagement with mobile apps and platforms. Multimedia Appendix 9 provides an

XSL•FO
RenderX

overview of these studies [22,24,29,33,38,43,45,49,52,57,61,66,71,72,74,77,80,92,93,95,101,107].

In general, these studies found mHealth tools and platforms to be usable, feasible, acceptable, and appreciated among users compared with traditional measures. For example, both older and younger patients with CMP and those who underwent joint surgery perceived that using an mHealth tool increased their independence and confidence in pain management [22,38,80]. One study reported that the prescription of therapeutic exercises via a smartphone app is feasible and well-accepted among patients of all ages [112]. Seven studies showed the long-term (>3-month follow-up) feasibility and acceptability of mHealth tools [24,33,38,49,52,72,101]. Specifically, with long-term follow-up, patients with knee or hip OA seemed to have preferences for goals related to physical activity and nutrition rather than for goals related to vitality and education [71]. Patients with OA undergoing primary hip or knee arthroplasty particularly appreciated the mHealth tools empowering patients, facilitating transitions from specialized hospital care to primary care, reducing unplanned contacts with the health system, and reducing overall health costs, proving to be cost-effective [38]. Another study performed in a real-world setting with a large (N=10,264) and diverse population experiencing CMP found that mHealth was accepted and considered especially useful for pain reduction [62]. The majority of studies included in this review focused on the patient as the end user of mHealth tools, although some also evaluated acceptability and perceptions from the perspective of HCPs. Features of mHealth tools such as automated reminders, messages with educational and motivational content, healthy living challenges, and wireless transmission of data contributed to increased self-care awareness and knowledge about CMP.

Discussion

Main Findings

This review showed that interventions based on mHealth systems have beneficial effects on adherence and clinical outcomes for individuals with CMP. Thus, this scientific evidence suggests that these mHealth systems could be promising alternatives for CMP self-management through multimodal approaches. The evidence presented here indicates that while the potential of mHealth tools is high, their results during implementation and execution are nevertheless mixed.

Mobile apps are the most widely reported mHealth tool for interventions, which have been successfully used to facilitate adherence to CMP management and improve clinical outcomes [114,115]. The freedom and portability of mobile devices combined with the advanced capacity to facilitate 2-way communication and collect and analyze data for a real-time response offer enormous potential to both patients and providers [116,117]. Owing to their abilities for automation, personalization, and easy integration into existing health systems, mobile apps are less operator-dependent and are less reliant on processes to facilitate the active and time-consuming exchange of information compared to traditional tools. However, apps may have a limitation in terms of difficulty of use for the older population with minimal technology experience or familiarity, and there is clearly room for improvement. Moreover, there is a lack of scientific and health professional support in many available mHealth systems, highlighting the need for developing appropriate apps based on the well-recognized guidelines in the management of CMP [113].

There is a growing recognition of the need for digital technologies to improve access to age-appropriate resources and personalized support for cocare [118-120]. Unlike previous reviews in this field that only included RCTs or cohort studies, this systematic review also included 13 qualitative studies (including questionnaire, interview, and discussion studies). These informative studies used qualitative methods that yielded rich data that can be used to better understand how and why mHealth tools impact adherence behaviors and clinical outcomes. Qualitative data can also enable patient-physician discussions regarding modifiable self-management options based on the perspectives and needs of patients, HCPs, or both groups. Moreover, user feedback can inform hypotheses that can then be tested. Research that seeks to understand how and why mHealth works will deliver on the broader promise of mHealth. Future mHealth tools will be able to draw on the knowledge generated when discrete hypotheses around the relative importance of, for example, patient-provider communication, optimal user interfaces, or targeted motivational messages are tested. These informative studies could lead to better mHealth tools that deliver improved health outcomes.

Implications and Future Directions

This review found that the usability, feasibility, and acceptability of mHealth tools for CMP management and adherence to different programs were generally high among both patients and providers. mHealth offers a way to address different barriers to care and to reduce health disparities from both patient and HCP perspectives. However, only one article published in this field over the last 25 years mentioned that the mHealth app prototype was codeveloped by patients and HCPs [87]. There will be more opportunities to codesign mHealth tools in the future. Undoubtedly, innovative mHealth tools could unintentionally increase health disparities due to unequal access to technology. There is also recognition that unequal access to, use of, and knowledge of information can influence the uptake and use of mHealth tools. These inequalities and the needs from target user groups should be taken into consideration early in the design and development of future mHealth tools. However, none of the studies included in this review addressed systematic differences in usability between diverse patient groups and geographical areas. Future research can be designed to better understand these differences and to encourage the development of mHealth tools that address the needs of diverse patient groups and populations living in regions with different levels of economic development.

The high prevalence of CMP globally coupled with the advantages of providing help through apps offers opportunities to help countless people who may be looking to the potential of mHealth to lessen the burden of their pain. One key aspect of this potential involves an increase in cost-effectiveness and expanded outreach of pain management. Of note, only one study included in the review specifically focused on the issue of

XSL•FO
RenderX

cost-effectiveness [60]. Furthermore, only one study presented the postmarketing observational data after a follow-up duration of 9 months [96]. Rigorous cost-effectiveness analyses will be necessary to demonstrate not only the health impact but also the value of investing in these innovations. Future studies evaluating the cost-effectiveness of mHealth tools are indeed needed, since we can see that pockets of mHealth innovations are expanding around the globe (Figure 3). Besides cost, language, and literacy barriers, availability and connectivity issues are also potential barriers to consider when developing these types of mHealth tools. Nevertheless, the strong attachment people have to mobile phones and the tendency to carry them everywhere open up opportunities for continuous symptom monitoring and connecting patients with providers outside of health care facilities.

Limitations and Strengths

There are some limitations of this study. First, we did not perform a meta-analysis and we did not weigh the quality of evidence or study design against reported results. Second, only the English literature was included, and the sample size of the included studies varied substantially from 1 in a case report to 10,264 in a cohort study. Third, the diversity of study designs, objectives, and outcome measures made clear comparisons among studies difficult, and the quality of evidence was deemed to be variable. However, this is the first 25-year systematic review focusing on evidence collected from January 1999 to December 2023 regarding the impact of mHealth on CMP management adherence. The main strengths of this review are that we included a diverse array of study designs; assessed both self-management and clinical outcomes; and incorporated the nascent literature regarding mHealth feasibility, usability, and acceptability.

Conclusions

mHealth is a potential high-impact tool to improve health outcomes among those with CMP through supporting adherence to personalized or tailored self-management programs for pain. Further evaluation of mHealth tools is needed, especially research that informs the cost-effectiveness of these tools. More innovation, optimization, and high-quality research in mHealth has the potential to transform the promise of mHealth technology into the reality of improved health care delivery and outcomes.

Acknowledgments

We would like to express our gratitude for the support from Hong Kong Jockey Club. Funding was provided by the Hong Kong Jockey Club Charities Trust.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 24 KB - jmir_v26i1e59358_app1.docx]

Multimedia Appendix 2
Inclusion criteria for study selection and mHealth tools.
[DOCX File , 24 KB - jmir_v26i1e59358_app2.docx]

Multimedia Appendix 3
Eligible studies.
[DOCX File , 55 KB - jmir_v26i1e59358_app3.docx]

Multimedia Appendix 4
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow for study selection.
[DOCX File , 58 KB - jmir_v26i1e59358_app4.docx]

Multimedia Appendix 5
Exluded studies with reasons.
[DOCX File , 42 KB - jmir_v26i1e59358_app5.docx]

Multimedia Appendix 6
Four types of mobile health tools.
[DOCX File , 15 KB - jmir_v26i1e59358_app6.docx]

Multimedia Appendix 7

XSL•FO
RenderX

Included studies investigating the impact of mobile health tools on adherence.
[XLSX File (Microsoft Excel File), 26 KB - jmir_v26i1e59358_app7.xlsx]

Multimedia Appendix 8
Includied studies investigating the impact of mobile health tools on clinical outcomes.
[XLSX File (Microsoft Excel File), 34 KB - jmir_v26i1e59358_app8.xlsx]

Multimedia Appendix 9
Included studies investigating the usability, feasibility, and acceptability of mobile health tools for patients with chronic musculoskeletal pain.
[XLSX File (Microsoft Excel File), 23 KB - jmir_v26i1e59358_app9.xlsx]

Multimedia Appendix 10
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 67 KB - jmir_v26i1e59358_app10.pdf]

References

1. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. A classification of chronic pain for ICD-11. Pain 2015 Jun;156(6):1003-1007 [FREE Full text] [doi: 10.1097/j.pain.0000000000000160] [Medline: 25844555]
2. Miaskowski C, Blyth F, Nicosia F, Haan M, Keefe F, Smith A, et al. A biopsychosocial model of chronic pain for older adults. Pain Med 2020 Oct 01;21(9):1793-1805. [doi: 10.1093/pm/pnz329] [Medline: 31846035]
3. Cheatle MD. Biopsychosocial approach to assessing and managing patients with chronic pain. Med Clin North Am 2016 Jan;100(1):43-53. [doi: 10.1016/j.mcna.2015.08.007] [Medline: 26614718]
4. Ampiah PK, Hendrick P, Moffatt F, Ahenkorah J. Operationalisation of a biopsychosocial approach for the non-pharmacological management of patients with chronic musculoskeletal pain in low- and middle-income countries: a systematic review. Musculoskeletal Care 2020 Oct;18(3):227-244. [doi: 10.1002/msc.1462] [Medline: 32056363]
5. Right to pain relief. International Association for the Study of Pain. 2004. URL: https://www.iasp-pain.org/GlobalYear/RighttoPainRelief [accessed 2024-07-31]
6. The burden of musculoskeletal diseases in the United States, third edition. United States Bone and Joint Initiative. 2014. URL: http://www.boneandjointburden.org [accessed 2024-07-31]
7. Fact Sheet: Aging in the United States. Population Reference Bureau. URL: http://www.prb.org/Publications/Media-Guides/2016/aging-unitedstates-fact-sheet.aspx [accessed 2024-07-31]
8. Forecast number of mobile users worldwide from 2020 to 2025. Statista. URL: https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/) [accessed 2024-07-31]
9. Devan H, Farmery D, Peebles L, Grainger R. Evaluation of self-management support functions in apps for people with persistent pain: systematic review. JMIR Mhealth Uhealth 2019 Mar 12;7(2):e13080 [FREE Full text] [doi: 10.2196/13080] [Medline: 30747715]
10. Pfeifer A, Uddin R, Schröder-Pfeifer P, Holl F, Swoboda W, Schiltenwolf M. Mobile application-based interventions for chronic pain patients: a systematic review and meta-analysis of effectiveness. J Clin Med 2020 Dec 05;9(11):3557 [FREE Full text] [doi: 10.3390/jcm9113557] [Medline: 33167300]
11. Du S, Liu W, Cai S, Hu Y, Dong J. The efficacy of e-health in the self-management of chronic low back pain: a meta analysis. Int J Nurs Stud 2020 Jul;106:103507. [doi: 10.1016/j.ijnurstu.2019.103507] [Medline: 32320936]
12. Thurnheer SE, Gravestock I, Pichierri G, Steurer J, Burgstaller JM. Benefits of mobile apps in pain management: systematic review. JMIR Mhealth Uhealth 2018 Oct 22;6(10):e11231 [FREE Full text] [doi: 10.2196/11231] [Medline: 30348633]
13. Viswanathan M, Golin CE, Jones CD, Ashok M, Blalock SJ, Wines RCM, et al. Interventions to improve adherence to self-administered medications for chronic diseases in the United States: a systematic review. Ann Intern Med 2012 Dec 04;157(11):785-795 [FREE Full text] [doi: 10.7326/0003-4819-157-11-201212040-00538] [Medline: 22964778]
14. Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc 2011 May;86(4):304-314 [FREE Full text] [doi: 10.4065/mcp.2010.0575] [Medline: 21389250]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
16. Foundation IDs for chronic pain diagnoses: Reference guide. IASP Classification for Chronic Pain in ICD11. URL: https://links.lww.com/PAIN/A658 [accessed 2024-03-16]
17. WHO Global Observatory for eHealth. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://iris.who.int/handle/10665/44607 [accessed 2024-03-16]
18. Norris SL, Glasgow RE, Engelgau MM, O'Connor PJ, McCulloch D. Chronic disease management. Dis Manage Health Outcomes 2003;11(8):477-488. [doi: 10.2165/00115677-200311080-00001]

XSL•FO
RenderX

19. ICD-11: International Classification of Diseases (11th revision). World Health Organization. URL: https://icd.who.int/ [accessed 2024-03-18]
20. Øverås CK, Nilsen TIL, Nicholl BI, Rughani G, Wood K, Søgaard K, et al. Multimorbidity and co-occurring musculoskeletal pain do not modify the effect of the SELFBACK app on low back pain-related disability. BMC Med 2022 Mar 08;20(1):53 [FREE Full text] [doi: 10.1186/s12916-022-02237-z] [Medline: 35130898]
21. Abadiyan F, Hadadnezhad M, Khosrokiani Z, Letafatkar A, Akhshik H. Adding a smartphone app to global postural re-education to improve neck pain, posture, quality of life, and endurance in people with nonspecific neck pain: a randomized controlled trial. Trials 2021 May 12;22(1):274 [FREE Full text] [doi: 10.1186/s13063-021-05214-8] [Medline: 33845880]
22. Ackerman IN, Bucknill A, Page RS, Broughton NS, Roberts C, Cavka B, et al. Preferences for disease-related education and support among younger people with hip or knee osteoarthritis. Arthritis Care Res 2017 May;69(4):499-508. [doi: 10.1002/acr.22950] [Medline: 27273912]
23. Alasfour M, Almarwani M. The effect of innovative smartphone application on adherence to a home-based exercise programs for female older adults with knee osteoarthritis in Saudi Arabia: a randomized controlled trial. Disabil Rehabil 2022 Jul 25;44(11):2420-2427. [doi: 10.1080/09638288.2020.1836268] [Medline: 33103499]
24. Amorim AB, Pappas E, Simic M, Ferreira ML, Jennings M, Tiedemann A, et al. Integrating mobile-health, health coaching, and physical activity to reduce the burden of chronic low back pain trial (IMPACT): a pilot randomised controlled trial. BMC Musculoskelet Disord 2019 Mar 11;20(1):71 [FREE Full text] [doi: 10.1186/s12891-019-2454-y] [Medline: 30744606]
25. Anan T, Kajiki S, Oka H, Fujii T, Kawamata K, Mori K, et al. Effects of an artificial intelligence-assisted health program on workers with neck/shoulder pain/stiffness and low back pain: randomized controlled trial. JMIR Mhealth Uhealth 2021 Oct 24;9(9):e27535 [FREE Full text] [doi: 10.2196/27535] [Medline: 34559054]
26. Arensman R, Kloek C, Pisters M, Koppenaal T, Ostelo R, Veenhof C. Patient perspectives on using a smartphone app to support home-based exercise during physical therapy treatment: qualitative study. JMIR Hum Factors 2022 Oct 13;9(3):e35316 [FREE Full text] [doi: 10.2196/35316] [Medline: 36098993]
27. Arfaei Chitkar SS, Mohaddes Hakkak HR, Saadati H, Hosseini SH, Jafari Y, Ganji R. The effect of mobile-app-based instruction on the physical function of female patients with knee osteoarthritis: a parallel randomized controlled trial. BMC Womens Health 2021 Oct 14;21(1):333 [FREE Full text] [doi: 10.1186/s12905-021-01451-w] [Medline: 34521400]
28. Östlind E, Ekvall Hansson E, Eek F, Stigmar K. Experiences of activity monitoring and perceptions of digital support among working individuals with hip and knee osteoarthritis - a focus group study. BMC Public Health 2022 Aug 30;22(1):1641 [FREE Full text] [doi: 10.1186/s12889-022-14065-0] [Medline: 36042425]
29. Bailey JF, Agarwal V, Zheng P, Smuck M, Fredericson M, Kennedy DJ, et al. Digital care for chronic musculoskeletal pain: 10,000 participant longitudinal cohort study. J Med Internet Res 2020 May 11;22(5):e18250 [FREE Full text] [doi: 10.2196/18250] [Medline: 32208358]
30. Barber T, Sharif B, Teare S, Miller J, Shewchuk B, Green LA, et al. Qualitative study to elicit patients' and primary care physicians' perspectives on the use of a self-management mobile health application for knee osteoarthritis. BMJ Open 2019 Mar 01;9(1):e024016 [FREE Full text] [doi: 10.1136/bmjopen-2018-024016] [Medline: 30782723]
31. Bechler U, Springer B, Rueckl K, Rolvien T, Boettner F. Can a simple iPad app improve C-arm based component position in anterior THA? Arch Orthop Trauma Surg 2021 Aug 13;141(8):1401-1409. [doi: 10.1007/s00402-021-03807-1] [Medline: 33582865]
32. Bellamy N, Patel B, Davis T, Dennison S. Electronic data capture using the Womac NRS 3.1 Index (m-Womac): a pilot study of repeated independent remote data capture in OA. Inflammopharmacology 2010 Jul 27;18(3):107-111. [doi: 10.1007/s10787-010-0040-x] [Medline: 20422296]
33. Bellamy N, Wilson C, Hendrikz J, Whitehouse SL, Patel B, Dennison S, et al. Osteoarthritis Index delivered by mobile phone (m-WOMAC) is valid, reliable, and responsive. J Clin Epidemiol 2011 Mar;64(2):182-190 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.03.013] [Medline: 20609562]
34. Beresford L, Norwood T. Can physical therapy deliver clinically meaningful improvements in pain and function through a mobile app? An observational retrospective study. Arch Rehabil Res Clin Transl 2022 Jul;4(2):100186 [FREE Full text] [doi: 10.1016/j.arrct.2022.100186] [Medline: 35756979]
35. Biebl JT, Huber S, Rykala M, Kraft E, Lorenz A. Attitudes and expectations of health care professionals toward app-based therapy in patients with osteoarthritis of the hip or knee: questionnaire study. JMIR Mhealth Uhealth 2020 Oct 28;8(10):e21704 [FREE Full text] [doi: 10.2196/21704] [Medline: 33112255]
36. Biebl JT, Rykala M, Strobel M, Kaur Bollinger P, Ulm B, Kraft E, et al. App-based feedback for rehabilitation exercise correction in patients with knee or hip osteoarthritis: prospective cohort study. J Med Internet Res 2021 Jul 13;23(7):e26658 [FREE Full text] [doi: 10.2196/26658] [Medline: 34255677]
37. Chhabra HS, Sharma S, Verma S. Smartphone app in self-management of chronic low back pain: a randomized controlled trial. Eur Spine J 2018 Dec 15;27(11):2862-2874. [doi: 10.1007/s00586-018-5788-5] [Medline: 30324496]
38. Colomina J, Drudis R, Torra M, Pallisó F, Massip M, Vargiu E, CONNECARE-Lleida Group. Implementing mHealth-enabled integrated care for complex chronic patients with osteoarthritis undergoing primary hip or knee arthroplasty: prospective, two-arm, parallel trial. J Med Internet Res 2021 Oct 02;23(9):e28320 [FREE Full text] [doi: 10.2196/28320] [Medline: 34473068]

XSL•FO
RenderX

39. Correia FD, Nogueira A, Magalhães I, Guimarães J, Moreira M, Barradas I, et al. Medium-term outcomes of digital versus conventional home-based rehabilitation after total knee arthroplasty: prospective, parallel-group feasibility study. JMIR Rehabil Assist Technol 2019 Mar 28;6(1):e13111 [FREE Full text] [doi: 10.2196/13111] [Medline: 30816849]
40. Crawford DA, Duwelius PJ, Sneller MA, Morris MJ, Hurst JM, Berend KR, et al. 2021 Mark Coventry Award: Use of a smartphone-based care platform after primary partial and total knee arthroplasty: a prospective randomized controlled trial. Bone Joint J 2021 Jul;103-B(6 Supple A):3-12. [doi: 10.1302/0301-620X.103B6.BJJ-2020-2352.R1] [Medline: 34053272]
41. Crawford DA, Lombardi AV, Berend KR, Huddleston JI, Peters CL, DeHaan A, et al. Early outcomes of primary total hip arthroplasty with use of a smartphone-based care platform: a prospective randomized controlled trial. Bone Joint J 2021 Jul;103-B(7 Supple B):91-97. [doi: 10.1302/0301-620X.103B7.BJJ-2020-2402.R1] [Medline: 34192907]
42. Dasa V, Skrepnik NV, Petersen D, Delanois RE. A novel mobile app-based neuromuscular electrical stimulation therapy for the management of knee osteoarthritis: results from an extension study of a randomized, double-blind, sham-controlled, multicenter trial. J Am Acad Orthop Surg Glob Res Rev 2022 Oct 01;6(9):e22.00115 [FREE Full text] [doi: 10.5435/JAAOSGlobal-D-22-00115] [Medline: 36094457]
43. de Brito Macedo L, Borges D, Melo S, da Costa K, de Oliveira Sousa C, Brasileiro J. Reliability and concurrent validity of a mobile application to measure thoracolumbar range of motion in low back pain patients. BMR 2020 Jan 13;33(1):145-151. [doi: 10.3233/bmr-181396]
44. Delgado AD, Salazar SI, Rozaieski K, Putrino D, Tabacof L. Engagement in an mHealth-guided exercise therapy program is associated with reductions in chronic musculoskeletal pain. Am J Phys Med Rehabil 2023 Dec 01;102(11):984-989. [doi: 10.1097/PHM.0000000000002257] [Medline: 37026894]
45. Guétin S, Diego ED, Mohy F, Adolphe C, Hoareau G, Touchon J, et al. A patient-controlled, smartphone-based music intervention to reduce pain—a multi-center observational study of patients with chronic pain. Eur J Integr Med 2016 Jun;8(3):182-187. [doi: 10.1016/j.eujim.2016.01.002]
46. Han J, Graham JH, Snyder DI, Alfieri T. Long-term use of wearable health technology by chronic pain patients. Clin J Pain 2022 Dec 01;38(12):701-710 [FREE Full text] [doi: 10.1097/AJP.0000000000001076] [Medline: 36198095]
47. Hardt S, Schulz MRG, Pfitzner T, Wassilew G, Horstmann H, Liodakis E, et al. Improved early outcome after TKA through an app-based active muscle training programme-a randomized-controlled trial. Knee Surg Sports Traumatol Arthrosc 2018 Dec 27;26(11):3429-3437. [doi: 10.1007/s00167-018-4918-2] [Medline: 29589050]
48. Hartmann R, Avermann F, Zalpour C, Griefahn A. Impact of an AI app-based exercise program for people with low back pain compared to standard care: a longitudinal cohort-study. Health Sci Rep 2023 Jan 12;6(1):e1060 [FREE Full text] [doi: 10.1002/hsr2.1060] [Medline: 36660258]
49. Hoogland J, Wijnen A, Munsterman T, Gerritsma CL, Dijkstra B, Zijlstra WP, et al. Feasibility and patient experience of a home-based rehabilitation program driven by a tablet app and mobility monitoring for patients after a total hip arthroplasty. JMIR Mhealth Uhealth 2019 Jan 31;7(1):e10342 [FREE Full text] [doi: 10.2196/10342] [Medline: 30702438]
50. Huber S, Priebe JA, Baumann K, Plidschun A, Schiessl C, Tölle TR. Treatment of low back pain with a digital multidisciplinary pain treatment app: short-term results. JMIR Rehabil Assist Technol 2017 Dec 04;4(2):e11 [FREE Full text] [doi: 10.2196/rehab.9032] [Medline: 29203460]
51. Itoh N, Mishima H, Yoshida Y, Yoshida M, Oka H, Matsudaira K. Evaluation of the effect of patient education and strengthening exercise therapy using a mobile messaging app on work productivity in Japanese patients with chronic low back pain: open-label, randomized, parallel-group trial. JMIR Mhealth Uhealth 2022 May 16;10(5):e35867 [FREE Full text] [doi: 10.2196/35867] [Medline: 35576560]
52. Kerckhove N, Delage N, Cambier S, Cantagrel N, Serra E, Marcaillou F, et al. eDOL mHealth app and web platform for self-monitoring and medical follow-up of patients with chronic pain: observational feasibility study. JMIR Form Res 2022 Mar 02;6(3):e30052 [FREE Full text] [doi: 10.2196/30052] [Medline: 35234654]
53. Kravitz RL, Schmid CH, Marois M, Wilsey B, Ward D, Hays RD, et al. Effect of mobile device-supported single-patient multi-crossover trials on treatment of chronic musculoskeletal pain: a randomized clinical trial. JAMA Intern Med 2018 Oct 01;178(10):1368-1377 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3981] [Medline: 30193253]
54. Krkoska P, Vlazna D, Sladeckova M, Minarikova J, Barusova T, Batalik L, et al. Adherence and effect of home-based rehabilitation with telemonitoring support in patients with chronic non-specific low back pain: a pilot study. Int J Environ Res Public Health 2023 Jan 13;20(2):1504 [FREE Full text] [doi: 10.3390/ijerph20021504] [Medline: 36674258]
55. Kurtz SM, Higgs GB, Chen Z, Koshut WJ, Tarazi JM, Sherman AE, et al. Patient perceptions of wearable and smartphone technologies for remote outcome monitoring in patients who have hip osteoarthritis or arthroplasties. J Arthroplasty 2022 Jul;37(7S):S488-S492. [doi: 10.1016/j.arth.2022.02.026] [Medline: 35277311]
56. Lambert TE, Harvey LA, Avdalis C, Chen LW, Jeyalingam S, Pratt CA, et al. An app with remote support achieves better adherence to home exercise programs than paper handouts in people with musculoskeletal conditions: a randomised trial. J Physiother 2017 Jul;63(3):161-167 [FREE Full text] [doi: 10.1016/j.jphys.2017.05.015] [Medline: 28662834]
57. Lebleu J, Pauwels A, Anract P, Parratte S, Van Overschelde P, Van Onsem S. Digital rehabilitation after knee arthroplasty: a multi-center prospective longitudinal cohort study. J Pers Med 2023 May 13;13(5):824 [FREE Full text] [doi: 10.3390/jpm13050824] [Medline: 37240994]

XSL•FO
RenderX

58. Lee M, Lee SH, Kim T, Yoo H, Kim SH, Suh D, et al. Feasibility of a smartphone-based exercise program for office workers with neck pain: an individualized approach using a self-classification algorithm. Arch Phys Med Rehabil 2017 Jan;98(1):80-87. [doi: 10.1016/j.apmr.2016.09.002] [Medline: 27693421]
59. Lee J, Lee M, Lim T, Kim T, Kim S, Suh D, et al. Effectiveness of an application-based neck exercise as a pain management tool for office workers with chronic neck pain and functional disability: a pilot randomized trial. Eur J Integr Med 2017 Jun;12:87-92. [doi: 10.1016/j.eujim.2017.04.012]
60. Lewkowicz D, Wohlbrandt AM, Bottinger E. Digital therapeutic care apps with decision-support interventions for people with low back pain in Germany: cost-effectiveness analysis. JMIR Mhealth Uhealth 2022 Mar 07;10(2):e35042 [FREE Full text] [doi: 10.2196/35042] [Medline: 35129454]
61. Lin W, Burke L, Schlenk EA, Yeh CH. Use of an ecological momentary assessment application to assess the effects of auricular point acupressure for chronic low back pain. Comput Inform Nurs 2019 May;37(5):276-282. [doi: 10.1097/CIN.0000000000000478] [Medline: 31094917]
62. Lo WLA, Lei D, Li L, Huang DF, Tong K. The perceived benefits of an artificial intelligence-embedded mobile app implementing evidence-based guidelines for the self-management of chronic neck and back pain: observational study. JMIR Mhealth Uhealth 2018 Dec 26;6(11):e198 [FREE Full text] [doi: 10.2196/mhealth.8127] [Medline: 30478019]
63. Marcuzzi A, Nordstoga AL, Bach K, Aasdahl L, Nilsen TIL, Bardal EM, et al. Effect of an artificial intelligence-based self-management app on musculoskeletal health in patients with neck and/or low back pain referred to specialist care: a randomized clinical trial. JAMA Netw Open 2023 Jul 01;6(6):e2320400 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.20400] [Medline: 37368401]
64. Mbada CE, Olaoye MI, Dada OO, Ayanniyi O, Johnson OE, Odole AC, et al. Comparative efficacy of clinic-based and telerehabilitation application of Mckenzie therapy in chronic low-back pain. Int J Telerehabil 2019 Jun 12;11(1):41-58 [FREE Full text] [doi: 10.5195/ijt.2019.6260] [Medline: 31341546]
65. Neumann-Langen MV, Ochs BG, Lützner J, Postler A, Kirschberg J, Sehat K, et al. Musculoskeletal rehabilitation: new perspectives in postoperative care following total knee arthroplasty using an external motion sensor and a smartphone application for remote monitoring. J Clin Med 2023 Dec 18;12(22):7163 [FREE Full text] [doi: 10.3390/jcm12227163] [Medline: 38002775]
66. Nordstoga AL, Bach K, Sani S, Wiratunga N, Mork PJ, Villumsen M, et al. Usability and acceptability of an app (SELFBACK) to support self-management of low back pain: mixed methods study. JMIR Rehabil Assist Technol 2020 Oct 09;7(2):e18729 [FREE Full text] [doi: 10.2196/18729] [Medline: 32902393]
67. Osama Al Saadawy B, Abdo N, Embaby E, Rehan Youssef A. Validity and reliability of smartphones in measuring joint position sense among asymptomatic individuals and patients with knee osteoarthritis: a cross-sectional study. Knee 2021 Mar;29:313-322. [doi: 10.1016/j.knee.2021.02.012] [Medline: 33677156]
68. Pach D, Blödt S, Wang J, Keller T, Bergmann B, Rogge AA, et al. App-based relaxation exercises for patients with chronic neck pain: pragmatic randomized trial. JMIR Mhealth Uhealth 2022 Jan 07;10(1):e31482 [FREE Full text] [doi: 10.2196/31482] [Medline: 34994708]
69. Park C, Yi C, Choi WJ, Lim H, Yoon HU, You SH. Long-term effects of deep-learning digital therapeutics on pain, movement control, and preliminary cost-effectiveness in low back pain: a randomized controlled trial. Digit Health 2023 Dec 03;9:20552076231217817 [FREE Full text] [doi: 10.1177/20552076231217817] [Medline: 38053732]
70. Pelle T, Bevers K, van der Palen J, van den Hoogen F, van den Ende C. Effect of the dr. Bart application on healthcare use and clinical outcomes in people with osteoarthritis of the knee and/or hip in the Netherlands; a randomized controlled trial. Osteoarthritis Cartilage 2020 May;28(4):418-427 [FREE Full text] [doi: 10.1016/j.joca.2020.02.831] [Medline: 32119972]
71. Pelle T, van der Palen J, de Graaf F, van den Hoogen FHJ, Bevers K, van den Ende CHM. Use and usability of the dr. Bart app and its relation with health care utilisation and clinical outcomes in people with knee and/or hip osteoarthritis. BMC Health Serv Res 2021 May 10;21(1):444 [FREE Full text] [doi: 10.1186/s12913-021-06440-1] [Medline: 33971861]
72. Peterson S. Telerehabilitation booster sessions and remote patient monitoring in the management of chronic low back pain: a case series. Physiother Theory Pract 2018 May;34(5):393-402. [doi: 10.1080/09593985.2017.1401190] [Medline: 29125371]
73. Pourahmadi MR, Bagheri R, Taghipour M, Takamjani IE, Sarrafzadeh J, Mohseni-Bandpei MA. A new iPhone application for measuring active craniocervical range of motion in patients with non-specific neck pain: a reliability and validity study. Spine J 2018 Mar;18(3):447-457. [doi: 10.1016/j.spinee.2017.08.229] [Medline: 28890223]
74. Pourahmadi M, Momeni E, Mohseni N, Hesarikia H, Ghanjal A, Shamsoddini A. The reliability and concurrent validity of a new iPhone® application for measuring active lumbar spine flexion and extension range of motion in patients with low back pain. Physiother Theory Pract 2021 Jan 13;37(1):204-217. [doi: 10.1080/09593985.2019.1616017] [Medline: 31081417]
75. Rabbi M, Aung MS, Gay G, Reid MC, Choudhury T. Feasibility and acceptability of mobile phone-based auto-personalized physical activity recommendations for chronic pain self-management: pilot study on adults. J Med Internet Res 2018 Oct 26;20(10):e10147 [FREE Full text] [doi: 10.2196/10147] [Medline: 30368433]

XSL•FO
RenderX

76. Rafiq MT, Abdul Hamid MS, Hafiz E. The effect of rehabilitation protocol using mobile health in overweight and obese patients with knee osteoarthritis: a clinical trial. Adv Rheumatol 2021 Oct 24;61(1):63. [doi: 10.1186/s42358-021-00221-4] [Medline: 34689837]
77. Ramkumar PN, Haeberle HS, Ramanathan D, Cantrell WA, Navarro SM, Mont MA, et al. Remote patient monitoring using mobile health for total knee arthroplasty: validation of a wearable and machine learning-based surveillance platform. J Arthroplasty 2019 Oct;34(10):2253-2259. [doi: 10.1016/j.arth.2019.05.021] [Medline: 31128890]
78. Ramos MM, Carnaz L, Mattiello SM, Karduna AR, Zanca GG. Shoulder and elbow joint position sense assessment using a mobile app in subjects with and without shoulder pain - between-days reliability. Phys Ther Sport 2019 May;37:157-163. [doi: 10.1016/j.ptsp.2019.03.016] [Medline: 30978602]
79. Rodríguez Sánchez-Laulhé P, Biscarri-Carbonero Á, Suero-Pineda A, Luque-Romero LG, Barrero García FJ, Blanquero J, et al. The effects of a mobile app-delivered intervention in people with symptomatic hand osteoarthritis: a pragmatic randomized controlled trial. Eur J Phys Rehabil Med 2023 Mar;59(1):54-64 [FREE Full text] [doi: 10.23736/S1973-9087.22.07744-9] [Medline: 36633498]
80. Rouzaud Laborde C, Cenko E, Mardini MT, Nerella S, Kheirkhahan M, Ranka S, et al. Satisfaction, usability, and compliance with the use of smartwatches for ecological momentary assessment of knee osteoarthritis symptoms in older adults: usability study. JMIR Aging 2021 Jul 14;4(3):e24553 [FREE Full text] [doi: 10.2196/24553] [Medline: 34259638]
81. Rughani G, Nilsen TIL, Wood K, Mair FS, Hartvigsen J, Mork PJ, et al. The selfBACK artificial intelligence-based smartphone app can improve low back pain outcome even in patients with high levels of depression or stress. Eur J Pain 2023 May 27;27(5):568-579. [doi: 10.1002/ejp.2080] [Medline: 36680381]
82. Sandal LF, Bach K, Øverås CK, Svendsen MJ, Dalager T, Stejnicher Drongstrup Jensen J, et al. Effectiveness of app-delivered, tailored self-management support for adults with lower back pain-related disability: a selfBACK randomized clinical trial. JAMA Intern Med 2021 Oct 01;181(10):1288-1296 [FREE Full text] [doi: 10.1001/jamainternmed.2021.4097] [Medline: 34338710]
83. Sax OC, Gesheff MG, Mahajan A, Patel N, Andrews T, Jreisat A, et al. A novel mobile app-based neuromuscular electrical stimulation therapy for improvement of knee pain, stiffness, and function in knee osteoarthritis: a randomized trial. Arthroplast Today 2022 Jul;15:125-131 [FREE Full text] [doi: 10.1016/j.artd.2022.03.007] [Medline: 35514364]
84. Selter A, Tsangouri C, Ali SB, Freed D, Vatchinsky A, Kizer J, et al. An mHealth app for self-management of chronic lower back pain (Limbr): pilot study. JMIR Mhealth Uhealth 2018 Oct 17;6(9):e179 [FREE Full text] [doi: 10.2196/mhealth.8256] [Medline: 30224333]
85. Shah N, Borrelli B, Kumar D. Perceptions about smartphone-based interventions to promote physical activity in inactive adults with knee pain - a qualitative study. Disabil Rehabil Assist Technol 2023 Oct 24:1-8. [doi: 10.1080/17483107.2023.2272854] [Medline: 37873670]
86. Shebib R, Bailey JF, Smittenaar P, Perez DA, Mecklenburg G, Hunter S. Randomized controlled trial of a 12-week digital care program in improving low back pain. NPJ Digit Med 2019 Jan 07;2(1):1. [doi: 10.1038/s41746-018-0076-7] [Medline: 31304351]
87. Shewchuk B, Green LA, Barber T, Miller J, Teare S, Campbell-Scherer D, et al. Patients' use of mobile health for self-management of knee osteoarthritis: results of a 6-week pilot study. JMIR Form Res 2021 Dec 25;5(11):e30495 [FREE Full text] [doi: 10.2196/30495] [Medline: 34842526]
88. Sitges C, Terrasa JL, García-Dopico N, Segur-Ferrer J, Velasco-Roldán O, Crespí-Palmer J, et al. An educational and exercise mobile phone-based intervention to elicit electrophysiological changes and to improve psychological functioning in adults with nonspecific chronic low back pain (BackFit app): nonrandomized clinical trial. JMIR Mhealth Uhealth 2022 Mar 15;10(3):e29171 [FREE Full text] [doi: 10.2196/29171] [Medline: 35289758]
89. Skrepnik N, Spitzer A, Altman R, Hoekstra J, Stewart J, Toselli R. Assessing the impact of a novel smartphone application compared with standard follow-up on mobility of patients with knee osteoarthritis following treatment with Hylan G-F 20: a randomized controlled trial. JMIR Mhealth Uhealth 2017 May 09;5(5):e64 [FREE Full text] [doi: 10.2196/mhealth.7179] [Medline: 28487266]
90. Slater H, Jordan JE, Chua J, Schütze R, Wark JD, Briggs AM. Young people's experiences of persistent musculoskeletal pain, needs, gaps and perceptions about the role of digital technologies to support their co-care: a qualitative study. BMJ Open 2016 Dec 09;6(12):e014007 [FREE Full text] [doi: 10.1136/bmjopen-2016-014007] [Medline: 27940635]
91. Slater H, Stinson JN, Jordan JE, Chua J, Low B, Lalloo C, et al. Evaluation of digital technologies tailored to support young people's self-management of musculoskeletal pain: mixed methods study. J Med Internet Res 2020 Jul 05;22(6):e18315 [FREE Full text] [doi: 10.2196/18315] [Medline: 32442143]
92. Støme LN, Pripp AH, Kværner JS, Kvaerner KJ. Acceptability, usability and utility of a personalised application in promoting behavioural change in patients with osteoarthritis: a feasibility study in Norway. BMJ Open 2019 Jan 28;9(1):e021608 [FREE Full text] [doi: 10.1136/bmjopen-2018-021608] [Medline: 30696666]
93. Suso-Ribera C, Castilla D, Zaragozá I, Ribera-Canudas MV, Botella C, García-Palacios A. Validity, reliability, feasibility, and usefulness of pain monitor: a multidimensional smartphone app for daily monitoring of adults with heterogenous chronic pain. Clin J Pain 2018 Oct;34(10):900-908. [doi: 10.1097/AJP.0000000000000618] [Medline: 29659375]

XSL•FO
RenderX

94. Suso-Ribera C, Castilla D, Zaragozá I, Mesas Á, Server A, Medel J, et al. Telemonitoring in chronic pain management using smartphone apps: a randomized controlled trial comparing usual assessment against app-based monitoring with and without clinical alarms. Int J Environ Res Public Health 2020 Oct 09;17(18):6568 [FREE Full text] [doi: 10.3390/ijerph17186568] [Medline: 32916983]
95. Svendsen MJ, Nicholl BI, Mair FS, Wood K, Rasmussen CDN, Stochkendahl MJ. One size does not fit all: Participants' experiences of the selfBACK app to support self-management of low back pain-a qualitative interview study. Chiropr Man Therap 2022 Oct 03;30(1):41 [FREE Full text] [doi: 10.1186/s12998-022-00452-2] [Medline: 36192724]
96. Teepe GW, Kowatsch T, Hans FP, Benning L. Postmarketing follow-up of a digital home exercise program for back, hip, and knee pain: retrospective observational study with a time-series and matched-pair analysis. J Med Internet Res 2023 Mar 27;25:e43775 [FREE Full text] [doi: 10.2196/43775] [Medline: 36848211]
97. Thiengwittayaporn S, Wattanapreechanon P, Sakon P, Peethong A, Ratisoontorn N, Charoenphandhu N, et al. Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial. Arch Orthop Trauma Surg 2023 Mar 28;143(2):729-738 [FREE Full text] [doi: 10.1007/s00402-021-04149-8] [Medline: 34453570]
98. Thongtipmak S, Buranruk O, Eungpinichpong W, Konharn K. Immediate effects and acceptability of an application-based stretching exercise incorporating deep slow breathing for neck pain self-management. Healthc Inform Res 2020 Jan;26(1):50-60 [FREE Full text] [doi: 10.4258/hir.2020.26.1.50] [Medline: 32082700]
99. Toelle TR, Utpadel-Fischler DA, Haas K, Priebe JA. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. NPJ Digit Med 2019 May 03;2(1):34. [doi: 10.1038/s41746-019-0109-x] [Medline: 31304380]
100. Tripuraneni KR, Foran JRH, Munson NR, Racca NE, Carothers JT. A smartwatch paired with a mobile application provides postoperative self-directed rehabilitation without compromising total knee arthroplasty outcomes: a randomized controlled trial. J Arthroplasty 2021 Dec;36(12):3888-3893. [doi: 10.1016/j.arth.2021.08.007] [Medline: 34462184]
101. Vad VB, Madrazo-Ibarra A, Estrin D, Pollak JP, Carroll KM, Vojta D, et al. "Back Rx, a personalized mobile phone application for discogenic chronic low back pain: a prospective pilot study". BMC Musculoskelet Disord 2022 Oct 19;23(1):923 [FREE Full text] [doi: 10.1186/s12891-022-05883-9] [Medline: 36261825]
102. VanWye WR, Hoover DL. Management of a patient's gait abnormality using smartphone technology in-clinic for improved qualitative analysis: a case report. Physiother Theory Pract 2018 May;34(5):403-410. [doi: 10.1080/09593985.2017.1419326] [Medline: 29308956]
103. Weise H, Zenner B, Schmiedchen B, Benning L, Bulitta M, Schmitz D, et al. The effect of an app-based home exercise program on self-reported pain intensity in unspecific and degenerative back pain: pragmatic open-label randomized controlled trial. J Med Internet Res 2022 Oct 28;24(10):e41899 [FREE Full text] [doi: 10.2196/41899] [Medline: 36215327]
104. Yamamoto Y, Murata Y, Tanaka N, Shigemura T, Maruyama J, Nakane R, et al. Mobile application for home exercise adherence in patients with knee osteoarthritis: a pilot study. Medicine 2022 Oct 21;101(42):e31181 [FREE Full text] [doi: 10.1097/MD.0000000000031181] [Medline: 36281120]
105. Yang J, Wei Q, Ge Y, Meng L, Zhao M. Smartphone-based remote self-management of chronic low back pain: a preliminary study. J Healthc Eng 2019 Feb 06;2019:4632946. [doi: 10.1155/2019/4632946] [Medline: 30881606]
106. Yeh CH, Kawi J, Grant L, Huang X, Wu H, Hardwicke RL, et al. Self-guided smartphone application to manage chronic musculoskeletal pain: a randomized, controlled pilot trial. Int J Environ Res Public Health 2022 Dec 11;19(22):14875 [FREE Full text] [doi: 10.3390/ijerph192214875] [Medline: 36429591]
107. Zhuo LX, Macedo LG. Feasibility and convergent validity of an activity tracker for low back pain within a clinical study: cross-sectional study. JMIR Rehabil Assist Technol 2021 Mar 26;8(1):e18942 [FREE Full text] [doi: 10.2196/18942] [Medline: 33769301]
108. Thompson R, Novikov D, Cizmic Z, Feng JE, Fideler K, Sayeed Z, et al. Arthrofibrosis after total knee arthroplasty: pathophysiology, diagnosis, and management. Orthop Clin North Am 2019 Jul;50(3):269-279. [doi: 10.1016/j.ocl.2019.02.005] [Medline: 31084828]
109. Yoon RS, Nellans KW, Geller JA, Kim AD, Jacobs MR, Macaulay W. Patient education before hip or knee arthroplasty lowers length of stay. J Arthroplasty 2010 Jul;25(4):547-551. [doi: 10.1016/j.arth.2009.03.012] [Medline: 19427164]
110. Garrison SR, Schneider KE, Singh M, Pogodzinski J. Preoperative physical therapy results in shorter length of stay and discharge disposition following total knee arthroplasty: a retrospective study. J Rehabil Med Clin Commun 2019;2:1000017. [doi: 10.2340/20030711-1000017] [Medline: 33884118]
111. Moyer R, Ikert K, Long K, Marsh J. The value of preoperative exercise and education for patients undergoing total hip and knee arthroplasty: a systematic review and meta-analysis. JBJS Rev 2017 Dec;5(12):e2. [doi: 10.2106/JBJS.RVW.17.00015] [Medline: 29232265]
112. Hasenöhrl T, Windschnurer T, Dorotka R, Ambrozy C, Crevenna R. Prescription of individual therapeutic exercises via smartphone app for patients suffering from non-specific back pain: a qualitative feasibility and quantitative pilot study. Wien Klin Wochenschr 2020 Mar;132(5-6):115-123 [FREE Full text] [doi: 10.1007/s00508-020-01616-x] [Medline: 32060724]

XSL•FO
RenderX

113. Osborn CY, Mulvaney SA. Development and feasibility of a text messaging and interactive voice response intervention for low-income, diverse adults with type 2 diabetes mellitus. J Diabetes Sci Technol 2013 May 01;7(3):612-622 [FREE Full text] [doi: 10.1177/193229681300700305] [Medline: 23759393]
114. Sadler S, Gerrard J, Searle A, Lanting S, West M, Wilson R, et al. The use of mHealth apps for the assessment and management of diabetes-related foot health outcomes: systematic review. J Med Internet Res 2023 Oct 04;25:e47608 [FREE Full text] [doi: 10.2196/47608] [Medline: 37792467]
115. Liu F, Song T, Yu P, Deng N, Guan Y, Yang Y, et al. Efficacy of an mHealth app to support patients' self-management of hypertension: randomized controlled trial. J Med Internet Res 2023 Dec 19;25:e43809 [FREE Full text] [doi: 10.2196/43809] [Medline: 38113071]
116. Chen L, Zhang D, Li T, Liu S, Hua J, Cai W. Effect of a mobile app-based urinary incontinence self-management intervention among pregnant women in China: pragmatic randomized controlled trial. J Med Internet Res 2023 Jul 27;25:e43528 [FREE Full text] [doi: 10.2196/43528] [Medline: 37368465]
117. Nagino K, Okumura Y, Akasaki Y, Fujio K, Huang T, Sung J, et al. Smartphone app-based and paper-based patient-reported outcomes using a disease-specific questionnaire for dry eye disease: randomized crossover equivalence study. J Med Internet Res 2023 Aug 03;25:e42638 [FREE Full text] [doi: 10.2196/42638] [Medline: 37535409]
118. Schliemann D, Tan MM, Hoe WMK, Mohan D, Taib NA, Donnelly M, et al. mHealth interventions to improve cancer screening and early detection: scoping review of reviews. J Med Internet Res 2022 Aug 15;24(8):e36316 [FREE Full text] [doi: 10.2196/36316] [Medline: 35969450]
119. Yang Y, Boulton E, Todd C. Measurement of adherence to mHealth physical activity interventions and exploration of the factors that affect the adherence: scoping review and proposed framework. J Med Internet Res 2022 Jul 08;24(6):e30817 [FREE Full text] [doi: 10.2196/30817] [Medline: 35675111]
120. Lu Q, Schulz PJ. Physician perspectives on internet-informed patients: systematic review. J Med Internet Res 2024 Jul 06;26:e47620 [FREE Full text] [doi: 10.2196/47620] [Medline: 38842920]

Abbreviations

CLBP: chronic low back pain
CMP: chronic musculoskeletal pain
CNP: chronic neck pain
HCP: health care provider
ICD-11: International Classification of Diseases, 11th edition
MeSH: Medical Subject Headings
mHealth: mobile health
OA: osteoarthritis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
QoL: quality of life
RCT: randomized controlled trial

Edited by G Eysenbach; submitted 10.04.24; peer-reviewed by E Ross, N Kerckhove; comments to author 04.06.24; revised version received 18.06.24; accepted 16.07.24; published 16.08.24.

Please cite as:
Shi JLH, Sit RWS
Impact of 25 Years of Mobile Health Tools for Pain Management in Patients With Chronic Musculoskeletal Pain: Systematic Review
J Med Internet Res 2024;26:e59358
URL: https://www.jmir.org/2024/1/e59358
doi:10.2196/59358
PMID:

©Jenny Lin-Hong Shi, Regina Wing-Shan Sit. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Unintended Consequences of Telehealth in Australia: Critical Interpretive Synthesis

Sagda Osman[1], MSc; Kate Churruca[1], PhD; Louise A Ellis[1], PhD; Dan Luo[1,2], PhD; Jeffrey Braithwaite[1], PhD

[1]Centre for Healthcare Resilience and Implementation Science, Australian Institute of Health Innovation, Macquarie University, North Ryde, Australia
[2]The Daffodil Centre, Sydney, Australia

Corresponding Author:
Sagda Osman, MSc
Centre for Healthcare Resilience and Implementation Science
Australian Institute of Health Innovation
Macquarie University
75 Talavera Rd
North Ryde, 2113
Australia
Phone: 61 (02) 9850 2400
Email: sagda.osman@mq.edu.au

Abstract

Background: Despite more than 2 decades of telehealth use in Australia and the rapid uptake during the COVID-19 pandemic, little is known about its unintended consequences beyond its planned and intended outcomes.

Objective: The aim of this review was to synthesize evidence on the unintended consequences of telehealth use in Australia to clarify its impact beyond its planned and intended outcomes.

Methods: We conducted a search of 4 electronic databases: Ovid MEDLINE, Ovid Embase, EBSCO CINAHL, and Scopus. A critical interpretive synthesis approach was adopted for its flexibility and interpretive nature. We extracted data about study characteristics and the types and models of telehealth services. The extracted unintended consequences were coded and mapped into the domains and dimensions of the Australian Health Performance Framework.

Results: Of the 4241 records identified by the search, 94 (2.22%) studies were eligible for data extraction and analysis. Of these 94 studies, 23 (24%) reported largely positive unintended consequences of telehealth associated with health status, while 6 (6%) noted a potential negative impact of telehealth on socioeconomic status. The findings of 4 (4%) of the 94 studies highlighted societal and financial consequences of telehealth beyond the health system. Almost all studies (93/94, 99%) reported unintended consequences under the 5 dimensions of the Australian Health Performance Framework.

Conclusions: Our synthesis offers a framework for understanding the unintended consequences of the use of telehealth as an alternative to in-person care in Australia. While we have documented many unintended benefits of telehealth use, our findings also shed light on many challenges of delivering care via telehealth across different domains and dimensions. These findings hold significant practice and policy-making implications for ensuring safe and high-quality care delivery via telehealth.

(J Med Internet Res 2024;26:e57848) doi:10.2196/57848

KEYWORDS

telehealth; telemedicine; unintended consequences; digital health; eHealth; critical interpretive synthesis; review methodology; literature review; Australia

Introduction

Background

The COVID-19 pandemic has resulted in a rapid increase in telehealth use globally. During the pandemic, telehealth emerged as a promising model of care to improve health care delivery and increase access for underserved communities. In Australia, during the first year of the pandemic in 2020, telehealth use grew almost exponentially because many health services rapidly moved to provide their services online [1-4]. Throughout the pandemic and beyond, a high consumer preference has been reported for telehealth as an alternative to standard care when it is clinically appropriate and safe, with several reported benefits such as flexibility, convenience, and time and cost

XSL•FO
RenderX

savings [5-9]. This pattern of a dramatic increase in the uptake of telehealth during the COVID-19 pandemic and, subsequently, its sustained use over time is similar across many other countries, such as the United States, the United Kingdom, and New Zealand [10,11].

Telehealth is broadly defined as using information and communication technologies to provide health care at a distance [12]. This definition includes models where a physician-patient interaction, whether synchronous or asynchronous, is not necessarily required, such as in patient monitoring, online support groups, or consumer mobile apps [12]. However, in this review, we focus primarily on synchronous telehealth or synchronous telehealth in conjunction with asynchronous models that act as an alternative to standard in-person care; this includes one-to-one patient-clinician interactions and models where clinician-clinician interactions appertain to case management.

Unintended consequences are the unplanned and often unforeseen outcomes of introducing an intervention, such as telehealth [13,14]. These outcomes can be (1) positive outcomes (accidental benefits), (2) adverse outcomes (drawbacks), or (3) neutral outcomes with no intrinsic positive or negative value [13,15]. To provide safe and effective care via telehealth, an understanding of its unintended or unexpected consequences is warranted to mitigate drawbacks, optimize unexpected benefits, and understand system-wide and long-term impact. There has thus far been limited research investigating this issue, with only 1 secondary analysis carried out in Canada [14]. Guided by the model proposed by Bloomrosen et al [16] for investigating the unintended consequences of health information technologies, the study by Alami et al [14] documented findings on the unintended consequences of telehealth across various domains. However, the Canadian data sources were primarily evaluation documents that did not account for published empirical research beyond the 10 projects investigated. Furthermore, despite defining unintended consequences as positive, neutral, or negative unforeseen outcomes, the reported consequences were primarily negative, such as rigidity of technology to accommodate existing workflows and processes, cognitive overload, invasion of patients' privacy, increase in administrative workload, and the additional costs associated with telehealth. Few neutral consequences were documented in the study, such as the restructuring of hierarchical relationships and the emergence of new modes of clinical practice [14].

Objectives

A comprehensive synthesis of primary and secondary empirical studies on this subject is thus still lacking. In this review, we aimed to close this knowledge gap by synthesizing findings on the unintended consequences of telehealth to fully understand the impact of telehealth, focusing not only on the microlevel (individual users) and the mesolevel (organizational processes and workflows) but also on a policy macrolevel or national level. The objective of the review was to investigate the current state of knowledge on the multilevel unintended consequences of telehealth implementation in the Australian health system. We used a critical interpretive synthesis (CIS) [17] approach due to the diversity of telehealth literature and the lack of consensus on terms and concepts. This review is part of a larger research project using systems sciences as a conceptual lens to understand telehealth's unintended consequences [13]. This conceptual framework informed the design and analysis of this review, emphasizing the rippling effects of telehealth implementation across the complex health care system.

Research Questions

The following research questions (RQs) were formulated to achieve the review objective:

- RQ1: How is telehealth implementation being evaluated or assessed?
- RQ2: What are the findings on the types and models of telehealth services implemented?
- RQ3: What are the findings on the unintended consequences of telehealth implementation?

While the first 2 RQs were used to extract study characteristics, the third RQ served as an initial guide for the CIS. In line with the principles of this methodology, our inquiry evolved and was refined throughout the review process, as detailed in the Methods section.

We focused our review on studies conducted within Australia to provide a picture of unintended consequences within a single, contained health care system. Health systems vary across different countries in terms of structure, funding models, and policies, but Australia represents a high-income country with a well-resourced hybrid health system that deals with fragmentation and challenges similar to those in comparable international health systems [18], making our findings relevant to such systems.

Methods

The CIS Method

A CIS offers an appropriate alternative to a standard systematic review. Its exploratory nature allows for flexibility in including diverse study types [17], including studies that do not focus directly on the unintended consequences of telehealth or ostensibly on the topic of unintended consequences but could still inform the analysis. Moreover, CIS provided us with the interpretive freedom necessary to build a framework and draw conclusions from studies that reported on the unintended consequences of telehealth without necessarily using the term "unintended consequences." This interpretive flexibility is necessary when building a knowledge base and synthesizing evidence where the literature is diverse and complex [17].

Search Strategy

A search strategy was developed in consultation with a clinical librarian to search the databases Ovid MEDLINE, Embase, EBSCO CINAHL, and Scopus using keywords and Medical Subject Headings terms related to "telehealth," its "implementation," and "unintended consequences" in "Australia" (Multimedia Appendix 1). We did not apply date limits to the search to ensure a comprehensive account of the available evidence. The search was conducted in November 2022.

XSL•FO
RenderX

Eligibility Criteria

Qualitative and mixed methods allow for the exploration of issues and variables that might be unexpected and therefore remain unmeasured in quantitative studies. They provide richness and depth, facilitating a better understanding of any unintended consequences of telehealth beyond the planned outcomes [19]; as such, they served as our primary data sources. The inclusion and exclusion criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Published in a peer-reviewed journal
- Examined synchronous telehealth (telephone or video) or synchronous telehealth in conjunction with asynchronous models for service delivery as part of the Australian health system
- Qualitative or mixed methods evaluations that examined the implementation of synchronous telehealth or synchronous telehealth in conjunction with asynchronous models or its unintended consequences in both routine practice or evaluation of specific programs or quantitative studies in which the main focus was the unintended outcomes of telehealth
- Unlimited in setting or specialty because the goal was to identify the unintended consequences of synchronous telehealth as a modality and an alternative model of care regardless of the specialty or setting; however, we extracted setting and specialty data for each study included
- Published in English and available in full text

Exclusion criteria

- Reviews, protocols, opinion pieces, commentaries, perspectives, conference proceedings, and books
- Focused solely on asynchronous telehealth modalities: store and forward, patient monitoring, mobile apps, studies on telehealth for education or professional development purposes only, or studies assessing telehealth readiness
- Carried out across multiple countries
- Not available in full text or not in English
- Examining after-hours helpline and triage programs (these were excluded because such programs primarily serve as entry points for assessment and referral services rather than as replacements for standard care [20,21])

Study Screening

To facilitate screening and duplicate removal, all references were downloaded to a reference manager (EndNote 20; Clarivate) and then uploaded to the web-based screening platform Rayyan (Rayyan Systems Inc). The first author (SO) removed all duplicates, and then 2 reviewers (SO and KC) independently completed a blind review of a random sample of titles and abstracts (2%) against the aforementioned inclusion and exclusion criteria. Cohen κ was calculated (0.82; near-perfect agreement) to determine interrater reliability, and the few discrepancies in the inclusion or exclusion decisions were discussed between the 2 reviewers (SO and KC). When necessary, a third reviewer (LAE) was consulted until a consensus was reached. A single reviewer (SO) screened the remainder of the titles and abstracts against the inclusion and exclusion criteria and conducted the full-text screening of the included references. As there is no search protocol for conducting a CIS, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Data Extraction

Study characteristics and findings on the types and models of telehealth services were extracted into a custom template in NVivo 20 (Lumivero) and Excel (Microsoft Corp). We extracted theoretical frameworks or models only if they underpinned the design and analysis of the study, with sufficient detail provided to explain their application. Textbox 2 summarizes the data items captured to answer RQs 1 and 2. To answer RQ 3, we identified any unintended benefits or drawbacks by examining the reported results in each included study, with the relevant text extracted verbatim and added to NVivo for coding and analysis.

XSL•FO
RenderX

Textbox 2. Data items extracted.

Research questions and data items

- How is telehealth implementation being evaluated or assessed?
  - Study design and methods used
  - Data collection tools
  - Theories and frameworks used to understand the impact of telehealth
  - Locations and regions of studies
  - Types and number of participants
- What are the findings on the types and models of telehealth services implemented?
  - The types of telehealth services implemented
  - Settings and specialties
  - Models of care implemented
  - Whether it was synchronous or a mix of synchronous and asynchronous models
  - Modalities of telehealth (video vs telephone)
  - If the study was conducted before or after the COVID-19 pandemic
- What are the findings on the unintended consequences of telehealth?
  - Any identified unintended consequences
  - Whether they are positive, negative, or neutral

Dixon-Woods et al [17] point out that the precise definitions of many constructs in a CIS may evolve during the review process. Our research process exemplified this approach because we began data extraction without a rigid definition of unintended consequences. Instead, we initially extracted all benefits and drawbacks of telehealth from included studies; for instance, we initially considered *reduced travel time for patients* an unintended positive consequence. However, in all included studies, the intended goal of telehealth implementation, whether explicit or implied, was to improve or sustain access to health care services. This definition encompassed using telehealth to ensure continuity of access during the COVID-19 pandemic and to improve access for patients with barriers such as geographic and mobility disadvantages. Therefore, we conducted another round of analysis to refine our understanding of unintended consequences, excluding benefits related to the reduced burden of accessing care in terms of money and time because these were apparently intended outcomes from the studies. Hence, in all instances, an unintended consequence was subsequently defined as either (1) any benefit of telehealth implementation other than improving or sustaining patients' access to health care services or (2) any drawback because drawbacks are usually not planned or intended although they can be foreseen or expected.

While we recorded whether each study was conducted before or after the COVID-19 pandemic, our data extraction focused on unintended consequences that were not explicitly tied to the unique circumstances of the pandemic. Therefore, we did not extract or include transient consequences solely attributable to the COVID-19 context (eg, consequences related to extreme social distancing measures). Our rationale was that these transient consequences might have limited applicability to long-term, routine telehealth implementation in typical care provision. This methodological approach ensured that our analysis focused on enduring issues relevant to long-term, routine telehealth implementation, regardless of whether the study was conducted before or during the COVID-19 pandemic. Nevertheless, by documenting the temporal context (before or after the pandemic) for each study in Multimedia Appendix 2 [6-8,22-112], we provide the necessary information to interpret the results in light of the implementation context.

Quality Appraisal

We wanted to be inclusive in our review and not omit any study based on quality issues; therefore, we purposely did not use a structured tool to appraise studies. This is in agreement with the growing case against the exclusion of relevant qualitative studies in reviews on quality grounds alone rather than relevance because they can nevertheless inform and contribute to the understanding as well as the richness of the final findings of a review [17,113,114].

Data Synthesis and Analysis

To analyze the extracted data, an iterative coding process was used, switching between inductive and deductive coding and following a best-fit framework-based synthesis of evidence [115-118]. This approach allowed us to identify relevant frameworks for analysis as we coded, while minimizing the risk of missing some relevant evidence that did not fit an a priori framework, thus capturing all unintended consequences extracted from the included studies. As we undertook this process, our initial broad question "What are the unintended consequences of telehealth?" evolved into "How do unintended

XSL•FO
RenderX

consequences of telehealth manifest across different domains of health care performance?" because it became apparent that telehealth's unintended consequences could be conceptualized in terms of accessibility, appropriateness, safety, efficiency, effectiveness, and continuity of care. We recognized that these concepts aligned closely with the definitions in the Australian Health Performance Framework (AHPF) [119], a framework with which we were familiar from previous work in the Australian health care context, which is described in detail in the following subsection. We chose to use the AHPF domains instead of defining entirely new constructs as typical when conducting a CIS for the following reasons: (1) the AHPF provided a robust, established framework that could accommodate the complexity of our findings; (2) its use allows for better comparison with other telehealth studies in the Australian context; and (3) it offers a more accessible account of the evidence for policy makers and practitioners.

Our initial inductive coding process resulted in 137 codes; after an iterative process of refining, merging, and deleting redundant codes, 82 (59.9%) unique codes were deductively mapped and translated into the domains and dimensions of the AHPF. As an example to demonstrate this iterative inductive-deductive process, we initially coded a wide range of positive and negative effects of telehealth related to continuity of care, such as *facilitating multidisciplinary care* and *undermining rapport building*. As we analyzed the data, we recognized that this theme aligned with the *continuity of care* dimension in the AHPF. However, we found it necessary to modify the framework to better summarize and capture consistent variations and nuances in the findings within each dimension (Figure 1); for example, we added subcategories under *continuity of care*, such as *interpersonal continuity* and *management continuity*, to capture telehealth's unintended consequences related to the *patient-clinician relationship* and *care management across teams*. Thus, we adapted the existing AHPF to document the full impact of telehealth and its unintended consequences within the Australian health system. This approach aligns with the view of Dixon-Woods et al [17] of CIS as a flexible approach that can be adapted to the needs of the RQ, rather than a rigid, prescriptive methodology. NVivo (version 20; Lumivero LLC) was used to facilitate coding and analysis.

Figure 1. Domains (in gray), dimensions (in brown), and subcategories (in blue) of the extracted consequences mapped into the domains and dimensions of the Australian Health Performance Framework.

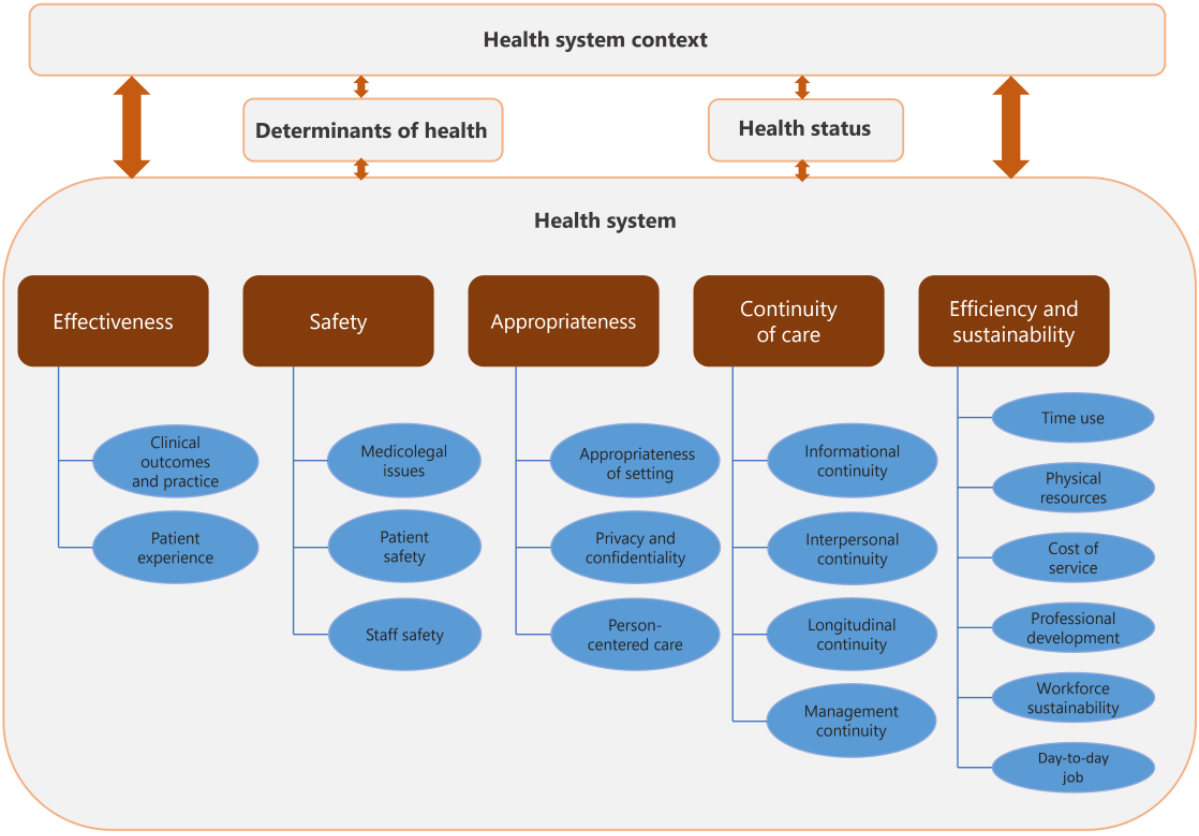

Mapping the Findings to the AHPF Domains

The AHPF supports reporting by different health organizations at different levels to assess the performance of health care delivery in Australia, focusing on four key domains (in gray in Figure 1): (1) determinants of health, (2) health status, (3) health system, and (4) health system context [119]. Within each domain, there are dimensions to guide the assessment of that specific domain; for example, in the health system domain, the AHPF uses accessibility, effectiveness, efficiency and sustainability, safety, appropriateness, and continuity of care

XSL•FO
RenderX

as dimensions to guide the assessment of the performance of the Australian health system.

We mapped all extracted consequences into the 4 key domains of the AHPF. Due to the greater abundance of findings falling under the health system domain than under other domains, the system-level consequences constituted most of the evidence presented in this review and necessitated the use of both dimensions and subcategories (shown in brown and blue in Figure 1, respectively), as discussed previously. To map these extracted consequences to the health system domain of the AHPF, we primarily used the definitions provided in the framework for each dimension. Where there was no clear definition, we supplemented with definitions from other sources, as summarized in Textbox 3. While we used the AHPF as an organizing framework, our approach went beyond simple mapping to achieve synthesis. We integrated findings across the included studies to generate new insights into how unintended consequences manifest in telehealth implementation; for instance, by synthesizing the findings related to effectiveness and appropriateness, we developed a new understanding of how telehealth can improve health outcomes by facilitating the involvement of carers in care delivery. By situating our findings within the AHPF, we emphasize telehealth's overarching impact across various health care performance domains, offering a systems-level perspective on telehealth's unintended consequences.

Textbox 3. Definitions of the Australian Health Performance Framework health system dimensions as used in data analysis.

Definitions and dimensions

- Accessibility is the ability to access affordable, convenient care at the right time and place, taking into account the various needs and circumstances of patients [119,120]; benefits in terms of time, burden, and costs of accessing care were not included in the final synthesis because they were considered intended benefits (discussed in the Data Extraction section above)
- Continuity of care refers to uninterrupted care or service across programs, practitioners, and levels over time [121]; as such, consequences under this dimension include findings on care planning and coordination, information sharing between providers, facilitation of multidisciplinary care, frequency of visits, and interpersonal exchange between patients and their providers [121,122]
- Appropriateness is defined as providing care that is person-centered and culturally sensitive, allowing patients to be involved in the decision-making regarding their health choices; when care is appropriate, patients are encouraged to share their experiences and provide feedback on the care they receive without fear of consequences, thus ensuring their dignity and the confidentiality of their health information during and after receiving care [120]
- Safety refers to the reduction or avoidance of harm due to, and during, health care provision or from the environment in which care is delivered [120]; any findings about patients' safety, staff safety, and medicolegal issues resulting from lack of safety in health care were mapped under this dimension
- Effectiveness refers to achieving the desired clinical outcomes, taking into consideration patients' perspectives; as such, patient-reported outcomes and patient-reported experiences constitute an essential component of evaluating health care effectiveness [120]; moreover, indicators such as preventable hospitalizations, preventable deaths, and screening and immunization rates are all measures of health care effectiveness [119]; any consequences of telehealth that support or undermine desirable clinical outcomes were mapped under this dimension
- Efficiency and sustainability refers to achieving desirable clinical outcomes at the minimum cost possible and optimal use of health care resources. such as workforce and physical resources, all while innovating to meet increasing health care demands without overstretching the workforce [120]; therefore, findings on telehealth impact on the cost of service, health workforce recruitment and retention, and supporting clinicians were mapped under this dimension

Results

Study Selection and Characteristics

Of the 4241 records retrieved, 2340 (55.18%) were screened after 1901 (44.82%) duplicates were removed. After title and abstract screening, 2125 (90.81%) of the 2340 studies were excluded, resulting in 215 (9.19%) studies progressing to full-text review. Studies for which abstracts were unavailable progressed directly to full-text review. In addition, survey studies that fulfilled the inclusion criteria progressed to full-text screening because they could have open-ended sections with relevant data on unintended consequences. At the full-text review stage, of the 215 studies, 121 (56.3%) were excluded, leaving 94 (43.7%) studies (Multimedia Appendix 2) that evaluated telehealth programs in Australia and were eligible for data extraction and analysis. Figure 2 summarized the PRISMA search and screening protocol followed and the number of studies at each stage.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) search and screening protocol for the literature review on the unintended consequences of telehealth.

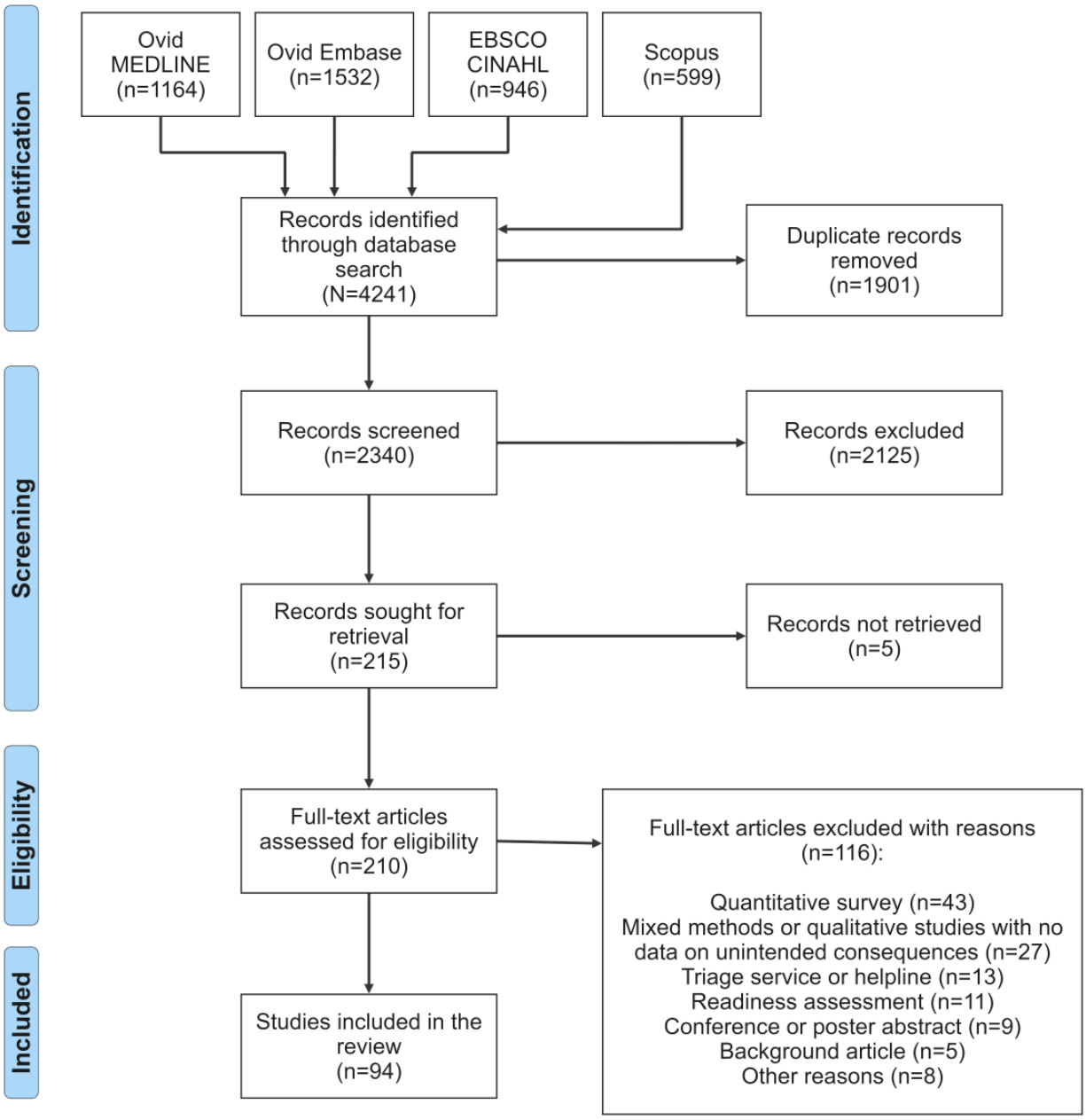

More than half of the included studies (50/94, 53%) were published in the last 3 years (after the onset of the COVID-19 pandemic; Table 1). Nearly a third of the studies (29/94, 31%) were conducted in multiple states or territories across Australia, with the remainder being conducted in Queensland (23/94, 24%), Victoria (18/94, 19%), New South Wales (14/94, 15%), South Australia (5/94, 5%), Western Australia (3/94, 3%), Tasmania (1/94, 1%), and the Northern Territory (1/94, 1%). The fewest studies were conducted in a metropolitan setting (18/94, 19%), with most being conducted in rural or regional settings (32/94, 34%) or in mixed settings (31/94, 33%). Of the 94 studies, 13 (14%) lacked data about whether the study was in metropolitan, rural, or regional settings.

XSL•FO
RenderX

Table 1. Number of studies per year (N=94).

Year	Studies, n (%)
2000	2 (2)
2002	1 (1)
2003	1 (1)
2009	1 (1)
2012	5 (5)
2014	4 (4)
2015	5 (5)
2016	5 (5)
2017	1 (1)
2018	3 (3)
2019	5 (5)
2020	10 (11)
2021	18 (19)
2022	32 (34)
2023	1 (1)

Most of the included studies (79/94, 84%) evaluated a synchronous-only model, whereas 16% (15/94) assessed a telehealth model using both synchronous and asynchronous modalities. Almost a third of the included studies (28/94, 30%) examined telehealth in various allied health settings, with the remainder being conducted in other primary care settings (8/94, 9%), mental health (14/94, 15%), or other specialties (oncology 7/94, 7%; palliative care 4/94, 4%; rheumatology 3/94, 3%; emergency medicine 3/94, 3%; surgery 3/94, 3%; hematology 2/94, 2%; geriatrics 2/94, 2%; gynecology and reproductive medicine 2/94, 2%; pulmonology 1/94, 1%; hepatology 1/94, 1%; nephrology 1/94, 1%; endocrinology 1/94, 1%; cardiology 1/94, 1%; anesthesia 1/94, 1%; radiology 1/94, 1%; multiple 11/94, 12%). The majority of the included studies evaluated a patient-to-provider telehealth model (71/94, 76%) whereas the remainder of the studies assessed provider-to-provider models (11/94, 12%), group therapy (3/94, 3%), or a combination of models (9/94, 10%), as shown in Table 2.

XSL•FO
RenderX

Table 2. Summary of characteristics of the included studies (N=94).

Study characteristics	Studies, n (%)
Methods	
Mixed methods	50 (53)
Qualitative	42 (45)
Quantitative	2 (2)
States or territories	
Queensland	23 (24)
Victoria	18 (19)
New South Wales	14 (15)
South Australia	5 (5)
Western Australia	3 (3)
Tasmania	1 (1)
Northern Territory	1 (1)
Multiple	29 (31)
Regions	
Regional or rural	32 (34)
Mixed	31 (33)
Metropolitan	18 (19)
Not reported	13 (14)
Type of telehealth	
Synchronous	79 (84)
Synchronous+asynchronous	15 (16)
Modalities	
Video	54 (57)
Video+telephone	34 (36)
Telephone	5 (5)
Not reported	1 (1)
Specialties	
Allied health	28 (30)
Mental health (psychology and psychiatry)	14 (15)
Primary care	8 (9)
Other specialist services	44 (47)
Participant groups	
Health care providers	46 (49)
Consumers	23 (24)
Health care providers+consumers	22 (23)
Not applicable	3 (3)
Models of care	
Patient to provider	71 (76)
Provider to provider	11 (12)
Group therapy	3 (3)
Multiple	9 (10)

XSL•FO
RenderX

Approximately half of the studies investigated the perspectives of providers in a range of different roles (46/94, 49%), nearly a quarter examined the views of consumers and carers (23/94, 24%), and 23% (22/94) included the perspectives of both consumers and health care providers. A few of the studies (3/94, 3%) did not recruit participants: they either examined existing documents or analyzed routinely collected data. Only 14 (15%) of the 94 studies used theoretical frameworks or models to underpin the study design and analysis, with the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability Framework [123] being the most commonly used framework among these 14 studies (n=3, 21%).

Unintended Consequences of Telehealth in Australia

The unintended consequences of telehealth extracted from the included studies were mapped into the 4 domains of the AHPF, as described in the following subsections and summarized in Figure 3.

Figure 3. The unintended consequences of telehealth across the Australian Health Performance Framework (AHPF) domains and dimensions and the newly devised subcategories across each dimension. The AHPF domains are interrelated; for example, health status and determinants of health are influenced by, and influence, the dimensions of the health system and the health system context, while the health system influences the broader context within which it exists and vice versa.

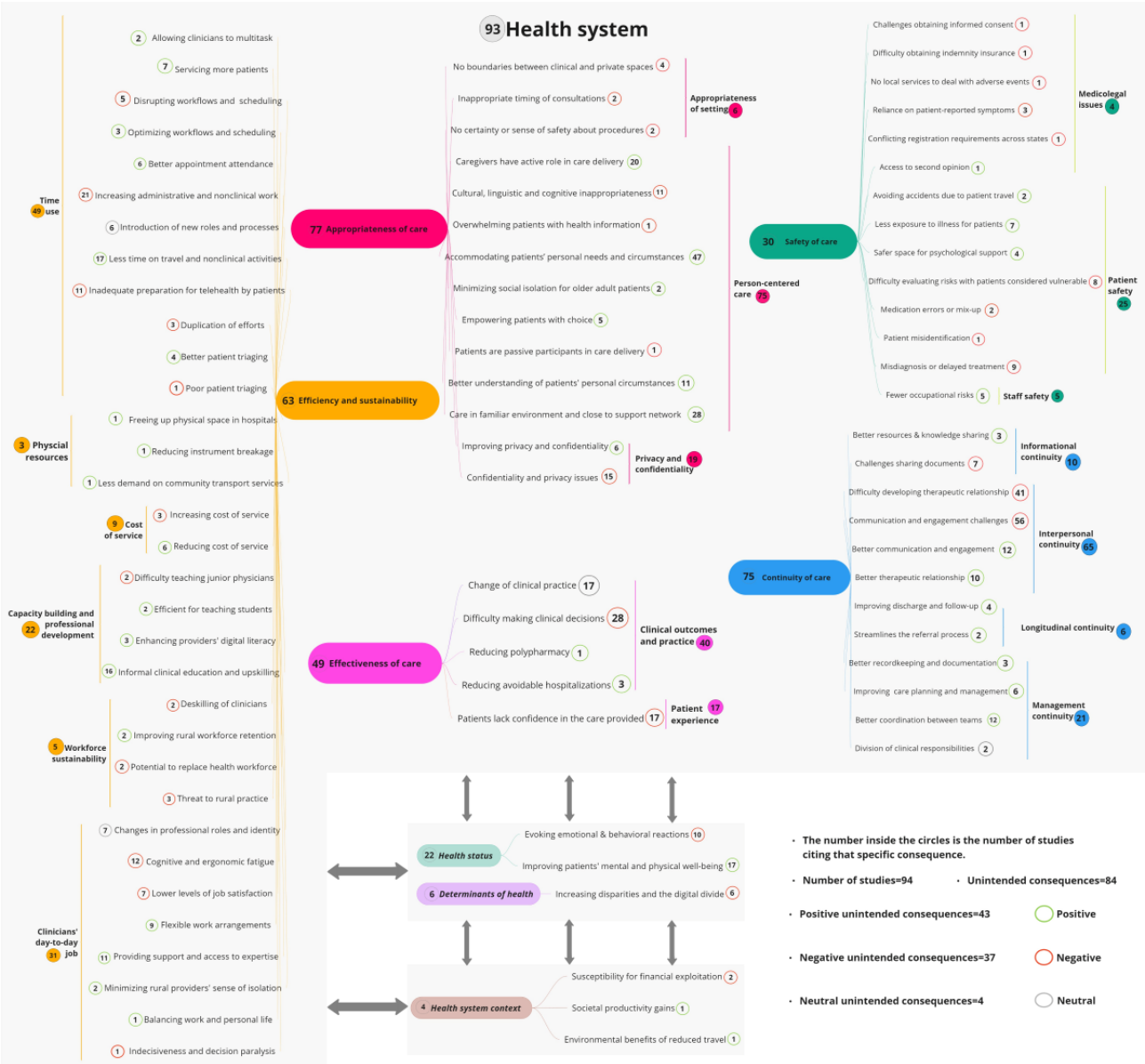

Health Status

Under this domain, nearly a fourth of the studies (23/94, 24%) reported unintended consequences of telehealth associated with health status, with most of the studies (17/23, 74%) reporting positive consequences; for example, 71% (12/17) of these studies reported that telehealth improved patients' mental and physical well-being, mostly because patients were in the comfort of their homes where they felt more relaxed, less vulnerable, and in control compared to being in an unfamiliar or clinical environment [8,22-32]. Some of the studies (6/17, 35%) highlighted other contributing factors to this improved well-being, such as the reduction in travel-induced anxiety and fatigue [24,33-36], as well as the social support resulting from connecting with people with similar experiences or challenges in group therapy [37]. However, several of the studies (10/23,

XSL•FO
RenderX

43%) highlighted some negative unintended consequences of telehealth, such as evoking behavioral reactions in children [25,28,38] and negative emotions from seeing one's image on the screen [31,32,34,39], including for those with, for example, eating disorders [31,32] or hearing aids [39]. In some of the studies (5/23, 22%), telehealth was found to provoke anxiety and discomfort in patients [6,25,40,41].

Determinants of Health

Of the 94 studies, 6 (6%) reported a potential negative impact of telehealth on individuals' socioeconomic status and other determinants of health. Specifically, 5 (83%) of these 6 studies reported concerns about the potential of telehealth to increase the digital divide due to the increased financial burden on consumers, resulting from internet and equipment costs [40,42-45]. In 2 (33%) of the 6 studies, lack of digital literacy [44], particularly among the older populations [30], was highlighted as another factor that could increase this digital gap.

The Health System Context

The findings of 4 (4%) of the 94 studies under the health system context emphasized telehealth's impact on factors external to, and generally beyond, the direct influence of the health system. Although 2 (50%) of these 4 studies reported positive outcomes of telehealth associated with this domain, such as societal productivity gain [46] and the environmental benefits of reduced travel [47], 2 (50%) studies highlighted the potential for financial exploitation through telehealth [48,49].

The Health System

This domain accounts for a majority of the reported unintended consequences of telehealth synthesized in this review. Of the 94 included studies, 93 (99%) reported findings with unintended consequences falling across the dimensions of the health system domain as detailed in the following subsections.

Accessibility

There were no reported unintended consequences under this dimension because all extracted outcomes were related to reducing the cost and time of accessing care and, as such, were considered intended benefits, as indicated earlier.

Safety

In 30 (32%) of the 93 studies, the safety of care provision was shown to be impacted by telehealth, with more than half (n=18, 60%) noting 9 negative consequences. In 10 (56%) of these 18 studies, telehealth was reported to jeopardize patient safety due to patient misidentification [45], medication errors [41,45], the risk of misdiagnosis [29,45,47,50-55], or delayed treatment [29,45]. Of these 10 studies, 6 (60%) highlighted some factors that give rise to these risks of misdiagnosis and delayed treatments in telehealth, such as lack of nonvisual cues [45,47,51,52], patients not preparing properly for a telehealth consultation as they would for an in-person consultation [45], and the lack of a physical examination [45,51,53,54]. According to 8 (44%) of the 18 studies noting negative consequences, the lack of a physical examination was believed to have led to overreliance on patient-reported symptoms [45,47,51,53,55], potentially increasing medicolegal liabilities [50,56,57]. Moreover, the findings of 3 (17%) of the 18 studies suggest that the use of telehealth could lead to medicolegal liabilities due to conflicting state registration requirements, potentially causing uncertainty surrounding professional liability in case of clinical incidents [58]; a lack of local services to deal with adverse events [50]; and challenges obtaining informed consent from patients due to miscommunication challenges [57]. Difficulties in evaluating, identifying, and addressing risk through telehealth were reported in 8 (44%) of the 18 studies involving patients considered vulnerable, including those who have experienced domestic violence or are at risk for suicide [31,32,40,45,52,59-61].

By contrast, 16 (53%) of the 30 studies noted 5 positive telehealth outcomes associated with health care safety. Of these 16 studies, 5 (31%) demonstrated that telehealth improved staff safety by reducing occupational risks associated with home visits [24,50,56], work-related travel [50], exposure to illness [62], and carrying work equipment [22]; and 13 (81%) highlighted the potential of telehealth to improve patient safety by reducing accidents resulting from patient transportation [50,63] and minimizing exposure to illness [29,36,64], which was particularly beneficial for patients with a compromised immune system [7,22,65,66]. According to 4 (25%) of the 16 studies, telehealth provided a safe space for psychological support for patients considered vulnerable [30,31,40,67]. Finally, the findings from 1 (6%) of the 16 studies suggested that telehealth has the potential to reduce medicolegal liabilities by reducing accidents while transporting patients, minimizing the risk of misdiagnosis by providing additional expertise through second opinions, and providing better incident auditing [50].

Appropriateness

Of the 93 studies, 77 (83%) documented several unintended consequences under the appropriateness dimension in terms of person-centered care, the appropriateness of the setting, and the privacy and confidentiality of patients.

Of these 77 studies, 75 (97%) reported telehealth consequences—primarily positive—related to person-centered care. Several telehealth benefits were documented in numerous studies (60/75, 80%), such as minimizing social isolation for older adult patients [66,68], allowing patients to receive care in a familiar and comfortable environment [7,8,23-26,28-30,32,33,64,68-77], allowing patients to receive care close to their support network [27,69,78,79], and accommodating patients' personal needs and circumstances [6-8,22,24,26,28-30,32,34-37,40,43,45,47,48,50-52,55,59-61,64,65,67-70,75-77,80-91]. These factors helped patients to feel relaxed and encouraged them to open up more easily [36,52,70-72,80]. In 5 (7%) of the 75 studies, telehealth was reported to empower patients and reinforce their autonomy [31,39,50,77,82], in contrast to findings from a study in which several patients described their role as "passive" during telehealth consultations compared to in-person consultations [51].

In several of the studies (11/75, 15%), clinicians shared their views on how telehealth allows them to see patients in their personal spaces, enabling them to provide more personalized care [8,32,42,55,61,67,73,75,80,82,92]. A number of studies (20/75, 27%) highlighted that telehealth facilitated the involvement of family members and carers and supported them

XSL•FO
RenderX

in having an active role in care delivery [8,22,25,26,29,32,34,38,56,62,63,67,77,79,85,93-97]. However, this increased involvement was viewed differently by clinicians and carers; some carers complained about the increased burden and responsibilities [22,38,56,93]. By contrast, clinicians were anxious about losing control due to reliance on carers to perform tasks they would have otherwise performed themselves in an in-person consultations if not for the limitations of delivering care via telehealth [25,94]. In addition, clinicians complained about parents' noncompliance to given instructions, which affected the quality of care provided during telehealth consultations [25,94].

A number of studies (12/75, 16%) pointed out other consequences of telehealth that undermined person-centered care. In 7 (58%) of these 12 studies, clinicians raised concerns about the cultural, cognitive, and social inappropriateness of telehealth for various groups such as Indigenous communities [47,83], patients with mental or cognitive impairment [49,83,97], patients with hearing impairments [29,30], patients who were acutely unwell [83,97], and rural and remote communities where an essential part of delivering care to patients is knowing their social circumstances [44]. Furthermore, according to the findings from 8 (67%) of the 12 studies, communication challenges were reported in telehealth consultations involving non–English-speaking patients or patients who required interpreters to be present [29,30,49,52,57,64,83,98]. In 1 (8%) of the 12 studies, clinicians shared their concerns that telehealth patients were being overserviced and overloaded with information relating to their health [42].

Of the 77 studies, 4 (5%) reported findings about the appropriateness of telehealth as a health care delivery setting compared to physical or brick-and-mortar settings. In 2 (50%) of these 4 studies, patients shared that telehealth removed the sense of safety and certainty that came from being surrounded by health professionals in physical settings where they felt more supported and had multiple opportunities to raise any concerns or issues [31,51]. A different study reported scheduling confusion and inappropriate timing of telehealth consultations [29], which was supported by findings from another study in which patients shared their frustration that clinicians prioritized in-person consultations and fitted telehealth consultations in between, leading to scheduling confusion and inappropriate timing of consultations [64].

Telehealth consequences related to patient privacy and confidentiality were documented in 19 (25%) of the 77 studies, with negative consequences being reported in most of them (n=16, 84%). Of the 19 studies, 8 (42%) reported clinicians' concerns about patients' privacy and confidentiality in telehealth consultations due to lack of private spaces to discuss sensitive matters, especially for patients considered vulnerable [28,40,42,45,50,52,54,89]. Similarly, in 5 (26%) of the 19 studies, telehealth was viewed as blurring the boundaries between private and clinical spaces [28,29,31,32], thereby invading the privacy of both patients [28,32,89] and clinicians [29]. In 5 (26%) of the 19 studies, patients were reportedly concerned about their privacy and confidentiality in telehealth consultations due to a distrust of technology [32], fear of cybersecurity breaches [64], and the possibility that other people might overhear their conversations [29,31,36]. According to 2 (11%) of the 19 studies, errors such as clinicians calling the wrong person [45] or patients mistakenly joining other patients' consultations [33] were reported as jeopardizing privacy during telehealth consultations. In addition, 2 (11%) of the 19 studies documented clinicians' views on the inappropriateness of a telehealth physical examination because patients get undressed in front of the camera [47,54].

These issues notwithstanding, 6 (32%) of the 19 studies reported positive consequences of telehealth associated with patient privacy and confidentiality. Of these 6 studies, 4 (67%) reported that telehealth allowed patients to seek care anonymously—in contrast to in-person care where patients may be seen or questioned by acquaintances when seeking care [52,61,83,99]. This feature of telehealth was particularly valued in small and Indigenous communities and when the topic was sensitive because patients had the option to seek care outside their community [61]. In 2 (33%) of the 6 studies, patients shared that telehealth allowed them to talk freely in the privacy of their homes compared to hospitals and other clinical settings where there is a risk of other people overhearing their conversations [72], all while controlling the level of exchange they prefer with the ability to turn the video on and off [31].

Effectiveness

Of the 93 studies, 49 (53%) reported telehealth's unintended consequences pertaining to the effectiveness of care in terms of clinical practice and clinical outcomes, as well as patient experience. Documented in 28 (57%) of the 49 studies, the most cited negative unintended consequence of telehealth under this dimension was the clinicians' inability to make clinical decisions when delivering care via telehealth, potentially impacting clinical outcomes [22,24,25,28-30,32,44,45,47,51-56,59,62,65-67,80,84,92,97,98,100,101]. The factors contributing to these challenges in clinical decision-making were presented in a number of studies (26/49, 53%), such as a lack of nonverbal cues as well as communication challenges [22,24,25,28-30,45,51,52,56,66,80,92], the inability to examine patients in person [29,30,32,45,47,53-56,59,65,66,84,92,97,98,101], and a lack of local and contextual understanding of the patients' circumstances, particularly when providing care via telehealth to rural and regional patients [44,100].

According to a number of studies (17/49, 35%), telehealth forced clinicians to change and adapt their clinical practices and processes to fit the new mode of delivery [8,22,29,31,45,49,50,52,56,59-61,66,91-93,95]. In several cases (10/17, 59%), this adaptation was believed to be a consequence of the inherent limitations of telehealth, such as the inability to examine patients in person [29,92], a lack of nonverbal cues that affect communication and rapport building [8,45,91,92,95], and the inability of the health professional to guide and provide physical support to patients or encourage participation [22,60,93]. In addition, in 2 (12%) of the 17 studies, clinicians adapted their clinical practices to ensure patients' privacy during telehealth consultations [91] and to mitigate the risk of medicolegal liabilities [50]. Regardless, in a few studies (4/49, 8%), telehealth was viewed as potentially improving clinical

XSL•FO
RenderX

effectiveness by reducing avoidable hospitalizations [63,83,102] and polypharmacy [63].

Unintended consequences in relation to patient-reported experiences of receiving care via telehealth were documented in 17 (35%) of the 49 studies in which patients shared concerns and a lack of confidence in the care they received and questioned whether it was as thorough and accurate as in-person care [6,7,25,29,32,33,35,36,38,41,44,45,51,64,69,84,101]. Their primary concerns included that clinicians may have missed something when delivering care via telehealth due to the lack of an in-person physical examination and physical interactions [6,7,33,35,36,38,44,51,64,69,84,101], communication challenges [6,32,44,64], and telehealth consultations being rushed by clinicians [6,36,41,64].

Continuity of Care

Of the 93 studies, 75 (81%) documented telehealth's unintended consequences associated with the 4 categories of continuity of care: informational continuity, interpersonal continuity, longitudinal continuity, and management continuity.

On the basis of the findings of 7 (9%) of these 75 studies, informational continuity was hindered by the use of telehealth, such as when sharing documents, test results, and prescriptions with patients [29,41,69,98,103] or with other clinicians to assist in clinical decision-making [29,65,104]. However, these findings are in contrast to what was reported in a study in which clinicians reported better access to, and sharing of, resources [22]. Likewise, in 2 (3%) of the 75 studies, telehealth was believed to improve informal knowledge sharing between providers to support care delivery [50,105].

In a number of studies (18/75, 24%), telehealth was viewed as facilitating case management and multidisciplinary care because it improved care coordination and communication between teams [30,32,42,48,50,63,79,84,86,88,102,105]; supported better recordkeeping and documentation [30,50,103]; and improved care planning [24,32,45,55,64,65], which was particularly beneficial in the management of chronic and complex conditions [45,64,65]. In addition, the findings from 6 (8%) of the 75 studies suggested that telehealth supported longitudinal continuity across the continuum of care by streamlining the referral process [63,100], allowing clinicians to make informed discharge decisions [24,84], and improving the follow-up experience for patients [45,47]. Finally, a division of clinical responsibilities, role reassignment, and changes in power dynamics were noted as unintended consequences of telehealth when multiple clinicians were involved in case management [50,106].

Of the 75 studies, 65 (87%) reported unintended consequences of telehealth related to interpersonal continuity of care in terms of interpersonal exchange and therapeutic relationship building. Specifically, in 56 (86%) of the 65 studies, telehealth was viewed as hindering effective communication and patient engagement [7,22-25,28-33,35-38,40-45,47-49,51,52,55-58,60-62,64,66,67,71,75,76,81,83,84,90,91,93-99,101-103,107,108], largely due to the lack of nonverbal information [7,22,24,25,28-33,35-38,40-45,47-49,51,52,55-58,61,64,66,71,76,81,84,90,91,93,95,101,107,108] and, to a lesser extent, due to the increase in distractions during telehealth consultations compared to in-person consultations [22,25,32,38,40,62,67,75,91,93,94,96], particularly in the case of children or young patients [22,25,32,38,40,62,67,75,93,94,96]. Nevertheless, these findings are in contrast to those reported in 12 (18%) of the 65 studies that documented an improvement in communication and patients' engagement in telehealth [22,28,34-36,40,50,52,70,71,84], primarily due to patients being comfortable in their homes and more inclined to open up about sensitive health issues [36,40,52,70,71], fewer distractions for both patients and clinicians [22,34,75], and the increased interest in technology among children [35]. Of the 12 studies, 2 (17%) suggested that patients opened up more easily and were less inhibited over telehealth because they were unaffected by clinicians' presence and any nonverbal cues that may have signaled patients to stop talking [50,84].

In 41 (63%) of the 65 studies reporting unintended consequences of telehealth related to interpersonal continuity of care, telehealth interactions were noted as impersonal, thus undermining rapport building and affecting the therapeutic patient-provider relationship [6,7,22,24,28-31,35,36,38-42,44,45,47-52,56,60,61,64,68,76,80,81,84,90,91,93-95,97,98,101,107]. However, contrary to these findings, 10 (15%) of the 65 studies reported better therapeutic relationship building in telehealth as a consequence of the improved communication and engagement in telehealth [42,75] as well as the increased frequency of contact between patients and their providers over extended periods of time [26,29,68,71,81,84,85,99].

Efficiency and Sustainability

Of the 93 studies, 63 (68%) reported unintended consequences of telehealth associated with the efficiency and sustainability of health care in terms of its impact on health workforce and resources use. According to the findings from 11 (17%) of these 63 studies, health professionals reported that telehealth supported them in their daily tasks by providing access to a second opinion or specialized expertise [40,44,50,58,68,77,83,85,108-110] and thereby minimized their feeling of isolation [50,109], potentially improving rural workforce retention [44,50]. In addition, other unintended benefits of telehealth for clinicians were documented in many of the studies (11/63, 17%), such as improved digital literacy and confidence using technology [37,68,111], improved work-life balance [22], and increased flexibility in terms of the time and location of consultations [24,29,37,40,68,82,83,88,94]. However, according to a study, this increased flexibility made possible by telehealth caused decision paralysis for clinicians [31].

On the basis of the findings from 11 (17%) of the 63 studies, telehealth caused health care providers cognitive fatigue [22,29,31,32,45,51,52,55,80,93,94] due to the increased energy demand to maintain patient engagement [22,32], the higher concentration required due to the lack of nonverbal cues [29,52], the new routines and administrative work resulting from delivering care via telehealth [93,94], and the constant anticipation of what comes next and the adaptation to accommodate these new routines and changes [31,45,51]. In 1 (2%) of the 63 studies, telehealth was reported to have caused

XSL•FO
RenderX

ergonomic fatigue due to the extended periods providers spent sitting when providing care via telehealth [92]. As documented in 7 (11%) of the 63 studies, telehealth was viewed as less rewarding and less satisfying to clinicians than in-person care delivery due to the absence of the personal touch and therapeutic connection [29,51,53,55,80,84,92]. Finally, 7 (11%) of the 63 studies noted a consequence of telehealth where it was reported to affect clinicians' roles and professional identity [49,50,85,92,93,106,110].

A number of studies (23/63, 37%) documented both positive and negative unintended consequences of telehealth associated with workforce sustainability and professional development. In particular, 2 (9%) of the 23 studies reported concerns about telehealth deskilling metropolitan clinicians providing care to rural regions via telehealth due to a lack of the hands-on experience necessary to build clinical skills [78] as well as rural clinicians due to a lack of opportunities to practice [57]. Thus, as noted in 3 (13%) of the 23 studies, telehealth could potentially threaten the viability and sustainability of rural practice by leading to the loss of rural patients to metropolitan providers [44,50,57]. Furthermore, in 2 (9%) of the 23 studies, telehealth was viewed as a technological replacement for health professionals [81], potentially leading to shortages in the workforce and exacerbating health care accessibility issues in the long term [50]. Nevertheless, as evidenced by a number of studies (18/23, 78%), telehealth was reportedly efficient for student supervision [68,82], improved clinicians' access to informal training and education, and supported capacity building and continuing professional development for rural health professionals [32,42,50,58,63,77-79,83,85,88,99,100,102,105,112]. However, 2 (9%) of the 23 studies presented contradictory findings because several clinicians shared difficulties teaching junior physicians and students via telehealth [29,80].

Of the 63 studies, 51 (81%) documented 12 unintended consequences pertaining to the impact of telehealth on the use of resources in health care. On the basis of evidence from a number of studies (29/51, 57%), telehealth enhanced productivity, improved the timeliness of care, and enabled better use of physical resources; for instance, in 20 (69%) of the 29 studies, telehealth was reported to have saved clinicians' time [29,49,56,58,88,102], particularly time spent on commuting or traveling to clinical facilities or patients' homes [7,22,50,62,65,68,82,83,89,103,106] and social and nonclinical activities [89], allowing them to service more patients [24,37,62,66,68,82,88]. Other reported telehealth benefits associated with time use included allowing providers to multitask [43,92], improving appointment scheduling [22,82] and attendance by patients [22,24,29,68,84,98], speeding up the referral process [100], and streamlining patient triage [24,55,65,107]. According to 6 (12%) of the 51 studies, telehealth can potentially reduce health care system expenditures due to reduced cancellation rates [22], saving clinicians' travel time [58,79,83], avoiding unnecessary patient transportation [112], and reducing reimbursements of patient travel [88]. Finally, in 3 (6%) of the 51 studies, telehealth was reported to have resulted in efficient use of the physical resources essential for health care delivery because it freed up space in hospitals [92], lowered demand for community transport services [68], and reduced instrument breakage [22]. By contrast, many studies (34/51, 67%) noted unintended drawbacks of telehealth related to the degradation of efficiency in health care delivery. As evidenced in 21 (62%) of the 34 studies, this degradation in efficiency was primarily due to the increased time spent by clinicians on administrative and nonclinical workloads [22,31,37,43,45,47,49,51,53,55,59,62,65,68,81,83,93,94,96,104,112]. Furthermore, 16 (47%) of the 34 studies reported other sources of inefficiencies in telehealth, such as the duplication of efforts by both patients and clinicians [41,45,53]; poor triaging [80]; impaired and disrupted workflows [29,45,47-49,64]; and patients not preparing properly for a telehealth consultation as they would for an in-person consultation, wasting the clinician's time and theirs [29,30,32,45,49,51,52,67,76,80,84]. According to 3 (9%) of the 34 studies, inefficiencies in telehealth care delivery added to health care delivery costs, making telehealth cost-ineffective and financially unviable for providers [53,55,57]. Finally, the findings from 6 (18%) of the 34 studies noted telehealth's impact on health care efficiency because it added new processes and workflows or adapted existing ones [30,45,49,93,96,111].

Discussion

Principal Findings

Our synthesis of studies evaluating telehealth in Australia suggests that there are both positive and negative unintended consequences, often affecting the same aspect of health care. This conclusion was particularly evident in how telehealth impacted patients' privacy because several of the included studies (4/94, 4%) documented a positive impact of telehealth on privacy in small and Indigenous remote communities because it allowed anonymity [52,61,83,99]. By contrast, other included studies (8/94, 9%) reported privacy and confidentiality concerns, particularly for groups considered vulnerable, such as children and those who have experienced domestic violence [28,40,42,45,50,52,54,89], as a result of the blurred boundaries between personal and therapeutic or clinical spaces. These findings are in line with another issue highlighted in our review that telehealth compromised the safety of domestic violence victims and patients who are at risk of suicide because it was challenging to assess risks in the surrounding environment via telehealth [32,40,45,52,59-61]. These issues notwithstanding, a number of studies (13/94, 14%) in our review reported positive consequences of telehealth on patient safety due to various factors appertaining to the provision of care at home and in a familiar environment, as well as the elimination of unnecessary travel [7,22,29-31,36,40,50,63-67].

While the findings from our review present telehealth as a model that largely supports a person-centered approach to care delivery, the included studies also recorded challenges undermining person-centered care delivery via telehealth. Notably, telehealth was often reported to be inappropriate and unfit for the personal and cultural needs of certain groups, such as culturally and linguistically diverse communities [29,30,49,52,57,64,98], Indigenous patients [47,83], and patients with cognitive and hearing impairments [29,30,49]. The positive impact of

XSL•FO
RenderX

telehealth on longitudinal and management continuity of care featured prominently throughout our review. However, telehealth was viewed as a hindrance to interpersonal continuity of care and challenging for rapport building, and as such, telehealth seems better suited as a complementary model rather than a substitute for in-person care because strong interpersonal therapeutic relationships are necessary for safe, efficient, and person-centered care.

At times, telehealth resulted in neutral consequences with no clear intrinsic positive or negative attributes. Identifying these consequences can provide significant leverage to influence and improve care delivery via telehealth because they were often intermediate consequences that could lead to long-term positive or negative outcomes; for example, we categorized the introduction of new workflows and processes as a neutral consequence because it is not inherently positive or negative. However, the findings from multiple studies (5/94, 5%) in our review indicated increased workload and inefficiencies in service delivery due to these new workflows and processes [45,49,59,65,93]. Alternatively, according to 1 (1%) of the 94 included studies, telehealth provided an opportunity to redesign service delivery because it improved existing processes [105]. Another noteworthy issue is the differentiation between actual and potential outcomes because some of the studies (12/94, 13%) highlighted several potential unintended consequences—in addition to actual consequences—which were often voiced as concerns by stakeholders. These were often long-term consequences associated with workforce sustainability [50,57,78,81], the determinants of health [30,40,42-45], and the broader financial context within which the health system operates [48,49]. Including these potential outcomes in our review enables policy makers and health care planners to anticipate and prepare for future challenges and opportunities, facilitating proactive risk mitigation and benefit maximization, which is particularly important in the rapidly evolving field of telehealth.

International Relevance

While our study primarily focused on the Australian context, the unintended consequences of telehealth implementation identified in this review have broader international relevance. Many of our findings resonate with research conducted in countries with similar health care systems or those facing comparable geographic challenges. These findings suggest that certain unintended consequences of telehealth may be inherent to the modality itself rather than being solely context specific; for instance, a recent review, which included studies from multiple countries, corroborates our findings on telehealth's ability to empower patients, foster a perceived safe environment, and enhance family involvement [124]. While the review focused on telehealth for pain management, the findings from our review show that these patterns extend beyond pain management to encompass other specialties such as geriatric care, mental health services, and allied health disciplines, underscoring telehealth's capacity to facilitate a person-centered approach to health care delivery. Furthermore, the potential of telehealth to address significant contributors to clinicians' burnout, as highlighted in a Canadian study [125], corroborates our findings on the unexpected positive impacts of telehealth on clinicians' work-life balance, underscoring the importance of considering telehealth not only as a tool for patient care but also as a potential strategy for improving health care workforce sustainability.

Conversely, our review also uncovered negative unintended consequences of telehealth that align with international findings. The cognitive and physical fatigue experienced by clinicians due to increased administrative workload and heightened concentration demands, as identified in our Australian-focused review, are mirrored in an international scoping review [126]. These findings suggest that the challenges of adapting to telehealth modalities are a universal experience for health care providers across different health care systems. More critically, our findings highlight telehealth's potential to exacerbate existing health care disparities, a concern corroborated by research from the United States, suggesting that specific populations—particularly older adults, those with lower socioeconomic status, and non–English-speaking patients—faced challenges in accessing telehealth services during the COVID-19 pandemic [127]. This alarming trend underscores the crucial need to prioritize equity of access in telehealth implementation, regardless of the health care system context.

While many of our findings have international parallels, some of our results may be more specific to the Australian context; for instance, our observations regarding providing support and informal training for rural clinicians through telehealth might be particularly relevant to Australia's unique geographic and demographic circumstances. While countries such as Canada or the United States might face similar rural health care challenges, the specific implementation and impact of telehealth in these areas may differ due to variations in health care systems, policies, and cultural factors. Therefore, caution is needed when interpreting and translating these findings to other health care systems.

Implications for Practice and Policy Making

One challenge in telehealth consultations that was featured repeatedly in multiple studies (17/94, 18%) in our review is the inability to examine patients in person [29,30,32,45,47,53-56,59,65,66,84,92,97,98,101]. Some of the reported ways to mitigate the risk of wrong assessment due to this lack of in-person physical examination included selecting appropriate patients and cases for telehealth [50], spending extra time eliciting information from patients [45], lowering the diagnostic threshold [45,50], and conducting initial in-person assessments to gain a better understanding of the patient's condition and circumstances [50,54]. Evidently, clinicians rely on their discretion when determining which cases or conditions are appropriate and safe for telehealth and what risk mitigation strategies to use. Professional bodies should frequently release evidence-based guidelines on which aspects of health care delivery and what cases are appropriate and safe for telehealth to assist clinicians in decision-making and ensure standardization. Recently, the Medical Board of Australia released telehealth guidelines incorporating feedback from professionals and the community [128]. However, these guidelines lack details on cases, conditions, or aspects of care

XSL•FO
RenderX

that are more suited to telehealth than in-person care; in addition, they do not provide recommendations on risk mitigation strategies beyond the technological and procedural aspects and routine precautions that apply to the provision of care in general, regardless of modality.

Adopting a hybrid mode where providers combine telehealth and in-person care may lead to higher job satisfaction for clinicians because this mode can offer the rewarding feeling of providing care face-to-face that is lacking when delivering care via telehealth [29,51,53,55,84,92,93], while balancing work and personal life [22]. As such, the hybrid delivery mode could reduce burnout and help attract and retain health professionals in rural and regional areas that experience health workforce shortages. However, an unintended consequence of telehealth that was identified in our review is its potential to threaten rural practice [50,57], necessitating safeguards to ensure that rural and regional areas receive funding, training opportunities, and the on-the-ground workforce necessary to ensure the quality and safety of health care delivered to these communities. Furthermore, the fact that challenges were reported when delivering care to rural and regional patients due to a lack of context and local knowledge [44,100] underscores the need for clinicians to dedicate time to acquiring contextual knowledge and understanding patients' specific circumstances to build solid therapeutic relationships and ensure the relevance of the clinical advice provided.

One issue that was featured in our review is the ambiguity regarding clinical roles and responsibilities when multiple clinicians are involved in care delivery via telehealth [50,58,106], suggesting the need for policy or laws to govern care provision via telehealth when multiple clinicians are involved, especially across various states or territories. We will need to specify roles and clinical responsibilities and be ready to deal with any medicolegal issues that may arise in such situations. Finally, perhaps the most concerning finding is the potential of telehealth to widen the digital divide, increasing disparities in health care access among specific groups such as older patients or those with limited digital literacy skills [30,45]. Initiatives to improve patients' digital literacy [129,130] and provide patient information and guidelines on preparing for telehealth consultations [131,132] may support inclusive and equitable health care access via telehealth.

Strengths and Limitations

The strength of this review is the inclusion of diverse qualitative, quantitative, and mixed methods studies from 4 databases, ensuring a comprehensive account of the unintended consequences of telehealth for the Australian health system. Our review translated these findings on the unintended consequences of telehealth into the AHPF dimensions, outlining a framework for assessing telehealth impact across the various health system domains and beyond and enabling comparison and standardization. Moreover, by systematically reviewing and categorizing unintended consequences, our study provides insights into which consequences are most frequently reported in the literature. This quantification helps prioritize areas for further research and policy attention, especially because recent policy changes and government funding have signaled a push toward incorporating telehealth as a standard modality of health service provision [133]. Thus, evidence-based insights such as those generated by our review are crucial for offering detailed recommendations for practitioners and policy makers on improving health care delivery via telehealth. To our knowledge, this review is the first attempt to synthesize evidence on the unintended consequences of telehealth, not only in Australia but also internationally.

While a CIS has strong characteristics such as flexibility and interpretive freedom in the inclusion and synthesis of results, it can be a double-edged sword in cases such as our review where there are few to no published studies explicitly focused on the topic. Therefore, the evidence extracted from the included studies depends entirely on the authors' interpretation of what can be considered an unintended consequence of telehealth, possibly leading to the exclusion of some relevant evidence. In our review, this negative impact was mitigated by the significant number of studies included, ensuring comprehensiveness. Furthermore, to minimize the impact of subjectivity when extracting data, we used explicit definitions for what constituted an *unintended* versus an *intended* consequence to guide our data extraction.

While our review identified overarching patterns and themes that cut across different telehealth modalities and specialties, a limitation of this study is the broad categorization of telehealth modalities without distinguishing between videoconferencing and telephone-based telehealth or between specialty-related consequences and the generalization of videoconferencing technologies. This approach may obscure important nuances specific to different modalities, technologies, or specialties; for instance, videoconferencing may present unique challenges related to technology use and nonverbal communication that are not present in telephone-based telehealth. These unique challenges underscore the need for more detailed analysis. Similarly, certain specialties may encounter distinct unintended consequences based on their specific care requirements.

Moreover, we did not distinguish between the various models of videoconferencing telehealth, such as dedicated videoconferencing units with pan, tilt, and zoom capabilities; computer-based videoconferencing; and mobile phone–based videoconferencing apps. These technologies may have unique characteristics that could lead to specific unintended consequences. While our approach offers valuable insights into general trends across telehealth implementations, it may not capture these modality-, specialty-, or technology-specific nuances.

Future Directions

There is still a dearth of studies focusing solely on the unintended consequences of telehealth. One plausible explanation could be that telehealth uptake only expanded recently, and long-term outcomes often take time to emerge. In addition, designing a study to investigate unintended consequences using conventional methods can be challenging. To capture the broader impact of telehealth and grasp the complex dynamics that could lead to the emergence of these consequences, studies with designs suited to study the complex nature of telehealth implementation are needed. Recent work

XSL•FO
RenderX

proposed a complexity-informed and system-thinking–guided approach to investigate such unintended consequences, emphasizing the interconnectedness between the various actors involved in telehealth implementation [13]. Such theoretical frameworks can provide a holistic understanding of the situation compared to the linear reductionist approach that is often used to study and evaluate telehealth programs.

There is little to no research examining the impact of telehealth beyond the health system and how telehealth influences, and is influenced by, factors such as the determinants of health and the broader health context, such as financial and societal aspects. We found that only 6 (6%) of the 94 studies touched upon the issue, and it was a secondary observation briefly noted by participants [30,40,42-45]. One noteworthy finding of our review is the potential of telehealth to threaten rural practice [50,57], suggesting a need for further research that investigates the impact of telehealth on rural workforce sustainability and rural practice viability. This is particularly important given the complexities of rural health and rural practice [1,134] and the growing perception of telehealth as a favorable solution to health care inaccessibility in rural areas [50,135-137]. Furthermore, additional research is necessary to investigate the safety of health care provision via telehealth and to examine and compare the effectiveness of telehealth in terms of clinical outcomes across various settings, population groups, and specialties.

Further research is needed to address some of the aforementioned limitations regarding the broad categorization of telehealth modalities. Such research should focus on conducting comparative analyses between different telehealth modalities and exploring how unintended consequences manifest across various specialties. These focused studies could provide a more granular understanding of how the choice of the telehealth modality and the specific medical context influence the nature and prevalence of unintended consequences. This would further enhance our understanding of telehealth implementation and guide more tailored strategies for mitigating potential negative outcomes in specific telehealth contexts. Future research could also build on our findings by conducting a more granular analysis of unintended consequences associated with specific videoconferencing technologies. Such studies could explore how factors such as image quality, ease of use, mobility, and specific features (eg, screen-sharing and recording capabilities) influence the nature and prevalence of unintended consequences. Moreover, as telehealth technologies continue to evolve rapidly, ongoing research will be crucial to understanding the implications of emerging videoconferencing tools and platforms and exploring consequences related to augmented reality features, artificial intelligence–assisted communication, or integration with other medical devices and systems.

Conclusions

The unintended consequences of telehealth synthesized in our review provide a framework for understanding the full impact of telehealth across the health care system and beyond. Identifying these consequences offers various opportunities to more fully leverage the advantages of telehealth while mitigating any potential harm, ultimately sustaining its adoption beyond the COVID-19 pandemic for safe and high-quality care across different settings and population groups. Planning and implementing a telehealth project is a complex undertaking, and while it is not entirely possible to plan and anticipate every possible unintended outcome, the consequences presented in our review provide a road map for planning, implementing, and scaling up telehealth projects to realize their full potential.

Acknowledgments

The authors would like to thank and acknowledge Jeremy Cullis for assistance with the development of the search strategy. This work was supported by a Macquarie University Research Excellence Scholarship (20224427) awarded to SO for PhD research. DL was affiliated with the Australian Institute of Health Innovation at Macquarie University at the time of the study and is currently affiliated with the Daffodil Centre.

Authors' Contributions

SO conceived and designed the study, conducted database search, screened studies, extracted data, performed data analysis and synthesis, wrote the original draft, and reviewed and edited subsequent drafts. KC contributed to abstract screening, reviewed and edited subsequent drafts, and provided guidance and supervision throughout the study. LAE, DL, and JB reviewed and edited subsequent drafts and provided guidance and supervision throughout the study. All authors approved the final version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 31 KB - jmir_v26i1e57848_app1.docx]

Multimedia Appendix 2
List of selected characteristics of the included studies and extracted unintended consequences.

XSL•FO
RenderX

[XLSX File (Microsoft Excel File), 25 KB - jmir_v26i1e57848_app2.xlsx]

References

1. Thomas E, Gallagher R, Grace SL. Future-proofing cardiac rehabilitation: transitioning services to telehealth during COVID-19. Eur J Prev Cardiol 2021 Jul 10;28(7):e35-e36 [FREE Full text] [doi: 10.1177/2047487320922926] [Medline: 33611474]
2. Zhou X, Snoswell CL, Harding LE, Bambling M, Edirippulige S, Bai X, et al. The role of telehealth in reducing the mental health burden from COVID-19. Telemed J E Health 2020 Apr;26(4):377-379 [FREE Full text] [doi: 10.1089/tmj.2020.0068] [Medline: 32202977]
3. Snoswell CL, Caffery LJ, Haydon HM, Thomas EE, Smith AC. Telehealth uptake in general practice as a result of the coronavirus (COVID-19) pandemic. Aust Health Rev 2020 Aug 28;44(5):737-740. [doi: 10.1071/ah20183]
4. Snoswell CL, Caffery LJ, Hobson G, Taylor ML, Mendis R, Haydon HM, et al. Telehealth and coronavirus: Medicare Benefits Schedule (MBS) activity in Australia. Centre for Online Health, The University of Queensland. 2024 Apr 16. URL: https://coh.centre.uq.edu.au/telehealth-and-coronavirus-medicare-benefits-schedule-mbs-activity-australia [accessed 2024-08-12]
5. Zurynski Y, Ellis LA, Dammery G, Smith CL, Halim N, Ansell J, et al. The voice of Australian health consumers: the 2021 Australian health consumer sentiment survey. NHMRC Partnership Centre for Health System Sustainability, Consumers Health Forum of Australia. 2022. URL: https://healthsystemsustainability.com.au/wp-content/uploads/2022/03/PCHSS_ConsumerSentimentSurveyReport_FINAL3.pdf [accessed 2023-11-15]
6. Adams L, Lester S, Hoon E, van der Haak H, Proudman C, Hall C, et al. Patient satisfaction and acceptability with telehealth at specialist medical outpatient clinics during the COVID-19 pandemic in Australia. Intern Med J 2021 Jul;51(7):1028-1037 [FREE Full text] [doi: 10.1111/imj.15205] [Medline: 34213046]
7. Barton CJ, Ezzat AM, Merolli M, Williams CM, Haines T, Mehta N, et al. "It's second best": a mixed-methods evaluation of the experiences and attitudes of people with musculoskeletal pain towards physiotherapist delivered telehealth during the COVID-19 pandemic. Musculoskelet Sci Pract 2022 Apr;58:102500 [FREE Full text] [doi: 10.1016/j.msksp.2021.102500] [Medline: 35074694]
8. Fogarty A, Savopoulos P, Seymour M, Cox A, Williams K, Petrie S, et al. Providing therapeutic services to women and children who have experienced intimate partner violence during the COVID-19 pandemic: challenges and learnings. Child Abuse Negl 2022 Aug;130(Pt 1):105365 [FREE Full text] [doi: 10.1016/j.chiabu.2021.105365] [Medline: 34686360]
9. Savira F, Robinson S, Toll K, Spark L, Thomas E, Nesbitt J, et al. Consumer preferences for telehealth in Australia: a discrete choice experiment. PLoS One 2023 Mar 31;18(3):e0283821 [FREE Full text] [doi: 10.1371/journal.pone.0283821] [Medline: 37000814]
10. Thomas EE, Haydon HM, Mehrotra A, Caffery LJ, Snoswell CL, Banbury A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2022 May;28(4):301-308. [doi: 10.1177/1357633X20960638] [Medline: 32985380]
11. Omboni S, Padwal RS, Alessa T, Benczúr B, Green BB, Hubbard I, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 04;1:7-35 [FREE Full text] [doi: 10.20517/ch.2021.03] [Medline: 35233563]
12. Telehealth. National Institutes of Health ,National Institute of Biomedical Imaging and Bioengineering. URL: https://www.nibib.nih.gov/science-education/science-topics/telehealth#pid-1421 [accessed 2023-11-15]
13. Osman S, Churruca K, Ellis LA, Braithwaite J. Systems archetypes to investigate the unintended consequences of telehealth in rural Australia: a systems thinking approach to telehealth evaluation and policymaking. Int J Health Plann Manage 2024 Mar;39(2):204-219. [doi: 10.1002/hpm.3741] [Medline: 37974503]
14. Alami H, Gagnon MP, Fortin JP. Some multidimensional unintended consequences of telehealth utilization: a multi-project evaluation synthesis. Int J Health Policy Manag 2019 Jun 01;8(6):337-352 [FREE Full text] [doi: 10.15171/ijhpm.2019.12] [Medline: 31256566]
15. Baert P. Unintended consequences: a typology and examples. Int Sociol 1991 Jun;6(2):201-210. [doi: 10.1177/026858091006002006]
16. Bloomrosen M, Starren J, Lorenzi NM, Ash JS, Patel VL, Shortliffe EH. Anticipating and addressing the unintended consequences of health IT and policy: a report from the AMIA 2009 Health Policy Meeting. J Am Med Inform Assoc 2011;18(1):82-90 [FREE Full text] [doi: 10.1136/jamia.2010.007567] [Medline: 21169620]
17. Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J, et al. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. BMC Med Res Methodol 2006 Jul 26;6:35 [FREE Full text] [doi: 10.1186/1471-2288-6-35] [Medline: 16872487]
18. Dixit SK, Sambasivan M. A review of the Australian healthcare system: a policy perspective. SAGE Open Med 2018 Apr 12;6:2050312118769211 [FREE Full text] [doi: 10.1177/2050312118769211] [Medline: 29686869]
19. Jabeen S. Unintended outcomes evaluation approach: a plausible way to evaluate unintended outcomes of social development programmes. Eval Program Plann 2018 Jun;68:262-274. [doi: 10.1016/j.evalprogplan.2017.09.005] [Medline: 28965770]

XSL•FO
RenderX

20. Keatinge D, Rawlings K. Outcomes of a nurse-led telephone triage service in Australia. Int J Nurs Pract 2005 Feb;11(1):5-12. [doi: 10.1111/j.1440-172X.2005.00495.x] [Medline: 15610339]
21. Knight K, Kenny A, Endacott R. Assessing clinical urgency via telephone in rural Australia. Nurs Health Sci 2015 Jun;17(2):201-207. [doi: 10.1111/nhs.12161] [Medline: 25472637]
22. Baker FA, Tamplin J. Music therapy service provision via telehealth in response to covid-19 restrictions: a survey of Australian practitioners and consumers. Aust J Music Ther 2021 Jan 01;32(1):1-24 [FREE Full text]
23. Banbury A, Parkinson L, Gordon S, Wood D. Implementing a peer-support programme by group videoconferencing for isolated carers of people with dementia. J Telemed Telecare 2019 Oct;25(9):572-577. [doi: 10.1177/1357633X19873793] [Medline: 31631761]
24. Chapman JE, Ponsford J, Bagot KL, Cadilhac DA, Gardner B, Stolwyk RJ. The use of videoconferencing in clinical neuropsychology practice: a mixed methods evaluation of neuropsychologists' experiences and views. Aust Psychol 2021 Mar 11;55(6):618-633. [doi: 10.1111/ap.12471]
25. Gibbs V, Cai RY, Aldridge F, Wong M. Autism assessment via telehealth during the Covid 19 pandemic: experiences and perspectives of autistic adults, parents/carers and clinicians. Res Autism Spectr Disord 2021 Oct;88:101859 [FREE Full text] [doi: 10.1016/j.rasd.2021.101859] [Medline: 34512797]
26. Hines M, Bulkeley K, Dudley S, Cameron S, Lincoln M. Delivering quality allied health services to children with complex disability via telepractice: lessons learned from four case studies. J Dev Phys Disabil 2019 Feb 16;31(5):593-609. [doi: 10.1007/s10882-019-09662-8]
27. Lee JJ, Burbury K, Underhill C, Harris S, Shackleton K, McBurnie J, et al. Exploring Australian regional cancer patients' experiences of clinical trial participation via telehealth. J Telemed Telecare 2022 Aug;28(7):508-516. [doi: 10.1177/1357633X20950180] [Medline: 32811275]
28. McQueen M, Strauss P, Lin A, Freeman J, Hill N, Finlay-Jones A, et al. Mind the distance: experiences of non-face-to-face child and youth mental health services during COVID-19 social distancing restrictions in Western Australia. Aust Psychol 2022 Jun 09;57(5):301-314. [doi: 10.1080/00050067.2022.2078649]
29. Smith SJ, Smith AB, Kennett W, Vinod SK. Exploring cancer patients', caregivers', and clinicians' utilisation and experiences of telehealth services during COVID-19: a qualitative study. Patient Educ Couns 2022 Oct;105(10):3134-3142 [FREE Full text] [doi: 10.1016/j.pec.2022.06.001] [Medline: 35688719]
30. Smyth L, Roushdy S, Jeyasingham J, Whitbread J, O'Brien P, Lloyd C, et al. Clinician perspectives on rapid transition to telehealth during COVID-19 in Australia - a qualitative study. Aust Health Rev 2023 Feb;47(1):92-99. [doi: 10.1071/AH22037] [Medline: 36261136]
31. Venville A, O'Connor S, Roeschlein H, Ennals P, McLoughlan G, Thomas N. Mental health service user and worker experiences of psychosocial support via telehealth through the COVID-19 pandemic: qualitative study. JMIR Ment Health 2021 Aug 12;8(8):e29671 [FREE Full text] [doi: 10.2196/29671] [Medline: 34182461]
32. Chatterton ML, Marangu E, Clancy EM, Mackay M, Gu E, Moylan S, et al. Telehealth service delivery in an Australian regional mental health service during COVID-19: a mixed methods analysis. Int J Ment Health Syst 2022 Aug 19;16(1):43 [FREE Full text] [doi: 10.1186/s13033-022-00553-8] [Medline: 35986332]
33. Brown J, Doherty D, Claus AP, Gilbert K, Nielsen M. In a pandemic that limits contact, can videoconferencing enable interdisciplinary persistent pain services and what are the patient's perspectives? Arch Phys Med Rehabil 2022 Mar;103(3):418-423 [FREE Full text] [doi: 10.1016/j.apmr.2021.10.018] [Medline: 34762855]
34. Devadula S, Langbecker D, Vecchio P, Tesiram J, Meiklejohn J, Benham H. Tele-rheumatology to regional hospital outpatient clinics: patient perspectives on a new model of care. Telemed J E Health 2020 Jul;26(7):912-919. [doi: 10.1089/tmj.2019.0111] [Medline: 31682204]
35. Fairweather GC, Lincoln MA, Ramsden R. Speech-language pathology teletherapy in rural and remote educational settings: decreasing service inequities. Int J Speech Lang Pathol 2016 Dec;18(6):592-602. [doi: 10.3109/17549507.2016.1143973] [Medline: 27063692]
36. Van Dam PJ, Caney D, Turner RC, Griffin P, Dwyer M, Prior S. Telehealth experiences of vulnerable clients living in Tasmania. Aust J Rural Health 2022 Apr;30(2):188-196 [FREE Full text] [doi: 10.1111/ajr.12835] [Medline: 35076149]
37. Pitt R, Hill AJ, Theodoros D, Russell T. “I definitely think it’s a feasible and worthwhile option”: perspectives of speech-language pathologists providing online aphasia group therapy. Aphasiology 2018 Jun 08;32(9):1031-1053. [doi: 10.1080/02687038.2018.1482403]
38. Filbay S, Bennell KL, Morello R, Smith L, Hinman RS, Lawford BJ. Exploring experiences with telehealth-delivered allied healthcare services for people with permanent and significant disabilities funded through a national insurance scheme: a qualitative study examining challenges and suggestions to improve services. BMJ Open 2022 Sep 14;12(9):e065600 [FREE Full text] [doi: 10.1136/bmjopen-2022-065600] [Medline: 36104129]
39. Thrum M, Driscoll C, Keogh T. Investigating the satisfaction of clinicians and clients in a teleaudiology trial. J Hear Sci 2018 Dec 31;8(4):34-47. [doi: 10.17430/1003211]
40. Ayres L, Pelkowitz L, Simon P, Thompson SC. Necessity as the catalyst of change: exploring client and provider perspectives of accelerated implementation of telehealth by a regional Australian community service organisation during COVID-19

XSL•FO
RenderX

restrictions. Int J Environ Res Public Health 2021 Oct 30;18(21):11433 [FREE Full text] [doi: 10.3390/ijerph182111433] [Medline: 34769949]
41. Isautier JM, Copp T, Ayre J, Cvejic E, Meyerowitz-Katz G, Batcup C, et al. People's experiences and satisfaction with telehealth during the COVID-19 pandemic in Australia: cross-sectional survey study. J Med Internet Res 2020 Dec 10;22(12):e24531 [FREE Full text] [doi: 10.2196/24531] [Medline: 33156806]
42. Bradford NK, Young J, Armfield NR, Herbert A, Smith AC. Home telehealth and paediatric palliative care: clinician perceptions of what is stopping us? BMC Palliat Care 2014 Jun 16;13:29 [FREE Full text] [doi: 10.1186/1472-684X-13-29] [Medline: 24963287]
43. Smithson R, Roche E, Wicker C. Virtual models of chronic disease management: lessons from the experiences of virtual care during the COVID-19 response. Aust Health Rev 2021 Jun;45(3):311-316. [doi: 10.1071/AH20190] [Medline: 33583487]
44. Sutarsa IN, Kasim R, Steward B, Bain-Donohue S, Slimings C, Hall Dykgraaf S, et al. Implications of telehealth services for healthcare delivery and access in rural and remote communities: perceptions of patients and general practitioners. Aust J Prim Health 2022 Dec;28(6):522-528. [doi: 10.1071/PY21162] [Medline: 35918783]
45. White J, Byles J, Walley T. The qualitative experience of telehealth access and clinical encounters in Australian healthcare during COVID-19: implications for policy. Health Res Policy Syst 2022 Jan 15;20(1):9 [FREE Full text] [doi: 10.1186/s12961-021-00812-z] [Medline: 35033107]
46. Snoswell CL, Smith AC, Page M, Scuffham P, Caffery LJ. Quantifying the societal benefits from telehealth: productivity and reduced travel. Value Health Reg Issues 2022 Mar;28:61-66 [FREE Full text] [doi: 10.1016/j.vhri.2021.07.007] [Medline: 34800833]
47. Mackenzie L, Noble N, Proietto A, Jones J, Norton G, Palazzi K. Acceptability and feasibility of telehealth outpatient video-link consultations: a national cross-sectional survey of surgeons prior to the COVID-19 pandemic. Aust J Rural Health 2023 Apr;31(2):244-255 [FREE Full text] [doi: 10.1111/ajr.12940] [Medline: 36326168]
48. De Guzman KR, Snoswell CL, Giles CM, Smith AC, Haydon HM. GP perceptions of telehealth services in Australia: a qualitative study. BJGP Open 2022 Mar 22;6(1):BJGPO.2021.0182 [FREE Full text] [doi: 10.3399/BJGPO.2021.0182] [Medline: 34819294]
49. Thomas EE, de Camargo Catapan S, Haydon HM, Barras M, Snoswell C. Exploring factors of uneven use of telehealth among outpatient pharmacy clinics during COVID-19: a multi-method study. Res Social Adm Pharm 2022 Sep;18(9):3602-3611 [FREE Full text] [doi: 10.1016/j.sapharm.2022.02.003] [Medline: 35183460]
50. Wade VA, Eliott JA, Hiller JE. A qualitative study of ethical, medico-legal and clinical governance matters in Australian telehealth services. J Telemed Telecare 2012 Mar;18(2):109-114. [doi: 10.1258/jtt.2011.110808] [Medline: 22267306]
51. Aung E, Pasanen L, LeGautier R, McLachlan SA, Collins A, Philip J. The role of telehealth in oncology care: a qualitative exploration of patient and clinician perspectives. Eur J Cancer Care (Engl) 2022 Mar;31(2):e13563. [doi: 10.1111/ecc.13563] [Medline: 35150180]
52. Butt Z, Kirsten L, Beatty L, Kelly B, Dhillon H, Shaw JM. Barriers and enablers to implementing telehealth consultations in psycho-oncology. Psychooncology 2022 Aug;31(8):1365-1373 [FREE Full text] [doi: 10.1002/pon.5939] [Medline: 35460322]
53. Jhaveri D, Emeto TI, Alele FO, Strom A, Benham H. Use of telemedicine for rheumatology practice in Queensland, Australia: experiences before and during the COVID-19 pandemic. Intern Med J 2022 Oct;52(10):1685-1690. [doi: 10.1111/imj.15706] [Medline: 35112769]
54. McPherson K, Nahon I. Telehealth and the provision of pelvic health physiotherapy in regional, rural and remote Australia. Aust N Z Continence J 2021;27(3):66-70. [doi: 10.33235/anzcj.27.3.66-70]
55. Thomas EE, Taylor ML, Ward EC, Hwang R, Cook R, Ross JA, et al. Beyond forced telehealth adoption: a framework to sustain telehealth among allied health services. J Telemed Telecare 2024 Apr;30(3):559-569 [FREE Full text] [doi: 10.1177/1357633X221074499] [Medline: 35130099]
56. Collier A, Morgan DD, Swetenham K, To TH, Currow DC, Tieman JJ. Implementation of a pilot telehealth programme in community palliative care: a qualitative study of clinicians' perspectives. Palliat Med 2016 Apr;30(4):409-417. [doi: 10.1177/0269216315600113] [Medline: 26290500]
57. Wiadji E, Mackenzie L, Reeder P, Gani JS, Carroll R, Smith S, et al. Utilization of telehealth by surgeons during the COVID 19 pandemic in Australia: lessons learnt. ANZ J Surg 2021 Apr;91(4):507-514 [FREE Full text] [doi: 10.1111/ans.16693] [Medline: 33634949]
58. Olver IN, Selva-Nayagam S. Evaluation of a telemedicine link between Darwin and Adelaide to facilitate cancer management. Telemed J 2000;6(2):213-218. [doi: 10.1089/107830200415144] [Medline: 10957733]
59. Eapen V, Dadich A, Balachandran S, Dani A, Howari R, Sequeria AZ, et al. E-mental health in child psychiatry during COVID-19: an initial attitudinal study. Australas Psychiatry 2021 Oct;29(5):498-503. [doi: 10.1177/10398562211022748] [Medline: 34126787]
60. Fogarty A, Jones A, Seymour M, Savopoulos P, Evans K, O'Brien J, et al. The parenting skill development and education service: telehealth support for families at risk of child maltreatment during the COVID‐19 pandemic. Child Fam Soc Work 2021 Dec 02;27(3):392-404. [doi: 10.1111/cfs.12890]

XSL•FO
RenderX

61. Knott V, Habota T, Mallan K. Attitudes of Australian psychologists towards the delivery of therapy via video conferencing technology. Aust Psychol 2021 Mar 11;55(6):606-617. [doi: 10.1111/ap.12464]
62. Johnsson G, Bulkeley K. Practitioner and service user perspectives on the rapid shift to teletherapy for individuals on the autism spectrum as a result of COVID-19. Int J Environ Res Public Health 2021 Nov 11;18(22):11812 [FREE Full text] [doi: 10.3390/ijerph182211812] [Medline: 34831567]
63. Haydon HM, Caffery LJ, Snoswell CL, Thomas EE, Taylor M, Budge M, et al. Optimising specialist geriatric medicine services by telehealth. J Telemed Telecare 2021 Dec;27(10):674-679. [doi: 10.1177/1357633X211041859] [Medline: 34726995]
64. Toll K, Spark L, Neo B, Norman R, Elliott S, Wells L, et al. Consumer preferences, experiences, and attitudes towards telehealth: qualitative evidence from Australia. PLoS One 2022 Aug 31;17(8):e0273935 [FREE Full text] [doi: 10.1371/journal.pone.0273935] [Medline: 36044536]
65. Hardie RA, Thomas J, Li J, Pearce C, Georgiou A. General practice perspective on the use of telehealth during the COVID-19 pandemic in Australia using an Action Research approach: a qualitative study. BMJ Open 2022 Oct 27;12(10):e063179 [FREE Full text] [doi: 10.1136/bmjopen-2022-063179] [Medline: 36302573]
66. James S, Ashley C, Williams A, Desborough J, Mcinnes S, Calma K, et al. Experiences of Australian primary healthcare nurses in using telehealth during COVID-19: a qualitative study. BMJ Open 2021 Aug 06;11(8):e049095 [FREE Full text] [doi: 10.1136/bmjopen-2021-049095] [Medline: 34362804]
67. Sutherland R, Hodge A, Chan E, Silove N. Clinician experiences using standardised language assessments via telehealth. Int J Speech Lang Pathol 2021 Dec;23(6):569-578. [doi: 10.1080/17549507.2021.1903079] [Medline: 34000937]
68. Swales M, Theodoros D, Hill AJ, Russell T. Speech-language pathologists' perceptions of the use of telepractice in the delivery of services to people with Parkinson's disease: a national pilot survey. Int J Speech Lang Pathol 2020 Aug;22(4):387-398. [doi: 10.1080/17549507.2019.1650110] [Medline: 31416341]
69. Auret K, Pikora T, Pola M. Specialist haematology consultation services in regional Western Australia: evaluating a model combining telehealth and onsite clinics. Intern Med J 2022 Mar;52(3):451-457. [doi: 10.1111/imj.15127] [Medline: 33156576]
70. Fix L, Seymour JW, Sandhu MV, Melville C, Mazza D, Thompson TA. At-home telemedicine for medical abortion in Australia: a qualitative study of patient experiences and recommendations. BMJ Sex Reprod Health 2020 Jul;46(3):172-176 [FREE Full text] [doi: 10.1136/bmjsrh-2020-200612] [Medline: 32665231]
71. Javanparast S, Roeger L, Kwok Y, Reed RL. The experience of Australian general practice patients at high risk of poor health outcomes with telehealth during the COVID-19 pandemic: a qualitative study. BMC Fam Pract 2021 Apr 08;22(1):69 [FREE Full text] [doi: 10.1186/s12875-021-01408-w] [Medline: 33832422]
72. Kelly JT, Warner MM, Conley M, Reidlinger DP, Hoffmann T, Craig J, et al. Feasibility and acceptability of telehealth coaching to promote healthy eating in chronic kidney disease: a mixed-methods process evaluation. BMJ Open 2019 Jan 29;9(1):e024551 [FREE Full text] [doi: 10.1136/bmjopen-2018-024551] [Medline: 30700482]
73. Kohlhoff J, Cibralic S, Horswood D, Turnell A, Maiuolo M, Morgan S. Feasibility and acceptability of internet-delivered parent-child interaction therapy for rural Australian families: a qualitative investigation. Rural Remote Health 2020 Jan;20(1):5306 [FREE Full text] [doi: 10.22605/RRH5306] [Medline: 31917594]
74. McGrath P. Technology-based patient consultations: research findings from haematology patients in regional, rural and remote queensland. Patient 2015 Apr;8(2):199-206. [doi: 10.1007/s40271-014-0074-z] [Medline: 25005862]
75. O'Brian S, Smith K, Onslow M. Webcam delivery of the Lidcombe program for early stuttering: a phase I clinical trial. J Speech Lang Hear Res 2014 Jun 01;57(3):825-830. [doi: 10.1044/2014_JSLHR-S-13-0094] [Medline: 24686834]
76. Owen N. Feasibility and acceptability of using telehealth for early intervention parent counselling. Adv Ment Health 2019 Oct 18;18(1):39-49. [doi: 10.1080/18387357.2019.1679026]
77. Raatz MK, Ward EC, Marshall J. Telepractice for the delivery of pediatric feeding services: a survey of practice investigating clinician perceptions and current service models in Australia. Dysphagia 2020 Apr;35(2):378-388. [doi: 10.1007/s00455-019-10042-9] [Medline: 31363846]
78. Jhaveri D, Larkins S, Kelly J, Sabesan S. Remote chemotherapy supervision model for rural cancer care: perspectives of health professionals. Eur J Cancer Care (Engl) 2016 Jan;25(1):93-98. [doi: 10.1111/ecc.12309] [Medline: 25871852]
79. Mooi JK, Whop LJ, Valery PC, Sabesan SS. Teleoncology for indigenous patients: the responses of patients and health workers. Aust J Rural Health 2012 Oct;20(5):265-269. [doi: 10.1111/j.1440-1584.2012.01302.x] [Medline: 22998201]
80. Callisaya ML, Lee AH, Khushu A. Rapid implementation of telehealth in geriatric outpatient clinics due to COVID-19. Intern Med J 2021 Jul;51(7):1151-1155 [FREE Full text] [doi: 10.1111/imj.15306] [Medline: 34143563]
81. Cartledge S, Rawstorn JC, Tran M, Ryan P, Howden EJ, Jackson A. Telehealth is here to stay but not without challenges: a consultation of cardiac rehabilitation clinicians during COVID-19 in Victoria, Australia. Eur J Cardiovasc Nurs 2022 Aug 29;21(6):548-558 [FREE Full text] [doi: 10.1093/eurjcn/zvab118] [Medline: 34935940]
82. Hill AJ, Miller LE. A survey of the clinical use of telehealth in speech-language pathology across Australia. J Clin Pract Speech Lang Pathol 2012;14(3):110-117 [FREE Full text]

XSL•FO
RenderX

83. Newman L, Bidargaddi N, Schrader G. Service providers' experiences of using a telehealth network 12 months after digitisation of a large Australian rural mental health service. Int J Med Inform 2016 Oct;94:8-20. [doi: 10.1016/j.ijmedinf.2016.05.006] [Medline: 27573307]
84. Pasanen L, Le Gautier R, Wong A, Wawryk O, Collins A, Schwetlik S, et al. Telehealth in outpatient delivery of palliative care: a qualitative study of patient and physician views. Palliat Support Care 2022 Jul 12;21(6):980-987. [doi: 10.1017/s1478951522000670]
85. Phillips D, Matheson L, Pain T, Kingston GA. Evaluation of an occupational therapy led paediatric burns telehealth review clinic: exploring the experience of family/carers and clinicians. Rural Remote Health 2022 Feb;22(1):6887 [FREE Full text] [doi: 10.22605/RRH6887] [Medline: 35138867]
86. Roberts S, Spain B, Hicks C, London J, Tay S. Telemedicine in the Northern territory: an assessment of patient perceptions in the preoperative anaesthetic clinic. Aust J Rural Health 2015 Jun;23(3):136-141. [doi: 10.1111/ajr.12140] [Medline: 25615954]
87. Shiner CT, Vratsistas-Curto A, Bramah V, McDonell K, Mahoney AE, Sweeney S, et al. Assessing unmet rehabilitation needs and the feasibility of a telehealth rehabilitation consultation service for road trauma survivors recently discharged from hospital. Disabil Rehabil 2022 Jul;44(15):3795-3804. [doi: 10.1080/09638288.2021.1887377] [Medline: 33605180]
88. Snoswell CL, Caffery LJ, Haydon HM, Wickramasinghe SI, Crumblin K, Smith AC. A cost-consequence analysis comparing patient travel, outreach, and telehealth clinic models for a specialist diabetes service to Indigenous people in Queensland. J Telemed Telecare 2019 Oct;25(9):537-544. [doi: 10.1177/1357633X19873239] [Medline: 31631756]
89. Wade VA, Karnon J, Eliott JA, Hiller JE. Home videophones improve direct observation in tuberculosis treatment: a mixed methods evaluation. PLoS One 2012;7(11):e50155 [FREE Full text] [doi: 10.1371/journal.pone.0050155] [Medline: 23226243]
90. Wiadji E, Mackenzie L, Reeder P, Gani JS, Ahmadi S, Carroll R, et al. Patient perceptions of surgical telehealth consultations during the COVID 19 pandemic in Australia: lessons for future implementation. ANZ J Surg 2021 Sep 21;91(9):1662-1667 [FREE Full text] [doi: 10.1111/ans.17020] [Medline: 34152673]
91. Bassilios B, Pirkis J, King K, Fletcher J, Blashki G, Burgess P. Evaluation of an Australian primary care telephone cognitive behavioural therapy pilot. Aust J Prim Health 2014;20(1):62-73. [doi: 10.1071/PY12111] [Medline: 23217591]
92. Haines KJ, Sawyer A, McKinnon C, Donovan A, Michael C, Cimoli C, et al. Barriers and enablers to telehealth use by physiotherapists during the COVID-19 pandemic. Physiotherapy 2023 Mar;118:12-19 [FREE Full text] [doi: 10.1016/j.physio.2022.09.003] [Medline: 36308980]
93. Campbell J, Theodoros D, Russell T, Hartley N, Gillespie N. Role and process change and satisfaction with an educational and developmental psychologist telehealth service for rural children. Educ Dev Psychol 2021 May 03;38(1):143-157. [doi: 10.1080/20590776.2021.1915097]
94. Erickson S, Bridgman K, Furlong L, Stark H. Speech-language pathologist perspectives of the implementation of telepractice-delivered stuttering treatment for school-age children. Lang Speech Hear Serv Sch 2022 Jan 05;53(1):30-43. [doi: 10.1044/2021_LSHSS-20-00167] [Medline: 34752153]
95. Hines M, Lincoln M, Ramsden R, Martinovich J, Fairweather C. Speech pathologists' perspectives on transitioning to telepractice: what factors promote acceptance? J Telemed Telecare 2015 Dec;21(8):469-473. [doi: 10.1177/1357633X15604555] [Medline: 26377120]
96. Sutherland R, Hodge A, Chan E, Silove N. Barriers and facilitators: clinicians' opinions and experiences of telehealth before and after their use of a telehealth platform for child language assessment. Int J Lang Commun Disord 2021 Nov;56(6):1263-1277. [doi: 10.1111/1460-6984.12666] [Medline: 34455670]
97. Shannon MM, Callum SM, Callisaya ML. Uncovering healthcare staff attitudes to the rapid deployment of telehealth in Victoria, 2020-2021: a 12-month telehealth experience. Intern Med J 2023 Jun;53(6):1018-1026. [doi: 10.1111/imj.15750] [Medline: 35289486]
98. Mclaughlin EJ, Ellett LC, Readman E, Mooney S. Telehealth for gynaecology outpatients during the COVID-19 pandemic: patient and clinician experiences. Aust N Z J Obstet Gynaecol 2022 Aug;62(4):553-559 [FREE Full text] [doi: 10.1111/ajo.13510] [Medline: 35322405]
99. Fairweather GC, Lincoln MA, Ramsden R. Speech-language pathology telehealth in rural and remote schools: the experience of school executive and therapy assistants. Rural Remote Health 2017;17(3):4225 [FREE Full text] [doi: 10.22605/RRH4225] [Medline: 28918643]
100. Bagot KL, Cadilhac DA, Vu M, Moss K, Bladin CF, VST Collaborators. Telemedicine in the acute health setting: a disruptive innovation for specialists (an example from stroke). J Telemed Telecare 2015 Dec;21(8):443-448. [doi: 10.1177/1357633X15610722] [Medline: 26556058]
101. Johnsson G, Kerslake R, Crook S. Delivering allied health services to regional and remote participants on the autism spectrum via video-conferencing technology: lessons learned. Rural Remote Health 2019 Sep;19(3):5358 [FREE Full text] [doi: 10.22605/RRH5358] [Medline: 31474113]
102. Buist A, Coman G, Silvas A, Burrows G. An evaluation of the telepsychiatry programme in Victoria, Australia. J Telemed Telecare 2000;6(4):216-221. [doi: 10.1258/1357633001935383] [Medline: 11027122]

XSL•FO
RenderX

103. Rainsford S, Hall Dykgraaf S, Phillips C. Effectiveness of telehealth palliative care Needs Rounds in rural residential aged care during the COVID-19 pandemic: a hybrid effectiveness-implementation study. Aust J Rural Health 2022 Feb;30(1):108-114 [FREE Full text] [doi: 10.1111/ajr.12789] [Medline: 34757687]
104. Martin-Khan M, Fatehi F, Kezilas M, Lucas K, Gray LC, Smith AC. Establishing a centralised telehealth service increases telehealth activity at a tertiary hospital. BMC Health Serv Res 2015 Dec 03;15:534 [FREE Full text] [doi: 10.1186/s12913-015-1180-x] [Medline: 26630965]
105. Bagot KL, Bladin CF, Vu M, Kim J, Hand PJ, Campbell B, et al. Exploring the benefits of a stroke telemedicine programme: an organisational and societal perspective. J Telemed Telecare 2016 Dec;22(8):489-494. [doi: 10.1177/1357633X16673695] [Medline: 27799453]
106. Zilliacus E, Meiser B, Lobb E, Barlow-Stewart K, Tucker K. A balancing act--telehealth cancer genetics and practitioners' experiences of a triadic consultation. J Genet Couns 2009 Dec;18(6):598-605. [doi: 10.1007/s10897-009-9247-7] [Medline: 19798555]
107. Harrison JD, Durcinoska I, Butow PN, White K, Solomon MJ, Young JM. Localized versus centralized nurse-delivered telephone services for people in follow up for cancer: opinions of cancer clinicians. Asia Pac J Clin Oncol 2014 Jun;10(2):175-182. [doi: 10.1111/ajco.12082] [Medline: 23656616]
108. Robinson S. Video-conferencing: under-used by rural general practitioners. Aust Health Rev 2002;25(6):131-135. [doi: 10.1071/ah020131a] [Medline: 12536872]
109. Guilfoyle C, Wootton R, Hassall S, Offer J, Warren M, Smith D, et al. User satisfaction with allied health services delivered to residential facilities via videoconferencing. J Telemed Telecare 2003;9 Suppl 1:S52-S54. [doi: 10.1258/135763303322196349] [Medline: 12952723]
110. Sharpe K, Elcock M, Aitken P, Furyk J. The use of telehealth to assist remote hospital resuscitation and aeromedical retrieval tasking: a 12-month case review. J Telemed Telecare 2012 Jul;18(5):260-266. [doi: 10.1258/jtt.2011.110906] [Medline: 22302764]
111. Johnston K, Smith D, Preston R, Evans R, Carlisle K, Lengren J, et al. "From the technology came the idea": safe implementation and operation of a high quality teleradiology model increasing access to timely breast cancer assessment services for women in rural Australia. BMC Health Serv Res 2020 Nov 30;20(1):1103 [FREE Full text] [doi: 10.1186/s12913-020-05922-y] [Medline: 33256724]
112. Neuhaus M, Langbecker D, Caffery LJ, Taylor M, Garner L, Williams G, et al. Telementoring for hepatitis C treatment in correctional facilities. J Telemed Telecare 2018 Dec;24(10):690-696. [doi: 10.1177/1357633X18795361] [Medline: 30343659]
113. Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B, et al. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. J Health Serv Res Policy 2007 Jan;12(1):42-47. [doi: 10.1258/135581907779497486] [Medline: 17244397]
114. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
115. Carroll C, Booth A, Leaviss J, Rick J. "Best fit" framework synthesis: refining the method. BMC Med Res Methodol 2013 Mar 13;13:37 [FREE Full text] [doi: 10.1186/1471-2288-13-37] [Medline: 23497061]
116. Carroll C, Booth A, Cooper K. A worked example of "best fit" framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. BMC Med Res Methodol 2011 Mar 16;11:29 [FREE Full text] [doi: 10.1186/1471-2288-11-29] [Medline: 21410933]
117. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: The Qualitative Researcher's Companion. Thousand Oaks, CA: SAGE Publications; 2002.
118. Dixon-Woods M. Using framework-based synthesis for conducting reviews of qualitative studies. BMC Med 2011 Apr 14;9:39 [FREE Full text] [doi: 10.1186/1741-7015-9-39] [Medline: 21492447]
119. Australia's health performance framework. Australian Institute of Health and Welfare, Australian Government. URL: https://www.aihw.gov.au/reports-data/australias-health-performance/australias-health-performance-framework [accessed 2023-10-29]
120. The Australian health performance framework. The National Health Information and Performance Principal Committee. 2017 Sep. URL: https://www.aihw.gov.au/getmedia/1c95574c-ac07-4126-8b7c-31eb29d9b381/oos318_attachment-1.pdf.aspx [accessed 2024-08-12]
121. Continuity and coordination of care: a practice brief to support implementation of the WHO Framework on integrated people-centred health services. World Health Organization. 2018 Nov 07. URL: https://www.who.int/publications/i/item/9789241514033 [accessed 2024-08-12]
122. Deeny S, Gardner T, Al-Zaidy S, Barker I, Steventon A. Reducing hospital admissions by improving continuity of care in general practice. The Health Foundation. 2017 Feb. URL: https://www.health.org.uk/publications/reducing-hospital-admissions-by-improving-continuity-of-care-in-general-practice# [accessed 2024-08-12]
123. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]

XSL•FO
RenderX

124. Fernandes LG, Devan H, Fioratti I, Kamper SJ, Williams CM, Saragiotto BT. At my own pace, space, and place: a systematic review of qualitative studies of enablers and barriers to telehealth interventions for people with chronic pain. Pain 2022 Feb 01;163(2):e165-e181. [doi: 10.1097/j.pain.0000000000002364] [Medline: 34433776]
125. Appireddy R, Khan S, Leaver C, Martin C, Jin A, Durafourt BA, et al. Home virtual visits for outpatient follow-up stroke care: cross-sectional study. J Med Internet Res 2019 Oct 07;21(10):e13734 [FREE Full text] [doi: 10.2196/13734] [Medline: 31593536]
126. Hilty DM, Armstrong CM, Smout SA, Crawford A, Maheu MM, Drude KP, et al. Findings and guidelines on provider technology, fatigue, and well-being: scoping review. J Med Internet Res 2022 May 25;24(5):e34451 [FREE Full text] [doi: 10.2196/34451] [Medline: 35612880]
127. Eberly LA, Khatana SA, Nathan AS, Snider C, Julien HM, Deleener ME, et al. Telemedicine outpatient cardiovascular care during the COVID-19 pandemic: bridging or opening the digital divide? Circulation 2020 Aug 04;142(5):510-512 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.048185] [Medline: 32510987]
128. Telehealth consultations with patients. Medical Board Ahpra. URL: https://www.medicalboard.gov.au/Codes-Guidelines-Policies/Telehealth-consultations-with-patients.aspx [accessed 2024-01-16]
129. Digital health. Good Things Foundation Australia. URL: https://goodthingsaustralia.org/our-programs/digital-health/ [accessed 2024-08-04]
130. About Canberra health literacy hub. Canberra Health Literacy Hub. URL: https://cbrhl.org.au/about-canberra-health-literacy/ [accessed 2024-01-15]
131. How to prepare for a telehealth consultation. Cancer Council. 2020 Apr 30. URL: https://www.cancer.org.au/blog/how-to-prepare-for-a-telehealth-consultation [accessed 2024-01-15]
132. Preparing for a virtual appointment. Agency for Clinical Innovation. URL: https://aci.health.nsw.gov.au/__data/assets/pdf_file/0008/654857/Preparing-for-a-virtual-appointment.pdf [accessed 2024-01-15]
133. MBS telehealth services - from 1 July 2024. Department of Health and Aged Care, Australian Government. URL: https://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Factsheet-Telehealth-Updates-April%202023 [accessed 2024-07-02]
134. Osman S, Churruca K, Ellis LA, Braithwaite J. Beyond the planned and expected: the unintended consequences of telehealth in rural and remote Australia through a complexity lens. Med J Aust 2024 Jun 03;220(10):496-498. [doi: 10.5694/mja2.52294] [Medline: 38703008]
135. Moffatt JJ, Eley DS. The reported benefits of telehealth for rural Australians. Aust Health Rev 2010 Aug;34(3):276-281. [doi: 10.1071/AH09794] [Medline: 20797357]
136. Gagnon MP, Duplantie J, Fortin JP, Landry R. Exploring the effects of telehealth on medical human resources supply: a qualitative case study in remote regions. BMC Health Serv Res 2007 Jan 11;7(1):6 [FREE Full text] [doi: 10.1186/1472-6963-7-6] [Medline: 17217534]
137. Calleja Z, Job J, Jackson C. Offsite primary care providers using telehealth to support a sustainable workforce in rural and remote general practice: a rapid review of the literature. Aust J Rural Health 2023 Feb;31(1):5-18. [doi: 10.1111/ajr.12920] [Medline: 36037328]

Abbreviations

AHPF: Australian Health Performance Framework
CIS: critical interpretive synthesis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

Edited by G Eysenbach; submitted 27.02.24; peer-reviewed by V Minichiello, M Mars, B Wolfe-Piggott; comments to author 11.06.24; revised version received 01.07.24; accepted 16.07.24; published 27.08.24.

Please cite as:
Osman S, Churruca K, Ellis LA, Luo D, Braithwaite J
The Unintended Consequences of Telehealth in Australia: Critical Interpretive Synthesis
J Med Internet Res 2024;26:e57848
URL: https://www.jmir.org/2024/1/e57848
doi:10.2196/57848
PMID:39190446

©Sagda Osman, Kate Churruca, Louise A Ellis, Dan Luo, Jeffrey Braithwaite. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.08.2024. This is an open-access article distributed under the terms of the Creative

XSL•FO
RenderX

Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Deceptively Simple yet Profoundly Impactful: Text Messaging Interventions to Support Health

Brian Suffoletto[1], MS, MD

Department of Emergency Medicine, Stanford University, Palo Alto, CA, United States

Corresponding Author:
Brian Suffoletto, MS, MD
Department of Emergency Medicine
Stanford University
3145 Porter Drive Wing B
Palo Alto, CA, 94035
United States
Phone: 1 412 901 6892
Email: suffbp@stanford.edu

Abstract

This paper examines the use of text message (SMS) interventions for health-related behavioral support. It first outlines the historical progress in SMS intervention research publications and the variety of funds from US government agencies. A narrative review follows, highlighting the effectiveness of SMS interventions in key health areas, such as physical activity, diet and weight loss, mental health, and substance use, based on published meta-analyses. It then outlines advantages of text messaging compared to other digital modalities, including the real-time capability to collect information and deliver microdoses of intervention support. Crucial design elements are proposed to optimize effectiveness and longitudinal engagement across communication strategies, psychological foundations, and behavior change tactics. We then discuss advanced functionalities, such as the potential for generative artificial intelligence to improve user interaction. Finally, major challenges to implementation are highlighted, including the absence of a dedicated commercial platform, privacy and security concerns with SMS technology, difficulties integrating SMS interventions with medical informatics systems, and concerns about user engagement. Proposed solutions aim to facilitate the broader application and effectiveness of SMS interventions. Our hope is that these insights can assist researchers and practitioners in using SMS interventions to improve health outcomes and reducing disparities.

(J Med Internet Res 2024;26:e58726) doi:10.2196/58726

KEYWORDS

SMS intervention; behavior; intervention; review; text messaging; SMS; interventions; behaviors; behaviour; behaviours; effectiveness; development; impact; narrative review; physical activity; diet; weight loss; mental health; substance use; meta-analysis; chatbot; chatbots; large language model; LLM; large language models; mobile phone

Introduction

SMS text messaging interventions have emerged as a significant tool in the landscape of digital health, bridging the expanding gap between health care provider constraints and patient needs. While deceptively simple in design, SMS interventions are grounded in a sophisticated blend of communication strategies, psychological insights, and technological advances, all of which have been shown to positively shape health behaviors. This paper aims to provide a comprehensive overview of the SMS interventions' trajectory, distilling the evidence of their impact, discussing the critical elements of successful interventions, and exploring future directions for research and application.

Our examination begins with a retrospection of the emergence and evolution of SMS interventions, starting from the early 2000s when digital communication first showed promise for supporting health behavior change. Our paper highlights the broad scope of funding from US government agencies and the expanding research in this field. A narrative review then spotlights the success of SMS interventions across critical health domains—physical activity, diet and weight loss, mental health, and substance use—drawing from recent meta-analyses. The paper explores the specific aspects of SMS interventions that contribute to their effectiveness with a focus on the real-time capability of SMS interventions to collect information and deliver microdoses of intervention support.

The paper then discusses crucial SMS intervention design considerations across categories of communication strategies,

XSL•FO
RenderX

psychological underpinnings, behavior change techniques (BCTs), and enhanced functionalities. It also identifies key unanswered questions that future research needs to address. It then discusses the evolution of SMS interventions from rule-based systems (eg, branching algorithms) to generative artificial intelligence systems. It focuses on recent advancements in large language models (LLMs), which present new opportunities for enhancing the interactivity and personalization of SMS interventions, promising to elevate the user experience and engagement further.

Finally, the paper discusses existing SMS intervention implementation challenges, including transitioning SMS interventions from research prototypes to widely implemented health interventions, the absence of a centralized platform for dissemination akin to app stores, legal and ethical concerns around data security and patient privacy, and the intricacies of integrating SMS interventions with existing electronic health record systems. The paper then offers suggestions for overcoming such barriers by combining innovation in technology development, policy formulation, and practice implementation.

Our goal was to equip researchers, health care practitioners, and policy makers with knowledge on how to effectively use SMS interventions, aiming to improve health outcomes and reduce health care access and quality disparities. By offering a thorough analysis and forward-looking perspective, this paper contributes to the broader conversation on SMS interventions' role in health promotion and disease prevention.

Growth of SMS Intervention Scholarship and Funding

To identify peer-reviewed publications on SMS interventions across key health-related categories, we conducted a comprehensive search on PubMed. This search included categories of physical activity, diet and weight loss, mental health, substance use, disease management, and medication adherence. Multimedia Appendix 1 lists the PubMed search terms. As shown in Figure 1, the history of SMS interventions for behavioral support starts in the early 2000s and appears to peak in 2021, with >250 peer-reviewed papers published that year. The data show growth across key health-related categories. The most substantial growth was observed in the "disease management and medication adherence" category, underscoring the perceived efficacy and value of SMS interventions in addressing daily health-related behaviors. While the number of studies in some categories, such as physical activity and diet and weight loss, has remained relatively low, there has been a consistent increase in research on mental health, substance use, reproductive and maternal health, and appointment attendance using SMS interventions. Conversely, the relatively slower increase in research on physical activity and diet and weight loss may reflect the complexities of influencing lifestyle behaviors through digital means alone or the saturation of research in these areas.

To identify US national funding trends for SMS intervention projects, we queried the National Institutes of Health Reporter website using the search term "text message AND intervention." We report the diversity of SMS intervention projects from 2007 to 2023 from the top 16 divisions of NHI. The National Cancer Institute (NCI) leads with 312 projects, followed by the National Institute of Mental Health with 274 projects and the National Institute on Drug Abuse with 239 projects. Other notable institutes include the National Institute of Diabetes and Digestive and Kidney Diseases; the National Heart, Lung, and Blood Institute; the Eunice Kennedy Shriver National Institute of Child Health and Human Development; and the National Institute on Minority Health and Health Disparities, all administering >170 projects each.

The major themes of NCI funding for SMS interventions to date include smoking cessation for specific groups such as pregnant women, screening for cervical and colorectal cancers in underserved areas, improvement of adherence to cancer treatments, and cancer prevention initiatives such as human papillomavirus vaccination. The National Institute of Mental Health has focused on SMS interventions for postpsychiatric emergency care, HIV-related health behaviors, and mental health condition monitoring. The National Institute on Drug Abuse has targeted funding at managing substance use disorders, emphasizing treatment engagement and recovery support. The National Institute of Diabetes and Digestive and Kidney Diseases funded SMS interventions aimed at diabetes management, obesity prevention, and kidney disease care. The National Heart, Lung, and Blood Institute invested in SMS interventions for medication adherence in sickle cell disease, hypertension and asthma management, and lifestyle modifications to reduce cardiovascular risk. Finally, the National Institute on Minority Health and Health Disparities supported a wide array of SMS interventions addressing health disparities, including SMS interventions for enhancing cancer screening attendance, obesity prevention in Hispanic women, health optimization in taxi drivers, workplace inclusion, physical activity in African American cancer survivors, sugary beverage reduction, maximizing COVID-19 vaccination uptake, and smoking cessation among immigrant populations.

XSL•FO
RenderX

Figure 1. Number of published peer-reviewed studies of SMS interventions by health-related categories and year.

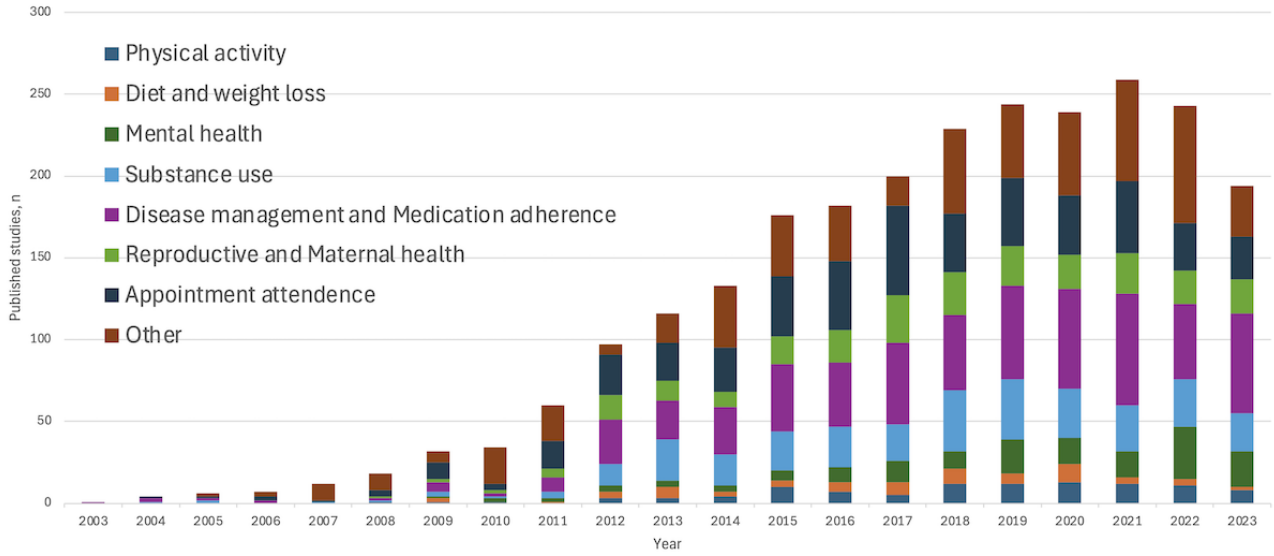

Narrative Review of SMS Interventions Across Key Health-Related Categories

Overview

In this section, we review recent meta-analyses on the impact of SMS interventions within key health areas, pinpointing components that bolster intervention efficacy. Our analysis specifically targeted peer-reviewed research published after the foundational reviews by Head et al [1] in 2013, Hall et al [2] in 2015, and Armanasco et al [3] in 2017. Head et al [1] analyzed 19 studies involving 5958 participants and determined that SMS interventions across various health promotion contexts achieved an average effect size of 0.33 (95% CI 0.27-0.39). A broader examination by Hall et al [2] found a consistent pattern of positive impacts of SMS interventions on diabetes self-management, physical activity, weight loss, smoking cessation, and medication adherence. The assessment by Armanasco et al [3] of the sustained impact of 7 SMS interventions during the postintervention period revealed a modest yet meaningful overall maintenance effect of Cohen *d*=0.17 (95% CI 0.03-0.31), underscoring the enduring benefits of SMS interventions beyond active engagement periods.

Physical Activity

In 2020, Smith et al [4] conducted a meta-analysis of 10 SMS interventions aimed at enhancing physical activity and found a significant increase in objectively measured steps per day after intervention among participants who received SMS interventions, with an effect size of Cohen *d*=0.38 (95% CI 0.19-0.58). The analysis indicated that interventions incorporating multiple components and tailored messages and those targeted at medical populations tended to yield larger, although not statistically significant, effect sizes. In addition, the study revealed that SMS interventions featuring personalized messages were potentially more effective in promoting physical activity than those using generic text messages (Cohen *d*=0.39 vs 0.25). Similarly, interventions that combined text messaging with other intervention components were more effective than text message–only interventions (Cohen *d*=0.37 vs 0.27).

Diet and Weight Loss

In 2015, Siopis et al [5] conducted a meta-analysis of 6 studies of SMS interventions targeting weight loss, finding that participants in SMS intervention programs lost significantly more weight (–2.17 kg, 95% CI –3.41 to –0.93) than those in control groups over 8 to 12 weeks. The design of these SMS interventions varied considerably, with intervention lengths ranging from 1 to 24 months and text messaging frequencies varying from daily to biweekly. In 2020, Skinner et al [6] analyzed 12 weight loss interventions involving 1977 participants and observed a mean weight loss difference of –2.28 kg (95% CI –3.17 to –1.36) favoring the SMS intervention groups. Among studies focusing on maintaining weight loss involving 3728 participants, the mean difference in weight change was –0.68 kg (95% CI –1.31 to –0.05) favoring SMS interventions. During postintervention follow-ups, the effect was –0.57 kg (95% CI –1.67 to 0.53), suggesting sustained SMS intervention benefits. These variations in weight loss outcomes were not clearly associated with specific intervention characteristics such as duration, text messaging frequency, theoretical basis, interactivity, or personalization. In 2020, Partridge et al [7] reviewed 8 studies targeting adolescents (767 participants), finding that SMS interventions led to greater reductions in BMI ranging from 1.3% to 4.5% compared to controls by the end of follow-up. These studies exhibited diversity in SMS intervention intensity, duration and content and behavioral strategies used in SMS interventions, underscoring the heterogeneity within intervention research on weight loss and management.

Mental Health

Within the domain of mental health, most SMS interventions tested focused on depression. In 2019, Senanayake et al [8] conducted a meta-analysis of 7 trials (845 patients; adults: n=664, 78.6%; adolescents: n=181, 21.4%), finding that the standardized mean reduction in depression scores for SMS interventions was 0.23 (95% CI –0.02- to 0.48). In 2020, Cox et al [9] examined 7 SMS interventions (1918 participants) and found, through pooled analysis, nonsignificant differences in depressive symptom scores between the SMS intervention and

XSL•FO
RenderX

control groups (mean –0.27, 95% CI –0.54 to 0.01). Reductions in depression were greater when text message content targeted mental well-being and mood improvement and used cognitive behavioral therapy and when the message frequency was ≥2 times per week. Evidence for other mental health issues remains less robustly studied. In 2020, D'Arcey et al [10] performed a systematic review of 15 studies focused on patients with schizophrenia. Although effect sizes were not presented, they commented that most SMS interventions demonstrated positive effects on dimensions of engagement such as medication adherence, clinic attendance, and therapeutic alliance.

Substance Use

Within the domain of substance use, the most studied behavioral focus for SMS interventions is smoking cessation. In 2016, Scott-Sheldon et al [11] conducted a meta-analysis evaluating the effectiveness of SMS interventions on smoking cessation, analyzing 22 interventions with 15,593 participants from 10 countries. The analysis revealed that individuals who received SMS interventions were significantly more likely to quit smoking compared to control groups, as evidenced by various smoking abstinence measures, including 7-day point prevalence (odds ratio [OR] 1.38, 95% CI 1.22-1.55) and continuous abstinence (OR 1.63, 95% CI 1.19-2.24). In 2023, Zhou et al [12] focused on young adults aged 16 to 30 years and found, across 7 studies, that SMS interventions were effective in promoting 7-day point prevalence abstinence with a risk ratio of 1.83 (95% CI 1.34-2.48).

Following smoking, alcohol misuse is the next most common substance use disorder addressed by SMS interventions. In 2021, Bendtsen et al [13] performed a meta-analysis of 26 trials (7376 participants) examining SMS interventions and found a significant decrease in alcohol consumption among intervention groups compared to controls (Hedges g=0.17, 95% CI 0.10-0.23). Subgroup analyses indicated more substantial effects for SMS interventions that used personalized messages (Hedges g=0.25), had a high frequency of messaging (Hedges g=0.24), and were grounded in behavioral theories (Hedges g=0.22).

Medication Adherence

In 2016, Thakkar et al [14] conducted a meta-analysis of 16 trials testing SMS interventions on medication adherence in chronic disease involving 2742 participants. The findings indicate that SMS interventions significantly enhanced medication adherence (OR 2.11, 95% CI 1.52-2.93). Notably, the effectiveness of SMS interventions did not vary based on the duration of the intervention; the type of chronic disease targeted; or specific features of the text messages, such as personalization, 2-way communication, or the frequency of text messages sent.

Several subpopulations have been the focus of meta-analyses for medication adherence SMS interventions.

For HIV in adolescents, a systematic review by Mehra et al [15] in 2021 included 7 studies with a total of 987 participants, where sample sizes ranged from 14 to 332. Of these 7 studies, 5 (71%) demonstrated a positive impact of SMS interventions in improving adherence. The pooled mean difference between the intervention and the control groups was 0.05 (95% CI –0.08 to 0.17).

For serious mental illness, in 2022, Simon et al [16] analyzed 9 unique trials involving 937 participants. The duration of these studies varied from 30 days to 18 months, with text message frequency ranging from twice a week to 12 times a day. Of these 9 studies, 7 (78%) reported statistically significant improvements in medication adherence and at least 1 clinical outcome.

For vascular disease, in 2017, Adler et al [17] examined 7 randomized trials testing SMS interventions for patients with established arterial occlusive events among 1310 participants. Although heterogeneity precluded meta-analyses, 6 (86%) out of 7 trials showed a significant effect on medication adherence across a range of definitions.

Summary

This narrative review highlights the substantial and growing evidence base for the effectiveness of SMS interventions across a range of health behaviors and outcomes. Meta-analyses consistently demonstrate small to moderate positive effects of SMS interventions on physical activity, diet and weight loss, mental health, substance use (particularly smoking and alcohol), and medication adherence. There is also emerging evidence for the benefits of SMS interventions in subpopulations such as adolescents and individuals with serious mental illness. However, the optimal design features of SMS interventions (eg, frequency, duration, and personalization) vary across studies and health domains, suggesting the need for further research to identify the most effective intervention components for specific behaviors and populations. In addition, more studies are needed to assess the long-term maintenance of effects and the potential for SMS interventions to help older adults. Despite these limitations, the current evidence supports the effectiveness of the use of SMS interventions in promoting health behavior change and improving clinical outcomes.

What Makes SMS Interventions Uniquely Useful?

SMS interventions continue to be a powerful tool for health behavior support due to their widespread use and unique functioning as a communication tool. One key reason is that text messaging remains a highly preferred form of communication. In a national sample of 1019 adults across the United States and Canada in 2023, 70% (n=713) of people reported that texting is the fastest way to reach them, 74% read every text they receive, and 62% read texts within a few minutes after receiving them [18]. Another key advantage of text messaging is its efficacy in collecting longitudinal and contextual information from individuals and delivering microdoses of intervention support. Text messaging functions as a "push" technology, delivering messages directly to individuals without any effort on their part [19], which has been shown to be a powerful engagement tool [20]. This contrasts with other digital communication modalities, such as email or web-based messaging, which require users to proactively check for messages. In contrast to text messaging, smartphone app

XSL•FO
RenderX

notifications can be turned off, and wearable technologies are not consistently worn [21].

The real-time data collection aspect of SMS interventions provides unique insights into the mechanisms of behavior change. Historically, behavioral intervention studies used sparse measurements taken immediately before and after the delivery of an intervention to examine mediators of distal behavior change [22]. This approach has constrained the ability to identify the dynamic and nonlinear processes through which behavior change likely occurs [23]. Text messaging allows for the low-barrier monitoring of participants, capturing detailed, time-stamped information about their behaviors, emotions, and environmental contexts. This results in a richer data set compared to traditional methods, which often rely on retrospective self-reports prone to recall bias [24]. Dynamic models of behavior change derived from the longitudinal data collected via 2-way SMS interventions so far have begun to offer new insights into behavior change mechanisms [24,25] have begun to offer new insights into behavior change mechanisms.

In summary, the real-time ability to collect information and deliver microdoses of intervention support is the secret superpower of SMS interventions. For researchers, it offers a unique opportunity to explore the intricate time dependencies between various precursors and health behaviors, enhancing our understanding of behavior change mechanisms. For users, it provides immediate, context-relevant support, fostering sustained engagement and promoting long-term health behavior change.

Key SMS Intervention Design Considerations

Overview

In this section, we review concepts to consider when designing SMS interventions, focusing on communication strategies, psychological theories, behavior change strategies, and adjunctive functionality. In Table 1 and the sections below, we summarize current best practices based on existing research or personal experience and highlight important open questions and areas where knowledge remains limited or where special consideration should be made. Understanding current best practices for design features and gaps in knowledge is crucial for developers seeking to create effective, engaging, and evidence-based SMS interventions.

XSL•FO
RenderX

Table 1. Open questions.

Feature and best practice	Open questions
Communication Strategies	
Onboarding	
Set expectations for the user, including intensity, duration, and what can and what cannot be responded to in real time.	How much baseline information should be captured to allow tailoring and personalization?
Embodiment	
Identifying programs with professional affiliation or credible source.	Is there an ethical obligation to remind users they are communicating with an automated computer program?
Personalization	
Personalize messages based on individual characteristics, such as name.	Should one use age-specific lingo?
Timing	
Determine the optimal timing of messages based on the user's daily routines and preferences and the target behavior or health condition.	What are the most effective strategies for determining the optimal timing of messages for different health behaviors, user characteristics, and intervention goals?
Consider the user's time zone and sleep schedule when scheduling messages to avoid sending them at inappropriate or disruptive times.	What are the trade-offs between highly personalized timing (eg, based on individual schedules) and more standardized timing (eg, based on general population norms), and how can these be balanced in the intervention design?
Allow users to customize or adjust the timing of messages based on their individual needs and preferences.	How can the timing of messages be dynamically adapted based on the user's changing needs, preferences, and contexts over time?
Intensity	
Consider using higher-frequency messaging during critical periods (eg, at the beginning of the intervention or during challenging times).	How can the frequency of messages be personalized based on individual participant characteristics and engagement patterns?
Allow participants to adjust the frequency of messages based on their needs and preferences.	What are the best practices for managing message fatigue and maintaining participant engagement over time?
Ensure that the frequency of messages does not become overwhelming or lead to message fatigue.	Does allowing voluntary "vacations" from behavioral programs provide long-term engagement benefits?
Use data analytics to monitor engagement and adjust frequency accordingly.	Engagement may not always equate to behavioral progress. What techniques can be used to determine the difference?
Duration	
Determine the appropriate duration of the intervention based on the complexity of the target behavior, the goals of the intervention, and the needs of the target population.	What is the optimal duration of SMS interventions for different health behaviors, populations, and settings, and how can this be determined empirically?
Plan for the gradual tapering or fading of intervention intensity to promote self-sufficiency and prevent relapse.	What are the most effective strategies for promoting the long-term maintenance of behavior change after the initial intervention period?
Consider offering booster sessions or periodic check-ins to reinforce skills and maintain motivation after the initial intervention period.	How often should one check in? What exactly should boosters look like?
Psychological theories	
General	
Select evidence-based psychological theories that are relevant to the target health behavior and population (eg, health belief model, theory of planned behavior, social cognitive theory, and self-determination theory).	What is the optimal way to translate psychological theories into effective SMS interventions?
Incorporate key constructs from the selected theories into the intervention design, such as self-efficacy, social support, and motivation.	How can multiple theories be integrated into a coherent intervention design?
Consider using a combination of theories to address different aspects of behavior change.	What is the role of emergent theories (eg, JITAIs[a]) in optimizing SMS interventions?
Behavior change techniques	

XSL•FO
RenderX

Feature and best practice	Open questions
Self-monitoring	
Encourage participants to regularly track their health behaviors by providing simple and user-friendly tools for self-monitoring.	How can self-monitoring be made more accessible and engaging for individuals with low health literacy or limited access to technology? How to identify individuals who experience an outsized negative affect due to the increased self-focused attention due to self-monitoring?
Tailor the frequency and timing of self-monitoring prompts based on participants' preferences and adherence patterns.	How can self-monitoring be sustained over the long term, and what are the best strategies for preventing participant burnout?
Emphasize the importance of consistent and accurate self-monitoring for behavior change and goal achievement.	What are the most effective ways to integrate self-monitoring data from multiple sources (eg, text messages, wearable devices, electronic health records) to provide comprehensive feedback and support?
Consider using gamification techniques to make self-monitoring more engaging and rewarding.	What types of gamification elements (eg, points, badges, leaderboards, challenges, levels) are most likely to resonate with the target audience and what role does social interaction play in the gamification strategy (eg, team-based challenges, peer encouragement)?
Goals setting and action planning	
Assist participants in setting SMART[b] goals aligned with their health behavior objectives.	What is the optimal balance between participant autonomy and intervention guidance in setting and monitoring goals?
Break down long-term goals into smaller, actionable steps (ie, action planning) that can be easily monitored and achieved through text message support.	How can SMS interventions be designed to help participants overcome common barriers to goal achievement, such as time constraints, the lack of resources, or competing priorities?
Provide regular prompts and reminders to monitor goal progress and encourage participants to report their accomplishments.	What is the role of habit formation and automaticity in sustaining goal-directed behaviors, and how can SMS interventions support these processes?
Adapt goals and action plans based on participants' progress and feedback.	What are the potential unintended consequences of goal-setting and monitoring (eg, feelings of failure or guilt), and how can these be mitigated through text message support?
Feedback	
Offer personalized feedback and encouragement based on goal progress, celebrating successes and providing support during setbacks.	How can SMS interventions balance the need for providing regular feedback with the risk of overwhelming or discouraging participants?
Provide timely, specific, and actionable feedback on participants' reported behaviors, such as progress toward goals, adherence to recommended actions, or trends over time.	What is the role of machine learning and adaptive algorithms in generating personalized and context-aware feedback messages?
Use a combination of positive reinforcement (eg, praise for successes) and constructive feedback (eg, suggestions for improvement) to maintain motivation and engagement.	How can the effectiveness of different feedback strategies be evaluated and compared across SMS intervention designs and health behavior domains?
Offer feedback in a nonjudgmental and supportive tone, emphasizing the participant's efforts and progress.	What are the potential unintended consequences of providing feedback on behavior (eg, increased anxiety or self-criticism), and how can these be mitigated through supportive messaging and resources?
Functionality	
Human helpers	
Identify key social support networks and human resources that can enhance the effectiveness of text messaging interventions, such as family members, health care providers, peer support groups, and research staff.	What are the optimal roles and responsibilities of different types of human helpers in SMS supporting text messaging interventions, and how can these be defined and communicated effectively?
Use text messaging to facilitate communication and coordination between users and their human helpers, such as sharing progress updates, requesting support, or scheduling appointments.	How can human helpers be trained and motivated to provide high-quality, consistent, and empathetic support to users, particularly over long periods?
Tailor the involvement of human helpers to the user's individual needs, preferences, and social context, and allow for flexibility and adaptability over time.	What are the potential risks and challenges of involving human helpers in text messaging interventions, such as breaches of confidentiality, boundary violations, or inconsistent support, and how can these be mitigated?
Multiplatform	

XSL•FO
RenderX

Feature and best practice	Open questions
Consider integrating text messaging with other digital health platforms, such as web pages or mobile apps, to provide a comprehensive and consistent user experience.	How can data from multiple platforms be effectively integrated and analyzed to provide a holistic view of user engagement and progress?
Ensure that the text messaging component complements and enhances the functionality of other platforms, rather than duplicating or conflicting with them.	What are the technical, logistical, and financial challenges of developing and maintaining a multiplatform intervention, and how can these be addressed?
Provide clear guidance to users on how to navigate between different platforms and access the full range of intervention features.	How to best assess engagement across various digital platforms?
Sensor integration	
Explore opportunities to integrate text messaging with data from body-worn sensors (eg, Fitbits [Fitbit Inc]), ambient sensors (eg, sensors in the home or car), and connected devices (eg, Alexa [Amazon.com, Inc]) to provide more personalized and context-aware support,	What types of sensors and connected devices are the most effective and acceptable for different health behaviors and populations, and how can these be seamlessly integrated with text messaging?
Use sensor data to trigger JITAIs delivered via text message based on the user's current state and context.	How can sensor data be used to develop more precise and personalized models of behavior change, and what are the implications for intervention design?
Ensure that the integration of sensor data is transparent, secure, and aligned with user preferences and expectations.	How can the validity, reliability, and meaningfulness of sensor data be ensured, and what are the implications for intervention fidelity and effectiveness?

[a]JITAI: just-in-time adaptive intervention.

[b]SMART: specific, measurable, achievable, relevant, and time-bound.

Communication Strategies

Overview

One key reason why SMS interventions are so effective for relationship building is their ability to simulate a human-like dialogue. The way in which information is conveyed, the tone of the messages, and the timing and frequency of delivery can all have a significant impact on users' engagement; motivation; and, ultimately, behavior change outcomes. This section explores several key communication strategies in SMS intervention design, including onboarding, embodiment, personalization, timing, intensity, and duration. By examining current best practices and identifying open questions in each of these areas, we aim to shed light on the critical role of communication in creating engaging, persuasive, and effective SMS interventions.

Onboarding

Clear communication during the onboarding process can help set expectations, build rapport, and increase motivation for engagement with the SMS intervention. Defining expectations is the first critical design component of a well-functioning SMS intervention and should include, at minimum, conveying information about the SMS intervention's intensity and duration. Additional information captured during the onboarding process could be used to personalize the material.

Embodiment

Identifying the program as originating from a credible source can enhance trust and engagement [26]. Regarding transparency, a major open question is whether and how SMS intervention designers should disclose that the program is automated [27], explicitly reminding users that there is no human provenance when there is none or articulating what the limits of human involvement are.

Personalization

Personalizing messages with user characteristics such as name is likely to contribute to engagement [28]. By using natural language, SMS interventions can create a sense of warmth, empathy, and personalization that resonates with users. Beyond this, it remains unknown as to how personalization such as the use of cohort-specific identifiers impacts engagement and effectiveness. It is, however, crucial for SMS interventions to avoid labeling someone as having a disease or condition [29], which can lead to the internalization of negative stereotypes and self-stigma, which can erode self-esteem and the motivation to engage in healthy behaviors [30].

Timing

For many behaviors, knowing when to interact with users may be as crucial as knowing what to do. Research on habit formation and context-dependent learning [31] suggests that behaviors are strongly influenced by the contexts and cues in which they occur. By delivering interventions at specific moments and in specific contexts, it may be possible to interrupt impending undesired habits or promote the formation of positive habits. However, identifying when these critical moments are likely to occur presents a significant challenge. Some behaviors may exhibit predictable patterns at the population level, such as increased sedentary behavior during the workday or decreased medication adherence on weekends. Other behaviors, however, may show high within-person variability, making it more challenging to predict when an individual will need support. Depending on the behavior and the necessity to control the timing of messages, consideration could be made to allow users to set their own preferred times for receiving messages.

XSL•FO
RenderX

Intensity (Dose)

The key to optimizing SMS intervention intensity is to provide sufficient support and guidance to promote behavior change while avoiding overwhelming or burdening the individual. The optimal messaging frequency may vary depending on the nature of the target behavior and its temporal dynamics. For behaviors that occur at short intervals, such as smoking or snacking, more frequent interactions may be necessary to provide timely support and reinforcement. In contrast, for behaviors that occur at longer time scales, such as physical activity or medication adherence, a more spaced-out cadence of messaging may be appropriate. However, it is important to recognize that simply increasing the intensity of messaging, particularly for challenging or complex behaviors, may not always lead to better outcomes [32]. In fact, this approach could backfire by overwhelming or frustrating users, leading to disengagement or even reactance [33]. An open question is how to taper the frequency or intensity of messages over time to gradually transfer the responsibility for behavior change from the intervention to the individual. Another related question is how often, for frequent behaviors, one must interact to be effective.

Duration

As with all behavioral interventions, key considerations for SMS interventions include how long they must run to be effective. Research suggests that it takes an average of 66 days for a new behavior to become a habit, but this can vary widely depending on the complexity of the behavior and individual differences [34]. There is also evidence that the duration of engagement with SMS interventions does not necessarily translate to more favorable outcomes [35,36]. In our experience, the relationship between SMS intervention engagement and effectiveness is U-shaped: individuals who show the greatest effects either use the program the least (indicating they have adopted the behavior and do not need the support) or the most (indicating their commitment to the work of behavior change). A fixed duration may be appropriate for interventions targeting short-term goals or behaviors with a clear end point, such as completing a vaccination schedule or recovering from surgery. However, for interventions targeting complex, long-term behavior changes, such as adopting a healthy diet or increasing physical activity, a more flexible and individualized approach may be necessary.

Psychological Theories

The incorporation of psychological theories into the design of SMS interventions has been a topic of growing interest. The meta-analysis by Head et al [1] and the systematic review by Hall et al [2] both found that SMS interventions based on theories were significantly more effective in promoting health behaviors than those not based on theories. Several specific theories have been commonly used in the design of SMS interventions. The transtheoretical model [37], which posits that individuals move through distinct stages of change when modifying health behaviors, has been successfully applied in SMS interventions for physical activity and smoking cessation [38]. Social cognitive theory [39], which emphasizes the interplay among personal, behavioral, and environmental factors in shaping behavior, has also been frequently used in SMS interventions for various health behaviors [38]. Other theories that have informed SMS intervention design include the health belief model [40], which focuses on individuals' perceptions of health risks and benefits, and the theory of planned behavior [41], which emphasizes the role of attitudes, subjective norms, and perceived behavioral control in shaping intentions and behaviors [42].

While evidence supports the use of these traditional theories in SMS intervention design, some researchers have argued that existing theories may need to be adapted or integrated to better suit the unique context of mobile interventions [43]. For example, existing theories could be updated and adapted to take into account the dynamic and interactive nature in which behaviors play out in the real world, as well as the potential for real-time interventions to adapt to those time- and context-dependent needs [44]. For example, mathematical equations informed by control theory [45] could be used to make predictions at a given time point for a given individual based on prior data from the individual to tailor some aspect of intervention content. Testing of such dynamical models is in the early stages [46,47]; thus, it remains unknown whether the complexity pays off in requisite effects.

BCTs Used in SMS Interventions

Self-monitoring, goal-setting, action planning, and feedback are 3 key BCTs that have been widely incorporated into behavioral interventions in general [22,48] and digital interventions (including SMS interventions) specifically [49].

Self-Monitoring

Many behaviors, especially when they become habitual, are below the level of conscious recognition [50]. The act of self-monitoring itself can raise awareness of one's behavior and motivate change. SMS interventions offer a convenient and accessible platform for prompting and assisting individuals in bringing behaviors, emotions, and desires to awareness. This can involve prompting individuals to reflect before a given behavior (to raise awareness of motives and expectancies) or following a behavior (to promote active reflection). It can also have varying levels of specificity, from simply prompting a reflection on a behavior (eg, "Think about how exercising made you feel?") to requesting specific data (eg, "Please reply with the number of steps you took today."). The frequency and format of self-monitoring prompts should be tailored to individual preferences and needs weighed against program requirements for tailoring. An open question remains how to identify individuals who experience an outsized negative affect due to the increased self-focused attention due to self-monitoring [51].

Goal-Setting and Action Planning

Goal-setting and *action planning* are closely related BCTs that assist behavior change [52]. A meta-analysis of 141 studies testing 384 effect sizes (16,523 participants) found a positive unique effect of goal-setting across a range of behaviors (Cohen d=0.34, 95% CI 0.28-0.41) [53]. SMS interventions should ideally be designed with specific, measurable, achievable, relevant, and time-bound goals [31]. Relatedly, action planning should ideally involve the concept of implementation intentions: the more specific someone can be with *when*, *where*, and *how*

XSL•FO
RenderX

a goal will be achieved, the more likely it is to be accomplished [54]. SMS interventions can provide templates or examples to help users develop effective goals and action plans and can send reminders or prompts to encourage adherence to the plans. An open question remains how best to support autonomy in self-determined goals [55] while pushing individuals who may not be intrinsically motivated toward progress.

Performance Feedback

The principles of cybernetics and control theory [45] suggest that feedback operates as part of a closed-loop system, wherein individuals compare their current behavior to their goals and adjust their actions accordingly. SMS interventions can leverage this feedback loop by providing regular, timely feedback and encouraging users to reflect on their progress and adjust as needed. Feedback can help individuals gauge their progress, identify areas for improvement, and adjust their strategies as needed. In SMS interventions, behavioral feedback can be descriptive (eg, "You walked an average of 8,000 steps per day this week.") or evaluative (eg, "Great job! You exceeded your goal of 7,000 steps per day this week!"). The frequency, tone, and format of feedback messages can be tailored to individual preferences and needs to maximize engagement and effectiveness. One open question is how best to balance the need for providing regular feedback with the risk of overwhelming or discouraging participants. Another open question is how to avoid behavioral disengagement [56] when an individual exhibits persistent goal failures.

Enhanced Functionalities

Human Helpers

Automated systems for delivering SMS interventions offer several advantages, including scalability and reproducibility. However, a key limitation of automated SMS interventions is their inability to provide supportive accountability, which can be crucial for patient engagement and support [57]. To address this, researchers have explored incorporating human helpers, such as health care professionals, significant others, and peers, into SMS interventions.

One early example of incorporating professionals into SMS interventions was a study by Rodgers et al [58], which evaluated a smoking cessation program where participants received personalized text messages from a team of counselors. The counselors used a web-based interface to monitor participant progress and tailor the content and timing of messages based on individual needs. To capitalize on the strengths of both human and automated approaches, some SMS interventions have adopted a hybrid model. Aguilera et al [59] evaluated a hybrid SMS intervention for depression that used an automated system to send daily mood-tracking prompts and self-management tips while also allowing participants to request additional support from a mental health coach via an text message. This hybrid approach aimed to strike a balance between the efficiency of automation and the benefits of human interaction.

Dyadic SMS interventions, which involve significant others, have been studied primarily in adolescent populations, addressing conditions such as asthma [60], diabetes [61], and mental health conditions [62]. In adult populations, dyadic SMS interventions have been explored for promoting physical activity [63]. Best practices for implementing dyadic SMS interventions include identifying key social support networks, providing training and guidance to human helpers, facilitating communication between users and helpers via text messaging, and tailoring the involvement of helpers to individual needs and preferences. However, several open questions remain regarding the optimal roles and responsibilities of different types of human helpers; strategies for ensuring the quality and consistency of support; and approaches to addressing potential risks and challenges, such as breaches of confidentiality or inconsistent support. Future research should aim to address these questions to further optimize the integration of human support into SMS interventions.

Multiplatform Integration

There is evidence that using multiple modes of digital communication (ie, combining text messaging with other digital communication platforms, such as web pages or mobile apps) can enhance health behavior change [64]. Best practices include ensuring that the text messaging component complements and enhances the functionality of other platforms, providing clear guidance on navigating between platforms, and using data from multiple sources to inform personalization. For example, there are many self-report behavioral instruments composed of more than a few questions that would be tedious if deployed through text messaging. In addition, there are forms of feedback such as interactive graphs that require other digital modalities. There remain open questions around the optimal combination and sequencing of communication channels, the challenges of developing and maintaining multiplatform interventions that can communicate with one another, and challenges of determining which components are necessary for effectiveness.

Integration With Sensors and Connected Devices

Integrating text messaging with data from body-worn and ambient sensors presents an opportunity to deliver more personalized and context-aware support. For example, studies have found promising results when using SMS interventions integrated with activity monitors [65], weight scales [66], physiological monitors (eg, blood glucose [67] and blood pressure [68]), and location-monitoring technology (GPS) [69]. It is certain that the use of wearable sensors [70] and ambient environmental sensors [71] will play a larger role in health monitoring and behavioral interventions. Open questions remain regarding the most effective types of sensors for different health behaviors, the level of certainty needed for inferred behaviors, the development of precise and personalized behavior change models, the optimal algorithms for triggering interventions or support, and the ethical and social implications of using sensor data.

Future Opportunities: LLMs as SMS Intervention Conversational Agents

Overview

To date, SMS interventions have primarily used rule-based systems that process simple client inputs through branching

XSL•FO
RenderX

algorithms to generate prewritten text responses. For example, a system might compare a reported alcohol consumption quantity against a preset threshold to trigger specific messages about drinking behavior. Although efforts to enhance interactivity through conversational agents such as chatbots have shown promise in promoting behavior change across various domains [72], including mental health [73] and substance use [74,75], these systems often lack the adaptability required for deep therapeutic engagement.

LLMs as Therapeutic Conversational Agents

The advent of LLMs [76] marks a potentially transformative advancement in conducting automated therapeutic conversations. LLMs, such as ChatGPT (OpenAI), are deep learning models trained on extensive textual data capable of understanding the context and relationships between words and generating responsive, novel text that is coherent and contextually appropriate [77]. These models can be fine-tuned after training for specific tasks or domains, enhancing their adaptability for diverse applications [78], including psychotherapy [79]. Emerging studies highlight LLMs' capabilities in areas such as emotional awareness [80], social awareness [81], empathy [82], creativity [83], and reasoning [84]. Preliminary findings suggest that ChatGPT can engage positively in therapeutic conversations [85], actively listen, provide validation, and suggest coping strategies [86]. Trials comparing rule-based and generative counseling are still in early stages [87].

Training LLMs

When developing LLM-based SMS interventions for health behavior support, researchers are advised to prioritize pretraining with domain-specific data sets to guarantee that the LLM produces accurate and relevant responses. This essential step involves compiling a diverse data set, including successful SMS interventions' text libraries, contributions from human experts, counseling manuals, and psychological theories. These types of data sets should be used to familiarize the LLM with effective language patterns, intervention strategies, and the appropriate tone for health behavior change. The training process should be carefully customized to stress the significance of clarity, empathy, and motivational encouragement, which are crucial for impactful health communication. In addition, the integration of both preformed and real-time responses is crucial, as real-time LLM-generated content can offer a dynamic conversational experience, while preformed replies ensure consistency and accuracy for sensitive topics.

Addressing the Unpredictability of LLMs

Due to their generative nature, LLM-based conversations can vary in quality and appropriateness. We suggest several key strategies to minimize untoward communication events. The first line of defense involves *LLM guardrails* [88], custom configurations within the LLM that prevent the generation of harmful or inappropriate content, including filters for sensitive topics and checks for alignment with therapeutic best practices. The second strategy is to incorporate a *human-in-the-loop system* [89], where a trained research team member reviews all messages and can edit and clarify any undesired message. These techniques ensure that LLM-based SMS interventions are both effective in promoting health behaviors and in conformance with high standards of quality and ethical practice.

Future Challenges: Bridging SMS Interventions From Research to Real-World Application

Currently, several SMS interventions are available to the US public, offering evidence-based support across various health domains. Notable examples include Text4baby, which provides maternal health information; VEText, which offers appointment reminders to US veterans; and Smokefree, which offers smoking cessation support. While these programs demonstrate the efficacy of SMS interventions, they are primarily accessible through government and foundation channels, underscoring the need for enhanced and centralized dissemination methods.

SMS interventions face several key challenges in transitioning from research environments to widespread real-world implementation and effectiveness. Unlike apps, which benefit from commercial platforms such as app stores facilitating consumer access and enabling monetization, SMS interventions lack a marketplace. This gap restricts commercial innovation and the potential for broader adoption. Furthermore, health systems, ideally positioned to leverage SMS interventions for enhancing patient engagement, encounter barriers due to outdated legal apprehensions regarding the security of text message communication. Concerns about the potential for unauthorized access to sensitive patient data and the technical difficulties of integrating SMS functionalities with existing health informatics systems further complicate implementation efforts, often requiring significant technological investment. Finally, there is the concern that outside the context of a controlled trial, SMS intervention effectiveness will be hampered by low longitudinal user engagement.

To address these challenges and foster the adoption of SMS interventions, several strategies are proposed. First, establishing a dedicated commercial portal for SMS interventions could mirror the app store model, offering a platform for users to access evidence-based interventions. This would not only enhance visibility but also stimulate commercial interest and development. Second, adopting standardized informed consent procedures would enable patients to opt into SMS interventions knowingly and willingly, mitigating legal and ethical concerns. In addition, formulating comprehensive policies and guidelines for digital communication within health systems can promote consistency, compliance, and confidence in the security of text message exchanges. Examples of health systems successfully implementing SMS interventions in routine health care are, however, growing. For example, Bressman et al [90] reported on the implementation of a 30-day postdischarge intervention using automated texting to supplement the standard of care within a single primary care practice in Penn Medicine. They found that the adjusted OR for an emergency department visit was 0.77 (95% CI 0.45-1.30) and for a readmission was 0.45 (95% CI 0.23-0.86). Educating health care providers on the value and utility of SMS interventions could prompt them to prescribe these interventions as part of a treatment plan, much

XSL•FO
RenderX

like any conventional medication. Finally, we believe that engagement with SMS interventions can be optimized through thoughtful design focused on matching user needs with support provision. For example, we have found that for SMS interventions of longer duration, voluntary vacations or breaks from SMS interventions could reduce attrition [91] and accommodate self-determination [92].

Conclusions

In this paper, we have attempted to capture the essence of SMS interventions, recognizing their fundamental yet often underestimated role in shaping health behaviors. Their apparent simplicity masks a capacity for delivering health interventions that are both nuanced and highly personalized, effectively transmitting information, offering support, and bridging gaps in health care services. The true power of text messaging emerges through its steady and subtle influence over time rather than through immediate, dramatic effects.

From the initial tests in the early 2000s to a peak of scholarly activity in the early 2020s, SMS interventions have demonstrated a broad spectrum of application across critical health behaviors and outcomes, underscored by a steady growth in both academic interest and funding. The narrative review further solidifies the evidence base for SMS interventions, showcasing their effectiveness in promoting physical activity, aiding weight loss, supporting mental health, reducing substance use, and enhancing medication adherence, with nuanced effects across diverse populations and health domains. We posit that this broad and consistent effectiveness is largely due to the ongoing widespread use of text messaging to communicate and unique functioning of text messaging as a communication tool.

The exploration of SMS interventions' key features reveals a delicate balance between technological advancement and human-centric design principles, underscoring the importance of personalization, timing, intensity, and the integration of psychological theories. These considerations are pivotal in designing SMS interventions that not only engage users but also instigate meaningful behavioral change. The emergence of generative artificial intelligence presents an exciting frontier, promising to enrich SMS interventions with more natural and adaptive conversational capabilities, provided that ethical standards and quality assurance mechanisms are rigorously maintained.

Despite all this support and promise, the commercialization of SMS interventions, their integration into health care systems, and their engagement outside controlled studies remain ongoing challenges. However, the potential of SMS interventions to bridge the communication gap between patients and providers is immense. With a strategic emphasis on developing a marketplace, patient consent procedures, policy development, medical education, and optimizing user engagement, SMS interventions could revolutionize health care delivery, enhancing patient engagement and support across the continuum of care.

As we navigate the future of health communication, the insights garnered from 2 decades of SMS intervention research illuminate the path forward, highlighting the need for continued innovation, integration, and interdisciplinary collaboration. As different messaging platforms arise, such as WhatsApp (Meta Platforms, Inc), Facebook Messenger (Meta Platforms, Inc), WeChat (Tencent Inc), and Snapchat (Snap Inc), existing evidence from SMS interventions can inform design. By harnessing the simplicity, adaptability, and depth of SMS interventions, we can continue to shape the landscapes of health behavior and support, making strides toward improved health outcomes and reduced disparities in the years to come.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PubMed search terms.
[DOCX File , 19 KB - jmir_v26i1e58726_app1.docx]

References

1. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013 Nov;97:41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
2. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015 Mar 18;36:393-415 [FREE Full text] [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
3. Armanasco AA, Miller YD, Fjeldsoe BS, Marshall AL. Preventive health behavior change text message interventions: a meta-analysis. Am J Prev Med 2017 Mar;52(3):391-402. [doi: 10.1016/j.amepre.2016.10.042] [Medline: 28073656]
4. Smith DM, Duque L, Huffman JC, Healy BC, Celano CM. Text message interventions for physical activity: a systematic review and meta-analysis. Am J Prev Med 2020 Jan;58(1):142-151 [FREE Full text] [doi: 10.1016/j.amepre.2019.08.014] [Medline: 31759805]
5. Siopis G, Chey T, Allman-Farinelli M. A systematic review and meta-analysis of interventions for weight management using text messaging. J Hum Nutr Diet 2015 Feb;28 Suppl 2:1-15. [doi: 10.1111/jhn.12207] [Medline: 24480032]

XSL•FO
RenderX

6. Skinner R, Gonet V, Currie S, Hoddinott P, Dombrowski SU. A systematic review with meta-analyses of text message-delivered behaviour change interventions for weight loss and weight loss maintenance. Obes Rev 2020 Jun;21(6):e12999. [doi: 10.1111/obr.12999] [Medline: 32043809]
7. Partridge SR, Raeside R, Singleton A, Hyun K, Redfern J. Effectiveness of text message interventions for weight management in adolescents: systematic review. JMIR Mhealth Uhealth 2020 May 26;8(5):e15849 [FREE Full text] [doi: 10.2196/15849] [Medline: 32348264]
8. Senanayake B, Wickramasinghe SI, Chatfield MD, Hansen J, Edirippulige S, Smith AC. Effectiveness of text messaging interventions for the management of depression: a systematic review and meta-analysis. J Telemed Telecare 2019 Oct;25(9):513-523. [doi: 10.1177/1357633X19875852] [Medline: 31631764]
9. Cox KL, Allida SM, Hackett ML. Text messages to reduce depressive symptoms: do they work and what makes them effective? A systematic review. Health Educ J 2020 Oct 08;80(3):253-271 [FREE Full text] [doi: 10.1177/0017896920959368]
10. D'Arcey J, Collaton J, Kozloff N, Voineskos AN, Kidd SA, Foussias G. The use of text messaging to improve clinical engagement for individuals with psychosis: systematic review. JMIR Ment Health 2020 Apr 02;7(4):e16993 [FREE Full text] [doi: 10.2196/16993] [Medline: 32238334]
11. Scott-Sheldon LA, Lantini R, Jennings EG, Thind H, Rosen RK, Salmoirago-Blotcher E, et al. Text messaging-based interventions for smoking cessation: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016 May 20;4(2):e49 [FREE Full text] [doi: 10.2196/mhealth.5436] [Medline: 27207211]
12. Zhou X, Wei X, Cheng A, Liu Z, Su Z, Li J, et al. Mobile phone-based interventions for smoking cessation among young people: systematic review and meta-analysis. JMIR Mhealth Uhealth 2023 Sep 12;11:e48253 [FREE Full text] [doi: 10.2196/48253] [Medline: 37706482]
13. Bendtsen M, McCambridge J, Åsberg K, Bendtsen P. Text messaging interventions for reducing alcohol consumption among risky drinkers: systematic review and meta-analysis. Addiction 2021 May;116(5):1021-1033 [FREE Full text] [doi: 10.1111/add.15294] [Medline: 33047865]
14. Thakkar J, Kurup R, Laba TL, Santo K, Thiagalingam A, Rodgers A, et al. Mobile telephone text messaging for medication adherence in chronic disease: a meta-analysis. JAMA Intern Med 2016 Mar;176(3):340-349. [doi: 10.1001/jamainternmed.2015.7667] [Medline: 26831740]
15. Mehra N, Tunje A, Hallström IK, Jerene D. Effectiveness of mobile phone text message reminder interventions to improve adherence to antiretroviral therapy among adolescents living with HIV: a systematic review and meta-analysis. PLoS One 2021 Jul 22;16(7):e0254890 [FREE Full text] [doi: 10.1371/journal.pone.0254890] [Medline: 34293033]
16. Simon E, Edwards AM, Sajatovic M, Jain N, Montoya JL, Levin JB. Systematic literature review of text messaging interventions to promote medication adherence among people with serious mental illness. Psychiatr Serv 2022 Oct 01;73(10):1153-1164 [FREE Full text] [doi: 10.1176/appi.ps.202100634] [Medline: 35959534]
17. Adler AJ, Martin N, Mariani J, Tajer CD, Owolabi OO, Free C, et al. Mobile phone text messaging to improve medication adherence in secondary prevention of cardiovascular disease. Cochrane Database Syst Rev 2017 Apr 29;4(4):CD011851 [FREE Full text] [doi: 10.1002/14651858.CD011851.pub2] [Medline: 28455948]
18. 2024 State of business texting report. Text Request. URL: https://www.textrequest.com/ebooks/state-of-business-texting-2024 [accessed 2024-08-19]
19. Suffoletto B. Text message behavioral interventions: from here to where? Curr Opin Psychol 2016 Feb 01;9:16-21 [FREE Full text] [doi: 10.1016/j.copsyc.2015.09.012] [Medline: 26665157]
20. Bidargaddi N, Almirall D, Murphy S, Nahum-Shani I, Kovalcik M, Pituch T, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR Mhealth Uhealth 2018 Nov 29;6(11):e10123 [FREE Full text] [doi: 10.2196/10123] [Medline: 30497999]
21. Patel M, Pavic A, Goodwin VA. Wearable inertial sensors to measure gait and posture characteristic differences in older adult fallers and non-fallers: a scoping review. Gait Posture 2020 Feb;76:110-121 [FREE Full text] [doi: 10.1016/j.gaitpost.2019.10.039] [Medline: 31756666]
22. Carey RN, Connell LE, Johnston M, Rothman AJ, de Bruin M, Kelly MP, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2019 Jul 17;53(8):693-707 [FREE Full text] [doi: 10.1093/abm/kay078] [Medline: 30304386]
23. Hayes AM, Laurenceau JP, Feldman G, Strauss JL, Cardaciotto L. Change is not always linear: the study of nonlinear and discontinuous patterns of change in psychotherapy. Clin Psychol Rev 2007 Jul;27(6):715-723 [FREE Full text] [doi: 10.1016/j.cpr.2007.01.008] [Medline: 17316941]
24. Suffoletto B, Chung T. Goal commitment and goal confidence mediate the path between desire to get drunk and binge drinking among young adults receiving text message goal-related feedback. Alcohol Clin Exp Res (Hoboken) 2023 May 09;47(5):1010-1021. [doi: 10.1111/acer.15065] [Medline: 36974483]
25. Mason M, Mennis J, Way T, Floyd Campbell L. Real-time readiness to quit and peer smoking within a text message intervention for adolescent smokers: modeling mechanisms of change. J Subst Abuse Treat 2015 Dec;59:67-73. [doi: 10.1016/j.jsat.2015.07.009] [Medline: 26297323]
26. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and agenda for future research. J Med Internet Res 2017 Jun 19;19(6):e218 [FREE Full text] [doi: 10.2196/jmir.7579] [Medline: 28630033]

XSL•FO
RenderX

27. Markham A, Buchanan E. Ethical decision-making and internet research: recommendations from the AoIR Ethics Working Committee (version 2.0). Association of Internet Researchers. 2012. URL: https://aoir.org/reports/ethics2.pdf [accessed 2024-08-19]
28. Rimer BK, Kreuter MW. Advancing tailored health communication: a persuasion and message effects perspective. J Commun 2006 Aug;56(s1):S184-S201. [doi: 10.1111/j.1460-2466.2006.00289.x]
29. Scambler G. Health-related stigma. Sociol Health Illn 2009 Apr;31(3):441-455 [FREE Full text] [doi: 10.1111/j.1467-9566.2009.01161.x] [Medline: 19366430]
30. Corrigan PW, Larson JE, Rüsch N. Self-stigma and the "why try" effect: impact on life goals and evidence-based practices. World Psychiatry 2009 Jun 12;8(2):75-81 [FREE Full text] [doi: 10.1002/j.2051-5545.2009.tb00218.x] [Medline: 19516923]
31. Wood W, Neal DT. A new look at habits and the habit-goal interface. Psychol Rev 2007;114(4):843-863. [doi: 10.1037/0033-295x.114.4.843]
32. Conroy DE, Wu J, Lee AM, Brunke-Reese D, Lagoa CM. Dose-response relations between the frequency of two types of momentary feedback prompts and daily physical activity. Health Psychol 2023 Mar;42(3):151-160 [FREE Full text] [doi: 10.1037/hea0001271] [Medline: 36862471]
33. Brehm JW. A Theory of Psychological Reactance. Cambridge, MA: Academic Press; 1966.
34. Lally P, van Jaarsveld CH, Potts HW, Wardle J. How are habits formed: modelling habit formation in the real world. Euro J Social Psych 2009 Jul 16;40(6):998-1009. [doi: 10.1002/ejsp.674]
35. Fjeldsoe BS, Goode AD, Job J, Eakin EG, Spilsbury KL, Winkler E. Dose and engagement during an extended contact physical activity and dietary behavior change intervention delivered via tailored text messaging: exploring relationships with behavioral outcomes. Int J Behav Nutr Phys Act 2021 Sep 07;18(1):119 [FREE Full text] [doi: 10.1186/s12966-021-01179-8] [Medline: 34493307]
36. Suffoletto B, Chung T, Muench F, Monti P, Clark DB. A text message intervention with adaptive goal support to reduce alcohol consumption among non-treatment-seeking young adults: non-randomized clinical trial with voluntary length of enrollment. JMIR Mhealth Uhealth 2018 Feb 16;6(2):e35 [FREE Full text] [doi: 10.2196/mhealth.8530] [Medline: 29453191]
37. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983;51(3):390-395. [doi: 10.1037//0022-006X.51.3.390]
38. Orr JA, King RJ. Mobile phone SMS messages can enhance healthy behaviour: a meta-analysis of randomised controlled trials. Health Psychol Rev 2015;9(4):397-416. [doi: 10.1080/17437199.2015.1022847] [Medline: 25739668]
39. Bandura A. Social cognitive theory of self-regulation. Organ Behav Hum Decis Process 1991 Dec;50(2):248-287. [doi: 10.1016/0749-5978(91)90022-L]
40. Rosenstock IM. The health belief model and preventive health behavior. Health Educ Monogr 1974 Dec 01;2(4):354-386 [FREE Full text] [doi: 10.1177/109019817400200405]
41. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
42. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344 [FREE Full text] [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
43. Riley WT, Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R. Health behavior models in the age of mobile interventions: are our theories up to the task? Transl Behav Med 2011 Mar;1(1):53-71 [FREE Full text] [doi: 10.1007/s13142-011-0021-7] [Medline: 21796270]
44. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
45. Carver CS, Scheier MF. Control theory: a useful conceptual framework for personality-social, clinical, and health psychology. Psychol Bull 1982;92(1):111-135. [doi: 10.1037//0033-2909.92.1.111]
46. Park J, Kim M, El Mistiri M, Kha R, Banerjee S, Gotzian L, et al. Advancing understanding of just-in-time states for supporting physical activity (project JustWalk JITAI): protocol for a system id study of just-in-time adaptive interventions. JMIR Res Protoc 2023 Sep 26;12:e52161 [FREE Full text] [doi: 10.2196/52161] [Medline: 37751237]
47. Daryabeygi-Khotbehsara R, Shariful Islam SM, Dunstan D, McVicar J, Abdelrazek M, Maddison R. Smartphone-based interventions to reduce sedentary behavior and promote physical activity using integrated dynamic models: systematic review. J Med Internet Res 2021 Sep 13;23(9):e26315 [FREE Full text] [doi: 10.2196/26315] [Medline: 34515637]
48. Michie S, Wood CE, Johnston M, Abraham C, Francis JJ, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technol Assess 2015 Nov;19(99):1-188 [FREE Full text] [doi: 10.3310/hta19990] [Medline: 26616119]
49. Garnett C, Crane D, Brown J, Kaner E, Beyer F, Muirhead C, et al. Reported theory use by digital interventions for hazardous and harmful alcohol consumption, and association with effectiveness: meta-regression. J Med Internet Res 2018 Feb 28;20(2):e69 [FREE Full text] [doi: 10.2196/jmir.8807] [Medline: 29490895]

XSL•FO
RenderX

50. Verplanken B, Orbell S. Reflections on past behavior: a self-report index of habit strength. J Appl Soc Psychol 2006 Jul 31;33(6):1313-1330 [FREE Full text] [doi: 10.1111/j.1559-1816.2003.tb01951.x]
51. Mor N, Winquist J. Self-focused attention and negative affect: a meta-analysis. Psychol Bull 2002 Jul;128(4):638-662. [doi: 10.1037/0033-2909.128.4.638] [Medline: 12081086]
52. Locke EA, Latham GP. Building a practically useful theory of goal setting and task motivation. A 35-year odyssey. Am Psychol 2002 Sep;57(9):705-717. [doi: 10.1037//0003-066x.57.9.705] [Medline: 12237980]
53. Epton T, Currie S, Armitage CJ. Unique effects of setting goals on behavior change: systematic review and meta-analysis. J Consult Clin Psychol 2017 Dec;85(12):1182-1198. [doi: 10.1037/ccp0000260] [Medline: 29189034]
54. Bailey RR. Goal setting and action planning for health behavior change. Am J Lifestyle Med 2019 Sep 13;13(6):615-618 [FREE Full text] [doi: 10.1177/1559827617729634] [Medline: 31662729]
55. Ryan RM, Deci EL. Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will? J Pers 2006 Dec;74(6):1557-1585. [doi: 10.1111/j.1467-6494.2006.00420.x] [Medline: 17083658]
56. Wrosch C, Scheier MF, Miller GE, Schulz R, Carver CS. Adaptive self-regulation of unattainable goals: goal disengagement, goal reengagement, and subjective well-being. Pers Soc Psychol Bull 2003 Dec 02;29(12):1494-1508. [doi: 10.1177/0146167203256921] [Medline: 15018681]
57. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
58. Rodgers A, Corbett T, Bramley D, Riddell T, Wills M, Lin RB, et al. Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. Tob Control 2005 Aug;14(4):255-261 [FREE Full text] [doi: 10.1136/tc.2005.011577] [Medline: 16046689]
59. Aguilera A, Bruehlman-Senecal E, Demasi O, Avila P. Automated text messaging as an adjunct to cognitive behavioral therapy for depression: a clinical trial. J Med Internet Res 2017 May 08;19(5):e148 [FREE Full text] [doi: 10.2196/jmir.6914] [Medline: 28483742]
60. Coker TR, Mitchell SJ, Lowry SJ, Klein EJ, Stout JW, Brown JC, et al. Text2Breathe: text-message intervention for parent communication and pediatric asthma. Acad Pediatr 2023 Jan;23(1):123-129 [FREE Full text] [doi: 10.1016/j.acap.2022.05.004] [Medline: 35577281]
61. Mayberry LS, Berg CA, Harper KJ, Osborn CY. The design, usability, and feasibility of a family-focused diabetes self-care support mHealth intervention for diverse, low-income adults with type 2 diabetes. J Diabetes Res 2016;2016:7586385 [FREE Full text] [doi: 10.1155/2016/7586385] [Medline: 27891524]
62. Suffoletto B, Goldstein T, Brent D. A text message intervention for adolescents with depression and their parents or caregivers to overcome cognitive barriers to mental health treatment initiation: focus groups and pilot trial. JMIR Form Res 2021 Nov 09;5(11):e30580 [FREE Full text] [doi: 10.2196/30580] [Medline: 34751665]
63. Berli C, Stadler G, Inauen J, Scholz U. Action control in dyads: a randomized controlled trial to promote physical activity in everyday life. Soc Sci Med 2016 Aug;163:89-97 [FREE Full text] [doi: 10.1016/j.socscimed.2016.07.003] [Medline: 27421075]
64. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
65. Agboola S, Jethwani K, Lopez L, Searl M, O'Keefe S, Kvedar J. Text to move: a randomized controlled trial of a text-messaging program to improve physical activity behaviors in patients with type 2 diabetes mellitus. J Med Internet Res 2016 Nov 18;18(11):e307 [FREE Full text] [doi: 10.2196/jmir.6439] [Medline: 27864165]
66. Godino JG, Golaszewski NM, Norman GJ, Rock CL, Griswold WG, Arredondo E, et al. Text messaging and brief phone calls for weight loss in overweight and obese English- and Spanish-speaking adults: a 1-year, parallel-group, randomized controlled trial. PLoS Med 2019 Sep;16(9):e1002917 [FREE Full text] [doi: 10.1371/journal.pmed.1002917] [Medline: 31553725]
67. Capozza K, Woolsey S, Georgsson M, Black J, Bello N, Lence C, et al. Going mobile with diabetes support: a randomized study of a text message-based personalized behavioral intervention for type 2 diabetes self-care. Diabetes Spectr 2015 May;28(2):83-91 [FREE Full text] [doi: 10.2337/diaspect.28.2.83] [Medline: 25987806]
68. Muldoon MF, Einhorn J, Yabes JG, Burton D, Irizarry T, Basse J, et al. Randomized feasibility trial of a digital intervention for hypertension self-management. J Hum Hypertens 2022 Aug 08;36(8):718-725. [doi: 10.1038/s41371-021-00574-9] [Medline: 34239050]
69. Tobin K, Heidari O, Volpi C, Sodder S, Duncan D. Use of geofencing interventions in population health research: a scoping review. BMJ Open 2023 Aug 03;13(8):e069374 [FREE Full text] [doi: 10.1136/bmjopen-2022-069374] [Medline: 37536963]
70. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on wearable technology for connected health: scoping review of research trends, advances, and barriers. J Med Internet Res 2019 Sep 05;21(9):e14017 [FREE Full text] [doi: 10.2196/14017] [Medline: 31489843]
71. Yin K, Laranjo L, Tong HL, Lau AY, Kocaballi AB, Martin P, et al. Context-aware systems for chronic disease patients: scoping review. J Med Internet Res 2019 Jun 17;21(6):e10896 [FREE Full text] [doi: 10.2196/10896] [Medline: 31210138]

XSL•FO
RenderX

72. Martinengo L, Jabir AI, Goh WW, Lo NY, Ho MH, Kowatsch T, et al. Conversational agents in health care: scoping review of their behavior change techniques and underpinning theory. J Med Internet Res 2022 Oct 03;24(10):e39243 [FREE Full text] [doi: 10.2196/39243] [Medline: 36190749]
73. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
74. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend 2021 Oct 01;227:108986 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108986] [Medline: 34507061]
75. Dulin P, Mertz R, Edwards A, King D. Contrasting a mobile app with a conversational chatbot for reducing alcohol consumption: randomized controlled pilot trial. JMIR Form Res 2022 May 16;6(5):e33037 [FREE Full text] [doi: 10.2196/33037] [Medline: 35576569]
76. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv Preprint posted online June 12, 2017 [FREE Full text]
77. ChatGPT homepage. ChatGPT. URL: https://chatgpt.com/ [accessed 2023-12-07]
78. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Jan 16;55(9):1-35. [doi: 10.1145/3560815]
79. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 02;3(1):12 [FREE Full text] [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
80. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058 [FREE Full text] [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
81. Sufyan NS, Fadhel FH, Alkhathami SS, Mukhadi JYA. Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists. Front Psychol 2024 Feb 2;15:1353022 [FREE Full text] [doi: 10.3389/fpsyg.2024.1353022] [Medline: 38379623]
82. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
83. Stevenson C, Smal I, Baas M, Grasman R, van der Maas H. Putting GPT-3's creativity to the (alternative uses) test. arXiv Preprint posted online June 10, 2022 [FREE Full text]
84. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. Nat Hum Behav 2023 Sep 31;7(9):1526-1541. [doi: 10.1038/s41562-023-01659-w] [Medline: 37524930]
85. Das A, Selek S, Warner AR, Zuo X, Hu Y, Kuttichi Keloth V, et al. Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In: Proceedings of the 21st Workshop on Biomedical Language Processing. 2022 Presented at: BioNLP 2022; May 26, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.bionlp-1.27]
86. Eshghie M, Eshghie M. ChatGPT as a therapist assistant: a suitability study. SSRN Preprint posted online May 1, 2023 [FREE Full text] [doi: 10.2139/ssrn.4423832]
87. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. arXiv Preprint posted online May 23, 2023 [FREE Full text]
88. Dong Y, Mu R, Jin G, Qi Y, Hu J, Zhao X, et al. Building guardrails for large language models. arXiv Preprint posted online February 2, 2024 [FREE Full text]
89. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 2022 Aug 17;56:3005-3054. [doi: 10.1007/s10462-022-10246-w]
90. Bressman E, Long JA, Honig K, Zee J, McGlaughlin N, Jointer C, et al. Evaluation of an automated text message-based program to reduce use of acute health care resources after hospital discharge. JAMA Netw Open 2022 Oct 03;5(10):e2238293 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.38293] [Medline: 36287564]
91. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
92. Deci E, Ryan R. Self-determination theory. In: Handbook of Theories of Social Psychology: Volume 1. Thousand Oaks, CA: SAGE Publications Ltd; 2012:416-437.

Abbreviations

BCT: behavior change technique
LLM: large language model
OR: odds ratio

XSL•FO
RenderX

Edited by G Eysenbach; submitted 22.03.24; peer-reviewed by K Wiseman, E Burner; comments to author 19.05.24; revised version received 30.05.24; accepted 15.07.24; published 27.08.24.

Please cite as:
Suffoletto B
Deceptively Simple yet Profoundly Impactful: Text Messaging Interventions to Support Health
J Med Internet Res 2024;26:e58726
URL: https://www.jmir.org/2024/1/e58726
doi:10.2196/58726
PMID:39190427

©Brian Suffoletto. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Health Applications (DiGAs) on a Fast Track: Insights From a Data-Driven Analysis of Prescribable Digital Therapeutics in Germany From 2020 to Mid-2024

Moritz Goeldner[1], Prof Dr; Sara Gehder[1], MSc

Working Group for Data-Driven Innovation, Hamburg University of Technology, Hamburg, Germany

Corresponding Author:
Moritz Goeldner, Prof Dr
Working Group for Data-Driven Innovation
Hamburg University of Technology
Am Schwarzenberg-Campus 4
Hamburg, 21073
Germany
Phone: 49 40428784777
Email: moritz.goeldner@tuhh.de

Abstract

Background: This study aimed to analyze the rapidly evolving ecosystem of digital health applications (*Digitale Gesundheitsanwendung*; DiGAs) in Germany, spurred by the 2019 Digital Healthcare Act. With over 73 million people in Germany now having access to DiGAs, these prescribable digital health apps and web-based applications represent a substantial stride in health care modernization, supporting both patients and health care providers with digital solutions for disease management and care improvement.

Objective: Through a data-driven approach, this research aimed to unpack the complexities of DiGA market dynamics, economic factors, and clinical evidence, offering insights into their impact over the past years.

Methods: The analysis draws from a range of public data sources, including the DiGA directory, statutory health insurance reports, app store feedback, and clinical study results.

Results: As of July 1, 2024, there are 56 DiGAs listed by the Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte), divided into 35 permanently and 21 preliminarily listed applications. Our findings reveal that a majority of DiGAs extend beyond the intended 1-year period to achieve permanent listing, reflecting the extensive effort required to demonstrate clinical efficacy. Economic analysis uncovered a dynamic pricing landscape, with initial prices ranging from approximately €200 to €700 (€1=US $1.07), averaging at a median of €514 for a 3-month DiGA prescription. Following negotiations or arbitration board decisions, prices typically see a 50% reduction, settling at a median of €221. Prescription data offer valuable insights into DiGA acceptance, with total prescriptions jumping from around 41,000 in the first period to 209,000 in the latest reporting period. The analysis of the top 15 DiGAs, representing 82% of the total prescriptions, shows that these best-performing apps receive from a minimum of 8 to a maximum of 77 daily prescriptions, with native apps and early market entrants achieving higher rates. Clinical evidence from all 35 permanently listed DiGAs indicates a uniform preference for randomized controlled trials to validate primary end points, with no noteworthy use of alternative study designs encouraged in the Digital Healthcare Act and related regulations. Moreover, all evaluated DiGAs focused on medical benefits, with health status improvement as a key end point, suggesting an underuse of patient-relevant structural and procedural improvement in demonstrating health care impact.

Conclusions: This study highlights the growth and challenges within the DiGA sector, suggesting areas for future research, such as the exploration of new study designs and the potential impact of patient-relevant structural and procedural improvements. For DiGA manufacturers, the strategic advantage of early market entry is emphasized. Overall, this paper underscores the evolving landscape of digital health, advocating for a nuanced understanding of digital health technology integration in Germany and beyond.

(J Med Internet Res 2024;26:e59013) doi:10.2196/59013

XSL•FO
RenderX

KEYWORDS

digital health application; DiGA; data-driven analysis; clinical evidence; health economics; positive care effect; medical benefit; patient-relevant structural and procedural improvements; pSVV; digital health care act

Introduction

Background

While Germany continues to navigate the complex challenges of digital transformation within its health care system, the reimbursement of digital medical devices by statutory health insurance represents a pivotal milestone in this journey. The 2019 Digital Healthcare Act (Digitale-Versorgung-Gesetz) marks a substantial legislative milestone, incorporating digital health applications (German: *Digitale Gesundheitsanwendung*; DiGA) into the benefits catalog of statutory health insurance in Germany [1-4]. Since September 2020, the first DiGAs have been officially listed in the DiGA directory of the Federal Institute for Drugs and Medical Devices (German: Bundesinstitut für Arzneimittel und Medizinprodukte; BfArM), thereby becoming accessible to more than 73 million individuals in Germany. DiGAs are digital medical devices of risk classes I and IIa, or IIb (a recent regulatory update—the Digital Law—has made class IIb medical devices eligible for DiGA listing, albeit imposing more stringent criteria for demonstrating evidence and proving the positive care effect) that are primarily based on digital technologies and aim at assisting both insured individuals and health care professionals by enhancing the detection, monitoring, treatment, or alleviation of diseases, as well as the identification of, treatment of, alleviation of, or compensation for injuries or disabilities. DiGAs listed in the DiGA directory may be prescribed by physicians and psychotherapists or directly provided by the statutory health insurances under certain requirements [5,6].

Compared with existing approval procedures in health care, the legislator has integrated 3 regulatory innovations. First, a 12-month trial phase allows DiGA manufacturers to initially introduce their DiGA to the market on a provisional basis and subsequently provide the final proof of effectiveness [6]. During this phase, the price for using the DiGA can be set by the manufacturer within defined maximum limits. Second, the approval was designed as a fast-track-procedure, meaning the BfArM's review of the application takes a maximum of 3 months. Third, the concept of "therapeutic benefit," crucial in drug approval, has been expanded to the more comprehensive concept of a "positive care effect" (German: *Positiver Versorgungseffekt*; pVE). This can be demonstrated either by proving a "medical benefit" (German: *medizinischer Nutzen*), that is, an improvement in health status or enhancement of quality of life, or a "patient-relevant structural and procedural improvement" (German: *patientenrelevante Struktur- und Verfahrensverbesserung*; pSVV) focusing on optimizing the interaction between patients and health care providers as well as supporting the patient [6-8]. To prove the positive care effect, each DiGA requires conducting a clinical trial within Germany. In addition, a preliminary study demonstrating the DiGA's basic features is required before preliminary listing as well.

After almost 4 years into the implementation of the DiGA framework, we seek to investigate the evolving landscape of the DiGA market. Other European countries, such as France and Belgium, are adopting comparable approaches for digital medical device reimbursement [9,10], though they have not reached the same scale yet. Thus, our primary aim was to unpack the dynamics that have shaped this market in Germany, quantifying the economic aspects of DiGAs, and identifying the characteristics of successful DiGA implementations based on the data available. Moreover, this study aimed to provide a comprehensive understanding of the clinical evidence surrounding DiGA that has been generated and was assessed by BfArM.

Objectives

To achieve these objectives, we adopt a data-driven approach, leveraging publicly available data sets to craft a coherent and detailed picture of the DiGA ecosystem. This methodology allows us to explore various facets of the DiGA landscape, from market trends and economic impact to clinical effectiveness and regulatory outcomes. Through this analysis, we aimed to uncover insights into the factors that contribute to the success of DiGA, the challenges they face in the market, and the broader implications for health care professionals, policy makers, and patients within the (digital) health care system.

Methods

To provide a comprehensive analysis of the DiGA ecosystem in Germany, our study leveraged publicly available data sources. The following data sources were obtained during the analysis.

DiGA Directory Data

Access to this data was facilitated via the DiGA application programming interface provided by BfArM, enabling up-to-date insights into the range and characteristics of each individual DiGA ever listed. The data were accessed using a custom-developed script using the cloud–based integrated development environment Replit.

Annual Reports of Statutory Health Insurance Funds

We analyzed 3 annual reports on DiGAs from the National Association of Statutory Health Insurance Funds (German: GKV-SV), which provided valuable information on the adoption rates and prescription trends for DiGAs [11-13]. It is important to consider that this data set does not include privately insured patients who use DiGAs, potentially underrepresenting total use. In Germany, about 10% of the population is privately insured [14].

App Store Ratings

To gauge user reception and satisfaction, we accessed app rating data from the Apple App Store (Apple, Inc) [15-17]. This was achieved through an application programming interface and a custom-developed Google Sheets (Alphabet, Inc) script, allowing for the automated collection of ratings data. Similarly,

XSL•FO
RenderX

app ratings from the Google Play Store were manually accessed to complement the data gathered from the Apple App Store, ensuring a comprehensive view of user feedback across major mobile platforms [18].

Accessing Clinical Studies

We conducted manual searches for all studies pertaining to permanently listed DiGAs that were identified as relevant for the "assessment decision by the BfArM" according to the DiGA directory. These studies were registered either in the German Clinical Trials Register, ClinicalTrials.gov (maintained by the National Library of Medicine at the National Institutes of Health), or the International Standard Randomised Controlled Trial Number registry (maintained by BioMed Central, part of Springer Nature). In addition, respective publications, study reports and protocols for each study were reviewed to assess the clinical evidence base supporting the use of these DiGAs. Our analysis encompassed a thorough review of study designs, measuring instruments, and dropout rates, analyzing the ratio of observation duration, sample size, and the length of the trial phase, and comparing age and gender of the participants with demographic data.

To ensure a uniform perspective across the collected data, our analysis specifically used information from the DiGA directory as of July 1, 2024. This approach provided us with a comprehensive data set encompassing 3 full years, in addition to approximately 3 months and a few days, since the first DiGA was introduced in late September 2020 and 6 months in 2024. It is important to note, however, that the data from the GKV-SV, particularly concerning the number of prescriptions, were segmented into 3 terms, with each term spanning from October to September of the following year. This discrepancy in the data collection periods has been carefully addressed in our analysis to ensure clarity and accuracy in our findings.

All data were consolidated into a single database. Basic calculations were performed using SPSS Statistics (version 29; IBM, Corp), and visualizations were created with flourish.studio (Canva UK Operations Ltd) and ThinkCell (think-cell Software GmbH). Given the nascent state of the DiGA market and the limited number of apps available, our study did not use statistical analysis. The market's current size does not support such an approach; however, this is anticipated to change as the market matures. Consequently, our results are primarily descriptive, offering insights into the present landscape of DiGAs, their use, user feedback, and the clinical evidence supporting their efficacy and utility.

Ethical Considerations

Ethics approval does not apply to this study, as we evaluated publicly available, deidentified data on DiGA prescriptions [11-13] and did not involve any human participants during data collection.

Results

Descriptive Data on DiGAs

As of July 1, 2024, a total of 64 DiGAs have been approved by the BfArM. Currently, 35 DiGAs are permanently listed, 21 are preliminarily listed (still required to submit clinical evidence for permanent listing), and 8 DiGAs have been removed from the directory. Among the 35 permanently listed DiGAs, 11 (31%) were initially approved for permanent listing, whereas 24 (69%) initially received preliminary approval and were later granted permanent listing. While the regulator allows up to 1 year for this process, an extension to 2 years is permissible (Figure 1).

Figure 1. Overview of digital health application (Digitale Gesundheitsanwendung; DiGA) categories and preliminary or permanent listings. Only 1 therapeutic category is indicated for each DiGA.

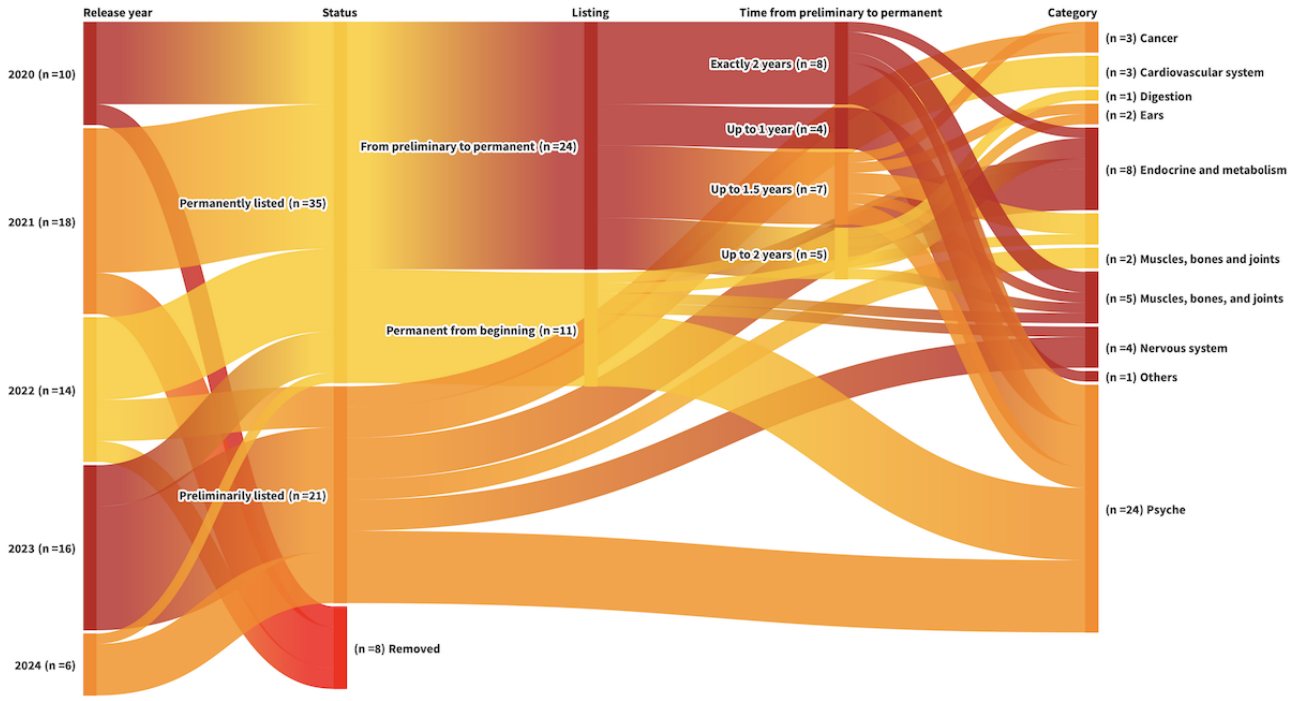

XSL•FO
RenderX

Our analysis indicated that most DiGAs necessitate considerably longer than 1 year to complete this transition, with one-third precisely requiring 2 years (including the period for BfArM to assess the submitted data). Moving forward, we have excluded the 8 removed DiGAs from all subsequent analyses, concentrating on the 56 DiGAs that remained accessible.

The therapeutic areas for DiGAs span 12 categories as defined by the BfArM, with the largest being the *psyche* category. This category includes 26 DiGAs designed to address various psychological conditions, predominantly through cognitive behavioral therapy. The *endocrine and metabolism* category follows with 8 DiGAs. The *muscles, bones, and joints* category is represented by 7 DiGAs, and *nervous system* with 4 DiGAs (notably, 5 DiGAs are linked to 2 categories each, resulting in a total of 61 categorizations).

The new concept of positive care effect is the pivotal point in the assessment of clinical evidence for DiGA. Despite the regulator introducing a new framework, the pSVV, its uptake remained modest. Among the 56 DiGAs, 55 (98%) had a medical benefit as primary end point, while only 1 (2%) DiGA (ProHerz, ProCarement GmbH) had a pSVV as primary end point. In addition, 10 (18%) DiGAs with a medical benefit as primary end point incorporated pSVV as an additional end point in their studies.

To qualify for DiGA listing, a product must be certified as a medical device. Among the 56 DiGAs, 32 (57%) were still recognized under the Medical Device Directive (MDD) as class I medical devices—a regulation that officially expired in May 2021, with a current transition period in effect. The remaining 24 (43%) DiGAs complied with the Medical Device Regulation (MDR), the updated European standardized regulation where the assessment process for digital medical devices has changed (refer to rule 11 of MDR), resulting in a more rigorous review of their risk classes [19]. Within this group, 21 (87%) were classified as class I medical devices, and 3 (13%) were class IIa medical devices, which necessitate certification from a notified body for approval. Until July 2024, no class IIb medical device was certified as a DiGA, because applications for DiGAs in this risk class only became possible after the new regulation took effect at the end of March 2024. While MDR certification certainly enhances the safety of (digital) medical devices by requiring premarket testing and continuous postmarket surveillance of medical devices, it also represents a significant financial and time-consuming burden, particularly when it comes to class IIa or IIb medical devices because notified bodies are a veritable bottleneck within this process [19,20]. Variations were observed in relation to the federal states hosting the headquarters of DiGA manufacturers. From Hamburg, 20 DiGAs (from 8 manufacturers) emerge, with an approximately equal distribution between MDR (11 DiGAs) and MDD (9 DiGAs) classifications. In Berlin, 11 DiGAs (from 6 manufacturers) fall exclusively under the category of MDD-regulated medical devices, with only 1 DiGA listed in 2024 being classified under the MDR. Such disparities are not evident in other federal states. While we found no reason for this discrepancy in our data, we assumed that the state authorities in the city-states of Berlin and Hamburg have different approaches to accepting class I MDR medical devices under rule 11 of MDR.

DiGAs are mandated to use digital technologies, which means they are not restricted to app formats alone. However, our observations show that 27 DiGAs, originating from 24 manufacturers, were available as native apps (unlike web applications, native apps are developed for a specific platform, such as the Apple App Store or the Google Play Store) in both the Apple App Store and Google Play Store. In addition, 10 DiGAs, from 6 manufacturers, provided access through both app and web application formats, whereas 19 DiGAs, from 7 manufacturers, were exclusively accessible via websites.

Analysis of Initial and Final Prices

During the first year following the initial listing, whether preliminary or permanent, DiGA manufacturers have considerable freedom in setting prices, subject to certain criteria based on the demonstrated positive care effect (medical benefit, pSVV, or both), the *International Classification of Diseases, Tenth Edition* (*ICD-10*) category, and the pricing of existing DiGAs within that category [2,21]. Upon final listing—or after 12 months for DiGAs already permanently listed—price negotiations with the GKV-SV commence, with 3 meetings planned to reach an agreement. Should these negotiations fail, an arbitration board is engaged to reach a definitive decision. Thus, not all permanently listed DiGAs have an established final price yet. As of July 1, 2024, among the 35 permanently listed DiGAs in our data set, 30 (86%) had received their final pricing. These final prices were markedly less than the initial prices, spanning from 29% to 93% of the initial price, averaging at 50% of the initial amount.

DiGA manufacturers also have the option to adjust (ie, increase) the price once within the first year, although this occurs infrequently and was not further considered in our data set. In addition, our analysis did not account for varying prices between the initial and subsequent prescriptions, as this pricing strategy is seldom implemented.

Typically, a DiGA is prescribed for 3 months of use. To date, only 1 preliminary listed DiGA, levidex, offered a lifetime license at a substantially higher cost of €2077.40 (€1=US $1.07). This DiGA will be excluded from the following analysis. The prices for the rest of the DiGAs were effective for a period of 3 months and ranged from approximately €200 to €700, with a median price of €535.49. One can observe that the initial pricing has trended upward over time (Figure 2). The median was €487 in the first year, increased to €567 in 2022, and adjusted to €547 in 2024 (Table 1).

XSL•FO
RenderX

Figure 2. Initial prices of digital health applications (Digitale Gesundheitsanwendung [DiGA]; n=55) plotted per year of initial listing. One preliminary listed DiGA (levidex) with a lifetime license was excluded.

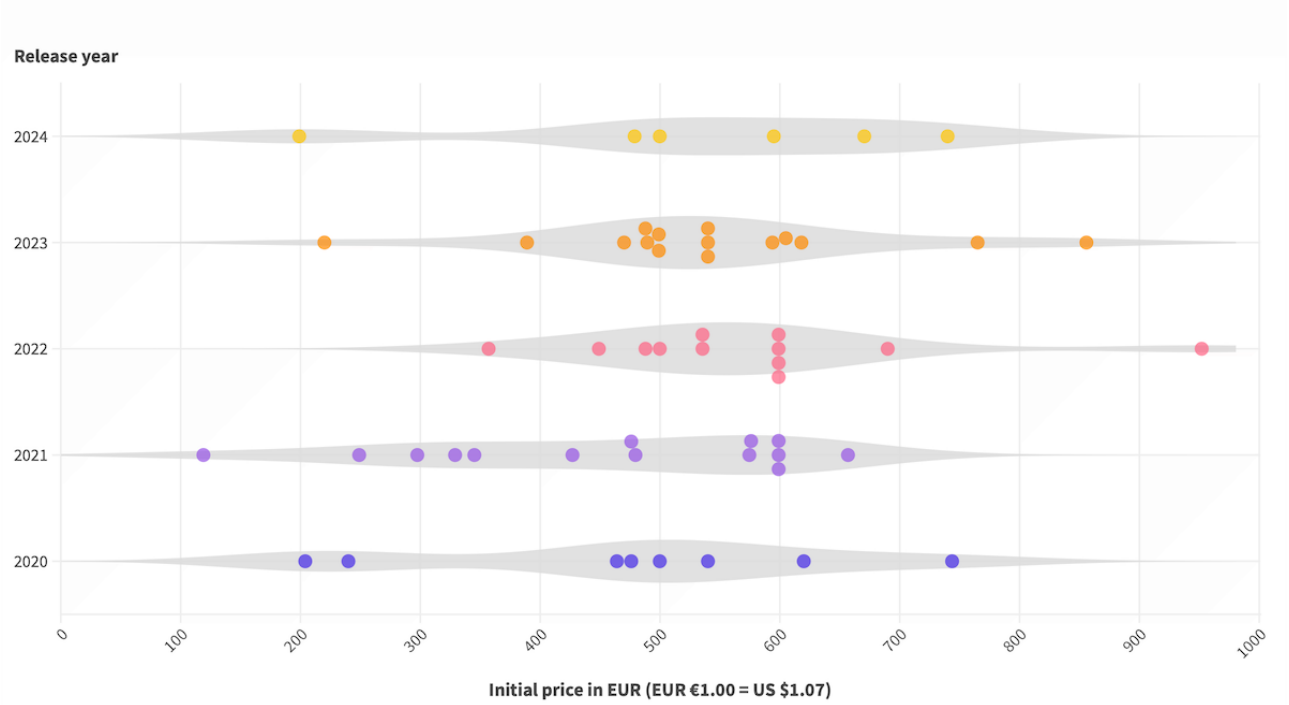

Table 1. Initial and final prices of digital health applications (DiGAs) per release year. One preliminary listed DiGA (levidex) with a lifetime license was excluded.

Release year	Price (€[a]), mean (SD)	Price (€), median (IQR)
Initial prices		
2020 (n=8)	473 (180)	488 (296-600)
2021 (n=14)	452 (162)	478 (321-599)
2022 (n=12)	575 (147)	567 (491-599)
2023 (n=15)	541 (147)	540 (488-605)
2024 (n=6)	531 (190)	547 (409-688)
All years (n=55)	514 (162)	535 (464-599)
Final prices		
2020 (n=8)	219 (16)	219 (209-229)
2021 (n=13)	219 (21)	223 (204-235)
2022 (n=6)	233 (9)	232 (224-238)
2023 (n=3)	228 (6)	232 (221-232)
All years (n=30)	221 (17)	224 (216-235)

[a]€1=US $1.07.

Similarly, the prices negotiated or determined by the arbitration board have seen a gradual rise, moving from a median of €219 for DiGAs initially listed in 2020 to a median of €232 for those listed in 2023 for the first time (refer to Figure 3 for an overview). However, the data reveal low variance from the mean values of final prices.

XSL•FO
RenderX

Figure 3. Final prices of digital health applications (Digitale Gesundheitsanwendung [DiGA]; n=30) plotted per year of initial listing.

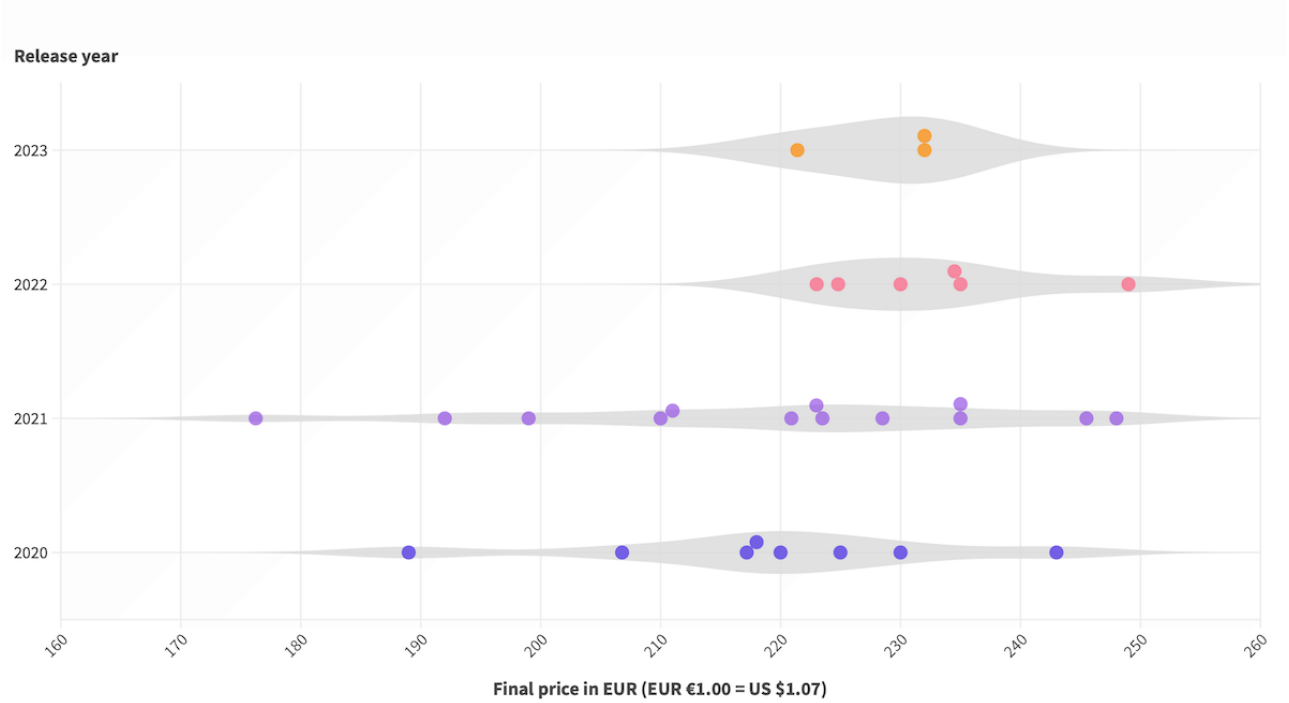

In our subsequent analysis, we investigate whether the chosen positive care effect (ie, medical benefit or pSVV) impacts pricing. Initial prices showed a statistically insignificant difference (we used a Mann-Whitney *U* test as the data is not normally distributed; *P*=.69) in median values: €518 (IQR €398-€599) for DiGAs without pSVV compared with €479 (IQR €449-€576) for DiGAs with pSVV (Figure 4).

Figure 4. Initial prices of digital health applications (Digitale Gesundheitsanwendung [DiGA]; n=55) and final prices (n=30) with and without patient-relevant structural and procedural improvements (Patientenrelevante Struktur- und Verfahrensverbesserungen; pSVV).

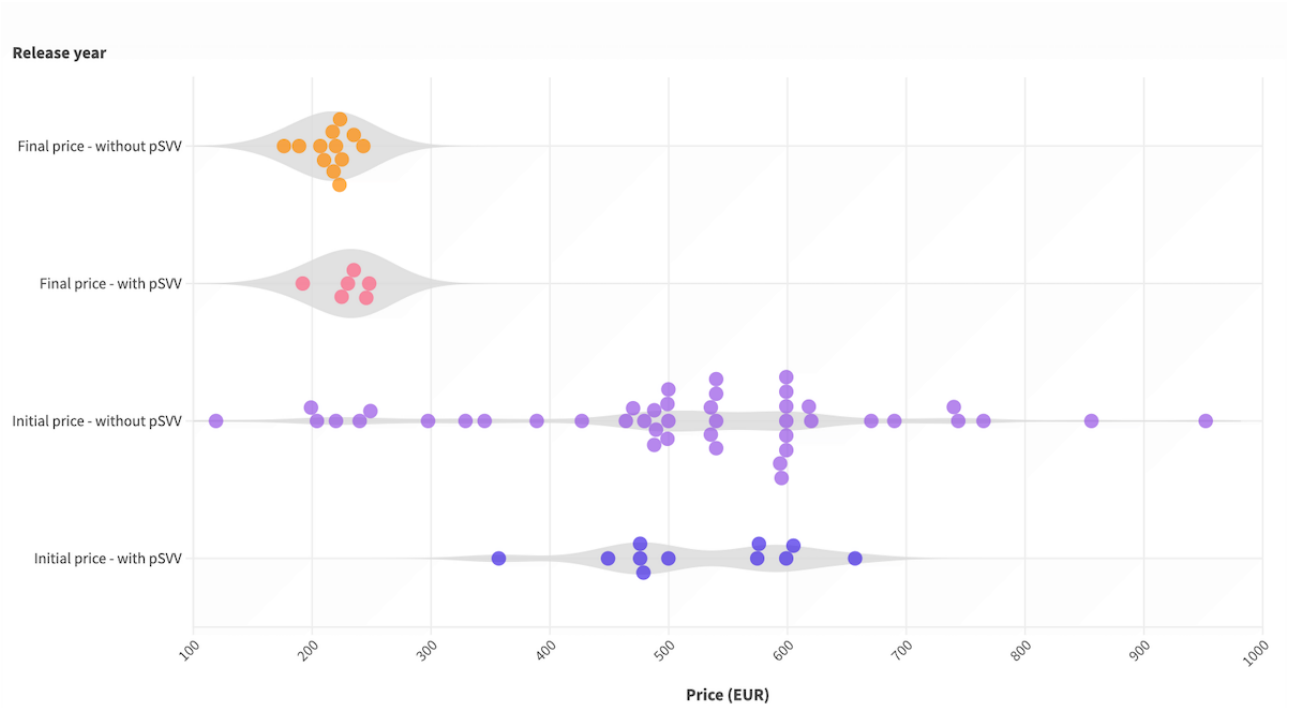

For the final prices, this insignificant difference was reversed; the median was €223 (IQR €212-€232) for DiGAs without pSVV and €232 (IQR €217-€246) for those that include a pSVV.

Analysis of Prescriptions

Although prescription numbers for DiGAs likely offer some of the most compelling insights, these data are not publicly available for all DiGAs. The only official figures are annually published by the GKV-SV, which releases rounded prescription

XSL•FO
RenderX

numbers. The total number of prescriptions was approximately 41,000 in the first reporting period (September 2020-September 2021), increased to 124,000 in the subsequent period (October 2021-September 2022), and reached 209,000 prescriptions in the latest period (October 2022-September 2023). Altogether, exactly 374,377 DiGA prescriptions were recorded from the initial availability of the first DiGA in late September 2020 until September 30, 2023. This total number also includes prescriptions of now delisted DiGAs.

However, although data on individual DiGA prescriptions were accessible for all DiGAs exceeding 100 total prescriptions in the first reporting period, this criterion tightened in later periods. In the subsequent period, only DiGAs with over 1000 total prescriptions throughout their life span were featured. The threshold was further elevated in the most recent report to include solely DiGAs with more than 5000 prescriptions over their life span. By the end of September 2023, 15 DiGAs surpassed the threshold of 5000 prescriptions over their lifetime. Hence, our subsequent analyses focused exclusively on these top 15 DiGAs. When available, data from 2021 and 2022 were also incorporated. To ensure consistency and comparability, we calculated the average daily prescriptions for these 15 DiGAs by taking the number of days listed in the directory from the first listing to the end of the reporting period of the GKV-SV at the end of the third quarter of 2023.

Within the top 15 DiGAs, we observed considerable variability in prescriptions, ranging from 77 per day for zanadio (Sidekick Health Germany GmbH) to 8 per day for Invirto (Sympatient GmbH). Among these top 15, 8 (53%) DiGAs were released in the first 12 months following the DiGA directory's launch, 6 (40%) during the second period, and only 1 (7%) DiGA (Endo-App, Endo Health GmbH) during the latest reporting period by the GKV-SV. The average year-on-year growth for these apps across their individual first 2 reporting period was approximately 270%. However, between the second and third reporting periods, the growth rate for these apps decreased to 140% (Figure 5). For example, the growth rate for zanadio was 237% between the first two periods and 28% between the second two periods. Similarly, the growth rates for Invirto were 238% and 86%, respectively (Figure 5). In total, 308,000 prescriptions were made for these top 15 DiGAs, a substantial share (82%) of the total 374,377 prescriptions.

Figure 5. Average daily prescriptions per digital health application (Digitale Gesundheitsanwendung [DiGA]; n=15).

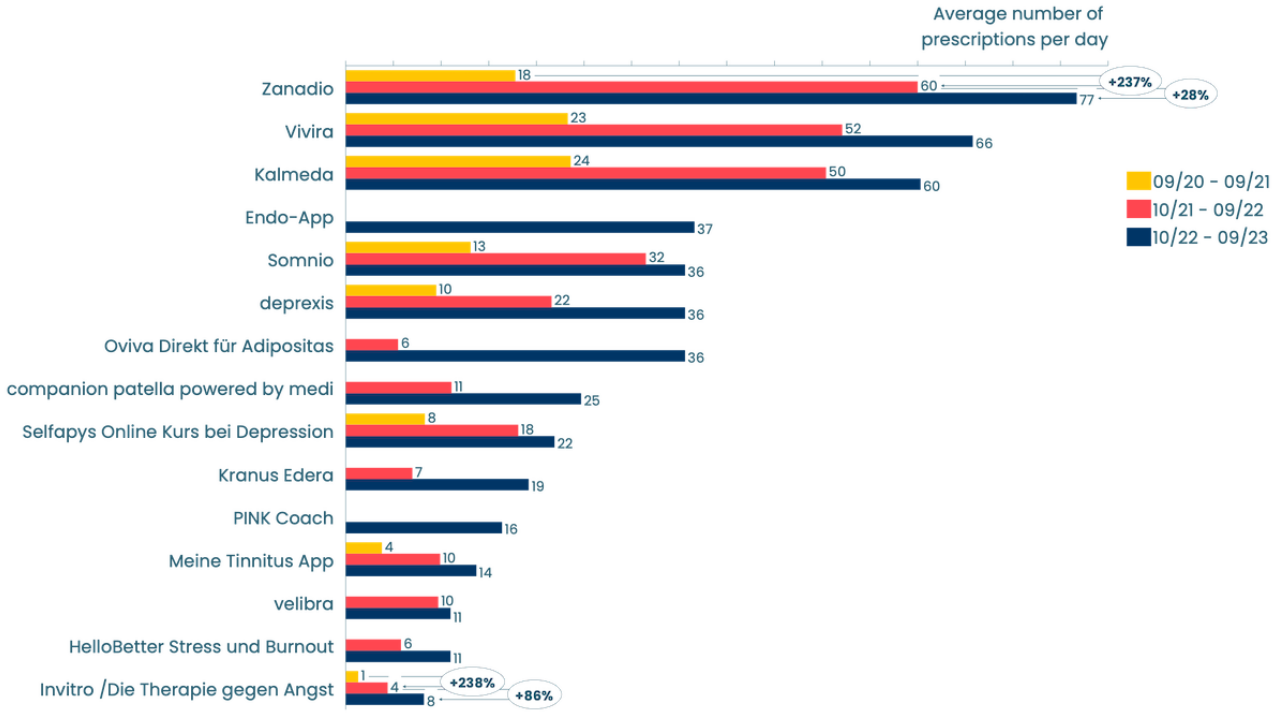

DiGAs are mandated to use digital technologies, which means they are not restricted to app formats alone. However, our observations showed that 27 DiGAs, originating from 24 manufacturers, were available as native apps in both the Apple App Store and Google Play Store. In addition, 10 DiGAs from 6 manufacturers provided access through both app and web-app formats, whereas 19 DiGAs from 7 manufacturers were exclusively accessible via websites. Within the top 15 DiGAs, it is notable that 9 DiGAs available solely as apps accounted for 210,000 prescriptions. In contrast, 2 DiGAs offering both app and web application formats garnered 46,000 prescriptions, and 4 DiGAs accessible only through web applications achieved 52,000 prescriptions.

While DiGA prescriptions are commonly associated with those issued by medical doctors or psychotherapists, there is an alternative pathway for patients to receive a DiGA directly from their statutory health insurance. Typically, health insurance providers may request proof that the patient has the condition targeted by the DiGA, though this is not an official requirement. In exploring the ratio of DiGAs distributed by health insurances versus total prescriptions, there are notable differences between the different DiGAs. Notably, Endo-App and Kranus Edera (Kranus Health GmbH), both categorized under *genitals*,

XSL•FO
RenderX

kidneys, and urinary tract, lead health insurance distributions at 40% and 28%, respectively, followed by Oviva Direkt (Oviva AG) at 23%. Conversely, at the spectrum's lower end, Vivira (ViViRA Health Lab GmbH) and Companion Patella (medi GmbH), focusing on muscles, bones, and joints, each saw only 5% of their distributions through statutory health insurance (Table 2).

Table 2. Digital health application (DiGA) prescriptions based on GKV-SV reporting [11-13].

DiGA name	Prescriptions from September 2020 to September 2021 (n=28,172), n (%)	Prescriptions from October 2021 to September 2022 (n=135,331), n (%)	Prescriptions from October 2022 to September 2023 (n=239,046), n (%)	Authorization by statutory health insurance (%)
Zanadio	6100 (21.65)	21,900 (16.18)	28,000 (11.71)	11
Vivira	8000 (28.4)	19,000 (14.04)	24,000 (10.04)	5
Kalmeda	8600 (30.53)	18,400 (13.6)	22,000 (9.2)	8
Deprexis	2100 (7.45)	7900 (5.84)	13,000 (5.44)	8
Endo-App	__[a]	—	13,000 (5.44)	40
Oviva Direkt	—	2000 (1.48)	13,000 (5.44)	23
Somnio	4500 (15.97)	11,500 (8.5)	13,000 (5.44)	11
Companion Patella	—	4000 (2.96)	9000 (3.76)	5
Selfapys Depression	2400 (8.52)	6600 (4.88)	8000 (3.35)	11
Kranus Edera	—	2000 (1.48)	7000 (2.93)	28
PINK Coach	—	—	6000 (2.51)	13
velibra	1400 (4.97)	3600 (2.66)	5000 (2.09)	8
HelloBetter Stress und Burnout	—	2000 (1.48)	4000 (1.67)	13
Meine Tinnitus App	—	2000 (1.48)	4000 (1.67)	6
Invitro	400 (1.42)	1600 (1.18)	3000 (1.18)	10

[a]Not available.

Finally, we also wanted to take into account the user's reviews in the app stores as well as the information if the DiGA was the first within its *ICD-10* category on the market. Therefore, we can only consider the 11 (73%) out of the top 15 DiGAs that are listed as a native smartphone app in the 2 major app stores. App store reviews were collected from both, Apple App Store and Google Play Store and combined, taking the average review (on a 5-star scale) and the number of reviews into account. Again, we used average prescriptions per day to have better comparability.

Unsurprisingly, it was observed that apps with a higher number of prescriptions tended to garner more reviews. Regarding the quality reflected in a 5-star rating system, Vivira, Oviva Direkt, somnio (mementor DE GmbH), and zanadio stood out with the highest reviews, scoring 4.6, 4.5, 4.5, and 4.4, respectively (Figure 6). On the lower spectrum, Kalmeda (3.3; Pohl-Boskamp Digital Health GmbH) and Selfapy (3.4; Selfapy GmbH; Selfapy provides a single app encompassing all their DiGAs, meaning reviews cannot be uniquely attributed to the app Selfapy Depression alone) were notable.

In addition, we observed that DiGAs received more average prescriptions per day when they were first to the market than when they were second in various categories (Figure 7). This pattern was evident with zanadio leading ahead of Oviva Direkt, Kalmeda outpacing Meine Tinnitus App (BAYOOCARE GmbH), and—on a lower level—velibra (GAIA AG) ahead of Invirto. Only Deprexis (GAIA AG) performed slightly better than Selfapy Depression in terms of prescriptions, but both apps were—similar to velibra and Invirto—released only 2 months apart. While many factors might have contributed to this trend, the first-to-market advantage is worth noting, particularly if there is a long period between being first and second to market.

XSL•FO
RenderX

Figure 6. Average app reviews in app stores (x-axis) and daily prescriptions of digital health applications (Digitale Gesundheitsanwendung [DiGA]; y-axis) among the top 15 that are available as native apps. The size of the bubbles indicates the number of app reviews.

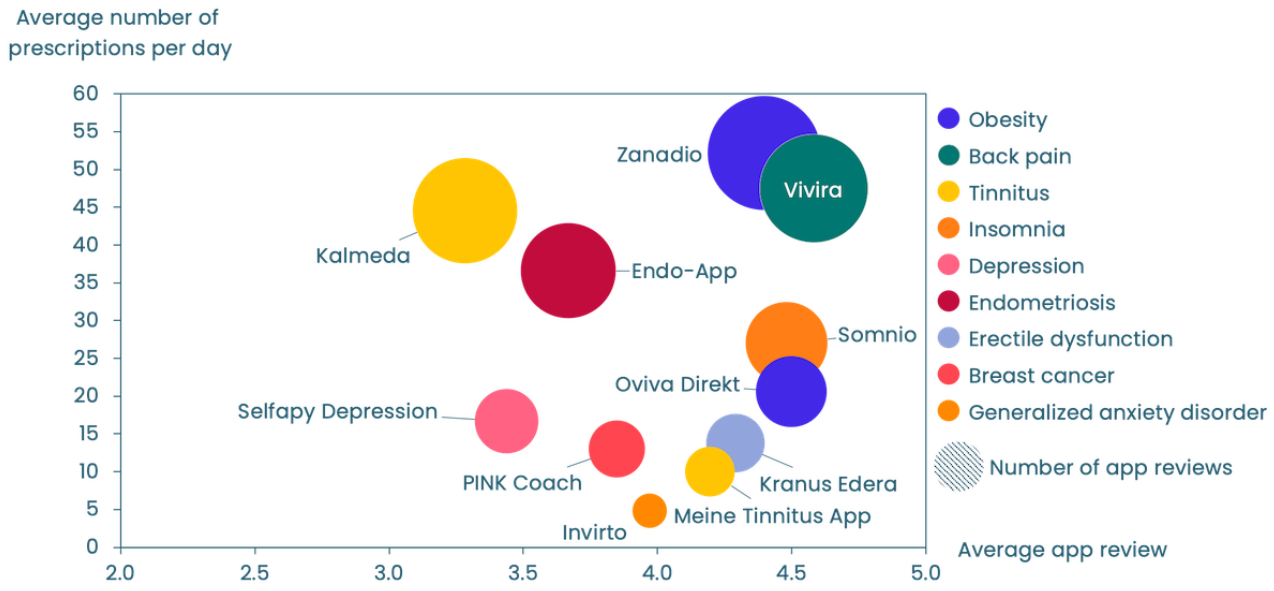

Figure 7. Average prescriptions per day (y-axis) and launch date (x-axis). The bubble size is proportionate to the average number of prescriptions per day.

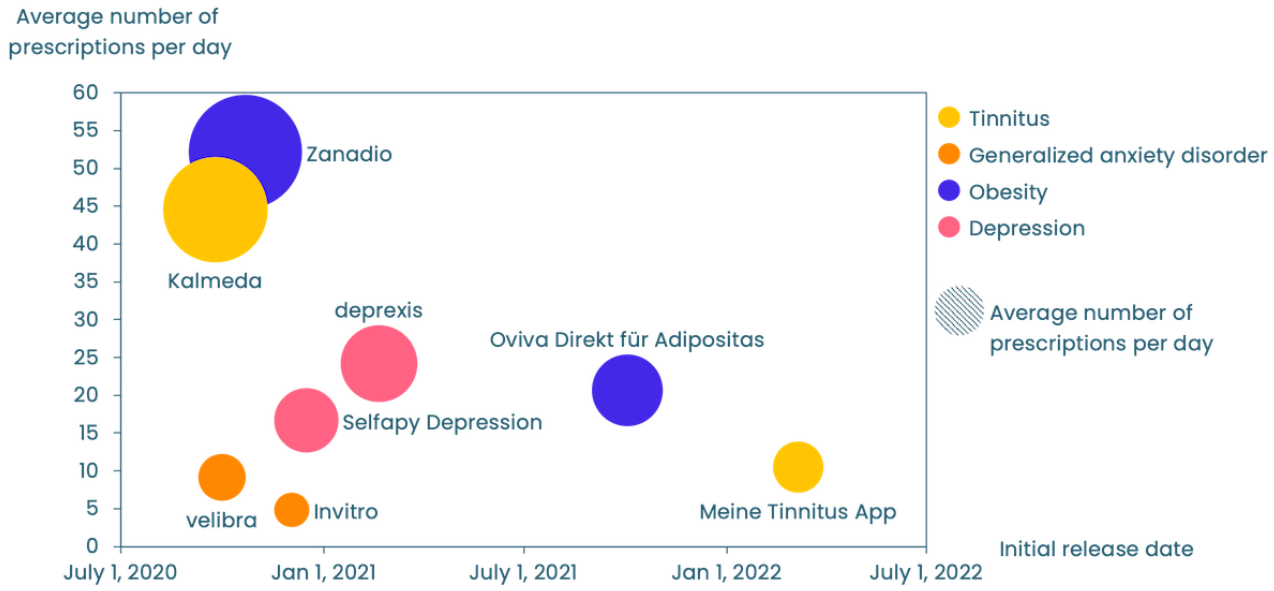

Analysis of Clinical Trial Data

To demonstrate a pVE for permanent listing of a DiGA, manufacturers must conduct a clinical trial in Germany that shows one or more end points related to the positive care effect. In this section, we analyzed all permanently listed DiGAs in our sample (refer to Multimedia Appendix 1 for an overview).

Among the 35 DiGAs analyzed, all have proven at least 1 medical benefit, with each opting for *improvement of the state of health* as primary end point. Notably, 23 (66%) of these 35 DiGAs demonstrated this single positive care effect exclusively. Three more DiGAs demonstrated both, *improvement of the state of health* and *improvement in quality of life*. The remaining 9 DiGAs used pSVV as an additional end point in their studies.

In total, 5 DiGAs have evidenced a pSVV along with *improvement of the state of health*, covering aspects such as *health competence*, *reduction of therapy-related efforts and burdens on patients and their relatives*, *alignment of treatment with guidelines and recognized standards*, and *patient sovereignty*. Furthermore, 3 more DiGAs were able to show 2 medical benefits along with a pSVV (with all 3 opting for *patient sovereignty*), and finally, 1 DiGA demonstrated 2 medical benefits in conjunction with 2 additional pSVV (ie, *health competence* and *coping with illness-related difficulties in everyday life*).

Following the permanent listing of a DiGA, manufacturers are required to make available to the public, within 1 year, the study results demonstrating the pVE of the DiGA. As of July 2024, for all 35 permanently listed DiGAs in our data set, at least 1

XSL•FO
RenderX

registered clinical trial was mentioned in the "assessment decision of BfArM" of the DiGA directory. Furthermore, 1 of these 35 DiGAs (Deprexis) cited 2 clinical trials—hence, the subsequent analyses refer to a total of 36 registered trials.

All 36 trials were registered as randomized controlled trials. The study type was registered as interventional across all 36 studies. In 35 (97%) of the studies, the allocation of participants to either the control group or the intervention group was conducted through "parallel distribution," while 1 (3%) study used a "factorial" design. Furthermore, 15 (42%) studies used masking, 19 (53%) explicitly did not, and 2 (6%) studies did not clarify their use of masking. Where implemented, masking usually concealed the group affiliation (ie, control or intervention) from either the outcome assessors, the statisticians, or both. Exceptionally, 1 study kept patients unaware of their group membership.

In the 36 conducted trials, 22 different primary end points (such as *depression symptoms*, *tinnitus exposure*, or *pain intensity*) were used. To evidence these primary end points, 29 different measurement instruments (mostly patient-reported outcome measures, but also measures such as weight or HbA_{1C}) were used. Furthermore, 6 end points were used by >1 DiGA manufacturer—*depression symptoms*, *anxiety symptoms*, *pain intensity*, *body weight*, *reduction of disease-specific symptom severity*, and *tinnitus burden*.

Yet, these primary end points were measured differently by DiGA manufacturers—only within the end points *depression symptoms* and *tinnitus burden*, 2 psychometric tests for measuring the severity of depression (Beck Depression Inventory and Patient Health Questionnaire) and a Tinnitus Questionnaire (Mini-TF-12) were used 2 or 3 times, respectively. Thus, most DiGA manufacturers opted for different measures for the primary end point, even if the same end point has been previously assessed by another DiGA. For secondary end points, we see a variety of measures that were used across DiGA manufacturers.

In total, 22 (61%) of the 36 studies used a waitlist control group, while 12 (33%) studies used *standard of care* as the control. Two DiGAs implemented control groups using apps with reduced functionality (sham DiGAs), featuring versions of the actual DiGA that provided only medically useful tips or survey instruments. There was no obvious pattern linking the design of these control groups with the indication area of the DiGA (Figure 8).

Figure 8. Control groups in clinical studies (n=36) per disease category. Two digital health applications (Digitale Gesundheitsanwendung; DiGAs) are associated with 2 disease categories. In this figure, both are located within the muscle, bones, and joints category; the second category was psyche (waitlist) and injuries (standard of care).

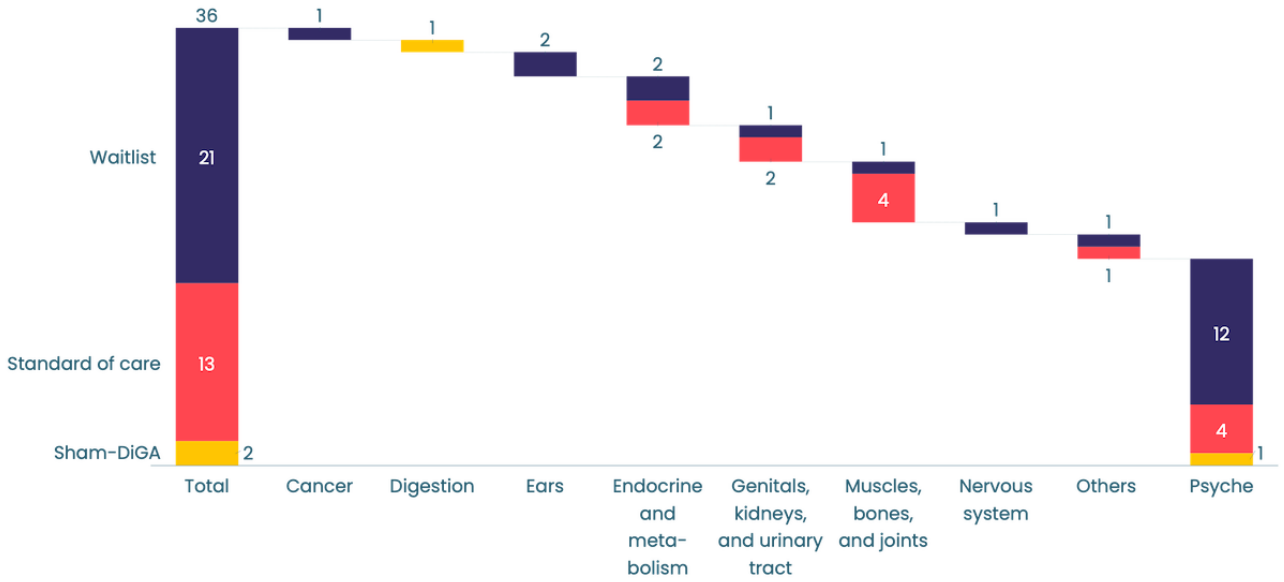

In total, 30 (86%) out of the 35 permanently listed DiGAs recommend a use duration of 3 months. Furthermore, 20 (57%) of these DiGAs correspondingly used an observation period of 3 months in their studies, 5 (14%) of them used a shorter observation period, and 5 (14%) used a longer one. In addition, 2 (6%) DiGAs suggested a use duration exceeding 3 months; 1 (3%) mirrored this with a longer study period, whereas the other did not. Two DiGAs advised a shorter use duration than 3 months and aligned their study duration accordingly.

The description of the study design varied widely among the 36 trials reviewed. A total of 13 (37%) DiGAs included detailed information with a link to a scientific publication in the study registry. Furthermore, 8 (23%) DiGAs provided a study report within the registry and 3 (9%) DiGAs published their study protocol. In addition, for 2 (6%) DiGAs, links to scientific publications were available in the DiGA directory.

For all 36 studies, the sample size information was provided, ranging from as low as 56 participants to as high as 1237, with the median being 239 participants. There has been a noticeable increase in sample size over the years: DiGAs that were permanently listed within the first 3 months in 2020 had a median of 139 (IQR 56-139) participants (combining both intervention and control groups), which rose to a median of 264 (IQR 211-810) participants in 2021, decreased slightly to 187

XSL•FO
RenderX

(IQR 151-256) in 2022, and then increased again to 253 (IQR 161-387) in 2023 and 220 (IQR 184-386) in 2024 (Figure 9).

Data on the duration of the clinical study were accessible for 30 trials, spanning from 4 months at the minimum to 54 months at the maximum, with a median duration of 17.5 months. Data on the dropout rate were available for 27 of the studies, indicating a median dropout rate of 17.5% among participants. Gender distribution was reported in 22 studies, revealing that approximately 63% of all study participants for which the gender was reported were female, 37% male, and less than 1% identified as diverse. The average age of participants, provided in 15 studies, had a median of 41.3 years. In comparison, the 2024 annual report of GKV-SV presented a gender distribution of 71% female and 29% male among actual DiGA users, with an average user age of 54 years.

Figure 9. Year of permanent digital health application (Digitale Gesundheitsanwendung; DiGA) listing versus sample size in clinical trials (n=36).

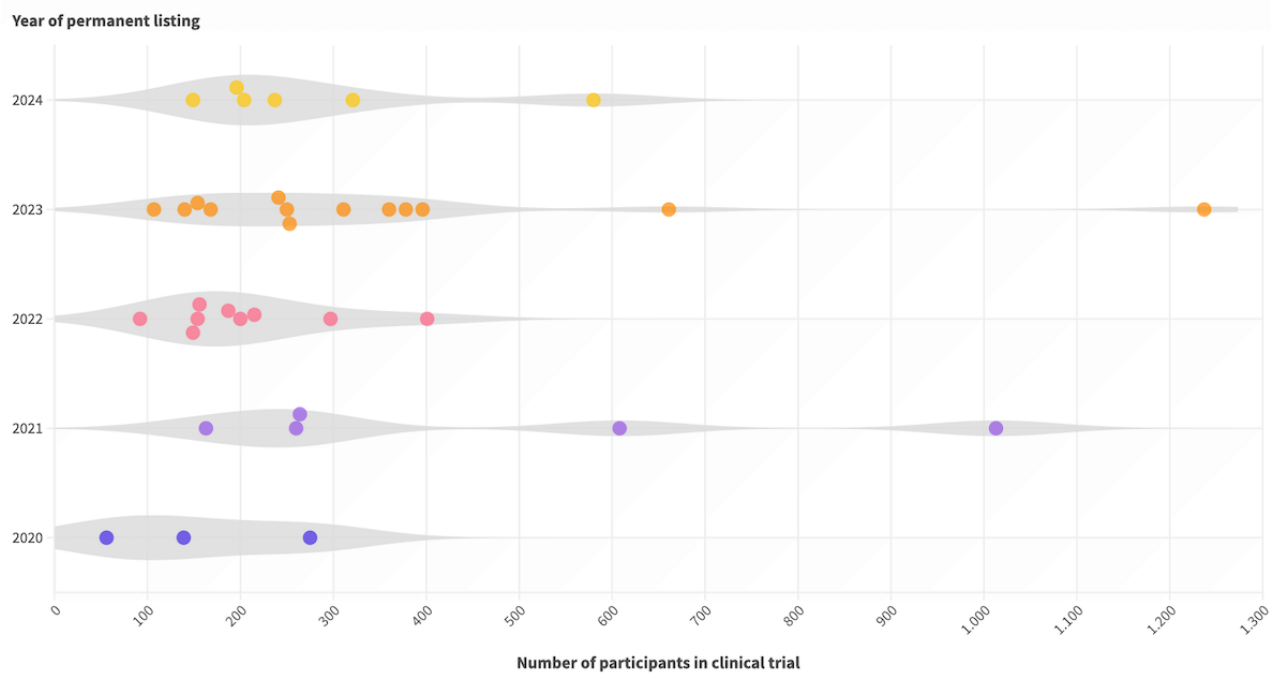

Discussion

Principal Findings and Comparison With Prior Work

In summary, our results paint a picture of the emerging DiGA ecosystem in Germany. Through a data-driven examination, this paper uncovered the complexities and challenges in market dynamics, pricing strategies, prescription trends, and clinical evidence, offering valuable implications for researchers, DiGA manufacturers, regulatory bodies and health care professionals.

As of July 1, 2024, BfArM has approved 64 DiGAs, with 35 being permanently listed, 21 being preliminarily listed, and 8 that were subsequently removed from the directory. This study's analysis, focusing on the 56 currently listed DiGAs, highlights the challenges in transitioning from preliminary to permanent listing. While 75% (24/32) of the DiGAs that were initially preliminary listed eventually achieved permanent listing, they often exceeded the 1-year timeline provided for this transition, with a third requiring exactly 2 years to complete the process. Although we do not have a rationale for this long timeframe in our data set, we anticipate that the process of evidence generation for digital medical devices is lengthy and that finding more appropriate methodologies for this purpose will require more academic attention in the future [4,6,22,23]. Our analysis reveals a wide therapeutic application of DiGAs across 12 categories, with psychological conditions being the most addressed, followed by endocrine and metabolic disorders. Despite the introduction of the positive care effect concept to encompass both medical benefits and patient-relevant structural and procedural improvements, the adoption of pSVV remains limited. Our economic analysis shows a steady year-on-year increase in both the initial and final prices of newly listed DiGAs. The appropriateness of these prices remains a topic of ongoing debate [24]. Looking ahead, the introduction of performance-based pricing models is anticipated to address these concerns by aligning costs more closely with the value and effectiveness of DiGAs [2,4].

Prescription analysis reveals a substantial, yet not overwhelming increase in DiGA use, attributed to both the introduction of new DiGAs to the market and the growth of existing DiGAs. Overall, the 15 most successful account for approximately 82% of total prescriptions. The year-on-year growth rate for DiGA prescriptions among the top 15 has shown a decline. Being first to market appears to be a critical factor for achieving substantial sales, although other elements, such as product quality, perceptions of health care professionals on DiGA [25,26], and the extent of sales efforts, obviously play vital roles but could not be analyzed in this study. While the current analysis has not yielded statistically significant results, the expanding market and the anticipated increase in the number of DiGAs suggest that statistically significant findings are likely to emerge in the near future.

XSL•FO
RenderX

The validity of study designs specifically for DiGAs remains a hotly debated topic [27]. The uniformity in choosing randomized controlled trials over alternative study designs for the primary clinical end point underscores an area ripe for exploration. Researchers are encouraged to investigate the impact of new study designs on the validation of digital health technologies [28,29], also with respect to continuous data evaluation [2]. Clinical trials for permanently listed DiGAs uniformly focused on medical benefits as positive care effect, with a universal emphasis on the *improvement of the state of health*. Thus, clinical trials for DiGAs so far adhere closely to the established methodologies common in the pharmaceutical sector. In the analysis of clinical studies, it was noted that the type and extent of data points made available to the public vary considerably. For example, although the gender ratio of study participants must be submitted to the BfArM for assessment, only 22 (61%) out of the 36 studies have made this information available to the public. In terms of advancing methodologies specifically for DiGAs, it would be beneficial to mandate the comprehensive publication of data points to be able to report relevant trends.

The modest uptake of pSVV as a novel regulatory concept highlights a gap in the market [30]. This calls for further research into the barriers to adoption and the potential of pSVV to broaden the scope of valued outcomes in digital health, potentially enhancing the effectiveness and patient-centeredness of DiGAs [7,31,32]. The hesitancy to adopt this new concept is also a reminder for regulatory bodies that regulatory innovation also requires accompanying measures during the introduction period.

With a median duration of 17.5 months for clinical trials, it logically follows that the initially targeted timeframe of 1 year from preliminary to permanent listing is not achievable for most manufacturers. However, our data also show that the fast-track procedure and preliminary listing of DiGA are a way of mitigating some of the challenges associated with randomized controlled trials [22,24,33]. Remarkably, 75% (24/32) of the DiGAs successfully transitioned from provisional to permanent listing. Furthermore, while more DiGAs will come to the market in the following years, this provides a fertile ground for analyzing DiGA's therapeutic and economic impact together using a holistic model [34].

This paper not only sheds light on the current state of DiGA in Germany, but also signals areas for future investigation and strategic adjustment. By highlighting the need for diversified clinical study designs, a deeper exploration into regulatory innovations, such as pSVV and their adoptions, and holistic economic analysis, it offers a road map for enhancing the integration, effectiveness, and acceptance of digital health technologies within the German health care system and beyond.

Implications for DiGA Manufacturers and Health Care Professionals

While causality cannot be definitively established, our findings suggest that DiGAs available as native apps generally see higher prescription numbers. Intriguingly, DiGA manufacturers with a larger portfolio in the market often offer web applications only, which tend to be less successful than those provided by manufacturers specializing on one or a few native apps. Having this in mind, DiGA manufacturers should consider developing native apps from the beginning. In addition, examining the implications of being first to market within specific *ICD-10* categories offers a unique lens through which to assess competitive advantage and market penetration strategies for digital health solutions. Considering that only 15 DiGAs have reached a milestone of 5000 prescriptions over their lifetime, it is crucial for DiGA manufacturers to recognize that listing in the DiGA directory does not guarantee automatic prescription success.

For health care professionals, it is essential to understand that DiGAs are subject to clinical trials akin to those required for pharmaceuticals. Moreover, the transition from preliminary to permanent listing fails for only a small number of DiGAs, highlighting the stringent evaluation process by the BfArM even at the initial listing stage. Despite public and media concerns about the potential risks associated with preliminarily listed DiGAs, our findings do not support the notion that these CE-certified, preliminarily listed DiGAs fail to deliver on their promised benefits. In addition, our findings indicate relatively positive user feedback on DiGAs, aligning with the extant research [17].

Limitations and Further Research

The insights from this study are constrained by the limited public availability of detailed prescription data beyond the top 15 DiGAs and the variability in reporting standards across DiGA clinical trials. Obviously, individuals outside the statutory health insurance system and even those outside of Germany can purchase DiGAs privately. However, due to the absence of publicly accessible data on these transactions, we were unable to incorporate this information into our analysis.

Further health economics analyses, such as correlating prescription data with the actual number of affected individuals in Germany, would be valuable as well. However, conducting these comprehensive calculations requires a larger and more robust data set with respect to prescriptions. Similarly, due to the limited number of DiGAs currently available, statistically robust calculations were not feasible at this stage. As the market continues to expand, it is anticipated that such analyses will become possible in the future. Finally, we did not assess the quality and robustness of the clinical studies in our sample [7,35] as our goal was to map rather than clinically assess the emerging DiGA ecosystem.

Furthermore, we acknowledge that the presented findings do not address all discussions and criticisms that are being debated within the German statutory health care system. For instance, it is not possible, on the basis of currently available public data, to scientifically comment on the economic impact of DiGAs on various stakeholders and doubts regarding the actual uptake and adherence to DiGAs after prescription [11-13,36-38]. We anticipate that the introduction of performance-based pricing components from 2026 onward, aimed at measuring use, patient satisfaction, and patient health outcomes, may provide usable data for future research in this area. In addition, the integration of economic aspects into the approval and pricing process of DiGAs, like the benefit assessment for pharmaceuticals [39], could be considered in this context.

XSL•FO
RenderX

Conclusions

As part of the 25th anniversary issue of the *Journal of Medical Internet Research*, this paper explores the thrilling early years of DiGAs in Germany, which marked an important stage in the digital transformation of health care. While DiGAs have not yet accelerated on a fast track to widespread adoption, their incremental positive influence on health care professionals and patients hints at a promising future. With their modest current uptake, there is a clear call for intensified efforts to boost their use in the coming years, ensuring that the best of digital health in Germany, Europe, and beyond is yet to be realized. The limited adoption of pSVV and the present monotony in clinical trial design are prominent areas for further research and potential regulatory attention.

It is encouraging to see that legislators have already adapted the DiGA fast-track procedure twice so far in 2024 (Digital Law) and in 2021 (Digital Modernization of Care and Nursing Act [German: *Digitale-Versorgung - und - Pflege - Modernisierungs - Gesetz*]) and has thus gradually adapted it in the spirit of agile legislation. Particularly noteworthy are the extended periods for permanent listing, the permission to also approve class IIb medical devices, and the fact that a path to performance–based pricing components has been introduced [4]. It is anticipated that further changes will be forthcoming, and this is a highly respectable approach that better reflects the reality of *digital* medical devices.

Enhancing transparency in DiGA use data and promoting methodological variety in clinical trials are crucial steps toward leveraging the full potential of digital health technologies for patient care. The insights provided by this study not only enrich our understanding of the DiGA landscape, but also underscore the necessity for continuous evaluation and adaptation within this fast-evolving sector. As DiGAs slowly start to make a positive impact, it is evident that the journey toward identifying and establishing study designs for digital health applications, which are broadly accepted across the market, is far from completion. The advancement of scientific dialogue is pivotal, serving as a crucial element in securing acceptance by both market participants and regulatory bodies.

Acknowledgments

The authors are grateful to the 2 anonymous reviewers whose insightful comments and suggestions contributed substantially to the quality of this manuscript.

Publishing-fee support was provided by the Funding Programme Open Access Publishing of Hamburg University of Technology (TUHH). MG has received funding from the European Union's Horizon 2020 research and innovation program, within the OpenInnoTrain project under Marie Skłowdowska-Curie grant agreement number 823971. The content of this publication does not reflect the official opinions of the European Union. Responsibility for the information and views expressed in the publication lies entirely with the authors.

Generative artificial intelligence (ChatGPT 4, OpenAI) was used to incorporate minor enhancements to sentence structure and wording throughout the publication. All data, analyses, and conclusions were exclusively developed and derived by the authors themselves, ensuring the originality and integrity of our research contributions.

Conflicts of Interest

MG, having previously worked as a consultant for DiGA manufacturers, declares that his past consultancy roles had no bearing on the findings of this research. SG, currently employed by a statutory health insurance company in Germany, also declares that her professional engagement with the insurance company did not influence the content or outcomes of this manuscript.

Multimedia Appendix 1
Overview of clinical trials of all permanently listed digital health applications (Digitale Gesundheitsanwendung; DiGAs) as of July 1, 2024.
[PDF File (Adobe PDF File), 36 KB - jmir_v26i1e59013_app1.pdf]

References

1. Stern AD, Matthies H, Hagen J, Brönnecke JB, Debatin JF. Want to see the future of digital health tools? Look to Germany. Harvard Business Review. 2020 Dec 02. URL: https://hbr.org/2020/12/want-to-see-the-future-of-digital-health-tools-look-to-germany [accessed 2024-04-29]
2. Brönneke JB, Herr A, Reif S, Stern AD. Dynamic HTA for digital health solutions: opportunities and challenges for patient-centered evaluation. Int J Technol Assess Health Care 2023 Nov 17;39(1):e72. [doi: 10.1017/S0266462323002726] [Medline: 37973549]
3. Giebel GD, Speckemeier C, Abels C, Börchers K, Wasem J, Blase N, et al. Problems and barriers related to the use of digital health applications: protocol for a scoping review. JMIR Res Protoc 2022 Apr 21;11(4):e32702 [FREE Full text] [doi: 10.2196/32702] [Medline: 35451979]

XSL•FO
RenderX

4. Schmidt L, Pawlitzki M, Renard BY, Meuth SG, Masanneck L. The three-year evolution of Germany's digital therapeutics reimbursement program and its path forward. NPJ Digit Med 2024 May 24;7(1):139 [FREE Full text] [doi: 10.1038/s41746-024-01137-1] [Medline: 38789620]
5. Mäder M, Timpel P, Schönfelder T, Militzer-Horstmann C, Scheibe S, Heinrich R, et al. Evidence requirements of permanently listed digital health applications (DiGA) and their implementation in the German DiGA directory: an analysis. BMC Health Serv Res 2023 Apr 17;23(1):369 [FREE Full text] [doi: 10.1186/s12913-023-09287-w] [Medline: 37069592]
6. Stern AD, Brönneke J, Debatin JF, Hagen J, Matthies H, Patel S, et al. Advancing digital health applications: priorities for innovation in real-world evidence generation. Lancet Digit Health 2022 Mar;4(3):e200-e206 [FREE Full text] [doi: 10.1016/S2589-7500(21)00292-2] [Medline: 35216754]
7. Lantzsch H, Eckhardt H, Campione A, Busse R, Henschke C. Digital health applications and the fast-track pathway to public health coverage in Germany: challenges and opportunities based on first results. BMC Health Serv Res 2022 Sep 21;22(1):1182 [FREE Full text] [doi: 10.1186/s12913-022-08500-6] [Medline: 36131288]
8. Wangler J, Jansky M. Two years of approved digital health applications in Germany - perspectives and experiences of general practitioners with an affinity for their use. Eur J Gen Pract 2023 Dec;29(1):2186396 [FREE Full text] [doi: 10.1080/13814788.2023.2186396] [Medline: 36919621]
9. van Kessel R, Srivastava D, Kyriopoulos I, Monti G, Novillo-Ortiz D, Milman R, et al. Digital health reimbursement strategies of 8 European countries and Israel: scoping review and policy mapping. JMIR Mhealth Uhealth 2023 Sep 29;11:e49003 [FREE Full text] [doi: 10.2196/49003] [Medline: 37773610]
10. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. NPJ Digit Med 2023 Mar 10;6(1):38 [FREE Full text] [doi: 10.1038/s41746-023-00777-z] [Medline: 36899073]
11. Annual report of GKV-SV on DiGA. GKV Umbrella Association. 2023. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2023_DiGA_Bericht_GKV-Spitzenverband.pdf [accessed 2024-04-29]
12. Annual report of GKV-SV on DiGA. GKV Umbrella Association. 2022. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2022_DiGA_Bericht_BMG.pdf [accessed 2024-04-29]
13. Annual report of GKV-SV on DiGA. GKV Umbrella Association. 2021. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2021_DiGA_Bericht_final_barrierefrei.pdf [accessed 2024-04-29]
14. Anzahl der Mitglieder und Versicherten der gesetzlichen und privaten Krankenversicherung in den Jahren 2017 bis 2023. Statista. URL: https://de.statista.com/statistik/daten/studie/155823/umfrage/gkv-pkv-mitglieder-und-versichertenzahl-im-vergleich/ [accessed 2024-04-29]
15. Biviji R, Williams KS, Vest JR, Dixon BE, Cullen T, Harle CA. Consumer perspectives on maternal and infant health apps: qualitative content analysis. J Med Internet Res 2021 Sep 01;23(9):e27403 [FREE Full text] [doi: 10.2196/27403] [Medline: 34468323]
16. Albrecht U, Hasenfuß G, von Jan U. Description of cardiological apps from the German app store: semiautomated retrospective app store analysis. JMIR Mhealth Uhealth 2018 Nov 20;6(11):e11753 [FREE Full text] [doi: 10.2196/11753] [Medline: 30459140]
17. Uncovska M, Freitag B, Meister S, Fehring L. Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. NPJ Digit Med 2023 Jun 21;6(1):115 [FREE Full text] [doi: 10.1038/s41746-023-00862-3] [Medline: 37344556]
18. Sadare O, Melvin T, Harvey H, Vollebregt E, Gilbert S. Can Apple and Google continue as health app gatekeepers as well as distributors and developers? Npj Digit Med 2023 Jan 31;6(1):8. [doi: 10.1038/S41746-023-00754-6]
19. Keutzer L, Simonsson US. Medical device apps: an introduction to regulatory affairs for developers. JMIR Mhealth Uhealth 2020 Jun 26;8(6):e17567 [FREE Full text] [doi: 10.2196/17567] [Medline: 32589154]
20. Barth T, Goeldner M, Spitzenberger F. Einfluss von regulatorischen Anforderungen auf Innovationen in der Medizintechnik am beispiel der europäischen medical device regulation („MDR“) und des nationalen Digitale-Versorgung-Gesetzes („DVG“). In: Bodemann M, Fellner W, Just V, editors. Zukunftsfähigkeit durch Innovation, Digitalisierung und Technologien: Geschäftsmodelle und Unternehmenspraxis im Wandel. Cham, Switzerland: Springer; 2021:223-252.
21. 2024. Framework agreement of GKV-SV and the manufacturer's representatives. 2024 Feb 28. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/DiGA_Rahmenvereinbarung_nach_134_Abs_4_und_5_SGB_V_Stand_27.02.2024.pdf [accessed 2024-03-11]
22. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions—a call for innovative evidence generation approaches. Npj Digit Med 2020 Aug 27;3(1):55. [doi: 10.1038/S41746-020-00314-2]
23. Torous J, Stern AD, Bourgeois FT. Regulatory considerations to keep pace with innovation in digital health products. NPJ Digit Med 2022 Aug 19;5(1):121 [FREE Full text] [doi: 10.1038/s41746-022-00668-9] [Medline: 35986056]
24. Gensorowsky D, Witte J, Batram M, Greiner W. Market access and value-based pricing of digital health applications in Germany. Cost Eff Resour Alloc 2022 Jun 13;20(1):25 [FREE Full text] [doi: 10.1186/s12962-022-00359-y] [Medline: 35698135]
25. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]

XSL•FO
RenderX

26. Schroeder T, Haug M, Georgiou A, Seaman K, Gewald H. Evidence of how physicians and their patients adopt mHealth apps in Germany: exploratory qualitative study. JMIR Mhealth Uhealth 2024 Jan 17;12:e48345 [FREE Full text] [doi: 10.2196/48345] [Medline: 38231550]
27. Giebel GD, Speckemeier C, Abels C, Plescher F, Börchers K, Wasem J, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808 [FREE Full text] [doi: 10.2196/43808] [Medline: 37171838]
28. Dick S, O'Connor Y, Heavin C. Approaches to mobile health evaluation: a comparative study. Inf Syst Manag 2019 Nov 27;37(1):75-92. [doi: 10.1080/10580530.2020.1696550]
29. Sharma A, Harrington RA, McClellan MB, Turakhia MP, Eapen ZJ, Steinhubl S, et al. Using digital health technology to better generate evidence and deliver evidence-based care. J Am Coll Cardiol 2018 Jun 12;71(23):2680-2690 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.523] [Medline: 29880129]
30. Gensorowsky D, Lampe D, Hasemann L, Düvel J, Greiner W. "Alternative study designs" for the evaluation of digital health applications - a real alternative? Z Evid Fortbild Qual Gesundhwes 2021 Apr;161:33-41. [doi: 10.1016/j.zefq.2021.01.006] [Medline: 33642251]
31. Cohen AB, Mathews SC, Dorsey ER, Bates DW, Safavi K. Direct-to-consumer digital health. Lancet Digit Health 2020 Apr;2(4):e163-e165 [FREE Full text] [doi: 10.1016/S2589-7500(20)30057-1] [Medline: 33328077]
32. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019 Oct 17;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
33. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 09;4(3):e107 [FREE Full text] [doi: 10.2196/mhealth.5720] [Medline: 27613084]
34. Kählke F, Buntrock C, Smit F, Ebert DD. Systematic review of economic evaluations for internet- and mobile-based interventions for mental health problems. NPJ Digit Med 2022 Nov 23;5(1):175 [FREE Full text] [doi: 10.1038/s41746-022-00702-w] [Medline: 36424463]
35. Kolominsky-Rabas PL, Tauscher M, Gerlach R, Perleth M, Dietzel N. How robust are studies of currently permanently included digital health applications (DiGA)? Methodological quality of studies demonstrating positive health care effects of DiGA. Z Evid Fortbild Qual Gesundhwes 2022 Dec;175:1-16 [FREE Full text] [doi: 10.1016/j.zefq.2022.09.008] [Medline: 36437182]
36. Digitale Gesundheitsanwendungen (DiGA): Mehr Transparenz für mehr Akzeptanz. BARMER Institut für Gesundheitssystemforschung. 2024. URL: https://www.bifg.de/media/dl/ePaper/bifg-ePaper-DiGA.pdf [accessed 2024-07-04]
37. Marktentwicklung Digitaler Gesundheitsanwendungen (DiGA-Report). Spitzenverband Digitale Gesundheitsversorgung. 2024. URL: https://digitalversorgt.de/wp-content/uploads/2024/01/DiGA-Report-2023-SVDGV.pdf [accessed 2024-07-04]
38. DiGA-report II 2024. Techniker Krankenkasse. 2024. URL: https://www.tk.de/presse/themen/digitale-gesundheit/digitaler-fortschritt/diga-report-2-2024-2125138?tkcm=ab [accessed 2024-07-03]
39. AMNOG – Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V. Gemeinsamer Bundesausschuss. 2024. URL: https://www.g-ba.de/themen/arzneimittel/arzneimittel-richtlinie-anlagen/nutzenbewertung-35a/ [accessed 2024-04-29]

Abbreviations

BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte (Federal Institute for Drugs and Medical Devices)
DiGA: Digitale Gesundheitsanwendung (digital health application)
GKV-SV: National Association of Statutory Health Insurance Funds
ICD-10: International Classification of Diseases, Tenth Edition
MDD: Medical Device Directive
MDR: Medical Device Regulation
PHQ-9: Patient Health Questionnaire
pSVV: patientenrelevante Struktur- und Verfahrensverbesserung (patient-relevant improvement of structure and processes)
pVE: Positiver Versorgungseffekt (positive care effect)

XSL•FO
RenderX

Edited by G Eysenbach; submitted 30.03.24; peer-reviewed by R Mathias, L Schmidt; comments to author 24.06.24; revised version received 12.07.24; accepted 16.07.24; published 29.08.24.

Please cite as:
Goeldner M, Gehder S
Digital Health Applications (DiGAs) on a Fast Track: Insights From a Data-Driven Analysis of Prescribable Digital Therapeutics in Germany From 2020 to Mid-2024
J Med Internet Res 2024;26:e59013
URL: https://www.jmir.org/2024/1/e59013
doi:10.2196/59013
PMID:

©Moritz Goeldner, Sara Gehder. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of the Veterans Health Administration's Digital Divide Consult for Tablet Distribution and Telehealth Adoption: Cohort Study

Jacqueline M Ferguson[1], PhD; James Van Campen[1], MS; Cindie Slightam[1], MPH; Liberty Greene[1,2], MS, MEd; Leonie Heyworth[3,4], MD, MPH; Donna M Zulman[1,2], MD, MS

[1]Center for Innovation to Implementation, Veterans Affairs Palo Alto Healthcare System, Menlo Park, CA, United States
[2]Division of Primary Care and Population Health, Stanford School of Medicine, Palo Alto, CA, United States
[3]Office of Connected Care/Telehealth, Department of Veterans Affairs Central Office, Veterans Health Administration, Washington, DC, United States
[4]Department of Medicine, University of California, San Diego School of Medicine, San Diego, CA, United States

Corresponding Author:
Jacqueline M Ferguson, PhD
Center for Innovation to Implementation
Veterans Affairs Palo Alto Healthcare System
795 Willow Road
MDP-152
Menlo Park, CA, 94025
United States
Phone: 1 (650) 493 5000
Email: Jacqueline.Ferguson@va.gov

Abstract

Background: Video telehealth offers a mechanism to help Veterans Health Administration (VHA) patients overcome health care access barriers; however, many veterans lack a suitable device and sufficient internet connectivity. To address disparities in technology access, VHA established a Connected Device Program that offers veterans loaned video-capable tablets and internet service. In 2020, VHA introduced a national Digital Divide Consult to facilitate and standardize referrals for this resource.

Objective: We sought to evaluate the reach and impact of VHA's Connected Device Program, leveraging Digital Divide Consult data to determine whether resources are supporting veterans with health care needs and access barriers.

Methods: We examined the reach of VHA's Connected Device Program using national secondary data from VHA's electronic health records among 119,926 tablet recipients who received a tablet (April 1, 2020, to February 28, 2023) and 683,219 veterans from the general VHA population. We assessed changes in tablet recipients' demographic and clinical characteristics before and after implementation of the Digital Divide Consult compared with the general VHA population. We examined the impact of tablets and the consult on adoption of telehealth (ie, video visit use and number of visits) adjusting for differences between tablet recipients and the general VHA population. Finally, we evaluated consult implementation by assessing the use of video-based services by tablet referral reason.

Results: Common reasons for tablet referral included mental health diagnoses (50,367/79,230, 63.9%), distance from a VHA facility >30 miles (17,228/79,230, 21.7%), and social isolation (16,161/79,230, 20.4%). Moreover, 63.0% (49,925/79,230) of individuals who received a tablet after implementation of the Digital Divide Consult had a video visit in the first 6 months of tablet receipt. Some consult reasons were associated with a higher-than-average percentage of video telehealth use, including enrollment in evidence-based mental health programs (74.8% [830/1100] with video use), living >30 miles from a VHA facility (68.3% [10,557/17,228] with video use), and having a mental health diagnosis (68.1% [34,301/50,367] with video use). Tablet recipients had nearly 3 times the likelihood of having a video visit within a month once provided a tablet compared to the general VHA population, with an adjusted risk ratio of 2.95 (95% CI 2.91-2.99) before consult implementation and 2.73 (95% CI 2.70-2.76) after consult implementation. Analyses of telehealth adoption suggested that veterans receiving tablets for mental health care and evidence-based programs have higher rates of video visits, while those who are homebound or receiving tablets for hospice have higher rates of nonuse.

XSL•FO
RenderX

Conclusions: This evaluation of VHA's Connected Device Program suggests that tablets are facilitating video-based care among veterans with complex needs. Standardization of referrals through the Digital Divide Consult has created opportunities to identify groups of tablet recipients with lower telehealth adoption rates who might benefit from a targeted intervention.

(J Med Internet Res 2024;26:e59089) doi:10.2196/59089

KEYWORDS

veterans; health care access; video-based care; telehealth; barriers to care; telemedicine

Introduction

Access to health care is a foundational priority for the US Veterans Health Administration (VHA) [1]. Many veterans experience barriers that impede their use of VHA clinical and social services, including geographic and transportation difficulties, physical and mental health challenges, and socioeconomic stressors [2]. Approximately 2.7 million VHA veterans live in rural or insular areas [3], a scenario that can compound other access barriers.

Video telehealth (ie, delivery of clinical care via video visits) offers a mechanism to overcome health care access barriers. This is particularly the case in a setting, such as VHA, a national health care system in which veterans can receive care from clinicians in remote locations. VHA's investments in telehealth date back over 20 years [4], a history that facilitated widespread expansion of telehealth during the COVID-19 pandemic [5]. However, video telehealth requires that patients have a suitable device and internet connectivity. By some estimates, 15% of veterans lack home internet [6] and nearly one-third of rural veterans do not access the internet at home [3]. This unequal access to digital technology—or "digital divide"—can translate to disparities in health care access [7]. VHA studies during the COVID-19 pandemic found evidence of a persistent digital divide [8], with a lower likelihood of video visits occurring among VHA patients in rural areas, those who are older, and those with a history of homelessness [9,10].

To enhance access to video telehealth, policymakers and VHA leaders have taken several steps, including removing geographic licensure barriers for providers and establishing partnerships with select major mobile carriers to reduce data fees associated with the use of VHA's video telehealth platform [8,11]. In 2016, VHA's Office of Rural Health (ORH) and Office of Connected Care (OCC) launched the Connected Device Program to distribute video-enabled tablets to veterans who had a clinical need that could be met through video telehealth, but who did not have their own device or a suitable data plan [12,13]. To support the ongoing evaluation and improvement of this program, VHA's Quality Enhancement Research Initiative (QUERI) and ORH funded the Virtual Access QUERI in 2018 [14]. Previous analyses supported by this partnered evaluation have examined tablet distribution patterns; veterans' experiences with the tablets; and the impact of tablets on mental health care access, continuity, and outcomes [15-19]. One evaluation found that rural veterans are at risk for suicide who received a VHA-issued tablet (compared to those who did not) experienced increases in mental health service use, as well as a 36% reduction in the likelihood of suicide-related emergency department visits [19]. Other assessments by the Virtual Access QUERI have included a survey of tablet recipients to understand their experience with and the remaining barriers to using video telehealth after the provision of a tablet [17,20,21].

The Virtual Access QUERI's partnership with OCC informed the design and implementation of the Digital Divide Consult, a new referral process for veterans to qualify for devices. Implemented in VHA in September 2020, this new nationwide standardized consult e-referred veterans to VHA social workers to assess the need and eligibility for the VHA-loaned tablet program or for other federal subsidies for a device and connectivity. The standardized consult offered an opportunity to examine the characteristics of tablet recipients and tablet referral patterns in greater detail across the entirety of VHA. As part of the Virtual Access QUERI, we leveraged the implementation of the Digital Divide Consult to conduct an in-depth assessment of the following: (1) tablet recipients' demographic and clinical characteristics; (2) the monthly use of video care among veterans with a tablet compared to the monthly use among the general VHA population (ie, veterans without a tablet), including changes in use patterns before and after consult implementation; (3) provider-selected reasons for tablet referral according to Digital Divide Consult data; and (4) the use of video-based services by tablet referral reason.

Methods

VHA Digital Divide Consult

The Digital Divide Consult is a format-standardized template in VHA's computerized patient record system (CPRS). Through this consult template, any VHA clinical staff member who manages a veteran's care can request the provision of a video-enabled device and internet service for the veteran. The provider may also request a peripheral device for use during video appointments (such as a blood pressure monitor or pulse oximeter). After the consult is placed, a social worker reviews the consult, evaluates the veteran's need for a device and internet service, and places an order to have a device with a data plan shipped directly to the veteran. In addition, the social worker can help the veteran apply for other federal broadband programs [22], including EveryoneOn (discontinued June 1 2024) [23] and Lifeline [24].

Prior to implementation of the Digital Divide Consult (mandatory by September 15, 2020), VHA did not have a nationally standardized process for tablet distribution. While there were explicit criteria for tablets, the national program offices (ORH and OCC) could not reliably track distribution reasons or adherence to the criteria until the implementation of the Digital Divide Consult [13]. Through the Digital Divide Consult, the provider must indicate that the veteran meets one

XSL•FO
RenderX

or more criteria for a tablet, which is henceforth referred to as Digital Divide Consult reasons. The Digital Divide Consult expands on the 4 existing consult reasons and includes more detailed device eligibility criteria compared to the original consult (Table 1). In addition, the health care provider placing the consult is required to attest that the veteran will require a video telehealth appointment in the next 90 days.

Table 1. Comparison of consult criteria before the Digital Divide Consult.

Pre-Digital Divide Consult criteria	Digital Divide Consult criteria
Transportation issues	• No DAV[a] van or Veterans Transportation Network access • Difficulty accessing public transportation • No car or access to a ride
Distance or geography	• Veteran lives >30 miles from a VHA[b] medical facility
Homebound or difficulty leaving home	• Work, school, or caregiver commitments make in-person visits challenging • Health issues make the patient homebound or make in-person visits challenging • Any mental health diagnosis • Participation in an evidence-based protocol for mental health therapy • Veteran has been hospitalized in the past 90 days • Social isolation
Other reasons described by the provider	• Cost of attending in-person visits is prohibitive • Difficulty attending visits at a VHA facility (eg, psychological distress and immuno-compromised state) • Documented disruptive behavior at a VHA facility • Homeless veteran or enrolled in HUD-VASH[c] • The patient does not meet any of the criteria (other reasons described by the provider)

[a]DAV: Disabled American Veterans.

[b]VHA: Veterans Health Administration.

[c]HUD-VASH: US Department of Housing and Urban Development-VA Supportive Housing.

Veteran Study Sample

Veteran patient data and outpatient visit history were extracted from the VHA Corporate Data Warehouse (CDW), a repository for VHA electronic health records [25]. The evaluation focused on 119,926 tablet recipients who had at least one outpatient VHA visit in 2019. In order to compare tablet recipients to the general VHA population, we also identified a random sample of 683,219 veterans with at least one VHA outpatient visit in 2019, who had not received a tablet. Overall, the study cohort comprised 803,145 veterans.

Tablet recipients were limited to those who received a tablet between April 01, 2020, and February 28, 2023, using the most recent tablet shipment date. This study sample excluded veterans with missing data on their zip code (required for drive time and rurality of residence) or priority enrollment group and those who had less than 1 month of follow-up (ie, tablet recipients who died within 1 month of tablet receipt or veterans from the general VHA population who died in April 2020). Additional details on sample selection and exclusion criteria are available in Multimedia Appendix 1.

Outcomes

Our outcomes were the occurrence of an outpatient video visit and the number of outpatient video visits in each month between April 1, 2020, and August 31, 2023. As a secondary analysis, we evaluated phone (eg, audio-only) visits. Though video visits are the preferred modality for telehealth at VHA [5], telehealth visits can also be offered via phone owing to patient's limited broadband access or provider-patient preference [26]. Outpatient video-based and phone-based VHA health care visits were identified using VHA's Managerial Cost Accounting (MCA) Stop Codes [9]. Veterans were censored in the month of death. We focused posttablet follow-up among tablet recipients to the first 6 months (180 days) of tablet receipt. This timeframe has been used in previous evaluations [15,19,27] because it spans the period of the highest telehealth engagement after tablet receipt and because devices are retrieved on a quarterly basis if they have not been used in the past 90 days.

Veteran Patient Characteristics

To compare differences between tablet recipients and the general VHA population and to adjust for confounding factors required to ensure exchangeability between the 2 groups, we obtained veteran patient characteristics that have been associated with health care use [9,10,28], using data from a 1-year baseline period (March 11, 2019, to March 10, 2020). The selection of a baseline period in this evaluation is complicated by eligibility criteria (ie, tablet receipt) being met at multiple times as a veteran could be referred for a tablet at any time between April 01, 2020, and February 28, 2023. We selected an unbiased choice of time zero [29] and defined the baseline period prior to the first eligible time (ie, April 01, 2020) and prior to the start of the COVID-19 pandemic.

Characteristics assessed included age, sex, race, ethnicity, marital status, rurality of the home, VHA priority enrollment category, drive time to a primary site of care, number of chronic conditions, and presence of a mental health condition. Race and

XSL•FO
RenderX

ethnicity were the most frequent self-identified classifications in patient health records. Missing data on race, ethnicity, or marital status were treated as distinct categories. Veterans' home addresses were classified into urban, rural, and highly rural designations based on the zip code of their home address [6,10]. We incorporated information from VHA's priority-based enrollment system, which categorizes veterans into 1 of 8 groups based on their disability rating, income, recent military service, and other factors [6]. We report priority groups in 4 categories: high disability, low or moderate disability, low income, and enrolled without special considerations. Twenty-eight chronic conditions, including 10 mental health conditions, were identified using ICD-10 codes [9,30-33].

To assess the provider-selected criteria for tablet receipt in the Digital Divide Consult, among all veterans (including those without a Digital Divide Consult), we assessed 4 veteran patient characteristics from electronic health care records that have been shown to be associated with differential use of video-based care [10,27]. A history of housing instability and history of homelessness were defined following an adapted methodology from Tsai et al [34] (Multimedia Appendix 1). A history of hospitalization was defined as an inpatient admission in the 180 days before March 1, 2020. A high risk for suicide was derived from the "High Risk for Suicide Patient Record Flag" (HRS-PRF) in patient electronic health care records between March 1, 2019, and March 2020. This patient record flag is used after a suicide risk evaluation or when a patient has had a recent suicide attempt, suicidal preparatory behavior, or suicidal ideation (Multimedia Appendix 1).

We also assessed prior use (eg, during the baseline period) of VHA outpatient care, including prior use of video-based, phone-based, and remote patient monitoring health care (ie, remote blood pressure monitoring) and prior use of primary care, mental health, specialty, and diagnostic or ancillary care [9,10]. See Multimedia Appendix 1 for additional information on the assessment of all veteran patient characteristics.

Digital Divide Consults were identified in the CDW and linked to tablet shipment dates obtained from VHA's Denver Acquisitions and Logistics Center, which oversees tablet distribution [5,13]. For pre-post analyses of the Digital Divide Consult implementation, we defined the preperiod as April 1, 2020, to October 1, 2020, and the postperiod as October 1, 2020, through the end of follow-up, August 31, 2023. We used October 1, 2020, rather than September 15, 2020 (effective date of consult implementation) to best capture shipped tablets generated from the new consult (ie, to account for the lag between the consult close date and tablet shipment).

Nationwide Trends in Telehealth and Tablet Distribution in the Study Sample

To describe trends in tablet distribution and nationwide trends in telehealth use (ie, beyond the analytic sample), we tallied all video and phone outpatient care visits nationwide in VHA between January 5, 2019, and September 2, 2023, among primary care, mental health, and subspecialty services. We graphically represented the locally estimated scatter plot smoothed (LOESS) weekly total number of video and phone visits nationwide (ie, line graph) overlaid on a bar graph representing the number of tablets shipped in each week for the 119,926 tablet recipients in the analytic sample.

Statistical Analysis: Evaluation of VHA's Connected Device Program and Digital Divide Consult

We performed several analyses to evaluate the VHA's Connected Device Program and Digital Divide Consult

Sociodemographic and Clinical Characteristics of Tablet Recipients and the General VHA Population

We have described the sociodemographic and clinical characteristics of tablet recipients (all and those who received tablets before and after Digital Divide Consult implementation), and compared tablet recipients with the general VHA population.

Evaluation of the Digital Divide Consult and Video Telehealth Use

To examine how the Digital Divide Consult influenced video telehealth use while accounting for temporal differences induced by the COVID-19 pandemic, we estimated monthly video visit use as defined by (1) the likelihood of any video visit in a month and (2) the count of video visits in a month. We assigned each person-month between April 01, 2020, and August 31, 2023, to the following categories: category A, person-months without a tablet (includes person-months from the general VHA population and person-months of tablet recipients *before* receiving a tablet); category B, person-months with a tablet among veterans who received their tablet before the implementation of the Digital Divide Consult; and category C, person-months with a tablet among veterans who received their tablet after the implementation of the Digital Divide Consult. Each model had fixed effects for the calendar month, and person-months without a tablet were set as the reference group. Therefore, our models compared the monthly use of video care in any given month between (1) category B divided by category A and (2) category C divided by category A.

By including a fixed effect for the calendar month and by comparing the monthly use of video care in the same month between person-months with a tablet and person-months without a tablet, we accounted for temporal differences in video care use induced by the COVID-19 pandemic. We accounted for differences between tablet recipients and the general VHA population by adjusting for the key patient demographics associated with differential use of video telehealth (ie, age, sex, rurality, etc) and the 4 characteristics similar to those captured in the Digital Divide Consult as described earlier (eg, history of housing instability, history of homelessness, history of hospitalization, and high risk for a suicide patient risk flag).

We estimated adjusted risk ratios (RRs) and the predicted probability (average marginal effects) of any video telehealth use in a month using generalized linear Poisson models with a log link function. Adjusted RRs can be interpreted as the relative likelihood of a video visit in a given month between months without a tablet and months with a tablet. We also estimated the adjusted incidence rate ratios (IRRs) for video telehealth and predicted the count of video visits per month (average marginal effects) using generalized linear negative binomial models with a log link function, which may be interpreted

XSL•FO
RenderX

similarly. All models report 95% CIs and have standard errors clustered at the patient level.

Video-Based Services by Referral Reason in the Digital Divide Consult

To evaluate video-based services by referral reason in the Digital Divide Consult, we restricted the study population of 91,196 veterans who received a tablet after the Digital Divide Consult to 79,230 veterans who had a consult completed no later than 6 months before the tablet ship date and no more than 30 days after the tablet ship date (to account for delayed consult completion). We have reported the percentage of postconsult tablet recipients with any video visit. We have reported the unadjusted mean number of video visits during the first 6 months of tablet possession among the 79,230 veterans (including veterans with no or zero video visits) using *t*-tests. We estimated the adjusted mean number of video visits using a negative binomial model for each Digital Divide Consult reason, with adjustment only for age at tablet receipt, gender, race, ethnicity, and rurality of the home.

Finally, as a sensitivity analysis to identify if an increased number of consult reasons was associated with increased video use, we also evaluated the association between the number of consult reasons selected and the probability of a video visit and predicted number of video visits in a month using a Poisson model and a negative binomial model, respectively. This analysis was minimally adjusted for age at tablet receipt, race, ethnicity, gender, and rurality of the home.

All statistical analyses were conducted in Stata 18 (StataCorp, LLC).

Ethical Considerations

This evaluation was conducted as part of the Virtual Access QUERI, which is designated as quality improvement by VA's Office of Rural Health and VA Research Administration. The institutional review board at the Stanford Research Compliance Office determined this evaluation does not meet the requirement of research or clinical investigation per Federal Regulations 45CFR 46.104 [35] and VA 38CFR 16.104 [36].

Results

Nationwide Trends in Telehealth and Tablet Distribution in the Study Sample

A total of 119,926 veterans received a tablet between April 1, 2020, and February 28, 2023, representing just over 2% of veterans who are actively using VHA health care. VHA tablet distribution increased in line with the expansion of video and phone-based care in early 2020 following the start of the COVID-19 pandemic and decreased as the pandemic continued (Figure 1). The largest number of tablets was distributed in late summer and early fall of 2020, with a peak of 2280 tablets shipped the week of September 21, 2020 (Figure 1).

XSL•FO
RenderX

Figure 1. Smoothed weekly trends in video and phone visits and counts of weekly tablet distribution in the Veterans Health Administration (VHA) between January 5, 2019, and September 2, 2023. The vertical dashed line represents March 11, 2020 (WHO declares COVID-19 a pandemic). The vertical dash-dot line represents May 11, 2023 (end of federal COVID-19 public health emergency declaration).

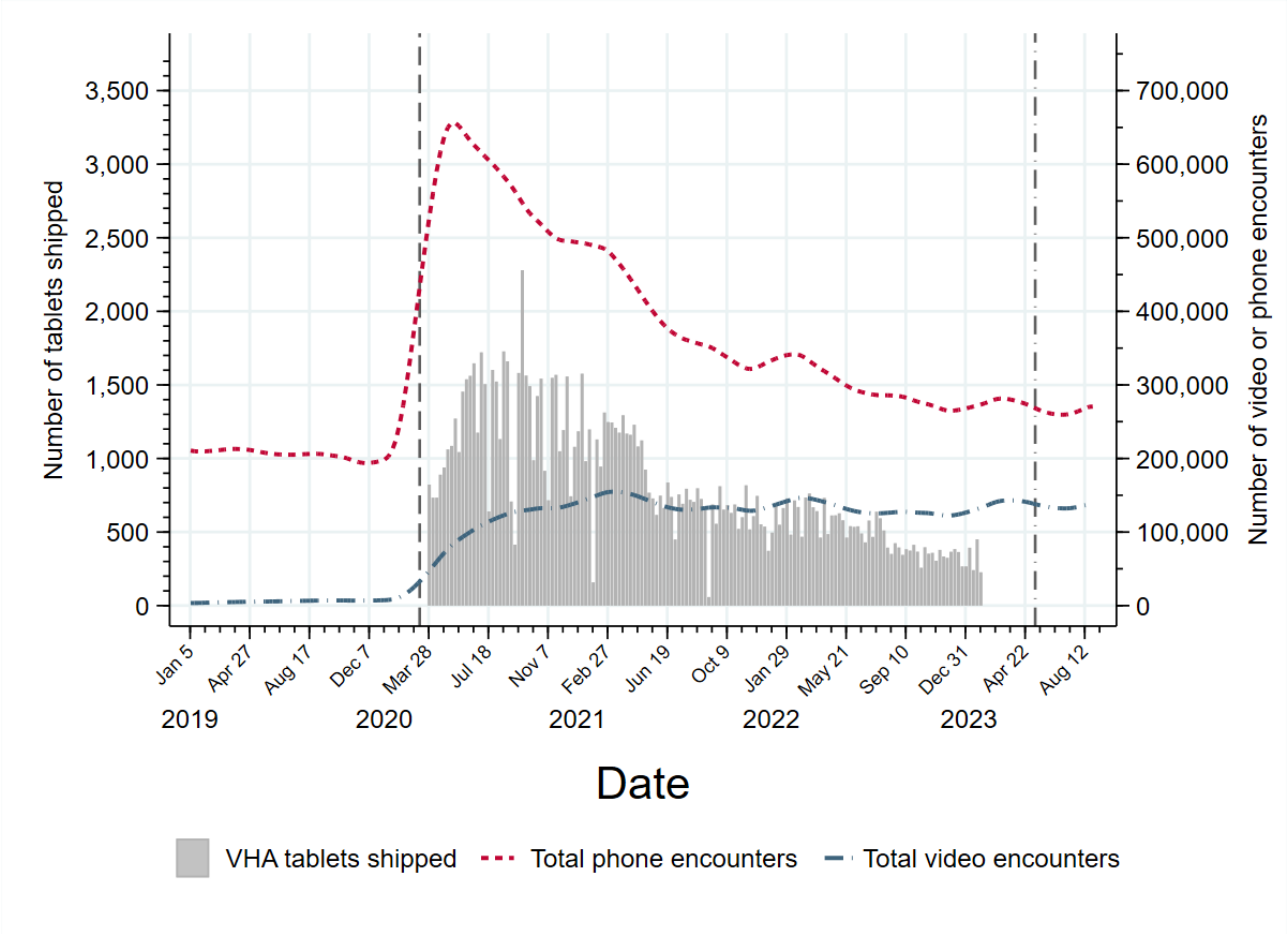

Sociodemographic and Clinical Characteristics of Tablet Recipients and the General VHA Population

Similar to the general VHA population, veterans who received a tablet were predominately white, male, and living in urban areas (Table 2). Veterans who received a tablet were on average 2 years younger than the general population of veterans (mean age: 64 vs 66 years) and on average had 2 more chronic conditions (4 vs 2). The proportion of tablet recipients who were black was nearly twice the proportion among the general VHA population (34,901/119,926, 29.1% vs 121,666/683,219, 17.8%). Larger proportions of tablet recipients than the general VHA population were living in urban areas (81,985/119,926, 68.4% vs 447,235/683,219, 65.5%), were classified as low income (30,889/119,926, 25.8% vs 122,003/683,219, 17.9%) or high disability in VHA enrollment priority groups (59,744/119,926, 49.8% vs 279,530/683,219, 40.9%), or were single, divorced, or widowed (70,007/119,926, 58.4% vs 290,989/683,219, 42.6%). A smaller proportion of tablet recipients than the general VHA population were of Hispanic or Latino ethnicity (6139/119,926, 5.1% vs 47,468/683,219, 6.9%). Tablet recipients had twice the prevalence of a mental health condition (82,066/119,926, 68.4%), 3 times the prevalence of hospitalization (14,226/119,926, 11.9%), and over 4 times the prevalence of a recent history of housing instability (3361/119,926, 2.8%), homelessness (23,002/119,926, 19.2%), or a suicide flag in their health care records (3014/119,926, 2.5%) compared to the general VHA population (Table 2). Tablet recipients also had higher levels of prior engagement with the VHA health care system at baseline than the general VHA population (eg, higher proportion of any use of VHA outpatient visits across all modalities and care types) (Multimedia Appendix 2).

XSL•FO
RenderX

Table 2. Demographic and clinical characteristics of Veterans Health Administration (VHA) tablet recipients and a 5:1 probability sample of the general VHA population (N=803,145) active in VHA care.

Variable	General VHA[a] population (N=683,219), n (%)	All tablet recipients (N=119,926), n (%)	Tablet recipients before Digital Divide Consult implementation (N=28,730), n (%)	Tablet recipients after Digital Divide Consult implementation (N=91,196), n (%)
Age (years)				
18-44	124,356 (18.2)	14,430 (12.0)	3189 (11.1)	11,241 (12.3)
45-64	200,718 (29.4)	48,199 (40.2)	10,556 (36.7)	37,643 (41.3)
≥65	358,145 (52.4)	57,297 (47.8)	14,985 (52.2)	42,312 (46.4)
Sex				
Male	619,267 (90.6)	104,890 (87.5)	25,492 (88.7)	79,398 (87.1)
Female	62,310 (9.1)	14,254 (11.9)	3067 (10.7)	11,187 (12.3)
Gender diverse	1642 (0.2)	782 (0.7)	171 (0.6)	611 (0.7)
Race				
Asian	6354 (0.9)	1212 (1.0)	289 (1.0)	923 (1.0)
American Indian or Alaska Native	8303 (1.2)	765 (0.6)	190 (0.7)	575 (0.6)
Black	121,666 (17.8)	34,901 (29.1)	7336 (25.5)	27,565 (30.2)
Native Hawaiian or Pacific Islander	6446 (0.9)	1069 (0.9)	281 (1.0)	788 (0.9)
Missing	46,719 (6.8)	6082 (5.1)	1434 (5.0)	4648 (5.1)
White	493,731 (72.3)	75,897 (63.3)	19,200 (66.8)	56,697 (62.2)
Ethnicity				
Not Hispanic or Latino	608,765 (89.1)	110,336 (92.0)	26,542 (92.4)	83,794 (91.9)
Hispanic or Latino	47,468 (6.9)	6139 (5.1)	1372 (4.8)	4767 (5.2)
Missing	26,986 (3.9)	3451 (2.9)	816 (2.8)	2635 (2.9)
Marital status				
Single, divorced, or widowed	290,989 (42.6)	70,007 (58.4)	16,580 (57.7)	53,427 (58.6)
Married	381,997 (55.9)	48,934 (40.8)	11,917 (41.5)	37,017 (40.6)
Unknown	10,233 (1.5)	985 (0.8)	233 (0.8)	752 (0.8)
Rural or urban status				
Urban	447,235 (65.5)	81,985 (68.4)	19,663 (68.4)	62,322 (68.3)
Rural	227,082 (33.2)	36,840 (30.7)	8825 (30.7)	28,015 (30.7)
Highly rural	8902 (1.3)	1101 (0.9)	242 (0.8)	859 (0.9)
Priority enrollment category				
No special enrollment	114,620 (16.8)	8589 (7.2)	2100 (7.3)	6489 (7.1)
Low income	122,003 (17.9)	30,889 (25.8)	7558 (26.3)	23,331 (25.6)
Low or moderate disability	167,066 (24.5)	20,704 (17.3)	4962 (17.3)	15,742 (17.3)
High disability	279,530 (40.9)	59,744 (49.8)	14,110 (49.1)	45,634 (50.0)
Drive time to the primary site of care (min)				
<30	533,772 (78.1)	96,786 (80.7)	23,586 (82.1)	73,200 (80.3)
≥30	141,575 (20.7)	22,332 (18.6)	4966 (17.3)	17,366 (19.0)
Unknown	7872 (1.2)	808 (0.7)	178 (0.6)	630 (0.7)
Number of chronic conditions				
0-3	470,628 (68.9)	48,221 (40.2)	10,355 (36.0)	37,866 (41.5)
4-6	175,048 (25.6)	49,865 (41.6)	12,299 (42.8)	37,566 (41.2)

XSL•FO
RenderX

Variable	General VHA[a] population (N=683,219), n (%)	All tablet recipients (N=119,926), n (%)	Tablet recipients before Digital Divide Consult implementation (N=28,730), n (%)	Tablet recipients after Digital Divide Consult implementation (N=91,196), n (%)
7-9	32,350 (4.7)	17,660 (14.7)	4790 (16.7)	12,870 (14.1)
≥10	2544 (0.4)	2225 (1.9)	688 (2.4)	1537 (1.7)
Missing	2649 (0.4)	1955 (1.6)	598 (2.1)	1357 (1.5)
Mental health condition	228,538 (33.5)	82,066 (68.4)	19,916 (69.3)	62,150 (68.1)
History of hospitalization (in the 180 days before March 1, 2020)	25,692 (3.8)	14,226 (11.9)	3902 (13.6)	10,324 (11.3)
Housing instability	4068 (0.6)	3361 (2.8)	860 (3.0)	2501 (2.7)
Homelessness	25,809 (3.8)	23,002 (19.2)	6175 (21.5)	16,827 (18.5)
High risk for suicide (any suicide electronic health care record flag in the 12 months before March 1, 2020)	3069 (0.4)	3014 (2.5)	725 (2.5)	2289 (2.5)

[a]VHA: Veterans Health Administration.

Evaluation of the Digital Divide Consult and Video Telehealth Use

The Digital Divide Consult was implemented in September 2020, but there was variation in the implementation by the VHA medical facility. By December 2020, 130 out of 139 (93.5%) VHA medical centers were using the Digital Divide Consult. Among all 139 VHA medical centers, 99 (71.2%) provided a Digital Divide Consult to ≥80% of tablet recipients during the last 3 months of 2020. A year after the Digital Divide Consult implementation, 131 (94%) VHA medical centers provided a Digital Divide Consult to ≥80% of tablet recipients. Among the 91,196 tablet recipients after consult implementation, 86.9% (n=79,230) had an identifiable Digital Divide Consult completed.

Among tablet recipients, those who received their tablets after consult implementation were largely similar to those who received their tablets before (Table 2). All veterans who received a tablet had nearly 3 times the likelihood of having a video visit in a month compared to the general VHA population, with an adjusted RR of 2.95 (95% CI 2.91-2.99) before consult implementation and 2.73 (95% CI 2.70-2.76) after consult implementation (Table 3). This corresponds to a predicted probability of having a video visit in a month of 0.21 (95% CI 0.20-0.21) and 0.19 (95% CI 0.19-0.19) among tablet recipients before and after consult implementation, respectively, and 0.07 (95% CI 0.069-0.090) among the general VHA population. Tablet recipients also had a higher predicted average count of video visits per month, at an IRR of 4.47 (95% CI 4.36-4.59) and 4.27 (95% CI 4.20-4.34) among tablet recipients before and after consult implementation, respectively, compared to the general VHA population. This corresponds to an average of 3 visits every 6 months among tablet recipients compared to less than 1 visit among the general VHA population.

Table 3. Monthly video visit comparing the general Veterans Health Administration population, tablet recipients before Digital Divide Consult implementation, and tablet recipients after Digital Divide Consult implementation.

Group[a]	Likelihood of any video visit per month[b]		Count of video visits per month[b]	
	Risk ratio (95% CI)	Predicted probability of a video visit (95% CI)	Incidence rate ratio (95% CI)	Predicted number of video visits per month (95% CI)
General VHA[c] population	1.00 (reference)	0.07 (0.069-0.070)	1.00 (reference)	0.12 (0.12-0.12)
Tablet recipients before consult	2.95 (2.91-2.99)	0.21 (0.20-0.21)	4.47 (4.36-4.59)	0.55 (0.53-0.56)
Tablet recipients after consult	2.73 (2.70-2.76)	0.19 (0.19-0.19)	4.27 (4.20-4.34)	0.52 (0.52-0.53)

[a]For the tablet cohort, we report video visits in the first 7 months of tablet receipt but exclude the first month of tablet receipt from models as tablet assignment occurred in these months and we did not want to attribute tablet assignment-related visits to the tablet-associated outcome; therefore, we had a maximum of 6 months of follow-up after tablet receipt among tablet recipients.

[b]Models were adjusted for age at tablet receipt; gender; race; ethnicity; rurality of the home; VHA enrollment priority category; number of chronic conditions; presence of a mental health condition; current marital status; drive time to the closest primary VHA facility; any hospitalization in the 180 days before March 1, 2020; any suicide flag in the 12 months before March 1, 2020; history of housing instability; history of homelessness; history of video, phone, and remote patient monitoring in 2019; history of primary care, mental health, specialty care use, and other VHA care (diagnostic/ancillary care) in 2019; and VHA medical center. Each model also included fixed effects for the calendar month and standard errors accounted for clustering at the patient level.

[c]VHA: Veterans Health Administration.

XSL•FO
RenderX

In a secondary analysis assessing the use of phone care, veterans who received a tablet before the implementation of the Digital Divide Consult had 1.26 (95% CI 1.25-1.27) times the likelihood of having a phone-based visit, while veterans who received their tablet after consult implementation had a RR of 1.48 (95% CI 1.47-1.49) compared to the general VHA population (Multimedia Appendix 3). Veterans with tablets also had a higher predicted number of phone visits, with 0.73 (95% CI 0.72-0.74) phone visits per month among veterans who received their tablet before consult implementation and 0.83 (95% CI 0.83-0.84) phone visits per month among those who received their tablet after consult implementation in the first 6 months of tablet possession compared to 0.44 (95% CI 0.44-0.44) visits per month among the general VHA population.

Video-Based Services by Referral Reason in the Digital Divide Consult

Among the 79,230 veterans who received a tablet with a Digital Divide Consult, the most common reason for receiving a tablet was having a mental health diagnosis (50,367/79,230, 63.6%), followed by living more than 30 miles away (17,228/79,230, 21.7%) and social isolation (16,161/79,230, 20.4%) (Table 4; Multimedia Appendix 4). The mean number of criteria checked was 1.97 (25th percentile-75th percentile: 1-3). Approximately 5407 veterans received a tablet with a consult stating they met no specific criteria or with no consult reason selected. A manual review of these consults highlighted additional reasons that veterans were referred for tablets, including avoiding health care facilities due to COVID-19, perceptions that a person might benefit from a device to increase patient-provider communication, and enrollment in a VHA health care program that relies on virtual appointments such as the VHA MOVE weight loss program.

On average, 63.0% (49,925/79,230) of individuals who received a tablet with a Digital Divide Consult had a video visit in the first 6 months of tablet receipt. Some Digital Divide Consult reasons were associated with a higher proportion of tablet recipients using video telehealth (Table 4). These reasons included being enrolled in evidence-based mental health programs (74.8% [830/1100] with any video use), living more than 30 miles away from a VHA center (68.3% [10,557/17,228] with any video use), having a mental health diagnosis (68.1% [34,301/50,367] with any video use), and finding in-person visits challenging (65.1% [4215/6472] with any video use). Other reasons were associated with a lower proportion of tablet recipients using video telehealth, such as disruptive behavior (56.9% [230/404] with any video use), being in hospice (55.2% [219/397] with any video use), and not meeting any of the specific criteria at the consult (58.4% [3159/5407] with any video use).

The mean number of video visits in the first 6 months of tablet receipt also varied by consult reason selected (Table 4). Consult reasons associated with a higher adjusted mean number of video visits per month included evidence-based mental health (7.58), social isolation (5.44), and any mental health condition (5.33). The mean number of visits in the first 6 months of tablet receipt among tablet recipients with no criteria selected was 4.23 (95% CI 3.98-4.48), and the mean number of visits with any of the criteria selected was 4.77 (95% CI 4.69-4.84) (*t*-test for difference $P<.01$). As the mean numbers can be influenced by outliers of video telehealth use, we also describe the proportion of tablet recipients having no visits, 1-3 visits, and 4 or more visits in a 6-month period in Multimedia Appendix 5. There was no relationship between the probability of having a video visit in the first 6 months and the number of consult reasons selected; however, there was a slightly increasing linear relationship between the number of consult reasons selected and the predicted number of video visits in a month (Multimedia Appendix 6).

XSL•FO
RenderX

Table 4. Percentage of Veterans Health Administration tablet recipients with any video use and mean number of video visits by consult reason (Digital Divide Consult) (N=79,230).

Digital Divide Consult reason[a]	Tablet recipients reporting a given Digital Divide Consult reason, n (%)	Percentage of tablet recipients with any video use[b], n (%)	Unadjusted mean number of video visits[c] in 6 months by reason reported, value (95% CI)	Adjusted mean number of video visits[d] in 6 months by reason reported, value (95% CI)
Enrolled in evidence-based mental health[e]	1100 (1.4)	830 (74.8)	8.60 (7.82-9.39)	7.58 (7.09-8.09)
Lives more than 30 miles from a VHA facility[f]	17,228 (21.7)	10,557 (68.3)	4.01 (3.89-4.14)	4.53 (4.45-4.62)
Any mental health diagnosis	50,367 (63.6)	34,301 (68.1)	5.70 (5.61-5.80)	5.33 (5.28-5.39)
Work, school, or caregiver commitments make in-person visits challenging	6472 (8.2)	4215 (65.1)	4.62 (4.40-4.84)	4.33 (4.21-4.45)
Homeless veteran or enrolled in the HUD-VASH[g] Program	6434 (8.1)	3988 (62.0)	4.97 (4.72-5.21)	4.21 (4.10-4.33)
Social isolation	16,161 (20.4)	10,016 (62.0)	5.24 (5.07-5.41)	5.44 (5.34-5.53)
Difficulty attending a VHA facility (ie, immunocompromised state or psychological distress)	13,324 (16.8)	8165 (61.3)	4.39 (4.23-4.55)	4.66 (4.57-4.75)
Cost of attending in-person is prohibitive	4608 (5.8)	2809 (61.0)	4.68 (4.40-4.96)	4.74 (4.68-4.89)
Hospitalized in the last 90 days	6401 (8.1)	3823 (59.7)	5.01 (4.74-5.23)	4.87 (4.74-5.01)
No specific criteria or no criteria selected	5407 (6.8)	3159 (58.4)	4.23 (3.98-4.48)	4.40 (4.27-4.54)
Health issues make the veteran homebound	13,280 (16.8)	7739 (58.3)	3.21 (3.09-3.33)	4.13 (4.04-4.22)
Difficulty with public transportation	9354 (11.8)	5383 (57.6)	3.79 (3.61-3.98)	4.28 (4.18-4.38)
No car or access to a ride	7983 (10.1)	4575 (57.3)	4.36 (4.13-4.59)	4.30 (4.19-4.41)
No veteran transport or DAV[h]	2743 (3.5)	1563 (57.0)	3.88 (3.56-4.21)	4.23 (4.05-4.42)
Documented disruptive behavior at VHA	404 (0.5)	230 (56.9)	3.99 (3.15-4.82)	3.56 (3.17-3.96)
Veteran has a hospice diagnosis[i]	397 (0.5)	219 (55.2)	3.52 (2.76-4.28)	3.87 (3.43-4.31)

[a]More than one criterion could be selected (mean number of criteria checked was 1.97 [25th percentile-75th percentile: 1-3]).

[b]Results are sorted by ranked percentage of tablet recipients with any video use.

[c]Unadjusted mean number of video visits was calculated using the *t*-test.

[d]The adjusted mean number of video visits is the average marginal effect where the independent variable was the selection of one of the Digital Divide Consult reasons (ie, reasons were not mutually adjusted for) and the dependent variable was the count of video visits (including veterans with no or zero video visits). Models were adjusted for age at tablet receipt, gender, race, ethnicity, and rurality of the home.

[e]Analytic sample for the "evidence-based mental health" consult reason has been restricted to veterans who received their tablets after March 1, 2020, when this reason was added (N=17,746).

[f]VHA: Veterans Health Administration.

[g]HUD-VASH: US Department of Housing and Urban Development-VA Supportive Housing.

[h]DAV: Disabled American Veterans.

[i]Analytic sample for the "hospice" consult reason has been restricted to veterans who received their tablets after May 1, 2021, when this reason was added (N=43,933).

Discussion

Since 2016, VHA's Connected Device Program has provided tablets to over 180,000 veterans. To our knowledge, this unique program was the first of its size in the United States to bridge the digital divide by offering patients a device to engage in video telehealth use. Findings from this evaluation suggest that the program has reached many veterans from groups that have experienced health care access disparities, including women veterans, gender-diverse veterans, veterans who are black, veterans who have a history of homelessness, and those with low-income priority status in VHA. While these patient characteristics are not explicitly targeted as a criterion in the Digital Divide Consult, supporting veterans who have historically faced health care access barriers and disparities aligns with VHA's Office of Health Equity. However, the underuse of tablets (only 63% of recipients after the Digital Divide Consult had video use within 6 months) indicates an opportunity for additional implementation strategies to ensure that those who receive this resource have the support they need to engage in video telehealth use.

In this evaluation, 68.4% (81,985/119,926) of tablet recipients lived in urban areas compared to a previous evaluation, where 47% of tablet recipients between May 1, 2016, and September

XSL•FO
RenderX

30, 2017, lived in urban areas [13]. This shift is likely a result of the COVID-19 pandemic's restriction on in-person care and the expansion of video-based care availability in multiple care services [9,37,38]. An evaluation of the characteristics of tablet recipients before and during the COVID-19 pandemic found that 55% of prepandemic tablet recipients were urban-dwelling compared with 68% during the pandemic but before the implementation of the Digital Divide Consult, which closely matches our result of 68.4% [15].

This evaluation also found that the tablet population has become more racially diverse, older, and more likely to be single compared with earlier evaluations [13,15]. This may be a reflection of the broadening of eligibility reasons for a tablet receipt in the Digital Divide Consult. Between 2016 and 2017, the eligibility criteria for access barriers included only distance or geography, transportation issues, difficulty leaving home, and other reasons described by the provider [13]. The increased tablet allotment to veterans with a history of hospitalization, a suicide risk flag, a history of mental health conditions, and a disability-related priority status suggests that tablets are directed to veterans with complex medical needs as intended and may be a direct result of the inclusion of these complex medical needs as a criterion for tablet receipt. Interestingly, the most common reason for tablet recipients was the presence of a mental health condition. This increased distribution of tablets among this population may be a consequence of a higher uptake of video-based care among mental health service providers [38].

The Digital Divide Consult was designed by OCC as a mechanism to standardize screening and referrals for VHA-issued tablets and internet service [39]. This consult improved the device-lending process by introducing a social worker assessment to identify and resolve barriers to care such as internet connectivity needs. We found that tablet receipt was associated with elevated likelihoods and rates of using video telehealth. Yet, there was a small but statistically significant decrease in the likelihood and rate of video telehealth use before versus after implementation of the Digital Divide Consult. Tablet recipients before the consult had 2.95 times the likelihood of a video visit and 4.47 times the incidence rate compared with the general VHA population in the same month. The RR of 2.73 and IRR of 4.27 were slightly smaller among tablet recipients after the consult compared with the general VHA population in the same month. While we adjusted for the calendar month in our analysis and made comparisons within a given month, we may not have fully accounted for temporal differences such as video care availability in VHA across preconsult and postconsult months. As seen in Figure 1, the preconsult period of April 1, 2020, to October 1, 2020, was marked by abnormally high availability of video care. In comparison, the postconsult period of October 1, 2020, through the end of follow-up, August 31, 2023, was marked by a reduction in video care and an increase in in-person care availability, establishing the new equilibrium of telemedicine following the COVID-19 pandemic [38]. Even so, the difference between pre- and postconsult tablet recipients is likely clinically nonsignificant as it translates to a difference of an additional visit every 3 years, assuming the rate of visits remains the same over time.

Veterans who received a tablet also had increased engagement in phone visits compared to the general VHA population, suggesting that tablets may increase veteran engagement with VHA beyond video visits through increased patient-provider communication as tablets are not set up to complete phone calls. This is consistent with prior work that demonstrated providing tablets to veterans also increased engagement in health information and management tasks in the veterans' My Health*e*Vet patient portal [40].

Despite the successes of VHA's Connected Device Program, challenges remain. The rate of video visit use among tablet recipients is lower than initially expected, and this pattern has been observed in tablet distribution programs outside VHA [41], suggesting a need for interventions to ensure that all veterans who receive VHA-issued tablets are able to use them [17]. In 2021, OCC launched the Connected Devices Support Program to build digital literacy among recipients of VHA-issued tablets. As a part of this program, a technician offers device education to veteran recipients of loaned tablets and a video test call ahead of the first video visit with their VHA care team. Other findings, such as the observed variation in video visit use among veterans who met specific criteria for a tablet, can inform implementation strategies within clinical services. For example, veterans who were referred for tablets to participate in a specific treatment protocol (ie, evidence-based mental health) were much more likely to have a subsequent video visit than those who were referred for more general reasons such as transportation barriers. This suggests that coupling tablet distribution with a specific treatment program or service might increase the likelihood of tablet use. The increased predicted number of video visits among evidence-based mental health and mental health–related consult reasons may reflect the underlying treatment protocols. For example, mental health visits in programs like the Intensive Community Mental Health Recovery (ICMHR) or Substance Use Disorder Intensive Outpatient Programs occur at a high intensity, corresponding to an average frequency of 7 to 15 visits per month [42], while primary care visits occur much less frequently, leading to a differential predicted number of visits by consult reasons.

Several limitations should be considered when interpreting these findings. First, the analysis does not provide information on whether video visits occurred specifically through the tablets. As veterans were screened for the possession of a video-enabled device, we assume that a subsequent video visit occurred using the VHA-provided tablet. Moreover, the study's focus on video may have overlooked other activities that could potentially contribute to veterans' health care and health navigation, such as their interaction with the VHA's patient portal and My HealtheVet, or the use of various health-related mobile apps. Additional research from the Virtual Access QUERI indicates that tablets increase engagement in patient portals, which may reduce the need for a subsequent visit (eg, medication management requests through messaging rather than an appointment).

Second, our results may be limited as the general veteran population may be different from veterans who received a tablet and, more specifically, those who received a tablet through the Digital Divide Consult (ie, exchangeability assumption). In our

XSL•FO
RenderX

analysis, we accounted for differences between the 2 populations by identifying and adjusting for key patient demographics associated with the differential use of video telehealth (eg, age, history of VHA use, and rurality of the home) [10] and characteristics similar to those captured in the Digital Divide Consult. We identified electronic health record equivalents for Digital Divide Consult reasons such as drive time to a VHA facility, mental health conditions, history of hospitalization, housing instability, history of homelessness, and suicide risk. We were unable to identify some Digital Divide Consult reasons that referred to social needs as these are not captured in the current electronic health record (eg, social isolation, difficulty attending a VHA facility, cost of attending is prohibitive, and no car or access to VHA). Thus, our results may be biased by nonexchangeability between the study groups. An alternative to our statistical adjustment would be emulating multiple nested trials, with each having a different start of follow-up (n=40 for each calendar month for tablet receipt), which is complicated to implement [43].

Additionally, our choice of the baseline period for veteran demographics and characteristics may influence our results. We chose a baseline period prior to the time of the first tablet receipt for the entire population, which may be less accurate for tablet recipients in the later years of our evaluation. However, most of these variables, such as race, rurality, and gender, are unlikely to change over the course of the study. However, 4 variables in our analysis represent more acute and temporally relevant windows. Changes in these variables, such as a new occurrence of suicide risk or homelessness, which may have increased in the acute phase of the pandemic, may bias our adjustment of baseline differences in use between tablet recipients and the general VHA population. Despite this, our approach is relatively robust as we compared general trends in patient demographics and video care use and as the sample size of the comparator population was sufficiently large (temporal variations in these 4 confounding variables will not unduly influence our analysis). This is a limitation of our approach and of observational data in general when eligibility criteria for the population of interest are met at multiple times [29].

In prior evaluations of the Connected Devices Program, statistical matching on patient characteristics has been used to compare the use of VHA health care among tablet recipients with more similar veterans from the general VHA population [18,19]. In this evaluation, matching was not used, as one of our primary questions was to compare the distribution of population characteristics among tablet recipients with the general VHA population. Matching these patient characteristics would prevent us from comparing these distributions between the 2 groups in our first objective.

Finally, we considered veterans who received tablets after the implementation of the Digital Divide Consult as part of the “postdigital divide” group regardless of whether there was an associated Digital Divide Consult for each patient. Among the postconsult group, 90% of the consults were identified, leading to an assumption of an intent-to-treat framework, which may result in underestimating the actual relationship between Digital Divide Consult implementation and the use of video or phone care. The frequency of tablet recipients after consult with an identifiable consult increased with time, indicating a growing use of the Digital Divide Consult. These limitations should be considered when interpreting the study’s findings and their implications for health care policy and practice.

In summary, the VHA’s Digital Divide Consult offers information about why veterans receive tablets and the degree to which different subgroups, who are often left out of technology initiatives, may use tablets. Findings from this evaluation suggest that VHA’s tablet distribution effort is reaching veterans from a number of high-need groups and has likely facilitated the use of video telehealth. Ongoing evaluation activities are examining the cost implications of VHA’s Connected Device Program, the experiences of tablet recipients who do not use or who underuse their tablets, and the use of tablets for engagement in health care outside of video visits. With the expansion of video telehealth as a modality for health care delivery, VHA’s Connected Device Program can serve as a model for other health care systems that aim to expand access for individuals who lack their own devices.

Acknowledgments

The authors thank Camila Chaudhary, the US Department of Veterans Affairs (VA) Office of Connected Care, the VA Office of Rural Health, and the Virtual Access Quality Enhancement Research Initiative (QUERI) team for their help in supporting this evaluation. Funding was provided by the US Department of VA Office of Rural Health. This work was supported by VA QUERI PEI-18-205. JMF was directly supported by VA QUERI to Advance Diversity in Implementation Leadership (ADIL) supplementary to QUERI PEI-18-205. The views expressed are those of the authors, and the contents do not represent the views of the US Department of Veterans Affairs or the US Government.

Data Availability

Owing to US Department of Veterans Affairs (VA) regulations and our ethics agreements, the analytic data sets used for this study are not permitted to leave the VA security system without a data use agreement. This limitation is consistent with other studies based on VA data. However, VA data are made freely available to researchers with an approved VA study protocol. For more information, please visit [44] or contact the VA Information Resource Center at VIReC@va.gov.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed sampling methodology and description of variable composition and definition.
[DOCX File , 21 KB - jmir_v26i1e59089_app1.docx]

Multimedia Appendix 2
History of Veterans Health Administration (VHA) outpatient health care use (March 11, 2019, to March 10, 2020) among tablet recipients and the general VHA population.
[DOCX File , 18 KB - jmir_v26i1e59089_app2.docx]

Multimedia Appendix 3
Monthly phone visit use comparing the general Veterans Health Administration population, tablet recipients before Digital Divide Consult, and tablet recipients after Digital Divide Consult.
[DOCX File , 15 KB - jmir_v26i1e59089_app3.docx]

Multimedia Appendix 4
Prevalence of the reasons for the Digital Divide Consult, ordered by frequency selected, among 79,230 tablet recipients. Multiple criteria can be selected for one consult. Commitments that make in-person visits challenging include school, work, or caregiver commitments. Difficulty attending VHA included reasons like immunocompromised state or psychological distress. Hospice consult and evidence-based consult reasons were added in May 1, 2021, and March 1, 2020, respectively. DAV: Disabled American Veterans; HUD-VASH, US Department of Housing and Urban Development-Veterans Affairs Supportive Housing; VHA: Veterans Health Administration.
[PNG File , 194 KB - jmir_v26i1e59089_app4.png]

Multimedia Appendix 5
Mean count of video visits among veterans with a Digital Divide Consult (N=79,230) by consult reason.
[DOCX File , 18 KB - jmir_v26i1e59089_app5.docx]

Multimedia Appendix 6
Adjusted association between the number of consult reasons selected and the probability of any video use and change in the predicted number of video visits in a month among veterans with a Digital Divide Consult (N=79,230).
[DOCX File , 16 KB - jmir_v26i1e59089_app6.docx]

References

1. VA Health Care Priorities. Veterans Affairs. URL: https://www.va.gov/health/priorities/ [accessed 2023-11-10]
2. Retirement and Socioeconomic Characteristics of Aged Veterans: Differences by Education and Race/Ethnicity. Social Security Administration. URL: https://www.ssa.gov/policy/docs/ssb/v79n1/v79n1p1.html [accessed 2023-11-10]
3. Rural Veterans: Rural Veteran Health Care Challenges. U.S. Department of Veterans Affairs. URL: https://www.ruralhealth.va.gov/aboutus/ruralvets.asp [accessed 2023-11-10]
4. VA Telehealth Services celebrates 20 years. VA News. URL: https://news.va.gov/118570/va-telehealth-services-celebrates-20-years/ [accessed 2023-11-10]
5. Heyworth L, Kirsh S, Zulman DM, Ferguson JM, Kizer K. Expanding access through virtual care: The VA's early experience with Covid-19. NEJM Catal Innov Care Deliv 2020 Jul 01;1(4):11 [FREE Full text]
6. Department of Veterans Affairs. 2019 Survey of Veteran Enrollees? Health and Use of Health Care- Data Findings Report. Veterans Affairs. URL: https://www.va.gov/HEALTHPOLICYPLANNING/SOE2019/2019_Enrollee_Data_Findings_Report-March_2020_508_Compliant.pdf [accessed 2024-08-20]
7. America's Digital Divide. U.S. Congress Joint Economic Committee. URL: https://www.jec.senate.gov/public/_cache/files/ff7b3d0b-bc00-4498-9f9d-3e56ef95088f/the-digital-divide-.pdf [accessed 2024-08-20]
8. O'Shea AMJ, Howren MB, Mulligan K, Haraldsson B, Shahnazi A, Kaboli PJ. Quantifying the digital divide: Associations of broadband internet with tele-mental health access before and during the COVID-19 pandemic. J Gen Intern Med 2023 Jul;38(Suppl 3):832-840 [FREE Full text] [doi: 10.1007/s11606-023-08120-8] [Medline: 37340258]
9. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the Veterans Health Administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc 2021 Mar 01;28(3):453-462 [FREE Full text] [doi: 10.1093/jamia/ocaa284] [Medline: 33125032]

XSL•FO
RenderX

10. Ferguson JM, Wray CM, Jacobs J, Greene L, Wagner TH, Odden MC, et al. Variation in initial and continued use of primary, mental health, and specialty video care among Veterans. Health Serv Res 2023 Apr;58(2):402-414 [FREE Full text] [doi: 10.1111/1475-6773.14098] [Medline: 36345235]
11. John S. McCain III, Daniel K. Akaka, and Samuel R. Johnson VA Maintaining Internal Systems and Strengthening Integrated Outside Networks Act of 2018 (S. 2372). Congress. URL: https://www.congress.gov/115/bills/s2372/BILLS-115s2372enr.pdf [accessed 2023-11-10]
12. Bridging the Digital Divide. Telehealth VA. URL: https://telehealth.va.gov/digital-divide [accessed 2023-11-10]
13. Zulman DM, Wong EP, Slightam C, Gregory A, Jacobs JC, Kimerling R, et al. Making connections: nationwide implementation of video telehealth tablets to address access barriers in veterans. JAMIA Open 2019 Oct;2(3):323-329 [FREE Full text] [doi: 10.1093/jamiaopen/ooz024] [Medline: 32766533]
14. Evaluating implementation of VA issued tablets for high-risk Veterans. US Department of Veterans Affairs. URL: https://www.queri.research.va.gov/centers/VideoTablets.pdf [accessed 2024-08-20]
15. Dhanani Z, Ferguson JM, Van Campen J, Slightam C, Jacobs JC, Heyworth L, et al. Overcoming access barriers for veterans: Cohort study of the distribution and use of Veterans Affairs' video-enabled tablets before and during the COVID-19 pandemic. J Med Internet Res 2023 Jan 26;25:e42563 [FREE Full text] [doi: 10.2196/42563] [Medline: 36630650]
16. Garvin LA, Hu J, Slightam C, McInnes DK, Zulman DM. Use of video telehealth tablets to increase access for veterans experiencing homelessness. J Gen Intern Med 2021 Aug;36(8):2274-2282 [FREE Full text] [doi: 10.1007/s11606-021-06900-8] [Medline: 34027612]
17. Slightam C, Gregory AJ, Hu J, Jacobs J, Gurmessa T, Kimerling R, et al. Patient perceptions of video visits using Veterans Affairs telehealth tablets: Survey study. J Med Internet Res 2020 Apr 15;22(4):e15682 [FREE Full text] [doi: 10.2196/15682] [Medline: 32293573]
18. Jacobs JC, Blonigen DM, Kimerling R, Slightam C, Gregory AJ, Gurmessa T, et al. Increasing mental health care access, continuity, and efficiency for veterans through telehealth with video tablets. Psychiatr Serv 2019 Nov 01;70(11):976-982. [doi: 10.1176/appi.ps.201900104] [Medline: 31378193]
19. Gujral K, Van Campen J, Jacobs J, Kimerling R, Blonigen D, Zulman DM. Mental health service use, suicide behavior, and emergency department visits among rural US veterans who received video-enabled tablets during the COVID-19 pandemic. JAMA Netw Open 2022 Apr 01;5(4):e226250 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.6250] [Medline: 35385088]
20. Wray CM, Ferguson JM, Greene L, Griffin A, Van Campen J, O'Shea AM, et al. Overcoming obstacles: barriers to virtual care use among video-enabled tablet recipients in the Veterans Health Administration. J Gen Intern Med 2024 Mar;39(4):549-556 [FREE Full text] [doi: 10.1007/s11606-023-08468-x] [Medline: 37914909]
21. Slightam C, SooHoo S, Greene L, Zulman DM, Kimerling R. Development and validation of a measure to assess patient experiences with video care encounters. JAMA Netw Open 2024 Apr 01;7(4):e245277 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.5277] [Medline: 38578639]
22. Affordable Connectivity Program. Federal Communications Commission. URL: https://www.fcc.gov/acp [accessed 2023-11-10]
23. EveryoneOn. URL: https://www.everyoneon.org/ [accessed 2023-11-10]
24. Lifeline Program for Low-Income Consumers. Federal Communications Commission. URL: https://www.fcc.gov/general/lifeline-program-low-income-consumers [accessed 2023-11-10]
25. Federal Register Volume 79, Issue 17 (January 27, 2014). GovInfo. URL: https://www.govinfo.gov/app/details/FR-2014-01-27/2014-01497/summary [accessed 2024-08-20]
26. Connolly SL, Charness ME, Miller CJ. To increase patient use of video telehealth, look to clinicians. Health Serv Res 2023 Feb;58(1):5-8 [FREE Full text] [doi: 10.1111/1475-6773.14041] [Medline: 35904201]
27. Dhanani Z, Ferguson JM, Van Campen J, Slightam C, Heyworth L, Zulman DM. Adoption and sustained use of primary care video visits among veterans with VA video-enabled tablets. J Med Syst 2024 Jan 30;48(1):16. [doi: 10.1007/s10916-024-02035-5] [Medline: 38289373]
28. Kakani P, Sorensen A, Quinton JK, Han M, Ong MK, Kamdar N, et al. Patient characteristics associated with telemedicine use at a large academic health system before and after COVID-19. J Gen Intern Med 2021 Apr;36(4):1166-1168 [FREE Full text] [doi: 10.1007/s11606-020-06544-0] [Medline: 33506387]
29. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016 Apr 15;183(8):758-764 [FREE Full text] [doi: 10.1093/aje/kwv254] [Medline: 26994063]
30. Yu W, Ravelo A, Wagner TH, Phibbs CS, Bhandari A, Chen S, et al. Prevalence and costs of chronic conditions in the VA health care system. Med Care Res Rev 2003 Sep;60(3 Suppl):146S-167S. [doi: 10.1177/1077558703257000] [Medline: 15095551]
31. Fishman P, Von Korff M, Lozano P, Hecht J. Chronic care costs in managed care. Health Aff (Millwood) 1997;16(3):239-247. [doi: 10.1377/hlthaff.16.3.239] [Medline: 9141341]
32. Ray GT, Collin F, Lieu T, Fireman B, Colby CJ, Quesenberry CP, et al. The cost of health conditions in a health maintenance organization. Med Care Res Rev 2000 Mar;57(1):92-109. [doi: 10.1177/107755870005700106] [Medline: 10705704]

XSL•FO
RenderX

33. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Committee to Evaluate the Department of Veterans Affairs Mental Health Services. Evaluation of the Department of Veterans Affairs Mental Health Services. Washington, DC: National Academies Press; 2018.
34. Tsai J, Szymkowiak D, Jutkowitz E. Developing an operational definition of housing instability and homelessness in Veterans Health Administration's medical records. PLoS One 2022;17(12):e0279973 [FREE Full text] [doi: 10.1371/journal.pone.0279973] [Medline: 36584201]
35. Code of Federal Regulations: Basic HHS Policy for Protection of Human Research Subjects. U.S. Department of Health and Human Services. 2018 Jun 19. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html#46.104
36. Protection of Human Subjects. Code of Federal Regulations: A point in time eCFR system. 2024 Mar 09. URL: https://www.ecfr.gov/current/title-38/chapter-I/part-16
37. Heyworth L, Shah N, Galpin K. 20 years of telehealth in the Veterans Health Administration: Taking stock of our past and charting our future. J Gen Intern Med 2024 Feb;39(Suppl 1):5-8 [FREE Full text] [doi: 10.1007/s11606-024-08617-w] [Medline: 38378981]
38. Ferguson JM, Wray CM, Van Campen J, Zulman DM. A new equilibrium for telemedicine: prevalence of in-person, video-based, and telephone-based care in the Veterans Health Administration, 2019-2023. Ann Intern Med 2024 Feb;177(2):262-264. [doi: 10.7326/M23-2644] [Medline: 38252945]
39. VA telehealth services through the Digital Divide Consult. VA News. URL: https://news.va.gov/104490/va-telehealth-services-through-the-digital-divide-consult/ [accessed 2023-11-10]
40. Griffin AC, Troszak LK, Van Campen J, Midboe AM, Zulman DM. Tablet distribution to veterans: an opportunity to increase patient portal adoption and use. J Am Med Inform Assoc 2022 Dec 13;30(1):73-82 [FREE Full text] [doi: 10.1093/jamia/ocac195] [Medline: 36269168]
41. Loizos M, Baim-Lance A, Ornstein KA, Lubetsky S, Salas-Belkin L, Chavez S, et al. If you give them away, it still may not work: Challenges to video telehealth device use among the urban homebound. J Appl Gerontol 2023 Sep;42(9):1896-1902 [FREE Full text] [doi: 10.1177/07334648231170144] [Medline: 37070328]
42. Evidence-Based Therapy. US Department of Veterans Affairs. URL: https://www.mentalhealth.va.gov/get-help/treatment/ebt.asp [accessed 2023-11-10]
43. Danaei G, Rodríguez LAG, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. Stat Methods Med Res 2013 Feb;22(1):70-96 [FREE Full text] [doi: 10.1177/0962280211403603] [Medline: 22016461]
44. VA Information Resource Center. U.S. Department of Veterans Affairs. URL: https://www.virec.research.va.gov [accessed 2024-08-22]

Abbreviations

CDW: Corporate Data Warehouse
IRR: incidence rate ratio
OCC: Office of Connected Care
ORH: Office of Rural Health
QUERI: Quality Enhancement Research Initiative
RR: risk ratio
VHA: Veterans Health Administration

Edited by G Eysenbach; submitted 02.04.24; peer-reviewed by H Shoemaker, E Vandersteen Bailey, ME Price; comments to author 25.05.24; revised version received 22.07.24; accepted 09.08.24; published 09.09.24.

Please cite as:
Ferguson JM, Van Campen J, Slightam C, Greene L, Heyworth L, Zulman DM
Evaluation of the Veterans Health Administration's Digital Divide Consult for Tablet Distribution and Telehealth Adoption: Cohort Study
J Med Internet Res 2024;26:e59089
URL: https://www.jmir.org/2024/1/e59089
doi:10.2196/59089
PMID:39250183

©Jacqueline M Ferguson, James Van Campen, Cindie Slightam, Liberty Greene, Leonie Heyworth, Donna M Zulman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Research Into Digital Health Intervention for Mental Health: 25-Year Retrospective on the Ethical and Legal Challenges

Charlotte L Hall[1,2,3], PhD; Aislinn D Gómez Bergin[3,4], PhD; Stefan Rennick-Egglestone[2,5], PhD

[1]Institute of Mental Health, Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[2]National Institute for Health and Care Research (NIHR) Nottingham Biomedical Research Centre, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

[3]National Institute for Health and Care Research (NIHR) MindTech HealthTech Research Centre, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

[4]Responsible AI UK, School of Computer Science, University of Nottingham, Nottingham, United Kingdom

[5]School of Health Sciences, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Stefan Rennick-Egglestone, PhD
School of Health Sciences
Institute of Mental Health
University of Nottingham
University of Nottingham Innovation Park
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 11582 ext 30926
Email: stefan.egglestone@nottingham.ac.uk

Abstract

Digital mental health interventions are routinely integrated into mental health services internationally and can contribute to reducing the global mental health treatment gap identified by the World Health Organization. Research teams designing and delivering evaluations frequently invest substantial effort in deliberating on ethical and legal challenges around digital mental health interventions. In this article, we reflect on our own research experience with digital mental health intervention design and evaluation to identify 8 of the most critical challenges that we or others have faced, and that have ethical or legal consequences. These include: (1) harm caused by online recruitment work; (2) monitoring of intervention safety; (3) exclusion of specific demographic or clinical groups; (4) inadequate robustness of effectiveness and cost-effectiveness findings; (5) adequately conceptualizing and supporting engagement and adherence; (6) structural barriers to implementation; (7) data protection and intellectual property; and (8) regulatory ambiguity relating to digital mental health interventions that are medical devices. As we describe these challenges, we have highlighted serious consequences that can or have occurred, such as substantial delays to studies if regulations around Software as a Medical Device (SaMD) are not fully understood, or if regulations change substantially during the study lifecycle. Collectively, the challenges we have identified highlight a substantial body of required knowledge and expertise, either within the team or through access to external experts. Ensuring access to knowledge requires careful planning and adequate financial resources (for example, paying public contributors to engage in debate on critical ethical issues or paying for legal opinions on regulatory issues). Access to such resources can be planned for on a per-study basis and enabled through funding proposals. However, organizations regularly engaged in the development and evaluation of digital mental health interventions should consider creating or supporting structures such as advisory groups that can retain necessary competencies, such as in medical device regulation.

(J Med Internet Res 2024;26:e58939) doi:10.2196/58939

KEYWORDS

digital mental health intervention; research ethics; compliance; regulation; digital health; mobile health; mhealth; intervention; interventions; mental health; retrospective; ethical; legal; challenge; challenges

XSL•FO
RenderX

Introduction

The World Health Organization has identified a treatment gap for mental health [1], which is present across all regions and for all mental health conditions [2]. For some conditions and in some settings, the treatment gap is increasing. For example, the depression treatment gap in the United States increased between 2015 and 2019 due to an increased depression prevalence in younger people, with no commensurate increase in treatment [3].

Digital mental health interventions (DMHIs) are now routinely integrated into mental health services internationally and can extend provision to those not currently using mental health services [4]. DMHIs can be used to support the prevention, assessment, diagnosis, treatment, and management of mental health conditions. For this review, we focus mainly on treatment and management. DMHIs can enable the delivery of treatment outside of psychiatric settings, a factor identified by a World Psychiatric Association member survey as extremely important for improving access to mental health care, particularly for health conditions other than schizophrenia and those experienced in childhood [5]. Even in the case of schizophrenia and mental health conditions experienced in childhood, studies have shown that treatment through DMHIs can be safe and effective in nonpsychiatric settings [6,7]; hence, we would argue that they are relevant transdiagnostically.

DMHIs can be more easily scalable than nondigital treatments, potentially narrowing the global treatment gap, particularly given a rapid increase in access to computing and networking technologies globally, even in low-resource settings [8]. However, despite this potential for scalability, the majority of DMHIs are not widely integrated into health care systems [9], and varying access to mobile technologies continues to be reported as a barrier to their use to narrow the treatment gap in low- and middle-income settings [10]. Furthermore, for the subset of DMHIs that integrate artificial intelligence (AI), a systematic review has highlighted biases in the mental health conditions examined in evaluations and deficits in the methodological quality of research. These examples point to continuing barriers that limit integration into evidence-based practice [11].

Over the past quarter century, the DMHI landscape has swiftly progressed, enabled by rapid technological advances during this period. This has resulted in an ever-increasing range of DMHIs, including those delivered as mobile apps, internet-delivered therapy, AI, extended and virtual reality, and remote measurement technology (to name just a few). The mode of digital complexity of these DMHIs may vary from low (eg, the digital platform used as a tool for delivery such as videoconferencing) to high (eg, personalization of therapeutic content through machine learning technologies). The level of human input into these DMHI may also vary from fully self-directed (eg, mental health app with no clinical input), to asynchronous motivational support (eg, a therapist provides motivational texts to encourage engagement with a digital tool), through to fully supported (eg, real-time delivered clinician therapy via videoconferencing) [12].

Many DMHIs are publicly available, often for free or at a low cost, such as through public app stores. Most app store offerings have no scientific support or involvement in their development process [13], with the consequence that some can recommend harmful strategies, such as consuming strong alcohol during a bipolar episode [14]. However, DMHIs have been an important focus of mental health research, including high-quality randomized controlled trials (RCTs). For example, in a recent review of trials included in the International Standard Randomized Controlled Trial Numbers (ISRCTN) registry, we identified 23 DMHI trials that had published findings [15]. Since DMHIs are delivered using digital technologies, their evaluations are frequently delivered using technology as well. In online trials, some or all trial procedures are delivered online, frequently (but not always) on the web, which brings a range of methodological challenges [16]. DMHIs often have a small effect size; a meta-analysis of evaluations of DMHIs for depression in adults found a pooled effect size of Cohen d=0.33 (95% CI 0.11-0.55) in evaluations of unsupervised DMHIs [17]. Since a small anticipated effect size necessitates a large target sample, just one challenge in online trials can be finding sufficient eligible participants.

The authors have published and contributed to RCTs of DMHI evaluations, including novel interventions for tics [18,19], attention deficit hyperactivity disorder (ADHD) [20], virtual reality therapy for people with psychosis [7], and the use of mental health recovery narratives as an intervention [21]. We have observed that research teams designing and delivering evaluations frequently invest substantial effort in deliberating on ethical and legal challenges around DMHIs. Conducting our own evaluations ethically and legally has been one of our primary concerns. This focus on ethics and legality can arise because evaluations of DMHIs introduce challenges that are less present in evaluations of more traditional clinical-led interventions, such as less or no contact with clinical teams and the impact this has on monitoring participants (for example, to identify adverse events). DMHIs can result in harm through different mechanisms, such as usage by people experiencing a desperate need for mental health improvement who are unsuited to a particular offering [22] but face a lower barrier to entry to a DMHI in a web-based trial with a self-rated inclusion assessment. Legal challenges can arise when there are rapid changes in international law, such as around data protection arrangements for data transfers between jurisdictions [23] and when regulatory systems are challenged by the rapidly changing landscape of DMHI provision (eg, as AI technologies are integrated) [24].

As we reflect on the past 25 years and the current position, we draw on our own research experience of DMHI design and evaluation and our knowledge of the field to identify ethical and legal concerns that we believe are of critical importance to address. We endeavor to support an informed discourse and guide future research designs to promote the ethically sound and legally resilient integration of DMHIs into mental health care provision.

XSL•FO
RenderX

Harm Caused by Online Recruitment Work

Online evaluations of DMHI often use online recruitment mechanisms. These can enable the rapid recruitment of participants [16], including people not currently using formal health services, which is important for external validity if DMHIs are intended to extend service reach [4]. However, some forms of online recruitment can cause harm to people with mental health conditions, which means that the design of recruitment strategies is an ethical issue.

Some researchers have chosen to "lurk" on various web-based discussion forums, by joining these forums as a member solely to share recruitment material with other members. Their presence can compromise the safety of these spaces [25]. For example, author Rennick-Egglestone, who is a longstanding member of a mental health support group for people experiencing cyclothymia, has seen peers leave the group permanently when researchers join to share recruitment material, due to a fear that group interactions may be treated as data in research studies. The British Sociological Association has published general guidance on researching web-based forums, which provides a broad overview of thought and practice in this space [26]. The British Psychological Society has provided a guide to how its Code of Human Research Ethics can be interpreted in internet-mediated research [27]. Both highlight that the distinction between public and private space is increasingly blurred in web-based interactions, leading to challenging decisions for researchers about appropriate methods.

Harm can also be caused if potential participants misunderstand online recruitment material, and hence feel misled about the purpose of a study they have chosen to participate in. This may be more likely in web-based evaluations (eg, if study information such as participant information sheets is disseminated over the internet) in a context wherein study team members are not routinely available to explain its contents. We have explored the development of mandatory recruitment principles for online evaluations by developing and using exemplar principles for a web-based trial, which were intended to promote safe practices by trial workers [28]. These included explicit instructions not to lurk but instead to approach the gatekeepers of online groups with requests to share recruitment material and direction on minimum content for each published recruitment message [28].

Frequently, participants identified through web-based recruitment mechanisms are not asked for formal identity verification. This is in keeping with guidance from US [29] and UK [30] regulators on procedures for clinical trials that do not evaluate investigational medicinal products, commonly referred to as non-CTIMPs (non–Clinical Trials of Investigational Medicinal Products). There is developing evidence that "fake participation" is a risk to the integrity of studies recruiting online that do not formally confirm identity. For example, Hydock (2017) reported that a "small but nontrivial portion of participants in online survey studies misrepresented their identity for the chance of financial gain" [19]. How to identify and respond to repeat registration is an ethical consideration, as misrepresentation by participants can distort findings [31]. This could amount to a misuse of the effort contributed by real trial participants and the misuse of public funds through subsequent commissioning decisions.

Without formal identity verification, DMHIs may also be accessed by people for whom they are not suitable, which raises both legal and ethical concerns, such as if DMHIs intended for adults are accessed by children. Legally, this is a developing space, with the UK Online Safety Act (2023) and the US Children's Online Personal Protection Act (1998) both offering children under the age of 13 greater protection in the online space (eg, restricting the data that is allowed to be collected without parental consent [32]). The latter was reinforced in the United States via the Children's Online Privacy Protection Rule introduced in 2013 [33]. Studies that inadvertently include data provided by younger children may therefore be challenged on the legality of their processes.

Some DMHIs can contain mental health material with the potential to cause distress [21], which may be unsuitable for people such as children lacking the capacity to process the distress of others. These considerations take place in a context where campaign groups across the globe have called for tighter restrictions in response to the terrible loss of children and young people to suicide attributed to the negative impact of web-based platforms [34].

Monitoring of Intervention Safety

DMHIs can harm people with mental health conditions. The harm can arise through mechanisms such as the physical act of engaging with the intervention (eg, cybersickness in the case of virtual reality), or through interactions between the type of content or information given and the mental health experiences of the technology user. For example, in the United States, the National Eating Disorders Association was forced to remove an AI chatbot servicing a helpline when it began to provide recommendations, such as limiting caloric intake that might reasonably be expected to exacerbate conditions for people experiencing disordered eating [35]. Most generally, harm due to DMHIs might be experienced as one or more *adverse events*, some of which can be serious, including hospitalization due to substantial distress associated with intervention usage [21]. A growing body of evidence suggests that, historically, adverse events caused by DMHIs have not been well recorded either in research evaluations or in post market surveillance of DMHI usage [15]. For example, a recent review of 23 DMHI trials included in the ISRCTN registry revealed that although most (69%) have published one or more statements on adverse event identification, few (26%) had reported on the number of adverse events in their main findings [15], and few had examined whether these events were serious (26%) or related to the intervention (17%). Another review on DMHI showed that only 65% (15/23) of studies (which included trials and postmarket surveillance studies) explored risks to safety. The authors noted that this was an improvement compared with the findings of a review in 2017, whereby none of the 9 included papers assessed safety risks [36].

XSL•FO
RenderX

Despite this increase, there is a clear need to improve the way harms are monitored, both during formal research evaluations and as technologies are implemented into practice. It is critical that harms are understood to inform the benefit/risk ratio and to make informed decisions on whether a product should be implemented in health care settings, as well as on forms of support required for safe use. Recording adverse events in online evaluations brings new challenges, particularly if there are no routinely planned interactions with health care workers where adverse events could be enquired about or reported. The likelihood of adverse events being spontaneously reported is limited in online interventions, even where reporting mechanisms are provided [15]. Suggestions have been raised to facilitate harm and event identification, including ensuring independent review committees include technology experts [15], ensuring adverse events are statistically compared between arms in a trial, and using information on harms collected during an evaluation to inform postmarket surveillance [37]. There is also the potential to integrate features into DMHIs to support the recording of harms, including developing real-time monitoring of symptoms that raise alerts when participants reach predefined cut-offs. This requires evaluation to understand its acceptability, feasibility, and effectiveness [15].

Exclusion of Specific Demographic or Clinical Groups

It is an important ethical principle that access to evidence-based health care is equitable and does not disadvantage certain groups or patient characteristics. For example, in the United Kingdom and the United States, over 80% of adults and young people own a smartphone [38-40]. However, access to a mobile data plan is likely to be less in populations that are typically underserved by research and health care services (with example populations in the United States including Black, Hispanic, older, and less educated populations) [39]; hence, evaluations and implementation of DMHIs risk exacerbating existing health inequalities [41]. Although it is anticipated that digital disparities will reduce over time, the importance of digital literacy should not be overlooked [40]. Some training projects have been developed to address issues around digital literacy, including the Digital Outreach for Obtaining Resources and Skills program, which provides both in-person and web-based support in digital literacy [42]. The use of "digital navigators," who are trained members of the clinic team providing technology support [43], while not placing additional burden on core-delivery clinic members, is a potentially promising way forward and has been implemented successfully in DMHI evaluations [44]. It is also critical that clinicians and health care providers are efficient at integrating technology into practice to increase access to DMHI, but clinicians may worry about the capacity of clinical systems to respond to risks created by DHIs [45].

Equally, it is critical that the DMHIs are designed and evaluated in the cultural context where they will be used [46]. The majority of research on DMHI is delivered in Westernized countries with relatively high incomes; studies have suggested that the effectiveness of these interventions may not directly translate to different cultures [47,48]. For example, issues such as the concept of "depression" not translating across cultures and languages [49] highlight the critical importance of language and context [50]. Although there is a growing interest in culturally adapted DMHIs, to date there has been no research that directly compares culturally adapted versus nonadapted interventions to explore if this improves uptake and effectiveness [48]. Given the potential that DMHIs have in providing scalable health care solutions, this requires further research.

Finally, trials of DMHIs often do not include participants who may be of higher risk, such as those with suicidality or severe mental health problems. A recent review found that 13 out of 23 trials screened out patients who had more serious and enduring mental health illnesses [15]. Another review found that the main method to mitigate risk in trials of DMHIs was to exclude high-risk populations [37]. Whereas establishing proof-of-concept and indications of efficacy and safety on a restricted sample is ethically appropriate, it is important that when larger, pragmatic trials or research designs are conducted before wide-scale rollout, these larger studies should include samples that are reflective of the population of interest, including all relevant characteristics and comorbidities. The severity of mental health problems is frequently underdiagnosed; hence, evaluations should consider that DMHIs may be used by people whose mental health conditions are much more severe than for whom the intervention was intended, such as people diagnosed with mild-to-moderate depression, but who are experiencing a period of severe depression [22].

Some categories of DMHI may embed a risk of discrimination on personal characteristics that are protected by law. One example is AI technologies, where evidence shows that algorithmic biases can lead to discrimination based on characteristics such as gender and ethnicity [51], which are frequently protected in national legislation [52]. Laws provide protection for individuals against discrimination, but the challenge is often understanding when discrimination has happened, as the algorithms used in interventions often can be difficult to interpret and explain [51].

Inadequate Robustness of Effectiveness and Cost-Effectiveness Findings

The adequacy of study designs to fully assess intervention effectiveness and cost-effectiveness is a key challenge to the field. If trial designs are not sufficiently robust to answer the intended question, it is possible that the reported benefits of an intervention may be inflated, resulting in interventions that are not effective or fit-for-purpose being recommended for routine use. Consequently, engaging in these interventions may result in users delaying access or being unable to access more suitable interventions, and potentially represents a waste of health service resources, hence the adequacy of evaluation design is inherently an ethical issue.

One significant challenge lies in selecting the most appropriate comparator for RCTs. Potential comparators include active therapeutic interventions delivered nondigitally (eg, face-to-face therapy) or different active interventions (eg, cognitive-behavioral therapy versus exposure therapy) delivered

XSL•FO
RenderX

digitally. Alternatively, comparators may be nonactive controls such as waitlists or no intervention. It may be impossible for participants to be masked to treatment allocation in a DMHI evaluation, raising the risk of bias. Reviews have suggested that comparisons to nonactive controls may lead to larger treatment effects than comparisons to active controls [53,54].

Trials of DMHI frequently encounter high dropout rates (failure to follow participants up to the primary and secondary end points) or fail to recruit the target number of participants, leading to reduced statistical power and a lower probability of identifying effects. Analysis of data from the National Institute for Health and Care Research (NIHR) Health Technology Assessment database revealed that from 2010 to 2020, 60% of DMHI trials did not meet their recruitment targets, and 56% did not meet their follow-up targets [55]. This shortfall may result in recommendations based on insufficient research to draw definitive conclusions. It raises an ethical concern about wasting the time of participants who have engaged in the research but to no scientific avail.

A significant challenge within DMHIs is the speed at which technology evolves and progresses, versus the slow and careful study designs we use, predicated on pharmaceutical research. By the time RCTs are complete, the technology itself may have required significant changes to be made within the intervention. This raises the question as to whether we should be considering alternative study designs (eg, rapid evaluations of specific therapeutic components or evaluating the application of therapeutic modalities through different mediums rather than specific interventions).

A systematic review has shown that current study designs often neglect to assess the cost-effectiveness of interventions [56]. Understanding cost-effectiveness is crucial for evaluating an intervention's benefits to the health care system and enabling informed decisions by health care providers regarding investment in digital technology over other treatment modalities. Moreover, if trials do not explore cost-effectiveness, they may fail to record other health resource utilization outside the intervention, which is essential for understanding clinical benefits between treatment groups.

Adequately Conceptualizing and Supporting Engagement and Adherence

Most DMHIs require some form of active engagement from their users. Some interventions, such as Beating the Blues [57], are built on an interaction approach in which users are expected to work through a structured program of activities intended to offer therapeutic benefits, often intended to last for weeks or months. Developing DMHIs with the potential to sustain engagement is a significant challenge; research studies and real-world implementation experiences often report low levels of engagement [55]. Where engagement is low or irregular, and engagement with a structured program of activities is intended, this is sometimes conceptualized by study teams as poor adherence to the intended technology interaction. Poor adherence to structured programs is frequently reported in research evaluations [55].

Challenges around engagement and adherence are ethical issues for research teams. For example, an automated DMHI may present a user experience that is off-putting due to poor design. This might cause harm if a person with a mental health condition perceives it as a personal failing that they could not sustain the expected engagement, rather than a technology design failure. Whether to include human support in study designs is an important question. Studies have consistently demonstrated that guided interventions, which involve human support through asynchronous messages or real-time interaction, have demonstrated higher adherence and greater clinical effectiveness compared to unguided interventions [58], but there are settings in which they may not be practical (eg, due to resource limitations) or where they may cause harm, such as where the expected user base is experiencing profound social anxieties [22].

Engagement with a DMHI is often difficult for researchers to quantify. Often, a range of DMHI interaction behaviors may need to be considered, including frequency or duration of interaction with the DMHI, and metrics of “active” engagement,” such as the number of clicks on links, pages visited, or interactive tasks completed [59]. It can be difficult for researchers to define which is the best metric to use. One approach is to conduct a principal component analysis to create a comprehensive composite measure [59]. Some researchers are beginning to explore more subjective constructs of engagement, including aspects such as knowing that the intervention is available whenever it is needed, and satisfaction and usability [60].

Despite its importance, research evaluations of internet-delivered interventions often neglect to investigate the impact of usage level on intervention outcome. A systematic review found that only 21% of included trials had conducted statistical analyses to account for differences in usage [61], leading to a lack of knowledge on how much engagement with a DMHI is sufficient. Even if an understanding of the necessary level of engagement to provide benefits is established in a tightly controlled research setting, achieving this in a real-world setting can be challenging, as research settings often offer greater levels of participant support than in real-world implementation. For example, when the serious game “SPARX” aimed at supporting low mood in children and young people was nationally rolled out in New Zealand, only between 2% and 5% of young people completed the full program [62], whereas considerably higher completion rates had been found in the trial [63].

There is also a danger of research teams misconceptualizing lack of usage as poor adherence, in cases where users have made a sensible choice to avoid usage of a technology that provides no benefits or is actively harming them. This can occur if teams take an ego-centric approach to conceptualizing their offerings as of central importance to their users [64]. For a recent trial examining the benefits of receiving access to a collection of mental health recovery narratives [21], we chose not to require a minimum level of usage, as our empirically developed change model had established that access to narratives could cause harm, so we wanted to enable users to self-manage engagement, which could include no engagement [65,66]. Healthy engagement was supported by a range of safety strategies,

XSL•FO
RenderX

including the use of content warnings [66]. Hence, in this example, we did not conceptualize low engagement as a lack of adherence [21].

Structural Barriers to Implementation

DMHIs often struggle to bridge the gap from a research product to implementation in health care services or broader society [67]. This is an ethical issue, as research evaluations regularly require hundreds of participants, large teams of evaluators and support workers, and access to substantial amounts of public funds, all of which are underused if this transition is not made. If DMHIs are both clinically effective and cost-effective but not implemented, people with mental health problems may not be receiving optimal care. Currently, there are several DMHIs that are available for use in health service settings that have not undergone independent evaluation [9]. Conversely, there are several well-validated DMHIs that may never be implemented in routine health care settings [68]. This illustrates a possible disconnect between research-led efforts to evaluate interventions, and the needs of clinical practice, which is a major barrier to service user benefit.

There is likely to be a broad range of factors underpinning this disconnect, and addressing these is critical to reducing the underuse of research evaluations of DMHIs. The fast-paced nature of technological progress can mean DMHIs can become obsolete or unacceptable to users before they reach the implementation stage [53] due to the slow progress of the research lifecycle. Research teams may lack the institutional support or commercial knowledge needed to implement DMHIs that they have evaluated, and work to implement technologies may be at odds with funder or institutional metrics used to evaluate individual academic performance, particularly where the regulatory burden is high. A possible solution is government-funded infrastructure to support the integration of academic, clinical, and industry expertise, but substantial resources may be required.

A clinically grounded challenge is that health care professionals may not have received preregistration training or continuous professional development in DMHIs and how to work with them clinically, and hence may lack the competence or confidence to integrate them into care. For example, an interview study with Norwegian general practitioners interested in computerized cognitive behavioral therapy revealed their concerns that they had insufficient knowledge to effectively recommend these to their service users [69]. We have argued that health care practitioners need training to enable their own psychological safety when providing therapies online through videoconferencing services, including due to the lesser availability of immediate emergency response in the event of suicidality [70].

Data Protection and Intellectual Property

The deployment and evaluation of DMHIs typically involve processing personal and sensitive data, sometimes with processing through third parties such as web-server hosts. In many jurisdictions, these data will be protected by legislation. For example, in the European Union (EU), Article 8 of the European Convention on Human Rights and Articles 7 and 8 of the EU Charter of Fundamental Rights enshrine privacy as a fundamental human right, and national implementations of the General Data Protection Regulations provide a specific legal framework for data processing.

Web-based DMHIs, in particular, have the potential to be evaluated in one jurisdiction but incorporate components hosted in another, frequently requiring close attention to the navigation of legal requirements, especially when the countries involved have different standards of data protection [71]. Additional challenges are present when computer vision technologies such as augmented reality are used, as data about third parties may be collected without their consent. Service users and clinicians have raised concerns about data security in DMHIs and the potential for unauthorized breaches of sensitive data [71-73]. For health care services, the introduction of these technologies may present too much uncertainty in terms of their legal requirements. This can contribute to challenges with the implementation of DMHIs. If users are encouraged to create and share content as part of working with a DMHI, this introduces additional considerations around how such content is licensed and used [74].

As evaluators, we have all invested substantial effort in data management planning around DMHIs, including identifying how to ethically and legally work with the data that is generated through our evaluations. This requires access to specific competencies such as knowledge of data protection laws. These competencies require work to maintain, as regulations around data processing can change rapidly (eg, when the EU Safe Harbor scheme for cross-border data processing was replaced by Privacy Shield [23]). Legal and ethical issues are present even for data that is publicly available. In 2015, the Samaritans released an app that identified tweets associated with suicidal ideation and alerted friends of those tweeting. This was rapidly withdrawn following interventions by privacy campaigners [75]. The UK Information Commissioner's Office subsequently found that it was unlikely to be compatible with UK data protection law [76]. More recently, the US Federal Trade Commission fined BetterHelp, a platform connecting people to therapists, $7.8 million for sharing IP and email addresses as well as sensitive mental health information to third parties [77].

There are similar challenges around intellectual property, as DMHIs may emerge from collaborative efforts involving multiple partners, and they may be built on underlying technologies such as software libraries produced by third parties, sometimes with restrictive licensing conditions that can preclude commercial use. Successfully addressing these issues requires a proactive stance from the outset of development. We advocate seeking guidance from legal or technology transfer teams early in the conceptual phase to mitigate risks. Moreover, there is a pressing need to enhance researchers' skills in translational research and foster collaborative endeavors between stakeholders to effectively support the future implementation of DMHI.

XSL•FO
RenderX

Regulatory Ambiguity Relating to DMHIs That Are Medical Devices

Several organizations across the globe regulate hardware and software artifacts classified as medical devices. These include the European Medicines Authority, the Medicines and Health Care Products Regulatory Agency (MHRA) in the United Kingdom, the Food and Drug Administration (FDA) in the United States, and the Therapeutic Goods Administration (TGA) in Australia. These organizations typically have legislative powers to ensure that medical devices are safe and effective, which can include requiring manufacturers to submit information on their approach to the safe storage and processing of data. How this is enacted and how medical devices are defined may differ, and some countries remain without regulatory authorities. Where medical devices are regulated, classification is usually based on the "intended use/purpose," as defined by the manufacturer or developer. For example, the EU's classifications for medical devices are outlined in Textbox 1 [78].

Textbox 1. European Union (EU) classifications for medical devices.

"Medical device" means any instrument, apparatus, appliance, software, implant, reagent, material, or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:

- diagnosis, prevention, monitoring, prediction, prognosis, treatment, or alleviation of disease;
- diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,
- investigation, replacement, or modification of the anatomy or of a physiological or pathological process or state;
- providing information by means of in vitro examination of specimens derived from the human body, including organ, blood, and tissue donations;
- and which does not achieve its principal intended action by pharmacological, immunological, or metabolic means, in or on the human body, but which may be assisted in its function by such means.

This definition explicitly recognizes the concept of Software as a Medical Device (SaMD), which is of particular relevance to DMHI developers and evaluators. The use of terms such as "disease" and "disability" have often left medical device classification ambiguous for SaMD intended for digital mental health purposes. In some jurisdictions, amendments have been made to reduce ambiguity. For example, the FDA has separated out a category of "low-risk devices" intended to target wellness or well-being, by offering specific internal staff guidance on how they should be regulated [79]. This perhaps creates a risk of DMHI developers classifying their tools as "wellness" devices rather than "treatments" to reduce regulatory burden [40]. In the United Kingdom, the MHRA is engaged in a change program on SaMD regulation but with a component examining arrangements needed so that AI technologies are effectively regulated [24].

Resolving ambiguity around whether our interventions are medical devices has sometimes taken substantial effort for us as evaluators (eg, through seeking opinions from sponsors and legal experts), and we are aware of at least one major research program that faced a delay of over a year when its intervention was unexpectedly classified as a medical device. We also wonder whether DMHIs are underclassified as medical devices; our review of DMHI trials found that only 2 of 23 interventions had a medical device classification [15]. Given the ambiguity of regulation, there is a potential for the medical status of DMHIs to change between the awarding of funding and the start of evaluation. We have found that regulatory guidance can be more ambiguous where DMHIs are not classified as medical devices, perhaps due to a regulatory perception that harm is less likely. For these DMHIs, authors ADGB and CLH are working to produce adverse event reporting guidelines that complement those of the MHRA.

Regulations that define when a medical device needs to be clinically investigated can also create ambiguity for DMHIs. For example, in MHRA regulation, clinical investigation is not needed if there is sufficient evidence available for an intervention that is equivalent [80]. Equivalency might be straightforward to assess for medication but less clear for DMHIs (eg, does equivalency hold if the same therapeutic intervention is delivered but through a different device?). Within MHRA regulation, further clinical investigation is needed when there has been a significant change to an intervention. However, current regulations may be unclear for AI-based DMHI interventions that learn and hence evolve; therefore, regulation is currently being examined in this space [24]. Finally, DMHIs are sometimes used outside their intended purpose (eg, if a self-management intervention is evaluated to treat a condition such as depression). These "secondary uses" for digital technologies appear to be relatively common and may produce harms that are not recorded (eg, if participants drop out of an evaluation due to being given an inappropriate technology and are not followed up).

Conclusion

Over the past 25 years, the rapid evolution of both DMHI-based research and encompassing regulatory systems has meant that study teams are faced with multifaceted ethical and legal challenges to resolve. We have drawn on our own experiences to describe 8 of the most pertinent challenges, but there will be others, and some concerns will specifically apply to particular types or forms of DMHI. Collectively, the challenges we have identified highlight a substantial body of knowledge and expertise that is required on the part of study teams, either within the team or through access to external experts. The challenges that we present also illustrate potentially serious negative consequences if sufficient resources are not available to study teams, including challenges regarding the legality of data collection procedures that inadvertently collect data from children, harm caused to members of the public through

XSL•FO
RenderX

unethical recruitment methods, and substantial delays to studies that discover medical devices legislation applies to their intervention. Our focus was predominately on treatment; however, a similar reflection of the challenges of DMHI for prevention, assessment, or diagnosis is also warranted.

In many cases, understanding how to respond to these challenges requires collaboration between industry representatives, academics, public contributors, and clinicians. Each of these stakeholder groups can bring unique expertise and perspectives, and by leveraging their collective insights, innovative solutions can be developed to promote the legal and ethical use of DMHIs. To support success, such collaborations need to be planned early, with relevant collaborators and financial resources integrated into funding proposals, or other mechanisms for collecting study resources. For example, resources to pay for legal opinions are frequently not included in proposals but may be necessary to support successful negotiation of regulatory processes.

Organizations conducting health research might consider how to create structures that develop and maintain the necessary competence to resolve the challenges that we have identified. For example, organizations that are frequently engaged in DMHI evaluation might consider developing specific competence on medical device regulations, including the capacity to conduct horizon-scanning activities to identify possible changes (given that regulations can change between proposal submission and evaluation conclusion). Organizations might also consider developing communities of public members interested in public involvement in research. Collectively, we have found discussions with public members essential to resolve the most challenging ethical issues, such as around the management of narrator identity in our trial of a DMHI integrating mental health recovery narratives [66].

Acknowledgments

Authors CLH and SRE were supported by the UK National Institute for Health and Care Research (NIHR) through the NIHR Nottingham Biomedical Research Centre (NIHR203310). CLH was also supported through the NIHR MindTech Health Research Centre. AGB was supported by the Engineering and Physical Sciences Research Council through the UK Responsible Artificial Intelligence (RAI; grant EP/Y009800/1). The views expressed here are those of the authors and not necessarily those of the NIHR or the UK Department of Health and Social Care.

Conflicts of Interest

None declared.

References

1. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866 [FREE Full text] [Medline: 15640922]
2. Demyttenaere K, Bruffaerts R, Posada-Villa J, Gasquet I, Kovess V, Lepine JP, et al. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 2004 Jun 2;291(21):2581-2590. [doi: 10.1001/jama.291.21.2581] [Medline: 15173149]
3. Goodwin RD, Dierker LC, Wu M, Galea S, Hoven CW, Weinberger AH. Trends in US depression prevalence From 2015 to 2020: the widening treatment gap. Am J Prev Med 2022 Nov;63(5):726-733 [FREE Full text] [doi: 10.1016/j.amepre.2022.05.014] [Medline: 36272761]
4. Rennick-Egglestone S, Newby C, Robinson C, Yeo C, Ng F, Elliott RA, et al. Differences between online trial participants who have used statutory mental health services and those who have not: analysis of baseline data from 2 pragmatic trials of a digital health intervention. J Med Internet Res 2023 Jun 27;25:e44687 [FREE Full text] [doi: 10.2196/44687] [Medline: 37368471]
5. Patel V, Maj M, Flisher AJ, De SMJ, Koschorke M, Prince M, WPA ZonalMember Society Representatives. Reducing the treatment gap for mental disorders: a WPA survey. World Psychiatry 2010 Oct;9(3):169-176 [FREE Full text] [Medline: 20975864]
6. Lockwood J, Williams L, Martin JL, Rathee M, Hill C. Effectiveness, user engagement and experience, and safety of a mobile app (Lumi Nova) delivering exposure-based cognitive behavioral therapy strategies to manage anxiety in children via immersive gaming technology: preliminary evaluation study. JMIR Ment Health 2022 Jan 24;9(1):e29008 [FREE Full text] [doi: 10.2196/29008] [Medline: 35072644]
7. Freeman D, Lambe S, Kabir T, Petit A, Rosebrock L, Yu L, gameChange Trial Group. Automated virtual reality therapy to treat agoraphobic avoidance and distress in patients with psychosis (gameChange): a multicentre, parallel-group, single-blind, randomised, controlled trial in England with mediation and moderation analyses. Lancet Psychiatry 2022 May;9(5):375-388 [FREE Full text] [doi: 10.1016/S2215-0366(22)00060-8] [Medline: 35395204]
8. The mobile economy: Sub-Saharan Africa. GSM Association. London, UK: GSM Association URL: https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/09/GSMA_MobileEconomy2020_SSA_Eng.pdf [accessed 2024-07-27]
9. Schueller SM, Torous J. Scaling evidence-based treatments through digital mental health. Am Psychol 2020 Nov;75(8):1093-1104 [FREE Full text] [doi: 10.1037/amp0000654] [Medline: 33252947]

XSL•FO
RenderX

10. Rodriguez-Villa E, Naslund J, Keshavan M, Patel V, Torous J. Making mental health more accessible in light of COVID-19: Scalable digital health with digital navigators in low and middle-income countries. Asian J Psychiatr 2020 Dec;54:102433. [doi: 10.1016/j.ajp.2020.102433] [Medline: 33271713]
11. Tornero-Costa R, Martinez-Millana A, Azzopardi-Muscat N, Lazeri L, Traver V, Novillo-Ortiz D. Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review. JMIR Ment Health 2023 Feb 02;10:e42045 [FREE Full text] [doi: 10.2196/42045] [Medline: 36729567]
12. Sanderson C, Kouzoupi N, Hall CL. Technology Matters: The human touch in a digital age - a blended approach in mental healthcare delivery with children and young people. Child Adolesc Ment Health 2020 May 27;25(2):120-122. [doi: 10.1111/camh.12385] [Medline: 32307846]
13. Neary M, Schueller SM. State of the field of mental health apps. Cogn Behav Pract 2018 Nov;25(4):531-537 [FREE Full text] [doi: 10.1016/j.cbpra.2018.01.002] [Medline: 33100810]
14. Nicholas J, Larsen ME, Proudfoot J, Christensen H. Mobile apps for bipolar disorder: a systematic review of features and content quality. J Med Internet Res 2015;17(8):e198 [FREE Full text] [doi: 10.2196/jmir.4581] [Medline: 26283290]
15. Gómez Bergin AD, Valentine AZ, Rennick-Egglestone S, Slade M, Hollis C, Hall CL. Identifying and categorizing adverse events in trials of digital mental health interventions: narrative scoping review of trials in the International Standard Randomized Controlled Trial Number Registry. JMIR Ment Health 2023 Feb 22;10:e42501 [FREE Full text] [doi: 10.2196/42501] [Medline: 36811940]
16. Murray E, Khadjesari Z, White IR, Kalaitzaki E, Godfrey C, McCambridge J, et al. Methodological challenges in online trials. J Med Internet Res 2009;11(2):e9 [FREE Full text] [doi: 10.2196/jmir.1052] [Medline: 19403465]
17. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
18. Hollis C, Hall CL, Jones R, Marston L, Novere ML, Hunter R, et al. Therapist-supported online remote behavioural intervention for tics in children and adolescents in England (ORBIT): a multicentre, parallel group, single-blind, randomised controlled trial. Lancet Psychiatry 2021 Oct;8(10):871-882 [FREE Full text] [doi: 10.1016/S2215-0366(21)00235-2] [Medline: 34480868]
19. Hall CL, Davies EB, Andrén P, Murphy T, Bennett S, Brown BJ, ORBIT Trial team. Investigating a therapist-guided, parent-assisted remote digital behavioural intervention for tics in children and adolescents-'Online Remote Behavioural Intervention for Tics' (ORBIT) trial: protocol of an internal pilot study and single-blind randomised controlled trial. BMJ Open 2019 Jan 03;9(1):e027583 [FREE Full text] [doi: 10.1136/bmjopen-2018-027583] [Medline: 30610027]
20. Hall CL, Walker GM, Valentine AZ, Guo B, Kaylor-Hughes C, James M, et al. Protocol investigating the clinical utility of an objective measure of activity and attention (QbTest) on diagnostic and treatment decision-making in children and young people with ADHD-'Assessing QbTest Utility in ADHD' (AQUA): a randomised controlled trial. BMJ Open 2014 Dec 01;4(12):e006838 [FREE Full text] [doi: 10.1136/bmjopen-2014-006838] [Medline: 25448628]
21. Slade M, Rennick-Egglestone S, Elliott RA, Newby C, Robinson C, Gavan SP, NEON study group. Effectiveness and cost-effectiveness of online recorded recovery narratives in improving quality of life for people with non-psychotic mental health problems: a pragmatic randomized controlled trial. World Psychiatry 2024 Feb;23(1):101-112 [FREE Full text] [doi: 10.1002/wps.21176] [Medline: 38214639]
22. Rennick-Egglestone S, Knowles S, Toms G, Bee P, Lovell K, Bower P. Health technologies' in the wild': experiences of engagement with computerised CBT. : ACM; 2016 Presented at: 2016 CHI Conference on Human Factors in Computing Systems; May 7-12; San Jose. [doi: 10.1145/2858036.2858128]
23. Weiss M, Archick K. US-EU data privacy: from safe harbor to privacy shield. Congressional Research Service. Washington, USA: Congressional Research Service; 2016 May. URL: https://sgp.fas.org/crs/misc/R44257.pdf [accessed 2024-09-03]
24. Software and AI as a Medical Device Change Programme - Roadmap. Gov.UK. 2023. URL: https://tinyurl.com/yckwd47w [accessed 2024-09-02]
25. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001 Nov 10;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
26. Sugiura L. Researching online forums: ethics case study. British Sociological Association. Durham, UK: British Sociological Association; 2016. URL: https://www.britsoc.co.uk/media/24834/j000208_researching_online_forums_-cs1-_v3.pdf [accessed 2024-09-02]
27. Hewson C, Buchanan T. Ethics guidelines for internet-mediated research. British Psychological Society. London, UK: The British Psychological Society; 2013. URL: https://www.bps.org.uk/guideline/ethics-guidelines-internet-mediated-research [accessed 2024-09-03]
28. Rennick-Egglestone S. Principles for the production and dissemination of recruitment material for three clinical trials of an online intervention. Trials 2021 Jul 09;22(1):441 [FREE Full text] [doi: 10.1186/s13063-021-05412-4] [Medline: 34243794]
29. Office of the Commissioner, Office of Clinical Policy and Programs, Office of Clinical Policy, Office of Good Clinical Practice, Center for Drug Evaluation and Research, Center for Devices and Radiological Health. Use of Electronic Informed

XSL•FO
RenderX

Consent in Clinical Investigations: Questions and Answers. Food and Drug Administration. Washington, D.C: Food and Drug Administration; 2016. URL: https://www.fda.gov/media/116850/download [accessed 2024-09-04]
30. Health Research Authority, Medicines and Healthcare Products Regulatory Agency. Joint statement on seeking consent by electronic methods v1.2. Health Research Authority. 2018 Sep. URL: https://www.hra.nhs.uk/about-us/news-updates/hra-and-mhra-publish-joint-statement-seeking-and-documenting-consent-using-electronic-methods-econsent/ [accessed 2024-09-02]
31. Robinson C, Newby C, Rennick-Egglestone S, Llewellyn-Beardsley J, Ng F, Elliott RA, et al. Statistical analysis plans for two randomised controlled trials of the Narrative Experiences Online (NEON) Intervention: impact of receiving recorded mental health recovery narratives on quality of life in people experiencing psychosis (NEON) and people experiencing non-psychosis mental health problems (NEON-O). Trials 2023 May 20;24(1):343 [FREE Full text] [doi: 10.1186/s13063-023-07246-8] [Medline: 37210551]
32. Song H, Holzer A. Online safety laws by country. J Responsible Innov 2023;3(3):3 [FREE Full text]
33. Children's Online Privacy Protection Rule ("COPPA"). Federal Trade Commission.: Federal Trade Commission; 1998. URL: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa [accessed 2024-09-03]
34. Bell J. Harmful or helpful? The role of the internet in self-harming and suicidal behaviour in young people. Ment Health Rev J 2014 Mar 05;19(1):61-71. [doi: 10.1108/mhrj-05-2013-0019]
35. NEDA suspends AI chatbot for giving harmful eating disorder advice. Psychiatrist. URL: https://www.psychiatrist.com/news/neda-suspends-ai-chatbot-for-giving-harmful-eating-disorder-advice/ [accessed 2024-09-02]
36. Bell IH, Lim MH, Rossell SL, Thomas N. Ecological momentary assessment and intervention in the treatment of psychotic disorders: a systematic review. Psychiatr Serv 2017 Nov 01;68(11):1172-1181. [doi: 10.1176/appi.ps.201600523] [Medline: 28669284]
37. Taher R, Hsu C, Hampshire C, Fialho C, Heaysman C, Stahl D, et al. The safety of digital mental health interventions: systematic review and recommendations. JMIR Ment Health 2023 Oct 09;10:e47433 [FREE Full text] [doi: 10.2196/47433] [Medline: 37812471]
38. Bergin A, Davies EB. Technology Matters: mental health apps - separating the wheat from the chaff. Child Adolesc Ment Health 2020 Feb;25(1):51-53 [FREE Full text] [doi: 10.1111/camh.12363] [Medline: 32063750]
39. Roberts ET, Mehrotra A. Assessment of disparities in digital access among medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020 Oct 01;180(10):1386-1389. [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]
40. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
41. Helsper EJ. A corresponding fields model for the links between social and digital exclusion. Commun Theor 2012 Oct 15;22(4):403-426. [doi: 10.1111/j.1468-2885.2012.01416.x]
42. Hoffman L, Wisniewski H, Hays R, Henson P, Vaidyam A, Hendel V, et al. Digital Opportunities for Outcomes in Recovery Services (DOORS): a pragmatic hands-on group approach toward increasing digital health and smartphone competencies, autonomy, relatedness, and alliance for those with serious mental illness. J Psychiatr Pract 2020 Mar;26(2):80-88. [doi: 10.1097/PRA.0000000000000450] [Medline: 32134881]
43. Connolly SL, Kuhn E, Possemato K, Torous J. Digital clinics and mobile technology implementation for mental health care. Curr Psychiatry Rep 2021 May 07;23(7):38 [FREE Full text] [doi: 10.1007/s11920-021-01254-8] [Medline: 33961135]
44. Carpenter-Song E, Acquilano SC, Noel V, Al-Abdulmunem M, Torous J, Drake RE. Individualized intervention to support mental health recovery through implementation of digital tools into clinical care: feasibility study. Community Ment Health J 2022 Jan;58(1):99-110 [FREE Full text] [doi: 10.1007/s10597-021-00798-6] [Medline: 33611684]
45. Roe J, Brown S, Yeo C, Rennick-Egglestone S, Repper J, Ng F, et al. Opportunities, enablers, and barriers to the use of recorded recovery narratives in clinical settings. Front Psychiatry 2020 Oct 30;11:589731 [FREE Full text] [doi: 10.3389/fpsyt.2020.589731] [Medline: 33192738]
46. Groen G, Jörns-Presentati A, Dessauvagie A, Seedat S, van den Heuvel LL, Suliman S, et al. Development of a mobile application for detection of adolescent mental health problems and feasibility assessment with primary health care workers. Issues Ment Health Nurs 2022 Nov;43(11):1046-1055. [doi: 10.1080/01612840.2022.2124003] [Medline: 36205922]
47. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, et al. Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. Clin Psychol Rev 2018 Jul;63:80-92. [doi: 10.1016/j.cpr.2018.06.007] [Medline: 29940401]
48. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
49. Kokanovic R, Dowrick C, Butler E, Herrman H, Gunn J. Lay accounts of depression amongst Anglo-Australian residents and East African refugees. Soc Sci Med 2008 Jan;66(2):454-466. [doi: 10.1016/j.socscimed.2007.08.019] [Medline: 17923179]

XSL•FO
RenderX

50. Loewenthal D, Mohamed A, Mukhopadhyay S, Ganesh K, Thomas R. Reducing the barriers to accessing psychological therapies for Bengali, Urdu, Tamil and Somali communities in the UK: some implications for training, policy and practice. Br J Guid Counc 2012 Feb;40(1):43-66. [doi: 10.1080/03069885.2011.621519]
51. Bartoletti I, Xenidis R. Study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination. Council of Europe. Brussels, Belgium: Council of Europe; 2023. URL: https://tinyurl.com/yaethk4n [accessed 2024-09-03]
52. Kotera Y, Rennick-Egglestone S, Ng F, Llewellyn-Beardsley J, Ali Y, Newby C, et al. Assessing diversity and inclusivity is the next frontier in mental health recovery narrative research and practice. JMIR Ment Health 2023 Apr 17;10:e44601 [FREE Full text] [doi: 10.2196/44601] [Medline: 37067882]
53. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual Research Review: Digital health interventions for children and young people with mental health problems: a systematic and meta-review. J Child Psychol Psychiatry 2016 Dec 10. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
54. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
55. Hall CL, Sanderson C, Brown BJ, Andrén P, Bennett S, Chamberlain LR, et al. Opportunities and challenges of delivering digital clinical trials: lessons learned from a randomised controlled trial of an online behavioural intervention for children and young people. Trials 2020 Dec 09;21(1):1011 [FREE Full text] [doi: 10.1186/s13063-020-04902-1] [Medline: 33298127]
56. Valentine AZ, Brown BJ, Groom MJ, Young E, Hollis C, Hall CL. A systematic review evaluating the implementation of technologies to assess, monitor and treat neurodevelopmental disorders: A map of the current evidence. Clin Psychol Rev 2020 Aug;80:101870 [FREE Full text] [doi: 10.1016/j.cpr.2020.101870] [Medline: 32712216]
57. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, REEACT Team. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627 [FREE Full text] [doi: 10.1136/bmj.h5627] [Medline: 26559241]
58. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
59. Khan K, Hollis C, Hall CL, Murray E, Davies EB, Andrén P, et al. Fidelity of delivery and contextual factors influencing children's level of engagement: process evaluation of the online remote behavioral intervention for tics trial. J Med Internet Res 2021 Jun 21;23(6):e25470. [doi: 10.2196/25470]
60. Graham AK, Kwasny MJ, Lattie EG, Greene CJ, Gupta NV, Reddy M, et al. Targeting subjective engagement in experimental therapeutics for digital mental health interventions. Internet Interv 2021 Sep;25:100403 [FREE Full text] [doi: 10.1016/j.invent.2021.100403] [Medline: 34401363]
61. Koneska E, Appelbe D, Williamson PR, Dodd S. Usage metrics of web-based interventions evaluated in randomized controlled trials: systematic review. J Med Internet Res 2020 Apr 16;22(4):e15474 [FREE Full text] [doi: 10.2196/15474] [Medline: 32297870]
62. Malatest International. Evaluation of SPARX. New Zealand Ministry of Health. URL: https://www.health.govt.nz/publications/evaluation-of-sparx [accessed 2024-09-02]
63. Merry SN, Stasiak K, Shepherd M, Frampton C, Fleming T, Lucassen MFG. The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: randomised controlled non-inferiority trial. BMJ 2012 Apr 18;344:e2598 [FREE Full text] [doi: 10.1136/bmj.e2598] [Medline: 22517917]
64. Rennick-Egglestone S, Mawson S. Homes of stroke survivors are a challenging environment for rehabilitation technologies. JMIR Rehabil Assist Technol 2021 Jun 17;8(2):e12029 [FREE Full text] [doi: 10.2196/12029] [Medline: 34137728]
65. Rennick-Egglestone S, Ramsay A, McGranahan R, Llewellyn-Beardsley J, Hui A, Pollock K, et al. The impact of mental health recovery narratives on recipients experiencing mental health problems: Qualitative analysis and change model. PLoS One 2019;14(12):e0226201 [FREE Full text] [doi: 10.1371/journal.pone.0226201] [Medline: 31834902]
66. Slade M, Rennick-Egglestone S, Llewellyn-Beardsley J, Yeo C, Roe J, Bailey S, et al. Recorded mental health recovery narratives as a resource for people affected by mental health problems: development of the Narrative Experiences Online (NEON) intervention. JMIR Form Res 2021 May 27;5(5):e24417 [FREE Full text] [doi: 10.2196/24417] [Medline: 34042595]
67. Liu M, Schueller SM. Moving evidence-based mental health interventions into practice: implementation of digital mental health interventions. Curr Treat Options Psych 2023 Oct 03;10(4):333-345. [doi: 10.1007/s40501-023-00298-2]
68. Bergin AD, Vallejos EP, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. NPJ Digit Med 2020;3:133 [FREE Full text] [doi: 10.1038/s41746-020-00339-7] [Medline: 33083568]
69. Wilhelmsen M, Høifødt RS, Kolstrup N, Waterloo K, Eisemann M, Chenhall R, et al. Norwegian general practitioners' perspectives on implementation of a guided web-based cognitive behavioral therapy for depression: a qualitative study. J Med Internet Res 2014;16(9):e208 [FREE Full text] [doi: 10.2196/jmir.3556] [Medline: 25208886]

XSL•FO
RenderX

70. Rennick-Egglestone S. Health workers need tech training for themselves and their patients. Times Higher Education. 2022. URL: https://www.timeshighereducation.com/campus/health-workers-need-tech-training-themselves-and-their-patients [accessed 2024-09-02]
71. Gooding P. Mapping the rise of digital mental health technologies: Emerging issues for law and society. Int J Law Psychiatry 2019;67:101498. [doi: 10.1016/j.ijlp.2019.101498] [Medline: 31785726]
72. Lustgarten SD, Garrison YL, Sinnard MT, Flynn AW. Digital privacy in mental healthcare: current issues and recommendations for technology use. Curr Opin Psychol 2020 Dec;36:25-31 [FREE Full text] [doi: 10.1016/j.copsyc.2020.03.012] [Medline: 32361651]
73. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
74. Perez Vallejos E, Koene A, Carter CJ, Hunt D, Woodard C, Urquhart L, et al. Accessing online data for youth mental health research: meeting the ethical challenges. Philos. Technol 2017 Oct 12;32(1):87-110. [doi: 10.1007/s13347-017-0286-y]
75. Hsin H, Torous J, Roberts L. An adjuvant role for mobile health in psychiatry. JAMA Psychiatry 2016 Feb;73(2):103-104. [doi: 10.1001/jamapsychiatry.2015.2839] [Medline: 26747695]
76. Information Commissioner's Annual Report and Financial Statements 2014/15. Information Commissioner's Office. London, UK: Information Commissioner's Office; 2015. URL: https://ico.org.uk/media/about-the-ico/documents/1431982/annual-report-2014-15.pdf [accessed 2024-09-02]
77. Fair L. FTC says online counseling service BetterHelp pushed people into handing over health information and broke its privacy promises. Federal Trade Commission. Washington, DC, USA: Federal Trade Commission; 2023 Mar 3. URL: https://tinyurl.com/kfxtkmzp [accessed 2024-09-02]
78. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009, and Repealing Council Directives 90/385/EEC and 93/42/EEC. EUR-Lex.: European Union URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32017R0745 [accessed 2024-09-02]
79. General wellness: policy for low risk devices. guidance for industry and Food and Drug Administration staff. US Food and Drug Administration. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2024-09-02]
80. Keutzer L, Simonsson US. Medical device apps: an introduction to regulatory affairs for developers. JMIR Mhealth Uhealth 2020 Jun 26;8(6):e17567 [FREE Full text] [doi: 10.2196/17567] [Medline: 32589154]

Abbreviations

ADHD: attention deficit hyperactivity disorder
AI: artificial intelligence
CTIMP: Clinical Trial of an Investigational Medicinal Product
DMHI: digital mental health intervention
EU: European Union
FDA: Food and Drug Administration
ISRCTN: International Standard Randomized Controlled Trial Number
MHRA: Medicines and Healthcare Products Regulatory Agency
NIHR: National Institute for Health and Care Research
RCT: randomized controlled trial
SaMD: Software as a Medical Device
TGA: Therapeutic Goods Administration

Edited by G Eysenbach; submitted 28.03.24; peer-reviewed by J Adler, A Smith; comments to author 11.05.24; revised version received 15.05.24; accepted 16.07.24; published 09.09.24.

Please cite as:
Hall CL, Gómez Bergin AD, Rennick-Egglestone S
Research Into Digital Health Intervention for Mental Health: 25-Year Retrospective on the Ethical and Legal Challenges
J Med Internet Res 2024;26:e58939
URL: https://www.jmir.org/2024/1/e58939
doi:10.2196/58939
PMID:

XSL•FO
RenderX

©Charlotte L Hall, Aislinn D Gómez Bergin, Stefan Rennick-Egglestone. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

A 25-Year Retrospective of the Use of AI for Diagnosing Acute Stroke: Systematic Review

Zhaoxin Wang[1]; Wenwen Yang[1], PhD; Zhengyu Li[1]; Ze Rong[1]; Xing Wang[1]; Jincong Han[1]; Lei Ma[1], PhD

Nantong University, Nantong, China

Corresponding Author:
Lei Ma, PhD
Nantong University
9# Seyuan Road
Chongchuan District
Nantong, 226019
China
Phone: 86 18860970645
Email: mlmyhero@163.com

Abstract

Background: Stroke is a leading cause of death and disability worldwide. Rapid and accurate diagnosis is crucial for minimizing brain damage and optimizing treatment plans.

Objective: This review aims to summarize the methods of artificial intelligence (AI)–assisted stroke diagnosis over the past 25 years, providing an overview of performance metrics and algorithm development trends. It also delves into existing issues and future prospects, intending to offer a comprehensive reference for clinical practice.

Methods: A total of 50 representative articles published between 1999 and 2024 on using AI technology for stroke prevention and diagnosis were systematically selected and analyzed in detail.

Results: AI-assisted stroke diagnosis has made significant advances in stroke lesion segmentation and classification, stroke risk prediction, and stroke prognosis. Before 2012, research mainly focused on segmentation using traditional thresholding and heuristic techniques. From 2012 to 2016, the focus shifted to machine learning (ML)–based approaches. After 2016, the emphasis moved to deep learning (DL), which brought significant improvements in accuracy. In stroke lesion segmentation and classification as well as stroke risk prediction, DL has shown superiority over ML. In stroke prognosis, both DL and ML have shown good performance.

Conclusions: Over the past 25 years, AI technology has shown promising performance in stroke diagnosis.

(J Med Internet Res 2024;26:e59711) doi:10.2196/59711

KEYWORDS

acute stroke; artificial intelligence; AI; machine learning; deep learning; stroke lesion segmentation and classification; stroke prediction; stroke prognosis

Introduction

Background

Stroke is a global public health issue, ranking as the second leading cause of death and the third leading cause of disability and death. One in every 4 people aged >25 years will experience a stroke in their lifetime. Stroke is responsible for 11.6% of deaths, and its incidence, mortality, and disability rates are on the rise [1-3].

Acute stroke refers to the clinical pathological state caused by the acute disruption of cerebral blood vessels. It can result in either the interruption of blood supply to the brain or the rupture of brain vessels, leading to damage to brain tissue. Ischemic strokes account for approximately 80% of all strokes, while hemorrhagic strokes make up approximately 20% [4]. Ischemic stroke is caused by reduced blood flow or blockage in the cerebral vessels, leading to oxygen and blood deprivation in brain tissue. Hemorrhagic stroke results from bleeding due to the rupture of cerebral vessels, with symptoms typically appearing within minutes and potentially leading to severe neurological deficits [5].

Over the past 25 years, artificial intelligence (AI) technology has achieved remarkable progress across various domains, notably in medical diagnostics [6]. Early AI applications

XSL•FO
RenderX

primarily used rule-based systems and machine learning (ML) models to analyze medical data and predict outcomes. With the advent of deep learning (DL), AI's ability to handle complex medical imaging data improved significantly. DL models, such as convolutional neural networks (CNNs), have been particularly effective in automatically identifying stroke lesions and detecting stroke risk factors in medical images. These models use large data sets to learn intricate patterns and features, improving diagnostic accuracy and efficiency. In addition, by integrating vast amounts of clinical data with imaging data, AI can now predict patient prognosis, assess functional recovery, and forecast treatment outcomes with greater precision [7] (Figure 1).

The swift and precise analysis of imaging data by AI offers detailed insights into the patient's condition, significantly supporting physicians in diagnostic decision-making and treatment planning. These AI models promote a deeper understanding of disease progression, enabling the development of personalized rehabilitation plans and thereby improving long-term patient outcomes [8].

Computed tomography (CT) and magnetic resonance imaging (MRI) are the most commonly recommended imaging methods for the clinical diagnosis of acute stroke. MRI has higher clarity and clinical sensitivity, offering better soft tissue contrast [9]. MRI imaging techniques used for brain examinations include T1-weighted imaging (T1WI), T2-weighted imaging (T2WI), fluid-attenuated inversion recovery (FLAIR), and diffusion-weighted imaging (DWI). T1WI provides excellent anatomical detail, but it is not very sensitive to early changes in acute stroke. T2WI can highlight increased water content in brain tissue and is commonly used for evaluating the subacute and chronic stages of stroke, although the images may be blurry and have artifacts. FLAIR can suppress cerebrospinal fluid signals, making it suitable for detecting white matter lesions and identifying lesions near the cerebrospinal fluid pathways. However, it has longer scanning times and is not sensitive enough for detecting small infarcts. DWI is highly sensitive to early acute cerebral ischemia, but DWI images have lower resolution and are not sufficiently sensitive to small infarcts [10]. Compared to MRI imaging, CT is faster and commonly used for detecting early signs of infarction. CT angiography (CTA) can provide information about vascular occlusion to guide treatment decisions, while CT perfusion (CTP) imaging can assess the extent of ischemic core and penumbra areas [11]. The 2018 American Heart Association and American Stroke Association guidelines indicate that noncontrast CT (NCCT) and CTA are recommended within 6 hours of acute stroke onset, while MRI and CTP are recommended for the 6- to 24-hour window [12,13] (Figure 2).

Ultrasound examination, with its advantages of fast imaging speed, lack of radiation, and lower cost, is commonly used in clinical practice for cardiac and carotid artery assessments [14,15] (Figure 3). For cerebral small vessel evaluation, T2WI and FLAIR-weighted MRI are commonly used in clinical settings [16].

Figure 1. Artificial intelligence used for acute stroke diagnosis over the past 25 years. CNN: convolutional neural network; kNN: k-nearest neighbors; NB: naive Bayes; ResNet: residual network; RF: random forest; SAM: Segment Anything Model; SVM: support vector machine.

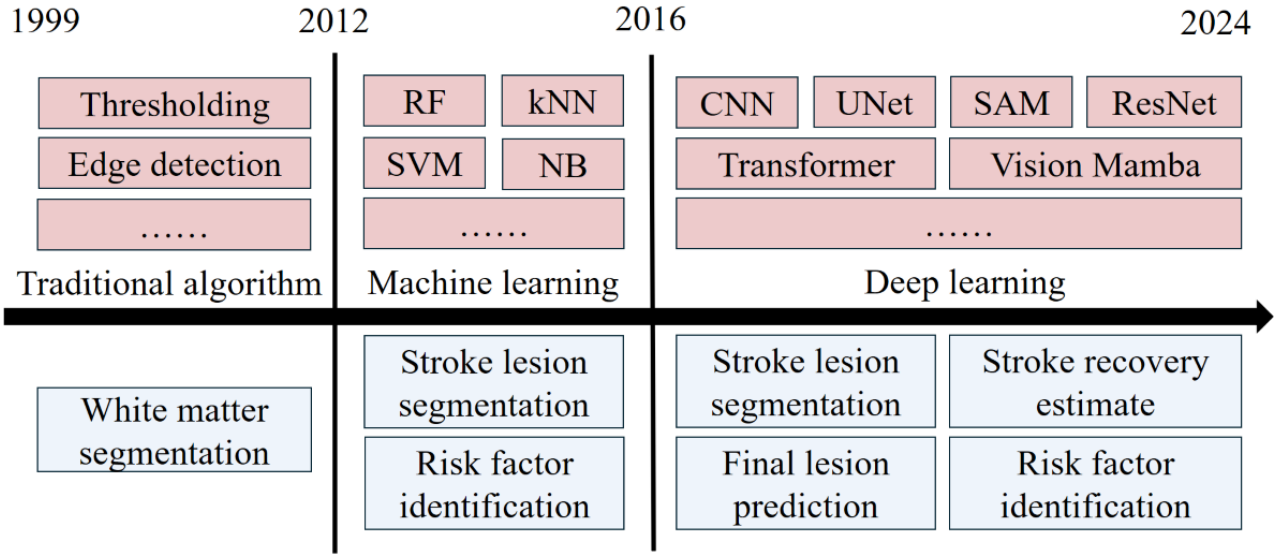

XSL•FO
RenderX

Figure 2. Stroke lesions identified using different imaging modalities. (A) Noncontrast computed tomography (NCCT). (B) CT perfusion (CTP). (C) T1-weighted imaging (T1WI). (D) T2-weighted imaging (T2WI). (E) Fluid-attenuated inversion recovery (FLAIR). (F) Diffusion-weighted imaging (DWI). (G) NCCT (annotated). (H) CTP (annotated). (I) T1WI (annotated). (J) T2WI (annotated). (K) FLAIR (annotated). (L) DWI (annotated). Stroke lesion areas are marked in blue, and the images are not paired.

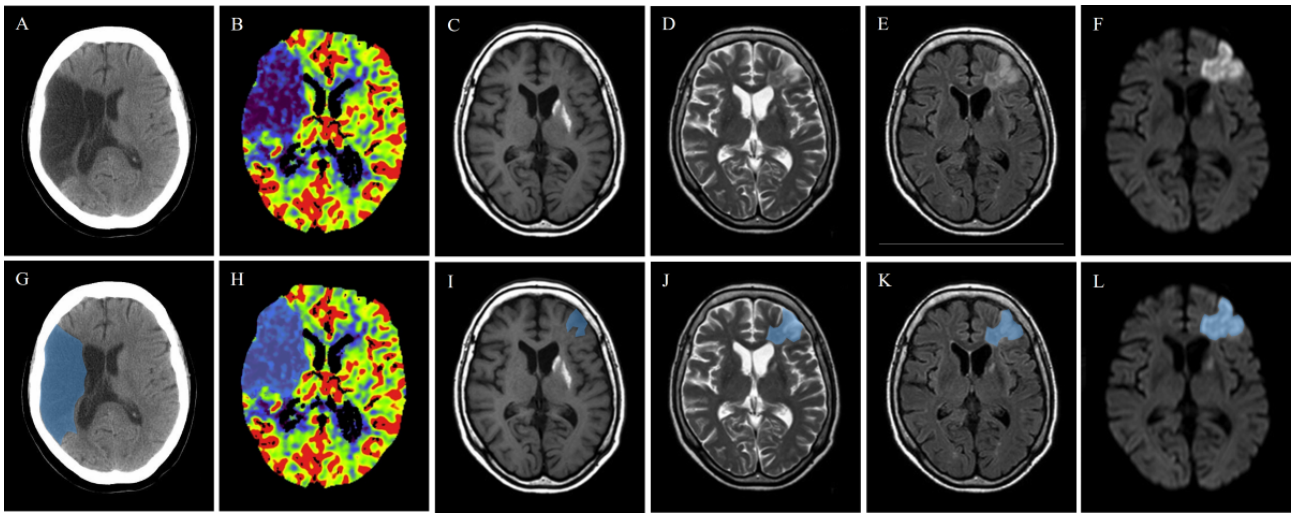

Figure 3. Cardiac ultrasound (US) and carotid US. (A) Cardiac US (2 heart chambers). (B) Cardiac US (4 heart chambers). (C) 3D carotid US. (D) 2D carotid US.

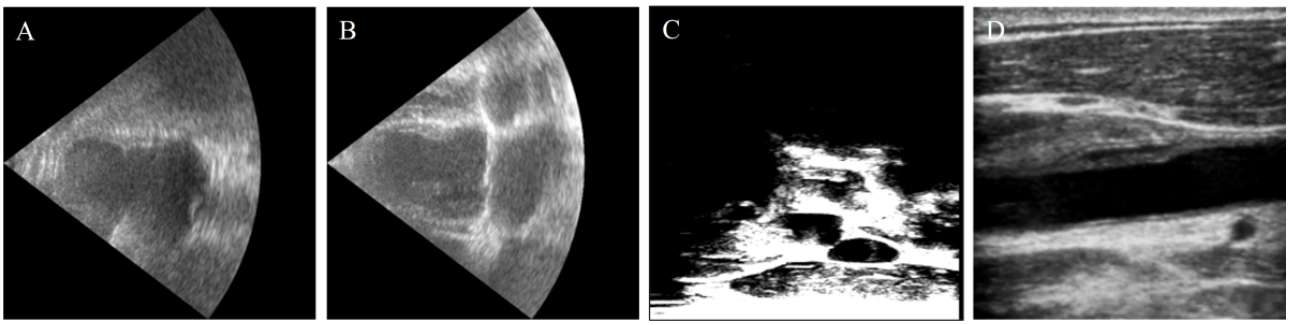

Main Contributions of This Review

The main contributions of this review are as follows: first, it provides a comprehensive summary of the common applications of AI in the diagnosis of acute stroke. Second, it outlines the development trends of AI algorithms for stroke lesion segmentation and classification, stroke risk prediction, and stroke prognosis from 1999 to 2024. This includes an analysis of data sources, types of algorithms, outcome metrics, and qualitative results described in each included article. Third, it discusses the clinical significance of the findings, existing challenges, and future research directions in this rapidly advancing field. The review aims to provide researchers and clinicians with insights into the current state of acute stroke diagnosis based on DL, offering a comprehensive reference for clinical practice.

Methods

Overview

For this systematic review, we adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [17]. We conducted a comprehensive literature search from January 1999 to February 2024 across the Google Scholar, PubMed, IEEE Xplore, Web of Science, and SpringerLink databases. The Boolean search string used was as follows: (ABSTRACT ("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "CNN") AND ABSTRACT ("ischemic stroke" OR "hemorrhagic stroke" OR "acute stroke" OR "stroke") AND ABSTRACT ("magnetic resonance imaging" OR "MRI" OR "Computed Tomography" OR "CT" OR "Ultrasound" OR "US")).

Evaluation Metrics

To assist readers in gaining a clearer understanding of these articles, we will define the model evaluation metrics featured in this review (Textbox 1). Common metrics include the Dice similarity coefficient (DSC) [18], Hausdorff distance [19], area under the curve (AUC) [20], accuracy [21], positive predictive value [22], recall [22], and specificity [22]. For image segmentation tasks, the main focus is on the DSC and Hausdorff distance metrics. For image classification tasks and prediction tasks, the main focus is on the accuracy, positive predictive value, recall, specificity, and AUC metrics (Table 1).

In this review, we also examine the statistical measures of significance and uncertainty, robustness or sensitivity analysis, methods for explainability or interpretability, and their validation (including validation or testing on external data), if reported in the studies [23].

XSL•FO
RenderX

Textbox 1. Acronyms with their definitions.

Acronyms and definitions

- AF: atrial fibrillation
- AI: artificial intelligence
- ASPECTS: Alberta Stroke Program Early CT Score
- AUC: area under the curve
- CAS: carotid artery stenosis
- CE: cardioembolism
- CNN: convolutional neural network
- CSVD: cerebral small vessel disease
- CT: computed tomography
- CTA: computed tomography angiography
- CTP: computed tomography perfusion
- DL: deep learning
- DNN: deep neural network
- DSC: Dice similarity coefficient
- DT: decision tree
- DWI: diffusion-weighted imaging
- FCN: fully convolutional network
- FLAIR: fluid-attenuated inversion recovery
- GRU: gated recurrent unit
- HD: Hausdorff distance
- kNN: k-nearest neighbors
- LA: left atrium
- LAA: large artery atherosclerosis
- LLM: large language model
- MHCA: multihead cross-attention
- ML: machine learning
- MRF: Markov random field
- MRI: magnetic resonance imaging
- NB: naive Bayes
- NCCT: noncontrast computed tomography
- PCA: principal component analysis
- PPV: positive predictive value
- RBF: radial basis function
- RBM: restricted Boltzmann machine
- RF: random forest
- ROI: region of interest
- SAM: Segment Anything Model
- SAO: small artery occlusion
- SE: squeeze and excitation
- STRIVE-1: Standards for Reporting Vascular Changes on Neuroimaging-1
- SVM: support vector machine
- T1WI: T1-weighted imaging

XSL•FO
RenderX

- T2WI: T2-weighted imaging
- US: ultrasound
- VIT: vision transformer
- VMamba: Vision Mamba
- WMH: white matter hyperintensities

Table 1. Equations and significance for evaluation metrics.

Metric	Equation	Significance
Segmentation metrics		
DSC[a] [18]		The overlap between segmentation results and labels; X and Y represent the segmentation result and label, respectively
HD[b] [19]		The maximum boundary distance between results and labels; denotes the directed Hausdorff distance from set A to set B
Classification and prediction metrics		
AUC[c] [20]	—[d]	Threshold-independent classification performance
Accuracy [21]		The proportion of correct predictions among all cases; TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative
PPV[e] [22]		The proportion of TPs among predicted positives
Recall[f] [22]		The proportion of TPs among all actual positives
Specificity [22]		The proportion of TNs among all actual negatives

[a]DSC: Dice similarity coefficient.
[b]HD: Hausdorff distance.
[c]AUC: area under the curve.
[d]Not applicable.
[e]PPV: positive predictive value (can also be referred to as *precision*).
[f]Recall can also be referred to as *sensitivity*.

Results

Literature Search Results

Overview

In the initial identification phase of the PRISMA flowchart (Figure 4), we identified 219 articles. During the screening phase, we reviewed the titles and abstracts of these publications to remove duplicates. We also excluded articles involving animals, other lesion types (eg, skin and lung), other stroke types (eg, perinatal stroke and chronic stroke), and other imaging modalities (eg, digital subtraction angiography). In addition, books, theses, review articles, and studies that did not propose an automated method based on ML or DL were excluded.

In the final inclusion phase, of the initial 219 articles, we retained 50 (22.8%) unique articles. These articles were thoroughly read to assess their eligibility for inclusion in this review. Articles were selected if they (1) relied on ML-based or DL-based automated methods and (2) included clinical records or medical images.

Segmentation and Classification of Stroke Lesions

Overview

Imaging examinations are typically included in the admission tests for patients with acute stroke [24,25]. Segmentation of stroke lesions can determine lesion volume and lesion location [26]. Classification of stroke lesions allows for the rapid identification of stroke type, aiding in patient triage [27]. This supports disease diagnosis, assists physicians in formulating treatment plans based on the patient's clinical presentation, and provides guidance for surgical or pharmacological interventions (Table 2).

XSL•FO
RenderX

Figure 4. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for systematic filtering and selection of articles.

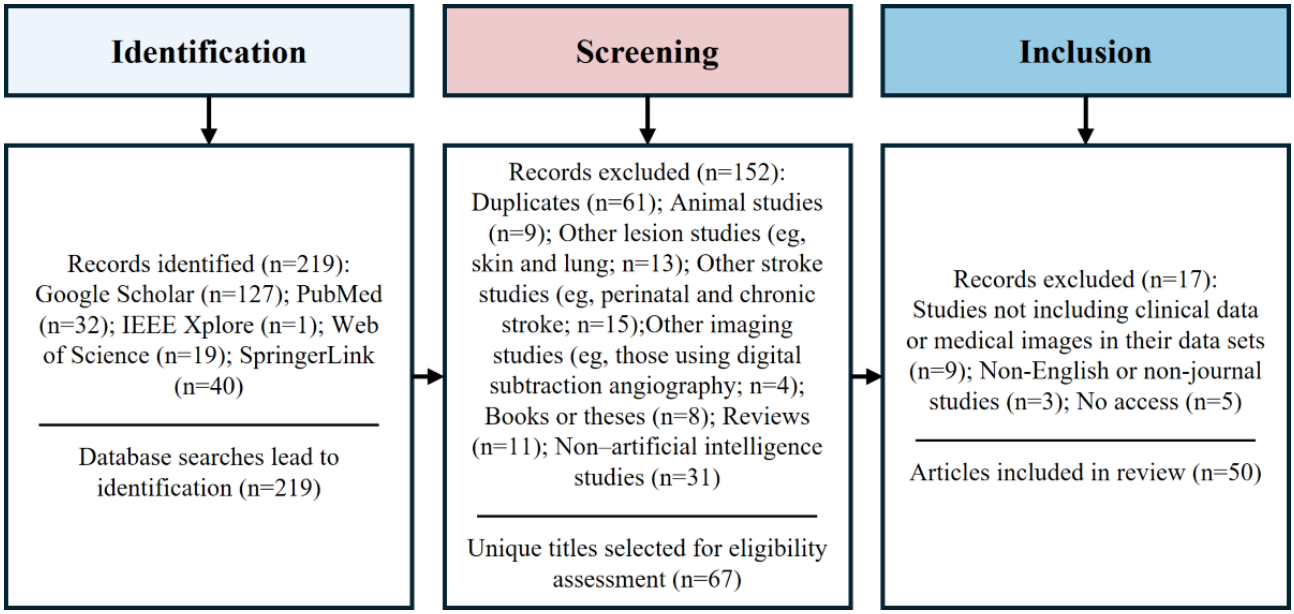

XSL•FO
RenderX

Table 2. Methods and performance of stroke lesion segmentation and classification.

Study, year	Image type	Data set, n	Methods	Performance	Merits
Machine learning (MRI[a])					
Maier et al [28], 2014	Multimodal MRI	36	SVM[b]	DSC[c]: 0.74	Uses SVM to extract local features
Mitra et al [29], 2014	Multimodal MRI	36	B-MRF[d], RF[e]	DSC: mean 0.39 (SD 0.22)	Multistage algorithm; fuses T1WI[f], T2WI[g], FLAIR[h], and DWI[i]
Maier et al [30], 2015	Multimodal MRI	37	RF	DSC: 0.65; HD[j]: 28.61 mm	Uses extra trees for voxel-level classification; incorporates intensity-derived image features
Pustina et al [31], 2016	T1WI	105	RF	DSC: mean 0.696, SD 0.16; HD: mean 17.9, SD 9.8 mm	Combines RF with multiresolution neighborhood data analysis
Griffis et al [32], 2016	T1WI	30	GNB[k]	DSC: 0.66	Incorporates probabilistic tissue segmentation; incorporates image algebra
Deep learning (CT[l])					
Goncharov et al [33], 2018	CTP[m]	156	UNet	DSC: mean 0.53, SD 0.01; HD: mean 23.29, SD 0.02 mm	Integrates UNet and T-Net (a variation of the UNet architecture); incorporates residual blocks
Clèrigues et al [34], 2019	CTP	156	CNN[n]	DSC: mean 0.49, SD 0.31; HD: mean 11.3, SD 31.6 mm	Uses symmetric modality augmentation; applies uncertainty filtering
Wang et al [35], 2020	Multimodal CT	156	CNN	DSC: mean 0.51, SD 0.31; HD: mean 0.55, SD 0.34 mm	Uses an innovative loss function; extracts features from CTA[o] images, synthesized pseudo-DWI
Kuang et al [36], 2021	NCCT[p]	230	CNN	DSC: 0.448	Combines CNN and classification network; introduces the ASPECTS[q] scale
Soltanpour et al [37], 2021	CTP	156	UNet	DSC: 0.69	Uses multiscale convolutional blocks, incorporates contralateral CT images; uses time-to-maximum heat maps
Luo et al [38], 2021	NCCT	293	Transformer, ResNet[r]-50	DSC: 0.7358	Combines ResNet-50 and transformer; introduces multihead cross-attention module
De Vries et al [39], 2023	CTP	156	UNet	DSC: 0.564	Incorporates symmetry-aware spatiotemporal convolutional structure; uses the dynamics of cerebral microperfusion; incorporates attention in skip connections
Kuang et al [40], 2024	NCCT	482	Transformer, CNN	DSC 0.5407	Introduces hybrid CNN transformer; introduces bilateral difference learning
Deep learning (MRI)					
Chen et al [41], 2017	DWI	741	CNN	DSC: 0.67	Multistage algorithm; introduces multiscale concept
Zhang et al [42], 2018	DWI	242	DenseNet	DSC: 0.7913	Combines dense connections and multiscale context
Karthik et al [43], 2019	Multimodal MRI	28	FCN[s]	DSC: 0.70	Uses leaky rectified linear unit in last 2 layers
Clèrigues et al [44], 2020	Multimodal MRI	236	UNet	DSC: 0.59, SD 0.31; HD: 0.84, SD 0.10 mm	Uses symmetric modality enhancement; uses dynamic weighted loss functions
Wu et al [45], 2023	Multimodal MRI	490	Transformer	DSC: 0.7258	Uses point-wise rendering to compute boundary

XSL•FO
RenderX

Study, year	Image type	Data set, n	Methods	Performance	Merits
Wu et al [46], 2023	Multimodal MRI	490	Transformer, CNN	DSC: 0.7368	Multistage algorithm; proposes 2 edge enhancement modules
Soh et al [47], 2023	T1WI	239	Transformer, UNet	DSC: 0.737	Comprises 2 parallel pipelines
Stroke classification					
Adam et al [48], 2016	NCCT	400	DT[t], CNN	AUC[u]: 0.994; recall: 0.987	Compares kNN[v] and DT
Gautam and Raman [49], 2021	NCCT	900	CNN	ACC: 98.33	Uses multifocus image fusion preprocessing
Chen et al [50], 2022	NCCT	400	CNN	Recall: 0.99	Uses hyperparameter optimization; uses transfer learning; compares optimized CNN and ResNet-50

[a]MRI: magnetic resonance imaging.
[b]SVM: support vector machine.
[c]DSC: Dice similarity coefficient.
[d]B-MRF: Bayesian-Markov random field.
[e]RF: random forest.
[f]T1WI: T1-weighted imaging.
[g]T2WI: T2-weighted imaging.
[h]FLAIR: fluid-attenuated inversion recovery.
[i]DWI: diffusion-weighted imaging.
[j]HD: Hausdorff distance.
[k]GNB: Gaussian naive Bayes.
[l]CT: computed tomography.
[m]CTP: computed tomography perfusion.
[n]CNN: convolutional neural network.
[o]CTA: computed tomography angiography.
[p]NCCT: noncontrast computed tomography.
[q]ASPECTS: Alberta Stroke Program Early Computed Tomography Score.
[r]ResNet: residual network.
[s]FCN: fully convolutional network.
[t]DT: decision tree.
[u]AUC: area under the curve.
[v]kNN: k-nearest neighbors.

Stroke Lesion Segmentation Based on ML

Stroke lesion segmentation is a challenging task due to the blurred boundaries between stroke lesions and normal tissues, particularly in hemorrhagic strokes. MRI images offer greater contrast and clarity compared to CT scans, allowing for more clearly delineated observation of stroke lesion boundaries [51].

Traditional thresholding and heuristic algorithms are inadequate for accurately segmenting stroke lesions. Before 2012, there was limited research on stroke lesion segmentation [52,53].

From 2012 to 2016, stroke lesion segmentation algorithms were primarily based on ML algorithms, with common segmentation methods mainly involving support vector machines (SVMs) [54], random forests (RFs) [55], and naive Bayes [56]. Due to the limited feature extraction capabilities of ML segmentation algorithms during this period, these algorithms primarily used MRI images.

In 2014, Maier et al [28] proposed an image segmentation method for ischemic stroke lesions based on SVMs using multimodal MRI images. The authors used SVMs to extract local features from the multimodal MRI data [28]. In the same year, Mitra et al [29] used a Bayesian-Markov random field for preliminary FLAIR image classification, then analyzed multimodal MRI data (T1WI, T2WI, FLAIR, and DWI) and contextually relevant features using an RF to identify likely lesion areas.

One year later, Maier et al [30] proposed another method for segmenting ischemic stroke lesions in multimodal MRI images using extra tree forests. This approach involved using extra trees for voxel-level classification and incorporated intensity-derived image features [30].

In 2016, Pustina et al [31] proposed an automatic segmentation algorithm for stroke lesions in T1WI images, combining RFs with multiresolution neighborhood data analysis to enhance efficiency and reduce observer dependency. In the same year, Griffis et al [32] used a stroke lesion segmentation method based

XSL•FO
RenderX

on the Gaussian naive Bayes classifier. By using probabilistic tissue segmentation and image algebra to create feature maps, the authors encoded information about missing and abnormal tissues and used the leave-one-out method for cross-validation [32].

Stroke Lesion Segmentation Based on DL

After 2016, there was a leap in the capabilities of DL for medical segmentation and feature extraction. DL has since outperformed ML as the main force in stroke image segmentation [57]. Due to the improved feature extraction capabilities of DL, algorithms for stroke lesion segmentation based on CT scans have emerged.

In 2018, Goncharov et al [33] proposed a CNN framework integrating UNet and T-Net (a variation of the UNet architecture). In each network, convolutional layers were replaced with residual blocks, and an initial convolution layer was added to enhance the learnability of the deep network [33].

In 2019, Clèrigues et al [34] proposed a 3D CTP segmentation method based on a 2D asymmetrical residual CNN. This method addresses issues such as small sample size, class imbalance, and overfitting by enhancing training, using symmetric modality augmentation, and applying uncertainty filtering, thereby improving performance and achieving fast inference [34].

In 2020, Wang et al [35] proposed a multimodal 3D CTP segmentation method based on CNNs. The authors extracted features from CTA images, synthesized pseudo-DWI, and used an innovative loss function to focus on lesion areas, thereby improving segmentation accuracy [35].

In 2021, Kuang et al [36] proposed a DL approach named EIS-Net, which combines a 3D triple CNN and a multiregion classification network for automatic segmentation of early ischemic stroke and Alberta Stroke Program Early CT Score scoring on NCCT in patients with acute ischemic stroke. In the same year, Soltanpour et al [37] proposed an improved version of the UNet network called MultiResUNet for the automatic segmentation of ischemic stroke lesions in CTP images. The algorithm uses multiscale convolutional blocks and incorporates contralateral CT images as references, estimating lesion locations using time-to-maximum heat maps [37]. Luo et al [38] proposed an ischemic stroke lesion segmentation method called UCATR for NCCT images. This method combines residual network (ResNet)-50 (a CNN that is 50 layers deep) and transformer encoders and introduces a multihead cross-attention module in the decoder to improve the accuracy of spatial information recovery [38].

In 2023, de Vries et al [39] proposed a model called Perf-UNet for segmenting the infarct core area from CTP source data. The model uses a symmetry-aware spatiotemporal convolutional structure, leveraging the dynamics of cerebral microperfusion, and incorporates attention modules in the skip connections [39].

In 2024, Kuang et al [40] proposed a hybrid CNN-transformer network based on circular feature interactions and bilateral difference learning. This network introduces a hybrid CNN-transformer in the encoder, a recurrent feature interaction module, and a shared CNN decoder with bilateral difference learning modules [40].

Compared to CT-based stroke lesion segmentation, MRI-based stroke lesion segmentation achieves higher segmentation accuracy.

In 2017, Chen et al [41] proposed a CNN network composed of a deconvolutional network for initial segmentation and a multiscale convolutional label evaluation network for eliminating false positives in small lesions.

In 2018, Zhang et al [42] proposed FC-DenseNet for segmenting acute stroke lesions in DWI medical images. This model combines dense connections and multiscale context on top of a 3D CNN to address common issues in DWI, such as noise, artifacts, and variations in lesion size and location [42].

In 2019, Karthik et al [43] proposed an improved fully convolutional network that applies leaky rectified linear unit activation in the last 2 layers of the UNet architecture. This approach precisely reconstructs ischemic lesions and allows the network to learn additional features not considered in the original UNet architecture [43].

In 2020, Clèrigues et al [44] addressed class imbalance issues by using symmetric modality enhancement, balanced training sample strategies, and dynamic weighted loss functions in the UNet architecture.

In 2023, Wu et al [45] introduced a lesion boundary–rendering method named TransRender. This method uses transformers to capture global information during the encoding phase and uses multiple renderings to effectively map encoding features of different levels to the original spatial resolution. The method adaptively selects points to compute boundary features based on point-wise rendering, supervising the rendering module to generate points and continuously refine uncertain regions [45]. In the same year, Wu et al [46] proposed a 2-phase brain multimodal MRI lesion segmentation method named W-Net. This method uses a CNN and a transformer as backbone networks, introduces a boundary deformation module and a boundary constraint module to address boundary ambiguity, and designs a multitask learning loss function to optimize W-Net from both regional and boundary perspectives [46]. Soh et al [47] proposed an algorithm named hybrid UNet transformer. This algorithm comprises 2 parallel pipelines, where the transformer-based pipeline uses feature maps from the intermediate layers of the UNet-based pipeline to enhance feature extraction capabilities [47].

Stroke Classification

The admission evaluation for patients with acute stroke typically includes stroke classification. Patients with acute ischemic stroke who are eligible for thrombolytic therapy should be referred to a thrombolytic center for treatment. For patients who are not suitable for thrombolytic therapy, conservative treatment should be provided upon admission. Patients with hemorrhagic stroke require rapid intervention to minimize brain damage and improve survival chances [24].

Acute ischemic stroke classification is typically based on CT imaging [58]. In CT images, ischemic stroke lesions appear as low-density regions (dark areas), while hemorrhagic stroke

XSL•FO
RenderX

lesions appear as high-density regions (white areas; Figure 5) [48-50].

Adam et al [48] proposed a stroke lesion classification model based on decision tree and k-nearest neighbors (kNN) algorithms. The study found that the decision tree algorithm outperformed kNN in classification [48].

Gautam and Raman [49] proposed a CNN model for classifying stroke CT images. The authors improved the quality of CT images by using multifocus image fusion preprocessing and then inputted the processed images into a 13-layer CNN architecture for classification [49].

Chen et al [50] used hyperparameter optimization and transfer learning to identify stroke conditions in brain CT images. The optimized CNN and ResNet-50 models demonstrated high accuracy in distinguishing between normal, hemorrhagic stroke, ischemic stroke, and other lesions. Although ResNet-50 exhibited the highest accuracy, it required more processing time [50].

Figure 5. Ischemic stroke and hemorrhagic stroke. (A) Ischemic stroke. (B) Ischemic stroke (annotated). (C) Hemorrhagic stroke. (D) Hemorrhagic stroke (annotated). Stroke lesion areas are marked in blue.

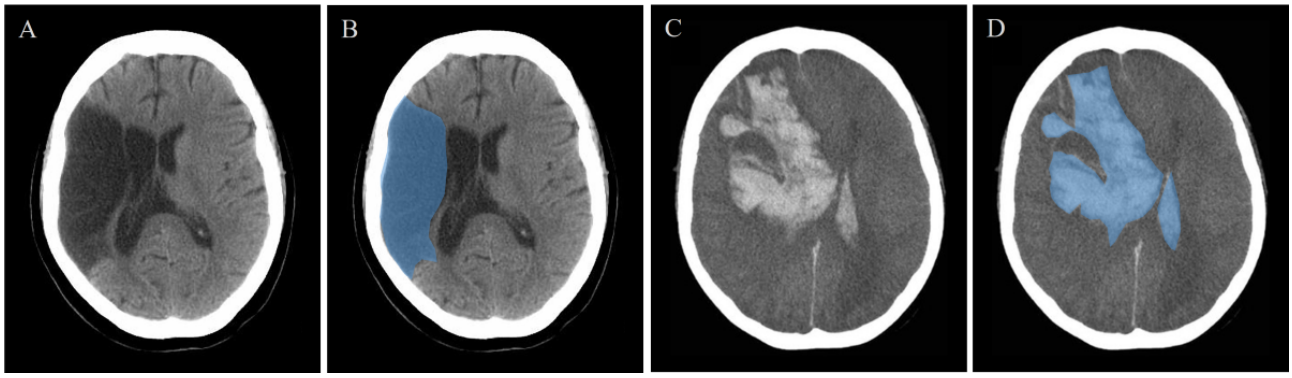

In-Depth Analysis of Stroke Lesion Segmentation and Classification

In clinical practice, the choice between using MRI and CT depends on the specific condition of the patient. CT has lower sensitivity in identifying early changes of acute ischemic stroke but demonstrates higher sensitivity in detecting characteristic high-density lesions of hemorrhagic stroke [59]. Compared to CT, MRI has higher contrast and a stronger ability to differentiate between brain tissues. The DWI sequence can detect minor changes in brain tissue just minutes after ischemia occurs. MRI also provides clearer identification of small hemorrhages and better differentiation between hemorrhagic and ischemic areas, along with detailed information on the extent and type of brain tissue damage after a stroke [60].

Before 2012, brain segmentation algorithms were primarily based on traditional methods for brain white matter segmentation. There are few studies focusing on the segmentation of stroke lesions in this period [52,53]. Between 2012 and 2016, segmentation algorithms for ischemic stroke lesions were primarily based on ML. Due to the blurred boundaries between stroke lesions and surrounding normal tissue in CT images, algorithms from this period mainly focused on MRI imaging. The accuracy of segmentation algorithms during this time generally did not exceed 0.75. These algorithms were complex, had weak generalization capabilities, and could not achieve the level of accuracy of manual segmentation. After 2016, with the development of UNet, DL has shown superior performance in the segmentation of stroke lesions [57]. As DL enables more in-depth feature extraction from medical images, significantly enhancing the extraction of effective features, an increasing number of researchers have started focusing on the segmentation of ischemic stroke lesions in CT images. However, the accuracy of these segmentation algorithms based on CT commonly did not surpass 0.75. Algorithms for segmenting ischemic stroke lesions based on MRI have shown better accuracy, enough to be comparable to manual segmentation. Some algorithms have achieved accuracies of up to 0.85. To facilitate the triage of patients with acute stroke, due to the urgency of the condition, the classification of stroke types primarily relies on CT images [58]. Current algorithms have achieved high accuracies, generally exceeding 0.90, indicating promising application prospects.

Data preprocessing is important for medical image segmentation. Preprocessing aims to standardize image quality, enhance valuable features, and reduce segmentation difficulties. Clèrigues et al [34] and Clèrigues et al [44] used symmetry-enhanced modalities, leveraging the high bilateral symmetry exhibited by the human brain in its natural state. This approach provides clearer and more accurate information for subsequent steps such as feature extraction, lesion detection, and image segmentation. For complex segmentation tasks such as stroke lesion segmentation, the introduction of multimodal approaches can significantly enhance algorithm performance by extracting complementary information from different imaging techniques. The studies by Mitra et al [29] and Wang et al [35] are particularly representative. Mitra et al [29] used multimodal MRI data, including T1WI, T2WI, FLAIR, and DWI, along with context-aware features. Wang et al [35] extracted features from CTA images, synthesized pseudo-DWI, for use in CTP segmentation methods.

In addition, some studies used multistage algorithms, combining coarse and fine-grained approaches to achieve better segmentation performance; for instance, Mitra et al [29] used a Bayesian-Markov random field model for preliminary classification of FLAIR images and used RFs for multimodal MRI segmentation. Chen et al [41] used a deconvolutional

XSL•FO
RenderX

network for initial segmentation and used a multiscale convolutional label evaluation network to evaluate and eliminate false positives of small lesions, achieving promising results on large data sets and demonstrating the algorithm's generalization ability.

Since 2021, various attention mechanisms and transformers have been widely applied in the field of medical image segmentation, leading to further improvements in the performance of stroke lesion segmentation algorithms. Transformers leverage self-attention mechanisms, enabling neural networks to capture global information and exhibit excellent performance when processing long sequence data [61]. De Vries et al [39] introduced attention mechanisms on the skip connections between the encoder and decoder, reducing the time dimension and only propagating the most informative features. Luo et al [38], Kuang et al [40], Wu et al [45], Wu et al [46], and Soh et al [47] all incorporated the concept of transformers, leading to significant improvements in algorithm performance.

Stroke Risk Prediction

Overview

Clinical methods for stroke risk prediction include clinical risk scores (eg, Framingham Stroke Risk Profile [62]), genetic risk analysis [63], serum biomarker detection [64], and imaging analysis [65]. Compared to other methods, imaging analysis provides direct anatomical and functional information, allowing clinicians to observe pathological conditions directly. In addition, imaging data can quantify stroke risk factors such as atrial fibrillation (AF), carotid artery stenosis (CAS), and cerebral small vessel disease, which are crucial for stroke risk prediction [66]. Furthermore, imaging techniques can monitor disease progression and treatment effects in real time, enabling clinicians to adjust treatment strategies promptly. Due to these advantages of imaging techniques, AI-based stroke risk prediction using medical imaging has become a major focus of research (Table 3).

XSL•FO
RenderX

Table 3. Methods and performance of stroke prediction.

Study, year	Image type	Data set, n	Methods	Performance	Merits
AI [a]-assisted detection of CAS[b]					
Alam et al [67], 2015	US[c]	200	RBF[d]	Accuracy: 0.982	Combines fuzzy c-means clustering; incorporates noise-handling methods
Araki et al [68], 2017	US	407	SVM[e]	Accuracy: 0.9428	Uses spatial domain filtering
Al-Mohannadi et al [69], 2021	US	100	Encoder-decoder	DSC[f]: 0.7423	Introduces edge operators
Mi et al [70], 2021	US	430	UNet	DSC: 0.780	A 3-branch network design; uses prior knowledge of the carotid artery
Latha et al [71], 2022	US	361	CapsuleNet	Accuracy: 0.967	Compares multiple DL[g] and ML[h] models
Jiang et al [72], 2023	US	244	CNN[i]	DSC: mean 0.927, SD 0.054	Introduces the concept of centerline extraction
AI-assisted detection of AF[j] and LA[k] enlargement					
Degel et al [73], 2018	US	161	CNN	DSC: mean 0.75, SD 0.17; HD[l]: mean 4.46, SD 2.73 mm	Introduces shape priors; introduces adversarial learning
Moradi et al [74], 2019	US	1137	UNet	DSC: mean 0.945, SD 0.12; HD: mean 1.62, SD 0.05 mm	Combines a pyramid network with UNet; uses the Niblack global thresholding method
Leclerc et al [75], 2020	US	2000	UNet	HD: mean 5.9, SD 3.6 mm	Cascaded network; presents a novel attention mechanism
Liu et al [76], 2021	US	3500	UNet	DSC: 0.949; HD: 4.33 mm	Proposes a pyramid local attention module; proposes a label coherence learning mechanism
Li et al [77], 2023	US	1764	ResNet[m]	DSC: mean 0.959, SD 0.002; HD: mean 0.422, SD 0.314 mm	Introduces residual and branched networks; introduces dilated convolutions; incorporates multiscale concepts in the decoder
AI-assisted detection of WMH[n]					
Liu et al [78], 2020	MRI[o]	75	CNN	DSC: 93.49	Cascaded network; uses dense blocks and dilated blocks; uses multiscale processing
Park et al [79], 2021	MRI	170	UNet	DSC: mean 0.8107, SD 0.0203	Incorporates an auxiliary classifier; applies the highlight foreground method; uses multimodal MRI images
Li et al [80], 2021	MRI	40	UNet	DSC: mean 0.736, SD 0.074	Introduces squeeze-and-excitation blocks and dense connection
Huang et al [81], 2023	MRI	455	VNet	DSC: mean 0.81, SD 0.06	Proposes the LSLoss[p] loss function
Farkhani et al [82], 2024	MRI	430	UNet	DSC: 0.8559	Incorporates consecutive self-attention encoders

[a]AI: artificial intelligence.
[b]CAS: carotid artery stenosis.
[c]US: ultrasound.
[d]RBF: radial basis function.
[e]SVM: support vector machine.
[f]DSC: Dice similarity coefficient.
[g]DL: deep learning.
[h]ML: machine learning.
[i]CNN: convolutional neural network.
[j]AF: atrial fibrillation.
[k]LA: left atrium.

XSL•FO
RenderX

[l]HD: Hausdorff distance.

[m]ResNet: residual network.

[n]WMH: white matter hyperintensities.

[o]MRI: magnetic resonance imaging.

[p]LSLoss: level set loss.

Risk Factors of Stroke

Stroke can be divided into 3 distinct subtypes with well-defined etiologies: large artery atherosclerosis, cardioembolism, and small artery occlusion (Figure 6) [66].

Large artery atherosclerosis-type stroke is typically caused by atherosclerotic stenosis in the carotid and intracranial large arteries. For patients with mild stenosis, conservative treatment is recommended. For those with severe stenosis, carotid endarterectomy surgery is performed to prevent the occurrence of ischemic stroke [71].

Cardioembolism-type stroke is caused by cardiogenic diseases, where emboli from the heart and aortic arch dislodge and circulate, leading to cerebral artery embolism [83]. Studies have shown that left atrium (LA) enlargement may lead to myocardial fibrosis and disruption of atrial muscle bundles, resulting in blood stasis within the atrium and reduced atrial contraction capacity. This can lead to AF, thereby increasing the risk of cardioembolic stroke [84].

Small artery occlusion-type stroke, also known as lacunar stroke, is a pathological condition caused by disease in the small penetrating arteries of the deep regions of the cerebral hemispheres or brainstem, leading to ischemic necrosis of the brain tissue supplied by these arteries [85]. The international Standards for Reporting Vascular Changes on Neuroimaging-1 clearly state that white matter hyperintensities (WMH) are neuroimaging indicators of small vessel disease and can predict the occurrence of stroke [16].

Figure 6. Risk factors and etiology of stroke. (A) Hemorrhagic stroke and ischemic stroke. (B) Large artery atherosclerosis–type stroke and its risk factors. (C) Cardioembolism-type stroke and its risk factors. (D) Small artery occlusion–type stroke and its risk factors. LA: left atrium.

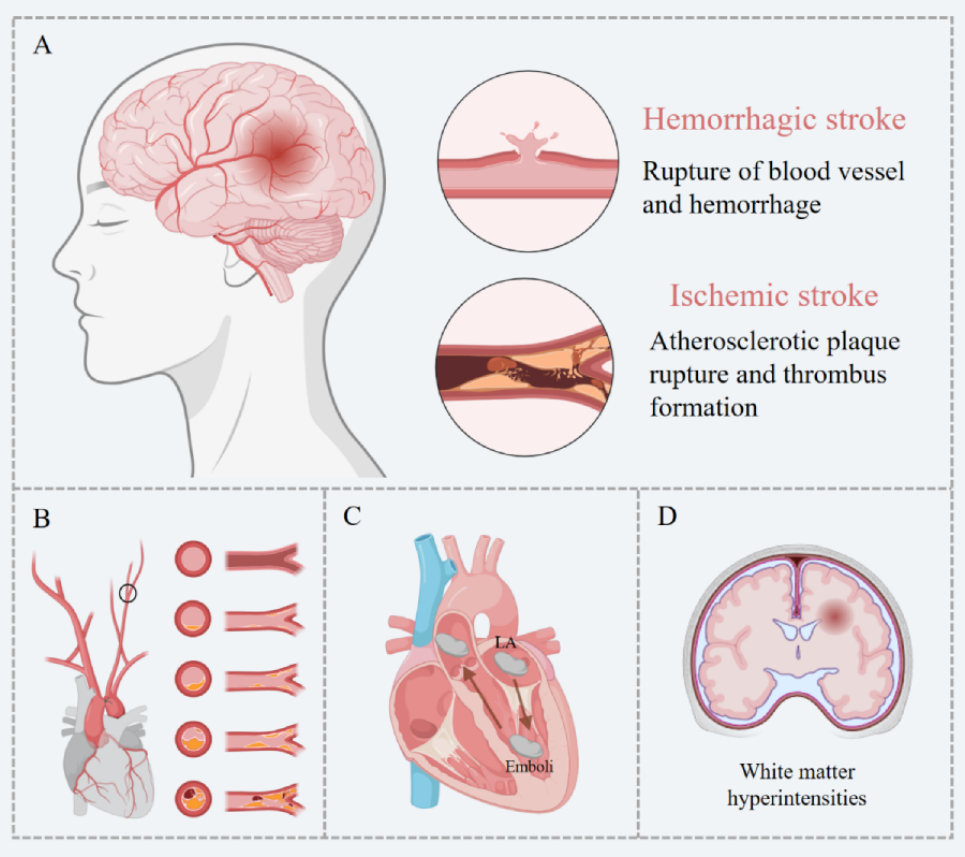

AI-Assisted Detection of CAS

The most commonly used imaging method for the detection of CAS in clinical practice is ultrasound (Figure 7) [86,87].

In 2015, Alam et al [67] proposed a robust fuzzy radial basis function network that combines fuzzy c-means clustering and a radial basis function network, incorporating spatial information and a smoothing parameter to handle noise.

In 2017, Araki et al [68] combined improved spatial domain filtering techniques with ML to use the morphological differences in carotid artery walls from ultrasound images. The authors used the improved spatial domain filtering techniques to remove image noise and used SVMs for classification [68].

In 2021, Al-Mohannadi et al [69] introduced morphological operations in an encoder-decoder structure to reduce the impact of noise on the results, using Sobel gradient direction images

XSL•FO
RenderX

and Prewitt gradient direction images as inputs. In the same year, Mi et al [70] designed an MBFF-Net (a multibranch feature fusion network) with 3 branches. The first 2 branches extract plaque features at multiple scales and different contexts, while the third branch uses prior information [70].

In 2022, Latha et al [88] compared various ML and DL algorithms for carotid artery segmentation. Among the ML algorithms, RF performed the best, while CapsuleNet outperformed other DL algorithms [88].

In 2023, Jiang and Chiu [72] integrated a centerline extraction network and a dual-stream centerline-guided network into a 3D UNet. The centerline extraction network generates centerline heat maps to indicate the position of the carotid artery centerline, while the centerline-guided network segments the 3D ultrasound images based on the centerline heat maps [72].

Figure 7. Carotid artery stenosis (CAS). (A) Healthy person. (B) Healthy person (annotated). (C) Patient with CAS. (D) Patient with CAS (annotated). Lumens are marked in blue.

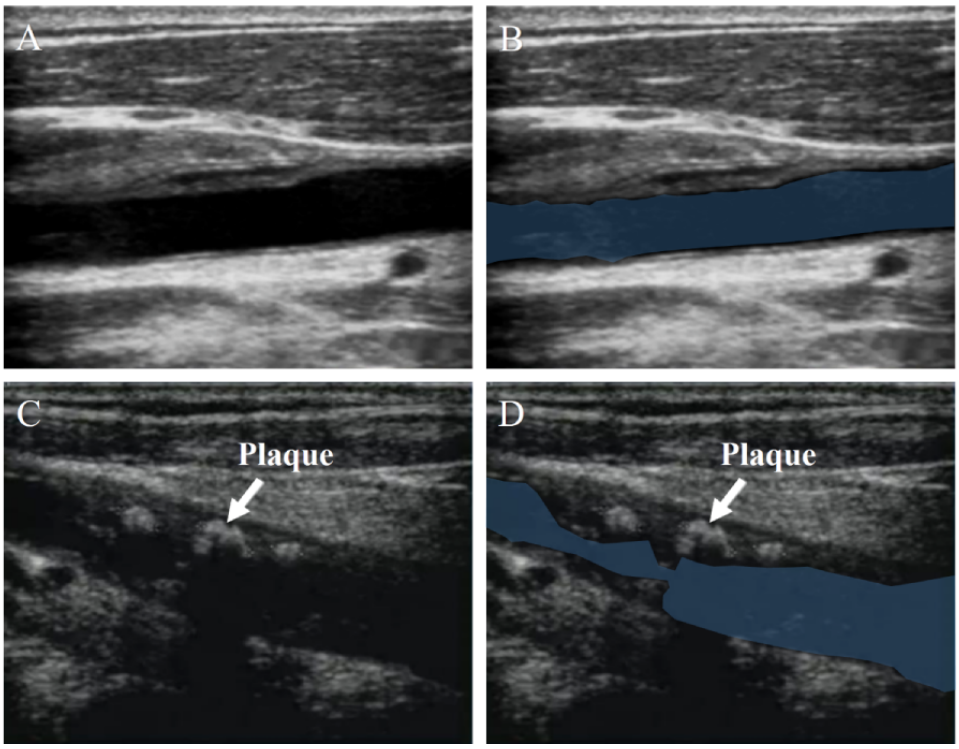

AI-Assisted Detection of AF and LA Enlargement

Similar to CAS, echocardiography is commonly used clinically to detect stroke-related heart disease, such as AF and LA enlargement (Figure 8) [15].

In 2018, Degel et al [73] proposed an algorithm that combines DL, shape priors, and adversarial learning to achieve automatic segmentation of the LA in 3D ultrasound images. This algorithm uses VNet for 3D volume segmentation and trains an autoencoder network to constrain the segmentation results to the shape of the LA [73].

In 2019, Moradi et al [74] proposed MFP-UNet, which combines a pyramid network with the UNet model, adding 2 downsampling layers to extract more detailed features from the images. The authors also used the Niblack global thresholding method to preprocess the cardiac ultrasound images to enhance contrast [74].

In 2020, Leclerc et al [75] proposed an RUNet composed of 2 UNets. The first UNet obtains the region of interest by dilating the preliminary results, and the second UNet is used to predict the final segmentation. RUNet outperformed other networks in reducing geometric and anatomical prediction anomalies [75].

In 2021, Liu et al [76] introduced a pyramid-shaped local attention mechanism in UNet to capture contextual information and designed a label consistency learning mechanism to improve the classification accuracy of edge pixels in cardiac ultrasound images.

In 2023, Li et al [77] proposed a multitask DL model for cardiac ultrasound segmentation and key point detection, named EchoEFNet. The network introduces ResNets, branched networks, and dilated convolutions in the encoder part and incorporates multiscale concepts in the decoder part, combining both low-level and high-level features [77].

XSL•FO
RenderX

Figure 8. The presence of left atrium (LA) enlargement. (A) Healthy person. (B) Healthy person (annotated). (C) Patient with LA enlargement. (D) Patient with LA enlargement (annotated). LA areas are marked in blue. LV: left ventricle; RA: right atrium; RV: right ventricle.

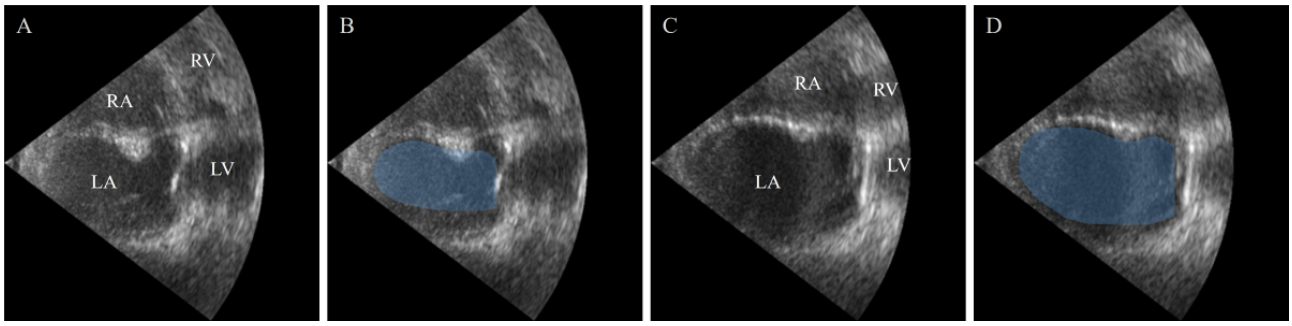

AI-Assisted Detection of WMH

MRI is superior to CT in terms of soft tissue contrast, allowing for better differentiation between white matter and gray matter in the brain. Therefore, MRI is more commonly used clinically to detect cerebral small vessel disease and WMH (Figure 9) [89,90].

In 2020, Liu et al [78] proposed the M2DCNN, a network composed of 2 symmetrical U-shaped subnetworks. This network uses dense blocks, dilated blocks, and multiscale processing to improve feature extraction and prediction capabilities. In addition, it introduces a loss function based on the similarity between lesions and background [78].

In 2021, Park et al [79] proposed a UNet with multiscale foreground highlighting, incorporating an auxiliary classifier in the middle layers of the decoder for deep supervision training. The authors used multiscale label images and applied the highlight foreground method to enhance foreground voxels [79].

In 2021, Li et al [80] introduced squeeze-and-excitation blocks and dense connections into the UNet architecture to better capture spatial and multiscale semantic information.

In 2023, Huang et al [81] proposed the level set loss (LSLoss) loss function. LSLoss consists of 4 loss terms, each designed to optimize foreground loss, background loss, region of interest loss, and divergence loss, to improve the accuracy of segmentation results [81].

In 2024, Farkhani et al [82] proposed the volumetric segmentation of WMH using transformer architecture, which enhances the UNet by using consecutive self-attention encoders to capture spatial dependencies in the data.

Figure 9. The presence of white matter hyperintensities (WMH). (A) Patient with minor WMH. (B) Patient with minor WMH (annotated). (C) Patient with extensive WMH. (D) Patient with extensive WMH (annotated). WMH areas are marked in blue.

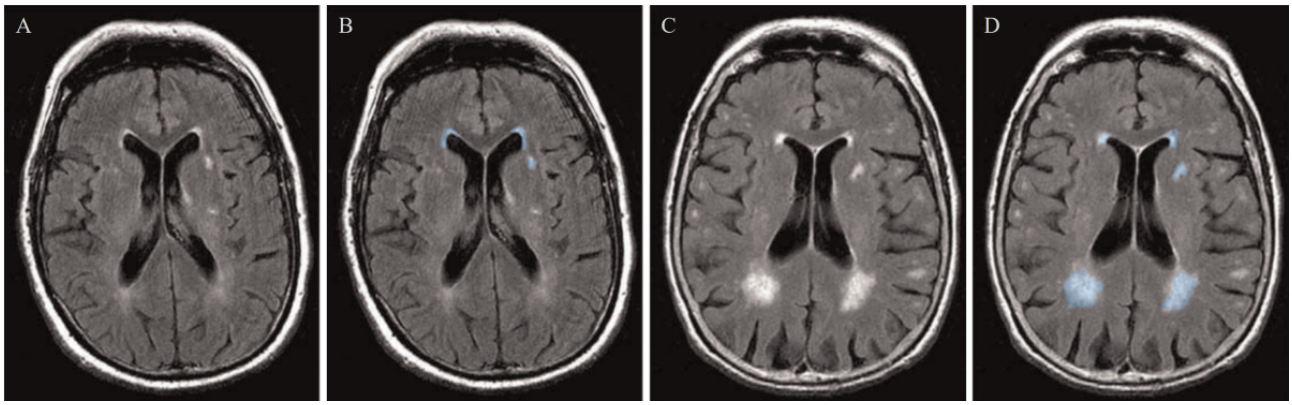

In-Depth Analysis of Stroke Risk Prediction

Carotid and cardiac ultrasound segmentation can assist physicians in quantitatively assessing the condition of a patient's heart and carotid arteries, providing important references for evaluating the risk of stroke. However, ultrasound image segmentation is a complex task. Compared to CT and MRI images, ultrasound images have lower contrast, noisy and blurry boundaries, and numerous artifacts, making ultrasound medical image segmentation more challenging.

For ultrasound image segmentation and classification, DL generally outperforms ML. Among ML algorithms, RF, with its capability to handle multifeature data, performs the best in handling these tasks with high accuracy and robustness. By contrast, other ML algorithms such as SVM and kNN perform well in certain specific tasks but are generally not as stable and accurate as RF. This is similar to the excellent performance of RF in stroke lesion segmentation and classification as discussed in the Segmentation and Classification of Stroke Lesions subsection, further demonstrating RF's superior performance in image segmentation and classification tasks [88].

To address the significant noise in ultrasound images, many researchers have incorporated denoising techniques into their algorithms. Alam et al [67] introduced spatial information and smoothing parameters to handle noise, and Araki et al [68] used spatial filtering techniques. In addition, due to the high noise and blurry boundaries in ultrasound images, advanced architectures such as transformers may not necessarily

XSL•FO
RenderX

outperform traditional convolutional blocks. In the comparative experiments in the study by Liu et al [76], traditional architectures such as UNet, UNet++, and ResNet performed better than Swin-UNet. Although Swin-UNet, with its advanced transformer architecture, significantly improved feature extraction and generalization capabilities, it exhibited overfitting in the segmentation of edge regions due to the substantial noise and blurred edges in ultrasound images.

The structures of carotid and cardiac ultrasound images are relatively fixed, and the introduction of morphological operations can better optimize these fixed structures. This effectively reduces the impact of noise and small artifacts, improves segmentation accuracy, enhances structural features, and increases robustness. Araki et al [68] used the morphological characteristics of the carotid artery wall to specifically improve spatial filtering, while Al-Mohannadi et al [69] introduced Sobel and Prewitt operators to make the model more attentive to edge information, aiding in the identification and segmentation of the boundaries of the carotid artery and the heart. Mi et al [70] and Degel et al [73] incorporated shape prior information, and Liu et al [76] introduced a label consistency learning mechanism. These 3 studies used the concept of templates to constrain the segmentation results and reduce missegmentation. For the segmentation of carotid vessels, centerline extraction can accurately determine the center position of the vessels, providing precise references for subsequent segmentation and feature extraction. This can largely ignore small noise and artifacts in the image, providing more stable segmentation results [72].

For WMH segmentation, many researchers have attempted to use attention mechanisms to enhance the model's focus on small targets. The attention mechanism allows the model to selectively focus on important regions of the image based on the task requirements, reducing dependence on irrelevant information [78]. Farkhani et al [82] achieved the precise capture of key regions through self-attention and hybrid transformer mechanisms, contributing to the improvement of WMH segmentation accuracy.

In the task of WMH segmentation, there is a significant class imbalance due to the large difference in the number of pixels between white matter and non–white matter regions. The WMH segmentation model tends to predict the more numerous classes while ignoring the minority classes, leading to overfitting. To address this issue, Liu et al [78] and Li et al [80] introduced dense blocks, which reduce the number of training parameters and alleviate the problem of vanishing gradients.

For small target segmentation, optimizing the loss function is a promising approach. In the study by Huang et al [81], an LSLoss function was proposed to train the segmentation network. This method allows for training without actual ground truth labels, helping to avoid overfitting the training data.

Stroke Prognosis

Overview

Stroke prognosis refers to predicting long-term outcomes such as recovery status, quality of life, and survival rates of patients based on medical imaging and clinical dates. AI technology can quantify the extent of brain tissue damage, assess patients' prognostic situations, and provide benchmarks for long-term rehabilitation (Table 4) [91].

XSL•FO
RenderX

Table 4. Methods and performance of stroke prognosis.

Study, year	Image type	Data set, n	Methods	Performance	Merits
Estimation of final infarct lesion and penumbra					
McKinley et al [92], 2017	MRI[a]	80	RF[b]	AUC[c]: mean 0.94, SD 0.08	An automated method to estimate the penumbra; uses multimodal MRI images
Nielsen et al [93], 2018	MRI	222	CNN[d]	AUC: mean 0.88, SD 0.12	Integrates time to maximum; uses apparent diffusion coefficient threshold
Yu et al [94], 2020	MRI	182	UNet	AUC: 0.92; DSC[e]: 0.58	Does not depend on reperfusion information
Pinto et al [95], 2021	MRI	75	CNN	DSC: 0.38; HD[f]: 29.21 mm	Accounts for the underlying cerebral blood flow; combines RBM[g] and GRU[h]
Wong et al [96], 2022	MRI	100	UNet	DSC: mean 0.88, SD 0.01; accuracy: mean 0.75, SD 0.03; AUC: mean 0.80, SD 0.03	Incorporates rotation and reflection variants; incorporates clinical indicators; predicts the modified Rankin Scale outcomes
Estimation of stroke recovery outcomes					
Bentley et al [97], 2014	CT[i], CRs[j]	116	SVM[k]	AUC: 0.744	Multimodal CRs
Monteiro et al [98], 2018	CRs	541	RF	AUC: mean 0.808, SD 0.085	Compares RF, XGBoost[l], and logistic regression
Cheon et al [99], 2019	CRs	15,099	DNN[m]	AUC: 0.8348	Incorporates PCA[n] into the DNN
Kuang et al [100], 2019	CT	257	RF	AUC: 0.89	Introduces the ASPECTS[o] scale
Scrutinio et al [101], 2020	CRs	1207	RF	AUC: 0.928; accuracy: 0.863	Implements the synthetic minority oversampling technique
Brugnara et al [102], 2020	CT, CRs	246	ML[p]	AUC: 0.856; accuracy: 0.804	Integrates multimodal CRs; predicts the modified Rankin Scale outcomes

[a]MRI: magnetic resonance imaging.
[b]RF: random forest.
[c]AUC: area under the curve.
[d]CNN: convolutional neural network.
[e]DSC: Dice similarity coefficient.
[f]HD: Hausdorff distance.
[g]RBM: restricted Boltzmann machine.
[h]GRU: gated recurrent unit.
[i]CT: computed tomography.
[j]CR: clinical record.
[k]SVM: support vector machine.
[l]XGBoost: extreme gradient boosting.
[m]DNN: deep neural network.
[n]PCA: principal component analysis.
[o]ASPECTS: Alberta Stroke Program Early Computed Tomography Score.
[p]ML: machine learning.

Estimation of Final Infarct Lesions and Penumbra

Acute stroke lesions tend to increase in size over time, leading to increased damage to brain tissue. To address this issue, some researchers have used AI to predict the final infarct lesions and the penumbra [103].

In 2017, McKinley et al [92] proposed an RF-based algorithm for penumbra prediction in multimodal MRI, called FASTER. This algorithm first extracts various statistical metrics from perfusion and diffusion imaging in the MRI data and then uses a decision forest algorithm to predict the penumbra [92].

In 2018, Nielsen et al [93] used CNNdeep to predict final infarct lesion, outperforming traditional general linear models, CNNs based on time to maximum, and apparent diffusion coefficient threshold methods.

In 2020, Yu et al [94] proposed a UNet-based model for predicting the final infarct lesions in patients with acute ischemic stroke. This model is the first DL-based stroke infarct prediction

XSL•FO
RenderX

model that does not rely on reperfusion information. The volume error between the predicted results and the ground truth is only 9 mm^3 [94].

In 2021, Pinto et al [95] proposed a method combining restricted Boltzmann machine and gated recurrent unit to predict the final stroke lesion in patients with acute ischemic stroke after 90 days. Each restricted Boltzmann machine learns features from different MRI parameter maps, which are then merged with the original maps and input into a convolutional and recurrent neural network architecture for supervised learning [95].

In 2022, Wong et al [96] developed a DL-based method for predicting the final infarct volume from MRI images. This approach uses networks with rotation and reflection variants and incorporates clinical variables to predict the 90-day modified Rankin Scale score [96].

Estimation of Stroke Recovery Outcomes

Clinical data can assist in estimating stroke recovery. Lesion volume and lesion location are critical factors in predicting recovery. Patients with extensive brain damage or strokes affecting critical brain areas generally have a poorer prognosis. Integrating patient clinical records with medical imaging can enhance the estimation of recovery outcomes [104].

Bentley et al [97] used SVMs to predict whether patients with acute ischemic stroke would develop symptomatic intracerebral hemorrhage after thrombolysis. Through the analysis of clinical records and brain CT images of 116 patients with acute stroke, the authors found that SVMs outperformed traditional prediction scoring systems in predicting symptomatic intracerebral hemorrhage [97].

Monteiro et al [98] investigated the use of ML methods to predict functional recovery in patients with ischemic stroke 3 months after the event. The study examined the performance of 5 different classifiers, among which RF exhibited the best performance [98].

Cheon et al [99] constructed a deep neural network model combined with principal component analysis for feature extraction. The model demonstrated high predictive ability based on clinical records from 15,099 patients with stroke [99].

Kuang et al [100] proposed an automated method for Alberta Stroke Program Early CT Score scoring from NCCT images of patients with acute ischemic stroke. This method uses texture features to train an RF classifier and performs well in identifying large areas of early infarction [100].

Scrutinio et al [101] used an RF algorithm enhanced with synthetic minority oversampling technique to predict the 3-year mortality rate of patients with severe stroke, significantly outperforming the logistic regression model.

Brugnara et al [102] integrated clinical records, multimodal imaging, and angiographic features to predict the modified Rankin Scale outcomes of patients with acute ischemic stroke after endovascular treatment. With the increase in medical information modalities, the model performance gradually improves [102].

In-Depth Analysis of Stroke Prognosis

For stroke prognosis, both DL and ML have their respective advantages [105]. DL models require large amounts of data for training to learn sufficiently complex features and avoid overfitting. This also explains why the deep neural network in the study by Cheon et al [99] achieved better performance after being trained on 15,099 clinical records, although acquiring a large amount of high-quality clinical data for stroke prognosis may be challenging.

By contrast, traditional ML methods have demonstrated greater robustness and efficiency in handling stroke prognosis tasks. Studies have shown that when handling stroke prognosis tasks, RF and its variant models have shown better performance than other ML algorithms [98,100,101]. This is mainly due to the robustness of RF models against data noise.

For the estimation of the final infarct lesion and penumbra, although the existing estimation may not achieve high DSC values, metrics such as volume error and AUC have reached high levels, indicating good practical value. In the future, through continual optimization and improvement of algorithms, the clinical translation prospects of AI in final infarct estimation will be even broader.

Discussion

Principal Findings

Over the past 25 years, significant advances have been made in AI-assisted stroke diagnosis. From 1999 to 2012, development was constrained by the limited performance of traditional thresholding and heuristic algorithms, resulting in low diagnostic accuracy and restricted clinical applicability. From 2012 to 2016, the advent of, and the advances made by, ML algorithms significantly expanded research in AI-assisted stroke diagnosis. The implementation of advanced statistical models and classifiers enabled more detailed analysis of stroke imaging data and improved diagnostic precision. After 2016, DL techniques, particularly CNNs, demonstrated significant improvements over traditional methods and ML algorithms [106].

AI has significantly enhanced the accuracy and efficiency of stroke lesion segmentation and classification, stroke risk prediction (eg, assessment of LA enlargement, CAS, and WMH), and stroke prognosis (eg, prediction of stroke recovery outcomes, final infarct lesion, and penumbra). Stroke lesion segmentation and classification enable more precise diagnoses, reduce the time to diagnosis, and improve patient treatment outcomes. Stroke risk prediction allows for timely intervention, potentially reducing stroke occurrence. Stroke prognosis helps physicians develop personalized rehabilitation plans, optimize treatment strategies in real time, and provide more accurate prognostic information to patients and their families. Integrating AI into clinical practice enables health care providers to use advanced data analysis and pattern recognition capabilities to streamline workflows, reduce diagnostic errors, enhance the diagnostic process, and improve the continuous monitoring and management of patients with stroke [107].

XSL•FO
RenderX

However, notwithstanding the aforementioned advances, several challenges remain in clinical applications. First, stroke diagnosis is a complex process that requires the integration of multimodal information (eg, medical imaging, genomic data, and clinical records) for comprehensive evaluation. However, the effective integration of multimodal data with diverse formats is challenging. Second, AI algorithms lack effective clinical validation. The generalization ability of AI models is limited, with many algorithms performing inconsistently across different data sets and imaging devices. In addition, given the inherent "black box" nature of AI algorithms, it is essential to develop transparent and interpretable models to help clinicians build trust in AI-assisted diagnosis. Finally, the use of AI in stroke diagnosis presents privacy and data security issues. Traditional data-sharing and model training methods may lead to information leakage.

The recent emergence of models such as the Segment Anything Model (SAM) [108], Vision Mamba (VMamba) [109], and multimodal large language models (LLMs) [110] has provided new directions for stroke diagnosis. SAM boasts a large parameter count and introduces pretrained weights from vision transformer, resulting in excellent generalization performance for medical image segmentation [108]. MedSAM, a fine-tuned version of the original SAM tailored for medical images, shows a 22.51% improvement in DSC on zero-shot medical image segmentation tasks compared to the untuned SAM [111]. nnSAM integrates pretrained SAM encoders and UNet encoders in parallel, combining robust feature extraction capabilities with better adaptability to different image segmentation tasks [112]. SAM performs well on external data sets from various imaging devices without further training, demonstrating potential to address the issue of limited generalization ability.

Multimodal LLMs can effectively integrate multimodal clinical data, enabling a more comprehensive understanding of patient conditions [113,114]. Gu et al [115] used LLMs for stroke quantitative evaluation using multimodal data, achieving an interrater agreement of 0.823 with expert ratings, demonstrating outstanding performance and application prospects. Meanwhile, VMamba represents a further improvement and upgrade of the transformer architecture. It can capture global information while maintaining linear computational complexity, addressing the performance bottleneck caused by the quadratic computational complexity of transformers [109]. Whether VMamba can achieve a leap in performance for stroke diagnosis similar to transformers remains to be seen and will require further evaluation over time.

Federated learning is able to address data privacy and security issues, allowing different institutions to collaboratively train AI models without sharing raw data. Federated learning not only protects patient privacy but also facilitates cross-institutional data sharing and collaboration, improving the generalization ability and accuracy of the models [116].

Overall, standardization is a key factor in promoting the application of AI in the medical field. It is necessary to establish unified evaluation standards and operational procedures to ensure that different institutions and researchers can consistently use AI technologies. In addition, medical professionals need continuous education and training on AI to better understand and trust the technology, which will enable them to apply it more effectively in clinical practice [117].

Conclusions

This paper reviewed the current status, challenges, and development trends of AI in acute stroke lesion segmentation and classification, stroke risk prediction, and stroke prognosis over the past 25 years. AI-assisted stroke diagnosis has now shown good performance, assisting physicians in making rapid diagnoses and improving patient outcomes. With the development of new AI technologies such as SAM, LLMs, and VMamba, AI-assisted diagnosis of acute stroke is expected to achieve higher accuracy and stability in the future.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA_2020_checklist.
[PDF File (Adobe PDF File), 64 KB - jmir_v26i1e59711_app1.pdf]

References

1. Tu WJ, Wang LD, Special Writing Group of China Stroke Surveillance Report. China stroke surveillance report 2021. Mil Med Res 2023 Jul 19;10(1):33 [FREE Full text] [doi: 10.1186/s40779-023-00463-x] [Medline: 37468952]
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021 Oct;20(10):795-820 [FREE Full text] [doi: 10.1016/S1474-4422(21)00252-0] [Medline: 34487721]
3. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. Int J Stroke 2022 Jan 05;17(1):18-29. [doi: 10.1177/17474930211065917] [Medline: 34986727]
4. Xu H, He K, Hu R, Ge Y, Li X, Ni F, et al. Identifying key biomarkers and immune infiltration in female patients with ischemic stroke based on weighted gene co-expression network analysis. Neural Plast 2022 Apr 8;2022:5379876 [FREE Full text] [doi: 10.1155/2022/5379876] [Medline: 35432523]

XSL•FO
RenderX

5. Morotti A, Poli L, Costa P. Acute stroke. Semin Neurol 2019 Feb 11;39(1):61-72. [doi: 10.1055/s-0038-1676992] [Medline: 30743293]
6. Murray NM, Unberath M, Hager GD, Hui FK. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. J Neurointerv Surg 2020 Feb;12(2):156-164. [doi: 10.1136/neurintsurg-2019-015135] [Medline: 31594798]
7. Zhao Z, Zhang Y, Su J, Yang L, Pang L, Gao Y, et al. A comprehensive review for artificial intelligence on neuroimaging in rehabilitation of ischemic stroke. Front Neurol 2024 Mar 28;15:1367854 [FREE Full text] [doi: 10.3389/fneur.2024.1367854] [Medline: 38606275]
8. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
9. van Beek EJ, Kuhl C, Anzai Y, Desmond P, Ehman RL, Gong Q, et al. Value of MRI in medicine: more than just another test? J Magn Reson Imaging 2019 Jun 25;49(7):e14-e25 [FREE Full text] [doi: 10.1002/jmri.26211] [Medline: 30145852]
10. El-Brashy MA, Mohamed HH, Ebied OM. Role of magnetic resonant diffusion-weighted imaging in evaluation of acute cerebral stroke. Menoufia Med J 2014;27(4). [doi: 10.4103/1110-2098.149744]
11. Wintermark M, Albers GW, Broderick JP, Demchuk AM, Fiebach JB, Fiehler J, et al. Acute stroke imaging research roadmap II. Stroke 2013 Sep;44(9):2628-2639 [FREE Full text] [doi: 10.1161/STROKEAHA.113.002015] [Medline: 23860298]
12. Leslie-Mazwi T, Chandra RV, Fraser JF, Hoh B, Baxter BW, Albuquerque FC, et al. AHA/ASA 2018 AIS guidelines: impact and opportunity for endovascular stroke care. J Neurointerv Surg 2018 Sep;10(9):813-817. [doi: 10.1136/neurintsurg-2018-013911] [Medline: 29807886]
13. Haggenmüller B, Kreiser K, Sollmann N, Huber M, Vogele D, Schmidt SA, et al. Pictorial review on imaging findings in cerebral CTP in patients with acute stroke and its mimics: a primer for general radiologists. Diagnostics (Basel) 2023 Jan 26;13(3):447 [FREE Full text] [doi: 10.3390/diagnostics13030447] [Medline: 36766552]
14. Wang Y, Yao Y. Application of artificial intelligence methods in carotid artery segmentation: a review. IEEE Access 2023 Feb 6;11:13846-13858. [doi: 10.1109/ACCESS.2023.3243162]
15. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging 2019 Sep;38(9):2198-2210. [doi: 10.1109/TMI.2019.2900516] [Medline: 30802851]
16. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol 2013 Aug;12(8):822-838. [doi: 10.1016/s1474-4422(13)70124-8]
17. Parums DV. Editorial: review articles, systematic reviews, meta-analysis, and the updated preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 guidelines. Med Sci Monit 2021 Aug 23;27:e934475 [FREE Full text] [doi: 10.12659/MSM.934475] [Medline: 34421116]
18. Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. Int J Radiat Oncol Biol Phys 2018 Jun 01;101(2):468-478 [FREE Full text] [doi: 10.1016/j.ijrobp.2018.01.114] [Medline: 29559291]
19. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993 Sep;15(9):850-863. [doi: 10.1109/34.232073]
20. Purves RD. Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). J Pharmacokinet Biopharm 1992 Jun 1;20(3):211-226. [doi: 10.1007/BF01062525] [Medline: 1522479]
21. Bishop CM. Pattern Recognition and Machine Learning. New York, NY: Springer; 2006.
22. Shreffler J, Huecker MR. Diagnostic testing accuracy: sensitivity, specificity, predictive values and likelihood ratios. In: StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; 2023.
23. Kim DY, Oh HW, Suh CH. Reporting quality of research studies on AI applications in medical images according to the CLAIM guidelines in a radiology journal with a strong prominence in Asia. Korean J Radiol 2023 Dec;24(12):1179-1189 [FREE Full text] [doi: 10.3348/kjr.2023.1027] [Medline: 38016678]
24. Lima FO, Silva GS, Furie KL, Frankel MR, Lev MH, Camargo É, et al. Field assessment stroke triage for emergency destination: a simple and accurate prehospital scale to detect large vessel occlusion strokes. Stroke 2016 Aug;47(8):1997-2002 [FREE Full text] [doi: 10.1161/STROKEAHA.116.013301] [Medline: 27364531]
25. Bouchez L, Sztajzel R, Vargas MI, Machi P, Kulcsar Z, Poletti PA, et al. CT imaging selection in acute stroke. Eur J Radiol 2017 Nov;96:153-161. [doi: 10.1016/j.ejrad.2016.10.026] [Medline: 28277288]
26. Guerrero R, Qin C, Oktay O, Bowles C, Chen L, Joules R, et al. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. Neuroimage Clin 2018;17:918-934 [FREE Full text] [doi: 10.1016/j.nicl.2017.12.022] [Medline: 29527496]

XSL•FO
RenderX

27. Amarenco P, Bogousslavsky J, Caplan LR, Donnan GA, Hennerici MG. Classification of stroke subtypes. Cerebrovasc Dis 2009 Apr 3;27(5):493-501 [FREE Full text] [doi: 10.1159/000210432] [Medline: 19342825]
28. Maier O, Wilms M, Gablentz J, Krämer U, Handels H. Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers. In: Proceedings of the SPIE Medical Imaging 2014. 2014 Presented at: SPIE 2014; February 16-18, 2014; San Diego, CA. [doi: 10.1117/12.2043494]
29. Mitra J, Bourgeat P, Fripp J, Ghose S, Rose S, Salvado O, et al. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. Neuroimage 2014 Sep;98:324-335. [doi: 10.1016/j.neuroimage.2014.04.056] [Medline: 24793830]
30. Maier O, Wilms M, von der Gablentz J, Krämer UM, Münte TF, Handels H. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. J Neurosci Methods 2015 Jan 30;240:89-100. [doi: 10.1016/j.jneumeth.2014.11.011] [Medline: 25448384]
31. Pustina D, Coslett HB, Turkeltaub PE, Tustison N, Schwartz MF, Avants B. Automated segmentation of chronic stroke lesions using LINDA: lesion identification with neighborhood data analysis. Hum Brain Mapp 2016 Apr 12;37(4):1405-1421 [FREE Full text] [doi: 10.1002/hbm.23110] [Medline: 26756101]
32. Griffis JC, Allendorfer JB, Szaflarski JP. Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. J Neurosci Methods 2016 Jan 15;257:97-108 [FREE Full text] [doi: 10.1016/j.jneumeth.2015.09.019] [Medline: 26432931]
33. Goncharov M, Pisov M, Belyaev M. Ensembling convolutional neural networks for ischemic stroke lesion segmentation based on CT images. In: The Eleventh International Conference И Т и С 2018. 2018 Presented at: И Т и С 2018; August 20-25, 2018; Novosibirsk, Russia.
34. Clèrigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Lladó X. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. Comput Biol Med 2019 Dec;115:103487 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.103487] [Medline: 31629272]
35. Wang G, Song T, Dong Q, Cui M, Huang N, Zhang S. Automatic ischemic stroke lesion segmentation from computed tomography perfusion images by image synthesis and attention-based deep neural networks. Med Image Anal 2020 Oct;65:101787. [doi: 10.1016/j.media.2020.101787] [Medline: 32712524]
36. Kuang H, Menon BK, Sohn SI, Qiu W. EIS-Net: segmenting early infarct and scoring ASPECTS simultaneously on non-contrast CT of patients with acute ischemic stroke. Med Image Anal 2021 May;70:101984. [doi: 10.1016/j.media.2021.101984] [Medline: 33676101]
37. Soltanpour M, Greiner R, Boulanger P, Buck B. Improvement of automatic ischemic stroke lesion segmentation in CT perfusion maps using a learned deep neural network. Comput Biol Med 2021 Oct;137:104849. [doi: 10.1016/j.compbiomed.2021.104849] [Medline: 34530336]
38. Luo C, Zhang J, Chen X, Tang Y, Weng X, Xu F. UCATR: based on CNN and transformer encoding and cross-attention decoding for lesion segmentation of acute ischemic stroke in non-contrast computed tomography images. In: Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2021 Presented at: EMBC 2021; November 01-05, 2021; Virtual Event. [doi: 10.1109/embc46164.2021.9630336]
39. de Vries L, Emmer BJ, Majoie CB, Marquering HA, Gavves E. PerfU-Net: baseline infarct estimation from CT perfusion source data for acute ischemic stroke. Med Image Anal 2023 Apr;85:102749 [FREE Full text] [doi: 10.1016/j.media.2023.102749] [Medline: 36731276]
40. Kuang H, Wang Y, Liu J, Wang J, Cao Q, Hu B, et al. Hybrid CNN-transformer network with circular feature interaction for acute ischemic stroke lesion segmentation on non-contrast CT scans. IEEE Trans Med Imaging 2024 Jun;43(6):2303-2316. [doi: 10.1109/tmi.2024.3362879]
41. Chen L, Bentley P, Rueckert D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. Neuroimage Clin 2017;15:633-643 [FREE Full text] [doi: 10.1016/j.nicl.2017.06.016] [Medline: 28664034]
42. Zhang R, Zhao L, Lou W, Abrigo JM, Mok VC, Chu WC, et al. Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional DenseNets. IEEE Trans Med Imaging 2018 Sep;37(9):2149-2160. [doi: 10.1109/TMI.2018.2821244] [Medline: 29994088]
43. Karthik R, Gupta U, Jha A, Rajalakshmi R, Menaka R. A deep supervised approach for ischemic lesion segmentation from multimodal MRI using fully convolutional network. Appl Soft Comput 2019 Nov;84:105685. [doi: 10.1016/j.asoc.2019.105685]
44. Clèrigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Lladó X. Acute and sub-acute stroke lesion segmentation from multimodal MRI. Comput Methods Programs Biomed 2020 Oct;194:105521. [doi: 10.1016/j.cmpb.2020.105521] [Medline: 32434099]
45. Wu Z, Zhang X, Li F, Wang S, Li J. TransRender: a transformer-based boundary rendering segmentation network for stroke lesions. Front Neurosci 2023;17:1259677 [FREE Full text] [doi: 10.3389/fnins.2023.1259677] [Medline: 37901438]
46. Wu Z, Zhang X, Li F, Wang S, Huang L, Li J. W-Net: a boundary-enhanced segmentation network for stroke lesions. Expert Syst Appl 2023 Nov;230:120637. [doi: 10.1016/j.eswa.2023.120637]
47. Soh WK, Yuen HY, Rajapakse JC. HUT: Hybrid UNet transformer for brain lesion and tumour segmentation. Heliyon 2023 Dec;9(12):e22412 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e22412] [Medline: 38046150]

XSL•FO
RenderX

48. Yahiya SY, Yousif A, Bakri MB. Classification of ischemic stroke using machine learning algorithms. Int J Comput Appl 2016 Sep 15;149(10):26-31 [FREE Full text] [doi: 10.5120/ijca2016911607]
49. Gautam A, Raman B. Towards effective classification of brain hemorrhagic and ischemic stroke using CNN. Biomed Signal Process Control 2021 Jan;63:102178. [doi: 10.1016/j.bspc.2020.102178]
50. Chen YT, Chen YL, Chen YY, Huang YT, Wong HF, Yan JL, et al. Deep learning-based brain computed tomography image classification with hyperparameter optimization through transfer learning for stroke. Diagnostics (Basel) 2022 Mar 25;12(4):807 [FREE Full text] [doi: 10.3390/diagnostics12040807] [Medline: 35453855]
51. Gillebert CR, Humphreys GW, Mantini D. Automated delineation of stroke lesions using brain CT images. Neuroimage Clin 2014;4:540-548 [FREE Full text] [doi: 10.1016/j.nicl.2014.03.009] [Medline: 24818079]
52. Matesin M, Loncaric S, Petravic D. A rule-based approach to stroke lesion analysis from CT brain images. In: Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces. 2001 Presented at: ISPA 2001; June 19-21, 2001; Pula, Croatia. [doi: 10.1109/ispa.2001.938631]
53. Shen S, Szameitat AJ, Sterr A. An improved lesion detection approach based on similarity measurement between fuzzy intensity segmentation and spatial probability maps. Magn Reson Imaging 2010 Feb;28(2):245-254. [doi: 10.1016/j.mri.2009.06.007] [Medline: 19695812]
54. Dayang P, Kouyim Meli AS. Evaluation of image segmentation algorithms for plant disease detection. Int J Image Graph Signal Process 2021 Oct 08;13(5):14-26. [doi: 10.5815/ijigsp.2021.05.02]
55. Breiman L. Random forests. Mach Learn 200;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
56. Rish I. An empirical study of the naive Bayes classifier. In: Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. 2001 Presented at: IJCAI 2001; August 4-6, 2001; Seattle, WA.
57. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustainability 2021 Jan 25;13(3):1224. [doi: 10.3390/su13031224]
58. Inamdar MA, Raghavendra U, Gudigar A, Chakole Y, Hegde A, Menon GR, et al. A review on computer aided diagnosis of acute brain stroke. Sensors (Basel) 2021 Dec 20;21(24):8507 [FREE Full text] [doi: 10.3390/s21248507] [Medline: 34960599]
59. Chawla M, Sharma S, Sivaswamy J, Kishore L. A method for automatic detection and classification of stroke from brain CT images. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:3581-3584. [doi: 10.1109/IEMBS.2009.5335289] [Medline: 19965232]
60. Barber PA, Darby DG, Desmond PM, Gerraty RP, Yang Q, Li T, et al. Identification of major ischemic change. Diffusion-weighted imaging versus computed tomography. Stroke 1999 Oct;30(10):2059-2065. [doi: 10.1161/01.str.30.10.2059] [Medline: 10512907]
61. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. arXiv Preprint posted online February 27, 2021.
62. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. Stroke 1991 Mar;22(3):312-318. [doi: 10.1161/01.str.22.3.312] [Medline: 2003301]
63. Boehme AK, Esenwa C, Elkind MS. Stroke risk factors, genetics, and prevention. Circ Res 2017 Feb 03;120(3):472-495 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.308398] [Medline: 28154098]
64. Jaroslav P, Christian R, Stefan O, Alexander Z, Zepper P, Holger P, et al. Evaluation of serum biomarkers for patients at increased risk of stroke. Int J Vasc Med 2012;2012:906954 [FREE Full text] [doi: 10.1155/2012/906954] [Medline: 22506118]
65. Saba L, Saam T, Jäger HR, Yuan C, Hatsukami TS, Saloner D, et al. Imaging biomarkers of vulnerable carotid plaques for stroke risk prediction and their potential clinical implications. Lancet Neurol 2019 Jun;18(6):559-572. [doi: 10.1016/S1474-4422(19)30035-3] [Medline: 30954372]
66. Adams HPJ, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 1993 Jan;24(1):35-41. [doi: 10.1161/01.str.24.1.35] [Medline: 7678184]
67. Alam J, Hassan M, Khan A, Chaudhry A. Robust fuzzy RBF network based image segmentation and intelligent decision making system for carotid artery ultrasound images. Neurocomputing 2015 Mar;151:745-755. [doi: 10.1016/j.neucom.2014.10.027]
68. Araki T, Jain PK, Suri HS, Londhe ND, Ikeda N, El-Baz A, et al. Stroke risk stratification and its validation using ultrasonic echolucent carotid wall plaque morphology: a machine learning paradigm. Comput Biol Med 2017 Jan 01;80:77-96. [doi: 10.1016/j.compbiomed.2016.11.011] [Medline: 27915126]
69. Al-Mohannadi A, Al-Maadeed S, Elharrouss O, Sadasivuni KK. Encoder-decoder architecture for ultrasound IMC segmentation and cIMT measurement. Sensors (Basel) 2021 Oct 14;21(20):6839 [FREE Full text] [doi: 10.3390/s21206839] [Medline: 34696054]
70. Mi S, Bao Q, Wei Z, Xu F, Yang W. MBFF-Net: multi-branch feature fusion network for carotid plaque segmentation in ultrasound. In: Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. 2021 Presented at: MICCAI 2021; September 27-October 1, 2021; Strasbourg, France. [doi: 10.1007/978-3-030-87240-3_30]

XSL•FO
RenderX

71. Flaherty ML, Kissela B, Khoury JC, Alwell K, Moomaw CJ, Woo D, et al. Carotid artery stenosis as a cause of stroke. Neuroepidemiology 2013 Oct 11;40(1):36-41 [FREE Full text] [doi: 10.1159/000341410] [Medline: 23075828]
72. Jiang M, Chiu B. A dual-stream centerline-guided network for segmentation of the common and internal carotid arteries from 3D ultrasound images. IEEE Trans Med Imaging 2023 Sep;42(9):2690-2705. [doi: 10.1109/TMI.2023.3263537] [Medline: 37015114]
73. Degel MA, Navab N, Albarqouni S. Domain and geometry agnostic CNNs for left atrium segmentation in 3D ultrasound. In: Proceedings of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention. 2018 Presented at: MICCAI 2018; September 16-20, 2018; Granada, Spain. [doi: 10.1007/978-3-030-00937-3_72]
74. Moradi S, Oghli MG, Alizadehasl A, Shiri I, Oveisi N, Oveisi M, et al. MFP-Unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Phys Med 2019 Nov;67:58-69. [doi: 10.1016/j.ejmp.2019.10.001] [Medline: 31671333]
75. Leclerc S, Smistad E, Grenier T, Lartizien C, Ostvik A, Cervenansky F, et al. RU-Net: a refining segmentation network for 2D echocardiography. In: Proceedings of the IEEE International Ultrasonics Symposium. 2019 Presented at: IUS 2019; October 6-9, 2019; Glasgow, UK.
76. Liu F, Wang K, Liu D, Yang X, Tian J. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. Med Image Anal 2021 Jan;67:101873 [FREE Full text] [doi: 10.1016/j.media.2020.101873] [Medline: 33129143]
77. Li H, Wang Y, Qu M, Cao P, Feng C, Yang J. EchoEFNet: multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography. Comput Biol Med 2023 Apr;156:106705 [FREE Full text] [doi: 10.1016/j.compbiomed.2023.106705] [Medline: 36863190]
78. Liu L, Chen S, Zhu X, Zhao XM, Wu FX, Wang J. Deep convolutional neural network for accurate segmentation and quantification of white matter hyperintensities. Neurocomputing 2020 Apr;384:231-242. [doi: 10.1016/j.neucom.2019.12.050]
79. Park G, Hong J, Duffy BA, Lee JM, Kim H. White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds. Neuroimage 2021 Aug 15;237:118140 [FREE Full text] [doi: 10.1016/j.neuroimage.2021.118140] [Medline: 33957235]
80. Li X, Zhao Y, Jiang J, Cheng J, Zhu W, Wu Z, et al. White matter hyperintensities segmentation using an ensemble of neural networks. Hum Brain Mapp 2022 Feb 15;43(3):929-939 [FREE Full text] [doi: 10.1002/hbm.25695] [Medline: 34704337]
81. Huang F, Xia P, Vardhanabhuti V, Hui SK, Lau KK, Ka-Fung Mak H, et al. Semisupervised white matter hyperintensities segmentation on MRI. Hum Brain Mapp 2023 Mar 10;44(4):1344-1358 [FREE Full text] [doi: 10.1002/hbm.26109] [Medline: 36214210]
82. Farkhani S, Demnitz N, Boraxbekk CJ, Lundell H, Siebner HR, Petersen ET, et al. End-to-end volumetric segmentation of white matter hyperintensities using deep learning. Comput Methods Programs Biomed 2024 Mar;245:108008 [FREE Full text] [doi: 10.1016/j.cmpb.2024.108008] [Medline: 38290291]
83. Raghib MF, Mutzenbach JS, Rösler C, Otto F, Coy MM, Müller-Thies-Broussalis E, et al. Acute treatment of stroke due to spontaneous calcified cerebral emboli causing large vessel occlusion. J Clin Neurosci 2018 Jan;47:56-61. [doi: 10.1016/j.jocn.2017.10.042] [Medline: 29102234]
84. Casaclang-Verzosa G, Gersh BJ, Tsang TS. Structural and functional remodeling of the left atrium: clinical and therapeutic implications for atrial fibrillation. J Am Coll Cardiol 2008 Jan 01;51(1):1-11 [FREE Full text] [doi: 10.1016/j.jacc.2007.09.026] [Medline: 18174029]
85. Caprio FZ, Lin C. Unusual causes of ischemic stroke and transient ischemic attack. In: Hankey GJ, Macleod M, Gorelick PB, Chen C, Caprio FZ, Mattle H, editors. Warlow's Stroke: Practical Management, Fourth Edition. Hoboken, NJ: John Wiley & Sons; 2019.
86. Alexandratou M, Papachristodoulou A, Li X, Partovi S, Davidhi A, Rafailidis V, et al. Advances in noninvasive carotid wall imaging with ultrasound: a narrative review. J Clin Med 2022 Oct 20;11(20):6196 [FREE Full text] [doi: 10.3390/jcm11206196] [Medline: 36294515]
87. Jaff MR, Goldmakher GV, Lev MH, Romero JM. Imaging of the carotid arteries: the role of duplex ultrasonography, magnetic resonance arteriography, and computerized tomographic arteriography. Vasc Med 2008 Nov;13(4):281-292. [doi: 10.1177/1358863X08091971] [Medline: 18940905]
88. Latha S, Muthu P, Lai KW, Khalil A, Dhanalakshmi S. Performance analysis of machine learning and deep learning architectures on early stroke detection using carotid artery ultrasound images. Front Aging Neurosci 2021 Jan 27;13:828214 [FREE Full text] [doi: 10.3389/fnagi.2021.828214] [Medline: 35153728]
89. Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 2010 Jul 26;341:c3666 [FREE Full text] [doi: 10.1136/bmj.c3666] [Medline: 20660506]
90. Das AS, Regenhardt RW, Vernooij MW, Blacker D, Charidimou A, Viswanathan A. Asymptomatic cerebral small vessel disease: insights from population-based studies. J Stroke 2019 May;21(2):121-138 [FREE Full text] [doi: 10.5853/jos.2018.03608] [Medline: 30991799]

XSL•FO
RenderX

91. Ertl M, Meisinger C, Linseisen J, Baumeister SE, Zickler P, Naumann M. Long-term outcomes in patients with stroke after in-hospital treatment-study protocol of the prospective stroke cohort Augsburg (SCHANA Study). Medicina (Kaunas) 2020 Jun 06;56(6):280 [FREE Full text] [doi: 10.3390/medicina56060280] [Medline: 32517235]
92. McKinley R, Häni L, Gralla J, El-Koussy M, Bauer S, Arnold M, et al. Fully automated stroke tissue estimation using random forest classifiers (FASTER). J Cereb Blood Flow Metab 2017 Aug;37(8):2728-2741 [FREE Full text] [doi: 10.1177/0271678X16674221] [Medline: 27798267]
93. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Stroke 2018 Jun;49(6):1394-1401. [doi: 10.1161/strokeaha.117.019740]
94. Yu Y, Xie Y, Thamm T, Gong E, Ouyang J, Huang C, et al. Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging. JAMA Netw Open 2020 Mar 02;3(3):e200772 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0772] [Medline: 32163165]
95. Pinto A, Pereira S, Meier R, Wiest R, Alves V, Reyes M, et al. Combining unsupervised and supervised learning for predicting the final stroke lesion. Med Image Anal 2021 Apr;69:101888. [doi: 10.1016/j.media.2020.101888] [Medline: 33387909]
96. Wong KK, Cummock JS, Li G, Ghosh R, Xu P, Volpi JJ, et al. Automatic segmentation in acute ischemic stroke: prognostic significance of topological stroke volumes on stroke outcome. Stroke 2022 Sep;53(9):2896-2905. [doi: 10.1161/STROKEAHA.121.037982] [Medline: 35545938]
97. Bentley P, Ganesalingam J, Carlton Jones AL, Mahady K, Epton S, Rinne P, et al. Prediction of stroke thrombolysis outcome using CT brain machine learning. Neuroimage Clin 2014;4:635-640 [FREE Full text] [doi: 10.1016/j.nicl.2014.02.003] [Medline: 24936414]
98. Monteiro M, Fonseca AC, Freitas AT, Pinho e Melo T, Francisco AP, Ferro JM, et al. Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM Trans Comput Biol Bioinf 2018 Nov 1;15(6):1953-1959. [doi: 10.1109/tcbb.2018.2811471]
99. Cheon S, Kim J, Lim J. The use of deep learning to predict stroke patient mortality. Int J Environ Res Public Health 2019 May 28;16(11):1876 [FREE Full text] [doi: 10.3390/ijerph16111876] [Medline: 31141892]
100. Kuang H, Najm M, Chakraborty D, Maraj N, Sohn SI, Goyal M, et al. Automated ASPECTS on noncontrast CT scans in patients with acute ischemic stroke using machine learning. AJNR Am J Neuroradiol 2019 Jan;40(1):33-38 [FREE Full text] [doi: 10.3174/ajnr.A5889] [Medline: 30498017]
101. Scrutinio D, Ricciardi C, Donisi L, Losavio E, Battista P, Guida P, et al. Machine learning to predict mortality after rehabilitation among patients with severe stroke. Sci Rep 2020 Nov 18;10(1):20127 [FREE Full text] [doi: 10.1038/s41598-020-77243-3] [Medline: 33208913]
102. Brugnara G, Neuberger U, Mahmutoglu MA, Foltyn M, Herweh C, Nagel S, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. Stroke 2020 Dec;51(12):3541-3551. [doi: 10.1161/strokeaha.120.030287]
103. Desmurget M, Bonnetblanc F, Duffau H. Contrasting acute and slow-growing lesions: a new door to brain plasticity. Brain 2007 Apr;130(Pt 4):898-914. [doi: 10.1093/brain/awl300] [Medline: 17121742]
104. Beloosesky Y, Streifler JY, Burstin A, Grinblat J. The importance of brain infarct size and location in predicting outcome after stroke. Age Ageing 1995 Nov;24(6):515-518. [doi: 10.1093/ageing/24.6.515] [Medline: 8588543]
105. Fang G, Huang Z, Wang Z. Predicting ischemic stroke outcome using deep learning approaches. Front Genet 2021 Jan 24;12:827522 [FREE Full text] [doi: 10.3389/fgene.2021.827522] [Medline: 35140746]
106. Bivard A, Churilov L, Parsons M. Artificial intelligence for decision support in acute stroke - current roles and potential. Nat Rev Neurol 2020 Oct 24;16(10):575-585. [doi: 10.1038/s41582-020-0390-y] [Medline: 32839584]
107. Recht MP, Dewey M, Dreyer K, Langlotz C, Niessen W, Prainsack B, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol 2020 Jun;30(6):3576-3584. [doi: 10.1007/s00330-020-06672-5] [Medline: 32064565]
108. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: an experimental study. Med Image Anal 2023 Oct;89:102918. [doi: 10.1016/j.media.2023.102918] [Medline: 37595404]
109. Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X. Vision mamba: efficient visual representation learning with bidirectional state space model. arXiv Preprint posted online January 17, 2024.
110. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023 Nov 02;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]
111. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024 Jan 22;15(1):654 [FREE Full text] [doi: 10.1038/s41467-024-44824-z] [Medline: 38253604]
112. Li Y, Jing B, Li Z, Wang J, Zhang Y. nnSAM: plug-and-play segment anything model improves nnUNet performance. arXiv Preprint posted online September 29, 2023 [FREE Full text]
113. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024 Jan 09;12(Suppl 1):e002583 [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]

XSL•FO
RenderX

114. Chen Z, Balan MM, Brown K. Language models are few-shot learners for prognostic prediction. arXiv Preprint posted online February 24, 2023.
115. Gu Z, He X, Yu P, Jia W, Yang X, Peng G, et al. Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain-adaptive pre-trained large language model. Artif Intell Med 2024 Apr;150:102822 [FREE Full text] [doi: 10.1016/j.artmed.2024.102822] [Medline: 38553162]
116. Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. Comput Ind Eng 2020 Nov;149:106854. [doi: 10.1016/j.cie.2020.106854]
117. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MD, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 04;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]

Abbreviations

AF: atrial fibrillation
AI: artificial intelligence
AUC: area under the curve
CAS: carotid artery stenosis
CNN: convolutional neural network
CT: computed tomography
CTA: computed tomography angiography
CTP: computed tomography perfusion
DL: deep learning
DSC: Dice similarity coefficient
DWI: diffusion-weighted imaging
FLAIR: fluid-attenuated inversion recovery
kNN: k-nearest neighbors
LA: left atrium
LLM: large language model
LSLoss: level set loss
ML: machine learning
MRI: magnetic resonance imaging
NCCT: noncontrast computed tomography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ResNet: residual network
RF: random forest
SAM: Segment Anything Model
SVM: support vector machine
T1WI: T1-weighted imaging
T2WI: T2-weighted imaging
VMamba: Vision Mamba
WMH: white matter hyperintensities

Edited by G Eysenbach; submitted 20.04.24; peer-reviewed by J Qin, Y Zhang, S Sakilay, K Gray; comments to author 08.06.24; revised version received 25.06.24; accepted 15.07.24; published 10.09.24.

Please cite as:
Wang Z, Yang W, Li Z, Rong Z, Wang X, Han J, Ma L
A 25-Year Retrospective of the Use of AI for Diagnosing Acute Stroke: Systematic Review
J Med Internet Res 2024;26:e59711
URL: https://www.jmir.org/2024/1/e59711
doi:10.2196/59711
PMID:

©Zhaoxin Wang, Wenwen Yang, Zhengyu Li, Ze Rong, Xing Wang, Jincong Han, Lei Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Data Analytics in Physical Activity Studies With Accelerometers: Scoping Review

Ya-Ting Liang[1,2], MS; Charlotte Wang[2,3], PhD; Chuhsing Kate Hsiao[2,3], PhD

[1]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
[2]Institute of Health Data Analytics and Statistics, College of Public Health, National Taiwan University, Taipei, Taiwan
[3]Master of Public Health Program, College of Public Health, National Taiwan University, Taipei, Taiwan

Corresponding Author:
Chuhsing Kate Hsiao, PhD
Institute of Health Data Analytics and Statistics
College of Public Health
National Taiwan University
No.17, Xu-Zhou Rd
Taipei, 10055
Taiwan
Phone: 886 2 33668032
Email: ckhsiao@ntu.edu.tw

Abstract

Background: Monitoring free-living physical activity (PA) through wearable devices enables the real-time assessment of activity features associated with health outcomes and provision of treatment recommendations and adjustments. The conclusions of studies on PA and health depend crucially on reliable statistical analyses of digital data. Data analytics, however, are challenging due to the various metrics adopted for measuring PA, different aims of studies, and complex temporal variations within variables. The application, interpretation, and appropriateness of these analytical tools have yet to be summarized.

Objective: This research aimed to review studies that used analytical methods for analyzing PA monitored by accelerometers. Specifically, this review addressed three questions: (1) What metrics are used to describe an individual's free-living daily PA? (2) What are the current analytical tools for analyzing PA data, particularly under the aims of classification, association with health outcomes, and prediction of health events? and (3) What challenges exist in the analyses, and what recommendations for future research are suggested regarding the use of statistical methods in various research tasks?

Methods: This scoping review was conducted following an existing framework to map research studies by exploring the information about PA. Three databases, PubMed, IEEE Xplore, and the ACM Digital Library, were searched in February 2024 to identify related publications. Eligible articles were classification, association, or prediction studies involving human PA monitored through wearable accelerometers.

Results: After screening 1312 articles, 428 (32.62%) eligible studies were identified and categorized into at least 1 of the following 3 thematic categories: classification (75/428, 17.5%), association (342/428, 79.9%), and prediction (32/428, 7.5%). Most articles (414/428, 96.7%) derived PA variables from 3D acceleration, rather than 1D acceleration. All eligible articles (428/428, 100%) considered PA metrics represented in the time domain, while a small fraction (16/428, 3.7%) also considered PA metrics in the frequency domain. The number of studies evaluating the influence of PA on health conditions has increased greatly. Among the studies in our review, regression-type models were the most prevalent (373/428, 87.1%). The machine learning approach for classification research is also gaining popularity (32/75, 43%). In addition to summary statistics of PA, several recent studies used tools to incorporate PA trajectories and account for temporal patterns, including longitudinal data analysis with repeated PA measurements and functional data analysis with PA as a continuum for time-varying association (68/428, 15.9%).

Conclusions: Summary metrics can quickly provide descriptions of the strength, frequency, and duration of individuals' overall PA. When the distribution and profile of PA need to be evaluated or detected, considering PA metrics as longitudinal or functional data can provide detailed information and improve the understanding of the role PA plays in health. Depending on the research goal, appropriate analytical tools can ensure the reliability of the scientific findings.

(J Med Internet Res 2024;26:e59497) doi:10.2196/59497

XSL•FO
RenderX

KEYWORDS

accelerometer; association; behavioral study; classification; digital biomarkers; digital health; physical activity; prediction; statistical method; wearable

Introduction

Physical Activity

Physical activity (PA), defined as any body movement that results in energy expenditure [1], has been acknowledged as a crucial factor in health management and disease prevention [1-3]. The energy expenditure due to moderate and vigorous PAs, such as brisk walking and running, is greater than that due to sedentary behaviors such as rest. These moderate and vigorous PAs are beneficial for health [1]. Many studies, including clinical trials, have focused on their effectiveness in treating and preventing chronic diseases, as well as promoting general health, well-being, and longevity [4-8]. Information about PA can be collected through questionnaires and digital devices. While questionnaires may suffer from recall bias, digital devices offer a more objective evaluation. The quantified summary statistics of the values recorded by these devices are often called digital biomarkers [9]. In a scoping review of 31 systematic reviews of clinical evidence involving digital biomarkers, wearables were the most commonly applied measurement tool (22/31, 71%), and PA was the most prevalent intervention (16/31, 52%) [4].

Wearable Devices

Wearable devices can continuously monitor an individual's free-living PA patterns. With the growing emphasis on health consciousness, there has been a notable surge in studies focused on PA, driven by the popularity of activity-monitoring wearables [3]. Earlier studies commonly used uniaxial accelerometer devices that measure only vertical axis acceleration in *g*-units, corresponding to the acceleration due to gravity (9.81 m/s^2) [10-13]. However, in recent decades, triaxial accelerometers operating in 3 orthogonal dimensions (a_x, a_y, and a_z) have gained preference due to their ability to offer a broad coverage, thus providing a more comprehensive understanding of overall human activity [14]. Among triaxial accelerometers, wrist-worn wearables are predominant, as devices worn on other body parts such as waists, hips, or thighs may cause inconvenience during movement [3]. Table 1 lists only triaxial wearable accelerometers used in >1 article reviewed here, as well as the placement positions of each device. More than half of them are research-grade wearables (eg, ActiGraph [ActiGraph, LLC], GENEActiv [ActivInsights], Axivity [Axivity Ltd], and RT3 [Stayhealthy, Inc]), while some are commercial wearables (eg, Fitbit Watch [Fitbit Inc]). This review will not discuss the comparison of different devices. Readers interested in the practical use, reliability, and validity of accelerometers are referred to published reviews [15-20].

Table 1. Triaxial wearable accelerometers that appeared in >1 reviewed article among the 428 articles.

Brand and model	Frequency (Hz)	Software	Manufacturer	References, n (%)
ActiGraph GT3X, GT9X, wGT3X-BT, wGT3X+, GT3X+, and ActiTrainer	30-100	ActiLife research	ActiGraph, LLC	180 (42.1)
GENEActiv original	10-100	GENEActiv PC	ActivInsight	24 (5.6)
Axivity AX3	12.5-3200	OMGUI	Axivity Ltd	37 (8.6)
RT3	0.017-1	RT3 ASSIST	Stayhealthy, Inc	2 (0.5)
activPAL activPAL3 and activePAL4	20-80	activPAL software	PAL Technologies Ltd	7 (1.6)
Actical	32	Actical software	Mini-Mitter, Respironics	30 (7)
Actiwatch	0.25-32	Actiware	Philips Respironics	5 (1.2)
Xiaomi Mi Band 2	__[a]	Master for Mi Band app	Xiaomi Corp	2 (0.5)
TracmorD	0.3-20	DirectLife Activity Monitor software	Philips DirectLife	5 (1.2)
Hookie AM20	100	—	Traxmeet Ltd	6 (1.4)
UKK AM30 and RM42	100	—	UKK Institute	3 (0.7)
Kenz Lifecorder EX	32	Lifecorder transfer software	Suzuken Co, Ltd	3 (0.7)
Shimmer3	1-512	ShimmerCapture	Shimmer Research Ltd	2 (0.7)
Fitbit	1-256	Fitabase server	Fitbit Inc	6 (1.4)

[a]Not available.

Such individual-centered PA information is usually collected by wearables at a frequency of 30 to 100 observations per second (Hz). Most devices do not have built-in software to output the collected data. However, several external programs such as ActiLife and GGIR [21] are available for analysis. When each person is monitored for multiple days in a free-living

XSL•FO
RenderX

environment, the data volume becomes enormous, leading to difficulties in analysis. Various digital variables that summarize daily activities have been proposed. Some are good for representing the total intensity, such as the summation or average of activity intensity (AI), and some are appropriate for the overall heterogeneity, such as variation and range, in either the time domain or the frequency domain [22]. These summary variables, however, do not catch the temporal patterns inherent in individual PAs. Other variables containing time information have been proposed and implemented in analyses. Part of this review will focus on these digital features.

Previous reviews have often emphasized the applications of wearables in health studies, focusing on the validity and reliability of these devices in evaluating PA and their use in clinical research [23-26]. While these reviews are comprehensive in terms of the practical use of PA metrics as digital data, they do not address the evolving trends in the analytical methods used to analyze such data. With the prevalence of wearables and the exponential increase in PA data volume, scientists now have the capacity to describe daily PA in a more sophisticated manner and examine whether PA trajectories vary across conditions or groups. Some advanced analytical tools have been proposed to address these tasks and will be discussed here.

In this review, we first describe the metrics currently used to summarize daily activities, particularly those transformed based on 3 acceleration axes. Common statistics derived from single-axis data, either in the time or frequency domain, are presented in Textbox 1 and will not be further discussed here. Second, we review the application of innovative PA metrics in statistical models aimed at different research goals, including the classification of activities, examination of the association of PA with health behavior or outcomes, and prediction of future events. Third, we discuss the limitations of current tools and propose directions for future analysis.

Textbox 1. Summary metrics and features calculated in time or frequency domain.

Domains, metrics, and features

- Time, uniaxial: mean, variance, SD, percentile, minimum, maximum, range
- Time, between 2 uniaxials: correlation between 2 axials, covariance between 2 axials
- Frequency: dominant frequency, peak power, spectral energy, entropy, and spectrum centroid

Methods

Protocol and Registration

The methods used in this review align with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist shown in Multimedia Appendix 1 [27]. The Arksey and O'Malley [28] framework was used after making methodological adaptations for this review. Protocols used in this scoping review are detailed in the following sections.

Eligibility Criteria

The eligibility criteria for article reviews have been established (Multimedia Appendix 1). The inclusion criteria were as follows: (1) the PA data reported on were obtained from human participants (adults or children, with disease or healthy) through a wearable accelerometer and (2) the purpose of the study containing the analysis of accelerometer data was to capture human activity patterns and influence on health.

The exclusion criteria included (1) articles not written in English; (2) articles incorporating other physiological activities detected by electrical signals (eg, electrocardiography); (3) articles conducting only statistical testing or containing no statistical model; (4) articles not focusing on human activity patterns; (5) articles not of original research, including systematic reviews, scoping reviews, books, and book chapters; (6) articles focusing on the reliability of wearable devices; and (7) articles addressing the upstream preprocessing of accelerometer data or measurement design (eg, the choice of epoch length of acceleration data or the placement of devices).

Information Sources

To collect the potentially relevant articles, the search was conducted in each of the following 3 publication databases: PubMed, IEEE Xplore, and ACM Digital Library. The final search was performed on February 29, 2024. As the main objective of this scoping review was to provide an overview of analytical methods used for the analysis of PA data in recent years, the search was constrained to articles published between 2000 and 2024. The full search strategies used in the 3 databases can be found in Multimedia Appendix 2.

Results

Selection of Evidence Sources and Data Extraction

After removing duplicate articles from the initial 1312 articles obtained by searching the 3 databases, screening for study goals based on title and abstract, and assessing eligibility (Figure 1), a final set of 428 articles was included in this review (Multimedia Appendix 3 [5,13,29-454]). Of them, 407 (95.1%) articles were from PubMed, 20 (4.7%) articles were from IEEE Xplore, and 1 (0.2%) article was from the ACM Digital Library. A total of 2 independent reviewers examined the content of each paper to determine which of the 3 categories of studies (activity classification, association, and prediction) it fit into and recorded the analytical tools used. Any inconsistencies in judgment were discussed and resolved through consensus. Following this, other study variables such as data domain, PA metrics, and summary statistics were examined.

The extracted data included brands and other details of the accelerometers used, research focus, context, analytic methods, and outcome variables. The extraction process was conducted

XSL•FO
RenderX

using a Microsoft Excel (Microsoft Corp) sheet designed by the reviewer YTL. This sheet consolidated the context of each article, capturing information on the brand of the device, PA variables, research issues addressed, activity features examined, statistical methodologies applied, and study outcomes.

Figure 1. Flowchart of the selection strategy.

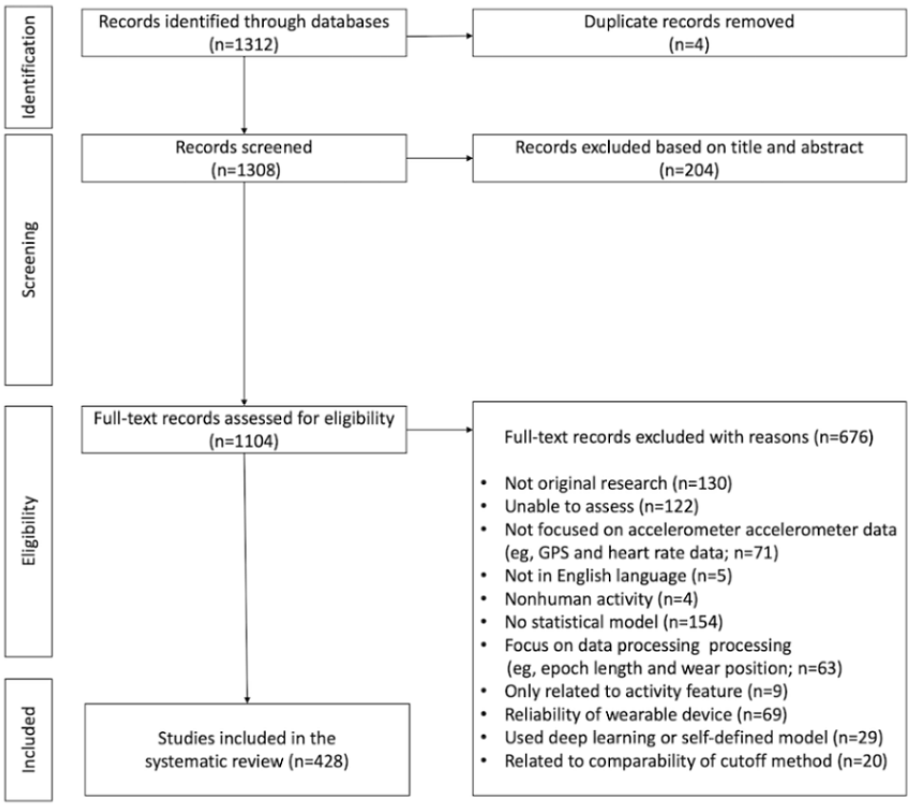

Measurements From Accelerometers

Most wearable accelerometers measure acceleration in 3 dimensions (a_x, a_y, a_z), with some also equipped with gyroscopes to assess body angular rates in degrees per second across the 3 axes. The reliability and accuracy of various brands, including ActiGraph, GENEActive, and Axtivity, have been compared and reviewed [15-20]. Because acceleration measurements are more common than gyroscope measurements in most devices, they have become the primary variable for evaluating body movement.

These accelerometer devices provide either raw accelerations in 3 dimensions or a combined metric as activity count (AC) per epoch produced by the manufacturer's proprietary software (Table 1) or by other free programs such as GGIR [21]. Raw accelerations have been favored over AC for 3 reasons. First, it is uncertain whether the definition of AC across brands is the same. Second, the development of free software such as GGIR [21] has relieved the computational difficulty and burden on researchers of analyzing acceleration measurements. Third, providing raw acceleration data makes them available for reuse in other types of studies.

Accelerations in 3 dimensions can be analyzed either separately or after being transformed into a single variable, as shown in Figure 2. Both uniaxial acceleration and transformed triaxial metric data can be used in analysis to classify activity types, explore association, and predict health outcomes. Figure 2 outlines the procedures from the collection of raw data to analysis of uniaxial or triaxial acceleration data. These analyses can be categorized further, as displayed in Figure 3. In both the time and frequency domains, 29 (6.8%; 29 uniaxial in time + 13 uniaxial in frequency – 13 duplicates) of 428 studies examined the uniaxial acceleration separately for its simplicity and straightforwardness. A total of 340 (79.4%; 74 transform triaxial in time + 4 transform triaxial in frequency + 275 summary of transform triaxial in time – 13 duplicates) of 428 studies used a triaxial acceleration metric combining the acceleration data from 3 dimensions in the time domain or frequency domain, or used a summary-type metric based on these transformations. These include the length of the acceleration vector (a_x, a_y, a_z), functions of the vector, amplitudes of Fourier-transformed acceleration waves, and summary statistics of these metrics. Analyses based on transformed variables have become prevalent in recent years because they combine information from all 3 acceleration axes. We refer to the variables transformed from the 3 raw accelerations as *transformed metrics*, which can be represented in the time domain or frequency domain. Some articles (80/428, 18.7%) used indirect acceleration metrics such as step count, AC, and walking time.

XSL•FO
RenderX

Figure 2. Outline of procedures from the collection of raw data to the application of analytical methods. PA: physical activity; ROC: receiver operating characteristic; VM: vector magnitude.

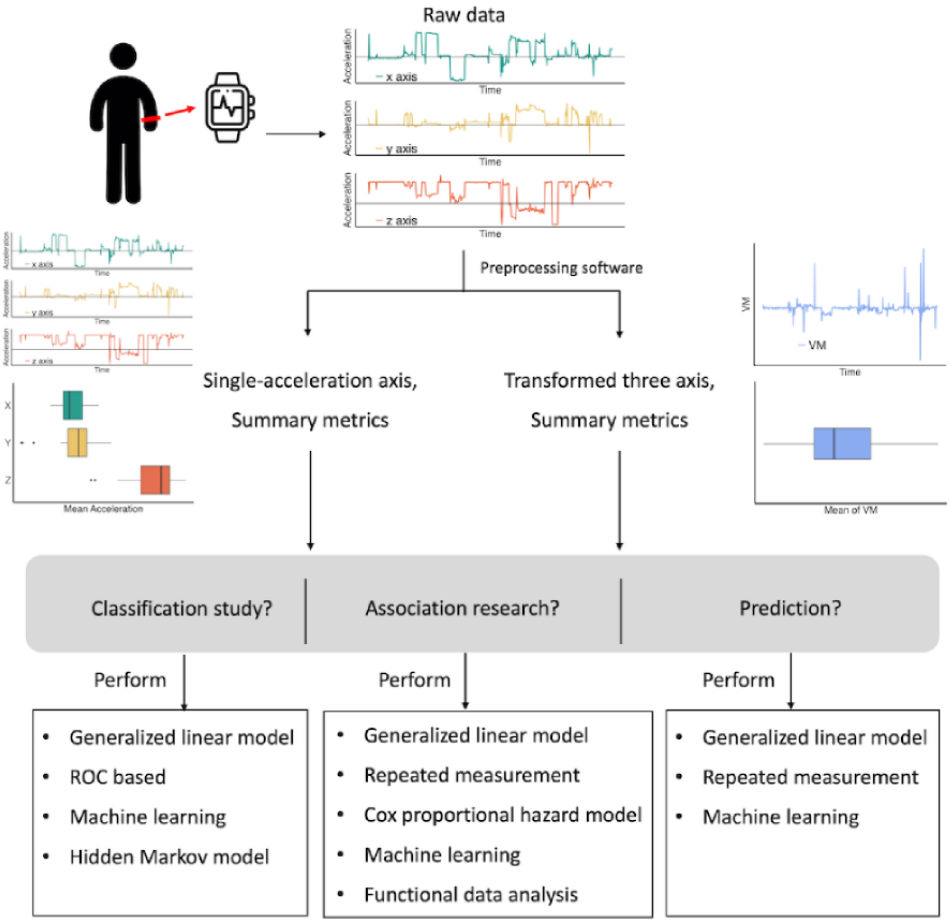

Figure 3. The left column lists the number of articles that used the corresponding physical activity (PA) metrics in analysis, the middle lists the number of articles focusing on each topic, and the right column lists the analytical tools adopted in studies. Note that an article may be counted more than once if it qualified >1 criterion. ROC: receiver operating characteristic.

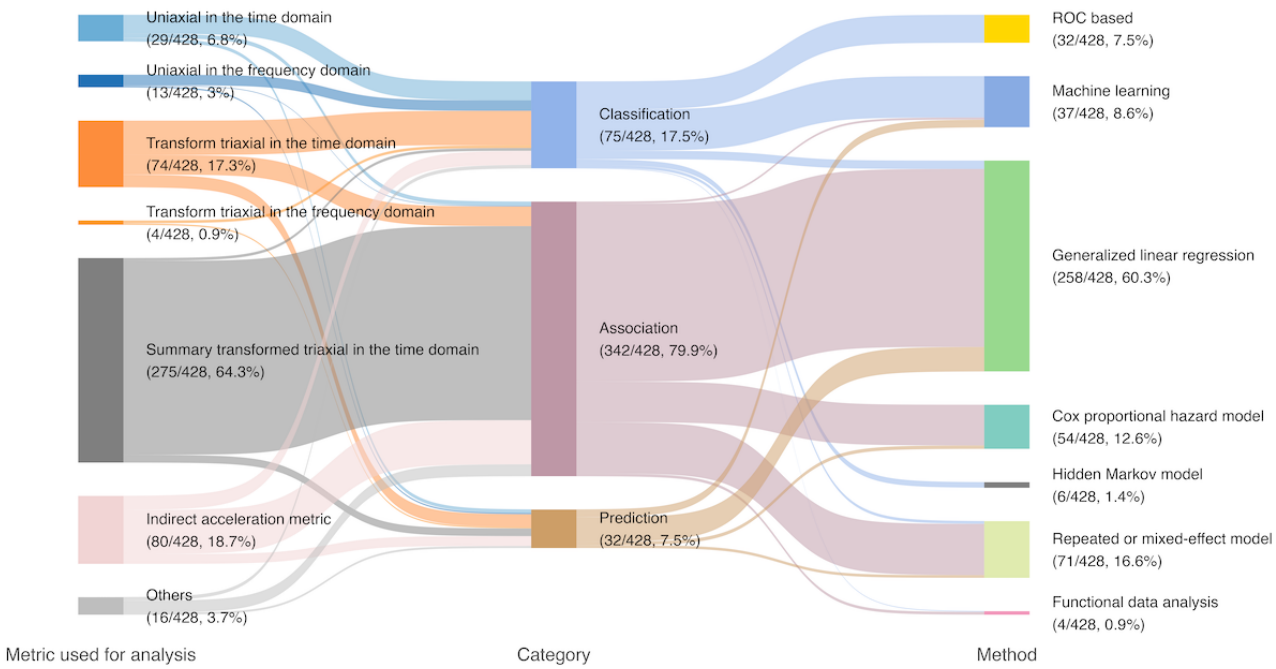

XSL•FO
RenderX

Metrics in Analysis: Transformed Metrics in the Time Domain

Several transformed metrics have been developed as functions of the 3 accelerations (a_x, a_y, a_z), including the Euclidean norm minus one (ENMO), signal magnitude area, angle-z, and PA index (Table 2). These metrics either focus on the vector magnitude (VM) of the acceleration (magnitude type) or the variability in the acceleration. For instance, the first developed metric was the straightforward VM based on the 3 uniaxial accelerations at time t:

$$\mathrm{VM} = [\, a_x(t)^2 + a_y(t)^2 + a_z(t)^2 \,]^{1/2} \quad \textbf{(1)}$$

Table 2. Transformed metrics in time domain and their summary quantities calculated from 3 accelerations (a_x, a_y, a_z).

Metric, feature, and summary	Definition	Reference
ENMO[a]	• $\max\{[\, a_x(t)^2 + a_y(t)^2 + a_z(t)^2]^{1/2} - 1, 0\}$	[455]
Angle-z	• $\tan^{-1}\{\, a_z(t) / [\, a_x(t)^2 + a_y(t)^2]^{1/2}\} \times 180 / \pi$	[8]
PAI or AI[b]	• $\mathrm{AI}(t,H) = [\max\{\, (\mathrm{var}(a_x(t,H)) + \mathrm{var}(a_y(t,H)) + \mathrm{var}(a_z(t,H)) - \mathrm{sum.var}) / 3, 0\}]^{1/2}$ • $\mathrm{sum.var} = \mathrm{var}(a_x) + \mathrm{var}(a_y) + \mathrm{var}(a_z)$	[456]
SMA[c]	• sum of integration of $\lvert a_x(t)\rvert$, $\lvert a_y(t)\rvert$ and $\lvert a_z(t)\rvert$ from $t = T$ to $T + 1$ seconds	[29]
VM or SVM[d]	• $[\, a_x(t)^2 + a_y(t)^2 + a_z(t)^2 \,]^{1/2}$	[30,31]
MVPA[e]	• number of times when the transformed metrics > threshold	[34-36]
L5[f]	• The means of the 5 least active hours during the day	[457]
M10[g]	• The means of the 10 most active hours during the day	[457]
IV[h]	• $\{\text{mean of } (x_h - \mathrm{mean}(x_h))^2 \text{ from } h = 1 \text{ to } p\} / \{\text{mean of } (x_i - \mathrm{mean}(x_i))^2 \text{ from } i = 1 \text{ to } N\}$ • N: the total number of data points • p: the number of data points per day • x_h: values of each hour from the mean 24-hour profile • x_i: each given hour of raw data	[457]
IS[i]	• $\{\text{mean of } (x_i - x_{i-1})^2 \text{ from } i = 1 \text{ to } N\} / \{\text{mean of } (x_i - \mathrm{mean}(x_i))^2 \text{ from } i = 1 \text{ to } N\}$ • N: the total number of data points • x_i: the value of a given hour	[457]
Acrophase	• The time at which the peak of a rhythm occurs	[458]

[a]ENMO: Euclidean norm minus one.
[b]PAI or AI: physical activity index or activity intensity.
[c]SMA: signal magnitude area.
[d]SVM: vector magnitude or signal vector magnitude.
[e]MVPA: moderate-to-vigorous physical activity.
[f]L5: least active 5-hour period.
[g]M10: most active 10-hour period.
[h]IV: intradaily variability.
[i]IS: interdaily stability.

The second developed metric, ENMO, is intuitive because it combines the strength of acceleration in gravity in the 3 axes, where the effect of gravity is removed from the formula by subtracting 1 [455]:

$$\mathrm{ENMO} = \max\{[\, a_x(t)^2 + a_y(t)^2 + a_z(t)^2 \,]^{1/2} - 1, 0\} \quad \textbf{(2)}$$

Its calculation can be performed with the R (R Foundation) package *GGIR* [21]. ENMO does not depend on the sampling frequency and is noted for its higher sensitivity to vertical and horizontal accelerations [455]. It has been widely used in studies for classifying specific movements and sleep stages [459,460]. Studies have showed that ENMO performs better at indicating moderate to heavy activities but is less robust at indicating sedentary and light activities compared to other measurements [456].

Other metrics similar to ENMO in calculating the “average” or “degree” of accelerations are signal magnitude area [29] and signal VM [30,31]. These metrics are all representations of the

XSL•FO
RenderX

average type. These metrics can be further transformed to imply certain levels of activity. For instance, the moderate-to-vigorous PA (MVPA) measure contains cutoff values for ENMO to determine whether the activity is light, moderate, or vigorous [32,33]. In studies where the classification of an individual's activity is of interest, the cutoff values need to be determined because the values may be person specific.

Another magnitude-type transformed metric is the heuristic algorithm looking at the distribution of change in z-angle proposed by van Hees et al [8]:

$$\text{Angle-z} = \tan^{-1}\{ a_z(t) / [a_x(t)^2 + a_y(t)^2]^{1/2}\} \times 180 / \pi \quad \textbf{(3)}$$

As the name suggests, this transformed metric emphasizes the movement in the direction vertical to the x-y plane, and the use of a_z increases the sensitivity for the detection of activities along this direction [460]. The heuristic algorithm looking at the distribution of change in z-angle was originally used to detect the sleep period time window. Its reliability has been demonstrated through comparisons with polysomnography [8], and it has been commonly used to investigate sleep conditions. For example, it was used by Jones et al [461,462] to differentiate sleep status, and it was combined with genomic data to explore significant sleep factors.

In contrast, PA index (PAI), also known as AI, examines the variability of accelerations, rather than their magnitude. The PAI metric combines the individual-specific variation in each dimension in a time interval of length H starting at time t as a summary metric [456]:

$$\text{PAI}(t,H) = [\max\{ (\text{var}(a_x(t,H)) + \text{var}(a_y(t,H)) + \text{var}(a_z(t,H)) - \text{sum.var}) / 3, 0\}]^{1/2} \quad \textbf{(4)}$$

Here, the normalization quantity sum.var = var(a_x) + var(a_y) + var(a_z), summation of 3 variances, is the systematic noise variance that can be computed from raw accelerations when the device is not moving. This PAI can be further modified by dividing the normalization quantity, becoming the relative AI, as opposed to the original version in the absolute scale. This AI is advantageous because of its additive and rotational invariance properties, indicating that PAI is a constant when an individual performs the same activity during the period. PAI can be applied in, among other areas, the surgical recovery process of women who have undergone mastectomies and evaluation of physical intensity during powered wheelchair mobility in people with severe dyskinetic cerebral palsy [463,464].

Metrics in Analysis: Metrics Summarizing PA in the Time Domain

Metrics such as those discussed earlier are all computed in the time domain. In other words, they preserve time information and can be used further to decipher longitudinal patterns. These repeated time-series measurements can be easily included in the generalized linear mixed-effects models (GLMMs) for additional analysis, or they can be summarized or categorized to simplify the time effect. One such summary statistic is the threshold that differentiates light activity from MVPA, with the time spent in MVPA also considered as a summary PA feature. Examples of MVPA include dancing, gardening, walking briskly, jogging, and running. To be categorized in this group, an activity usually needs to reach a certain intensity evaluated by metabolic equivalent of task (MET), such as the maximum or peak oxygen uptake or heart rate [34-36]. For example, one unit of MET is the amount of oxygen consumed or energy used per kg per minute while sitting at rest. Activities corresponding to MET values falling within a specified interval are categorized as MVPAs. The boundaries or thresholds of this interval need to be determined first. The concept of MVPA is easy to understand and often referenced when exercise is prescribed to people for health benefits; the definition of its thresholds, however, can be individual specific because the thresholds are related to individualized homeostasis disturbance [465].

Other summary metrics compare PA data within the same person to indicate a person's rest-activity rhythm, including nonparametric circadian rhythm measures such as interdaily stability (IS), intradaily variability (IV), least active 5-hour period, most active 10-hour period, and the acrophase in cosinor analysis [458]. Both IS and IV are useful in describing rest and activity rhythms: IS provides information on how stable the rest-activity pattern is between different days, and IV quantifies the stability of the fragmentation in the activity pattern between time intervals [457].

These summary metrics effectively capture and quantify the characteristics of AI and circadian rhythms, enhancing the understanding of their relation to health and behavior [37,462]. However, defining thresholds for PA intensity can be both population and person specific [465]. In addition, these summary statistics may suffer from information loss because daily PA is a continuum [38]. They focus on activities in fragments of the day rather than the overall PA profile throughout the day.

Metrics in Analysis: Metrics Summarizing Temporal Patterns in the Frequency Domain

One way to explore daily continuous temporal trends and trajectories is via the Fourier transformation, allowing metrics to be decomposed into periodic curves expressed by both frequency and amplitude. These longitudinal waves in the frequency domain are threshold free. For instance, ENMO, angle-z, and AI can be represented in the frequency domain to examine their patterns over time. Among the 428 reviewed articles, this frequency representation was considered in 4 (0.9%) articles for triaxial accelerations, while 13 (3%) adopted frequency domain representation in uniaxial acceleration.

We use $x(t)$ to denote the transformed metric in the time domain, where t indexes the time. For example, $x(0)$, $x(1)$,..., $x(t)$,..., and $x(T_1)$ represent the ENMO value observed throughout the day. Then, these $x(t)$s can be converted to $X^F(0)$, $X^F(1)$, $X^F(2)$,..., and $X^F(T_1)$, where each one is a combination of sinusoid functions. In other words, these $X^F(k)$s are all periodic curves with frequency k and amplitude $||X^F(k)||$. These new curves are in the same unit as the original $x(t)$, and in the case of ENMO, the unit is g:

$$X^F(k) = \text{sum}\{x(t) \times \exp(-2\pi ikt / T)\} \text{ over } t \text{ from } 0 \text{ to } T-1 \quad \textbf{(5)}$$

XSL•FO
RenderX

For example, $X^F(2)$ implies that this individual performed an activity of amplitude $||X^F(2)||$ with frequency 2. Usually, the waves with large strength indicate dominant activities in a day, while the waves with low frequencies can imply regular rhythms such as commuting to and from work or napping habits. The computation of this transformation can be made easy with the fast Fourier transformation (FFT) in R with the package *stats*.

The advantages of the FFT are 3-fold. First, it simplifies the interpretation of periodic activities using frequency and strength. Second, as the PA curve is dominated by a few low-frequency waves, each curve can be represented with a small number of waves, significantly reducing data dimension. Third, these dominant waves can be considered as digital features of an individual's PA and incorporated into further analyses, such as classification [39,40].

The FFT-based variables are useful for periodic activities, especially when activity metrics from multiple days are collected to establish an individual's daily cycles. However, the FFT approach can be inappropriate if the original PA lacks periodic patterns. An alternative way would be to relax this constraint when summarizing the PA trajectory.

Metrics in Analysis: Metrics of Temporal Patterns in the Time Domain

An alternative threshold free approach to capture the temporal dynamics inherent in accelerometer-based measures involves considering these measures as curves within infinite-dimensional spaces, treating these collected measures as functional data. Functional data analysis (FDA) has emerged as a powerful tool for analyzing PA data, enabling the continuous variation of PA patterns throughout the monitoring period, without the need to assume cycles. Furthermore, considering PA as a function of time facilitates the assessment of time-varying effects of covariates such as BMI, expanded disability status scale, and lung function [38,41,42,466-469]. Among the 428 screened articles, 4 (0.9%) used such features in PA analysis.

In FDA, the observed transformed metrics in a day from the same individual, $x(0)$, $x(1)$,..., $x(t)$,..., and $x(T1)$ are considered the discretized observations from an unknown continuous curve, $Y(t)$. The patterns inherent among the multiple curves, $Y(t)$s, from multiple individuals can be explored via functional principal component analysis (FPCA), an analogy to principal component analysis for multivariate data [470,471]. These extracted principal component functions are called eigenfunctions and can indicate the decomposed daily temporal patterns. Each of the original curves is now associated with these eigenfunctions through a score. For instance, the i-th score of a daily curve now corresponds to its representation along with the i-th function. A large positive score indicates greater similarity in pattern and magnitude with this function. In other words, these scores become indications of the temporal trajectories.

In addition, similar to FFT, the top few principal component functions in FPCA can explain most of the variability in the data. This greatly reduces the original data size to a few dominant functions, which can describe the major activity patterns in the population. In other words, the original data set can be approximated with only a few functions and scores with this FPCA approach. FPCA has previously been used in modeling yearly temperature curves and constructing growth curves of children; however, few studies have considered the application of FPCA and FDA in PA studies [43,469]. Another group of articles incorporated within-person and between-person heterogeneity into their analysis using 2-level FPCA or multilevel FPCA [470,472]. One level of scores represents the person-to-person variation, while the second level represents the day-to-day variation [470,472]. FDA involves smoothing techniques, FPCA, functional linear regression, and clustering or classification methods for analyzing data observed over continuous domains. It has been adopted mostly in classification research, which we will discuss in the next sections.

Activity Classification Studies With Threshold-Based Methods

Most early studies involving PA measured by accelerometers were classification studies that classified activity types or intensities, such as sleep-wake or moderate or vigorous activities. Of the 428 reviewed studies, 75 (17.5%) fall into this category (middle panel in Figure 3 and Multimedia Appendix 3). The main 2 topics in these classification studies are the classification of specific activities and sleep-wake stages. Specific activities for classification include daily activities (sitting, walking, and lying) [13,44,45], housework (washing dishes and dust mopping) [46,456], fall detection [473,474], and rehabilitation activities [475,476], while sleep stages for classification include wake, light sleep, and deep sleep stages [8,460,477]. Each activity type corresponds to a range of AI that must be determined first. In other words, the threshold values of these intensity intervals are necessary to classify activities. Once the cutoff values of the PA metrics are selected, several models are available for classification. Among the 75 classification studies reviewed, researchers opted for the threshold-based method (32/75, 43%), machine learning (ML; 32/75, 43%), regression (11/75, 15%), and the hidden Markov model (HMM; 6/75, 8%). These numbers are indicated above the edges connecting the corresponding nodes in Figure 4.

Threshold-based methods are commonly used to identify activities with different intensities after specific cutoff points of PA metrics are determined. These cutoff values are often established by the receiver operating characteristic (ROC) curves based on metrics transformed from the triaxial accelerations and aggregated over a specific period, such as a week. The AI values are then categorized into multiple levels or ranges. For instance, this approach was applied in studies monitoring the daily behaviors of children who were instructed to engage in specific activities [47-49]. The intensity level of each activity is determined based on its characteristics, followed by the use of an ROC curve to evaluate the differentiation between target behaviors.

Threshold-based classification offers the advantage of easy calculation and interpretation. However, this method may lack robustness because AI levels can be affected by various factors, such as the placement of the device on the body and differences among populations, and are highly dependent on the individual, limiting their generalizability [50,51,465]. If a large amount of

XSL•FO
RenderX

data is collected from each individual during specific activities, individual-specific thresholds can be established and used to classify this individual's behavior. This generalization can be particularly useful when considering PA as a form of treatment.

Figure 4. Network association plot of the articles. The number inside the node indicates the number of articles, and the number by the edge indicates the number of articles containing terms in both connecting nodes. ROC: receiver operating characteristic.

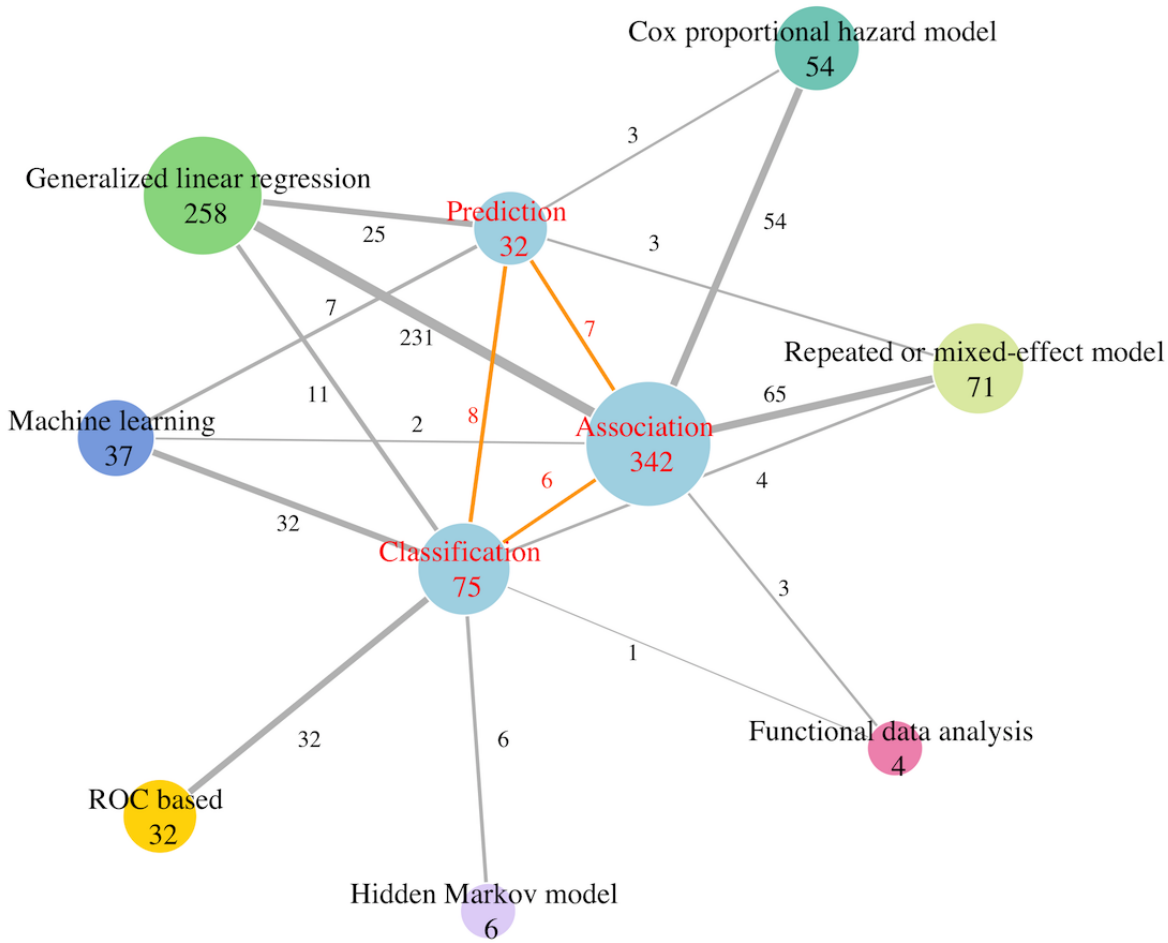

Activity Classification Studies With Regression Models

Rather than determining thresholds using ROC curves, several studies opt for linear regression models [44,50,51]. Such models allow for individualized threshold values, selected to correspond to a predetermined energy expenditure through a linear relationship established in the regression model. These values are then used in classification tasks, particularly beneficial for those considering PA for older adults or individuals with injuries or diseases [44,50,51]. The rationale behind this design is that the range of physical intensity may differ in populations with specific conditions, necessitating the determination of thresholds tailored to their characteristics. Therefore, these studies first apply linear regression to establish the threshold values between light activity and MVPA in settings such as wheelchair use or treadmill exercise and subsequently evaluate classification using these derived threshold values.

The use of the study-specific thresholds for PA metrics helps mitigate the influence of sampling variation and allows for the incorporation of confounding factors such as age, gender, and health status when establishing the cutoff values. Consequently, the conclusions drawn from such studies are more reliable and applicable to research involving similar populations. However, conducting this type of study requires a larger sample size and repeated measurements of PA to establish, calibrate, and validate the threshold values.

Activity Classification Studies With ML Models

To avoid the determination of thresholds a priori, other studies have formulated activity classification as a task of activity recognition and resolved this task with supervised ML approaches [13,44,52,53,478-480] or unsupervised methods such as those using k-means and Gaussian mixture models [54]. Almost half (32/75, 43%) of the classification studies reviewed here adopted the ML approach (Figure 4). ML tools, including decision trees, random forest, logistic regression, naive Bayes, and multilayer neural networks, have been used to perform recognition among walking, jogging, going upstairs, going downstairs, sitting, standing, and lying at rest [13,52]. Except for 2 (2/32, 6%) studies based on the National Health and Nutrition Examination Survey [54,317], 2 (6%) on children or adults [53,480], 2 (6%) on older adults (aged >60 and >70 years, respectively [55,56]), and 2 (6%) on participants with diseases (spinal cord injury [51] and cognitive impairment [57]), the

XSL•FO
RenderX

remaining articles (24/32, 75%) either used public data sets, such as the experimental data collected during the DARPA'98 Intrusion Detection Evaluation program [478], or recruited convenient samples, such as community residents aged between 20 and 89 years [52], to investigate whether ML algorithms can differentiate between activity types and intensities across groups. In addition, in 7 (22%) out of the 32 articles, participants were asked to wear the device for 1 to 14 days to record their free-living activities, and the data were used to cluster activity types, intensities, or sleep-wake stages. In another 23 (72%) articles, participants were asked to perform specific movements for 30 to 60 minutes, and the data were used to differentiate specific activities.

The performance of these tools was compared using criteria such as F_1-scores, confusion matrices, and accuracy rates. No consistent conclusion of the best tool could be drawn. A logistic regression model with or without cross-validation offered the best performance in some studies [478,480], while multilayer neural networks performed the best in others [479].

Several factors may have affected the inconsistency in performance. First, the sample sizes in these studies were not large, potentially resulting in greater sampling variation and susceptibility to individual variability. Moreover, the ML approaches typically assumed independence among activity data for each epoch, overlooking potential correlations that may have existed. Such interdependencies and associations within the data can be essential when tackling classification problems [481]. In addition, it is worth noting that when identifying specific activities, using individual's directional accelerations may yield better results compared to using the transformed metric of the triaxial accelerations [482].

Activity Classification Studies With HMMs

Another classification approach that does not rely on predetermined thresholds is the HMM. Widely used in analyzing time series or correlated data, the HMM originated in the field of speech recognition, where it demonstrated effectiveness in handling sequences of feature vectors. On the basis of augmenting the Markov chain, the fundamental assumption of the HMM is that the observed events depend on hidden states that are not directly observable. A typical HMM consists of essential components, including distinct observations, hidden states, a state transition model, emission probabilities, and an initial state distribution. These components collectively contribute to the modeling and analysis of underlying processes captured by the HMM [29,58-60,483,484].

The application steps for the HMM can be summarized as follows. The first step involves conducting a study to collect labeled data consisting of accelerometer counts and their corresponding activities. These labeled data are then used to estimate the parameters in the HMM classification model. Next, the trained model is applied to unlabeled activity data to estimate the activity type with a probability [29,58-60,483,484]. In the context of activity recognition, the HMM is constructed in 2 different ways. The first one involves a single HMM model for all types of activities to be classified [29,58-60,483,484]. The second one builds an HMM for each specific activity, taking each activity as a distinct entity [483]. These 2 approaches represent different strategies in activity recognition.

The use of the HMM comes with both advantages and limitations. One of the advantages is its ability to leverage the inherent autocorrelation in temporally close activities, which enhances estimation strength in classification tasks [59]. Consequently, the HMM tends to achieve a higher accuracy in classifying types of activities compared to threshold-based methods [484]. Second, HMM provides an estimated probability for each activity label, offering a quantification of uncertainty in estimation. However, parametric HMMs can be computationally expensive due to the large number of parameters involved. In addition, a substantial amount of training data is typically required to ensure stability in parameter estimation. Furthermore, determining the optimal hyperparameter settings before training the model can pose challenges [485].

Studies on the Association Between PA and Health Conditions With Regression Models

In addition to classifying activities, there has been a growing emphasis on researching the relationship between PA and health outcomes. Figure 5 illustrates the distribution of articles over the years, with association research consistently comprising the majority (Figure 5). Specifically, association research accounts for 79.9% (342/428) of the reviewed literature (Figure 3 and Multimedia Appendix 3). The studied health outcomes encompass a wide range, including cardiovascular health [61,62], metabolic markers [486], mental well-being [63,64], chronic disease risk [5], and overall mortality [65,66]. By examining the relationship between PA and health outcomes, scientists can gain insights into the potential benefits and mechanisms through which PA impacts overall well-being and disease prevention. Such studies highlight the broader implications and significance of PA in promoting optimal health while controlling confounding factors.

When investigating and interpreting the association between PA and health conditions, regression models are commonly used as powerful tools. The results can provide evidence to identify protective or risk factors of health conditions. In addition, the regression approach can be directly generalized to health outcomes with either discrete or continuous values. Depending on whether the PA metrics are summary statistics over a period such a week or are included as time series in analysis, these regression models can be grouped into the ones using summary-type PA metrics as longitudinal PA and ones emphasizing PA temporal trajectories in FDA, as discussed in subsequent sections.

XSL•FO
RenderX

Figure 5. The value on the top of each bar is the sum of the number of articles per topic per year. The solid black line indicates the number of articles considering longitudinal patterns in physical activity (PA).

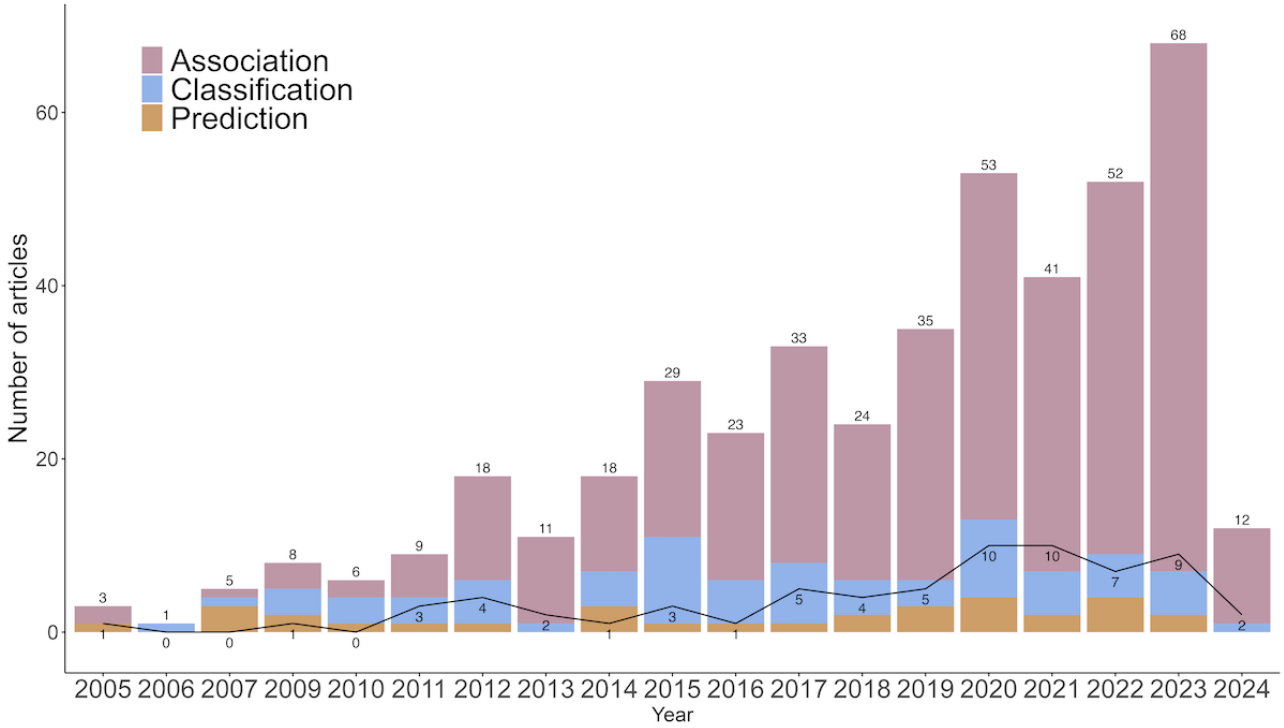

Regression Models With Summary PA Metrics

In regression models of this type, the primary aim is to evaluate the impact of overall PA on specific health outcomes. Therefore, summary PA metrics are frequently adopted as explanatory variables. One commonly used summary statistic is the time spent in MVPA per day or week. This metric represents the duration of activities with higher intensity. With this time of MVPA included as a predictor in regression models, studies have revealed beneficial effects of PA on various health conditions, such as healthy aging [487] and reduced frailty [67]. Similar predictors include the times spent in each of the different specified levels of activity (such as low PA, vigorous PA, and sedentary time). Because these variables are correlated with each other, multivariate pattern analysis to handle the multicollinearity was used to examine their association with metabolic health measures such as the cardiometabolic variables [68,69]. Another study used a quantile regression model with the Pittsburgh Fatigability Scale score as the outcome variable and considered the 24-hour rest-activity rhythm as a summary of PA to examine its association with physical fatigability in older adults [70].

Depending on the outcome variable, either a linear regression model or a generalized linear model can be included for analysis, with or without random effects for individual influence. When the health outcome is a time-to-event measurement, a Cox regression model would be used to assess the association. Cox regression models are used for survival analysis focusing on the estimated hazard ratios of predictor variables (eg, PA levels) and the time-to-event outcomes (eg, disease occurrence or mortality). Examples include the incidence of cardiovascular disease [71,72] and all-cause mortality [65,73]. By adjusting for covariates in the model, such as age, sex, smoking status, and comorbidities, researchers can better evaluate the impact of PA on the survival time of the outcome.

The regression models described earlier provide a direct interpretation of the relationship between overall PA patterns and health conditions. However, they do not account for temporal variations, such as longitudinal trends, as the variable can indicate PA changes across time. In other words, these models overlook the variability in time inherent in the original PA metrics such as ENMO, angle-z, and AI. To address this limitation, alternative regression models designed for longitudinal or functional data would offer a more suitable approach.

Regression Models With PA Metrics as Repeated Measurements

When the study objective is to examine the impact of PA temporal patterns on health outcomes, the identification of trends across time and the use of PA metrics indexed by time are essential. Extracting information from repeatedly measured PA is crucial for achieving this goal. Therefore, the modeling approach must account for correlation in the longitudinally measured metrics. Various statistical models for longitudinal data analysis have been developed to incorporate within-person variation, including repeated-measures ANOVA and mixed-effects models. These models were used in 65 (19%) out of the 342 association studies reviewed, as indicated by the line connecting these 2 nodes in Figure 4.

Repeated-measures ANOVA is a statistical method primarily used to analyze differences among ≥3 dependent samples, with the assumption that different individuals are independent from each other, while different visits or time points for the same individual exhibit correlation. For instance, in a school-based behavior intervention study, 2-way ANOVA was applied to

XSL•FO
RenderX

analyze the effect of treatment, time, and their interaction on various PA variables, such as the intensity, frequency, and duration of PA [488]. These PA variables were considered repeated measurements, as they were collected before and after the intervention. This study concluded that the intervention positively impacted the psychological constructs and the PA levels of the adolescents involved. Similarly, in a study on children's motor skills, repeated-measures multivariate ANOVA (MANOVA) was used to explore skill competence and PA [489].

The calculation of the covariance matrix in repeated-measures MANOVA assumes compound symmetry, implying that variances are equal across time points, and the correlations between any 2 time points are constant. However, this assumption may not hold true if the study spans a long period. Other requirements for repeated-measures MANOVA include complete cases, where no data are missing, and covariates are time-independent. Both conditions may be compromised due to factors such as loss to follow-up, device malfunction or misplacement, or variations in human PA rhythms.

The linear mixed-effects model and GLMM can accommodate continuous or discrete longitudinal response variables. The response variable can be the time spent in various levels of PA, in MVPAs across seasons, or in categories of sleep [74,75]. In a bidirectional association study on adolescents, the linear mixed-effects model was used to assess the effect of night sleep on the PA performed the following day, and the GLMM investigated the effect of daytime PA on night sleep [74]. Both models used the random intercept to account for the day-to-day association at the individual level while adjusting for demographic variables as the fixed effects. When modeling the within-person correlation, some studies chose to model via the correlation matrix, rather than the random effect. For instance, a generalized linear model was used in a cohort study of aging workers to assess the effect of neighborhood socioeconomic disadvantages and greenness on leisure-time PA [76]. The time spent was monitored during working and nonworking days. The within-person variation in the repeatedly measured PA time was included in the covariance structure, in contrast to the person-specific or cluster-specific random effect in mixed-effects models. The quantified association can then be estimated by generalized estimating equations, as demonstrated in studies of frailty and daily PA [77,490].

Regression Models With PA Metrics as Functions of Time

An alternative approach to modeling PA recognizes it as a continuum and considers the collected PA measurements as discretized observations from a continuous curve. This involves applying FDA in regression models to establish association. Unlike the amplitudes from the FFTs or the scores from FPCA discussed earlier, FDA-based regression models can capture time-varying relationships between 2 variables, both of which can be functions of time. The use of this approach has gained much attention (3, 0.9% out of 342 association studies in Figure 4). Functional regression models can be categorized into three groups, depending on whether the response or covariates are functional or vector data: (1) functional responses with functional covariates, (2) functional responses with scalar covariates, and (3) scalar responses with functional covariates [491].

Function-on-function regression is the model for a functional response involving ≥1 functional covariates [38]. The formula for this model can be expressed as follows:

$$y_i(t) = \beta_0(t) + \int \beta(t) x_i(t) dt + \varepsilon_i(t) \quad \textbf{(6)}$$

Here, the response, covariate, and coefficient are all time-varying functions. In a study investigating the association between PA and lung function, both the flow-volume curve (FVC) and PA of recruited smokers and nonsmokers were monitored through time [38]. The FVC, measured in mL, indicates the volume of air one can blow out after full inspiration. The FVC was denoted as $y_i(t)$, and the PA of the i-th person monitored through time was denoted as $x_i(t)$. This model allows researchers to investigate how the functional response changes in relation to ≥1 functional covariates, providing valuable insights into the complex associations between functional data.

Function-on-scalar regression (FoSR) allows for the analysis of the relationship between a functional response and scalar covariate. It enables researchers to examine how functional predictors, such as activity patterns, relate to scalar covariates, such as health indicators or disease risk [78,466,468,492]. The model equation of FoSR can be expressed as follows:

$$y_i(t) = \beta_0(t) + \beta_1(t) x_i + \varepsilon_i(t) \quad \textbf{(7)}$$

Here, the only scalar is the covariate x_i. FoSR was applied in studies investigating the relationships between 24-hour rest or activity rhythms (the response function y) and scalar covariates (x) such as clinical characteristics and demographics [42]. In addition, FoSR was used in a study of children's behavior [468] and a study of activity patterns in individuals with different mental statuses [466]. By leveraging FoSR, these studies evaluated the dynamic relationships between functional responses and scalar covariates over time, providing important findings for understanding various health and behavioral outcomes.

Scalar-on-function regression studies the relationship between a scalar response and functional covariates [493]:

$$y_i = \beta_0 + \int \beta(t) x_i(t) dt + \varepsilon_i \quad \textbf{(8)}$$

Different from the above-mentioned FoSR, the association of interest is now between the scalar response y_i and the functional covariate $x_i(t)$. For instance, the percentage of time spent at or above a certain intensity was used as a functional predictor $x_i(t)$ to investigate the influence of this PA profile on the same person's biological aging y_i, a scalar response [41].

The concept of scalar-on-function regression has been extended for types of accelerometer data analysis. For example, the scalar-on-time-by-distribution regression model was proposed to study the association between the scalar response and temporally local distributional information in wearable data [467]. In another study, researchers introduced the scalar-on-quantile-function regression model, which used the

XSL•FO
RenderX

user-specific quantile function as a predictor to analyze the distributional nature of accelerometer data [494].

Prediction of Health Conditions With PA Variables

Of the 428 studies reviewed, 32 (7.5%) were prediction studies (Multimedia Appendix 3), involving the prediction of health-related outcomes or types of PAs [79-82]. Prediction can be derived based on statistical regression models or ML approaches. Once the association between covariates and response is established with the regression models discussed in earlier sections, new responses can be predicted based on newly observed values of predictors [495]. To predict all-cause mortality, a study applied sparse partial least square regression as the ML tool to determine the importance of 21 PA and sedentary behavior features and to construct a composite score based on their relevance [83]. Similarly, PA measures such as the total AC and active-to-sedentary transition probability were identified in logistic regression models as strong stand-alone predictors of 5-year all-cause mortality based on the National Health and Nutrition Examination Survey study, and these PA features can improve the prediction performance of traditional risk factors [43]. The scores derived from FPCA on PA counts were also considered in predicting mortality, sleep efficiency, and decreases in cognition [43,471].

Prediction studies forecasting activity-related energy expenditure are also prevalent [496-499]. Many studies of this category have relied on the explained variance R^2 from linear Pearson correlation analysis or the nonlinear Spearman rank correlation to determine the interpretability of PA variables such as step counts per day or VM counts per day [496]. It is important to note that such predictions are often population specific [84,497-500], and caution is warranted when generalizing these results to children, adults, and patients.

Discussion

Principal Findings

There has been a notable increase in the use of accelerometer-based wearable devices to study the association between PA and health. These devices, worn in a free-living environment, offer objective and dynamic information about an individual's activity patterns, allowing researchers to delve into various health issues. Selecting appropriate activity metrics and analytical approaches is crucial once study aims are established. Our review indicates that these aims typically involve classification, association, and prediction, while commonly used PA metrics often comprise VM values. These metrics can be computed as summary statistics to indicate the overall information regarding PA frequency, intensity, and duration. Alternatively, these metrics can be constructed as longitudinal or functional data, offering insights into PA profiles and time-varying effects. This perspective has attracted attention in recent literature, where the emphasis has shifted toward predominant PA trajectories and temporal variation. We anticipate a continued surge in the use of accelerometer-based wearable devices for PA monitoring, alongside the development of more advanced analytical tools to tackle increasingly complex scientific inquiries.

Limitations

This review did not include comparison and validation studies between different brands of devices and various PA metrics. Such studies may be crucial in determining whether data collected from different sources can be merged and provide information on data calibration methods. The findings from these studies may be essential when integrating various databases is needed. However, this task extends beyond the current scope of this review. Future research could initiate by gathering studies that compare different brands and subsequently apply network meta-analysis methodologies if the number of studies is not small.

Studies about the validation of PA metrics are few because this requires a gold standard that may not be easy to derive or set up. Studies that relied on self-report questionnaires may draw criticisms of recall bias and lack of reliability. The comparison between different PA metrics is not straightforward either. One solution would be to wear multiple devices that record different metrics and examine their consistency. Nevertheless, as long as the same device is worn by all participants in the study and the number of participants is large, the analysis of PA metrics can still offer insights into the effect of PA on outcomes of interest.

Other limitations of this review pertain to the choice of keywords used to screen articles. Because "physical activity" was a required phrase, articles containing only the word "activity" were excluded. For instance, studies focusing on sleep quality variation or sleep disturbances monitored by wearables may not be included if they did not emphasize the pattern or intensity of PA. However, research using accelerometers to detect sleep quality and efficiency, particularly for evaluating the association between sleep and cognition functions or chronic diseases, has garnered significant interest in recent years [471,492,501,502]. In addition, this review did not categorize articles based on study designs. Further categorizing them into cross-sectional, observational, or clinical studies could provide insights into the necessity and application of the analytical tools adopted in these studies. Similarly, this review did not focus on the use of wearables in clinical settings. However, a quick comparison of the selected articles from 2005 to 2009 with those from 2022 to 2023 revealed that only 2 (15%) out of 13 studies in the former period used accelerometers in clinical settings, while 23 (19.7%) out of 117 studies in the latter period did so. This difference demonstrates the increasing acceptance and adoption of accelerometers within the clinical community.

Conclusions

PA has been shown to benefit human health and prevent diseases, and wearable accelerometers offer a cost-efficient means to monitor an individual's daily PA. Data generated by these wearable devices can be used as either summary-type PA metrics or continuous PA functions in ML algorithms or statistical models to address scientific questions regarding the effects of various PAs in digital health studies. Such studies can focus on increasing or maintaining cognitive functions in participants with neurodegenerative conditions, decreasing the risk of chronic diseases such as cardiovascular diseases, and maintaining physical fitness in general populations. Moreover,

XSL•FO
RenderX

research on the use of PA for personalized health is expected to increase. For instance, studies leveraging biobank databases that contain both PA and genetic information have been conducted to investigate their impacts on health outcomes such as psychiatric disorders and circadian rhythms [30,85,458] As analytical tools continue to evolve alongside technological advancements, this review delineates the dimensions, developments, and transformations within methodological frameworks providing foundational guidelines for future research on PA and health.

Acknowledgments

This work was partially supported by Ministry of Science and Technology, Taiwan (110-2314-B-002-078-MY3).

Authors' Contributions

YTL drafted the manuscript and designed the Microsoft Excel file for managing the reviewed articles. CW and CKH revised the manuscript. All authors participated in developing the review's strategy and reviewed the selected articles.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 109 KB - jmir_v26i1e59497_app1.docx]

Multimedia Appendix 2
Key literature search terms and search results.
[DOCX File , 19 KB - jmir_v26i1e59497_app2.docx]

Multimedia Appendix 3
References of articles included in the review.
[DOCX File , 248 KB - jmir_v26i1e59497_app3.docx]

References

1. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 1985 Apr;100(2):126-131 [FREE Full text] [Medline: 3920711]
2. Physical activity and health; a report of the Surgeon General : executive summary. Centers for Disease Control and Prevention. 1996. URL: https://stacks.cdc.gov/view/cdc/13242 [accessed 2024-03-31]
3. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med 2014 Jul;48(13):1019-1023 [FREE Full text] [doi: 10.1136/bjsports-2014-093546] [Medline: 24782483]
4. Motahari-Nezhad H, Fgaier M, Mahdi Abid M, Péntek M, Gulácsi L, Zrubka Z. Digital biomarker-based studies: scoping review of systematic reviews. JMIR Mhealth Uhealth 2022 Oct 24;10(10):e35722 [FREE Full text] [doi: 10.2196/35722] [Medline: 36279171]
5. Barker J, Smith Byrne K, Doherty A, Foster C, Rahimi K, Ramakrishnan R, et al. Physical activity of UK adults with chronic disease: cross-sectional analysis of accelerometer-measured physical activity in 96 706 UK Biobank participants. Int J Epidemiol 2019 Aug 01;48(4):1167-1174 [FREE Full text] [doi: 10.1093/ije/dyy294] [Medline: 30721947]
6. Ramakrishnan R, Doherty A, Smith-Byrne K, Rahimi K, Bennett D, Woodward M, et al. Accelerometer measured physical activity and the incidence of cardiovascular disease: Evidence from the UK Biobank cohort study. PLoS Med 2021 Jan 12;18(1):e1003487. [doi: 10.1371/journal.pmed.1003487]
7. van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. PLoS One 2015;10(11):e0142533 [FREE Full text] [doi: 10.1371/journal.pone.0142533] [Medline: 26569414]
8. van Hees VT, Sabia S, Jones SE, Wood AR, Anderson KN, Kivimäki M, et al. Estimating sleep parameters using an accelerometer without sleep diary. Sci Rep 2018 Aug 28;8(1):12975 [FREE Full text] [doi: 10.1038/s41598-018-31266-z] [Medline: 30154500]
9. Vasudevan S, Saha A, Tarver ME, Patel B. Digital biomarkers: convergence of digital health technologies and biomarkers. NPJ Digit Med 2022 Mar 25;5(1):36 [FREE Full text] [doi: 10.1038/s41746-022-00583-z] [Medline: 35338234]
10. Lee M, Khan AM, Kim J, Cho Y, Kim T. A single tri-axial accelerometer-based real-time personal life log system capable of activity classification and exercise information generation. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:1390-1393. [doi: 10.1109/IEMBS.2010.5626729] [Medline: 21096339]

XSL•FO
RenderX

11. Vallance JK, Winkler EA, Gardiner PA, Healy GN, Lynch BM, Owen N. Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005-2006). Prev Med 2011 Oct;53(4-5):284-288. [doi: 10.1016/j.ypmed.2011.07.013] [Medline: 21820466]
12. Moy ML, Garshick E, Matthess KR, Lew R, Reilly JJ. Accuracy of uniaxial accelerometer in chronic obstructive pulmonary disease. J Rehabil Res Dev 2008;45(4):611-617 [FREE Full text] [doi: 10.1682/jrrd.2007.09.0147] [Medline: 18712646]
13. Long X, Yin B, Aarts RM. Single-accelerometer-based daily physical activity classification. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:6107-6110. [doi: 10.1109/IEMBS.2009.5334925] [Medline: 19965261]
14. Chen KY, Bassett DR. The technology of accelerometry-based activity monitors: current and future. Med Sci Sports Exerc 2005 Nov;37(11 Suppl):S490-S500. [Medline: 16294112]
15. Liu F, Wanigatunga AA, Schrack JA. Assessment of physical activity in adults using wrist accelerometers. Epidemiol Rev 2022 Jan 14;43(1):65-93 [FREE Full text] [doi: 10.1093/epirev/mxab004] [Medline: 34215874]
16. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA accelerometer. Med Sci Sports Exerc 2011 Jun;43(6):1085-1093. [doi: 10.1249/MSS.0b013e31820513be] [Medline: 21088628]
17. Murphy SL. Review of physical activity measurement using accelerometers in older adults: considerations for research design and conduct. Prev Med 2009 Feb;48(2):108-114. [doi: 10.1016/j.ypmed.2008.12.001] [Medline: 19111780]
18. Godinho C, Domingos J, Cunha G, Santos AT, Fernandes RM, Abreu D, et al. A systematic review of the characteristics and validity of monitoring technologies to assess Parkinson's disease. J Neuroeng Rehabil 2016 Mar 12;13:24 [FREE Full text] [doi: 10.1186/s12984-016-0136-7] [Medline: 26969628]
19. Kelly LA, McMillan DG, Anderson A, Fippinger M, Fillerup G, Rider J. Validity of actigraphs uniaxial and triaxial accelerometers for assessment of physical activity in adults in laboratory conditions. BMC Med Phys 2013 Nov 26;13(1):5 [FREE Full text] [doi: 10.1186/1756-6649-13-5] [Medline: 24279826]
20. Montoye AH, Conger SA, Connolly CP, Imboden MT, Nelson MB, Bock JM, et al. Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings. Meas Phys Educ Exerc Sci 2017 Jun 22;21(4):223-234. [doi: 10.1080/1091367x.2017.1337638]
21. Migueles J, Rowlands A, Huber F, Sabia S, van HV. GGIR: a research community driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. J Meas Phys Behav Sep 2019;2(3):188-196. [doi: 10.1123/jmpb.2018-0063]
22. Avci A, Bosch S, Marin-Perianu M, Marin-Perianu R, Havinga P. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey. In: Proceedings of the 23th International Conference on Architecture of Computing Systems. 2010 Presented at: ACS '10; February 22-23, 2010; Hannover, Germany p. 1-10 URL: https://ieeexplore.ieee.org/document/5759000 [doi: 10.1109/iswc.2010.5665858]
23. Plasqui G, Bonomi AG, Westerterp KR. Daily physical activity assessment with accelerometers: new insights and validation studies. Obes Rev 2013 Jun;14(6):451-462. [doi: 10.1111/obr.12021] [Medline: 23398786]
24. Garatachea N, Torres LG, González GJ. Physical activity and energy expenditure measurements using accelerometers in older adults. Nutr Hosp 2010;25(2):224-230 [FREE Full text] [Medline: 20449530]
25. Rao PJ, Phan K, Maharaj MM, Pelletier MH, Walsh WR, Mobbs RJ. Accelerometers for objective evaluation of physical activity following spine surgery. J Clin Neurosci 2016 Apr;26:14-18. [doi: 10.1016/j.jocn.2015.05.064] [Medline: 26765766]
26. Broderick JM, Ryan J, O'Donnell DM, Hussey J. A guide to assessing physical activity using accelerometry in cancer patients. Support Care Cancer 2014 Apr;22(4):1121-1130. [doi: 10.1007/s00520-013-2102-2] [Medline: 24389829]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
29. Li A, Ji L, Wang S, Wu J. Physical activity classification using a single triaxial accelerometer based on HMM. In: Proceedings of the 2010 IET International Conference on Wireless Sensor Network. 2010 Presented at: IET-WSN '10; November 15-17, 2010; Beijing, China p. 155-160 URL: https://ieeexplore.ieee.org/document/5741087 [doi: 10.1049/cp.2010.1045]
30. Dennison CA, Legge SE, Bracher-Smith M, Menzies G, Escott-Price V, Smith DJ, et al. Association of genetic liability for psychiatric disorders with accelerometer-assessed physical activity in the UK Biobank. PLoS One 2021;16(3):e0249189 [FREE Full text] [doi: 10.1371/journal.pone.0249189] [Medline: 33770123]
31. Nawab KA, Storey BC, Staplin N, Walmsley R, Haynes R, Sutherland S, et al. Accelerometer-measured physical activity and functional behaviours among people on dialysis. Clin Kidney J 2021 Mar;14(3):950-958 [FREE Full text] [doi: 10.1093/ckj/sfaa045] [Medline: 33777379]
32. McDevitt B, Connolly J, Duddy D, Doherty R, Condell J. Preliminary investigations of the validity and interinstrument reliability for classification of accelerometer physical activity cut-points against indirect caliometry in healthy adults. In: Proceedings of the 33rd Irish Signals and Systems Conference. 2022 Presented at: ISSC '22; June 9-10, 2022; Cork, Ireland p. 1-9 URL: https://ieeexplore.ieee.org/document/9826199 [doi: 10.1109/issc55427.2022.9826199]

XSL•FO
RenderX

33. Migueles JH, Cadenas-Sanchez C, Alcantara JM, Leal-Martín J, Mañas A, Ara I, et al. Calibration and cross-validation of accelerometer cut-points to classify sedentary time and physical activity from hip and non-dominant and dominant wrists in older adults. Sensors (Basel) 2021 May 11;21(10):3326 [FREE Full text] [doi: 10.3390/s21103326] [Medline: 34064790]
34. Agiovlasitis S, Ballenger BK, Schultz EE, Du Q, Motl RW. Calibration of hip accelerometers for measuring physical activity and sedentary behaviours in adults with Down syndrome. J Intellect Disabil Res 2023 Feb;67(2):172-181 [FREE Full text] [doi: 10.1111/jir.13002] [Medline: 36514268]
35. Pulsford RM, Cortina-Borja M, Rich C, Kinnafick F, Dezateux C, Griffiths LJ. Actigraph accelerometer-defined boundaries for sedentary behaviour and physical activity intensities in 7 year old children. PLoS One 2011;6(8):e21822 [FREE Full text] [doi: 10.1371/journal.pone.0021822] [Medline: 21853021]
36. Roscoe CM, James RS, Duncan MJ. Calibration of GENEActiv accelerometer wrist cut-points for the assessment of physical activity intensity of preschool aged children. Eur J Pediatr 2017 Aug 3;176(8):1093-1098. [doi: 10.1007/s00431-017-2948-2] [Medline: 28674825]
37. Qi G, Dutta D, Leroux A, Ray D, Muschelli J, Crainiceanu C, et al. Genome-wide association studies of 27 accelerometry-derived physical activity measurements identified novel loci and genetic mechanisms. Genet Epidemiol 2022 Mar;46(2):122-138 [FREE Full text] [doi: 10.1002/gepi.22441] [Medline: 35043453]
38. Benadjaoud MA, Menai M, van Hees VT, Zipunnikov V, Regnaux J, Kivimäki M, et al. The association between accelerometer-assessed physical activity and respiratory function in older adults differs between smokers and non-smokers. Sci Rep 2019 Jul 16;9(1):10270 [FREE Full text] [doi: 10.1038/s41598-019-46771-y] [Medline: 31311982]
39. Chuang FC, Wang JS, Yang YT, Kao TP. A wearable activity sensor system and its physical activity classification scheme. In: Proceedings of the 2012 International Joint Conference on Neural Networks. 2012 Presented at: IJCNN '12; June 10-15, 2012; Brisbane, Australia p. 1-6 URL: https://ieeexplore.ieee.org/document/6252581 [doi: 10.1109/ijcnn.2012.6252581]
40. Zhang S, Rowlands AV, Murray P, Hurst TL. Physical activity classification using the GENEA wrist-worn accelerometer. Med Sci Sports Exerc 2012 Apr;44(4):742-748. [doi: 10.1249/MSS.0b013e31823bf95c] [Medline: 21988935]
41. Banker M, Song PX. Supervised learning of physical activity features from functional accelerometer data. IEEE J Biomed Health Inform 2023 Dec;27(12):5710-5721. [doi: 10.1109/JBHI.2023.3318205] [Medline: 37738184]
42. Huang W, Hao L, Wu X, Yu X, Cui E, Leroux A. Gender difference in "second-shift" physical activity: new insights from analyzing accelerometry data in a nationally representative sample. SSM Popul Health 2023 Dec;24:101536 [FREE Full text] [doi: 10.1016/j.ssmph.2023.101536] [Medline: 37927817]
43. Tabacu L, Ledbetter M, Leroux A, Crainiceanu C, Smirnova E. Quantifying the varying predictive value of physical activity measures obtained from wearable accelerometers on all-cause mortality over short to medium time horizons in NHANES 2003-2006. Sensors (Basel) 2020 Dec 22;21(1):4 [FREE Full text] [doi: 10.3390/s21010004] [Medline: 33374911]
44. Bonomi AG, Goris AH, Yin B, Westerterp KR. Detection of type, duration, and intensity of physical activity using an accelerometer. Med Sci Sports Exerc 2009 Sep;41(9):1770-1777. [doi: 10.1249/MSS.0b013e3181a24536] [Medline: 19657292]
45. Stewart T, Narayanan A, Hedayatrad L, Neville J, Mackay L, Duncan S. A dual-accelerometer system for classifying physical activity in children and adults. Med Sci Sports Exerc 2018 Dec;50(12):2595-2602. [doi: 10.1249/MSS.0000000000001717] [Medline: 30048411]
46. García-Massó X, Serra-Añó P, Gonzalez LM, Ye-Lin Y, Prats-Boluda G, Garcia-Casado J. Identifying physical activity type in manual wheelchair users with spinal cord injury by means of accelerometers. Spinal Cord 2015 Oct;53(10):772-777. [doi: 10.1038/sc.2015.81] [Medline: 25987002]
47. Chandler JL, Brazendale K, Beets MW, Mealing BA. Classification of physical activity intensities using a wrist-worn accelerometer in 8-12-year-old children. Pediatr Obes 2016 Apr 20;11(2):120-127. [doi: 10.1111/ijpo.12033] [Medline: 25893950]
48. Duncan MJ, Roscoe CM, Faghy M, Tallis J, Eyre EL. Estimating physical activity in children aged 8-11 years using accelerometry: contributions from fundamental movement skills and different accelerometer placements. Front Physiol 2019;10:242 [FREE Full text] [doi: 10.3389/fphys.2019.00242] [Medline: 30936837]
49. Gába A, Dygrýn J, Mitáš J, Jakubec L, Frömel K. Effect of accelerometer cut-off points on the recommended level of physical activity for obesity prevention in children. PLoS One 2016;11(10):e0164282 [FREE Full text] [doi: 10.1371/journal.pone.0164282] [Medline: 27723835]
50. Bammann K, Thomson NK, Albrecht BM, Buchan DS, Easton C. Generation and validation of ActiGraph GT3X+ accelerometer cut-points for assessing physical activity intensity in older adults. The OUTDOOR ACTIVE validation study. PLoS One 2021;16(6):e0252615 [FREE Full text] [doi: 10.1371/journal.pone.0252615] [Medline: 34081715]
51. McCracken LA, Ma JK, Voss C, Chan FH, Martin Ginis KA, West CR. Wrist accelerometry for physical activity measurement in individuals with spinal cord injury-a need for individually calibrated cut-points. Arch Phys Med Rehabil 2018 Apr;99(4):684-689. [doi: 10.1016/j.apmr.2017.10.024] [Medline: 29222006]
52. Mardini MT, Bai C, Wanigatunga AA, Saldana S, Casanova R, Manini TM. Age differences in estimating physical activity by wrist accelerometry using machine learning. Sensors (Basel) 2021 May 12;21(10):3352 [FREE Full text] [doi: 10.3390/s21103352] [Medline: 34065906]

XSL•FO
RenderX

53. Xie C, Bi S, Dong M, Li L, Chi S. Walking recognition method for physical activity analysis system of child based on wearable accelerometer. In: Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics. 2017 Presented at: ROBIO '17; December 5-8, 2017; Macau, Macao p. 2439-2443 URL: https://ieeexplore.ieee.org/document/8324785 [doi: 10.1109/robio.2017.8324785]
54. Shim J, Fleisch E, Barata F. precise segmentation of U.S. adults from 24-hour wearable-based physical activity profiles using machine learning clustering. In: Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics. 2023 Presented at: ICHI '23; June 26-29, 2023; New York, NY p. 498-500 URL: https://www.computer.org/csdl/proceedings-article/ichi/2023/026300a498/1SN7x53nsVG [doi: 10.1109/ichi57859.2023.00083]
55. Kos M, Bogdan M, Glynn NW, Harezlak J. Classification of human physical activity based on raw accelerometry data via spherical coordinate transformation. Stat Med 2020 Sep 30;39(22):2901-2920 [FREE Full text] [doi: 10.1002/sim.8582] [Medline: 32478905]
56. Wullems JA, Verschueren SM, Degens H, Morse CI, Onambélé GL. Performance of thigh-mounted triaxial accelerometer algorithms in objective quantification of sedentary behaviour and physical activity in older adults. PLoS One 2017 Nov 20;12(11):e0188215 [FREE Full text] [doi: 10.1371/journal.pone.0188215] [Medline: 29155839]
57. Qi J, Liang H, Chen J, Peng X, Newcombe L, Yang P. A hybrid hierarchical model for accessing physical activity recognition towards free-living environments. In: Proceedings of the 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing. 2020 Presented at: ISPA-BD '20; December 17-19, 2020; Exeter, UK p. 1342-1347 URL: https://ieeexplore.ieee.org/document/9443709 [doi: 10.1109/ispa-bdcloud-socialcom-sustaincom51426.2020.00199]
58. Mannini A, Sabatini AM. Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors (Basel) 2010;10(2):1154-1175 [FREE Full text] [doi: 10.3390/s100201154] [Medline: 22205862]
59. Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. Med Sci Sports Exerc 2006 Sep;38(9):1626-1634. [doi: 10.1249/01.mss.0000227542.43669.45] [Medline: 16960524]
60. Thornton CB, Kolehmainen N, Nazarpour K. Using unsupervised machine learning to quantify physical activity from accelerometry in a diverse and rapidly changing population. PLOS Digit Health 2023 Apr;2(4):e0000220 [FREE Full text] [doi: 10.1371/journal.pdig.0000220] [Medline: 37018183]
61. Cochrane SK, Chen S, Fitzgerald JD, Dodson JA, Fielding RA, King AC, LIFE Study Research Group. Association of accelerometry-measured physical activity and cardiovascular events in mobility-limited older adults: the LIFE (Lifestyle Interventions and Independence for Elders) study. J Am Heart Assoc 2017 Dec 02;6(12):e007215 [FREE Full text] [doi: 10.1161/JAHA.117.007215] [Medline: 29197830]
62. Vásquez PM, Durazo-Arvizu RA, Marquez DX, Argos M, Lamar M, Odoms-Young A, et al. Association of accelerometer-measured physical activity and cardiovascular health in the Hispanic community health study/study of Latinos (HCHS/SOL). Hisp Health Care Int 2022 Mar;20(1):15-24. [doi: 10.1177/1540415320985581] [Medline: 33685281]
63. Appelqvist-Schmidlechner K, Raitanen J, Vasankari T, Kyröläinen H, Häkkinen A, Honkanen T, et al. Relationship between accelerometer-based physical activity, sedentary behavior, and mental health in young finnish men. Front Public Health 2022;10:820852 [FREE Full text] [doi: 10.3389/fpubh.2022.820852] [Medline: 35252097]
64. Ghorbani S, Afshari M, Eckelt M, Dana A, Bund A. Associations between physical activity and mental health in iranian adolescents during the COVID-19 pandemic: an accelerometer-based study. Children (Basel) 2021 Nov 07;8(11):1022 [FREE Full text] [doi: 10.3390/children8111022] [Medline: 34828736]
65. Dohrn I, Sjöström M, Kwak L, Oja P, Hagströmer M. Accelerometer-measured sedentary time and physical activity—a 15 year follow-up of mortality in a Swedish population-based cohort. J Sci Med Sport 2018 Jul;21(7):702-707. [doi: 10.1016/j.jsams.2017.10.035]
66. Strain T, Wijndaele K, Dempsey PC, Sharp SJ, Pearce M, Jeon J, et al. Wearable-device-measured physical activity and future health risk. Nat Med 2020 Sep;26(9):1385-1391. [doi: 10.1038/s41591-020-1012-3] [Medline: 32807930]
67. Mañas A, Del Pozo-Cruz B, Guadalupe-Grau A, Marín-Puyalto J, Alfaro-Acha A, Rodríguez-Mañas L, et al. Reallocating accelerometer-assessed sedentary time to light or moderate- to vigorous-intensity physical activity reduces frailty levels in older adults: an isotemporal substitution approach in the TSHA study. J Am Med Dir Assoc 2018 Feb;19(2):185.e1-185.e6. [doi: 10.1016/j.jamda.2017.11.003] [Medline: 29269096]
68. Aadland E, Andersen LB, Resaland GK, Kvalheim OM. Interpretation of multivariate association patterns between multicollinear physical activity accelerometry data and cardiometabolic health in children-a tutorial. Metabolites 2019 Jul 02;9(7):129 [FREE Full text] [doi: 10.3390/metabo9070129] [Medline: 31269708]
69. Aadland E, Kvalheim OM, Hansen BH, Kriemler S, Ried-Larsen M, Wedderkopp N, International Children's Accelerometry Database (ICAD) Collaborators. The multivariate physical activity signature associated with metabolic health in children and youth: an International Children's Accelerometry Database (ICAD) analysis. Prev Med 2020 Dec;141:106266 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106266] [Medline: 33022325]
70. Graves JL, Qiao YS, Moored KD, Boudreau RM, Venditti EM, Krafty RT, et al. Profiles of accelerometry-derived physical activity are related to perceived physical fatigability in older adults. Sensors (Basel) 2021 Mar 02;21(5):1718 [FREE Full text] [doi: 10.3390/s21051718] [Medline: 33801352]

XSL•FO
RenderX

71. Del Pozo Cruz B, Ahmadi M, Inan-Eroglu E, Huang B, Stamatakis E. Prospective associations of accelerometer-assessed physical activity with mortality and incidence of cardiovascular disease among adults with hypertension: the UK Biobank study. J Am Heart Assoc 2022 Mar 15;11(6):e023290 [FREE Full text] [doi: 10.1161/JAHA.121.023290] [Medline: 35253444]
72. Dohrn I, Welmer A, Hagströmer M. Accelerometry-assessed physical activity and sedentary time and associations with chronic disease and hospital visits - a prospective cohort study with 15 years follow-up. Int J Behav Nutr Phys Act 2019 Dec 09;16(1):125 [FREE Full text] [doi: 10.1186/s12966-019-0878-2] [Medline: 31818303]
73. LaMonte MJ, Buchner DM, Rillamas-Sun E, Di C, Evenson KR, Bellettiere J, et al. Accelerometer-measured physical activity and mortality in women aged 63 to 99. J Am Geriatr Soc 2018 May;66(5):886-894 [FREE Full text] [doi: 10.1111/jgs.15201] [Medline: 29143320]
74. Berninger N, Knell G, Gabriel KP, Plasqui G, Crutzen R, Hoor GT. Bidirectional day-to-day associations of reported sleep duration with accelerometer measured physical activity and sedentary time among Dutch adolescents: an observational study. J Meas Phys Behav 2020 Dec;3(4):304-314 [FREE Full text] [doi: 10.1123/jmpb.2020-0010] [Medline: 35665029]
75. Sit CH, Huang WY, Yu JJ, McKenzie TL. Accelerometer-assessed physical activity and sedentary time at school for children with disabilities: seasonal variation. Int J Environ Res Public Health 2019 Aug 30;16(17):3163 [FREE Full text] [doi: 10.3390/ijerph16173163] [Medline: 31480233]
76. Halonen JI, Pulakka A, Pentti J, Kallio M, Koskela S, Kivimäki M, et al. Cross-sectional associations of neighbourhood socioeconomic disadvantage and greenness with accelerometer-measured leisure-time physical activity in a cohort of ageing workers. BMJ Open 2020 Aug 16;10(8):e038673 [FREE Full text] [doi: 10.1136/bmjopen-2020-038673] [Medline: 32801206]
77. Oliver M, Schluter PJ, Schofield GM, Paterson J. Factors related to accelerometer-derived physical activity in Pacific children aged 6 years. Asia Pac J Public Health 2011 Jan;23(1):44-56. [doi: 10.1177/1010539510370992] [Medline: 20484244]
78. Keller JL, Tian F, Fitzgerald KC, Mische L, Ritter J, Costello MG, et al. Using real-world accelerometry-derived diurnal patterns of physical activity to evaluate disability in multiple sclerosis. J Rehabil Assist Technol Eng 2022 Jan 12;9:20556683211067362 [FREE Full text] [doi: 10.1177/20556683211067362] [Medline: 35070348]
79. Assah FK, Ekelund U, Brage S, Corder K, Wright A, Mbanya JC, et al. Predicting physical activity energy expenditure using accelerometry in adults from sub-Sahara Africa. Obesity (Silver Spring) 2009 Aug;17(8):1588-1595 [FREE Full text] [doi: 10.1038/oby.2009.39] [Medline: 19247268]
80. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas 2014 Nov;35(11):2191-2203 [FREE Full text] [doi: 10.1088/0967-3334/35/11/2191] [Medline: 25340969]
81. Ladlow P, Nightingale TE, McGuigan MP, Bennett AN, Phillip R, Bilzon JL. Impact of anatomical placement of an accelerometer on prediction of physical activity energy expenditure in lower-limb amputees. PLoS One 2017;12(10):e0185731 [FREE Full text] [doi: 10.1371/journal.pone.0185731] [Medline: 28982199]
82. Smirnova E, Leroux A, Cao Q, Tabacu L, Zipunnikov V, Crainiceanu C, et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National Health and Nutritional Examination Survey 2003-2006. J Gerontol A Biol Sci Med Sci 2020 Sep 16;75(9):1779-1785 [FREE Full text] [doi: 10.1093/gerona/glz193] [Medline: 31504213]
83. Chen M, Landré B, Marques-Vidal P, van Hees VT, van Gennip AC, Bloomberg M, et al. Identification of physical activity and sedentary behaviour dimensions that predict mortality risk in older adults: development of a machine learning model in the Whitehall II accelerometer sub-study and external validation in the CoLaus study. EClinicalMedicine 2023 Jan;55:101773 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101773] [Medline: 36568684]
84. White T, Westgate K, Wareham NJ, Brage S. Estimation of physical activity energy expenditure during free-living from wrist accelerometry in UK adults. PLoS One 2016;11(12):e0167472 [FREE Full text] [doi: 10.1371/journal.pone.0167472] [Medline: 27936024]
85. Luo M, Yu C, Del Pozo Cruz B, Chen L, Ding D. Accelerometer-measured intensity-specific physical activity, genetic risk and incident type 2 diabetes: a prospective cohort study. Br J Sports Med 2023 Oct;57(19):1257-1264 [FREE Full text] [doi: 10.1136/bjsports-2022-106653] [Medline: 37277158]
86. Aadland E, Kvalheim OM, Anderssen SA, Resaland GK, Andersen LB. Multicollinear physical activity accelerometry data and associations to cardiometabolic health: challenges, pitfalls, and potential solutions. Int J Behav Nutr Phys Act 2019 Aug 27;16(1):74 [FREE Full text] [doi: 10.1186/s12966-019-0836-z] [Medline: 31455305]
87. Aadland E, Steene-Johannessen J. The use of individual cut points from treadmill walking to assess free-living moderate to vigorous physical activity in obese subjects by accelerometry: is it useful? BMC Med Res Methodol 2012 Dec 15;12:172 [FREE Full text] [doi: 10.1186/1471-2288-12-172] [Medline: 23152980]
88. Agarwal V, Smuck M, Shah NH. Quantifying the relative change in physical activity after total knee arthroplasty using accelerometer based measurements. AMIA Jt Summits Transl Sci Proc 2017;2017:463-472 [FREE Full text] [Medline: 28815146]

XSL•FO
RenderX

89. Agbaje AO. Associations of accelerometer-based sedentary time, light physical activity and moderate-to-vigorous physical activity with resting cardiac structure and function in adolescents according to sex, fat mass, lean mass, BMI, and hypertensive status. Scand J Med Sci Sports 2023 Aug;33(8):1399-1411 [FREE Full text] [doi: 10.1111/sms.14365] [Medline: 37035905]
90. Agbaje AO. Mediating effect of fat mass, lean mass, blood pressure and insulin resistance on the associations of accelerometer-based sedentary time and physical activity with arterial stiffness, carotid IMT and carotid elasticity in 1574 adolescents. J Hum Hypertens 2024 May;38(5):393-403 [FREE Full text] [doi: 10.1038/s41371-024-00905-6] [Medline: 38409590]
91. Agbaje AO, Perng W, Tuomainen T. Effects of accelerometer-based sedentary time and physical activity on DEXA-measured fat mass in 6059 children. Nat Commun 2023 Dec 12;14(1):8232 [FREE Full text] [doi: 10.1038/s41467-023-43316-w] [Medline: 38086810]
92. Aguiar B, Silva J, Rocha T, Carneiro S, Sousa I. Monitoring physical activity and energy expenditure with smartphones. In: Proceedings of the 2014 International Conference on Biomedical and Health Informatics. 2014 Presented at: BHI '14; June 1-4, 2014; Valencia, Spain p. 664-667 URL: https://ieeexplore.ieee.org/document/6864451 [doi: 10.1109/bhi.2014.6864451]
93. Ahn H, Choi E, Rhee T, Choi J, Lee K, Kwon S, et al. Accelerometer-derived physical activity and the risk of death, heart failure, and stroke in patients with atrial fibrillation: a prospective study from UK Biobank. Br J Sports Med 2024 May 02;58(8):427-434. [doi: 10.1136/bjsports-2023-106862] [Medline: 38418213]
94. Aittasalo M, Vähä-Ypyä H, Vasankari T, Husu P, Jussila A, Sievänen H. Mean amplitude deviation calculated from raw acceleration data: a novel method for classifying the intensity of adolescents' physical activity irrespective of accelerometer brand. BMC Sports Sci Med Rehabil 2015;7:18 [FREE Full text] [doi: 10.1186/s13102-015-0010-0] [Medline: 26251724]
95. Ajja R, Wikkeling-Scott LF, Brazendale K, Hijazi R, Abdulle A. Accelerometer measured physical activity patterns of children during segmented school day in Abu Dhabi. BMC Pediatr 2021 May 17;21(1):182 [FREE Full text] [doi: 10.1186/s12887-021-02639-7] [Medline: 33865344]
96. Albrecht BM, Stalling I, Recke C, Bammann K. Accelerometer-assessed outdoor physical activity is associated with meteorological conditions among older adults: cross-sectional results from the OUTDOOR ACTIVE study. PLoS One 2020;15(1):e0228053 [FREE Full text] [doi: 10.1371/journal.pone.0228053] [Medline: 31978178]
97. Ali H, Messina E, Bisiani R. Subject-dependent physical activity recognition model framework with a semi-supervised clustering approach. In: Proceedings of the 2013 European Modelling Symposium. 2013 Presented at: EMS '13; November 20-22, 2013; Manchester, UK p. 42-47 URL: https://ieeexplore.ieee.org/document/6779819 [doi: 10.1109/ems.2013.7]
98. Allen CL, Montes E, Hoang T, Romo T, Peña J, Navarro J. Can stereotype threat and lift visual messages affect subsequent physical activity? Evidence from a controlled experiment using accelerometers. Health Commun 2023 Dec 08:1-12. [doi: 10.1080/10410236.2023.2277573] [Medline: 37941378]
99. Alsamman RA, Shousha TM, Faris ME, Abdelrahim DN, Arumugam A. Association of sociodemographic, anthropometric, and sleep quality factors with accelerometer-measured sitting and physical activity times among Emirati working women during the COVID-19 pandemic: a cross-sectional study. Womens Health (Lond) 2024;20:17455057231225539 [FREE Full text] [doi: 10.1177/17455057231225539] [Medline: 38279820]
100. Anastasopoulou P, Tansella M, Stumpp J, Shammas L, Hey S. Classification of human physical activity and energy expenditure estimation by accelerometry and barometry. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:6451-6454. [doi: 10.1109/EMBC.2012.6347471] [Medline: 23367406]
101. Andersen MB, Ostenfeld EB, Fuglsang J, Møller M, Daugaard M, Ovesen PG. Maternal prepregnancy body mass index and physical activity during pregnancy assessed by accelerometer. Am J Obstet Gynecol MFM 2020 Dec;2(4):100182. [doi: 10.1016/j.ajogmf.2020.100182] [Medline: 33345908]
102. Andersson C, Lyass A, Larson MG, Spartano NL, Vita JA, Benjamin EJ, et al. Physical activity measured by accelerometry and its associations with cardiac structure and vascular function in young and middle-aged adults. J Am Heart Assoc 2015 Mar 19;4(3):e001528 [FREE Full text] [doi: 10.1161/JAHA.114.001528] [Medline: 25792127]
103. Armstrong B, Covington LB, Hager ER, Black MM. Objective sleep and physical activity using 24-hour ankle-worn accelerometry among toddlers from low-income families. Sleep Health 2019 Oct;5(5):459-465 [FREE Full text] [doi: 10.1016/j.sleh.2019.04.005] [Medline: 31171491]
104. Arteaga D, Donnelly T, Crum K, Markham L, Killian M, Burnette WB, et al. Assessing physical activity using accelerometers in youth with Duchenne muscular dystrophy. J Neuromuscul Dis 2020;7(3):331-342 [FREE Full text] [doi: 10.3233/JND-200478] [Medline: 32417792]
105. Augustin NH, Mattocks C, Cooper AR, Ness AR, Faraway JJ. Modelling fat mass as a function of weekly physical activity profiles measured by actigraph accelerometers. Physiol Meas 2012 Dec;33(11):1831-1839. [doi: 10.1088/0967-3334/33/11/1831] [Medline: 23110964]
106. Ayabe M, Kumahara H. Effect of handling breaks on estimation of heart rate responses to bouts of physical activity among young women: an accelerometer research issue. Gait Posture 2020 Oct;81:1-6. [doi: 10.1016/j.gaitpost.2020.06.032] [Medline: 32645577]

XSL•FO
RenderX

107. Baldanzi G, Sayols-Baixeras S, Ekblom-Bak E, Ekblom Ö, Dekkers KF, Hammar U, et al. Accelerometer-based physical activity is associated with the gut microbiota in 8416 individuals in SCAPIS. EBioMedicine 2024 Feb;100:104989 [FREE Full text] [doi: 10.1016/j.ebiom.2024.104989] [Medline: 38301483]
108. Ballin M, Antonsson O, Rosenqvist V, Nordström P, Nordström A. Association of dog ownership with accelerometer-measured physical activity and daily steps in 70-year-old individuals: a population-based cross-sectional study. BMC Public Health 2021 Dec 21;21(1):2313 [FREE Full text] [doi: 10.1186/s12889-021-12401-4] [Medline: 34933682]
109. Bangen KJ, Calcetas AT, Thomas KR, Wierenga C, Smith CN, Bordyug M, et al. Greater accelerometer-measured physical activity is associated with better cognition and cerebrovascular health in older adults. J Int Neuropsychol Soc 2023 Feb 15;29(9):859-869. [doi: 10.1017/s1355617723000140]
110. Baque E, Sakzewski L, Trost SG, Boyd RN, Barber L. Validity of accelerometry to measure physical activity intensity in children with an acquired brain injury. Pediatr Phys Ther 2017 Oct;29(4):322-329. [doi: 10.1097/PEP.0000000000000439] [Medline: 28953176]
111. Barrett S, Begg S, O Halloran P, Kingsley M. The effect of a physical activity coaching intervention on accelerometer-measured sedentary behaviours in insufficiently physically active ambulatory hospital patients. Int J Environ Res Public Health 2021 May 22;18(11):5543 [FREE Full text] [doi: 10.3390/ijerph18115543] [Medline: 34067292]
112. Baumann S, Groß S, Voigt L, Ullrich A, Weymar F, Schwaneberg T, et al. Pitfalls in accelerometer-based measurement of physical activity: the presence of reactivity in an adult population. Scand J Med Sci Sports 2017 Oct 12;28(3):1056-1063. [doi: 10.1111/sms.12977]
113. Baumann S, Guertler D, Weymar F, Bahls M, Dörr M, van den Berg N, et al. Do accelerometer-based physical activity patterns differentially affect cardiorespiratory fitness? A growth mixture modeling approach. J Behav Med 2020 Mar;43(1):99-107. [doi: 10.1007/s10865-019-00069-6] [Medline: 31190167]
114. Bayartai M, Määttä J, Karppinen J, Oura P, Takatalo J, Auvinen J, et al. Association of accelerometer-measured physical activity, back static muscular endurance and abdominal obesity with radicular pain and non-specific low back pain. Sci Rep 2023 May 12;13(1):7736 [FREE Full text] [doi: 10.1038/s41598-023-34733-4] [Medline: 37173344]
115. Bayartai M, Taulaniemi A, Tokola K, Vähä-Ypyä H, Parkkari J, Husu P, et al. Role of the interaction between lumbar kinematics and accelerometer-measured physical activity in bodily pain, physical functioning and work ability among health care workers with low back pain. J Electromyogr Kinesiol 2023 May;69:102744. [doi: 10.1016/j.jelekin.2023.102744] [Medline: 36680881]
116. Bergh IH, van Stralen MM, Grydeland M, Bjelland M, Lien N, Andersen LF, et al. Exploring mediators of accelerometer assessed physical activity in young adolescents in the health in adolescents study - a group randomized controlled trial. BMC Public Health 2012 Oct 21;12:814 [FREE Full text] [doi: 10.1186/1471-2458-12-814] [Medline: 22995043]
117. Biswas A, Chen C, Prince SA, Smith PM, Mustard CA. Daily accelerometer-measured physical activity patterns and associations with cardiometabolic health among Canadian working adults. Health Rep 2023 Mar 15;34(3):15-29 [FREE Full text] [doi: 10.25318/82-003-x202300300002-eng] [Medline: 36921073]
118. Bonomi AG, Plasqui G, Goris AH, Westerterp KR. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. J Appl Physiol (1985) 2009 Sep;107(3):655-661 [FREE Full text] [doi: 10.1152/japplphysiol.00150.2009] [Medline: 19556460]
119. Browne RA, Macêdo GA, Cabral LL, Oliveira GT, Vivas A, Fontes EB, et al. Initial impact of the COVID-19 pandemic on physical activity and sedentary behavior in hypertensive older adults: an accelerometer-based analysis. Exp Gerontol 2020 Dec;142:111121 [FREE Full text] [doi: 10.1016/j.exger.2020.111121] [Medline: 33132145]
120. Carlson RH, Huebner DR, Hoarty CA, Whittington J, Haynatzki G, Balas MC, et al. Treadmill gait speeds correlate with physical activity counts measured by cell phone accelerometers. Gait Posture 2012 Jun;36(2):241-248 [FREE Full text] [doi: 10.1016/j.gaitpost.2012.02.025] [Medline: 22475727]
121. Carvalho EV, Reboredo MM, Gomes EP, Teixeira DR, Roberti NC, Mendes JO, et al. Physical activity in daily life assessed by an accelerometer in kidney transplant recipients and hemodialysis patients. Transplant Proc 2014;46(6):1713-1717 [FREE Full text] [doi: 10.1016/j.transproceed.2014.05.019] [Medline: 25131019]
122. Cassidy S, Fuller H, Chau J, Catt M, Bauman A, Trenell MI. Accelerometer-derived physical activity in those with cardio-metabolic disease compared to healthy adults: a UK Biobank study of 52,556 participants. Acta Diabetol 2018 Sep;55(9):975-979 [FREE Full text] [doi: 10.1007/s00592-018-1161-8] [Medline: 29808390]
123. Chan JA, Bosma H, Drosinou C, Timmermans EJ, Savelberg H, Schaper N, et al. Association of perceived and objective neighborhood walkability with accelerometer-measured physical activity and sedentary time in the Maastricht study. Scand J Med Sci Sports 2023 Dec;33(11):2313-2322. [doi: 10.1111/sms.14455] [Medline: 37489093]
124. Chen G, Qi Q, Hua S, Moon J, Spartano NL, Vasan RS, et al. Accelerometer-assessed physical activity and incident diabetes in a population covering the adult life span: the Hispanic community health study/study of Latinos. Am J Clin Nutr 2020 Dec 11;112(5):1318-1327 [FREE Full text] [doi: 10.1093/ajcn/nqaa232] [Medline: 32910816]
125. Chen T, Honda T, Chen S, Narazaki K, Kumagai S. Dose-response association between accelerometer-assessed physical activity and incidence of functional disability in older Japanese adults: a 6-year prospective study. J Gerontol A Biol Sci Med Sci 2020 Oct 16;75(9):1763-1770 [FREE Full text] [doi: 10.1093/gerona/glaa046] [Medline: 32134454]

XSL•FO
RenderX

126. Cheng TS, Brage S, van Sluijs EM, Ong KK. Pre-pubertal accelerometer-assessed physical activity and timing of puberty in British boys and girls: the millennium cohort study. Int J Epidemiol 2023 Oct 05;52(5):1316-1327 [FREE Full text] [doi: 10.1093/ije/dyad063] [Medline: 37208864]
127. Choe EY, He D, Sun G. Trading-off transit and non-transit physical activity among older people: evidence from longitudinal accelerometer data of a natural experiment study. J Urban Health 2023 May;100(2):408-417 [FREE Full text] [doi: 10.1007/s11524-022-00709-4] [Medline: 36656439]
128. Cohn-Schwartz E, Khalaila R. Accelerometer-assessed physical activity and cognitive performance among European adults aged 50+: the mediating effects of social contacts and depressive symptoms. Healthcare (Basel) 2022 Dec 14;10(11):2279 [FREE Full text] [doi: 10.3390/healthcare10112279] [Medline: 36421603]
129. Cooper AR, Goodman A, Page AS, Sherar LB, Esliger DW, van Sluijs EM, et al. Objectively measured physical activity and sedentary time in youth: the International children's accelerometry database (ICAD). Int J Behav Nutr Phys Act 2015 Sep 17;12:113 [FREE Full text] [doi: 10.1186/s12966-015-0274-5] [Medline: 26377803]
130. Corcoran MP, Chui KK, White DK, Reid KF, Kirn D, Nelson ME, et al. Accelerometer assessment of physical activity and its association with physical function in older adults residing at assisted care facilities. J Nutr Health Aging 2016;20(7):752-758. [doi: 10.1007/s12603-015-0640-7] [Medline: 27499309]
131. Corder K, Sharp SJ, Atkin AJ, Andersen LB, Cardon G, Page A, International Children's Accelerometry Database (ICAD) Collaborators. Age-related patterns of vigorous-intensity physical activity in youth: the International Children's Accelerometry Database. Prev Med Rep 2016 Dec;4:17-22 [FREE Full text] [doi: 10.1016/j.pmedr.2016.05.006] [Medline: 27413656]
132. Crouter SE, Flynn JI, Bassett DR. Estimating physical activity in youth using a wrist accelerometer. Med Sci Sports Exerc 2015 May;47(5):944-951 [FREE Full text] [doi: 10.1249/MSS.0000000000000502] [Medline: 25207928]
133. Crouter SE, Oody JF, Bassett DR. Estimating physical activity in youth using an ankle accelerometer. J Sports Sci 2018 Oct 08;36(19):2265-2271 [FREE Full text] [doi: 10.1080/02640414.2018.1449091] [Medline: 29517959]
134. D'Silva A, Bebb G, Boyle T, Johnson ST, Vallance JK. Demographic and clinical correlates of accelerometer assessed physical activity and sedentary time in lung cancer survivors. Psychooncology 2018 Mar;27(3):1042-1049. [doi: 10.1002/pon.4608] [Medline: 29226994]
135. da SI, van HV, Ramires VV, Knuth AG, Bielemann RM, Ekelund U, et al. Physical activity levels in three Brazilian birth cohorts as assessed with raw triaxial wrist accelerometry. Int J Epidemiol 2014 Dec;43(6):1959-1968 [FREE Full text] [doi: 10.1093/ije/dyu203] [Medline: 25361583]
136. da Silva SG, Evenson KR, da Silva ICM, Mendes MA, Domingues MR, da Silveira MF, et al. Correlates of accelerometer-assessed physical activity in pregnancy-the 2015 Pelotas (Brazil) birth cohort study. Scand J Med Sci Sports 2018 Aug;28(8):1934-1945 [FREE Full text] [doi: 10.1111/sms.13083] [Medline: 29542188]
137. Dankel SJ, Loenneke JP, Loprinzi PD. Combined associations of muscle-strengthening activities and accelerometer-assessed physical activity on multimorbidity: findings from NHANES. Am J Health Promot 2017 Jul;31(4):274-277. [doi: 10.4278/ajhp.150520-QUAN-894] [Medline: 26730562]
138. DasMahapatra P, Chiauzzi E, Bhalerao R, Rhodes J. Free-living physical activity monitoring in adult US patients with multiple sclerosis using a consumer wearable device. Digit Biomark 2018;2(1):47-63 [FREE Full text] [doi: 10.1159/000488040] [Medline: 32095756]
139. Davoudi A, Ozrazgat-Baslanti T, Tighe P, Bihorac A, Rashidi P. Pain and physical activity association in critically Ill patients. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:5696-5699 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9176227] [Medline: 33019268]
140. de Vries EA, Heijenbrok-Kal MH, van Kooten F, Giurgiu M, Ribbers GM, van den Berg-Emons RJ, et al. Unraveling the interplay between daily life fatigue and physical activity after subarachnoid hemorrhage: an ecological momentary assessment and accelerometry study. J Neuroeng Rehabil 2023 Oct 26;20(1):127 [FREE Full text] [doi: 10.1186/s12984-023-01241-5] [Medline: 37752550]
141. Delgado-Gonzalo R, Celka P, Renevey P, Dasen S, Sola J, Bertschi M. Physical activity profiling: activity-specific step counting and energy expenditure models using 3D wrist acceleration. In: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2015 Presented at: EMBC '15; August 25-29, 2015; Milan, Italy p. 25-29 URL: https://ieeexplore.ieee.org/document/7320271 [doi: 10.1109/embc.2015.7320271]
142. Diaz KM, Krupka DJ, Chang MJ, Kronish IM, Moise N, Goldsmith J, et al. Wrist-based cut-points for moderate- and vigorous-intensity physical activity for the actical accelerometer in adults. J Sports Sci 2018 Jan;36(2):206-212 [FREE Full text] [doi: 10.1080/02640414.2017.1293279] [Medline: 28282744]
143. Dibben GO, Gandhi MM, Taylor RS, Dalal HM, Metcalf B, Doherty P, et al. Physical activity assessment by accelerometry in people with heart failure. BMC Sports Sci Med Rehabil 2020;12:47 [FREE Full text] [doi: 10.1186/s13102-020-00196-7] [Medline: 32817798]
144. Dijk-Huisman HC, Senden R, Smeets MH, Marcellis RG, Magdelijns FJ, Lenssen AF. The effect of a smartphone app with an accelerometer on the physical activity behavior of hospitalized patients: a randomized controlled trial. Sensors (Basel) 2023 Oct 25;23(21):8704 [FREE Full text] [doi: 10.3390/s23218704] [Medline: 37960404]

XSL•FO
RenderX

145. Diniz-Sousa F, Veras L, Ribeiro JC, Boppre G, Devezas V, Santos-Sousa H, et al. Accelerometry calibration in people with class II-III obesity: energy expenditure prediction and physical activity intensity identification. Gait Posture 2020 Mar;76:104-109. [doi: 10.1016/j.gaitpost.2019.11.008] [Medline: 31756665]
146. Do B, Zink J, Mason TB, Belcher BR, Dunton GF. Physical activity and sedentary time among mothers of school-aged children: differences in accelerometer-derived pattern metrics by demographic, employment, and household factors. Womens Health Issues 2022;32(5):490-498 [FREE Full text] [doi: 10.1016/j.whi.2022.03.005] [Medline: 35491346]
147. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank study. PLoS One 2017;12(2):e0169649 [FREE Full text] [doi: 10.1371/journal.pone.0169649] [Medline: 28146576]
148. Domazet SL, Tarp J, Thomsen RW, Højlund K, Stidsen JV, Brønd JC, et al. Accelerometer-derived physical activity and sedentary behaviors in individuals with newly diagnosed type 2 diabetes: a cross-sectional study from the Danish nationwide DD2 cohort. Front Sports Act Living 2022;4:1089579 [FREE Full text] [doi: 10.3389/fspor.2022.1089579] [Medline: 36761371]
149. Domingos C, Picó-Pérez M, Magalhães R, Moreira M, Sousa N, Pêgo JM, et al. Free-living physical activity measured with a wearable device is associated with larger hippocampus volume and greater functional connectivity in healthy older adults: an observational, cross-sectional study in Northern Portugal. Front Aging Neurosci 2021;13:729060 [FREE Full text] [doi: 10.3389/fnagi.2021.729060] [Medline: 34916921]
150. Dougherty EN, Randall I, Haedt-Matt AA, Pila E, Smith K, Wang S, et al. Accelerometer-based physical activity and shape and weight concerns among youth with overweight and obesity: a pilot exploratory ecological momentary assessment study. Child Obes 2024 May;20(4):236-242. [doi: 10.1089/chi.2022.0236] [Medline: 37253094]
151. Douma JAJ, Verdonck-de Leeuw IM, Leemans CR, Jansen F, Langendijk JA, Baatenburg de Jong RJ, et al. Demographic, clinical and lifestyle-related correlates of accelerometer assessed physical activity and fitness in newly diagnosed patients with head and neck cancer. Acta Oncol 2020 Mar;59(3):342-350. [doi: 10.1080/0284186X.2019.1675906] [Medline: 31608747]
152. Dugas LR, Kliethermes S, Plange-Rhule J, Tong L, Bovet P, Forrester TE, et al. Accelerometer-measured physical activity is not associated with two-year weight change in African-origin adults from five diverse populations. PeerJ 2017;5:e2902 [FREE Full text] [doi: 10.7717/peerj.2902] [Medline: 28133575]
153. Edwards MK, Loprinzi PD. Associations between accelerometer-assessed sedentary behavior, physical activity and objectively-measured cardiorespiratory fitness with red blood cell distribution width. Int J Cardiol 2016 Oct 15;221:755-758. [doi: 10.1016/j.ijcard.2016.07.137] [Medline: 27428316]
154. Ehakeem A, Gregson CL, Tobias JH, Lawlor DA. Age at puberty and accelerometer-measured physical activity: findings from two independent UK cohorts. Ann Hum Biol 2020 Jul;47(4):391-399. [doi: 10.1080/03014460.2019.1707284] [Medline: 32380867]
155. Ekblom-Bak E, Börjesson M, Ekblom Ö, Angerås O, Bergman F, Berntsson C, et al. Accelerometer derived physical activity and subclinical coronary and carotid atherosclerosis: cross-sectional analyses in 22 703 middle-aged men and women in the SCAPIS study. BMJ Open 2023 Dec 23;13(11):e073380 [FREE Full text] [doi: 10.1136/bmjopen-2023-073380] [Medline: 37996228]
156. Ekblom-Bak E, Börjesson M, Bergman F, Bergström G, Dahlin-Almevall A, Drake I, Ekblom. Accelerometer derived physical activity patterns in 27.890 middle-aged adults: the SCAPIS cohort study. Scand J Med Sci Sports 2022 May;32(5):866-880 [FREE Full text] [doi: 10.1111/sms.14131] [Medline: 35080270]
157. Ekstedt M, Nyberg G, Ingre M, Ekblom Ö, Marcus C. Sleep, physical activity and BMI in six to ten-year-old children measured by accelerometry: a cross-sectional study. Int J Behav Nutr Phys Act 2013 Jul 22;10:82 [FREE Full text] [doi: 10.1186/1479-5868-10-82] [Medline: 23800204]
158. Elhakeem A, Hannam K, Deere KC, Hartley A, Clark EM, Moss C, et al. Associations of lifetime walking and weight bearing exercise with accelerometer-measured high impact physical activity in later life. Prev Med Rep 2017 Dec;8:183-189 [FREE Full text] [doi: 10.1016/j.pmedr.2017.10.011] [Medline: 29134173]
159. Evenson KR, Herring AH, Wen F. Accelerometry-assessed latent class patterns of physical activity and sedentary behavior with mortality. Am J Prev Med 2017 Feb;52(2):135-143 [FREE Full text] [doi: 10.1016/j.amepre.2016.10.033] [Medline: 28109457]
160. Evenson KR, Wen F, Herring AH, Di C, LaMonte MJ, Tinker LF, et al. Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91 years: the women's health initiative OPACH calibration study. Prev Med Rep 2015;2:750-756 [FREE Full text] [doi: 10.1016/j.pmedr.2015.08.021] [Medline: 26527313]
161. Ezeugwu V, Klaren RE, A Hubbard E, Manns P, Motl RW. Mobility disability and the pattern of accelerometer-derived sedentary and physical activity behaviors in people with multiple sclerosis. Prev Med Rep 2015;2:241-246 [FREE Full text] [doi: 10.1016/j.pmedr.2015.03.007] [Medline: 26844077]
162. Fairclough SJ, Rowlands AV, Del Pozo Cruz B, Crotti M, Foweather L, Graves LE, et al. Reference values for wrist-worn accelerometer physical activity metrics in England children and adolescents. Int J Behav Nutr Phys Act 2023 Mar 25;20(1):35 [FREE Full text] [doi: 10.1186/s12966-023-01435-z] [Medline: 36964597]

XSL•FO
RenderX

163. Farooq M, Sazonov E. A novel wearable device for food intake and physical activity recognition. Sensors 2016 Jul 11;16(7):1067. [doi: 10.3390/S16071067]
164. Feinglass J, Lee J, Semanik P, Song J, Dunlop D, Chang R. The effects of daily weather on accelerometer-measured physical activity. J Phys Act Health 2011 Oct;8(7):934-943 [FREE Full text] [doi: 10.1123/jpah.8.7.934] [Medline: 21885884]
165. Fernández-Verdejo R, Alcantara JM, Galgani JE, Acosta FM, Migueles JH, Amaro-Gahete FJ, et al. Deciphering the constrained total energy expenditure model in humans by associating accelerometer-measured physical activity from wrist and hip. Sci Rep 2021 Jul 10;11(1):12302 [FREE Full text] [doi: 10.1038/s41598-021-91750-x] [Medline: 34112912]
166. Foldager F, Jørgensen PB, Tønning LU, Petersen ET, Jakobsen SS, Vainorius D, et al. The relationship between muscle power, functional performance, accelerometer-based measurement of physical activity and patient-reported outcomes in patients with hip osteoarthritis: a cross-sectional study. Musculoskelet Sci Pract 2022 Dec;62:102678 [FREE Full text] [doi: 10.1016/j.msksp.2022.102678] [Medline: 36335851]
167. Foong YC, Chherawala N, Aitken D, Scott D, Winzenberg T, Jones G. Accelerometer-determined physical activity, muscle mass, and leg strength in community-dwelling older adults. J Cachexia Sarcopenia Muscle 2016 Jun;7(3):275-283 [FREE Full text] [doi: 10.1002/jcsm.12065] [Medline: 27239404]
168. Fox FA, Diers K, Lee H, Mayr A, Reuter M, Breteler MM, et al. Association between accelerometer-derived physical activity measurements and brain structure: a population-based cohort study. Neurology 2022 Sep 13;99(11):e1202-e1215 [FREE Full text] [doi: 10.1212/WNL.0000000000200884] [Medline: 35918154]
169. Francis SL, Simmering JE, Polgreen LA, Evans NJ, Hosteng KR, Carr LJ, et al. Gamifying accelerometer use increases physical activity levels of individuals pre-disposed to type II diabetes. Prev Med Rep 2021 Oct;23:101426 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101426] [Medline: 34178586]
170. Frehlich L, Turin TC, Doyle-Baker PK, Lang JJ, McCormack GR. Mediation analysis of the associations between neighbourhood walkability and greenness, accelerometer-measured physical activity, and health-related fitness in urban dwelling Canadians. Prev Med 2024 Jan;178:107792. [doi: 10.1016/j.ypmed.2023.107792] [Medline: 38052331]
171. Fridolfsson J, Arvidsson D, Ekblom-Bak E, Ekblom Ö, Bergström G, Börjesson M. Accelerometer-measured absolute versus relative physical activity intensity: cross-sectional associations with cardiometabolic health in midlife. BMC Public Health 2023 Dec 24;23(1):2322 [FREE Full text] [doi: 10.1186/s12889-023-17281-4] [Medline: 37996871]
172. Fridolfsson J, Börjesson M, Ekblom-Bak E, Ekblom Ö, Arvidsson D. Stronger association between high intensity physical activity and cardiometabolic health with improved assessment of the full intensity range using accelerometry. Sensors (Basel) 2020 Mar 18;20(4):1118 [FREE Full text] [doi: 10.3390/s20041118] [Medline: 32085652]
173. Frith E, Loprinzi PD. Accelerometer-assessed light-intensity physical activity and mortality among those with mobility limitations. Disabil Health J 2018 May;11(2):298-300. [doi: 10.1016/j.dhjo.2017.08.007] [Medline: 28877860]
174. Full KM, Moran K, Carlson J, Godbole S, Natarajan L, Hipp A, et al. Latent profile analysis of accelerometer-measured sleep, physical activity, and sedentary time and differences in health characteristics in adult women. PLoS One 2019;14(6):e0218595 [FREE Full text] [doi: 10.1371/journal.pone.0218595] [Medline: 31247051]
175. Full KM, Whitaker KM, Pettee Gabriel K, Lewis CE, Sternfeld B, Sidney S, et al. Cardiovascular risk and functional burden at midlife: prospective associations of isotemporal reallocations of accelerometer-measured physical activity and sedentary time in the CARDIA study. Prev Med 2021 Oct;150:106626 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106626] [Medline: 34019927]
176. Gába A, Mitáš J, Jakubec L. Associations between accelerometer-measured physical activity and body fatness in school-aged children. Environ Health Prev Med 2017 May 28;22(1):43 [FREE Full text] [doi: 10.1186/s12199-017-0629-4] [Medline: 29165128]
177. Gao N, Shao W, Salim FD. Predicting personality traits from physical activity intensity. Computer 2019 Jul;52(7):47-56. [doi: 10.1109/mc.2019.2913751]
178. Gao Y, Hua S, Mok Y, Salameh M, Qi Q, Chen G, et al. Joint associations of peripheral artery disease and accelerometry-based physical activity with mortality: the Hispanic community health study/study of Latinos (HCHS/SOL). Atherosclerosis 2022 May;347:55-62 [FREE Full text] [doi: 10.1016/j.atherosclerosis.2022.03.008] [Medline: 35334347]
179. Garay JL, Barreira TV, Wang Q, Brutsaert TD. Size at birth and accelerometer-measured physical activity or sedentary behavior in healthy term-born adults. Am J Hum Biol 2022 Jul;34(6):e23717 [FREE Full text] [doi: 10.1002/ajhb.23717] [Medline: 34978113]
180. Garnotel M, Bastian T, Romero-Ugalde HM, Maire A, Dugas J, Zahariev A, et al. Prior automatic posture and activity identification improves physical activity energy expenditure prediction from hip-worn triaxial accelerometry. J Appl Physiol (1985) 2018 Mar 01;124(3):780-790 [FREE Full text] [doi: 10.1152/japplphysiol.00556.2017] [Medline: 29191980]
181. Glass NL, Bellettiere J, Jain P, LaMonte MJ, LaCroix AZ, Women's Health Initiative. Evaluation of light physical activity measured by accelerometry and mobility disability during a 6-year follow-up in older women. JAMA Netw Open 2021 Mar 01;4(2):e210005 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0005] [Medline: 33620446]
182. Gråstén A, Huhtiniemi M, Hakonen H, Jaakkola T. Development of accelerometer-based light to vigorous physical activity in fitness profiles of school-aged children. Scand J Med Sci Sports 2021 Dec;31(12):2343-2355. [doi: 10.1111/sms.14056] [Medline: 34536979]

XSL•FO
RenderX

183. Gråstén A, Huhtiniemi M, Kolunsarka I, Jaakkola T. Developmental associations of accelerometer measured moderate-to-vigorous physical activity and sedentary time with cardiorespiratory fitness in schoolchildren. J Sci Med Sport 2022 Dec;25(11):884-889 [FREE Full text] [doi: 10.1016/j.jsams.2022.08.015] [Medline: 36137921]
184. Gupta N, Dencker-Larsen S, Lund Rasmussen C, McGregor D, Rasmussen CD, Thorsen SV, et al. The physical activity paradox revisited: a prospective study on compositional accelerometer data and long-term sickness absence. Int J Behav Nutr Phys Act 2020 Jul 20;17(1):93 [FREE Full text] [doi: 10.1186/s12966-020-00988-7] [Medline: 32690043]
185. Gupta N, Heiden M, Aadahl M, Korshøj M, Jørgensen MB, Holtermann A. What is the effect on obesity indicators from replacing prolonged sedentary time with brief sedentary bouts, standing and different types of physical activity during working days? A cross-sectional accelerometer-based study among blue-collar workers. PLoS One 2016;11(5):e0154935 [FREE Full text] [doi: 10.1371/journal.pone.0154935] [Medline: 27187777]
186. Guthrie N, Bradlyn A, Thompson SK, Yen S, Haritatos J, Dillon F, et al. Development of an accelerometer-linked online intervention system to promote physical activity in adolescents. PLoS One 2015;10(5):e0128639 [FREE Full text] [doi: 10.1371/journal.pone.0128639] [Medline: 26010359]
187. Gutiérrez-Hervás A, Cortés-Castell E, Juste-Ruíz M, Palazón-Bru A, Gil-Guillén V, Rizo-Baeza MM. Physical activity values in two- to seven-year-old children measured by accelerometer over five consecutive 24-hour days. Nutr Hosp 2018 Jul 27;35(3):527-532. [doi: 10.20960/nh.1403] [Medline: 29974757]
188. Ha L, Wakefield CE, Diaz C, Mizrahi D, Signorelli C, Yacef K, et al. Patterns of physical activity and sedentary behavior in child and adolescent cancer survivors assessed using wrist accelerometry: a cluster analysis approach. Health Informatics J 2023;29(4):14604582231212525 [FREE Full text] [doi: 10.1177/14604582231212525] [Medline: 37903362]
189. Haapala EA, Rantalainen T, Hesketh KD, Rodda CP, Duckham RL. Accelerometer-based osteogenic indices, moderate-to-vigorous and vigorous physical activity, and bone traits in adolescents. J Musculoskelet Neuronal Interact 2022 Dec 01;22(4):514-523 [FREE Full text] [Medline: 36458389]
190. Hachenberger J, Teuber Z, Li Y, Abkai L, Wild E, Lemola S. Investigating associations between physical activity, stress experience, and affective wellbeing during an examination period using experience sampling and accelerometry. Sci Rep 2023 May 31;13(1):8808 [FREE Full text] [doi: 10.1038/s41598-023-35987-8] [Medline: 37258597]
191. Hager ER, Gormley CE, Latta LW, Treuth MS, Caulfield LE, Black MM. Toddler physical activity study: laboratory and community studies to evaluate accelerometer validity and correlates. BMC Public Health 2016 Oct 06;16(1):936 [FREE Full text] [doi: 10.1186/s12889-016-3569-9] [Medline: 27600404]
192. Hagströmer M, Kwak L, Oja P, Sjöström M. A 6 year longitudinal study of accelerometer-measured physical activity and sedentary time in Swedish adults. J Sci Med Sport 2015 Oct;18(5):553-557. [doi: 10.1016/j.jsams.2014.07.012] [Medline: 25277849]
193. Hajna S, Brage S, Dalton A, Griffin SJ, Jones AP, Khaw K, et al. Cross-sectional and prospective associations between active living environments and accelerometer-assessed physical activity in the EPIC-Norfolk cohort. Health Place 2021 Jan;67:102490 [FREE Full text] [doi: 10.1016/j.healthplace.2020.102490] [Medline: 33321456]
194. Hale JL, Knell G, Swartz MD, Shiroma EJ, Ellis T, Lee I, et al. Association of Parkinson's disease status with accelerometer-derived physical activity and sedentary behavior in older women: the Women's Health Study (WHS). Prev Med Rep 2023 Oct;35:102361 [FREE Full text] [doi: 10.1016/j.pmedr.2023.102361] [Medline: 37584064]
195. Halloway S, Arfanakis K, Wilbur J, Schoeny ME, Pressler SJ. Accelerometer physical activity is associated with greater Gray matter volumes in older adults without dementia or mild cognitive impairment. J Gerontol B Psychol Sci Soc Sci 2019 Oct 15;74(7):1142-1151 [FREE Full text] [doi: 10.1093/geronb/gby010] [Medline: 29432610]
196. Hamasaki H, Noda M, Moriyama S, Yoshikawa R, Katsuyama H, Sako A, et al. Daily physical activity assessed by a triaxial accelerometer is beneficially associated with waist circumference, serum triglycerides, and insulin resistance in Japanese patients with prediabetes or untreated early type 2 diabetes. J Diabetes Res 2015;2015:526201 [FREE Full text] [doi: 10.1155/2015/526201] [Medline: 26064983]
197. Hansen BH, Kolle E, Dyrstad SM, Holme I, Anderssen SA. Accelerometer-determined physical activity in adults and older people. Med Sci Sports Exerc 2012 Feb;44(2):266-272. [doi: 10.1249/MSS.0b013e31822cb354] [Medline: 21796052]
198. Hazizi AS, Aina MB, Mohd NM, Zaitun Y, Hamid JJ, Tabata I. Accelerometer-determined physical activity level among government employees in Penang, Malaysia. Malays J Nutr 2012 Apr;18(1):57-66. [Medline: 23713230]
199. Healy GN, Winkler EA, Brakenridge CL, Reeves MM, Eakin EG. Accelerometer-derived sedentary and physical activity time in overweight/obese adults with type 2 diabetes: cross-sectional associations with cardiometabolic biomarkers. PLoS One 2015;10(3):e0119140 [FREE Full text] [doi: 10.1371/journal.pone.0119140] [Medline: 25775249]
200. Heitmann KA, Løchen M, Hopstock LA, Stylidis M, Welde B, Schirmer H, et al. Cross-sectional associations between accelerometry-measured physical activity, left atrial size, and indices of left ventricular diastolic dysfunction: the Tromsø study. Prev Med Rep 2021 Mar;21:101290 [FREE Full text] [doi: 10.1016/j.pmedr.2020.101290] [Medline: 33425668]
201. Helgadóttir B, Forsell Y, Ekblom Ö. Physical activity patterns of people affected by depressive and anxiety disorders as measured by accelerometers: a cross-sectional study. PLoS One 2015;10(1):e0115894 [FREE Full text] [doi: 10.1371/journal.pone.0115894] [Medline: 25585123]

XSL•FO
RenderX

202. Hernando C, Hernando C, Collado EJ, Panizo N, Martinez-Navarro I, Hernando B. Establishing cut-points for physical activity classification using triaxial accelerometer in middle-aged recreational marathoners. PLoS One 2018;13(8):e0202815 [FREE Full text] [doi: 10.1371/journal.pone.0202815] [Medline: 30157271]
203. Hesketh KR, Evenson KR, Stroo M, Clancy SM, Østbye T, Benjamin-Neelon SE. Physical activity and sedentary behavior during pregnancy and postpartum, measured using hip and wrist-worn accelerometers. Prev Med Rep 2018 Jul;10:337-345 [FREE Full text] [doi: 10.1016/j.pmedr.2018.04.012] [Medline: 29868389]
204. Hikihara Y, Tanaka C, Oshima Y, Ohkawara K, Ishikawa-Takata K, Tanaka S. Prediction models discriminating between nonlocomotive and locomotive activities in children using a triaxial accelerometer with a gravity-removal physical activity classification algorithm. PLoS One 2014;9(4):e94940 [FREE Full text] [doi: 10.1371/journal.pone.0094940] [Medline: 24755646]
205. Ho EC, Hawkley L, Dale W, Waite L, Huisingh-Scheetz M. Social capital predicts accelerometry-measured physical activity among older adults in the U.S.: a cross-sectional study in the National Social Life, Health, and Aging project. BMC Public Health 2018 Jul 27;18(1):804 [FREE Full text] [doi: 10.1186/s12889-018-5664-6] [Medline: 29945588]
206. Hooker SA, Masters KS. Purpose in life is associated with physical activity measured by accelerometer. J Health Psychol 2016 Jun 07;21(6):962-971. [doi: 10.1177/1359105314542822] [Medline: 25104777]
207. Hooker SP, Diaz KM, Blair SN, Colabianchi N, Hutto B, McDonnell MN, et al. Association of accelerometer-measured sedentary time and physical activity with risk of stroke among US adults. JAMA Netw Open 2022 Jun 01;5(6):e2215385 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.15385] [Medline: 35657625]
208. Hossain SS, Lazar DM, Begum M. Ordinal statistical models of physical activity levels from accelerometer data. Int J Exerc Sci 2021;14(7):338-357 [FREE Full text] [Medline: 34055179]
209. Hsiao C, Wen C, Yen H, Hsueh M, Liao Y. Association between accelerometer-measured light-intensity physical activity and cognitive function in older adults. J Nutr Health Aging 2022;26(3):230-235 [FREE Full text] [doi: 10.1007/s12603-022-1749-0] [Medline: 35297464]
210. Hughes RL, Pindus DM, Khan NA, Burd NA, Holscher HD. Associations between accelerometer-measured physical activity and fecal microbiota in adults with overweight and obesity. Med Sci Sports Exerc 2023 May 01;55(4):680-689 [FREE Full text] [doi: 10.1249/MSS.0000000000003096] [Medline: 36728974]
211. Huisingh-Scheetz M, Wroblewski K, Kocherginsky M, Huang E, Dale W, Waite L, et al. The relationship between physical activity and frailty among U.S. older adults based on hourly accelerometry data. J Gerontol A Biol Sci Med Sci 2018 Apr 17;73(5):622-629 [FREE Full text] [doi: 10.1093/gerona/glx208] [Medline: 29106478]
212. Hyde ET, LaCroix AZ, Evenson KR, Howard AG, Anuskiewicz B, Di C, et al. Accelerometer-measured physical activity and postmenopausal breast cancer incidence in the women's health accelerometry collaboration. Cancer 2023 May 15;129(10):1579-1590 [FREE Full text] [doi: 10.1002/cncr.34699] [Medline: 36812131]
213. Hylkema TH, Brouwer S, Kooijman CM, De Vries AJ, Breukelman F, Dekker H, et al. Accelerometer measured sedentary and physical activity behaviors of working patients after total knee arthroplasty, and their compensation between occupational and leisure time. J Occup Rehabil 2021 Jul;31(2):350-359 [FREE Full text] [doi: 10.1007/s10926-020-09924-9] [Medline: 32946009]
214. Ikeda E, Guagliano JM, Atkin AJ, Sherar LB, Ekelund U, Hansen B, International Children's Accelerometry Database (ICAD) Collaborators. Cross-sectional and longitudinal associations of active travel, organised sport and physical education with accelerometer-assessed moderate-to-vigorous physical activity in young people: the international children's accelerometry database. Int J Behav Nutr Phys Act 2022 May 02;19(1):41 [FREE Full text] [doi: 10.1186/s12966-022-01282-4] [Medline: 35366914]
215. Ishikawa-Takata K, Kaneko K, Koizumi K, Ito C. Comparison of physical activity energy expenditure in Japanese adolescents assessed by EW4800P triaxial accelerometry and the doubly labelled water method. Br J Nutr 2013 Oct;110(7):1347-1355. [doi: 10.1017/S0007114513000603] [Medline: 23544366]
216. Izawa KP, Watanabe S, Hiraki K, Morio Y, Kasahara Y, Takeichi N, et al. Determination of the effectiveness of accelerometer use in the promotion of physical activity in cardiac patients: a randomized controlled trial. Arch Phys Med Rehabil 2012 Nov;93(11):1896-1902. [doi: 10.1016/j.apmr.2012.06.015] [Medline: 22750166]
217. Jaeschke L, Steinbrecher A, Boeing H, Gastell S, Ahrens W, Berger K, et al. Factors associated with habitual time spent in different physical activity intensities using multiday accelerometry. Sci Rep 2020 Jan 21;10(1):774 [FREE Full text] [doi: 10.1038/s41598-020-57648-w] [Medline: 31964962]
218. Jago R, Salway R, House D, Walker R, Emm-Collison L, Sansum K, et al. Short and medium-term effects of the COVID-19 lockdowns on child and parent accelerometer-measured physical activity and sedentary time: a natural experiment. Int J Behav Nutr Phys Act 2023 May 27;20(1):42 [FREE Full text] [doi: 10.1186/s12966-023-01441-1] [Medline: 37101270]
219. Jain RK, Vokes T. Physical activity as measured by accelerometer in NHANES 2005-2006 is associated with better bone density and trabecular bone score in older adults. Arch Osteoporos 2019 Mar 02;14(1):29. [doi: 10.1007/s11657-019-0583-4] [Medline: 30826896]
220. Janney CA, Fagiolini A, Swartz HA, Jakicic JM, Holleman RG, Richardson CR. Are adults with bipolar disorder active? Objectively measured physical activity and sedentary behavior using accelerometry. J Affect Disord 2014 Jan;152-154:498-504 [FREE Full text] [doi: 10.1016/j.jad.2013.09.009] [Medline: 24095103]

XSL•FO
RenderX

221. Jeng B, Cederberg KL, Lai B, Sasaki JE, Bamman MM, Motl RW. Wrist-based accelerometer cut-points for quantifying moderate-to-vigorous intensity physical activity in Parkinson's disease. Gait Posture 2022 Jan;91:235-239 [FREE Full text] [doi: 10.1016/j.gaitpost.2021.10.027] [Medline: 34749075]
222. Jeran S, Steinbrecher A, Haas V, Mähler A, Boschmann M, Westerterp KR, et al. Prediction of activity-related energy expenditure under free-living conditions using accelerometer-derived physical activity. Sci Rep 2022 Oct 04;12(1):16578 [FREE Full text] [doi: 10.1038/s41598-022-20639-0] [Medline: 36195647]
223. Jin X, Chen Y, Feng H, Zhou M, Chan JW, Liu Y, et al. Association of accelerometer-measured sleep duration and different intensities of physical activity with incident type 2 diabetes in a population-based cohort study. J Sport Health Sci 2024 Mar;13(2):222-232 [FREE Full text] [doi: 10.1016/j.jshs.2023.03.001] [Medline: 36871624]
224. Kim J, Nakamura T, Kikuchi H, Yoshiuchi K, Yamamoto Y. Co-variation of depressive mood and spontaneous physical activity evaluated by ecological momentary assessment in major depressive disorder. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:6635-6638. [doi: 10.1109/EMBC.2014.6945149] [Medline: 25571517]
225. Júdice PB, Hetherington-Rauth M, Northstone K, Andersen LB, Wedderkopp N, Ekelund U, International Children's Accelerometry Database (ICAD) Collaborators. Changes in physical activity and sedentary patterns on cardiometabolic outcomes in the transition to adolescence: international children's accelerometry database 2.0. J Pediatr 2020 Oct;225:166-73.e1. [doi: 10.1016/j.jpeds.2020.06.018] [Medline: 32553870]
226. Jung ME, Locke SR, Bourne JE, Beauchamp MR, Lee T, Singer J, et al. Cardiorespiratory fitness and accelerometer-determined physical activity following one year of free-living high-intensity interval training and moderate-intensity continuous training: a randomized trial. Int J Behav Nutr Phys Act 2020 Feb 26;17(1):25 [FREE Full text] [doi: 10.1186/s12966-020-00933-8] [Medline: 32102667]
227. Junno J, Keisu A, Niemelä M, Modarress Julin M, Korpelainen R, Jämsä T, et al. Accelerometer-measured physical activity is associated with knee breadth in middle-aged Finns - a population-based study. BMC Musculoskelet Disord 2022 May 31;23(1):517 [FREE Full text] [doi: 10.1186/s12891-022-05475-7] [Medline: 35642051]
228. Junttila HE, Vaaramo MM, Huikari SM, Kari JT, Leinonen A, Farrahi V, et al. Association of accelerometer-measured physical activity and midlife income: a Northern Finland Birth Cohort 1966 study. Scand J Med Sci Sports 2023 Oct;33(9):1765-1778. [doi: 10.1111/sms.14421] [Medline: 37272147]
229. Jussila A, Husu P, Vähä-Ypyä H, Tokola K, Kokko S, Sievänen H, et al. Accelerometer-measured physical activity levels and patterns vary in an age- and sex-dependent fashion among Finnish children and adolescents. Int J Environ Res Public Health 2022 Jul 06;19(11):6950 [FREE Full text] [doi: 10.3390/ijerph19116950] [Medline: 35682533]
230. Kanai M, Izawa KP, Kobayashi M, Onishi A, Kubo H, Nozoe M, et al. Effect of accelerometer-based feedback on physical activity in hospitalized patients with ischemic stroke: a randomized controlled trial. Clin Rehabil 2018 Aug;32(8):1047-1056. [doi: 10.1177/0269215518755841] [Medline: 29400070]
231. Kantomaa MT, Tikanmäki M, Kankaanpää A, Vääräsmäki M, Sipola-Leppänen M, Ekelund U, et al. Accelerometer-measured physical activity and sedentary time differ according to education level in young adults. PLoS One 2016;11(7):e0158902 [FREE Full text] [doi: 10.1371/journal.pone.0158902] [Medline: 27403958]
232. Ketels M, Rasmussen CL, Korshøj M, Gupta N, De Bacquer D, Holtermann A, et al. The relation between domain-specific physical behaviour and cardiorespiratory fitness: a cross-sectional compositional data analysis on the physical activity health paradox using accelerometer-assessed data. Int J Environ Res Public Health 2020 Oct 29;17(21):7929 [FREE Full text] [doi: 10.3390/ijerph17217929] [Medline: 33137943]
233. Khurshid S, Al-Alusi MA, Churchill TW, Guseh JS, Ellinor PT. Accelerometer-derived "weekend warrior" physical activity and incident cardiovascular disease. JAMA 2023 Jul 18;330(3):247-252. [doi: 10.1001/jama.2023.10875] [Medline: 37462704]
234. Khurshid S, Weng L, Al-Alusi MA, Halford JL, Haimovich JS, Benjamin EJ, et al. Accelerometer-derived physical activity and risk of atrial fibrillation. Eur Heart J 2021 Jul 01;42(25):2472-2483 [FREE Full text] [doi: 10.1093/eurheartj/ehab250] [Medline: 34037209]
235. Khurshid S, Weng L, Nauffal V, Pirruccello JP, Venn RA, Al-Alusi MA, et al. Wearable accelerometer-derived physical activity and incident disease. NPJ Digit Med 2022 Sep 02;5(1):131 [FREE Full text] [doi: 10.1038/s41746-022-00676-9] [Medline: 36056190]
236. Kijima T, Akai K, Amagasa S, Inoue S, Yamagata S, Ishibashi Y, et al. Accelerometer-measured physical activity and posture among older adults in assisted-living residences. SAGE Open Med 2024;12:20503121231220798 [FREE Full text] [doi: 10.1177/20503121231220798] [Medline: 38186563]
237. Kim Y, Wijndaele K, Sharp SJ, Strain T, Pearce M, White T, et al. Specific physical activities, sedentary behaviours and sleep as long-term predictors of accelerometer-measured physical activity in 91,648 adults: a prospective cohort study. Int J Behav Nutr Phys Act 2019 May 07;16(1):41 [FREE Full text] [doi: 10.1186/s12966-019-0802-9] [Medline: 31064403]
238. Kinnunen TI, Tennant PW, McParlin C, Poston L, Robson SC, Bell R. Agreement between pedometer and accelerometer in measuring physical activity in overweight and obese pregnant women. BMC Public Health 2011 Jul 27;11:501 [FREE Full text] [doi: 10.1186/1471-2458-11-501] [Medline: 21703033]

XSL•FO
RenderX

239. Kong S, Park HY, Kang D, Lee JK, Lee G, Kwon OJ, et al. Seasonal variation in physical activity among preoperative patients with lung cancer determined using a wearable device. J Clin Med 2020 Jan 27;9(2):349 [FREE Full text] [doi: 10.3390/jcm9020349] [Medline: 32012720]
240. Kretzschmar M, Lin W, Nardo L, Joseph GB, Dunlop DD, Heilmeier U, et al. Association of physical activity measured by accelerometer, knee joint abnormalities, and cartilage T2 measurements obtained from 3T magnetic resonance imaging: data from the osteoarthritis initiative. Arthritis Care Res (Hoboken) 2015 Oct;67(9):1272-1280 [FREE Full text] [doi: 10.1002/acr.22586] [Medline: 25777255]
241. Kruisdijk F, Deenik J, Tenback D, Tak E, Beekman A, van Harten P, et al. Accelerometer-measured sedentary behaviour and physical activity of inpatients with severe mental illness. Psychiatry Res 2017 Aug;254:67-74. [doi: 10.1016/j.psychres.2017.04.035] [Medline: 28456024]
242. Kulinski JP, Khera A, Ayers CR, Das SR, de Lemos JA, Blair SN, et al. Association between cardiorespiratory fitness and accelerometer-derived physical activity and sedentary time in the general population. Mayo Clin Proc 2014 Aug;89(8):1063-1071 [FREE Full text] [doi: 10.1016/j.mayocp.2014.04.019] [Medline: 25012770]
243. Kumar D, Thanikkal A, Krishnamurthy P, Chen X, Zhang P. Accelerometer-based alcohol consumption detection from physical activity. In: Proceedings of the 2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications. 2021 Presented at: WiMob '21; October 11-13, 2021; Bologna, Italy p. 415-418 URL: https://ieeexplore.ieee.org/document/9606257 [doi: 10.1109/wimob52687.2021.9606257]
244. Kuritz A, Mall C, Schnitzius M, Mess F. Physical activity and sedentary behavior of children in afterschool programs: an accelerometer-based analysis in full-day and half-day elementary schools in Germany. Front Public Health 2020;8:463 [FREE Full text] [doi: 10.3389/fpubh.2020.00463] [Medline: 32984249]
245. Kwon S, Andersen LB, Grøntved A, Kolle E, Cardon G, Davey R, ICAD Collaborators. A closer look at the relationship among accelerometer-based physical activity metrics: ICAD pooled data. Int J Behav Nutr Phys Act 2019 May 29;16(1):40 [FREE Full text] [doi: 10.1186/s12966-019-0801-x] [Medline: 31036032]
246. Kwon S, Janz KF, International Children's Accelerometry Database (ICAD) Collaborators. Tracking of accelerometry-measured physical activity during childhood: ICAD pooled analysis. Int J Behav Nutr Phys Act 2012 Jun 07;9:68 [FREE Full text] [doi: 10.1186/1479-5868-9-68] [Medline: 22676230]
247. LaCroix AZ, Bellettiere J, Rillamas-Sun E, Di C, Evenson KR, Lewis CE, Women's Health Initiative (WHI). Association of light physical activity measured by accelerometry and incidence of coronary heart disease and cardiovascular disease in older women. JAMA Netw Open 2019 Mar 01;2(3):e190419 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0419] [Medline: 30874775]
248. LaMonte MJ, LaCroix AZ, Nguyen S, Evenson KR, Di C, Stefanick ML, et al. Accelerometer-measured physical activity, sedentary time, and heart failure risk in women aged 63 to 99 years. JAMA Cardiol 2024 May 01;9(4):336-345. [doi: 10.1001/jamacardio.2023.5692] [Medline: 38381446]
249. LaMonte MJ, Lewis CE, Buchner DM, Evenson KR, Rillamas-Sun E, Di C, et al. Both light intensity and moderate-to-vigorous physical activity measured by accelerometry are favorably associated with cardiometabolic risk factors in older women: the objective physical activity and cardiovascular health (OPACH) study. J Am Heart Assoc 2017 Oct 17;6(10):e007064 [FREE Full text] [doi: 10.1161/JAHA.117.007064] [Medline: 29042429]
250. Länsitie M, Kangas M, Jokelainen J, Venojärvi M, Vaaramo E, Härkönen P, et al. Association between accelerometer-measured physical activity, glucose metabolism, and waist circumference in older adults. Diabetes Res Clin Pract 2021 Aug;178:108937 [FREE Full text] [doi: 10.1016/j.diabres.2021.108937] [Medline: 34217770]
251. Länsitie M, Kangas M, Jokelainen J, Venojärvi M, Timonen M, Keinänen-Kiukaanniemi S, et al. Cardiovascular disease risk and all-cause mortality associated with accelerometer-measured physical activity and sedentary time a prospective population-based study in older adults. BMC Geriatr 2022 Oct 05;22(1):729 [FREE Full text] [doi: 10.1186/s12877-022-03414-8] [Medline: 36064345]
252. Lawman HG, Horn ML, Wilson DK, Pate RR. A multilevel approach to examining time-specific effects in accelerometer-assessed physical activity. J Sci Med Sport 2015 Dec;18(6):667-672 [FREE Full text] [doi: 10.1016/j.jsams.2014.09.003] [Medline: 25245427]
253. Le Cornu Q, Chen M, van Hees V, Léger D, Fayosse A, Yerramalla MS, et al. Association of physical activity, sedentary behaviour, and daylight exposure with sleep in an ageing population: findings from the Whitehall accelerometer sub-study. Int J Behav Nutr Phys Act 2022 Dec 09;19(1):144 [FREE Full text] [doi: 10.1186/s12966-022-01391-0] [Medline: 36494722]
254. Lee I, Shiroma EJ, Evenson KR, Kamada M, LaCroix AZ, Buring JE. Accelerometer-measured physical activity and sedentary behavior in relation to all-cause mortality: the women's health study. Circulation 2018 Dec 09;137(2):203-205 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.117.031300] [Medline: 29109088]
255. Lee J, Walker ME, Gabriel KP, Vasan RS, Xanthakis V. Associations of accelerometer-measured physical activity and sedentary time with chronic kidney disease: The Framingham Heart Study. PLoS One 2020;15(6):e0234825 [FREE Full text] [doi: 10.1371/journal.pone.0234825] [Medline: 32542048]
256. Lee PH. Examining non-linear associations between accelerometer-measured physical activity, sedentary behavior, and all-cause mortality using segmented cox regression. Front Physiol 2016;7:272 [FREE Full text] [doi: 10.3389/fphys.2016.00272] [Medline: 27445859]

XSL•FO
RenderX

257. Lee S, Kim SC, Gim J, Park SJ, Seo SH, Kim SJ, et al. Accelerometer-derived physical activity analysis of elderly osteoarthritis patients. Musculoskelet Sci Pract 2023 Aug;66:102808. [doi: 10.1016/j.msksp.2023.102808] [Medline: 37352763]
258. Leininger B, Schulz C, Gao Z, Bronfort G, Evans R, Pope Z, et al. Accelerometer-determined physical activity and clinical low back pain measures in adolescents with chronic or subacute recurrent low back pain. J Orthop Sports Phys Ther 2017 Oct;47(10):769-774 [FREE Full text] [doi: 10.2519/jospt.2017.7345] [Medline: 28898136]
259. Lemberg GM, Riso E, Fjørtoft I, Kjønniksen L, Kull M, Mäestu E. School children's physical activity and preferred activities during outdoor recess in Estonia: using accelerometers, recess observation, and schoolyard mapping. Children (Basel) 2023 May 10;10(4):702 [FREE Full text] [doi: 10.3390/children10040702] [Medline: 37189951]
260. Li L, Nakamura T. An epidemiological sleep study based on a large-scale physical activity database. In: Proceedings of the 2019 IEEE 1st Global Conference on Life Sciences and Technologies. 2019 Presented at: LifeTech '19; March 12-14, 2019; Osaka, Japan p. 292-293 URL: https://ieeexplore.ieee.org/document/8883989 [doi: 10.1109/lifetech.2019.8883989]
261. Li S, Howard JT, Sosa ET, Cordova A, Parra-Medina D, Yin Z. Calibrating wrist-worn accelerometers for physical activity assessment in preschoolers: machine learning approaches. JMIR Form Res 2020 Aug 31;4(8):e16727 [FREE Full text] [doi: 10.2196/16727] [Medline: 32667893]
262. Li T, Zong G, Peng P, Wang S, Cheng B. Accelerometer-measured physical activity and sample-based frailty in older women: does pattern really matter? Front Public Health 2023;11:1304279 [FREE Full text] [doi: 10.3389/fpubh.2023.1304279] [Medline: 38332942]
263. Li X, Kearney PM, Keane E, Harrington JM, Fitzgerald AP. levels and sociodemographic correlates of accelerometer-based physical activity in Irish children: a cross-sectional study. J Epidemiol Community Health 2017 Jul;71(6):521-527. [doi: 10.1136/jech-2016-207691] [Medline: 28130391]
264. Liang YY, Feng H, Chen Y, Jin X, Xue H, Zhou M, et al. Joint association of physical activity and sleep duration with risk of all-cause and cause-specific mortality: a population-based cohort study using accelerometry. Eur J Prev Cardiol 2023 Jul 12;30(9):832-843. [doi: 10.1093/eurjpc/zwad060] [Medline: 36990109]
265. Lin C, Lin K, Hsueh M, Liao Y. Associations of accelerometer-measured sedentary behavior and physical activity with sleep in older adults. J Formos Med Assoc (Forthcoming) 2023 Aug 14 [FREE Full text] [doi: 10.1016/j.jfma.2023.08.001] [Medline: 37586972]
266. Lin H, Hartley P, Forsyth F, Pilling M, Hobbs FD, Taylor CJ, et al. Clinical and demographic correlates of accelerometer-measured physical activity in participants enrolled in the OPTIMISE HFpEF study. Eur J Cardiovasc Nurs 2022 Jan 11;21(1):67-75. [doi: 10.1093/eurjcn/zvab028] [Medline: 33837414]
267. Lin H, Lynk N, Moore LL, Cabral HJ, Heffernan KS, Dumas AK, et al. A pragmatic approach to the comparison of wrist-based cutpoints of physical activity intensity for the MotionWatch8 accelerometer in children. PLoS One 2020;15(6):e0234725 [FREE Full text] [doi: 10.1371/journal.pone.0234725] [Medline: 32559207]
268. Liu M, Gan X, Ye Z, Zhang Y, He P, Zhou C, et al. Association of accelerometer-measured physical activity intensity, sedentary time, and exercise time with incident Parkinson's disease. NPJ Digit Med 2023 Nov 28;6(1):224 [FREE Full text] [doi: 10.1038/s41746-023-00969-7] [Medline: 38017114]
269. Liu M, Zhang Y, Zhang Y, He P, Zhou C, Ye Z, et al. Association of accelerometer-measured physical activity and its change with progression to chronic kidney disease in adults with type 2 diabetes and overweight/obesity. Br J Sports Med 2024 Mar 13;58(6):313-319. [doi: 10.1136/bjsports-2023-107564] [Medline: 38320851]
270. Long MT, Pedley A, Massaro JM, Hoffmann U, Esliger DW, Vasan RS, et al. Hepatic steatosis is associated with lower levels of physical activity measured via accelerometry. Obesity (Silver Spring) 2015 Jul;23(6):1259-1266 [FREE Full text] [doi: 10.1002/oby.21058] [Medline: 25959049]
271. Loprinzi PD. Accelerometer-determined physical activity and mortality in a national prospective cohort study of adults at high risk of a first atherosclerotic cardiovascular disease event. Int J Cardiol 2016 Jan 01;202:417-418. [doi: 10.1016/j.ijcard.2015.09.061] [Medline: 26432494]
272. Loprinzi PD. Accelerometer-determined physical activity and all-cause mortality in a national prospective cohort study of hypertensive adults. J Hypertens 2016 May;34(5):848-852. [doi: 10.1097/HJH.0000000000000869] [Medline: 26828782]
273. Loprinzi PD, Addoh O. Accelerometer-determined physical activity and all-cause mortality in a national prospective cohort study of adults post-acute stroke. Am J Health Promot 2018 Jan;32(1):24-27. [doi: 10.1177/0890117117720061] [Medline: 28718295]
274. Loprinzi PD, Frith E. Accelerometer-assessed physical activity and school absenteeism due to illness or injury among children and adolescents: NHANES 2003 to 2006. Am J Health Promot 2018 Mar;32(3):571-577. [doi: 10.1177/0890117116684241] [Medline: 29431510]
275. Loprinzi PD, Gilham B, Cardinal BJ. Association between accelerometer-assessed physical activity and objectively measured hearing sensitivity among U.S. adults with diabetes. Res Q Exerc Sport 2014 Oct;85(3):390-397. [doi: 10.1080/02701367.2014.930404] [Medline: 25141090]
276. Loprinzi PD, Joyner C. Accelerometer-determined physical activity and mortality in a national prospective cohort study: considerations by visual acuity. Prev Med 2016 Jul;87:18-21. [doi: 10.1016/j.ypmed.2016.02.005] [Medline: 26861750]

XSL•FO
RenderX

277. Loprinzi PD, Lee H, Gilham B, Cardinal BJ. Association between accelerometer-assessed physical activity and tinnitus, NHANES 2005-2006. Res Q Exerc Sport 2013 Jun;84(2):177-185. [doi: 10.1080/02701367.2013.784840] [Medline: 23930543]
278. Loprinzi PD, Sheffield J, Tyo BM, Fittipaldi-Wert J. Accelerometer-determined physical activity, mobility disability, and health. Disabil Health J 2014 Oct;7(4):419-425. [doi: 10.1016/j.dhjo.2014.05.005] [Medline: 25224982]
279. Loprinzi PD, Smit E, Lin FR, Gilham B, Ramulu PY. Accelerometer-assessed physical activity and objectively determined dual sensory impairment in US adults. Mayo Clin Proc 2013 Jul;88(7):690-696. [doi: 10.1016/j.mayocp.2013.04.008] [Medline: 23751983]
280. Loprinzi P, Smit E, Lee H, Crespo C, Andersen R, Blair SN. The "fit but fat" paradigm addressed using accelerometer-determined physical activity data. N Am J Med Sci 2014 Jul;6(7):295-301 [FREE Full text] [doi: 10.4103/1947-2714.136901] [Medline: 25077076]
281. Loyen A, Clarke-Cornwell AM, Anderssen SA, Hagströmer M, Sardinha LB, Sundquist K, et al. Sedentary time and physical activity surveillance through accelerometer pooling in four European countries. Sports Med 2017 Jul 10;47(7):1421-1435 [FREE Full text] [doi: 10.1007/s40279-016-0658-y] [Medline: 27943147]
282. Lund K, Larsen MD, Knudsen T, Kjeldsen J, Nielsen RG, Brage S, et al. Physical activity measured by accelerometry in paediatric and young adult patients with inflammatory bowel disease. BMC Gastroenterol 2022 Jul 07;22(1):290 [FREE Full text] [doi: 10.1186/s12876-022-02358-y] [Medline: 35672666]
283. Lund Rasmussen C, Palarea-Albaladejo J, Korshøj M, Gupta N, Nabe-Nielsen K, Holtermann A, et al. Is high aerobic workload at work associated with leisure time physical activity and sedentary behaviour among blue-collar workers? A compositional data analysis based on accelerometer data. PLoS One 2019;14(6):e0217024 [FREE Full text] [doi: 10.1371/journal.pone.0217024] [Medline: 31170169]
284. Ma J, Kim H, Kim J. Isotemporal substitution analysis of accelerometer-derived sedentary behavior and physical activity on cardiometabolic health in Korean adults: a population-based cross-sectional study. Int J Environ Res Public Health 2021 Oct 22;18(21):11102 [FREE Full text] [doi: 10.3390/ijerph182111102] [Medline: 34769622]
285. Machado-Rodrigues AM, Rodrigues D, Gama A, Nogueira H, Silva MG, Mascarenhas LP, et al. Tri-axial accelerometer-assessed physical activity and its association with weight status in a sample of elementary-school children. Obes Res Clin Pract 2023;17(3):192-197. [doi: 10.1016/j.orcp.2023.04.003] [Medline: 37087316]
286. Madimenos FC, Snodgrass JJ, Blackwell AD, Liebert MA, Sugiyama LS. Physical activity in an indigenous Ecuadorian forager-horticulturalist population as measured using accelerometry. Am J Hum Biol 2011;23(4):488-497 [FREE Full text] [doi: 10.1002/ajhb.21163] [Medline: 21538650]
287. Maes I, Van Dyck D, Van Cauwenberg J, Mertens L. Age-related differences in the associations of physical environmental factors and psychosocial factors with accelerometer-assessed physical activity. Health Place 2021 Jan;67:102492. [doi: 10.1016/j.healthplace.2020.102492] [Medline: 33316601]
288. Malek ME, Norman Å, Elinder LS, Patterson E, Nyberg G. Relationships between physical activity parenting practices and children's activity measured by accelerometry with children's activity style as a moderator-a cross sectional study. Children (Basel) 2022 Mar 12;9(2):248 [FREE Full text] [doi: 10.3390/children9020248] [Medline: 35204968]
289. Mañas A, Pozo-Cruz BD, Rodríguez-Gómez I, Losa-Reyna J, Rodríguez-Mañas L, García-García FJ, et al. Can physical activity offset the detrimental consequences of sedentary time on frailty? A moderation analysis in 749 older adults measured with accelerometers. J Am Med Dir Assoc 2019 May;20(5):634-8.e1. [doi: 10.1016/j.jamda.2018.12.012] [Medline: 30738823]
290. Manglani HR, Phansikar M, Duraney EJ, McKenna MR, Canter R, Nicholas JA, et al. Accelerometry measures of physical activity and sedentary behavior: associations with cognitive functioning in MS. Mult Scler Relat Disord 2023 Dec;79:104963. [doi: 10.1016/j.msard.2023.104963] [Medline: 37690438]
291. Manns P, Ezeugwu V, Armijo-Olivo S, Vallance J, Healy GN. Accelerometer-derived pattern of sedentary and physical activity time in persons with mobility disability: national health and nutrition examination survey 2003 to 2006. J Am Geriatr Soc 2015 Jul;63(7):1314-1323. [doi: 10.1111/jgs.13490] [Medline: 26173621]
292. Maslow AL, Colabianchi N. Youth physical activity resource use and activity measured by accelerometry. Am J Health Behav 2011;35(2):219-227 [FREE Full text] [doi: 10.5993/ajhb.35.2.9] [Medline: 21204684]
293. Matabuena M, Félix P, Hammouri ZA, Mota J, Del Pozo Cruz B. Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES. Aging Clin Exp Res 2022 Dec;34(12):3107-3114 [FREE Full text] [doi: 10.1007/s40520-022-02260-3] [Medline: 36183279]
294. Matsuzawa R, Matsunaga A, Kutsuna T, Ishii A, Abe Y, Yoneki K, et al. Association of habitual physical activity measured by an accelerometer with high-density lipoprotein cholesterol levels in maintenance hemodialysis patients. ScientificWorldJournal 2013;2013:780783 [FREE Full text] [doi: 10.1155/2013/780783] [Medline: 24453904]
295. Matthews CE, Keadle SK, Troiano RP, Kahle L, Koster A, Brychta R, et al. Accelerometer-measured dose-response for physical activity, sedentary time, and mortality in US adults. Am J Clin Nutr 2016 Nov;104(5):1424-1432 [FREE Full text] [doi: 10.3945/ajcn.116.135129] [Medline: 27707702]

XSL•FO
RenderX

296. Maylor BD, Edwardson CL, Clarke-Cornwell AM, Davies MJ, Dawkins NP, Dunstan DW, et al. Physical activity assessed by wrist and thigh worn accelerometry and associations with cardiometabolic health. Sensors (Basel) 2023 Aug 23;23(17):7353 [FREE Full text] [doi: 10.3390/s23177353] [Medline: 37687813]
297. McDonald L, Oguz M, Carroll R, Thakkar P, Yang F, Dhalwani N, et al. Comparison of accelerometer-derived physical activity levels between individuals with and without cancer: a UK Biobank study. Future Oncol 2019 Dec;15(33):3763-3774. [doi: 10.2217/fon-2019-0443] [Medline: 31637942]
298. McGarty AM, Penpraze V, Melville CA. Calibration and cross-validation of the ActiGraph wGT3X+ accelerometer for the estimation of physical activity intensity in children with intellectual disabilities. PLoS One 2016;11(10):e0164928 [FREE Full text] [doi: 10.1371/journal.pone.0164928] [Medline: 27760219]
299. McLellan G, Arthur R, Donnelly S, Buchan DS. Segmented sedentary time and physical activity patterns throughout the week from wrist-worn ActiGraph GT3X+ accelerometers among children 7-12 years old. J Sport Health Sci 2020 Mar;9(2):179-188 [FREE Full text] [doi: 10.1016/j.jshs.2019.02.005] [Medline: 32099726]
300. Melin M, Hagerman I, Gonon A, Gustafsson T, Rullman E. Variability in physical activity assessed with accelerometer is an independent predictor of mortality in CHF patients. PLoS One 2016;11(4):e0153036 [FREE Full text] [doi: 10.1371/journal.pone.0153036] [Medline: 27054323]
301. Memari AH, Ghaheri B, Ziaee V, Kordi R, Hafizi S, Moshayedi P. Physical activity in children and adolescents with autism assessed by triaxial accelerometry. Pediatr Obes 2013 May;8(2):150-158. [doi: 10.1111/j.2047-6310.2012.00101.x] [Medline: 23042790]
302. Micklesfield LK, Westgate K, Smith A, Kufe CN, Mendham AE, Lindsay T, et al. Physical activity behaviors of a middle-age South African cohort as determined by integrated hip and thigh accelerometry. Med Sci Sports Exerc 2022 Oct 01;54(9):1493-1505 [FREE Full text] [doi: 10.1249/MSS.0000000000002940] [Medline: 35658390]
303. Midorikawa T, Tanaka S, Kaneko K, Koizumi K, Ishikawa-Takata K, Futami J, et al. Evaluation of low-intensity physical activity by triaxial accelerometry. Obesity (Silver Spring) 2007 Dec;15(12):3031-3038 [FREE Full text] [doi: 10.1038/oby.2007.361] [Medline: 18198312]
304. Miller NE, Strath SJ, Swartz AM, Cashin SE. Estimating absolute and relative physical activity intensity across age via accelerometry in adults. J Aging Phys Act 2010 Apr;18(2):158-170 [FREE Full text] [doi: 10.1123/japa.18.2.158] [Medline: 20440028]
305. Mossavar-Rahmani Y, Lin J, Pan S, Song RJ, Xue X, Spartano NL, et al. Characterizing longitudinal change in accelerometry-based moderate-to-vigorous physical activity in the Hispanic community health study/study of Latinos and the Framingham heart study. BMC Public Health 2023 Aug 24;23(1):1614 [FREE Full text] [doi: 10.1186/s12889-023-16442-9] [Medline: 37620824]
306. Moy ML, Matthess K, Stolzmann K, Reilly J, Garshick E. Free-living physical activity in COPD: assessment with accelerometer and activity checklist. J Rehabil Res Dev 2009;46(2):277-286 [FREE Full text] [doi: 10.1682/jrrd.2008.07.0083] [Medline: 19533541]
307. Murabito JM, Pedley A, Massaro JM, Vasan RS, Esliger D, Blease SJ, et al. Moderate-to-vigorous physical activity with accelerometry is associated with visceral adipose tissue in adults. J Am Heart Assoc 2015 Mar 03;4(3):e001379 [FREE Full text] [doi: 10.1161/JAHA.114.001379] [Medline: 25736442]
308. Nagayoshi S, Oshima Y, Ando T, Aoyama T, Nakae S, Usui C, et al. Validity of estimating physical activity intensity using a triaxial accelerometer in healthy adults and older adults. BMJ Open Sport Exerc Med 2019;5(1):e000592 [FREE Full text] [doi: 10.1136/bmjsem-2019-000592] [Medline: 31749982]
309. Nakabazzi B, Wachira LM, Oyeyemi AL, Ssenyonga R, Onywera VO. Parental-perceived home and neighborhood environmental correlates of accelerometer-measured physical activity among school-going children in Uganda. PLOS Glob Public Health 2021;1(12):e0000089 [FREE Full text] [doi: 10.1371/journal.pgph.0000089] [Medline: 36962124]
310. Nam S, Jeon S, Ash G, Whittemore R, Vlahov D. Racial discrimination, sedentary time, and physical activity in African Americans: quantitative study combining ecological momentary assessment and accelerometers. JMIR Form Res 2021 Jul 07;5(6):e25687 [FREE Full text] [doi: 10.2196/25687] [Medline: 34096870]
311. Neil-Sztramko SE, Rafn BS, Gotay CC, Campbell KL. Determining activity count cut-points for measurement of physical activity using the Actiwatch2 accelerometer. Physiol Behav 2017 May 01;173:95-100. [doi: 10.1016/j.physbeh.2017.01.026] [Medline: 28108333]
312. Nero H, Benka Wallén M, Franzén E, Ståhle A, Hagströmer M. Accelerometer cut points for physical activity assessment of older adults with Parkinson's disease. PLoS One 2015;10(9):e0135899 [FREE Full text] [doi: 10.1371/journal.pone.0135899] [Medline: 26332765]
313. Nguyen S, LaCroix AZ, Hayden KM, Di C, Palta P, Stefanick ML, et al. Accelerometer-measured physical activity and sitting with incident mild cognitive impairment or probable dementia among older women. Alzheimers Dement 2023 Jul;19(7):3041-3054 [FREE Full text] [doi: 10.1002/alz.12908] [Medline: 36695426]
314. Noor MH, Salcic Z, Wang KI. Dynamic sliding window method for physical activity recognition using a single tri-axial accelerometer. In: Proceedings of the 2015 IEEE 10th Conference on Industrial Electronics and Applications. 2015 Presented at: ICIEA '15; June 15-17, 2015; Auckland, New Zealand p. 102-105 URL: https://ieeexplore.ieee.org/document/7334092 [doi: 10.1109/iciea.2015.7334092]

XSL•FO
RenderX

315. Nordstrøm M, Hansen BH, Paus B, Kolset SO. Accelerometer-determined physical activity and walking capacity in persons with Down syndrome, Williams syndrome and Prader-Willi syndrome. Res Dev Disabil 2013 Dec;34(12):4395-4403 [FREE Full text] [doi: 10.1016/j.ridd.2013.09.021] [Medline: 24139709]
316. Nørgaard M, Lomholt JJ, Thastum M, Herlin M, Twilt M, Herlin T. Accelerometer-assessed daily physical activity in relation to pain cognition in juvenile idiopathic arthritis. Scand J Rheumatol 2017 Jan;46(1):22-26. [doi: 10.3109/03009742.2016.1160146] [Medline: 27308924]
317. Nu U, Touati T, Buddhadev S, Sun R, Smuck M, Song I. Who is physically active? Classification and analysis of physical activity using NHANES data. In: Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence. 2020 Presented at: SSCI '20; December 1-4, 2020; Canberra. Australia p. 527-533 URL: https://ieeexplore.ieee.org/document/9308353 [doi: 10.1109/ssci47803.2020.9308353]
318. Ohkawara K, Oshima Y, Hikihara Y, Ishikawa-Takata K, Tabata I, Tanaka S. Real-time estimation of daily physical activity intensity by a triaxial accelerometer and a gravity-removal classification algorithm. Br J Nutr 2011 Jun;105(11):1681-1691. [doi: 10.1017/S0007114510005441] [Medline: 21262061]
319. Orme MW, Steiner MC, Morgan MD, Kingsnorth AP, Esliger DW, Singh SJ, et al. 24-hour accelerometry in COPD: exploring physical activity, sedentary behavior, sleep and clinical characteristics. Int J Chron Obstruct Pulmon Dis 2019;14:419-430 [FREE Full text] [doi: 10.2147/COPD.S183029] [Medline: 30863042]
320. Ortlieb S, Dias A, Gorzelniak L, Nowak D, Karrasch S, Peters A, KORA Study Group. Exploring patterns of accelerometry-assessed physical activity in elderly people. Int J Behav Nutr Phys Act 2014 Mar 28;11(1):28 [FREE Full text] [doi: 10.1186/1479-5868-11-28] [Medline: 24575796]
321. Ortlieb S, Gorzelniak L, Nowak D, Strobl R, Grill E, Thorand B, et al. Associations between multiple accelerometry-assessed physical activity parameters and selected health outcomes in elderly people--results from the KORA-age study. PLoS One 2014;9(11):e111206 [FREE Full text] [doi: 10.1371/journal.pone.0111206] [Medline: 25372399]
322. Parada H, McDonald E, Bellettiere J, Evenson KR, LaMonte MJ, LaCroix AZ. Associations of accelerometer-measured physical activity and physical activity-related cancer incidence in older women: results from the WHI OPACH Study. Br J Cancer 2020 May;122(9):1409-1416 [FREE Full text] [doi: 10.1038/s41416-020-0753-6] [Medline: 32139875]
323. Park J, Ishikawa-Takata K, Tanaka S, Mekata Y, Tabata I. Effects of walking speed and step frequency on estimation of physical activity using accelerometers. J Physiol Anthropol 2011;30(3):119-127 [FREE Full text] [doi: 10.2114/jpa2.30.119] [Medline: 21636955]
324. Pärkkä J, Ermes M, Antila K, van Gils M, Mänttäri A, Nieminen H. Estimating intensity of physical activity: a comparison of wearable accelerometer and gyro sensors and 3 sensor locations. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:1511-1514. [doi: 10.1109/IEMBS.2007.4352588] [Medline: 18002254]
325. Paschali AA, Goodrick GK, Kalantzi-Azizi A, Papadatou D, Balasubramanyam A. Accelerometer feedback to promote physical activity in adults with type 2 diabetes: a pilot study. Percept Mot Skills 2005 Mar;100(1):61-68. [doi: 10.2466/pms.100.1.61-68] [Medline: 15773694]
326. Peter-Marske KM, Evenson KR, Moore CC, Cuthbertson CC, Howard AG, Shiroma EJ, et al. Association of accelerometer-measured physical activity and sedentary behavior with incident cardiovascular disease, myocardial infarction, and ischemic stroke: the women's health study. J Am Heart Assoc 2023 Apr 04;12(7):e028180 [FREE Full text] [doi: 10.1161/JAHA.122.028180] [Medline: 36974744]
327. Peters TM, Moore SC, Xiang YB, Yang G, Shu XO, Ekelund U, et al. Accelerometer-measured physical activity in Chinese adults. Am J Prev Med 2010 Jun;38(6):583-591 [FREE Full text] [doi: 10.1016/j.amepre.2010.02.012] [Medline: 20494234]
328. Petersen TL, Brønd JC, Kristensen PL, Aadland E, Grøntved A, Jepsen R. Resemblance in accelerometer-assessed physical activity in families with children: the Lolland-Falster Health Study. Int J Behav Nutr Phys Act 2020 Dec 04;17(1):161 [FREE Full text] [doi: 10.1186/s12966-020-01067-7] [Medline: 33276796]
329. Pettee Gabriel K, Karvonen-Gutierrez CA, Colvin AB, Ylitalo KR, Whitaker KM, Lange-Maia BS, et al. Associations of accelerometer-determined sedentary behavior and physical activity with physical performance outcomes by race/ethnicity in older women. Prev Med Rep 2021 Oct;23:101408 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101408] [Medline: 34123715]
330. Pettee Gabriel K, Sternfeld B, Shiroma EJ, Pérez A, Cheung J, Lee I. Bidirectional associations of accelerometer-determined sedentary behavior and physical activity with reported time in bed: women's health study. Sleep Health 2017 Mar;3(1):49-55 [FREE Full text] [doi: 10.1016/j.sleh.2016.10.001] [Medline: 28346151]
331. Phan A, Askim T, Lydersen S, Indredavik B, Wethal T. Accelerometer-measured physical activity at 3 months as a predictor of symptoms of depression and anxiety 1 year after stroke: a multicentre prospective cohort study in central Norway. J Rehabil Med 2023 Dec 16;55:jrm12309 [FREE Full text] [doi: 10.2340/jrm.v55.12309] [Medline: 37970656]
332. Plasqui G, Joosen AM, Kester AD, Goris AH, Westerterp KR. Measuring free-living energy expenditure and physical activity with triaxial accelerometry. Obes Res 2005 Aug;13(8):1363-1369 [FREE Full text] [doi: 10.1038/oby.2005.165] [Medline: 16129718]
333. Posis AIB, Bellettiere J, Salem RM, LaMonte MJ, Manson JE, Casanova R, et al. Associations of accelerometer-measured physical activity and sedentary time with all-cause mortality by genetic predisposition for longevity. J Aging Phys Act 2023 May 01;31(2):265-275 [FREE Full text] [doi: 10.1123/japa.2022-0067] [Medline: 36002033]

XSL•FO
RenderX

334. Prioreschi A, Hodkinson B, Tikly M, McVeigh JA. Changes in physical activity measured by accelerometry following initiation of DMARD therapy in rheumatoid arthritis. Rheumatology (Oxford) 2014 May;53(5):923-926. [doi: 10.1093/rheumatology/ket457] [Medline: 24459221]
335. Pugh ME, Buchowski MS, Robbins IM, Newman JH, Hemnes AR. Physical activity limitation as measured by accelerometry in pulmonary arterial hypertension. Chest 2012 Dec;142(6):1391-1398 [FREE Full text] [doi: 10.1378/chest.12-0150] [Medline: 22576635]
336. Qiu S, Xing Z. Association between accelerometer-derived physical activity and incident cardiac arrest. Europace 2023 Dec 06;25(12):euad353 [FREE Full text] [doi: 10.1093/europace/euad353] [Medline: 38016070]
337. Rahman H, Carrault G, Ge D, Amoud H, Prioux J, Le FA. Ambulatory physical activity representation and classification using spectral distances approach. In: Proceedings of the 2015 International Conference on Advances in Biomedical Engineering. 2015 Presented at: ICABME '15; September 16-18, 2015; Beirut, Lebanon p. 69-72 URL: https://ieeexplore.ieee.org/document/7323253 [doi: 10.1109/icabme.2015.7323253]
338. Ramulu PY, Maul E, Hochberg C, Chan ES, Ferrucci L, Friedman DS. Real-world assessment of physical activity in glaucoma using an accelerometer. Ophthalmology 2012 Jul;119(6):1159-1166 [FREE Full text] [doi: 10.1016/j.ophtha.2012.01.013] [Medline: 22386950]
339. Ricketts HC, Buchan DS, Steffensen F, Chaudhuri R, Baker JS, Cowan DC. Physical activity levels in asthma: relationship with disease severity, body mass index and novel accelerometer-derived metrics. J Asthma 2023 May;60(4):824-834 [FREE Full text] [doi: 10.1080/02770903.2022.2102037] [Medline: 35876843]
340. Robertson MC, Green CE, Liao Y, Durand CP, Basen-Engquist KM. Self-efficacy and physical activity in overweight and obese adults participating in a worksite weight loss intervention: multistate modeling of wearable device data. Cancer Epidemiol Biomarkers Prev 2020 Apr;29(4):769-776 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0907] [Medline: 31871110]
341. Roscoe CM, James RS, Duncan MJ. Accelerometer-based physical activity levels differ between week and weekend days in British preschool children. J Funct Morphol Kinesiol 2019 Oct 12;4(3):65 [FREE Full text] [doi: 10.3390/jfmk4030065] [Medline: 33467380]
342. Rowlands AV, Dempsey PC, Gillies C, Kloecker DE, Razieh C, Chudasama Y, et al. Association between accelerometer-assessed physical activity and severity of COVID-19 in UK biobank. Mayo Clin Proc Innov Qual Outcomes 2021 Dec;5(6):997-1007 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2021.08.011] [Medline: 34430796]
343. Rowlands AV, Henson JJ, Coull NA, Edwardson CL, Brady E, Hall A, et al. The impact of COVID-19 restrictions on accelerometer-assessed physical activity and sleep in individuals with type 2 diabetes. Diabet Med 2021 Oct;38(10):e14549 [FREE Full text] [doi: 10.1111/dme.14549] [Medline: 33650112]
344. Rowlands AV, van Hees VT, Dawkins NP, Maylor BD, Plekhanova T, Henson J, et al. Accelerometer-assessed physical activity in people with type 2 diabetes: accounting for sleep when determining associations with markers of health. Sensors (Basel) 2023 Jul 07;23(12):5382 [FREE Full text] [doi: 10.3390/s23125382] [Medline: 37420551]
345. Rubin DS, Huisingh-Scheetz M, Hung A, Ward RP, Nagele P, Arena R, et al. Accuracy of physical function questions to predict moderate-vigorous physical activity as measured by Hip Accelerometry. Anesthesiology 2019 Dec;131(5):992-1003 [FREE Full text] [doi: 10.1097/ALN.0000000000002911] [Medline: 31490293]
346. Salerno EA, Saint-Maurice PF, Wan F, Peterson LL, Park Y, Cao Y, et al. Prospective associations between accelerometry-derived physical activity and sedentary behaviors and mortality among cancer survivors. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad007 [FREE Full text] [doi: 10.1093/jncics/pkad007] [Medline: 36786414]
347. Salvo D, Torres C, Villa U, Rivera JA, Sarmiento OL, Reis RS, et al. Accelerometer-based physical activity levels among Mexican adults and their relation with sociodemographic characteristics and BMI: a cross-sectional study. Int J Behav Nutr Phys Act 2015 Jul 20;12:79 [FREE Full text] [doi: 10.1186/s12966-015-0243-z] [Medline: 26088430]
348. Salway R, Foster C, de Vocht F, Tibbitts B, Emm-Collison L, House D, et al. Accelerometer-measured physical activity and sedentary time among children and their parents in the UK before and after COVID-19 lockdowns: a natural experiment. Int J Behav Nutr Phys Act 2022 May 16;19(1):51 [FREE Full text] [doi: 10.1186/s12966-022-01290-4] [Medline: 35570265]
349. Sandroff BM, Motl RW, Kam JP, Pula JH. Accelerometer measured physical activity and the integrity of the anterior visual pathway in multiple sclerosis. Mult Scler Relat Disord 2014 Jan;3(1):117-122. [doi: 10.1016/j.msard.2013.06.014] [Medline: 25877982]
350. Sasaki S, Nakamura K, Ukawa S, Okada E, Amagasa S, Inoue S, et al. Association of accelerometer-measured physical activity with kidney function in a Japanese population: the DOSANCO Health Study. BMC Nephrol 2022 Jan 03;23(1):7 [FREE Full text] [doi: 10.1186/s12882-021-02635-0] [Medline: 34979979]
351. Savikangas T, Suominen TH, Alén M, Rantalainen T, Sipilä S. Changes in femoral neck bone mineral density and structural strength during a 12-month multicomponent exercise intervention among older adults - does accelerometer-measured physical activity matter? Bone 2024 Jan;178:116951 [FREE Full text] [doi: 10.1016/j.bone.2023.116951] [Medline: 37913888]
352. Schumacher BT, Bellettiere J, LaMonte MJ, Evenson KR, Di C, Lee I, et al. Accelerometer-measured daily steps, physical function, and subsequent fall risk in older women: the objective physical activity and cardiovascular disease in older women study. J Aging Phys Act 2022 Aug 01;30(4):635-645 [FREE Full text] [doi: 10.1123/japa.2021-0159] [Medline: 34627127]

XSL•FO
RenderX

353. Schwendinger F, Wagner J, Infanger D, Schmidt-Trucksäss A, Knaier R. Methodological aspects for accelerometer-based assessment of physical activity in heart failure and health. BMC Med Res Methodol 2021 Dec 14;21(1):251 [FREE Full text] [doi: 10.1186/s12874-021-01350-6] [Medline: 34775952]
354. Scott D, Johansson J, Gandham A, Ebeling PR, Nordstrom P, Nordstrom A. Associations of accelerometer-determined physical activity and sedentary behavior with sarcopenia and incident falls over 12 months in community-dwelling Swedish older adults. J Sport Health Sci 2021 Oct;10(5):577-584 [FREE Full text] [doi: 10.1016/j.jshs.2020.01.006] [Medline: 34088651]
355. Seguin-Fowler RA, LaCroix AZ, LaMonte MJ, Liu J, Maddock JE, Rethorst CD, et al. Association of neighborhood walk score with accelerometer-measured physical activity varies by neighborhood socioeconomic status in older women. Prev Med Rep 2022 Oct;29:101931 [FREE Full text] [doi: 10.1016/j.pmedr.2022.101931] [Medline: 36161128]
356. Serra MC, Balraj E, DiSanzo BL, Ivey FM, Hafer-Macko CE, Treuth MS, et al. Validating accelerometry as a measure of physical activity and energy expenditure in chronic stroke. Top Stroke Rehabil 2017 Jan;24(1):18-23 [FREE Full text] [doi: 10.1080/10749357.2016.1183866] [Medline: 27322733]
357. Sevil M, Rashid M, Maloney Z, Hajizadeh I, Samadi S, Askari MR, et al. Determining physical activity characteristics from wristband data for use in automated insulin delivery systems. IEEE Sens J 2020 Nov 1;20(21):12859-12870 [FREE Full text] [doi: 10.1109/jsen.2020.3000772] [Medline: 33100923]
358. Shadyab AH, LaMonte MJ, Kooperberg C, Reiner AP, Carty CL, Manini TM, et al. Association of accelerometer-measured physical activity with leukocyte telomere length among older women. J Gerontol A Biol Sci Med Sci 2017 Oct 12;72(11):1532-1537 [FREE Full text] [doi: 10.1093/gerona/glx037] [Medline: 28329327]
359. Sharma B, Obeid J, DeMatteo C, Noseworthy MD, Timmons BW. New insights into accelerometer-measured habitual physical activity and sedentary time during early recovery in pediatric concussion. Pediatr Exerc Sci 2024 May 01;36(2):58-65 [FREE Full text] [doi: 10.1123/pes.2023-0016] [Medline: 37591503]
360. Shin SW, Kim H, Kang YH, Kim J. Accelerometer-derived physical activity and sedentary behavior patterns among Korean adults. Phys Act Nutr 2023 Jul;27(2):25-33 [FREE Full text] [doi: 10.20463/pan.2023.0015] [Medline: 37583069]
361. Shin YE, Choi WH, Shin TM. Physical activity recognition based on rotated acceleration data using quaternion in sedentary behavior : a preliminary study. In: Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014 Presented at: EMBC '14; August 26-30, 2014; Chicago, IL p. 26-30 URL: https://ieeexplore.ieee.org/document/6944741 [doi: 10.1109/embc.2014.6944741]
362. Shwetar Y, Huang Z, Veerubhotla A, Knezevic S, Hong E, Spungen AM, et al. Predicting physical activity intensity using raw accelerometer signals in manual wheelchair users with spinal cord injury. Spinal Cord 2022 Mar;60(2):149-156. [doi: 10.1038/s41393-021-00728-z] [Medline: 34819608]
363. Si K, Cao Z, Liu Q, Yang Y, Dai Q, Yao Y, et al. Accelerometer-measured physical activity, sedentary behavior, and risk of incident pelvic organ prolapse: a prospective cohort study in the UK Biobank. Int J Behav Nutr Phys Act 2024 Mar 02;21(1):12 [FREE Full text] [doi: 10.1186/s12966-024-01559-w] [Medline: 38308373]
364. Siddique J, Aaby D, Montag SE, Sidney S, Sternfeld B, Welch WA, et al. Individualized relative-intensity physical activity accelerometer cut points. Med Sci Sports Exerc 2020 Mar;52(2):398-407 [FREE Full text] [doi: 10.1249/MSS.0000000000002153] [Medline: 31524826]
365. Silva RD, Cabral LL, Browne RA, Lemos TM, Alves CP, Crochemore-Silva I, et al. Joint associations of accelerometer-measured physical activity and sedentary time with cardiometabolic risk in older adults: a cross-sectional study. Exp Gerontol 2022 Aug;165:111839. [doi: 10.1016/j.exger.2022.111839] [Medline: 35609715]
366. Simón-Vicente L, Rivadeneyra-Posadas J, Soto-Célix M, Raya-González J, Castillo D, Calvo S, García-Bustillo, et al. Accelerometer cut-points for physical activity assessment in adults with mild to moderate Huntington's disease: a cross-sectional multicentre study. Int J Environ Res Public Health 2022 Dec 11;19(22):14834 [FREE Full text] [doi: 10.3390/ijerph192214834] [Medline: 36429552]
367. Sjöros T, Vähä-Ypyä H, Laine S, Garthwaite T, Lahesmaa M, Laurila SM, et al. Both sedentary time and physical activity are associated with cardiometabolic health in overweight adults in a 1 month accelerometer measurement. Sci Rep 2020 Nov 25;10(1):20578 [FREE Full text] [doi: 10.1038/s41598-020-77637-3] [Medline: 33239818]
368. Slootmaker SM, Chinapaw MJ, Seidell JC, van MW, Schuit AJ. Accelerometers and Internet for physical activity promotion in youth? Feasibility and effectiveness of a minimal intervention [ISRCTN93896459]. Prev Med 2010 Jul;51(1):31-36. [doi: 10.1016/j.ypmed.2010.03.015] [Medline: 20380847]
369. Smith DM, DeCaro JA, Murphy SL, Parmelee PA. Momentary reports of fatigue predict physical activity level: wrist, waist, and combined accelerometry. J Aging Health 2020 Oct;32(9):921-925. [doi: 10.1177/0898264319863609] [Medline: 31319748]
370. Smith HA, Storti KL, Arena VC, Kriska AM, Gabriel KK, Sutton-Tyrrell K, et al. Associations between accelerometer-derived physical activity and regional adiposity in young men and women. Obesity (Silver Spring) 2013 Jul;21(6):1299-1305 [FREE Full text] [doi: 10.1002/oby.20308] [Medline: 23408709]
371. Smith ID, Ross LM, Gabaldon JR, Holdgate N, Pieper CF, Ning TC, et al. The relation of accelerometer-measured physical activity and serum uric acid using the national health and nutrition survey (NHANES) 2003-2004. Front Sports Act Living 2021;3:775398 [FREE Full text] [doi: 10.3389/fspor.2021.775398] [Medline: 35098119]

XSL•FO
RenderX

372. Smith KE, O'Connor SM, Mason TB, Wang S, Dzubur E, Crosby RD, et al. Associations between objective physical activity and emotional eating among adiposity-discordant siblings using ecological momentary assessment and accelerometers. Pediatr Obes 2021 Mar;16(3):e12720 [FREE Full text] [doi: 10.1111/ijpo.12720] [Medline: 32881329]
373. Spartano NL, Davis-Plourde KL, Himali JJ, Andersson C, Pase MP, Maillard P, et al. Association of accelerometer-measured light-intensity physical activity with brain volume: the Framingham heart study. JAMA Netw Open 2019 May 05;2(4):e192745 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2745] [Medline: 31002329]
374. Spartano NL, Demissie S, Himali JJ, Dukes KA, Murabito JM, Vasan RS, et al. Accelerometer-determined physical activity and cognitive function in middle-aged and older adults from two generations of the Framingham Heart Study. Alzheimers Dement (N Y) 2019;5:618-626 [FREE Full text] [doi: 10.1016/j.trci.2019.08.007] [Medline: 31660424]
375. Spartano NL, Heffernan KS, Dumas AK, Gump BB. Accelerometer-determined physical activity and the cardiovascular response to mental stress in children. J Sci Med Sport 2017 Jan;20(1):60-65 [FREE Full text] [doi: 10.1016/j.jsams.2016.05.008] [Medline: 27283342]
376. Spartano NL, Wang R, Yang Q, Chernofsky A, Murabito JM, Vasan RS, et al. Association of accelerometer-measured physical activity and sedentary time with epigenetic markers of aging. Med Sci Sports Exerc 2023 Mar 01;55(2):264-272 [FREE Full text] [doi: 10.1249/MSS.0000000000003041] [Medline: 36107108]
377. Stamatakis E, Ahmadi MN, Friedenreich CM, Blodgett JM, Koster A, Holtermann A, et al. Vigorous intermittent lifestyle physical activity and cancer incidence among nonexercising adults: the UK biobank accelerometry study. JAMA Oncol 2023 Oct 01;9(9):1255-1259 [FREE Full text] [doi: 10.1001/jamaoncol.2023.1830] [Medline: 37498576]
378. Stamatakis E, Ahmadi MN, Gill JMR, Thøgersen-Ntoumani C, Gibala MJ, Doherty A, et al. Association of wearable device-measured vigorous intermittent lifestyle physical activity with mortality. Nat Med 2022 Dec;28(12):2521-2529 [FREE Full text] [doi: 10.1038/s41591-022-02100-x] [Medline: 36482104]
379. Stevens ML, Crowley P, Rasmussen CL, Hallman DM, Mortensen OS, Nygård C, et al. Accelerometer-measured physical activity at work and need for recovery: a compositional analysis of cross-sectional data. Ann Work Expo Health 2020 Mar 20;64(2):138-151 [FREE Full text] [doi: 10.1093/annweh/wxz095] [Medline: 31879769]
380. Straight CR, Ward-Ritacco CL, Evans EM. Association between accelerometer-measured physical activity and muscle capacity in middle-aged postmenopausal women. Menopause 2015 Dec;22(11):1204-1211. [doi: 10.1097/GME.0000000000000447] [Medline: 25783471]
381. Stubbs B, Chen L, Chung M, Ku P. Physical activity ameliorates the association between sedentary behavior and cardiometabolic risk among inpatients with schizophrenia: a comparison versus controls using accelerometry. Compr Psychiatry 2017 May;74:144-150. [doi: 10.1016/j.comppsych.2017.01.010] [Medline: 28167327]
382. Sufrinko AM, Howie EK, Elbin RJ, Collins MW, Kontos AP. A preliminary investigation of accelerometer-derived sleep and physical activity following sport-related concussion. J Head Trauma Rehabil 2018;33(5):E64-E74. [doi: 10.1097/HTR.0000000000000387] [Medline: 29601343]
383. Sutin AR, Luchetti M, Stephan Y, Terracciano A. Meaning in life and accelerometer-measured physical activity: association based on 67,038 UK biobank participants. Ment Health Phys Act 2021 Oct;21:100412 [FREE Full text] [doi: 10.1016/j.mhpa.2021.100412] [Medline: 35340340]
384. Sutin AR, Stephan Y, Kekäläinen T, Luchetti M, Terracciano A. Purpose in life and accelerometer-measured physical activity among older adults. Psychol Health 2023 May 18:1-15. [doi: 10.1080/08870446.2023.2200414] [Medline: 37073429]
385. Sweegers MG, Buffart LM, Huijsmans RJ, Konings IR, van Zweeden AA, Brug J, et al. From accelerometer output to physical activity intensities in breast cancer patients. J Sci Med Sport 2020 Mar;23(2):176-181. [doi: 10.1016/j.jsams.2019.09.001] [Medline: 31537492]
386. Takae R, Hatamoto Y, Yasukata J, Kose Y, Komiyama T, Ikenaga M, et al. Association of lower-extremity muscle performance and physical activity level and intensity in middle-aged and older adults: a doubly labeled water and accelerometer study. J Nutr Health Aging 2020;24(9):1023-1030. [doi: 10.1007/s12603-020-1449-6] [Medline: 33155632]
387. Tanaka C, Hikihara Y, Ando T, Oshima Y, Usui C, Ohgi Y, et al. Prediction of physical activity intensity with accelerometry in young children. Int J Environ Res Public Health 2019 Mar 15;16(6):931 [FREE Full text] [doi: 10.3390/ijerph16060931] [Medline: 30875871]
388. Tanaka C, Shikano A, Imai N, Chong KH, Howard SJ, Tanabe K, et al. Accelerometer-measured physical activity and sedentary time among children in Japan before and during COVID-19: a cross-sectional and longitudinal analysis. Int J Environ Res Public Health 2023 Jan 09;20(2):1130 [FREE Full text] [doi: 10.3390/ijerph20021130] [Medline: 36673886]
389. Tanaka C, Tanaka S. Daily physical activity in Japanese preschool children evaluated by triaxial accelerometry: the relationship between period of engagement in moderate-to-vigorous physical activity and daily step counts. J Physiol Anthropol 2009 Dec;28(6):283-288 [FREE Full text] [doi: 10.2114/jpa2.28.283] [Medline: 20009376]
390. Tanaka C, Tanaka S, Kawahara J, Midorikawa T. Triaxial accelerometry for assessment of physical activity in young children. Obesity (Silver Spring) 2007 May;15(5):1233-1241 [FREE Full text] [doi: 10.1038/oby.2007.145] [Medline: 17495200]
391. Tanaka S, Ishikawa-Takata K, Nakae S, Sasaki S. Prediction of the physical activity level of community-dwelling older Japanese adults with a triaxial accelerometer containing a classification algorithm for ambulatory and non-ambulatory activities. Sensors (Basel) 2023 May 22;23(10):4960 [FREE Full text] [doi: 10.3390/s23104960] [Medline: 37430874]

XSL•FO
RenderX

392. Tao K, Liu W, Xiong S, Ken L, Zeng N, Peng Q, et al. Associations between self-determined motivation, accelerometer-determined physical activity, and quality of life in Chinese college students. Int J Environ Res Public Health 2019 Aug 16;16(16):2941 [FREE Full text] [doi: 10.3390/ijerph16162941] [Medline: 31426280]
393. Tarp J, Bugge A, Andersen LB, Sardinha LB, Ekelund U, Brage S, et al. Does adiposity mediate the relationship between physical activity and biological risk factors in youth?: a cross-sectional study from the International Children's Accelerometry Database (ICAD). Int J Obes (Lond) (Forthcoming) 2017 Oct 03 [FREE Full text] [doi: 10.1038/ijo.2017.241] [Medline: 29087387]
394. Tarp J, Hansen BH, Fagerland MW, Steene-Johannessen J, Anderssen SA, Ekelund U. Accelerometer-measured physical activity and sedentary time in a cohort of US adults followed for up to 13 years: the influence of removing early follow-up on associations with mortality. Int J Behav Nutr Phys Act 2020 Mar 14;17(1):39 [FREE Full text] [doi: 10.1186/s12966-020-00945-4] [Medline: 32169059]
395. Tasheva P, Kraege V, Vollenweider P, Roulet G, Méan M, Marques-Vidal P. Accelerometry assessed physical activity of older adults hospitalized with acute medical illness - an observational study. BMC Geriatr 2020 Oct 02;20(1):382 [FREE Full text] [doi: 10.1186/s12877-020-01763-w] [Medline: 33008378]
396. Taylor RW, Williams SM, Farmer VL, Taylor BJ. Changes in physical activity over time in young children: a longitudinal study using accelerometers. PLoS One 2013;8(11):e81567 [FREE Full text] [doi: 10.1371/journal.pone.0081567] [Medline: 24282607]
397. Thraen-Borowski KM, Gennuso KP, Cadmus-Bertram L. Accelerometer-derived physical activity and sedentary time by cancer type in the United States. PLoS One 2017;12(8):e0182554 [FREE Full text] [doi: 10.1371/journal.pone.0182554] [Medline: 28806753]
398. Thralls KJ, Godbole S, Manini TM, Johnson E, Natarajan L, Kerr J. A comparison of accelerometry analysis methods for physical activity in older adult women and associations with health outcomes over time. J Sports Sci 2019 Oct;37(20):2309-2317 [FREE Full text] [doi: 10.1080/02640414.2019.1631080] [Medline: 31195893]
399. Toivo K, Vähä-Ypyä H, Kannus P, Tokola K, Alanko L, Heinonen OJ, et al. Physical activity measured by accelerometry among adolescents participating in sports clubs and non-participating peers. Eur J Sport Sci 2023 Jul;23(7):1426-1434 [FREE Full text] [doi: 10.1080/17461391.2022.2103740] [Medline: 35861140]
400. Toomey CM, Whittaker JL, Doyle-Baker PK, Emery CA. Does a history of youth sport-related knee injury still impact accelerometer-measured levels of physical activity after 3-12 years? Phys Ther Sport 2022 May;55:90-97. [doi: 10.1016/j.ptsp.2022.03.003] [Medline: 35290947]
401. Trinh L, Motl RW, Roberts SA, Gibbons T, McAuley E. Estimation of physical activity intensity cut-points using accelerometry in breast cancer survivors and age-matched controls. Eur J Cancer Care (Engl) 2019 Sep;28(5):e13090. [doi: 10.1111/ecc.13090] [Medline: 31106924]
402. Trost SG, Brookes DS, Ahmadi MN. Evaluation of wrist accelerometer cut-points for classifying physical activity intensity in youth. Front Digit Health 2022;4:884307 [FREE Full text] [doi: 10.3389/fdgth.2022.884307] [Medline: 35585912]
403. Trumpeter NN, Lawman HG, Wilson DK, Pate RR, Van Horn ML, Tate AK. Accelerometry cut points for physical activity in underserved African Americans. Int J Behav Nutr Phys Act 2012 Jul 14;9:73 [FREE Full text] [doi: 10.1186/1479-5868-9-73] [Medline: 22697280]
404. Tsai L, Boyle E, Buhl SF, Kock G, Brønd JC, Visser M, et al. Associations between appetite, physical activity and sedentary behaviour from hip- and wrist-worn accelerometers in community-dwelling older adults. Geriatr Gerontol Int 2023 Jul;23(6):411-417. [doi: 10.1111/ggi.14588] [Medline: 37098733]
405. Tsunoda K, Kitano N, Kai Y, Jindo T, Uchida K, Arao T. Dose-response relationships of accelerometer-measured sedentary behaviour and physical activity with non-alcoholic fatty liver disease. Aliment Pharmacol Ther 2021 Dec;54(10):1330-1339. [doi: 10.1111/apt.16631] [Medline: 34633105]
406. Vaara JP, Kyröläinen H, Vasankari T, Kainulainen H, Raitanen J, Kujala UM. Accelerometer-based sedentary time, physical activity, and serum metabolome in young men. Metabolites 2022 Jul 27;12(8):700 [FREE Full text] [doi: 10.3390/metabo12080700] [Medline: 36005572]
407. Vähä-Ypyä H, Vasankari T, Husu P, Mänttäri A, Vuorimaa T, Suni J, et al. Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (MAD). PLoS One 2015;10(8):e0134813 [FREE Full text] [doi: 10.1371/journal.pone.0134813] [Medline: 26292225]
408. Valenti G, Camps SG, Verhoef SP, Bonomi AG, Westerterp KR. Validating measures of free-living physical activity in overweight and obese subjects using an accelerometer. Int J Obes (Lond) 2014 Jul;38(7):1011-1014. [doi: 10.1038/ijo.2013.195] [Medline: 24166066]
409. van Bakergem M, Sommer EC, Heerman WJ, Hipp JA, Barkin SL. Objective reports versus subjective perceptions of crime and their relationships to accelerometer-measured physical activity in Hispanic caretaker-child dyads. Prev Med 2017 Feb;95 Suppl:S68-S74 [FREE Full text] [doi: 10.1016/j.ypmed.2016.12.001] [Medline: 27939263]
410. van Ballegooijen AJ, van der Ploeg HP, Visser M. Daily sedentary time and physical activity as assessed by accelerometry and their correlates in older adults. Eur Rev Aging Phys Act 2019;16:3. [doi: 10.1186/s11556-019-0210-9] [Medline: 30820261]

XSL•FO
RenderX

411. van der Zee-Neuen A, Wirth W, Hösl K, Osterbrink J, Eckstein F. The association of physical activity and depression in patients with, or at risk of, osteoarthritis is captured equally well by patient reported outcomes (PROs) and accelerometer measurements - analyses of data from the Osteoarthritis Initiative. Semin Arthritis Rheum 2019 Dec;49(3):325-330. [doi: 10.1016/j.semarthrit.2019.05.009] [Medline: 31248586]
412. Van Dyck D, Barnett A, Cerin E, Conway TL, Esteban-Cornejo I, Hinckson E, et al. Associations of accelerometer measured school- and non-school based physical activity and sedentary time with body mass index: IPEN adolescent study. Int J Behav Nutr Phys Act 2022 Jul 14;19(1):85 [FREE Full text] [doi: 10.1186/s12966-022-01324-x] [Medline: 35836235]
413. Van Dyck D, Cerin E, De Bourdeaudhuij I, Salvo D, Christiansen LB, Macfarlane D, et al. Moderating effects of age, gender and education on the associations of perceived neighborhood environment attributes with accelerometer-based physical activity: the IPEN adult study. Health Place 2015 Nov;36:65-73 [FREE Full text] [doi: 10.1016/j.healthplace.2015.09.007] [Medline: 26454247]
414. van Eijk RP, Bakers JN, Bunte TM, de Fockert AJ, Eijkemans MJ, van den Berg LH. Accelerometry for remote monitoring of physical activity in amyotrophic lateral sclerosis: a longitudinal cohort study. J Neurol 2019 Oct;266(10):2387-2395 [FREE Full text] [doi: 10.1007/s00415-019-09427-5] [Medline: 31187191]
415. van GS, Boonen A, van DH, Heuft L, Luime J, Spoorenberg A, et al. Accelerometer quantification of physical activity and activity patterns in patients with ankylosing spondylitis and population controls. J Rheumatol 2015 Dec;42(12):2369-2375. [doi: 10.3899/jrheum.150015] [Medline: 26523021]
416. van Landingham SW, Willis JR, Vitale S, Ramulu PY. Visual field loss and accelerometer-measured physical activity in the United States. Ophthalmology 2012 Dec;119(12):2486-2492. [doi: 10.1016/j.ophtha.2012.06.034] [Medline: 22892152]
417. van Langenberg DR, Papandony MC, Gibson PR. Sleep and physical activity measured by accelerometry in Crohn's disease. Aliment Pharmacol Ther 2015 May;41(10):991-1004. [doi: 10.1111/apt.13160] [Medline: 25783784]
418. Vandercappellen EJ, Koster A, Savelberg HH, Eussen SJ, Dagnelie PC, Schram MT, et al. Accelerometer-derived physical activity and sedentary time and cardiac biomarkers: the Maastricht Study. Front Cardiovasc Med 2023;10:1081713 [FREE Full text] [doi: 10.3389/fcvm.2023.1081713] [Medline: 37187790]
419. Vanhelst J, Béghin L, Turck D, Gottrand F. New validated thresholds for various intensities of physical activity in adolescents using the Actigraph accelerometer. Int J Rehabil Res 2011 Jun;34(2):175-177. [doi: 10.1097/MRR.0b013e328340129e] [Medline: 20890210]
420. Vanhelst J, Hurdiel R, Mikulovic J, Bui-Xuân G, Fardy P, Theunynck D, et al. Validation of the Vivago wrist-worn accelerometer in the assessment of physical activity. BMC Public Health 2012;12:690 [FREE Full text] [doi: 10.1186/1471-2458-12-690] [Medline: 22913286]
421. Vásquez E, Strizich G, Isasi CR, Echeverria SE, Sotres-Alvarez D, Evenson KR, et al. Is there a relationship between accelerometer-assessed physical activity and sedentary behavior and cognitive function in US Hispanic/Latino adults? The Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Prev Med 2017 Oct;103:43-48. [doi: 10.1016/j.ypmed.2017.07.024] [Medline: 28765082]
422. Vásquez PM, Tarraf W, Chai A, Doza A, Sotres-Alvarez D, Diaz KM, et al. Accelerometer-measured latent physical activity profiles and neurocognition among middle-aged and older Hispanic/Latino Adults in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). J Gerontol B Psychol Sci Soc Sci 2022 Dec 29;77(12):e263-e278 [FREE Full text] [doi: 10.1093/geronb/gbac161] [Medline: 36219450]
423. Verbestel V, De HS, Bammann K, Barba G, Hadjigeorgiou C, Eiben G, IDEFICS Consortium. Are context-specific measures of parental-reported physical activity and sedentary behaviour associated with accelerometer data in 2-9-year-old European children? Public Health Nutr 2015 Apr;18(5):860-868. [doi: 10.1017/S136898001400086X] [Medline: 24887315]
424. Verma VK, Lin W. Machine learning-based 30-day hospital readmission predictions for COPD patients using physical activity data of daily living with accelerometer-based device. Biosensors (Basel) 2022 Aug 05;12(8):605 [FREE Full text] [doi: 10.3390/bios12080605] [Medline: 36005000]
425. Vorwerg Y, Petroff D, Kiess W, Blüher S. Physical activity in 3-6 year old children measured by SenseWear Pro®: direct accelerometry in the course of the week and relation to weight status, media consumption, and socioeconomic factors. PLoS One 2013;8(4):e60619 [FREE Full text] [doi: 10.1371/journal.pone.0060619] [Medline: 23573273]
426. Wang Y, Wattelez G, Frayon S, Caillaud C, Galy O, Yacef K. ABIPA: ARIMA-based integration of accelerometer-based physical activity for adolescent weight status prediction. ACM Trans Comput Healthcare 2023 Mar 30;4(1):1-19. [doi: 10.1145/3561611]
427. Watts EL, Saint-Maurice PF, Doherty A, Fensom GK, Freeman JR, Gorzelitz JS, et al. Association of accelerometer-measured physical activity level with risks of hospitalization for 25 common health conditions in UK adults. JAMA Netw Open 2023 Mar 01;6(2):e2256186 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.56186] [Medline: 36795414]
428. Weaver RG, Crimarco A, Brusseau TA, Webster CA, Burns RD, Hannon JC. Accelerometry-derived physical activity of first through third grade children during the segmented school day. J Sch Health 2016 Oct;86(10):726-733. [doi: 10.1111/josh.12426] [Medline: 27619763]
429. Wendt A, da Silva IC, Gonçalves H, Menezes A, Barros F, Wehrmeister FC. Short-term effect of physical activity on sleep health: a population-based study using accelerometry. J Sport Health Sci 2022 Oct;11(5):630-638 [FREE Full text] [doi: 10.1016/j.jshs.2020.04.007] [Medline: 32422346]

XSL•FO
RenderX

430. Werneck AO, Jago R, Kriemler S, Andersen LB, Wedderkopp N, Northstone K, International Children's Accelerometry Database (ICAD) Collaborators. Association of change in the school travel mode with changes in different physical activity intensities and sedentary time: a international children's accelerometry database study. Prev Med 2021 Dec;153:106862 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106862] [Medline: 34710443]
431. Whitaker KM, Pettee Gabriel K, Buman MP, Pereira MA, Jacobs DR, Reis JP, et al. Associations of accelerometer-measured sedentary time and physical activity with prospectively assessed cardiometabolic risk factors: the CARDIA study. J Am Heart Assoc 2019 Jan 08;8(1):e010212 [FREE Full text] [doi: 10.1161/JAHA.118.010212] [Medline: 30616480]
432. Whitaker KM, Pettee Gabriel K, Laddu D, White DK, Sidney S, Sternfeld B, et al. Bidirectional associations of accelerometer measured sedentary behavior and physical activity with knee pain, stiffness, and physical function: the CARDIA study. Prev Med Rep 2021 Jul;22:101348 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101348] [Medline: 33816086]
433. Whitaker KM, Xiao Q, Pettee Gabriel K, Gordon Larsen P, Jacobs DR, Sidney S, et al. Perceived and objective characteristics of the neighborhood environment are associated with accelerometer-measured sedentary time and physical activity, the CARDIA Study. Prev Med 2019 Jul;123:242-249 [FREE Full text] [doi: 10.1016/j.ypmed.2019.03.039] [Medline: 30940573]
434. Whitaker KM, Zhang D, Pettee Gabriel K, Ahrens M, Sternfeld B, Sidney S, et al. Longitudinal associations of midlife accelerometer determined sedentary behavior and physical activity with cognitive function: the CARDIA study. J Am Heart Assoc 2021 Mar 02;10(3):e018350 [FREE Full text] [doi: 10.1161/JAHA.120.018350] [Medline: 33470140]
435. Willis JR, Jefferys JL, Vitale S, Ramulu PY. Visual impairment, uncorrected refractive error, and accelerometer-defined physical activity in the United States. Arch Ophthalmol 2012 Mar;130(3):329-335. [doi: 10.1001/archopthalmol.2011.1773] [Medline: 22411662]
436. Wu J, Olson JL, Brunke-Reese D, Lagoa CM, Conroy DE. Wearable device adherence among insufficiently-active young adults is independent of identity and motivation for physical activity. J Behav Med 2024 May;47(2):197-206. [doi: 10.1007/s10865-023-00444-4] [Medline: 37642938]
437. Wu Y, Goodrich JM, Dolinoy DC, Sánchez BN, Ruiz-Narváez EA, Banker M, et al. Accelerometer-measured physical activity, reproductive hormones, and DNA methylation. Med Sci Sports Exerc 2020 Mar;52(3):598-607 [FREE Full text] [doi: 10.1249/MSS.0000000000002175] [Medline: 31652236]
438. Wu Z, Wang Z, Hu B, Zhang X, Zhang F, Wang H, et al. Relationships of accelerometer-based measured objective physical activity and sedentary behaviour with cognitive function: a comparative cross-sectional study of China's elderly population. BMC Geriatr 2020 May 22;20(1):149 [FREE Full text] [doi: 10.1186/s12877-020-01521-y] [Medline: 32321436]
439. Yang X, Dinh A, Chen L. Implementation of a weararable real-time system for physical activity recognition based on Naive Bayes classifier. In: Proceedings of the 2010 International Conference on Bioinformatics and Biomedical Technology. 2010 Presented at: ICBBT '10; April 16-18, 2010; Chengdu, China p. 101-105 URL: https://ieeexplore.ieee.org/document/5479000 [doi: 10.1109/icbbt.2010.5479000]
440. Yamaga Y, Svensson T, Chung U, Svensson AK. Association between metabolic syndrome status and daily physical activity measured by a wearable device in Japanese office workers. Int J Environ Res Public Health 2023 Mar 28;20(5):4315 [FREE Full text] [doi: 10.3390/ijerph20054315] [Medline: 36901325]
441. Yang C, Wang S, Wang W, Belcher BR, Dunton GF. Day-level associations of physical activity and sedentary time in mother-child dyads across three years: a multi-wave longitudinal study using accelerometers. J Behav Med 2022 Oct;45(5):702-715 [FREE Full text] [doi: 10.1007/s10865-022-00335-0] [Medline: 35753007]
442. Yang W, Li MH, Yu JJ, Wong SH, Sum RK, Sit CH. The associations between accelerometer-measured physical activity levels and mental health in children and adolescents with intellectual disabilities during the COVID-19 pandemic. J Autism Dev Disord 2023 Dec;53(12):4809-4821 [FREE Full text] [doi: 10.1007/s10803-022-05777-8] [Medline: 36181649]
443. Yen H, Liao Y, Huang H. Smart wearable device users' behavior is essential for physical activity improvement. Int J Behav Med 2022 Jun;29(3):278-285. [doi: 10.1007/s12529-021-10013-1] [Medline: 34363130]
444. Yerrakalva D, Hajna S, Wijndaele K, Dempsey PC, Westgate K, Wareham N, et al. Bidirectional associations of accelerometer-assessed physical activity and sedentary time with physical function among older English adults: the EPIC-Norfolk cohort study. Eur J Ageing 2022 Dec;19(4):1507-1517 [FREE Full text] [doi: 10.1007/s10433-022-00733-y] [Medline: 36506675]
445. Yi L, Mason TB, Yang C, Chu D, Dunton GF. Longitudinal associations between neighborhood park and open space access and children's accelerometer-assessed measured physical activity: the evidence from the MATCH study. J Phys Act Health 2021 Oct 01;18(9):1058-1066 [FREE Full text] [doi: 10.1123/jpah.2021-0177] [Medline: 34198261]
446. Yoshioka M, Ayabe M, Yahiro T, Higuchi H, Higaki Y, St-Amand J, et al. Long-period accelerometer monitoring shows the role of physical activity in overweight and obesity. Int J Obes (Lond) 2005 May;29(5):502-508. [doi: 10.1038/sj.ijo.0802891] [Medline: 15672105]
447. You Y, Chen Y, Wang X, Wei M, Zhang Q, Cao Q. Accelerometer-measured physical activity patterns are associated with phenotypic age: isotemporal substitution effects. Heliyon 2023 Oct;9(9):e19158 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e19158] [Medline: 37810111]
448. Yu X, Hao L, Crainiceanu C, Leroux A. Occupational determinants of physical activity at work: evidence from wearable accelerometer in 2005-2006 NHANES. SSM Popul Health 2022 Mar;17:100989 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100989] [Medline: 34977325]

XSL•FO
RenderX

449. Zhang X, Yang Z, Du L, Xiong C, Wang Z, Pan Y, et al. Associations of accelerometer-measured physical activity, sedentary behavior with self-reported insomnia in older women: does pattern matter? Sleep Med 2023 May;104:58-63. [doi: 10.1016/j.sleep.2023.02.015] [Medline: 36906996]
450. Zhong Q, Zhou R, Huang Y, Chen H, Liu H, Huang Z, et al. The independent and joint association of accelerometer-measured physical activity and sedentary time with dementia: a cohort study in the UK Biobank. Int J Behav Nutr Phys Act 2023 May 17;20(1):59 [FREE Full text] [doi: 10.1186/s12966-023-01464-8] [Medline: 37198574]
451. Zhou W, Veliz PT, Smith EML, Chen W, Reddy RM, Larson JL. Comparison of pre-diagnosis physical activity and its correlates between lung and other cancer patients: accelerometer data from the UK Biobank prospective cohort. Int J Environ Res Public Health 2023 Jan 05;20(2):1001 [FREE Full text] [doi: 10.3390/ijerph20021001] [Medline: 36673757]
452. Zijlstra WM, Ploegstra M, Vissia-Kazemier T, Roofthooft MT, Sarvaas GD, Bartelds B, et al. Physical activity in pediatric pulmonary arterial hypertension measured by accelerometry. a candidate clinical endpoint. Am J Respir Crit Care Med 2017 Dec 15;196(2):220-227. [doi: 10.1164/rccm.201608-1576OC] [Medline: 28178426]
453. Zisko N, Nauman J, Sandbakk SB, Aspvik NP, Salvesen Ø, Carlsen T, et al. Absolute and relative accelerometer thresholds for determining the association between physical activity and metabolic syndrome in the older adults: the Generation-100 study. BMC Geriatr 2017 May 16;17(1):109 [FREE Full text] [doi: 10.1186/s12877-017-0497-1] [Medline: 28511695]
454. Zlatar ZZ, Hays CC, Mestre Z, Campbell LM, Meloy MJ, Bangen KJ, et al. Dose-dependent association of accelerometer-measured physical activity and sedentary time with brain perfusion in aging. Exp Gerontol 2019 Oct 01;125:110679 [FREE Full text] [doi: 10.1016/j.exger.2019.110679] [Medline: 31382010]
455. van Hees VT, Gorzelniak L, Dean LE, Eder M, Pias M, Taherian S, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. PLoS One 2013;8(4):e61691 [FREE Full text] [doi: 10.1371/journal.pone.0061691] [Medline: 23626718]
456. Bai J, Di C, Xiao L, Evenson KR, LaCroix AZ, Crainiceanu CM, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. PLoS One 2016;11(8):e0160644 [FREE Full text] [doi: 10.1371/journal.pone.0160644] [Medline: 27513333]
457. Danilevicz IM, van Hees VT, van der Heide FC, Jacob L, Landré B, Benadjaoud MA, et al. Measures of fragmentation of rest activity patterns: mathematical properties and interpretability based on accelerometer real life data. BMC Med Res Methodol 2024 Jun 07;24(1):132 [FREE Full text] [doi: 10.1186/s12874-024-02255-w] [Medline: 38849718]
458. Hare DJ, Jones S, Evershed K. A comparative study of circadian rhythm functioning and sleep in people with Asperger syndrome. Autism 2006 Nov;10(6):565-575. [doi: 10.1177/1362361306068509] [Medline: 17088273]
459. Hurter L, Fairclough SJ, Knowles ZR, Porcellato LA, Cooper-Ryan AM, Boddy LM. Establishing raw acceleration thresholds to classify sedentary and stationary behaviour in children. Children (Basel) 2018 Dec 19;5(12):172 [FREE Full text] [doi: 10.3390/children5120172] [Medline: 30572683]
460. Sundararajan K, Georgievska S, Te Lindert BH, Gehrman PR, Ramautar J, Mazzotti DR, et al. Sleep classification from wrist-worn accelerometer data using random forests. Sci Rep 2021 Jan 08;11(1):24 [FREE Full text] [doi: 10.1038/s41598-020-79217-x] [Medline: 33420133]
461. Jones SE, Lane JM, Wood AR, van Hees VT, Tyrrell J, Beaumont RN, et al. Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. Nat Commun 2019 Jan 29;10(1):343 [FREE Full text] [doi: 10.1038/s41467-018-08259-7] [Medline: 30696823]
462. Jones SE, van Hees VT, Mazzotti DR, Marques-Vidal P, Sabia S, van der Spek A, et al. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. Nat Commun 2019 Apr 05;10(1):1585 [FREE Full text] [doi: 10.1038/s41467-019-09576-1] [Medline: 30952852]
463. Abraham B, Duong D, George M, Greene M, Kepley J, Rosenburg K. Using patient-generated health data to facilitate preoperative decision making for breast cancer patients. In: Proceedings of the 2017 Conference on Systems and Information Engineering Design Symposium. 2017 Presented at: SIEDS '17; April 28, 2017; Charlottesville, VA p. 271-276 URL: https://ieeexplore.ieee.org/document/7937730 [doi: 10.1109/sieds.2017.7937730]
464. Bekteshi S, Nica IG, Gakopoulos S, Konings M, Maes R, Cuyvers B, et al. Exercise load and physical activity intensity in relation to dystonia and choreoathetosis during powered wheelchair mobility in children and youth with dyskinetic cerebral palsy. Disabil Rehabil 2022 Aug;44(17):4794-4805. [doi: 10.1080/09638288.2021.1921064] [Medline: 33970729]
465. Jamnick NA, Pettitt RW, Granata C, Pyne DB, Bishop DJ. An examination and critique of current methods to determine exercise intensity. Sports Med 2020 Oct;50(10):1729-1756. [doi: 10.1007/s40279-020-01322-8] [Medline: 32729096]
466. Difrancesco S, Riese H, Merikangas KR, Shou H, Zipunnikov V, Antypa N, et al. Sociodemographic, health and lifestyle, sampling, and mental health determinants of 24-hour motor activity patterns: observational study. J Med Internet Res 2021 Feb 17;23(2):e20700 [FREE Full text] [doi: 10.2196/20700] [Medline: 33595445]
467. Ghosal R, Varma VR, Volfson D, Urbanek J, Hausdorff JM, Watts A, et al. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in Alzheimer's disease. Sci Rep 2022 Jul 07;12(1):11558 [FREE Full text] [doi: 10.1038/s41598-022-15528-5] [Medline: 35798763]
468. Goldsmith J, Liu X, Jacobson JS, Rundle A. New insights into activity patterns in children, found using functional data analyses. Med Sci Sports Exerc 2016 Sep;48(9):1723-1729 [FREE Full text] [doi: 10.1249/MSS.0000000000000968] [Medline: 27183122]

XSL•FO
RenderX

469. Zeitzer JM, David R, Friedman L, Mulin E, Garcia R, Wang J, et al. Phenotyping apathy in individuals with Alzheimer disease using functional principal component analysis. Am J Geriatr Psychiatry 2013 Apr;21(4):391-397 [FREE Full text] [doi: 10.1016/j.jagp.2012.12.012] [Medline: 23498386]
470. Lin W, Zou J, Di C, Sears DD, Rock CL, Natarajan L. Longitudinal associations between timing of physical activity accumulation and health: application of functional data methods. Stat Biosci 2023 Jul;15(2):309-329 [FREE Full text] [doi: 10.1007/s12561-022-09359-1] [Medline: 37383028]
471. Zeitzer JM, Blackwell T, Hoffman AR, Cummings S, Ancoli-Israel S, Stone K, Osteoporotic Fractures in Men (MrOS) Study Research Group. Daily patterns of accelerometer activity predict changes in sleep, cognition, and mortality in older men. J Gerontol A Biol Sci Med Sci 2018 Apr 17;73(5):682-687 [FREE Full text] [doi: 10.1093/gerona/glw250] [Medline: 28158467]
472. Goldsmith J, Zipunnikov V, Schrack J. Generalized multilevel function-on-scalar regression and principal component analysis. Biometrics 2015 Jun;71(2):344-353 [FREE Full text] [doi: 10.1111/biom.12278] [Medline: 25620473]
473. Bagalà F, Becker C, Cappello A, Chiari L, Aminian K, Hausdorff JM, et al. Evaluation of accelerometer-based fall detection algorithms on real-world falls. PLoS One 2012;7(5):e37062 [FREE Full text] [doi: 10.1371/journal.pone.0037062] [Medline: 22615890]
474. Li Q, Stankovic JA, Hanson MA, Barth A, Lach J, Zhou G. Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In: Proceedings of the 6th International Workshop on Wearable and Implantable Body Sensor Networks. 2009 Presented at: BSN '09; June 3-5, 2009; Berkeley, CA p. 138-143 URL: https://ieeexplore.ieee.org/document/5226903 [doi: 10.1109/bsn.2009.46]
475. Bavan L, Surmacz K, Beard D, Mellon S, Rees J. Adherence monitoring of rehabilitation exercise with inertial sensors: a clinical validation study. Gait Posture 2019 May;70:211-217. [doi: 10.1016/j.gaitpost.2019.03.008] [Medline: 30903993]
476. Bidargaddi N, Sarela A, Klingbeil L, Karunanithi M. Detecting walking activity in cardiac rehabilitation by using accelerometer. In: Proceedings of the 3rd International Conference on Intelligent Sensors, Sensor Networks and Information. 2007 Presented at: ISSNIP '07; December 3-6, 2007; Melbourne, Australia p. 555-560 URL: https://ieeexplore.ieee.org/document/4496903 [doi: 10.1109/issnip.2007.4496903]
477. Boe AJ, McGee Koch LL, O'Brien MK, Shawen N, Rogers JA, Lieber RL, et al. Automating sleep stage classification using wireless, wearable sensors. NPJ Digit Med 2019;2:131 [FREE Full text] [doi: 10.1038/s41746-019-0210-1] [Medline: 31886412]
478. Gaur S, Gupta GP. Framework for monitoring and recognition of the activities for elderly people from accelerometer sensor data using apache spark. In: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications. 2019 Presented at: ICDSMLA '19; March 29-30, 2019; Singapore, Singapore p. 29-30 URL: https://tinyurl.com/yejz6sce [doi: 10.1007/978-981-15-1420-3_79]
479. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. SIGKDD Explor Newsl 2011 Mar 31;12(2):74-82. [doi: 10.1145/1964897.1964918]
480. Zdravevski E, Risteska Stojkoska B, Standl M, Schulz H. Automatic machine-learning based identification of jogging periods from accelerometer measurements of adolescents under field conditions. PLoS One 2017;12(9):e0184216 [FREE Full text] [doi: 10.1371/journal.pone.0184216] [Medline: 28880923]
481. Attal F, Mohammed S, Dedabrishvili M, Chamroukhi F, Oukhellou L, Amirat Y. Physical human activity recognition using wearable sensors. Sensors (Basel) 2015 Dec 11;15(12):31314-31338 [FREE Full text] [doi: 10.3390/s151229858] [Medline: 26690450]
482. Javed AR, Sarwar MU, Khan S, Iwendi C, Mittal M, Kumar N. Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition. Sensors (Basel) 2020 Apr 14;20(8):2216 [FREE Full text] [doi: 10.3390/s20082216] [Medline: 32295298]
483. Abreu M, Barandas M, Leonardo R, Gamboa H. Detailed human activity recognition based on multiple HMM. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 2019 Presented at: BIOSTEC '19; February 22-24, 2019; Prague, Czech Republic p. 171-178 URL: https://www.scitepress.org/Papers/2019/73869/73869.pdf [doi: 10.5220/0007386901710178]
484. Witowski V, Foraita R, Pitsiladis Y, Pigeot I, Wirsik N. Using hidden markov models to improve quantifying physical activity in accelerometer data - a simulation study. PLoS One 2014;9(12):e114089 [FREE Full text] [doi: 10.1371/journal.pone.0114089] [Medline: 25464514]
485. Scheffer C, Engelbrecht H, Heyns PS. A comparative evaluation of neural networks and hidden Markov models for monitoring turning tool wear. Neural Comput Appl 2005 Jun 14;14(4):325-336. [doi: 10.1007/s00521-005-0469-9]
486. Metcalf BS, Jeffery AN, Hosking J, Voss LD, Sattar N, Wilkin TJ. Objectively measured physical activity and its association with adiponectin and other novel metabolic markers: a longitudinal study in children (EarlyBird 38). Diabetes Care 2009 Mar;32(3):468-473 [FREE Full text] [doi: 10.2337/dc08-1329] [Medline: 19033408]
487. Menai M, van Hees VT, Elbaz A, Kivimaki M, Singh-Manoux A, Sabia S. Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. Sci Rep 2017 Apr 03;8:45772 [FREE Full text] [doi: 10.1038/srep45772] [Medline: 28367987]

XSL•FO
RenderX

488. Zhang Y, Yin Y, Liu J, Yang M, Liu Z, Ma X. Impact of combined theory-based intervention on psychological effects and physical activity among Chinese adolescents. Int J Environ Res Public Health 2020 Apr 27;17(9):3026 [FREE Full text] [doi: 10.3390/ijerph17093026] [Medline: 32349260]
489. Lee J, Zhang T, Chu TL, Gu X. Effects of a need-supportive motor skill intervention on children's motor skill competence and physical activity. Children (Basel) 2020 Mar 17;7(3):21 [FREE Full text] [doi: 10.3390/children7030021] [Medline: 32192010]
490. Yuki A, Otsuka R, Tange C, Nishita Y, Tomida M, Ando F, et al. Daily physical activity predicts frailty development among community-dwelling older Japanese adults. J Am Med Dir Assoc 2019 Aug;20(8):1032-1036. [doi: 10.1016/j.jamda.2019.01.001] [Medline: 30792108]
491. Ramsay JO, Silverman BW. Functional Data Analysis. New York, NY: Springer; 2005.
492. Spira AP, Zipunnikov V, Raman R, Choi J, Di J, Bai J, et al. Brain amyloid burden, sleep, and 24-hour rest/activity rhythms: screening findings from the anti-amyloid treatment in asymptomatic Alzheimer's and longitudinal evaluation of amyloid risk and neurodegeneration studies. Sleep Adv 2021;2(1):zpab015 [FREE Full text] [doi: 10.1093/sleepadvances/zpab015] [Medline: 34661109]
493. Reiss PT, Goldsmith J, Shang HL, Ogden RT. Methods for scalar-on-function regression. Int Stat Rev 2017 Aug;85(2):228-249 [FREE Full text] [doi: 10.1111/insr.12163] [Medline: 28919663]
494. Ghosal R, Varma VR, Volfson D, Hillel I, Urbanek J, Hausdorff JM, et al. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's disease. Biostatistics 2023 Jul 14;24(3):539-561 [FREE Full text] [doi: 10.1093/biostatistics/kxab041] [Medline: 36519565]
495. Tang L, Halloran S, Shi JQ, Guan Y, Cao C, Eyre J. Evaluating upper limb function after stroke using the free-living accelerometer data. Stat Methods Med Res 2020 Nov;29(11):3249-3264. [doi: 10.1177/0962280220922259] [Medline: 32441206]
496. Jeran S, Steinbrecher A, Pischon T. Prediction of activity-related energy expenditure using accelerometer-derived physical activity under free-living conditions: a systematic review. Int J Obes (Lond) 2016 Aug;40(8):1187-1197. [doi: 10.1038/ijo.2016.14] [Medline: 27163747]
497. Nightingale TE, Walhin J, Thompson D, Bilzon JL. Influence of accelerometer type and placement on physical activity energy expenditure prediction in manual wheelchair users. PLoS One 2015;10(5):e0126086 [FREE Full text] [doi: 10.1371/journal.pone.0126086] [Medline: 25955304]
498. Nightingale TE, Walhin JP, Thompson D, Bilzon JL. Predicting physical activity energy expenditure in wheelchair users with a multisensor device. BMJ Open Sport Exerc Med 2015 Aug;1(1):bmjsem-2015-000008 [FREE Full text] [doi: 10.1136/bmjsem-2015-000008] [Medline: 27900111]
499. Sardinha LB, Júdice PB. Usefulness of motion sensors to estimate energy expenditure in children and adults: a narrative review of studies using DLW. Eur J Clin Nutr 2017 Mar;71(3):331-339. [doi: 10.1038/ejcn.2017.2] [Medline: 28145419]
500. Dabare P, Wickramasinghe P, Waidyatilaka I, Devi S, Kurpad AV, Samaranayake D, et al. Prediction equation for physical activity energy expenditure in 11-13-year-old Sri Lankan children. Nutrients 2023 Feb 10;15(4):906 [FREE Full text] [doi: 10.3390/nu15040906] [Medline: 36839264]
501. Sakal C, Li T, Li J, Yang C, Li X. Association between sleep efficiency variability and cognition among older adults: cross-sectional accelerometer study. JMIR Aging 2024 Apr 04;7:e54353 [FREE Full text] [doi: 10.2196/54353] [Medline: 38596863]
502. McBeth J, Dixon WG, Moore SM, Hellman B, James B, Kyle SD, et al. Sleep disturbance and quality of life in rheumatoid arthritis: prospective mHealth study. J Med Internet Res 2022 Apr 22;24(4):e32825 [FREE Full text] [doi: 10.2196/32825] [Medline: 35451978]

Abbreviations

AC: activity count
AI: activity intensity
ENMO: Euclidean norm minus one
FDA: functional data analysis
FFT: fast Fourier transformation
FoSR: function-on-scalar regression
FPCA: functional principal component analysis
FVC: flow-volume curve
GLMM: generalized linear mixed-effects model
HMM: hidden Markov model
IS: interdaily stability
IV: intradaily variability
MANOVA: multivariate ANOVA
MET: metabolic equivalent of task

XSL•FO
RenderX

ML: machine learning
MVPA: moderate-to-vigorous physical activity
PA: physical activity
PAI: physical activity index
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
ROC: receiver operating characteristic
VM: vector magnitude

Edited by G Eysenbach; submitted 14.04.24; peer-reviewed by P Bergman, C Sakal, U Sinha; comments to author 14.05.24; revised version received 27.05.24; accepted 16.07.24; published 11.09.24.

Please cite as:
Liang YT, Wang C, Hsiao CK
Data Analytics in Physical Activity Studies With Accelerometers: Scoping Review
J Med Internet Res 2024;26:e59497
URL: https://www.jmir.org/2024/1/e59497
doi:10.2196/59497
PMID:

©Ya-Ting Liang, Charlotte Wang, Chuhsing Kate Hsiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

From Fax to Secure File Transfer Protocol: The 25-Year Evolution of Real-Time Syndromic Surveillance in England

Alex J Elliot[1,2], PhD; Helen E Hughes[1], PhD; Sally E Harcourt[1], MPH; Sue Smith[1], MPH; Paul Loveridge[1], BSc; Roger A Morbey[1,2], PhD; Amardeep Bains[1], BSc; Obaghe Edeghere[1], MBBS; Natalia R Jones[2,3], PhD; Daniel Todkill[1], MBChB; Gillian E Smith[1,2], MBBS

[1]Real-time Syndromic Surveillance Team, UK Health Security Agency, Birmingham, United Kingdom

[2]NIHR Health Protection Research Unit in Emergency Preparedness and Response, King's College London, London, United Kingdom

[3]School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom

Corresponding Author:
Alex J Elliot, PhD
Real-time Syndromic Surveillance Team
UK Health Security Agency
23 Stephenson Street
Birmingham, B2 4BH
United Kingdom
Phone: 44 1212329211
Email: alex.elliot@ukhsa.gov.uk

Abstract

The purpose of syndromic surveillance is to provide early warning of public health incidents, real-time situational awareness during incidents and emergencies, and reassurance of the lack of impact on the population, particularly during mass gatherings. The United Kingdom Health Security Agency (UKHSA) currently coordinates a real-time syndromic surveillance service that encompasses 6 national syndromic surveillance systems reporting on daily health care usage across England. Each working day, UKHSA analyzes syndromic data from over 200,000 daily patient encounters with the National Health Service, monitoring over 140 unique syndromic indicators, risk assessing over 50 daily statistical exceedances, and taking and recommending public health action on these daily. This English syndromic surveillance service had its origins as a small exploratory pilot in a single region of England in 1999 involving a new pilot telehealth service, initially reporting only on “cold or flu” calls. This pilot showed the value of syndromic surveillance in England, providing advanced warning of the start of seasonal influenza activity over existing laboratory-based surveillance systems. Since this initial pilot, a program of real-time syndromic surveillance has evolved from the single-system, -region, -indicator pilot (using manual data transfer methods) to an all-hazard, multisystem, automated national service. The suite of systems now monitors a wide range of syndromes, from acute respiratory illness to diarrhea to cardiac conditions, and is widely used in routine public health surveillance and for monitoring seasonal respiratory disease and incidents such as the COVID-19 pandemic. Here, we describe the 25-year evolution of the English syndromic surveillance system, focusing on the expansion and improvements in data sources and data management, the technological and digital enablers, and novel methods of data analytics and visualization.

(J Med Internet Res 2024;26:e58704) doi:10.2196/58704

KEYWORDS

epidemiology; population surveillance; sentinel surveillance; public health surveillance; bioterrorism; mass gathering; pandemics

Introduction

The early 2000s saw the dawn of a new era of public health threat: bioterrorism. In the United States, the September 11, 2001, attacks and anthrax mail incidents underlined a heightened terrorist threat [1,2]. The particular emphasis on the potential for the deliberate and malicious release of highly pathogenic organisms (as highlighted by the anthrax mail incident) raised awareness and concern around bioterrorism to new levels. In the United States, a novel form of public health surveillance had emerged, with an increasing focus on providing timely sources of health intelligence to strengthen monitoring of the health impact of potential bioterror attacks. This form of surveillance was termed “syndromic” as the data capture systems underpinning it primarily focused on monitoring the presentation of chief complaints, symptoms, or provisional diagnoses

XSL•FO
RenderX

presented to health care providers, rather than laboratory-confirmed cases of the disease [3-5]. Syndromic surveillance can now be defined as "the (near) real-time collection, analysis, interpretation and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human public-health threats which require effective public health action" [6].

The authors of this viewpoint paper share a combined work experience of over 150 years coordinating syndromic surveillance in England. Here we look back over the last 25 years and reflect on the evolution of syndromic surveillance in England, "from fax to SFTP (secure file transfer protocol)." We outline the journey that English syndromic surveillance has taken with respect to changes in technology, methodology, drivers and enablers, and epidemiology.

Syndromic Surveillance in England: The Early Years

In England, syndromic surveillance first started in the late 1990s as a small, opportunistic project, led by a single public health epidemiologist working in a regional field epidemiology team in the West Midlands. At the time, there had been a focus on sentinel general practitioner (GP; family physicians available to patients generally during working hours) network surveillance, which, while not classically considered "syndromic," is now considered to be part of the "syndromic toolkit."

Within the United Kingdom, the National Health Service (NHS) is a publicly funded health care system that provides free health care services at the point of use for most residents. In 1998, a new NHS telephone health helpline (NHS Direct) was being piloted across parts of England [7,8]. A number of factors converged to start the English syndromic surveillance journey: a regional director of public health with an innovative approach to public health and a public health epidemiologist with an interest in novel surveillance were working together in the West Midlands, one of the regions where the new telehealth system was being piloted. The following question was asked: "can NHS Direct call data be used for influenza surveillance to improve timeliness over existing influenza surveillance systems?"

The answer and outcome of this project was "yes"; the telehealth data showed promise for public health surveillance [9]. While NHS Direct calls specific to patients with presenting "cold or flu" symptoms were higher at times when known influenza was circulating, the timing of the increase in activity was in advance of the increase in traditional (laboratory-based) reporting for influenza [10,11]. In England, this provided a "lightbulb moment" for syndromic surveillance and the benefits it could bring to the practice of public health surveillance. It was the initial impetus for developing a national program of work over the next 2 decades.

The Major Drivers for Development

As a new and relatively unknown form of public health surveillance at that time, the early development of syndromic surveillance in England initially struggled to attract resources and support for expansion. However, over the last 2 decades, several drivers have highlighted the importance, benefits, and ongoing need for syndromic surveillance. These drivers often involved major public health "incidents" (where an unexpected organism, environmental, or other hazard has been identified but the aim is to assess the hazard and risk and prevent impact on health, eg, an outbreak) where syndromic surveillance was used to support the public health response, thereby highlighting its impact in the response to public health emergencies.

As previously described, syndromic surveillance in England was initially conceived during the late 1990s. However, its use in major public health incidents and "events" (a predictable future situation that is likely to have an impact on the health of the population, eg, a mass gathering) have strengthened confidence in and understanding of the use of syndromic surveillance, enabling the development and continuous improvement of a suite of syndromic surveillance systems over the decades.

In 2005, a large explosion and fire (Buncefield Oil Depot, Hertfordshire, United Kingdom) posed a significant and very visible threat to public health [12]. Sizeable parts of South-East England (including the Greater London urban conurbation) were exposed to a large smoke plume with a potential for air pollution causing harm to the population. At the time, the near real-time monitoring of NHS Direct telehealth calls and GP consultations were used to assess changes in health care–seeking behaviors in areas affected by the smoke plume or increases in presentations of syndromes potentially linked to smoke exposure, for example, breathing difficulty and asthma [13]. The findings from the regular analysis of data from these syndromic surveillance systems and the use of the data in the dynamic risk assessments all provided reassurance of the lack of impact on the public's health to the national incident management team. This incident demonstrated the use of syndromic time-series data in providing trend and historical context (the observed over expected) and near real-time (daily) delivery and analysis (here of daily telehealth data) to support a major incident response.

Another major early driver in the evolution of the English syndromic surveillance service was the 2009 influenza pandemic. At this time, England's Health Protection Agency (HPA) operated 2 national syndromic surveillance systems: the continuing NHS Direct telehealth system and a new expanded GP consultation surveillance system. In addition to insights from laboratory, hospital, and mortality surveillance, syndromic surveillance monitored a set of relevant indicators that provided intelligence into how the pandemic was progressing. Syndromic surveillance was able to identify and monitor very early local activity, which was initially focused in urban and deprived districts of Birmingham, in the West Midlands region of England, before syndromic surveillance then subsequently monitored the wider spread of disease throughout the country [14].

One of the biggest drivers for the expansion of the English syndromic surveillance service was the London 2012 Olympic and Paralympic Games. In total, 2 new national syndromic

XSL•FO
RenderX

surveillance systems were commissioned to provide daily data on health care–seeking trends of the population: the Emergency Department Syndromic Surveillance System (EDSSS) and the GP out-of-hours system [15,16]. A new bespoke statistical exceedance algorithm was developed to provide automated statistical alarms on syndromic data [17]. A novel risk assessment process was developed and implemented to standardize the review of syndromic exceedances and other key epidemiological signals, facilitate timely and consistent decision making including timely alerting of partners, and support better documentation of decision-making [18]. Underpinning this development program were increased resources, in the form of additional people in the team to deliver the new developments, and money to resource the expansion of new systems. A key legacy of the Games was securing a recurring increased budget to maintain the new developments, systems, ways of working, and analytical methods. The London 2012 enhanced syndromic surveillance response became "business as usual" going forward, thus providing a greater ability to support routine and seasonal surveillance of key public health threats [15].

Other major events have helped demonstrate the use and importance of syndromic surveillance in the public health system. These have included the public health response to an ash cloud from an Icelandic volcanic eruption [19]; population health monitoring during the 2022 Commonwealth Games; geopolitical events, for example, the G8 Summit [20]; the COVID-19 pandemic [21,22]; and assessments of the health impact of air pollution [23].

The Evolution of Syndromic Surveillance Systems

In England, syndromic surveillance systems have been traditionally developed at a country level rather than at the region or state level. This is possible due to the advantage of having a NHS that is freely accessible by the whole population at the point of need. The national syndromic surveillance service in England has evolved from a single system with one source of data to 6 national real-time systems with several data sources.

However, one limitation of this approach is that syndromic surveillance systems are vulnerable to changes within the NHS. The adoption of different clinical patient management systems, organizational changes, improvements to clinical guidance, and updates to clinical systems can all affect the data collected and its availability for syndromic surveillance purposes. In the last 25 years there have been multiple major changes to the syndromic surveillance systems used in England and some of the data sources underpinning those systems. Altogether, there have been 13 different incarnations of the systems and their evolving data providers (Figure 1).

The NHS Direct national telehealth service ran for over 10 years but was replaced by a new, free-to-access telehealth service, but in 2013 was replaced by a new, free to access telehealth service, NHS 111. On the face of it, syndromic surveillance outputs spanning NHS Direct and NHS 111 have remained largely similar, reporting on a selection of syndromic indicators, for example, cold or flu and cough [24]. However, the logistical work in moving between different telehealth services was substantial, requiring engagement with different stakeholders, governance requirements, data sharing agreements, data architecture and pathways, reporting tools, and adjustment of statistical baselines.

As NHS patient services evolve, syndromic surveillance systems are flexible and can capture new and potentially useful sources of health care data. An example is the NHS 111 online service, which launched nationally in 2018 as an alternative point of access for nonurgent care support [25]. This provided a unique opportunity for syndromic surveillance to capture a novel set of digital data from patients potentially not using other established NHS services. The NHS 111 online syndromic surveillance system was launched during the COVID-19 pandemic to bolster situational awareness and now supports day-to-day surveillance of both infectious and noninfectious hazards [26].

Another significant change over the years was the EDSSS. This started as a small sentinel emergency department (ED) surveillance network of approximately 30 EDs across England [15]. Initial pilot work was completed in 2011 in time for its use during the London 2012 Olympic and Paralympic Games [15]. The development of EDSSS was a close collaboration between the HPA and the Royal College of Emergency Medicine, the national professional body for emergency medicine in the United kingdom. EDSSS provided benefits for both partners: HPA developed the first ED syndromic surveillance system in England to contribute to its public health surveillance function, and the Royal College of Emergency Medicine was able to pilot a national clinical minimum dataset among EDs that participated in the EDSSS. This work provided proof of principle that the code-set used for recording clinical information could be standardized nationally [15]. The early sentinel EDSSS provided the basis for the development of the national Emergency Care Data Set, which has improved data capture and reporting across EDs in England [27]. The Emergency Care Data Set provided an opportunity to revolutionize ED syndromic surveillance in England, with the fundamental gain of EDSSS changing from a small sentinel system to a near-national system, collecting data from 150 EDs across England [28,29].

GP syndromic surveillance has also evolved. The aim of this surveillance is to monitor trends in the consulting behavior of patients using the underlying clinical codes entered into the patient record (by the GP) to track consultation rates among the registered population. In England, there are several different providers of GP electronic patient management systems. The GP syndromic surveillance system has evolved from using data from a single provider to now using data from multiple providers.

XSL•FO
RenderX

Figure 1. The evolution of syndromic surveillance systems and the underlying data streams used in the English syndromic surveillance system, 1999-2024. GP: general practitioner; NHS: National Health Service.

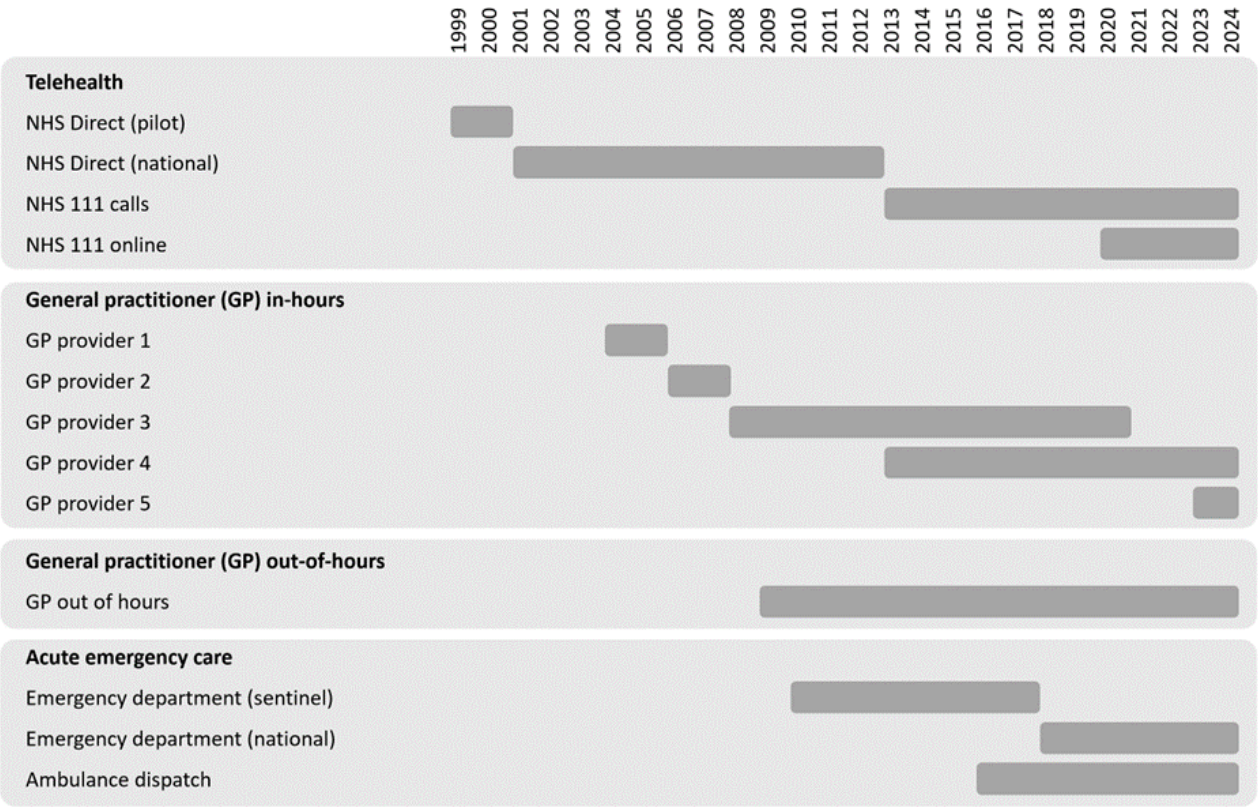

Another notable advancement during the last 25 years has been the adoption of electronic patient records across the NHS. The vast majority of patient attendances are now recorded electronically, removing the need for paper-based records to be digitalized retrospectively. This facilitates real-time surveillance as the record is immediately available for interrogation. Furthermore, the coding of patient episodes has improved over the years, as standardized coding systems have been introduced. In England, SNOMED CT (Systematized Medical Nomenclature for Medicine–Clinical Terminology) is now the NHS-wide clinical coding standard used across primary and secondary care services [30]. The result of this for syndromic surveillance is the harmonization of code lists across systems and the ability to develop expertise and knowledge of a single coding system.

The availability of electronic patient records also allows for the inclusion of more complex and varied data fields within the syndromic dataset. In England, the syndromic data extracted from patient records has expanded and evolved. In the early years, simple counts of patient episodes with a syndromic code were extracted. Now, much more detailed and granular information (still retaining anonymity) can be routinely extracted or derived, including (1) severity measures, (2) treatments, (3) ethnicity, (4) investigations, and (5) measures of deprivation [22,31,32]. These additional fields can be vital during investigations of potential threats, incidents, or events, allowing more information to feed into the epidemiological assessment of the situation, which can provide further insights into the characteristics of the affected population, allowing for improved targeted actions to be taken.

One consequence of the continual evolution and expansion of the current United Kingdom Health Security Agency (UKHSA) syndromic surveillance program is that data volumes have increased substantially. In the beginning, a pilot system, in one area of England, resulted in a small selection of indicators being analyzed each week. Now, each day, UKHSA analyzes syndromic data from over 200,000 daily NHS patient encounters, monitoring over 140 unique syndromic indicators, risk assessing more than 50 daily statistical exceedances, and alerting and taking public health action on those each day. This has required an advancement in data analytics methods to manage the expansion of data and the number of factors under investigation. Early methods included statistical process control charts and standardized incidence ratios to determine the unusual activity that needed further investigation. The bespoke statistical methods developed for the London 2012 Olympic and Paralympic Games are used in the live daily syndromic system today. This statistical method was developed to provide a single robust approach for all syndromic surveillance systems, including any future new data sources. It is sophisticated enough to automatically adjust for changes in data provider volumes, amend scale accordingly (from national to local signals), incorporate day-of-the-week effects, including public holidays, and detect both spikes and significant changes compared with previous years [17].

A further consequence of changing system data feeds (eg, due to operational or contractual reasons; Figure 1) is the challenge of analyzing data from different sources over the longer term. Following a transition to a different syndromic data feed in a syndromic surveillance system, it is possible to compare trends over the time periods covering the 2 different feeds; however,

XSL•FO
RenderX

changes in service provision at the new health care provider, different coding, or patient-presenting behaviors can make comparisons challenging. Statistical techniques can be adopted to "correct" the new data to the reporting levels of the old source; however, this also comes with difficulties and risks. Furthermore, over the 25-year span of the UKHSA syndromic surveillance program, health care services, patient health care–seeking behavior, and the general health of the population have changed significantly over time. These factors also make comparisons of contemporary and historical data challenging, as baseline levels will be different, and it is therefore difficult to interpret data and draw conclusions.

The Evolution of Technology

The early syndromic surveillance developments in England, including small-scale sentinel GP surveillance and the pilot of the NHS Direct telehealth call system, were underpinned by technologically crude data transfer methods. The surveillance of NHS Direct call data was initially established based on the faxing of weekly reports from the data provider. Each faxed report contained the number of calls for a small number of key syndromic indicators, for example, cold or flu. Data were manually entered from the physical fax into a spreadsheet. Before the advent of "big data" and "cloud-based" technologies, faxes were a standard method for securely sharing data between the data provider and the public health agency.

In England, the first advances in syndromic data-sharing technology came through routine emailing of syndromic data. At first, this continued as a manual process requiring human intervention to both send and receive the email; however, eventually, this process was adapted to use automated email procedures, removing the need for human involvement in the process. Asking busy NHS services to manually email syndromic data reports each day was not sustainable. Very quickly, automation became a standard for establishing syndromic surveillance systems in England. The subsequent development of automated processes for importing data into a database at the public health organization also removed manual resource requirements and the risk of transcription errors from the manual process.

During the 2000s, emailing of data evolved into more sophisticated transfer methods, including SFTP, only possible as the storage of data evolved from workbooks to databases. More recently, new NHS-based secure data transfer methods, for example, Message Exchange for Social Care and Health, have once again revolutionized the technology underpinning syndromic surveillance [33]. Overall, each of these technological advancements has made it possible to transfer larger and more complex sets of data more securely and reliably.

This increase in data volumes and complexity has driven the need for bigger and more efficient data storage solutions. Therefore, there has developed a need for expanded data storage capacity. During the early years, syndromic data storage consisted of Excel worksheets (Microsoft Corporation). However, as the number of systems increased and historical data accumulated, these systems quickly outgrew the capacity that Excel could provide. SQL database platforms became the standard for storing and interrogating data. To this day, SQL databases feed other investigation and statistical analysis tools that are used for analyzing and visualizing data. Similarly, in the early years, data analysis and visualization were exclusively limited to Excel. However, this has evolved, and now a suite of different data analytics tools is used for analysis and visualization. Open access tools, such as R Studio (Posit team), provide a wide range of analytical and statistical tools with the ability to deliver outputs and visualizations. In UKHSA, syndromic surveillance data are now processed using SQL for storage, R Studio for statistical exceedances, and a combination of Microsoft Excel, R Studio, and Power BI (Microsoft Corporation) for interrogating data through dashboards and other digital tools.

In England, the format for reporting and disseminating syndromic surveillance data and intelligence is one area that has remained relatively unchanged as other surveillance stages have evolved. The early years saw the development of static reports, published in PDF format, containing a collection of surveillance data charts, key messages, caveats and limitations, and background information to the syndromic system [34]. This approach remained largely the same to this day [35]. However, the COVID-19 pandemic saw an appetite for disseminating public health surveillance data and intelligence through dashboards, and for UKHSA, this is becoming the standard approach to sharing and reporting surveillance insights and intelligence. This provides a more dynamic user-friendly platform with the ability to customize data interrogation and reports and extract underlying data. For UKHSA syndromic surveillance, this is also the future direction, and plans are underway to integrate key syndromic outputs into the UKHSA data dashboard [36].

Syndromic Collaboration

Over the last 25 years, collaboration has been vital in the successful development and expansion of real-time syndromic surveillance in England. It cannot be understated how important collaboration with data provider organizations has been to the successful evolution of our syndromic surveillance service. Often, these relationships and collaborations are built not on contractual or transactional relationships, and purchasing of data, but on a shared understanding of the importance of the work and the public health benefit and impact. Several of the provider collaborations in England have been established now for 10-15 years. During this time, relationships, understanding, and trust are built, which are essential for a positive and fruitful collaboration and successful syndromic surveillance system. This factor is often overlooked in favor of a focus on data and technology, but without the "people," the best data and technology platforms do not automatically mean success [37].

Outside of the data provider relationships, the English syndromic surveillance service has benefited from a valuable network of international collaborators who have helped shape the program. The International Society for Disease Surveillance operated from 2005 to 2019, providing excellent contacts for the English syndromic program and linking the team with a range of technical and expert syndromic support networks. The annual

XSL•FO
RenderX

International Society for Disease Surveillance conference provided a platform for collaboration, sharing innovation and experience, and showcasing excellence [38].

In Europe, the wider development and implementation of syndromic surveillance has been less common across individual countries compared with the growth across individual states in the United States. France, however, was an early European innovator in syndromic surveillance, providing a blueprint from which the English ED syndromic surveillance was developed. The 2015 European-wide "Triple-S" project provided a framework for syndromic surveillance in Europe, which led to a collaborative approach to coordinating syndromic surveillance across the continent [6]. Despite Triple-S providing an impetus for developing syndromic surveillance across Europe, the further development of a European-wide syndromic surveillance network could not take advantage of the momentum of Triple-S due to a lack of funding. However, an informal cross-country collaborative Euro-syndromic network remains a legacy of Triple-S, with cross-country collaboration across several European countries and the US Centers for Disease Control and Prevention [23,39].

Closer to home, syndromic work across the United Kingdom's devolved administrations (England, Scotland, Wales, and Northern Ireland) has been productive over the years. Joint working has delivered cross-border projects and strengthened the UK response to events such as large mass gatherings [20,39]. There are continuing UK collaborations to share knowledge and harmonize syndromic approaches and systems to provide a better UK-focused syndromic approach in the future.

The Next 25 Years

Over the course of the professional careers of the authors, syndromic surveillance in England has evolved from the exploration of a novel surveillance pilot project to a mainstay source of national public health intelligence. It has developed from an innovative source of data examining seasonal influenza surveillance only to an all-hazard surveillance service, supporting a wide range of infectious and noncommunicable disease surveillance programs.

However, syndromic surveillance cannot "sit on its laurels" and stay still. Over the next 25 years, syndromic surveillance must move with the times and continue to evolve at pace to tackle future emerging health threats. Syndromic surveillance needs to continue to innovate and take advantage of new advances in technology and analytical methodologies. Importantly, this evolution must also prioritize the "human element" to develop and train the next generation of leaders and experts in syndromic surveillance to drive forward these future programs.

One key priority must be the standardization and harmonization of systems across countries and continents. Currently, different countries operate different suites of syndromic surveillance systems, even where there are similarities between countries with respect to the type of surveillance undertaken. For example, within ED syndromic surveillance systems, although the general concept of an ED may broadly be the same in all locations (providing acute emergency care), there are disparities in the syndromes monitored (which may be due to the availability of the clinical information and coding formats used); the methods used for analysis and interpretation of results; and the outputs delivered to both public health professionals, health care providers in general, as well as to the public. While there will always be a level of autonomy required across each country, there are opportunities to develop minimum and common datasets to aid the interpretation of syndromic data across countries rather than between countries.

Delivering such harmonization will require the strengthening of international networks and the formation of a network of excellence, with those leading exponents of syndromic surveillance enabling knowledge mobilization and delivering plans for strengthening harmonization. However, such networks require resources, support, and dedication to deliver a structured program of development. There is also a role for syndromic surveillance in supporting the model of a national biosurveillance network, which has collaborative surveillance at its heart, given its multisource and multisectoral surveillance approach [40]. Existing international networks, such as the World Health Organization (WHO) Epidemic Intelligence from Open Sources initiative, may also play a future role in fostering these collaborative networks and nurturing technological and methodological advances [41].

Syndromic surveillance is on a new wave of technology-driven growth, which will almost certainly be underpinned by cloud-based storage and machine learning among others. While each of these represents an opportunity to improve surveillance, we must not get distracted from the basics of epidemiological surveillance and analysis and the fundamental aims and objectives of the surveillance we undertake.

Data protection, security, and governance have also played an influencing role in syndromic surveillance over the last 2 decades. On reflection, during the early stages of syndromic surveillance, there was a lower "threshold" of control around the use of patient data for public health research and surveillance. However, in more modern times, there is a wider understanding of the ethics and risks (but also the benefits) of using patient data, and as such, the governance framework has been strengthened around surveillance systems and public health organizations as a whole. In the general population, this has also led to a far greater understanding about how an individual's health data are used. The controls and restrictions around the use of patient (and particularly identifiable) data are now far stricter. While this direction of travel has been a necessary and positive move, it does mean that for syndromic surveillance (where anonymized health record data are used the majority of the time), it is far more difficult and complex to access the health data that underpin the systems.

One final future aspiration of UKHSA syndromic surveillance is to support and facilitate the expansion of syndromic surveillance in resource-poor countries. This is still an area where syndromic surveillance in general has poor uptake, likely due to multiple factors, such as resource constraints issues, underdeveloped health and public health systems, and limited capacity and capability to establish and maintain syndromic surveillance systems. There are future opportunities to engage

XSL•FO
RenderX

in knowledge exchange with other countries and expand work already ongoing to spread the benefits of syndromic surveillance globally [42,43].

Summary

Over the last 25 years, the English national syndromic surveillance program has evolved from a small pilot project to a national real-time multisystem all-hazard program. This journey started with an innovative public health doctor (GES) taking a local idea, which, thanks to the initial collaboration with local public health experts, has grown into a national surveillance program with global partnership. It has been an exemplar of how opportunity, chance, innovation, technology, and public health expertise can come together to materialize into a world-leading service, supporting complex and varied public health incidents and events.

Acknowledgments

We thank the current syndromic data providers, that are NHS 111 and National Health Service (NHS) England; out-of-hours providers submitting data to the general practitioner out-of-hours system; emergency department clinicians and NHS Trusts and NHS England supporting Emergency Department Syndromic Surveillance System; and participating The Phoenix Partnership and the Oxford-Royal College of General Practitioner Clinical Informatics Digital Hub practices supporting general practitioner in-hours, ambulance trusts, and the Association of the Ambulance Chief Executives. We also thank the wider United Kingdom Health Security Agency (UKHSA) Real-Time Syndromic Surveillance Team for their technical expertise in delivering the daily syndromic service. We acknowledge Maureen Baker, Brian McCloskey, Sue Ibbotson, Rod Griffiths, Mary O'Mahony, and Duncan Cooper for historical support of the English national syndromic surveillance service and their contribution to a wider piece of work documenting the history of syndromic surveillance in England. AJE, RAM, NRJ, and GES are affiliated with the National Institute for Health and Care Research Health Protection Research Unit (NIHR HPRU) in Emergency Preparedness and Response at King's College London in partnership with UKHSA. AJE, GES, and HEH are affiliated with the NIHR HPRU in Gastrointestinal Infections at the University of Liverpool in partnership with UKHSA. The views expressed in this article are those of the authors and not necessarily those of the NIHR, UKHSA, or the Department of Health and Social Care.

Authors' Contributions

AJE and GES conceptualized the work and drafted the initial manuscript. AJE, HEH, SEH, SS, PL, RAM, AB, NRJ, OE, DT, and GES reviewed and revised subsequent drafts and approved the final version for submission.

Conflicts of Interest

None declared.

References

1. Das D, Weiss D, Mostashari F, Treadwell T, McQuiston J, Hutwagner L, et al. Enhanced drop-in syndromic surveillance in New York City following September 11, 2001. J Urban Health 2003 Jun;80(2 Suppl 1):i76-i88 [FREE Full text] [doi: 10.1007/pl00022318] [Medline: 12791782]
2. Borio L, Frank D, Mani V, Chiriboga C, Pollanen M, Ripple M, et al. Death due to bioterrorism-related inhalational anthrax: report of 2 patients. JAMA 2001 Nov 28;286(20):2554-2559. [doi: 10.1001/jama.286.20.2554] [Medline: 11722269]
3. Heffernan R, Mostashari F, Das D, Besculides M, Rodriguez C, Greenko J, et al. New York City syndromic surveillance systems. MMWR Suppl 2004;53:23-27 [FREE Full text] [Medline: 15714622]
4. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis 2004;10(5):858-864. [doi: 10.3201/eid1005.030646] [Medline: 15200820]
5. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health 2001;1:9 [FREE Full text] [doi: 10.1186/1471-2458-1-9] [Medline: 11722798]
6. Triple S Project. Assessment of syndromic surveillance in Europe. Lancet 2011;378(9806):1833-1834. [doi: 10.1016/S0140-6736(11)60834-9] [Medline: 22118433]
7. Pencheon D. NHS direct: managing demand. BMJ 1998;316(7126):215-216 [FREE Full text] [doi: 10.1136/bmj.316.7126.215a] [Medline: 9468696]
8. Munro J, Nicholl J, O'Cathain A, Knowles E. Impact of NHS direct on demand for immediate care: observational study. BMJ 2000;321(7254):150-153 [FREE Full text] [doi: 10.1136/bmj.321.7254.150] [Medline: 10894694]
9. Harcourt SE, Smith GE, Hollyoak V, Joseph CA, Chaloner R, Rehman Y, et al. Can calls to NHS direct be used for syndromic surveillance? Commun Dis Public Health 2001;4(3):178-182. [Medline: 11732356]
10. Cooper DL, Verlander NQ, Elliot AJ, Joseph CA, Smith GE. Can syndromic thresholds provide early warning of national influenza outbreaks? J Public Health (Oxf) 2009;31(1):17-25. [doi: 10.1093/pubmed/fdm068] [Medline: 18032426]
11. Cooper DL, Smith GE, Hollyoak VA, Joseph CA, Johnson L, Chaloner R. Use of NHS direct calls for surveillance of influenza--a second year's experience. Commun Dis Public Health 2002;5(2):127-131. [Medline: 12166298]

XSL•FO
RenderX

12. Mohan R, Walton HA, Thomson D, Webster H, Wilkinson P, Grundy C, et al. The buncefield oil depot fire of 2005: potential air-pollution health impacts under alternative meteorological scenarios. PLoS Curr 2012;4:RRN1300 [FREE Full text] [doi: 10.1371/currents.RRN1300] [Medline: 22453896]
13. Smith G, Hippisley-Cox J, Harcourt S, Heaps M, Painter M, Porter A, et al. Developing a national primary care-based early warning system for health protection--a surveillance tool for the future? Analysis of routinely collected data. J Public Health (Oxf) 2007;29(1):75-82. [doi: 10.1093/pubmed/fdl078] [Medline: 17158478]
14. Smith S, Smith G, Olowokure B, Ibbotson S, Foord D, Maguire H, et al. Early spread of the 2009 influenza A(H1N1) pandemic in the United Kingdom--use of local syndromic data, May-August 2009. Euro Surveill 2011;16(3):19771 [FREE Full text] [Medline: 21262185]
15. Elliot AJ, Hughes HE, Hughes TC, Locker TE, Shannon T, Heyworth J, et al. Establishing an emergency department syndromic surveillance system to support the London 2012 Olympic and Paralympic games. Emerg Med J 2012;29(12):954-960. [doi: 10.1136/emermed-2011-200684] [Medline: 22366039]
16. Harcourt SE, Fletcher J, Loveridge P, Bains A, Morbey R, Yeates A, et al. Developing a new syndromic surveillance system for the London 2012 Olympic and Paralympic Games. Epidemiol Infect 2012;140(12):2152-2156 [FREE Full text] [doi: 10.1017/S0950268812001781] [Medline: 22892324]
17. Morbey RA, Elliot AJ, Charlett A, Verlander NQ, Andrews N, Smith GE. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England. Bioinformatics 2015;31(22):3660-3665. [doi: 10.1093/bioinformatics/btv418] [Medline: 26198105]
18. Smith GE, Elliot AJ, Ibbotson S, Morbey R, Edeghere O, Hawker J, et al. Novel public health risk assessment process developed to support syndromic surveillance for the 2012 Olympic and Paralympic Games. J Public Health (Oxf) 2017;39(3):e111-e117 [FREE Full text] [doi: 10.1093/pubmed/fdw054] [Medline: 27451417]
19. Elliot AJ, Singh N, Loveridge P, Harcourt S, Smith S, Pnaiser R, et al. Syndromic surveillance to assess the potential public health impact of the Icelandic volcanic ash plume across the United Kingdom, April 2010. Euro Surveill 2010;15(23):19583 [FREE Full text] [Medline: 20546694]
20. Meyer N, McMenamin J, Robertson C, Donaghy M, Allardice G, Cooper D. A multi-data source surveillance system to detect a bioterrorism attack during the G8 summit in Scotland. Epidemiol Infect 2008;136(7):876-885 [FREE Full text] [doi: 10.1017/S0950268807009132] [Medline: 17678556]
21. Elliot AJ, Harcourt SE, Hughes HE, Loveridge P, Morbey RA, Smith S, et al. The COVID-19 pandemic: a new challenge for syndromic surveillance. Epidemiol Infect 2020;148:e122 [FREE Full text] [doi: 10.1017/S0950268820001314] [Medline: 32614283]
22. Hughes HE, Hughes TC, Morbey R, Challen K, Oliver I, Smith GE, et al. Emergency department use during COVID-19 as described by syndromic surveillance. Emerg Med J 2020;37(10):600-604 [FREE Full text] [doi: 10.1136/emermed-2020-209980] [Medline: 32948621]
23. Hughes HE, Morbey R, Fouillet A, Caserio-Schönemann C, Dobney A, Hughes TC, et al. Retrospective observational study of emergency department syndromic surveillance data during air pollution episodes across London and Paris in 2014. BMJ Open 2018;8(4):e018732 [FREE Full text] [doi: 10.1136/bmjopen-2017-018732] [Medline: 29674360]
24. Harcourt SE, Morbey RA, Loveridge P, Carrilho L, Baynham D, Povey E, et al. Developing and validating a new national remote health advice syndromic surveillance system in England. J Public Health (Oxf) 2017 Mar 01;39(1):184-192 [FREE Full text] [doi: 10.1093/pubmed/fdw013] [Medline: 26956114]
25. Armstrong S. The apps attempting to transfer NHS 111 online. BMJ 2018;360:k156. [doi: 10.1136/bmj.k156] [Medline: 29335297]
26. Elliot AJ, Bennett CD, Hughes HE, Morbey RA, Todkill D, Thompson R, et al. Spike in asthma healthcare presentations in eastern England during June 2021: a retrospective observational study using syndromic surveillance data. Int J Environ Res Public Health 2021;18(23):12353 [FREE Full text] [doi: 10.3390/ijerph182312353] [Medline: 34886084]
27. Emergency care data set (ECDS). NHS England. 2024. URL: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/emergency-care-data-set-ecds [accessed 2024-03-25]
28. Morbey R, Hughes H, Smith G, Challen K, Hughes TC, Elliot AJ. Potential added value of the new emergency care dataset to ED-based public health surveillance in England: an initial concept analysis. Emerg Med J 2019;36(8):459-464. [doi: 10.1136/emermed-2018-208323] [Medline: 31253597]
29. Hughes HE, Hughes TC, Haile A, Smith GE, McCloskey B, Elliot AJ. Syndromic surveillance revolution? Public health benefits of modernizing the emergency care patient health record in England. Public Health Rep 2017;132(1_suppl):12S-15S [FREE Full text] [doi: 10.1177/0033354917706955] [Medline: 28692387]
30. SNOMED CT.: NHS England; 2024. URL: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct [accessed 2024-04-16]
31. Smith S, Morbey R, de Lusignan S, Pebody RG, Smith GE, Elliot AJ. Investigating regional variation of respiratory infections in a general practice syndromic surveillance system. J Public Health (Oxf) 2021;43(2):e153-e160. [doi: 10.1093/pubmed/fdaa014] [Medline: 32009178]

XSL•FO
RenderX

32. Todkill D, Loveridge P, Elliot AJ, Morbey R, Lusignan SD, Edeghere O, et al. Socioeconomic and geographical variation in general practitioner consultations for allergic rhinitis in England, 2003-2014: an observational study. BMJ Open 2017;7(8):e017038 [FREE Full text] [doi: 10.1136/bmjopen-2017-017038] [Medline: 28801431]
33. Message exchange for social care and health.: NHS England; 2024. URL: https://digital.nhs.uk/services/message-exchange-for-social-care-and-health-mesh [accessed 2024-03-25]
34. NHS direct syndromic surveillance bulletin.: Health Protection Agency; 2010. URL: https://webarchive.nationalarchives.gov.uk/ukgwa/20100703062652/http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/RealtimeSyndromicSurveillance/PrimaryCareSurveillanceNHSDirectSyndromicSurveillance/ [accessed 2024-03-25]
35. Remote health advice syndromic surveillance system.: UK Health Security Agency; 2024. URL: https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses#remote-health-advice-syndromic-surveillance-system [accessed 2024-03-25]
36. UKHSA data dashboard.: UK Health Security Agency; 2024. URL: https://ukhsa-dashboard.data.gov.uk/ [accessed 2024-03-25]
37. Smith GE, Elliot AJ, Lake I, Edeghere O, Morbey R, Catchpole M, et al. Syndromic surveillance: two decades experience of sustainable systems - its people not just data!. Epidemiol Infect 2019 Jan;147:e101 [FREE Full text] [doi: 10.1017/S0950268819000074] [Medline: 30869042]
38. Baer A, Olson D, Brownstein J. Editorial: International Society for Disease Surveillance Conference 2010: enhancing the synergy between research, informatics, and practice in public health. Emerg Health Threats J 2011;4:7172 [FREE Full text] [doi: 10.3402/ehtj.v4i0.7172] [Medline: 24149037]
39. Hughes HE, Colón-González FJ, Fouillet A, Elliot AJ, Caserio-Schonemann C, Hughes TC, et al. The influence of a major sporting event upon emergency department attendances; a retrospective cross-national European study. PLoS One 2018;13(6):e0198665 [FREE Full text] [doi: 10.1371/journal.pone.0198665] [Medline: 29898000]
40. Defining collaborative surveillance: a core concept for strengthening the global architecture for health emergency preparedness, response, and resilience (HEPR). World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240074064 [accessed 2024-04-16]
41. Yanagawa M, Lorenzo JC, Fukusumi M, Shimada T, Kasamatsu A, Ota M, et al. Use of epidemic intelligence from open sources for global event-based surveillance of infectious diseases for the Tokyo 2020 olympic and paralympic games. Western Pac Surveill Response J 2022;13(3):1-7 [FREE Full text] [doi: 10.5365/wpsar.2022.13.3.959] [Medline: 36688182]
42. May L, Chretien JP, Pavlin JA. Beyond traditional surveillance: applying syndromic surveillance to developing settings--opportunities and challenges. BMC Public Health 2009;9:242 [FREE Full text] [doi: 10.1186/1471-2458-9-242] [Medline: 19607669]
43. Diwan V, Agnihotri D, Hulth A. Collecting syndromic surveillance data by mobile phone in rural India: implementation and feasibility. Glob Health Action 2015;8:26608 [FREE Full text] [doi: 10.3402/gha.v8.26608] [Medline: 25843499]

Abbreviations

ED: emergency department
EDSSS: Emergency Department Syndromic Surveillance System
GP: general practitioner
HPA: Health Protection Agency
NHS: National Health Service
SFTP: secure file transfer protocol
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms
UKHSA: United Kingdom Health Security Agency
WHO: World Health Organization

Edited by G Eysenbach; submitted 18.04.24; peer-reviewed by J Mellor, U Hoang; comments to author 10.05.24; revised version received 04.06.24; accepted 11.06.24; published 17.09.24.

Please cite as:
Elliot AJ, Hughes HE, Harcourt SE, Smith S, Loveridge P, Morbey RA, Bains A, Edeghere O, Jones NR, Todkill D, Smith GE
From Fax to Secure File Transfer Protocol: The 25-Year Evolution of Real-Time Syndromic Surveillance in England
J Med Internet Res 2024;26:e58704
URL: https://www.jmir.org/2024/1/e58704
doi:10.2196/58704
PMID:

XSL•FO
RenderX

©Alex J Elliot, Helen E Hughes, Sally E Harcourt, Sue Smith, Paul Loveridge, Roger A Morbey, Amardeep Bains, Obaghe Edeghere, Natalia R Jones, Daniel Todkill, Gillian E Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

We Have Spent Time, Money, and Effort Making Self-Help Digital Mental Health Interventions: Is Anyone Going to Come to the Party?

Skye Fitzpatrick[1], PhD; Alexander O Crenshaw[2], PhD; Victoria Donkin[3], BA; Alexis Collins[3], BA; Angela Xiang[3], BA; Elizabeth A Earle[1], BA; Kamya Goenka[1], BA; Sonya Varma[1], MA; Julianne Bushe[3], BA; Tara McFadden[4], MEd; Andrea Librado[4], PsyD; Candice Monson[3], PhD

[1]York University, Toronto, ON, Canada
[2]Kennesaw State University, Kennesaw, GA, United States
[3]University of Waterloo, Waterloo, ON, Canada
[4]Atlas Institute for Veterans and Families, Ottawa, ON, Canada

Corresponding Author:
Skye Fitzpatrick, PhD
York University
4700 Keele St
Toronto, ON, M3J1P3
Canada
Phone: 1 4167362100 ext 66214
Email: skyefitz@yorku.ca

Abstract

Although efficacious psychotherapies exist, a limited number of mental health care providers and significant demand make their accessibility a fundamental problem. Clinical researchers, funders, and investors alike have converged on self-help digital mental health interventions (self-help DMHIs) as a low-cost, low-burden, and broadly scalable solution to the global mental health burden. Consequently, exorbitant financial and time-based resources have been invested in developing, testing, and disseminating these interventions. However, the public's assumed desirability for self-help DMHIs by experts has largely proceeded without question. This commentary critically evaluates whether self-help DMHIs can, and will, reach their purported potential as a solution to the public burden of mental illness, with an emphasis on evaluating their real-world desirability. Our review finds that self-help DMHIs are often perceived as less desirable and credible than in-person treatments, with lower usage rates and, perhaps accordingly, clinical trials testing self-help DMHIs suffering from widespread recruitment challenges. We highlight two fundamental challenges that may be interfering with the desirability of, and engagement in, self-help DMHIs: (1) difficulty competing with technology companies that have advantages in resources, marketing, and user experience design (but may not be delivering evidence-based interventions) and (2) difficulty retaining (vs initially attracting) users. We discuss a range of potential solutions, including highlighting self-help DMHIs in public mental health awareness campaigns; public education about evidence-based interventions that can guide consumers to appropriate self-help DMHI selection; increased financial and expert support to clinical researchers for marketing, design, and user experience in self-help DMHI development; increased involvement of stakeholders in the design of self-help DMHIs; and investing in more research on ways to improve retention (versus initial engagement). We suggest that, through these efforts, self-help DMHIs may fully realize their promise for reducing the global burden of mental illness.

(J Med Internet Res 2024;26:e58198) doi:10.2196/58198

KEYWORDS

online interventions; self-help; digital interventions; mental health; psychotherapy; intervention desirability

XSL•FO
RenderX

We Have Spent Time, Money, and Effort Making Online Self-Help Interventions: Is Anyone Going to Use Them?

Since their inception, psychotherapy interventions have offered inherently individual-level solutions for what are societally pervasive problems of mental illness. For example, early psychoanalytic interventions were offered to members of the Austrian elite and required exorbitant provider resources (eg, up to 6 to 10 hours per week for 6 months to 3 years) [1]. Given this, it is probably safe to say that "accessible mental health for all" was not the original framework upon which many psychological interventions were built. Although individually beneficial, the inadequacy of psychotherapy in combating the global burden of mental illness is clear; more than 70% of people with mental illness do not receive any mental health intervention [2]. In response to this problem, contemporary interventionists have sought to translate inaccessible psychological interventions for mass consumption. Given their low-cost, low-burden, and broadly scalable potential, one proposed solution that appears particularly promising is self-help digital mental health interventions (self-help DMHIs).

Self-help DMHIs are a class of interventions that are self-directed and delivered over online platforms. They often combine psychoeducational materials (in text, video, or audio format) and explicit practice and implementation of psychotherapeutic concepts in a structured format. They may also leverage web browsers, apps, or other forms of digital technologies (eg, texting) as their mode of delivery. Self-help DMHIs may or may not involve some minimal contact with a paraprofessional (eg, a "coach") but, crucially, do not involve extensive one-on-one time with a mental health care professional. We argue that, once a mental health care professional becomes involved in delivering an intervention, that intervention is no longer "self-help." Owing to their perceived potential, researcher, clinician, funder, and investor interest in self-help DMHIs has flourished [3], and decades of work has now been conducted in this arena. Although some experts suggest that self-help DMHIs may represent a scalable mental health solution [2,4-6], it is less clear whether they are actually desirable to the public. As the time, money, and research invested into self-help DMHIs grows, this commentary aims to take stock of what is known about their desirability and consider whether their promise has been (or can be) realized. First, we present a discussion of why self-help DMHIs may be a compelling mental health solution; review the way that they have amassed attention, resources, and investment from multiple sectors; and identify later-revealed problems in their uptake and desirability. Our review of these issues largely centers on a North American context, and we encourage more empirical attention toward whether the financial and empirical interest in self-help DMHIs, as well as difficulties with their uptake, is moderated by region of the world. We then discuss potential reasons for desirability and uptake issues in self-help DMHIs and present possible solutions.

The Rise (and Potential Fall?) of Self-Help DMHIs: A Party Planning Story

Let's Have a Party: Self-Help DMHIs Are the Future (and the Present)!

Self-help DMHIs have long been purported to be a central solution to the joint problems of mental health burden and lack of available evidence-based mental health interventions [2,4-6]. There are several reasons for this optimism. First, self-help DMHIs' reliance on technology and focus on scalability means that once built, they are relatively low-cost to maintain compared to the number of people they can serve. They may not require the resources of highly educated mental health care professionals or, if they do, typically require significantly less than face-to-face psychotherapeutic interventions. The lower clinician time and associated cost make self-help DMHIs easier to disseminate to the masses than face-to-face psychotherapeutic interventions. Second, self-help DMHIs can usually be immediately used anywhere the internet can reach. This fact means that the traditional geographic barriers that limit access to trained mental health providers can be partially overcome [5]. Third, self-help DMHIs offer the promise of addressing stigma-related barriers that may prohibit help-seeking from some people who are uncomfortable meeting with health care professionals. For example, research suggests that younger populations, including university students and especially young men, are more likely to seek online mental health support than see a mental health care professional, as it is considered to be less stigmatizing and more anonymous [7,8]. Finally, generations of people who have been exposed to the internet from childhood are now in adulthood and may be acclimatized to meeting an array of needs (eg, shopping, entertainment, socializing, and health care) via web-based applications. It therefore seems natural that some would want to meet their mental health needs in the same way. For these reasons, self-help DMHIs appear to be a clear and inevitable current and next step for mental health intervention.

The apparent potential of self-help DMHIs has grown since the onset of the COVID-19 pandemic, which has further amplified the need for low-cost, virtually delivered interventions. Rates of depression, anxiety, substance use disorder, and posttraumatic stress rose throughout the pandemic [9-11], placing an added burden on health care systems that were already strained under the weight of the virus and its sequelae. Mental health interventions that can be delivered with minimal burden to the health care system were even more sorely needed. Moreover, social distancing precautions limited the accessibility of mental health care providers in face-to-face contexts, resulting in a significant uptake in virtual or remote-delivered care. Although these changes may be partially reverting back as the pandemic recedes in public consciousness, much of the shift to virtual mental health care has endured and is expected to continue [12]. For all of these reasons, self-help DMHIs may be "the future" of mental health, if not its present.

XSL•FO
RenderX

Let's Make It a Big Party: Heavy Investment in Self-Help DMHIs

Widespread belief in the promise of self-help DMHIs is demonstrated by the exorbitant time, money, resources, and intellect that has been invested in their development. The number of peer-reviewed academic publications related to self-help DMHIs more than doubled in the 10 years from 2012-2021 compared to over the previous 10 years using the following search terms on PsycInfo in October 2022: noft((online or internet or web)) AND noft(psychotherap* or intervention or treatment or therap*) AND noft(self-help or self-guided). NOFT = anywhere except full text. These initiatives have also been backed by substantial funding. For example, our review of the National Institutes of Health RePORTER funding award database [13] found that over US $27.5 billion in grant funding across 66,201 projects was allocated to the study of self-help DMHIs in the fiscal years between 2012 and 2022, using the following search terms: limited to the following agencies: National Institute of Mental Health, National Institute on Alcohol Abuse and Alcoholism, Eunice Kennedy Shriver National Institute of Child Health and Human Development, and National Institute on Drug Abuse) on the National Institutes of Health RePORTER in August 2022: (online or remote or virtual or web-based or internet) AND (self-help or self help or self-management or self management or self-guided or self guided or self-directed or self directed).

However, clinical researchers are not alone in recognizing the promise and potential profit of self-help DMHIs. A 2021 report that analyzed market trends and investments from 2017 to 2020 suggested that the market value of mental health-related apps was US $5.2 billion, and one forecast predicted a compound annual growth rate of 16.5% from 2022 to 2030, bringing the forecasted revenue to US $17.5 billion by 2030 [14,15]. Clearly, self-help DMHIs have been recognized as not only potentially beneficial but also highly profitable. These findings document that there has been a great deal of attention, money, time, and intellect from academic, private, and industrial spheres devoted to the development of self-help DMHIs, at least in a North American context. It is possible that such a surge of financial interest in other areas of the world is more or less apparent than this, and research investigating whether a region of the world moderates the relative increase in funding for self-help DMHI work in recent years is needed.

The Party Is Starting, but Where Are the Guests? Do People Actually Want Self-Help DMHIs?

The findings reviewed above suggest that the promise of self-help DMHIs is obvious to experts and investors. In fact, such promise is so seemingly obvious that many of the assumptions underpinning it may not have been adequately tested. Indeed, the notion that self-help DMHIs are actually desirable and feasible to deliver to people is an empirical question whose answer is not yet clear.

Service users consider immediate access and low wait times to be important when receiving treatment for mental health. While users expect face-to-face therapies to meet their needs on most of these factors and demonstrate a greater preference for traditional services over self-help DMHIs [16-19], self-help DMHIs are generally perceived as more convenient and accessible than traditional methods of therapy [17]. Further, potential consumers are also aligned with experts in viewing several additional potential benefits of self-help DMHIs, such as lower costs, reduced embarrassment or stigma, and avoiding wait-times, as well as other benefits such as the ability to track progress and interactivity [20]. Some have argued that such perceived advantages of self-help DMHIs may persuade users to engage with them [21] and, indeed, preferences for self-help DMHIs relative to traditional services increase when people are asked to consider the long wait times for the latter [18].

However, the perceived benefits of self-help DMHIs may not translate to their actual use. Research demonstrates that the proportion of users stating that they would likely use a self-help DMHI for their mental health is significantly greater than the relative proportion of the same users who report having a preference for it over traditional services, or an intention to actually use a self-help DMHI when given other choices [16,22,23]. Therefore, there may be a discrepancy between people's hypothetical interest in self-help DMHIs, and their actual likelihood of engaging with it in their real lives. Such a discrepancy may be accounted for by negative perceptions of self-help DMHIs. For example, individuals perceive self-help DMHIs to be less credible relative to face-to-face interventions [17,24]. Further, while consumers recognize potential benefits of self-help DMHIs, they also have several concerns. In one study of adolescents, over half of the sample considered several perceived characteristics of self-help DMHIs to be moderately problematic, including not being able to ask questions, information feeling too general, lacking therapist support, and privacy concerns [20]. Finally, while self-help DMHIs may have the potential to reach people with elevated self-stigma related to their mental health, research suggests that, as with face-to-face interventions [25], self-stigma predicts a lower likelihood of engaging with self-help DMHIs [26].

In summary, although self-help DMHIs are widely perceived as more accessible than face-to-face mental health services, it is not yet clear whether such a perception will lead to their use. Research on the desirability of self-help DMHIs suggests that people generally do not perceive them as comparably desirable or credible, and appear less likely to use them, relative to face-to-face services. However, this body of research is small, especially compared to the relative body of self-help DMHI intervention research.

Desirability of Self-Help DMHIs in Clinical Trial Research

In order to further assess potential desirability issues in self-help DMHIs, our team conducted a review of self-help DMHI trials registered on ClinicalTrials.gov from the years 2016-2021 to determine (1) what percentage of self-help DMHI trials met recruitment number targets and (2) when studies did not meet recruitment targets, the extent to which they fell short. Although answers to these questions do not directly test desirability, they may indicate whether researchers tend to have a more difficult time engaging users in self-help DMHIs than they initially anticipated. Also, we note that ClinicalTrials.gov is a US-based website that particularly houses registered trials from North

XSL•FO
RenderX

American researchers. We selected this website as it is a particularly well-established and large database for clinical trials. However, while a review of all potential global registration sources was beyond the scope of this paper, we acknowledge that these results may not be generalizable to other regions of the world. We encourage further research into whether underrecruiting issues for self-help DMHIs generalize to other world regions.

On ClinicalTrials.gov, we used the search following terms: (“online” OR “remote” OR “web-based” OR “virtual” OR “internet”) AND (“self-help” OR “self-management” OR “self-guided” OR “self-directed”). The search focused on trials in the United States only and the study had to be marked as “completed.” Search results were then reviewed by one of the authors to determine if they met the criteria for a self-help DMHI. Uncertain classifications were presented to a group of six authors and decided by joint agreement. We referred to the earliest version of the record to determine the intended sample size and the most recent version of the record for the final sample size. We verified the final sample size through a Google Scholar search for publications linked to the clinical trials identifier. Figure 1 presents a visual representation of our findings. In total, 177 records were found, 70 of which met the criteria as a self-help DMHI. Of those 70 trials, 38 (54.3%) successfully recruited the planned sample size originally intended (the average target sample size was 220 individuals, SD 302). Of the 32 trials that did not meet recruitment targets, they recruited 65.1% of their intended sample on average.

Notably, this search did not account for trials that extended the recruitment period past the initially planned end date (eg, due to recruitment difficulties), so it may underestimate recruitment challenges. However, these findings indicate that trialists’ original estimates of the number of people who will be interested in, eligible for, and initiate participation in, the self-help DMHIs that they are testing seem to be consistently inflated. Certainly, any clinical trial failing to meet its recruitment targets is common, whether a self-help DMHI or not. Moreover, as we did not gather recruitment statistics for registered psychological intervention trials that are not DMHIs, it is difficult to clearly compare these recruitment outcomes with in-person psychotherapy research. Indeed, it is possible that DMHIs have larger recruitment targets (anticipating smaller effect sizes) which drives recruitment issues, rather than poor desirability of the intervention itself. However, unlike in-person psychotherapy trials, self-help DMHIs can often have a much broader catchment area due to not requiring participants to come into a laboratory or clinic and being able to recruit beyond the states or provinces wherein a mental health care provider is licensed [27,28]. Such consistent underrecruiting, even given these advantages, may indicate that self-help DMHIs are not as desirable as interventionists assume. These findings are concerning in a context wherein so many self-help DMHIs have already been built, tested, and are ready to be delivered.

XSL•FO
RenderX

Figure 1. Visual representation of the number of completed self-help DMHI trials registered on ClinicalTrials.gov that met recruitment targets. Note: Of the trials that did not meet their recruitment targets, the percentage of their intended sample size recruited ranged from 14.4% to 96%. DMHI: digital mental health intervention. Note: Of the trials did that did not meet their recruitment targets, the percentage of their intended sample size recruited ranged from 14.4% to 96%.

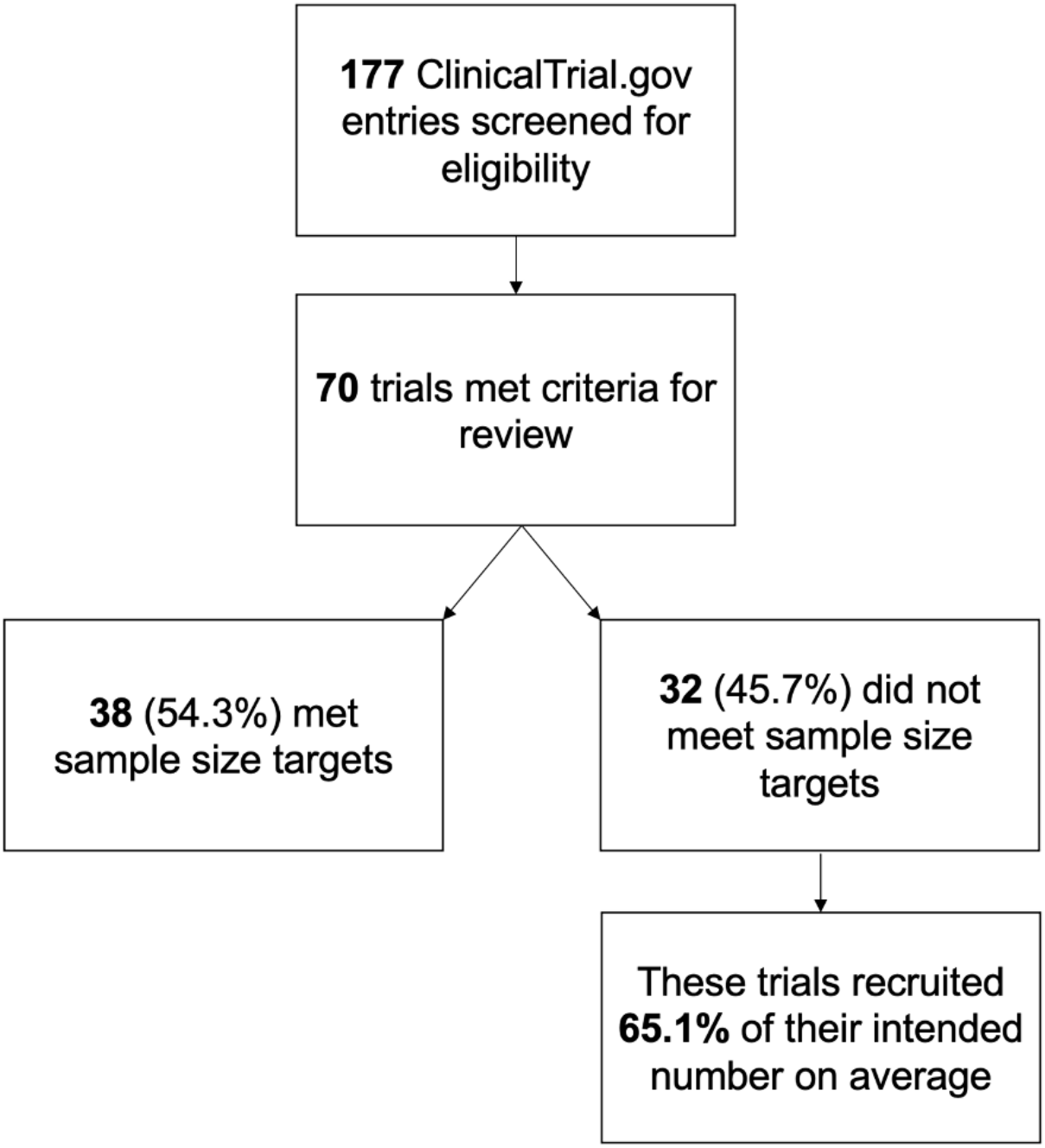

Why Might People Not Want Self-Help DMHIs, and What Can We Do About It?

Overview

Our review of clinical trials, in combination with previously reviewed research, suggests that self-help DMHIs may have a desirability issue. If self-help DMHIs are to be the solution to global difficulties disseminating mental health intervention to the vast numbers of people who need them, then it is critical that people actually want and use them. Below we outline two key potential problems that may be interfering with the uptake of self-help DMHIs, as well as potential solutions. Some of these solutions include implementation considerations that offer process support around the intervention itself and can strengthen uptake and retention. This list is not exhaustive but is intended to facilitate early anticipation of potential desirability issues in the self-help DMHI field and generate potential solutions so that the promise of these interventions may be realized.

Problem 1: People Are Going to Someone Else's Party

As previously discussed, the self-help DMHI marketplace is becoming increasingly saturated, which may limit the desirability and perceived credibility of individual self-help DMHIs. Most critically, it may limit the desirability of self-help DMHIs that are evidence-based. Many technology companies

XSL•FO
RenderX

that have produced self-help DMHI products have a wealth of experience in building attractive and appealing products, backed by experts in marketing, engineering, user experience, gamification, and web design. However, the extent to which mental health experts are involved in these products is unclear and, consequently, the extent to which many self-help DMHI offerings are evidence-based and likely to be efficacious is also mixed [29]. Unfortunately, consumers lack education regarding what interventions are or are not evidence-based. Indeed, Becker et al [30] found that only 2 out of 53 mental health care consumers had ever heard of the term "evidence-based practice," and only 20% of community members can accurately define evidence-based mental health care when asked [31]. Further, studies suggest that individuals have misconceptions about the characteristics of evidence-based practice, most commonly believing that such interventions are inflexible or not person-centered [30,32,33]. Based on this research, there is a clear lack of knowledge in public consciousness regarding evidence-based and empirically supported mental health care practices, which may lead individuals to take up self-help DMHIs based on other factors. Branding, marketing, and user experience (UX) may carry a disproportionate amount of weight in determining which services are taken up by consumers, but do not necessarily translate to the extent to which those services will be efficacious. Further, grant-funded clinical researchers best positioned to produce evidence-based tools may have a small fraction of the financial and human resources to invest in increasing the attractiveness of a product (eg, marketing and web design) compared with large technology companies. Consequently, evidence-based self-help DMHIs may get lost amidst an array of offerings, with several more seemingly attractive (but potentially not evidence-based) options "soaking up all of the oxygen." This vast competition for consumers in the self-help DMHI space may dilute the desirability of evidence-based interventions, many of which cannot compete in their capacity to draw in consumers. Moreover, individuals or people they know may try popular but not evidence-based self-help DMHIs with little benefit. Consequently, they may conclude that all self-help DMHIs are unlikely to be helpful, damaging the perceived credibility and desirability of evidence-based offerings.

Potential Solutions to Problem 1

Increasing the perceived credibility of self-help DMHIs, and consumers' capacity to discern between them, requires a shift in public education tactics related to mental health. In recent years, public mental health awareness campaigns have sought to reduce stigma related to mental illness by encouraging people to talk publicly and openly about their psychological struggles and seek help in the form of medication or psychotherapy if needed (eg, Bell Let's Talk, Jansport's #LightentheLoad, Instagram's #HereforYou, Maybelline New York—Brave Together). Undoubtedly, combating mental health stigma is critical in helping people gain access to the interventions they may need. However, broadly encouraging people to pursue psychiatric and psychological interventions may inadvertently mislead the public because there is a massive deficit in the availability of mental health interventions. For example, in the province that several of the authors live in, waitlists for publicly-funded psychotherapies were up to 2.5 years for youths and children in 2020 [34]. Although helping the public desire psychotherapy may improve the uptake of psychological interventions generally, there are not nearly enough mental health care providers to actually meet the demand that the destigmatizing campaigns seek to foment [5]. Therefore, a shift in public framing from solely seeking mental health support through psychiatric or individual psychotherapy interventions to include other supports across the intervention spectrum (eg, peer support, psychoeducation, and mental health literacy) is needed. Increasing awareness about the efficacy of self-help DMHIs as part of these public destigmatizing efforts is critical. Indeed, greater knowledge about self-help DMHIs is also associated with greater perceived helpfulness of them [35].

Second, awareness campaigns that encourage people to seek support from any mental health care provider or self-help DMHI are insufficient. Therefore, in addition to public efforts to encourage help-seeking generally, significant education regarding evidence-based mental health practices, empirically supported interventions, how efficacious self-help DMHIs work, and how consumers can determine the extent to which the self-help DMHI that they are considering is evidence-based are critical. Indeed, participants who receive more information about self-help DMHIs are more likely to engage in these programs in the future [36]. However, currently, participants report having little to no knowledge about them [36]. Moreover, when designing informational materials, co-creation with those who have lived experience increases the relevancy of those materials and can motivate consumers to seek out additional information about the subject in question [37].

Third, the first and second solutions mentioned earlier shift the burden of determining which interventions will be safe and efficacious to the consumer. It is also possible for self-help DMHI products to be regulated by governmental bodies. For example, work is being undertaken in the United Kingdom to evaluate factors that should be considered when regulating mental health apps [38], and at least some mental health apps in the United States have received review and approval as medical devices by the Food and Drug Administration [39]. More efforts on behalf of regulatory bodies are needed to protect the public from self-help DMHIs that may be harmful or unhelpful and identify ones that may be efficacious.

Fourth, clinical researchers seeking to construct self-help DMHIs derived from evidence-based practices or treatments may need support in constructing platforms that can compete with attractive, well-financed, but not empirically supported ones. This can therefore be an opportunity for interdisciplinary thinking and collaboration alongside professionals in other sectors such as marketing, UX, user interface design, and gamification. Those seeking grants to support the development of self-help DMHIs are advised to build salaries for such experts into their budgets or to collaborate with existing self-help DMHI developers to leverage and customize existing platforms to save costs. Moreover, funders may require education regarding why such personnel are essential to the successful building and delivery of self-help DMHIs.

XSL•FO
RenderX

Fifth, data suggests that there are some modifiable factors that can make a self-help DMHI more desirable to consumers. For example, consumers report a preference for online programs that offer brief, spaced-out content sessions instead of fewer, longer ones. Some studies suggest that they also tend to prefer self-help DMHIs that use text and images to deliver key concepts instead of video [16]. However, such findings may be moderated by the intended audience, as research with youths suggests that they prefer dynamic, interactive content to static information (eg, text) [40], and a study in a German sample generally suggested that the mode of delivery (eg, text, audio, video, or game) of a mental health intervention was not perceived to be particularly important relative to other factors [41]. Reviewing and conducting extensive desirability research about specific elements of self-help DMHIs with one's intended audience to identify what would be appealing, attractive, and helpful to them may help researchers design more engaging tools. Further, it can be helpful to learn from those with lived experience regarding what might act as a barrier to uptake to inform the design of the self-help DMHI and identify if there are additional elements that would support its access and use [42]. Grant budgets also need to allocate funds to compensate these individuals for their time.

Problem 2: Guests Stop by the Party, but They Don't Stick Around

In addition to working to initially attract people to self-help DMHIs, issues also arise in retaining them. Rates of retention in self-help DMHIs vary widely across studies. For instance, recent reviews found that retention levels ranged from 40% to 100% for mental health self-help DMHIs [43] and from 14% to 96% for wellness self-help DMHIs [44]. It is important to note that retention is measured differently across studies (eg, completing all vs half of the intervention modules) [44]. However, real-world engagement of self-help DMHI apps from user uptake data is reportedly lower than in controlled studies of similar interventions [45]. Indeed, the median retention rates for Android self-help DMHI apps after 15 and 30 days is between 3% and 4% [46], suggesting that participants may be less likely to engage in longer interventions, especially in self-paced contexts where progress is not consistently monitored as it is in research studies. Indeed, perhaps the benefits of self-help DMHIs in being accessible and easy to sign up for may also translate to issues with their capacity to retain participants, as they may require less motivation and commitment to pursue than standard face-to-face interventions. This lack of continued engagement in a self-help DMHI may hamper its efficacy and, in turn, damage the credibility and desirability of future self-help DMHIs.

Potential Solutions to Problem 2

To our knowledge, little attention has been paid to ways to retain, versus engage, users in self-help DMHIs. Indeed, self-help DMHIs that are built with the predominant goal of generating revenue may be incentivized to focus on getting users to buy or sign up for them, but not actually to use them. Although some studies have examined which individual characteristics predict self-help DMHI retention [47-49], few have directly tested methods to enhance it. Wojtowicz et al [49] showed that receiving phone-based coaching, compared to email-based coaching, predicted completion of modules in one online program. Similarly, self-help DMHIs incorporating human feedback and in-app mood monitoring have lower dropout rates than those that do not [50]. These studies suggest that some human involvement alongside self-help DMHIs, in addition to progress monitoring, may enhance retention. A review of the implementation literature could also offer a helpful starting point for factors that can influence retention. For example, research in telehealth interventions has demonstrated the need to ensure the fit or appropriateness of consumer-defined needs, consumer skills and capacity, and the structure of their daily lives [51]. The applicability of these factors and others noted as important for retention in the implementation literature, could be readily applied to and tested with self-help DMHIs.

In addition to these features, the content of the intervention itself may promote or reduce engagement. Geraghty et al [52] compared participants who received a self-help DMHI for body dissatisfaction containing therapeutic content alone (ie, cognitive behavioral techniques of self-monitoring and cognitive restructuring) to those who received therapeutic content within the format of a gratitude intervention. Participants enrolled in the gratitude intervention were two times more likely to complete the intervention compared to those who completed the intervention containing only therapeutic content. More research is needed to identify which forms of content may be optimal for enhancing engagement across self-help DMHIs and why.

In addition, it is well documented in the psychotherapy literature that individuals' motivation and readiness for change informs treatment outcomes [53] and the psychotherapy landscape offers a range of tools that can increase readiness to change (eg, motivational interviewing) [54]. If the ease of signing up for self-help DMHIs means that they are able to initially attract users who vary in their motivation for change, then self-help DMHIs present both unique problems and opportunities. While they may be more likely to "lose" users who exhibited low motivation for change, they may also be uniquely poised to initially engage these users in the first place. Designing self-help DMHIs based on the premise that users may not be motivated to engage in them, and deliberately introducing interventions early with the sole purpose of increasing motivation, may be useful in promoting retention. More testing on factors that can increase motivation to engage with self-help DMHIs is therefore needed.

Moreover, technical elements of self-help DMHIs may be important to promoting user engagement and retention. People have reported that interventions with gamification features and interactive content are more enjoyable, while standard psychoeducation and visually unappealing interfaces with technological glitches are unsurprisingly rated as less attractive [55]. Therefore, the few studies that exist suggest that there are several key humans, content, and technical features that can be incorporated into self-help DMHIs to enhance their capacity to engage and retain. However, many empirical questions remain regarding best practices in promoting continued engagement in self-help DMHIs, and realizing their promise requires research to this end.

XSL•FO
RenderX

Finally, the aforementioned solutions are predicated on the notion that people should be retained. Indeed, burgeoning literature has instead focused on the development and testing of interventions that are designed without a presumption of repeat engagement (ie, single-session interventions) [56-61]. Single-session interventions overcome access issues by limiting the amount of time a mental health care provider is required to spend on one client (ie, one session vs several) and can be delivered over the internet [59], expanding their access to geographically remote or otherwise difficult-to-reach groups. Both self-administered, digital, and self-help digital single-session interventions have been developed [59-64]. Indeed, in a meta-analysis, while therapist-delivered single-session interventions yielded a larger effect size on youth psychological outcomes than self-help ones, this difference was not statistically significant [56]. More single-session approaches could be applied to the self-help DMHI landscape to address retention issues. Moreover, online single-session interventions exemplify another strategy of expanding access to mental health care which may complement self-help DMHI efforts.

Conclusions

The mental health field faces a crisis in access to the psychotherapeutic interventions that have been developing within it for decades. While self-help DMHIs have the potential to be part of an effective response to this crisis, their promise is not fully realized because relatively little attention has been paid to whether and how consumers actually want them compared to the exorbitant amount of time, money, and labor that has been devoted to creating them. It is as if the self-help DMHI field has gone to great lengths and expense to throw a large party; venues have been booked, fancy cakes have been ordered, and entertainment has been scheduled, but nobody asked whether people can get to the venue, what food they want to or can eat, and what entertainment they like. Moreover, nobody collected RSVPs (Répondez s'il vous plaît). Is everyone going to show up?

This commentary is not intended to discourage the pursuit of self-help DMHI development and testing, but rather to encourage researchers to first establish a foundation of work required to make them successful in the long-term. This involves studying whether, when, and how self-help DMHIs are desirable to target populations, and working with experts in other content areas (eg, those with lived experience, marketing experts, UX experts, and implementation scientists and practitioners) to enhance their engagement, desirability, and retention. It also involves a deliberate shift in public education efforts to delineate evidence-based self-help DMHIs from other offerings that are currently saturating the marketplace but may not be helpful. Ultimately, with a more deliberate, thoughtful approach to self-help DMHI development and testing that carefully considers issues of desirability, engagement, and retention, we believe that self-help DMHIs may fully realize their promise as leading “the future” of mental health intervention.

Acknowledgments

This work was supported by grants from the Canadian Institute for Military and Veterans Health Research True Patriot Love Research Initiative (6023235-985815), the Canadian Department of National Defence (IDEaS_CP-0971), and the Canadian Institutes for Health Research (MS3-173098). These organizations were not directly involved in conceptualizing or writing this manuscript. It was also supported by the Atlas Institute for Veterans and Families (144555), and members from this organization have been involved in the development of this manuscript as coauthors.

Conflicts of Interest

CM receives royalties from Guilford Press related to the publication of the treatment manual from which Couple HOPES was adapted. CM and SF receive payment for providing training to clinicians in psychotherapies for posttraumatic stress disorder, borderline personality disorder, and related conditions. CM is the chief executive officer and founder of Nellie Health—a private mental health intervention company—and SF owns equity in this company.

References

1. Tögel C. Sigmund Freud's practice: visits and consultation, psychoanalyses, remuneration. Psychoanal Q 2009;78(4):1033-1058. [doi: 10.1002/j.2167-4086.2009.tb00426.x] [Medline: 19928437]
2. World mental health report: transforming mental health for all. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2022-06-16]
3. Lal S. E-mental health: promising advancements in policy, research, and practice. Healthc Manage Forum 2019;32(2):56-62 [FREE Full text] [doi: 10.1177/0840470418818583] [Medline: 30739487]
4. Borgueta AM, Purvis CK, Newman MG. Navigating the ethics of internet-guided self-help interventions. Clin Psychol 2018;25(2):e12235 [FREE Full text] [doi: 10.1111/cpsp.12235] [Medline: 33664555]
5. Kazdin AE, Blase SL. Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspect Psychol Sci 2011;6(1):21-37. [doi: 10.1177/1745691610393527] [Medline: 26162113]
6. Richards D. Self-help: empowering service users or aiding cash strapped mental health services? J Ment Health 2004;13(2):117-123. [doi: 10.1080/09638230410001669246]

XSL•FO
RenderX

7. Ellis LA, Collin P, Hurley PJ, Davenport TA, Burns JM, Hickie IB. Young men's attitudes and behaviour in relation to mental health and technology: implications for the development of online mental health services. BMC Psychiatry 2013;13(1):119 [FREE Full text] [doi: 10.1186/1471-244X-13-119] [Medline: 23601273]
8. Ryan ML, Shochet IM, Stallman H. Universal online interventions might engage psychologically distressed university students who are unlikely to seek formal help. Adv Ment Health 2010;9(1):73-83. [doi: 10.5172/jamh.9.1.73]
9. Dragioti E, Li H, Tsitsas G, Lee KH, Choi J, Kim J, et al. A large-scale meta-analytic atlas of mental health problems prevalence during the COVID-19 early pandemic. J Med Virol 2022;94(5):1935-1949 [FREE Full text] [doi: 10.1002/jmv.27549] [Medline: 34958144]
10. Horigian VE, Schmidt RD, Feaster DJ. Loneliness, mental health, and substance use among US young adults during COVID-19. J Psychoactive Drugs 2021;53(1):1-9. [doi: 10.1080/02791072.2020.1836435] [Medline: 33111650]
11. Krishnamoorthy Y, Nagarajan R, Saya GK, Menon V. Prevalence of psychological morbidities among general population, healthcare workers and COVID-19 patients amidst the COVID-19 pandemic: a systematic review and meta-analysis. Psychiatry Res 2020;293:113382 [FREE Full text] [doi: 10.1016/j.psychres.2020.113382] [Medline: 32829073]
12. Bierbooms JJPA, van Haaren M, IJsselsteijn WA, de Kort YAW, Feijt M, Bongers IMB. Integration of online treatment into the "New Normal" in mental health care in post-COVID-19 times: exploratory qualitative study. JMIR Form Res 2020;4(10):e21344 [FREE Full text] [doi: 10.2196/21344] [Medline: 33001835]
13. Unpublished raw data on the research portfolio of online reporting tools expenditures and results system. National Institutes of Health. URL: https://reporter.nih.gov/ [accessed 2022-08-26]
14. Clay RA. Mental health apps are gaining traction. American Psychological Association. 2021. URL: https://www.apa.org/monitor/2021/01/trends-mental-health-apps [accessed 2022-11-16]
15. Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/mental-health-apps-market-report [accessed 2022-11-16]
16. Batterham PJ, Calear AL. Preferences for internet-based mental health interventions in an adult online sample: findings from an online community survey. JMIR Ment Health 2017;4(2):e26 [FREE Full text] [doi: 10.2196/mental.7722] [Medline: 28666976]
17. Musiat P, Goldstone P, Tarrier N. Understanding the acceptability of e-mental health—attitudes and expectations towards computerised self-help treatments for mental health problems. BMC Psychiatry 2014;14:109 [FREE Full text] [doi: 10.1186/1471-244X-14-109] [Medline: 24725765]
18. Peynenburg VA, Mehta S, Hadjistavropoulos HD. Postsecondary student perceptions and preferences for the treatment of depression and anxiety: comparison of internet-delivered cognitive behaviour therapy to face-to-face cognitive behaviour therapy and medication. Can J Behav Sci 2020;52(3):220-230. [doi: 10.1037/cbs0000165]
19. Soucy JN, Hadjistavropoulos HD. Treatment acceptability and preferences for managing severe health anxiety: perceptions of internet-delivered cognitive behaviour therapy among primary care patients. J Behav Ther Exp Psychiatry 2017;57:14-24. [doi: 10.1016/j.jbtep.2017.02.002] [Medline: 28242411]
20. Knechtel LM, Erickson CA. Who’s logging on? Differing attitudes about online therapy. J Technol Hum Serv 2020;39(1):24-42. [doi: 10.1080/15228835.2020.1833810]
21. Wallin EEK, Mattsson S, Olsson EMG. The preference for internet-based psychological interventions by individuals without past or current use of mental health treatment delivered online: a survey study with mixed-methods analysis. JMIR Ment Health 2016;3(2):e25 [FREE Full text] [doi: 10.2196/mental.5324] [Medline: 27302200]
22. March S, Day J, Ritchie G, Rowe A, Gough J, Hall T, et al. Attitudes toward e-mental health services in a community sample of adults: online survey. J Med Internet Res 2018;20(2):e59 [FREE Full text] [doi: 10.2196/jmir.9109] [Medline: 29459357]
23. Apolinário-Hagen J, Fritsche L, Bierhals C, Salewski C. Improving attitudes toward e-mental health services in the general population via psychoeducational information material: a randomized controlled trial. Internet Interv 2018;12:141-149 [FREE Full text] [doi: 10.1016/j.invent.2017.12.002] [Medline: 30135778]
24. Apolinário-Hagen J, Kemper J, Stürmer C. Public acceptability of e-mental health treatment services for psychological problems: a scoping review. JMIR Ment Health 2017;4(2):e10 [FREE Full text] [doi: 10.2196/mental.6186] [Medline: 28373153]
25. Lannin DG, Vogel DL, Brenner RE, Abraham WT, Heath PJ. Does self-stigma reduce the probability of seeking mental health information? J Couns Psychol 2016;63(3):351-358. [doi: 10.1037/cou0000108] [Medline: 26323042]
26. Crisp DA, Griffiths KM. Participating in online mental health interventions: Who is most likely to sign up and why? Depress Res Treat 2014;2014:790457 [FREE Full text] [doi: 10.1155/2014/790457] [Medline: 24804089]
27. Doss BD, Knopp K, Roddy MK, Rothman K, Hatch SG, Rhoades GK. Online programs improve relationship functioning for distressed low-income couples: results from a nationwide randomized controlled trial. J Consult Clin Psychol 2020;88(4):283-294. [doi: 10.1037/ccp0000479] [Medline: 32134290]
28. Sommers-Spijkerman MPJ, Trompetter HR, Schreurs KMG, Bohlmeijer ET. Compassion-focused therapy as guided self-help for enhancing public mental health: a randomized controlled trial. J Consult Clin Psychol 2018;86(2):101-115. [doi: 10.1037/ccp0000268] [Medline: 29265836]

XSL•FO
RenderX

29. Schueller SM, Torous J. Scaling evidence-based treatments through digital mental health. Am Psychol 2020;75(8):1093-1104 [FREE Full text] [doi: 10.1037/amp0000654] [Medline: 33252947]
30. Becker SJ, Spirito A, Vanmali R. Perceptions of 'evidence-based practice' among the consumers of adolescent substance use treatment. Health Educ J 2016;75(3):358-369 [FREE Full text] [doi: 10.1177/0017896915581061] [Medline: 27087698]
31. Mora Ringle VA, Walsh LM, Maxwell CA, Smith AM, Grossman RA, Becker SJ, et al. Understanding of evidence-based mental health care and the perceived importance of scientific information in a sample of U.S. adults. J Clin Psychol 2020;76(1):161-175. [doi: 10.1002/jclp.22856] [Medline: 31491053]
32. Carman KL, Maurer M, Yegian JM, Dardess P, McGee J, Evers M, et al. Evidence that consumers are skeptical about evidence-based health care. Health Aff 2010;29(7):1400-1406. [doi: 10.1377/hlthaff.2009.0296] [Medline: 20522522]
33. Tanenbaum SJ. Perspectives on evidence-based practice from consumers in the US public mental health system. J Eval Clin Pract 2008;14(5):699-706. [doi: 10.1111/j.1365-2753.2008.01020.x] [Medline: 19018898]
34. Kids can't wait: 2020 report on wait lists and wait times for child and youth mental health care in Ontario. Children's Mental Health Ontario. 2020. URL: https://cmho.org/wp-content/uploads/CMHO-Report-WaitTimes-2020.pdf [accessed 2022-12-01]
35. Sweeney GM, Donovan CL, March S, Forbes Y. Logging into therapy: adolescent perceptions of online therapies for mental health problems. Internet Interv 2019;15:93-99 [FREE Full text] [doi: 10.1016/j.invent.2016.12.001] [Medline: 30792959]
36. Casey LM, Joy A, Clough BA. The impact of information on attitudes toward e-mental health services. Cyberpsychol Behav Soc Netw 2013;16(8):593-598. [doi: 10.1089/cyber.2012.0515] [Medline: 23679567]
37. Becker SJ, Helseth SA, Tavares TL, Squires DD, Clark MA, Zeithaml VA, et al. User-informed marketing versus standard description to drive demand for evidence-based therapy: a randomized controlled trial. Am Psychol 2020;75(8):1038-1051 [FREE Full text] [doi: 10.1037/amp0000635] [Medline: 33252943]
38. Medicines and Healthcare Products Regulatory Agency. Mental health funding of £1.8m welcomed by MHRA and nice to explore regulation of digital mental health tools. GOV.UK. 2022. URL: https://www.gov.uk/government/news/mental-health-funding-of-18m-welcomed-by-mhra-and-nice-to-explore-regulation-of-digital-mental-health-tools [accessed 2024-07-22]
39. Middleton J. First smartphone app to gain FDA approval for depression. American Academy of Family Physicians. 2024. URL: https://tinyurl.com/4a57hfxr [accessed 2024-07-22]
40. Michel T, Tachtler F, Slovak P, Fitzpatrick G. A review of youth mental health promotion apps towards their fit with youth media preferences. EAI Endor Trans Perv Health Technol 2019;5(17):161419 [FREE Full text] [doi: 10.4108/eai.13-7-2018.161419]
41. Phillips EA, Himmler SF, Schreyögg J. Preferences for e-mental health interventions in Germany: a discrete choice experiment. Value Health 2021;24(3):421-430 [FREE Full text] [doi: 10.1016/j.jval.2020.09.018] [Medline: 33641777]
42. Graham AK, Lattie EG, Powell BJ, Lyon AR, Smith JD, Schueller SM, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020;75(8):1080-1092 [FREE Full text] [doi: 10.1037/amp0000686] [Medline: 33252946]
43. Babbage CM, Jackson GM, Davies EB, Nixon E. Self-help digital interventions targeted at improving psychological well-being in young people with perceived or clinically diagnosed reduced well-being: systematic review. JMIR Ment Health 2022;9(8):e25716 [FREE Full text] [doi: 10.2196/25716] [Medline: 36018675]
44. Lipschitz JM, Van Boxtel R, Torous J, Firth J, Lebovitz JG, Burdick KE, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022;24(10):e39204 [FREE Full text] [doi: 10.2196/39204] [Medline: 36240001]
45. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
46. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
47. Carpenter KM, Stoner SA, Mundt JM, Stoelb B. An online self-help CBT intervention for chronic lower back pain. Clin J Pain 2012;28(1):14-22 [FREE Full text] [doi: 10.1097/AJP.0b013e31822363db] [Medline: 21681084]
48. Gill S, Contreras O, Muñoz RF, Leykin Y. Participant retention in an automated online monthly depression rescreening program: patterns and predictors. Internet Interv 2014;1(1):20-25 [FREE Full text] [doi: 10.1016/j.invent.2014.02.003] [Medline: 25045623]
49. Wojtowicz M, Day V, McGrath PJ. Predictors of participant retention in a guided online self-help program for university students: prospective cohort study. J Med Internet Res 2013;15(5):e96 [FREE Full text] [doi: 10.2196/jmir.2323] [Medline: 23697614]
50. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
51. Vassilev I, Rowsell A, Pope C, Kennedy A, O'Cathain A, Salisbury C, et al. Assessing the implementability of telehealth interventions for self-management support: a realist review. Implement Sci 2015;10:59 [FREE Full text] [doi: 10.1186/s13012-015-0238-9] [Medline: 25906822]

XSL•FO
RenderX

52. Geraghty AWA, Wood AM, Hyland ME. Attrition from self-directed interventions: investigating the relationship between psychological predictors, intervention content and dropout from a body dissatisfaction intervention. Soc Sci Med 2010;71(1):30-37. [doi: 10.1016/j.socscimed.2010.03.007] [Medline: 20400220]
53. Krebs P, Norcross JC, Nicholson JM, Prochaska JO. Stages of change and psychotherapy outcomes: a review and meta-analysis. J Clin Psychol 2018;74(11):1964-1979. [doi: 10.1002/jclp.22683] [Medline: 30335193]
54. Lundahl BW, Kunz C, Brownell C, Tollefson D, Burke BL. A meta-analysis of motivational interviewing: twenty-five years of empirical studies. Res Soc Work Pract 2010;20(2):137-160 [FREE Full text] [doi: 10.1177/1049731509347850]
55. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't Work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
56. Schleider JL, Weisz JR. Little treatments, promising effects? Meta-Analysis of single-session interventions for youth psychiatric problems. J Am Acad Child Adolesc Psychiatry 2017;56(2):107-115. [doi: 10.1016/j.jaac.2016.11.007] [Medline: 28117056]
57. Loades ME, Schleider JL. Technology matters: online, self-help single session interventions could expand current provision, improving early access to help for young people with depression symptoms, including minority groups. Child Adolesc Ment Health 2023;28(4):559-561. [doi: 10.1111/camh.12659] [Medline: 37415054]
58. Schleider JL, Smith AC, Ahuvia I. Realizing the untapped promise of single-session interventions for eating disorders. Int J Eat Disord 2023;56(5):853-863 [FREE Full text] [doi: 10.1002/eat.23920] [Medline: 36815724]
59. Schleider JL, Mullarkey MC, Fox KR, Dobias ML, Shroff A, Hart EA, et al. A randomized trial of online single-session interventions for adolescent depression during COVID-19. Nat Hum Behav 2022;6(2):258-268 [FREE Full text] [doi: 10.1038/s41562-021-01235-0] [Medline: 34887544]
60. Walton MA, Chermack ST, Shope JT, Bingham CR, Zimmerman MA, Blow FC, et al. Effects of a brief intervention for reducing violence and alcohol misuse among adolescents: a randomized controlled trial. JAMA 2010;304(5):527-535 [FREE Full text] [doi: 10.1001/jama.2010.1066] [Medline: 20682932]
61. Miu AS, Yeager DS. Preventing symptoms of depression by teaching adolescents that people can change. Clin Psychol Sci 2014;3(5):726-743 [FREE Full text] [doi: 10.1177/2167702614548317]
62. Messer M, Fuller-Tyszkiewicz M, Liu C, Anderson C, Linardon J. A randomized controlled trial of an online single session intervention for body image in individuals with recurrent binge eating. Int J Eat Disord 2024;57(8):1668-1679 [FREE Full text] [doi: 10.1002/eat.24213] [Medline: 38619359]
63. Dobias ML, Schleider JL, Jans L, Fox KR. An online, single-session intervention for adolescent self-injurious thoughts and behaviors: results from a randomized trial. Behav Res Ther 2021;147:103983. [doi: 10.1016/j.brat.2021.103983] [Medline: 34688102]
64. Mullarkey M, Dobias M, Sung J, Ahuvia I, Shumake J, Beevers C, et al. Web-based single session intervention for perceived control over anxiety during COVID-19: randomized controlled trial. JMIR Ment Health 2022;9(4):e33473 [FREE Full text] [doi: 10.2196/33473] [Medline: 35230962]

Abbreviations

DMHI: digital mental health intervention
RSVP: Répondez s'il vous plaît
UX: user experience

Edited by G Eysenbach; submitted 08.03.24; peer-reviewed by M Loades, J Freyer-Adam; comments to author 01.05.24; revised version received 22.07.24; accepted 10.08.24; published 19.09.24.

Please cite as:
Fitzpatrick S, Crenshaw AO, Donkin V, Collins A, Xiang A, Earle EA, Goenka K, Varma S, Bushe J, McFadden T, Librado A, Monson C
We Have Spent Time, Money, and Effort Making Self-Help Digital Mental Health Interventions: Is Anyone Going to Come to the Party?
J Med Internet Res 2024;26:e58198
URL: https://www.jmir.org/2024/1/e58198
doi:10.2196/58198
PMID:

©Skye Fitzpatrick, Alexander O Crenshaw, Victoria Donkin, Alexis Collins, Angela Xiang, Elizabeth A Earle, Kamya Goenka, Sonya Varma, Julianne Bushe, Tara McFadden, Andrea Librado, Candice Monson. Originally published in the Journal of Medical

XSL•FO
RenderX

Internet Research (https://www.jmir.org), 19.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development Trends and Prospects of Technology-Based Solutions for Health Challenges in Aging Over the Past 25 Years: Bibliometric Analysis

Lu Liu[1*], MD, PhD; Xiu-Ling Wang[2*], BD, RN; Nuo Cheng[3*], MM, RN; Fu-Min Yu[4*], MD, PhD; Hui-Jun Li[5,6], BD; Yang Mu[7], BD; Yonghui Yuan[8], MD, PhD; Jia-Xin Dong[9], BD; Yu-Dan Wu[9], BD; Da-Xin Gong[10,11], MD, PhD; Shuang Wang[12], MD, PhD; Guang-Wei Zhang[10,11,12], MD, PhD

[1]Department of Cardiac Surgery, The First Hospital of China Medical University, Shenyang, China

[2]Department of Cardiology, The First Hospital of China Medical University, Shenyang, China

[3]Department of Gynecology, The First Hospital of China Medical University, Shenyang, China

[4]Department of Cardiovascular Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou, China

[5]Shenyang Medical & Film Science and Technology Co. Ltd., Shenyang, China

[6]Enduring Medicine Smart Innovation Research Institute, Shenyang, China

[7]Goodwill Information Technology Co., Ltd., Beijing, China

[8]Cancer Hospital of China Medical University, Liaoning Cancer Hospital & Institute, Clinical Research Center for Malignant Tumor of Liaoning Province, Shenyang, China

[9]Department of Nursing Faculty, China Medical University, Shenyang, China

[10]Smart Hospital Management Department, The First Hospital of China Medical University, Shenyang, China

[11]Internet Hospital Branch, Chinese Research Hospital Association, Beijing, China

[12]Department of General Practice, The First Hospital of China Medical University, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Guang-Wei Zhang, MD, PhD
Smart Hospital Management Department
The First Hospital of China Medical University
155 Nanjingbei St
Shenyang, 110001
China
Phone: 86 024 83283350
Email: gwzhang@cmu.edu.cn

Abstract

Background: As the global population ages, we witness a broad scientific and technological revolution tailored to meet the health challenges of older adults. Over the past 25 years, technological innovations, ranging from advanced medical devices to user-friendly mobile apps, are transforming the way we address these challenges, offering new avenues to enhance the quality of life and well-being of the aging demographic.

Objective: This study aimed to systematically review the development trends in technology for managing and caring for the health of older adults over the past 25 years and to project future development prospects.

Methods: We conducted a comprehensive bibliometric analysis of literatures related to technology-based solutions for health challenges in aging, published up to March 18, 2024. The search was performed using the Web of Science Core Collection, covering a span from 1999 to 2024. Our search strategy was designed to capture a broad spectrum of terms associated with aging, health challenges specific to older adults, and technological interventions.

Results: A total of 1133 publications were found in the Web of Science Core Collection. The publication trend over these 25 years showed a gradual but fluctuating increase. The United States was the most productive country and participated in international collaboration most frequently. The predominant keywords identified through this analysis included “dementia,” “telemedicine,” “older-adults,” “telehealth,” and “care.” The keywords with citation bursts included “telemedicine” and “digital health.”

XSL•FO
RenderX

Conclusions: The scientific and technological revolution has significantly improved older adult health management, particularly in chronic disease monitoring, mobility, and social connectivity. The momentum for innovation continues to build, with future research likely to focus on predictive analytics and personalized health care solutions, further enhancing older adults' independence and quality of life.

(J Med Internet Res 2024;26:e63367) doi:10.2196/63367

KEYWORDS

bibliometrics; CiteSpace; VOSviewer; visualization; aging health; technological innovations; tech-based; technology-based; technology; health challenges; challenges; trends; older adults; older adult; ageing; aging; elder; elderly; older person; older people; gerontology; geriatric; geriatrics; remote; remote monitoring; monitoring; surveillance; artificial intelligence; AI; AI-driven; innovation; innovations; health management; telemedicine; remote care

Introduction

As the global population aging trend continues to intensify, the health issues of older adults have increasingly become the focus of global attention [1]. As a pivotal force in the scientific and technological revolution addressing global aging, a diverse array of technologies is reshaping how we meet the health challenges of older adults [2]. By harnessing technological innovations, ranging from cutting-edge medical devices to user-friendly mobile apps, novel avenues to enhance the quality of life and well-being of the aging demographic are unveiled [3]. Among these advancements are remote health monitoring systems capable of real-time tracking of vital signs, assistive technologies fostering independence in daily activities, and the integration of artificial intelligence (AI) and data analytics, revolutionizing health care delivery and enabling tailored approaches to personalized medicine for older adults [4-6]. Moreover, the indispensability of telemedicine and virtual health care solutions has been underscored, particularly in the aftermath of the COVID-19 pandemic, highlighting their crucial role in delivering remote care and consultations [7,8]. These technologies empower older adults to access health care services from the convenience of their homes, curtailing the necessity for physical clinic visits and mitigating exposure to infectious diseases [9]. When an accident occurs, older adults can seek emergency help remotely, and nurses can observe and understand the situation through videos, providing targeted medical assistance [10]. In essence, technology stands as a transformative catalyst in confronting the health complexities associated with aging, offering innovative remedies to enhance quality of life and foster autonomy within the older adult populace [11].

Recent reports have begun to focus on the application of science and technology in the management of older adult health. Chan et al [12] discussed the use of home health monitoring systems in older adult health management through a scoping review. Hong et al [13] provided an overview of the development of smart home technologies for older adults through bibliometric and scientometric analyses. Yenişehir et al [14] analyzed AI technologies for preventing falls among older adults. While these reports demonstrated the potential of technology applications across various aspects, they often focus on specific technologies or address specific health areas for older adults. However, there remains a lack of research exploring a more comprehensive technological revolution across the broader field of older adult health.

The bibliometric analysis methodology equips us with a robust tool set for exploration within this rapidly evolving field. By quantitatively assessing published literature, this method effectively delineates the research hot spots and elucidates the knowledge structures that define scientific activities. It enables us to trace the developmental trajectory and pivotal transformations of digital health and the technological revolution in addressing the health challenges of aging. Notably, over the past 25 years, the integration of technology in older adult health care has seen extensive growth and diversification, with numerous innovative solutions being introduced [15]. This paper aims to comprehensively review the development trends of technology in aging health management and care during this period and look forward to future development prospects. By systematically analyzing the evolution paths of technological solutions, existing achievements, and challenges faced, we will provide a comprehensive understanding of technology applications in the health of older adults. We will also explore possible future development directions to provide important reference and inspiration for promoting the further development of the field of health technology for older adults.

Methods

Data Collection

Web of Science stands as the foremost research platform encompassing hard sciences, social sciences, arts, and humanities information [16,17]. Additionally, it serves as the authoritative global citation database, collaborating with the world's most reputable publishers. We searched the Web of Science Core Collection and selected the Science Citation Index Expanded and Social Sciences Citation Index databases. The search strategy was meticulously designed to encompass a comprehensive array of terms associated with aging, health challenges specific to older adults, and technological interventions that address these issues. The search formula is shown in Textbox 1.

The search period ranged from January 1, 1999, to March 18, 2024; this period was chosen to capture the developments in the field over the last 25 years. The inclusion criteria were strictly articles published in English, focusing on original research and reviews, to ensure the quality and relevance of the data extracted.

XSL•FO
RenderX

Textbox 1. Search formula.

(TS=("aging" OR "ageing" OR "elderly" OR "seniors" OR "older adults" OR "gerontology") AND TS=("health problems" OR "health issues" OR "healthcare challenges" OR "geriatric health" OR "chronic diseases" OR "mobility issues" OR "cognitive declines" OR "mental health issues" OR "physical health problems" OR "dementia" OR "Alzheimer's" OR "osteoporosis" OR "arthritis" OR "cardiovascular diseases" OR "diabetes" OR "falls" OR "hearing loss" OR "vision impairment" OR "loneliness" OR "social isolation" OR "well-being" OR "quality of life") AND TS=("technological solutions" OR "technology responses" OR "digital health" OR "health technologies" OR "e-health" OR "m-health" OR "telehealth" OR "telemedicine" OR "wearable technologies" OR "assistive technologies" OR "smart healthcare solutions" OR "AI in healthcare" OR "information technology in healthcare" OR "robotic assistance" OR "virtual reality in healthcare" OR "augmented reality in healthcare" OR "IoT in healthcare" OR "internet of things in healthcare" OR "machine learning in healthcare" OR "data analytics in healthcare" OR "blockchain in healthcare" OR "genomic medicine" OR "personalized medicine" OR "remote patient monitoring" OR "predictive analytics" OR "health informatics" OR "biotechnology" OR "nanotechnology")).

Data Analysis

Microsoft Office Excel 2019, VOSviewer (v.1.6.20), and CiteSpace (v.6.3.R1 Advanced) were used to analyze all 1172 files. VOSviewer, a bibliometric software, was developed in 2009 by van Eck and Waltman [18] at Leiden University's Center for Science and Technology Studies in the Netherlands. This Java-based software is free to use and is highly proficient in visualizing and analyzing large-scale data [18]. CiteSpace, created by Chen [19], is a bibliometric and visual analysis tool used for delving into collaborations, priorities, internal structures, focal areas, and potential trends within distinct domains.

Results

Annual Scientific Production

Our search strategy uncovered 1172 publications from 1999 to 2024, with 1133 being research or review articles, including 827 (73%) research articles and 306 (27%) review articles. The publication trend over the last 25 years showed a gradual but fluctuating increase, starting with single-digit publications in 1999 to a peak of 195 publications in 2022. This upward trajectory experienced a slight decline to 191 publications in 2023. The detailed evolution of these publication numbers, capturing both the growth and the fluctuations, is comprehensively documented in Figure 1.

Figure 1. Annual scientific production in technology-based solutions for health challenges in aging.

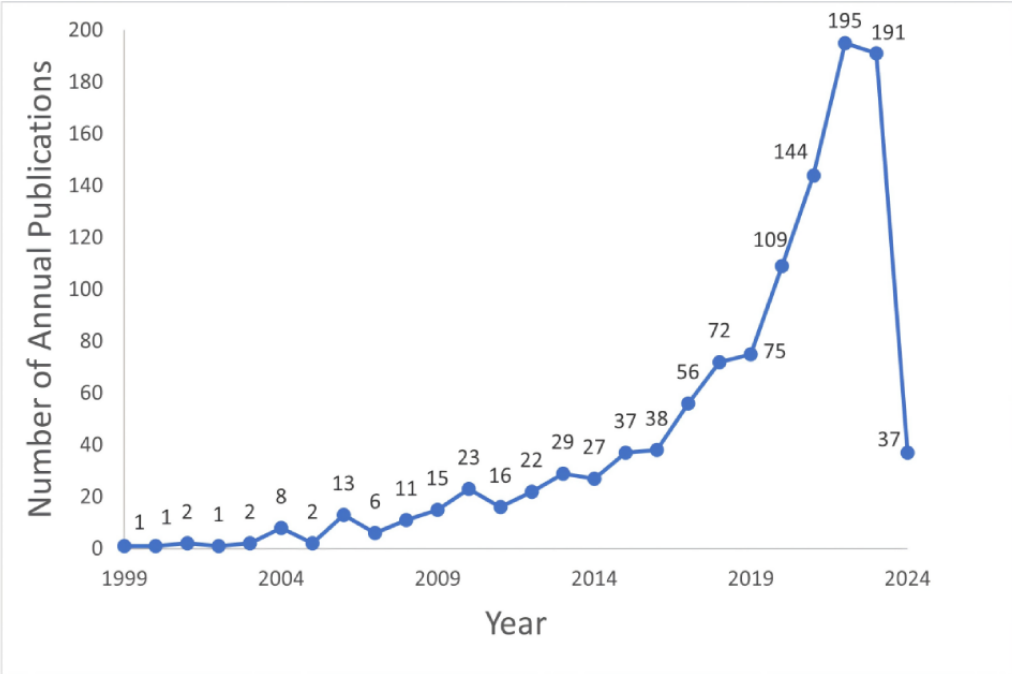

National and Institutional Contributions to Scientific Research and Collaborative Endeavors

We describe the scientific contributions of each country to research on technological responses to health issues in aging (Figure 2 and Table 1). The top 10 countries most interested in technological solutions to aging health issues, based on the count of publications and their centrality in the research network, are the United States (404 publications, centrality of 0.20), Italy (112 publications, centrality of 0.13), England (102 publications, centrality of 0.13), Canada (91 publications, centrality of 0.06), Australia (81 publications, centrality of 0.11), France (81 publications, centrality of 0.07), the Netherlands (74 publications, centrality of 0.06), Spain (74 publications, centrality of 0.07), People's Republic of China (72 publications, centrality of 0.05), and Germany (70 publications, centrality of 0.03). The United States, Italy, India, Japan, and Australia have a prominent centrality in the network, a distinction visually emphasized by pink circles in Figure 2. These circles not only highlight the critical roles these countries play but also underscore their extensive interconnections with other nations. This analysis highlights the leading roles these countries play

XSL•FO
RenderX

in advancing research focused on addressing health challenges associated with aging through technological innovations.

Figure 2. (A) National and (B) institutional contributions to scientific research and collaborative endeavors.

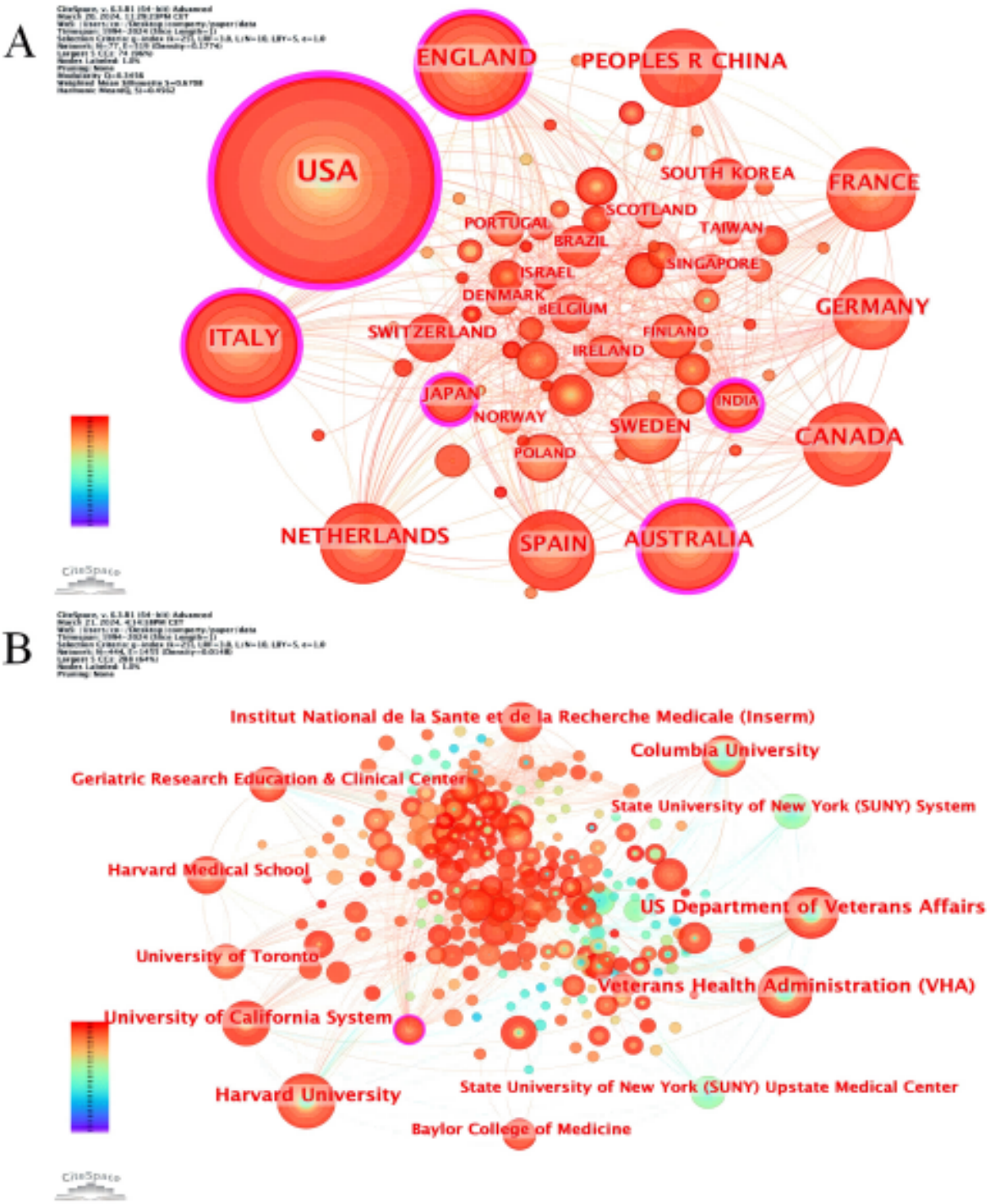

XSL•FO
RenderX

Table 1. National and institutional contributions to scientific research and collaborative endeavors.

Country or institution	Rank	Publications, n	Centrality
Country			
United States	1	404	0.20
Italy	2	112	0.13
England	3	102	0.13
Canada	4	91	0.06
Australia	5	81	0.11
France	6	81	0.07
The Netherlands	7	74	0.06
Spain	8	74	0.07
People's Republic of China	9	72	0.05
Germany	10	70	0.03
Institution			
Harvard University	1	44	0.08
US Department of Veterans Affairs	2	41	0.04
Veterans Health Administration (VHA)	3	41	0.04
University of California System	4	33	0.09
Columbia University	5	31	0.06
Institut National de la Sante et de la Recherche Medicale (INSERM)	6	25	0.04
Harvard Medical School	7	24	0.01
Geriatric Research Education & Clinical Center	8	23	0.09
University of Toronto	9	22	0.02
State University of New York (SUNY) System	10	20	0.01

We also present an analysis of the scientific contributions of various institutions to research on technological responses to health issues in aging (Figure 2 and Table 1). Among the top institutions demonstrating notable interest in technological solutions for aging-related health issues, Harvard University stands out with 44 publications and a centrality of 0.08. Following closely are the US Department of Veterans Affairs and Veterans Health Administration, each with 41 publications and a centrality of 0.04. The University of California System has contributed 33 publications with a centrality of 0.09, while Columbia University follows with 31 publications and a centrality of 0.06. Notably, *Institut National de la Sante et de la Recherche Medicale* has produced 25 publications, maintaining a centrality of 0.04. Harvard Medical School and the Geriatric Research Education & Clinical Center have 24 and 23 publications, respectively, with centrality values of 0.01 and 0.09. Additionally, the University of Toronto has contributed 22 publications with a centrality of 0.02, whereas the State University of New York System has produced 20 publications with a centrality of 0.01.

Publications Analysis Based on Top Cited Articles

In the retrieved collection, a total of 1133 papers were cited 52,704 times. The top 10 most cited papers, detailed in Table 2, showcase significant contributions to the field. The most cited article, "Lack of exercise is a major cause of chronic diseases" [20], published in *Comprehensive Physiology* in 2012, received 1405 citations. This highlights the critical impact of physical inactivity on health. Among the top 20 most cited papers, journals such as *Comprehensive Physiology*, *Computer Methods and Programs in Biomedicine*, and *IEEE Transactions on Biomedical Engineering* have each made notable contributions, reflecting the interdisciplinary nature of research impacting health and technology. The diversity in topics among these highly cited papers reflects the evolving landscape of health-related research, addressing contemporary challenges such as the COVID-19 pandemic and the aging population.

XSL•FO
RenderX

Table 2. The top 10 most cited papers.

Title	Citations, n	Journal	Year
Lack of exercise is a major cause of chronic diseases [20]	1405	*Comprehensive Physiology*	2012
A review of smart homes- present state and future challenges [21]	565	*Computer Methods and Programs in Biomedicine*	2008
Unobtrusive sensing and wearable devices for health informatics [22]	502	*IEEE Transactions on Biomedical Engineering*	2014
Mobile-health: a review of current state in 2015 [23]	489	*Journal of Biomedical Informatics*	2015
Impact of social isolation due to COVID-19 on health in older people: mental and physical effects and recommendations [24]	446	*Journal of Nutrition Health & Aging*	2020
The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use [25]	426	*Journal of Medical Internet Research*	2013
Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the Better Effectiveness After Transition -- Heart Failure (BEAT-HF) randomized clinical trial [26]	412	*JAMA Internal Medicine*	2016
Age-related diseases and clinical and public health implications for the 85 years old and over population [27]	399	*Frontiers in Public Health*	2017
Older adults' attitudes towards and perceptions of 'smart home' technologies: a pilot study [28]	396	*Medical Informatics and the Internet in Medicine*	2004
Smart homes and home health monitoring technologies for older adults: a systematic review [29]	336	*International Journal of Medical Informatics*	2016

Publications Analysis Based on Authors

In the landscape of this research field, a comprehensive analysis reveals that 1173 papers have been published with contributions from 6086 authors, detailed in Table 3. Among these contributors, Ruth S Weinstock and Steven Shea stand out with 16 publications each, followed by Joseph P Eimicke with 9 publications. These figures highlight not only the prolific nature of certain researchers but also the depth of collaboration and interdisciplinary engagement within the community. The citation metrics further illuminate the impact of these contributions, with Ruth S Weinstock's works receiving 613 citations, indicating a significant influence within the academic sphere.

Delving deeper into the data, other notable contributors such as Roberto Bernabei, with 6 publications garnering 394 citations, demonstrate the high quality and relevance of research in this area. This detailed examination of publications and citations reveals a vibrant network of scholarly communication, characterized by a rich tapestry of collaboration and intellectual exchange. The distribution of citations, with individuals like Roberto Bernabei achieving an average of 65.67 citations per document, underscores the substantial reach and engagement of the research outputs in this field, which is documented in detail in Table 3.

Table 3. Publications analysis based on authors.

Author	Documents, n	Citations, n	Citations per document
Bijan Najafi	16	266	16.63
Steven Shea	16	603	37.69
Ruth S Weinstock	16	613	38.31
Walter Palmas	10	423	42.30
Joseph P Eimicke	9	474	52.67
Rebeca Izquierdo	9	408	45.33
Jeanne A Teresi	8	405	50.63
Roberto Bernabei	6	394	65.67
Philip C Morin	6	159	26.50
Anne Tiedemann	6	19	3.17
Paula M Trief	6	292	48.67

XSL•FO
RenderX

Keyword Analysis

In this comprehensive analysis, keywords were meticulously selected from the titles and content of research papers, establishing a set of standardized terms that succinctly capture the core themes of the papers' subject matter. Through the application of VOSviewer, a total of 4690 keywords were identified. Notably, 188 of these keywords appeared more than 10 times, and 102 keywords were mentioned over 20 times. A thorough visual analysis was performed on the top 100 keywords by frequency, offering insights into the most prevalent topics of discussion and research. The analysis was further refined to keywords appearing more than 5 times in Figure 3 to improve clarity and accessibility. The visual mapping of these keywords illustrates the connections between them, with the thickness of the lines between nodes indicating the strength of their associations. This visualization underscores the frequency with which certain keywords co-occur within the same literature, revealing patterns of thematic relevance and interdisciplinary connections.

The predominant keywords identified through this analysis included "dementia," "telemedicine," "older-adults," "telehealth," "care," "people," "health," "older adults," "technology," "covid-19," "alzheimer's-disease," "aging," "digital health," "intervention," and "risk." These terms not only highlight the core subjects of the analyzed literature but also reflect significant trends and areas of interest within the research community. The emphasis on terms such as "dementia," "telemedicine," and "covid-19" points to a strong focus on health care challenges, technological advancements in medical care, and the impact of global health crises. Similarly, the frequent mention of "older adults" and "aging" underscores the growing attention to geriatric care and the health issues associated with aging populations.

Keyword clustering reveals the underlying structure of knowledge within a particular field of research, helping to categorize and define its boundaries. The clusters obtained after analysis mainly included "social isolation," "mild cognitive impairment," "assistive technology," "health," "personalized medicine," "chronic disease," "home health monitoring," "ambient assisted living," "physical activity," and "e-health program."

Finally, we statistically analyzed the strongest citation bursts of the top 10 keywords (Figure 3). The blue line represents the time interval, and the red line represents the time period when the keyword outbreak was found [30]. The first keyword with citation bursts initially surfaced in 2000 (telemedicine) and remained the most influential keyword until 2016. Conversely, the most recent keyword with citation bursts emerged in 2021 (digital health) and maintain its significance until 2022.

Figure 3. Keyword analysis. (A) Keywords appearing more than 5 times as shown by overlay visualization. (B) Keywords appearing more than 5 times as shown by density visualization. (C) Keyword cluster analysis. (D) The strongest citation bursts of the top 10 keywords.

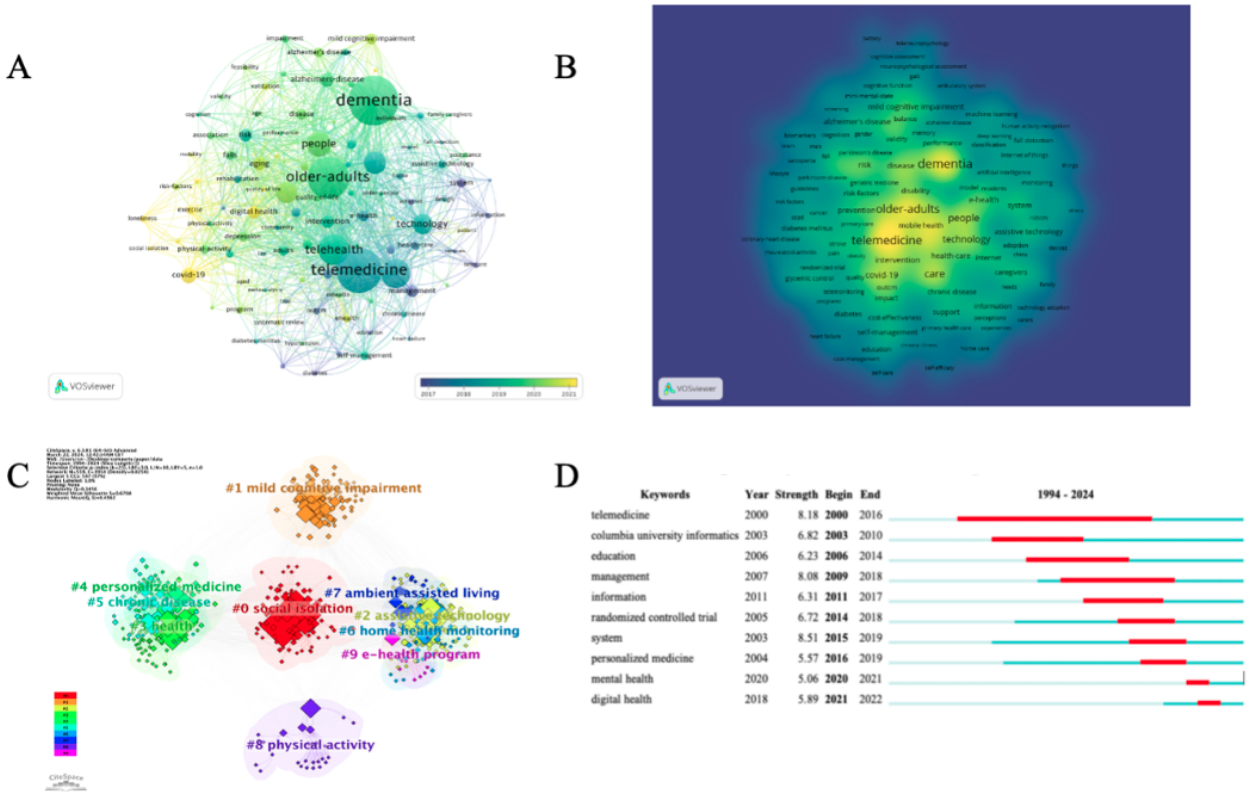

XSL•FO
RenderX

Discussion

Principal Findings

In this research, we used bibliometric analysis techniques for a comprehensive examination of literatures related to the topic of smart health care published prior to March 18, 2024. Our analysis shows that the field of technology-based solutions for health challenges in aging is gradually becoming the focus of widespread attention. In the past 5 years, there has been a sharp increasing trend. Notably, the prevalence of research-oriented literatures significantly outweighs that of reviews. This illustrates the complexity of the field that needs to be addressed, as the research literature often provides original research results and new insights.

Among the top 10 publishing countries, comprising 2 from the Americas, 6 from Europe, 1 from Asia, and 1 from Oceania, their collective output constitutes 99% of the total articles. These countries not only boast high gross domestic products but also house diverse aging populations, providing a rich backdrop for research into older adult health technology. For example, the United States and Italy not only possess robust economies but are also confronted with profound aging-related challenges. This compels them to prioritize and allocate significant funding to research initiatives aimed at mitigating these issues. Similarly, England, Canada, and Australia leverage their economic strengths with strategic investments in health technology, effectively addressing the needs of their expanding older adult populations. Furthermore, the central role these countries play in international research collaborations underlines their influence in shaping global research agendas. Their economic capacity enables higher research capabilities and outputs, while the extensive aging of their populations demands innovative solutions to increasingly complex health challenges. On the other hand, emerging economies such as China, despite their lower overall research outputs, are experiencing rapid growth in this area. This is propelled by urgent demographic shifts and an increasing governmental focus on health care innovation [31]. This diverse international landscape showcases the pivotal role of economic and demographic factors in driving the direction and intensity of research into aging and health technology.

Keywords serve as indicators of the central themes and primary content within a paper, thus offering a concise representation of the research's focal points. Our use of cluster analysis based on these keywords has enabled the identification of 10 distinct categories. Furthermore, by analyzing the occurrence of explosive words, we gain deeper insights into the intricate nuances and broad spectrum of the older adult health technology field. The current research hot spots focus on older adults with cognitive impairment, social isolation, and chronic diseases. Interventions include assistive technology, personalized medical solutions and eHealth projects, implementing remote consultation, education, monitoring, triage, care, and health management for older adults.

Social isolation can lead to an increase in mental and physical health problems, which is an important health challenge for older adults, especially those with cognitive conditions, so taking steps to reduce social isolation is critical to the health of older adults [32,33]. Assistive technology plays an important role in this regard. For example, ambient assisted living technology can provide a convenient and safe home environment and reduce the burden of life for older adults [34]. Virtual reality technology and computer-aided technology can detect very subtle cognitive impairment, help diagnose earlier, and prevent or delay the occurrence of dementia [35]. In addition, personalized medical solutions and eHealth projects through computer, iPad, or smartphone applications can also provide customized medical services to older adults to meet their individual health needs, thereby improving their quality of life [36-39].

On the other hand, older adults often face challenges with chronic conditions that require ongoing health monitoring and management [40]. Home health monitoring and telemedicine solutions provide older adults with convenient health management methods and help timely intervention and treatment of chronic diseases [41]. Additionally, by encouraging older adults to participate in physical activity and providing remote guidance and monitoring, their physical and mental health can be improved, thereby reducing the risk of chronic disease [42,43]. To sum up, by combining various means such as assistive technology, personalized medicine, health monitoring, and education, the health level of older adults can be comprehensively improved and the process of healthy aging can be promoted.

The analysis results show that the focus of the research shifted from "telemedicine" to "digital health." Digital health emphasizes more on the health and well-being of individuals and populations rather the diseases and patients [44]; has a broad scope; and includes the use of wearable devices, mobile health, telehealth, health information technology, and telemedicine [45]. This paradigm shift reflects an evolving health care approach where technology extends beyond remote clinical interactions to a comprehensive system that enhances personalized health management. Digital health integrates a variety of innovative technologies that facilitate not only remote monitoring but also proactive health management strategies. These tools empower patients through real-time data, engaging them in their own health management and improving the responsiveness of health care systems to their needs.

Indeed, the technological revolution has played a significant role in addressing key areas such as chronic disease management, mobility support, social connectivity, and cognitive enhancement. Chronic disease management has seen innovative applications of AI and machine learning to predict and manage episodes in conditions like diabetes and heart disease, reducing hospital visits and improving quality of life [46,47]. Mobility impairments, often a barrier to independent living, are being tackled with advanced robotics and smart wearables that assist movement and monitor falls [48,49]. Social isolation, exacerbated by physical distancing during the COVID-19 pandemic, is being mitigated through virtual reality platforms that allow social interactions and community building without physical presence [50]. Furthermore, cognitive decline—one of the most challenging aspects of aging—is being tackled with cognitive training apps and platforms that stimulate brain activity and decelerate the progression of diseases like

XSL•FO
RenderX

Alzheimer disease [51]. These technology-driven innovations underscore the transformative role of digital health in enhancing health management for older adults.

Our research findings reveal a significant shift toward a more integrated and personalized approach in older adult health management through the application of digital health technologies. This shift from traditional telemedicine to a comprehensive digital health system emphasizes a greater focus on enhancing individual health outcomes and overall well-being, rather than solely treating diseases. Our analysis underscores the transformative role of technology in health care, particularly through innovations such as AI, smart wearables, virtual reality, and ambient assisted living technologies, which collectively improve the quality of life and independence of the older adult population. Future research should aim to address the barriers to adoption and implement strategies that ensure these technological solutions are accessible, equitable, and tailored to the diverse needs of aging populations worldwide. Additionally, further studies are needed to evaluate the long-term impacts of these technologies on health outcomes and to understand better how they can be effectively integrated into standard care practices. This will not only help in advancing the field of digital health but also in shaping policies and practices that support healthy aging in various socioeconomic contexts.

While this study provides valuable insights, it is imperative to acknowledge its limitations. Primarily, our analysis draws upon data solely from the Science Citation Index Expanded and Social Sciences Citation Index databases, thus excluding studies from non–Sciences Citation Index journals, alternative databases, or conference presentations from our assessment. Additionally, it is essential to recognize the temporal influence on citations, whereby newly published works may receive fewer citations compared to earlier ones over time.

Conclusion

With the increasing trend of global population aging, technology-based solutions in older adults have developed rapidly, especially in the past decade. In the past 25 years, the focuses of the research are on older adults with chronic diseases, cognitive impairment, and social isolation, and some progress has been made in remote monitoring, education, personalized medicine, care, and ambient assisted living. The recent research focus has shifted toward digital health, indicating health and prevention. More research can be conducted in areas such as prediction, prevention, health management, and appropriate technological innovation in the future. At the same time, international cooperation should continue to be strengthened. Resource-limited countries can learn from the methods and strategies of resource-rich countries, combine their own characteristics, and provide more support for older adults.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020YFC2006401 and 2020YFC2006406), the Public Livelihood Foundation of the Department of Science and Technology of Liaoning Province (2021JH2/10300014), Science and Technology Projects in Liaoning Province (2023JH2/20200056 and 2023JH26/10300016), and National College Students' Innovation and Entrepreneurship Training Program (202110159005).

Authors' Contributions

LL, NC, XLW, FMY, and GWZ contributed to the concept and design. HJL and YM contributed to the acquisition of data. YHY, JXD, and YDW contributed to the analysis of data. LL, XLW, NC, and FMY prepared the first draft of the manuscript, and SW, DXG, and GWZ revised it. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Gu D, Andreev K, Dupre ME. Major trends in population growth around the world. China CDC Wkly 2021 Jul 09;3(28):604-613 [FREE Full text] [doi: 10.46234/ccdcw2021.160] [Medline: 34594946]
2. Penno E, Gauld R. Change, connectivity, and challenge: exploring the role of health technology in shaping health care for aging populations in Asia Pacific. Health Syst Reform 2017 Jul 03;3(3):224-235. [doi: 10.1080/23288604.2017.1340927] [Medline: 31514665]
3. Zhou B, Deng Q, Zhou S, Zhuo D. Health care in future community: innovatively discover and respond to the needs of today's seniors. Front Public Health 2023 Dec 12;11:1302493 [FREE Full text] [doi: 10.3389/fpubh.2023.1302493] [Medline: 38152669]
4. Wang Z, Yang Z, Dong T. A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time. Sensors (Basel) 2017 Feb 10;17(2):341 [FREE Full text] [doi: 10.3390/s17020341] [Medline: 28208620]
5. Ambagtsheer R, Shafiabady N, Dent E, Seiboth C, Beilby J. The application of artificial intelligence (AI) techniques to identify frailty within a residential aged care administrative data set. Int J Med Inform 2020 Apr;136:104094. [doi: 10.1016/j.ijmedinf.2020.104094] [Medline: 32058264]

XSL•FO
RenderX

6. Marston HR, Samuels J. A review of age friendly virtual assistive technologies and their effect on daily living for carers and dependent adults. Healthcare (Basel) 2019 Mar 21;7(1):49 [FREE Full text] [doi: 10.3390/healthcare7010049] [Medline: 30901898]
7. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021 Apr 19;7:20552076211009396 [FREE Full text] [doi: 10.1177/20552076211009396] [Medline: 33959378]
8. Ludwig W, Wolf K, Duwenkamp C, Gusew N, Hellrung N, Marschollek M, et al. Health-enabling technologies for the elderly--an overview of services based on a literature review. Comput Methods Programs Biomed 2012 May;106(2):70-78. [doi: 10.1016/j.cmpb.2011.11.001] [Medline: 22115611]
9. Gupta D, Bhatt S, Gupta M, Tosun AS. Future smart connected communities to fight COVID-19 outbreak. Internet Things (Amst) 2021 Mar;13:100342 [FREE Full text] [doi: 10.1016/j.iot.2020.100342] [Medline: 38620326]
10. Op den Buijs J, Simons M, Golas S, Fischer N, Felsted J, Schertzer L, et al. Predictive modeling of 30-day emergency hospital transport of patients using a personal emergency response system: prognostic retrospective study. JMIR Med Inform 2018 Nov 27;6(4):e49 [FREE Full text] [doi: 10.2196/medinform.9907] [Medline: 30482741]
11. Abraham C, Nishihara E, Akiyama M. Transforming healthcare with information technology in Japan: a review of policy, people, and progress. Int J Med Inform 2011 Mar;80(3):157-170. [doi: 10.1016/j.ijmedinf.2011.01.002] [Medline: 21292546]
12. Chan A, Cohen R, Robinson K, Bhardwaj D, Gregson G, Jutai JW, et al. Evidence and user considerations of home health monitoring for older adults: scoping review. JMIR Aging 2022 Nov 28;5(4):e40079 [FREE Full text] [doi: 10.2196/40079] [Medline: 36441572]
13. Hong Y, Wang Z, Cho JY. Global research trends on smart homes for older adults: bibliometric and scientometric analyses. Int J Environ Res Public Health 2022 Nov 10;19(22):14821 [FREE Full text] [doi: 10.3390/ijerph192214821] [Medline: 36429540]
14. Yenişehir S. Artificial intelligence based on falling in older people: a bibliometric analysis. Aging Med (Milton) 2024 Apr 09;7(2):162-170 [FREE Full text] [doi: 10.1002/agm2.12302] [Medline: 38725694]
15. Garçon L, Khasnabis C, Walker L, Nakatani Y, Lapitan J, Borg J, et al. Medical and assistive health technology: meeting the needs of aging populations. Gerontologist 2016 Apr 18;56 Suppl 2(Suppl 2):S293-S302. [doi: 10.1093/geront/gnw005] [Medline: 26994268]
16. Li T, Yang A, Liu G, Zou S, Chen Y, Ni B, et al. Status quo and research trends of craniopharyngioma research: a 10-year bibliometric analyses (from 2011 to 2020). Front Oncol 2021 Sep 30;11:744308 [FREE Full text] [doi: 10.3389/fonc.2021.744308] [Medline: 34660308]
17. He Z, Dai L, Zuo Y, Chen Y, Wang H, Zeng H. Hotspots and frontiers in pulmonary arterial hypertension research: a bibliometric and visualization analysis from 2011 to 2020. Bioengineered 2022 Jun 26;13(6):14667-14680 [FREE Full text] [doi: 10.1080/21655979.2022.2100064] [Medline: 35880647]
18. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug 31;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
19. Chen C. Searching for intellectual turning points: progressive knowledge domain visualization. Proc Natl Acad Sci U S A 2004 Apr 06;101 Suppl 1(Suppl 1):5303-5310 [FREE Full text] [doi: 10.1073/pnas.0307513100] [Medline: 14724295]
20. Booth FW, Roberts CK, Laye MJ. Lack of exercise is a major cause of chronic diseases. Compr Physiol 2012 Apr;2(2):1143-1211 [FREE Full text] [doi: 10.1002/cphy.c110025] [Medline: 23798298]
21. Chan M, Estève D, Escriba C, Campo E. A review of smart homes- present state and future challenges. Comput Methods Programs Biomed 2008 Jul;91(1):55-81. [doi: 10.1016/j.cmpb.2008.02.001] [Medline: 18367286]
22. Zheng YL, Ding XR, Poon CCY, Lo BPL, Zhang H, Zhou XL, et al. Unobtrusive sensing and wearable devices for health informatics. IEEE Trans Biomed Eng 2014 May;61(5):1538-1554 [FREE Full text] [doi: 10.1109/TBME.2014.2309951] [Medline: 24759283]
23. Silva BMC, Rodrigues JJPC, de la Torre Díez I, López-Coronado M, Saleem K. Mobile-health: a review of current state in 2015. J Biomed Inform 2015 Aug;56:265-272 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.003] [Medline: 26071682]
24. Sepúlveda-Loyola W, Rodríguez-Sánchez I, Pérez-Rodríguez P, Ganz F, Torralba R, Oliveira DV, et al. Impact of social isolation due to COVID-19 on health in older people: mental and physical effects and recommendations. J Nutr Health Aging 2020;24(9):938-947 [FREE Full text] [doi: 10.1007/s12603-020-1469-2] [Medline: 33155618]
25. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: Internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 02;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
26. Ong MK, Romano PS, Edgington S, Aronow HU, Auerbach AD, Black JT, et al. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the Better Effectiveness After Transition -- Heart Failure (BEAT-HF) randomized clinical trial. JAMA Intern Med 2016 Mar;176(3):310-318 [FREE Full text] [doi: 10.1001/jamainternmed.2015.7712] [Medline: 26857383]
27. Jaul E, Barron J. Age-related diseases and clinical and public health implications for the 85 years old and over population. Front Public Health 2017 Dec 11;5:335 [FREE Full text] [doi: 10.3389/fpubh.2017.00335] [Medline: 29312916]

XSL•FO
RenderX

28. Demiris G, Rantz M, Aud M, Marek K, Tyrer H, Skubic M, et al. Older adults' attitudes towards and perceptions of "smart home" technologies: a pilot study. Med Inform Internet Med 2004 Jun;29(2):87-94. [doi: 10.1080/14639230410001684387] [Medline: 15370989]
29. Liu L, Stroulia E, Nikolaidis I, Miguel-Cruz A, Rios Rincon A. Smart homes and home health monitoring technologies for older adults: a systematic review. Int J Med Inform 2016 Jul;91:44-59. [doi: 10.1016/j.ijmedinf.2016.04.007] [Medline: 27185508]
30. Liu J, Wang L, Fang H, Wang X, Wu L, Zhang J. Home-based cardiac rehabilitation: a review of bibliometric studies and visual analysis of CiteSpace (2012-2021). Medicine (Baltimore) 2022 Dec 09;101(49):e31788 [FREE Full text] [doi: 10.1097/MD.0000000000031788] [Medline: 36626492]
31. Fang EF, Xie C, Schenkel JA, Wu C, Long Q, Cui H, et al. A research agenda for ageing in China in the 21st century (2nd edition): focusing on basic and translational research, long-term care, policy and social networks. Ageing Res Rev 2020 Dec;64:101174 [FREE Full text] [doi: 10.1016/j.arr.2020.101174] [Medline: 32971255]
32. Leigh-Hunt N, Bagguley D, Bash K, Turner V, Turnbull S, Valtorta N, et al. An overview of systematic reviews on the public health consequences of social isolation and loneliness. Public Health 2017 Nov;152:157-171. [doi: 10.1016/j.puhe.2017.07.035] [Medline: 28915435]
33. Eshkoor SA, Hamid TA, Mun CY, Ng CK. Mild cognitive impairment and its management in older people. Clin Interv Aging 2015 Apr 10;10:687-693 [FREE Full text] [doi: 10.2147/CIA.S73922] [Medline: 25914527]
34. Nilsson MY, Andersson S, Magnusson L, Hanson E. Ambient assisted living technology-mediated interventions for older people and their informal carers in the context of healthy ageing: a scoping review. Health Sci Rep 2021 Mar 28;4(1):e225 [FREE Full text] [doi: 10.1002/hsr2.225] [Medline: 33392394]
35. Lunardini F, Luperto M, Romeo M, Basilico N, Daniele K, Azzolino D, et al. Supervised digital neuropsychological tests for cognitive decline in older adults: usability and clinical validity study. JMIR Mhealth Uhealth 2020 Sep 21;8(9):e17963 [FREE Full text] [doi: 10.2196/17963] [Medline: 32955442]
36. Ekstedt M, Kirsebom M, Lindqvist G, Kneck Å, Frykholm O, Flink M, et al. Design and development of an eHealth service for collaborative self-management among older adults with chronic diseases: a theory-driven user-centered approach. Int J Environ Res Public Health 2021 Dec 30;19(1):391 [FREE Full text] [doi: 10.3390/ijerph19010391] [Medline: 35010652]
37. Rutz M, Gerlach M, Schmeer R, Gaugisch P, Bauer A, Wolff D, et al. Providing knowledge and support for caring relatives with the smartphone - the MoCaB project [Article in German]. Pflege 2019;32(6):305-314. [doi: 10.1024/1012-5302/a000695] [Medline: 31389305]
38. Steiner B, Zippel-Schultz B, Popa A, Hellrung N, Szczesny S, Möller C, et al. CASEPLUS-SimPat: an intersectoral web-based case management system for multimorbid dementia patients. J Med Syst 2020 Feb 08;44(3):63 [FREE Full text] [doi: 10.1007/s10916-020-1533-9] [Medline: 32036462]
39. Xie B, Champion JD, Kwak J, Fleischmann KR. Mobile health, information preferences, and surrogate decision-making preferences of family caregivers of people with dementia in rural Hispanic communities: cross-sectional questionnaire study. J Med Internet Res 2018 Dec 10;20(12):e11682 [FREE Full text] [doi: 10.2196/11682] [Medline: 30530450]
40. Maresova P, Javanmardi E, Barakovic S, Barakovic Husic J, Tomsone S, Krejcar O, et al. Consequences of chronic diseases and other limitations associated with old age - a scoping review. BMC Public Health 2019 Nov 01;19(1):1431 [FREE Full text] [doi: 10.1186/s12889-019-7762-5] [Medline: 31675997]
41. Mohammed KI, Zaidan AA, Zaidan BB, Albahri OS, Alsalem MA, Albahri AS, et al. Real-time remote-health monitoring systems: a review on patients prioritisation for multiple-chronic diseases, taxonomy analysis, concerns and solution procedure. J Med Syst 2019 Jun 11;43(7):223. [doi: 10.1007/s10916-019-1362-x] [Medline: 31187288]
42. Collado-Mateo D, Lavín-Pérez AM, Peñacoba C, del Coso J, Leyton-Román M, Luque-Casado A, et al. Key factors associated with adherence to physical exercise in patients with chronic diseases and older adults: an umbrella review. Int J Environ Res Public Health 2021 Feb 19;18(4):2023 [FREE Full text] [doi: 10.3390/ijerph18042023] [Medline: 33669679]
43. Krebs P, Shtaynberger J, McCabe M, Iocolano M, Williams K, Shuk E, et al. An eHealth intervention to increase physical activity and healthy eating in older adult cancer survivors: summative evaluation results. JMIR Cancer 2017 Mar 01;3(1):e4 [FREE Full text] [doi: 10.2196/cancer.6435] [Medline: 28410171]
44. Fatehi F, Samadbeik M, Kazemi A. What is digital health? review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
45. Ronquillo Y, Meyers A, Korvek SJ. Digital health. StatPearls. Treasure Island, FL: StatPearls Publishing; 2023 May 1. URL: https://www.ncbi.nlm.nih.gov/books/NBK470260/ [accessed 2024-09-10]
46. Bin C, Li Q, Tang J, Dai C, Jiang T, Xie X, et al. Machine learning models for predicting the risk factor of carotid plaque in cardiovascular disease. Front Cardiovasc Med 2023 Sep 22;10:1178782 [FREE Full text] [doi: 10.3389/fcvm.2023.1178782] [Medline: 37808888]
47. Tokita J, Lam D, Vega A, Wang S, Amoruso L, Muller T, et al. A real-world precision medicine program including the KidneyIntelX test effectively changes management decisions and outcomes for patients with early-stage diabetic kidney disease. J Prim Care Community Health 2024;15:21501319231223437 [FREE Full text] [doi: 10.1177/21501319231223437] [Medline: 38185870]

XSL•FO
RenderX

48. Antos S, Danilovich MK, Eisenstein AR, Gordon KE, Kording KP. Smartwatches can detect walker and cane use in older adults. Innov Aging 2019 Jan;3(1):igz008 [FREE Full text] [doi: 10.1093/geroni/igz008] [Medline: 31025002]
49. Zhou H, Al-Ali F, Rahemi H, Kulkarni N, Hamad A, Ibrahim R, et al. Hemodialysis impact on motor function beyond aging and diabetes-objectively assessing gait and balance by wearable technology. Sensors (Basel) 2018 Nov 14;18(11):3939 [FREE Full text] [doi: 10.3390/s18113939] [Medline: 30441843]
50. Kenyon K, Kinakh V, Harrison J. Social virtual reality helps to reduce feelings of loneliness and social anxiety during the COVID-19 pandemic. Sci Rep 2023 Nov 07;13(1):19282 [FREE Full text] [doi: 10.1038/s41598-023-46494-1] [Medline: 37935718]
51. Tomaszewski Farias S, Fox J, Dulaney H, Chan M, Namboodiri S, Harvey DJ, et al. Memory support training and lifestyle modifications to promote healthy aging in persons at risk for Alzheimer's disease: a digital application supported intervention (Brain Boosters). BMC Geriatr 2023 Dec 21;23(1):881 [FREE Full text] [doi: 10.1186/s12877-023-04574-x] [Medline: 38129775]

Abbreviations

AI: artificial intelligence

Edited by G Eysenbach; submitted 18.06.24; peer-reviewed by XC Liu, A Sydor; comments to author 30.07.24; revised version received 14.08.24; accepted 31.08.24; published 20.09.24.

Please cite as:
Liu L, Wang XL, Cheng N, Yu FM, Li HJ, Mu Y, Yuan Y, Dong JX, Wu YD, Gong DX, Wang S, Zhang GW
Development Trends and Prospects of Technology-Based Solutions for Health Challenges in Aging Over the Past 25 Years: Bibliometric Analysis
J Med Internet Res 2024;26:e63367
URL: https://www.jmir.org/2024/1/e63367
doi:10.2196/63367
PMID:39238480

©Lu Liu, Xiu-Ling Wang, Nuo Cheng, Fu-Min Yu, Hui-Jun Li, Yang Mu, Yonghui Yuan, Jia-Xin Dong, Yu-Dan Wu, Da-Xin Gong, Shuang Wang, Guang-Wei Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Equity in Digital Mental Health Interventions in the United States: Where to Next?

Athena Robinson[1], PhD; Megan Flom[1], PhD; Valerie L Forman-Hoffman[1], MPH, PhD; Trina Histon[1], PhD; Monique Levy[1], MS; Alison Darcy[1], PhD; Toluwalase Ajayi[2], MD; David C Mohr[3], PhD; Paul Wicks[4], PhD; Carolyn Greene[5], PhD; Robert M Montgomery[1], MA

[1]Woebot Health, San Francisco, CA, United States

[2]Joan & Irwin Jacobs Center for Health Innovation, University of California, San Diego, San Diego, CA, United States

[3]Center for Behavioral Intervention Technologies, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[4]Sano Genetics, London, United Kingdom

[5]United States Department of Veterans Affairs, Mann-Grandstaff Veterans Affairs Medical Center, Spokane, WA, United States

Corresponding Author:
Athena Robinson, PhD
Woebot Health
535 Mission St Fl 14
San Francisco, CA, 94105
United States
Phone: 1 707 563 1759
Email: scicomms@woebothealth.com

Abstract

Health care technologies have the ability to bridge or hinder equitable care. Advocates of digital mental health interventions (DMHIs) report that such technologies are poised to reduce the documented gross health care inequities that have plagued generations of people seeking care in the United States. This is due to a multitude of factors such as their potential to revolutionize access; mitigate logistical barriers to in-person mental health care; and leverage patient inputs to formulate tailored, responsive, and personalized experiences. Although we agree with the potential of DMHIs to advance health equity, we articulate several steps essential to mobilize and sustain meaningful forward progression in this endeavor, reflecting on decades of research and learnings drawn from multiple fields of expertise and real-world experience. First, DMHI manufacturers must build diversity, equity, inclusion, and belonging (DEIB) processes into the full spectrum of product evolution itself (eg, product design, evidence generation) as well as into the fabric of internal company practices (eg, talent recruitment, communication principles, and advisory boards). Second, awareness of the DEIB efforts—or lack thereof—in DMHI research trials is needed to refine and optimize future study design for inclusivity as well as proactively address potential barriers to doing so. Trials should incorporate thoughtful, inclusive, and creative approaches to recruitment, enrollment, and measurement of social determinants of health and self-identity, as well as a prioritization of planned and exploratory analyses examining outcomes across various groups of people. Third, mental health care advocacy, research funding policies, and local and federal legislation can advance these pursuits, with directives from the US Preventive Services Taskforce, National Institutes of Health, and Food and Drug Administration applied as poignant examples. For products with artificial intelligence/machine learning, maintaining a “human in the loop” as well as prespecified and adaptive analytic frameworks to monitor and remediate potential algorithmic bias can reduce the risk of increasing inequity. Last, but certainly not least, is a call for partnership and transparency within and across ecosystems (academic, industry, payer, provider, regulatory agencies, and value-based care organizations) to reliably build health equity into real-world DMHI product deployments and evidence-generation strategies. All these considerations should also extend into the context of an equity-informed commercial strategy for DMHI manufacturers and health care organizations alike. The potential to advance health equity in innovation with DMHI is apparent. We advocate the field’s thoughtful and evergreen advancement in inclusivity, thereby redefining the mental health care experience for this generation and those to come.

(J Med Internet Res 2024;26:e59939) doi:10.2196/59939

KEYWORDS

Digital Mental Health Interventions; mental health; health equity; access to health care; health plan implementations

XSL•FO
RenderX

Introduction

The past few decades have marked significant momentum in the digital mental health field. More than 350,000 health-related mobile apps are available and a significant portion of these are specifically related to mental health support [1]. Enthusiasts postulate that digital mental health interventions (DMHIs) may eliminate commonplace access barriers [2,3] while also providing evidence-based mental health care that yields meaningful outcomes [4,5]. Moreover, there is a tenable undercurrent of hope, combined with appropriate and poignant questioning, on whether or not DMHI can genuinely and reliably reduce health care disparities. "Techquity" is a recently coined colloquial term that refers to this potential to either bridge or hinder equitable health care [6,7].

Problems with access to mental health care, both historical and present, are well documented across the United States and are especially pronounced within minoritized and underserved populations [8-13]. A key and increasingly common approach to understanding these inequities is through the study of social determinants of health (SDOH), which focuses on how the various circumstances in which people live affect their health and well-being [14,15]. Groups of individuals affected by systemic inequities in mental health care, in terms of both prevalence and care access, have been identified via various SDOH factors including, but not limited to, people who experience stigma or discrimination because of race, ethnicity, class, gender, sexual orientation (eg, LGBTQIA+ [Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Asexual+]); language or another aspect of identity; those with low household income or educational attainment; those who are underemployed or unemployed; older adults; those living in rural communities; members of Medicare or Medicaid; uninsured or underinsured people; veterans; and people living with disabilities [16-20]. When considering DMHIs, we must also acknowledge the "digital divide," as many individuals may lack access, consistent connectivity, skills, or trust to engage with digital mental health technologies, further contributing to and even widening existing inequities, as was the case during the COVID-19 pandemic [21-23].

The arc of *how* to progress toward bridging equitable health care is multifaceted, complex, dynamic, and of course, evergreen. Scholars have contributed important theoretical and applied work toward addressing inequities in digital health and DMHI, focusing on a range of topics from development to dissemination [24-28]. This viewpoint, in the spirit of the overarching special issue theme "Reflecting on Transformative Technologies, Interventions, Methods, and Policy Issues," aims to contribute to the conversation on health equity in US-based DMHIs in 2 ways. First, by referencing extant frameworks and best practices, and second, by articulating perspectives across key stages of product and organizational maturation, from initial product development to commercialization and health care ecosystem adoption (Figure 1; see also [14,25,28-41]). The present authorship collaboration will do this by drawing upon its collective expertise in several relevant domains, including scientific research in academia and industry, DMHI commercialization and deployment strategies from both industry and health care system perspectives, DMHI product development, and organizational leadership. We also include salient illustrative examples from real-world development at Woebot Health, a US-based company that manufactures DMHIs, with which a number of the authors are affiliated.

Figure 1. Integral elements toward building, sustaining, and advancing equitable mental health care and corresponding recommended reading. DMHI: digital mental health intervention; LLM: large language model; PIDAR: Partner, Identify, Demonstrate, Access, Report; USPSTF: US Preventive Services Taskforce.

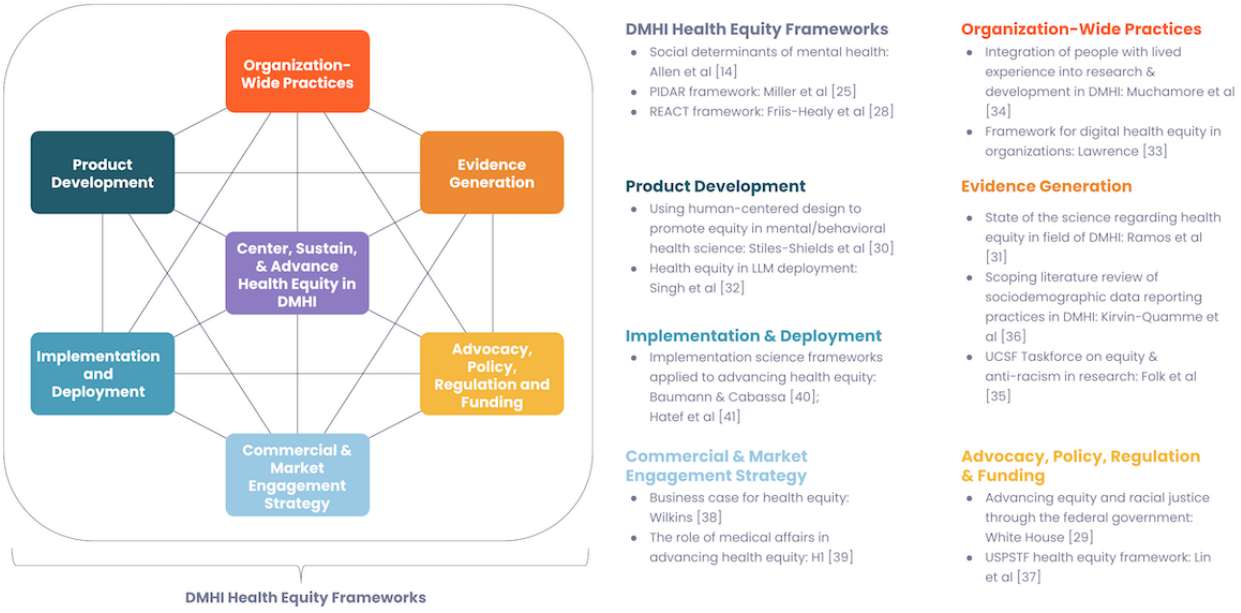

XSL•FO
RenderX

To begin to articulate the arc, we first focus on key elements in product and intervention development. Available development frameworks call for requisite community involvement, partnership, and trust throughout as well as embedded technological support for varied digital literacy levels [25,28]. Beyond development, multiple DMHI app evaluation frameworks have been published [42-44], some with specific callouts for the import of diversity equity and inclusion therein [45]. Moreover, industry manufacturers and academicians alike are encouraged to consider their company's or lab's integration of diversity, equity, inclusion, and belonging (DEIB) principles into talent recruitment, workstreams, and advisory boards. Next, we discuss considerations in evidence generation. Several recent systematic reviews and meta-analyses suggest that DMHI is effective for addressing symptoms of depression and anxiety [46-48], including across the age range from adolescents [49] to older adults [50]. However, simultaneous calls to action for enhanced diversity in clinical trials underscored the ongoing need for inclusive research [29,51], calling into question whether these conclusions may be biased due to unrepresentative samples [52]. Methodological strategies to facilitate inclusive clinical trial designs will be considered. The adoption of evidence-based DMHIs has been encouraged by the American Medical Association [53] and American Psychiatric Association [54], perhaps augmenting existing options for mental health care [55]. We also highlight the need for various ecosystems (academic, industry, payer, provider, and value-based care organizations) to articulate their requirements for health equity to be a part of a manufacturer's DMHI product and evidence-generation portfolio before adoption or deployment within a system. Specification of considerations of an equity-informed commercialization strategy for evidence-based DHMI will be discussed.

Given the promising yet relatively nascent status of the DMHI field, we encourage transparent and open conversation and a collective articulation to help specify and manifest an evergreen process of monitoring for and reducing mental health care inequities.

Product Development

As noted in the World Health Organization's Global Strategy on Digital Health 2020-2025, "When planning and prioritizing digital health interventions...the specific potential of digital technologies to promote health equity should be leveraged. Designed properly, digital solutions can propel inclusiveness as digital connectivity can transcend physical barriers" [56]. As any creation usually bears some imprint of the creator, a key starting principle for building an inclusive service is ensuring that the team that builds it represents a diverse set of perspectives and backgrounds, which include those of the intended users. Equity must also be baked into the product development life cycle. Similar to the European Union's General Data Protection Regulation (GDPR) principle of "Privacy by Design," we may think about this as "Equity by Design." This entails explicit exploration of whether equity is likely to be achieved early on in design sessions, identifying any potentially problematic pitfalls and addressing them systematically. Human-centered design roots itself in the philosophy that thorough discovery and understanding of human needs optimize product design to meet those needs, as well as making the experience meaningful and satisfying [57]. Some human-centered design frameworks have been tailored to address DMHIs broadly [58], while others specifically target improving the health equity of DMHI [30,59]. For example, human-centered design approaches, which organically focus on inviting and hearing needs at hand, should be purposefully extended to include voices from marginalized communities [30,60], including communities with diverse digital literacy levels [61]. The PIDAR (Partner, Identify, Demonstrate, Access, Report) framework, developed by the Society of Behavioral Medicine's Health Equity Special Interest Group, highlights partnerships as both integral to avoiding increased health inequities in DMHI and a fundamental first step in an equity-centric intervention development approach [25]. Methodologies for community-based participatory research have been articulated and incorporate strategies such as rapid prototyping and iteration for efficient colearning, as well as key human-centric and empathy-first principles of commitment, collaboration, and listening [62-64].

As more DMHIs begin to integrate aspects of artificial intelligence (AI), machine learning (ML), and large language models (LLMs), issues surrounding health equity in the rapidly evolving world of AI/ML are particularly salient to this conversation. These technologies hold great potential to, for example, enhance a conversational agent's ability to understand natural language inputs as well as generate nuanced and tailored responses. However, there are a number of risks associated with the use of LLMs in mental health, both general and specific to health equity concerns. General concerns are LLM's potential to "hallucinate" or generate false or misleading information as well as the limitations imposed by training data, which can be biased, inaccurate, or even actively harmful [65]. Equity-related LLM risks include their potential to exacerbate or amplify disparities if algorithms are constructed using data that reflect historical or societal biases and inequities; additionally, they increase skepticism and alienation of minoritized individuals due to lack of transparency and perception of bias [66,67]. Furthermore, many LLMs are trained primarily on a corpus of English sources, and may only be available in certain languages (eg, English), creating potential biases and risks for nonprimary language learners in understanding and interpreting these interventions, or directly excluding large swaths of people who do not speak certain languages [31,68,69]. Health equity–centered AI development frameworks are emerging to proactively address such risks [70-72]. For example, frameworks focused on the integration of LLMs into health care products recommend that LLMs are (1) thoroughly evaluated for equity-related risks, (2) rigorously monitored during deployment, (3) integrated into systems with humans-in-the-loop (eg, providers or users) to guard against biased algorithmic drift, and (4) trained in collaboration with impacted communities or populations they are intended for, all of which can help mitigate the exacerbation of potential biases [32,61,72]. Certainly, more research as well as the development and refinement of regulatory frameworks for responsible AI utilization in DMHI is called for overall [73]. President Biden's Executive Order highlights the need to protect the "rights and safety of the public" through

XSL•FO
RenderX

(1) strengthening AI governance, (2) advancing responsible AI innovation, and (3) managing risk from the use of AI [74].

Organizationwide Practices

Embedding DEIB practices into the fabric of organizations, from corporate objectives to hiring practices and more, has been advocated for openly, and is another key element of the larger techquity conversation with implications for DMHIs [33,75]. There is strong support for the benefits that a diverse workforce can bring to an organization in terms of key business outcomes (eg, innovation and revenue generation), as well as aligning with what talent may seek within a company's culture [76,77]. For mental health–focused organizations, the inclusion of people with lived experience of mental health challenges is especially critical, as their involvement in all aspects of DMHI research, development, policy, and practice is fundamental to promoting equity [34,78]. Diversity at both the inception of digital health innovation (eg, within design teams at DMHI manufacturing organizations) and in health care systems deploying such innovations (eg, accountable care organizations or integrated delivery networks) is part of the holistic picture [33]. Organizational commitment to various endeavors and working groups (eg, podcasts, blogs, hiring practices, educational seminars, advisory boards, health benefits, and employee well-being programs, task forces) that host and elevate conversations on equity promote recognition of its fundamental importance as well as set a regular cadence of expected updates or outputs [79-82]. Examples are the DMHI industry commitment to supporting their own employees' mental health via a variety of DEIB efforts (ie, Headspace [83]), the establishment of a clinical diversity advisory board (ie, Woebot Health [80]), and podcast interviews with industry leaders to discuss key DEIB topics (ie, Meeting of the Minds Podcast; Health Nonprofit Digital Marketing [82,84]). Powerful work led by the University of California San Francisco's Taskforce on Equity and Anti-Racism in Research underscored several core learnings from their endeavors, among them a clear call for a diverse workforce across the full membership of the research team, as well as ongoing education and support for DEIB in research [35]. The National Institutes of Health (NIH) offers awards dedicated to engaging undergraduate students from diverse backgrounds in biomedical research [85,86] and Rush Education and Career Center Hub (REACH) offers support in STEM (Science, Technology, Engineering, Mathematics) research to students from preschool age onward [87]. Both of these highlight direct efforts to diversify researchers of the future.

Evidence Generation

Enhancing Health Equity in Evidence Generation for DMHIs: Frameworks and Methodological Considerations

The generation of evidence of the feasibility, acceptability, efficacy, effectiveness, and safety of DMHIs provides another critical opportunity for the consideration of health equity. Existing frameworks that have been used to explore the centering of health equity in research are PIDAR, RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance), and others, many of which highlight partnerships with diverse stakeholders and potential users, consideration of how access to technology and digital literacy affect eligibility, and how data are collected and disseminated [25,26,31]. In addition to these frameworks, we offer further methodological considerations to center evidence generation in health equity, drawing on our experience in implementing these practices.

Meta-analytic research has generally supported DMHI's efficacy, feasibility, and acceptability [46-48]. However, despite the anticipated promise of DMHIs in addressing the health equity gap, little is known about outcomes across various sociodemographic subgroups and whether the trials recruited diverse samples; additionally, only a few studies have specifically focused on interventions intended for minoritized populations [31,88]. For example, a recent systematic literature review spearheaded by Woebot Health highlighted the underreporting of sociodemographic characteristics in clinical trials evaluating DMHIs and pointed to even fewer reporting any results by sociodemographic characteristics [36]. To examine the real impact of DMHIs on health equity, evidence-generation methodology needs to be thoughtfully designed and implemented.

Recruitment and Trial Design

First, the recruitment of diverse samples is the foundation upon which evidence generation rests. A variety of tactics can be used to recruit diverse samples, including the use of culturally sensitive recruitment materials; community outreach; thoughtful partnerships with socially, economically, geographically, or racially diverse health care systems; resources to support those with limited access to the internet; and diverse research and recruitment teams [25,28]. For example, in a recent research collaboration between Scripps Translational Science Institute and Woebot Health, utilization of such methods facilitated successful recruitment and enrollment of the a priori target of approximately 50% of a sample self-identifying as individuals from categories noted as historically underrepresented in biomedical research, including racial/ethnic minorities and rural/nonmetropolitan area residents, among others [89]. Real-world partnerships are particularly important for inviting collaboration with marginalized groups and exploring the potential of the DMHI to address health equity, with the necessity of real-world evidence outlined as an important component in several frameworks on health equity within DMHIs [26,28,31]. In these cases, clinician-referred recruitment should be managed with care and not utilized as a stand-alone recruitment strategy. Medical research has shown that clinicians and researchers may selectively exclude certain minority participants due to biases or assumptions about their interest in research, ability to follow study protocols, and likelihood of dropping out, among others [90]. Indeed, a large review [91] of minority participation in health research demonstrated that while minorities are willing to participate, they are less likely to be recruited. Where possible, health systems, patient registries, or other common recruitment sources should consider an "opt-out" method for study recruitment invitations. Specifically, upon entry to the system/registry, patients consent by default to be contacted for research opportunities unless they

XSL•FO
RenderX

choose to "opt out" of such communications. This might allow for a greater diversity of individuals to be invited to participate in studies in the first place, without compromising their agency to decide which particular studies to apply for. To guide recruitment, specific targets for race, ethnicity, sexual orientation, and gender, among others, should be determined a priori based on the desired end users and relevant prevalence rates. While attempting to create targets based on census data or population prevalence is one viable option, we suggest aligning targets more closely with the prevalence rates of mental health problems in specific sociodemographic groups to more accurately capture real-world needs [92]. Additionally, to have a sufficient sample size to examine group differences, recruiting large enough samples or overrecruiting for a particular group(s) of interest may be necessary.

Sociodemographic Surveys and Data Collection

Second, sociodemographic surveys need to be thoughtfully designed and consistently implemented. Sociodemographic reporting should include culturally sensitive and inclusive response options to questions on race, ethnicity, sexual orientation, and gender identity, as well as questions related to SDOH such as food and housing insecurity [93]. To promote honest disclosure, participants should be given the opportunity to respond electronically and be assured of the privacy and protection of their sensitive data. Of note, it is often more challenging to collect comprehensive sociodemographic and SDOH data in the context of deployment outside of research settings. Thoughtful approaches to the collection and use of these data should be agreed upon by stakeholders and tested with end users to ensure that factors such as assessment burden, privacy concerns, and potential cultural differences in the appropriateness of questionnaire content are appropriately considered.

Analysis and Outcomes

Third, analyzing and reporting on sociodemographic characteristics and outcomes across subgroups are integral to examining the impact of DMHIs on health equity. At a minimum, DMHI research should provide the sociodemographic characteristics of their sample. Additionally, consumer-based DMHI data (outside the research context) should be presented in dashboards and reports that include a breakdown of sociodemographic characteristics, where possible and sample size permitting. Ideally, outcomes are also reported by sociodemographic groups of interest, or included as covariates in models to better understand their potential impact on or association with outcomes (see, eg, [94]). However, given the small sample sizes in many sociodemographic subgroups of interest, traditional hypothesis-confirming statistical methods of significance testing and modeling may not be possible. For such cases, we suggest focusing on exploratory, descriptive, and hypothesis-generating approaches such as within-group effect sizes (eg, the effect size of a symptom change score in non–Hispanic Black participants) and between-group effect sizes (eg, the effect size of the difference between symptom change scores in non–Hispanic Black and non–Hispanic White participants) where possible. Consolidating certain subgroups into another broader group may also be necessary in certain cases (eg, collapsing "genderqueer," "nonbinary," "agender," and "different identity" into "other" for gender identity; see [95]), ideally to reach a sample size of at least 20-25 per subgroup [96], although greater care must be taken in the interpretation of these less precise groupings. For this exploratory approach, replication is key to identifying consistent and reliable patterns, and these analyses should be prespecified and conducted for every study even if only for internal purposes. Data from these exploratory analyses can be used to identify potential questions and hypotheses for future research, provide preliminary data for power analyses and study planning, and inform intervention improvements and possible precision targets. Researchers should also go beyond the assessment of differences in outcomes by sociodemographic characteristics, and consider how they can identify, capture, and analyze the structural factors (eg, discrimination, exclusionary policies, digital redlining) that may be responsible for creating or exacerbating inequities [97,98].

The outcomes assessed should be comprehensive and include efficacy (eg, symptom change), engagement (eg, app use metrics, therapeutic alliance), satisfaction and feasibility (including trust and perception of bias), and safety. Moreover, the assessed outcomes must reflect not only the priorities and values of the researchers, but also those of the participants. Doing so allows for a more nuanced understanding of the impact of DMHIs on various groups of people, and is necessary for a precision-based research and development approach that can be leveraged to address health equity and create better products. Qualitative data (eg, free-text questions about app experience, participant interviews) should be solicited, as they illuminate the nuance of experience across different groups of people and can reveal previously unknown issues or concerns among users.

Safety, in particular, is an important outcome to examine in the context of health equity [99]. Despite this, not all DMHI trials provide safety data and there is little standardization in how adverse events are captured and reported [100]. This may be, to a certain extent, because different types of DMHIs require varying degrees of safety capturing and oversight. Trials also vary considerably in the level of clinical support to mitigate and manage potential adverse events. Safety measurement must align with each intervention's risk and capabilities to address safety concerns, while being as consistent as possible across DMHIs. Lastly, when generating internal safety reports for each trial, safety events and any other related data should be provided alongside sociodemographic data to allow exploration of prevalence rates within different groups of people.

All of these equity-centering research practices, from recruitment and data collection to analysis and safety considerations, should also be integrated into iterative internal product development processes and user experience (UX) research as well as external evidence-generation efforts.

Advocacy, Policy, Regulation, and Funding

Shifting our view "up" to another level, we turn to considerations of advocacy, policy, regulation, and funding,

XSL•FO
RenderX

which may unlock paths to address inequities unavailable at the individual or even company level [27]. For example, the translation of mental health evidence–generation findings into new policies intended to improve health equity in mental health care has been bolstered by the implementation of advocacy efforts and new policy, regulation, and funding efforts at both the federal and local levels [52,101,102]. Advocacy efforts such as lobbying and grassroots campaigns focused on improving the equitable distribution of mental health services and awareness of mental health disparities have been critical in advancing policy change and allocation of resources as well [103,104].

In part due to these advocacy efforts, the past several decades have brought about various legislative initiatives focused on parity, community mental health initiatives, workforce diversity, and trauma-informed care that have served to improve mental health equity. Perhaps the most groundbreaking example of federal parity legislation in the past 25 years was the implementation of the Mental Health Parity and Addiction Equity Act of 2008 [105] which required insurance companies to offer mental health and substance use disorder care coverage similar to that of medical care, paving the way for equitable access to the support and services for those in need. Broader efforts to expand internet access have also evolved in recent years. These initiatives include broadband expansion that uses public or private funding to build the infrastructure needed to deliver internet to areas without current capabilities [106], net neutrality regulations established in 2003 [107], public Wi-Fi funded by governments or municipalities, and digital literacy programs that teach the fundamentals of using the internet [108]. The ability of users to connect online to DMHIs has been greatly enhanced by these collective efforts.

The regulatory oversight of medical devices by the Food and Drug Administration (FDA) can also impact digital mental health equity. Clinical trial diversity initiatives, regulatory science research and policies, public health education initiatives, postmarket surveillance activities, and drug/device presubmission interactions and approval processes have the ability to support or leave vulnerable populations in need of additional digital mental health care [109-111]. For example, following the onset of the COVID-19 pandemic and the nationwide public health emergency, the FDA announced in April 2020 that it would temporarily waive certain requirements for digital therapeutics targeting psychiatric disorders. This measure aimed to provide urgently needed mental health support tools to a struggling population, although it expired in November 2023 [112].

In addition, various organizations such as the American Psychological Association (APA) [101], the National Institute of Mental Health (NIMH) [52], the US Preventive Services Taskforce (USPSTF) [37], the Substance Abuse and Mental Health Services Administration (SAMHSA) [102], and others have provided support for recommendations and funding that have advanced mental health equity. For example, the USPSTF recently developed a framework to incorporate health equity into each step of their recommendation-making process. The NIMH's ongoing Laboratories to Optimize Digital Health program seeks to fund innovative research projects that study ways to increase access, efficacy and effectiveness, and quality of DMHIs, particularly among those who experience health disparities [113]. Beyond the critical research focused on interventions to help more vulnerable populations, researchers should also focus on understanding individual-level differences in the prevalence, root causes, and mechanisms of expression of different mental health challenges [114-116]. Additional funding opportunities should be made available to train underrepresented minorities, develop culturally competent interventions, and determine the impact of the implementation of interventions or policies within underserved communities and health care settings. Trends in research that require the involvement of patients with lived experiences, interventions integrating peer support, and community-based participatory research [63] highlight the importance of including multiple stakeholders to identify priorities, develop and test interventions, and inform policy. These efforts reflect a commitment to advocating for and enacting policies that guide, regulate, and fund work that ensures a wide variety of users can access and benefit from DMHIs. Progress in these areas will be crucial in advancing equity across the field of DMHI in the future.

Commercial and Market Engagement Strategy

For-profit companies and commercial endeavors represent an opportunity to accelerate the traction of equity-related efforts in mental health. Company success in terms of market penetration and financial growth are well aligned and in many cases dependent on the factors described earlier (DEIB maturity in the workforce, more inclusive practices in product design and evidence generation, more representative and true to real-world evidence gathering, and reinforcement of equity-related policy). Today's market rewards and reinforces equity-centered practices as customers are seeking products that meet sophisticated requirements in this regard. Deloitte estimated that their life sciences clients could garner a 10× return on their investment by bolstering equitable care access in underserved markets [38] and entrepreneurs focused on mental health equity are attracting early support and funding [117]. That said, challenges exist for for-profit ventures as well. For example, the cost of providing DMHI access to "hard to reach" populations (eg, those without a smartphone or internet access, or who are members of groups too small to amass commercial attention) may be greater, impacting profitability and sustainability of outreach efforts (see Galea [118] for an analogous case related to the greater cost of vaccinating certain vulnerable populations). Moreover, in cases where end users, and not an intermediary organization (eg, an employer, health plan, or health care provider), directly bear some or all of the cost for access to a DMHI, existing financial inequities may perpetuate or exacerbate inequities in mental health access and outcomes [119]. In both these cases, commercial stakeholders may consider steps such as advocating for ring-fenced funding, or allocation of protected funds that are specifically and exclusively intended for the purpose of reaching historically minoritized and underserved populations. Importantly, the purpose of these protected funds would not be to optimize customer acquisition costs, but rather to optimize representation.

XSL•FO
RenderX

Commercial and medical affairs functions in health care organizations such as DMHI and pharmaceutical manufacturers play a vital role in translating and reinforcing progress from science/policy out to the market and market demands back to internal groups [120]. Customer listening allows firms to collect very clear requirements on how products need to function and be deployed to address real-world equity-related challenges, helping product and evidence teams prioritize their road maps [39]. In the other direction, commercial and medical affairs teams showcase emerging work by science and product, building confidence and momentum in the market to go after equity-related unmet needs with more urgency.

Strong and creative commercial and medical affairs functions can also help build differentiated paths to market by developing tailored marketing, enrollment, and payment models to meet the specific needs of underserved populations. Customers who struggle to meet quality standards or equity requirements may be more inclined to pursue digital solutions aggressively. Manufacturers can foster more effective deployments by offering configurable solutions and flexibility in implementation and marketing. Additionally, not all DMHI manufacturers operate as for-profit ventures; exploring other models for addressing health inequities, such as the US State Department's "Public Health 3.0" initiative [121] and nonprofit business models [122], is crucial for scaling new products.

Equity-Informed Implementation Science in Real-World Deployments

Although it appears as the final section of this paper, the importance of thoughtful consideration of health equity within and across DMHI ecosystem deployments cannot be underscored enough. This view is critical because it spans the individual, familial, and community levels, as well as the larger encompassing "ecosystem" organizations, such as health care systems or provider networks, which may be in unique positions to operate across multiple levels of SDOH to address inequities [27]. Deployment is also where "the rubber meets the road," and the actual impact of DMHIs is realized.

Implementation science frameworks offer considerations into how evidence-supported interventions are "implemented" in routine practice in naturalistic, real-world contexts; and may, for example, recommend evaluation of factors such as immediate and sustained adoption of the intervention, cost, and barriers therein [123]. Such real-world data are of paramount importance in determining not only the real-world effectiveness of DMHIs, but also integral for informing and refining approaches to deployment processes (ie, provider training and patient-facing information). Moreover, such data create an opportunity for monitoring if the DHMI or its deployment characteristics are closing or exacerbating equity gaps, paving the way for deployment adjustments accordingly.

Ecosystems purchasing or deploying DMHIs (eg, payer, provider, value-based care organizations) are encouraged to transparently and reliably emphasize the need for health equity to be included in a manufacturer's DMHI product and evidence-generation portfolio before adoption or deployment within their system. To begin, existing implementation science frameworks have been adapted to focus on health equity and can provide thoughtful questions for ecosystems and deployment partners to consider. These may include an early focus on "reach" (ie, who is included and excluded in different deployment strategies), proactive design and selection of interventions for vulnerable populations, and measuring and monitoring outcomes through the lens of addressing inequities [40]. Ecosystems may also consider adopting a health equity–informed implementation science framework [33], such as the Evidence- and Consensus-Based Digital Healthcare Equity Framework [41], or may choose to create and implement their own, as demonstrated by health care organizations such as Elevance [124], Kaiser Permanente [125], Reliant Medical Group [126], and the Centers for Medicare and Medicaid Services [127].

One example of how health equity might be considered in the deployment of a DMHI within a health care ecosystem is through the collection, monitoring, and evaluation of SDOH data. DMHIs (or their underlying platform infrastructure) could be designed to thoughtfully gather information about specific SDOH factors from patients. Such data could inform further risk assessment or clinical follow-ups, and facilitate the offering of available resources provided by the system to the patient. Processes like this may enable earlier intervention, more precise tailoring of interventions, and greater utilization of available support and resources for those in greatest need. This approach has the potential to benefit underserved individuals, DMHI manufacturers, and health care systems as the shift toward whole-person care evolves.

New models of care that support DMHI implementation are also emerging in postpandemic health care delivery: one such role is the digital navigator. This role facilitated the rapid transition to telehealth in the early days of the pandemic as a means to support patients accessing platforms to access care. The digital navigator can access which phone and data plan a patient has to optimally match them with a DMHI that takes the whole-person's life context into account. This role is one method for addressing the friction from referral to activation that is often experienced in real-world deployments [128]. Other promising examples of alternative care models are stepped care [129,130] and collaborative care [131,132], both of which have accrued growing bodies of support in the literature, and which offer opportunities for DMHI integration and implementation. As these new models of care mature, data exchange to support the optimal levels of human and digital touchpoints can support the broader use of DMHIs in populations that may not traditionally seek care for mental health.

Conclusions

The potential to advance health equity through innovation in DMHIs is apparent, but by no means guaranteed. Here, utilizing decades of learning across multiple domains of expertise, we explored the progress and opportunities within the DMHI field to address mental health care inequities and articulated several steps essential to mobilize and sustain meaningful forward progression in this endeavor. While notable progress has been

XSL•FO
RenderX

made over the past 25 years, we advocate for the field's thoughtful and evergreen advancement in inclusivity and equity, thereby continuing to redefine the mental health care experience for this generation and those to come.

Acknowledgments

We thank Adam Platt, MSc, for coordinating scientific communications and facilitating the organization of the manuscript preparation and submission, as well as Katherine McGuckin, MS, for conceptual and drafting contributions to the "Commercial and Market Engagement Strategy" section. We also thank Julie Ross and Robbert Zusterzeel for their review and comments, and Stephanie Rapoport for helping to design the summary figure. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Funding to publish this article as open access was provided by Woebot Health.

Authors' Contributions

AR, in collaboration with RMM, MF, and AD, conceptualized the piece and outlined it. The introduction was drafted by AR and RMM. The "Product Development" section was drafted by AR with support from RMM and AD. The "Organizationwide Practices" section was drafted by AR. The "Evidence Generation" section was drafted by MF, with support from RMM and VLFH. The "Advocacy, Policy, Regulation, and Funding" section was drafted by VLFH with support from RMM. The "Commercial and Market Engagement Strategy" section was drafted by ML, AR, and RMM. The "Equity-Informed Implementation Science in Real-World Deployments" section was drafted by AR, RMM, and TH. TA, DCM, PW, and CG reviewed and provided contributions and feedback throughout. All authors contributed to multiple draft revisions, with ownership for incorporation of all author edits and flow of the overall manuscript championed by RMM and AR. All authors read and approved the submitted version.

Conflicts of Interest

AR, MF, VLFH, AD, TH, ML, and RMM are all employees of Woebot Health. DCM has accepted honoraria and consulting fees from Boehringer-Ingelheim, Optum Behavioral Health, Centerstone Research Institute, and the One Mind Foundation; royalties from Oxford Press; and has an ownership interest in Adaptive Health, Inc. TA has nothing to declare. PW is an associate editor at the Journal of Medical Internet Research and is on the editorial advisory boards of BMC Medicine, The Patient, and Digital Biomarkers; was a former consultant to Woebot Health; and is currently employed by Sano Genetics and holds stock/options in Sano Genetics. CG is a consultant to Woebot Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. IQVIA Institute for Human Data Science. Digital health trends 2021: innovation, evidence, regulation, and adoption. IQVIA. 2021. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/digital-health-trends-2021 [accessed 2024-04-23]
2. Radez J, Reardon T, Creswell C, Lawrence PJ, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry 2021 Feb;30(2):183-211 [FREE Full text] [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
3. National Institutes of Mental Health (NIMH). Technology and the future of mental health treatment. NIMH. 2023. URL: https://www.nimh.nih.gov/health/topics/technology-and-the-future-of-mental-health-treatment [accessed 2024-04-23]
4. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022 May 12;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
5. Faria M, Zin STP, Chestnov R, Novak AM, Lev-Ari S, Snyder M. Mental health for all: the case for investing in digital mental health to improve global outcomes, access, and innovation in low-resource settings. J Clin Med 2023 Oct 25;12(21):6735 [FREE Full text] [doi: 10.3390/jcm12216735] [Medline: 37959201]
6. Clark CR, Akdas Y, Wilkins CH, Rhee K, Johnson KB, Bates DW, et al. TechQuity is an imperative for health and technology business: let's work together to achieve it. J Am Med Inform Assoc 2021 Aug 13;28(9):2013-2016 [FREE Full text] [doi: 10.1093/jamia/ocab103] [Medline: 34157112]
7. Ipsos, The HLTH Foundation. Ipsos. 2022. URL: https://www.ipsos.com/sites/default/files/ct/publication/documents/2022-03/Ipsos-HLTH-Path-to-techquity_0.pdf [accessed 2024-04-23]
8. Cook BL, Trinh N, Li Z, Hou SS, Progovac AM. Trends in racial-ethnic disparities in access to mental health care, 2004-2012. Psychiatr Serv 2017 Jan 01;68(1):9-16 [FREE Full text] [doi: 10.1176/appi.ps.201500453] [Medline: 27476805]
9. Cook BL, Hou SS, Lee-Tauler SY, Progovac AM, Samson F, Sanchez MJ. A review of mental health and mental health care disparities research: 2011-2014. Med Care Res Rev 2019 Dec;76(6):683-710. [doi: 10.1177/1077558718780592] [Medline: 29877136]

XSL•FO
RenderX

10. Erikson CE, Dent RB, Park YH, Luo Q. Historic redlining and contemporary behavioral health workforce disparities. JAMA Netw Open 2022 Apr 01;5(4):e229494. [doi: 10.1001/jamanetworkopen.2022.9494] [Medline: 35482310]
11. Institute of Medicine (US). Introduction. In: How Far Have We Come in Reducing Health Disparities? Progress Since 2000: Workshop Summary. Washington, DC: National Academies Press; 2012:1.
12. Salmond S, Dorsen C. Time to reflect and take action on health disparities and health inequities. Orthop Nurs 2022;41(2):64-85 [FREE Full text] [doi: 10.1097/NOR.0000000000000828] [Medline: 35358124]
13. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
14. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. Int Rev Psychiatry 2014 Aug;26(4):392-407. [doi: 10.3109/09540261.2014.928270] [Medline: 25137105]
15. Marmot M. Social determinants of health inequalities. Lancet 2005;365(9464):1099-1104 [FREE Full text] [doi: 10.1016/S0140-6736(05)71146-6] [Medline: 15781105]
16. National Institute of Mental Health (NIMH). Medicaid data show wide differences in mental health care in the United States. National Institute of Mental Health (NIMH). 2023. URL: https://www.nimh.nih.gov/news/science-news/2023/medicaid-data-show-wide-differences-in-mental-health-care-in-the-united-states [accessed 2024-04-23]
17. Williams ND, Turpin RE, Akré ERL, Boekeloo BO, Fish JN. Disparities in mental health care access among persons differing in sexual identity: nationally representative findings. Psychiatr Serv 2022 Apr 01;73(4):456-459 [FREE Full text] [doi: 10.1176/appi.ps.202100045] [Medline: 34407630]
18. Kirkbride JB, Anglin DM, Colman I, Dykxhoorn J, Jones PB, Patalay P, et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry 2024 Feb;23(1):58-90 [FREE Full text] [doi: 10.1002/wps.21160] [Medline: 38214615]
19. Compton MT, Shim RS. The social determinants of mental health. FOC 2015 Oct;13(4):419-425. [doi: 10.1176/appi.focus.20150017]
20. Alegría M, NeMoyer A, Falgàs Bagué I, Wang Y, Alvarez K. Social determinants of mental health: where we are and where we need to go. Curr Psychiatry Rep 2018 Sep 17;20(11):95 [FREE Full text] [doi: 10.1007/s11920-018-0969-9] [Medline: 30221308]
21. Digital access: a super determinant of health. Substance Abuse and Mental Health Services (SAMHSA). 2023. URL: https://www.samhsa.gov/blog/digital-access-super-determinant-health [accessed 2024-04-23]
22. Beaunoyer E, Dupéré S, Guitton MJ. COVID-19 and digital inequalities: reciprocal impacts and mitigation strategies. Comput Human Behav 2020 Oct;111:106424 [FREE Full text] [doi: 10.1016/j.chb.2020.106424] [Medline: 32398890]
23. Spanakis P, Peckham E, Mathers A, Shiers D, Gilbody S. The digital divide: amplifying health inequalities for people with severe mental illness in the time of COVID-19. Br J Psychiatry 2021 Oct;219(4):529-531 [FREE Full text] [doi: 10.1192/bjp.2021.56] [Medline: 35048887]
24. Simmons D, Kaganoff S, Drasser K. Building toward equity: a working model for digital health. Rock Health. 2021 Aug. URL: https://rockhealth.com/insights/building-toward-equity-a-working-model-for-digital-health/ [accessed 2024-04-23]
25. Miller SJ, Sly JR, Alcaraz KI, Ashing K, Christy SM, Gonzalez B, et al. Equity and behavioral digital health interventions: strategies to improve benefit and reach. Transl Behav Med 2023 Jun 09;13(6):400-405 [FREE Full text] [doi: 10.1093/tbm/ibad010] [Medline: 36940409]
26. Jaworski BK, Webb Hooper M, Aklin WM, Jean-Francois B, Elwood WN, Belis D, et al. Advancing digital health equity: directions for behavioral and social science research. Transl Behav Med 2023 Apr 03;13(3):132-139 [FREE Full text] [doi: 10.1093/tbm/ibac088] [Medline: 36318232]
27. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
28. Friis-Healy EA, Nagy GA, Kollins SH. It is time to REACT: opportunities for digital mental health apps to reduce mental health disparities in racially and ethnically minoritized groups. JMIR Ment Health 2021 Jan 26;8(1):e25456 [FREE Full text] [doi: 10.2196/25456] [Medline: 33406050]
29. Advancing equity and racial justice through the federal government. The White House. Washington, DC: The White House URL: https://www.whitehouse.gov/equity/ [accessed 2024-04-23]
30. Stiles-Shields C, Cummings C, Montague E, Plevinsky JM, Psihogios AM, Williams KDA. A call to action: using and extending human-centered design methodologies to improve mental and behavioral health equity. Front Digit Health 2022;4:848052. [doi: 10.3389/fdgth.2022.848052] [Medline: 35547091]
31. Ramos G, Hernandez-Ramos R, Taylor M, Schueller SM. State of the science: using digital mental health interventions to extend the impact of psychological services. Behavior Therapy 2024 Apr:e1 (forthcoming). [doi: 10.1016/j.beth.2024.04.004]
32. Singh N, Lawrence K, Richardson S, Mann DM. Centering health equity in large language model deployment. PLOS Digit Health 2023 Oct;2(10):e0000367 [FREE Full text] [doi: 10.1371/journal.pdig.0000367] [Medline: 37874780]
33. Lawrence K. Digital health equity. In: Lawrence K, Linwood SL, editors. Digital Health. Brisbane, Queensland, Australia: Exon Publications; 2022.

XSL•FO
RenderX

34. Muchamore I, Karanikolas P, Gooding P. How lived experience expertise shapes research and development in digital mental health: a review of literature and insights. Wellcome Open Research. 2024. URL: https://wellcomeopenresearch.org/documents/9-373 [accessed 2024-07-19]
35. Folk JB, Carter SP, Jagsi R, Mangurian C. Title IX offices at top United States universities with medical centers: staffing, investigations and transparency. Office of Diversity and Outreach UCSF. San Francisco, CA: Office of Diversity and Outreach UCSF; 2023. URL: https://research.ucsf.edu/sites/g/files/tkssra2931/f/UCSF_Task_Force_Report_Final_Anti-Racism_in_Research_IDEA.pdf [accessed 2024-04-24]
36. Kirvin-Quamme A, Kissinger J, Quinlan L, Montgomery R, Chernenok M, Pirner MC, et al. Common practices for sociodemographic data reporting in digital mental health intervention research: a scoping review. BMJ Open 2024 Feb 12;14(2):e078029 [FREE Full text] [doi: 10.1136/bmjopen-2023-078029] [Medline: 38346876]
37. Lin JS, Webber EM, Bean SI, Evans CV. Development of a health equity framework for the US Preventive Services Task Force. JAMA Netw Open 2024 Mar 04;7(3):e241875 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.1875] [Medline: 38466305]
38. Wilkins D. The business case for health equity in life sciences. Deloitte. 2022. URL: https://www2.deloitte.com/us/en/blog/health-care-blog/2022/the-business-case-for-health-equity-in-life-sciences.html [accessed 2024-04-24]
39. Equitable healthcare starts with medical affairs. H1. 2022. URL: https://h1.co/blog/a-different-approach-and-a-wider-scope-equitable-healthcare-starts-with-medical-affairs/ [accessed 2024-04-24]
40. Baumann AA, Cabassa LJ. Reframing implementation science to address inequities in healthcare delivery. BMC Health Serv Res 2020 Mar 12;20(1):190 [FREE Full text] [doi: 10.1186/s12913-020-4975-3] [Medline: 32164706]
41. Hatef E, Austin M, Scholle S, Buckley B. Evidence- and consensus-based digital healthcare equity framework: a practical guide for implementation. Agency for Healthcare Research and Quality. Rockville, MD: Agency for Healthcare Research and Quality (US); 2024 Feb. URL: https://digital.ahrq.gov/ahrq-funded-projects/creating-digital-healthcare-equity-framework-accompanying-guide-its-use#nav-publications [accessed 2024-04-24]
42. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open 2021 Mar 19;11(3):e047001 [FREE Full text] [doi: 10.1136/bmjopen-2020-047001] [Medline: 33741674]
43. Neary M, Schueller SM. State of the field of mental health apps. Cogn Behav Pract 2018 Nov;25(4):531-537 [FREE Full text] [doi: 10.1016/j.cbpra.2018.01.002] [Medline: 33100810]
44. Silberman J, Wicks P, Patel S, Sarlati S, Park S, Korolev IO, Evidence DEFINED Workgroup, et al. Rigorous and rapid evidence assessment in digital health with the evidence DEFINED framework. NPJ Digit Med 2023 May 31;6(1):101 [FREE Full text] [doi: 10.1038/s41746-023-00836-5] [Medline: 37258851]
45. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021 Dec;147:103990 [FREE Full text] [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
46. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
47. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
48. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023 Jan;53(2):567-579. [doi: 10.1017/S0033291721001999] [Medline: 34047264]
49. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
50. Riadi I, Kervin L, Dhillon S, Teo K, Churchill R, Card KG, et al. Digital interventions for depression and anxiety in older adults: a systematic review of randomised controlled trials. Lancet Healthy Longev 2022 Aug;3(8):e558-e571 [FREE Full text] [doi: 10.1016/S2666-7568(22)00121-0] [Medline: 36102765]
51. Food and Drug Administration (FDA). Diversity action plans to improve enrollment of participants from underrepresented populations in clinical studies guidance for industry. FDA. Silver Spring, MD: Food and Drug Administration URL: https://www.fda.gov/media/157635/download [accessed 2024-04-23]
52. National Institute of Mental Health (NIMH). NIMH's approach to mental health disparities research. NIMH. Bethesda, MD: NIMH URL: https://www.nimh.nih.gov/about/organization/od/odwd/nimhs-approach-to-mental-health-disparities-research [accessed 2024-04-23]
53. Berg S. AMA unveils playbook to speed digital health adoption. American Medical Association. 2018. URL: https://www.ama-assn.org/practice-management/digital/ama-unveils-playbook-speed-digital-health-adoption [accessed 2024-04-23]
54. The app evaluation model. American Psychiatric Association. URL: https://www.psychiatry.org:443/psychiatrists/practice/mental-health-apps/the-app-evaluation-model [accessed 2024-04-23]

XSL•FO
RenderX

55. Bond RR, Mulvenna MD, Potts C, O'Neill S, Ennis E, Torous J. Digital transformation of mental health services. Npj Ment Health Res 2023 Aug 22;2(1):13 [FREE Full text] [doi: 10.1038/s44184-023-00033-y] [Medline: 38609479]
56. World Health Organization (WHO). Global strategy on digital health 2020-2025. WHO. Geneva, Switzerland: WHO; 2021. URL: https://iris.who.int/bitstream/handle/10665/344249/9789240020924-eng.pdf [accessed 2024-04-23]
57. What is human-centered design (HCD)? The Interaction Design Foundation. URL: https://www.interaction-design.org/literature/topics/human-centered-design [accessed 2024-04-23]
58. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun 07;9(6):e35591 [FREE Full text] [doi: 10.2196/35591] [Medline: 35671081]
59. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
60. Porche MV, Folk JB, Tolou-Shams M, Fortuna LR. Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. Front Psychiatry 2022;13:867460 [FREE Full text] [doi: 10.3389/fpsyt.2022.867460] [Medline: 35530032]
61. Figueroa CA, Murayama H, Amorim PC, White A, Quiterio A, Luo T, et al. Applying the digital health social justice guide. Front Digit Health 2022;4:807886. [doi: 10.3389/fdgth.2022.807886] [Medline: 35295620]
62. Unertl KM, Schaefbauer CL, Campbell TR, Senteio C, Siek KA, Bakken S, et al. Integrating community-based participatory research and informatics approaches to improve the engagement and health of underserved populations. J Am Med Inform Assoc 2016 Jan;23(1):60-73 [FREE Full text] [doi: 10.1093/jamia/ocv094] [Medline: 26228766]
63. Israel BA, Coombe CM, Cheezum RR, Schulz AJ, McGranaghan RJ, Lichtenstein R, et al. Community-based participatory research: a capacity-building approach for policy advocacy aimed at eliminating health disparities. Am J Public Health 2010 Nov;100(11):2094-2102. [doi: 10.2105/AJPH.2009.170506] [Medline: 20864728]
64. Chen E, Leos C, Kowitt SD, Moracco KE. Enhancing community-based participatory research through human-centered design strategies. Health Promot Pract 2020 Jan;21(1):37-48. [doi: 10.1177/1524839919850557] [Medline: 31131633]
65. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023 Nov;26(1):e300884 [FREE Full text] [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
66. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023 Oct 20;6(1):195 [FREE Full text] [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]
67. Timmons AC, Duong JB, Simo Fiallo N, Lee T, Vo HPQ, Ahle MW, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci 2023 Sep;18(5):1062-1096 [FREE Full text] [doi: 10.1177/17456916221134490] [Medline: 36490369]
68. Chua L, Ghazi B, Huang Y, Kamath P, Kumar R, Manurangsi P, et al. Crosslingual capabilities and knowledge barriers in multilingual large language models. arXiv :1 Preprint posted online on June 23, 2024. [doi: 10.48550/arXiv.2406.16135]
69. Shen L, Tan W, Chen S, Chen Y, Zhang J, Xu H, et al. The language barrier: dissecting safety challenges of LLMs in multilingual contexts. arXiv :1 Preprint posted online on January 23, 2024. [doi: 10.48550/arXiv.2401.13136]
70. Chen I, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 2019 Feb 01;21(2):E167-E179 [FREE Full text] [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]
71. Green BL, Murphy A, Robinson E. Accelerating health disparities research with artificial intelligence. Front Digit Health 2024;6:1330160 [FREE Full text] [doi: 10.3389/fdgth.2024.1330160] [Medline: 38322109]
72. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Foundational Principles of Ophthalmic Imaging/Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D.C., et al. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med 2023 Sep 12;6(1):170 [FREE Full text] [doi: 10.1038/s41746-023-00913-9] [Medline: 37700029]
73. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
74. Young S. Advancing governance, innovation, and risk management for agency use of artificial intelligence. Executive Office of the President Office of Management and Budget. 2024. URL: https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf [accessed 2024-04-23]
75. Nundy S, Cooper LA, Kelsay E. Employers can do more to advance health equity. Harvard Business Review. 2023. URL: https://hbr.org/2023/01/employers-can-do-more-to-advance-health-equity [accessed 2024-04-23]
76. The Deloitte Millennial Survey 2018. Deloitte. 2018. URL: https://www2.deloitte.com/tr/en/pages/about-deloitte/articles/millennialsurvey-2018.html [accessed 2024-04-23]
77. Eswaran V. The business case for diversity in the workplace is now overwhelming. World Economic Forum. 2019. URL: https://www.weforum.org/agenda/2019/04/business-case-for-diversity-in-the-workplace/ [accessed 2024-04-23]
78. Hawke LD, Sheikhan NY, Jones N, Slade M, Soklaridis S, Wells S, et al. Embedding lived experience into mental health academic research organizations: critical reflections. Health Expect 2022 Oct;25(5):2299-2305 [FREE Full text] [doi: 10.1111/hex.13586] [Medline: 35999670]

XSL•FO
RenderX

79. Michalak C, Jackson M. Supporting the well-being of your underrepresented employees. Harvard Business Review. 2022. URL: https://hbr.org/2022/03/supporting-the-well-being-of-your-underrepresented-employees [accessed 2024-04-23]
80. Woebot Health Forms Clinical Diversity Advisory Board. Woebot Health. 2021. URL: https://woebothealth.com/woebot-health-forms-clinical-diversity-advisory-board/ [accessed 2024-04-23]
81. Delogne MC. Woebot Health to present abstracts on diversity in digital mental health interventions at ABCT 2022. Woebot Health. 2024. URL: https://woebothealth.com/woebot-health-to-present-abstracts-on-diversity-in-digital-mental-health-interventions-at-abct-2022/ [accessed 2024-04-23]
82. Hemphill C. 8 building blocks of health equity strategy. Woebot Health. URL: https://woebothealth.com/8-building-blocks-of-health-equity-strategy/ [accessed 2024-04-23]
83. Making mental health a part of your DEIB strategy. Headspace. 2024. URL: https://get.headspace.com/guide-making-mental-health-a-part-of-your-deib-strategy [accessed 2024-04-23]
84. Hemphill C. What does racism cost our healthcare system? Woebot Health. URL: https://woebothealth.com/what-does-racism-cost-our-healthcare-system/ [accessed 2024-04-24]
85. NIMHD Research Diversity Supplement Application Guidelines. National Institute on Minority Health and Health Disparities. URL: https://www.nimhd.nih.gov/programs/extramural/training-career-dev/research-supplements/diversity-supplements.html [accessed 2024-04-24]
86. Welcome to NIGMS. National Institute of General Medical Sciences. URL: https://nigms.nih.gov/ [accessed 2024-04-24]
87. Empower minds. Create opportunities. REACH. URL: https://reachatrush.org/ [accessed 2024-04-24]
88. Ramos G, Chavira DA. Use of technology to provide mental health care for racial and ethnic minorities: evidence, promise, and challenges. Cognitive and Behavioral Practice 2022 Feb;29(1):15-40. [doi: 10.1016/j.cbpra.2019.10.004]
89. Bullard L, Rapoport S, Ariniello L, Delgado F, Waalen J, Williams A, et al. Utilization of community-based participatory research strategies in a postpartum digital health trial. 2024 Presented at: Annual Postpartum Support International Conference; July 26-28, 2024; Washington, DC.
90. Joseph G, Dohan D. Diversity of participants in clinical trials in an academic medical center: the role of the 'Good Study Patient?'. Cancer 2009 Feb 01;115(3):608-615 [FREE Full text] [doi: 10.1002/cncr.24028] [Medline: 19127544]
91. Wendler D, Kington R, Madans J, Van Wye G, Christ-Schmidt H, Pratt LA, et al. Are racial and ethnic minorities less willing to participate in health research? PLoS Med 2006 Feb;3(2):e19 [FREE Full text] [doi: 10.1371/journal.pmed.0030019] [Medline: 16318411]
92. Cullen MR, Lemeshow AR, Amaro S, Bandera EV, Cooper LA, Kawachi I, et al. A framework for setting enrollment goals to ensure participant diversity in sponsored clinical trials in the United States. Contemp Clin Trials 2023 Jun;129:107184 [FREE Full text] [doi: 10.1016/j.cct.2023.107184] [Medline: 37054773]
93. Call CC, Eckstrand KL, Kasparek SW, Boness CL, Blatt L, Jamal-Orozco N, Scholars for Elevating EquityDiversity (SEED). An ethics and social-justice approach to collecting and using demographic data for psychological researchers. Perspect Psychol Sci 2023 Sep;18(5):979-995 [FREE Full text] [doi: 10.1177/17456916221137350] [Medline: 36459692]
94. Boucher E, Honomichl R, Ward H, Powell T, Stoeckl SE, Parks A. The effects of a digital well-being intervention on older adults: retrospective analysis of real-world user data. JMIR Aging 2022 Sep 02;5(3):e39851 [FREE Full text] [doi: 10.2196/39851] [Medline: 36053569]
95. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
96. Placzek M, Friede T. Clinical trials with nested subgroups: analysis, sample size determination and internal pilot studies. Stat Methods Med Res 2018 Nov;27(11):3286-3303. [doi: 10.1177/0962280217696116] [Medline: 29298604]
97. McCall T, Asuzu K, Oladele CR, Leung TI, Wang KH. A socio-ecological approach to addressing digital redlining in the United States: a call to action for health equity. Front Digit Health 2022;4:897250 [FREE Full text] [doi: 10.3389/fdgth.2022.897250] [Medline: 35924138]
98. Egede LE, Ruiz R, Mosley-Johnson E, Aguilar-Gaxiola SA, Corbie GM, Wilkins CH, et al. Laying the groundwork to make diversity, equity, and inclusion front and center in clinical and translational research. J Clin Transl Sci 2023;7(1):e95 [FREE Full text] [doi: 10.1017/cts.2023.524] [Medline: 37125065]
99. Martinez-Martin N. Envisioning a path toward equitable and effective digital mental health. AJOB Neurosci 2022;13(3):196-198 [FREE Full text] [doi: 10.1080/21507740.2022.2082597] [Medline: 35797130]
100. Taher R, Hsu C, Hampshire C, Fialho C, Heaysman C, Stahl D, et al. The safety of digital mental health interventions: systematic review and recommendations. JMIR Ment Health 2023 Oct 09;10:e47433 [FREE Full text] [doi: 10.2196/47433] [Medline: 37812471]
101. Health equity. American Psychological Association. URL: https://www.apaservices.org/advocacy/issues/health-equity [accessed 2024-04-24]
102. Behavioral health equity. SAMHSA - Substance Abuse and Mental Health Services. URL: https://www.samhsa.gov/behavioral-health-equity [accessed 2024-04-24]
103. Advocate for change. NAMI. URL: https://www.nami.org/advocacy/advocate-for-change/ [accessed 2024-04-24]
104. Advocacy network. Mental Health America. URL: https://mhanational.org/issues/advocacy-network [accessed 2024-04-24]

XSL•FO
RenderX

105. The mental health parity and addiction equity act (MHPAEA). Centers for Medicare & Medicaid Services. 2013. URL: https://www.cms.gov/marketplace/private-health-insurance/mental-health-parity-addiction-equity [accessed 2024-04-24]
106. Cohen ES, editor. Broadband internet: access, regulation, and policy. New York, NY: Nova Science Publishers; 2008.
107. Wu T. Network neutrality, broadband discrimination. SSRN Journal 2003;2:141. [doi: 10.2139/ssrn.388863]
108. Rolan G, Denison T, Mackenzie C. "Broadband, digital literacy and public libraries: the Mill Park story". Library Hi Tech News 2015;6:15-18. [doi: 10.1108/lhtn-03-2015-0020]
109. Clinical trial diversity. FDA. URL: https://www.fda.gov/consumers/minority-health-and-health-equity/clinical-trial-diversity [accessed 2024-04-24]
110. Spinner J, Araojo RR. FDA's strategies to close the health equity gap among diverse populations. J Prim Care Community Health 2021;12:21501327211000232 [FREE Full text] [doi: 10.1177/21501327211000232] [Medline: 33733919]
111. Grabman G, Tenenbaum C. FDA regulation must uphold women's health. Food & Drug Law Institute (FDLI). 2022. URL: https://www.fdli.org/wp-content/uploads/2022/12/Grabman-Tenenbaum-FDLJ-77-3.pdf [accessed 2024-08-30]
112. Food and Drug Administration (FDA). Enforcement policy for digital health devices for treating psychiatric disorders during the Coronavirus Disease 2019 (COVID-19): public health emergency guidance for industry and Food and Drug Administration staff. FDA. 2020. URL: https://www.regulations.gov/document/FDA-2020-D-1138-0068 [accessed 2024-04-24]
113. PAR-22-154: Laboratories to optimize digital health. National Institutes of Mental Health. 2023. URL: https://grants.nih.gov/grants/guide/pa-files/PAR-22-154.html [accessed 2024-04-24]
114. Kessler RC, Wang PS. The descriptive epidemiology of commonly occurring mental disorders in the United States. Annu Rev Public Health 2008;29:115-129. [doi: 10.1146/annurev.publhealth.29.020907.090847] [Medline: 18348707]
115. Hossain MM, Tasnim S, Sultana A, Faizah F, Mazumder H, Zou L, et al. Epidemiology of mental health problems in COVID-19: a review. F1000Res 2020;9:636 [FREE Full text] [doi: 10.12688/f1000research.24457.1] [Medline: 33093946]
116. Ohlsson H, Kendler KS. Applying causal inference methods in psychiatric epidemiology: a review. JAMA Psychiatry 2020 Jun 01;77(6):637-644 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.3758] [Medline: 31825494]
117. Meet New Profit's 2023 mental health equity catalyze cohort. New Profit. 2023 May 22. URL: https://www.newprofit.org/go/meet-new-profits-2023-mental-health-equity-catalyze-cohort/ [accessed 2024-04-24]
118. Galea S. The price of health equity. JAMA Health Forum 2021 Apr 01;2(4):e210720 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.0720] [Medline: 36218807]
119. Knifton L, Inglis G. Poverty and mental health: policy, practice and research implications. BJPsych Bull 2020 Oct;44(5):193-196 [FREE Full text] [doi: 10.1192/bjb.2020.78] [Medline: 32744210]
120. Soliman W. ACMA. 2021. URL: https://medicalaffairsspecialist.org/blog/diversity-inclusion-clinical-trials-medical-affairs [accessed 2024-04-24]
121. DeSalvo KB, Wang YC, Harris A, Auerbach J, Koo D, O'Carroll P. Public Health 3.0: a call to action for public health to meet the challenges of the 21st century. Prev Chronic Dis 2017 Sep 07;14:E78 [FREE Full text] [doi: 10.5888/pcd14.170017] [Medline: 28880837]
122. Dwyer M. Nonprofit leaders see potential to improve health equity through cross-sector partnerships. Pew. 2022. URL: https://pew.org/3zMdVT3 [accessed 2024-04-24]
123. Bauer MS, Damschroder L, Hagedorn H, Smith J, Kilbourne AM. An introduction to implementation science for the non-specialist. BMC Psychol 2015 Sep 16;3(1):32 [FREE Full text] [doi: 10.1186/s40359-015-0089-9] [Medline: 26376626]
124. Efforts to close the digital health equity divide may improve access to healthcare. ElevanceHealth. 2023. URL: https://www.elevancehealth.com/our-approach-to-health/health-equity/efforts-to-close-the-digital-health-equity-divide [accessed 2024-04-24]
125. Mordecai D, Histon T, Neuwirth E, Heisler WS, Kraft A, Bang Y, et al. How Kaiser Permanente Created a Mental Health and Wellness Digital Ecosystem. NEJM Catalyst 2021 Jan;2(1) [FREE Full text] [doi: 10.1056/cat.20.0295]
126. Youn SJ, Jaso B, Eyllon M, Sah P, Hoyler G, Barnes JB, et al. Leveraging implementation science to integrate digital mental health interventions as part of routine care in a practice research network. Adm Policy Ment Health 2024 May;51(3):348-357. [doi: 10.1007/s10488-023-01292-9] [Medline: 37615809]
127. CMS framework for health equity. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/priorities/health-equity/minority-health/equity-programs/framework [accessed 2024-04-24]
128. Chen K, Lane E, Burns J, Macrynikola N, Chang S, Torous J. The digital navigator: standardizing human technology support in app-integrated clinical care. Telemed J E Health 2024 Jun;30(7):e1963-e1970. [doi: 10.1089/tmj.2024.0023] [Medline: 38574251]
129. Nicholas J, Ringland KE, Graham AK, Knapp AA, Lattie EG, Kwasny MJ, et al. Stepping up: predictors of 'stepping' within an iCBT stepped-care intervention for depression. Int J Environ Res Public Health 2019 Nov 25;16(23):4689 [FREE Full text] [doi: 10.3390/ijerph16234689] [Medline: 31775297]
130. Nordgreen T, Haug T, Öst LG, Andersson G, Carlbring P, Kvale G, et al. Stepped care versus direct face-to-face cognitive behavior therapy for social anxiety disorder and panic disorder: a randomized effectiveness trial. Behav Ther 2016 Mar;47(2):166-183. [doi: 10.1016/j.beth.2015.10.004] [Medline: 26956650]

XSL•FO
RenderX

131. Adewuya AO, Momodu O, Olibamoyo O, Adegbaju A, Adesoji O, Adegbokun A. The effectiveness and acceptability of mobile telephone adherence support for management of depression in the Mental Health in Primary Care (MeHPriC) project, Lagos, Nigeria: A pilot cluster randomised controlled trial. J Affect Disord 2019 Jun 15;253:118-125. [doi: 10.1016/j.jad.2019.04.025] [Medline: 31035212]
132. Jonassaint CR, Belnap BH, Huang Y, Karp JF, Abebe KZ, Rollman BL. Racial differences in the effectiveness of internet-delivered mental health care. J Gen Intern Med 2020 Feb;35(2):490-497 [FREE Full text] [doi: 10.1007/s11606-019-05542-1] [Medline: 31745855]

Abbreviations

AI: artificial intelligence
APA: American Psychological Association
DEIB: diversity, equity, inclusion, and belonging
DMHI: digital mental health intervention
FDA: Food and Drug Administration
GDPR: General Data Protection Regulation
LGBTQIA+: Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Asexual+
LLM: large language model
ML: machine learning
NIH: National Institutes of Health
NIMH: National Institute of Mental Health
PIDAR: Partner, Identify, Demonstrate, Access, Report
REACH: Rush Education and Career Center Hub
RE-AIM: Reach, Effectiveness, Adoption, Implementation, Maintenance
SAMHSA: Substance Abuse and Mental Health Services Administration
SDOH: social determinants of health
STEM: Science, Technology, Engineering, Mathematics
USPSTF: US Preventive Services Taskforce

Edited by G Eysenbach; submitted 30.04.24; peer-reviewed by S Miller, R Bhattacharya, D Ellis, J Murphy; comments to author 21.06.24; revised version received 07.08.24; accepted 10.08.24; published 24.09.24.

Please cite as:
Robinson A, Flom M, Forman-Hoffman VL, Histon T, Levy M, Darcy A, Ajayi T, Mohr DC, Wicks P, Greene C, Montgomery RM
Equity in Digital Mental Health Interventions in the United States: Where to Next?
J Med Internet Res 2024;26:e59939
URL: https://www.jmir.org/2024/1/e59939
doi:10.2196/59939
PMID:39316436

©Athena Robinson, Megan Flom, Valerie L Forman-Hoffman, Trina Histon, Monique Levy, Alison Darcy, Toluwalase Ajayi, David C Mohr, Paul Wicks, Carolyn Greene, Robert M Montgomery. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Evolutionary Trends in the Adoption, Adaptation, and Abandonment of Mobile Health Technologies: Viewpoint Based on 25 Years of Research

Jennifer Portz[1,2], MSW, PhD; Susan Moore[2], MSPH, PhD; Sheana Bull[2], MPH, PhD

[1]Division of General Internal Medicine, School of Medicine, University of Colorado, Aurora, CO, United States
[2]mHealth Impact Lab, Colorado School of Public Health, University of Colorado, Aurora, CO, United States

Corresponding Author:
Jennifer Portz, MSW, PhD
Division of General Internal Medicine
School of Medicine
University of Colorado
Mailstop B119
Aurora, CO, 80045
United States
Phone: 1 303 724 8856
Email: jennifer.portz@CUAnschutz.edu

Abstract

Over the past quarter-century, mobile health (mHealth) technologies have experienced significant changes in adoption rates, adaptation strategies, and instances of abandonment. Understanding the underlying factors driving these trends is essential for optimizing the design, implementation, and sustainability of interventions using these technologies. The evolution of mHealth adoption has followed a progressive trajectory, starting with cautious exploration and later accelerating due to technological advancements, increased smartphone penetration, and growing acceptance of digital health solutions by both health care providers and patients. However, alongside widespread adoption, challenges related to usability, interoperability, privacy concerns, and socioeconomic disparities have emerged, necessitating ongoing adaptation efforts. While many mHealth initiatives have successfully adapted to address these challenges, technology abandonment remains common, often due to unsustainable business models, inadequate user engagement, and insufficient evidence of effectiveness. This paper utilizes the Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework to examine the interplay between the academic and industry sectors in patterns of adoption, adaptation, and abandonment, using 3 major mHealth innovations as examples: health-related SMS text messaging, mobile apps and wearables, and social media for health communication. Health SMS text messaging has demonstrated significant potential as a tool for health promotion, disease management, and patient engagement. The proliferation of mobile apps and devices has facilitated a shift from in-person and in-clinic practices to mobile- and wearable-centric solutions, encompassing everything from simple activity trackers to advanced health monitoring devices. Social media, initially characterized by basic text-based interactions in chat rooms and online forums, underwent a paradigm shift with the emergence of platforms such as MySpace and Facebook. This transition ushered in an era of mass communication through social media. The rise of microblogging and visually focused platforms such as Twitter(now X), Instagram, Snapchat, and TikTok, along with the integration of live streaming and augmented reality features, exemplifies the ongoing innovation within the social media landscape. Over the past 25 years, there have been remarkable strides in the adoption and adaptation of mHealth technologies, driven by technological innovation and a growing recognition of their potential to revolutionize health care delivery. Each mobile technology uniquely enhances public health and health care by catering to different user needs. SMS text messaging offers wide accessibility and proven effectiveness, while mobile apps and wearables provide comprehensive functionalities for more in-depth health management. Social media platforms amplify these efforts with their vast reach and community-building potential, making it essential to select the right tool for specific health interventions to maximize impact and engagement. Nevertheless, continued efforts are needed to address persistent challenges and mitigate instances of abandonment, ensuring that mHealth interventions reach their full potential in improving health outcomes and advancing equitable access to care.

(J Med Internet Res 2024;26:e62790) doi:10.2196/62790

XSL•FO
RenderX

KEYWORDS

technology adoption; mobile health; SMS text messaging; mobile apps; wearables; social media

Introduction

mHealth, short for mobile health, represents a transformative approach to health care delivery by harnessing the power of mobile devices such as smartphones, tablets, and wearables to enhance communication, collect health data, and monitor well-being [1]. This technology has the potential to revolutionize health care services by offering personalized care, improving access, reducing costs, and empowering patients in managing their health [2]. With the proliferation of smartphones, even among traditionally underserved populations, and the widespread availability of internet access, the stage is set for the broad adoption of mHealth solutions [3,4]. However, the full potential of mHealth has yet to be realized [5].

At the heart of mHealth's potential is the theory of diffusion of innovations, which explains how new ideas, technologies, and practices spread within a society or organization [6]. In the context of mHealth, this diffusion process involves the adoption of mHealth technologies by various stakeholders, including health care providers, patients, and policy makers. Several factors influence the diffusion of mHealth innovations. First, the perceived benefits of these technologies play a crucial role. Innovations that promise to improve health care access, efficiency, and outcomes are more likely to be embraced by users. Second, the ease of use and accessibility of mHealth solutions are critical determinants of their adoption. Particularly among marginalized populations, including racial and ethnic minorities, individuals with low socioeconomic status, older adults, and those with disabilities, ensuring that mHealth tools are user-friendly and accessible is essential for equitable adoption.

Despite the promise of mHealth, challenges remain in achieving widespread adoption. Concerns regarding usability, privacy, data security, and the potential replacement of in-person care with virtual interactions have led to hesitancy among both patients and providers [7]. Additionally, interoperability issues between mHealth apps and electronic health record systems pose challenges for seamless integration into clinical workflows, potentially increasing provider burden [8]. Furthermore, the complexity of reimbursement and regulatory mechanisms presents obstacles to the large-scale implementation of mHealth solutions [9]. Without clear guidelines on reimbursement for mHealth services and adherence to regulatory requirements, health care organizations may hesitate to invest in these technologies. While mHealth holds tremendous promise for revolutionizing health care delivery, its widespread adoption depends on addressing implementation challenges and ensuring that innovations are accessible, user-friendly, and aligned with regulatory and reimbursement frameworks. By leveraging the principles of diffusion of innovations and addressing barriers to adoption, mHealth has the potential to transform health care and public health, improving outcomes for individuals, families, and communities.

In this viewpoint, we use the innovation-diffusion Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework [10] to examine the dynamic relationship between the academic and industry sectors concerning adoption, adaptation, and abandonment patterns of 3 prominent mHealth innovations: health-related SMS text messaging, mobile apps and wearables for chronic disease management, and the use of social media for health communication. By utilizing these innovations as case studies, the paper will explore their potential in health promotion, disease management, and patient engagement. This viewpoint aims to provide insights into the ongoing trends in innovation and adoption within the mHealth landscape.

Innovation and Diffusion

To consolidate various innovation diffusion theories, the NASSS framework emerged. It aims to guide the development of novel technologies that are conducive to widespread scalability while also identifying emerging technologies with limited potential [10]. NASSS represents a comprehensive, health care–specific paradigm that underscores the interplay among individual factors (micro), organizational dynamics (meso), and overarching policy frameworks (macro) in facilitating or hindering technology adoption. By harnessing the analytical power of NASSS, insights into technological shortcomings can be gathered, enhancing preparedness for implementation. NASSS notably delineates several key domains: (1) the nature of the health condition(s); (2) the technological features and types of data collected; (3) the value proposition for patients, industry, and clinical stakeholders; (4) the composite adopter system, including health care professionals, patients, and informal caregivers; (5) organizational structures such as capacity and readiness to adopt; (6) broader institutional, societal, and cultural contexts; and (7) the dynamic interactions and reciprocal adaptations among these domains over time. While barriers to adoption and adaptation are categorized as simple, complicated, or complex, technology solutions addressing complex challenges that span multiple domains rarely achieve mainstream acceptance. A summary of our viewpoint is provided in Table 1.

XSL•FO
RenderX

Table 1. Viewpoint summary by NASSS[a] dimension.

NASSS dimension	SMS text messaging	Mobile health and wearables	Social media
Condition	• Effectiveness in promoting healthy behaviors, smoking cessation, weight management, and chronic condition self-management. • SMS text messaging was invented in 1984; widely used by 1999; popular by 2000; and now an average of 32 texts/day. • Limitations in addressing cultural differences and sociocultural context.	• Mobile apps offer health tracking, personalized coaching, educational resources, behavioral interventions, and remote monitoring. Wearables monitor metrics, such as physical activity, sleep, heart rate, and nutrition, integrated into mobile apps. • The early 2000s saw a rise in health-related mobile apps with basic tracking functions. Now over 350,000 health apps are available. • Limitations in addressing cultural differences and sociocultural context.	• Technology for communication and information sharing among virtual communities. • Early systems such as bulletin board systems, Usenet[b], and Internet Relay Chat (1980s); web-based communities such as webrings and GeoCities (1990s); first social media platform: Six Degrees (1997). Now over 5 billion users worldwide, representing 62.3% of the global population. • Limitations in addressing cultural differences and sociocultural context.
Technology	• SMS text messaging is highly accessible, easy to implement, and scalable in health programs. Automated systems can be developed in-house or through third-party vendors. • Difficulty in verifying SMS text message comprehension and engagement.	• Mobile apps require diverse technologies and skills, developed for operating systems such as iOS and Android. • Security measures are crucial for protecting user data, and regular updates and maintenance are vital for app reliability and user experience.	• Varies in communication modalities (text, image, audio, and video). • Central to social media; includes replies, shares, likes, and other engagement types, often permission-based.
Value proposition	• Cost-effective promotion: SMS text messaging programs are low-cost. Health care systems can reach thousands with minimal cost per SMS text message. • Effectiveness: Texting interventions can be effective, but impact is typically small to medium; scaling is necessary for significant impact.	• User empowerment: Apps and wearables offer personalized tools for health management, fostering autonomy, self-efficacy, and empowerment. • Bidirectional engagement: Personalization, gamification, and social support enhance user engagement and motivation.	• Extensive reach: Ability to tailor messages to specific communities and transmit information rapidly. • Cost-effective promotion: Low-cost messaging with potential for viral spread. • Data analysis: Identifies emerging health concerns, traces outbreaks, and tracks disease spread.
Adopters	• Health care providers: Providers endorse SMS text messaging but are reluctant to engage directly due to concerns about boundaries and being overwhelmed with SMS text messages. Automated systems allow delegation, limiting direct provider-patient communication. • Challenges: Older adults and the visually impaired engage less with texting. Limited use of literacy-sensitive and multilingual SMS text messaging. Limited evidence of SMS text messaging's impact on caregivers.	• Widespread use: Mobile apps and wearables are used across diverse populations, but adoption varies due to clinical, demographic, and technological factors. • Challenges: Challenges include technology literacy, cost, security concerns, accessibility for older adults and low-income communities, and limited internet connectivity in rural areas.	• Widespread use: Nearly ubiquitous among technology users; billions on platforms such as Facebook. • Challenges: Adoption varies by platform and demographics; health campaigns should tailor messages to specific populations.
Organization	• Organizations use SMS text messaging for appointment reminders and medication refills. Regulatory barriers limit the widespread use of prevention and self-management programs. • Barriers to making texting a standard of care include resource constraints, motivation, and regulatory issues.	• Facilitated by seamless integration into workflows. Engagement with user feedback promotes effective adoption in clinical care settings. • Challenges including organizational resistance, limited resources, data privacy concerns, and interoperability issues hinder adoption.	• Engagement is affected by policies on who can post and what content is shareable. • Involves content experts and communications experts; slow approval processes can hinder timely responses. • Requires understanding of cultural contexts and platform-specific norms for successful strategies.

XSL•FO
RenderX

NASSS dimension	SMS text messaging	Mobile health and wearables	Social media
Wider system	• Regulatory challenges, security and consent, and potential for sociocultural tailoring of SMS text messages, yet set up is costly and complex.	• Regulatory challenges, interoperability with existing systems, digital divide, and resistance from health care providers.	• Approachability and trust of social media by health care systems; navigating health policies alongside social media policies is complex; and the importance of maintaining accountability, combating misinformation, and addressing harassment in social media.
Embedding and adaptation over time	• Interest in AI[c]-enabled chatbots as a sophisticated iteration of SMS text messaging. AI chatbots offer interactive and personalized communication but face similar implementation challenges as current texting programs.	• Requires collaboration, resource allocation, long-term funding, and viable business models to expand initiatives and reach broader populations effectively.	• Health communication strategies must consider potential negative impacts to avoid unintended consequences.

[a]NASSS: Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability.

[b]Usenet: User's Network.

[c]AI: artificial intelligence.

SMS Text Messaging

Condition

SMS text messaging can be used to promote healthy behaviors and address various conditions and comorbidities in a culturally responsive manner. Although SMS text messaging was invented in 1984, it was not widely adopted until 1999, when users began texting across different phone carrier networks. By 2000, SMS text messaging had become popular, with SMS text message volumes averaging around 35 texts per person per month in the United States [11]. The earliest evidence of SMS text messaging's efficacy for health emerged in the first decade of the 21st century, with publications demonstrating its potential impact on self-management of type 1 diabetes, drinking behaviors, and medical appointment adherence [12-14]. The popularity of SMS text messaging in daily life has continued to accelerate, with people now receiving an average of 32 texts per day (range 16-128 texts per day). A lower number of SMS text messages is typically sent among older individuals, while younger people tend to receive higher volumes [15]. With this increased popularity, there has been a concomitant rise in studies exploring how SMS text messages can impact health behaviors, prevention, and the self-management of both chronic and infectious conditions. We now have a robust body of literature, including meta-analyses, demonstrating that SMS text messaging can be effective in facilitating healthy behaviors [16-22]. Numerous meta-analyses have focused on smoking cessation, weight management, medication and appointment adherence [23-28], and self-management of chronic conditions [29-32]. While this literature is extensive, researchers have consistently acknowledged limitations in using SMS text messaging to create robust programs that incorporate cultural differences and the sociocultural context of patients and populations. However, there are notable exceptions that demonstrate the feasibility of integrating cultural context into SMS text messaging programs. Examples include a smoking cessation program for Māori in New Zealand [33] and another focused on promoting healthy behaviors in low- and middle-income country settings [34].

Technology

SMS text messaging is a highly accessible and easy-to-implement technology, which enhances its potential for scalability and widespread dissemination in health care. It does not require complex technological architecture. Automated programs can be developed in-house or utilized through third-party vendors contracted with health care delivery systems. However, it can be challenging to obtain clear evidence that people are reading and understanding SMS text messages. Automated systems can document that SMS text messages are sent and received, but they cannot confirm that they have been read or that people comprehend them [35]. This is particularly true for unidirectional SMS text messages, where people cannot respond to an SMS text message or their responses are not cataloged. Given that smartphone ownership with texting capability is nearly universal in the United States [36], and texting behaviors are widespread among smartphone users [15], there is no additional knowledge required for people to access and engage with SMS text messages. However, we note that there are barriers for some people in accessing and using SMS text messaging, which we will discuss further below.

Value Proposition

The value proposition of SMS text messaging can only be realized when programs are scaled. Developers stand to gain significantly from building SMS text messaging programs due to the relative simplicity of coding such systems, which results in relatively low production costs. However, this low cost also means that a return on their development investment will be achieved only when thousands of users access and utilize these systems [37]. Health care delivery systems also face very low costs in implementing texting programs; SMS text messages are sent and received for just a few cents and can easily reach thousands or even tens of thousands of patients through texting campaigns. While SMS text messaging interventions have been shown to be effective, their effects are generally small and occasionally moderate [38]. Texting campaigns can only be impactful when scaled to thousands of users. The widespread use and growing popularity of texting as a daily communication

XSL•FO
RenderX

medium mean that it may become increasingly difficult for health care delivery systems to send SMS text messages that resonate or compete for attention against the dozens, or even hundreds, of SMS text messages individuals receive each day [15].

Adopters

While people have adopted SMS text messaging with enthusiasm, challenges with reach and engagement remain. Although health care providers may support the use of SMS text messaging from their clinics and care delivery organizations to patients, they often show reluctance or resistance to directly engage with patients via SMS text messages. Providers cite concerns such as being inundated with SMS text messages from patients and having professional boundaries violated, such as being expected to respond to SMS text messages at night and on weekends [39]. The widespread use of automated systems to send SMS text messages to patients across entire health care systems allows providers to delegate texting tasks to other staff, keeping their involvement limited to a more tangential role. While this approach may effectively manage patient-provider communication, it also means that providers might not see the SMS text messages patients are sending, potentially missing opportunities to better understand the daily challenges patients face with prevention and self-management behaviors.

While texting is highly popular and accessible at the population level, some groups, such as older adults and individuals with visual impairments, engage less with SMS text messages due to accessibility concerns, such as font size or limited access to text-to-speech features [40]. Ongoing critiques of SMS text messaging literacy suggest that greater accessibility could be achieved by paying more attention to literacy levels [41]. Additionally, there is a lack of widespread use of multilingual SMS text messaging, which could improve accessibility for non-English speakers [42]. Much of the literature on SMS text messaging in health has focused on the impacts of interventions at the individual level. There is less evidence showing the impact on caregivers [43], which represents a missed opportunity given that caregivers could help older adults or individuals with impairments access and improve care with support from SMS text messages.

Organization

While organizations have adopted SMS text messaging for certain tasks, there is a lack of consistent adoption and implementation of prevention and self-management programs across patient lifespans. There are multiple examples of organizations using SMS text messaging to remind patients to schedule and keep appointments and to pick up or refill prescribed medications [23]. While the capacity to innovate may have driven earlier adoption of SMS text messaging systems, organizations still face challenges in widespread use for diverse prevention and self-management programs due to regulatory barriers, which we discuss below. The ability for organizations to develop texting programs in-house may be less relevant given the widespread availability of texting platforms from third-party vendors. Organizations can make texting campaigns attractive and valuable when they focus on key objectives, such as increasing childhood vaccine coverage [44] or screening for colorectal cancer [45]. However, the widespread institutionalization of systems integrating texting for primary prevention, health behavior, and self-management as a standard of care remains elusive. It is unclear whether this is due to resource constraints, motivation, willingness to change, or other unknown factors.

Wider System

Understanding the context of texting within the framework of policies and regulations is crucial. While individual organizations and health care delivery systems might be able to implement texting as a standard of care for specific use cases (eg, to encourage flu vaccine uptake), several regulatory requirements limit its widespread institutionalization as envisioned. The Health Insurance Portability and Accountability Act (HIPAA) requirements for confidentiality and privacy restrict what can be shared in outbound patient messages and necessitate security measures that may complicate access via cell phones. Additionally, the security protocols in electronic health records have created firewalls that make direct texting to patients difficult without complex integrations with patient portals. Organizational adherence to HIPAA is influenced by a recent ruling from the Federal Communications Commission, which requires organizations sending SMS text messages to obtain explicit permission from individuals. This adds an additional layer of effort for organizations to enroll patients in SMS text messaging programs and may bias program participation by excluding those who cannot be reached or choose not to opt-in [46]. While automated texting has the potential to enhance attention to sociocultural variability by allowing individual tailoring to patients, setting up the necessary infrastructure and programming can be daunting and initially costly for health care delivery systems. We will elaborate on this topic in the following paragraphs.

Embedding and Adaptation Over Time

Embedding and adapting texting over time requires aligning incentives for health care organizations. As discussed in the preceding paragraphs, there is substantial evidence that SMS text messaging can improve health behaviors and outcomes. Given that this evidence has been available for nearly 2 decades, what prevents texting from being embedded in health care delivery systems as a standard of care to support healthy behaviors and self-management across the lifespan? The regulatory challenges mentioned above, the complexity of integrating texting with electronic health records, and concerns about overwhelming patients and providers with texting fatigue are key reasons for the lack of universal texting integrations despite their potential. Additionally, there is insufficient alignment of incentives for organizations to use texting alongside other interventions. Numerous incentives within value-based care frameworks can be enhanced through texting interventions, such as improving colorectal, breast, and cervical cancer screening; screening for diabetes, hypertension, and cholesterol; and increasing flu, COVID-19, and childhood vaccine coverage.

While our assessment here has focused specifically on texting, it is important to note the growing interest in exploring the potential of more advanced SMS text messaging technologies,

XSL•FO
RenderX

such as artificially intelligent (AI) chatbots [47-50]. AI-enabled chatbots advance automated communication beyond unidirectional texts and bidirectional texts with fixed response options (eg, text "1" if you need information on smoking cessation). Unlike these methods, AI-enabled chatbots allow users to initiate and direct their questions, providing a more personalized interaction. These systems interpret and classify queries using advanced language models—large neural networks trained on extensive text corpora and fine-tuned for accurate classification in domain-specific text processing. They can engage in conversations with patients on topics of their choice. However, embedding these advanced systems into health care delivery systems will face the same challenges as current texting programs.

Mobile Apps and Wearables

Condition

Mobile apps and wearables offer versatile functionalities for promoting healthy behavior and managing various health conditions and comorbidities. The earliest health-related mobile apps emerged in the early 2000s with the advent of smartphones and the integration of basic health features [51]. These early apps primarily focused on basic health tracking, such as recording physical activity, monitoring nutrition intake, and tracking vital signs (eg, heart rate and blood pressure) [52]. Modern mobile apps typically offer a broader range of functionalities beyond SMS text messaging. They often include features for health tracking, personalized coaching, access to educational resources, behavioral interventions, and remote health monitoring, among others. Combined with wearables equipped with sensor technology, these tools enable comprehensive health tracking and monitoring of metrics such as physical activity, sleep patterns, heart rate, and nutrition intake. This empowers individuals to gain insights into their health status and make informed decisions [53,54]. There are over 350,000 health-related mobile apps available, with approximately 250 new apps released daily, accounting for 47% of all apps [55,56]. Health-related app use varies by individual characteristics, but an estimated 70% of adults in the United States have used an app to track their health [57]. Early wearables, such as pedometers and heart rate monitors, primarily targeted individuals interested in exercise and fitness. By the mid-2010s, the rise of smartwatches and fitness trackers led to the integration of wearable data directly into mobile apps [58,59].

The scientific evaluation of effectiveness and safety has not kept pace with the rapid expansion of the commercial health-related app market, creating a disconnect between scientific literature and available apps [50,60]. Mobile apps developed and tested in clinical settings are rarely widely available in app stores, while commercially developed apps often lack high-quality effectiveness testing [61-63]. Despite the growing interest and proliferation of apps across popular platforms, as well as increased research focus on their evaluation, their influence on outcomes related to chronic diseases remains uncertain [64,65]. Systematic reviews and meta-analyses have produced varied findings regarding the impact of apps on clinical outcomes, self-management, and behavior change. This inconsistency raises concerns about the ability of scientific research to keep pace with rapid technological advancements. Many trials suffer from biases, such as high attrition rates and incomplete outcome data, and the effects on health outcomes and participant engagement often decline over time in longer-duration studies [66]. Existing evidence suggests that while mobile apps might offer some benefit compared with having no intervention at all, there is limited high-quality evidence demonstrating their effectiveness in enhancing health outcomes. This contrasts with lower-tech alternatives, such as SMS text messaging, which has a more robust evidence base. Similarly, wearables are increasingly used for various conditions and health topics [67]. They are reliable for measuring steps and heart rate but may be less accurate for capturing other sensor data, such as energy expenditure [68]. The impact of wearables on health outcomes such as exercise, weight, and biomarkers in populations with chronic conditions is also mixed [69].

Technology

Developing a mobile app requires a diverse set of technologies, tools, and skills to create a functional, secure, and user-friendly app that meets the needs and expectations of its target audience [70]. Mobile apps are built on operating systems such as iOS by Apple Inc. and Android by Google (Alphabet Inc.). While these systems have similarities, they require different development kits and tools. Developers choose between them based on factors such as target audience, market share, and monetization models, which can impact an app's success. There are 4 major development approaches: native, cross-platform native, hybrid, and progressive web apps, each with its own advantages and disadvantages [71]. Native apps are platform-specific and written in languages provided by the platform. Cross-platform apps, by contrast, compile into native code. Hybrid apps use web technologies bundled into app installation packages, while progressive web apps offer an app-like experience through browsers. Developers may prefer hybrid or cross-platform approaches due to the specialized skill set required for native development. These approaches utilize web technologies and frameworks within containers to mimic native apps. Backend development involves creating and managing services that support the mobile front end, known as an application programming interface (API). Cloud-based backend services are increasingly popular compared with on-site solutions, as they allow developers to focus on app features without worrying about additional infrastructure. The mobile front end, which is the user-facing aspect of the app, includes the visual and interactive elements, typically residing on devices or accessed through browsers. Development involves various team members, including designers and developers. Front ends interact with backend services via APIs, accessing data and updating backend systems. APIs facilitate communication between components and often integrate third-party functionalities, such as payment processing and social media integration, which enhances the user experience and enables cross-platform compatibility [72].

mHealth apps generate a wide range of data types to help users monitor and manage their health [73]. These include biometric

XSL•FO
RenderX

data, such as heart rate and blood pressure; physical activity data, such as steps and calories burned; sleep data for tracking sleep patterns and quality; and nutrition data for logging food intake and monitoring dietary habits. Additionally, medication adherence apps record medication schedules and doses [74], symptom tracking apps allow users to monitor health indicators and communicate with health care providers [75], and health record apps provide easy access to personal electronic health records [76]. Developers build backend services or leverage databases from third-party providers, which are typically cloud-based but can also be on-site, to improve efficiency and productivity. Security measures safeguard user data through encryption techniques and secure authentication methods. Efficient database management dictates how an app stores, retrieves, and manages data, ensuring seamless user experiences and system reliability. Regular updates and maintenance are essential for sustaining app performance and delivering a reliable user experience amid evolving user needs and technological advancements.

Value Proposition

The health-related mobile app and wearable market features a wide range of developers, from established health care organizations and technology companies to start-ups and independent developers. Monetization models include free apps with in-app purchases, subscription-based models, and paid apps. However, the majority of apps and wearables are currently not reimbursed by insurance or Medicare [77,78]. The market is highly competitive, with a continuous influx of new apps and features. The value of mobile apps and wearables for patients and consumers lies in their ability to empower individuals by providing personalized resources and tools for effective health management. These technologies are generally well-perceived by users, offering convenient access to health-related information and services and allowing individuals to monitor their health [79]. Personalization and tailoring, along with features such as gamification and social support, enhance engagement and motivation, encouraging users to adopt healthier habits and achieve their goals. Ultimately, mHealth apps and wearables empower individuals to actively participate in their health care, fostering autonomy, self-efficacy, and empowerment in managing their well-being.

Adopters

People use mobile apps and wearables across different ages, genders, races, ethnicities, and health systems [36]. However, several factors contribute to the nonadoption of mHealth apps. Usage and access vary by clinical and demographic factors, with challenges including technology literacy, interest, cost, and security concerns [7]. Older adults and individuals with sensory impairments may have difficulty navigating touch interfaces or small screens, while low-income communities face obstacles related to the affordability of smartphones and data plans [80,81]. Additionally, rural populations may struggle with limited internet connectivity, non-English speakers may find comprehension challenging, and individuals with limited digital literacy skills or access to technical support may have difficulties effectively using mobile apps. These factors impact adoption rates among diverse user populations.

In addition to adoption challenges, barriers include insufficient early engagement with users to ensure successful app installation, a lack of direct communication with health care professionals, and limited interactive features. Concerns about technical issues and data security also deter sustained engagement. App usage initiation often depended on the ease of downloading and the individual's motivation. Factors influencing engagement included perceptions of the app's benefits and drawbacks, its relevance and quality, and the incorporation of behavior change techniques to promote health. User engagement is also influenced by individual health goals, user feedback, timely reminders, in-app social support, and coaching, with users having the potential to disengage and reengage as needed [82].

Organization

Organizational barriers and facilitators play a significant role in the adoption and success of mobile apps in clinical care. Key barriers include staff resistance to change, limited resources, concerns about data privacy, interoperability challenges with existing systems, and the need for additional training [8]. Conversely, facilitators include strong leadership support, clear communication and education about benefits, seamless integration into workflows, evidence-based practice demonstrations, and user feedback engagement. Addressing these factors can optimize the adoption and integration of mobile apps into clinical care settings, promoting their effective use for improved patient outcomes.

Wider System

Addressing system barriers requires a multifaceted approach involving collaboration among policy makers, health care providers, technology developers, and community stakeholders to create supportive regulatory frameworks, enhance infrastructure, improve digital literacy, and foster a culture of innovation and acceptance within the health care ecosystem. Key barriers to implementing and scaling health-related mobile apps and wearables include regulatory and legal challenges, particularly concerning data privacy, security, and compliance with medical device regulations across various regions. Additionally, interoperability issues with existing health care systems and electronic health records pose technical hurdles that demand complex solutions and coordination. Disparities in access to technology and digital literacy skills further contribute to a digital divide, limiting the effectiveness of mobile apps among underserved populations. Resistance from health care providers, along with insufficient training and support, hinders adoption, while low user engagement and adherence rates highlight the need to sustain motivation and address usability issues. Achieving sustainability and scalability requires careful resource allocation, long-term funding, and the development of viable business models to expand mobile app and wearable initiatives beyond the pilot and efficacy phases, effectively reaching broader populations.

XSL•FO
RenderX

Social Media

Condition

Social media broadly refers to the use of technology to connect individuals for communication and information sharing within virtual communities. The origins of social media, both as a concept and as a practice, can be traced to early internet systems with group communication features, such as bulletin board systems, USENET newsgroups, and online messaging services such as Internet Relay Chat. These platforms, popular in the 1980s, enabled communities of users with shared interests to communicate with one another. Early 1990s web-based distributed communities, such as webrings and GeoCities (GeoCities/Yahoo!), also represent early forms of social media. Webrings linked websites with similar content through embedded referential links, based on user self-identification and requests for inclusion to a "ringmaster." GeoCities allowed users to define themselves through virtual geographic neighborhoods. However, the first platform explicitly designed to create and establish online social networks of family, friends, and acquaintances was Six Degrees (MacroView/YouthStream Media Networks), launched in 1997. Since then, global social media use has surged to over 5 billion users in 2024, representing 62.3% of the world's population [83].

Technology

Social media platforms are diverse, each focusing on different modalities for group-oriented communication. Content is published as posts, which may include text, images, audio, or video, depending on the platform. A key aspect of social media is the ability for users to interact with others' posts. These interactions are typically permission-based and can be restricted to a closed-circle community or made available for public engagement, depending on platform settings and user preferences. Interaction types are as varied as content types and may include replies, reposts or shares with or without additional content, and acknowledgments of engagement such as likes, dislikes, and favorites. Examples of historical and current social media platforms include those focused on long-form essay and blog posting (eg, LiveJournal [Rambler Media Group], Blogger [Google LLC/Alphabet Inc.], Medium [A Medium Corporation]), short-post microblogging sites with posts originally limited to SMS text message length and remaining generally brief (eg, BlueSky [Bluesky Social, PBC], X [formerly Twitter], Mastodon [Mastodon gGmbH], Plurk [Plurk, Inc.]), image and video-oriented platforms (eg, Instagram [Meta Platforms], Pinterest [Pinterest, Inc.], YouTube [Google LLC], TikTok [ByteDance]), instant messaging platforms (eg, AOL Instant Messenger [AOL], Snapchat [Snap Inc.], QQ [Shenzhen Tencent Computer System Co., Ltd.], WhatsApp [Meta Platforms]), and multimodal social media sites (eg, Friendster [Friendster Labs Inc.], Myspace [Viant Technology LLC], Tumblr [Automattic], Baidu [Baidu, Inc.], Facebook [Meta Platforms]).

Value Proposition

The primary value of social media for health communication, promotion, and intervention lies in its extensive reach, the ability to tailor messaging to specific community or user group interests at scale, and the speed of information transmission among platform members. Social media offers a relatively low-cost means of message promotion, with the potential for broad audience reach through reposting, thereby requiring minimal investment. The concept of "going viral," where a social media post rapidly spreads from user to user, and even across platforms, at a speed and scope far beyond the initial audience, is often considered a measure of significant success in information dissemination. Another area of high value is the data analysis of social media content. Such analyses have been successfully used to identify emerging health concerns and health beliefs held at the population level among specific groups [84,85], trace outbreaks of foodborne illness [86], and track the spread of infectious diseases such as influenza and COVID-19 [87,88].

Adopters

Individual adoption of social media use is so widespread as to be nearly ubiquitous among technology users. As of 2019, there were 3.5 billion people online, of whom 2.4 billion used Facebook, which at the time was the dominant social media platform. This means that 2 out of every 3 people online use social media [89]. Community formation on social media platforms may include individuals who know each other from offline environments or those who solely share similar interests and characteristics online. Trust and influence are also significant factors in social media adoption. Trust among the initial group of contacts who see an original post is extended across networks with each interaction through friend-of-a-friend social acceptance. Parasocial relationships with celebrities and influencers, established through the perception of interacting with their posts as if interacting directly with the person themselves, lead to additional credence being placed in social media content from such individuals. This, in turn, results in increased adoption of promoted messaging. However, social media adoption differs greatly by platform type and user demographic. Health promotion and outreach campaigns, patient recruitment initiatives, or community organizing involving social media can be extremely effective. However, they should consider the population distributions and primary languages of various platforms and tailor messages accordingly to achieve maximum reach and benefit [90].

Organization

Engagement with social media platforms and systems for health communication at organizational levels can be influenced by institutional policies regarding who is authorized to make and respond to posts and what content is permissible for sharing with a public audience. Content experts who develop messaging and communications experts who tailor it for social media may be different individuals or teams. Often, the process of obtaining group or institutional approvals for specific content is much slower than the speed at which messages spread across social media. This delay can be problematic if required authorizations prevent an organization from responding effectively to questions or comments, or from addressing negative replies in a timely manner. In situations where a topic might be perceived as sensitive or controversial, a lack of response can be both detrimental and damaging. Understanding the cultural contexts

XSL•FO
RenderX

of social media use in general, as well as the unique dynamics of specific platform environments, is essential for successfully developing and implementing health communication strategies on these networks.

Wider System

In addition to the factors noted at the organizational level, health-related institutions at governmental levels often possess or are perceived to possess an authority that can make navigating questions of approachability and trust challenging. This is especially true in a social media environment where there must be a balance between adopting a less formal manner of communication and maintaining the formality required to accurately represent the organization and convey health-related content [91]. Differing policies among social media platforms can be challenging to navigate alongside health policies and regulations for data privacy and security, such as HIPAA. Important ethical considerations also arise, including how to maintain accountability and responsibility for content, combat misinformation, and respond to aggressive and harassing behaviors in social media spaces, whether directed at those who post health information or those who react to it.

Embedding and Adaptation Over Time

In the last 25 years, social media has evolved from a niche technology to a central means of engaging communities and the public. Today, the majority of users under the age of 30 years consider social media sources to be as valid as, or even more valid than, traditional news media [92]. The challenge is that the same factors driving the positive aspects of social media can also lead to negative ones. For instance, the rapid pace of information distribution facilitates the spread of both misinformation and disinformation [93,94]. Social media interventions have proven successful in improving mental health status through interpersonal connections and increasing self-esteem and belonging. However, social media can also promote poor self-image through comparisons with artificially high standards, serve as a source of extreme stress, and potentially worsen mental health status [95]. Future work in using social media for health communication and intervention must be mindful of these considerations to avoid unintended consequences.

Discussion

By applying the NASSS framework, this paper elucidates the complex interplay among individual, organizational, and societal factors influencing the adoption and implementation of mHealth technologies. The framework provides a comprehensive perspective to understand the dynamics of technology adoption in health care over the last 25 years. Each mobile technology offers distinct advantages and disadvantages for promoting public health and health care services. SMS text messaging is widely accessible and effective for reaching large populations, especially in low-resource settings. Its simplicity has proven successful in behavior change and medication adherence. However, SMS text messaging is limited in functionality, supporting only basic text communication, which may be less engaging for users seeking interactive or multimedia experiences. Mobile apps and wearables offer a comprehensive range of functionalities, including health tracking, personalized coaching, and integration with wearables, making them ideal for in-depth health management. Yet, they face higher adoption barriers due to the need for smartphones, app literacy, and potential costs, with user engagement often declining over time. Social media platforms, with their extensive reach and rapid information dissemination capabilities, are valuable for building communities and promoting health interventions. However, they carry risks such as the spread of misinformation and require careful management of privacy and data security. Each mobile technology serves different user needs and contexts, highlighting the importance of choosing the appropriate tool for specific health interventions.

Our retrospective discussion highlights the potential of SMS text messaging, mobile apps, wearables, and social media to promote healthy behavior and manage diverse health conditions, underscored by their accessibility and ease of implementation. However, challenges such as provider reluctance, accessibility barriers, and regulatory constraints impede the widespread adoption and integration of mHealth into health care delivery systems. SMS text messaging, while effective, often leads to abandonment due to its limited interactivity, as users may seek more dynamic tools. Mobile apps, despite their comprehensive features, face adoption challenges, including the need for smartphones, app literacy, and potential costs, with user engagement often declining if the apps are not well-designed or integrated into clinical or daily workflows. Social media campaigns struggle with user engagement due to misinformation risks, privacy concerns, and the need for constant content management, which can lead to distrust and abandonment.

Moreover, we emphasize the importance of addressing sociocultural variability and literacy concerns to enhance the inclusivity and effectiveness of mHealth interventions. While these technologies offer personalized resources and empower users to actively participate in their health care, challenges such as limited scientific evidence, technological complexity, and organizational barriers hinder their widespread adoption and impact. Moreover, disparities in access, digital literacy, and user engagement underscore the need for a holistic approach to address systemic barriers and optimize the effectiveness of mHealth interventions.

The last 25 years have seen remarkable strides in adopting and adapting mHealth technologies, driven by technological innovation and a growing recognition of their potential to revolutionize health care delivery. Nevertheless, continued efforts are needed to address persistent challenges and mitigate instances of abandonment, ensuring that mHealth interventions can fully realize their potential in improving health outcomes and advancing equitable access to care.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Singh K, Landman A. Mobile health. In: Sheikh A, Cresswell K, Wright A, Bates D, editors. Key Advances in Clinical Informatics. Cambridge, MA: Academic Press; 2017:183-196.
2. Free C, Phillips G, Watson L, Galli L, Felix L, Edwards P, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. PLoS Med 2013;10(1):e1001363 [FREE Full text] [doi: 10.1371/journal.pmed.1001363] [Medline: 23458994]
3. van Veen T, Binz S, Muminovic M, Chaudhry K, Rose K, Calo S, et al. Potential of mobile health technology to reduce health disparities in underserved communities. West J Emerg Med 2019 Aug 06;20(5):799-802 [FREE Full text] [doi: 10.5811/westjem.2019.6.41911] [Medline: 31539337]
4. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med 2015 Apr 15;7(283):283rv3 [FREE Full text] [doi: 10.1126/scitranslmed.aaa3487] [Medline: 25877894]
5. Ku JP, Sim I. Mobile Health: making the leap to research and clinics. NPJ Digit Med 2021 May 14;4(1):83 [FREE Full text] [doi: 10.1038/s41746-021-00454-z] [Medline: 33990671]
6. Rogers E. Diffusion of Innovations. 3rd ed. London, UK: Free Press; 2003.
7. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical factors affecting patients' adoption of mobile health tools: systematic literature review and narrative synthesis. JMIR Mhealth Uhealth 2022 May 05;10(5):e36284 [FREE Full text] [doi: 10.2196/36284] [Medline: 35318189]
8. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding clinicians' adoption of mobile health tools: a qualitative review of the most used frameworks. JMIR Mhealth Uhealth 2020 Jul 06;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
9. Zakerabasali S, Ayyoubzadeh SM, Baniasadi T, Yazdani A, Abhari S. Mobile health technology and healthcare providers: systemic barriers to adoption. Healthc Inform Res 2021 Oct;27(4):267-278 [FREE Full text] [doi: 10.4258/hir.2021.27.4.267] [Medline: 34788907]
10. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
11. Kuczynski K. When did texting start? Text message and smartphone history. MessageDesk. URL: https://www.messagedesk.com/blog/text-messaging-history-timeline-evolution#:~:text=Text%20messaging%20started%20in%201999,per%20month%20in%20the%20U.S [accessed 2024-05-09]
12. Downer SR, Meara JG, Da Costa AC, Sethuraman K. SMS text messaging improves outpatient attendance. Aust Health Rev 2006 Aug;30(3):389-396. [doi: 10.1071/ah060389] [Medline: 16879098]
13. Franklin VL, Waller A, Pagliari C, Greene SA. A randomized controlled trial of Sweet Talk, a text-messaging system to support young people with diabetes. Diabet Med 2006 Dec;23(12):1332-1338. [doi: 10.1111/j.1464-5491.2006.01989.x] [Medline: 17116184]
14. Weitzel JA, Bernhardt JM, Usdan S, Mays D, Glanz K. Using wireless handheld computers and tailored text messaging to reduce negative consequences of drinking alcohol. J Stud Alcohol Drugs 2007 Jul;68(4):534-537. [doi: 10.15288/jsad.2007.68.534] [Medline: 17568957]
15. Blagojevi I. Texting Statistics. 99 Firms. URL: https://99firms.com/blog/texting-statistics/#gref [accessed 2024-05-09]
16. Boland V, Stockings E, Mattick R, McRobbie H, Brown J, Courtney R. The methodological quality and effectiveness of technology-based smoking cessation interventions for disadvantaged groups: a systematic review and meta-analysis. Nicotine Tob Res 2018 Feb 07;20(3):276-285. [doi: 10.1093/ntr/ntw391] [Medline: 28034998]
17. Scott-Sheldon LAJ, Lantini R, Jennings EG, Thind H, Rosen RK, Salmoirago-Blotcher E, et al. Text messaging-based interventions for smoking cessation: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016 May 20;4(2):e49 [FREE Full text] [doi: 10.2196/mhealth.5436] [Medline: 27207211]
18. Siopis G, Chey T, Allman-Farinelli M. A systematic review and meta-analysis of interventions for weight management using text messaging. J Hum Nutr Diet 2015 Feb;28 Suppl 2:1-15. [doi: 10.1111/jhn.12207] [Medline: 24480032]
19. Spohr SA, Nandy R, Gandhiraj D, Vemulapalli A, Anne S, Walters ST. Efficacy of SMS text message interventions for smoking cessation: a meta-analysis. J Subst Abuse Treat 2015 Sep;56:1-10. [doi: 10.1016/j.jsat.2015.01.011] [Medline: 25720333]
20. Bendtsen M. Text messaging interventions for reducing alcohol consumption among harmful and hazardous drinkers: protocol for a systematic review and meta-analysis. JMIR Res Protoc 2019 Apr 23;8(4):e12898 [FREE Full text] [doi: 10.2196/12898] [Medline: 31012866]
21. Whitemore R, Leonardi-Bee J, Naughton F, Sutton S, Cooper S, Parrott S, et al. Effectiveness and cost-effectiveness of a tailored text-message programme (MiQuit) for smoking cessation in pregnancy: study protocol for a randomised controlled

XSL•FO
RenderX

trial (RCT) and meta-analysis. Trials 2019 May 22;20(1):280 [FREE Full text] [doi: 10.1186/s13063-019-3341-4] [Medline: 31118090]
22. Schippers M, Adam PCG, Smolenski DJ, Wong HTH, de Wit JBF. A meta-analysis of overall effects of weight loss interventions delivered via mobile phones and effect size differences according to delivery mode, personal contact, and intervention intensity and duration. Obes Rev 2017 Apr;18(4):450-459. [doi: 10.1111/obr.12492] [Medline: 28187246]
23. Boksmati N, Butler-Henderson K, Anderson K, Sahama T. The effectiveness of SMS reminders on appointment attendance: a meta-analysis. J Med Syst 2016 Apr;40(4):90. [doi: 10.1007/s10916-016-0452-2] [Medline: 26852337]
24. Finitsis DJ, Pellowski JA, Johnson BT. Text message intervention designs to promote adherence to antiretroviral therapy (ART): a meta-analysis of randomized controlled trials. PLoS One 2014;9(2):e88166 [FREE Full text] [doi: 10.1371/journal.pone.0088166] [Medline: 24505411]
25. Park LG, Howie-Esquivel J, Dracup K. A quantitative systematic review of the efficacy of mobile phone interventions to improve medication adherence. J Adv Nurs 2014 Sep;70(9):1932-1953. [doi: 10.1111/jan.12400] [Medline: 24689978]
26. Nglazi MD, Bekker L, Wood R, Hussey GD, Wiysonge CS. Mobile phone text messaging for promoting adherence to anti-tuberculosis treatment: a systematic review protocol. Syst Rev 2013 Jan 16;2:6 [FREE Full text] [doi: 10.1186/2046-4053-2-6] [Medline: 23324135]
27. Farmer AJ, McSharry J, Rowbotham S, McGowan L, Ricci-Cabello I, French DP. Effects of interventions promoting monitoring of medication use and brief messaging on medication adherence for people with Type 2 diabetes: a systematic review of randomized trials. Diabet Med 2016 May;33(5):565-579. [doi: 10.1111/dme.12987] [Medline: 26470750]
28. Horvath T, Azman H, Kennedy GE, Rutherford GW. Mobile phone text messaging for promoting adherence to antiretroviral therapy in patients with HIV infection. Cochrane Database Syst Rev 2012 Mar 14;2012(3):CD009756 [FREE Full text] [doi: 10.1002/14651858.CD009756] [Medline: 22419345]
29. Whittemore R, Siverly L, Wischik DL, Whitehouse CR. An umbrella review of text message programs for adults with type 2 diabetes. Diabetes Educ 2020 Dec;46(6):514-526. [doi: 10.1177/0145721720965464] [Medline: 33353511]
30. de Jongh T, Gurol-Urganci I, Vodopivec-Jamsek V, Car J, Atun R. Mobile phone messaging for facilitating self-management of long-term illnesses. Cochrane Database Syst Rev 2012 Dec 12;12(12):CD007459 [FREE Full text] [doi: 10.1002/14651858.CD007459.pub2] [Medline: 23235644]
31. Wan Y, Wu X, Kou Y. The impact of text message on self-management for coronary heart disease: a meta-analysis of randomized controlled trials. Heart Surg Forum 2020 Feb 10;23(1):E018-E024. [doi: 10.1532/hsf.2519] [Medline: 32118537]
32. Arambepola C, Ricci-Cabello I, Manikavasagam P, Roberts N, French DP, Farmer A. The impact of automated brief messages promoting lifestyle changes delivered via mobile devices to people with type 2 diabetes: a systematic literature review and meta-analysis of controlled trials. J Med Internet Res 2016 Apr 19;18(4):e86 [FREE Full text] [doi: 10.2196/jmir.5425] [Medline: 27095386]
33. Bramley D, Riddell T, Whittaker R, Corbett T, Lin R, Wills M, et al. Smoking cessation using mobile phone text messaging is as effective in Maori as non-Maori. N Z Med J 2005 Jun 03;118(1216):U1494. [Medline: 15937529]
34. Materia FT, Smyth JM, Puoane T, Tsolekile L, Goggin K, Kodish SR, et al. Implementing text-messaging to support and enhance delivery of health behavior change interventions in low- to middle-income countries: case study of the Lifestyle Africa intervention. BMC Public Health 2023 Aug 10;23(1):1526 [FREE Full text] [doi: 10.1186/s12889-023-16388-y] [Medline: 37563595]
35. MacDougall S, Jerrott S, Clark S, Campbell LA, Murphy A, Wozney L. Text message interventions in adolescent mental health and addiction services: scoping review. JMIR Ment Health 2021 Jan 08;8(1):e16508 [FREE Full text] [doi: 10.2196/16508] [Medline: 33416504]
36. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/#:~:text=The%20vast%20majority%20of%20Americans,a%20cellphone%20of%20some%20kind [accessed 2024-05-09]
37. Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler N, et al. Powering population health research: considerations for plausible and actionable effect sizes. SSM Popul Health 2021 Jun;14:100789 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100789] [Medline: 33898730]
38. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013 Nov;97:41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
39. Chen MA, Hollenberg JP, Michelen W, Peterson JC, Casalino LP. Patient care outside of office visits: a primary care physician time study. J Gen Intern Med 2011 Jan;26(1):58-63 [FREE Full text] [doi: 10.1007/s11606-010-1494-7] [Medline: 20811956]
40. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022 Mar 09;12(3):e054561 [FREE Full text] [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]
41. Siedner MJ, Santorino D, Haberer JE, Bangsberg DR. Know your audience: predictors of success for a patient-centered texting app to augment linkage to HIV care in rural Uganda. J Med Internet Res 2015 Mar 24;17(3):e78 [FREE Full text] [doi: 10.2196/jmir.3859] [Medline: 25831269]
42. Patton SR, Coffman MJ, De Haven MJ, Miller C, Krinner LM. Text message intervention for Latino adults to improve diabetes outcomes. Hisp Health Care Int 2022 Dec;20(4):248-255. [doi: 10.1177/15404153221084610] [Medline: 35274994]

XSL•FO
RenderX

43. Martindale-Adams JL, Clark CD, Martin JR, Henderson CR, Nichols LO. Text messages to support caregivers in a health care system: development and pilot and national rollout evaluation. J Particip Med 2022 Oct 17;14(1):e35318 [FREE Full text] [doi: 10.2196/35318] [Medline: 36251368]
44. Yunusa U, Garba SN, Umar AB, Idris SH, Bello UL, Abdulrashid I, et al. Mobile phone reminders for enhancing uptake, completeness and timeliness of routine childhood immunization in low and middle income countries: a systematic review and meta-analysis. Vaccine 2021 Jan 08;39(2):209-221. [doi: 10.1016/j.vaccine.2020.11.043] [Medline: 33277058]
45. Elepaño A, Fusingan AS, Yasay E, Sahagun JA. Mobile health interventions for improving colorectal cancer screening rates: a systematic review and meta-analysis. Asian Pac J Cancer Prev 2021 Oct 01;22(10):3093-3099 [FREE Full text] [doi: 10.31557/APJCP.2021.22.10.3093] [Medline: 34710983]
46. Hutnick A, Wainwright J. Calling and texting regulations tightening: FCC adopts 1TCPA consent and do-not-text rules. Kelley Drye. URL: https://www.kelleydrye.com/viewpoints/blogs/ad-law-access/calling-and-texting-regulations-tightening-fcc-adopts-11-tcpa-consent-and-do-not-text-rules [accessed 2024-05-09]
47. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano Serrano ME, Martinez-Suberviola FJ, Dejal@ Group. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter, controlled, randomized clinical trial in primary care. JMIR Mhealth Uhealth 2022 Jun 27;10(6):e34273 [FREE Full text] [doi: 10.2196/34273] [Medline: 35759328]
48. Oh K, Lee D, Ko B, Hyeon J, Choi H. Empathy bot: conversational service for psychiatric counseling with chat assistant. Stud Health Technol Inform 2017;245:1235. [Medline: 29295322]
49. Jones OT, Calanzani N, Saji S, Duffy SW, Emery J, Hamilton W, et al. Artificial intelligence techniques that may be applied to primary care data to facilitate earlier diagnosis of cancer: systematic review. J Med Internet Res 2021 Mar 03;23(3):e23483 [FREE Full text] [doi: 10.2196/23483] [Medline: 33656443]
50. Millenson M, Baldwin J, Zipperer L, Singh H. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis (Berl) 2018 Sep 25;5(3):95-105 [FREE Full text] [doi: 10.1515/dx-2018-0009] [Medline: 30032130]
51. Mobile health apps: timeline. Medical Device Network. 2020. URL: https://www.medicaldevice-network.com/features/mobile-health-apps-mhealth/?cf-view [accessed 2024-05-09]
52. Kao C, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM R 2017 May;9(5S):S106-S115. [doi: 10.1016/j.pmrj.2017.02.018] [Medline: 28527495]
53. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
54. Marvel F, Dowell P, Mossburg S. Emergence of application-based health care. Agency for Healthcare Research and Quality. URL: https://psnet.ahrq.gov/perspective/emergence-application-based-health [accessed 2024-05-09]
55. How mobile health apps are improving health care: saving costs, saving lives. MindSea. URL: https://mindsea.com/health-apps/ [accessed 2024-05-09]
56. Digital health trends 2021: innovation, evidence, regulation and adoption. IQVIA. 2021 Jul. URL: https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/digital-health-trends-2021/iqvia-institute-digital-health-trends-2021.pdf [accessed 2024-05-09]
57. Health apps usage statistics. Vicert. San Francisco, CA: Vicert Digital Health Solutions; 2024 Oct. URL: https://www.vicert.com/blog/health-apps-usage-statistics [accessed 2024-05-09]
58. Irwin M. From pedometers to Fitbit: the journey of activity trackers. Irwins megastore. 2021. URL: https://irwinsmegastore.ie/blogs/tech/history-of-activity-trackers-from-pedometers-to-fitbit-infographic [accessed 2024-06-01]
59. Parrott M, Ruyak J, Liguori G. The history of exercise equipment. ACSM's Health and Fitness Journal 2020;24(6):5-8. [doi: 10.1249/fit.0000000000000612]
60. Grundy QH, Wang Z, Bero LA. Challenges in assessing mobile health app quality: a systematic review of prevalent and innovative methods. Am J Prev Med 2016 Dec;51(6):1051-1059. [doi: 10.1016/j.amepre.2016.07.009] [Medline: 27659122]
61. Gasteiger N, Dowding D, Norman G, McGarrigle L, Eost-Telling C, Jones D, et al. Conducting a systematic review and evaluation of commercially available mobile applications (apps) on a health-related topic: the TECH approach and a step-by-step methodological guide. BMJ Open 2023 Jun 12;13(6):e073283 [FREE Full text] [doi: 10.1136/bmjopen-2023-073283] [Medline: 37308269]
62. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. NPJ Digit Med 2018 May 9;1(1):12 [FREE Full text] [doi: 10.1038/s41746-018-0021-9] [Medline: 31304297]
63. Roberts AE, Davenport TA, Wong T, Moon H, Hickie IB, LaMonica HM. Evaluating the quality and safety of health-related apps and e-tools: adapting the Mobile App Rating Scale and developing a quality assurance protocol. Internet Interventions 2021 Apr;24:100379 [FREE Full text] [doi: 10.1016/j.invent.2021.100379] [Medline: 33777705]
64. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021 Jan 11;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
65. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020 Jan 13;3(1):4. [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]

XSL•FO
RenderX

66. Cucciniello M, Petracca F, Ciani O, Tarricone R. Development features and study characteristics of mobile health apps in the management of chronic conditions: a systematic review of randomised trials. NPJ Digit Med 2021 Oct 05;4(1):144 [FREE Full text] [doi: 10.1038/s41746-021-00517-1] [Medline: 34611287]
67. Barteit S, Boudo V, Ouedraogo A, Zabré P, Ouremi L, Sié A, et al. Feasibility, acceptability and validation of wearable devices for climate change and health research in the low-resource contexts of Burkina Faso and Kenya: study protocol. PLoS One 2021 Sep 30;16(9):e0257170 [FREE Full text] [doi: 10.1371/journal.pone.0257170] [Medline: 34591893]
68. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
69. Mattison G, Canfell O, Forrester D, Dobbins C, Smith D, Töyräs J, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 01;24(7):e36690 [FREE Full text] [doi: 10.2196/36690] [Medline: 35776492]
70. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
71. What's the difference between web apps, native apps, and hybrid apps? Amazon Web Services. URL: https://aws.amazon.com/compare/the-difference-between-web-apps-native-apps-and-hybrid-apps/#:~:text=In%20a%20native%20app%2C%20your,native%20app%20shell%20or%20container [accessed 2024-05-09]
72. What is mobile application development? Amazon Web Services. URL: https://aws.amazon.com/mobile/mobile-application-development/ [accessed 2024-05-09]
73. Fischer F, Kleen S. Possibilities, problems, and perspectives of data collection by mobile apps in longitudinal epidemiological studies: scoping review. J Med Internet Res 2021 Jan 22;23(1):e17691 [FREE Full text] [doi: 10.2196/17691] [Medline: 33480850]
74. Santo K, Richtering SS, Chalmers J, Thiagalingam A, Chow CK, Redfern J. Mobile phone apps to improve medication adherence: a systematic stepwise process to identify high-quality apps. JMIR Mhealth Uhealth 2016 Dec 02;4(4):e132 [FREE Full text] [doi: 10.2196/mhealth.6742] [Medline: 27913373]
75. Cruz S, Fernandes C, Magalhães B. A scoping review of mobile apps for use with palliative patients in the context of home care. Int J Med Inform 2023 Sep;177:105166. [doi: 10.1016/j.ijmedinf.2023.105166] [Medline: 37527596]
76. Johnson C, Richwine C, Patel V. Individuals' access and use of patient portals and smartphone health apps, 2020. Office of the National Coordinator for Health IT. 2021. URL: https://www.healthit.gov/sites/default/files/page/2021-09/HINTS_2020_Consumer_Data_Brief.pdf [accessed 2024-05-09]
77. Weinstein RS, Lopez AM, Joseph BA, Erps KA, Holcomb M, Barker GP, et al. Telemedicine, telehealth, and mobile health applications that work: opportunities and barriers. Am J Med 2014 Mar;127(3):183-187. [doi: 10.1016/j.amjmed.2013.09.032] [Medline: 24384059]
78. Powell AC, Bowman MB, Harbin HT. Reimbursement of apps for mental health: findings from interviews. JMIR Ment Health 2019 Aug 06;6(8):e14724 [FREE Full text] [doi: 10.2196/14724] [Medline: 31389336]
79. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019 Jul 10;7(7):e13817 [FREE Full text] [doi: 10.2196/13817] [Medline: 31293246]
80. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024 Feb 16;24(1):506 [FREE Full text] [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
81. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023 Sep 05;9:20552076231198702 [FREE Full text] [doi: 10.1177/20552076231198702] [Medline: 37691766]
82. Moungui HC, Nana-Djeunga HC, Anyiang CF, Cano M, Ruiz Postigo JA, Carrion C. Dissemination strategies for mHealth apps: systematic review. JMIR Mhealth Uhealth 2024 Jan 05;12:e50293 [FREE Full text] [doi: 10.2196/50293] [Medline: 38180796]
83. Kemp S. Digital 2024: global overview report. DataReportal. 2024. URL: https://datareportal.com/reports/digital-2024-global-overview-report [accessed 2024-05-09]
84. Meadows CZ, Tang L, Liu W. Twitter message types, health beliefs, and vaccine attitudes during the 2015 measles outbreak in California. Am J Infect Control 2019 Nov;47(11):1314-1318. [doi: 10.1016/j.ajic.2019.05.007] [Medline: 31266661]
85. Gkotsis G, Mueller C, Dobson R, Hubbard T, Dutta R. Mining social media data to study the consequences of dementia diagnosis on caregivers and relatives. Dement Geriatr Cogn Disord 2020 Aug 27;49(3):295-302 [FREE Full text] [doi: 10.1159/000509123] [Medline: 32854092]
86. Tao D, Hu R, Zhang D, Laber J, Lapsley A, Kwan T, et al. A novel foodborne illness detection and web application tool based on social media. Foods 2023 Jul 20;12(14):2769 [FREE Full text] [doi: 10.3390/foods12142769] [Medline: 37509861]
87. Huang X, Wang S, Zhang M, Hu T, Hohl A, She B, et al. Social media mining under the COVID-19 context: progress, challenges, and opportunities. Int J Appl Earth Obs Geoinf 2022 Sep;113:102967 [FREE Full text] [doi: 10.1016/j.jag.2022.102967] [Medline: 36035895]

XSL•FO
RenderX

88. Alessa A, Faezipour M. A review of influenza detection and prediction through social networking sites. Theor Biol Med Model 2018 Feb 01;15(1):2 [FREE Full text] [doi: 10.1186/s12976-017-0074-5] [Medline: 29386017]
89. Ortiz-Ospina E. The rise of social media. Our World in Data. 2019. URL: https://ourworldindata.org/rise-of-social-media [accessed 2024-05-09]
90. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020 Oct 26;22(10):e22810 [FREE Full text] [doi: 10.2196/22810] [Medline: 33104015]
91. Mergel I. Social media adoption: toward a representative, responsive or interactive government. New York, NY: ACM; 2014 Presented at: The 15th Annual International Conference on Digital Government Research; June 18-21, 2014; Aguascalientes, Mexico p. 163-170. [doi: 10.1145/2612733.2612740]
92. Social Media Fact Sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2024-05-09]
93. Cascini F, Pantovic A, Al-Ajlouni YA, Failla G, Puleo V, Melnyk A, et al. Social media and attitudes towards a COVID-19 vaccination: a systematic review of the literature. EClinicalMedicine 2022 Jun;48:101454 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101454] [Medline: 35611343]
94. Wilson SL, Wiysonge C. Social media and vaccine hesitancy. BMJ Glob Health 2020 Oct;5(10):004206 [FREE Full text] [doi: 10.1136/bmjgh-2020-004206] [Medline: 33097547]
95. Zsila, Reyes MES. BMC Psychol 2023 Jul 06;11(1):201 [FREE Full text] [doi: 10.1186/s40359-023-01243-x] [Medline: 37415227]

Abbreviations

AI: artificial intelligence
API: application programming interface
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health
NASSS: Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability

Edited by G Eysenbach; submitted 31.05.24; peer-reviewed by S Iribarren, M Stach; comments to author 17.07.24; revised version received 14.08.24; accepted 05.09.24; published 27.09.24.

Please cite as:
Portz J, Moore S, Bull S
Evolutionary Trends in the Adoption, Adaptation, and Abandonment of Mobile Health Technologies: Viewpoint Based on 25 Years of Research
J Med Internet Res 2024;26:e62790
URL: https://www.jmir.org/2024/1/e62790
doi:10.2196/62790
PMID:39331463

©Jennifer Portz, Susan Moore, Sheana Bull. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Hospital Is Not the Home: Lessons From Implementing Remote Technology to Support Acute Inpatient and Transitional Care in the Home in the United States and United Kingdom

Matt Wilkes[1], MBChB, PhD; Annabel Kramer[1], BSc; Juliana Pugmire[1], MPH, DrPH; Christopher Pilkington[1], EdM, MCP; Benjamin Zaniello[1], MD, MPH; Nicole Zahradka[1], PhD

Best Buy Health, Boston, MA, United States

Corresponding Author:
Matt Wilkes, MBChB, PhD
Best Buy Health
2 Seaport Lane
Boston, MA, 02210
United States
Phone: 1 415 941 5734
Email: matt.wilkes@bestbuy.com

Abstract

The COVID-19 pandemic, patient preference, and economic opportunity are shifting acute care from the hospital to the home, supported by the transformation in remote monitoring technology. Monitoring patients with digital medical devices gives unprecedented insight into their physiology. However, this technology does not exist in a vacuum. Distinguishing pathology from physiological variability, user error, or device limitations is challenging. In a hospital, patients are monitored in a contrived environment. Monitoring at home instead captures activities of daily living alongside patients' trajectory of disease and recovery. Both settings make for "noisy" data. However, we are familiar with hospital noise, accounting for it in our practice and perceptions of normal. Home monitoring as a diagnostic intervention introduces a new set of downstream consequences, dependent on device, cadence of collection, adherence, duration, alarm thresholds, and escalation criteria. We must accept greater ambiguity and contextualize vital signs. All devices balance accuracy with acceptability, so compromises are inevitable and perfect data should not be expected. Alarms must be specific as well as sensitive, balancing clinical risk with capacity for response. By setting expectations around data from the home, we can smooth the adoption of remote monitoring and accelerate the transition of acute care.

(J Med Internet Res 2024;26:e58888) doi:10.2196/58888

KEYWORDS

telemedicine; implementation science; hospital-to-home transition; remote patient monitoring; digital health; transition of care; accuracy; acceptability

Introduction

Remote patient monitoring (RPM) has been evolving for over 60 years. In 1961, NASA developed the technology to monitor the Project Mercury astronauts' electrocardiography, temperature, and respiration rate in flight [1]. Since then, RPM has developed in parallel with health care technology, but the SARS-CoV-2 pandemic was an inflection point in demand. Significant volumes of acute and transitional care moved from in person to virtual, and from the hospital to the home [2]. Continued government support via telehealth reimbursement and the Centers for Medicare and Medicaid Services Acute Hospital Care at Home Waiver has been a critical enabler, but so has provider and patient preferences: remote care can ease access, and care at home for many conditions is seen as preferable to a clinic or hospital visit [3,4]. Economically, it is an enticing proposition with the potential to reduce cost of care, free up inpatient capacity, and reduce readmissions [5]. Remotely monitoring patients with digital medical devices can give unprecedented insight into their physiology, and RPM is now a core part of programs as health care delivery transitions to the home [6].

These devices measure familiar vital signs such as heart rate, respiratory rate, oxygen saturation, and temperature. However, the advantages of acute care at home, particularly patients' ability to perform their activities of daily living, mean that data from the home are not the same as those from the hospital.

XSL•FO
RenderX

Indeed, in this viewpoint, we argue that their superficial similarity mandates a conscious adjustment in our expectations and practice of monitoring. Clinicians, researchers, and innovators building on data gathered from home monitoring must understand the substrate of their work. A planned implementation, including appropriate education of all involved, is key to the adoption of new technologies in health care [7]. Without educating clinicians about the differences between hospital and home monitoring, and giving them strategies to practice effectively, they can quickly lose confidence in the technology. We have implemented intermittent and continuous home monitoring in acute care settings (predominantly hospital at home and oncology) in 21 US health systems (22,264 patients) between July 2019 and 2024, and 22 UK National Health Service trusts (11,862 patients) between December 2017 and July 2024. The median (IQR) length of stay between patients first transmitting vital signs to discharge was 3.9 (1.9-12.9) days in the United States and 7.3 (4.6-12.2) days in the United Kingdom. Our implementations have typically included the provision and configuration of proprietary and third-party remote monitoring devices, electronic medical record integration, care pathway design, staff training, logistics, and program troubleshooting and evaluation. In addition, our US nursing triage service has provided direct monitoring for 5515 patients in hospital-at-home and acute oncology settings since December 2021. In this viewpoint, we highlight some underappreciated differences between home and hospital monitoring and share our best practices, ranging from the philosophical to the tactical, to smooth implementation and make effective use of this promising technology.

Hospitals are highly contrived environments. Trained practitioners collect vital signs with patients being awake, deliberately positioned, and motionless [8]. The architecture of the hospital floor, the social roles of clinician and patient, and the ingrained rituals of care dictate the interaction. In contrast, monitoring in the home is more naturalistic. Patients live and act according to their condition, social role, resources, and environment. This lack of restriction is, in many ways, the "point" of care in the home, key to its therapeutic intent and the richness of its monitoring data [9,10]. Both the hospital and the home make for "noisy" data. However, we are more familiar with hospital noise, so we subconsciously account for it in our practice and our perceptions of "normal" vital signs [11,12]. When there is ambiguity, we can more easily physically examine the patient. The noise of the home is novel, and the patient is out of our immediate reach so we cannot visually contextualize it. We must actively familiarize ourselves with this new approach to monitoring if we are to use it to its full potential.

Accuracy Versus Acceptability

Care in the home takes measurement out of clinicians' hands. The patient collects their own vital signs or passively wears the equipment unsupervised. Both carry the possibility of user error. We must also accept compromises in the technology itself. Many hospital monitoring devices are heavy, expensive, complex, wired, and unintuitive. They are not suitable for the home. Home monitoring equipment must be untethered, less obtrusive, and more intuitive for patient adherence [13]. Trading accuracy for adherence is the price of home monitoring, and a balance must be struck between accuracy and acceptability depending on the population and goals of the monitoring (Figure 1). For example, a finger pulse oximeter is not practical to wear constantly. So, we gather oxygen saturation and heart rate from watches, armbands, or patches instead. These devices reflect light from the skin instead of transmitting it through the digital arteries of the finger like a hospital probe. The available signal is reduced, rendering them more vulnerable to motion artifact, high ambient light, or changes in skin perfusion [14]. When validated in laboratory conditions, both finger probes and wearables can perform equally well, but the laboratory is not the home either.

Figure 1. The accuracy-acceptability spectrum. HCP: health care provider; ICU: intensive care unit.

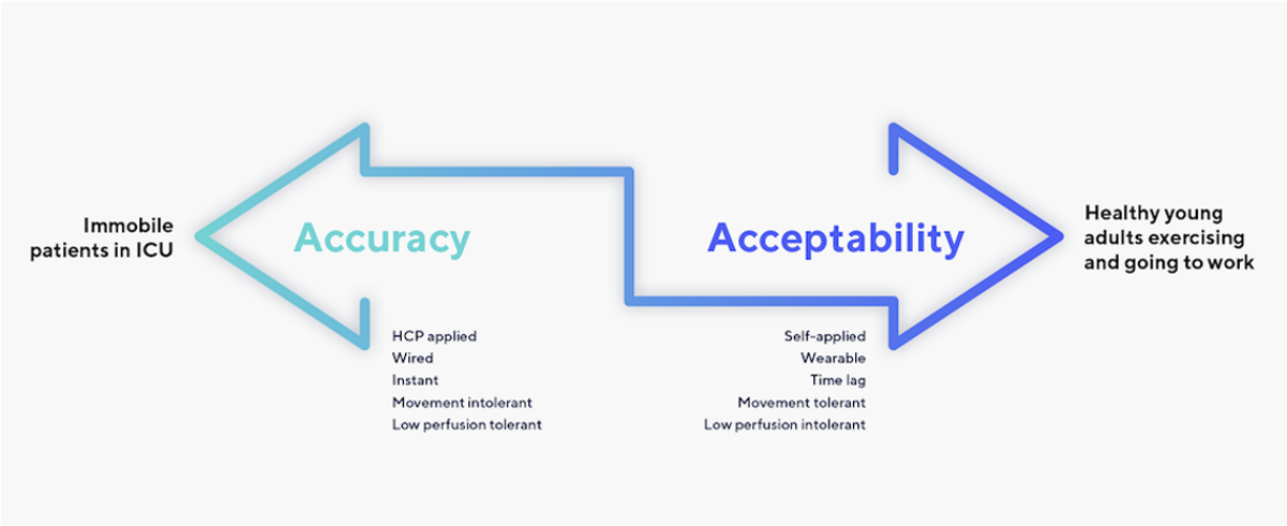

XSL•FO
RenderX

Continuous Versus Intermittent

Vital sign monitoring in the home can be continuous (via stick-on patches or armbands), intermittent (eg, finger pulse oximeters or blood pressure cuffs), or both. Continuous monitors collect data passively (the patient just wears them), while intermittent monitors require a patient to initiate a reading. Continuous monitoring has historically been the province of critical care and the operating room. When untethered continuous monitoring became possible, the ease of passive data collection and the superficial reassurance of "ICU-style monitoring" made it an appealing solution for acute care in the home. However, its utility is an evolving question: it remains unclear whether (1) the advantages of passive data collection and reduced interobserver variability outweigh the challenges of managing high data volumes and (2) whether the opportunity to detect deterioration more rapidly exceeds the consequences of incidental findings that subsequently prove innocuous [12,15,16].

Continuous monitoring requires effective handling of large volumes of data [17]. Data are typically aggregated into windows of time (eg, "median SpO_2 over the last hour") to make them digestible and facilitate transmission to the electronic health record (EHR). Longer windows of aggregation mean fewer data points to triage with greater suppression of outliers and variability (more specificity and less sensitivity), which helps mitigate noise from activities of daily living but can mask brief but clinically significant periods of deterioration. Aggregation also introduces time lags compared to spot monitoring [18]. These time lags mean that discrepancies between continuous and intermittent data streams frequently arise, even if both are functioning correctly, with the potential for loss of confidence in the technology if not well understood by users.

Figure 2 illustrates this scenario. The patient was an older adult man in a US hospital-at-home program, with cellulitis of the lower limb. He had been sent home with a continuous monitor (Current Health Gen 2; Best Buy Health), and his main caregiver had independently purchased an over-the-counter finger pulse oximeter. The raw monitoring data (Figure 2C) indicated 3 episodes of hypoventilation and hypoxia, after a postural change that was visible on the motion sensors (Figure 2A). Figure 2B illustrates the misunderstanding that followed. The continuous data were aggregated into medians over time for ease of recording in the EHR (Figure 2B; the blue bars represent this aggregation). During sleep, the patient rolled onto one side and their oxygen saturation levels decreased. The nursing triage service called his caregiver, who woke the patient, sat him up, and encouraged him to take deep breaths. The patient's saturation levels rapidly rose again. The patient's caregiver also asked the patient to put on their finger probe and the patient's oxygen saturation levels were normal (Figure 2B; pink dot). The continuous monitor, on a longer time aggregation, lagged and still showed hypoxia before normalizing. The caregiver queried this apparent discrepancy, which resulted in an investigation. On close of examination of the continuous data, it appeared the patient had sleep apnea. This was confirmed by a subsequent sleep study. The continuous monitor was dismissed as inaccurate, with the more familiar finger probe treated as the "reference." Yet, it was the continuous monitor, with its capacity for detailed review of raw signal data, that identified the pathology. This example also illustrated the potential for incidental findings with continuous monitoring, leading to downstream investigations that, in the case of this nonagenarian patient, were unlikely to affect his overall prognosis.

XSL•FO
RenderX

Figure 2. (A) Data from motion-sensing accelerometers and gyroscopes; (B) the basis of the discrepancy between continuous and spot data: the blue bars represent aggregated continuous data from the wearable, lagging behind the spot monitoring by the finger probe (pink dot); (C) raw monitoring data illustrating 3 episodes of hypoventilation and hypoxia, after a postural change recorded by the motion-sensing accelerometers and gyroscopes. HCP: health care provider; Resp.: respiratory; SpO2: oxygen saturation.

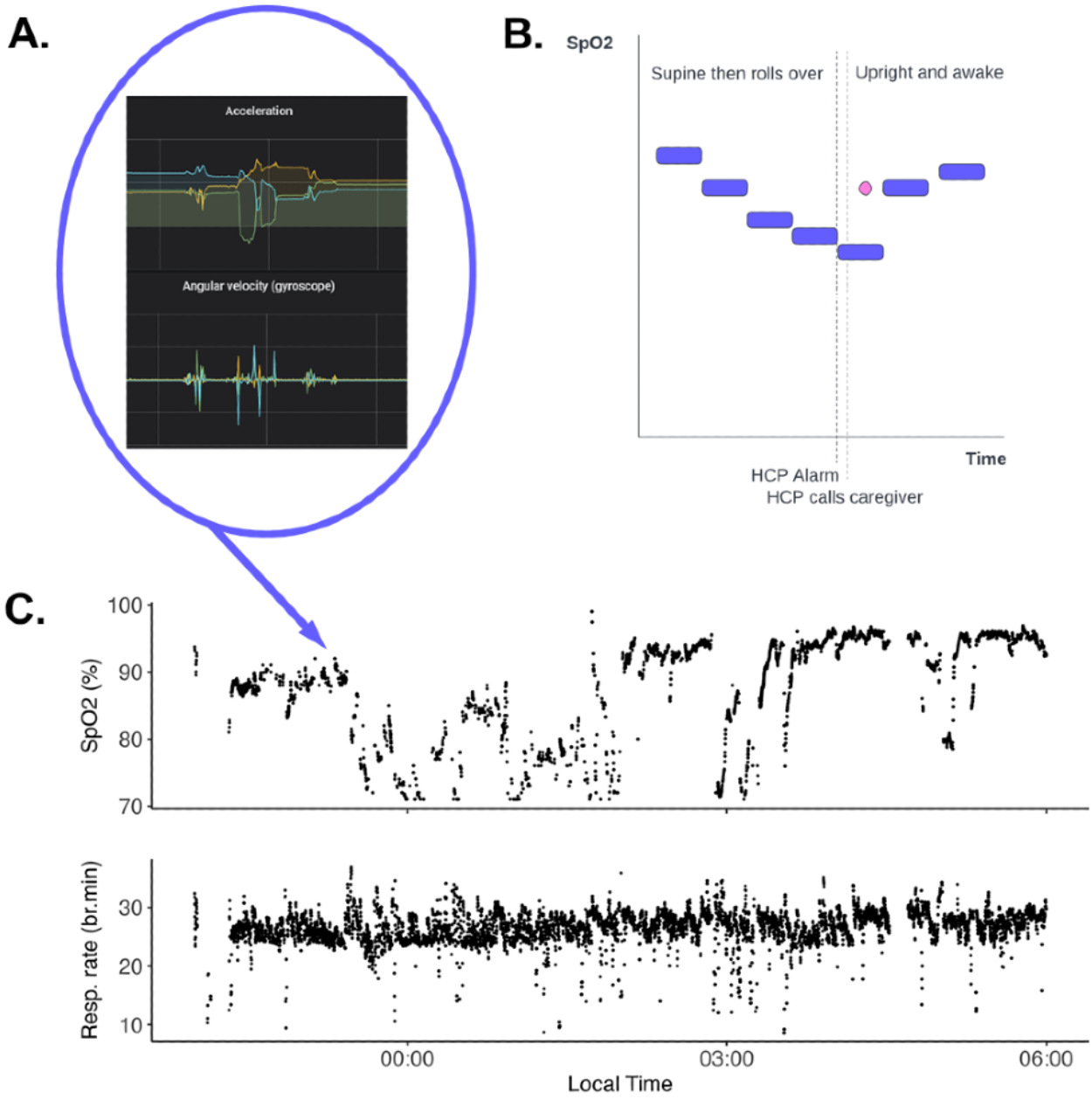

Interpreting Variability

When a monitoring device records an unexpected value, either in the hospital or in the home, we must parse whether it represents genuine clinical change, natural variation, user behavior, user error, device error, or a combination of the above (Figure 3) [19]. It begs the question of what is a "normal" or expected reading. Humans exhibit inter- and intraindividual variability, even when healthy and at rest. We have predictable patterns of variation in vital signs, for example, circadian rhythms and menstrual cycles, as well as unpredictable patterns based on activity and internal and external stimuli [20,21]. These are further complicated by our trajectories of disease or recovery [22]. In the hospital, variation is suppressed by intermittent spot monitoring under contrived circumstances, or its cause is obvious: for example, a tachycardia might be due to a patient being suddenly awoken, just having taken medication, standing up, or receiving bad news. A data stream from the home lacks this context, so our challenge becomes distinguishing between variability and pathology. In intermittent monitoring, there may be too few data points to do this easily, and in continuous monitoring, there may be too many [18]. Nonetheless, trends in variability are becoming increasingly important indicators of health and disease [22,23]. These are readily suppressed in the hospital but better preserved in the home.

XSL•FO
RenderX

Figure 3. Variations in systolic and diastolic blood pressure readings during the rollout of an Food and Drug Administration 510(k)–cleared cuff in the home (5418 patients). The gray bars are the total number of daily readings (n=159,645), the blue bars are readings where the systolic variance from the previous day is >20 mm Hg (median 20%, IQR 19%-22% of readings), and the red bars are readings where the diastolic variance is >20 mm Hg (median 8%, IQR 7%-10% of readings). For reference, a systematic review found clinical blood pressure measurement to vary from true resting blood pressure, from –23.6 to +33 mm Hg for systolic blood pressure and from –14 to +23 mm Hg for diastolic blood pressure, due to patient, device, procedural, and observer factors, indicating that this problem is not unique to the home.

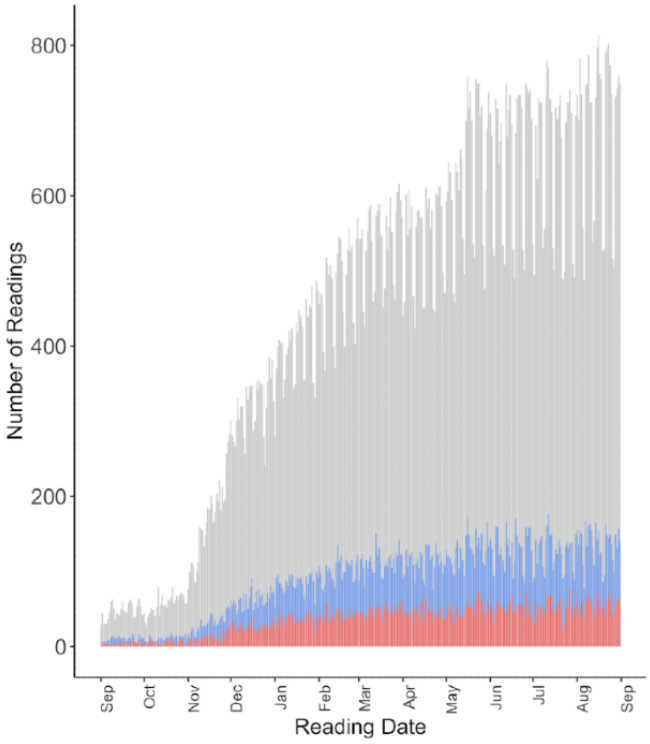

Monitoring Successfully in the Home: Lessons Learned

Overview

We believe that monitoring a patient in their own environment during acute and transitional care can be less intrusive and more informative than in the hospital. However, we need to implement home monitoring collaboratively and iteratively, being mindful of the technology and the needs of the service, while managing the expectations of all involved [24]. The following recommendations are based on our experience.

Choose and Test Technology to Achieve a Specific Goal

It is important to select technology based on the program's goals, rather than simply adopting what is available, previously purchased, or the norm on the hospital floor. Then, the technology should be actively studied on how it performs remotely. Regulatory clearance (eg, Conformité Européenne [CE] marking or Food and Drug Administration 510(k) clearance) is the minimum bar for any clinical device, but the laboratory validation studies required for clearance give little indication of how well it will perform in the home. Pragmatic testing in the context in which the technology will be used is essential to ensure it will be fit for purpose. This includes ensuring equitable provision for patients with low levels of digital literacy or access [25]. Challenges such as the provision of infrastructure or subsidization of access are more easily addressed on a governmental level. However, technology providers can work to design accessible standards, develop specific training programs, incorporate translation services into platform communications, and ensure that cultural sensitivities are respected [26]. Building in free-roaming cellular connectivity to devices, rather than relying solely on patients' own internet connections, is an effective way to allay concerns about incurred costs. In all cases, user experience testing is key to validating assumptions about the target population. For example, one might have assumed that older patients would struggle with remote monitoring technology. In a comparison between 316 patients aged over 75 years and 541 younger patients using remote monitoring equipment, the older group did indeed rate themselves as more likely to avoid technology or delay adoption (82% vs 56%) and scored the technology as marginally less easy to use (ease-of-use score 5.5 vs 6.2) [27]. However, the older patients exhibited higher adherence to wearable use (95.3% vs 93.3%; $P<.001$) and equivalent adherence to remote monitoring tasks such as blood pressure measurement. The obvious implication was that patients should not be excluded because of age, but the more nuanced finding was that patient education should be focused on building confidence in the technology as much as skill acquisition.

Once a candidate technology is chosen, how it will work in the broader ecosystem should be considered, including interoperability, cybersecurity, and the flow of aggregated data into the EHR. Specifically, this may include procedures to annotate or amend data flowing from the home into the EHR if the data are missing or erroneous. This is not always straightforward. As Figure 1 illustrated, there is a compromise between accuracy and acceptability: no wearable device will

XSL•FO
RenderX

ever be completely accurate all the time and may occasionally transmit incorrect or artifactual data to the medical record. This is also true of manual measures in the hospital, which have considerable interobserver variability, but these are more of a "cultural norm," and in the home, there are usually no additional measures to corroborate the data from the wearable [19]. One approach is to annotate remotely collected vital signs with a measure of confidence that the vital signs accurately reflect the patient's physiology, based on the strength and regularity of the waveform data, the degree of interference (eg, from movement), and the percentage of data available in each window of time.

Set Expectations Early and Educate Throughout the Program

Before implementing remote monitoring, it is important to educate hospital clinicians to expect different and more variable data. Gaps in continuous data from the home are normal. Perfect accuracy or coverage is not expected (or necessary) to make good decisions, even in acute care. Our nursing team has been able to safely monitor 5515 patients at home with a median (IQR) adherence to continuous monitoring devices of 79% (77%-81%; defined as the percentage time data transmitted over length of stay)—in other words, with approximately 20% of data "missing." This remains the case even for the most vulnerable populations: in a cohort of 40 patients with hematological malignancies receiving chimeric antigen receptor T-cell (CAR-T) therapy managed acutely at home (median length of stay 14.7, IQR 10.4-27.6 days), overall adherence to continuous monitoring devices was 79% (IQR 68%-88%) [28]. Median (IQR) survey completion was 50% (7.7%-80%) and blood pressure adherence was 95% (69.8%-100%). A total of 25 patients developed cytokine release syndrome while at home but were safely managed, despite incomplete data. We advise repeated training sessions as clinicians gain familiarity with this new stream of data. Patients and caregivers should be trained to use the equipment properly, accounting for their knowledge, motivation, skills, systems, and behaviors [13]. If the collected vital signs are visible to patients and carers, then a sufficient degree of education and annotation should be insured so they can react constructively to their data without undue concern [25].

Prescribe a Dose of Monitoring

It is recommended to move away from the hospital practice of applying "standard monitoring." Instead, clinicians should be encouraged to consider home monitoring as a diagnostic intervention with downstream clinical, operational, and financial consequences. An appropriate "dose" of monitoring should be prescribed to capture meaningful change. A monitoring prescription might be used for intermittent or continuous data capture, specifying a bespoke cadence of data collection, adherence, duration, alarm thresholds, escalation, and discharge criteria. The prescription may need to be adjusted over time to reflect the patient's clinical trajectory. As home monitoring matures, we can evolve dose-response curves to titrate monitoring across different use cases. This dovetails with the previous point regarding education and setting expectations. In the outpatient CAR-T program described above, 30 patients triggered a total of 670 clinical alarms (0.5 per patient-day) over 4 months of monitoring [28]. These alarms prompted 1102 calls by monitoring nurses to patients, 63% of which were during nonclinic hours. The alarms were set to trigger when the median of a particular vital sign, aggregated over 30 or 60 minutes, breached a set threshold. When a 15-minute aggregation window was tried in an analogous program, 6 patients triggered a total of 436 alarms in 1 month without a notable difference in outcomes to the first program.

Individualize Alarms

It is beneficial to individualize alarms in the home. In the hospital, a patient's baseline vital signs, especially their variability, will be suppressed as they rest in bed. In the home, there will be more variation. We advise using alarm thresholds and trends tailored to individual patients and based on their clinical condition, risk, and social circumstances, rather than generic hospital "normal" values [29]. Their baseline values should be established early in their home stay, and then the focus should be on deviations from that baseline, rather than absolute values [30]. This recommendation is perhaps best illustrated with an inpatient example, where a "gold standard" comparator (standard nursing care) can be used. In a study of 28 patients receiving CAR-T therapy in the hospital with a remote monitoring device worn as a "black box," a simple alarm set at baseline temperature plus 2 SDs allowed the detection of temperature disturbances indicative of cytokine release syndrome a median of 184 (IQR 53-312) minutes earlier than standard care [31]. By comparison, the use of a fixed temperature threshold only improved detection by a median of 22 (IQR –43 to 113.5) minutes without a significant increase in specificity (Figure 4 [32]). Using trend-based measures is also an approach to deal with the variability in the peripheral devices described in Figure 3. While Figure 3 concerned blood pressure, in a systematic review and meta-analysis of 75 studies, Niven et al [33] demonstrated strikingly broad limits of agreement between peripheral and central measures of temperature, especially among adults with fever (–1.44 to 1.46 °C). In other words, peripheral measures of temperature in 2 patients with the same core temperature could vary by ~3 °C. This supports moving to a trend-based diagnosis from the fixed hospital threshold of 38 °C in febrile illness.

XSL•FO
RenderX

Figure 4. Axillary temperature plot of a patient receiving immunotherapy in the 24 hours before a diagnosis of CRS. The green dashed line represents an alarm of individual baseline temperature plus 2 SDs (Ti), and the pink line represents a fixed alarm at 38 °C (Tf). The arrow mark is the time the nurse recorded CRS onset (SoC). The use of a continuous monitoring system led to earlier detection than standard nursing care, as did an individualized threshold compared to a fixed threshold of 38 °C. CRS: cytokine release syndrome; SoC: standard of care.

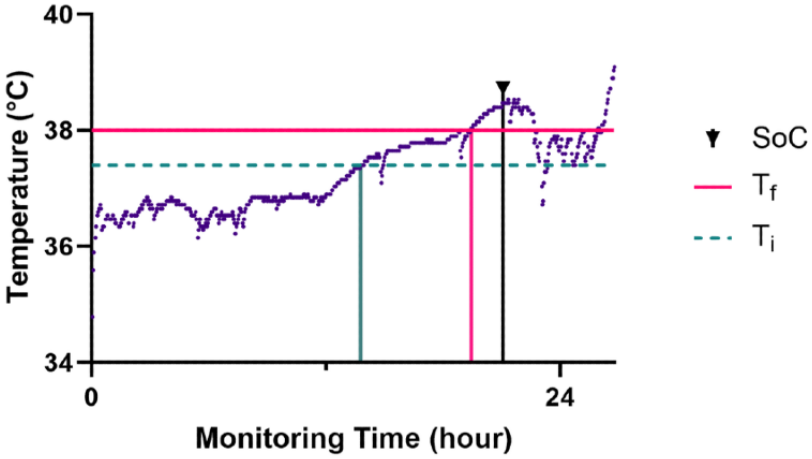

Balance Sensitivity, Specificity, Risk, and Capacity to Respond

In the hospital, we can use alarms that are purely sensitive (high risk of false positives and low risk of false negatives). Alarm fatigue is an issue, but it is straightforward to examine the patient following an alarm [34]. In contrast, a false alarm in the home may entail waking the patient or caregiver or dispatching a paramedic unnecessarily. We must therefore introduce more specificity. However, if we push either sensitivity or specificity too far, there will be a loss of trust from the patients and clinicians, financial penalties, and degradation of clinical care, as in the oncology example above (Figure 5).

We must find a balance of sensitivity and specificity, while being mindful of our clinical risk tolerances and our capacity to respond to alarms. All decisions should be based on new and variable data. Accepting this balance allows us to give patients space to recover in their own home and grants us a rich stream of diagnostic information. It is the crux of the mindset shift to home monitoring. It also requires establishing a feedback loop to differentiate useful and nonuseful alarms, adjusting the balance of sensitivity and specificity accordingly. This titration can be challenging, as there may be a gap between an alarm sounding and its utility being understood. Furthermore, those monitoring the data, assessing the patient, and modifying alarms might all be different individuals in different locations. In all cases, escalation criteria should be agreed in advance, being mindful that clinicians receiving the referral may not be intimately familiar with the nuances we have described above. We suggest erring toward sensitive settings at the start and increasing specificity as confidence grows.

Figure 5. Balancing sensitivity and specificity with clinical risk.

XSL•FO
RenderX

Conduct "Rounds" Regularly on Patients

It is important to not rely solely on alarms. Instead, teams should conduct "rounds" on a virtual care patient's data as often as their condition warrants to become familiar with the patterns in their vital signs. Our monitoring team conducts a "round" on patients every 4-6 hours, as they might on the hospital floor. A "round" involves reviewing the patient's vital signs and any asynchronous communications. This helps nurses establish patterns that differentiate the routine from the pathological, as well as picking up issues like incorrect device wear. Nurses also use surrogates for examining a patient. For example, compensatory changes in other vital signs, measures of motion, postural change, speed of vital sign recovery following activities of daily living, and adherence to monitoring all paint a picture of a patient in the home. These observations can then be fed back to patients' care teams during multidisciplinary "huddles" [35]. Our nurses have anecdotally reported that hypoxia and rising temperatures are most frequently spotted early by rounding, particularly when the patient's health care providers have chosen alarms that maximize specificity. As one nurse described:

> *During rounding on one hospital at home patient, I noted her SpO2 had been gradually decreasing - but no other vital sign was doing the same [and she had not yet triggered an alarm]. I called her and she was short of breath…she no idea what to do with herself. I called the responsible clinician and was told to place her supplementary oxygen on immediately. I did this and stayed on with her until we had a sustained acceptable spot check (and she wasn't huffing and puffing). She was educated by site the next day on supplementary oxygen use.*

Acknowledge the Stressors and Set Appropriate Standards of Care

Above all, we must recognize the learning curve and acknowledge the initial uncertainty and stress for care teams with this new stream of data. Clinicians should be monitored as they develop meaningful relationships with their RPM patients [36]. With time, they will develop confidence to not disturb patients, such as the patient referred to in Figure 2, with every fluctuation in vital signs—even in the acute setting. While desirable, this is also medicolegally vulnerable position, as the clinicians would opt not to respond to values traditionally considered "abnormal" in hospital, to give patients the space to recover. Malpractice is judged as deviation from an accepted standard of care. While there are now national figures for escalation and mortality rates (6.2% and 0.5%, respectively, in the United States), there are no standards of care for responding to data from the home [2,37]. If we default to hospital standards of care (rapid response to every change in vital signs and alerting senior staff), we either undermine care at home by overcontacting patients or risk a monitoring care team being in legal jeopardy for appearing to be unduly lax. So, we must set new standards for care at home, supporting clinicians in their decisions, building facilities for them to easily document their judgment calls, and constructing achievable escalation pathways around them. Standards and governance need to develop alongside a program as it evolves, with experienced clinicians mentoring those joining the program. Given all the limitations and subtleties we have described, we advise caution in entrusting too much triage, too early, to machine learning alone.

Consider Data Ownership, Evaluation, and Labeling Before Deployment

There is often insufficient clarity around data ownership in an RPM program. There is a need to establish early on what data belong to the patient, the technology provider, and the health system [25]. This should be considered for both identifiable and deidentified data. This is particularly pertinent in the context of program evaluation and health care research. The most useful data are typically those labeled with symptoms or behavior by the patient, with quality and completeness by the technology provider, and with clinical context by the health system. If all 3 cannot be physically and legally united, then evaluators and researchers would not hold all pieces of the puzzle.

Conclusions

We all recognize that hospitals are often not ideal healing environments due to their architecture and processes (bright lights, other patients, vital checks, and overhead codes) and that shifting acute care to the home may positively impact patient satisfaction, clinical outcomes, and health care economics. However, the home can be as noisy as the hospital, only for different reasons. Patients' vital signs are products of these environments, and as clinicians and researchers, we must interpret them in their proper context. Monitoring patients in the home where they fall ill, recover, and thrive is a powerful tool to understand health and disease, but differences to hospital monitoring must be understood, expected, acknowledged, and—we believe—celebrated in the next 25 years of RPM's evolution.

Acknowledgments

We gratefully acknowledge Tracey Baker, Nick Calafat, Addie Eversmeyer-Marquez, Kim Hayden, Dan McQueen, Jason Melnyk, Deidre Rolli, Katie Williams, and Adam Wolfberg for sharing their experience of remote patient monitoring implementation and critically reviewing this manuscript.

Conflicts of Interest

All authors are employees of Best Buy Health, receiving salary, and have stock in Best Buy Inc.

References

XSL•FO
RenderX

1. Carpentier WR, Charles JB, Shelhamer M, Hackler AS, Johnson TL, Domingo CMM, et al. Biomedical findings from NASA's Project Mercury: a case series. NPJ Microgravity 2018;4:6 [FREE Full text] [doi: 10.1038/s41526-018-0040-5] [Medline: 29560415]
2. Levine DM, Souza J, Schnipper JL, Tsai TC, Leff B, Landon BE. Acute hospital care at home in the United States: the early national experience. Ann Intern Med 2024 Jan;177(1):109-110 [FREE Full text] [doi: 10.7326/M23-2264] [Medline: 38190713]
3. Leff B, Burton L, Mader S, Naughton B, Burl J, Clark R, et al. Satisfaction with hospital at home care. J Am Geriatr Soc 2006 Sep 07;54(9):1355-1363. [doi: 10.1111/j.1532-5415.2006.00855.x] [Medline: 16970642]
4. de Jong L, Zeidler J, Damm K. A systematic review to identify the use of stated preference research in the field of older adult care. Eur J Ageing 2022 Dec 07;19(4):1005-1056 [FREE Full text] [doi: 10.1007/s10433-022-00738-7] [Medline: 36506676]
5. Levine DM, Ouchi K, Blanchfield B, Saenz A, Burke K, Paz M, et al. Hospital-level care at home for acutely ill adults: a randomized controlled trial. Ann Intern Med 2020 Jan 21;172(2):77-85. [doi: 10.7326/M19-0600] [Medline: 31842232]
6. Vegesna A, Tran M, Angelaccio M, Arcona S. Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed J E Health 2017 Jan;23(1):3-17 [FREE Full text] [doi: 10.1089/tmj.2016.0051] [Medline: 27116181]
7. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
8. Evans D, Hodgkinson B, Berry J. Vital signs in hospital patients: a systematic review. Int J Nurs Stud 2001 Dec;38(6):643-650. [doi: 10.1016/s0020-7489(00)00119-x] [Medline: 11602268]
9. Caplan GA, Coconis J, Board N, Sayers A, Woods J. Does home treatment affect delirium? a randomised controlled trial of rehabilitation of elderly and care at home or usual treatment (The REACH-OUT trial). Age Ageing 2006 Jan;35(1):53-60. [doi: 10.1093/ageing/afi206] [Medline: 16239239]
10. Arsenault-Lapierre G, Henein M, Gaid D, Le Berre M, Gore G, Vedel I. Hospital-at-home interventions vs in-hospital stay for patients with chronic disease who present to the emergency department: a systematic review and meta-analysis. JAMA Netw Open 2021 Jun 01;4(6):e2111568 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11568] [Medline: 34100939]
11. Lawrence K, Singh N, Jonassen Z, Groom LL, Alfaro Arias V, Mandal S, et al. Operational implementation of remote patient monitoring within a large ambulatory health system: multimethod qualitative case study. JMIR Hum Factors 2023 Jul 27;10:e45166 [FREE Full text] [doi: 10.2196/45166] [Medline: 37498668]
12. Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. Ann Emerg Med 2002 Mar;39(3):233-237. [doi: 10.1067/mem.2002.122017] [Medline: 11867974]
13. Eze CE, Dorsch MP, Coe AB, Lester CA, Buis LR, Farris KB. Facilitators and barriers to blood pressure telemonitoring: a mixed-methods study. Digit Health 2023 Jul 25;9:20552076231187585 [FREE Full text] [doi: 10.1177/20552076231187585] [Medline: 37529536]
14. Lee H, Ko H, Lee J. Reflectance pulse oximetry: practical issues and limitations. ICT Express 2016 Dec;2(4):195-198. [doi: 10.1016/j.icte.2016.10.004]
15. Dressler R, Dryer MM, Coletti C, Mahoney D, Doorey AJ. Altering overuse of cardiac telemetry in non-intensive care unit settings by hardwiring the use of American Heart Association guidelines. JAMA Intern Med 2014 Nov;174(11):1852-1854. [doi: 10.1001/jamainternmed.2014.4491] [Medline: 25243419]
16. Leenen JPL, Leerentveld C, van Dijk JD, van Westreenen HL, Schoonhoven L, Patijn GA. Current evidence for continuous vital signs monitoring by wearable wireless devices in hospitalized adults: systematic review. J Med Internet Res 2020 Jun 17;22(6):e18636 [FREE Full text] [doi: 10.2196/18636] [Medline: 32469323]
17. Waleed M, Kamal T, Um TW, Hafeez A, Habib B, Skouby KE. Unlocking insights in IoT-based patient monitoring: methods for encompassing large-data challenges. Sensors (Basel) 2023 Jul 28;23(15):6760 [FREE Full text] [doi: 10.3390/s23156760] [Medline: 37571543]
18. Zahradka N, Geoghan S, Watson H, Goldberg E, Wolfberg A, Wilkes M. Assessment of remote vital sign monitoring and alarms in a real-world healthcare at home dataset. Bioengineering (Basel) 2022 Dec 28;10(1):37 [FREE Full text] [doi: 10.3390/bioengineering10010037] [Medline: 36671610]
19. Kallioinen N, Hill A, Horswill MS, Ward HE, Watson MO. Sources of inaccuracy in the measurement of adult patients' resting blood pressure in clinical settings: a systematic review. J Hypertens 2017 Mar;35(3):421-441 [FREE Full text] [doi: 10.1097/HJH.0000000000001197] [Medline: 27977471]
20. Honkalampi K, Järvelin-Pasanen S, Tarvainen MP, Saaranen T, Vauhkonen A, Kupari S, et al. Heart rate variability and chronotype - a systematic review. Chronobiol Int 2021 Dec 16;38(12):1786-1796. [doi: 10.1080/07420528.2021.1939363] [Medline: 34130562]
21. Bhogal AS, Mani AR. Pattern analysis of oxygen saturation variability in healthy individuals: entropy of pulse oximetry signals carries information about mean oxygen saturation. Front Physiol 2017 Aug 02;8:555 [FREE Full text] [doi: 10.3389/fphys.2017.00555] [Medline: 28824451]

XSL•FO
RenderX

22. Al Rajeh A, Bhogal AS, Zhang Y, Costello JT, Hurst JR, Mani AR. Application of oxygen saturation variability analysis for the detection of exacerbation in individuals with COPD: a proof-of-concept study. Physiol Rep 2021 Dec;9(23):e15132 [FREE Full text] [doi: 10.14814/phy2.15132] [Medline: 34851045]
23. Schutte AE, Kollias A, Stergiou GS. Blood pressure and its variability: classic and novel measurement techniques. Nat Rev Cardiol 2022 Oct 19;19(10):643-654 [FREE Full text] [doi: 10.1038/s41569-022-00690-0] [Medline: 35440738]
24. Miranda R, Oliveira MD, Nicola P, Baptista FM, Albuquerque I. Towards a framework for implementing remote patient monitoring from an integrated care perspective: a scoping review. Int J Health Policy Manag 2023 Jan 30;12:7299 [FREE Full text] [doi: 10.34172/ijhpm.2023.7299] [Medline: 37579426]
25. Ginsburg GS, Picard RW, Friend SH. Key issues as wearable digital health technologies enter clinical care. N Engl J Med 2024 Mar 21;390(12):1118-1127. [doi: 10.1056/NEJMra2307160] [Medline: 38507754]
26. Jerjes W, Harding D. Telemedicine in the post-COVID era: balancing accessibility, equity, and sustainability in primary healthcare. Front Digit Health 2024 Aug 21;6:1432871 [FREE Full text] [doi: 10.3389/fdgth.2024.1432871] [Medline: 39233772]
27. Pugmire J, Wilkes M, Kramer A, Zaniello B, Zahradka N. 2225 Geriatric patients in virtual wards; technology adoption style, ease of use, and adherence to remote monitoring technology. Age Ageing 2024 Aug 8;53(Supplement_3):afae139.080. [doi: 10.1093/ageing/afae139.080]
28. Cox T, Zahradka N, Billups RL. Standardization of outpatient care after CAR-T therapy across a large cell therapy network-through technology and decentralized virtual nurses: preliminary results. Blood 2023 Nov 2;142(Supplement 1):254. [doi: 10.1182/blood-2023-187253]
29. van Rossum MC, Vlaskamp LB, Posthuma LM, Visscher MJ, Breteler MJM, Hermens HJ, et al. Adaptive threshold-based alarm strategies for continuous vital signs monitoring. J Clin Monit Comput 2022 Apr 11;36(2):407-417 [FREE Full text] [doi: 10.1007/s10877-021-00666-4] [Medline: 33575922]
30. Brekke IJ, Puntervoll LH, Pedersen PB, Kellett J, Brabrand M. The value of vital sign trends in predicting and monitoring clinical deterioration: a systematic review. PLoS One 2019 Jan 15;14(1):e0210875 [FREE Full text] [doi: 10.1371/journal.pone.0210875] [Medline: 30645637]
31. Rajeeve S, Zahradka N, Wilkes M, Pan D, Calafat NJ, Serebryakova K, et al. Early and consistent CRS detection using wearable device for remote patient monitoring following CAR-T therapy in relapsed/refractory multiple myeloma (RRMM): early results of an investigator-initiated trial. Blood 2023 Nov 2;142(Supplement 1):1007. [doi: 10.1182/blood-2023-181936]
32. Rajeeve S, Zahradka N. Moving oncology treatment home with remote monitoring [Podium presentation]. 2023 Presented at: DPharm Conference; September 20-22,2023; Boston, MA.
33. Niven DJ, Gaudet JE, Laupland KB, Mrklas KJ, Roberts DJ, Stelfox HT. Accuracy of peripheral thermometers for estimating temperature: a systematic review and meta-analysis. Ann Intern Med 2015 Nov 17;163(10):768-777. [doi: 10.7326/M15-1150] [Medline: 26571241]
34. National patient safety goals. The Joint Commission. 2021. URL: https://tinyurl.com/33rn2sws [accessed 2024-10-02]
35. Mashaw A, Johnson E, Shulman E, Rastogi R, Varner K, Womack J, et al. Advanced care at home at scale in an integrated health care system. Am J Manag Care 2023 Dec 01;29(12):e357-e364 [FREE Full text] [doi: 10.37765/ajmc.2023.89469] [Medline: 38170526]
36. Pugmire J, Wilkes M, Wolfberg A, Walters R, Zahradka N. A structured qualitative evaluation finds that nurses experience an improved sense of connection with their patients in a virtual COVID-19 pilot program. 2022 Oct 27 Presented at: 2022 American Academy of Home Care Medicine Annual Meeting; October 28-29, 2022; Orlando, FL URL: https://tinyurl.com/mwdre9sm [doi: 10.2196/37567]
37. Brody AA, Dorfman E, Caspers CG, Sadarangani TR. What's next for Hospital at Home Programs in the United States: a clarion call for permanent, person-centered solutions. J Am Geriatr Soc 2023 Jan 02;71(1):11-14 [FREE Full text] [doi: 10.1111/jgs.18089] [Medline: 36321658]

Abbreviations

CAR-T: chimeric antigen receptor T-cell
CE: Conformité Européenne
EHR: electronic health record
RPM: remote patient monitoring

XSL•FO
RenderX

Edited by G Eysenbach; submitted 27.03.24; peer-reviewed by J Goh, S Peterson; comments to author 07.08.24; revised version received 26.08.24; accepted 25.09.24; published 11.10.24.

Please cite as:
Wilkes M, Kramer A, Pugmire J, Pilkington C, Zaniello B, Zahradka N
Hospital Is Not the Home: Lessons From Implementing Remote Technology to Support Acute Inpatient and Transitional Care in the Home in the United States and United Kingdom
J Med Internet Res 2024;26:e58888
URL: https://www.jmir.org/2024/1/e58888
doi:10.2196/58888
PMID:39331537

©Matt Wilkes, Annabel Kramer, Juliana Pugmire, Christopher Pilkington, Benjamin Zaniello, Nicole Zahradka. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Primary Care Informatics: Vitalizing the Bedrock of Health Care

Jacqueline Guan-Ting You[1,2], MD; Tiffany I Leung[3,4], MD, MPH; Deepti Pandita[5], MD; Matthew Sakumoto[6], MD

[1]Clinical Informatics, Mass General Brigham, Somerville, MA, United States
[2]Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States
[3]Department of Internal Medicine (adjunct), Southern Illinois University School of Medicine, Springfield, IL, United States
[4]JMIR Publications, Inc., Toronto, ON, Canada
[5]Department of Medicine, University of California Irvine, Irvine, CA, United States
[6]Department of Medicine, University of California San Francisco, San Francisco, CA, United States

Corresponding Author:
Matthew Sakumoto, MD
Department of Medicine
University of California San Francisco
533 Parnassus Avenue, U127
San Francisco, CA, 94143
United States
Phone: 1 4154761000
Email: matthew.sakumoto@ucsf.edu

Abstract

Primary care informatics (PCI) professionals address workflow and technology solutions in a wide spectrum of health, ranging from optimizing the experience of the individual patient in the clinic room to supporting the health of populations and augmenting the work of frontline primary care clinical teams. PCI overlaps uniquely with 2 disciplines with an impact on societal health—primary care and health informatics. Primary care is a gateway to health care access and aims to synthesize and coordinate numerous, complex elements of patients' health and medical care in a holistic manner. However, over the past 25 years, primary care has become a specialty in crisis: in a post–COVID-19 world, workforce shortages, clinician burnout, and continuing challenges in health care access all contribute to difficulties in sustaining primary care. Informatics professionals are poised to change this trajectory. In this viewpoint, we aim to inform readers of the discipline of PCI and its importance in the design, support, and maintenance of essential primary care services. Although this work focuses on primary care in the United States, which includes general internal medicine, family medicine, and pediatrics (and depending on definition, includes specialties such as obstetrics and gynecology), many of the principles outlined can also be applied to comparable health care services and settings in other countries. We highlight (1) common global challenges in primary care, (2) recent trends in the evolution of PCI (personalized medicine, population health, social drivers of health, and team-based care), and (3) opportunities to move forward PCI with current and emerging technologies using the 4Cs of primary care framework. In summary, PCI offers important contributions to health care and the informatics field, and there are many opportunities for informatics professionals to enhance the primary care experience for patients, families, and their care teams.

(J Med Internet Res 2024;26:e60081) doi:10.2196/60081

KEYWORDS

health care delivery; primary care; primary health care; primary prevention; quality of health care; holistic care; holistic medicine; people-centric care; person-centric care; medical informatics applications; primary care informatics; medical informatics; health informatics; information science; data science

Introduction

Introduction to Primary Care

Primary care is defined by the World Health Organization as "a model of care that supports first-contact, accessible, continuous, comprehensive and coordinated person-focused care" while achieving equitable access to care and emphasizing population health [1]. Primary care is one of the only sections of health care shown to have positive impacts on health outcomes and population health, emphasizing its fundamental role in the health and well-being of the public [2]. Primary care specialties can include internal medicine, family medicine, and

XSL•FO
RenderX

pediatrics [3,4]; some definitions also include obstetrics and gynecology [5,6].

Primary care is the backbone of health care and is crucial to the health of populations. One international survey indicates that about 75% of respondents had a usual source of primary care [7]. Recent US data indicate that 55% of public insurance (Medicaid) beneficiaries primarily visited primary care clinicians and 36% of patients predominantly visited specialists [8]. Having a usual source of care, such as primary care, is associated with increased preventive screenings such as blood pressure measurements and diabetes and cholesterol screening [7]. Access to primary care has been correlated with lower avoidable hospitalization rates [9]. Loss of primary care has been associated with increased utilization of specialty, urgent care, and emergency care among older adults [10].

Stange et al [11] refer to primary care as an essential bridge between the health of individuals and population health, "situated at the unsettled boundary region between the messy but meaningful lives of individuals, families, and communities, and the health care, public health, and social systems designed to support them." Areas where primary care can collaborate with public health and play a role in population health include but are not limited to chronic disease management, communicable diseases, and maternal-child health [12]. This role as a mediator also includes facilitating care with medical specialists and social services [13]. Because of its unique position in facilitating numerous interactions between different parties, primary care serves as a central hub for various data streams and care workflows; informatics solutions in this discipline therefore must match the field's unique needs.

Introduction to Primary Care Informatics

Primary care informatics (PCI) augments the crucial work that primary care provides to populations. In 2003, the Primary Care Informatics Working Group under the American Medical Informatics Association defined PCI as the "application of information technology to improve the practice, education, and research of primary care, and the development of informatics tools appropriate for primary care practice" [14]. At the time, PCI was emerging globally to address the use of technology in areas such as chronic disease management. de Lusignan [15] argued that PCI was not just a discipline of technology but rather a "science" dedicated to the promotion of "patient-centered primary medical care" with a biopsychosocial model of health; PCI was distinct from other areas of health informatics due to the use of heuristics-based clinical management and patient-centered consultations. The Primary Care Informatics Working Group of the International Medical Informatics Association was formed in the early 2000s [15], with the aims of addressing real-world primary care data and ontologies for research; technology integration in primary care; and ethical, legal, and social issues related to primary care [16].

PCI professionals bring unique expertise at the nexus of primary care and informatics, well before the COVID-19 pandemic. We define PCI professionals as inclusive of clinicians, researchers, technology professionals, experts, and others who engage in PCI. A few examples of PCI issues included the impact of electronic health records (EHRs) on primary care [17-19], benefits of health information exchanges in primary care [20], evidence on health care workers' perceptions of mobile health in primary care [21], and risks and benefits of artificial intelligence (AI) in primary care [22]. The COVID-19 pandemic put a spotlight on the central role of primary care clinicians globally to mobilize and continue to provide access to care and public health care services to populations all over the world during lockdowns and additional restrictions. PCI professionals rapidly deployed and scaled up telehealth visits and identified and addressed inequities in the digital divide (such as those due to internet access) [23]. Telehealth and telemedicine positively impacted primary care by providing improvements in key areas including continuity, accessibility, care coordination, comprehensiveness, access, better appointment times, and efficiency [24]. In response to the COVID-19 pandemic, additional opportunities were highlighted to further strengthen systems, including vaccination and disease prevention, disease management, public health surveillance, and pandemic preparedness [25].

In this viewpoint paper, we highlight issues uniquely serviced by PCI, distinct from health informatics as a whole, discussing aspects of PCI that have evolved since the early 2000s to mark the 25th anniversary of *JMIR Publications* and the *Journal of Medical Internet Research*. In particular, in the wake of COVID-19, there have been changes to primary care practice pertinent to PCI, including challenges to clinician well-being and digital health transformation in domains such as telehealth [26].

Common Challenges in Primary Care

Challenges may vary for different regions of the same country; for example, in the United States, access to primary care for adults [27,28] and pediatric patients [28] is more challenging in rural than urban settings. Differences in primary care worldwide range from physician consultation time [29] to funding and structure [30]. Countries and geographic regions have different needs in primary care—for example, Japan has universal health care and an aging population with challenges around chronic disease management [31], while in sub-Saharan Africa, the disease burden consists of both chronic diseases as well as infectious diseases [32]. In many primary care practices, unlike most specialties, outcomes are addressed at both the individual patient and population levels; at times, these perspectives are not always aligned, for example, when individual values or beliefs may differ from the population or public health recommendations, or when guidelines do not fit a patient's clinical presentation or health conditions. Yet, the responsibility of addressing these differences in shared decision-making can fall to primary care clinicians to reconcile with patients.

A recent report on US primary care identified the lack of usability and interoperability of digital health tools as barriers to health care organizations and clinicians unlocking the full possibilities of health information technology—"real potentials of digital health to aggregate a wide array of medical, environmental, biological, and social data; make meaningful sense of information; automate care; make care proactive and

XSL•FO
RenderX

not reactive; enhance health equity; and enable population health monitoring and management have been barely explored" [2]. Similarly, recent international physician survey data suggest the majority of primary care physicians are not able to adequately coordinate care with specialists and hospitals [33], in part due to poor data access and communication, a gap that PCI professionals can help address.

Also, there are troubling themes in recent primary care challenges that cross borders, such as clinician burnout [34], aging populations [31], chronic disease prevalence [32], significant administrative work [35], and mismatch of patient demand and clinician supply [32], which in turn threaten the health of communities globally. However, the authors are optimistic that greater awareness of the opportunities and changes possible through intentional applications of PCI, building on the last 25 years of evidence-based advancements and development in the field, can pave the way forward for continued primary care, and consequently, overall health system improvements. The authors share perspectives reflecting their current and prior primary care experience in the United States, Europe, and Asia, although they expect that some themes described may be applicable to certain health care systems and settings elsewhere.

Evolution of PCI

Overview

PCI has evolved to address some of these new challenges. de Lusignan [15] highlighted 3 core areas of PCI, including heuristic decision-making, patient-centered care, and the biopsychosocial model. Building on the principles de Lusignan [15] highlighted, recent trends in primary care with PCI at its core include personalized medicine, analytics- and data-driven population health, social drivers of health, and team-based care. Heuristic-based care is now bolstered by point-of-care analytics and data-driven decisions derived from population health analytics; patient-centered care is informed by personalized medicine; and the biopsychosocial model is strengthened by improved awareness and screening of social determinants. The planted seeds of PCI in the early 2000s have grown and flourished with new technologies and new models of primary care.

Personalized Medicine, Augmented Personalized Decision-Making, and Customization of Care

Primary care clinicians often engage in shared decision-making—described in one model as a four-step process of (1) recognizing the opportunity to make a decision with patient and clinician, (2) introducing a decision aid, (3) having a conversation, and (4) receiving care [36]. This decision-making allows for more personalized and arguably patient-centered care. Personalized medicine has the potential to provide new resources and tools to facilitate shared decision-making.

The growth of genetics knowledge has offered new opportunities to consider how genomic medicine may potentially be integrated into primary care—for example, a real-world pilot with large gene sequencing panels in primary care [37] versus the integration of genetics counselors into primary care teams [38]. Furthermore, the nexus of precision medicine and AI will enable improved care—such as more precise medication dosing, improved treatment recommendations based on comorbidities, or improved risk prediction for diseases [39]—areas in which PCI professionals can inform research and subsequent practical implementation. AI also has the potential to blur the lines between personalized medicine and population health—for example, with precision population analytics, creating "precision cohorts" similar to a primary care clinician's patients can help better inform which treatment option to use based on population-level data [40].

Digital health tools, such as wearables and mobile health apps, can provide patient-centered insights and shift primary care toward proactive yet targeted "precision prevention" [41]. These technologies can drive customized care with shared decision-making and allow primary care to pivot toward primary prevention and not just chronic disease management. Along with the personalized treatment of health conditions, patients are better able to customize their care experience; patients are more empowered to access care flexibly outside of brick-and-mortar clinics through mediums such as telehealth [42] and patient portal messages [43].

Population-Based Approach Driven by Big Data and Analytics

Primary care is uniquely positioned to become a potential hub for better management of the vast amounts of unused health care data across clinical settings. Each hospital produces an average of 50 petabytes of data annually with an estimated 97% of data remaining unused [44]. EHRs, wearables, and other technologies have expanded modalities [45] and the volume of health data [46]. Health information exchanges can augment the data of individual systems, powering descriptive, predictive, and prescriptive analytics to facilitate quality, improved risk management and clinical decision-making at the population level [47]. Primary care, where longitudinal chronic care and multidisease management take place, can be a vital central hub for patient and population health data. Wagner's [48] widely adopted chronic care model includes information systems as a core component of providing high-quality chronic disease care. Data-driven improvement using computer-based systems is also considered a building block for a high-performing primary care practice [49]. While extensive research on predictive machine learning has been performed in areas such as cardiac disease, mental health, communicable diseases, and perinatal outcomes, work needs to be done to take into account model usability in the real world [50], as well as practical considerations such as external validation for pragmatic clinical decision-making [51].

Furthermore, as technologies, like generative AI, are implemented across health care [52,53], PCI professionals have and continue to lead the charge in steering the primary care and informatics communities toward equitable [54-56] and ethical [57] approaches to implementing these technologies. For example, experts spanning PCI, policy, bioethics, and philosophy for the largest US medical specialty organization [58] recently called for ethical, accountable, equitable, transparent, and environmentally aware use of AI in health care

XSL•FO
RenderX

[59]. PCI professionals can advocate for appropriate safeguards balanced with clinically reasonable regulations to avoid some of the pitfalls associated with current EHR administrative work [60].

Population health can encompass chronic disease management—examples of PCI initiatives in this area include chronic care management and remote patient monitoring programs, which provide continuous support across the care continuum [61]. Remote patient monitoring is most often applied in chronic disease management for cardiology, endocrinology, pulmonary, and geriatrics care [62]; beyond expanding the interface modalities and underlying AI models [63] powering this technology, PCI can add value by improving the patient and clinician experience with these technologies [64]. Even as analytics and associated technologies for addressing population health continue to evolve, PCI professionals can help ground these technologies in a sociotechnical design or implementation approach with appropriate outcome measures (such as patient-reported outcome measures) [65].

Social Drivers of Health

As part of comprehensive, culturally sensitive, and patient-centered care, PCI professionals address social determinants of health (SDOH), also referred to as social drivers of health, in EHR design or improvements [66-68]. SDOH includes areas such as education, food, housing, income, and other nonmedical factors that contribute to health [69]. Recent attention to these drivers has advanced the traditional biopsychosocial model of care. The longitudinal relationship of patients with primary care clinicians, as opposed to the episodic nature of other specialties, allows for greater patient trust in disclosing SDOH information during primary care visits. Documenting and visualizing SDOH gaps increases the likelihood of addressing these factors that greatly impact patients' well-being [70]. PCI professionals can identify and present resources to both patient and clinician, matching the right intervention with the right patient to maximize benefit to the patient and community.

Other examples of relevant determinants within SDOH also include commercial determinants of health (industry entities shaping health choices in areas such as food, alcohol, tobacco use, pharmaceuticals, and media consumption) [71,72]; political determinants of health (sociopolitical events correlating with disparities in health outcomes) [73]; and geographic location (using geographical information systems to measure disease incidence and prevalence, health risk, community access to health services, and community health profiling) [74]. In addition, health informatics researchers and PCI professionals play a role in addressing immigrant and refugee health, for example, in building appropriate record systems to accommodate the international needs of this vulnerable population [75]. There is a significant intersection between PCI and the emerging field of social informatics [66], which deals with data from a breadth of sources to advance social and medical care. Furthermore, as PCI implements patient-facing solutions (eg, in digital health), the field should aim for meaningful, equitable solutions built on community engagement [76], which can allow for culturally relevant informatics tools that engage and advocate for underrepresented populations [77].

Informatics to Enable Dynamic, Flexible Team-Based Care

Primary care teams can vary in composition but can include nurses, medical assistants, health coaches, behavioral health specialists, and pharmacists supporting primary care providers, often coupled with clinical workflows such as standing orders or protocols [78]. There is evidence that these team members can provide effective care in particular areas of primary care; for example, primary care nurses provide effective care in areas such as chronic disease management, care management, and health promotion [79]. Team-based primary care has been correlated with improved quality measures, such as depression and diabetes care adherence, and reduced emergency department visits and hospitalizations [80]. This team-based care is a key component of care management, supporting higher-risk patients with better disease management and care navigation; care management in turn is vital to promoting value-based-care—care that is patient-centered and improves clinical outcomes while avoiding unnecessary costs [81].

Health information technology has been shown to be an important part of effective team-based primary care. [82] For example, the advanced team-based care model integrates technology into a care team model with enhanced clinical workflows and roles, including a care team coordinator whose responsibilities include history taking during clinic visits, coordination of preventive health screenings, clinician-supervised laboratory test or order placement and documentation; this model improves patient outcomes such as cervical cancer screening and immunizations and allows physicians or advanced practice providers to focus on communication, clinical decision-making, and management of a greater number of health issues [83]. PCI professionals also can introduce new tools to team-based care such as secure messaging for interprofessional communication [84,85]. With increases in telehealth since the COVID-19 pandemic, PCI professionals have been core to establishing new care team models specific to telehealth, taking into account teams that are often remotely located [86].

Opportunities in PCI

Overview

A broad range of digital health interventions for prevention in primary care exist, ranging from EHR-based solutions to telehealth, that have proven efficacy but have not been widely adopted or are primarily clinician facing [87,88]. We break down current and future state technologies that have the potential to transform the primary care experience.

We apply the 4Cs of primary care (alternatively known as the 4 tenets or 4 pillars of primary care) framework [89,90] to different domains and tasks encountered in primary care and map these domains and tasks to health informatics tools (Table 1). Some of these domains may encompass multiple pillars.

XSL•FO
RenderX

Table 1. 4Cs of primary care, primary care domains and tasks, and corresponding informatics tools or aspects.

4 Cs of primary care and primary care domain or task	Health informatics tools or aspects
Contact	
Patient engagement and patient education	Mobile health apps, text messaging, telehealth, medical extended reality, personal health records, patient portals
Comprehensiveness	
Chronic disease management	Remote patient monitoring, EHR[a] design, telehealth visits, mobile health
Population health management	Clinical decision support, machine learning analytics, disease registries
Social determinants of health	Clinical decision support, EHR design
Coordination	
Care team or clinician-clinician coordination	Interoperability, EHR design, electronic consults
Documentation or administrative tasks	Ambient listening documentation, generative AI[b] chart summarization
Population health management	Precision population analytics, predictive analytics, health information exchange
Collaboration with public health	Data aggregation, interoperability, data reporting, learning health systems
Continuity	
Patient-clinician communication	Patient portal messaging, telehealth

[a]EHR: electronic health record.

[b]AI: artificial intelligence.

Contact

The first C, contact, refers to patients accessing primary care services. With the necessitated movement to telemedicine and telehealth services during the COVID-19 pandemic, telehealth adoption among certain patients was rapid [91]. After the pandemic, there is much work to be done to address gaps in telehealth access, especially among those of lower socioeconomic status [42].

Health care systems must rethink how individuals access health care, and PCI practitioners are poised to bridge gaps in access and creatively engage patients in care. Text messaging for medication adherence in chronic diseases [92] and mobile health have engaged patients with chronic health conditions [93,94]. On the other hand, medical extended reality is a growing field overlapping with clinical informatics; its applications span from behavioral health to integration with procedural specialties [95]. Patient education is a key component of primary care visits; virtual reality has been studied for patient education [96], and in the future, could be a tool in PCI for patient education and engagement. Deep learning can also improve patient instructions by “translating” language to an appropriate reading or health literacy level [97]. While benchmarking suggests large language models’ multilingual functions do not perform well enough compared to commercial translation services [98], large language models could someday create high-quality multilingual patient education materials.

Comprehensiveness

The second C, comprehensiveness, refers to patient-centered management of the breadth of conditions and a balance of prevention and treatment.

To provide comprehensive care, primary care clinicians and other care team members must be able to find the right information for the right patient (or even groups of patients) at the right time. Interoperable and usable EHRs are key to timely and appropriate data access. However, current EHR usability is often poor and associated with clinician burnout [99]. Similarly, a recent survey of US family physicians demonstrates poor satisfaction with interoperability in primary care [100]. PCI professionals, in collaboration with experts in these domains, can help facilitate improved clinical workflows, data infrastructure, and user-friendly designs to advance comprehensive care.

PCI professionals can also augment population health management, where clinical decision support has been shown to positively impact screenings for cancer, infectious diseases, and cardiovascular risk factors [101]. Health information management systems through health information exchanges have been associated with better care outcomes [102]; data analytics in primary care are correlated with decreased emergency department visits and increased care engagement [103]. Disease registries, such as those for chronic kidney disease [104], can assist primary care teams in better managing chronic diseases at a population level. Other digital health interventions can help with preventive care, such as immunization and cancer screening, and improve clinical outcomes [87]. As mentioned earlier, the integration of social drivers of health into informatics tools and clinical workflows will also improve patient-centered care.

Coordination

The third C, coordination, refers to bringing together different parts of the health care system and serving as a bridge between primary care and specialty care. Primary care serves as a contact point to other parts of the health system, including medical specialists and social services.

XSL•FO
RenderX

Care coordination has been positively correlated with improved care measures such as diabetes, cardiovascular, and cancer screening [105]. Electronic consults (eConsults) thus far have had mixed outcomes [106] but still have the potential to assist with coordination between primary care and specialists and address access to specialty care.

Care coordination is inextricably linked with communication. Interoperability and facile interfaces for communication can improve interactions between different care team members, and when coupled with timely and convenient access to data, are powerful in advancing patient care. National learning health systems have been posited as a framework to advance interoperability and digital health innovation [107] for the public good.

Coordination includes administrative tasks around patient care such as documentation of patient encounters. Generative AI solutions, such as ambient AI [108] and medical record summarization [109], have the potential to alleviate the documentation burden in this area and lead to more concise clinician-clinician communication.

Continuity

The fourth C, continuity, has highly variable definitions, especially in the context of team-based care, but tends to refer to the patient's ability to receive care that is centered around relationships and has a sense of longitudinality. Primary care continuity correlates with decreased patient mortality [110]. As mentioned previously, telehealth has provided new touchpoints for primary care practitioners and their patients. Patient portals [43] and patient access to notes [111] have both been shown to increase patients' understanding of their health, which arguably in turn can be an opportunity for primary care teams to engage patients in their health over time and build trust with patients. Continuity can also be extended to transitions of care between care settings that intersect with primary care, such as inpatient hospitalization, where there is a current dearth of patient-clinician communication tools [112]—this is an area where PCI can provide new bridges for continuity.

Discussion

Primary care is the bedrock of health care and PCI in turn can vitalize this bedrock—contributing to societal health and well-being by improving existing workflows or thoughtfully integrating new technologies into care. PCI melds the learnings of primary care and informatics but is also able to add distinctive value to both fields by augmenting the core tenets of primary care with biopsychosocial-driven technologies. At a time when the global primary care workforce is facing immense challenges, PCI professionals have a key role to play in supporting their primary care colleagues, disseminating best practices learned in primary care to other informatics colleagues, and adapting appropriate tools from other branches of biomedical and health informatics into primary care. PCI is evolving as care is delivered across a greater spectrum of settings and patients are empowered with new digital health tools to engage in their health. They can benefit from new opportunities and tools for proactive preventive care rather than reactive chronic disease management. Ultimately, PCI has been an effective area within primary care and will continue to be critical in driving improved patient and clinician experiences for safe, equitable, effective, patient-centered, and team-based care.

Acknowledgments

JGTY is supported by the National Library of Medicine/National Institutes of Health grant (T15LM007092). Ryan Jelinek contributed to the conceptualization and writing of an earlier version of this manuscript.

Authors' Contributions

JGTY and MS contributed to writing the original draft. JGTY, MS, TIL, and DP contributed to the conceptualization, writing, reviewing, and editing of the manuscript. JGTY contributed to project administration. MS contributed to supervision.

Conflicts of Interest

TIL is the scientific editorial director at JMIR Publications and reports consulting fees from Plushcare, Inc; however, TIL had no involvement in peer review, editorial review, or any aspects of editorial processing of this manuscript. MS reports personal fees from Clearstep Health, Carbon Health, Matter Health, and Nabla Health.

References

1. World Health Organization. Primary care. 2024. URL: https://www.who.int/teams/integrated-health-services/clinical-services-and-systems/primary-care [accessed 2024-02-25]
2. Robinson SK, Meisnere M, Phillips RLJ, McCauley L. Implementing High-Quality Primary Care: Rebuilding the Foundation of Health Care. Washington, DC: National Academies Press; 2021:45-62.
3. Primary care. American Academy of Family Physicians. 2024. URL: https://www.aafp.org/about/policies/all/primary-care.html [accessed 2024-07-13]
4. 2022 physician specialty data report executive summary. Association of American Medical Colleges. 2022. URL: https://www.aamc.org/data-reports/data/2022-physician-specialty-data-report-executive-summary [accessed 2024-07-13]
5. Brotherton SE, Rockey PH, Etzel SI. US graduate medical education, 2004-2005: trends in primary care specialties. JAMA 2005;294(9):1075-1082. [doi: 10.1001/jama.294.9.1075] [Medline: 16145028]

XSL•FO
RenderX

6. Jeffe DB, Whelan AJ, Andriole DA. Primary care specialty choices of United States medical graduates, 1997-2006. Acad Med 2010;85(6):947-958. [doi: 10.1097/ACM.0b013e3181dbe77d] [Medline: 20505392]
7. Croke K, Moshabela M, Kapoor NR, Doubova SV, Garcia-Elorrio E, HaileMariam D, et al. Primary health care in practice: usual source of care and health system performance across 14 countries. Lancet Glob Health 2024;12(1):e134-e144 [FREE Full text] [doi: 10.1016/S2214-109X(23)00513-2] [Medline: 38096885]
8. Romaire MA. Use of primary care and specialty providers: findings from the medical expenditure panel survey. J Gen Intern Med 2020;35(7):2003-2009 [FREE Full text] [doi: 10.1007/s11606-020-05773-7] [Medline: 32291713]
9. Rosano A, Loha CA, Falvo R, van der Zee J, Ricciardi W, Guasticchi G, et al. The relationship between avoidable hospitalization and accessibility to primary care: a systematic review. Eur J Public Health 2013;23(3):356-360. [doi: 10.1093/eurpub/cks053] [Medline: 22645236]
10. Sabety AH, Jena AB, Barnett ML. Changes in health care use and outcomes after turnover in primary care. JAMA Intern Med 2021;181(2):186-194 [FREE Full text] [doi: 10.1001/jamainternmed.2020.6288] [Medline: 33196767]
11. Stange KC, Miller WL, Etz RS. The role of primary care in improving population health. Milbank Q 2023;101(S1):795-840 [FREE Full text] [doi: 10.1111/1468-0009.12638] [Medline: 37096603]
12. Martin-Misener R, Valaitis R, Wong ST, Macdonald M, Meagher-Stewart D, Kaczorowski J, et al. A scoping literature review of collaboration between primary care and public health. Prim Health Care Res Dev 2012;13(4):327-346. [doi: 10.1017/S1463423611000491] [Medline: 22353204]
13. Doty MM, Tikkanen R, Shah A, Schneider EC. Primary care physicians' role in coordinating medical and health-related social needs in eleven countries. Health Aff (Millwood) 2020;39(1):115-123. [doi: 10.1377/hlthaff.2019.01088] [Medline: 31821045]
14. Little DR, Zapp JA, Mullins HC, Zuckerman AE, Teasdale S, Johnson KB. Moving toward a United States strategic plan in primary care informatics: a white paper of the primary care informatics working group, American Medical Informatics Association. Inform Prim Care 2003;11(2):89-94 [FREE Full text] [doi: 10.14236/jhi.v11i2.556] [Medline: 14567875]
15. de Lusignan S. What is primary care informatics? J Am Med Inform Assoc 2003;10(4):304-309 [FREE Full text] [doi: 10.1197/jamia.M1187] [Medline: 12668690]
16. Primary health care informatics WG. International Medical Informatics Association. 2024. URL: https://imia-medinfo.org/wp/primary-health-care-informatics/ [accessed 2024-04-29]
17. Pizziferri L, Kittler AF, Volk LA, Honour MM, Gupta S, Wang S, et al. Primary care physician time utilization before and after implementation of an electronic health record: a time-motion study. J Biomed Inform 2005;38(3):176-188 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.009] [Medline: 15896691]
18. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009;78(1):22-31. [doi: 10.1016/j.ijmedinf.2008.06.005] [Medline: 18644745]
19. Huang MZ, Gibson CJ, Terry AL. Measuring electronic health record use in primary care: a scoping review. Appl Clin Inform 2018;9(1):15-33 [FREE Full text] [doi: 10.1055/s-0037-1615807] [Medline: 29320797]
20. Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices. J Am Board Fam Med 2010;23(5):655-670 [FREE Full text] [doi: 10.3122/jabfm.2010.05.090192] [Medline: 20823361]
21. Odendaal WA, Watkins JA, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020;3(3):CD011942 [FREE Full text] [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
22. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform 2019;28(1):41-46 [FREE Full text] [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
23. Liaw ST, Kuziemsky C, Schreiber R, Jonnagaddala J, Liyanage H, Chittalia A, et al. Primary care informatics response to COVID-19 pandemic: adaptation, progress, and lessons from four countries with high ICT development. Yearb Med Inform 2021;30(1):44-55 [FREE Full text] [doi: 10.1055/s-0041-1726489] [Medline: 33882603]
24. Silva CRDV, Lopes RH, de Goes Bay O, Martiniano CS, Fuentealba-Torres M, Arcêncio RA, et al. Digital health opportunities to improve primary health care in the context of COVID-19: scoping review. JMIR Hum Factors 2022;9(2):e35380 [FREE Full text] [doi: 10.2196/35380] [Medline: 35319466]
25. Kuziemsky C, Liaw ST, Leston M, Pearce C, Jonnagaddala J, de Lusignan S. Towards equitable and resilient digital primary care systems: an international comparison and insight for moving forward. Yearb Med Inform 2022;31(1):47-59 [FREE Full text] [doi: 10.1055/s-0042-1742502] [Medline: 35654430]
26. Khalil-Khan A, Khan MA. The impact of COVID-19 on primary care: a scoping review. Cureus 2023;15(1):e33241 [FREE Full text] [doi: 10.7759/cureus.33241] [Medline: 36618499]
27. McGrail MR, Wingrove PM, Petterson SM, Bazemore AW. Mobility of US rural primary care physicians during 2000-2014. Ann Fam Med 2017;15(4):322-328 [FREE Full text] [doi: 10.1370/afm.2096] [Medline: 28694267]
28. Ramesh T, Yu H. US pediatric primary care physician workforce in rural areas, 2010 to 2020. JAMA Netw Open 2023;6(9):e2333467 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.33467] [Medline: 37703020]

XSL•FO
RenderX

29. Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open 2017;7(10):e017902 [FREE Full text] [doi: 10.1136/bmjopen-2017-017902] [Medline: 29118053]
30. Hanson K, Brikci N, Erlangga D, Alebachew A, de Allegri M, Balabanova D, et al. The Lancet Global Health commission on financing primary health care: putting people at the centre. Lancet Glob Health 2022;10(5):e715-e772 [FREE Full text] [doi: 10.1016/S2214-109X(22)00005-5] [Medline: 35390342]
31. Kato D, Ryu H, Matsumoto T, Abe K, Kaneko M, Ko M, et al. Building primary care in Japan: literature review. J Gen Fam Med 2019;20(5):170-179 [FREE Full text] [doi: 10.1002/jgf2.252] [Medline: 31516802]
32. Mash R, Howe A, Olayemi O, Makwero M, Ray S, Zerihun M, et al. Reflections on family medicine and primary healthcare in sub-Saharan Africa. BMJ Glob Health 2018;3(Suppl 3):e000662 [FREE Full text] [doi: 10.1136/bmjgh-2017-000662] [Medline: 29765778]
33. Gumas ED, Lewis C, Horstman C, Gunja MZ. Finger on the pulse: the state of primary care in the U.S. and nine other countries. The Commonwealth Fund. 2024 Mar 28. URL: https://www.commonwealthfund.org/publications/issue-briefs/2024/mar/finger-on-pulse-primary-care-us-nine-countries [accessed 2024-07-06]
34. Gunja MZ, Gumas ED, Williams RD, Doty MM, Shah A, Fields K. Stressed out and burned out: the global primary care crisis. The Commonwealth Fund. 2022 Nov 17. URL: https://www.commonwealthfund.org/publications/issue-briefs/2022/nov/stressed-out-burned-out-2022-international-survey-primary-care-physicians [accessed 2024-02-25]
35. Sinnott C, Dorban-Hall B, Dixon-Woods M. Tackling the crisis in general practice. BMJ 2023 May 02;381:966 [FREE Full text] [doi: 10.1136/bmj.p966] [Medline: 37130600]
36. Friedberg MW, van Busum K, Wexler R, Bowen M, Schneider EC. A demonstration of shared decision making in primary care highlights barriers to adoption and potential remedies. Health Aff (Millwood) 2013;32(2):268-275. [doi: 10.1377/hlthaff.2012.1084] [Medline: 23381519]
37. Wildin RS, Giummo CA, Reiter AW, Peterson TC, Leonard DGB. Primary care implementation of genomic population health screening using a large gene sequencing panel. Front Genet 2022;13:867334 [FREE Full text] [doi: 10.3389/fgene.2022.867334] [Medline: 35547253]
38. Borle K, Michaels NJ, Evans DR, Elliott AM, Price M, Austin J. Advancing the quintuple aim for health care improvement through the integration of genetic counselors into primary care. Am J Med 2023;136(12):1136-1138. [doi: 10.1016/j.amjmed.2023.08.017] [Medline: 37699497]
39. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
40. Tang PC, Miller S, Stavropoulos H, Kartoun U, Zambrano J, Ng K. Precision population analytics: population management at the point-of-care. J Am Med Inform Assoc 2021;28(3):588-595 [FREE Full text] [doi: 10.1093/jamia/ocaa247] [Medline: 33180897]
41. Canfell OJ, Littlewood R, Burton-Jones A, Sullivan C. Digital health and precision prevention: shifting from disease-centred care to consumer-centred health. Aust Health Rev 2022;46(3):279-283. [doi: 10.1071/AH21063] [Medline: 34882538]
42. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
43. Carini E, Villani L, Pezzullo AM, Gentili A, Barbara A, Ricciardi W, et al. The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: updated systematic literature review. J Med Internet Res 2021;23(9):e26189 [FREE Full text] [doi: 10.2196/26189] [Medline: 34494966]
44. Moore J, Dias GY. How to harness the power of health data to improve patient outcomes. World Economic Forum. 2024 May 01. URL: https://www.weforum.org/agenda/2024/01/how-to-harness-health-data-to-improve-patient-outcomes-wef24/ [accessed 2024-07-06]
45. Rajendran S, Pan W, Sabuncu MR, Chen Y, Zhou J, Wang F. Learning across diverse biomedical data modalities and cohorts: challenges and opportunities for innovation. Patterns 2024;5(2):100913 [FREE Full text] [doi: 10.1016/j.patter.2023.100913] [Medline: 38370129]
46. Dinov ID. Volume and value of big healthcare data. J Med Stat Inform 2016;4:3 [FREE Full text] [doi: 10.7243/2053-7662-4-3] [Medline: 26998309]
47. Tallman EF, Richardson D, Rogow TM, Kendrick DC, Dixon BE. Chapter 18—Leveraging HIE to facilitate large-scale data analytics. In: Health Information Exchange. Washington, DC: Academic Press; 2023:399-421.
48. Wagner EH. Chronic disease management: what will it take to improve care for chronic illness? Eff Clin Pract 1998;1(1):2-4. [Medline: 10345255]
49. Bodenheimer T, Ghorob A, Willard-Grace R, Grumbach K. The 10 building blocks of high-performing primary care. Ann Fam Med 2014;12(2):166-171 [FREE Full text] [doi: 10.1370/afm.1616] [Medline: 24615313]
50. Morgenstern JD, Buajitti E, O'Neill M, Piggott T, Goel V, Fridman D, et al. Predicting population health with machine learning: a scoping review. BMJ Open 2020;10(10):e037860 [FREE Full text] [doi: 10.1136/bmjopen-2020-037860] [Medline: 33109649]

XSL•FO
RenderX

51. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak 2021;21(1):54 [FREE Full text] [doi: 10.1186/s12911-021-01403-2] [Medline: 33588830]
52. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med 2023;3(1):141 [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
53. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]
54. Clark CR, Wilkins CH, Rodriguez JA, Preininger AM, Harris J, DesAutels S, et al. Health care equity in the use of advanced analytics and artificial intelligence technologies in primary care. J Gen Intern Med 2021;36(10):3188-3193 [FREE Full text] [doi: 10.1007/s11606-021-06846-x] [Medline: 34027610]
55. Dankwa-Mullan I, Scheufele EL, Matheny ME, Quintana Y, Chapman WW, Jackson G, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. J Health Care Poor Underserved 2021;32(2):300-317. [doi: 10.1353/hpu.2021.0065]
56. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
57. Liaw ST, Liyanage H, Kuziemsky C, Terry AL, Schreiber R, Jonnagaddala J, et al. Ethical use of electronic health record data and artificial intelligence: recommendations of the primary care informatics working group of the International Medical Informatics Association. Yearb Med Inform 2020;29(1):51-57 [FREE Full text] [doi: 10.1055/s-0040-1701980] [Medline: 32303098]
58. Who we are and what we do. American College of Physicians. 2024. URL: https://www.acponline.org/about-acp/who-we-are [accessed 2024-07-13]
59. Daneshvar N, Pandita D, Erickson S, Sulmasy LS, DeCamp M, ACP Medical Informatics Committeethe Ethics, ProfessionalismHuman Rights Committee. Artificial intelligence in the provision of health care: an American college of physicians policy position paper. Ann Intern Med 2024;177(7):964-967 [FREE Full text] [doi: 10.7326/M24-0146] [Medline: 38830215]
60. Levy DR, Sloss EA, Chartash D, Corley ST, Mishuris RG, Rosenbloom ST, et al. Reflections on the documentation burden reduction AMIA plenary session through the lens of 25 × 5. Appl Clin Inform 2023;14(1):11-15 [FREE Full text] [doi: 10.1055/a-1976-2052] [Medline: 36356593]
61. Donohue D. A primary care answer to a pandemic: keeping a population of patients safe at home through chronic care management and remote patient monitoring. Am J Lifestyle Med 2020;14(6):595-601 [FREE Full text] [doi: 10.1177/1559827620935382] [Medline: 33117101]
62. Farias FACD, Dagostini CM, Bicca YDA, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed e-Health 2020;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
63. Shaik T, Tao X, Higgins N, Li L, Gururajan R, Zhou X, et al. Remote patient monitoring using artificial intelligence: current state, applications, and challenges. WIREs Data Min Knowl Discovery 2023;13(2):e1485. [doi: 10.1002/widm.1485]
64. Pannunzio V, Ornelas HCM, Gurung P, van Kooten R, Snelders D, van Os H, et al. Patient and staff experience of remote patient monitoring-what to measure and how: systematic review. J Med Internet Res 2024;26:e48463 [FREE Full text] [doi: 10.2196/48463] [Medline: 38648090]
65. Leung TI, van Merode GG. Value-based health care supported by data science. In: Fundamentals of Clinical Data Science. Cham, Switzerland: Springer International Publishing; 2019:193-212.
66. Pantell MS, Adler-Milstein J, Wang MD, Prather AA, Adler NE, Gottlieb LM. A call for social informatics. J Am Med Inform Assoc 2020;27(11):1798-1801 [FREE Full text] [doi: 10.1093/jamia/ocaa175] [Medline: 33202020]
67. Byhoff E, Kangovi S, Berkowitz SA, DeCamp M, Dzeng E, Earnest M, et al. A society of general internal medicine position statement on the internists' role in social determinants of health. J Gen Intern Med 2020;35(9):2721-2727 [FREE Full text] [doi: 10.1007/s11606-020-05934-8] [Medline: 32519320]
68. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington DC: National Academies Press; 2015.
69. Social determinants of health. World Health Organization. 2024. URL: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1 [accessed 2024-07-14]
70. Andermann A, CLEAR Collaboration. Taking action on the social determinants of health in clinical practice: a framework for health professionals. CMAJ 2016;188(17-18):E474-E483 [FREE Full text] [doi: 10.1503/cmaj.160177] [Medline: 27503870]
71. Even D, Abdalla SM, Maani N, Galea S. News media as a commercial determinant of health. Lancet Glob Health 2024;12(8):e1365-e1369 [FREE Full text] [doi: 10.1016/S2214-109X(24)00191-8] [Medline: 38908393]
72. Kickbusch I, Allen L, Franz C. The commercial determinants of health. Lancet Global Health 2016;4(12):e895-e896. [doi: 10.1016/s2214-109x(16)30217-0]
73. Montoya-Williams D, Fuentes-Afflick E. Political determinants of population health. JAMA Netw Open 2019;2(7):e197063 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.7063] [Medline: 31322684]

XSL•FO
RenderX

74. Shaw N, McGuire S. Understanding the use of geographical information systems (GIS) in health informatics research: a review. J Innov Health Inform 2017;24(2):940 [FREE Full text] [doi: 10.14236/jhi.v24i2.940] [Medline: 28749317]
75. Chiesa V, Chiarenza A, Mosca D, Rechel B. Health records for migrants and refugees: a systematic review. Health Policy 2019;123(9):888-900 [FREE Full text] [doi: 10.1016/j.healthpol.2019.07.018] [Medline: 31439455]
76. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
77. Rajamani G, Espinosa PR, Rosas LG. Intersection of health informatics tools and community engagement in health-related research to reduce health inequities: scoping review. J Particip Med 2021;13(3):e30062 [FREE Full text] [doi: 10.2196/30062] [Medline: 34797214]
78. Wagner EH, Flinter M, Hsu C, Cromp D, Austin BT, Etz R, et al. Effective team-based primary care: observations from innovative practices. BMC Fam Pract 2017;18(1):13 [FREE Full text] [doi: 10.1186/s12875-017-0590-8] [Medline: 28148227]
79. Keleher H, Parker R, Abdulwadud O, Francis K. Systematic review of the effectiveness of primary care nursing. Int J Nurs Pract 2009;15(1):16-24. [doi: 10.1111/j.1440-172X.2008.01726.x] [Medline: 19187165]
80. Reiss-Brennan B, Brunisholz KD, Dredge C, Briot P, Grazier K, Wilcox A, et al. Association of integrated team-based care with health care quality, utilization, and cost. JAMA 2016;316(8):826-834. [doi: 10.1001/jama.2016.11232] [Medline: 27552616]
81. Wang P, Vienneau M, Vogeli C, Schiavoni K, Jubelt L, Mendu ML. Reframing value-based care management: beyond cost reduction and toward patient centeredness. JAMA Health Forum 2023;4(6):e231502 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.1502] [Medline: 37327007]
82. Liu L, Chien AT, Singer SJ. Enabling system functionalities of primary care practices for team dynamics in transformation to team-based care: a qualitative comparative analysis (QCA). Healthcare (Basel) 2023;11(14):2018 [FREE Full text] [doi: 10.3390/healthcare11142018] [Medline: 37510459]
83. Cliff BQ, Siegel N, Panzer J, Deis E, Patel A, Edmiston C, et al. Effects of advanced team-based care on care processes and health measures in a federally qualified health center. J Ambul Care Manage 2024;47(1):33-42. [doi: 10.1097/JAC.0000000000000484] [Medline: 37994512]
84. Baratta LR, Harford D, Sinsky CA, Kannampallil T, Lou SS. Characterizing the patterns of electronic health record-integrated secure messaging use: cross-sectional study. J Med Internet Res 2023;25:e48583 [FREE Full text] [doi: 10.2196/48583] [Medline: 37801359]
85. Kwan B, Bell JF, Longhurst CA, Goldhaber NH, Clay B. Implementation of an electronic health record-integrated instant messaging system in an academic health system. J Am Med Inform Assoc 2024;31(4):997-1000. [doi: 10.1093/jamia/ocad253] [Medline: 38287641]
86. Sakumoto M, Brosnan M, Gutierrez J. Key considerations for virtual team-based primary care. Society of General Internal Medicine. 2022. URL: https://www.sgim.org/article/key-considerations-for-virtual-team-based-primary-care/ [accessed 2024-07-07]
87. Willis VC, Craig KJT, Jabbarpour Y, Scheufele EL, Arriaga YE, Ajinkya M, et al. Digital health interventions to enhance prevention in primary care: scoping review. JMIR Med Inform 2022;10(1):e33518 [FREE Full text] [doi: 10.2196/33518] [Medline: 35060909]
88. Jimenez G, Matchar D, Koh CHG, van der Kleij R, Chavannes NH, Car J. The role of health technologies in multicomponent primary care interventions: systematic review. J Med Internet Res 2021;23(1):e20195 [FREE Full text] [doi: 10.2196/20195] [Medline: 33427676]
89. Jimenez G, Matchar D, Koh GCH, Tyagi S, van der Kleij RMJJ, Chavannes NH, et al. Revisiting the four core functions (4Cs) of primary care: operational definitions and complexities. Prim Health Care Res Dev 2021;22:e68 [FREE Full text] [doi: 10.1017/S1463423621000669] [Medline: 34753531]
90. Prates ML, Machado JC, Silva LSD, Avelar PS, Prates LL, Mendonça ET, et al. Performance of primary health care according to PCATool instrument: a systematic review. Cien Saude Colet 2017;22(6):1881-1893 [FREE Full text] [doi: 10.1590/1413-81232017226.14282016] [Medline: 28614508]
91. Weiner JP, Bandeian S, Hatef E, Lans D, Liu A, Lemke KW. In-person and telehealth ambulatory contacts and costs in a large US insured cohort before and during the COVID-19 pandemic. JAMA Netw Open 2021;4(3):e212618 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2618] [Medline: 33755167]
92. Thakkar J, Kurup R, Laba TL, Santo K, Thiagalingam A, Rodgers A, et al. Mobile telephone text messaging for medication adherence in chronic disease: a meta-analysis. JAMA Intern Med 2016;176(3):340-349. [doi: 10.1001/jamainternmed.2015.7667] [Medline: 26831740]
93. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020;22(12):e21687 [FREE Full text] [doi: 10.2196/21687] [Medline: 33295292]

XSL•FO
RenderX

94. Broderick A, Haque F. Mobile health and patient engagement in the safety net: a survey of community health centers and clinics. The Commonwealth Fund. 2015 May 13. URL: https://www.commonwealthfund.org/publications/issue-briefs/2015/may/mobile-health-and-patient-engagement-safety-net-survey-community [accessed 2024-04-25]
95. Spiegel BM, Rizzo A, Persky S, Liran O, Wiederhold B, Woods S, et al. What is medical extended reality? A taxonomy defining the current breadth and depth of an evolving field. J Med Ext Real 2024;1(1):4-12 [FREE Full text] [doi: 10.1089/jmxr.2023.0012] [Medline: 38505474]
96. Urlings J, Sezer S, Ter Laan M, Bartels R, Maal T, Boogaarts J, et al. The role and effectiveness of augmented reality in patient education: a systematic review of the literature. Patient Educ Couns 2022;105(7):1917-1927 [FREE Full text] [doi: 10.1016/j.pec.2022.03.005] [Medline: 35341611]
97. Oniani D, Sreekumar S, DeAlmeida R, DeAlmeida D, Hui V, Lee YJ, et al. Toward improving health literacy in patient education materials with neural machine translation models. AMIA Jt Summits Transl Sci Proc 2023;2023:418-426 [FREE Full text] [Medline: 37350905]
98. Zhu W, Liu H, Dong Q. Multilingual machine translation with large language models: empirical results and analysis. 2024 Presented at: 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics; June 16-21, 2024; Mexico City, Mexico p. 2765-2781 URL: https://aclanthology.org/2024.findings-naacl.176.pdf [doi: 10.18653/v1/2024.findings-naacl.176]
99. Melnick ER, West CP, Nath B, Cipriano PF, Peterson C, Satele DV, et al. The association between perceived electronic health record usability and professional burnout among US nurses. J Am Med Inform Assoc 2021;28(8):1632-1641 [FREE Full text] [doi: 10.1093/jamia/ocab059] [Medline: 33871018]
100. Everson J, Hendrix N, Phillips RL, Adler-Milstein J, Bazemore A, Patel V. Primary care physicians' satisfaction with interoperable health information technology. JAMA Netw Open 2024;7(3):e243793 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.3793] [Medline: 38530309]
101. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021;18(16):8435 [FREE Full text] [doi: 10.3390/ijerph18168435] [Medline: 34444182]
102. Porreca D, Yaraghi N. The impact of population health analytics on health care quality and efficacy among CPC+ participants. Milbank Memorial Fund. 2022. URL: https://www.milbank.org/wp-content/uploads/2022/10/WNY_Report2.pdf [accessed 2024-09-18]
103. López-Martínez F, Núñez-Valdez ER, García-Díaz V, Bursac Z. A case study for a big data and machine learning platform to improve medical decision support in population health management. Algorithms 2020;13(4):102. [doi: 10.3390/a13040102]
104. Mendu ML, Ahmed S, Maron JK, Rao SK, Chaguturu SK, May MF, et al. Development of an electronic health record-based chronic kidney disease registry to promote population health management. BMC Nephrol 2019;20(1):72 [FREE Full text] [doi: 10.1186/s12882-019-1260-y] [Medline: 30823871]
105. Elliott MN, Adams JL, Klein DJ, Haviland AM, Beckett MK, Hays RD, et al. Patient-reported care coordination is associated with better performance on clinical care measures. J Gen Intern Med 2021;36(12):3665-3671 [FREE Full text] [doi: 10.1007/s11606-021-07122-8] [Medline: 34545472]
106. Vimalananda VG, Orlander JD, Afable MK, Fincke BG, Solch AK, Rinne ST, et al. Electronic consultations (E-consults) and their outcomes: a systematic review. J Am Med Inform Assoc 2020;27(3):471-479 [FREE Full text] [doi: 10.1093/jamia/ocz185] [Medline: 31621847]
107. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. Lancet Digit Health 2021;3(6):e383-e396 [FREE Full text] [doi: 10.1016/S2589-7500(21)00005-4] [Medline: 33967002]
108. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst 2024;5(3):CAT.23.0404. [doi: 10.1056/cat.23.0404]
109. Bhasker S, Bruce D, Lamb J, Stein G. Tackling healthcare's biggest burdens with generative AI. McKinsey & Company. 2023 Jul 10. URL: https://www.mckinsey.com/industries/healthcare/our-insights/tackling-healthcares-biggest-burdens-with-generative-ai [accessed 2024-04-25]
110. Baker R, Freeman GK, Haggerty JL, Bankart MJ, Nockels KH. Primary medical care continuity and patient mortality: a systematic review. Br J Gen Pract 2020;70(698):e600-e611 [FREE Full text] [doi: 10.3399/bjgp20X712289] [Medline: 32784220]
111. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
112. Spencer RA, Punia HS. A scoping review of communication tools applicable to patients and their primary care providers after discharge from hospital. Patient Educ Couns 2021;104(7):1681-1703. [doi: 10.1016/j.pec.2020.12.010] [Medline: 33446366]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
EHR: electronic health record
PCI: primary care informatics
SDOH: social determinants of health

Edited by G Eysenbach; submitted 30.04.24; peer-reviewed by D Little, B Hoyt; comments to author 11.05.24; revised version received 16.07.24; accepted 31.08.24; published 15.10.24.

Please cite as:
You JGT, Leung TI, Pandita D, Sakumoto M
Primary Care Informatics: Vitalizing the Bedrock of Health Care
J Med Internet Res 2024;26:e60081
URL: https://www.jmir.org/2024/1/e60081
doi:10.2196/60081
PMID:

©Jacqueline Guan-Ting You, Tiffany I Leung, Deepti Pandita, Matthew Sakumoto. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mapping the Evolution of Digital Health Research: Bibliometric Overview of Research Hotspots, Trends, and Collaboration of Publications in JMIR (1999-2024)

Jing Hu[1,2,3*], BScPT, MS, DPT; Chong Li[1,2,3*], PhD; Yanlei Ge[1,2,3], MSc; Jingyi Yang[1,2,3], MSc; Siyi Zhu[1,2,3], MD, PhD; Chengqi He[1,2,3], MD, PhD

[1]Rehabilitation Medicine Center, West China Hospital, Sichuan University, Chengdu, China

[2]Key Laboratory of Rehabilitation Medicine in Sichuan Province, West China Hospital, Sichuan University, Chengdu, China

[3]School of Rehabilitation Sciences, West China School of Medicine, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Chengqi He, MD, PhD
Rehabilitation Medicine Center
West China Hospital
Sichuan University
#37 Guoxue Alley
Wuhou District
Chengdu, 610041
China
Phone: 86 28 8542 2847
Fax: 86 28 8542 2847
Email: hxkfhcq2015@126.com

Abstract

Background: While bibliometric studies of individual journals have been conducted, to the best of our knowledge, bibliometric mapping has not yet been utilized to analyze the literature published by the *Journal of Medical Internet Research* (*JMIR*).

Objective: In celebration of the journal's 25th anniversary, this study aimed to review the entire collection of *JMIR* publications from 1999 to 2024 and provide a comprehensive overview of the main publication characteristics.

Methods: This study included papers published in *JMIR* during the 25-year period from 1999 to 2024. The data were analyzed using CiteSpace, VOSviewer, and the "Bibliometrix" package in R. Through descriptive bibliometrics, we examined the dynamics and trend patterns of *JMIR* literature production and identified the most prolific authors, papers, institutions, and countries. Bibliometric maps were used to visualize the content of published articles and to identify the most prominent research terms and topics, along with their evolution. A bibliometric network map was constructed to determine the hot research topics over the past 25 years.

Results: This study revealed positive trends in literature production, with both the total number of publications and the average number of citations increasing over the years. And the global COVID-19 pandemic induced an explosive rise in the number of publications in *JMIR*. The most productive institutions were predominantly from the United States, which ranked highest in successful publications within the journal. The editor-in-chief of *JMIR* was identified as a pioneer in this field. The thematic analysis indicated that the most prolific topics aligned with the primary aims and scope of the journal. Currently and in the foreseeable future, the main themes of *JMIR* include "artificial intelligence," "patient empowerment," and "victimization."

Conclusions: This bibliometric study highlighted significant contributions to digital health by identifying key research trends, themes, influential authors, and collaborations. The findings underscore the necessity to enhance publications from developing countries, improve gender diversity among authors, and expand the range of research topics explored in the journal.

(J Med Internet Res 2024;26:e58987) doi:10.2196/58987

KEYWORDS

JMIR; bibliometric analysis; ehealth; digital health; medical informatics; health informatics; open science; publishing

XSL•FO
RenderX

Introduction

The *Journal of Medical Internet Research* (*JMIR*) is a renowned peer-reviewed journal that specializes in health informatics and digital health. It addresses a broad spectrum of topics, including emerging technologies, medical devices, apps, engineering, telehealth, and informatics applications for patient education, prevention, population health, and clinical care. Professor Gunther Eysenbach, a renowned expert in digital health and eHealth, established the journal *JMIR* in 1999 and has since served as its Editor-in-Chief. He has also defined digital health as health services and information delivered or enhanced through the internet and related technologies [1]. Because of its interdisciplinary nature, digital health spans various fields, including medical sciences, computer sciences, behavioral sciences, social sciences, communication sciences, psychology, library sciences, informatics, human-computer interaction studies, and related areas. To drive and accelerate the advancement of digital health, *JMIR* provides valuable insights across all these topics. The inaugural edition was released in August 1999, with a second publication later that same year. Beginning in 2022, *JMIR* has published 12 issues annually. Starting in 2023, the publications have been organized into annual volumes without individual issue numbers. Under JMIR Publications, Inc., 26 volumes have been published, with 1 volume released each year. Additionally, *JMIR* has curated 129 thematic collections to organize publications on specific digital health topics. A variety of metrics consistently place the journal among the top in its field. According to the SCImago Journal Rank powered by Scopus, *JMIR* is ranked in the top 25% (Q1) of the Medicine: Health Informatics discipline, with a ranking of 8th out of 123 (93rd percentile). Based on the 2023 Journal Citation Indicator from Clarivate, the journal ranks 5th out of 31 journals in the Medical Informatics category (85.5th percentile). In the Health Care Sciences and Services category, it is ranked 3rd out of 105 journals, placing it in the 97.6th percentile. Additionally, the journal has an impact factor of 7.4. Based on the scholarly influence of publications over the past 5 years, Google Scholar has ranked *JMIR* as the top journal in the field of Medical Informatics.

When a journal celebrates an anniversary or special occasion, it is customary to organize a specific academic undertaking, such as a review [2], an editorial [3], or a bibliometric study [4]. Computer-assisted statistics and software have significantly advanced the practice of bibliometrics. These powerful tools enable the summarization of journal contributions and cooperation patterns, as well as the analysis of large volumes of scholarly publications to understand the knowledge structure and developmental trends across various disciplines [5]. A retrospective bibliometric analysis of the European Journal of Innovation Management from 1998 to 2021 was conducted by Bamel et al [6]; Martorell Cunill et al [7] performed a bibliometric analysis for the International Journal of Hospitality Management's 35th anniversary [7]; and Nath et al [8] examined publications in Community Dentistry and Oral Epidemiology from 1973 to 2022.

Given *JMIR*'s prominent position in the digital health sector, conducting a bibliometric analysis of this journal can provide valuable insights into the knowledge framework of the field. The readers, editors, and reviewers of *JMIR*, along with academics researching digital health, will gain a clear understanding of the current boundaries of the subject, as well as the key contributors, emerging topics, and evolving focal points.

To celebrate *JMIR*'s silver jubilee (25 years), this paper aims to retrospectively review the journal's productivity, impact, emerging knowledge, and contributions to the digital health discipline since its establishment. This study analyzes the impact of the journal by examining its publication and citation patterns. It also identifies the most influential authors, institutions, and nations associated with the journal, along with an analysis of the authors' keywords. To achieve these objectives, articles published in *JMIR* from its inception to the present were retrieved and analyzed. Three bibliometric analytical tools—CiteSpace, VOSviewer, and the R package Bibliometrix—were utilized to comprehensively assess the academic structure of *JMIR* publications. This study not only summarizes the established academic structure but also predicts potential future research focuses for *JMIR*.

Methods

Search Strategy

The publications spanning from January 1999 to February 2024 were obtained from the Web of Science Core Collection (WoSCC) database on February 22, 2024. The research used the strategies outlined in the source title and included articles and reviews. We excluded letters, revisions, proceedings papers, editorial materials, and retracted publications. The investigation was conducted independently by 2 researchers (JH and CL) to ensure the reliability and accuracy of the data collection process.

Data Extraction and Analysis

The publications were collected and retrieved in the "Plain Text" format of "Full Records and References" from WoSCC. General information about *JMIR* publications was obtained from the WoSCC database. The bibliometric analysis of the included studies was conducted using CiteSpace 6.3 R1 Advanced, VOSviewer 1.6.19, and the R package Bibliometrix 4.1.2. This analysis focused on publications from January 1999 to December 2023. Publications from 2024 were discussed separately due to incomplete citation data, which could introduce bias into the analysis.

The software CiteSpace 6.3 R1 Advanced [9] was used to conduct a cocitation analysis, which included authors and references. Additionally, the software was used to create dual maps of journals and to discover citation bursts for keywords and authors. Clustering of keywords with labels generated from the keywords was performed to identify research themes and visualize the intellectual structure. Furthermore, a timeline view of keywords was created to visualize keyword evolution and identify trend shifts. The settings for CiteSpace were configured as follows: the period was set from January 1999 to December 2023, with each time slice representing a 1-year period. The strength of the linkages was measured using the cosine method, and the scope of the links was limited to each time slice. The

XSL•FO
RenderX

selection criterion used was the g-index, with a value of k=25. Additionally, the network was pruned using the algorithm for pruning sliced networks and minimum spanning trees.

VOSviewer (version 1.6.19) [10] was used to analyze collaboration networks of countries, institutions, and authors through coauthorship analysis, with a minimum threshold of 5 documents per subject. VOSviewer also facilitated the analysis of keyword co-occurrence, using a minimum occurrence threshold of 5 keywords. The full count method was applied for all analyses. Network visualization, overlay visualization, and density visualization were used to depict the connections, temporal changes, and concentration of items within the network. The LinLog/modularity method was used to enhance the clarity and interpretability of the visualizations, facilitating a comprehensive understanding of collaboration patterns and keyword relationships. Additionally, the total link strength value was calculated for each analysis to measure connectivity and importance within the network [11].

The R package Bibliometrix (version 4.1.2) [12] was utilized to compute Lotka's law [13] and analyze multiple-country publications (MCPs). The visualizations included collaborative globe maps of countries and trending subjects of keywords. Three-field plots were used to visualize the relationships and connections between countries, authors, and keywords, with the left field representing countries, the middle field representing authors, and the right field representing authors' keywords. A thematic analysis of terminology was conducted based on "keyword plus" using a Walktrap clustering algorithm to identify the major research themes, their relative importance, interconnections, and evolving trends.

Interpretation of Charts

Node

A node is defined as a single entity, such as a country, institution, author, or keyword. A connection between nodes signifies the presence of a co-occurrence relationship.

Betweenness Centrality

Betweenness centrality is a metric used to quantify the significance of nodes within a network. A higher betweenness centrality indicates greater connectivity relevance for the nodes [14].

Cluster View

A cluster view was created using the likelihood ratio statistic and keyword lists of cited articles within each cluster. The cluster name was automatically determined by selecting the keyword with the strongest correlation for each cluster. The effectiveness of the graph drawing was assessed using the modularity Q value and the weighted mean silhouette value (S). Clustering results were deemed satisfactory if the modularity Q was greater than 0.3 [15] and the homogeneity of the cluster structure was verified when the mean S was greater than 0.5 [14].

Dual-Map Overlap

The dual-map overlay visualizes the distribution and citations of works across different disciplines. The left panel displays the journals that are being cited, while the right panel presents the journals that cite them. The citation paths are indicated by colored stripes in the middle [16].

Burst Detection

Burst detection involves identifying changes in the frequency of keywords, which may occur when there is a significant shift in the number of citations.

Results

Global Trend in Publication Outputs and Citations

The study's workflow, including the search strategy and method, is depicted in Multimedia Appendix 1. The search conducted using the WoSCC database resulted in a total of 7780 documents. After excluding 137 correlations, 2 proceeding papers, 71 editorial materials, 2 retracted publications, and 120 letters, 7448 publications were included in the data set. This data set involved the collaboration of 32,232 authors. Out of these publications, 6306 (84.67%) were categorized as articles, while 1142 (15.33%) were classified as reviews. General information about the papers is presented in Multimedia Appendix 2. The average annual publication count is 297.92. According to the "Web of Science Citation Report," there have been a total of 226,105 citations for these articles, resulting in an average of 30.36 citations per article and 9044.2 citations annually.

Figure 1 illustrates the patterns of yearly publications and citation values. The number of annual publications exhibited a consistent and gradual increase during the first 25 years of the century, followed by a sudden spike in 2022. The number of citations showed continuous and gradual growth until it reached a point of stability in the past 2 years. The cumulative H-index of *JMIR* is 177.

XSL•FO
RenderX

Figure 1. The number of annual publications (bar chart) and the number of annual citations (line chart) from 1999 to 2023 (source: the Web of Science Core Collection). The explosion symbol in the graphic signifies the significant increase in research activities during the global COVID-19 pandemic.

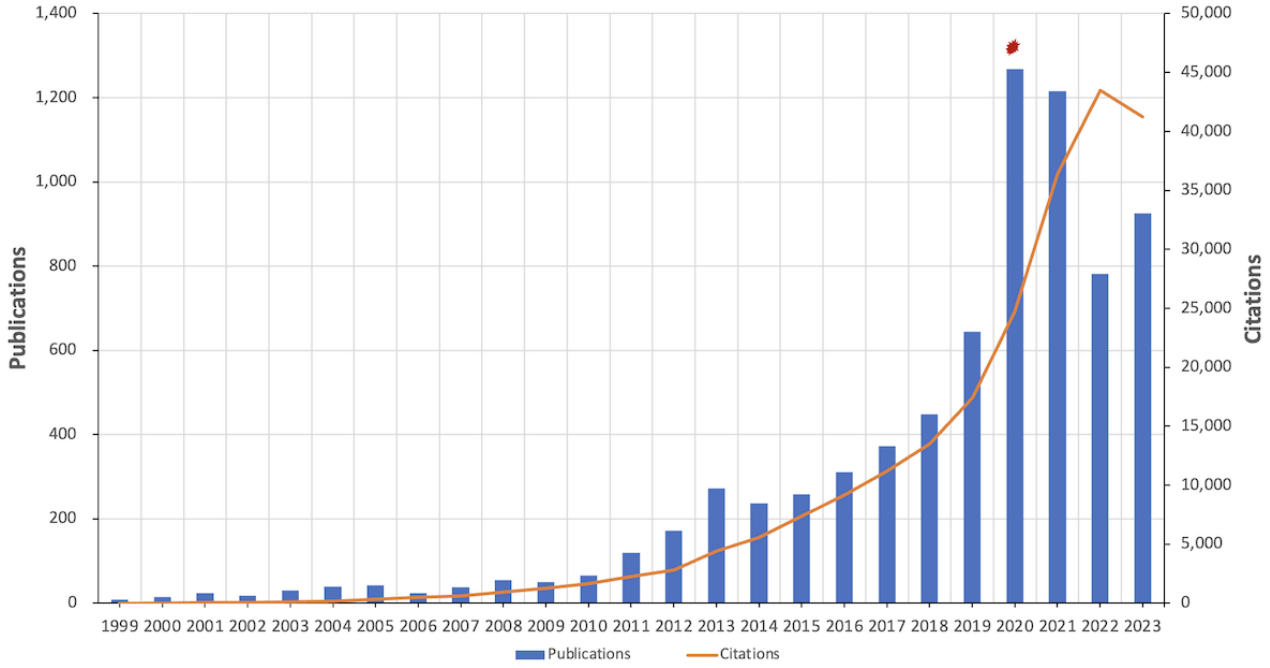

Analysis of Authoritative Country/Region

A total of 120 countries contributed to publishing in *JMIR*. Table 1/Figure 2A presents the rankings of the 10 most prolific countries based on their total publication (TP). The United States ranked first with 2769 publications, accounting for 37.18% of the total 7448 publications. The United Kingdom followed with 949 publications, representing 12.74%, while China came in third with 798 publications, accounting for 10.71%. The United States stands out for having the highest number of publications and total citations (TC=83,973), along with an impressive H-index of 128. These figures reflect the country's significant contributions to research in the field of digital health. Figure 2B displays the top 10 most productive nations, and the data suggest that the United States experienced a more notable rise in productivity compared with other countries. Outside the top 10 nations in terms of publications, the Netherlands had the highest number of citations per publication (average citation [AC]=38.51), followed by Australia (AC=35.65) and the United Kingdom (AC=35.25). Meanwhile, the United States had the most citations overall (AC=35.65; Table 1).

The ratio of MCPs measures the proportion of multiple-country publications among total publications, which highlights international cooperation in research. Among the top 10 most prolific nations, Switzerland (82/163, 50.3%), Spain (61/154, 39.6%), and Germany (148/384, 38.5%) have the highest ratios of MCPs for global collaboration, as shown in Multimedia Appendix 3. Although the United States (326/2156, 15.1%), the United Kingdom (197/619, 31.8%), and China (279/808, 34.5%) ranked as the top 3 nations in terms of the volume of publications involving multiple countries, their MCP ratios were not the highest due to the sheer volume of publications they have produced.

Based on the collaboration world map of countries (Figure 2C), the United States had the highest degree of collaboration with China, with a frequency of 230. This was followed by the United Kingdom, with a frequency of 166, and Canada, with a frequency of 146, as shown in Multimedia Appendix 4. Multiple countries have made significant contributions or engaged in collaboration with scholars in *JMIR* over varying durations. The United States, Canada, the Netherlands, and Australia have exhibited high levels of activity, with an average publication year of 2018. With advancements in computer and internet technologies, research in digital health is attracting increasing attention from scholars. Over the past 3 years, researchers from China, South Korea, and Brazil have demonstrated significant academic engagement and collaboration with other nations (Multimedia Appendix 5).

XSL•FO
RenderX

Figure 2. Analysis of countries/regions engaged in publications in the Journal of Medical Internet Research (JMIR). (A) Top 10 countries/regions with the largest number of publications (source: WoSCC); (B) top 10 countries/regions with the largest number of publications over time (source: WoSCC); (C) countries' collaboration world map (source: the R package Bibliometrix). WoSCC: Web of Science Core Collection.

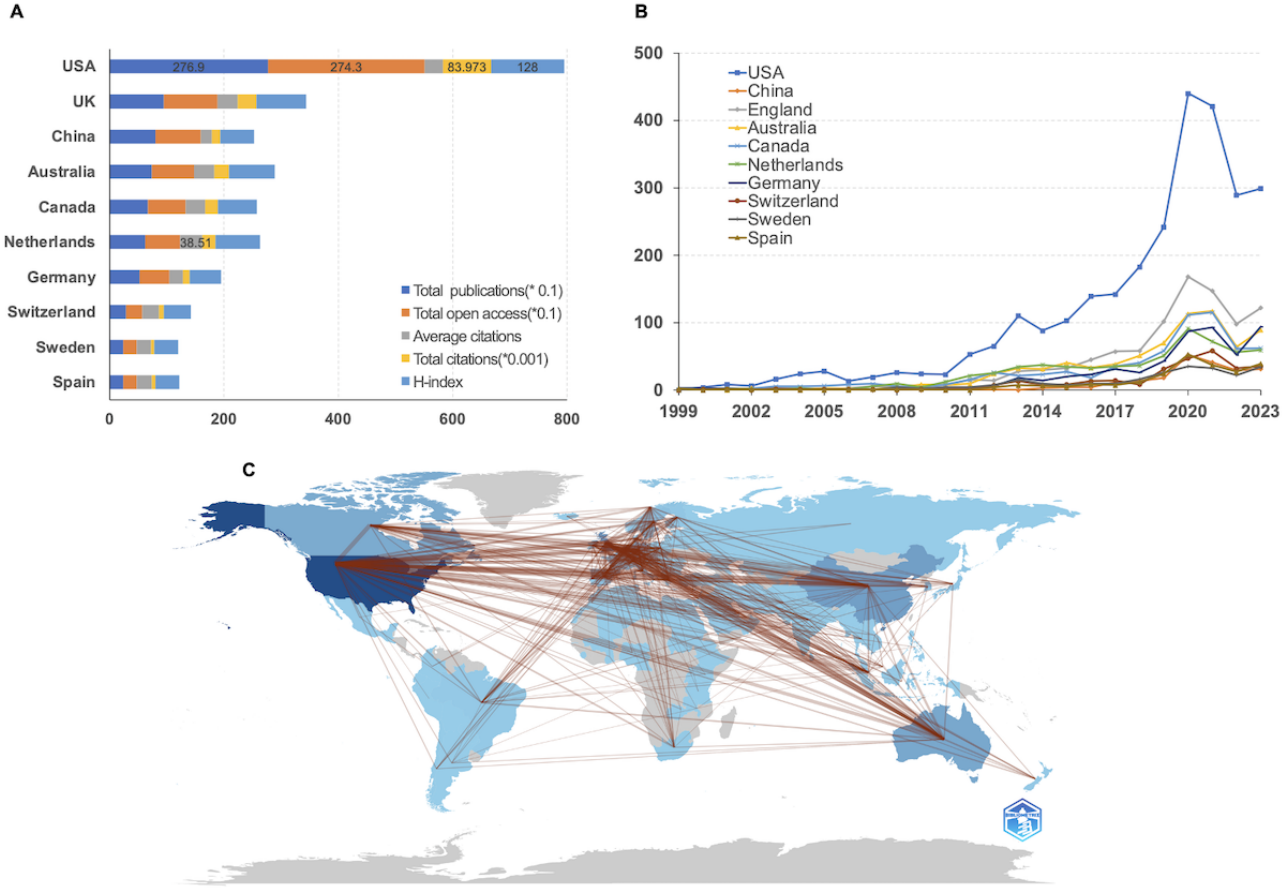

Table 1. Top 10 countries in terms of publications (source from WoSCC[a]).

Country	Total publication, n	Total open access, n	Average citations	Total citations, n	H-index
United States	2769	2743	31.79	83973	128
United Kingdom	949	940	35.25	32614	87
China	798	794	19.49	14955	59
Australia	740	733	35.65	25691	80
Canada	667	660	34.35	22494	68
The Netherlands	619	615	38.51	22870	78
Germany	524	522	23.27	11857	55
Switzerland	286	285	29.51	8280	47
Sweden	237	235	25	5793	42
Spain	236	235	26.73	6225	42

[a]WoSCC: WoSCC: Web of Science Core Collection.

Analysis of Publishing Institution

A total of 6689 institutions contributed to publications in *JMIR*. Figure 3A displays the top 10 institutions based on the number of papers. The 3 leading institutions are the University of California System, with 351 publications (4.71%); Harvard University, with 328 publications (4.40%); and the University of Toronto, with 283 publications (3.80%). The trend in institutions' contributions to *JMIR*, in terms of the number of research articles published over time, has remained similar across institutions, as shown in Figure 3B. The level of collaboration among institutions is substantial, as depicted in Figure 3C. The University of Toronto had the highest level of collaboration with several institutions, including the University Health Network Toronto (link strength=78), Women's College Hospital (link strength=31), and the Centre for Addiction and Mental Health (link strength=28). Meanwhile, Harvard Medical

XSL•FO
RenderX

School and Harvard University engage in extensive relationships with other scholarly institutions, as indicated in Multimedia Appendix 6. Multimedia Appendix 7 displays a timeline view of institutional collaboration. Harvard Medical School, one of the most productive institutions, has recently been active in *JMIR*, with an average publication year of 2020. Several scholarly institutions in Asia have also been actively involved in and made significant contributions to *JMIR* in recent years, including Peking University, Sun Yat-sen University, and National Taiwan University.

Figure 3. Analysis of institutions engaged in publications in the Journal of Medical Internet Research (JMIR). (A) Top 10 institutions with the largest number of publications (source: WoSCC); (B) top 10 institutions with the largest number of publications over time (source: WoSCC); (C) institutions' collaboration network (source: VOSviewer). WoSCC: Web of Science Core Collection.

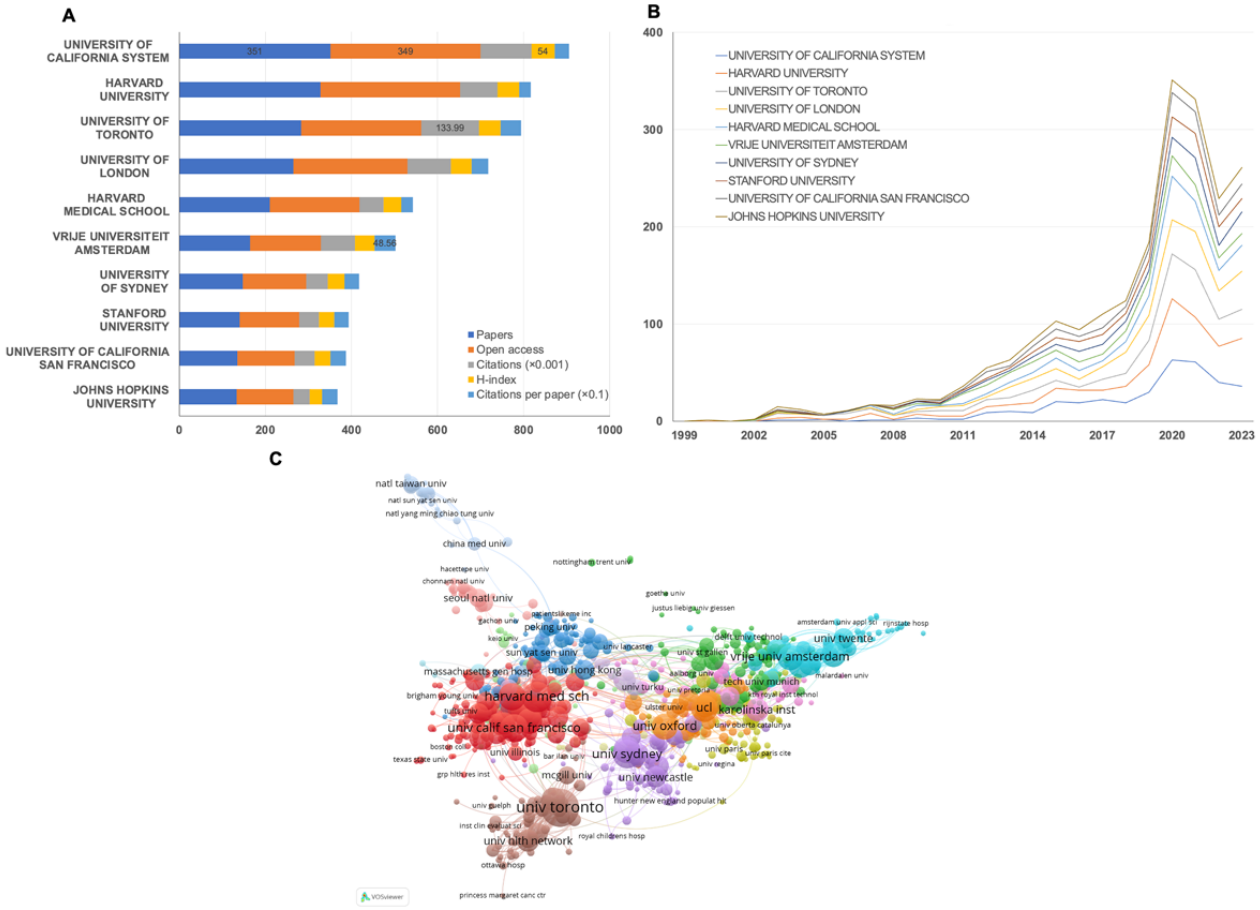

Analysis of Author Distribution

The authors demonstrating the highest levels of productivity were Heleen Riper (TP=41) from Vrije Universiteit Amsterdam, Hein de Vries (TP=39) from Maastricht University, and Jinseok Lee (TP=39) from Kyung Hee University. Data regarding the top 10 authors with the highest publication frequency were retrieved from the WoSCC database and are presented in Multimedia Appendix 8. Three out of the 10 most prolific authors hailed from the Netherlands. Additionally, an analysis was conducted on the worldwide academic influence of *JMIR* authors. Multimedia Appendix 9 displays information on the top 25 influential writers based on the strength of citation bursts. Gunther Eysenbach had the most significant burst in citations since 1999, with a strength of 11 from 1999 to 2011. Hein De Vries had the highest burst strength of 12.29 from 2012 to 2017. Tobias Kowatsch (strength=8.11, 2021-2024) and Felix Balzer (strength=3.76, 2020-2024) were the most recent authors with strong citation bursts. Figure 4A and Multimedia Appendix 10 display the author's coauthorship network mapping and its strength. Heleen Riper collaborated extensively with other authors, demonstrating a total link strength of 100. The average publication year of their work was 2016. Hein de Vries and Lorainne Tudor Car, both affiliated with Nanyang Technological University, also have significant scholarly connections with other authors. Hein de Vries has a total link strength of 81, while Lorainne Tudor Car has a total link strength of 76. Additionally, Lorainne Tudor Car has remained academically active, with an average publication year of 2020. Cocitation analysis identified Gunther Eysenbach, Helen Christensen, and Gerhard Andersson as the 3 most frequently cited authors in connection with other *JMIR* authors (Figure 4B/Multimedia Appendix 11). The most significant cocitation link emerged between Gunther Eysenbach and Susannah Fox, with a link strength of 355. According to the authors' local citation map (Figure 4C), Gunther Eysenbach (local citations=536), Helen Christensen (local citations=493), and Julia EWC van Gemert (local citations=429) were the most frequently cited by other *JMIR* authors. A 3-field map shown in Figure 4D helps visualize the links between the authors, nations, and keywords. Most authors from the Netherlands who have made exceptional contributions to publishing in *JMIR* focus on themes such as eHealth, COVID-19 impacts, and mobile phone apps. Lotka's law was assessed by quantifying

XSL•FO
RenderX

the authors' publication frequency in *JMIR* as a measure of their productivity (Multimedia Appendix 12). The data revealed that 77.54% (24,993/32,232) of the authors contributed to *JMIR* publications by submitting at least one paper, while 13.45% (4336/32,232) published a second article.

Figure 4. Analysis of authors engaged in publications in the Journal of Medical Internet Research (JMIR). (A) Coauthorship network of authors (source: VOSviewer); (B) cocitation network of authors (source: VOSviewer); (C) authors' local citation map (source: the R package Bibliometrix); (D) the 3-field plot—middle field: authors; left field: country; right field: keywords (source: the R package Bibliometrix).

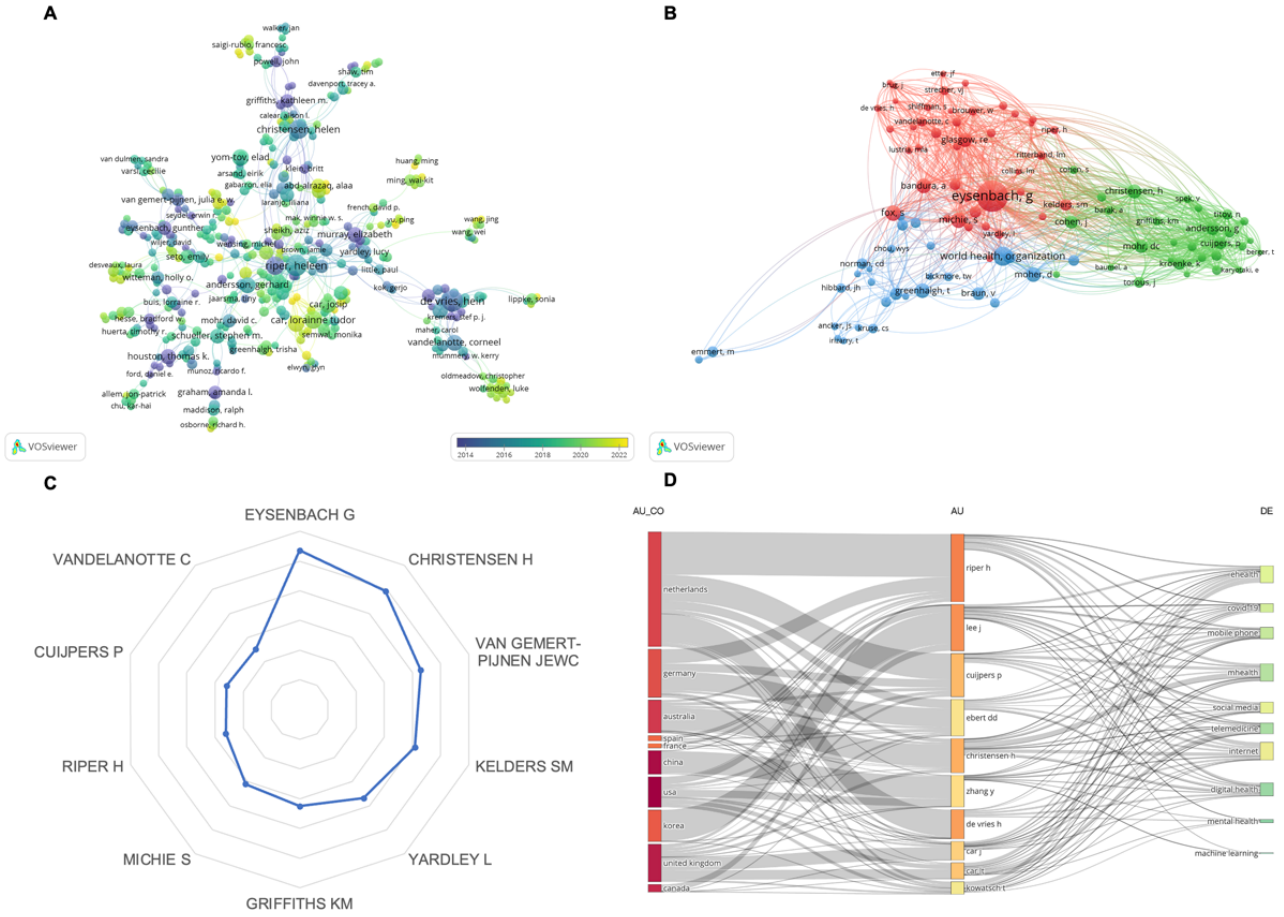

Analysis of Journal Distribution

Our investigation revealed that a total of 11,896 journals were referenced by publications in *JMIR*. Multimedia Appendix 13 presents a concise overview of the 25 most frequently cited publications in *JMIR*. In addition to *JMIR* itself (n=5982), the most commonly cited academic journals include the International Journal of Environmental Research and Public Health (n=2538), PLoS One (n=1915), and JMIR mHealth and uHealth (n=1641). In terms of the distribution of articles citing *JMIR* publications, the majority of citations originated from *JMIR* itself (n=24,714), followed by PLoS One (n=4562) and the Journal of the American Medical Informatics Association (n=3877). Multimedia Appendix 14 provides additional frequently cited sources. According to data from the Web of Science, most articles citing *JMIR* publications can be classified into the categories of Health Care Sciences Services (21,539/120,399, 17.89%, articles citing *JMIR* publications), Medical Informatics (17,144/120,399, 14.24%), and Public Environmental Occupational Health (17,072/120,399, 14.18%). The dual-map overlay in Figure 5 illustrates the distribution of topics among journals. This overlay consists of 2 sides: the citation channels indicate that research published in health, nursing, medicine, psychology, education, and social journals is frequently cited by studies published in *JMIR*, which falls under the category of medicine, medical, and clinical journals.

XSL•FO
RenderX

Figure 5. The dual - map overlay of journals in the field of frailty and depression research (source: CiteSpace).

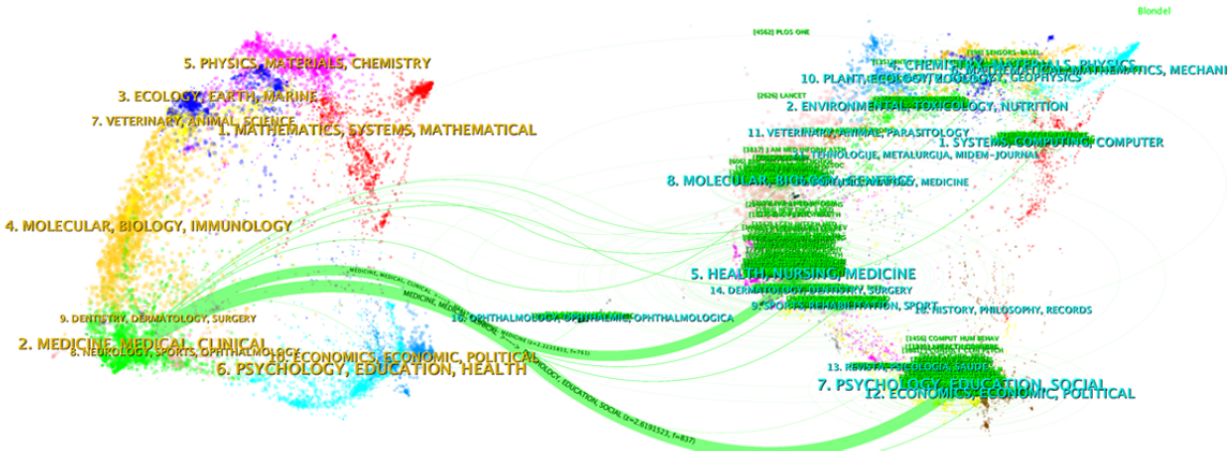

Analysis of Highly Impact Publications

Table 2 highlights the 10 most widely referenced publications from *JMIR*, which have had a significant academic influence on a global scale. Among these, 5 are reviews and 5 are original articles. Four of the publications report on randomized controlled trials [17-20]. Three publications explored the topic of mental health [19-21]. Additionally, the significant role of compliance in digital health care is discussed in 5 out of the 10 studies [17-20,22]. An article published in 2005, which examined attrition in eHealth trials [22], has the highest number of citations (TC=1615), averaging approximately 80.75 citations per year. The most recent publication among the top 10 notable articles is by Andras Lanczky et al [23], which describes a web-based tool for conducting univariate and multivariate Cox proportional hazards survival analyses using data from transcriptomic, proteomic, metabolomic, or genomic investigations. This publication had the highest citation density of 182.5 among all the publications. Another notable citation density was found in a review examining the impact of the COVID-19 epidemic on the mental health of college students, which had a citation density of 180 [21]. Multimedia Appendix 15 (see also [17-19,22,24-29]) provides a summary of the 10 most frequently cited publications by local authors. The intellectual foundation of *JMIR* is further enriched by publications from other sources, with data regarding the top 10 references available in Multimedia Appendix 16 (see also [17,18,22,30-36]). Among the 10 highly cited articles, 5 were published in *JMIR*, and the authors predominantly referenced statements related to systematic reviews and meta-analyses [30].

XSL•FO
RenderX

Table 2. The top 10 high-impact articles published between 1999 and 2023 (sourced from the Web of Science Core Collection).

No	Reference	Type	First author, year	Last author	Main find	Total citations, n	Country	Duration (years)	Citation density
1	[22]	Article	Gunther Eysenbach, 2005	Single author	This article advocates for the necessity of a "science of attrition," which entails the development of models to understand the termination of eHealth apps and the occurrence of participants dropping out of eHealth trials.	1615	Canada	20	80.75
2	[17]	Review	Thomas L. Webb,2010	Susan Michie	This review assessed theory and behavior change strategies to determine the best internet-based treatments to promote health behavior change. A new scoring scheme was also created for measuring internet-based intervention modes of delivery and linking modes to effect sizes.	1582	England	15	105.47
3	[24]	Review	Anne Moorhead, 2013	Ciska Hoving	Social media is a strong tool for collaboration and social engagement. Social media for health communication has many benefits, but the information provided must be checked for quality and dependability and users' confidentiality and privacy must be safeguarded.	1264	North Ireland	12	105.33
4	[25]	Article	Cameron D. Norman, 2006	Harvey A. Skinner	A model of eHealth literacy was presented, covering the abilities customers need to profit directly from eHealth. Health practitioners can use a resource list and a literacy-type profile with examples of patient-client issues to increase literacy across each domain.	1232	Canada	19	64.84
5	[21]	Article	Changwon Son, 2020	Farzan Sasangohao	The COVID-19 pandemic introduced increased levels of stress, anxiety, and depression in higher education.	900	The United States	5	180
6	[37]	Article	Trisha Greenhalgh, 2017	Sara Shaw	A Nonadopting, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework can help predict and evaluate the success of a technology-supported health or social care program.	871	The United Kingdom	8	108.88

XSL•FO
RenderX

No	Reference	Type	First author, year	Last author	Main find	Total citations, n	Country	Duration (years)	Citation density
7	[18]	Review	Saskia M. Kelders, 2012	Julia E. W. C. an Gemert-Pijnen	Although health care settings differ in intervention characteristics, they do not predict adherence. Technology and interaction differences predict adherence.	765	The Netherlands	13	58.85
8	[19]	Review	Helen Christensen, 2009	Louise Farrer	Randomized controlled trials of web interventions have lower dropout rates than open access websites. Internet intervention research and behavioral health literature both require theoretical models of adherence. Disease-related anxiety and sadness need more research.	749	Australia	16	46.81
9	[23]	Article	Andras Lanczky, 2012	Balazs Gyorffy	This study introduced a web-based tool capable of performing univariate and multivariate Cox proportional hazards survival analyses using data generated by genomic, transcriptomic, proteomic, or metabolomic studies.	730	Hungary	4	182.5
10	[20]	Review	Tara Donker, 2013	Helen Christensen	Apps for mental health have the potential to be very beneficial and increase treatment accessibility. The few evidence-based mental health apps that are now in the public domain need to be identified, and the public has to be made aware of this.	696	Australia	12	58.0

Analysis of Keywords Co-occurrence

Keyword co-occurrence analysis is a powerful bibliometric tool that accurately identifies current issues within the scientific knowledge structure. We conducted a co-occurrence analysis of keywords in *JMIR* by examining the frequency of their occurrences (Figure 6A). The 5 most frequently used keywords were internet (n=1524), eHealth (n=810), social media (n=904), COVID-19 (n=784), and health (n=770). The term "quality" exhibited the highest degree of centrality, with a value of 1.11. Additionally, a total of 12 distinct clusters were generated using the log-likelihood ratio to examine keyword co-occurrence (Figure 6B and Multimedia Appendix 17). The largest cluster (#0), designated "electronic health records," comprised 321 members and had a silhouette value of 0.566. The most frequently highlighted members of this cluster included patient portals and electronic health records. The second cluster, classified as "mental health," boasted a silhouette value of 0.67 and consisted of 181 members, with mental health, depression, and anxiety being the most commonly mentioned topics. The third cluster (#2), named "hypertension," comprised 152 nodes and had a silhouette value of 0.722. The 3 most frequently mentioned terms in this cluster were hypertension, diabetes, and medication adherence. The investigation of cluster #0, which focuses on "electronic health records," was initiated by the authors of *JMIR* in 1999 and has continued to the present day. After clustering, the timeline view of the keywords was analyzed and is shown in Figure 6C. The size of the nodes is directly proportional to the frequency of keyword occurrences. The color of the clusters indicates the evolution of research emphasis over time, with the most recently established connections represented in red or yellow and positioned at the top of the image. Initially, a significant number of researchers focused on the topic of the internet, followed by numerous studies examining diabetes and COVID-19. However, over the past 2 years, research has increasingly concentrated on electronic medical records and hypertension.

XSL•FO
RenderX

Figure 6. Analysis of keywords of Journal of Medical Internet Research (JMIR) publications: (A) keyword co-occurrence network (source: VOSviewer); (B) keyword cluster analysis (source: CiteSpace); (C) timeline view of keywords (source: CiteSpace).

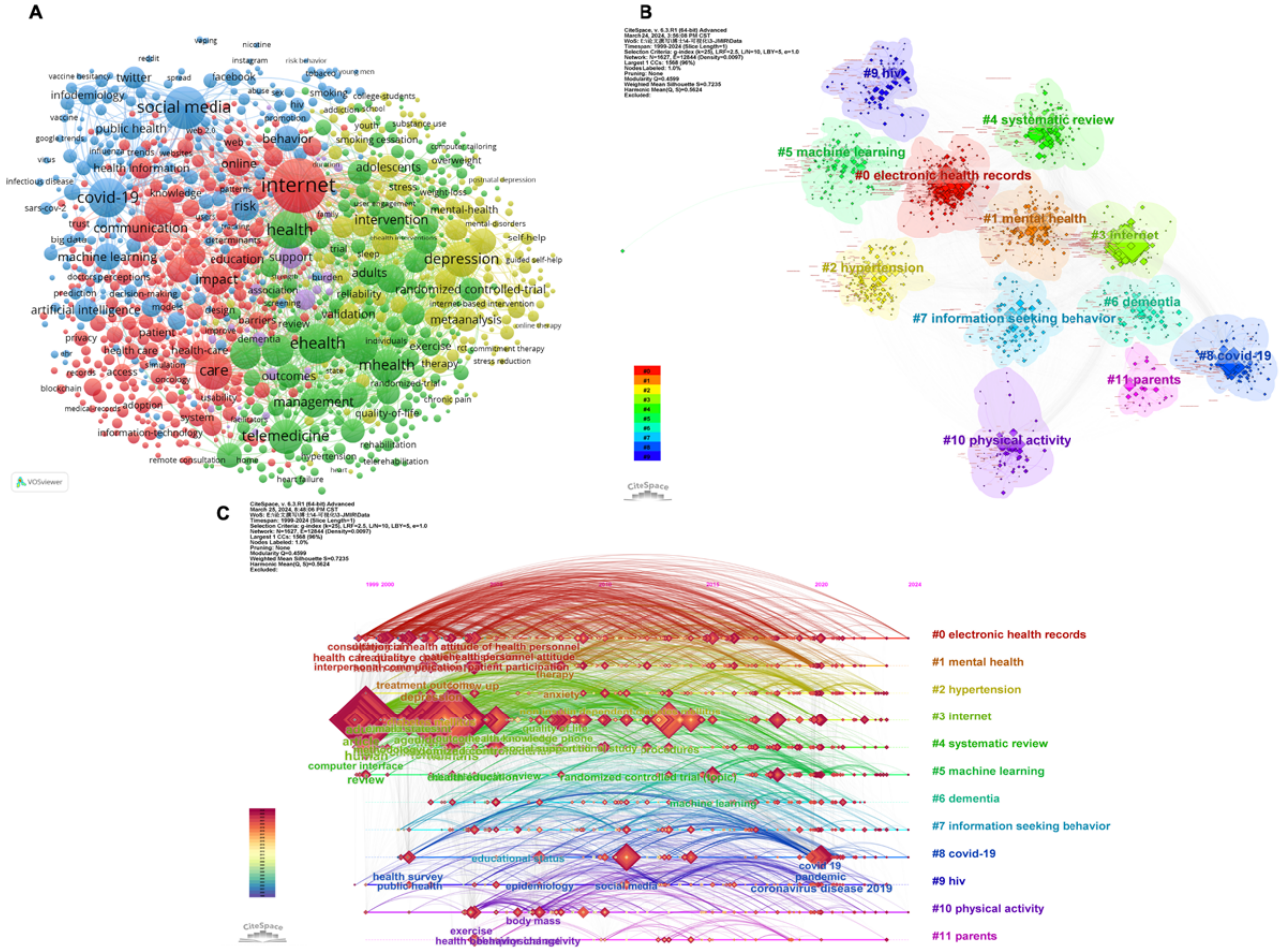

Changes in Trends of Research

The study conducted a thematic terminology analysis (Figure 7A) to examine the main issues present in *JMIR* publications. The x-axis represents the centrality of network clusters, indicating the degree of connectivity with other clusters in the graph, which reflects the significance of a research topic. The y-axis represents density, quantifying the internal robustness and growth of a cluster network's subject [38]. Clusters in the motor themes exhibit high centrality and density, suggesting that these topics are well-developed and essential for structuring research. Currently, the fields of digital health—specifically through social media and mobile phones—and public health, including infodemiology and COVID-19, are positioned in the motor themes quadrant. This indicates that extensive studies have been conducted on these issues, which are directly linked to the central theme of *JMIR*. Themes located in the emerging or declining quadrants exhibit both low centrality and low density, indicating that they are not well-developed and are on the margins. Investigations and preventive efforts targeting HIV and mental health issues, such as depression and anxiety, were situated in the third quadrant, suggesting that research in these areas is still in its nascent phase. Meanwhile, research on the application of artificial intelligence (AI)—including deep learning and machine learning—was positioned in the space between the first and third quadrants, indicating that while it remains in its early stages, it has a strong connection to the significant domain of digital health. Research on electronic health records and health informatics, positioned within the basic themes, has demonstrated that these areas require collaborative research across multiple disciplines [39]. Figure 7B illustrates the top 25 keywords with the strongest citation bursts. The keyword "quality" experienced the most significant increase in citations since 1999, with a burst strength of 14.12 from 2000 to 2011. From 2003 to 2016, "internet" exhibited the highest burst strength at 96.16. The latest trending keywords include "mobile phone" (n=70.38, 2021-2023) and "digital health" (n=66.91, 2021-2023). Additionally, the examination of trending topics revealed that "artificial intelligence," "patient empowerment," and "victimization" are emerging trends within this field (Figure 7C).

XSL•FO
RenderX

Figure 7. Analysis of keywords engaged in frailty and depression research. (A) thematic terminology analysis of keywords (source: the R package Bibliometrix); (B) the top 15 keywords with the strongest citation bursts (source: CiteSpace); (C) Trend topics of keywords (source: the R package Bibliometrix).

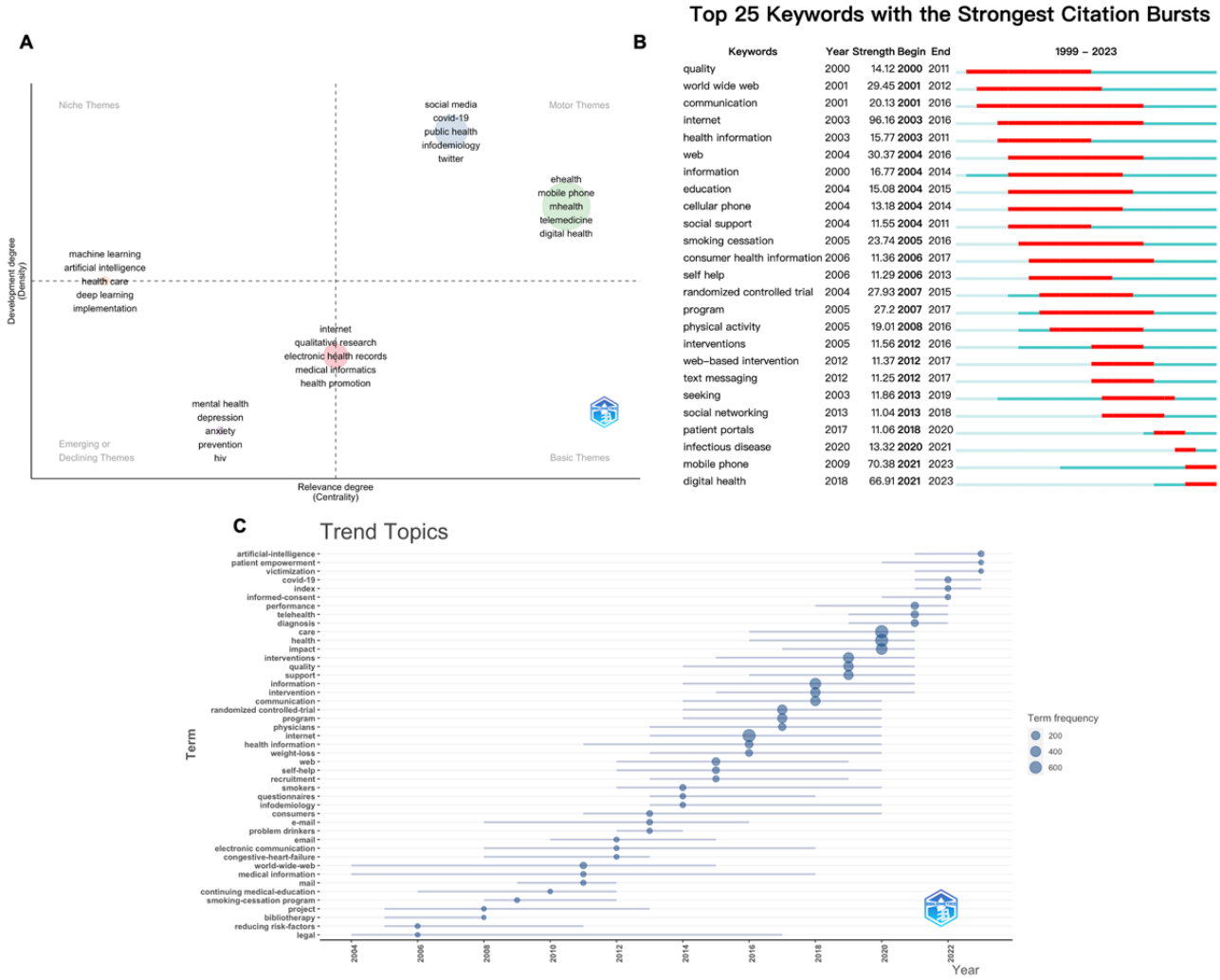

The Most Recently Published Articles in JMIR

To monitor the publishing patterns in *JMIR*, we conducted a search and analysis of recently published articles from 2024. Eight publications from *JMIR* were collected during the literature retrieval on February 22, 2024. A 3-field plot illustrated insights regarding the authors, countries, and keywords associated with these newly published articles (Multimedia Appendix 18). Among these, 5 papers focused on the implementation of telehealth programs in various contexts [40-44]. Jiang et al [40] and Bazzano et al [44] investigated chronic health conditions, while Chang et al [41] and Chen et al [43] focused on the impact of telehealth on cardiovascular health. In addition to examining the application of telehealth programs, Jiang et al [40] and Lim et al [45] researched the development and deployment of conversational agents. These software programs significantly enhance communication in health care, improving its overall effectiveness. The recently published articles also provided valuable insights for stakeholders on efficiently managing income and expenditures while optimizing telehealth initiatives [41,42,44].

Discussion

Principal Findings

To commemorate the *JMIR*'s 25th anniversary of publication in 2024, we conducted an advanced bibliometric study. The objective of this study was to analyze the bibliometric characteristics of the journal from 1999 to 2024. This involved investigating patterns and changes in research literature production, as well as identifying the most productive countries, institutions, authors, research terms, and topics, along with their evolution over time. Furthermore, we were interested in identifying emerging subjects through keyword analysis and trend tracking.

Our descriptive bibliometric analysis revealed that *JMIR* primarily focused on original articles (6306/7448, 84.67%; Multimedia Appendix 2) and review articles (1142/7448, 15.33%; Multimedia Appendix 2) related to health services and digital health. Before 2010, the journal published approximately the same number of articles each year. However, beginning in 2011, there was a significant increase in annual publication volume, which continued through 2023. *JMIR*'s annual publication volume peaked in 2022 and continued to rise into

XSL•FO
RenderX

2023 (n=309.75). The slight decrease in the annual number of articles from 2021 to 2024 may be attributed to the emergence of COVID-19 and the subsequent economic downturn. Additionally, *JMIR*'s increasing emphasis on publishing higher-quality articles could contribute to this trend. The mean number of citations per article was 30.36, significantly exceeding the journal's impact factor (7.4). This suggests that the articles published in *JMIR* are of high quality and have been extensively referenced by scholars worldwide. The correlation between the number of citations and the number of annual publications indicates that the journal consistently produces high-quality papers. Consequently, the volume of articles published has a substantial impact on the annual citation count.

In terms of the volume of articles published and the number of active institutions, the majority of contributing countries are developed nations with advanced levels of ecological and technological development. The United States has made the most significant contribution to *JMIR*, accounting for approximately 37.18% (2769/7448; Table 1) of all publications. Furthermore, 4 institutions based in the United States are among the top 10 most productive institutions. In terms of national productivity, the United Kingdom and China ranked second and third, respectively, with contributions of 12.74% (949/7448) and 10.71% (798/7448). Two institutions from the United Kingdom were included in the top 10 most productive institutions, while China had no relevant institutions in this category. Despite the presence of numerous research institutions in China, these findings suggest that there has been insufficient emphasis on digital health. The investigation also revealed that these publications were biased toward industrialized, English-speaking nations, potentially limiting the journal's influence and applicability for practitioners and scholars in other regions. In terms of article units, Switzerland (82/163, 50.3%), Spain (61/154, 39.6%), and Germany (148/384, 38.5%) had the highest MCP ratios (Multimedia Appendix 3), significantly surpassing that of the United States (326/2156, 15.1%). This suggests that these 3 countries were more engaged in multi-institutional collaboration. The national collaboration map indicated that the United States and China had the closest cooperation, with a frequency of 230, suggesting that the United States should enhance its multicountry exchanges and collaborative efforts.

Heleen Riper, a researcher from Vrije Universiteit Amsterdam, was the most productive author and collaborated extensively with other writers. However, the strongest cocitation link was identified between Gunther Eysenbach and Susannah Fox, with a link strength of 355. Notably, Gunther Eysenbach experienced the most significant increase in citations between 1999 and 2011, achieving a strength score of 11. Tobias Kowatsch is currently the most active author, with a strength score of 8.11 between 2021 and 2024. According to the authors' local citation map, Gunther Eysenbach is identified as the most influential author, with a total of 536 local citations. These results indicate that Eysenbach has significant influence in this domain and is dedicated to fostering international exchanges and collaborations in the digital health and health care sectors. As the editor-in-chief of *JMIR*, Gunther Eysenbach is affiliated with the University of Victoria. His research interests primarily revolve around health care, with a specific focus on health policy, eHealth, and consumer health informatics [22,46-49]. The analysis of the 3-field plot revealed that the most prominent authors originated from the Netherlands. These researchers mainly concentrated on eHealth [50-53], particularly emphasizing its impact on mental health [54-57] and the application of mobile phones [51,58] in their studies. Multiple analyses indicate that women authors remain inadequately represented, and this gender disparity may be attributed to unconscious bias in the peer review process, as well as insufficient empowerment and exploration of women's contributions in academia. This underscores the need for additional support for significant female authors in the future [59,60].

The top 10 globally cited publications (Table 2) and locally cited publications (Multimedia Appendix 15) from *JMIR* were summarized. Local citations refer to references from within *JMIR*, while global citations encompass citations across the entire academic literature. Local citations evaluate impact within a specific field, whereas global citations assess overall influence across disciplines [61]. The overlap between these 2 lists highlights the significance and relevance of certain publications within both the broader academic community and the specific field addressed by *JMIR*. For instance, the article titled "The Law of Attrition," published by editor Gunther Eysenbach in 2005 [22], has been cited 1615 times (WoSCC), ranking at the top of both lists. It emphasizes the importance of establishing a "science of consumption" and developing models to understand the termination of eHealth apps and the withdrawal of participants from eHealth trials. This work was essential in advancing and establishing the eHealth model. A subsequent review conducted in 2010 evaluated various theoretical and behavioral change methodologies to identify the most effective internet-based treatments for promoting healthy behavioral change [17]. A novel scoring system has been developed to assess internet-based interventions and efficient connection models [37]. This framework serves as a guide for the future implementation of internet-based therapies. In 2020, Son et al [21] published a study titled "Effects of COVID-19 on College Students' Mental Health in the United States: Interview Survey Study," which highlighted the importance of prioritizing the mental well-being of students affected by the pandemic [21]. Further analysis revealed that the papers with the highest citations among the top 10 primarily focused on mental health, internet-based tools, and eHealth.

Keywords serve as valuable tools for identifying core themes, primary study subjects, and content within a specific domain. Based on the co-occurrence analysis conducted using VOSviewer, the themes most extensively researched in *JMIR* were "internet" (n=1524), "eHealth" (n=901), and "social media" (n=810). Additionally, the analysis revealed 4 clusters, each represented by different colors. "Internet" emerged as the keyword with the highest frequency and was most commonly associated with other keywords. COVID-19 was the most prominent disease of concern in this journal and demonstrated the strongest connection with "social media." This finding aligns with the perspective that COVID-19 has accelerated the development and adoption of digital technologies, including

XSL•FO
RenderX

information dissemination, health communication, data collection, community support, behavioral insights, and the promotion of digital health tools [62,63]. *JMIR* is particularly focused on how public health crises impact human health and the potential benefits of digital health. Since the onset of COVID-19, *JMIR* has published 3 themes related to the epidemic. According to the clustering results from CiteSpace, the prominent keyword clusters included "electronic health records," "mental health," and "internet." However, *JMIR* has also featured articles addressing "hypertension," "dementia," and "HIV." The silhouette scores for all clusters were greater than 0.5, indicating highly credible clusters [64]. To remain relevant and impactful in the rapidly expanding field of digital health, *JMIR* may need to broaden the range of subjects it investigates. Based on the thematic terminology analysis of keywords from the R package Bibliometrix, the themes of "machine learning," "health care," and "implementation" were found to be underdeveloped and marginalized. These themes represent specific research areas that have achieved a certain level of maturity but still require significant contributions to the field. By contrast, the research themes of "depression," "anxiety," "prevention," and "HIV" are either experiencing notable advancements or a decline in research interest. The COVID-19 pandemic and HIV significantly impact global health, with anxiety and despair closely associated with these crises [65,66]. An examination of keywords has shown that the emergence of digital health, driven by the COVID-19 pandemic, has garnered significant attention in recent years (2021-2023).

The current and future key focal points of *JMIR* include "artificial intelligence," "patient empowerment," and "victimization." With the advancement of computer science and technology, AI is praised as a possible agent of change that will revolutionize the field of medicine as well as digital health. *JMIR* has published 7 AI-related themes, encompassing areas such as communication, detection, diagnosis, treatment, and analysis. It has been observed that AI's capabilities significantly enhance diagnostic accuracy, personalize treatment, improve operational efficiency, and support clinical decision-making. These advancements in digital health are leading to improved patient outcomes, enhanced care delivery, and increased operational efficiencies within the health care system [67,68]. Furthermore, digital health technologies play a pivotal role in enhancing patient empowerment by providing tools and resources that enable patients to actively manage their health [69]. AI further promotes patient engagement and empowerment by delivering individualized health services, including education, decision-making aids, and interactive support. The contribution of digital health to patient empowerment is an area of growing interest for *JMIR*, as evidenced by the recent publication of 122 articles utilizing this keyword. Another trending topic is "victimization" with *JMIR* recently publishing 67 related articles. Risks in both daily and digital environments underscore the importance of patient empowerment. *JMIR* integrates digital health into support systems for persons who experienced violence to enhance the accessibility, effectiveness, and safety of digital health services. There is a growing concern and interest in these topics, reflecting their increasing relevance in the field of digital health.

We propose several strategies to enhance the impact and relevance of *JMIR*: (1) advocating for the adoption of novel and advanced research methods, along with cutting-edge statistical techniques for data analysis, such as data linkage, machine learning, and AI; (2) fostering journal submissions by establishing partnerships with universities and research institutions, or by creating country-specific special issues to broaden the journal's international scope and include articles from a more diverse range of geographical locations; (3) promoting author diversity to ensure a wider representation of ideas and perspectives in the journal; and (4) prioritizing research on topics that address emerging themes and encourage cross-disciplinary investigations.

Limitations

This study uses a bibliometric approach to comprehensively examine the *JMIR* literature published between 1999 and 2024. However, this study has certain limitations. All publication data were obtained from the WoSCC, which means that some publications may have been inadvertently overlooked. Additionally, the accuracy of the findings may be affected when converting the data format to merge documents from multiple databases, as each database may have distinct characteristics. Nonetheless, the WoSCC database is considered the most representative, as it encompasses a majority of the prominent and essential scholarly journals worldwide. Therefore, we assert that the current research findings accurately reflect the global state of this discipline. Another unavoidable issue is the lack of transparency in the algorithms used by bibliometric software, which may lead to algorithmic bias. To address this concern, 3 distinct bibliometric analysis software programs were utilized to generate a comprehensive set of information.

Conclusions

A bibliometric analysis conducted from 1999 to 2024 found that *JMIR* published a total of 7780 articles authored by 32,232 individuals. The journal featured a diverse group of authors from various continents; however, the United States emerged as the most prolific country, with its institutions making significant contributions. The number of articles published in *JMIR* increased steadily, with a notable spike in 2022. The highest average number of publications, 309.75, was recorded in 2023. The study also identifies the most productive and innovative subjects published in *JMIR*, highlighting promising avenues for future research. Additionally, this analysis provides valuable insights into the journal's history and current standing, offering readers and potential contributors essential information for future development.

XSL•FO
RenderX

Acknowledgments

This study was supported by the Department of Science and Technology of Sichuan Province (grant 2024NSFSC1575) and the 135 Project of West China Hospital (grant ZYGD18018). The funders played no role in the design, conduct, or reporting of this study.

Authors' Contributions

JH and CL conceived the idea of the manuscript, collected the publications, conducted the bibliometric analysis, and drafted the manuscript. YLG, JYY, and SYZ contributed to drafting and critically reviewed the manuscript. JH and CQH acquired the funding. CQH contributed to supervision and program administration. All authors read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart of bibliometric analysis.
[PNG File , 170 KB - jmir_v26i1e58987_app1.png]

Multimedia Appendix 2
A descriptive summary of JMIR publications from 1999 to 2024 (source: the R package Bibliometrix).
[DOCX File , 16 KB - jmir_v26i1e58987_app2.docx]

Multimedia Appendix 3
Corresponding author's countries (source: the R package Bibliometrix).
[DOCX File , 16 KB - jmir_v26i1e58987_app3.docx]

Multimedia Appendix 4
The top 25 countries' collaboration (source: the R package Bibliometrix).
[DOCX File , 16 KB - jmir_v26i1e58987_app4.docx]

Multimedia Appendix 5
The timeline view of coauthorship analysis network of country (source: VOSviewer).
[PNG File , 893 KB - jmir_v26i1e58987_app5.png]

Multimedia Appendix 6
Collaborations among scholar institutions (source: VOSviewer).
[DOCX File , 16 KB - jmir_v26i1e58987_app6.docx]

Multimedia Appendix 7
The timeline view of coauthorship analysis network of institutions (source: VOSviewer).
[PNG File , 728 KB - jmir_v26i1e58987_app7.png]

Multimedia Appendix 8
The top 10 most frequently published authors (source: Web of Science Core Collection).
[DOCX File , 17 KB - jmir_v26i1e58987_app8.docx]

Multimedia Appendix 9
The top 25 authors with the strongest citation bursts (source: CiteSpace).
[PNG File , 332 KB - jmir_v26i1e58987_app9.png]

Multimedia Appendix 10
The top 10 most frequently collaborated authors (source: VOSviewer).
[DOCX File , 16 KB - jmir_v26i1e58987_app10.docx]

Multimedia Appendix 11
The top 10 influential authors among cocitation analysis (source: VOSviewer).

XSL•FO
RenderX

[DOCX File , 16 KB - jmir_v26i1e58987_app11.docx]

Multimedia Appendix 12
Author's productivity through Lotka's law (source: the R package Bibliometrix).
[DOCX File , 17 KB - jmir_v26i1e58987_app12.docx]

Multimedia Appendix 13
The top 25 most citing sources by the JMIR authors (source: Web of Science Core Collection).
[DOCX File , 16 KB - jmir_v26i1e58987_app13.docx]

Multimedia Appendix 14
The top 10 most locally cited sources (source: the R package Bibliometrix).
[PNG File , 100 KB - jmir_v26i1e58987_app14.png]

Multimedia Appendix 15
The top 10 most locally cited documents (source: the R package Bibliometrix).
[DOCX File , 24 KB - jmir_v26i1e58987_app15.docx]

Multimedia Appendix 16
The top 10 citing references by the JMIR authors (source: Web of Science Core Collection).
[DOCX File , 25 KB - jmir_v26i1e58987_app16.docx]

Multimedia Appendix 17
Summary of 11 clusters analysis for keywords (source: CiteSpace).
[DOCX File , 19 KB - jmir_v26i1e58987_app17.docx]

Multimedia Appendix 18
The 3-field plot graph of the 8 publications in 2024 (source: the R package Bibliometrix).
[PNG File , 243 KB - jmir_v26i1e58987_app18.png]

References

1. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
2. Michael Besser G, Pfeiffer RF, Thorner MO. Anniversary review: 50 years since the discovery of bromocriptine. Eur J Endocrinol 2018 Aug;179(2):R69-R75. [doi: 10.1530/EJE-18-0378] [Medline: 29752299]
3. Daws LC. 75th year anniversary: past and future-Editorial. Pharmacol Rev 2023 Dec 15;76(1):1-6. [doi: 10.1124/pharmrev.123.000989] [Medline: 38101933]
4. Kamdem JP, Duarte AE, Lima KRR, Rocha JBT, Hassan W, Barros LM, et al. Research trends in food chemistry: A bibliometric review of its 40 years anniversary (1976-2016). Food Chem 2019 Oct 01;294:448-457. [doi: 10.1016/j.foodchem.2019.05.021] [Medline: 31126486]
5. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim W. How to conduct a bibliometric analysis: an overview and guidelines. Journal of Business Research 2021 Sep;133:285-296 [FREE Full text] [doi: 10.1016/j.jbusres.2021.04.070]
6. Bamel N, Kumar S, Bamel U, Lim WM, Sureka R. The state of the art of innovation management: insights from a retrospective review of the European Journal of Innovation Management. EJIM 2022 Sep 22;27(3):825-850. [doi: 10.1108/ejim-07-2022-0361]
7. Martorell Cunill O, Socias Salvá A, Otero Gonzalez L, Mulet-Forteza C. Thirty-fifth anniversary of the International Journal of Hospitality Management: a bibliometric overview. International Journal of Hospitality Management 2019 Apr;78:89-101 [FREE Full text] [doi: 10.1016/j.ijhm.2018.10.013]
8. Nath S, Thomson WM, Baker SR, Jamieson LM. A bibliometric analysis of community dentistry and oral epidemiology: fifty years of publications. Community Dent Oral Epidemiol 2024 Apr;52(2):171-180. [doi: 10.1111/cdoe.12910] [Medline: 37798876]
9. Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci 2005 Dec 14;57(3):359-377. [doi: 10.1002/asi.20317]
10. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
11. van Eck N, Waltman L. Visualizing bibliometric networks. In: Measuring Scholarly Impact: Methods and Practice. Cham, Switzerland: Springer International Publishing; Sep 24, 2014:285-320.

XSL•FO
RenderX

12. Arruda H, Silva ER, Lessa M, Proença D, Bartholo R. VOSviewer and Bibliometrix. J Med Libr Assoc 2022 Jul 01;110(3):392-395 [FREE Full text] [doi: 10.5195/jmla.2022.1434] [Medline: 36589296]
13. Kawamura M, Thomas CD, Tsurumoto A, Sasahara H, Kawaguchi Y. Lotka's law and productivity index of authors in a scientific journal. J Oral Sci 2000 Jun;42(2):75-78. [doi: 10.2334/josnusd.42.75] [Medline: 10989588]
14. Chen C. Science mapping: a systematic review of the literature. Journal of Data and Information Science 2017 Mar 21;2(2):1-40 [FREE Full text] [doi: 10.1515/jdis-2017-0006]
15. Chen YA, Liang S, Shih Y, Chen C, Lee Y, Chang L, et al. Epidemiological and genetic correlates of severe acute respiratory syndrome coronavirus infection in the hospital with the highest nosocomial infection rate in Taiwan in 2003. J Clin Microbiol 2006 Feb;44(2):359-365 [FREE Full text] [doi: 10.1128/JCM.44.2.359-365.2006] [Medline: 16455884]
16. Chen C, Leydesdorff L. Patterns of connections and movements in dual‐map overlays: a new method of publication portfolio analysis. Asso for Info Science & Tech 2013 Nov 01;65(2):334-351 [FREE Full text] [doi: 10.1002/asi.22968]
17. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
18. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
19. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
20. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
21. Son C, Hegde S, Smith A, Wang X, Sasangohar F. Effects of COVID-19 on college students' mental health in the United States: interview survey study. J Med Internet Res 2020 Sep 03;22(9):e21279 [FREE Full text] [doi: 10.2196/21279] [Medline: 32805704]
22. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
23. Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. J Med Internet Res 2021 Jul 26;23(7):e27633 [FREE Full text] [doi: 10.2196/27633] [Medline: 34309564]
24. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
25. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
26. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
27. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
28. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
29. Brouwer W, Kroeze W, Crutzen R, de Nooijer J, de Vries NK, Brug J, et al. Which intervention characteristics are related to more exposure to internet-delivered healthy lifestyle promotion interventions? A systematic review. J Med Internet Res 2011 Jan 06;13(1):e2 [FREE Full text] [doi: 10.2196/jmir.1639] [Medline: 21212045]
30. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
31. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
33. Erdfelder E, Faul F, Buchner A. GPOWER: a general power analysis program. Behavior Research Methods, Instruments, & Computers 1996 Mar;28(1):1-11. [doi: 10.3758/BF03203630]
34. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

35. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
36. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
37. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
38. Cobo M, Martínez M, Gutiérrez-Salcedo M, Fujita H, Herrera-Viedma E. 25 years at Knowledge-Based Systems: a bibliometric analysis. Knowledge-Based Systems 2015 May;80:3-13 [FREE Full text] [doi: 10.1016/j.knosys.2014.12.035]
39. Agbo FJ, Oyelere SS, Suhonen J, Tukiainen M. Scientific production and thematic breakthroughs in smart learning environments: a bibliometric analysis. Smart Learn Environ 2021 Jan 15;8(1):1-25. [doi: 10.1186/s40561-020-00145-4]
40. Jiang Z, Huang X, Wang Z, Liu Y, Huang L, Luo X. Embodied conversational agents for chronic diseases: scoping review. J Med Internet Res 2024 Jan 09;26:e47134 [FREE Full text] [doi: 10.2196/47134] [Medline: 38194260]
41. Chang HY, Wu HW, Hung CS, Chen YH, Huang C, Yang L, et al. Costs and cardiovascular benefits of a fourth-generation synchronous telehealth program on mortality and cardiovascular outcomes for patients with atrial fibrillation: retrospective cohort study. J Med Internet Res 2024 Jan 08;26:e48748 [FREE Full text] [doi: 10.2196/48748] [Medline: 38190237]
42. He X, Cui F, Lyu M, Sun D, Zhang X, Shi J, et al. Key factors influencing the operationalization and effectiveness of telemedicine services in Henan Province, China: cross-sectional analysis. J Med Internet Res 2024 Jan 05;26:e45020 [FREE Full text] [doi: 10.2196/45020] [Medline: 38180795]
43. Chen Y, Turkson-Ocran R, Koirala B, Davidson PM, Commodore-Mensah Y, Himmelfarb CD. Association between the composite cardiovascular risk and mHealth use among adults in the 2017-2020 Health Information National Trends Survey: cross-sectional study. J Med Internet Res 2024 Jan 04;26:e46277 [FREE Full text] [doi: 10.2196/46277] [Medline: 38175685]
44. Bazzano AN, Patel T, Nauman E, Cernigliaro D, Shi L. Optimizing telehealth for diabetes management in the deep south of the United States: qualitative study of barriers and facilitators on the patient and clinician journey. J Med Internet Res 2024 Jan 01;26:e43583 [FREE Full text] [doi: 10.2196/43583] [Medline: 37976468]
45. Lim WA, Custodio R, Sunga M, Amoranto AJ, Sarmiento RF. General characteristics and design taxonomy of chatbots for COVID-19: systematic review. J Med Internet Res 2024 Jan 05;26:e43112 [FREE Full text] [doi: 10.2196/43112] [Medline: 38064638]
46. Leung TI, de Azevedo Cardoso T, Mavragani A, Eysenbach G. Best practices for using AI tools as an author, peer reviewer, or editor. J Med Internet Res 2023 Aug 31;25:e51584 [FREE Full text] [doi: 10.2196/51584] [Medline: 37651164]
47. Grajales FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
48. Wiljer D, Urowitz S, Apatu E, DeLenardo C, Eysenbach G, Harth T, Canadian Committee for Patient Accessible Health Records. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res 2008 Oct 31;10(4):e34 [FREE Full text] [doi: 10.2196/jmir.1061] [Medline: 18974036]
49. Eysenbach G. Infodemiology: The epidemiology of (mis)information. Am J Med 2002 Dec 15;113(9):763-765. [doi: 10.1016/s0002-9343(02)01473-0] [Medline: 12517369]
50. Ebert DD, Franke M, Zarski A, Berking M, Riper H, Cuijpers P, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021 Dec 22;23(12):e22107 [FREE Full text] [doi: 10.2196/22107] [Medline: 34941541]
51. Zarski A, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146 [FREE Full text] [doi: 10.2196/jmir.4493] [Medline: 27357528]
52. Boß L, Lehr D, Reis D, Vis C, Riper H, Berking M, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234 [FREE Full text] [doi: 10.2196/jmir.5952] [Medline: 27582341]
53. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
54. Warmerdam L, van Straten A, Twisk J, Riper H, Cuijpers P. Internet-based treatment for adults with depressive symptoms: randomized controlled trial. J Med Internet Res 2008 Nov 20;10(4):e44 [FREE Full text] [doi: 10.2196/jmir.1094] [Medline: 19033149]
55. Etzelmueller A, Vis C, Karyotaki E, Baumeister H, Titov N, Berking M, et al. Effects of internet-based cognitive behavioral therapy in routine care for adults in treatment for depression and anxiety: systematic review and meta-analysis. J Med Internet Res 2020 Aug 31;22(8):e18100 [FREE Full text] [doi: 10.2196/18100] [Medline: 32865497]

XSL•FO
RenderX

56. Warmerdam L, Smit F, van Straten A, Riper H, Cuijpers P. Cost-utility and cost-effectiveness of internet-based treatment for adults with depressive symptoms: randomized trial. J Med Internet Res 2010 Dec 19;12(5):e53 [FREE Full text] [doi: 10.2196/jmir.1436] [Medline: 21169166]
57. Bolier L, Haverman M, Kramer J, Westerhof GJ, Riper H, Walburg JA, et al. An Internet-based intervention to promote mental fitness for mildly depressed adults: randomized controlled trial. J Med Internet Res 2013 Sep 16;15(9):e200 [FREE Full text] [doi: 10.2196/jmir.2603] [Medline: 24041479]
58. Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. J Med Internet Res 2016 Mar 29;18(3):e72 [FREE Full text] [doi: 10.2196/jmir.5505] [Medline: 27025287]
59. Sullivan S, Posada A, Hawkins M, Higueras G. Where are the female experts? Perceptions about the absence of female presenters at a Digital Health Conference in Bolivia. Mhealth 2020;6:37 [FREE Full text] [doi: 10.21037/mhealth-2019-di-12] [Medline: 33437834]
60. The Lancet Digital Health. Empowering women in health technology. The Lancet Digital Health 2022 Mar;4(3):e149. [doi: 10.1016/s2589-7500(22)00028-0]
61. Saravanan P, Rajeswari S, Kumar JA, Rajasimman M, Rajamohan N. Bibliometric analysis and recent trends on MXene research - a comprehensive review. Chemosphere 2022 Jan;286(Pt 3):131873. [doi: 10.1016/j.chemosphere.2021.131873] [Medline: 34411934]
62. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
63. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
64. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 Nov;20:53-65 [FREE Full text] [doi: 10.1016/0377-0427(87)90125-7]
65. Mazza MG, De Lorenzo R, Conte C, Poletti S, Vai B, Bollettini I, COVID-19 BioB Outpatient Clinic Study group, et al. Anxiety and depression in COVID-19 survivors: role of inflammatory and clinical predictors. Brain Behav Immun 2020 Oct;89:594-600 [FREE Full text] [doi: 10.1016/j.bbi.2020.07.037] [Medline: 32738287]
66. Nightingale S, Ances B, Cinque P, Dravid A, Dreyer AJ, Gisslén M, et al. Cognitive impairment in people living with HIV: consensus recommendations for a new approach. Nat Rev Neurol 2023 Jul;19(7):424-433. [doi: 10.1038/s41582-023-00813-2] [Medline: 37311873]
67. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022 Oct 05;24(10):e40238 [FREE Full text] [doi: 10.2196/40238] [Medline: 36197712]
68. Gilbert S, Anderson S, Daumer M, Li P, Melvin T, Williams R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. J Med Internet Res 2023 Apr 14;25:e43682 [FREE Full text] [doi: 10.2196/43682] [Medline: 37058329]
69. Calvillo J, Román I, Roa LM. How technology is empowering patients? A literature review. Health Expect 2015 Oct;18(5):643-652 [FREE Full text] [doi: 10.1111/hex.12089] [Medline: 23711169]

Abbreviations

AC: average citation
AI: artificial intelligence
JMIR: Journal of Medical Internet Research
MCP: multiple-country publication
TC: total citation
TP: total publication
WoSCC: Web of Science Core Collection

Edited by G Eysenbach; submitted 30.03.24; peer-reviewed by C Zhang, JH Chien, J Zhou, R Xu; comments to author 23.05.24; revised version received 28.05.24; accepted 05.09.24; published 17.10.24.

Please cite as:
Hu J, Li C, Ge Y, Yang J, Zhu S, He C
Mapping the Evolution of Digital Health Research: Bibliometric Overview of Research Hotspots, Trends, and Collaboration of Publications in JMIR (1999-2024)
J Med Internet Res 2024;26:e58987
URL: https://www.jmir.org/2024/1/e58987
doi:10.2196/58987
PMID:

XSL•FO
RenderX

©Jing Hu, Chong Li, Yanlei Ge, Jingyi Yang, Siyi Zhu, Chengqi He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Transitioning Perspectives in Digital Health Through Phenomenology Integration

Maddalena Fiordelli[1], PhD

Institute of Public Health, Faculty of Biomedical Sciences, Università della Svizzera italiana, Lugano, Switzerland

Corresponding Author:
Maddalena Fiordelli, PhD
Institute of Public Health, Faculty of Biomedical Sciences
Università della Svizzera italiana
Via Buffi 13
Lugano, 6900
Switzerland
Phone: 41 586664139
Email: maddalena.fiordelli@usi.ch

Abstract

The evolution of digital health, from its early days as eHealth to its current expansive scope, reflects a significant transformation in health care delivery and management. This transition underscores the integration of digital technologies across the health continuum from prevention and diagnosis to treatment and rehabilitation. The emergence of digital health has introduced innovative solutions but also posed challenges, particularly in aligning technological advancements with health needs, human experiences, and ethical considerations. This position paper aims to explore the integration of phenomenology in digital health, advocating for a paradigm that emphasizes the centrality of human experience in the design and implementation of digital health solutions. It specifically seeks to address challenges related to relevance for individuals who "speak" different languages, ensuring long-term use, addressing digital and health literacy, coordinating various sources, and navigating ethical issues in the rapidly evolving digital health landscape. Drawing upon years of research and practical experience in communication technologies and health, this paper uses a reflective approach to examine the intersection of digital health and phenomenology. It reviews the historical development of digital health, identifies the challenges faced during its evolution, and discusses the potential of phenomenological methods to enhance user-centered design and ethical practices in digital health. The integration of phenomenology into digital health facilitates a deeper understanding of user experiences, enabling the development of more responsive and ethical digital health solutions. Participatory design models, informed by phenomenological perspectives, offer a pathway to bridge the gap between technological innovation and human-centric health care. The paper highlights successful practices in digital health development, including mobile apps for vaccination decision-making and platforms for managing chronic conditions, illustrating the benefits of a phenomenological approach. Transitioning perspectives in digital health through phenomenology integration represents a critical step toward realizing the full potential of digital technologies in health care.

(J Med Internet Res 2024;26:e62691) doi:10.2196/62691

KEYWORDS

eHealth; digital health; phenomenology; phenomenological; participatory; health communication; health information; active listening; lived experience

The Promise of Digital Health

In 2001, this journal published an editorial listing the 10 characteristics, the 10 "e's" of eHealth, and providing a definition that was later extensively cited. The definition referred to eHealth as "an emerging field at the intersection of medical informatics, public health and business" [1]. But the multidisciplinary nature of eHealth was just a part of this visionary definition, which further underscored that "the term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide." The potential and the promises encapsulated in this term were scrutinized for years to come. Today, the term seems far away, as it has been gradually replaced by the broader and more modern expression "digital health." The latter better represents a complex reality by encompassing in the definition fast technological advancements together with social changes. Revisiting the 1970s reveals the emergence of various terminologies (eg, telematic and

XSL•FO
RenderX

telehealth) [2], which have been stepwise encompassed in an interdisciplinary field and endeavor. According to World Health Organization (WHO) definition, "Digital health expands the concept of eHealth to include digital consumers" [3]; therefore, to reflect a part of the process, the end user can no longer be neglected but should become the core.

Digital health inherits from eHealth the disruptive potential of going beyond the technical development: it holds many promises. Digital health serves disparate purposes according to the health frameworks adopted, technological solutions used, foreseen aims, and contexts of adoption. If we consider the care continuum—from promotion to primary, secondary, and tertiary prevention, treatment, and recovery—digital health solutions exist for each stage, though with varying degrees of emphasis [4]. In terms of technological applications, we can enumerate many categories, such as geospatial and environmental, knowledge generators, imaging, personalized therapeutics, and so forth [5]. Digital health also started pervading not only different health settings, such as classical health care contexts like hospitals, specialized clinics, and private practice, but also citizens' daily life and occupational environments (eg, well-being apps). Looking at the communication that digital health solutions facilitate or at the health communication aims they realize, we can enumerate the provision of information and motivational support, as well as patient engagement, facilitation of peer networking, and the possibility to meet individual patient needs in ways otherwise impossible [6].

The Challenges Encountered by Digital Health

But beyond these promises, numerous challenges emerge, especially during the implementation phase of digital health solutions (Textbox 1). The discrepancy between market development and research outcomes is a primary challenge, already documented in systematic reviews 10 years ago [7]. The reality is that the market is pushing forward, but the consequence is that we often do not have solutions that are based on either theory or evidence where effectiveness remains unclear.

Textbox 1. Challenges to the promises of digital health.

- Discrepancy between market development and research outcomes
- Quality and accessibility of information
- Health and digital literacy
- Multidisciplinary nature and communication challenges
- Integration into daily routines
- Coordination of stakeholders
- Sustainability
- Ethical concerns

Digital health develops following the evolution of information and communication technologies. This major change makes information, which is at the heart of health and health care, more accessible to everyone and increasingly present. However, as neologism like *infodemic* and recent scientific endeavor clearly showed [8], not all this information is of good quality, and this quality needs to be discerned at various levels.

Other crucial issues are related to the competence of the population, whether it be health literacy or digital health literacy [9]. Such issues are mostly investigated in specific subgroups of the population. We often discuss the digital divide among older adults in a simplistic way, viewing it primarily as a barrier to the adoption of digital health. However, current evidence is mixed, and older adults have demonstrated the ability to use technology effectively when the advantages and purposes are clear to them [10]. This example highlights that matters of literacy need to be framed in a broader and more complex context.

Digital health is multidisciplinary, and, in its development and implementation, it involves multiple audiences who speak different languages. In this context, health communication plays an important role as it has the potential to design proper strategies and to build the necessary common ground for reciprocal understanding [11]. Digital health solutions are attractive and innovative, but their adoption fails to endure over time [12,13] because it entails proper integration into daily habits, and this is a significant change that often gets unattended.

When promising digital health solutions have to be implemented, there are various audiences and sources that need to be coordinated, and the perspectives of these stakeholders must be taken into consideration [14]. If the intricacies of their power, legitimacy, and urgency are not considered, they might hinder the process [15]. Finally, 2 major chapters relate to the sustainability of technology and solutions on the one hand and ethical issues on the other. What often constitute a barrier to broad adoption of digital health solutions are indeed ethical matters [16] that might scare citizen and institutions: concerns about privacy, data governance, and so forth.

We could yet borrow the words of Gary Kreps [17] from a decade ago: "There is a long way to go for digital health information systems to reach their incredible potential." If we compare the initial expectations about digital health, which were simplified perspectives that made us imagine easily achievable goals, with this more realistic outlook, some spontaneous questions arise: can we truly keep the initial promise? And if so, what are the concrete steps to do so?

XSL•FO
RenderX

The Intersection of Digital Health and Phenomenology Through the Progressive Integration of Participatory Approaches

The chronological analysis of information and communication technologies and their development provides an additional perspective. Initially, the focus was almost exclusively on the content to be conveyed and its structure. When the web started its broader diffusion in the 90s, the approach to design was pretty much content based [18]. The web constituted an unprecedented opportunity of delivering information in different format, and the highest priority from the communication perspective was the one of the "few" senders back then, who decided how to frame and convey the message in the best way conceived. However, over the years, there has been a significant evolution, with increasing emphasis on the end user in the design process. It was about 10 years later that user-centered design approaches started to mushroom and eventually developed into in co-design, and later cocreation approaches [19].

This progression has led to the integration of participatory design models [20], highlighting the approach and sensitivity that are typical of the humanities and social sciences [21]. In addition to being participatory, this approach adopts a phenomenological perspective, valuing the experiences and voices of the user and even the patient. Wang and colleagues [22], in their umbrella review, outline a framework that facilitates the integration of digital experience at every stage of the development of digital health technologies and in all its iterations.

Phenomenology is the philosophical approach aimed at studying phenomena as they are experienced by individuals. Adopting a phenomenological approach therefore largely entails the incorporation of lived experience into the investigation [23]. This entails using the Husserlian lens of phenomenology [24], which underscores the necessity of *epochè* as an essential methodological approach for the examination of experiences. This is crucial in the development of conscious digital health solutions. By suspending preconceptions and assumptions, this methodology enables a critical examination and understanding of users' lived experiences, thereby uncovering their genuine needs and challenge. This approach ensures that digital health solutions are authentically responsive to human contexts, fostering the development of empathetic, user-centered technologies that integrate seamlessly into daily life. Incorporating phenomenology into the development process facilitates the creation of innovative, ethically sound digital health tools that enhance the complexity of human experiences and lead to meaningful health outcomes.

The arguments presented in this paragraph illustrate how phenomenology could and should be integrated into the whole recurring cycle of design, development, testing, and implementation of a digital health solution. This contribution advocates for a paradigm that emphasizes the centrality of human experience in the design and implementation of digital health solutions [21]. In that, participatory design models constitute the approach adopted by successful practices in the digital health solution development and implementation cycle.

Potential Pitfalls to the Integration

However, it is important to be aware that participatory approaches adoption for phenomenological integration can face various challenges. The first challenge concerns the complexity of identifying and involving all those who are directly or indirectly affected by the proposed technology. The direct and indirect beneficiaries or users of a digital health solution could be many, and their involvement is resource intensive. As it is for more classical research participation, involvement of the public into digital health solutions design, development, and implementation could face challenges in involving people [25,26]. It is essential to ensure the necessary resources to support participatory methods, which require a greater commitment in terms of time and reflection. This includes ensuring that those engaging with participatory methods hold the required sensitivity and competencies.

The second challenge addresses the need to balance the tension between the co-design phase as a separate entity and the existence of continuous and iterative co-design, as illustrated in the previous framework. So far, several attempts have been made throughout the process, but a systematic methodology has not yet been developed, possibly due to the inherently flexible nature of participatory methods [27]. Moreover, this approach is relatively recent and not yet widely shared. Acknowledging key methods, such as Q Methodology and phenomenological interviews [28-31] to name a few, which facilitate the integration of lived experiences, is a crucial first step toward systematizing this approach.

A third challenge is suggested by post phenomenological analysis and mediation theory [32], which demonstrate that technologies in an embodiment relation alter users by transforming their actions, motivations, and perceptions and making it impossible to revert to a pretechnological state without losing these changes. This highlights even further the importance of reflexivity in those who are collecting and analyzing the lived experiences of all those concerned with the digital health solution and, particularly, the end users [33].

Finally, there is the inherent tension between the 2 dimensions of digital health, namely, the health care sector and the technology sector. This tension manifests in terms of differences in implementation timelines and objectives and often complicates direct dialogue between the 2 spheres [5]. This poses many issues when developing and most of all implementing digital health, as it represents the root to the resistances from health care professionals [30]. In that, the listening and the integration of stakeholders' perspectives remain though the only viable solution, as health care professional and technology developers' lived experiences are essential to incorporate.

The Empirical Demonstration of a Viable Option

The content of this work is informed by the practices I share in the following, which represent direct working experiences of the past 10 years of my research career. I have had the privilege

XSL•FO
RenderX

of participating in various projects for the development, adaptation, and evaluation of digital health interventions. The chronological trajectory of my experiences also reflects the evolving sensitivity toward participatory design methods as a way of integrating phenomenology, highlighting and witnessing their maturation over time. These 3 case studies exemplify that.

The first practice concerns the development of a mobile app aimed at helping parents inform themselves and make decisions about their children's vaccinations [34,35]. This application was conceived following a top-down approach, based on existing evidence and theories. However, during the evaluation phase, the use of mixed methods allowed us to capture the direct experiences of users. Within Wan and colleagues' [22] design evaluation framework, we were able to address the challenge of long-term use through careful listening to end user experiences. Two different experiences are evidenced by 2 quotes (Textbox 2), representative of shared experiences that support or criticize the application, paving the way for a potential new iteration of the technological solution's design. This case study illustrates a limited yet valuable investigation of lived experiences, the results of which can improve digital user experience autonomy [22].

The second practice involves making the design phase participatory from the outset. For the creation of a prototype app for the prevention and self-management of pressure ulcers in patients with spinal cord injuries, we shared all decisions regarding content and functionality design with the various stakeholders involved: from health care professionals to individuals with spinal cord injuries. A comparison with the lived experiences of health care professionals allowed us to select scientific evidence where there was ambiguity or when the complexity of cases made it necessary [36]. Meanwhile, in ideation workshops, people with spinal cord injuries shared their significant experiences and the needs that emerged regarding the prevention of pressure ulcers [37]. This enabled researchers and developers to gain a deeper understanding of their reality, addressing challenges we have mentioned such as population literacy and the coordination of various sources. Before reaching the usability test phase, the development process was richly informed by multidisciplinary participatory methods (Figure 1) [37,38], from user experience design to health communication. These methods leveraged the collection of lived experiences to guide the decision-making process throughout all stages of the design, impacting personalization, information and content, navigation, and visualization [22].

Adequately preparing a complex intervention for its implementation requires several research steps [39] that in our experience have taken more than 3 years to complete. This first involves adapting the content and delivery methods to the target population and culture, and second, testing its usability and effects to implement the program in the real-world context. This solution is *iSupport for dementia*, a support program for caregivers of people with dementia developed by a group of experts for the WHO in a generic version, which my colleagues and I have recently adapted, developed, and tested for the Swiss context in 3 national languages and in the form of a mobile app, desktop version, and paper manual. In this project, our goal was to advance from a less deep level of participation, such as that defined as "informing," toward a more advanced form of collaboration and empowerment [25]. To achieve this objective, in addition to using mixed methods to include diverse experiences, we created a community advisory board [40], thus establishing a partnership that lasted throughout the project's duration (Figure 2) [38,41]. Community advisory board members included caregivers, associations and government representatives, researchers, and technology developers. They were actively involved as industry experts throughout the entire research process, providing constant feedback on iSupport content and research strategies. This included participation in annual meetings, preliminary content review, support in selection, and active participation throughout the entire study. Thanks to this collaboration, which followed the entire project's development, in accordance with the framework proposed by Wang et al [22], we were able to address challenges such as cultural and linguistic diversity and facilitate long-term adoption, developing a solution capable of meeting the needs of the involved audience. The multistakeholder participatory design served the purpose of refining both the design and the evaluation of the digital health solution. This integrated participatory approach has currently become a distinctive element of our public health research projects.

Textbox 2. Quotes from participants in the experiment of Morbiquiz.

Quote 1:

- "Whether videos were present or not, it wouldn't have made any difference. Because you could see this mother speaking, recounting her experience, but.... But I think that if there had been more videos with, let's say, different opinions, from different mothers, it might have been more...more instructive, more of an overall picture...." (11225, both interviews [34])

Quote 2:

- "It's a call to play, it's a call to action. It's so interesting to me, when you open the first question, I mean, now we have so many tools to navigate online and find the right answer, right? In fact, it invites you...to understand, read, analyze, correct? Then you give your answer. If it's correct, great. You're happy that what you saw was correct and you deepen your knowledge with the answer you receive. If it's wrong, then you start to question the source you consulted, right? This challenge needs to be emphasized. It means getting involved, seeking information, and ultimately becoming active." (11051, both interviews [34])

XSL•FO
RenderX

Figure 1. Co-design and development approach (reproduced from Amann et al [37], which is published under Creative Commons Attribution 4.0 International License [38]). SCI: spinal cord injury.

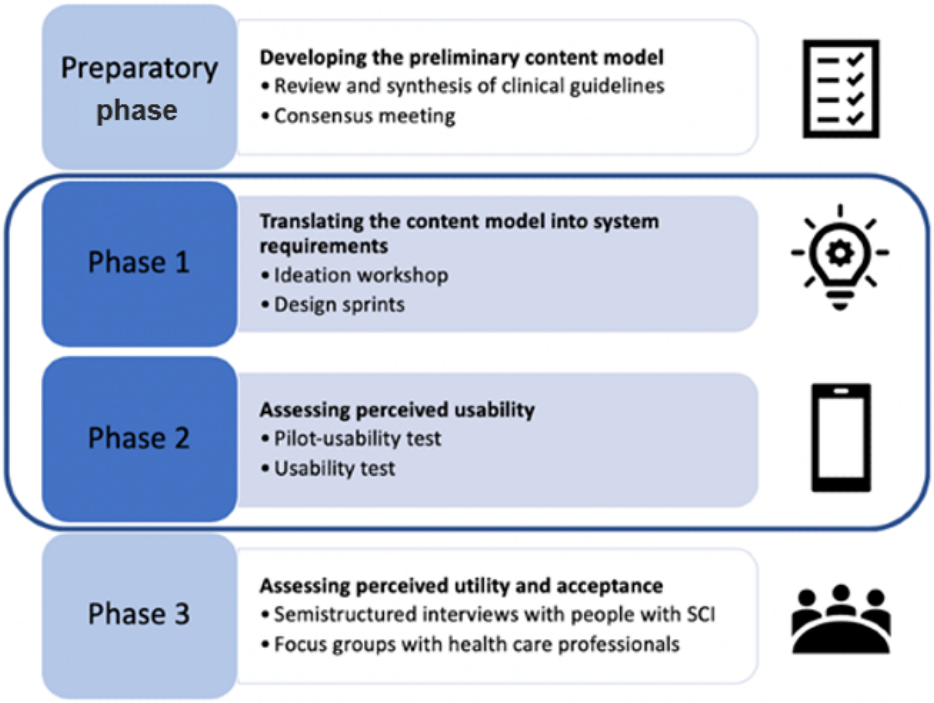

Figure 2. Flowchart of the adaptation process of iSupport in Switzerland (reproduced from Messina et al [40], which is published under Creative Commons Attribution 4.0 International License [38]). CAB: community advisory board; WHO: World Health Organization.

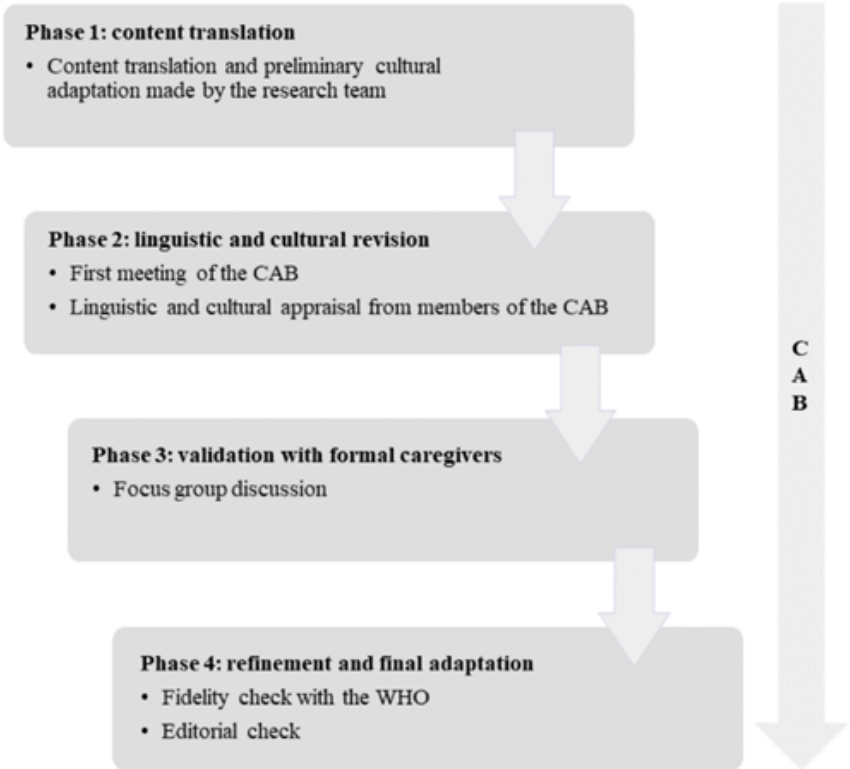

This contribution underscores the importance of investigating the lived experiences of all those involved into digital health solutions development and use as an effective tool for communication and persuasion in their design, development, and implementation. Favoring the listening of the different experiences emerges as a powerful means to establish empathy among the various stakeholders involved in the process. It highlights the need to value experiential knowledge to address and overcome common challenges in digital health, thus helping to fulfill the promises initially made. Furthermore, it emphasizes that this approach not only enhances social connections but also improves interactions in the context of developing solutions for digital health. The focus is on the importance of working on digital health with the many people involved, rather than simply

XSL•FO
RenderX

working for them. This promotes a more participatory and user-centered approach, aiming to enhance collaboration and more effectively meet the needs of those involved in the development and use of digital health solutions. As this is accomplished, conducting further research to evaluate the effectiveness of integrating phenomenology into digital health solutions through participatory methods would significantly enhance the existing body of knowledge in the field.

Conclusions

Transitioning perspectives in digital health through phenomenology integration represents a critical step toward realizing the full potential of digital technologies in health care. By centering the development and evaluation of digital health solutions on human experiences, we can overcome existing challenges and fulfill the promise of digital health to improve care outcomes and user engagement. Thanks to the sensitivity borrowed from the health communication perspective, this position paper calls for a collaborative effort among researchers, practitioners, and stakeholders to embrace phenomenological methods in digital health, ensuring that technological advancements enhance, rather than complicate, the health care experience.

Acknowledgments

I sincerely thank all those who have shared their experiences to realize the projects I have shown, and the colleagues who embraced these participatory approaches. Without you, this paper could not be conceived.

Conflicts of Interest

None declared.

References

1. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
2. Lovett JE, Bashshur RL. Telemedicine in the USA. Telecommunications Policy 1979;3(1):3-14. [doi: 10.1016/0308-5961(79)90019-3]
3. World Health Organization. Digital health EURO. URL: https://www.who.int/europe/health-topics/digital-health [accessed 2024-05-24]
4. Cohen AB, Dorsey ER, Mathews SC, Bates DW, Safavi K. A digital health industry cohort across the health continuum. NPJ Digit Med 2020;3:68 [FREE Full text] [doi: 10.1038/s41746-020-0276-9] [Medline: 32411829]
5. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
6. Robinson A, Oksuz U, Slight R, Slight S, Husband A. Digital and mobile technologies to promote physical health behavior change and provide psychological support for patients undergoing elective surgery: meta-ethnography and systematic review. JMIR Mhealth Uhealth 2020;8(12):e19237 [FREE Full text] [doi: 10.2196/19237] [Medline: 33258787]
7. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: a decade of evolution. J Med Internet Res 2013;15(5):e95 [FREE Full text] [doi: 10.2196/jmir.2430] [Medline: 23697600]
8. Fiordelli M, Diviani N. Granting access to information is not enough: towards an integrated concept of health information acquisition. In: Wac K, Wulfovich S, editors. Quantifying Quality of Life: Incorporating Daily Life Into Medicine. Cham, Switzerland: Springer International Publishing; 2022:501-510.
9. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022;24(7):e35816 [FREE Full text] [doi: 10.2196/35816] [Medline: 35793141]
10. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024;24(1):506 [FREE Full text] [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
11. Hu Y. Health communication research in the digital age: a systematic review. J Commun Healthc 2015;8(4):260-288. [doi: 10.1080/17538068.2015.1107308]
12. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2:38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
13. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024;26:e54375 [FREE Full text] [doi: 10.2196/54375] [Medline: 38787601]
14. Lampariello R, Ancellin-Panzani S. Mastering stakeholders' engagement to reach national scale, sustainability and wide adoption of digital health initiatives: lessons learnt from Burkina Faso. Fam Med Community Health 2021;9(3):e000959 [FREE Full text] [doi: 10.1136/fmch-2021-000959] [Medline: 34144970]
15. Broekhuis M, Weering MD, Schuit C, Schürz S, van Velsen L. Designing a stakeholder-inclusive service model for an eHealth service to support older adults in an active and social life. BMC Health Serv Res 2021;21(1):654 [FREE Full text] [doi: 10.1186/s12913-021-06597-9] [Medline: 34225745]

XSL•FO
RenderX

16. Whitelaw S, Pellegrini DM, Mamas MA, Cowie M, van Spall HGC. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
17. Kreps GL. Achieving the promise of digital health information systems. J Public Health Res 2014;3(3):471 [FREE Full text] [doi: 10.4081/jphr.2014.471] [Medline: 25553317]
18. Garzotto F, Paolini P, Schwabe D. HDM—a model-based approach to hypertext application design. ACM Trans Inf Syst 1993;11(1):1-26. [doi: 10.1145/151480.151483]
19. Kang X, Kang J, Chen W. Conceptualization and research progress on web-based product co-design. Informatics 2020;7(3):30. [doi: 10.3390/informatics7030030]
20. Nguyen AM, Rivera AM, Gualtieri L. A new health care paradigm: the power of digital health and e-patients. Mayo Clin Proc Digit Health 2023;1(3):203-209. [doi: 10.1016/j.mcpdig.2023.04.005]
21. Papoutsi C, Wherton J, Shaw S, Morrison C, Greenhalgh T. Putting the social back into sociotechnical: case studies of co-design in digital health. J Am Med Inform Assoc 2021;28(2):284-293 [FREE Full text] [doi: 10.1093/jamia/ocaa197] [Medline: 33043359]
22. Wang T, Giunti G, Melles M, Goossens R. Digital patient experience: umbrella systematic review. J Med Internet Res 2022;24(8):e37952 [FREE Full text] [doi: 10.2196/37952] [Medline: 35925651]
23. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019;8(2):90-97 [FREE Full text] [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
24. Tassone BG. The relevance of Husserl's phenomenological exploration of interiority to contemporary epistemology. Palgrave Commun 2017;3(1):11. [doi: 10.1057/palcomms.2017.66]
25. Vaughn LM, Jacquez F. Participatory research methods—choice points in the research process. J Participatory Res Methods 2020;1(1). [doi: 10.35844/001c.13244]
26. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
27. Messiha K, Chinapaw MJM, Ket HCFF, An Q, Anand-Kumar V, Longworth GR, et al. Systematic review of contemporary theories used for co-creation, co-design and co-production in public health. J Public Health (Oxf) 2023;45(3):723-737 [FREE Full text] [doi: 10.1093/pubmed/fdad046] [Medline: 37147918]
28. Burton SV, Valenta AL, Starren J, Abraham J, Nelson T, Kochendorfer K, et al. Examining perspectives on the adoption and use of computer-based patient-reported outcomes among clinicians and health professionals: a Q methodology study. J Am Med Inform Assoc 2022;29(3):443-452 [FREE Full text] [doi: 10.1093/jamia/ocab257] [Medline: 34871423]
29. Callens C, Verhoest K, Klijn EH, Nõmmik S, Pina V, Brogaard L. How Service Users Envision their Engagement in Processes of Collaborative Innovation: A Q-Methodological Study on User Involvement in eHealth Collaborations. Public Policy and Administration 2023:1-25 [FREE Full text] [doi: 10.1177/09520767231170298]
30. Tomasella F, Morgan HM. "Sometimes I don't have a pulse … and I'm still alive!" Interviews with healthcare professionals to explore their experiences of and views on population-based digital health technologies. Digit Health 2021;7:20552076211018366 [FREE Full text] [doi: 10.1177/20552076211018366] [Medline: 34104464]
31. Cummings M, Bradley J, Teal G. Patient co-design of digital health storytelling tools for multimorbidity: a phenomenological study. Health Expect 2022;25(6):3073-3084 [FREE Full text] [doi: 10.1111/hex.13614] [Medline: 36165104]
32. Liberati N. Emotions and Digital Technologies. HUMANA.MENTE Journal of Philosophical Studies 2019;12(36):292-309 [FREE Full text]
33. Finlay L. "Outing" the researcher: the provenance, process, and practice of reflexivity. Qual Health Res 2002;12(4):531-545. [doi: 10.1177/104973202129120052] [Medline: 11939252]
34. Fadda M, Galimberti E, Fiordelli M, Schulz PJ. Evaluation of a mobile phone-based intervention to increase parents' knowledge about the measles-mumps-rubella vaccination and their psychological empowerment: mixed-method approach. JMIR Mhealth Uhealth 2018;6(3):e59 [FREE Full text] [doi: 10.2196/mhealth.8263] [Medline: 29514772]
35. Fadda M, Galimberti E, Fiordelli M, Romanò L, Zanetti A, Schulz PJ. Effectiveness of a smartphone app to increase parents' knowledge and empowerment in the MMR vaccination decision: a randomized controlled trial. Hum Vaccin Immunother 2017;13(11):2512-2521 [FREE Full text] [doi: 10.1080/21645515.2017.1360456] [Medline: 29125783]
36. Fiordelli M, Zanini C, Amann J, Scheel-Sailer A, Brach M, Stucki G, et al. Selecting evidence-based content for inclusion in self-management apps for pressure injuries in individuals with spinal cord injury: participatory design study. JMIR Mhealth Uhealth 2020;8(5):e15818 [FREE Full text] [doi: 10.2196/15818] [Medline: 32432559]
37. Amann J, Fiordelli M, Brach M, Bertschy S, Scheel-Sailer A, Rubinelli S. Co-designing a self-management app prototype to support people with spinal cord injury in the prevention of pressure injuries: mixed methods study. JMIR Mhealth Uhealth 2020;8(7):e18018 [FREE Full text] [doi: 10.2196/18018] [Medline: 32673241]
38. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-10-21]

XSL•FO
RenderX

39. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of medical research council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
40. Matthews AK, Anderson EE, Willis M, Castillo A, Choure W. A community engagement advisory board as a strategy to improve research engagement and build institutional capacity for community-engaged research. J Clin Transl Sci 2018;2(2):66-72 [FREE Full text] [doi: 10.1017/cts.2018.14] [Medline: 31660220]
41. Messina A, Amati R, Annoni AM, Bano B, Albanese E, Fiordelli M. Culturally adapting the world health organization digital intervention for family caregivers of people with dementia (iSupport): community-based participatory approach. JMIR Form Res 2024;8:e46941 [FREE Full text] [doi: 10.2196/46941] [Medline: 38265857]

Abbreviations

WHO: World Health Organization

Edited by G Eysenbach; submitted 29.05.24; peer-reviewed by N Liberati, U Sinha; comments to author 18.07.24; revised version received 08.08.24; accepted 31.08.24; published 23.10.24.

Please cite as:
Fiordelli M
Transitioning Perspectives in Digital Health Through Phenomenology Integration
J Med Internet Res 2024;26:e62691
URL: https://www.jmir.org/2024/1/e62691
doi:10.2196/62691
PMID:

©Maddalena Fiordelli. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The Value of Smartwatches in the Health Care Sector for Monitoring, Nudging, and Predicting: Viewpoint on 25 Years of Research

Charlotte Köhler[1], PhD; Alexander Bartschke[2], MSc; Daniel Fürstenau[3,4], PhD; Thorsten Schaaf[3], PhD; Eduardo Salgado-Baez[2,3,5], MD

[1]Department for Data Science & Decision Support, European University Viadrina, Frankfurt (Oder), Germany
[2]Core Unit Digital Medicine & Interoperability, Berlin Institute of Health @ Charité, Charité - Universitätsmedizin Berlin, Berlin, Germany
[3]Institute of Medical Informatics, Charité - Universitätsmedizin Berlin, Berlin, Germany
[4]School of Business & Economics, Freie Universität Berlin, Berlin, Germany
[5]Department of Anesthesiology & Intensive Care Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Eduardo Salgado-Baez, MD
Institute of Medical Informatics
Charité - Universitätsmedizin Berlin
Charitéplatz 1
Berlin, 10117
Germany
Phone: 49 30450631167
Email: eduardo.salgado@charite.de

Abstract

We propose a categorization of smartwatch use in the health care sector into 3 key functional domains: monitoring, nudging, and predicting. Monitoring involves using smartwatches within medical treatments to track health data, nudging pertains to individual use for health purposes outside a particular medical setting, and predicting involves using aggregated user data to train machine learning algorithms to predict health outcomes. Each domain offers unique contributions to health care, yet there is a lack of nuanced discussion in existing research. This paper not only provides an overview of recent technological advancements in consumer smartwatches but also explores the 3 domains in detail, culminating in a comprehensive summary that anticipates the future value and impact of smartwatches in health care. By dissecting the interconnected challenges and potentials, this paper aims to enhance the understanding and effective deployment of smartwatches in value-based health care.

(J Med Internet Res 2024;26:e58936) doi:10.2196/58936

KEYWORDS

consumer devices; smartwatches; value-based health care; monitoring; nudging; predicting; mobile phone

Introduction

The consumer smartwatch market has experienced significant growth during the past years. Initially perceived as lifestyle accessories, smartwatches have transformed into advanced health monitoring devices. Equipped with various sensors, smartwatches capture vital health metrics, such as heart rate, oxygen saturation, fitness levels, and sleep quality. Remarkably, studies suggest that between 76% and 90% of adults are willing to share wearable data with their physician [1-3]. This positions smartwatches as valuable tools for personal health management, providing an unprecedented opportunity for patients to generate and manage their health data actively.

Smartwatches are expected to have a transformative impact on health care, with studies forecasting global cost savings of approximately US $200 billion due to wearable technology [4]. Integrating smartwatch data into electronic health records (EHRs) is a growing trend, enhancing patient monitoring and personalized care. In the United States and Canada, efforts have been focused on incorporating biometric and behavioral patient-generated health data into EHRs [5,6]. These endeavors include developing application programming interfaces (APIs) to address interoperability challenges and to understand social determinants of health better. In Europe, initiatives to integrate smartwatch data also encompass patient portals and shared decision-making tools [5]. A significant development is

XSL•FO
RenderX

Germany's recent legislation, which allows patients to transfer health data collected by smartwatches directly into their health records [7].

Numerous studies explore the integration of smartwatches into health care [8-14]. Yet, many studies concentrate on particular facets or stand-alone applications of smartwatches, often lacking a comprehensive framework that encompasses the diverse challenges and opportunities these devices present. This paper aims to bridge this gap by categorizing and evaluating these challenges within various contexts. By adopting this structured approach, we seek to provide a holistic perspective on the nuances of using smartwatches in health care settings. The goal is to enhance the foundational understanding of these complexities, thereby enabling more strategic and informed decisions regarding the adoption of smartwatches in the pursuit of value-based health care solutions. Furthermore, we hope this framework will guide future studies in clearly identifying the health perspectives they address and in exploring the value and challenges within each specific domain.

We propose categorizing consumer smartwatches into 3 key functional domains: monitoring, nudging, and predicting. Monitoring involves tracking patient health data within medical treatments and providing valuable supplementary insights from data generated outside medical facilities. Nudging refers to individuals using smartwatches for personal health tracking outside of clinical settings, offering significant public health benefits due to self-motivated health tracking. Predicting involves leveraging data from multiple users to train machine learning algorithms for predicting health outcomes, representing a promising area.

Monitoring offers the promise of reaching remote patients and gathering detailed health insights to enhance medical treatment quality. Yet, the effectiveness of this approach hinges on the trust in data-sharing practices and the integrity of the data collected, areas where consumer smartwatches may fall short. In the nudging domain, smartwatches are seen as powerful tools for encouraging healthier behaviors across various demographics, thus playing a significant role in addressing widespread health issues and enhancing public health. However, the responsibility of leveraging smartwatches effectively in nudging scenarios rests on the individual. Predictive use of smartwatches is a nascent yet promising field. Previously, patient data were only collected after diagnosis or during physician visits. Now, smartwatches collect data from both healthy and unhealthy individuals, providing unprecedented insights into disease and health maintenance. However, this area also raises significant data privacy and ethical concerns.

In The Evolution of Smartwatches in Tracking Health section, we will provide a concise overview of the latest advancements in smartwatch technology. Subsequently, we will present the 3 distinct perspectives of smartwatch use, monitoring, nudging, and predicting, exploring each in detail. Finally, we will offer a comprehensive summary, outlining the anticipated value and impact of smartwatches in the health care sector moving forward.

The Evolution of Smartwatches in Tracking Health

In recent years, the smartwatch market has witnessed a significant surge in sales globally. From generating approximately US $12.67 billion in revenue in 2017, the industry's earnings soared to US $44.94 billion by 2023, with projections estimating a climb to US $62.46 billion by 2028 [15]. In 2022, Apple led the market as the top-selling manufacturer with their Apple Watch, followed by Garmin's popular models such as Forerunner and Venu, Fitbit with its Sense smartwatch (as well as together with Google on the Pixel Watch), and Samsung with its Galaxy Watches [15].

All these smartwatches are equipped with various sensors that gather data, which are then analyzed and transformed into health features through each manufacturer's proprietary algorithms. These sensors often include gyroscopes, magnetometers, barometers or altimeters, GPS, and photoplethysmograms. The study by Henriksen et al [16] provides an exhaustive examination of sensors integrated into smartwatches over time and across different manufacturers. This section will focus on the health features derived from these sensors (passive tracking) while acknowledging that smartwatches also allow users to enter self-reported data (active tracking), such as moods or menstrual cycles, enriching overall health estimation. Promised health features encompass *activity tracking* (eg, daily steps, running, and cycling), *heart rate monitoring* via scanning wrist blood flow, and *electrocardiogram (ECG)*—a newer addition that has received attention for its capacity to detect heart conditions [17] and has thus attracted considerable medical research interest [18,19]. *Fall detection* mechanisms identify abrupt positional changes using accelerometers and gyroscopes. Continuous measurement of blood oxygen and sleep tracking are interconnected features using reflectance photoplethysmography, where optical sensors analyze light rebounding from the wrist [20]. Although sleep tracking has long been available on smartwatches, it is the integration of oxygen level analysis that now enables a truly comprehensive evaluation of sleep quality. This feature is particularly advantageous for managing conditions such as sleep apnea, marking a significant enhancement in sleep health monitoring [21]. The latest sensor addition measures skin *temperature*, supporting passive *cycle and ovulation tracking*—marking a gender-specific health feature innovation.

Figure 1 illustrates the evolution and adoption of health features across the 4 leading smartwatch brands: Apple, Samsung, Fitbit, and Garmin. A darker shade of blue indicates the year a feature was introduced. Apple stands out as a pioneer, consistently adding new health features and maintaining the most comprehensive offerings. Samsung, initially slower to integrate health features, has rapidly expanded its capabilities since the 2018 launch of the Galaxy Watch, aligning more closely with Apple's offerings. Fitbit and Garmin, both early adopters of activity tracking, have been slower to incorporate newer health features such as fall detection and ECG. However, they have led in exercise-related innovations, with Fitbit introducing blood

XSL•FO
RenderX

oxygen measurement alongside Apple and Garmin excelling in sleep tracking and oxygen level monitoring.

For all the manufacturers mentioned earlier, a smartwatch alone is insufficient for gathering health insights. Users must also connect the watch to a smartphone, install an app, and accept the manufacturers' data and privacy terms to access detailed health data. This raises concerns about users' limited control over how their data are handled and used [22]. A detailed study on the privacy issues and potential security threats is provided by Cilliers [23]. While there are numerous health tracking features, it can be assumed that the primary goal of these manufacturers is commercial success, and marketing health improvements serves merely as a means to this end.

As manufacturers continue to enhance their products with more health features, the line between lifestyle accessories and essential health tools becomes increasingly blurred. However, if manufacturers wish to market a tracking feature for medical purposes, they must obtain regulatory approval to classify the function as a medical device [24]. This classification ensures that the features meet the necessary standards for medical use, which is crucial for both consumer safety and product credibility. However, obtaining approval from regulatory bodies such as the Food and Drug Administration in the United States and securing CE marking in Europe is a time-intensive process that can delay the introduction of new features to the market. For example, when Apple introduced the ECG feature, it initially received clearance only in the United States, leaving users in other countries without access at the outset [25]. Consequently, manufacturers try to sidestep these hurdles whenever possible by marketing their functions as fitness and wellness tools rather than for medical purposes, although these features have significant health relevance beyond mere fitness tracking. As of 2024, the development of smartwatches is focused on incorporating several new innovative health features. Functionalities for measuring blood glucose, monitoring blood pressure, and detecting alcohol levels are expected in the upcoming models [26]. These anticipated features highlight an ongoing trend of increasing smartwatches' relevance in health care.

There is also a growing interest in developing open-source smartwatches, which offer greater control over data privacy and security compared to commercial models and can be tailored for specific health conditions. For instance, Bianchini et al [27] discussed an open-source smartwatch designed for dermatitis management. In addition, smaller manufacturers such as PKvitality and Mindpax [28] have developed smartwatches specifically for continuous blood glucose monitoring and improving treatment outcomes for mental illnesses, respectively.

Figure 1. Overview of health tracking features introduced. ECG: electrocardiogram.

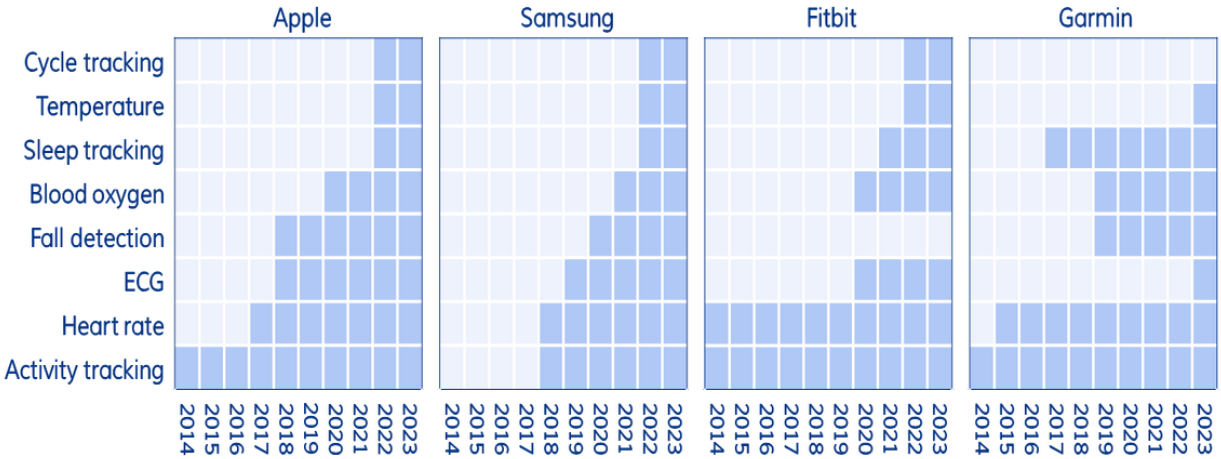

Differentiating the Value of Smartwatches in Health Care

Overview

Critical considerations for using smartwatches in health care include validity (correct measurements) and reliability (consistent measurements); data privacy, ownership, and security; device usability; reimbursement; and the technical integration of data into clinical workflows. Several studies illuminate aspects of these considerations. Validation and reliability studies such as those conducted by Fuller et al [29], Lauterbach et al [30], and Awolusi et al [31] evaluate the accuracy of consumer smartwatches in measuring health metrics. Concerns around data privacy and security are addressed in research by Ching and Singh [32] and Perez and Zeadally [33]. The usability and patient experience with smartwatches are explored in studies by Manini et al [11], Dehghani [34], and Pal et al [35] and the alignment with reimbursement models is explored in the study by Smuck et al [36], while the potential for integrating smartwatch data into treatment is discussed by Dunn et al [37] and Kheirkhahan et al [38]. Notably, the significance of each factor differs depending on its applications in health care. For example, while the precision of step counts may be less critical in nudging scenarios aimed at increasing daily activity, the accuracy of ECG measurements is paramount when used as a determinant for emergency medical care. Our differentiation in smartwatch use—monitoring, nudging, and predicting—provides a framework for evaluating which aspects are more or less crucial depending on the intended health care application.

In the Monitoring section, we explore the role of smartwatches in patient monitoring, highlighting its benefits and challenges

XSL•FO
RenderX

in health care. The Nudging section explores how smartwatches can influence health behaviors through nudging. The predictive capabilities of smartwatches, aimed at predicting health outcomes, are examined in the Predicting section. To illustrate these concepts, Figures 2, 3, and 4 offer visual representations of the three domains. The diagrams highlight the interactions between the individual and the smartwatch, the processes involved in collecting and storing health data, and the healthcare sector's involvement in leveraging this information.

Figure 2. Overview of smartwatch apps in health care (monitoring).

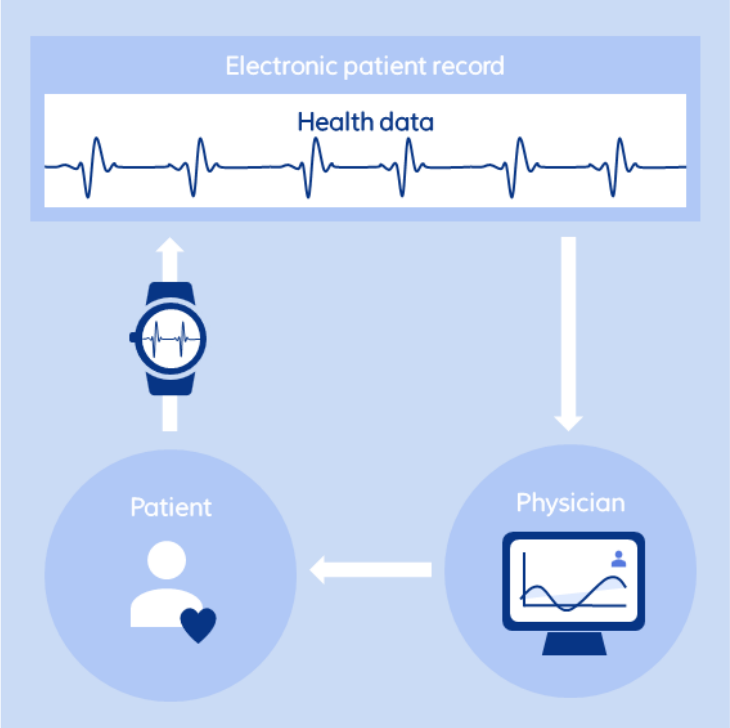

Figure 3. Overview of smartwatch apps in health care (nudging).

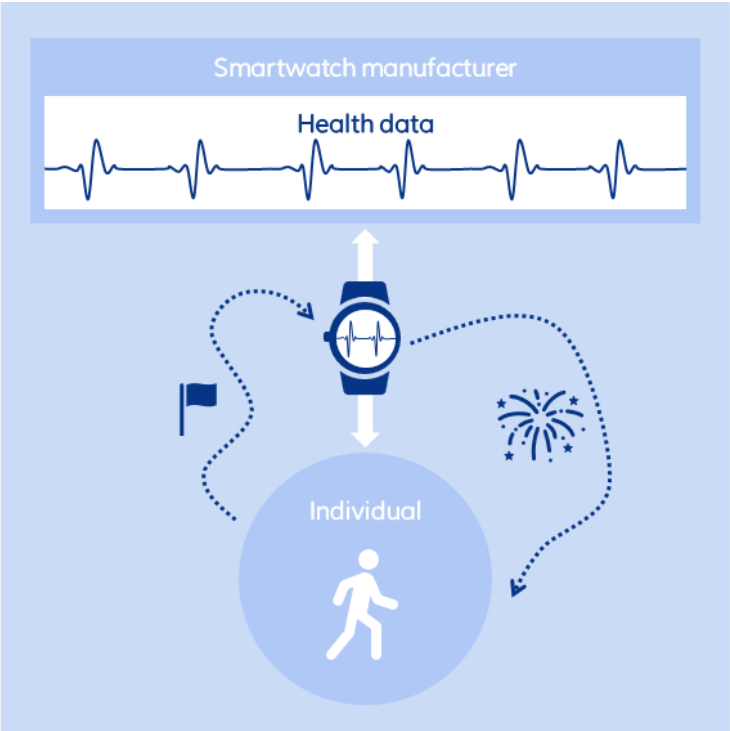

XSL•FO
RenderX

Figure 4. Overview of smartwatch apps in health care (predicting).

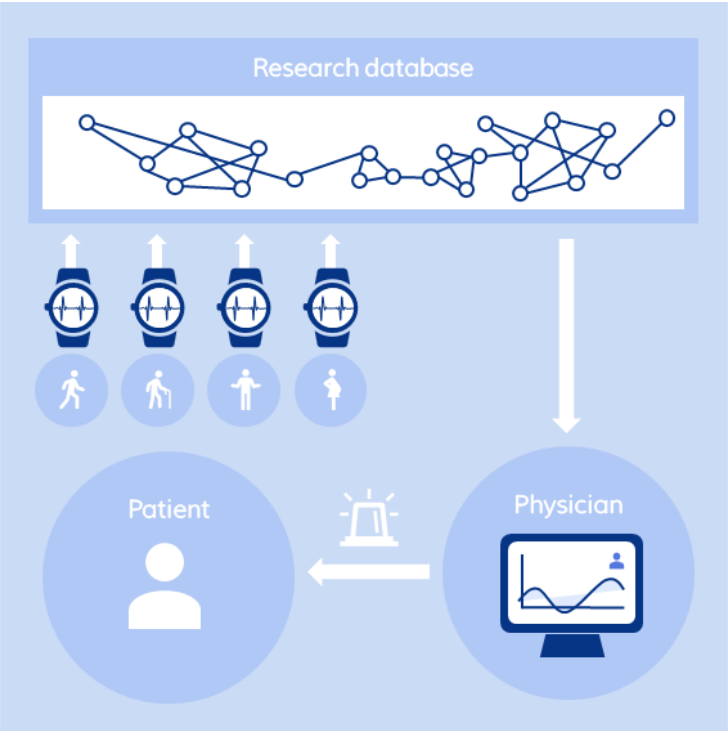

Monitoring

Overview

Smartwatches provide continuous patient monitoring at low costs and enable the gathering of patient-generated data despite physical separation from medical facilities under normal life conditions, providing health care providers with insights previously unattainable or only accessible through patient self-reporting. For instance, if a physician inquires about a patient's sleep over recent weeks, traditional answers would rely on subjective estimates. In contrast, a smartwatch can supply objective, unbiased data detailing both the quantity and quality of sleep, as well as its correlation with other vitals measured concomitantly. This significant advancement in the precision and reliability of health assessments is undeniable and represents a transformative shift in the delivery of care. In addition, the ability of smartwatches to automatically transmit vital sign assessment and movement to the EHR remotely can free up staff from manually collecting vital signs, allowing physicians and nurses to spend more time with hands-on care and avoiding patients' unnecessary visits to health institutions.

To implement effective monitoring of a patient's vital data, the process begins with establishing and discussing targeted health outcomes with the patient, ensuring alignment with their ongoing therapy. The technical capabilities of potential smartwatches can be discussed; if the patient already possess a smartwatch, its suitability for monitoring specific biometrics is checked. Subsequently, the appropriate sensor data from the smartwatch are selected to match these outcomes, and a precise data collection interval is set. Data are synchronized with a medical database, for instance, stored within the patient's EHR. Health care providers are then granted access to a dashboard offering in-depth data analyses, thus providing critical insights into the patient's health status. Regular follow-up consultations enable the health care team to monitor progress toward the established goals. Any deviations from the expected outcomes are explored jointly with the patient to adjust the treatment plan as needed. This cyclical and collaborative process champions a patient-centric approach, ensuring that the monitoring initiative is customized to suit the individual's unique health requirements and circumstances effectively (Figure 2).

In the subsequent paragraphs, we will initially explore the significance of smartwatches within the monitoring domain, both as a stand-alone and in comparison to other domains. This will be followed by examining the challenges and obstacles linked to using smartwatches for monitoring purposes.

Supervised Smartwatch Use to Provide Access to Health Data That Were Previously Unavailable, Costly, or Difficult to Obtain

Using smartwatches as monitoring tools is predominantly advantageous in 3 scenarios. First, smartwatches introduce the capability to gather health data that were previously either unavailable, costly, or difficult to obtain. Traditionally, continuous health monitoring has been confined to hospital stays or medical practices. Given the high costs, minimizing in-patient duration is a priority. Accessing patient parameters before or after hospital admission could significantly enhance care quality by making previously uncollected data available to physicians. An example can be found in the study by Randazzo et al [39], where patients were equipped with a custom-developed smartwatch that is designed to collect essential vital signs, including ECG, heart rate, blood oxygen levels, and temperature. The primary goal is to facilitate remote monitoring by hospitals, enabling earlier discharge of patients while still ensuring close observation of their biometrics. Aside from hospital stays, other monitoring methods, such as visiting a sleep laboratory for sleep apnea detection, can be inconvenient

XSL•FO
RenderX

and burdensome for patients. In their study, Chen et al [40] demonstrated that a Huawei smartwatch can distinguish between normal sleep patterns and sleep apnea events. Although the study acknowledged the smartwatch's limitation in providing more detailed classifications, it emphasized that even this initial level of differentiation can contribute to reducing the number of hospitalizations required. Fall detection for the older adults is among the most studied monitoring applications of smartwatches (refer to studies by Mauldin et al [41], Casilari and Oviedo-Jiménez [42], and Brew et al [43]). The severity of falls, particularly those occurring at night or when individuals are alone, is significantly increased due to the individuals' inability to seek help. This often results in prolonged periods of distress, as they may have to endure pain for an extended time before the incident is noticed. In addition to monitoring and fall detection, smartwatches can also assist in fall risk assessment and prevention [44] (refer to the Predicting section).

Second, smartwatches are revolutionizing health data accessibility, particularly for underserved populations and areas where traditional monitoring is challenging due to geography, economy, or limited digital health literacy. Research by Sheng et al [45] and Ibeneme et al [46] underscores these disparities. Enhanced by mobile software, smartwatches are crafting new pathways for care, reaching beyond typical consumer bases. For instance, the Apple Watch is favored by families with older members for its emergency services and health data–sharing capabilities, as noted by Tai et al [47] and Strauss et al [48]. Similarly, it is proving beneficial for children with conditions such as cardiovascular diseases or autism, offering precise monitoring that rivals traditional methods [49-51]. Furthermore, smartwatches are essential in settings such as extreme environments, remote workspaces, and professional sports, delivering critical insights into users' physiological states and safety [52-54]. This leap forward in monitoring technology is reshaping health care delivery, particularly for remote or rural communities, highlighting an urgent need for expanded research into how telemedicine and wearable technology can synergize to enhance care for all.

Third, monitoring solutions can be particularly beneficial when enhanced data on a patient's status could either improve care quality or alleviate the illness's burden. This is particularly crucial for individuals managing chronic conditions that significantly impact daily life. Chronic diseases such as diabetes or mental health issues exemplify areas where smartwatch monitoring can make a meaningful difference. Chakrabarti et al [55] highlighted the utility of smartwatches for patients with diabetes by monitoring symptoms such as frequent urination and sleep disturbances, which are challenging side effects of the condition. In mental health care [56], there is growing interest in using smartwatches to predict and manage psychotic disorder relapses and to monitor behavioral health. The use of smartwatches in the monitoring domain has a prevalent advantage compared to the other domains discussed later, which is the supervision of a health care professional when patients interact and use the smartwatch. Spender et al [57] claim that there are several reliability issues of the data collected with smartwatches that can be connected to wrong wearing behavior or choice of smartwatch, such as the specific device used, any changes to the device during the monitoring period (such as updates), the manner of device use, and how the device is worn. When a patient is under the supervision of a physician, potential errors in measurement and use can be more effectively monitored, and patients can be better educated on the proper use of the smartwatch. This supervision ensures that data collected for health monitoring are more accurate and reliable, mitigating issues related to device inconsistencies or incorrect use. Smartwatches reduce treatment costs across various applications; even in cases where the integration of these tools increases costs, the improvements in patient quality still deem these tools overall effective from a value-based health care perspective [58].

The digital capabilities of smartwatches enable health care providers to extend their reach to more patients, particularly in remote areas or specific patient groups. Consequently, this leads to fewer in-person visits and more effective use of health care resources. Leveraging smartwatches to monitor patients with chronic illnesses presents a cost-effective solution, potentially mitigating high treatment expenses. However, economic evaluations must be conducted carefully, tailored to each setting, and compared against other telemedicine alternatives. Nevertheless, the level of patient data detail gathered by these tools is unmatched. In addition, the introduction of smartwatches as monitoring tools in health care can significantly enhance the delivery of patient-centered care. These devices have the potential to empower patients to actively engage in decision-making processes about their health in accordance with their values and preferences. This increases their agency in managing different aspects of their health and illness, helping to reduce fear, and facilitating the involvement of family and friends when needed. Smartwatches, being a low-cost tool and often subsidized by health insurance companies, can contribute to providing more equitable access to personal health management allowing health care professionals to observe, prevent, or respond to health issues based on data generated in everyday life outside medical facilities, enhancing access to care and reducing patient vulnerability.

Smartwatch Use in Treatment: Trust in the Treatment Process, Data Sharing, and Data Quality

The quality of vital parameters collected is crucial for effective health monitoring. Generally, smartwatches receive positive validation for their measurement accuracy. For instance, recent studies such as the study conducted by Zarak et al [59] report that ECG measurement accuracy in consumer smartwatches ranges from 65% to 99% and the study by Jiang et al [60] indicates that the accuracy of saturation of peripheral oxygen (SpO_2) measurements varies between 90% and 96%. However, the real-world applicability of these devices in health settings is questionable, as most validation tests are conducted in laboratory settings [29]. For example, step counts can become inaccurate for patients using gait aids [61], and there is diminished performance in smartwatches for individuals with darker skin tones [62]. This may lead to patients' reluctance to trust these devices and physicians' reservations about recommending smartwatch use.

XSL•FO
RenderX

Another crucial aspect of fostering trust in new technology is considering the data privacy of collected health data. Patients must be informed that smartwatches gather highly personal health information, which, despite being treated with the utmost sensitivity, poses privacy concerns. Therefore, it is essential for patients, in collaboration with their physicians, to conduct a thorough cost-benefit analysis when considering the use of smartwatches for continuous health monitoring. This process is akin to evaluating standard treatments, such as medications, where the potential side effects are balanced against the anticipated health benefits in consultation with health care providers. The issue of data privacy does not belong to a specific medical domain; however, it creates uncertainty over who bears the responsibility of educating patients on this vital matter. This underscores the need for clear guidelines and educational efforts to ensure patients can make informed decisions about using smartwatches in their health care regimen [63].

Nudging

Overview

With nudging, we refer to the process where individuals analyze their own data to actively improve their health. Nudging involves subtly guiding individuals toward behavior change, with the aim of fostering long-term habit modifications that result in improved health outcomes [64]. It entails using psychological strategies rather than direct medical interventions to influence user behavior as shown in the study by Marchiori et al [65]. Smartwatches play a significant role in this process by tracking fitness parameters and providing motivational cues, enhancing individuals' physical and psychological self-awareness. This heightened awareness can lead to greater recognition of health-promoting behaviors, such as improved nutrition habits, even if the smartwatch's sensor technology may not directly measure these behaviors [66]. Unlike in the monitoring settings, no direct supervision or contextual integration in the health care setting is typically associated with nudging using smartwatches. However, incentives from health care providers, such as health insurers or physicians, may exist to encourage the adoption of smartwatches (Figure 3).

Smartwatch providers strategically incorporate health-related benefits and outcomes into their marketing efforts and design decisions. For instance, the smartwatch interface may include visual rewards to incentivize users to achieve activity goals, adhere to sleep and nutritional schedules, and engage in mindfulness exercises. Sangameswaran [67] showed that smartwatches, being worn close to the body and serving as a personal device, offer effortless access to information and use subtle methods, such as haptics, vibrations, and light, to capture the wearer's attention. These features make smartwatches a powerful tool for driving behavior changes throughout the day. For example, the Apple Watch uses a "firework" visualization as a reward when users successfully reach their daily goals for steps, exercise, and standing. Similarly, Garmin watches provide users with information about their stress levels, emphasizing the significance of mental awareness and well-being. Smartwatches can guide users through breathing exercises, enabling them to incorporate meditation practices into their daily lives. An overview of related digital nudges can be found in the study by Bergram et al [68].

In the following sections, we consider the benefits and challenges in the nudging domain.

Effect of Smartwatches on Healthy Behaviors Across a Broad Demographic

Smartwatches, with their low barrier to entry, are well-positioned to impact a broad demographic, particularly in public health domains. They hold significant potential in managing lifestyle-related diseases—commonly associated with high stress, inadequate sleep, and insufficient physical activity that are prevalent across diverse populations.

First, smartwatches serve as educational tools promoting a healthier lifestyle. While a trusted physician-patient interaction is invaluable for lifestyle guidance, smartwatches offer similar, continuous, and motivational support. Conditions such as hypertension and obesity, widespread in Western societies, can be alleviated through increased physical activity, an area where smartwatches have shown efficacy. Research indicates that smartwatch ownership often leads to a significant uptick in physical activity. For instance, Ferguson et al [69] documented an increase of up to 40 minutes in daily walking time, while Li et al [70] emphasized smartwatches' positive effect on user activity levels. Smartwatches facilitate personal goal setting and progress tracking, engaging users in their health journey. This self-driven approach empowers individuals to take charge of their health, fostering a sense of achievement and personal responsibility.

Second, smartwatches provide nudging for individuals who lack external motivation, such as the older adults, who may experience loneliness and thus have diminished social incentives to remain active. In situations without external encouragement, smartwatches adopt a supportive role, offering controlled encouragement. Studies such as the study by Chung et al [71] show that older participants feel motivated by their smartwatches to be more active, illustrating the devices' capability to fulfill a motivational void. Boateng et al [72] developed a smartwatch step-counting app specifically tailored for the older adults to engage them in physical activity. In addition, the study by Rosales et al [73] revealed the intrinsic motivation among the older adults to use smartwatches, associating the device with feeling "cool" and youthful.

Similar to the monitoring domain, significant potential cost savings are anticipated in both the monitoring and nudging scenarios. However, accurately quantifying the financial benefits of health nudging via smartwatches presents challenges, mainly due to the long-term nature of these effects and their interaction with other health initiatives. Although measuring the direct health outcomes of nudging can be complex, the broader impact on public health is anticipated to be considerable. This perspective is supported by actions from major German health insurers who offer reimbursements for smartwatches, typically cautious in supporting preventive measures, thereby acknowledging their utility in health interventions. On average, these insurers provide a subsidy of approximately 300 Euros (US $328) toward the purchase of smartwatches, demonstrating

XSL•FO
RenderX

a recognition of their value in enhancing health care outcomes and saving health-related costs [74].

Distribution of Health Benefits of Smartwatches and the Onus of Using a Smartwatch for Healthy Behavior

In addressing the challenges of smartwatch use for health-related nudging, maintaining user engagement is paramount. The role of usability, which includes technical quality, design, and affordability, is critical in this regard. As Chuah et al [75] suggest, user motivation to achieve personal goals greatly influences continued smartwatch use. However, there is a mismatch between the demographics of typical smartwatch users and those who would benefit most from their health applications. For instance, the user group with the highest use is between 25 to 34 years, a demographic at lower risk of lifestyle-related diseases. In contrast, for older individuals, who could greatly benefit from such technology, smartwatch use is considerably lower. This use pattern highlights the need for targeted strategies to increase smartwatch adoption among older populations, despite disparities in digital literacy, who stand to gain significantly in terms of health monitoring and management [76]. Similarly, the study by Helsen et al [77] highlights a pronounced connection between runners and smartwatch use. This implies that individuals already engaged in healthy activities such as running are more inclined to use smartwatches. Therefore, these devices may be predominantly adopted by those who have already committed to a health-conscious lifestyle. While owning a smartwatch can facilitate increased activity levels, its effectiveness hinges on the individual's preexisting motivation to adopt life-changing health behaviors. Constantiou et al [78] highlight the individualized use and impact of smartwatches on health tracking. Users' reactions to their performance data can vary greatly; unsatisfactory feedback might lead to discouragement and reduced use, whereas those receiving positive feedback are likely to find smartwatches particularly beneficial, reinforcing their already positive health behaviors. This distinction underscores the personalized nature of smartwatch utility and its varied effects on individual health tracking and motivation efforts.

However, not only the equitable promotion, design, and adoption among different user groups but also the technical fairness of smartwatches raise important considerations. It has been observed that the performance of smartwatch sensors can vary significantly across various skin tones, and their accuracy tends to decrease in individuals with higher BMI [79]. This variability in sensor performance not only raises questions about the inclusivity and fairness of these devices but also about their effectiveness and reliability in health monitoring across diverse populations. This issue underscores the need for more inclusive design and testing processes in the development of smartwatch technology to ensure that it serves a broad demographic, regardless of their physical characteristics.

Another factor that should not be overlooked is that smartwatches, though expected and discussed to bring major health improvements, can also be harmful and decline the health of people. This is particularly discussed for mental health, that is, patients experiencing anxiety, for which the permanent monitoring of biometrics can have a negative effect [80]. In the context of nudging, where users interact solely with the smartwatch without direct supervision from health care providers, there exists a risk associated with the use, interpretation, and implementation of smartwatch recommendations. These risks could potentially lead to harmful behaviors or increase expectations, pressure, or stress. Without professional guidance, users might misinterpret health data or recommendations, leading to decisions that could adversely affect their health rather than improve it.

Another significant challenge is the requirement for users to accept privacy and data processing terms set by manufacturers, often after the purchase. This situation can leave users susceptible, having no option but to agree to these terms to use the device. It places the onus of data management and interpretation on the users themselves, without a collective regulatory body overseeing manufacturer practices in health applications. The widespread use of smartwatches for health purposes thus imposes a responsibility on providers to guard against manipulative or harmful nudging practices [81]. This situation calls for stringent measures in data ownership, privacy, and transparency about the nudging mechanisms used, particularly because manufacturers' primary focus is on commercial success rather than patient health.

Predicting

Overview

In the domain of prediction, we focus on detecting potential health risks and predicting future health outcomes for individual patients [82]. We identify 2 main directions that strengthen the use of predictive methods with smartwatch data. First, the ease of continuous data collection from patients enriches the understanding of their health statuses. Using these data can improve predictions of treatment outcomes, drawing from patient's own historical data or that of others with similar conditions. This approach is closely aligned with the monitoring domain, as it leverages patient-monitored data to predict risks and outcomes. Second, health data collection has traditionally been concentrated around visits to health care providers, primarily capturing data from those actively receiving treatment. This method leaves a significant gap as healthy individuals who do not need to visit physicians remain underrepresented in health datasets. Relying exclusively on data from individuals with a disease can result in biased insights; including data from healthy individuals provides a more balanced view, helping to understand why certain people remain healthy while others develop illnesses.

The outlined procedure in the prediction domain starts by consolidating data from individuals, collected with the smartwatches, into a medical research database. As an initial step, the desired outcome must be defined; this could be a classification task (eg, "cancer" vs "healthy" as in the study by Hou et al [83]) or a regression task (like calculating a risk score as in the study by Goldstein et al [84]). Given that smartwatches collect time-series data, the outcome of interest could also be a change in the observations over a period or a time output itself (eg, predicting relapses as in the study by Moshe et al [85]). Once the outcome has been formulated, the appropriate prediction model must be selected. In this context, supervised

XSL•FO
RenderX

machine learning, a major subset of artificial intelligence (AI) algorithms, is used extensively for making predictions. These algorithms are trained on one segment of the data and tested on another, ensuring they can effectively identify patterns and make accurate predictions based on new, unseen data. However, there is often a trade-off between model performance and explainability. For example, decision trees provide a clear set of decision rules, which makes the decision-making process understandable and increases the trust in results; however, they typically only perform good prediction results for simpler datasets. In contrast, deep learning algorithms, which use billions of decision nodes, often lack transparency in explaining outcomes but can yield superior outcomes even for complex prediction scenarios.

When developing a predictive model, evaluation metrics are crucial for determining which model to apply. One important metric is accuracy, which indicates the overall correctness of predictions. In medical predictions, however, there is often a higher interest in not just the general prediction accuracy but also the accuracy of specific predicted classes and potential mistakes made by the model. For example, consequences of incorrect predictions vary significantly (eg, mistakenly telling a patient they are healthy when they need treatment is generally worse than the reverse). Therefore, sensitivity and specificity, precision, and recall, which measure how well positive and negative classes are predicted, are commonly considered as well.

Insights derived from these analyses are then reviewed by health care professionals or embedded in clinical decision support systems (CDSSs), facilitating improved patient health outcomes (Figure 4). The study by Steyerberg and Vergouwe [86] outlines the requirements for developing prediction models in a clinical environment. Many of these steps align with established data science principles, and it is particularly noteworthy that the authors evaluate specific metrics, such as the model's suitability and usefulness for a particular CDSS.

We will explore the most advantageous scenarios for using smartwatches in predictive applications, followed by an examination of the related challenges.

Using Predictive Methods for Major Advances in Health, Particularly in Preventive Care and Risk Estimation

Predictive analytics can significantly enhance the quality of care for patients in treatment management by using smartwatch data for informed decision-making. For instance, Stojancic et al [87] used smartwatch data to predict pain scores in patients with sickle cell disease, thereby facilitating more effective pain management. Similarly, Liu et al [88] applied predictive analytics to anticipate mortality events in palliative care using smartwatch data, easing the emotional burden for patients and their families by providing some certainty in a challenging situation. In addition, smartwatches are gaining recognition for managing mental health conditions, particularly those with high-relapse risks such as schizophrenia, as explored in the study by Fonseka and Woo [89]. These examples illustrate that for patients already undergoing treatment and monitored via smartwatch, the application of predictive algorithms can further enhance the quality of care. In addition, Brasil et al [90] highlighted that applying AI to analyze large datasets can address key challenges in rare disease management with limited patient populations and geographical dispersion. Brasil et al [91] provided a comprehensive review of machine learning algorithms tailored to rare disease studies.

The most intriguing and still uncertain aspect of predictive analytics is its potential to predict conditions in individuals not yet undergoing treatment, thereby playing a significant preventive role. This potential unfolds in 2 key ways: first, by diagnosing diseases that are already present but undetected as well as identifying risks for diseases not yet manifest; and second, by analyzing data from individuals with a disease at an aggregated level to optimize health care resource assignment and capacity management.

For the first time, predictions allowing for a timely diagnosis could make treatment possible before an irreversible stage is reached. This might be particularly relevant for diseases that are asymptomatic in the early stages [92]. Many examples can be found for cardiovascular diseases. In their systematic review, Nazarian et al [93] showed that based on smartwatch data cardiac arrhythmias can be predicted with an average accuracy of 97%. A comprehensive review of both custom-built and commercial smartwatches for detecting cardiovascular disease risk is detailed in the study by Moshawrab et al [94], which highlights a rapid increase in research over the past 3 years, pointing to a promising future for deeper insights in this area. Furthermore, Mizuno et al [95] emphasize the importance of using these devices for cardiovascular disease risk prediction. The authors suggest that in cases of heightened risk, continuous monitoring combined with lifestyle changes, akin to the nudging scenario presented earlier, can be particularly effective. An overview of additional diagnostic applications of wearables, including in psychological and neurological diseases, liver diseases, metabolic diseases, and sleep disorders, is provided in the study by Chakrabarti et al [55]. The most notable example of using smartwatch data at an aggregated level for resource management occurred during the COVID-19 pandemic. For instance, Föll et al [96] demonstrated how smartwatches could be used as risk-scoring tools in general wards for patients with COVID-19, optimizing resource allocation to those most in need. Similarly, Esmaeilpour et al [97] developed an alert system using smartwatches to detect respiratory viral infections and manage available capacities more effectively. Similar ideas can be found in the studies by Zhu et al [98], Mishra et al [99], Kallel et al [100], and Abbasi [101]. A notable initiative in Germany [102] solicited smartwatch data from the public, leveraging it to predict and identify COVID-19 outbreaks nationwide without direct testing or patient contact. This project highlighted the smartwatch's significant role not only in detecting outbreaks but also as a strategic tool for policy makers in pandemic management. Despite its success, the project concluded in 2022, and a new platform for data donation has not yet been established. To effectively harness the value of these data and implement predictive measures in health care, regulatory approval and both public and private reimbursement are essential, not only for COVID-19 but also for future public health crises [103].

XSL•FO
RenderX

The Value of Predictive Smartwatch Use and Its Implementation Into Feasible Solutions

After having discussed how leveraging smartwatch data for predictive analytics can aid not only in treating diseases but also in preventing them. However, to date, we do not see feasible large-scale implementation of collecting and using smartwatch data for predictions in health care.

A major challenge in using smartwatch data for predictive analytics is the privacy and ethics of data collection. We have already highlighted the need for patients to balance the benefits of sharing their smartwatch data against potential risks in the Monitoring section. In addition, in this section, we have discussed the importance of collecting data not only from patients but also from healthy individuals. For the latter, the direct benefits of sharing their personal data are not always clear, raising questions about whether the risks of data sharing are justified. Informed user consent in sharing their personal data for predictive purposes is critical, highlighting a significant responsibility and a gap in effectively educating the public about data sharing [63]. Both digital literacy and, more specifically, data literacy [104], are crucial for understanding the implications of data sharing. Gupta et al [105] observed that older and more conservative participants were less inclined to share their data, whereas Seifert and Vandelanotte [106] found that those comfortable with technology and higher incomes were more willing. Moreover, Gupta et al [105] emphasized that the willingness to share data extends beyond simply signing a consent form; factors such as data deletion policies, regulatory oversight, and transparency are also crucial for individuals to agree to share their data. With rapid advancements in AI, it becomes challenging for individuals to agree to data sharing without knowing the potential future uses of their data [63]. Data privacy measures necessitate a focus on data minimization and anonymization. However, it is essential to recognize that excessive anonymization can strip away valuable contextual information vital for making accurate predictions, as demonstrated by Hicks et al [107], who showed that activity and health patterns can vary significantly across different demographics.

Given these substantial challenges, data donation models could provide a viable solution by collecting data only from individuals who voluntarily choose to participate, significantly increasing the likelihood that they understand the implications of their decision. Gomez Ortega et al [104] proposed a framework for establishing a data donation platform, detailing strategies for setting up the platform and engaging users to continuously evaluate and reconsider their participation, thereby enabling ethical data sharing. The initiative undertaken by the Robert Koch Institute during the COVID-19 pandemic is a prime example, where smartwatch data donations were solicited to track potential outbreaks, demonstrating this method's ability to reach a broader demographic than traditional research methods [102]. However, Strotbaum et al [108] argued that the successful establishment of such a platform raises long-term issues that have yet to be resolved. These include data anonymization, primary and secondary uses of data, benefits for donors, and access to these datasets—challenges that persist from the perspectives of both researchers and policy makers. Solutions and regulations for these issues need to be developed in the future.

Discussion

In exploring smartwatches' impacts on health, we dissected their roles across monitoring, nudging, and prediction domains. This section consolidates those insights, providing a summarizing lens to juxtapose and compare challenges across these areas. We specifically focus on the predominant challenges highlighted in the Differentiating the Value of Smartwatches in Health Care section: the validity and reliability of measurements, data privacy and security, device usability, and technical integration.

Challenges related to the *validity and reliability* of data from smartwatches are significant, as these devices rely on algorithms that are often nontransparent and proprietary to their manufacturers. Conducting empirical studies to assess the accuracy of devices from leading brands could help provide some reassurance on using these in a medical setting [29,59,60]. However, due to the fast-paced release cycle of new models, by the time studies such as the study by Fuller et al [29] are published, the evaluated models are often no longer produced or available to consumers. As this strategy essentially follows a "test afterward" approach, a more proactive alternative could involve exerting stricter regulations on manufacturers to disclose their algorithms or establishing uniform testing standards that all manufacturers must meet before releasing new models. We also see potential in specifically engineered smartwatches for medical applications by smaller manufacturers. These devices usually adhere to more stringent regulations and offer greater transparency about their algorithms. However, they cannot compete with the range of functions offered by major brands such as Apple, Garmin, and Samsung. Particularly in the domains of nudging and predicting, where strong motivation to wear the device is necessary, we see no alternative but to rely on these established major smartwatch providers.

Data privacy and security pose substantial challenges in the context of sensitive health data collection by smartwatches. The burden of data protection often falls on patients, without a straightforward way to circumvent the storage of their data on manufacturers' servers. In both monitoring and prediction scenarios, where consumer smartwatches are used, the data are typically replicated to health care providers or research databases. Nonetheless, users are required to consent to the sharing of their data with manufacturers. Within the nudging domain, the data exclusively reside with the manufacturers, introducing significant risks. As outlined in the previous section, the value of data collected by smartwatches is immense. While the health care system is subject to stringent regulations for data storage, manufacturers may store data indefinitely, potentially including sensitive information, such as the frequency of falls, mental health issues, or heart attacks. While major manufacturers such as Apple, Samsung, and Garmin currently guarantee a safe environment for personal health data [109-111], past incidents of data breaches [112] show that there is always a risk. Furthermore, users may share their data with third-party apps on their smartphones that do not implement the same level

XSL•FO
RenderX

of security measures, distributing highly sensitive health data to many providers and increasing the risk of a breach at some point.

This scenario underscores the need for rigorous data protection measures and scrutiny over how health data collected by smartwatches is stored, used, and shared in commercial domains. However, with increasing medical relevance, manufacturers might be incentivized to provide alternatives. For instance, Garmin is exploring the use of localized data transmission hubs that store data solely on the user's device and allow direct transfer only into the EHR, thereby reducing the risk of unauthorized access to the data.

The *usability* of smartwatches impacts both the quantity and quality of data collected. According to the study by Ding et al [113], enhancing usability in medical settings could be achieved through simpler interfaces and extended battery life. While smartwatches designed as lifestyle accessories are advantageous, those crafted specifically as medical devices often lack the design and motivational features of mainstream smartwatches. Usability becomes crucial when motivation is necessary, or health benefits are not direct and long-term and hence not immediately apparent. Therefore, in nudging and predicting scenarios, encouraging users to wear their smartwatches consistently is vital. In contrast, in monitoring scenarios, where wearing a smartwatch is directly connected to personal health improvements, users may be more inclined to tolerate less user-friendly devices, provided they deliver the anticipated health outcomes. In considering the usability of smartwatch technology, it is crucial to highlight the issue of *access*. It is worth remembering that consumer smartwatches aim to maximize sales within their target demographic, yet this group does not always align with those who could most benefit from the health improvements offered by the technology. There is a notable use bias toward White individuals, adults aged 18 to 50 years, and those with higher education levels [114]. Despite the potential of smartwatches as a low-cost, accessible health care solution, their actual use across different demographics is uneven. Financial barriers are one obstacle which health care providers can address to some extent. However, educational efforts are also essential to integrate smartwatches into a broader range of individuals' health routines, ensuring more equitable access and utilization.

Integration stands out as a pivotal element in smartwatch applications across monitoring, nudging, and predicting. It is crucial to differentiate between device connectivity and data interoperability. For connectivity, smartwatches use various protocols such as Wi-Fi, Bluetooth, NFC, ZigBee, ANT+, Cellular, and USB to transfer data, chosen based on data volume, transmission frequency, and impact on battery life. Internet-enabled devices, often smartphones, serve as gateways for data transmission to caregivers, facilitating Cloud connections and communication. However, most smartwatches are compatible with a few specific platforms, such as Apple's watchOS, Google's Wear OS, and other proprietary systems [115]. Cross-platform interoperability faces hurdles; for instance, while Wear OS devices can pair with iOS, functionality is reduced, and features such as ECG or blood pressure monitoring may be unavailable. Apple's watchOS, designed exclusively for iOS integration, further limits system openness and compatibility with other devices and platforms. Health data exchange is tightly regulated, necessitating compliance with international standards for integration into EHR. Achieving semantic interoperability involves adopting universal terminologies such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Logical Observation Identifiers Names and Codes (LOINC), while syntactic interoperability requires standards such as Health Level 7 Fast Healthcare Interoperability Resources. Although there has been progress in health data standardization, aligning smartwatch data with these protocols remains a challenge, even with efforts such as the standards developed by the Institute of Electrical and Electronics Engineers or the guidelines provided by the Food and Drug Administration [116].

While our discussion has explored the different areas of smartwatch applications separately, we have yet to encounter a solution that satisfactorily addresses all 3 areas. Ideally, a single smartwatch that meets all these challenges and requirements would significantly enhance overall health and blur the lines between these categories. Until such a solution is available, it is crucial that users and patients are thoroughly informed about the benefits and risks associated with using wearables for health data tracking. This is particularly pertinent in the nudging domain, where individuals independently attempt to interpret health data using a device on their wrist. Nudging domain, a gap in health care knowledge exists, which future developments must consider. Bridging this gap is essential to ensure that individuals have the necessary understanding and tools to effectively use smartwatches in managing their health.

Conclusions

In this paper, we have explored the multifaceted value of consumer smartwatches within the health care sector, categorizing their utility into monitoring, nudging, and predicting domains and highlighting their potential in transforming health and illness management. Monitoring leverages smartwatches for gathering patients' data, including vital signs and activity data. Nudging pertains to the capability of smartwatches to influence users toward adopting healthier lifestyles, such as increasing physical activity, which can significantly enhance their long-term health. Predicting extends these applications by using vast amounts of data collected from numerous users to train AI models and implement CDSSs based on them. For each domain, we have outlined the associated benefits and challenges, presenting a comprehensive analysis of these elements.

While numerous challenges persist in integrating smartwatches into health care, our findings indicate that these challenges vary in intensity and risk across the different domains of monitoring, nudging, and predicting. Overall, we would like to emphasize the following aspects.

First, in the literature, smartwatches are often discussed in general terms regarding their potential and challenges, which can create the impression that the obstacles to their use are insurmountable. However, our contribution aims to demonstrate that challenges significantly vary across different domains. A more nuanced understanding of domain-specific challenges can

XSL•FO
RenderX

help stakeholders realize that not all issues are equally relevant, thereby facilitating the targeted implementation of smartwatches. We advocate for a clear distinction in the future use of smartwatches within medical settings, specifying the particular domain of application as we have suggested. This approach will ensure that the use of smartwatches is appropriately tailored to the relevant medical context. By adopting this perspective, smartwatches should not be viewed as a universal tool. While they may be versatile in certain cases, it is crucial to first consider the specific scenarios in which they will be integrated and used to maximize their effectiveness and relevance in the delivery of health care.

Second, we have demonstrated the immense potential value of smartwatches for health across various domains. However, despite numerous studies highlighting this value, these solutions are rarely implemented in individuals' daily lives. To achieve safe and beneficial integration into health care settings, it is crucial to develop practical strategies. This involves creating methods for individual-level integration, such as personalized health management apps that use smartwatch data, improving interoperability with existing health care systems, and training health care providers on the effective use of smartwatch data. In addition, exploring the broader potential of collecting extensive health data from the population with minimal effort, as demonstrated during the COVID-19 pandemic, could significantly enhance public health monitoring and response.

Third, data privacy is a critical concern when it comes to the use of smartwatches. While these devices can collect a variety of health metrics and serve as diagnostic tools, most consumer smartwatches lack the standards required for securely handling these private data. As a result, the health data collected by smartwatches are one of the most unprotected sources of health information, despite its detailed and sensitive nature. Currently, the responsibility for protecting these data is largely placed on individuals who use smartwatches. To address this issue, there is an urgent need for standardized data protection protocols, stronger encryption methods, and greater involvement from organizations and regulatory bodies to ensure robust safeguards are in place, thereby relieving individuals of this burden.

While smartwatches have been a part of the consumer landscape for years, their integration into health care is still new. Following the study by Heidel and Hagist [117], their long-term effects must first be demonstrated throughout the next years. Future economic studies will be crucial for assessing smartwatches from a value-based perspective, aiming to quantify their demonstrated impact on enhancing care quality and determining the costs required to realize this value. The technological landscape of smartwatches is rapidly advancing. At present, various innovative features are under development for future market introduction. Significantly, the progression of sensors tailored for health data collection stands out, with upcoming functionalities including blood glucose and blood pressure monitoring. This ongoing evolution undoubtedly establishes smartwatches as crucial components of health care delivery now and moving forward.

Acknowledgments

ChatGPT (version 4.0; OpenAI) was used to revise language and grammar to enhance the clarity of the manuscript and ensure its coherence and readability.

Authors' Contributions

All authors contributed to the main conceptual ideas and the proof outline. CK took the lead in drafting the manuscript, with specific sections coauthored by ES-B (Introduction, The Evolution of Smartwatches in Tracking Health, and Monitoring sections), TS (Nudging section), and AB (Discussion section). DF played a pivotal role by providing vital structural feedback and improvements in the manuscript, thereby enhancing its overall coherence and structure. Together, all authors actively participated in refining and approving the final manuscript.

Conflicts of Interest

None declared.

References

1. Share of respondents who would prefer the following methods for sharing wearable health data with doctors in the United States in 2023. Statista. URL: https://www.statista.com/statistics/1424173/methods-for-sharing-wearable-health-data-in-the-us/ [accessed 2024-08-16]
2. Connecting patients for better health: 2016. Canada Health Infoway. 2016 Oct 12. URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/3152-connecting-patients-for-better-health-2016 [accessed 2024-08-16]
3. Accenture study finds growing demand for digital health services revolutionizing delivery models: patients, doctors + machines. Accenture. 2018 Mar 6. URL: https://newsroom.accenture.com/news/2018/accenture-study-finds-growing-demand-for-digital-health-services-revolutionizing-delivery-models-patients-doctors-machines [accessed 2024-07-31]
4. Pando A. Wearable health technologies and their impact on the health industry. Forbes. 2019 May 2. URL: https://www.forbes.com/councils/forbestechcouncil/2019/05/02/wearable-health-technologies-and-their-impact-on-the-health-industry/ [accessed 2024-07-31]

XSL•FO
RenderX

5. Dinh-Le C, Chuang R, Chokshi S, Mann D. Wearable health technology and electronic health record integration: scoping review and future directions. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e12861 [FREE Full text] [doi: 10.2196/12861] [Medline: 31512582]
6. Wark K, Cheung K, Wolter E, Avey JP. "Engaging stakeholders in integrating social determinants of health into electronic health records: a scoping review". Int J Circumpolar Health 2021 Dec 12;80(1):1943983 [FREE Full text] [doi: 10.1080/22423982.2021.1943983] [Medline: 34252016]
7. Tagesordnungen und Termine. Bundesrat. 2024. URL: https://www.bundesrat.de/DE/plenum/bundesrat-kompakt/24/1041/1041-pk.html?nn=4352766#top-4 [accessed 2024-07-31]
8. Bin KJ, De Pretto LR, Sanchez FB, Battistella LR. Digital platform to continuously monitor patients using a smartwatch: preliminary report. JMIR Form Res 2022 Sep 15;6(9):e40468 [FREE Full text] [doi: 10.2196/40468] [Medline: 36107471]
9. Selder JL, Te Kolste HJ, Twisk J, Schijven M, Gielen W, Allaart CP. Accuracy of a standalone atrial fibrillation detection algorithm added to a popular wristband and smartwatch: prospective diagnostic accuracy study. J Med Internet Res 2023 May 26;25:e44642 [FREE Full text] [doi: 10.2196/44642] [Medline: 37234033]
10. Ding EY, Han D, Whitcomb C, Bashar SK, Adaramola O, Soni A, et al. Accuracy and usability of a novel algorithm for detection of irregular pulse using a smartwatch among older adults: observational study. JMIR Cardio 2019 May 15;3(1):e13850 [FREE Full text] [doi: 10.2196/13850] [Medline: 31758787]
11. Manini TM, Mendoza T, Battula M, Davoudi A, Kheirkhahan M, Young ME, et al. Perception of older adults toward smartwatch technology for assessing pain and related patient-reported outcomes: pilot study. JMIR Mhealth Uhealth 2019 Mar 26;7(3):e10044 [FREE Full text] [doi: 10.2196/10044] [Medline: 30912756]
12. Elzinga WO, Prins S, Borghans LG, Gal P, Vargas GA, Groeneveld GJ, et al. Detection of clenbuterol-induced changes in heart rate using at-home recorded smartwatch data: randomized controlled trial. JMIR Form Res 2021 Dec 30;5(12):e31890 [FREE Full text] [doi: 10.2196/31890] [Medline: 34967757]
13. Merschel S, Reinhardt L. Analyzability of photoplethysmographic smartwatch data by the preventicus heartbeats algorithm during everyday life: feasibility study. JMIR Form Res 2022 Mar 28;6(3):e29479 [FREE Full text] [doi: 10.2196/29479] [Medline: 35343902]
14. Nelson EE, Rousseau MA, Black TA, George MN, Rashid RM. Smartwatch technology in medicine: a call for future dermatologic research. JMIR Dermatol 2023 Oct 16;6:e47252 [FREE Full text] [doi: 10.2196/47252] [Medline: 37843896]
15. Smartwatches - worldwide. Statista. URL: https://www.statista.com/outlook/hmo/digital-health/digital-fitness-well-being/fitness-trackers/smartwatches/worldwide [accessed 2024-07-31]
16. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, Muzny M, Hartvigsen G, Hopstock LA, et al. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. J Med Internet Res 2018 Mar 22;20(3):e110 [FREE Full text] [doi: 10.2196/jmir.9157] [Medline: 29567635]
17. Bangert B. Local 12: Apple Watch saves local woman's life from deadly blood clot. University of Cincinnati. 2023 Jun 16. URL: https://www.uc.edu/news/articles/2023/06/local-12--apple-watch-saves-local-womans-life-from-deadly-blood-clot.html [accessed 2024-07-31]
18. Fan J, Dai H, Guo Y, Xu J, Wang L, Jiang J, et al. Smartwatch-detected arrhythmias in patients after transcatheter aortic valve replacement (TAVR): analysis of the SMART TAVR trial. J Med Internet Res 2024 Jul 19;26:e41843 [FREE Full text] [doi: 10.2196/41843] [Medline: 39028996]
19. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917 [FREE Full text] [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
20. Zhang Z, Khatami R. Can we trust the oxygen saturation measured by consumer smartwatches? Lancet Respir Med 2022 May;10(5):e47-e48. [doi: 10.1016/s2213-2600(22)00103-5]
21. Zhang Z, Qi M, Hügli G, Khatami R. The challenges and pitfalls of detecting sleep hypopnea using a wearable optical sensor: comparative study. J Med Internet Res 2021 Jul 29;23(7):e24171 [FREE Full text] [doi: 10.2196/24171] [Medline: 34326039]
22. Ernst CP, Ernst AW. The influence of privacy risk on smartwatch usage. In: Proceedings of the Twenty-Second Americas Conference on Information Systems. 2016 Presented at: AMCIS 2016; August 11-14, 2016; San Diego, CA.
23. Cilliers L. Wearable devices in healthcare: privacy and information security issues. Health Inf Manag 2020 May 30;49(2-3):150-156. [doi: 10.1177/1833358319851684] [Medline: 31146589]
24. Robeznieks A. Wearables, the FDA and patient advice: what physicians should know. American Medical Association. 2019 Jul 16. URL: https://tinyurl.com/4kfecaxk [accessed 2024-08-19]
25. Phelan D. Apple Watch heart rate features now live in 19 more countries. Forbes. 2019 Mar 28. URL: https://www.forbes.com/sites/davidphelan/2019/03/28/apple-watch-series-4-ecg-ekg-feature-goes-live-in-19-more-countries-including-u-k-how-to-use-it/ [accessed 2024-08-19]
26. Gurman M. Apple makes major progress on no-prick blood glucose tracking for its watch. Bloomberg. 2023 Feb 24. URL: https://www.bloomberg.com/news/articles/2023-02-22/apple-watch-blood-glucose-monitor-could-revolutionize-diabetes-care-aapl [accessed 2024-07-31]

XSL•FO
RenderX

27. Bianchini M, Bolzan P, Parini B, Maffei S, Cipriani F. Wear-to-care. Co-designing the next wave of open wearables in the healthcare sector. In: Proceedings of the Second EAI International Conference on Wearables in Healthcare. 2020 Presented at: HealthWear 2020; December 10-11, 2020; Virtual Event. [doi: 10.1007/978-3-030-76066-3_10]
28. AKTIBIPO-VALIDATION: a study to validate benefits of Mindpax system in managing bipolar disorder (VAL1M). National Institutes of Health, National Library of Medicine. URL: https://clinicaltrials.gov/study/NCT04945057 [accessed 2024-07-31]
29. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
30. Lauterbach CJ, Romano PA, Greisler LA, Brindle RA, Ford KR, Kuennen MR. Accuracy and reliability of commercial wrist-worn pulse oximeter during normobaric hypoxia exposure under resting conditions. Res Q Exerc Sport 2021 Sep;92(3):549-558. [doi: 10.1080/02701367.2020.1759768] [Medline: 32633688]
31. Awolusi I, Marks E, Hallowell M. Wearable technology for personalized construction safety monitoring and trending: review of applicable devices. Autom Constr 2018 Jan;85:96-106. [doi: 10.1016/j.autcon.2017.10.010]
32. Ching KW, Singh MM. Wearable technology devices security and privacy vulnerability analysis. Int J Netw Secur Appl 2016 May 30;8(3):19-30. [doi: 10.5121/ijnsa.2016.8302]
33. Perez AJ, Zeadally S. Privacy issues and solutions for consumer wearables. IT Prof 2017 Jun 16;20(4):46-56 [FREE Full text] [doi: 10.1109/MITP.2017.265105905]
34. Dehghani M. Exploring the motivational factors on continuous usage intention of smartwatches among actual users. Behav Inf Technol 2018 Jan 17;37(2):145-158. [doi: 10.1080/0144929x.2018.1424246]
35. Pal D, Tassanaviboon A, Arpnikanondt C, Papasratorn B. Quality of experience of smart-wearables: from fitness-bands to smartwatches. IEEE Consumer Electron Mag 2020 Jan 1;9(1):49-53. [doi: 10.1109/mce.2019.2941462]
36. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
37. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018 Sep 27;15(5):429-448. [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
38. Kheirkhahan M, Nair S, Davoudi A, Rashidi P, Wanigatunga AA, Corbett DB, et al. A smartwatch-based framework for real-time and online assessment and mobility monitoring. J Biomed Inform 2019 Jan;89:29-40 [FREE Full text] [doi: 10.1016/j.jbi.2018.11.003] [Medline: 30414474]
39. Randazzo V, Pasero E, Navaretti S. VITAL-ECG: a portable wearable hospital. In: Proceedings of the IEEE Sensors Applications Symposium (SAS). 2018 Presented at: SAS 2018; March 12-14, 2018; Seoul, Korea (South). [doi: 10.1109/sas.2018.8336776]
40. Chen X, Xiao Y, Tang Y, Fernandez-Mendoza J, Cao G. ApneaDetector: detecting sleep apnea with smartwatches. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Jun 24;5(2):1-22. [doi: 10.1145/3463514]
41. Mauldin TR, Canby ME, Metsis V, Ngu AH, Rivera CC. SmartFall: a smartwatch-based fall detection system using deep learning. Sensors (Basel) 2018 Oct 09;18(10):3363 [FREE Full text] [doi: 10.3390/s18103363] [Medline: 30304768]
42. Casilari E, Oviedo-Jiménez MA. Automatic fall detection system based on the combined use of a smartphone and a smartwatch. PLoS One 2015 Nov 11;10(11):e0140929 [FREE Full text] [doi: 10.1371/journal.pone.0140929] [Medline: 26560737]
43. Brew B, Faux SG, Blanchard E. Effectiveness of a smartwatch app in detecting induced falls: observational study. JMIR Form Res 2022 Mar 21;6(3):e30121 [FREE Full text] [doi: 10.2196/30121] [Medline: 35311686]
44. Haescher M, Matthies, DJ, Srinivasan K, Bieber G. Mobile assisted living: smartwatch-based fall risk assessment for elderly people. In: Proceedings of the 5th International Workshop on Sensor-based Activity Recognition and Interaction. 2018 Presented at: iWOAR '18; September 20-21, 2018; Berlin, Germany. [doi: 10.1145/3266157.3266210]
45. Sheng Y, Doyle J, Bond R, Jaiswal R, Gavin S, Dinsmore J. Home-based digital health technologies for older adults to self-manage multiple chronic conditions: a data-informed analysis of user engagement from a longitudinal trial. Digit Health 2022;8:20552076221125957 [FREE Full text] [doi: 10.1177/20552076221125957] [Medline: 36171962]
46. Ibeneme S, Karamagi H, Muneene D, Goswami K, Chisaka N, Okeibunor J. Strengthening health systems using innovative digital health technologies in Africa. Front Digit Health 2022 Mar 31;4:854339 [FREE Full text] [doi: 10.3389/fdgth.2022.854339] [Medline: 35434700]
47. Tai KY, Chiang DL, Chen TS, Shen VR, Lai F, Lin FY. Smart Fall Prediction for Elderly Care Using iPhone and Apple Watch. Wireless Pers Commun 2020 Apr 15;114(1):347-365. [doi: 10.1007/s11277-020-07366-3]
48. Strauss DH, Davoodi NM, Healy M, Metts CL, Merchant RC, Banskota S, et al. The geriatric acute and post-acute fall prevention intervention (GAPcare) II to assess the use of the Apple Watch in older emergency department patients with falls: protocol for a mixed methods study. JMIR Res Protoc 2021 Apr 01;10(4):e24455 [FREE Full text] [doi: 10.2196/24455] [Medline: 33792553]
49. Zahedivash A, Chubb H, Giacone H, Boramanand N, Dubin AM, Trela A, et al. Utility of the Apple Watch® for identifying arrhythmias in children. J Am Coll Cardiol 2023 Mar 7;81(8):1563. [doi: 10.1016/S0735-1097(23)02007-7]

XSL•FO
RenderX

50. Yee-Ming Li JY, Kwok SY, Tsao S, Hoi-Yan Chung C, Hing-Sang Wong W, Cheung YF. Detection of QT interval prolongation using Apple Watch electrocardiogram in children and adolescents with congenital long QT syndrome. Int J Cardiol Heart Vasc 2023 Aug;47:101232 [FREE Full text] [doi: 10.1016/j.ijcha.2023.101232] [Medline: 37346232]
51. Amiri A, Peltier N, Goldberg C, Sun Y, Nathan A, Hiremath SV, et al. WearSense: detecting autism stereotypic behaviors through smartwatches. Healthcare (Basel) 2017 Feb 28;5(1):11 [FREE Full text] [doi: 10.3390/healthcare5010011] [Medline: 28264474]
52. Aguado-García A, Arroyo-Valerio A, Escobedo G, Bueno-Hernández N, Olguín-Rodríguez PV, Müller MF, et al. Opportune warning of COVID-19 in a Mexican health care worker cohort: discrete beta distribution entropy of smartwatch physiological records. Biomed Signal Process Control 2023 Jul;84:104975 [FREE Full text] [doi: 10.1016/j.bspc.2023.104975] [Medline: 37125410]
53. Düking P, Van Hooren B, Sperlich B. Assessment of peak oxygen uptake with a smartwatch and its usefulness for training of runners. Int J Sports Med 2022 Jun 30;43(7):642-647 [FREE Full text] [doi: 10.1055/a-1686-9068] [Medline: 35094376]
54. Brunner G, Melnyk D, Sigfússon B, Wattenhofer R. Swimming style recognition and lap counting using a smartwatch and deep learning. In: Proceedings of the 2019 ACM International Symposium on Wearable Computers. 2019 Presented at: ISWC '19; September 9-13, 2019; London, UK. [doi: 10.1145/3341163.3347719]
55. Chakrabarti S, Biswas N, Jones LD, Ashili S. Role of smartwatches in diabetes patient care. Arch Diab Obes 2022;4(2):431-432. [doi: 10.32474/ADO.2022.04.000185]
56. Hilty DM, Armstrong CM, Luxton DD, Gentry MT, Krupinski EA. A scoping review of sensors, wearables, and remote monitoring for behavioral health: uses, outcomes, clinical competencies, and research directions. J Technol Behav Sci 2021 Mar 03;6:278-313. [doi: 10.1007/s41347-021-00199-2]
57. Spender A, Bullen C, Altmann-Richer L, Cripps J, Duffy R, Falkous C, et al. Wearables and the internet of things: considerations for the life and health insurance industry. Br Actuar J 2019 Jul 18;24:e22. [doi: 10.1017/s1357321719000072]
58. De Sario Velasquez GD, Borna S, Maniaci MJ, Coffey JD, Haider CR, Demaerschalk BM, et al. Economic perspective of the use of wearables in health care: a systematic review. Mayo Clin Proc Digit Health 2024 Sep;2(3):299-317. [doi: 10.1016/j.mcpdig.2024.05.003]
59. Zarak MS, Khan SA, Majeed H, Yasinzai AQ, Hamzazai W, Chung D, et al. Systematic review of validation studies for the use of wearable smartwatches in the screening of atrial fibrillation. Int J Arrhythm 2024 Jun 26;25:11. [doi: 10.1186/s42444-024-00118-5]
60. Jiang Y, Spies C, Magin J, Bhosai SJ, Snyder L, Dunn J. Investigating the accuracy of blood oxygen saturation measurements in common consumer smartwatches. PLOS Digit Health 2023 Jul 12;2(7):e0000296. [doi: 10.1371/journal.pdig.0000296]
61. Kooner P, Baskaran S, Gibbs V, Wein S, Dimentberg R, Albers A. Commercially available activity monitors such as the fitbit charge and Apple Watch show poor validity in patients with gait aids after total knee arthroplasty. J Orthop Surg Res 2024 Jul 15;19(1):404 [FREE Full text] [doi: 10.1186/s13018-024-04892-9] [Medline: 39004751]
62. Koerber D, Khan S, Shamsheri T, Kirubarajan A, Mehta S. Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: a systematic review. J Racial Ethn Health Disparities 2023 Dec 14;10(6):2676-2684 [FREE Full text] [doi: 10.1007/s40615-022-01446-9] [Medline: 36376641]
63. Andreotta AJ, Kirkham N, Rizzi M. AI, big data, and the future of consent. AI Soc 2021 Aug 30;37:1715-1728. [doi: 10.1007/s00146-021-01262-5]
64. Hansen PG, Skov LR, Skov KL. Making healthy choices easier: regulation versus nudging. Annu Rev Public Health 2016 Mar 18;37(1):237-251. [doi: 10.1146/annurev-publhealth-032315-021537] [Medline: 26735430]
65. Marchiori DR, Adriaanse MA, De Ridder DT. Unresolved questions in nudging research: putting the psychology back in nudging. Soc Pers Psychol Compass 2017 Jan;11(1):e12297. [doi: 10.1111/spc3.12297]
66. Gupta M, Sinha N, Singh P, Chuah SH. Gender differences in the wearable preferences, device and advertising value perceptions: smartwatches vs. fitness trackers. Int J Technol Mark 2020;14(2):199. [doi: 10.1504/ijtmkt.2020.110127]
67. Sangameswaran S. Design implications for design of mind-body technology for adolescent sleep. University of Washington. 2023. URL: https://digital.lib.washington.edu/researchworks/items/5b96fa8e-dc97-44b8-a1e8-aa59c708b973 [accessed 2024-10-08]
68. Bergram K, Djokovic M, Bezençon V, Holzer A. The digital landscape of nudging: a systematic literature review of empirical research on digital nudges. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3517638]
69. Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. The Lancet Digit Health 2022 Aug;4(8):E615-E626. [doi: 10.1016/s2589-7500(22)00111-x]
70. Li C, Chen X, Bi X. Wearable activity trackers for promoting physical activity: a systematic meta-analytic review. Int J Med Inform 2021 Aug;152:104487. [doi: 10.1016/j.ijmedinf.2021.104487] [Medline: 34020170]
71. Chung J, Brakey HR, Reeder B, Myers O, Demiris G. Community-dwelling older adults' acceptance of smartwatches for health and location tracking. Int J Older People Nurs 2023 Jan 12;18(1):e12490 [FREE Full text] [doi: 10.1111/opn.12490] [Medline: 35818900]

XSL•FO
RenderX

72. Boateng G, Petersen CL, Kotz D, Fortuna KL, Masutani R, Batsis JA. A smartwatch step-counting app for older adults: development and evaluation study. JMIR Aging 2022 Aug 10;5(3):e33845 [FREE Full text] [doi: 10.2196/33845] [Medline: 35947445]
73. Rosales A, Fernández-Ardèvol M, Comunello F, Mulargia S, Ferran-Ferrer N. Older people and smartwatches, initial experiences. Profesional De La Información 2017 Jun 11;26(3):457-463. [doi: 10.3145/epi.2017.may.12]
74. Zuschuss zu Apple-Watch oder fitness-tracker. Krankenkassen. Deutschland. URL: https://www.krankenkassen.de/gesetzliche-krankenkassen/leistungen-gesetzliche-krankenkassen/apps/apple-watch/ [accessed 2024-07-31]
75. Chuah SH, Rauschnabel PA, Krey N, Nguyen B, Ramayah T, Lade S. Wearable technologies: the role of usefulness and visibility in smartwatch adoption. Comput Hum Behav 2016 Dec;65:276-284. [doi: 10.1016/j.chb.2016.07.047]
76. Shewale R. Smartwatch statistics 2024: worldwide market data. DemandSage. 2024 Mar 24. URL: https://www.demandsage.com/smartwatch-statistics/ [accessed 2024-07-31]
77. Helsen K, Janssen M, Vos S, Scheerder J. Two of a kind? Similarities and differences between runners and walkers in sociodemographic characteristics, sports related characteristics and wearable usage. Int J Environ Res Public Health 2022 Jul 29;19(15):9284 [FREE Full text] [doi: 10.3390/ijerph19159284] [Medline: 35954641]
78. Constantiou I, Mukkamala A, Sjöklint M, Trier M. Engaging with self-tracking applications: how do users respond to their performance data? Eur J Inf Syst 2022 May 30;32(6):941-961. [doi: 10.1080/0960085x.2022.2081096]
79. Ajmal, Boonya-Ananta T, Rodriguez AJ, Du Le VN, Ramella-Roman JC. Monte Carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (PPG) signal. Biomed Opt Express 2021 Nov 10;12(12):7445. [doi: 10.1364/boe.439893]
80. Rosman L, Gehi A, Lampert R. When smartwatches contribute to health anxiety in patients with atrial fibrillation. Cardiovasc Digit Health J 2020 Jul;1(1):9-10 [FREE Full text] [doi: 10.1016/j.cvdhj.2020.06.004] [Medline: 34386784]
81. Maier-Rigaud R, Böning SL. Nudging towards a healthier life? Conceptualising the role of health apps and wearables between empowerment and manipulation. In: Proceedings of the International Conference on Consumer Research. 2016 Presented at: ICCR 2016; September 26-27, 2016; Bonn, Germany.
82. Janani SR, Subramanian R, Karthik S, Vimalarani C. Healthcare monitoring using machine learning based data analytics. Int J Comput Commun Control 2023 Feb 09;18(1) [FREE Full text] [doi: 10.15837/ijccc.2023.1.4973]
83. Hou C, Zhong X, He P, Xu B, Diao S, Yi F, et al. Predicting breast cancer in Chinese women using machine learning techniques: algorithm development. JMIR Med Inform 2020 Jun 08;8(6):e17364 [FREE Full text] [doi: 10.2196/17364] [Medline: 32510459]
84. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814 [FREE Full text] [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]
85. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021 Jan 28;12:625247 [FREE Full text] [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
86. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014 Aug 01;35(29):1925-1931 [FREE Full text] [doi: 10.1093/eurheartj/ehu207] [Medline: 24898551]
87. Stojancic RS, Subramaniam A, Vuong C, Utkarsh K, Golbasi N, Fernandez O, et al. Predicting pain in people with sickle cell disease in the day hospital using the commercial wearable Apple Watch: feasibility study. JMIR Form Res 2023 Mar 14;7:e45355 [FREE Full text] [doi: 10.2196/45355] [Medline: 36917171]
88. Liu JH, Shih CY, Huang HL, Peng JK, Cheng SY, Tsai JS, et al. Evaluating the potential of machine learning and wearable devices in end-of-life care in predicting 7-day death events among patients with terminal cancer: cohort study. J Med Internet Res 2023 Aug 18;25:e47366 [FREE Full text] [doi: 10.2196/47366] [Medline: 37594793]
89. Fonseka LN, Woo BK. Wearables in schizophrenia: update on current and future clinical applications. JMIR Mhealth Uhealth 2022 Apr 07;10(4):e35600 [FREE Full text] [doi: 10.2196/35600] [Medline: 35389361]
90. Brasil S, Pascoal C, Francisco R, Dos Reis Ferreira V, Videira PA, Valadão AG. Artificial intelligence (AI) in rare diseases: is the future brighter? Genes (Basel) 2019 Nov 27;10(12):978 [FREE Full text] [doi: 10.3390/genes10120978] [Medline: 31783696]
91. Brasil S, Neves CJ, Rijoff T, Falcão M, Valadão G, Videira PA, et al. Artificial intelligence in epigenetic studies: shedding light on rare diseases. Front Mol Biosci 2021 May 5;8:648012 [FREE Full text] [doi: 10.3389/fmolb.2021.648012] [Medline: 34026829]
92. Colombage L, Amarasiri T, Sanjeewani T, Senevirathne C, Panchendrarajan R. SmartCare: detecting heart failure and diabetes using smartwatch. In: Proceedings of the IEEE International Conference on Smart Internet of Things (SmartIoT). 2022 Presented at: SmartIoT 2022; August 19-21, 2022; Suzhou, China. [doi: 10.1109/smartiot55134.2022.00013]
93. Nazarian S, Lam K, Darzi A, Ashrafian H. Diagnostic accuracy of smartwatches for the detection of cardiac arrhythmia: systematic review and meta-analysis. J Med Internet Res 2021 Aug 27;23(8):e28974 [FREE Full text] [doi: 10.2196/28974] [Medline: 34448706]

XSL•FO
RenderX

94. Moshawrab M, Adda M, Bouzouane A, Ibrahim H, Raad A. Smart wearables for the detection of cardiovascular diseases: a systematic literature review. Sensors 2023 Jan 11;23(2):828. [doi: 10.3390/S23020828]
95. Mizuno A, Changolkar S, Patel MS. Wearable devices to monitor and reduce the risk of cardiovascular disease: evidence and opportunities. Annu Rev Med 2021 Jan 27;72(1):459-471. [doi: 10.1146/annurev-med-050919-031534] [Medline: 32886543]
96. Föll S, Lison A, Maritsch M, Klingberg K, Lehmann V, Züger T, et al. A scalable risk-scoring system based on consumer-grade wearables for inpatients with COVID-19: statistical analysis and model development. JMIR Form Res 2022 Jun 21;6(6):e35717 [FREE Full text] [doi: 10.2196/35717] [Medline: 35613417]
97. Esmaeilpour Z, Natarajan A, Su HW, Faranesh A, Friel C, Zanos TP, et al. Detection of common respiratory infections, including COVID-19, using consumer wearable devices in health care workers: prospective model validation study. JMIR Form Res 2024 Jul 17;8:e53716 [FREE Full text] [doi: 10.2196/53716] [Medline: 39018555]
98. Zhu T, Watkinson P, Clifton DA. Smartwatch data help detect COVID-19. Nat Biomed Eng 2020 Dec;4(12):1125-1127. [doi: 10.1038/s41551-020-00659-9] [Medline: 33293722]
99. Mishra T, Wang M, Metwally AA, Bogu GK, Brooks AW, Bahmani A, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec;4(12):1208-1220 [FREE Full text] [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
100. Kallel A, Rekik M, Khemakhem M. Hybrid-based framework for COVID-19 prediction via federated machine learning models. J Supercomput 2022;78(5):7078-7105 [FREE Full text] [doi: 10.1007/s11227-021-04166-9] [Medline: 34754141]
101. Abbasi J. Using smartwatch data to detect COVID-19 cases early. JAMA 2020 Dec 08;324(22):2247. [doi: 10.1001/jama.2020.23696] [Medline: 33289811]
102. We want to better understand COVID-19 and its long-term effects. With you, together. Corona-Datenspende. URL: https://corona-datenspende.github.io/en/ [accessed 2024-07-31]
103. Raza MM, Venkatesh KP, Kvedar JC. Intelligent risk prediction in public health using wearable device data. NPJ Digit Med 2022 Oct 13;5(1):153 [FREE Full text] [doi: 10.1038/s41746-022-00701-x] [Medline: 36229593]
104. Gomez Ortega A, Bourgeois J, Hutiri WT, Kortuem G. Beyond data transactions: a framework for meaningfully informed data donation. AI Soc 2023 Aug 30. [doi: 10.1007/s00146-023-01755-5]
105. Gupta R, Iyengar R, Sharma M, Cannuscio CC, Merchant RM, Asch DA, et al. Consumer views on privacy protections and sharing of personal digital health information. JAMA Netw Open 2023 Mar 01;6(3):e231305 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.1305] [Medline: 36862410]
106. Seifert A, Vandelanotte C. The use of wearables and health apps and the willingness to share self-collected data among older adults. Aging Health Res 2021 Sep;1(3):100032. [doi: 10.1016/j.ahr.2021.100032]
107. Hicks JL, Althoff T, Sosic R, Kuhar P, Bostjancic B, King AC, et al. Best practices for analyzing large-scale health data from wearables and smartphone apps. NPJ Digit Med 2019;2:45 [FREE Full text] [doi: 10.1038/s41746-019-0121-1] [Medline: 31304391]
108. Strotbaum V, Pobiruchin M, Schreiweis B, Wiesner M, Strahwald B. Your data is gold – data donation for better healthcare? it - Information Technology 2019;61(5-6):219-229. [doi: 10.1515/itit-2019-0024]
109. About the privacy and security of your health records. Apple. URL: https://support.apple.com/en-ca/111755 [accessed 2024-08-16]
110. How does Samsung Health protect my data. Samsung. URL: https://www.samsung.com/levant/support/apps-services/how-does-samsung-health-protect-my-data/ [accessed 2024-08-16]
111. Keeping data safe at Garmin. Garmin. URL: https://www.garmin.com/en-US/legal/security/ [accessed 2024-08-16]
112. Lewis D. iCloud data breach: hacking and celebrity photos. Forbes. 2014 Sep 22. URL: https://www.forbes.com/sites/davelewis/2014/09/02/icloud-data-breach-hacking-and-nude-celebrity-photos/ [accessed 2024-08-16]
113. Ding EY, CastañedaAvila M, Tran KV, Mehawej J, Filippaios A, Paul T, et al. Usability of a smartwatch for atrial fibrillation detection in older adults after stroke. Cardiovasc Digit Health J 2022 Jun;3(3):126-135 [FREE Full text] [doi: 10.1016/j.cvdhj.2022.03.003] [Medline: 35720675]
114. Chandrasekaran R, Katthula V, Moustakas E. Patterns of use and key predictors for the use of wearable health care devices by US adults: insights from a national survey. J Med Internet Res 2020 Oct 16;22(10):e22443 [FREE Full text] [doi: 10.2196/22443] [Medline: 33064083]
115. de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago JM. Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. Sensors (Basel) 2016 Sep 21;16(9):1538 [FREE Full text] [doi: 10.3390/s16091538] [Medline: 27657081]
116. Foster KR, Torous J. The opportunity and obstacles for smartwatches and wearable sensors. IEEE Pulse 2019 Jan;10(1):22-25. [doi: 10.1109/mpuls.2018.2885832]
117. Heidel A, Hagist C. Potential benefits and risks resulting from the introduction of health apps and wearables into the German statutory health care system: scoping review. JMIR Mhealth Uhealth 2020 Sep 23;8(9):e16444 [FREE Full text] [doi: 10.2196/16444] [Medline: 32965231]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
API: application programming interface
CDSS: clinical decision support system
ECG: electrocardiogram
EHR: electronic health record
LOINC: Logical Observation Identifiers Names and Codes
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SpO_2: saturation of peripheral oxygen

Edited by G Eysenbach; submitted 28.03.24; peer-reviewed by P Sethuramalingam, ME Manaa; comments to author 17.07.24; revised version received 19.08.24; accepted 31.08.24; published 25.10.24.

Please cite as:
Köhler C, Bartschke A, Fürstenau D, Schaaf T, Salgado-Baez E
The Value of Smartwatches in the Health Care Sector for Monitoring, Nudging, and Predicting: Viewpoint on 25 Years of Research
J Med Internet Res 2024;26:e58936
URL: https://www.jmir.org/2024/1/e58936
doi:10.2196/58936
PMID:39356287

©Charlotte Köhler, Alexander Bartschke, Daniel Fürstenau, Thorsten Schaaf, Eduardo Salgado-Baez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Evolution of Health Information Technology for Enhanced Patient-Centric Care in the United States: Data-Driven Descriptive Study

Wesley Barker[1*], MS; Wei Chang[1*], MPH; Jordan Everson[1*], MPP, PhD; Meghan Gabriel[1*], PhD; Vaishali Patel[1*], MPH, PhD; Chelsea Richwine[1*], PhD; Catherine Strawley[1*], MPH

Office of the Assistant Secretary for Technology Policy, US Department of Health and Human Services, Washington, DC, United States

[*]all authors contributed equally

Corresponding Author:
Meghan Gabriel, PhD
Office of the Assistant Secretary for Technology Policy
US Department of Health and Human Services
330 C St SW
Floor 7
Washington, DC, 20201
United States
Phone: 1 202 465 0597
Email: meghan.gabriel@hhs.gov

Abstract

Background: Health information technology (health IT) has revolutionized health care in the United States through interoperable clinical care data exchange, e-prescribing, electronic public health reporting, and electronic patient access to health information.

Objective: This study aims to examine progress in health IT adoption and its alignment with the Office of the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health IT (ASTP's) mission to enhance health care through data access and exchange.

Methods: This study leverages data on end users of health IT to capture trends in engagement in interoperable clinical care data exchange (ability to find, send, receive, and integrate information from outside organizations), e-prescribing, electronic public health reporting, and capabilities to enable patient access to electronic health information. Data were primarily sourced from the American Hospital Association Annual Survey IT Supplement (2008 to 2023), Surescripts e-prescribing use data (2008 to 2023), the National Cancer Institute's Health Information National Trends Survey (2014 to 2022), and the National Center for Health Statistics' National Electronic Health Records Survey (2009 to 2023).

Results: Since 2009, there has been a 10-fold increase in electronic health record (EHR) use among hospitals and a 5-fold increase among physicians. This enabled the interoperable exchange of electronic health information, e- prescribing, electronic public health data exchange, and the means for patients and their caregivers to access crucial personal health information digitally. As of 2023, 70% of hospitals are interoperable, with many providers integrated within EHR systems. Nearly all pharmacies and 92% of prescribers possess e-prescribing capabilities, an 85%-point increase since 2008. In 2013, 40% of hospitals and one-third of physicians allowed patients to view their online medical records. Patient access has improved, with 97% of hospitals and 65% of physicians possessing EHRs that enable patients to access their online medical records. As of 2022, three-fourths of individuals report being offered access to patient portals, and over half (57%) report engaging with their health information through their patient portal. Electronic public health reporting has also seen an increase, with most hospitals and physicians actively engaged in key reporting types.

Conclusions: Federal incentives have contributed to the widespread adoption of EHRs and broad digitization in health care, while efforts to promote interoperability have encouraged collaboration across health care entities. As a result, interoperable clinical care data exchange, e-prescribing, electronic public health reporting, and patient access to health information have grown substantially over the past quarter century and have been shown to improve health care outcomes. However, interoperability hurdles, usability issues, data security concerns, and inequitable patient access persist. Addressing these issues will require collaborative efforts among stakeholders to promote data standardization, implement governance structures, and establish robust health information exchange networks.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59791) doi:10.2196/59791

KEYWORDS

interoperability; e-prescribing; electronic public health reporting; patient access to health information; electronic health records; health IT

Introduction

The growth of health information technology (health IT) adoption and use in the United States over the years has led to better health care delivery, improved health outcomes, and enhanced patient engagement, therefore supporting patient-centric care [1]. This growth in adoption and use of health IT has been supported by numerous regulations and initiatives, including federal initiatives and legislation as well as quality improvement efforts centering around value-based care. The Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009 spurred increases in the use of certified electronic health record (EHR) technology by hospitals and health care professionals [2], reflecting a collaborative effort involving public-private partnerships and various grant programs to enhance the technological foundation of health care facilities nationwide [3]. The HITECH Act was authorized by the Office of the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health IT (ASTP) to develop the Health IT Certification Program and charged the Centers for Medicare and Medicaid Services (CMS) to establish the EHR Incentive Program. CMS used these programs to measure performance related to data capture and sharing for Stage 1 in 2011-2014 and performance related to clinical processes for Stage 2 in 2014 [4,5].

The enactment of the 21st Century Cures Act [6] built upon this momentum and introduced specific provisions to increase the exchange and availability of health information. This act emphasized the critical importance of interoperability and set forth measures to combat information blocking, ensuring that health care data could be exchanged and used more freely and effectively. The Cures Act final rule established further requirements to help mitigate information blocking and support health information exchange [6]. Authorized by the Cures Act, the Trusted Exchange Framework and Common Agreement (TEFCA) [7] was published in 2022, with a go-live that initiated Qualified Health Information Network applications and then designations in late 2023. TEFCA established governance, policy, and technical requirements for interoperability; connectivity to safely exchange information to improve patient care; and support for patient access to health care information. These efforts collectively focus on resolving persistent challenges for interoperability as the health care sector continues to encounter obstacles, underscoring the complexities and challenges in health IT implementation and use [8]. In 2023, the Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Rule [9] continued to implement provisions of the Cures Act to make further headway in the advancement of patient access, interoperability, and standards. Spurred on by these efforts, the advancement of health IT has brought about transformation across the US health care system.

Key advancements in health IT have paved the way for a more efficient and safer health care environment. Despite ongoing complexities, these developments have facilitated the real-time sharing of patient information among health care providers, ultimately leading to improved patient health outcomes. This paper examines the successes and challenges faced in health IT implementation and contextualizes these within the broad landscape of regulation, legislation, and initiatives put forward in the United States. Health IT, as an umbrella term, covers a wide array of technologies designed to improve health care outcomes [10]. Therefore, we focus on a broad set of technologies that have been the focus of public policymaking in support of patient-centric care: interoperable clinical care data exchange, electronic prescribing, public health reporting, and patient access to health information. This data driven analysis offers new insights by providing a critical appraisal of the health IT advances that policy has facilitated and areas where policymakers can further support patient-centric care.

Methods

This analysis describes end-user adoption and use of health IT, focusing on interoperable clinical care data exchange capabilities, e-prescribing practices, electronic public health reporting, and patient access to electronic health information [11]. In the context of this paper, interoperable exchange is measured by the capabilities to find, send, receive, and integrate patient health information electronically for hospitals and office-based physicians. To provide a comprehensive view of the evolving health IT landscape for enhanced patient-centric care, we extracted and analyzed data from 2008 to 2023 from various sources (Table S1 in Multimedia Appendix 1).

The data for assessing interoperability and EHR adoption among hospitals were derived from the American Hospital Association (AHA) IT supplement [12] to the AHA Annual Survey, a nationally representative survey of the US hospitals. Respondents were the Chief Information Officer or the person most knowledgeable about their hospitals' health IT capabilities. The AHA and ASTP have collaborated to monitor the adoption and use of health IT in the US hospitals since 2008. Annual response rates to the IT supplement vary, but average greater than 50% of nonfederal acute care hospitals in the US. Nonfederal acute care hospitals were the focus of our analyses due to their receipt of EHR incentive payments.

The data for assessing interoperability and EHR adoption among physicians were derived from the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics' National Electronic Health Records Survey [13,14]. These data span from 2009 to 2021 and track trends among nonfederally employed office-based physicians in adopting and using EHRs for direct patient care. This survey excludes nondirect care providers such as radiologists and pathologists and provides

XSL•FO
RenderX

valuable insights into the EHR functionalities used in outpatient settings. Annual responses to the survey vary. In 2021, a total of 1875 responses were weighted to reflect national estimates for approximately 403,013 office-based physicians in the United States.

For e-prescribing trends, Surescripts transactional data from 2008 to 2023 [15-17], a prominent e-prescription network in the United States, were used. This network captures data across a wide array of pharmacies, and is used by a majority of chain, franchise, or independently owned pharmacies in the US. Surescripts routed 134 million e-prescribing messages in 2008 and 2.5 billion in 2023.

The National Cancer Institute's Health Information National Trends Survey (HINTS) [18] were used to measure engagement with health IT, tracking how people access and use health information, their use of information technology for health management, and their engagement level in health-related behaviors. This nationally representative survey offers insights into public engagement with health IT and the evolving trends in this domain. The survey has been fielded since 2003; data used for this analysis are from 2014-2022. Data were collected from 3677 respondents in 2014 and 6252 respondents in 2022, and responses were weighted to generate national estimates.

The University of California San Francisco (UCSF) National Physician Health IT Survey was used to measure primary care physicians who reported to immunization information services (IIS) through their EHR. This nationally representative survey of office-based physicians was conducted in partnership with the UCSF, ASTP, and the American Board of Family Medicine in 2022 [19,20]. Data from 1375 respondents specializing in family medicine, internal medicine, obstetrics and gynecology, pediatrics, and other primary care specialties and whose email addresses were listed in the Definitive Healthcare data were collected. Analyses focused upon physicians who used EHRs and who worked in primary care for nonfederal employers. Responses were weighted to generate national estimates.

The CMS established the Promoting Interoperability (PI) Program (formerly the EHR Incentive Program) in 2011 to encourage certain providers and hospitals to adopt, use, and demonstrate the meaningful use of certified electronic health record technologies [21]. These data were used to estimate hospital and physician adoption of electronic capabilities and enablement to submit immunizations, syndromic surveillance, and laboratory reports to public health jurisdictions.

Results

Overview

Many physicians and hospitals benefitted from federal incentive payments, authorized by the HITECH Act, to implement and use certified health IT, shown by the surge in certified health IT adoption among office-based physicians and nonfederal acute care hospitals. EHR adoption and use, underpinning interoperable clinical care data exchange, e-prescribing, electronic public health reporting, and patient access to health information increased over the past 2 decades. Since 2008 there was an 87%-point increase in EHR use among hospitals and a 61%-point increase among physicians (Figure 1).

Figure 1. Electronic health record adoption among hospitals and physicians. Blue line: Percent of office-based physicians; orange line: Percent of nonfederal acute care hospitals. Sources: American Hospital Association Information Technology Supplement (2008-2021) and National Center for Health Statistics Ambulatory Care Survey (2008-2011) and National Electronic Health Record Survey (2012-2021).

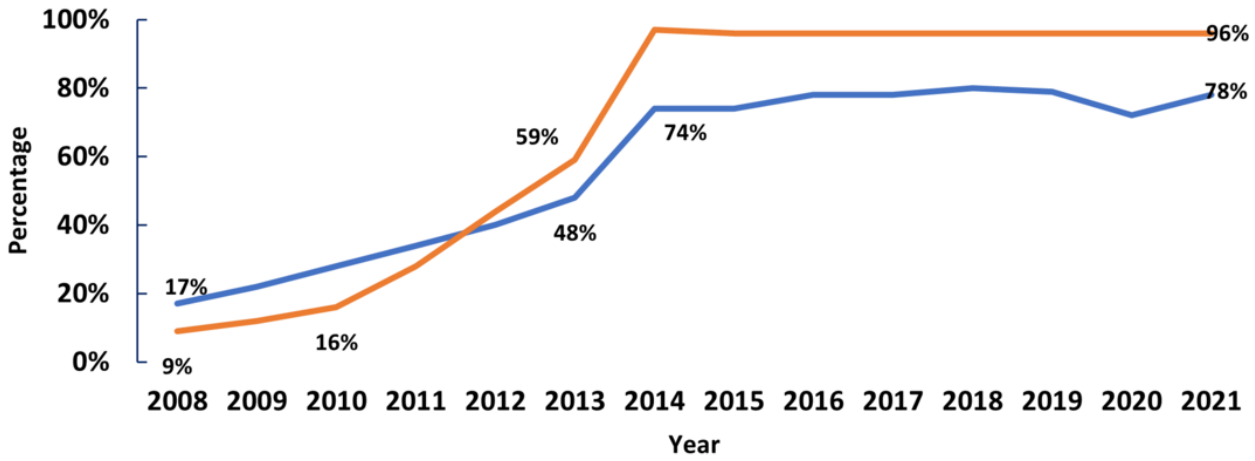

In Figure 1, 2008-2013 includes hospitals and physicians that have adopted an EHR system that integrates patient data, medication tracking, clinician notes, and diagnostic results. 2014 to 2021 includes hospitals and physicians that have adopted a certified EHR system that meets the capability, functionality, and security requirements adopted by the US Department of Health and Human Services. This graph represents data collected from office-based physicians and nonfederal acute care hospitals.

Interoperable Clinical Care Data Exchange

Following this period of rapid digitization, there has been progress related to interoperable data exchange among hospitals (Figure 2). From 2014 to 2023, the engagement in all 4

XSL•FO
RenderX

interoperable exchange domains (find, send, receive, and integrate patient health information electronically) increased. During this decade, the ability to find information increased by 75%, from 48% in 2014 to 84% in 2023. The ability to send information increased by 18%, from 78% in 2014 to 92% in 2023. The ability to receive information increased by 55%, from 56% in 2014 to 87% in 2023. Finally, the ability to integrate information increased by 95%, from 40% to 78% in 2023. In 2014, 23% of hospitals were engaged in all 4 interoperable exchange domains either routinely or sometimes. By 2023, 70% of hospitals sometimes or routinely engaged in all 4 interoperable exchange domains. Notably, while over half of hospitals were not fully interoperable in 2018, by 2023, this declined to 30%, indicating an improvement in interoperable exchange among hospitals.

Figure 2. Hospitals engaging in interoperable exchange of electronic health information. Blue line: Send; orange line: Receive; green line: Find; yellow line: Integrate; Blue dotted line: Hospital engagement in all 4 domains of interoperability. Source: 2014-2023 American Hospital Association Information Technology Supplement. These data were collected from nonfederal acute care hospitals.

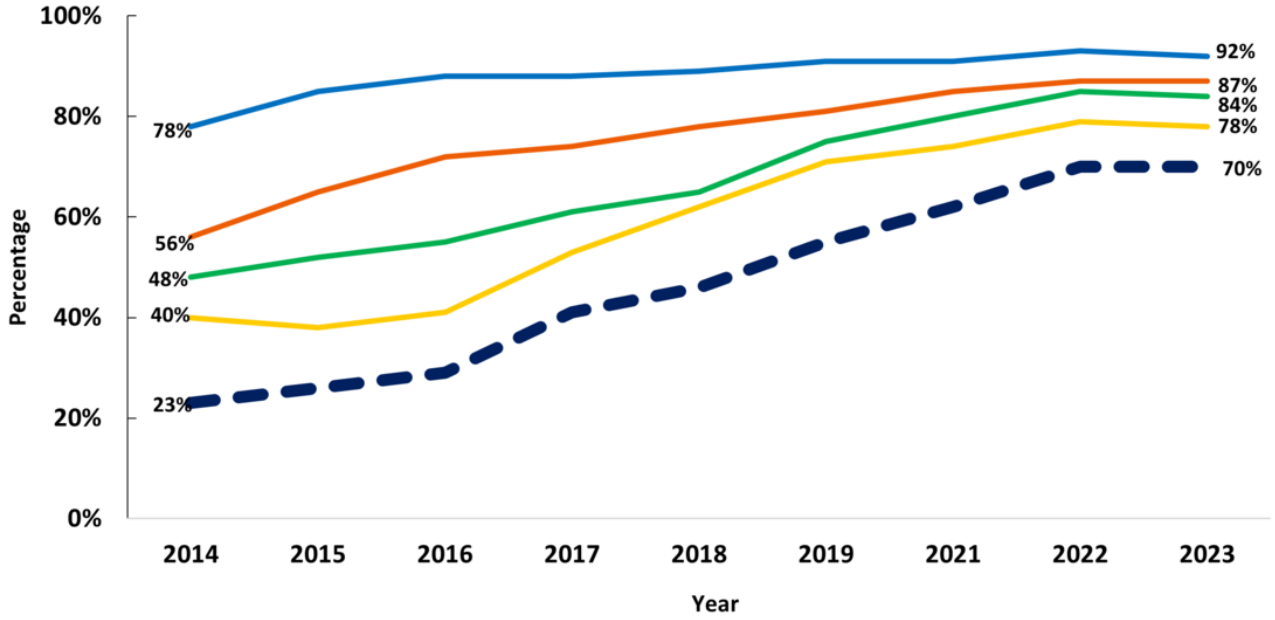

For office-based physicians, there has been modest progress related to interoperable exchange (Figure 3). From 2015 to 2021, like hospitals, office-based physicians' engagement in all 4 interoperable exchange domains increased, most notably, the abilities to find and receive electronic information. During this 7-year period, the ability to find information increased by 44%, the ability to receive information increased by 39%, the ability to send information increased by 3%, and the ability to integrate information increased by 23% among office-based physicians. By 2021, 16% of office-based physicians were engaged in all 4 interoperable exchange domains, with a 78% increase in engagement from 2015 to 2021.

XSL•FO
RenderX

Figure 3. Physicians Engaging in Interoperable Exchange of Electronic Health Information. Green line: Send; orange line: Receive; blue line: Find; yellow line: Integrate; blue dotted line: Physician engagement in all 4 domains of interoperability. Source: 2015-2021 National Electronic Health Record Survey. These data are collected from office-based physicians.

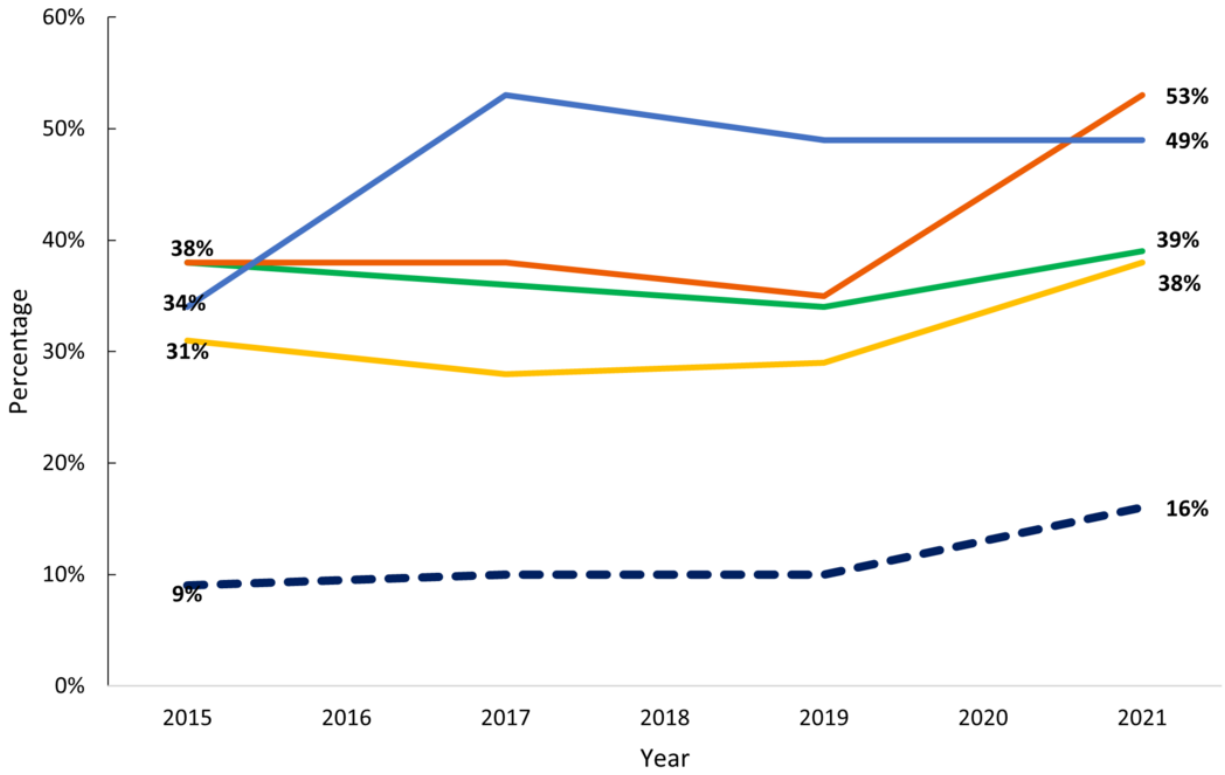

Electronic Prescribing

The percentage of prescribers who e-prescribe has grown from 7% in 2008 to 92% in 2021, an 85%-point increase. Now, there is near-universal adoption among pharmacies, and 92% of prescribers are enabled for e-prescribing (Figure 4). In addition, the electronic prescribing of controlled substances (EPCS) has increased since first allowed in 2010 (Figure 5). Early on, uptake of EPCS was slow, from 0.05% in 2012 to just over 4% in 2015. By 2017, about a third of office-based physicians who prescribe were using EPCS, and as of 2021, nearly three-fourths of them were using EPCS.

XSL•FO
RenderX

Figure 4. Electronic prescribing use among prescribers. Source: Analysis of Electronic Prescribing Transaction Data.

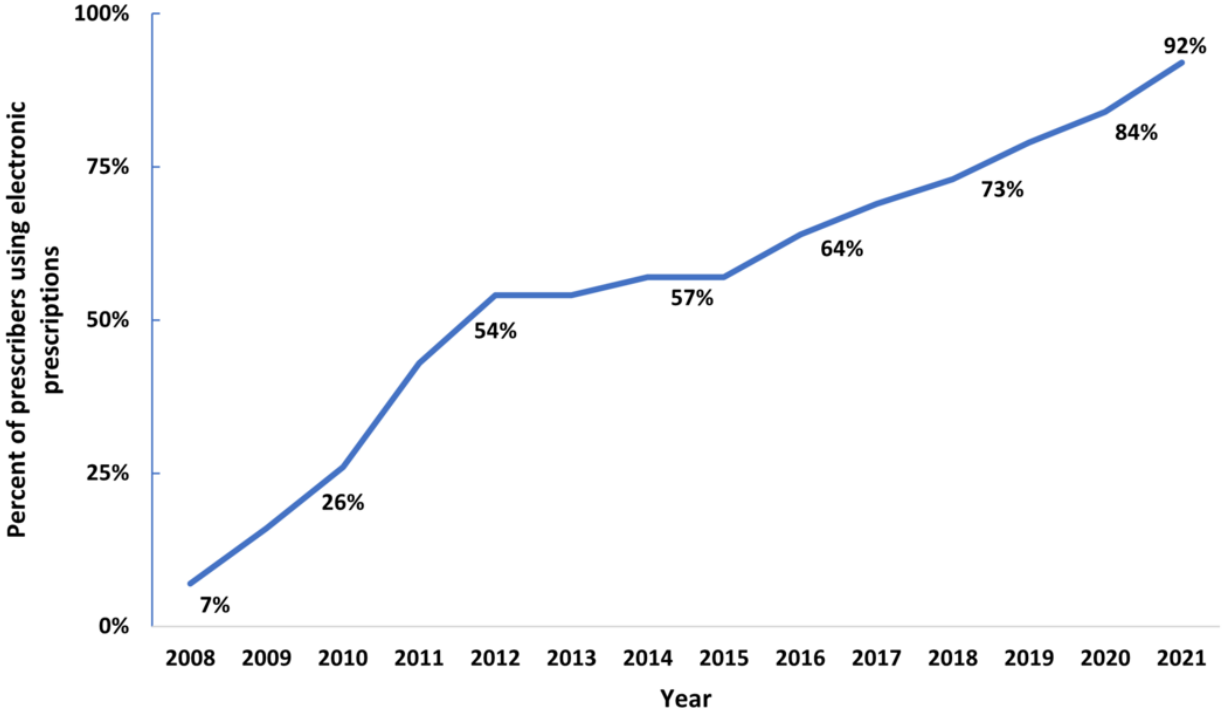

Figure 5. Electronic prescribing of controlled substances. Sources: EPCS (electronic prescribing of controlled substances) Transactional Data (2012-2015, prescribers enabled for ECPS) and National Center for Health Statistics Ambulatory Care Survey National Electronic Health Record Survey (2017-2021, percent of office-based physicians using EPCS).

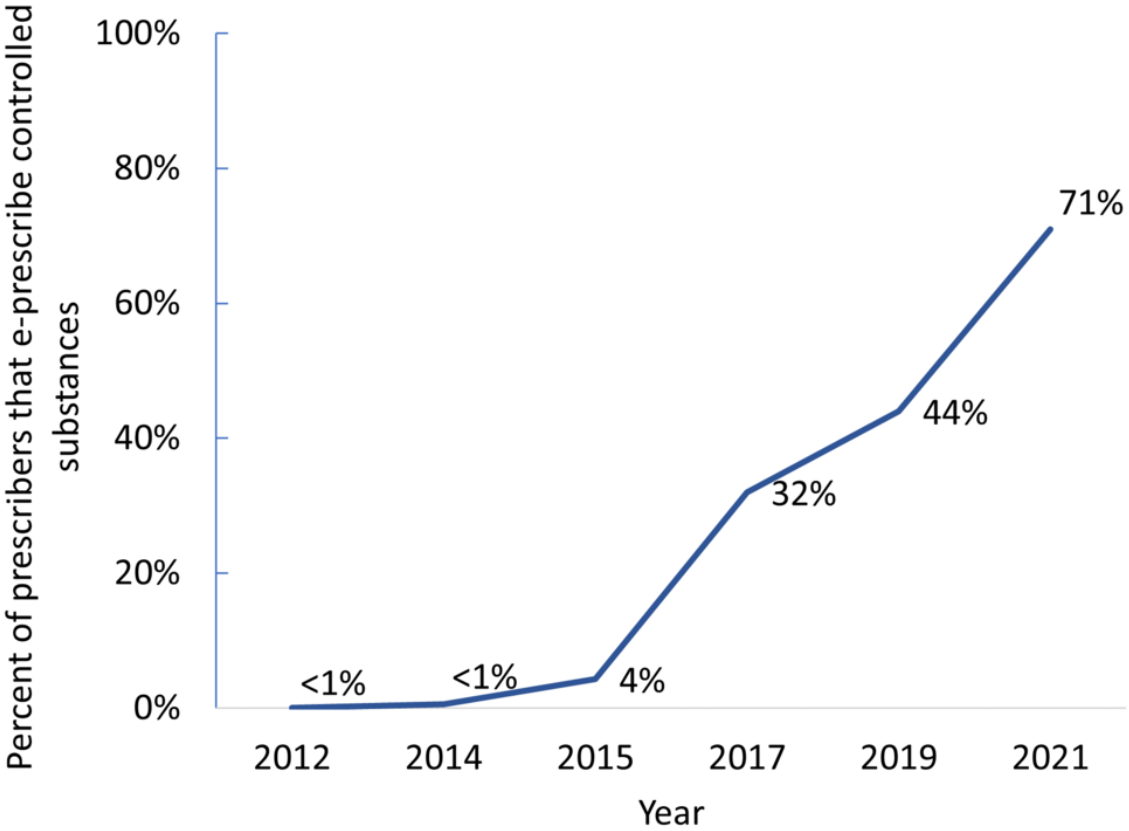

Electronic Public Health Reporting

Public health reporting capabilities among hospitals and primary care physicians have increased over the past decade (Textbox 1). An analysis of CMS's Promoting Interoperability Program data shows that in 2012, among eligible hospitals required to use certified health IT to report public health data electronically, 63% could electronically submit immunization data to their

XSL•FO
RenderX

state IIS, 57% could electronically report laboratory results, and 55% could electronically report syndromic surveillance data to public health agencies. For primary care physicians participating in the program, over half (57%) were able to electronically submit immunization data to their state IIS.

Textbox 1. Public health reporting among hospitals and physicians.

2012

- 63% of participating hospitals enabled to report immunization data to public health agencies.
- 57% of participating hospitals enabled to report laboratory results to public health agencies.
- 55% of participating hospitals enabled to report syndromic surveillance data to public health agencies.
- 57% of participating primary care physicians enabled to report to immunization information services.

2022

- 90% of hospitals enabled to report immunization data to public health agencies.
- 85% of hospitals enabled to report laboratory results to public health agencies.
- 86% of hospitals enabled to report syndromic surveillance data to public health agencies.
- 74% of primary care physicians who viewed immunization data in their electronic health record report data to immunization information services.

By 2022, according to the AHA Health IT Supplement, rates of hospitals' electronic reporting increased for syndromic surveillance (86%) and laboratory reporting (85%) and immunization registry reporting (90%). In addition, as of 2022, about three-fourths of primary care physicians who viewed immunization data in their EHR indicated their primary outpatient EHR reported data to their state IIS.

Patient Access to Health Information

Over the past decade, progress has been made in enabling patient access to electronic health information (Figure 6). In 2012, 24% of nonfederal acute care hospitals offering inpatient care had health IT capable of enabling patients to view their online medical records. As of 2017, 97% of nonfederal acute care hospitals providing inpatient care possessed health IT that enabled patients to view their electronic health information online, predominantly through patient portals or smartphone health apps. This percentage has remained stable throughout the subsequent years, and nearly all (97%) of hospitals offering inpatient care in 2023 enabled patient access to their data.

Figure 6. Access to patient electronic health information. Blue line: Hospital capability to enable patient electronic access to health information. Sources: NEHRS (National Electronic Health Record Survey), AHA (American Hospital Association) IT Supplement, and HINTS (Health Information National Trends Survey) Surveys. Blue line: Hospital capability to enable patient electronic access to health information, grey line: Individuals offered a patient portal by a provider or insurer, yellow line: Individuals offered and accessed a patient portal.

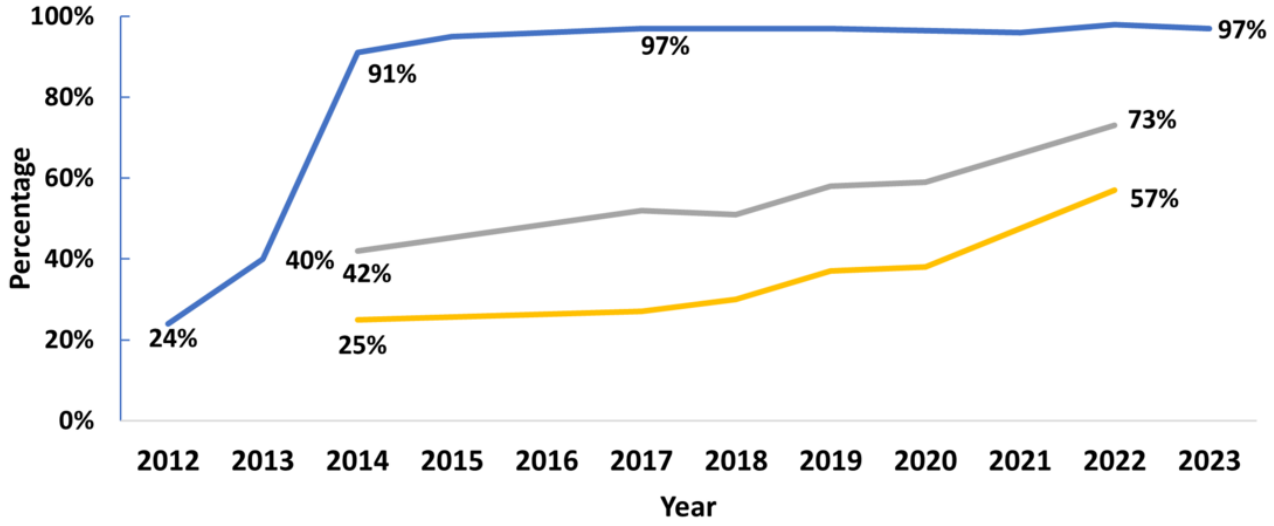

In 2014, 42% of individuals were offered a patient portal by their provider or insurer. In 2022, 73% of individuals reported being offered online access to their medical records by health care providers or insurers, representing a 74% increase since 2014. Only a quarter of individuals were offered and accessed their data in 2014. However, by 2022, the percentage of individuals who were offered and access their online medical records rose to 57%, a 128% increase.

XSL•FO
RenderX

Discussion

Principal Findings

Our analysis provides insights into the outcomes of ongoing efforts to enhance health care delivery through advanced technology and data exchange. High adoption rates for health information technologies among health care professionals and hospitals since 2009 underscore the critical role of federal incentives and collaborative efforts [22-24]. We also highlight the need for continued efforts to overcome the challenges in health IT implementation and ensure that all patients have access to high-quality, patient-centric care. The rise in the adoption of these functionalities reflects the success of initiatives and programs designed to transition health care from mostly paper-based to digitally native. The rise in the adoption of these functionalities reflects a broader trend toward greater patient involvement in their health and care.

Interoperable Clinical Care Data Exchange

The growth in interoperable exchange among health care providers signifies a transformative era in health care communication and data exchange [25]. These findings show that both hospitals and office-based physicians alike have made progress in interoperable exchange and the adoption and use of health IT. However, office-based physicians have consistently lagged behind hospitals. With the modest progress in interoperable exchange among physicians, it is important to continually monitor growth to inform targeted initiatives and support. Furthermore, previous analyses examining variation in interoperable exchange across both hospitals and physicians show that providers with greater resources outpace those with fewer resources, and the latter group continue to face challenges that need to be addressed [26-28].

Despite marked growth in interoperable exchange among hospitals, the majority of hospitals continue to experience interoperability challenges, and lower-resourced hospitals reported challenges with using application programming interfaces (APIs) for more efficient information exchange [29]. Also, more than half of physicians experienced positive benefits from engaging in electronic exchange of patient health information in areas related to practice quality, care coordination, and efficiencies, yet most physicians also experienced barriers to exchange [30]. Although, a recent survey of family medicine physicians revealed that the majority (91%) find it at least somewhat easy to use external information for patient care, around 8 out of 10 reported facing challenges in locating important information and integrating it into their EHR. These data point to the need to explore interventions that support data integration to ensure that accessing and using patient information from external sources becomes seamless.

Despite the investments of the past 2 decades, more work to enable clinical data exchange remains. Of particular importance is data standardization and the seamless exchange of information across diverse health systems. The continuing evolution of technology and legislative and regulatory support are essential to address these barriers and advance interoperable exchange, as this is a critical factor in the integration and advancement of health IT, allowing health care providers to share patient information in real time, leading to improved patient outcomes [31,32]. In addition, necessary information is not always available at the point of care and hospitals may not always be able to share patient information with external providers including other hospitals, ambulatory care, long-term care providers, or behavioral health, even when they have the capability [33]. This can lead to providers frequently missing patient records, as a recent survey of family medicine providers found that a third frequently encountered this issue [28]. Improvements that come with enhanced interoperability will play a role in providing patient information that is needed and when it is needed (at the point of care).

E-Prescribing

The rise in e-prescribing adoption illustrates another shift toward digital health care processes to help improve prescription accuracy, efficiency, and patient safety [27,34]. Research suggests that adherence to early e-prescribing thresholds set forth by the Medicare Promoting Interoperability Program is associated with reduced rates of adverse drug events, indicating the potential of e-prescribing to improve medication accuracy and patient adherence [35]. The integration of e-prescribing into EHRs and the advancement of EPCS have been crucial in this development. The use of e-prescribing is now nearly universal, and the use of EPCS [36] is growing rapidly, and being facilitated by state and federal initiatives [37] and the integration of systems within EHRs.

However, there are still areas for improvement, particularly in ensuring uniform adoption across all health care settings. Future efforts must focus on overcoming these challenges to fully realize the benefits of e-prescribing in enhancing patient care. The use of EPCS has contributed to safer prescribing practices and aiding in combating the opioid crisis. Also, prescription drug monitoring programs [37] (PDMPs) have become integral tools in addressing the opioid epidemic, offering real-time data to aid health care providers in making informed prescribing decisions. While PDMPs have shown promise in reducing inappropriate prescribing, their effectiveness is amplified when integrated with EHRs, highlighting the importance of ease of access and use for health care professionals [38-40]. Recent studies confirmed that physicians frequently use PDMPs [37] and report benefits beyond reduced prescribing, including improved clinical decision-making and overall patient care. The growth of EPCS and other related tools and initiatives, such as PDMPs, have affected improvement on patient-centric care by reducing inappropriate [41] controlled substance prescribing practices and prescription overdoses. While the adoption and use of e-prescribing of medications inclusive of controlled substances has been hailed as a success, there is still room for improvement. Updating the National Council for Prescription Drug Programs Script standards to align with Medicare and other federal programs would help to ensure a unified approach across our health care system [42]. This alignment would help to ensure that these tools enhance patient care function optimally and provide accurate, efficient, and compliant e-prescribing and monitoring of controlled substances.

Furthermore, to reduce the burden associated with medication before approvals, the Health Information Technology Advisory

XSL•FO
RenderX

Committee [43] Taskforce on Pharmacy Interoperability and Emerging Therapeutics [44] has recommended that electronic previous authorizations be included with Health IT Certification requirements and to support better decision-making during prescribing. The implementation of real-time benefit tools [45] is important, as reflected by legislative requirements [46], and will help improve prescription affordability and transparency. To further enhance patient-centric care, it may be beneficial for providers and pharmacies to include additional data elements, such as the indication or reason why a medication was prescribed, on e-prescriptions [47]. This will improve communication between the pharmacist and patient regarding the intended use of the medication and ensure safe and effective care. These updates and improvements to e-prescribing are set to transform it into a more integrated tool for medication management in support of patient-centric care. It would help to improve bidirectional access between parties, uphold patient privacy and security, and address existing gaps in prescription services.

Electronic Public Health Reporting

The adoption of electronic public health reporting capabilities for both hospitals and physicians has changed the landscape of public health surveillance and response, driven by legislative frameworks such as the HITECH Act and furthered by the CMS PI Programs [26,48-50]. The transition from optional to mandatory reporting [51] on public health measures for providers under these programs has increased electronic public health reporting capabilities, streamlining the submission of public health data, and equipping public health agencies with more timely and accurate data [52] needed for disease surveillance and response efforts [2].

Despite this progress, providers continue to face challenges to electronic public health reporting due to a lack of data standardization and use of different vocabulary standards, high implementation costs related to interfaces and data submission or transmission, technical complexity of interfaces, and difficulty extracting relevant information from the EHR [50]. Furthermore, many providers cite public health agencies' lack of capacity to electronically receive information as a major barrier to public health information exchange [26,53]. Varying technical complexities of sending and receiving systems across institutions persist, highlighting a need for comprehensive strategies to enhance system uniformity and functionality to achieve interoperable public health data systems. Nevertheless, the ongoing digitization of public health reporting has facilitated a notable improvement in data exchange and real-time monitoring capabilities.

Public health reporting is crucial for rapid response and effective public health management, as evidenced by the increased adoption rates among hospitals for electronic submission of immunization, syndromic surveillance, and laboratory data and among physicians for IIS reporting efforts. A similar trend is seen among physicians, among whom there was early uptake of exchange, and rates have remained steady since then [19,54]. Recent data show that around half of primary care physicians exchange information with their state's IIS, but among those who viewed immunization data in their EHR, approximately three-fourths of them were able to access and report data to their IIS [19]. Rates of awareness of their EHR's reporting capabilities and overall satisfaction among primary care physicians varied, and the data show that those physicians that exchanged with their IIS through their primary outpatient EHR have a higher rate of satisfaction than those using an outside portal, paper, or fax, showing the perceived benefits of integrating these capabilities within the EHR [19].

Federal mandates to report public health data electronically and underlying Health IT Certification Program requirements that standardize these processes across disparate health IT have played a role in advancing electronic reporting. The mandates led to an uptake in electronically reporting key public health data, underscoring the influence of federal policy in enhancing public health reporting infrastructures. Despite advancements driven by federal requirements and coordination with state, tribal, local, and territorial health departments, interoperable exchange issues and the integration of reporting functionalities within existing health IT present ongoing challenges, highlighting the need for continued collaboration among stakeholders to foster a more integrated and efficient public health reporting ecosystem.

Though progress has been made in electronic public health reporting, addressing existing challenges remains critical for further developing efficient, standardized, and universally accessible public health reporting systems. Enhancing electronic reporting capabilities across all health care settings, particularly among office-based physicians, is essential for achieving comprehensive public health surveillance and ensuring a coordinated response to public health emergencies. These challenges underscore the importance of continued support and resource allocation from federal initiatives to address underlying barriers to the interoperable exchange of public health data. The CDC's Data Modernization Initiative is one such effort aimed at modernizing public health infrastructure needed to facilitate the flow of timely, actionable data to inform public health decision-making and developing a trained public health workforce to support evolving data needs. The CDC's Public Health Data Strategy highlights specific areas of focus for Data Modernization Initiative investment and outlines specific actions and goals, including strengthening the core of public health data and advancing open and interoperable public health data, to ensure critical core data elements can be shared between health care providers and public health [55].

Patient Access to Electronic Health Information

In recent years, advancements in health IT have notably enhanced patient access to electronic health information, underpinned by provisions from the 21st Century Cures Act to reduce the effort for patients to electronically access their health information and increased demand due to the COVID-19 pandemic's impact on health care delivery. The transition to EHRs facilitated by federal mandates fostered an informed and engaged patient population. The adoption of certified health IT has not only streamlined patient engagement but also enhanced the quality of health care delivery. The surge in telehealth and patient portal usage from 2020 to 2022 [56] exemplifies the role that external factors, such as the COVID-19 pandemic, have

XSL•FO
RenderX

played in accelerating the demand for electronic access to health information. As of 2022, most individuals who accessed their data did so to view test results, view clinical notes, or download health information [56]. This period underscored the use of patient portals and health applications in providing crucial health care services and information, thereby contributing [57] to an increase in their adoption and use.

These developments signify a pivot toward more patient-centric health care, with increased patient engagement and demand for transparent, accessible health information [58]. Data indicate that nearly all hospitals report the ability to enable patients to access their information online. Expanded health IT adoption, however, illuminates persistent access and usage differences across different patient groups, accentuated by varying levels of education, income, and ethnicity [59]. These differences highlight the critical need [60] for targeted interventions to bridge gaps and ensure that advancements in health IT benefit all sections of the population equitably. Recent data show that family physicians who practiced in rural areas, that did not have staff and linkages to community programs to address patients' social needs, and that treated a large proportion of vulnerable patients were less likely to be satisfied with access to patient information from external providers [28]. Therefore, the inability to access and use information from external sources creates friction and barriers to patient-centric care.

The shift toward app-based access to health information introduces new dimensions to the health IT landscape, offering increased convenience and patient engagement [56]. However, this shift [61] also raises new challenges when it comes to data privacy, security, and interoperability. As mandated by the Cures Act Final Rule [6], implementing standards-based APIs represents a step forward in addressing these challenges, promoting safer and more efficient access to health information. Addressing these multifaceted challenges and opportunities is imperative for advancing a more inclusive, effective, and patient-centric health care system.

Raising the Bar to Support Patient-Centric Care

Ultimately, health IT has contributed to more patient-centric care in the United States [62], but more work remains to ensure that information is exchanged across the health care delivery system and that exchanged information is easily usable by humans and machines. Enhancing standards for data sharing, expanding transparency and accountability, promoting open APIs, addressing information blocking, and leveraging lessons from the COVID-19 pandemic will help to further facilitate and support patient-centric care.

For example, data standards, such as the United States Core Data for Interoperability [63], which was established in the Cures Act Final Rule [64], includes data elements, classes, and standard code set versions that are continuously and transparently updated through an annual, public process. The implementation of these updated standards and frameworks are integral to support seamless exchange of data elements needed for patient-centric care. In addition, the Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Rule supports provisions from the Cures Act and sets first of its kind transparency requirements for artificial intelligence used in health IT [9]. The establishment of TEFCA, and the adoption of secure, standardized APIs that can be accessed and used without special effort will help to facilitate and support necessary data sharing to promote patient-centric care.

Since the start of this decade, federal policymakers have focused on initiatives and policies to address persistent challenges with interoperable exchange. The implementation of provisions of the 21st Century Cures Act [65], such as defining information blocking, establishing disincentives, and clarifying valid exceptions are initiatives to help support interoperable exchange [64,66]. In addition, insights from the COVID-19 pandemic tell us that although interoperable exchange has not been fully achieved, the rapid adoption and implementation of new digital health and virtual care models, which were vital during the pandemic, were made possible. Due to this, rates of patient engagement and electronic public health reporting have improved, and further advancements to interoperability and transparency through standards and policies that support the access, use, and exchange of patient information are helping to move the needle forward.

As technology advances, it is essential to update and standardize the way health IT applications launch and interact. A specific focus on security, data accessibility, and interoperable exchange will help to ensure that patients can access their electronic health information. These efforts are meant to broaden the playing field, inviting innovators and new models for data sharing, clinical decision-making, and patient access to their electronic health information. The rapid digitization of the past 2 decades has created new opportunities in health care technology and delivery.

This analysis provides a high-level overview of advancements in health IT that have revolutionized patient-centric care in the United States. Continuous measurement and evaluation of health IT policy and program implementation have been vital to identify barriers and help formulate ways to address these challenges. Due to the complex nature and inherent difficulties in quantifying interoperable exchange among physicians, it is important to explore methods for accurate measurement and to address these challenges. This will enable more comprehensive evaluations of interoperability. Furthermore, advancement of indices or tools to holistically measure interoperability and related capabilities could be used to assess the impacts of policies, inform future policy opportunities, and identify areas for targeted support, including performance gaps and health equity concerns. As we move forward into the next quarter-century, it is now the time to be bold and help support tools and processes to ensure patient-centric care [67].

Conclusion

The United States has experienced growth in enhanced information exchange, e-prescribing, electronic public health reporting, and patient access to health information, which have transformed health care delivery and improved patient outcomes. While these achievements are noteworthy, it is essential to continue the collaborative efforts, policy support, and technological innovation to overcome the persistent obstacles, such as nonstandardized health IT systems and financial burdens.

XSL•FO
RenderX

Legislation such as the HITECH Act and 21st Century Cures Act, have played a vital role in addressing these challenges.

Addressing social determinants and ensuring equitable technology access remain central to bridging the digital divide and achieving a comprehensive health care transformation. In the next 25 years, we should shift our focus from health IT adoption to infrastructure improvements necessary for patient-centric care. It is crucial to establish a system where patient data can flow safely and seamlessly. The commitment to advancing health IT will be indispensable in overcoming existing challenges and harnessing the full potential of technological innovations in health care as we transition toward a more integrated and patient-centric health care framework. The ongoing evolution of health IT offers a future where health care delivery is efficient, accessible, and tailored to the needs of every patient, creating another quarter-century of improvements in public health and patient care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data sources.
[DOCX File , 15 KB - jmir_v26i1e59791_app1.docx]

References

1. The Evolution of Patient-Centered Care: Patient Access to Their Health Data. The evolution of patient-centered care: patient access to their health data. Washington, DC: HealthITBuzz; 2024 Jan 23. URL: https://www.healthit.gov/buzz-blog/patient-access/the-evolution-of-patient-centered-care-patient-access-to-their-health-data [accessed 2024-09-03]
2. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
3. DeSalvo KB, Mertz K. Broadening the view of interoperability to include person-centeredness. J Gen Intern Med 2015;30(Suppl 1):S1-S2 [FREE Full text] [doi: 10.1007/s11606-014-3096-2] [Medline: 25480726]
4. Promoting Interoperability Programs. Promoting interoperability programs. Baltimore, MD: U.S. Centers for Medicare & Medicaid Services; 2024 Aug 15. URL: https://www.cms.gov/medicare/regulations-guidance/promoting-interoperability-programs [accessed 2024-09-03]
5. Meaningful Use. Meaningful use. Washington, DC: HealthIT.gov; 2013 Jun 01. URL: https://www.healthit.gov/faq/what-meaningful-use [accessed 2024-09-03]
6. ONC's Cures Act Final Rule. ONC's cures act final rule. Washington, DC: ASTP/ONC; 2024 Mar 24. URL: https://www.healthit.gov/topic/oncs-cures-act-final-rule [accessed 2024-09-03]
7. Trusted Exchange Framework and Common Agreement (TEFCA). Trusted exchange framework and common Agreement (TEFCA). Washington, DC: HealthIT.gov; 2024 Aug 01. URL: https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca [accessed 2024-09-03]
8. Lye CT, Forman HP, Daniel JG, Krumholz HM. The 21st century cures act and electronic health records one year later: will patients see the benefits? J Am Med Inform Assoc 2018;25(9):1218-1220 [FREE Full text] [doi: 10.1093/jamia/ocy065] [Medline: 30184156]
9. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1) Final Rule. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing (HTI-1) final rule. Washington, DC: HealthIT.gov; 2024 Mar 07. URL: https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program [accessed 2024-09-03]
10. Kruse CS, Beane A. Health information technology continues to show positive effect on medical outcomes: systematic review. J Med Internet Res 2018;20(2):e41 [FREE Full text] [doi: 10.2196/jmir.8793] [Medline: 29402759]
11. Mason C. Understand the four levels of interoperability in healthcare.: Wolters Kluwer; 2023. URL: https://www.wolterskluwer.com/en/expert-insights/understand-the-four-levels-of-interoperability-in-healthcare [accessed 2024-09-03]
12. Data from: AHA healthcare IT database. AHA Data & Insights. 2024. URL: https://www.ahadata.com/aha-healthcare-it-database [accessed 2024-09-03]
13. 2021 National Electronic Health Records Survey (NEHRS) Public Use File Documentation.: Centers for Disease Control and Prevention; 2022. URL: https://www.cdc.gov/nchs/data/nehrs/nehrs2021doc-508.pdf [accessed 2024-09-03]
14. Statistics NCFH. National Electronic Health Records Survey. URL: https://www.cdc.gov/nchs/nehrs/about.htm [accessed 2024-09-03]
15. Surescripts. 2023 national progress rReport: tracking the impact of health intelligence sharing across the United States. URL: https://surescripts.com/why-surescripts/our-impact/national-progress-report [accessed 2024-09-03]
16. Hufstader Gabriel M, Yang Y, Vaidya V, Wilkins TL. Adoption of electronic prescribing for controlled substances among providers and pharmacies. Am J Manag Care 2014;20(11 Spec No. 17):SP541-SP546 [FREE Full text] [Medline: 25811828]

XSL•FO
RenderX

17. Gabriel ME, Furukawa MF, Vaidya V. Emerging and encouraging trends in e-prescribing adoption among providers and pharmacies. Am J Manag Care 2013;19(9):760-764 [FREE Full text] [Medline: 24304258]
18. National Cancer Institute. What is HINTS?.: National Institutes of Health URL: https://hints.cancer.gov/ [accessed 2024-09-03]
19. Richwine C SC. Electronic Access to Immunization Information among Primary Care Physicians.: HealthIT.gov; 2024. URL: https://www.healthit.gov/data/data-briefs/electronic-access-immunization-information-among-primary-care-physicians#Appendix_Table_A2_70 [accessed 2024-09-03]
20. Hendrix N, Maisel N, Everson J, Patel V, Bazemore A, Rotenstein LS, et al. Impact of response bias in three surveys on primary care providers' experiences with electronic health records. J Am Med Inform Assoc 2024;31(8):1754-1762. [doi: 10.1093/jamia/ocae148] [Medline: 38894620]
21. Promoting interoperability programs.: Centers for Medicare & Medicaid Services URL: https://www.cms.gov/medicare/regulations-guidance/promoting-interoperability-programs [accessed 2024-09-03]
22. King J, Patel V, Furukawa M. Physician adoption of electronic health record technology to meet meaningful use objectives: 2009-2012.: Office of the National Coordinator for Health Information Technology; 2012. URL: https://www.healthit.gov/data/data-briefs/physician-adoption-electronic-health-record-technology-meet-meaningful-use [accessed 2024-09-03]
23. Charles D, King J, Furukawa M, Patel V. Hospital adoption of electronic health record technology to meet meaningful use objectives: 2008-2012.: Office of the National Coordinator for Health Information Technology; 2013. URL: https://www.healthit.gov/sites/default/files/oncdatabrief10final.pdf [accessed 2024-09-03]
24. Charles D, Swain M, Patel V. Interoperability among U.S. non-federal acute care hospitals.: Office of the National Coordinator for Health Information Technology; 2015. URL: https://www.healthit.gov/sites/default/files/briefs/onc_databrief25_interoperabilityv16final_081115.pdf [accessed 2024-09-03]
25. Charles D, Gabriel M, Furukawa M. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2013.: Office of the National Coordinator for Health Information Technology; 2014. URL: https://www.healthit.gov/sites/default/files/oncdatabrief16.pdf [accessed 2024-09-03]
26. Richwine C. Progress and ongoing challenges to electronic public health reporting among non-federal acute care hospitals.: Office of the National Coordinator for Health Information Technology; 2023. URL: https://www.healthit.gov/data/data-briefs/progress-and-ongoing-challenges-electronic-public-health-reporting-among-non [accessed 2024-09-03]
27. Pylypchuk Y, Everson J. Interoperability and methods of exchange among hospitals in 2021.: Office of the National Coordinator for Health Information Technology; 2023. URL: https://www.healthit.gov/sites/default/files/2023-01/DB64_Interop_and_Methods_of_Exchange_Among_Hosp.pdf [accessed 2024-09-03]
28. Everson J, Hendrix N, Phillips RL, Adler-Milstein J, Bazemore A, Patel V. Primary care physicians' satisfaction with interoperable health information technology. JAMA Netw Open 2024;7(3):e243793 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.3793] [Medline: 38530309]
29. Hospital use of APIs to enable data sharing between EHRs and apps.: Office of the National Coordinator for Health Information Technology; 2023. URL: https://www.healthit.gov/data/data-briefs/hospital-use-apis-enable-data-sharing-between-ehrs-and-apps [accessed 2024-09-03]
30. Pylypchuk YEJ, Charles D, Patel V. Interoperability among office-based physicians in 2015, 2017, and 2019.: Office of the National Coordinator for Health Information Technology; 2022. URL: https://www.healthit.gov/data/data-briefs/interoperability-among-office-based-physicians-2019 [accessed 2024-09-03]
31. Richardson JE, Abramson EL, Kaushal R. The value of health information exchange. JHL 2012;4:17-23 [FREE Full text] [doi: 10.2147/jhl.s16438]
32. Brenner SK, Kaushal R, Grinspan Z, Joyce C, Kim I, Allard RJ, et al. Effects of health information technology on patient outcomes: a systematic review. J Am Med Inform Assoc 2016;23(5):1016-1036 [FREE Full text] [doi: 10.1093/jamia/ocv138] [Medline: 26568607]
33. Interoperable exchange of patient health information among U.S. hospitals: 2023.: Office of the National Coordinator for Health Information Technology; 2024. URL: https://www.healthit.gov/data/data-briefs/interoperable-exchange-patient-health-information-among-us-hospitals-2023#:~:text=Overall%2C%20in%202023%2C%2071%25hospitals%20often%20used%20ha%20information [accessed 2024-09-03]
34. Hufstader M, Swain M, Furukawa M. State variation in e-prescribing trends in the United States.: Office of the National Coordinator for Health Information Technology; 2012. URL: https://www.healthit.gov/sites/default/files/briefs/us_e-prescribingtrends_onc_brief_4_nov2012.pdf [accessed 2024-09-03]
35. Powers C, Gabriel M, Encinosa W, Mostashari F, Bynum J. Meaningful use stage 2 e-prescribing threshold and adverse drug events in the medicare part D population with diabetes. J Am Med Inform Assoc 2015;22(5):1094-1098 [FREE Full text] [doi: 10.1093/jamia/ocv036] [Medline: 25948698]
36. Abouk R, Powell D. Can electronic prescribing mandates reduce opioid-related overdoses? Econ Hum Biol 2021;42:101000 [FREE Full text] [doi: 10.1016/j.ehb.2021.101000] [Medline: 33865194]
37. Richwine C, Everson J. Electronic prescribing of controlled substances and use of prescription drug monitoring programs among office-based physicians, 2019-2021.: Office of the National Coordinator for Health Information Technology; 2023. URL: https://www.healthit.gov/sites/default/files/2023-01/

XSL•FO
RenderX

Electronic%20Prescribing%20of%20Controlled%20Substances%20and%20Use%20of%20PDMP-Data-brief_508.pdf [accessed 2024-09-03]
38. Parasrampuria S, Blanco M, Barker W. Electronic prescribing of controlled substances (EPCS) among office-based physicians, 2017.: Office of the National Coordinator for Health Information Technology; 2019. URL: https://www.healthit.gov/data/data-briefs/electronic-prescribing-controlled-substances-epcs-among-office-based-physicians#:~:text=In%20,2017%2C%20,32%20percent%20of,prescriptions%2C%20electronically%20prescribe%20controlled%20substances [accessed 2024-09-03]
39. Hufstader GM, Smith JY, Sow M, Joseph S, Wilkins TL. Electronic prescribing of controlled substances: a tool to help promote better patient care. American Journal of Pharmacy Benefits 2016;8(5):185-189.
40. Mehta S, Brown W, Ferguson E, Najera J, Pantell M. The association between prescription drug monitoring programs and controlled substance prescribing: a cross-sectional study using data from 2019 National Electronic Health Records Survey. J Am Med Inform Assoc 2023;30(6):1042-1046 [FREE Full text] [doi: 10.1093/jamia/ocad053] [Medline: 37011637]
41. Mauri AI, Townsend TN, Haffajee RL. The association of state opioid misuse prevention policies with patient- and provider-related outcomes: a scoping review. Milbank Q 2020;98(1):57-105 [FREE Full text] [doi: 10.1111/1468-0009.12436] [Medline: 31800142]
42. E-Prescribing standards and requirements.: Centers for Medicare & Medicaid Services URL: https://www.cms.gov/medicare/regulations-guidance/electronic-prescribing/adopted-standard-and-transactions [accessed 2024-09-03]
43. Recommendations to the National Coordinator for Health IT.: Office of the National Coordinator for Health Information Technology URL: https://www.healthit.gov/topic/federal-advisory-committees/recommendations-national-coordinator-health-it [accessed 2024-09-03]
44. Miri A, Briggs-Malonson M. Final Report of the Health Information Technology Advisory Committee on Pharmacy Interoperability and Emerging Therapeutics.: ONC HealthIT Advisory Committee; 2023. URL: https://www.healthit.gov/sites/default/files/page/2023-11/2023-11-09_PhIET_TF_2023_Recommendations_Transmittal_Letter_508.pdf [accessed 2024-09-03]
45. Everson J, Dusetzina SB. Real-time prescription benefit tools-the promise and peril. JAMA Intern Med 2022;182(11):1137-1138. [doi: 10.1001/jamainternmed.2022.3962] [Medline: 36094566]
46. Changes to medicare advantage and part D will provide better coverage, more access and improved transparency for medicare beneficiaries.: Centers for Medicare & Medicaid Services; 2021. URL: https://www.cms.gov/newsroom/press-releases/changes-medicare-advantage-and-part-d-will-provide-better-coverage-more-access-and-improved [accessed 2024-09-03]
47. Rupp MT, Warholak TL, Murcko AC. Indication or diagnosis should be required on prescriptions. J Manag Care Spec Pharm 2021;27(8):1136-1139 [FREE Full text] [doi: 10.18553/jmcp.2021.27.8.1136] [Medline: 34337989]
48. Office of the National Coordinator for Health Information Technology. U.S. hospital adoption of computerized capabilities to meet meaningful use stage 2 objectives. Health IT Quick-Stat #23. 2014. URL: https://www.healthit.gov/data/quickstats/hospital-adoption-meaningful-use-stage-2-functionalities [accessed 2024-09-03]
49. Heisey-Grove D, Chaput D, Daniel J. Hospital reporting on meaningful use public health measures in 2014.: Office of the National Coordinator for Health Information Technology; 2015. URL: https://www.healthit.gov/data/data-briefs/hospital-reporting-meaningful-use-public-health-measures-2014 [accessed 2024-09-03]
50. Richwine C, Everson J, Patel V. Electronic public health reporting among non-federal acute care hospitals during the COVID-19 pandemic, 2021.: Office of the National Coordinator for Health Information Technology; 2022. URL: https://www.healthit.gov/data/data-briefs/electronic-public-health-reporting-among-non-federal-acute-care-hospitals-during#:~:text=FINDINGS&text=More%20than%20three%2Dquarters%20of,reporting%20(81%25)%20in%202021 [accessed 2024-09-03]
51. Myers E, Smith J. Federal agencies align to promote public health reporting.: Health IT; 2021. URL: https://www.healthit.gov/buzz-blog/health-it/federal-agencies-align-to-promote-public-health-reporting [accessed 2024-09-03]
52. Dixon BE, Zhang Z, Arno JN, Revere D, Joseph Gibson P, Grannis SJ. Improving notifiable disease case reporting through electronic information exchange-facilitated decision support: a controlled before-and-after trial. Public Health Rep 2020;135(3):401-410 [FREE Full text] [doi: 10.1177/0033354920914318] [Medline: 32250707]
53. Holmgren AJ, Apathy NC, Adler-Milstein J. Barriers to hospital electronic public health reporting and implications for the COVID-19 pandemic. J Am Med Inform Assoc 2020;27(8):1306-1309 [FREE Full text] [doi: 10.1093/jamia/ocaa112] [Medline: 32442266]
54. Electronic reporting to immunization information services (IIS) among medicare eligible professionals, 2011-2014.: Office of the National Coordinator for Health IT URL: https://www.healthit.gov/data/quickstats/electronic-reporting-immunization-information-services-iis-among-medicare-eligible [accessed 2024-09-03]
55. Data Modernization Initative.: Centers for Disease Control and Prevention; 2024. URL: https://www.cdc.gov/surveillance/data-modernization/index.html [accessed 2024-09-03]
56. Strawley C, Richwine C. Individuals' access and use of patient portals and smartphone health apps, 2022.: Office of the National Coordinator for Health Information Technology; 2023. URL: https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2022 [accessed 2024-09-03]

XSL•FO
RenderX

57. Barker W, Richwine C. Patient usage of apps to access online medical records. JAMA Netw Open 2023;6(11):e2343312 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.43312] [Medline: 37962892]
58. Richwine C, Johnson C, Patel V. Disparities in patient portal access and the role of providers in encouraging access and use. J Am Med Inform Assoc 2023;30(2):308-317 [FREE Full text] [doi: 10.1093/jamia/ocac227] [Medline: 36451262]
59. Richwine C. Progress and persistent disparities in patient access to electronic health information. JAMA Health Forum 2023;4(11):e233883 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.3883] [Medline: 37948063]
60. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021;4(1):52 [FREE Full text] [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
61. Senft N, Butler E, Everson J. Growing disparities in patient-provider messaging: trend analysis before and after supportive policy. J Med Internet Res 2019;21(10):e14976 [FREE Full text] [doi: 10.2196/14976] [Medline: 31593539]
62. Everson J. Health information technology. Oxford Research Encyclopedia of Economics and Finance 2019. [doi: 10.1093/acrefore/9780190625979.013.102]
63. Marquard B, Haas E, Dolin G. US core implementation guide.: HL7 International URL: https://build.fhir.org/ig/HL7/US-Core/ [accessed 2024-09-03]
64. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program.: Office of the National Coordinator for Health Information Technology; 2020. URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2024-09-03]
65. Public Law 114–255.: Authenticated US Government Information GPO; 2016. URL: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf [accessed 2024-09-03]
66. Grants, Contracts, and Other Agreements: Fraud and Abuse; Information Blocking; Office of Inspector General's Civil Money Penalty Rules.: Federal Register; 2023. URL: https://www.federalregister.gov/documents/2023/07/03/2023-13851/grants-contracts-and-other-agreements-fraud-and-abuse-information-blocking-office-of-inspector [accessed 2024-09-03]
67. Farr C. Digital health companies: it's time to be bold.: Second Opinion; 2024. URL: https://secondopinion.media/p/digital-health-companies-its-time [accessed 2024-09-03]

Abbreviations

AHA: American Hospital Association
API: application programming interface
ASTP: Office of the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health IT
CDC: Centers for Disease Control and Prevention
CMS: Centers for Medicare and Medicaid Services
EHR: electronic health record
EPCS: electronic prescribing of controlled substances
HITECH: Health Information Technology for Economic and Clinical Health
IIS: immunization information service
PDMP: prescription drug monitoring program
PI: Promoting Interoperability
TEFCA: Trusted Exchange Framework and Common Agreement
UCSF: University of California San Francisco

Edited by G Eysenbach; submitted 29.04.24; peer-reviewed by U Sinha, U Shrivastava; comments to author 23.06.24; revised version received 12.07.24; accepted 09.08.24; published 28.10.24.

Please cite as:
Barker W, Chang W, Everson J, Gabriel M, Patel V, Richwine C, Strawley C
The Evolution of Health Information Technology for Enhanced Patient-Centric Care in the United States: Data-Driven Descriptive Study
J Med Internet Res 2024;26:e59791
URL: https://www.jmir.org/2024/1/e59791
doi:10.2196/59791
PMID:39466303

©Wesley Barker, Wei Chang, Jordan Everson, Meghan Gabriel, Vaishali Patel, Chelsea Richwine, Catherine Strawley. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits

XSL•FO
RenderX

unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Advancing the United Nations Sustainable Development Goals Through Digital Health Research: 25 Years of Contributions From the Journal of Medical Internet Research

Raghu Raman[1*], PhD; Monica Singhania[2*], PhD; Prema Nedungadi[1*], PhD

[1]Amrita School of Business, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Kerala, India
[2]Faculty of Management Studies, University of Delhi, New Delhi, India
[*]all authors contributed equally

Corresponding Author:
Raghu Raman, PhD
Amrita School of Business
Amrita School of Computing, Amrita Vishwa Vidyapeetham
Amritapuri
Kerala, 690525
India
Phone: 91 9895028779
Email: raghu@amrita.edu

(J Med Internet Res 2024;26:e60025) doi:10.2196/60025

KEYWORDS
sustainable development goal; topic modeling; public health; surveillance; gender equality; non-communicable disease; social media; COVID-19; SARS-CoV-2; coronavirus; machine learning; artificial intelligence; AI; digital health

Introduction

Celebrating 25 years of the *Journal of Medical Internet Research*'s influence, we examine its contributions to the United Nations (UN) sustainable development goals (SDGs) [1]. Our study explores the interconnectedness of the SDGs using the SDG framework [2]. We aimed to:

- Analyze *Journal of Medical Internet Research* publications using a topic modeling approach to identify predominant topics;
- Map these topics to relevant SDGs to illustrate the journal's impact on global health and equity;
- Demonstrate the interdisciplinary nature of digital health research and its relevance to sustainable development.

Methods

Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we conducted a comprehensive search on Scopus for articles published between 1999 and 2023 and found 8124 publications. Using inbuilt SDG mapping queries, we narrowed the results to 3550 publications (Multimedia Appendix 1) to identify relevant topics and their alignment with the SDGs [3]. The analysis used BERTopic, an advanced topic modeling technique, to review 25 years of *Journal of Medical Internet Research* publications. We used co-citation mapping to visualize SDG linkages and BERTopic to identify semantic connections [4].

Results

Figure 1 presents a network visualization of SDG interconnectivity to highlight the interdisciplinary nature of the research published in the *Journal of Medical Internet Research* [5]. Central to this network is SDG 3 (Good Health and Well-Being), which shows strong ties to SDGs 5 (Gender Equality), 10 (Reduced Inequalities), 16 (Peace, Justice, and Strong Institutions), and 9 (Industry, Innovation, and Infrastructure). These linkages highlight how advancements in health contribute broadly to gender equality, social equity, justice, and economic development, indicating a holistic approach that integrates health with key SDGs.

Figure 2 shows major topics mapped to SDGs. SDG 3, which accounts for 79.8% of the publications, indicates that the journal primarily focused on health and well-being. Other notably featured SDGs include 9 and 10, with significant contributions to innovation in health care and reducing inequalities. SDGs 5 and 16 are also highlighted, supporting gender equality and the development of peaceful societies. Less-represented SDGs, such as SDG 4 (Quality Education) and SGD 6 (Clean Water and Sanitation), suggest expanding research areas, with environmental sustainability goals showing potential for future focus.

XSL•FO
RenderX

Figure 1. Interdisciplinary research published in the *Journal of Medical Internet Research* as a network of the United Nations sustainable development goals.

Figure 2. Major topics of works published in the *Journal of Medical Internet Research* mapped to the United Nations SDGs. AI: artificial intelligence; SDG: sustainable development goal.

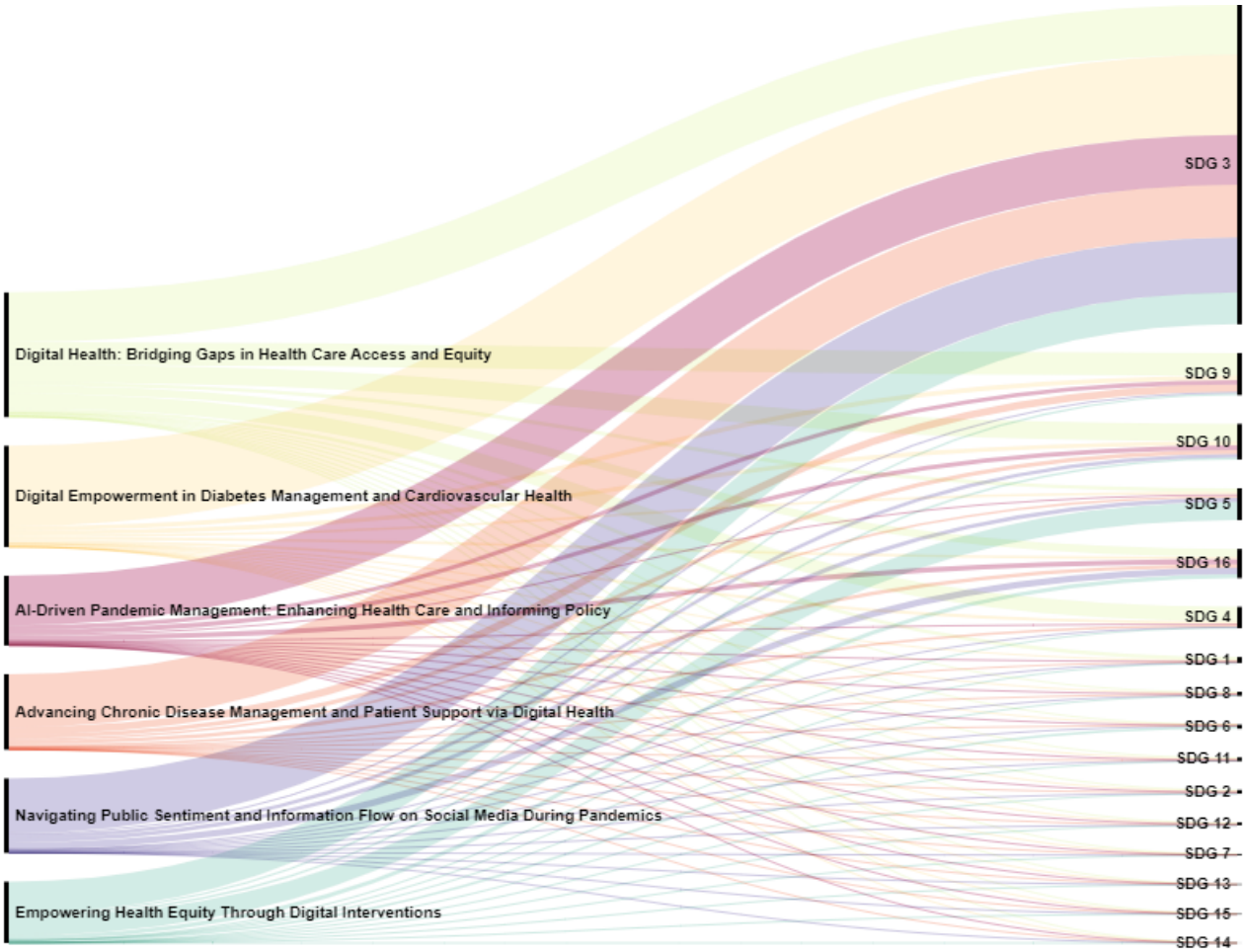

XSL•FO
RenderX

Discussion

Principal Findings

The 6 topics with significant keywords in the *Journal of Medical Internet Research* collectively emphasize the use of digital health to address diverse global health challenges, aligning with several SDGs (Multimedia Appendix 2). The significance of these topics is emphasized in studies such as those on chronic disease management [6,7], artificial intelligence (AI)–driven pandemic management [8,9], health care access [10], diabetes and cardiovascular health management [11], social media impact during pandemics [12], and health equity through digital interventions [13,14]. The topics below outline the diverse ways research from this journal has contributed to the SDGs:

- Advancing chronic disease management and patient support via digital health:
  - Highlights how digital interventions support patients with chronic disease management from diagnosis to recovery, enhancing autonomy and treatment adherence
  - Aligns with SDG 3
- AI-driven pandemic management:
  - Showcases the use of AI in enhancing health care delivery and informing public health policies during pandemics
  - Aligns with SDGs 3 and 9
- Digital health to bridge gaps in health care access and equity:
  - Use of digital health technologies to make health care more accessible and equitable
  - Supports SDGs 10 and 16 by improving access to quality care for underserved populations
- Empowering health equity through digital interventions
  - Centers on reducing health disparities through digital interventions (eg, HIV prevention and health care for marginalized populations)
  - Supports SDG 3.3 in combating communicable diseases
- Digital empowerment in diabetes management and cardiovascular health:
  - Uses digital platforms to improve self-care, clinical outcomes, and health literacy in managing diabetes and cardiovascular diseases
  - Aligns with SDG 3.4 in reducing mortality from noncommunicable diseases
- Navigating public sentiment and information flow on social media during pandemics:
  - Explores how social media can shape public sentiment and spread information during health crises, including how misinformation affects public behavior and governance

Digitization can negatively impact the SDGs. COVID-19 worsened digital inequity, deepening divides in racialized communities and limiting access to essential services [15]. Significant challenges and financial losses in web-based research with underserved populations exist, emphasizing digitalization's negative impact on data quality and resource efficiency [16]. To enhance the contributions of the *Journal of Medical Internet Research* to SDG 4 (Quality Education) and SDG 6 (Clean Water and Sanitation), we recommend prioritizing interdisciplinary research, launching special issues on these topics, forming partnerships with educational and environmental organizations, promoting specific funding opportunities, and organizing workshops.

Limitations and Conclusion

Our study is constrained by the selected database and SDG mapping approach, which may limit its comprehensiveness. Additionally, while BERTopic modeling offers robust analysis, its inherent limitations and the subjective nature of topic interpretation could lead to oversimplifications and biases, necessitating expert review and validation. Despite these limitations, this study underscores the pivotal role of the *Journal of Medical Internet Research* in advancing the UN SDGs.

Data Availability

The dataset generated and analyzed during this study is available in Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dataset generated and analyzed during this study.
[XLSX File (Microsoft Excel File), 5833 KB - jmir_v26i1e60025_app1.xlsx]

Multimedia Appendix 2
Major topics, top keywords, and SDG focus. AI: artificial intelligence; SDG: sustainable development goal.
[PNG File , 108 KB - jmir_v26i1e60025_app2.png]

References

XSL•FO
RenderX

1. Transforming our world: The 2030 agenda for sustainable development. United Nations. URL: https://sdgs.un.org/2030agenda [accessed 2024-10-22]
2. Raman R, Lathabhai H, Pattnaik D, Kumar C, Nedungadi P. Research contribution of bibliometric studies related to sustainable development goals and sustainability. Discov Sustain 2024 Jan 15;5(1). [doi: 10.1007/s43621-024-00182-w]
3. Raman R, Lathabhai H, Mandal S, Kumar C, Nedungadi P. Contribution of business research to sustainable development goals: bibliometrics and science mapping analysis. Sustainability 2023 Aug 29;15(17):12982. [doi: 10.3390/su151712982]
4. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv Preprint posted online Mar 11, 2022 [FREE Full text] [doi: 10.48550/arXiv.2203.05794]
5. Le Blanc D. Towards integration at last? The sustainable development goals as a network of targets. Sustainable Development 2015 Apr 10;23(3):176-187. [doi: 10.1002/sd.1582]
6. Lee T, Sheu S, Chang H, Hung Y, Tseng L, Chou S, et al. Developing a web-based comic for newly diagnosed women with breast cancer: an action research approach. J Med Internet Res 2019 Feb 04;21(2):e10716 [FREE Full text] [doi: 10.2196/10716] [Medline: 30714947]
7. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52 [FREE Full text] [doi: 10.2196/jmir.3951] [Medline: 25803266]
8. Ko H, Chung H, Kang WS, Kim KW, Shin Y, Kang SJ, et al. COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation. J Med Internet Res 2020 Jun 29;22(6):e19569 [FREE Full text] [doi: 10.2196/19569] [Medline: 32568730]
9. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
10. Kontos E, Blake KD, Chou WS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
11. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My Diabetes Coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 05;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]
12. Park S, Suh Y. A comprehensive analysis of COVID-19 vaccine discourse by vaccine brand on Twitter in Korea: topic and sentiment analysis. J Med Internet Res 2023 Jan 31;25:e42623 [FREE Full text] [doi: 10.2196/42623] [Medline: 36603153]
13. Luo Q, Wu Z, Mi G, Xu J, Scott SR. Using HIV risk self-assessment tools to increase HIV testing in men who have sex with men in Beijing, China: app-based randomized controlled trial. J Med Internet Res 2023 Sep 01;25:e45262 [FREE Full text] [doi: 10.2196/45262] [Medline: 37656500]
14. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020 Jun 02;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]
15. Turin TC, Subroto S, Raihan MMH, Koch K, Wiles R, Ruttan E, et al. Identifying challenges, enabling practices, and reviewing existing policies regarding digital equity and digital divide toward smart and healthy cities: protocol for an integrative review. JMIR Res Protoc 2022 Dec 08;11(12):e40068 [FREE Full text] [doi: 10.2196/40068] [Medline: 36480264]
16. Jackson AM, Woo J, Olson M, Dalisay F, Pokhrel P, Muller CJ, et al. Methodological challenges in web-based qualitative research with medically underserved populations. J Med Internet Res 2023 Mar 30;25:e44086 [FREE Full text] [doi: 10.2196/44086] [Medline: 36995748]

Abbreviations

AI: artificial intelligence
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SDG: sustainable development goal
UN: United Nations

XSL•FO
RenderX

Edited by G Eysenbach; submitted 29.04.24; peer-reviewed by S Ibeneme, AK Sahu; comments to author 27.06.24; revised version received 29.06.24; accepted 10.08.24; published 04.11.24.

Please cite as:
Raman R, Singhania M, Nedungadi P
Advancing the United Nations Sustainable Development Goals Through Digital Health Research: 25 Years of Contributions From the Journal of Medical Internet Research
J Med Internet Res 2024;26:e60025
URL: https://www.jmir.org/2024/1/e60025
doi:10.2196/60025
PMID:

©Raghu Raman, Monica Singhania, Prema Nedungadi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Journey From Nonimmersive to Immersive Multiuser Applications in Mental Health Care: Systematic Review

Iveta Fajnerova[1,2], PhD; Lukáš Hejtmánek[1,3], PhD; Michal Sedlák[1], PhD; Markéta Jablonská[1], MA; Anna Francová[1,2], MA; Pavla Stopková[1,2], MD, PhD

[1]Research Center for Virtual Reality in Mental Health and Neuroscience, National Institute of Mental Health, Klecany, Czech Republic

[2]Third Faculty of Medicine, Charles University, Prague, Czech Republic

[3]Faculty of Humanities, Charles University, Prague, Czech Republic

Corresponding Author:
Iveta Fajnerova, PhD
Research Center for Virtual Reality in Mental Health and Neuroscience
National Institute of Mental Health
Topolová 748
Klecany, 250 67
Czech Republic
Phone: 420 283 088 478
Email: Iveta.fajnerova@nudz.cz

Abstract

Background: Over the past 25 years, the development of multiuser applications has seen considerable advancements and challenges. The technological development in this field has emerged from simple chat rooms through videoconferencing tools to the creation of complex, interactive, and often multisensory virtual worlds. These multiuser technologies have gradually found their way into mental health care, where they are used in both dyadic counseling and group interventions. However, some limitations in hardware capabilities, user experience designs, and scalability may have hindered the effectiveness of these applications.

Objective: This systematic review aims at summarizing the progress made and the potential future directions in this field while evaluating various factors and perspectives relevant to remote multiuser interventions.

Methods: The systematic review was performed based on a Web of Science and PubMed database search covering articles in English, published from January 1999 to March 2024, related to multiuser mental health interventions. Several inclusion and exclusion criteria were determined before and during the records screening process, which was performed in several steps.

Results: We identified 49 records exploring multiuser applications in mental health care, ranging from text-based interventions to interventions set in fully immersive environments. The number of publications exploring this topic has been growing since 2015, with a large increase during the COVID-19 pandemic. Most digital interventions were delivered in the form of videoconferencing, with only a few implementing immersive environments. The studies used professional or peer-supported group interventions or a combination of both approaches. The research studies targeted diverse groups and topics, from nursing mothers to psychiatric disorders or various minority groups. Most group sessions occurred weekly, or in the case of the peer-support groups, often with a flexible schedule.

Conclusions: We identified many benefits to multiuser digital interventions for mental health care. These approaches provide distributed, always available, and affordable peer support that can be used to deliver necessary help to people living outside of areas where in-person interventions are easily available. While immersive virtual environments have become a common tool in many areas of psychiatric care, such as exposure therapy, our results suggest that this technology in multiuser settings is still in its early stages. Most identified studies investigated mainstream technologies, such as videoconferencing or text-based support, substituting the immersive experience for convenience and ease of use. While many studies discuss useful features of virtual environments in group interventions, such as anonymity or stronger engagement with the group, we discuss persisting issues with these technologies, which currently prevent their full adoption.

(J Med Internet Res 2024;26:e60441) doi:10.2196/60441

XSL•FO
RenderX

KEYWORDS

digital health; mental health care; clinical interventions; multiuser; immersive; virtual reality; VR; app; mental health; online tools; synthesis; mobile phone; PRISMA

Introduction

Background

Thousands of studies are exploring self-care applications, which deliver immediate on-demand psychological help to people otherwise waiting on in-person therapy [1,2]. There are a multitude of research projects aimed at exploring the efficiency of online tools or virtual environments for addressing anxiety [2] or affective disorders [3]. However, most of the current investigation into technologies and mental health care is focused on single-user experiences [4-6].

One of the determining factors of effective psychotherapy is establishing a functional relationship between the therapist and the client [7]. In the case of group interventions, the establishment of mutual communication between clients in the group plays a crucial role [8]. However, while technological advances have modernized, simplified, and increased the availability of therapeutic interventions to people who might have had difficulties accessing them in the past, these tools are largely solitary. Meanwhile, multiuser tools and virtual environments could provide useful intervention programs for a therapeutic dyad or a whole group (3 or more participants [9,10]).

There are a number of social groups meeting and supporting each other in various areas related to mental disorders and mental health, such as postnatal depression, addiction, or suicidal tendencies [11-14]. People search for peer support not only on social media but also in different virtual environments, where they form dedicated groups. Such support groups have appeared in many popular video games, such as World of Warcraft (Blizzard Entertainment), Second Life (Linden Research, Inc), or Minecraft (Mojang studios) [15,16] or recently in virtual reality (VR) enabled social applications, especially VRChat (VRChat, Inc) [17-20]. Currently, we can already recognize a whole range of commercial VR applications designed as platforms for mutual communication and interaction in virtual environments, often referred to as the *metaverse* [21], such as Meta Horizon workrooms (Meta Inc) [22], Rec room (Meta Inc) [23], or Glue (Meta Inc) [24].

This Review

This review aims to focus on the applications and procedures of mental health care interventions that use such multiuser platforms, where professionals or peers work with clients side by side and which use technology as a means to connect, not to substitute the human contact (ie, self-care apps or artificial intelligence (AI) chatbots [25,26]).

We set off to answer the following questions: (1) what has changed in the field of remote group therapy in the past 25 years, (2) what technologies have been tested and deemed functional, (3) which groups seem to benefit most from such interventions, and (4) what new platforms, including immersive VR, have to offer with regard to future mental health support.

Group Therapy and Its Role in Mental Health Care

Group psychotherapy or group therapy is a form of psychotherapy that involves treating a small number of clients under the supervision of one or more therapists [27]. A broader concept of group therapy refers also to support groups for people with a variety of mental health conditions usually led by peers aimed at psychoeducation and skills training [28].

Group psychotherapy is based not only on the interaction of therapists with a group of patients but also on the interactions between patients or clients themselves [29]. In the original concept of psychodynamic and interpersonal group therapy, the group dynamics, that is, relationships and interactions between members and the therapist, are used for therapeutic purposes [29,30]. The group context and group process are considered the main mechanisms of change by developing, exploring, and examining interpersonal relationships within the group.

Group psychotherapy offers several advantages over an individual therapy format [31]. First, more patients can reach the treatment at the same time, at a reasonable cost, and one therapist may interact with several patients at the same time. This is highly important given the fact that in some countries, only a small number of psychiatric patients actually receive adequate psychological care [32].

Second, the experience of being a part of the group itself, sharing personal experiences, and meeting other people with similar mental health problems can have therapeutic effects [33], as is the case of self-care or support groups. Other treatment factors specific to group therapy that emerge from the interpersonal setting include feeling connected to the world and being respected and valued by others [34]. Other specific treatment factors were proposed, including vicarious and interpersonal learning, social support, experiencing universality, altruism, fostering hope, and a sense of belonging and relatedness [35]. Group therapy may therefore provide a source of corrective relational experiences [36,37]. Some studies have shown better ratings of group therapeutic factors during ongoing treatment, as well as a relationship to treatment outcome for patients with anxiety disorders, although systematic evaluation of these factors is inadequate [38-40]. Therapeutic factors and group processes in online groups appear, thus far, to be the same as those in traditional group interventions [41,42].

Needless to say, there are disadvantages to group therapy. Participants must be willing to share their personal experiences and issues. Some patients consider this format so challenging that it can prevent their engagement in the group, which becomes an obstacle to successful treatment [43]. The group setting also provides less time and opportunities to deal with individual topics, and some patients may feel overlooked [44]. Organization of group meetings is more difficult and managing the therapeutic group requires additional skills and specific training from the therapist compared with individual therapy [43].

XSL•FO
RenderX

In the context of cognitive behavioral therapy (CBT), the effectiveness of group therapy is well-supported [45]. The results of a meta-analytic review indicated that anxiety prevention group programs can be a promising strategy for reducing the incidence rates of anxiety disorders [46]. Researchers have not found substantial differences between group and individual CBT for treatment of anxiety disorders in adults [47,48], as well as in youth [49,50]. Group CBT was also superior to waitlist control groups or produced equivalent results when compared with other active treatments, that is, individual psychotherapy and pharmacotherapy [47,51,52].

Online Communication in Health Care: From Chat Rooms to Virtual Worlds

Digital tools have considerably affected how people manage their mental health. The accessibility of online social support has been gradually rising due to free social media and communication platforms [53,54]. People look on the web for information and knowledge about their condition, but they also seek guidance and support from peers and professionals [53,55]. Over the past 25 years, there has been a remarkable evolution in how people connect and communicate in online spaces in general. The rapid technological development is clearly visible in the transition from text-based chat rooms to videoconferencing, which represents a substantial shift in the way we interact with others on the web. The recent emergence of "virtual worlds" and immersive virtual devices provides another crucial milestone in digital interactions. So how have these technologies been used in mental health care over the years?

In the late 1990s and early 2000s, online communication primarily revolved around text–based chat rooms. These platforms allowed users to engage in real time conversations with others who shared similar interests or demographics. However, interactions in chat rooms were limited to text, and there was minimal visual or auditory engagement. This approach has been mostly used in the form of peer-support groups and peer-to-peer communities, mostly addressing depression and social support, in general [56]. Even now, chat rooms are extensively used in various forms. Social networking sites have gained massive user bases, presenting an opportunity to deliver internet-based mental health interventions to many people. A review covering the application of social networking sites in mental health care [57] pointed out that such an approach, mostly aimed at mental health literacy or specific symptoms (depression), shows high acceptability and engagement. Despite promising results, there is a lack of high-quality evidence supporting its effectiveness.

Voice calls became more prevalent in online spaces during the mid-2000s. A voice added a new dimension to online interactions, enhancing the sense of presence and immediacy and allowing for certain nonverbal cues (eg, voice intonation, speech speed, and pauses). An even bigger improvement came in the late 2000s when videoconferencing (eg, Skype, Zoom, and Microsoft Teams) became available. Videoconferencing provides a more immersive experience, by allowing participants to see each other's facial expressions, gestures, and partially, their body language, all in real time. Videoconferencing is also simple to use, and therefore, it has become a popular platform in remote health care group interactions. A systematic review of home-based support groups delivered via videoconferencing showed that this approach is feasible, as it can replicate group processes, such as bonding or cohesiveness with outcomes similar to in-person groups [58]. It demonstrated the ability to engage with people with similar problems; improve accessibility to peer groups; and development of health-related knowledge, skills, and insights provided during the intervention.

Another essential advancement came with the development of "virtual worlds" or virtual environments. These were originally designed as games (eg, Second Life, World of Warcraft, and Minecraft), but at its core, they can be seen as social platforms as well. Besides the often implemented voice and chat-based messaging, these worlds also allow for interpersonal interactions, such as approaching somebody or avoiding them, exchanging items, or emoting. Although people usually start playing for fun, these games have a lot of downtime or routine work that does not require constant attention, which invites socializing with other players. For these reasons, virtual worlds provide a well-featured place where support groups can organically appear [16].

In recent years, a range of extended reality technologies have gained traction, offering even more immersive and interactive experiences in virtual worlds. These immersive technologies include VR, augmented reality (AR), and mixed reality (MR). AR and MR lenses like Pokémon GO and Snapchat overlay digital content onto the real world, blending virtual and physical experiences. Immersive VR headsets (head-mounted displays [HMDs]) like HTC Vive and Meta (Oculus) Quest allow users to fully enter virtual worlds and interact with the environment and other people as well. Virtual worlds (eg, metaverse) and immersive headsets offer novel opportunities for delivering group–based mental health interventions. These technologies can simulate real-life social interactions, provide engaging and interactive experiences, and enhance the sense of presence and connection among participants [59,60].

The importance of internet-based multiuser tools and platforms for health care has been enhanced during the COVID-19 pandemic (beginning to emerge worldwide in 2020), which has exacerbated social isolation and mental health challenges [61], particularly among marginalized communities (eg, older mothers of newborns, cultural minorities, and people with various psychiatric disorders). These groups often face unique challenges that contribute to feelings of isolation, including limited access to resources, societal stigma, and changes in social support networks. Therefore, we assume that the occurrence of tools supporting social communication in mental health care during COVID-19 has increased rapidly during the pandemic [62,63], as well as the number of extended reality applications focused on mental health in general.

Objectives of the Systematic Review

This systematic review aimed to investigate the use of online multiuser interventions to support mental health that facilitate communication in dyadic and group interventions in the last 25 years. The main objective was to systematically explore and interpret evidence about these types of interventions, specifically

XSL•FO
RenderX

about methods and systems (multiuser technologies) used. This reflects the absence of review studies that look into this phenomenon comprehensively and not only with respect to the selected technology (eg, chat rooms). The second objective was to evaluate acceptability, suitability, and safety of multiuser technologies, and identify potential gaps and opportunities for future research. Specifically, this review aimed to identify potential benefits and disadvantages of the available multiuser technologies, especially with regard to their use in therapeutic or peer-supported group interventions. Furthermore, our aim was to understand whether the potential of immersive technologies has manifested itself in an increase in studies using VR or AR multiuser interventions, and what barriers might prevent it.

Methods

Review Guidelines

This systematic review includes primary sources related to the use of multiuser interventions in the context of mental health care. It follows the principles of the updated PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guideline [64,65].

Inclusion and Exclusion Criteria

Several inclusion and exclusion criteria were determined before and during the records screening process. The database search covered articles in English, published from January 1999 to March 2024, related to multiuser mental health interventions. For the identification of records, the inclusion criteria were established as follows: (1) the record should be related to mental health care, (2) the intervention should be conducted in 2 or more people, and (3) the intervention itself should be administered or mediated via digital technology. Opinion papers, research protocols, review papers, and papers without original research were excluded.

Search Strategy

Two academic database sources, Web of Science and PubMed, were used for the systematic review search using the search queries listed in Table 1.

Both queries performed the search in the title, abstract, and keywords of the database records. Specifying keywords in the queries for the term “multiuser” was crucial to incorporate studies that used inconsistent terminology. By adding a wide range of alternative keywords, such as, “multi patient,” “dyadic,” or “group therapy,” the number of results from the databases dramatically increased. The export from the databases was conducted on April 09, 2024. All records were screened based on the predefined inclusion criteria.

Table 1. Database search queries for records identification. YP: Year published; TS: Topic.

Database	Search queries
Web of Science	PY=(1999-2024) AND TS=((virtual OR digital OR app OR “computer program”) AND (mental OR psychological OR psychiatric OR depress* OR anxiety) AND (therapy OR teletherapy OR psychotherapy OR intervention OR treatment) AND (multi-user OR multiuser OR “multi user” OR multi-patient OR multipatient OR “multi patient” OR multi-participant OR multiparticipant OR “multi participant” OR multi-respondent OR multirespondent OR “multi respondent” OR multi-proband OR multiproband OR “multi proband” OR dyadic OR triadic OR collaborative OR cooperative OR “group therap*” OR “group teletherap*” OR “group intervention*” OR “group treatment*” OR “group support*” OR “therapy group*” OR “teletherapy group*” OR “intervention group*” OR “treatment group*” OR “support group*”))
PubMed	(1999:2024[dp]) AND (virtual[Title/Abstract] OR digital[Title/Abstract] OR app[Title/Abstract] OR “computer program”[Title/Abstract]) AND (mental[Title/Abstract] OR psychological[Title/Abstract] OR psychiatric[Title/Abstract] OR depress*[Title/Abstract] OR anxiety[Title/Abstract]) AND (therapy[Title/Abstract] OR teletherapy[Title/Abstract] OR psychotherapy[Title/Abstract] OR intervention[Title/Abstract] OR treatment[Title/Abstract]) AND (multi-user[Title/Abstract] OR multiuser[Title/Abstract] OR “multi user”[Title/Abstract] OR multi-patient[Title/Abstract] OR multipatient[Title/Abstract] OR “multi patient”[Title/Abstract] OR multi-participant[Title/Abstract] OR multiparticipant[Title/Abstract] OR “multi participant”[Title/Abstract] OR multi-respondent[Title/Abstract] OR multirespondent[Title/Abstract] OR “multi respondent”[Title/Abstract] OR multi-proband[Title/Abstract] OR multiproband[Title/Abstract] OR “multi proband”[Title/Abstract] OR dyadic[Title/Abstract] OR triadic[Title/Abstract] OR collaborative[Title/Abstract] OR cooperative[Title/Abstract] OR “group therap*”[Title/Abstract] OR “group teletherap*”[Title/Abstract] OR “group intervention*”[Title/Abstract] OR “group treatment*”[Title/Abstract] OR “group support*”[Title/Abstract] OR “therapy group*”[Title/Abstract] OR “teletherapy group*”[Title/Abstract] OR “intervention group*”[Title/Abstract] OR “treatment group*”[Title/Abstract] OR “support group*”[Title/Abstract])

Data Selection Process and Records Eligibility

Out of 2679 records obtained from the databases, 1540 (57.48%) were from Web of Science and 1139 (42.52%) were from PubMed. These were processed by automation tools; 987 (36.84%) were identified as duplicates and 56 (2.09%) were marked as ineligible as they were incomplete or did not fit the publication type criteria (eg, poster abstracts). This resulted in 1636 (61.07%) records being moved from the identification phase to the screening phase.

The screening phase consisted of 3 manually conducted screenings with the help of a self-hosted NocoDB database viewer [66]. The records were assessed by 5 reviewers consisting of experienced scientific researchers of our team. In screenings 1 and 2, the reviewers read through the abstracts of all records and tagged them with predefined categories based on the inclusion and exclusion criteria. Screening 3 consisted of reading and assessing full-texts of the records.

Screening 1 started with 1636 (61.07%) records, and each record was assessed by a 1 reviewer (from 5 reviewers in total). The

XSL•FO
RenderX

listed reasons for excluding the record were: (1) if it was not aimed at mental health (eg, focused on the use of technologies in physical therapy), (2) it did not include any multiuser experiences (eg, use of self-care apps), (3) it was not delivered through digital mediums (eg, digital medium used for scheduling sessions or data recording, not for therapy), (4) it was discovered to be a duplicate, (5) it was a protocol or an opinion without data or (6) other, unspecified reasons. This process resulted in the exclusion of 1310 (48.9%) records.

Screening 2 started with 326 (12.17%) records and each record was assessed by multiple reviewers to achieve objectivity. Every record that was tagged by at least 3 reviewers as not matching inclusion criteria was excluded. The exclusion criteria were similar to those in screening 1. However, to limit the number of records, we decided to exclude records that dominantly focused on mental health care using internet-based tools in response to the COVID-19 pandemic in healthy populations. We also excluded records focusing solely on secondary caregivers (eg, peer support groups for parents of psychiatric patients), and interventions using AI chatbots. This resulted in 171 (6.38%) records being excluded.

There was an intermediate step in the screening process. It started with 155 (5.79%) records and consisted of filtering out the records that were tagged as being reviews during screening 1 and 2. This resulted in the exclusion of 83 (3.1%) records.

Screening 3 started with 72 (2.69%) records. In this phase, every author focused on extracting relevant information through independent full-text reviews. During this process, another 23 (0.86%) records were excluded, the most common reason being that the intervention was not in fact multiuser or that the record only described the opinions or insights of the authors.

In total, 49 (1.83%) records were considered eligible and included in our systematic review. The full-texts of these records were read and summarized, each by a single researcher, and information about procedures and study designs, means of intervention administration (technology, hardware), target groups, sample sizes, outcome measures, and other parameters were extracted.

Results

Standardization of Terminology

The performed systematic review had to overcome considerable complications related to unclearly defined or not established terminology regarding multiuser online tools. Studies using these technologies adopted very broad terminology based on the context of internet-based platforms used, often referring to virtual environments ranging from simple web-based forums to virtual worlds. We decided to distinguish these technologies as chat rooms, voice calls, videoconferences, and virtual worlds. The virtual worlds include both low-immersive desktop use to immersive visualization using VR headsets. A much bigger challenge was the search for applications that are designed for use by multiple users, allowing communication and interaction. Here, the terminology was quite diverse, with the term multiuser appearing only in a few cases. Some of the applications refer to communication or collective applications, but most of the records simply referred to group interventions or therapies, which alone made the search strategy very difficult, as the term “group” is often used in connection with the research method (experimental, therapeutic, or control group). Thus, the search required a very laborious elimination of all false-positively selected articles that did not address any group activities. This inconsistency in terminology regarding the term “multiuser” creates a potential negative selection bias when searching for records in the scientific body of knowledge. Despite the elaborated query incorporating many possible variations of the terms used for multiuser technologies, some records could be potentially missed if the research teams did not use all the terms included in our search query.

Multiuser Interventions in Mental Health Care

An overview of the records selection process is documented in the PRISMA 2020 flow diagram (Figure 1). The systematic search led to 49 (1.83%) records that present studies aimed at dyadic and group interventions. Even though we covered 25 years of progress, it was apparent that most of the published works appeared after 2015 (only 2 records appeared before). This suggests that even though chat rooms and videoconferencing were available for more than a decade sooner, they did not find their way to clinical care as fast as could be expected.

As mentioned before, we were interested both in dyadic interventions, allowing communication between the patient (client) and the professional (expert), or peer and group interventions, allowing interaction of 3 or more participants. Out of 49 studies, our search identified 10 (20%) studies that applied purely dyadic interventions, 36 (73%) studies used purely group interventions, and 3 (6%) studies combined both approaches. For practical purposes, henceforth, we will refer only to 2 categories—“group interventions” (consisting of purely group interventions and interventions combining group and dyadic approaches), and “dyadic interventions” (consisting of purely dyadic interventions). A complete list of studies using group interventions [9, 67-104] are presented in Table S1 in Multimedia Appendix 1, while the list of dyadic interventions [10, 105-113] are presented in Table S2 in Multimedia Appendix 1.

In the next subsections, we analyzed the data listed in Tables S1 and S2 in Multimedia Appendix 1 in greater detail to show technology trends in the use of multiuser interventions in mental health care. All the descriptive statistics are calculated from the entirety of 49 studies, adding up to 100%.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for the overview of the selection process.

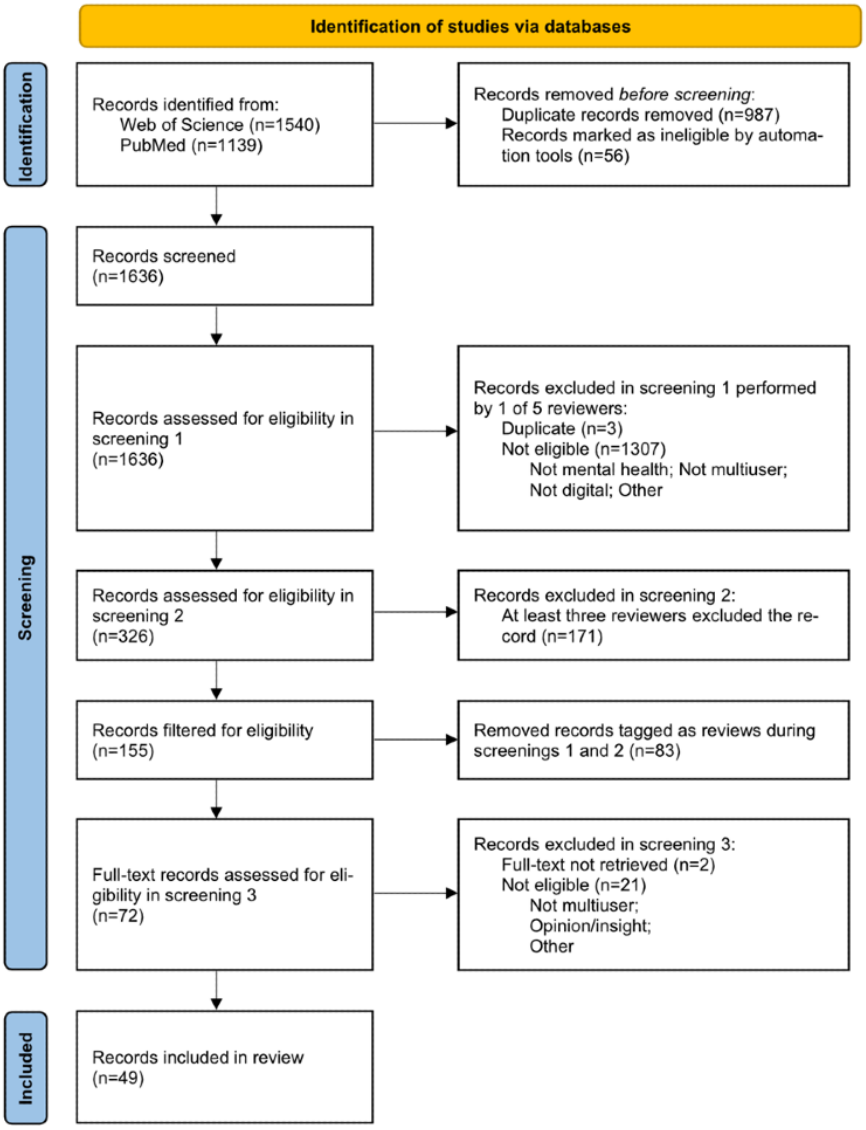

Hardware and Software Used in the Studies

In terms of the hardware devices, most of the multiuser interventions used mobile phones (14/49, 29%), computer monitors (11/49, 22%), or their combination (8/49, 16%). Some of the studies did not specify devices used by the target population (13/49, 27%). Importantly, out of 49 studies, only 3 (6%) studies used HMDs (VR headsets). The comparison of the hardware used in the group and dyadic interventions is shown in Figure 2.

In terms of technological tools, the reviewed studies used various multiuser platforms. Some of the studies used only one technology while others used a combination of technologies. For cases where only one technology was used, chat rooms (text form of communication, social network groups, and forums) were used in 17 of 49 (35%) studies, voice calls were not used, videoconferences were used in 13 (27%) studies, and the most advanced technology of virtual worlds was used only in 5 (10%) studies. In the other 14 (29%) studies, the combination of multiple technological tools was used. The comparison of the multiuser technology used in the group and dyadic interventions is shown in Figure 3.

XSL•FO
RenderX

Figure 2. Hardware used in reviewed studies.

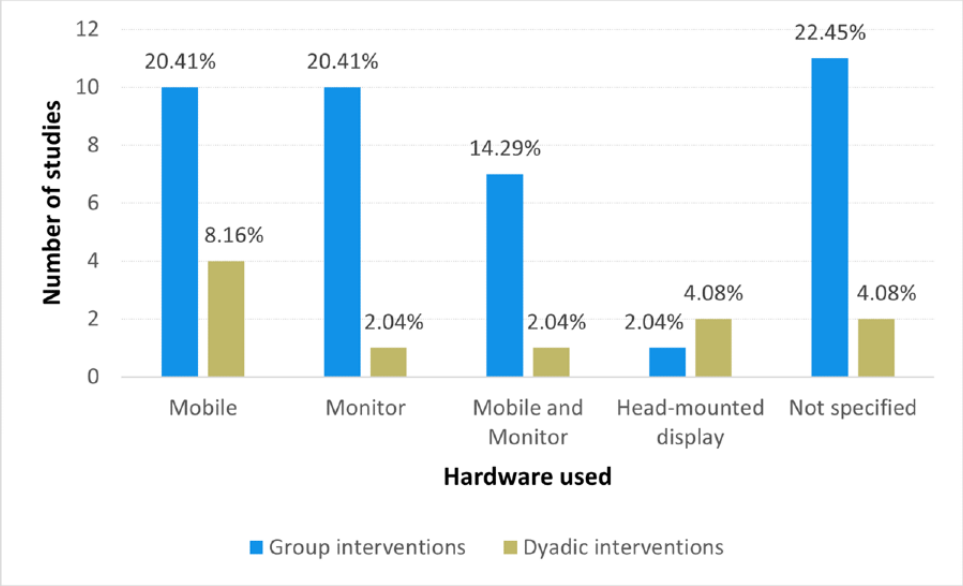

Figure 3. Multiuser technology used in reviewed studies.

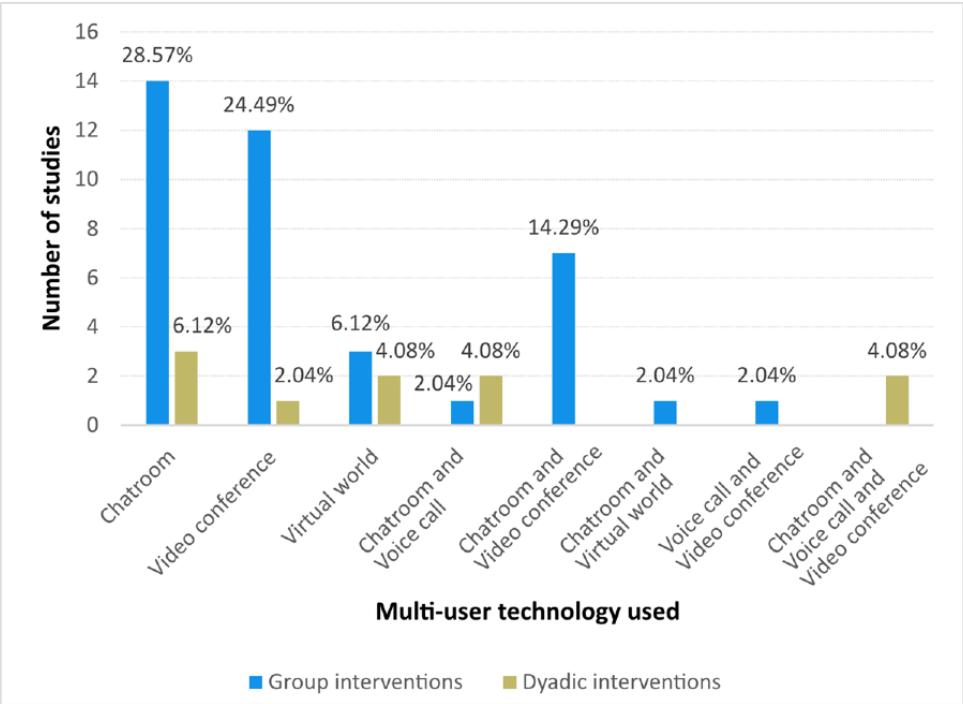

Type of Support Provided in the Studies

When comparing who moderated the interventions, the studies usually used trained professionals to lead the interventions (23/49, 47%), peer support was used in a few studies (11/49, 22%), while other studies used a combined approach (15/49, 31%). The comparison of the interpersonal source of support in the group and dyadic interventions is shown in Figure 4.

In terms of frequency, the digital interventions allowed variable (both synchronous and asynchronous) communication, combining multiuser interventions with self-care mobile health apps in some cases. Therefore, we refer only to the frequency of multiuser interventions specifically. The interventions were conducted either daily (6/49, 12%), weekly (24/49, 49%), or in an unscheduled or asynchronous manner allowing the participants to connect at any time—mostly in cases of support chats (18/49, 37%). In addition, there was 1 study that used the group intervention only once (1/49, 2%). The comparison of the meeting frequency in the group and dyadic interventions is shown in Figure 5.

XSL•FO
RenderX

Figure 4. Interpersonal source of support in reviewed studies.

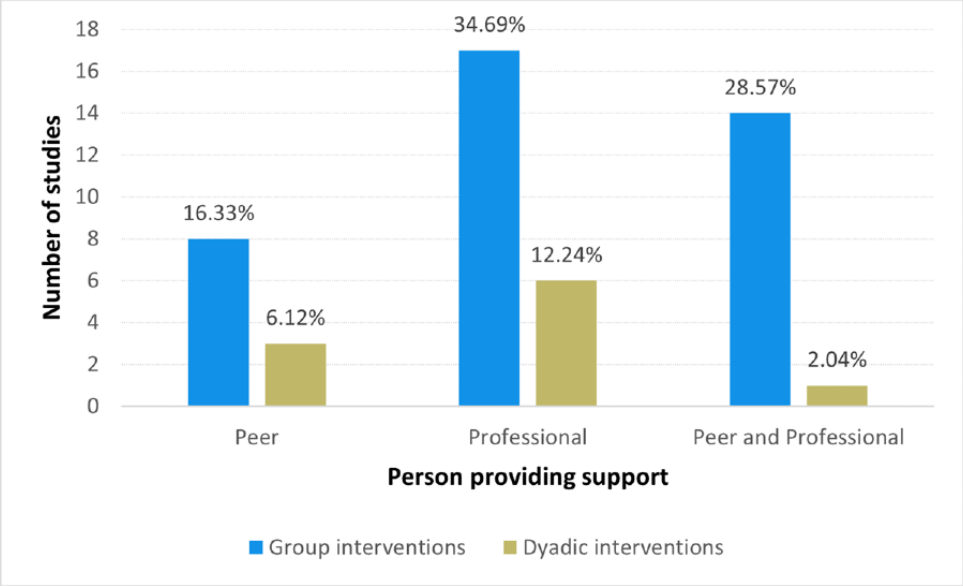

Figure 5. Meeting frequency in reviewed studies.

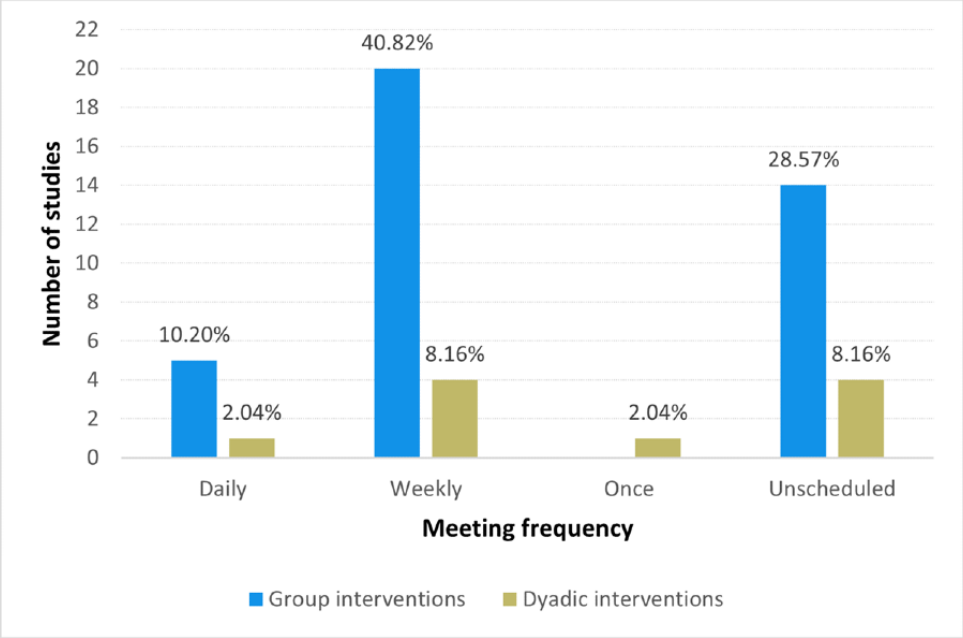

Clinical Methods and Outcome Measures

Most of the studies were aimed at support groups and stress reduction in different age and target groups. The portion of peer-supported and professional-supported intervention groups is described in detail in the previous paragraph and Figure 4. The interventions that were led by professionals typically involved clinicians, psychotherapists, facilitators, or coaches. The interventions that allowed peer support usually combined this approach with a professional-led component of the intervention.

When led by professionals, the reviewed interventions typically applied CBT approach adapted for online group therapy or other approaches like dialectic behavioral therapy, art therapy, or hypnosis. The peer support groups typically focused on parenting, addiction, and stress reduction methods.

The study designs used in the reviewed records were mostly randomized controlled trials (17/49, 35%), pilot or feasibility studies (10/49, 20%), quasi-experimental studies (8/49, 16%), and qualitative research (9/49, 18%). Out of 49 studies, 2 (4%) studies combined quantitative and qualitative approaches in mixed methods design, and 3 (6%) other studies were in the form of a report, an exploratory study, and an open trial.

The reviewed studies focused on various categories of target populations (similar for group and dyadic interventions) while addressing topics relevant to them. These prevalently included pregnant women or nursing mothers (covering topics such as parenting, self-esteem, or self-efficacy); adolescents, young adults, and university students (covering topics such as stress prevention and reduction); patients with eating disorders, depression, or post-traumatic stress disorder (psychoeducation and symptom reduction); and patients with addiction disorders

XSL•FO
RenderX

(eg, smoking cessation). The less represented target groups included minorities (eg, lesbian, gay, bisexual, transgender, intersex, queer or questioning, asexual [LGBTQA+]), patients with psychotic disorders (recovery), and children with autism spectrum disorders (training of social interactions).

The sample size across studies ranged from 3 to hundreds of participants depending on the type of the study (qualitative vs quantitative). The size of different intervention groups also varied, but in most studies, the groups led by a professional included 2 to 10 participants present synchronously. When the intervention was asynchronous, or only peer support was provided, the group size was not always specified or corresponded to the size of the whole sample.

Regarding the outcome measures used, most group and dyadic studies used scales for depression and anxiety, social support, self-esteem and coping, and a variety of stress measures (including physiological markers), and health scales. Some of the studies used additional (or as a main tool in the case of qualitative studies) qualitative interviews with participants. Feasibility studies often included measures of engagement and acceptability of the tested interventions (user experience).

In terms of the clinical effectiveness of the tested digital multiuser interventions, the randomized controlled trials supported the usefulness of these interventions, reporting overall positive results. When compared with waitlists or minimal care, the interventions showed substantial symptom reduction (eg, depression and anxiety) using videoconferencing on mobile phones or computer monitors [74,107,108]; using chat rooms (on computers) [111]; or using a combination of a chat room, voice call, and videoconference either using mobile phones [108] or unspecified hardware [110]. The records showed comparable effects of multiuser interventions in response to standard procedures in a study using videoconference (on computer monitors) [78] and in a study using chat rooms and voice calls (on mobile phones) [105]. Long-term benefits were observed in 3 studies using chat rooms either on mobile phones [76], mobile phones and computers [101], or unspecified hardware [106]. However, some studies reported only partial symptom reduction. This was the case across studies using various technologies, ranging from chat rooms on mobile phones [83], voice calls and videoconferencing on mobile phones [91], and virtual worlds visualized on computer monitors [52,94]. Few studies using chat rooms were unable to confirm effects in the targeted outcomes due to missing or limited data when used on computer monitors [86], mobile phones [97], or using unspecified hardware [95].

Discussion

Principal Findings

New technologies allowing multiuser communication and interaction offer innovative solutions for delivering mental health care remotely. They play a crucial role in addressing social isolation and mental health disparities by providing accessible, culturally sensitive, and innovative group-based mental health care to be conducted virtually while overcoming barriers related to geographical distance and mobility.

Overall, the technological innovations leading to a transition from chat rooms (including social network sites) to voice calls, videoconferences, and later to immersive virtual environments, reflect the ongoing evolution of internet-based communication technologies, driven by advancements in internet connectivity, software development, and user demands for richer and more engaging virtual interactions with other humans. This is also reflected in the use of multiuser technologies in clinical research focused on mental health reviewed in this paper.

Benefits of Digital Multiuser Interventions

The reviewed technologies enable various forms of mental support, ranging from general peer support groups and peer-led interventions to therapist-led group sessions that provide a convenient and accessible alternative to traditional in-person services. Interventions using freely available web-based tools or existing mobile health applications typically require fewer resources than in-person meetings, making them more affordable for organizations and participants.

Online therapy meetings also eliminate geographical barriers, allowing people from different locations to connect easily. Participants can join group sessions from the comfort of their homes, making it more convenient for patients while eliminating physical transportation barriers [84,94,102] (eg, rural environments, older people, and those with a physical disability), overcoming other limitations and stressors created by the mental conditions that should be addressed in the therapy (eg, agoraphobia or social phobia). In some studies, participants provided qualitative feedback and one of the advantages mentioned was that digital communication helps them overcome the fear and stigma of judgment when discussing mental health issues and other sensitive topics. They also perceived the digital platform as facilitating personal reflection [67]. It was also suggested that online group-based interventions and support groups foster community and peer support, allowing participants to connect with others who share similar experiences and challenges. Peer-led support online groups can be especially beneficial, providing a safe space to share experiences, exchange coping strategies, and receive peer validation and encouragement, as needed.

Moreover, online platforms can be designed to meet the cultural and linguistic needs of minority populations, offering support groups and therapy sessions facilitated by professionals who understand their unique experiences and backgrounds. Culturally sensitive interventions [75,89,90] can help reduce stigma and increase engagement in mental health care among marginalized communities.

All these factors explain why many research studies report good acceptability and usability [67,84,85] of multiuser technologies in the target populations, potentially increasing the adherence of patients. By leveraging these technologies, mental health care providers can bridge the gap in access to services and support the well-being of many underserved populations.

Disadvantages of Digital Multiuser Interventions

There are several disadvantages to digital interventions. These are related to the unwanted effect that technology might have on participants, but there are also unique issues tied to their

XSL•FO
RenderX

development and use by health care providers. Some study participants expressed that they would have preferred face-to-face meetings [84,94] and considered virtual settings to be uncomfortable. Indeed, online group interventions are not the same as face-to-face meetings. They present some specific obstacles and challenges that should be compensated for or acknowledged as weaknesses [61].

The most crucial factor undermining the therapeutic efficiency of online interventions could be the absence of body-to-body interaction [61] and limited nonverbal communication. Internet-based meetings often lack the richness of face-to-face interactions, and participants miss out on many nonverbal cues, such as body language and facial expressions (especially in chat rooms and voice calls).

Participants also highlighted various technical issues with the hardware or software. Connectivity problems (network or phone access issues), unclear audio or video, audio and video glitches, and other technical difficulties can easily disrupt the flow of online meetings and hinder effective communication [85].

As the intervention can be carried out from anywhere, it lacks a controlled environment, so background noises often interrupt the meetings [85]. Furthermore, as the meetings happen often at the patients' homes, they may be more prone to everyday distractions, not focusing fully on the meeting, leading to decreased engagement. Overall, building rapport and fostering a sense of togetherness can be more challenging in virtual spaces compared with in-person interactions.

There are also serious security concerns regarding any treatment happening online, as internet-based meetings may be vulnerable to privacy breaches and security threats, especially when using commercial software without adequate safeguards.

It is also important to add that these technologies, especially the immersive ones, require a new set of skills from health care workers. The therapist must be knowledgeable about the software, potential issues, and how to solve them, as having IT support might not be possible during private sessions.

Internet-based chats, voice calls, and videoconferencing have become so common that it could be expected that each health care worker and patient can use or can learn to use them. However, setting up immersive devices and accessing virtual worlds can be much more challenging for new users not familiar with this technology. This can lead to the patients focusing more on the technology than the session itself.

Finally, there are considerable problems to overcome with regard to developing or hosting one's own multiuser platform. Using existing solutions, such as social platforms, dedicated software, games, or videoconferencing can be costly, but it also raises questions about privacy, data collection, and data retention, which need to be discussed. These existing technologies could be more suited for peer support groups with no affiliation to the health care providers.

For these reasons, many of the reviewed studies have built new applications and solutions, but such development often requires initial and ongoing investments in terms of software development and the cost of maintaining servers. Any application, whether on a mobile phone or in VR, which allows multiuser engagement, requires a backend service to transfer data between users constantly. The applications must also be regularly updated to abide by new security standards or with new browsers and operating systems being released. All this can be a complicated process that is not easily implemented within the organizational structure of health care institutions.

Increasing Immersion in Multiuser Interventions

It seems that multiuser digital interactions are slowly making their way into mental health care. This is apparent from our findings, which demonstrates that over the years the number of studies using these technologies is increasing. We have observed that chat rooms and videoconferencing still represent the most prevalent methods used in mental health care performed in groups, at least represented in research studies. The popularity of chat rooms is probably due to their low barrier of entry and because they allow both synchronous and asynchronous communication. This makes them especially useful in design protocols where self-care methods and psychoeducation tools are combined with on-demand support from peers or professionals [99].

However, purely text-based online support does not offer many important aspects of in-person group interventions, such as speech tonality or facial expressions. These can be achieved using videoconferencing software, which has become very popular in health care. Participants can hear and see each other and react to a multitude of nonverbal cues. It allows participants to freely express themselves, which may increase interactions and invite active participation. As videoconferencing is typically synchronous, it also promotes a sense of presence in the group. For these reasons, along with relative ease of use, videoconferencing seems to be highly prevalent in group therapeutic settings.

By contrast, videoconferencing is still just a series of individual camera streams and lacks the spatial and interactive aspects of human communication (eg, we cannot look directly at the person we are talking to in the group).

In contrast, multiuser virtual world environments enable us to create a common space that the participants share in the same moment, allowing for greater engagement or more natural conversations. Virtual worlds offer more naturalistic social interactions. People can move toward or away from one another, and they can look at each other while talking. The software can incorporate hand gestures or full-body tracking. However, immersive visualization of the avatar's body might present an obstacle if not correctly synchronized with the participant's body movements. This can be a problem for full-body avatars as feet are generally not tracked by commercial HMDs. It has been shown that half-body avatars with head and hand tracking present a good option for social interactions and are able to induce good (co)presence in the multiuser VR environment [109,114].

Some current VR headsets, such as Meta Quest Pro or Vision Pro, also allow recognition of facial expressions and eye movements (gaze direction and blinking), which are then translated to the avatar.

XSL•FO
RenderX

These immersive features have been shown to enhance a sense of social presence [100] or copresence [109]. Participants reported feeling like being in the same room with others, even though they met only in a virtual environment [84,100]. This may potentially increase the feeling of togetherness and thus provide an experience similar to the one provided in face-to-face group communication.

In addition, VR worlds (eg, Second Life [115] or Innerworld [116]) allow participants to customize their avatars, fostering their creativity and self-expression that could occur during in-person interactions [102], while guaranteeing anonymity, if needed. This customization allows participants to stay fully anonymous by selecting a generic avatar and even changing their real voice to an artificial one. On the contrary, they can convey selected personal aspects about themselves, such as sex, race, or age through the avatar's physical features and clothing. Therefore, the virtual worlds protect the participants' identity and privacy while keeping some of the nonverbal-communication cues still available [109].

Importantly, the anonymity given by the avatar allows participants to disclose emotions and concerns more freely and participate in social interactions with minimal risk of rejection [100]. However, we need to be mindful of the Proteus effect, which shows that the behavior in virtual worlds tends to mimic the characteristics of the virtual avatar chosen [117]. Herrera et al [114] reported that the behavioral realism of avatars used in the virtual environment also affects their nonverbal behavior in the physical world. This factor, often omitted in the current studies, could have both beneficial effects (eg, in increasing self-confidence and courage) and negative influence (eg, an increase in antisocial behavior toward other members of the group), and should be considered in the future usage of VR avatars in multiuser clinical interventions.

Finally, the professional or peer moderating the group session also has perfect control over the content of meetings in virtual worlds, often missing in other internet-based communication tools [61], enabling them to create a safe and supportive environment for each participant.

However, some constraints should be considered in VR interventions, mainly due to technical demands, as virtual worlds may require specialized hardware and software to access them. These technical requirements could be a barrier for some individuals and are probably why, despite many advantages, VR interventions are not so frequently used in clinical settings yet. Entering the virtual world does not require a VR headset (HMD) in general, as it can be presented on a traditional computer screen, at the expense of a lower immersion level. However, accessing virtual worlds using any device can be challenging for new users not familiar with this technology (eg, almost a third of the new users in the study by Nosek et al [102], described Second Life as "somewhat difficult" to learn).

In contrast, highly immersive virtual devices such as HMDs, provide a higher sense of presence for the participants that could facilitate the therapeutic process. An observational study using the Innerworld VR application [116] suggested that a higher level of immersion with HMD could potentiate the anticipated symptom changes in contrast to devices with lower immersion. However, higher immersion might be potentially associated with some adverse effects, such as cybersickness symptoms (eg, headache or dizziness). A recent systematic review analyzed the side effects reported using the Simulator Sickness Questionnaire in 55 documented research articles on VR clinical interventions and concluded that side effects (related mainly to disorientation, nausea, and oculomotor disturbances) were reported more frequently with HMDs than in desktop systems [118]. Another systematic review conducted on 73 studies with VR and AR interventions shows that 45 of these studies failed to mention or measure adverse effects completely [119], which is alarming. Cybersickness should be evaluated using standardized measures and minimized with regard to the potential negative impact (eg, worsening clinical symptoms or an increased fall risk [119]). The benefits of the applied intervention should always outweigh the risks and challenges associated with the technology. Immersive group interventions in anxiety disorders should also be monitored for potential ambiguities in cybersickness evaluation as these symptoms might be mediated by provoked anxiety [120], and standardized measures, such as Simulator Sickness Questionnaire should be adjusted in this target group accordingly, as suggested by Bouchard et al [121].

By contrast, advanced HMD devices could enable stimulation through multiple sensory modalities, by incorporating features, such as haptic feedback, spatial audio, and realistic environmental elements that can further enhance immersion and user engagement which should be addressed in future studies. Moreover, the shift toward MR technology enables us to benefit from social collaboration and interactions within the comfort of familiar surroundings. This combination of virtual and physical environments prepares the ground for highly social and colocated experiences, which could be applied not only remotely but also to in-person therapies [122].

Overall, immersion in the context of multiusers VR interventions could enhance social interactions and collaboration among participants [9]. By creating shared virtual spaces where users can interact and engage with each other, multiuser VR applications can foster a sense of community and support among individuals seeking mental health interventions. Interventions in virtual worlds are also reported to be more enjoyable [94].

However, the accessibility of this fast-developing immersive technology to clinical facilities and clinical populations represents a huge disadvantage. Despite the notable reduced financial costs of acquiring HMDs, the rapid obsolescence of the technology from the perspective of manufacturers constantly creating new devices greatly complicates the development of clinical applications. Developers must place considerable emphasis on multi-platform software solutions compatible with different brands and iterations of headsets, which is even more challenging in multiuser solutions.

Limitations

The main study limitation is the inability to synthesize the selected records. The variability of target groups, study designs, methods, and outcome measures used in the reviewed records prevented us from a rigorous synthesis of the reported studies, which were not easily comparable. As our main objective was

XSL•FO
RenderX

to address the rate of occurrence or popularity of multiuser digital interventions for mental health care, we did not focus in detail on the efficiency of individual applications and the synthesis of their clinical outcomes. The large variability could also lead to a publication bias. When so many different technologies and research methodologies are employed, some studies might observe positive results by chance, and without replication, such findings might not be refuted. We have observed most of the studies reporting positive outcomes, but we cannot be certain if this is due to the efficiency of the method, or due to a publication bias skewed toward positive results.

Another limitation is related to the query used for the database search of the selected records. The inconsistency in the terminology used in the reviewed articles, particularly in the context of "multiuser" technologies, creates a potential negative selection bias that could lead to certain publications not being registered in our search. For example, the term "cognitive behavioral immersion" has been used in a few studies testing the Innerworld application [21,116] not included in the systematic search.

Conclusions and Future Directions

Overall, participants in the multiuser digital interventions experienced positive emotional, social, and clinical outcomes. Although internet-based group meetings and virtual worlds offer unique advantages and opportunities for social connections, they also present challenges that must be addressed to ensure effective communication and engagement during group interventions. Balancing the benefits of accessibility and convenience with the limitations of technology and social interaction remains a key consideration in leveraging these platforms for mental health support and peer mediation.

To allow for systematic research, future studies should follow certain principles. One of the most important factors is a clear unification of terminology for multiuser interventions. Detailed reporting of methods is also important, especially with regard to the applied devices and procedures used during these interventions, including the frequency and synchronicity of sessions and standardized outcome measures, which then allows for valid evaluation of observed effects in case of feasibility studies and clinical trials.

As technology evolves, we can expect further innovations in how people connect and collaborate in virtual spaces, shaping the future of remote communication and collaboration. On the basis of the technological progress in VR and MR devices and multiuser virtual environments, the potential benefits of these platforms could be used in mental health care in the near future. Recent developments have shown promise in leveraging the immersive nature of VR to create engaging and effective mental health interventions in multiuser setups. In the future, virtual spaces tailored to specific mental health conditions are likely to be developed. In group VR therapy, an environment where all patients or clients engage in a shared task could be used, similar to individual VR-based exposure therapy, wherein patients with social phobia might practice public speaking in a virtual auditorium with other patients being able to observe and interact, thus learn from each other's experiences and strategies and experience a sense of support. Current applications also do not offer many opportunities and benefits of virtual spaces, such as using stimuli presented in the environment (eg, a virtual whiteboard). However, the technology must be studied in detail with regard to its limitations and challenges mentioned previously, as these may lead to a reduction or complete suppression of the expected therapeutic effects.

Future research studies addressing the feasibility and effectiveness of VR-based multiuser interventions should also address some of the gaps identified in this review. Particularly, neither the effects of customized avatars compared with real identities nor the Proteus effect was studied in reported clinical studies. Moreover, various levels of immersion used to visualize virtual environments were not directly compared in previous studies. Thus, it is not clear whether immersive VR provides clear benefits with regards to the clinical effectiveness of multiuser interventions that would diminish or outweigh any barriers that this technology certainly has so far.

The introduction of AI chatbots represents the next logical step in digital interventions. AI technology is not yet sufficiently advanced to be fully comparable to human communication, especially when simulating experts and peers. Errors and glitches in verbal and nonverbal communication, and other limitations, such as missing personalization or simulated emotions and empathy [123], might affect intervention outcomes, especially in terms of adherence rate and long-term benefits [124]. Nevertheless, further technological developments in this area will surely lead to significant advances over time, especially in dyadic communication agents providing scalable and accessible first-contact intervention in mental health care [125]. In addition, there will surely be an emphasis on the creation of ethical frameworks to guide the development and use of these technologies and to address issues, such as patients' privacy or potential biases in AI algorithms [126].

Acknowledgments

The study was supported by the project Research of Excellence on Digital Technologies and Wellbeing (grant CZ.02.01.01/00/22_008/0004583) cofinanced by the European Union.

Data Availability

The datasets generated during and analyzed during this study and code used to generate them are available in the GitHub repository [127].

XSL•FO
RenderX

Authors' Contributions

All authors contributed to the conceptualization and design of the systematic review. MS was responsible for the preparation of the database search queries. LH was responsible for data merging, and setup of the sorting tool used by all evaluators during the screening procedure. IF, LH, MS, MJ, and AF were responsible for the screening of all selected records for eligibility. LH and MJ were responsible for the systematic synthesis of the presented results. IF, PS, and MS wrote the manuscript draft. MJ and LH reviewed and edited the revised manuscript versions. IF was responsible for funding acquisition and study supervision. All authors read and revised the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete list of 49 reviewed studies using group digital interventions.
[PDF File (Adobe PDF File), 718 KB - jmir_v26i1e60441_app1.pdf]

Multimedia Appendix 2
PRISMA checklist.
[PDF File (Adobe PDF File), 68 KB - jmir_v26i1e60441_app2.pdf]

References

1. Chan S, Li L, Torous J, Gratzer D, Yellowlees PM. Review and implementation of self-help and automated tools in mental health care. Psychiatr Clin North Am 2019 Dec;42(4):597-609. [doi: 10.1016/j.psc.2019.07.001] [Medline: 31672210]
2. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023 Jan 28;53(2):567-579 [FREE Full text] [doi: 10.1017/S0033291721001999] [Medline: 34047264]
3. Malakouti SK, Rasouli N, Rezaeian M, Nojomi M, Ghanbari B, Shahraki Mohammadi A. Effectiveness of self-help mobile telephone applications (apps) for suicide prevention: a systematic review. Med J Islam Repub Iran 2020;34:85 [FREE Full text] [doi: 10.34171/mjiri.34.85] [Medline: 33306048]
4. Chan S, Li L, Torous J, Gratzer D, Yellowlees PM. Review of use of asynchronous technologies incorporated in mental health care. Curr Psychiatry Rep 2018 Aug 28;20(10):85. [doi: 10.1007/s11920-018-0954-3] [Medline: 30155593]
5. Tang Y, Gierc M, Lam RW, Liu S, Faulkner G. The effectiveness of internet-guided self-help interventions to promote physical activity among individuals with depression: systematic review. JMIR Ment Health 2022 Dec 12;9(12):e38049 [FREE Full text] [doi: 10.2196/38049] [Medline: 36508243]
6. Arif R, Ashraf S, Bhatt K, Shah K. A literature review examining virtual reality exposure therapy for individuals diagnosed with social anxiety disorder. J Nerv Ment Dis 2023;211(10):729-734. [doi: 10.1097/nmd.0000000000001698]
7. Norcross JC, Lambert MJ. Psychotherapy relationships that work III. Psychotherapy (Chic) 2018 Dec;55(4):303-315. [doi: 10.1037/pst0000193] [Medline: 30335448]
8. Rothke S. The role of interpersonal feedback in group psychotherapy. Int J Group Psychother 1986 Apr 29;36(2):225-240. [doi: 10.1080/00207284.1986.11491450] [Medline: 3733292]
9. Galliers J, Wilson S, Marshall J, Talbot R, Devane N, Booth T, et al. Experiencing EVA park, a multi-user virtual world for people with aphasia. ACM Trans Access Comput 2017 Oct 05;10(4):1-24. [doi: 10.1145/3134227]
10. Crowell C, Sayis B, Benitez JP, Pares N. Mixed reality, full-body interactive experience to encourage social initiation for autism: comparison with a control nondigital intervention. Cyberpsychol Behav Soc Netw 2020 Jan 01;23(1):5-9. [doi: 10.1089/cyber.2019.0115] [Medline: 31644332]
11. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020 Apr 03;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
12. Ali K, Farrer L, Gulliver A, Griffiths KM. Online peer-to-peer support for young people with mental health problems: a systematic review. JMIR Ment Health 2015 May 19;2(2):e19 [FREE Full text] [doi: 10.2196/mental.4418] [Medline: 26543923]
13. Liblub S, Pringle K, McLaughlin K, Cummins A. Peer support and mobile health for perinatal mental health: a scoping review. Birth 2024 Sep 24;51(3):484-496. [doi: 10.1111/birt.12814] [Medline: 38268345]
14. Repper J, Carter T. A review of the literature on peer support in mental health services. J Ment Health 2011 Aug;20(4):392-411. [doi: 10.3109/09638237.2011.583947] [Medline: 21770786]
15. Longman H, O'Connor E, Obst P. The effect of social support derived from world of warcraft on negative psychological symptoms. Cyberpsychol Behav 2009 Oct;12(5):563-566. [doi: 10.1089/cpb.2009.0001] [Medline: 19817567]
16. Ringland K. "Autsome": fostering an autistic identity in an online Minecraft community for youth with autism. Inf Contemp Soc (2019) 2019 Apr;11420:132-143 [FREE Full text] [doi: 10.1007/978-3-030-15742-5_12] [Medline: 31157326]

XSL•FO
RenderX

17. Wang M. Social VR: a new form of social communication in the future or a beautiful illusion? J Phys Conf Ser 2020 Apr 01;1518(1):012032. [doi: 10.1088/1742-6596/1518/1/012032]
18. Kelley B. The rise of the 'quarantine bar simulator': the uses and gratifications of social VR during the COVID-19 pandemic. In: Proceedings of the 4th International Conference on Information and Computer Technologies. 2021 Presented at: ICICT' 21; March 11-14, 2021; Hilo, HI p. 216-221 URL: https://ieeexplore.ieee.org/document/9476918 [doi: 10.1109/icict52872.2021.00043]
19. Deighan MT, Ayobi A, O'Kane A. Social virtual reality as a mental health tool: how people use VRChat to support social connectedness and wellbeing. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany p. 1-13 URL: https://dl.acm.org/doi/10.1145/3544548.3581103 [doi: 10.1145/3544548.3581103]
20. Zamanifard S, Robb A. Social virtual reality is my therapist: overcoming social anxiety disorder through using social virtual reality. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI EA '23; April 23-28, 2023; Hamburg, Germany p. 1-6 URL: https://dl.acm.org/doi/10.1145/3544549.3585888 [doi: 10.1145/3544549.3585888]
21. Robinson N, Mahapatra A, Jean-Baptiste B, Mallard A, Yang A, Hollon SD, et al. Cognitive behavioral immersion for substance use disorders: a feasibility and pilot study of a peer-based coaching program in the metaverse. Games Health J 2023 Oct 01;12(5):397-404 [FREE Full text] [doi: 10.1089/g4h.2022.0214] [Medline: 37294540]
22. Horizon workrooms virtual office and meetings. Meta. URL: https://forwork.meta.com/horizon-workrooms/ [accessed 2024-07-30]
23. Home page. Rec Room. URL: https://recroom.com/ [accessed 2024-07-30]
24. Havu J, Glukhova P. Glue is a virtual collaboration platform for teams who need remote meetings to be as great as face-to-face meetings. Glue. URL: https://www.glue.work/ [accessed 2024-07-30]
25. Clough BA, Casey LM. Technological adjuncts to enhance current psychotherapy practices: a review. Clin Psychol Rev 2011 Apr;31(3):279-292 [FREE Full text] [doi: 10.1016/j.cpr.2010.12.008] [Medline: 21382535]
26. Imel ZE, Caperton DD, Tanana M, Atkins DC. Technology-enhanced human interaction in psychotherapy. J Couns Psychol 2017 Jul;64(4):385-393 [FREE Full text] [doi: 10.1037/cou0000213] [Medline: 28318277]
27. Malhotra A, Baker J. Group therapy. StatPearls. 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK549812/ [accessed 2024-04-29]
28. Worrall H, Schweizer R, Marks E, Yuan L, Lloyd C, Ramjan R. The effectiveness of support groups: a literature review. Ment Health Soc Incl 2018 Apr 09;22(2):85-93. [doi: 10.1108/mhsi-12-2017-0055]
29. Stein A. The nature and significance of interaction in group psychotherapy. Int J Group Psychother 1970 Apr 29;20(2):153-162. [doi: 10.1080/00207284.1970.11492071] [Medline: 5423652]
30. Durkin HE. Toward a common basis for group dynamics. Int J Group Psychother 2015 Oct 29;7(2):115-130. [doi: 10.1080/00207284.1957.11508725]
31. Guttmacher JA, Birk L. Group therapy: what specific therapeutic advantages? Compr Psychiatry 1971 Nov;12(6):546-556. [doi: 10.1016/0010-440x(71)90037-x] [Medline: 5117074]
32. Weisberg RB, Beard C, Moitra E, Dyck I, Keller MB. Adequacy of treatment received by primary care patients with anxiety disorders. Depress Anxiety 2014 May;31(5):443-450 [FREE Full text] [doi: 10.1002/da.22209] [Medline: 24190762]
33. Marogna C, Caccamo F. Analysis of the process in brief psychotherapy group: the role of therapeutic factors. ResPsy 2015 Mar 01;17(1):43-51. [doi: 10.4081/ripppo.2014.161]
34. Yalom ID, Leszcz M. Theory and Practice of Group Psychotherapy. 5th Edition. New York, NY: Basic Books; 2005.
35. Fuhriman A, Burlingame GM. Consistency of matter: a comparative analysis of individual and group process variables. Couns Psychol 1990 Jan 01;18(1):6-63. [doi: 10.1177/0011000090181002]
36. Caccamo F, Ghedin S, Marogna C. Evaluation of burnout and alexithymia in a group of nurses in palliative care: a pilot study. G Ital Med Lav Ergon 2017 Dec;39(4):249-255. [Medline: 29916571]
37. Caccamo F, Saltini S, Carella E, Carlon R, Marogna C, Sava V. The measure of effectiveness of a short-term 2-week intensive cardiac rehabilitation program in decreasing levels of anxiety and depression. Monaldi Arch Chest Dis 2018 Jan 29;88(1):858 [FREE Full text] [doi: 10.4081/monaldi.2018.858] [Medline: 29557570]
38. Behenck A, Wesner AC, Finkler D, Heldt E. Contribution of group therapeutic factors to the outcome of cognitive-behavioral therapy for patients with panic disorder. Arch Psychiatr Nurs 2017 Apr;31(2):142-146. [doi: 10.1016/j.apnu.2016.09.001] [Medline: 28359425]
39. Choi YH, Park KH. Therapeutic factors of cognitive behavioral group treatment for social phobia. J Korean Med Sci 2006 Apr;21(2):333-336 [FREE Full text] [doi: 10.3346/jkms.2006.21.2.333] [Medline: 16614524]
40. Taube-Schiff M, Suvak MK, Antony MM, Bieling PJ, McCabe RE. Group cohesion in cognitive-behavioral group therapy for social phobia. Behav Res Ther 2007 Apr;45(4):687-698. [doi: 10.1016/j.brat.2006.06.004] [Medline: 16928359]
41. Heckman TG, Heckman BD, Anderson T, Bianco JA, Sutton M, Lovejoy TI. Common factors and depressive symptom relief trajectories in group teletherapy for persons ageing with HIV. Clin Psychol Psychother 2017 Jan 04;24(1):139-148. [doi: 10.1002/cpp.1989] [Medline: 26538241]

XSL•FO
RenderX

42. Brouzos A, Vassilopoulos SP, Stavrou V, Baourda VC, Tassi C, Brouzou KO. Therapeutic factors and member satisfaction in an online group intervention during the COVID-19 pandemic. J Technol Behav Sci 2021 Sep 26;6(4):609-619 [FREE Full text] [doi: 10.1007/s41347-021-00216-4] [Medline: 34604505]
43. Strauss B, Spangenberg L, Brähler E, Bormann B. Attitudes towards (psychotherapy) groups: results of a survey in a representative sample. Int J Group Psychother 2015 Jul 24;65(3):410-430. [doi: 10.1521/ijgp_2014_64_001] [Medline: 24963534]
44. Shechtman Z, Kiezel A. Why do people prefer individual therapy over group therapy? Int J Group Psychother 2016 Oct 07;66(4):571-591. [doi: 10.1080/00207284.2016.1180042] [Medline: 38475636]
45. Wolgensinger L. Cognitive behavioral group therapy for anxiety: recent developments. Dialogues Clin Neurosci 2022 Apr 01;17(3):347-351. [doi: 10.31887/dcns.2015.17.3/lwolgensinger]
46. Fisak Jr BJ, Richard D, Mann A. The prevention of child and adolescent anxiety: a meta-analytic review. Prev Sci 2011 Sep;12(3):255-268. [doi: 10.1007/s11121-011-0210-0] [Medline: 21437675]
47. Barkowski S, Schwartze D, Strauss B, Burlingame GM, Rosendahl J. Efficacy of group psychotherapy for anxiety disorders: a systematic review and meta-analysis. Psychother Res 2020 Nov 24;30(8):965-982. [doi: 10.1080/10503307.2020.1729440] [Medline: 32093586]
48. Mayo-Wilson E, Dias S, Mavranezouli I, Kew K, Clark DM, Ades AE, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. Lancet Psychiatry 2014 Oct;1(5):368-376. [doi: 10.1016/s2215-0366(14)70329-3]
49. Kodal A, Fjermestad K, Bjelland I, Gjestad R, Öst LG, Bjaastad JF, et al. Long-term effectiveness of cognitive behavioral therapy for youth with anxiety disorders. J Anxiety Disord 2018 Jan;53:58-67 [FREE Full text] [doi: 10.1016/j.janxdis.2017.11.003] [Medline: 29195188]
50. Sigurvinsdóttir AL, Jensínudóttir KB, Baldvinsdóttir KD, Smárason O, Skarphedinsson G. Effectiveness of cognitive behavioral therapy (CBT) for child and adolescent anxiety disorders across different CBT modalities and comparisons: a systematic review and meta-analysis. Nord J Psychiatry 2020 Apr 18;74(3):168-180. [doi: 10.1080/08039488.2019.1686653] [Medline: 31738631]
51. Burlingame GM, Seebeck JD, Janis RA, Whitcomb KE, Barkowski S, Rosendahl J, et al. Outcome differences between individual and group formats when identical and nonidentical treatments, patients, and doses are compared: a 25-year meta-analytic perspective. Psychotherapy (Chic) 2016 Dec;53(4):446-461. [doi: 10.1037/pst0000090] [Medline: 27918191]
52. Wersebe H, Sijbrandij M, Cuijpers P. Correction: psychological group-treatments of social anxiety disorder: a meta-analysis. PLoS ONE 2013 Dec 6;8(12):12. [doi: 10.1371/annotation/5f2f7ff4-ecfd-4a41-a162-34c1dd0c962a]
53. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Apr;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
54. Marques IC, Ferreira JJ. Digital transformation in the area of health: systematic review of 45 years of evolution. Health Technol 2019 Dec 13;10(3):575-586. [doi: 10.1007/S12553-019-00402-8]
55. Shalaby RA, Agyapong VI. Peer support in mental health: literature review. JMIR Ment Health 2020 Jun 09;7(6):e15572 [FREE Full text] [doi: 10.2196/15572] [Medline: 32357127]
56. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166 [FREE Full text] [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
57. Ridout B, Campbell A. The use of social networking sites in mental health interventions for young people: systematic review. J Med Internet Res 2018 Dec 18;20(12):e12244 [FREE Full text] [doi: 10.2196/12244] [Medline: 30563811]
58. Banbury A, Nancarrow S, Dart J, Gray L, Parkinson L. Telehealth interventions delivering home-based support group videoconferencing: systematic review. J Med Internet Res 2018 Feb 02;20(2):e25 [FREE Full text] [doi: 10.2196/jmir.8090] [Medline: 29396387]
59. Dincelli E, Yayla A. Immersive virtual reality in the age of the Metaverse: a hybrid-narrative review based on the technology affordance perspective. J Strateg Inf Syst 2022 Jun;31(2):101717. [doi: 10.1016/j.jsis.2022.101717]
60. Ghanbarzadeh R, Ghapanchi AH, Blumenstein M, Talaei-Khoei A. A decade of research on the use of three-dimensional virtual worlds in health care: a systematic literature review. J Med Internet Res 2014 Feb 18;16(2):e47 [FREE Full text] [doi: 10.2196/jmir.3097] [Medline: 24550130]
61. Weinberg H. Obstacles, challenges, and benefits of online group psychotherapy. Am J Psychother 2021 Jun 01;74(2):83-88. [doi: 10.1176/appi.psychotherapy.20200034] [Medline: 33525914]
62. Bell IH, Pot-Kolder R, Rizzo A, Rus-Calafell M, Cardi V, Cella M, et al. Advances in the use of virtual reality to treat mental health conditions. Nat Rev Psychol 2024 Jul 10;3(8):552-567. [doi: 10.1038/s44159-024-00334-9]
63. Pons P, Navas-Medrano S, Soler-Dominguez JL. Extended reality for mental health: current trends and future challenges. Front Comput Sci 2022 Nov 18;4:1-13. [doi: 10.3389/fcomp.2022.1034307]
64. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

65. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
66. NocoDB Cloud. URL: https://nocodb.com [accessed 2024-05-10]
67. Ronen K, Gewali A, Dachelet K, White E, Jean-Baptiste M, Evans YN, et al. Acceptability and utility of a digital group intervention to prevent perinatal depression in youths via interactive maternal group for information and emotional support (IMAGINE): pilot cohort study. JMIR Form Res 2024 Feb 02;8:e51066 [FREE Full text] [doi: 10.2196/51066] [Medline: 38306159]
68. Pokowitz EL, Prakash N, Planaj D, Oprandi S, Deldin PJ. Mood lifters for graduate students and young adults: a mixed-methods investigation into mechanisms of change in online group therapy. Behav Sci (Basel) 2024 Mar 20;14(3):252 [FREE Full text] [doi: 10.3390/bs14030252] [Medline: 38540555]
69. Penwell TE, Smith M, Ortiz SN, Brooks G, Thompson-Brenner H. Traditional versus virtual partial hospital programme for eating disorders: feasibility and preliminary comparison of effects. Eur Eat Disord Rev 2024 Mar 07;32(2):163-178. [doi: 10.1002/erv.3031] [Medline: 37677002]
70. Conroy K, Kehrer SM, Georgiadis C, Hare M, Ringle VM, Shaw AM. Learning from adolescents and caregivers to enhance acceptability and engagement within virtual dialectical behavior therapy for adolescents skills groups: a qualitative study. Child Psychiatry Hum Dev (Forthcoming) 2024 Jan 06. [doi: 10.1007/s10578-023-01641-7] [Medline: 38183575]
71. Zimmerman M, D'Avanzato C, King BT. Telehealth treatment of patients with major depressive disorder during the COVID-19 pandemic: comparative safety, patient satisfaction, and effectiveness to prepandemic in-person treatment. J Affect Disord 2023 Feb 15;323:624-630 [FREE Full text] [doi: 10.1016/j.jad.2022.12.015] [Medline: 36521663]
72. Xie EB, Freeman M, Penner-Goeke L, Reynolds K, Lebel C, Giesbrecht GF, et al. Building Emotional Awareness and Mental Health (BEAM): an open-pilot and feasibility study of a digital mental health and parenting intervention for mothers of infants. Pilot Feasibility Stud 2023 Feb 18;9(1):27 [FREE Full text] [doi: 10.1186/s40814-023-01245-x] [Medline: 36800982]
73. Wright E, Martinovic J, de Camps Meschino D, Barker LC, Philipp DA, Israel A, et al. A virtual mother-infant postpartum psychotherapy group for mothers with a history of adverse childhood experiences: open-label feasibility study. BMC Psychiatry 2023 Dec 18;23(1):950 [FREE Full text] [doi: 10.1186/s12888-023-05444-x] [Medline: 38110902]
74. Wang C, Darbari I, Tolaymat M, Quezada S, Allen J, Cross RK. Randomized controlled pilot study of feasibility and effectiveness of peer led remote mindfulness‐based art workshops on stress, anxiety, and depression in medical students. Psychol Sch 2022 Sep 30;60(6):1744-1759. [doi: 10.1002/pits.22809]
75. Hunt X, Jivan DC, Naslund JA, Breet E, Bantjes J. South African university students' experiences of online group cognitive behavioural therapy: implications for delivering digital mental health interventions to young people. Glob Ment Health (Camb) 2023 Jul 28;10:e45 [FREE Full text] [doi: 10.1017/gmh.2023.39] [Medline: 37854416]
76. Hollis C, Hall CL, Khan K, Le Novere M, Marston L, Jones R, et al. Online remote behavioural intervention for tics in 9- to 17-year-olds: the ORBIT RCT with embedded process and economic evaluation. Health Technol Assess 2023 Oct:1-120. [doi: 10.3310/cpms3211]
77. Gerson J, Tawde P, Ghiasian G, Salwen-Deremer JK. Patients' experiences with virtual group gut-directed hypnotherapy: a qualitative study. Front Med (Lausanne) 2023 Feb 22;10:1066452 [FREE Full text] [doi: 10.3389/fmed.2023.1066452] [Medline: 36910502]
78. Finnerty R, McWeeny S, Trainor L. Online group music therapy: proactive management of undergraduate students' stress and anxiety. Front Psychiatry 2023 Apr 21;14:1183311 [FREE Full text] [doi: 10.3389/fpsyt.2023.1183311] [Medline: 37151974]
79. Brownstone LM, Hunsicker MJ, Palazzolo LP, Dinneen JL, Kelly DA, Stennes J, et al. “Sharing lived experience”: describing a virtual counselor-facilitated LGBTQ+ support group for disordered eating. Psychol Sex Orientat Gend Divers 2023 Aug 10 [FREE Full text] [doi: 10.1037/sgd0000660]
80. Zaccari B, Loftis JM, Haywood T, Hubbard K, Clark J, Kelly UA. Synchronous telehealth yoga and cognitive processing group therapies for women veterans with posttraumatic stress disorder: a multisite randomized controlled trial adapted for COVID-19. Telemed J E Health 2022 Mar 29;28(11):1642-1650 [FREE Full text] [doi: 10.1089/tmj.2021.0612] [Medline: 35357957]
81. Mi RZ, Kornfield R, Shah DV, Maus A, Gustafson DH. Intraindividual, dyadic, and network communication in a digital health intervention: distinguishing message exposure from message production. Health Commun 2022 Apr 25;37(4):397-408 [FREE Full text] [doi: 10.1080/10410236.2020.1846273] [Medline: 33238732]
82. Mendelson D, Thibaudeau É, Sauvé G, Lavigne KM, Bowie CR, Menon M, et al. Remote group therapies for cognitive health in schizophrenia-spectrum disorders: feasible, acceptable, engaging. Schizophr Res Cogn 2022 Jun;28:100230 [FREE Full text] [doi: 10.1016/j.scog.2021.100230] [Medline: 35242604]
83. Kruzan KP, Whitlock J, Bazarova NN, Bhandari A, Chapman J. Use of a mobile peer support app among young people with nonsuicidal self-injury: small-scale randomized controlled trial. JMIR Form Res 2022 Jan 10;6(1):e26526 [FREE Full text] [doi: 10.2196/26526] [Medline: 35006076]

XSL•FO
RenderX

84. Fell L, Goshe B, Traeger L, Perez G, Iannuzzi D, Park E, et al. Acceptability of a virtual mind-body group intervention for teen siblings of children with autism spectrum disorder. J Autism Dev Disord 2022 Dec 03;52(12):5243-5252. [doi: 10.1007/s10803-022-05500-7] [Medline: 35657446]
85. El Ayadi AM, Duggal M, Bagga R, Singh P, Kumar V, Ahuja A, et al. A mobile education and social support group intervention for improving postpartum health in northern India: development and usability study. JMIR Form Res 2022 Jun 29;6(6):e34087 [FREE Full text] [doi: 10.2196/34087] [Medline: 35767348]
86. Ali K, Fassnacht DB, Farrer LM, Rieger E, Moessner M, Bauer S, et al. Recruitment, adherence and attrition challenges in internet-based indicated prevention programs for eating disorders: lessons learned from a randomised controlled trial of ProYouth OZ. J Eat Disord 2022 Jan 04;10(1):1 [FREE Full text] [doi: 10.1186/s40337-021-00520-7] [Medline: 34983675]
87. Yeshua-Katz D. The role of communication affordances in post-traumatic stress disorder Facebook and WhatsApp support groups. Int J Environ Res Public Health 2021 Apr 26;18(9):4576 [FREE Full text] [doi: 10.3390/ijerph18094576] [Medline: 33925904]
88. Dabit S, Quraishi S, Jordan J, Biagianti B. Improving social functioning in people with schizophrenia-spectrum disorders via mobile experimental interventions: results from the CLIMB pilot trial. Schizophr Res Cogn 2021 Dec;26:100211 [FREE Full text] [doi: 10.1016/j.scog.2021.100211] [Medline: 34381699]
89. Craig SL, Leung VW, Pascoe R, Pang N, Iacono G, Austin A, et al. AFFIRM online: utilising an affirmative cognitive-behavioural digital intervention to improve mental health, access, and engagement among LGBTQA+ youth and young adults. Int J Environ Res Public Health 2021 Feb 05;18(4):1541 [FREE Full text] [doi: 10.3390/ijerph18041541] [Medline: 33562876]
90. Craig SL, Iacono G, Pascoe R, Austin A. Adapting clinical skills to telehealth: applications of affirmative cognitive-behavioral therapy with LGBTQ+ youth. Clin Soc Work J 2021 Mar 02;49(4):471-483 [FREE Full text] [doi: 10.1007/s10615-021-00796-x] [Medline: 33678921]
91. Abedishargh N, Ramezani Farani A, Gharraee B, Farahani H. Effectiveness of internet-based cognitive behavioral therapy in weight loss, stress, anxiety, and depression via virtual group therapy. Iran J Psychiatry Behav Sci 2021 Sep 01;15(3):e113096. [doi: 10.5812/ijpbs.113096]
92. Tsai HH, Cheng CY, Shieh WY, Chang YC. Effects of a smartphone-based videoconferencing program for older nursing home residents on depression, loneliness, and quality of life: a quasi-experimental study. BMC Geriatr 2020 Jan 28;20(1):27 [FREE Full text] [doi: 10.1186/s12877-020-1426-2] [Medline: 31992217]
93. Strand M, Eng LS, Gammon D. Combining online and offline peer support groups in community mental health care settings: a qualitative study of service users' experiences. Int J Ment Health Syst 2020;14:39 [FREE Full text] [doi: 10.1186/s13033-020-00370-x] [Medline: 32514303]
94. Robinson-Whelen S, Hughes RB, Taylor HB, Markley R, Vega JC, Nosek TM, et al. Promoting psychological health in women with SCI: development of an online self-esteem intervention. Disabil Health J 2020 Apr;13(2):100867. [doi: 10.1016/j.dhjo.2019.100867] [Medline: 31757776]
95. Vogel EA, Thrul J, Humfleet GL, Delucchi KL, Ramo DE. Smoking cessation intervention trial outcomes for sexual and gender minority young adults. Health Psychol 2019 Jan;38(1):12-20 [FREE Full text] [doi: 10.1037/hea0000698] [Medline: 30489104]
96. Tait RJ, Paz Castro R, Kirkman JJ, Moore JC, Schaub MP. A digital intervention addressing alcohol use problems (the "Daybreak" program): quasi-experimental randomized controlled trial. J Med Internet Res 2019 Sep 04;21(9):e14967 [FREE Full text] [doi: 10.2196/14967] [Medline: 31486406]
97. Sawyer A, Kaim A, Le HN, McDonald D, Mittinty M, Lynch J, et al. The effectiveness of an app-based nurse-moderated program for new mothers with depression and parenting problems (eMums plus): pragmatic randomized controlled trial. J Med Internet Res 2019 Jun 04;21(6):e13689 [FREE Full text] [doi: 10.2196/13689] [Medline: 31165715]
98. Campbell A, Ridout B, Amon K, Navarro P, Collyer B, Dalgleish J. A customized social network platform (kids helpline circles) for delivering group counseling to young people experiencing family discord that impacts their well-being: exploratory study. J Med Internet Res 2019 Dec 20;21(12):e16176 [FREE Full text] [doi: 10.2196/16176] [Medline: 31859671]
99. Granado-Font E, Ferré-Grau C, Rey-Reñones C, Pons-Vigués M, Pujol Ribera E, Berenguera A, et al. Coping strategies and social support in a mobile phone chat app designed to support smoking cessation: qualitative analysis. JMIR Mhealth Uhealth 2018 Dec 20;6(12):e11071 [FREE Full text] [doi: 10.2196/11071] [Medline: 30573445]
100. Knowles LM, Stelzer EM, Jovel KS, O'Connor MF. A pilot study of virtual support for grief: feasibility, acceptability, and preliminary outcomes. Comput Human Behav 2017 Aug;73:650-658. [doi: 10.1016/J.CHB.2017.04.005]
101. Carolan S, Harris PR, Greenwood K, Cavanagh K. Increasing engagement with an occupational digital stress management program through the use of an online facilitated discussion group: results of a pilot randomised controlled trial. Internet Interv 2017 Dec;10:1-11 [FREE Full text] [doi: 10.1016/j.invent.2017.08.001] [Medline: 30135747]
102. Nosek MA, Robinson-Whelen S, Hughes RB, Nosek TM. An internet-based virtual reality intervention for enhancing self-esteem in women with disabilities: results of a feasibility study. Rehabil Psychol 2016 Nov;61(4):358-370. [doi: 10.1037/rep0000107] [Medline: 27709978]

XSL•FO
RenderX

103. Bohleber L, Crameri A, Eich-Stierli B, Telesko R, von WA. Can we foster a culture of peer support and promote mental health in adolescence using a web-based app? A control group study. JMIR Ment Health 2016 Sep 23;3(3):e45 [FREE Full text] [doi: 10.2196/mental.5597] [Medline: 27663691]
104. Lemma A, Fonagy P. Feasibility study of a psychodynamic online group intervention for depression. Psychoanal Psychother 2013 Jul;30(3):367-380. [doi: 10.1037/a0033239]
105. Schefft C, Krämer R, Haaf R, Jedeck D, Schumacher A, Köhler S. Evaluation of the internet-based intervention "Selfapy" in participants with unipolar depression and the impact on quality of life: a randomized, parallel group study. Qual Life Res 2024 May 25;33(5):1275-1286. [doi: 10.1007/s11136-024-03606-2] [Medline: 38403818]
106. Yeo G, Loo G, Oon M, Pang R, Ho D. A digital peer support platform to translate online peer support for emerging adult mental well-being: randomized controlled trial. JMIR Ment Health 2023 Apr 18;10:e43956 [FREE Full text] [doi: 10.2196/43956] [Medline: 36756843]
107. Bozkurt MA, Cesur B. The effect of the application on the self-efficacy and anxiety levels of mothers: randomized controlled trial. Digit Health 2023 Apr 25;9:20552076231169840 [FREE Full text] [doi: 10.1177/20552076231169840] [Medline: 37124329]
108. Arakawa Y, Haseda M, Inoue K, Nishioka D, Kino S, Nishi D, et al. Effectiveness of mHealth consultation services for preventing postpartum depressive symptoms: a randomized clinical trial. BMC Med 2023 Jun 26;21(1):221 [FREE Full text] [doi: 10.1186/s12916-023-02918-3] [Medline: 37365535]
109. Li C, Yip PY. Remote arts therapy in collaborative virtual environment: a pilot case study. Front Virtual Real 2023 May 12;4:1-6. [doi: 10.3389/frvir.2023.1059278]
110. Shorey S, Chee CY, Ng ED, Lau Y, Dennis C, Chan YH. Evaluation of a technology-based peer-support intervention program for preventing postnatal depression (part 1): randomized controlled trial. J Med Internet Res 2019 Aug 29;21(8):e12410 [FREE Full text] [doi: 10.2196/12410] [Medline: 31469084]
111. Topooco N, Byléhn S, Dahlström Nysäter E, Holmlund J, Lindegaard J, Johansson S, et al. Evaluating the efficacy of internet-delivered cognitive behavioral therapy blended with synchronous chat sessions to treat adolescent depression: randomized controlled trial. J Med Internet Res 2019 Nov 01;21(11):e13393 [FREE Full text] [doi: 10.2196/13393] [Medline: 31682572]
112. Shorey S, Ng ED. Evaluation of a technology-based peer-support intervention program for preventing postnatal depression (part 2): qualitative study. J Med Internet Res 2019 Aug 29;21(8):e12915 [FREE Full text] [doi: 10.2196/12915] [Medline: 31469080]
113. Depp CA, Mausbach B, Granholm E, Cardenas V, Ben-Zeev D, Patterson TL, et al. Mobile interventions for severe mental illness: design and preliminary data from three approaches. J Nerv Ment Dis 2010;198(10):715-721. [doi: 10.1097/nmd.0b013e3181f49ea3]
114. Herrera F, Oh SY, Bailenson JN. Effect of behavioral realism on social interactions inside collaborative virtual environments. Presence 2020;27:163-182. [doi: 10.1162/pres_a_00324]
115. Second Life. URL: http://secondlife.com/ [accessed 2024-05-10]
116. Ezawa ID, Hollon SD, Robinson N. Examining predictors of depression and anxiety symptom change in cognitive behavioral immersion: observational study. JMIR Ment Health 2023 Jul 14;10:e42377 [FREE Full text] [doi: 10.2196/42377] [Medline: 37450322]
117. Yee N, Bailenson JN. The difference between being and seeing: the relative contribution of self-perception and priming to behavioral changes via digital self-representation. Media Psychology 2009 May 19;12(2):195-209. [doi: 10.1080/15213260902849943]
118. Simón-Vicente L, Rodríguez-Cano S, Delgado-Benito V, Ausín-Villaverde V, Cubo Delgado E. Cybersickness. A systematic literature review of adverse effects related to virtual reality. Neurología 2022 Jul [FREE Full text] [doi: 10.1016/j.nrl.2022.04.009]
119. Lundin RM, Yeap Y, Menkes DB. Adverse effects of virtual and augmented reality interventions in psychiatry: systematic review. JMIR Ment Health 2023 May 05;10:e43240 [FREE Full text] [doi: 10.2196/43240] [Medline: 37145841]
120. Pot-Kolder R, Veling W, Counotte J, van der Gaag M. Anxiety partially mediates cybersickness symptoms in immersive virtual reality environments. Cyberpsychol Behav Soc Netw 2018 Mar;21(3):187-193. [doi: 10.1089/cyber.2017.0082] [Medline: 29356575]
121. Bouchard S, Berthiaume M, Robillard G, Forget H, Daudelin-Peltier C, Renaud P, et al. Arguing in favor of revising the factor structure when assessing side effects induced by immersions in virtual reality. Front Psychiatry 2021 Nov 5;12:739742 [FREE Full text] [doi: 10.3389/fpsyt.2021.739742] [Medline: 34803762]
122. Navas-Medrano S, Soler-Dominguez JL, Pons P. Mixed Reality for a collective and adaptive mental health metaverse. Front Psychiatry 2023 Jan 5;14:1272783 [FREE Full text] [doi: 10.3389/fpsyt.2023.1272783] [Medline: 38250268]
123. Possati LM. Psychoanalyzing artificial intelligence: the case of Replika. AI Soc 2023;38(4):1725-1738 [FREE Full text] [doi: 10.1007/s00146-021-01379-7]
124. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord 2024 Jul 01;356:459-469. [doi: 10.1016/j.jad.2024.04.057] [Medline: 38631422]

XSL•FO
RenderX

125. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec 31;18(sup1):37-49 [FREE Full text] [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
126. Khawaja Z, Bélisle-Pipon JC. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. Front Digit Health 2023 Nov 8;5:1278186 [FREE Full text] [doi: 10.3389/fdgth.2023.1278186] [Medline: 38026836]
127. Hejtmánek L. BrainVR/manuscript-multiuser-jimr: publication JIMR. Zenodo. URL: https://zenodo.org/records/13888256 [accessed 2024-04-29]

Abbreviations

AI: artificial intelligence
AR: augmented reality
CBT: cognitive behavioral therapy
HMD: head-mounted display
LGBTQA+: lesbian, gay, bisexual, transgender, intersex, queer or questioning, asexual
MR: mixed reality
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
VR: virtual reality

Edited by G Eysenbach; submitted 10.05.24; peer-reviewed by J Ferrer Costa, D Patel; comments to author 09.06.24; revised version received 31.07.24; accepted 31.08.24; published 07.11.24.

Please cite as:
Fajnerova I, Hejtmánek L, Sedlák M, Jablonská M, Francová A, Stopková P
The Journey From Nonimmersive to Immersive Multiuser Applications in Mental Health Care: Systematic Review
J Med Internet Res 2024;26:e60441
URL: https://www.jmir.org/2024/1/e60441
doi:10.2196/60441
PMID:

©Iveta Fajnerova, Lukáš Hejtmánek, Michal Sedlák, Markéta Jablonská, Anna Francová, Pavla Stopková. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Decoding the Digital Pulse: Bibliometric Analysis of 25 Years in Digital Health Research Through the Journal of Medical Internet Research

Robert Kaczmarczyk[1], MD; Theresa Isabelle Wilhelm[2], MD; Jonas Roos[3], MD; Ron Martin[4], MD

[1]Department of Dermatology and Allergy, Technical University of Munich, Munich, Germany
[2]Eye Center—Medical Center, Faculty of Medicine, Albert-Ludwigs-University of Freiburg, Freiburg, Germany
[3]Department of Orthopedics and Trauma Surgery, University Hospital of Bonn, Bonn, Germany
[4]Department of Plastic and Hand Surgery, Burn Care Center, BG Klinikum Bergmannstrost Halle, Halle, Germany

Corresponding Author:
Theresa Isabelle Wilhelm, MD
Eye Center—Medical Center
Faculty of Medicine
Albert-Ludwigs-University of Freiburg
Killianstraße 5
Freiburg, 79106
Germany
Phone: 49 76127040020
Email: theresa.wilhelm@uniklinik-freiburg.de

Abstract

Background: As the digital health landscape continues to evolve, analyzing the progress and direction of the field can yield valuable insights. The *Journal of Medical Internet Research* (JMIR) has been at the forefront of disseminating digital health research since 1999. A comprehensive network analysis of JMIR publications can help illuminate the evolution and trends in digital medicine over the past 25 years.

Objective: This study aims to conduct a detailed network analysis of JMIR's publications to uncover the growth patterns, dominant themes, and potential future trajectories in digital health research.

Methods: We retrieved 8068 JMIR papers from PubMed using the Biopython library. Keyword metrics were assessed using accuracy, recall, and F_1-scores to evaluate the effectiveness of keyword identification from Claude 3 Opus and Gemini 1.5 Pro in addition to 2 conventional natural language processing methods using key bidirectional encoder representations from transformers. Future trends for 2024-2026 were predicted using Claude 3 Opus, Google's Time Series Foundation Model, autoregressive integrated moving average, exponential smoothing, and Prophet. Network visualization techniques were used to represent and analyze the complex relationships between collaborating countries, paper types, and keyword co-occurrence.

Results: JMIR's publication volume showed consistent growth, with a peak in 2020. The United States dominated country contributions, with China showing a notable increase in recent years. Keyword analysis from 1999 to 2023 showed significant thematic shifts, from an early internet and digital health focus to the dominance of COVID-19 and advanced technologies such as machine learning. Predictions for 2024-2026 suggest an increased focus on artificial intelligence, digital health, and mental health.

Conclusions: Network analysis of JMIR publications provides a macroscopic view of the evolution of the digital health field. The journal's trajectory reflects broader technological advances and shifting research priorities, including the impact of the COVID-19 pandemic. The predicted trends underscore the growing importance of computational technology in future health care research and practice. The findings from JMIR provide a glimpse into the future of digital medicine, suggesting a robust integration of artificial intelligence and continued emphasis on mental health in the postpandemic era.

(J Med Internet Res 2024;26:e60057) doi:10.2196/60057

XSL•FO
RenderX

KEYWORDS

digital health; JMIR publication analysis; network analysis; artificial intelligence; AI; large language models; eHealth; Claude 3 Opus; digital; digital technology; digital intervention; machine learning; natural language processing; NLP; deep learning; algorithm; model; analytics; practical model; pandemic; postpandemic era; mobile phone

Introduction

The *Journal of Medical Internet Research* (JMIR), founded in 1999 as the first international scientific peer-reviewed journal focusing on health care research, information, and communication via internet and intranet technologies [1], pioneered several innovations. These included the introduction of new copyright frameworks and the adoption of electronic-only publishing and open-access content [2]. Over the past 25 years, digital health technologies such as telemedicine and wearable biosensors have significantly advanced health care by improving access to information and improving patient monitoring [3-5]. In addition, the use of artificial intelligence (AI) in clinical environments has brought significant advances, such as the ability to automate image analysis. This technology facilitates the early detection of diseases, for example, diabetic retinopathy through retinal images [6,7]. These advances promise to improve personalized medicine, preventative care, and patient outcomes, and the future is likely to continue to be shaped by technological innovation [8,9].

On the occasion of the 25th anniversary of JMIR, we aimed to analyze the development, main topics, and future trends of digital medicine over the past 25 years using the example of JMIR's publications in a meta-analysis. For this purpose, network analysis serves as an effective tool to process and visualize complex data and to uncover key features and relationships within interconnected systems [10-12]. Key variables impacting a topic can be identified by mapping keywords, countries, or other study-specific elements as nodes, with their connections as edges in a network. This visualization helps systematically uncover core themes and reveal influential factors and interactions among various research elements. This approach has been effectively applied in dermatological research to analyze trends through congress data [13,14]. This method enriches our understanding of the structure and evolution of a research field and provides new insights into how individual studies contribute to the broader scientific discourse.

In this study, we therefore conducted a comprehensive network analysis of JMIR's publications over the last quarter century to illuminate the evolution, prevailing trends, and future paths of digital health. By thoroughly examining paper types, paper titles, countries, MeSH (Medical Subject Headings) terms, and keywords, we constructed a complex picture of the digital health landscape, highlighting key themes, breakthroughs, and the global distribution of papers. The method aimed to identify critical junctures in the development of digital health, including the rise of telemedicine, the proliferation of mobile health apps, and the introduction of AI. With this, we provide a comprehensive overview of the evolution of the field and insights into emerging trends that will redefine the future of health care.

Methods

Ethical Considerations

This study exclusively used publicly accessible, secondary data from internet sources (PubMed) and did not involve human participants or any sensitive personal information. As such, it did not require ethics board approval.

Data Acquisition

We conducted a comprehensive review of papers published over 25 years (1999-2023) in JMIR, which is fully indexed by PubMed. Using the Biopython (version 1.83; Bioinformatics Open Source Project [15]) library, we retrieved a dataset of 8068 papers as of March 20, 2024, using the search query "("Journal of Medical Internet Research"[Journal])" and the first entry in the PubMed history as the reference year for the following analyses.

MeSH Terms and Keyword Analysis

Following the initial exploratory data visualization, we encountered challenges due to incomplete data and a high frequency of generic terms like "Humans" within the MeSH terms (Multimedia Appendix 1). Consequently, Anthropic's most advanced large language model, Claude 3 Opus (version 20240229 [16]) with the Anthropic Python application programming interface library (version 0.19.1 [17]) as well as Gemini 1.5 Pro (February 2024) using Google's Generative-AI Python application programming interface library [18], was used to derive keywords based on the paper titles and abstracts. We prompted the model to generate a concise JSON list of 6 to 12 significant keywords for each paper. Keywords were adjusted to be concise and in Title Case, as specified by the prompt requirements (Multimedia Appendix 2).

Evaluation of Keyword Metrics

We assessed keyword quality from several sources, including PubMed MeSH terms, author keywords, Claude 3 Opus keywords, Gemini 1.5 Pro (February 2024), and 2 conventional natural language processing (NLP) methods using key bidirectional encoder representations from transformers [19] with the scientific NLP models en_core_sci_lg and en_core_sci_md [20], against the reference derived from paper titles and abstracts in addition to the MeSH terms and author keywords or the MeSH terms and author keywords alone. All text data were converted to lowercase for case-insensitive matching. Keywords were aggregated into a unified set, and presence or absence matrices were constructed for each source. Precision, recall, and F_1-score were calculated to evaluate the effectiveness of each source in identifying keywords.

Statistical Analysis

To account for fluctuations in the number of papers published each year, we first calculated the annual percentage of keyword occurrences and subsequently computed the average of these

XSL•FO
RenderX

annual percentages for each defined period (1999-2007, 2008-2013, 2014-2019, and 2020-2023). This method ensures that years with fewer publications do not disproportionately affect the analysis within each period. The time frames were arbitrarily chosen based on initial exploratory analysis and took into consideration the rising number of publications, which justified the inclusion of more years in earlier time frames. Additionally, the onset of the COVID-19 pandemic in 2020 represented a natural turning point.

Future Trend Prediction

To evaluate the performance of different keyword trend prediction methods, we used various metrics including mean absolute error (MAE), mean squared error (MSE), root-mean-squared error (RMSE), R^2, and mean absolute percentage error (MAPE). The models assessed were Claude 3 Opus, autoregressive integrated moving average (ARIMA), exponential smoothing (ES), Facebook's Prophet, and the Foundation Model Time Series Foundation Model (TimesFM) by Google and tested against the actual keywords (MeSH+author keywords set) for the first 7 months of 2024.

Network Visualization

Network visualization was performed using Gephi (version 0.10.1; Gephi Consortium [21]). Co-occurrence matrices were generated for keywords appearing together in a paper, for combined assignments of different paper types to a paper, and for collaborating locations contributing to a paper. The generation of co-occurrence matrices was done by importing CSV-based data of keywords, paper types, or contributing countries into Gephi, recognizing edges between entities that co-occur on a line-by-line basis, that is, in a paper. Nodes were thus represented by the keywords, paper types, or contributing countries. Edges corresponded to the co-occurrence of nodes within a paper. Node size and labeling were proportional to node frequency. Edge strength, as indicated by the line thickness between 2 nodes, was set in relation to the weight of the edge, that is, the co-occurrence frequency. To ensure readability, we implemented a minimum keyword frequency filter ranging from 3 to 30, depending on the total number of keywords in the 4 subsets of years (see figure descriptions). The network compositions were created using the Fruchterman-Reingold layout [22] and the ForceAtlas algorithm [23]. While the Fruchterman-Reingold layout excels at providing a balanced distribution and minimizing edge crossings, making it suitable for intuitive and interpretable visualization of connected nodes in close proximity, the ForceAtlas layout effectively emphasizes community structures and efficiently handles large networks, making it particularly effective for dense and complex network visualizations. We therefore applied the Fruchterman-Reingold layout to visually cluster related keywords and paper types, where its ability to maintain readability and coherence was critical, and used the ForceAtlas to visualize the network of contributing countries, where the algorithm's strength in cluster detection was particularly useful in revealing the underlying community structure. To ensure readability and optimize visualization, network parameters were manually adjusted, including minimum node size (10.0), maximum node size (100.0), attraction force (15.0), area (10,000.0), node opacity (70.0), edge thickness (3.0), and edge opacity (15.0) for the Fruchterman-Reingold graphs and repulsion strength (200.0), attraction strength (20.0), maximum displacement (10.0), auto-stabilization strength (80.0), attraction force (30.0), node opacity (1000.0), edge thickness (3.0), and edge opacity (30.0) for the ForceAtlas graph. Edges were colored to match the nodes they connected, based on the determined modularity class. Nodes in the same class were assigned the same color. Modularity, which measures the ability of the network to partition into distinct communities [24], was used for this determination.

Results

Publication Trends and Global Contribution Analysis in JMIR

JMIR has experienced significant publication growth over the past 25 years, starting with 15 papers in 1999 and reaching a peak of 1567 in 2020. The journal consistently increased its publication volume each year, exceeding 100 papers per year since 2011. However, there was a decline to 491 papers in 2023. Our analysis included 8068 papers, with an average presence of 9 (SD 4.2) MeSH terms and 6.5 (SD 4.3) keywords per paper. Notably, 212 (2.6%) papers had no MeSH terms, and 1012 (12.5%) papers had no keywords.

Our review of paper types in the publication database reveals that "Journal Article" is the most common category (n=7806, 96.8%), followed by "Research Support, Non-U.S. Gov't" (n=4069, 50.4%), "Randomized Controlled Trials" (n=1076, 13.3%), "Review" (n=987, 12.2%), and "Systematic Reviews" (n=824, 10.2%). In combined article classifications, "Journal Article" paired with "Research Support, Non-U.S. Gov't" is most frequent (n=2109, 26.1%), followed by "Journal Article" alone (n=1995, 24.7%; Figure 1).

Over the years, the distribution of scientific paper contributions by country exhibited significant changes (Table 1). From 1999 to 2007, Canada was in the lead (n=41, 12.8%), followed by the United States (n=18, 5.6%). In the period of 2008-2013, the United States took the lead (n=172, 18.1%, +12.5%), and in 2014-2019, its contribution increased (n=1059, 26.5%, +8.4%). The most recent data from 2020 to 2023 show the United States with a reduced share (n=1378, 20.1%, –6.4%), while China's participation increased (n=609, 8.9%, +3.9%). The United Kingdom (top 2 to 7) and Australia (top 3 to 4) remained consistently among the top contributors over all periods. Emerging contributors such as Singapore (n=119, 1.7%) and Hong Kong (n=140, 2%) highlight the expanding global diversity in research authorship.

The network analysis of the contributing countries over the entire period from 1999 to 2023 delineates the pivotal role of regions characterized by sustained high levels of scientific output and robust collaborations with other regions. Notably, countries such as the United States, the United Kingdom, and China emerge as central nodes within clusters of identical modularity classes. Meanwhile, peripheral nations such as Australia assume a mediating function, attributable to their close

XSL•FO
RenderX

affiliations with adjacent clusters, exemplified by strong connections to the United States, China, and Europe (Figure 2).

Figure 1. Weighted network visualization (Fruchterman-Reingold layout) of all paper types from 1999 to 2023 corresponding to 44 nodes and 205 edges and a modularity of 0.157 with 4 communities found (nodes of the same color correspond to the same community).

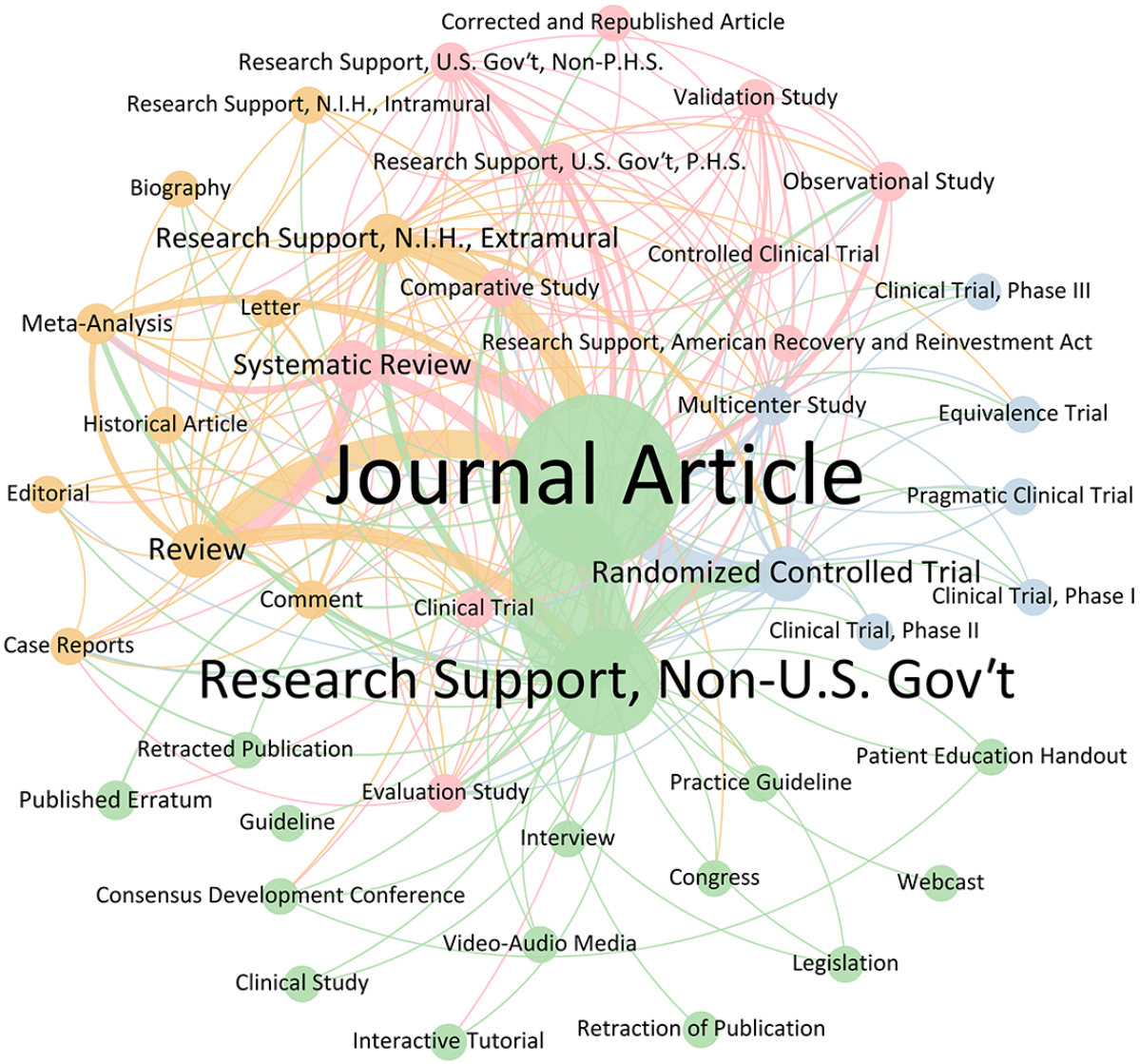

XSL•FO
RenderX

Table 1. Trends in country contribution to JMIR publications (1999-2023)[a].

1999-2007 (n=319)		2008-2013 (n=933, +614, +129.5%)			2014-2019 (n=2864, +1931, +207%)			2020-2023 (n=3952, +1088, +38%)		
Country	Values, n (%)	Country	Values, n (%)	Trend (%)	Country	Values, n (%)	Trend (%)	Country	Values, n (%)	Trend (%)
Canada	41 (12.8)	United States	172 (18.1)	+12.5	United States	1059 (26.5)	+8.4	United States	1378 (20.1)	–6.4
United States	18 (5.6)	Netherlands	122 (12.9)	+8.5	United Kingdom	407 (10.2)	+3.2	China	609 (8.9)	+3.9
Netherlands	14 (4.4)	Australia	80 (8.4)	+5	Australia	306 (7.7)	–0.8	United Kingdom	534 (7.8)	–2.4
Australia	11 (3.4)	Canada	76 (8)	–4.8	Netherlands	250 (6.3)	–6.6	Australia	335 (4.9)	–2.8
Germany	10 (3.1)	United Kingdom	66 (7)	+4.5	Canada	224 (5.6)	–2.4	Canada	330 (4.8)	–0.8
Denmark	8 (2.5)	Sweden	37 (3.9)	+3.3	China	198 (5)	+3.5	Germany	308 (4.5)	0
United Kingdom	8 (2.5)	Germany	24 (2.5)	–0.6	Germany	181 (4.5)	+2	Netherlands	237 (3.5)	–2.8
Greece	7 (2.2)	Norway	22 (2.3)	+0.5	Switzerland	88 (2.2)	+0.1	Switzerland	168 (2.4)	+0.2
Norway	6 (1.9)	Switzerland	20 (2.1)	+0.6	Sweden	83 (2.1)	–1.8	Spain	149 (2.2)	+0.2
Switzerland	5 (1.6)	Japan	14 (1.5)	+1.2	Spain	77 (1.9)	+0.7	Hong Kong	140 (2)	+0.7
Spain	4 (1.2)	China	14 (1.5)	+1.2	Norway	66 (1.7)	–0.7	France	133 (1.9)	+0.4
New Zealand	3 (0.9)	Spain	12 (1.3)	0	France	61 (1.5)	+1	Singapore	119 (1.7)	+0.3
Italy	3 (0.9)	Finland	10 (1.1)	New	Singapore	57 (1.4)	New	Sweden	113 (1.6)	–0.4
Nigeria	2 (0.6)	Italy	10 (1.1)	+0.1	Hong Kong	54 (1.4)	+0.6	Italy	102 (1.5)	+0.6
Niger	2 (0.6)	Hong Kong	7 (0.7)	New	Denmark	51 (1.3)	+0.6	Japan	82 (1.2)	+0.3

[a]This table displays the distribution and trends of country contributions to the JMIR over 4 time periods. The countries are listed in descending order based on the number of publications (n) for each period accompanied by their corresponding percentage (%) of the total publications within that time frame and the trend that is indicating the percentage point change compared to the previous period. Significant shifts include the increasing dominance of the United States, a marked rise in publications from China in the most recent period, and the consistent presence of countries such as the United Kingdom, Australia, and Canada throughout the 25 years.

XSL•FO
RenderX

Figure 2. Weighted network visualization (ForceAtlas layout) of country contributions from 1999 to 2023 corresponding to 125 nodes and 1055 edges and a modularity of 0.112 with 9 communities found (countries of the same color correspond to the same community).

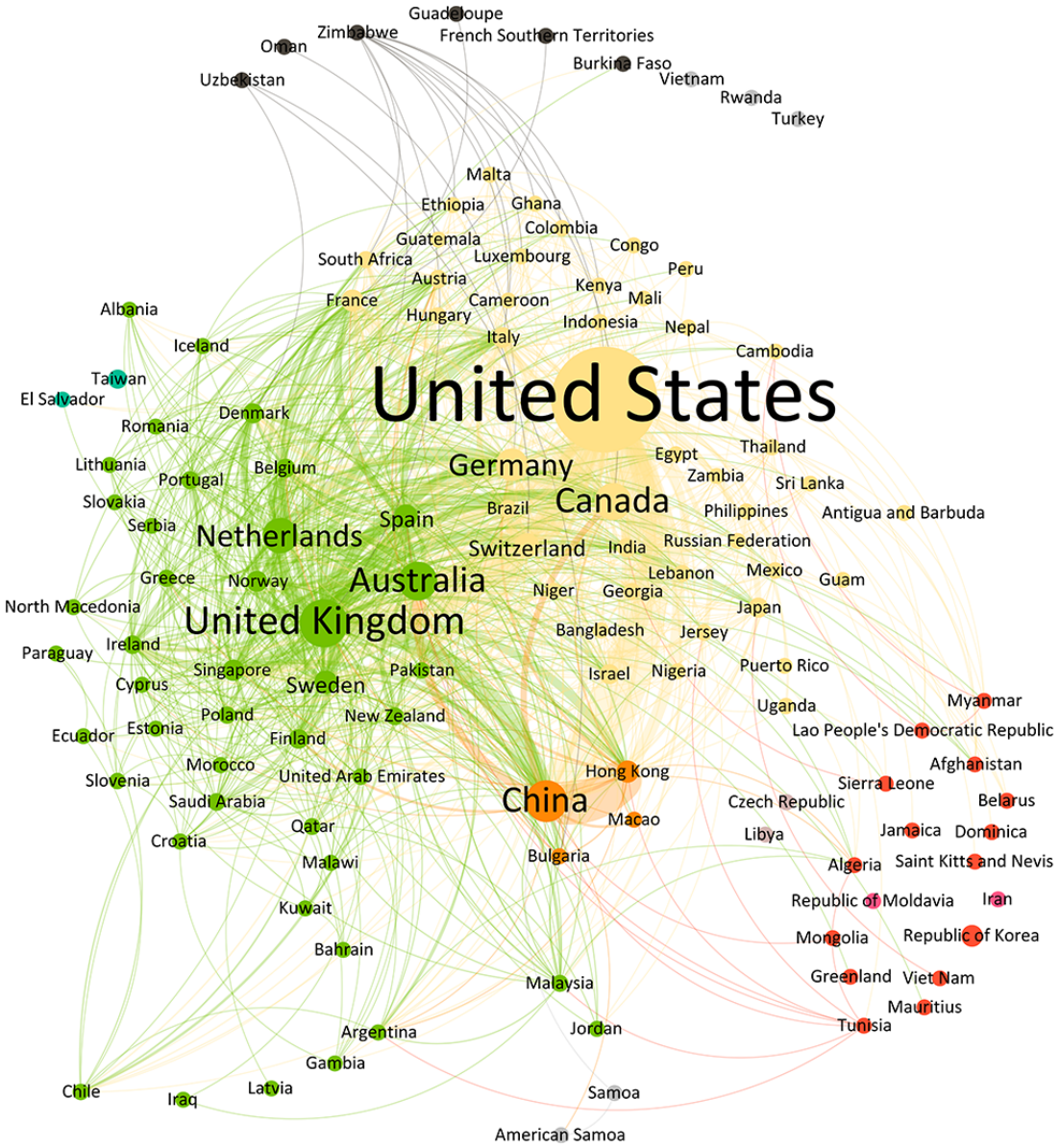

Keyword Benchmark

The comparative analysis of keyword sets from different sources highlighted variations in accuracy, recall, and F_1-scores. The Claude 3 Opus keyword set demonstrated the highest accuracy (0.84) and F_1-score (0.23), suggesting effective keyword identification compared to Gemini 1.5 Pro and conventional NLP (Table 2). However, it is important to note that these metrics are based on keyword occurrences in at least 1 of the MeSH, keywords, paper titles, or abstract sets, which is only a rough estimate of the true keyword relevance. Despite this limitation, the comprehensive coverage and consistent performance of the extracted Claude 3 Opus keywords and the absence of missing data led to its selection for further analysis.

XSL•FO
RenderX

Table 2. Evaluation of keyword quality across different methods and sources[a].

Category and method	n=7904 (98%)[b]				n=8068 (100%)	
	Reference: MeSH+author keywords[c]+titles+abstracts[d]	Reference: MeSH+author keywords[c]			Keywords per paper, mean (SD)	Words per keyword, mean (SD)
	Accuracy	Accuracy	Recall	F_1-score		
Reference						
MeSH+keywords	1.00	1.00	1.00	1.00	14.52 (5.63)	1.70 (0.38)
Reference (split)						
MeSH	0.99	0.99	0.64	0.78	9.00 (4.20)	1.63 (0.40)
Keywords	0.89	0.89	0.44	0.59	6.51 (4.34)	1.61 (0.79)
Large language model						
Claude 3 Opus keywords	0.84	0.28	0.20	0.23	9.83 (1.29)	1.82 (0.27)
Gemini 1.5 Pro keywords	0.72	0.19	0.14	0.16	9.35 (3.26)	1.87 (0.63)
Conventional NLP[e] using KeyBERT[f]						
spacy-en_core_sci_md	0.80	0.00	0.00	0.00	9.00 (0.07)	2.89 (0.14)
spacy-en_core_sci_lg	0.80	0.00	0.00	0.00	9.00 (0.07)	2.88 (0.14)

[a]Comparison of keyword quality of various methods and sources, including PubMed MeSH terms, author keywords, large language model–generated keywords (Claude 3 Opus and Gemini 1.5 Pro), and conventional NLP methods (using KeyBERT with different models). The reference set is constructed from MeSH terms, author keywords, titles, and abstracts or only MeSH terms and author keywords. The metrics presented include accuracy, precision, recall, and F_1-score. Additionally, the table provides the mean and SD of the number of keywords per paper and the mean number of words per keyword. Claude 3 Opus demonstrates superior performance with higher accuracy (0.84) and F_1-score (0.23) compared to Gemini 1.5 Pro and conventional keyword extraction using KeyBERT.

[b]A total of 7904 (98%) represents the number of papers with available reference categories (MeSH terms, author keywords, or both). This subset was used for calculating accuracy, precision, recall, and F_1-metrics.

[c]For author keywords, an exact case-insensitive string match between the generated keyword and any item in the reference list (MeSH terms or author keywords) was considered correct.

[d]For titles and abstracts, a keyword was considered correct if it appeared as a case-insensitive substring within the paper's title or abstract text.

[e]NLP: natural language processing.

[f]KeyBERT: key bidirectional encoder representations from transformers.

Keyword Trend Analysis With Claude Opus 3 (1999-2023)

The keyword analysis from 1999 to 2023 showed significant thematic shifts across 4 periods (Table 3):

- 1999-2007 (n=319 papers): The primary keywords are "Internet" (n=121, 41.9%) and "Health Information" (n=50, 16.7%), followed by "Randomized Controlled Trial" (n=20, 5.2%) and "eHealth" (n=18, 5%), highlighting early focuses on basic internet applications in health care (Figure 3A).
- 2008-2013 (n=933 papers): Trending keywords comprise "Randomized Controlled Trial" (n=129, 13.9%, +8.7%), "Physical Activity" (n=62, 7.6%, +5.4%), "Web-Based Intervention" (n=78, 7.6%, +5.8%), and "Depression" (n=62, 6.9%, +4.3%), whereas the more generic term "Internet" (n=91, 12.3%, –29.5%) is on the decline (Figure 3B).
- 2014-2019 (n=2864 papers): "Social Media" (n=356, 12.8%, +8.9%) and "Randomized Controlled Trial" (n=302, 11.1%, –2.8%) were the most extracted keywords, though the latter with a slight decrease. "Systematic Review" (n=197, 6.7%, +3.4%) and "Mental Health" (n=183, 6%, +1.9%) are under the trending keywords (Figure 3C).
- 2020-2023 (n=3952 papers): The impact of the pandemic led to "COVID-19" (n=980, 22.9%, +22.9%) being the most prominent keyword, accompanied by related terms like "Pandemic" (n=559, 12.8%, +12.8%), "Social Media" (n=460, 11.7%, –1.1%), and "Digital Health" (n=350, 9.4%, +5.5%; Figure 3D). The integration of advanced technology in medical research in the form of "Artificial Intelligence" is on the rise (n=259, 7.6%, +6.6%; Figure 3D).

XSL•FO
RenderX

Table 3. Future trend prediction benchmark for the year 2024 across several methods[a].

Model	MAE	MSE	RMSE	MAPE
Claude 3 Opus	*1.85* [b]	*8.63*	*2.94*	80.09
TimesFM	1.86	9.24	3.04	*79.03*
ARIMA	1.90	8.85	2.98	82.36
ES	1.96	9.03	3.01	86.13
Prophet	4.41	217.90	14.76	232.30

[a]This table presents the benchmark results of various models for future keyword trend prediction. The performance metrics include mean absolute error (MAE), mean squared error (MSE), root-mean-squared error (RMSE), and mean absolute percentage error (MAPE). Claude 3 Opus, Time Series Foundation Model (TimesFM), autoregressive integrated moving average (ARIMA), exponential smoothing (ES), and Prophet were evaluated against the actual keyword data (the MeSH and author keywords are considered ground truth) for the first 7 months of 2024. Claude 3 Opus exhibited the lowest errors across most metrics, indicating its superior accuracy in predicting keyword trends.

[b]Best results are marked in italics format.

Figure 3. Weighted network visualization (Fruchterman-Reingold layout) of the most frequent paper keywords in the 4 periods (A) 1999-2007, (B) 2008-2013, (C) 2014-2019, and (D) 2020-2023 with a minimum keyword frequency of (A) 3, (B) 9, (C) 25, and (D) 30. This corresponded to (A) 1675 nodes (n=201, 12% visible), 11,600 edges (n=1624, 14% visible), and 5 found communities with a modularity of 0.32; (B) 31,400 nodes (n=157, 0.5% visible), 69,900 edges (n=2796, 4% visible), and 5 found communities with a modularity of 0.28; (C) 13,900 nodes (n=139, 1% visible), 113,575 edges (n=4543, 4% visible), and 5 found communities with a modularity of 0.2; and (D) 16,300 nodes (n=163, 1% visible), 147,575 edges (n=5903, 4% visible), and 4 found communities with a modularity of 0.2.

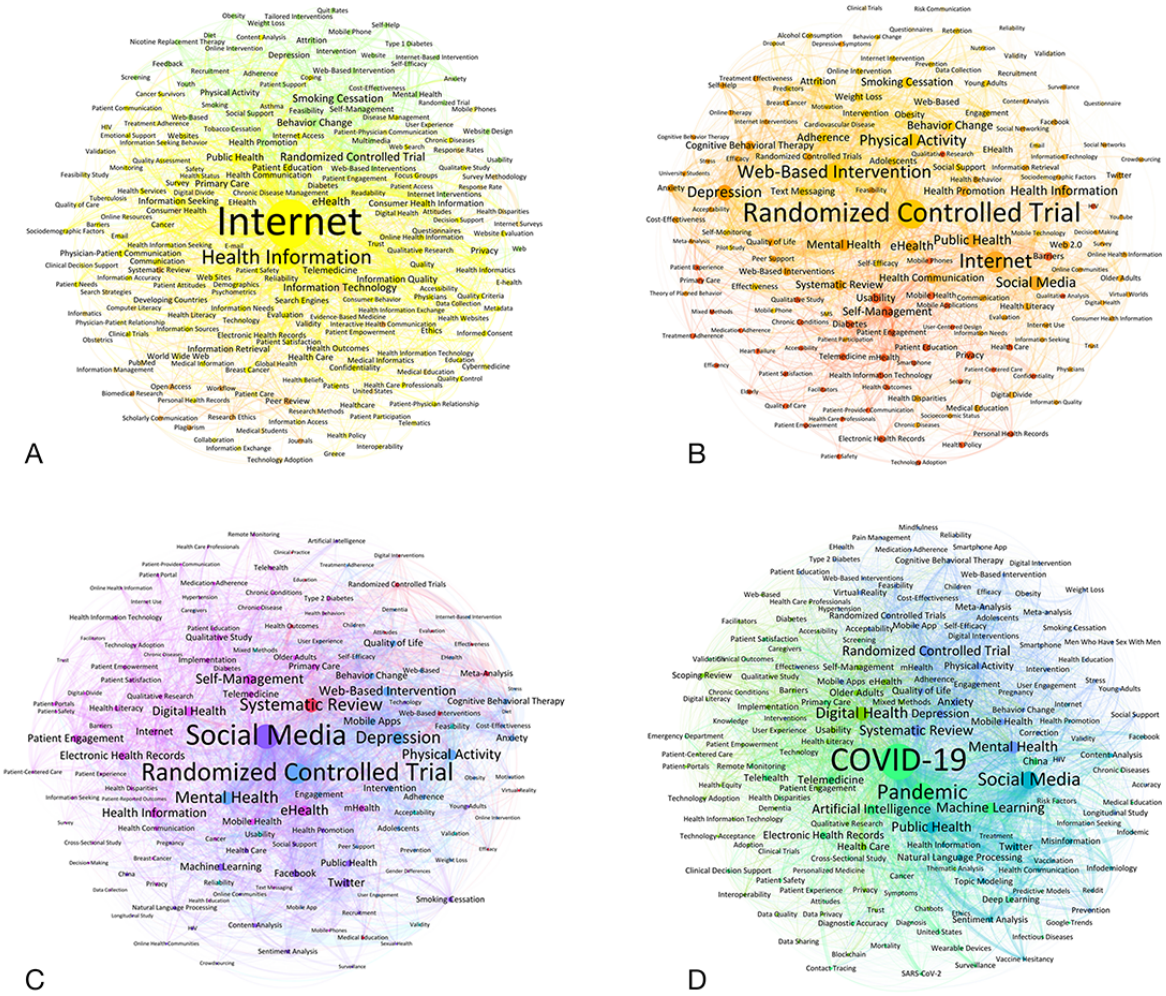

Future Trend Prediction Benchmark

Claude 3 Opus achieved the lowest MAE at 1.85, the lowest MSE at 8.62, and the lowest RMSE at 2.94. TimesFM showed the best performance in terms of MAPE with a value of 79.03, slightly outperforming Claude 3 Opus in this metric. ARIMA, ES, and Prophet exhibited higher errors across all metrics compared to Claude 3 Opus and TimesFM. Specifically, Claude 3 Opus demonstrated significant improvements over other models. Compared to ARIMA, Claude 3 Opus showed a 2.59%

XSL•FO
RenderX

improvement in MAE, a 2.55% improvement in MSE, a 1.28% improvement in RMSE, and a 4.04% improvement in MAPE. When compared to ES, Claude 3 Opus had a 5.59% improvement in MAE, a 4.50% improvement in MSE, a 2.28% improvement in RMSE, and an 8.24% improvement in MAPE. The Prophet model lagged considerably behind, with Claude 3 Opus showing improvements of 57.98% in MAE, 96.04% in MSE, 80.10% in RMSE, and 65.98% in MAPE. Although TimesFM performed slightly better in terms of MAPE, Claude 3 Opus still showed a minimal 0.61% improvement in MAE, a 6.65% improvement in MSE, and a 3.38% improvement in RMSE over TimesFM (Table 3).

Predicted Keyword Trends by Claude Opus 3 (2024-2026)

For the period from 2024 to 2026, we have leveraged Claude 3 Opus, which emerged as the top-performing model in 3 of 4 categories in our 2024 future trend prediction benchmark (Table 3). Claude 3 Opus predictions indicate an increased focus on “Artificial Intelligence” (17.8%, +10.2%) and “Digital Health” (13.2%, +3.8%), as also seen in the trending keywords “ChatGPT” (11.8%, +10.6%), “GPT-4” (7.8%, +7.4%), “Large Language Models” (9.1%, +8.3%), and “Natural Language Processing” (7.5%, +3.7%). There is a further reduction of “COVID-19” specific research (10.2%, –12.7%). “Public Health” is expected to decline (5.6%, –1.1%; Table 4).

XSL•FO
RenderX

Table 4. Evolution of keyword prevalence in JMIR publications (1999-2026)[a].

Claude 3 Opus–derived keywords from PubMed paper titles and abstracts											Prediction		
1999-2007 (n=319)		2008-2013 (n=933)			2014-2019 (n=2864)			2020-2023 (n=3952)			2024-2026[b]		
Keyword list	Values, n (%)	Keyword list	Values, n (%)	Trend (%)	Keyword list	Values, n (%)	Trend (%)	Keyword list	Values, n (%)	Trend (%)	Keyword list	Values, %	Trend (%)
Internet	121 (41.9)	Randomized Controlled Trial	129 (13.9)	8.7	Social Media	356 (12.8)	8.9	COVID-19	980 (22.9)	22.9	Artificial Intelligence	17.8	+10.2
Health Information	50 (16.7)	Internet	91 (12.3)	–29.5	Randomized Controlled Trial	302 (11.1)	–2.8	Pandemic	559 (12.8)	12.8	Digital Health	13.2	+3.8
Randomized Controlled Trial	20 (5.2)	Physical Activity	62 (7.6)	5.4	Depression	196 (6.8)	–0.1	Social Media	460 (11.7)	–1.1	Machine Learning	12.7	+5.1
eHealth	18 (5)	Web-Based Intervention	78 (7.6)	5.8	Systematic Review	197 (6.7)	3.4	Digital Health	350 (9.4)	5.5	Social Media	12.1	+0.4
Information Quality	12 (4.8)	Depression	62 (6.9)	4.3	Mental Health	183 (6)	1.9	Systematic Review	288 (7.6)	0.9	ChatGPT	11.8	+10.6
Information Technology	17 (4.7)	eHealth	47 (6.8)	1.8	Physical Activity	150 (5.8)	–1.8	Machine Learning	281 (7.6)	4.6	COVID-19	10.2	–12.7
Public Health	14 (4.5)	Health Information	48 (5.8)	–10.9	eHealth	149 (5.3)	–1.5	Artificial Intelligence	259 (7.6)	6.6	Systematic Review	9.6	+2
Smoking Cessation	18 (4.4)	Public Health	52 (5.2)	0.7	Web-Based Intervention	132 (5.1)	–2.5	Mental Health	296 (7.4)	1.4	Large Language Models	9.1	+8.3
World Wide Web	8 (3.9)	Behavior Change	41 (5.1)	1.5	Self-Management	146 (5)	1.2	Randomized Controlled Trial	272 (7.3)	–3.8	Randomized Controlled Trial	8.5	+1.2
Information Retrieval	10 (3.9)	Smoking Cessation	39 (4.5)	0.1	Health Information	126 (4.7)	–1	Public Health	285 (6.7)	3.4	Mental Health	8.2	+0.8
Telemedicine	12 (3.9)	Health Promotion	34 (4.3)	1.7	Twitter	115 (4.4)	3.2	Depression	206 (5.4)	–1.3	GPT-4	7.8	+7.4
Primary Care	13 (3.9)	Mental Health	42 (4.1)	2.3	Internet	99 (4.2)	–8.2	Telemedicine	186 (4.5)	1.7	Natural Language Processing	7.5	+3.7
Behavior Change	15 (3.6)	Web 2.0	25 (4)	+4	Digital Health	127 (3.9)	3	Electronic Health Records	166 (4.3)	0.7	Depression	7.1	+1.7
Patient Education	11 (3.2)	Social Media	57 (3.9)	+3.9	Facebook	92 (3.7)	2.6	Anxiety	171 (4.2)	1.4	Health Care	6	+2.1
Ethics	8 (3.2)	Health Communication	34 (3.8)	1.7	Patient Engagement	105 (3.6)	1.3	Twitter	159 (4.1)	–0.3	Public Health	5.6	–1.1

XSL•FO
RenderX

Claude 3 Opus–derived keywords from PubMed paper titles and abstracts											Prediction		
1999-2007 (n=319)		2008-2013 (n=933)			2014-2019 (n=2864)			2020-2023 (n=3952)			2024-2026[b]		
Keyword list	Values, n (%)	Keyword list	Values, n (%)	Trend (%)	Keyword list	Values, n (%)	Trend (%)	Keyword list	Values, n (%)	Trend (%)	Keyword list	Values, %	Trend (%)
Health Care	10 (3.2)	Self-Management	40 (3.8)	1.8	Electronic Health Records	109 (3.5)	1.2	Health Care	142 (3.9)	2.1	Microsoft Bing Chat	5.2	+5.1
Privacy	10 (3.2)	Usability	34 (3.7)	3	Mobile Apps	97 (3.4)	3	Natural Language Processing	132 (3.8)	2.4	Telehealth	5	+1.9
Information Seeking	11 (3.1)	Adherence	38 (3.7)	2.3	Behavior Change	89 (3.3)	–1.7	China	146 (3.3)	1.9	Deep Learning	4.8	+2
Consumer Health Information	10 (2.8)	Web-Based Interventions	24 (3.6)	2.1	Public Health	89 (3.3)	–1.9	Physical Activity	134 (3.3)	–2.5	Anxiety	4.7	+0.5
Search Engines	8 (2.7)	Obesity	26 (3.4)	2.7	Intervention	88 (3.2)	0.8	Older Adults	126 (3.2)	1.1	Electronic Health Records	4.7	+0.4

[a]This table presents a longitudinal analysis of keyword prevalence (generated from paper titles and abstracts by Claude 3 Opus) in the JMIR across 5 distinct periods from 1999 to a predictive analysis for 2024-2026. For each period, the top-ranking keywords are listed by their occurrence frequency (n), alongside the mean percentage (%) of papers they were featured in. The keywords are listed in descending order based on the mean percentage of keyword occurrence. Trends are calculated as the percentage point change from the previous period, reflecting shifts in research focus. Key observations include the initial prominence of "Internet," which gave way to "Randomized Controlled Trial" and later to "COVID-19" during the pandemic, with "Artificial Intelligence" predicted to lead in the coming years.

[b]For the 2024-2026 period, no occurrence frequency (n) is provided as Claude 3 Opus, without the knowledge of the total frequency, only provided a relative distribution in percentages.

Discussion

Principal Findings

The main objective of this study was to analyze the digital transformation in the field of medicine over the last 25 years using JMIR as an example. Our analysis offers a detailed view of the journal's growth and the shifting focus of its papers.

JMIR has seen a steady increase in publications annually, peaking at 1567 papers in 2020. This trend underscores the journal's expanding role in the digital health field and aligns with the broader rise in digital medicine technologies and applications, particularly from 2010 onward with advances in digital health and mobile health [25]. The surge in 2020 is attributed to the COVID-19 pandemic, which spurred an urgent expansion in digital health solutions like remote monitoring [26] as well as general changes in handling medical research funding, for example, "research community mobilized to submit and review grants more rapidly than ever before" [27].

The peak reflects JMIR's pivotal role in publishing timely research during this crisis. The decline in papers by 2023 suggests a possible stabilization or emerging challenges in the field, such as funding variations or research topic saturation, a trend that is also becoming apparent globally. For example, the number of publications in PubMed under the search terms eHealth [28] or mobile health [29] also shows a peak value in 2021 with a continuous decline since then. Another possible reason could be the flood of publications on COVID-19 as a global focus of interest. Currently, there is a lack of scientific studies offering explanations for this phenomenon.

Emergence of Internet and eHealth in Medical Research (1999-2007)

During the early years of the JMIR, the focus was primarily on the broad use of the internet in health contexts, evident from "Internet" being the most frequent keyword. There was a strong emphasis on the provision and dissemination of health information ("Health Information") and the use of the internet for interventions, as indicated by the presence of keywords like "Smoking Cessation" and "Behavior Change." Several studies of the population's use of the internet for health information have been conducted during this period [30-32]. The use of the internet for medical research has also already been scientifically investigated, and early warnings have been given about the consequences that problems can become worse [33]. Nowadays, it is controversial, but it has been shown that internet research can lead to a better understanding of symptoms and diagnoses [34].

In particular, the use of questionnaires as a scientific tool was discussed, and it was critically noted that the population still must familiarize itself with the use of the internet [35,36]. As a result, the term eHealth appears and seeks a further definition [37]. Whereas today, it is defined as a set of technologies applied via the internet to improve quality of life and facilitate health

XSL•FO
RenderX

care [38]. This shows the beginnings of digital medicine and scientific research. Working titles such as "The WWW of the World Wide Web: Who, What, and Why?" show the demands on science but also the confrontation with the new topic [39]. Additionally, terms such as "Information Technology" emphasize the technological perspective. This era marks the foundational phase of digital health, characterized by the initial exploration of the internet's applications in health care and patient education.

Behavioral Health and eHealth Expansion (2008-2013)

This era marked a notable shift toward more targeted applications and methodologies in the realm of digital health research. The prominence of "Randomized Controlled Trial" as a leading keyword underscores an intensifying commitment to rigorous scientific evaluation of digital health interventions. The rise of "Web-Based Intervention" and "Social Media" as key terms highlights the broadening of platforms and approaches used in health initiatives, suggesting a move toward more web-based and user-engaged strategies. Especially various works on the use of smartphones show a trend toward individualized digital medicine [40-42]. The decline in the dominance of the "Internet" keyword (from n=121, 41.9% in 1999-2007 to n=91, 12.3%, –29.5% in 2008-2013) indicates its ubiquitous presence rather than a novel aspect to be noted. During this period, mental health received increased attention, with "Depression" emerging as a significant focus area, emphasizing the expansion of digital strategies to address mental health issues.

Social Media, Mental Health Focus, and Systematic Reviews (2014-2019)

The focus shifted significantly toward social media and systematic reviews, indicating a growing recognition of these platforms' impact on public health and research dissemination. Keywords like "Digital Health" and "Electronic Health Records" suggest a deeper integration of advanced technologies in health services. Mental health remained a consistent theme, with "Depression" and "Mental Health" being prominent. The appearance of "Twitter" as a keyword underscores the specific platforms that researchers are studying within the social media context.

Dominance of COVID-19 and Advanced Technologies (2020-2023)

In the latest period, the overwhelming influence of the COVID-19 pandemic is evident, with "COVID-19" and "Pandemic" being the most significant keywords. The emergence of "Machine Learning" and "Artificial Intelligence" as top keywords highlights a leap toward more sophisticated technologies in health research. The continuity of "Digital Health" and "Telemedicine" underscores the accelerated adoption of remote and digital health technologies in response to the pandemic. An increasing focus on "Mental Health" and "Anxiety" reflects the impact of challenging times, including the pandemic and social isolation. Another notable development in the period from 2020 to 2023 is the significant increase in publications from China. This rise underscores China's growing commitment and interest in digital medicine. Even before 2020, there was a steady increase in Chinese research contributions, indicating an increasing prioritization of digital health technologies in Chinese medical research. According to a study by Liang et al [43] published in the *Journal of Healthcare Engineering*, China has a higher publication volume in the medical informatics field's professional journals compared to Western counterparts, with these publications also demonstrating significant practical application value. This trend is also reflected in the increasing integration of digital technologies into China's health care system, affecting not only research but also the practical implementation of innovative digital solutions.

Anticipated AI Focus in Digital Medicine (2024-2026)

The projected keyword trends predicted by Claude 3 Opus signal a stronger emphasis on computational technologies in health care, with "Artificial Intelligence" (17.8%, +10.2%) and "Machine Learning" (12.7%, +5.1%) at the forefront. These trends indicate ongoing innovation in intelligent systems for enhancing health care delivery. "Digital Health" (13.2%, +3.8%) continues to be a mainstay, demonstrating the sector's growth and the further integration of technology in health care practices. The expected prominence of "ChatGPT" (11.8%, +10.6%) and "Large Language Models" (9.1%, +8.3%) highlights the role of advanced language processing in health communication and data analysis, though the dominance of ChatGPT might get diminished through the raise of proprietary and open-source competitors [16,44,45]. While "COVID-19" (10.2%, –12.7%) remains significant, its reduced trend suggests a shift toward a broader, postpandemic research landscape. Steady interest in "Mental Health" (8.2%, +0.8%) reflects the ongoing commitment to addressing psychological well-being through digital means and maybe in coping with pandemic or postpandemic psychological traumata [46,47].

Limitations

This study's findings are subject to several limitations. The keyword generation through the Claude 3 Opus model might not always reflect the full depth of the papers due to the challenges of capturing complex academic content. Consequently, the keywords, though the best according to our benchmark, could include terms with limited relevance. Our analysis is confined to JMIR, which, while a leader in digital health, may have publication and editorial preferences that differ from other journals. As such, the conclusions, particularly regarding trends and foci in digital health research, may not be generalizable across the broader scientific community. Predictive analyses for the period 2024-2026 are inherently speculative and reliant on current trends, which are subject to rapid and unpredictable shifts in the field of digital health technology. These predictions, therefore, should be interpreted with caution. Our focus on JMIR content, predominantly English language and sourced from PubMed, may not represent the global scale of digital health research or capture the full spectrum of contributions from non-English sources and developing regions. This limitation could skew the representation of global research contributions and advancements. Despite these considerations, this study provides a systematic and detailed examination of digital health research evolution over the past 25 years using JMIR as a lens to observe key developments and offering a perspective on the potential future directions of the field.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge support by the Open Access Publication Fund of the University of Freiburg. Grammarly (Grammarly, Inc) and GPT-4 were used for language improvement and general paper revision. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Data Availability

All data generated or analyzed during this study are included in this published paper and in Multimedia Appendices 2 and 3. The raw data used in this study are publicly available through PubMed.

Authors' Contributions

RK, JR, and RM were involved in the conceptualization and methodology of the study. TIW, RK, and RM curated the data, while RK and RM performed the formal analysis. RK was responsible for project administration and software development. JR and RK were responsible for supervision. TIW and RK provided resources for the study. Validation was carried out by RM, TIW, and RK. RK and RM created the visualizations for the manuscript. The writing process involved JR, RK, TIW, and RM, with RK and JR preparing the original draft and TIW and RM taking part in the review and editing process. All 4 authors (RK, TIW, RM, and JR) have directly accessed and verified the underlying data reported in the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Top 10 MeSH (Medical Subject Headings) and keywords for given periods.
[DOCX File , 21 KB - jmir_v26i1e60057_app1.docx]

Multimedia Appendix 2
Claude 3 Opus keyword prediction prompts.
[DOCX File , 15 KB - jmir_v26i1e60057_app2.docx]

Multimedia Appendix 3
All scripts and data used in this study for reproducibility.
[ZIP File (Zip Archive), 11941 KB - jmir_v26i1e60057_app3.zip]

References

1. Eysenbach G. Welcome to the Journal of Medical Internet Research. J Med Internet Res 1999;1(1):e5. [doi: 10.2196/jmir.1]
2. Eysenbach G. 10 years experience with pioneering open access publishing in health informatics: the Journal of Medical Internet Research (JMIR). Stud Health Technol Inform 2010;160(Pt 2):1329-1333. [Medline: 20841900]
3. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
4. Uddin R, Koo I. Real-time remote patient monitoring: a review of biosensors integrated with multi-hop IoT systems via cloud connectivity. Appl Sci 2024;14(5):1876. [doi: 10.3390/app14051876]
5. Kalasin S, Surareungchai W. Challenges of emerging wearable sensors for remote monitoring toward telemedicine healthcare. Anal Chem 2023;95(3):1733-1784. [doi: 10.1021/acs.analchem.2c02642] [Medline: 36629753]
6. Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF. Automated assessment of diabetic retinal image quality based on clarity and field definition. Invest Ophthalmol Vis Sci 2006;47(3):1120-1125. [doi: 10.1167/iovs.05-1155] [Medline: 16505050]
7. Rana M, Bhushan M. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimed Tools Appl 2022:1-39 [FREE Full text] [doi: 10.1007/s11042-022-14305-w] [Medline: 36588765]
8. Rogozea L, Leasu F, Repanovici A, Baritz M. Ethics, robotics and medicine development. 2010 Presented at: Proceedings of the 9th WSEAS International Conference on Signal Processing, Robotics and Automation Citeseer; February 20-22, 2010; Cambridge, United Kingdom p. 1-264 URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=06b4988589b5a07250b9329ce4d024d567ee5d1a
9. Thimbleby H. Technology and the future of healthcare. J Public Health Res 2013;2(3):e28 [FREE Full text] [doi: 10.4081/jphr.2013.e28] [Medline: 25170499]

XSL•FO
RenderX

10. Dwyer T. Network visualization as a higher-order visual analysis tool. IEEE Comput Graph Appl 2016;36(6):78-85. [doi: 10.1109/MCG.2016.117] [Medline: 27893370]
11. Hevey D. Network analysis: a brief overview and tutorial. Health Psychol Behav Med 2018;6(1):301-328 [FREE Full text] [doi: 10.1080/21642850.2018.1521283] [Medline: 34040834]
12. Pokorny JJ, Norman A, Zanesco AP, Bauer-Wu S, Sahdra BK, Saron CD. Network analysis for the visualization and analysis of qualitative data. Psychol Methods 2018;23(1):169-183. [doi: 10.1037/met0000129] [Medline: 28569530]
13. Kaczmarczyk R, King F, Biedermann T, Zink A. What's driving dermatology? Contribution title analysis of the largest German Dermatology Congress 2019. Digit Health 2021;7:20552076211012138 [FREE Full text] [doi: 10.1177/20552076211012138] [Medline: 33996141]
14. Kaczmarczyk R, King F, Tizek L, Biedermann T, Zink A. Visualising the past to plan the future: a network analysis of the largest European dermatology conference. Eur J Dermatol 2021;31(2):161-169. [doi: 10.1684/ejd.2021.4010] [Medline: 34001467]
15. Cock PJA, Antao T, Chang J, Chapman B, Cox C, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25(11):1422-1423 [FREE Full text] [doi: 10.1093/bioinformatics/btp163] [Medline: 19304878]
16. Introducing the next generation of Claude. Anthropic PBC. URL: https://www.anthropic.com/news/claude-3-family [accessed 2024-04-29]
17. anthropics/anthropic-sdk-python. GitHub. 2024. URL: https://github.com/anthropics/anthropic-sdk-python [accessed 2024-04-29]
18. google/generative-ai-python. GitHub. 2024. URL: https://github.com/google/generative-ai-python [accessed 2024-03-11]
19. Grootendorst M. MaartenGr/KeyBERT. GitHub. 2024. URL: https://github.com/MaartenGr/KeyBERT [accessed 2024-08-01]
20. scispacy. GitHub. URL: https://allenai.github.io/scispacy/ [accessed 2024-08-01]
21. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Proc Int AAAI Conf Weblogs Soc Media 2009;3(1):361-362. [doi: 10.1609/icwsm.v3i1.13937]
22. Fruchterman TMJ, Reingold EM. Graph drawing by force‐directed placement. Softw Pract Exp 2006;21(11):1129-1164. [doi: 10.1002/spe.4380211102]
23. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 2014;9(6):e98679 [FREE Full text] [doi: 10.1371/journal.pone.0098679] [Medline: 24914678]
24. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech 2008;2008(10):P10008 [FREE Full text] [doi: 10.1088/1742-5468/2008/10/P10008]
25. Ahmadvand A, Kavanagh D, Clark M, Drennan J, Nissen L. Trends and visibility of "Digital Health" as a keyword in articles by JMIR publications in the new millennium: bibliographic-bibliometric analysis. J Med Internet Res 2019;21(12):e10477 [FREE Full text] [doi: 10.2196/10477] [Medline: 31855190]
26. Bokolo Anthony Jnr. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020;44(7):132 [FREE Full text] [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
27. Chinnery PF, Pearce JJ, Kinsey AM, Jenkinson JM, Wells G, Watt FM. How COVID-19 has changed medical research funding. Interface Focus 2021;11(6):20210025 [FREE Full text] [doi: 10.1098/rsfs.2021.0025] [Medline: 34956595]
28. PubMed eHealth Search. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/?term=ehealth [accessed 2024-04-29]
29. PubMed Mobile Health Search. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/?term=mobile+health [accessed 2024-04-29]
30. Lambert SD, Loiselle CG. Health information seeking behavior. Qual Health Res 2007;17(8):1006-1019. [doi: 10.1177/1049732307305199] [Medline: 17928475]
31. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, Dumitru RC, Pudule I, Santana S, et al. European citizens' use of E-health services: a study of seven countries. BMC Public Health 2007;7:53 [FREE Full text] [doi: 10.1186/1471-2458-7-53] [Medline: 17425798]
32. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007;102(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
33. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why are health care interventions delivered over the internet? A systematic review of the published literature. J Med Internet Res 2006;8(2):e10 [FREE Full text] [doi: 10.2196/jmir.8.2.e10] [Medline: 16867965]
34. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. BJGP Open 2017;1(2):bjgpopen17X100833 [FREE Full text] [doi: 10.3399/bjgpopen17X100833] [Medline: 30564661]
35. Kongsved SM, Basnov M, Holm-Christensen K, Hjollund NH. Response rate and completeness of questionnaires: a randomized study of internet versus paper-and-pencil versions. J Med Internet Res 2007;9(3):e25 [FREE Full text] [doi: 10.2196/jmir.9.3.e25] [Medline: 17942387]
36. Eysenbach G, Wyatt J. Using the internet for surveys and health research. J Med Internet Res 2002;4(2):E13 [FREE Full text] [doi: 10.2196/jmir.4.2.e13] [Medline: 12554560]

XSL•FO
RenderX

37. Oh H, Rizo C, Enkin M, Jadad A. What is eHealth (3): a systematic review of published definitions. J Med Internet Res 2005;7(1):e1 [FREE Full text] [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]
38. da Fonseca MH, Kovaleski F, Picinin CT, Pedroso B, Rubbo P. E-Health practices and technologies: a systematic review from 2014 to 2019. Healthcare (Basel) 2021;9(9):1192 [FREE Full text] [doi: 10.3390/healthcare9091192] [Medline: 34574966]
39. Powell J, Clarke A. The WWW of the World Wide Web: who, what, and why? J Med Internet Res 2002;4(1):e4 [FREE Full text] [doi: 10.2196/jmir.4.1.e4] [Medline: 11956036]
40. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. J Med Internet Res 2013 Jun 14;15(6):e120 [FREE Full text] [doi: 10.2196/jmir.2600] [Medline: 23770578]
41. Dennison L, Morrison L, Conway G, Yardley L. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. J Med Internet Res 2013;15(4):e86 [FREE Full text] [doi: 10.2196/jmir.2583] [Medline: 23598614]
42. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
43. Liang J, Zhang Z, Fan L, Shen D, Chen Z, Xu J, et al. A comparison of the development of medical informatics in China and that in Western countries from 2008 to 2018: a bibliometric analysis of official journal publications. J Healthc Eng 2020;2020:8822311 [FREE Full text] [doi: 10.1155/2020/8822311] [Medline: 33101616]
44. Introducing Meta Llama 3: the most capable openly available LLM to date. Meta AI. URL: https://ai.meta.com/blog/meta-llama-3/ [accessed 2024-04-29]
45. Source: Tiny but mighty: the Phi-3 small language models with big potential. Microsoft. URL: https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/ [accessed 2024-04-29]
46. Kaubisch LT, Reck C, von Tettenborn A, Woll CFJ. The COVID-19 pandemic as a traumatic event and the associated psychological impact on families—a systematic review. J Affect Disord 2022;319:27-39 [FREE Full text] [doi: 10.1016/j.jad.2022.08.109] [Medline: 36089074]
47. Costa ACDS, Menon V, Phadke R, Dapke K, Miranda AV, Ahmad S, et al. Mental health in the post COVID-19 era: future perspectives. Einstein (Sao Paulo) 2022;20:eCE6760 [FREE Full text] [doi: 10.31744/einstein_journal/2022CE6760] [Medline: 35584448]

Abbreviations

AI: artificial intelligence
ARIMA: autoregressive integrated moving average
ES: exponential smoothing
MAE: mean absolute error
MAPE: mean absolute percentage error
MeSH: Medical Subject Headings
MSE: mean squared error
NLP: natural language processing
RMSE: root-mean-squared error
TimesFM: Time Series Foundation Model

Edited by G Eysenbach; submitted 30.04.24; peer-reviewed by CA Almenara, Z Lu; comments to author 03.07.24; revised version received 01.08.24; accepted 31.08.24; published 15.11.24.

Please cite as:
Kaczmarczyk R, Wilhelm TI, Roos J, Martin R
Decoding the Digital Pulse: Bibliometric Analysis of 25 Years in Digital Health Research Through the Journal of Medical Internet Research
J Med Internet Res 2024;26:e60057
URL: https://www.jmir.org/2024/1/e60057
doi:10.2196/60057
PMID:

©Robert Kaczmarczyk, Theresa Isabelle Wilhelm, Jonas Roos, Ron Martin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative

XSL•FO
RenderX

Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges

David Owen[1], MSc; Amy J Lynham[2], PhD; Sophie E Smart[2], PhD; Antonio F Pardiñas[2*], PhD; Jose Camacho Collados[1*], PhD

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

[2]Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, United Kingdom

[*]these authors contributed equally

Corresponding Author:
David Owen, MSc
School of Computer Science and Informatics
Cardiff University
Abacws
Senghennydd Road
Cardiff, CF24 4AG
United Kingdom
Phone: 44 (0)29 2087 4812
Email: owendw1@cardiff.ac.uk

Abstract

Background: Mental health disorders are currently the main contributor to poor quality of life and years lived with disability. Symptoms common to many mental health disorders lead to impairments or changes in the use of language, which are observable in the routine use of social media. Detection of these linguistic cues has been explored throughout the last quarter century, but interest and methodological development have burgeoned following the COVID-19 pandemic. The next decade may see the development of reliable methods for predicting mental health status using social media data. This might have implications for clinical practice and public health policy, particularly in the context of early intervention in mental health care.

Objective: This study aims to examine the state of the art in methods for predicting mental health statuses of social media users. Our focus is the development of artificial intelligence–driven methods, particularly natural language processing, for analyzing large volumes of written text. This study details constraints affecting research in this area. These include the dearth of high-quality public datasets for methodological benchmarking and the need to adopt ethical and privacy frameworks acknowledging the stigma experienced by those with a mental illness.

Methods: A Google Scholar search yielded peer-reviewed articles dated between 1999 and 2024. We manually grouped the articles by 4 primary areas of interest: datasets on social media and mental health, methods for predicting mental health status, longitudinal analyses of mental health, and ethical aspects of the data and analysis of mental health. Selected articles from these groups formed our narrative review.

Results: Larger datasets with precise dates of participants' diagnoses are needed to support the development of methods for predicting mental health status, particularly in severe disorders such as schizophrenia. Inviting users to donate their social media data for research purposes could help overcome widespread ethical and privacy concerns. In any event, multimodal methods for predicting mental health status appear likely to provide advancements that may not be achievable using natural language processing alone.

Conclusions: Multimodal methods for predicting mental health status from voice, image, and video-based social media data need to be further developed before they may be considered for adoption in health care, medical support, or as consumer-facing products. Such methods are likely to garner greater public confidence in their efficacy than those that rely on text alone. To achieve this, more high-quality social media datasets need to be made available and privacy concerns regarding the use of these data must be formally addressed. A social media platform feature that invites users to share their data upon publication is a possible solution. Finally, a review of literature studying the effects of social media use on a user's depression and anxiety is merited.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e59225) doi:10.2196/59225

KEYWORDS

mental health; depression; anxiety; schizophrenia; social media; natural language processing; narrative review

Introduction

Background

The Global Burden of Disease study (1990-2019) reports that anxiety disorders, major depressive disorder, and schizophrenia are the main drivers of years lived with disability and disability-adjusted life years across all age groups worldwide [1]. These mental health conditions are a sizable burden on the global population and public health systems. To help alleviate these problems, early intervention is essential [2].

The experiences of those with mental health disorders are often recounted on social media [3]. More broadly, users of Facebook and Reddit express favorable and adverse life events through the medium of text [4,5], and pictorial expressions of sensitive topics such as illness or hardship are becoming increasingly common through image-focused platforms such as Instagram [6]. As a result, methods that harness social media data for prediction of the mental health status of users have burgeoned [7-9]. Research has also spiked following the COVID-19 pandemic [10] and has become a truly interdisciplinary pursuit involving not only computer scientists but also psychologists, psychiatrists, and neuroscientists [11]. The broad idea behind this field is that models underpinned by artificial intelligence (AI) can "predict" a person's "mental health status" (refer to the study by Chancellor et al [12] for a discussion on the meaning of these terms in this literature). A branch of AI that is most appropriate for these methods is natural language processing (NLP), which uses computational techniques to learn, understand, and produce human language content [13]. Text-based dialogue systems, for example, have become a mainstay of NLP research. Their use in assisting people with neurocognitive disorders or mental health conditions is a popular application area. An early system, ELIZA [14], dates back to 1966. It purported to perform the role of a psychotherapist in conversation with a patient and has influenced the design of modern conversational agents such as ChatGPT [15]. In 2024, the potential for adults with dementia to adopt ChatGPT as a memory aid has been explored; it may be able to provide reminders of names, dates, and events, thus easing anxiousness [16]. The mining of text data to help assess a person's mental state has also followed from pre–21st century work. The *Whissell Dictionary of Affect in Language* [17], compiled in 1989 and now available on the web [18], can be used to estimate the mood conveyed in a body of text. This has given rise to modern methods for predicting the mental health status of social media users. Indeed, the huge volume of human language content available on the web, for example in Facebook and Reddit postings, fits very well the technical constraints of NLP techniques and can be straightforwardly processed into model inputs.

Some of the earliest attempts at predicting the mental health statuses of members of web-based communities were done without AI, through manual review of postings and classic statistical analyses. For example, in November 1999, psychiatrists monitored the general psychiatry subforum of the Norwegian web-based forum Doktoronline [19,20]. They observed that users who wrote negatively about their mental health by expressing sadness or resignation typically received positive and constructive responses from other users. Subsequently, these users often sought social support in their local communities. This corroborated previous findings showing that web-based community participation can have positive, real-life consequences for individuals [21,22], a motivation for later attempts at developing automatic health care intervention methods. Haker et al [23] examined the writings of web-based forum users who self-disclosed diagnoses of schizophrenia. They too noted that users with schizophrenia benefited by receiving advice from other users about medications and approaching health care professionals, as well as by receiving empathy and support.

The advent of social media platforms such as Facebook provided further locations for discussion about mental health disorders. Moreno et al [24] recognized that instances of major depressive disorder (depression hereafter) can be challenging to identify, particularly in older adolescents. So, between 2009 and 2010 they sought Facebook profiles of freshman students whose status updates referenced depression symptoms. Such students were then contacted and those who were willing were clinically screened to determine a diagnosis of depression. Students displaying depression symptoms in their status updates were more than twice as likely to be at risk for depression. Furthermore, the status updates referencing depression symptoms were often found to be a means of gathering support or attention, yet the students showed reluctance in seeking help in person. Thus, it was recognized that Facebook depression disclosures could be harnessed to identify those who might have unmet needs for mental health care. This provided an explicit motivation for improving the methods for predicting this disorder early in its course.

This Study

Due to the large volume of literature that exists in this area, which swelled during the COVID-19 pandemic, a review is timely. In this study, we focused on methods that concern the detection of language features presented in the texts of user social media postings. The main aim of our review was to ascertain state-of-the-art methodologies for detecting linguistic features that can be attributed to mental illnesses. This includes cataloging datasets containing "ground truth" (gold standard) labels of mental health status [12], which are available to help fine-tune these methodologies. Ground truths may be obtained from electronic health records (EHRs), clinical questionnaires, or self-disclosure statements of a mental health diagnosis (eg, "I was diagnosed with depression"). We then examined how these methodologies integrate the temporal stochasticity of mental states as reflected by longitudinal studies. We also identified common technical and ethical constraints met in the

XSL•FO
RenderX

development of the reviewed studies. Finally, we will form recommendations for the future direction of AI-based research on mental health.

Methods

Overview

We used Google Scholar to seek peer-reviewed articles published between January 1999 and February 2024. This literature search engine was selected because it is considered the most comprehensive search engine in academia [25-27]. It offers particularly extensive coverage of computer science and informatics, which is the primary discipline of literature that forms this review, outperforming databases like Scopus [28]. Our search aimed to retrieve literature covering the 3 main mental health burdens reported by the Global Burden of Disease study [1]: depression, anxiety, and schizophrenia, which are all common mental disorders. The articles then underwent a manual selection exercise to assign each of them to 1 of the 4 different subject areas that cover important and distinct aspects around mental health research in social media: datasets on social media and mental health, methods for predicting mental health status, longitudinal analyses of mental health, and ethical aspects of the data and analysis of mental health. These subject areas, described in more detail in Textbox 1, underpin the aims of this review described in the Introduction section.

Textbox 1. The subject areas covered in this narrative review.

Datasets on social media and mental health

To develop methods for predicting mental health status or conducting longitudinal analyses, carefully constructed social media datasets are required. We identify publicly available datasets that support this work and the challenges met in constructing them.

Methods for predicting mental health status

Approaches may consider how to detect mental health disorders in social media users and measure attributes of those disorders, such as their severities. We examine the progress in this area against a backdrop of evolving natural language processing technologies.

Longitudinal analyses of mental health

One's mental health state is fluid. We review attempts to gauge mental health state changes at both an individual level and population level. The former may assist in directing personalized health care to people at risk, while the latter may help inform public health policy.

Ethical aspects of the data and analysis of mental health

Research activities in the domain of predicting mental health status inevitably involve the acquisition and processing of personal data. We study the concerns reported among the general population and how they may be ameliorated.

Literature Search and Selection Strategy

A detailed exposition of the literature search and selection strategy, which is informed by Ferrari [29], now follows.

The search string in Textbox 2 was deployed to search article titles. 4479 articles were returned.

The articles then underwent a 4-stage manual sifting exercise as shown in Figure 1.

Textbox 2. Search string used in Google Scholar.

(intitle:"mental health"

OR intitle:"mental illness"

OR intitle:"mental disorder"

OR intitle:"psychiatric disorder"

OR intitle:"depression"

OR intitle:"anxiety"

OR intitle:"schizophrenia")

AND

(intitle:"social media"

OR intitle:"forum"

OR intitle:"forums"

OR intitle:"facebook"

OR intitle:"twitter"

OR intitle:"reddit")

XSL•FO
RenderX

Figure 1. The 4-stage sifting process for selecting articles for inclusion in the narrative review.

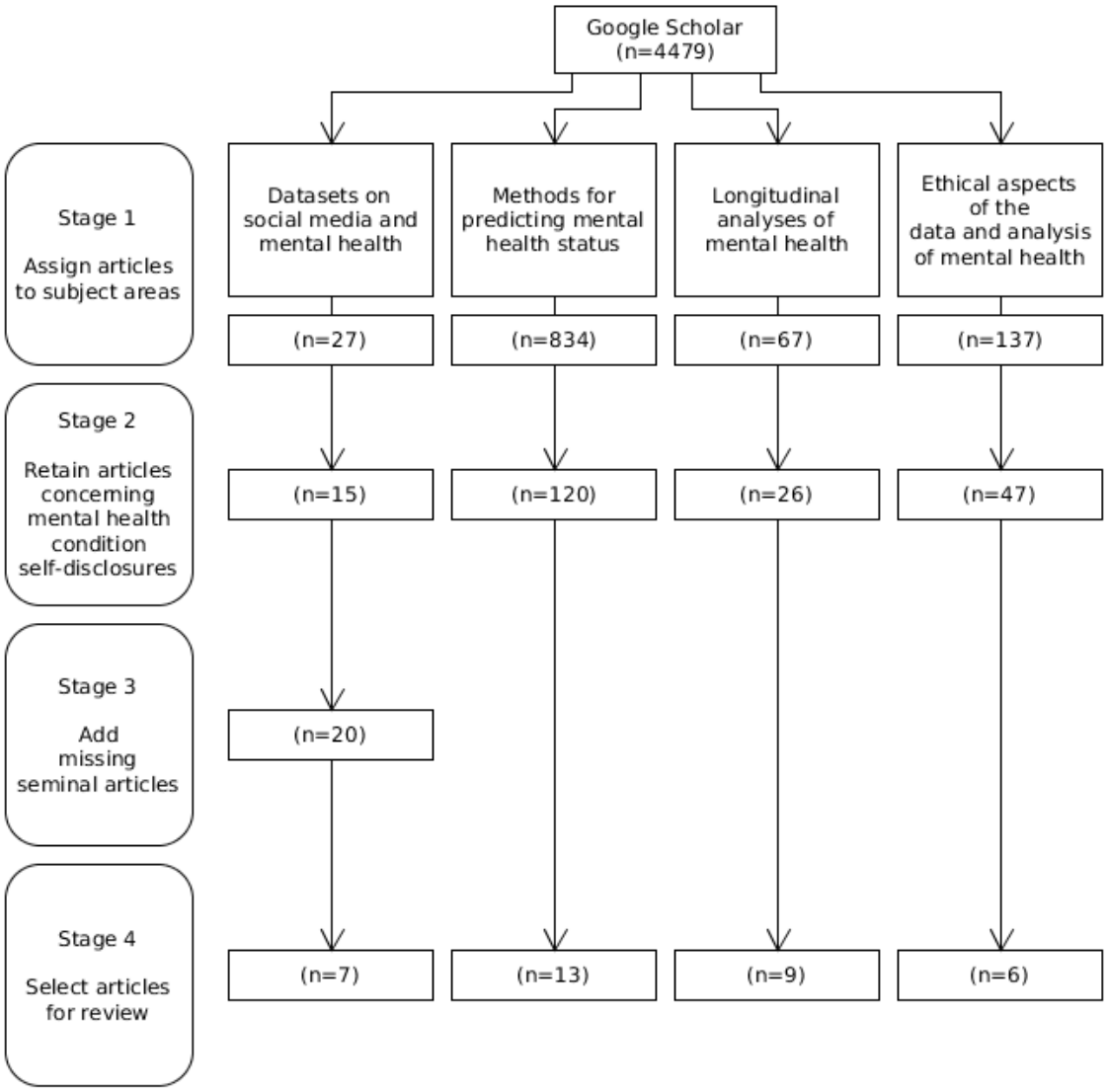

In the first stage, the title and abstract of each article was inspected so that it could be assigned to 1 of the 4 different subject areas shown in Textbox 1. Articles that did not relate to any of the 4 areas of interest were discarded. Duplicate results, preprints, presentation slides, posters, and non–English language articles were also discarded. A total of 1065 articles remained after this stage.

In the second stage, bodies of candidate articles for the 4 subject areas underwent inspection. Only articles where the participants of the studies self-disclosed a diagnosis of depression, anxiety, or schizophrenia were retained. The purpose was to ensure that only studies that used this ground truth were considered.

A third stage was performed that affected only the datasets on social media and mental health subject area. Due to our inherent knowledge of the subject area, we recognized that following the second stage, 5 seminal papers were absent. These articles were not retrieved in the Google Scholar search (Textbox 2). Their omission appeared due to their titles containing only implicit references to social media platforms. For example, the article “RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses” does not explicitly mention that it concerns social media data. Following completion of the third stage, we arrived at 20 articles in this subject area.

A fourth and final stage involved selecting the articles for review in each subject area. The selection process identified the most pertinent articles over a broad timespan that concerned depression, anxiety, and schizophrenia.

A comprehensive listing of the articles considered at each stage of the exercise is provided in Multimedia Appendix 1.

Results

Overview

Following the 4-stage manual sifting exercise, 35 articles across the 4 subject areas were finally selected for review. The content of these articles covered research activity undertaken between 1999 and 2024 and influential events such as the COVID-19 pandemic. Table 1 describes the articles that were finally included for review. The format of this table is drawn from Szeto et al [30]. A narrative review of these articles is presented in the following 4 sections, which cover each of our 4 subject areas.

XSL•FO
RenderX

Table 1. Articles across the 4 subject areas that were selected for review.

Subject area and article title	Year published	Study population	Summary
Subject area: datasets on social media and mental health			
Predicting Depression via Social Media [31]	2013	Posts by 476 Twitter users who self-reported a diagnosis of depression between September 2011 and June 2012	Development of methods for dataset construction via crowdsourcing and quantifying users' depressive language use during the year before their diagnosis
Quantifying Mental Health Signals in Twitter [32]	2014	Posts published between 2008 and 2013 by 6696 Twitter users with a self-stated diagnosis of a mental health disorder: 394 with bipolar disorder, 441 with depression, 244 with PTSD[a], 159 with seasonal affective disorder, and 5728 controls	Development and evaluation of a method for swift and inexpensive capture of data about a range of mental illnesses
Depression and Self-Harm Risk Assessment in Online Forums [33]	2017	Posts published between January 2006 and October 2016 by 9210 Reddit users with a self-stated diagnosis of depression and 107,274 controls	Development and evaluation of a method for recognizing users with depression from their language use alone
RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses [34]	2018	Self-reported depression diagnosis posts by 598 Reddit users published between June 2009 and October 2016	Development of methods for rule-based time extraction of depression diagnosis dates and mental health condition state classification
SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions [35]	2018	Posts published between January 2006 and December 2017 by 385,476 Reddit users with a self-stated diagnosis of a mental health disorder: 10,098 users with ADHD[b], 8783 users with anxiety, 2911 users with autism, 6434 users with bipolar disorder, 14,139 users with depression, 598 users with eating disorders, 2336 with OCD[c], 2894 with PTSD, 1331 with schizophrenia, and 335,952 controls	Development of methods for recognizing self-reported mental health condition diagnoses and obtaining high-quality labeled data automatically, rather than manually
Mental Health Surveillance over Social Media with Digital Cohorts [36]	2019	Randomly selected posts belonging to 48,000 US Twitter users	Development of methods for automatically inferring characteristics, including gender, ethnicity, and location of randomly collected Twitter users
Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview) [37]	2021	Posts published between November 2009 and October 2020 by 80 Reddit users who completed a BDI-II[d] questionnaire	Development of methods for determining the severity of depression in Reddit users
Subject area: methods for predicting mental health status			
Social Media as a Measurement Tool of Depression in Populations [3]	2013	Posts by 117 Twitter users who indicated that they have clinical depression with onset between September 2011 and June 2012 and 157 controls	Development of methods for determining a social media depression index that may serve to gauge levels of depression in populations
Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter [38]	2015	Posts by approximately 2000 Twitter users, of whom approximately 600 self-identified as having been clinically diagnosed with depression	Investigation into the use of topic models to analyze linguistic signals for detecting depression
Quantifying the Language of Schizophrenia in Social Media [39]	2015	Posts published between 2008 and 2015 by 174 Twitter users who self-reported a diagnosis of schizophrenia	Development of methods for analyzing how the language of schizophrenia can aid in identifying and getting help to people with schizophrenia
Recognizing Depression from Twitter Activity [40]	2015	CES-D[e] questionnaire responses and posts by 209 Twitter users	Development of methods for extracting and using features from the activity histories of Twitter users to estimate the presence of depression
A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals [41]	2017	Posts published between 2012 and 2016 by 146 Twitter users who self-disclosed a diagnosis of schizophrenia and 146 controls	Development of methods for combining linguistic features of Twitter content with clinical appraisals to form a diagnostic tool for identifying individuals with schizophrenia

XSL•FO
RenderX

Subject area and article title	Year published	Study population	Summary
Detecting depression and mental illness on social media: an integrative review [42]	2017	43 peer-reviewed articles	Literature review of methods for predicting mental illness using social media
Forecasting the onset and course of mental illness with Twitter data [43]	2017	Posts by 105 Twitter users who had a diagnosis of depression and 99 controls. Also, posts by 174 Twitter users who had a diagnosis of PTSD	Development of models to predict the emergence of depression in Twitter users
A text classification framework for simple and effective early depression detection over social media streams [44]	2019	Posts by 135 Reddit users who have depression and 752 controls	Development of a text classification approach for early risk detection concerning social media users with depression, with an emphasis on explainable AI[f]
Towards Preemptive Detection of Depression and Anxiety in Twitter [45]	2020	Posts by 548 Twitter users who self-disclosed having either depression or anxiety and 4102 controls	Development of an LM[g]-based approach for early detection of depression in Twitter users
A Transformers Approach to Detect Depression in Social Media [46]	2021	Posts by 4000 Reddit users who self-disclosed having depression and 4000 controls	Development of transformer-based models for detecting depression in social media users
Characterisation of Mental Health Conditions in Social Media Using Deep Learning Techniques [47]	2022	77 peer-reviewed articles	Literature review of research concerning DL[h] techniques for identifying various mental health conditions from social media data
Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media [48]	2023	Posts by 3107 Reddit users	Development of methods for generating synthetic social media data for subsequent use in transformer-based language model depression detection
Prompt-based mental health screening from social media text [49]	2024	Posts by 1684 Twitter users who self-reported a diagnosis of depression and 11,788 controls	Development of methods that use LLM[i] prompting as an aid to mental health screening in social media text
Subject area: longitudinal analyses of mental health			
Feeling bad on Facebook: Depression disclosures by college students on a social networking site [24]	2011	Facebook profiles of 200 university students	Development of methods for determining associations between displayed depression symptoms on Facebook and other demographic or Facebook use characteristics
Towards Assessing Changes in Degree of Depression through Facebook [50]	2014	Status updates and survey responses of 28,749 Facebook users collected between June 2009 and March 2011	Development of a regression model to predict users' degrees of depression based on their Facebook status updates
Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language [51]	2017	Posts by 3680 Twitter users with a self-stated diagnosis of a mental health condition: 2271 with generalized anxiety disorder, 687 with eating disorders, 247 prone to panic attacks, 318 with schizophrenia, and 157 who have attempted suicide	An investigation of textual patterns in Tweet sequences occurring over short time windows to ascertain their suitability in quantifying psychological phenomena
Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study [52]	2018	Posts by 203 Twitter users who self-identified as having schizophrenia and 173 controls	An exploration of the feasibility of monitoring web-based discussions about suicide among Twitter users who self-identify as having schizophrenia
Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates [53]	2018	Status updates and depression severity ratings of 29 Facebook users and 49 Twitter users	A study of the associations between depression severity and emotion word expression on Facebook and Twitter status updates
What about Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions [54]	2018	Posts by 585 Twitter users who self-reported a diagnosis of depression and 6596 controls	Development of a method for identifying users with or at risk of depression by incorporating measures of 8 emotions as features from Twitter posts over time, including a temporal analysis of these features

XSL•FO
RenderX

Subject area and article title	Year published	Study population	Summary
Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study [55]	2021	Posts of 2575 Twitter users who self-disclosed a diagnosis of depression and 2575 controls	Development of transformer-based DL language models to identify users with depression from their everyday language and to monitor the fluctuation of their depression levels
Using language in social media posts to study the network dynamics of depression longitudinally [56]	2022	Posts by 946 Twitter users who self-reported the dates of any depressive episodes in the past 12 months and the severity of their current depressive symptoms	An investigation into the association between depression severity and text features in Twitter posts
Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation [57]	2023	Posts by 56 Reddit users who self-reported a diagnosis of depression and 168 controls	An investigation to determine the time points in the posting history of a person with depression, that are most indicative of their depression
Subject area: ethical aspects of the data and analysis of mental health			
Effectiveness of Social Media Interventions for People With Schizophrenia: A Systematic Review and Meta-Analysis [58]	2016	2 peer-reviewed publications	Literature review of the effectiveness of social media interventions for supporting people with schizophrenia
Ethical issues in using Twitter for population-level depression monitoring: a qualitative study [59]	2016	16 Twitter users with a self-reported diagnosis of depression who participated in a series of focus groups and 10 controls	Cross-sectional survey study of public attitudes toward using Twitter data for mental health monitoring
Social media, big data, and mental health: current advances and ethical implications [60]	2016	62 peer-reviewed articles	Literature review of work that uses social media "big data," NLP[j], and ML[k] for mental health surveillance and the ethical considerations therein
Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media [61]	2019	55 peer-reviewed articles	Literature review of how scientific articles represent human research participants in human-centered ML
Ethics and Privacy in Social Media Research for Mental Health [62]	2020	35 peer-reviewed articles	Literature review of research that uses social media data in the context of mental health, with reference to the challenges in relation to consent, privacy, and use of such data
Understanding the Role of Social Media–Based Mental Health Support Among College Students: Survey and Semistructured Interviews [63]	2021	101 US university students aged 18 to 24	Web-based survey followed by semistructured interviews to investigate into whether and how social media platforms help meet university students' mental health needs in terms of the social support that they offer

[a]PTSD: posttraumatic stress disorder.
[b]ADHD: attention-deficit/hyperactivity disorder.
[c]OCD: obsessive-compulsive disorder.
[d]BDI-II: Beck Depression Inventory-II.
[e]CES-D: Center for Epidemiologic Studies Depression Scale.
[f]AI: artificial intelligence.
[g]LM: language model.
[h]DL: deep learning.
[i]LLM: large language model.
[j]NLP: natural language processing.
[k]ML: machine learning.

Datasets on Social Media and Mental Health

To develop methods for predicting mental health status, access to high-quality datasets is essential. De Choudhury et al [31] observed in 2013 that previous research had relied heavily on small, homogeneous samples of individuals who gave retrospective self-reports about their mental health, often via surveys. The authors also recognized that a person's posting activity on social media could provide time-stamped insights into their psychological state. To this end, they used

XSL•FO
RenderX

crowdsourcing to compile a dataset of tweets belonging to 476 Twitter users who self-reported a diagnosis of depression. The data were subsequently used to analyze linguistic and behavioral patterns, such as symptom mentions and diurnal activity, respectively. While the data were deemed high quality by Coppersmith et al [32], they pointed to the limited size and scope of these data in terms of self-reported diagnoses, which needed to be obtained by manual completion of a questionnaire, namely the Center for Epidemiologic Studies Depression Scale screening test. Therefore, they proposed an automated method for labeled dataset construction, which sought self-reports of mental illness diagnoses on Twitter such as "I was diagnosed with depression." Their yield of >5000 different users conveying such statements between 2008 and 2013 indicated that a low-cost and low-resource method for data collection was possible. However, the authors acknowledged some limitations. First, only Twitter users were captured, a sample not likely representative of the general population but in this sense, similar to other social media datasets. Second, it was not possible to verify that the self-stated diagnoses were genuine or captured the same psychopathology as clinical diagnoses. For example, population biobank data have shown self-reported depression to be less heritable (ie, less of its variance in the population can be attributed to genetic factors) than diagnostically ascertained depression [64]. Nevertheless, this approach has ostensibly provided the foundation for several publicly available and widely used mental health datasets.

Yates et al [33] developed the Reddit Self-reported Depression Diagnosis (RSDD) dataset, which contains the posting histories of 9210 users with a diagnosis of depression revealed by self-report statements, like the ones described earlier. Further populated with 107,274 users without depression for control purposes, RSDD has become an often-used resource in the development of methods for predicting depression [65-70]. It has also propagated the development of 2 sister datasets, RSDD-Time [34] and Self-reported Mental Health Diagnoses (SMHD) [35]. The former was conceived by MacAvaney et al [34] after recognizing that research had largely not examined the temporality of mental health diagnoses. They randomly selected 598 posts from the RSDD dataset that contained the self-reported diagnosis statement of a user with depression and manually annotated them to denote when the diagnosis occurred. Owen et al [57] successfully exploited RSDD and RSDD-Time in a longitudinal study that evidenced a relationship between selected time spans before diagnosis and the sentiment a user exhibits in their postings. However, because many of the annotations in RSDD-Time denote that the diagnosis dates of many of the users cannot be estimated with a reasonable degree of accuracy (eg, the user merely stated that their depression diagnosis occurred "in the past"), the findings were predicated on the posting histories of only 56 users with depression. This highlights a need for much larger datasets where the dates of depression diagnoses are denoted with a high degree of accuracy.

SMHD, meanwhile, was born out of a desire for datasets covering a broad range of mental health disorders. It provided a platform for the development of methods concerning not only depression [71,72] but also suicidal ideation [73], schizophrenia [74], and even multiclass experimental setups involving combinations of anxiety, eating disorders, attention-deficit/hyperactivity disorder, bipolar disorder, and posttraumatic stress disorder [7,75-77]. It was also intended that a wider range of higher positive predictive value patterns be used to collect a greater volume of users who were diagnosed. Such patterns detect diagnosis keywords relevant to each disorder, drawn from the *Diagnostic and Statistical Manual of Mental Disorders* [78]. As a result, SMHD contains 20,406 users who were diagnosed and 335,952 matched controls. Despite these strengths, RSDD and SMHD are limited in their scope because they do not include posts made in mental health subreddits. It is recognized that language used in dedicated mental health subreddits systematically differs from the rest of Reddit [79]. This, and the limitation that they used only simple text patterns such as "I was diagnosed with depression" to collect users with mental health disorders, must be consistently considered in research work as it may introduce a bias to any models developed [80].

Other biases also exist in social media data. For example, most social media platforms, including Facebook, Twitter, and Instagram, have more male users [81]. There is also evidence to suggest that people with higher levels of education and household income are more frequent social media users [82]. To address such biases and improve the representativeness of social media datasets, Amir et al [36] considered a cohort-based approach to dataset construction. That is, they developed a demographic inference pipeline, which sought Twitter users and identified their age, gender, ethnicity, and location to create a subsample that was representative of the wider population. They then leveraged an existing model [83] to ascertain the prevalence of depression and posttraumatic stress disorder across the 48,000 users collected. This is in contrast to identifying users based on self-reported diagnosis statement patterns, which as mentioned, is another potential source of bias. The authors proposed that such use of surveillance-based methods could aid the identification of population-level trends in disorder prevalence. However, they also acknowledged that proper evaluation of these patterns would require disentangling the ways in which social media datasets differ from representative samples of the underlying population. In any case, further development and adoption of surveillance-based methods are constrained by privacy and ethical considerations. For example, it would surely require the permission of social media users before their data could be automatically sought and analyzed en masse, particularly in relation to personally identifiable information (age, gender, ethnicity, and health status). We explore these matters in more depth in the Ethical Aspects of the Data and Analysis of Mental Health section.

Finally in this section, we mention the work of the eRisk Lab [37], which touched upon another important dimension in the support of methods for predicting mental health status. Their 2021 dataset, which comprised Reddit posting histories belonging to 80 users, was accompanied by ground truth data that can aid in the development of methods for gauging the severity of depression. Recorded against each user was a completed Beck Depression Inventory-II (BDI-II) questionnaire, which categorizes the severity of their depression (ranging from

XSL•FO
RenderX

minimal to severe). While the dataset proved useful in designing methods for finding associations between language features in the users' postings and their depression severities, the ground truth BDI-II questionnaires provided only the depression severities at the terminuses of the users' posting histories. Because the state of one's mental health is somewhat fluid [34], the dataset may contain users whose depression may have long passed. This is plausible given that one user in the dataset had a posting history spanning >10 years, although it should be noted that this is an anomaly, with the median posting history in the dataset being just >1 year. Furthermore, the dataset's small size in terms of number of users is a major constraint [84,85]. This highlights the difficulty in obtaining copious ground truth data that are traditionally collected via confidential questionnaires.

Table 2 summarizes some important features of the datasets discussed in this section, including the platform, contents, compilation year, acquisition inquiries information, and article title.

Table 2. Datasets discussed in this review that may be obtained from their authors.

Dataset	Platform	Contents	Year compiled	Acquisition inquiries	Article
RSDD[a]	Reddit	116,484 users: 9210 with depression and 107,274 controls	2014	RSDD dataset [86]	Depression and Self-Harm Risk Assessment in Online Forums [33]
RSDD-Time	Reddit	598 users with depression	2018	ir@Georgetown—resources [87]	RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses [34]
SMHD[b]	Reddit	385,476 users: 10,098 with ADHD[c], 8783 with anxiety, 2911 with autism, 6434 with bipolar disorder, 14,139 with depression, 598 with eating disorders, 2336 with OCD[d], 2894 with PTSD[e], 1331 with schizophrenia, and 335,952 controls	2018	ir@Georgetown—resources—SMHD [88]	SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions [35]
2015 Computational Linguistics and Clinical Psychology Shared Task	Twitter	1746 users: 477 with depression, 396 with PTSD, and 873 controls	2015	CLPsych 2015 shared task evaluation [89]	Mental Health Surveillance over Social Media with Digital Cohorts [36]
eRisk 2021 Text Research Collection	Reddit	80 users who completed a BDI-II[f] questionnaire	2021	eRisk 2021 text research collection [90]	Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview) [37]

[a]RSDD: Reddit Self-reported Depression Diagnosis.

[b]SMHD: self-reported mental health diagnoses.

[c]ADHD: attention-deficit/hyperactivity disorder.

[d]OCD: obsessive-compulsive disorder.

[e]PTSD: posttraumatic stress disorder.

[f]BDI-II: Beck Depression Inventory-II.

Methods for Predicting Mental Health Status

Background

The methods covered in this review are supported by machine learning (ML). As there is a broad terminology concerning ML, we introduce the relevant terms in Table 3.

XSL•FO
RenderX

Table 3. Machine learning (ML) terms used in this review.

Term	Description
Data representation	
LDA[a] [91]	A technique that can examine a group of documents and produce a series of words, known as a topic, that characterizes those documents. For example, "anatomy, dissection, genomes" may form the topic of a collection of biomedical documents.
LIWC[b] [92]	A text analysis technique that can infer the emotion conveyed in text (eg, positive or negative).
Ontology [93]	A graphical representation of knowledge that is both human-readable and machine-readable. For example, a biomedical ontology might show how different neurological signs and symptoms may be linked to relevant diseases.
Data augmentation [94]	The methods used to increase the size of a dataset by adding slightly modified copies of existing items in the dataset.
Algorithms	
Supervised learning [95]	A type of ML[c] algorithm analogous to human learning from past experiences to gain new knowledge to improve our ability to perform real-world tasks.
SVM[d] [96]	A supervised ML algorithm that learns by assigning labels to objects and can be used, for example, to recognize fraudulent credit card activity.
Random forest [97]	A supervised ML algorithm that combines the output of multiple decision trees to reach a single result.
DL[e] [98]	A type of ML algorithm (supervised or unsupervised) that can produce complex models from data without features (eg, LIWC) needing to be derived as input.
Pretrained models	
LM[f] [99]	An LM is a probability distribution over words or word sequences. LMs learn to predict text that might come before and after other text and thus are used in tasks such as predicting text when writing an email.
BERT[g] [100]	An LM that examines words within text by considering both left-to-right and right-to-left contexts.
ALBERT[h] [101]	A lightweight alternative to BERT that is suitable for use where less computing power is available.
MentalBERT [102]	An LM designed specifically to aid NLP[i] tasks in the mental health care research community.
MentalRoBERTa [102]	An alternative to MentalBERT that can perform predictions in longer left-to-right and right-to-left contexts.
LLM[j] [103]	Large-scale LM designed for NLP tasks such as producing complex text.
GPT [104]	A family of neural network (in that they mimic the workings of the human brain) models that support AI[k]-driven applications for creating content such as text, images, or sound.
ChatGPT [105]	A chatting robot that can provide a detailed response to a question or instruction.
Performance metrics	
Positive predictive value [98]	Of the instances in a dataset predicted by an ML algorithm to have a certain label, positive predictive value denotes how many of them indeed have that label. This is often referred to as precision in the ML literature.
Sensitivity [106]	Of the instances in a dataset with a particular label, sensitivity denotes how many of them were predicted correctly by an ML algorithm. Sensitivity is also known as recall.
F_1-score [107]	The harmonic mean of positive predictive value and sensitivity.
AUROC[l] [108]	Denotes an ML algorithm's performance in terms of distinguishing between labels.

[a]LDA: latent Dirichlet allocation.
[b]LIWC: Linguistic Inquiry and Word Count.
[c]ML: machine learning.
[d]SVM: support vector machine.
[e]DL: deep learning.
[f]LM: language model.
[g]BERT: Bidirectional Encoder Representations from Transformers.
[h]ALBERT: A Lite Bidirectional Encoder Representations from Transformers.
[i]NLP: natural language processing.
[j]LLM: large language model.
[k]AI: artificial intelligence.
[l]AUROC: area under the receiver operating characteristic.

XSL•FO
RenderX

Traditional ML Approaches

In 2013, methods for predicting mental health status from social media data began to emerge [12] and have often involved interdisciplinary teams of computer scientists and clinical psychologists. De Choudhury et al [3] were proponents of supervised learning methods for predicting depression among populations. Exploiting post-level and user-level features from a crowdsourced Twitter dataset, they developed the social media depression index. To do this they used a support vector machine (SVM). The social media depression index could be used to determine the degree of depression manifested by users in their daily tweets. In a US demographic population study, they observed that women were 1.5 times more likely to express signs of depression on social media than men, which marginally exceeded findings from epidemiological surveys on formal diagnoses that suggest the figure to be 1.3 [109]. The overestimation was linked to the greater emotional expressivity of women [110], suggesting that methods more sensitive to language use could help develop more robust models. Such methods include topic modeling via latent Dirichlet allocation (LDA). While this approach has also been used for predicting depression in Twitter users [32], its results have to be taken cautiously as its dataset, in terms of users with depression and control users, was not deemed a representative sample of the population [38]. Later work used LDA-derived features as input to an SVM classifier to discern between users with depression and control users on Twitter [40]. Although the effectiveness of the topic-driven approach was demonstrated to some extent, only a modest result of 35% sensitivity was achieved. In a similar experimental setup for the prediction of depression in Twitter users [43], another traditional ML algorithm, random forests, was deployed using Linguistic Inquiry and Word Count (LIWC) features derived from post text. A commendable area under the receiver operating characteristic score of 87% was achieved and the method was validated by the collection of the mental health histories of its 204 participants via the Center for Epidemiologic Studies Depression Scale questionnaire. Tsugawa et al [40] acknowledged that emerging deep learning algorithms could well advance the methods in this area and were likely to inform future work. We explore these algorithms in the next subsection, Language Models and Transformers. A contemporaneous review also concluded that advances in NLP and ML were making the prospect of large-scale screening of social media for at-risk individuals a near-future possibility [42]. It also cited 2 studies that were influential in dataset design methods [31,32] that we discussed in the Datasets on Social Media and Mental Health section as being likely to help realize this.

By 2019, interest in methods for early prediction of depression had developed due to the recognition that they could help people receive the health care and social support they need sooner than they otherwise might [44]. Burdisso et al [44] designed an algorithm named SS3 that would calculate the degree to which some given text belonged to a certain category. While it could be generalized to any domain, in this case, it was used to classify depressed and control users of the longhand forum Reddit. It demonstrated superior early risk classification performance across several different experimental settings when compared to baselines computed using more traditional algorithms such as SVM. It also demonstrated significantly faster computation times; approximately 20 times faster than SVM. A further aim of SS3 was to provide explainability [111] for its classification decisions. It could display pertinent excerpts of a user's Reddit text, such as "Fact is, I was feeling really depressed and wanting to kill myself," which may assist clinicians. This transparency could not be gleaned from traditional "black box" algorithms such as SVM. SS3 was also hailed as a low-resource method, because, unlike SVM, it does not necessarily need to process the entire input text before returning its classification decision. However, it was acknowledged that because it examines each word of the input text in a singleton fashion, it would not consider potentially crucial 2-word phrases such as "kill myself" in a classification decision.

Language Models and Transformers

The capabilities of language models (LMs) had become well understood in NLP by the start of the 2020s. So, further to the work conducted by Burdisso et al [44], Bidirectional Encoder Representations from Transformers (BERT) and A Lite BERT (ALBERT) were deployed in an early depression prediction task involving tweets that denoted whether a user was with depression or anxiety or with no disorder [45]. As BERT and ALBERT necessarily consider the context of each word they encounter in a classification task, the consideration of n-word phrases is inevitable, thus addressing a matter highlighted by Burdisso et al [44]. In an experimental setting where users with depression and control users were balanced, an F_1-score of 77% was achieved using BERT, compared to an SVM baseline of 65%. In an imbalanced dataset however, which is a more accurate representation of real-world scenarios where these tools could be applied, BERT achieved an F_1-score of 74% compared to SVM's score of 75%. Malviya et al [46] performed a similar experiment where individual posts in a Reddit dataset would be classified as depressed or nondepressed in nature by BERT and traditional baseline algorithms [46]. Once again, strong BERT performance was observed in a balanced experimental setting, therefore, strengthening evidence that further research is needed before LMs could be deployed for this prediction task in more realistic, imbalanced settings. Suggestions include generating synthetic instances to create balance [112] and resampling [113]. A review of deep learning approaches to mental health prediction [47] that postdates both studies [45,46] echoed the need for further work involving much larger datasets while acknowledging the impact of existing datasets that we have already highlighted [33,35].

Some of the most recent methods have harnessed generative AI, principally using GPT [104]. The arrival of generative AI has enhanced opportunities in this domain. We have already noted that the use of quality data is crucial in the pursuit of methods for predicting mental health status. Such data are often scarce and have given rise to data augmentation techniques [94,114]. A slightly different approach involves synthesizing data derived from existing data [115]. In an annual workshop task, a participating team used ChatGPT to synthesize data that would help develop models for identifying BDI-II-recognized depression symptoms conveyed in Reddit posts [48]. Several

XSL•FO
RenderX

thousand apparently suitable texts were generated. For example, to the BDI-II response "I am so sad or unhappy that I can't stand it," ChatGPT formed the text "I'm so overwhelmed by sadness that I can barely function anymore." However, it was found that models for linking such texts to appropriate BDI-II responses performed more strongly with respect to real data rather than their synthesized counterparts. It was suggested that the synthesized texts were overly detailed and complex, thus confounding LMs used in the subsequent classification exercise. One LM used was MentalRoBERTa [102], which is trained on real Reddit data. More judicious use of ChatGPT such that it produces less detailed texts that are more semantically similar to the BDI-II responses was proposed as follow-up work. A further use of a GPT has been in the automatic trisection [49] of the SetembroBR Twitter corpus of users with depression and control users [116]. The GPT was prompted to label each tweet as having either high, medium, or low relevance to mental health. The labeled dataset was then used as an input to a bag-of-words classifier and its prediction performance was compared with that of a BERT-derived baseline produced by an earlier study [117]. While this approach was markedly low resource and improved the baseline result by 5% in terms of sensitivity, it was acknowledged that improved prompting of the GPT, perhaps by using a more formal definition of depression, might see further improved sensitivity. Therefore, large LM (LLM) supported GPTs have shown potential for aiding mental health prediction in a variety of ways. For that potential to be fully realized, computer scientists need to consider how GPT prompting techniques can be optimized in each context.

Considerations for Schizophrenia

Finally in this section, we examine the literature's coverage of schizophrenia. In a 2015 study by Mitchell et al [39], LDA was applied in a Twitter dataset with the goal of distinguishing between users with schizophrenia and controls. Key findings were that irrealis mood (denoted by the use of uncertain terms such as "think" or "believe") [118] and flat affect (due to lack of emoticon use) [119] were prevalent in the posts of people with schizophrenia. A limitation of their dataset was that users' self-statements of schizophrenia diagnoses could not be verified, which is a problem in this field of research as psychotic symptoms might preclude people from believing in their diagnoses [120,121]. In any case, people with schizophrenia may be reluctant to disclose their diagnoses on social media because they are likely to receive stigmatized responses [122,123]. Birnbaum et al [41] attempted more accurate identification on Twitter using a human-machine partnered approach. Self-reported schizophrenia statements were scrutinized for their authenticity by a psychiatrist and a graduate-level mental health clinician. The ML-derived model subsequently developed was able to distinguish between users with schizophrenia and controls with 87% sensitivity. Despite this, the authors acknowledged that truly confirming the diagnosis of a user who makes a self-disclosure statement is not possible without access to the user's EHRs.

Longitudinal Analyses of Mental Health

Studies discussed so far have tended to predict a person's mental health at a particular point in time. However, a person's mental health state is not static [34]. Indeed, it has been argued that inferences derived from sample-level "snapshots" of mental health states might not lead to reliable predictions of the individual-level variation in these states through time [124]. Therefore, research has also examined temporal profiles of mental health disorders and symptoms. A 2011 study considered US college students' Facebook status updates and their potential for exhibiting content that may reveal symptoms of depression [24]. It was noted that opportunities for recognition and treatment of depression were being missed, particularly among college students [125,126]. Therefore, Facebook, a social media platform that had become well-established among the student population [127], presented innovative opportunities to identify college students at risk. A manual exercise saw the collection of Facebook status updates of 200 students that spanned 1 year. Human annotators then scrutinized each post, denoting a depressive symptom if deemed present according to the *Diagnostic and Statistical Manual of Mental Disorders* criteria [128]. A quarter of profiles exhibited at least 1 depressive symptom (as inferred through the use of terms like "hopeless" or "giving up"). This evidence that Facebook may allow the identification of at-risk students would be a precursor to future longitudinal analyses.

Schwartz et al [50] sought to gauge how the level of depression changes among Facebook users during a calendar year. Their method involved the extraction of 1-to-3-word terms, LDA-derived topics, and LIWC categories from the status updates of >28,000 users. A regression model was developed that indicated a significantly higher degree of depression among users during winter months than in summer months, which is compatible with observations made in the psychiatry literature [129]. A baseline model that considered only the average sentiment across each user's status updates was outperformed in terms of accuracy almost 3-fold, although the optimal model only exceeded 30% [130]. By comparison, Loveys et al [51] conducted experiments predicting mental health statuses during much shorter time spans, hours in fact. Tweets belonging to >2500 users who self-stated a diagnosis of either anxiety or schizophrenia were automatically labeled with either positive, neutral, or negative sentiment. For each user, the changes (or otherwise) in terms of sentiment across 3 subsequent tweets that occur within any 3-hour window were observed. These observations were dubbed "micropatterns." It was noted that users with schizophrenia were less likely to show emotional variability between tweets than control users, which perhaps demonstrates a deficit in affective expression, a known schizophrenia symptom [131]. Users with anxiety were less likely to make consecutive positive tweets than controls, again consistent with psychological findings [132]. However, the micropatterns did not contain sufficient details to indicate the severity of the mental health disorders but enriching the automatic labeling process by considering linguistic features other than sentiment (eg, terms that may be mapped to specific symptoms) may help in this respect.

XSL•FO
RenderX

Emotions and their changing nature over a series of web-based postings have also been studied. Seabrook et al [53] considered whether "emotion dynamics" in Twitter and Facebook may provide early indicators for depression risk. The feasibility of using emotion variability and instability as an indicator of depression severity, measured by the Patient Health Questionnaire-9 [133], was explored. It was hypothesized that self-reported depression severity would be positively associated with negative emotion word variability and instability across status updates. Status updates and depression severity ratings of 29 Facebook users and 49 Twitter users were collected. MoodPrism [134] would gauge the emotion of their status updates and the severity of depression (via Patient Health Questionnaire-9) over a 1-year period. Results suggested that instability in the negative emotion expressed on Facebook provides insight into the presence of depression symptoms for social media users. Also, greater variability of negative emotion expression on Twitter may, in fact, be protective for mental health. However, these observations were constrained by the users' tweets being unavailable for manual inspection due to privacy reasons. Therefore, no manual verification was possible, and the results are essentially unreproducible. Another study from 2018 also considered emotion expressions on Twitter for their use in predicting depression [54]. In total, 8 basic emotions (anger, disgust, fear, happiness, sadness, surprise, shame, and confusion) were sought in the tweets of 585 users with depression across a 4-month period. The average intensity of each emotion was calculated via the EMOTIVE ontology [135] and used in a time-series analysis of each user. This analysis in turn helped build ML-based classifiers for labeling previously unseen Twitter users as being either depressed or not. In the best-performing setup with a random forests classifier, 87% sensitivity using temporal features was achieved compared to 71% using simple LIWC features. This suggests that the changes in an individual's emotions over time show potential in identifying users with depression. Fine-grained consideration of the language used in tweets, such as tentative (eg, "maybe") and temporal-related terms, may not only predict its presence but also its severity [56].

The emergence of transformer-based LMs coincided with the onset of the COVID-19 pandemic. It was no coincidence that interest grew in methods for monitoring population-level depression on social media at that time and that LMs would feature. In one study, tweets dated between March 3 and May 22, 2020, were collected regarding users who self-disclosed having depression [55]. The goal was to develop a model for monitoring the fluctuation of depression levels of different groups as COVID-19 propagated. Using the BERT-like model XLNet [136] and a geographical aggregation of users in the dataset, they demonstrated how depression levels fluctuated between the aforementioned dates in New York, California, Florida, and the United States as a whole. It was observed that depression levels in all 4 geographical areas were similar during the pandemic, with a steady increase after the announcement of the United States National Emergency on March 13, a modest decrease after April 23, followed by a steep increase after May 10. The overall depression score of Florida was substantially lower than the United States average and the other 2 states, possibly because it has a lower depression level overall compared to the average US level irrespective of the pandemic. These findings were constrained by the fact that only Twitter users were considered, who therefore are not fully representative of the population. In a further use of LMs, Owen et al [57] aimed to determine how far in advance of a Reddit user's depression diagnosis their postings were most indicative of their condition. Overall, 56 users with depression and 168 controls were acquired from an intersection of the RSDD [33] and RSDD-Time datasets [34]. BERT and a specialist LM, MentalBERT [102], considered all user posts in increasingly large temporal bands up to 24 weeks (approximately 6 months) before the diagnosis dates of users with depression. The LMs achieved F_1-scores of 0.726 and 0.715, respectively, when 12 weeks of postings were considered, suggesting therefore that the most poignant language used by users with depression occurs in the final 3 months before their eventual diagnosis. The reason for the specialist LM performing less effectively than its general counterpart may be explained by the fact that the former is trained on text found in mental health subreddits, and such postings are not included in RSDD. Findings were tempered by the fact that the diagnosis dates were mere estimates, as explained in the discussion of RSDD-Time in the Datasets on Social Media and Mental Health section. In any case, it was posited that a multimodal classification approach might provide more robust results. For example, a Reddit user's upvotes or downvotes for posts may also be predictive of their mental health state.

We conclude this section by again exploring what the literature has covered in the realm of schizophrenia. Hswen et al [52] investigated the language used by Twitter users with schizophrenia to observe whether it would help assess suicide risk [52]. They examined the frequency of suicide-related tweets, paying particular attention to the times of such tweets. They hypothesized that Twitter users who self-identify as having schizophrenia would be significantly more likely to post tweets containing suicide terms when compared to Twitter users from the general population, thereby reflecting the elevated risk of suicide observed among individuals with schizophrenia in real-world settings. The tweets of 203 users with schizophrenia and 173 control users covering a 200-day period were collected. Only tweets that contained the words suicide or suicidal were targeted because, perhaps not surprisingly, the term suicide is frequently contained in suicide-related conversations [137,138]. Crucially, the time of day of each tweet was recorded. A logistic regression model predicted that the users with schizophrenia showed significantly greater odds of tweeting about suicide compared with control users (odds ratio 2.15, 95% CI 1.42-3.28). Considering the times of tweets, the frequency of conversations about suicide on Twitter correlated significantly with discussions about depression and anxiety, another trend that is consistent with established data [139,140]. However, similar to the studies discussed previously [39,41], the inability to be able to verify the diagnoses of the users with schizophrenia was cited as a main limitation.

XSL•FO
RenderX

Ethical Aspects of the Data and Analysis of Mental Health

When constructing datasets, developing methods, and performing longitudinal analyses to aid mental health prediction, people's privacy ought to be considered. In 2016, Mikal et al [59] sought to determine the attitudes of Twitter users toward the platform's use in population health monitoring. Their qualitative study focused on depression. A focus group was formed of Twitter users, some of whom had previously received a diagnosis of depression while others had not. The group was canvassed for their opinions on the prospect of machine-driven health monitoring and their privacy expectations thereon. Broadly speaking, participants were supportive of the use of publicly available data for health monitoring activities, provided that the user identities were concealed. The concerns about the reliability of methods that use crude keyword searches and the misleading findings they could yield were also noted. An incorrect labeling of depression for a user whose identity is revealed would be considered stigmatizing according to participants. The study was only indicative because the group comprised just 26 Twitter users of a narrow demographic (predominantly male with an average age of 26.9 years). However, a concurrent study by Conway and O'Connor [60] gleaned further evidence of fears regarding such stigmatization.

Nicholas et al [62] address similar privacy matters. They note that the introduction of the General Data Protection Regulation in Europe and popular scandals such as Cambridge Analytica's use of Facebook data brought data privacy into sharp focus. User concerns are many and varied. Some users fear that the research findings may affect credit card applications [141] and employment prospects and attract stigma [142]. Fears are compounded by evidence that deidentified data can be reidentified using materials published alongside research articles [143]. Indeed, the desire for anonymity appears particularly widely held, which echoes the findings by Mikal et al [59] and is reinforced by Vornholt and De Choudhury [63]. Therefore, obtaining explicit user consent for the use of their data is considered crucial. A possible route is via acceptance of social media platform terms and conditions. However, as these may not be read and understood [144], this may not constitute informed consent. One solution is to explicitly invite users to donate their social media data for research purposes [145]. Another proposal is a feature that enables users to opt in or out of their data being used as they post it [146].

A matter has also been identified regarding the terminology used in this area of mental health research. Chancellor et al [61] reviewed how human participants are referred to in literature for predicting mental health status using social media data. Common traits were seen across 55 articles. For example, introductions often refer to human participants as "individuals" and "people," but technical sections then refer to them as "samples" and "data," respectively. It is argued that this may present risks to scientific rigor and the populations the research aims to help. Inconsistent terminology may cause misunderstandings regarding study design thus affecting reproducibility of results. Depersonalization and dehumanization may be another byproduct [147]. This may cause individuals and communities to become stigmatized, echoing the findings of the studies discussed earlier. To alleviate this, it is suggested that more human-centered methods such as participatory design should be considered where interviews and field studies are conducted. However, this is at odds with the challenges highlighted in the Datasets on Social Media and Mental Health section where acquiring sizable datasets through such methods is largely intractable [32].

With respect to schizophrenia, Välimäki et al [58] determined via their review that the perceptions and risks of social media interventions are largely unexplored. However, there are suggestions that some clinicians fear that the use of web-based peer support without professional moderation may cause anxiety in the bearer of the disorder [148]. Cognitive deficits in people with schizophrenia can inhibit the development of digital skills [149], evidencing clinicians' misgivings.

Discussion

Principal Findings

We have seen that there is growing interest in methods for predicting mental health status using social media data, particularly those that involve NLP. Enthusiasm has been notable since the COVID-19 pandemic when interest in remote monitoring of individual- and population-level mental states grew. Indeed, the search strategy followed for this review yielded more articles in the years 2020 to 2021 than in the previous 20 years; 917 and 903, respectively (Multimedia Appendix 1). Methods have progressed from those that use features from text as input to traditional ML algorithms, to increasingly sophisticated approaches using transformer-based LMs and now, LLMs. The research community has endeavored to provide social media data to support this work and to do so in ways that are increasingly sensitive to ethical and privacy concerns of the participants involved.

Our review has not only shown depression to be the most common condition reported in publicly available datasets, it also highlights the need for much larger samples where contextual information on this and other conditions, such as a date for the diagnosis and not just its presence, is denoted to a high degree of accuracy. Having such data would likely strengthen results found in longitudinal studies, most of which have focused on depression as well, providing more opportunities for predictions before an eventual diagnosis is formalized [57]. Obtaining such ground truth data via traditional confidential questionnaires is time-consuming and intrusive from the participant's point of view [134]. A solution may involve obtaining consented access to EHRs to accompany the users' social media postings, as piloted by Eichstaedt et al [150]. Indeed, this means of verification is crucial in studies that consider schizophrenia because diagnosis self-disclosure statements, although having high sensitivity [151], may lack specificity [120,121]. In any case, social media data obtained also needs to be broadened to better support NLP methods. For example, Reddit datasets should routinely include postings from mental health subreddits in addition to other subreddits [79]. This would help ensure that LMs pretrained on such data are less prone to biases that may dampen the effectiveness of methods developed thereon [80]. LLM-driven technologies,

XSL•FO
RenderX

such as ChatGPT and its successors, will likely underpin methods in the immediate future. However, a fledgling attempt involving Reddit posts found that models were better able to detect BDI-II-measured depression symptoms using authentic data rather than LLM (GPT-3) synthesized data [48]. It was suggested that improved prompt manipulation is needed to produce synthesized data that are less stilted. Another role for LLMs may be in the automatic labeling of mental health dataset instances. Ramos dos Santos and Paraboni [49] produced evidence that an LLM (GPT-3.5) can perform promisingly (72% sensitivity) when distinguishing between tweets of users who may have depression and users who likely do not. LLMs may eventually offer a far less costly alternative to dataset labeling than manual approaches. Psychiatry literature suggests LLM performance in these settings could be improved by prescribing potentially time-consuming trials to learn what prompts are best suited for specific tasks [152,153]. Instruction fine-tuning is one such proposal for improving LLM performance. LLMs including GPTs are trained on very large, nondomain specific datasets such as Wikipedia. However, further training an LLM on smaller, domain-specific datasets may enhance its performance in that domain. For example, when comparing the performance of a nonfine-tuned LLM and its fine-tuned counterpart, Xu et al [154] measured a 23.4% increase in accuracy across 6 different mental health prediction tasks involving Reddit data. However, fine-tuning ought to be performed using a wider range of domain-specific datasets, which is advisable to reduce biases in the resulting LLM.

With respect to population-level and individual-level longitudinal studies, we found the analysis of emotions conveyed in social media posts to be an underrepresented topic of research in this area. Consideration of fine-grained language features may also help to better predict depression severity over time [56]. In fact, the most promising approaches will probably involve those that augment NLP; multimodal methods that consider nontext features from social media activities are expected to help provide richer findings. In Twitter, this may involve consideration of user geolocations and profile images. For example, Ghosh et al [155] attempted to distinguish between users with depression and users without depression by considering their profile images and the text of their profile descriptions. A classifier that used features from the profile image outperformed a baseline classifier that used only features from the profile description by approximately 10% in terms of the F_1-score. While profile images may be predictive of users' mental health statuses to an extent, there are confounding factors that these multimodal methods must address. For example, people with depression are likely on social media platforms to display positive-looking pictures (including profile images) as opposed to negative-looking ones, according to Ghosh et al [155]. This perhaps counterintuitive phenomenon has been dubbed "smiling depression" and training of multimodal models with larger, labeled datasets is needed so that they may become more discerning in these conditions. Semwal et al [156] have also evidenced in similar experimental settings that information contained in tweet text and profile images complement one another and ought to be used in alliance. They recorded that their multimodal model outperformed their best-performing textual and image-only models by 3.5% and 27.1%, respectively, in terms of F_1-score. Therefore, the conclusion was that images seem to contain significant information regarding a user's mental health status, thus motivating further study in mental health status prediction. Meanwhile in Reddit, multimodal methods may involve time-aware consideration of user posts. One study considered the relative time between posts as a feature for distinguishing between Reddit and Twitter users with and without depression [157]. Obtaining an F_1-score of 0.93 with Reddit and 0.87 with Twitter, it was concluded that a time-aware approach to classification is more effective where posting frequency is relatively high. The supposition is that the concise nature of Twitter posts, compared to the often much lengthier posts on Reddit, contributes to users posting more frequently on Twitter. A further study considered a multimodal approach with emphasis on emojis, again in the task of distinguishing between users with and without depression on both Twitter and Reddit [158]. With F_1-scores of 0.80 and 0.95 being achieved for Twitter and Reddit, respectively, it could be concluded, given the 2 studies that have just been outlined, that different multimodal approaches will be suitable for different platforms.

The advent of multimodal approaches may also help allay a privacy-related concern that our review has brought to the fore. The public has expressed concerns about methods for predicting mental health status that harness primitive keyword searches due to the risk of unreliable output. Naturally, a social media user may be affronted at receiving an incorrect diagnosis of depression, anxiety, or schizophrenia [59,60]. Multimodal approaches that more accurately capture people's real-life behaviors are thus being pursued [159]. It is not only methods that need to improve to gain public confidence; more fundamentally, the means of collecting data for use in any study need to be more explicit and have user consent. Inviting users to grant access to their social media data for research purposes on a large scale, perhaps at the point that they publish a social media posting, could become widespread [160]. However, such invitations must be accessible to a wide demographic. Privacy literacy, which describes one's understanding of the risks of sharing information on social media, is considered more prevalent among women than men, for example [161].

Finally, our literature search returned many articles that consider the effects of social media use on a user's levels of depression and anxiety (Multimedia Appendix 1). A primary hypothesis, greatly debated in specialist literature [162], is that extended or otherwise distinct patterns of social media use may cause or exacerbate these mental health disorders. This was not the subject area of this study, but our results on the volume of published articles suggest that this related matter perhaps merits a review of its own.

Potential Clinical Applications

With reference to the research covered in this review, we now consider the potential clinical applications of using AI on social media data. These include (1) evaluating data at a population level to inform health care delivery and policy making, (2) identifying and providing access to support and interventions for those at risk of developing mental health problems, and (3) monitoring existing individual patients to detect and intervene

XSL•FO
RenderX

at early signs of relapse [163]. The third application area was underrepresented in methods for predicting mental health status literature.

At a population level, AI and NLP may be used to navigate large volumes of data to inform clinical needs in a particular area, to identify changing patterns of mental illness across populations and time, to better understand patients' experiences and perceptions of health services, and to identify patterns of risky behaviors among certain demographics (eg, young people accessing accounts linked to proanorexia or encouraging self-harm). As noted earlier, NLP was used to evaluate large volumes of social media data during the COVID-19 pandemic and identify the specific concerns of people living with mental illness, including health anxieties, loneliness, and suicidality [164]. This type of information can be used to inform resource allocation in health services and the development of government policies. Crucially, this analysis can be performed relatively quickly (particularly compared to traditional research methods), which is essential during periods of instability, such as a public health crisis, where decisions need to be made rapidly.

At an individual level, AI may be used to identify people at risk of or living with mental health problems and enable organizations to provide early intervention support. There are some concerns regarding consent, data use, and privacy, as noted in the Ethical Aspects of the Data and Analysis of Mental Health section. Interestingly, while both young people and mental health professionals somewhat agree that social media companies should use AI to proactively detect users at risk of suicide or self-harm and signpost them helpful information and resources, they felt more strongly that AI capabilities should be used to promote helpful content such as psychoeducation [165]. In addition, there are logistical challenges to doing this, such as how individual data collected by global platforms can be harnessed by localized health care providers to support care.

Despite these challenges, social media has been proven to be a useful tool to identify relevant individuals for research, including delivering interventions to young people living with eating disorders [166] and who have been exposed to suicide (eg, a friend or family member had died by suicide or attempted suicide) [165] and using Facebook data to detect relapse in patients with schizophrenia [167]. As an example, Birnbaum et al [167] used LIWC on extracted Facebook archives and concurrent medical records for participants with psychosis. Researchers built an individual-centric classifier to predict re-admission to the hospital due to exacerbation of psychotic symptoms. However, the sensitivity of the prediction model was low (38%) indicating that the algorithm only identified a small proportion of all those who relapsed. Furthermore, the algorithm was applied to retrospective Facebook archives and paired with retrospective medical records, all with explicit consent from participants. The use of social media data to prospectively predict relapse in patients is likely to be considerably more challenging. As the authors noted, patients may change their social media behavior if they are aware that they are being actively monitored by their care team.

While the AI-driven mental health status prediction methods outlined may appear to lend themselves readily for use in clinical practice, there are limitations that need to be addressed before they are adopted. A chief limitation, as already mentioned in this review, is the likelihood of bias in methods based on data that do not represent diverse populations [168,169]. Thus, they may not be able to account, for example, for the fact that mental health conditions may present differently in different people. This is challenging to overcome because the field of mental health care is limited in its access to large, high-quality datasets. Compounding these limitations is the fact that the underlying biological processes of mental health disorders are still poorly understood meaning that models must be bootstrapped from observations rather than be derived from first principles. Indeed, the nature of decision-making in mental health care can be far more complex than that of other clinical areas. Indicative of this is the fact that the specific and objective task of tumor identification from an image is already successfully supported by AI-driven methods [170]. Mental health care therefore desires AI-driven methods that are transparent, explainable, and able to provide guidance to clinicians [26,169,171].

Limitations

We have reviewed the literature in what we deemed 4 chief areas in the realm of predicting mental health status. There are opportunities for greater depth of coverage in these areas and they could be the subject of review articles of their own. There is also scope for a greater breadth of coverage that could fuel follow-up studies. For example, our coverage has primarily considered research related to NLP, with occasional deference to multimodal alternatives. Visual computing provides techniques applicable to data from predominantly image-based platforms, such as Instagram [6,9]. Experts in computer vision may therefore be able to provide greater insight here.

Being a narrative review, the nature of article selection and analysis is somewhat subjective. To mollify this, we used a well-defined search and selection process that borrowed features often used in systematic reviews [29] (Methods section).

In addition, we only considered articles in which the participants of the studies self-reported a diagnosis of depression, anxiety, or schizophrenia; however, more widely any sort of information garnered from a social media posting should be treated like a self-report. While this confers a certainty that the input reflects the experiences and beliefs of the social media user, providing the opportunity to automatically accrue large datasets that have information about mental health statuses, this approach also has weaknesses that have been explored in the psychopathological literature [172]. For example, compared to a manually compiled and curated dataset, there are likely to be more false positive instances of nearly any common diagnosis, although there may also be false negative instances or controls that do in fact bear a mental health disorder [32] are also possible. In the case of schizophrenia, the condition itself might be partly responsible for the unreliability of self-reports, creating an even larger weakness for automatically constructed datasets as previously highlighted.

We should also mention that the social media platforms covered in this review, including Facebook, Twitter, and Reddit, are ostensibly English-language platforms. This coverage is perhaps by virtue of our literature search and selection strategy, which

XSL•FO
RenderX

excluded non-English language articles. Therefore, we acknowledge that the findings presented in this paper may well not apply to non-English language platforms such as Weibo [173] and VK [174], which are Chinese and Russian language platforms, respectively. A complementary narrative review that considers social media platforms concerning these languages and cultures could form future work.

Finally, we highlight a theme that has recurred throughout this review, which is that of biases in predicting mental health status research. Addressing these biases, or at least being aware of them, is crucial for ensuring accurate and generalizable findings. This review has concerned predominantly English-language social media platforms, which in turn, largely reflect Western culture. Therefore, when such findings are reported in the literature it must be ceded that they might not generalize to social media platforms that predominantly reflect Eastern culture. In any case, there are other platform-related biases that must be considered; certain platforms may be used largely by certain demographics. We have already noted that on platforms such as Facebook, Twitter, and Instagram, male users are in the majority [81] and that social media users are generally well-educated and affluent [82]. Cohort-based strategies for dataset construction have been trialed to account for these biases [36]. There are also user-oriented biases that may distort datasets. A user's posting habits may change over time and convey a distorted view of their life and experiences [50]. This behavior may be influenced by reports published in traditional print media on the negative consequences of social media use [175]. It may also be influenced by the proliferation of use-limiting tools, which encourage users to choose carefully the personal information they share on social media platforms [176]. On a collective scale, certain users may post content significantly more frequently than others, creating imbalances in datasets and subsequent models. This is evident in 2 of the datasets we have covered [86,90]. Data augmentation is one approach that may alleviate this problem [94,114], while another includes data synthesis via LLMs [48]. Lastly, we should mention confirmation bias, which involves people's tendency to seek data that support their beliefs and ignore or distort data contradicting them [177]. Wherever possible, a selection of appropriate datasets ought to be used in experimental setups so that conclusions are better balanced. In general, it is suggested that future research in the domain of mental health status prediction should seek and report data biases to enhance the reliability of findings [27].

Conclusions

The research area of predicting mental health status is receiving much attention, particularly in recent years. The COVID-19 era appears to have been the catalyst for the expanding interest. Further work needs to be completed with respect to methods for predicting mental health status before they may be considered sufficiently reliable for clinical purposes. We have documented public misgivings about text-only approaches, particularly those that rely on keyword searches. We have also acknowledged that image-based social media platforms such as Instagram are in wide use. Therefore, to help gain public confidence, methods will likely need to be multimodal. That is, they will need to generalize to text-, voice-, image-, and video-based social media data. The pursuit is merited to help relieve strain on health care and mental health services. In fact, the integration of automated early health care intervention methods and traditional methods may be advantageous.

This work cannot take place in a vacuum; however, due consideration must be given to the ethical concerns regarding the collection and use of social media users' data. Consent from users needs to be sought, perhaps by providing them with the opportunity to donate their social media data or by allowing them to choose to share their data for research purposes on a post-by-post basis. In any event, the purposes of collecting such data ought to be made clear to users through transparent data use agreements. Then, when data are subsequently compiled into datasets for public release, anonymization of the user accounts they contain is essential.

Acknowledgments

AJL was supported by the National Centre for Mental Health, a collaboration between Cardiff, Swansea, and Bangor Universities, funded by the Welsh government through Health and Care Research Wales. SES and AFP were supported by the European Union's Horizon 2020 research and innovation program under grant agreement 964874. AFP was also supported by a Medical Research Council Programme Grant (MR/Y004094/1). JCC was supported by a United Kingdom Research and Innovation Future Leaders Fellowship.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Verbose exposition of the 4-stage literature selection process.
[XLSX File (Microsoft Excel File), 272 KB - jmir_v26i1e59225_app1.xlsx]

XSL•FO
RenderX

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella MP, SET-DEP Group. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. J Affect Disord 2016 Jul 01;198:96-101. [doi: 10.1016/j.jad.2016.03.025] [Medline: 27015158]
3. De CM, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference. 2013 Presented at: WebSci '13; May 2-4, 2013; Paris, France p. 47-56 URL: https://dl.acm.org/doi/10.1145/2464464.2464480 [doi: 10.1145/2464464.2464480]
4. Ammari T, Schoenebeck S. Networked empowerment on Facebook groups for parents of children with special needs. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15; April 18-23, 2015; Seoul, Republic of Korea p. 2805-2814 URL: https://dl.acm.org/doi/10.1145/2702123.2702324 [doi: 10.1145/2702123.2702324]
5. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in Reddit social media forum. IEEE Access 2019;7:44883-44893. [doi: 10.1109/access.2019.2909180]
6. Andalibi N, Ozturk P, Forte A. Sensitive self-disclosures, responses, and social support on Instagram: the case of #Depression. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computi. 2017 Presented at: CSCW '17; February 25-March 1, 2017; Portland, OR p. 1485-1500 URL: https://dl.acm.org/doi/10.1145/2998181.2998243 [doi: 10.1145/2998181.2998243]
7. Dinu A, Moldovan A. Automatic detection and classification of mental illnesses from general social media texts. In: Proceedings of the 2021 International Conference on Recent Advances in Natural Language Processing. 2021 Sep Presented at: RANLP '21; Sep 1-3, 2021; Virtual Event p. 358-366 URL: https://aclanthology.org/2021.ranlp-1.41.pdf [doi: 10.26615/978-954-452-072-4_041]
8. Singh A, Singh J. Automation of detection of social network mental disorders – a review. IOP Conf Ser Mater Sci Eng 2021 Jan 01;1022(1):012008 [FREE Full text] [doi: 10.1088/1757-899x/1022/1/012008]
9. Muhammad KA. Unveiling the emotional and psychological states of Instagram users: a deep learning approach to mental health analysis. Inf Sci Lett 2023 May 01;12(5):1877-1890. [doi: 10.18576/isl/120531]
10. Ganguly C, Nayak S, Gupta A. Mental health impact of COVID-19 and machine learning applications in combating mental disorders: a review. In: Jain S, Pandey K, Jain P, Seng KP, editors. Artificial Intelligence, Machine Learning, and Mental Health in Pandemics. New York, NY: Academic Press; Jan 2022:1-51.
11. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry 2020 Jun;7(6):547-560 [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]
12. Chancellor S, Birnbaum M, Caine E, Silenzio V, De CM. A taxonomy of ethical tensions in inferring mental health states from social media. In: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT* '19; January 29-31, 2019; Atlanta, GA p. 79-88 URL: https://dl.acm.org/doi/10.1145/3287560.3287587 [doi: 10.1145/3287560.3287587]
13. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
14. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 1966 Jan;9(1):36-45. [doi: 10.1145/365153.365168]
15. Rajaraman V. From ELIZA to ChatGPT: history of human-computer conversation. Reson 2023 Jun 21;28(6):889-905. [doi: 10.1007/s12045-023-1620-6]
16. Gilman ES, Kot S, Engineer M, Dixon E. Training adults with mild to moderate dementia in ChatGPT: exploring best practices. In: Proceedings of the 29th International Conference on Intelligent User Interface. 2024 Presented at: IUI '24; March 18-21, 2024; Greenville, SC p. 101-106 URL: https://dl.acm.org/doi/abs/10.1145/3640544.3645230 [doi: 10.1145/3640544.3645230]
17. Whissell CM. The dictionary of affect in language. In: Plutchik R, Kellerman H, editors. The Measurement of Emotions. New York, NY: Academic Press; 1989:113-131.
18. Whissell dictionary of affect in language. Spirituality & The Brain. URL: https://www.god-helmet.com/wp/whissel-dictionary-of-affect/index.htm [accessed 2024-08-02]
19. Welcome to Dokteronline. Dokteronline. URL: https://www.dokteronline.com/ [accessed 2024-02-29]
20. Johnsen JK, Rosenvinge JH, Gammon D. Online group interaction and mental health: an analysis of three online discussion forums. Scand J Psychol 2002 Dec;43(5):445-449. [doi: 10.1111/1467-9450.00313] [Medline: 12500784]
21. McKenna KY, Bargh JA. Coming out in the age of the internet: identity "demarginalization" through virtual group participation. Pers Soc Psychol 1998;75(3):681-694. [doi: 10.1037//0022-3514.75.3.681]

XSL•FO
RenderX

22. Thorn P, La Sala L, Hetrick S, Rice S, Lamblin M, Robinson J. Motivations and perceived harms and benefits of online communication about self-harm: an interview study with young people. Digit Health 2023;9:20552076231176689 [FREE Full text] [doi: 10.1177/20552076231176689] [Medline: 37252260]
23. Haker H, Lauber C, Rössler W. Internet forums: a self-help approach for individuals with schizophrenia? Acta Psychiatr Scand 2005 Dec;112(6):474-477. [doi: 10.1111/j.1600-0447.2005.00662.x] [Medline: 16279878]
24. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety 2011 Jun;28(6):447-455 [FREE Full text] [doi: 10.1002/da.20805] [Medline: 21400639]
25. Winn JG, Hao T, Hardaway JW, Oh H. Redefining searching in non-medical sciences systematic reviews: the ascendance of Google Scholar as the primary database. J Libr Inf Sci 2024 Jun 16;5:1. [doi: 10.1177/09610006241256393]
26. Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics 2021;126(1):871-906 [FREE Full text] [doi: 10.1007/s11192-020-03690-4] [Medline: 32981987]
27. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 2018 Nov 10;118(1):177-214. [doi: 10.1007/s11192-018-2958-5]
28. Martín-Martín A, Orduna-Malea E, Thelwall M, López-Cózar E. Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. J Informetr 2018 Nov;12(4):1160-1177. [doi: 10.1016/j.joi.2018.09.002]
29. Ferrari R. Writing narrative style literature reviews. Med Writ 2015 Dec 23;24(4):230-235. [doi: 10.1179/2047480615z.000000000329]
30. Szeto MD, Barber C, Ranpariya VK, Anderson J, Hatch J, Ward J, et al. Emojis and Emoticons in health care and dermatology communication: narrative review. JMIR Dermatol 2022;5(3):e33851 [FREE Full text] [doi: 10.2196/33851] [Medline: 36405493]
31. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc Int AAAI Conf Weblogs Soc Media 2021 Aug 03;7(1):128-137. [doi: 10.1609/icwsm.v7i1.14432]
32. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014 Presented at: CLCP-LSCR '14; June 27, 2014; Baltimore, MD p. 51-60 URL: https://aclanthology.org/W14-3207.pdf [doi: 10.3115/v1/w14-3207]
33. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: EMNLP '17; September 7-11, 2017; Copenhagen, Denmark p. 2968-2978 URL: https://aclanthology.org/D17-1322.pdf [doi: 10.18653/v1/d17-1322]
34. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, et al. RSDD-time: temporal annotation of self-reported mental health diagnoses. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology. 2018 Presented at: CLPsych '18; June 5, 2018; New Orleans, LA p. 168-173 URL: https://aclanthology.org/W18-0618.pdf [doi: 10.18653/v1/w18-0618]
35. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: Proceedings of the 27th International Conference on Computational Linguistics. 2018 Presented at: COLING '18; August 20-26, 2018; Santa Fe, MX p. 1485-1497 URL: https://aclanthology.org/C18-1126.pdf [doi: 10.18653/v1/w18-0618]
36. Amir S, Dredze M, Ayers J. Mental health surveillance over social media with digital cohorts. In: Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology. 2019 Jun Presented at: CLPsych '19; June 6, 2019; Minneapolis, MN p. 114-120 URL: https://aclanthology.org/W19-3013.pdf [doi: 10.18653/v1/w19-3013]
37. Parapar J, Martín-Rodilla P, Losada D, Crestani F. Overview of eRisk 2021: early risk prediction on the internet. In: Proceedings of the 12th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction. 2021 Presented at: CLEF '21; September 21-24, 2021; Virtual Event p. 324-344 URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_22 [doi: 10.1007/978-3-030-85251-1_22]
38. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen V, Boyd-Graber J. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology. 2015 Presented at: CLPsych '15; June 5, 2015; Denver, CO p. 99-107 URL: https://aclanthology.org/W15-1212.pdf [doi: 10.3115/v1/w15-1212]
39. Mitchell M, Hollingshead K, Coppersmith G. Quantifying the language of schizophrenia in social media. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology. 2015 Presented at: CLPsych '15; June 5, 2015; Denver, CO p. 11-20 URL: https://aclanthology.org/W15-1202.pdf [doi: 10.3115/v1/w15-1202]
40. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing depression from Twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI EA '15; April 18-23, 2015; Seoul, Republic of Korea p. 3187-3196 URL: https://dl.acm.org/doi/10.1145/2702123.2702280 [doi: 10.1145/2702123.2702280]

XSL•FO
RenderX

41. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. J Med Internet Res 2017 Aug 14;19(8):e289 [FREE Full text] [doi: 10.2196/jmir.7956] [Medline: 28807891]
42. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 2017 Dec;18:43-49. [doi: 10.1016/j.cobeha.2017.07.005]
43. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep 2017 Oct 11;7(1):13006 [FREE Full text] [doi: 10.1038/s41598-017-12961-9] [Medline: 29021528]
44. Burdisso SG, Errecalde M, Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. Expert Syst Appl 2019 Nov;133:182-197. [doi: 10.1016/j.eswa.2019.05.023]
45. Owen D, Camacho-Collados J, Anke LE. Towards preemptive detection of depression and anxiety in Twitter. In: Proceedings of the 5th Social Media Mining for Health Applications. 2020 Presented at: SMM4H '20; December 12, 2020; Barcelona, Spain p. 82-89 URL: https://aclanthology.org/2020.smm4h-1.12.pdf
46. Malviya K, Roy B, Saritha S. A transformers approach to detect depression in social media. In: Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems. 2021 Presented at: ICAIS '21; March 25-27, 2021; Coimbatore, India p. 718-723 URL: https://ieeexplore.ieee.org/document/9395943 [doi: 10.1109/icais50930.2021.9395943]
47. Sharma T, Panchendrarajan R, Saxena A. Characterisation of mental health conditions in social media using deep learning techniques. In: Hong TP, Serrano-Estrada L, Saxena A, Biswas A, editors. Deep Learning for Social Media Data Analytics. Cham, Switzerland: Springer; 2022:157-176.
48. Bucur AM. Utilizing ChatGPT generated data to retrieve depression symptoms from social media. arXiv Preprint posted online July 5, 2023 [FREE Full text]
49. Ramos dos Santos W, Paraboni I. Prompt-based mental health screening from social media text. arXiv [FREE Full text] Preprint posted online January 11, 2024 [doi: 10.5753/brasnam.2024.1879]
50. Schwartz H, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, et al. Towards assessing changes in degree of depression through Facebook. In: Proceedings of the 2014 Workshop on Computational Linguistics and Clinical Psychology. 2014 Presented at: CLPsych '14; June 27, 2014; Baltimore, MD p. 118-125 URL: https://aclanthology.org/W14-3214.pdf [doi: 10.3115/v1/w14-3214]
51. Loveys K, Crutchley P, Wyatt E, Coppersmith G. Small but mighty: affective micropatterns for quantifying mental health from social media language. In: Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology. 2017 Presented at: CLPsych '17; August 3, 2017; Vancouver, BC p. 85-95 URL: https://aclanthology.org/W17-3110.pdf [doi: 10.18653/v1/w17-3110]
52. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Monitoring online discussions about suicide among Twitter users with schizophrenia: exploratory study. JMIR Ment Health 2018 Dec 13;5(4):e11483 [FREE Full text] [doi: 10.2196/11483] [Medline: 30545811]
53. Seabrook EM, Kern ML, Fulcher BD, Rickard NS. Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. J Med Internet Res 2018 May 08;20(5):e168 [FREE Full text] [doi: 10.2196/jmir.9267] [Medline: 29739736]
54. Chen X, Sykora M, Jackson T, Elayan S. What about mood swings? Identifying depression on Twitter with temporal measures of emotions. In: Proceedings of the 2018 Web Conference. 2018 Presented at: WWW '18; April 23-27, 2018; Lyon, France p. 1653-1660 URL: https://dl.acm.org/doi/fullHtml/10.1145/3184558.3191624 [doi: 10.1145/3184558.3191624]
55. Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring depression trends on Twitter during the COVID-19 pandemic: observational study. JMIR Infodemiology 2021;1(1):e26769 [FREE Full text] [doi: 10.2196/26769] [Medline: 34458682]
56. Kelley SW, Gillan CM. Using language in social media posts to study the network dynamics of depression longitudinally. Nat Commun 2022 Feb 15;13(1):870 [FREE Full text] [doi: 10.1038/s41467-022-28513-3] [Medline: 35169166]
57. Owen D, Antypas D, Hassoulas A, Pardiñas AF, Espinosa-Anke L, Collados JC. Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. JMIR AI 2023 Mar 24;2:e41205 [FREE Full text] [doi: 10.2196/41205] [Medline: 37525646]
58. Välimäki M, Athanasopoulou C, Lahti M, Adams CE. Effectiveness of social media interventions for people with schizophrenia: a systematic review and meta-analysis. J Med Internet Res 2016 Apr 22;18(4):e92 [FREE Full text] [doi: 10.2196/jmir.5385] [Medline: 27105939]
59. Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC Med Ethics 2016 Apr 14;17:22 [FREE Full text] [doi: 10.1186/s12910-016-0105-5] [Medline: 27080238]
60. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
61. Chancellor S, Baumer EP, De Choudhury M. Who is the "human" in human-centered machine learning: the case of predicting mental health from social media. Proc ACM Hum Comput Interact 2019 Nov 07;3(CSCW):1-32. [doi: 10.1145/3359249]
62. Nicholas J, Onie S, Larsen ME. Ethics and privacy in social media research for mental health. Current Psychiatry Reports 2020 Nov 23;22(12):84. [doi: 10.1007/s11920-020-01205-9] [Medline: 33225404]

XSL•FO
RenderX

63. Vornholt P, De Choudhury M. Understanding the role of social media-based mental health support among college students: survey and semistructured interviews. JMIR Ment Health 2021 Jul 12;8(7):e24512 [FREE Full text] [doi: 10.2196/24512] [Medline: 34255701]
64. Cai N, Revez JA, Adams MJ, Andlauer TF, Breen G, Byrne EM, MDD Working Group of the Psychiatric Genomics Consortium, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet 2020 Apr;52(4):437-447 [FREE Full text] [doi: 10.1038/s41588-020-0594-5] [Medline: 32231276]
65. Cong Q, Feng Z, Li F, Xiang Y, Rao G, Tao C. X-A-BiLSTM: a deep learning approach for depression detection in imbalanced data. In: Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. 2018 Presented at: BIBM '18; December 3-6, 2018; Madrid, Spain p. 1624-1627 URL: https://ieeexplore.ieee.org/document/8621230 [doi: 10.1109/bibm.2018.8621230]
66. Marnauzs S, Kalita J. A domain independent social media depression detection model. In: The 33rd AAAI Conference on Artificial Intelligence. 2019 Presented at: AI '19; January 27-February 1, 2019; Honolulu, HI p. 50-55 URL: https://faculty.uccs.edu/jkalita/wp-content/uploads/sites/45/2024/02/00Proceedings2019Optimized.pdf#page=62
67. Trifan A, Semeraro D, Drake J, Bukowski R, Oliveira JL. Social media mining for postpartum depression prediction. Stud Health Technol Inform 2020 Jun 16;270:1391-1392. [doi: 10.3233/SHTI200457] [Medline: 32570674]
68. Bucur AM, Cosma A, Dinu LP. Early risk detection of pathological gambling, self-harm and depression using BERT. arXiv Preprint posted online June 30, 2021 [FREE Full text]
69. Ali J, Ngo DQ, Bhattacharjee A, Maiti T, Singh T, Mei J. Depression detection: text augmentation for robustness to label noise in self-reports. In: Bertolaso M, Capone L, Rodríguez-Lluesma C, editors. Digital Humanism: A Human-Centric Approach to Digital Technologies. Cham, Switzerland: Springer; 2022:81-103.
70. Kulkarni H, MacAvaney S, Goharian N, Frieder O. Knowledge augmentation for early depression detection. In: Proceedings of the 7th International Workshop on Health Intelligence. 2023 Presented at: W3PHAI '23; February 13-14, 2023; Washington, DC p. 175-191. [doi: 10.1007/978-3-031-36938-4_14]
71. Souza VB, Nobre J, Becker K. Characterization of anxiety, depression, and their comorbidity from texts of social networks. In: Proceedings of the 35th Brazilian Symposium on Databases. 2022 Presented at: SBBD '20; September 28-October 1, 2020; Virtual Event p. 121-132 URL: https://sol.sbc.org.br/index.php/sbbd/article/view/13630/13478 [doi: 10.5753/sbbd.2020.13630]
72. Chen Z, Yang R, Fu S, Zong N, Liu H, Huang M. Detecting reddit users with depression using a hybrid neural network SBERT-CNN. In: Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics. 2023 Presented at: ICHI '23; June 26-29, 2023; Houston, TX p. 193-199 URL: https://www.computer.org/csdl/proceedings-article/ichi/2023/026300a193/1SN7B0FKEmI [doi: 10.1109/ichi57859.2023.00035]
73. Buddhitha P, Inkpen D. Multi-task learning to detect suicide ideation and mental disorders among social media users. Front Res Metr Anal 2023;8:1152535 [FREE Full text] [doi: 10.3389/frma.2023.1152535] [Medline: 37138946]
74. Khan M, Sakib S, Habib A, Hossain M. A machine learning and deep learning approach to classify mental illness with the collaboration of natural language processing. In: Proceedings of the 2022 International Conference on Information and Communication Technology for Development. 2022 Presented at: ICICTD '22; July 29-30, 2022; Khulna, Bangladesh p. 83-94 URL: https://link.springer.com/chapter/10.1007/978-981-19-7528-8_7 [doi: 10.1007/978-981-19-7528-8_7]
75. Zanwar S, Wiechmann D, Qiao Y, Kerz E. Exploring hybrid and ensemble models for multiclass prediction of mental health status on social media. In: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis. 2022 Presented at: LOUHI '22; December 7, 2022; Abu Dhabi, United Arab Emirates p. 184-196 URL: https://aclanthology.org/2022.louhi-1.21.pdf [doi: 10.18653/v1/2022.louhi-1.21]
76. Sekulic I, Strube M. Adapting deep learning methods for mental health prediction on social media. In: Proceedings of the 5th Workshop on Noisy User-generated Text. 2019 Presented at: WNUT '19; November 4, 2019; Hong Kong, China p. 322-327 URL: https://aclanthology.org/D19-5542.pdf [doi: 10.18653/v1/d19-5542]
77. Borba de Souza V, Campos Nobre J, Becker K. DAC stacking: a deep learning ensemble to classify anxiety, depression, and their comorbidity from Reddit texts. IEEE J Biomed Health Inform 2022 Jul;26(7):3303-3311. [doi: 10.1109/JBHI.2022.3151589] [Medline: 35230959]
78. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). Washington, DC: American Psychiatric Association; 2013.
79. Ireland M, Iserman M. Within and between-person differences in language used across anxiety support and neutral reddit communities. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology. 2018 Presented at: CLPsych '18; June 5, 2018; New Orleans, LA p. 182-193 URL: https://aclanthology.org/W18-0620.pdf [doi: 10.18653/v1/w18-0620]
80. Harrigian K, Aguirre C, Dredze M. Do models of mental health based on social media data generalize? In: Proceedings of the 2020 Findings of the Association for Computational Linguistics. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual Event p. 3774-3788 URL: https://aclanthology.org/2020.findings-emnlp.337.pdf [doi: 10.18653/v1/2020.findings-emnlp.337]
81. The 2024 social media demographics guide. Khoros. URL: https://khoros.com/resources/social-media-demographics-guide [accessed 2024-04-29]

XSL•FO
RenderX

82. Hruska J, Maresova P. Use of social media platforms among adults in the United States—behavior on social media. Societies 2020 Mar 23;10(1):27. [doi: 10.3390/soc10010027]
83. Amir S, Coppersmith G, Carvalho P, Silva M, Wallace B. Quantifying mental health from social media with neural user embeddings. arXiv Preprint posted online April 30, 2017 [FREE Full text]
84. Wu S, Qiu Z. A RoBERTa-based model on measuring the severity of the signs of depression. In: Proceedings of the 2021 Conference and Labs of the Evaluation Forum. 2021 Presented at: CLEF '21; September 21-24, 2021; Bucharest, Romania p. 1-10 URL: https://ceur-ws.org/Vol-2936/paper-86.pdf
85. Inkpen D, Skaik R, Buddhitha P, Angelov D, Fredenburgh M. uOttawa at eRisk 2021: automatic filling of the Beck's depression inventory questionnaire using deep learning. In: Proceedings of the 2021 Conference and Labs of the Evaluation Forum. 2021 Presented at: CLEF '21; September 21-24, 2021; Bucharest, Romania p. 966-980 URL: https://ceur-ws.org/Vol-2936/paper-79.pdf
86. Reddit Self-reported Depression Diagnosis (RSDD) dataset. Georgetown IR Lab. 2023. URL: https://georgetown-ir-lab.github.io/emnlp17-depression/ [accessed 2024-04-05]
87. Resources. Georgetown IR Lab. 2023. URL: https://ir.cs.georgetown.edu/resources/ [accessed 2024-04-05]
88. Resources - SMHD. Georgetown IR Lab. 2023. URL: https://ir.cs.georgetown.edu/resources/smhd.html [accessed 2024-04-05]
89. Dredze M. CLPsych 2015 shared task evaluation. Johns Hopkins Whiting School of Engineering. URL: https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/ [accessed 2024-04-05]
90. eRisk 2021 text research collection. Bibtex. URL: https://tec.citius.usc.es/ir/code/eRisk2021.html [accessed 2024-04-05]
91. What is topic modeling?: a beginner's guide. Levity. URL: https://levity.ai/blog/what-is-topic-modeling [accessed 2024-04-29]
92. Pennebaker J, Boyd R, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. The University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2024-04-29]
93. Hier DB, Brint SU. A neuro-ontology for the neurological examination. BMC Med Inform Decis Mak 2020 Mar 04;20(1):47 [FREE Full text] [doi: 10.1186/s12911-020-1066-7] [Medline: 32131804]
94. Chen J, Tam D, Raffel C, Bansal M, Yang D. An empirical survey of data augmentation for limited data learning in nlp. Trans Assoc Comput Linguist 2023;11:191-211. [doi: 10.1162/tacl_a_00542]
95. Liu B. Supervised learning. In: Liu B, editor. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Cham, Switzerland: Springer; 2011.
96. Noble WS. What is a support vector machine? Nat Biotechnol 2006 Dec;24(12):1565-1567. [doi: 10.1038/nbt1206-1565] [Medline: 17160063]
97. Breiman L. Random forests. Mach Learn 2001;45:5-32. [doi: 10.1023/A:1010933404324]
98. Zhong G, Wang L, Ling X, Dong J. An overview on data representation learning: from traditional feature learning to recent deep learning. J Fin Data Sci 2016 Dec;2(4):265-278. [doi: 10.1016/j.jfds.2017.05.001]
99. What is BERT and how is it used in AI? H2O.ai. 2023. URL: https://h2o.ai/wiki/bert/ [accessed 2024-02-18]
100. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL-HLT '19; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf
101. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: a lite bert for self-supervised learning of language representations. arXiv Preprint posted online September 26, 2019 [FREE Full text]
102. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: publicly available pretrained language models for mental healthcare. In: Proceedings of the 13th Language Resources and Evaluation Conference. 2022 Presented at: LREC '22; June 20-25, 2022; Marseille, France p. 7184-7190 URL: https://aclanthology.org/2022.lrec-1.778.pdf
103. Jo A. What ChatGPT and generative AI mean for science. Nature 2023 Feb 9;614(1):214-216 [FREE Full text]
104. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020 Presented at: NIPS '20; December 6-12, 2020; Vancouver, BC p. 1877-1901 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883
105. Wu T, He S, Liu J, Sun S, Liu K, Han Q, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sin 2023 May;10(5):1122-1136. [doi: 10.1109/jas.2023.123618]
106. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (Kaunas) 2021 May 16;57(5):503 [FREE Full text] [doi: 10.3390/medicina57050503] [Medline: 34065637]
107. Shung KP. Accuracy, precision, recall or F1? Towards Data Science. 2018. URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [accessed 2024-04-18]
108. Narkhede S. Understanding AUC - ROC curve. Towards Data Science. 2018. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [accessed 2024-08-18]
109. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, National Comorbidity Survey Replication. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 2003 Jun 18;289(23):3095-3105. [doi: 10.1001/jama.289.23.3095] [Medline: 12813115]

XSL•FO
RenderX

110. Fujita F, Diener E, Sandvik E. Gender differences in negative affect and well-being: the case for emotional intensity. J Pers Soc Psychol 1991 Sep;61(3):427-434. [doi: 10.1037//0022-3514.61.3.427] [Medline: 1941513]
111. Software and systems engineering "software testing" part 11: guidelines on the testing of AI-based systems. International Organization for Standardization. 2023. URL: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:29119:-11:ed-1:v1:en:term:3.1.31 [accessed 2024-04-12]
112. Shaikh S, Daudpota SM, Imran AS, Kastrati Z. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Appl Sci (Basel) 2021 Jan 19;11(2):869. [doi: 10.3390/app11020869]
113. Younes Y, Mathiak B. Handling class imbalance when detecting dataset mentions with pre-trained language models. In: Proceedings of the 5th International Conference on Natural Language and Speech Processing. 2022 Presented at: ICNLSP '22; December 16-17, 2022; Trento, Italy p. 79-88 URL: https://aclanthology.org/2022.icnlsp-1.9.pdf
114. Bayer M, Kaufhold MA, Buchhold B, Keller M, Dallmeyer J, Reuter C. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. Int J Mach Learn Cybern 2022 Apr 12;14(1):135-150 [FREE Full text] [doi: 10.1007/s13042-022-01553-3] [Medline: 35432623]
115. Li B, Hou Y, Che W. Data augmentation approaches in natural language processing: a survey. AI Open 2022;3:71-90. [doi: 10.1016/j.aiopen.2022.03.001]
116. Santos WR, de Oliveira RL, Paraboni I. SetembroBR: a social media corpus for depression and anxiety disorder prediction. Lang Resour Eval 2023 Jan 11;58(1):273-300. [doi: 10.1007/s10579-022-09633-0]
117. Santos W, Yoon S, Paraboni I. Mental health prediction from social media text using mixture of experts. IEEE Latin Am Trans 2023 Jun;21(6):723-729. [doi: 10.1109/tla.2023.10172137]
118. Andrews GJ. Co-creating health's lively, moving frontiers: brief observations on the facets and possibilities of non-representational theory. Health Place 2014 Nov;30:165-170. [doi: 10.1016/j.healthplace.2014.09.002] [Medline: 25282125]
119. Gur RE, Kohler CG, Ragland JD, Siegel SJ, Lesko K, Bilker WB, et al. Flat affect in schizophrenia: relation to emotion processing and neurocognitive measures. Schizophr Bull 2006 Apr;32(2):279-287 [FREE Full text] [doi: 10.1093/schbul/sbj041] [Medline: 16452608]
120. Rickelman BL. Anosognosia in individuals with schizophrenia: toward recovery of insight. Issues Ment Health Nurs 2004;25(3):227-242. [doi: 10.1080/01612840490274741] [Medline: 14965844]
121. Schizophrenia. National Alliance on Mental Illness. 2023. URL: https://www.nami.org/About-Mental-Illness/Mental-Health-Conditions/Schizophrenia [accessed 2024-03-19]
122. Li A, Jiao D, Liu X, Zhu T. A comparison of the psycholinguistic styles of schizophrenia-related stigma and depression-related stigma on social media: content analysis. J Med Internet Res 2020 Apr 21;22(4):e16470 [FREE Full text] [doi: 10.2196/16470] [Medline: 32314969]
123. Robinson P, Turk D, Jilka S, Cella M. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Soc Psychiatry Psychiatr Epidemiol 2019 Jan;54(1):51-58 [FREE Full text] [doi: 10.1007/s00127-018-1571-5] [Medline: 30069754]
124. Meyer-Lindenberg A. The non-ergodic nature of mental health and psychiatric disorders: implications for biomarker and diagnostic research. World Psychiatry 2023 Jun;22(2):272-274 [FREE Full text] [doi: 10.1002/wps.21086] [Medline: 37159352]
125. Zivin K, Eisenberg D, Gollust SE, Golberstein E. Persistence of mental health problems and needs in a college student population. J Affect Disord 2009 Oct;117(3):180-185. [doi: 10.1016/j.jad.2009.01.001] [Medline: 19178949]
126. Eisenberg D, Golberstein E, Gollust SE. Help-seeking and access to mental health care in a university student population. Med Care 2007 Jul;45(7):594-601. [doi: 10.1097/MLR.0b013e31803bb4c1] [Medline: 17571007]
127. Burke M, Kraut R, Marlow C. Social capital on Facebook: differentiating uses and users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011 Presented at: CHI '11; May 7-12, 2011; Vancouver, BC p. 571-580 URL: https://dl.acm.org/doi/10.1145/1978942.1979023 [doi: 10.1145/1978942.1979023]
128. Bell CC. DSM-IV: diagnostic and statistical manual of mental disorders. JAMA 1994 Sep 14;272(10):828-829. [doi: 10.1001/jama.1994.03520100096046]
129. Westrin A, Lam RW. Seasonal affective disorder: a clinical update. Ann Clin Psychiatry 2007;19(4):239-246. [doi: 10.1080/10401230701653476] [Medline: 18058281]
130. Mohammad S, Kiritchenko S, Zhu X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM) & Proceedings of the 7th International Workshop on Semantic Evaluation. 2013 Presented at: SIGLEX '13; June 14-15, 2013; Atlanta, GA p. 321-327 URL: https://aclanthology.org/S13-2053.pdf [doi: 10.3115/v1/s14-2077]
131. Messinger JW, Trémeau F, Antonius D, Mendelsohn E, Prudent V, Stanford AD, et al. Avolition and expressive deficits capture negative symptom phenomenology: implications for DSM-5 and schizophrenia research. Clin Psychol Rev 2011 Feb;31(1):161-168 [FREE Full text] [doi: 10.1016/j.cpr.2010.09.002] [Medline: 20889248]
132. Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van IJzendoorn MH. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psychol Bull 2007 Jan;133(1):1-24. [doi: 10.1037/0033-2909.133.1.1] [Medline: 17201568]

XSL•FO
RenderX

133. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
134. Rickard N, Arjmand H, Bakker D, Seabrook E. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. JMIR Ment Health 2016 Nov 23;3(4):e49 [FREE Full text] [doi: 10.2196/mental.6202] [Medline: 27881358]
135. Sykora MD, Jackson TW, O'Brien A, Elayan S. Emotive ontology: extracting fine-grained emotions from terse, informal messages. Int J Comput Sci Inf Sec 2013;8(2):6 [FREE Full text]
136. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv Preprint posted online June 19, 2019 [FREE Full text]
137. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, et al. Tracking suicide risk factors through Twitter in the US. Crisis 2014;35(1):51-59. [doi: 10.1027/0227-5910/a000234] [Medline: 24121153]
138. Fodeh S, Goulet J, Brandt C, Hamada A. Leveraging Twitter to better identify suicide risk. Proc Mach Learn Res 2017;69:1-7 [FREE Full text]
139. Hawton K, Sutton L, Haw C, Sinclair J, Deeks JJ. Schizophrenia and suicide: systematic review of risk factors. Br J Psychiatry 2005 Jul;187:9-20. [doi: 10.1192/bjp.187.1.9] [Medline: 15994566]
140. Sareen J, Cox BJ, Afifi TO, de Graaf R, Asmundson GJ, ten Have M, et al. Anxiety disorders and risk for suicidal ideation and suicide attempts: a population-based longitudinal study of adults. Arch Gen Psychiatry 2005 Nov;62(11):1249-1257. [doi: 10.1001/archpsyc.62.11.1249] [Medline: 16275812]
141. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 06;19(6):e195 [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]
142. Ford E, Curlewis K, Wongkoblap A, Curcin V. Public opinions on using social media content to identify users with depression and target mental health care advertising: mixed methods survey. JMIR Ment Health 2019 Nov 13;6(11):e12942 [FREE Full text] [doi: 10.2196/12942] [Medline: 31719022]
143. Liu G, Wang C, Peng K, Huang H, Li Y, Cheng W. SocInf: membership inference attacks on social media health data with machine learning. IEEE Trans Comput Soc Syst 2019 Oct;6(5):907-921. [doi: 10.1109/tcss.2019.2916086]
144. Reidenberg JR, Breaux T, Cranor LF, French B, Grannis A, Graves J, et al. Disagreeable privacy policies: mismatches between meaning and users understanding. SSRN Journal Preprint posted online March 31, 2014 [FREE Full text] [doi: 10.2139/ssrn.2418297]
145. Sleigh J. Experiences of donating personal data to mental health research: an explorative anthropological study. Biomed Inform Insights 2018 Jun 27;10:1178222618785131 [FREE Full text] [doi: 10.1177/1178222618785131] [Medline: 30013355]
146. Beninger K, Fry A, Jago N, Lepps H, Nass L, Silvester H. Research using social media; users' views. NatCen Social Research. 2014 Feb 20. URL: https://www.researchgate.net/profile/Kelsey-Beninger/publication/261551701_Research_using_Social_Media_Users'_Views/links/0c96053497fed9ac11000000/Research-using-Social-Media-Users-Views.pdf [accessed 2024-02-18]
147. Haslam N. Dehumanization: an integrative review. Pers Soc Psychol Rev 2006;10(3):252-264. [doi: 10.1207/s15327957pspr1003_4] [Medline: 16859440]
148. Kaplan K, Salzer MS, Solomon P, Brusilovskiy E, Cousounis P. Internet peer support for individuals with psychiatric disabilities: a randomized controlled trial. Soc Sci Med 2011 Jan;72(1):54-62. [doi: 10.1016/j.socscimed.2010.09.037] [Medline: 21112682]
149. Spanakis P, Wadman R, Walker L, Heron P, Mathers A, Baker J, et al. Measuring the digital divide among people with severe mental ill health using the essential digital skills framework. Perspect Public Health 2024 Jan;144(1):21-30 [FREE Full text] [doi: 10.1177/17579139221106399] [Medline: 35929589]
150. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A 2018 Oct 30;115(44):11203-11208 [FREE Full text] [doi: 10.1073/pnas.1802331115] [Medline: 30322910]
151. Woolway GE, Legge SE, Lynham A, Smart SE, Hubbard L, Daniel ER, et al. Assessing the validity of a self-reported clinical diagnosis of schizophrenia. medRxiv Preprint posted online December 8, 2023 [FREE Full text] [doi: 10.1101/2023.12.06.23299622] [Medline: 38106032]
152. Cheng S, Chang C, Chang W, Wang H, Liang C, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci 2023 Nov;77(11):592-596 [FREE Full text] [doi: 10.1111/pcn.13588] [Medline: 37612880]
153. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell 2023 Oct;6:20-25. [doi: 10.21037/jmai-23-71]
154. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-llm: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol 2024 Mar 06;8(1):1-32. [doi: 10.1145/3643540]
155. Ghosh S, Ekbal A, Bhattacharyya P. What does your bio say? Inferring Twitter users' depression status from multimodal profile information using deep learning. IEEE Trans Comput Soc Syst 2022 Oct;9(5):1484-1494. [doi: 10.1109/tcss.2021.3116242]

XSL•FO
RenderX

156. Semwal N, Suri M, Chaudhary D, Gorton I, Kumar B. Multimodal analysis and modality fusion for detection of depression from Twitter data. In: Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence. 2023 Presented at: AAAI '23; February 25-March 4, 2023; Philadelphia, PA p. 1-5 URL: https://amulyayadav.github.io/AI4SG2023/images/31.pdf
157. Bucur AM, Cosma A, Rosso P, Dinu LP. It's just a matter of time: detecting depression with time-enriched multimodal transformers. In: Proceedings of the 45th European Conference on Information Retrieval on Advances in Information Retrieval. 2023 Presented at: ECIR '23; April 2-6, 2023; Dublin, Ireland p. 200-215 URL: https://link.springer.com/chapter/10.1007/978-3-031-28244-7_13 [doi: 10.1007/978-3-031-28244-7_13]
158. Zhang H, Wang H, Han S, Li W, Zhuang L. Detecting depression tendency with multimodal features. Comput Methods Programs Biomed 2023 Oct;240:107702. [doi: 10.1016/j.cmpb.2023.107702] [Medline: 37531689]
159. Khoo LS, Lim MK, Chong CY, McNaney R. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. Sensors (Basel) 2024 Jan 06;24(2):348. [doi: 10.3390/s24020348] [Medline: 38257440]
160. van Driel II, Giachanou A, Pouwels JL, Boeschoten L, Beyens I, Valkenburg PM. Promises and pitfalls of social media data donations. Commun Methods Meas 2022 Sep 12;16(4):266-282. [doi: 10.1080/19312458.2022.2109608]
161. Choi S. Privacy literacy on social media: its predictors and outcomes. Int J Hum Comput Interact 2022 Apr 18;39(1):217-232. [doi: 10.1080/10447318.2022.2041892]
162. Orben A, Przybylski AK. The association between adolescent well-being and digital technology use. Nat Hum Behav 2019 Feb;3(2):173-182. [doi: 10.1038/s41562-018-0506-1] [Medline: 30944443]
163. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
164. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: observational study. J Med Internet Res 2020 Oct 12;22(10):e22635 [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
165. La Sala L, Pirkis J, Cooper C, Hill NT, Lamblin M, Rajaram G, et al. Acceptability and potential impact of the #chatsafe suicide postvention response among young people who have been exposed to suicide: pilot study. JMIR Hum Factors 2023 May 19;10:e44535 [FREE Full text] [doi: 10.2196/44535] [Medline: 37204854]
166. Kasson E, Vázquez MM, Doroshenko C, Fitzsimmons-Craft EE, Wilfley DE, Taylor CB, et al. Exploring social media recruitment strategies and preliminary acceptability of an mHealth tool for teens with eating disorders. Int J Environ Res Public Health 2021 Jul 28;18(15):7979 [FREE Full text] [doi: 10.3390/ijerph18157979] [Medline: 34360270]
167. Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, R Van Meter A, De Choudhury M, et al. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. NPJ Schizophr 2019 Oct 07;5(1):17 [FREE Full text] [doi: 10.1038/s41537-019-0085-9] [Medline: 31591400]
168. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
169. Singh V, Sarkar S, Gaur V, Grover S, Singh OP. Clinical practice guidelines on using artificial intelligence and gadgets for mental health and well-being. Indian J Psychiatry 2024 Jan;66(Suppl 2):S414-S419 [FREE Full text] [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_926_23] [Medline: 38445270]
170. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim H, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Sep;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
171. Maddox TM, Rumsfeld JS, Payne PR. Questions for artificial intelligence in health care. JAMA 2019 Jan;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
172. Tiego J, Martin EA, DeYoung CG, Hagan K, Cooper SE, Pasion R, HiTOP Neurobiological Foundations Work Group. Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. Nat Ment Health 2023 May;1(5):304-315 [FREE Full text] [doi: 10.1038/s44220-023-00057-5] [Medline: 37251494]
173. Home page. Weibo. 2023. URL: https://m.weibo.cn/ [accessed 2024-07-21]
174. Home page. VK. 2023. URL: https://vk.com/ [accessed 2024-07-21]
175. Just how harmful is social media? Our experts weigh-in. Columbia University Mailman School of Public Health. 2021 Sep 27. URL: https://www.publichealth.columbia.edu/news/just-how-harmful-social-media-our-experts-weigh [accessed 2024-08-31]
176. Kozyreva A, Lewandowsky S, Hertwig R. Citizens versus the internet: confronting digital challenges with cognitive tools. Psychol Sci Public Interest 2020 Dec;21(3):103-156 [FREE Full text] [doi: 10.1177/1529100620946707] [Medline: 33325331]
177. Peters U. What is the function of confirmation bias? Erkenn 2020 Apr 20;87(3):1351-1376. [doi: 10.1007/s10670-020-00252-1]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
ALBERT: A Lite Bidirectional Encoder Representations from Transformers
BDI-II: Beck Depression Inventory-II
BERT: Bidirectional Encoder Representations from Transformers
EHR: electronic health record
LDA: latent Dirichlet allocation
LIWC: Linguistic Inquiry and Word Count
LLM: large language model
LM: language model
ML: machine learning
NLP: natural language processing
RSDD: Reddit Self-reported Depression Diagnosis
SMHD: self-reported mental health diagnoses
SVM: support vector machine

Edited by G Eysenbach; submitted 23.04.24; peer-reviewed by Y Hua, KW Tay; comments to author 05.07.24; revised version received 08.09.24; accepted 01.10.24; published 15.11.24.

Please cite as:
Owen D, Lynham AJ, Smart SE, Pardiñas AF, Camacho Collados J
AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges
J Med Internet Res 2024;26:e59225
URL: https://www.jmir.org/2024/1/e59225
doi:10.2196/59225
PMID:

©David Owen, Amy J Lynham, Sophie E Smart, Antonio F Pardiñas, Jose Camacho Collados. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A 25-Year Retrospective of Health IT Infrastructure Building: The Example of the Catalonia Region

Jordi Piera-Jiménez[1,2,3], MBA, PhD; Gerard Carot-Sans[1,2], PhD; Marina Ramiro-Pareta[2], PhD; Maria Mercedes Nogueras[2,4], PhD; Júlia Folguera-Profitós[1,2], MSc; Pepi Ródenas[1], MSc; Alba Jiménez-Rueda[2], BSc; Thais de Pando Navarro[1,2], MSc; Josep Antoni Mira Palacios[5], BSc; Joan Carles Fajardo[6], BSc; Joan Ustrell Campillo[5], BSc; Emili Vela[1,2], BSc; David Monterde[2,6], BSc; Damià Valero-Bover[1,2], MSc; Tara Bonet[1,2], MSc; Guillermo Tarrasó-Urios[1,2], PhD; Roser Cantenys-Sabà[1,2], MSc; Pau Fabregat-Fabregat[1,2], BSc; Beatriz Gómez Oliveros[2], MSc; Jesús Berdún[7], MSc; Xabier Michelena[2,8], MD, PhD; Isaac Cano[9,10,11], PhD; Rubèn González-Colom[9,10], PhD; Josep Roca[9,10,11], MD, PhD; Oscar Solans[6], MD; Caridad Pontes[2,7,12], MD, PhD; Pol Pérez-Sust[1,13], MSc

[1]Catalan Health Service, Barcelona, Spain
[2]Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain
[3]Faculty of Informatics, Multimedia and Telecommunications, Universitat Oberta de Catalunya, Barcelona, Spain
[4]Agency of Health Quality and Assessment of Catalonia, Barcelona, Spain
[5]Center for Telecommunications and Information Technologies, Barcelona, Spain
[6]Catalan Institute of Health, Barcelona, Spain
[7]Hospital de la Santa Creu i Sant Pau, Barcelona, Spain
[8]Rheumatology Unit, Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain
[9]Hospital Clínic de Barcelona, Barcelona, Spain
[10]Institut d'Investigacions Biomèdiques August Pi I Sunyer (IDIBAPS), Barcelona, Spain
[11]Universitat de Barcelona, Barcelona, Spain
[12]Universitat Autònoma de Barcelona, Barcelona, Spain
[13]Catalan Department of Health, Barcelona, Spain

Corresponding Author:
Jordi Piera-Jiménez, MBA, PhD
Catalan Health Service
Gran Via de les Corts Catalanes 587
Barcelona, 08007
Spain
Phone: 34 934643013
Email: jpiera@catsalut.cat

Abstract

Over the past decades, health care systems have significantly evolved due to aging populations, chronic diseases, and higher-quality care expectations. Concurrently with the added health care needs, information and communications technology advancements have transformed health care delivery. Technologies such as telemedicine, electronic health records, and mobile health apps promise enhanced accessibility, efficiency, and patient outcomes, leading to more personalized, data-driven care. However, organizational, political, and cultural barriers and the fragmented approach to health information management are challenging the integration of these technologies to effectively support health care delivery. This fragmentation collides with the need for integrated care pathways that focus on holistic health and wellness. Catalonia (northeast Spain), a region of 8 million people with universal health care coverage and a single public health insurer but highly heterogeneous health care service providers, has experienced outstanding digitalization and integration of health information over the past 25 years, when the first transition from paper to digital support occurred. This Viewpoint describes the implementation of health ITs at a system level, discusses the hits and misses encountered in this journey, and frames this regional implementation within the global context. We present the architectures and use trends of the health information platforms over time. This provides insightful information that can be used by other systems worldwide in the never-ending transformation of health care structure and services.

XSL•FO
RenderX

(J Med Internet Res 2024;26:e58933) doi:10.2196/58933

KEYWORDS

health ITs; eHealth; integrated care; open platforms; interoperability; Catalonia; digitalization; health care structure; health care delivery; integrated pathway; integrated treatment plan; process management

Introduction

The past 25 years have witnessed a progressive shift in the health care needs of the population. Factors such as aging, increased prevalence of chronic diseases, and heightened expectations for quality care have driven this evolution. The fast-paced changes in health care needs require an integrated and comprehensive approach that addresses the multifaceted health requirements of individuals [1].

Worldwide, health care systems have been continuously looking to improve the quality and efficiency of health care delivery. The late 1990s and early 2000s saw a shift from mechanistic, top-down approaches with minimal impact on outcomes to dynamic models recognizing health care as a complex adaptive system with a growing emphasis on multidisciplinary care, partnership working, and disease management [2]. Population-based approaches emerged, targeting specific groups such as older adults and integrating social care, education, and housing to address broader determinants of health [3-7].

In the last decade, the focus on integrated care has further evolved with advancements in technology and global trends such as advances in precision medicine. Strengthening primary care and moving toward a whole-of-system approach that links different levels of the health system has become crucial [8-10]. Functional integration, involving enhanced communication, improved care coordination, and robust clinical information flows, supports integrated care partnerships [11,12]. Initiatives such as accountable care organizations in the United States, where groups of health care providers work together to improve quality and reduce costs for Medicare patients by coordinating care and sharing in the savings achieved, are a good example of this [13-15]. Furthermore, the use of digital tools and health information systems has facilitated better care coordination and patient management, leading to value-based care models that incentivize quality and outcomes [16]. With this idea in mind, the National Health Service in the United Kingdom has recently launched the integrated care systems (ICSs) initiative to better coordinate services across health and social care, aiming for more efficient, patient-centered delivery [17]. Another example of this is the Dutch vision document on network medicine, which emphasizes the importance of connected care networks to address complex patient needs comprehensively [18]. The United States also emphasizes patient-centered medical homes and patient-centered medical neighborhoods to foster comprehensive, patient-focused care [19,20].

In parallel, there has been an unprecedented evolution in information and communications technologies across all life domains—including health care—with a potential to revolutionize health care delivery. Innovations in telemedicine, electronic health records (EHRs), digital diagnostic tools, and mobile health apps exemplify how technology can enhance health care accessibility, efficiency, and patient outcomes [21,22]. These developments suggest a future in which health care is more personalized, data driven, and accessible than ever before.

Despite these technological advancements, organizational, political, and cultural constraints have hampered the effective integration of new information and communication technology solutions into health systems worldwide [23]. A notable issue identified elsewhere is the prevailing siloed approach to health care information management [16,24]. Most health care organizations still operate with fragmented systems, hindering the seamless collection, storage, and sharing of health information [16]. This compartmentalization jeopardizes the deployment of resilient health care models that can deliver person-centered care in an increasingly demanding context [25] (Figure 1 [25]).

This Viewpoint showcases the development and implementation of health ITs (HITs) in Catalonia over the past 25 years and frames them in the global context of digital health transformation. Owing to the early adoption of digital solutions, Catalonia represents a useful case study to describe the multiple interventions that can be implemented in health information systems at the country level and appraise the pros and cons of each of them. Therefore, the aim of this Viewpoint was to provide insightful information to countries that are endeavoring their journey through the digitalization of HITs. After describing the structural context of the Catalan health system and providing a chronological overview of the digital transformation of HITs, we thoroughly describe each of the HIT solutions, grouped according to their overall purpose into data integration for clinical support, patient-system interaction, and health care management and public health.

XSL•FO
RenderX

Figure 1. Paradigm shift in the health care delivery model—from vertical siloed care (A) to collaborative care (B; adapted from the work by Amelung et al [25]).

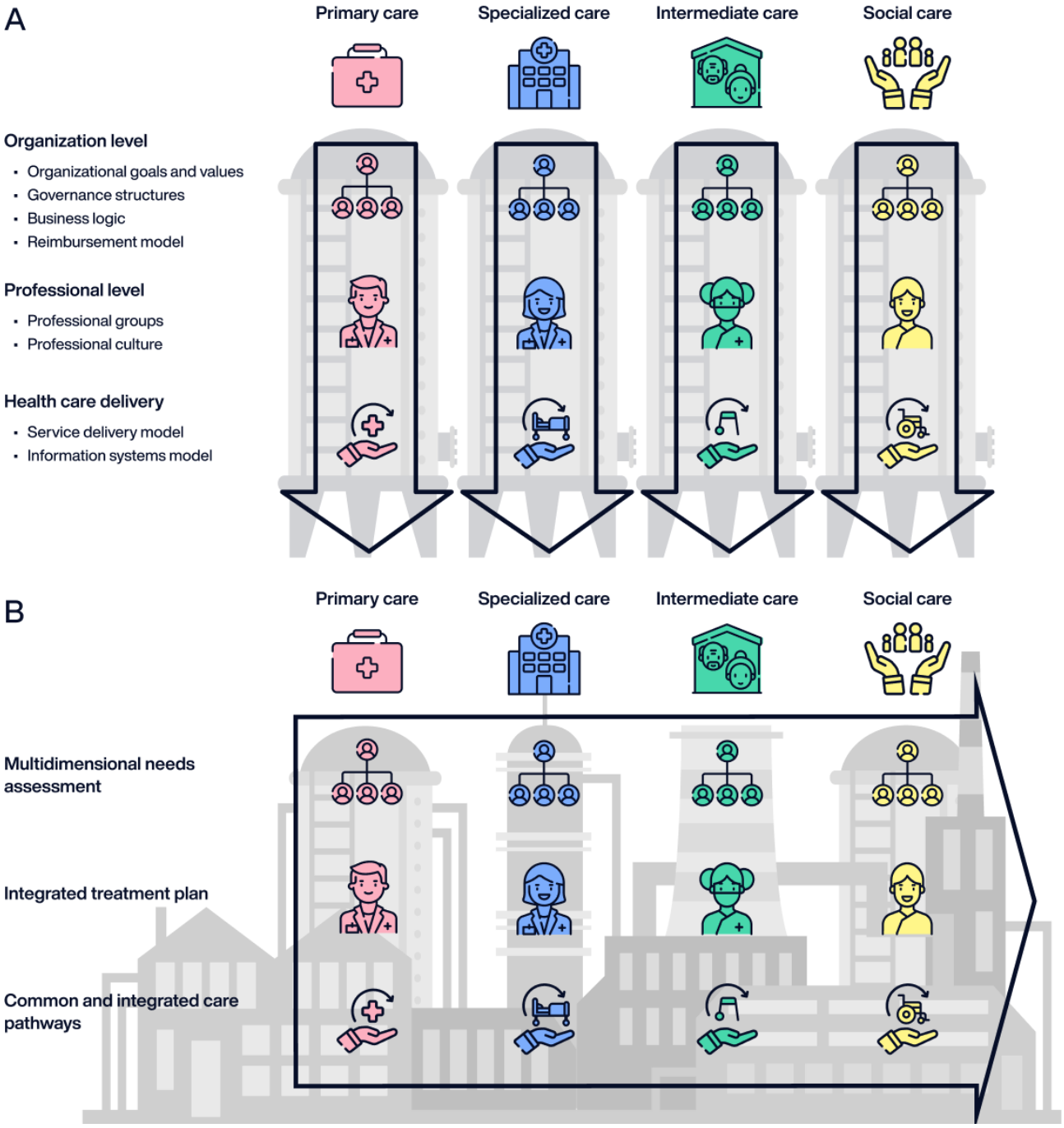

Context and Setting

Catalonia is an autonomous community in Spain with a population of approximately 8 million people distributed across an area of 32,114 km^2, primarily concentrated in urban areas. As in many other countries in Europe [26], the Catalan population has experienced an aging shift over the past few decades [27]. The health care system in Catalonia is integrated into the Spanish National Health System (NHS), primarily funded through taxation, and provides universal health care coverage to the entire population. Services under the Spanish NHS are generally provided free of charge at the point of delivery [28]. Nevertheless, according to the Catalan Health Survey, 32.5% of the Catalan citizens have a private voluntary insurance, mostly to cover complementary services to those under the statutory health system, such as reduced waiting times for specialized care and access to services that are restricted within the public benefits package [29].

In Catalonia, health care services are delivered through a network of providers that operate at four levels of care: (1) primary care (369 facilities); (2) specialized care, including public and semiprivate (publicly subsidized) hospitals (69 facilities); (3) intermediate and long-term care (96 facilities); and (4) mental health care (42 facilities) [30]. Health care in Catalonia is organized into 10 health regions, each with its network of facilities, which enables tailoring health care services

XSL•FO
RenderX

to the demographic characteristics of each area. Outpatient drugs are prescribed by physicians, either general practitioners (GPs) or specialists providing care through public coverage, and dispensed at private community pharmacies based on a copayment system [28] (Multimedia Appendix 1).

Catalonia exhibits a high degree of digital advancement, as evidenced by its 58% digital maturity index (a measure of the region's effective use of digital technologies to improve processes and services), a high satisfaction rate among e–government service users, and an estimated internet penetration rate of >93% [31].

The Never-Ending Metamorphosis

Overview

According to Vial [32], digital transformation can be defined as a process that aims to improve an entity by triggering "significant changes to its properties through combinations of information, computing, communication, and connectivity technologies." Digital transformation processes embody the dynamic and ongoing evolution of organizational operations, culture, and end-user experiences through the integration of digital technology [33]. This journey is inherently complex and multifaceted, reflecting the continuous innovation and adaptation that characterize the digital era. Digital transformations are not linear or predictable; instead, they are iterative, requiring constant adaptation to each cycle of technological advancement in response to new and changing market demands [34].

In the past 25 years, the landscape of HIT in Catalonia has been progressively reshaped through a transformation process that is still ongoing and started right after the first digitalization of medical records in the late 90s. Even though the health care sector has been slower than other industries in the adoption of technology, much has changed over the past 25 years. This section dives into the metamorphosis of the health care sector in Catalonia, catalyzed by the adoption of HIT within the period ranging from 1998 to 2023. First, we present the state of the health care sector at the beginning of the period and systematically describe the transformation journey and the drivers of transformation leading to the comprehensive eHealth platform (Figure 2).

Although the digital transformation milestones occurred subsequently on a timeline, for clarity and consistency, we present the results of this retrospective grouped into three categories of HIT infrastructures aimed at (1) integrating data to support clinical decision-making, (2) improving the interaction between patients and the health care system, and (3) health care management and public health.

Briefly, the journey kickoff was the introduction of administrative systems designed to handle patient appointments, billing, and other clerical tasks. The adoption of electronic medical records (EMRs) followed this first step, endeavoring an implementation process that was fostered by the Spanish entry into the European Union in 1986 and spanned the 90s. These early iterations of EMRs consisted of digitalizing patient records within specific health care settings, with limited—if not missing—data sharing capabilities. This marks the initial stage of the analyzed period, characterized by stand-alone solutions designed for serving single health care provider organizations. While these systems represented a significant step forward in managing patient information digitally, their isolated nature posed substantial challenges for the continuity of care.

Figure 2. Transformation road map in Catalonia, including the different systems comprising the Catalan eHealth platform for the period of 1998 to 2023. PROM: patient-reported outcome measure.

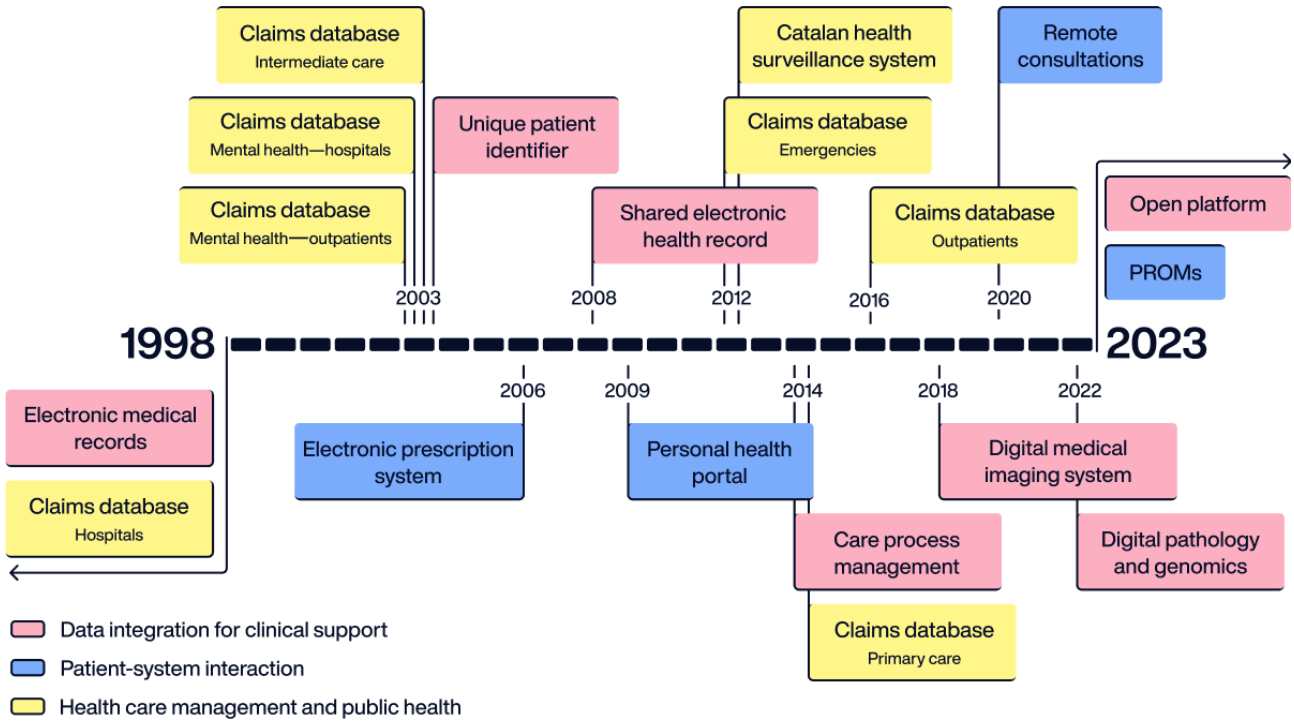

XSL•FO
RenderX

Data Integration for Clinical Support

Overview

Providing health care professionals with comprehensive information on patients has been for years the backbone of the digitalization of health information systems. For this information to be useful and safe, it should integrate the health data of a given patient regardless of the health care setting or place where they were created [35]. The universal health care coverage, paired with a unique patient identification number, was a mainstay for this approach. However, our health care landscape, with many providers with different health information systems, hampered the full integration of all patients' data and required implementing alternative solutions, such as shared health records or federated platforms.

Unique Patient Identifiers

In Spain, the cornerstone for integrating health information was laid in 2003 with the introduction of the individual health card (in Catalan, Targeta Sanitaria Individual), associated with a unique patient identifier code and established as a requisite for nonurgent health care access within the Spanish NHS [36]. This initiative also led to the creation of a protected database that collects essential information about the insured population. In Catalonia, this registry was named Central Registry of Persons Insured (in Catalan, Registre Central d'Assegurats) and contains essential demographic information needed for assigning health care resources to all Catalan inhabitants.

The interoperability of the unique patient identifier system is supported by a robust technical infrastructure that includes a communications network, standards for devices and cards, and an XML-based messaging system. This design allows for the fluid exchange of health information between regions, guaranteeing that users receive care and their prescriptions across Spain, thus reinforcing Spanish NHS cohesion and quality [37].

Catalonia, recognizing the potential of unique patient identifiers, has been transitioning toward a master patient index since 2020. This system aims to track all individuals receiving care within the region extending beyond Catalan and Spanish inhabitants (ie, including tourists and other people being served by the public health services). Built upon the Fast Healthcare Interoperability Resources (FHIR) standard, the master patient index allows for flexible data management and accurate patient matching across different providers within the ecosystem.

Shared EHR

At the center of the Catalan eHealth platform is the shared EHR of Catalonia (locally known as Història Clínica Compartida de Catalunya in Catalan), established in 2008. The shared EHR was founded on the Catalan Law 21/2000 about information rights, patient autonomy, and clinical documentation, which established "the objective to advance towards a unique medical record per patient" [38]. The main objective of this record is to improve the quality of health care delivery by providing an instrument that aids health care professionals in clinical decision-making by enabling access to relevant health information across Catalan health care providers [39].

From a technical perspective, the shared EHR was designed considering the fragmentation of health information systems in the Catalan health care ecosystem. Therefore, it was built by pooling together information collected by multiple EMRs (ie, 30 in total) in a central repository. Locally, information is collected following different standards but normalized according to international messaging and semantic interoperability standards (ie, Health Level 7 Clinical Document Architecture [HL7-CDA], Systemized Nomenclature of Medicine–Clinical Terms [SNOMED-CT], and Local Observation Identifiers Names and Codes [LOINC]) when transferred to the central record. The unique patient identifier works as a link for connecting information from different sources to a given patient. Service provider organizations can consume push-and-pull web services, allowing them to seamlessly integrate the information contained within their clinical workstations. For those who do not have the capabilities to integrate the information into their EMRs, a web-based viewer is provided.

From a security perspective, the record is only accessible by the homologated public service provider organizations through the virtual private network of the Catalan public health system. Security assertion markup language tokens are provided to service providers, and a 3-level security schema is followed (ie, patient identification, provider identification, and type of information), ensuring full traceability and legal compliance.

Since its inception, the shared EHR has been increasing the quantity and quality of the information collected, moving from semistructured documents into more granular information. Table S1 in Multimedia Appendix 1 provides an overview of the type of information contained and data volume. On December 31, 2023, the storage capacity of the database was 60 TB (including auditing). This system allows health care professionals to track their patients throughout the health care system, reviewing reports generated in other health care facilities (eg, due to emergency consultations, specialized interventions and tests, or regular consultations done in health care centers other than the regular one). Although the benefits of this record have not been formally assessed, it was deployed with the objective to reduce duplicated assessments, aid in diagnosis, and improve the continuity of care across the Catalan health care ecosystem.

Care Process Management

The care process management platform (locally known as Integrador de Serveis de Salut in Catalan), established in 2014, aimed to revolutionize the way in which health and social care services are coordinated and delivered, building on the principle that person-centered health care delivery often extends beyond the edges of a single organization [40,41]. The platform was designed by the Catalan Health Service, which maintains the specifications, and is provided free of charge to all service provider organizations within the public health care ecosystem in Catalonia. By providing a common language and configurable workflows, the platform enables diverse health care entities to interact efficiently, share critical information, and deliver coordinated care in a standardized way.

Functionally, the care process management platform targets key areas to improve the health care experience for all stakeholders. Citizens are expected to experience smoother transitions when

XSL•FO
RenderX

crossing organizational boundaries, reduced duplicated tests, and improved patient safety and security. Health care providers benefit from improved interoperability and integration with existing information systems, facilitating better care delivery and enabling a holistic view of patient health. Health services are expected to see enhanced resource planning, reduced waiting times, and an improved use of resources.

Technically, the platform is built on a foundation of interoperability and uses a range of technologies to ensure seamless communication and data sharing across different systems and stakeholders. Key technical features include a message broker for centralized messaging, rule engine for business logic, monitoring of tools for real-time data tracking, a configurable form tool, and a terminology server to standardize health care terminology across systems (Figure 3).

Figure 3. Architecture and main components of the care process management platform.

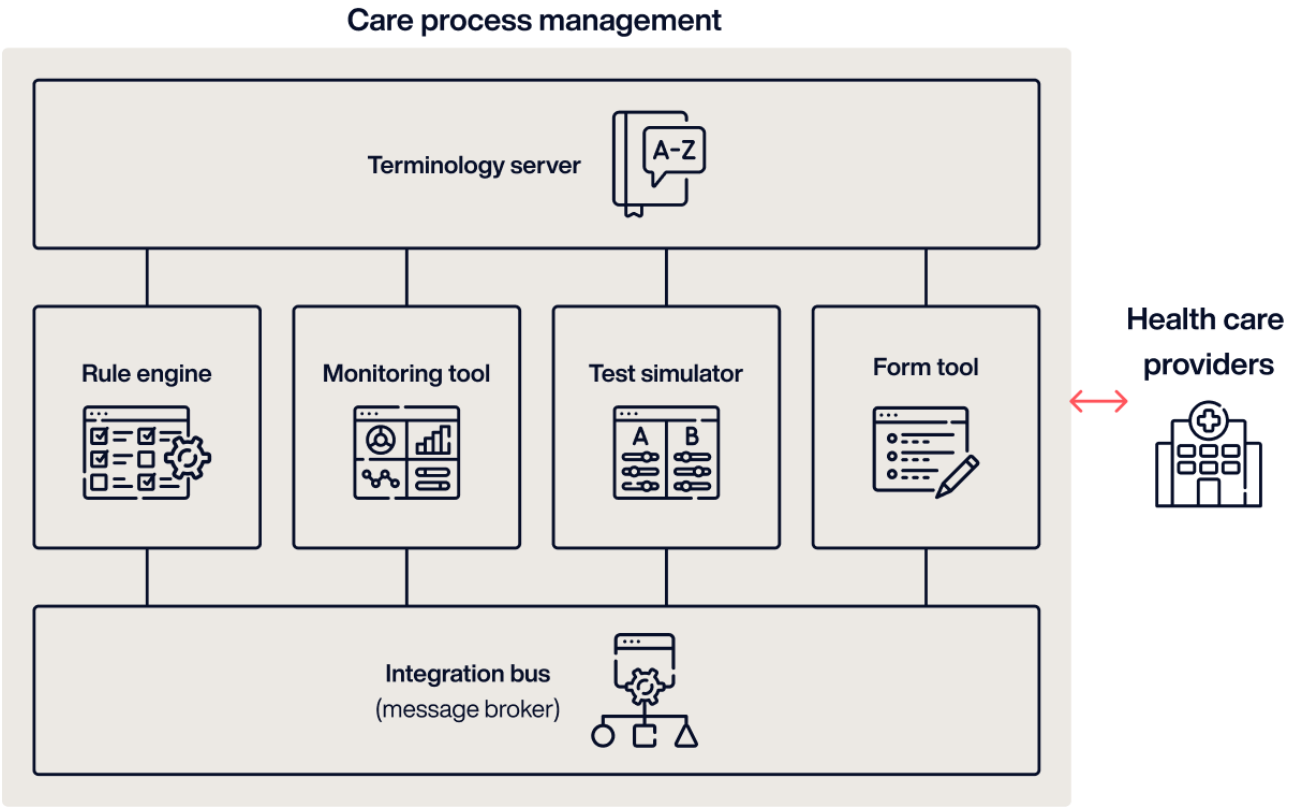

Interoperability is achieved through three main layers: (1) semantic, ensuring consistent meaning and use of data; (2) syntactic, standardizing the structure of communication through Health Level Seven messaging; and (3) technical, defining the communication channels and standards for data exchange. This layered approach ensures that information flows smoothly and meaningfully across different health care and social services, supporting better decision-making and facilitating continuous and integrated care.

The implementation of care process management platform spans various health care settings, including mental health centers, primary and specialized care, social care facilities, and emergency services. However, the model is mainly used as a referral system, and to date, it does not allow for setting integrated care workflows across providers.

Federated Platforms

While initially relying on a decentralized system (ie, every provider having its own infrastructure), challenges such as data fragmentation and barriers to innovation prompted a shift toward a federated model. In this model, the Catalan eHealth platform provides a central infrastructure to which all service provider organizations connect, avoiding the need for them to support big storage facilities, particularly for managing structured data such as medical images, pathology, and genomics. This federated approach overcomes important hurdles for data sharing and interoperability, such as diagnosis and treatment delay, and, at the same time, allows for economies of scale in terms of costs of data storage and security. In the future, this model will likely facilitate implementing machine learning algorithms for clinical decision support across the region due to higher volumes of data stored in a single repository.

Both the Digital Medical Imaging System of Catalonia (SIMDCAT) and the Digital Pathology Transformation of the Catalan Health Institute initiatives [42] are prime examples of how digital tools and platforms are integrated into this centralized model to improve the management and accessibility of medical images and pathology data (Figure 4). Following the same strategy, Catalonia is currently working on the construction of a federated network for genomics, starting with precision oncology [43]. The strategic focus on precision oncology aligns with the Catalan Health Service's investment in specific genomic panels for primary use, underscoring the importance of robust tracking for all interventions in this specialized area. This prioritization reflects an understanding of opportunity cost in health care resource allocation; by

XSL•FO
RenderX

focusing on precision oncology, Catalonia is maximizing the potential benefits of its investment in genomic technologies while acknowledging that resources directed there cannot be simultaneously invested in other areas. This approach allows for targeted implementation and evaluation of the federated genomics network in a high-impact area before potential expansion to other genomic applications.

Figure 4. Example architecture of a federated information platform—Digital Medical Imaging System of Catalonia and Digital Pathology Transformation of the Catalan Health Institute. AI: artificial intelligence; LAN: local area network; NHS: National Health Service; VPN: virtual private network.

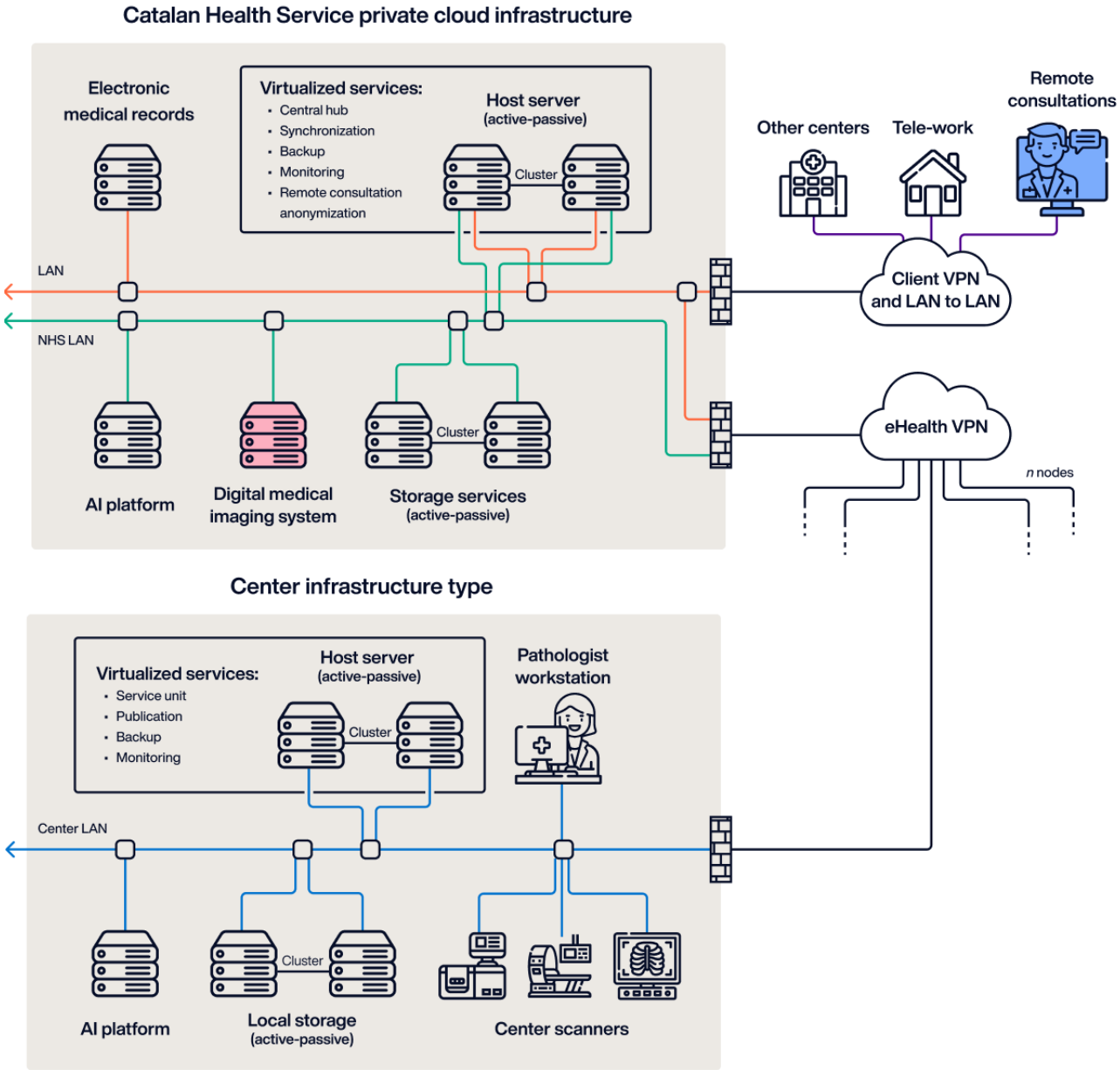

For reference, the current storage capacity of the Digital Pathology Transformation of the Catalan Health Institute amounts to 2.4 PB, corresponding to 37 million pathology exams (approximately 1 PB per year). Correspondingly, the SIMDCAT platform contains 3.7 PB of 118 diagnostic imaging technique examinations. Results from computed tomography and radiography, digital radiography, ultrasound, electrocardiography, mammography, and magnetic resonance account for 90% of the total examinations. Table S2 in Multimedia Appendix 1 provides the entire list of diagnostic imaging techniques and their volume currently stored in the SIMDCAT platform.

The improvement of communications in the region and the deployment of cloud computing have been the main technological driver for this transformation, which not only improves the operational aspects of health care delivery but also sets the stage for future innovations that can further enhance the effectiveness and efficiency of health care services in the region.

Patient-System Interaction

Overview

The shift from diagnosis-centered to person-centered care and the increasing involvement of patients in their care are among the most important changes in health care routines in the past

XSL•FO
RenderX

few years and are acknowledged as high-quality practices for future health care systems [44,45]. This approach requires allowing patients access to their own health care data, thus adding an additional challenge for health information infrastructures.

The drug prescription system and the personal health portal were significant breakthroughs in our digitalization journey that enhanced the interaction between patients and the health care system. The consolidation of these 2 milestones allowed us to build more innovative tools, such as remote consultations and patient-reported outcome measures (PROMs).

Personal Health Portal

With the development of the My health platform in 2009, Catalonia was among the leading regions in early implementation of patient health portals [46]. My health is a comprehensive tool that allows citizens to access and manage their personal health information. Users of this platform have the convenience of viewing their medical records web-based, which includes details such as physicians' notes, test results, medications, and information about their vaccinations and allergies. One of the key services of the platform is the ability to manage appointments with health care professionals. Users can schedule, reschedule, or cancel their appointments through the platform, making health care more accessible and user-friendly [47].

The platform also provides users with important information about their health plans, including what services and coverage they have as well as test results, including laboratory assessments and reports of diagnosis techniques. This can be particularly helpful in understanding and navigating the health care system. In addition, My health includes a variety of health resources and educational materials, offering guidance and information on preventive care and general health maintenance.

Technologically, the platform is a subset of all the objective information in the shared EHR of Catalonia. This integration ensures that users have a comprehensive and up-to-date view of their health record and current treatments. This aspect of the platform underscores the broader move toward eHealth solutions that aim to enhance the efficiency, personalization, and accessibility of health services.

Even though the My health platform was deployed in 2009, its use was low among the population. The COVID-19 pandemic was a great opportunity to boost its use by moving added-value services into it [48,49]. The addition of electronic prescriptions and sick leave forms, COVID-19 digital certificates and tests results, and immunizations made the number of registered users explode (increasing from 600,000 in March 2020 to 5.8 million in December 2023). Currently, a median of 376,000 users access the system on a daily basis.

Electronic Prescription System

Although originally developed for invoicing purposes, the electronic prescription system has dramatically changed patients' routines, shifting from regular face-to-face visits to their GP to collect a handful of messy handwritten papers to structured medication plans accessible by the GPs, specialists, pharmacists, patients, and the public health insurer.

In Spain, drug prescription is tightly controlled, limiting the issue of one medication per prescription and up to 3 months of advanced prescriptions [50]. Initially, a paper-based system required patients to visit their GP quarterly to obtain multiple prescription forms for each long-term medication, which they would then take to pharmacists for dispensing, often receiving all 3 months' supply at once. This handwritten process risked medication errors and made it difficult to monitor patient medication plans, increasing the chance of dangerous drug interactions, especially with the addition of acute medications. Moreover, despite the advice to pick up medications monthly, many patients collected them all at once, leading to waste if medication plans changed.

In the 2000s, the Catalan Health Service led the first deployment in Spain of a single electronic prescription system (locally known as Sistema Integrat de Recepta Electrònica in Catalan), aimed to be used for outpatient prescriptions across the entire public health care ecosystem in Catalonia. Through liaison with the Catalan College of Pharmacists, a mirrored platform was built to support dispensation and data collection from the 3200 community pharmacies for invoicing purposes. The College of Pharmacists worked to progressively harmonize the different versions of pharmacy management software to ensure connection and performance of the dispensation and invoicing process through the electronic prescription service (Figure 5). The process was piloted in 2006 and came into effect in 2007, reaching universal coverage 2 years later in 2009 [51-53].

The electronic prescription system allowed professionals to prescribe into a single, shared, and traceable medication plan and apply prescription support systems with basic checks for safety and drug interactions. The national regulation was updated to cover electronic prescription, allowing for the programming of successive monthly prescriptions for chronic medications for up to 1 year (automatically released to patients once a month), thus reducing the number of medical visits solely aimed at collecting new paper prescriptions while preventing patients from collecting an excess of packages in advance [50]. During the COVID-19 crisis, the electronic prescription system platform was also used for prescribing and invoicing antigen rapid-diagnosis tests done at community pharmacies for mass-testing strategies in schools and other community settings. The COVID-19 pandemic also encouraged ruling out paper for medication plans, with an estimate saving of 46 million paper sheets (ie, 600 tons of paper) [54]. The Catalan electronic prescription system is fully connected to those of the other Spanish regions and with the European ePrescription and eDispensation initiative [55].

Hospital medication has traditionally received less attention regarding information integration due to the localized activity and higher control by hospital care. However, outpatient medication prescribed from hospitals has growth, resulting in an increasingly high treatment burden and economic impact. In the mid-2010s, a dedicated registry (Registre de Pacients i Tractaments in Catalan) was created with a twofold purpose: to (1) keep a track record of the use and main clinical results of

XSL•FO
RenderX

medications and (2) provide a tool for supporting patient-level invoicing of medication by hospitals. This registry has allowed for the assessment of quality and effectiveness for clinical and managerial purposes as well as the implementation of several outcome-based drug access agreements [56].

The next step in HIT related to drug prescription is to integrate all prescription information regardless of the setting (ie, in-hospital, outpatient, or community) into a single therapeutic plan that allows for safety surveillance and knowledge-based management of therapeutics. This strategy is being implemented within the Information Systems Master Plan of the Catalan Health Service [57].

Figure 5. Information flow within the electronic prescription system and general use figures in 2023.

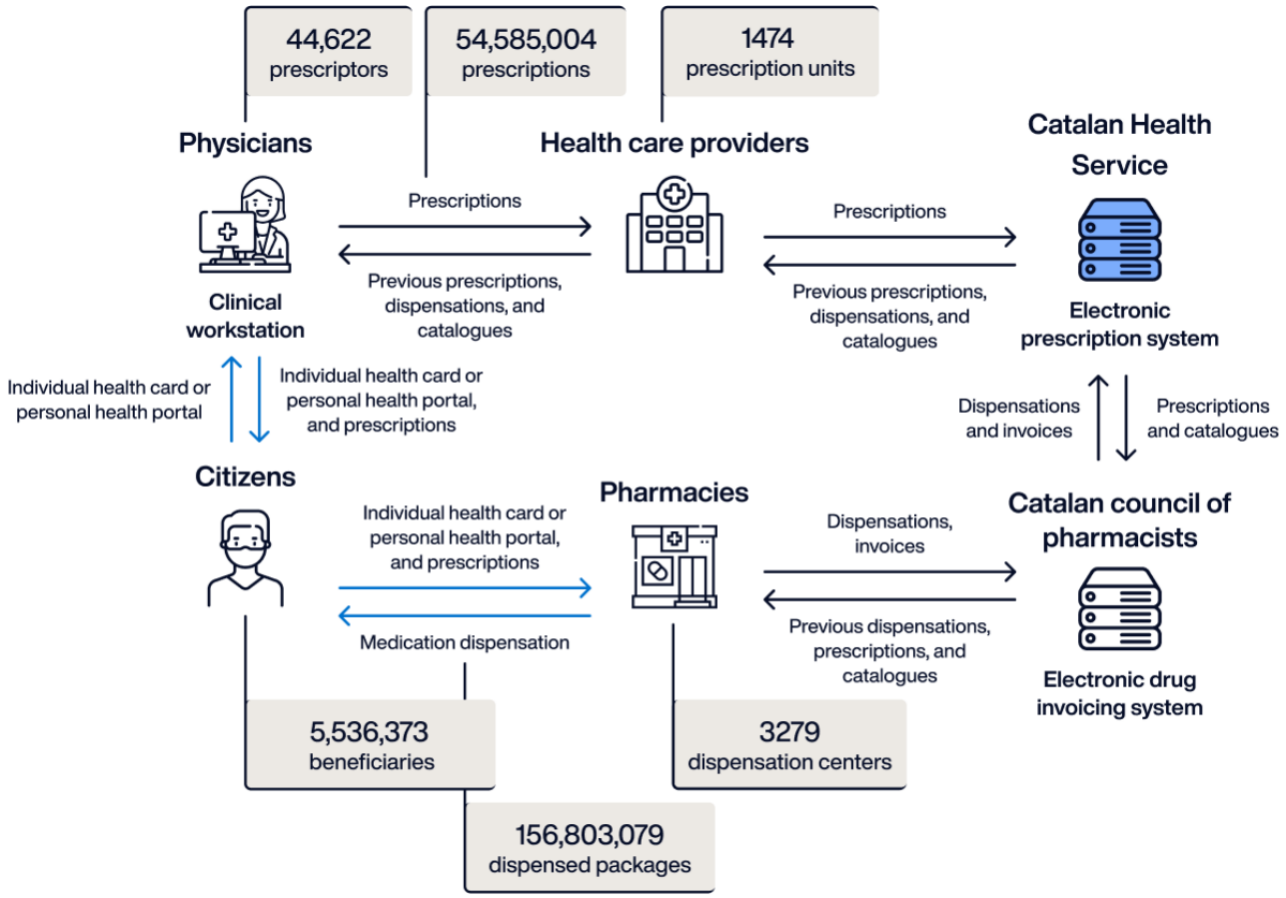

Remote Consultations

In 2015, Catalonia witnessed an important transition to remote consultations, which began with the implementation of eConsultation, a digital platform that enables asynchronous communication between patients and health care professionals, facilitating health care continuity. eConsultation allows for 2-way asynchronous communication, enabling the exchange of messages, medical documents, reports, and test results between patients and health care providers [58,59]. This transition from traditional face-to-face interactions to a more flexible and efficient digital interface marks a significant evolution in health care delivery methods.

During the early stages of eConsultation implementation, this modality of asynchronous consultation was used by young patients, mostly women, with simple health issues [59]. Health care professionals, including physicians and nurses, mainly used eConsultation for consultative and managerial tasks [58]. The onset of the COVID-19 pandemic, which led to nationwide lockdowns dictated to contain the spread of SARS-CoV-2, catalyzed a significant uptick in eConsultation use, facilitating its broader acceptance and integration into the daily practices of health care professionals (Figure 6). The intensive use of eConsultation, including proactive patient communication, issue of sick leave certifications, prescription renewal, and addressing general medical inquiries remotely, shifted toward a hybrid model that melds traditional health care with telemedicine approaches and heralds a new era in health care delivery.

XSL•FO
RenderX

Figure 6. Monthly evolution of patient visits to health care professionals in Catalonia (January 2020-January 2024).

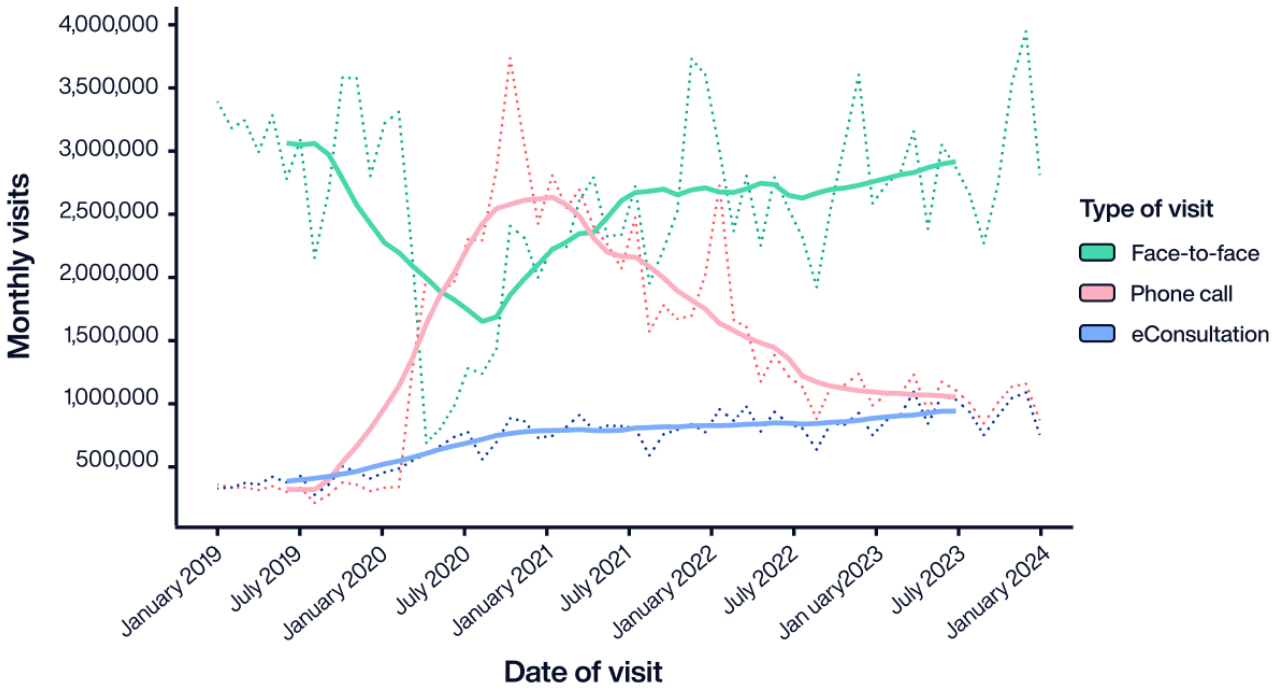

The COVID-19 crisis boosted not only asynchronous remote consultations but also video consultations. The accelerated and hasty implementation of remote consultations increased the risk of reducing the quality of health care delivery. Therefore, Catalonia endeavored an initiative (ie, the "Latitud" model) for an evidence-based implementation of remote consultations driven by a multidisciplinary team including experts from the health care, legal, and eHealth domains. Using an iterative process, which spanned May 2020 to June 2020, this team developed and disseminated evidence-based guidelines for health care professionals regarding the use of remote consultations [60]. The recommendations for professionals and organizations regarding the use of remote consultations are available in the Multimedia Appendix 1. This use framework is currently being assessed in a randomized controlled trial of synchronous remote consultations in the outpatient setting [61].

According to retrospective reports, remote consultations solved 88% of the cases, indicating a high capacity for reducing the work burden of health care centers [62]. This efficiency was particularly pervasive in the management of common issues such as test results, medical inquiries, and repeat prescriptions, which streamlined interactions and reduced unnecessary in-person visits. Of note, the reduction in face-to-face visits was concomitant with an increase in the demand for health care support due to its ease of access, potentially adding to the GPs' workloads. Aside from the increased workload, limited digital literacy among some patient groups and the need for appropriately integrating telemedicine activities into professional workflows were identified as important challenges for the implementation and persistence of remote consultations in Catalonia [59].

Electronic Prescription of PROMs

PROMs are standardized questionnaires collecting patients' health outcomes, directly provided by them with neither input from nor interpretation by health care professionals [63]. PROMs, often used for diagnosis or follow-up purposes, include a wide range of outcomes, such as symptoms, quality of life, psychological features, functional status, social issues, response to treatment, or patient satisfaction [64]. Electronic PROMs (ePROMs) allow for data collection via the internet on devices, increasing response rates. Furthermore, electronic collection of PROMs facilitates automated integration into EMRs, aiding clinical decision-making and promoting continuity of care. In recent years, ePROM prescription has been increasingly used in specialized care [65-68]; conversely, their use in primary care, where patient profiles are more heterogeneous, is still scarce [63,69,70].

In November 2022, an ePROM prescription module was implemented in the workstation of health care professionals at the system level. The design and implementation were driven by a panel of experts, including health care professionals and managers, and members of the internet-based PROM library in Spain (BiblioPRO) [71]. Implementation was carried out in three phases: (1) pilot implementation of 4 ePROMs in the primary care workstation of 11 preselected primary care centers, (2) scale-up across the whole primary care network, and (3) integration of the tool into the hospital's workstations. Further enhancements of the tool were introduced in the months following implementation, including reminders to patients, integration into the personal health portal, integration into the shared EHR viewer, and addition of new ePROMs (Figure 7).

XSL•FO
RenderX

Figure 7. Progression of electronic patient-reported outcome measure (ePROM) prescription from pilot implementation to February 29, 2024. The dotted line shows the weekly crude rate, and the full line shows the 7-day moving average. The vertical lines represent key implementation events. EMR: electronic medical record.

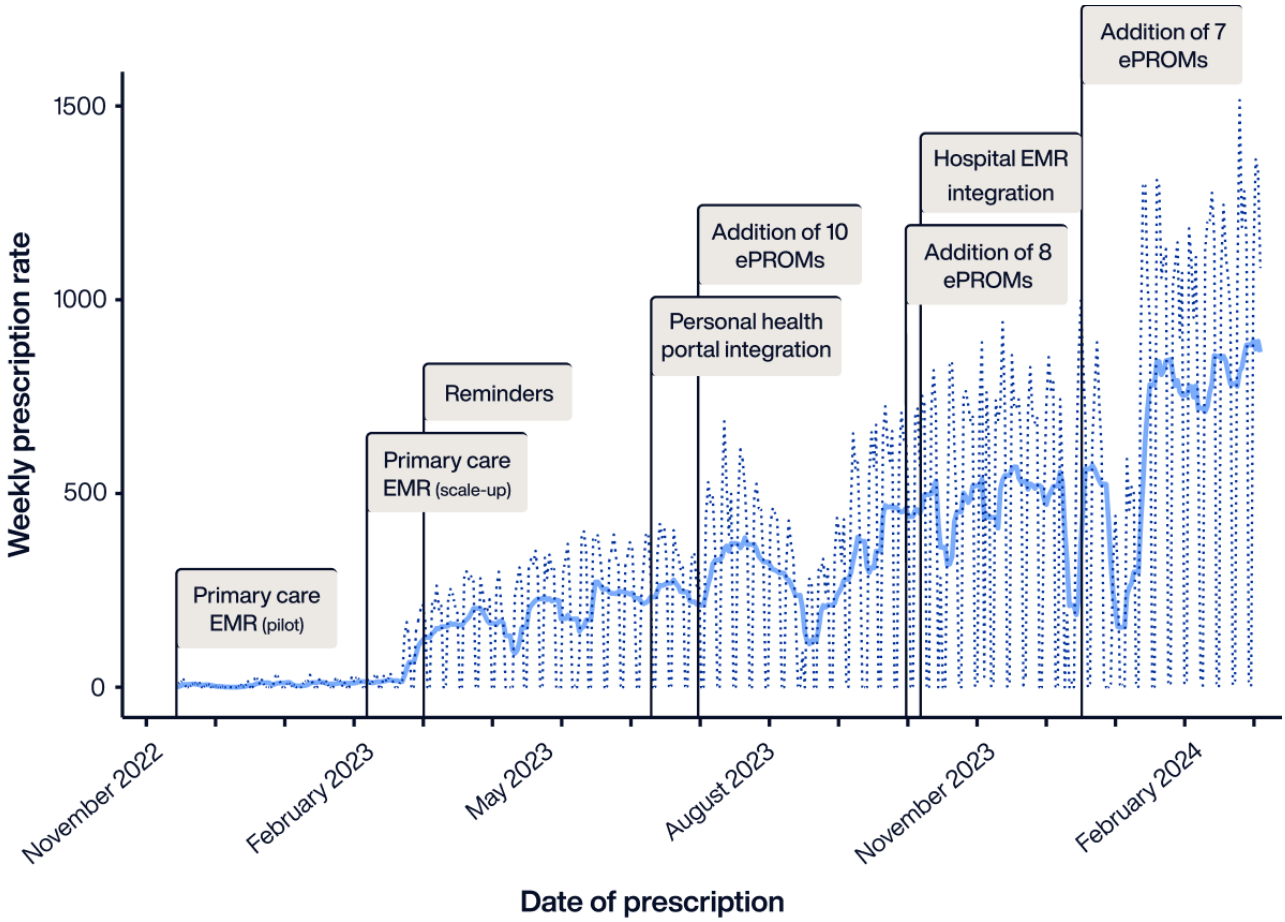

Currently, health care professionals in our system can prescribe ePROMs among a set of 29 ePROMs for several conditions (Table S3 in Multimedia Appendix 1). The prescription of ePROMs has increased since their implementation in the Catalan health care system (Figure 6). As of February 29, 2024, a total of 142,979 ePROMs had been prescribed by 6259 health care professionals, with 112,424 of these ePROMs being completed by patients, resulting in an average response rate of 78.7%. Therefore, in addition to supporting clinical decisions in routine care, ePROMs provide a significant amount of health information from the catchment population, paving the way for secondary use of data for benchmarking or research.

Health Care Management and Public Health

Overview

Data collected with the primary purpose of health care delivery hold high potential for knowledge-based decision-making in nonclinical settings. Important activities for the quality of health, such as resource allocation, targeted public health or population-based assessment, and monitoring are currently based on the secondary use of clinical data. The single public insurer for health care services in Catalonia encouraged building registries and tools for benchmarking, invoicing, and monitoring the health status of the population.

Claims Registries

The earliest stages of the centralization of health care information were driven by a Spanish directive regarding the registry of specialized health care activities, which was prompted by the European initiative of defining minimum basic datasets (locally known as Conjunt Mínim Bàsic de Dades in Catalan) of hospital and outpatient activities as a foundation for the management of health care services [72]. Different claims registries were consecutively created to cover several stakeholders and care levels, including primary care, hospitalization, emergency room, emergency transportation, specialized outpatient care, intermediate care facilities, mental health hospitals, and outpatient mental health.

The claim registries differ in the type of information collected, although they share a similar structure that includes a unique identification number common for all registries of the Catalan Health Service. Similarly, the updating frequency depends on the purpose of each registry (Table 1).

The claim registries have an automated data validation system aimed at identifying inconsistencies between variables (eg, age-diagnosis correspondence and temporal sequences) and undergo external audits periodically to ensure provider payment accuracy

XSL•FO
RenderX

Table 1. Minimum basic datasets established by the Catalan Health Service.

Registry	Care level	Type of information	Variables, n	Year launched
Hospitalization	Specialized care	Admissions to hospitals, including stays and procedures	159	1992
Primary care	Primary care	Visits to primary care services (nurses or physicians)	82	2014
Emergency	Specialized care	Admissions to the emergency room	65	2012
Outpatient mental health	Specialized care	Outpatient visits to mental health professionals	75	2003
Mental health hospitals	Specialized care	Admissions to mental health hospitals	52	2003
Intermediate care	Intermediate care	Admissions to intermediate care and skilled nursing facilities	Palliative care: 31 and long-term care: 169	2003
Outpatients	Specialized care	Diagnoses and procedures of outpatient visits	72	2016

From Benchmarking to Population Health Assessment

The systematic collection of health information for health care claims, which included demographics, diagnoses, and use of health care resources, allowed for the development of registries aimed at public health surveillance. In 2011, the Catalan Health Surveillance System was created as an integrated platform periodically gathering data from claims registries and drug billing registries. This integration allowed for a comprehensive view of Catalan patients in a single database that included demographic characteristics (ie, age, sex, and socioeconomic status), clinical characteristics (ie, all chronic and acute diagnoses), and resource use (hospitalization, primary care visits, emergency department visits, skilled nursing facilities, palliative care, mental health services, pharmacy dispensation, outpatient visits to specialists, home hospitalizations, medical transportation [urgent and nonurgent], ambulatory rehabilitation, respiratory therapies, and dialysis).

The Catalan Health Surveillance System set up 2 important resources: the platform of modules for the follow-up of quality indicators (Mòduls per al Seguiment d'Indicadors de Qualitat in Catalan) and the Adjusted Morbidity Group (AMG) case-mix tool [73].

The modules for the follow-up of quality indicators are a business intelligence tool for the Catalan health care system, providing key performance and morbidity indicators at the hospital and emergency room levels. It offers data on demographics, health care interactions, and the prevalence of the top 20 chronic diseases available for the overall population and by health care area. In addition, it includes health care quality metrics such as in-hospital mortality, complications, readmissions, and preventable hospitalizations for each area.

More recently, the integration of indicators on morbidity and resource use has evolved toward public business intelligence for benchmarking, in which citizens and providers can review several key performance indicators among primary care teams and hospitals [74].

AMG is a case-mix tool that considers all possible chronic conditions (and recent acute diagnoses) among the diagnosis code groups of the International Classification of Diseases, 10th Revision, Clinical Modification [73,75]. On the basis of a model that accounts for differences in key health outcomes, the tool provides a morbidity index representing the clinical complexity of patients and allocates each patient to a morbidity group based on their complexity. The numerical index can be used to stratify the population into health risk groups. The AMG tool has shown high performance in predicting key health outcomes [7,73,76] and has been used as an adjusting variable to summarize the comorbidity burden in many research projects and internal benchmark analyses to account for the clinical complexity of individuals tended to by health care teams or providers being compared in the settings of research and resource allocation [77,78]. In the COVID-19 setting, a similar approach was used to develop a specific risk tool that predicted hospitalization or death beyond age [79].

Discussion and Future

The transformation journey of the Catalan health care system over the past 25 years illustrates the efforts toward a paradigm of health care delivery also envisioned by other countries that is more integrated, efficient, and person centered. This transformation has occurred concomitantly with the adoption of HIT.

Revisiting the international examples mentioned in the Introduction section provides valuable context for the Catalan HIT journey. The ICSs of the National Health Service in the United Kingdom share similarities with Catalonia's approach, particularly in their aim to better coordinate services across health and social care. However, while the UK ICS is a relatively recent initiative, Catalonia's integrated care efforts, supported by its HIT infrastructure, have been evolving for >2 decades. This longer-term perspective offers insights into the challenges and benefits of this integration effort, which have been extensively reported elsewhere [78,80-84]. The Dutch vision on network medicine emphasizes connected care networks to address complex patient needs, an approach that aligns closely with Catalonia's care process management platform. The implementation of this concept in Catalonia through tangible HIT solutions provides a practical example of how such a vision can be feasible. In comparison to the US patient-centered medical homes and neighborhoods, which focus on comprehensive, patient-focused care primarily in primary care settings, Catalonia's approach extends this person-centric model across the entire health care continuum, facilitated by its integrated HIT systems. The Catalan experience, with integrated platforms such as the shared EHR, the care process management

XSL•FO
RenderX

platform, or the federated platforms, demonstrates how a unified HIT infrastructure can support and enhance these person-centered care models on a system-wide scale despite the absence of a common health information system.

Catalonia's early adoption of EHRs in the late 1990s and early 2000s placed the region at the forefront of the global EHR movement, aligning with trends observed in other European countries, although the pace of adoption varied significantly worldwide. A study by Jha et al [85] found wide variation in EHR adoption rates across 7 nations as of 2005. The Netherlands led with nearly 100% adoption in primary care, a trend that has continued and expanded to other countries [85]. As of 2016, countries such as Australia, the Netherlands, New Zealand, Israel, and Germany had achieved near-universal EHR adoption among physicians [85]. However, these rates of EHR adoption have not been mirrored in all high-income nations. Canada and the United States, for instance, had implementation rates as low as 25% by 2009 [86]. In response to this lag, the US federal government took decisive action through the Health Information Technology for Economic and Clinical Health Act, committing unprecedented resources of up to US $27 billion over 10 years to support the meaningful adoption and use of EMRs and EHRs through incentive payments [87]. This significant investment accelerated EHR adoption in the United States, highlighting the importance of governmental support and financial incentives in driving technological transformation in health care. In addition to governmental support, the progress of Catalonia in this area was facilitated by its unified health care model and single public insurer system, which provided a more coordinated environment for technology implementation compared to more fragmented health care systems. This structural advantage allowed Catalonia to achieve high adoption rates without the need for large-scale federal incentive programs such as those implemented in the United States.

Aside from the broad adoption of EHRs, Catalonia has made significant strides in patient-professional interaction through its "My Health" platform [46]. In this regard, it is important to also consider the advancements made by other European countries, particularly Denmark and Estonia, in this area. Denmark has been a pioneer in developing patient portals. The Danish eHealth portal [88], launched in 2003, provides citizens with comprehensive access to their health data and a range of eHealth services. According to Kierkegaard [89], by 2010, over 85% of the Danish population had used the portal. The portal allows patients to access their full medical records, book appointments, renew prescriptions, and communicate with health care providers. Denmark's success in this area is attributed to its long-standing focus on eHealth and strong coordination between national and regional health authorities [89]. Estonia, known for its advanced digital society, has also made remarkable progress in eHealth. The Estonian eHealth system, launched in 2008, includes a patient portal that gives citizens access to their health records, e-prescriptions, and various e-services [90,91]. According to a case study conducted by Kivekäs et al [91] in 2017, a total of 47% of the Estonian population had accessed their health data through the portal. Estonia's approach is notable for its integration of health data with other digital services through the country's X-Road data exchange layer, providing a high level of interoperability and security. Both Denmark and Estonia's experiences highlight the importance of strong national eHealth strategies, robust digital infrastructure, and a culture of digital innovation in driving the adoption and use of patient portals [92].

In our experience, in Catalonia, one of the critical drivers of this transformation has been the political commitment to a unified health care delivery model, facilitated by a single public insurer with direct economic leverage on service provider organizations. This foundational feature enabled the implementation of key infrastructures, such as the unique patient identifier, which became a cornerstone for subsequent integrations and innovations in the health care ecosystem. The lack of such a unique identifier, explicitly discouraged in some countries such as the United States, requires algorithmic approaches to information integration that can lead to errors and reduced quality of care [37]. On the other hand, unique identifiers have raised concerns regarding the privacy of patient information [37]. The historical willingness of Spanish governments to have a single and public health insurer that operates through a unique identifier has significantly facilitated the creation of HIT structures with a person-centered perspective and is a cornerstone of the Catalan eHealth platform.

The unique patient identifier was a key facilitator of the shared EHR launched in 2008, relatively early compared to many other regions. This initiative represents a significant step toward integrated care, aligning with broader international trends in health care digitalization and service integration. Denmark, often cited as a leader in HIT, launched its national health portal with shared EHR access in 2003. The Danish health data network allows health care providers to access patient data across different care settings, facilitating continuity of care and reducing duplication of services. By 2010, nearly all primary care physicians in Denmark were using EMRs and sharing data through the national network [89]. In the United Kingdom, the National Health Service has been working toward a more integrated approach to health records and care delivery. The National Health Service Long Term Plan, published in 2019, sets out a vision for ICSs across England. A key component of this strategy is the development of shared care records across health and social care providers within each ICS [17]. The aim is to enable seamless information sharing between different care settings, improving coordination and patient outcomes. While implementation is still ongoing, early adopter regions such as Greater Manchester have made significant progress in developing shared care records across their ICSs [17]. Similarly, in the United States, health information exchanges have been developed to facilitate the sharing of electronic health information across organizations within a region, community, or hospital system. For example, the Indiana Health Information Exchange, one of the oldest and largest in the United States, connects >100 hospitals and 14,000 practices, allowing providers to access patient data from multiple sources [93]. Australia has also made strides in this area with the My Health Record system, a national EHR system launched in 2012. By 2019, over 90% of Australians had a My Health Record, allowing them to share their health information with health care providers across the country [94].

XSL•FO
RenderX

The Catalan shared EHR system stands out for its comprehensive approach, integrating data from multiple providers across the region. The system allows health care professionals to access relevant patient information regardless of where it was generated within the Catalan health care system. This level of integration has been challenging to achieve in many other health care systems, particularly those with more fragmented provider landscapes. However, the implementation of shared EHR systems is not without challenges. Issues of data privacy, security, and interoperability remain significant concerns worldwide. A study by Fragidis and Chatzoglou [95] highlighted that, while many countries have made progress in implementing national EHR systems, barriers such as lack of standardization, resistance to change, and concerns about data protection continue to hinder full adoption and use. Despite these challenges, the trend toward shared EHRs continues to gain momentum worldwide, driven by the potential benefits of improved care coordination, reduced medical errors, and more efficient health care delivery.

Catalonia's implementation of remote consultations in 2015 positioned the region at the forefront of telemedicine adoption, a trend that was dramatically accelerated by the COVID-19 pandemic. Initially, remote consultations were primarily used by younger patients for straightforward health concerns; however, the pandemic prompted a significant expansion of their use, becoming a crucial tool for maintaining health care services [48,59]. This shift mirrored global trends, with many countries rapidly adopting telemedicine to ensure continuity of care while minimizing viral transmission risks [96]. The response of the Catalan health system to this surge included the development of the "Latitud" model, an evidence-based framework for implementing remote consultations, addressing challenges such as technical issues and the need for physical examinations identified in other contexts [60,97]. The effectiveness of remote consultations in Catalonia has been promising, with retrospective reports indicating the resolution of 88% of cases, demonstrating a high potential for reducing health care center workloads [62]. However, challenges remain, including limited digital literacy among some patient groups and the need to integrate telemedicine effectively into professional workflows [59]. Looking forward, health care systems, including in Catalonia, are likely to adopt a hybrid model blending in-person and remote care. The Catalonia experience offers valuable insights for other health care systems, highlighting both the potential of remote consultations to improve health care accessibility and efficiency and the importance of evidence-based implementation strategies to ensure quality of care in this new paradigm.

The limitations of the Catalan digital health infrastructure became increasingly apparent as the demand for more integrated and patient-centered services grew [5,84,98,99]. The fragmented nature of health care data, confined within proprietary systems and inaccessible across different health care settings, led to inefficiencies, effort duplication, increased costs, and barriers to the scaling up of innovation. Moreover, the rising expectations of patients, who increasingly sought more involvement in their health care decisions and access to their health information, underscored the need for a more open and connected approach [100]. Finally, it is worth mentioning that, despite the large amount of data routinely collected and centralized by the Catalan Health Service, many decisions regarding planning and resource allocation are not yet data driven and most health indicators of the Catalan Department of Health still rely on population surveys, as they did years before [30]. This scenario suggests an underuse of the data that goes beyond technological barriers.

Recognizing these challenges, Catalonia embarked on a journey toward the open platform paradigm in health as defined in the digital health strategy for Catalonia [57]. The idea of the open platform is a forward-thinking approach first defined by the Apperta Foundation, which emphasizes the importance of interoperability, open standards, patient-centered care, and the strategic use of digital technologies to enhance health outcomes [101]. It entails a shift away from proprietary, siloed systems toward a more integrated, flexible, and user-friendly digital health ecosystem. This paradigm advocates for the idea that health care should be accessible, transparent, and seamlessly connected, enabling the free flow of health data to support patient care, research, and innovation. The underpinning concept is that information systems for patient data collection and clinical information share the need for an information architecture (ie, International Organization for Standardization 18308:2011) based on standardized or logical information models [102]. This element is key for a satisfying experience for users—both professionals and citizens—in all their interactions with the health system tools, facilitating updates and innovations in EHR-related applications used by health care providers, and centralizing patients' personal health data. The technological progress makes it essential to push the current deployment of HIT from interoperability to shared and common vocabularies and ontologies, achieving full data liquidity between organizations and systems through the meaningful use of different standards [103] (Figure 8).

XSL•FO
RenderX

Figure 8. Architecture of the envisioned open platform of the Catalan Health Service (adapted from Defining an Open Platform—Apperta Foundation). API: application programming interface.

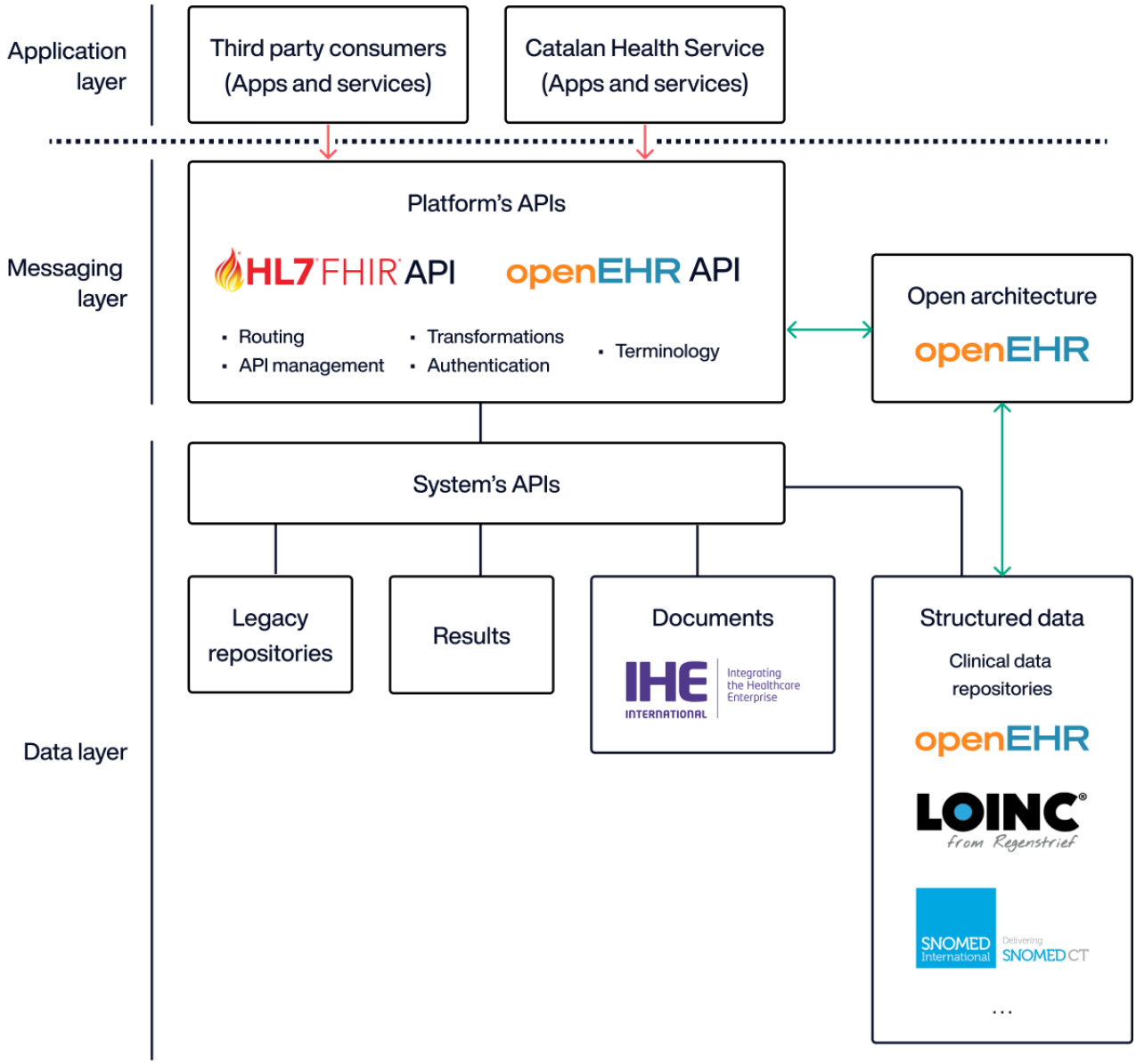

Looking ahead into the future, Catalonia will continue to evolve its HIT infrastructure on the open platform paradigm. The focus will be on further enhancing integration, leveraging advanced technologies, and improving patient engagement. A key priority is the seamless integration of health and social care data, which would support a more holistic approach to patient care and enable better management of complex, chronic conditions. Concurrently, there is significant potential in harnessing artificial intelligence to improve clinical decision support systems and refine population health management strategies. These advanced analytics could provide invaluable insights for preventive care and resource allocation. Patient empowerment remains a crucial area for development, with opportunities to expand and improve digital tools that facilitate active participation in health care decisions, including better telemedicine platforms and more intuitive interfaces for accessing health care services. In addition, addressing any remaining interoperability challenges is vital to ensure seamless data flow across all health care settings, from primary care to specialized hospital services. This would not only improve care coordination but also support more comprehensive research and public health initiatives.

As final remarks, for countries embarking on their HIT journey, the Catalan experience offers several key considerations:

- Political and structural foundations—a unified health care model and strong political commitment are crucial for successful implementation.
- Phased approach—start with foundational elements such as unique patient identifiers and basic EHR systems before moving to more advanced integrations.
- User-centric design—prioritize the needs of both health care providers and patients in system design to ensure adoption and effectiveness.
- Data governance—establish robust data privacy and security measures from the outset to build trust and ensure compliance with regulations.

XSL•FO
RenderX

- Interoperability—adopt international standards for data persistence and exchange to facilitate future integrations and avoid vendor lock-in.
- Change management—invest in training and support for health care providers to ensure smooth adoption of new technologies.
- Continuous evaluation—regularly assess the impact of HIT implementations on health care quality, efficiency, and patient outcomes to guide future developments.

Conclusions

The evolution of HIT in Catalonia over the past quarter of a century provides valuable insights and lessons for other regions aiming to navigate the complex landscape of digital health care transformation. However, the effectiveness of such strategies is contingent on a robust primary care network and a unified health care model. Hence, the scope of our analysis and applicability of most of the solutions described in this paper are particularly suited to countries with universal health care coverage, a unique patient identification number, and a strong network of primary care centers.

The future of health care digitalization in Catalonia and potentially elsewhere hinges on overcoming legacy data fragmentation and moving toward a genuinely integrated health and social information system that streamlines the strategic use of health data. This transition is not just about adopting new technologies but also about a paradigm shift in health care delivery, moving from referrals and transferability of responsibility toward a collaboration emphasizing a holistic approach to patient care for improved health outcomes.

Acknowledgments

This work is a summary and a retrospective analysis of the deployment of health IT solutions across the Catalan Health Service within the past 25 years. Hence, none of the solutions described herein would have been possible without the commitment of many professionals (current and past) who, nevertheless, do not meet the authorship criteria for this work. The authors would like to thank all of them for making this journey possible. All authors are civil servants from the government of Catalonia and affiliated entities.

Authors' Contributions

The Viewpoint was conceptualized by JP-J. PR and RC-S were responsible for data curation and JF-P and RC-S conducted the formal analysis of summary measures reported. The original draft was written by JP-J, GC-S, MR-P, CP, and MMN and subsequently revised and edited by JP-J, GC-S, MR-P, MMN, JF-P, PR, AJ-R, TdPN, JAMP, JCF, JUC, EV, DM, DV-B, TB, GTU, RC-S, PF-F, BGO, JB, XM, IC, RG-C, JR, OS, CP, and PP-S. Figures were prepared by MR-P. All authors have read and agreed to the submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information on the drug prescription system, recommendations for the use of remote consultations, and supplementary tables.
[PDF File (Adobe PDF File), 4698 KB - jmir_v26i1e58933_app1.pdf]

References

1. Shahzad M, Upshur R, Donnelly P, Bharmal A, Wei X, Feng P, et al. A population-based approach to integrated healthcare delivery: a scoping review of clinical care and public health collaboration. BMC Public Health 2019 Jun 07;19(1):708. [doi: 10.1186/s12889-019-7002-z] [Medline: 31174501]
2. Evans JM, Baker RG, Berta W, Barnsley J. The evolution of integrated health care strategies. Adv Health Care Manag 2013;15:125-161. [doi: 10.1108/s1474-8231(2013)0000015011] [Medline: 24749215]
3. Maddux FW, McMurray S, Nissenson AR. Toward population management in an integrated care model. Blood Purif 2013;36(3-4):152-159. [doi: 10.1159/000356095] [Medline: 24496184]
4. Vedel I, Monette M, Beland F, Monette J, Bergman H. Ten years of integrated care: backwards and forwards. The case of the province of Québec, Canada. Int J Integr Care 2011 Jan 21;11 Spec Ed(Special 10th Anniversary Edition):e004. [doi: 10.5334/ijic.574] [Medline: 21677842]
5. Baltaxe E, Cano I, Herranz C, Barberan-Garcia A, Hernandez C, Alonso A, et al. Evaluation of integrated care services in Catalonia: population-based and service-based real-life deployment protocols. BMC Health Serv Res 2019 Jun 11;19(1):370. [doi: 10.1186/s12913-019-4174-2] [Medline: 31185997]
6. Cano I, Dueñas-Espín I, Hernandez C, de Batlle J, Benavent J, Contel JC, et al. Protocol for regional implementation of community-based collaborative management of complex chronic patients. NPJ Prim Care Respir Med 2017 Jul 14;27(1):44. [doi: 10.1038/s41533-017-0043-9] [Medline: 28710482]

XSL•FO
RenderX

7. Dueñas-Espín I, Vela E, Pauws S, Bescos C, Cano I, Cleries M, et al. Proposals for enhanced health risk assessment and stratification in an integrated care scenario. BMJ Open 2016 Apr 15;6(4):e010301. [doi: 10.1136/bmjopen-2015-010301] [Medline: 27084274]
8. Kringos DS, Boerma WG, Hutchinson A, van der Zee J, Groenewegen PP. The breadth of primary care: a systematic literature review of its core dimensions. BMC Health Serv Res 2010 Mar 13;10:65 [FREE Full text] [doi: 10.1186/1472-6963-10-65] [Medline: 20226084]
9. Valentijn PP, Schepman SM, Opheij W, Bruijnzeels MA. Understanding integrated care: a comprehensive conceptual framework based on the integrative functions of primary care. Int J Integr Care 2013 Mar 22;13(1):e010. [doi: 10.5334/ijic.886] [Medline: 23687482]
10. Piera-Jiménez J, Dedeu T, Pagliari C, Trupec T. Strengthening primary health care in Europe with digital solutions. Aten Primaria 2024 Oct;56(10):102904. [doi: 10.1016/j.aprim.2024.102904] [Medline: 38692228]
11. de Stampa M, Vedel I, Bergman H, Novella J, Lechowski L, Ankri J, et al. Opening the black box of clinical collaboration in integrated care models for frail, elderly patients. Gerontologist 2013 Apr;53(2):313-325. [doi: 10.1093/geront/gns081] [Medline: 22961463]
12. Grudniewicz A, Gray CS, Boeckxstaens P, De Maeseneer J, Mold J. Operationalizing the chronic care model with goal-oriented care. Patient 2023 Nov;16(6):569-578. [doi: 10.1007/s40271-023-00645-8] [Medline: 37642918]
13. Fisher ES, Shortell SM. Accountable care organizations: accountable for what, to whom, and how. JAMA 2010 Oct 20;304(15):1715-1716. [doi: 10.1001/jama.2010.1513] [Medline: 20959584]
14. Barnes AJ, Unruh L, Chukmaitov A, van Ginneken E. Accountable care organizations in the USA: types, developments and challenges. Health Policy 2014 Oct;118(1):1-7. [doi: 10.1016/j.healthpol.2014.07.019] [Medline: 25145942]
15. Kaufman BG, Spivack BS, Stearns SC, Song PH, O'Brien EC. Impact of accountable care organizations on utilization, care, and outcomes: a systematic review. Med Care Res Rev 2019 Jun 12;76(3):255-290. [doi: 10.1177/1077558717745916] [Medline: 29231131]
16. Feeley TW, Landman Z, Porter ME. The agenda for the next generation of health care information technology. NEJM Catal 2020 May;1(3):22. [doi: 10.1056/cat.20.0132]
17. Charles A, Wenzel L, Kershaw M, Ham C, Walsh N. A year of integrated care systems: reviewing the journey so far. The King's Fund. 2018. URL: https://www.kingsfund.org.uk/insight-and-analysis/reports/year-integrated-care-systems [accessed 2024-04-29]
18. Medical specialist 2025. Dutch Federation of Medical Specialists. 2017. URL: https://demedischspecialist.nl/sites/default/files/FMS_visiedoc_MS2025%28eng%29_2017_PL_v02%28lr%29.pdf [accessed 2024-08-06]
19. Spatz C, Bricker P, Gabbay R. The patient-centered medical neighborhood: transformation of specialty care. Am J Med Qual 2014 Aug 21;29(4):344-349. [doi: 10.1177/1062860613498812] [Medline: 23966551]
20. O'Dell ML. What is a patient-centered medical home? Mo Med 2016;113(4):301-304. [Medline: 30228482]
21. Global strategy on digital health 2020-2025. World Health Organization (WHO). URL: https://iris.who.int/bitstream/handle/10665/344249/9789240020924-eng.pdf?sequence=1 [accessed 2024-03-20]
22. Health in the 21st century. Putting data to work for stronger health systems. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/health/health-in-the-21st-century-e3b23f8e-en.htm [accessed 2024-03-20]
23. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
24. Hajek AM. Breaking down clinical silos in healthcare. Front Health Serv Manage 2013;29(4):45-50 [FREE Full text] [doi: 10.1097/01974520-201304000-00006]
25. Amelung V, Stein V, Goodwin N, Balicer R, Nolte E, Suter E. Handbook Integrated Care. Cham, Switzerland: Springer; 2017.
26. Grundy EM, Murphy M. Population ageing in Europe. In: Michel JP, Beattie BL, Martin FC, Walston J, editors. Oxford Textbook of Geriatric Medicine. 3rd edition. Oxfordshire, UK: Oxford University Press; 2017:11-18.
27. Basic indicators of Catalonia: population by age group. Idescat. URL: https://www.idescat.cat/indicadors/?id=basics&n=10329&lang=en [accessed 2024-08-02]
28. Bernal-Delgado E, Garcia-Armesto S, Oliva J, Sanchez Martinez FI, Repullo JR, Pena-Longobardo LM, et al. Spain: health system review. Health Syst Transit 2018 May;20(2):1-179. [Medline: 30277216]
29. de Salut D, de Catalunya G. L'estat de salut, els comportaments relacionats amb la salut i l'ús de serveis sanitaris a Catalunya, 2023: resum executiu dels principals resultats de l'ESCA del 2023. Generalitat de Catalunya, Departament de Salut. 2023. URL: https://scientiasalut.gencat.cat/bitstream/handle/11351/11405/enquesta_salut_catalunya_resum_executiu_ca_2023.pdf?sequence=8&isAllowed=y [accessed 2024-08-06]
30. Catalan health plan 2021-2025. Catalan Government. 2020. URL: https://hdl.handle.net/11351/7948 [accessed 2024-08-06]
31. OECD digital economy outlook 2024 (volume 1). Organisation for Economic Co-operation and Development. 2024. URL: https://www.oecd.org/en/publications/2024/05/oecd-digital-economy-outlook-2024-volume-1_d30a04c9.html [accessed 2024-04-29]

XSL•FO
RenderX

32. Vial G. Understanding digital transformation: a review and a research agenda. J Strateg Inf Syst 2019 Jun;28(2):118-144. [doi: 10.1016/j.jsis.2019.01.003]
33. Ismail MH, Khater M, Zaki M. Digital business transformation and strategy: what do we know so far? University of Cambridge. URL: https://cambridgeservicealliance.eng.cam.ac.uk/system/files/documents/2017NovPaper_Mariam.pdf [accessed 2024-03-19]
34. Hanelt A, Bohnsack R, Marz D, Antunes Marante C. A systematic review of the literature on digital transformation: insights and implications for strategy and organizational change. J Manag Stud 2020 Oct 12;58(5):1159-1197. [doi: 10.1111/joms.12639]
35. Information for health: a strategy for building the national health information infrastructure: report and recommendations from the national committee on vital and health statistics. U. S. Department of Health and Human Services. URL: https://aspe.hhs.gov/reports/information-health-strategy-building-national-health-information-infrastructure-0 [accessed 2024-04-29]
36. Ley 16/2003, de 28 de mayo, de cohesión y calidad del Sistema Nacional de Salud. Spanish Government. URL: https://www.boe.es/buscar/act.php?id=BOE-A-2003-10715 [accessed 2024-03-19]
37. Riplinger L, Piera-Jiménez J, Dooling JP. Patient identification techniques - approaches, implications, and findings. Yearb Med Inform 2020 Aug 21;29(1):81-86 [FREE Full text] [doi: 10.1055/s-0040-1701984] [Medline: 32823300]
38. LLEI 21/2000, de 29 de desembre, sobre els drets d'informació concernent la salut i l'autonomia del pacient, i la documentació clínica. Catalan Government. URL: https://portaljuridic.gencat.cat/ca/document-del-pjur/?documentId=246194 [accessed 2024-03-21]
39. Marimon-Suñol S, Rovira-Barberà M, Acedo-Anta M, Nozal-Baldajos MA, Guanyabens-Calvet J. Historia Clínica Compartida en Cataluña. Med Clin 2010 Jan;134:45-48. [doi: 10.1016/s0025-7753(10)70009-9]
40. Eason K, Waterson P. The implications of e-health system delivery strategies for integrated healthcare: lessons from England. Int J Med Inform 2013 May;82(5):e96-106. [doi: 10.1016/j.ijmedinf.2012.11.004] [Medline: 23266062]
41. Auschra C. Barriers to the integration of care in inter-organisational settings: a literature review. Int J Integr Care 2018 Jan 16;18(1):5 [FREE Full text] [doi: 10.5334/ijic.3068] [Medline: 29632455]
42. Temprana-Salvador J, López-García P, Castellví Vives J, de Haro L, Ballesta E, Rojas Abusleme M, et al. DigiPatICS: digital pathology transformation of the Catalan health institute network of 8 hospitals-planification, implementation, and preliminary results. Diagnostics (Basel) 2022 Mar 30;12(4):852 [FREE Full text] [doi: 10.3390/diagnostics12040852] [Medline: 35453900]
43. Propostes per avançar en el model de Medicina Personalitzada de Precisió (MPP) a Catalunya. Consorci de Salut i Social de Catalunya. URL: https://www.consorci.org/media/upload/arxius/publicacions/WEB_informe_MPP_v01_2022_02_17.pdf [accessed 2024-08-06]
44. Brito JP, Kunnerman M, Montori VW. Shared decision making. BMJ Best Practice. URL: https://bestpractice.bmj.com/info/toolkit/practise-ebm/shared-decision [accessed 2024-03-25]
45. Regional digital health action plan for the WHO European region 2023-2030. World Health Organization (WHO). URL: https://apps.who.int/iris/handle/10665/360950 [accessed 2024-03-25]
46. Wong BL, Maaß L, Vodden A, van Kessel R, Sorbello S, Buttigieg S, European Public Health Association (EUPHA) Digital Health Section. The dawn of digital public health in Europe: implications for public health policy and practice. Lancet Reg Health Eur 2022 Mar;14:100316. [doi: 10.1016/j.lanepe.2022.100316] [Medline: 35132399]
47. Garcia-Ribatallada E. La Meva Salut: strategic digital tool to transform the relation between the user and the health system based on a user-centred approach. In: proceedings of the 2021 44th World Hospital Congress. 2021 Presented at: WHC '21; November 8-11, 2021; Barcelona, Spain URL: https://hdl.handle.net/11351/6944
48. Pérez Sust P, Solans O, Fajardo JC, Medina Peralta M, Rodenas P, Gabaldà J, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 04;6(2):e19106 [FREE Full text] [doi: 10.2196/19106] [Medline: 32339998]
49. Fidel Kinori SG, Carot-Sans G, Cuartero A, Valero-Bover D, Roma Monfa R, Garcia E, et al. A web-based app for emotional management during the COVID-19 pandemic: platform development and retrospective analysis of its use throughout two waves of the outbreak in Spain. JMIR Form Res 2022 Mar 31;6(3):e27402 [FREE Full text] [doi: 10.2196/27402] [Medline: 35142638]
50. Real Decreto 1718/2010, de 17 de diciembre, sobre receta médica y órdenes de dispensación. Spanish Government. URL: https://www.boe.es/eli/es/rd/2010/12/17/1718/con [accessed 2024-03-19]
51. Modol JR. Maintaining the pharmacy model: the Catalan electronic prescription infrastructure. In: Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P, editors. Information Infrastructures within European Health Care: Working with the Installed Base. Cham, Switzerland: Springer; 2017:55-72.
52. What is an electronic prescription and how does it work? Catalan Health Service. URL: https://catsalut.gencat.cat/ca/serveis-sanitaris/atencio-farmaceutica/recepta-electronica/que-es-i-com-funciona/index.html#googtrans(ca|en) [accessed 2024-03-19]
53. Gilabert-Perramon A, López-Calahorra P, Escoda-Geli N, Salvadó-Trias C. Receta electrónica en Cataluña (Rec@t): una herramienta de salud. Med Clin 2010 Jan;134:49-55. [doi: 10.1016/S0025-7753(10)70010-5]

XSL•FO
RenderX

54. Atenció primària de l'Institut Català de la Salut. Catalan Institute of Health. 2021 Dec 31. URL: https://ics.gencat.cat/web/.content/gestio-ambiental/declaracion-ambiental-annex-VII-INSTITUT-CATALA-DE-LA-SALUT.pdf [accessed 2024-10-14]
55. Electronic cross-border health services. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/electronic-cross-border-health-services_en [accessed 2024-03-19]
56. Roig Izquierdo M, Prat Casanovas MA, Gorgas Torner MQ, Pontes García C. Registro de pacientes y tratamientos de medicamentos hospitalarios en Cataluña: 10 años de datos clínicos. Med Clin (Barc) 2020 Mar 13;154(5):185-191 [FREE Full text] [doi: 10.1016/j.medcli.2019.09.009] [Medline: 31759696]
57. The Catalan information systems master plan. Ministry of Health (Catalan Government). 2017. URL: https://salutweb.gencat.cat/web/.content/_ambits-actuacio/Linies-dactuacio/tic/pdsis/pdsis-en.pdf [accessed 2024-03-19]
58. Saigí-Rubió F, Vidal-Alaball J, Torrent-Sellens J, Jiménez-Zarco A, López Segui F, Carrasco Hernandez M, et al. Determinants of Catalan public primary care professionals' intention to use digital clinical consultations (eConsulta) in the post-COVID-19 context: optical illusion or permanent transformation? J Med Internet Res 2021 May 31;23(6):e28944 [FREE Full text] [doi: 10.2196/28944] [Medline: 34097638]
59. Solans O, Vidal-Alaball J, Roig Cabo P, Mora N, Coma E, Bonet Simó JM, et al. Characteristics of citizens and their use of teleconsultations in primary care in the catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629 [FREE Full text] [doi: 10.2196/28629] [Medline: 33970867]
60. Piera-Jiménez J, Berdún J, Solans Ò, Forment A, Ramalho L, Baiget J, et al. Facilitating the implementation of remote consultations: development of evidence-informed visual guidelines. Implementation Sci 2021 Dec 20;16(S2):104. [doi: 10.1186/s13012-021-01163-7]
61. Rodríguez-Molinero A, Carot-Sans G, Escrig R, Tebé C, Arce J, Pérez-López C, ECASeT Research Group. Study protocol of a randomized controlled trial to assess safety of teleconsultation compared with face-to-face consultation: the ECASeT study. Trials 2023 Dec 08;24(1):797 [FREE Full text] [doi: 10.1186/s13063-023-07679-1] [Medline: 38066614]
62. López Seguí F, Vidal-Alaball J, Sagarra Castro M, García-Altés A, García Cuyàs F. General practitioners' perceptions of whether teleconsultations reduce the number of face-to-face visits in the catalan public primary care system: retrospective cross-sectional study. J Med Internet Res 2020 Mar 16;22(3):e14478 [FREE Full text] [doi: 10.2196/14478] [Medline: 32175914]
63. Sasseville M, Supper W, Gartner J, Layani G, Amil S, Sheffield P, et al. Clinical integration of digital patient-reported outcome measures in primary health care for chronic disease management: protocol for a systematic review. JMIR Res Protoc 2023 Aug 18;12:e48155 [FREE Full text] [doi: 10.2196/48155] [Medline: 37594780]
64. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient‐reported outcome measures (PROMs): a review of generic and condition‐specific measures and a discussion of trends and issues. Health Expect 2021 May 05;24(4):1015-1024. [doi: 10.1111/hex.13254]
65. Consolo L, Castellini G, Cilluffo S, Basile I, Lusignani M. Electronic patient-reported outcomes (e-PROMs) in palliative cancer care: a scoping review. J Patient Rep Outcomes 2022 Sep 23;6(1):102 [FREE Full text] [doi: 10.1186/s41687-022-00509-z] [Medline: 36138279]
66. Basch E, Schrag D, Henson S, Jansen J, Ginos B, Stover AM, et al. Effect of electronic symptom monitoring on patient-reported outcomes among patients with metastatic cancer: a randomized clinical trial. JAMA 2022 Jun 28;327(24):2413-2422 [FREE Full text] [doi: 10.1001/jama.2022.9265] [Medline: 35661856]
67. Steinbeck V, Langenberger B, Schöner L, Wittich L, Klauser W, Mayer M, et al. Electronic patient-reported outcome monitoring to improve quality of life after joint replacement: secondary analysis of a randomized clinical trial. JAMA Netw Open 2023 Sep 05;6(9):e2331301 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.31301] [Medline: 37656459]
68. Leahy AB, Feudtner C, Basch E. Symptom monitoring in pediatric oncology using patient-reported outcomes: why, how, and where next. Patient 2018 Apr 25;11(2):147-153 [FREE Full text] [doi: 10.1007/s40271-017-0279-z] [Medline: 29071524]
69. Murphy M, Hollinghurst S, Salisbury C. Identification, description and appraisal of generic PROMs for primary care: a systematic review. BMC Fam Pract 2018 Mar 15;19(1):41 [FREE Full text] [doi: 10.1186/s12875-018-0722-9] [Medline: 29544455]
70. Watson L, Qi S, DeIure A, Link C, Chmielewski L, Hildebrand A, et al. Using autoregressive integrated moving average (ARIMA) modelling to forecast symptom complexity in an ambulatory oncology clinic: harnessing predictive analytics and patient-reported outcomes. Int J Environ Res Public Health 2021 Aug 07;18(16):8365 [FREE Full text] [doi: 10.3390/ijerph18168365] [Medline: 34444115]
71. BiblioPRO. Ministry of Science Innovation and Universities (Spanish Government). URL: https://www.bibliopro.org/index.htm [accessed 2024-03-19]
72. Real Decreto 69/2015, de 6 de febrero, por el que se regula el Registro de Actividad de Atención Sanitaria Especializada. Spanish Government. URL: https://www.boe.es/buscar/pdf/2015/BOE-A-2015-1235-consolidado.pdf [accessed 2024-03-25]
73. Vela E, Clèries M, Monterde D, Carot-Sans G, Coca M, Valero-Bover D, et al. Performance of quantitative measures of multimorbidity: a population-based retrospective analysis. BMC Public Health 2021 Oct 18;21(1):1881 [FREE Full text] [doi: 10.1186/s12889-021-11922-2] [Medline: 34663289]

XSL•FO
RenderX

74. Central Results. Health Quality and Assessment Agency of Catalonia. URL: https://aquas.gencat.cat/ca/fem/central-resultats/index.html#googtrans(ca|en) [accessed 2024-08-06]
75. Monterde D, Vela E, Clèries M, García Eroles L, Pérez Sust P. Validez de los grupos de morbilidad ajustados respecto a los clinical risk groups en el ámbito de la atención primaria. Aten Primaria 2019 Mar;51(3):153-161. [doi: 10.1016/J.APRIM.2017.09.012]
76. Monterde D, Vela E, Clèries M, Garcia-Eroles L, Roca J, Pérez-Sust P. Multimorbidity as a predictor of health service utilization in primary care: a registry-based study of the Catalan population. BMC Fam Pract 2020 Feb 17;21(1):39 [FREE Full text] [doi: 10.1186/s12875-020-01104-1] [Medline: 32066377]
77. Santaeugènia SJ, Contel JC, Vela E, Cleries M, Amil P, Melendo-Azuela EM, et al. Characteristics and service utilization by complex chronic and advanced chronic patients in catalonia: a retrospective seven-year cohort-based study of an implemented chronic care program. Int J Environ Res Public Health 2021 Sep 08;18(18):9473 [FREE Full text] [doi: 10.3390/ijerph18189473] [Medline: 34574394]
78. González-Colom R, Carot-Sans G, Vela E, Espallargues M, Hernández C, Jiménez FX, et al. Five years of Hospital at Home adoption in Catalonia: impact, challenges, and proposals for quality assurance. BMC Health Serv Res 2024 Feb 01;24(1):154 [FREE Full text] [doi: 10.1186/s12913-024-10603-1] [Medline: 38297234]
79. Vela E, Carot-Sans G, Clèries M, Monterde D, Acebes X, Comella A, et al. Development and validation of a population-based risk stratification model for severe COVID-19 in the general population. Sci Rep 2022 Feb 28;12(1):3277 [FREE Full text] [doi: 10.1038/s41598-022-07138-y] [Medline: 35228558]
80. Herranz C, Gómez A, Hernández C, González-Colom R, Carles Contel J, Cano I, et al. Toward sustainable adoption of integrated care for prevention of unplanned hospitalizations: a qualitative analysis. Int J Integr Care 2024 Jun 28;24(2):28 [FREE Full text] [doi: 10.5334/ijic.7724] [Medline: 38948163]
81. Baltaxe E, Cano I, Risco R, Sebio R, Dana F, Laxe S, et al. Role of co-creation for large-scale sustainable adoption of digitally supported integrated care: prehabilitation as use case. Int J Integr Care 2022 Oct 07;22(4):1 [FREE Full text] [doi: 10.5334/ijic.6503] [Medline: 36304784]
82. Baltaxe E, Embid C, Aumatell E, Martínez M, Barberan-Garcia A, Kelly J, et al. Integrated care intervention supported by a mobile health tool for patients using noninvasive ventilation at home: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 13;8(4):e16395 [FREE Full text] [doi: 10.2196/16395] [Medline: 32281941]
83. Cano I, Alonso A, Hernandez C, Burgos F, Barberan-Garcia A, Roldan J, et al. An adaptive case management system to support integrated care services: lessons learned from the NEXES project. J Biomed Inform 2015 Jun;55:11-22 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.011] [Medline: 25796455]
84. Vela E, Plaza A, Carot-Sans G, Contel JC, Salvat-Plana M, Fabà M, et al. Data and care integration for post-acute intensive care program of stroke patients: effectiveness assessment using a disease-matched comparator cohort. J Integr Care 2022 Sep 12;30(4):434-455. [doi: 10.1108/jica-12-2021-0062]
85. Jha AK, Doolan D, Grandt D, Scott T, Bates DW. The use of health information technology in seven nations. Int J Med Inform 2008 Dec;77(12):848-854. [doi: 10.1016/j.ijmedinf.2008.06.007] [Medline: 18657471]
86. Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: a more personalized healthcare and public health policy. Health Policy Technol 2017 Mar;6(1):20-25. [doi: 10.1016/j.hlpt.2016.08.002]
87. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
88. Din digitale indgang til sundhedsvæsnet. Sundhed. URL: https://www.sundhed.dk/ [accessed 2024-04-29]
89. Kierkegaard P. eHealth in Denmark: a case study. J Med Syst 2013 Dec;37(6):9991. [doi: 10.1007/s10916-013-9991-y] [Medline: 24166019]
90. Tiik M, Ross P. Patient opportunities in the Estonian Electronic Health Record system. Stud Health Technol Inform 2010;156:171-177. [Medline: 20543352]
91. Kivekäs E, Enlund H, Borycki E, Saranto K. General practitioners' attitudes towards electronic prescribing and the use of the national prescription centre. J Eval Clin Pract 2016 Oct;22(5):816-825. [doi: 10.1111/jep.12548] [Medline: 27121026]
92. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018 Mar;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
93. Vest JR, Unruh MA, Freedman S, Simon K. Health systems' use of enterprise health information exchange vs single electronic health record vendor environments and unplanned readmissions. J Am Med Inform Assoc 2019 Oct 01;26(10):989-998 [FREE Full text] [doi: 10.1093/jamia/ocz116] [Medline: 31348514]
94. Annual report 2022-23. Australian Digital Health Agency. 2023. URL: https://www.digitalhealth.gov.au/sites/default/files/documents/australian-digital-health-agency-annual-report-2022-23.pdf [accessed 2024-08-06]
95. Fragidis LL, Chatzoglou PD. Implementation of a nationwide electronic health record (EHR). Int J Health Care Qual Assur 2018 Mar 12;31(2):116-130. [doi: 10.1108/ijhcqa-09-2016-0136]
96. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]

XSL•FO
RenderX

97. Murphy M, Scott LJ, Salisbury C, Turner A, Scott A, Denholm R, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021 Jan 17;71(704):e166-e177. [doi: 10.3399/bjgp.2020.0948]
98. González-Colom R, Carot-Sans G, Vela E, Espallargues M, Hernández C, Jiménez FX, et al. Five years of Hospital at Home adoption in Catalonia: impact and challenges. BMC Health Serv Res BioMed Central 2024 Feb 01;24(154):154. [doi: 10.1186/s12913-024-10603-1]
99. Piera-Jiménez J, Daugbjerg S, Stafylas P, Meyer I, Müller S, Lewis L, et al. BeyondSilos, a telehealth-enhanced integrated care model in the domiciliary setting for older patients: observational prospective cohort study for effectiveness and cost-effectiveness assessments. JMIR Med Inform 2020 Oct 06;8(10):e20938 [FREE Full text] [doi: 10.2196/20938] [Medline: 33021490]
100. El-Haddad C, Hegazi I, Hu W. Understanding patient expectations of health care: a qualitative study. J Patient Exp 2020 Dec 28;7(6):1724-1731 [FREE Full text] [doi: 10.1177/2374373520921692] [Medline: 33457636]
101. Defining an open platform. Apperta Foundation. URL: https://apperta.org/openplatforms/ [accessed 2024-03-25]
102. Pedrera-Jiménez M, García-Barrio N, Frid S, Moner D, Boscá-Tomás D, Lozano-Rubí R, et al. Can OpenEHR, ISO 13606, and HL7 FHIR work together? An agnostic approach for the selection and application of electronic health record standards to the next-generation health data spaces. J Med Internet Res 2023 Dec 28;25:e48702 [FREE Full text] [doi: 10.2196/48702] [Medline: 38153779]
103. Tsafnat G, Dunscombe R, Gabriel D, Grieve G, Reich C. Converge or collide? Making sense of a plethora of open data standards in health care. J Med Internet Res 2024 Apr 09;26:e55779 [FREE Full text] [doi: 10.2196/55779] [Medline: 38593431]

Abbreviations

AMG: Adjusted Morbidity Group
EHR: electronic health record
EMR: electronic medical record
ePROM: electronic patient-reported outcome measure
FHIR: Fast Healthcare Interoperability Resources
GP: general practitioner
HIT: health information technology
HL7-CDA: Health Level 7 Clinical Document Architecture
ICS: integrated care system
LOINC: Local Observation Identifiers Names and Codes
NHS: National Health System
PROM: patient-reported outcome measure
SIMDCAT: Digital Medical Imaging System of Catalonia
SNOMED-CT: Systemized Nomenclature of Medicine–Clinical Terms

Edited by G Eysenbach; submitted 28.03.24; peer-reviewed by S Liu, P Kokol, L Elvas; comments to author 19.07.24; revised version received 09.08.24; accepted 25.09.24; published 18.11.24.

Please cite as:
Piera-Jiménez J, Carot-Sans G, Ramiro-Pareta M, Nogueras MM, Folguera-Profitós J, Ródenas P, Jiménez-Rueda A, de Pando Navarro T, Mira Palacios JA, Fajardo JC, Ustrell Campillo J, Vela E, Monterde D, Valero-Bover D, Bonet T, Tarrasó-Urios G, Cantenys-Sabà R, Fabregat-Fabregat P, Gómez Oliveros B, Berdún J, Michelena X, Cano I, González-Colom R, Roca J, Solans O, Pontes C, Pérez-Sust P
A 25-Year Retrospective of Health IT Infrastructure Building: The Example of the Catalonia Region
J Med Internet Res 2024;26:e58933
URL: https://www.jmir.org/2024/1/e58933
doi:10.2196/58933
PMID:39556831

©Jordi Piera-Jiménez, Gerard Carot-Sans, Marina Ramiro-Pareta, Maria Mercedes Nogueras, Júlia Folguera-Profitós, Pepi Ródenas, Alba Jiménez-Rueda, Thais de Pando Navarro, Josep Antoni Mira Palacios, Joan Carles Fajardo, Joan Ustrell Campillo, Emili Vela, David Monterde, Damià Valero-Bover, Tara Bonet, Guillermo Tarrasó-Urios, Roser Cantenys-Sabà, Pau Fabregat-Fabregat, Beatriz Gómez Oliveros, Jesús Berdún, Xabier Michelena, Isaac Cano, Rubèn González-Colom, Josep Roca, Oscar Solans, Caridad Pontes, Pol Pérez-Sust. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution

XSL•FO
RenderX

License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Perspectives of Digital Health Innovations in Low- and Middle-Income Health Care Systems From South and Southeast Asia

Siyan Yi[1,2,3], MHSc, MD, PhD; Esabelle Lo Yan Yam[1,4], MSc; Kochukoshy Cheruvettolil[5], MA; Eleni Linos[6], DrPH; Anshika Gupta[7], MPH; Latha Palaniappan[8], MD; Nitya Rajeshuni[9], MD, MS; Kiran Gopal Vaska[7], MPA; Kevin Schulman[10], MD; Karen N Eggleston[11], PhD

[1]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore
[2]KHANA Center for Population Health Research, Phnom Penh, Cambodia
[3]Public Health Program, Touro University California, Vallejo, CA, United States
[4]College of Health and Medicine, Australian National University, Canberra, Australia
[5]Bill & Melinda Gates Foundation, Seattle, WA, United States
[6]Center for Digital Health, School of Medicine, Stanford University, Stanford, CA, United States
[7]National Health Authority, Ministry of Health and Family Welfare, New Delhi, India
[8]Division of Primary Care and Population Health, School of Medicine, Stanford University, Stanford, CA, United States
[9]Department of Pediatrics, School of Medicine, Stanford University, Stanford, CA, United States
[10]Clinical Excellence Research Center, School of Medicine, Stanford University, Stanford, CA, United States
[11]Shorenstein Asia-Pacific Research Center, Freeman Spogli Institute for International Studies, Stanford University, Stanford, CA, United States

Corresponding Author:
Siyan Yi, MHSc, MD, PhD
Saw Swee Hock School of Public Health
National University of Singapore and National University Health System
#10-01, 12 Science Drive 2
Singapore, 117549
Singapore
Phone: 65 6516 6914
Fax: 65 6779 1489
Email: siyan@nus.edu.sg

Abstract

Digital health innovations have emerged globally as a transformative force for addressing health system challenges, particularly in resource-constrained settings. The COVID-19 pandemic underscored the critical importance of these innovations for enhancing public health. In South and Southeast Asia, a region known for its cultural diversity and complex health care landscape, digital health innovations present a dynamic interplay of challenges and opportunities. We advocate for ongoing research built into system development and an evidence-based strategy focusing on designing and scaling national digital health infrastructures combined with a vibrant ecosystem or “marketplace” of local experiments generating shared experience about what works in which settings. As the global digital health revolution unfolds, the perspectives drawn from South and Southeast Asia—including the importance of local partnerships—may provide valuable insights for shaping future strategies and informing similar initiatives in low- and middle-income countries, contributing to effective digital health strategies across diverse global health contexts.

(J Med Internet Res 2024;26:e57612) doi:10.2196/57612

KEYWORDS

digital health innovations; public health; South and Southeast Asia; health care challenges; low- and middle-income countries; LMICs; global health; health AI; artificial intelligence; public health responses; global health contexts; digital health

XSL•FO
RenderX

Introduction

The rapid development of digital health—encompassing technologies like mobile health (mHealth), telemedicine, digital health information systems, Internet of Things, machine learning and artificial intelligence (AI), personalized digital nudges, large language models, and generative AI—has ushered in a new era of health care possibilities [1]. Digital health technology has been proposed as an effective solution to address health care access, cost, and quality issues. While the promise of digital health is universal, the context in which it is applied varies across communities and regions. This viewpoint paper critically examines the diverse challenges and opportunities encountered in digital health innovations in low- and middle-income countries (LMICs), drawing on perspectives from the experiences of health systems in South and Southeast Asia, a vibrant and multifarious region.

Just as many have highlighted the advantages of a "human-in-the-loop" approach to AI, we emphasize the potential benefits of a "researcher-in-the-loop" approach to digital health innovations in LMICs. This approach aims to ensure objective evaluation, continuous improvement, and effective dissemination of successes and failures, ultimately enhancing patient outcomes while reducing disparities, costs, and the burden and burnout of health professionals. A trilateral team comprising policy makers, researchers, and development partners can expedite the identification and dissemination of best practices across different LMIC settings. Additionally, partnering with trusted global organizations or catalytic funders—such as the Bill & Melinda Gates Foundation (BMGF), National Institutes of Health Fogarty International Center, Abdul Latif Jameel Poverty Action Lab, and the International Initiative for Impact Evaluation—can provide essential resources to low-resource settings, mitigate risks during the initial adoption phase, and support adapting interventions to the local context before scaling up to larger populations. This collaboration also helps build a regional or global "toolkit" of use cases and experiences that can benefit other LMICs [2].

The following sections will first articulate our analytic framework and examine the foundational elements of digital health innovations, focusing on their architecture, standards, and issues related to privacy and security. Next, we will explore the digital health solutions marketplace in South and Southeast Asia, emphasizing the importance of local community partnerships and the need for a continuous cycle of learning and improvement through impact evaluation, research, and development. We emphasize the need for cultural sensitivity and prioritizing equity to ensure that digital health solutions are effectively adapted to local contexts.

Improving the Social Net Benefit From Digital Health Innovation: Analytic Framework

Our analytic framework considers balancing the social benefits of digital health innovations with the social costs of adoption, recognizing (1) the positive spillovers from evidence generation and (2) the opportunity costs of not adopting innovations that might otherwise have improved health and equity at affordable cost. Critically, the social benefits should consider the incremental global public good of collecting evidence and experience about what works in diverse local settings. If these spillover benefits for other LMICs are ignored, resource constraints will lead to underinvestment in digital health innovation for low-resource communities. To leverage additional resources to align with social benefit (to internalize this externality, in economics jargon), partners can support local experimentation and evidence generation. Regional or global partners can help leverage existing technology platforms and standards to generate economies of scale and global public goods. Cutting-edge applications can be tailored and redeployed in multiple low-resource settings, allowing LMICs to decrease implementation costs, foster the development of reusable tools and methodologies, and nurture culturally relevant communities of practice.

To balance social benefit and cost while taking account of these spillovers, ongoing research should be built into system development. We advocate for a 2-pronged top-down and bottom-up innovation approach: an evidence-based strategy focusing on designing and scaling national digital health infrastructures combined with a vibrant ecosystem or "marketplace" of local experiments generating shared experience. Importantly, evidence about failures and successes is essential to reduce the costs of deploying ineffective strategies.

This dual approach of top-down design and bottom-up experimentation, enabled through careful evaluation of effectiveness and impact, allows economies with different starting conditions to seize opportunities to "leapfrog" toward more robust, resilient health systems fitting their contexts rather than imitate the development path of any given current high-income country or region. This theoretical framework emphasizes that the incremental social net benefit of digital innovations depends critically upon the "starting point" and desired "end point"—both of which rely on local context and require local partner input to accurately assess benefits and costs.

For example, the starting point for any "leapfrog" in technology may differ in light of significant technological disparities—"the digital divide." Despite the rapidly advancing technological landscape in LMICs, substantial disparities in digital access and digital literacy persist. Across South and Southeast Asia, less than half of the population in Myanmar uses the internet, and only 2% of people in Cambodia and the Lao People's Democratic Republic have a fixed broadband subscription compared to an average of 4% in LMICs [3]. Cambodia has low digital literacy [4], whereas in countries like Singapore, digital literacy and technology use are relatively high, even among older adults [5]. Digital literacy can impact one's ability and attitude toward using digital technology [5,6]. The scarcity of technological resources in LMICs can also hamper the introduction, adoption, and integration of digital health interventions into the broader health system and compound existing inequities in access to health care, potentially exacerbating disparities in health outcomes [7,8].

Yet, we emphasize that the incremental benefits of digital technologies could be significant—even transformational—in

XSL•FO
RenderX

LMICs. By providing access to quality care in remote settings, the incremental benefits could be considerably higher than in high-income health systems if sufficient oversight and stewardship uphold the tenet "First, do no harm." The ethical complications of denying access to the social value of digital health innovations cannot be ignored.

Local partnerships can reduce the social costs of developing appropriate digital technologies and enhance their social benefits by creating trust for engagement and evidence aligned with local community goals. To emphasize the net benefit of understanding existing traditions or policies before attempting to change them, we use the concept of "Chesterton's Fence" based on the writer Chesterton [9]: "If you come across a fence blocking a road and do not know why it was built, you should not remove it until you understand its purpose." Digital health innovations should not blindly "leapfrog" Chesterton's Fence: reforms or changes should not be made until the reasons for the current situation are fully understood. Incorporating historical knowledge and avoiding hasty or uninformed decision-making can enhance the net social value of changes.

For balancing the social benefits and costs of digital health innovation, Chesterton's Fence suggests the importance of actively involving local communities, respecting local decision-making processes, and partnering with local governments, nongovernmental organizations (NGOs), and other stakeholders to tailor proposed solutions to the cultural, economic, and environmental context. Implementing small-scale pilot projects can provide valuable insights, allowing for adjustments before broader application, considering both the direct and indirect consequences for the community. Moreover, the social benefits and costs should be measured over a forward-looking time frame, not limited to short-term results. Sustainable change often requires investments in local skills and institutions to ensure long-term effectiveness. Patience and persistence are vital in these efforts to harness the potential of digital technology to improve the health of disadvantaged individuals and communities. Ongoing evaluation and willingness to adapt strategies based on feedback are critical for maximizing the social net benefit of digital health innovations.

Data Architecture, Standards, Privacy, and Security

For digital health technology to have a long-lasting impact on health systems, technology must be developed around an extensible architecture that can be efficiently scaled for new use cases and apps. While we are at the very early stages of technology adoption, it is essential to understand how critical these foundational concepts are to a proper digital solution for health care. An essential aspect of this adaptable architecture is incorporating data governance to facilitate the development of use cases and technology deployment.

Electronic medical records have been used to streamline patient data management and improve health care service delivery, exemplifying digital health's potential to improve health care efficiency [10]. The Indonesian and Singaporean governments are making electronic medical records accessible to the population [11,12], while an electronic information registry for routine immunization is being introduced in Cambodia and the Lao People's Democratic Republic. However, with specific data standards for organization and sharing, data elements may be digitally exchanged between clinical systems.

The process of exchanging digital information is known as interoperability. Most high-income countries have either implemented or are in the process of introducing legislation mandating interoperability [13]. However, many legacy health information technology solutions are designed primarily for providers, hospitals, or clinics, which supports organizational goals but presents challenges for data sharing and exchange as patients move across different providers. In contrast, only some LMICs have achieved comparable levels of interoperability. Nevertheless, LMICs may have the most promising opportunities to "leapfrog" older standards by directly adopting a patient-centered data architecture.

Rather than developing a data architecture that mirrors the health ministry organizational chart, policy makers should consider technology solutions centered on patients, such as personal health records. In Singapore, the government has launched the Next Generation Electronic Medical Record system to consolidate data and harmonize processes across health establishments, including private care providers, providing a longitudinal view of patient's health data [14]. Such efforts to integrate digital health solutions into existing health care frameworks hold promise in optimizing resource allocation and health care delivery. Mobile phones have greatly facilitated personal health records but could include cloud-based models or an innovative blockchain model such as that used in Estonia [15]. Personal health records could have distinct advantages for patients, such as having patient records available when making clinical decisions.

Launched in 2021, India's national Ayushman Bharat Digital Mission (ABDM) was built based on a decade of digitalization experiences in the national identity system, personal digital records, and payment systems (Figure 1[16-19]). ABDM uses the Ayushman Bharat Health Account as a digital health identity number that can be shared freely while keeping the underlying information secure. Over 268,000 health professionals have been included in the national registry, irrespective of the medical systems they practice. More than 229,000 health facilities have been registered with the facility registry [20].

A key lesson from India's experiences with ABDM is building an integrated data ecosystem instead of disparate information systems. The integration will be crucial to support personalized health care and evidence-based policy making. The overall data architecture at a country or market level should also anticipate the interest in developing these tools and their challenges. Individual tools should anticipate updates to the overall data architecture and regulation at the country and market level.

Apart from data architecture for storage and exchange, data standards vary across countries, leading to incomparability [21]. The same concept (eg, blood glucose) may be represented in various ways from one setting to the next. The World Health Organization has developed a data architecture to support countries' capacity to collect, manage, analyze, and use health data from population-based sources such as household surveys and institutional sources such as health facilities [22]. Similarly,

XSL•FO
RenderX

Digital Health Europe attempts to develop comprehensive data standards, including international patient summaries that provide primary medical data for supporting cross-border care [23]. The European case of regional harmonization is a model for regional organizations such as the Association of Southeast Asian Nations, a political and economic union of 10 Southeast Asian states, to work toward fostering data exchange and improving data comparability between countries.

Finally, moving from paper to digital records raises the risk of data breaches and unauthorized access to sensitive health data, which poses ethical and legal dilemmas. Some countries have tightened regulations to increase protection. Singapore's Personal Data Protection Act is a regulatory model for safeguarding personal health data [24]. In Thailand, Indonesia, and Vietnam, data protection was embedded in several laws before its personal protection law was passed in 2019, 2022, and 2023, respectively [25-27]. India has adopted the Digital Personal Data Protection Act, which empowers individuals and specifies requirements for any data fiduciary involved in handling personal data, as well as for the National Health Authority, which serves as a consent manager rather than a data repository [28]. India's ABDM also includes a consent artifact that empowers users to decide how, when, where, and what record types they allow to be shared between one health provider and another (Figure 2).

However, inadequate data protection regulations in some countries risk the security and privacy of health data [29], which could undermine individuals' willingness to engage with digital health platforms. An example is the apprehension surrounding Indonesia's new Health Bill, which allows for collecting and using health data, including genomic data, outside the country [30]. Clear guidelines for data collection, storage, and sharing, along with measures to safeguard sensitive health information, are critical.

These architectural elements are designed for 3 primary functions: interoperability of health records, interoperability of services through the unified health interface for all providers, and interoperability of health claims so that hospitals or other providers are manageable when dealing with multiple payers to submit claims. Together, these layers create a citizen-centric view of public health by building the infrastructure underpinning a vibrant ecosystem. Featuring 3 main exchanges—the Unified Health Interface, the Health Information Exchange consent manager, and the Health Claims Exchange—this system has succeeded in integrating more than 1280 different public and private apps and laying the foundation of digital health public infrastructure for ongoing public and private innovation (Figure 2). India continues to explore strategies to drive further adoption and social value.

Figure 1. India's digitalization policies, strategies, and experiences [16-19].

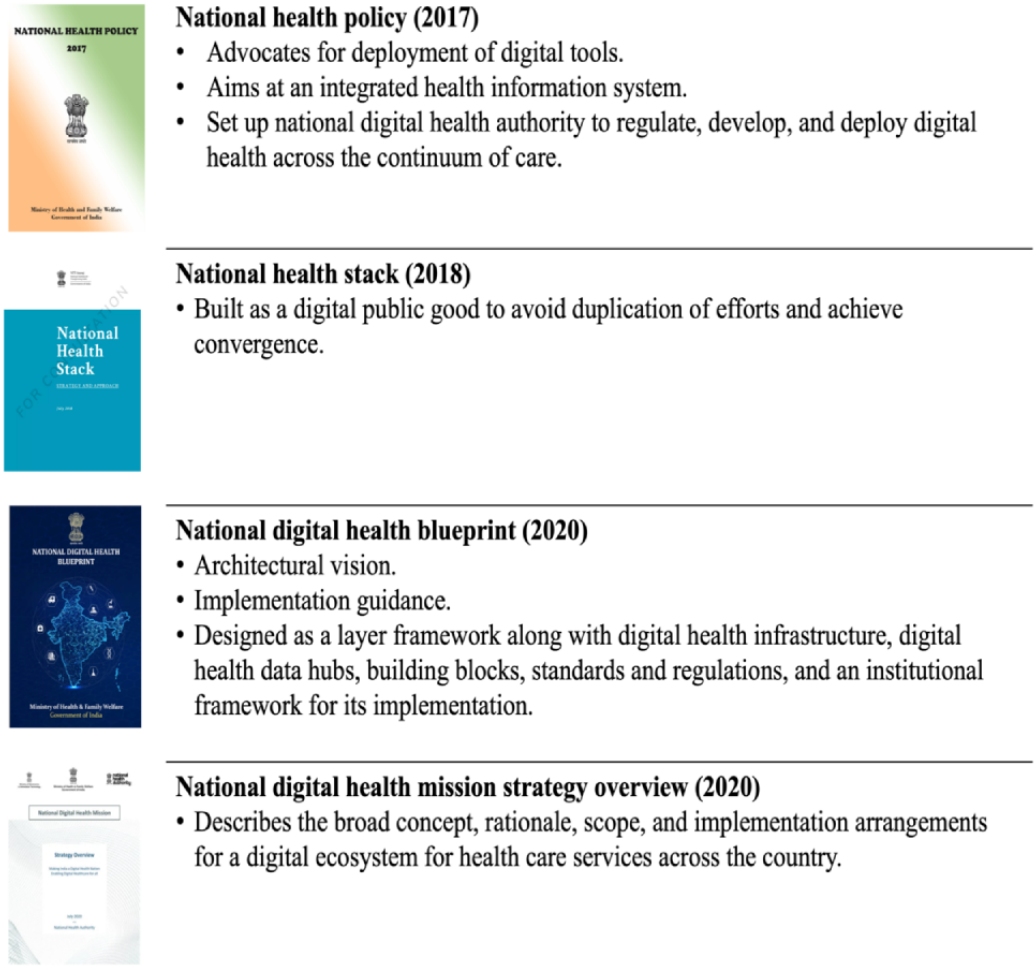

XSL•FO
RenderX

Figure 2. India's ABDM data architectural elements and underpinning services and features. ABDM: Ayushman Bharat Digital Mission; ABHA: Ayushman Bharat Health Account; API: application programming interface; UPI: Unified Payments Interface.

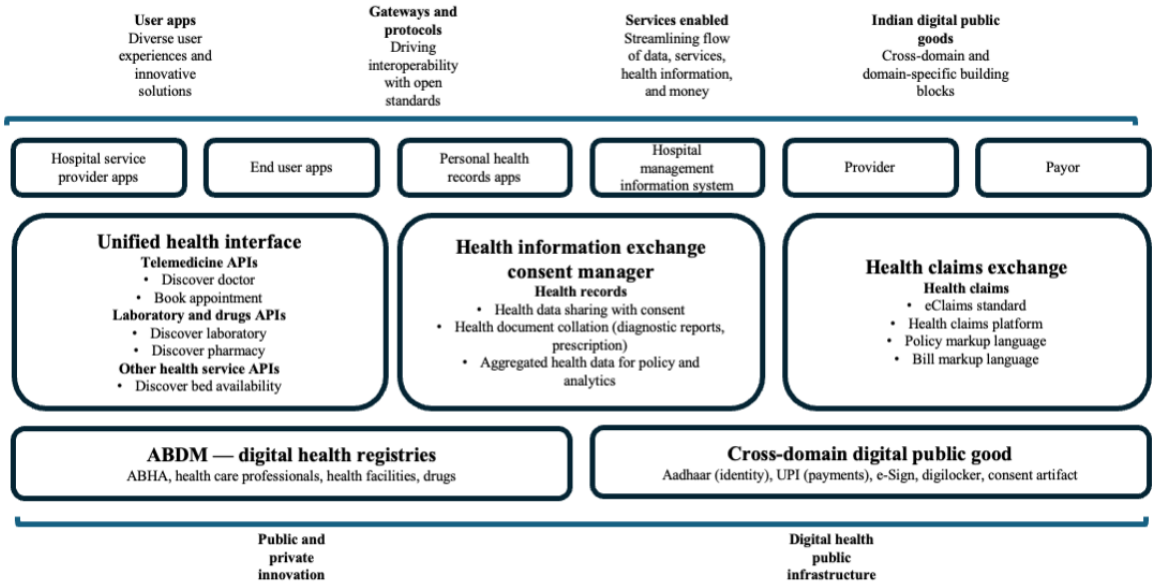

Marketplace—Digital Health Solutions

The genuine excitement surrounding digital health care solutions stems not from data architecture but from the services they enable. Data architectures are not a final use case for a technology; they are merely an enabler of a broader digital health care services ecosystem. Digital services use an underlying data architecture to provide patient information, tools, reports, and recommendations. Digital services can be built on a standard data architecture or stand-alone solutions. The former requires a means to access and potentially update or add to the data, while the latter runs the risk of creating noninteroperable data silos. We advocate for adopting both approaches while systematically assessing the latter and designing in anticipation of interoperable solutions in the near future. Digital solutions must be tested to ensure they perform as expected, including concepts such as the user interface; data access, storage, manipulation, and reporting; and clinical validity.

A vibrant ecosystem of apps shown to be effective in different settings, in turn, lays the foundation for a "marketing strategy" to disseminate and encourage the take-up of digital health solutions such as digital health records among individuals and organizations. Building the personal health report system means little if patients and providers do not engage. To drive meaningful adoption and use, the digital health ecosystem must create value for every player or stakeholder, fostering trust and spurring curiosity about the possibilities for more manageable, cheaper, higher quality, more accessible, and more resilient services. The patients and providers need evidence from use cases that improve their lives, transforming inertia and fear of new technology into excitement and fear of missing out. Digital solutions must also easily integrate into existing workflows and systems, reducing the barriers to uptake.

For example, the Indian government started by ensuring that all the government programs, such as ABDM-enabled health and laboratory management information systems for government hospitals, are ABDM-enabled. Such deployment creates the demonstration space to study integration into existing workflows and make these low-cost "plug and play" solutions available to all participants so that even the smallest local provider need not invest much time and money to adopt them.

While the goal is to avoid creating siloed platforms—an HIV platform, a TB platform, and so on—as much as possible, health systems may benefit from a vibrant ecosystem or "marketplace" of local apps and experiments generating shared experience about what works in diverse settings to enhance access while reducing disparities. One noteworthy example from Southeast Asia comes from a recent randomized controlled trial that demonstrated the benefits of digital health interventions for accessing health care and support services among female entertainment workers, a social and gender-marginalized population in Cambodia [31]. The study used a mobile intervention to connect the vulnerable population to critical health and gender-based violence services, showcasing the transformative potential of digital health in addressing pressing societal challenges even without a national-level integrated digital health ecosystem.

The potential of digital health to bridge geographical gaps, extend health care services to underserved populations, and support the health system in moving care from hospitals to communities is promising. In archipelagic countries like Indonesia and the Philippines, as well as in countries such as Cambodia, where 75% of the population lives in rural areas [3], digital health interventions such as teleconsultation can be used to improve health care accessibility. Digital health additionally holds potential in countries reliant on community health workers to bridge geographic gaps. For example, Accredited Social Health Activist workers in India, who are employed by the Ministry of Health and Family Welfare through India's National Rural Health Mission, have successfully leveraged mHealth for disease surveillance and treatment in a country where low-cost smartphones are increasingly prevalent [32,33].

XSL•FO
RenderX

A mobile phone can be considered a valid medical device equipped with multiple sensors: a camera, a microphone, a screen, a global positioning system, and data. Organizations like BMGF are working with local partners to develop apps leveraging the ubiquity of mobile phones to enhance care in low-income settings, such as scanning to assess a newborn's anthropometrics and monitor malnourishment. mHealth has been proven effective to varying degrees in health education, disease screening, and linkages to care and treatment services [34-38], while wearable devices can increase physical activity and improve health [39], and Internet of Things–enabled systems can support community eldercare [40]. An app for counseling women about contraceptive methods showed that for the informed adoption of unfamiliar technologies, personalized information could be just as effective and far cheaper than subsidies [41]. The power of social media as a platform for health education also holds considerable promise both for large-scale public health campaigns and specifically tailored interventions that are culturally sensitive and build on trusted community figures or institutions [42].

AI has demonstrated significant potential in supporting precision health, including the use of large language models and generative AI for diagnosing complex medical conditions [43]. For example, integrating a deep learning algorithm into Thailand's national diabetic retinopathy screening program has shown that this system can provide real-time detection of diabetic retinopathy with accuracy comparable to that of retina specialists in community-based settings [44]. AI-enabled solutions for precision public health extend beyond diagnostics to clinical decision support and health care delivery, particularly in low-resource settings where staffing, training, funding, and equipment are often limited. While AI solutions may rely on effective collaboration with academic and industry partners, NGOs can also play a crucial role in data acquisition and technology adoption, as detailed in the following section [45].

Ultimately, digital health solutions aim to augment, not replace, the services provided by frontline health workers—we need to avoid the "Turing Trap" [46] and focus on innovations that augment the social value created by the existing health workforce. In the most recent and rigorous studies, generative AI and related technologies tend to boost the productivity of new and less-experienced health workers more than their more-experienced counterparts [47]. The potential for AI and other digital health assistants to upskill and extend the limited health professional resources of LMICs underscores the social value of experimenting and assessing real-time effectiveness without waiting for the design of national architecture where it is lacking.

Digital health innovation also needs to consider the financing levers in both the public and private sectors to understand the feasibility of adopting and scaling up digital health interventions. A private philanthropy like BMGF can provide catalytic funding to derisk the technology early and gather evidence about what works in practice before turning over implementation at scale to governments and civil society. For example, BMGF successfully helped deploy a platform developed in India in 2 other countries, Cambodia and Nigeria [48,49]. Using technologies to support health system strengthening through health financing also holds promise for increasing health care efficiency, accountability, and transparency [50]. Digital technologies have been used to collect health insurance premiums and taxes through mobile wallet services, support pay-for-performance strategic purchasing models, and raise awareness of benefit packages and cost schedules [50].

Local Partnerships for Successful Digital Health Implementation

The architecture on which digital technologies are built may be universal, but the services they provide must be localized to address local needs and be sufficiently engaging to promote adoption. One example of the importance of tailoring digital health innovations to the local context comes from maternal and child health in India [51]. Many Indians, especially in rural areas, are not born in their parents' houses or even their parents' towns but rather in their maternal grandparents' houses or communities. Typically, in the third trimester of pregnancy, the pregnant woman will move to her parents' house, and the baby will be born there. Since the third trimester is a critical time for the mother and child, customized software supported by the BMGF alerts community workers to gather 2 addresses for each pregnant woman and then sends an alert to the maternal grandparents' community in the third trimester. This allows for follow-up care and management continuity, ensuring that health care workers in each jurisdiction are aware of incoming cases, thereby reducing the number of pregnant women lost to follow-up at this critical time.

Local champions and partnerships are critical in providing cultural nuances, context, and advocacy, and in some instances, in creating a network of users and customers for realizing the full social benefit of digital technologies. NGO Arogya World led one example of local champions and partnerships enabling successful digital health implementation in India. Arogya World has developed and implemented several diabetes prevention programs that have successfully incorporated digital tools. Their program mDiabetes, a text messaging intervention delivering diabetes education, significantly improved health behaviors and nutrition [52]. The Healthy Schools Program, a school-based intervention targeting 11- to 14-year-olds, incorporated digital curricula in light of the COVID-19 pandemic. Healthy Schools Program achieved a broader reach through digital campaigns, yielding notable gains in knowledge about diabetes, its risk factors, and behaviors spanning nutritious food consumption and physical activity.

Another example comes from the NGO Noora Health, which has successfully leveraged digital health across India, Bangladesh, and Indonesia to improve access to caregiver training using multimedia. Successful features of their programs include mobile chat services and digital curricula integrated into comprehensive care delivery models that use local health care systems. Through meaningful technology adoption and partnerships with local governments, policy makers, and community health workers, NGOs like Arogya World and Noora Health are well-poised to launch large-scale public health campaigns that cater to increasingly technology-savvy populations.

XSL•FO
RenderX

Local partnerships are crucial in successfully implementing community-based digital health interventions, as evidenced by the Mobile Link and i-MoMCARE projects in Cambodia. Led by the Khmer HIV/AIDS NGO Alliance, a leading local NGO, the Mobile Link used a participatory approach to involve representatives of the female entertainment worker community, civil society organizations, and policy makers in formative studies and intervention development, implementation, and evaluation [31]. This partnership with local stakeholders has significantly increased access to services and improved health outcomes in the target population. Similarly, i-MoMCARE leverages primary health care providers, community health workers (known as village health support groups in Cambodia), a local university, and the national maternal and child health program to provide maternal and child health education and reminders via a mobile app [49]. This enables real-time data collection and personalized care plans, ensuring timely antenatal, delivery, and postnatal care for mothers and children in rural Cambodia. These partnerships facilitate trust, cultural relevance, and sustainability, ultimately enhancing the effectiveness and reach of digital health technologies. Additionally, the involvement of trusted and reliable local sources of information improves the credibility and acceptance of these digital health interventions. A rigorous impact evaluation is part of the design for i-MoMCARE so that this experience can contribute evidence sound for other low-resource settings globally.

Local champions and advocates provide the social legitimacy necessary for the adoption and effective use of digital technologies for health and well-being. This is particularly important in the realm of AI, where apps still need to be thoroughly tested at the population level and where security breaches have undermined public confidence. In some contexts, these technologies could even be seen as tools for surveillance and propaganda. Social legitimacy thus seeds the trust required for the adoption and effective use of the technology. For example, a large-scale randomized controlled trial in West Bengal, India, found that SMS text messages from West Bengal native and 2019 Nobel laureate Abhijit Banerjee doubled the reporting of health symptoms to community health workers and had other positive spillover impacts on their communities [53]. This example illustrates the power of communication technology and potentially social media as platforms for health education and public health information for large-scale campaigns and specifically tailored interventions that build on trusted community figures or institutions.

Digital Health Innovations—A Cycle of Learning and Improvement

The rapid proliferation of digital health technologies requires rigorous research and impact evaluation to generate the evidence to assist LMICs in identifying the most effective strategies. Mobile phones and their voice and AI apps provide many opportunities to enhance individual health knowledge and deliver helpful information. They also call for careful ongoing research to address inappropriate advice or information concerns. For example, standard smartphone voice systems (eg, Siri or S Voice) have failed to recognize issues about mental health, interpersonal violence, or physical health [54]. The accuracy of AI-generated information about diabetes is also questionable [55]. Many ethical issues need to be addressed when adopting monitoring devices or contactless sensors and contact-based wearable devices installed in health care settings, otherwise called “ambient intelligence,” to collect data in health and elderly care [56]. AI or machine learning also relies on models that require significant amounts of data for development and attention to the potential errors introduced by data that may be flawed or biased [57]. Validation of the models across health care markets and populations is essential [58].

Digital health research requires an interdisciplinary team of researchers in public health, data science, information technology, social sciences, and engineering. Researchers must work with policy makers and end users such as health care professionals, communities, and individuals in designing interventions that are problem- or need-driven and set in the broader policy environment. The principle of equity must guide the innovation cycle to prevent inadvertently deepening health disparities through underdiagnosis and undertreatment of individuals historically with limited access to care or who face biases due to language, race, or education. For example, using medical spending to predict medical needs can result in significant racial bias [59]. Strategies to reach marginalized populations, including migrants and underserved communities, are imperative. Initiatives such as using mobile phones to deliver health information to key HIV population communities in Cambodia reflect a commitment to advancing health equity [31,60].

In addition, ensuring that participants comprehensively understand the implications of digital health research, especially in linguistically and culturally diverse contexts, is paramount. Countries in the region have taken steps to develop culturally sensitive informed consent processes to uphold ethical standards and respect individuals’ autonomy [30]. The digitization experience of the banking and financial industry in the region can also offer valuable lessons for digital health research and interventions. The banking sector has learned that retaining a human touch in the technological interface with customers is essential [34]. In some countries where public health services have been underused due to the lack of trust in the system, the introduction of digital health interventions must be carefully designed to instill confidence in the intent, quality, and integrity of the technology [35]. Retaining a human element in digital health interventions could be one way to instill confidence.

The key to success will be alacrity in the stewardship of health sectors to address priority population health needs and improve equity. Policies should support the generation of evidence assessing digital health apps so that patients, providers, and policy makers can ask and answer the right questions in a suitable time frame to enable a virtuous cycle of learning and improvement. In LMICs, where governments lack resources to try many new things, private philanthropy can bring resources, take the risk, and generate evidence about “what works” and scalability.

While some may perceive the term “leapfrogging” as imposing a development paradigm that prioritizes Western values and homogenizes diverse cultures, we suggest that each culture and economy may choose its leapfrog developmental path tailored

XSL•FO
RenderX

to its unique circumstances. Although space limitations preclude a full exploration of alternative development models within this brief perspective piece, future research should prioritize an in-depth discussion of related theories and evidence. This includes conducting cost-benefit analyses of heterogeneous digital health interventions across LMIC contexts.

Conclusions

Digital health programs and research in South and Southeast Asia illustrate that innovation can deliver a social net benefit to communities while generating positive spillovers for other LMICs based on a comprehensive understanding of technological, cultural, and ethical dimensions of the intervention and community context. While technological challenges persist, opportunities for supplementing traditional health care, strengthening health systems, and enhancing epidemic preparedness hold promise. At the current unprecedented time in humanity's digital transformation, these discussions are topical and urgent. Ethical considerations underscore the importance of equitable access, cultural sensitivity, robust data governance, and partnering with local communities. Collaborative efforts among researchers, policy makers, health care professionals, and civic organizations or other community representatives within and outside the health sector will be instrumental in harnessing digital health's potential to improve health outcomes and promote health equity in LMICs.

Acknowledgments

The authors express their gratitude to the staff members, particularly Ms Lisa Lee, at the Asia Health Program, Shorenstein Asia-Pacific Research Center, Freeman Spogli Institute for International Studies, Stanford University, for their outstanding contribution to organizing a seminar series on Asia-Pacific Digital Health Innovation, during which the collaboration for this review work was conceived.

Conflicts of Interest

None declared.

References

1. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-10-26]
2. Asia-Pacific Digital Health Innovation: technology, trust, and the role of APEC. Standard APARC. URL: https://aparc.fsi.stanford.edu/events/asia-pacific-digital-health-innovation-technology-trust-and-role-apec [accessed 2023-10-18]
3. World Bank open data. World Bank. URL: https://data.worldbank.org [accessed 2024-10-19]
4. Kusumastuti A, Nuryani AF. Digital literacy levels in ASEAN (comparative study on ASEAN countries). 2020 Presented at: Proceedings of the 13th International Interdisciplinary Studies Seminar; October 30-31, 2019; Malang, Indonesia. [doi: 10.4108/eai.23-10-2019.2293047]
5. Digital literacy among older adults in Singapore. Singapore Management University. 2023. URL: https://rosa.smu.edu.sg/sites/rosa.smu.edu.sg/files/Briefs/May%20Jun23/Digital%20Literacy_Final2.pdf [accessed 2024-10-26]
6. Tegegne MD, Tilahun B, Mamuye A, Kerie H, Nurhussien F, Zemen E, et al. Digital literacy level and associated factors among health professionals in a referral and teaching hospital: an implication for future digital health systems implementation. Front Public Health 2023;11:1130894 [FREE Full text] [doi: 10.3389/fpubh.2023.1130894] [Medline: 37113180]
7. Yu J, Meng S. Impacts of the internet on health inequality and healthcare access: a cross-country study. Front Public Health 2022;10:935608 [FREE Full text] [doi: 10.3389/fpubh.2022.935608] [Medline: 35757602]
8. Makri A. Bridging the digital divide in health care. Lancet Digit Health 2019;1(5):e204-e205. [doi: 10.1016/s2589-7500(19)30111-6]
9. Chesterton GK. The Thing: Why I Am a Catholic (Illustrated). North America, Europe: Aeterna Press; 1929.
10. Dornan L, Pinyopornpanish K, Jirapornchaoren W, Hashmi A, Dejkriengkraikul N, Angkurawaranon C. Utilisation of electronic health records for public health in Asia: a review of success factors and potential challenges. Biomed Res Int 2019;2019:7341841 [FREE Full text] [doi: 10.1155/2019/7341841] [Medline: 31360723]
11. About National Electronic Health Record (NEHR). Synapxe. URL: https://www.synapxe.sg/healthtech/national-programmes/national-electronic-health-record-nehr [accessed 2024-10-19]
12. Novianti. Implementation of Electronic Medical Record System in Indonesia Viewed from the Perspective of Legal Certainty. ijebss 2024 Mar 22;2(04):1114-1122. [doi: 10.58451/ijebss.v2i04.145]
13. Health at a Glance: OECD Indicators. Paris: OECD Publishing; 2023.
14. NHG Corporate Yearbook Financial Year 2021/22. National Healthcare Group. 2022. URL: https://corp.nhg.com.sg/Pages/Digital%20Version%202022-2023/index.html [accessed 2024-10-26]
15. Heston TF. A case study in blockchain healthcare innovation. Authorea Working Paper No AUTHOREA_213011_364. 2017. URL: https://ssrn.com/abstract=3077455 [accessed 2023-12-06]
16. National Health Policy 2017. Ministry of Health and Family Welfare, Government of India.: New Delhi: National Portal of India; 2017. URL: https://mohfw.gov.in/sites/default/files/9147562941489753121.pdf [accessed 2024-11-19]

XSL•FO
RenderX

17. National Health Stack: Strategy and Approach. NITI Aayog. New Delhi: Ayushman Bharat Digital Mission. 2018. URL: https://abdm.gov.in:8081/uploads/NHS_Strategy_and_Approach_1_89e2dd8f87.pdf [accessed 2024-11-19]
18. National Digital Health Blueprint. Ministry of Health and Family Welfare, Government of India, New Delhi: Ayushman Bharat Digital Mission;. 2017. URL: https://abdm.gov.in:8081/uploads/ndhb_1_56ec695bc8.pdf [accessed 2024-11-19]
19. National Digital Health Mission. National Health Authority, Government of India, New Delhi: NITI Aayog. 2020. URL: https://www.niti.gov.in/sites/default/files/2023-02/ndhm_strategy_overview.pdf [accessed 2024-11-19]
20. ABDM dashboard India. National Health Authority, Ministry of Health and Family Welfare, Government of India. URL: https://dashboard.abdm.gov.in/abdm/ [accessed 2024-10-19]
21. Liverani M, Teng S, Le MS, Coker R. Sharing public health data and information across borders: lessons from Southeast Asia. Global Health 2018;14(1):94 [FREE Full text] [doi: 10.1186/s12992-018-0415-0] [Medline: 30268139]
22. Data collection and analysis tools. World Health Organization. 1950. URL: https://www.who.int/data/data-collection-tools [accessed 2024-10-26]
23. Health data standards. The European Institute for Innovation through Health Data (i~HD). URL: https://www.i-hd.eu/health-standards/what-are-standards/health-data-standards/ [accessed 2024-10-19]
24. Personal Data Protection Act 2012. 2012. URL: https://sso.agc.gov.sg/Act/PDPA2012 [accessed 2024-10-19]
25. Protect Personal Data, 13/2023/ND-CP. URL: https://www.forvismazars.com/vn/en/insights/technical-updates/newsletters-alerts/vietnam-tax-legal-updates/legal-alert-june-2023 [accessed 2024-10-26]
26. Personal Data Protection Act, B.E. 2562, 136. Sect. 69. URL: https://cyrilla.org/entity/sl9175g71u?file=15887704957529yl1t08r6ge.pdf&page=1 [accessed 2019-05-19]
27. Law (UU) Number 27 of 2022 concerning the Protection of Personal Data. URL: https://en.mkri.id/news/details/2023-02-13/Govt:_Law_on_Personal_Data_Protection_Provides_Legal_Protection [accessed 2024-10-26]
28. Burman A. Understanding India's new data protection law New Delhi. Carnegie India. 2023. URL: https://carnegieindia.org/2023/10/03/understanding-india-s-new-data-protection-law-pub-90624#:~:text=The%202023%20act%20creates%2C%20for,clearly%20enumerated%20in%20the%20law [accessed 2024-10-19]
29. Suvannaphakdy S. Better safeguards needed for trusted data use in ASEAN countries. Yusof Ishak Institute. 2022. URL: https://www.iseas.edu.sg/wp-content/uploads/2022/06/ISEAS_Perspective_2022_67.pdf [accessed 2024-10-30]
30. Yulisman L. New draft law in Indonesia may be putting people's health data at risk: experts. The Straits Times. 2023. URL: https://www.straitstimes.com/asia/se-asia/new-draft-law-in-indonesia-may-be-putting-people-s-health-data-at-risk-experts [accessed 2024-10-26]
31. Brody C, Chhoun P, Tuot S, Fehrenbacher AE, Moran A, Swendeman D, et al. A mobile intervention to link young female entertainment workers in Cambodia to health and gender-based violence services: randomized controlled trial. J Med Internet Res 2022;24(1):e27696 [FREE Full text] [doi: 10.2196/27696] [Medline: 34982716]
32. Srinidhi V, Karachiwala B, Iyer A, Reddy B, Mathrani V, Madhiwalla N, et al. ASHA Kirana: when digital technology empowered front-line health workers. BMJ Glob Health 2021;6(Suppl 5):e005039 [FREE Full text] [doi: 10.1136/bmjgh-2021-005039] [Medline: 34548289]
33. Bajpai N, Beriya A, Biberman J, Sharma A, Wadhwa M. Delivering next-generation public services through mobile technology. Center for Sustainable Development. 2022. URL: https://csd.columbia.edu/sites/default/files/content/docs/ICT%20India/Papers/ICT_India_Working_Paper_65.pdf [accessed 2024-10-19]
34. Pengput A, Schwartz DG. Telemedicine in Southeast Asia: a systematic review. Telemed J E Health 2022;28(12):1711-1733. [doi: 10.1089/tmj.2021.0516] [Medline: 35417250]
35. Garg PR, Uppal L, Mehra S, Mehra D. Mobile health app for self-learning on HIV prevention knowledge and services among a young Indonesian key population: cohort study. JMIR Mhealth Uhealth 2020;8(9):e17646 [FREE Full text] [doi: 10.2196/17646] [Medline: 32896831]
36. Patel A, Praveen D, Maharani A, Oceandy D, Pilard Q, Kohli MPS, et al. Association of multifaceted mobile technology-enabled primary care intervention with cardiovascular disease risk management in rural Indonesia. JAMA Cardiol 2019;4(10):978-986 [FREE Full text] [doi: 10.1001/jamacardio.2019.2974] [Medline: 31461123]
37. Mao Y, Lin W, Wen J, Chen G. Impact and efficacy of mobile health intervention in the management of diabetes and hypertension: a systematic review and meta-analysis. BMJ Open Diabetes Res Care 2020;8(1):e001225 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001225] [Medline: 32988849]
38. Feroz A, Perveen S, Aftab W. Role of mHealth applications for improving antenatal and postnatal care in low and middle income countries: a systematic review. BMC Health Serv Res 2017;17(1):704 [FREE Full text] [doi: 10.1186/s12913-017-2664-7] [Medline: 29115992]
39. Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022;4(8):e615-e626 [FREE Full text] [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]
40. Valera ACL, Wei WL, Tan HP, Tan HX, Liang H. Real world, large scale IoT systems for community eldercare: experiences and lessons learned. In: Elderly Care: Options, Challenges and Trends. New York, NY: Nova Science; 2018:53-79.

XSL•FO
RenderX

41. Athey S, Bergstrom K, Hadad V, Jamison JC, Özler B, Parisotto L, et al. Can personalized digital counseling improve consumer search for modern contraceptive methods? Sci Adv 2023;9(40):eadg4420 [FREE Full text] [doi: 10.1126/sciadv.adg4420] [Medline: 37801502]
42. de Vere Hunt I, Linos E. Social media for public health: framework for social media-based public health campaigns. J Med Internet Res 2022;24(12):e42179 [FREE Full text] [doi: 10.2196/42179] [Medline: 36515995]
43. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2023;1(1):AIp2300031. [doi: 10.1056/aip2300031]
44. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. Lancet Digit Health 2022;4(4):e235-e244 [FREE Full text] [doi: 10.1016/S2589-7500(22)00017-6] [Medline: 35272972]
45. Williams D, Hornung H, Nadimpalli A, Peery A. Deep learning and its application for healthcare delivery in low and middle income countries. Front Artif Intell 2021;4:553987 [FREE Full text] [doi: 10.3389/frai.2021.553987] [Medline: 33997772]
46. Brynjolfsson E. The Turing Trap: the promise and peril of human-like artificial intelligence. Daedalus 2022;151(2):272-287. [doi: 10.1162/daed_a_01915]
47. Brynjolfsson E, Li D, Raymond LR. Generative AI at work. Working paper 31161. National Bureau of Economic Research. 2023. URL: https://www.nber.org/papers/w31161 [accessed 2024-10-26]
48. Modi D, Dholakia N, Gopalan R, Venkatraman S, Dave K, Shah S, et al. mHealth intervention "ImTeCHO" to improve delivery of maternal, neonatal, and child care services—a cluster-randomized trial in tribal areas of Gujarat, India. PLoS Med 2019;16(10):e1002939 [FREE Full text] [doi: 10.1371/journal.pmed.1002939] [Medline: 31647821]
49. Saing CH, Ung M, Suy S, Oy S, Dary C, Yam ELY, et al. i-MoMCARE: innovative mobile technology for maternal and child health care in Cambodia—study protocol of a cluster randomized controlled trial. Trials 2023;24(1):692 [FREE Full text] [doi: 10.1186/s13063-023-07724-z] [Medline: 37880782]
50. Digital technologies for health financing: what are the benefits and risks for UHC?: Some initial reflections. Contract No.: No. 19. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240031005#:~:text=To%20achieve%20progress%20towards%20UHC,and%20making%20purchasing%20more%20strategic [accessed 2024-10-26]
51. Singh K, Walters MR. Use of mHealth in promoting maternal and child health in "BIMARU" states of India "A health system strengthening strategy": systematic literature review. PLOS Digit Health 2024;3(2):e0000403 [FREE Full text] [doi: 10.1371/journal.pdig.0000403] [Medline: 38306391]
52. Pfammatter A, Spring B, Saligram N, Davé R, Gowda A, Blais L, et al. mHealth intervention to improve diabetes risk behaviors in india: a prospective, parallel group cohort study. J Med Internet Res 2016;18(8):e207 [FREE Full text] [doi: 10.2196/jmir.5712] [Medline: 27496271]
53. Banerjee A, Alsan M, Breza E, Chandrasekhar AG, Chowdhury A, Duflo E, et al. Messages on Covid-19 prevention in India increased symptoms reporting and adherence to preventive behaviors among 25 million recipients with similar effects on non-recipient members of their communities. National Bureau of Economic Research Working Paper 27496. 2020. URL: https://www.nber.org/system/files/working_papers/w27496/w27496.pdf [accessed 2024-10-26]
54. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Intern Med 2016;176(5):619-625 [FREE Full text] [doi: 10.1001/jamainternmed.2016.0400] [Medline: 26974260]
55. Hong J, Kikuta NT, Simos A, Tsai S, Lin B, Rodriguez F, et al. Testing the appropriateness of diabetes prevention and care information given by the online conversational AI ChatGPT. Clin Diabetes 2023;41(4):549-552. [doi: 10.2337/cd23-0026] [Medline: 37849522]
56. Martinez-Martin N, Luo Z, Kaushal A, Adeli E, Haque A, Kelly SS, et al. Ethical issues in using ambient intelligence in health-care settings. Lancet Digit Health 2021;3(2):e115-e123 [FREE Full text] [doi: 10.1016/S2589-7500(20)30275-2] [Medline: 33358138]
57. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
58. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
59. Mullainathan S, Obermeyer Z. On the inequity of predicting A while hoping for B. AEA Papers Proc 2021;111:37-42. [doi: 10.1257/pandp.20211078]
60. Evaluation of the MStyle program for men who have sex with men in Cambodia. HIV Innovate and Evaluate Project. 2016. URL: https://pdf.usaid.gov/pdf_docs/PA00MF6V.pdf [accessed 2016-06-13]

Abbreviations

ABDM: Ayushman Bharat Digital Mission
AI: artificial intelligence
BMGF: Bill & Melinda Gates Foundation

XSL•FO
RenderX

LMIC: low- and middle-income country
mHealth: mobile health
NGO: nongovernmental organization

Edited by G Eysenbach; submitted 21.02.24; peer-reviewed by G Akrong, MH Shah, G Agbozo, A Ray; comments to author 26.06.24; revised version received 30.07.24; accepted 10.10.24; published 25.11.24.

Please cite as:
Yi S, Yam ELY, Cheruvettolil K, Linos E, Gupta A, Palaniappan L, Rajeshuni N, Vaska KG, Schulman K, Eggleston KN
Perspectives of Digital Health Innovations in Low- and Middle-Income Health Care Systems From South and Southeast Asia
J Med Internet Res 2024;26:e57612
URL: https://www.jmir.org/2024/1/e57612
doi:10.2196/57612
PMID:

©Siyan Yi, Esabelle Lo Yan Yam, Kochukoshy Cheruvettolil, Eleni Linos, Anshika Gupta, Latha Palaniappan, Nitya Rajeshuni, Kiran Gopal Vaska, Kevin Schulman, Karen N Eggleston. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Ten Myths About the Effect of Social Media Use on Well-Being

Jeffrey A Hall[1], PhD

Department of Communication Studies, University of Kansas, Lawrence, KS, United States

Corresponding Author:
Jeffrey A Hall, PhD
Department of Communication Studies
University of Kansas
Bailey Hall
1440 Jayhawk Boulevard
Lawrence, KS, 66045
United States
Phone: 1 7858641082
Email: hallj@ku.edu

Abstract

This viewpoint reviews the empirical evidence regarding the association between social media use and well-being, including life satisfaction and affective well-being, and the association between social media use and ill-being, including loneliness, anxiety, and depressive symptomology. To frame this discussion, this viewpoint will present 10 widely believed myths about social media, each drawn from popular discourse on the topic. In rebuttal, this viewpoint will offer a warranted claim supported by the research. The goal is to bring popular beliefs into dialogue with state-of-the-art quantitative social scientific evidence. It is the intention of this viewpoint to provide a more accurate and nuanced claim to challenge each myth. This viewpoint will bring attention to the importance of using rigorous scientific evidence to inform public debates about social media use and well-being, especially among adolescents and young adults.

(J Med Internet Res 2024;26:e59585) doi:10.2196/59585

KEYWORDS

social media; well-being; health promotion; depressive disorder; depression; anxiety; adolescent; mental health

Claims about the harms of social media are everywhere. Many are taken as fact. In the face of this certainty, it is useful to remember that it was once taken for granted that serial dramas on the radio, comic books, going to the cinema, and arcade games were all once considered to be undeniably harmful, particularly for the youth [1]. This is not to discount the legitimate interests of the public and policy makers to understand whether and how social media influence well-being. Locating the latest and most accurate research information is challenging. Evidence on the topic is endlessly updated, reviewed, and debated by researchers from fields ranging from psychology and communication through education and computer science to neuroscience and economics.

To understand this new form of an old controversy and to summarize complex and scattered academic evidence, this paper will use a myth versus warranted claim structure. This paper will present 10 myths about the harms of social media. Each myth is directly quoted or paraphrased from public discourse, including op-eds, podcasts, claims of politicians, and news headlines. In response to each myth, a warranted claim supported by peer-reviewed research will be offered. Meta-analyses, studies with large or representative samples, and studies using longitudinal methods will be prioritized as sources for supporting and justifying each warranted claim.

Social media are platforms that enable searchable and publicly distributed content, including text, photos, memes, and videos [2]. Social media are often identified by a branded platform name (eg, Instagram). Users create content and this content is transmitted in a decentralized fashion. Users—to varying degrees depending on the platform—decide what is shared and reshared. Platforms vary considerably as to whether it is normative for users to share content, or whether the content is typically distributed by prominent accounts or users (eg, influencers or content creators). Meta-analyses suggest that social network sites, particularly Facebook, are the most studied social media [2-4]. Thus, most meta-analytic claims are drawn from evidence from decades of research on Facebook use, primarily among American college students [2-4]. Social media platforms that distribute video content (eg, TikTok or YouTube) have recently become quite popular [5], but are less well-represented in the research on well-being. Whenever possible, this viewpoint will focus on time using social media,

XSL•FO
RenderX

rather than screen time generally, which often includes TV, gaming, and internet use.

In this paper, well-being will be defined as including both eudaemonic and hedonic well-being. While the former focuses on meaning, connection, and life purpose, the latter focuses on pleasure, enjoyment, and entertainment. Both well-being and ill-being will be examined. Well-being is typically measured in social media research as life satisfaction, positive emotions, and the absence of negative emotions, while ill-being is typically measured as the presence of mental health symptomology (eg, anxiety, depression, or loneliness) [4,6].

This paper will focus on the effects of social media on the well-being of users as an aggregate. Social science research reports effect sizes, typically expressed as a correlation or mean difference. By nature, the goal of quantitative social science is to document the association between 2 (or more) variables, not to document the singular experience of each participant in the study. Thus, trends in the data, reported as aggregate effects, are not certainties that equally apply to all people. No association between 2 variables is equally true for every single person in the sample. Thus, it can be true that the warranted claims are accurate for the whole and that they do not perfectly align with every individual's experience.

Furthermore, this paper will not examine cases of real-world harms conducted through or enabled by social media. Social media are used to facilitate the sale of illegal substances, to coordinate sex trafficking, to plan terrorism, and to send death threats. Each activity is illegal whether conducted through a mobile device or social media platform or on landline telephone or mailed post. When this paper speaks about the effects of social media, it will focus on its typical and mundane uses (ie, how most people use social media most of the time), not social media when used as a telecommunications channel to commit illegal acts. If the reader's concerns about social media are focused on illegal activities, this paper will not be informative. If the reader feels that the harms of social media when used to commit illegal acts outweigh the influence of mundane and typical uses of social media on users' well-being, then this paper will not be persuasive.

- *Myth one: There is undeniable evidence that time spent on social media has a toxic effect on its users.*
- *Warranted claim one: Time spent on social media does not have a strong effect on the well-being of its users.*

There are 2 decades of research on the association between social media and well-being. A comprehensive meta-analysis that reviewed 226 studies published between 2006 and 2018, including 275,728 participants and 1279 effect sizes, reported a weighted mean effect size of r=.01, which is no different from 0 [4]. An analysis of 3 data sets, including 355,000 adolescents, found that the association between social media use and well-being accounts for, at most, 0.4% of the variance in well-being, which the authors conclude is of "little practical value" [7]. Another large study of adolescent users concluded that the association was "too small to merit substantial scientific discussion" [8]. A longitudinal study that measured social media use through an app installed on participants' mobile devices found no associations between any measures of Facebook use and loneliness or depression over time [9].

Along with cross-sectional and longitudinal studies, there are research studies that have tested what happens when people stop using social media, but they are fewer in number. Of the 20 that tested the effects of a digital detox (70% of the 20 specifically restricted social media use), the majority found no effects or mixed findings [10]. The authors concluded that "*mixed findings exist* [i.e., some benefits and some harms of abstaining], but no clear answer can be given yet" [10]. Some social media interventions have studied the change in global well-being after a very short period (eg, 7-20 minutes) [11], which calls into question whether any change could even have occurred in that interval. By comparison, studies that require participants to abstain from social media for a week or more report no changes in daily loneliness, affective well-being, and positive or negative affect [12-14].

One could also ask the question "How much would a person have to reduce their social media use to improve their well-being?" Two studies sought an answer. The first estimated that the change would have to be substantial: "5 hours 8 minutes of daily device-based engagement [would have to occur] before caregivers would be able to notice subjectively significant variations in psychosocial functioning" of their child [15]. Another study of Spanish adolescents estimated that to experience a gain in well-being, social media use would need to be reduced by 10 hours a day [16]. Both estimates far exceed the amount of time typical users spend on social media. Among adolescents and young adults, objectively measured social media time ranges between 2 and 2.5 hours a day [17,18]. For adults, objectively measured daily social media use is only 30-45 minutes [5,19].

Maybe social media use causes stress. When stress is measured through biological methods (eg, cortisol in a saliva sample), studies have found that social media use decreases stress [20,21] or has no association with stress [22]. Substantial reductions in screen time, including time on social media, have no effect on biologically measured stress [23]. One study [24] compared the argument that social media cause stress with the argument that social media are sought out after being stressed. It concluded that stress seems to precede social media use rather than being caused by it [24].

Overall, research suggests that screen time, including social media use, plays little to no role in the well-being of most users. Its effects are not conclusively harmful or toxic in terms of decreasing well-being, increasing ill-being, or causing stress. Rather, the effects of social media on users are negligible but heterogeneous (ie, the experience varies between individuals) [25]. These 2 claims are not inconsistent; it can both be the case that social media do not influence the well-being of most users, but might help or hurt some users in some circumstances.

- *Myth two: Social media addiction is pervasive and harmful.*
- *Warranted claim two: Experts disagree on whether social media addiction exists, what the diagnostic criteria are, and how it should be measured.*

XSL•FO
RenderX

Most Americans (56.9%) believe they are addicted to their smartphones [26], but many experts would disagree with that self-diagnosis. For over 20 years, there has been research on *internet* addiction, which has since accommodated new technologies (ie, the smartphone and social media). However, the *Diagnostic and Statistical Manual of Mental Disorders, fifth edition* (*DSM-5*) does not list internet addiction [27] as a diagnosis and it has not since included social media addiction. Part of the challenge lies in good measurement.

The measurement of technology-based addiction is incredibly fraught. These measures are deeply problematic and major researchers have discounted their value [4,6,28]. The most concerning issue is the tautological relationship between cause and consequence [5,28-30]. Essentially, a person cannot report having a technology addiction without simultaneously reporting a mental health problem [28-30]. One comprehensive review [30] concluded that the measures of technology-based addiction have invalid and arbitrarily designed instruments and major inconsistencies between instruments in terms of conceptual definition, assessment items, and diagnostic criteria. These measurement problems also influence the interpretation of meta-analyses. Specifically, associations between ill-being and internet or social media addiction are substantially stronger than associations between ill-being and time on social media [4,6,31]. Thus, social media addiction research is misleading due to the comorbidity of cause and consequence and poor overall measurement.

The Digital Wellness Lab associated with Harvard Medical School recommends using the phrase “problematic interactive media use” while cautioning that it is not clear whether media use is a cause of an underlying mental health challenge or a manifestation of it [32]. In other words, a large amount of time spent using the internet, smartphones, social media, or video games is probably externalization of psychological disorders, but is unlikely to be the root cause of such disorders [30].

Although a shift from “addiction” to “problematic use” is important, the existing measures of problematic social media use also need further development. Problematic use measures have similar problems as addiction measures, namely that high social media use and mental health impairment are measured using the same items [5,29]. Some researchers have called for a complete reconceptualization of problematic social media use [4,29,30]. Until a valid and reliable measure of problematic use emerges, studies using those instruments should be interpreted with the above limitations in mind.

The myth of social media addiction has major policy implications. Calling something an addiction means it should be *treated* as an addiction. One clear example of how the language of addiction frames public deliberation is the comparison between tobacco use (and tobacco companies) and social media use (and technology companies). This framing justifies the proposed solutions. The argument goes that if social media use is as harmful and addictive as tobacco, then policy makers should combat it as a public health crisis. Indeed, it is common for policy makers and the public to use the drug metaphor to explain how social media function and what the government should do about it [33]. Yet, experts agree that social media addiction is not a diagnosable addiction (ie, it lacks the necessary characteristics of addiction) [27], and, even if it were an addiction, there are no valid measures to document its prevalence. This does not dismiss the possibility that problematic use exists. Rather, to study problematic use, we need better conceptualization and measurement [30] to guide policy decisions and treatment for individuals [32].

- *Myth three: Spending more time on social media will inevitably make users depressed, anxious, sad, and lonely.*
- *Warranted claim three: Over time, declines in well-being are associated with increased social media use.*

The third myth is centered on the question, “Which comes first?” There are 4 possible sequential orders of the association between social media use and well-being over time: changes in social media use decrease/increase future well-being exclusively; changes in well-being increase/decrease future social media use exclusively; both changes affect each other; change in one is not predictive of change in the other. The myth is that social media use *increases* ill-being (or *decreases* well-being) in the future.

The empirical evidence supports the sequential order opposite to this myth: declines in well-being precede increased future social media use. A meta-analysis of longitudinal studies on social media and well-being found that changes in well-being (eg, eudaemonic well-being, hedonic well-being, and social well-being) are associated with changes in social media use, but not the other way around [4]. A longitudinal analysis of 84,000 UK adults aged 10-80 years from 2011 to 2018 found that drops in well-being were followed by more social media use over time across all age groups [34]. The same study, however, found that increases in social media were associated with future declines in well-being for a few adolescent age groups. Other research has found that increased depression predicts future social media use, particularly for adolescents, but not vice versa [35]. Among adolescents, social media use tended to increase after reductions in friendship quality, suggesting that interpersonal struggles precipitate future social media use [36]. Another study found that both declining life satisfaction and higher psychological distress corresponded with increased future social media use among adults [37].

It is important to note that other studies have found no relationship between well-being and social media use over time. One study, including thousands of Dutch adults over 6 years (2012-2017), found no association in either direction [38]. Another identified a bidirectional relationship, but concluded that these associations were so small as to be trivial [39]. Yet another suggested that loneliness and social media use are unrelated over time for adolescents [36]. Sensitive to the difference between well-being and ill-being, a meta-analysis reported that changes in ill-being (ie, anxiety or depression) were unrelated to social media use over time [4].

Another way to frame the discussion is not to focus on change over the years, but to consider shorter time frames, such as within or between days. There are several reasons why people might turn to social media when feeling down, lonesome, or anxious: it is always available, it is distracting, and it could provide a way to get in touch with responsive others [40,41].

XSL•FO
RenderX

But, social media may not be effective at getting one's needs met. Adolescents commonly use technology to cope with negative emotions, but social media use does not appear to contribute to effective emotional regulation between days [42]. Similarly, people who tend to rely on social media to cope with loneliness do not seem to get their needs met over days or months [43,44]. Thus, people may turn to social media when stressed or down, but social media use does not appear to be effective at helping individuals cope with those emotions.

Focusing on different ways of using social media, a 3-year longitudinal study of late adolescents in Spain examined the use of social media as an escape from problems. Escapism was associated with more time on social media and future declines in well-being [16]. These 3 concepts (ie, escapism, time on social media, and negative outcomes) showed a circular relationship with one another. Perhaps social media functions like a social snack—temporarily redirecting or distracting an individual from their negative affect or loneliness, but failing to fully satisfy their needs or cope with the origin of their negative affect [40,45,46]. For example, during the pandemic, turning to social media to cope with loneliness increased loneliness over time, accounting for the frequency of face-to-face and video chat conversations [47]. Unfortunately, very few studies have focused on how variation in well-being works in tandem with using social media to cope with negative feelings or to escape one's problems [43]. Thus, more research is needed to determine if some platforms or specific patterns of use are effective for coping or getting one's needs met.

Although evidence suggests social media use is a poor coping strategy, this is not equivalent to saying that it is directly harmful. Rather, many researchers have suggested that social media use, especially to escape or distract an individual from their problems, is a misdirected solution [4,16,24,42,43,45-50]. This is particularly concerning for adolescents who are already at high risk for ill-being; building a strong habit of social media use to cope may contribute to poor self-regulation [48-50]. Furthermore, because there is some evidence of a bidirectional relationship over time among some adolescents [34], it is possible that using social media to cope could exacerbate existing social, emotional, or mental health challenges. By contrast, because several large longitudinal studies have found nonexistent or trivial bidirectional relationships between social media use and well-being, it is important that future research studies test the "exacerbation explanation" further.

- *Myth four: Social media are the main cause of the problems teens are facing.*
- *Warranted claim four: Preexisting vulnerabilities (eg, poverty, mental health, lack of family support) are associated with both adolescent social media use and adolescent ill-being.*

Some would argue that social media are a primary cause of ill-being because adolescents who are lonely or depressed use social media frequently. When 2 things co-occur frequently, social scientists try to explain why this happens. In pursuit of this goal, researchers have begun to identify what sorts of adolescents are heavy social media users and experience high ill-being. In other words, social scientists look for a third variable that explains both things. If the heaviest social media users also experience high degrees of psychological dysfunction, but *forsome other reason than social media*, then it could explain the small, but significant, association between social media and ill-being [28,31].

Various preexisting socioeconomic and psychological vulnerabilities (eg, poverty, poor mental health, or lack of family support) precede or co-occur with social media use. Families who are struggling to get by financially and experience frequent stress are more likely to have children with heavier device engagement [13]. A longitudinal study of over 12,000 UK adolescents found that lower family support was associated with greater social media use across time [50]. In the same study, social media use itself did not predict changes in ill-being or well-being [50]. Compared with social media use, family struggles were much more strongly associated with ill-being. An 8-year longitudinal investigation found that when an adolescent was in a riskier home environment (eg, low parental involvement) or was at risk by disposition (eg, low behavior self-regulation), then social media use was related to greater depression over time [51]. Similarly, in lower education and income families, social media use is more prevalent than other types of media or internet use [52]. Adolescents who report problematic smartphone use are more likely to experience stress, poorer health, and less belonging at school [52]. The same study concluded that economically disadvantaged youth report more spillover from online experiences to serious offline problems, including face-to-face confrontations, physical fights, and getting into trouble at school.

By contrast, social media use has little effect on well-resourced users—it may even enhance their well-being. Social media use is associated with greater well-being when people are already socially well-connected [53]. Similarly, adolescents with high well-being are less influenced by long durations of smartphone use [18].

These findings are bolstered by research that separates between-person effects from within-person effects. Between-person effects occur when certain types of people engage in specific media practices. For example, people who are unemployed and socially isolated tend to watch a lot of TV [54]. By comparison, within-person effects occur when media use explains variance in well-being beyond what could be attributable to the between-person effects. Within-person effects in the prior example would mean no matter who the person is, frequent TV watching would be associated with social isolation. When within-person effects are found for media, it rules out the argument that the characteristics of people are the whole story. Research documenting within-person effects has found no association between well-being and social media use [38,55-58]. These studies were conducted across countries, various time frames (days to decades), using large samples, and with longitudinal designs. One meta-analysis of social media use and depression among adolescents confirmed the presence of between-person effects, but not within-person effects [59].

By comparison, there is strong evidence that the association between social media use and well-being lies at the between-person level. This means different types of people

XSL•FO
RenderX

engage in different social media practices, rather than social media use *causing* harm. One way to interpret this is to consider the people who have the available time and motivation to be on social media for extensive periods. An abundance of time and a lack of desirable options for leisure or connection may reflect preexisting psychological dysfunction as well as social media users' socioeconomic or socioemotional context.

Therefore, knowing a person's background provides a considerable amount of information about their media habits. Whether coming from a "promote the positive" or "mitigate the negative" perspective, researchers have concluded that more attention should be paid to *who* uses social media and *how* they use it. From a digital thriving perspective, researchers have embraced the importance of context in understanding social media use: "We do not interact with our devices in a vacuum, however: The interactive and dialogical nature of digital media implies that our use of them cannot be considered in separation from our social context." [60]. Researchers operating from a harm mitigation perspective argue that "the conversation surrounding the legislation of social media should focus on determining which segments of the population would benefit from external regulation of social media platforms (e.g., those with psychologically vulnerable dispositions) instead of focusing overtly on the unrealistic end-goal of benefiting all segments of the population by implementing blanket policies." [58].

- *Myth five: Compared with other harms, the harm of social media use is far greater.*
- *Warranted claim five: Once the primary predictors of well- and ill-being are accounted for, social media use is a negligible factor in explaining variance in well- and ill-being.*

According to popular opinion, it is social media's fault that bad things are getting worse. For this to be true, among all the factors that cause harm to a person, social media use would have to be a big one. While the fourth myth-claim pair explores the co-occurrence of social media use and a person's family, socioeconomic, and psychological struggles, the fifth pair explores how important social media use is in the big picture of a person's overall ill- and well-being.

To proceed, a little background is needed. Early research on social media typically measured social media use (usually the amount of time spent) and then measured global ill-being (eg, loneliness and anxiety) [5], with measures of eudaemonic well-being being relatively uncommon [61]. In these studies, social media use was not measured side-by-side with known indicators of ill- and well-being. Study design influences what we can learn and know. Earlier study designs answered the question "Is social media use associated with ill-being?" More rarely used study designs answer the question "Is social media a stronger or weaker indicator of well- and ill-being compared with other factors?" Combined with a reliance on cross-sectional data, early research on social media could not establish the comparative or relative effect of social media—there was not enough information. Only recently have studies begun to measure social media use side-by-side with other indicators of well- and ill-being.

Social media use explains very little variance in well- and ill-being, compared with other factors. A large study that compared the association between social media use and well-being along with other characteristics of adolescents concluded that "the association of well-being with regularly eating potatoes was nearly as negative as the association with technology use, and wearing glasses was more negatively associated with well-being." [6]. A large longitudinal study of adolescents concluded that social media are among the least important factors in predicting changes in mental health [50]. A weighted national sample of over 20,000 Americans found that social media use was one of the least valuable predictors of loneliness—and, in this study, more Facebook use was associated with less loneliness [62].

What about social media use compared against other media (eg, TV or the internet)? Studies that sought to answer this question suggest that no form of media use is any more or less predictive of well-being than any other [55-57]. Simply, no form of media consumption seems to matter that much in the big picture. One group of authors concluded "These findings are incompatible with the dose-response model that underlies societal discourse" of social media harms [56]. In other words, once considered in the context of other factors, the equation *X amount of social media use leads to Y amount of harm* does not hold up.

What about over time—does media play an important role there? Long-term longitudinal studies (eg, over several years) are well-equipped to explore gradual changes in ill- and well-being. Such studies suggest that social media use does not matter much [6,7,50,57]. Thus, it is very unlikely that social media use plays much of a role in changes in eudaemonic well-being, life satisfaction, or mental health, which are typically stable and change only gradually [63].

Overall, social media use is a tiny (perhaps nonexistent) factor in predicting ill- and well-being. Other major factors—and even some smaller factors such as wearing glasses—play a bigger role than social media use in understanding whether a person experiences greater well-being or less ill-being. Global well- and ill-being change over long periods or because of major changes in life circumstances; they are probably not meaningfully affected by a person's preference of media.

- *Myth six: The adoption of social media, especially on mobile devices, perfectly coincides with the beginning of the contemporary adolescent mental health and loneliness crises.*
- *Warranted claim six: Longitudinal studies do not support the conclusion that the adoption of mobile or social media preceded or caused declines in adolescent mental health or the adolescent loneliness epidemic.*

Time spent socializing has been on the decline for at least 30 years in several countries across the global north [54,64-66]. Although social time has declined across demographic groups, unmarried, unemployed adult men appear to have experienced the greatest increase in time alone in the United States in the past 50 years [54,64]. Loneliness is typically high among late adolescents and young adults because it is associated with the developmental challenges of emerging adulthood. However, there has been a meaningful rise in loneliness among adolescents

XSL•FO
RenderX

and young adults in the global north [64,66]. In response to these trends, it has been argued that there is a direct relationship between the adoption of social media and smartphones (~2011-2014) and subsequent increases in ill-being. Some people have concluded that social and mobile media have directly caused the rise in loneliness and decline in mental health among the youth.

Longitudinal findings do not support this myth. One study of US adolescents from 2009 to 2017 found that changes in depression were not related to prior social media use or between-year changes in social media use [67]. Rather, this study found between-level associations between social media use and mental health that are consistent with those discussed above (see myth five). A study of German adolescents conducted over the decade in question found no evidence that social media use influenced future depression [68]. An 8-wave longitudinal study (2009-2016) of thousands of adolescents in the United Kingdom measured social media use during the school day and found that between-year associations between social media use and well-being were trivial [39]. A meta-analysis of studies from 2006 to 2018 found that the size of the association between social media use and well-being decreased over those 12 years [4]. In other words, social media use had a weaker association with well-being as studies were conducted closer to the present day. The same study found no drop-off in well-being after 2012.

By contrast, a preregistered meta-analysis of loneliness over 50 years found that loneliness has indeed increased for young adults (mean age of 21 years), but its rise preceded the introduction of social media and the adoption of smartphones by nearly a decade [69]. This means loneliness was increasing and time spent socializing was decreasing well before the widespread adoption of social and mobile media.

Another way to challenge this myth is on a country-by-country basis. One study [70], using data that spanned from 2008 to 2019, examined the portion of active Facebook users in 72 countries (as a percent of 13- to 34-year-olds in the population) in relation to well-being (ie, life satisfaction and negative and positive psychological experiences). Facebook usage data were provided by Meta Platforms, Inc. and the Gallup World Poll provided the measures of well-being. Including data from nearly 1 million people, there was no evidence that rates of adoption of Facebook were associated with changes in well-being for young people. In fact, this study found that an increase in positive experiences was associated with greater Facebook adoption within countries. The same study further demonstrated that Facebook adoption was unrelated to changes in mental ill-being (ie, anxiety, depression, or self-harm) within countries [70]. This study is particularly notable as it used objective measures of social media adoption, which is rare in social media and well-being research [28,63]. Similar to the above cross-country results, 2 other longitudinal studies of adult social media users found that greater social and mobile media use was associated with increases in well-being in subsequent years [71,72].

It is reasonable to argue that declining well-being and rising loneliness may influence social media use (see myth three). It is consistent with the evidence that individuals who find themselves in less social circumstances or who seek to cope with isolation or loneliness may turn to social media to distract, escape, or connect [18,43,44,47]. However, these 3 purposes yield very different outcomes. The use of social media to reach out and communicate with loved ones is most likely to help ameliorate loneliness [18,40,41]. By contrast, social media use to escape or cope with negative emotions may not be effective [18,42-44,46]. Social media does not have to be the cause of declining mental health or increasing loneliness to be an unsuitable coping mechanism for individuals who are already depressed or lonely.

- *Myth seven: Social media are the reason people don't spend time together.*
- *Warranted claim seven: Social media use does not cause people to stop people from talking to each other face-to-face, rather they are used to help people keep in touch when face-to-face interactions decrease.*

This myth has a name when studied by academic researchers: the social displacement hypothesis. This hypothesis states that rising rates of technology use cause declining rates of face-to-face communication, subsequently reducing well-being. Over the last 20 years, there have been rising rates of social media use and declining amounts of social time across the global north [64,65]. Where does social media time come from? If borrowed from face-to-face social time, it is reasonable to argue that this could reduce well-being. If borrowed from other media, it is unlikely to affect well-being [55,56].

Social media time is probably displacing time spent on the internet and watching TV, just as internet use once borrowed time from TV and other media [73-75]. Patterns of media use by US adolescents between 2006 and 2016 suggest that social media use climbed while all other forms of media use fell (eg, books, magazines, newsprint, or movies at the theatre) [74]. More recently, social media platforms have been used to access broadcast, streaming, and influencer or creator content. From 2018 to 2023, about 45 minutes of entertainment content shifted from being watched on a TV to being watched on a mobile device, often through social media platforms [5,76]. Thus, a substantial portion of mobile screen time is spent doing what might be called "watching TV" on a personal device. These days much of this TV-like content is distributed through social media platforms (eg, TikTok, YouTube).

The social displacement of face-to-face interactions is rarely tested using experimental designs (eg, randomly assigning participants to increase social media use). One study tested displacement by reporting what activities people reintroduce when they abstain from social media for weeks at a time [73]. When abstaining, people mainly spent more time browsing the internet, working, and engaging in household activities [73]. Another study found that after 2-4 weeks of Facebook abstinence, participants reported watching more TV and using the internet more during their time away [73]. Two studies found that when abstaining from a platform, people simply redirect their communication and entertainment to other social media platforms [12,73]. In fact, 3 studies found evidence contrary to social displacement: on days that participants abstained from social media, they engaged in fewer social interactions with

XSL•FO
RenderX

weak ties [73,77] and engaged in fewer phone calls, texts, and emails [78].

Rather than displacing social interaction, online communication mirrors offline communication. Social media are used to communicate by already social people [79-81]. A longitudinal study of Dutch adults found that face-to-face communication is associated with social media use, and, over time, increased social media use does not decrease time talking to close others [81]. Another German study found that active social media use is positively associated with face-to-face communication 6 months later [82]. A longitudinal study conducted in the early years of social media adoption (2009-2011) found that social media use was not associated with reduced frequency of face-to-face contact within years or in the future [71]. An 8-year longitudinal study of Norwegian children and adolescents (10-18 years old) found that social media use predicted more time with friends offline in the future [83]. The same study revealed that using social media did not affect the development of social skills (ie, communicating online did not harm the development of social skills in general) [83]. These findings are inconsistent with the social displacement hypothesis.

Social media use, particularly texting and group chats, may compensate for a lack of opportunities to engage in face-to-face interaction, such as long-distance friends and family. On days with fewer face-to-face interactions, people turn to smartphones to communicate [84]. When opportunities for in-person contact decline, people use social and mobile media to connect [41].

- *Myth eight: Teens using their smartphones around each other is a sign of a disconnected and discontented generation.*
- *Warranted claim eight: The effects of co-present mobile use are highly situational and influenced by social norms.*

Using one's mobile phone in the presence of other people is called co-present mobile use or "phubbing" (phone + snubbing). Co-present device use is not social media use per se, as mobile devices are used for a variety of reasons, increasingly for streaming services, watching sports, and online videos [5,19]. However, a great deal of mobile phone use is social media use as well [5,19,76]. Research on phubbing often uses experimental methods to study how people feel during co-present device use, particularly during first conversations between strangers. A meta-analysis confirmed that co-present device use has several negative social consequences during first interactions, leading to the interaction partner having a negative impression of the phubber [85]. There is longitudinal evidence that chronic co-present mobile use affects adults' romantic relationships. Individuals who perceive their partners to be overly attentive to their smartphones tend to have fewer positive interactions and more negative moods [86,87].

There is evidence that social media use may diminish well-being when potential conversational partners are nearby. If young adults use social media in the company of close friends and family, they experience lower well-being than when they use social media when by themselves [58]. Perhaps, using social media in good company is an opportunity cost [58,81,88]. Yet, it is also possible that people become more attentive to their devices when they are in socially uncomfortable circumstances [40]. However, there is some evidence that challenges the presumptive harms of co-occurring face-to-face and social media use. One study found that engaging in both within the same period is associated with higher affective well-being among adolescents, compared with face-to-face communication alone [89]. In other words, when adolescents reported that they had both face-to-face and virtual social interactions, they reported the highest momentary well-being.

There are other important contextual factors to consider. One observational study suggested that co-present mobile device use is quite common in real life and has negligible effects on relationships [90]. Using a mobile device near a friend while waiting for a class to start does not lead to attentional conflict for either person and it has no association with relational intimacy. Most of the time, the people waiting together do not even notice that a device has been used. Furthermore, it is quite common for people who enjoy each other's company to share the content of their mobile devices with one another [91,92]. In such circumstances, co-present device use is a way to share something funny or interesting. After all, watching media together—whether a movie or a meme—can be an interpersonal experience.

Overall, co-present device use may negatively influence relationships, first impressions, conversation quality, and, perhaps, momentary well-being. The negative effects are most apparent when people are in a location where it is not appropriate to use a mobile device, people have high-quality relationship partners to talk to [58], and when people use their mobile device to the exclusion of their conversation partner(s) without offering an explanation or apology [40]. As a consequence, the harms or benefits of phubbing are best understood in the context of interpersonal conversational norms.

- *Myth nine: The solution is to quit or ban social media.*
- *Warranted claim nine: The benefits of social media abstinence vary by person and by patterns of use.*

Social or mobile media detox, reduction, abstinence, or prohibition are frequently offered as solutions to their presumptive harms. Logically, if something is bad for you, then quitting it is a good solution. Recommendations vary from school day bans to dinner table or bedtime bans, to uninstalling all social media apps. Some bans and reductions of social media use are self-driven and consistent with personal goals. To the degree that an individual wishes to use social media less, researchers and practitioners should help them achieve their goals. After all, successful media management promotes personal autonomy [60]. Other bans, however, offer no exceptions for personal patterns of use as they are imposed by governments, schools, or other institutions. The effects of social media prohibition by institutions (eg, a school-wide ban) on their members are understudied and remain a crucial topic for future research. However, there is good reason to believe that social media abstinence is unlikely to achieve the desired aims for all users equally.

One reason for the lack of efficacy is that there are benefits associated with social media use. The use of social network sites, particularly Facebook, is associated with online social capital [93]. When compared with online games and general

XSL•FO
RenderX

internet use, social media use is particularly beneficial for cultivating bridging social capital, which provides access to unique information and opportunities [3]. Through social network sites, users have access to diverse sources and can ask for or lookup information or aid. A recent research review confirmed the positive association between social media use and online social capital across the world, including the global south (eg, Pakistan, Iran, Palestine, Bangladesh, and China) [94]. When social media enable access to online communities, abstinence from social media could be a loss of access to information, aid, and support. Furthermore, abstinence would be a loss for people who have offline communities that expect group members to be available on social media, including friendship groups, workplaces, teams, or volunteer organizations. Leaving social media means not being informed about one's offline community too [40,60,77]. It is important to note that this online availability may create its own pressure to be constantly responsive and available, which may interfere with users' autonomy and well-being by creating digital stress [17,40,60].

A second benefit is associated with engaging in one-on-one or one-to-few conversations, such as texting or group chats through social media (eg, Snapchat, WhatsApp, or Facebook Messenger). Texting, sharing content, and responding to others' content (sometimes called *active* social media use) may be weakly associated with global, eudaemonic, and social well-being [4,28]. A recent meta-analysis suggested that the difference in effect between passive and active social media use is probably negligible, however, partly because both effects are very small [95]. It may depend on whom a social media user is communicating with [40]. Among Spanish young adults, using social media to communicate with friends and family is associated with greater well-being over time [16]. Chatting online and giving feedback to others are both associated with positive emotions and affective well-being for Dutch and German adolescents [25,96]. Indeed, texting frequency is associated with global well-being and reduced loneliness [9,43] and is particularly important for maintaining long-distance romantic relationships and friendships [97-100]. For users of social media in supportive social environments, texting is associated with reduced depression in longitudinal research [101]. In fact, a moderate amount of texting has been shown to offer a protective effect for some adolescents, even for those in socially disadvantaged environments [101]. This suggests blanket bans on social media use, especially ones that prevent mobile texting and social media direct messaging, will not help those who are most vulnerable.

The third benefit of social media use is in the domain of hedonic well-being. Social media content is often enjoyable, funny, and inspiring [102]. Italian adolescents have found that smartphone use is associated with short-lived hedonic pleasure [18]. Social media use may influence eudaemonic well-being, particularly when users seek content that is inspiring, motivational, and growth oriented, such as learning new skills [60,102].

The consequences of quitting social media can be found in studies that experimentally assign people to reduce use or abstain altogether. In interventions for social media use that require abstinence, 26% of 39 studies reported declines in well-being after leaving social media, particularly in the domain of social connection [11]. Meta-analyses suggest that social media use is more strongly associated with social well-being than it is negatively associated with anxiety or depression. These associations are drawn from reviews that include social media addiction studies, which tend to inflate negative effect sizes [4].

An exemplary summary [103] of when and why abstaining from social media would be helpful concluded that the benefits of quitting are not straightforward—even for people who want to. People feel guilty about using social media because they believe that it is harmful or unproductive, or both [103,104]. Part of the negative feelings associated with mindlessly scrolling is due to the guilt about doing so [105]. Therefore, successfully abstaining from social media can bring about a boost of well-being ("I stopped something that is bad for me" or "I accomplished my goals") [13]. But, failure to break a strong habit can lead to negative self-evaluations ("I failed to achieve my goal") [103]. This could be particularly pernicious if self-control failure is how individuals arrived at the conclusion that they need to digitally detox in the first place. For example, mindless scrolling is most strongly associated with goal conflict for those with the least self-control [105]. It is likely that self-control failure both contributes to frustrations with social media and makes it difficult to break well-established media habits. When people try to restrict their use by changing their notifications or smartphone-specific reminders, they experience more anxiety, worry, and check their mobile devices even more [103]. This suggests that technological solutions may not provide a quick fix—quitting an ingrained habit is difficult and aversive.

Finally, it is important to point out that restricting mobile communication may be particularly harmful to those who need access to their online community the most. Approaches developed from the need to belong framework have suggested that mediated communication with loved ones is better than no communication, especially when people are going through times of social deprivation [40,41,99]. People who increased mediated communication at the beginning of the pandemic reported significantly more closeness to friends and family compared with those who did not [47]. As an important cautionary note, this benefit does not seem to apply to all types of social media use [43,46,57]. Whether texting and active social media use are sufficient to meet individuals' fundamental need to belong or to ameliorate loneliness over longer periods (eg, weeks, months, or years) requires further study.

- *Myth ten: We do not need another study on social media.*
- *Warranted claim ten: The research on the harms or benefits of social media must continue as platforms, features, habits, contexts, and users constantly change.*

Some prominent researchers and advocates speak definitively and without reservation about the harms of social media. They often find an audience who agree with their warnings. After all, belief in the harms of social media is widespread. Other researchers offer nuanced, hedged, and complex responses to the question of harm [103,106,107]. Embracing the ideas espoused in this paper—for example, that social media are a small player in a big stage of adolescent development and do not have a causal effect on well-being—will be unsatisfying to

XSL•FO
RenderX

many people. It may fly in the face of many adolescents' and parents' personal experiences or beliefs. Clear certainty of an obvious villain is intuitively more satisfying than answers that resemble "it's complicated" and "we need more research." It is completely understandable if the public is frustrated with the state of social media research (or the contents of this paper).

Researchers themselves disagree about how to interpret the results. Where some see substantial or clearly detrimental results, others see trivial or meaningless effect sizes [6,28]. Two well-designed studies of the same phenomenon will find different results. This is due to the nature of social science; each study contributes a little more information to an incomplete puzzle. Furthermore, social media are always changing, which means the puzzle picture itself is in flux as researchers seek to assemble it. By the nature of science and by the nature of the topic, research on social media will never be definitive, unchanging, and 100% accurate. It will always be probable and contingent. And researchers *should* always be open to being wrong. Good research is open to criticism and improvement and recognizes contingencies and nuance. The warranted claims offered in this paper may prove wrong as more evidence mounts. As unsatisfying as it may be to policy makers and the public, this is how good science works. In the meantime, we need more research, with better designs, and better theory to explain when and why social media might affect well-being.

Vanden Abeele and her colleagues [33] offered an intuitive way to think about social media. They present 3 metaphors about its nature: a drug (ie, the addiction framework), a demon (ie, insidious platform design), or a donut. They argue that the final metaphor is most useful and compelling. A donut may not be the best source of nutrition, but whether it is bad for a person depends on the characteristics of the person and the context. If it is wanted and not a substitute for nutritious food, then it is probably fine. On any given day, eating a donut is not going to change the trajectory of a person's life. Eating donuts, no matter how delicious, will not make a person happy or healthy in the long run. Eating a lot of donuts is a bad coping mechanism for real problems. A habit of excessive donut eating in response to real interpersonal or mental health struggles will create new problems. But this does not mean that donuts caused the initial unhappiness.

The donut metaphor is also useful because it is just one food on a menu of options that vary in healthfulness. Social media use is just one way of occupying time from a menu of media options that occupy a lot of people's leisure time [54]. Even if we believe that social media are harmful, is it any more harmful than TV as a source of amusement or distraction? Probably not [55-57]. This comparison is especially relevant as TikTok and YouTube become more popular and as more people stream more TV content through social media.

The benefits or harms of social media depend on what a person needs, the choices a person has, what their media diet is, and the degree to which their habit is chronic and excessive [33,60]. To be clear, social media use is probably more harmful to some people than others [6,103], such as those with poor executive functioning, those with poor social and emotional resources, and those for whom social media are the only good option for finding connection and meaning [52,58,60]. Personality does not seem to be a source of this moderation, but global well-being matters quite a bit [57,58]. For example, a large study of social media use among late adolescents found that individuals with higher depression and lower life satisfaction tended to have a more negative experience with social media [58]. Lonely people experience more loneliness on days with more social media use, while well-connected people tend to feel more connected on days of more social media use [53]. Beyond these preexisting vulnerabilities, social media are a considerable source of distraction for most adolescents, particularly among those with low executive functioning [108].

The donut metaphor directs future research to develop a better understanding of *when* and *for whom* social media have negative effects. It also cautions against researchers (or the public) assuming that a donut is not food for someone who is hungry, or social media are not a source of amusement and connection for someone who is alone. Conversations about hyper-palatable foods, such as donuts, are instructive. We may be quick to criticize people who eat chips and processed foods or drink soda, but we must keep in mind that those foods may be the cheapest and most accessible form of nutrition a person has (eg, they are living in food deserts). Although it is easy, and may feel morally righteous, to tell a person who is depressed or lonely that they should not use social media to cope, we must consider what other forms of social nutrition that person has available to them. This perspective suggests that research should study ways to help individuals—as well as parents and teachers—build better coping mechanisms and ways to manage their media habits [109]. Blaming social media distracts policy makers and stakeholders from focusing on the root causes of loneliness, mental health struggles, and a lack of life satisfaction or meaning [1].

Finally, the donut metaphor applies to the benefits of social media as well as its harms. It makes sense that donuts made from higher-quality content are probably healthier (and more delicious). Social media content changes depending on whom someone is communicating with or watching. However, researchers do not know if the benefits of using social media are due to predisposition or existing social resources rather than content. Are there specific patterns of social media use that would be beneficial for everyone, or are positive social media habits nontransferrable between people? Although intuitive, encouraging more active and less passive use does not seem to be a solution [95]. Establishing the nature of beneficial social media use might provide clarity about how to formulate interventions and promote healthy mobile and social media habits.

In the bigger picture, we must accept that social media use is something that humans created and continually recreate. The collapse of once-popular social media platforms is a good reminder that no platform is permanent. For the user, social media habits exist at the intersection of agency, preferences, predisposition, context, and leisure. Rather than approaching social media as something that should be forbidden if it is not beneficial, it is better to consider social media as a choice of how people use their leisure time. For most people, spending time using social media is neither beneficial nor harmful—it is

XSL•FO
RenderX

merely a small piece of a bigger whole. Although it is not a great idea to make a compulsive habit out of social media (or donuts) and doing so will not solve preexisting challenges or vulnerabilities, social media do not cause the harms purported in the 10 myths either.

Conflicts of Interest

None declared.

References

1. Orben A. The Sisyphean cycle of technology panics. Perspect Psychol Sci 2020 Sep;15(5):1143-1157. [doi: 10.1177/1745691620919372] [Medline: 32603635]
2. Stoycheff E, Liu J, Wibowo K, Nanni D. What have we learned about social media by studying Facebook? A decade in review. New Media & Society 2017 Mar 01;19(6):968-980 [FREE Full text] [doi: 10.1177/1461444817695745]
3. Domahidi E. The associations between online media use and users? perceived social resources: a meta-analysis. J Comp-Med Comm 2018;23:181-200 [FREE Full text] [doi: 10.1093/jcmc/zmy007]
4. Hancock J, Liu S, Luo M, Mieczkowski H. Social media and psychological well-being. In: The Psychology of Technology: Social Science Research in the Age of Big Data. New York, NY: American Psychological Association; 2022:195-238.
5. Benes R, Keating C, Lis J, Orozco O. Time Spent With Connected Devices 2023. Santa Clara, CA: Intel Inc; 2023.
6. Meier A, Reinecke L. Computer-mediated communication, social media, and mental health: a conceptual and empirical meta-review. Communication Research 2020 Oct 21;48(8):1182-1209. [doi: 10.1177/0093650220958224]
7. Orben A, Przybylski A. The association between adolescent well-being and digital technology use. Nat Hum Behav 2019 Feb;3(2):173-182 [FREE Full text] [doi: 10.1038/s41562-018-0506-1] [Medline: 30944443]
8. Orben A, Przybylski A. Screens, teens, and psychological well-being: evidence from three time-use-diary studies. Psychol Sci 2019 May;30(5):682-696 [FREE Full text] [doi: 10.1177/0956797619830329] [Medline: 30939250]
9. Dissing A, Hulvej Rod N, Gerds T, Lund R. Smartphone interactions and mental well-being in young adults: a longitudinal study based on objective high-resolution smartphone data. Scand J Public Health 2021 May;49(3):325-332. [doi: 10.1177/1403494820920418] [Medline: 32536319]
10. Radtke T, Apel T, Schenkel K, Keller J, von Lindern E. Digital detox: an effective solution in the smartphone era? A systematic literature review. Mobile Media & Communication 2021 Jul 15;10(2):190-215 [FREE Full text] [doi: 10.1177/20501579211028647]
11. Herriman Z, Taylor AM, Roberts RM. Interventions to reduce the negative impact of online highly visual social networking site use on mental health outcomes: a scoping review. Psychology of Popular Media 2024 Jan;13(1):111-139. [doi: 10.1037/ppm0000455]
12. Allcott H, Braghieri L, Eichmeyer S, Gentzkow M. The welfare effects of social media. Ame Econ Rev 2020 Mar;110(3):629-676 [FREE Full text] [doi: 10.1257/aer.20190658]
13. Hall J, Xing C, Ross E, Johnson R. Experimentally manipulating social media abstinence: results of a four-week diary study. Media Psychology 2019 Nov 05;24(2):259-275. [doi: 10.1080/15213269.2019.1688171]
14. Stieger S, Lewetz D. A week without using social media: results from an ecological momentary intervention study using smartphones. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):618-624. [doi: 10.1089/cyber.2018.0070] [Medline: 30334650]
15. Przybylski A, Orben A, Weinstein N. How much is too much? Examining the relationship between digital screen engagement and psychosocial functioning in a confirmatory cohort study. J Am Acad Child Adolesc Psychiatry 2020 Sep;59(9):1080-1088 [FREE Full text] [doi: 10.1016/j.jaac.2019.06.017] [Medline: 31400437]
16. García-Manglano J, Fernández A, Serrano C, López-Madrigal C, Fernández-Duval G, de la Rosa Fernández-Pacheco P, et al. Social media and mental health: the role of interpersonal relationships and social media use motivations, in a nationally representative, longitudinal sample of Spanish emerging adults. Journal of Social and Personal Relationships 2024 Mar 26;41(5):1157-1182 [FREE Full text] [doi: 10.1177/02654075241230248]
17. Steele R, Khetawat D, Christofferson J, Hall J. Concurrent validity of self-reported social media use in adolescents and young adults: associations with objective data and psychosocial functioning. J Psychopathol Behav Assess 2022 Dec 29;45(1):97-108 [FREE Full text] [doi: 10.1007/s10862-022-10013-9]
18. Marciano L, Driver C, Schulz P, Camerini A. Dynamics of adolescents' smartphone use and well-being are positive but ephemeral. Sci Rep 2022 Jan 25;12(1):1316 [FREE Full text] [doi: 10.1038/s41598-022-05291-y] [Medline: 35079056]
19. Deng T, Kanthawala S, Meng J, Peng W, Kononova A, Hao Q, et al. Measuring smartphone usage and task switching with log tracking and self-reports. Mobile Media & Communication 2018 Apr 25;7(1):3-23. [doi: 10.1177/2050157918761491]
20. Oppenheimer S, Bond L, Smith C. Social media does not elicit a physiological stress response as measured by heart rate and salivary cortisol over 20-minute sessions of cell phone use. PLoS One 2024;19(4):e0298553. [doi: 10.1371/journal.pone.0298553] [Medline: 38568926]
21. Vanman E, Baker R, Tobin S. The burden of online friends: the effects of giving up Facebook on stress and well-being. J Soc Psychol 2018;158(4):496-507 [FREE Full text] [doi: 10.1080/00224545.2018.1453467] [Medline: 29558267]

XSL•FO
RenderX

22. Tomczyk S, Hoferichter F. Associations between social media use, psychological stress, well-being, and alpha-amylase levels in adolescents. Journal of Stress, Anxiety, Trauma, and Resilience 2022;1(2):26-37 [FREE Full text]
23. Pedersen J, Rasmussen M, Sørensen SO, Mortensen S, Olesen L, Brage S, et al. Effects of limiting digital screen use on well-being, mood, and biomarkers of stress in adults. Npj Ment Health Res 2022;1(1):14 [FREE Full text] [doi: 10.1038/s44184-022-00015-6] [Medline: 37521498]
24. Lee D, Jiang T, Crocker J, Way B. Can inflammation predict social media use? Linking a biological marker of systemic inflammation with social media use among college students and middle-aged adults. Brain Behav Immun 2023 Aug;112:1-10. [doi: 10.1016/j.bbi.2023.05.010] [Medline: 37224891]
25. Beyens I, Pouwels J, van Driel II, Keijsers L, Valkenburg PM. The effect of social media on well-being differs from adolescent to adolescent. Sci Rep 2020 Jul 01;10(1):10763 [FREE Full text] [doi: 10.1038/s41598-020-67727-7] [Medline: 32612108]
26. Reviews.org. URL: https://www.reviews.org/mobile/cell-phone-addiction/ [accessed 2023-11-02]
27. Pies R. Should DSM-V designate "internet addiction" a mental disorder? Psychiatry (Edgmont) 2009 Feb;6(2):31-37 [FREE Full text] [Medline: 19724746]
28. Valkenburg P. Social media use and well-being: what we know and what we need to know. Curr Opin Psychol 2022 Jun;45:101294 [FREE Full text] [doi: 10.1016/j.copsyc.2021.12.006] [Medline: 35016087]
29. Tokunaga R, Rains S. A review and meta-analysis examining conceptual and operational definitions of problematic internet use. Hum Commun Res 2016 Mar 01;42(2):165-199 [FREE Full text] [doi: 10.1111/hcre.12075]
30. Panova T, Carbonell X. Is smartphone addiction really an addiction? J Behav Addict 2018 Jun 01;7(2):252-259 [FREE Full text] [doi: 10.1556/2006.7.2018.49] [Medline: 29895183]
31. Cunningham S, Hudson C, Harkness K. Social media and depression symptoms: a meta-analysis. Res Child Adolesc Psychopathol 2021 Feb;49(2):241-253. [doi: 10.1007/s10802-020-00715-7] [Medline: 33404948]
32. Digital Wellness Lab. URL: https://digitalwellnesslab.org/cimaid/ [accessed 2023-11-02]
33. Vanden Abeele MMP, Halfmann A, Lee E. Drug, demon, or donut? Theorizing the relationship between social media use, digital well-being and digital disconnection. Curr Opin Psychol 2022 Jun;45:101295. [doi: 10.1016/j.copsyc.2021.12.007] [Medline: 35123383]
34. Orben A, Przybylski A, Blakemore S, Kievit R. Windows of developmental sensitivity to social media. Nat Commun 2022 Mar 28;13(1):1649 [FREE Full text] [doi: 10.1038/s41467-022-29296-3] [Medline: 35347142]
35. Heffer T, Good M, Daly O, MacDonell E, Willoughby T. The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: an empirical reply to Twenge et al. (2018). Clinical Psychological Science 2019 Jan 29;7(3):462-470 [FREE Full text] [doi: 10.1177/2167702618812727]
36. Lawrence D, Hunter S, Cunneen R, Houghton S, Zadow C, Rosenberg M, et al. Reciprocal relationships between trajectories of loneliness and screen media use during adolescence. J Child Fam Stud 2021 Aug 12;31(5):1306-1317. [doi: 10.1007/s10826-021-02066-3]
37. Jarman H, McLean S, Paxton S, Sibley C, Marques M. Examination of the temporal sequence between social media use and well-being in a representative sample of adults. Soc Psychiatry Psychiatr Epidemiol 2023 Aug;58(8):1247-1258 [FREE Full text] [doi: 10.1007/s00127-022-02363-2] [Medline: 36121488]
38. Stavrova O, Denissen J. Does using social media jeopardize well-being? The importance of separating within- from between-person effects. Social Psychological and Personality Science 2020 Aug 03;12(6):964-973. [doi: 10.1177/1948550620944304]
39. Orben A, Dienlin T, Przybylski AK. Social media's enduring effect on adolescent life satisfaction. Proc Natl Acad Sci U S A 2019 May 21;116(21):10226-10228. [doi: 10.1073/pnas.1902058116] [Medline: 31061122]
40. Hall JA. Relating Through Technology. New York, NY: Cambridge University Press; 2020.
41. Hall JA, Woszidlo A. Families communicating through technology. In: Handbook of Family Communication. New York, NY: Routledge; Dec 07, 2021:407-442.
42. Scott R, Zimmer-Gembeck MJ, Gardner A, Hawes T, Modecki K, Duffy A, et al. Daily use of digital technologies to feel better: adolescents' digital emotion regulation, emotions, loneliness, and recovery, considering prior emotional problems. J Adolesc 2024 Apr;96(3):539-550. [doi: 10.1002/jad.12259] [Medline: 37811912]
43. Hall J, Pennington N, Merolla A. Which mediated social interactions satisfy the need to belong? J Comp-Med Comm 2022 Nov 12;28(1):zmac026. [doi: 10.1093/jcmc/zmac026]
44. Thomas V, Balzer Carr B, Azmitia M, Whittaker S. Alone and online: understanding the relationships between social media, solitude, and psychological adjustment. Psychology of Popular Media 2021 Apr;10(2):201-211. [doi: 10.1037/ppm0000287]
45. Wang Z, Tchernev J, Solloway T. A dynamic longitudinal examination of social media use, needs, and gratifications among college students. Computers in Human Behavior 2012 Sep;28(5):1829-1839 [FREE Full text] [doi: 10.1016/j.chb.2012.05.001]
46. Bonsaksen T, Ruffolo M, Price D, Leung J, Thygesen H, Lamph G, et al. Associations between social media use and loneliness in a cross-national population: do motives for social media use matter? Health Psychol Behav Med 2023;11(1):2158089 [FREE Full text] [doi: 10.1080/21642850.2022.2158089] [Medline: 36618890]

XSL•FO
RenderX

47. Jütte D, Hennig-Thurau T, Cziehso G, Sattler H. When the antidote is the poison: investigating the relationship between people's social media usage and loneliness when face-to-face communication is restricted. PLoS One 2024;19(2):e0296423 [FREE Full text] [doi: 10.1371/journal.pone.0296423] [Medline: 38335211]
48. George M, Russell M, Piontak J, Odgers C. Concurrent and subsequent associations between daily digital technology use and high-risk adolescents' mental health symptoms. Child Dev 2018 Jan;89(1):78-88 [FREE Full text] [doi: 10.1111/cdev.12819] [Medline: 28466466]
49. Shensa A, Sidani J, Escobar-Viera C, Switzer G, Primack B, Choukas-Bradley S. Emotional support from social media and face-to-face relationships: associations with depression risk among young adults. J Affect Disord 2020 Jan 01;260:38-44 [FREE Full text] [doi: 10.1016/j.jad.2019.08.092] [Medline: 31493637]
50. Panayiotou M, Black L, Carmichael-Murphy P, Qualter P, Humphrey N. Time spent on social media among the least influential factors in adolescent mental health: preliminary results from a panel network analysis. Nat. Mental Health 2023 May 08;1(5):316-326 [FREE Full text] [doi: 10.1038/s44220-023-00063-7]
51. Magis-Weinberg L, Ballonoff Suleiman A, Dahl R. Context, development, and digital media: implications for very young adolescents in LMICs. Front Psychol 2021;12:632713 [FREE Full text] [doi: 10.3389/fpsyg.2021.632713] [Medline: 33967899]
52. George M, Jensen M, Russell M, Gassman-Pines A, Copeland W, Hoyle R, et al. Young adolescents' digital technology use, perceived impairments, and well-being in a representative sample. J Pediatr 2020 Apr;219:180-187 [FREE Full text] [doi: 10.1016/j.jpeds.2019.12.002] [Medline: 32057438]
53. Hall JA, Pennington N. Three tests of compensation by mobile and social media: how change in social resources over time moderates the effect of use on well-being. 2025 Presented at: International Communication Association; June 12-16, 2025; Denver, CO.
54. Rinderknecht R, Negraia D, Lohmann S, Zagheni E. Understanding the growth of solitary leisure in the U.S. 1965-2018 (MPIDR Working Paper WP 2023-025). Max Plank Institute. 2023. URL: https://www.demogr.mpg.de/papers/working/wp-2023-025.pdf [accessed 2024-11-19]
55. Johannes N, Nguyen T, Weinstein N, Przybylski A. Objective, subjective, and accurate reporting of social media use: no evidence that daily social media use correlates with personality traits, motivational states, or well-being. Technology, Mind, and Behavior 2021 Aug 12;2(2):1-14 [FREE Full text] [doi: 10.1037/tmb0000035]
56. Johannes N, Dienlin T, Bakhshi H, Przybylski A. No effect of different types of media on well-being. Sci Rep 2022 Jan 06;12(1):61 [FREE Full text] [doi: 10.1038/s41598-021-03218-7] [Medline: 34992220]
57. Kuper N, Kroencke L, Harari G, Denissen J. Who benefits from which activity? On the relations between personality traits, leisure activities, and well-being. J Pers Soc Psychol 2023 Jul;125(1):141-172. [doi: 10.1037/pspp0000438] [Medline: 36326676]
58. Vaid S, Kroencke L, Roshanaei M, Talaifar S, Hancock J, Back M, et al. Variation in social media sensitivity across people and contexts. Sci Rep 2024 Mar 19;14(1):6571 [FREE Full text] [doi: 10.1038/s41598-024-55064-y] [Medline: 38503817]
59. Liu M, Kamper-DeMarco K, Zhang J, Xiao J, Dong D, Xue P. Time spent on social media and risk of depression in adolescents: a dose-response meta-analysis. Int J Environ Res Public Health 2022 Apr 24;19(9):5164 [FREE Full text] [doi: 10.3390/ijerph19095164] [Medline: 35564559]
60. Vanden AM. Digital wellbeing as a dynamic construct. Comm Theory 2021;31(4):932-955. [doi: 10.1093/ct/qtaa024]
61. Bekalu M, Sato T, Viswanath K. Conceptualizing and measuring social media use in health and well-being studies: systematic review. J Med Internet Res 2023 May 10;25:e43191 [FREE Full text] [doi: 10.2196/43191] [Medline: 37163319]
62. Bruce L, Wu J, Lustig S, Russell D, Nemecek D. Loneliness in the United States: a 2018 National Panel Survey of demographic, structural, cognitive, and behavioral characteristics. Am J Health Promot 2019 Nov;33(8):1123-1133 [FREE Full text] [doi: 10.1177/0890117119856551] [Medline: 31203639]
63. Dienlin T, Johannes N. The impact of digital technology use on adolescent well-being . Dialogues in Clinical Neuroscience 2022 Apr 01;22(2):135-142 [FREE Full text] [doi: 10.31887/dcns.2020.22.2/tdienlin]
64. Atalay E. A twenty-first century of solitude? Time alone and together in the United States: Working papers of the Federal Reserve Bank, PA. Federal Reserve Bank of Philadelphia. 2022. URL: https://www.philadelphiafed.org/-/media/frbp/assets/working-papers/2022/wp22-11.pdf; [accessed 2022-05-11]
65. Hall J, Liu D. Social media use, social displacement, and well-being. Curr Opin Psychol 2022 Aug;46:101339. [doi: 10.1016/j.copsyc.2022.101339] [Medline: 35395533]
66. Twenge J, Spitzberg B, Campbell W. Less in-person social interaction with peers among U.S. adolescents in the 21st century and links to loneliness. Journal of Social and Personal Relationships 2019 Mar 19;36(6):1892-1913 [FREE Full text] [doi: 10.1177/0265407519836170]
67. Coyne SM, Rogers AA, Zurcher JD, Stockdale L, Booth M. Does time spent using social media impact mental health?: An eight year longitudinal study. Computers in Human Behavior 2020 Mar;104:106160 [FREE Full text] [doi: 10.1016/j.chb.2019.106160]
68. Schemer C, Masur P, Geiß S, Müller P, Schäfer S. The impact of internet and social media use on well-being: a longitudinal analysis of adolescents across nine years. J Comp-Med Comm 2021;28:1-21. [doi: 10.1093/jcmc/zmaa]

XSL•FO
RenderX

69. Buecker S, Mund M, Chwastek S, Sostmann M, Luhmann M. Is loneliness in emerging adults increasing over time? A preregistered cross-temporal meta-analysis and systematic review. Psychol Bull 2021 Aug;147(8):787-805. [doi: 10.1037/bul0000332] [Medline: 34898234]
70. Vuorre M, Przybylski A. Estimating the association between Facebook adoption and well-being in 72 countries. R Soc Open Sci 2023 Aug;10(8):221451-221422 [FREE Full text] [doi: 10.1098/rsos.221451] [Medline: 37564066]
71. Hall J, Kearney M, Xing C. Two tests of social displacement through social media use. Information, Communication & Society 2018 Feb 02;22(10):1396-1413 [FREE Full text] [doi: 10.1080/1369118x.2018.1430162]
72. Hampton K. Social media and change in psychological distress over time: the role of social causation. J Comp-Med Comm 2019;24(5):205-222. [doi: 10.1093/jcmc/zmz010]
73. Hall J, Johnson R, Ross E. Where does the time go? An experimental test of what social media displaces and displaced activities' associations with affective well-being and quality of day. New Media & Society 2018 Oct 23;21(3):674-692 [FREE Full text] [doi: 10.1177/1461444818804775]
74. Twenge J, Martin G, Spitzberg B. rends in U.S. Adolescents' media use, 1976–2016: the rise of digital media, the decline of TV, and the (near) demise of print. Psychology of Popular Media Culture 2019 Oct;8(4):329-345 [FREE Full text] [doi: 10.1037/ppm0000203]
75. Vilhelmson B, Thulin E, Elldér E. Where does time spent on the Internet come from? Tracing the influence of information and communications technology use on daily activities. Information, Communication & Society 2016 Mar 27;20(2):250-263 [FREE Full text] [doi: 10.1080/1369118x.2016.1164741]
76. Rideout V, Peebles A, Mann S, Robb M. Common Sense census: media use by tweens and teens, 2021. Common Sense Media. 2021. URL: https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2021 [accessed 2021-04-11]
77. Baym N, Wagman K, Persaud C. Mindfully scrolling: rethinking Facebook after time deactivated. Social Media + Society 2020 May 13;6(2):6 [FREE Full text] [doi: 10.1177/2056305120919105]
78. Przybylski A, Nguyen T, Law W, Weinstein N. Does taking a short break from social media have a positive effect on well-being? Evidence from three preregistered field experiments. J Technol Behav Sci 2021 Feb 07;6(3):507-514 [FREE Full text] [doi: 10.1007/s41347-020-00189-w]
79. Requena F, Ayuso L. Individualism or complementarity? The effect of digital personal networks on face-to-face personal networks. Information, Communication & Society 2018 Jun 05;22(14):2097-2111 [FREE Full text] [doi: 10.1080/1369118x.2018.1477968]
80. Stepanikova I, Nie N, He X. Time on the internet at home, loneliness, and life satisfaction: evidence from panel time-diary data. Computers in Human Behavior 2010 May;26(3):329-338 [FREE Full text] [doi: 10.1016/j.chb.2009.11.002]
81. Pouwels J, Valkenburg P, Beyens I, van Driel II, Keijsers L. Social media use and friendship closeness in adolescents' daily lives: an experience sampling study. Dev Psychol 2021 Feb;57(2):309-323. [doi: 10.1037/dev0001148] [Medline: 33539132]
82. Dienlin T, Masur P, Trepte S. Reinforcement or displacement? The reciprocity of FtF, IM, and SNS communication and their effects on loneliness and life satisfaction. J Comput-Mediat Comm 2017 Feb 16;22(2):71-87 [FREE Full text] [doi: 10.1111/jcc4.12183]
83. Steinsbekk S, Bjørklund O, Valkenburg P, Nesi J, Wichstrøm L. The new social landscape: relationships among social media use, social skills, and offline friendships from age 10-18 years. Comput Human Behav 2024 Jul;156:108235. [doi: 10.1016/j.chb.2024.108235] [Medline: 39513051]
84. Verduyn P, Schulte-Strathaus J, Kross E, Hülsheger U. When do smartphones displace face-to-face interactions and what to do about it? Computers in Human Behavior 2021 Jan;114(6):106550-106632. [doi: 10.1016/j.chb.2020.106550] [Medline: 29786478]
85. Courtright J, Caplan S. A meta-analysis of mobile phone use and presence. Hum Comm & Tech 2020;1:20-35. [doi: 10.17161/hct.v1i2.13412]
86. Halpern D, Katz J. Texting's consequences for romantic relationships: a cross-lagged analysis highlights its risks. Computers in Human Behavior 2017 Jun;71:386-394 [FREE Full text] [doi: 10.1016/j.chb.2017.01.051]
87. McDaniel B, Drouin M. Daily technology interruptions and emotional and relational well-being. Comput Human Behav 2019 Oct;99:1-8 [FREE Full text] [doi: 10.1016/j.chb.2019.04.027] [Medline: 32831467]
88. Pouwels J, Valkenburg P, Beyens I, van Driel II, Keijsers L. Some socially poor but also some socially rich adolescents feel closer to their friends after using social media. Sci Rep 2021 Oct 27;11(1):21176. [doi: 10.1038/s41598-021-99034-0] [Medline: 34707197]
89. Achterhof R, Kirtley O, Schneider M, Hagemann N, Hermans K, Hiekkaranta A, et al. Adolescents' real-time social and affective experiences of online and face-to-face interactions. Computers in Human Behavior 2022 Apr;129:107159 [FREE Full text] [doi: 10.1016/j.chb.2021.107159]
90. Vanden Abeele M, Hendrickson A, Pollmann M, Ling R. Phubbing behavior in conversations and its relation to perceived conversation intimacy and distraction: an exploratory observation study. Computers in Human Behavior 2019 Nov;100:35-47 [FREE Full text] [doi: 10.1016/j.chb.2019.06.004]
91. Brown B, O'hara K, Mcgregor M, Mcmillan D. Text in talk. ACM Trans Comput-Hum Interact 2018 Jan 22;24(6):1-25 [FREE Full text] [doi: 10.1145/3152419]

XSL•FO
RenderX

92. Vanden Abeele M, Antheunis M, Schouten A. The effect of mobile messaging during a conversation on impression formation and interaction quality. Computers in Human Behavior 2016 Sep;62:562-569 [FREE Full text] [doi: 10.1016/j.chb.2016.04.005]
93. Liu D, Ainsworth S, Baumeister R. A meta-analysis of social networking online and social capital. Review of General Psychology 2016 Dec 01;20(4):369-391 [FREE Full text] [doi: 10.1037/gpr0000091]
94. Ahmad Z, Soroya S, Mahmood K. Bridging social capital through the use of social networking sites: a systematic literature review. Journal of Human Behavior in the Social Environment 2022 Apr 20;33(4):473-489 [FREE Full text] [doi: 10.1080/10911359.2022.2064025]
95. Godard R, Holtzman S. Are active and passive social media use related to mental health, wellbeing, and social support outcomes? A meta-analysis of 141 studies. J Comp Med Comm 2024;29(1):1-14. [doi: 10.1093/jcmc/zmad055]
96. Wenninger H, Krasnova H, Buxmann P. Understanding the role of social networking sites in the subjective well-being of users: a diary study. European Journal of Information Systems 2018 Aug 02;28(2):126-148 [FREE Full text] [doi: 10.1080/0960085x.2018.1496883]
97. Kahlow J, Coker M, Richards R. The multimodal nature of Snapchat in close relationships: toward a social presence-based theoretical framework. Computers in Human Behavior 2020 Oct;111:106409 [FREE Full text] [doi: 10.1016/j.chb.2020.106409]
98. Ruppel E, Burke T, Cherney M. Channel complementarity and multiplexity in long-distance friends' patterns of communication technology use. New Media & Society 2017 Mar 23;20(4):1564-1579 [FREE Full text] [doi: 10.1177/1461444817699995]
99. Holtzman S, Kushlev K, Wozny A, Godard R. Long-distance texting: text messaging is linked with higher relationship satisfaction in long-distance relationships. J Soc Pers Relat 2021 Dec;38(12):3543-3565 [FREE Full text] [doi: 10.1177/02654075211043296] [Medline: 34924671]
100. Sharabi, Roaché D, Pusateri K. Texting toward intimacy: relational quality, length, and motivations in textual relationships. Communication Studies 2019 Jul 01;70(5):601-619 [FREE Full text] [doi: 10.1080/10510974.2019.1634117]
101. Dyer J, Coyne S, Gale M, Sheppard JA. A person-centered approach to understanding the long-term relationship between social media usedepression across adolescence. J Adol 2025:1 (forthcoming).
102. Raney AA, Janicke-Bowles SH, Oliver MB, Dale KR. Introduction to Positive Media Psychology. New York, NY: Routledge; Dec 29, 2020.
103. Vanden AM, Vandebosch H, Koster E, De Leyn T, Gaeveren KV, van TT, et al. Why, how, when, and for whom does digital disconnection work? A process-based framework of digital disconnection. Communication Theory 2024 Feb 1;34(1):3-17 [FREE Full text] [doi: 10.1093/ct/qtad016]
104. Lee A, Hancock J. Social media mindsets: a new approach to understanding social media use and psychological well-being. J Comp-Med Comm 2024;29(1):1-24. [doi: 10.1093/jcmc/zmad048]
105. de Segovia Vicente D, Van Gaeveren GK, Murphy S, Vanden Abelee M. Does mindless scrolling hamper well-being? Combining ESM and log-data to examine the link between mindless scrolling, goal conflict, guilt, and daily well-being. J Comp-Med Comm 2024;29(1):1-21. [doi: 10.1093/jcmc/zmad056]
106. Prinstein M, Nesi J, Telzer E. J Child Psychol Psychiatry 2020 Mar;61(3):349-352. [doi: 10.1111/jcpp.13219] [Medline: 32064633]
107. Odgers CL. The great rewiring: is social media really behind an epidemic of teenage mental illness? Nature 2024 Mar 29;628(8006):29-30 [FREE Full text] [doi: 10.1038/d41586-024-00902-2]
108. Siebers T, Beyens I, Pouwels J, Valkenburg P. Explaining variation in adolescents' social media-related distraction: the role of social connectivity and disconnectivity factors. Curr Psychol 2022 Nov 24;42(34):1-14 [FREE Full text] [doi: 10.1007/s12144-022-03844-y] [Medline: 36468162]
109. Weinstein E, James C. Behind Their Screens. Boston, MA: MIT Press; Aug 16, 2022.

Abbreviations

DSM-5*: *Diagnostic and Statistical Manual of Mental Disorders, fifth edition

Edited by G Eysenbach; submitted 16.04.24; peer-reviewed by M Vanden Abeele, K Burnell; comments to author 17.07.24; revised version received 30.07.24; accepted 03.10.24; published 25.11.24.

Please cite as:
Hall JA
Ten Myths About the Effect of Social Media Use on Well-Being
J Med Internet Res 2024;26:e59585
URL: https://www.jmir.org/2024/1/e59585
doi:10.2196/59585
PMID:

XSL•FO
RenderX

©Jeffrey A Hall. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Telemonitoring for Chronic Heart Failure: Narrative Review of the 20-Year Journey From Concept to Standard Care in Germany

Sebastian Spethmann[1,2,3], MD; Gerhard Hindricks[1,2], MD; Kerstin Koehler[1,2], MD; Stefan Stoerk[4], MD; Christiane E Angermann[5,6], MD; Michael Böhm[7], MD; Birgit Assmus[8,9], MD; Sebastian Winkler[10], MD; Martin Möckel[2,11], MD; Mirja Mittermaier[2,12], MD; Monika Lelgemann[13], MD, MSc; Daniel Reuter[13]; Ralph Bosch[14], MD; Alexander Albrecht[15], MD; Stephan von Haehling[16,17], MD; Thomas M Helms[18,19], MD; Stefan Sack[20], MD; Tarek Bekfani[21], MD; Jan Wolfgang Gröschel[1,22,23], MD; Magdalena Koehler[24,25]; Christoph Melzer[2], MD; Jan Wintrich[7], MD; Bettina Zippel-Schultz[18]; Georg Ertl[6], MD; Claus Vogelmeier[26], MD; Nikolaos Dagres[1,2], MD; Jasmin Zernikow[1,2*], MD; Friedrich Koehler[1,2*], MD

[1]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité (DHZC), Berlin, Germany

[2]Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[3]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Germany

[4]Department Clinical Research & Epidemiology, Comprehensive Heart Failure Centre, Department of Internal Medicine I, University Hospital Würzburg, Würzburg, Germany

[5]Comprehensive Heart Failure Centre, University Hospital Würzburg, Würzburg, Germany

[6]Department for Internal Medicine I, University Hospital Würzburg, Würzburg, Germany

[7]Department of Internal Medicine II, Cardiology, Marien Hospital Herne, Ruhr University Bochum, Bochum, Germany., Herne, Germany

[8]Department of Cardiology and Angiology, University of Giessen, Giessen, Germany

[9]DZHK (German Centre for Cardiovascular Research), Partner Site Rhein/Main, Kerckhoff Heart Center, Bad Nauheim, Germany

[10]Department of Internal Medicine and Cardiology, BG Klinikum Unfallkrankenhaus Berlin, Berlin, Germany

[11]Department of Emergency and Acute Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany

[12]Department of Infectious Diseases, Respiratory Medicine and Critical Care, Charité-Universitätsmedizin Berlin, Berlin, Germany

[13]Gemeinsamer Bundesausschuss, Berlin, Germany

[14]Cardio Centrum Ludwigsburg-Bietigheim, Ludwigsburg, Germany

[15]Kardiologische Gemeinschaftspraxis Wilmersdorfer Strasse 62, Vivantes Klinikum Am Urban, Berlin, Germany

[16]Department of Cardiology and Pneumology, University of Göttingen Medical Center, Göttingen, Germany

[17]German Center for Cardiovascular Research (DZHK), Partner Site Göttingen, Göttingen, Germany

[18]German Foundation for the Chronically Ill, Berlin, Germany

[19]Peri Cor Cardiology Working Group / associated UCSF, Hamburg, Germany

[20]Department of Cardiology, Pneumology, and Internal Intensive Care Medicine, Schwabing Hospital, Academic Municipal Hospital Munich, Munich, Germany

[21]Department of Cardiology and Angiology, Universitätsklinikum Magdeburg, Magdeburg, Germany

[22]ECRC Experimental and Clinical Research Center, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[23]Working Group on Cardiovascular Magnetic Resonance, Experimental and Clinical Research Center, a joint cooperation between Charité Medical Faculty and the Max-Delbrück Center for Molecular Medicine, Berlin, Germany

[24]Department of Preventive Sports Medicine and Sports Cardiology, University Hospital 'Klinikum rechts der Isar', School of Medicine, Technical University Munich, Munich, Germany

[25]Department of Acute Medicine, Donauspital Wien, Wien, Germany

[26]Department of Internal Medicine, Pulmonary and Critical Care Medicine, Member of the German Center for Lung Research (DZL), University of Marburg, Marburg, Germany

[*]these authors contributed equally

Corresponding Author:
Sebastian Spethmann, MD
Department of Cardiology, Angiology and Intensive Care Medicine
Deutsches Herzzentrum der Charité (DHZC)
Chariteplatz 1
Berlin, 10117

XSL•FO
RenderX

Germany
Phone: 49 30450513142
Fax: 49 304507513235
Email: Sebastian.spethmann@dhzc-charite.de

Abstract

Background: Chronic heart failure (CHF) is a major cause of morbidity and mortality worldwide, placing a significant burden on health care systems. The concept of telemedicine for CHF was first introduced in the late 1990s, and since 2010, studies have demonstrated its potential to improve patient outcomes and reduce health care costs. Over the following decade, technological advancements and changes in health care policy led to the development of more sophisticated telemedicine solutions for CHF, including remote patient management through invasive or noninvasive telemonitoring devices, mobile apps, and virtual consultations. Years of public funding in Germany have generated evidence that remote patient management improves outcomes for patients with CHF, such as quality of life, and reduces hospital admissions. Based on these data, the Federal Joint Committee (Gemeinsamer Bundesausschuss; G-BA) decided, independently of the current European Society of Cardiology recommendations, to incorporate telemedicine as a standard digital intervention for high-risk patients with reduced left ventricular ejection fraction in Germany in 2020.

Objective: This review aims to illustrate the journey from the initial concept through pioneering studies that led to telemedicine's integration into standard care, and to share current experiences that have positioned Germany as a leader in cardiovascular telemedicine.

Methods: We review and discuss existing literature and evidence on the development and implementation of telemonitoring for CHF in Germany over the past 20 years. Relevant studies, reports, and guidelines were identified through a comprehensive search of electronic databases, including PubMed, Google Scholar, and specialized journals focused on CHF telemonitoring.

Results: Pioneering studies, such as the TIM-HF2 (Telemedical Interventional Management in Heart Failure II) and IN-TIME (Influence of Home Monitoring on Mortality and Morbidity in Heart Failure Patients with Impaired Left Ventricular Function) trials, demonstrated the effectiveness of remote patient management applications for patients with CHF in Germany and their applicability to current practices involving both invasive and noninvasive methods. Collaborations between researchers and technology developers overcame barriers, leading to sustainable improvements in patient care. Ongoing research on artificial intelligence applications for prioritizing and interpreting individual health data will continue to transform digital health care.

Conclusions: The establishment of telemedical care for patients with HF across Europe is likely to benefit from experiences in Germany, where significant improvements have been achieved in the care of patients with HF.

(J Med Internet Res 2024;26:e63391) doi:10.2196/63391

KEYWORDS

telemedicine; e-counseling; heart decompensation; Europe; patient care management

Introduction

The Origins of Telemedicine

The history of telemedicine is closely tied to the technical innovations of its era, beginning in the 19th century with Samuel Morse's invention of the telegraph and Morse code system. The introduction of the telephone by Philipp Reis in 1861, and its subsequent development by Alexander Graham Bell in 1876, opened new possibilities for telemedicine, enabling remote diagnosis for the first time. In 1911, radio technology was first used in maritime shipping for telemedical consultations. Since 1931, Cuxhaven Hospital in Germany has provided medical advice to seafarers via radio [1].

Over the past 150 years, advances in telecommunications have driven new telemedical applications. However, it is primarily in the last 20 years that telemedicine has been increasingly used for the outpatient care of chronically ill patients. Since the advent of smartphones in 2007, telemedicine has become part of daily life, supporting both asynchronous care—through electronic exchange of telemedical data—and real-time consultations via video or telephone conferencing.

Telemedicine thus has 2 primary application scenarios: on the one hand, it facilitates professional exchanges between geographically distant health care providers (known as "doc-to-doc telecardiology"). On the other hand, it enables a direct connection between doctors and patients within the patient's home environment, using information and communication tools (referred to as "doc-to-patient telecardiology").

In Germany, the implementation of telemonitoring was notably delayed due to restrictive legislation, including a long-standing legal ban on remote treatment, which was only eased in 2018. Although direct doctor-patient contact remains the standard practice, telemedical cocare is now an available option. However, a complete replacement of outpatient care with telemedicine remains excluded.

All other fundamental requirements—such as the personal delivery of services, adherence to specialist standards, medical

XSL•FO
RenderX

duty of care, patient confidentiality, and the obligation to inform patients of treatment risks—must also be upheld in telecardiological cotreatment and are outlined in a treatment contract. A particular requirement is that patients, most of whom lack medical training, must be able to perform simple diagnostic procedures (eg, measuring blood pressure and recording electrocardiograms [ECGs]) with adequate precision.

Telemedicine in Heart Failure: A Life Prolonging and Economic Digital Measure

Cardiac decompensation is the most common and prognostically significant complication of chronic heart failure (CHF), with an average survival time of about 2.5 years following its first occurrence [2]. Each day, approximately 1250 patients are admitted to hospitals in Germany due to cardiac decompensation [3]. Since 2005, it has been the leading single diagnosis for hospital admissions, representing a substantial economic burden on the health care system [4].

The decline in cardiac function often begins gradually and can progress to life-threatening decompensation, marked by changes in specific vital parameters. With modern sensor technology, early shifts in these vital signs can be detected remotely and transmitted immediately to the attending physician, enabling an effective doctor-to-patient telemedicine care model. The first evidence of the efficacy of home telemonitoring as a preventive strategy was provided by the 2005 TEN-HMS study (Trans-European Network-Home-Care Management System). The study demonstrated a reduction in the average hospital stay by 6 days and a significantly lower 1-year mortality rate (29%) compared with usual care (45%) in patients with CHF and a left ventricular ejection fraction (LVEF) below 40% [5].

Telemonitoring can be categorized as either invasive or noninvasive, based on the type of sensors used to record vital signs (Figure 1). In invasive telemonitoring, disease-specific vital signs are monitored using active or passive implanted devices. In noninvasive telemonitoring, patients record vital signs daily using external home measuring devices, such as scales, blood pressure monitors, external ECG devices, or wearables. Through active self-measurement, patients provide treating physicians with up-to-date information on their current health status.

Figure 1. Selected types of sensors used in telemonitoring. CIED: cardiac implantable electronic device.

Among the 6 landmark studies (Figure 2) on the efficacy of cardiovascular telemedicine, 3 were conducted in Germany, 2 of which received public funding. These studies ultimately verified telemedicine as an independent treatment method. As a direct result, the German Federal Joint Committee (Gemeinsamer Bundesausschuss; G-BA) has included telemedicine as a digital intervention for high-risk patients with heart failure (HF) with reduced left ventricular ejection fraction in the benefits catalog in Germany.

XSL•FO
RenderX

Figure 2. Highlighted trials and randomized controlled trials (RCTs) on telemonitoring from 2009 to present.

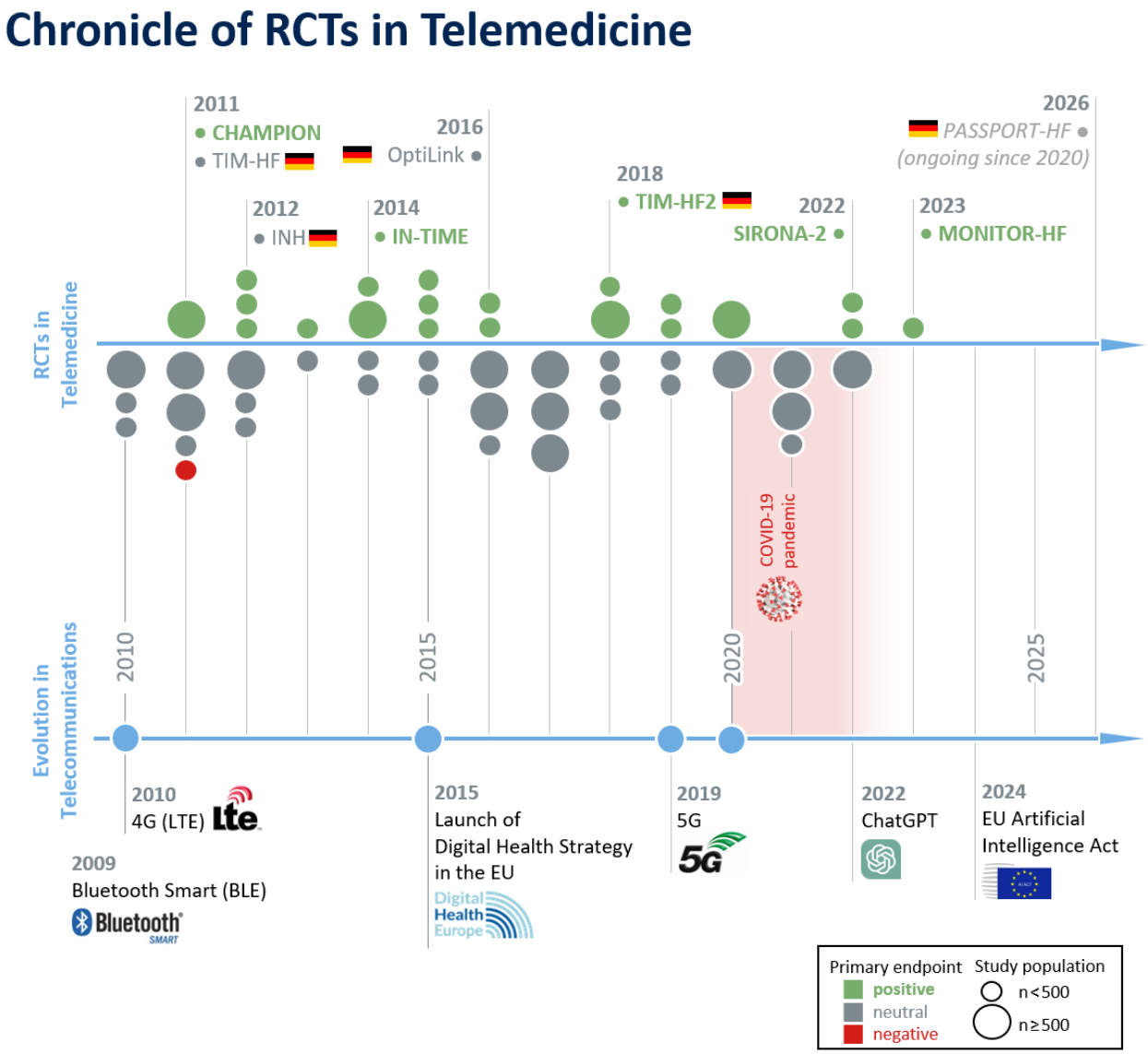

This overview presents the German perspective, tracing the journey from the initial study idea through evidence generation to implementation in standard care. It reports on current experiences in Germany and offers insights into future developments.

Methods

This narrative review synthesizes existing literature and evidence regarding the development and implementation of telemonitoring for CHF in Germany over the past 20 years. Relevant studies, reports, and guidelines were identified through a comprehensive search of electronic databases, including PubMed, Google Scholar, and specific journals focused on telemonitoring for CHF.

The selection of literature was guided by its relevance to the topic, emphasizing major studies, clinical trials, and policy papers that significantly contributed to the evolution of telemedicine in HF care. Sources were not restricted to randomized controlled trials; observational studies, expert opinions, and case studies were also included to provide a comprehensive understanding of the field.

The review process involved critically analyzing and summarizing the findings, focusing on key milestones, technological advancements, and their impact on clinical practice. Additionally, insights from ongoing and future research initiatives were incorporated to highlight emerging trends and potential future directions in telemedicine for HF care.

Results

Evidence on the Efficacy of Noninvasive Telemonitoring

Overview

Noninvasive devices, such as scales, blood pressure monitors, ECG devices, and wearables, have demonstrated benefits in multiparametric data transmission, ease of use, integrative monitoring of comorbidities (eg, diabetes), and cost-effectiveness.

XSL•FO
RenderX

Evidence for the efficacy of noninvasive cardiovascular telemonitoring first emerged in the mid-2000s, leading to a differentiation between trials based on invasive or noninvasive sensor technology. Three major landmark trials—INH (Interdisciplinary Network for Heart Failure), TIM-HF (Telemedical Interventional Monitoring in Heart Failure Study), and TIM-HF2 (Telemedical Interventional Management in Heart Failure II)—were conducted in Germany and received public funding from the German Ministry of Education and Research, as well as the Federal Ministry for Economic Affairs and Climate Protection.

The INH-Study Program (2004-2022)

The randomized, controlled, multicenter INH (Interdisciplinary Network for Heart Failure) study program (ISRCTN23325295) investigated the effects of regular telephone-based monitoring and education provided by HF nurses as part of an intensified care approach by a specialized HF center. The remote patient management strategy emphasized patient self-monitoring of vital parameters and HF symptoms. The first phase of the trial assessed the 6-month risk of mortality and morbidity in 715 patients with an LVEF of ≤40%, all of whom had been hospitalized for acute decompensated HF.

The composite primary outcome for time to all-cause death or rehospitalization was neutral between the telemonitoring group (n=352) and the usual care control group (n=363). This observation was influenced by a significant reduction in all-cause mortality to 38% in the remote patient management group, which was associated with a trending increase in hospitalization rates in the intervention arm. Additionally, patients in the telemonitoring group demonstrated improvements in quality of life, functional class, and adherence to medical therapy.

A long-term evaluation of the INH trial (the “E-INH study”) in a larger population (n=1022) demonstrated that telephone-based monitoring over an 18-month period led to a significant reduction in all-cause and cardiovascular mortality after 120 months in the group of survivors who had received remote patient management (33% vs 40%; *P*=.04). Additionally, HF-related and cardiovascular hospitalizations were significantly reduced at 18, 36, and 60 months by –25%, –29%, and –30%, respectively. As previously noted, the telemonitoring procedures did not include automated transmission of patient-recorded vital signs. Instead, they relied on postdischarge telephone monitoring, where specialized nurses reviewed patients’ self-monitored results. These nurses also provided information, education, and guidance on optimal medical therapies; advised on appropriate actions in case of suspected HF decompensation; supported self-adjustment of diuretic agents; and coordinated specialist care as needed.

Overall, the INH study program suggests that telephone-based monitoring and education alone can improve short-term quality of life and mortality, with a sustained long-term impact demonstrated by reduced all-cause and cardiovascular mortality, fewer hospitalizations, and improved quality of life [6,7].

The TIM-HF Study Program (2005-2024)

The TIM-HF study (NCT00543881) and the TIM-HF2 study (NCT01878630) were large, randomized, controlled, open, parallel telemedical trials that investigated the efficacy of noninvasive, multiparameter telemonitoring in a population with moderate to severe HF (New York Heart Association [NYHA] class II or III).

Both studies were fully publicly funded. The core element of this telemedical care model was a telemedical center staffed with experienced HF specialists and nurses, offering regular and emergency telemedical services 24 hours a day, 7 days a week, functioning as a “virtual emergency department” [8].

The TIM-HF trial was a telemedical randomized controlled trial conducted between 2008 and 2010 across 4 German federal states: Berlin, Brandenburg, Saxony-Anhalt, and Baden-Württemberg. The main inclusion criteria were an LVEF below 35% and a history of HF hospitalization within 24 months before randomization. The primary endpoint was all-cause mortality. The study concluded on a predetermined date (April 30, 2010), resulting in varying follow-up periods for each patient, with a minimum of 12 months and a maximum of 28 months (mean follow-up 21.5 months, SD 7.2 months). The results for the primary endpoint were neutral [9]. However, a post hoc analysis identified a subset of patients who may benefit most from telemonitoring [10].

TIM-HF2 was a nationwide telemedical randomized controlled trial conducted between 2013 and 2018 across 13 German federal states. The main inclusion criteria were informed by the post hoc analysis of the TIM-HF study. Patients were enrolled regardless of their LVEF status if they had experienced an HF-related hospitalization within 12 months before randomization. A total of 1538 patients were randomized equally into either a telemedical intervention arm or a control arm. The study follow-up period was 1 year. The primary endpoint was defined as “days lost due to unplanned cardiovascular hospitalization or death within 1 year.” The secondary endpoint was total mortality. Telemedical intervention showed a significant benefit in the primary endpoint, with 17.8 days lost in the telemedical group compared with 24.2 days in the control group (*P*=.046). Additionally, there was a nearly 30% reduction in 1-year mortality in the telemedical group, with a mortality rate of 7.9% per 100 patient-years (95% CI 6.14-10.10) compared with 11.3% per 100 patient-years in the control group (95% CI 9.21-13.95). This corresponded to a hazard ratio of 0.70 (95% CI 0.50-0.96; *P*=.028) [11].

The 2 subsequent TIM-HF studies offered robust evidence supporting the efficacy of noninvasive telemonitoring in a specific subgroup of patients with HF.

Evidence on Invasive Telemonitoring Using Cardiac Implantable Electronic Devices

Overview

Active implants, also known as cardiac implantable electronic devices (CIEDs), not only serve their primary function of treating cardiac arrhythmias through defibrillators or cardiac resynchronization therapy, but also provide an option for

XSL•FO
RenderX

telemonitoring. In the context of telemonitoring, these devices periodically transmit data on physical activity, submit ECG readings, and may detect changes in pulmonary impedance. However, this active home monitoring application is limited to a specific subgroup of patients with HF who meet the eligibility criteria for CIED therapy.

The IN-TIME (Influence of Home Monitoring on Mortality and Morbidity in Heart Failure Patients with Impaired Left Ventricular Function) landmark trial marked the beginning of an era characterized by technology-based, industry-funded study designs, moving away from reliance on governmental funding. This shift has paved the way for the generation of fundamental evidence and has further shaped the research landscape in telemedicine. As a result, companies providing device therapy have become involved as principal investigating parties in telemedical research.

The IN-TIME Trial (2007-2014)

The IN-TIME trial (NCT00538356) is currently the only study that has investigated device-based multiparameter telemonitoring using active CIEDs and has reported a positive primary endpoint for a composite score that includes all-cause mortality.

This international multicenter randomized trial was conducted at 36 centers across Israel (2 sites), Australia (1 site), and Europe (33 sites), including 26 sites in Germany, and was funded by BIOTRONIK SE & Co. KG. A total of 664 patients with congestive heart failure classified as NYHA class II-III, with an LVEF of less than 35%, who had undergone implantation of a dual-chamber CIED, either an implantable cardiac defibrillator (ICD) or a cardiac resynchronization therapy with a defibrillator (CRT-D), were randomized in a 1:1 ratio to either the standard care control arm or the device-based telemonitoring in addition to standard care arm.

The telemonitoring software installed in the CIEDs transmitted daily tracked data to the manufacturer's online platform. These data included not only arrhythmic events but also parameters related to the device's functionality. For the intervention group, any subsequent abnormalities were further forwarded to the study centers, where HF nurses conducted structured interviews to assess typical HF symptoms and discussed potential interventions with physicians.

With the primary endpoint defined as a "modified Packer score," which includes the incidence rate of all-cause mortality, HF-related hospitalizations, changes in NYHA class, and quality of life assessments, the intervention group receiving telemonitoring demonstrated a significantly lower occurrence rate of the composite clinical endpoint compared with the control arm (18.9% vs 27.2%, respectively; P=.013). Importantly, total mortality was significantly lower in the telemonitoring arm compared with the control arm (3.0% vs 8.2%; hazard ratio 0.36; P=.004). However, the rates of hospital admissions for worsening HF did not differ significantly between the 2 groups. In a post hoc analysis, no interactions were found between the effect of telemonitoring and device type (ICD or CRT-D), NYHA class at enrollment, or age (≤67 or >67 years), indicating a consistently significant impact of telemonitoring across all subgroups.

Thus, the IN-TIME trial demonstrated the beneficial impact of device-based telemonitoring in patients with HF through daily data transmission. Consequently, these results may not be directly transferable to remote monitoring platforms that utilize less frequent data transmissions [12].

The OptiLink-HF Trial (2009-2015)

In the invasive telemonitoring OptiLink-HF study (Optimization of Heart Failure Management using Medtronic OptiVol Fluid Status Monitoring and CareLink Network; NCT00769457), a total of 1002 patients were enrolled following implantation of an ICD or CRT-ICD [13]. These CIEDs also included the capability to measure intrathoracic impedance. In the event of a significant decrease in thoracic impedance, indicative of pulmonary fluid overload, the devices transmitted fluid threshold crossing alerts.

Patient enrollment occurred across 65 centers in Germany from October 2008 to April 2013. The mandatory inclusion criteria for participants included CHF classified as NYHA class II or III, an LVEF of less than 35%, and at least one of the following: an HF hospitalization within the past 12 months, the use of intravenous diuretic medication for outpatient treatment, or an increase in natriuretic peptide levels within the preceding 30 days.

Patients were randomized in a 1:1 ratio, with the telemedicine function either turned on or off. Physicians were required to respond within 2 working days using a prespecified intervention algorithm [14]. During these phone contacts, discussions included changes in body weight, medication adjustments, and symptoms of decongestion. The initiation of therapeutic measures was at the discretion of the physicians, allowing for a range of responses from no action to reprogramming the fluid threshold crossing alert threshold. If fluid threshold crossing alerts continued for more than 12 days, patients were prompted to attend an in-office or in-hospital visit for further evaluation.

The follow-up period for all study participants was set at 18 months, with an option for reinformed consent to extend follow-up until the last patient completed their study visit. This led to a mean follow-up duration of 22.9 months for the participants.

The primary endpoint of the study was the rate of death from any cause and the rate of cardiovascular hospitalizations. Notably, the primary composite endpoint and its individual components were not achieved. Major limitations of the study included technical issues, with 24% of the impedance alarms failing to transmit to the treating physicians. Additionally, inappropriate responses to the alerts were observed in 40% of cases, which included instances of no patient contact at all (12.4%); patient contact exceeding the 2-working-day threshold (10.8%), particularly during weekends and public holidays (4.0%); and a lack of appropriate medical intervention (32.8%).

Interestingly, a post hoc analysis of the OptiLink-HF trial suggested that impedance-based telemonitoring could have significant prognostic effects when alert transmissions were managed appropriately [15]. Additionally, the intervention was found to significantly decrease the primary endpoint in patients with preserved renal function, while it did not demonstrate the

XSL•FO
RenderX

same benefit in those with chronic kidney disease [16]. This highlights the critical importance of timely and effective management of alert transmissions, especially in these vulnerable patients [17].

Trial-Based Evolution of Invasive Hemodynamic Telemonitoring

Overview

The alterations in intracardiac filling pressures, which serve as the ideal surrogate for incipient congestion, are not well reflected by changes in body weight. Invasive hemodynamic telemonitoring provides the advantage of earlier detection of impending cardiac decompensation events [18]. However, regardless of the underlying technical devices used, reliable and prompt integration into the workflows of telemedical centers is considered crucial for achieving therapeutic effectiveness.

Intracardiac filling pressures can be monitored by tracking the pulmonary artery pressure (PAP), right ventricle, or left atrium. These implants enable the transmission of tracked values in a "passive" manner upon query with a wireless external readout device.

The CardioMEMS HF System

The electrode- and battery-free PAP sensor CardioMEMS (Abbott Laboratories) is implanted in a proximal branch of the pulmonary artery and calibrated to the individual patient. Changes in pulsatile flow are converted into a pressure derivative and wirelessly transmitted to a remote dashboard, such as that at a telemedical center. Daily measurements must be initiated actively by the patient, and specialized staff regularly interrogate the PAP trends accumulating over time to determine whether and which actions need to be taken. In the Abbott-funded CHAMPION trial (NCT00531661) [19], conducted in the United States and Canada, 550 patients in NYHA functional class III, regardless of LVEF, had PAP sensors implanted. The CHAMPION trial reported a 39% risk reduction in HF-related hospitalizations after 6 months in patients who received physician-led interventions based on the assessed PAP values compared with the control group [19]. These effects were sustained after 12 months and remained consistent in the nonrandomized 3-year extended follow-up period [20]. The prospective MEMS-HF registry (CardioMEMS European Monitoring Study for Heart Failure) had an observational design and yielded comparable effects when utilized within the European health care system [21].

These findings were corroborated by the pre–COVID-19 impact analysis of the GUIDE-HF trial (NCT03387813), which studied 1000 patients in NYHA class II or III from the United States [22], and the recent MONITOR-HF trial (NTR7673), conducted in 348 NYHA class III patients in the Netherlands [23]. Both trials indicated a sizeable benefit of PAP-guided management on HF hospitalization rates compared with the control group.

The CardioMEMS HF System in Germany: The PASSPORT-HF Trial

As aforementioned, the G-BA evaluates medical innovations for inclusion in the medical services covered by health insurance. Given the major differences in health care systems between the United States and Germany, the G-BA made a novel decision to directly fund a telemedicine randomized controlled trial across 50 study centers. The still ongoing PASSPORT-HF trial (Pulmonary Artery Sensor System Pressure Monitoring to Improve Heart Failure Outcomes; NCT04398654) aims to recruit 554 patients in NYHA class III who have experienced an HF hospitalization within the last 12 months.

PASSPORT-HF is investigating whether PAP-guided care, supported by HF nurses and utilizing daily monitoring of PAP values with the CardioMEMS system (Abbott), reduces the number of unplanned HF-related hospitalizations or all-cause mortality after 12 months.

In the PASSPORT-HF trial, patients with CHF are being randomized in a 1:1 ratio to either device implantation or, in contrast to the CHAMPION trial, no device implantation in the control arm. Both groups will receive structured telephone support from HF nurses in addition to guideline-compliant pharmacotherapy, which exceeds standard practice in Germany. Therefore, PASSPORT-HF aims to evaluate whether the previously reported benefits of PAP-guided hemodynamic monitoring as a novel HF management tool in routine outpatient telemedical care can be replicated in the German health care setting. Initial study results are expected in 2026 [24].

Recent Innovations: The CordellaHF Management System

The *Cordella HF Management System* (Cordella, Endotronix Inc.) comprises an implantable, battery-free PAP sensor and an external handheld patient reader. This system was evaluated in the SIRONA study program (NCT03375710).

The Cordella HF Management System is the first to facilitate the simultaneous recording of both invasive PAP data and externally measured parameters, such as blood pressure, heart rate, body weight, and oxygen saturation. This capability offers a more comprehensive overview of a patient's compensation status to the HF care team. Additionally, the system enables patients to access their trending data through an easy-to-use handheld device for PAP measurement. Unlike the CardioMEMS HF system, the Cordella HF Management System allows for PAP measurements to be taken in various body positions and under stress conditions. This flexibility is particularly important, as patients have expressed a clear preference for measurements conducted while seated [25].

The *Cordella system* was initially evaluated in a first-in-human study involving 15 patients, where the PAP difference measured 90 days after implantation showed no significant difference between the values obtained from the Cordella sensor and those from right heart catheterization. Notably, patient compliance with daily remote monitoring exceeded 98% [26]. Following this, the SIRONA-2 multicenter, single-arm, open-label study (NCT03375710) included 70 patients with HF in NYHA class III across 4 sites in Ireland and Belgium, as well as 3 centers in Germany.

The primary efficacy endpoint that compared the *Cordella* PAP sensor measurements with established catheter-derived reference pressure at 90 days was met (mean PAP difference 0.0-2.9 mmHg, P=.003), remaining in good agreement for up to 12

XSL•FO
RenderX

months of observation. There were no reports of sensor failures or deaths within the 90-day period [25], and improvements in NYHA class and 6-minute walk distance were observed [27]. Hence, the *Cordella* management system appears feasible and safe and is currently under investigation in the PROACTIVE-HF single-arm study (NCT04089059), which has prespecified safety and effectiveness endpoints to provide objective evidence of a similar risk-benefit profile to the *CardioMEMS* system in 450 patients in NYHA class III (Figure 2).

The Regulatory Process: From Trials to Standard Health Care in Germany

Telemonitoring in patients with CHF, as a digitally supported type of care, has been recognized as an independent examination and treatment method by the G-BA in March 2021 [28]. Consequently, telemonitoring became a standard health care service for this patient group in Germany [29].

Since 2009, health insurance has been compulsory for all German residents, as ruled by the German Insurance Contract Act [30]. As of 2022, approximately 87% of residents are covered by statutory health insurance, while about 10% have private insurance [31]. Both types offer telemonitoring as a service; however, private insurers define new aspects regarding current indication criteria (Textbox 1).

Textbox 1. The current inclusion criteria for heart failure telemonitoring (according to Gemeinsamer Bundesausschuss decision and reimbursement for private insurance).

- New York Heart Association II-III stage heart failure with a left ventricular ejection fraction <40%.
- Implanted cardiac device (implantable cardiac defibrillator or cardiac resynchronization therapy with a pacemaker/defibrillator) or heart failure hospitalization due to cardiac decompensation in the past year.
- Present treatment according to current guidelines (guideline-directed medical therapy).
- Absence of identifiable factors that compromise the transfer of monitoring data or that would interfere with patient self-management.
- If other prerequisites are fulfilled: patients with private health insurance who have chronic heart failure and exhibit a left ventricular ejection fraction >40%; also including hospitalization for decompensated heart failure within the last 12 months.

Reimbursement by statutory health insurance companies requires a prior assessment of benefit, necessity, and cost-effectiveness. For the evaluation of digital care services such as telemonitoring, the G-BA has the statutory task of conducting a qualitative benefit assessment. The submission of research projects for assessment by the G-BA is subject to a particularly strict “1-shot” strategy, meaning that projects can only be submitted once.

Research on the efficacy of telehealth interventions in the care of patients with CHF is inconsistent. Reasons for the differing clinical endpoints are variations in the telemedical systems used (invasive measurement sensors vs noninvasive monitoring) and different concepts of guided care, such as the involvement of HF nurses. Additionally, the response times to abnormally transmitted values varied, with some studies defining responses during regular consultation hours while others provided 24/7 care through a telemedicine center.

Despite these differences in study design, the German TIM-HF2 landmark trial, among others, identified high-risk patients with a 12-month history of HF-related hospitalization as the population that benefits most from telemedical cocare.

The fact that there is no treatment benefit in the absence of such a recent HF hospitalization is an important evidence-based selection criterion for the indication of telemonitoring, thereby avoiding unnecessary resource use [29].

Since 2016, the Institute for Quality and Efficiency in Health Care (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen [IQWiG]), as the highest joint self-governing body of the German health care system, has been investigating the benefits of invasive telemonitoring with active implants and noninvasive telemonitoring in CHF on behalf of the G-BA. The aforementioned IN-TIME trial for invasive telemonitoring and the German TIM-HF and TIM-HF2 trials for noninvasive telemonitoring formed the basis of an assessment regarding additional health care benefits [11,12,32]. In its assessment of October 2019, the IQWiG identified a statistically significant reduction in cardiovascular mortality through telemonitoring as a sufficient additional benefit [33].

According to the aforementioned G-BA decision, telemonitoring is currently only indicated for patients with CHF with reduced LVEF. This selection criterion is inconsistent with current trial data, which also revealed a benefit for patients with CHF with preserved LVEF [11]. A reevaluation by the G-BA and, ultimately, by the IQWiG for translation into medical care is currently pending.

The Role of Medical Scientific Societies in Telemonitoring Implementation Processes

The German Cardiac Society (Deutsche Gesellschaft für Kardiologie [DGK]) provides an important basis for the development of collaborative projects in telemedicine for patients with CHF. In 2005, the DGK founded a telemedicine working group (Working Group 33) that also addresses the certification criteria for telemedicine centers [34]. Accordingly, a position paper on the certification processes ensuring quality standards for telemedical centers, aligned with legal requirements, was published in 2022 [35].

A thorough understanding of the available technical capabilities is essential for the effective collaboration of different study groups and the development of joint projects. Therefore, the working group on CHF (Working Group 10) developed guidelines for patients planning long-distance travel [36]. As a

XSL•FO
RenderX

straightforward and intersecting approach to telemedical surveillance, remote monitoring of implanted devices (such as pacemakers, ICDs, and implantable cardiac monitors) has been recommended and further implemented by the European Society of Cardiology (ESC) [37-39].

Another important step toward representing professional interests at the national level was the establishment of the "Digital Transformation in Internal Medicine" Commission of the German Society of Internal Medicine (Deutsche Gesellschaft für Innere Medizin [DGIM]) in 2019. The DGIM, founded in 1882, is the largest medical society in Germany, steeped in tradition with over 30,000 members [40-42]. First, in 2014, the DGIM published guiding principles for the implementation of telemedical service provision [43]. The commission further serves its self-conception as a platform to impart literacy, offer training opportunities, and provide scientific support services.

From a European perspective, the ESC defined telemedical care concepts in a position statement in 2015 as 1 of 7 application areas of eHealth in cardiology [44]. As a result of technical developments and updated evidence, this paper was considered preliminary; therefore, the *ESC Digital Health Committee* was established in 2016 and is continuously discussing scientific updates in this area. The recent guidelines of the ESC on the diagnosis of acute HF and CHF [45] differentiate between hemodynamic, invasive telemonitoring using implants that measure PAP and noninvasive telemonitoring with external measuring devices. This distinction has led to the issuance of class II-B recommendations for both telemedical technologies.

International Cooperations

The German research and development project focused on telemedicine for patients with HF has always been linked to international advancements in this area. A notable example is the German-Austrian project "Telemed5000" (project number FKZ-01MD19014A), which explored the applicability of artificial intelligence (AI) within the workflow of a telemedical center. With a budget of approximately €2.4 million (US $2.6 million), this project was funded by the German Federal Ministry of Economic Affairs and Climate Action and cofunded by the Austrian Federal Ministry of Climate Action, Environment, Energy, Mobility, Innovation, and Technology [46].

There has been and continues to be an intensive international exchange regarding scientific results related to telemedicine in patients with HF, as well as insights from Germany's implementation process, particularly with neighboring European Union countries. For instance, the *German Foundation for the Chronically Ill* (Deutsche Stiftung für chronisch Kranke) organized a roundtable of experts from 7 European countries—France, Germany, Italy, Poland, Spain, the Netherlands, and the United Kingdom—in December 2023. This gathering aimed to review the current state of telemedicine in the participating countries, facilitate mutual learning, and provide impetus for European cooperation.

Establishing reliable regulations, overcoming regional differences, redefining roles and processes, personalizing health care services, promoting innovation and research, utilizing AI, and efficiently managing and safeguarding health care data have been identified as key levers for the further development of telemedicine.

Current Practice in Germany

An estimated 150,000-200,000 patients with CHF in Germany are currently eligible for telemedical cocare [47]. Additionally, the G-BA in Germany has defined criteria for the inclusion of current cardiovascular telemonitoring (Textbox 1).

Accordingly, care is provided by the primary physician (either a primary care doctor or a specialist in internal medicine or cardiology) in cooperation with a consulting physician (cardiologist) at the telemedicine center. Therefore, the entire care process is designed for outpatient settings only and is completely separated from in-hospital stakeholders.

Cardiologists who wish to provide telemonitoring must hold a license for treating patients within the framework of public health insurance. Additionally, the establishment of a telemonitoring center requires the presence of a technical infrastructure and a license for the follow-up of CIEDs [28,48].

After a patient has consented to telemonitoring, qualified HF nurses conduct training on the proper use of the issued equipment. This training is essential for ensuring adequate patient adherence, which forms the foundation for accurate measurements and subsequent data transfer. By contrast, the monitoring of CIEDs is automated, with individually programmed alerts transmitted as soon as a connection is established between the CIED and the transmitter. An 80% data transmission rate is mandatory and must be reported by the telemedicine center annually. The center is required to evaluate patient data at least 5 days a week or daily for severely ill patients under "intensified monitoring." If a response is necessary, such as a change in medication, the telemedicine center is obligated to contact the primary care physician or interact directly with the patient if the physician is unavailable. These responses must adhere to specified time frames: the processing of an alert must not exceed 24 hours, while the transmission of a subsequent clinical need for action should reach the primary attending physician within a maximum of 48 hours (Figure 3).

XSL•FO
RenderX

Figure 3. Components of telemonitoring in heart failure according to the G-BA (Federal Joint Committee/Gemeinsamer Bundesausschuss) approval in Germany.

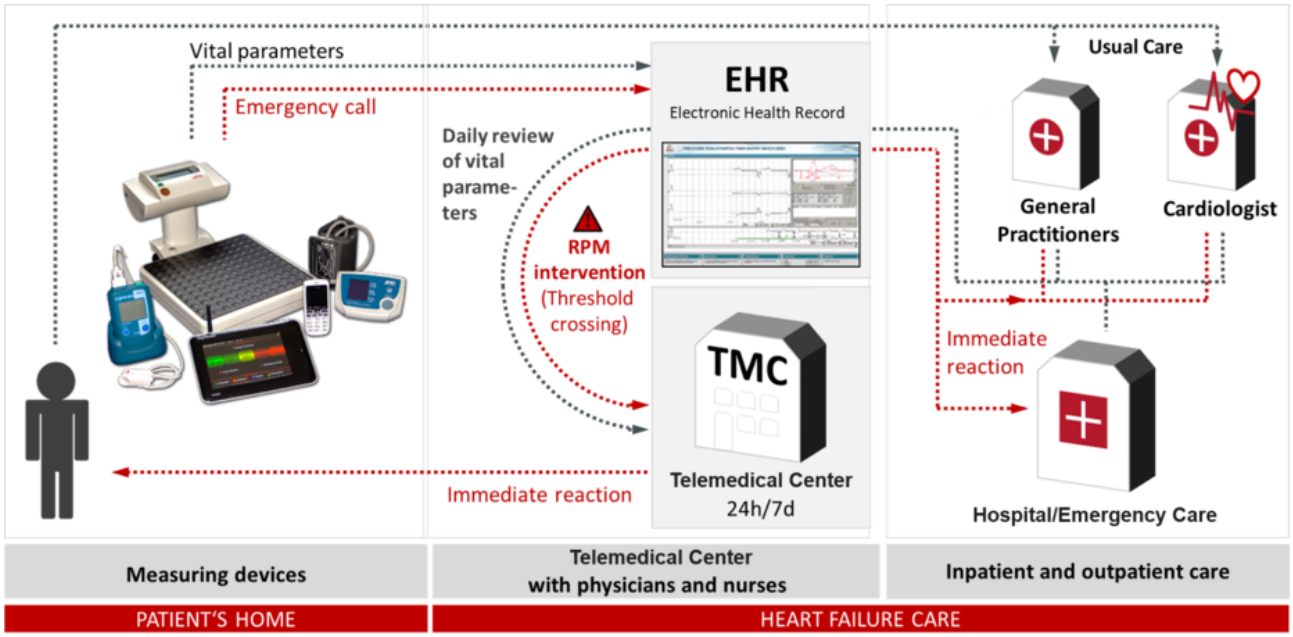

Furthermore, a quality assurance program for telemonitoring has been implemented, focusing on adherence to guidelines, daily data review, alarm management, and tracking hospitalizations for decompensated HF and other relevant scenarios.

Further Digital Applications in Germany

Digital health applications offer prescription-enabled mobile apps that differ from classic telecardiology by providing daily evaluations of transferred values and deriving subsequent actions. These apps constitute the third pillar of patient management, incorporating mobile health (mHealth), and primarily aim to empower patients to manage chronic diseases through lifestyle and behavioral modifications.

The central website of the German Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte [BfArM]) offers a directory of available digital health apps [49] categorized into 12 different medical fields, providing relevant information for both patients and health care providers. To date, only 1 Digital Health Application (DiGA), "ProHerz," has been provisionally approved for patients with HF. However, it is recognized that long-term engagement with digital apps is achieved by only a small number of users; for instance, the 30-day retention rates for patients using mental health apps were only 3.3% [50].

To enhance adherence, future "DiGAs 2.0" that incorporate AI applications promise feasible transmission of vital parameters tracked by wearable devices. This advancement could facilitate more effective therapies in the remote analysis of HF compensation but necessitates thorough scientific evaluation. The initial narrative of a DiGA designed for sole use by patients is currently being reevaluated to explore higher scalability for remote monitoring and a more detailed integration into the medical practices of general practitioners and specialists.

Outlook on Upcoming Challenges

Unmet Questions of Remote Patient Monitoring

Telemonitoring is an effective technology regarding availability, accessibility, and affordability. However, improvements in outcomes for patients with CHF through telemedicine have not been universally demonstrated across all available technologies and devices. Refining the current evidence remains challenging due to unresolved questions about the most appropriate monitoring technology for individual patients, the optimal time to initiate telemonitoring, and, most importantly, the duration of monitoring following cardiac decompensation [51].

Furthermore, there is a need to incorporate current evidence for special patient populations, such as patients with HF with preserved ejection fraction [52], into the catalog of services provided by statutory health insurance without requiring a complete restart of the authorization process.

In contrast to the introduction of a new drug therapy, transferring a digital therapy from a controlled study setup to routine health care is significantly more complex. To facilitate the introduction of noninvasive telemedicine in HF, the publicly funded prospective phase IV registry study, "TIM-HF4" (Telemedical Interventional Management in Heart Failure 4), will commence in 2025. This study aims to investigate patient acceptance and the effectiveness of noninvasive telemonitoring in HF under real-life conditions, focusing on outcomes such as repeated hospitalizations due to HF, death from any cause, and health economic criteria.

The transferability of telemonitoring to other cardiological indications has not been sufficiently explored in scientific trials to date. However, telemonitoring could be beneficial in clinical settings by improving discharge processes, such as accelerating patient discharge following rhythmological interventions and enhancing aftercare. Notably, telemedical supervision may facilitate faster and more reliable uptitration of

XSL•FO
RenderX

guideline-directed medical therapy, taking into account patients' clinical tolerability and vital signs. The STRONG-HF trial highlighted this potential, demonstrating a significantly reduced risk of HF readmission, and alleviated symptoms when guideline-directed medical therapy was uptitrated within 2 weeks of discharge, alongside monitoring of the clinical status and evaluation of laboratory values [53].

In the context of a prehospital setting, the ongoing randomized controlled study, ResKriVer-TAVI, funded by the Federal Ministry for Economic Affairs and Climate Action (German Clinical Trials Register DRKS00027842), is investigating whether telemedical interventional management enhances clinical outcomes, specifically cardiovascular hospitalization and death from any cause, in patients with aortic stenosis awaiting transcatheter aortic valve implantation [54].

While feasibility studies demonstrate the proof of concept for the latest technologies, there is a critical need for randomized trials to evaluate the clinical efficacy of cutting-edge telemonitoring devices and their eventual integration into routine practice. As it stands, it remains unclear whether noninvasive measurements obtained through wearable devices are equally effective compared with implanted cardiac devices. Additionally, the potential advantages of combining methods that integrate multiple sensors have yet to be tested against a single-sensor approach for superiority.

AI-Supported Applications in Telemonitoring and Remote Patient Management

As the number of patients eligible for telemedical care continues to grow, machine learning tools hold significant promise for addressing the challenges of numerical scaling and the preprocessing of complex health data from diverse sources. Currently, triaging vital parameters is an established practice. In the future, advancements in pattern recognition are expected to enhance prioritization and improve the prediction of individual decompensation thresholds.

Nevertheless, AI applications are currently not eligible for standard cardiovascular telemedicine care. With the implementation of the European AI Act in March 2024 [55], clear transparency requirements and specific obligations for higher-risk AI applications will create a wide range of opportunities for research into their applications in telemonitoring patient care.

AI is also a promising tool for identifying novel HF parameters: AI applications could integrate smartwatch-based data with established intracardiac and extracardiac device monitoring data. A real-time analysis of wearable-derived intrinsic hemodynamic changes (such as heart rhythm, heart rate, level of activity, pedometer data, and oxygen saturation) could further enhance patient-related outcomes by providing instant alerts to the telemedical center.

The telemedical sensor system will likely be complemented by automated speech analysis to detect pulmonary congestion, as promising data showed a prediction of HF events up to 3 weeks in advance [56]. The "Cordio HearO" smartphone app was introduced in a pilot study including 40 patients with acute cardiac decompensation [57]. This innovative strategy is currently investigated as part of the "Telemed5000-Stimme" observational study (DRKS00020763) in patients on regular hemodialysis [58] in Germany, as well as the "VAMP-HF" study (AI-Based Voice Analysis for Monitoring Patients Hospitalized With Acute Decompensated Heart Failure) in patients with acute decompensated HF at the Deutsches Herzzentrum der Charité and the US Mayo Clinic [59].

Besides their promising transformative potential, the implementation of AI tools in cardiovascular telemonitoring remains a challenge in terms of ethical scenarios and data protection.

Discussion

The development of telemedicine for the care of patients with HF in Germany has been an extremely long and difficult process, paved by public funding. Important milestones are major pioneering studies such as the TIM-HF2 and IN-TIME trials, which demonstrated the effectiveness of telemedicine approaches using various technological methodologies. The success of these initiatives was largely due to the close cooperation between researchers and technology developers, who created innovative solutions for clinical practice. One significant hurdle that had to be overcome was the traditionally strong separation between inpatient and outpatient care settings in Germany.

With the gradual introduction of telemedical services for patients with HF, research in this area has not come to an end. Instead, the knowledge gained will be used to continuously improve care and advance the digital transformation of the health care system. Care research approaches, such as the forthcoming TIM-HF4 study, play a crucial role in evaluating the efficacy of noninvasive telemonitoring in a real-life setting, particularly regarding health economic criteria.

As a future approach, the establishment of telemedical care for patients with HF across Europe is likely to benefit from the experiences gained in Germany. Despite the lengthy journey, these efforts have resulted in significant improvements in the care of patients with HF.

Acknowledgments

We thank Volker Möller for the figure designs.

Authors' Contributions

S Spethmann and FK equally drafted the manuscript. GH, KK, MB, BA, M Moecke, M Mittermaier, S Stoerk, ML, DR, RB, AA, SvH, TMH, CEA, S Sack, TB, JWG, MK, GE, CV, SW, CM, JW, BZS, ND, JZ, and FK contributing textual content. FK

XSL•FO
RenderX

und S Spethmann supervised manuscript writing and provided continuous guidance. All authors read and approved the final manuscript.

Conflicts of Interest

Stefan von Haehling has been a paid consultant for and/or received honoraria payments from AstraZeneca, Bayer, Boehringer Ingelheim, BRAHMS, Chugai, Edwards Lifesciences, Grünenthal, Helsinn, Hexal, Novartis, Novo Nordisk, Pharmacosmos, Respicardia, Roche, Servier, Sorin, and Vifor. S.v.H. reports research support from Amgen, Boehringer Ingelheim, IMI, and the German Center for Cardiovascular Research (DZHK).
Thomas M. Helms does not have a financial interest / arrangement or affiliation with one or more organizations that could be perceived as a real or apparent conflict of interest in the context of the subject of this paper.
Ralph Bosch received lecture honoraria and consultant remuneration from BIOTRONIK and Medtronic.
Birgit Assmus received speaking and / or consulting fees outside the submitted work from the following companies: Abbot, AstraZeneca, Bayer, Boehringer-Ingelheim, BMS, Edwards, Novartis, NovoNordisk, Pfizer, Vifor CSL, ZOLL.
Friedrich Koehler received research grants from the German Federal Ministry of Education and Research (BMBF) as well as the German Federal Ministry of Economic Affairs and Climate Action (BMWK). He received consultant remuneration from Biotronik.

References

1. Flesche CW, Jalowy A, Inselmann G. Telemedicine in the maritime environment--hightech with a fine tradition. Med Klin (Munich) 2004 Mar 15;99(3):163-168. [doi: 10.1007/s00063-004-1026-5] [Medline: 15024490]
2. Setoguchi S, Stevenson LW, Schneeweiss S. Repeated hospitalizations predict mortality in the community population with heart failure. Am Heart J 2007 Aug;154(2):260-266. [doi: 10.1016/j.ahj.2007.01.041] [Medline: 17643574]
3. Statistisches Bundesamt. Diagnosedaten der patienten und patientinnen in krankenhäusern (einschl. sterbe- und stundenfälle) - fachserie 12 reihe 6.2.1 - 2016 (letzte ausgabe - berichtsweise eingestellt). Statistisches Bundesamt. 2016. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Publikationen/Downloads-Krankenhaeuser/diagnosedaten-krankenhaus-2120621167005.html [accessed 2024-06-01]
4. Neumann T, Biermann J, Erbel R, Neumann A, Wasem J, Ertl G. Heart failure: the commonest reason for hospital admission in Germany: medical and economic perspectives. Dtsch Arztebl Int 2009;106(16):269-275. [doi: 10.3238/arztebl.2009.0566]
5. Cleland JGF, Louis AA, Rigby AS, Janssens U, Balk AHMM, TEN-HMS Investigators. Noninvasive home telemonitoring for patients with heart failure at high risk of recurrent admission and death: the Trans-European Network-Home-Care Management System (TEN-HMS) study. J Am Coll Cardiol 2005 May 17;45(10):1654-1664 [FREE Full text] [doi: 10.1016/j.jacc.2005.01.050] [Medline: 15893183]
6. Angermann CE, Störk S, Gelbrich G, Faller H, Jahns R, Frantz S, Competence Network Heart Failure. Mode of action and effects of standardized collaborative disease management on mortality and morbidity in patients with systolic heart failure: the Interdisciplinary Network for Heart Failure (INH) study. Circ Heart Fail 2012 Jan;5(1):25-35. [doi: 10.1161/CIRCHEARTFAILURE.111.962969] [Medline: 21956192]
7. Angermann CE, Sehner S, Faller H, Güder G, Morbach C, Frantz S, INH Study Groupof the Competence Network Heart Failure. Longer-term effects of remote patient management following hospital discharge after acute systolic heart failure: the randomized E-INH trial. JACC Heart Fail 2023 Feb;11(2):191-206 [FREE Full text] [doi: 10.1016/j.jchf.2022.10.016] [Medline: 36718715]
8. Winkler S, Koehler K, Prescher S, Koehler M, Kirwan BA, Tajsic M, et al. Is 24/7 remote patient management in heart failure necessary? Results of the telemedical emergency service used in the TIM-HF and in the TIM-HF2 trials. ESC Heart Fail 2021 Oct;8(5):3613-3620 [FREE Full text] [doi: 10.1002/ehf2.13413] [Medline: 34182596]
9. Koehler F, Winkler S, Schieber M, Sechtem U, Stangl K, Böhm M, Telemedical Interventional Monitoring in Heart Failure Investigators. Impact of remote telemedical management on mortality and hospitalizations in ambulatory patients with chronic heart failure: the telemedical interventional monitoring in heart failure study. Circulation 2011 May 03;123(17):1873-1880. [doi: 10.1161/CIRCULATIONAHA.111.018473] [Medline: 21444883]
10. Koehler F, Winkler S, Schieber M, Sechtem U, Stangl K, Böhm M, et al. Telemedicine in heart failure: pre-specified and exploratory subgroup analyses from the TIM-HF trial. Int J Cardiol 2012 Nov 29;161(3):143-150. [doi: 10.1016/j.ijcard.2011.09.007] [Medline: 21982700]
11. Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Kirwan BA, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. Lancet 2018 Sep 22;392(10152):1047-1057. [doi: 10.1016/S0140-6736(18)31880-4] [Medline: 30153985]
12. Hindricks G, Taborsky M, Glikson M, Heinrich U, Schumacher B, Katz A, IN-TIME study group*. Implant-based multiparameter telemonitoring of patients with heart failure (IN-TIME): a randomised controlled trial. Lancet 2014 Aug 16;384(9943):583-590. [doi: 10.1016/S0140-6736(14)61176-4] [Medline: 25131977]
13. Böhm M, Drexler H, Oswald H, Rybak K, Bosch R, Butter C, OptiLink HF Study Investigators. Fluid status telemedicine alerts for heart failure: a randomized controlled trial. Eur Heart J 2016 Nov 01;37(41):3154-3163. [doi: 10.1093/eurheartj/ehw099] [Medline: 26984864]

XSL•FO
RenderX

14. Brachmann J, Böhm M, Rybak K, Klein G, Butter C, Klemm H, OptiLink HF Study Executive BoardInvestigators. Fluid status monitoring with a wireless network to reduce cardiovascular-related hospitalizations and mortality in heart failure: rationale and design of the OptiLink HF Study (Optimization of Heart Failure Management using OptiVol Fluid Status Monitoring and CareLink). Eur J Heart Fail 2011 Jul;13(7):796-804 [FREE Full text] [doi: 10.1093/eurjhf/hfr045] [Medline: 21555324]
15. Wintrich J, Pavlicek V, Brachmann J, Bosch R, Butter C, Oswald H, et al. Remote monitoring with appropriate reaction to alerts was associated with improved outcomes in chronic heart failure: results from the OptiLink HF study. Circ Arrhythm Electrophysiol 2021 Jan;14(1):e008693. [doi: 10.1161/CIRCEP.120.008693] [Medline: 33301362]
16. Wintrich J, Pavlicek V, Brachmann J, Bosch R, Butter C, Oswald H, et al. Differences in management of telemedicine alerts on weekdays and public holidays: results from the OptiLink heart failure trial. J Telemed Telecare 2024 Jan;30(1):173-179. [doi: 10.1177/1357633X211039398] [Medline: 34524925]
17. Wintrich J, Pavlicek V, Brachmann J, Bosch R, Butter C, Oswald H, et al. Impedance-based remote monitoring in patients with heart failure and concomitant chronic kidney disease. ESC Heart Fail 2023 Oct;10(5):3011-3018 [FREE Full text] [doi: 10.1002/ehf2.14387] [Medline: 37537796]
18. Zile MR, Bennett TD, St John Sutton M, Cho YK, Adamson PB, Aaron MF, et al. Transition from chronic compensated to acute decompensated heart failure: pathophysiological insights obtained from continuous monitoring of intracardiac pressures. Circulation 2008 Sep 30;118(14):1433-1441. [doi: 10.1161/CIRCULATIONAHA.108.783910] [Medline: 18794390]
19. Abraham WT, Adamson PB, Bourge RC, Aaron MF, Costanzo MR, Stevenson LW, CHAMPION Trial Study Group. Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial. Lancet 2011 Feb 19;377(9766):658-666. [doi: 10.1016/S0140-6736(11)60101-3] [Medline: 21315441]
20. Abraham WT, Stevenson LW, Bourge RC, Lindenfeld JA, Bauman JG, Adamson PB, CHAMPION Trial Study Group. Sustained efficacy of pulmonary artery pressure to guide adjustment of chronic heart failure therapy: complete follow-up results from the CHAMPION randomised trial. Lancet 2016 Jan 30;387(10017):453-461. [doi: 10.1016/S0140-6736(15)00723-0] [Medline: 26560249]
21. Angermann CE, Assmus B, Anker SD, Asselbergs FW, Brachmann J, Brett ME, MEMS-HF Investigators. Pulmonary artery pressure-guided therapy in ambulatory patients with symptomatic heart failure: the CardioMEMS European Monitoring Study for Heart Failure (MEMS-HF). Eur J Heart Fail 2020 Oct;22(10):1891-1901. [doi: 10.1002/ejhf.1943] [Medline: 32592227]
22. Lindenfeld J, Zile MR, Desai AS, Bhatt K, Ducharme A, Horstmanshof D, et al. Haemodynamic-guided management of heart failure (GUIDE-HF): a randomised controlled trial. Lancet 2021 Sep 11;398(10304):991-1001. [doi: 10.1016/S0140-6736(21)01754-2] [Medline: 34461042]
23. Brugts JJ, Radhoe SP, Clephas PRD, Aydin D, van Gent MWF, Szymanski MK, MONITOR-HF investigators. Remote haemodynamic monitoring of pulmonary artery pressures in patients with chronic heart failure (MONITOR-HF): a randomised clinical trial. Lancet 2023 Jun 24;401(10394):2113-2123. [doi: 10.1016/S0140-6736(23)00923-6] [Medline: 37220768]
24. Störk S, Bernhardt A, Böhm M, Brachmann J, Dagres N, Frantz S, et al. Pulmonary artery sensor system pressure monitoring to improve heart failure outcomes (PASSPORT-HF): rationale and design of the PASSPORT-HF multicenter randomized clinical trial. Clin Res Cardiol 2022 Nov;111(11):1245-1255 [FREE Full text] [doi: 10.1007/s00392-022-01987-3] [Medline: 35246723]
25. Sharif F, Rosenkranz S, Bartunek J, Kempf T, Assmus B, Mahon NG, et al. Safety and efficacy of a wireless pulmonary artery pressure sensor: primary endpoint results of the SIRONA 2 clinical trial. ESC Heart Fail 2022 Oct;9(5):2862-2872 [FREE Full text] [doi: 10.1002/ehf2.14006] [Medline: 35686479]
26. Mullens W, Sharif F, Dupont M, Rothman AMK, Wijns W. Digital health care solution for proactive heart failure management with the Cordella Heart Failure System: results of the SIRONA first-in-human study. Eur J Heart Fail 2020 Oct;22(10):1912-1919. [doi: 10.1002/ejhf.1870] [Medline: 32476191]
27. Sharif F, Rosenkranz S, Bartunek J, Kempf T, Aßmus B, Mahon NG, et al. Twelve-month follow-up results from the SIRONA 2 clinical trial. ESC Heart Fail 2024 Apr;11(2):1133-1143 [FREE Full text] [doi: 10.1002/ehf2.14657] [Medline: 38271076]
28. Beschluss des Gemeinsamen Bundesausschusses über eine änderung der richtlinie methoden vertragsärztliche versorgung: telemonitoring bei herzinsuffizienz. Gemeinsamer Bundesausschuss. 2020. URL: https://www.g-ba.de/beschluesse/4648/ [accessed 2024-05-14]
29. Spethmann S, Köhler F. Telemedicine in chronic heart failure-From clinical studies to standard care. Internist (Berl) 2022 Mar;63(3):266-273. [doi: 10.1007/s00108-022-01268-1] [Medline: 35138432]
30. Versicherungsvertragsgesetz vom 23. November 2007 (BGBl. I S. 2631), das zuletzt durch Artikel 4 des Gesetzes vom 11. April 2024 (BGBl. 2024 I Nr. 119) geändert worden ist. Gesetze im Internet. 2024 Apr 11. URL: https://www.gesetze-im-internet.de/vvg_2008/BJNR263110007.html [accessed 2024-06-09]
31. vdek-basisdaten des gesundheitswesens in Deutschland 2024. Verband der Ersatzkassen e. V. URL: https://www.vdek.com/presse/daten.html#urheberrecht [accessed 2024-06-09]

XSL•FO
RenderX

32. Koehler J, Stengel A, Hofmann T, Wegscheider K, Koehler K, Sehner S, et al. Telemonitoring in patients with chronic heart failure and moderate depressed symptoms: results of the Telemedical Interventional Monitoring in Heart Failure (TIM-HF) study. Eur J Heart Fail 2021 Jan;23(1):186-194 [FREE Full text] [doi: 10.1002/ejhf.2025] [Medline: 33063412]
33. Dierks C, Nitz G, Grau U, Mehlitz A. Datengestütztes, zeitnahes management in zusammenarbeit mit einem ärztlichen telemedizinischen zentrum für patientinnen und patienten mit einer fortgeschrittenen herzinsuffizienz - rapid report. In: IQWiG und Industrie – Rechtliche Fragen zum Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Berlin/Heidelberg, Germany: Springer; Apr 03, 2008.
34. Helms TM, Perings C, Sommer P, Köhler F, Frey N, von Haehling S, et al. Positionspapier zur zertifizierung von telemedizinzentren. Kardiologe 2021 Dec 06;16(1):6-20. [doi: 10.1007/s12181-021-00522-4]
35. Helms T, Perings C, Sommer P, Köhler F, Frey N, von Haehling S, et al. Positionspapier zur zertifizierung von telemedizinzentren. Kardiologe 2021 Dec 06;16(1):6-20 [FREE Full text] [doi: 10.1007/s12181-021-00522-4]
36. von Haehling S, Birner C, Dworatzek E, Frantz S, Hellenkamp K, Israel CW, et al. Travelling with heart failure: risk assessment and practical recommendations. Nat Rev Cardiol 2022 May;19(5):302-313. [doi: 10.1038/s41569-021-00643-z] [Medline: 34992256]
37. European Society of Cardiology (ESC), European Heart Rhythm Association (EHRA), Brignole M, Auricchio A, Baron-Esquivias G, Bordachar P, et al. 2013 ESC guidelines on cardiac pacing and cardiac resynchronization therapy: the task force on cardiac pacing and resynchronization therapy of the European Society of Cardiology (ESC). Developed in collaboration with the European Heart Rhythm Association (EHRA). Europace 2013 Aug;15(8):1070-1118. [doi: 10.1093/europace/eut206] [Medline: 23801827]
38. Dubner S, Auricchio A, Steinberg JS, Vardas P, Stone P, Brugada J, et al. ISHNE/EHRA expert consensus on remote monitoring of cardiovascular implantable electronic devices (CIEDs). Europace 2012 Feb;14(2):278-293 [FREE Full text] [doi: 10.1093/europace/eur303] [Medline: 22232544]
39. Slotwiner D, Varma N, Akar JG, Annas G, Beardsall M, Fogel RI, et al. HRS Expert Consensus Statement on remote interrogation and monitoring for cardiovascular implantable electronic devices. Heart Rhythm 2015 Jul;12(7):e69-100. [doi: 10.1016/j.hrthm.2015.05.008] [Medline: 25981148]
40. DGIM:History. Die Deutsche Gesellschaft für Innere Medizin e.V. (DGIM). URL: https://www.dgim.de/ueber-uns/historie/ [accessed 2024-06-06]
41. Vogelmeier CF, Köhler F. "Digitalization in internal medicine" - not "digitalization of internal medicine" - a promising perspective from non-digital natives. Dtsch Med Wochenschr 2019 Apr;144(7):425. [doi: 10.1055/a-0746-5964] [Medline: 30925594]
42. Deutsche Gesellschaft für Innere Medizin. Ordentliche Mitglieder. Deutsche Gesellschaft für Innere Medizin (DGIM). URL: https://www.dgim.de/mitglieder/ordentliche-mitgliedschaft/#:~:text=Werden%20Sie%20Mitglied%20in%20der,wissenschaftlichen%20Fachgesellschaften%20in%20Europa%20z%C3%A4hlt [accessed 2024-06-10]
43. DGIM:Positionspapiere/Stellungsnahmen. Deutsche Gesellschaft für Innere Medizin (DGIM). URL: https://www.dgim.de/veroeffentlichungen/positionspapiere/stellungnahmen/ [accessed 2024-06-06]
44. Cowie MR, Bax J, Bruining N, Cleland JG, Koehler F, Malik M, et al. e-Health: a position statement of the European Society of Cardiology. Eur Heart J 2016 Jan 01;37(1):63-66 [FREE Full text] [doi: 10.1093/eurheartj/ehv416] [Medline: 26303835]
45. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, ESC Scientific Document Group. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021 Sep 21;42(36):3599-3726 [FREE Full text] [doi: 10.1093/eurheartj/ehab368] [Medline: 34447992]
46. Köhler F. Telemed5000 Projekt: Telemedizin: Klinische Studien: Telemed5000-HF. URL: https://www.telemed5000.de/telemedizin/klinische-studien/telemed5000-hf/ [accessed 2024-05-12]
47. Schneider D, Köhler K, Köhler F. Telemedicine in cardiology - what is new? Dtsch Med Wochenschr 2023 Jun;148(12):767-773. [doi: 10.1055/a-1928-0362] [Medline: 37257479]
48. Vereinbarung von qualitätssicherungsmaßnahmen nach § 135 Abs. 2 SGB V zur kontrolle von aktiven kardialen rhythmusimplantaten (qualitätssicherungsvereinbarung rhythmusimplantat-kontrolle). KBV - Kassenärztliche Bundesvereinigung. URL: https://www.kbv.de/media/sp/Rhythmusimplantat-Kontrolle.pdf [accessed 2024-05-14]
49. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). URL: https://diga.bfarm.de/de [accessed 2024-11-02]
50. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
51. Bekfani T, Fudim M, Cleland JGF, Jorbenadze A, von Haehling S, Lorber A, et al. A current and future outlook on upcoming technologies in remote monitoring of patients with heart failure. Eur J Heart Fail 2021 Jan;23(1):175-185 [FREE Full text] [doi: 10.1002/ejhf.2033] [Medline: 33111389]
52. Kerwagen F, Koehler K, Vettorazzi E, Stangl V, Koehler M, Halle M, et al. Remote patient management of heart failure across the ejection fraction spectrum: a pre-specified analysis of the TIM-HF2 trial. Eur J Heart Fail 2023 Sep;25(9):1671-1681 [FREE Full text] [doi: 10.1002/ejhf.2948] [Medline: 37368507]

XSL•FO
RenderX

53. Mebazaa A, Davison B, Chioncel O, Cohen-Solal A, Diaz R, Filippatos G, et al. Safety, tolerability and efficacy of up-titration of guideline-directed medical therapies for acute heart failure (STRONG-HF): a multinational, open-label, randomised, trial. Lancet 2022 Dec 03;400(10367):1938-1952. [doi: 10.1016/S0140-6736(22)02076-1] [Medline: 36356631]
54. Mattig I, Koehler K, Barzen G, Hiddemann M, Kugel E, Roemmelt C, et al. Telemedical management in patients waiting for transcatheter aortic valve implantation: the ResKriVer-TAVI study design. Front Cardiovasc Med 2023;10:1352592 [FREE Full text] [doi: 10.3389/fcvm.2023.1352592] [Medline: 38322273]
55. Gesetz über künstliche Intelligenz: Legislative Entschließung des Europäischen Parlaments vom 13. März 2024 zu dem Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). Europäisches Parlament. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_DE.pdf [accessed 2024-06-10]
56. Abraham WT. The HearO Community Study: remote speech analysis to predict worsening heart failure events. ESC 365. 2022 May 21. URL: https://esc365.escardio.org/presentation/250301 [accessed 2024-06-15]
57. Amir O, Abraham WT, Azzam ZS, Berger G, Anker SD, Pinney SP, et al. Remote speech analysis in the evaluation of hospitalized patients with acute decompensated heart failure. JACC Heart Fail 2022 Jan;10(1):41-49 [FREE Full text] [doi: 10.1016/j.jchf.2021.08.008] [Medline: 34969496]
58. Koehler F. Telemed5000 Projekt: telemedizin: klinische studien: Telemed5000-Stimme. TELEMED5000. URL: https://www.telemed5000.de/telemedizin/klinische-studien/telemed5000-stimme/ [accessed 2024-05-12]
59. Künstliche intelligenz erkennt herzerkrankungen am klang der stimme: neue studie am DHZC. Informationsdienst Wissenschaft. 2024 Apr 3. URL: https://www.dhzb.de/presse/news/detailansicht-meldungen/ansicht/pressedetail/eine-fabel-zweimal-taeglich [accessed 2024-05-18]

Abbreviations

AI: artificial intelligence
BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte
CHF: chronic heart failure
CIED: cardiac implantable electronic device
CRT-D: cardiac resynchronization therapy with a defibrillator
DGIM: Deutsche Gesellschaft für Innere Medizin
DGK: Deutsche Gesellschaft für Kardiologie
DiGA: Digital Health Application
ECG: electrocardiogram
ECG: European Society of Cardiology
G-BA: Federal Joint Committee (Gemeinsamer Bundesausschuss)
HF: heart failure
INH: Interdisciplinary Network for Heart Failure
IN-TIME: Influence of Home Monitoring on Mortality and Morbidity in Heart Failure Patients with Impaired Left Ventricular Function
IQWiG: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
LVEF: left ventricular ejection fraction
mHealth: mobile health
NYHA: New York Heart Association
PAP: pulmonary artery pressure
TIM-HF: Telemedical Interventional Monitoring in Heart Failure Study
TIM-HF2: Telemedical Interventional Management in Heart Failure II

Edited by G Eysenbach; submitted 18.06.24; peer-reviewed by EP Greenup, M Nomali; comments to author 24.07.24; revised version received 13.08.24; accepted 26.09.24; published 04.12.24.

Please cite as:
Spethmann S, Hindricks G, Koehler K, Stoerk S, Angermann CE, Böhm M, Assmus B, Winkler S, Möckel M, Mittermaier M, Lelgemann M, Reuter D, Bosch R, Albrecht A, von Haehling S, Helms TM, Sack S, Bekfani T, Gröschel JW, Koehler M, Melzer C, Wintrich J, Zippel-Schultz B, Ertl G, Vogelmeier C, Dagres N, Zernikow J, Koehler F
Telemonitoring for Chronic Heart Failure: Narrative Review of the 20-Year Journey From Concept to Standard Care in Germany
J Med Internet Res 2024;26:e63391
URL: https://www.jmir.org/2024/1/e63391
doi:10.2196/63391
PMID:

XSL•FO
RenderX

©Sebastian Spethmann, Gerhard Hindricks, Kerstin Koehler, Stefan Stoerk, Christiane E Angermann, Michael Böhm, Birgit Assmus, Sebastian Winkler, Martin Möckel, Mirja Mittermaier, Monika Lelgemann, Daniel Reuter, Ralph Bosch, Alexander Albrecht, Stephan von Haehling, Thomas M Helms, Stefan Sack, Tarek Bekfani, Jan Wolfgang Gröschel, Magdalena Koehler, Christoph Melzer, Jan Wintrich, Bettina Zippel-Schultz, Georg Ertl, Claus Vogelmeier, Nikolaos Dagres, Jasmin Zernikow, Friedrich Koehler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Don't Forget the Humble Text Message: 25 Years of Text Messaging in Health

Rosie Dobson[1,2], PhD; Robyn Whittaker[1,2], PhD; Lorien C Abroms[3], DSc; Dale Bramley[2], MBChB; Caroline Free[4], PhD; Hayden McRobbie[5,6], PhD; Melanie Stowell[1], MSc; Anthony Rodgers[7], PhD

[1]School of Population Health, University of Auckland, Auckland, New Zealand

[2]Service Improvement and Innovation, Te Whatu Ora, Auckland, New Zealand

[3]Department of Prevention and Community Health, Milken Institute School of Public Health, The George Washington University, Washington, DC, DC, United States

[4]Clinical Trials Unit, London School of Hygiene & Tropical Medicine, London, United Kingdom

[5]National Public Health Service, Te Whatu Ora, Auckland, New Zealand

[6]Faculty of Medicine & Health, University of New South Wales, Sydney, Australia

[7]The George Institute for Global Health, Sydney, Australia

Corresponding Author:
Rosie Dobson, PhD
School of Population Health
University of Auckland
Private Bag 92019
Auckland, 1142
New Zealand
Phone: 64 93737599
Email: r.dobson@auckland.ac.nz

Abstract

Since the early studies exploring the use of SMS text messaging for health intervention, text messaging has played a pivotal role in the advancement of mobile health. As an intervention modality, text messaging has provided vital learnings for the design and delivery of interventions, particularly in low-resource settings. Despite the advances in technology over the last 25 years, text messaging is still being used in largely the same way to deliver health information, behavior change interventions, and support. The strong, consistent evidence for the benefits of this type of intervention has made text messaging a routine part of health interventions around the world. Key to its success is its simplicity, alongside the benefit of being arguably the most accessible form of consumer digital health intervention. Text message interventions are well suited for public health interventions due to their low cost, vast reach, frequent use, high read rates, and ability to be tailored and personalized. Furthermore, the nature of text messaging interventions makes them ideal for the delivery of multilingual, culturally tailored interventions, which is important in the context of increasing cultural diversity in many countries internationally. Indeed, studies assessing text message–based health interventions have shown them to be effective across sociodemographic and ethnic groups and have led to their adoption into national-level health promotion programs. With a growing focus on artificial intelligence, robotics, sensors, and other advances in digital health, there is an opportunity to integrate these technologies into text messaging programs. Simultaneously, it is essential that equity remains at the forefront for digital health researchers, developers, and implementers. Ensuring digital health solutions address inequities in health experienced across the world while taking action to maximize digital inclusion will ensure the true potential of digital health is realized. Text messaging has the potential to continue to play a pivotal role in the delivery of equitable digital health tools to communities around the world for many years to come. Further new technologies can build on the humble text message, leveraging its success to advance the field of digital health. This Viewpoint presents a retrospective of text messaging in health, drawing on the example of text message–based interventions for smoking cessation, and presents evidence for the continued relevance of this mobile health modality in 2025 and beyond.

(J Med Internet Res 2024;26:e59888) doi:10.2196/59888

KEYWORDS

text messaging; messaging; SMS; texting; mHealth; mobile health

XSL•FO
RenderX

Introduction

In December 1992, the first SMS text message was sent. It contained the words "Merry Christmas" and was sent by a 22-year-old software programmer from the United Kingdom [1]. It took several more years for text messaging to become a standard means of communication from mobile phone to mobile phone. Now, as the most widely used feature on mobile phones, more than 5 billion people send and receive text messages globally [1]. With the rise in messaging apps such as WhatsApp and WeChat, the end of the text message was envisioned, but contrary to expectations, text messages have still remained a key form of communication, with around 40 billion messages sent every day around the world [1]. Their ubiquity is not restricted to high-income contexts, with mobile phone penetration high internationally across demographic groups [2]. People tend to not ignore text messages, with 95% of all text messages opened and read within 3 minutes of being received [3], enabling text messages to be effective prompts. In comparison, email open rates average just 20% [4]. Text messages also have the greatest response rate—209% higher than phone calls, email, or social media apps like Facebook [3]. Coupled with their low cost to implement, this makes the humble text message a valuable tool for health services [5].

Text messaging has been used to support or deliver health services and programs for approximately 25 years [5]. Early research agendas included testing the use of text messaging to communicate health information, such as coordination of health professionals' schedules, patient appointment reminders, delivery of patient test results, and monitoring of posttreatment side effects [6]. The first published study using text messaging for health promotion is said to be from 2002 [5,7]; it consisted of a program to support young people in managing their asthma with daily reminders to use an inhaler, health education tips, and safety messages [8]. Publications on the topic increased in the following years, with the first randomized controlled trial (RCT) published in 2005 [9], and the first systematic review on text messaging interventions for health published in 2009 [6,7]. This review assessed findings from 14 studies, most of which focused on text messaging to support clinical care (eg, diabetes self-management [6]).

This Viewpoint presents a reflection on text messaging in health and discusses why this modality is still relevant today despite the rapid advances in technology in health. We will use an example of a text message–based intervention for smoking cessation (mCessation), one of the first areas to use text messaging as a stand-alone health program with good success—and one that our team have been involved in for 25 years [9].

The Example of mCessation

Our foray into this work began when Prof Anthony Rodgers noticed that people instinctively reached for either their cigarettes or their mobile phones after a meal, and the idea of replacing one with the other to help people quit smoking was born. Professor Rodgers started piecing together funding (including funding from the National Heart Foundation of New Zealand, the Cancer Society of New Zealand, Vodafone New Zealand, and Alcatel) and a team who used the then-standard "brief intervention" training to turn what cessation counselors would say into short, snappy text messages (with a 160-character maximum for SMS text messages at the time). This eventually led to the first published RCT of a health intervention, Stop Smoking by Mobile Phone (STOMP), delivered solely by text messages in 2005 [9]. This national RCT of 1100 people across New Zealand demonstrated a doubling of quit rates compared with no support regardless of age, socioeconomic status, geographic location, or level of addiction. It was important to show that STOMP was as effective for Māori—the Indigenous population of New Zealand—as non-Māori due to the high smoking rates and health consequences in this population, as well as the obligations of the health service to uphold Te Tiriti o Waitangi (the Treaty of Waitangi) [10]. Indeed, the intervention was shown to be as effective for Māori as non-Māori at increasing short-term (6 weeks) self-reported quit rates.

This study represents the start of a large program of work in mCessation that continues today. The first steps were working with the New Zealand Ministry of Health and the Quit Group, a New Zealand service delivering smoking cessation programs, to implement the intervention as a national smoking cessation program. Real-world evaluations of this national service found it achieved quit rates comparable to the National Quitline but was less expensive to deliver [11]. At the same time there was an international collaboration with the London School of Hygiene & Tropical Medicine (LSHTM) to adapt the messages for a UK population, pilot test [12], and then conduct a full RCT across England (funded by the UK Medical Research Council, Cancer Research UK, and Primary Care Research Networks). The text-messaging intervention, txt2stop, was delivered from Auckland, New Zealand, to 2915 intervention group participants recruited by the LSHTM [13]. This high-quality RCT won the Royal College of General Practitioners paper of the year award in 2011 [14]. The trial repeated New Zealand's findings of doubling quit rates and included biochemical verification of quitting and a health economic analysis [15] that showed it to be one of the most cost-effective health interventions and potentially cost-saving for the National Health Service.

From there, the number of studies increased, and we conducted the first (and following) Cochrane systematic reviews of the use of mobile phones for smoking cessation [16-19]. This confirmed the short-term effectiveness of text message program trials in self-reported quitting. The most recent revision of the Cochrane review, completed in 2019, found moderate-certainty evidence that text message–based smoking cessation interventions result in greater quit rates at 6-12 months than minimal smoking cessation support, as well as interventions that use text message–based interventions in addition to other forms of support, compared to other forms of support alone [19]. Studies have also documented promising results for populations at higher risk for adverse health outcomes, such as pregnant smokers [20] and smokers who are HIV positive [21], as well as for smokers in low- and middle-income countries [22-24].

XSL•FO
RenderX

In 2015 the World Health Organization (WHO) picked up mCessation as the first program in its "Be He@lthy, Be Mobile" global initiative with the International Telecommunication Union, citing the demonstrated effectiveness of studies like STOMP and txt2stop as a rationale for scaling up mobile phone–based support for smoking cessation [25]. The handbook provides a detailed work plan and set of messages from the US National Cancer Institute. The WHO is working directly with several countries (eg, Burkina Faso, China, and the Philippines) to help them implement large-scale smoking cessation programs [26]. One example is India, which implemented a nationwide mCessation program in 2015 and saw a self-reported quit rate of 19% and a quit attempt rate of 66% among registered subscribers [27].

There have been many more studies collated in updates of the Cochrane reviews. These have tended to show the same results over time and support further adoption of large-scale mCessation programs, such as those offered by the US National Cancer Institute [28,29]. Moreover, despite newer technologies in this space, text messaging–based smoking cessation has not been surpassed by other forms of digital health, such as app-based interventions—though more research is needed to adequately compare the relative effectiveness of digital modalities for smoking cessation [19,30].

But the success of text messaging hasn't just been limited to smoking cessation. Other areas of health have caught on, with success shown in studies assessing text messaging to support medication and appointment adherence [31,32], maternal health [33], diabetes management [34], sexual health [35,36], cancer screening [37] and blood pressure control [38], to name a few.

So Why Is the Simple Text Message So Useful for the Delivery of Health Interventions?

The reasons for this are likely multifaceted but may include text messaging's reach into nearly all population groups, its timeliness, its seamless integration into daily habits (eg, checking one's mobile phone), its alignment with key behavior change techniques, and its ease of adaptation and personalization.

Text messaging remains the most equitable way of reaching people regardless of type of phone, socioeconomic status, or geographic location. In much of the world, it is free to receive text messages and a smartphone is not required to receive a text message. Further, recipient replies to text messages can be "zero-rated" by health program providers so that they don't cost the individual, ensuring access across socioeconomic groups. Even so, evidence shows that text message campaigns are a cost-effective way to reach people [15,39].

Mobile phones have become integral to everyday life for the majority of people on the planet. Furthermore, people are remarkably well trained to look at their text messages within minutes of receiving them, and unlike notifications from apps or platforms, they do not require high levels of digital literacy and are relatively difficult to disregard. With a device so integral to daily life, using this modality for health delivery and health promotion and behavior change interventions makes sense.

Text messages align well with behavior change techniques that have been tried and tested over time [40]. Unprompted, brief text messages are useful as reminders of the user's behavior change goal, providing positive reinforcement of engagement with health behaviors, delivering health information in simple and digestible bite-sized amounts, and providing motivation and support without requiring the end user to actively engage or seek further information or support to maintain engagement [41]. It is well established that personalized health information alone is not sufficient to change behavior [42], and therefore the success of text messaging likely lies in the interplay between the content and the delivery mechanism—that is, just the fact that messages can be scheduled to arrive regularly, at key times or randomly, and in one word or glance they can be associated with the behavior in question. Indeed, prior research has shown that merely their ability to remind people of the targeted behavior change goal is highly valued by users [43]. Beyond the content and delivery mechanism, other intervention characteristics (eg, duration and dosage) may also be key factors in the success of text messaging interventions. Further, there is growing evidence for the positive impact of participatory design methods in the development of text messaging interventions (eg, Dobson et al [44]) that allow for intervention characteristics to be individually tailored.

The nature of text message–based interventions makes them ideal for the delivery of culturally tailored and personalized interventions, which has been identified as a key element of effective technology-based smoking cessation programs [30,45]. They can be adapted for different populations and contexts, translated into multiple languages, and personalized with peoples' names and key information. An example of this is in maternal health, where a text message–based program developed by our team delivers free, culturally tailored, and personalized information and support to pregnant individuals and their families. Participants can receive texts during pregnancy through the first 2 years of a child's life [46]. The program was originally designed to increase reach to populations in the Auckland region that were being missed by traditional New Zealand maternal support. There are 16 different versions of the Text for MATernal and Child Health (TextMATCH) program, accommodating different languages, cultures, and relationships to the child (mother or family member); this has resulted in a library of over 6000 messages. The program is tailored to the recipient's culture, language preference, and stage of pregnancy or baby's age, and it is personalized to the individual by using information such as the person's name and their baby's name. Since being implemented 10 years ago, over 8000 women and family members have received the program, and it has been adapted for other regions in New Zealand [47].

Many of the advantages of text messaging are also mirrored in app-based messaging, such as on WeChat, WhatsApp, and Facebook Messenger. Although less consistently used across demographic groups, many of these tools are highly used in targeted populations (eg, WeChat in China), making them arguably more useful for health interventions in these populations than traditional text messaging [48]. But there are

XSL•FO
RenderX

a number of differences between app-based messaging and text messaging that raise concerns related to equity and reach, such as the need for data or Wi-Fi, device capacity requirements, data privacy, and, in many cases, the requirement for a subscription to an additional product [49]. Furthermore, text messages are less likely to be muted than app-based messages [50], as people increasingly experience "notification fatigue" from apps [51].

Although the evidence for text messaging appears strong, particularly in smoking cessation, it is important to note that not all studies of text messaging interventions have had consistent positive results (eg, Redfern et al [52]). There is a need for research to understand the mechanisms for success with this modality, including the degree to which the different aspects of the design, including the content, tailoring, and personalization, impact effectiveness and engagement, as well as understanding the most appropriate use cases and contexts for implementation.

The Future of Text Messaging in Health

So the simple text message works—as all those businesses who use it for appointment reminders will tell us and the wealth of published studies show us. But despite the well-established evidence for text messaging in health, there is always pressure to focus on newer technologies. While the true benefit of digital health lies in its future potential for more accessible and personalized health care, many of those in highest need may be unlikely to be able to access these newer technologies, such as wearables, at this time. These continued advancements risk excluding those with limited digital access, literacy, confidence, and motivation, thus exacerbating existing inequities in health outcomes [53,54]. To ensure all people are able to benefit from digital health, it is essential that equity remains at the forefront among digital health researchers, developers, and implementers. While the barriers to digital inclusion are addressed for these newer modalities, text messaging will continue to play a pivotal role in the delivery of equitable digital health tools to communities around the world. We believe that this is due not only to the technology's low cost, vast reach, and low digital literacy requirements, but also its simplicity, which continues to appeal to end users among both consumers of health care and health professionals.

There is significant potential to update the simple text message—not by replacing it but by adding to the power of it. Examples are becoming more frequent, such as multimodal interventions incorporating newer technologies, such as wearables with text messaging [55,56], the development of algorithms to further enhance and personalize traditional text messaging programs [57,58], and algorithm and artificial intelligence–based tools using text messaging to alert or prompt health professionals or patients [59,60]. For example, a member of our team is developing a large language model text messaging program for smoking cessation powered by ChatGPT (OpenAI Inc). Such a program will provide users with both automated text messages and text messages from a digital chatbot coach. These types of interventions show promise in their ability to increase the complexity and sophistication of the content on the back end while maintaining the simplicity and accessibility of text messages for end users. Targeted research assessing the effectiveness of multimodal interventions is needed to understand whether and how these are more effective than interventions relying on text messages alone [61,62].

Conclusion

In conclusion, we recommend that digital health researchers and developers do not overlook this simple yet effective modality in their search for the ultimate technological weapon in health. We challenge researchers to harness the learnings from this unassuming modality in the pursuit of equitable, accessible, and effective digital health solutions.

Conflicts of Interest

LCA received Text2Quit royalties between 2012 and 2022.

References

1. "Merry Christmas": the 30th anniversary of the first text message. Vodafone. URL: https://www.vodafone.co.uk/newscentre/features/merry-christmas-the-30th-anniversary-of-the-first-text-message/ [accessed 2024-12-09]
2. Mobile phone ownership. International Telecommunication Union. URL: https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-mobile-phone-ownership/ [accessed 2024-12-09]
3. Worldwide texting statistics. Vermont State Highway Safety Office. 2018. URL: http://shso.vermont.gov/content/worldwide-texting-statistics [accessed 2024-12-09]
4. Bourque C. Why strategic SMS marketing matters. Forbes Business Council. URL: https://www.forbes.com/sites/forbesbusinesscouncil/2023/11/10/why-strategic-sms-marketing-matters/?sh=64c651754995 [accessed 2024-12-09]
5. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015 Mar 18;36:393-415 [FREE Full text] [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
6. Fjeldsoe BS, Marshall AL, Miller YD. Behavior change interventions delivered by mobile telephone short-message service. Am J Prev Med 2009 Feb;36(2):165-173. [doi: 10.1016/j.amepre.2008.09.040] [Medline: 19135907]
7. Head KJ, Noar SM, Iannarino NT, Harrington NG. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013 Nov;97:41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
8. Neville R, Greene A, McLeod J, Tracey A, Surie J. Mobile phone text messaging can help young people manage asthma. BMJ 2002 Sep 14;325(7364):600 [FREE Full text] [doi: 10.1136/bmj.325.7364.600/a] [Medline: 12228151]

XSL•FO
RenderX

9. Rodgers A, Corbett T, Bramley D, Riddell T, Wills M, Lin R, et al. Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. Tob Control 2005 Aug;14(4):255-261 [FREE Full text] [doi: 10.1136/tc.2005.011577] [Medline: 16046689]
10. Bramley D, Riddell T, Whittaker R, Corbett T, Lin R, Wills M, et al. Smoking cessation using mobile phone text messaging is as effective in Maori as non-Maori. N Z Med J 2005 Jun 03;118(1216):U1494. [Medline: 15937529]
11. Milne K, Bowler S, Li J, Salmon P. Evaluation of the first year of the Txt2Quit service. The Quit Group. URL: https://quit.org.nz/info-resources/FINAL-2008-09-Txt2Quit-evaluation-report-20090731.pdf [accessed 2024-12-09]
12. Free C, Whittaker R, Knight R, Abramsky T, Rodgers A, Roberts IG. Txt2stop: a pilot randomised controlled trial of mobile phone-based smoking cessation support. Tob Control 2009 Apr;18(2):88-91. [doi: 10.1136/tc.2008.026146] [Medline: 19318534]
13. Free C, Knight R, Robertson S, Whittaker R, Edwards P, Zhou W, et al. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. Lancet 2011 Jul 02;378(9785):49-55 [FREE Full text] [doi: 10.1016/S0140-6736(11)60701-0] [Medline: 21722952]
14. Sullivan F. 2011 RCGP and Novartis research paper of the year. Br J Gen Pract 2012 Aug 01;62(601):432-433. [doi: 10.3399/bjgp12x653697]
15. Guerriero C, Cairns J, Roberts I, Rodgers A, Whittaker R, Free C. The cost-effectiveness of smoking cessation support delivered by mobile phone text messaging: Txt2stop. Eur J Health Econ 2013 Oct;14(5):789-797 [FREE Full text] [doi: 10.1007/s10198-012-0424-5] [Medline: 22961230]
16. Whittaker R, Borland R, Bullen C, Lin RB, McRobbie H, Rodgers A. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2009 Oct 07(4):CD006611. [doi: 10.1002/14651858.CD006611.pub2] [Medline: 19821377]
17. Whittaker R, McRobbie H, Bullen C, Borland R, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2012 Nov 14;11:CD006611. [doi: 10.1002/14651858.CD006611.pub3] [Medline: 23152238]
18. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016 Apr 10;4(4):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
19. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. Cochrane Database Syst Rev 2019 Oct 22;10(10):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub5] [Medline: 31638271]
20. Abroms LC, Johnson PR, Leavitt LE, Cleary SD, Bushar J, Brandon TH, et al. A randomized trial of text messaging for smoking cessation in pregnant women. Am J Prev Med 2017 Dec;53(6):781-790 [FREE Full text] [doi: 10.1016/j.amepre.2017.08.002] [Medline: 28982527]
21. Kierstead EC, Harvey E, Sanchez D, Horn K, Abroms LC, Spielberg F, et al. A pilot randomized controlled trial of a tailored smoking cessation program for people living with HIV in the Washington, D.C. metropolitan area. BMC Res Notes 2021 Jan 06;14(1):2 [FREE Full text] [doi: 10.1186/s13104-020-05417-3] [Medline: 33407848]
22. Jiang N, Nguyen N, Siman N, Cleland CM, Nguyen T, Doan HT, et al. Adaptation and assessment of a text messaging smoking cessation intervention in Vietnam: pilot randomized controlled trial. JMIR Mhealth Uhealth 2021 Oct 08;9(10):e27478 [FREE Full text] [doi: 10.2196/27478] [Medline: 34623318]
23. Krishnan N, Gu J, Abroms LC. Mobile phone-based messaging for tobacco cessation in low and middle-income countries: A systematic review. Addict Behav 2021 Feb;113:106676. [doi: 10.1016/j.addbeh.2020.106676] [Medline: 33038676]
24. Whittaker R, Umali E, Tanielu H, McCool J. TXTTaofitapaa: pilot trial of a Samoan mobile phone smoking cessation programme. J Glob Health Rep 2019;3:e2019035. [doi: 10.29392//001C.11965]
25. Be He@lthy be mobile: a handbook on how to implement mTobaccoCessation. World Health Organization; International Telecommunication Union. 2016. URL: https://iris.who.int/bitstream/handle/10665/251719/9789241549813-eng.pdf?sequence=1 [accessed 2024-12-09]
26. mTobaccoCessation. World Health Organization. URL: https://www.who.int/campaigns/world-no-tobacco-day/2021/quitting-toolkit/text-message-support/mtobaccocessation [accessed 2024-12-09]
27. Gopinathan P, Kaur J, Joshi S, Prasad VM, Pujari S, Panda P, et al. Self-reported quit rates and quit attempts among subscribers of a mobile text messaging-based tobacco cessation programme in India. BMJ Innov 2018 Jul 25;4(4):147-154. [doi: 10.1136/BMJINNOV-2018-000285]
28. Prutzman YM, Wiseman KP, Grady MA, Budenz A, Grenen EG, Vercammen LK, et al. Using digital technologies to reach tobacco users who want to quit: evidence from the National Cancer Institute's Smokefree.gov initiative. Am J Prev Med 2021 Mar;60(3 Suppl 2):S172-S184 [FREE Full text] [doi: 10.1016/j.amepre.2020.08.008] [Medline: 33663705]
29. Abroms LC, Boal AL, Simmens SJ, Mendel JA, Windsor RA. A randomized trial of Text2Quit: a text messaging program for smoking cessation. Am J Prev Med 2014 Sep;47(3):242-250 [FREE Full text] [doi: 10.1016/j.amepre.2014.04.010] [Medline: 24913220]
30. Fang YE, Zhang Z, Wang R, Yang B, Chen C, Nisa C, et al. Effectiveness of eHealth smoking cessation interventions: systematic review and meta-analysis. J Med Internet Res 2023 Jul 28;25:e45111 [FREE Full text] [doi: 10.2196/45111] [Medline: 37505802]

XSL•FO
RenderX

31. Ershad Sarabi R, Sadoughi F, Jamshidi Orak R, Bahaadinbeigy K. The effectiveness of mobile phone text messaging in improving medication adherence for patients with chronic diseases: a systematic review. Iran Red Crescent Med J 2016 May;18(5):e25183 [FREE Full text] [doi: 10.5812/ircmj.25183] [Medline: 27437126]
32. Middleton T, Constantino M, McGill M, D'Souza M, Twigg SM, Wu T, et al. An enhanced SMS text message-based support and reminder program for young adults with type 2 diabetes (TEXT2U): randomized controlled trial. J Med Internet Res 2021 Oct 21;23(10):e27263 [FREE Full text] [doi: 10.2196/27263] [Medline: 34524102]
33. Poorman E, Gazmararian J, Parker RM, Yang B, Elon L. Use of text messaging for maternal and infant health: a systematic review of the literature. Matern Child Health J 2015 May;19(5):969-989. [doi: 10.1007/s10995-014-1595-8] [Medline: 25081242]
34. Haider R, Sudini L, Chow CK, Cheung NW. Mobile phone text messaging in improving glycaemic control for patients with type 2 diabetes mellitus: A systematic review and meta-analysis. Diabetes Res Clin Pract 2019 Apr;150:27-37. [doi: 10.1016/j.diabres.2019.02.022] [Medline: 30822496]
35. Rinehart DJ, Leslie S, Durfee MJ, Stowell M, Cox-Martin M, Thomas-Gale T, et al. Acceptability and efficacy of a sexual health texting intervention designed to support adolescent females. Acad Pediatr 2020;20(4):475-484. [doi: 10.1016/j.acap.2019.09.004] [Medline: 31560971]
36. Free C, Palmer M, McCarthy O, Jerome L, Berendes S, Knight M, et al. Effectiveness of a behavioural intervention delivered by text messages (safetxt) on sexually transmitted reinfections in people aged 16-24 years: randomised controlled trial. BMJ 2022 Sep 28;378:e070351 [FREE Full text] [doi: 10.1136/bmj-2022-070351] [Medline: 36170988]
37. Uy C, Lopez J, Trinh-Shevrin C, Kwon SC, Sherman SE, Liang PS. Text messaging interventions on cancer screening rates: a systematic review. J Med Internet Res 2017 Aug 24;19(8):e296 [FREE Full text] [doi: 10.2196/jmir.7893] [Medline: 28838885]
38. Tam HL, Wong EML, Cheung K, Chung SF. Effectiveness of text messaging interventions on blood pressure control among patients with hypertension: systematic review of randomized controlled trials. JMIR Mhealth Uhealth 2021 Sep 22;9(9):e24527 [FREE Full text] [doi: 10.2196/24527] [Medline: 34550078]
39. Burn E, Nghiem S, Jan S, Redfern J, Rodgers A, Thiagalingam A, et al. Cost-effectiveness of a text message programme for the prevention of recurrent cardiovascular events. Heart 2017 Jun;103(12):893-894 [FREE Full text] [doi: 10.1136/heartjnl-2016-310195] [Medline: 28235776]
40. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
41. Abroms LC, Whittaker R, Free C, Mendel Van Alstyne J, Schindler-Ruwisch JM. Developing and pretesting a text messaging program for health behavior change: recommended steps. JMIR Mhealth Uhealth 2015 Dec 21;3(4):e107 [FREE Full text] [doi: 10.2196/mhealth.4917] [Medline: 26690917]
42. French DP, Cameron E, Benton JS, Deaton C, Harvie M. Can communicating personalised disease risk promote healthy behaviour change? A systematic review of systematic reviews. Ann Behav Med 2017 Oct;51(5):718-729 [FREE Full text] [doi: 10.1007/s12160-017-9895-z] [Medline: 28290066]
43. Abroms LC, Heminger CL, Boal AL, Van Alstyne JM, Krishnan N. Text2Quit: An analysis of user experiences with a mobile smoking cessation program. J Smok Cessat 2020 Mar;15(1):23-28 [FREE Full text] [doi: 10.1017/jsc.2019.22] [Medline: 33777241]
44. Dobson R, Stowell M, Tane T, Gardiner C, McCool J, Ni L, et al. Tackling anxiety through innovation: Development and pilot study of an in-the-moment messaging intervention for young adults. Digit Health 2024;10:20552076241283245 [FREE Full text] [doi: 10.1177/20552076241283245] [Medline: 39314815]
45. Bendotti H, Lawler S, Ireland D, Gartner C, Hides L, Marshall H. What do people want in a smoking cessation app? An analysis of user reviews and app quality. Nicotine Tob Res 2022 Feb 01;24(2):169-177. [doi: 10.1093/ntr/ntab174] [Medline: 34460922]
46. Dobson R, Whittaker R, Bartley H, Connor A, Chen R, Ross M, et al. Development of a culturally tailored text message maternal health program: TextMATCH. JMIR Mhealth Uhealth 2017 Apr 20;5(4):e49 [FREE Full text] [doi: 10.2196/mhealth.7205] [Medline: 28428159]
47. Dobson R, Uri A, Whittaker R, Evison K, Umali E, McRobbie H. Is opt-out enrolment acceptable for low-risk digital health services? J Prim Health Care 2022 Dec;14(4):368-371. [doi: 10.1071/HC22088] [Medline: 36592779]
48. Su Z, Wei X, Cheng A, Zhou X, Li J, Qin R, et al. Utilization and effectiveness of a message-based tobacco cessation program (mCessation) in the Chinese general population: longitudinal, real-world study. J Med Internet Res 2023 May 02;25:e44840 [FREE Full text] [doi: 10.2196/44840] [Medline: 37129934]
49. McCarthy OL, Mavodza C, Chikwari CD, Dauya E, Tembo M, Hlabangana P, et al. Adapting an evidence-based contraceptive behavioural intervention delivered by mobile phone for young people in Zimbabwe. BMC Health Serv Res 2022 Jan 25;22(1):106 [FREE Full text] [doi: 10.1186/s12913-022-07501-9] [Medline: 35078457]
50. Smith AC, Fowler LA, Graham AK, Jaworski BK, Firebaugh M, Monterubio GE, et al. Digital overload among college students: implications for mental health app use. Social Sciences 2021 Jul 21;10(8):279. [doi: 10.3390/socsci10080279]

XSL•FO
RenderX

51. Travers M. A psychologist offers 2 solutions to "digital fatigue" and "notification overload". Forbes. URL: https://www.forbes.com/sites/traversmark/2023/09/09/a-psychologist-offers-2-solutions-to-digital-fatigue-and-notification-overload/?sh=75eb11bc41b9 [accessed 2024-12-09]
52. Redfern J, Tu Q, Hyun K, Hollings M, Hafiz N, Zwack C, et al. Mobile phone text messaging for medication adherence in secondary prevention of cardiovascular disease. Cochrane Database Syst Rev 2024 Mar 27;3(3):CD011851 [FREE Full text] [doi: 10.1002/14651858.CD011851.pub3] [Medline: 38533994]
53. Lyles CR, Wachter RM, Sarkar U. Focusing on digital health equity. JAMA 2021 Nov 09;326(18):1795-1796. [doi: 10.1001/jama.2021.18459] [Medline: 34677577]
54. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52 [FREE Full text] [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
55. Whittaker R, Dobson R, Candy S, Tane T, Burrowes K, Reeve J, et al. Mobile pulmonary rehabilitation: feasibility of delivery by a mobile phone-based program. Front Comput Sci 2021 Mar 23;3:1-9. [doi: 10.3389/fcomp.2021.546960]
56. Blok AC, Sadasivam RS, Amante DJ, Kamberi A, Flahive J, Morley J, et al. Gamification to motivate the unmotivated smoker: the "take a break" digital health intervention. Games Health J 2019 Aug;8(4):275-284 [FREE Full text] [doi: 10.1089/g4h.2018.0076] [Medline: 31219347]
57. Celano CM, Massey C, Long J, Kim S, Velasquez O, Healy BC, et al. An adaptive, algorithm-based text message intervention to promote health behavior adherence in type 2 diabetes: treatment development and proof-of-concept trial. J Diabetes Sci Technol 2023 Mar;17(2):364-373 [FREE Full text] [doi: 10.1177/19322968211065067] [Medline: 34911398]
58. Kim EK, Brown LA, Seltzer EK, Hartzell-Leggin D, Borodyanskaya YL, Andy UU, et al. Development of a patient-centered text message-based platform for the self-management of interstitial cystitis/bladder pain syndrome symptoms. Neurourol Urodyn 2023 Feb;42(2):510-522 [FREE Full text] [doi: 10.1002/nau.25115] [Medline: 36519701]
59. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med 2019 Nov;47(11):1485-1492 [FREE Full text] [doi: 10.1097/CCM.0000000000003891] [Medline: 31389839]
60. Lowres N, Duckworth A, Redfern J, Thiagalingam A, Chow CK. Use of a machine learning program to correctly triage incoming text messaging replies from a cardiovascular text-based secondary prevention program: feasibility study. JMIR Mhealth Uhealth 2020 Jun 16;8(6):e19200 [FREE Full text] [doi: 10.2196/19200] [Medline: 32543439]
61. Hwang M, Jiang Y. Personalization in digital health interventions for older adults with cancer: A scoping review. J Geriatr Oncol 2023 Nov;14(8):101652. [doi: 10.1016/j.jgo.2023.101652] [Medline: 37866009]
62. Hrynyschyn R, Prediger C, Stock C, Helmer SM. Evaluation methods applied to digital health interventions: what is being used beyond randomised controlled trials?-a scoping review. Int J Environ Res Public Health 2022 Apr 25;19(9):5221 [FREE Full text] [doi: 10.3390/ijerph19095221] [Medline: 35564616]

Abbreviations

LSHTM: London School of Hygiene & Tropical Medicine
RCT: randomized controlled trial
STOMP: Stop Smoking by Mobile Phone
TextMATCH: Text for MATernal and Child Health
WHO: World Health Organization

Edited by G Eysenbach; submitted 30.04.24; peer-reviewed by Y Bartlett, P Cupertino, J Redfern; comments to author 25.11.24; revised version received 02.12.24; accepted 02.12.24; published 17.12.24.

Please cite as:
Dobson R, Whittaker R, Abroms LC, Bramley D, Free C, McRobbie H, Stowell M, Rodgers A
Don't Forget the Humble Text Message: 25 Years of Text Messaging in Health
J Med Internet Res 2024;26:e59888
URL: https://www.jmir.org/2024/1/e59888
doi:10.2196/59888
PMID:

©Rosie Dobson, Robyn Whittaker, Lorien C Abroms, Dale Bramley, Caroline Free, Hayden McRobbie, Melanie Stowell, Anthony Rodgers. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Looking Back on Digital Medical Education Over the Last 25 Years and Looking to the Future: Narrative Review

Oluwadamilola Ogundiya[1], iBSc; Thahmina Jasmine Rahman[1], iBSc; Ioan Valnarov-Boulter[1], iBSc; Tim Michael Young[1], MBBS, PhD, PFHEA

Queen Square Institute of Neurology, University College London, London, United Kingdom

Corresponding Author:
Tim Michael Young, MBBS, PhD, PFHEA
Queen Square Institute of Neurology
University College London
No/7 Queen Square
London, WC1N 3BG
United Kingdom
Phone: 44 203 1082781
Email: t.young@ucl.ac.uk

Abstract

Background: The last 25 years have seen enormous progression in digital technologies across the whole of the health service, including health education. The rapid evolution and use of web-based and digital techniques have been significantly transforming this field since the beginning of the new millennium. These advancements continue to progress swiftly, even more so after the COVID-19 pandemic.

Objective: This narrative review aims to outline and discuss the developments that have taken place in digital medical education across the defined time frame. In addition, evidence for potential opportunities and challenges facing digital medical education in the near future was collated for analysis.

Methods: Literature reviews were conducted using PubMed, Web of Science Core Collection, Scopus, Google Scholar, and Embase. The participants and learners in this study included medical students, physicians in training or continuing professional development, nurses, paramedics, and patients.

Results: Evidence of the significant steps in the development of digital medical education in the past 25 years was presented and analyzed in terms of application, impact, and implications for the future. The results were grouped into the following themes for discussion: learning management systems; telemedicine (in digital medical education); mobile health; big data analytics; the metaverse, augmented reality, and virtual reality; the COVID-19 pandemic; artificial intelligence; and ethics and cybersecurity.

Conclusions: Major changes and developments in digital medical education have occurred from around the start of the new millennium. Key steps in this journey include technical developments in teleconferencing and learning management systems, along with a marked increase in mobile device use for accessing learning over this time. While the pace of evolution in digital medical education accelerated during the COVID-19 pandemic, further rapid progress has continued since the resolution of the pandemic. Many of these changes are currently being widely used in health education and other fields, such as augmented reality, virtual reality, and artificial intelligence, providing significant future potential. The opportunities these technologies offer must be balanced against the associated challenges in areas such as cybersecurity, the integrity of web-based assessments, ethics, and issues of digital privacy to ensure that digital medical education continues to thrive in the future.

(J Med Internet Res 2024;26:e60312) doi:10.2196/60312

KEYWORDS

digital health; digital medical education; health education; medical education; mobile phone; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Background

The last 25 years have seen striking digital technological advances in the internet, now accessed by many of us daily in everyday life through multiple devices, including smartphones. This period also provides an interesting unit of time over which changes in digital medical education can be viewed, as it spans a very significant portion of an individual health worker's career. A medical student who started their studies in 1999 may now be an experienced consultant or general practitioner (GP). What purpose does looking back at such a time frame have for us now? While research is typically forward looking, knowledge acquisition is not an inexorable upward curve of progress. Knowledge can be lost unless we look for it. Without deliberate reflection, valuable insights can be lost in the rapid pace of discovery.

Historically, knowledge acquisition has not always followed an inexorable upward curve of progress. In 1901, a strange device, the Antikythera mechanism, was retrieved from an ancient sunken Roman shipwreck just off the Greek island of Antikythera [1]. This complicated, geared analog computer was capable of predicting planetary movements and eclipses. It was >2000 years old with no known precursor, and nothing near its complexity would be produced for >1000 years. The invention and its secrets were lost to history in the shipwreck. In our modern era, despite vast stores of information, we may still face similar risks. For example, entering the search term *digital medical education* on Google Scholar, restricted to 2023 to 2024, yields >1300 articles. Thus, new discoveries run the risk of being submerged under a sea of data, justifying reviews that provide a contemporaneous overview of digital medical education and identify areas for further progress, such as this one. Therefore, it is necessary to examine the ever-changing state of digital medical education, including the emerging technologies that help define this era.

While digital health concerns the use of technology for improving health and well-being [2], the term *digital medical education*, as used in our study, transcends this definition to focus instead on the use of digital technologies in medical education. In contrast to traditional teaching methods, digital medical education leverages digital technologies, such as teleconferencing and learning management, to more efficiently create and manage web-based courses and allow the educational experience to span beyond the traditional four walls of the classroom.

To contextualize the evolution of digital medical education, it is instructive to briefly consider the terminology underpinning it. Terms such as *distance learning*, *web-based learning*, *e-learning*, and *digital learning* have often been used interchangeably, but there are important differences, especially when considering their origins and development. For example, e-learning, although it may be carried out remotely or in person with a teacher, does not always necessitate an internet connection, as demonstrated by offline computer-based learning, which was prevalent in the late 20th century. By contrast, web-based learning is inherently dependent on a working internet connection. Originating from the digital encoding of signals, digital learning, initially described computer-based technologies, though its context expanded with the widespread popularity of digital televisions around the year 2000. Similarly, while the etymology of telemedicine implies the provision of medical care remotely [3,4], the term is also extensively applied to medical education. Looking back at the last few decades, we can see that such terms continue to be in a state of flux regarding the popularity of use. For example, in Figure 1 [5], the inflection point of popularity for the term *telemedicine* in the late 1990s coincides with the first appearance of the term *e-learning*, suggesting the latter may have evolved in part as a simple contraction of the former.

Figure 1. Frequency of the use of phrases related to e-learning by year. Google Ngram illustration showing the frequency of use by year as a percentage of appearance in the extensive corpus of scanned Google Books [5].

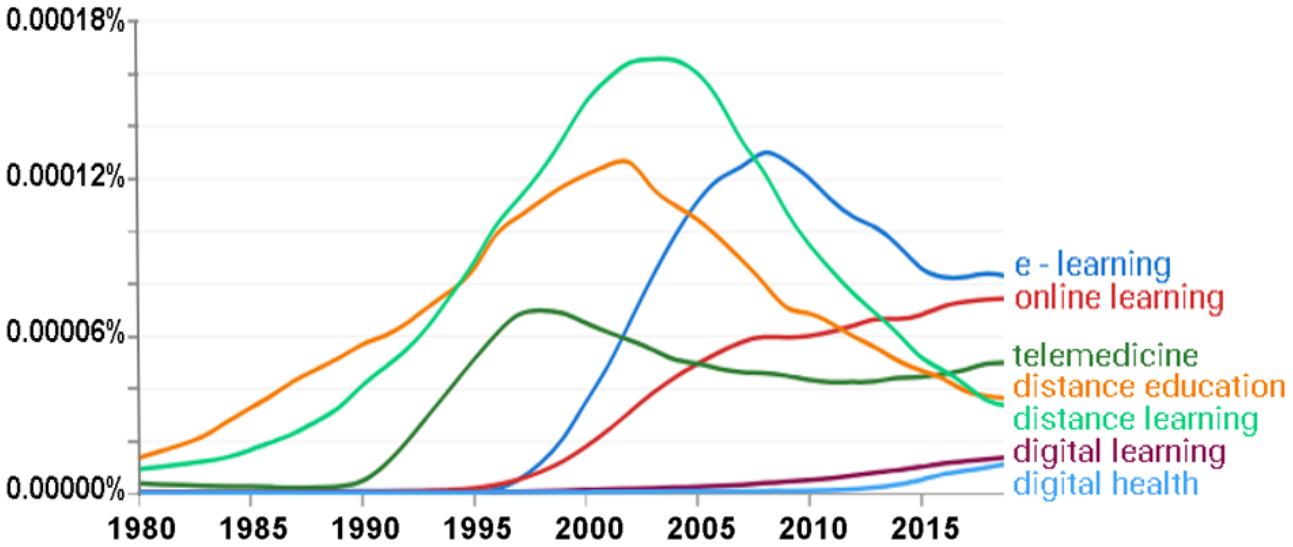

XSL•FO
RenderX

The effectiveness of digital medical education broadly seems to have equivalence to traditional face-to-face approaches [6]. Allied with this, the enormous and expanding economic importance is apparent. The global e-learning market, estimated at US $6.6 billion by the International Data Corporation in 2002, is believed to have grown to >30 times that value since then [7,8].

Distance learning, where the educator and learner are significantly separated in space, has a long background history of use. By the late 19th century, major universities provided degrees via distance learning, and by the early 20th century, the development of radio and the emergence of television allowed for new opportunities for distance learning. Programmed Logic for Automated Teaching Operations (PLATO) was the first computer-mediated learning environment, and thus, the first medium used for digital learning [9]. PLATO was developed as a working model in 1960, with a distance learning (>25 miles) app demonstrated using phone connections by the next year [9]. Learners at the University of Illinois using PLATO were able to not only create but also edit digital images, helped in part by the use of touch-sensitive plasma screens in 1966 [9]. The students taught via PLATO, in its very first simulation (a health education case of acute myocardial infarction), outperformed those taught with conventional face-to-face methods, and by the mid-60s, the first fully web-based university course was developed [9]. By 1984, the Open University, established in the United Kingdom in the late 1960s, provided education to 60,000 undergraduate students from their homes [10]. Through the use of the Cyclops computer system, students and tutors could interact via a teleconferencing network, displaying great success [10,11].

The evolution of digital learning accelerated with the introduction of the US Advanced Research Projects Agency Network (ARPANET) in 1969, which was a packet-switched computer network that linked different computers together [12]. Although the absolute precedence of ARPANET has been disputed, the development of ARPANET laid the foundations for what would become the internet [13]. Initially, ARPANET developed into facilitating connections between networks ("internets") and then to *the* internet in the early 1980s. Subsequent global reach was achieved through the Web from the early 1990s onward, with web servers and web pages rapidly increasing in number. By the start of our designated review period, the Web (retrospectively termed Web 1.0) was characterized by its fixed search and read-only capabilities [14]. This initial static phase laid the groundwork for future innovations in digital education.

Digital Medical Health in 1999

Before we cover the progress in digital medical education over the last 25 years, it would be helpful to first establish the baseline in 1999. Formal continuing medical education (CME) has been an important cornerstone in physician education for >50 years and can provide a useful surrogate marker when considering how widespread and extensive digital medical education is. By 1999, approximately 277,000 physicians in the United States (about a third of the total) were using the web, and 19% of CME credits were earned digitally, including CD-ROMS and web access [15]. While this suggests that digital delivery of CME was already quite common in 1999, this was largely delivered in a passive form using simple text or graphics. Marked global disparities of access were apparent; another article published that year found that the overwhelming majority of medical school websites mentioning CME were based in the United States, and only a quarter of identified CME websites offered web-based CMEs at that time [16]. Technical barriers, such as software and server problems, were common, with approximately 5% of the identified sites found to be inaccessible due to this [16]. Even for the accessible web-based CME sites, only a few gave immediate feedback to the user [16].

In an article in the first issue of JMIR, we found that the UK National Database of Telemedicine was launched just 2 months before 1999, providing information on a web page to anyone working in the field of telemedicine, including in education, even outside of the United Kingdom [17]. A study published in 1999 assessed the use of this web page and reported that there were only 120 visits a week to this national web page, despite it having been publicized extensively [17]. Rather than the real-time web-based learner feedback that we are used to today, the authors described an analysis of keywords used in search engines in an attempt to make their website easier for users to find [17].

Patient digital medical education also existed in 1999. For example, GPs in Scotland described their own customized patient-oriented websites [18]. The resultant internet documents were then published on servers located at their local health board, allowing viewing via National Health Service Scotland intranet servers. The educational material could also be viewed by patients via public servers if they had an internet connection. For these Scottish GPs, a template had to be modified specifically for their GP practice website, and a facilitator was required to be available on-site to help with any difficulties [18]. Elsewhere, limitations of digital medical education in 1999 were noted, including concerns about the great variability of graphics, response time, and the lack of visualization between students and educators [19,20].

Such articles were largely published as inspiring exceptions rather than reflecting the norm at that time. Matheson [15] noted in 1999 that "Only a small fraction of the revolutionary changes in medical education made possible by the World Wide Web have yet been realized." Medical educators of that time considered the potential opportunities of the digital environment to rethink course structure and duration [21]. However, concerns about the unstoppable pace of change were also expressed as a "tsunami of change about to break over our heads" [21]. Such hopes and fears from the past can speak to us now as we look ahead to the rapid expansion of technologies, such as artificial intelligence (AI) and the metaverse, and in turn, contemplate the future of digital medical education.

Our study broadly encompassed digital health including web-based learning and e-learning and the concept of distance learning. In addition, studies involving blended (also known as hybrid) delivery, where health education is given as a mix of digital (typically distance) learning and traditional face-to-face

XSL•FO
RenderX

teaching, were included. The scope included publications covering a broad range of health learners.

These groups were selected because they represent key stages in the continuum of medical education and reflect cohorts where the impact of digital medical education has been particularly significant. Furthermore, they include the professionals who play a pivotal role in the construction and delivery of health courses themselves and individuals who, as a result of rapid advancements in health care technologies, have become increasingly reliant on digital tools for education and the delivery of health care to patients. The review did not specifically focus on digital pedagogies that were beyond the scope of our study and that have been covered elsewhere [22].

Aims

This study aimed to perform a literature search to identify relevant publications that covered digital medical education and its associated technologies over the last 25 years. The further aims were to analyze the results with a thematic approach as a narrative review, including hypotheses on future possible developments in digital medical education. This approach was used to provide an accessible overview of the evolving and wide-ranging field of digital medical education.

Methods

In this narrative review, we considered technological and educational factors that have developed in digital medical education over the last 25 years using literature searches. The focus was on advancements in digital tools and teaching strategies that have influenced the effectiveness and delivery of digital medical education. Textbox 1 outlines the inclusion and exclusion criteria used for the literature search.

Literature reviews were conducted to support this review using PubMed, Web of Science Core Collection, Scopus, Google Scholar, and Embase. The first 3 databases offer a long search string length, while Embase has recently been shown to be associated with a high number of unique references in reviews. The biomedical and health science emphasis of PubMed and Embase is balanced by the broader scientific coverage of Web of Science, Scopus, and Google Scholar [23,24].

The original literature searches were conducted between April 16, 2024, and August 31, 2024. The review process was iterative, allowing the inclusion of additional sources as they emerged during the study search period with the specific search terms agreed by the authors. The collected literature was analyzed thematically to identify common themes and key developments, which are described and highlighted in this paper.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for literature searches used in this review.

Inclusion criteria

- Participants from original articles
  - Medical students
  - Trainee physicians
  - Physicians undergoing continuing professional development
  - Nurses
  - Paramedical staff
  - Patients or members of the public
- Topics
  - Digital medical education
  - Health education
  - Medical education
  - e-Learning
  - Telemedicine, distance learning, and web-based learning
- Publication dates of articles: from 1999 to 2024
- Language of publication: English
- Search terms
  - “health education”
  - “digital medical education”
  - “medical education”
  - “e-learning”
  - “distance learning”
  - “learning management systems”
  - “artificial intelligence”
  - “big data analytics”
  - “augmented reality”
  - “virtual reality”
  - “metaverse”
  - “mobile health”
  - “COVID-19”
  - “cybersecurity”
  - “ethics” and “digital health”
  - “ethics” or “digital health”
  - “telemedicine”
  - “online learning”

Exclusion criteria

- Topics: focus on digital pedagogy
- Publication dates of articles: before 1999
- Language of publication: non-English language

XSL•FO
RenderX

Results

Overview

The literature search yielded 128 articles, which were used in the study to develop the narrative review. The results aligned with at least 1 of 8 different themes. These were learning management systems (LMSs); telemedicine (in digital medical education); mobile health (mHealth); big data analytics (BDA); the metaverse, augmented reality (AR), and virtual reality (VR); the COVID-19 pandemic; AI; and ethics and cybersecurity. The results are presented and analyzed sequentially for each of these themed subgroups. For each theme, the results are used to hypothesize the potential relevance for the further development of digital medical education in the near future.

LMSs and Digital Health Courses

LMSs, or web-based learning environments, consist of specialized e-learning software that facilitates comprehensive course management [25]. While the concept of LMS is not new, its use in higher education, including digital medical education, has markedly expanded over the last 25 years. LMS allows the automated coordination of digital educational courses and associated resources to deliver education in a web-based environment [25]. Practically, educational material can be produced, posted, and edited together with other resource materials, such as electronic journals [26]. In turn, the learners can submit work, which is then marked, the results collated, and feedback sent to the learners via the same platform. Learners increasingly rely on a broad spectrum of e-learning tools to supplement their education. LMSs provide centralized platforms that enable more efficient management of such educational content and digital resources, and by 2016, they were considered key components in digital medical education [25]. LMSs serve as a bridge between traditional teaching strategies and modern e-learning technologies, and their use has allowed entire courses to be structured and developed, providing the potential for large-scale open access to health education through massive open online courses (MOOCs) as well [27]. MOOCs were developed during the US Open Education Resources movement in the early 2000s [27]. They incorporate a decentralized open design, encouraging interaction between learners, and have seen rapid growth in the last 16 years [27]. Since 2008, the launch of MOOCs has built on the improvements seen in Web 2.0. MOOCs underwent rapid expansion from 2012 onward, with both Udacity and edX (education platforms led by an American organization, and Harvard and Massachusetts Institute of Technology, respectively, offering internet-based courses and degrees) founded that year [28].

The parallel evolution of the Web to Web 2.0 in 2004 has been of central importance to the further development of LMS, which allowed for read-and-write capabilities rather than read-only capabilities, making it much more interactive [14]. Initially, it was feared that this new, participatory version of the Web might threaten the dominance of LMS in digital education. Instead, this proved to be a key progression from a digital medical education perspective, improving dynamic communication among learners, their web-based courses, and their tutors. The gradual introduction of Web 3.0 (also termed Web3) over the last decade has been associated with a less centralized structure and allows greater use of more complex semantics by users without having to break these down into simpler components [14,28]. Web 3.0 also offers a more flexible schema of data storage and interactions between data [14]. While Web 1.0 was considered “read-only” and Web 2.0 was “read-write,” Web 3.0 has been described as “read-write-own,” allowing some degree of digital ownership [28].

The development of LMS from the early 2000s reflects academic concerns that web-based courses significantly increase educators’ workload and could only effectively handle ≤20 students per course [29]. However, LMSs have since evolved to support a larger number of students remotely, significantly increasing the capacity for distance learning while simultaneously reducing the overall workload for instructors. Examples of this include Canvas (created in 2002), Blackboard (2004), and Moodle (2011), the latter being an example of open-source software widely used in digital medical education [30]. Being open source, Moodle can run on most operating systems, including Macintosh, Windows, and Linux [26]. Likewise, Moodle’s flexibility allows educators to customize it to reflect the complexity and diversity of medical curricula, and alongside its ease of integration with external programs and cost-effectiveness, it is often the LMS of choice in medical education [31]. These platforms have revolutionized medical education by providing flexible, personalized, and interactive learning spaces that offer discussion forums for students, the setting of assessments, and a communication channel for students [31].

Videoconferencing using LMS allows for synchronous interactions across continents and real-time health care student feedback on questions, teaching, and even on the course itself [32]. The development of teleconferencing software, such as Skype (founded in 2003), Google Hangouts (2005), Zoom (2011), and Microsoft Teams (2017), has revolutionized the delivery of health education, offering increasingly sophisticated tools for remote, live, and 2-way interactions [29]. Beyond sharing the live classroom experience at a distance, teleconferencing has allowed for live streaming surgical procedures, grand rounds, and even the remote proctoring of medical assessments [32,33]. Teleconferencing software is primarily used for videoconferencing; however, other modalities include audioconferencing and audio-graphic teleconferencing**.** Disruptions to medical education practices during the COVID-19 pandemic highlighted the demand for videoconferencing tools, accelerating their adoption, which will be covered more comprehensively in the *Impact of the COVID-19 Pandemic on Digital Medical Education* subsection [34]. The use of videoconferencing and LMS digital platforms in digital medical education has evidence of effectiveness, with a systematic review published by the World Health Organization in 2015 finding that e-learning in medical education from the year 2000 was overall equivalent in terms of skill and knowledge acquisition compared to traditional face-to-face courses [35].

LMSs also offer health care students the ability to customize and adapt their educational experience by pursuing extra content for topics of interest or by engaging with content that better caters to their personal learning style [36]. These additionally

XSL•FO
RenderX

provide flexibility for medical students by allowing the use of asynchronous learning materials, such as case studies, prerecorded lectures, and question banks, accommodating varied lifestyles and schedules [37]. These characteristics improve satisfaction and engagement with medical education when used alongside traditional teaching methods while also helping to build skills in autonomous regulation and independence when trying to meet learning goals [38]. Furthermore, it is easier to update and restructure content on LMS, keeping medical educational materials more current than traditional syllabi and textbooks, which cannot be readily reviewed regularly.

The interactive dimension of many health courses places a particular emphasis on enhancing interactions with students despite the web-based delivery. The evolution of LMS has played a significant role in the practice of blended learning, which has been adopted into the curriculum of many institutions for medical education [39]. Blended learning has been demonstrated to yield significantly better knowledge acquisition in health education compared to traditional techniques [40]. The use of e-learning may actually be more effective in the accrual of factual knowledge than the use of traditional approaches [41]. More clinical scenarios, such as examinations and communication skills, have traditionally been taught through direct patient interactions or simulation teaching, either with manikins or standardized patients (typically actors). However, more recently it has been shown that the use of LMS platforms can also effectively create web-based simulation training in health education [42].

Digital question banks on LMS and participatory e-learning content have become some of the most popular revision methods for medical students, with studies demonstrating that these may be used by up to 92% of students [43]. Many use the concept of gamification by using instantaneous feedback from quizzes and the idea of “streaks,” which encourage students to maintain daily engagement with the platform [44]. These features are thought to increase student motivation and make it easier to track and visualize progress. Furthermore, digital platforms may use competitive aspects, such as leaderboards, badges, points, and challenges, allowing medical students to compare their progress with each other and may make learning more enjoyable [45]. These e-learning materials have also demonstrated efficacy in the continued education of health professionals without any evidence that these methods of learning affect professional behaviors or patient outcomes [46]. However, a scoping review evaluating digital medical education for medical students between 2000 and 2019 found that such courses varied in terms of delivery and assessment, were mainly delivered as electives, and often lacked robust evaluation [47].

From a wider perspective, digital education platforms have demonstrated the potential to help reduce geographic and financial barriers to medical education by allowing educators to share knowledge across borders. There is evidence that LMS can improve educational opportunities for medical students in resource-constrained lower-income countries [48]; however, inequalities in internet connectivity, technology, and language barriers could potentially worsen the disparities in health knowledge. It is important that e-learning content aligns with medical curricula and is standardized. This is achieved by developing quality assurance and guidelines for these platforms through collaborations with accrediting institutional bodies [49].

Moreover, evidence suggests that medical students find e-learning platforms to be effective in improving understanding and are associated with high satisfaction scores [50]. Health care student satisfaction overall with LMS use in medical education appears to be high, with >94% of medical and nursing students recently showing satisfaction with the LMS used in China [50]. Therefore, it is likely that these digital platforms will continue to play an important role in the educational toolkit of modern-day health care students.

Looking forward, several areas for the development of existing LMSs used in medical health education can be identified and broadly classified into either technical or use opportunities. On the technical side, the interface of LMS currently used can be potentially confusing to both the learners and the educators themselves. More intuitive interface development of LMS interfaces could be used to counter this potential challenge [31]. Increasing the use of a single sign-in for both LMS and institutional emails would be in line with student feedback and could help with communication via the LMS themselves [25]. In a similar manner, installation and support systems to help users optimize their use of Moodle as a free open-source platform have been identified as areas that might practically be improved to help educators [31]. The potential security vulnerabilities of LMS, particularly for open-source versions such as Moodle, are discussed in more detail in the *Ethics and Cybersecurity in Digital Health Platform* section.

The use of LMS is another area where future developments have been anticipated [31]. Although the original development of LMS such as Moodle had incorporated social constructivist pedagogy [26], there has been a danger that the practical administrative aspects of LMS could predominate. Since the COVID-19 pandemic, this has recently been challenged as an area requiring optimization to move forward [51]. Looking forward, the identities of the 3 main components of a working LMS, namely the learner, the educator, and the LMS system itself, may need to change to establish authentic web-based teaching and learning. In a recent study regarding the LMS Blackboard, only a few of the 53 core study sites analyzed appeared to be at higher levels of identity changes. In essence, most sites were used mainly as digital depository with very limited or infrequent communication from the educators. Such studies have championed the change of LMS from being a predominant interaction between the learner and the LMS to a learning community with educators and learners repositioned as coparticipants [51].

Telemedicine and Digital Medical Education

Telemedicine allows the transfer of medical services from health professionals where distance may have otherwise been a limiting factor, and it has been used across specialties [52]. In addition, telemedicine has an important role in health education by providing learners an insight into real-world telehealth scenarios and training them in the appropriate medical skills. It is this educational aspect of telemedicine that was considered in this section. The main methods for the delivery of telehealth training in digital medical education range from traditional lectures to

XSL•FO
RenderX

simulation-based education [53] and cover a broad range of skills. In the training for telesurgery, for example, medical students can observe procedures and enter associated case-based discussions. In the first decade of the new millennium, despite becoming part of the medical curriculum in France in 2009, teaching telemedicine to medical students elsewhere was still limited [54]. Beyond medical schools, professionals have also seen the benefits of teleconferencing in telemedicine; for example, as a result of distance learning being available, continued professional development opportunities are more accessible to professionals who would normally have been limited by distance, financial burden, or time [33]. Training health professionals often requires more active manipulation of tools with live feedback from remote educators, a form of interactive and immersive teaching method that is generally well tolerated [54]. In addition to ensuring that users are familiar with how to use telemedicine to deliver effective medical care [55], learners should be trained on how best to maintain patient confidentiality and privacy in telemedicine [52,56].

In the late 1990s, training in telemedicine was in its developmental stages [57,58], with its limited use in training at the undergraduate and postgraduate levels [58]. Where training was available, programs were aimed toward practicing medical professionals and were often specific to a form of telemedicine [56,58]. In addition to direct training, telemedicine has been used to provide telementoring, for example, to surgeons in training [59]. In recent years, more medical training curricula have shifted to include training and education on telemedicine [52,58]. By 2019, just before the COVID-19 pandemic, 25% of American medical schools included telemedicine in their curricula [55]. Goals for telemedicine training include improving decision-making and coordination to enable learners in their turn to be able to deliver effective health care [60].

Looking forward, a major challenge for the use of telemedicine in digital medical education is that it cannot fully replicate in-person rapport building, body language cues [52], and clinical skills essential for medical students, despite the efforts of existing nonverbal communication training platforms [61]. Thus, while it has been suggested that the greatest benefits are to be gained when training is commenced in earlier stages of medical education [62], greater integration with traditional teaching may provide the best balance.

Further challenges to anticipate include the expansion of telemedicine for health education more equitably, particularly targeting lower-income countries and rural areas within high-income countries [52,63]. This is particularly important as remote areas in lower-income countries may have the most to gain from telemedicine-oriented digital education [64,65]. Work focusing on sub-Saharan Africa, containing 33 of the 48 lowest-income countries in the world, has provided hope for ways forward to achieve [64,66]. While the first decade of the millennium carried great hope that international African networks, such as the Fundamental of Modern Telemedicine for Africa and the Réseauen Afrique Francophone pour la Telemedicine, might help narrow the gap in telemedicine education across Africa, accessibility is still suboptimal even in 2024 [64]. Ongoing challenges contributing to this include access to the internet, computers, and social media, while only a few medical practitioners were even aware of telemedicine training, as reported in a 2024 study based in Ethiopia [64]. Important factors identified in this study were access to computers and the internet, while a gender discrepancy was identified, where, among the practitioners, there is a greater knowledge of telemedicine among men. Improving existing reimbursement models and fostering more coordinated partnerships between governments, the private sector, and communities in telehealth education have been identified as important steps forward in the coming years [65]. Therefore, rather than additional technological advances, it seems likely that the advances in infrastructure will be one of the key opportunities to further develop the great potential of telemedicine digital medical education in the coming years [65].

mHealth and Digital Medical Education

The last 25 years have seen a surge in the use of mobile technology worldwide [67-69]. The increased use of mHealth apps [70], particularly within digital medical education, partially reflects the ease of access they can provide. While mHealth apps have often been used for medical practice [71], they have also been widely used in medical education [72,73]. The increased use of mHealth apps for medical education in recent years [74] reflects the call for such expansion in the early 2010s [68]. Examples of mHealth use in digital medical education include web-based textbooks, podcasts, medical calculators, web-based lectures [68,72,73], and internet-based anatomical models used to visualize structures in 3D and medical scans [68]. Simulation apps, such as Touch Surgery [75], have been used to effectively learn, consolidate, and practice surgical maneuvers and clinical skills such as cardiopulmonary resuscitation [76,77]. Furthermore, mHealth apps can facilitate collaborative learning, CME, and clinical decision-making [68,72,73]. In addition to providing education for health care students and professionals, the ease of access to mHealth apps allows them to be used for the health education of patients as well [68]. This educational content can range from teaching patients rehabilitation for chronic obstructive pulmonary disease [78,79] to providing general explanations for their symptoms.

Both within and outside lecture halls and clinics, mobile devices have proven to be a valuable resource for accessing educational content where distance, resources, or financial situations may have previously been a limiting factor [68,73,75,80]. While some apps have paid components [68,76], the low costs of many apps lessen the potential for financial barriers for many learners [68,72]. To maximize the educational benefit and maintain user interest, mHealth apps have implemented strategies, including interactive quizzes, instant feedback, and gamification, which use game-like features, such as point scoring, leaderboards, and challenges to make the learning more engaging [81].

As with many digital health resources, concerns for privacy and confidentiality should be addressed [72]. This can, in part, be mitigated through regular updates of antivirus software; however, the risk of a mobile device being stolen or lost remains. Another challenge may be that the convenience and subsequent reliance on mHealth educational resources could compromise the standard and depth of internal thought that is required in medical practice [72]. Furthermore, physical limitations, such

XSL•FO
RenderX

as screen size [68] and obstruction of the device screen during interactive elements [76], can restrict the ways mHealth apps are used for medical education.

Looking ahead, while mHealth apps may be widely available [71,80], a concern among users is regarding the accuracy of results [72]. For example, it was noted in 2012 that >44% of apps giving information on cancer provided data that were not scientifically validated [67]. There is not yet a standardized protocol for establishing the quality of mHealth apps, and the development of such protocols would not only aid design but also allow for more consistent quality control [81]. In the near future, the increasing use of AI, explored more comprehensively subsequently, is likely to simulate complex medical scenarios using mHealth [81].

Further expansion of the reach of mHealth education apps geographically will be an important goal once concerns about equitable access to technology have been addressed. One important aspect regarding this is the need for apps to be available on multiple operating systems [68,81]. Accessibility to mHealth in lower-income nations may superficially appear less problematic than telemedicine, with nearly 4 billion smartphone users recently estimated globally and wide penetration even in many lower-income countries [82]. However, such figures can be deceptive as the full functionality of smartphones requires a good internet connection, which is often problematic in lower-income countries [64]. Similarly, even where rapid uptake of mHealth education apps is accomplished, retention and ongoing participation of users have been more problematic [82]. A recent paper has pointed to the opportunity for improved incorporation of the complexities of learner engagement at the development stage of mHealth apps [82]. This concept has been shown to be effective in a recent study involving the co-design of an mHealth app with the learners, in this case, medical students [83]. Building on such personalized models for mHealth learners in the future could more closely align with the use of similar apps designed for patients [68] and may lead to greater engagement and agency among learners.

BDA in Digital Medical Education

Through ever-increasing interactions with digital technologies, users generate large amounts of data. BDA describes the process of collecting and interpreting these large datasets from multiple sources to gain meaningful information [84]. The history of big data may be outlined in terms of the volume of typical datasets over time. The late 1990s, at the beginning of our chosen review period, have been considered the milestone when the terabyte (a billion bytes of data) started giving way to petabytes (1000 terabytes), while the second decade of the new millennium has been viewed as the transition from petabytes to the exabyte (1000 petabytes) [85]. BDA use is widespread across industries and has been associated with reduced costs, improved productivity, and an impact on decision-making [86]. In the last decade, developments have been made in the use of BDA in health care, supporting improvements in patient monitoring, data-driven diagnosis, management of hospital resources to improve service quality, and evidence-based decision-making at both a public health policy level and a local level [87]. However, applications of BDA in digital medical education have only begun to accelerate in more recent years [88]. BDA holds great potential to enhance learning, although a recent systematic review showed that analytics were most frequently used to simply capture the number of connections made by students to the learning material, while areas such as feedback and at-risk intervention were yet to be fully used [89].

A key quality of BDA lies in its ability to support decisions. Changes to curricula, course structures, and admission policy can be visualized by their impacts on performance data [90], allowing program directors to swiftly notice patterns and respond to them. In addition, accumulating performance data could help generate benchmarks for programs as an additional metric for ensuring quality standards. On an individual level, BDA can help to identify or predict students and trainees who are struggling or at risk of underachieving, allowing extra intervention to be provided for them [91]. However, there has been some debate regarding the psychological consequences of using predictive analytics in this way and whether it may potentially exacerbate a student's struggles by prematurely labeling them as falling below a certain metric [92].

Furthermore, digital medical education platforms and institutions may use BDA as a form of feedback by using data patterns on engagement instead of relying on manual questionnaires, helping to save time and potentially producing more valuable insights into where improvements need to be made in health education [93]. For example, data on whether medical students engage with digital content (eg, clicking hyperlinks or magnifying images on course web pages) have demonstrated value in predicting which students would perform better at examinations [94]. With focused use, BDA can facilitate and optimize learning methods, helping learners to understand specific medical content and aiding in the personalization of the education process. Performance data on clinical skills, such as catheter insertion, can create a baseline to judge when a student is ready to perform such skills on a real patient while also providing the student with reassurance and confidence when they are ready [90]. Data collected about learners should be fed back to them, closing the loop, and allowing them to have a greater understanding of how data are being used and how they will benefit them [95]. One way of doing this is by creating personalized dashboards for students, which systematically deliver and help visualize performance data [96].

Looking to the near future, an ambitious endeavor of BDA may be possible from the combination of data on clinical information in hospitals with data from medical education, to gain an insight into how different medical curricula and educational models impact patient outcomes [90]. However, it may be difficult to interpret these data-generated patterns, as many other factors can affect clinical outcomes, including team decisions, local practices, and technological disparities, making it hard to attribute results to individual educational programs [97]. Other challenges associated with BDA still require consideration. Emerging patterns have been shown to be open to misinterpretation, as data may simply be correlated and not causally linked [90]. In addition, as with other large datasets, there is the potential trap of finding examples of significant patterns that may not necessarily be meaningful [98]. Traces of

XSL•FO
RenderX

a learner's digital data, reflecting their behaviors on the web, may potentially be gathered as part of big data when using digital health technologies in education [99]. It is still unclear what the exact limits on the use of such data would be in the future; however, it is an important area to be cognizant of, given the increased use of such technologies in digital medical education [100].

The Metaverse, AR, and VR in Digital Medical Education

The metaverse is a 3D web-based world in which users can interact both with other users and with digital objects, often, but not exclusively, experienced through AR and VR [28]. While the metaverse is not in itself a new concept, the more recent rapid development of associated tools, such as blockchain and nonfungible tokens as well as the close association with the evolving Web 3.0 in the last decade [101], have created considerable opportunities across digital environments, particularly in digital medical education. The decentralized concept of Web 3.0 provides the potential for greater dissemination of education while also posing challenges for education providers in terms of possible loss of some control over content and ensuring that legal procedures, from data protection to copyright and accessibility laws, are strictly followed. At the same time, individual educators and educational companies now have increasing opportunities to interact with learners directly using the blockchain features of Web 3.0 to monetize the services they provide, potentially bypassing traditional educational institutions [28].

Earlier integrations of the metaverse in medical education involved the use of web-based classrooms, where avatars represent and perform actions on behalf of the user. These 3D worlds have been reproduced on a computer monitor without the need for headsets or lenses. A platform exploring these internet-based worlds for digital medical education includes Second Life, developed by Linden Lab in 2003, which provides educational materials on various medical topics and offers consumers access to health information [102]. Another promising platform for integrating metaverse platforms into medical education is the open-source OpenSimulator project [103]. This platform has provided teaching in multiple medical specialties, with base models that have been repurposed and customized to create web-based medical center scenarios.

As the metaverse is built around extended reality technology aimed at making an immersive experience for users, the more focused terms AR and VR are important to consider [28]. AR involves a digital overlay on top of the real world, whereas VR involves full immersion into the digital sphere [104]. Both AR and VR can be realized via head-mounted devices [105] or digital glasses [106], among other technologies, and represent just a few of the many ways in which the metaverse is experienced. An important distinction between the metaverse and some of the experiences offered by AR and VR lies in the nature of the interactions with other humans. While the metaverse is inherently social, providing a shared digital environment for multiple users, AR and VR can also enable an individual to interact with immersive environments independently without the need for social contact.

AR has shown significant value in training techniques across numerous medical specialties and is seen frequently in surgery [107]. For example, as a urological surgery simulation tool, Google Glass demonstrated positive outcomes in medical students and urological surgeons across various stages of training [106]. This training provided a sense of reality in the training environment, which was particularly favored by the younger participants. AR has also shown benefits for students studying anatomy, with AR textbooks assisting students in learning more effectively compared to standard textbooks [107]. Furthermore, students taught with AR were found to have better performance in long-term spatial anatomy knowledge compared to those taught traditionally [107].

In terms of the efficiency of VR, a randomized controlled trial carried out with Japanese nursing students noted no significant differences between individuals using VR and those taught with traditional methods in terms of skills, knowledge, or confidence [108]. The only statistically significant finding was a higher satisfaction rate in students who were taught traditionally, likely due to the sickness associated with VR adjustment. The study was limited to a small sample size and only tested third-year students, making it difficult to determine its wider applicability. Similar results were seen in a study on undergraduate paramedic students, with no significant differences between the results of those who had lecture-based teaching and those who had VR-based teaching [109]. However, a systematic review evaluating immersive VR use in nursing education [110] highlighted that there is generally a positive response to VR-assisted learning among nursing students. This included reduced anxiety in those performing skills in real life with prior VR practice due to 360-degree simulations that resembled real life more closely than simulation situations with manikins or lectures. A systematic review conducted by Foronda et al [111] found similar results, with the efficiency of VR typically not being statistically different from traditional teaching methods. However, it was also noted that most trials using VR only used it as an intervention once, and there was a lack of data showing how technology faired to traditional teachings across multiple sessions. VR simulators have shown significance compared to controls in providing greater self-confidence for surgical residents in dentistry [112]. Furthermore, it was identified that residents who used VR navigation software had higher accuracy in marking their first implantation sites during the surgical stage of dental implantation [112]. A recent meta-analysis performed by Kyaw et al [113] showed that teaching health professionals with VR interventions that are mostly interactive resulted in a considerable improvement in skillset and, to a lesser extent, improvement in knowledge compared to traditional methods. Both AR and VR are promising as active learning methods for teaching anatomy to undergraduate students. Combining these methods into extended reality teaching was perceived by health care students as more useful than traditional methods [114].

There are some device-related downsides to both AR and VR, as they are primarily dependent on the associated headsets or digital glasses, and device development in the coming years should consider these issues. For example, users who already wear spectacles report discomfort when wearing such devices on top of them [106]. Device use may also be associated with

XSL•FO
RenderX

nausea, motion sickness (sometimes termed "cybersickness") [104], technical difficulties, and stress, and it is unclear whether such symptoms resolve as a participant becomes accustomed to the experience over time or whether they are long-lasting symptoms [104,105]. There is a need to develop clear guidance for potential users of AR and VR to determine whether any preexisting health conditions might need additional support or to take precautions before using these devices [104]. Notably, currently, there is a lack of universal standardization or established guidelines encompassing VR, the metaverse, and AR.

AR and VR devices currently require regular charging, which can be inconvenient when needed for long periods and may benefit from future improvements in battery technology. Furthermore, there are concerns about accessibility to such education in low- and middle-income countries, despite having populations that might particularly benefit from AR- and VR-based education due to the unequal distribution of health professionals. Inequalities and limited accessibility have been noted in many countries, including higher-income ones, with internal disparities between urban and rural areas [105]. As mentioned in the *Telemedicine and Digital Medical Education* section, lower-income countries and isolated rural areas, even in higher-income countries, may stand to benefit the most from the metaverse if it becomes accessible in more remote areas in the near future. As technology and equipment become more readily available, the overall cost of AR and VR should decrease, improving accessibility in low- and middle-income countries.

There are some scenarios where students' confidence in AR or VR education techniques may be misplaced, as a systematic review by Baniasadi et al [104] suggested that a challenge of remote VR education may be a lack of direct supervision and evaluation of student performance. In addition, there may be limited technical support for those who may find it difficult to manage the complexity of VR independently. Increasing student engagement with AR or VR teaching by gamifying digital learning materials may also increase social connectivity, helping combat the risk of social isolation in remote learning, and seem likely to be important areas for expansion in the future [115].

Impact of the COVID-19 Pandemic on Digital Medical Education

The COVID-19 pandemic, which began in early 2020, impacted nearly 1.6 billion learners, that is, 94% of all students worldwide at the time, with enforced social distancing necessitating rapid and significant changes in education, particularly with a pivot from face-to-face to remote digital delivery [116]. For many educators, who might previously only have been accustomed to teaching face-to-face, the rapid change required to convert to distance learning with digital delivery created tremendous challenges [117]. The rapid upskilling of these individuals in fundamental digital skills, such as the use of videoconferencing, was vital across many courses globally, including in health education. For most educators, the immediate practicalities of rapidly upskilling educators and changing delivery to digital formats had to take precedence over exact pedagogies [116]. Preexisting providers of digital medical education were in a position to help their peers adapt. There were significant challenges faced by students as well [118]. For those who had originally applied for completely face-to-face courses, the experience of distance learning may have differed significantly from what was anticipated. We should not underestimate the very significant added stresses of the time [119]. For example, in many countries, lockdowns added restrictions on movement and concerns about risk to personal health or the health of loved ones. This multitude of factors makes it difficult, even in retrospect, to fully analyze the effectiveness of digital medical health education during the pandemic. However, what does seem certain is that the COVID-19 pandemic was associated with a much wider subsequent dissemination of digital health teaching [120].

Much of the early literature published during the time of the COVID-19 pandemic focused on the adaptations used by medical health educators to pivot to the digital environment. Many of these specifically looked at the practical steps that were required, including training on acceptable behaviors expected during these digital meetings, especially using videoconferencing [120]. There appear to be more studies conducted in the United States that report the direct use of web-based discussions with a patient as part of the digital education delivery to students compared to those in the United Kingdom and Europe [120]. Prolonged use of videoconferencing was associated with learner fatigue, in some cases, postulated to stem from critical self-analysis of the learner's face appearing on the screen during teaching [121].

The enforced change of educational delivery to digital platforms away from classrooms may have risked some social isolation, especially in university student populations where much socialization occurs during in-person classes [118]. Various techniques have been used during videoconferencing to help improve focus and engagement between students and educators and among students themselves. Regular chunking of information and the use of breakout rooms may have benefited some in this regard. With apparently less forthcoming students, sometimes referred to as "lurkers" [120], these techniques may have been effective. The interactive behavior of web-based learners has sometimes been sharply divided into "the workers, the lurkers, and the shirkers" [122]. However, the role of more passively "lurking" in the background during videoconferencing calls may, in fact, be used by most attendees early on, as they adjust to the meeting and before they can participate more fully [123].

Despite the challenges of the time, student surveys during the pandemic revealed some positive feedback regarding their enforced web-based medical education [124]. Since the last of the lockdowns in most countries was lifted, beginning in March 2021 in England [125], the evidence overall indicates that there has been increased use of digital education methods since prepandemic times [120]. For example, in the United States, the total number of students enrolled in degree-granting postsecondary institutions who had ≥1 of their courses delivered on the web was >14 million (75% of the total) at the height of the COVID-19 pandemic in autumn 2020, but this number was still >11 million (60% of the total) in autumn 2021, well after nationwide lockdowns had been lifted [126]. This was still

XSL•FO
RenderX

substantially above the prepandemic 2019 level when 42% of postbaccalaureate (postgraduate) students were enrolled in ≥1 distance education course [127].

Many learning institutes had not only upskilled educators but also invested in the technology to permit digital learning [128]. In addition, the sudden enforced changes in education during the COVID-19 pandemic may have made changes in previously accepted higher-education structures easier to gain ground. For example, we are noticing a rapid rise since the COVID-19 pandemic in alternative courses to formal higher-education degrees, such as microcredentials and digital badges, with which digital medical education could be used to offer greater flexibility to learners than with traditional degree courses, even compared to those with a modular structure [129].

AI in Digital Medical Education

The term "artificial intelligence" (later abbreviated as AI) was first coined in 1955 and loosely described machines that could use language to solve problems, create abstract concepts, and improve themselves in doing so [130]. However, it is in the last decade that dramatic developments in this field have led to both great potential and great challenges for digital medical education. AI has been used in areas as diverse as medical practice management, patient monitoring, diagnostics, and device integration and within medical education in a range of fields, including radiology, dermatology, and surgery [131,132]. The use of AI tools in digital health has been particularly impactful, with AI tools demonstrating the capacity to outperform even highly experienced professionals, highlighting its potential to reshape health care education by optimizing training and providing decision support tools [133]. Despite this, the use of AI in digital medical education has, to date, lagged somewhat behind its use in medical practice.

Large language models, such as ChatGPT, have become very widely used in the last 2 years across many disciplines. Despite several potential challenges, they can quickly offer a very brief summary of many long-established facts in medicine [134]. Consequently, they can be used partially when preparing medical education, including the generation of patient cases and practice questions. Because of plausible errors and dataset time limitations, the current use of such large language model studies may be less effective for students than experienced educators who could more easily spot errors. Examples of how medical educators might use tools, such as ChatGPT, have included internet-based patient simulation, quizzes for medical students, summarizing research articles, and generating brief curricula for digital health learners [134]. However, the rapid production of relevant material by large language models gives rise to concerns that some students might use them to help answer examination questions. In 2023, it was shown that ChatGPT could perform as well as a typical third-year medical student on Step 1 and Step 2 in the US Medical Licensing Examination, and further performance improvements are likely [135]. Nevertheless, students might benefit from interacting with ChatGPT in a dialogue to explore concepts when learning independently or with peers [135]. It is currently not possible to detect with certainty whether a student who submitted their work had used AI to answer examination questions, and this is especially a concern with noninvigilated examinations. While tools such as the generative textual likelihood ratio can help humans detect work produced using large language models, the high stakes of plagiarism for students can make even 99% accuracy seem insufficient to confidently declare that AI models were used [136]. Hence, teaching institutions have to reconsider the assessment environment for students with invigilated face-to-face or distance learning examinations in real time. Furthermore, alternative approaches, such as higher marks for critiquing, use of modern references, or interpretation of medical images by students, may currently help, but even these may become less challenging for large language models in the near future [137]. A similar concern in digital medical education extends beyond students. A recent study demonstrated that it was possible to use ChatGPT (GPT-3) to deliberately produce a fraudulent paper nearly 2000 words long with 17 references in just 1 hour. While there were some errors, such as false references and so-called "hallucinations," [134] which an expert might detect, this detection rate may decrease with time as large language models continue to improve.

AI text-to-image generation of high-quality images depicting physical appearances has proved possible with the addition of diffusion models over the last decade, allowing the potential production of illustrations even of rare medical conditions [138]. For example, images of arthritis, a potential thyroid mass, and hypothyroidism have been produced as entirely novel images not based on traditional photographs of patients [139,140]. The images of the face, in particular, may offer great promise as an alternative to traditional patient photographs, where the very important issues of consent and confidentiality may limit the availability of images. Such AI text-to-image generation can also rapidly produce novel pictures to illustrate case studies for teaching. Moreover, many tools are now widely available to produce new, short videos from AI text-to-video prompts. However, most tools to date mainly use speaking avatars (with some degree of apparent lip-synching) on a background of very short traditional stock video clips, which are typically linked together via AI to create a finished video. These tools certainly have the potential to enliven a short text teaching session. There is also rapidly evolving work on creating genuinely new moving images without the use of stock videos. OpenAI, in early 2023, demonstrated some high-quality short videos using Sora, which may be more widely available later in the year and potentially then be used for digital medical education [141].

Looking forward, pressing challenges to address regarding AI include clarifying the copyright of AI-generated text and images, including the limits of such images and the lack of clarity about the original dataset of images used to train the generative adversarial networks through which AI image-generating tools are developed. Other potential issues, such as accuracy, and potential perpetuation of stereotypes (including skin color and perceived gender) have been raised [142]. Similar concerns have been raised for text-to-video generation, with the possibility that inaccurately portrayed medical images or deliberately misleading videos (including "deep fake" images) may be produced [141]. The establishment of a global quality control body of AI-generated products could substantially help to manage challenges such as these. Watermarking of work

XSL•FO
RenderX

produced using large language models and text-to-image generation began in 2024, although this may not be impregnable, and further development in the coming years is expected [136]. Despite these considerable cautions, the field of AI text-to-image and text-to-video generation seems to have great potential for the near future of digital medical education while recognizing the important challenges that have been mentioned in this field. This marks AI as one of the biggest potential areas of growth in digital medical education looking to the immediate future.

Ethics and Cybersecurity in Digital Medical Education

The ethical and legal considerations of digital medical education have not been explored in extensive depth in the currently available published literature; however, some extrapolation is possible from work looking primarily at digital health and is highly relevant given the many advances in digital education in the last 25 years. A recent review found that disparities in digital access, especially among users with different ethnicities or income statuses, contributed significantly to eHealth use and exacerbated social health inequalities [143]. Likewise, digital medical education may be a disproportionately less available option for racial or ethnic minority groups, older adults, and those with limited health literacy [143]. This raises concerns regarding the fairness of accessibility of educational resources across all demographics. The open access model, used successfully in many medical journals, might provide a model to follow while trying to develop greater equality digital medical education, potentially building on some of the challenges that still exist in some areas of open access [144].

Health organizations do not always have the level of cybersecurity needed to prevent unwarranted access to personal patient information [145]; therefore, educational content created from patient information or sensitive data may not be appropriate to use without adequate safeguards. This should be borne in mind when obtaining consent from patients for the use of such material (eg, medical photography) in education. Safeguards should be put in place to restrict the inappropriate distribution of such content to unintended personnel via means such as social media [146]. This is particularly the case in digital education, where images and videos might be stored for long-term use, including downloading, manipulating, and uploading [147]. The use of encryption software may aid in securing data with sensitive information, with multiple layers of encryption providing stronger protection [148]. This may help overcome some ethical concerns regarding data storage, as not only text but also some photos and videos can be anonymized. General Data Protection Regulation compliance is important in digital medical education, with aspects such as obtaining informed consent data minimization, and anonymization, all being factored in [149]. Securing password protection of accounts and software or creating personalized links for individuals may help ensure that data are only accessed by people with given log-ins, helping to regulate those who can access educational software that includes patient images or data. In addition to ensuring that data protection protocols are followed, it is also fundamental that the significance and impact of data protection is taught to health care students who are increasingly accessing health and education digitally. However, it is currently unclear how widely digital health competencies have been achieved among health workers in dealing with aspects such as electronic health records [150,151].

The immediate future holds significant cybersecurity concerns. Potential security breaches can reach any area of digital health care courses, even its fundamental learning platforms. For example, concerns about potential vulnerabilities of LMSs, especially open-source versions such as Moodle, have been highlighted. As LMSs are used to form the main foundation of many digital medical education courses, this is a key area that will need ongoing vigilance [26,31]. As Moodle was designed to store user data in caches, these could theoretically be targeted by an attacker to launch an attack with the subsequent session [26]. Even with limited chances to access the system, brute force cyberattacks have also been highlighted as a potential vulnerability [26].

Digital medical education platforms may have some potential vulnerabilities, which have been categorized as falling into 4 main groupings [26]. Authentication challenges may occur if an attacker uses the “forgot your password” option. If session tokens are sent in response without proper encryption, security may be breached. Availability issues may involve denial of service instigated by either flooding or logic attacks. Confidentiality attacks attempt to access existing confidential data on the learning system, while integrity attacks attempt to modify or delete stored data [26]. Suggested improvements, including the use of Completely Automated Public Turing test to tell Computers and Humans Apart when logging in to help counter brute force attacks and the use of a secure sockets layer to help prevent hijacking of sessions, have not yet been fully implemented, and increasing the use of such technologies in the immediate future would assist with digital course security [26]. While such technical improvements would further enhance digital security, it seems likely that human error is important in many cybersecurity breaches as with digital health care [152]. Therefore, further developments of tailored educator training to avoid breaches are likely to be an effective approach in countering the increasing risk to cybersecurity [152].

One possible risk of the increasing use of digital medical education is that it might contribute to digital addiction, potentially affecting 25% of the general population, especially in instances of the metaverse, AR, and VR being integrated into health education through game-like apps [153]. There are concerns that technology overuse due to the introduction of digital medical education might exacerbate social isolation and reduce opportunities for health care students to benefit from the positive role modeling of professional behaviors, such as empathy and communication skills [118]. Although student response systems, such as Socrative, use games to improve learner engagement, Kahoot! was the first to be designed based on a theory of intrinsic motivation [154]. Kahoot! was only realized as a platform in 2013; however, by 2019, >2.5 billion individuals had played it, underlying the great importance, both of potential benefits and ongoing vigilance to avoid potential digital addiction with student response systems [154].

XSL•FO
RenderX

Discussion

Innovations and Future Directions

Digital medical education has evolved dramatically in the past 25 years. While we have examined key components of this journey, some overarching themes emerge. Overall, a paradigm shift has occurred during this time, seeing a change from didactic, expository digital medical education to more active learning models. Technologies incorporated during this time include LMS, medical simulation and role-playing with VR and AR, and the metaverse. The resulting advancements in ways that medical education can be created, managed, and experienced have offered pointers for further possible developments in digital medical education. The increasing use of the internet and mobile apps are just a few examples of the many technological advancements that are reshaping how health care professionals are educated, increasing learning efficiency and offering uniquely personalized learning experiences. Digital medical education continues to expand before us, propelled by the rapid recent advances in these digital tools and underlying components, such as AI.

Developments in digital medical education have seen significant growth in recent years, driven by the emergence of generative AI and advancements in AR and VR, which have enabled immersive, simulation-based learning. Looking to the future, we can anticipate further integration of AI-associated learning systems; the use of wearable technology to provide more immersive experiences and real-time feedback; and the use of blockchain, for example, to store, encrypt, and verify certifications. Furthermore, future synthesis of AI may include improved personalized analytics for students, shaping an era where students have a more transparent understanding of their progress, competency in learning a skill, and their personal learning styles.

Challenges and Opportunities

However, several key challenges require ongoing attention. These include the need for improved accessibility to relevant technologies and, in some areas, improved internet connectivity. Ethical challenges also exist on several fronts, especially surrounding consent, confidentiality, and ownership of the web-based material used in digital medical education. While a boon for many, this relentless progress in digital medical education may leave some learners, and even some educators, with a feeling of being unable to keep pace with the evolving digital landscape, causing fear and stress [119,120]. To address this, institutions should prioritize accessible training programs, mentorship opportunities, and user-friendly interfaces to build confidence and competence.

Despite concerns about the rapid advances in technologies underlying digital medical education today, these changes should be embraced rather than feared, giving us further opportunities to support and empower our learners. However, at its heart, digital medical education will always be more than the technology and rather should enhance ways that we, as educators, can reach, support, motivate, and empower our learners.

Conflicts of Interest

None declared.

References

1. David N. The Antikythera mechanism: its dating and place in the history of technology. J Mediterr Archaeol 2017 Jul 03;30(1):85-104. [doi: 10.1558/jmea.32915]
2. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
3. Banay GL. An introduction to medical terminology I. Greek and Latin derivations. Bull Med Libr Assoc 1948 Jan;36(1):1-27 [FREE Full text] [Medline: 16016791]
4. Charen T. The etymology of medicine. Bull Med Libr Assoc 1951 Jul;39(3):216-221 [FREE Full text] [Medline: 14848629]
5. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Google Books Team, et al. Quantitative analysis of culture using millions of digitized books. Science 2011 Jan 14;331(6014):176-182 [FREE Full text] [doi: 10.1126/science.1199644] [Medline: 21163965]
6. McCutcheon K, Lohan M, Traynor M, Martin D. A systematic review evaluating the impact of online or blended learning vs. face-to-face learning of clinical skills in undergraduate nurse education. J Adv Nurs 2015 Feb;71(2):255-270. [doi: 10.1111/jan.12509] [Medline: 25134985]
7. Hayashi A, Chen C, Ryan T, Wu J. The role of social presence and moderating role of computer self efficacy in predicting the continuance usage of e-learning systems. J Inf Syst Educ 2004 Jan;15(2).
8. Wang ZY, Zhang LJ, Liu YH, Jiang WX, Tang SL, Liu XY. Process evaluation of e-learning in continuing medical education: evidence from the China-Gates Foundation Tuberculosis Control Program. Infect Dis Poverty 2021 Mar 10;10(1):23 [FREE Full text] [doi: 10.1186/s40249-021-00810-x] [Medline: 33750423]
9. Cope B, Kalantzis M. A little history of e-learning: finding new ways to learn in the PLATO computer education system, 1959–1976. History Educ 2023 Jan 16;52(6):905-936. [doi: 10.1080/0046760x.2022.2141353]
10. Howe AF, McConnell D. Teaching electronics at a distance by cyclops: an assessment of the telewriting system. Int J Electric Eng Educ 1984 Jul 01;21(3):237-249. [doi: 10.1177/002072098402100308]
11. Börje H. The role of technology in distance education. Instr Sci 1985 Aug;14(1):87-93. [doi: 10.1007/bf00052439]

XSL•FO
RenderX

12. Paloque-Bergès C, Schafer V. Arpanet (1969–2019). Internet Hist 2019 Jan 16;3(1):1-14. [doi: 10.1080/24701475.2018.1560921]
13. Bollmer G. Networks before the internet. JCMS 2019;59(1):142-148. [doi: 10.1353/cj.2019.0071]
14. Lassila O, Hendler J. Embracing "Web 3.0". IEEE Internet Comput 2007 May;11(3):90-93. [doi: 10.1109/mic.2007.52]
15. Matheson GO. Medical education: the on-line revolution. Phys Sportsmed 1999 Sep;27(9):5. [doi: 10.3810/psm.1999.09.989] [Medline: 20086741]
16. Peterson M. Continuing medical education on the internet: state of the art. J Contin Educ Health Prof 1999;19(4):242-249 [FREE Full text] [doi: 10.1002/chp.1340190410]
17. Briggs J. SYD1/444: assessing the readership of the UK national database of telemedicine. J Med Internet Res 1999 Sep 19;1:e126 [FREE Full text] [doi: 10.2196/jmir.1]
18. Neville RG, Ricketts IW, McCowan C, Hoskins G, Fenton I, Newell D. PHL4/411: getting practice on the internet. J Med Internet Res 1999 Sep 19;1(suppl1):e85 [FREE Full text] [doi: 10.2196/jmir.1.suppl1.e85]
19. Lum EH, Gross TJ. Telemedical education: teaching spirometry on the internet. Am J Physiol 1999 Jun;276(6 Pt 2):S55-S61. [doi: 10.1152/advances.1999.276.6.S55] [Medline: 16211667]
20. Diaz DP, Cartnal RB. Students' learning styles in two classes: online distance learning and equivalent on-campus. Coll Teach 2010 Mar 25;47(4):130-135 [FREE Full text] [doi: 10.1080/87567559909595802]
21. Lairson TD. Rethinking the “course” in an online world. Campus Wide Inf Syst 1999;16(5):189-190. [doi: 10.1108/10650749910300807]
22. Srinivasa KG, Kurni M, Saritha K. Pedagogy for e-learning. In: Learning, Teaching, and Assessment Methods for Contemporary Learners. Singapore, Singapore: Springer; 2022.
23. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods 2020 Mar;11(2):181-217 [FREE Full text] [doi: 10.1002/jrsm.1378] [Medline: 31614060]
24. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 06;6(1):245 [FREE Full text] [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]
25. Back DA, Behringer F, Haberstroh N, Ehlers JP, Sostmann K, Peters H. Learning management system and e-learning tools: an experience of medical students' usage and expectations. Int J Med Educ 2016 Aug 20;7:267-273 [FREE Full text] [doi: 10.5116/ijme.57a5.f0f5] [Medline: 27544782]
26. Kumar S, Dutta K. Investigation on security in LMS Moodle. Int J Inf Technol Knowl Manag 2011;4(1):233-238.
27. Veluvali P, Surisetti J. Learning management system for greater learner engagement in higher education—a review. High Educ Future 2021 Oct 27;9(1):107-121. [doi: 10.1177/23476311211049855]
28. Fan S, Yecies B, Zhou ZI, Shen J. Challenges and opportunities for the web 3.0 metaverse turn in education. IEEE Trans Learn Technol 2024;17:1989-2004. [doi: 10.1109/tlt.2024.3385505]
29. Hopper KB. Education, teleconferencing, and distance learning in respiratory care. Respir Care 2004 Apr;49(4):410-20; discussion 421 [FREE Full text] [Medline: 15085813]
30. Huynh R. The role of e-learning in medical education. Acad Med 2017 Apr;92(4):430. [doi: 10.1097/ACM.0000000000001596] [Medline: 28350599]
31. Jeong S, Hwang H. Do we need Moodle in medical education? A review of its impact and utility. Kosin Med J 2023 Sep;38(3):159-168. [doi: 10.7180/kmj.23.139]
32. Young T, Abdool K, Broadbent DK, Kok CY. Engaging student reviewers on the clinical neurology by distance learning course. In: Student Empowerment in Higher Education. Reflecting on Teaching Practice and Learner Engagement. Berlin, Germany: Logos Verlag Berlin GmbH; 2020:423-436.
33. Lamba P. Teleconferencing in medical education: a useful tool. Australas Med J 2011;4(8):442-447 [FREE Full text] [doi: 10.4066/AMJ.2011.823] [Medline: 23393532]
34. Lieberman JA, Nester T, Emrich B, Staley EM, Bourassa LA, Tsang HC. Coping with COVID-19. Am J Clin Pathol 2021 Jan 04;155(1):79-86 [FREE Full text] [doi: 10.1093/ajcp/aqaa152] [Medline: 32876671]
35. eLearning for undergraduate health professional education: a systematic review informing a radical transformation of health workforce development. World Health Organization. 2015. URL: https://iris.who.int/handle/10665/330089 [accessed 2024-05-01]
36. El-Sabagh HA. Adaptive e-learning environment based on learning styles and its impact on development students' engagement. Int J Educ Technol High Educ 2021 Oct 01;18(1) [FREE Full text] [doi: 10.1186/s41239-021-00289-4]
37. Kokoç M. Flexibility in e-learning: modelling its relation to behavioural engagement and academic performance. Themes eLearn 2019;12:1-16 [FREE Full text]
38. Shearer RL, Park E. Theory to practice in instructional design. In: Handbook of Distance Education. Milton Park, UK: Routledge; 2018.
39. Hege I, Tolks D, Adler M, Härtl A. Blended learning: ten tips on how to implement it into a curriculum in healthcare education. GMS J Med Educ 2020;37(5):Doc45 [FREE Full text] [doi: 10.3205/zma001338] [Medline: 32984504]

XSL•FO
RenderX

40. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504 [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
41. Zehry K, Halder N, Theodosiou L. E-Learning in medical education in the United Kingdom. Procedia Soc Behav Sci 2011;15:3163-3167. [doi: 10.1016/j.sbspro.2011.04.265]
42. Verkuyl M, Violato E, Harder N, Southam T, Lavoie-Tremblay M, Goldsworthy S, et al. Virtual simulation in healthcare education: a multi-professional, pan-Canadian evaluation. Adv Simul (Lond) 2024 Jan 10;9(1):3 [FREE Full text] [doi: 10.1186/s41077-023-00276-x] [Medline: 38200615]
43. Wynter L, Burgess A, Kalman E, Heron JE, Bleasel J. Medical students: what educational resources are they using? BMC Med Educ 2019 Jan 25;19(1):36 [FREE Full text] [doi: 10.1186/s12909-019-1462-9] [Medline: 30683084]
44. Krishnamurthy K, Selvaraj N, Gupta P, Cyriac B, Dhurairaj P, Abdullah A, et al. Benefits of gamification in medical education. Clin Anat 2022 Sep 08;35(6):795-807. [doi: 10.1002/ca.23916] [Medline: 35637557]
45. Watsjold BK, Cosimini M, Mui P, Chan TM. Much ado about gaming: an educator's guide to serious games and gamification in medical education. AEM Educ Train 2022 Aug;6(4):e10794 [FREE Full text] [doi: 10.1002/aet2.10794] [Medline: 36034886]
46. Tudor Car L, Soong A, Kyaw BM, Chua KL, Low-Beer N, Majeed A. Health professions digital education on clinical practice guidelines: a systematic review by Digital Health Education collaboration. BMC Med 2019 Jul 18;17(1):139 [FREE Full text] [doi: 10.1186/s12916-019-1370-1] [Medline: 31315642]
47. Tudor Car L, Kyaw BM, Nannan Panday RS, van der Kleij R, Chavannes N, Majeed A, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275 [FREE Full text] [doi: 10.2196/28275] [Medline: 34287206]
48. Frehywot S, Vovides Y, Talib Z, Mikhail N, Ross H, Wohltjen H, et al. E-learning in medical education in resource constrained low- and middle-income countries. Hum Resour Health 2013 Feb 04;11:4 [FREE Full text] [doi: 10.1186/1478-4491-11-4] [Medline: 23379467]
49. Soulé H, Warrick T. Defining 21st century readiness for all students: what we know and how to get there. Psychol Aesthet Creat Arts 2015 May;9(2):178-186 [FREE Full text] [doi: 10.1037/aca0000017]
50. Chen S, Morgado M, Jiang H, Mendes JJ, Guan J, Proença L. Medical and nursing students' satisfaction with e-learning platforms during the COVID-19 pandemic: initial findings of an experimental project in China. Heliyon 2024 Feb 29;10(4):e26233 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26233] [Medline: 38404766]
51. Kwon S, Kim W, Bae C, Cho M, Lee S, Dreamson N. The identity changes in online learning and teaching: instructors, learners, and learning management systems. Int J Educ Technol High Educ 2021;18(1):67 [FREE Full text] [doi: 10.1186/s41239-021-00304-8] [Medline: 34977335]
52. Jumreornvong O, Yang E, Race J, Appel J. Telemedicine and medical education in the age of COVID-19. Acad Med 2020 Dec;95(12):1838-1843 [FREE Full text] [doi: 10.1097/ACM.0000000000003711] [Medline: 32889946]
53. Waseh S, Dicker AP. Telemedicine training in undergraduate medical education: mixed-methods review. JMIR Med Educ 2019 Apr 08;5(1):e12515 [FREE Full text] [doi: 10.2196/12515] [Medline: 30958269]
54. Budakoğlu I, Sayılır M, Kıyak YS, Coşkun Ö, Kula S. Telemedicine curriculum in undergraduate medical education: a systematic search and review. Health Technol (Berl) 2021;11(4):773-781 [FREE Full text] [doi: 10.1007/s12553-021-00559-1] [Medline: 33996380]
55. Jonas CE, Durning SJ, Zebrowski C, Cimino F. An interdisciplinary, multi-institution telehealth course for third-year medical students. Acad Med 2019 Jun;94(6):833-837. [doi: 10.1097/ACM.0000000000002701] [Medline: 30870152]
56. Jarvis-Selinger S, Chan E, Payne R, Plohman K, Ho K. Clinical telehealth across the disciplines: lessons learned. Telemed J E Health 2008 Sep;14(7):720-725. [doi: 10.1089/tmj.2007.0108] [Medline: 18817503]
57. Singh G, O'Donoghue J, Soon CK. Telemedicine: issues and implications. Technol Health Care 2002;10(1):1-10. [Medline: 11847443]
58. Angaran DM. Telemedicine and telepharmacy: current status and future implications. Am J Health Syst Pharm 1999 Jul 15;56(14):1405-1426. [doi: 10.1093/ajhp/56.14.1405] [Medline: 10428449]
59. Inumpudi A, Srinivas M, Gupta DK. Telemedicine in pediatric surgery. Pediatr Surg Int 2001 Jul 1;17(5-6):436-441. [doi: 10.1007/s003830000528] [Medline: 11527184]
60. Guze PA. Using technology to meet the challenges of medical education. Trans Am Clin Climatol Assoc 2015;126:260-270 [FREE Full text] [Medline: 26330687]
61. Liu C, Lim RL, McCabe KL, Taylor S, Calvo RA. A web-based telehealth training platform incorporating automated nonverbal behavior feedback for teaching communication skills to medical students: a randomized crossover study. J Med Internet Res 2016 Sep 12;18(9):e246 [FREE Full text] [doi: 10.2196/jmir.6299] [Medline: 27619564]
62. Boyers LN, Schultz A, Baceviciene R, Blaney S, Marvi N, Dellavalle RP, et al. Teledermatology as an educational tool for teaching dermatology to residents and medical students. Telemed J E Health 2015 Apr;21(4):312-314 [FREE Full text] [doi: 10.1089/tmj.2014.0101] [Medline: 25635528]
63. Yaghobian S, Ohannessian R, Mathieu-Fritz A, Moulin T. National survey of telemedicine education and training in medical schools in France. J Telemed Telecare 2020 Jun;26(5):303-308. [doi: 10.1177/1357633X18820374] [Medline: 30602352]

XSL•FO
RenderX

64. Shibabaw AA, Chereka AA, Walle AD, Demsash AW, Dube GN, Dubale AT, et al. Knowledge of telemedicine and its associated factors among health professional in Ethiopia: a systematic review and meta-analysis. PLoS One 2024;19(4):e0301044 [FREE Full text] [doi: 10.1371/journal.pone.0301044] [Medline: 38635662]
65. Cuadros DF, Huang Q, Mathenjwa T, Gareta D, Devi C, Musuka G. Unlocking the potential of telehealth in Africa for HIV: opportunities, challenges, and pathways to equitable healthcare delivery. Front Digit Health 2024;6:1278223 [FREE Full text] [doi: 10.3389/fdgth.2024.1278223] [Medline: 38500968]
66. Wamala DS, Augustine K. A meta-analysis of telemedicine success in Africa. J Pathol Inform 2013;4:6 [FREE Full text] [doi: 10.4103/2153-3539.112686] [Medline: 23858382]
67. Pandey A, Hasan S, Dubey D, Sarangi S. Smartphone apps as a source of cancer information: changing trends in health information-seeking behavior. J Cancer Educ 2013 Mar;28(1):138-142. [doi: 10.1007/s13187-012-0446-9] [Medline: 23275239]
68. Mosa AS, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 2012 Jul 10;12:67 [FREE Full text] [doi: 10.1186/1472-6947-12-67] [Medline: 22781312]
69. Maaß L, Freye M, Pan CC, Dassow HH, Niess J, Jahnel T. The definitions of health apps and medical apps from the perspective of public health and law: qualitative analysis of an interdisciplinary literature overview. JMIR Mhealth Uhealth 2022 Oct 31;10(10):e37980 [FREE Full text] [doi: 10.2196/37980] [Medline: 36315221]
70. Lewis TL, Boissaud-Cooke MA, Aungst TD, Eysenbach G. Consensus on use of the term "app" versus "application" for reporting of mHealth research. J Med Internet Res 2014 Jul 17;16(7):e174; discussion e174 [FREE Full text] [doi: 10.2196/jmir.3460] [Medline: 25033233]
71. Murfin M. Know your apps: an evidence-based approach to evaluation of mobile clinical applications. J Physician Assist Educ 2013;24(3):38-40. [doi: 10.1097/01367895-201324030-00008] [Medline: 24261171]
72. Wallace S, Clark M, White J. 'It's on my iPhone': attitudes to the use of mobile computing devices in medical education, a mixed-methods study. BMJ Open 2012;2(4):e001099 [FREE Full text] [doi: 10.1136/bmjopen-2012-001099] [Medline: 22923627]
73. Ozdalga E, Ozdalga A, Ahuja N. The smartphone in medicine: a review of current and potential use among physicians and students. J Med Internet Res 2012 Sep 27;14(5):e128 [FREE Full text] [doi: 10.2196/jmir.1994] [Medline: 23017375]
74. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
75. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
76. Semeraro F, Taggi F, Tammaro G, Imbriaco G, Marchetti L, Cerchiari EL. iCPR: a new application of high-quality cardiopulmonary resuscitation training. Resuscitation 2011 Apr;82(4):436-441. [doi: 10.1016/j.resuscitation.2010.11.023] [Medline: 21227560]
77. Low D, Clark N, Soar J, Padkin A, Stoneham A, Perkins GD, et al. A randomised control trial to determine if use of the iResus© application on a smart phone improves the performance of an advanced life support provider in a simulated medical emergency. Anaesthesia 2011 Apr;66(4):255-262 [FREE Full text] [doi: 10.1111/j.1365-2044.2011.06649.x] [Medline: 21401537]
78. Marshall A, Medvedev O, Antonov A. Use of a smartphone for improved self-management of pulmonary rehabilitation. Int J Telemed Appl 2008;2008:753064 [FREE Full text] [doi: 10.1155/2008/753064] [Medline: 18615186]
79. Spielmanns M, Gloeckl R, Jarosch I, Leitl D, Schneeberger T, Boeselt T, et al. Using a smartphone application maintains physical activity following pulmonary rehabilitation in patients with COPD: a randomised controlled trial. Thorax 2023 May;78(5):442-450 [FREE Full text] [doi: 10.1136/thoraxjnl-2021-218338] [Medline: 35450945]
80. O'Donovan J, Bersin A, O'Donovan C. The effectiveness of mobile health (mHealth) technologies to train healthcare professionals in developing countries: a review of the literature. BMJ Innov 2014 Dec 12;1(1):33-36 [FREE Full text] [doi: 10.1136/bmjinnov-2014-000013]
81. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MD, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 04;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]
82. Monachelli R, Davis SW, Barnard A, Longmire M, Docherty JP, Oakley-Girvan I. Designing mHealth apps to incorporate evidence-based techniques for prolonging user engagement. Interact J Med Res 2024 Mar 26;13:e51974 [FREE Full text] [doi: 10.2196/51974] [Medline: 38416858]
83. Dederichs M, Nitsch FJ, Apolinário-Hagen J. Piloting an innovative concept of e-mental health and mHealth workshops with medical students using a participatory co-design approach and app prototyping: case study. JMIR Med Educ 2022 Jan 10;8(1):e32017 [FREE Full text] [doi: 10.2196/32017] [Medline: 35006085]
84. Caspari-Sadeghi S. Learning assessment in the age of big data: learning analytics in higher education. Cogent Educ 2022 Dec 31;10(1) [FREE Full text] [doi: 10.1080/2331186x.2022.2162697]
85. Lydia EL, Swarup MB, Laxmi MV. A literature inspection on big data analytics. Int J Innov Res Eng Manag 2016;3. [doi: 10.21276/ijirem]

XSL•FO
RenderX

86. Rahul K, Banyal RK, Arora N. A systematic review on big data applications and scope for industrial processing and healthcare sectors. J Big Data 2023 Aug 27;10:133. [doi: 10.1186/s40537-023-00808-2]
87. Khanra S, Dhir A, Islam AK, Mäntymäki M. Big data analytics in healthcare: a systematic literature review. Enterp Inf Syst 2020 Oct 12;14(7):878-912. [doi: 10.1080/17517575.2020.1812005]
88. Chan T, Sebok-Syer S, Thoma B, Wise A, Sherbino J, Pusic M. Learning analytics in medical education assessment: the past, the present, and the future. AEM Educ Train 2018 Apr 22;2(2):178-187 [FREE Full text] [doi: 10.1002/aet2.10087] [Medline: 30051086]
89. Bojic I, Mammadova M, Ang CS, Teo WL, Diordieva C, Pienkowska A, et al. Empowering health care education through learning analytics: in-depth scoping review. J Med Internet Res 2023 May 17;25:e41671 [FREE Full text] [doi: 10.2196/41671] [Medline: 37195746]
90. Ellaway RH, Pusic MV, Galbraith RM, Cameron T. Developing the role of big data and analytics in health professional education. Med Teach 2014 Mar;36(3):216-222. [doi: 10.3109/0142159X.2014.874553] [Medline: 24491226]
91. Saqr M, Fors U, Tedre M. How learning analytics can early predict under-achieving students in a blended medical education course. Med Teach 2017 Jul;39(7):757-767. [doi: 10.1080/0142159X.2017.1309376] [Medline: 28421894]
92. Thoma B, Ellaway RH, Chan TM. From utopia through dystopia: charting a course for learning analytics in competency-based medical education. Acad Med 2021 Jul 01;96(7S):S89-S95. [doi: 10.1097/ACM.0000000000004092] [Medline: 34183609]
93. Borakati A. Evaluation of an international medical e-learning course with natural language processing and machine learning. BMC Med Educ 2021 Mar 25;21(1):181 [FREE Full text] [doi: 10.1186/s12909-021-02609-8] [Medline: 33766037]
94. Cirigliano MM, Guthrie CD, Pusic MV. Click-level learning analytics in an online medical education learning platform. Teach Learn Med 2020 May 12;32(4):410-421. [doi: 10.1080/10401334.2020.1754216] [Medline: 32397923]
95. Clow D. The learning analytics cycle: closing the loop effectively. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. 2012 Presented at: LAK '12; April 29-May 2, 2012; Vancouver, BC URL: https://doi.org/10.1145/2330601.2330636 [doi: 10.1145/2330601.2330636]
96. Boscardin C, Fergus KB, Hellevig B, Hauer KE. Twelve tips to promote successful development of a learner performance dashboard within a medical education program. Med Teach 2018 Aug;40(8):855-861. [doi: 10.1080/0142159X.2017.1396306] [Medline: 29117744]
97. Triola MM, Hawkins RE, Skochelak SE. The time is now: using graduates' practice data to drive medical education reform. Acad Med 2018 Jun;93(6):826-828. [doi: 10.1097/ACM.0000000000002176] [Medline: 29443719]
98. Wise AF, Shaffer DW. Why theory matters more than ever in the age of big data. Learn Anal 2015 Dec 07;2(2):5-13. [doi: 10.18608/jla.2015.22.2]
99. Nebeker C, Torous J, Bartlett Ellis RJ. Building the case for actionable ethics in digital health research supported by artificial intelligence. BMC Med 2019 Jul 17;17(1):137 [FREE Full text] [doi: 10.1186/s12916-019-1377-7] [Medline: 31311535]
100. Thoma B, Warm E, Hamstra SJ, Cavalcanti R, Pusic M, Shaw T, et al. Next steps in the implementation of learning analytics in medical education: consensus from an international cohort of medical educators. J Grad Med Educ 2020 Jun;12(3):303-311 [FREE Full text] [doi: 10.4300/JGME-D-19-00493.1] [Medline: 32595850]
101. Herath HM, Mittal M, Kataria A. Navigating the metaverse: a technical review of emerging virtual worlds. WIREs Data Mining Knowl Discov 2024 Mar 29;14(4):e1538. [doi: 10.1002/widm.1538]
102. Boulos MN, Hetherington L, Wheeler S. Second life: an overview of the potential of 3-D virtual worlds in medical and health education. Health Info Libr J 2007 Dec;24(4):233-245 [FREE Full text] [doi: 10.1111/j.1471-1842.2007.00733.x] [Medline: 18005298]
103. Dafli EL, Vegoudakis KI, Pappas C, Bamidis PD. Re-use and exchange of an Opensim platform based learning environment among different medical specialties for clinical scenarios. In: Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine. 2009 Presented at: ITAB 2009; November 4-7, 2009; Larnaka, Cyprus. [doi: 10.1109/itab.2009.5394369]
104. Baniasadi T, Ayyoubzadeh SM, Mohammadzadeh N. Challenges and practical considerations in applying virtual reality in medical education and treatment. Oman Med J 2020 May 19;35(3):e125 [FREE Full text] [doi: 10.5001/omj.2020.43] [Medline: 32489677]
105. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021 Jul 08;9(3):e29080 [FREE Full text] [doi: 10.2196/29080] [Medline: 34255668]
106. Alrishan Alzouebi I, Saad S, Farmer T, Green S. Is the use of augmented reality-assisted surgery beneficial in urological education? A systematic review. Curr Urol 2021 Sep;15(3):148-152 [FREE Full text] [doi: 10.1097/CU9.0000000000000036] [Medline: 34552454]
107. Parsons D, MacCallum K. Current perspectives on augmented reality in medical education: applications, affordances and limitations. Adv Med Educ Pract 2021 Jan;Volume 12:77-91. [doi: 10.2147/amep.s249891]
108. Babaita AO, Kako M, Teramoto C, Okamoto M, Hayashi Y, Ohshimo S, et al. Face-to-face versus 360° VR video: a comparative study of two teaching methods in nursing education. BMC Nurs 2024 Mar 25;23(1):199 [FREE Full text] [doi: 10.1186/s12912-024-01866-4] [Medline: 38523319]

XSL•FO
RenderX

109. Behmadi S, Asadi F, Okhovati M, Ershad Sarabi R. Virtual reality-based medical education versus lecture-based method in teaching start triage lessons in emergency medical students: virtual reality in medical education. J Adv Med Educ Prof 2022 Jan;10(1):48-53 [FREE Full text] [doi: 10.30476/JAMP.2021.89269.1370] [Medline: 34981005]
110. Choi J, Thompson CE, Choi J, Waddill CB, Choi S. Effectiveness of immersive virtual reality in nursing education. Nurse Educ 2021 Oct 12;47(3):E57-E61. [doi: 10.1097/nne.0000000000001117]
111. Foronda CL, Gonzalez L, Meese MM, Slamon N, Baluyot M, Lee J, et al. A comparison of virtual reality to traditional simulation in health professions education: a systematic review. Simul Healthc 2024 Jan 01;19(1S):S90-S97. [doi: 10.1097/SIH.0000000000000745] [Medline: 37651101]
112. Koolivand H, Shooreshi MM, Safari-Faramani R, Borji M, Mansoory MS, Moradpoor H, et al. Comparison of the effectiveness of virtual reality-based education and conventional teaching methods in dental education: a systematic review. BMC Med Educ 2024 Jan 03;24(1):8 [FREE Full text] [doi: 10.1186/s12909-023-04954-2] [Medline: 38172742]
113. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959 [FREE Full text] [doi: 10.2196/12959] [Medline: 30668519]
114. García-Robles P, Cortés-Pérez I, Nieto-Escámez FA, García-López H, Obrero-Gaitán E, Osuna-Pérez MC. Immersive virtual reality and augmented reality in anatomy education: a systematic review and meta-analysis. Anat Sci Educ 2024;17(3):514-528. [doi: 10.1002/ase.2397] [Medline: 38344900]
115. Lewis KO, Popov V, Fatima SS. From static web to metaverse: reinventing medical education in the post-pandemic era. Ann Med 2024 Dec;56(1):2305694 [FREE Full text] [doi: 10.1080/07853890.2024.2305694] [Medline: 38261592]
116. Pokhrel S, Chhetri R. A literature review on impact of COVID-19 pandemic on teaching and learning. High Educ Future 2021 Jan 19;8(1):133-141. [doi: 10.1177/2347631120983481]
117. Espino-Díaz L, Fernandez-Caminero G, Hernandez-Lloret CM, Gonzalez-Gonzalez H, Alvarez-Castillo JL. Analyzing the impact of COVID-19 on education professionals. Toward a paradigm shift: ICT and neuroeducation as a binomial of action. Sustainability 2020 Jul 14;12(14):5646. [doi: 10.3390/su12145646]
118. Krebs C, Quiroga-Garza A, Pennefather P, Elizondo-Omaña RE. Ethics behind technology-enhanced medical education and the effects of the COVID-19 pandemic. Eur J Anat 2021;25(4) [FREE Full text]
119. Gopika JS, Rekha RV. Awareness and use of digital learning before and during COVID-19. Int J Educ Reform 2023 May 08 [FREE Full text] [doi: 10.1177/10567879231173389]
120. Voss M, Geniets A, Winters N. Strategies for digital clinical teaching during the COVID pandemic: a scoping review. Med Sci Educ 2024 Feb;34(1):219-235 [FREE Full text] [doi: 10.1007/s40670-023-01894-w] [Medline: 38510387]
121. Ratan R, Miller DB, Bailenson JN. Facial appearance dissatisfaction explains differences in zoom fatigue. Cyberpsychol Behav Soc Netw 2022 Feb;25(2):124-129 [FREE Full text] [doi: 10.1089/cyber.2021.0112] [Medline: 34842445]
122. Taylor J. Teaching and learning online: the workers, the lurkers and the shirkers. In: Proceedings of the 2nd Conference on Research in Distance and Adult Learning in Asia. 2002 Presented at: CRIDALA 2002; June 5-7, 2002; Hong Kong, China.
123. Bozkurt A, Koutropoulos A, Singh L, Honeychurch S. On lurking: multiple perspectives on lurking within an educational community. Int High Educ 2020 Jan;44:100709. [doi: 10.1016/j.iheduc.2019.100709]
124. Lollobrigida M, Ottolenghi L, Corridore D, Pingitore G, Damiano C, Serafini G, et al. Student evaluation of distance learning during the COVID-19 pandemic: a cross-sectional survey on medical, dental, and healthcare students at Sapienza University of Rome. Int J Environ Res Public Health 2022 Aug 19;19(16):10351 [FREE Full text] [doi: 10.3390/ijerph191610351] [Medline: 36011985]
125. Prestige E, Stander J, Wei Y. Covid lockdowns in the UK: estimating their effects on transmission. Signif (Oxf) 2022 Apr;19(2):14-18 [FREE Full text] [doi: 10.1111/1740-9713.01628] [Medline: 35601697]
126. Number and percentage of students enrolled in degree-granting postsecondary institutions, by distance education participation, location of student, level of enrollment, and control and level of institution: fall 2020 and fall 2021. National Center for Education Statistics. 2022. URL: https://nces.ed.gov/programs/digest/d22/tables/dt22_311.15.asp [accessed 2024-12-17]
127. Distance learning. National Center for Education Statistics. 2023. URL: https://nces.ed.gov/fastfacts/display.asp?id=80&ref=knowledgeecology.me [accessed 2024-12-17]
128. Zancajo A, Verger A, Bolea P. Digitalization and beyond: the effects of COVID-19 on post-pandemic educational policy and delivery in Europe. Policy Soc 2022;41(1):111-128. [doi: 10.1093/polsoc/puab016]
129. Vordenberg SE, Fusco NM, Ward KE, Darley A, Brady JH, Culhane NS, et al. An integrative review of micro-credentials and digital badges for pharmacy educators. Am J Pharm Educ 2024 Mar;88(3):100660. [doi: 10.1016/j.ajpe.2024.100660] [Medline: 38272238]
130. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence. AI Mag 1955;27(4):12. [doi: 10.1609/aimag.v27i4.1904]
131. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. J Gen Intern Med 2019 Aug;34(8):1626-1630 [FREE Full text] [doi: 10.1007/s11606-019-05035-1] [Medline: 31090027]
132. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410 [FREE Full text] [doi: 10.5334/pme.954] [Medline: 37868075]

XSL•FO
RenderX

133. Plesner LL, Müller FC, Nybing JD, Laustrup LC, Rasmussen F, Nielsen OW, et al. Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology 2023 May;307(3):e222268. [doi: 10.1148/radiol.222268] [Medline: 36880947]
134. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
135. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the united states medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
136. Prajapati M, Baliarsingh SK, Dora C, Bhoi A, Hota J, Mohanty JP. Detection of AI-generated text using large language model. In: Proceedings of the International Conference on Emerging Systems and Intelligent Computing. 2024 Presented at: ESIC 2024; February 9-10, 2024; Bhubaneswar, India. [doi: 10.1109/esic60604.2024.10481602]
137. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 2024 Feb 07;28(1):206-211. [doi: 10.1111/eje.12937] [Medline: 37550893]
138. Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, Hacihaliloglu I, et al. Diffusion models in medical imaging: a comprehensive survey. Med Image Anal 2023 Aug;88:102846. [doi: 10.1016/j.media.2023.102846] [Medline: 37295311]
139. Burr P, Kumar A, Young T. The potential of AI text-to-image generation in medical education: the educator and students' perspective. In: Beckingham S, Lawrence J, Powell S, Hartley P, editors. Using Generative AI Effectively in Higher Education. Milton Park, UK: Routledge; 2024.
140. Kumar A, Burr P, Young TM. Using AI text-to-image generation to create novel illustrations for medical education: current limitations as illustrated by hypothyroidism and Horner syndrome. JMIR Med Educ 2024 Feb 22;10:e52155 [FREE Full text] [doi: 10.2196/52155] [Medline: 38386400]
141. Waisberg E, Ong J, Masalkhi M, Lee AG. Concerns with OpenAI's Sora in medicine. Ann Biomed Eng 2024 Aug;52(8):1932-1934. [doi: 10.1007/s10439-024-03505-0] [Medline: 38558354]
142. Thomas RJ, Thomson TJ. What does a journalist look like? Visualizing journalistic roles through AI. Digit Journalism 2023 Jul 07:1-23. [doi: 10.1080/21670811.2023.2229883]
143. Latulippe K, Hamel C, Giroux D. Social health inequalities and eHealth: a literature review with qualitative synthesis of theoretical and empirical studies. J Med Internet Res 2017 Apr 27;19(4):e136 [FREE Full text] [doi: 10.2196/jmir.6731] [Medline: 28450271]
144. Logullo P, de Beyer JA, Kirtley S, Schlüssel MM, Collins GS. Open access journal publication in health and medical research and open science: benefits, challenges and limitations. BMJ Evid Based Med 2024 Jul 23;29(4):223-228 [FREE Full text] [doi: 10.1136/bmjebm-2022-112126] [Medline: 37770125]
145. Vukotich G. Healthcare and cybersecurity: taking a zero trust approach. Health Serv Insights 2023;16:11786329231187826 [FREE Full text] [doi: 10.1177/11786329231187826] [Medline: 37485022]
146. Keet K, Kramer B. Advances in digital technology in teaching human anatomy: ethical predicaments. Adv Exp Med Biol 2022;1388:173-191. [doi: 10.1007/978-3-031-10889-1_8] [Medline: 36104621]
147. Giansanti D. Cybersecurity and the digital-health: the challenge of this millennium. Healthcare (Basel) 2021 Jan 11;9(1):62 [FREE Full text] [doi: 10.3390/healthcare9010062] [Medline: 33440612]
148. Masters K. Ethics in medical education digital scholarship: AMEE Guide No. 134. Med Teach 2020 Mar;42(3):252-265. [doi: 10.1080/0142159X.2019.1695043] [Medline: 31835957]
149. Hussein R, Wurhofer D, Strumegger EM, Stainer-Hochgatterer A, Kulnik ST, Crutzen R, et al. General data protection regulation (GDPR) toolkit for digital health. Stud Health Technol Inform 2022 Jun 06;290:222-226. [doi: 10.3233/SHTI220066] [Medline: 35673005]
150. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022:10.31478/202206e [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
151. Jimenez G, Spinazze P, Matchar D, Koh Choon Huat G, van der Kleij RM, Chavannes NH, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
152. Jerry-Egemba N. Safe and sound: strengthening cybersecurity in healthcare through robust staff educational programs. Healthc Manage Forum 2024 Jan;37(1):21-25. [doi: 10.1177/08404704231194577] [Medline: 37725069]
153. Meng SQ, Cheng JL, Li YY, Yang XQ, Zheng JW, Chang XW, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
154. Wang AI, Tahir R. The effect of using Kahoot! for learning – a literature review. Comput Educ 2020 May;149:103818 [FREE Full text] [doi: 10.1016/j.compedu.2020.103818]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

AR: augmented reality
ARPANET: Advanced Research Projects Agency Network
BDA: big data analytics
CME: continuing medical education
GP: general practitioner
LMS: learning management system
mHealth: mobile health
MOOC: massive open online course
PLATO: Programmed Logic for Automated Teaching Operations
VR: virtual reality

Edited by G Eysenbach; submitted 07.05.24; peer-reviewed by F Sendra-Portero, KL Mauco; comments to author 11.07.24; revised version received 05.09.24; accepted 04.12.24; published 19.12.24.

Please cite as:
Ogundiya O, Rahman TJ, Valnarov-Boulter I, Young TM
Looking Back on Digital Medical Education Over the Last 25 Years and Looking to the Future: Narrative Review
J Med Internet Res 2024;26:e60312
URL: https://www.jmir.org/2024/1/e60312
doi:10.2196/60312
PMID:

©Oluwadamilola Ogundiya, Thahmina Jasmine Rahman, Ioan Valnarov-Boulter, Tim Michael Young. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Decades in the Making: The Evolution of Digital Health Research Infrastructure Through Synthetic Data, Common Data Models, and Federated Learning

Jodie A Austin[1,2], BPharm, DipClinPharm, PhD; Elton H Lobo[1], BE, MEng, PhD; Mahnaz Samadbeik[1,3], HIM, PhD; Teyl Engstrom[1], BMath, BBus, MEpi; Reji Philip[1,2], BSc, MCA, MIT; Jason D Pole[1,4], BHSc (Hons), MScEpi, PhD; Clair M Sullivan[1,5], MBBS (Hons), MD

[1]Queensland Digital Health Centre, Centre for Health Services Research, The University of Queensland, Brisbane, Australia
[2]The Office of the Chief Clinical Information Officer, eHealth Queensland, Brisbane, Australia
[3]Social Determinants of Health Research Center, School of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran
[4]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[5]Endocrinology Department, Royal Brisbane and Women's Hospital, Metro North Hospital and Health Service, Queensland Health, Brisbane, Australia

Corresponding Author:
Jodie A Austin, BPharm, DipClinPharm, PhD
Queensland Digital Health Centre
Centre for Health Services Research
The University of Queensland
Level 5, UQ Health Sciences Building
Fig Tree Cres
Brisbane, 4029
Australia
Phone: 61 7 3176 5530
Email: j.austin1@uq.edu.au

Abstract

Traditionally, medical research is based on randomized controlled trials (RCTs) for interventions such as drugs and operative procedures. However, increasingly, there is a need for health research to evolve. RCTs are expensive to run, are generally formulated with a single research question in mind, and analyze a limited dataset for a restricted period. Progressively, health decision makers are focusing on real-world data (RWD) to deliver large-scale longitudinal insights that are actionable. RWD are collected as part of routine care in real time using digital health infrastructure. For example, understanding the effectiveness of an intervention could be enhanced by combining evidence from RCTs with RWD, providing insights into long-term outcomes in real-life situations. Clinicians and researchers struggle in the digital era to harness RWD for digital health research in an efficient and ethically and morally appropriate manner. This struggle encompasses challenges such as ensuring data quality, integrating diverse sources, establishing governance policies, ensuring regulatory compliance, developing analytical capabilities, and translating insights into actionable strategies. The same way that drug trials require infrastructure to support their conduct, digital health also necessitates new and disruptive research data infrastructure. Novel methods such as common data models, federated learning, and synthetic data generation are emerging to enhance the utility of research using RWD, which are often siloed across health systems. A continued focus on data privacy and ethical compliance remains. The past 25 years have seen a notable shift from an emphasis on RCTs as the only source of practice-guiding clinical evidence to the inclusion of modern-day methods harnessing RWD. This paper describes the evolution of synthetic data, common data models, and federated learning supported by strong cross-sector collaboration to support digital health research. Lessons learned are offered as a model for other jurisdictions with similar RWD infrastructure requirements.

(J Med Internet Res 2024;26:e58637) doi:10.2196/58637

KEYWORDS

real-world data; digital health research; synthetic data; common data models; federated learning; university-industry collaboration

XSL•FO
RenderX

Background

While randomized controlled trials (RCTs) have long been accepted as the gold standard in evidence-based medicine, increasingly, there is a need to evolve this practice [1]. Well-designed RCTs are ideal for investigating the safety and efficacy of an intervention in a highly controlled setting, for example, treatment effects in drug development [2]. RCTs can fail to demonstrate the effectiveness of the intervention under complex, "real-world," dynamic conditions [3]. This can have serious cost implications for health systems when the outcomes promised under RCT conditions fail to deliver during postmarket surveillance [4]. Increasingly, health decision makers are focusing on real-world data (RWD) to deliver large-scale longitudinal insights that are actionable. RWD are collected as part of routine care in real time using digital health infrastructure [3,5]. Modern-day health research can capitalize on the benefits of RWD with a focus on translating the findings into clinical practice. Together, the findings generated through RCTs and RWD can bridge evidence gaps to support regulatory decision-making [6]. RWD "can provide valuable complementary evidence by answering important questions on treatment effects in clinical practice that are not answered by RCTs" [7]. Perspectives in medical research regarding RCTs as the only source of practice-guiding clinical evidence need to evolve. Certainly, the use of RWD for regulatory decision-making must address key considerations to ensure that the evidence generated is fit for purpose. This includes evaluation of data relevancy and quality, including accuracy, completeness, provenance, and transparency of RWD processing [8]. Steps to address these considerations are evident in the frameworks and policies emerging over the past decade, for example, to support the Food and Drug Administration (FDA) with harnessing RWD for postmarket safety surveillance [9]. Both data obtained through RCTs and RWD have their strengths and weaknesses (Textbox 1), further emphasizing a complementary approach to both methods in modern-day health research.

Textbox 1. Comparing data capture methods for randomized controlled trials (RCTs) versus real-world data (RWD).

Data capture for RCTs

- Demonstrate efficacy under controlled conditions (internal validity)
- Describe effect and causal relationships between an intervention and an outcome
- Data collected in a controlled and scheduled manner in accordance with the clinical trial
- Collected specifically to answer a small number of questions
- Other data regarding comorbidities may be incomplete or contain recall bias
- Intervention compared to either placebo or selected alternative
- Quality assessment tools used to review risk of bias resulting from imperfect RCT methodology
- Data elements centered on a specific research question with limited longitudinal insights

RWD

- Demonstrate effectiveness under real-world conditions (external validity)
- Describe the association or correlation between an intervention and an outcome
- Can be used to derive causal relationships but entail strong assumptions and rigorous methods, including evaluation of the RWD relevancy and quality
- Data often offer the advantage of being available in real time or near real time (recency of data capture)
- Provide a comprehensive picture of the patient (including details of the illness and social determinants)
- The same data used for clinical care are used for research purposes, noting that RWD can be subject to other forms of bias; for example, the care received may be a function of socioeconomic resources
- No control arm or intervention compared to standard treatment or care
- Evaluation of data quality is necessary to ensure accuracy, completeness, provenance, and transparency of processing
- Data assets may offer fragmented real-world trajectories across health systems

The interest in RWD for medical research has coincided with the rapid expansion of health IT (HIT), generating vast volumes of digital data through a myriad of sources. These include electronic medical records (EMRs), personal health records, wearable devices, mobile health, registries, and administrative data (such as claims and billing activities) [10]. However, the massive amounts of data now generated across various health care systems and platforms pose challenges in data integration and interoperability. The European Commission's funding initiatives, such as Horizon Europe and the Innovative Health Initiative, emphasize the importance of cross-sector collaboration and data integration to foster improved interoperability and advance health care research [11,12]. Other challenges faced by RWD capture for research include privacy and confidentiality concerns [13]. Using RWD for research requires the secondary use of the data for purposes other than those for which they were originally collected [14]. Ethical and governance considerations must reflect both social license and

XSL•FO
RenderX

privacy-protecting regulations. However, a difficulty faced by researchers in the digital era is conforming to regulatory frameworks established before digitization. While efforts are underway to integrate access to RWD for secondary use into updated legislation, novel methods are necessary to harness "big data" for digital health research. The same way that drug trials require infrastructure such as research nurses to support their conduct, digital health and the use of RWD also have research infrastructure needs [15]. These are not yet present in most academic institutions.

Health care research urgently requires the transformative power of data and HIT. Solutions are emerging to capture RWD siloed across HIT systems while addressing critical challenges such as interoperability, privacy, security, and effectiveness. This paper describes the rapid evolution of the medical research landscape and the ongoing development of modern-day research infrastructure. Such methods include common data models (CDMs) [16], federated learning (FL) [17], and synthetic data generation [18] supported by strong cross-sector collaboration. These novel methods are explored and, in turn, lessons learned are offered as a model for other jurisdictions with similar RWD infrastructure requirements.

Methodology

Health data collection methods have undergone significant evolution alongside the widespread adoption of HIT systems, EMRs, and other digital health technologies. To comprehensively understand this evolution, we conducted a review and perspective study, tracing the progression from traditional data capture methods such as RCTs to the integration of RWD into medical research. Our objective was to provide both a retrospective examination and a forward-looking perspective on the evolution of research infrastructure for digital health over the past 25 years. In our methodology, we outlined the trends and strategies identified through the rapid review to overcome barriers to using RWD and enhance health research infrastructure. We emphasized the incorporation of all available health data resources to ensure a comprehensive analysis, with continued attention to data privacy, ethical compliance in digital health, and mitigation of disclosure risk.

The Right Data for the Right Problem

Overview

To support the evolution of modern-day digital health research, a multifaceted approach, including synthetic data generation, mapping to CDMs, FL, and enablers to promote RWD extraction for research, is proposed. Figure 1 conceptualizes such an approach using CDM frameworks to support access to routinely collected health data, synthetic data generation, and FL infrastructure. Such an approach provides flexibility, offering the right data for the right problem at hand. Scenarios will always exist in research that require the extraction of identifiable or potentially reidentifiable patient information from data repositories for research purposes. In such circumstances, while the clinical validity of the data is high, so, too, can be the disclosure risk. Strict adherence to ethics and governance research protocols is essential. However, in recent years, there has been growing interest in alternative methods to harness RWD while minimizing disclosure risk. Methods to support RWD access in a deidentified manner, standardizing terminologies and mitigating the need for data sharing outside of enterprise structures are of particular focus. In doing so, the need to access identifiable or potentially reidentifiable patient health care data is minimized. The strategies identified to deliver each alternative method, balancing privacy concerns against clinical usefulness, are outlined in Figure 1.

XSL•FO
RenderX

Figure 1. Approaches to accessing data for modern-day health research. AI: artificial intelligence; CDM: common data model.

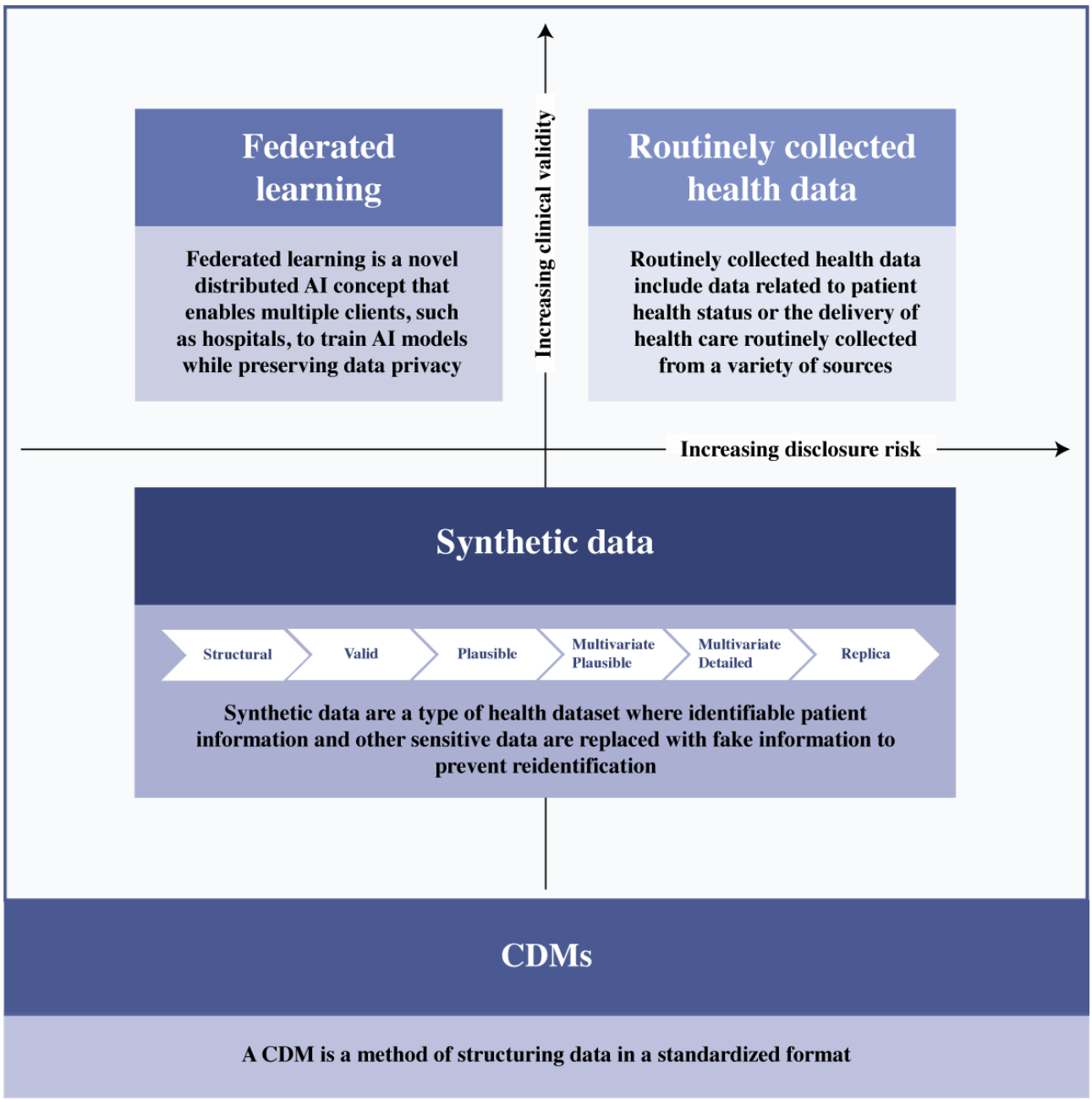

Goal 1: Synthetic Data Generation

Historically, accessing RWD has been associated with many challenges, such as laborious data access and consent procedures [19], particularly in environments in which privacy protection is prioritized and public scrutiny of digital privacy is rising [20]. Synthetic datasets, generated by a model to represent essential aspects of RWD [21], have been proposed to offer a solution for both privacy concerns and the need for widespread data access for analysis [22].

Synthetic datasets are generally classified into 3 broad categories: fully synthetic, partially synthetic, and hybrid [23]. Fully synthetic datasets entirely synthesize data without original values, ensuring privacy but compromising data validity [24-26]. In contrast, partially synthetic datasets replace selected attributes with synthetic values to preserve privacy while retaining original data, which is beneficial for imputing missing values [24-26]. Hybrid synthetic datasets combine original and synthetic data for strong privacy preservation, increasing data validity to help achieve a balance between privacy and fidelity [24-26]. However, there is a more detailed classification by the UK Office for National Statistics, which describes synthetic data in 6 levels [27], as shown in Figure 1. On the basis of this classification, a synthetic structural dataset (lowest level), developed solely from metadata, lacks clinical value and disclosure risk but is suitable only for basic code testing [27]. Conversely, a replica-level synthetically augmented dataset (highest level), which preserves format, structure, and patterns, offers high analytical value but increases disclosure risks due to its similarity to the original data [27]. The selection of synthetic data would depend on the nature of the application.

The use of synthetic data has a long-standing history dating back to the early stages of computing [28]. The early

XSL•FO
RenderX

foundational work of Stanislaw Ulam and John von Neumann in the 1940s, particularly focusing on the Monte Carlo simulation technique [29], is one such example. However, the notion of fabricating synthetic data to ensure valid statistical inferences and uphold disclosure control was first suggested by Rubin (as cited in the work by Raghunathan [22]) as a discussion of the work by Jabine (as cited in the work by Raghunathan [22]). Over time, the generation of synthetic data has moved from the use of statistical methods (eg, multiple data imputation and Bayesian bootstrap) [23] to more robust algorithms [30] due to the rise of several novel tools and services [23]. An early example is the synthetic minority oversampling technique algorithm, where synthetic data points are generated by selecting a predetermined number of neighbors for each underrepresented instance, randomly choosing some minority class instances, and creating artificial observations along the line between the selected minority instance and its closest neighbors [31]. This algorithm underwent maturation over time, leading to the emergence of several variants [32-35], which predominately focused on continuous variables but failed to identify nominal features when applied to datasets with categorical features, necessitating the creation of new labels for these attributes [36].

The introduction of deep learning methodologies, exemplified by the inception of variational autoencoders in 2013 and generative adversarial networks (GANs) in 2014, catalyzed the evolution of more promising paradigms in the domain of synthetic data generation [37]. GANs, most importantly [37], had the potential to generate synthetic data without direct engagement with the original dataset, a feature with potential implications for reducing disclosure risk [38]. The GAN model first proposed by Goodfellow et al [38] considers simultaneously training two neural network models: (1) a generative model that captures the data distribution and (2) a discriminative model that determines where the sample is generated from the model or data distribution (Figure 2) [39]. Initially, the generative model commences with noise inputs, devoid of access to the training or original dataset, relying on feedback from the discriminative model to generate a data sample [39]. Currently, GANs have gained a lot of interest due to their capability to produce high-quality synthetic data that closely match real data, especially in health care applications [40], including (1) forecasting and planning, (2) design and evaluation of new health technology and algorithms, (3) data augmentation, (4) testing and benchmarking, and (5) education [41].

Figure 2. Generative adversarial network model.

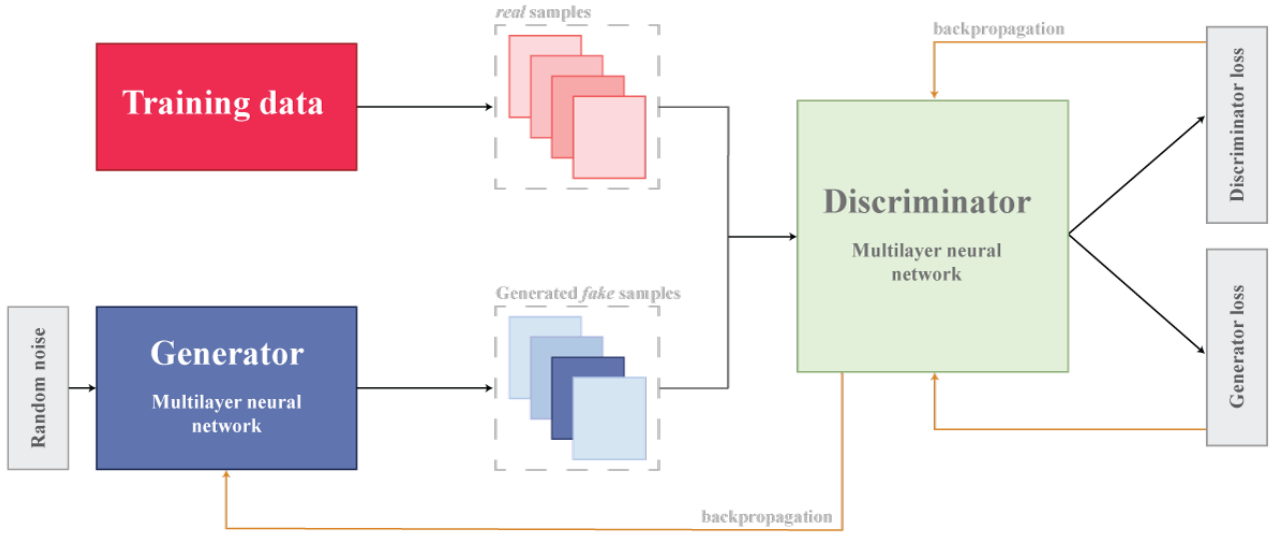

In the domain of published literature, GAN models are frequently discussed for their role in generating synthetic data [42-46]. However, various applications and services are now accessible for creating synthetic data tailored specifically for health care applications [23]. Among these tools are Synthea, implemented in Java; *DataSynthesizer* and *SynSys*, which are Python packages; and *synthpop* and *simPop*, both packages based on R [30,47,48]. *Synthea* uses the PADARSER (publicly available data approach to the realistic synthetic electronic health record) framework for synthetic data generation, relying on publicly available datasets instead of real electronic health records (EHRs) [49]. The framework emphasizes (1) using health statistics, (2) assuming no access to real EHRs, (3) integrating clinical guidelines, and (4) ensuring realistic properties in synthetic EHRs, as shown in Figure 3 [49].

synthpop uses regression trees for generating variables in a synthetic population but cannot handle complex data structures such as sophisticated sampling designs or hierarchical clusters (eg, individuals within households) [50], whereas *simPop* focuses on a modular object-oriented concept that uses various approaches, such as calibration through iterative proportional fitting and simulated annealing and modeling or data fusion through logistic regression, to generate a synthetic population [50]. In contrast, *DataSynthesizer* and *SynSys* use real patient data for the generation of synthetic datasets. For example, the *DataSynthesizer* includes 3 key modules for the generation of synthetic data: *DataDescriber*, which analyzes attribute types and distributions while preserving privacy; *DataGenerator*, which uses this analysis to create synthetic data; and Model Inspector, which provides an intuitive summary for evaluation and adjustment of parameters [51]. *SynSys* uses real data to train Markov and regression models to generate more realistic synthetic data, as shown in Figure 4 [30].

XSL•FO
RenderX

Figure 3. PADARSER (publicly available data approach to the realistic synthetic electronic health record) framework reproduced from Walonoski J et al [49], which is published under Creative Commons Attribution 4.0 International License [52]. EHR: electronic health record.

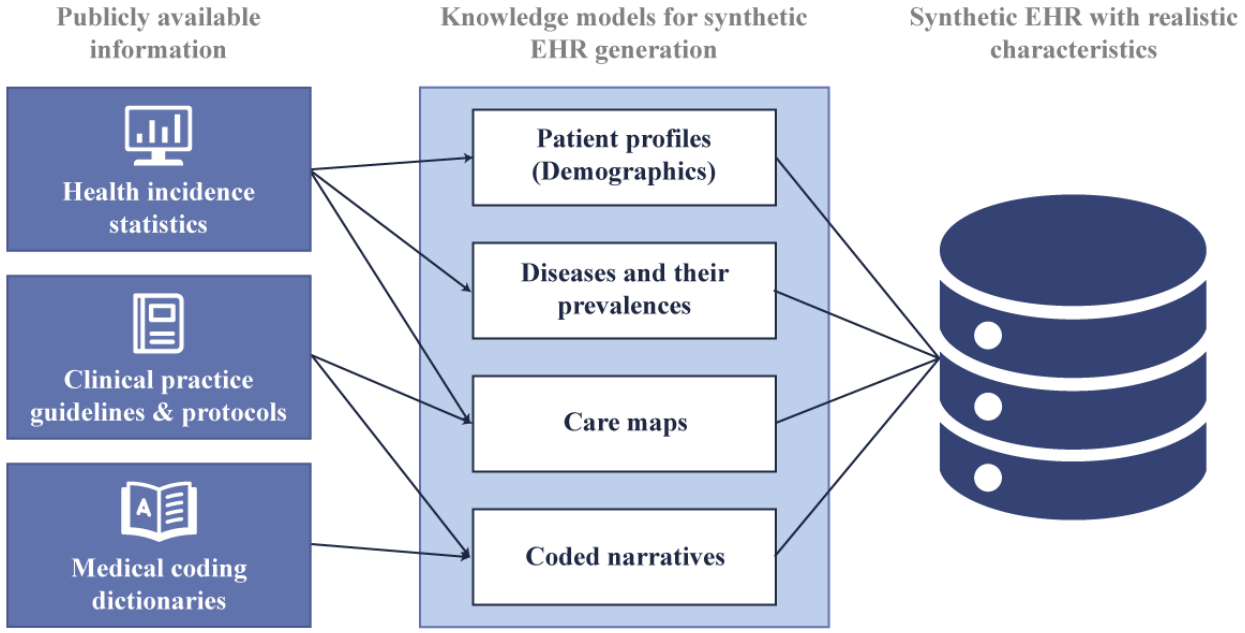

XSL•FO
RenderX

Figure 4. SynSys model adapted from Dahmen J et al [30], which is published under Creative Commons Attribution 4.0 International License [53].

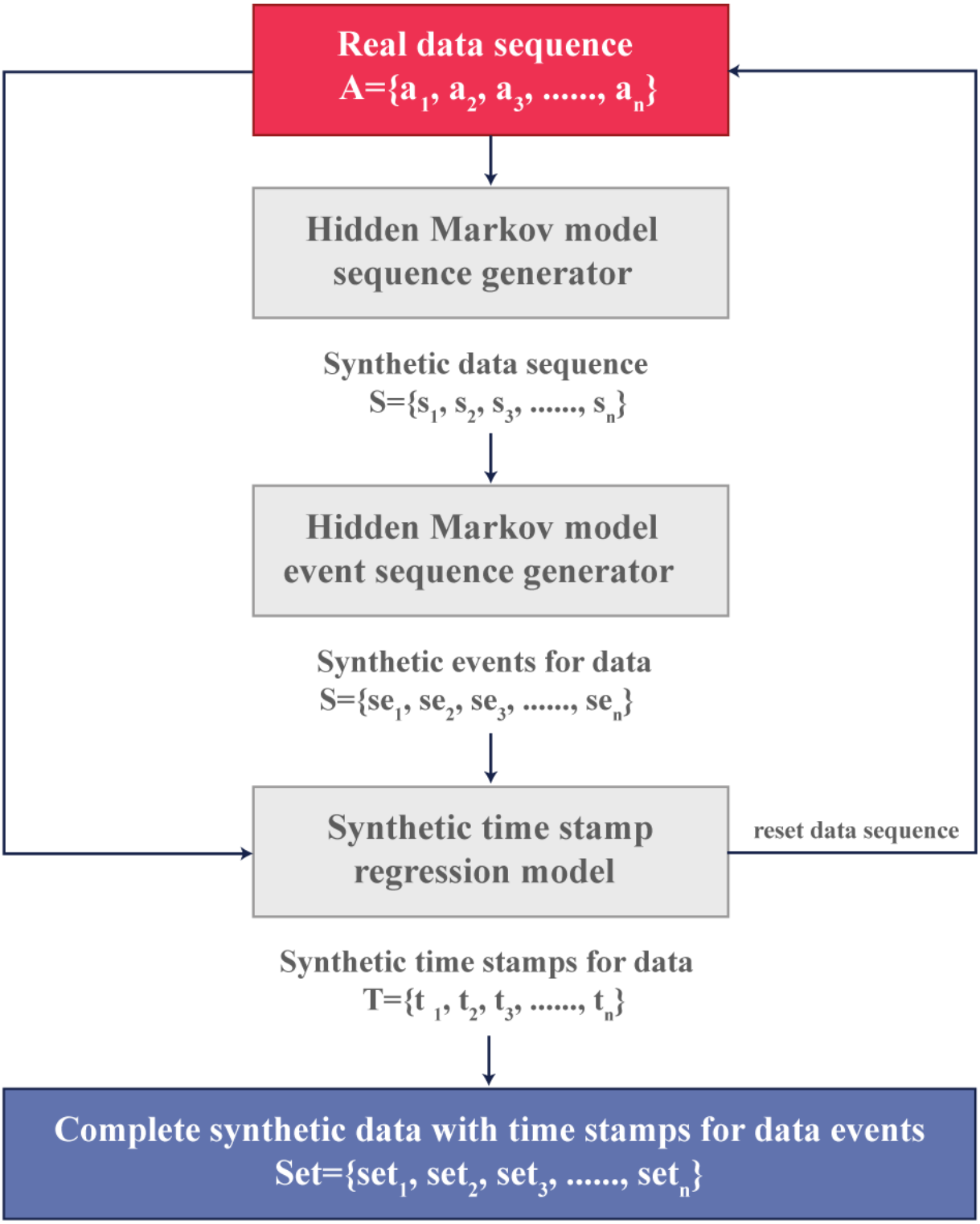

Goal 2: CDMs

Sharing clinical data, including clinical trial data, for research is increasingly recognized as an efficient way to advance scientific knowledge [54]. However, the sharing of clinical data in health care is not without its challenges, with research highlighting concerns related to privacy, security, and interoperability [55]. While literature exists with regard to mitigating privacy and security issues in clinical data sharing for research purposes, interoperability issues persist [56]. One potential solution that has been touted to limit issues related to interoperability are CDMs [55].

CDMs are commonly used in research to enable the exchange or sharing of datasets for specific purposes [57]. The objective of a CDM is to streamline the conversion of data from diverse databases into a consistent format with standardized terminology, thereby enabling systematic analysis [58]. Over the past decade, several CDMs have been collaboratively

XSL•FO
RenderX

developed and risen to the level of de facto standards for clinical research data. These include the Health Care Systems Research Network (formerly known as the HMO Research Network) Virtual Data Warehouse, the National Patient-Centered Clinical Research Network CDM, the Observational Medical Outcomes Partnership (OMOP) CDM, the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model, and the Sentinel CDM [59].

The CDISC was one of the oldest known CDMs, established in 1998, and has been pivotal in streamlining clinical data acquisition, interchange, and submission processes. With its 12 domains (Textbox 2) and unique variable naming conventions, the CDISC ensures clarity and consistency in data representation. However, it mainly aims to provide guidelines rather than imposing strict data collection requirements, allowing for flexibility for different study designs and objectives [60].

Textbox 2. Clinical Data Interchange Standards Consortium domains and their data structures [60].

Demographics: 1 record per subject
Disposition: 1 record per subject
Exposure: 1 record per subject per phase or dose
Adverse events: 1 record per subject per adverse event
Concomitant medications: 1 record per subject per medication
Serum chemistry: 1 record per subject per visit per measurement
Hematology: 1 record per subject per visit per measurement
Urinalysis: 1 record per subject per visit per measurement
Electrocardiogram: 1 record per subject per visit
Vital signs: 1 record per subject per visit (per position)
Physical examination: 1 record per subject per examination, body system, or finding
Medical history: 1 record per subject per examination, body system, or condition

Another significant CDM, Sentinel, initiated as part of the FDA's Sentinel Initiative to monitor FDA-regulated medical products on a national scale [61]. It uses standardized concept codes with 19 tables (Textbox 3) [62], although users may need to map data due to variations in coding systems [63]. On the other hand, the Health Care Systems Research Network Virtual Data Warehouse aims to centralize data extraction and loading processes across 17 health care systems in the United States [64]. Its comprehensive structure comprises 7 content areas and >450 variables spread across 18 tables, as illustrated in Figure 5 [64], enhancing research efficiency by consolidating data management efforts [64].

XSL•FO
RenderX

Figure 5. Health Care Systems Research Network Virtual Data Warehouse common data model, modified from Ross TR et al [62], which is published under a Creative Commons Attribution 4.0 International License [65]. AHFS: American hospital formulary service; DX: diagnostic; EverNDC: Ever National Drug Code; GPI: generic product identifier; LOINC: Logical Observation Identifiers Names and Codes; MD: medical doctor; NDC: National Drug Code; Rx: prescription.

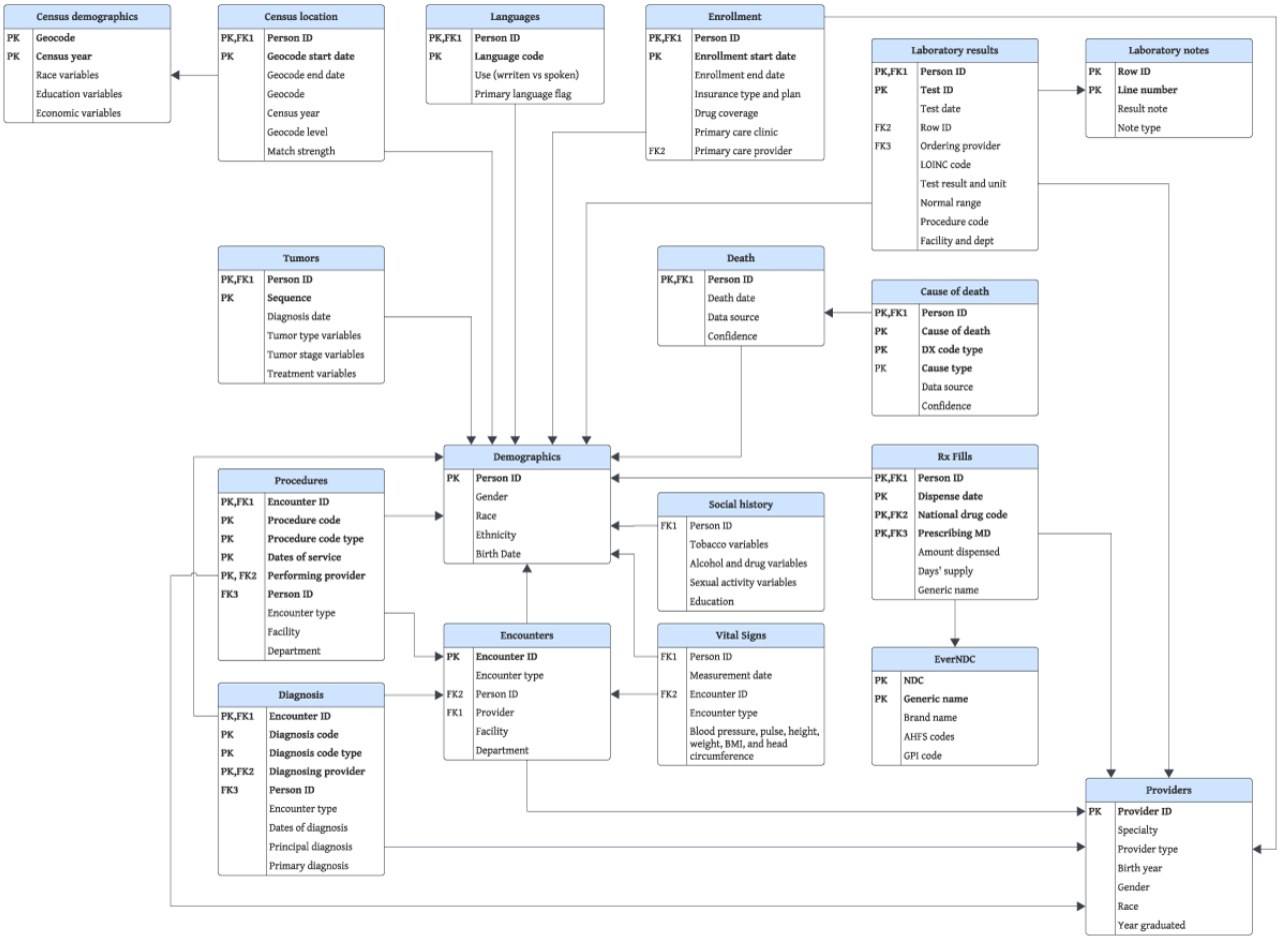

XSL•FO
RenderX

Textbox 3. Sentinel Common Data Model [62].

Administrative data

- Enrollment
- Demographic
- Dispensing
- Encounter
- Diagnosis
- Procedure
- Prescribing

Mother-infant linkage data

- Mother-infant linkage

Auxiliary data

- Facility
- Provider

Feature engineering data

- Feature engineering

Registry data

- Death
- Cause of death
- State vaccine

Inpatient data

- Inpatient pharmacy
- Inpatient transfusion

Clinical data

- Laboratory test results
- Vital signs

Patient-reported measure (PRM) data

- PRM survey
- PRM survey response

The National Patient-Centered Clinical Research Network was implemented to support patient-centered studies and stands out for its expansive data coverage, storing information from >100 million individuals [66] in a common format across its 23 interconnected tables [67]. It incorporates actual dates and a unique patient identifier for efficient data navigation, ensuring data integrity and facilitating comprehensive analysis [68]. In addition, the Observational Health Data Sciences and Informatics program focused on standardizing medical data representation across diverse source systems [69]. With its OMOP CDM comprising 18 tables [70], the Observational Health Data Sciences and Informatics program integrates data from >100 databases worldwide [71], addressing the need for standardized EHR data and consistent patient-level information in observational databases [69], as shown in Figure 6 [72].

XSL•FO
RenderX

Figure 6. Observational Medical Outcomes Partnership Common Data Model, reproduced from Jiang G et al [72], which is published under Creative Commons Attribution 4.0 International License [52].

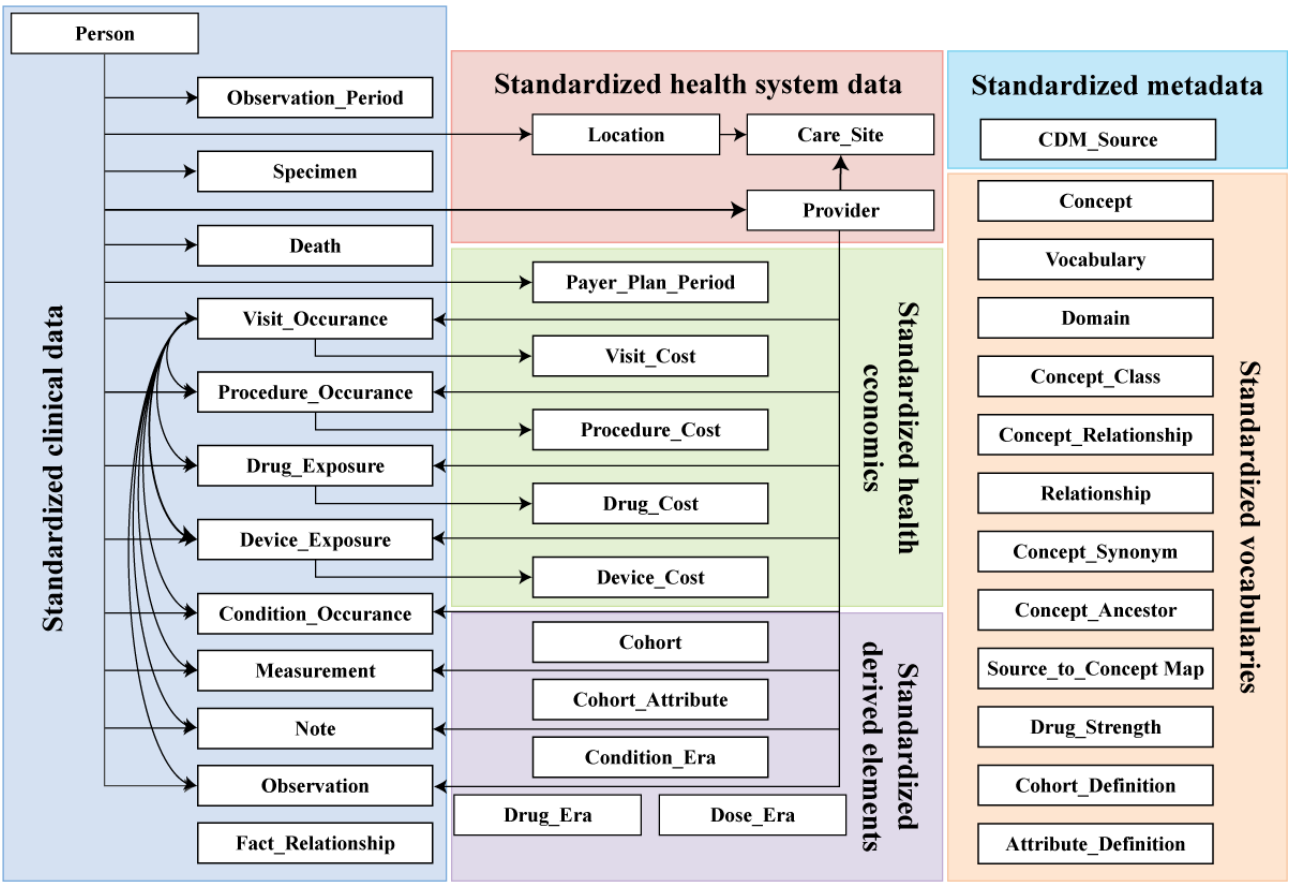

Figure 7. Extract, transform, and load process adapted from the work published by Abd Al-Rahman SQ et al [73], under the CC-BY-SA license [74].

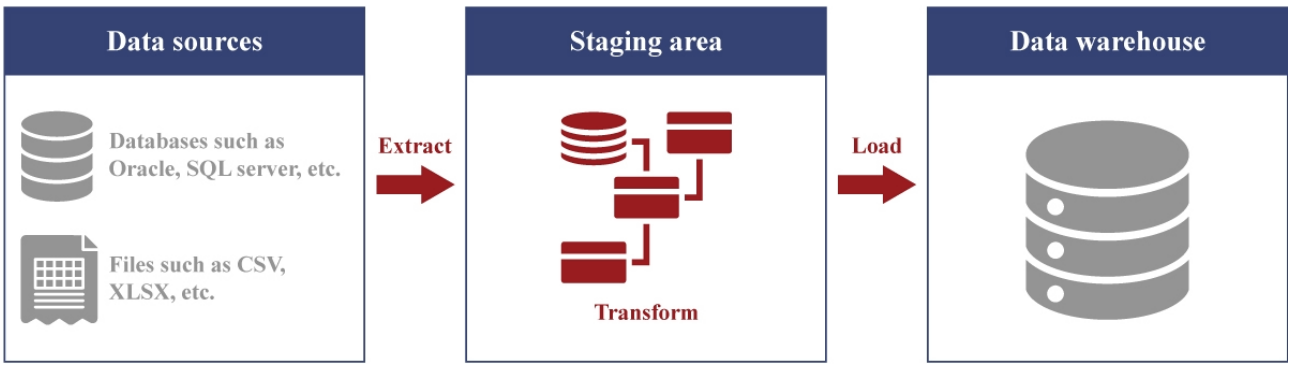

While most approaches follow a different structure for storing their data, most of these models use an extract, transform, and load (ETL) process to map the data from the source database to the target structure, as shown in Figure 7 [73]. The source database may come from hospital-wide systems (such as Epic EMRs, Oracle Health, and Meditech [75]) or departmental systems (such as MOSAIQ by Elekta [76], ARIA by Varian [76], picture archiving and communication systems [77], pathology or laboratory information systems [78], and others).

The ETL process operates via 3 principal stages: extraction, transformation, and loading [79]. Extraction refers to retrieving data from relevant sources, often in file formats such as CSV [79], relational databases such as MySQL [79], nonrelational databases such as NoSQL [80], graph databases such as Neo4j [81], or accessed through Representational State Transfer clients [79]. Transformation entails the refinement and adaptation of the data to conform to the prescribed schema, encompassing tasks such as normalization, deduplication, and quality validation procedures [79]. This stage may also involve aligning the data with standardized terminologies such as the Systemized Nomenclature of Medicine–Clinical Terms or the International Classification of Diseases to ensure semantic consistency and interoperability across systems [82] or understanding preexisting standards (such as Digital Imaging and Communications in Medicine [83], the National Council for Prescription Drug Programs SCRIPT standard [84], and so on) toward mapping relevant information. Loading involves the transfer of the refined data into operational databases, data marts, or data warehouses for subsequent use [79].

XSL•FO
RenderX

Goal 3: FL

Traditional centralized machine learning (ML) approaches face privacy and security risks [85] and limited predictive accuracy due to single-source data constraints [86]. To limit these challenges, FL has emerged as a solution by facilitating distributed model training on local devices. Google introduced FL in 2016, which uses distributed learning platforms to leverage enhanced computational abilities of devices, connect devices executing local training models, and facilitate cooperation among devices to build consensus global models of learning [87]. FL offers a secure and efficient approach to analyzing fragmented health care data [88]. This decentralized approach reduces the risk of data exposure and vulnerability to cyberattacks [89].

Over the past years, there has been a notable trend regarding how medical data are processed and used. EMRs play an important role in health care data collection and retrieval. However, strict regulations on data sharing necessitate the anonymization of sensitive patient attributes [90]. Health care organizations face challenges in aggregating clinical records for deep learning models due to privacy, data ownership, and legal concerns. Balancing data protection with leveraging collective knowledge is challenging [88]. In health care, FL initiatives are emerging as a privacy-enhancing approach to artificial intelligence and ML. These initiatives aim to collaboratively train predictive models across various institutions without centralizing sensitive personal data [91]. Recently, FL has been applied to the health care domain and life science industry, addressing the need for high-quality models in ML applications [92,93]. The FL paradigm has gained popularity for its scalable and privacy-preserving approach to joint training across federated health data repositories [85,93,94]. FL develops ML models over distributed datasets in locations such as hospitals, laboratories, and mobile devices, ensuring data privacy [88]. FL aims to overcome barriers associated with transferring sensitive clinical data to a central repository in conventional centralized artificial intelligence and ML models [85]. This approach allows for training of ML models on distributed client nodes, preserving the privacy and integrity of patient data [85]. The core concept of FL involves sharing only the parameters of the ML model being trained rather than sharing the actual data [94-96].

The FL methodology involves a network of nodes, each sharing models instead of raw training data with the central server. FL is conducted iteratively as follows. Initially, the server distributes the current global ML model parameters to all participating edge nodes. Each node then uses its locally stored data samples to update its own model based on the received parameters. Subsequently, each node transmits its updated model parameters back to the server. The server performs a global aggregation operation, combining and weighting the model parameters received from each node to generate a new set of global model parameters. This process is iterated multiple times until convergence. Importantly, at no stage do the nodes share their training data with each other or the central server, enhancing privacy and reducing bandwidth use [87,97,98].

Goal 4: Cross-Sector Collaboration (Enablers to Promote RWD Access for Research)

The methods outlined previously provide novel approaches to RWD (or simulated RWD) access to promote digital health research. While these methods may meet most digital health research requests, access to ethically approved identifiable RWD cannot be dismissed. However, a conundrum in the digital era, with EMRs now generating vast volumes of health care data, is the limited skilled informaticians trained in data extraction and analysis. The Joint Science Academies Statement on Global Issues specific to “Digital Health and the Learning Health System” noted the basic requirement of developing and cultivating a digital health workforce, stating that “the training challenge for leveraging digital health is vast—in health care, public health and biomedical science” [99]. Those trained in data extraction are often focused on the operational activities of the health care organization. Support is needed to streamline RWD extraction for digital health research. Assigning domain experts to handle the manual data extraction steps to support researchers with access to medical RWD is necessary [100]. Academia-industry digital health collaborations can leverage uniquely skilled resources and networks to benefit both sectors [101]. Embedding staff with affiliations to both the university and health care sectors is one potential method. To overcome barriers related to university-industry collaboration, an environment fostering the missions of both sectors is necessary [102]. Being cognizant of the notable differences between the primary cross-sector objectives is necessary, for example, feasible timelines and balancing competing demands [103]. This approach is explored further in the use case below.

Use Case

In reviewing the evolution of digital techniques used to harness RWD, consideration must be given to the application of such methods to support modern-day research. An illustrative use case is provided in this section to offer a forward-looking perspective on where such techniques may be headed.

A center dedicated to digital health research was established in Queensland, Australia. The center spanned 6 university faculties, collaborating with external government and industry partners. To overcome the challenges of harnessing RWD for research, the center established a service offering a multifaceted approach to RWD access (Figure 1). The needs and current and future state of each research infrastructure goal have been summarized in Table 1.

XSL•FO
RenderX

Table 1. The needs and current and future state of the research infrastructure goals of a center dedicated to digital health research established in Queensland, Australia.

	Synthetic data	CDMs[a]	FL[b]	Routinely collected health data (EMR[c])
Needs	• Access to large-scale RWDd for research, minimizing the risk of patient disclosure • Support for EMR training and education • Support for clinical analytics tool development • Development of AIe pipelines before RWD access	• Ability to collaborate on research both nationally and internationally using disparate clinical and administrative datasets • Access to large-scale RWD for research, minimizing the risk of patient disclosure • Reduction of burden on informaticians to run bespoke research data extracts	• Promote RWD sharing across organizations while maintaining data privacy • Enable model training without centralizing sensitive data, preserving individual user privacy • Keep data local and share only model updates, minimizing the risk of data breaches	• Support clinicians and researchers with access to RWD • Reduce burden on informaticians to run bespoke research data extracts through the establishment of a dedicated team working across sectors
Current state	• Semirepresentative data displaying univariant distributions sourced from publicly available health statistics	• Local statewide EMR data transformed to the OMOPf CDM within the training (nonproduction) environment [104]	• Established governance, ethics, and data custodian approvals • Health databases relevant to a specific chronic disease use case standardized to the OMOP CDM, synthetically generated and shared with FL clients to test FL model	• Established team of clinical informaticians and data engineers holding conjoint positions across both the university and health care sectors • Contractual agreements established to demarcate the roles and responsibilities of the conjoint staff members accessing dual networks
Future state	• Representative synthetic data mirroring multivariant distributions from the local statewide EMR	• Local statewide EMR data transformed to the OMOP CDM within the production environment	• Technology established with the ability to demonstrate privacy and security using synthetic data • Scalable and reliable infrastructure for FL in health care designed to handle large volumes of data from diverse sources • Establishment of national infrastructure for FL in digital health to generate new models of care	• Continued expansion of the service to promote digital health research, including data extraction beyond the statewide EMR

[a]CDM: common data model.

[b]FL: federated learning.

[c]EMR: electronic medical record.

[d]RWD: real-world data.

[e]AI: artificial intelligence.

[f]OMOP: Observational Medical Outcomes Partnership.

The infrastructure goals highlighted in Table 1 draw upon techniques emerging in recent decades through the maturation of digital health technologies and strong cross-sector collaborations. The use case signifies how organizations are joining forces to advance modern-day research through RWD capture. No individual goal was deemed superior, yet through commitment to drive each approach to RWD access (Figure 1), this dedicated service is a method for providing researchers with the right data for the right problem.

Discussion

Overview

The evolution of digital health has seen many health care organizations shifting beyond the foundational levels of implementation to established methods of harnessing RWD to promote a learning health system [105]. A learning health system needs academic inquiry brought close to the routinely generated health care data, yet data security and privacy must remain paramount. While the clinical validity of the data is always greatest via direct access and extraction from the data source, so, too, is the disclosure risk. Novel methods have emerged and evolved to support access to RWD for modern-day health care research. Application of these techniques over time

XSL•FO
RenderX

has provided an opportunity to reflect on the emerging needs, including the strengths and weaknesses of each goal and the future directions. In addition, the lessons learned for the described digital health research center case in point (Table 1) are included for each goal in the following sections.

Synthetic Data Strengths, Weaknesses, and Lessons Learned

Synthetic data generation has made significant advancements in recent decades, from statistical methods to robust algorithms and established applications and services tailored to synthetic data generation for health care needs. The synthetic data created by the various models have the potential to reduce costs and accelerate data generation [106]. As such, synthetic data can have numerous applications in health care, such as estimating the impact of policies, augmenting ML algorithms, and improving predictive public health models [29]. Although synthetic data hold promise, significant work needs to be done to make them a clear option to replace RWD [107]. The reason for this conundrum is the lack of a clear understanding as to whether such a dataset can be used for decision-making or whether the final analysis would require original data [108]. Locally, the use of semirepresentative synthetic datasets (Table 1) has been effective in supporting researchers with projects less reliant on accurate representations within the synthetic data to enable research to progress while awaiting the necessary approvals to access production data. Example projects include the support of qualitative focus group sessions to co-design clinical analytics tools or development of the infrastructure for future FL projects. Work continues to explore whether similar results and accurate conclusions can be drawn from representative synthetic data when compared to RWD, with some demonstrating promising results [109,110].

Synthetic data are not free from bias [111], privacy [112], and data quality assessment [41] issues. Bias, inherent in human society, especially affects marginalized groups and is reflected in data access and generation [113]. This poses a risk with ML algorithm adoption, potentially perpetuating or amplifying societal biases [111]. Regarding privacy, while synthetic data have been claimed to be a potential solution for mitigating privacy concerns, Stadler et al [112] highlight that synthetic datasets often contain residual information from their training data, making them vulnerable to ML-based attacks that can reveal features preserved by the generative model. However, it is challenging to predict the type of information retained in synthetic data or the specific features targeted by adversaries, thereby complicating the assessment of the privacy benefits provided by synthetic data generation [112]. In addition, Stadler et al [112] explain that differential privacy, a technique used in synthetic data generation to inject noise into the original statistical information for enhanced privacy [114], provides limited defense against ML-based inference attacks, particularly for high-dimensional datasets [112]. The evaluation of data quality is another such issue, which remains an open challenge [115]. The problem arises from the absence of a standardized quality metric, which impedes fair and definitive comparisons between methods, consequently affecting the selection of an appropriate approach [41]. As a consequence of these issues, there is a crucial need for tailored regulations on synthetic data use in medicine and health care to ensure quality and minimize potential risks [116].

Synthetic data frequently reside in a regulatory gray zone concerning their use [117], and existing data protection laws such as the General Data Protection Regulation and Health Insurance Portability and Accountability Act (HIPAA) have constraints in adequately addressing all potential risks linked to synthetic data [29]. For instance, HIPAA's privacy rule considers the creation of deidentified data as a health care operation, thus exempting them from the need for patient consent, a principle similarly applied in the General Data Protection Regulation [117]. However, synthetic health data, while not deidentified, closely replicate real data, raising questions about whether they should be classified as protected health information and require informed consent and research ethics review [117]. Some studies have demonstrated the use of synthetic data in research, eliminating the need for an ethics review [118]. Whether this is a scalable future direction for synthetic data use in research remains to be seen.

CDM Strengths, Weaknesses, and Lessons Learned

The past 2 decades have seen the emergence of numerous CDMs to support collaborative health care research through data standardization. For example, the use of the OMOP CDM to conduct observational studies has grown extensively in recent years (from 14 publications in 2016 to 57 publications in 2020) [119], and its utility has been demonstrated in numerous, large-scale, multinational studies, such as estimating comparative drug safety and effectiveness [120-122]. The benefits are obvious for observational research in the digital era, when research questions can be addressed through combining databases with different underlying models, different information types, and different coding systems. What must not be overlooked is the potential for different biases to exist within different datasets and these nuances to be lost during translation to the CDM. Due to the complex transformations between sources and targets with varying schemas, databases, and technologies, the ETL implementations are considered prone to faults or issues [123].

According to Nwokeji and Matovu [124], these issues include complexity, cost, data heterogeneity, lack of automation, maintenance, standardization, and time. First, the growing complexity of data structures presents formidable obstacles to devising streamlined strategies [124]. In addition, the cost-intensive nature of ETL solution development imposes significant financial burdens [120]. Data heterogeneity, stemming from diverse sources and formats, further complicates the integration process [124]. Many existing ETL solutions continue to rely on manual procedures or necessitate human intervention, indicating an incomplete transition toward automation [124]. A lesson learned through the local mapping of the statewide EMR to the OMOP CDM within a nonproduction environment [104] (Table 1) highlighted the requirement for a joint clinical and technical venture. Establishing appropriate governance structures with input from clinical and technical staff is necessary to clearly articulate and endorse CDM implementation and ongoing maintenance decisions. Maintenance of ETL solutions is rendered demanding

XSL•FO
RenderX

by the variety of data schemas and the dynamic nature of application requirements [124]. Furthermore, the absence of standardized methodologies for modeling ETL processes and executing workflows exacerbates these challenges [124]. Finally, the protracted process of designing, developing, implementing, and executing ETL solutions entails considerable time investments [124]. Despite these challenges, a multitude of commercial tools, including Microsoft SQL Server Integration Services, Oracle Warehouse Builder, IBM InfoSphere, and Informatica PowerCenter, alongside open-source alternatives such as Talend Open Studio and Pentaho Kettle, serve to facilitate and simplify these processes [79].

To address interoperability issues, the use of CDMs continues to expand within the health domain. Areas of future focus include the ongoing development of CDMs, their vocabularies, and tools to support their use. Further work is warranted to establish guidelines for CDM development [125] and achieving consensus on governance practices across institutions using RWD for secondary purposes [104].

FL Strengths, Weaknesses, and Lessons Learned

Of the goals discussed, FL is the most recent technique emerging in the field of RWD access. This technology allows for learnings to be obtained from health data across organizations and locations without attempting traditional integration [87,97]. The adoption of FL in the health care domain addresses the challenges of data privacy, confidentiality, and security while still enabling efficient model training [126]. Existing works on FL in the health sector reveal a diverse range of applications categorized into prognosis, diagnosis, and clinical workflow. Prognosis-related applications encompass endeavors such as stroke prediction and prevention, brain data meta-analysis, and brain tumor segmentation [88,127,128]. Diagnosis-related applications include COVID-19 diagnosis, morphometry for Alzheimer disease, and heart disease predictions from EHRs [88,129,130]. In addition to prognosis and diagnosis, FL holds significant potential in optimizing clinical workflows within the health care sector. These applications encompass various aspects, such as drug sensitivity prediction, integration of medical data, and clinical decision support systems [88,131,132]. These advancements highlight FL in streamlining clinical workflow efficiencies, enhancing patient care, and fostering innovation in health care delivery [88]. The application of FL demonstrates its potential to enhance health care outcomes while preserving data privacy and security, highlighting the significance of interdisciplinary research and innovative solutions in advancing FL across scientific domains.

Despite the numerous advantages of FL, this methodology presents several challenges that must be addressed for its effective implementation in scientific settings. The challenges facing FL can be categorized into several critical domains. First, privacy and security concerns arise from compromised servers or clients, potentially jeopardizing data integrity and confidentiality, with active and passive attacks posing threats to overall data security [87,88,91,133,134]. The distributed nature of FL gives rise to potential new privacy and security issues that must be avoided, including the leakage of sensitive patient information (privacy) and poisoning of data (security) [135]. Second, communication bottlenecks exacerbate these challenges, hindering seamless data exchange between clients and servers and raising issues regarding network state and protocol efficacy [87,88,91]. Third, addressing the heterogeneity in data distribution poses significant challenges, particularly in handling nonindependent and non–identically distributed data [88,91,136]. Fourth, the rising computing costs, especially considering the varied capabilities of devices, highlight the critical need to address challenges related to asymmetric computing and mitigate concerns regarding energy consumption in scenarios involving on-device training [85,88,91]. Moreover, the reliability of central servers responsible for managing local training and updates is also uncertain, increasing the likelihood of data leakage and security breaches [87,88,137]. Finally, the development of new FL computing frameworks, which include redundant servers, hardware accelerators, and decentralized training models, necessitates a comprehensive and thorough investigation [87,138]. These multifaceted challenges highlight the urgent need for interdisciplinary research and innovative solutions to facilitate the successful implementation and advancement of FL across scientific domains.

FL offers a novel approach to collaborative training across health care data repositories, bypassing the need for data sharing and safeguarding sensitive medical information [139]. In the use case provided in this paper (Table 1), the process involved a combination of approaches. Standardizing the data from health databases such as EMRs and health registries via a CDM was necessary, including provision to the FL client to then test the FL model using a synthetically generated dataset. This innovative method has the potential to address various health care issues by using distributed datasets across health care facilities. By doing so, it creates opportunities for pioneering research and business opportunities in the future of health care. Researchers will focus on integrating FL into upcoming medical devices such as intelligent implants and wearables. This will lead to the development of new eHealth services, improving patient well-being.

Personalization is key in preventive health care and chronic disease management through tailored interventions. It is expected that FL will drive precision medicine and elevate health care standards in the coming years. FL also stands to transform health care delivery, offering improved precision, accessibility, and patient-centered care [85,88,97,139].

Looking ahead, challenges such as ensuring data quality and incorporating expert knowledge into FL models need attention. Designing effective incentive mechanisms is crucial to encourage users of mobile and wearable devices to participate in the FL process. This participation involves these devices collecting high-quality data locally, training local models, and sharing model updates with a central server.

Cross-Sector Collaboration: Enablers to Promote RWD Access for Research

Digitization can accelerate RWD access through the novel technical methods emerging in recent decades. However, a holistic approach is necessary to support modern-day research in a system as multifaceted as that of digital health. The types of collaboration between the university and industry or health

XSL•FO
RenderX

care sectors to drive digital transformation are varied [140]. Human factors are as important as the technologies themselves. Rybnicek and Königsgruber [141] identified 4 categories to drive the success of these cross-sector collaborations: institutional factors, relationship factors, output factors, and framework factors. The illustrative use case (Table 1) supports this approach. Embedding staff members across both types of organizations with access to both academic and health care networks and governed by the policies and procedures of the health care sector was key to supporting RWD access for research. Contractual agreements were critical to outline the key roles and responsibilities of conjoint staff, the governance frameworks by which they must abide, and clear reporting lines across both organizations. Colocation was deemed essential to build the relationship and trust. This takes both time and commitment from both sectors. As organizations continue to strive for advancements in HITs, it is the interpersonal relations that are fostering this growth. "As much as we talk about technology, at the end of the day collaboration is about people" [140].

Conclusions

The past 25 years have seen a maturation in digital health at large. HITs are opening new and efficient ways to deliver patient care. This evolution of patient care delivery and its ability to digitally capture data through routine care has underpinned the progression of medical research techniques. A shift in perspective is necessary, moving away from the emphasis on RCTs as the only source of practice-guiding clinical evidence to include the use of RWD. Novel methods are necessary to harness the vast volumes of RWD now generated through these digital platforms. Techniques such as synthetic data generation, CDMs, FL, and collaborations between the health care and university sectors all support this common goal. Appropriate policies and frameworks are essential to address the challenges of using RWD for research. We demonstrated how, by mapping health care data to a CDM and generating a synthetic dataset, these approaches facilitate the establishment of FL infrastructure, highlighting the interoperability of these methodologies across various research environments. To achieve a learning health system, a new and disruptive research infrastructure must be established, maintained, and enhanced to expedite the translation of research findings into clinical practice. This infrastructure, equipped with emerging digital health techniques and supported by strong cross-sector collaborations, advances research by enabling more effective RWD capture, providing researchers with "the right data for the right problem."

Conflicts of Interest

None declared.

References

1. Bondemark L, Ruf S. Randomized controlled trial: the gold standard or an unobtainable fallacy? Eur J Orthod 2015 Oct 01;37(5):457-461. [doi: 10.1093/ejo/cjv046] [Medline: 26136438]
2. Wieseler B, Neyt M, Kaiser T, Hulstaert F, Windeler J. Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy? BMJ 2023 Mar 02;380:e073100 [FREE Full text] [doi: 10.1136/bmj-2022-073100] [Medline: 36863730]
3. Chodankar D. Introduction to real-world evidence studies. Perspect Clin Res 2021;12(3):171-174 [FREE Full text] [doi: 10.4103/picr.picr_62_21] [Medline: 34386383]
4. Pihlstrom BL, Curran AE, Voelker HT, Kingman A. Randomized controlled trials: what are they and who needs them? Periodontol 2000 2012 Jun;59(1):14-31. [doi: 10.1111/j.1600-0757.2011.00439.x] [Medline: 22507057]
5. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
6. Morales DR, Arlett P. RCTs and real world evidence are complementary, not alternatives. BMJ 2023 Apr 03;381:736. [doi: 10.1136/bmj.p736] [Medline: 37011918]
7. Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative, Franklin JM, Desai RJ, Feldman W, et al. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. JAMA 2023 Apr 25;329(16):1376-1385 [FREE Full text] [doi: 10.1001/jama.2023.4221] [Medline: 37097356]
8. Daniel GS, Bryan J, McClellan M, Romine M, Frank K, Silcox C. Characterizing RWD quality and relevancy for regulatory purposes. Duke-Margolis Center. 2018 Oct 01. URL: https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing_rwd.pdf [accessed 2024-09-20]
9. Berger M, Daniel G, Frank K, Hernandez A, McClellan M, Okun S, et al. A framework for regulatory use of real-world evidence. Duke-Margolis Center. 2017 Sep 13. URL: https://healthpolicy.duke.edu/sites/default/files/2020-08/rwe_white_paper_2017.09.06.pdf [accessed 2024-09-20]
10. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 05;22(1):287 [FREE Full text] [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
11. Innovative health initiative. European Commission. URL: https://research-and-innovation.ec.europa.eu/research-area/health/innovative-health-initiative_en [accessed 2024-09-25]
12. Innovative Health Initiative launches first five projects. Innovative Health Initiative. URL: https://www.ihi.europa.eu/news-events/newsroom/innovative-health-initiative-launches-first-five-projects [accessed 2024-09-25]

XSL•FO
RenderX

13. Verma A, Bhattacharya P, Patel Y, Shah K, Tanwar S, Khan B. Data localization and privacy-preserving healthcare for big data applications: architecture and future directions. In: Proceedings of Emerging Technologies for Computing, Communication and Smart Cities. 2021 Presented at: ETCCS 2021; August 21-22, 2021; Punjab, India. [doi: 10.1007/978-981-19-0284-0_18]
14. Näher AF, Vorisek CN, Klopfenstein SA, Lehne M, Thun S, Alsalamah S, et al. Secondary data for global health digitalisation. Lancet Digit Health 2023 Feb;5(2):e93-101 [FREE Full text] [doi: 10.1016/S2589-7500(22)00195-9] [Medline: 36707190]
15. Togo K, Yonemoto N. Real world data and data science in medical research: present and future. Jpn J Stat Data Sci 2022 Apr 13;5(2):769-781 [FREE Full text] [doi: 10.1007/s42081-022-00156-0] [Medline: 35437515]
16. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021 Mar;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
17. Savage N. Synthetic data could be better than real data. Nature 2023 Apr 27. [doi: 10.1038/d41586-023-01445-8] [Medline: 37106108]
18. Nikolentzos G, Vazirgiannis M, Xypolopoulos C, Lingman M, Brandt EG. Synthetic electronic health records generated with variational graph autoencoders. NPJ Digit Med 2023 Apr 29;6(1):83 [FREE Full text] [doi: 10.1038/s41746-023-00822-x] [Medline: 37120594]
19. Bietz MJ, Bloss CS, Calvert S, Godino JG, Gregory J, Claffey MP, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc 2016 Apr;23(e1):e42-e48 [FREE Full text] [doi: 10.1093/jamia/ocv118] [Medline: 26335984]
20. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest 2020 Feb 03;130(2):565-574 [FREE Full text] [doi: 10.1172/JCI129197] [Medline: 32011317]
21. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic data -- what, why and how? arXiv. Preprint posted online on May 6, 2022 2024. [doi: 10.48550/arXiv.2205.03257]
22. Raghunathan TE. Synthetic data. Annu Rev Stat Appl 2021 Mar 07;8(1):129-140. [doi: 10.1146/annurev-statistics-040720-031848]
23. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health 2023 Jan;2(1):e0000082 [FREE Full text] [doi: 10.1371/journal.pdig.0000082] [Medline: 36812604]
24. Domingo-Ferrer J, Montes F. Privacy in Statistical Databases: International Conference, PSD 2022, Paris, France, September 21–23, 2022, Proceedings. Cham, Switzerland: Springer; 2022.
25. Hernandez-Matamoros A, Fujita H, Perez-Meana H. A novel approach to create synthetic biomedical signals using BiRNN. Inf Sci 2020 Dec;541:218-241. [doi: 10.1016/j.ins.2020.06.019]
26. Sano N. Synthetic data by principal component analysis. In: 2020 International Conference on Data Mining Workshops. 2020 Presented at: ICDMW; November 17-20, 2020; Sorrento, Italy. [doi: 10.1109/ICDMW51313.2020.00023]
27. ONS methodology working paper series number 16 - synthetic data pilot. Office for National Statistics. URL: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot [accessed 2024-12-02]
28. Nikolenko SI. Synthetic Data for Deep Learning. Cham, Switzerland: Springer; 2021.
29. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023 Oct 09;6(1):186 [FREE Full text] [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]
30. Dahmen J, Cook D. SynSys: a synthetic data generation system for healthcare applications. Sensors (Basel) 2019 Mar 08;19(5):1181 [FREE Full text] [doi: 10.3390/s19051181] [Medline: 30857130]
31. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
32. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Proceedings of the International Conference on Intelligent Computing. 2005 Presented at: ICIC 2005; August 23-26, 2005; Hefei, China. [doi: 10.1007/11538059_91]
33. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 2004 Jun;6(1):20-29. [doi: 10.1145/1007730.1007735]
34. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the IEEE International Joint Conference on Neural Networks. 2008 Presented at: IJCNN 2008; June 1-8, 2008; Hong Kong, China. [doi: 10.1109/ijcnn.2008.4633969]
35. Torres FR, Carrasco-Ochoa JA, Martínez-Trinidad JF. SMOTE-D a deterministic version of SMOTE. In: Proceedings of the 8th Mexican Conference on Pattern Recognition. 2016 Presented at: MCPR 2016; June 22-25, 2016; Guanajuato, Mexico. [doi: 10.1007/978-3-319-39393-3_18]
36. Mukherjee M, Khushi M. SMOTE-ENC: a novel SMOTE-based method to generate synthetic data for nominal and continuous features. Appl Syst Innov 2021 Mar 02;4(1):18. [doi: 10.3390/asi4010018]
37. Figueira A, Vaz B. Survey on synthetic data generation, evaluation methods and GANs. Mathematics 2022 Aug 02;10(15):2733. [doi: 10.3390/math10152733]

XSL•FO
RenderX

38. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020 Oct 22;63(11):139-144. [doi: 10.1145/3422622]
39. Little RJ. Statistical analysis of masked data. J Off Stat 1993;9(2):407-426 [FREE Full text]
40. Ghosheh G, Li J, Zhu T. A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation measures and data sources. arXiv. Preprint posted online on March 14, 2022 2024 [FREE Full text]
41. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: state of the art in health care domain. Comput Sci Rev 2023 May;48:100546. [doi: 10.1016/j.cosrev.2023.100546]
42. Rashidian S, Wang F, Moffitt R, Garcia V, Dutt A, Chang W, et al. SMOOTH-GAN: towards sharp and smooth synthetic EHR data generation. In: Proceedings of the 18th International Conference on Artificial Intelligence in Medicine. 2020 Presented at: AIME 2020; August 25-28, 2020; Minneapolis, MN. [doi: 10.1007/978-3-030-59137-3_4]
43. Imtiaz S, Arsalan M, Vlassov V, Sadre R. Synthetic and private smart health care data generation using GANs. In: Proceedings of the 2021 International Conference on Computer Communications and Networks. 2021 Presented at: ICCCN 2021; July 19-22, 2021; Athens, Greece. [doi: 10.1109/icccn52240.2021.9522203]
44. Abedi M, Hempel L, Sadeghi S, Kirsten T. GAN-based approaches for generating structured data in the medical domain. Appl Sci 2022 Jul 13;12(14):7075. [doi: 10.3390/app12147075]
45. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. In: Proceedings of the IEEE 15th International Symposium on Biomedical Imaging. 2018 Presented at: ISBI 2018; April 4-7, 2018; Washington, DC. [doi: 10.1109/isbi.2018.8363576]
46. Torfi A, Fox EA. CorGAN: correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. arXiv. Preprint posted online on January 25, 2020 2024. [doi: 10.48550/arXiv.2001.09346]
47. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 2020 May 07;20(1):108 [FREE Full text] [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]
48. Dankar FK, Ibrahim M. Fake it till you make it: guidelines for effective synthetic data generation. Appl Sci 2021 Feb 28;11(5):2158. [doi: 10.3390/app11052158]
49. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018 Mar 01;25(3):230-238 [FREE Full text] [doi: 10.1093/jamia/ocx079] [Medline: 29025144]
50. Templ M, Meindl B, Kowarik A, Dupriez O. Simulation of synthetic complex data: the R package simPop. J Stat Softw 2017;79(10):1-38. [doi: 10.18637/jss.v079.i10]
51. Ping H, Stoyanovich J, Howe B. DataSynthesizer: privacy-preserving synthetic datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management. 2017 Presented at: SSDBM '17; June 27-29, 2017; Chicago, IL. [doi: 10.1145/3085504.3091117]
52. Attribution-Non-Commercial 4.0 International (CC BY-NC 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by-nc/4.0/deed.en [accessed 2024-12-19]
53. Attribution 4.0 International (CC BY4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/deed.en [accessed 2024-12-19]
54. Kalkman S, Mostert M, Udo-Beauvisage N, van Delden JJ, van Thiel GJ. Responsible data sharing in a big data-driven translational research platform: lessons learned. BMC Med Inform Decis Mak 2019 Dec 30;19(1):283 [FREE Full text] [doi: 10.1186/s12911-019-1001-y] [Medline: 31888593]
55. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inform 2020 Jul;107:103421 [FREE Full text] [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]
56. Aneja S, Avesta A, Xu H, Machado LO. Clinical informatics approaches to facilitate cancer data sharing. Yearb Med Inform 2023 Aug;32(1):104-110 [FREE Full text] [doi: 10.1055/s-0043-1768721] [Medline: 37414028]
57. Li B, Tsui R. How to improve the reuse of clinical data-- openEHR and OMOP CDM. J Phys Conf Ser 2020 Oct 01;1624:032041. [doi: 10.1088/1742-6596/1624/3/032041]
58. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
59. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
60. Wood FEJ, Fitzsimmons MJ. Clinical data interchange standards consortium (CDISC) standards and their implementation in a clinical data management system. Drug Inf J 2001 Dec 30;35:853-862. [doi: 10.1177/009286150103500323]
61. Kawai AT, Martin D, Henrickson SE, Goff A, Reidy M, Santiago D, et al. Validation of febrile seizures identified in the sentinel post-licensure rapid immunization safety monitoring program. Vaccine 2019 Jul 09;37(30):4172-4176. [doi: 10.1016/j.vaccine.2019.05.042] [Medline: 31186192]
62. Sentinel common data model. Sentinel Initiative. URL: https://www.sentinelinitiative.org/sites/default/files/Sentinel%20Common%20Data%20Model_01102024.PNG [accessed 2024-03-02]

XSL•FO
RenderX

63. Ogunyemi OI, Meeker D, Kim HE, Ashish N, Farzaneh S, Boxwala A. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. Med Care 2013 Aug;51(8 Suppl 3):S45-S52. [doi: 10.1097/MLR.0b013e31829b1e0b] [Medline: 23774519]
64. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO research network virtual data warehouse: a public data model to support collaboration. EGEMS (Wash DC) 2014;2(1):1049 [FREE Full text] [doi: 10.13063/2327-9214.1049] [Medline: 25848584]
65. Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by-nc-nd/4.0/ [accessed 2024-12-19]
66. Toh S, Rasmussen-Torvik LJ, Harmata EE, Pardee R, Saizan R, Malanga E, et al. The national patient-centered clinical research network (PCORnet) bariatric study cohort: rationale, methods, and baseline characteristics. JMIR Res Protoc 2017 Dec 05;6(12):e222 [FREE Full text] [doi: 10.2196/resprot.8323] [Medline: 29208590]
67. Yu Y, Zong N, Wen A, Liu S, Stone DJ, Knaack D, et al. Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration. J Biomed Inform 2022 Mar;127:104002 [FREE Full text] [doi: 10.1016/j.jbi.2022.104002] [Medline: 35077901]
68. Hossain MS. Design and implementation of serverless architecture for i2b2 on AWS cloud and Snowflake data warehouse. University of Missouri. 2023. URL: https://mospace.umsystem.edu/xmlui/handle/10355/96163 [accessed 2024-09-20]
69. Carus J, Nürnberg S, Ückert F, Schlüter C, Bartels S. Mapping cancer registry data to the episode domain of the observational medical outcomes partnership model (OMOP). Appl Sci 2022 Apr 15;12(8):4010. [doi: 10.3390/app12084010]
70. Makadia R, Ryan PB. Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. EGEMS (Wash DC) 2014;2(1):1110 [FREE Full text] [doi: 10.13063/2327-9214.1110] [Medline: 25848597]
71. Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart JB, et al. Transforming anesthesia data into the observational medical outcomes partnership common data model: development and usability study. J Med Internet Res 2021 Oct 29;23(10):e29259 [FREE Full text] [doi: 10.2196/29259] [Medline: 34714250]
72. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR. A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. Stud Health Technol Inform 2017;245:887-891 [FREE Full text] [Medline: 29295227]
73. Abd Al-Rahman SQ, Hasan EH, Sagheer AM. Design and implementation of the web (extract, transform, load) process in data warehouse application. IAES Int J Artif Intell 2023 Jun 01;12(2):765. [doi: 10.11591/ijai.v12.i2.pp765-775]
74. Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by-sa/4.0/ [accessed 2024-12-19]
75. Beauvais B, Kruse CS, Fulton L, Shanmugam R, Ramamonjiarivelo Z, Brooks M. Association of electronic health record vendors with hospital financial and quality performance: retrospective data analysis. J Med Internet Res 2021 Apr 14;23(4):e23961 [FREE Full text] [doi: 10.2196/23961] [Medline: 33851924]
76. Kirrmann S, Gainey M, Röhner F, Hall M, Bruggmoser G, Schmucker M, et al. Visualization of data in radiotherapy using web services for optimization of workflow. Radiat Oncol 2015 Jan 20;10(1):22 [FREE Full text] [doi: 10.1186/s13014-014-0322-3] [Medline: 25601225]
77. Huang HK, Taira RK. Infrastructure design of a picture archiving and communication system. AJR Am J Roentgenol 1992 Apr;158(4):743-749. [doi: 10.2214/ajr.158.4.1546584] [Medline: 1546584]
78. Sinard J. Pathology LIS: relationship to institutional systems. In: Practical Pathology Informatics. New York, NY: Springer; 2006:173-206.
79. Bansal SK, Kagemann S. Integrating big data: a semantic extract-transform-load framework. Computer 2015 Mar;48(3):42-50. [doi: 10.1109/mc.2015.76]
80. Yangui R, Nabli A, Gargouri F. ETL based framework for NoSQL warehousing. In: Proceedings of the 14th European, Mediterranean, and Middle Eastern Conference. 2017 Presented at: EMCIS 2017; September 7-8, 2017; Coimbra, Portugal. [doi: 10.1007/978-3-319-65930-5_4]
81. Baghal A. Leveraging graph models to design acute kidney injury disease research data warehouse. In: Proceedings of the Sixth International Conference on Social Networks Analysis, Management and Security. 2019 Presented at: SNAMS 2019; October 22-25, 2019; Granada, Spain. [doi: 10.1109/snams.2019.8931838]
82. Burrows E, Razzaghi H, Utidjian L, Bailey L. Standardizing clinical diagnoses: evaluating alternate terminology selection. AMIA Jt Summits Transl Sci Proc 2020 May;2020:71-79 [FREE Full text] [Medline: 32477625]
83. Bidgood WDJ, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Inform Assoc 1997 May 01;4(3):199-212. [doi: 10.1136/jamia.1997.0040199]
84. Dhavle AA, Rupp MT. Towards creating the perfect electronic prescription. J Am Med Inform Assoc 2015 Apr;22(e1):e7-12. [doi: 10.1136/amiajnl-2014-002738] [Medline: 25038197]
85. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020 Sep 14;3(1):119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]
86. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. J Healthc Inform Res 2021;5(1):1-19 [FREE Full text] [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]

XSL•FO
RenderX

87. Krishnan S, Anand AJ, Srinivasan R, Kavitha R, Suresh S. Handbook on Federated Learning: Advances, Applications and Opportunities. Boca Raton, FL: CRC Press; 2023.
88. Joshi M, Pal A, Sankarasubbu M. Federated learning for healthcare domain - pipeline, applications and challenges. ACM Trans Comput Healthcare 2022 Nov 03;3(4):1-36. [doi: 10.1145/3533708]
89. Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: a systematic literature review. ACM Trans Internet Technol 2021 Jun 02;21(2):1-31. [doi: 10.1145/3412357]
90. Olatunji IE, Rauch J, Katzensteiner M, Khosla M. A review of anonymization for healthcare data. Big Data 2022 Mar 10. [doi: 10.1089/big.2021.0169] [Medline: 35271377]
91. Dhade P, Shirke P. Federated learning for healthcare: a comprehensive review. Eng Proc 2023;59(1):230. [doi: 10.3390/engproc2023059230]
92. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol 2022 May 03;13(4):1-23. [doi: 10.1145/3501813]
93. Cremonesi F, Planat V, Kalokyri V, Kondylakis H, Sanavia T, Miguel Mateos Resinas V, et al. The need for multimodal health data modeling: a practical approach for a federated-learning healthcare platform. J Biomed Inform 2023 May;141:104338 [FREE Full text] [doi: 10.1016/j.jbi.2023.104338] [Medline: 37023843]
94. Guendouzi BS, Ouchani S, EL Assaad H, EL Zaher M. A systematic review of federated learning: challenges, aggregation methods, and development tools. J Netw Comput Appl 2023 Nov;220:103714. [doi: 10.1016/j.jnca.2023.103714]
95. Berghout T, Benbouzid M, Bentrcia T, Lim WH, Amirat Y. Federated learning for condition monitoring of industrial processes: a review on fault diagnosis methods, challenges, and prospects. Electronics 2022 Dec 29;12(1):158. [doi: 10.3390/electronics12010158]
96. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on February 17, 2016 2024. [doi: 10.1002/9781119845041.ch10]
97. Yang Q, Fan L, Yu H. Federated Learning: Privacy and Incentive. Cham, Switzerland: Springer; 2020.
98. Wang Y. Performance enhancement schemes and effective incentives for federated learning [thesis]. University of Ottawa. 2021 Nov 16. URL: https://ruor.uottawa.ca/items/fb7bb89a-0b36-4116-99c0-cd03e0c6517e [accessed 2024-12-09]
99. 2020 digital health and learning health system. National Academies. 2020 May. URL: https://www.nationalacademies.org/documents/link/LF5F46A0F4D8D3765F6DBFFA9DD3EA606B9CCD57CD98/file/DC3D040E01F6D31C3E3ECB934FB1EF25E407789B8DE2 [accessed 2024-10-10]
100. Gehrmann J, Herczog E, Decker S, Beyan O. What prevents us from reusing medical real-world data in research. Sci Data 2023 Jul 13;10(1):459 [FREE Full text] [doi: 10.1038/s41597-023-02361-2] [Medline: 37443164]
101. Liu C, Shao S, Liu C, Bennett GG, Prvu Bettger J, Yan LL. Academia-industry digital health collaborations: a cross-cultural analysis of barriers and facilitators. Digit Health 2019 Sep 26;5:2055207619878627 [FREE Full text] [doi: 10.1177/2055207619878627] [Medline: 31632684]
102. Awasthy R, Flint S, Sankarnarayana R, Jones RL. A framework to improve university–industry collaboration. J Industry Univ Collab 2020 Feb 23;2(1):49-62. [doi: 10.1108/jiuc-09-2019-0016]
103. Hingle M, Patrick H, Sacher PM, Sweet CC. The intersection of behavioral science and digital health: the case for academic-industry partnerships. Health Educ Behav 2019 Feb 24;46(1):5-9. [doi: 10.1177/1090198118788600] [Medline: 30041556]
104. Hallinan CM, Ward R, Hart GK, Sullivan C, Pratt N, Ng AP, et al. Seamless EMR data access: integrated governance, digital health and the OMOP-CDM. BMJ Health Care Inform 2024 Feb 21;31(1):e100953 [FREE Full text] [doi: 10.1136/bmjhci-2023-100953] [Medline: 38387992]
105. Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, et al. Scalable collaborative infrastructure for a learning healthcare system (SCILHS): architecture. J Am Med Inform Assoc 2014;21(4):615-620 [FREE Full text] [doi: 10.1136/amiajnl-2014-002727] [Medline: 24821734]
106. Gal MS, Lynskey O. Synthetic data: legal implications of the data-generation revolution. Iowa Law Rev 2024;109(3). [doi: 10.2139/ssrn.4414385]
107. Alloza C, Knox B, Raad H, Aguilà M, Coakley C, Mohrova Z, et al. A case for synthetic data in regulatory decision-making in Europe. Clin Pharmacol Ther 2023 Oct;114(4):795-801. [doi: 10.1002/cpt.3001] [Medline: 37441734]
108. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med 2022;1(1):e000167 [FREE Full text] [doi: 10.1136/bmjmed-2022-000167] [Medline: 36936569]
109. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open 2021 Apr 16;11(4):e043497 [FREE Full text] [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]
110. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. JMIR Med Inform 2020 Feb 20;8(2):e16492 [FREE Full text] [doi: 10.2196/16492] [Medline: 32130148]
111. Baumann J, Castelnovo A, Crupi R, Inverardi N, Regoli D. Bias on demand: a modelling framework that generates synthetic data with bias. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023 Presented at: FAccT '23; June 12-15, 2023; Chicago, IL. [doi: 10.1145/3593013.3594058]

XSL•FO
RenderX

112. Stadler T, Oprisanu B, Troncoso C. Synthetic data -- anonymisation groundhog day. arXiv. Preprint posted online on November 13, 2020 2024. [doi: 10.1109/msp.2004.9]
113. Draghi B, Wang Z, Myles P, Tucker A. Identifying and handling data bias within primary healthcare data using synthetic data generators. Heliyon 2024 Jan 30;10(2):e24164 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e24164] [Medline: 38288010]
114. Rosenblatt L, Liu X, Pouyanfar S, de Leon E, Desai A, Allen J. Differentially private synthetic data: applied evaluations and enhancements. arXiv. Preprint posted online on June 12, 2017 2024 [FREE Full text]
115. Platzer M, Reutterer T. Holdout-based empirical assessment of mixed-type synthetic data. Front Big Data 2021;4:679939 [FREE Full text] [doi: 10.3389/fdata.2021.679939] [Medline: 34268491]
116. Lu Y, Shen M, Wang H, Wang X, van Rechem C, Fu T, et al. Machine learning for synthetic data generation: a review. arXiv. Preprint posted online on February 8, 2023 2024. [doi: 10.48550/arXiv.2302.04062]
117. Tsao SF, Sharma K, Noor H, Forster A, Chen H. Health synthetic data to enable health learning system and innovation: a scoping review. Stud Health Technol Inform 2023 May 18;302:53-57. [doi: 10.3233/SHTI230063] [Medline: 37203608]
118. Guo A, Foraker RE, MacGregor RM, Masood FM, Cupps BP, Pasque MK. The use of synthetic electronic health record data and deep learning to improve timing of high-risk heart failure surgical intervention by predicting proximity to catastrophic decompensation. Front Digit Health 2020 Dec 7;2:576945 [FREE Full text] [doi: 10.3389/fdgth.2020.576945] [Medline: 34713050]
119. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP - a scoping review. Stud Health Technol Inform 2021 Sep 21;283:95-103. [doi: 10.3233/SHTI210546] [Medline: 34545824]
120. Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, et al. Opioid use, postoperative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study. Lancet Rheumatol 2019 Dec;1(4):e229-e236. [doi: 10.1016/S2665-9913(19)30075-X] [Medline: 38229379]
121. Lane JC, Weaver J, Kostka K, Duarte-Salles T, Abrahao MT, Alghoul H, et al. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. The Lancet Rheumatology 2020 Nov;2(11):e698-e711. [doi: 10.1016/s2665-9913(20)30276-9]
122. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019 Nov 16;394(10211):1816-1826 [FREE Full text] [doi: 10.1016/S0140-6736(19)32317-7] [Medline: 31668726]
123. Homayouni H, Ghosh S, Ray I. An approach for testing the extract-transform-load process in data warehouse systems. In: Proceedings of the 22nd International Database Engineering & Applications Symposium. 2018 Presented at: IDEAS '18; June 18-20, 2018; Villa San Giovanni, Italy. [doi: 10.1145/3216122.3216149]
124. Nwokeji JC, Matovu R. A systematic literature review on big data extraction, transformation and loading (ETL). In: Arai K, editor. Intelligent Computing. Cham, Switzerland: Springer; 2021.
125. Ahmadi N, Zoch M, Kelbert P, Noll R, Schaaf J, Wolfien M, et al. Methods used in the development of common data models for health data: scoping review. JMIR Med Inform 2023 Aug 03;11:e45116 [FREE Full text] [doi: 10.2196/45116] [Medline: 37535410]
126. Mbunge E, Muchemwa B, Jiyane S, Batani J. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. Global Health J 2021 Dec;5(4):169-177. [doi: 10.1016/j.glohj.2021.11.008]
127. Ju C, Zhao R, Sun J, Wei X, Zhao B, Liu Y, et al. Privacy-preserving technology to help millions of people: federated prediction model for stroke prevention. arXiv. Preprint posted online on June 15, 2020 2024. [doi: 10.48550/arXiv.2006.10517]
128. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Demiralp, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015 Oct;34(10):1993-2024 [FREE Full text] [doi: 10.1109/TMI.2014.2377694] [Medline: 25494501]
129. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018 Apr;112:59-67 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.01.007] [Medline: 29500022]
130. Wu X, Liang Z, Wang J. FedMed: a federated learning framework for language modeling. Sensors (Basel) 2020 Jul 21;20(14):4048 [FREE Full text] [doi: 10.3390/s20144048] [Medline: 32708152]
131. Honkela A, Das M, Nieminen A, Dikmen O, Kaski S. Efficient differentially private learning improves drug sensitivity prediction. Biol Direct 2018 Feb 06;13(1):1 [FREE Full text] [doi: 10.1186/s13062-017-0203-4] [Medline: 29409513]
132. Valdes G, Simone CB2, Chen J, Lin A, Yom SS, Pattison AJ, et al. Clinical decision support of radiotherapy treatment planning: a data-driven machine learning strategy for patient-specific dosimetric decision making. Radiother Oncol 2017 Dec;125(3):392-397. [doi: 10.1016/j.radonc.2017.10.014] [Medline: 29162279]
133. Lyu L, Yu H, Ma X, Chen C, Sun L, Zhao J, et al. Privacy and robustness in federated learning: attacks and defenses. IEEE Trans Neural Netw Learn Syst 2024 Jul;35(7):8726-8746. [doi: 10.1109/TNNLS.2022.3216981] [Medline: 36355741]
134. Yang Q, Huang A, Fan L, Chan CS, Lim JH, Ng KW, et al. Federated learning with privacy-preserving and model IP-right-protection. Mach Intell Res 2023 Jan 10;20(1):19-37. [doi: 10.1007/s11633-022-1343-2]

XSL•FO
RenderX

135. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Nitin Bhagoji A, et al. Advances and open problems in federated learning. FNT Mach Learn 2021;14(1–2):1-210. [doi: 10.1561/2200000083]
136. Abdelmoniem AM, Ho CY, Papageorgiou P, Canini M. A comprehensive empirical study of heterogeneity in federated learning. IEEE Internet Things J 2023 Aug 15;10(16):14071-14083 [FREE Full text] [doi: 10.1109/JIOT.2023.3250275]
137. Truong N, Sun K, Wang S, Guitton F, Guo Y. Privacy preservation in federated learning: an insightful survey from the GDPR perspective. Comput Secur 2021 Nov;110:102402. [doi: 10.1016/j.cose.2021.102402]
138. Abreha HG, Hayajneh M, Serhani MA. Federated learning in edge computing: a systematic survey. Sensors (Basel) 2022 Jan 07;22(2):450 [FREE Full text] [doi: 10.3390/s22020450] [Medline: 35062410]
139. Kumar Y, Singla R. Federated learning systems for healthcare: perspective and recent progress. In: Rehman MH, Gaber MM, editors. Federated Learning Systems. Cham, Switzerland: Springer; 2021.
140. Evans N, Miklosik A, Du JT. University-industry collaboration as a driver of digital transformation: types, benefits and enablers. Heliyon 2023 Oct;9(10):e21017 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e21017] [Medline: 37867890]
141. Rybnicek R, Königsgruber R. What makes industry–university collaboration succeed? A systematic review of the literature. J Bus Econ 2018 Sep 12;89(2):221-250. [doi: 10.1007/s11573-018-0916-6]

Abbreviations

CDISC: Clinical Data Interchange Standards Consortium
CDM: common data model
EHR: electronic health record
EMR: electronic medical record
ETL: extract, transform, and load
FDA: Food and Drug Administration
FL: federated learning
GAN: generative adversarial network
HIPAA: Health Insurance Portability and Accountability Act
HIT: health IT
ML: machine learning
OMOP: Observational Medical Outcomes Partnership
PADARSER: publicly available data approach to the realistic synthetic electronic health record
RCT: randomized controlled trial
RWD: real-world data

Edited by G Eysenbach; submitted 26.03.24; peer-reviewed by V Rudrapatna, A Kremer, SC You, M Field; comments to author 15.09.24; revised version received 04.10.24; accepted 30.11.24; published 20.12.24.

Please cite as:
Austin JA, Lobo EH, Samadbeik M, Engstrom T, Philip R, Pole JD, Sullivan CM
Decades in the Making: The Evolution of Digital Health Research Infrastructure Through Synthetic Data, Common Data Models, and Federated Learning
J Med Internet Res 2024;26:e58637
URL: https://www.jmir.org/2024/1/e58637
doi:10.2196/58637
PMID:39705072

©Jodie A Austin, Elton H Lobo, Mahnaz Samadbeik, Teyl Engstrom, Reji Philip, Jason D Pole, Clair M Sullivan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX